Visual Computing Design for Explainable, Spatially-aware Machine Learning

Andrew Wentzel
M.S., University of Illinois Chicago, 2019
B.Eng., The Cooper Union, 2016

Dissertation

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
at the Graduate College of the University of Illinois at Chicago, 2025

Chicago, Illinois

Defense Committee:
Dr. G.Elisabeta Marai, *Chair and Advisor*
Dr. Fabio Miranda
Dr. Xinhua Zhang
Dr. Guadalupe Canahuate, University of Iowa
Dr. Renata Raidou, TU Wien

Acknowledgments

I would like to thank the other members of the Electronic Visualization laboratory. I would like specifically to thank Juan Trelles Trabucco Trelles for helping to teach me most of what I know about software engineering and my fellow inmates Carla Floricel Sosea, Nafiul Nipu, and Sanjana Srabanti for their support through the years, as well as Dana and Lance Long at the EVL for their very patient logistical and technical support.

I would also like to thank my Advisor and the team at the MD Anderson Cancer Center and the University of Iowa, namely Guadalupe Canahuate, Xinhua Zhang, David Fuller, Abdallah S.R. Mohamed, Mohamed Naser, and Serageldin Attia, for being excellent collaborators in our work and often serving as the "humans" for the "Human-Centered Design" part of my work.

Finally, I would like to thank my Mother, Lisa Wentzel, my late Grandparents Patrick and Barbara Obrien, and my brothers Matt and Jake for their emotional support and willingness to drive me places. Finally, I want to thank my cat Patches for carefully supervising me when working at home.

Contributions of Authors

This thesis consists of work from 5 published papers. For all papers, G.E. Marai helped with the research direction, and the drafting and editing of the paper and conference presentation as my advisor.

1. *Chapter 2* In "Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration" [345], the interface was based on the master's thesis work of Peter Hanula, who designed most of the stylized 3-d radiation dose plot and controls for selecting organs and patients, as well as did some preliminary work experimenting with similarity methods; Tim Luciani helped in the design of some of the figures and editing the final paper; Baher Elgohari, and Hesham Elhalawani helped with providing the data; Guadalupe Canahuate, David Vock and Clifton David Fuller provided feedback on the interface and analysis methodology.

2. *Chapter 3* In "MOTIV: Visual Exploration of Moral Framing in Social Media" [348], Lauren Levine and Andrew Rojecki provided annotations and moral foundation labels for the tweets; Vipul Dhariwal, Abari Bhattacharya, and Barbara Di Eugenio helped create the Twitter dataset and identify the process for finding tweet stance and relevance; and Elena Zheleva and Zahra Fatemi helped with sentiment analysis. All coauthors helped with requirements gathering and providing feedback for the interface.

3. *Chapter 4* In "DASS Good: Explainable Data Mining of Spatial Cohort Data" [344], Carla Floricel helped with design choices on the interface, namely layout and color choice, as well as helped edit the paper; Lisanne Van Dijk helped gathr and preprocessing the patient dose-volume data; and Guadalupe Canahuate, Mohamed A Naser, Abdallah S.R. Mohamed, and C.D. fuller provided design requirements and feedback for the interface and helped with drafting the clinical paper.

4. *Chapter 5* In "Explainable Spatial Clustering: Leveraging Spatial Data in Radiation Oncology" [343], all listed coauthors helped with requirements gathering and feedback,

as well as drafting clinical papers. Guadalupe Canahuate also helped test different clustering methods for the associated papers. The graph-based lymph node designs and dendrograms were designed by Tim Luciani.

5. *Chapter 6* In "DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer" [342], the reinforcement learning models and formulation of the decision-making sequence were extensions of models created by Xinhua Zhang's group. Serageldin Attia helped with the data collection and feedback on the case studies. Serageldin Attia and Clifton David Fuller provided the initial interviews for the requirements gathering, as well as getting us access to other clinicians for the user workshop. All coauthors helped with providing feedback on the interface designs.

Contents

List of Figures

List of Tables

Summary

Research in visual computing (VC) focuses on integrating visual representations of datasets with computing systems, thus leveraging the complementary abilities of humans and machines. The resulting VC systems may include machine learning (ML) techniques, which have remarkable abilities in terms of processing vast amounts of data. However, creating VC+ML systems that can operate on spatio-temporal data is a largely unexplored, difficult topic. Challenges in this area include: capturing domain characterizations for spatial VC+ML problems, in particular when serving both ML model builders and ML model clients; creating explainable VC+ML models that can operate on spatial data, including measures that can capture spatial structure similarity; designing visual encodings for ML models that use spatial information and that enhance the modeler and client understanding of the model; and measuring the effect of deploying such VC+ML systems in practice.

Based on several multi-year collaborations, in this dissertation I will document several instances where spatial and explainable VC+ML solutions are required, and what role spatiality plays in these situations. Throughout this process, I extract and describe design requirements for explainable spatial VC+ML systems, and I identify necessary VC advances, including spatially-aware similarity measures and spatial visual explainability. First, I construct spatially-aware similarity measures to support ML problems in precision oncology, patient risk stratification in data mining, explainable interactive regression modeling for hypothesis testing in social science, and digital twin interfaces for cancer treatment planning. Second, I introduce novel encodings for 3-dimensional and geospatial data and explore how to integrate such encodings with ML model explanations. Finally, I implement and evaluate several VC+ML systems and several resulting models that are explicitly designed to integrate domain-specific structural spatial information and unsupervised ML. Throughout this process, I also articulate design lessons related to model actionability.

Chapter 1

Introduction

1.1 Motivation

Many real-world machine learning (ML) applications require combining the strengths of human analysts who have specialized knowledge, with computational systems that excel at statistical analysis and large-scale data processing. This is especially true in the case of statistical modeling with spatial data, where information about the coordinates of each data item benefits from domain expert knowledge. For example, in many location-dependent cancers, the location of the tumors and the specific pattern in which the disease spreads have an impact on the patient's treatment, such as when applying radiation therapy, as well as on the treatment outcomes. At the same time, growing clinical data repositories should make it possible to leverage treatment and outcome data from past patients who had similar characteristics. Identifying similar patients and supporting physicians in reasoning about them would require a blend of spatially-aware machine learning methods and domain expert input.

Computing with location-dependent oncology data is an example of a problem that relies on spatial data, or data where items are tied to either 2 to 3-dimensional coordinates, or have an intrinsic 2-3 dimensional geometry. Visual computing (VC) studies computing with spatial data, often in conjunction with domain-expert interaction. VC research often integrates image analysis, computer graphics, and data visualization approaches to help domain scientists reason about these spatial data. As shown in the oncology example above, the acquisition of ever larger datasets, in conjunction with advances in data mining and machine learning (ML) techniques, have motivated a shift towards combining the strengths

of VC and ML. However, traditional research in VC+ML has focused on abstract data without spatial coordinates. In contrast, spatiality in VC+ML, which is often essential when dealing with human-subject data such as medical images or population geospatial data, is under-explored.

The specific research question this dissertation aims to address are:

How do we integrate spatial data into explainable VC+ML systems?

This involves several sub-challenges:

1. Domain Characterization: What role does spatial data play in VC+ML? How is spatial ML being used by domain experts?
2. What strategies can we use for the design of models for spatial VC problems that consider requirements from both ML model builders and model clients in a collaborative setting?
3. How do we model similarity between sets of spatial features?
4. How do we design visual encodings for these spatial datasets, and how do we explain these spatial machine learning model predictions to clients?
5. How do we measure the effect of deploying these spatial VC+ML systems in practice?

To answer this, I will detail findings several design studies that include: 1) domain characterization for VC systems for spatial data that integrate ML strategies; 2) strategies for the design of spatial VC problems and novel spatial encodings that consider requirements from both ML model builders and model clients in collaborative settings; 3) the design of integrated VC+ML systems that support interactive exploration of spatial data and model refinement; 4) lessons on how to design spatially-aware ML algorithms that consider context-dependent spatial features as well as explainability requirements; and 5) evaluations of these

integrated systems in practice through qualitative feedback, quantitative gains in model performance, and insights not possible without the use of custom VC+ML solutions.

1.2 Contributions

This dissertation examines the issue of spatial data integration in VC+ML systems. First, I examine and characterize the use context of spatial VC+ML systems in several application domains, focusing on both front-end requirements and modeling requirements as a cohesive unit. I then describe the design and implementation of several spatially-aware VC+ML systems, with a focus on algorithms for supporting spatially-aware ML, unsupervised learning, and reinforcement learning problems. Next, I detail how these models are integrated into VC systems through the design of visual encodings that capture model behavior as well as important spatial structure. I then discuss design considerations from both a model building perspective, and how this can be extended to domain experts and end users. Finally, I measure the effect of these spatial VC+ML systems in practice, through the use of domain expert feedback, model improvement, and results otherwise not accessible through standard nonspatial approaches.

The remainder of these chapters are structured by domain application, as described below.

1. Chapter 2 [345] presents a VC+ML system in the context of radiation therapy (RT) planning in oncology. After characterizing the application domain, the chapter discusses the context of predicting the treatment plans for a new patient using CT scan data and plans from an existing patient cohort. Central to this work is the introduction of a novel topological similarity measure TSSIM to support unsupervised ML over the spatial data, and the use of a custom stylized 3-dimensional plots for visualizing a simplified representation of each patient's spatial RT plan. We apply this approach to cohort stratification and show this system can automatically identify similar patients, and use their spatial data to predict the RT plan for a new patient.

2. Chapter 3 introduces a VC+ML system (MOTIV) [347, 348] for analyzing social me-

dia microblog data in social-science. MOTIV leverages geospatial data and generalized additive models (GAMs), a type of statistical modeling. The chapter characterizes the application domain, then our use of custom data visual encodings and interactive modeling to help discover patterns between large-scale geospatial information and complex temporal patterns. We evaluate MOTIV through qualitative and quantitative feedback, as well as published insights that were not attainable through our collaborators' standard approaches.

3. Chapter 4 describes a VC+ML system (DASS) [344] for designing unsupervised ML models (clustering and stratification) to predict symptom severity in head and neck cancer patients treated with radiation therapy. The chapter provides a domain characterization for the ML modeling process, and the spatial data underneath. It then describes 2D visual maps that capture important 3D spatiality to support model explanations during model development, as well as the design of "model explanations" meant to reduce complex models such as clustering to more familiar and simplified models. We evaluate this system through a quantitative and qualitative analysis.

4. Chapter 5 reflects on the model development of spatial clusters for clinical oncology, and on the process of explaining these spatial models to biomedical researchers [343]. We distill general design goals for clinical spatial cluster analysis.

5. Chapter 6 presents a clinical decision support system that allows clinicians to analyze nuanced risk models for new patients and decide on an optimal treatment plan, with a focus on model explainability for supervised deep learning models with spatial and temporal components. This work is specifically targeted at use by domain experts in practical applications and not domain experts.

6. Chapter 7 summarizes my work and discusses possible future research opportunities.

Chapters 2, 4, 5, and 6 are completed, published work that came out of a collaboration with the MD Anderson Cancer Center. Chapter 3 is a published work using social media

data in collaboration with social scientists and natural language processing researchers at the University of Illinois-Chicago.

1.3 Background and Terminology

1.3.1 Terminology

Precision Medicine seeks to identify cohorts of similar patients, and examine which treatments maximize survival while minimizing side effects in that cohort, and prescribe a similar treatment for the new patients based on the subgroup they belong to. A core component of this paradigm is the ability to identify similar patients.

Personalized Medicine is medicine designed specifically for a specific patient. This is often used synonymously with Precision Medicine, but here refers to, for example, treatment based on parametric models that consider only the current patient's diagnostic data. This is often considered the "gold standard" of treatment, but is generally at higher risk of relying on models that overfit due to the complexity of creating models that can capture the complex nature of patient health.

Digital Twins are a general concept for the virtual representation of a real world object or process that can be used to simulate and predict behavior. Standard applications of digital twins are models used for building management systems and simulations of engineering and climate systems. For this work, we mainly discuss digital twins in the concept of simulating individuals in the context of personalized medicine.

Decision Support Tools, in the context of medicine, are computation systems that are designed to provide clinicians with personalized information about a patient, to provide insight that improves their ability to make treatment decisions. In most contexts, this refers specifically to applications of AI tools combined with interfaces, where personalized patient outcome predictions or treatment recommendations are shown to a clinician, who makes the final decision on the treatment for the patient [111].

Supervised machine learning refers to a broad class of statistical methods where models are made that predict a specific outcome given a set of input data, and we explicitly use

the outcome in the data used for training in the development of the model. For most cases discussed in this dissertation, I consider *classification* problems, where we build models that predict the relative likelihood of a set of yes/no outcomes. Common methods used for classification are logistic regression, decision trees, and neural networks.

Unsupervised machine learning is ML that does not explicitly use a training target in the data, and instead automatically finds patterns in the data. Common methods include clustering, dimensionality reduction, and rule mining. K-nearest-neighbors (KNNs) are models that rely on finding similar data points in the training data to a new input, and are often considered to be semi-supervised machine learning as they do not fit a model to the data.

Explainability is a broad concept in machine learning. For the purpose of this dissertation explainability that refers the ability to allow a user to understand either 1) The factors that the model considers when making a decision in general, 2) The most important factors that influenced an individual model decision, 3) the minimal set of factors that would lead to a different decision for an input (counterfactual), or 4) A logical process or set of rules that approximates the decision logic of the model that is understandable to a lay-user with high-fidelity, for an individual decision.

Spatial Data refers to data where each data item has at least one of two different features: an associated coordinate in an n-dimensional Euclidean space (usually 2 or 3-dimensional), or an associated geometry. Data with a coordinate but no associated geometry is considered to be "point-like". For this dissertation, we focus on spatial data with associated geometries, where point-like data is often aggregated within these regions.

1.3.2 Explainable ML

Explainable Machine Learning encompasses a wide range of concepts, and there is still no widely adopted vocabulary for the techniques and evaluation methods. Several recent review papers have attempted to present a semantic classification of different types of user interpretability with regard to the methodology and models used [81, 306, 379], while other works have proposed frameworks for discussing explainable AI in terms of the goals and

end-users [4, 16, 168, 268]. Zachary Lipton described the notion of "transparency" that is most closely analogous to the popular notion of model interpretability, where they break methods down into 3 components: (1) simulatability, the ability of a person to "reasonably" reproduce the results of the model; (2) decomposability, where each of the model's input, internal parameters, and mapping function can be individually described; (3) and algorithmic transparency, which describes how well we can describe different properties of the underlying algorithm, such as mathematical guarantees of convergence [179]. In response to the GDPR's requirement of a "right to an explanation", some work has come out that attempts to establish guidelines for an acceptable explanation. Doshi et al. [79] proposed that a reasonable explanation should be able to provide either (1) a layman explanation of the "factors" used and their relative contribution to a final decision, or (2) answers to questions related to which factors would need to change to alter a decision (counterfactuals).

An adjacent concept to interpretability is the growing interest in "trust" in AI. Jian et al. proposed a machine learning "trust score" [139], which measured the agreement between predictions of nearest neighbor predictions. Holliday et al. [124] found evidence that explanations were important in developing user trust in intelligent systems, but when equivalent explanations are provided, model performance is more important than model transparency, and human-centered research has shifted towards looking at trustworthy explanations, rather than trustworthy models. User-centered evaluations have emerged that also explore different factors that affect how much users trust an explanation of a model, such as how different aspects of recommender systems affect perceived trust [21, 160]. Recently, Davis et al. proposed a framework for empirical experiments in trustworthy machine learning that rely on measures of utility - or how well an explanation supports human decision-making [71], rather than trust. They highlight the issue that over-trusting faulty explanations can be detrimental, and highlight the importance of "appropriate trust". We will broadly categorize AI models as globally interpretable, locally interpretable, and black-box models for this work.

Interpretable Models

Globally interpretable models are those that can be understood in a global context, and are often referred to as white-box or glass-box models. These include linear and logistic regression methods, which have a clear input-output relationship that can be understood via a monotonic relationship between input and output values. The coefficients of these models are also usable as surrogates for covariate effect size [235], making them widely adopted in controlled experiments. Globally interpretable models also have the benefit of allowing for different degrees of global inference on the data, which is important for applications where one of the goals is model fairness or knowledge discovery.

Rule-based systems and decision trees are often cited as one of the most transparent models, as they can be easily described via a global deliberative process, and are often visualized as a flowchart style tree. Naive Bayes methods are also often considered inherently explainable, although they grow very complex with high-dimensional and continuous variables, making them approach a 'black box' model for many applications. Recently, more interest has been given to generalized additive models, where the final prediction is treated as a linear combination of functions for each input feature. These models can be seen as a generalization of linear regression, allowing for more flexibility in the models while maintaining most of the inherent interpretability, as each feature's contribution to a prediction can be treated as a 1 or 2-dimensional shape function depending on if interaction terms are considered [183].

Locally interpretable models encompass those that can be inherently explained given a specific set of data. The most evident example here is k-nearest-neighbors, where we can completely explain a prediction by showing the neighbors of a particular data point, but a global explanation of the model is elusive. Bayes nets and other probabilistic graphical models can also be considered locally interpretable, as the structure of the model can be visualized as a graph, but probability distributions are often conditioned upon input data [367].

Black Box Models

While the above models have some intrinsic interpretability, most references to 'Explain-

able AI' generally refer to methods of interpreting black-box models that have opaque inner workings. Popular methods in these categories include nonlinear support vector machines (SVMs), random forests, deep neural networks (DNNs), and matrix factorization models. Many papers have looked at frameworks for interpreting general black-box models by analyzing their input-output relationships. Earlier methods have used mimic models, which train a globally interpretable model to predict the output of a black-box model [33]. Other ways of generating global explanations of models involve feature importance measures, which give an idea of how much information a given feature provides a model. These measures include general dropout or permutation importance, which describes how much a model's performance is affected by removing a features [10], Shapley values [66], or mean information gain in random forests [162]. Koh et al. [153] proposed a general framework for describing the influence that individual data points have on a black-box model prediction.

Due to the difficulty in creating global explanations with black-box models, many techniques have focused on creating local explanations for specific predictions of a black-box model. Ribeiro et al. proposed a framework for creating locally interpretable model explanations (LIME) by creating a white box model from data sampled locally around a given point, allowing for local fidelity [268]. Rule-based systems and Bayes nets have also been used to explain predictions given by matrix factorization recommender systems [250, 279] or random forests [114]. Other works have proposed methods of creating locally more locally meaningful feature importances for explaining, including 'anchors' [269], which finds the minimum subset of features needed to be constant for a prediction to be stable. Other work has tried to create generalized ways of calculating feature importance, including methods based on shapley values from game theory [188]. TreeExplainer introduced a way to calculate Shapley values for predictions from random forest [187]. Little work here has explored unsupervised models or clustering. For deep learning, methods have been developed for introspection into specific model types. Model architectures can be visualized through model graphs [356] which have been deployed in common deep learning libraries. Many

methods have been proposed for using attention to visualize which sections of an image are most responsible for a particular class prediction in computer convolutional neural networks [46, 283]. Feature classification techniques have also been proposed to visualize the activation of intermediate layers in CNNs [241]. Semantic dictionaries have also been used in image classification, where representative images are generated to visualize network neurons. Semantic dictionaries have been used to show which representative images have most similar activations in the network to a given image [242, 307]. Similar concepts have been used in neural language models, where the most important previous words are highlighted [57, 75].

1.3.3 Visual Computing

Our work exists within this framework as a series of "design studies", which leverages design approaches from Human-computer interaction for the visual design aspects and attempts to find present more concrete findings from domain-specific applications which are placed into existing frameworks in order to support domain transfer. We focus on Activity Centered Design, which emphasizes a design approach that breaks down user requirements into activities and tasks that compose those activities, which domain experts engage in, which are then presented in our work. Similar paradigms include the critical decision method [64] for extracting needs from domain experts, and Action Design Research [210] which suggests a general framework for conducting collaborative design studies through a four. These approaches usually present stakeholders as the basis for the studies. In contrast, data-first design studies [243], which describes a framework where real-world datasets serve as the foundation of a design study, and appropriate stakeholders are identified only after data has been acquired. While we are primarily concerned with stakeholder risks, elements of data-first design arise in instances when goals are poorly defined or restricted by data availability, as often occurs in modeling studies. In terms of design aims, the most similar work in this area is work on visualization needs for oncologists in other specialties [263]. However, we focus on unique aspects of model steering that merges elements of visual design and medical XAI in a way that is under-explored.

The standard framework for the design of data-visualization is the nested model [227], which breaks the design process into 1) "domain problem characterization", 2) "data abstraction design", 3) "encoding and interaction design", and 4) "algorithm design". Many works have since proposed extensions of the nested model such as McKenna's et al.'s [212] proposed a framework of four *design activities*: understand, ideate, design, and deploy, with examples of each that they link to the different stages of the nested model. Wang et al. [329] proposed a variation of the nested model for the design of XAI systems that focus on the design of visual explanations of the models rather than statistical data visualizations. Such work often focuses on data visualization techniques aimed at allowing users to reason about the data to gain insights or aid in decision-making.

Tangential to the visual design process for visual computing is the study of how people interact with data visualization systems for data analysis. A common framework for data analysts is the sensemaking process [172, 254, 255], which typically broken down into a "foraging" loop where patterns are identified, and a "sensemaking" loop where explanations are generated for the explanations. In terms of integrating the sensemaking loop into data visualization, a common loop is a top-down approach is the idea of "overview first, zoom and filter, details on demand" [286]. For domain-specific workflows, a common approach is an alternative bottom-up approach where users are shown interesting details and then general context is given [185]. In reality, many systems such as the ones we design use a mixture of bottom-up and top-down workflows that are supported via the use of an overview+details design paradigm as described by Cockburn et al. [65], wherein we use separate linked views for analyzing the data in different granularities.

In terms of creating value from design studies, Lee et al. [173] proposes several possible "contributions" to the field of design. When considering design study rigor, Meyer and Dykes [216] suggest three "topics" for design studies: visualization idioms, design guidelines, and problem domains, and propose a framework with 6 aspects for evaluating rigor: informedness, reflexiveness, abundant, plausibility, resonance, and transparency. Of these, our

work focuses on *New domain and problem* descriptions, focused on domain specific applications of visualization for spatial XAI, as well all *design guidelines* and occasionally *lessons from failure*. While not the main focus of this work, secondary outcomes are *visualization idioms* in the form of *data abstractions* and novel *visual representations* that arise from the novel requirements of the problem domain. Additionally, this work will ultimately advocate for the use of greater *reflexiveness* in the design of XAI algorithms themselves by actively examining how the choice of designs influence how models are built and received through researchers.

VC + ML

One approach for supporting better VC is with the integration of machine learning and explainable machine learning systems (XML) with visual computing. These approaches are complementary, as ML systems are increasingly being applied to gain insights into real-world problems using large amounts of data that are intractable to human users, but may fall into pitfalls due to their limited ability to learn context. As a result, strides have been made in integrating XML and VC into systems such as medical diagnosis and treatment, finance, and crime prevention. Policymakers have been pursuing better methods of creating XML from both performance and ethical standpoints [98]. This has been spearheaded by initiatives such as DARPA's 2016 XAI initiative [115], and the inclusion of the "right to be informed" in the EU's General Data Protection Regulation, which insists that decisions made through AI have sufficient transparency to allow a third party to understand and override a decision [145]. This has pushed a growing interest in improving not just the models, but the methods with which we interact with these models in order to best create truly transparent models. Approaches for improving models through XAI have integrated visualization tools for both analyzing and steering models in ways that can't be done easily through traditional machine learning approaches, with significant success in areas such as large-scale computer vision models and black box model building for data scientists [352].

While such ML applications have value in improving the accuracy and speed of many human-centered tasks, there are significant challenges in the way of properly developing and deploying such models. In the case of human centered applications, models can easily learn erroneous relationships in the training data that can result in poor or unfair decision-making that may go uncorrected. Such issues can arise from challenges born from the high-dimensional nature of the data with many confounders, social contagion effects [59], bias present in the training data [69], or the use of training objectives that don't align perfectly with the stated goals of the model [370] such as the use of expression detection models for emotion or deceit detection [20]. Results of the misuse of these high-stakes algorithms can have strong negative effects, such as poor or biased decision-making, or poor adoption in settings when there isn't sufficient trust in models.

Many visualization approaches have been created for generating explanations of ML models. Several works have given overviews of XAI systems and techniques in various domains [8, 222, 228], including XAI specific to medical domains [317].

Domain-specific interfaces for exploring white-box models are common in areas that involve decisions that affect humans. RegressionExplorer [76] described the design of a system for inspecting regression models built by biostatisticians. Other systems have looked at interactive rule mining for health record data [289]. Clusterphile [44] presented a system for tuning clustering algorithms for arbitrary data. Analytic interfaces have also been developed for working with specific deep model architectures, including generative adversarial networks [326], and deep reinforcement learning [325].

Several systems have also taken a model-agnostic approach that relies on post-hoc analytic techniques. These systems offer user-guided exploration of the input-output relationships of variables to understand black-box models without performing model introspection [157,371]. Liu et al. [180] created a system that used topological data analysis techniques to visualize the prediction error of black-box models within the input space. FairVis [37] designed a model agnostic interface around the specific goal of identifying sources of intersectional bias

in the prediction algorithm.

VC + ML for Vis

Related work in spatial visual computing has been applied to other medical applications, such as medical imaging and clinical decision support systems. For visual computing on individuals, Kohn et al. [167] presented an early version of linked views for neurosurgical planning along with a method for clustering brain fibers based on the number of shared voxels between each track. Similar work has applied 3d visualization with linked views for clinical decision planning support [278] for individual patients. Pandey et al. [247] presented a method of using topology-preserving 2-d mapping of 3-dimensional brain data, similar to work we do with 2-dimensional mappings of anatomical information. My work differs from these in that I incorporate spatially aware machine learning methods. In contrast, Mistelbauer et al. [220] presented a collaborative workflow for visualizing aortic blood flow with algorithms for automatic prediction of aortic dissection risk using 3d modeling. In contrast, my work explores cohort-level modeling and XAI methods in the risk prediction methods. For group-level analysis, Bernard et al. [22] presented methods of cohort analysis of prostate cancer with statistical measures to support pattern finding in the data. Raidou et al. [261] presented a series of works in visual computing for radiotherapy planning, with steps including imaging, tissue characterization, segmentation, dose planning, and tumor control probability modeling. Other works by the game group have looked at applying clustering to 3-dimensional bladder data for use in clinical predictive models [100, 262] with a focus on shape variability and uncertainty. In contrast, our work focuses on incorporating the spatial methods directly into the models, and focuses more on explainability and trust in the visual computing models.

On the machine learning side, Wenskovitch et al [339] presented a series of works on non-spatial methods of visualization and unsupervised machine learning. Wang et al. [329] proposed a domain characterization of VC+Explainable ML design, applied to drug discov-

ery. However, these works don't focus on spatiality in the machine learning and visualization aspects beyond standard CNN approaches.

Chapter 2

(TSSIM) Visual Spatial Case-based Reasoning for Radiation Plan Prediction

One important application of spatial ML+VC is in computing over medical imaging for cancer patients. A particular challenge in this domain is the design of spatially aware topological similarity measures for unsupervised learning, as well as the design of visual encodings that can communicate the topological information to clients.

This chapter describes a visual computing approach to radiation therapy (RT) planning, based on spatial similarity within a patient cohort. In radiotherapy for head and neck cancer treatment, dosage to organs at risk surrounding a tumor is a large cause of treatment toxicity. Along with the availability of patient repositories, this situation has led to clinician interest in understanding and predicting RT outcomes based on previously treated similar patients. To enable this type of analysis, we introduce a novel topology-based spatial similarity measure, T-SSIM, and a predictive algorithm based on this similarity measure. We couple the algorithm with a visual steering interface that intertwines visual encodings for the spatial data and statistical results, including a novel parallel-marker encoding that is spatially aware. We report quantitative results on a cohort of 165 patients, as well as a qualitative evaluation with domain experts in radiation oncology, data management, biostatistics, and medical imaging, who are collaborating remotely.

The content of this chapter was originally presented at the 2019 IEEE Vis Conference and published in Transactions on Visualization and Computer Graphics [345].

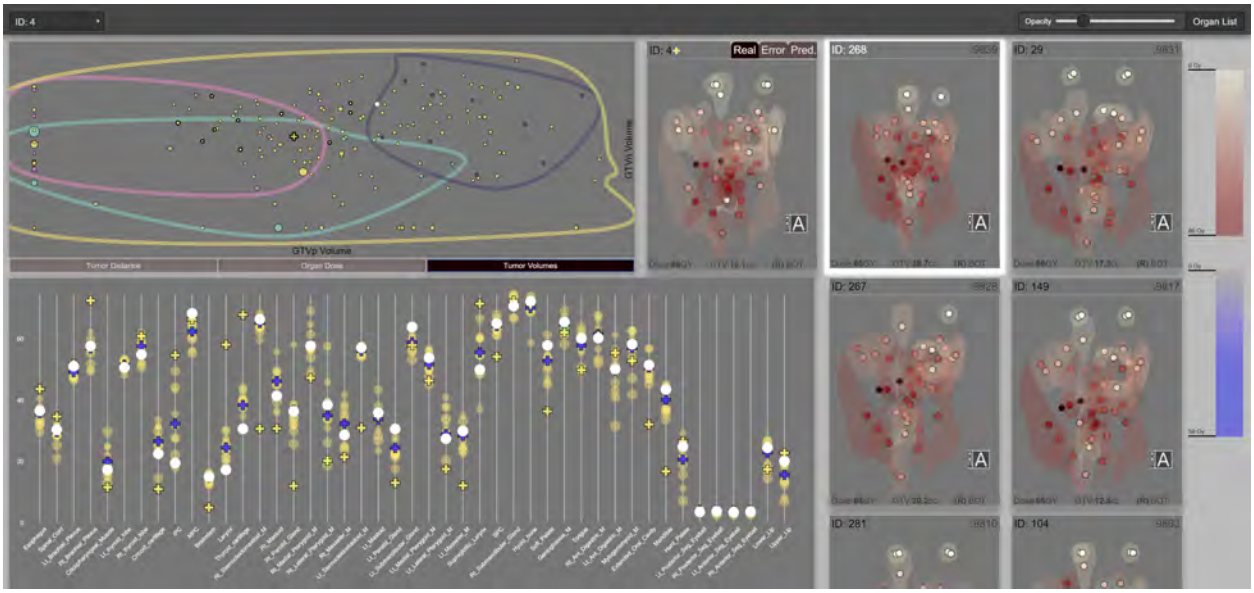


Figure 2.1: Visual computing for cohort-based radiation therapy (RT) prediction. A stylized 3D view of the predicted radiation plan of the current patient is placed centrally; top pale markers (front and back of eyes) receive the least radiation; tumors (black markers) receive the most. Additional RT views show the most similar patients under our novel T-SSIM measure, who contribute to the prediction; the most similar patient is currently highlighted (white) for comparison. A scatterplot (left) shows 4 clusters generated through the T-SSIM measure, with the current (cross) and comparison patient highlighted. A parallel-marker encoding (bottom) shows the predicted (blue cross) per-organ dose distribution within the context of the most similar patients; spatially collocated organs are in contiguous sections of the x-axis.

2.1 Introduction

Modern radiation therapy (RT) has seen large advancements in the application of computational approaches for imaging and rendering structural data of a patient. However, once this information is extracted, the field requires a high level of human expertise and a tremendous amount of effort to create and develop personalized, high-quality treatment plans. For example, head and neck radiotherapy plans take as long as a week, which, given that aggressive tumors double in 30 days, deteriorates the chances of tumor control and patient survival [237]. Furthermore, radiotherapy plans also affect organs located nearby a tumor, resulting in significant toxicity (side effects) and loss of quality-of-life. There is no current method to predict toxicity before the development of the plan.

With the emergence of large patient RT data repositories, there is growing interest in leveraging these repositories to computationally predict the dose distribution and toxicity for a patient before the actual RT plan is created. Under a healthcare model termed "precision

medicine," such predictions would be based on outcomes registered for past patients with similar characteristics. These characteristics include the location of the tumor relative to the nearby organs at risk, which heavily influences the development of radiotherapy plans.

However, due to a lack of computational methodology to handle spatial similarity, radiation oncology clinicians rely solely on structural visual information from medical images, prior knowledge, and memory to guide the development of radiation plans and to forecast toxicity. This approach is not scalable.

In this work, we present a visual computing approach to RT planning, based on spatial similarity within a patient cohort. This approach introduces a novel spatial measure, T-SSIM, based on tumor-to-organs distance and organ volume, and its application in a novel predictive algorithm for dose distribution. The resulting algorithms are integrated with visual steering to support the algorithm development in a remote collaborative setting, as well as deriving insight into the role of spatial information. Specifically, the contributions of this paper are: 1) a novel hybrid topological-structural similarity measure for spatial data, inspired by an image fidelity technique; 2) the development of a predictive algorithm for RT-dose distribution, based on this spatial similarity; 3) the design and implementation of an interface to guide the development of these algorithms, including a novel parallel-marker visual encoding which is spatially-aware; 4) the application of these algorithms and design to the emerging field of precision oncology RT planning, along with a description of this novel domain; 5) a quantitative and qualitative evaluation with collaborating domain experts.

2.2 Related Work

Related work consists of other projects that study spatial similarity measures, visual integration of spatial biomedical data with nonspatial data, and visual steering to assist in model development.

Spatial Similarity Approaches in bioinformatics, pathology and oncology [161,252,340,364] facilitate spatial similarity by encoding spatial relationships through graph-based techniques.

Unlike in our case, the underlying graphs are often small or constructed manually by clinicians [324, 374]. A second class of methods, based on 3D shape-based similarity, have been successful in shape retrieval applications in computer vision [47, 131]. These methods typically experiment with artificial models such as CAD models or 3D scanner output, and focus on classifying models of very different shapes. These methods fall short of distinguishing anatomical objects within the same class unless the objects have easily identifiable structures, such as the mandible and outer body contour [277, 300]. In our case, structures are in the same class and do not have easily identifiable features. A third class of methods seeks to apply deep-learning to narrow versions of the similarity problem. For example, Nguyen et al. [237] use deep-learning to predict dose distribution over a small set of organs in a cohort that had received the same type of RT plan, using tumor dosage and masks for organ 3D volumes. However, to date, no method has looked at automatically quantifying spatial similarity between patients for a large number of organs or a variety of treatments, or at presenting the prediction methodology in a way that can be understood by clinicians, as we do.

Visualizing Biomedical Data and Nonspatial Data Through established surface extraction and rendering algorithms, scientific visualization of biomedical data has been able to gradually shift its research focus towards visual computing [182], integration of nonspatial data [142], and new technologies. For example, instead of rendering magnetic resonance data from scratch, Nunes et al. [239] focused on analysis, by linking existing medical imaging software (MITK [355]) with statistical views of metabolic data to support delineation of target volumes in RT planning. In recent RT plan visualization research, Patel et al. [249] use virtual reality (VR) to visualize RT plans, allowing 3D structure visualization with hue and opacity, as typically done in desktop applications. Ward et al. [335] describe a VR system for radiation planning that allows the user to alter beam positions. Although these and other works have led to advances in viewing and planning specific radiotherapy plans in detail, none of these works seek to compare RT plans between patients or make predictions. Two

other works [259, 262] have proposed visual tools for the exploration of uncertain tumor control probabilities in the prostate, and dose delivery accuracy as a function of bladder shape analysis, respectively. These works do not consider spatial similarity, surrounding organs at risk, or the RT plan as a whole.

In terms of spatial-nonspatial data integration, two prevailing paradigms for integrating spatial and non-spatial features exist: overlays and multiple coordinated views (sometimes called linked views) [197]. In biomedical scientific visualization, an overlay approach [28, 298, 299] is commonly used when the non-spatial feature is scalar. As the non-spatial data becomes more complex (connectivity, clusters, dynamic characteristics, other statistics), the linked-view paradigm [7, 25, 140] becomes prevalent. Several reports [197, 202, 205] further support the use of coordinated views in collaborative tasks which involve multiple users with complementary expertise. Other, more recent approaches [190, 205, 238] use a hybrid approach that consists of both overlays and linked views. We follow a similar hybrid approach to support the exploration of RT plan data.

Visual Steering for Model Development Visual steering (or integrated problem-solving environments) is a top problem in scientific visualization [142]. Under this research umbrella, visualization tools for predictive model development have been developed for domain-specific applications. Naqa et al. [86] built a visualization tool to help create statistical models for dose-toxicity outcomes for specific organs, using a combination of statistical views and model controls. Unlike our work, their project assumed that the dose-distribution was already known, and was restricted to individual organs. Poco et al. [256] built a system for visualizing and developing similarity measures in environmental data, but focused on abstracted views for improving the measures without referring to underlying spatial patterns, as we do. Kwon et al. [165] provided a generic method for clustering model development, and used it for the development of patient similarity when diagnosing heart failure, but with no spatial data included. Visual steering tools based on multiple coordinated views appear also in visual encoding design [208], engineering [270, 337], epidemiology [205], cell signaling [287], and

artificial intelligence [211]; some of these works emphasize visually adjusting a simulation as it progresses, while others couple the steering with off-line processes. These methods differ from our goals in the key consideration of the problem space. We are interested in developing predictive models using RT medical data, which has unique requirements related to spatial and statistical data.

2.3 Methods

2.3.1 Domain Background and Problem

In head and neck cancer treatment, RT is often used as a primary or secondary treatment for patients. Radiation oncology relies heavily on the use of imaging in order to obtain information about the patient's tumor and surrounding organs. Traditionally, data acquisition is accomplished via magnetic resonance imaging (MRI), computational tomography (CT), or ultrasound. These techniques provide 2d image slices across the target volume, that can then be segmented to identify organs of interest, and used in diagnostics and radiotherapy planning. Current planning techniques typically use these image overlaid with a color map, allowing clinicians to 'paint' the dose across the organ as a way to visualize the outcome of the different radiation plans [316].

In radiation therapy planning, a primary concern is limiting dose to organs at risk near the target volume, while maximizing tumor exposure. For example, a head and neck tumor may receive 66-72 Gy units of radiation, while nearby organs at risk ideally would receive lower amounts. Unfortunately, that is not always possible, and radiotherapy has been linked by several studies to organ damage and long-term toxicity (side effects), including xerostomia (permanent dry mouth), and swallowing complications [38, 178, 351]. In light of these considerations, high-precision methods have been developed that allow for complex, highly conformable radiotherapy plans to be developed and delivered. Intensity-modulated radiation therapy (IMRT) is one such method.

IMRT allows for delivering more precise dose distributions via multiple (5-9) different radiation beams, each with tunable intensity distribution [178]. The increased complexity

of these plans comes at the cost of longer planning time and heavy reliance on clinician knowledge [214]. IMRT plans are typically created through a mixture of expert knowledge and planning software, with repeated trial-and-error rounds and consultation between the planner expert and the physician regarding the plan quality and tradeoffs. Beams are typically set up at an expert-determined fixed height, in order to reduce the problem space for the optimization software. Constraints are given on the allowable doses to organs of interest, which typically take the form of maximum doses to organs at risk, and minimum constraints to the target dose [338]. As a result, the problem space and planning time are expensive, and there is keen interest in leveraging computational techniques in order to support predicting the outcome of the radiation plan earlier in the process.

At the same time, the high incidence of cancer cases has led to the creation of large repositories of patient data, along with their diagnosis scans, their respective RT plans, and treatment outcomes. Under the "precision medicine" healthcare model, practitioners seek to leverage these repositories in order to predict, for a specific patient, the most appropriate therapy course, along with the outcomes of that treatment. Unlike in personalized medicine, the precision medicine prediction is based on data collected from a cohort of similar patients in the repositories [202].

While cohort similarity based on abstract data (e.g. genetic sequence profile) is in general well researched in the statistics community, there is a general lack of spatial similarity methodology. In the domain characterization discussed in this work, our radiation oncology collaborators would like to be able to automatically retrieve, given the diagnostic scan of a new patient, a cohort of patients with similar tumor location. Currently, this is done based on clinician or institution memory alone, which is not scalable. Should such an automated similarity measure become available, the domain experts would then like to analyze the patterns in the RT plans of the patients within that cohort. Based on that information, they would like to predict the RT dose distribution for the new patient and its potential effects, without going through a detailed RT planning process from scratch. Because these tasks

and activities rely on the visual assessment of spatial similarity and prediction in terms of dose distribution over the head and neck organs, the problem stands to benefit from a visual computing solution.

We arrived at this domain characterization of precision RT planning through a two-year collaboration with a team of radiation oncologists and statisticians located at multiple geographical sites. During this collaboration, we (four visual computing researchers) held weekly remote meetings and quarterly in-person meetings with a group of four radiation oncologists, a data management specialist, and a statistician. To characterize this novel application domain and design a solution, we followed an Activity-Centered-Design paradigm (ACD) as described by Marai [200], coupled with team science principles for remote collaboration, previously described [202].

Design Process

We implemented the theoretical ACD paradigm through an iterative, multi-stage process. After identifying and confirming with our collaborators the main activities to be performed, the research team met weekly with the domain experts to collect feedback and refine user requirements for the design and algorithms. We used a quantitative methodology to assess the capabilities of the resulting solution, and a qualitative evaluation methodology via note-taking to assess client activities.

Data processing

The cohort data for this project is part of a repository of head and neck cancer patients from the MD Anderson Cancer Center that have received IMRT. Contrast-enhanced computed tomography (CECT) volume imaging data from the initial patient diagnoses were retrieved through commercially available contouring software [295]. Contours were manually segmented to extract primary (GTVp) and secondary nodal gross tumor volumes (GTVn), as well as volumes of interest in the prediction related to organs at risk. Each CECT image

was 512x512 pixels, with a slice thickness of 1.25-5mm. Connected tumor volumes were treated as one volume. After segmentation, we used a custom Matlab script to extract a list of structural features for each volume of interest: volume, centroid position, and distance between each volume of interest, including tumors. Distance was measured as the minimum distance between the two volumes. Dosimetric data on the minimum, mean, and maximum dose for each volume of interest was extracted from radiation plans. Additional data on each patient's treatment plan was also included, which included the patient's tumor laterality, tumor subsite, and prescribed dose. All patient data was anonymized; patients were coded using dummy ids.

45 organs of interest were identified as being of interest by our oncology collaborators, in addition to the primary and secondary tumor volumes. Of the candidate patients, only those with data on all 45 organs, and at least one primary or secondary tumor volumes were included. Since segmentation and labeling of the data were done manually for higher accuracy, some anomalies in the dataset were found after visual analysis. Patients with organ position or mean doses more than 3 standard deviations about the population average were flagged and analyzed alongside our collaborators using the visual computing solution, and those with likely erroneous radiation plans were also excluded. The selection criteria was demographics-agnostic to prevent selection bias. 165 patients (140 male, age 59+-8.75 years, tumor N-staging [13] 0th through IVth: 32, 18, 91, 6, and 18, respectively) out of 245 candidate patients were included in the final cohort. The data can be further filtered based on patient characteristics. Data was then post-processed in order to compute derived features used in the visual interface, create dose predictions, and label patients with clustering results, as described below.

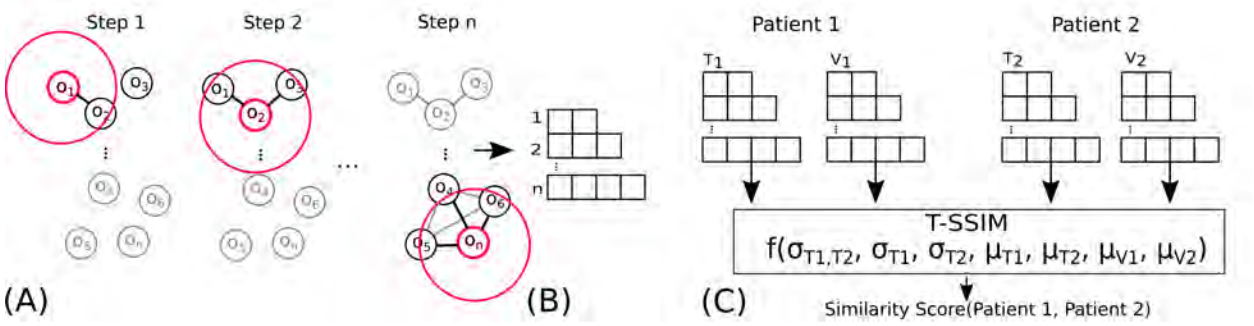


Figure 2.2: Construction of the spatial similarity measure. (A) A sliding window (a sphere, illustrated in 2D here) steps through the centroids of the organs, to identify nearby organs. (B) Each step in the sliding window constructs a variable-length vector based on the set of nearby organs (e.g., 2 organs in Step 1, 3 in Step 2, 4 in the n step). (C) We create two sets of vectors populated with tumor-organ distances and volumes, respectively, for each patient. These vectors are used as inputs into a similarity function (T-SSIM) to compare two patients. The vectors can be represented in matrix form (Section 2.3.2).

2.3.2 Algorithms

In order to support computing over images and 3D models (i.e., visual computing) for this project, we need to design appropriate algorithms for spatial similarity and prediction, described below.

T-SSIM Spatial Similarity Algorithm

In constructing a similarity algorithm special considerations need to be made for our problem. First, traditional methods of measuring similarity along feature vector representations, such as correlation or mean-squared-error, do not take into account the original structure inherent in the patient's anatomy. Second, neither shape-based techniques nor deep-learning techniques are a good match for this problem (Section 2.2). Third, the large number of organs-at-risk considered and the lack of clinician agreement makes infeasible the manual construction of a 3D graph structure based on the head and neck data. Fourth, an algorithmically constructed 3D graph-structure would have large edge cardinality, making graph-based matching algorithms infeasible. Because of these considerations, we arrived at a hybrid solution: 1) construct a topological structure based on organ adjacency; this structure will be common among all patients; 2) for each patient, generate two copies of the structure with tumor-to-organ distance data and volume data, respectively, specific to that patient;

3) define a similarity measure over these patient-specific data structures, inspired by image processing. Fig. 2.2 illustrates this process.

Our spatial similarity algorithm is inspired by the Structural Similarity Index (SSIM) [333], which is traditionally used to measure signal fidelity when comparing two images. Since the SSIM was designed for image processing, it takes advantage of an important assumption about the data: that pixel position serves as a direct analogue of spatial position. Because our data is already a reduced set of features (organs and tumors), rather than the original CECT images, this image-based assumption no longer holds. However, by reformulating the problem, we can use the spatial data we have to achieve the same effect, as described below. We refer to this novel reformulation as the Topological Structural Similarity Index, or T-SSIM.

In the original SSIM, a sliding window is used to calculate image similarity between the same local regions in two images. This local similarity is computed as:

$$SSIM(A,B) = \frac{(2\mu(A)\mu(B) + c_1)(2\sigma(A,B) + c_2)}{(\mu(A)^2 + \mu(B)^2 + c_1)(\sigma(A,A)^2 + \sigma(B,B)^2 + c_2)}$$

where $\mu(A)$ is the mean of matrix A, $\sigma(A,B)$ is the matrix covariance between two matrices A and B, and $\sigma(A,A)$ is the self-covariance of matrix A; c_1 and c_2 are small constants that are used for numerical stability. One of the reasons we use a local window is because image features and distortions are often space-variant. The window serves to isolate pixels within a certain distance from each other, so window size serves as a direct analog for actual distance. In contrast, our data is spatially bound to the centroids of each target volume. Thus, we need to find a way to encode the distance between the centroids, rather than a pixel distance. While the direct equivalent of a sliding window would be constructing a 3D area and sliding through different voxels, most of those voxels would be empty. Instead, we construct a topological equivalent.

In order to construct a topological equivalent to the SSIM image data, and create a sliding window analog, we need notation to describe when two volumes are within a window, for which we will use the concept of spatial adjacency. Let us define a matrix $\overline{D}^{|O|\times|O|}$, where

$d_{i,j} \in \overline{D}$ denotes the average distance between organs i and j across the cohort. We define two organs as being adjacent when the average distance between them is less than a certain distance d_{max}. Mathematically, we can write this as $o_j \sim o_i \, \forall \, o_j, o_i \in O \mid d_{i,j} < d_{max}$, where the $\sim$ operator denotes adjacency. If we consider our window to be a 3D sphere centered at a point, we can define all organs within the window as all the points adjacent to the center of the sphere (Fig. 2.2A). For efficiency, we will only consider the set of windows centered at each organ. Conversely, we can represent this set of windows as an adjacency matrix $M^{|O|\times|O|}$:

$$M_{i,j} = \begin{bmatrix} 1 & o_i \sim o_j \\ 0 & else \end{bmatrix}$$

In other words, the row M_i in our topological structure is a row representing all organs that are within a certain distance from organ i (Fig. 2.2-B). Via line search [294] so that the whole topological structure is connected, we found the optimal parameter d_{max} as 80mm for the window size. The topological structure is common across all patients.

The next element we need is pixel value analog. In our data, each organ is bound to several variables that could be used. Alternatively, we can compute similarity over multiple variables, and take a weighted average of them. The downside of such an approach would be that not all possible variables influence equally the final result, so using multiple values would require careful weighing of the values. To overcome this problem, we consider the underlying formulation of the SSIM.

The original SSIM formulation can alternatively be written as the composition of three functions for intensity (luminance), contrast, and structure. These components can be written as:

$$l(x,y) = \frac{2\mu(x)\mu(y) + c_1}{\mu(x)^2 + \mu(y)^2 + c_1}$$

$$c(x,y) = \frac{2\sigma(x)\sigma(y) + c_2}{\sigma(x)^2 + \sigma(y)^2}$$

$$s(x,y) = \frac{2\sigma(x,y) + c_2}{2\sigma(x)\sigma(y) + c_2}$$

using the same SSIM notations. This formulation allows us to combine multiple variables. While we found that the distances between the primary tumor and each organ provided good matches using the original SSIM formulation, we can augment that measure by considering the organ volume as another intensity channel.

For notation, let us consider the set of the organs adjacent to organ i, M_i, and patients A and B. Let us instantiate a copy of the topological structure with the matrix of tumor-organ distances $T^{|P|\times|O|}$ and another copy with the matrix of organ volumes $V^{|P|\times|O|}$ (Fig. 2.2-C), where $T_{i,j}$ represents the jth organ of the ith patient. We want to perform calculation over subsets of adjacent organs that we encoded in M. We can write each of these local subsets of values as $M_i \cdot T_j = T_j^{(i)}$ and $M_i \cdot T = V_j = V_j^{(i)}$. Put simply, $T_j^{(i)}$ is the set of tumor-organ distances for all the organs near organ i, for patient j. With this notation, we can now define local similarity as:

$$f_i(A, B) = l(T_A^i, T_B^i)\, l(V_A^i, V_B^i)\, c(T_A^i, T_B^i)\, s(T_A^i, T_B^i)$$

By summing up the local similarity scores along the entire set of organs, we obtain a similarity score for patient A and patient B. We can then generate a matrix of similarity scores $S^{|P|\times|P|}$, where each entry is:

$$S_{A,B} = \frac{\sum_{i=0}^{|O|} f_i(A, B)}{\mid O \mid}$$

Scores are normalized across the dataset to be between 0 and 1. In Fig. 2.1 right, note how this measure successfully retrieves patients with similar tumor location.

Prediction and Statistical Analysis

To predict a patient's dose distribution, we use a weighted k-nearest-neighbors algorithm, which is a common method of prediction in similarity-based health models [284]. The dose distribution prediction was calculated as the per-organ dose average of the k most similar patients:

$$Rad_{i,j}^{predicted} = \frac{\sum_{n \in N_i} S_{n,j} Rad_{n,j}}{\sum_{n \in N_i} S_{n,j}}$$

where $Rad^{|P| \times |O|}$ is a matrix of radiation doses across the cohort, $Rad_{i,j}$ denotes the radiation dose to the jth organ for the ith patient, and N_i is the set of the k most similar patients to patient i.

Even before applying clustering to this similarity matrix, we started noticing unusual groups of patients forming based on this similarity measure, and specific patterns of radiation distribution. An immediate goal became to perform clustering and statistical analysis using this spatial measure, and incorporate the resulting information: each patient was labeled with a cluster computed separately from the similarity measure, as discussed in Section 2.4. To allow for the dosimetric and tumor-organ distance data to be viewed across the whole dataset, principal component analysis (PCA) [99] was done on the matrix of radiation doses Rad and tumor-organ distance T. When making a prediction, only patients within the same cluster were considered. When analyzing the optimal number of matches for our prediction (Section 2.4), we found that the number varied with the size of the cluster, and making the parameter tunable for different clusters helped improve performance. After testing different parameters via line search [294], we found that a good number of matches to use was the square-root of the cluster size.

Because the input RT plans already consider maximum organ doses, and minimum target constraints [338], the predicted results fall within clinically acceptable ranges. All data processing, calculations for similarity, predicted dose, and PCA were computed offline, and information was exported as a JSON file for use in visual steering.

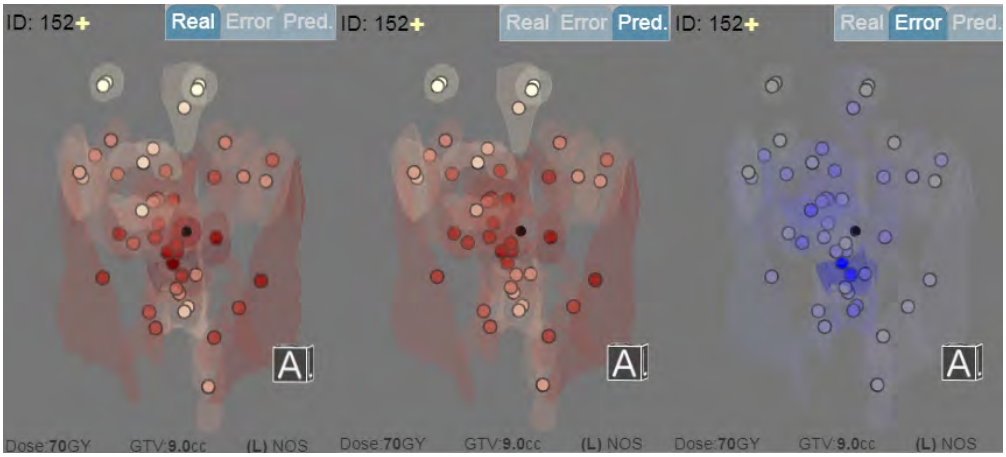


Figure 2.3: Three stylized views of the 3D radiation plan for Patient 152 showing the actual (left), the predicted (center), and the prediction error (right, in blue) in the radiation plan. Circular markers indicate the location of organs at risk, and black markers indicate the tumors. Red luminance is mapped to the radiation dose (higher dose mapped to darker shades) and blue luminance is mapped to error size. Transparent organ models are shown for context. The pale markers at the top correspond to the eyes, and the lowest marker is located down the spine.

2.3.3 Visual Steering Design

Once the visual computing algorithms are defined, a visual analysis interface enables the domain experts to steer the further development of these computation processes. By introducing an interactive visual steering component, we are able to leverage domain-specific knowledge, and support the discovery of patterns in the data. The visual analysis component of this application (Fig. 2.1) followed multiple design iterations, aligned with the similarity algorithm and prediction algorithm development. The final prototype design was designed to support the following activities (i.e., sets of tasks), derived from the domain characterization: (1) analyze the result of data clustering and similarity measures in the context of the entire cohort, and of spatial and dosimetric data, (2) analyze the inherently spatial dosimetric data extracted from the patients' scans and radiation therapy plans in a way that is visually intuitive to the domain experts, (3) compare those similar patients used in dose predictions, (4) analyze the result of our T-SSIM patient similarity measure, and (5) analyze the results of the dose prediction algorithm.

The final prototype comprises several linked views. We chose to use linked views because they allow visual scaffolding from familiar visual representations to less familiar encodings [197]. Unlike public health research, which is focused on cohorts, precision medicine

is about the treatment of a specific patient, so the entry point to the application is a search box for a specific patient within the cohort (the default is the first patient). Because radiation oncologists are familiar with RT plan renderings, a 3D stylized radiation plan of the selected patient is placed centrally on the screen (activities (2) and (3)). Additional RT views for the most similar patients put the patient in a local context, and allow users to assess how the prediction algorithm is being used concretely (activities (3), (4), and (5)). To support analysis within the cohort, and allow for clustering studies context (activity (1)), a scatterplot shows the clustering data among different dimensions that can be explored. Finally, we provide a novel encoding that allows for the local dose distribution of each organ of interest to be understood within the context of the k most similar patients (activities (3), (4), and (5)). By linking the views, we provide a way of allowing specific plans to be understood within context, and we support a variety of workflows for exploring the data. We describe each component in detail below.

Stylized Radiation Plan Renderings

Centrally in the visual interface is a stylized 3D rendering of the radiation plan for the selected patient (Fig. 2.3). Organs of interest are represented as circles drawn at the organs' centroids. In order to reduce issues with segmentation and allow the visualization to be rendered without requiring information on the entire 3D contours from the CECT scans, the organ shapes are represented using transparent, generic 3D VTK models, centered at the centroids of the target volumes. A slider changes the opacity level of the organ models in the radiation plan, as well as the color-scale to the right of the radiation plans. By combining centroid data and generalized models instead of full 3D contours, we effectively reduce the computational requirements of the system and minimize visual occlusion while still showing a recognizable 3D structure of the patient anatomy. We encode dose to each organ with the luminance of the respective centroid node and model; we encode larger doses with darker values. Gross Tumor Volumes (GTVs) are shown only as nodes located at the

tumor centroids, drawn in black, to make them identifiable, with a radiation-dose luminance border, as there are no corresponding 3D contours for these regions. Additionally, when both a primary tumor (GTVp) and secondary (GTVn) are present, a line segment is drawn between these nodes, to further emphasize their spatial relationship. These stylized 3D views, as well as a miniature cube with orientation labels (scene bottom-right corner), can be rotated in sync by direct manipulation to allow the user to more easily see specific areas while still being able to recall the current orientation quickly. Additional marks, labels, and details on demand display information about organ names, dosage, volume, and tumor location, to help correlate information across the views. This stylized 3D view was the result of several design iterations, ranging from highly stylized node-link renderings of the organs to fully-fledged volume renderings, and a variety of markers and labels to indicate current orientation and details.

Because one of the goals is to be able to analyze the result of the prediction algorithm, tabs above the radiation plan allow the user to change the view to the predicted plan, and to the prediction error in the plan. We encode the prediction error using a blue hue in order to distinguish which information is currently shown.

A separate, scrollable panel (Fig. 2.1-right) shows similar stylized 3D views for the nearest-neighbors of the selected patient, sorted by descending similarity. Allowing the user to control the matched radiation plans separately supports the placement of those plans near the selected patient plan for easier comparison. For these neighbor RT plans, the similarity score between the given patient and the currently selected patient is shown in the top-right corner. Two color scales, automatically populated to encode the upper bounds of the doses found in the dataset, serve as a visual reference for colormaps, as well as inform the user of the minimum and maximum mean dose, and prediction error in the data. A neutral gray background was used to allow for contrast with both colored visual encodings and black text [91], and to allow for white to be used for brushing and linking.

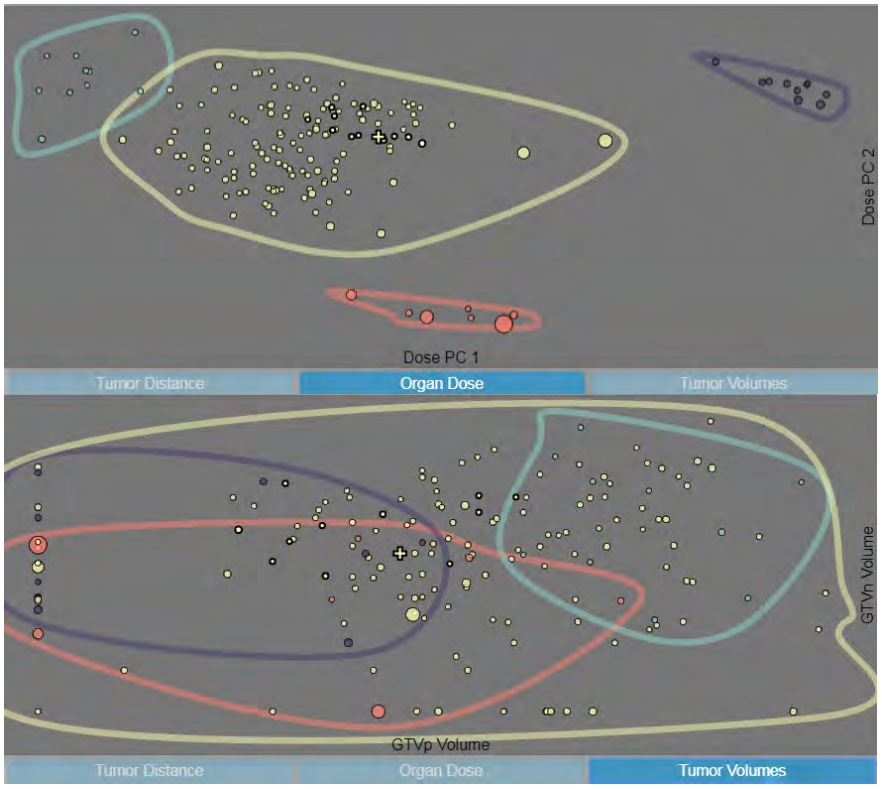


Figure 2.4: Two configurations of the scatterplot. The data can be plotted across the principal components of the radiation doses (top), primary and secondary tumor volumes (bottom), and principal components of the distances between each organ and the primary tumor volume (see Fig. 2.1 top left).

Scatterplot View

A main activity of interest to our collaborators was being able to analyze clustering results in the data. Additionally, We wanted a way to find correlations across the dataset to help identify where the largest prediction errors were occurring. Since the main data of interest was the relationship between spatial information and the radiation plan, followed by dose prediction, we selected the distances between the GTV and the 45 organs of interest and the dose information, respectively, as two of the feature spaces that could be viewed. For these feature spaces, PCA was done to project the 45 data dimensions to two. After several visual computing iterations and further discussion with collaborators (described in the Evaluation section), it was determined that tumor volume was also an important factor, and so it was included as an additional space. Since tumor volumes are usually categorized in 3-4 discrete stages, we used both the GTVp and GTVn volumes as proxy values to allow for better

discrimination among the cohort.

Patients in the scatterplot are color-coded according to cluster labels. The number of clusters shown was decided also through several design iterations, described in Section 2.4. In order to allow for easier perception of outliers, an envelope is drawn around each cluster. Animated transitions when changing the axis variables in the scatterplot allow for a visual understanding of how the different clusters are distributed across multiple dimensions (Fig. 2.4). Tooltips on the scatterplot allow the user to view the name, size, mean dose, and mean prediction error for the entire cluster.

Markers in the scatterplot are sized by the error in the radiation dose prediction for each patient to allow easy identification of patterns in prediction error, and to find outliers in the data. By default, patients are represented as semi-transparent circular markers, while a different shape is used for the patient in focus (a cross) so they can be more easily identified via pre-attentive cues. In an application of Tufte's layering and separation principle [311], patients used as matches for the selected patient are given a higher opacity and larger border so that they can be identified among the rest of the cohort. Additional tooltips allow the user to view the patient id, position, mean dose, prediction error, cluster, and current position in the scatterplot.

Parallel-Marker Plot for Organ Doses

While rendering the radiation plans in 3D provides an intuitive understanding of the relationship between the anatomical structure of the patient and the radiation plans, it proved insufficient for understanding the details of how the dose prediction was generated for each organ. Often, the dose distribution will vary significantly in a few organs across the cluster, while others, such as the brainstem and eyes, show little variance. In addition, a small number of matches means that a single outlier can strongly skew the distribution for certain organ predictions.

As a result, we wanted a way to explore and analyze the dose distribution across the

matches used for the prediction, while keeping track of spatially-collocated organs. Because predictions are based on a small number of patients at a time, traditional statistical plots such as box plots or violin plots are not appropriate for this task, as a single outlier would skew them in the data. Likewise, encodings that rely on size to encode distribution density require excessive screen real-estate to be visually discernible, which is infeasible when visualizing a large number (45 organs) of distributions.

Instead, we introduce a spatially-aware parallel marker encoding to fit our goals (Fig. 2.1 bottom). The encoding uses a parallel coordinate system, where the x-axis is divided into equal-length bins, each corresponding to one organ of interest in the radiation plan, not including GTVs. To encode spatial organization of anatomical marks, we started by grouping the 45 organs into 6 different categories (Throat, Oral Cavity and Jaw, Salivary Glands, Eyes, Brainstem & Spinal Cord, and Misc), which were determined after discussion with our radiation oncology collaborators at MD Anderson Cancer Center, and we laid out organs within each category contiguously along the x-axis. A vertical line is extended up the center of each bin to provide a visual reference. The order of the axes is fixed and based on the anatomical groups. The y-axis encodes dose, scaled based on the minimum and maximum dose found in the entire dataset. Moving the mouse into a bin highlights the vertical line for that bin, and brings up a tooltip giving the name of the organ, the predicted organ dose, and the actual organ dose for the currently active patient.

Within each bin, the dose to the specific organ is encoded by a marker for each patient considered for the current prediction. We chose to plot each patient point individually, given the relatively small number of points in each bin. By making makers semi-translucent, regions where several points overlap appear as more opaque, giving a visual indicator of density. The current patient is denoted by a different shape (cross), while matches are shown as dots and colored based on their clusters, maintaining consistency in color and shape with the encodings in the scatterplot. The predicted dose is also denoted by a cross marker, colored in blue. The size of dot markers is based on the computed similarity with

the given patient. This encoding serves as a visual metaphor, as larger dots carry more 'weight' in the prediction, and the predicted dose is effectively at the center-of-mass of the dots in each bin. We converged to this composite encoding after experimenting with and discarding parallel plot coordinate plots, as well as a variety of other axis encodings, markers, and channels.

The different views are linked through color, marker shapes, and brushing and linking. For example, when the user hovers over the encoding of another patient, all other encodings related to the same patient are highlighted in white (Fig. 2.1). Additionally, the user can select a patient to bring into focus by clicking on a point in the scatterplot or clicking on the patient ID label above their radiation plan. The data processing and algorithm for our system was implemented in python, using the NumPY library [318] for doing numerical computations, and Pandas [213] for data-processing. The front-end visualization was implemented as a web-based tool using HTML, CSS, and Javascript, with the three.js [36] and d3.js [27] libraries.

2.4 Evaluation and Results

Because of the visual computing nature of this project, we use a hybrid quantitative and qualitative evaluation methodology shaped along two case studies. We first present a case study of how visual analysis was used in conjunction with our similarity measure to help develop and improve the prediction algorithm. Along with this discussion, we present quantitative data about the prediction performance. In the second case study, we present a qualitative evaluation done with four senior domain experts in data mining, biostatistics, cancer medicine, and medical imaging.

2.4.1 Case Study: Algorithm Development

One of the topics of interest to our collaborators was understanding the importance of structural similarity in predicting radiation plans. However, traditional prediction methods are complicated by the fact that radiation plans can vary widely based on subjective planning

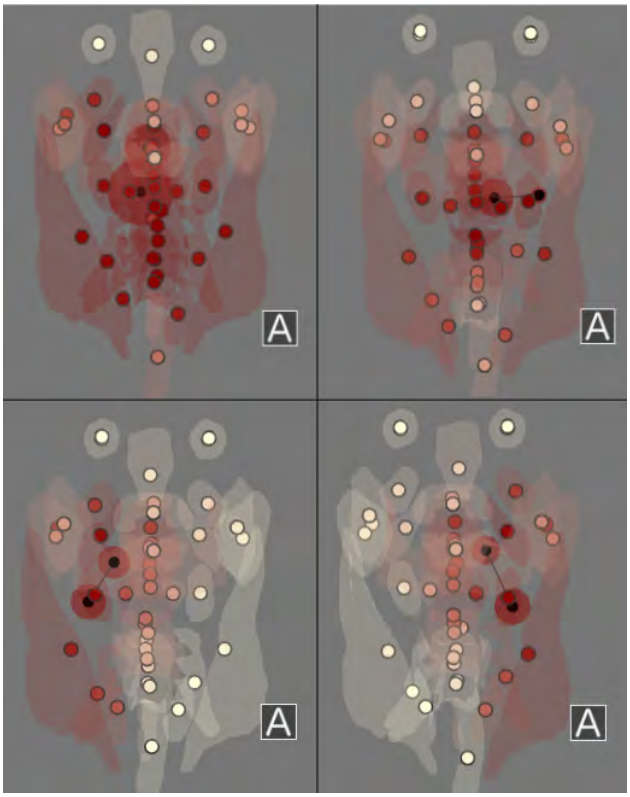


Figure 2.5: Example radiation plans for the 4 different patterns identified in the data. Top left: a plan with a higher dose to the lower-anterior throat. Top right: a plan with a 'standard' dose distribution, where radiation is lower in the throat and distributed to both the left and right sides of the head. Bottom right: a plan with dosing primarily to the right side of the head. Bottom left: a plan with dosing primarily to the left side of the head.

factors that can be patient-case, clinician, or institution specific. In this first analysis, we discuss the development and performance of our prediction algorithm in conjunction with this goal, demonstrate how insight from the visual computing tool was leveraged to help improve the prediction algorithm, and how visualization can be used to convey the results to clinicians to allow for better expert feedback in the algorithm design process.

We begin by first describing our measure for quantitatively assessing the success of the prediction algorithm. Given that for each patient in the cohort we have access to the actual RT plan for that patient, the accuracy of prediction across the cohort can be computed via leave-one-out validation, as follows: 1) for each patient in the cohort, use the tumor-to-organ distances and organ volumes to determine the most similar patients in the cohort via the T-SSIM similarity measure; 2) use the set of similar patients' RT plans to predict the dose distribution per organ (i.e. the RT plan) of the current patient; 3) compute and report the

prediction error as the difference between the predicted RT plan and the actual RT plan for that patient; 4) report the mean error across the cohort. In assessing error, we chose to compute the total absolute error for each patient. We decided on this measure over root mean squared error (RMSE), since the RMSE is typically done to more strongly punish outliers. Because we are interested in typical patterns for the patient, we are less interested in the effect that outliers have on the prediction.

Using the similarity measure and prediction algorithm without dividing the cohort into clusters, we initially found a mean prediction error of 16.68%, or 6.15 Grays (Gy), with a standard deviation of 9.31%. We compared this method to the naive method, where the predicted dose distribution is simply the average of the entire cohort. Using this naive method, we get a mean error of 20.62%, or 7.48 Gy with a standard deviation of 14.0%, which was suspiciously close to the performance of our initial prediction.

To better understand these results, the data and outliers were inspected using the visual steering tool. For each outlier prediction in the dataset, we inspected the k nearest neighbors selected for the prediction in the RT panel adjacent to the outlier patient. Where visual inspection did not pick up on subtle cues, the dose distribution plot was particularly useful in helping identify suspicious neighbor matches. Using the RT views, RT outliers were found to belong in three distinct pattern classes. Patients in these classes had larger errors, suggesting that they had peculiarities in their dose distribution that were not being captured by our similarity measure alone. RT plans for the patients in the 3 classes were analyzed and discussed with our radiation oncology collaborators and contrasted with patients with good predictions. In this manner, we identified four distinct patterns in how the RT plans were distributed (Fig. 2.5). This finding was subsequently confirmed in the scatterplot panel. The first, largest group was the 'standard' plan, recognized by our collaborators as most common for the cohort. Another group comprised a subset of the patients that received additional radiation to their lower throat, near the larynx. While surprised by this finding, our collaborators found this second RT plan type consistent with results reported by Amdur

et al. [11]. Amdur et al. discussed the choice of delivering additional irradiation to the larynx in patients and compared it to other methods of irradiation that largely avoid irradiating the larynx at all, leading to two potentially highly different dose distributions based on subjective choices made by the physician. The remaining two plan types were groups that appeared to have received highly unilateral radiation to only a specific side of their head, with the two groups corresponding to the two sides of the head. The radiation oncologists were enthusiastic and surprised by the power of the measure in making these findings possible. It was determined after discussion with our collaborators that the differences between the four plan types were likely due to radiation planning methods related to several other factors than tumor location, including the health of the patient, the tumor staging, and whether a biopsy had previously been done on the primary or secondary tumor.

Given this insight, we investigated introducing four clusters into the prediction, based on the different radiation plan archetypes found. This time, by only considering similar patients within the same cluster, our prediction error improved dropped to 12.3%, or 4.71 Gy, with a standard deviation of 4.43%. When normalized by prescribed dose, the total prediction error is 6.87% across the four clusters and for the 45 organs considered. Beyond the ability of the measure to identify the four RT classes, this prediction power was considered remarkable by our medical collaborators.

2.4.2 Case Study: Toxicity and Clustering Outlier

Because our project aims to support expert researchers in a specialized domain, we performed a remote qualitative evaluation with four senior domain experts, who are co-authors on this paper (GC, DV, BE, GM). The experts have backgrounds in data mining, biostatistics, radiation oncology, and medical imaging, respectively. All participants were familiar with the visual computing application throughout its development stages. Because of the experts' participation in the design process, the lack of an alternative existing system to solve the same problem, and in further accordance with the ACD paradigm, the evaluation was focused on the functionality of the application with respect to the target problem. Participants were

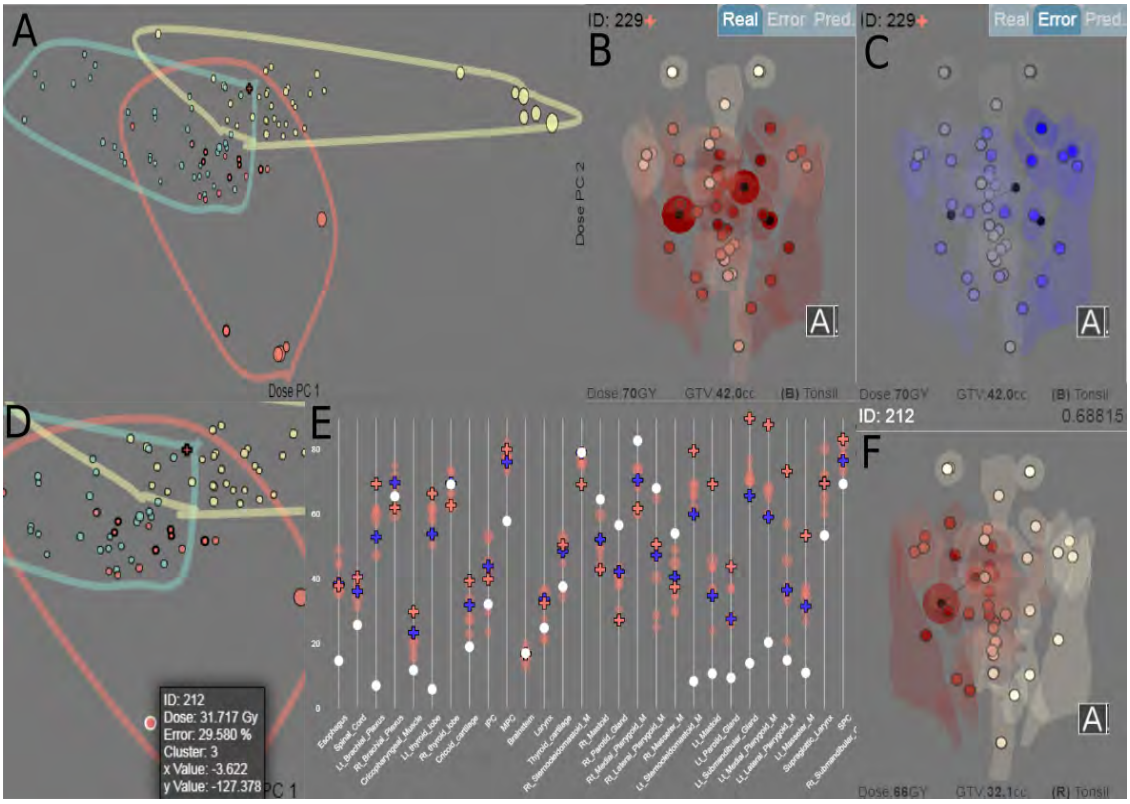


Figure 2.6: Snapshots of key moments during the qualitative evaluation. (A) Picture of the dose-PCA scatterplot on the reduced cohort using the clustering provided by GC. Clusters visibly divide the feature space despite being done without dose information. (B) RT plan for the patient being inspected (shown in (A) as the cross orange marker). (C) RT prediction error for the patient. Error rates are highest on the left side of the head. (D) Close up of the dose-distribution. One of the matches (highlighted) is significantly further from the other matches. (E) Parallel-marker dose plot of the patient and its matches. Doses from the suspicious match (highlighted) are significantly lower for several adjacent areas. (F) Radiation plan of the suspect patient, who received almost no radiation to the left side of their head.

given a briefing on the different components and basic functionality of the visual interface, and were encouraged to ask questions to guide the exploration of the data and results. The first author navigated the application with direct guidance from the participants, who were shown the same screen and were able to communicate with each other.

The main goal was to investigate whether our similarity measure can predict whether a patient will develop a particular toxicity (side-effects) after RT treatment, such as requiring the insertion of a feeding tube (FT). There are no current algorithms that can accomplish this type of prediction. The starting point of this investigation was a subset of 92 patients in the cohort for whom toxicity data was readily available. Collaborator GC had generated a clustering of this subset using our similarity measure, with the aim of correlating the tumor-

locations and RT plans with the toxicity data. The clustering had yielded three clusters, one of which was statistically correlated with the feeding tube toxicity.

The investigation (Fig. 2.6) started with the group examining the resulting clusters. Clustering had been done on the patient similarity scores provided by our similarity measure, and no expert (including GC) had seen the labeled results before in the context of the patient spatial information. The analysis started with the scatterplot visualizing the clusters, with targeted questions about the three PCA tabs. In the organ-dose plot, a collaborator noted that the clustering visibly divided the patients into separated groups. This was exciting to the group, given that the clustering had been done over the spatial similarity only, independent of dose. One of the visual computing researchers pointed out the cluster that was statistically correlated with the feeding tube outcome (turquoise cluster in Fig. 2.6.A).

Upon further inspection, the group noted that some of the matches within a different cluster (orange cluster in Fig. 2.6.A) were far apart in the organ-dose plot, while being close in tumor-organ distance plot. The group asked why that was, and proceeded to examine the RT views of that patient (Fig. 2.6.B), followed by the patient's predicted RT plan. Upon noticing spatially-localized higher prediction errors (Fig. 2.6.C), the group proceeded to examine the RT views of the nearest neighbors used to compute the prediction. By linking the view of each neighbor with the corresponding highlighted mark in the organ-dose scatterplot, the group was able to determine a suspicious match: while the tumor location in the neighbor was very similar to one in the patient under consideration, the two patients were apart in the dose-distribution plot (Fig. 2.6.D). A detailed investigation of the two patients and their match followed, this time using the parallel-marker plot (Fig. 2.6.E). One of the experts noted a localized difference in a contiguous subset of organs in the marker plot (last quartile of x-axis), and as the group circled back to the RT view of the match, they noticed that the neighbor RT featured a low dose to half of their head (Fig. 2.6.F). The expert in radiation oncology explained that the way the radiation plan was done could have been affected by a number of factors, such as if a biopsy had been performed on the patient's

lymph node. This led to a group discussion of the earlier case study and the usefulness of including a fourth cluster in the analysis, potential ways to incorporate more patients, and future plans to predict other toxicity outcomes based on the RT prediction.

An interesting result of this evaluation was the ability of the different domain experts to guide parts of the visualization and ask questions to each other. The collaborator with a background in data mining understood principal component analysis, and was able to explain the plot tabs to another expert. Instead of stopping the investigation with a convenient p-value finding, the group continued to examine the clustering that had generated that outcome, and were able to spot outliers and suggest improvements to the clustering. The medical imaging expert caught on the spatial dose pattern and explained it to the other specialists. When analyzing why two patients were being matched despite having notably different dose profiles with the clustering, the expert in radiation oncology provided the rest of the group with a clinical rationale for that fact. The statistician picked up on that interpretation, and suggested additional data collection. The group was able to efficiently use the whole system in order to make an important observation. Overall, we believe that this evaluation highlights a potential for visual computing methods such as these to support interdisciplinary collaboration more effectively.

2.5 Discussion and Conclusion

This work introduces a hybrid topological-image fidelity approach to creating an RT spatial similarity measure. Our results show that the resulting measure can successfully retrieve patients with similar tumor location. The similarity measure was then successfully used to make a valid prediction of RT dose distributions in a new head and neck cancer patient. The development of this measure and prediction algorithm was made possible through a visual steering approach, where a visual interface coupled with the spatial algorithms enabled us to identify and analyze situations where early algorithm versions failed. The same approach enabled us to identify four specific RT patterns in the data, and, in conjunction with the

spatial similarity measure, to improve prediction. When evaluated on a dataset of 165 patients, the prediction had low mean error: 4.71 Gy, compared to doses per organ as high as 70-90 Gy. We also observed low 4.43% standard deviation in the computed error, suggesting high certainty in our prediction. This type of certainty is particularly important when dealing with life-affecting patient outcomes. In conjunction with clustering, the spatial measure enabled detecting correlation between patient groups and a specific toxicity, paving the way towards precision medicine that leverages spatial information in patient data repositories.

Another result of this integrated approach is the ability to visually assess outliers and problems in the data. Since our data relies on segmentation of complex CECT images, problems in the data are to be expected. The high-dimensional nature of this data, combined with a relatively small dataset, makes outlier detection using traditional methods difficult. Additionally, automatic outlier detection methods are insufficient, since the presence of different clusters in the radiation plans means that new data could appear to be outliers, when in fact they are valid, but uncommon, RT plans, or that bad data can insidiously look 'normal'. However, by visualizing outliers, we were able to consult with experts in order to determine if the resulting anatomies and radiation plans are plausibly valid, or can be removed. For example, two patients in the cohort had several organs, including their eyes, positioned near the base of their throat. While these configurations are physiologically impossible, they were not detected in standard outlier detection, and even showed high similarity scores with each other.

Our qualitative evaluation also shows that an approach grounded in the ACD paradigm and visual scaffolding principles can lead to a satisfactory outcome for a difficult scientific problem. Using this approach, collaborators with a variety of complementary expertise were able to work together in order to gain insight into the relationship between spatial information and RT plans. A coordinated-views paradigm allowed us to leverage visual representations familiar to some of the experts, in order to expose those experts to novel or unfamiliar encodings. For example, oncologists were able to make connections between

RT volume renderings and the cluster and parallel-marker encodings. In the same vein, we note that our parallel-marker plot builds on familiar statistical plots while accommodating fewer samples and spatial contiguity. Because these visual encodings were developed through participatory design, we do not explicitly report feedback, which was enthusiastic, from our collaborators.

While our approach and spatial algorithms are generalizable to other problems in medicine and elsewhere, we note that there are limitations as well. First, details of an RT plan can change based on factors specific to the clinician and institution. For example, we have seen in our data that there are many cases where two patients are similar in terms of tumor location, but only one patient has highly-unilateral dosing. When generalizing a prediction method, we have to consider that other clusters could arise due to differences in the data, as well as technological and methodological differences between institutions. As a result, being able to inspect the data and leverage clinical knowledge is an essential function that can be accomplished through the use of visual computing.

Furthermore, while our current measure can encapsulate volumetric and spatial information, microscopic as well as higher-level information on the organ structure, such as shape and orientation, could be relevant and could also be included. Additionally, computing the similarity scores requires $| Cohort |^2$ computations, where $| Cohort |$ is the size of the cohort, and so it is done offline, while more sophisticated clustering methods are run off-site. This means that currently, online analysis can take place only once the results are generated. On 5 trials using a machine with 8GB DDR4 RAM and Intel i5-7200U 2.5GHz processor, the offline calculation took under 10 minutes (100.5s for processing and 476.5s for prediction, on average). This amount is negligible compared to the week-long IMRT planning process, which also requires medical professional input during the planning. In addition, our parallel-marker plot, which works well for 45 organs, has limited scalability to thousands of measurements. Finally, while our approach does not rely on learned parameters, we need to specify two meta-parameters: window size for computing organ adjacency, and an optimal

number of matches to use in the prediction, which may affect generalization. In our study, we found the optimal parameters via simple optimization [294].

In conclusion, We present a visual computing approach to support the development of a predictive algorithm to estimate radiotherapy plans in head and neck cancer patients. We present a novel, hybrid way of measuring anatomical similarity based on topology and measures of image fidelity. This similarity measure is then used in the emerging field of precision oncology, to retrieve patients in a cohort who are likely to have similar radiation plans and outcomes. By tightly coupling a visual analysis interface and a novel encoding with our algorithms, we derived valuable insight into the role that spatial information plays in radiation therapy planning, and were able to drive the development of the predictive algorithm. This visual steering approach is supported by employing coordinated views of spatial and nonspatial, statistical data. These views allowed domain experts in radiation oncology, statistics, data management and medical imaging to explore the data from different perspectives. Ultimately, the visual computing methodology presented in this paper enables calculations and insights into medical data that were otherwise not possible.

Acknowledgements

This work was supported by the National Institutes of Health [NCI-R01-CA214825, NCI-R01CA225190] and the National Science Foundation [CNS-1625941, CNS-1828265].

2.6 Chapter Conclusion

This work is an example of early applications of domain characterization and visual encoding design for spatial unsupervised machine learning with specialized anatomical data. Additionally, I propose a spatial similarity measure for the unsupervised model using clinical data which contains booth coordinates and spatial geometry, which were later integrated into the clustering used in chapter 3. This chapter focuses on proposing front-end solutions with an established dataset. In the next chapter, I will focus on introducing closed-loop model iteration directly into an interface that uses spatial data, while exploring how to perform

domain characterization and encoding when the project is still in the data collection and exploration stage, leading to rapidly changing design and data requirements in the context of COVID-19 data.

Chapter 3

(MOTIV) Transparent Data Mining and Inference for Social Media Data

This chapter discusses applications of Explainable Data Mining approaches for the Digital Humanities, with an emphasis on geospatial-temporal data. In particular, I will focus on dealing with the VC+ML sensemaking process for projects that are developed during the data-foraging stage of the sensemaking loop, and introduce examples of closed-loop model building integrated directly into the user interface.

In this chapter, we present a visual computing framework for analyzing moral rhetoric on social media around controversial topics. Using Moral Foundation Theory, we propose a methodology for deconstructing and visualizing the *when*, *where*, and *who* behind each of these moral dimensions as expressed in microblog data. We characterize the design of this framework, developed in collaboration with experts from language processing, communications, and causal inference. Our approach integrates microblog data with multiple sources of geospatial and temporal data, and leverages unsupervised machine learning (generalized additive models) to support collaborative hypothesis discovery and testing. We implement this approach in a system named MOTIV. We illustrate this approach on two problems, one related to Stay-at-home policies during the COVID-19 pandemic, and the other related to the Black Lives Matter movement. Through detailed case studies and discussions with collaborators, we identify several insights discovered regarding the different drivers of moral sentiment in social media. Our results indicate that this visual approach supports rapid, collaborative hypothesis testing, and can help give insights into the underlying moral values behind controversial political issues.

A shorter version of this paper was submitted to the 2023 IEEE Vis Workshop Vis4PanEmRes [347]. The full paper has been published in the Computer Graphics Forum [348]. Supplemental Material for this paper is listed at

`https://osf.io/ygkzn/?view_only=6310c0886938415391d977b8aae8b749`.

3.1 Introduction

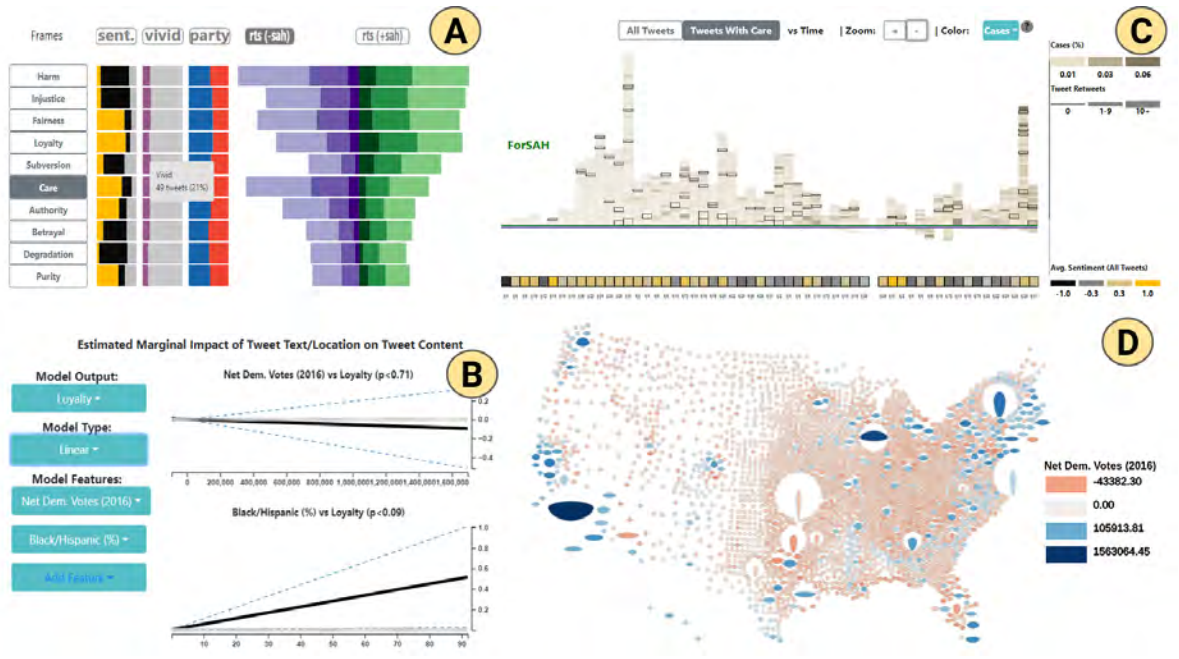


Figure 3.1: MOTIV visualization of moral framing on social media in 2020. (A) Summary panel showing tweet feature such as sentiment and tweets for or against the topic (B) Model building view showing inference scores for county votes within each county vs tweets expressing Loyalty, derived from a generalized linear model (C) Timeline of tweets, along with retweet count, COVID-19 cases, and sentiment. (lower bar). Counties from LA are shown in bold (D) Glyph-based map of counties showing 2016 voting history (color), voting age population (width), and tweets (height).

Social media has become a center of discussion of heated political discourse, ranging from the response to local government policy, to the rise of the #MeToo and #BlackLivesMatter protests in the US, and the shifting narratives that drove increasingly polarized reactions to the COVID-19 pandemic. The shift towards social media for discussing divisive political issues has made morality a vehicle for political messaging of all kinds, from social movements, misinformation and political propaganda through the use of modern moral panics [251]. In addition, the effect of the pandemic has inspired a renewed interest in understanding driving factors in the propagation of ideas on social media [2, 3]. Analyses of social media

discourse have attempted to either distill quantifiable text features that summarize popular topics [204,236] or identify the content spread by major influencers such as news outlets [40]. However, such basic text features often miss key information about users' motivations and personal values.

One approach that can help quantify users' motivations when considering social media dynamics is Moral Foundations Theory (MFT) [110]. Moral Foundation Theory is a psychological tool that proposes using a set of "Moral Foundations" as a basis to explain human reasoning. In this model different Moral Frames, such as Loyalty or Authority, can give insight into the nature of political discourse that is missing in traditional social media analysis approaches which consider only demographic and social factors. MFT has been applied to predicting social dynamics [30, 73], reaction to violent protests [105, 223], responses to hate speech [353], and reaction to appeals for charity [126, 331].

Social media analysis of ongoing topics gave several challenges. From a computational perspective, the short, informal nature of tweets, reliance on context and linked media in tweets, and the difficulty in understanding moral framing make traditional natural language processing (NLP) approaches such as semantic dictionaries and neural models relatively ineffective. For example, the sentence "Fauci said we should stay home!" could be a pro-SAH tweet that is expressing Authority by following an expert. On the other hand, this same sentence posted by a different individual may use this as an anti-SAH expression of Freedom, depending on the individuals' and their audience's feelings about Anthony Fauci. Therefore, meaningful analysis of social media data benefits from human-in-the-loop expert input and data visualization.

Additionally, the changing nature of the COVID-19 pandemic used a basis for this project produced many dynamic challenges to the design process. These ranged from the expected size of the dataset, to the features used, and difficulties with identifying retweets and multiple tweets between users, which quickly changed what was feasible with the data. Furthermore, Our collaborators come from a variety of backgrounds, and thus had different baseline ex-

pectations and workflows, and the short nature of the project meant that our collaboration had limited time to mature. While the design process was a challenge, it did yield several domain specific insights that were published by our collaborators in addition to our visualization work [93, 273]. Our design process had to meet significant challenges, from vague requirements to ongoing data foraging.

Visualization of MFT social media data poses several challenges. First, tweets may feature more than one MF, making succinct summarization difficult. In addition, because of the need to capture context in social media trends, the resulting data is large scale, and both temporal and geospatial. Last, the data is analyzed at multiple levels of detail, from high level trends in large corpuses to detailed content and local context. As a result, a solution needs to be able to handle a large number of different features while still maintaining an acceptable level of visual simplicity to make the system usable for clients with limited visual literacy.

In response to these challenges, we present a novel integrated visual framework for analyzing Moral Frames in social media. This framework is designed in collaboration with domain experts in NLP, machine learning, communications, and social science. Our collaborators are keenly interested in analyzing how differences in messaging affect public sentiment regarding controversial issues related to public health and welfare, in order to improve public messaging for social good. Our contributions are: 1) An analysis of the activities and workflows needed for the Moral Frame analysis of discourse; 2) The activity-centered design and implementation of MOTIV (Media Opinion Trend Inference and Visualization), a visual analysis system for exploring annotated, geotagged social media MF data; 3) An evaluation with domain experts in multiple fields; and 4) Lessons learned from the design process, with particular emphasis on working with an evolving dataset and during the data foraging and data understanding phases, as well as challenges when working with domain experts with limited visual literacy and different design goals.

3.2 Related Work and Background

Moral Foundation Theory is a model for analyzing social dynamics by identifying underlying "moral frames" implicit in the values expressed by individuals within a group. MFT was introduced by Graham et al., as a way of discussing the difference in moral values among groups. Graham's model used 5 (later expanded to 6) foundations, which are each split into positive (virtue) and negative (vice) orientations. For example, one frame is Care/Harm. "Care" is the virtue that is defined as "the need to help or protect oneself or others". "Harm" is the contrasting vice, which deals with "fear of damage or destruction to oneself or others" [109]. The 6 pairs of 12 Moral Frames are: Care — Harm, Loyalty — Betrayal, Authority — Subversion, Purity — Degradation, Fairness — Injustice, and Freedom — Oppression. Of these Moral Frames, Loyalty, Authority, and Purity are often referred to as "binding frames", which are the ones more strongly associated with conservatism. In contrast, Care and Fairness are "individualizing frames", which are associated with Liberalism. Freedom and Oppression are unique Moral Frames, proposed to better capture the viewpoints of Libertarians [132].

MFT has been used throughout social science in order to explore different values within groups, such as how individuals within different political parties may value different moral frames differently [155]. Moral arguments affect individuals' stance [266]. Moral foundations have also been tied to public health behaviors. For example, vaccine hesitancy is associated with Purity and Liberty, while pro-vaccine messaging focuses on Care and Harm [12]. Research has suggested that demographics influence moral framing, with women being more affected by Care/Harm, Injustice, and Purity [353], and Authority, Loyalty, and Purity are linked to conservative viewpoints in White Americans, but not African Americans [72].

Alternatives to MFT include the theory of Moral Motives [137], Dyadic Morality [281], and Relationship Regulation Theory [258]. We use MF theory over these alternatives as it provides the most diverse set of moral dimensions, and has been successfully used in popular textual frame analysis models in political communication [90]. Research has also shown that

moral judgment combined with emotion is a primary driver of viral spread in social networks and public health [122, 314]. MFT is thus a valuable tool for understanding how discourse develops around politically divisive issues on social media. In particular, we look at the MFT rhetoric surrounding politicized issues in the U.S, how moral valuation and stance relate to demographic factors, and the underlying Moral Foundations driving discussions on Twitter.

Social Media Analysis and Visualization Many studies have been done regarding social media responses to different topics [45,135]. Such studies include responses to the COVID-19 pandemic [6], feelings about public health policy such as vaccine mandates [78], and climate change [74].

Chen et al. [51] provide an overview of common visualization goals: visual monitor, feature extraction, event detection, anomaly detection, predictive analysis, and situational awareness. Guo et al. [117] provide an overview of event sequence data, including approaches to social media in terms of both collective and egocentric patterns, and lists challenges with social media vis. Our system uniquely merges aspects of feature extraction, event and anomaly detection and stance detections, with the integration of moral foundation theory and enriched demographic features for applications in politics and journalism. In relation to Guo et al., we deal with the challenge of multivariate event analysis - we need to identify both temporal and regional context when analyzing temporal changes in the tweet trends.

Several visualization systems have looked at how information spreads within communities on social media. Google+ Ripples [108] shows communities of Google+ users using Euler diagrams [272]. Visualization of social network information has also been used in journalism to visualize news coverage [219], and identify misinformation [70, 147].

Several systems have been built for real-time detections, such as visualizing information spread on social media using retweets and topic sentiment [40, 50], and real-time topic clustering [152], but do not incorporate geospatial information. In contrast, several systems have used integrated maps and text summarization for event detection [80] and disaster response [26]. However, these systems focus only on social media data, and do not analyze

moral frames or augment their data for more detailed analysis beyond simple sentiment.

For temporal analysis, other systems have focused on temporal progression of topics using sentence trees [127], timelines [358] and custom encodings [82]. Other works have explicitly focused on polarized topics [49] and stance detection [159]. Some work has integrated human-machine mixed analytics to detect "anomalous" threads [377] and bots [41], which both rely on a mixture of timeline visualization and glyphs. However, none of these systems tie their discourse to demographics or MFT framing.

Other methods have linked spatial and temporal information through methods such as Spatio-temporal clustering [322] and flow maps [150]. Other systems use linked views for monitoring events on Twitter [204], and journal articles [248]. However, no existing systems have incorporated other spatial information such as demographics or regional political ideologies.

3.3 Design

3.3.1 Design Process

MOTIV was developed via remote collaboration between four different research groups between May 2020 and February 2022 as the result of a RAPID [68] grant intended to fund projects to help inform and educate the public about COVID-19 safety measures. MOTIV was designed alongside the development of our dataset and required rapid updates to our design requirements and goals. Our design process is based on an Activity Centered Design domain characterization process [199], which we modified as a result of our irregular program circumstances.

The core group consisted of two researchers in communications, three NLP researchers, two researchers in causal inference in social media, and two visual computing researchers, all of whom are listed as co-authors. The team met remotely twice a month to discuss updates, identify project goals, discuss progress in data analysis, augment results from analyzing and annotating tweets, and produce visual representations based on data gathered from various sources.

At the beginning of the project, we used interviews and notes taken during meetings to develop the task analysis and project requirements, which were updated regularly at meetings as we explored potential data sources. Since our project was based on an emerging topic at the time (Stay-at-Home orders), we frequently worked on high-fidelity or functional prototypes of datasets as they were being developed and shared during group meetings. Results from these sessions guided future directions for the project and future datasets, and feedback informed updates to existing design requirements. Due to the nature of the data and collaboration, we focused on developing encodings that show as much data as possible and then refining them into simpler encodings that were more accessible to collaborators with moderate visual literacy (see supplement). Because the project aimed to support domain experts, not novice users, our design process focused on identifying existing workflows and activities performed by the domain experts, and building solutions that extend these activities.

3.3.2 Activity and Task Analysis

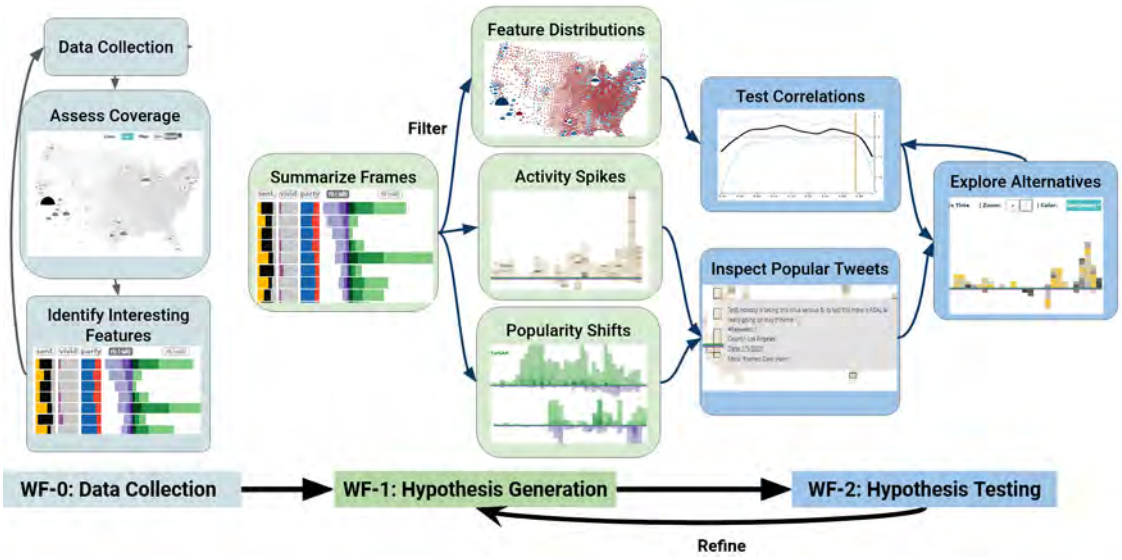


Figure 3.2: Workflows: (WF 0) Data foraging, where data is iteratively collected and analyzed to identify the quality of coverage and interesting features. (WF 1) Hypothesis generation, where moral frames are analyzed to identify interesting findings. A summary view is used to identify interesting frames, which are filtered and assessed in more detail. (WF 2) Hypothesis testing, where observations in (WF 1) are confirmed by drill-down or correlation testing. Insights are used to guide future investigations in (WF 1).

MOTIV was designed alongside collaborators in Communications and NLP, with an em-

phasis on supporting the Communications users in the final version of the interface, while earlier prototypes were intended to support NLP research in the development of a moral-frame annotated dataset.

In the beginning of this project, our aims were to support the development of a usable dataset along with collaborators in NLP and causal inference, which largely modeled the *foraging* loop in the sensemaking process [254]. Specifically, our collaborators, both computing experts and communication scientists, were interested in ways of generating a relevant tweet corpus, and assessing its geopolitical and temporal coverage.

Given this characterization, we found that our collaborators were interested in multiple, interrelated workflows (Fig. 3.2). *Foraging* (WF-0), is where all researchers assess the quality of the tweets, the coverage of the dataset in terms of moral foundations, time, location, and stance, and the distribution of potentially relevant features such as sentiment or vividness. *Hypothesis generation* (WF-1), is where researchers searched for major trends within the social media data, such as a general increase in the tweets about Liberty, and then developed hypothesis around potential causes of these trends, such as these Liberty tweets being driven by people from rural areas. Finally, during the *Hypothesis testing* phase (WF-2), researchers looked for ways to verify the causes of these trends, such as by looking at the correlation between population and Liberty tweets or investigating the events that co-occur with a spike in pro-Liberty tweets. The findings from the second stage would then feed back into WF-1.

During the data analysis state, we found that our interdisciplinary team's main interests ranged from examining how different socioeconomic and demographic factors relate to stance and moral framing with respect to controversial issues, as well as what textual factors such as "vividness" and "sentiment" affect tweet popularity. We found that our collaborators tended to model macro-level social dynamics as a feedback system, in which overall trends tended to be guided by two phenomena of interest: 1) grassroots memetic propagation of ideas in response to larger social movements, and 2) disruption events when a notable story or individual causes a shift in the online zeitgeist. Their research activities are thus

focused on identifying and explaining these types of phenomena and how Moral framing factors into them. In this way, MF serves as a lens to describe larger trends within social movements, while also serving as a reflection of how disruptive movements are viewed by others. We focused on identifying tasks that could not be done by individual researchers via their standard workflows:

- A1. *Summarize relationships between Moral Frames and demographic and political factors:* When investigating Moral Foundations on Twitter, our collaborators started with investigating key features and trends surrounding each Moral Frame (WF-0). They were interested in how political affiliation, tweet content, and popularity differed between each Moral Frame, and how this affected tweet stance and virality. Collaborators focused on investigating high-level relationships in the data using summarization before determining which MF to explore in detail (WF-1).

- A2. *Understand temporal trends:* Beyond high-level relationships, our team was very interested in exploring temporal trends in the popularity of each MF, with a focus on points when a topic would drastically change in popularity (WF-1), which could then be tied back to inciting events (WF-2). Our team was interested in the effect that changes in COVID-19 cases and lockdown orders had on MF trends, so a major requirement is the ability to include details of case rates alongside tweet popularity.

- A3. *Identify characteristics and Moral Frames of viral tweets:* Polarized discussion can be strongly affected by a small number of particularly viral ideas. Our collaborators were interested in identifying the most viral tweets, and their underlying Moral Frames. Identifying important tweets can help identify events or tweets that drive changes in temporal trends (WF-1). Additionally, identifying commonalities within viral tweets provides insights into potentially interesting features (WF-0), and how Moral Framing is viewed by different groups (WF-2).

- A4. *Understand the geographic distribution of each Frame within social context:* Moral

framing is heavily tied to political ideology and culture in literature, and regional differences are thus a major factor in how Moral Frames propagate within different groups. As a result, we wished to identify the geographical distribution of our tweets with different moral frames (WF-0), as well as overlap with factors such as income, political leanings, and COVID-19 cases (WF-1), with a focus on how Moral Frames vary based on the socioeconomic factors in the local area in response to state and country-wide mandates (WF-2).

- A5. *Verify hypotheses about meaningful relationships in the data:* Once hypotheses and potential relationships in the data were identified (WF-1), our collaborators often resorted to performing statistical testing to verify these findings. This was valuable for identifying features that would be useful for future models for our NLP researchers (WF-0), and validating findings from our communications researchers (WF-2). Thus, MOTIV needed to be able to identify causal effects while accounting for confounders, in a way that was immediately available to both analysts and non-analysts during sessions.

Non-functional requirements included online availability for remote collaborators. Because the datasets required expert annotations, our design needed to be usable with up to around 2,000 tweets, with potential to scale to larger datasets if better automated methods become available.

3.3.3 Data and Architecture

MOTIV was originally designed around a dataset of stay-at-home (SAH) tweets during the beginning of the COVID-19 pandemic, which is not claimed in this paper's contributions. Our dataset was gradually constructed and updated during the foraging stage of the project, with multiple data sources being added in gradually. Our Twitter corpus of geotagged annotated tweets using the US-SAH-MF corpus as described by Fatemi et al. [93], which we briefly summarize here. First, a sample of 87M tweets were taken from March 1 to June 30, 2020, from a dataset of COVID-19-related tweets [48]. Through interactive Latent Dirichlet

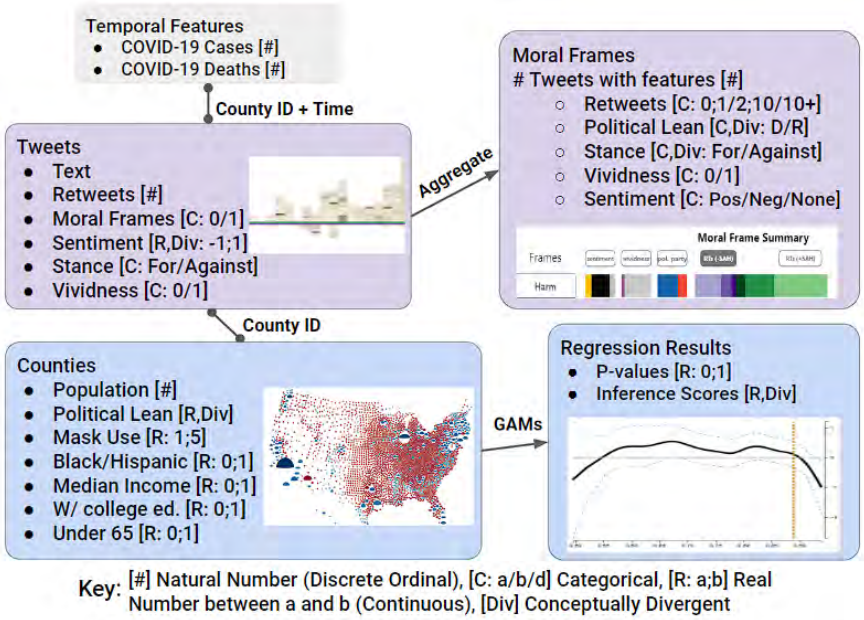


Figure 3.3: Data Abstraction. Tweets are labeled with textual features, and augmented with county-level data using the timestamp and geolocation data. Tweets are aggregated by MF and county for summarization. Generalized Regression Models model county demographics and aggregated tweet statistics to generate partial dependence plots in the inference views. County FIPS code is used to link all sections of the interface during brushing.

Allocation analysis, we extracted 20 topics, and identified 4 topics related to stay-at-home orders. The top 10 words in each topic, along with synonyms from Word-Net were used to sample 100 tweets with each keyword. We then identified manually which keywords contained at least 80% tweets relevant to SAH: home, open, quarantine, inside, and lockdown. These tweets were then hand-annotated by Moral Frame experts with the following information: 1) stance - if the tweet was in support of or against SAH, 2) whether the tweet contained specific or vivid descriptions (vividness), and 3) which of the 12 Moral Frames were expressed in the tweet, if any.

To reduce burden on our manual annotators, sentiment score was annotated using the sentiment analysis tool Vader [129], which uses a rule-based lexical system to determine whether the content in the tweet expressed positive emotions (e.g. happiness) or negative emotions (e.g. anger), without requiring training data. VADER has been shown to outperform other baselines, including human annotators, for sentiment analysis [267]. Scores of > 0.25 were defined as "positive", scores of < -0.25 were defined as "negative" sentiment, and middling scores were defined as "neutral".

We mapped the geolocation associated with each tweet to each of the 3113 US counties, excluding Antarctica, as follows. First, we obtained the bounding box of each geotagged tweet from the Twitter metadata. We then calculated the area of overlap between each bounding box and the borders for each county. Each tweet was assigned to the county with the highest percentage of overlap, and tweets that did not have at least 25% overlap with a single county were excluded. To avoid introducing bias by "guessing" stance or framing, we removed tweets that did not have a clear stance or moral framing. The result was 1483 geotagged tweets from the US that were determined to be relevant to SAH orders.

Overall, our system considers two data items: tweets, and counties, which are connected via geolocation. For tweets, we use the geotag to identify features taken from the corresponding count. On the county level, we incorporate 2018 census data [221,312], voting ratings for each county from 2018 [221], a self-rated mask usage survey from the New York Times [304], and COVID-19 cases and death rates for the time period covered by the dataset [141].

Political leaning is encoded as the number of votes for the Democratic Party minus the votes for the Republican Party in the 2016 presidential election, based on collaborator input. Because American voting patterns are largely polarized along the urban-rural continuum [280], we assume that, on average, regional trends can serve as a proxy for individual political beliefs. We also aggregate the total number of tweets with each moral frame and stance within each county, which results in 14 different continuous values for each county. Additional attributes selected during the foraging stage, and the data abstraction, are detailed in (Fig. 3.3).

Later on, MOTIV was further used to analyze a second dataset taken from the Moral Foundations Twitter corpus [125] to compare geotagged tweets associated with the #BlackLivesMatter (BLM) movement between 2014 and 2016. We encoded stance using hashtags: tweets that contain more hashtags for BLM to be in support, while tweets that contain more hashtags related to the All or Blue Lives Matter (ALM) to be opposed. Tweets that contained an equal number of hashtags for each side were excluded, as we could not be confident

in their stance. In total, we identified 1051 tweets in support of the BLM movement and 854 tweets in support of the ALM movement.

Data processing is implemented in python using the Flask and Pandas packages. The front-end is implemented as a web app using JavaScript with the d3.js and React libraries. Generalized additive models were implemented using the pyGam package.

3.3.4 Layout Design

To support the five main activities (A1-A5), and both foraging and hypothesis-related workflows, MOTIV uses multiple coordinated views which were developed gradually as the dataset was being developed. The four panels each support one main activity, whereas the view coordination supports insights into multiple dimensions of the data. The entrance to our interface is a Summarization Panel (Fig. 3.1-A) that shows sentiment, political party, stance and retweets aggregated by each frame. Once a frame of interest is identified, the analyst selects that frame, which loads detailed views in the other panels and filters tweets by moral frame (A1). To get an overview of temporal trends alongside COVID rates, we include a novel timeline view, which allows us to identify temporal trends and view tweet details in phase 2, as well as view temporal trends with secondary variables in WF 2 (Fig. 3.1-B) (A2, A3). To view geospatial trends, we use a novel glyph-based map that encodes demographics, MF popularity, and population for each county in the US (Fig. 3.1-D) (A5). Finally, an Inference panel allows for building predictive models and visualizes their partial dependence curves, which helps identify the relationship between individual features and demographics (Fig. 3.1-C) (A5).

We designed encodings to help capture foraging and hypothesis supporting patterns and outliers (A1-A4). To deal with the issue of "misleading" patterns, we then introduced an inference panel and tooltip details, to be used to validate findings with greater fidelity (A5). Additional linking and brushing highlights data items from the same region in linked views.

MOTIV was designed to support our collaborators' specific research needs, as opposed to novice users, and so our novel encodings benefit from participatory design and from

visual scaffolding [198]. Still, as our domain experts wished to be able to share the system with novice researchers in their groups, MOTIV also provides explicit legends and visual explanations on demand for custom encodings.

3.3.5 Summarization Panel

The summary view shows distributions of tweet-features that expressed a given Moral Frame (Fig. 3.1-A). We use rotated stacked bar-charts to encode tweet features such as stance and vividness. The stance stacked bar charts are further broken up by the number of retweets, to indicate the overall popularity of each Moral Frame. Bars are aligned horizontally with Moral Frames, to allow for a side-by-side comparison of part-to-whole relationships. The panel also supports sorting the order of frames based on each feature, to better show frames with the highest or lowest incidence of a specific feature value. The linked panels will update to filter by tweets with the selected Moral Frame.

Earlier design iterations included variants of parallel coordinate plots and correlation matrices. However, these were deemed to be unnecessary complex by our collaborators.

3.3.6 Timeline Panel

Visualizing temporal information is important for understanding how discourse evolves over time, how public sentiment evolves in response to major events, and how these events give context to unexpected patterns in the data. To support this type of analysis, we use a novel Timeline panel that encodes each tweet, as well as the tweet's date, popularity, stance, and geolocation data (such as COVID-19 rates) over time. Our timeline supports inspection at two levels of granularity: overall trends and major events, and the context of events by visualizing the details of popular tweets.

In the layout, the X-axis is mapped to time, and aggregated into "bins", or windows of time, depending on the total number of dates covered in the dataset. Individual tweets within each time window are encoded as tiles, which are stacked within each window of time (Fig. 3.4). Thus, each tile is positioned along the X-axis of the timeline according to

the tweet date. To capture differences in tweet stance, the center of the chart is bisected horizontally along the X-axis. Tweets that are in support of a topic (e.g., for SAH) are positioned above the center axis, whereas tweets opposed to that topic are placed below the line, to allow direct stance comparison.

To support tweet popularity analysis, we sort the tweets along the Y-axis based on their popularity, such that the most popular tweets are always close to the center axis of the timeline. Because our goal is to capture the *popularity* of each Moral Frame, and not simply the number of tweets, the height of each tweet is scaled according to the number of retweets, such that more popular tweets contribute more to the overall height of the timeline.

Finally, each tile is color-coded based on a user-selected demographic or tweet-specific features of interest. Details including text, location, and the Moral Frames expressed are provided via tooltip interaction. One can also filter the timeline to show only tweets expressing a certain Moral Frame. Selecting a tweet will highlight in the chart all tweets from the same county, as well as highlight the county the selected tweet is from, in the other panels (Fig. 3.7).

We arrived at this custom encoding after exploring several popular variants of sparklines and steam graphs using a larger set of tweets without geotags, where color encoded sentiment over time. While this approach helped identify high-level trends, it also prevented inspection of individual tweets, which is important for understanding the context behind spikes in tweets. Additionally, since some features like COVID-19 cases were dependent on both location and time, it was important to map case rates to individual tweets in the timeline without aggregation.

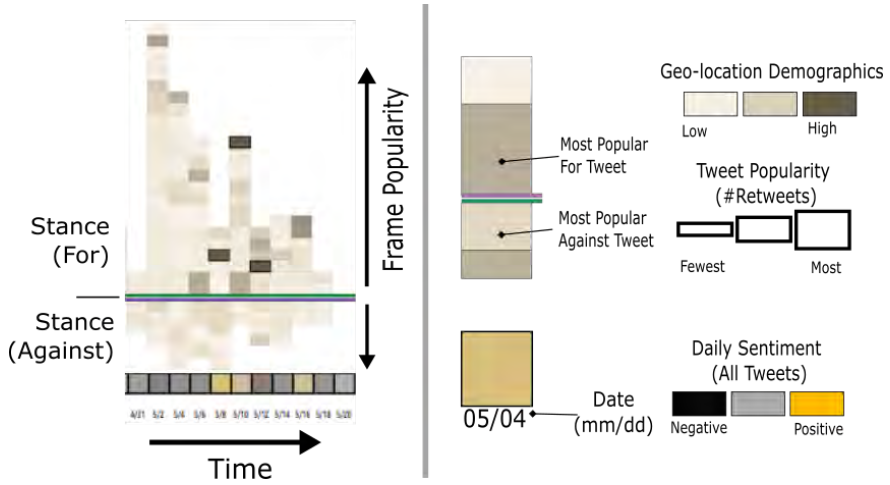


Figure 3.4: Outline of the timeline encoding. (Left) Timeline over a period of 10 time bins. Individual tiles encode tweets within the time bin. Tile height and position encode retweets and stance while color encodes a secondary variable. (Right) 4-tweet example encoding for a single time bin, annotated with the date in mm/dd format (05/04). A square tile in the bottom timeline shows the tweet date where color encodes sentiment score across all tweets for that date.

3.3.7 Geospatial Map Panel

A major task was understanding how cultural and socio-econonomic factors influenced the spread of moral frames within different regions. To accomplish this, we include a county-level map showing local demographics and distributions of tweets with each moral frame. We experimented with choropleth maps, glyph-based encodings, and hybrids between the two. Because of the number of multiple variables to encode, designing this panel was particularly challenging. We use a custom glyph-based map (Fig. 3.5). The custom glyph uses a solid color to encode demographics, while shape encodes tweet density and population. Each glyph is drawn as a distorted ellipse, whose width encodes county population, whereas the upper radius encodes tweets within the Moral Frame in support of the topic, and the lower radius encodes moral tweets opposed to the topic. The resulting glyph is similar to a star-chart with a variable radius. We chose to use an ellipse over diamond shapes through experimentation, as we found that differences in the exaggerated degree of convexity of the curves of outlier counties served as a better pre-attentive cue than glyphs that use straight edges. A force-directed layout is used to adjust the position to prevent overlap between counties. This layout and the white space generated by the glyph helps emphasize counties with uneven

tweet/population ratios. By comparing the size and shape of the glyph, one can easily identify and examine both major cities, and areas with a disproportionately high number of tweets with a given stance.

When encoding political votes, we use an equal-intensity color scheme for both parties. The chosen color scale is divergent, where the maximum values are relative to the largest minimum and maximum values, so the most polarized republican areas are the same intensity as the most polarized democratic areas. This is because U.S. major population centers have massive bias towards democratic votes, and thus all republican areas would become almost white. In contrast, the experts were interested in overall polarization. The differences in population are accounted for by differences in size of the glyphs.

Our initial designs centered around choropleths, which were familiar to our collaborators, for showing tweets and demographics simultaneously. We experimented with using a mixture of color blending and texture blending [119], and overlayed glyphs (circles or spikes) to represent multiple variables, as recommended by Ware et al. [336]. Additionally, we experimented with using different levels of aggregation, where counties within a single voting district were grouped together to approximate areas of equal population (see supplement). A prototype of this standard multivariate map is shown at `https://tehwentzel.github.io/covid_map/`. These prototypes were developed during the data foraging stage. However, we found that it was still difficult to discriminate details around cities with high populations and small county area, which were regions of interest. As a result, we found comparatively better feedback and more useful findings through the use of the glyph map. A basic chloropleth map is also shown in order to allow for users to start with simpler visualizations before using the glyph view.

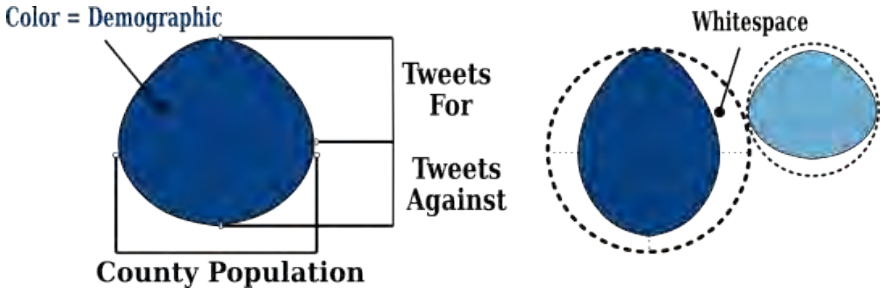


Figure 3.5: County map glyph: width encodes population, while the upper and lower radius encode tweets for and against the topic of interest that express a certain Moral Frame. Color encodes a user-defined variable, which is voting history in the example.

3.3.8 GAM Inference Panel

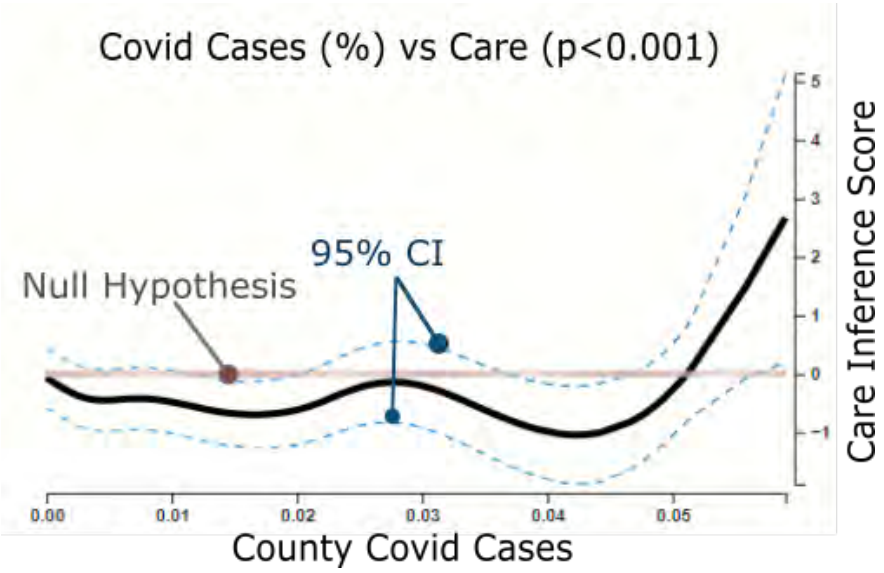


Figure 3.6: Partial-dependence plot showing the relationship between COVID-19 cases and the number of tweets expressing the Care frame. The Y-axis represents the direct effect of cases on tweets with care estimated by a generalized additive model. Blue lines represent the 95% confidence interval while the tan line shows the axis at Y=0, to provide a visual reference for the null hypothesis that the two variables are unrelated. The plot shows a majority of values below the grey (null) line, and a spike in Care tweets in counties with the highest COVID-19 rates, suggesting that a disproportionate amount of Care tweets come from a few counties with the highest case rates, but not from those with only a moderate number of cases.

One major design goal during hypothesis generation and testing to confirm visual analysis inferences via statistical tests (A5) when determining what factors influence moral stance and tweet popularity. This activity faces two major issues. First, many demographic factors, such as population and COVID-19 rates can serve as confounding variables. Second, the search space of potential confounders is too large to visualize all at once. To address these issues, we implemented an Inference panel that allows for interactive hypothesis testing.

Our Inference panel is centered around the use of generalized additive models (GAMS) [121]. GAMs are a class of predictive models. that treat the predicted variable as the sum of individual functions of input variables, written as:

$$y = f_0(x_0) + f_1(x_1) + ... + f_n(x_n) + intercept$$

where y is the variable being predicted, and $x_0, x_1, ...x_n$ is the set of n input variables. Each function $f_i(x_i)$ can be visualized individually as a "partial-dependence plot", allowing users to visualize the relationship between each variable, while accounting for relationships between correlated variables that are taken into account in the multivariate model.

Our implementation consists of a control panel for interactively building a predictive model, and the partial dependence plots of each input variable (Fig. 3.6). The control panel allows for the selection of the dependent variable being predicted, the input variables, and the type of shape function used in training the GAM. We included as potential predictors tweet-level features, such as the presence of a Moral Frame, or the number of retweets. Demographic factors, COVID-19 rates, and tweet content are included as potential input variables.

The model allows for either a spline or linear fit of the model. Spline curves allow for better representation of the distribution of the data, while linear models afford more accurate reporting of p-values to identify statistically significant relationships.

The choice to use GAMs and partial-dependence-plots was decided after many design iterations. Early in the project, we used clustering with user-defined demographic or textual features to automatically generate intersectional groups that could be displayed as a series of bar charts, star charts, or modified sankey-diagrams. However, collaborators felt that the implementation was too complex to interpret quickly when performing hypothesis testing. In contrast, we found that GAMs and partial dependence plots were more grounded in the existing knowledge of collaborators who used regression and line-graphs regularly in their research, while showing fewer features at one time to reduce cognitive load to users.

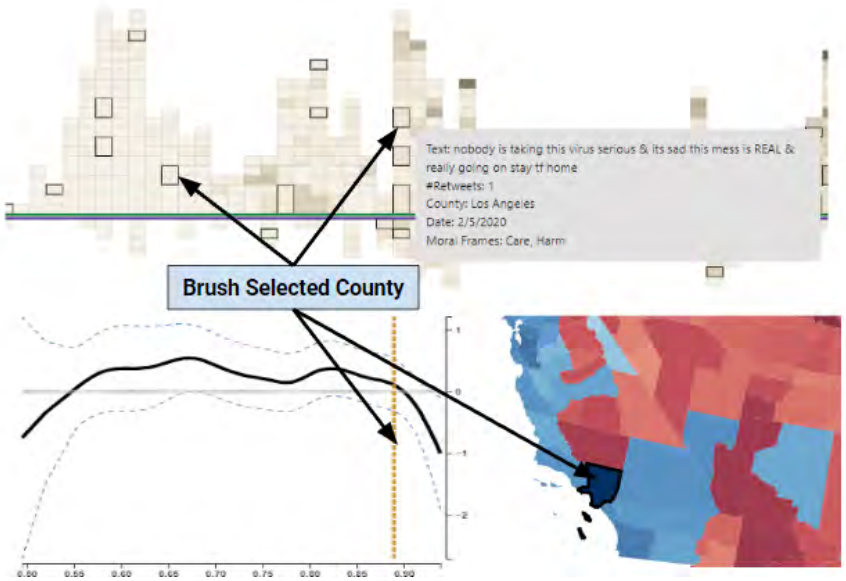


Figure 3.7: Examples of brushing and tooltip interactions in MOTIV: items associated with a selected county will be highlighted in the Map, Timeline, and Inference views; and the map and timeline views show additional details for counties or tweets via a tooltip when the user hovers the mouse over an item.

3.4 Evaluation

MOTIV has been adopted as a research tool by our collaborators in communications, NLP, and causal inference, with the intention of supporting insights into SAH policy application in the U.S. For this reason, we demonstrate its capabilities in two case studies, reported here in abbreviated form, which were performed over several months by our collaborators and later used in publications in our collaborators respective fields [93, 273]. In addition, we provide feedback from the target users.

Due to pandemic and work-from-home measures, the studies were completed remotely using the think-aloud technique with note-taking. We denote where these case studies correspond to activity workflows with the notation [WF]. Video summaries of the case studies are provided in the supplementary materials.

3.4.1 Stay-at-home Attitudes and Dominant Moral Frames

This case study focused on the creation of an annotated MF corpus and the subsequent analysis of Moral Frames as expressed in microblog data related to Stay at Home (SAH) orders in the U.S. (Fig. 3.8).

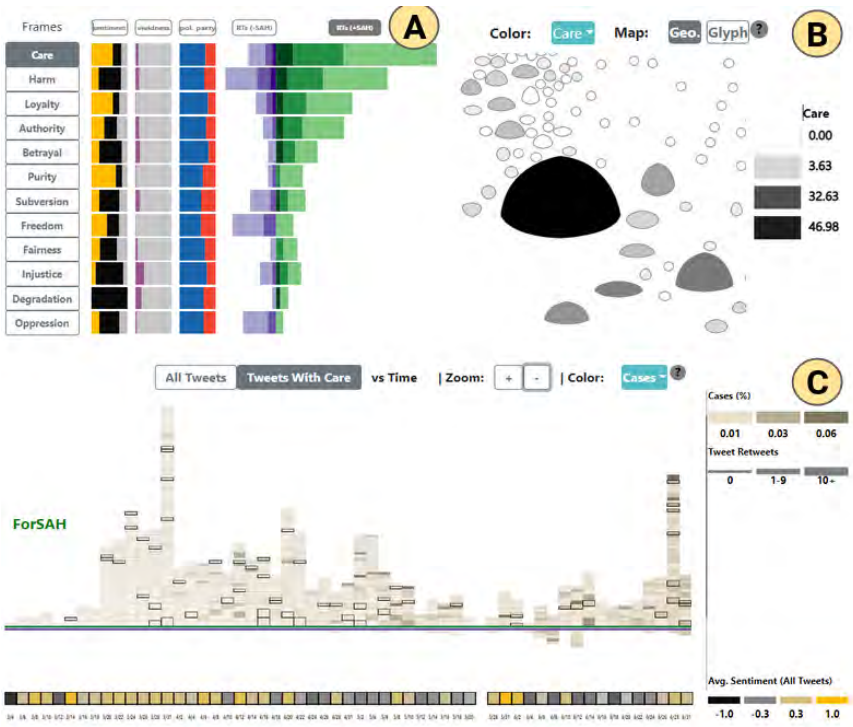


Figure 3.8: Overview for case study 1. (A) Summary panel of tweets in the SAH dataset, sorted by Popularity. Care and Harm are dominant, as all frames besides Freedom and Oppression are mostly for-SAH. (B) Glyph map of care-tweets by county, focused on L.A. (C) Timeline of tweets expressing Harm. Major peaks occur at the end of March and June, with a smaller peak in April.

Our collaborators were interested in which frames were dominant in the microblog data, as well as their vividness, popularity, sentiment, what temporal trends they followed, and the surrounding socioeconomic context around the tweets expressing each frame. Using the Summarization and Inference panels, the team confirmed relatively low popularity and a general lack of vividness across the corpus (<15% vividness). The MF with the highest average vividness was Injustice (6 vivid tweets out of 25) [WF 1]. Although the team had hypothesized a correlation between vividness and popularity, the Inference panel indicated a non-significant positive correlation ($p > .5$) [WF 2] .

By sorting the most popular frames, it became apparent that Care and Harm are the most popular frames expressed in Stay at Home tweets, and that they are both, surprisingly, predominantly in support of SAH orders. The communications experts noted that Care and Harm are complementary frames that form the virtue and vice around a single Moral Foundation, respectively, so this finding was intriguing. The group then noted that all

"virtues" such as Care were correlated with higher sentiment (yellow in the sentiment column) than all "vices" (black in the sentiment column), such as Harm [WF 1].

The group was then extremely surprised to note that, aside from Freedom and Oppression, most other frames were also in support of SAH orders (Purple bars showing for tweets were larger than Green bars in the summarization panel). These other frames were being expressed predominantly in democratic counties—even frames typically associated with conservative views, like Loyalty and Betrayal. Upon inspecting the timeline view, the group was able to confirm that most tweets are in support of SAH (predominantly above the centerline), and most tweets have low popularity (short tiles). In addition, they noted a correlation with increasing COVID-19 case numbers (darker tile shade), and overall more negative sentiment (more gray and black in the sentiment bar) as the pandemic evolves. By further examining individual tweets, they were able to determine that some viral tweets (taller tiles near the centerline) were, as expected, also vivid (e.g., *"Protesters attacking governors for stay at home orders. Claim it infringes upon their rights. Know what else infringes upon your rights? DEATH."*). Several other popular tweets reflected counter-intuitive information (e.g., the news that most of the NYC new COVID-19 cases were people following SAH orders), influencer SAH tweets, or, again, vivid pleas from overwhelmed nurses and doctors working in intensive care units [WF 1].

A visual computing researcher then noticed in the Timeline panel several spikes in the number of SAH tweets on March 31st, May 2nd, and July 28th, and a significant and surprising drop around May 28th. This sparked a vivid discussion involving the county map. Communications experts inferred the peaks corresponded to the beginning and end of several regional lockdowns, whereas the drop corresponded to the onset of social unrest related to the George Floyd events and Black Lives Matter (BLM) movement in the US [WF 1].

Based on the same Timeline panel, the group noticed the first wave of anti-quarantine (below the center x-axis) tweets, which, upon inspection in the Geospatial panel, appear to originate in counties with lower COVID-19 rates. Brushing the area around Los Angeles

in the county map, we noticed suburban counties (small counties surrounding large glyphs representing cities) had a higher Harm/Care tweet ratio (short, dark glyphs). The most senior communications expert hypothesized that tweets about Care originate mostly from large cities, whereas Harm is more evenly distributed among different suburban or rural populations [WF 1]. The group tested this hypothesis in the inference plot by showing the relationship between population and each frame in the Inference panel. Comparing both frames, the group found that Harm is indeed more prevalent in lower-population counties than Care (flat slope and smaller p-value) [WF 2].

The group concluded that the data collected was generally in favor of SAH orders, with increasing negative sentiment as pandemic fatigue set in. Although Care was predominant, most of the other frames expressed were also overall in support of SAH, with several interesting anomalies. They also noted the data was biased towards urban areas (large, tall glyphs in the geospatial map). Near the end of May, the BLM rhetoric appeared to have supplanted the SAH discourse, despite an expectation of increasing conservative or anti-SAH views due to pandemic fatigue. The team concluded that public policy messaging which had targeted Care-for-others appeared to have been effective [WF 2].

3.4.2 Moral Frames and Black Lives Matter

This second case study uses a subset of the Moral Foundations Twitter corpus [125] to compare tweets associated with the #BlackLivesMatter (BLM) movement and the #AllLivesMatter (ALM) movement between 2014 and 2016. The #BlackLivesMatter movement is a social movement that gained widespread popularity in 2014 in response to the disproportionate violence against African Americans, particularly by the police. The #AllLivesMatter movement, among other movements, arose as a critical response to the BLM movement. Both movements have become central to political discussions in the United States around issues such as police protections and criminal justice reforms, and played a role in the 2016 US presidential election [88]. Understanding the Moral Framework behind both movements can give insight into the driving forces behind these political movements.

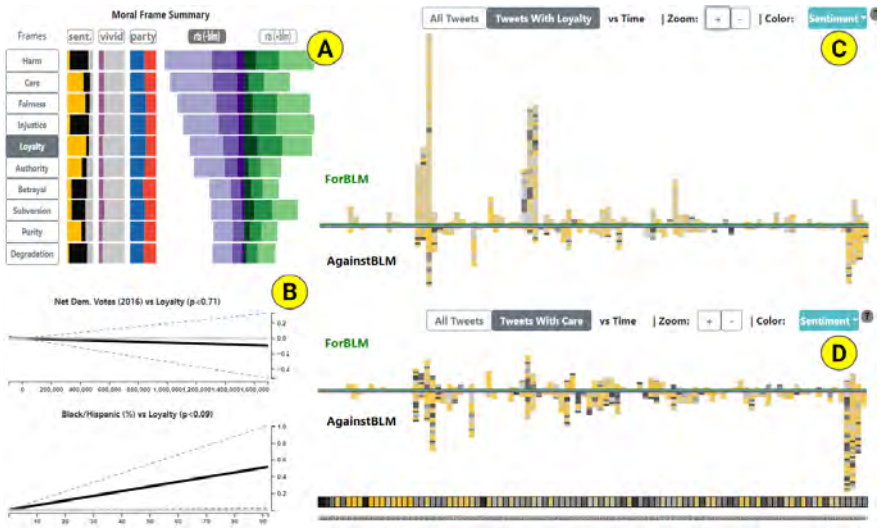


Figure 3.9: Overview of our BLM MF analysis. (A) Moral frame summary of tweets in the BLM dataset, sorted by percentage of tweets from democratic areas. Loyalty and Fairness are the dominant democratic frames, while betrayal is the most republican frame. (B) Correlations between demographics and frames. Republican votes are correlated with tweets for Authority, while the percentage of Black or Hispanic individuals is the strongest predictor of pro-Loyalty tweets. (C) Tweet timeline of pro-Loyalty tweets colored by tweet sentiment. Spikes in #BLM tweets occur around major protests. (D) Tweet timeline of pro-care tweets. A large spike in #bluelivesmatter tweets occurs during July 2016, in response to a police shooting in Dallas.

The team started the investigation with the Frame Summary Panel (Fig. 3.9-A) and sorting Moral Frames by political party. The frames most strongly associated with democratic areas (blue in the "party" column, top of the list) were Loyalty, Fairness, and Injustice. In contrast, Betrayal and Degradation were most often associated with more negative sentiment (black in the sentiment column) and republican areas (red in the "party" column, bottom of list) [WF 1].

The team also noted that despite being relatively balanced politically, a majority of tweets that express Care are in support of ALM (purple column larger than green column), which is unexpected, given that prior literature suggests that Care is more strongly associated with political liberals, as is the BLM movement. A communications researcher mentioned that Loyalty would be correlated with pro-BLM tweets since it is a "Binding Frame", and decided to explore further by viewing pro-Loyalty tweets in the Timeline panel (Fig. 3.9-C) [WF 1]. Four major spikes in activity can be seen, 3 of which are predominantly for BLM (more tweets above the center axis) and from relatively democratic areas (blue rectangles), while one is for ALM with a higher percentage of Republican areas (red rectangles). Investigating the

popular tweets from these time periods revealed the context behind these tweets: they are all tweets expressing solidarity for major protests related to police brutality: The Ferguson Protests [276], the 2015 Baltimore Protests [223], the 2015 Mizzou Protests [309], and the 2016 Dallas Protests in which 5 police officers were murdered [193] [WF 2].

Given the association between Care, political liberals, and SAH attitudes in our prior case studies, one researcher expressed interest in the fact that Care was not related to pro-BLM tweets "Care shows up in Republican areas, that's strange". In the timeline (Fig. 3.9-D), we see small spikes in activity around the Ferguson, Baltimore, and Dallas Protests. However, a visual computing researcher quickly noticed a large spike in tweets around the Dallas protest that are for ALM (below the center line) *"Oh, I see... Cops were killed in the protest. These people care for the cops ("blue lives") who were killed."* [WF 2].

Finally, the communications experts recalled that in the SAH analysis (first case study), Care was correlated with mask usage during COVID-19. Examining the counties expressing Care in this second study, they remarked on the shift in terms of geographical coverage: *"Care [in this second study] and Care [in the first study] is [not] correlated. That is counter-intuitive"* [WF 1]. Our collaborators theorized that this may reflect a shift in moral sentiment in partially republican areas between 2016 and after the 2020 pandemic, and a shift in priorities of the GOP rhetoric towards more Libertarian Rhetoric and away from Care [WF 2].

3.4.3 Expert Feedback

Overall, the domain experts found the MOTIV interface *"valuable on multiple fronts"*. In the data collection stage, they stated it helped them *"identify errors in the filtering process and refine queries"*. In the exploratory stage, it helped them *"understand the relationships between stance, moral frames, sentiment, and political affiliation"*. Through the map interface, the group stated they were *"able to identify some of the demographic biases in opinions"*. In later stages feedback was more explicit and supportive *"Oh, wow, the bubbles, that's *powerful*. I have to say, I am very impressed with the work. I am blown away by what you do.*

That's a really powerful graphic.". Finally, in the hypothesis generation stage, the interface helped the group *"find counterintuitive findings first (e.g., no moral frame was associated with only support or opposition to SAH attitudes) which we looked into further"*.

We asked the seven experts in NLP, Communications, and Causal Inference (CI) to rate the perceived usefulness of each component of MOTIV on a 5-point scale, as well as its helpfulness in identifying certain features: (T1) Popular frames. (T2) Relevant tweet features. (T3) Geopolitical/demographic trends. (T4) Tweet trends over time. Results and the respondent's domain expertise are shown in Fig. 3.10. The encodings were well received. Two collaborators diverged from the group in terms of map scores. These collaborators had been most active during the foraging stage where they had used primarily choropleths, and wished for a pop-up glyph explanation to facilitate visual scaffolding. According to this feedback, we added an on-demand glyph explanation. In addition, we asked the experts to rate specific system capabilities and helpful or not, and collected open-ended feedback. As indicated in Fig. 3.10, the experts had different foci and priorities served by the system.

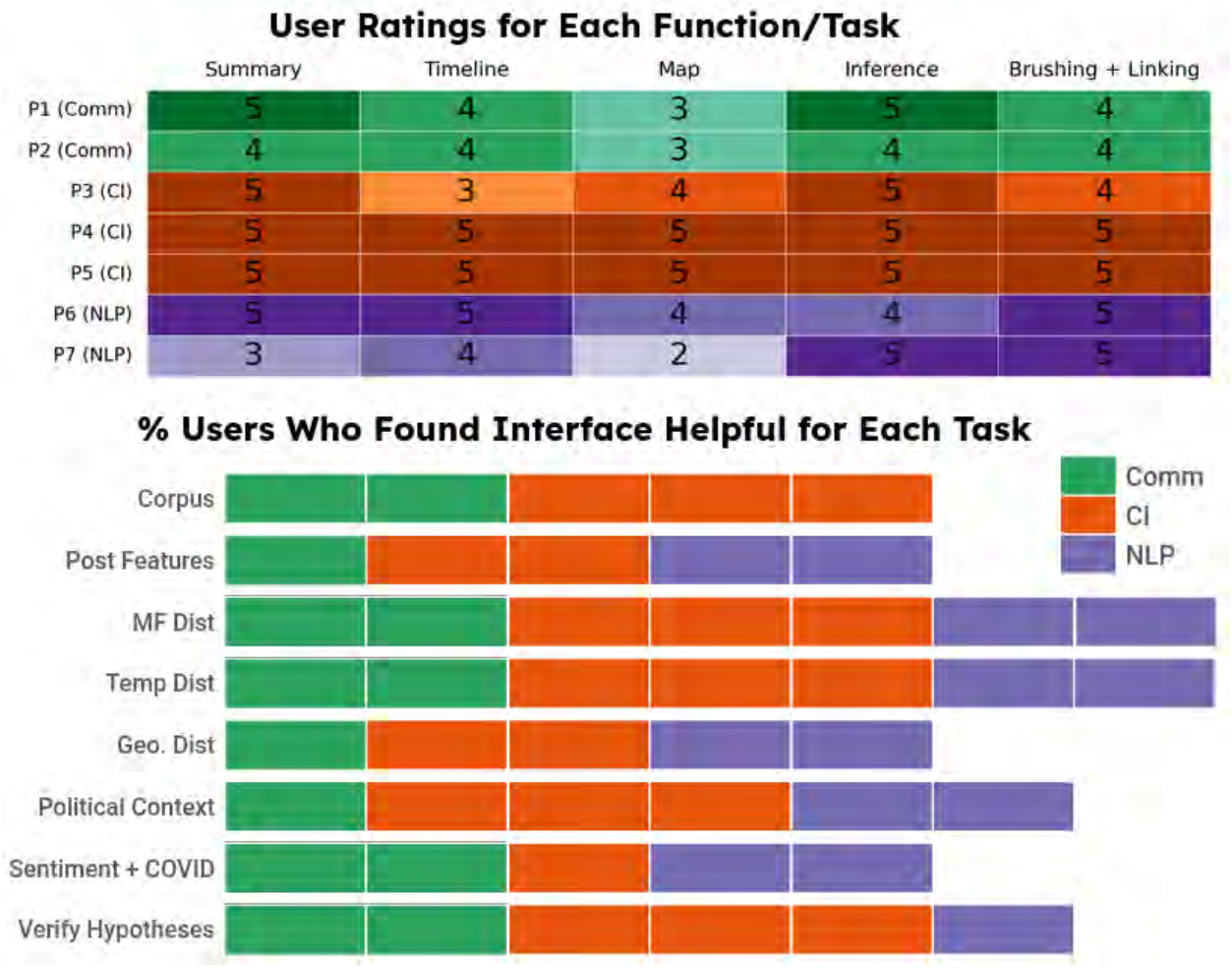


Figure 3.10: Results from user feedback questionnaire for users. (Top) Responses from the 5-point ratings for each component of the system. Hue corresponds to user's domain of expertise (Green = Communications, Orange = Causal Inference, Blue = Natural Language Processing), while luminance double-encodes user rating (darker = higher rating). (Bottom) % users who found the interface helpful for specific tasks.

3.5 Discussion and Conclusion

MOTIV was developed through participatory design over the course of a developing project. This project was built around a rapidly evolving study, in which the aims and data were constantly shifting during the entire course of the project. Despite these challenges, MOTIV's adoption as a research instrument by our collaborators is strong evidence of its value. The case studies and feedback we report further demonstrate that our approach, which blends data visualization, XAI, and social science, provides rich insights.

We thus discuss insights obtained from designing for these exceptional circumstances.

Collaborative Design During Data Foraging Our project was started at the height of the COVID-19 pandemic, and included multiple collaborators for different domains, with disparate design goals and an uncertain dataset. As a result, we found many challenges during the concurrent design and data foraging stage. Sedlmair [282] describe a common pitfall of design studies as starting a project before "real data is available", and working with tasks that are not well suited for design. In our case, we found that real data was present. However, due to the emerging nature of the pandemic, different parties disagreed repeatedly about the interesting aspects of the data and required tasks. As a result, the overall tasks and data considered for the visualization changed frequently during the data collection and prototyping stage, despite real data and suitable tasks being available from the start. Furthermore, our original topic changed so quickly that newly acquired data was considered "obsolete" (e.g. Twitter/X's free API is now no longer available). As a result, we saw significant benefit from making the final design highly flexible in terms of the design and user control for exploration of different variables, which made adapting the interface to other problems, such as BLM, useful. Last but not least, with respect to winnowing, due to the urgency of the pandemic project we felt that we did not have a choice to interrupt or withdraw from the project.

During the prototyping and implementation stage, we found the largest benefit to performing rapid updates with real data in order to draw out better conclusions from collaborators. Presenting data from one collaborator to the group allowed for better input from additional collaborators, such as when discussing the temporal changes in the quantity of the tweets. These temporal changes could be attributed to either changes in the natural language processing, or when tweets were inspected in more detail when analyzed by the communications experts. Also, due to the pandemic time-pressure, there was often not enough time to include legends, for example colormaps, which led to additional discussions.

In terms of collaboration, we worked with experts from multiple disciplines. A main issue was a lack of agreement between individuals about the scope of the project. Initially,

collaborators stated they were mainly interested in a basic COVID-19 dashboard alongside a county map of political affiliation. However, analyzing implied goals and workflows during lab meetings and discussions led to different directions. As a result, careful note-taking, and rapid development prototypes that used the actual data during the foraging stage in order to gain as much natural feedback as possible turned out to be a more effective way of soliciting design requirements. An additional pandemic challenge was related to the team not having had enough time to absorb team science principles.

Finally, when publishing results, the project met difficulties due to differing expectations: our collaborators and program officers expected valuable insights and a useful system, which the project provided, while reviewers alternatively anticipated large-scale corpuses, general or automated tools, or detailed validation of published NLP algorithms.

Visual Complexity During our design process, a common issue we faced was reconciling the visual complexity necessitated by our task requirements with the simplicity required by our collaborators. Earlier iterations were often complex, as our original tasks were focused on inspecting many variables, with little insight into which were most important. We found that an approach grounded in visual scaffolding [198] was most effective, where we built on encodings such as bar charts and choropleth maps that are familiar to users, and we added features and variations to address tasks that could not be met using traditional methods.

One example was the design of the glyph map. During prototyping, collaborators showed a strong preference for choropleth maps due to their familiarity with it over unfamiliar methods that could show multiple variables. The glyph-based encoding helped in this sense, and also helped with the identification of major cities and counties of interest. By inspecting the alternative choropleth map, the viewers were able to mentally anchor the location of different areas in the glyph map even after being distorted by the force-layout, yielding positive feedback.

Ultimately, we found our collaborators' visual literacy improved during the project. A focus on identifying well-motivated incremental changes to existing designs may help appli-

cations be more accessible and generalizable to domain experts and wider audiences.

Collaborative Hypothesis Testing A recurring challenge was to support the exploration of a large problem space to help identify interesting avenues for further investigation by collaborators. During the initial stages of our project, we found that a common workflow was that collaborators with computing backgrounds would share preliminary, exploratory data analysis using a variety of NLP techniques. Findings would be shared with experts in communications, who would identify potentially interesting findings. Follow-up statistical testing could then be performed to identify useful results.

In terms of transferability to other studies, our work captures challenges relevant to concurrent design for emerging, urgent problems, which differ from typical design experiences. Most user-centered design approaches focus on the workflows of individuals, which can generally be obtained based on input from the user themselves. We found benefit in characterizing the workflows that occur through interactions between domain experts across domains, and focusing on adding in visualization that helps support gaps in information sharing between groups. For example, the Inference panel helped share results between the workflows of our statistical researchers and those with backgrounds in communications and moral foundation theory.

Assumptions and Limitations MOTIV inherits the biases in the data we use. For example, most Twitter users tend to be younger and more democratic (liberal) than the average American. However, despite being an underrepresented sample, Twitter users are more politically active and may therefore be more likely to start discussions with others regarding political issues [354]. Additionally, we are limited by our reliance on regional demographics, which we assume correlate with tweet content on aggregate. Furthermore, we focus on high-precision keywords for tweet relevance, and thus it is possible that less precise or more obscure keywords are more popular with different demographics. As we don't have access to the underlying political affiliation of each individual, this remains a potential source of bias in the data. Despite this, our dataset is representative of the subset of US based

Twitter users who allow their location to be known. For these specific datasets representing this subgroup of the population, with stated and known representation limitations, our collaborators have described and shared these analyses and insights, which speaks to the relevance of these datasets to the field of social communication. While we only look at relatively small ($<$ 2000 tweets) datasets here, we have explored the use of MOTIV for a significantly larger set of 100,000+ non-geotagged tweets [93]. Ideally, future work would look into more robust models for moral foundation theory using larger datasets or leveraging pre-trained large language models, which became widespread after this project was completed.

In terms of generalizability, the summary, map, and inference views can scale to arbitrary sizes. However, the Timeline view would need adjustments, as it shows encodings for all tweets simultaneously. Based on prototyping (see supplemental materials), we found that variations of aggregated timelines and sparklines work well for showing trends, while context and influential tweets could be investigated on-demand by showing the most popular tweets within user-selected criteria at different time points of interest. The usage of improved automatic labeling could allow the system to work with arbitrary datasets, although current state-of-the-art models do not perform well compared to manual labeling [125].

In this work we introduced a novel methodology for exploring moral frame political discourse via analysis of social media. This approach is, in our collaborators' perspective, data-agnostic, and we show it can produce valuable insights on two separate datasets. The approach is not (nor should be) limited to social media datasets, and could apply to formal polls or surveys. Our approach draws on methods from data visualization, explainable machine learning, and social science to provide rich insights into how the public formulates arguments over social media. By integrating Moral Foundations' theory with custom visual encodings and interactions, we provide a novel and rich approach to Twitter data visualization systems. Through a detailed analysis in two case studies of tweets related to Stay-at-home orders in the U.S. during the COVID-19 pandemic and the Black Lives Matter movement, we show this approach can identify key events that affect the nature of politi-

cal discourse, even without the presence of explicit labels, in addition to insights into how moral values regarding politicized movements are disseminated by different social groups. We identify design lessons relevant to working with domain scientists who have limited visual literacy, yet are keenly interested in quick hypothesis generation and testing. While we focus on the application of a specific set of Moral Foundations to frame our analysis, our approach could also apply to problems centered around comparing classes of tweets, such as in topic modeling, or when clustering latent variables learned using recurrent neural networks.

3.6 Acknowledgments

We thank Juan Trelles and our other colleagues at the Electronic Visualization Laboratory for their technical and emotional support. This work was partially supported by NSF award IIS-2031095 and NIH award NCI-R01 CA258827.

Additional Material including previous submissions, videos of the system, additional details about the evaluation, and examples of designs made during the prototyping phase can be found at `https://osf.io/ygkzn/?view_only=6310c0886938415391d977b8aae8b749`

3.7 Chapter Conclusion

This chapter presents my work in attempting to incorporate interactive, spatial machine learning methods and model building directly into the visual computing system through the use of GAMs. Additionally, I explored design issues related to domain characterization and encoding design when dealing with developing datasets, in the context of geospatial + temporal Microblog and COVID-19 pandemic data by supporting more approaches for generalized exploration of larger design spaces, and discussed some issues in the design process with collaborative work that involves VC + ML. In my next work, I return to spatial unsupervised modeling and introduce methods of generating domain-specific model explanations and techniques for improving the workflows of collaborative model buildings by incorporating spatially aware information-scent into an integrated data visualization and visual model building system.

Chapter 4

(DASS) Actionable, Interactive Clustering of Spatial Radiation Therapy Plans

One important area of the VC+ML pipeline is leveraging interactive model development into the user workflows. This chapter extends the previous work with interactive model development by incorporating live model iteration through guided spatial feature and parameter tuning, as well as developing and encoding domain-specific simplified cluster model explanations. Developing applicable clinical machine learning models is a difficult task when the data includes spatial information, for example, radiation dose distributions across adjacent organs at risk. We describe the co-design of a modeling system, DASS, to support the hybrid human-machine development and validation of predictive models for estimating long-term toxicities related to radiotherapy doses in head and neck cancer patients. Developed in collaboration with domain experts in oncology and data mining, DASS incorporates human-in-the-loop visual steering, spatial data, and explainable AI to augment domain knowledge with automatic data mining. We demonstrate DASS with the development of two practical clinical stratification models and report feedback from domain experts. Finally, we describe the design lessons learned from this collaborative experience.

The contents of this chapter were originally presented at the 2023 IEEE Eurovis Conference and published in Computer Graphics Forum [344]. A companion paper reporting the spatial clusters built using this interface was published in Frontiers in Oncology [350].

4.1 Abstract

Developing applicable clinical machine learning models is a difficult task when the data includes spatial information, for example, radiation dose distributions across adjacent organs

at risk. We describe the co-design of a modeling system, DASS, to support the hybrid human-machine development and validation of predictive models for estimating long-term toxicities related to radiotherapy doses in head and neck cancer patients. Developed in collaboration with domain experts in oncology and data mining, DASS incorporates human-in-the-loop visual steering, spatial data, and explainable AI to augment domain knowledge with automatic data mining. We demonstrate DASS with the development of two practical clinical stratification models and report feedback from domain experts. Finally, we describe the design lessons learned from this collaborative experience.

4.2 Introduction

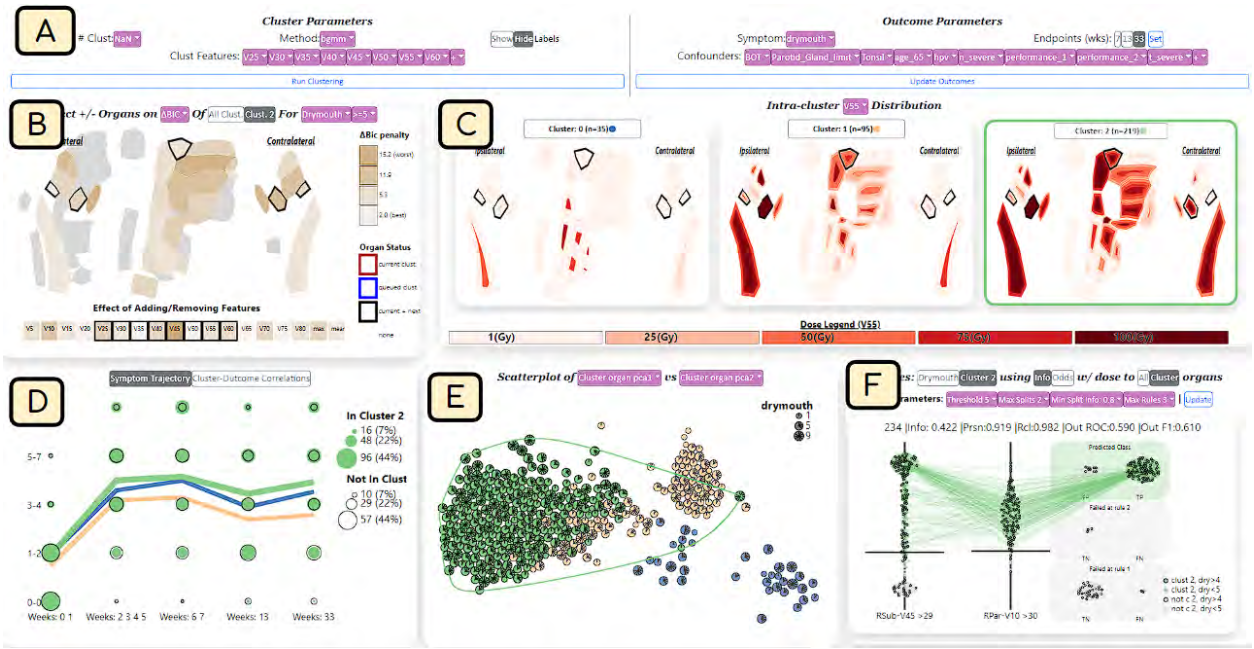


Figure 4.1: DASS interactive model building for head and neck cancer. A) Control panel for changing cluster parameters and the desired outcome. B) Additive Effect panel showing the effect of changing cluster features. C) Intra-cluster dose distribution plot. D) Outcome plot showing the symptom ratings of patients over time within each cluster. E) Stylized scatterplot showing cohort projections. F) Rule builder view, showing a rule-based mimic model that predicts patients in the selected cluster.

Precision radiotherapy (RT) is a medical paradigm that seeks to personalize cancer RT and care for an individual patient, based on data from cohorts of similar patients. Because for many site-specific cancers, the treatment depends on the location and spread of the disease, modern approaches to precision RT aim to leverage spatial patient-specific information such

as anatomical data drawn from CT scans [345, 346]. In conjunction with the cohort data, this information can then be used to improve patient outcomes such as survival or quality of life after treatment.

In this context, machine learning (ML) models are powerful tools for stratifying the cohort data in meaningful ways, for example into patient groups at high-risk versus low-risk of developing treatment-related symptoms. However, developing applicable clinical ML models for patient stratification is difficult when the data includes spatial information, for example, radiation dose distributions across adjacent organs at risk. In addition, while ML approaches often work well with large oncology data, automated model-building approaches using smaller cohorts often perform poorly when deployed in practice [207]. Furthermore, prediction using treatment plans and qualitative outcomes such as symptom ratings is particularly difficult. This results in simpler models that may underperform or complex models that are very likely to overfit. With advancements in explainable AI techniques, we can better probe models and iteratively find ways of improving models that properly leverage domain knowledge, helping us avoid issues with poor generalization and overfitting, while improving on standard statistical approaches. These combined issues make RT cohort modeling well-suited for a human-machine mixed-initiative system.

In this work, we present a visual steering approach for creating patient stratifications of head and neck cancer (HNC) patients based on 3-dimensional dose distributions to organs-at-risk, to separate patients at high risk of experiencing long-term side effects. Unlike the current state of the art, our approach supports interactively exploring and visualizing high-dimensional spatial dose distributions, the temporal analysis of RT cohort data, access to both individual patient data and patient distribution within a cluster, constructing unsupervised rule models to help explain the clusters, and iteratively refining and exploring parameters to create actionable stratifications. We implement this approach in *Dose Analytics and Symptom Stratifier* (DASS), a visual computing system designed to allow for the development and exploration of patient stratifications according to different symptoms of

interest. We describe two case studies of applying DASS and show how it has been used to improve existing outcome models. Finally, we provide design lessons gained through this collaborative visual steering design.

4.3 Background

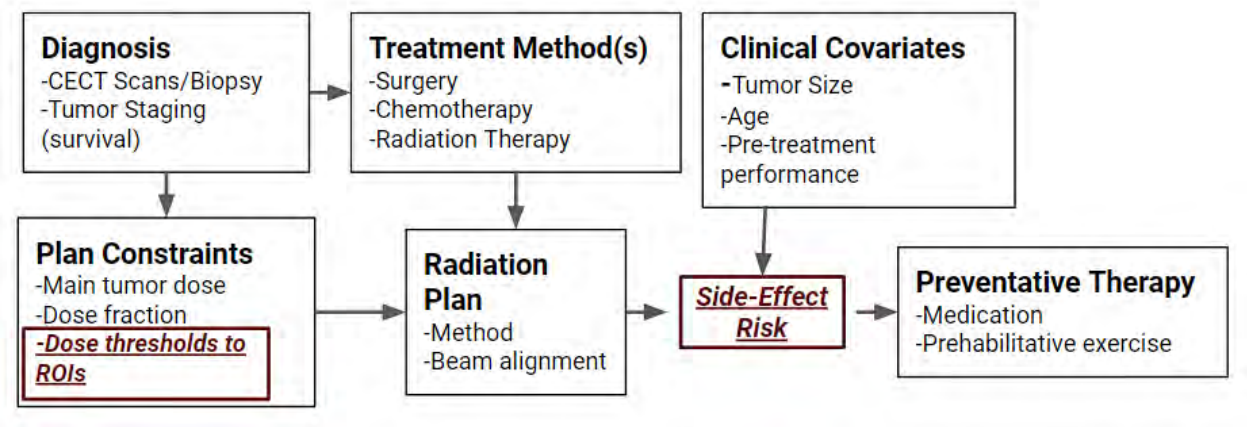


Figure 4.2: Physician Workflow. The main items of interest for this project are to help establish a stratification of patients' risk of certain side-effects from their radiation plan (Side-Effect red box), which can then be used to identify which patients require additional preventative treatment, as well as help identify dose thresholds (Dose thresholds red box) that can be input as soft-constraints during treatment planning.

Head and neck oncology has seen large increases in patient survival due to a shift from smoking-driven tumors to less aggressive HPV-driven tumors. This increase in survival has resulted in a shift in priorities towards increasing the quality of life of patients: radiation to organs near the primary tumor during treatment can lead to tissue damage, resulting in long-term side effects [89, 170]. Predicting when symptoms driven by spatial tissue damage occur is thus an understudied area of interest to oncologists, as it can help identify better treatment guidelines.

When performing the initial diagnosis, oncologists rely on patient history and clinical staging that rank the size and spread of the tumor [245] to determine the method of treatment to optimize patient survival. However, after the treatment methodology is established, predictive models are needed to identify patients that may need preventative treatment for serious side effects. Common criteria for treatment decisions are tumor location, T-staging, which is a 5-level stratification based on the size and penetration of the primary tumor into

surrounding tissue, and N-staging, which is a 4-stage stratification based on the size and spread of secondary tumors in the lymphatic system [245]. These criteria are combined with other factors to decide on treatment plans for patients, which may include radiation therapy and other treatments like concurrent chemotherapy. Modern treatment plans are typically chosen to maximize the chance of survival in patients. However, severe side effects (toxicities) are common in many patients as a result of damage to surrounding organs from radiation treatment [89]. A diagram of the clinical workflow is available in the supplementary material (Figure A1).

In particular, predicting tissue damage from radiation therapy in head and neck cancer (HNC) patients is a challenge due to the high number of treatment parameters and high number of organs that may factor into side effects. For example, drymouth is often caused by radiation damage to a subset of the salivary glands. Identifying when failure may occur is a difficult modeling task, in which one needs to consider the glands as a spatially interrelated system, as some may compensate for damage to other glands. Additionally, each organ may have a separate non-linear response to the radiation dose over time, and symptom severity varies throughout treatment. Furthermore, the large numbers of HNC patients in a cohort and the dimensionality of the data pose a challenge in terms of visual analysis. Finally, human modelers also require access to individual patient data, as well as to the patient distribution within a cluster to make informed inferences about patient outcomes.

4.4 Related Work

4.4.1 Visual Analysis of Cohort Data

Several applications of visual analysis have focused on different algorithmic approaches for clustering patients [206, 229, 346]. Visualization tools often extend these approaches by allowing human-in-the-loop analysis to identify sub-cohorts [24, 158, 376]. Other systems have focused on comparison of cohorts to discover differences in disease progression [194], genetics [106], cancer treatment disparities [288], but, unlike our work, these systems do not focus on model building.

Many systems use clustering [215] and dimensionality reduction [85, 240] on key features to guide explorations over high-dimensionality data. Some tools have looked at visual analytics for creating clusters with unstructured health data [44, 107, 164], while other systems incorporate temporal clustering methods [95, 116, 330, 369, 375]. However, these systems do not attempt to incorporate spatial information in their clustering models, as we do. Additionally, none of these systems link detailed treatment plans to qualitative patient outcomes in the cohorts, as we do.

4.4.2 Visualization of Medical Image Data

Work in visual computing with medical imaging often focuses on linking spatial features to external variables to support exploration for domain experts. Early work focused on visualizing spatial imaging data with open source tools (MITK [355]) and introduced integration of spatial and non-spatial linked views [112].

Specialized approaches have been developed to explore cohort features in other domains such as tissue imaging [92, 138, 334], neuroscience [14, 134, 143, 190], and lumbar spine features [56, 151].

Focusing on cohorts of RT data, BladderRunner [262] visualized cohorts of prostate cancer patients which used a mixture of T-SNE and Gaussian mean-shift clustering to group patients based on bladder shape. VAPOR [100] extended their work to consider RT-induced treatment toxicity. Other work has extended these results to explore uncertainty in RT data for visual analysis [113, 271] and predictive models [101]. However, these approaches do not deal with HNC oncology treatment, which has more complex treatment and symptom patterns but lower temporal variability.

Previous HNC work has used spatial data to cluster patients based on tumor spread to lymph nodes [186]. Many techniques rely on simplified representations of anatomical data to allow for better analysis of high-dimensional data [151,262,343,345]. While these works often deal with feature engineering, none of them focus on directly altering the model in parallel with the visual analysis, as we do. Additionally, we uniquely provide tools for validating the

feasibility of the underlying model's logic and embedding anatomical data directly into the system.

4.4.3 Visual Steering and Interactive Machine Learning

In the medical domain, several projects have developed visualization systems around the workflows of clinical model builders and biostatisticians with a focus on regression models [77]. Raidou et. al [259,260] proposed a tool for visual analysis of regression-based Tissue Complication Probability models, with a focus on uncertainty. However, these approaches do not focus on clustering or stratification models, as we do.

Other work has focused on actionable explanations for pre-built models for clinicians, such as normal tissue complication models [372], binary classifiers [53], case-based reasoning [201, 218], and black box models [52]. For explainable AI, DrugExplorer [329] proposed a model for user-centered XAI alongside a system for exploring graph-neural-networks for drug repurposing. However, none of these approaches tackle iterative probing and model development, or capturing spatial information in their data, as we do.

Additionally, our work uses interactive rule mining to help explain the clusters. Many systems have worked on aggregated visualization of rules [225, 290, 302, 303, 315, 362, 366], and used interactive rule mining to approximate more complex models [217]. Our approach differs from these in that we include a novel rule mining algorithm focused on matching clinical use cases, along with a novel visual encoding that allows for interactive parameter tuning.

4.5 Methods

The DASS design is rooted in our earlier experience with clinical stratification models that relied on forward search for feature selection for clustering [346]. Fully automated parameter searches yielded models that performed well on a single performance metric. However, when the clusters were inspected by clinical collaborators, they would often find issues with the organs used, such as organs that are completely unrelated to the outcome, or smaller organs

that they felt should be included. Thus, we introduced a human-in-the-loop forward search directly into our front-end alongside model explanations to help improve the process of iterating on our clusters.

User-guided search has two additional benefits. First, our clinician collaborators wished to specify desired characteristics of the models, which led to a need to explore multiple alternative outcomes or starting points based on these desired characteristics. Second, collaborator input is required when balancing model performance, the feasibility of the organs considered, and the number of organs considered. For example, we found that in one model, including both the soft and hard palate had identical effects on the outcome. Thus, the decision came down to the clinicians, who helped us identify which one was of more clinical importance.

Furthermore, in previous work, we attempted to find clusters through hyperparameter search or using predefined cluster features. However, we found that neither approach performed well. Automatic feature selection led to clusters that focus on organs that served as positional *indicator features*, such as the oral cavity [346], but are not causally linked to outcomes and resulted in model explanations that are not well-received. Notably, we found that the brainstem and brachial plexus nerves often appeared as predictors, despite clinicians noting that neither can be associated with any of the outcomes being predicted. Such models work well, but lack causal plausibility, which hinders adoption and cannot be generalized to treatment guidelines. The DASS design specifically addresses these problems through its back-end and front-end.

4.5.1 Data

Data were collected from a cohort of 349 HNC patients treated at the MD Anderson Cancer Center using Radiation Therapy, with or without chemotherapy, using a 7-week treatment course. We consider three types of data: spatial dosimetric data taken from the patient's treatment plan; unstructured clinical data taken from the patient's health record; and temporal information on the patient's self-reported side effects taken during and after treatment.

All values are positive ordinal values. Symptom ratings for individuals are discrete, while dose values are continuous.

Diagnostic images were taken at the time of diagnosis, and 40 organs of interest were segmented from these images and considered in the treatment plan. Dose treatment plans were extracted for each organ of each patient. We include 3-dimensional information on the cumulative dose received by each organ during treatment. We use the notation "VX" to denote the maximum dose that penetrates X% of the organ. For each organ, we consider the V5-V95 range in increments of 5, as well as the mean and maximum dose.

For outcomes, patients were asked to fill out an MD Anderson Symptom Inventory (MDASI) questionnaire [275]. This inventory includes self-reported symptoms for 28 different items, such as drymouth and pain, on a scale of 1-10. We also include secondary variables that may be used as confounders in the patient outcomes taken from electronic health record data, which we generally treat as binary confounding variables.

4.5.2 Collaboration

Our work was done as part of an ongoing collaboration between data scientists and research oncologists at three US sites. DASS was commissioned to serve first and foremost the needs of the model builders, but to also facilitate clinician input and feedback on the models. Remote meetings were held weekly, during which we would get feedback on designs, and update project goals based on feedback and current results. Examples of prototypes during this phase are included in the supplement Appendix B.

We followed an Activity-Centered Design (ACD) process [199], which is a methodology conceived to better support designing for domain experts by focusing on existing user workflows and activities. The approach has higher success rates in interdisciplinary settings than Human-Centered Design (63% versus 25%) [199]. We focused on the workflows around the development of clinically applicable models, as well as the associated data analysis and verification required to validate and publish the results.

4.5.3 Task Analysis

Modeling Requirements The goal of our project was to aid in the development of an interpretable decision-support tool for clinicians to help identify HNC patients at high risk of long-term severe (self-reported rating > 4 on a 10-point scale) side effects due to radiation damage. We focus on HNC patients as the sensitivity of organs in the head and neck makes detection of quality of life measures in survivors a difficult, under-explored application. Our collaborators were specifically interested in a model that could improve on existing clinical systems by incorporating sets of related organs that together support specific functions, and thus should be treated as a system.

Our system was designed to be used for asymmetric collaborative analysis, which would be handled by model-builders with expertise in the underlying algorithms, with clinicians providing input and feedback. Therefore, we identified requirements for the models themselves, as well as the steps needed to create and validate each model. For our models, we derived the following requirements:

Actionable: Usable in a practical setting. In a typical workflow, clinicians use risk stratifications that rank a patient's risk of survival, which are then integrated into a holistic treatment plan. As such, we require that our models output a simple ranking for each patient, as well as insights that are usable without access to the models. Access to individual patient data, as well as the patient distribution in each cluster, in terms of both doses and symptoms, was necessary.

Plausible: Generalize well to a real-world setting. The underlying features that lead to a patient being classified as high-risk must be easy to understand in their spatial context. The models must also place patients in the high-risk group because they received a high dose to a specific set of organs, and the set of organs considered must be mechanistically linked to the outcome of interest.

Transparent: Be easily probed, assess the plausibility of the models, and identify edge cases in the models. We also needed to be able to demonstrate the plausibility of the models

and explain its internal logic to readers with a clinical background.

Visualization Tasks Based on these requirements, we designed a dose-based stratification methodology that clustered 2D dose distributions to a set of organs and used the resulting patient clusters as a proxy for patient risk. Our visual front-end is designed around visual steering, which uses information scent and visual cues to guide our team through the process of selecting, validating, and refining the range of potential parameters for the models to balance different performance metrics and model plausibility. Because this task requires significant knowledge of the models when adjusting parameters, our interactive system is designed to be used directly by models builders and visual computer experts, with encodings designed to allow model builders to communicate intermediate results to clinicians and domain experts.

Through a series of iterative sessions where we developed models and discussed them with our collaborators, we identified the following Activities and Tasks for our visual interface:

- **A1** - Given a symptom, find optimal cluster parameters
  - **T1** Find organs causally related to the symptom of interest.
  - **T2** Identify a window in the dose-volume histogram that best stratifies the cohort.
  - **T3** Validate a choice of clustering algorithm and parameters
- **A2** - Validate that the logic of a model is causal and plausible
  - **T4** Examine the dose distribution of each cluster and where the doses differ.
  - **T5** Verify if the cluster with the highest symptom risk also has higher doses to the organs used in the clustering.
  - **T6** Identify confounders that may impact risk prediction.
  - **T7** Validate the predictive accuracy of the clusters.
- **A3** - Examine and explain individual clusters
  - **T8** Identify the organ doses that most distinguish each cluster.

– **T9** Evaluate differences in symptom trajectory between clusters over time.

A1 deals with the development of models, while activities A2 and A3 help to quantify the models and provide feedback to improve the parameters in A1. A2 is a requirement for clinical publishable findings, while A3 is important for identifying any insights that can be drawn from the final model. For example, once a model is validated, finding that the high-risk cluster for taste dysfunction tends to have a very high maximum dose to the tongue may indicate that future work should investigate the effect of tongue dose on outcomes in more detail.

4.5.4 Back-end Algorithms

Modeling. DASS allows selecting from a range of clustering algorithms: K-nearest-neighbors, Hierarchical clustering, spectral clustering, and a Gaussian Mixture Model. After several iterations, we converged to a Bayesian variant of a Gaussian mixture model for all cluster outcomes. Once a set of organs and a dose-volume histogram (DVH) is identified, these features are encoded as a vector for each patient of size *#organs * window-size*. Patient vectors are clustered, which are ranked based on the sum of the mean doses to each organ included in the cluster. Ideally, this will result in the highest rank cluster (high dose) being the most correlated with the outcome.

To evaluate the resulting models, we also need to specify a symptom and time point to use as the outcome of interest. We then convert ratings to a binary outcome using a severity threshold. After discussion with our collaborators, the default was a symptom rating above 4 out of 10 at 6 months after treatment.

Once our clusters and outcomes are identified, we perform multivariate correlation analysis using a likelihood ratio test (LRT) to assess the correlation between each cluster and the outcome of interest, using a set of clinical confounders interactively specified. The (LRT) builds a regression model with and without each cluster, and we can compare both models to assess the impact that each cluster has on the goodness of fit of the model.

From this, we can calculate an odds-ratio and statistical significance p-value for each cluster, as well as the Bayesian Information Change (BIC) [156]. BIC and AIC are estimates of the goodness of fit of a model that include a penalty for the number of variables considered, in order to prevent overfitting, where lower scores indicate better fits [156]. For BIC, reductions in score relative to a baseline model of at least 2 indicate reasonable evidence, while reductions of at least 6 indicate "strong" evidence of improvement [257]. This provides a set of different metrics for assessing the cluster quality in terms of stratifying the cohort.

In addition, to assessing the quality of the current clustering, we provide a forward search in which we alter the existing cluster parameters by adding or removing either a single organ or a single feature from the dose-volume histogram window. We then re-cluster the cohort, and evaluate the new p-value, AIC, and BIC for the new clusters, relative to that of the existing cluster. These metrics are used to provide information scent for users when performing a forward search of the data.

Rule Mining. To help explain the clusters, we designed a constrained rule mining algorithm and used it to generate a set of dose thresholds that work as a classifier. Our algorithm looks for splits among all dose features in the dataset to find a set that maximizes the mutual information between the splits and a binary outcome. This algorithm is designed to approximate standard rule mining, with the following additional constraints so that the results approximate the rules used by clinicians when specifying dose thresholds: 1) Monotonicity – the high-probability subset for each split in the data must either always be the group above or always in the group below the threshold; 2) Minimalism – The algorithm can only use one dose-feature for each organ; 3) Informative – each "rule" in the ruleset must have a minimal predictive value (user-set) on its own.

Specifically, the algorithm works as follows: 1) we calculate the mutual information gain between each feature split within each ROI (e.g. V40 to the Tongue > 40) and the binary outcome of interest; 2) of the resulting splits, we select the k most important splits; 3) for each of the k best rules, we test combinations of all other splits in step 1 that do not share

the same ROI, and calculate the new mutual information gain of the combined rules. Rules are combined using the AND operator (i.e. the patients must satisfy all rules); 4) steps 2-3 are repeated until no improvement is seen in the mutual information gain. To speed up the algorithm, pruning parameters used to speed up the search can be adjusted in the interface.

To speed up the algorithm our backend also allows for pruning parameters in the forward search by: 1) increasing the distance between dose values when testing thresholds (granularity); 2) limiting the candidate rule sets used when performing forward search to only the top k at each step; 3) limiting the maximum number of rules in a ruleset. These pruning parameters are pre-set based on testing but can be adjusted in the interface.

Implementation All data pre-processing and modeling is done using Python with NumPy, Pandas, and Flask for the back-end. Clustering and dimensionality reduction is performed using the scikit-learn package, while statistical tests use the statsmodels package. Our system frontend is implemented using React and D3.js.

4.6 Front-end Design

The DASS front-end (Fig. 4.1) is composed of 6 panels: a cluster dose view (Fig. 4.1-B) that shows the within-organ dose distribution for each cluster (A2), an additive effects view (Fig. 4.1-C) that shows the estimated impact of adding or removing features from the cluster on the specified outcome (A1), an outcome view (Fig. 4.1-D) that shows the different symptom ratings over time for each cluster (A2-A3), a configurable scatterplot view (Fig. 4.1-E) that shows a 2D projection of all the patients in either the dose or outcome space (A2), a rule view (Fig. 4.1-F) that shows a set of dose thresholds that best separates a cluster of interest (A1-A3), and a control panel (Fig. 4.1-A) that allows users to specify the cluster parameters and outcomes of interest. We arrived at this design following a parallel prototyping process, with multiple design alternatives and repeated feedback. This process is illustrated in the supplemental materials.

To better support the analytical workflow, we use a categorical color scale for cluster

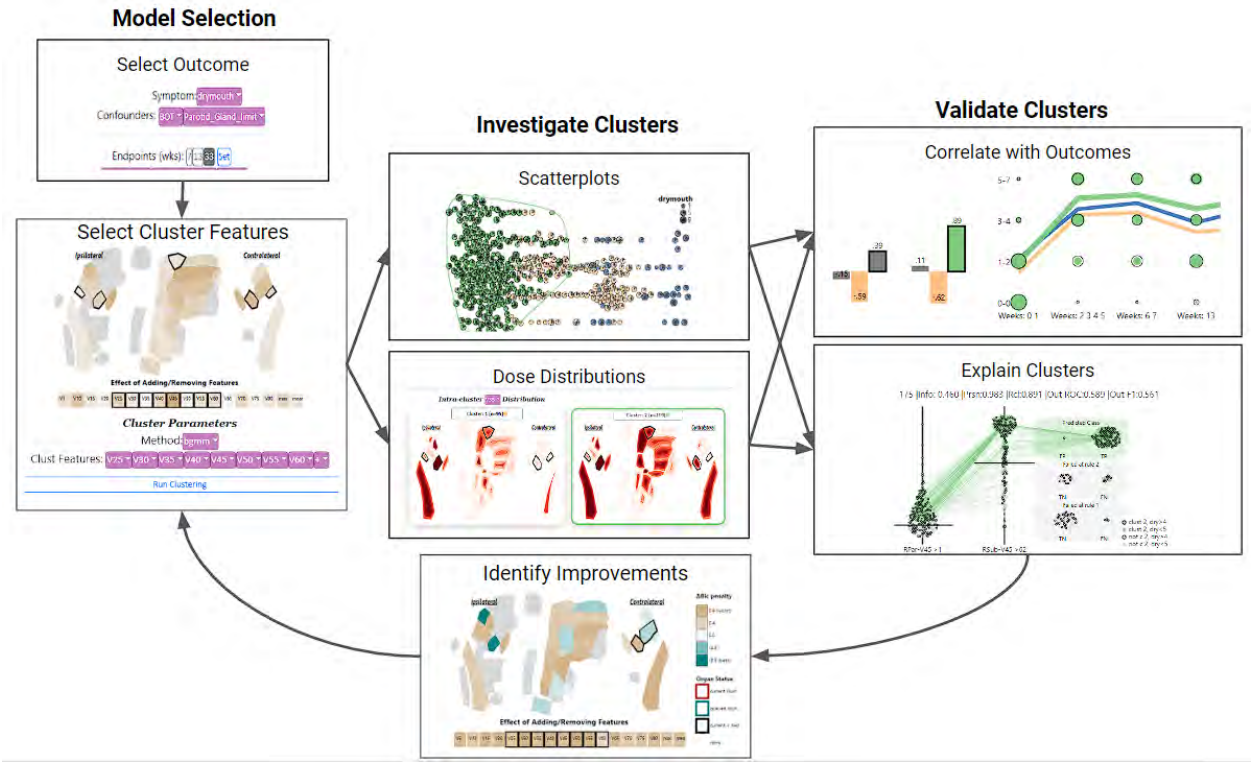


Figure 4.3: DASS model-building workflow. First, a desired outcome is selected, along with the initial clustering parameters, which can be drawn from prior literature or informed from information metrics in the explanation view. After clusters are generated, the inter-cluster distributions can be investigated using the dose distribution and configurable scatterplot views. Cluster performance can then be validated by investigating the inter-cluster symptom trajectories and correlations with outcomes. A rule-based classifier can also be used to produce explanations for the high or low-risk clusters based on dose thresholds. Once clusters are investigated, the Additive Effects panel can identify potential changes to the clustering parameters that could improve the model performance.

membership. Analysts can select a specific cluster, which is used to populate the temporal outcome and rule views, and brush in all other linked views. By default, DASS automatically selects for brushing the highest dose cluster, as this cluster was typically of the most interest to our clinicians.

4.6.1 Visual Scaffolding

When dealing with organ data, understanding the relative position of each organ is essential for analysis of the relationships between organs and side-effects. Specifically, dose values are correlated with location, and it is important to identify situations where organs may be linked to toxicities due to their centrality and proximity to nearby organs rather than being directly causally linked.

In previous work, we represented the set of organs as a stylized plot showing each organ

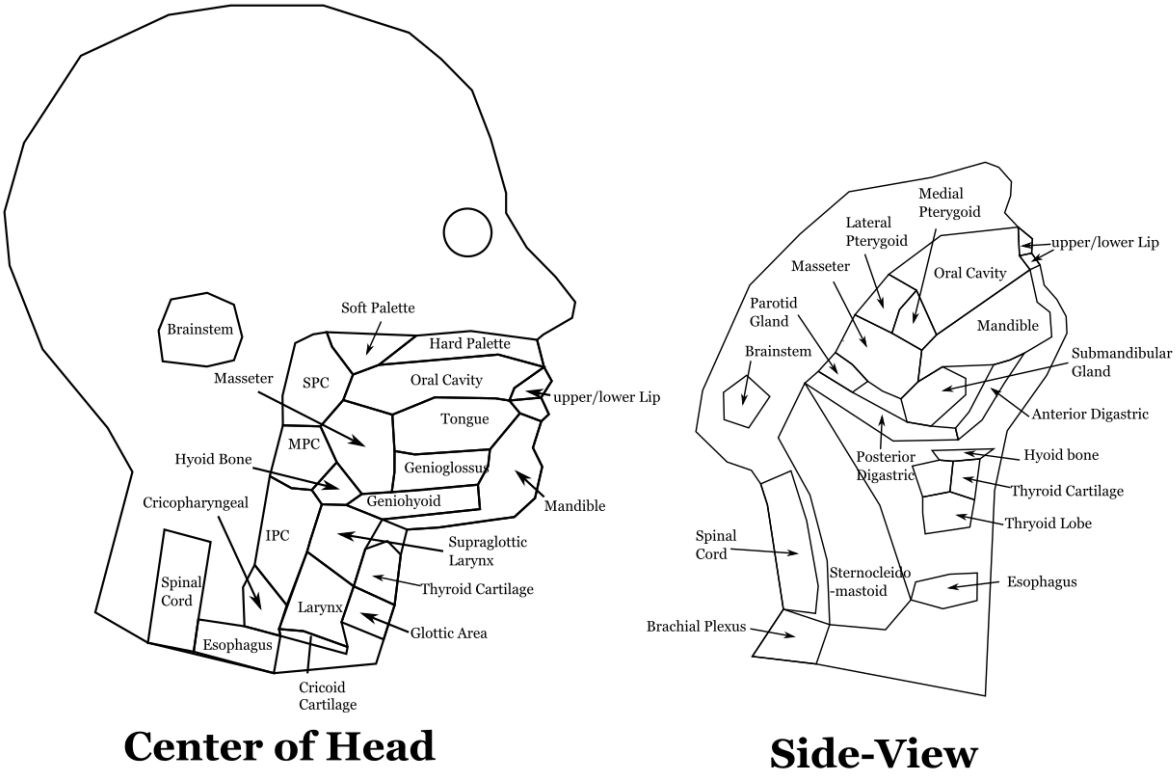


Figure 4.4: Development of the organ diagram used to encode organ-specific values. (Left) Organs that are located in the center of the head (Right) Bilateral organs that occur on both the left and right sides of the head.

as a plot in 3 dimensions [345]. However, we felt that this representation was limited in its usefulness, as it is difficult to identify organs that may be smaller and clustered together, but may be functionally important, such as salivary glands and smaller organs in the neck. Previous work has also shown that 2-dimensional maps of anatomical regions work well, and work well with clinicians who are typically trained to work with image slices and 2-dimensional anatomical drawings [343]. Expanding on this, we created a 2-dimensional representation of 45 organs used in our dataset based on existing anatomical drawings [94].

We then divided up the organs in the head into unilateral organs that sit along the mid-sagittal plane (e.g. tongue), and those that exist as a pair of organs on each side of the mid-sagittal plane (e.g. eyes), which are further subdivided into those on the same side as the primary tumor (ipsilateral side) and those on the opposite side of the primary tumor (contralateral side). This gives us three "groups" of organs along the center axis. For each region, we took tracings around organs of interest using multiple anatomical cross-sections.

We then overlaid all drawings, added in missing regions such as the spinal cord, and manually adjusted each contour to avoid overlap and regularize the size of each region. Adjustments were also made to ensure that regions were reasonably concave so that color gradients were visible. A diagram of the final drawing with all regions labeled is available in (Fig. 4.4).

4.6.2 Additive Effects Panel

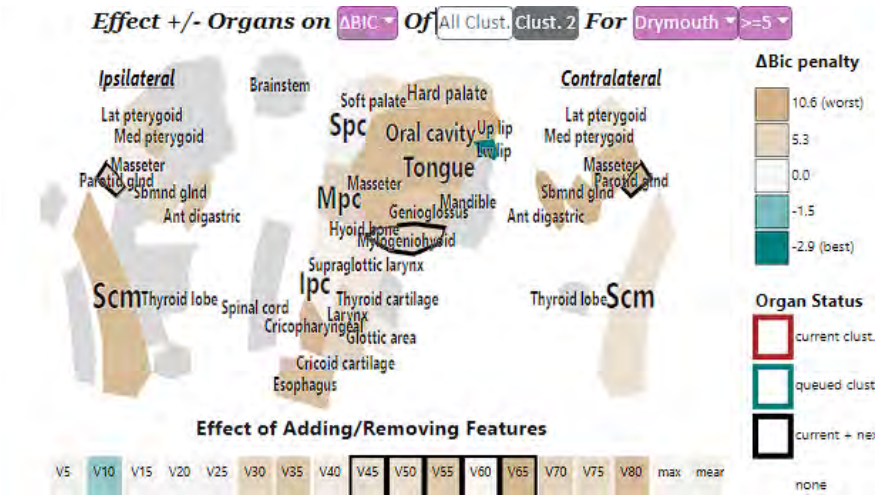


Figure 4.5: Additive Effects encoding showing a heat map of the organs and dose-features used in clustering. Color encodes the goodness of fit effect of adding (no or teal outline) or removing (dark black or brown outline) features to the clustering.

When working on model development (A1) our main task is to identify a set of organs to cluster once our desired outcome has been specified (Fig. 4.5). In this panel, we provide a forward search to estimate the effect of adding (for features not in the current clusters) or removing (for features in the current clusters) different organs or features from the clustering space on model performance (Section 4.5.4). We chose a beige-white-teal color scheme as we wanted to de-emphasize uninteresting (negative) results while still capturing the divergent nature of the results. Thus, we used beige as it has lower perceptual salience than the rest of DASS.

Since model developers may be interested in balancing performance between multiple outcomes, we allow choosing which information metric is used to encode color: BIC, AIC, or the t-statistic—-which we report as a change in p-value, as well as the inputs to the LRT test, and the threshold used to rank an outcome as “severe”.

Alternative designs relied on variations of heat-map and bar charts with effect sizes. However, these were replaced with the visual mapping approach, as we found that it helped to cue users about the approximate position and function of each organ when deciding on clinical relevance. Our collaborators also found that using similar layouts for the dose-cluster encoding and additive effects view reduced cognitive load and made the system more visually consistent.

4.6.3 Outcome Plot

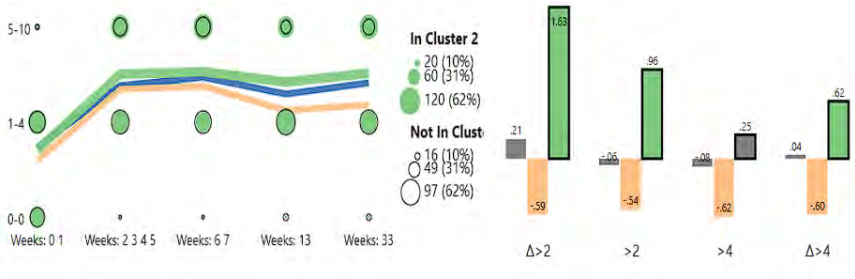


Figure 4.6: (Left) Plots showing the symptom ratings over time from the start of treatment for the specified symptom of interest, broken up by cluster. Circular markers encode the percentage of patients that experience a symptom at each level and time point, and help us estimate a patient's relative risk. Line charts show average ratings for each symptom. (Right) Bar chart showing the results of multivariate correlation tests for the clusters at different thresholds.

To support validation and iterative model improvement, it was important to show how outcomes vary within each cluster. This is important when ensuring, for example, that the cluster with the highest doses is actually capturing the high risk patients. To do this, we provide two types of encodings that show patient outcomes for each cluster: a temporal view of symptom ratings for the clusters, and a statistical bar-chart view showing the results of the likelihood ratio tests performed on each cluster for the outcome of interest.

Our temporal view uses a novel encoding (Fig. 4.6) to encode the trajectory of the symptom of interest across the entire treatment period for the patient clusters. This encoding has two components: a symbol grid, and a simple line chart. To reduce the complexity of the encoding, we first group the symptom ratings and treatment dates into bins (we selected five). In the symbol grid, we divide the patients into those in the selected cluster, and those not in the selected cluster (out of cluster). For each patient, we calculate the highest rating

for the symptom within the treatment dates before aggregating by cluster. We then calculate the percentage of patients from the selected cluster that fall in each rating + date bin. These percentages are encoded as circles on a grid, where the x-axis shows each date bin, and the y-axis encodes the symptom ratings. Size encodes the percentage of patients. Values for the in-cluster patients are shown as a saturated marker, while the out-of-cluster patients are shown as a black border marker. By comparing the markers, we can approximate the odds-ratio of a patient within the selected cluster having a symptom of a given severity at each time point.

In addition to the symbol grid, we overlay a line chart that shows the mean symptom value over time for each cluster. The line charts use cluster colors. A cluster chart can be clicked to select that cluster for more details.

The statistical bar chart view encodes the results of the LRT test (Section 4.5.4). This view is used for assessing how well a model performs while accounting for the specific outcomes and confounders. Cluster-outcome relationships that are statistically significant ($p < .05$) are shown using their categorical cluster color, while relationships that are not ($p > .05$) are shown in gray. The selected cluster for the interface is highlighted using a bold black border between the bars for that cluster.

4.6.4 Cluster Dose-Distribution Plots

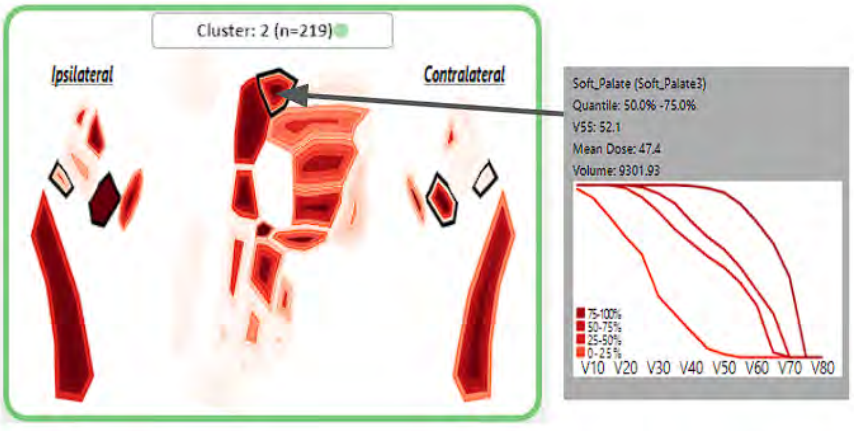


Figure 4.7: Per-organ dose distribution for a selected cluster. Color gradients shows within-cluster distributions. (Left) A tooltip shows the full dose-volume histogram for a brushed organ. Dotted area shows the value (V55) currently being shown in the heat map.

Once a reasonable set of cluster features has been identified, our first set of tasks involves investigating the dose distribution within each cluster (A2 T5-6). This is useful for identifying when the clusters are separating out patients with higher dose to other organs that were not included in the cluster inputs. To do this, we calculate the quartile ranges of a user-selected dose value within each cluster, for each organ. These values are then shown as a gradient heatmap using our 2-dimensional organ diagram using a sequential red color scheme (Section 4.6.1), where the innermost color represents the top quantile (80%) and the outer color represents the lower quantile (20%), allowing us to visualize the inter-organ dose distribution for each organ.

Interactions allow directly adding or removing organs from the cluster queue, as well as selecting a cluster to be used for brushing in other views. This facilitates the investigation of other aspects of the cluster in more detail.

To anchor the visual heatmap in the clinician's knowledge, we add a tooltip for each organ that can show the dose-volume histogram for each quantile for the selected organ and cluster (Fig. 4.7). This allows for a more detailed view of the entire histogram, while highlighting the relationship between the novel heatmap, and the standard dose-volume histogram that clinicians are familiar with.

4.6.5 Scatterplot

To visualize the distribution of patients across each cluster, we include a modified scatterplot panel that shows a 2-dimensional plot of the patients across two interactively-selected dimensions (Fig. 4.8). By default, we show the first two principal components of the features used to cluster the patients, but allow choosing to alternatively view higher order principal components, the principal components of the symptoms, or individual clinical or symptom ratings. Because we found that avoiding visual occlusion was more important than a high-fidelity projection, we use a force directed layout to remove overlap between glyphs.

Each patient in the scatterplot is encoded with a custom glyph that encodes its cluster membership, and the rating for the symptom of interest between 0 and 10. Each circular

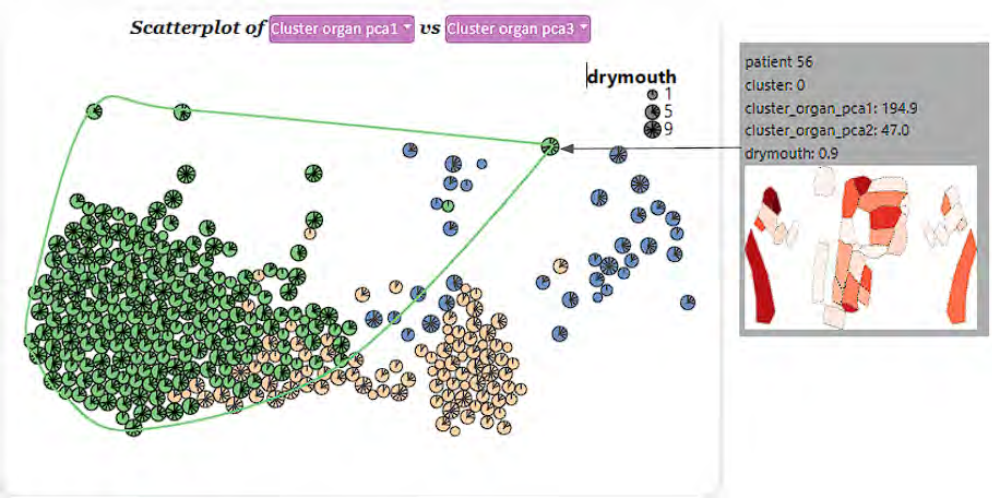


Figure 4.8: Stylized scatterplot. Patients are represented by a custom glyph that encodes the outcome of interest (late drymouth ratings) as marks extending radially. Markers are colored by cluster membership, and a contour is shown around the currently selected cluster. A tooltip (left) shows a heatmap of the dose applied.

glyph is encoded with ticks that extend in 32.7-degree intervals in a clockwise radial pattern, where the number of ticks corresponds to the symptom rating. Thus, a full "pinwheel" glyph represents a patient with a symptom rating of 10, while an empty circle represents a patient that does not experience the symptom. Because symptom ratings use discrete ordinal (integer) values, we can encode the exact ratings. We additionally scale the size of the glyph based on the symptom rating to support visual identification of small or high dose values.

Finally, we color code the glyphs based on their cluster membership. The selected cluster is brushed by giving the corresponding glyphs a higher opacity, and drawing a contour around the convex hull of the cluster in the scatterplot. By hovering the mouse over a patient glyph, the user can view a tooltip showing a plot of the given patient's received dose, and ratings for all symptoms over time. The dose to each organ is encoded for each patient using the organ diagram heat map (Section 4.6.4).

Previous designs used alternative projection methods with alternative projections and glyph encodings. However, we found that allowing inspection of individuals was more important than preserving location with perfect fidelity. In contrast, T-SNE avoided occlusion, but tended to produce visual clusters that did not correspond to the desired clusters. For

glyph design, we considered alternative shapes (e.g. diamonds or circles) for different levels of severity. However, collaborators found the use of color and shape confusing, while the use of ticks + size was better received, and we were able to identify the patient of most interest (very high and very low severity) fairly easily for further inspection.

4.6.6 Rule Builder

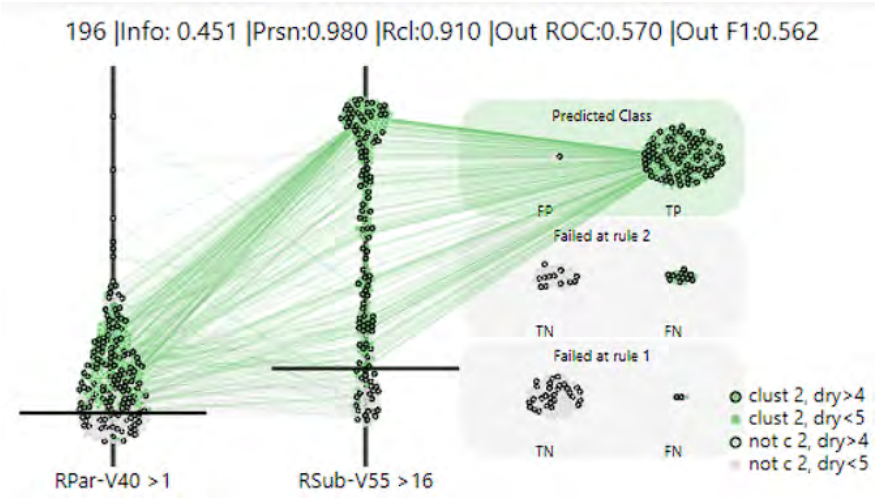


Figure 4.9: Ruleset encoding in the rule mining view. A swarm plot of the patients is shown for the feature used in each rule, with the first and most informative rule on the right. A horizontal line shows the cutoff thresholds used in the rule. Patients that pass a rule are then plotted in a swarm plot in the next rule on the right. The section on the right shows rule patients failed at, with patients that pass all rules at the top (green area). Patients in each section are divided to show the False Positives or False Negatives at each level. Lines connect markers for a patient across each sub-plot.

Once our clusters are built, one of our goals is to explain the clusters in terms that are familiar to clinicians. To accomplish this, we used a constrained rule mining algorithm (Section 4.5.4) to produce a set of dose thresholds such that the group of patients that meet these thresholds approximates the selected cluster. This approach was chosen as clinicians often work with dose thresholds when choosing treatment plans.

When a cluster of interest is selected, our algorithm finds a list of rule-sets that optimize the mutual information between the patients and the cluster of interest. We then generate a plot for each ruleset, and show the top rules in a list to the user. We also show the number of predicted positives, information gain, precision, recall, and f1 for predicting the true class above each plot.

Our novel rule encoding is based on a mixture of swarm plots and parallel coordinate plots that are modified to show the progressive filtering of each ruleset (Fig. 4.9). We encode each feature (e.g. V50 to the tongue) along the x-axis. We then map the y-axis to the dose value in grays. Patients are plotted along the y-axis based on their value for the given dose feature in the x-axis, and adjusted using a force-directed layout to avoid overlap. A horizontal line is then drawn at the threshold of the rule for the feature on each step of the x-axis. Patient marks are color-coded based on the selected cluster, while patients not in the selected cluster are gray.

To show the effect of additional rules, the features along the x-axis are ordered from left to right by the maximum information gain for its corresponding rule. In the first feature, we show all patients in the cohort. For additional features, we filter out all patients that do not satisfy rules from all previous features. The rightmost side of the encoding shows the patient groups stratified along the y-axis based on when they were filtered out of the ruleset. The set of patients that satisfy all rules is grouped at the top, while the set of patients that do not satisfy the first rule is grouped at the bottom. We further separate the final group by those in the true class (target cluster) and those not in the true class, allowing us to visualize the false positives and false negatives for each rule.

To provide a visual cue for how the rules are filtering the cohort along the x-axis, we provide lines that connect the undistorted locations of patients between axes, equivalent to a parallel coordinate plot with filtering. Once a patient is filtered out, we draw a line from the corresponding rule to the group on the right side. To prevent overlap, we only show the lines for the patients within one stratum at a time, which is changed by brushing a patient in the given strata. By default, we brush the group of patients that satisfy all rules (predicted positives).

4.7 Evaluation

The first and foremost value of DASS comes from its unique functionality and its ability to support clinical model development, which we illustrate via two case studies. These case studies, presented here in abbreviated form, illustrate the process of creating models for practical use, based on real clinical data. The case studies were performed via Zoom meetings with desktop sharing, with one of the data scientists piloting DASS and the group using the think-aloud methodology with note-taking. We furthermore collected and report qualitative feedback from clinical collaborators during these case studies.

As further evidence of the DASS functional value, we provide in the supplemental materials a quantitative evaluation of clusters generated with DASS against baseline ML clusters. The DASS clusters improve performance for drymouth, choking, and swallowing issues. Finally, with an eye towards the generalizability of DASS to other modeling problems, we collected additional feedback where eight data scientists rated the usefulness and usability of DASS.

Since the interactive model-building components are directly targeted at modelers, An additional quantitative comparison of our clusters against baseline ML clusters generated without DASS can be found in the supplemental materials.

Our dataset consists of 349 patients treated with radiation therapy for oropharyngeal cancer. These models have been generated with the help of DASS by four data scientists in our group over several months of remote collaboration. The models have shown improvements over baseline models, and have been favorably evaluated by three clinical oncologists.

4.7.1 Case Study 1

Our group was interested in identifying patients at high risk of developing drymouth at 6 months after treatment, a common side effect in HNC patients. In particular, the clinician analysts in the group wished to model the relationship between drymouth and the radiation dose applied to the salivary glands. The medical literature had established a few dose

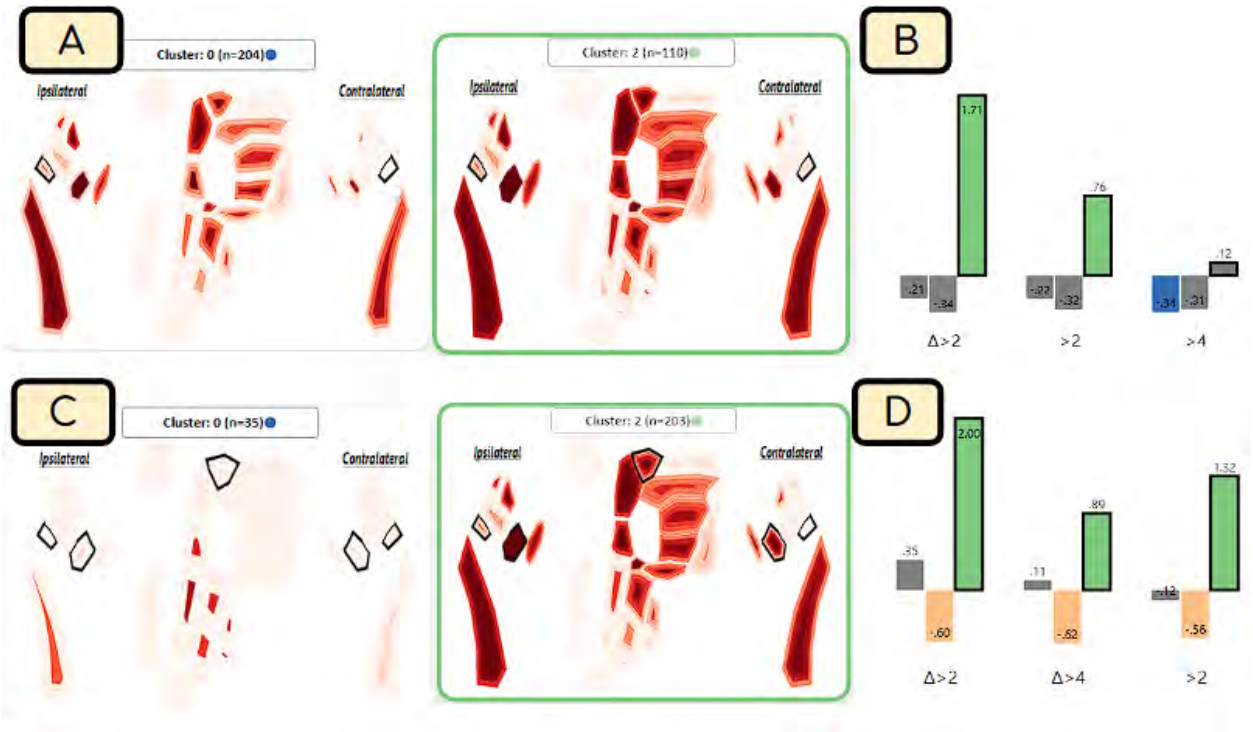


Figure 4.10: Case 1. (A) Low-and high-dose clusters using starting features. The low dose cluster includes several organs with high variance in the dose distribution. (B) Initial model performance. (C) Low- and high-dose clusters using the final model. Low dose cluster has a much lower variance, with only a few sets of outliers. (D) Final model performance measures. High-risk cluster is correlated with drymouth with a higher odds ratio than the initial clusters.

guidelines for parotid glands, but not for other salivary glands.

The model building process started by setting the parameters in the DASS control panel. Based on results from earlier work [345], the group set the initial clustering features to be V40-V55 doses to the ipsilateral and contralateral Parotid glands. Three clusters based on a Gaussian mixture model were investigated. Inspecting the initial clusters in the outcome plot, the analysts noticed that, as expected, there was a higher rate of drymouth in the highest dose cluster (Cluster 2 in Fig. 4.10), although the correlation was not significant for the desired threshold of > 5. Moving to the dose distribution plot, the group noted that the low and medium dose clusters tended to have a high-variation in the dose to certain organs, as indicated by the dark red inner contours and light outer contours to several organs (Fig. 4.10-A), suggesting that the model parameters did not differentiate the low dose patients well. Moving to the additive effects view, the model was iteratively adjusted to include the submandibular glands and soft palate, with a larger dose window (V30-V55).

After updating the model, the group noticed the clusters in the scatterplot panel achieved much better separation in the data (Fig. 4.10-D) compared to using just the parotid glands (Fig. 4.10-B). Returning to the dose cluster plots, the group also verified that the low dose cluster had a lower overall variance in the doses (Fig. 4.10-C).

Once the group achieved a set of features, the analysts aimed to verify the validity of the resulting model. Looking at the outcome panel, they noticed that while the high dose cluster was a strong predictor of drymouth, the low dose cluster had a high odds-ratio. Moving back to the scatterplot, and with the help of the oncologists, they inspected the patients in this low dose group, and noticed an interesting pattern: a number of patients had very high symptom ratings, and confirmed that none of their organs received notably high doses. Pivoting to the temporal outcome panel, the analysts further noted that this low-dose group had the highest incidence of severe drymouth at the start of treatment. After further discussion with the clinical collaborators, the group concluded that existing treatment plans try to minimize dose to the parotid glands, but not the submandibular glands, so the dose tends to be much lower in severe cases. The team theorized that there is likely a minor, but not full compensatory effect of the contralateral salivary glands when one set of salivary glands fails that should be explored later when investigating dose guidelines.

4.7.2 Case Study 2

This second case study dealt with the identification of patients at high risk of swallowing dysfunction, which is a less common outcome that is theorized to be related to damage to muscles in the mouth and throat. Swallowing disorders are also related to patients that require a feeding tube and weight loss, and thus it is an important outcome to avoid. High-risk patients can also be assigned prehabilitative therapy such as swallowing exercises as well.

To help identify a set of starting organs, the analysts inspected the rule mining view and set the desired outcome to be severe late swallowing using all available features (Fig. 4.11-A). By looking at the resulting rules, the group was able to identify the organs and dose features

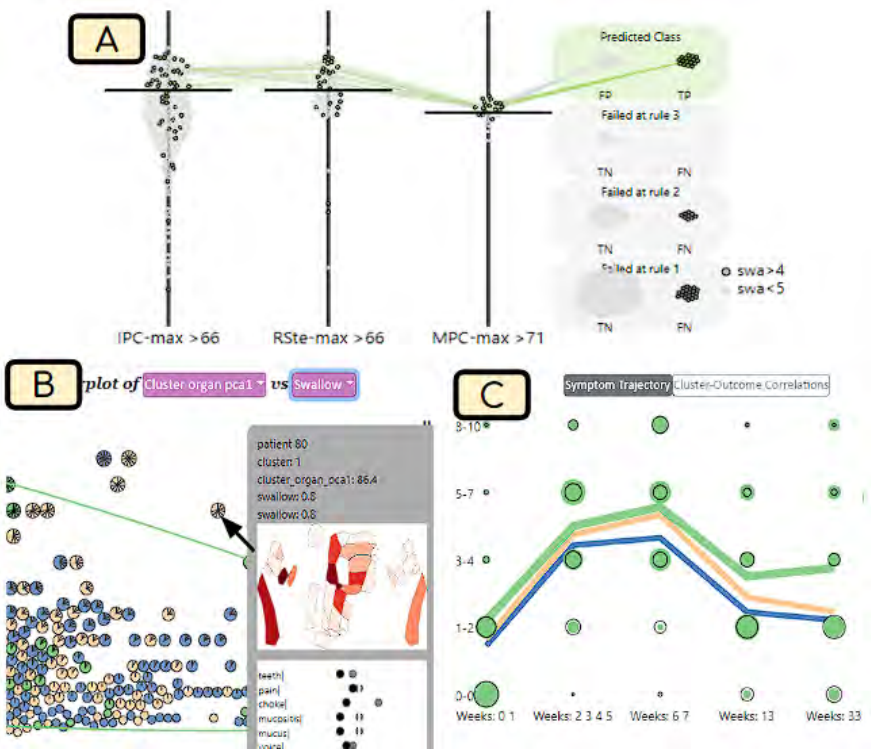


Figure 4.11: Case 2. (A) Rule mining results for predicting severe swallow dysfunction, which suggest using high doses to the pharyngeal constrictors. (B) Scatterplot of the first principal component of the cluster features vs swallow ratings. A tooltip highlights a case with severe swallowing in a low-dose cluster. (C) Outcome plot for the final clusters. High risk patients have similar ratings during treatment, but swallowing issues increase between 6 weeks and 6 months after treatment.

that best predicted severe swallowing, which allowed selecting a set of starting features for the cluster. Among the best splits were high dose depths (V55-V70) to the superior, medial, and inferior pharyngeal constrictors, which are key muscles used in swallowing, which were chosen as a starting point for the clusters. After running the clustering, the analysts inspected the outcome view and noticed that the initial clustering parameters were effectively separating the high-risk patients: this highest dose cluster had a significantly higher odds ratio of severe late swallowing (2.56) than other clusters (Fig. 4.11-C). Inspecting the cluster dose distribution view, it was noted that this high-dose cluster was noticeably smaller (n = 35) than the drymouth cluster and that the high-dose cluster tended to consistently have a much higher V55 to the IPC than other clusters.

Moving to the scatterplot, the analysts changed the dimensions to show the first principal component of the dose and swallowing ratings, which allowed identifying all patients with

high swallow dysfunction that were not in the cluster (Fig. 4.11-B). Using the tooltip, the group found some of these patients had high doses to the base of the tongue and upper larynx. The analysts then added the supraglottic larynx to the clustering parameters in hopes of capturing this group. The group then moved to the additive effects view, iteratively changed the dose window to include only the V55-V65, and added the esophagus, which is another major muscle used for swallowing in the base of the throat. After finalizing the parameter set, the analysts inspected the rule view to find the features that best distinguished the high-risk cluster. This high-risk cluster was easily distinguished using the V55 to the Inferior Pharyngeal Constrictor. Our clinical collaborators noted that all the pharyngeal constrictor muscles are located close together, and there exist guidelines for the dose for all of these muscles. Thus, a high IPC dose is likely a predictor of a high dose to all related organs. Additionally, the group discussed the fact that the dose threshold for swallowing was higher than drymouth, which may indicate that muscles are less sensitive to radiation relative to salivary glands.

4.7.3 General Usefulness and Usability Feedback

In addition to the case studies, which illustrate the DASS unique functionality, we collected qualitative and quantitative feedback from both collaborators and from modelers not affiliated with the project. All collaborators appreciated the functionality provided by DASS, and are in the process of publishing the resulting clinical models. Regarding the spatial cluster panel, our clinical collaborators found it intuitive and useful for inspecting dose distributions of organs of interest. Feedback on the rule mining algorithm was also positive, with oncologists remarking that it was "very useful", as it could "translate our results into practical applications". A data mining expert responded similarly to the additive effects panel, saying that it was a "nice, very nice way to explore the parameter space".

Additionally, we asked, via an anonymous online questionnaire, three senior data scientists in the group, who were not directly involved in the DASS design but participated in walkthroughs of the system, and five junior data scientists, who were not affiliated with the

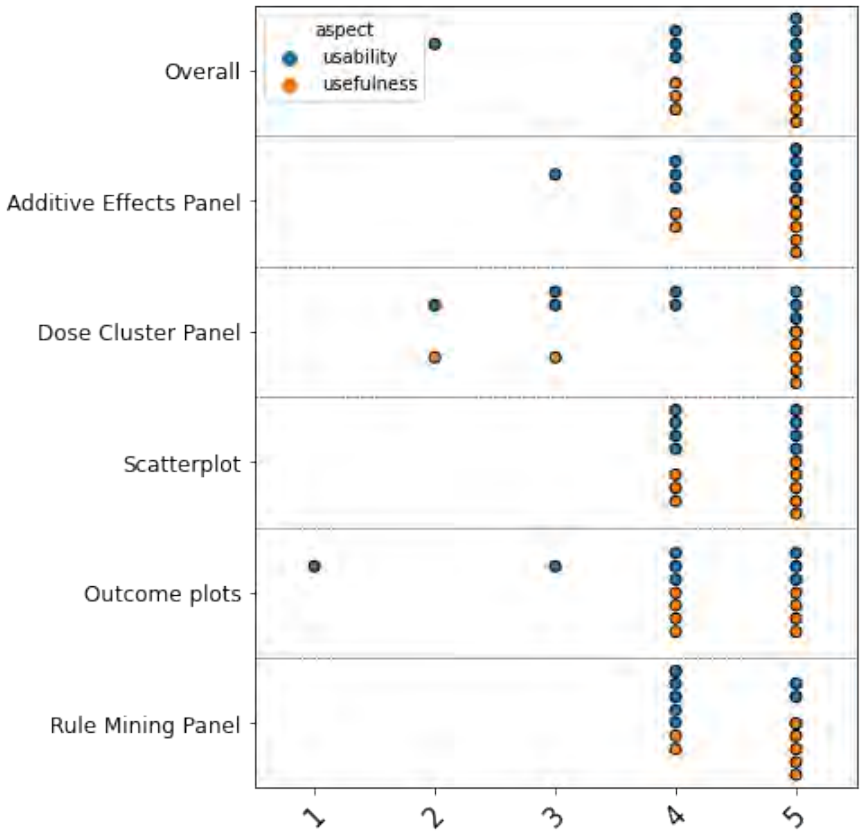


Figure 4.12: General DASS usability and usefulness.

project to rate the usefulness and usability of the whole system and of each component of the system on a Likert scale from 1 to 5. We specifically sought feedback from data scientists, with an eye towards generalizability, as modelers are the intended users of the interactive model-building components of the system. Results are shown in Fig. 4.12.

Feedback was very positive, with most ratings between 4-5, in particular for usefulness. Ratings for usability were slightly lower, as expected: some of the group experts clarified the group's narration during the model-building process was extremely useful, and they wished for visual help buttons replicating that experience on demand. Ratings from the junior data scientists not affiliated with the project were occasionally lower, in particular for the composite outcome marker plots and the dose cluster panel. Based on the qualitative feedback, the difference in these cases was directly related to the visual scaffolding and domain expertise which collaborators benefited from, as these plots were based on methods

used in RT planning and clinical biostatistics.

4.8 Discussion and Conclusion

Our design relies on three main principles for improving model development: 1) information-scent to guide model development (A1); 2) visual scaffolding to support bridging the information gap between what domain experts commonly deal with and what is needed to reason about the data (A2); 3) model explanations aimed at translating our novel approach to the types of simpler "models" use in practice (A3). Our case studies show how the system was effectively used to develop explainable models that outperformed our previous attempts at developing clinical models.

In terms of generalization, our design philosophy is most suited for applications where the training data is insufficient for automated inference tasks, but can be augmented through collaboration between domain experts and visual analysis experts during the model-building process, which requires continuous input from both parties during model-building to identify and reason about unexpected results. Additionally, the system is best suited for smaller, complex datasets. While some models can generalize knowledge by simply collecting more data, this is not feasible when doing small cohort analysis when data is rare and expensive to collect. Thus, these problems benefit the most from domain-expert input to help embed domain knowledge into the algorithm that can not be inferred from the data.

Below, we distill the design lessons gathered from this project when dealing with visual steering and explainable AI problems in collaboration with domain experts.

L1. *Explanation Scaffolding:* We extend the concept of visual scaffolding – gradually building to more complex visualizations from a more familiar one – to that of XAI-style model explanations. Specifically, we argue that model explanations should aim to translate more complex models into those that mimic how users commonly deal with the data. In our case, we used constrained rule mining in conjunction with visualizing intra-cluster dose distributions using a visual scaffolding approach. Other systems have used regression models which

are common in biostatistics. However, clinicians do not often reason about such models directly, so they are less useful in clinical practice.

L2. *Keep Model Goals Flexible:* When developing models, data scientists may work solely to optimize the performance in terms of easily measured outcomes [196], which leads to issues during collaboration with model end-users [370]. In practice, there is often a misalignment between what can easily be measured, and what makes a model useful in practice. In developing our models, we found that it was important to allow users to investigate a mixture of outcomes, in addition to qualitative factors such as model plausibility and complexity, which need to be leveraged against each other when deciding on the final model.

L3. *Encourage Skepticism:* One motivation in the design of our system was a recurring problem of designing models that performed well, whereas further probing revealed internal logic that appeared to be the result of biases and spurious correlations in the data. Despite this, our models were often received without skepticism when these issues were not brought up. This issue with over-trusting erroneous explanations has been suggested in early empirical studies [148, 360]. The communication gap between model builders and experts may result in dramatically over-trusting the models for both parties as they may be unable to identify issues in the models on their own. When dealing with XAI, designers should focus on promoting skepticism about the models by highlighting potential issues in the models, such as outliers and confounders, which can help highlight previously unknown issues in the models.

The main limitation of our system is the reliance on visualizations that require familiarity as well as knowledge of the underlying models and data, which is made possible by the long-term nature of our collaboration. While we use domain-specific designs for our visual scaffolding approach and model designs, the design philosophy can be generalized to other problems involving spatial data where model outputs can be translated into discrete groups, such as clustering and decision trees. In terms of scalability, our system requires 5-15 seconds to update new results for each cluster, depending on the number of clusters and rules mining

settings. Scaling to larger datasets may increase the required time, although this is still significantly faster than alternatives that do not use interactive steering. Visualization of individual patients in the Scatterplot and Rule view may also be difficult with very large cohorts.

In conclusion, we have presented an ML and visual steering system for clinical oncology symptom modeling with spatial data. We described the co-design of a clinical visual-steering system, and demonstrated its ability to support the creation of interpretable ML models for stratifying patients. Additionally, we presented a set of lessons learned for model co-development and model explanations for a hybrid, machine expert and human expert problem. We hope that these findings will help future designers create better, and more trustworthy models in high-stakes settings.

Acknowledgements Our work is supported by NIH awards NCI-R01-CA258827 and NLM-R01-LM012527, and NSF awards CDSE-1854815 and CNS-1828265.

4.9 Chapter Conclusion

DASS describes the design of a human-in-the-loop spatial clustering algorithm for 3-dimensional dose distributions and modeling interface for patient symptom outcomes. It presents several strategies for improving spatial clustering by using simplified model explanations and 2-dimensional representations of 3-dimensional dosimetric data that preserve relevant topological information for domain users. This paper also proposes the idea of model actionability in XAI systems for domain-specific applications and attempts to instill strategies for directly designing systems for incorporating these user goals in the design of the models themselves by providing spatially aware information scent into the integrated visual analytics. In our next paper, I will generalize design lessons from previous work, as well as work in writing clinical papers into a formalization of the design process of explainable spatial unsupervised modeling for clinical application.

Chapter 5

Explainable Spatial Clustering in Radiation Oncology for Domain Experts

Chapters 2 and 4 discussed different applications for spatial unsupervised machine learning in radiation therapy planning. This chapter extends these works and incorporates additional insights from the dissemination of those results through the publication of clinical results-focused papers to present a domain characterization of explainable unsupervised learning with spatial data.

Advances in data collection in radiation therapy have led to an abundance of opportunities for applying data mining and machine learning techniques to promote new data-driven insights. In light of these advances, supporting collaboration between machine learning experts and clinicians is important for facilitating better development and adoption of these models. Although many medical use cases rely on spatial data, where understanding and visualizing the underlying structure of the data is important, little is known about the interpretability of spatial clustering results by clinical audiences. In this work, we reflect on the design of visualizations for explaining novel approaches to clustering complex anatomical data from head and neck cancer patients. These visualizations were developed, through participatory design, for clinical audiences during a multi-year collaboration with radiation oncologists and statisticians. We distill this collaboration into a set of lessons learned for creating visual and explainable spatial clustering for clinical users.

The clinical studies this chapter is based on were published in Radiotherapy and Oncology [346, 349], while our discussion of the design lessons was published as a short paper and presented at the 2020 IEEE Vis Conference [343]. The initial designs of the lymph-node

encodings and dendrograms referenced in this chapter were created by Tim Luciani as part of his PhD dissertation [186].

5.1 Introduction

One of the most important applications of machine learning (ML) techniques to oncological healthcare is patient stratification. Stratification is the division of a patient population (group) into subgroups, or "strata". Each strata represents a particular section of that patient population. The strata are typically correlated with specific demographic or disease traits, and specific outcomes, including survival or side effects in response to specific treatments. The nature of patient stratification makes it well suited for clustering—an unsupervised data mining technique that groups patients based on some measure of distance between them. When the distance measure and clustering algorithm is well-chosen, clustering can generate novel insights and help discover previously undiscovered structure in the data.

Oncological data is often tied to a patient's anatomy, which complicates the construction of a similarity measure between patients and the selection of a clustering algorithm. In cancer patients, the spatial information of the tumor and surrounding anatomy is vital in deciding optimal treatment and forecasting patient endpoints. Thus, understanding the underlying spatial structure of the data during the clustering process is important. Despite a widespread interest in sophisticated clustering techniques for patient stratification, the adoption of clustering in oncology is stifled by the difficulty in understanding the inner workings of spatially-informed clustering.

In this work, we examine a participatory design of explanatory visual encodings born out of a long-term collaboration between oncology, data mining, and data visualization practitioners performing analysis on a cohort of head and neck cancer patients [201, 285]. Specifically, this work looks at interpreting clusters of stratified head and neck cancer patients based on secondary disease spread to the lymph nodes, with the goal of helping clinical users understand the strata and use them to help predict the toxicity outcome of disease treatment.

We reflect on the process of creating domain-specific visual encodings through participatory design to help "bridge the gap" between the data experts and healthcare experts [136]. We further explore obstacles and successes when creating visual encodings for interpreting data mining techniques, and for communicating with oncology experts with limited background in both visualization and in artificial intelligence.

5.2 Related Work

5.2.1 Explainable AI

Advances in machine learning and data mining has led to a recent increase in papers about interpreting artificial intelligence systems, largely grouped under the umbrella of 'Explainable AI' (XAI). XAI encompasses a wide range of concepts, and there is still no widely adopted vocabulary for the techniques and evaluation methods. Many systems visualization solutions have proposed model-agnostic solutions that rely only on inspecting the relationship between the data and predictions of a model [157, 180, 352, 371]. However, these are usually more applicable to model developers, and there is little adoption of these methods for end users of the models. Model-specific methods involve model introspection and include a variety of more traditional methods. Coefficients in linear models can be mapped to effect sizes for different features [235], making them popular in controlled experiment analysis. Other popular methods have been developed for visualizing decision trees [130] and bayes nets [168]. Model-specific methods allow for more robust statistical interpretations and causal analysis, which is an important factor for facilitating trust in lay-users. When models are too complicated to be understood via introspection, mimic models are often used where an interpretable model is made to approximate a black-box model. For applications where the goal is to interpret a given prediction, rather than the whole mode, instance based approaches where an explanation is built around a local subset of the data has been shown to provide better fidelity than building mimic models with the entire dataset [268, 269].

Cluster Explainability Interpretation and visualization of clusters is a common analysis task tightly integrated with dimensionality reduction in general, but is less understood

than traditional explainable AI (XAI) approaches, which are generally focused on supervised learning. A task analysis of 10 data analysts [31] included 3 tasks related to clusters: verifying clusters, naming clusters, and matching clusters to existing classes. General methods of cluster visualizing have typically been linked to low-dimensionality embedding, where classes are shown plotted in a 2 or 3-dimensional space, and cluster-membership is shown on top of the data in the lower-dimension space [9, 84, 339]. Hierarchical clustering methods, where clusters are iteratively created at different levels of granularity, have commonly been visualized as dendrograms. When dimensionality reduction isn't appropriate, general methods of multivariate data visualization are used, such as parallel coordinate plots [58] or specialized glyph encodings [42]. Other systems synthesize existing methods to support visual steering and clustering for scientists [43, 44, 215]. While some recent work has dealt with clustering ensemble geospatial data [189], we are not aware of any methods that deal explicitly with clustering anatomical or 3-d data as in this work.

5.2.2 AI in Healthcare

Most data-driven systems that are actively used in practice by clinicians emphasize simplicity and intuitiveness over performance. Widely used scoring systems have relied purely on clinical intuition [103], despite its sup-optimal performance in followup studies [149]. More successful medical scoring systems rely on regression methods with rounded coefficients [15, 224] despite the negative impact on performance [313]. Other commonly employed methods include recursive partition analysis (RPA) [19, 169, 176], or variations of logistic regression [63]. Many works have demonstrated improvement in clinical models by using more advanced methods, such as boosted trees and support vector machines in disease prediction [60, 265, 292], and deep learning models for predicting medical events [54, 264]. Despite the superior performance of many of these advanced models, they are difficult to deploy in practice and have poor adoption due to their lack of interpretability.

Vis in Healthcare Visualization approaches to healthcare problems often focus on supporting data exploration, rather than understanding predictive models [23, 34, 181]. Certain

systems for model exploration have been developed to aid in the development of regression models based on the workflows of biostatisticians [76, 274]. Other systems have applied visualization for clustering cancer data [215], and predicting infection spread in hospital wards [226]. For spatial data, Grossmann et al. [113] incorporated methods for visualizing clusters based on bladder shape to support a retrospective study on prostate cancer patients. Some works have attempted to identify design considerations when working with domain experts in healthcare [192, 263]. However, except for Raidou et al. [263], most of these considerations do not apply to clustering or spatial data, and are largely focused on analytics and electronic health record data. As a result, there is a dearth of papers discussing how to approach unsupervised XAI models to reach clinical audiences.

5.3 Background

In many cancer patients, tumors metastasize into the lymphatic system, causing lymph nodes to become "involved"—affected by secondary nodal tumors. The lymphatic system forms a complex chain of lymph nodes, and these secondary tumors spread along these chains to adjacent regions stochastically. Affected lymph nodes are a long-established factor in determining patient outcomes in head and neck cancer [174]. Current predictive systems use a staging system based on the size and number of nodal tumors, but miss more nuanced predictions about how the different patterns of nodal spread may affect toxicity outcomes [128,357]. No prior machine learning methods correctly handle this type of spatial data, due to a lack of spatial similarity measures [87, 186].

Our data comes from a cohort of 582 head and neck cancer patients collected retrospectively from the MD Anderson Cancer Center. All patients survived for at least 6 months after treatment. Data was collected on the presence of 2 severe side effects: feeding tube dependency, and aspiration - fluid in the lungs that requires removal. We mainly consider the presence of either of these side effects, which we define as radiation-associated dysphagia (RAD) [61]. The data also encodes the disease spread to 9 connected regions (denoted as

levels 1A-6) on each side of the head, along with the disconnected retropharyngeal lymph node (RP or RPN). Many patients in this cohort had unique patterns of disease spread to the lymph nodes.

The project consisted of 2 phases with distinct design requirements. In phase 1 (model development), we worked alongside six domain experts in radiation oncology, and two data analysts with data mining and biostatistics backgrounds, over four years. During this time we developed, validated, and deployed an anatomically-informed patient stratification method based on each patient's patterns of diseased lymph nodes [186]. To demonstrate the important role of spatiality, the stratification used only anatomical features. We met with representatives from this group up to three times per week via teleconferencing, as well as in quarterly face to face meetings. In phase 2 (model dissemination), our results needed to be analyzed and delivered to the larger radiation oncology community. In this stage, we received feedback from three additional radiation oncologists and two bioinformaticians with expertise in head and neck cancer. The final stratification approach is available to clinicians through an open-source interface [201]. Below, we reflect on the design process, which focused on an activity-centered design paradigm [199], along with feedback from the domain experts.

5.4 Model Development Phase

In phase 1, we worked to identify a meaningful, anatomically-informed distance measure between patients, as well as an appropriate method of clustering the patients. We developed an approach in which each side of the head was treated as a graph. Nodes in this graph corresponded with regions in the head that aligned with those used in existing oncology literature, and regions that were anatomically adjacent in the head were connected in the graph as an edge. Each patient was treated as two sub-graphs, one for each side of the head, containing only the nodes with nodal tumors. A distance measure based on these graphs then needed to be identified, alongside a clustering technique that led to meaningful clusters

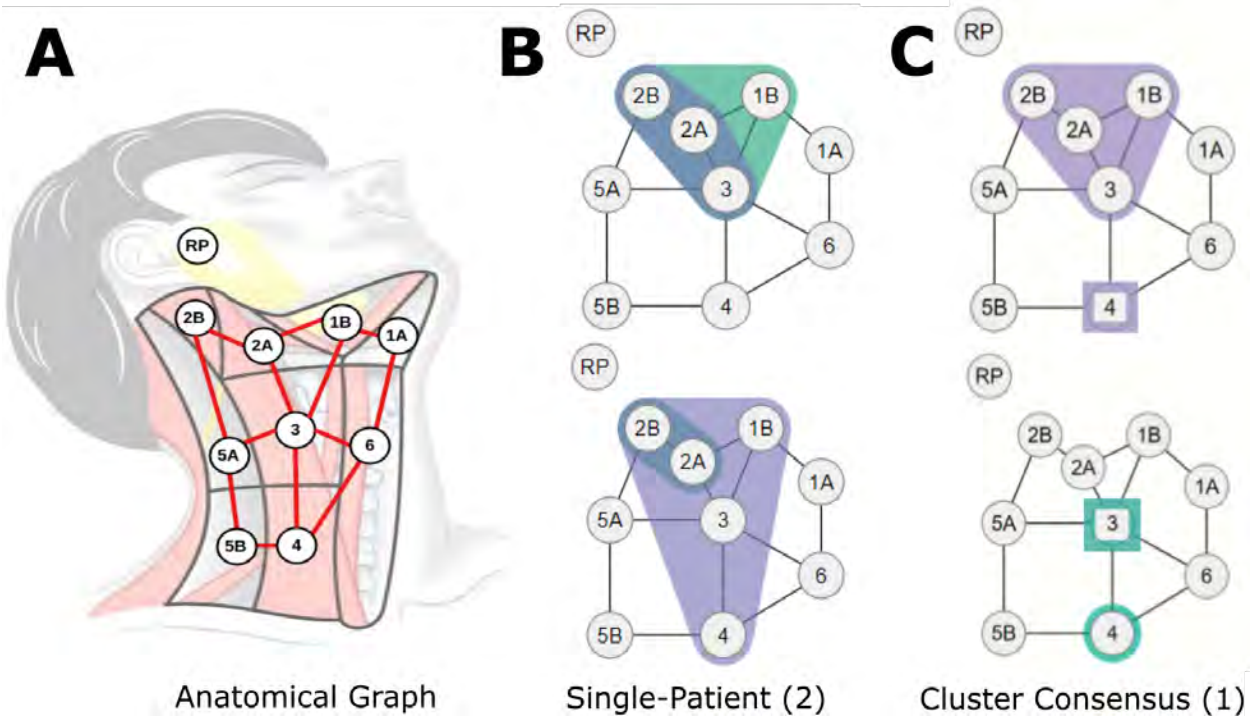


Figure 5.1: (A) Lymph nodes overlaid over a diagram of the neck. (B) Example graphs of diseased nodes for 2 individual patients (datapoint representation). (C) Example consensus graph for 1 cluster (cluster representation). The top-right graph shows disease spread with 66+% of patients on the right nodes in 1B, 2A, 2B, and 3, and disease in 1-33% of patients in right node 4. The bottom-right graph similarly indicates involvement of >66% and <33% of patients in left nodes 4 and 3, respectively.

(activity 2). Clustering was performed using only the spatial disease spread captured by the graph model. Because identifying relevant structures in oncological data is nontrivial, defining this methodology required iterative experimentation with different features, clustering techniques, numbers of clusters, and other parameters [308]. We identified the following activities that required visual support:

1. Identify and analyze the relevant spatial data features underlying one datapoint (i.e. patient).

2. Analyze the effects of different spatial similarity measures on clustering (i.e. why two patients are considered to be similar under a specific measure).

3. Analyze the representative patterns and pattern variation within each cluster.

Datapoint Representation The first design followed a graph metaphor to encode the diseased regions for each patient (activity 1). A compact graph that followed an anatomical map of lymph node chains for half the head (because the problem is symmetric) was used as a template for each patient (Fig. 5.1-A), based on ideas from biological network visualization [203,341]. For each patient, two envelopes were drawn over their diseased nodes. Green and purple envelopes were used for the left and right side of the head, respectively. Areas where envelopes overlap are shown in blue and denote regions where tumors occur on both sides of the head, which are of particular interest to oncologists (Fig. 5.1-B).

This design allowed for a compact representation of a complex spatial feature space, while following the mathematical intuition behind different distance measures. These graphs were incorporated into an interface that shows patients and compares them to their most similar matches. The compact representation was useful in identifying the spatial features of each datapoint, as well as interpreting distance between patients.

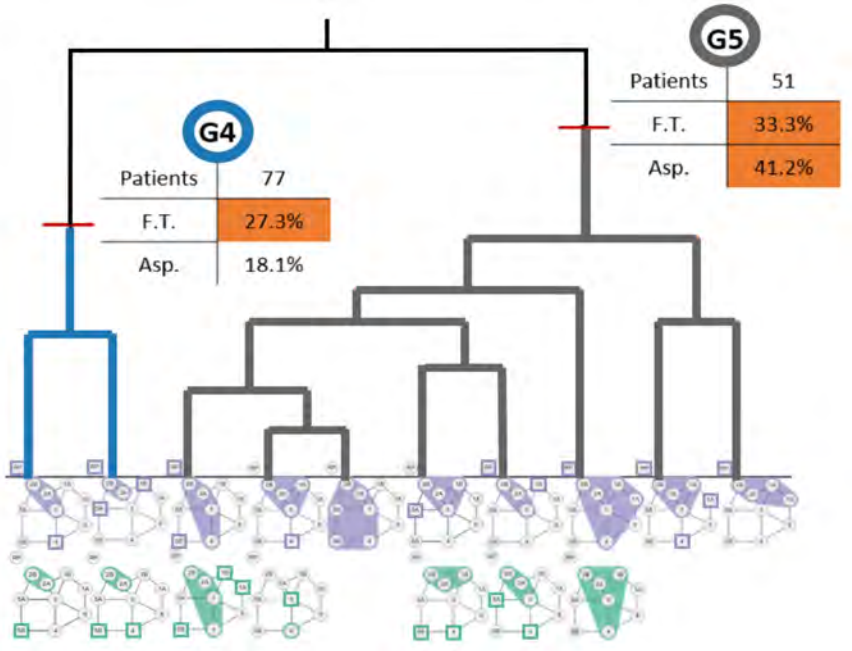


Figure 5.2: Part of an augmented dendrogram of lymph node clusters (clusters 1-3 not shown; the full dendrogram is available at **http://www.sciencedirect.com/science/article/pii/S2590177X20300019#f0050**. Leaves of the tree are smaller clusters that merge at higher levels according to the agglomerative clustering algorithm. Clusters are id-ed by colors in the graph. Clusters are further augmented with breakdowns of relevant clinical covariates of interest (F.T.: Feeding Tube; Asp.: Aspiration).

Cluster Representation In a first attempt to characterize each cluster, we selected a representative patient for each cluster: i.e., the patient closest to the cluster centroid (activity 3). The representative patient, however, did not capture any intra-cluster variability. Subsequently, we created a new representative encoding by placing the most commonly affected nodes for a cluster in a "consensus" graph. Nodes where $\frac{2}{3}$ of the patients in that cluster had nodal tumors were outlined in envelopes. However, in this new representation, it was unclear why certain clusters were not merged. In a third iteration, we added a different marker (squares) for nodes where less than $\frac{2}{3}$ of the patients in that cluster, but at least one patient had nodal tumors (Fig. 5.1-C). We used shape, rather than color, because hue already encoded disease laterality, and further intensity variation was not legible given the small scale.

However, at small scale, the markers and colors for multiple clusters became hard to distinguish. Additionally, outside clinicians and bioinformaticians mis-interpreted the third encoding as representing only one patient in that cluster, and in one case, as clusters containing identical patients. In the fourth design, two stacked graphs were used for each side of the head for each cluster, and visual scaffolding [198] was used to explain the progression from a single datapoint representation to the consensus graph. The consensus graphs were placed within dendrograms, which showed the consensus graphs of smaller component clusters within each larger cluster of interest (Fig. 5.2). To further clarify the hierarchical clustering process, we added explicit color-coding of the dendrograms, with labels and colors showing the cluster names and tracing the merging process, as well as small statistics tables showing the patient toxicity outcomes within each larger cluster.

5.5 Clinical Model Dissemination Phase

In the second phase, our results needed to be able to reach their intended audience: clinical radiation oncologists. While the methodological development was concerned with the clinical validity of the analysis, clinical readers are more concerned with significance of the

results, and place more importance on feasibility, trust in the underlying covariates, and the implications of the results [345, 346], rather than the methodology used, which had already been peer-reviewed [186]. In this phase, we used four clusters to align with existing staging systems, and the clustering still only considered spatial disease spread. In order to effectively communicate results, we identified the following activities to support:

1. Describe patient clusters from an anatomical perspective.
2. Identify each cluster's underlying structure.
3. Connect structural cluster differences to clinical covariates.
4. Explain plausible causal relationships between the clusters and correlated patient outcomes.

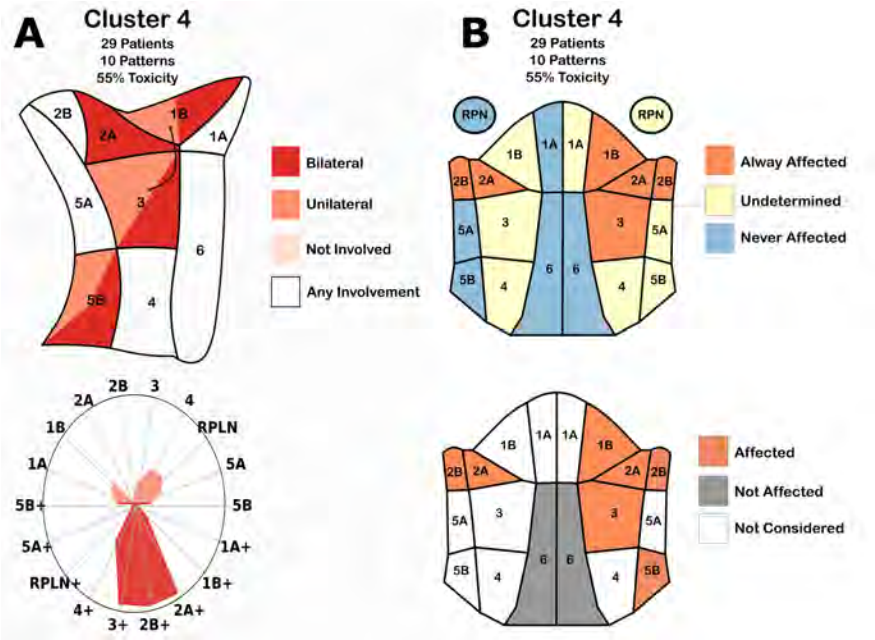


Figure 5.3: (A) Cluster conditionals. (Top-left) Map of the regions in the neck. Color indicates when the decision tree classified a patient into the cluster based on if the region had no disease (pale red), tumors in one side of the head (red), both sides of the head (dark red), or a combination of two options. (Bottom-left) Radar chart showing the percentage of patients in the cluster with nodal tumors in a given region. Color indicates the presence of tumors in exactly one (pale red) or two (dark red) sides of the head. (B) Second iteration of cluster conditionals. (Top-right) Membership diagram showing the regions in the head. Color indicates when all (red), a subset of (yellow), or none of (blue) the patients in a cluster had nodal tumors in a region. (Bottom-right) Decision-tree based diagram. Colors indicate when a decision tree classified a patient into that cluster.

Cluster Conditionals The first design relied on two synergistic encodings for each cluster. The first encoding expanded on the original anatomical diagram to show the most discriminative features in each cluster (conditionals). To do this, a decision tree was trained on the cohort to predict cluster membership with 100% accuracy using the number of sides of the head with a nodal tumor in each region of the head and neck, which could be 0 (no disease), 1 (unilateral disease), or 2 (bilateral disease). Because experts who had not participated in the methodology design process had trouble understanding the graph-based encoding, the set of variables considered sufficient to any patient in the training data into a given cluster was then encoded into an anatomical region diagram of one side of the neck (Fig. 5.3-A). By focusing on the regions that the decision tree considered, the diagram highlighted the regions that best identified the key differences between clusters, while omitting regions with commonalities between then, in order to support activities 2 and 3. The second encoding was a radar plot of the percentage of people in a cluster with either unilateral or bilateral disease spread in a given region of the neck. This representation allowed for a more detailed view of the overall distribution of tumors in each cluster (activity 1).

The initial cluster visualization design using trees was found to intuitively make sense to clinical collaborators. However, they had difficulty understanding the underlying explanation of the diagrams and how they were generated within the space of a figure caption, as they had limited experience with decision trees. Collaborators incorrectly assumed that all combinations of nodal disease in the diagrams were shared between all patients in a given cluster. Additionally, our collaborators pointed out that while the one-sided diagram of the neck was common for surgical applications, radiation oncologists often visualized the neck in terms of a front view that included both sides of the head simultaneously.

In the second design (Fig. 5.3-B), each cluster is encoded using a frontal view anatomical diagram. A red-yellow-blue categorical color scheme was used to mark which regions were diseased in all patients, some patients, or no patients within the cluster, respectively, following the original intuition of our collaborators. An additional anatomical diagram based

on the decision tree was included for each cluster below the membership diagrams. Since the new diagram included both sides of the head, color was used to show when the decision tree split the cluster based on the presence of disease (red), or absence of disease (gray) in a given region, while white regions were not considered in the model.

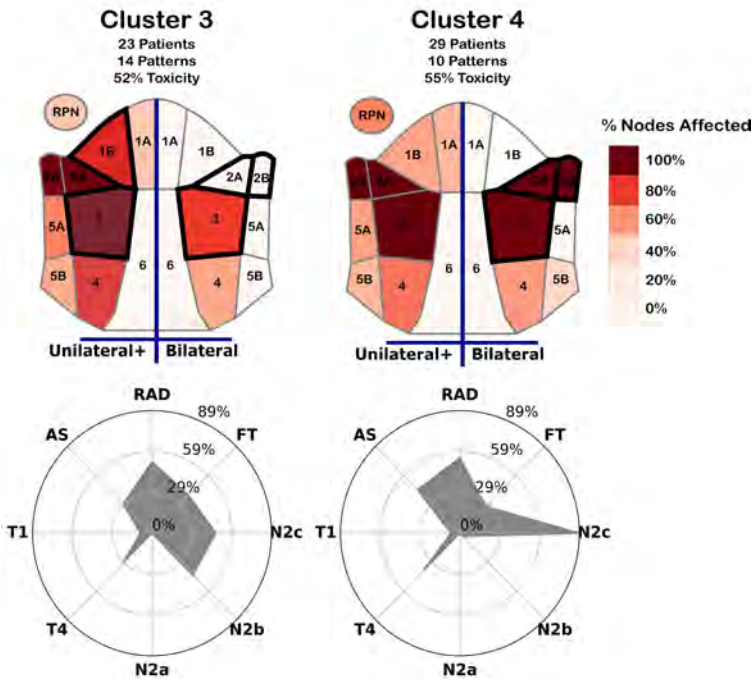


Figure 5.4: Designs for two high-risk cluster conditionals using heatmaps. (Top) Spatial heatmaps showing the portion of patients with nodal tumors in each region for at least one (left) or both (right) sides of the head. Regions most informative in determining cluster membership are outlined in a thick dark border. (Bottom) Radar charts showing the percentage of patients within the cluster with a given toxicity outcome (FT/RAD/AS), and those within an existing risk-staging group (T1/T4/N2a/N2b/N2c).

Cluster Membership The conditional designs were better-received by the clinicians, but difficulties in understanding the colormap and the lack of detail in the cluster membership made it challenging to correctly draw insights. To address these concerns, we designed a new heatmap diagram of the neck (Fig. 5.4), which used a sequential white-red color scheme to encode the number of patients in a cluster with disease in a given region (activity 1). We note that head and neck oncologists account for symmetry when discussing similar patients, and thus a symmetric encoding was a desired feature. A simplified decision tree was trained to identify the regions that contained the most information about cluster membership, which were outlined with a dark border in the heatmaps (activity 2). Additional labels

were included, to indicate the left/right sides of the diagram show unilateral vs. bilateral involvement, rather than the literal left/right sides of the head.

To help indicate the relationship between the clusters and other clinical data, covariates and outcomes that were the most interesting to clinicians were included in a radar chart alongside the heatmaps for each cluster. The inclusion of these data helped with the collaborators' ability to discuss potential relationships between the structure of the clusters and correlated outcomes (activities 3 and 4).

5.6 Design Lessons

Through the course of these iterations, we have distilled design lessons for interpretable clustering with spatial data.

L1. *Use visual scaffolding based on users' spatial background.* Spatial representations were, as expected, essential to understanding the clustering. Furthermore, encodings were better received when they mapped directly to the users' model of the problem, particularly when the users did not participate in the design. Using a graph-based encoding for the patient lymph node chains allowed us to draw parallels to graph theory, which was useful when testing similarity measures that were based on graph matching methods. In contrast, when designing for the wider oncology community, the encoding best received was created by visually scaffolding the graph directly onto an anatomical diagram of the neck from clinical literature.

L2. *Incorporate visual details specific to the user's activities.* When designing for the methodology development, we focused on developing the clustering algorithm and ensuring that the results were more meaningful than existing methods. Placing the cluster visualizations within a dendrogram allowed the users to scrutinize the inner workings of the clusters at different scales. In contrast, clinicians were more results-focused. Namely, their key interests focused on the spatial structure underlying each cluster, how the clusters related to outcomes and existing clinical categories, and if these correlations could be explained in a

way that was supported by clinical intuition. Thus, the design benefited from incorporating anatomical details and additional clinical covariates that were not considered when designing the model.

L3. *Show secondary variables and outcomes.* Design iterations that failed to include explicit labeling of results directly into the figure led to confusion. In the initial dendrograms, viewers had trouble connecting the clusters directly to other statistical analysis. For the clinical figures, collaborators often assumed that there were direct causal relationships between variables shown in the figure. In this case, it was useful to include potential confounding variables, to allow the readers to come up with alternative hypotheses.

L4. *Design for both interactive and static visualization.* In our experience, we started out with interactive designs aiming to assist a relatively small group of domain experts, who participated in the design process. Relatively quickly, it became obvious that the spatial clustering had to be explained to a broader audience that expected static visualizations, in the style of biomedical illustrations. Future works will stay closer to the illustrative style during the interactive model development phase, to reduce the cost of later redesign.

L5. *Build decision trees and conditionals to help explain spatial cluster differences.* When working with the broader audience, we found that the easiest way to explain cluster differences required explicit construction of decision trees, and "conditionals" based on the structure of the data—attempting to directly encode the differences was infeasible.

5.7 Conclusion

This work reflects on the process of designing visualizations for clustering with anatomical spatial data. These designs were developed in two phases over several years, using participatory design alongside collaborators with background in bioinformatics and radiation oncology. Through these designs iterations, we distill a set of lessons learned. While we focus on a particular problem, our design approach can be generalized to other type of cancer with spatially dependent data. These designs are part of a larger body of work borne out

of a multi-year collaboration with domain experts with anatomical cancer data. By incorporating additional insights from sibling projects, we aim to develop a comprehensive set of design guidelines for visualizing clusters of spatial data and effectively disseminating these results to domain expert audiences outside of the visualization community.

Acknowledgments This work is supported by the US National Institutes of Health, through awards NIH NCI-R01CA214825 and NIH NCI-R01CA2251.

5.8 Chapter Conclusion

This work is an effort to compile the results from multiple projects into a moral formal design characterization for VC+ML in clinical oncology work with spatial data. It presents the idea of actionability and domain sense as core goals in the design process. It presents a two-phase model for clinical research, which separates the model builders and model users. Additionally, we provide a domain characterization for unsupervised ML + VC in the clinical space. While we focus on only clustering here, future work may expand our work to consider supervised spatial ML as well.

Chapter 6

DITTO: A Visual Digital-twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer

The previous chapters focused mainly on developing spatially-aware VC+ML systems for the model-building and dissemination stage. These systems have targeted model builders, clinical collaborators, and in two cases wider biomedical audiences. While my work so far takes into consideration desirable characteristics of the system, such as actionability, it does not provide us with an opportunity to observe how these systems are received by domain experts from wider audiences. The work focuses on designing XAI for clinical practitioners, and explores the careful balance of model trust, as well as how to approach visual complexity for users without knowledge of the underlying model who have varying desired level of detail and visual literacy.

Digital twin models are of high interest to Head and Neck Cancer (HNC) oncologists, who have to navigate a series of complex treatment decisions that weigh the efficacy of tumor control against toxicity and mortality risks. Evaluating individual risk profiles necessitates a deeper understanding of the interplay between different factors such as patient health, spatial tumor location and spread, and risk of subsequent toxicities that can not be adequately captured through simple heuristics. To support clinicians in better understanding trade-offs when deciding on treatment courses, we developed DITTO, a digital-twin and visual computing system that allows clinicians to analyze detailed risk profiles for each patient, and decide on a treatment plan. DITTO relies on a sequential Deep Reinforcement Learning digital twin (DT) to deliver personalized risk of both long-term and short-term disease outcome and toxicity risk for HNC patients. Based on a participatory collaborative design

alongside oncologists, we also implement several visual explainability methods to promote clinical trust and encourage healthy skepticism when using our system. We evaluate the efficacy of DITTO through quantitative evaluation of performance and case studies with qualitative feedback. Finally, we discuss design lessons for developing clinical visual XAI applications for clinical end users.

The contents of this chapter were accepted for presentation at the IEEE Vis 2024 Conference and is published in IEEE Transactions of Visualization and Computer Graphics [342].

6.1 Introduction

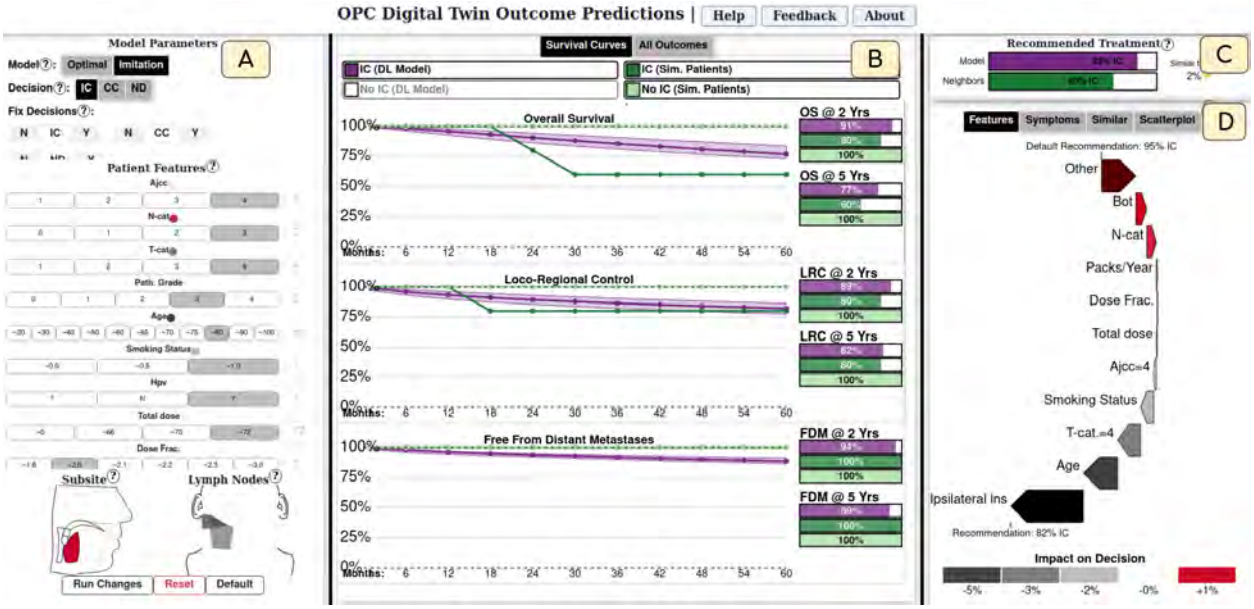


Figure 6.1: Overview of DITTO. (A) Input panel to alter model parameters and input patient features. (B) Temporal outcome risk plots for the patient based on different models and treatment groups. (C) Treatment recommendation based on the twin model and similar patients. (D) Auxiliary data panel, currently showing a waterfall plot of how each feature cumulatively contributes to the model decision.

Head and Neck Cancer (HNC) is a serious but treatable illness that affects up to 65,000 people each year in the United States alone. Care for HNC patients is a complex, multi-stage process that is dependent on the spatial location of the disease and its spread, and which includes potentially repeated cycles of surgery, chemotherapy, and radiation therapy. Determining the appropriate course of treatment for each patient is currently reliant on high level national guidelines and clinician cumulative experience. However, current guidelines do

not adequately address the wide range of individual patient responses to treatments or the dynamic adjustments clinicians must make in response. For example, treating patients with chemotherapy before radiation treatment may reduce the overall tumor size and therefore reduce the risk of severe long-term side effects, but may also increase mortality risk. As a result, there is exceptional interest in digital twin (DT) models of the treatment process to help HNC oncologists better understand the potential risks and benefits of different treatment decisions at each state in the treatment process. Digital twins are data-driven simulations of patients and how they respond to treatment, which can be used to tailor treatments for individual patients based on how they are expected to respond to different interventions. DTs require complex simulations of a patient's health at multiple points in treatment, and thus rely on models that are more complex than those typically used in clinical settings (e.g., logistic regression). Data visualization is an underutilized resource that can help clinicians interact more effectively with these digital twins.

Visualization for digital twins for subject-matter experts is an under-explored visualization challenge [244], with many additional challenges specific to HNC clinical decision-making. In terms of data, DTs consider multiple aspects of treatment, in addition to a combination of spatial and dynamic multivariate data to capture the patient state, which need to be visualized. In terms of outcomes, patient simulations yield dense, dynamic, and temporal outcome predictions, which need to be presented efficiently to users who may be interested in only a small subset of the resulting outcomes, depending on the context.

Furthermore, creating usable DT models also constitutes a visual explainable AI (XAI) challenge. While many XAI approaches have been developed for explaining models to model builders, less work has looked at the specific needs of model clients, who have unique requirements when considering both model performance and model explanations. For example, HNC clinical decisions may heavily depend on factors like spatial features and clinician experience, making simplification of results difficult. Issues with model explainability and actionability may be a factor in the low penetration of ML models in medicine (<2% [5])

beyond medical image analysis ML. Additionally, since existing models often contain biased or insufficiently diverse datasets to perfectly model the cohort, it is important to give recommendations that allow for model introspection and support appropriate trust in the recommendations while allowing physicians to identify cases when the model should be disregarded. Finally, complex model results need to be communicated to physicians while ensuring that the visualizations are sufficiently familiar so that they require minimal training.

In this work, we introduce a visual analysis interface for digital twins in oropharyngeal cancer treatment (DITTO). Our specific contributions are: 1) Requirements engineering of the factors that HNC oncologists consider when interacting with digital twin systems for treatment planning; 2) The design and implementation of a visual computing system with a dual digital-twin back-end, one twin (set of models) of the HNC patients, and one twin (set of models) of the HNC physician decisions; 3) The design of visual encodings for the visual computing front-end, with a focus on supporting clinicians and supporting both trust and skepticism in the models; and 4) A qualitative evaluation of the system with clinicians, resulting in visual digital twin design insights.

6.2 Related Work

6.2.1 Patient Risk Modeling

Research in head and neck (HNC) oncology focuses on evaluating ways of improving patient outcomes through changes in treatment. Current approaches have seen relatively high survival rates ($\sim 86\%$) in many HNC patients. As a result, current work often focuses on reducing side-effects (*toxicities* or *symptoms*) from treatment for patients with good survival probabilities. Earlier works have built interpretable models for predicting patient clinical outcomes for HNC patients such as survival and toxicity using clinical features [201], lymph node involvement [186, 349], tumor location [345, 346] and dose distributions [350], and radiomics [39]. This work is an extension of these approaches with a focus on temporally changing outcomes as well as intermediate treatment responses, which relies on more complex black-box models and post-hoc, instance based explanation methods for model in-

terpretability.

Risk modeling for patients with censored time-to-event outcome data like survival [175] is generally modeled using approaches such as cox proportional hazard models [305], non-parametric Kaplan-Meier analysis, and fully parametric models such as linear regression and survival trees [328, 368]. This work adapts a deep-learning approach to survival modeling called deep survival machines (DSMs), which use a fully parametric mixture of distributions fitted to the training data [232, 365]. Other approaches have adapted deep learning approaches to Cox proportional hazard models [234, 361] and attention-based transformer models for predicting survival [177]. However, none of these models account for differences in patient response during treatment.

In terms of Reinforcement learning, VA for interpretable RL is usually focused on targeting model builders [325, 327]. For clinical models, several systems have proposed attention weights for interpreting temporal neural networks [55,191]. In terms of visualization, RetainVis [163] focuses on exploring a recurrent neural network on temporal electronic health record data in patient cohorts. RMExplorer [166] uses subgroup statistics and feature attribution methods to explore model fairness in risk models.

More generally, DrugExplorer [329] proposed a general framework for XAI applied to drug discovery. In terms of presenting models to users, Suh et al [291] and Zitek et al [381] discuss strategies for communicating models to domain experts, but do not expand this to applications in decision support. Kaur et al. [148] showed that many users can "over trust" erroneous model explanations they don't understand properly. VISPUR [301] discusses methods of identifying spurious correlations in causal models, but do not focus on integrating domain expert knowledge.

6.2.2 Digital Twins

A digital twin is a digital model of a real-world system or process, that serves as the digital counterpart of it for practical purposes, such as simulation, integration, testing, monitoring, and maintenance. Although the term digital twin was introduced in 2010, visual steering of

detailed computer simulations (i.e., digital twins) has been used before for flood simulation planning [337] and VR applications for manufacturing [380].

In healthcare, limited work has been done in exploring digital twins for patients using dashboards [146, 171] and 3D models to visualize blood flow [209]. Digital twin tools have also made for simulating physicians [296]. Other approaches have built digital twins for radiation dosage adaptation [310], glioblastoma treatment [83], and emergency department management and [29], but do not integrate visualization or explainability. Marai et al. [202] developed a web visualization tool for HNC patient risk based on similar patients that allows for what-if analysis. Our work uniquely integrates visualization for both a digital twin and digital physicians. Additional, to our knowledge, there has been no interactive visual computing approach for digital twins that can also factor temporal decision-making.

6.2.3 Decision Support Systems

Relevant to this work is clinical decision support (CDSS) systems. Jacobs et al. discuss a CDSS for clinical depression [133]. Other systems have focused on identifying ways of supporting physician workflows for heart implants [363], critical care patients [373] and diabetes care [35]. Other work has focused on model building for CDSS Bayes networks [230], and integrating feature explanations to help train physicians in diagnostics [246]. More generally, a recent study has suggested that users are more likely to use AI recommendations for harder tasks [118]. Despite this, few visual systems have focused on decision recommendation in the context of explainable ML recommendations.

Several systems have been developed specifically to visually communicate risk prediction to clinical users or patients, although none of them focus on deep learning-driven personalized patient outcomes. A majority of these systems focus on variants of Kaplan Meier plots to communicate patient survival based on general diagnostic features [62,67,319]. Oncofunction [369] focuses on helping patients plan post-treatment symptoms. PROACT [120] found patients were primarily interested in time left and survival risk at different time points using simple visualization methods. Vromans et al. [323] found that some information seeking was

a coping mechanism for a percentage of the population, and personal quality-of-life measures were equally as important as patient survival. Floricel et al. [96, 97] uses temporal glyphs and Sankey diagrams to show clusters of patient symptoms over time. Other common tools have used Kaplan-Meier curves [319] and nomograms [102] for HNC and prostate cancer. However, as far as we know, no online systems yet include digital twins with temporal state outcomes, or use model explainability methods with patient-specific predictions.

6.3 Methods

6.3.1 Requirement Analysis

This project was developed as part of a multi-year collaborative research project between HNC oncology radiotherapists from the MD Anderson Cancer Center, machine learning experts at the University of Illinois Chicago and University of Iowa, and visualization researchers at the University of Illinois Chicago. We followed the Activity-Centered Design (ACD) methodology, which has higher success rates in interdisciplinary settings than Human-Centered Design (63% vs. 25%) based on a survey of design studies [199]. Requirements for both the interface and models were initially gathered through interviews with three research oncologists, and gradually refined during weekly meetings through multiple rounds of parallel prototyping and feedback over the course of several months.

To gather formative feedback, a version of the interface designed based on our initial requirements was presented to a group of 11 physicians with clinical experience within the HNC oncology group at the MD Anderson Cancer Center. Several participants were familiar with the underlying dataset, but none had participated in the design of DITTO. During the session, participants were given an overview of the system components before being given a demo of the system and an online link where they were allowed to interact with the system on their own. This was followed by an open-ended discussion and a feedback interview.

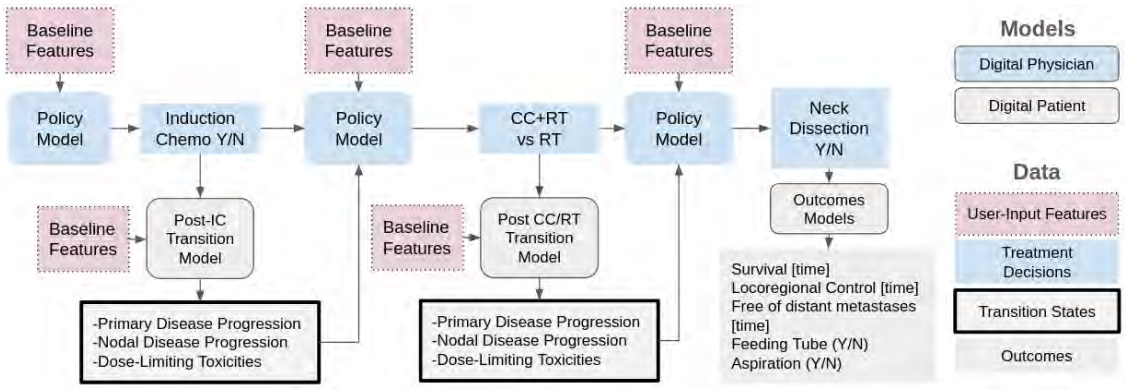


Figure 6.2: Overview of the treatment sequence simulated by the digital twin models, along with the features to be considered. Intermediate results of induction and concurrent chemotherapy are used as inputs into the next decision. Final outcomes are a mixture of time-to-event curves and fixed binary outcomes. The DT model is trained to make optimal decisions with respect to the final outcomes.

6.3.2 Data Abstraction

Our dataset uses the patient cohort described in Tardini et al. [297]. The cohort consists of 526 anonymized patients with squamous cell oropharyngeal tumors treated using definitive radiation therapy at the MD Anderson Cancer Center between 2003 and 2013. All data were collected after approval from the MDACC IRB (PA16-0303 and RCR03-0800). All patients included also had either recorded deaths or a minimum followup time of 4 years. Patients diagnostic data, treatment sequence, and outcomes were collected using EHR records.

Standard treatments for patients include a mixture of *surgery*, *chemotherapy*, and *radiation therapy* (RT). Chemotherapy can either be given before RT (induction - IC) or with RT (concurrent - CC). While real treatment can involve multiple rounds of each therapy, our simplified treatment sequence models the treatment process as 3 decisions: chemotherapy before RT (IC), chemotherapy concurrent with RT (CC), and neck-dissection (ND), a common surgery. These decisions are critical decision points, aligned with the standard-of-care [1]. The entire treatment sequence model is shown in Fig. 6.2. Baseline features include age (cont.), if the patient is male or female/nonbinary (binary); race (binary x3); which regions of the neck have affected lymph nodes gregoire2014delineation (binary x14); smoking status (never, former, current) (ord.); total radiation dose to the tumor (cont.) and dose per-visit (cont.); tumor staging (ord. x3); and tumor subsite (categorical x6). Tumor staging fea-

tures (T, N, and AJCC) are ordinal rankings of tumor severity used to decide on treatment regime based on tumor size and spread for the main tumor and nearby lymph nodes [13]. Race used a simplified grouping of demographics: White, African American/Black, Hispanic, and "Other", which is modeled as three one-hot variables in the data input. Minority inclusion reflects the demographics of the MD Anderson Cancer Center patient population which is approximately 85% Caucasian and 15% minority. Default gender is denoted as "male" in the model, to reflect the demographics of the patient population which is approximately 30% females and 70% males and did not have information on nonbinary individuals.

Each feature is associated with a feature-importance during modeling, based on how much it contributes to the twin model final decision, which is a value between 0 and 1.

After each decision, the patient response to treatment is modeled as a transition state in terms of the response (change in size) of the primary tumor, nodal tumors, and any dose-limiting toxicities (DLTs). Tumor response is categorized into 4 groups based on the amount of change in tumor size: progressive disease, stable disease, partial response, and complete response. DLT types considered in our model are: Hematological, Neurological, Dermatological, Gastrointestinal, or Other. All DLT categories with fewer than 3 instances in the dataset are grouped into the "other" class.

For temporal outcomes, we consider patient survival (OS), local-regional control (LRC), and distant control (FDM). For each of these, we collected whether the event occurred, as well as either the time of the event or last follow-up date. Additionally, we recorded whether the patient was hospitalized for a feeding tube (FT) or lung aspiration (AS) within 6 months after finishing treatment as binary toxicity outcomes. As an auxiliary outcome, we extracted symptom ratings from a separate dataset of 937 patients with self-reported outcomes after receiving radiation therapy [332], which is used in a secondary view to display possible symptom trajectories.

6.3.3 Digital Twins and Planning for Trust and Skepticism

One of our goals is to provide support for both trust and skepticism in the system recommendations. In prior work [344], we have discussed visualizing "counterfactuals", where the model recommendation and ground truth diverge, and adding cues to highlight when model predictions should be given more scrutiny. Because DITTO aims to provide treatment recommendations for a new patient, where the ground truth is not available, we implement instead "neighborhood-based models" that are shown alongside the treatment recommendation and predicted outcomes, to encourage both trust and skepticism in the digital twin recommendations. These neighbor-based models show outcomes from similar patients in the cohort and are described in Section 6.3.6.

Our core dual digital twin system is based on modeling of patient responses at each time point for a given patient, alongside modeling of the physician-recommended treatment. We specifically refer to the patient response models as the "Patient Simulator", and the predicted physician treatment decisions as the "Policy Model".

To further encourage trust and skepticism, and avoid reinforcing clinical bias, we planned to leverage and show recommendations from two deep learning models for the twin. In the context of modeling a physician we implement two approaches: imitation learning [378], which attempts to mimic what an expert would learn, and Deep Q Learning (DQN) [320], which attempts to find an optimal decision based on expected future losses. e also implement a preliminary imitation learning model for use by clinicians. We refer to the DQN strategy as the "Optimal Policy Model" and to the imitation learning strategy as the "Imitation Policy Model". For the purpose of the interface, viewers can select the specific strategy to be used, and examine results from that model strategy. These two supervised deep learning models (DQN and imitation) are "Digital Twins" of the physician decision process, in addition to the patient simulator.

In total, DITTO leverages and can show three recommendations: two based on the DTs of the physician decision process, and one based on the neighborhood models.

6.3.4 Task Analysis

Our system aims to help HNC oncology radiotherapists better understand the likely tradeoffs of adding other treatments to radiation therapy. Based on interviews, we found that clinicians generally consider treatment decisions at each stage individually, with a primary focus on identifying potential outcomes in terms of both immediate disease response, toxicity risk, and overall temporal outcomes up to 5 years. Individual interests, degrees of information seeking, and visual literacy varied based on the individual practitioner and their backgrounds. As a result, our system was designed to have flexibility, with the most prominent views being presented by default, and more detailed views available on demand.

Additionally, we have found during our collaboration that some clinicians trust their past clinical experience over neural networks and cohort-based reasoning, while others tend to trust the model even when the system does not make sense. As a result, our main design focuses on simultaneously showing results from both the supervised deep learning models used in the digital twin and neighbor-based models that use similar patients in the cohort (Sec. 6.3.6), to cue the user to have appropriate trust and skepticism in the system.

Based on interviews and clinical feedback during the prototyping stage, we arrived at the following task abstraction:

T1. Identify the risk profile of a patient given a treatment selection.

1. Display the temporal risk of negative outcomes for the individual patient using the digital twin
2. Identify the cumulative patient risk in terms of the cohort of similar patients in the dataset
3. Identify the ideal treatment plan for the patient
4. Compare the patient to similar patients based on treatment and diagnostic data
5. Display expected patient symptom profiles after radiation therapy

T2. Identify relative benefit of treatment at the given time point.

1. Display the potential gain in therapeutic efficacy in terms of survival, disease control, and additional side effects
2. Compare expected cumulative tumor control and survival to the probability of additional toxicity due to treatment for the patient
3. Display the risk of dose-limiting toxicity due to chemotherapy or treatment complications

T3. Identify the trustworthiness of the model predictions and recommended treatment

1. Show the cumulative impact of each attribute on the recommended treatment in terms of percentage confidence
2. Flag when the patient is an outlier in the cohort
3. Display confidence intervals for the patient outcome predictions
4. Compare the prediction of the DT and neighbor-based models

Nonfunctional Requirements In addition to tasks, we determined a number of nonfunctional requirements. DITTO needed to build via visual scaffolding [198] on encodings in existing clinical tools, such as Kaplan-Meier plots, barcharts, and cumulative distribution histograms. Additionally, several clinicians desired to be able to show these results to patients, and thus designs needed to avoid causing patent anxiety (i.e., scale survival should show risk always compared to 0). Finally, DITTO needed to be responsive and available online to be used by clinicians at any time, with minimal (< 5 seconds) time to produce results for a new patient.

During the workshop, two participants requested information about data provenance and the model details, including limitations, available in the interface. Additionally, participants asked for the patient inputs to always be visible, and to only render additional views once an input has been manually submitted. Our original design also included both survival plots and barcharts of all outcomes and predicted transition states at the same time, in addition to median time to event for each temporal outcome. However, clinicians stated that most

use-cases would focus on a smaller subset of results: the survival plots and survival at 2 and 5 years, with uncertainty values given, and that these designs should be centrally located, and additional results could be given on-demand as exact values.

6.3.5 Deep Reinforcement Learning Models

DITTO uses an extension of the dual digital twin system described in Tardini et al. [297]. The full system is shown in Fig. 6.2. Patients are assumed to follow a series of 3 binary decisions: Induction chemotherapy (IC), Concurrent chemotherapy (CC), and Neck Dissection (ND).

Our digital twin is composed of multiple sub-models at each state in the treatment sequence, which are shown in more detail in Fig. 6.2. For the purpose of this section, we define terminology when referring to each of these subcomponents. We call a model that predicts the patient's direct response to each treatment the "Transition Model", and the model that predicts long term temporal outcomes after definitive treatment is completed (i.e. survival and recurrence) the "Outcome Model". Following RL terminology, we refer to the model that simulates a physician as the "Policy Model". We have two versions of the policy model: The "Optimal Policy Model", and the "Imitation Policy Model", which attempts to predict the best treatment in terms of long term outcomes, and the treatment a physician would make, respectively. We only use one Policy model at a time, which is defined by the user. We use deep learning for all DT models due to their ability to deal with multimodal inputs with variable outputs and handle missing data [253]. The following section briefly discusses the details of each model.

To supplement the Digital Twin predictions, we show alternative predictions in the interface based on the most similar patients in the cohort at the given timepoint. We refer to this as the "Neighbor-based models" collectively, as we do not have to simulate responses at each step since all patients have ground truth decisions and patient responses available.

Below we briefly describe each model. Due to space constraints, full details, model parameters, and evaluation can be found in the supplemental material Appendix A.

Patient Simulator

To simulate the patient, we use a set of models to mimic intermediate response to treatment (transition models), and long-term response after treatment (outcome models).

Transition models predict patient response to treatment in terms of tumor shrinkage and severe toxicities from treatment. Specifically, we consider primary disease response (PD), and nodal disease response (ND), which are each 4 categorical ordinal variables, as well as 5 binary results for different types of dose-limiting toxicities (DLTs). For induction chemotherapy (IC), disease response is always assumed to be stable when no treatment is done.

For post-treatment outcomes, we predict a combination of temporal and static outcomes. We predict static outcomes using a deep neural network that predicts hospitalization due to two severe toxicities at up to 6 months after treatment: Aspiration (AS), and Feeding Tube insertion (FT). The temporal outcome model predicts cumulative patient risk over time for overall survival (OS), locoregional control (LRC), and distant metastases (FDM) for up to 5 years. Temporal risk models use a variant of deep survival machines (DSM) [232]. For all three outcomes, the DSM model returns a mixture of parametric log-normal distributions for the patient that can be used to provide a cumulative survival risk over time.

Because clinicians listed confidence intervals as important for reasoning about the model predictions (T3.3), all transition and outcome models are trained using dropout on the penultimate layer between 50% and 75% [104]. During evaluation, we re-run each prediction with random dropout at least 20 times, and then save the 95% confidence intervals for each prediction.

Policy Modeling

The patient simulator models and ground truth responses are used as the environment to train a digital physician (policy model) . The policy model is a deep-learning based

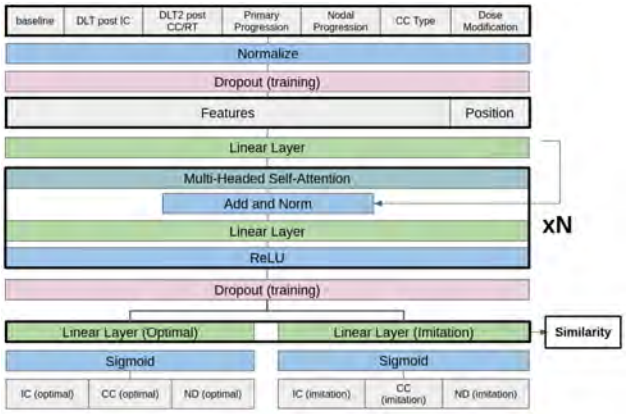


Figure 6.3: Architecture for the model used to simulate a physician decision. We use a shared embedding with a custom position token at each stage, followed by a separate layer for each output. Model activations for the penultimate layers are used when calculating similar patients. Policy models use a modified version of a transformer encoder that saves the cohort at each time point into memory at training time.

transformer encoder that predicts a binary treatment decision based on the baseline patient features, response to the previous treatment, previous decisions, and current timepoint.

Because we need to explain the policy model recommendations (T3), we use integrated gradients [293] to obtain feature importance for each decision relative to a baseline value. Integrated gradients was chosen as it satisfies the completeness axiom where attributions sum to the difference in the prediction between the baseline and actual recommendation, which was found to be easier to reason about with our clinicians. For our baseline, we assume the lowest possible rating for most ordinal attributes such as tumor staging or disease response, and the most common value for categorical attributes such as gender, ethnicity, and tumor subsite, as well as age and dose to the main tumor, based on feedback from clinicians and what they found most intuitive.

6.3.6 Neighbor-based Models

To provide an alternative model prediction to improve user trust (Section 6.3.4). we provide methods for estimating different patient outcomes using similar patients in the cohort, based on the embeddings taken from the final layer in the policy model for the given time-point and output. Our approach uses a modified variant of average treatment effect, which is used in causal modeling for finding predicted effects from treatment while correcting for confounders.

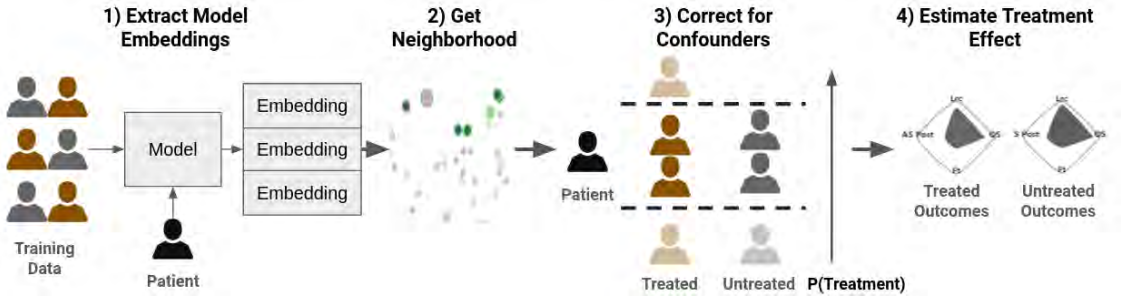


Figure 6.4: Diagram neighbor-based models when predicting patient outcomes. Model embeddings from the policy model are used to extract the most similar patients. Neighbors are filtered by their estimated likelihood of receiving treatment from the imitation model for those closest to the new patient. the difference between untreated and treated filtered neighbors can then be used to estimate impact of treatment.

For a new patient, we calculate a set of k patients whose embeddings are most similar at each time point in terms of embedding using euclidean distance. When predicting treatment policy (physician choices), we use a smaller subset of the $n, n < k$ most similar patients and report the percent of patients that received treatment. For other outcomes and patient response, we take from the k patients those with a predicted probability of receiving treatment that are within a certain value of the patient. We then calculate the relative prevalence of each outcome for the untreated and treated patients within this propensity-matched [17] group (Fig. 6.4). For our system, We calculate the value difference as a fixed percentage of the standard deviation of the logits of the propensity scores in the cohort, defined as:

$$cd = \alpha * \sqrt{\frac{1}{|X|}\sum_{x \in X}\left(\ln\left(\frac{p_x}{p_x - 1}\right) - \frac{1}{|X|}\sum_{k \in X}\left(\ln\left(\frac{p_k}{p_k - 1}\right)\right)\right)^2}$$

Where X is the cohort and p_n is the predicted probability of patient n receiving treatment. We use an α of .1 based on the suggested formula in [18], which is increase in increments of .1 until treated and untreated groups have at least 5 patients.

6.3.7 Implementation

Our back-end was implemented in Python using flask and pandas for data processing. Deep learning (digital twin) models use Pytorch, and deep survival machines use modified code taken from the auton-survival package [233]. Feature attributions were calculated using the Captum package [154]. Our system front-end uses react with d3.js. Our online interface requires approximately 3.6-4.5 seconds to return simulation results for a new patient with

two cores on an AMD EPYC 7452 Processor and requires 4GB of ram with 4 worker processes on the server, based on test queries for 10 random patients in the cohort. Specific model parameters were chosen via model tuning are given in the supplemental material.

6.4 Design

6.4.1 Layout and Workflow

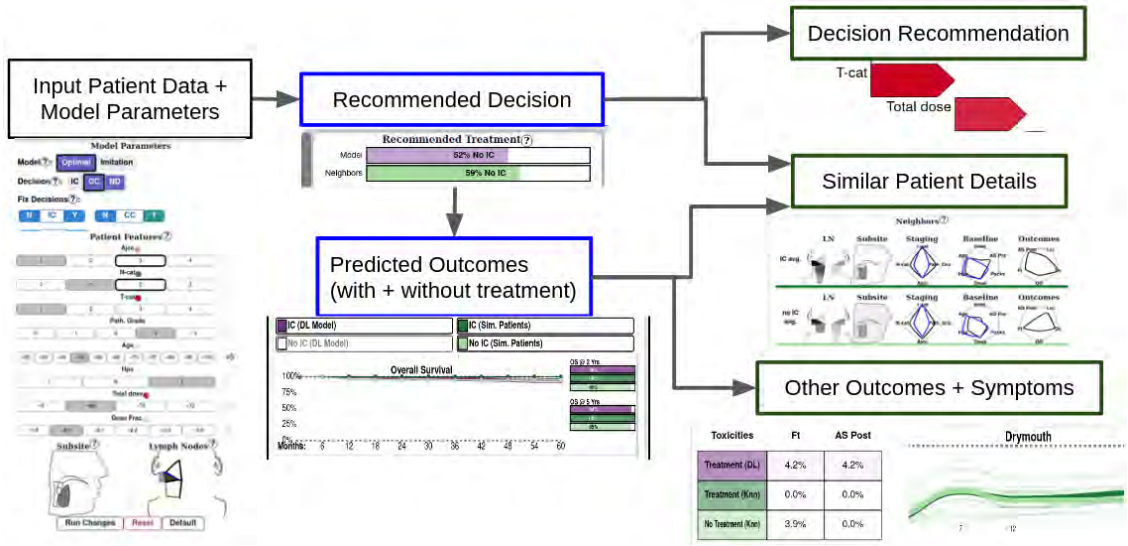


Figure 6.5: Diagram of user workflow using the interface. (Left) Users start by inputting patient features and setting the model parameters, including the treatment to be considered. (Center) Users start by viewing the most prominent information: treatment recommendations and long-term patient outcome risk plots for survival and disease recurrence. (Right) Users who wish for more information can view additional views such as model explanations, similar patients, and additional patient risk prediction results.

Our main system is divided into three main components: input, patient outcomes, and treatment recommendation + supplemental views (Fig. 6.5). First, an input panel on the left is used to change model and patient details (Fig. 6.1-A). To minimize cognitive load, we focus on only showing one, user-selected treatment (IC, CC, or ND) at a time. Users can optionally decide on other treatment decision when calculating future patient outcomes, with the policy model handling the other treatment decisions when nothing is input by the user. Next, central views show patient survival outcomes (Fig. 6.1-B) as well as the recommended treatment for the patient (Fig. 6.1-C). Finally, additional views are shown via tabs to users who have an interest in more detailed information, such as model feature explanations (Fig. 6.1-D), similar patients, additional outcomes, and predicted symptoms

ratings. These views are changed by toggling a set of buttons above the panel. Because many views are only of interest to certain users, we added functionality to resize width of each view via dragging the black vertical dividers, to allow users to expand auxiliary views as needed, while keeping the main goal of evaluating patient outcomes the main focus.

Whenever model predictions are shown in the interface, we present the deep-learning based Digital Twin predictions, and the neighbor based models. We use purple to encode Digital Twin predictions, and green to encode similar patient predictions.

6.4.2 User Input

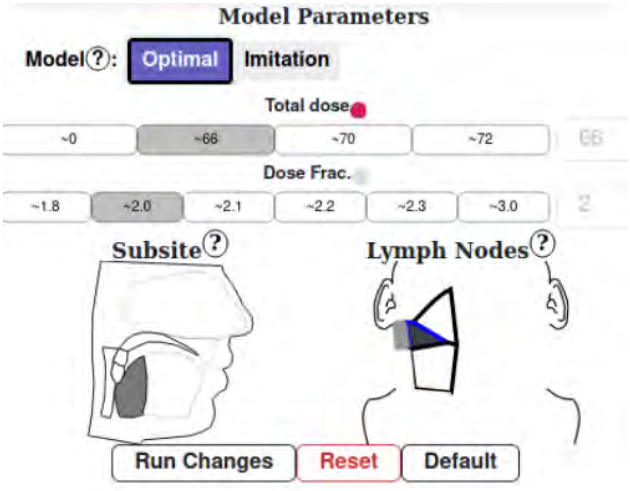


Figure 6.6: Examples of model and feature inputs for DITTO. (TOP) Toggleable model parameters. (Center) Unstructured feature inputs given as both buttons and free-form input. (Bottom) Spatial inputs for tumor subsite and affected Lymph Node levels. Colors next indicate feature importance in the current prediction.

The left panel allows inputting the relevant patient features and model parameters into the system. At the top, prompts are given for model input parameters: 1) whether the policy model should use the "optimal" or "imitation" strategy (Section 6.3.3); 2) what decision is being considered; and 3) if any of the other decisions in the system are assumed to be "fixed" (yes or no). By default, the decision is decided by the currently selected policy model's recommendation.

Below the model parameter input is a panel for the current patient (Fig. 6.6). By default, the average values for each feature are selected. We found that clinicians tend to think of continuous variables such as smoking pack-years and age in terms of discrete "bins"

therefore, all features are shown using categorical stylized radio buttons to make selection easier, with free-text inputs on the side that allow users to use specific values when desired. These values are checked for validity based on the feature. When analyzing concurrent chemotherapy or neck-dissection, users can either specify the patient's primary and nodal tumor response to the previous round of chemotherapy, or allow the system to estimate this response automatically.

For the spatial inputs: affected lymph nodes and tumor subsites, we allow users to directly interact with diagrams of the respective areas. The diagram for the lymph nodes was previously developed alongside clinicians in our work with explainable lymph node clustering [343]. The diagram of tumor subsites were adapted from diagrams created by the MD Anderson Cancer Center. See the Appendix B for a labeled description of each spatial diagram.

In addition to feature input, we include color cues in the feature attribution plot for each of the features, described in detail in 6.4.4 (T3.1). These are shown as colored dots next to each feature for nonspatial inputs, and as a color fill in the spatial features.

Because we do not want to re-run the computationally expensive simulation every time a feature or parameter is changed, a new simulation is run using the updated features once the user selects the "run changes" button at the bottom. Additional buttons reset the feature inputs to the last time the simulation was run, and load the default patient features.

6.4.3 Survival Plots and Outcomes

When collecting feedback from HNC clinicians at the MD Anderson cancer center, several clinicians suggested that users with less information seeking behavior will primarily be interested in seeing tumor control and survival risk for treated and untreated groups over time. As a result, we centrally place an outcomes view panel (Fig. 6.7) that shows the model predictions for all relevant endpoints in our system. By default, we show temporal plots for survival, local-regional control, and distant metastasis for the treated and untreated groups using the Digital Twin outcome models (T1.1) and neighbor predictions (T1.2), up to 60

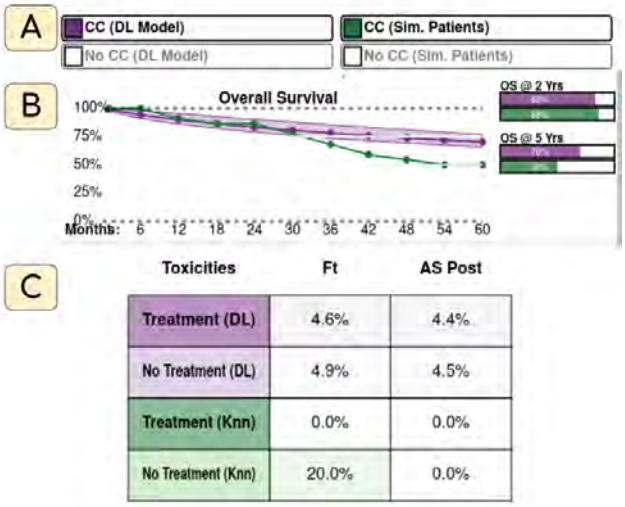


Figure 6.7: Image of survival curves for a patient based on different models. (A) Legend with toggle-able models and outcomes, currently showing only treatment groups. (B) Survival plot for a patient, showing prediction with concurrent chemotherapy and 95% CI based on the DSM model (purple) and similar patients (green), along with fixed probabilities at 2 and 5 years. (C) Alternative outcomes view showing tables of predicted probabilities for additional toxicities (Ft - Feeding Tube, AS Post - Aspiration Post-Treatment).

months post-treatment (Fig. 6.7-B). We also include 90% confidence intervals for Digital twin predictions as semi-transparent envelopes (T3.3). We chose to use temporal outcome plots as the main outcome plot, as oncologists often use variants of Kaplan Meier survival plots to assess patient risk. Additionally, the legend at the top can also be used to toggle off the visibility of certain models or treatment groups when the user only wants to see predictions for certain parameters (T3.4) (Fig. 6.7-A). Each output is color-coded, where hue encodes model group (Digital twin vs neighbor-based) and luminance encodes treatment group (darker for treated groups).

Because a subset of information-seeking clinicians were interested in more details regarding patient response, an alternative window (Fig. 6.7-C), shows static risk tables for all transition outcomes and temporal risk at 2 and 5 years for both Digital twin and neighbor-based predictions, for both the treated and untreated groups, for a total of 4 predictions each, via a toggle button (T2.1, T2.2, T2.3). This view relies on direct encoding of features. Additionally, each cell is color coded, with opacity encoding the risk percentage. These additional results were originally encoded as a barchart shown alongside the survival plots, but were moved to a simpler, more explicit table shown on demand based on clinician feed-

back as well as recent findings suggesting that tables with explicit values are less prone to confirmation bias when reasoning about the data [359].

6.4.4 Treatment Recommendation

The right panel of DITTO is devoted to more detailed model results, based on the varying requirements cited by different clinicians. We show the recommended treatment based on both the policy model, and similar patients at the top, in terms of a percentage between 0 and 100% for the suggested treatment (Fig. 6.1-C) (T1.3). To provide a cue as to how reliable the model recommendation is, we calculate the Mahalanobis distance between the patient embedding taken from the model for each time point and the rest of the cohort (T3.2). We then calculate the relative percentile of the distance for this patient relative to the rest of the cohort (e.g., 0 to 100%), which is shown next to the recommendation. We show a symbol (thumbs-up vs thumbs-down) based on if the percentile is below or above 75%, respectively. This feature was based on a specific clinician request for a cue regarding whether the new patient recommendation can be trusted based on the cohort being used. Our original design included a full histogram. However, during the workshop, several clinicians misread the histogram, as some assumed being in the middle was better and others assumed the left was better. Additionally, clinicians did not find seeing the distribution of the full training cohort useful, and thus recommended using a text rating.

Below the model recommendation, a panel shows additional custom model details. By default, the view shows a waterfall chart variant(Fig. 6.8). This view shows the cumulative impact of each attribute on the final decision in terms of percentage confidence in the given treatment on the x-axis (T3.1). The baseline shows the decision impact for a "default" patient, which is either the lowest possible value for ordinal (e.g., tumor staging) or continuous values, or the most common value for categorical features. We then show the impact of each feature as an error moving the decision along the x-axis. Because the integrated-gradients feature attribution method satisfies "completeness", the final position at the bottom is equal to the position of the final decision relative to the first decision point. Each bar is drawn

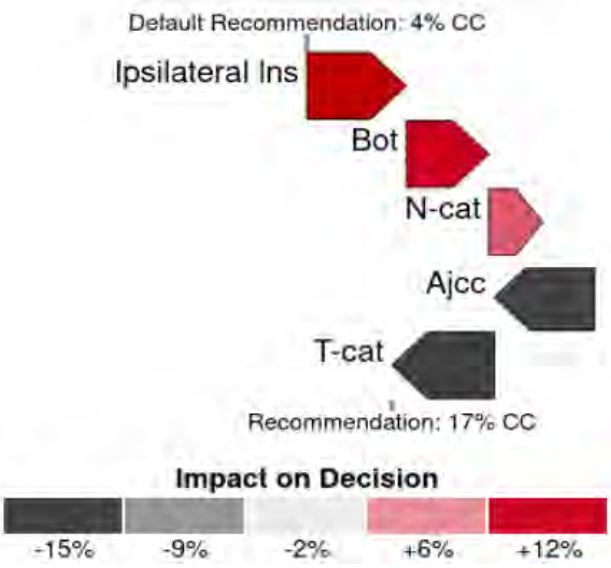


Figure 6.8: Truncated feature contribution waterfall plot showing how each feature contributes to the final model recommendation, relative to the default (median) patient. Color double-encodes attributions

as an error that uses a diverging color scheme to double-encode impact size. All values below a certain threshold (1%) are aggregated into an "other" value as they have negligible interest to users. Features are shown in order of positive impact from the top to the bottom. This view was finalized as waterfall charts are an established method of showing feature attributions [123], with the arrows and color encoding added to improve intuitiveness of the system. Additionally, it was very well received by clinicians during prototyping, and described as "very intuitive" by a collaborator with no prior experience with feature attributions.

6.4.5 Similar Patients

Based on interviews and previous experience with clinicians, many HNC oncologists are interested in using previous patients to reason about likely outcomes and the trustworthiness of the prediction and improve domain sense. As a result, we include an optional view that shows details on the similar patients used in the Average Treatment Effect estimates (Fig. 6.9) (T1.2, T1.4). The view shows feature summaries of each patient, as well as the average values for the treated and untreated groups. Each patient is encoded as a single row of patients. We show the tumor subsite and lymph nodes as heatmaps using the diagrams described in

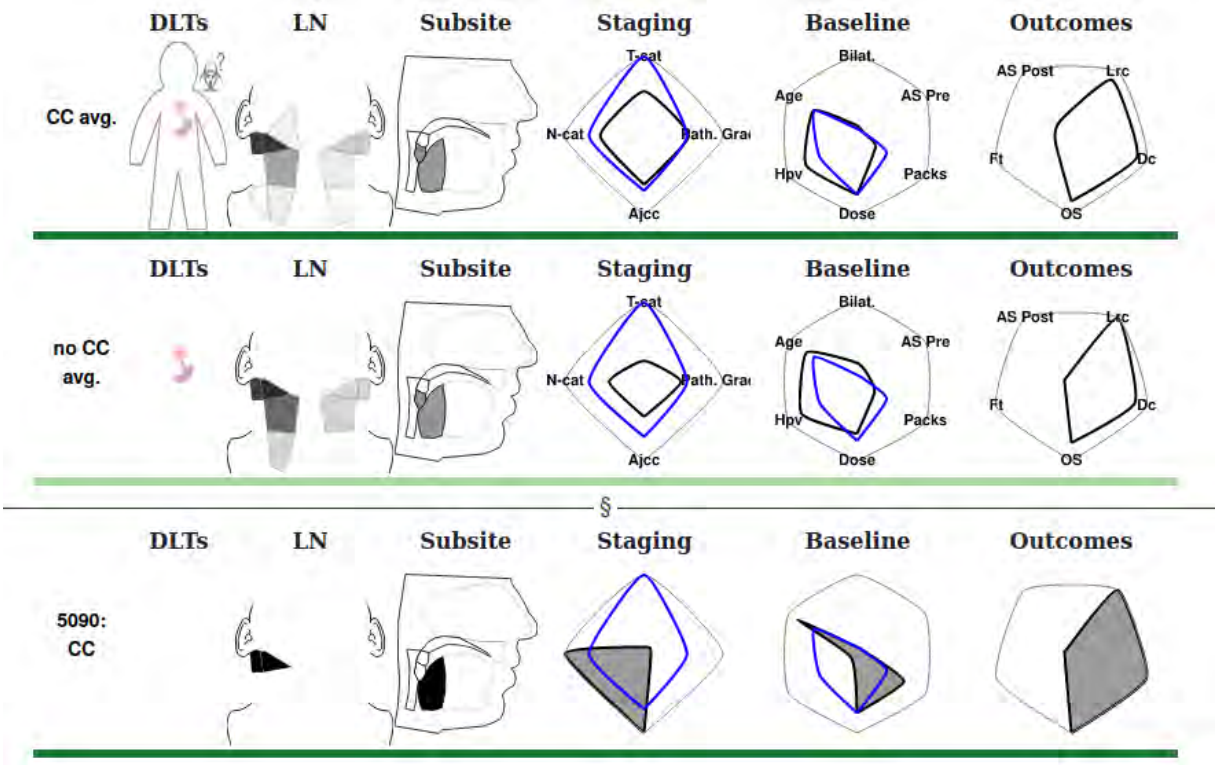


Figure 6.9: Similar patients view showing the row for average treatment group. Each row shows toxicities, lymph node involvement, tumor subsite, staging, demographics, 4 year outcomes. Blue lines indicate the input features of the current patient in the staging and baseline Kiviat diagrams.

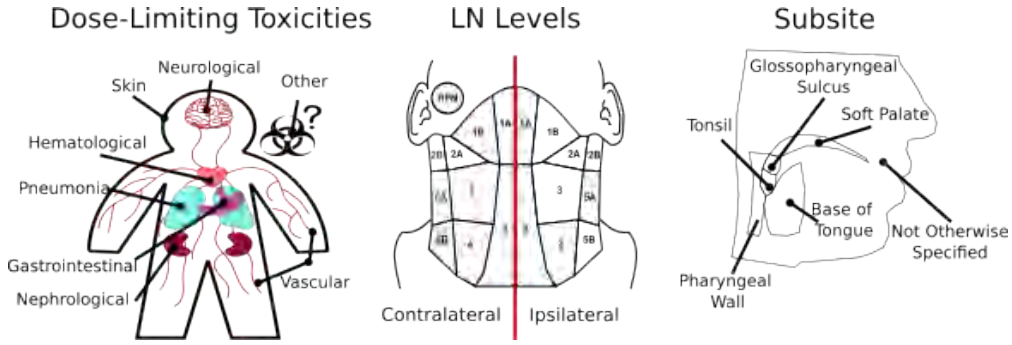


Figure 6.10: Diagrams used for spatial features in the visualization. (Left) Dose-limiting toxicities. (Center) Lymph node regional involvement. (Right) Primary tumor subsite. All regions not included in the diagram are considered "Not Otherwise Specified".

Section 6.4.2, as well as a diagram for and dose-limiting toxicity from the current treatment (Fig. 6.10). Additionally, we show three Kiviat charts with distributions of the most relevant features: diagnostic tumor staging (T-stage, N-stage, Overall Stage, and pathological grade), important clinical features (HPV, smoking status, age, etc.), and patient outcomes at 4 years (survival, local-regional control, distant control, aspiration, and feeding tube). The features for the current patient for non-outcome features are overlaid on top of each patient in blue,

to support comparison between the groups and the current patient. This design was based on prior work showing promising results for diagram based spatial encodings [343, 344], and radial charts to encode clinical features [195, 201] when displaying similar patients for clinicians, along with positive feedback from collaborators.

We use colored borders and labels to indicate which patients are in the treated and untreated groups. This view is included as it was found to be useful for clinicians that value inspecting individual patients, or identifying confounders that may impact the recommendation of the neighbor-based predictions. However, since many clinicians said this functionality was only a secondary concern, it is hidden by default.

6.4.6 Symptoms

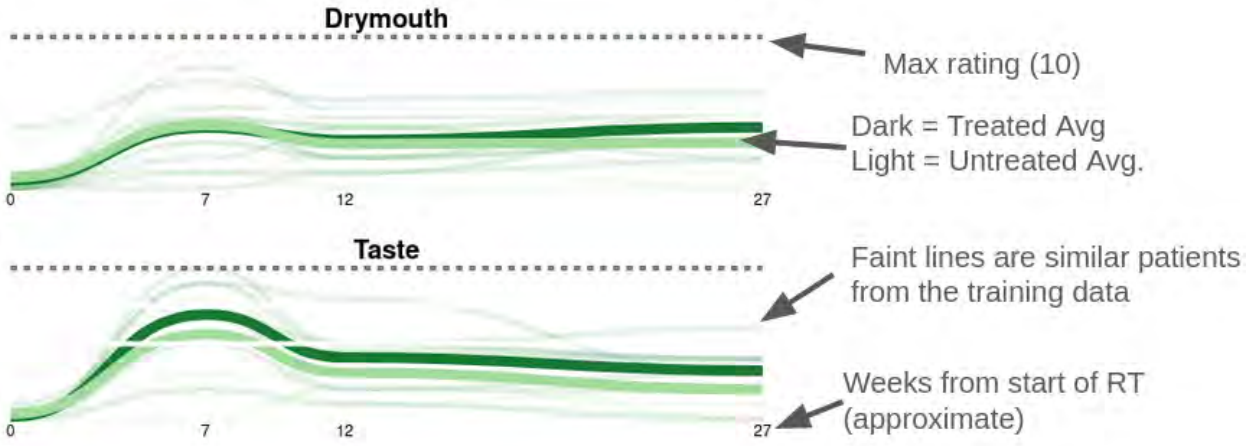


Figure 6.11: Symptoms prediction for a patient. Dark green indicates average of patients that receive a selected treatment, light green is average of patients that don't receive treatment. Faint lines indicate trajectories of the cohort patients used to make the prediction.

Finally, because several oncologists expressed a desire to see the effect of treatment on long-term subclinical side effects, we include a KNN-based symptom progression model for the patient (Fig. 6.11) (T1.5). Due to data constraints, this view only includes a neighbor-based model, with no deep-learning based model as the cohorts were different and we were unable to get a sufficiently accurate model. This view shows self reported symptom progression for 10 different symptoms for a period of 6 months after the start of radiation treatment. Each similar patient is shown as a faint line, and group median values for treated and un-

treated groups are shown as bold lines. Symptoms are ordered by mean rating at the end of the time-period, as clinicians are most interested in long-term side effects that are more likely to be permanent.

6.5 Qualitative Evaluation

A quantitative evaluation of the models used in the system is included in the supplementary materials. To further evaluate DITTO, we performed two case studies with two users: one HNC clinician with 9 years of experience, 4 years of which were at the MD Anderson Cancer center, along with one Data Mining researcher, to find out how oncologists interact with the system. The case studies covered the evaluation of a single patient each and were performed via Zoom meetings with desktop sharing. To assess how different model recommendations might affect the users, we selected one patient that had both the neighbor-based and DT model agree with the true patient recommendation (non-counterfactual) and a case where the neighbor-based and DT disagreed with each other (counterfactual). The policy model strategy was set to "Imitation" based on clinician preference. Qualitative feedback was collected via a debriefing interview derived from the System Usability Scale [32] structure.

6.5.1 Typical Recommendation

Our first case study was taken from an example patient where both the neighbor-based and Imitation policy model agreed with the clinical ground truth. Starting with the patient input, the patient was notable for having a high T-stage (large primary tumor) and "Not Otherwise Specified" tumor location, suggesting that the patient had a large, irregularly positioned tumor, and being African American. The clinician then moved to the treatment recommendation (Fig. 6.12-A) to confirm that the model recommendation lined up with the similar patients in the cohort, where 100% of patients receive chemotherapy. Looking at the feature importances for the policy model recommendation (T3.1), they noted that the most prominent features are the LN spread, the patient's race, the pathological grade, and the T-staging. While this finding mostly lined up with clinical reasoning, the impact of race was

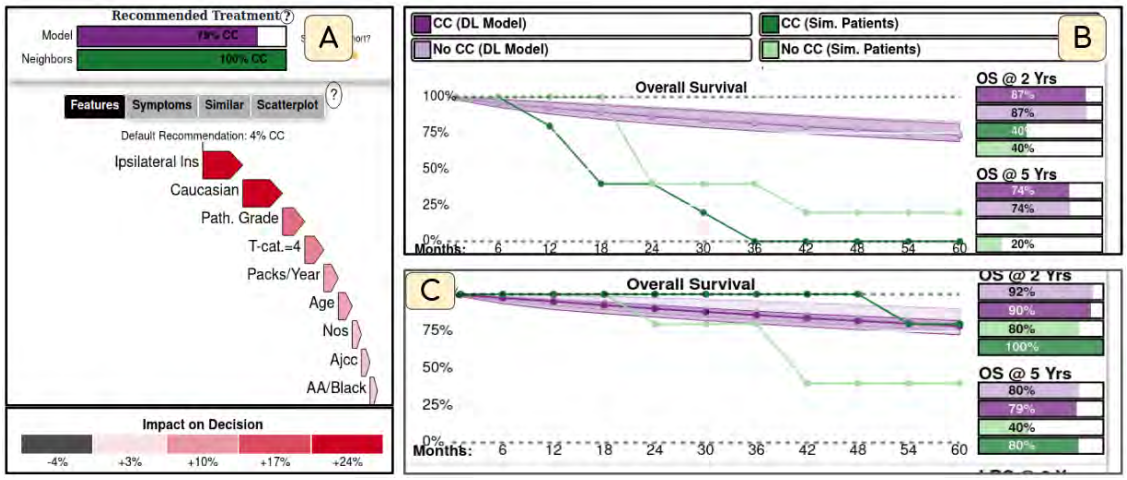


Figure 6.12: First case study. (A) Feature importance and recommendation (truncated) showing that LN spread and Race are the main predictors of the patient receiving CC. (B) Patient survival curves. Green lines show very low survival for both treated (dark green) and untreated (light green) groups, but high survival from the DSM model (purple). (C) Survival curves for the patient when their race is changed to "white/caucasian". Similar patients have much higher survival rates.

surprisingly high (+33% chance of CC).

In the survival outcomes (Fig. 6.12-B), they noted that there was a large discrepancy between the predicted survival, and those reported by the cohort (T3.4): only 40% of similar patients survived 2 years (T1.2), and none of the treated group survived 5 years despite a predicted survival rating of 89% with high confidence (T1.1, T3.3). Interestingly, outcomes were better in the untreated group. Looking at the similar patients, we could see higher T-stage and pathological grade in the CC group, which may account for the difference, although it was unclear if race also impacted this (T1.4).

Looking back at the issue of race, the group tested this patient by changing only their race to "white/Caucasian". Indeed, this changes both the predicted treatment from the Deep Policy model (73% no cc), and the patient outcomes, with significantly higher rates of predicted survival in the similar patients (Fig. 6.12-C) (T3.2). This led to a discussion on the use of race in the model, where we discussed issues of bias and confirmed that, indeed, race has an impact on physician treatment and patient outcomes, which requires further study [231]. Interestingly, the clinician also tested the "optimal" policy model, which only showed a minimal impact of race ($< 1\%$) on the recommended treatment, which was no CC. Notably, the optimal policy model listed the low pathological grade, tumor location, and

AJCC stage as reasons to not give CC, while the LN spread is given as the primary reason to give CC. Based on the predicted outcomes, we noticed a much higher risk of side-effects (5.9% chance increase in feeding tube and 3.6% chance increase in Aspiration), with non-significantly higher predicted chance of tumor response or control, which led to the no-CC recommendation (T2.2, T2.3), as well as slightly higher incidence of severe symptoms in the symptom plot for the CC group (T1.5).

6.5.2 Counterfactual Recommendation

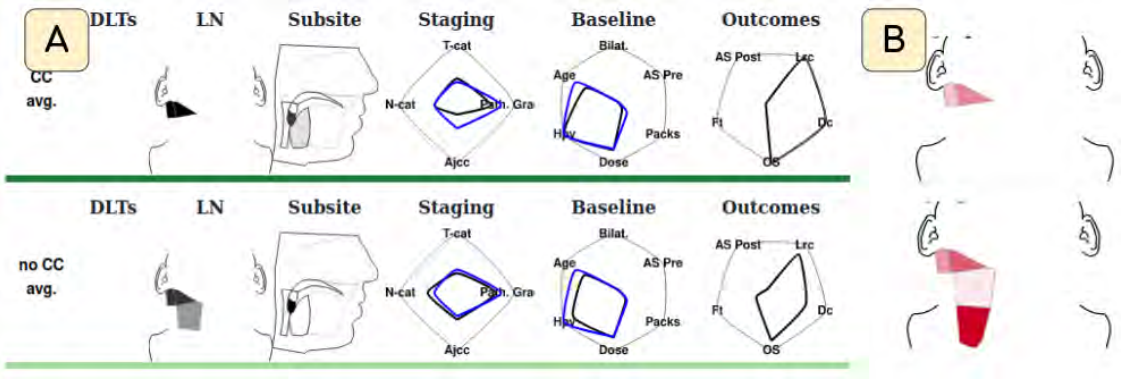


Figure 6.13: Second case study. (A) Average of treated and untreated groups. Treated patients have lower LN spread and staging, and higher survival rates, which is counter-intuitive. (B) Feature attributions for ipsilateral LN levels II (top) vs levels II-IV (bottom), where dark red encodes higher likelihood of receiving CC. Changing the patient to have LN level IV involvement significantly increases confidence that the patient should receive CC.

In this second case study, we examined a patient where the Deep Policy Model predicted no CC, while the most similar patients all received CC. In this case, the group noticed that patients had relatively low staging and low smoking, which the clinician confirmed lined up with the patient not needing CC in most cases (T1.3). They speculated the difference may be due to physician preference or other factors, such as features not accounted into the model but present in the lab notes (e.g., the patient having only one kidney). Additionally, they noted that in this case, the existing guidelines cite smoking and Lymph node levels as the main causal factors. We can see in the input LN diagram that the patient had LN levels II effects (Fig. 6.13-B, left), the most common levels, which have an impact on an increased chance of CC.

Noting in the outcome panel that there is a relatively high chance of survival for both

groups given the low risk (T1.1, T1.2), and similar risk profile for treated and untreated groups (T2.2), the clinician moved to the similar patient panel. Notably, both treated and untreated groups had similar characteristics, but the untreated group actually had more nodal extension to level 3, higher staging, a higher average smoking rate, and lower survival and tumor control after 4 years (Fig. 6.13-A) (T1.4). They noted that this may confirm that the difference in the cohort treatment may be due to the physician or other factors.

Moving to the input panel, the clinician tested the impact of the two changes given by the physician: lymph node extension to level IV, and smoking > 20 pack-years, and confirmed that with these changes the model indeed changed to predict CC (52% chance), and that LN level IV was a major factor in the change in treatment (Fig. 6.13-B) (T3.1).

6.5.3 Qualitative Feedback

Feedback from HNC clinicians at the MD Anderson Cancer Center was very positive, stating that the system was "really attractive" and "amazing". When asked about their favorite features of the interface, multiple participants stated that they liked the views of similar patients, as well as symptom progression in the auxiliary panels. They also felt that many clinicians would be more interested in just the outcomes in the center. In response to this feedback, we turned the additional panels into a separate on-demand view. The participants also found the feature attributions interesting, saying "I also like the neighborhood panel and the multiple outcomes together". Some were particularly interested in the lymph node involvement levels for similar patients, as well as how this relates to feature importance in the policy model. They were also able to identify possible sources of data bias in the predictions by looking at the treated and untreated groups. When asked about the usefulness of the simplified three-decision model, the most senior clinician commented that "The 3 decision points are critical decision points, aligned with the standard of care. We could get more granular, but it's a great start."

6.6 Discussion

Our results show that DITTO is an effective tool for treatment planning for HNC clinicians using a novel Digital Twin system. Clinician feedback was very positive, with a variety of "favorite" components and background, suggesting that DITTO can handle a variety of patient treatment goals. Additionally, while we had initial concern that the use of two model outputs would prove confusing, our case studies show that investigating model discrepancies indeed leads to interesting discussion into how patients should be treated and how physician preference or uncounted variables may impact certain recommendations.

6.6.1 Design Lessons

A majority of explainable ML work has focused on visualizations meant for model builders or clinical researchers to use in research context, while most clinician-facing systems focus on relatively simple models [62, 184, 319]. In this regard, this system is a novel attempt to deliver the results of a complex Digital Twin system to clinical end users. In particular, this work focused on two challenges: delivering many potential results to clinicians in a way that allows them in a way that is relatively accessible, and to find a way to balance encouraging oncologists to use the system while not overly relying on potentially incorrect predictions. We list here the specific design insights we've developed during this participatory design process.

L1. Use visual scaffolding. Our users were clinicians who had experience with risk modeling visualization. We found our best results by scaffolding, such as relying on temporal plots and spatial anatomical diagrams. Previous attempts at novel encodings such as histograms or unique glyphs were less successful with wider audiences.

L2. Account for different information-seeking needs. We found in our interviews and literature reviews that the degree of information seeking behavior, as well as attitudes towards different models, varied greatly between clinicians. For example, we found that some users were completely uninterested in seeing similar patients or a scatterplot of the training

cohort, while others listed the similar patients as their "favorite" part of the system and were able to identify potential confounder bias by looking at the similar patients. We also found that many users were primarily interested in seeing only the recommended treatment and time-to-survival, so this information could be communicated to patients, and felt additional features were distracting in the interface. As a result, we altered our design to afford these additional features in secondary tabs, and allowed for resizing of the different parts of the interface, while highlighting only the survival plots by default.

L3. Provide access to multiple models, and cues such as counterfactuals and confidence intervals to balance user expectations of the model. In using our system, we found that clinicians have a tendency to either fully trust or distrust a model in the absence of additional cues, and expressed a desire for "honesty" in terms of model confidence. To encourage users to "think slowly" [144, 148] about the model predictions, we relied on multiple cues: showing different model predictions side-by-side, using model confidence intervals when available, and placing the feature attribution plot prominently in the visualization. Still, there is necessarily a design tradeoff between interface simplicity, user acceptance of the model, and the number of additional cues. While many lay-users may prefer only being given a single prediction, we consider this a questionable design tradeoff with respect to XAI. As a result, we initially show users all model predictions, with the option to toggle off information.

6.6.2 Limitations and Future Work

In terms of limitations, our current models are limited by data availability and model performance. Our dataset requires modeling 19 different outcomes and transition state variables while relying on less than 600 patients from a single institution with limited demographic diversity. Furthermore, a more granular digital twin system could consider multiple rounds and dosages of chemotherapy and surgery. Our available dataset is also specific to oropharyngeal HNC patients. In terms of our interface, we focus on a limited group of HNC clinicians, a few of whom may have above-average visual literacy and information seeking behavior.

Our imitation learning model inherits existing treatment biases, and diversity shortcom-

ings in the training data. While the initial goal of the system is to reveal these biases as shown in our case studies, there is the potential for users to interpret these explanations as justification for biased reasoning when over-trusting the system. Regardless, this bias should not be reflected in the risk prediction or optimal model, which should theoretically contradict the treatment recommendation in such a case.

In terms of generalizability, the general approach can be applied to any similar treatment sequence that can be simplified into discrete decision stages, and a majority of our visualization system is domain-agnostic, except for the tumor subsite and lymph node spread diagrams. Regarding visualizations, our algorithms for uncertainty, feature attribution, and similarity are specific to deep learning classification, but these values can be obtained more generally through bootstrapping, Shapley values, and appropriate distance metrics, respectively, and can be visualized in the same way.

6.7 Conclusion

In conclusion, we have implemented a visual clinical decision support system based on a temporal deep-reinforcement learning model that is capable of simulating patient treatment outcomes. To our knowledge, this is one of the few attempts at an explainable AI focused interface for clinical users, as well as one of the first attempts at a visual interface to explore a dual digital twin system in a healthcare setting. Through our participatory design, we highlight several findings with a focus on balancing information density, usability, and encouraging appropriate trust for a variety of of end users. In our future work, we hope to evaluate this interface on a large range of clinical end users in practice, as well as extend our work to even more detailed decision-making that can consider more patient quality-of-life measures.

6.8 Chapter Conclusion

This chapter focuses on moving the target of explainable spatial models from model builders to domain subject-matter experts, which is a challenging and under-explored field of explain-

able ML. Particularly, we deal with the challenges of capturing appropriate trust, as well as communicating the results of several spatial-temporal modeling results to clinicians with varying requirements and levels of familiarity and trust with ML related systems. The work presented in this paper is part of an ongoing evaluation as the system and model are actively being deployed in practice to gather long-term feedback, which is an interesting avenue of future research.

Chapter 7

Discussion and Conclusion

7.1 Discussion

This dissertation details visual computing strategies for Visual Computer + Spatial Machine Learning applications, including task abstraction, model development, encoding design, and deployment of spatial VC+ML systems for several real-world applications. The approaches in my work aim to combine established research in machine learning, data visualization, and human-computer interaction by taking a user- and activity-centric approach to designing both the model and the interface used to explore and explain the models.

Returning to the sub-challenges, this dissertation seeks to address each of these issues:

Domain Characterization: In Chapter 2 I introduce the role of spatial tumor distributions and spatial anatomy radiation treatment planning. Chapters 4, 5 and 6 I further the role that both these anatomical tumor and radiation dose distributions have on patient outcomes. Chapter 3 discusses the role that geospatial data has in social science applications when attempting to identify the role of location on polarized political issues.

Design of Spatial ML models and Measuring Spatial Similarity: In Chapter 2 I introduce a k-nearest-neighbors model that uses a localized spatial topological similarity. In Chapters 4 and 5 I extend the concept of spatial similarity to develop clustering models with anatomical spatial inputs through a mixture of different methods, and emphasize the importance of incorporating visual steering and integrating domain knowledge into the feature selection process. In Chapter 3 I discuss generalized and linear regression models that incorporate socio-economic data about different regions in order to find associations

between users and political sentiment. Finally, in Chapter 6 I detail deep learning based reinforcement learning, classification, and deep survival models that incorporate the location and secondary spread of patient tumor location, as well as size through the use of diagnostic staging information.

Visual Encodings and Explanations of Spatial Models: In this work I detail a number of visual strategies for visually encoding spatial data. For example, stylized radiation plots Fig. 2.3 can be used to show simplified 3-dimensional distributions. When showing anatomical information where preserving depth and size is less important, 2-dimensional mappings of 3-d data can be useful for showing important information while highlighting smaller regions such as in Fig. 5.4. We can also use color gradients within heatmaps to show intra-organ distributions as shown in Fig. 4.7. When working with clinicians, many clients benefit from showing spatial visualizations of individual patients for case based reasoning, such as in Chapter 2 or Chapter 6. Finally, we can rely on combining these spatial encodings with standard feature attribution methods such as integrated gradients [293] in Chapter 6.

In addition, this work highlights some important findings for emphasizing encodings that cue users to the efficacy of individual model predictions, such as the multiple cues discussed in Chapter 6, i.e. using multiple models, flagging inputs that are outliers in the dataset using histogram distance metrics, including confidence intervals, and showing feature attributions to identify sources of bias or incorrect correlations in the model.

Measuring the impact of human-centered spatial ML systems: In this document I detail a number of different approaches for evaluating these systems in contexts where traditional empirical studies are not appropriate due to the novelty of the problems being addressed and the small target audiences. Particularly, we rely on qualitative and quantitative feedback and findings from clients through user questionnaires, case studies, interviews, and user workshops. In the case of visual steering systems focused on model building, we can also measure the value in terms of model improvement beyond models built using traditional methods. Finally, we mention the value of these systems in enabling insights that result in

publishable studies in the application domain such as in [346, 349, 350].

Overall, this work introduces several key concepts that are important when designing for spatial-temporal VC+ML in collaborative settings. First, the design of VC+ML systems needs to incorporate user needs into both the model itself, the explanations, and the interface used to interact with the model. These are often treated as separate concepts, which greatly limits the usability of many models in real settings. In our work, we show this through the use of k-nearest-neighbor models which allow us to show patients to clinicians and obtain contextual information that may be missing from the models, as well as the use of clustering that maps to the clinicians' experience using patient staging, which enables them to reason about a simplified "spatial risk" in addition to other information when making decisions.

Second, model explanation strategies need to map to existing user activities, and may be very specific to the domain application. This ties in with the proposed concept of "explanation scaffolding" in chapter 5, where explanations are usually grounded in existing domain knowledge. When designing these strategies, special considerations need to be taken for collaborative projects and users with more varied backgrounds, such as when working with clinicians, where we employed 2-dimensional heat maps to show intra-cluster distributions of spatial anatomical data, and rule mining tools to map these high risk groups to concepts similar to the thresholds used by radiologists. This can partially be handled by relying either on communication between expert groups during collaborative use of the system, or including a mixture of different explanations on-demand, as is done in chapter 7. However, these approaches come with trade-offs, as it introduces scope-creep into the design that may result in unmanageable visual complexity.

Third, incorporating domain knowledge into spatial models is a difficult task that often requires collaboration between users with domain knowledge and interfaces that are augmented with data mining and exploration systems, and incorporating pre-defined rules is often insufficient when exploring models. When relying on cohort based methods such as with KNN and clustering models, some of this issue is handled by ensuring that bounds of

the data are always seen in the cohort. However, in more complex situations such as feature selection, identifying causally-linked features and outcomes becomes more difficult, and thus this process benefits from collaborative visualization approaches in order to come to a mutual agreement on what results make sense from a data-centered and domain-centered perspective.

Finally, when developing models that are used in practice, it is important to be able to balance different qualitative and quantitative features to optimize for, rather than relying on a single metric. This generally requires human-in-the-loop interaction during the building process, along with visualizations to grasp the various important aspects of the model. For our work, we have explored models that in particular are designed to support Transparency, Actionability, Domain Sense, model Plausibility, and Trust, in addition to standard performance metrics. Specifically, we have discussed important considerations for all of these dimensions:

- **Transparency** requires that we specifically have to prioritize models that allow for introspection, which often may come into direct competition with model performance when considering, for example, a KNN and Deep learning models.
- **Actionability** requires carefully considering what we are predicting, and what we want to optimize when training the model. For example, looking purely at a single metric may be insufficient for clinical applications, where high recall may be a priority over high precision in a model that attempts to detect high risk patients.
- **Domain Sense** requires ensuring that the model and the visualizations can be scaffolded onto existing knowledge for the domain expert clients. This is particularly important for decision support tools, where we have to carefully ensure that the underlying logic of the model is reasonable to our clients.
- **Plausibility** requires a careful balance of features and outputs so that the underlying logic of the system is valid according to our end users. This ties into both Domain Sense

and Transparency, but also often requires input from model builders to understand what is being implied through our statistical insights.

- **Appropriate Trust**, which we define as when the client has an accurate mental model of the reliability of an individual model prediction, is an important aspect of how model clients interact with our models. Because of the high complexity and high risk of erroneous conclusions in many of the spatial models discussed in this work, we need to ensure that our models are persuasive, while identifying cues that can flag when a prediction might be wrong. In Chapter 4 and Chapter 5, we mention trust in terms of communicating model behavior and showing underlying model behavior in a way that allows clients to verify that the underlying logic of the system maps to their knowledge of the world, such as the function of different organs and their relationship to patient outcomes. We explore this in most detail in Chapter 6, where we propose a number of additional strategies: showing multiple model outcomes that rely on different underlying strategies, and showing confidence intervals generated through bootstrapping when showing predicted outcomes. Additionally, we incorporate outlier detection via Mahalanobis distance to flag when the input to the model may be too different from the training data for the model to have an accurate prediction.

In terms of limitations, the work described here generally relies on domain-specific designs. Despite this approach, we have distilled generalizable design lessons for visual computing applications. Additionally, since the work is largely targeted toward domain experts working on novel problems, our designs are focused on utility over general usability. Comprehensive empirical evaluations of the design choices for broad audiences are beyond the scope of this dissertation. Additionally, our datasets and evaluations are demographically limited, as our data and clients are limited to certain English-speaking demographics within the United States, and thus design choices such as choice of baseline visual literacy, semantic color associations, and treatment options may differ when applied to different areas. Finally, in terms of spatiality, all of our work focuses on data that can be aggregated within geometric

regions, such as discrete organ volumes, and thus additional considerations would need to be accounted for when applying any of these methods to continuous field data.

Potential future work could include further work in validating explainable decision support systems in the field. This work has explored the preliminary stages and formative feedback, but work in gathering results from a wider audience with repeated interactions with the system are ongoing. Additional extensions beyond this dissertation may include visualizations that examine patient outcomes at a more granular level. For example, extending the spatial work to consider voxel-level distributions of radiation dosage within organs for identifying outcomes, where convolutional networks could be used alongside human guidance to train meaningful embeddings of patient distributions. Additionally, our collaborators have been working on obtaining more granular information on treatment regimes and patient symptoms, which could be used to support more specific treatment planning while considering the quality of life for patients beyond survival and hospitalization.

In conclusion, the design of useful explainable models is a developing topic that is only narrowly explored, despite a recent boom in interest among certain academic circles. This thesis focuses on how we can incorporate XAI approaches alongside human-computer-interaction design to create better, more useful machine models by incorporating spatial-temporal trends and domain knowledge. However, it is clear that overall, the needs of VC+ML systems require domain-specificity. While this work is limited to two general domains: Head and Neck Cancer and social media analysis, there is a wealth of other areas and models that need to be considered in the future.

Chapter 8

Appendices

8.1 Appendix A: Chapter 3 (MOTIV) detailed user feedback

Qualitative Feedback

User Questionnaire

Likert scale (1-5) questions:

1. Considering the MOTIV interface, how useful is the Moral Frame Summarization panel, indicated by (A) above? (e.g., did it help you tell which MFs were most common? Whether most tweets were in favor of or against SAH? etc.)
2. Considering the MOTIV interface above, how useful is the Inference (Regression Plots) Panel (B)? (e.g., did it help verify what features were correlated, like Mask Usage and Care?)
3. Considering the MOTIV interface above, how useful is the Tweet Timeline panel (C)? (e.g., did it help you identify the most popular tweets? did it help you identify spikes in the # of tweets? etc.)
4. Considering the MOTIV interface above, how useful is the Map Panel (D)? (e.g., did it help you see the spatial distribution of tweets for a specific frame?)
5. Considering the MOTIV interface above, how useful was the fact that the different panels were coordinated via interaction and color? (e.g., did it help you connect when forSAH tweets were most present and where they were coming from?)
6. (T1) How useful was the system for identifying the most popular frames? (e.g., did it help you see Care was popular?)
7. (T2) How useful was the system for identifying relevant/present tweet features? (e.g., did it help you see that vividness was not particularly present? could you tell there was a mix of sentiments?)
8. (T3) How useful was the system for identifying geo-political or demographic trends? (e.g., did it help you see most of the tweets in the corpus were in urban areas?)
9. (T4) How useful was the system for identifying trends over time? (e.g., did it help you see there were several spikes in the timeline, and when?)

In what ways was MOTIV helpful to you? Check as many boxes as you wish (Yes/No):

- Helped me understand the corpus we were collecting/generating.
- Helped me understand which tweet features were present and which were not (e.g., vividness, virality).
- Helped me understand the moral frame distribution across the corpus.
- Helped me understand the temporal distribution of the tweets.
- Helped me understand the geographical distribution of tweets.
- Helped me understand the political context of the moral frame distribution.
- Helped me correlate the tweet sentiment with the COVID-19 cases over time.
- Helped me verify hypotheses about meaningful relationships in the data.

Open-ended:

- Any suggestions for further improvement?
- Any other feedback?

Questionnaire Results

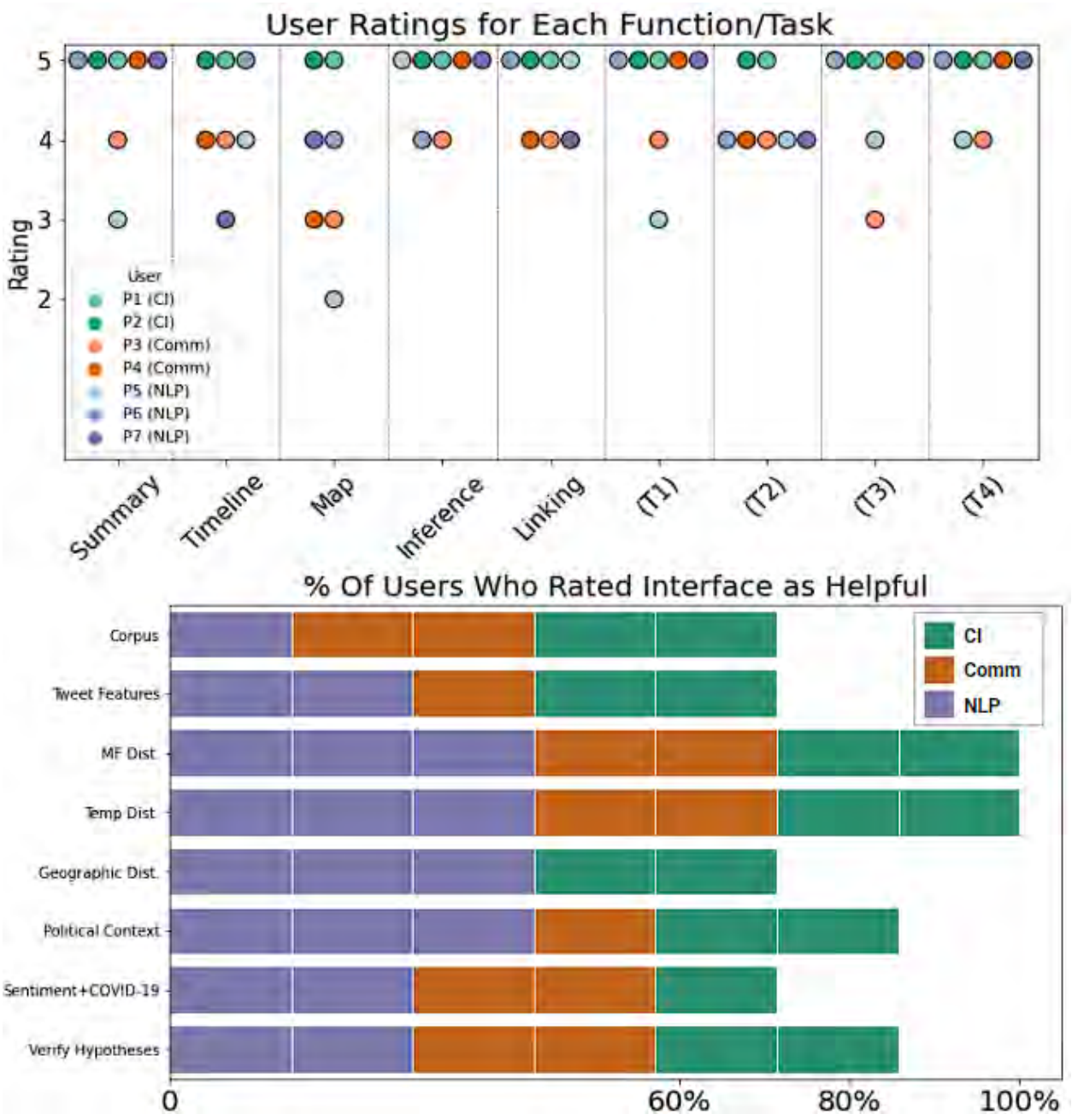
Likert Scale Questions

User	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9
P1 (NLP)	3	4	2	5	5	3	4	4	4
P2 (CI)	5	5	5	5	5	5	5	5	5
P3 (CI)	5	5	5	5	5	5	5	5	5
P4 (NLP)	5	5	4	4	5	5	4	5	5
P5 (Comm)	4	4	3	4	4	4	4	3	4
P6 (NLP)	5	3	4	5	4	5	4	5	5
P7 (Comm)	5	4	3	5	4	5	4	5	5

User	In what ways was MOTIV helpful to you? Check as many boxes as you wish.
P1 (NLP)	Helped me understand the corpus we were collecting/generating, Helped me understand which tweet features were present and which not (e.g., vividness, virality), Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me understand the geographical distribution of the tweets, Helped me understand the political context of the moral frame distribution

P2 (CI)	Helped me understand the corpus we were collecting/generating, Helped me understand which tweet features were present and which not (e.g., vividness, virality), Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me correlate the tweet sentiment with the Corona cases over time, Helped me understand the geographical distribution of the tweets, Helped me understand the political context of the moral frame distribution, Helped me verify hypotheses about meaningful relationships in the data
P3 (CI)	Helped me understand the corpus we were collecting/generating, Helped me understand which tweet features were present and which not (e.g., vividness, virality), Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me understand the geographical distribution of the tweets, Helped me understand the political context of the moral frame distribution, Helped me verify hypotheses about meaningful relationships in the data
P4 (NLP)	Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me correlate the tweet sentiment with the Corona cases over time, Helped me understand the geographical distribution of the tweets, Helped me understand the political context of the moral frame distribution, Helped me verify hypotheses about meaningful relationships in the data
P5 (Comm)	Helped me understand the corpus we were collecting/generating, Helped me understand which tweet features were present and which not (e.g., vividness, virality), Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me correlate the tweet sentiment with the Corona cases over time, Helped me verify hypotheses about meaningful relationships in the data
P6 (NLP)	Helped me understand which tweet features were present and which not (e.g., vividness,

	virality), Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me correlate the tweet sentiment with the Corona cases over time, Helped me understand the geographical distribution of the tweets, Helped me understand the political context of the moral frame distribution, Helped me verify hypotheses about meaningful relationships in the data
P7 (Comm)	Helped me understand the corpus we were collecting/generating, Helped me understand the moral frame distribution across the corpus, Helped me understand the temporal distribution of the tweets, Helped me correlate the tweet sentiment with the Corona cases over time, Helped me understand the political context of the moral frame distribution, Helped me verify hypotheses about meaningful relationships in the data



Any suggestions for further improvement?

- In panel B, it was not clear what are x-axis and y-axis. On top of the plot, you mentioned Covid cases vs Care. Does the y-axis show Care? If so, what does this number show about Care?
- [The map] was the most difficult to decipher, in my opinion. I'm not sure what exactly would make the map a little easier on the eyes. Perhaps some additional or clearer labelling in the key, to denote what the difference in the colors or patterns are, as well as what the difference in size or shape of the dots on the map mean. I do like that when you

hover on the dots, it gives you the specific information like the number of tweets and the demographic percentage.
- I think I would have been quite lost without specific examples of how the graphics would be useful (ie the "e.g." portion like "Considering the MOTIV interface above, how useful is the Map Panel (D)? (e.g., did it help you see the spatial distribution of tweets for a specific frame?))

Any other feedback

- The visualizations of the dataset are very helpful. However, the content of the interface is dense and one can miss some important parameters (For example, in the Timeline panel I did not notice the average sentiments visualization initially)
- This is some excellent work and generates great insights from the data
- n/a

8.2 Appendix B: Chapter 3 (MOTIV) extended case studies

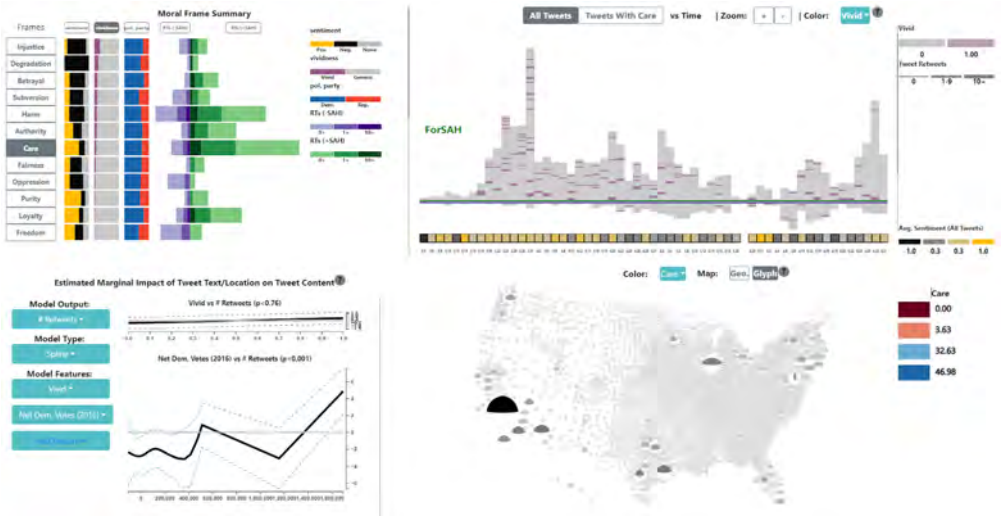
Case Studies

Here we include additional case studies using our system, as well as our original case studies with additional figures, which were removed from the main paper due to space constraints.

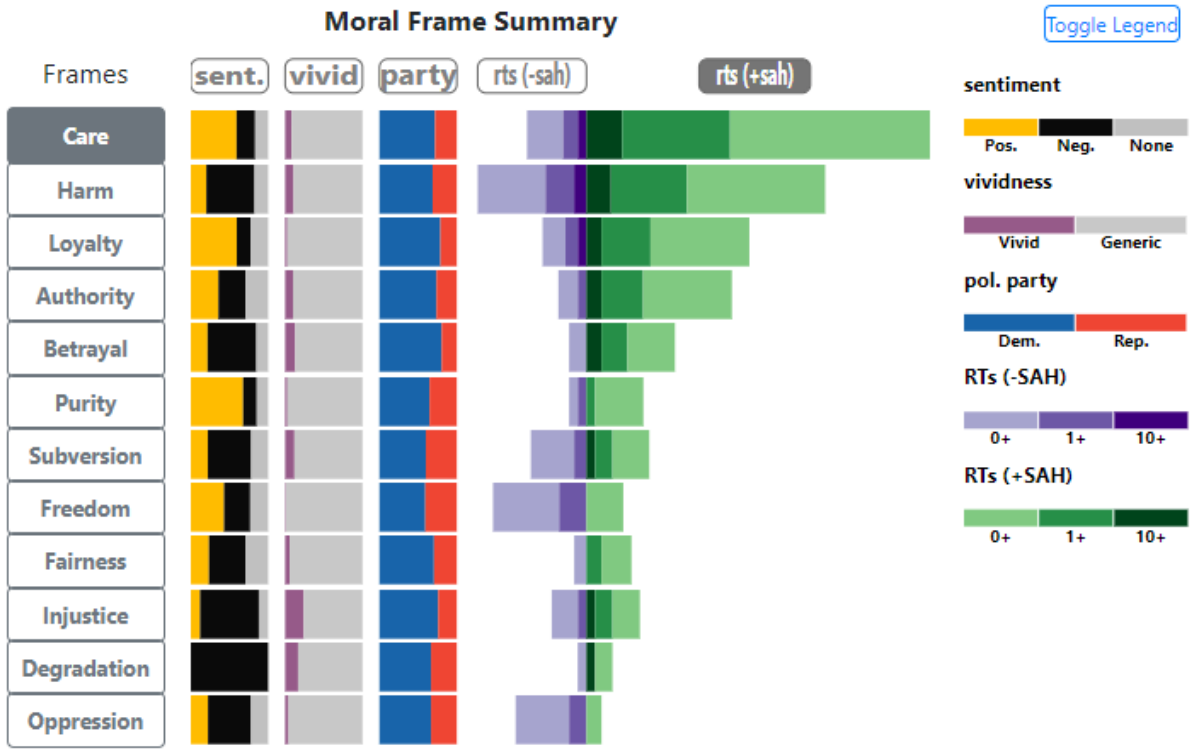
SAH Attitudes (Extended)

This exploratory case study focused on an analysis of Moral Frames as expressed in microblog data related to Stay at Home (SAH) orders in the U.S. Our collaborators were interested in which frames were dominant in the microblog data, and their vividness, popularity, sentiment, what temporal trends they followed, and the surrounding socioeconomic context around the tweets expressing each frame.

The investigation started by inspecting the distribution of the features in the data to assess for issues and biases in the data collection and labeling [WF 0]. Looking at the moral frame summary view, we noticed that there was a low number of viral tweets with 100+ tweets, as well as few tweets that were considered “vivid” - tweets that reference personal stories. We found that few frames had more than 15% vivid tweets, the highest being injustice with 6 out of 25 tweets being vivid. Investigating the inference view, we found a non-significant positive correlation ($p > .5$) between vividness and retweet count. When investigating the distribution of tweets, we found that a majority of tweets across most frames were localized to larger cities such as LA, Chicago, and regions around Houston, New York City, and Florida. Few tweets originated from rural areas, suggesting that future work should focus on getting a more comprehensive range of tweets from more diverse areas.

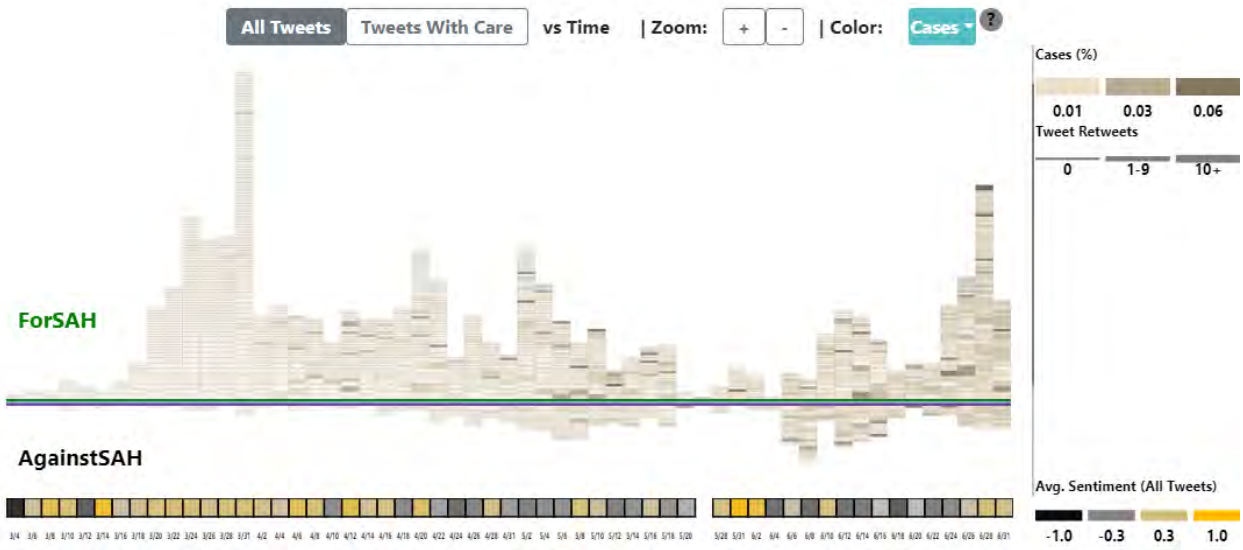


We then shifted our focus to analyzing the dominant moral frames. This investigation started by looking at the Moral Frame overview, by sorting the most popular frames [WF 1]. It became apparent that Care and Harm are the most popular frames expressed in Stay at Home tweets, and that they are both, surprisingly, predominantly in support of SAH orders. The communications experts noted that Care and Harm are complementary frames that form the virtue and vice around a single Moral Frame, respectively, so this finding was intriguing. The group then noted that all “virtues” such as Care were correlated with higher sentiment than all “vices", such as Harm [WF 1].



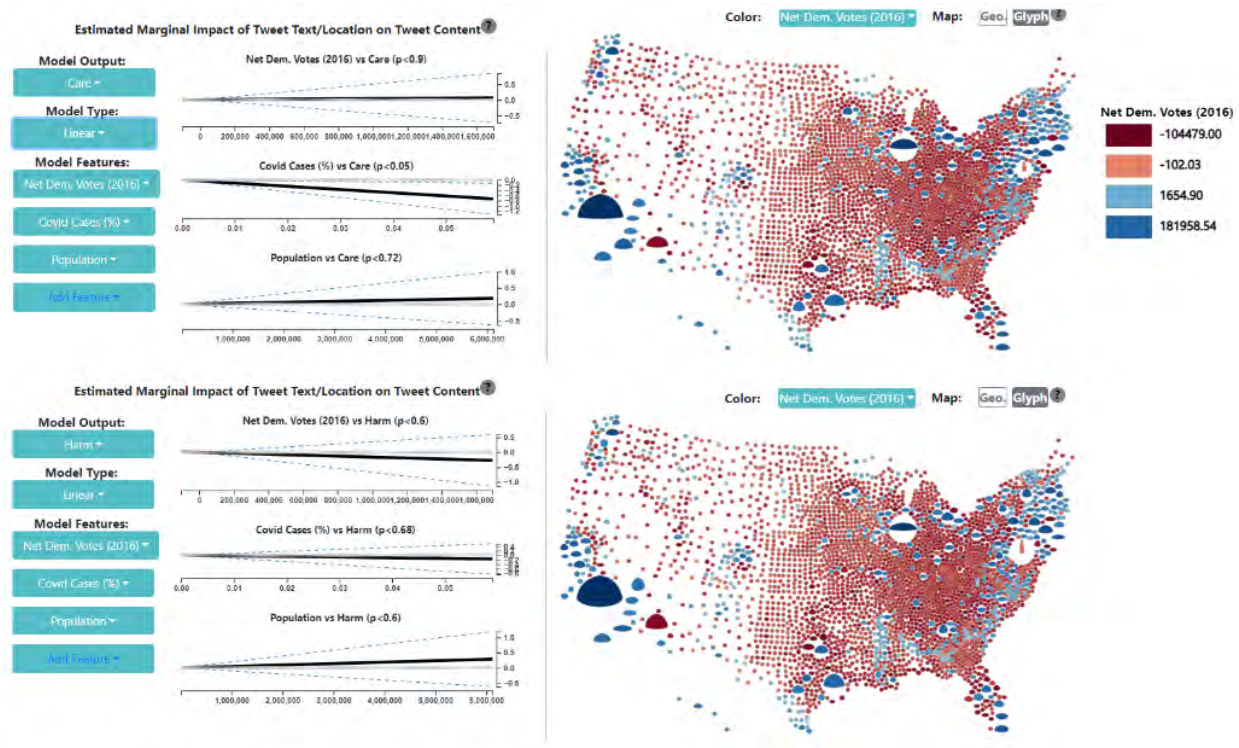
The group was then extremely surprised to note that, aside from Freedom and Oppression, most other frames were also in support of SAH orders [WF 1]. These other frames were being expressed predominantly in democratic counties—even frames typically associated with conservative views, like Loyalty and Betrayal. They were further surprised to notice, by interacting with the Summarization panel, the relatively low popularity of the tweets and a general lack of vividness [WF 1]. Upon inspecting the timeline view, the group was able to confirm that most tweets are in support of SAH (predominantly above the centerline), and most tweets have low popularity (short tiles). In addition, they noted a correlation with increasing COVID-19 case numbers (redder tile shade), and overall more negative sentiment (more gray and black in the sentiment bar) as the pandemic evolves. By further examining individual tweets, they were able to determine that the few viral tweets were, as expected, also vivid (e.g., \emph{"Protesters attacking governors for stay at home orders. Claim it infringes upon their rights. Know what else infringes upon your rights? DEATH."}). Several other popular tweets reflected counter-intuitive information (e.g., the news that most of the NYC new COVID-19

cases were people following SAH orders), influencer (e.g., Elon Musk) SAH tweets, or, again, vivid pleas from overwhelmed nurses and doctors working in intensive care units [WF 1].



A visual computing researcher then noticed in the Timeline panel several spikes in the number of SAH tweets on March 31st, May 2nd, and July 28th, and a significant and surprising drop around May 28th [WF 1]. This sparked a vivid discussion involving the county map. Communications experts inferred that the peaks corresponded to the beginning and end of several regional lockdowns, whereas the drop corresponded to the onset of social unrest related to the George Floyd events and Black Lives Matter (BLM) movement in the US [WF 1].

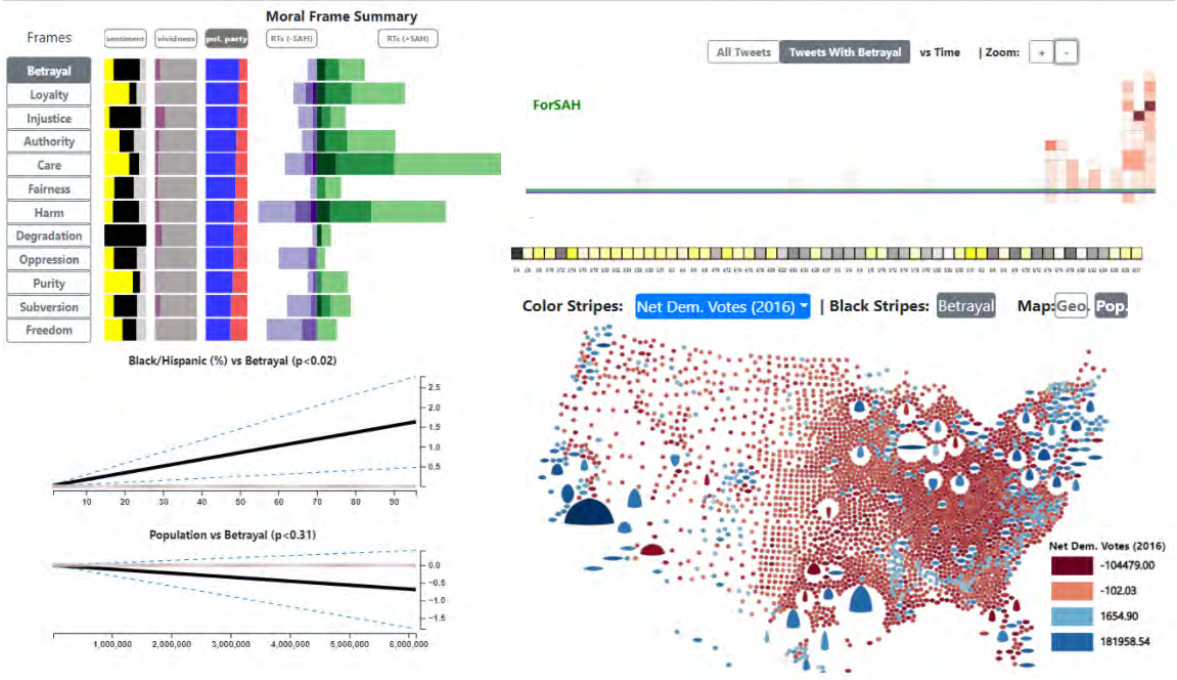
Based on the same Timeline panel, the group noticed the first wave of anti-quarantine tweets [WF 1], which, upon inspection in the Geospatial panel, appear to originate in counties with lower COVID-19 rates [WF 1]. Brushing the area around Los Angeles in the county map, we noticed suburban counties had a higher Harm/Care tweet ratio [WF 1]. The most senior communications expert hypothesized that tweets about Care originate mostly from large cities, whereas Harm is more evenly distributed about different suburban or rural populations [WF 1]. The group tested this hypothesis in the inference plot by showing the relationship between the population and each frame in the Inference panel [WF 2]. Comparing both frames, the group found that Harm is indeed more prevalent in lower-population counties than Care [WF 2].



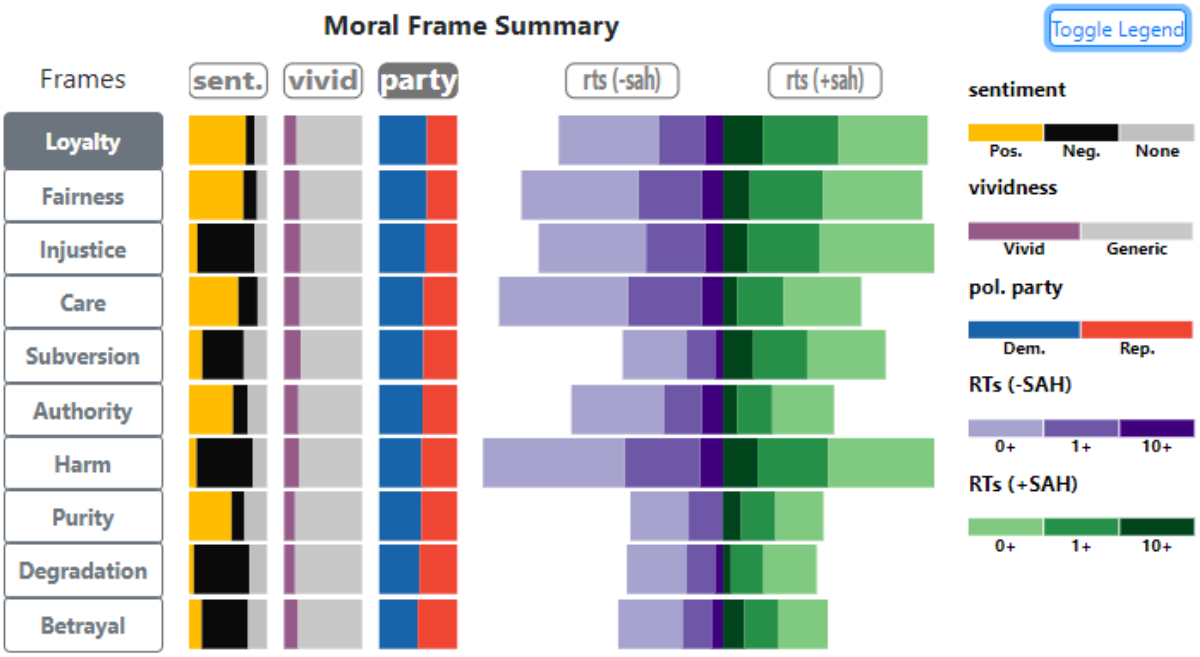
The group concluded that the microblog discourse was generally in favor of SAH orders, with increasing negative sentiment as pandemic fatigue set in. Although Care was predominant, most of the other frames expressed were also overall in support of SAH, with several interesting anomalies. They also noted the data was biased toward urban areas.

Near the end of May, the BLM rhetoric had nearly supplanted the SAH discourse, despite pandemic fatigue and an expectation of increasing conservative views. They concluded that the public policy messaging which had targeted Care-for-others had been overall effective [WF 2].

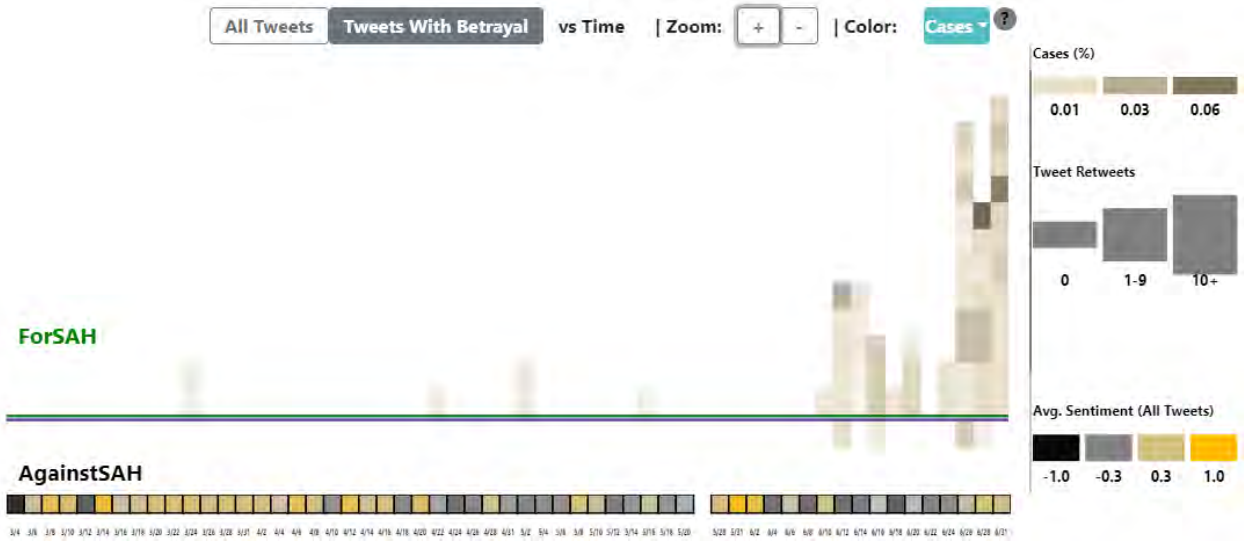
Political Association of SAH and Moral Frames



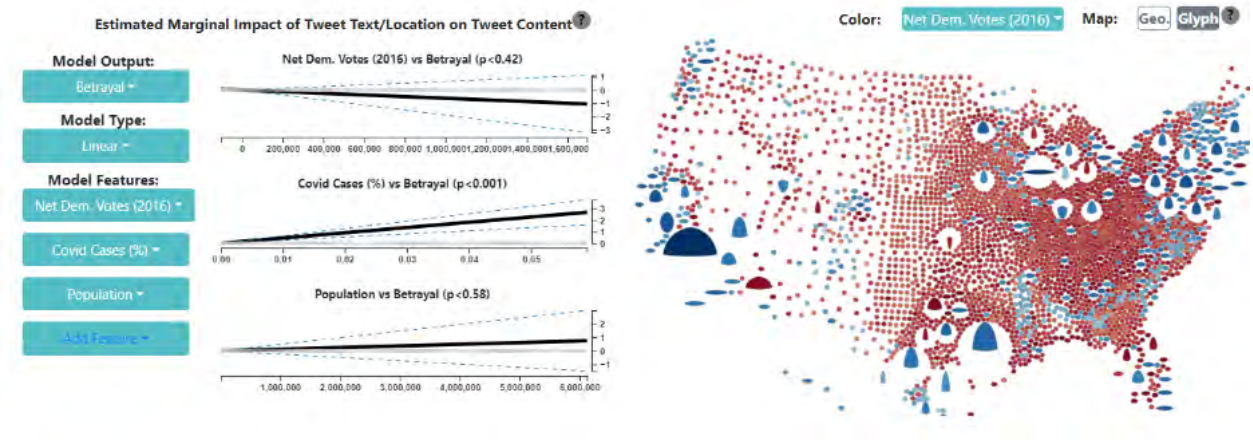
This case study examined how regional politics may affect moral stance and SAH discourse. Previous research has suggested that Care, Harm, and Fairness are more strongly valued by liberals. To test this hypothesis, the group started by sorting the Moral Frames in the Summarization panel by the percentage of tweets from Democratic counties [WF 1]. Unintuitively, the Betrayal frame rose to the top.



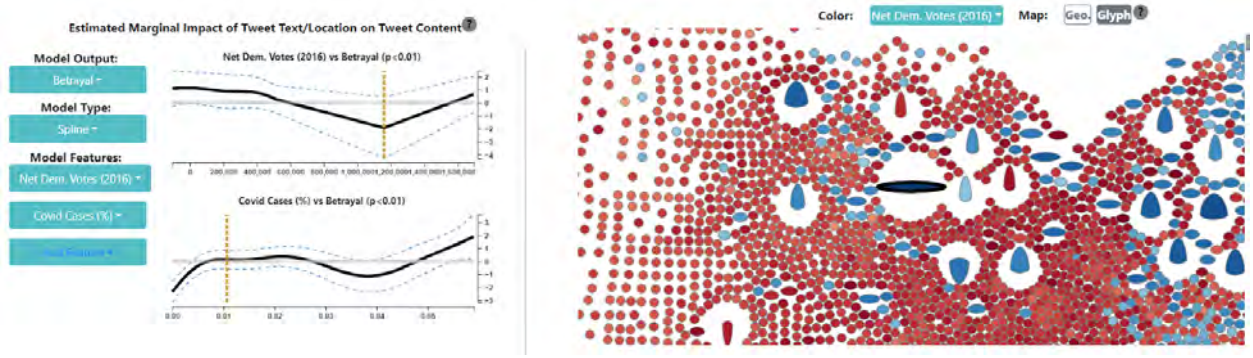
Closer inspection via filtering the Timeline panel [WF 1] revealed that all the tweets with the Betrayal frame rise close to the end of the timeline, right after the George Floyd protests in May, thus explaining the Democratic lean [WF 2]. The social scientists found it interesting that Betrayal (a moral judgment) did not seem in this case to carry Anger (an emotion), at least based on the viral spread: "*Anger is a typical driver of viral spread, and we don't see that here. [...] Were people upset about the BLM demonstrators breaking lockdowns?*", and then after a back-and-forth with the Timeline panel:"*Yeah, there's no correlation between Sentiment and Retweets. Interesting.*" The social scientists theorized then that the Betrayal frame was associated with tweets in major cities, which were known to be Democratic, around the time when lockdowns were being lifted [WF 1].



The group sought to confirm this theory by looking at the Inference panel, where they were surprised to see that net democratic votes and population are not correlated with tweets expressing Betrayal, but COVID-19 rates are [WF 2]. This finding suggested that these tweets may be a reaction to local policies after COVID-19 spikes. This observation led the group to inspect the map of net democratic votes vs. Betrayal tweets, where they saw hotspots in cities in Texas and Florida, as well as Democratic states that lifted lockdowns at this time. This suggests that Betrayal is expressed in overall democratic cities (blue) that are located *within* states that had less strict SAH orders (red), which was indeed a prominent issue in conservative states like Florida and Texas at the time [WF 2].



By examining the Inference panel for democratic votes vs. Betrayal, the most senior social scientist noticed a dip in the plot ("*Huh. There's a dip in Betrayal*."), then moved to the geomap, where they inspected abnormal glyphs, to see which correlated with the dip in the Inference panel. "I*t's Chicago [shows the dotted line lining up with the dip]. There are no tweets about Betrayal in Chicago... Betrayal is about Loyalty, and it's typically associated with Conservative sentiment. Chicago is not conservative, so the dip makes sense? [...] When we look at LA, which is also blue and democratic, well, they reopened in June 2020 and relaxed some rules, and people were not happy.*" We then theorized that the variance in betrayal response may be due to differences in quarantine policies at the beginning of June [WF 2].

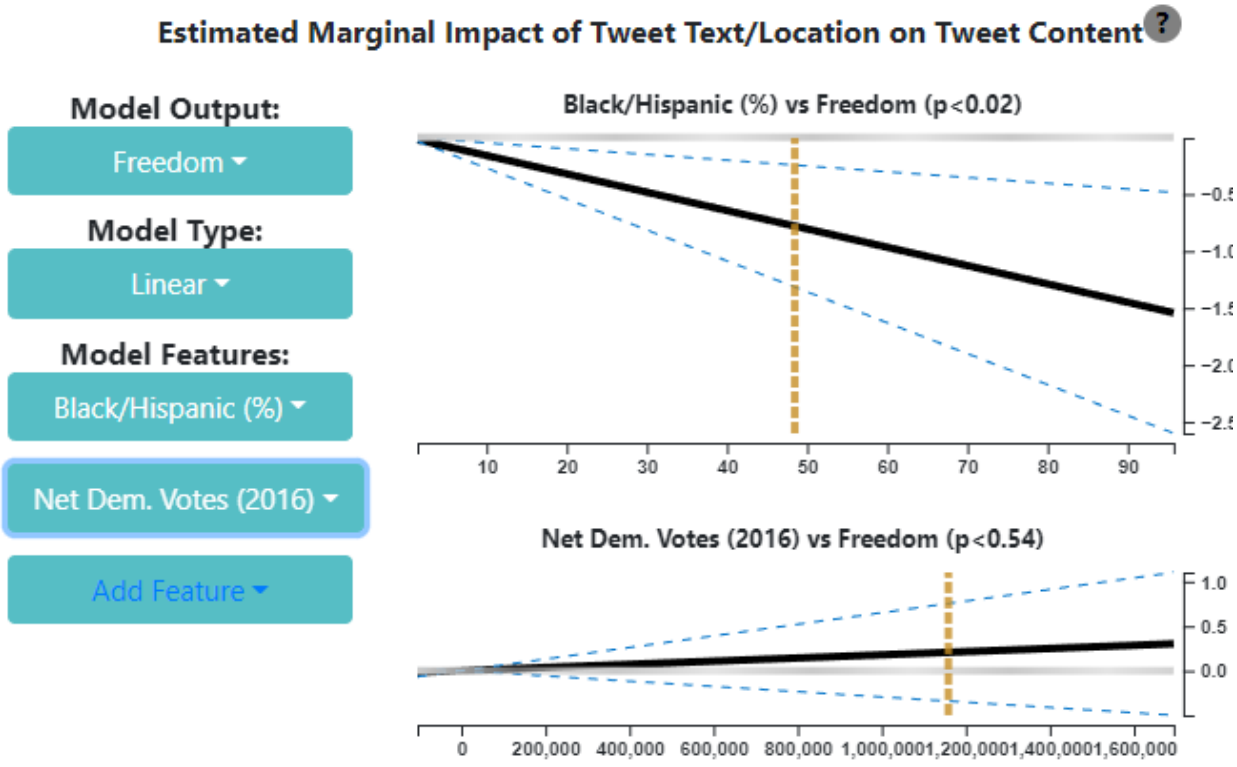


During a quick back and forth between the overview and the timeline, the group examined other frames: "*Look at Freedom—the mood tanks later in time. People must be tired of the pandemic.*"; "*Purity and Degradation, it's really surprising how they don't work in concert. Purity tends to be more about religion, and Degradation shows more negative sentiment, maybe that's why*" [WF 2]. "*Loyalty and Authority have similar patterns. But Authority is mostly at the start of the pandemic. Loyalty is expressed more after George Floyd. Interesting.*"

Returning to the Freedom frame, the group then focused on the Inference panel, where they examined the relationships between mask usage and Freedom, and found that mask usage was significantly lower in areas that tweeted about Freedom. Exploring other Moral Frames, they noticed that this was unique to the Freedom frame, and mask usage was instead positively

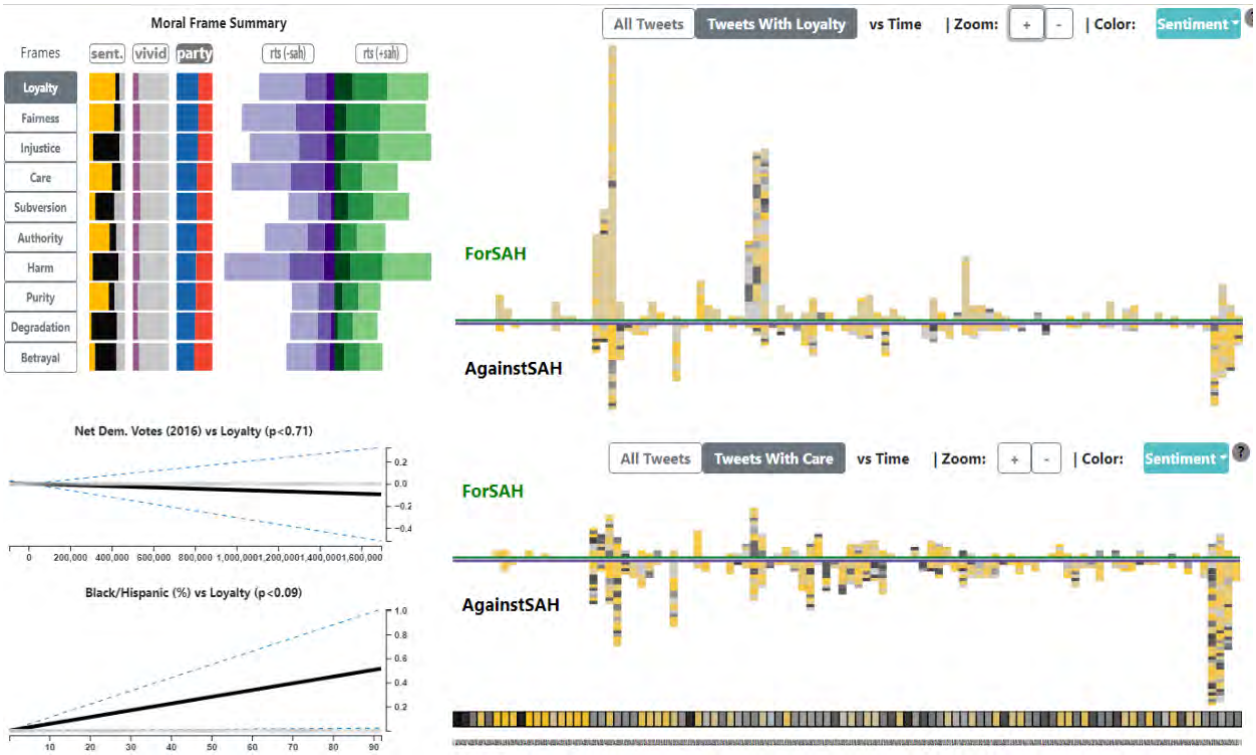
correlated with Care. “*Freedom is a controversial frame, it's more Libertarian than Republican. Libertarians believe in individual freedom, so they are not as opposed to abortion, yet are vaccine-opposed. They occupy a different political space than Democrats and Republicans, they just behave differently. [...] That's a good point. I think you're onto something here*.”

With a back and forth with the Summarization panel and the Geo-map. “*It's strange, Freedom correlates with whites, but not republicans. I did not expect that.*” After further discussion of the political spectrum, the group concluded that finer differentiation between Libertarians and the two dominant parties would be useful [WF 2].



The case study concluded that Moral Frames typically associated with conservative values were, in the case of SAH orders, still expressed in democratic areas. In other cases, the group concluded that finer political differentiation between traditional conservatives and Libertarians associated with the republican party in the US was needed. These findings reinforce the value of public policy SAH messaging targeting preferentially a more positive frame like Care.

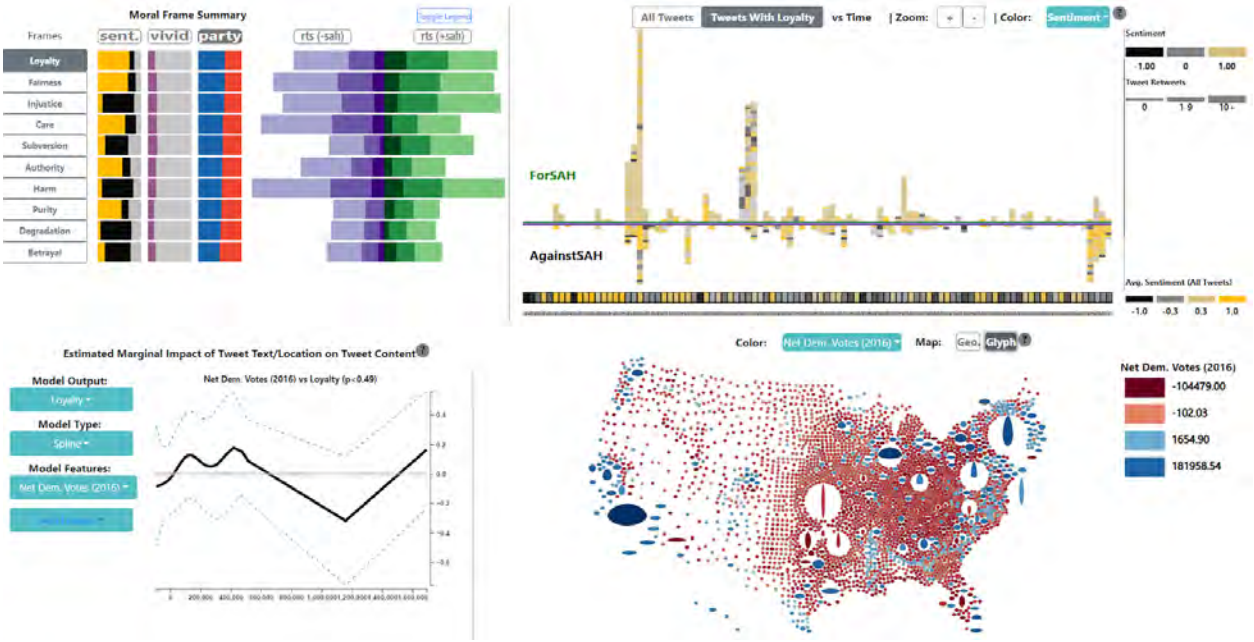
Black Lives Matter (Extended)



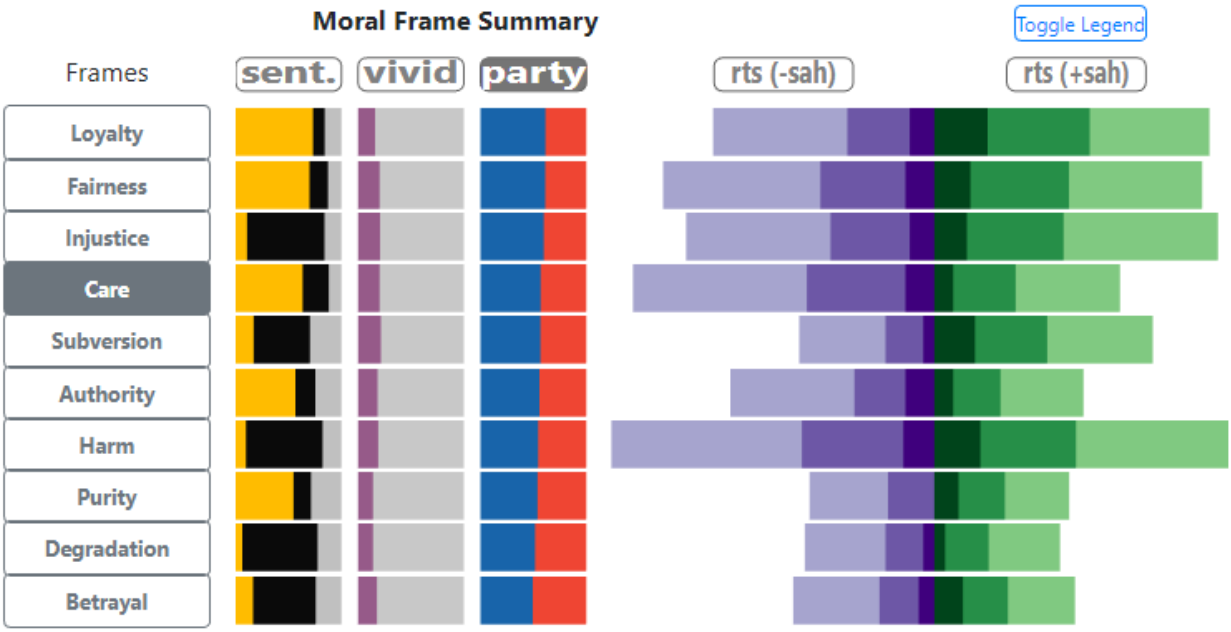
This case study uses a subset of the Moral Foundations Twitter corpus to compare tweets associated with the #BlackLivesMatter (BLM) movement and the #AllLivesMatter (ALM) movement between 2014 and 2016. The #BlackLivesMatter movement is a social movement that gained widespread popularity in 2014 in response to the disproportionate violence against African Americans, particularly by the police. The #AllLivesMatter movement, among other movements, arose as a critical response to the BLM movement. Both movements have become central to political discussions in the United States around issues such as police protections and criminal justice reforms and played a role in the 2016 presidential election. Understanding the Moral Framework behind both movements can give insight into the driving forces behind these political movements.

We consider tweets that contain more hashtags connected to BLM to be in support, while tweets that contain more hashtags related to the ALM movement are opposed. Tweets that contained an equal number of hashtags related to each movement were excluded. Vividness was annotated using a convolutional neural network that was trained on tweets from the hand-annotated stay-at-home tweet corpus. In total, we identified 1051 tweets in support of the BLM movement and 854 tweets in support of the ALM movement.

We started our investigation by looking at the Frame Summary View and sorting Moral Frames by political party [WF 1]. We could see that the frames most strongly associated with democratic areas are Loyalty, Fairness, and Injustice. In contrast, Betrayal and Degradation are most often associated with more negative sentiment and republican areas [WF 1].

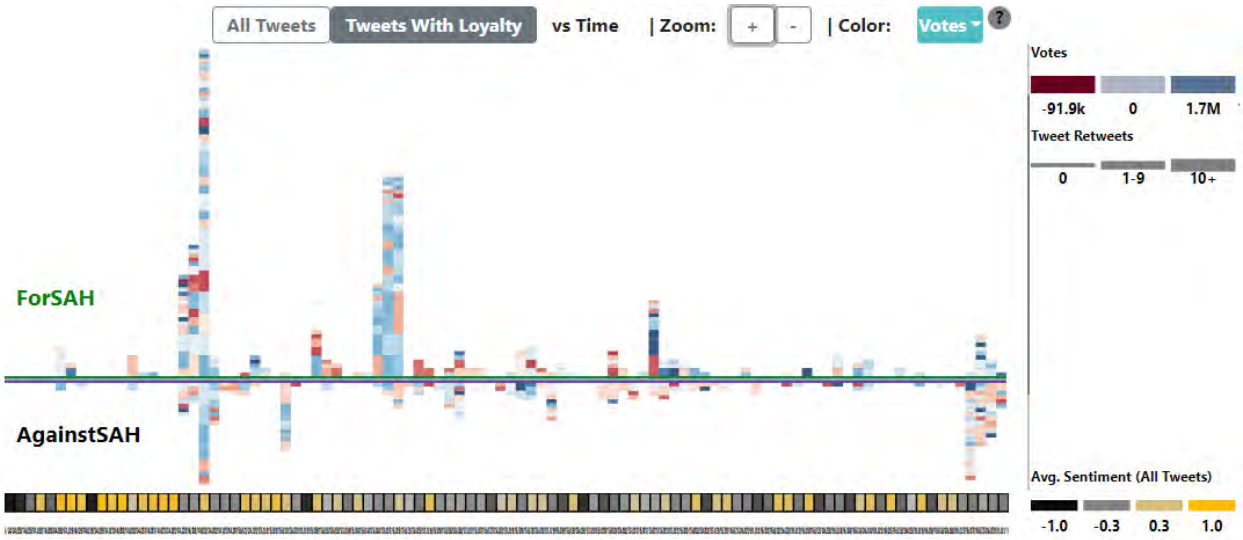


We also note that despite being relatively balanced politically, a majority of tweets that express Care are in support of ALM, which is unexpected given that prior literature suggests that Care is more strongly associated with political liberals, as is the BLM movement.

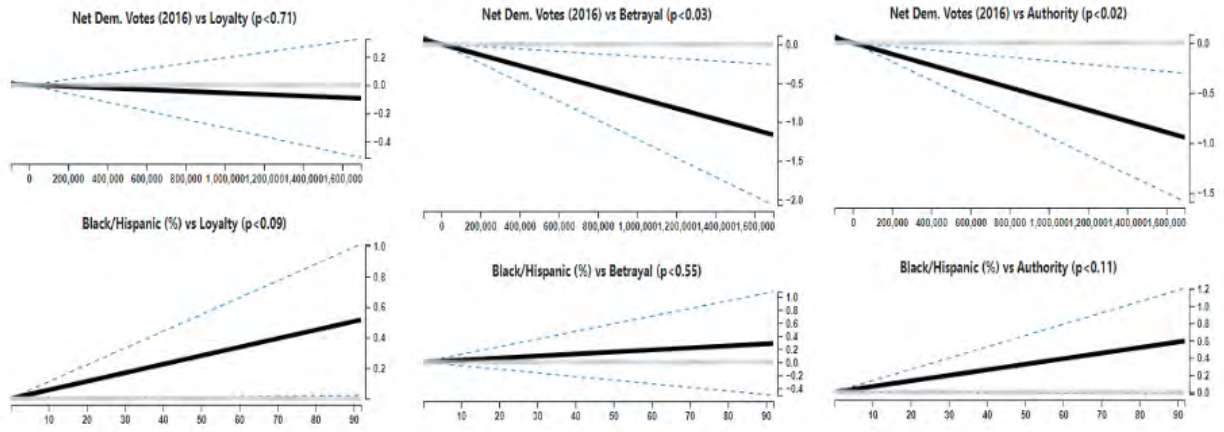


One researcher mentioned that the the literature points towards Injustice/Fairness being strongly associated with political liberals, but we did expect that Loyalty would be strongly correlated with pro-BLM tweets since it is a “Binding Frame”, and decided to explore further by viewing pro-Loyalty tweets in the timeline view [WF 1]. We can see 4 major spikes in activity, 3 of which are predominantly for BLM and from relatively democratic areas, while one is for ALM

with a higher percentage of Republican areas. Investigating the popular tweets from these time periods revealed the context behind these tweets: they are all tweets expressing solidarity for major protests: The Ferguson Protests, the 2015 Baltimore Protests, the 2015 Mizzou Protests, and the 2016 Dallas Protests in which 5 police officers were murdered [WF 2].

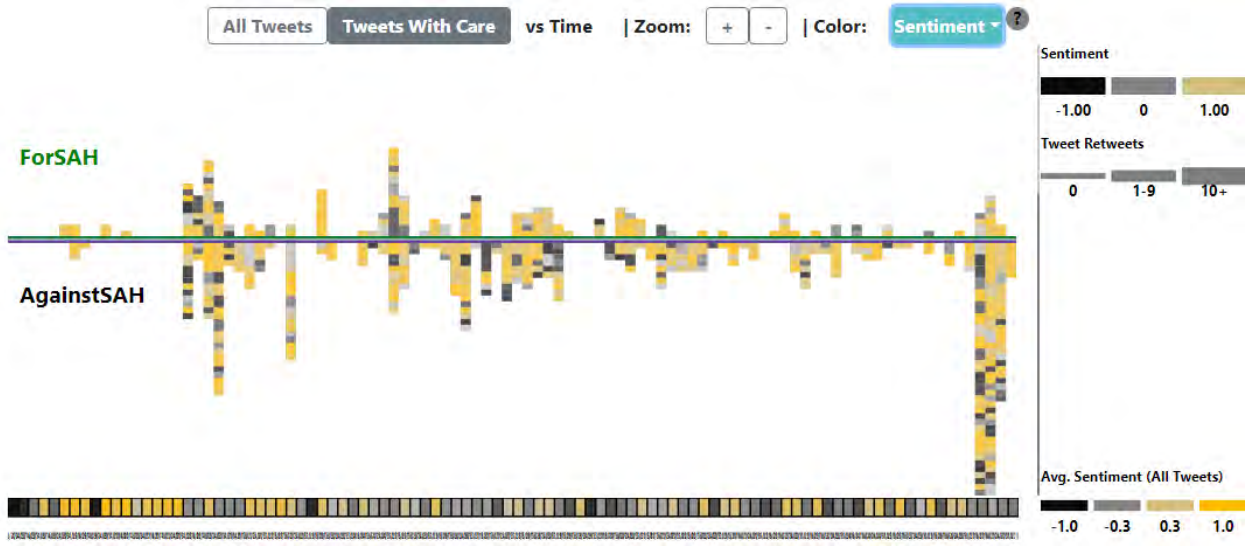


Looking at the Inference view, we theorize that Loyalty may be more related to race than political party [WF 1]. Comparing both political party and race, we see that political association with Loyalty and Fairness are largely accounted for by the portion of Black/Hispanic individuals. In contrast, appeals to Authority and Betrayal are more strongly associated with regions that are predominantly republican [WF 2].

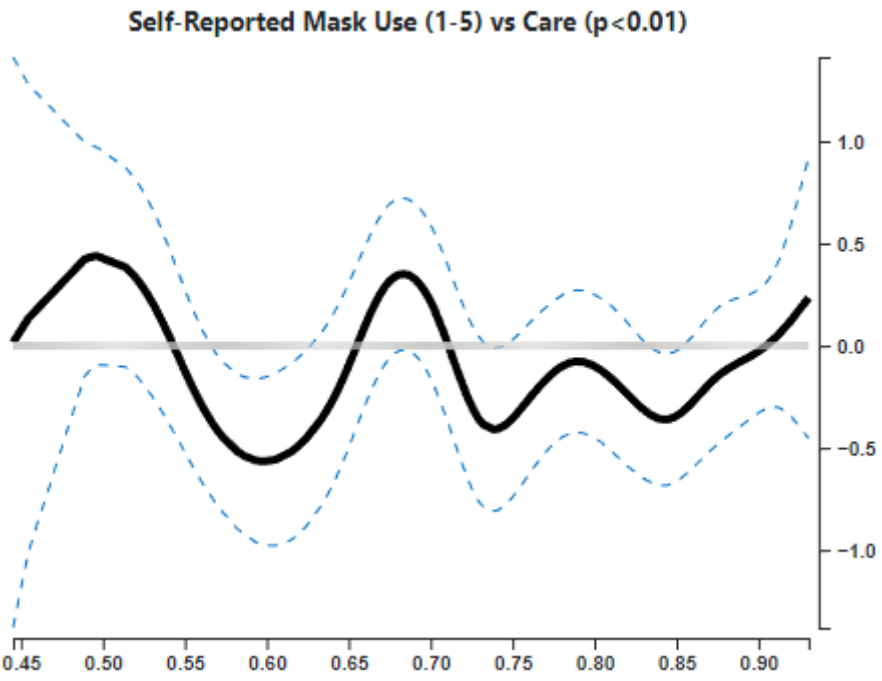


Given the association between Care, political liberals, and SAH attitudes in our prior case studies, one researcher expressed interest in the fact that Care was not related to pro-BLM tweets “Care shows up in Republican areas, that’s strange”. In the timeline, we see small spikes in activity around the Ferguson, Baltimore, and Dallas Protests. However, a visual computing researcher quickly noticed a large spike in tweets around the Dallas protest that are for ALM

"*Oh, I see… Cops were killed in the protest. These people care for the cops ("blue lives") who were killed.*" [WF 2]. We also find that despite Care being associated with positive sentiment in the Covid dataset, a large number of Care tweets express negative sentiment in the BLM dataset. By inspecting popular tweets, researchers noted that this may be a result of a large number of tweets discussing care in terms of reduction of harm, e.g. "*Rip to anyone killed from other people's ignorance…*", while most tweets with positive language are limited to solidarity with protestors.



Comparing these findings to mask usage during the pandemic, we see Care in the BLM dataset is not correlated with mask usage during COVID-19: "*Care and Mask Use are [not] correlated. That is counter-intuitive*", despite a strong association between Care and mask use in the SAH dataset [WF 1]. Our collaborators theorized that this may reflect a shift in moral sentiment in partially republican areas between 2016 and after the 2020 pandemic, and a shift in priorities of the GOP rhetoric towards more Libertarian Rhetoric and away from care [WF 2].


Self-Reported Mask Use (1-5) vs Care (p<0.01)
1.0
0.5
0.0
−0.5
−1.0
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90

8.3 Appendix C: Chapter 3 (MOTIV) Prototypes

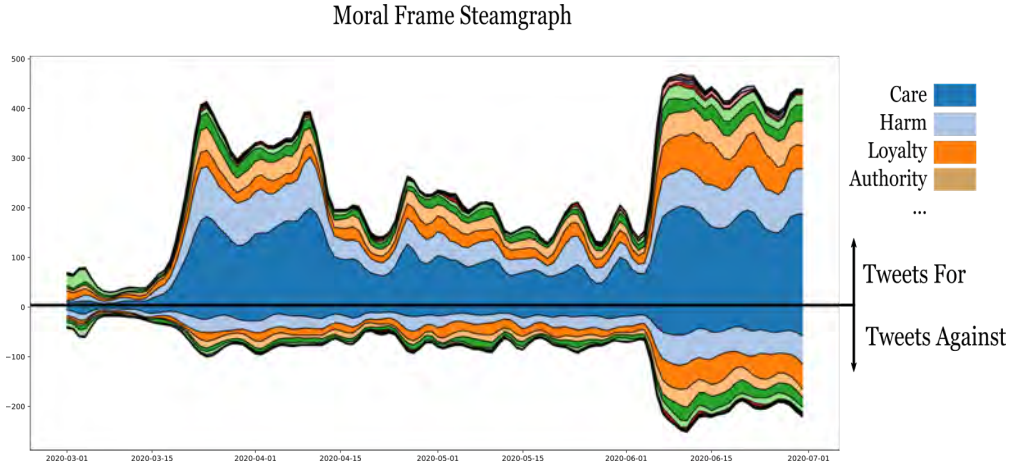
Prototype Designs

Temporal Designs

These temporal designs are alternatives that were explored while doing analysis with a larger set of tweets that did not contain geolocation information. These focus on scalability at the cost of allowing for interactive analysis of individual text or covid rates.

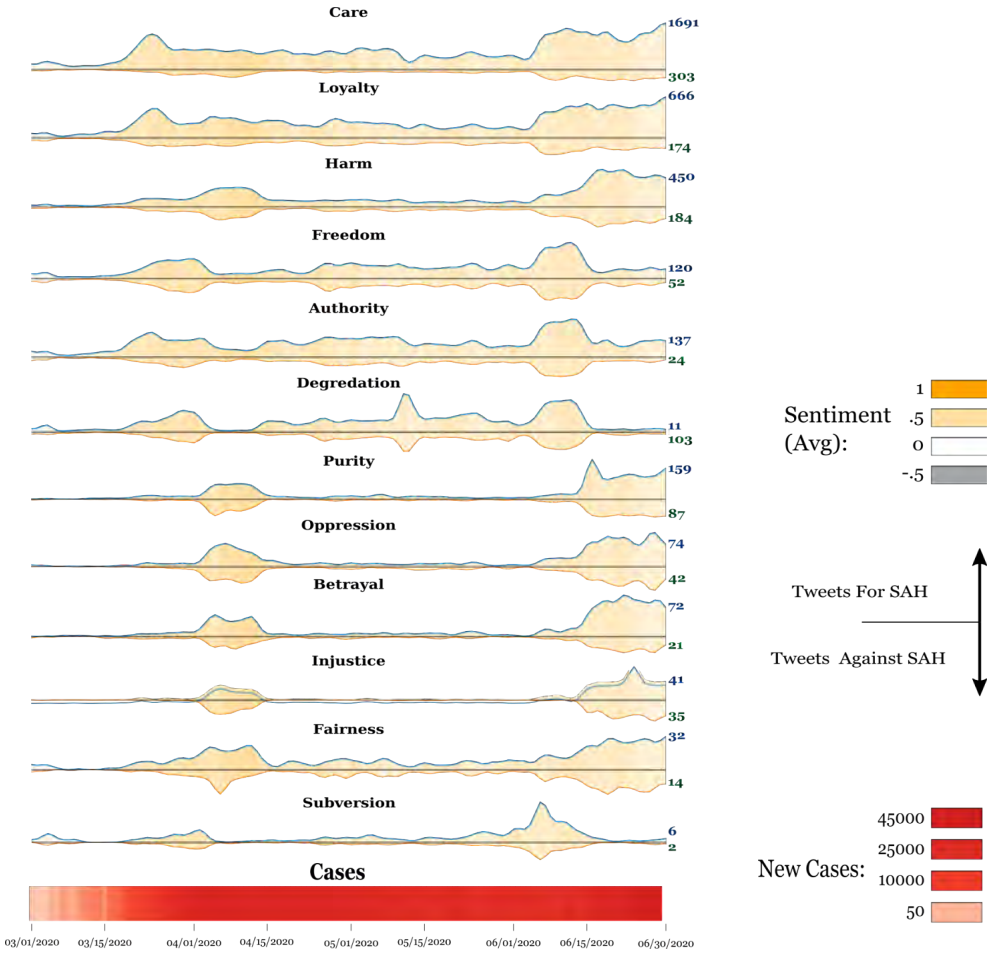
Steamgraph

This design was an early attempt to explore aggregated ways of visualizing the larger set of frames over time. This example is a streamgraph using a larger dataset of ~100,000 tweets that did not include geospatial information. Frames are ordered such that the frames with the most tweets are located closer to the center axis.



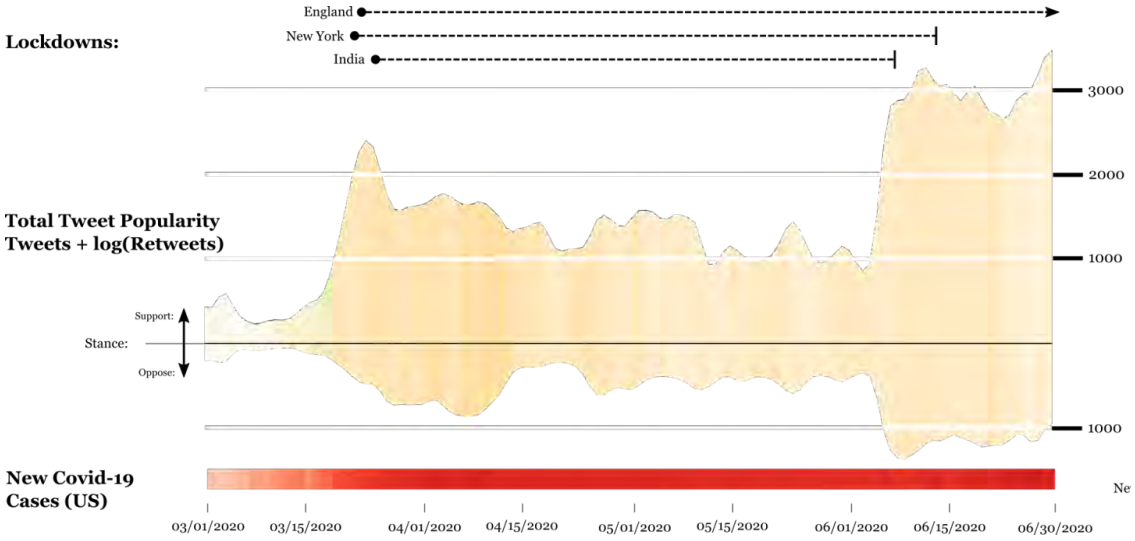
Small Multiples

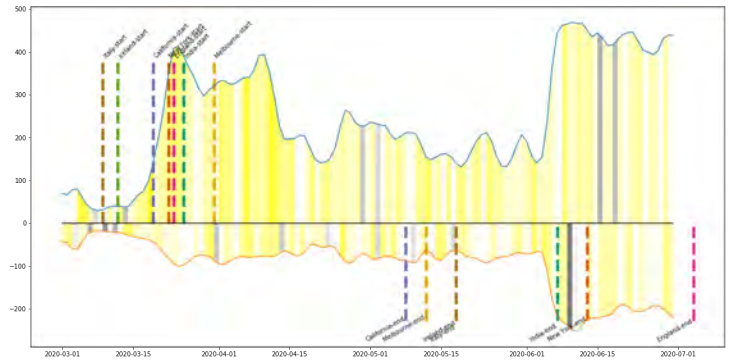
This encoding was a design used during the analysis of a larger dataset of non-geotagged tweets. We used small multiples to encode trends in tweet popularity over time. The color of each slice encodes the weighted average sentiment of all tweets within the time period. Overall cases are shown at the bottom. This encoding is useful for identifying periods of relative increased activity for individual frames. Time periods of note were explored offline by investigating popular tweets with a specific frame and stance during those time periods.



Aggregated Tweet Timeline

This encoding shows overall tweets of all frames. In this example, we manually noted the beginning and end of different prominent lockdowns, in an attempt to correlate changes in activity to different events. Time periods of note were explored offline by investigating popular tweets during those time periods.



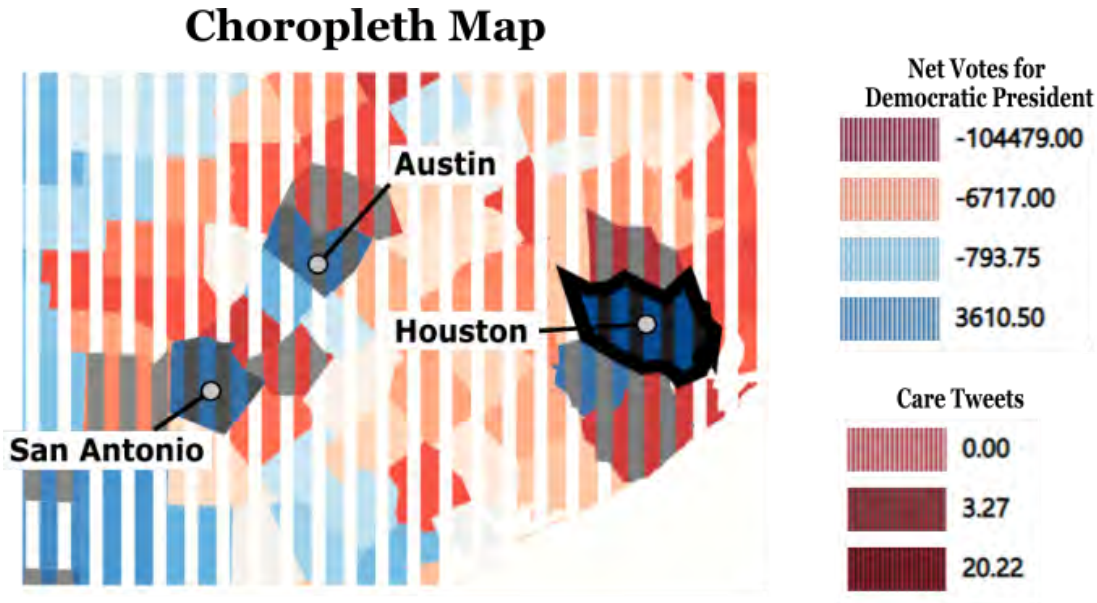


Geospatial Encodings

During the data foraging process, we investigated different ways of visualizing multiple demographic features for each county, as well as visualizing the distribution of tweets that were being collected. These are earlier approaches that we tested before using our final glyph-based map encoding.

Choropleth Texture Map

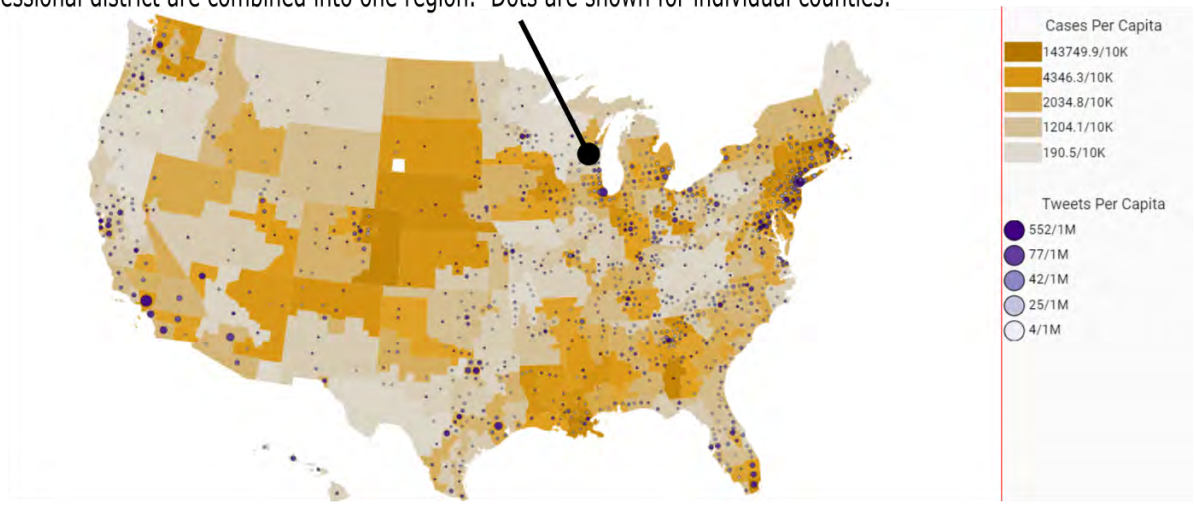
An earlier approach to encoding multiple values was encoding two variables as alternating stripes, where a demographic variable used a saturated hue, while a secondary variable used a grayscale hue. This was useful for finding areas with high values for both variables. However, we find it was less effective for our purposes than using the glyph encoding.



Choropleth + Glyph Map

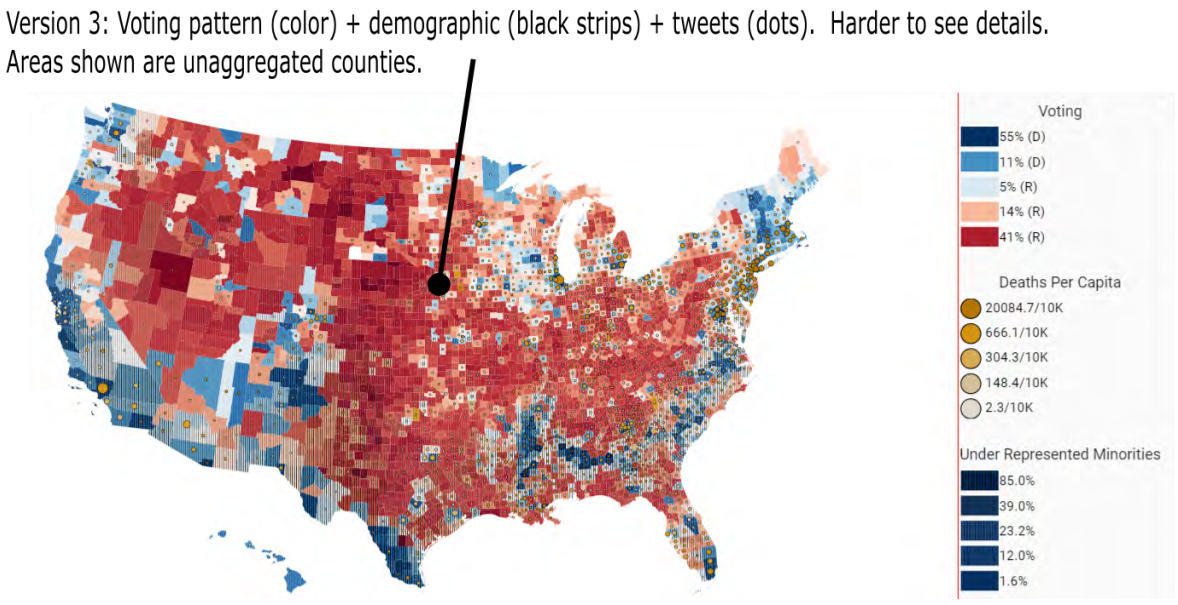
This earlier variant of a map encoded covid rates as a single saturated color that was aggregated by voting district, while tweets were shown as circles, aggregated by county. Tweet count was encoded as both circle shape and size. We found this was useful to avoid overlap with high tweet counts during the foraging process, since tweet data tended to be sparse.

Version 2: demographics are yellow, tweets are dots (more = larger, more viral = darker)
This granularity is counties aggregated such that those that have the largest mutually overlapping congressional district are combined into one region. Dots are shown for individual counties.



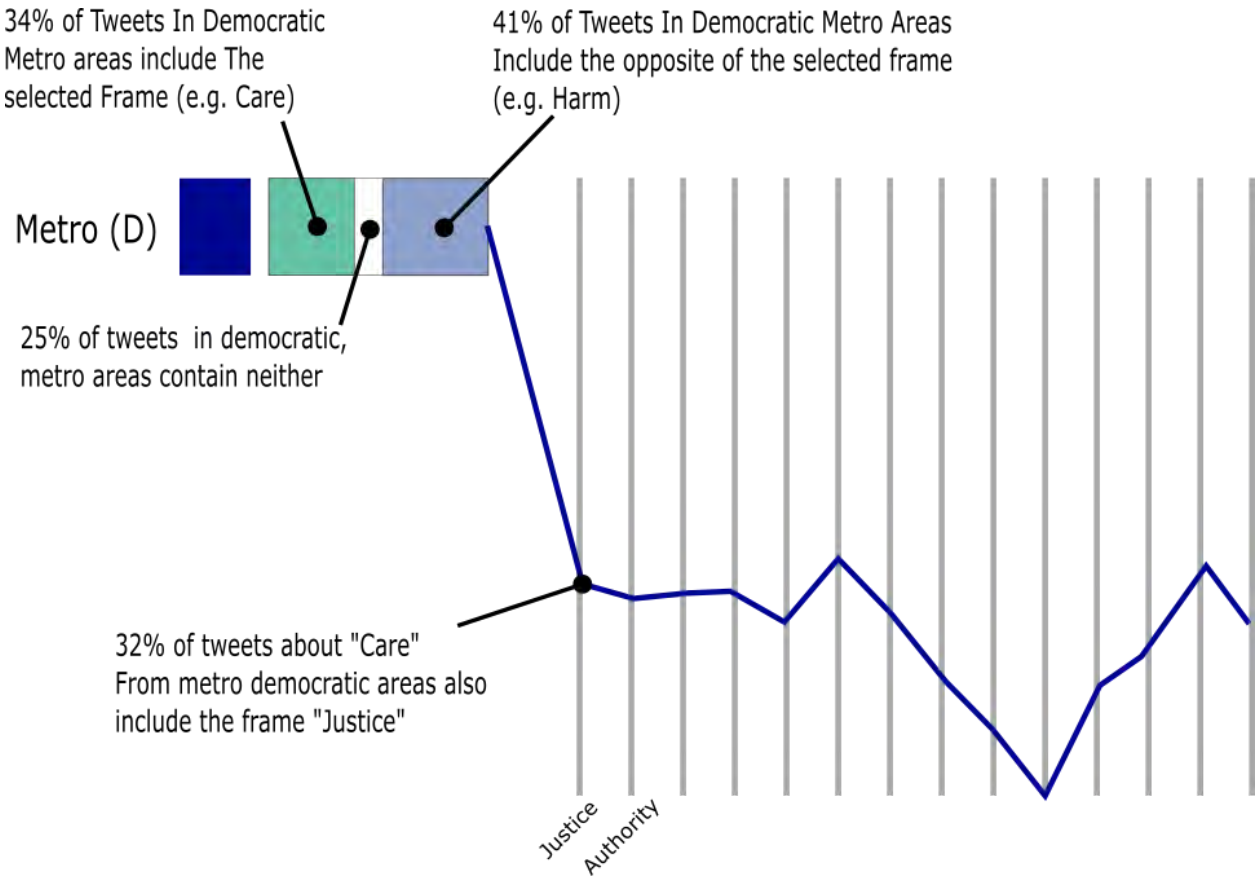
Choropleth Texture + Glyph map

This variant attempted to show three county level variables, using both textures and circle glyphs. While this approach was somewhat interesting, we found it was too difficult to discern interesting areas around high population areas.



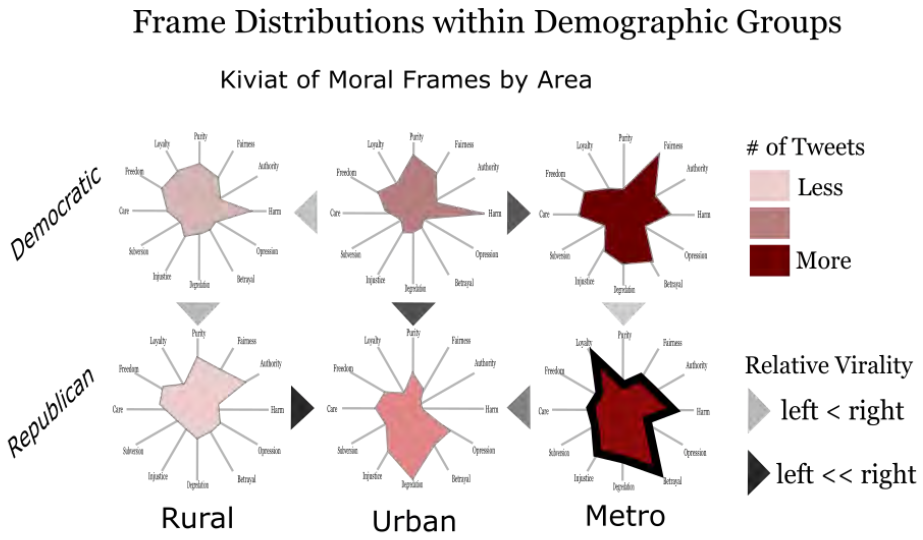
Glyphs + Parallel Coordinates

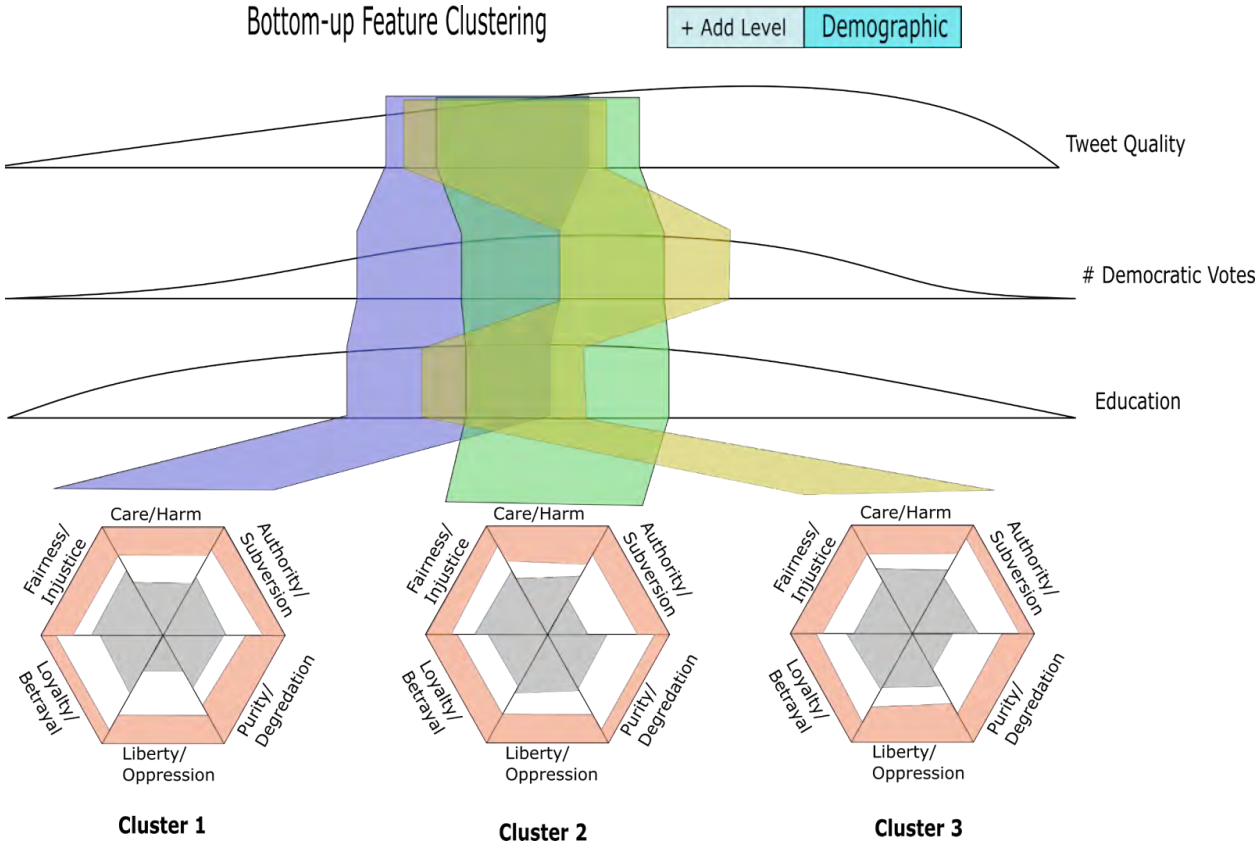
This prototype looked at a non-geospatial approach to demographic groups, by grouping regions based on the stratification of the rural-urban continuum codes and voting districts, since we were primarily interested in the difference between democratic cities and rural areas. This was scrapped as we also wanted to identify outlier counties that may be affected by regional policy differences.

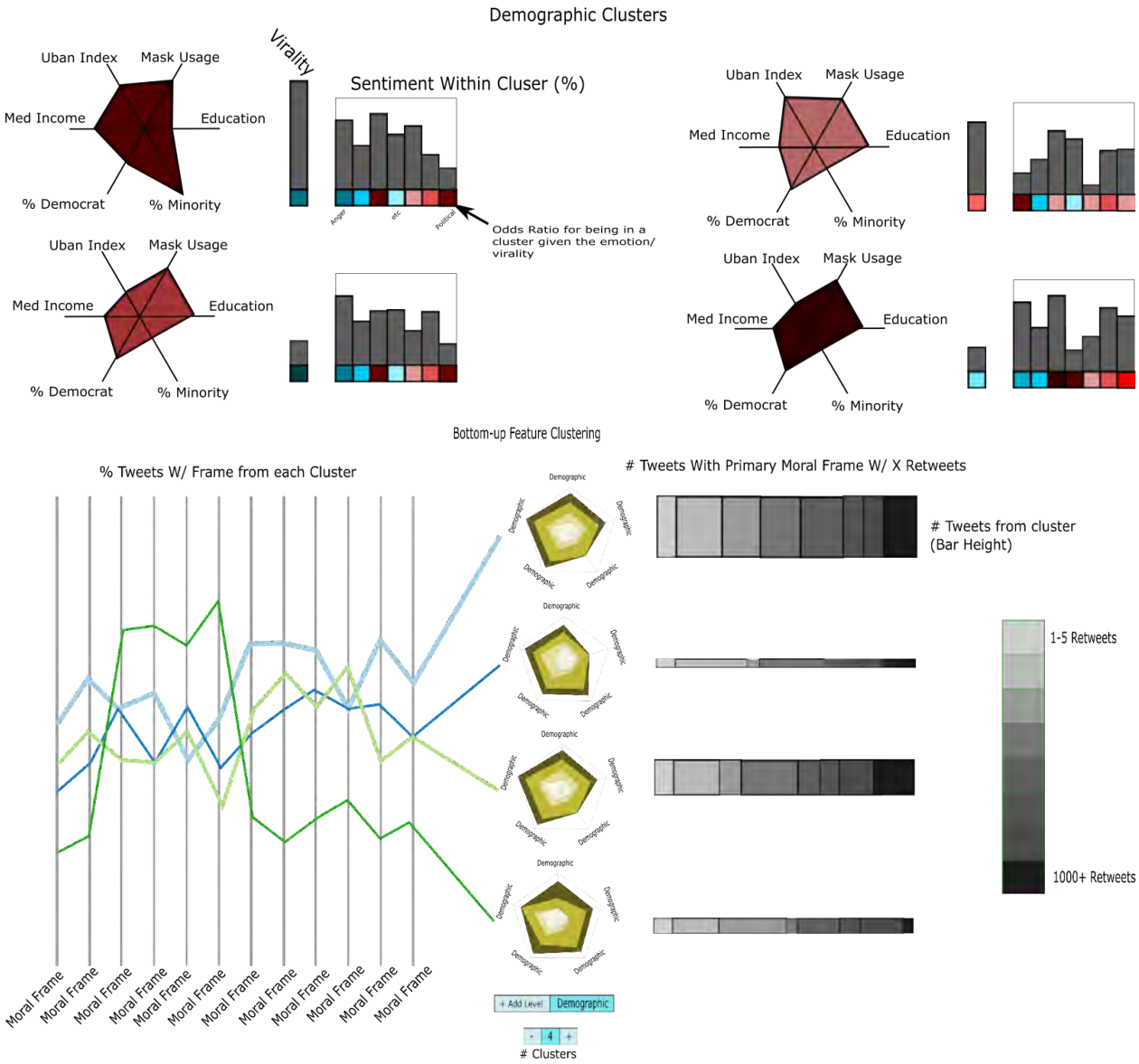


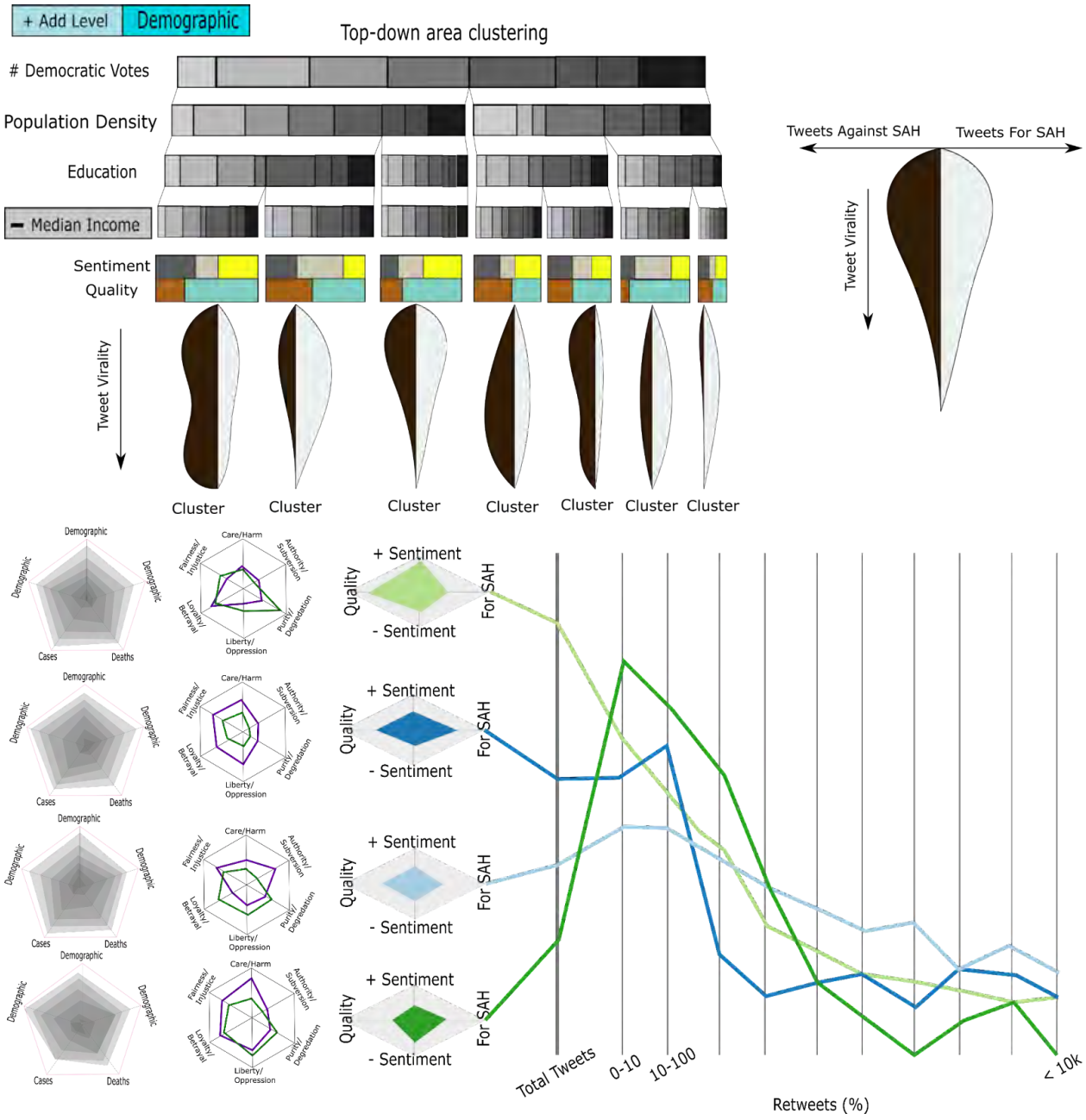
Demographic Clusters

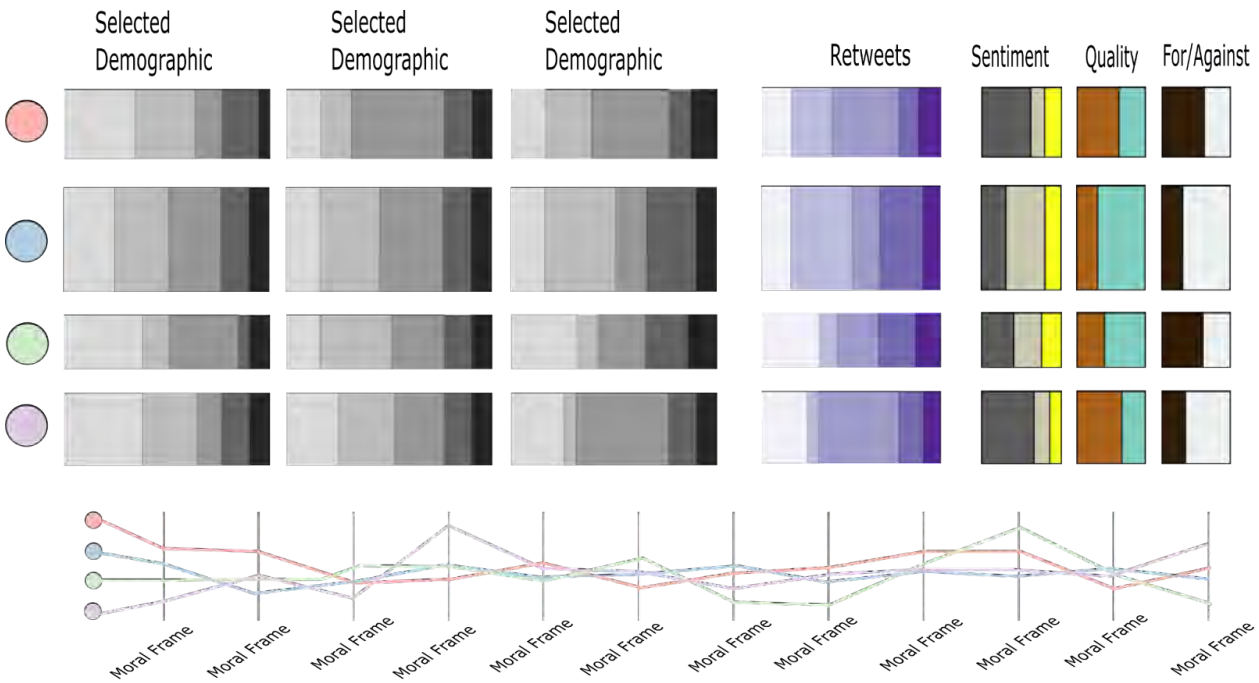
Our early approaches looked at ways of automatically clustering tweets and counties based on similar features, to simplify analysis and identify interesting patterns. Ultimately, these prototypes were not included, as we found the cognitive load of reasoning using clusters was too high for our collaborators, and we ultimately replaced this approach with our inference view, as our partial dependence plots allowed for a simpler explanation of the relationship between multiple demographic and tweet features.

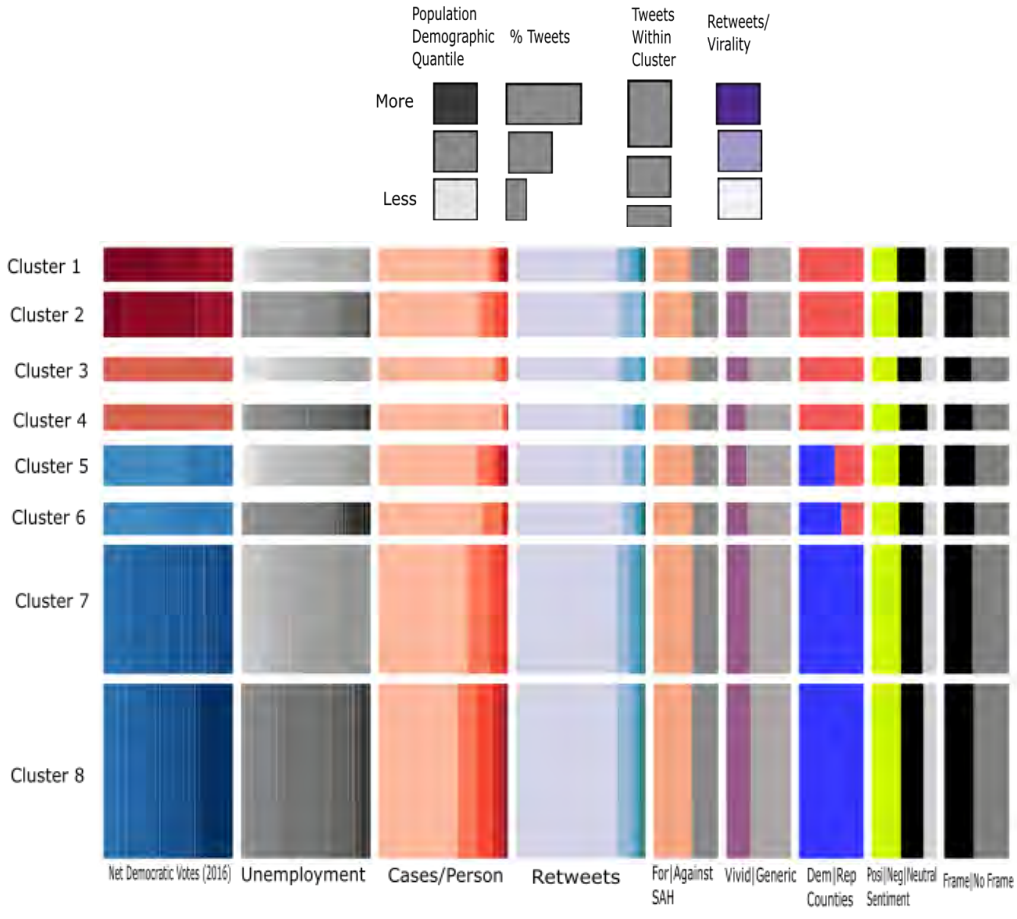

Frame Distributions within Demographic Groups
Kiviat of Moral Frames by Area
Democratic
Republican
# of Tweets
Less
More
Relative Virality
left < right
left << right
Rural
Urban
Metro


Bottom-up Feature Clustering
+ Add Level
Demographic
Tweet Quality
Democratic Votes
Education
Care/Harm
Fairness/ Injustice
Authority/ Subversion
Loyalty/ Betrayal
Purity/ Degredation
Liberty/ Oppression
Cluster 1
Cluster 2
Cluster 3


Demographic Clusters
Uban Index
Mask Usage
Med Income
Education
% Democrat
% Minority
Virality
Sentiment Within Cluser (%)
Anger
etc.
Political
Odds Ratio for being in a cluster given the emotion/ virality
Bottom-up Feature Clustering
% Tweets W/ Frame from each Cluster
Moral Frame
Tweets With Primary Moral Frame W/ X Retweets
Demographic
Tweets from cluster (Bar Height)
1-5 Retweets
1000+ Retweets
+ Add Level
Demographic
- 4 +
Clusters

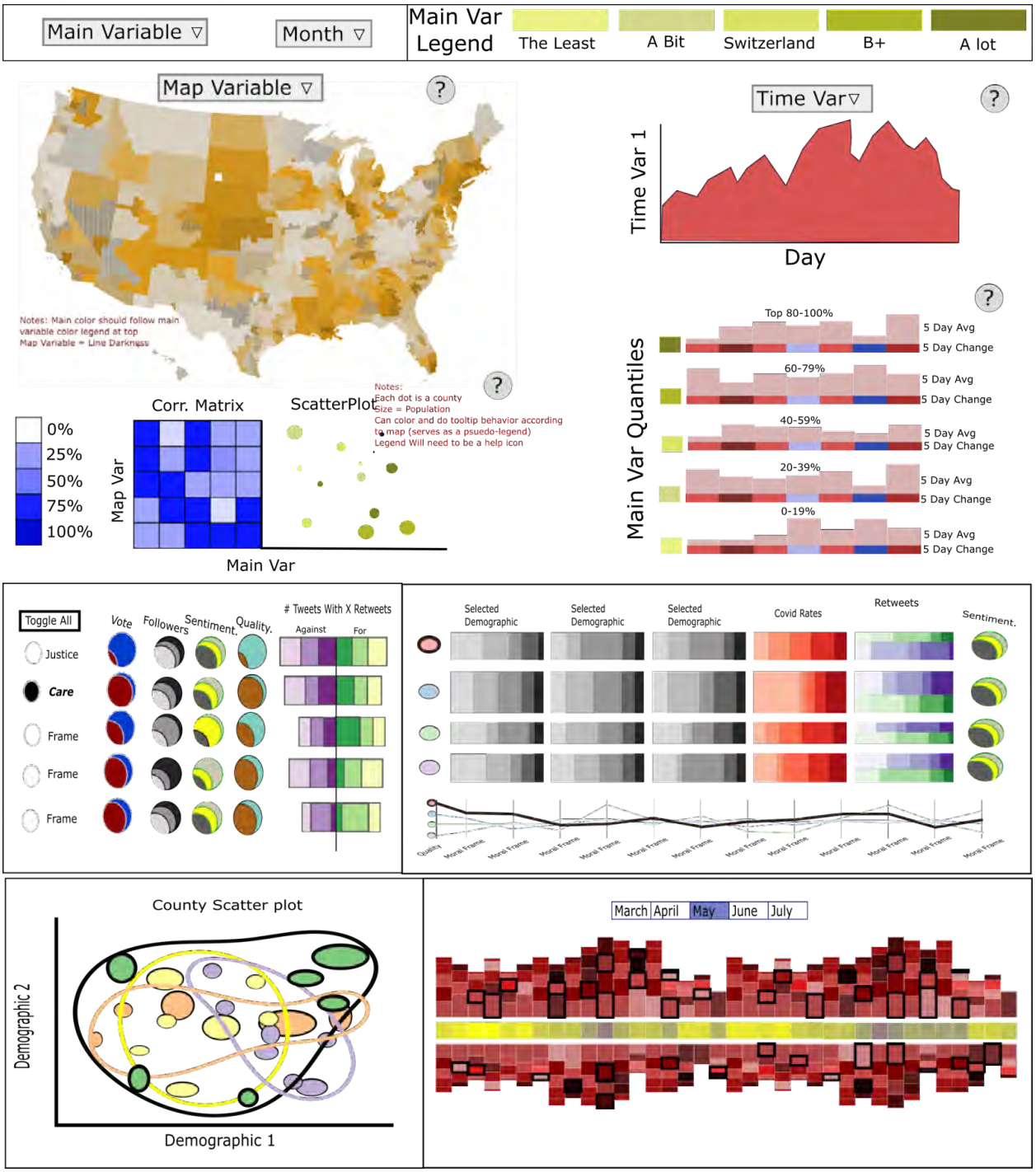

+ Add Level
Demographic
Top-down area clustering
Democratic Votes
Population Density
Education
Median Income
Sentiment
Quality
Tweet Virality
Cluster
Cluster
Cluster
Cluster
Cluster
Cluster
Cluster
Tweets Against SAH
Tweets For SAH
Tweet Virality
Demographic
Cases
Deaths
Care/Harm
Fairness/ Injustice
Loyalty/ Betrayal
Authority/ Subversion
Purity/ Degradation
Liberty/ Oppression
+ Sentiment
- Sentiment
Quality
For SAH
Total Tweets
0-10
10-100
< 10k
Retweets (%)

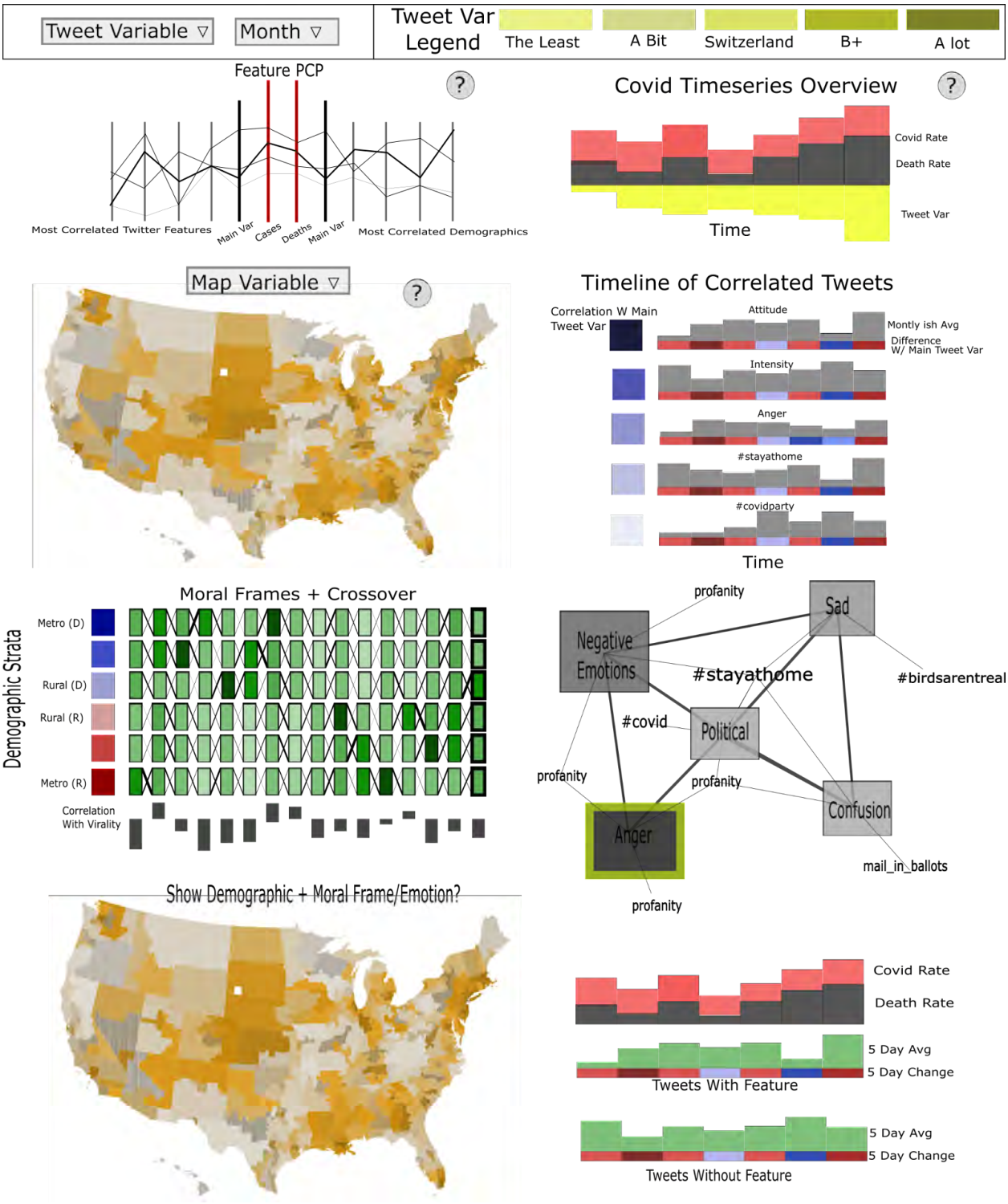

Selected Demographic
Selected Demographic
Selected Demographic
Retweets
Sentiment
Quality
For/Against
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame
Moral Frame


Population Demographic Quantile
% Tweets
Tweets Within Cluster
Retweets/ Virality
More
Less
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Net Democratic Votes (2016)
Unemployment
Cases/Person
Retweets
For|Against SAH
Vivid|Generic
Dem|Rep Counties
Posi|Neg|Neutral Sentiment
Frame|No Frame

Early Prototypes

These are early prototypes of our system, which focused on mockups that looks at different layouts, with the goal of eliciting what kinds of information and visualizations our collaborators were most interested in during the data foragin process.




Tweet Variable
Month
Tweet Var Legend
The Least
A Bit
Switzerland
B+
A lot
Feature PCP
Most Correlated Twitter Features
Main Var
Cases
Deaths
Main Var
Most Correlated Demographics
Covid Timeseries Overview
Covid Rate
Death Rate
Tweet Var
Time
Map Variable
Timeline of Correlated Tweets
Correlation W Main Tweet Var
Attitude
Montly ish Avg
Difference W/ Main Tweet Var
Intensity
Anger
#stayathome
#covidparty
Time
Moral Frames + Crossover
Demographic Strata
Metro (D)
Rural (D)
Rural (R)
Metro (R)
Correlation With Virality
profanity
Sad
Negative Emotions
#stayathome
#birdsarentreal
#covid
Political
profanity
profanity
Confusion
Anger
mail_in_ballots
profanity
Show Demographic + Moral Frame/Emotion?
Covid Rate
Death Rate
5 Day Avg
5 Day Change
Tweets With Feature
5 Day Avg
5 Day Change
Tweets Without Feature

8.4 Appendix D: Chapter 6 (DITTO) model Details and Evaluation

8.4.1 Model Details

Patient Simulator Models

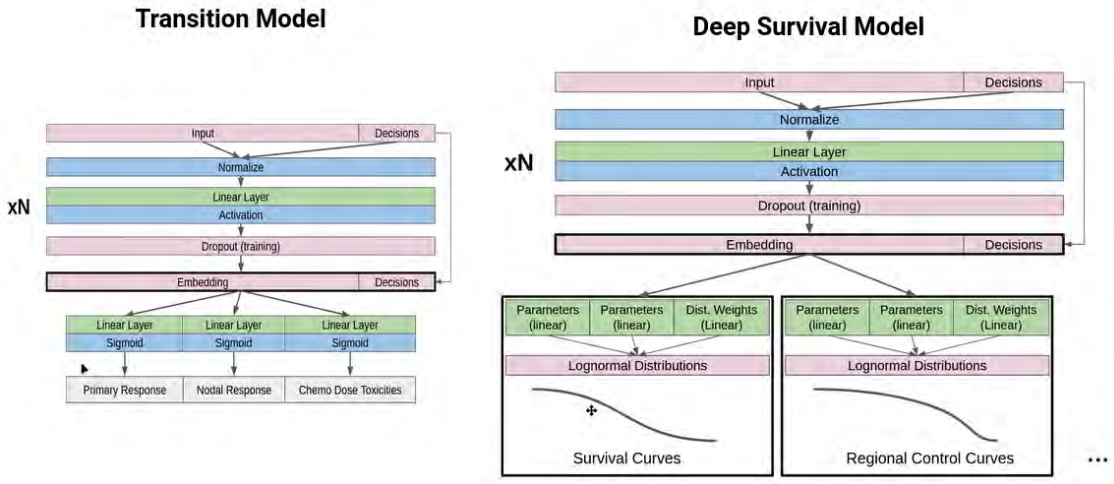


Figure 8.1: Architecture for the transition and outcome DSM models. Patient state and previous state treatment decision use a standard DNN with input dropout to improve the models' ability to deal with unknown data. The decision is concatenated to the penultimate layer in order to prevent the model from relying only on correlated features due to the use of dropout during training. DSM models predict a mixture of model parameters for each patient from a pre-trained set of user-defined number of mixtures.

To simulate the patient, we use a set of models to mimic intermediate response to treatment (transition models), and long-term response after treatment (outcome models). Fig. 8.1 Shows the architecture for the transition models and outcome DSM models. Each time-point uses a separate transition state model. For IC, we constrain the model to not allow for any tumor response when no IC is given, as this would indicate no treatment at this point.

Transition models predict patient response to treatment in terms of tumor shrinkage and severe toxicities from treatment. Specifically, we consider primary disease response (PD), and nodal disease response (ND), which are each 4 categorical ordinal variables, as well as 5 binary results for different types of dose-limiting toxicities (DLTS). For the case of Induction chemotherapy, disease response is always assumed to be stable when no treatment is done. Separate models are trained for post-IC and post-CC transitions, as this resulted in better performance.

For the outcome model, two separate models are used. The first is a deep neural network that predicts toxicity risk using binary variables: Aspiration (AS), and Feeding Tube (FT) at 6 months after treatment.

The second outcome model predicts cumulative patient risk over time for overall survival (OS), locoregional control (LRC), and distant metastases (FDM) for up to 5 years. Temporal risk models use a variant of deep survival machines (DSM) [232]. For all three outcomes, the DSM model returns a mixture of parametric log-normal distributions for the patient that can be used to provide a cumulative survival risk over time.

Because clinicians listed confidence intervals as important for reasoning about the model predictions (T3.3), all transition and outcome models are trained using dropout on the penultimate layer between 50% and 75%. During evaluation, we re-run each prediction with random dropout at least 20 times, and then save the 95% confidence intervals for each prediction. [104].

All models implemented in pytorch and trained using the Adam optimizer. Models were trained using early stopping until the validation loss stopped increasing for at least 10 epochs. Transition models, static outcome models, and Deep Survival Machines for temporal outcomes used a dropout of 10% on the input layer and 50% on the penultimate layer during training. Transition models and static outcome models used 2 hidden layers with an output size of 500 each. The DSM used a single hidden layer with a size of 100 and 6 different distributions for each outcome.

Policy Models

The patient simulator models and ground truth responses are used as the environment to train a digital physician (policy model) (Fig. 6.3). Because there is disagreement among users as to whether they prefer to see what a physician would do, or what the "best" choice should be, we jointly train two versions of the policy model: one that minimizes a combination of patient risks based on the patient simulator responses (optimal policy model), and one that

predicts what a physician would do based on the cohort data (imitation policy model).

Each policy model (optimal and imitation) is trained using a dual loss function: prediction of the ground truth (or optimal) decision sequence, and triplet loss. Triplet loss is included as it was found to increase model performance in terms of AUC and accuracy for the imitation model. Specifically, the loss for a given patient p at each epoch for a given output (optimal or imitation) is given by:

$$L(p) = w_1 \cdot \sum_{i=0}^{2} BCE(\hat{y_{p,i}}, y_{p,i}) + w_2 \cdot max\Big(d(a_p, b_p) - d(a_p, c_p) + 1, 0\Big)$$

Where y_i and $\hat{y}_i$ are the ground truth and predicted decisions, respectively. $d(\cdot, \cdot)$ is the Euclidean distance. a_p is the final hidden layer weight vector for the patient. b_p are the hidden weights for a randomly sampled patient with the same ground truth treatment sequence, and c_p is a randomly sampled patient with a different treatment sequence. w_1 and w_2 are user-decided weights. For our implementation, we use $w_1 = 1$ and $w_2 = .2$ for both outputs.

Both optimal and imitation policy models use shared layers until the penultimate layers, which are unique to each output, and are re-trained each epoch (Fig. 6.3). This allows for joint learning of important features from each other. To encourage our model to explicitly consider other patients in the cohort, our policy model architecture uses a transformer encoder and uses a position token to encode the temporal state of the patient. During evaluation on a new datapoint, the cohort data for the current state is used as the Query input of the multi-headed attention as described in Vaswani et al. [321].

Imitation policy model decisions are trained using the unaltered ground truth states in the data to predict the decision made by the clinician. The optimal model decision is, in contrast, trained on using random data augmentation on the pre-treatment variables for the patient for each epoch. Specifically, each column has a 25% probability of being pseudo-randomly shuffled in the training sample, and the predicted patient response to treatment using the deep learning transition models.

When determining the treatment sequence for the optimal decision, we calculate the decisions that minimize a combination of all predicted outcomes, given by:

$$L = w_{tox} \sum_{z \in Z} w_z P(z = 1) + w_s \sum_{o \in O} \frac{w_o}{\tilde{T}(o)}$$

Where $z \in Z$ is the set of binary outcomes (e.g toxicity, 4 year survival, 4 year locoregional control), $o \in O$ is the set of temporal survival outcomes (survival, locoregional control, distant control), $\tilde{T}(o)$ is the median predicted time-to-event of outcome o, and $w \in W$ a set of user defined weights for each aspect of the loss function.

Because we need to explain the policy model recommendations (T3), we use integrated gradients [293] to obtain feature importance for each decision relative to a baseline value. Integrated gradients was chosen as it satisfies the completeness axiom where attributions sum to the difference in the prediction between the baseline and actual recommendation, which was found to be easier to reason about with our clinicians. For our baseline, we assume the lowest possible rating for most ordinal attributes such as tumor staging or disease response, and the most common value for categorical attributes such as gender, ethnicity, and tumor subsite, as well as age and dose to the main tumor, based on feedback from clinicians and what they found most intuitive.

All models implemented in pytorch and trained using the Adam optimizer. Models were trained using early stopping until the validation loss stopped increasing for at least 10 epochs. Our policy model used an input dropout of 10% and 25% dropout on the final layers, with a single transformer encoder of size 1000 for the joint embedding, and a linear layer of size 20 for both the optimal and imitation model outputs.

KNN-based Symptom Prediction

Our symptom prediction model uses a different cohort of patient and relies on a KNN predictor using the embeddings taken from a model trained to predict symptom trajectories. Specifically, we trained a fully connected deep learning model to predict symptom ratings

for each symptom and each time point in the data. Time points considered were at 0, 7, 12, and 27 weeks after starting radiation therapy. Outputs were treated as independent values with a sigmoid loss function that was scaled to be between 0 and 10. Input features were gender, packs-years, HPV status, treatment dose and dose frction, race, tumor laterality, tumor subsite, T-category, N-category, and treatment decisions for IC and CC.

Patient embeddings for the cohort were taken from the model activations in the batch-normalized penultimate layer in the deep learning model. When predicting a new patient, we take the new patient's model embeddings and extract the 10 most similar patients, based on euclidean distance, from the embeddings of both the treated and untreated patients, respectively. Patient symptom profiles are taken for these patient seperately

During deep learning model training, we used an 80/20 train validation split on the data for parameter tuning, using the mean-squared-error loss (MSE). Missing symptom values were ignored in the loss function. All models were trained using the ADAM optimizer in pytorch using early stopping on the validation loss. Our final model used a single hidden layer of size 10 with the ReLU activation, followed by batch normalization, with no dropout.

8.4.2 Model Evaluation

We evaluated our system on a cohort of 536 patients. The dataset was split into a training cohort of 389 patients and an evaluation cohort of 147 patients before beginning the development of the models. The training sample was stratified in order to get a minimum of 3 patients with each endpoint, and treatment decision in the model. Because we could not achieve enough samples of patients with several dose limiting toxicities, all toxicities that were not present in both cohorts were aggregated into an "other" category for the purpose of modeling and evaluation. The features used for the entire cohort, excluding lymph node patterns, is shown in (Table 8.2), stratified by treatment sequence. An anova F-test was used to analyze correlations between each feature set and the treatment sequence, and p-values are included in the table.

Performance of the policy model with and without triplet loss is shown in (Section 8.4.2).

We see an increase in imitation model performance, with a slight decrease in "optimal" model performance for accuracy but increase in AUC. This is likely due to the heavy imbalance in the optimal outcomes: only 10.8% of cases recommended concurrent chemotherapy and 19% of cases recommended neck dissection, as rare events were predicted with higher prediction confidence. Given that a majority of users preferred to use the "imitation" model, the triplet model was used in practice.

In general, AUC tended to perform better than Accuracy in the optimal model, likely due to the heavy imbalance in the optimal outcomes: only 10.8% of cases recommended concurrent chemotherapy and 19% of cases recommended neck dissection. In general, model performance is comparable to similar outcome models from earlier studies, considering the added difficulty of optimizing for 23 different outcomes and 6 treatment decisions. Interestingly, our optimal model suggested induction chemotherapy followed by radiation alone a majority of the time, which contradicts the standard practice where concurrent chemotherapy is standard while induction is used for patients with very large tumor spread that needs to be reduced before applying radiation. However, the data is largely limited by confounders and lack of detailed information on how changes in patient's health affect treatment and outcomes. Additionally, we have been told that the specific grade of dose-limiting toxicity is an important factor in treatment and side effects, which our model does not consider.

Performance of transition models are shown in (Section 8.4.2). Because the outcomes we want to predict are often rare events, we compared default training performance with basic cross-entropy loss with a balanced loss function. Non-balanced models generally performed better in terms of AUC with similar accuracy.

To evaluate time series models, we calculate F1 and ROC AUC scores at 12, 24, 36, and 48 months after treatment (Table 8.4). We exclude longer periods, as we tend to have fewer followup data available after 48 months. OS, FDM and LRC models tend to have high F1 score but modest AUC scores, possibly due to the fact that failures are rare events in the data.

Decision	Optimal		Imitation	
	AUC	Accuracy	AUC	Accuracy
With Triplet Loss				
IC	0.84	0.58	0.79	0.88
CC	0.97	0.73	0.93	0.78
ND	0.95	0.79	0.90	0.81
No Triplet Loss				
IC	0.82	0.71	0.60	0.87
CC	0.96	0.91	0.74	0.78
ND	0.94	0.88	0.84	0.81

Table 8.1: Physician Simulator Policy Model Performace with and without use of triplet loss.

Treatment Sequence	CC	None	CC + ND	IC + CC	IC + CC + ND	IC	ND	IC + ND	P-Value
Count	223	57	51	100	36	45	11	13	1
HPV+	56.50%	80.70%	54.90%	50.00%	61.11%	42.22%	54.55%	61.54%	6.44E-03
HPV Unknown	6.28%	1.75%	7.84%	16.00%	11.11%	2.22%	18.18%	7.69%	
Age (Mean)	59.3	61.3	57.7	58.5	58.3	57.6	59.6	57.0	4.92E-01
Pack-years	17.6	10.5	18.9	17.6	21.8	15.4	16.7	4.8	1.83E-01
Male	88.34%	80.70%	92.16%	87.00%	91.67%	88.89%	81.82%	92.31%	6.76E-01
Smoker	19.28%	19.30%	35.29%	22.00%	22.22%	24.44%	18.18%	0.00%	2.38E-01
Former Smoker	42.15%	40.35%	29.41%	34.00%	33.33%	33.33%	54.55%	30.77%	
Bilateral	4.48%	3.51%	5.88%	4.00%	2.78%	2.22%	0.00%	0.00%	9.50E-01
T-category_1	18.83%	63.16%	5.88%	6.00%	13.89%	28.89%	54.55%	30.77%	1.30E-18
T-category_2	42.15%	33.33%	54.90%	33.00%	27.78%	48.89%	45.45%	61.54%	4.45E-02
T-category_3	24.66%	3.51%	21.57%	29.00%	27.78%	17.78%	0.00%	7.69%	3.25E-03
T-category_4	14.35%	0.00%	17.65%	32.00%	30.56%	4.44%	0.00%	0.00%	6.69E-08
N-category_1	52.91%	80.70%	52.94%	27.00%	16.67%	22.22%	63.64%	61.54%	4.96E-14
N-category_2	39.46%	12.28%	43.14%	65.00%	75.00%	73.33%	27.27%	38.46%	7.13E-14
N-category_3	1.79%	0.00%	0.00%	8.00%	8.33%	4.44%	0.00%	0.00%	1.91E-02
AJCC_2	15.25%	5.26%	15.69%	16.00%	22.22%	22.22%	9.09%	7.69%	1.66E-16
AJCC_3	9.42%	5.26%	13.73%	22.00%	25.00%	0.00%	18.18%	0.00%	2.10E-04
AJCC_4	36.77%	12.28%	39.22%	49.00%	38.89%	57.78%	18.18%	38.46%	5.90E-05
subsite_BOT	50.22%	35.09%	47.06%	56.00%	55.56%	57.78%	18.18%	46.15%	7.85E-02
subsite_GPS	0.90%	1.75%	1.96%	2.00%	8.33%	0.00%	0.00%	7.69%	7.50E-02

subsite_Soft palate	0.90%	1.75%	3.92%	1.00%	0.00%	0.00%	0.00%	0.00%	6.48E-01
subsite_Tonsil	41.26%	54.39%	41.18%	36.00%	33.33%	40.00%	81.82%	30.77%	4.78E-02
Pathological Grade_1	0.90%	0.00%	0.00%	3.00%	0.00%	0.00%	9.09%	0.00%	1.04E-01
Pathological Grade_2	28.25%	31.58%	27.45%	28.00%	33.33%	28.89%	45.45%	7.69%	6.63E-01
Pathological Grade_3	50.67%	54.39%	56.86%	48.00%	55.56%	46.67%	36.36%	61.54%	8.39E-01
Pathological Grade_4	0.90%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.69%	5.00E-02
White/Caucasion	93.27%	89.47%	96.08%	86.00%	86.11%	93.33%	90.91%	92.31%	3.57E-01
Aspiration Pre-Therapy	2.24%	0.00%	1.96%	7.00%	5.56%	2.22%	0.00%	0.00%	2.14E-01
Total Dose (gy)	68.99	66.86	69.47	69.36	69.33	67.42	68.05	67.23	1.05E-14
Dose Fractions	2.10	2.16	2.08	2.11	2.08	2.15	2.17	2.18	7.82E-05
Survival (Months)	76.26	71.24	80.52	74.10	74.47	87.80	98.57	97.21	1.76E-02
Locoregional control (Months)	74.46	67.63	67.28	71.36	63.85	86.54	72.10	94.07	2.50E-02
FDM (months)	74.69	71.00	77.95	72.38	69.56	84.34	94.86	97.21	5.02E-02
Overall Survival	75.34%	82.46%	70.59%	75.00%	61.11%	86.67%	63.64%	100.00%	4.36E-02
Locoregional Control	91.03%	92.98%	68.63%	85.00%	69.44%	91.11%	63.64%	84.62%	1.74E-05
FDM	89.69%	96.49%	86.27%	90.00%	80.56%	88.89%	72.73%	100.00%	1.25E-01
FT	17.49%	5.26%	21.57%	25.00%	38.89%	8.89%	18.18%	0.00%	4.80E-04
Aspiration Post-Therapy	17.49%	3.51%	25.49%	22.00%	41.67%	8.89%	18.18%	7.69%	1.87E-04
CR Primary	0.00%	0.00%	0.00%	34.00%	38.89%	66.67%	0.00%	46.15%	2.99E-50
PR Primary	0.00%	0.00%	0.00%	52.00%	55.56%	28.89%	0.00%	30.77%	2.02E-51
CR Nodal	0.00%	0.00%	0.00%	10.00%	2.78%	11.11%	0.00%	0.00%	1.79E-06
PR Nodal	0.00%	0.00%	0.00%	75.00%	88.89%	86.67%	0.00%	84.62%	1.21E-148
DLT after CC	0.00%	0.00%	0.00%	75.00%	50.00%	64.44%	0.00%	84.62%	0.00E+00
CR Primary 2	83.41%	91.23%	68.63%	90.00%	72.22%	91.11%	81.82%	92.31%	4.86E-03
PR Primary 2	16.14%	7.02%	25.49%	10.00%	19.44%	4.44%	18.18%	7.69%	3.56E-02
CR Nodal 2	52.02%	52.63%	17.65%	58.00%	19.44%	57.78%	9.09%	7.69%	1.38E-09
PR Nodal 2	43.95%	35.09%	80.39%	34.00%	77.78%	37.78%	90.91%	69.23%	4.46E-11
DLT After CC/RT	27.80%	0.00%	15.69%	29.00%	25.00%	0.00%	0.00%	0.00%	2.95E-07

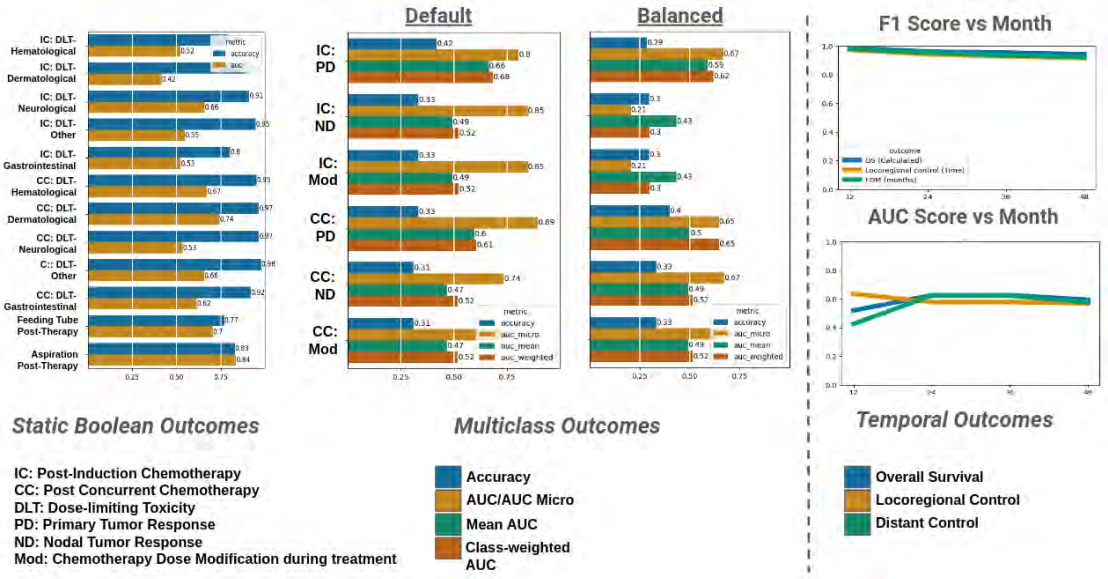


Figure 8.2: All Transition State Outcomes. (Right) Accuracy and AUC score for boolean outcomes such as toxicities. Models perform well in terms of accuracy and late toxicity (FT and Aspiration), but have mixed AUC results for dose-limiting toxicities due to the heavy imbalance in the data and low number of positive samples to learn from. (Center) Model performance for multi-class transition states (disease response and dose modification) using accuracy and micro, macro, and weighted AUC score for both unweighted and balanced loss weights. Models perforce best in terms of macro AUC score. Balanced models generally performed worse. (Right) F1 score and AUC score for temporal outcomes at 12, 24, 36, and 48 months after treatment. F1 scores tend to be very high while AUC scores stay around .6, likely due to issue with imbalanced data and incomplete censoring.

State	Outcome	Metric	Value
After IC	Primary Response	accuracy	0.404
	Primary Response	auc_micro	0.801
	Primary Response	auc_weighted	0.674
	Nodal Response	accuracy	0.333
	Nodal Response	auc_micro	0.853
	Nodal Response	auc_weighted	0.533
	Dose Modification	accuracy	0.333
	DLT_Gastrointestinal	accuracy	0.804
	DLT_Other	accuracy	0.946
	DLT_Dermatological	accuracy	0.893
	DLT_Hematological	accuracy	0.786
	DLT_Neurological	accuracy	0.911
	DLT_Gastrointestinal	auc	0.497
	DLT_Other	auc	0.415
	DLT_Dermatological	auc	0.420
	DLT_Hematological	auc	0.511
	DLT_Neurological	auc	0.557
After RT + CC	Primary Response	accuracy	0.333
	Primary Response	auc_micro	0.887
	Primary Response	auc_weighted	0.568
	Nodal Response	accuracy	0.372
	Nodal Response	auc_micro	0.756
	Nodal Response	auc_weighted	0.545
	DLT_Gastrointestinal	accuracy	0.918
	DLT_Other	accuracy	0.980
	DLT_Dermatological	accuracy	0.966
	DLT_Hematological	accuracy	0.952
	DLT_Neurological	accuracy	0.966
	DLT_Gastrointestinal	auc	0.564
	DLT_Other	auc	0.727
	DLT_Dermatological	auc	0.625
	DLT_Hematological	auc	0.613
	DLT_Neurological	auc	0.552
After All Treatment	Feeding Tube	accuracy	0.803
	Feeding Tube	auc	0.683
	Feeding Tube	f1	0.216
	Aspiration Post-therapy	accuracy	0.803
	Aspiration Post-therapy	auc	0.775
	Aspiration Post-therapy	f1	0.065

Table 8.3: Model Performance for all transition states and toxicity

outcome	months	metric	value
OS	12	AUC	0.52
		F1	0.99
	24	AUC	0.63
		F1	0.96
	36	AUC	0.64
		F1	0.95
	48	AUC	0.60
		F1	0.94
Locoregional Control	12	AUC	0.64
		F1	0.97
	24	AUC	0.56
		F1	0.94
	36	AUC	0.57
		F1	0.93
	48	AUC	0.57
		F1	0.92
Distant Control	12	AUC	0.42
		F1	0.98
	24	AUC	0.62
		F1	0.95
	36	AUC	0.62
		F1	0.93
	48	AUC	0.57
		F1	0.92

Table 8.4: Model Performance for Deep Survival Models at 12, 24, 36, and 48 months.

8.5 Appendix E: Chapter 6 (DITTO) Prototypes

8.5.1 Prototypes

Fig. 8.3 and Fig. 8.4 show early versions of the interface. Fig. 8.5 Shows an early version of the outcomes view in more detail.

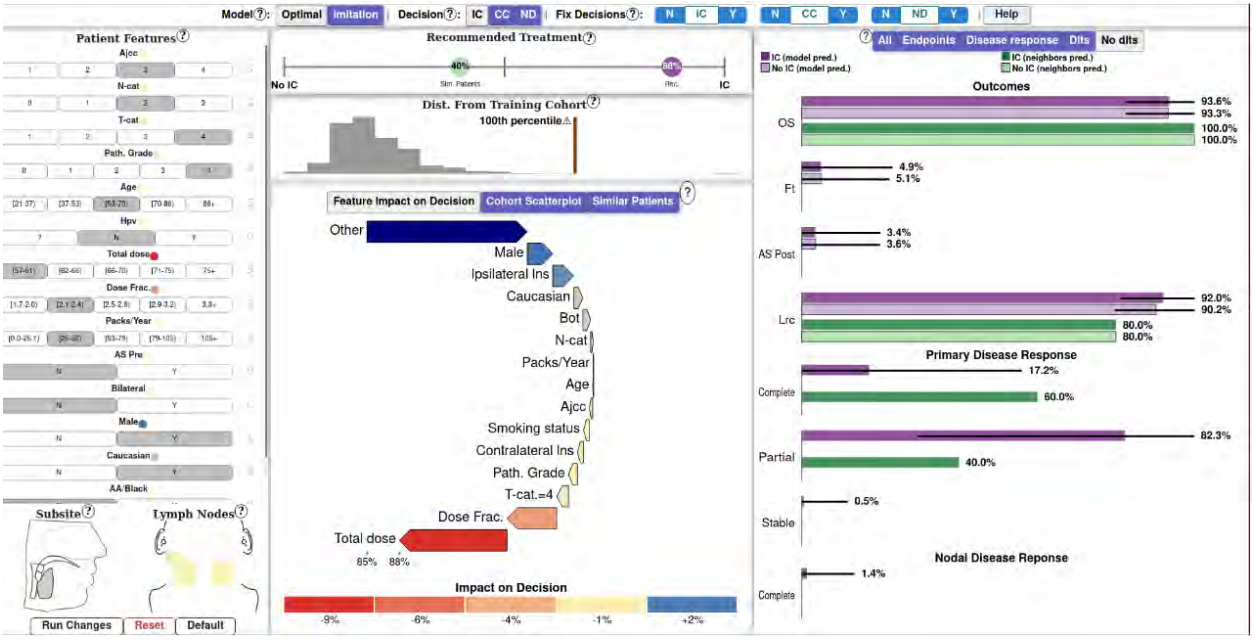


Figure 8.3: Early version of the interface before integrating temporal outcomes. In this version we used a different encoding for treatment recommendation that users found intuitive as it showed the raw model output as percentage of confidence in the patient receiving treatment. This version also showed an additional histogram of the Mahalanobis distances for the cohort. We also used a different color scheme. Additionally, outcomes were shown only as barcharts with a toggle to change the set of outcomes being shown (transition states, DLTs, or 4 year post-treatment outcomes). Model parameters were shown at the top instead of alongside the patient panel.

Figure 8.4: Early version of the interface before the workshop. In this version, the patient input panel was hidden in a "drawer" and could be pulled out via the grey section on the far left, once an initial patient was input. This version includes barcharts with alternative patient outcomes alongside temporal outcomes. Model parameters where shown at the top instead of alongside the patient panel.

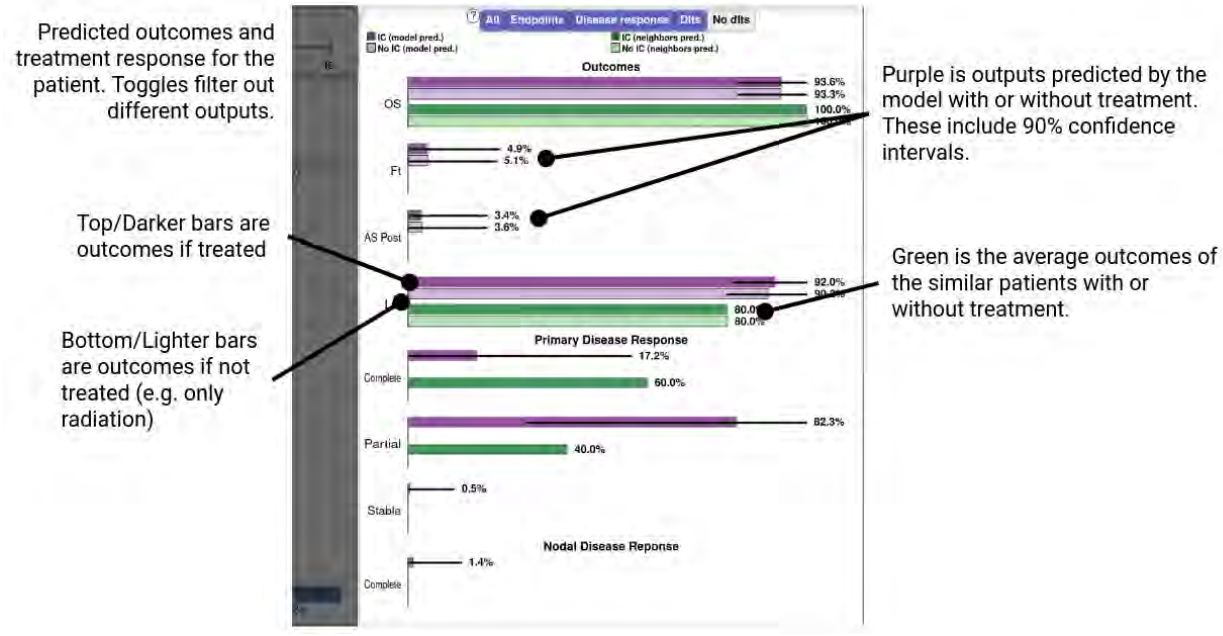


Figure 8.5: Early version of the outcome view. Our original variant used only static outcomes (4 year survival etc) and focused on barcharts of multiple symptoms, based on the original DT model which used binary outcomes only. This was altered after clinicians states that they were used to dealing with temporal risk plots when reasoning about risk profiles, which also required the addition of the Deep survival machine outcome models.

8.6 Appendix F: Copyright Permissions

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer

Andrew Wentzel , Serageldin Attia , Xinhua Zhang , Guadalupe Canahuate , Clifton David Fuller , G. Elisabeta Marai

Transactions on Visualization and Computer Graphics

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.

2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.

4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.

5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.

6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Andrew Wentzel — Signature

06-09-2024 — Date (dd-mm-yyyy)

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration
A Wentzel , P Hanula , T Luciani , B Elgohari , H Elhalawani , G Canahuate , D Vock , CD Fuller , Georgeta-Elisabeta Marai
Transactions on Visualization and Computer Graphics

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Andrew Wentzel — **Signature**

12-08-2019 — **Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Explainable Spatial Clustering: Leveraging Spatial Data in Radiation Oncology
Andrew Wentzel
2020 IEEE Visualization Conference (VIS)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

You have indicated that you DO wish to have video/audio recordings made of your conference presentation under terms and conditions set forth in "Consent and Release."

CONSENT AND RELEASE

1. In the event the author makes a presentation based upon the Work at a conference hosted or sponsored in whole or in part by the IEEE, the author, in consideration for his/her participation in the conference, hereby grants the IEEE the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive, in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IEEE and live or recorded broadcast of the Presentation during or after the conference.
2. In connection with the permission granted in Section 1, the author hereby grants IEEE the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IEEE from any claim based on right of privacy or publicity.

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Andrew Wentzel — **Signature**

14-09-2020 — **Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the

Exclusive License Agreement

COMPUTER GRAPHICS FORUM

Published by John Wiley & Sons Limited on behalf of the Eurographics Association
(together the "Owner")

Date: 4/10/2023

Contributor name: Andrew Wentzel

Contributor address: 842 W Taylor St #2032, Chicago, IL 60607

Manuscript number: 1091

Event/Conference name: EuroVis Conference 2023

Issue no. (if known): 41-3

Re: Manuscript entitled (the "Contribution")
DASS Good: Explainable Data Mining of Spatial Cohort Data

for publication in COMPUTER GRAPHICS FORUM (the "Journal")

published by JOHN WILEY & SONS LIMITED ("Wiley")

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable the Owner to disseminate your Contribution to the fullest extent, we need to have this Exclusive License Agreement executed. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected, this Agreement will be null and void. **Publication cannot proceed without a signed copy of this Agreement.**

A. COPYRIGHT

1. The Contributor grants to the Owner (jointly) an exclusive license of all rights of copyright in the Contribution during the full term of copyright and any extensions or renewals, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so. For the avoidance of doubt, "Contribution" is defined to only include the article submitted by the Contributor for publication in the Journal and does not extend to any supporting information submitted with or referred to in the Contribution ("Supporting Information"). To the extent that any Supporting Information is submitted to the Journal for online hosting by the Journal alongside the Contribution, the Owner is granted a perpetual, non-exclusive license to host and disseminate this Supporting Information for this purpose.

2. Reproduction, posting, transmission or other distribution or use of the final Contribution in whole or in part in any medium by the Contributor as permitted by this Agreement requires a citation to the Journal suitable in form and content as follows: (DOI, Title of Article, Contributor, Journal Title and Volume/ Issue, Copyright © [year], copyright owner as specified in the Journal, Publisher). Links to the final article on the publisher website are encouraged where appropriate.

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's employer, retains all proprietary rights other than an exclusive license of copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution.

C. PERMITTED USES BY CONTRIBUTOR

1. Submitted Version. The Owner licenses back the following rights to the Contributor in the version of the Contribution as originally submitted for publication (the "Submitted Version"):

a. The right to self-archive the Submitted Version on the Contributor's personal website, place in a not for profit subject-based preprint server or repository, or in a Scholarly Collaboration Network (SCN) which has signed up to the STM article sharing principles [http://www.stm-assoc.org/stm-consultations/scn-consultation-2015/](("Compliant SCNs"), or in the Contributor's company/ institutional repository or archive. This right extends to both intranets and the Internet. The Contributor may replace the Submitted Version with the Accepted Version, after any relevant embargo period as set out in paragraph C. 2(a) below has elapsed. The Contributor may wish to add a note about acceptance by the Journal and upon publication it is recommended that Contributors add a Digital Object Identifier (DOI) link back to the Final Published Version.

b. The right to transmit, print and share copies of the Submitted Version with colleagues, including via Compliant SCNs, provided that there is no systematic distribution of the submitted version, e.g. posting on a listserve, network (including SCNs which have not signed up to the STM sharing principles) or automated delivery.

2. Accepted Version. The Owner licenses back the following rights to the Contributor in the version of the Contribution that has been peer-reviewed and accepted for publication (the "Accepted Version"), but not the Final Published Version:

a. The right to self-archive the Accepted Version with Eurographics branding, as hosted on the Eurographics Digital Library http://diglib.eg.org/ , on the Contributor's personal website, in the Contributor's company/institutional repository or archive, in Compliant SCNs, and in not for profit subject-based repositories such as PubMed Central. There are separate arrangements with certain funding agencies governing reuse of the Accepted Version as set forth at the following website: http://www.wiley.com/go/funderstatement. The Contributor may not update the Accepted Version or replace it with the Final Published Version. The Accepted Version posted must contain a legend as follows: This is the accepted version of the following article: FULL CITE, which has been published in final form at http://onlinelibrary.wiley.com. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [http://olabout.wiley.com/WileyCDA/Section/id-820227.html].

b. The right to transmit, print and share copies of the Accepted Version with Eurographics branding with colleagues, including via Compliant SCNs (in private research groups only before the embargo and publicly after), provided that there is no systematic distribution of the Accepted Version, e.g. posting on a listserve, network (including SCNs which have not signed up to the STM sharing principles) or automated delivery.

3. Final Published Version. The Owner hereby licenses back to the Contributor the following rights with respect to the final published version of the Contribution (the "Final Published Version"):

a. *Copies for colleagues*. The personal right of the Contributor only to send or transmit individual copies of the Final Published Version in any format to colleagues upon their specific request, and to share copies in private sharing groups in Compliant SCNs, provided no fee is charged, and further-provided that there is no systematic external or public distribution of the Contribution, e.g. posting on a listserve, network or automated delivery.

b. *Re-use in other publications*. The right to re-use the Final Published Version or parts thereof for any publication authored or edited by the Contributor (excluding journal articles) where such re-used material constitutes less than half of the total material in such publication. In such case, any modifications must be accurately noted.

c. *Teaching duties*. The right to include the Final Published Version in teaching or training duties at the Contributor's institution/place of employment including in course packs, e-reserves, presentation at professional conferences, in-house training, or distance learning. The Final Published Version may not be used in seminars outside of normal teaching obligations (e.g. commercial seminars). Electronic posting of the Final Published Version in connection with teaching/training at the Contributor's company/institution is permitted subject to the implementation of reasonable access control mechanisms, such as user name and password. Posting the Final Published Version on the open Internet is not permitted.

d. *Oral presentations*. The right to make oral presentations based on the Final Published Version.

4. Article Abstracts, Figures, Tables, Artwork and Selected Text (up to 250 words).

a. Contributors may re-use unmodified abstracts for any non-commercial purpose. For on-line uses of the abstracts, the Owner encourages but does not require linking back to the Final Published Version.

b. Contributors may re-use figures, tables, artwork, and selected text up to 250 words from their Contributions, provided the following conditions are met:

(i) Full and accurate credit must be given to the Final Published Version.

(ii) Modifications to the figures and tables must be noted. Otherwise, no changes may be made.

(iii) The re-use may not be made for direct commercial purposes, or for financial consideration to the Contributor.

(iv) Nothing herein will permit dual publication in violation of journal ethical practices.

D. CONTRIBUTIONS OWNED BY EMPLOYER

1. If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire" in the course of employment), the Contribution is owned by the company/institution which must execute this Agreement (in addition to the Contributor's signature). In such case, the company/institution hereby grants an exclusive license to the Owner of all rights of copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph A above.

For company/institution -owned work, signatures cannot be collected electronically and so instead please print off this Agreement, ask the appropriate person in your company/institution to sign the Agreement as well as yourself in the space provided below, and email a scanned copy of the signed Agreement to the Journal production editor. For production editor contact details please visit the Journal's online author guidelines.

2. In addition to the rights specified as retained in paragraph B above and the rights granted back to the Contributor pursuant to paragraph C above, the Owner hereby grants back, without charge, to such company/institution, its subsidiaries and divisions, the right to make copies of and distribute the Final Published Version internally in print format or electronically on the Company's internal network. Copies so used may not be resold or distributed externally. However the company/institution may include information and text from the Final Published Version as part of an information package included with software or other products offered for sale or license or included in patent applications. Posting of the Final Published Version by the company/institution on a public access website may only be done with written permission, and payment of any applicable fee(s). Also, upon payment of the applicable reprint fee, the company/institution may distribute print copies of the Final Published Version externally.

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government, U.K. Government, and other government employees: see notes at end.)

Computer Graphics Forum

Published by Wiley and Eurographics - The European Association for Computer Graphics (together the "Owner")

LICENSE AGREEMENT FOR PUBLISHING CC BY-NC

Date: March 18, 2024

Responsible Corresponding Author (the "Author") name: Andrew Wentzel

Author email address:

Manuscript number: CGF-23-ORA-135.R1

Re: Manuscript or work entitled MOTIV: Visual Exploration of Moral Framing in Social Media (the "Contribution")

for publication in Computer Graphics Forum (the "Journal")

published by John Wiley & Sons Ltd ("Wiley")

Dear Author:

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable Wiley to disseminate your Contribution to the fullest extent, we need to have this License Agreement (the "Agreement") executed. If there are any co-authors of the Contribution ("Co-author"), you must obtain each Co-author's consent to the terms of this Agreement (including the rights granted to Owner) and obtain their signed written permission to execute this Agreement on behalf of the Co-author(s), and you must provide the written permission on request by the Owner or Wiley (where Wiley is not the Owner). If there are no such Co-authors, terms related to Co-author(s) in this Agreement do not apply. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected before publication, this Agreement will be null and void. **Publication cannot proceed without a signed copy of this Agreement and payment of the applicable article publication charge in full (without deduction of any taxes or fees).**

For good and valuable consideration, including the publishing services rendered by Wiley and the mutual covenants and agreements herein, the parties agree as follows:

A. TERMS OF USE

1. The Contribution will be made Open Access under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited and is not used for commercial purposes.

2. For an understanding of what is meant by the terms of the Creative Commons License, please refer to Wiley's Open Access Terms and Conditions (http://www.wileyauthors.com/OAA).

3. If any material contained in the Contribution is the output of Artificial Intelligence Generated Content (AIGC) tools, (a) such tools do not fulfil the role of, nor can they be listed as, an author of Contribution, (b) Author or Co-author will describe its use, transparently and in detail, in the methods, acknowledgement, or equivalent section of the Contribution (provided, however, no such description is needed for tools that are used to improve spelling, grammar, general editing), and (c) Author and each Co-author is responsible for the accuracy of any information provided by any AIGC tool and for referencing any supporting work on which that information depends. The final decision about whether use of an AIGC tool is appropriate or permissible lies with the Journal's editor or other party responsible for the publication's editorial policy.

4. Notwithstanding acceptance, the Owner or Wiley is permitted to require changes to the Contribution, including changes to the length of the Contribution. In addition, the Owner or Wiley is permitted to elect not to publish the Contribution, and/or permitted to retract, withdraw, or publish a correction or other notice for a contribution accepted for publication, if for any reason, in the Owner's or Wiley's reasonable judgment, such publication would be inconsistent with the Core Practices and associated guidelines set forth by the Committee on Publication Ethics (https://publicationethics.org/core-practices) or would result in legal liability, violation of Wiley's ethical guidelines, or violation of journal ethical practices.

5. Once a Contribution has been accepted for publication, an article publication charge ("APC") is due. The Author assumes responsibility for the APC, and no refunds will be issued. If Wiley decides not to publish the Contribution, no APC will be charged and the Author is free to submit the Contribution to any other journal from any other publisher.

B. LICENSE

In addition to the non-exclusive rights to the Contribution the Owner has under the CC BY-NC license, and subject to the full Retained Rights and Permitted Uses in paragraph C below, the Author and each Co-author hereby grants to the Owner, during the full term of the copyright and any extensions or renewals, an exclusive license of all rights of copyright in and to the Contribution that the Author and Co-author do not grant under the CC BY-NC license, and all rights therein, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, for commercial purposes, and to license or permit others to do so. In addition, the Author and each Co-Author hereby grants to the Owner, during the full term of copyright and any extensions or renewals, the exclusive, worldwide, irrevocable and fully transferable right to use and exploit the Contribution in any manner, including: the rights to reproduce, to distribute (for example in any book format or any digital format), to exhibit, and to make available to the public; the recitation performance, and presentation rights; the broadcasting rights; the rights of communication by video or audio recordings; the rights of communication of broadcasts and of works made available to the public. Such exclusive rights do not conflict with the rights granted to users under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0). "Contribution" means the article submitted by the Author for publication in the Journal (including any embedded rich media) and all subsequent versions. The definition of Contribution does not extend to any supporting information submitted with or referred to in the Contribution ("Supporting Information"). To the extent that any Supporting Information is submitted to the Journal, the Owner is granted a perpetual, non-exclusive license to publish, republish, transmit, sell, distribute and otherwise use this Supporting Information in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so. If the Contribution was shared as a preprint or as an accepted manuscript, the Author and each Co-author hereby grant to the Owner exclusivity as to all rights retained by the Author and by each Co-author in the preprint or the accepted manuscript.

C. RETAINED RIGHTS AND PERMITTED USES

1. Retained Rights. The Author and each Co-author or, if applicable, the Author's or Co-author's Employer, retains all proprietary rights, such as copyright (subject to the above-stated license granted to Owner and Creative Commons license), and patent rights in any process, procedure or article of manufacture described in the Contribution.

2. Final Published Version. To the extent the following rights are not permitted for all users under the CC BY-NC license, the Owner hereby licenses back to the Author and each Co-author the following rights with respect to the final published version of the Contribution (the "Final Published Version"):

> **a.** Distribution. The right to non-commercial distribution of the Final Published Version in any format through any means, provided no fee is charged.
>
> **b.** Re-use in other publications. The right to re-use the Final Published Version or parts thereof for any journal or book publication authored or edited by the Author or any Co-author where such re-used material constitutes less than half of the total material in such publication. In such case, any modifications must be accurately noted.
>
> **c.** Teaching duties. The right (and the right to grant colleagues at other academic institutions the right) to include the Final Published Version in teaching duties at the Author's or any Co-author's academic institution, including in academic course packs (which course packs may be sold by a local copy shop for academic courses), e-reserves, society and academic collections, in-house training, or distance learning. The Final Published Version may not be used in seminars outside of normal academic teaching obligations (for example, commercial seminars sponsored by pharmaceutical companies).
>
> **d.** Translations. The right to translate, and authorize their academic colleagues to translate, the Final Published Version for posting on the Author's, Co-author's, or academic colleague's personal website.
>
> **e.** Professional society and academic institution collections: The right to include the Final Published Version in a collection curated by a learned or professional society or academic institution, whether for a conference or another purpose, which may be sold by the society, as long as such collection is not sponsored or otherwise paid for by a commercial entity (for example, a pharmaceutical company).
>
> **f.** Nothing herein will permit dual publication in violation of journal ethical practices.

3. Article Abstracts, Figures, Tables, Artwork and Selected Text (up to 250 words). To the extent the following rights are not permitted for all users under the CC BY-NC license, the Owner hereby licenses back to the Author and each Co-author the following rights.

> **a.** The right to re-use unmodified abstracts for any non-commercial purpose. For online uses of the abstracts, the Owner encourages but does not require linking back to the Final Published Version.
>
> **b.** The right to re-use figures, tables, artwork, and selected text up to 250 words from their Contributions, provided the following conditions are met:
>
> > (i) Full and accurate credit must be given to the Final Published Version.
> >
> > (ii) Modifications to the figures and tables must be noted. Otherwise, no changes may be made.
> >
> > (iii) The re-use may not be made for direct commercial purposes, or for financial consideration to the Author or any Co-author.
> >
> > (iv) The re-use does not constitute dual publication in violation of journal ethical practices.

D. COPYRIGHT NOTICE

The Author, each Co-author, and the company/institution agree that any and all copies of the Final Published Version or any part thereof distributed or posted by them in print or electronic format as permitted will include the notice of copyright as stipulated in the Journal and a full citation to the Final Published Version of the Contribution in the Journal as published by Wiley.

E. CONTRIBUTIONS OWNED BY EMPLOYER

If the Contribution was written by the Author in the course of the Author's employment as a "work-made-for-hire," the Contribution is owned by the company/institution, which must execute this Agreement (in addition to the Author). In such case, the company/institution hereby grants to the Owner, during the full term of copyright, an exclusive license of all rights of copyright in and to the Contribution throughout the world as specified in the License in paragraph B above (and subject to the full Retained Rights and Permitted Uses in paragraph C above, which rights are available to academic institutions that own the Contribution).

F. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires.

(U.S. Government, U.K. Government, and other government employees: see terms at end.)

G. AUTHOR'S REPRESENTATIONS

The Author represents that: (i) if the Contribution has multiple authors, the Author has informed each Co-author of the terms of this Agreement (including the grant of rights to Owner in paragraph B above), has obtained their signed written permission to execute this Agreement on their behalf, and will provide such written permission on request by the Owner; (ii) the Author and each Co-author have the full power, authority and capability to enter into this Agreement, to grant the rights and license granted herein and to perform all obligations hereunder; (iii) neither the Author nor any Co-author has granted exclusive rights to, or transferred their copyright in, any version of the Contribution to any third party; (iv) the Contribution is the Author's and all Co-author's original work, all individuals identified as authors actually contributed to the Contribution, and all individuals who contributed are included; (v) the Contribution is submitted only to this Journal and has not been published before, has not been included in another manuscript, and is not currently under consideration or accepted for publication elsewhere; (vi) if excerpts from copyrighted works owned by third parties are included, the Author shall obtain written permission from the copyright owners for all uses as set forth in the standard permissions form and the Journal's Author Guidelines, and show credit to the sources in the Contribution; (vii) the Contribution and any submitted Supporting Information contain no libelous or unlawful statements, do not infringe upon the rights (including without limitation the copyright, patent or trademark rights) or the privacy of others, do not breach any confidentiality obligation, do not violate a contract or any law, do not contain material or instructions that might cause harm or injury, and only utilize data that has been obtained in accordance with applicable legal requirements and Journal policies; (viii) there are no conflicts of interest relating to the Contribution, except as disclosed; and (ix) if the Author or any Co-author is a resident of either Iran, Syria, Cuba, Crimea, North Korea, Donetsk or Luhansk, the Contribution has been prepared in the relevant resident's personal capacity during the course of their teaching or research work, in other words not as an official representative or otherwise on behalf of their relevant government or institution.

The Author represents that the following information will be clearly identified in the Contribution: (1) all financial and material support for the research and work; (2) any financial interests the Author or each Co-author may have in companies or other entities that have an interest in the information in the Contribution or any submitted Supporting Information (e.g., grants, advisory boards, employment, consultancies, contracts, honoraria, royalties, expert testimony, partnerships, or stock ownership); and (3) indication of no such financial interests if appropriate.

Cited Literature

[1] Treatment option overview for oropharyngeal cancer. https://www.cancer.gov/types/head-and-neck/hp/adult/oropharyngeal-treatment-pdq#_49. accessed July 22, 2024.

[2] Reflections on pandemic visualization. https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/24091. Accessed 2023-03-31.

[3] Framing in communication: From theories to computation. https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/22131?highlight=social%20media. Accessed 2023-03-31.

[4] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 2018.

[5] M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *Proc. ACM-BCB*, pp. 559–560, 2018. doi: 10.1109/ICHI.2018.00095

[6] L. M. Aiello, D. Quercia, K. Zhou, M. Constantinides, S. Šćepanović, and S. Joglekar. How epidemic psychology works on twitter: Evolution of responses to the covid-19 pandemic in the us. *Humanit Soc Sci Commun*, 8, 2021.

[7] A. K. Al-Awami, J. Beyer, D. Haehn, N. Kasthuri, J. W. Lichtman, H. Pfister, and M. Hadwiger. Neuroblocks–visual tracking of segmentation and proofreading for large connectomics projects. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 738–746, 2016.

[8] G. Alicioglu and B. Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520, 2022. doi: 10.1016/j.cag.2021.09.002

[9] B. Alper, N. Riche, G. Ramos, and M. Czerwinski. Design study of linesets, a novel set visualization technique. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2011.

[10] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 2010.

[11] R. J. Amdur, J. G. Li, C. Liu, R. W. Hinerman, and W. M. Mendenhall. Unnecessary laryngeal irradiation in the imrt era. *Head & Neck: J. Sci. Spec. Head & Neck*, pp. 257–264, 2004.

[12] A. B. Amin, R. A. Bednarczyk, C. E. Ray, K. J. Melchiori, J. Graham, J. R. Huntsinger, and S. B. Omer. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880, 2017. doi: 10.1038/s41562-017-0256-5

[13] M. Amin and al. *AJCC Cancer Staging Manual 8th edition*. Wiley Online Library, 12 2016.

[14] D. A. Angulo, C. Schneider, J. H. Oliver, N. Charpak, and J. T. Hernandez. A multi-facetted visual analytics tool for exploratory analysis of human brain and function datasets. *Front. neuroinformatics*, 10:36, 2016. doi: 10.3389/fninf.2016.00036

[15] E. M. Antman, M. Cohen, P. J. Bernink, C. H. McCabe, T. Horacek, G. Papuchis, B. Mautner, R. Corbalan, D. Radley, and E. Braunwald. The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama*, 2000.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Info. Fusion*, 2020. doi: 10.1016/j.inffus.2019.12.012

[17] P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[18] P. C. Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161, 2011.

[19] M. Banerjee, D. Biswas, W. Sakr, and D. P. Wood Jr. Recursive partitioning for prognostic grouping of patients with clinically localized prostate carcinoma. *Cancer: Interdisc. Int. J. American Cancer Soc.*, 2000.

[20] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. PMID: 31313636. doi: 10.1177/1529100619832930

[21] S. Berkovsky, R. Taib, and D. Conway. How to recommend? user trust factors in movie recommender systems. In *Proc. 22nd Int. Conf. on Int. User Inter.*, IUI '17, 14 pages, p. 287–300. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025171.3025209

[22] J. Bernard, T. May, D. Pehrke, T. Schlomm, and J. Kohlhammer. Visual Computing for Big Data Analysis in Prostate Cancer Research. In S. Bruckner and T. Ropinski, eds., *EG 2017 - Dirk Bartz Prize*. The Eurographics Association, 2017. doi: 10.2312/egm.20171044

[23] J. Bernard, D. Sessler, J. Kohlhammer, and R. A. Ruddle. Using dashboard networks to visualize multiple patient histories: a design study on post-operative prostate cancer. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2018.

[24] J. Bernard, D. Sessler, T. May, T. Schlomm, D. Pehrke, and J. Kohlhammer. A visual-interactive system for prostate cancer cohort analysis. *Comp. Grap. and App.*, 35(3):44–55, 2015. doi: 10.1109/MCG.2015.49

[25] J. Beyer, A. Al-Awami, N. Kasthuri, J. W. Lichtman, H. Pfister, and M. Hadwiger. Connectomeexplorer: Query-guided visual analysis of large volumetric neuroscience data. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2868–2877, 2013.

[26] H. Bosch et al. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *Trans. Vis. Comp. Graph.*, 19(12):2022–2031, 2013. doi: 10.1109/TVCG.2013.186

[27] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, 17(12):2301–2309, 2011.

[28] J. Böttger, A. Schäfer, G. Lohmann, A. Villringer, and D. S. Margulies. Three-dimensional mean-shift edge bundling for the visualization of functional connectivity in the brain. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 471–480, 2014.

[29] G. Bouleux, H. B. E. Haouzi, V. Cheutet, G. Demesure, W. Derigent, T. Moyaux, and L. Trilling. Requirements for a Digital Twin for an Emergency Department. In *Proc. SOHOMA*, pp. 130–141. Springer, Cham, Switzerland, February 2023. doi: 10.1007/978-3-031-24291-5_11

[30] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proc. of the Nat. Acad. of Sci.*, 114(28):7313–7318, 2017.

[31] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proc. 5th Work. Beyond Time Errors: Novel Eval. Meth. Vis.*, 2014.

[32] J. Brooke. *SUS – a quick and dirty usability scale*, pp. 189–194. 01 1996.

[33] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Disc. Data Mining*, 2006.

[34] A. A. Bui, D. R. Aberle, and H. Kangarloo. Timeline: visualizing integrated patient records. *IEEE Trans. Info. Techn Biomed.*, 2007.

[35] E. R. Burgess, I. Jankovic, M. Austin, N. Cai, A. Kapuścińska, S. Currie, J. M. Overhage, E. S. Poole, and J. Kaye. Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust. In *Proc. CHI*, CHI '23, article no. 15, 19 pages. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581251

[36] R. Cabello and al. Three. js. *URL: https://github. com/mrdoob/three. js*, 2010.

[37] A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *Confe. on Vis. Anal. Sci. and Tech. (VAST)*, pp. 46–56. IEEE, 2019. doi: 10.1109/VAST47406.2019.8986948

[38] H. B. Caglar, R. B. Tishler, M. Othus, E. Burke, Y. Li, et al. Dose to larynx predicts for swallowing complications after intensity-modulated radiotherapy. *Int. J. Rad. Onco., Bio., & Phys.*, pp. 1110–1118, 2008.

[39] G. Canahuate, A. Wentzel, A. S. R. Mohamed, L. V. van Dijk, D. M. Vock, B. Elgohari, H. Elhalawani, C. D. Fuller, and G. E. Marai. Spatially-aware clustering improves AJCC-8 risk stratification performance in oropharyngeal carcinomas. *Oral Oncol.*, 144:106460, September 2023. doi: 10.1016/j.oraloncology.2023.106460

[40] N. Cao et al. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Trans. Vis. Comp. Graph.*, 18(12):2649–2658, 2012. doi: 10.1109/TVCG.2012.291

[41] N. Cao et al. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *Trans. Vis. Comp. Graph.*, 22(1):280–289, 2016. doi: 10.1109/TVCG.2015.2467196

[42] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2011.

[43] R. Cava, C. M. D. S. Freitas, and M. Winckler. Clustervis: visualizing nodes attributes in multivariate graphs. In *Proc. Symp. App. Comp.*, 2017.

[44] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided visual clustering analysis. *Trans. Vis. Comp. Graph.*, 25(1):267–276, 2018. doi: 10.1109/TVCG.2018.2864477

[45] C.-H. Chang, M. Monselise, and C. C. Yang. What are people concerned about during the pandemic? detecting evolving topics about covid-19 from twitter. *J. Health. Info. Res.*, 5(1):70–97, 2021. doi: 10.1007/s41666-020-00083-3

[46] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conf. App. Comp. Vis. (WACV)*, 2018.

[47] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. In *Comp. Graph. For.*, pp. 223–232, 2003.

[48] E. Chen, K. Lerman, E. Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *Pub. Health and Surv.*, 6(2):e19273, 2020. doi: 10.2196/19273

[49] S. Chen et al. Co-bridges: Pair-wise visual connection and comparison for multi-item data streams. *Trans. Vis. Comp. Graph.*, 2020. doi: 10.1109/TVCG.2020.3030411

[50] S. Chen et al. R-map: A map metaphor for visualizing information reposting process in social media. *Trans. Vis. Comp. Graph.*, 26(1):1204–1214, 2020. doi: 10.1109/TVCG.2019.2934263

[51] S. Chen, L. Lin, and X. Yuan. Social media visual analytics. *Computer Graphics Forum*, 36(3):563–587, 2017. doi: 10.1111/cgf.13211

[52] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, and K. Veeramachaneni. Vbridge: Connecting the dots between features and data to explain healthcare models. *Trans. Vis. Comp. Graph.*, 28(1):378–388, 2021. doi: 10.1109/TVCG.2021.3114836

[53] F. Cheng, Y. Ming, and H. Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *Trans. Vis. Comp. Graph.*, 27(2):1438–1447, 2020. doi: 10.1109/TVCG.2020.3030342

[54] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. Gram: graph-based attention model for healthcare representation learning. In *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Disc. Data Mining*, 2017.

[55] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[56] J. Choi, S.-E. Lee, Y. Lee, E. Cho, S. Chang, and W.-K. Jeong. Dxplorer: A unified visualization framework for interactive dendritic spine analysis using 3d morphological features. *Trans. Vis. Comp. Graph.*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3116656

[57] M. Choi, C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proc. Conf Human Factors Comp. Sys. (CHI)*, 2019.

[58] S.-Y. Chou, S.-W. Lin, and C.-S. Yeh. Cluster identification with parallel coordinates. *Patt. Recog. Letters*, 1999.

[59] N. A. Christakis and J. H. Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

[60] T. Christensen, A. Frandsen, S. Glazier, J. Humpherys, and D. Kartchner. Machine learning methods for disease prediction with claims data. In *IEEE Int. Conf. Health. Inform. (ICHI)*, 2018.

[61] K. M. Christopherson, A. Ghosh, A. S. R. Mohamed, M. Kamal, G. B. Gunn, T. Dale, J. Kalpathy-Cramer, J. Messer, A. S. Garden, H. Elhalawani, et al. Chronic radiation-associated dysphagia in oropharyngeal cancer survivors: Towards age-adjusted dose constraints for deglutitive muscles. *Clinical Transl. Rad. Onco.*, 18:16–22, 2019.

[62] Y.-C. Chu, W.-T. Kuo, Y.-R. Cheng, C.-Y. Lee, C.-Y. Shiau, D.-C. Tarng, and F. Lai. A survival metadata analysis responsive tool (smart) for web-based analysis of patient survival and risk. *Scientific reports*, 8(1):12880, 2018.

[63] M. M. Churpek, T. C. Yuen, C. Winslow, A. A. Robicsek, D. O. Meltzer, R. D. Gibbons, and D. P. Edelson. Multicenter development and validation of a risk stratification tool for ward patients. *American J Respir. Crit. Care Med.*, 2014.

[64] L. Cibulski, E. Dimara, S. Hermawati, and J. Kohlhammer. Supporting domain characterization in visualization design studies with the critical decision method. In *IEEE 4th Workshop on Visualization Guidelines in Research, Design, and Education (VisGuides)*, pp. 8–15, 2022. doi: 10.1109/VisGuides57787.2022.00007

[65] A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv.*, 41(1), 2009. doi: 10.1145/1456650.1456652

[66] S. B. Cohen, E. Ruppin, and G. Dror. Feature selection based on the shapley value. In *IJCAI*, 2005.

[67] A. Corvò, H. S. G. Caballero, and M. A. Westenberg. Survivis: Visual analytics for interactive survival analysis. In *EuroVA@ EuroVis*, pp. 73–77, 2019.

[68] F. Crim. Dear colleague letter on the coronavirus disease 2019 (covid-19). `https://www.nsf.gov/pubs/2020/nsf20052/nsf20052.jsp`, 2020.

[69] D. Danks and A. J. London. Algorithmic bias in autonomous systems. In *Ijcai*, vol. 17, pp. 4691–4697, 2017.

[70] A. Dant and J. Richards. Behind the rumours: How we built our twitter riots interactive. *The Guardian*, 8, 2011. doi: uk/interactive/2011/dec/07/london-riots-twitter

[71] B. Davis, M. Glenski, W. Sealy, and D. Arendt. Measure utility, gain trust: Practical advice for xai researchers. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 1–8, 2020. doi: 10.1109/TREX51495.2020.00005

[72] D. E. Davis, K. Rice, D. R. Van Tongeren, J. N. Hook, C. DeBlaere, E. L. Worthington Jr, and E. Choe. The moral foundations hypothesis does not replicate well in black samples. *Journal of personality and social psychology*, 110(4):e23, 2016. doi: 10.1037/pspp0000056

[73] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, and J. Graham. Purity homophily in social networks. *J. of Experimental Psychology: General*, 145(3):366, 2016.

[74] N. Diakopoulos, A. X. Zhang, D. Elgesem, and A. Salway. Identifying and analyzing moral evaluation frames in climate change blog discourse. In *8th AAAI Conf. on Weblogs & Soc. Med.*, 2014.

[75] Y. Ding, Y. Liu, H. Luan, and M. Sun. Visualizing and understanding neural machine translation. In *Proc. 55th Meet. Assoc. Comp. Linguistics*, vol. 1, 2017.

[76] D. Dingen, M. van't Veer, P. Houthuizen, E. H. Mestrom, E. H. Korsten, A. R. Bouwman, and J. Van Wijk. RegressionExplorer: Interactive exploration of logistic regression models with subgroup analysis. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, 2018.

[77] D. Dingen, M. van't Veer, P. Houthuizen, E. H. J. Mestrom, E. H. Korsten, A. R. Bouwman, and J. van Wijk. Regressionexplorer: Interactive exploration of logistic regression models with subgroup analysis. *Trans. Vis. Comp. Graph.*, 25(1):246–255, 2019. doi: 10.1109/TVCG.2018.2865043

[78] C. Doogan et al. Public perceptions and attitudes toward covid-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of twitter data. *J. Med. Int. Res.*, 22:e21419, 2020. doi: 10.2196/21419

[79] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[80] W. Dou et al. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Trans. Vis. Comp. Graph.*, pp. 93–102, 2012. doi: 10.1109/VAST.2012.6400485

[81] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *Comm. ACM*, 2019.

[82] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *Trans. Vis. Comp. Graph.*, 16(6):1129–1138, 2010. doi: 10.1109/TVCG.2010.129

[83] A. Ebrahimi Zade, S. Shahabi Haghighi, and M. Soltani. Deep neural networks for neuro-oncology: Towards patient individualized design of chemo-radiation therapy for glioblastoma patients. *Journal of Biomedical Informatics*, 127:104006, 2022. doi: 10.1016/j.jbi.2022.104006

[84] A. Efrat, Y. Hu, S. G. Kobourov, and S. Pupyrev. Mapsets: visualizing embedded and clustered graphs. In *Inte. Symp. Graph Draw.* Springer, 2014.

[85] I. El Naqa, J. D. Bradley, and J. O. Deasy. Nonlinear kernel-based approaches for predicting normal tissue toxicities. In *IEEE Int. Conf. on Mach. Learn. and App.*, pp. 539–544, 2008. doi: 10.1109/ICMLA.2008.126

[86] I. El Naqa, G. Suneja, P. Lindsay, A. Hope, J. Alaly, et al. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose–volume outcome relationships. *Phys. Med. & Bio.*, p. 5719, 2006.

[87] H. Elhalawani, T. A. Lin, S. Volpe, A. S. Mohamed, A. L. White, J. Zafereo, A. J. Wong, J. E. Berends, S. AboHashem, B. Williams, et al. Machine learning applications in head and neck radiation oncology: lessons from open-source radiomics challenges. *Front. Onco.*, 8:294, 2018.

[88] J. Eligon. One slogan, many methods: Black lives matter enters politics. *The New York Times*, 18, 2015.

[89] B. Emami, J. Lyman, A. Brown, L. Cola, M. Goitein, J. Munzenrider, B. Shank, L. Solin, and M. Wesson. Tolerance of normal tissue to therapeutic irradiation. *Int. Jour. Rad. Onco. Bio. Phys.*, 21(1):109–122, 1991. doi: 10.1016/0360-3016(91)90171-y

[90] R. M. Entman and A. Rojecki. Freezing out the public: Elite and media framing of the us anti-nuclear movement. *Political Communication*, 1993.

[91] S. Evergreen and C. Metzner. Design principles for data visualization in evaluation. *New Directions for Evaluation*, 2013.

[92] M. Falk, A. Ynnerman, D. Treanor, and C. Lundström. Interactive visualization of 3d histopathology in native resolution. *Trans. Vis. Comp. Graph.*, 25(1):1008–1017, 2018. doi: 10.1109/TVCG.2018.2864816

[93] Z. Fatemi, A. Bhattacharya, E. Zheleva, B. Di Eugenio, V. Dhariwal, L. Levine, A. Rojecki, A. Wentzel, and G. Marai. Understanding stay-at-home attitudes through framing analysis of tweets. In *Proc. DSAA*, pp. 1–10, 8 2022. doi: 10.1109/DSAA54385.2022.10032455

[94] M. J. Fehrenbach and S. W. Herring. *Illustrated Anatomy of the Head and Neck*, vol. 5. Elsevier Health Sciences, 2015.

[95] C. Floricel, N. Nipu, M. Biggs, A. Wentzel, G. Canahuate, L. Van Dijk, A. Mohamed, C. D. Fuller, and G. E. Marai. Thalis: Human-machine analysis of longitudinal symptoms in cancer therapy. *Trans. Vis. Comp. Graph.*, 28(1):151–161, 2021. doi: 10.1109/TVCG.2021.3114810

[96] C. Floricel, N. Nipu, M. Biggs, A. Wentzel, G. Canahuate, L. Van Dijk, A. Mohamed, C. D. Fuller, and G. E. Marai. THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy. *Trans. Vis. Comp. Graph.*, 28(01):151–161, January 2022. doi: 10.1109/TVCG.2021.3114810

[97] C. Floricel, A. Wentzel, A. Mohamed, D. Fuller, G. Canahuate, and G. E. Marai. Roses have thorns: Understanding the downside of oncological care delivery through visual analytics and sequential rule mining. *IEEE transactions on visualization and computer graphics*, 2024.

[98] L. Floridi and J. Cowls. A Unified Framework of Five Principles for AI in Society. *Harvard Data Sci. Rev.*, 1(1), jul 1 2019. doi: 10.1162/99608f92.8cd550d1

[99] D. Frey and R. Pimentel. Principal component analysis and factor analysis. *Quant. Ethology*, pp. 219–245, 1978.

[100] K. Furmanová, N. Grossmann, L. P. Muren, O. Casares-Magaz, V. Moiseenko, J. P. Einck, M. E. Gröller, and R. G. Raidou. Vapor: visual analytics for the exploration of pelvic organ variability in radiotherapy. *Comp. & Graph.*, 91:25–38, 2020. doi: 10.1016/j.cag.2020.07.001

[101] K. Furmanová, L. P. Muren, O. Casares-Magaz, V. Moiseenko, J. P. Einck, S. Pilskog, and R. G. Raidou. Previs: Predictive visual analytics of anatomical variability for radiotherapy decision support. *Comp. & Graph.*, 97:126–138, 2021. doi: 10.1016/j.cag.2021.04.010

[102] A. Gafita, J. Calais, T. R. Grogan, B. Hadaschik, H. Wang, M. Weber, S. Sandhu, C. Kratochwil, R. Esfandiari, R. Tauber, et al. Nomograms to predict outcomes after 177lu-psma therapy in men with metastatic castration-resistant prostate cancer: an international, multicentre, retrospective study. *The Lancet Oncology*, 22(8):1115–1125, 2021.

[103] B. F. Gage, A. D. Waterman, W. Shannon, M. Boechler, M. W. Rich, and M. J. Radford. Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. *Jama*, 2001.

[104] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

[105] J. Giersch. Punishing campus protesters based on ideology. *Research & Politics*, 6(4):2053168019892129, 2019.

[106] M. Glueck, A. Gvozdik, F. Chevalier, A. Khan, M. Brudno, and D. Wigdor. Phenostacks: Cross-sectional cohort phenotype comparison visualizations. *Trans. Vis. Comp. Graph.*, 23(1):191–200, 2017. doi: 10.1109/TVCG.2016.2598469

[107] M. Glueck, M. P. Naeini, F. Doshi-Velez, F. Chevalier, A. Khan, D. Wigdor, and M. Brudno. Phenolines: Phenotype comparison visualizations for disease subtyping via topic models. *Trans. Vis. Comp. Graph.*, 24(1):371–381, 2017. doi: 10.1109/TVCG.2017.2745118

[108] Google. Google+ ripples. `https://services.google.com/fh/files/misc/ripples.pdf`, 2012.

[109] J. Graham et al. Moral foundations theory: The pragmatic validity of moral pluralism. In *Adv. in Exp. Soc. Psyc.*, vol. 47, pp. 55–130. Elsevier, 2013. doi: 10.1016/B978-0-12-407236-7.00002-4

[110] J. Graham, J. Haidt, M. Motyl, P. Meindl, C. Iskiwitch, and M. Mooijman. *Moral foundations theory.* Guilford Publications, 2018.

[111] M. M. Graham, M. T. James, and J. A. Spertus. Decision support tools: Realizing the potential to improve quality of care. *Canadian Journal of Cardiology*, 34(7):821–826, 2018. doi: 10.1016/j.cjca.2018.02.029

[112] D. L. Gresh, B. E. Rogowitz, R. L. Winslow, D. F. Scollan, and C. K. Yung. Weave: A system for visually linking 3-d and statistical visualizations, applied to cardiac simulation and measurement data. In *IEEE Vis.*, pp. 489–492, 2000. doi: 10.1109/VISUAL.2000.885739

[113] N. Grossmann, O. Casares-Magaz, L. P. Muren, V. Moiseenko, J. P. Einck, M. E. Gröller, and R. G. Raidou. Pelvis runner: Visualizing pelvic organ variability in a cohort of radiotherapy patients. In *Eurographics Work. on Vis. Comp. for Bio. and Med.*, pp. 69–78, 2019. doi: 10.2312/vcbm.20191233

[114] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local rule-based explanations of black box decision systems. *arXiv*, 2018. doi: 1805.10820

[115] D. Gunning. Darpa's explainable artificial intelligence (xai) program. In *Proc. 24th Int. Conf. on Int. User Inter.*, 1 pages, p. ii. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3301275.3308446

[116] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao. Eventthread: Visual summarization and stage analysis of event sequence data. *Trans. Vis. Comp. Graph.*, 24(1):56–65, 2017. doi: 10.1109/TVCG.2017.2745320

[117] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz, and N. Cao. Survey on visual analysis of event sequence data. *IEEE Trans. Vis. Comp. Graph.*, 28(12):5091–5112, 2022. doi: 10.1109/TVCG.2021.3100413

[118] S. Ha, S. Monadjemi, and A. Ottley. Guided By AI: Navigating Trust, Bias, and Data Exploration in AI-Guided Visual Analytics. *Computer Graphics Forum*, 43(3):e15108, June 2024. doi: 10.1111/cgf.15108

[119] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *Trans. Vis. and Comp. Graph.*, 13(6):1270–1277, 2007. doi: 10.1109/TVCG.2007.70623

[120] A. Hakone, L. Harrison, A. Ottley, N. Winters, C. Gutheil, P. K. J. Han, and R. Chang. Proact: Iterative design of a patient-centered visualization for effective prostate cancer health risk communication. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):601–610, 2017. doi: 10.1109/TVCG.2016.2598588

[121] T. Hastie and R. Tibshirani. *Generalized Additive Models*, vol. 1. Institute of Mathematical Statistics, 1986. doi: 10.1214/ss/1177013604

[122] I. Heimbach, B. Schiller, T. Strufe, and O. Hinz. Content virality on online social networks: Empirical evidence from twitter, facebook, and google+ on german news websites. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 39–47, 2015. doi: 10.1145/2700171.2791032

[123] F. Hohman et al. Gamut: A design probe to understand how data scientists understand machine learning models. In *Conf. Hum. Fact. Comp. Sys.*, 2019. doi: 10.1145/3290605.3300809

[124] D. Holliday, S. Wilson, and S. Stumpf. User trust in intelligent systems: A journey over time. In *Proc. 21st Int. Conf. on Int. User Inter.*, IUI '16, 5 pages, p. 164–168. Association for Computing Machinery, 2016. doi: 10.1145/2856767.2856811

[125] J. Hoover et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Soc. Psy. and Pers. Sci.*, 11(8):1057–1071, 2020. doi: 10.1177/1948550619876629

[126] J. Hoover, K. Johnson, R. Boghrati, J. Graham, M. Dehghani, and M. B. Donnellan. Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, 4(1), 2018. doi: 10.1525/collabra.129

[127] M. Hu, K. Wongsuphasawat, and J. Stasko. Visualizing social media content with sententree. *Trans. Vis. Comp. Graph.*, 23(1):621–630, 2017. doi: 10.1109/TVCG.2016.2598590

[128] S. H. Huang and B. O'Sullivan. Overview of the 8th edition tnm classification for head and neck cancer. *Curr. Treat. Op. Onco.*, 2017.

[129] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. Int. AAAI Conf. on Web & Soc. Med.*, vol. 8, 2014.

[130] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Dec. Supp. Sys.*, 2011.

[131] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Comp.-Aided Des.*, pp. 509–530, 2005.

[132] R. Iyer, S. Koleva, J. Graham, P. Ditto, and J. Haidt. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS One*, 2012. doi: 10.1371/journal.pone.0042366

[133] M. Jacobs, J. He, M. F. Pradier, B. Lam, A. C. Ahn, et al. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proc. CHI*, pp. 1–14. ACM, May 2021. doi: 10.1145/3411764.3445385

[134] W. M. Jainek, S. Born, D. Bartz, W. Straßer, and J. Fischer. Illustrative hybrid visualization and exploration of anatomical and functional brain data. *Comp. Graph. For.*, 27(3):855–862, 2008. doi: 10.1111/j.1467-8659.2008.01217.x

[135] H. Jang et al. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. *J. Med. Int. Res.*, 23(2):e25431, 2021. doi: 10.2196/25431

[136] S. Jänicke, P. Kaur, P. Kuzmicki, and J. Schmidt. Participatory visualization design as an approach to minimize the gap between research and application. In *Gap bet. Vis. Res. Vis. Soft. (VisGap)*. The Eurographics Association, 2020.

[137] R. Janoff-Bulman and N. C. Carnes. Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review*, 17(3):219–236, 2013.

[138] J. Jessup, R. Krueger, S. Warchol, J. Hoffer, J. Muhlich, C. C. Ritch, G. Gaglia, S. Coy, Y.-A. Chen, J.-R. Lin, S. Santagata, P. K. Sorger, and H. Pfister. Scope2screen: Focus+context techniques for pathology tumor assessment in multivariate image data. *Trans. Vis. Comp. Graph.*, 28(1):259–269, 2022. doi: 10.1109/TVCG.2021.3114786

[139] H. Jiang, B. Kim, M. Guan, and M. Gupta. To trust or not to trust a classifier. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 31, pp. 5541–5552. Curran Associates, Inc., 2018.

[140] R. Jianu, C. Demiralp, and D. Laidlaw. Exploring 3d dti fiber tracts with linked 2d representations. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1449–1456, 2009.

[141] John Hopkins University Center for Systems Science and Engineering. `https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data`, 2020.

[142] C. Johnson. Top scientific visualization research problems. *IEEE Cpmp. Graph. App. (CGA)*, pp. 13–17, 7 2004.

[143] D. Jönsson, A. Bergström, C. Forsell, R. Simon, M. Engström, S. Walter, A. Ynnerman, and I. Hotz. Visualneuro: A hypothesis formation and reasoning application for multi-variate brain cohort study data. *Comp. Graph. For.*, 39(6):392–407, 2020. doi: 10.1111/cgf.14045

[144] D. Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[145] M. E. Kaminski. The right to explanation, explained. *Berkeley Tech. LJ*, 34:189, 2019. doi: 10.2139/ssrn.3196985

[146] M. Karabacak, P. Jagtiani, A. Carrasquilla, I. M. Germano, and K. Margetis. Prognosis individualized: Survival predictions for who grade ii and iii gliomas with a machine learning-based web application. *NPJ Digital Medicine*, 6(1):200, 2023. doi: 10.1038/s41746-023-00948-y

[147] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. L. Arendt, S. Shaikh, and W. Dou. Vulnerable to Misinformation? Verifi! In *Proc. 24th Int. Conf. on Int. User Inter.* ACM, 2019. doi: 10.1145/3301275.3302320

[148] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proc. CHI*, 14 pages, p. 1–14, 2020. doi: 10.1145/3313831.3376219

[149] C. Keogh, E. Wallace, C. Dillon, B. D. Dimitrov, and T. Fahey. Validation of the chads2 clinical prediction rule to predict ischaemic stroke. *Thrombosis and haemostasis*, 2011.

[150] S. Kim et al. Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. *Trans. Vis. Comp. Graph.*, 24(3):1287–1300, 2018. doi: 10.1109/TVCG.2017.2666146

[151] P. Klemm, S. Oeltze-Jafra, K. Lawonn, K. Hegenscheid, H. Völzke, and B. Preim. Interactive visual analysis of image-centric cohort study data. *Trans. Vis. Comp. Graph.*, 20(12):1673–1682, 2014. doi: 10.1109/TVCG.2014.2346591

[152] J. Knittel, S. Koch, T. Tang, W. Chen, Y. Wu, S. Liu, and T. Ertl. Real-time visual analysis of high-volume social media posts. *Trans. Vis. Comp. Graph.*, 28(1):879–889, 2021. doi: 10.48550/arXiv.2108.03052

[153] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *Proc. the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[154] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[155] S. Koleva et al. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Res. in Pers.*, 46(2):184–194, 2012. doi: 10.1016/j.jrp.2012.01.006

[156] S. Konishi and G. Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

[157] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[158] J. Krause, A. Perer, and H. Stavropoulos. Supporting iterative cohort construction with visual temporal queries. *Trans. Vis. Comp. Graph.*, 22(1):91–100, 2016. doi: 10.1109/TVCG.2015.2467622

[159] K. Kucher, R. M. Martins, C. Paradis, and A. Kerren. Stancevis prime: Visual analysis of sentiment and stance in social media texts. *J. Vis.*, 23(6):1015–1034, 20 pages, dec 2020. doi: 10.1007/s12650-020-00684-5

[160] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symp. on Vis. Lang. and Hu. Cent. Comp.*, pp. 3–10, 2013. doi: 10.1109/VLHCC.2013.6645235

[161] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng. Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data. *J. of Dig. Imag.*, pp. 1025—1039, 2013.

[162] M. B. Kursa, W. R. Rudnicki, et al. Feature selection with the boruta package. *J Stat Softw*, 2010.

[163] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, 2019. doi: 10.1109/TVCG.2018.2865027

[164] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *Trans. Vis. Comp. Graph.*, 24(1):142–151, 2017. doi: 10.1109/TVCG.2017.2745085

[165] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, A. Perer, et al. Clustervision: Visual supervision of unsupervised clustering. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 142–151, 2018.

[166] B. C. Kwon, U. Kartoun, S. Khurshid, M. Yurochkin, S. Maity, D. G. Brockman, A. V. Khera, P. T. Ellinor, S. A. Lubitz, and K. Ng. Rmexplorer: A visual analytics approach to explore the performance and the fairness of disease risk models on population subgroups. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 50–54, 2022. doi: 10.1109/VIS54862.2022.00019

[167] A. Köhn, F. Weiler, J. Klein, O. Konrad, H. Hahn, and H.-O. Peitgen. State-of-the-Art Computer Graphics in Neurosurgical Planning and Risk Assessment. In P. Cignoni and J. Sochor, eds., *EG Short Papers*. The Eurographics Association, 2007. doi: 10.2312/egs.20071048

[168] C. Lacave and F. J. Díez. A review of explanation methods for bayesian networks. *Knowl. Eng. Review*, 2002.

[169] K. R. Lamborn, S. M. Chang, and M. D. Prados. Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis. *Neuro-onco.*, 2004.

[170] J. A. Langendijk, P. Doornaert, I. M. Verdonck-de Leeuw, C. R. Leemans, N. K. Aaronson, and B. J. Slotman. Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. *Jour. Clin. Onco.*, 26(22):3770–3776, 2008. doi: 10.1200/JCO.2007.14.6647

[171] M. W. Lauer-Schmaltz, I. Kerim, J. P. Hansen, G. M. Gulyás, and H. B. Andersen. Human digital twin-based interactive dashboards for informal caregivers of stroke patients. In *Proc. PETRA*, PETRA '23, 7 pages, p. 215–221. ACM, 2023. doi: 10.1145/3594806.3594824

[172] J. Lawrance, M. Burnett, R. Bellamy, C. Bogart, and C. Swart. Reactive information foraging for evolving goals. In *Proc. SIGCHI Conf. on Hu. Fac. in Comp. Sys.*, CHI '10. ACM, New York, NY, USA, 2010. doi: 10.1145/1753326.1753332

[173] B. Lee, K. Isaacs, D. A. Szafir, G. E. Marai, C. Turkay, M. Tory, S. Carpendale, and A. Endert. Broadening intellectual diversity in visualization research papers. *IEEE Comp. Graph. and App.*, 39(4):78–85, 2019. doi: 10.1109/MCG.2019.2914844

[174] C. R. Leemans, R. Tiwari, J. J. Nauta, I. V. D. Waal, and G. B. Snow. Recurrence at the primary site in head and neck cancer and the significance of neck lymph node metastases as a prognostic factor. *Cancer*, 1994.

[175] K.-M. Leung, R. M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104, 1997.

[176] J. Li, M. Wang, M. Won, E. G. Shaw, C. Coughlin, W. J. Curran Jr, and M. P. Mehta. Validation and simplification of the radiation therapy oncology group recursive partitioning analysis classification for glioblastoma. *Int. J. Rad. Onco., Bio., Phys.*, 2011.

[177] X. Li, V. Krivtsov, and K. Arora. Attention-based deep survival model for time series data. *Reliability Engineering & System Safety*, 217:108033, 2022. doi: 10.1016/j.ress.2021.108033

[178] C. C. Ling, C. Burman, C. S. Chui, G. J. Kutcher, and Leibel. Conformal radiation treatment of prostate cancer using inversely-planned intensity-modulated photon beams produced with dynamic multileaf collimation. *Int. J. Rad. Onco., Bio., Phys.*, pp. 721–730, 1996.

[179] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 27 pages, June 2018. doi: 10.1145/3236386.3241340

[180] S. Liu, D. Wang, D. Maljovec, R. Anirudh, J. J. Thiagarajan, S. A. Jacobs, B. C. Van Essen, D. Hysom, J.-S. Yeom, J. Gaffney, et al. Scalable topological data analysis and visualization for evaluating data-driven models in scientific applications. *IEEE Trans. Vis. Comp. Graph.*, 2019.

[181] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale. Timespan: Using visualization to explore temporal multi-dimensional data of stroke patients. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2015.

[182] B. Lorensen. On the death of visualization. In *Proc. Work. Vis. Res. Chal.*, 2004.

[183] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *Proc. 19th ACM SIGKDD int. conf. on Knowl. disc. and data mining*, pp. 623–631, 2013.

[184] M. Louise Davies. A New Personalized Oral Cancer Survival Calculator to Estimate Risk of Death From Both Oral Cancer and Other. *JAMA Otolar. Head. Neck Surg.*, 149(11):993–1000, November 2023. doi: 10.1001/jamaoto.2023.1975

[185] T. Luciani, A. Burks, C. Sugiyama, J. Komperda, and G. E. Marai. Details-first, show context, overview last: Supporting exploration of viscous fingers in large-scale ensemble simulations. *Trans. Vis. Comp. Graph.*, 2019. doi: 10.1109/TVCG.2018.2864849

[186] T. Luciani, A. Wentzel, B. Elgohari, H. Elhalawani, A. Mohamed, G. Canahuate, D. M. Vock, C. D. Fuller, and G. E. Marai. A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine. *J. of biomed. informatics*, 112:100067, 2020. doi: 10.1016/j.yjbinx.2020.100067

[187] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intell.*, 2020.

[188] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Adv. in Neu. Info. Proc. Sys.*, vol. 30, pp. 4765–4774. Curran Associates, Inc., 2017.

[189] B. Ma and A. Entezari. An interactive framework for visualization of weather forecast ensembles. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2018.

[190] C. Ma, F. Pellolio, D. A. Llano, K. A. Stebbings, R. V. Kenyon, and G. E. Marai. Rembrain: Exploring dynamic biospatial networks with mosaic matrices and mirror glyphs. *Elec. Imag.*, pp. 1–13, 2018. doi: 10.2352/J.ImagingSci.Technol.2017.61.6.060404

[191] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 833–840, 2020.

[192] R. G. C. Maack, D. Saur, H. Hagen, G. Scheuermann, and C. Gillman. Towards Closing the Gap of Medical Visualization Research and Clinical Daily Routine. In C. Gillmann, M. Krone, G. Reina, and T. Wischgoll, eds., *Gap bet. Vis. Res. Vis. Soft. (VisGap)*. The Eurographics Association, 2020.

[193] S. Maben, P. Endres-Parnell, and L. Helvie-Mason. Re (claiming) and re (naming)# mydallasis: An analysis of citizen framing of dallas after the 2016 ambush. *Southwestern Mass Communication Journal*, 35(1), 2019.

[194] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *20th Int Conf on Int. User Int.*, pp. 38–49, 2015. doi: 10.1145/2678025.2701407

[195] M. Mantovani, A. Wentzel, J. T. Trabucco, J. Michaelis, and G. E. Marai. Kiviat Defense: An Empirical Evaluation of Visual Encoding Effectiveness in Multivariate Data Similarity Detection. *J. Imag.Sci. Tech.*, 67:1–13, November 2023. doi: 10.2352/J.ImagingSci.Technol.2023.67.6.060406

[196] Y. Mao, D. Wang, M. Muller, K. R. Varshney, I. Baldini, C. Dugan, and A. Mojsilović. How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proc. ACM Human-Comp. Int.*, 3:1–23, 2019. doi: 10.1145/3361118

[197] G. Marai. Visual Scaffolding in Integrated Spatial and Nonspatial Visual Analysis. In *Int. EuroVis Work. Vis. Anal.*, pp. 1–5, 2015.

[198] G. E. Marai. Visual scaffolding in integrated spatial and nonspatial analysis. In *EuroVis Works. Vis. Ana. (EuroVA)*. The Eurographics Association, 2015. doi: 10.2312/eurova.20151097

[199] G. E. Marai. Activity-centered domain characterization for problem-driven scientific visualization. *Trans. Vis. Comp. Graph.*, 24(1):913–922, 2017. doi: 10.1109/TVCG.2017.2744459

[200] G. E. Marai. Activity-centered domain characterization for problem-driven scientific visualization. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 913–922, 2018.

[201] G. E. Marai, C. Ma, A. T. Burks, F. Pellolio, G. Canahuate, D. M. Vock, A. S. Mohamed, and C. D. Fuller. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *Trans. Vis. Comp. Graph.*, 25(4):1732–1745, 2018. doi: 10.1109/TVCG.2018.2817557

[202] G. E. Marai, C. Ma, A. T. Burks, F. Pellolio, G. Canahuate, D. M. Vock, A. S. R. Mohamed, and C. D. Fuller. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1732–1745, 4 2019. doi: 10.1109/TVCG.2018.2817557

[203] G. E. Marai, B. Pinaud, K. Bühler, A. Lex, and J. H. Morris. Ten simple rules to create biological network figures for communication, 2019. doi: doi.org/10.1371/journal.pcbi.1007244

[204] A. Marcus et al. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Conf. Hum. Fact. Comp. Sys.*, p. 227–236, 2011. doi: 10.1145/1978942.1978975

[205] A. Maries, N. Mays, M. O. Hunt, K. F. Wong, Layton, and G. E. Marai. Grace: A visual comparison framework for integrated spatial and non-spatial geriatric data. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2916–2925, 2013.

[206] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *2nd ACM SIGHIT Int. Health Info. Symp.*, p. 389–398. Association for Computing Machinery, 2012. doi: 10.1145/2110363.2110408

[207] J. S. Marwaha and J. C. Kvedar. Crossing the chasm from model performance to clinical impact: the need to improve implementation and evaluation of ai, 2022. doi: 10.1038/s41746-022-00572-2

[208] K. Matkovic, W. Freiler, D. Gracanin, and H. Hauser. Comvis: A coordinated multiple views system for prototyping new visualization technology. In *Proc. IEEE. Int. Conf. Info. Vis. (InfoVis)*, 7 2008.

[209] J. W. S. McCullough, R. A. Richardson, A. Patronis, R. Halver, R. Marshall, et al. Towards blood flow in the virtual human: efficient self-coupling of HemeLB. *Interface Focus*, 11(1):20190119, February 2021. doi: 10.1098/rsfs.2019.0119

[210] N. McCurdy, J. Dykes, and M. Meyer. Action design research and visualization design. In *Proc. Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Vis.*, BELIV '16, 9 pages, p. 10–18. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2993901.2993916

[211] S. McGregor, H. Buckingham, T. G. Dietterich, R. Houtman, C. Montgomery, and R. Metoyer. Facilitating testing and debugging of markov decision processes with interactive visualization. In *Proc. IEEE Symp. Visual Lang. Human-Centric Comp. (VL/HCC)*, 10 2015.

[212] S. McKenna, D. Mazur, J. Agutter, and M. Meyer. Design activity framework for visualization design. *Trans. Vis. Comp. Graph.*, 20(12):2191–2200, 2014. doi: 10.1109/TVCG.2014.2346331

[213] W. McKinney et al. Data structures for statistical computing in python. *Proc. 9th Python in Sci. Conf.*, pp. 51–56, 2010.

[214] W. M. Mendenhall, R. J. Amdur, and J. R. Palta. Intensity-modulated radiotherapy in the standard management of head and neck cancer: promises and pitfalls. *J. Clin. Onco.*, pp. 2618–2623, 2006.

[215] T. Metsalu and J. Vilo. Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic acids research*, 2015.

[216] M. Meyer and J. Dykes. Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):87–97, 2020. doi: 10.1109/TVCG.2019.2934539

[217] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *Trans. Vis. Comp. Graph.*, 25(1):342–352, 2018. doi: 10.1109/TVCG.2018.2864812

[218] Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren. Protosteer: Steering deep sequence model with prototypes. *trans. vis. comp. graph.*, 26(1):238–248, 2019. doi: 10.1109/TVCG.2019.2934267

[219] A. Mishra, S. Ginjpalli, and C. Bryan. News kaleidoscope: Visual investigation of coverage diversity in news event reporting. In *Pacific Vis.*, pp. 131–140. IEEE Computer Society, Los Alamitos, CA, USA, apr 2022. doi: 10.1109/PacificVis53943.2022.00022

[220] G. Mistelbauer, K. Bäumler, D. Mastrodicasa, L. D. Hahn, A. Pepe, V. Sandfort, V. Hinostroza, K. Ostendorf, A. Schroeder, A. M. Sailer, M. J. Willemink, S. Walters, B. Preim, and D. Fleischmann. Transdisciplinary Visualization of Aortic Dissections. In R. Raidou and T. Kuhlen, eds., *EuroVis 2023 - Dirk Bartz Prize*. The Eurographics Association, 2023. doi: 10.2312/evm.20231085

[221] MIT Election Data and Science Lab. `https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md`, 2020.

[222] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Inter. Intell. Sys. (TiiS)*, 11(3-4):1–45, 2021.

[223] M. Mooijman, J. Hoover, Y. Lin, H. Ji, and M. Dehghani. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396, 2018. doi: 10.1038/s41562-018-0353-0

[224] R. P. Moreno, P. G. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.-R. Le Gall, et al. Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive care med.*, 2005.

[225] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *Trans. Vis. Comp. Graph.*, 19(12):1962–1971, 2013. doi: 10.1109/TVCG.2013.125

[226] M. Müller, M. Petzold, M. Wunderlich, T. Baumgartl, M. Höhn, V. Eichel, N. Mutters, S. Scheithauer, M. Marschollek, and T. von Landesberger. Visual analysis for hospital infection control using a rnn model. In *EuroVis Work. Vis. Ana. (EuroVA)*. The Eurographics Association, 2020.

[227] T. Munzner. A nested model for visualization design and validation. *Trans. Vis. Comp. Graph.*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111

[228] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proc Nat. Acad. of Sci.*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116

[229] A. Mustaqeem, S. M. Anwar, and M. Majid. A modular cluster based collaborative recommender system for cardiac patients. *Art. Intel. Med.*, 102:101761, 2020. doi: 10.1016/j.artmed.2019.101761

[230] J. Müller-Sielaff, S. B. Beladi, S. W. Vrede, M. Meuschke, P. J. F. Lucas, J. M. A. Pijnenborg, and S. Oeltze-Jafra. Visual assistance in development and validation of bayesian networks for clinical decision support. *IEEE Transactions on Visualization and Computer Graphics*, 29(8):3602–3616, 2023. doi: 10.1109/TVCG.2022.3166071

[231] A. O. Naghavi, M. I. Echevarria, T. J. Strom, Y. A. Abuodeh, et al. Treatment delays, race, and outcomes in head and neck cancer. *Cancer Epidem.*, 45:18–25, December 2016. doi: 10.1016/j.canep.2016.09.005

[232] C. Nagpal, X. Li, and A. Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J. of Biomed. and Hea. Info.*, 25(8):3163–3175, 2021.

[233] C. Nagpal, W. Potosnak, and A. Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. *arXiv preprint arXiv:2204.07276*, 2022.

[234] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller. Deep cox mixtures for survival regression. In K. Jung, S. Yeung, M. Sendak, M. Sjoding, and R. Ranganath, eds., *Proceedings of the 6th Machine Learning for Healthcare Conference*, vol. 149 of *Proceedings of Machine Learning Research*, pp. 674–708. PMLR, 06–07 Aug 2021.

[235] S. Nakagawa and I. C. Cuthill. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews*, 2007.

[236] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By. Sentiment analysis on social media. In *Int. Conf. Adv. Soc. Net. Anal. & Min*, pp. 919–926, 2012. doi: 10.1109/ASONAM.2012.164

[237] D. Nguyen, X. Jia, D. Sher, M.-H. Lin, Z. Iqbal, et al. Three-dimensional radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture. *Phys. Med. & Bio.*, 05 2018.

[238] C. Nowke, M. Schmidt, S. J. van Albada, J. M. Eppler, R. Bakker, et al. Visnest—interactive analysis of neural activity data. In *Bio. Data Vis.*, pp. 65–72, 2013. doi: 10.1109/BioVis.2013.6664348

[239] M. Nunes, B. Rowland, M. Schlachter, S. Ken, K. Matkovic, et al. An integrated visual analysis system for fusing mr spectroscopy and multi-modal radiology imaging. In *IEEE Vis. Ana. Sci. Tech. (VAST)*, pp. 53–62, 2014. doi: 10.1109/VAST.2014.7042481

[240] S. Oeltze, H. Doleisch, H. Hauser, P. Muigg, and B. Preim. Interactive visual analysis of perfusion data. *Trans. Vis. Comp. Graph.*, 13(6):1392–1399, 2007. doi: 10.1109/TVCG.2007.70569

[241] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2017.

[242] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 2018.

[243] M. Oppermann and T. Munzner. Data-first visualization design studies. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV)*, pp. 74–80, 2020. doi: 10.1109/BELIV51497.2020.00016

[244] E. Oral, R. Chawla, M. Wijkstra, N. Mahyar, and E. Dimara. From information to choice: A critical inquiry into visualization tools for decision making. *Trans. Vis. Comp. Graph*, 30(01):359–369, jan 2024. doi: 10.1109/TVCG.2023.3326593

[245] B. O'Sullivan, S. H. Huang, J. Su, A. S. Garden, E. M. Sturgis, K. Dahlstrom, N. Lee, N. Riaz, X. Pei, S. A. Koyfman, et al. Development and validation of a staging system for hpv-related oropharyngeal cancer by the international collaboration on oropharyngeal cancer network for staging (icon-s): a multicentre cohort study. *The Lancet Onco.*, 17(4):440–451, 2016. doi: 10.1016/S1470-2045(15)00560-4

[246] Y. Ouyang, Y. Wu, H. Wang, C. Zhang, F. Cheng, C. Jiang, L. Jin, Y. Cao, and Q. Li. Leveraging historical medical records as a proxy via multimodal modeling and visualization to enrich medical diagnostic learning. *Trans. Vis. Comp. Graph*, 30(01):1238–1248, jan 2024. doi: 10.1109/TVCG.2023.3326929

[247] A. Pandey, H. Shukla, G. S. Young, L. Qin, A. A. Zamani, L. Hsu, R. Huang, C. Dunne, and M. A. Borkin. Cerebrovis: Designing an abstract yet spatially contextualized cerebral artery network visualization. *IEEE Trans. Vis. Comp. Graph (TVCG)*, 26(1):938–948, 2020. doi: 10.1109/TVCG.2019.2934402

[248] D. Park et al. Supporting comment moderators in identifying high quality online news comments. *Conf. Hum. Fact. Comp. Sys.*, p. 1114–1125, 2016. doi: 10.1145/2858036.2858389

[249] D. Patel, L. P. Muren, A. Mehus, Y. Kvinnsland, D. M. Ulvang, and K. P. Villanger. A virtual reality solution for evaluation of radiotherapy plans. *Radiotherapy & Onco.*, pp. 218–221, 2 2007.

[250] G. Peake and J. Wang. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proc. 24rd ACM SIGKDD Int. Conf. Knowledge Disc. Data Mining*, 2018.

[251] C. Pepin-Neff and A. Cohen. President trump's transgender moral panic. In *The Trump Administration*, pp. 219–234. Routledge, 2022.

[252] E. G. M. Petrakis and C. Faloutsos. Similarity searching in medical image databases. *IEEE Trans. Knowl. Data Eng.*, pp. 435–447, 1997.

[253] S. Phung, A. Kumar, and J. Kim. A deep learning technique for imputing missing healthcare data. In *Conf. IEEE EMBC*, pp. 6513–6516, 2019. doi: 10.1109/EMBC.2019.8856760

[254] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *int. conf. on int. anal.*, vol. 5, pp. 2–4, 2005.

[255] P. Pirolli, S. K. Card, and M. M. Van Der Wege. Visual information foraging in a focus + context visualization. In *Proc. CHI*, CHI '01, 8 pages, p. 506–513. ACM, 2001. doi: 10.1145/365024.365337

[256] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. R. Schwalm, D. N. Huntzinger, and others. Visual reconciliation of alternative similarity spaces in climate modeling. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1923–1932, 2014.

[257] A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, pp. 111–163, 1995. doi: 10 .2307/271063

[258] T. S. Rai and A. P. Fiske. Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, 118(1):57, 2011.

[259] R. Raidou, O. Casares-Magaz, L. Muren, U. van der Heide, J. Rørvik, M. Breeuwer, and A. Vilanova. Visual analysis of tumor control models for prediction of radiotherapy response. *Comp. Graph. For.*, 35(3):231–240, 2016. doi: 10.1111/cgf.12899

[260] R. Raidou, U. van der Heide, C. Dinh, G. Ghobadi, J. Kallehauge, M. Breeuwer, and A. Vilanova. Visual analytics for the exploration of tumor tissue characterization. *Comp. Graph. For.*, 34(3):11–20, 2015. doi: 10. 1111/cgf.12613

[261] R. G. Raidou, M. Breeuwer, and A. Vilanova. Visual Analytics for Digital Radiotherapy: Towards a Comprehensible Pipeline. In S. Bruckner and T. Ropinski, eds., *EG 2017 - Dirk Bartz Prize*. The Eurographics Association, 2017. doi: 10.2312/egm.20171042

[262] R. G. Raidou, O. Casares-Magaz, A. Amirkhanov, V. Moiseenko, L. P. Muren, J. P. Einck, A. Vilanova, and M. E. Gröller. Bladder runner: Visual analytics for the exploration of rt-induced bladder toxicity in a cohort study. In *Comp. Graph. For.*, vol. 37-3, pp. 205–216. Wiley Online Library, 2018. doi: 10.1111/cgf.13413

[263] R. G. Raidou, K. Furmanova, N. Grossmann, O. Casares-Magaz, V. Moiseenko, J. P. Einck, M. Gröller, and L. P. Muren. Lessons learnt from developing visual analytics applications for adaptive prostate cancer radiotherapy. In *Gap bet. Vis. Res. Vis. Soft. (VisGap)*. The Eurographics Association, 2020.

[264] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Med.*, 2018.

[265] N. S. Rajliwall, R. Davey, and G. Chetty. Cardiovascular risk prediction based on xgboost. In *5th Asia-Pacific World Congress Comp. Sci. Eng. (APWC on CSE)*. IEEE, 2018.

[266] R. Rezapour, L. Dinh, and J. Diesner. Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics. In *Proc. 32nd ACM Confe.on Hypertext and Soc. Med.*, pp. 177–188, 2021. doi: 10.1145/3465336.3475112

[267] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[268] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Disc. Data Mining*, 2016.

[269] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[270] H. Ribičić, J. Waser, R. Fuchs, G. Blöschl, and E. Gröller. Visual analysis and steering of flooding simulations. *Trans. Vis. Comp. Graph.*, 19(6):1062–1075, 2013. doi: 10.1109/TVCG.2012.175

[271] G. Ristovski, T. Preusser, H. K. Hahn, and L. Linsen. Uncertainty in medical visualization: Towards a taxonomy. *Comp. & Graph.*, 39:60–73, 2014. doi: 10.1016/j.cag.2013.10.015

[272] P. Rodgers, L. Zhang, and A. Fish. General euler diagram generation. In *Diagrams*, pp. 13–27, 2008. doi: 10. 1007/978-3-540-87730-1_6

[273] A. Rojecki et al. The moral imperatives of self-quarantining. *American Political Science Association Annual Meeting*, Oct 2021.

[274] D. Rojo Garcia, N. N. Htun, and K. Verbert. GaCoVi: a Correlation Visualization to Support Interpretability-Aware Feature Selection. *Proc. EuroVis 2020 Short Papers*, 2020.

[275] D. I. Rosenthal, T. R. Mendoza, M. S. Chambers, J. A. Asper, I. Gning, M. S. Kies, R. S. Weber, J. S. Lewin, A. S. Garden, K. K. Ang, et al. Measuring head and neck cancer symptom burden: the development and validation of the md anderson symptom inventory, head and neck module. *Head & Neck*, 29(10):923–931, 2007. doi: 10.1002/hed.20602

[276] R. Rothstein. The making of ferguson. *Journal of Affordable Housing & Community Development Law*, 24(2):165–204, 2015.

[277] S. Ruiz-Correa, R. W. Sze, H. J. Lin, L. G. Shapiro, M. L. Speltz, and M. L. Cunningham. Classifying craniosynostosis deformations by skull shape imaging. In *IEEE Symp. Comp.-Based Med. Sys. (CBMS)*, pp. 335–340, 2005.

[278] I. Rössling, J. Dornheim, L. Dornheim, A. Boehm, and B. Preim. The Tumor Therapy Manager and its Clinical Impact. In K. Buehler and A. Vilanova, eds., *Eurographics 2011 - Dirk Bartz Prize*. The Eurographics Association, 2011. doi: 10.2312/EG2011/med/001-004

[279] I. Sanchez, T. Rocktaschel, S. Riedel, and S. Singh. Towards extracting faithful and descriptive representations of latent variable models. In *AAAI Spr. Symp Knowl. Repr. Reason. (KRR): Integ. Symb. Neur. Appro.*, 2015.

[280] D. J. Scala and K. M. Johnson. Political polarization along the rural-urban continuum? the geography of the presidential vote, 2000–2016. *The ANNALS of the American Academy of Political and Social Science*, 672(1):162–184, 2017. doi: 10.1177/0002716217712696

[281] C. Schein and K. Gray. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1):32–70, 2018.

[282] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans. Vis. Comp. Graph.*, 18(12):2431–2440, 2012. doi: 10.1109/TVCG.2012.213

[283] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[284] A. Sharafoddini, J. A. Dubin, and J. Lee. Patient similarity in prediction models based on health data: A scoping review. *JMIR Med. Inform.*, p. e7, 3 2017.

[285] T. Sheu, D. M. Vock, A. S. Mohamed, N. Gross, C. Mulcahy, M. Zafereo, G. B. Gunn, A. S. Garden, P. Sevak, J. Phan, et al. Conditional survival analysis of patients with locally advanced laryngeal cancer: construction of a dynamic risk model and clinical nomogram. *Sci. reports*, 7:43928, 2017.

[286] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. on Vis. Lang.*, pp. 336–343, 1996. doi: 10.1109/VL.1996.545307

[287] A. M. Smith, W. Xu, Y. Sun, J. R. Faeder, and G. E. Marai. Rulebender: integrated modeling, simulation and visualization for rule-based intracellular biochemistry. *BMC Bioinform.*, p. S3, 2012.

[288] S. Srabanti, M. Tran, V. Achim, D. Fuller, G. Canahuate, F. Miranda, and G. E. Marai. A tale of two centers: Visual exploration of health disparities in cancer care. In *Pac. Vis. Symp. (PacificVis)*, pp. 101–110. IEEE, 2022. doi: 10.1109/pacificvis53943.2022.00019

[289] C. D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2014.

[290] D. Streeb, Y. Metz, U. Schlegel, B. Schneider, M. El-Assady, H. Neth, M. Chen, and D. A. Keim. Task-based visual interactive modeling: Decision trees and rule-based classifiers. *Trans. Vis. Comp. Graph.*, 28(9):3307–3323, 2022. doi: 10.1109/TVCG.2020.3045560

[291] A. Suh, G. Appleby, E. W. Anderson, L. Finelli, R. Chang, and D. Cashman. Are metrics enough? guidelines for communicating and visualizing predictive models to subject matter experts. *Trans. Vis. Comp. Graph.*, pp. 1–16, 2023. doi: 10.1109/TVCG.2023.3259341

[292] M. Sultana, A. Haider, and M. S. Uddin. Analysis of data mining techniques for heart disease prediction. In *3rd Int. Conf. Elec. Eng. Info. Commun. Tech. (ICEEICT)*, 2016.

[293] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

[294] W. Swann. A survey of non-linear optimization techniques. *FEBS letters*, 2(S1):S39–S55, 1969.

[295] V. M. Systems. Velocity, 2018.

[296] A. K. Talukder, E. Selg, and R. E. Haas. Physicians' Brain Digital Twin: Holistic Clinical & Biomedical Knowledge Graphs for Patient Safety and Value-Based Care to Prevent the Post-pandemic Healthcare Ecosystem Crisis. In *Knowledge Graphs and Semantic Web*, pp. 32–46. Springer, Cham, Switzerland, November 2022. doi: 10.1007/978-3-031-21422-6_3

[297] E. Tardini, X. Zhang, G. Canahuate, A. Wentzel, A. S. Mohamed, L. Van Dijk, C. D. Fuller, and G. E. Marai. Optimal treatment selection in sequential systemic and locoregional therapy of oropharyngeal squamous carcinomas: Deep q-learning with a patient-physician digital twin dyad. *Journal of medical Internet research*, 24(4):e29455, 2022.

[298] M. ten Caat, N. M. Maurits, and J. B. Roerdink. Functional unit maps for data-driven visualization of high-density eeg coherence. In *EuroVis*, pp. 259–266, 2007.

[299] M. Ten Caat, N. M. Maurits, and J. B. Roerdink. Data-driven visualization and group analysis of multichannel eeg coherence with functional units. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 756–771, 2008.

[300] C.-C. Teng, L. G. Shapiro, I. Kalet, C. Rutter, and R. Nurani. Head and neck cancer patient similarity based on anatomical structural geometry. In *IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, pp. 1140–1143, 2007.

[301] X. Teng, Y. Ahn, and Y. Lin. Vispur: Visual aids for identifying and interpreting spurious associations in data-driven decisions. *Trans. Vis. Comp. Graph.*, 30(01):219–229, jan 2024. doi: 10.1109/TVCG.2023.3326587

[302] S. T. Teoh and K.-L. Ma. Paintingclass: interactive construction, visualization and exploration of decision trees. In *Int. Conf. on Knowl. disc. and data min.*, pp. 667–672, 2003. doi: 10.1145/956750.956837

[303] S. T. Teoh and K.-L. Ma. Starclass: Interactive visual classification using star coordinates. In *Int. Conf. on Data Min.*, pp. 178–185. SIAM, 2003. doi: 10.1137/1.9781611972733.16

[304] The New York Times and Dynata. A detailed map of who is waring masks in the u.s. `https://github.com/nytimes/covid-19-data/blob/master/mask-use`, 2020.

[305] T. M. Therneau. Extending the cox model. In *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, pp. 51–84. Springer, 1997.

[306] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Net. and Learn. Sys.*, 32(11):4793–4813, 2021. doi: 10.1109/TNNLS.2020.3027314

[307] Y. Toda, F. Okura, et al. How convolutional neural networks diagnose plant disease. *Plant Phenomics*, 2019.

[308] J. Tosado, L. Zdilar, H. Elhalawani, B. Elgohari, D. M. Vock, G. E. Marai, C. Fuller, A. S. Mohamed, and G. Canahuate. Clustering of largely right-censored oropharyngeal head and neck cancer patients for discriminative groupings to improve outcome prediction. *Sci. reports*, 10(1):1–14, 2020.

[309] B. Trachtenberg. The 2015 university of missouri protests and their lessons for higher education policy and administration. *Ky. LJ*, 107:61, 2018.

[310] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. Ten Haken, and I. El Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med. Phys.*, 2017.

[311] E. R. Tufte. *Envisioning Information*. Envisioning Information. Graphics Press, 1990.

[312] United States Census Bureau. `https://www.census.gov/data/tables/time-series/demo/voting-and-registration/congressional-voting-tables.html`, 2020.

[313] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learn.*, 2016.

[314] S. Valenzuela, M. Piña, and J. Ramírez. Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of communication*, 67(5):803–826, 2017. doi: 10.1111/jcom.12325

[315] S. van den Elzen and J. J. van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *Conf. on Vis. Anal. Sci. and Tech. (VAST)*, pp. 151–160, 2011. doi: 10.1109/VAST.2011.6102453

[316] U. A. Van der Heide, A. C. Houweling, G. Groenendaal, R. G. Beets-Tan, and P. Lambin. Functional MRI for radiotherapy dose painting. *Magnetic Resonance Imag.*, pp. 1216–1223, 2012.

[317] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. doi: 10.1016/j.media.2022.102470

[318] S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Comp. Sci. & Engi.*, p. 22, 2011.

[319] L. V. van Dijk, A. S. Mohamed, S. Ahmed, N. Nipu, G. E. Marai, K. Wahid, N. M. Sijtsema, B. Gunn, A. S. Garden, A. Moreno, et al. Head and neck cancer predictive risk estimator to determine control and therapeutic outcomes of radiotherapy (hnc-predictor): development, international multi-institutional validation, and web implementation of clinic-ready model-based risk stratification for head and neck cancer. *European J. of Cancer*, 178:150–161, 2023. doi: 10.1016/j.ejca.2022.10.011

[320] H. van Hasselt, A. Guez, and D. Silver. Deep Reinforcement Learning with Double Q-Learning. *AAAI*, 30(1), March 2016. doi: 10.1609/aaai.v30i1.10295

[321] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[322] T. von Landesberger et al. Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *Trans. Vis. Comp. Graph.*, 22(1):11–20, 2016. doi: 10.1109/TVCG.2015.2468111

[323] R. D. Vromans, S. Hommes, F. J. Clouth, D. N. Lo-Fo-Wong, X. A. Verbeek, L. van de Poll-Franse, S. Pauws, and E. Krahmer. Need for numbers: assessing cancer survivors' needs for personalized and generic statistical information. *BMC Medical Informatics and Decision Making*, 22(1):1–14, 2022.

[324] F. Wang, J. Sun, and S. Ebadollahi. Integrating distance metrics learned from multiple experts and its application in patient similarity assessment. In *Proc. SIAM Int. Conf. Data Mining*, 4 2011.

[325] J. Wang, L. Gou, H.-W. Shen, and H. Yang. Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):288–298, 2019. doi: 10.1109/TVCG.2018.2864504

[326] J. Wang, L. Gou, H. Yang, and H. Shen. Ganviz: A visual analytics approach to understand the adversarial game. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2018. doi: 10.1109/TVCG.2018.2816223

[327] J. Wang, W. Zhang, H. Yang, C.-C. M. Yeh, and L. Wang. Visual analytics for rnn-based deep reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4141–4155, 2022. doi: 10.1109/TVCG.2021.3076749

[328] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6), article no. 110, 36 pages, feb 2019. doi: 10.1145/3214306

[329] Q. Wang, K. Huang, P. Chandak, M. Zitnik, and N. Gehlenborg. Extending the nested model for user-centric xai: A design study on gnn-based drug repurposing. *Trans. Vis. Comp. Graph.*, 29(1):1266–1276, 2022. doi: 10.1109/TVCG.2022.3209435

[330] Q. Wang, T. Mazor, T. A. Harbig, E. Cerami, and N. Gehlenborg. Threadstates: State-based visual analysis of disease progression. *Trans. Vis. Comp. Graph.*, 28(1):238–247, 2021. doi: 10.1109/TVCG.2021.3114840

[331] X. Wang. The role of the ingroup moral foundation on message responses: Two experiments on race and nationality. *Howard Journal of Communications*, pp. 1–17, 2021.

[332] Y. Wang, L. V. Dijk, A. S. R. Mohamed, M. Naser, et al. Improving prediction of late symptoms using lstm and patient-reported outcomes for head and neck cancer patients. In *Proc. ICHI*, pp. 292–300, 2023. doi: 10.1109/ICHI57859.2023.00047

[333] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Proc.*, pp. 600–612, 4 2004.

[334] S. Warchol, R. Krueger, A. J. Nirmal, G. Gaglia, J. Jessup, C. C. Ritch, J. Hoffer, J. Muhlich, M. L. Burger, T. Jacks, S. Santagata, P. K. Sorger, and H. Pfister. Visinity: Visual spatial neighborhood analysis for multiplexed tissue imaging data. *Trans. Vis. Comp. Graph.*, pp. 1–11, 1912. doi: 10.1109/TVCG.2022.3209378

[335] J. W. Ward, R. Phillips, T. Williams, C. Shang, L. Page, et al. Immersive visualization with automated collision detection for radiotherapy treatment planning. In *MMVR*, pp. 491–496, 2007.

[336] C. Ware et al. Designing pairs of colormaps for visualizing bivariate scalar fields. In *Com. Graph. For.*, pp. 49–53, 2020. doi: 10.2312/evs.20201047

[337] J. Waser, R. Fuchs, H. Ribičič, B. Schindler, G. Blöschl, and E. Gröller. World lines. *Trans. Vis. Comp. Graph.*, 16(6):1458–1467, 2010. doi: 10.1109/TVCG.2010.223

[338] S. Webb. Optimizing the planning of intensity-modulated radiotherapy. *Phys. Med. & Bio.*, pp. 2229–2246, 12 1994.

[339] J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, and C. North. Towards a systematic combination of dimension reduction and clustering in visual analytics. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2017.

[340] J. Wenskovitch, L. Harris, J. Tapia, J. Faeder, and G. Marai. Mosbie: A tool for comparison and analysis of rule-based biochemical models. *BMC Bioinform.*, pp. 1–22, 2014.

[341] J. E. Wenskovitch, L. A. Harris, J.-J. Tapia, J. R. Faeder, and G. E. Marai. MOSBIE: a tool for comparison and analysis of rule-based biochemical models. *BMC Bioinfo.*, 15(1):1–16, 2014.

[342] A. Wentzel, S. Attia, X. Zhang, G. Canahuate, C. Fuller, and G. Marai. Ditto: A visual digital twin for interventions and temporal treatment outcomes in head and neck cancer". *Trans. Vis. Comp. Graphics*, pp. 1–11, January 2025. doi: 10.1109/TVCG.2024.3456160

[343] A. Wentzel, G. Canahuate, L. V. Van Dijk, A. S. Mohamed, C. D. Fuller, and G. E. Marai. Explainable spatial clustering: Leveraging spatial data in radiation oncology. In *Vis. Conf. (short paper)*, pp. 281–285. IEEE, 2020. doi: 10.1109/VIS47514.2020.00063

[344] A. Wentzel, C. Floricel, G. Canahuate, M. Naser, A. S. Mohamed, C. D. Fuller, L. van Dijk, and G. E. Marai. DASS Good: Explainable Data Mining of Spatial Cohort Data. *Comp. Graph. For.*, 24(3), Jun 2023. doi: 10.1111/cgf.14830

[345] A. Wentzel, P. Hanula, T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D. Vock, C. D. Fuller, and G. E. Marai. Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration. *Trans. Vis. Comp. Graph.*, 26(1):949–959, 2019. doi: 10.1109/TVCG.2019.2934546

[346] A. Wentzel, P. Hanula, L. V. van Dijk, B. Elgohari, A. S. Mohamed, C. E. Cardenas, C. D. Fuller, D. M. Vock, G. Canahuate, and G. E. Marai. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. *Radiotherapy and Onc.*, 148:245–251, 2020. doi: 10.1016/j.radonc.2020.05.023

[347] A. Wentzel, L. Levine, V. Dhariwal, Z. Fatemi, B. Di Eugenio, A. Rojecki, E. Zheleva, and G. E. Marai. A Lens to Pandemic Stay at Home Attitudes. *IEEE Vis4PanDemRes Workshop*, August 2023. doi: 10.48550/arXiv.2308.13552

[348] A. Wentzel, L. Levine, V. Dhariwal, Z. Fatemi, B. D. Eugenio, A. Rojecki, E. Zheleva, and G. E. Marai. Motiv: Visual exploration of moral framing in social media. *Comp. Graph. For.*, 2024. doi: 10.1111/cgf.15072

[349] A. Wentzel, T. Luciani, L. V. van Dijk, N. Taku, B. Elgohari, A. S. Mohamed, G. Canahuate, C. D. Fuller, D. M. Vock, and G. Elisabeta Marai. Precision association of lymphatic disease spread with radiation-associated toxicity in oropharyngeal squamous carcinomas. *Radiotherapy & Onco.*, 161:152–158, 2021. doi: 10.1016/j.radonc.2021.06.016

[350] A. Wentzel, A. S. R. Mohamed, M. A. Naser, L. V. van Dijk, K. Hutcheson, A. M. Moreno, C. D. Fuller, G. Canahuate, and G. E. Marai. Multi-organ spatial stratification of 3-d dose distributions improves risk prediction of long-term self-reported severe symptoms in oropharyngeal cancer patients receiving radiotherapy: development of a pre-treatment decision support tool. *Frontiers in Oncology*, 13, 2023. doi: 10.3389/fonc.2023.1210087

[351] M. Werner-Wasik, E. Yorke, J. Deasy, J. Nam, and L. B. Marks. Radiation dose-volume effects in the esophagus. *Int. J. Rad. Onco., Bio., Phys.*, pp. S86–S93, 3 2010.

[352] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Trans. Vis. Comp. Graph.*, 2019.

[353] C. Wilhelm and S. Joeckel. Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles*, 80(7):381–392, 2019. doi: 10.1007/s11199-018-0941-5

[354] S. Wojcik and A. Hughes. Sizing up twitter users. Pew Research Center, April 2019.

[355] I. Wolf, M. Vetter, I. Wegner, M. Nolden, T. Bottger, et al. The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK. In *Med. Imag.: Vis., Image-Guided Proc., and Disp.*, pp. 16–28, 2004. doi: 10.1117/12.535112

[356] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mane, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2017.

[357] J. Wu, M. F. Gensheimer, N. Zhang, F. Han, R. Liang, Y. Qian, C. Zhang, N. Fischbein, E. L. Pollom, B. Beadle, et al. Integrating tumor and nodal imaging characteristics at baseline and mid-treatment computed tomography scans to predict distant metastasis in oropharyngeal cancer treated with concurrent chemoradiotherapy. *Int. J. Rad. Onco., Bio., Phys.*, 2019.

[358] Y. Wu et al. Opinionflow: Visual analysis of opinion diffusion on social media. *Trans. Vis. Comp. Graph.*, 20(12):1763–1772, 2014. doi: 10.1109/TVCG.2014.2346920

[359] C. Xiong, E. Lee-Robbins, I. Zhang, A. Gaba, and S. Franconeri. Reasoning affordances with tables and bar charts. *Trans. Vis. Comp. Graph*, pp. 1–13, 2022. doi: 10.1109/TVCG.2022.3232959

[360] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of causality in visualized data. *Trans. Vis. Comp. Graph.*, 26(1):853–862, 2019. doi: 10.1109/TVCG.2019.2934399

[361] L. Xu and C. Guo. Coxnam: An interpretable deep survival analysis model. *Expert Systems with Applications*, 227:120218, 2023. doi: 10.1016/j.eswa.2023.120218

[362] W. Xu, A. M. Smith, J. R. Faeder, and G. E. Marai. Rulebender: a visual interface for rule-based modeling. *Bioinformatics*, 27(12):1721–1722, 2011. doi: 10.2312/evs.20201067

[363] Q. Yang, A. Steinfeld, and J. Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 11 pages, p. 1–11. Association for Computing Machinery, 2019. doi: 10.1145/3290605.3300468

[364] B. Yener, C. Gunduz, and S. H. Gultekin. The cell graphs of cancer. *Bioinform.*, pp. i145–i151, 08 2004.

[365] S. Yousefi, F. Amrollahi, M. Amgad, C. Dong, J. E. Lewis, C. Song, D. A. Gutman, S. H. Halani, J. E. Velazquez Vega, D. J. Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):11707, 2017. doi: 10.1038/s41598-017-11817-6

[366] J. Yuan, O. Nov, and E. Bertini. An exploration and validation of visual factors in understanding classification rule sets. In *Vis. COnf. (short paper)*, pp. 6–10. IEEE, 2021. doi: 10.1109/VIS49827.2021.9623303

[367] J.-D. Zapata-Rivera, E. Neufeld, and J. Greer. Visualization of bayesian belief networks. In *IEEE Visualization Late Breaking Hot Topics Proceedings*, 1999.

[368] L. Zdilar, D. M. Vock, G. E. Marai, C. D. Fuller, A. S. Mohamed, H. Elhalawani, B. A. Elgohari, C. Tiras, A. Miller, and G. Canahuate. Evaluating the effect of right-censored end point transformation for radiomic feature selection of data from patients with oropharyngeal cancer. *JCO clinical cancer informatics*, 2:1–19, 2018.

[369] V. Zebralla, J. Müller, T. Wald, A. Boehm, G. Wichmann, T. Berger, K. Birnbaum, K. Heuermann, S. Oeltze-Jafra, T. Neumuth, et al. Obtaining patient-reported outcomes electronically with "oncofunction" in head and neck cancer patients during aftercare. *Frontiers in Oncology*, 10:549915, 2020.

[370] A. X. Zhang, M. Muller, and D. Wang. How do data science workers collaborate? roles, workflows, and tools. *Proc. ACM on Human-Comp. Int.*, 4:1–23, 2020. doi: 10.1145/3392826

[371] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE Trans. Vis. Comp. Graph. (TCVG)*, 2018.

[372] L. Zhang, M. Hub, C. Thieke, R. O. Floca, and C. P. Karger. A method to visualize the uncertainty of the prediction of radiobiological models. *Phys. Med.*, 29(5):556–561, 2013. doi: 10.1016/j.ejmp.2012.11.004

[373] M. Zhang, D. Ehrmann, M. Mazwi, D. Eytan, M. Ghassemi, and F. Chevalier. Get to the point! problem-based curated data views to augment care for critically ill patients. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, article no. 278, 13 pages. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3501887

[374] P. Zhang, F. Wang, J. Hu, and R. Sorrentino. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits Transl. Sci. Proc.*, p. 132, 2014.

[375] T. Zhang, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and E. Glassman. Interactive cohort analysis and hypothesis discovery by exploring temporal patterns in population-level health records. In *IEEE Vis. Ana. in Heal. (VAHC)*, pp. 14–18, 2021. doi: 10.1109/VAHC53616.2021.00007

[376] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Info. Vis.*, 14(4):289–307, 2015. doi: 10.1177/1473871614526077

[377] J. Zhao et al. Fluxflow: Visual analysis of anomalous information spreading on social media. *Trans. Vis. Comp. Graph.*, 20(12):1773–1782, 2014. doi: 10.1109/TVCG.2014.2346922

[378] B. Zheng, S. Verma, J. Zhou, I. W. Tsang, and F. Chen. Imitation Learning: Progress, Taxonomies and Challenges. *Trans. Neu. Net. Learn. Sys.*, pp. 1–16, October 2022. doi: 10.1109/TNNLS.2022.3213246

[379] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conf. Comp. Intell. Games (CIG)*, 2018.

[380] Z. Zhu, C. Liu, and X. Xu. Visualisation of the Digital Twin data in manufacturing by using Augmented Reality. *Proc. CIRP*, 81:898–903, January 2019. doi: 10.1016/j.procir.2019.03.223

[381] A. Zytek, D. Liu, R. Vaithianathan, and K. Veeramachaneni. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *Trans Vis. Comp. Graph*, 28(1):1161–1171, 11 pages, jan 2022. doi: 10.1109/TVCG.2021.3114864

Andrew Wentzel Vita

Education

University of Illinois, Chicago *August 2018-December 2024*
- MS Degree, *Computer Science (*earned 12/2019) 4.0 GPA

The Cooper Union *May 2016*
- Bachelor of Engineering*, Mechanical Engineering* 3.74 GPA

Experience

University of Illinois, Chicago
- Research Assistant*; Electronic Visualization Lab* *January 2019 - December 2024*
- Teaching Assistant - Visual Data Science (CS 526) *Fall 2024*

Epsilon
- PhD Intern, Decision Science and Visual Analytics *May 2024 - August 2024*

Northwestern University
- Architectural Drafter *Sept. 2016 - August 2018*

The Cooper Union
- Computer Center Supervisor *May 2013- May 2016*
- Research Assistant, Advanced Computational Fluid Dynamics Lab *June 2015 - June 2016*

First Author Papers

Andrew Wentzel, Serageldin Attia et al. *DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer,* IEEE Trans. Vis. Comp. Graphics, Jan 2025.

Andrew Wentzel, Lauren Levine, Vipu Dhariwal, Zarah Fatemi, Abara Bhattacharya, Barbara Di Eugenio, Andrew Rojecki, Elena Zheleva, G.Elisabeta Marai, *MOTIV: Visual Exploration of Moral Framing in Social Media.* Computer Graphics Forum, 2024.

Andrew Wentzel, Lauren Levine, et. al, *A Lens to Stay at Home Pandemic Attitudes.* pp. 1-5. IEEE Workshop on Visualization for Pandemic and Emergence Response, 2023.

Andrew Wentzel, Abdallah SR Mohamed, et. al, *Multi-organ spatial stratification of 3-D dose distributions improves risk prediction of long-term self-reported severe symptoms in oropharyngeal cancer patients receiving radiotherapy: development of a pre-treatment decision support tool.* Frontiers in Oncology, 2023.

Andrew Wentzel, Carla Floricel, et. al, *DASS Good: Explainable Data Mining of Spatial Cohort Data.* pp. 1-13. Computer Graphics Forum, 2023.

Andrew Wentzel, Timothy Luciani, et. al, *Precision association of lymphatic disease spread with radiation-associated toxicity in oropharyngeal squamous carcinomas.* pp. 1-7. Radiotherapy and Oncology, 2021.

Andrew Wentzel, Guadalupe Canahuate, et. al, *Explainable Spatial Clustering: Leveraging Spatial Data in Radiation Oncology.* pp. 1-5. IEEE Visualization Conference (VIS), 2020.

Andrew Wentzel, Peter Hanula, et. al, *Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy.* pp. 1-7. Radiotherapy and Oncology, 2020.

Andrew Wentzel, Peter Hanula, et. al, *Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration.* pp. 1-11. IEEE Trans. Vis. Comp. Graph, 2019.**(1)

(1) cited in: B. Preim, R. Raidou, N. Smit, K. Lawonn, "Visualization, Visual Analytics and Virtual Reality in Medicine" 1st Edition

Coauthor Papers

Juan Trelles, **Andrew Wentzel**, William Berrios, Hagit Shatkay, and G.E. Marai., *BI-LAVA: Biocuration With Hierarchical Image Labeling Through Active Learning and Visual Analytics.* Computer Graphics Forum e15261, 2024.

Mirko Mantovani, **Andrew Wentzel**, et al., *Kiviat Defense: An Empirical Evaluation of Visual Encoding Effectiveness in Multivariate Data Similarity Detection.* pp. 1-13. Journal of Imaging Science and Technology, 2023.

Guadalupe Canahuate, **Andrew Wentzel,** Abdallah SR Mohamed, Lisanne V. van Dikj, David M Vock, Baher Elgorhari, Hesham Elhalawani, Clifton D Fuller, G.E Marai, *Spatially-aware clustering improves AJCC-8 risk stratification performance in oropharyngeal carcinomas.* Oral Oncology, 2023.

Carla Floricel, **Andrew Wentzel,** et. al, *Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining.* pp. 1-11. Transactions on Visualization and Computer Graphics, 2023.

Zahra Fatemi, Abari Bhattacharya, **Andrew Wentzel,** et. al, *Understanding Stay-at-home Attitudes through Framing Analysis of Tweets.* IEEE 9th Int. Conf. on data Science and Advanced Analytics, 2022.

Elisa Tardini, Xinhua Zhang, Guadalupe Canahuate, **Andrew Wentzel,** et. al, *Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad.* pp. 1-21, J Med Internet Res, 2022.

Timothy Luciani, **Andrew Wentzel,** et. al, *A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine.* Journal of Biomedical Informatic, 2020.

Conference Presentations

DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer. IEEE Visualization Conference (VIS), 2024.

Kiviat Defense: An Empirical Evaluation of Visual Encoding Effectiveness in Multivariate Data Similarity Detection. Electronic Imaging - Visualization and Data Analysis Conference, 2024.

DASS Good: Explainable Data Mining of Spatial Cohort Data. EuroVis Conference, 2023.

Explainable Spatial Clustering: Leveraging Spatial Data in Radiation Oncology. IEEE Visualization Conference (VIS), 2020.

Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration. IEEE Visualization Conference (VIS), 2019.

Awards/Scholarships

UIC College of Engineering Exceptional Research Promise Award	*April 2024*
Northwestern Walter P. Murphy Fellowship	*Sep. 2017 - Mar. 2018*
Cooper Union, Full Tuition Scholarship	*2012-2016*
Cooper Union, Society of Military Engineering (SAME) Merit Scholarship	*2015*

Journals I Have Reviewed For

IEEE VIS Conference/Transactions on Visualization and Computer Graphics: 2022-1056, 2022-1400, 2022-1212, 2023-1109, 2023-1279, 2024-1707, 2024-1755
EuroVis Conference/Computer Graphics Forum: CGF-22-ORA-052, EuroVis-22-1273
Computers and Graphics: CAG-D-22-00541R2
Nature Scientific Reports: c3e51fe9-828b-4245-8f63-4bdf02-a8c383
Frontiers in Oncology: 1017033