

Problem-Driven Design Strategies for Scientific Visualization

by

Timothy Luciani

B.S. in Computer Science, University of Pittsburgh, PA. 2011

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:

Dr. G. Elisabeta Marai, *Chair and Advisor*

Dr. Andrew Johnson

Dr. Robert Kenyon

Dr. Angus Forbes, University of California, Santa Cruz

Dr. Klaus Mueller, Stony Brook University

Copyright by
Timothy Luciani
2019

In loving memory of
Basil "PeeWee" Luciani

ACKNOWLEDGMENTS

I thank my advisor, Dr. G. Elisabeta Marai, for her guidance and support throughout my dissertation. It's been a long journey. I would also like to thank my committee members Dr. Andrew Johnson, Dr. Robert Kenyon, Dr. Angus Forbes, and Dr. Klaus Mueller for their invaluable feedback on my work.

I gratefully acknowledge my collaborators, without whom none of my accomplishments would have been possible: Dr. Brian Cherinka, Dr. Alexandros Labrinidis, Dr. William Ray, Dr. Baher Elgohari, Dr. Hesham Elhalawani, Dr. Guadalupe Canahuate, Dr. David Vock, Dr. C. David Fuller, John Wenskovitch, and Jon Komperda. I also gratefully thank the generous support from the NSF GRFP and NIH.

I humbly thank my labmates at the Electronic Visualization Laboratory at UIC and elsewhere for their support. In particular, I thank Maxine Brown for helping to strengthen my focus towards my goals and for saving my sanity more than once. And to my friends: Alexis Miranda, Matt Philips, Sarah White, John Wenskovitch, Juan Trelles, Brian Dicks, Krishna Bharadwaj, Andrew Burks, Chihua Ma, Ganesh Jagadeesan, and Tim Kubal – I could not have dreamed for a more supportive group of companions.

I thank the loving support of my family: my mother, Cynthia Bailey, for her unconditional support and love, without which I would have surely failed; my father, Basil Luciani, who inspired my love of technology as well as, unknowingly, my desire to “poke it and see what happens”; and my brother, Brendon Luciani, for proving that any challenge can be overcome. Finally, I thank my grandparents, Audrey “WaWa” and Basil “Pappy” Luciani, for being my best escape even in my worst of times.

TBL

CONTRIBUTION OF AUTHORS

This dissertation is comprised of my previously published works, in collaboration with other authors. My direct contributions for each of these works are outlined at the beginning of each chapter. These contributions include:

Chapter 2 was originally published in the IEEE Transactions on Visualization and Computer Graphics (TVCG) Journal in 2014 as *Large-Scale Overlays and Trends: Visually Mining, Panning and Zooming the Observable Universe*. [1]. This version has been edited to be consistent with the rest of the dissertation. Coauthors on the original work include Brian Cherinka (BC), Daniel Oliphant (DO), Sean Myers (SM), W. Michael Wood-Vasey (MWV), Alexandros Labrinidis (AL), and G. Elisabeta Marai (GEM). The contributions from each author included: BC and MWV served as our observational astronomy domain experts, providing the theoretical underpinnings for the trend image design and provided support with the software testing and the design of the case studies; DO designed the first versions of the client-server architecture, the Data-Driven Spots, and the original image stitching algorithm behind Astroshelf; SM contributed to the implementation and refactoring of the front-end client and SkyView; AL served as a co-principal investigator on the project for the data management of the project and recommended the use of the S^2 -tree for indexing the images; GEM served a co-principal investigator on the project for the visualization aspects of the project and directed the top-level design, implementation and testing of the tool. My (TL) contributions to this work included the design and implementation of the pixel-based trend images, postage-stamp images, and interactive SkyView. Additionally, I worked with BC to project the various survey images (SDSS, FIRST, and LSST) for their stitching and cross-registration in

CONTRIBUTION OF AUTHORS (Continued)

the SkyView and helped design and develop the current version of the client-server architecture and AL and his data management team. I am the first author on this work.

Chapter 3 was originally published in the BMC Proceedings in 2014 as *FixingTIM: Interactive Exploration of Sequence and Structural Data to Identify Functional Mutations in Protein Families* [2]. This version has been edited to be consistent with the rest of the dissertation. Coauthors on this work include John Wenskovitch (JW), Koonwah Chen (KC), David Koes (DK), Timothy Travers (TT), and G. Elisabeta Marai (GEM). The contributions from each author included: JW implemented the sorting algorithms for the trend image and several of the data parsers; KC contributed the design and implementation of the sequence and residue views; DK and TT served as our structural biology domain experts and provided support with the software testing and the design of the case studies; GEM conceived this project, and directed the top-level design, implementation and testing of the tool. My (TL) contributions to this work included the design and implementation of the client-server architecture, the database back-end, as well as the 3D view and trend image of the visual interface. I am the first author on this work.

Chapter 4 is pending review in the Journal of Biomedical Informatics as *A Spatial Neighborhood Method for Computing Lymph Node Carcinoma Similarity in Precision Medicine*. Coauthors on this work include Baher Elgohari (BH), Hesham Elhalawani (HE), Guadalupe Canahuat (GC), David M. Vock (DV), C. David Fuller (CDF), and G. Elisabeta Marai (GEM). The contributions from each author included: BH, HE and CDF served as the radiation oncology domain experts for this work, providing feedback on the relevance of our spatial neighborhood approach; BH and HE were responsible for the data curation, and cleansing of the patient cohort; GC provided the data mining expertise on the project

CONTRIBUTION OF AUTHORS (Continued)

and contributed the hierarchical clustering implementation and Chi-Squared analysis; DV provided the statistical analysis expertise and suggested the use of the Fisher’s Exact Test of the Chi-Squared Test; GEM provided the visualization expertise and directed the top-level design, implementation, and testing of the approach. My (TL) contributions to this work included the design and implementation of the similarity measure, the compact graph visual representations, and the spatial-measure dendrogram. I am the first author on this work.

Chapter 5 was originally published in the IEEE Transactions on Visualization and Computer Graphics (TVCG) Journal in 2018 as *Details-first, Show Context, Overview last: Supporting Exploration of Viscous Fingers in Large-Scale Ensemble Simulations*. [3]. This version has been edited to be consistent with the rest of the dissertation. In particular, relevant sections that appeared in the original manuscript were relocated to the Introduction and Discussion chapters of this dissertation to help motivate this work and explain its contributions. Coauthors on this work include Andrew Burks (AB), Cassiano Sugiyama (CS), Jonathan Komperda (JK), and G. Elisabeta Marai (GEM). The contributions from each author included: AB and CS co-designed and developed the FingerFinder web-application, including the underlying visual encodings and algorithms, as National Science Foundation REUs under the direction of GEM and me; AB also designed the layout of the merge tree to reduce cluttering between finger tracks; JC served as the mechanical engineering expert for this project, provided theoretical and qualitative feedback on the Details-first model, and provided support with the software testing; GEM conceived this project and its theoretical framework, and directed the design, implementation and testing of the FingerFinder tool. My (TL) contributions to this work included the conception of the merge tree for the tracking of the features over time as well as contributions to the design and implementation of the Fin-

CONTRIBUTION OF AUTHORS (Continued)

gerFinder tool. Additionally, I worked with GEM and JC on developing the theoretical underpinnings of the Details-first approach. I am the first author on this work.

Permission to use these works appears in Appendix B.

TABLE OF CONTENTS

<u>CHAPTER</u>		<u>PAGE</u>
1	INTRODUCTION	1
1.1	Motivation	1
1.2	Relevant Visualization Paradigms	4
1.2.1	Overview-first paradigm	4
1.2.2	Search-first paradigm	5
1.2.3	Details-first paradigm	5
1.3	Key Terminology	6
1.3.1	Overview	7
1.3.2	Context	8
1.3.3	Detail	8
1.4	Scientific Workflow Theory	8
1.5	Challenges	9
1.6	Contributions Overview	12
2	LARGE-SCALE OVERLAYS AND TRENDS: VISUALLY MINING, PANNING AND ZOOMING THE OBSERVABLE UNIVERSE	15
2.1	Introduction	16
2.2	Domain Analysis	19
2.2.1	Data Analysis	21
2.2.2	Task Analysis	23
2.3	Related Work	24
2.4	Design and Implementation	27
2.4.1	Data Retrieval and Preprocessing	29
2.4.1.1	Catalog Data	31
2.4.2	Cross-Registration and Online Overlays	31
2.4.2.1	Sky Panoramas	31
2.4.2.2	DDS Overlays	33
2.4.2.3	Online Compositing	34
2.4.3	Interactive Trend Images	35
2.4.4	Rendering and Interaction	39
2.4.4.1	Panning and Zooming Large-Scale Overlays	39
2.4.4.2	Trend Image Interaction	40
2.4.4.3	Small Multiples	41
2.4.4.4	Linked Views	41
2.5	Results	42
2.5.1	Preprocessing	42

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
	2.5.2 Performance	43
	2.5.3 Case Study: UGC 08782 - A Dusty Elliptical	43
	2.5.4 Case Study: Trends in Type Ia Supernovae	45
	2.5.5 Case Study: Spectroscopic Analysis of Galaxies	48
	2.5.6 Domain-Expert Feedback	51
	2.6 Discussion and Conclusion	53
3	INTERACTIVE EXPLORATION OF SEQUENCE AND STRUCTURAL DATA TO IDENTIFY FUNCTIONAL MUTATIONS IN PROTEIN FAMILIES	56
	3.1 Background	57
	3.2 Methods	60
	3.2.1 Data and Task Analysis	60
	3.2.2 Client-Server Framework	61
	3.2.3 Visual Design	62
	3.2.3.1 Trend Image Panel	63
	3.2.3.2 Residue Viewer	66
	3.2.3.3 3D Viewer	66
	3.2.3.4 Protein Sequence Viewer	66
	3.3 Results and Discussion	66
	3.3.1 TIM protein-family exploration	67
	3.3.2 Structural Biologist Feedback	69
	3.3.3 BioVis Contest Organizer Feedback	71
	3.4 Conclusions	73
4	A SPATIAL NEIGHBORHOOD METHOD FOR COMPUTING LYMPH NODE CARCINOMA SIMILARITY IN PRECISION MEDICINE	74
	4.1 Introduction	75
	4.2 Methods	78
	4.2.1 Overview	78
	4.2.2 Patient Cohort	79
	4.2.3 Topological Map	80
	4.2.4 Similarity Computation	83
	4.2.4.1 Spatial and Non-Spatial Similarity Measures	84
	4.2.4.2 Hierarchical Clustering	87
	4.2.4.3 Statistical Analysis	89
	4.2.4.4 Visual Analysis	89
	4.3 Results	91
	4.3.1 Spatial vs Categorical Node Patient Categorization	91
	4.3.2 Domain Expert Feedback	93
	4.3.3 Hierarchical Clustering Analysis	93
	4.3.3.1 Measure Agreement	95

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
	4.3.4 Statistical Analysis Results	96
	4.3.5 Performance	96
	4.4 Discussion	97
	4.5 Conclusion	98
5	“DETAILS-FIRST, SHOW CONTEXT, OVERVIEW LAST”: SUPPORTING EXPLORATION OF VISCOUS FINGERS IN LARGE-SCALE ENSEMBLE SIMULATIONS	100
	5.1 Introduction	102
	5.2 Background and Related Work	104
	5.2.1 Spatial Features as Details	105
	5.2.2 Features and Soft-knowledge	105
	5.2.3 CFD Visualization	106
	5.2.3.1 Feature Extraction	106
	5.2.3.2 Feature Tracking	107
	5.2.3.3 Feature-based Filtering	107
	5.2.3.4 Ensemble Visualization	108
	5.3 Model Instantiation	109
	5.3.1 Constructive Example and Workflow Analysis	109
	5.3.1.1 Data and Tasks	110
	5.3.1.2 Model Perspective	110
	5.3.1.3 Scientific Workflow Analysis	111
	5.4 Finger Segmentation and Spatial Context Calculation	112
	5.4.1 Finger Visualization	113
	5.4.1.1 3D View and Context	113
	5.4.1.2 Vertical Slab and 2D View	114
	5.4.2 Finger Properties and Analysis	115
	5.5 Finger Tracking and Temporal Context Calculation	116
	5.5.1 Simulation Analysis	117
	5.5.1.1 Temporal Context Visualization and Filtering	117
	5.5.1.2 Time Chart View	118
	5.5.1.3 Finger Forest View	119
	5.5.2 Simulation Summarization and Ensemble Analysis	120
	5.5.2.1 Ensemble Analysis	120
	5.5.3 Design and Implementation	122
	5.6 Evaluation	125
	5.6.1 Domain Expert Scenarios	125
	5.6.1.1 Exploring Finger Formation	125
	5.6.1.2 Similar Simulation Analysis	127
	5.6.2 Domain Expert Feedback	128
	5.6.2.1 Instantiation Expert Feedback	128
	5.6.2.2 Theoretical Model Feedback	130

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
	5.7 Discussion	131
	5.7.1 Model Summary	131
	5.7.2 Relationship to Other Models and Theories	133
	5.8 Conclusion	133
6	DISCUSSION AND CONCLUSION	135
	6.1 Discussion	135
	6.2 Conclusion	137
	APPENDICES	143
	Appendix A	144
	Appendix B	145
	Appendix C	147
	CITED LITERATURE	149
	VITA	162

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	Observational Astronomy Data Analysis	20
II	Observational Astronomy Task Analysis	23
III	Function Protein Mutation Task Analysis	61
IV	Patient Characteristics and Post-therapy Side Effects	80
V	Toxicity Outcome Distributions of the Spatial-Metric Groups	93

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Munzner’s Nested Model of Visual Design and Evaluation	10
2	Cross-correlated large-scale overlays	17
3	Astroshelf client-server architecture	28
4	Astronomical Map Projections	31
5	Astroshelf Interface	34
6	Trend image of 100 quasi-stellar (quasar) objects	38
7	Quasar spectral plot	40
8	Case Study 2: SDSS J1352	44
9	Case Study 2: Trends in Type 1a Supernovae	46
10	Trend image for 200 galaxies, stars, and quasars around the Galactic North Pole	48
11	Search results from the Sloan Digital Sky Survey Catalog	53
12	FixingTIM visual interface	58
13	FixingTIM client-server architecture	62
14	Trend image coloring scheme	63
15	Visual comparison of scTIM/dTIM	65
16	Identified functional mutations	68
17	Pipeline detailing the Lymph Node Similarity Methodology	77
18	Topological Map and Graph Representation of the Head and Neck	81
19	Example Similarity Ranking of OPC Cancer Patients	86

LIST OF FIGURES (Continued)

<u>FIGURE</u>		<u>PAGE</u>
20	Spatial Similarity Measure Vector Construction	87
21	Spatial vs Categorical Node Patient Categorization	91
22	Spatial Similarity Dendrogram	94
23	FingerFinder Interface	106
24	Finger Calculation and Exploration Process Workflow Decomposition	109
25	Detail and Spatial Context View	114
26	Temporal Context Visualization	118
27	Kiviat Diagram Panel	121
28	Snapshots of Details-First Design Process	123

LIST OF ABBREVIATIONS

AGN	Active Galactic Nucleus
AJCC	American Joint Committee on Cancer
CECT	Contrast-Enhanced Computed Tomography
CFD	Computational Fluid Dynamics
DDS	Data-Driven-Spots
Dec	Declination
DR7	Data Release 7
dTIM	Defective Triose Phosphate Isomerase
FASTA	Fast-All
FIRST	FIRST Images of the Radio Sky at Twenty cm
FITS	Flexible Image Transport System
FPM	Finite Pointset Method
FT	Feeding Tube
g-r	Green - Red
HAC	Hierarchical Agglomerative Clustering
HNSCC	Head and Neck Squamous Cell Carcinom
HSV	Hue, Saturation, Value (color space)

LIST OF ABBREVIATIONS (Continued)

IMRT	Intensity-Modulated Radiation Therapy
LN	Lymph Node
LSST	Large Synoptic Survey Telescope
MCV	Multiple Coordinated Views
MDACC	MD Anderson Cancer Center
NIH	National Institute of Health
OPC	Oropharyngeal Carcinomas
PCP	Parallel Coordinates Plot
PDB	Protein Data Bank
Quasar	Quasi-Stellar Object
RA	Right Ascension
RP	Retropharyngeal
SAS	Science Archive Server
scTIM	Saccharomyces Cerevisiae Triose Phosphate Isomerase
SDSS	Sloan Digital Sky Survey
S&E	Science and Engineering
SIN	Slant Orthogonal Projection
SPC	Superior Pharyngeal Constrictor

LIST OF ABBREVIATIONS (Continued)

TAN	Gnomic Projection
TIM	Triose Phosphate Isomerase
WCS	World Coordinate System

SUMMARY

The scientific community faces an increasing amount of data due to advances in acquisition technologies. Because of the ubiquity of large data repositories in today's society, scientists have begun to analyze and explore the abundance of data now at their fingertips to both generate and test new hypotheses. However, spatial features often are an essential trait of these large-scale scientific datasets. While established visualization guidelines can provide recommendations that are generalizable enough to be applied to a variety of problems, they are often designed for abstract data (no spatial attributes) of moderate scale. Selecting an effective approach in these situations can be challenging for both novice and veteran researchers, especially when it may not be apparent wherein the data to begin the analysis or how to go about doing so. Visualization design strategies can help lessen this burden by guiding questions such as what data to show, how to display it, and if necessary, in what order to arrange it.

This dissertation examines design strategies for visualization collaborations involving spatial data; in particular, how the data and task abstractions, workflow processes, and user expertise affect the decisions behind the design strategies of visualization collaborations involving spatial data. These strategies are actualized through the design and implementation of four integrated systems that demonstrate their effectiveness across the four spatial data problems in science and engineering domains. The merits and limitations of this work are supported through an analysis of each domain problem by demonstrating the complexities involved with interdisciplinary research and the necessity of working directly with domain scientists and their data. These strategies can serve as a reference for researchers who are working on endeavors that similarly characterize to those described in this dissertation.

CHAPTER 1

INTRODUCTION

1.1 Motivation

As our society becomes increasingly sophisticated, so will the methods of space exploration, drug development, and treatment of patients in hospitals. Whether the goal is to discover how events in far off galaxies relate to those in our own, develop more effective drugs, or further personalize cancer treatment on a per-patient level, visualization serves as an invaluable interface between humans and data through which discovery and insight might be gained.

As a corollary to these progressions, the scientific community faces an increasing amount of data due to advances in acquisition technologies. The emergence of this fourth science paradigm as a new standard of scientific exploration and discovery to the existing three paradigms of science (empirical evidence, scientific theory, and computational science) has created a shift in how scientists are choosing to conduct their day-to-day research [4]. Because of the ubiquity of large data repositories in today's society, scientists have begun to analyze and explore the abundance of data now at their fingertips to both generate and test new hypotheses.

However, analyzing this wealth of information can be difficult for scientists, especially in situations when it may not be clear wherein the data to begin their analysis. In addition to the increase in scale, many of these datasets contain multiple aspects of the same data. These aspects provide unique, yet often complementary, views into the various attributes that define the dataset. For instance, a biological

dataset for identifying functional mutations throughout a family of proteins may consist of the captured sequence and 3D structural data for each of the proteins, as well as derived information on how these proteins relate to one another within a family and the summary statistics between the family members. Because of the heterogeneity of this dataset (i.e., spatial structures and non-spatial sequences, family statistics, etc.), an application with a single graphical view would not be sufficient to provide biologists with information for each of the different data aspects at once; in other words, the design of such an application would require more than one view of the data, simultaneously. This visualization design challenge is not unique only to biological domains; spatial features are also an essential trait of datasets in many other science and engineering (S&E) domains.

Additionally, the design of these applications must also facilitate the breadth of skill and expertise supplied by the members within a scientific collaboration. In endeavors that feature such an assortment of complementary user expertise, it is often beneficial that the collaborating members work together on the same data in the same interface. In these situations, this interface must facilitate the needs of each of the members, some of whom may wish to view different aspects of a dataset that contains a mixture of data elements with and without spatial data attributes (i.e., data with intrinsic cartesian coordinates). For example, in addition to providing views for the multiple aspects of data, the design of a visualization application for identifying functional protein mutations must also consider the various data abstractions, visual encodings, and interaction affordances that users with expertise in both molecular biology and bioinformatics would be most familiar [2].

However, selecting an effective design for any visualization collaboration can be challenging for novice and seasoned visualization researchers, alike. In S&E collaborations, determining how to begin

these endeavors is often complicated by the multiple aspects of the data and the varying expertise of target users. In this data-intensive world, visualization design strategies can help lessen the difficulty of new collaborations by guiding questions such as what data to show, how to show it, and if necessary, in what order to show it.

And yet, the concept of a *design strategy* does not carry a precise definition. Separately, the general dictionary defines *design* as the specification of elements to be created that details their aesthetics, function, and arrangement, and *strategy* as a plan of action devised to achieve a specific goal. In the visualization literature, the term *strategy* tends to follow its general definition while *design* has been defined multiple times. Munzner [5] defines design as a “creative process [...] to select one of many possible good choices from the backdrop of the far larger set of bad choices” and that it encapsulates the “generation and validation of data abstractions, visual encodings, and interaction mechanisms.” We note this definition does not include decisions related to the arrangement of visual elements, which is nevertheless discussed in later chapters of the textbook as an additional design decision that must be considered. We adopt a similar definition for design as Munzner and define a *design strategy* to be an actionable plan focused on generating appropriate and effective data abstractions and visual encodings that satisfy the requirements of the expert. By doing so, we extend the textbook definition to explicitly emphasize the *function* and *placement* of the visual abstractions within a target visual interface, the *interactions* between them, and how these decisions align with the execution of interrelated processes within the experts’ workflows.

With this definition in mind, this dissertation examines how the data and task abstractions, workflow processes, and user expertise affect design strategy decisions for S&E visualization collaborations that

involve spatial data. As we have already seen, these collaborations may not always provide designers a clear path in which to begin their analysis. In these situations, visualization guidelines can provide useful, general recommendations for how to approach these endeavors. Therefore, let us begin our investigation by first considering relevant guidelines that are often prescribed in the visualization literature as useful starting points in visualization design and the terminology behind them.

1.2 Relevant Visualization Paradigms

Over the years, research in the field of visualization has produced various design models and guidelines that aim to provide a useful starting point for designing visualization applications. A common goal of these approaches is the design of techniques that provide the user with an awareness of either the entire (*overview*) or reduced (*context*) information space to facilitate further investigation of regions of interest (*details*). In this section, we examine two well-established visualization paradigms as well as a third introduced later in this work (Chapter 5). Parts of this section originally appeared in our Details-first manuscript [3] presented in Chapter 5. For the sake of clarity, some of the paragraphs have been reworked for use here.

1.2.1 Overview-first paradigm

Among the well-established guidelines for how to design visual interfaces, few are as widely cited as Shneiderman’s 1996 Visual Information Seeking mantra: “Overview-first, zoom and filter, then details on demand” [6]. The mantra provides an intuitive guideline to the interplay between the need to first give the user a *broad awareness* of the dataset (known as an *overview*), the need to show information about each of the individual data points, and the appropriate stage of the analysis in which to do so [5]. Guidelines such as those offered by Shneiderman’s mantra (henceforth, Overview-first) provide

recommendations that are generalizable enough to be used by novice users and be applied to a variety of visualization problems with abstract data (no spatial attributes) of moderate scale.

1.2.2 Search-first paradigm

However, it is often the case in S&E endeavors that the target users are experts within the domain and are working with large-scale, multidimensional datasets. In these situations, creating an overview for top-down analysis may not be feasible. As further argued by van Ham and Perer [7] in their alternative “Search, Show context, Expand on demand” mantra for large graphs, there is also a significant class of scientific users who are not interested in global patterns in the data but have specific questions about one or several specific data points. This alternate approach to visual analysis is similar to the online map process (henceforth, Search-first), where search results provide the starting point for exploring local neighborhoods (known as the *context*). As a practical example, an astronomer who studies a class of quasars is typically not interested in an overview of the entire observable universe, but rather a subset of the data containing observations based on some search criteria [1].

1.2.3 Details-first paradigm

Last but not least, there are situations where providing an initial overview is not relevant or practical for users, while providing direct access to specific features (*details*) is paramount. However, details in S&E endeavors are often spatial features that do not have a precise definition. Instead, their identification relies on internalized knowledge in the domain expert’s head, without which these details cannot be searched against or aggregated to provide the user with a context or an overview. For example, in computational fluid dynamics (CFD), domain scientists often work on the same problem for months

and have a good mental overview of the underlying data [8]. Nevertheless, visualization textbooks only report on the Overview-first and the Search-first mantras [5].

For these situations, this work presents a novel, alternative approach in Chapter 5 – “Details-first, Show context, Overview last” – that supports situations where the main user workflow is centered around spatial or spatiotemporal feature analysis. Unlike the Overview-first and Search-first paradigms, the problem overview can only be observed as non-spatial summarization statistics of or across simulation runs. From an information theory perspective, Chen et al. [8] argue briefly that in such cases, having the direct ability to reach a detailed view (henceforth, Details-first) would reduce the cost of step-by-step zoom operations. Other arguments against first presenting global overviews to users are more practical. First, details describing spatial features may not carry a precise definition, and thus may not be readily available to generate an initial overview. Furthermore, creating an overview may also not be feasible, especially in the case of large-scale multidimensional datasets that are maintained at a centralized location and transferring it to multiple client machines is not an option [7; 9; 10]. Finally, in some scientific problems such as simulation ensemble visualization [11], the problem overview is not one spatial dataset, but a collection of datasets, whose summarization in an overview is not necessarily clear to the domain expert.

1.3 Key Terminology

Because the terms *overview*, *context*, and *detail* are overloaded in the visualization literature, we must first clarify the meaning and usage of each term in the scope of this work. Again, parts of this section originally appeared in our Details-first manuscript [3] presented in Chapter 5 and have been reworked for use here.

1.3.1 Overview

The meaning of *overview* is diverse in the visualization literature. As Hornbæk [12] notes, many authors [6; 13; 14] write about users gaining an overview of the information space, a process which Hornbæk identifies as “overviewing”. This process is akin to the design concept of “knowledge in the head” or “internalized knowledge” [15]. In this respect, Spence [13] defines overview exclusively in relation to the perceptual and cognitive processes through which an overview is acquired, rapidly and without any cognitive effort. Similarly, Tufte [14] discusses overview as an awareness of the content and structure of an information space, acquired by pre-attentive cues, information reception, and active creation.

Yet Greene et al. [16] and Shneiderman [6] also note that “an overview is constructed from, and represents, a collection of objects of interest”. Munzner’s [5] discussion of overviews touches on both aspects: “broad awareness of the entire information space [...] and all items”, but also “when the dataset is sufficiently large, some form of *reduce* action must be used in order to show everything at once.” Last, while Shneiderman [6] discusses overviewing in his mantra paper as “seeing the entire collection,” the mantra and subsequent examples refer to overview in the sense of a technical, user interface component (“knowledge in the world” [15]).

In this work, we adopt the Munzner dual definition. To distinguish between the two common uses, “spatial overview” denotes the spatial overview of one simulation (often internalized by domain experts), and “summarization overview” denotes a collection of objects of interest, constructed through reduction of the entire information space.

1.3.2 Context

In general, context can denote either 1) a global setting, or 2) a local circumstance. In the visualization literature, the concept is similarly used. When describing the concept of focus+context, Card et al. [17] equate context with overview (a global view at reduced detail). Doleisch et al. [18] also describe context as “the rest of the [spatial] data”, at a lower resolution, or in reduced style, (i.e., using translucency). More generally, Furnas [19] explains that context, conceptually, is “any presentation of an information structure” that helps the user “to extract meaning, to understand something about [another focused/original] structure.”

The van Ham and Perer [7] construction and usage of context is consistent with the locality aspect of the general vocabulary definition. In our work, context is defined similarly, along its locality aspect.

1.3.3 Detail

In general, detail denotes an individual feature, fact, or item. In the Overview-first paradigm [6], a detail is “implicitly defined in contrast to overview” [12]. Contrary to Shneiderman and Spence [13], Tufte does not contrast between overview and detail, and instead suggests that “to clarify, add detail.” Munzner describes “a more detailed view that shows a smaller number of data items with more information about each one” [5]. As described above, details are often spatial features in S&E visualization collaborations such as those found in this work.

1.4 Scientific Workflow Theory

So far we have examined three visualization paradigms – Overview-first, Search-first, and Details-first – that can offer guidance for new visualization collaborations. However, while some S&E problems involving spatial data pair well with a particular paradigm (e.g., the query-based workflows found in

observational astronomy lend themselves nicely to a Search-first design strategy), others do not have an obvious pairing when determining which approach to apply. In these situations, considering the problem from a scientific workflow theory [20] can help determine which paradigm, or multiple paradigms, best fit the spatial data problem at hand.

Casually speaking, a workflow is an abstract description of the tasks required for executing a particular real-world process, and the flow of information each of the tasks [21]. More specifically, scientific workflow theory borrows heavily from business workflow modeling and decomposes each workflow into three components [20]: data, control, and (human) resource components. For each workflow, the *data* component captures the information that is required during the execution of a workflow; the *control-flow* component describes the set of steps that make up the process and the way in which the thread of execution is routed between them; and the *resource* component identifies the people and facilities that actually carry out the process. By capturing these elements for a particular problem, the design components corresponding to overview, context, and details can often be easily identified. This decomposition can further assist with which approach to apply and when to apply it.

1.5 Challenges

However, even with all of the information presented in this chapter so far, developing visualization design strategies for S&E collaborations still presents several significant challenges. First, it is often difficult to determine the underlying requirements of domain scientists when developing design strategies for S&E problems involving spatial data. From vortices in flow simulations to the bonding sites on protein structures, spatial data is at the very core of most S&E collaborations. However, these datasets commonly describe information about physical, spatial structures that can be unique to the domain in

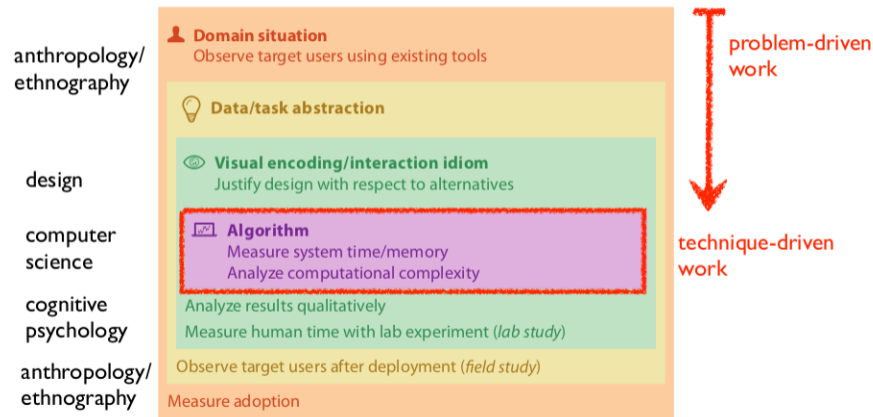


Figure 1. Spatial data problems must be evaluated following the three nested levels of problem-driven work [22].

which they originate. Many of the problems that confront scientists working with these datasets can also be unique in that the data and tasks are tightly coupled with the target domain. Furthermore, we have previously seen that these datasets – many of which are large-scale and heterogeneous – may also consist of multiple data attributes that, depending on the expertise of the target scientists, must be taken into account in the design. One approach to mitigate this challenge is to work together with domain scientists and their data to help determine which visualization design strategies best align with data and tasks of the target domain.

A second challenge when developing visualization design strategies lies in how we evaluate their success. Visualization theories and models span a wide range, from mathematical abstractions and frameworks to guidelines and novel interpretations of different aspects of the development of visualizations in particular contexts [23; 24]. As a result, validations of the resulting theories and models also

cover a range of approaches. For example, the Search-first mantra was introduced along with a constructive example in the domain of large graphs of citations [7], with reported usage cases and no domain expert feedback. Additionally, the visualization design literature has shown that a model or theory can also be acceptably supported by as little as one to a few concrete examples coming from the experience of one to a few authors [25; 26; 27; 22]. One notable exception can be found in Shneiderman's instantiation of the Overview-first mantra, where its "notable theoretical development" [28] of was not accompanied by supporting evidence [6].

Another common approach to validating a visualization design is to test it using the components of Munzner's nested model of visual design and evaluation (Figure 1) [22]. In problem-driven design, the success of the solution can only be evaluated once the nested levels of Munzner's model have been completed, from characterizing the domain to proposing the appropriate visual encodings. In short, this means that we must complete the nested model before we can observe success. At the same time, any of these components may influence the success of the outcome, and it may be difficult to tease out the influence of the layout strategy as opposed to the influence of the visual encodings. However, it may be possible to experiment only with the layout and observe the outcome.

Finally, extrapolating generalizable strategies from examples of successful collaborations can prove to be the most difficult challenge in the development of design strategies for S&E problems. Again, the close collaboration between visualization researchers and domain scientists can yield strategies that have been specifically designed to solve the target problem of the expert. These collaborations often produce solutions highly tailored to the domain and can create challenges for designers looking to emulate the same success. Unfortunately, such insight would require ethnographic studies to determine how the data

and task abstractions, user expertise, and workflow processes affect the decisions behind their presented design strategies. While we cannot entirely address this challenge in this dissertation on empirical evidence alone – four examples of successful design strategies cannot constitute proof – we can identify the commonalities between the results of our solutions using scientific workflow theory [20], which has been tested in other domains.

1.6 Contributions Overview

We ground our work by focusing on the design space of four specific spatial data problems – spectroscopic analysis of galaxies in observational astronomy (Chapter 2), functional mutation analysis in protein families in computational biology (Chapter 3), lymph node metastasis in radiation oncology (Chapter 4), and viscous finger evolution in mechanical engineering (Chapter 5). From an applications standpoint, our goal is to develop tools that solve real-world problems. In doing so, this dissertation describes the design decisions from domain characterization to evaluation regarding the data and tasks identified by their target user workflows and examines the use of the three discussed paradigms (Overview-first, Search-first, and Details-first) for developing the design strategies behind the each presented solution.

Specifically, this dissertation presents the following contributions: 1) descriptions of the tasks and data associated with specific problems related to spectroscopic analysis of galaxies in observational astronomy, functional mutation analysis in protein families in molecular biology, lymph node metastasis in radiation oncology, and viscous finger evolution in mechanical engineering; 2) several novel visual representations and techniques that demonstrate how the Overview-first and Search-first mantras can be used for spatial data analysis; 3) a “Details-first, Show context, Overview last” approach for the ex-

ploration of large-scale spatial data centered around feature analysis; 4) the design and implementation of four integrated systems that demonstrate the effectiveness of our design strategies across the four described domains situations; 5) the deployment and expert evaluation of these four integrated systems; 6) and a discussion of the merits, applicability and limitations of our presented design strategies. The design strategies presented throughout this dissertation can be referenced as guidance for similar spatial data problems.

The first contribution of this work – descriptions of the tasks and data associated with specific problems in the domains of observational astronomy, molecular biology, radiation oncology, and mechanical engineering – details the basis on which our presented solutions were built. These analyses may also serve as a starting point for researchers who are interested in or are currently working with similar spatial data problems as those described in this dissertation.

The second contribution of this work – several novel visual representations and techniques – demonstrates how the Overview-first and Search-first mantras can facilitate spatial data analysis. While the concepts of filtering and aggregating data to provide global and contextual visual representations are by no means new, they are not sufficiently researched in relation to spatial data attributes. Specifically, this work demonstrates techniques for filtering and aggregating these datasets according to their spatial properties to create effective visual representations.

The third contribution of this work – a Details-first approach – presents an alternative approach to the Overview-first and Search-first mantras that supports situations where the main user workflow is oriented along spatial or spatiotemporal feature analysis, while the problem overview can only be observed as non-spatial summarization statistics of or across simulation runs. We construct this approach

using theoretical evidence from scientific workflow theory and practical evidence from the domain of computational fluid dynamics to support our “Details-first, Show context, Overview” last exploration paradigm. Furthermore, we demonstrate the effectiveness of this approach on a large-scale, spatial dataset and examine its relationship to other design strategies.

The fourth and fifth contributions of this work – the design, implementation, deployment, and evaluation of four integrated applications – demonstrates how the three strategies (Overview-first, Search-first, and Details-first) can be applied to solve the real-world problems of real-world users in various spatial data, S&E domains. The integrated applications are developed in close collaboration with domain experts to solve complex domain-specific problems.

The last contribution – a discussion on the merits, applicability, and limitations – reflects on each design strategy in the context of the spatial data problem the collaboration aimed to solve and discusses potential future applications that our strategies might benefit as well as areas in our design that may require further research.

CHAPTER 2

LARGE-SCALE OVERLAYS AND TRENDS: VISUALLY MINING, PANNING AND ZOOMING THE OBSERVABLE UNIVERSE

This chapter was originally published in the IEEE Transactions on Visualization and Computer Graphics (TVCG) Journal © in 2014 [1]. This version has been edited to be consistent with the rest of the dissertation. Coauthors on the original work include Brian Cherinka (BC), Daniel Oliphant (DO), Sean Myers (SM), W. Michael Wood-Vasey (MWV), Alexandros Labrinidis (AL), and G. Elisabeta Marai (GEM). The contributions from each author included: BC and MWV served as our observational astronomy domain experts, providing the theoretical underpinnings for the trend image design and provided support with the software testing and the design of the case studies; DO designed the first versions of the client-server architecture, the Data-Driven Spots, and the original image stitching algorithm behind Astroshelf; SM contributed to the implementation and refactoring of the front-end client and SkyView; AL served as a co-principal investigator on the project for the data management of the project and recommended the use of the S^2 -tree for indexing the images; GEM served a co-principal investigator on the project for the visualization aspects of the project and directed the top-level design, implementation and testing of the tool. My (TL) contributions to this work included the design and implementation of the pixel-based trend images, postage-stamp images, and interactive SkyView. Additionally, I worked with BC to project the various survey images (SDSS, FIRST, and LSST) for their stitching and cross-registration in the SkyView and helped design and develop the current version of the client-server architecture and AL and his data management team. I am the first author on this work.

This chapter begins our investigation of spatial data design strategies with the problem domain of observational astronomy, where advances in telescope technology have resulted in astronomical workflows needing to handle data of both increasing scale and variety. While the data in this domain characterizes as both big volume (e.g., large numbers of observation and varying scales) and big variety (e.g., catalogs, spectra, and images), many observational astronomy workflows are designed around the goal of discovering correlations within moderate collections of observations; not the entire observable universe. In our collaboration with observational astronomer (coauthors BC and MWV), we found that their workflows typically began with query-based tasks to select, group, and browse collections of observations according to spatial data properties. Based on the execution order of these workflows, we decided that the Search-first visualization design paradigm most closely aligned with the tasks of our collaborating astronomers. Specifically, since the target workflows centered around the search and filter operations, we chose to design our solution with a layout that prioritized the tasks related to first reducing the information space before enabling any further exploration of the data. Following this approach, this chapter introduces a novel computational framework and web-based application, Astroshelf, that facilitates these workflow processes and assists in the visual integration, mining, and interactive navigation of large-scale collections of observations. To address challenges associated with scalability, we present a novel visual representation designed to assist astronomers in identifying trends in large collections of spatial observations.

2.1 Introduction

Advances in data acquisition technology enable astronomers to amass large collections of complementary data, ranging from large scale, gigabit images to spectroscopic measurements. With the insight

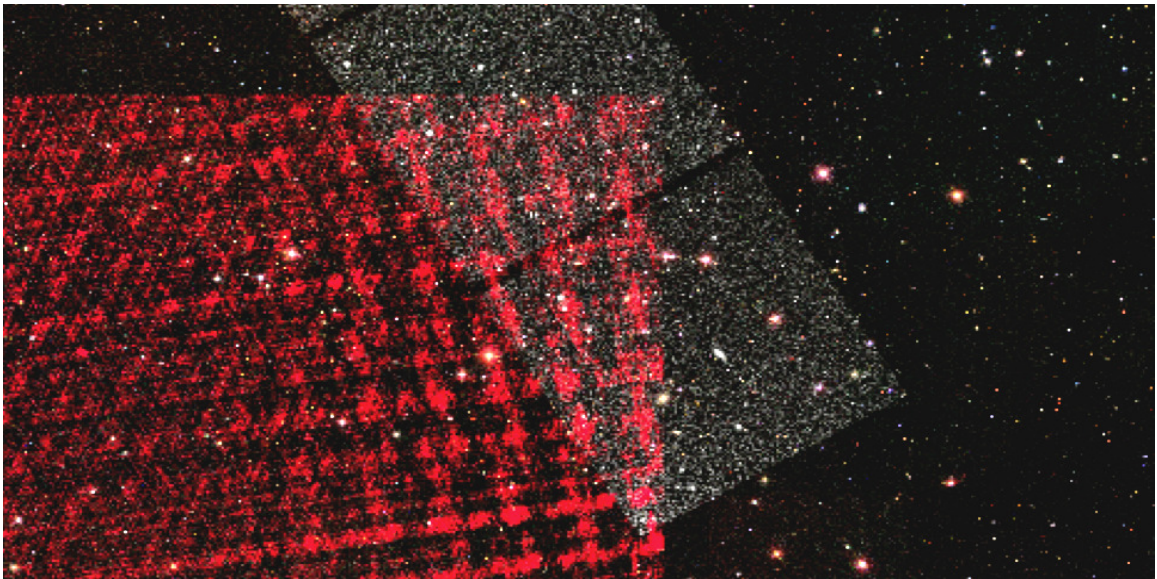


Figure 2. Cross-correlated large-scale overlays of optical observations, radio-emission observations, and simulation results from the SDSS sky survey (color-on-black, full-coverage overlay), the FIRST sky survey (red overlay to the left), and the LSST dataset (gray overlay, diagonal). Transparency can be interactively controlled for each overlay, enabling cross-spectrum analysis. Hardware-accelerated overlays coupled with a web-based client-server architecture allow panning and zooming of gigabit sky panoramas at interactive frame rates.

gained by these observations, researchers can better understand the happenings in our galaxy by studying similar events in distant ones.

However, astronomical workflows are becoming cumbersome due to the increasing scale and variety of data sources. Astronomers gather the data needed for a particular study by querying multiple surveys for images, cross-correlating complementary images of the same object or set of objects, and searching multiple catalogs for potential supporting details. Once collected, astronomers must then flip back and forth between these spatial and non-spatial details and images to gain context. This process is both

tedious and challenging, often requiring hours to complete. As a result, more time is spent gathering data rather than analyzing it.

Additionally, the data’s properties further complicate the design of informative representations and techniques. First, the Information Seeking Mantra [6] fails in this domain due to the volume and density of the data. Coincidentally, the tedium described in the previous workflow closely resembles the Van Ham-Perer mantra [7]; astronomers often perform manual, query-based data filtering operations before analysis. Secondly, the data’s innate spatial features uniquely define the domain’s tasks; for instance, galactic position and age – defined by a spatial attribute, redshift – help to determine morphological similarity. In conjunction, these properties pose challenges to the design and validation of visual representations and underlying spatial similarity.

Inspired by an analysis of observational astronomy workflows, we propose a web-based visual infrastructure for the interactive navigation and mining of large-scale, distributed, multi-layer geospatial data. We introduce an automated pipeline for cross-correlating image data from complementary surveys, and we enable the visual mining of catalog information in conjunction with the large scale image data (Figure 2). A spatially indexed, hardware-accelerated, client-server backbone allows fetching, displaying, panning and zooming of gigabit panoramas in real time.

The contributions of this work (extended from our Best Paper Runner-Up Award [29] at the IEEE Large Data Analysis and Visualization Symposium 2012) are as follows: 1) a formal analysis of the data and tasks specific to the observational astronomy domain; 2) the design of a client-server architecture for the interactive navigation of large scale, complementary astronomy observations; 3) two compact, scalable visual abstractions—hardware-accelerated pixel-based overlays and trend images—to enable

the interactive mining, panning and zooming of these data; 4) a web-based, cross-platform implementation of this approach; and 5) the application of this approach to observational astronomy data through three case studies.

2.2 Domain Analysis

Our first contribution is a formal analysis of the domain data and tasks. This analysis provides a problem-driven basis on which further visualizations and interactions can be built.

Astronomy surveys cover a wide area of sky by acquiring many smaller images — some of which may overlap — over their targeted region. A given survey usually only covers a small fraction of the whole sky. However, the advent of large telescopes like the Large Synoptic Survey Telescope will change dramatically, over the next decade, the scale of these surveys. Different surveys may or may not cover the same area of sky, resulting in possibly completely disparate or overlapping datasets. The Extended Groth Strip [30] for example, is one of the most observed regions of the sky, with upwards of eight different telescopes/surveys collecting data; this region is rich with multi-wavelength observations.

In our experiments we use data from three surveys, the Sloan Digital Sky Survey (SDSS), the Faint Images of the Radio Sky at Twenty Centimeters (FIRST), and simulated results from the Large Synoptic Survey Telescope (LSST). **SDSS** is an optical, wide-field, survey covering a quarter of the sky. Over the past ten years, it has imaged a half a billion galaxies and taken spectra for a half a million, providing a massive leap in the amount of astronomical data (roughly 15 TB raw image data, stored as 2048x1489 pixel field images). **FIRST** is a radio survey of the sky, following the same path as SDSS. FIRST also covers about a quarter of the sky and contains roughly a million discrete radio sources [31]. **LSST** is a future optical full-sky survey, along the same lines as SDSS but of unsurpassed scale. It will cover

TABLE I. Observational Astronomy Data Analysis

Field / Attribute	Data Type (per point)	Visual Mapping
<i>2D Fields</i>		
Optical	RGB Value	Color Overlay
Radio	RGB Value	Color Overlay
Simulated	Intensity	Color Overlay
<i>2D Field Attributes</i>		
Projection Scheme	Formula	Location on Unit Sphere
Stripe Information	Numeric Tuple	Individual Tile Image
<i>Catalogs</i>		
Table of Search Results	Alphanumeric Tuple	Data Driven Spots (DDS)
<i>Object Attributes</i>		
Identifiers	Alphanumeric Value	Detail-on-Demand
Coordinates	Numeric Tuple	Pixel Coordinate
Redshift	Numeric Value	Pixel Intensity
Wavelength & Flux	Numeric Array	Pixel Intensity
Spectra	Numeric Array	Trend Line & 2D Plot
Image	RGB Value	Small Multiple

~20,000 sq. degrees of the sky, scanning the entire sky (visible from the southern hemisphere) every 3 nights, in six photometric bands. LSST will image approximately 3 billion galaxies and will archive about 6.8 PB of images a year. As LSST has yet to acquire sky images, the LSST project has generated simulations of images of the sky to mimic and observe the observational prowess of the survey. Seven fields (189 unique image files), each covering ~10 sq. degrees, have been simulated.

2.2.1 Data Analysis

Astronomers use a variety of data formats to collect, organize, analyze, and share information about the observable Universe. The most common formats used are images and catalogs.

Images are rectangular snapshots (*tiles*) of regions of the sky, typically labeled with the spatial location of the region. Images in astronomy are usually stored as a Flexible Image Transport System (FITS) file. FITS files store image metadata in a human-readable ASCII header, and often include technical telescope details from when the image was taken. FITS files are extremely versatile, capable of storing non-image data such as spectra, 3D data cubes, multi-table databases, and catalog data.

Since the observable Universe is projected onto a sphere, the angle is the most natural unit to use in measuring positions of objects on the sky. Astronomers describe the coordinates of objects in Right Ascension (RA) and Declination (Dec). Similar to how longitude and latitude describe positions of objects on the Earth from a given reference point, right ascension and declination mark the position, in degrees, of an object with respect to the celestial equator.

Catalogs index all of the objects in a set of images. The catalogs contain spatial location information for every object imaged, along with any non-spatial properties collected or calculated from the observations (e.g. brightness, mass.) Each object in the catalog receives a unique identifier. Catalogs generated from the same survey will use the same unique object identifiers, making cross-matching within a survey straightforward. However, as is often the case, when the same object is observed in different surveys, it is assigned different identifiers for each catalog; this labeling makes cross-survey matching a non-trivial task. While the observed objects in each survey may not overlay exactly due

to variation in each telescope’s construction and parameters, they may still be physically and visually associated with each other (e.g. radio jets emanating from the center of a galaxy.)

Among the spatial and non-spatial object properties typically stored in catalogs, *spectra* play a particularly important role. Whereas imaging only captures broad features across the entire object, such as color or shape, spectra capture detailed information on the physical processes in and around the object, such as kinematics, temperature, distance from observer (redshift), and elemental content of gas associated with the object. Spectra specify the *wavelength* distribution of electromagnetic radiation emitted by a celestial object, as well as the *flux* (or intensity) of the object at those wavelengths. Large surveys typically acquire one spectrum per object, resulting in an extremely large number of spectra per survey; for example, SDSS has acquired spectra for ~ 1.6 million objects. Selecting particular classes of objects based on spectra and looking for trends can often lead to valuable insight about that object class.

Table I summarizes the data types typical of observational astronomy, as well as the visual mappings proposed in this work. In summary, the observational astronomy domain features large-scale, distributed, overlapping, multivariate datasets consisting of both spatial and non-spatial data: in a nutshell, data characterized by big volume and big variety. *Big volume* characteristics encompass: large image sizes (gigabit), impacting both rendering and interaction rates; fragmented images resulting in numerous image tiles; multiple scales; and large numbers of both observations and objects, often indexed in collections. *Big variety* characteristics include: data heterogeneity (e.g., catalogs, spectra and images); multiple data sources (surveys); and complementary domain expertise (e.g., expertise in super-

TABLE II. Observational Astronomy Task Analysis

Task	Visual/Interaction Mapping	Technical Challenge
T1 Pan and zoom in real-time	Panning & Zooming	Real-time infrastructure & interaction
T2 Analyze spatial distribution of objects	Object Overlays	Scalable visual abstraction
T3 Cross-correlate 2D image fields	Filtering on Overlays	Image cross-registration pipeline
T4 Identify trends & outliers in an object-collection	Interactive Trend Images	Visual abstraction
T5 Group objects according to properties	Linked Views	Interaction design
T6 Inspect object properties	Linked Views & Details-on-Demand	Visual design

novae as complementary to expertise in transient events). While the data is indexed by object location, uncertainties in the measured position make visual correlation particularly useful.

2.2.2 Task Analysis

The Universe is a complex structure with many physical processes governing its formation and evolution. While space-based telescopes can observe the full electromagnetic spectrum, cost and technical challenges preclude the design of a single all-purpose telescope. Instead, astronomers rely on many telescopes that observe specific regions of the electromagnetic spectrum and then cross-match the datasets to identify the same objects in each one. Astronomers must also manually seek out data related to a particular object.

Astronomical processes occur on many length scales, from small-scale features such as dust particles to large-scale features such as clusters and superclusters of galaxies. With observations usually pertaining to a specific scale at a time, it can be easy to lose the big picture of how all these processes are connected. Therefore it is advantageous to stitch multiple observations together to create a seamless zoomable image. This would allow astronomers to visually explore how stellar and galactic physical processes relate to the larger picture of galaxy groups and clusters.

Browsing massive astronomy collections of objects provides additional challenges. For example, selecting particular classes of objects based on spectra and looking for trends can often lead to valuable insight gained about that object class. However, when studying trends within astronomy objects of a specific class, typical efforts rely upon inspecting individual objects on an image in the sky or in a database table. The approach takes enormous amounts of time. Furthermore, outliers in the dataset can often skew scientific results and must be located and removed before any analysis can be performed. Typical hypotheses relate to identifying trends, common properties, outliers, and discrepancies in collections of objects. Typical operations relate to grouping, selecting and analyzing objects from a collection.

Table II summarizes the tasks and challenges typical to observational astronomy. In summary, the observational astronomer workflow involves both queries of the type *what – where – correlated-with-what* and tasks of the type *browse – group – analyze* over multiple surveys at multiple scales. In conjunction with the big volume and big-variety of the data, astronomers seek the ability to interact with and compare multi-field data for a large number of objects and images.

Last but not least, additional requirements gleaned from interviews referred to desired interaction rates, ease of use, learning curve, and cross-platform desirability.

2.3 Related Work

Multiple attempts have been made to facilitate the observational astronomy workflows. However, none integrate large scale distributed astronomy research data while attaining interactive visual mining, panning and zooming framerates.

Google Sky [32] is a primarily educational, interactive, scalable view of the Sloan Digital Sky Survey (SDSS). While it provides a friendly and clean interface, it also relies on local copies of the data to

the exclusion of multiple surveys; which is a limiting factor for astronomy researchers. An additional drawback is its inability to integrate and share catalog data from multiple datasets. The National Virtual Observatory [33] (NVO) is a service designed primarily for aggregating and cross-matching information from multiple surveys. While it provides some form of catalog cross-registration, the NVO has a cumbersome interface which lacks a much-needed interactive visual component. The World Wide Telescope [34] is a Microsoft Research, primarily educational project designed to allow users to view the Universe with a large, high resolution image of the sky. The ability to overlay multiple maps and visually cross-match objects is nonexistent. There is also a lack of connectivity with catalogs and other scientific data. Last, a variety of institutions have created web interfaces for accessing astronomical data, either for querying specific astronomy databases [35; 36], or for aggregating data on many objects from multiple catalogs [36; 37]. These interfaces either lack a visual interface entirely or they provide only a static sky image to view a few objects at a time. Visual overlays of cross-matched data are non-existent and the user interfaces require a steep learning curve.

Attempts to work with gigascale image data have been made in other domains, though none have been applied directly to observational astronomy. Saliency Assisted Navigation identifies areas of interest in gigapixel images [38]. Through preprocessing and filtering regions of interest, discernible locations in a scene can be presented interactively. Kopf et al. [39] and Machiraju et al. [40] have also developed systems for dealing with gigascale and terascale image data. While these systems have complementary strengths in terms of the storage and the scale of the data being manipulated, each was generally designed for local geospatial data, and not for distributed geospatial sources.

Architectures for multi-zoom large-scale visualizations have also been explored. Space-Scale Diagrams [41] have been used in many geospatial applications [42; 43; 44]; they serve as a basis for our navigational approach. However, earlier applications were not designed to handle the magnitude of data described in this work. The challenges of indexing astronomy data are discussed in Page’s indexing discussion [45]; we use a new indexing approach, based on a *Geohash* [46]. A step further, ZAME [47] has used GPU-accelerated rendering to deliver interactive framerates to multi-scale visualizations. While the ZAME approach is beneficial to client-based applications that are able to provide full graphics support, web-based applications like ours pose more stringent constraints (e.g., limits on how many textures can be passed to a shader at once). However, advances in web-based technologies have been gradually mitigating these constraints, and the advantages of web-based approaches (e.g., built-in cross-platform support and inherent access to online data collections) far outweigh these shortcomings. Furthermore, panning and zooming is a common problem among geospatial applications [48; 49; 50]. While many of these works focus on interactive techniques relevant to this project, the focus of this chapter is an efficient architecture for viewing and cross-correlating gigabit image data.

Presenting multivariate data visually is also common among geospatial applications. Oriented Slivers provides a method to visualize multivariate information simultaneously on a single 2D plane, but becomes easily cluttered as the dimensionality of the data rises [51]. Heat maps [52] alleviate this problem by assigning each value a temperature and producing a color map based on the resulting heat combinations. While particularly beneficial in giving a general overview of data over large areas, heat maps are less useful in identifying individual data points. The approach we adopt for overlaying in-

formation in the sky, Data Driven Spots (DDS), addresses both of these concerns via a pixel-based visualization [53].

The spectra data associated with sky objects are instantiations of ordered single-index tables; in which, however, the index itself is a property. Plain index tables have in general many possible visual mappings, from time-series line charts [54] and bar charts to graph-views, scatterplots [55], colored matrix cells [56] and 3D representations [57]. However, such mappings typically suffer from scalability issues. To address scalability concerns, we follow a pixel-based approach inspired by the compact representations of Keim [58]. Unlike existing pixel-based work which enables comparison through a small-multiple paradigm, however, our approach leverages alignment and resampling of the table data based on the index property. We further employ sorting properties of the object collection in order to generate a single, composite interactive image.

2.4 Design and Implementation

Based on the domain data and task analysis, we design a pipeline for the interactive exploration of observable astronomy data. Given the multiple, distributed sources of data, and the scale of the data, we follow a client-server model (Figure 3). The online processes of this architecture are user-demand driven and occur in real-time. While, in a certain sense, our work aims to create a “Scientific Google Sky”, we note that due to different requirements Google Sky uses a different—albeit unpublished—infrastructure, organization and implementation than our system.

Our server handles requests for image data and catalogs; it includes an offline module for preprocessing astronomical images. Where applicable, to enhance real-time panning and zooming (and thus help support task types T1 and T3), we assign prefixes to images, then organize and store them in a

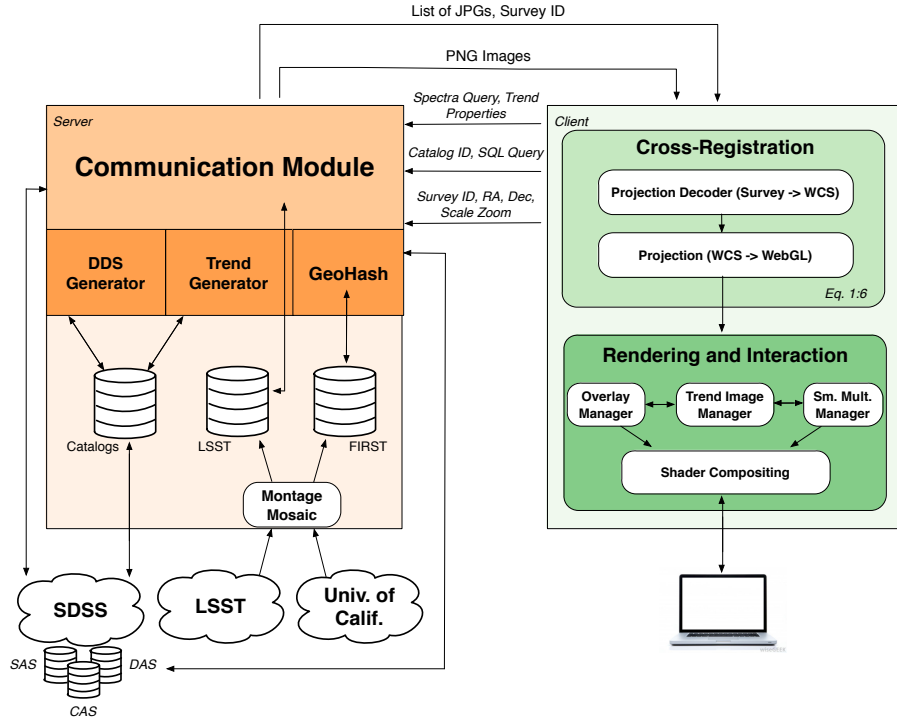


Figure 3. Client-server architecture for the interactive exploration of observable astronomy data. On the server side, our offline module (light orange) preprocesses raw astronomical datasets through *Montage Mosaic*; where applicable, we assign prefixes to offline images, organize and store them in a spatially indexed, prefix-matching structure (*Geohash*). Online server-modules abstract catalog results into visual DDS overlays (**DDS Generator**) and handle the construction of trend images (**Trend Generator**). The client handles the tile stitching and cross-registration process into Gigabit, zoomable panoramas, manages the rendering and interaction for the overlay view, the trend and small multiple views; and composites the images using hardware-accelerated shaders. The only third party tools are *Montage Mosaic* and the MongoDB powering the *Geohash*.

spatially-indexed, prefix-matching structure (*Geohash*). We store catalogs locally. Our online Data Driven Spots (DDS) module abstracts catalog results into an image overlay (task type T2). Our additional Trend Image interactive module allows the users to visually construct and browse collections of objects (tasks T4 and T5). Finally, our master Communication Module handles communication with the client, and interfaces with the data and the other server modules.

The client handles the requests from the user and the view management process. We support through a separate module the tile-stitching and cross-registration pipeline (T1 and T3). A second module controls the rendering and interaction processes through a web-based interface; ultimately, the module presents the stitched images and catalog results to the users in the form of gigabit panoramic overlays, interactive trend images, and small multiples (tasks T1 through T6). As the user navigates the sky, the client queries the server with the current field-of-view or desired catalog information.

Below we describe in detail the server and client modules which, independently and in conjunction, meet the technical challenges identified in Table II: real-time panning and zooming capabilities, an image cross-registration pipeline, and scalable visual abstractions and interactions. The online modules are implemented using web-based technologies, including WebGL, HTML5 and JavaScript.

2.4.1 Data Retrieval and Preprocessing

We perform image data retrieval and preprocessing on a per-survey basis. To ensure survey and dataset compatibility, we extract the RA/Dec coordinates for each image tile so that images from multiple surveys will be properly aligned.

For surveys which benefit from an online programmatic interface, like SDSS, our system implements simple scripts to access the data remotely. Image data for the SDSS survey are stored remotely through

various data releases; each release consists of FITS files and frame images. To access the survey data, our server sends SQL queries and fetches online images from the SDSS Data Access Server (DAS) and the SDSS Science Archive Server (SAS).

When a programmatic interface does not exist (e.g., FIRST or LSST) we fetch the sky images a-priori and store them locally. There are 30,500 FIRST image files, requiring 300GB storage. To ensure compatibility between surveys, the raw data is processed using the third-party tool Montage Mosaic [59], to extract the image data from the raw FITS format; the resulting images are named according to the RA/Dec center of the image.

In the case of the FIRST survey, we perform further optimization to reduce the rendering load when a large area of the sky is being viewed. To this end, we use custom Matlab code to generate a pyramid of image tiles, with four levels (number of levels empirically determined for demonstration purposes) of decreasing resolution. We obtain tiles through repeated Gaussian filtering followed by subsampling. We perform this entire preprocess once for the dataset, averaging a 30 second generation time per tile. Once the local images are preprocessed, we spatially index them for quick access into the geospatial index powered by MongoDB [46], an open source document-oriented NoSQL database system. We hash the coordinates of the image tiles as string-based prefixes through MongoDB’s *Geohash* table.

We do not map locally, however, LSST images to multiple levels of detail, since the domain experts anticipate a future programmatic online interface. Because this small LSST test dataset is privately owned and accessed, we perform the entire procedure a-priori and store locally all images. *Montage Mosaic* and *MongoDB* are the only third-party tools in our system.

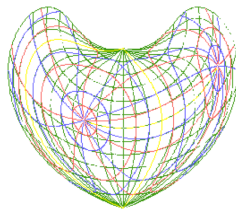


Figure 4. Depending on the survey, astronomical data can appear in different map projections (shown in different colors above). While about 25 different projections are common to astronomy, there is no limit to the number of possible projections available.

2.4.1.1 Catalog Data

Aside from sky survey panoramas, we create more specific overlays from user-performed searches over catalogs. We retrieve catalog data from the SDSS server and store the data locally into a MySQL database.

2.4.2 Cross-Registration and Online Overlays

To support task types T1 (pan and zoom), T2 (analyze spatial distributions) and T3 (cross-correlate images), we follow a cross-registration and online overlaying approach.

2.4.2.1 Sky Panoramas

We create sky panoramas (T1) and cross-correlate images (T3) by stitching together multiple astronomy images into a seamless, zoomable, pixel-based abstraction (Figure 5, center). Depending on the survey, astronomy images can appear in different map projections. While about 25 different projections are common, the number of possible projections is not limited (Figure 4). In our approach we use the World Coordinate System (WCS) specification [60]. Our custom code converts to WCS a variety of

image coordinates given in different projection schemes. SDSS’s projection scheme is Gnomonic (TAN), an azimuthal projection, given by equations 54 and 55 in [61]. The FIRST radio survey uses a Slant Orthographic projection (SIN), also an azimuthal projection, and is given by equations 59 and 60 in [61]. The LSST simulated dataset uses the TAN projection scheme, similar to SDSS.

For the actual cross-registration and stitching we project the images on a viewing sphere. Our sphere is an abstraction of the sky as viewed from Earth, with the camera located at the center of the sphere.

To overlay sky images for viewing, our next step is to convert the WCS coordinates into the native WebGL graphics coordinates. The standard WCS Cartesian coordinate system is a right-handed coordinate system with the positive x , y , and z axes pointing outward, to the right, and up, respectively. In the WCS spherical coordinate system, the angle θ increases clockwise starting from the positive z -axis, and the angle ϕ increases counter-clockwise starting from the positive x -axis. In contrast, in the right-handed WebGL graphics coordinate system the positive x , y , z , axes point to the right, up, and outwards, respectively. Furthermore, the angle θ increases clockwise starting from the negative x -axis, and the angle ϕ increases counter clockwise starting from the negative y -axis. Due to these differences between the standard and WebGL coordinate systems, a transformation has to be applied to convert from the world RA/Dec coordinates to the WebGL spherical and Cartesian graphics coordinates.

We derive this transformation as:

$$\phi = (90^\circ - Dec) \quad (2.1)$$

$$\theta = (270^\circ - RA) + 360^\circ \quad ; \text{ when } RA > 270^\circ \quad (2.2)$$

$$\theta = (270^\circ - RA) \quad ; \text{ when } RA \leq 270^\circ \quad (2.3)$$

$$x = \sin(\phi) * \cos(\theta) \quad (2.4)$$

$$y = \cos(\phi) \quad (2.5)$$

$$z = \sin(\phi) * \sin(\theta) \quad (2.6)$$

where Equation 2.1 spells out, for clarity, the *Declination* rotation into the WebGL angle ϕ , and Equations 2.2 and 2.3 spell out the *Right Ascension* rotation into the WebGL angle θ .

Following the above transformation, our approach maps sky images to the unit viewing sphere. Aside from projecting each image tile onto the sky, our cross-registration and overlay module also facilitates zooming in and out to account for telescope parameters, and changing transparency.

2.4.2.2 DDS Overlays

To support task type T2 (analyze spatial distribution of object), we generate custom Data-Driven-Spots overlays from catalog data. In response to the client requirements—desired resolution of the output image, the minimum and maximum RA/Dec values, attribute thresholds, desired color-mapping, and any other optional filters on the other parameters present in the catalog database — our server creates one or more new PNG images from the catalog database. The RA/Dec columns in each tuple

are used to position the drawing within the image (Figure 5, bottom left). The closer the value of the key attribute is to the maximum threshold, the brighter the pixel will be at that location. All data tuples are added to the images, which we then compress and return over the network to the client application.

To create a visual abstraction of multiple data sources (task T3), pixels are further composited on-line into transparent overlays using the WebGL GLSL fragment shader. WebGL has the advantage of performing computations exclusively on the client machine GPU, leaving the CPU available for user interaction. To optimize JavaScript memory use and texture loading we implement local garbage col-

lection; this optimization helps prevent HTML5 from bottlenecking interaction while rendering texture objects.

The client we implement receives the images, cross-registers them through the approach described above, computes the pixel-based overlay and displays it. The client finally renders the scene, where the visualization scene graph consists of the viewing sphere with the camera at the center looking out.

2.4.3 Interactive Trend Images

As outlined in Table II, grouping and regrouping spatial objects according to their properties is a common astronomy task (task types T4 and T5). The custom overlays we enable—based on catalog queries (2.4.2)—facilitate tasks which analyze scalar properties of the objects. Opacity-control further enables establishing correlations amongst multiple object properties.

However, some object properties are non-scalar, but ordered tuples or indexed-arrays for which the index itself is an object property. For example, object spectra are ordered tables of (key, value) pairs in which each key is a specific wavelength and the value is the flux at that wavelength. Analyzing such table properties across collections of objects, in order to establish trends or identify outliers, can be enormously time-consuming. We note that disparate features such as data artifacts that exist in a small set of spectra (amongst the larger pool) tend to be difficult to identify algorithmically or analytically, due to their ill-defined nature and randomness. To facilitate this process, we propose a second compact, pixel-based visual abstraction called an interactive trend image.

Trend images are a novel visual abstraction which relies on aligning and resampling property-indexed table data into a pixel-based representation. The abstraction further requires and leverages

sorting properties of the data across an object collection. The trend abstraction builds on the astronomy concept of a 2D composite image [62].

The input to the trend image abstraction is a collection of objects. Because of their nature, the objects are guaranteed to have at least one scalar property suitable for sorting, p_{sort} —the distance from Earth to each object. Each object also has an indexed-table property in which the keys (a.k.a. index) are themselves a property of the object. We map the values in each object’s table to a row of pixels, and then combine all the rows corresponding to the object collection into a trend image as described below.

Let N be the total number of objects in the collection. Let M be the total number of samples in each object’s table-property—for example, flux, $\rho_{1:M}^i$, indexed by ordered wavelength, $\lambda_{1:M}^i$; where i is the object index ($1 : N$). To generate the trend image, we resample and align the data across the object collection. We generate first a uniformly-spaced basis for the collection of objects returned by the client query. The uniform basis b ranges from the minimum and maximum keys over all the N queried objects, and it is incremented by the desired horizontal resolution r of the image (specified by the client):

$$b = \left[\min_{\substack{i=1:N \\ j=1:M}} \lambda_j^{(i)} : r : \max_{\substack{i=1:N \\ j=1:M}} \lambda_j^{(i)} \right] \quad (2.7)$$

We next generate for each object a normalized table of values, ρ_{norm} , through resampling and interpolation over the uniform basis:

$$\rho_{norm(k)}^{(i)} = \rho_{k-1}^{(i)} + (\rho_{k+1}^{(i)} - \rho_{k-1}^{(i)}) * \frac{(b_k - \lambda_{k-1}^{(i)})}{(\lambda_{k+1}^{(i)} - \lambda_{k-1}^{(i)})} \quad (2.8)$$

where $i = 1 : N$, $k = 1 : |b|$.

Next, a pixel row is generated for each object. Pixels are organized from left to right in the table order and are individually mapped to a color in the HSV space based on the property values in the normalized table.

For color encoding, we calculate the hue of each pixel, H , its saturation S and value V as follows:

$$H_k^i = (\gamma_{high} - \gamma_{low}) * (1 - (b_k^{(i)} - b_{min})/b_{max}) \quad (2.9)$$

$$S_k^i = 1 \quad (2.10)$$

$$V_k^i = (\rho_{norm(k)}^{(i)} - \rho_{low})/(\rho_{high} - \rho_{low}) \quad (2.11)$$

where $i = 1 : N$, $k = 1 : |b|$, and $(\gamma_{low}, \gamma_{high})$ and $(\rho_{low}, \rho_{high})$ are the user-desired color range, respectively the desired property-mapping range.

In the example above, H encodes the wavelength and V encodes the flux value. Other mappings are also possible: for example, encoding flux alone in a grayscale image (Figure 6); encoding rest-frame wavelength ($b_k^i/(1 + z^i)$, where z^i is the object redshift) as hue; or encoding magnitudes for objects which do not have spectra, but do have broadband photometric colors which span specific wavelength regions.

To collect data for an interactive trend image, our client sends to the server queries for sets of spatial objects, and their desired properties. The server Trend Generator module (Figure 3) fetches and processes the data for each object matching the query, and caches the object properties of interest into a

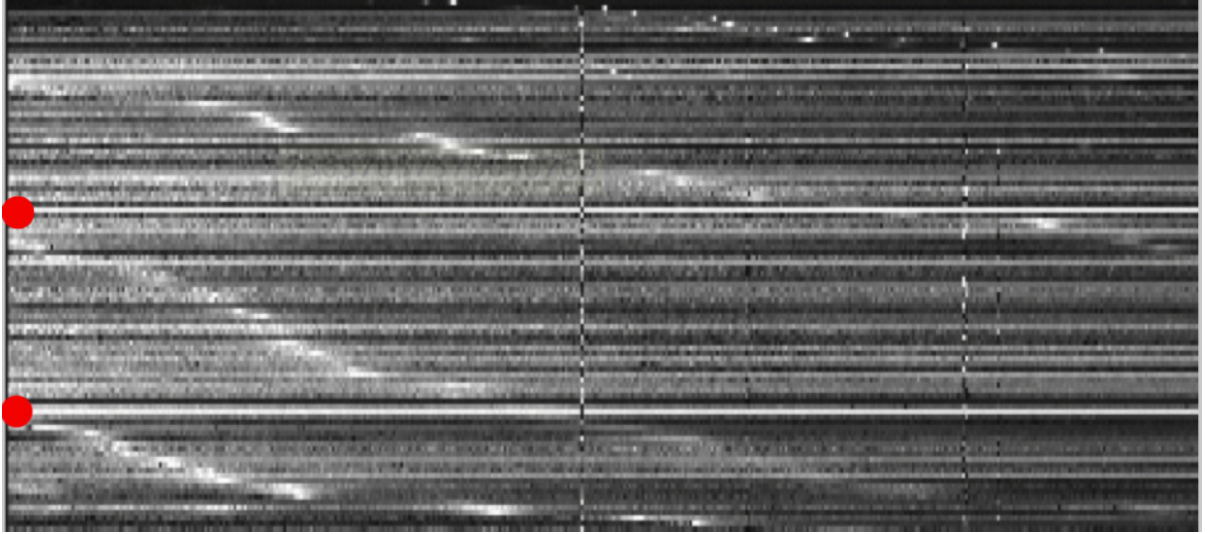


Figure 6. Trend image for a test data set containing 100 quasi-stellar (quasar) objects. Each pixel-row corresponds to the spectrum of a quasar object; quasars are sorted vertically according to their redshift (a distance-related measurement). Note how outliers—quasars with unusual spectra (marked in red)—immediately stand out.

local mySQL database. The user may also specify a color mapping range (γ_{high} and γ_{low}), the desired property-mapping range, and the desired horizontal resolution r of the trend image.

The server returns to the client the collection of pixel-based rows, as highly compressed PNG images. In turn, the client assembles the composite image from the individual rows using the sorting property p_{sort} —for example, the object distance from the observation point. The client also facilitates alternative sortings, where available, and further interaction with the data. Sorting the objects reveals trends in the data, and helps identify outliers. Further interaction through a fish-eye lens and details-on-demand enables the individual analysis of object properties.

Figure 6 shows an example grayscale trend image generated for a test dataset containing 100 quasi-stellar (quasar) objects—extremely remote and massive celestial objects. Since quasars are visually similar in appearance to dim stars, they are difficult to identify from examination of image overlays alone. Instead, astronomers differentiate quasars from other stars by analyzing their spectrum (Figure 7). The trend image shows the quasar spectra as horizontal rows, while the vertical sorting property is the quasar redshift. In this representation, key identifying features become apparent, such as trends in emission lines present in quasars (log-style curves in the figure). Outliers such as quasars with incorrect redshifts or with unusual spectra immediately stand out, as well.

The trend image abstractions allow the user to quickly, and intuitively, identify which spectra amongst the large dataset are unlike the others, or those that belong together in groups. Furthermore, as we show in Section 5, the identification of outliers from the trends is key in identifying both (1) new objects of interest that will bring insight; (2) problems with the previous steps in the data analysis and processing.

2.4.4 Rendering and Interaction

2.4.4.1 Panning and Zooming Large-Scale Overlays

To enable interactive panning and zooming, an Overlay Manager maintains the current viewing location and parameters, as well as a list of the image tiles currently in the view. The manager sends to the server requests for new images, when needed. Panning the view maps mouse motion to updates in the view range. Zooming also computes and maps the new scale to updates in the viewing range. If the updated range covers images that have not been fetched yet, the manager requests for those image

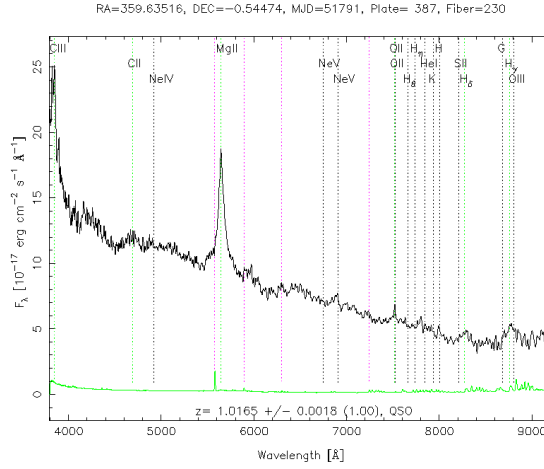


Figure 7. On-demand spectrum for a Figure 6 quasar object. The spectrum plot contains dotted lines to indicate the wavelengths at which atoms are absorbed or emitted.

tiles to be sent out to all overlays that are listening to the current view. Each of those requests is handled asynchronously.

Interactive trend image rendering and the on-demand spectra are handled as a separate process (Figure 5, right), in a side panel from the sky-view. Similar to the custom-overlay creation process, users can specify the parameters to be used in the construction of the trend image and then interact with the resulting set of spectra.

2.4.4.2 Trend Image Interaction

On the client side, we provide interaction techniques to further help support workflows related to object-group analysis. Hovering over the trend image highlights individual rows for inspection; a GLSL shader fish-eye lens can further be activated to magnify a selected row.

Details-on-demand (task T6) are also provided: such as the spectrum for that object, the object ID, and its coordinates. Selecting the plot opens a new tab in the browser that links to the object's SDSS reference page. This allows the user to drill-down for additional properties.

2.4.4.3 Small Multiples

The objects in the current collection may not be located in the same area of the sky, and thus may not be visible or distinguishable in a single gigabit overlay view. To facilitate the analysis of such collections (task T6), we provide an additional small multiples view of the collection of objects represented by the trend image. In order to retrieve a postage-stamp image of each object, the client sends the list of object coordinates to the server. The server then connects to the the SDSS Science Archive Server and requests cutout field images centered around the coordinates it received. Finally, the images are returned to the client and rendered in the small multiple view.

2.4.4.4 Linked Views

Dynamic queries and view linking further allow the users to refine the collection of objects represented by the trend image (task T5). Through query interaction in a linked panel the client can both filter out incorrect results as well as add new entries to the visual abstraction. Right-clicking on an object line in the trend view enables the user to jump to the object's image in the gigabit overlay panel, and thus make further inferences based on the object's celestial neighborhood.

Overall, the use of the trend pixel-based abstraction allows viewing and analyzing data for large collections of objects, while efficiently using the available screen real estate. Along with the zooming, filtering, details-on-demand, dynamic queries and linked-views interaction techniques described above, the small multiples representation aides in both identification and filtering of the results. Furthermore,

condensing the data from multiple FITS files into PNG images enables us to avoid the transfer of large FITS files to the client. The approach reduces thus the bandwidth usage between the client and server.

2.5 Results

In this section we report on the performance of our approach. We first measure the precomputation of the FIRST images stored on the backend of the pipeline; conversion to raw images, mosaicking, and reprojection. Next we report rendering speeds with varying amounts of image data presented to the user. We then present a case study where domain experts perform an overlay-based analysis with our tool and report their findings. A second case-study examines the benefits of interactive trend images to browsing and grouping tasks. The third and most complex case study follows an integrated workflow through our system. Finally, we report feedback from repeated evaluation with a group of five astronomy researchers, as well as from three astronomy workshops where the tool was demonstrated and made available to astronomers for testing. The workshops correspond to separate interest-based groups of astronomers associated with particular sky surveys; each workshop featured more than 30 participants. The implementation of our approach is in its beta release.

2.5.1 Preprocessing

Offline precomputation of the FIRST images is the most time consuming part of the pipeline; however, this stage only has to be performed once when the data is first acquired for a survey. Each image takes between 30 and 40 seconds to generate, with 20 seconds of the process dedicated to reprojecting the image into the WCS map projection. Depending on the sky coverage of the survey, this preprocessing can take anywhere from a week to a month. In the case of FIRST, it took fifteen days to compute all

of the images needed for tiles using a server running CentOS 6, Dual 6 Core processor at 24GHz, and 32 GB RAM.

2.5.2 Performance

The initial data retrieval and loading stage varies depending on the source the images arrive from. To retrieve FIRST images from our server, a loading time of 50-200ms is incurred for sizes varying between 400-700 KB. Retrieving LSST images from our server incurs a loading time between 200-400ms with sizes varying between 4-5 MB. Finally, SDSS loading times are slightly higher, typically incurring 750-1250ms with sizes varying between 60-70 KB. These speeds can vary greatly depending on the bandwidth and load of the SDSS servers at the time of use.

Trend image generation takes between 25 seconds (for a 50 object collection) to 130 seconds (for a 200 object collection), on a Macbook Pro, 2.3GHz Quad Core i7 with 8G of RAM. Returning the spectrum associated with an object takes between 180ms and 220ms, dependent on the speed of the network.

Once the images are fetched, the rendering speed hovers at 45 frames per second on a Windows 7 Machine, Quad Core i5, 16 GB RAM. This allows interactive panning and zooming to regions of interest. Our web-based implementation has been tested on multiple browsers such as Safari, Chrome and Firefox.

2.5.3 Case Study: UGC 08782 - A Dusty Elliptical

Figure 8 shows how the cross-correlation and interactive visual navigation of SDSS and FIRST can lead to immediate gains in astronomy when used in tandem. Thus, we based the following case study

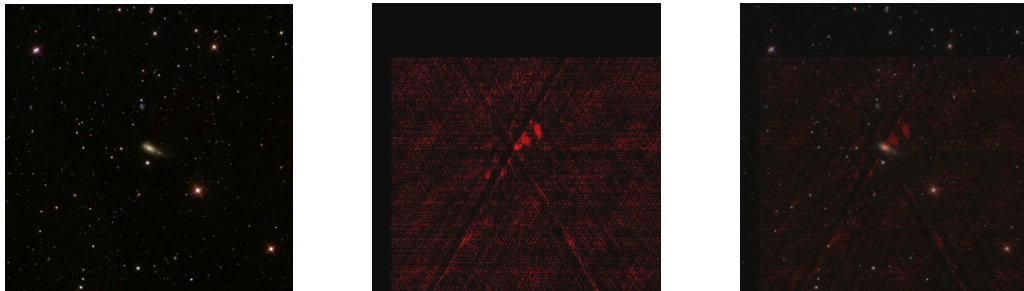


Figure 8. Images of UGC 08782 from two surveys. The left image shows an optical image of the galaxy from the SDSS while the center image shows a radio image of the same galaxy from FIRST. When overlaid in the right image, the connection between the two as radio emission emanating as jets from the central black hole of the galaxy becomes immediately clear.

on a previous discovery that was the direct result of the integration of these two imaging surveys. Two of our senior astronomy research collaborators provided us with this example.

Figure 8 (left) shows an optical image from UGC 08782, a bright elliptical galaxy at a redshift of 0.045. The morphology of this galaxy was originally ambiguous between a spiral and a dusty elliptical, exhibiting dust lanes and disturbed morphological features. Dusty ellipticals are often seen to show signatures of an active galactic nucleus (AGN) [63; 64]. Some of these AGN exhibit jets, which tend to be perpendicular to the dust lanes. One way to test if UGC 08782 fits these trends is by checking its SDSS spectrum, viewing the optical image, and searching for radio counterparts [65]. Figure 8 (center) shows radio observations from the FIRST survey of the same region, which detected several interesting features. The image in the radio looks quite different. There is a single bright point where the optical galaxy ought to be and two bright patches extending to the upper right. Due to the differing resolutions and sensitivities of the surveys, it is unclear looking at the individual images whether the

FIRST emission is from a unique object or associated with UGC 08782. Without our system, associating the FIRST emission with an optical counterpart would require manually searching optical catalogs for nearby objects and match on position, ranking by closest proximity; and then carefully overlaying the two locations using photo-editing software. The astronomers estimate the process would require 30 minutes to one hour, end to end.

In contrast, when the images are viewed together (Figure 8, right), in under one minute, using our online overlays, the association between these two sources from different surveys is immediate. The bright radio point source lines up on the center of the optical galaxy, as it would if it were the nucleus of the galaxy. The two patches of radio emission in the upper right appear to emanate from the central point source, as a radio jet might. Not only does the overlay allow for a more efficient cross-matching, it also provides a nice framework for understanding the physical processes observed in each survey and how those processes are connected to one another.

2.5.4 Case Study: Trends in Type Ia Supernovae

The second case-study, completed by two different senior astronomy researchers, showcases the benefits of trend images in the analysis of large collections of Type Ia (“one-A”) Supernovae. Supernovae are stars that are undergoing catastrophic explosions, which can be classified according to their spectra. In particular, the spectrum of a Type Ia Supernova is characterized by a lack of hydrogen lines, a strong absorption line at 6550 angstroms near maximum, and late-time spectrum iron-group emission lines. Identifying these spectrum-based features via direct catalog querying is, however, a laborious and intensive process. The process is further prone to errors such as inclusion of problematic data, or

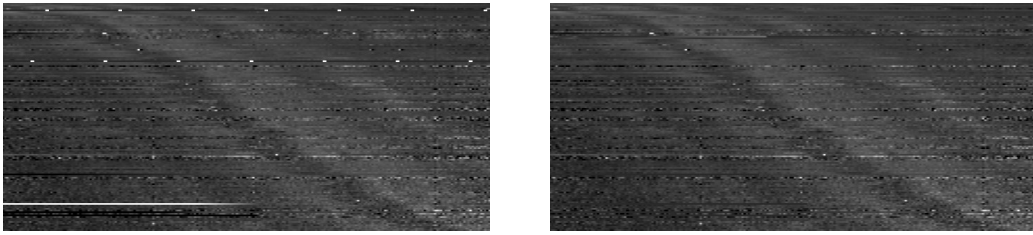


Figure 9. Two trend images generated for the Type Ia Supernovae case study. This interactive visualization allows users to see the spectrum of interest in order to identify unusual outliers. In the first image, 200 Type Ia Supernovae objects are ordered in increasing redshift top ($z=0.15$) to bottom ($z=1.0$) and increasing wavelength left to right. The general broad features of the Type Ia can be seen in the dark and light bands that represent characteristic features of explosions. These bands smoothly trace-back over redshift, indicating that these objects form a consistent class. The first trend image (left) shows clear evidence of potential outliers: objects with unusual spectra that do not match the group. Upon further inspection of these objects, the astronomer removed the confirmed outliers and generated the reduced dataset shown in right.

exclusion of a good object. To alleviate these shortcomings, the researchers were looking for ways to group, analyze trends, and regroup potential collections of Type Ia Supernovae.

The trend image mechanism enabled the researchers to quickly query the SDSS DR7 survey, then analyze and group 200 potential Type Ia Supernova objects (Figure 9). In these trend images, the spectrum of objects is plotted as intensity along the X axis while the objects themselves are sorted along Y according to their redshift. Redshift is the factor by which the wavelengths of light have been stretched as it travels by the expansion of the Universe and provides a sorting in terms of how far back in time we are looking for each object, or equivalently a sorting in distance. Sorting along the redshift property let the researchers immediately see how spectral features (such as emission lines from particular elements) move in observed wavelength with redshift, and immediately observe whether the strength of those

features is changing over time. Objects with incorrect redshifts or with unusual spectra immediately stand out because of the great mismatch with their neighbors.

Producing this first trend image from the raw observations (Figure 9, left) brought immediate attention to two distinct outlier classes: 1) the periodic bright spots in the otherwise grey lines (one line a few rows down from the top and another $1/3$ down from the top), as well as three half-length continuous lines near the bottom (one white, two black); and 2) the individual bright pixels in spectra scattered among the data. These classes are visibly noticeable as outliers to the rest of the spectra in the image that – according to the researchers – would have taken many man-hours to discover by combing through each spectrum in the dataset individually. Upon further examination of the individual spectra, the first class turned out to be a data reduction issue: unreasonable values in the released telescope data. The second class turned out to be an astrophysically interesting issue: emission lines from the region of the host galaxy underlying the supernovae.

After investigation and correction of the first issue, the researchers arrived at the new visualization in Figure 9 (right). This second trend image has the powerful property that features which are vertical lines (constant observed wavelength) represent information about detector problems or atmospheric lines, whereas the properties of the supernovae themselves follow the curves of constant rest wavelength seen in the dark and light bands. The smooth trends seen in this visual representation of supernova spectra confirm that they exhibit similar features with one another and can be grouped together as one class of objects. The researchers took about 10 minutes with our system to complete this case study, with most of the time spent examining the outlier spectra. They estimate that, without our system, the study would require on the order of weeks of work.

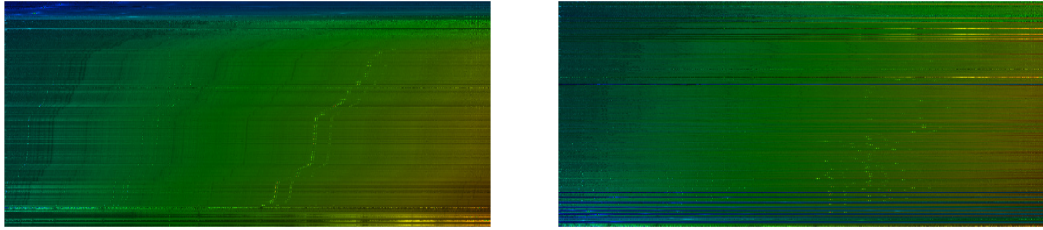


Figure 10. Trend image for 200 galaxies, stars, and quasars in the direction towards the Galactic North Pole, from the Sloan Digital Sky Survey. In the left image, each row is an object's spectrum, plotted along x in observed wavelength, and sorted along y in redshift, with smaller values at the bottom. Galaxies occupy the majority of the parameter space, in the central part of the image; stars are grouped in the noisier, bottom part; while quasars (top band) appear entirely blue. In the right image, the same object rows have been sorted by g-r color, with smaller (bluer) values at the bottom and larger (redder) values at the top. In this image, star-forming galaxies with strong emission lines features are blue in color, reflecting their young age. In contrast, elliptical galaxies tend to be redder, reflecting their older ages. At the very top, the researcher noticed a fair number of extremely red stars in the Milky Way North Pole.

2.5.5 Case Study: Spectroscopic Analysis of Galaxies

The third case study follows an integrated workflow through our system. In this study, a senior astronomer is interested in the spectroscopic analysis of galaxies, and in particular, in understanding how object colors, and absorption and emission features in spectra from different classes relate to each other. The astronomer began his analysis by examining first the properties of galaxy spectra in contrast to those of stars and quasars. To this end, he used our system to do a broad query over the Sloan Digital Sky Survey for 200 astronomical objects within 1 degree of RA, Dec: 191.0, 26.0, a region pointing towards the Galactic North Pole. Specific search parameters included the RA, Dec, redshift, and g-r color (a star with a high g-r color is redder than a star with a low g-r color.)

Continuing to use the system, the researcher then generated the trend image corresponding to the spectra in this resulting dataset (Figure 10, left). To explore the trends in the dataset, he sorted the trend image first by redshift. Sorting by redshift shows the stars, galaxies, and quasars existing in three distinct locations in this parameter space. Galaxies occupy the majority of the parameter space, in the central part of the image, with many clear absorption and emission features visible as dark, and bright lines, respectively, trending through the visual representation. For example, at the right of the image, the researcher could clearly see the strong H-alpha and SII emission lines in star-forming galaxies trending upwards and to the right. These features were also apparent in the one-dimensional spectra detail on demand. The sudden disappearance of the feature at higher redshift revealed a fair number of elliptical galaxies lacking these features. He noticed similar trends in the dark-banded absorption lines.

Stars and quasars occupy less of the parameter space in this particular region of the sky but were still instantly visible to the researcher in the trend image. The stars occupy the “noisier”, lower part of the image and make their presence known by “breaking up” the nice trends seen in the galaxy spectra. Zooming in on the stellar spectra revealed to the researcher the presence of vertical absorption lines bands, that exhibit no trend with redshift. The researcher recalled that the stars had a redshift close to 0, due to the stars location within the Milky Way Galaxy. The quasars are visible in the upper part of the image. As quasars are galaxies that predominantly exist at high redshifts, they looked entirely blue.

Sorting on g-r color (Figure 10, right) further revealed that the star-forming galaxies exhibiting strong emission lines features are also blue in color, reflecting their young age. In contrast, elliptical galaxies tended to be redder, reflecting their older ages. At the very top, the researcher noticed a fair number of extremely red stars in the Milky Way North Pole. He concluded these stars are thus likely

a part of the Milky Way halo, a spheroid surrounding the disk containing clusters of old, red stars that have existed since the earliest formations of the Milky Way.

Having explored the trends in the various objects found towards the Galactic North Pole, the astronomer decided to inspect the objects themselves along with their local environment by bringing up the thumbnail images of the dataset. He found the galaxies in this direction to be an average mix of galaxies exhibiting a wide variety of properties.

Curious about the positional relationship between the North Pole stars, the galaxies, and quasars, the astronomer then decided to display an overlay of all the objects in the dataset, color-coded by redshift, with marker differentiating between the three classes of objects. He found that all three classes appeared evenly distributed within the region specified in his query.

Finally, examining the full sky image of this region (see supplemental video in Appendix A), the astronomer visually noticed a large numbers of red stars in the halo of our Galaxy, supporting his deductions from the trend image. Wondering if any of the galaxies or quasars in the dataset are producing radio emissions, he overlaid the FIRST radio survey and explored the overlapping regions. Following these observations, with no noticeable overlapping features between the radio and optical images, the astronomer decided to explore, in a future study, the question of whether or not stars and galaxies in the directions toward the Galactic center and anti-center exhibit similar colors as seen in the direction of the Galactic North Pole.

The astronomer estimates that completing the present study in the absence of our system would have required weeks of collecting, examining, and grouping the data. Typically, researchers would have approached the problem by first identifying the type of data they are interested in through papers, cat-

alogs, or a perhaps a specialized interface. These data are specific to the question they are attempting to answer, and are seldom restricted to one particular survey. Our researcher and his group would then manually download locally all the data and/or catalogs satisfying their initial criteria. He would then display the imaging or spectroscopic data in a local software package, manually inspecting individual objects in this dataset one at a time, all the while marking outliers, objects of interest, or interesting patterns. Classification of spectroscopic objects is often done by the identification of specific emission- or absorption- line features visible in the spectrum. This approach is rather time-consuming, involving the initial data location and acquisition, combined with manual inspection of spectra for spectral-line identification. Manual inspection, on an individual basis, of an object’s spectral class could take anywhere from 1-30 seconds, depending on the astronomer’s goals. Cross-survey correlations, and marking interesting or outlying objects, for further analysis may often take much longer. Performing this set of tasks for hundreds of objects at a time was an unreasonable time commitment for the astronomer.

In contrast, collecting, viewing and visually analyzing the results in our system highlighted practically instantly many aspects of the spectroscopic dataset that would have been difficult to glean by examining the spectra one at a time. Completing this case study (highlighted in the supplemental video in Appendix A) with our system has only required 5 to 10 minutes. The astronomer has adopted our system as a research tool.

2.5.6 Domain-Expert Feedback

Feedback from repeat evaluation meetings showed enthusiasm for the tool. The domain experts considered the approach “an exciting and effective tool for visualizing all-sky surveys. Many of the tools required have been implemented effectively.” The ability to compare images of the sky taken at

different wavelengths simultaneously and to visually query catalogs was particularly appreciated, while the interactive navigation was considered on par with the much appreciated Google Sky interface. The interactive trend images attracted enthusiastic feedback. Astronomers stated that the interactive trend images allow them to more easily and quickly identify patterns and outliers in the data. The researchers are eager to use the tool in their research and in classrooms.

The workshop expert-users particularly appreciated the ability to combine separate sources of information without having to resort to cumbersome, external tools for image processing. As shown in the example in Figure 11, overlaying catalog search results visually further enables queries of the *what – where – correlated-with-what* type. In this example, more than 800 points resulting from searches over the Sloan Digital Sky Survey catalog are visualized efficiently using pixel-based overlays: two query results based on two different attributes are overlaid (red for redshift, blue for the focal ratio of the telescope; brighter intensities correspond to greater values), revealing vertical spatial patterns in conjunction to attribute overlaps. Figure 2 further shows three cross-correlated overlays (partial coverage shown in the figure solely for static illustration purposes) of optical observations, radio-emission observations, and simulation results from the SDSS sky survey, the FIRST sky survey, and the LSST dataset. Transparency can be interactively controlled for each overlay, enabling cross-spectrum analysis. The workshop researchers are interested in applying this prototype to specific problems such as browsing large sets of objects and galaxy identification. In toy demonstrations, the interactive trend images have already been used to *browse–group–analyze* several collections of objects, from galaxies to quasars; to great feedback and requests for immediate release to the astronomy community. Several astronomy research groups have expressed keen interest in integrating their data with our tool.

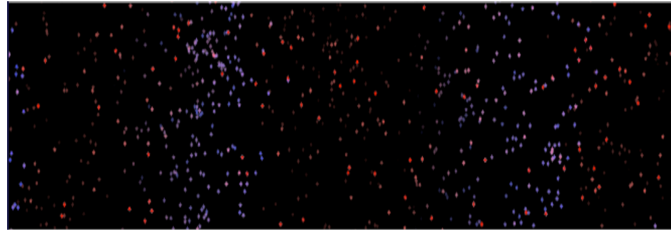


Figure 11. 831 points resulting from searches over the Sloan Digital Sky Survey catalog database are visualized efficiently using pixel-based overlays. Two query results based on two different attributes are overlaid (red for redshift, blue for the focal ratio of the telescope; brighter intensities correspond to greater values), revealing spatial patterns in conjunction to attribute overlaps.

2.6 Discussion and Conclusion

Our approach enables the visual cross-correlation of sky surveys taken at different wavelengths, as well as the visual querying of catalogs. Furthermore, the combination of prefix-matching indexing, a client-server backbone, and of pixel-based overlays makes possible the interactive exploration of large scale, complementary astronomy observations.

New surveys can be flexibly added to the system, provided they specify the raw image data and the projection information of the telescope in standard FITS files. For surveys which benefit from a programmatic interface, our system would implement a simple script to access the data from the online interface. If a programmatic interface does not exist, the images would first need to be downloaded, organized in indices, and stored on local servers.

Our results show that pixel-based overlays and geohashing have the potential to generate scalable, interactive, graphical representations of astronomy data. This approach may allow us to overcome bandwidth and screen-space current limitations in astronomy visualization. The advantages of this

approach are its versatility, flexible control on the client side, and visual scalability (to the pixel level), enabling the visual analysis of large datasets. Accessing graphics hardware through WebGL further provides the users with a rich, graphics-accelerated web experience.

The trend image visual abstractions naturally highlight the trends within objects of a given class. They also support the rapid identification of outlier objects in an object collection — be they outlier objects characterized by poor data/identifications, or outlier objects which have unusual physical properties. Interactive trend images similar to those depicted in Figure 9 may be constructed with almost any property of the object of interest — such as distance, color, or time since a transient event began — for the Y sorting. These interactive representations provide a rapid way to look for correlations between properties of objects, but also take advantage of the human eye’s ability to recognize patterns and detect outliers.

Finally, evaluation on three case studies, as well as overwhelmingly positive feedback from astronomers emphasize the benefits of this visual approach to the observational astronomy field. In terms of limitations, relying on streaming the data from remote sources is a concern as certain surveys do not provide programmatic access to their images.

In conclusion, we have introduced a novel approach to assist the interactive exploration and analysis of large-scale observational astronomy datasets. Our approach successfully integrates large-scale, distributed, multi-layer geospatial data while attaining interactive visual mining, panning and zooming framerates. From a technical perspective, we contribute a novel computing infrastructure to cross-register, cache, index, and present large-scale geospatial data at interactive rates. Large local image datasets are partitioned into a spatial index structure that allows prefix-matching of spatial objects and

regions. In conjunction with pixel-based overlays and trend images, this web-based approach allows fetching, displaying, panning and zooming of gigabit panoramas of the sky in real time. In our implementation, images from three surveys (SDSS, FIRST, and LSST), and catalog search results were visually cross-registered and integrated as overlays, allowing cross-spectrum analysis of astronomy observations.

From the application end, we contribute an analysis and model of the observational astronomy domain, as well as three case studies and an evaluation from domain experts. Astronomer feedback and testing indicates that our approach matches the interactivity of state-of-the-art, corporate educational tools, while having the power and flexibility needed to serve the observational astronomy research community. Being able to quickly aggregate and overlay data from multiple surveys brings immediate clarity to inherently complex phenomena, reducing time spent managing the data while allocating more time for science.

CHAPTER 3

INTERACTIVE EXPLORATION OF SEQUENCE AND STRUCTURAL DATA TO IDENTIFY FUNCTIONAL MUTATIONS IN PROTEIN FAMILIES

This chapter was originally published in the BMC Proceedings in 2014 [2]. This version has been edited to be consistent with the rest of the dissertation. Coauthors on this work include John Wenskovich (JW), Koonwah Chen (KC), David Koes (DK), Timothy Travers (TT), and G. Elisabeta Marai (GEM). The contributions from each author included: JW implemented the sorting algorithms for the trend image and several of the data parsers; KC contributed the design and implementation of the sequence and residue views; DK and TT served as our structural biology domain experts and provided support with the software testing and the design of the case studies; GEM conceived this project, and directed the top-level design, implementation and testing of the tool. My (TL) contributions to this work included the design and implementation of the client-server architecture, the database back-end, as well as the 3D view and trend image of the visual interface. I am the first author on this work.

This chapter continues our investigation into spatial data design strategies with the problem domain of computational biology. In particular, this chapter examines the challenges associated with discovering correlations between protein structure and functionality (i.e., why specific mutations to the protein structure cause dysfunction). In our collaboration with biologists (collaborators DK, TT), we found that many of these challenges coincided with those in observational astronomy domain (Chapter 2); the workflows that we uncovered sought to identify correlations within protein families that consisted of heterogeneous data (i.e., spatial protein structures with non-spatial sequence information). However, we

observed that unlike the spatial data in the presented observational astronomy problem, these workflows sought to preserve the structural information of the protein for determining if a mutation was likely to affect its function. Additionally, we found that while biologists performed similar workflows to identify protein mutations, the execution of these tasks within their workflows most closely aligned with their expertise in either molecular biology and bioinformatics. Therefore, our solution required a hybrid-design approach, incorporating both a 3D structure representation of a target protein and a novel visual abstraction of its closest family members. In this chapter, we present these visual representations as components of a novel visualization tool, FixingTIM, that we designed to help identify protein mutations and discover their effect on protein function. To do so, we designed the layout of FixingTIM to de-emphasize any single visual representation, enabling various starting points and facilitate the different workflow processes of the biologists.

3.1 Background

By determining the 3D structure and functionality of proteins, biologists can gain insight into the associated cellular processes, speed up the creation of pharmaceutical products, and develop drugs that are more effective in combating disease. A variety of protein-sequencing techniques are currently available; these techniques enable biologists to examine amino acid sequences. As amino acid sequence ultimately determines protein 3D structure, mining of sequence information may facilitate the discovery of correlations between protein structure and functionality. However, the vast number of proteins sequenced by scientists make interactive mining tools necessary in solving this problem.

To improve the exploration process, many efforts have been made, from folding the sequences through classification [66; 67], to tools for 3D view exploration [68] and to web-based applications

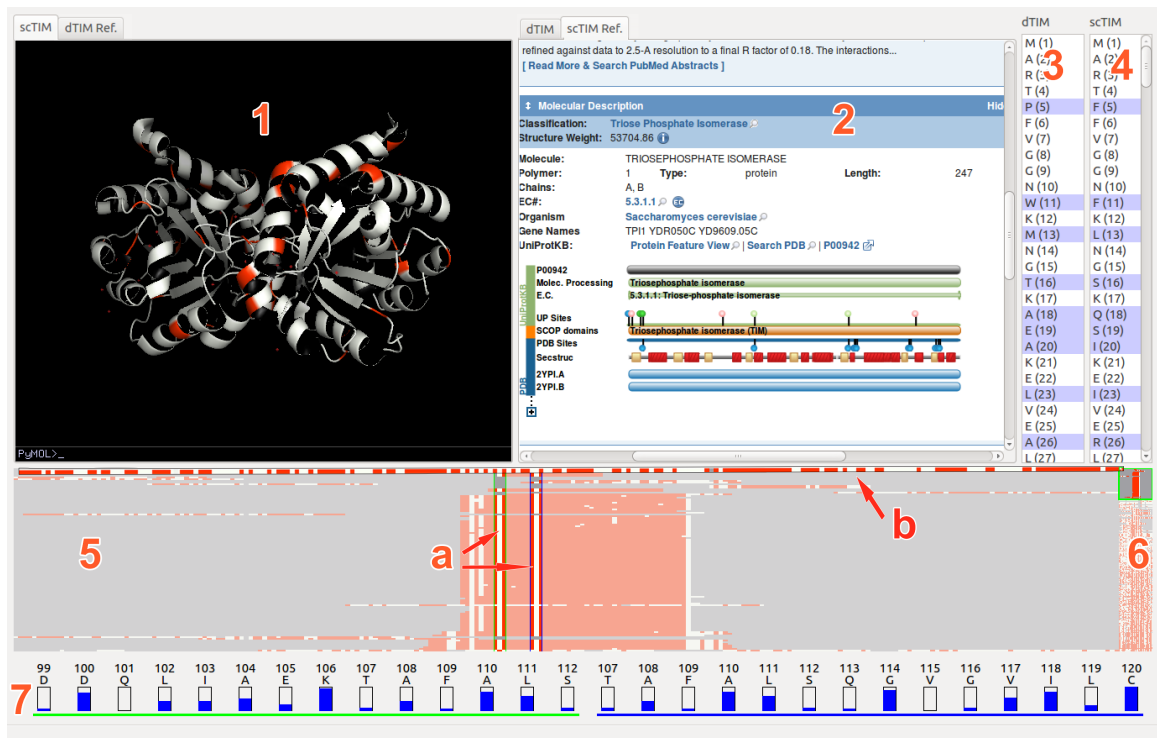


Figure 12. FixingTIM visual interface with four panels: a 3D view and reference information panel (1 and 2); a protein sequence viewer (3 and 4); a trend image panel for aggregating protein families (5 and 6) with two fragment paddles (5a) and a sequence paddle (5b); and a residue view for residue distribution information (7).

which present large amounts of information to the users [69]. Nevertheless, challenges in solving this mining problem remain, from addressing scalability to spatial and non-spatial data integration and to tool integration.

In all but trivial cases, the task of predicting functionality remains an open challenge. In particular, the question “is this mutation likely to affect the function of this protein?” remains elusive [70]. Despite efforts to improve protein analysis and bridge the “protein structure gap,” many approaches

continue to rely on multiple sequence alignments (MSA) to determine whether mutations affect function [71]. Overall, the reliance on this approach has largely stalled functionality prediction advancement for decades [70].

Similar to observational astronomy, the increased volume of the sequenced proteins hinders attempts to create a complete data overview without filtering operations and lends itself nicely to the Search-first mantra [7]. However, the omission of spatial information from the comparative analysis is chiefly to blame for this perceived failure of the MSA approach. Unlike astronomical data, the spatial structure and non-spatial sequence data aspects that describe a protein are not inherently linked [72] despite the spatial attributes being vital to the task of predicting functionality.

In this chapter, we introduce a novel visualization tool, FixingTIM (Figure 12), to help identify protein mutations across families of structural models, and to help discover the effect of these mutations on protein function. Following a rigorous data and task analysis, we pursue a client-server approach in which distributed data sources for 3D structure and non-spatial sequence information are integrated. To better address scalability concerns, we aggregate family-sequence data into a novel interactive pixel-based abstraction called a trend image. Interactive exploration, multiple linked views, and details on demand further allow the generation of hypotheses regarding structure and functionality correlations in a diverse and fragmented space. The tool is open source and publicly available at <http://visualizlab.org/fixingTIM>.

3.2 Methods

3.2.1 Data and Task Analysis

The design of our tool is informed by our domain data and task analysis. The data we consider in this application consists primarily of protein characteristics, where protein characteristics include structural information and amino acid sequence information. This data can be used to build a 3D representation of the protein that allows visualization of the atoms that comprise each residue in the protein as well as the bonds between these atoms.

The *protein structure* — determined theoretically or experimentally — is typically stored using a PDB file [73], alongside references to the studies that determined the structure of the proteins, the residue sequence (the sequence of amino acids that make up the protein), and the positions of each atom in 3D space. The structural data can be visually mapped to a 3D representation of the protein, which includes atoms, bonds, amino acids and protein chains. The *amino acid sequence* of each protein is typically stored in remote databases, for example, Uniprot [74]. Each sequence consists of a string of capital letters, each letter representing an amino acid (also called a residue) in the protein. To find regions of conservation within sequences belonging to the same protein family, these sequences can be aligned using a host of computational alignment tools, with gaps introduced to better align common sub-sequences present across the family. A particular sequence family may include special mutations, some functionally-defective. Finally, external web services [74] may provide additional relevant *metadata and data*, such as model-quality ratings provided by domain experts.

From the desirable features of a visual mining system (as indicated by the BioVis 2013 Data Contest), we focus on the tasks outlined in Table III.

TABLE III. Function Protein Mutation Task Analysis

Task	Visual/Interaction Mapping	Technical Challenge
T1 Generate 3D protein structures from sequence data	NA	Computational time
T2 Inspect 3D protein structures	Geometric Model & Attribute Overlays	Interaction design
T3 Link to online resource	Linked Views	Interaction design
T4 Compare a single protein to the rest of its family	Interactive Trend Images	Visual abstraction
T5 Identify sequence mutation locations on a family of proteins	Interactive Trend Images	Visual abstraction
T6 Examine multiple sequence alignments	Linked Views & Details-on-Demand	Visual design
T7 Highlight specific residue locations on the 3D protein structures	Linked Views & Details-on-Demand	Visual design
T8 Examine residue distribution across a protein family	Details-on-Demand	Visual abstraction

3.2.2 Client-Server Framework

Given the variety and distributed nature of relevant domain data, we design and implement an overall client-server architecture (Figure 13). Our server fetches and caches in a local MySQL database the protein sequences, alignment, and 3D structures from ModBase, Uniprot, and the National Institute of Health (NIH) BLAST server [75]. The server provides sequence and 3D structure data to the client. If a 3D structure does not exist for the protein, the server computes an approximate model using the Sali Lab Modeller toolkit [76].

Our client implements three core modules: a trend image module for exploring protein families; an interaction manager for viewing; and an external reference module for access to online catalogs. The three modules are linked, allowing for simultaneous interaction with the data in each abstraction.

The back-end of the tool is implemented in Python, C, and MySQL. The front-end of the tool uses Python as the primary development language, with Qt for the GUI and the PyMOL Molecular Graphics System [77] for rendering the protein structures.

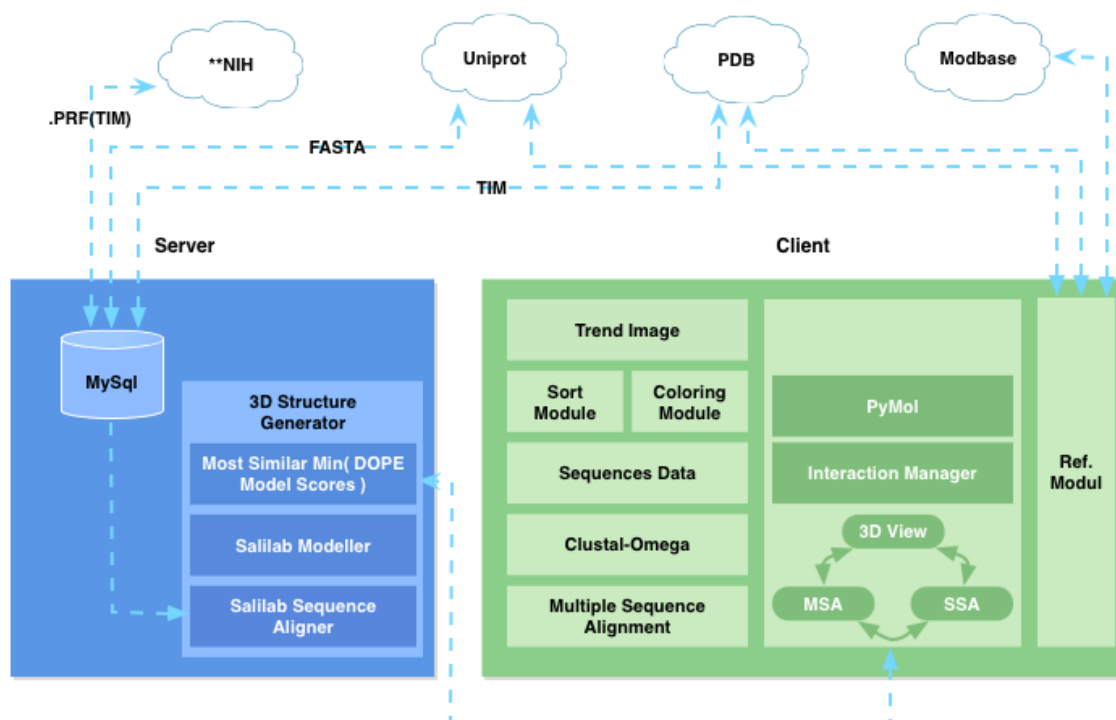



Figure 13. Client-server architecture diagram, with TIM-instantiation as an example application

3.2.3 Visual Design

Given the diversity and complementary nature of the data required, as well as the comparison nature of the domain-specific interactions, we pursue a linked multi-view top design. The visual interface consists of four linked panels (Figure 12): a tabbed 3D structure and reference viewing panel (two side-by-side views); a side-by-side protein sequence viewer; a trend image panel for exploring protein families; and a residue view for residue distribution information.



Frequency	Others	Most frequent fragment				
General chemical characteristics	Basic	Aliphatic	Hydroxyl or Sulfur-containing	Aromatic	Cyclic	Acidic and their Amides
Side-chain polarity	Basic polar	Nonpolar		Polar	Acidic polar	
Side-chain charge	Neutral	Negative		Positive		
Side-chain contact with polar solvent		Charged		Polar	Hydrophobic	
Common fragments with the comparison sequence		Common fragment				

Figure 14. Trend image coloring scheme based on residue properties and their occurrence frequency throughout the family. Each set of categorical residue properties (rows) are mapped to one of five colors (columns).

In this top design, the trend image view serves as the main anchor point of the interface. From this view, users can explore an entire protein family, and view the differences between family members. By right-clicking on a trend line, the user has the option of opening the structure file for this model in the 3D View, to compare it side-by-side with another model. Below, we describe each module in detail.

3.2.3.1 Trend Image Panel

The trend image view provides the ability to navigate and sort through large numbers of sequences. The trend-image is a pixel-based visual abstraction, in which each line represents the residues of a single protein sequence. The trend image summarizes an entire protein family, aligned by one of several sorting algorithms, and colored by one of several different color schemes (e.g., Figure 14). The trend image view contains paddles (fixed-width brushes) for the selection of subsequences from a full family of protein sequences. These paddles also link to the residue distribution view at the bottom of the tool;

this panel displays information about the distribution of amino acids, namely the fraction of proteins in the family that share the same residue as the selected protein.

A vertical overview pane (component 6 in Figure 12) provides a high-level view of the full dataset; while the two *fragment*-selection paddles allows narrowing the section of sequence considered for analysis and drilling for details. The sequence-selection paddle allows users to select a particular sequence. A selection event prompts the application to search for the 3D structure from online repositories; the 3D structure is presented if it already exists or it is generated on the fly if it does not.

We note that an earlier version of the software included a single fragment-selection paddle. However, upon repairing the defective protein (dTIM), the BioVis domain experts discovered two symmetric sequence pairings with identical mutations; once repaired, the function of the entire protein was restored. Identifying these unique mutations required the ability to examine two parts of the sequence in detail, simultaneously. Based on this information, we implemented two vertical selection paddles in the trend image. These paddles allows for the examination of two locations in the sequence for residue conservation simultaneously, which alleviates the burden of individual inspection to identify symmetric pairings of mutations.

To facilitate navigation of the trend image, we provide a set of sorting algorithms. The sorting algorithms calculate a weight for each sequence relative to one input member of a protein family, and then order the sequences by their respective weights. We provide sorting by using the following measures as weights: fragment frequency; edit distance; weighted edit distance; number or percentage of common residues; number or percentage of common residues without regard to sequence position; number of residue subsequences of length N in common; and edit distance on selected residues.

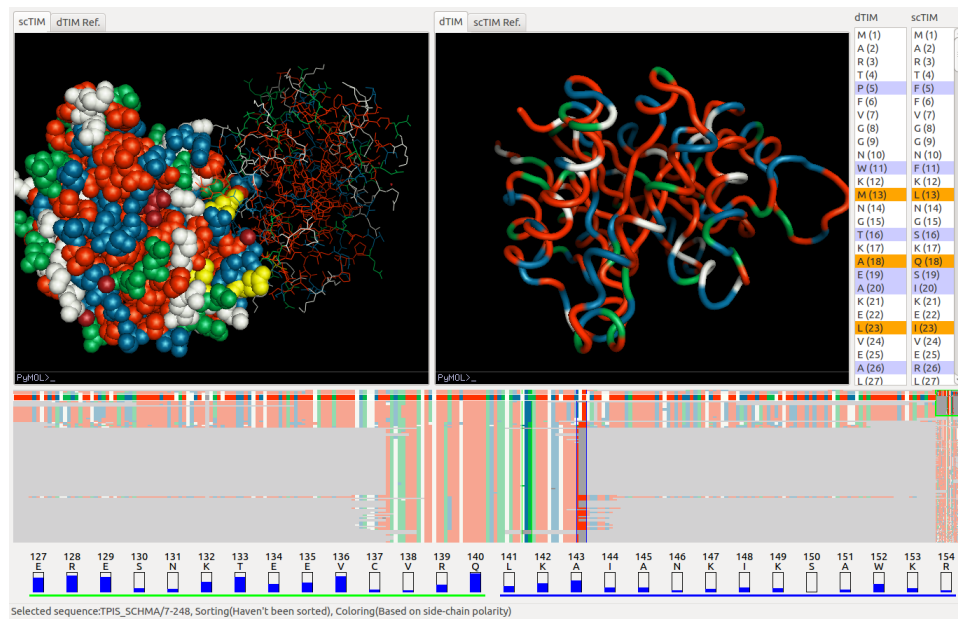


Figure 15. Visual Comparison of the defective protein (dTlM) and its closest functional family member (scTIM). The two volume views show the protein backbone of dTIM (left) and the CPK sphere representation of scTIM (right), respectively. The trend image is sorted by the number of common residues. A side-chain polarity coloring is applied, and the two vertical selection paddles are located around position 142. The two residues shown at the bottom (green underline, respectively blue underline) correspond to the two vertical paddle selections.

In addition to the sorting algorithms, we also provide a ColorBrewer [78] set of coloring schemes to highlight a subset of the residues of each sequence. In each color scheme, black is used for residues that are not included in the scheme, whereas white is used to represent spacing in the sequence alignment. The residues in each internal class are given the same color. The list of coloring schemes is as follows: fragment frequency; general chemical characteristics; side-chain polarity; side-chain charge; and side-chain contact with polar solvent.

3.2.3.2 Residue Viewer

The residue distribution view displays the fragment ID, fragment name, and the percentage of each residue type found in the same column (corresponding fragment in each sequence) for all sequences.

3.2.3.3 3D Viewer

The top left panel of the tool provides two 3D structural views—one for the target protein, and one for the source protein, as well as a tabbed reference tool providing information about each protein. The structures can be examined at both the amino acid sequence level and at the atomic level. Through panning, zooming, rotation and details-on-demand operations (synchronized between the two views), users can observe different aspects of the two 3D structures.

Alternatively, the tabbed reference viewer allows users to access information from three complementary online data repositories: Uniprot, ModBase [79], and the RCSB Protein Data Bank. ModBase, for example, provides links to other databases, as well as ribbon diagrams for various models in the current sequence, and quality-criteria quantifying the reliability of certain model aspects.

3.2.3.4 Protein Sequence Viewer

To link in sequence information, the sequence panel lists the residue sequences for the selected structural models, with the differences between the two sequences highlighted in blue. The residues are selectable, and the selections are reflected in the 3D structure view.

3.3 Results and Discussion

We demonstrate our tool on a TIM protein-family application. These proteins play an important role in efficient energy production and can be found in nearly every organism, including animals, fungi, plants, and bacteria. In this section we report on our experience using the tool for this application. We

follow with a formal evaluation by two structural biologists, expanded from our IEEE BioVis Data Contest Visualization Award-winning submission. We last report the feedback from the contest organizers.

3.3.1 TIM protein-family exploration

The application examines the scTIM protein (saccharomyces cerevisiae triosephosphate isomerase), a member of the TIM family that was mutated towards the family consensus: a number of amino acids in the sequence were replaced by the most common residue found at that location in the TIM family. The resulting amino acid sequence is dTIM. Unfortunately, dTIM is functionally defective - one or more of the modifications made to scTIM caused the protein to lose its metabolic transport properties. Identifying which modifications caused the loss of functionality is an interesting open research problem.

For this application, we obtained the scTIM PDB, the TIM family sequence data and alignment information from the Battelle Center for Mathematical Medicine, through www.biovis.net. We used the tool to fetch 28 additional PDB files from RCSB, and to further generate more than 620 PDB files from the provided sequence data. We used the database backend to link PDB and FASTA IDs for preprocessing, and added data from ModBase and Uniprot.

Using our tool, we start by identifying the differences between the dTIM and scTIM sequences. There are 49 different subsequences of residues, encompassing 104 residues modified, created, or deleted in the creation of dTIM. By selecting some or all of these residues in the protein sequence viewer, we can highlight their locations on both 3D structures (Figure 15). We can pan, zoom, and rotate the structures to more closely examine the distribution of these alterations on the protein structure. We can also adjust the rendering properties of the structure.

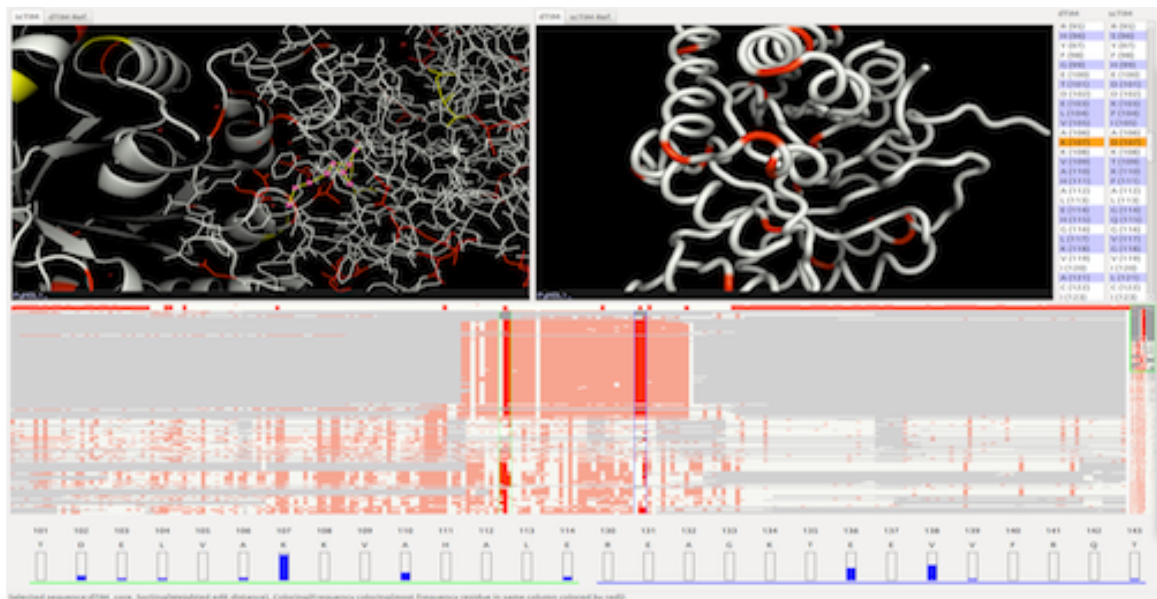


Figure 16. Two residue mutations that may restore function to the protein; residues (K)107(D) and (E)138(E). For both residues, dTIM shares the family consensus but differs from its parent, scTIM.

To determine which models from the TIM family are most similar to the original scTIM, we use the trend-image view in the lower panel. In Figure 15, we can quickly see, for example, that only a few sequences have the same fragment in position 142 with scTIM. A step further, selecting any of the sorting modes from the menu allows comparisons to be made to scTIM. For example, when sorting by common residues, we find that TPIS_HAEDU, the TIM protein homolog found in bacterial species *Haemophilus ducreyi*, shares the greatest number of residues with scTIM. Selecting a particular coloring method displays specific information for each residue.

Manipulating the vertical selection paddle allows us to explore subsequences of the full TIM sequence. Distribution information about residues in the highlighted subsequences are displayed below

the trend image and show the most common amino acid in the TIM family at each sequence index. The bars in the residue viewer that are nearly empty imply that very few members of the TIM family share the same residue as scTIM, making it an ideal candidate for mutation towards the family consensus.

Manipulating the horizontal selection paddle allows us to further explore the individual TIMs in the family, with a fish-eye lens expanding the selected row to more clearly show the residue sequence and coloring. Right-clicking on a selected row allows us to load the structure of that specific TIM into the structure view. If this TIM is unfamiliar to the user, a number of reference databases can be accessed.

In terms of limitations, while the trend image provides a scalable approach to viewing large amounts of sequence data, finding a particular sequence in a protein family remains a challenge. Similarly, attempting to code too much information into the color schemes results in an overload of colors, rendering the trend image unreadable and ineffective. A reduction in the number of colors restores readability to the view, at the cost of removing some information from the trend image.

3.3.2 Structural Biologist Feedback

Two senior structural biology researchers (collaborators DK and TT) have provided feedback and testing throughout the software development process. They are also providing the following example workflow through our system.

In this evaluation session, the researchers sought to explore the mutations in the BioVis Data Contest dataset. Given their structural biology background, the researchers began their analysis by loading and interacting with the 3D structures of the dTIM and scTIM proteins. Their interaction focused on searching for the residues that make up the active site and the protein-protein interface. In their estimation, these two sites were likely candidates for the location of functional mutations.

The researchers selected next the key residues in the 3D Viewer. This action highlights those residues in both the Trend Image Panel and the Protein Sequence Viewer. Using the Protein Sequence Viewer, the researchers identified which of these residues had changed in the conversion from scTIM to dTIM. For each different residue, the researchers returned to the 3D Viewer to inspect the structure of each of the residues and examine their interactions. At this stage the researchers did not use the remaining genomic data, as they were unsure of how to best use this information for the purpose of the contest. Under these circumstances, they identified and proposed two relevant mutations as the most likely candidates: Y101 (to E100) and D81 (to P80).

However, by using the trend image panel, the researchers were able to identify several further matching sequences in the trend image. Although during the evaluation session the trend alignment and residue numbering within a sequence were slightly off due to insertions and deletions in the sequence (later accounted for and corrected in the software), the most senior researcher was able to identify the same set of candidate mutations as captured in the case study below.

In the researchers assessment, it *“would be possible to come up with some reasonable hypotheses without using [this] tool, but it would definitely take more time.* In the default workflow, the researchers believe they would start by building a homology model for dTIM using Modeller and then align this model with the known structure for scTIM within PyMOL, followed by proposing a list of mutations that could be relevant, for example those localized to the active site. However, this approach would not leverage the information of the protein sequence family. To access this type of information without our tool, one could create a multiple sequence alignment using any of a number of online servers, then load the result in an alignment viewer such as JalView [80], and then go back and forth between the structures

in PyMOL and the alignment viewer, in order to refine the previous list of mutations. However, in the researchers opinion, this alternative approach would be tedious. As such, they particularly appreciated our tools integration of the capabilities of existing spatial and sequence viewers, along with other useful functionality built within our software, and the speedup to such workflows provided by our approach.

3.3.3 BioVis Contest Organizer Feedback

Feedback from the BioVis 2013 conference organizers further confirmed the ability of our tool to successfully identify the dysfunctional protein mutations. The experts hypothesized that the most harmful mutations to the protein existed in the active site—the area of the protein which is responsible for its function. Since dTIM was created by combining the mutations of its 640 family proteins, the Trend Image Panel was first observed by the experts in order to gauge the difference of scTIM to the rest of its family. When sorted by the weighted edit distance between scTIM and its protein family members, the trend image exposed five distinct residue locations where dTIM varied from scTIM, but was consistent with the rest of its family. These locations are in our assessment (A)58(G), (K)107(D), (E)138(L), (L)146(V), and (A)22(R) and (L)218(V). From these mutations, the 107, 138 and 146 residues are almost fully conserved throughout the entire family, but differ in scTIM to dTIM. While residue 138 looks promising, since it is very frequent across the entire family, closer inspection shows that a mutation did not occur between dTIM and scTIM. In contrast, mutation 58 is also highly conserved throughout the TIM family, but is also a mutation from scTIM to dTIM. Finally, the last two mutations are both symmetrical in position, at an offset of 22 from either end of the sequence; these residues are also highly conserved in the TIM family, but not between scTIM and dTIM, which indicates they are not responsible for the loss of functionality.

Further examining the 3D structure of the four remaining residues (excluding the two symmetric ones) from the candidate list above, we notice that they lie in or just outside of the active site. Again, this is where most chemical reactions occur, since the active site is the binding site of molecules. This observation brings us full circle to the earlier structural biologist feedback: the structural biology experts initially suspected the most damaging mutations would lie in the active site, and that restoring these mutations could restore functionality.

In terms of the tool features, the trend image and its sorting capabilities—based on our proposed metrics of similarity—were greatly appreciated by the IEEE BioVis domain experts evaluating the tool. A closer examination through the use of the 3D Model Viewer provided evidence that the residues identified above were part of the active site of the protein. This finding matched the initial hypothesis of where the most critical mutations existed, and demonstrates the benefits of combining spatial and non-spatial information in a single tool.

Without the aid of our tool, each sequence would have to be collected and aligned to determine the areas of residue variance. Once found, individual models of dTIM and scTIM have to be examined to locate the area where the mutations were occurring on the 3D structure. By using our tool, the experts were able to quickly identify the residue mutations and link them to the 3D model with a single action. Each expert biologist that tested our tool noted the ease in interaction between the different data representations—sequence alignment, model interaction, etc. These features, coupled the linked data views, proved very efficient for the task of identifying protein mutations. With the insights gained, the locations of where the sequence needs to be repaired were quickly identified.

3.4 Conclusions

The FixingTIM application shows that our tool can assist in the navigation of family of proteins, as well as in the exploration of individual protein structures. The side-by-side 3D views facilitate visual comparison, while the trend image abstraction provides an effective view and exploration of large collections of sequence data. Our tool successfully integrates multiple sources of information and both spatial and non-spatial data. Furthermore, a computational backbone facilitates sorting collections of sequences, as well as generates 3D structures for modified sequences.

In conclusion, we introduced a novel visualization tool that integrates 3D structural information and sequence information for a protein, with additional information from the multiple sequence alignment of the family of proteins with the same function, and with meta information extracted from the family data. In conjunction with domain expert knowledge, this interactive tool can help provide biophysical insight into why specific mutations affect function, and potentially suggest additional modifications to the protein that could be used to rescue functionality.

CHAPTER 4

A SPATIAL NEIGHBORHOOD METHOD FOR COMPUTING LYMPH NODE CARCINOMA SIMILARITY IN PRECISION MEDICINE

This chapter investigates a situation in “precision medicine” where clinicians aim to use big data patient repositories consisting of demographic and clinical characteristics, treatments, and outcomes to tailor therapy decision to the individual patient, based on data from patients who are similar to the one under consideration. Specifically, this chapter details our collaboration with radiation oncologists (CDF, BH, HE) from the MD Anderson Cancer Center (MDACC) to develop a methodology for comparing oropharyngeal carcinomas (OPC) cancer patients based on their spatial patterns of lymph nodal involvement to determine potential treatment strategies regarding both efficacy and toxicity outcomes. Because this spread of disease heavily influences the patients treatment and side-effects, patients with similar nodal involvement often experience similar post-therapy dysphagia side-effects known to arise due to radiation toxicity at various locations of the head and neck.

Unlike the tasks associated with the observed workflows in observational astronomy (Chapter 2) and molecular biology (Chapter 3), clinicians are not strictly interested in exploring the spatial patient data within these large repositories, but rather wish to extract specific spatial information and establish a relationship between corresponding attributes in different patients. Rather than adapting the two established design paradigms to facilitate spatial data exploration, we instead describe a strategy that abstracts the spatial data to work within these known paradigms. By constructing a 2D topological map of the lymph node regions in the human head and neck, we found that we can encode the structural information of the

nodal spread of involvement to build a spatial measure that captures patient similarity within a cohort. In this chapter, we show how the resulting similarity measure can be used with the Search-First paradigm to rank and query for similar patients based on their spatial correlates. We discuss how the strategies behind our design strategy – from the design of our topology-based spatial similarity measurement to the visual and statistical analysis used to validate it – captures groups of patients more susceptible to dysphagia toxicity based on their spatial pattern of nodal involvement.

This chapter has been edited to be consistent with the rest of the dissertation. Coauthors on this work include Baher Elgohari (BH), Hesham Elhalawani (HE), Guadalupe Canahuat (GC), David M. Vock (DV), C. David Fuller (CDF), and G. Elisabeta Marai (GEM). The contributions from each author included: BH, HE and CDF served as the radiation oncology domain experts for this work, providing feedback on the relevance of our spatial neighborhood approach; BH and HE were responsible for the data curation, and cleansing of the patient cohort; GC provided the data mining expertise on the project and contributed the hierarchical clustering implementation and Chi-Squared analysis; DV provided the statistical analysis expertise and suggested the use of the Fisher’s Exact Test of the Chi-Squared Test; GEM provided the visualization expertise and directed the top-level design, implementation, and testing of the approach. My (TL) contributions to this work included the design and implementation of the similarity measure, the compact graph visual representations, and the spatial-measure dendrogram. I am the first author on this work, which is pending review in the Journal of Biomedical Informatics.

4.1 Introduction

The United States National Cancer Institute estimated more than 51,000 people in the United States were diagnosed in 2018 with head and neck squamous cell carcinoma (HNSCC) [81]. Of these HNSCC

cases, more than 90% will result as oropharyngeal carcinomas (OPC), which include cancers of the larynx (voice box), pharynx (throat), lips, tongue, and nose [82; 83]. At the same time, the number of HNSCC cases makes possible the creation of big data repositories consisting of the demographic and clinical characteristics, treatments, and outcomes of patients undergoing radiotherapy. These repositories present opportunities towards informing and further personalizing treatment on a per-patient level, rather than relying on clinician experience or institutional memory alone [84; 85]. Under a healthcare model termed “precision medicine”, clinicians aim to use these patient repositories to tailor therapy decision to the individual patient, based on data from patients who are similar to the one under consideration. Currently, these correlates typically include age, performance status, clinical staging information, and sometimes genetics—attributes that can be statistically aggregated, matched and analyzed.

Yet, similar to most other cancer types, HNSCC treatment and side effects depend in large measure on the spatial location and spread of the cancer. In particular, for more than 50% of OPC patients, the treatment and side-effects are heavily influenced by the spread of disease to lymph nodes (LN) and their corresponding areas (levels), at risk for metastases. OPC generally metastasizes to regional LNs following the lymphatic drainage of the head and neck [86], often resulting in chains of affected LNs along the drainage pathway. These chains correspond to the spread of involvement to specific locations of the head and neck and are thus defined by their spatial attributes. Therefore, for those patients receiving intensity-modulated radiation therapy (IMRT), these chains represent additional targets that must receive radiation treatment. Further complicating matters, the soft tissue structures of the head and neck (organs, muscles, etc.) are highly susceptible to both direct and indirect radiation exposure [81], and the increased toxicity to specific regions has been shown to correlate with post-therapy quality of

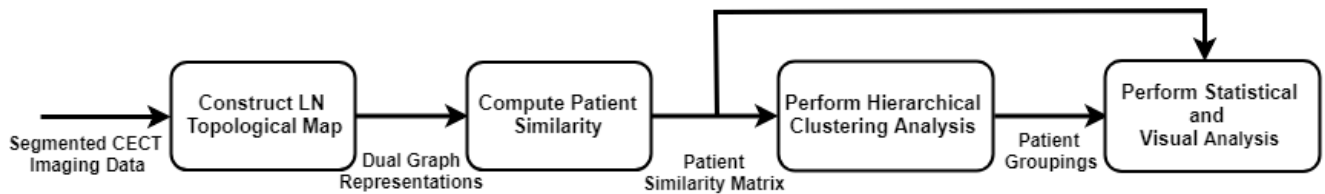


Figure 17. Pipeline detailing the steps and data flow of our presented methodology. After receiving the contrast-enhanced computed tomography (CECT) images from the clinicians, we construct a topological mapping of each patient’s involved nodes and the connections between them. The result matrices are used to compute similarity using a Tanimoto coefficient; hierarchical clustering is performed on the ranked patient scores to determine patient groups; statistical and visual analysis is performed on the groups to determine groups with higher toxicity outcome rates and validate the results.

life. For example, aspiration and dysphagia side-effects affect as many as 30%-50% of patients treated with IMRT [87]. Therefore, we believe that grouping patients by their patterns of nodal involvement spread can help improve treatment strategies regarding both efficacy and toxicity.

Because within a patient cohort there are many rare or unique combinations of spatial involvement chains, as we show in this chapter, the use of direct-match comparison techniques based solely on the oncological labeling of affected LN levels has limited applicability: categorical labeling of the nodes does not lead to meaningful groups of similar spatial (2D or 3D) patterns. Instead, taking into account both the metastasized nodes and the pathways that connect them has the potential to create meaningful, spatially similar groupings.

In this chapter, we present a novel modeling methodology for the comparison of OPC patients within a cohort, based on the patient nodal spread of involvement. We define a topological map, we construct

computational representations, and we introduce a novel graph-based measure to derive patient LN involvement similarity. To validate our approach, we apply our measure to a clinical cohort of 582 post-therapy OPC patients. We perform hierarchical clustering on the output of the similarity ranking to test for correlations with post-therapy toxicity. We contrast these spatial measure results against the results obtained using a categorical labeling of the nodes. Specifically, we hypothesize that the underlying spatial information contained within the chains of affected LN levels would significantly correlate with post-therapy dysphagia side-effects known to arise due to radiation toxicity at various locations of the head and neck. This approach would further allow for binning of patients in cohorts deemed by clinicians as significantly more informative than categorical binning.

4.2 Methods

4.2.1 Overview

Our method is constructed as follows (Figure 17): the LN levels for eligible patients are manually segmented from contrast-enhanced computed tomography imaging data. We then define and construct a LN topological map, based on the level location and its surrounding local neighborhood, and using the medical literature [88] and clinician input; because of left-right symmetry in the human head and neck, this is a 2D map with cells for each node region. To facilitate patient comparison using the spatial information, we next define and construct a dual-graph representation over the topological map; this representation captures the neighbor relationships among the lymph nodes. We use the graph representation to compute the pairwise similarity between each patient using a spatial measure. To compare and contrast the merits of the spatial measure, we also compute the categorical similarity among the patients; this categorical measure ignores spatial relationships in the data. Next, we perform hierarchi-

cal agglomerative clustering on the similarity output and compare the resulting patient groupings. The results are then presented to the clinicians for interpretation of the rankings and clusters of patients. Finally, we perform a statistical analysis to determine if our spatial measure is significantly correlated with post-treatment toxicity outcomes. We describe below in detail each component of this method.

4.2.2 Patient Cohort

Oropharyngeal cancer (OPC) patients who were treated at MD Anderson Cancer Center between 2005 and 2013 were retrospectively reviewed under an approved IRB protocol. Out of the 644 eligible patients who had a pathologically proven OPC, either with a positive biopsy or a surgical excision and received treatment (i.e., radiotherapy +/- chemotherapy) with a curative intent, 582 patients had affected lymph nodes and were included in this study. Affected lymph node (LN) levels were collected from contrast enhanced computed tomography (CECT) diagnostic scans which took place at patients' initial visit for staging and disease assessment. LN levels (retropharyngeal (RP), submental (Ia), submandibular (Ib), upper, medial and lower jugular (II, III, IV respectively) and level V a, b) were defined based on anatomical landmarks and were coded in relation to tumor position. Patients' relevant demographic, clinical, and toxicity data (toxicity of interest were feeding tube and aspiration at six months) were retrieved from electronic medical records.

Table IV shows the post-therapy side-effect counts and patient characteristics across the cohort. Of the 582 patients who underwent intensity-modulated radiotherapy, 163 patients suffered from either post-therapy dysphagia side-effects, with 95 (16.32%) patients reporting aspiration (breathing a foreign material to the airways, such as saliva) and 99 (17.01%) requiring a feeding tube six months after the end of radiotherapy treatment (Feeding Tube at 6 months).

TABLE IV. Patient Characteristics and Post-therapy Side Effects

Characteristics	N (%)
<i>Post-therapy Side Effect</i>	
Feeding tube at 6 mo.	99 (17.01%)
Aspiration	95 (16.32%)
No side effect	388 (66.67%)
<i>Gender</i>	
Male	512 (87.97%)
Female	70 (12.03%)
Median age at Diagnosis	57.8 years (range 20.95 - 88.47)
<i>HPV Status</i>	
Positive	360 (61.86%)
Negative	45 (7.73%)
Unknown	177 (30.41%)
<i>T-category (T)</i>	
Tx	1 (0.17%)
Tis	1 (0.17%)
T1	129 (22.16%)
T2	245 (42.10%)
T3	121 (20.79%)
T4	85 (14.61%)
<i>N-category (N)</i>	
N1	72 (12.37%)
N2	492 (84.54%)
N3	18 (3.09%)
<i>AJCC 7th Edition</i>	
III	68 (11.68%)
IV	514 (88.32%)

4.2.3 Topological Map

Spatial similarity has been facilitated in many domains such as mechanical engineering [89], bioinformatics [90], and oncology [91; 92] by encoding spatial relationships through either topology-based or shape-based techniques. These techniques have the ability to “exhibit common classes of descriptive

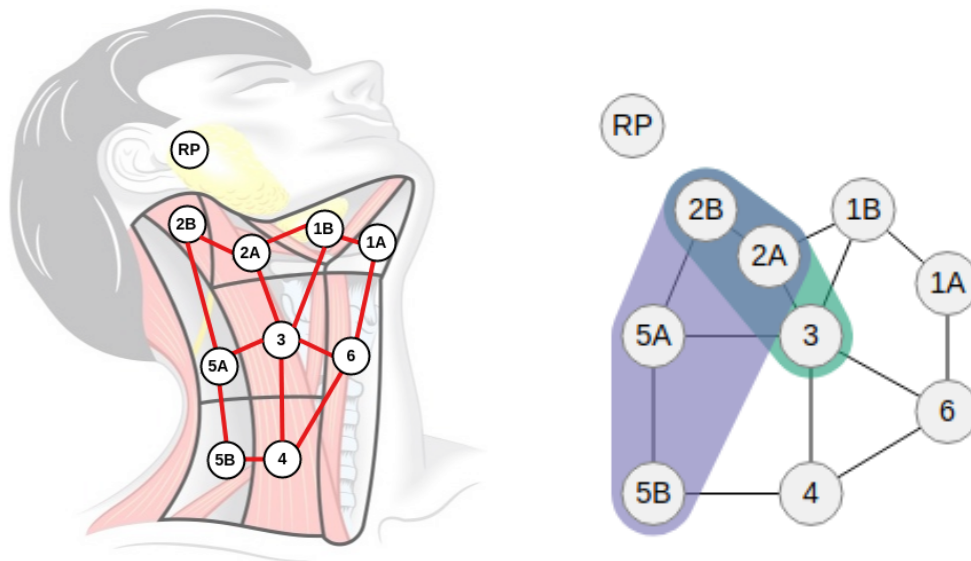


Figure 18. Topological map and graph representation. (Left) A topological map was defined over the lymph node regions (shown in gray), overlaid with a dual graph representation (red) of the map showing the connectivity between the lymph node levels. The Retropharyngeal (RP) lymph nodes are a group of nodes near the base of the skull and are disconnected from the dual graph because their involvement requires specialized treatment. (Right) A compact graph representation was derived from the red graph representation to visually illustrate metastasis over both sides of the head and neck, using symmetry and color to distinguish between left (green), right (purple), and bilateral (blue) involvement.

spatial (topological) features that are quantified by definition of computable measures” [93]. Both topology and shape-based techniques aim to extract spatial attributes, then establish a relationship between corresponding attributes in different patients. However, shape-similarity based methods tend to focus on classifying models of very different shapes, and 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 [91; 94; 95]. In our case, structures are in the same class and do not have easily identifiable features. However, OPC patient analysis presents an opportunity for topology-based techniques.

To this end, we first defined and constructed a 2D topological map over the LN levels, based on the consensus guidelines for the delineation of the head and neck [88], and using the left-right symmetry of the human head and neck and input from our clinician collaborators. Each cell in this topology (shown in gray in Figure 18 (left)) corresponds to an LN level in the human head and neck, based on the spatial location and local neighborhood of each level. Over this topology, we then defined a dual graph representation (shown in red in Figure 18 (left)), where each cell was represented as a node in an undirected graph, and edges were created between each pair of adjacent faces. Using this abstraction, a chain of involvement would follow the links between the adjacent faces; for example, the path connecting LN levels 2B-2A-3 corresponds to a lymph chain of involvement. We decided to place the Retropharyngeal (RP) LN, a LN group near the base of the skull, as a disconnected node in the graph (upper left) because metastasis to this group bears a poor prognosis to OPC patients and requires specialized treatment.

Finally, to account for both sides of the head and neck, the graph was encoded as an adjacency matrix where the upper and lower triangles correspond to the left and right side, respectively; Figure 18 (right) illustrates metastasis over both sides of the head and neck. We initialized the matrix so that each row and column corresponded to one of the LN levels in the graph and assigned the LN levels (involved, not-involved) to each element along the diagonal, as follows:

$$M_{i,i} = \begin{cases} 2, & \text{if } M_{l,i} \text{ AND } M_{r,i} \text{ are involved} \\ 1, & \text{if } M_{l,i} \text{ OR } M_{r,i} \text{ is involved} \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

where M_{ii} is the graph node corresponding to LN level i , and M_{ij} is an edge between graph nodes i and j . Furthermore, edges between two involved LNs nodes were assigned a value of 1 in the matrix, according to the dual graph in Figure 18 (left). Since the RP LN level appears as a disconnected node on the graph, we handle it as a special case and encoded its status via two boolean flags related to the left and right involvement. Therefore, the resulting matrix M has dimensions of 9x9, for the nine groups of lymph nodes that are connected in the graph representation.

For later analysis, we furthermore encode the laterality of nodal involvement for each patient using the position of their primary tumor: for patients with right-sided primary tumors, right-sided LNs are encoded as ipsilateral' structures with tumor on the right; for patients with left-sided primary tumors, left-sided LNs are encoded as contralateral' structures with tumor on the left.

4.2.4 Similarity Computation

We designed two similarity measures to investigate whether incorporating spatial information about the lymph node chains (i.e., the spatial location and neighborhood of the nodes involved) partitioned patients more meaningfully than only considering the level itself (i.e., non-spatial labels). Each measure was designed around the non-binary Tanimoto coefficient [96] using either: a) each patient's LN level involvement status only (i.e., only the affected nodes in the graph representation) to measure the non-spatial similarity or b) a combination of status and pathways (affected nodes and edges in the graph representation) to measure the spatial similarity. We chose the Tanimoto coefficient based on its ability to produce the most "meaningful" rankings for smaller, diverse graphs [97] when compared against subgraph [98] and substructure [99] measures.

4.2.4.1 Spatial and Non-Spatial Similarity Measures

After constructing the adjacency matrices M (Eq. 4.2.3), a vector was instantiated for each patient using the involvement status of their LN levels, as follows:

$$v_{p,i} = \begin{cases} 2, & \text{if } LN_{L_i} \text{ AND } LN_{R_i} \text{ are involved} \\ 1, & \text{if } LN_{L_i} \text{ OR } LN_{R_i} \text{ is involved} \\ 0, & \text{otherwise} \end{cases} \quad (4.2)$$

where $v_{p,i}$ is the vector element that corresponds to the involvement status of the left and right LN levels i for patient p . These values were extracted from the main diagonal of each patients' matrix M . Then, to incorporate the spatial information into the measure, additional elements were appended to the resulting vector to encode the edges to and from the involved LNs as defined by the topological map. We enumerated every pair of involved LN levels connected by an edge as a bigram [100] label and added them to the involvement vector $v_{p,i}$ (Eq. 4.2.4.1). We choose not to enumerate further than the two-node combinations because of the small number of nodes in the graph – if all n-grams were enumerated, the similarity distance between patients would increase, and the similarity score for partial pattern matches would decrease. Furthermore, permutations of each bigram are considered once (e.g., bigram permutations between LN levels 2A and 2B, 2A-2B and 2B-2A, are considered as being the same).

Once enumerated, the bigrams on the left- and right-side were encoded into the vector:

$$v_{p,B} = \begin{cases} 2, & \text{if } B_{L_{i,j}} \text{ AND } B_{R_{i,j}} \text{ are involved} \\ 1, & \text{if } B_{L_{i,j}} \text{ OR } B_{R_{i,j}} \text{ is involved} \\ 0, & \text{otherwise} \end{cases} \quad (4.3)$$

where $v_{p,B}$ is the vector element that corresponds to the combined left- (L) and right-side (R) bigrams B of involved LNs i and j for patient p . Overall, 13 bigram weights were added to the vector to represent the 26 bigrams on both sides of head and neck.

Next, the cohort was ranked in pairwise-fashion by computing the Tanimoto coefficient between each of the newly constructed vectors:

$$T(v_p, v_q) = \frac{v_p \cdot v_q}{\|v_p\|^2 + \|v_q\|^2 - v_p \cdot v_q} \quad (4.4)$$

where the function $T(v_p, v_q)$ returns the Tanimoto coefficient between the vectors v of patients p and q .

In order to examine the merits of the spatial measure, we likewise constructed a vector using only the involvement status of the LN level labels (Eq. 4.2.4.1) for the non-spatial (categorical) measure and again ranked the cohort in pairwise-fashion by computing a Tanimoto coefficient (Eq. 4.2.4.1).

To illustrate, in contrast, how these measures work, let us consider patients #14 and #245 from Figure 19 (top left). Patient #14 possesses a bilateral involvement between LN levels 2A, 2B, 3, 4, and 5B,



Figure 19. Example similarity ranking. Patient #14 (shown top left) is unique within the cohort, in that no other patient in the 582 patient cohort exhibits the same ten bilateral LN levels and RP involvement. Following Patient #14 are the seven closest-ranked patients (shown in left-right and top-down order) based on our spatial similarity measure. The two most similar patients share eight bilaterally involved LN levels; the next two have similar bilateral chains but either share fewer involved LN levels (Patient #10128) or possess two additional involved LN levels (Patient #84); while the last three similar patients have similar involvements but with significantly fewer LN levels.

and a unilateral involvement on one RP LN level, while Patient #245 possess a bilateral involvement between LN levels 2A, 2B, 3, and 4. Figure 20 illustrates the corresponding vectors that are constructed for the spatial (Figure 20a) and non-spatial (Figure 20b) measures using Eq. 4.2.4.1 and Eq. 4.2.4.1. Computing the Tanimoto coefficient (Eq. 4.2.4.1) between both sets of patient vectors results in a similarity score of 0.87 for the spatial measure and 0.76 for the non-spatial measure.

$$\begin{array}{cc}
 \mathbf{v}_{14} = \begin{array}{c} \text{RP} \\ 2\text{A} \\ \vdots \\ 5\text{B} \end{array} \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 2 \end{bmatrix} & \mathbf{v}_{245} = \begin{array}{c} \text{RP} \\ 2\text{A} \\ \vdots \\ 5\text{B} \end{array} \begin{bmatrix} 0 \\ 2 \\ \vdots \\ 0 \end{bmatrix} \\
 \text{(a) Spatial Similarity Measure} & \text{(b) Non-Spatial Similarity Measure}
 \end{array}$$

Figure 20. An illustration of the involvement vectors \mathbf{v} constructed for Patient #14 and Patient #245.

(a) The vectors \mathbf{v} constructed for the spatial similarity measure (Eq. 4.2.4.1 and 4.2.4.1). (b) The vectors \mathbf{v} constructed for the non-spatial measure (Eq. 4.2.4.1). Note that the spatial vectors (a) include bigrams while the non-spatial vectors (b) do not.

After ranking each patient, we construct two similarity matrices for the spatial (M_{Sp}) and non-spatial (M_{nSp}) measures, using the similarity scores between each patient pair in the cohort. The result of this step is a similarity matrix for each measure, with the number of rows/columns in each matrix equal to the number of patients in the repository. These matrices are then used in the hierarchical clustering analysis. The patient similarity computation was implemented in Python 2.7.

4.2.4.2 Hierarchical Clustering

Once a spatial measure is obtained, stepwise clustering techniques, such as hierarchical agglomerative clustering (HAC), are a quick yet practical approach to group similar subjects without a priori knowledge of the underlying data distribution [101; 102]. For example, recent studies [103; 104] have used hierarchical clustering to define anatomical subgroups of patients and test for clinical significance.

Furthermore, Bruse et al. [104] investigated which distance/linkage combinations would provide the most “clinical meaningfulness” when applied to a cohort of healthy and pathological aortic arches post-surgical repair patients. Their results show that hierarchical clustering using a Matthews correlation coefficient [105] combined with a weighted-linkage [106] function can yield significant patient sub-groups based on spatial features. While we define our own similarity measure in this chapter, we adopt the weighted-linkage function for determining the distance between the groups when performing our hierarchical clustering.

Following a bottom-up approach where each patient was first represented as a singleton cluster, we used a hierarchical agglomerative clustering (HAC) algorithm to iteratively combine clusters in a pairwise fashion, based on the computed similarity scores and linkage distance function. Based on the results from Bruse et al.’s study [104], we chose to use the weighted-linkage function [106] when determining the distance between clusters. At each iteration, the weighted-linkage function calculates the distance between every pair of clusters, i and j , by computing the arithmetic mean of distances (i.e., similarity scores) between all points in i and j . The algorithm then combines the “nearest” (smallest distance) two clusters and continues iterating until only a single cluster remains.

The resulting clustering output for the spatial measure was further summarized in a dendrogram, a tree-like abstraction which illustrates how similar clusters were grouped (x-axis) and at what level (y-axis) they merged. Finally, partitions of highly similar patients were formed by cutting the dendrogram at a specified level. This level was empirically determined based on the calculated expected values of toxicological outcomes, as described in the next section. Clustering was performed using the MATLAB r2018a machine learning toolbox [107].

4.2.4.3 Statistical Analysis

Results from hierarchical clustering are commonly summarized using a dendrogram, a tree-like structure that displays how the elements are partitioned into groups based on the computed similarity and linkage functions [108; 109]. We construct such a dendrogram as described below.

The patient groupings were compared using the Rand Index [110] to determine the measure of similarity between the two measures' clustering output. This measure quantifies the number of pairing agreements between two clusters into a frequency between 0.0 and 1.0, where a value of 0.0 indicates that the clusterings disagree on every pairing of samples and a value of 1.0 indicates that both clusterings are the same. Additionally, the Fisher's exact test [111] was performed on both clusterings using two toxicity binary variables (Y/N) provided with the cohort: the post-treatment aspiration symptoms and feeding-tube necessity at six-months. We chose the more computationally-expensive Fisher's exact test over the Chi-squared test because the high variation of nodal involvement patterns within the cohort yields small numbers of expected values within each group (e.g., for a clustering with $k = 6$ clusters). While when using Chi-squared the number of expected values for each group should be at least 5, to guarantee the significance of the p-value (otherwise a small p-value could be in fact not significant), Fisher's exact test works well on small numbers of samples. Using Fisher's test, the most significant grouping was for $k = 6$ as the number of clusters, and so both clusterings were cut at the $k = 6$ level. Statistical tests were performed using the MATLAB 2018a statistical toolbox [107].

4.2.4.4 Visual Analysis

To facilitate the assessment of our approach by clinicians, we have constructed an application to help interpret the abstracted nodal involvement of each patient in the cohort in the context of the computed

similarity between patients. The visual interface (Figure 19) consists of small multiple representations of the abstract topological map (Section 4.2.3) and control menus which allow a specific patient to be selected and viewed. To keep the representations compact, only one side of the head and neck was abstracted; color was used to distinguish between left (green), right (purple), and bilateral (blue) involvement. The visual interface was implemented using the web technologies JavaScript, HTML, CSS, and the D3 [112] JavaScript library to quickly provide our collaborators (CDF, HE, BE) with access to ongoing results (cross-platform, no installation required, etc.) and to facilitate future integration of this application into the SMART-ACT [113] environment.

Additionally, we created an informational dendrogram (Figure 22) to convey the patient clustering and statistical analysis results to the collaborating radiation oncologists (collaborators CDF, HE, BE). Side-effect statistics are displayed atop each of the groups formed by the $k = 6$ horizontal cut. The most frequently occurring involvement pattern for each cluster was determined based on the consensus nodal spread of each cluster along the x-axis (at $y=0$) and is shown in miniature at the bottom of each cluster along the bottom x-axis. The consensus was determined based on a two-thirds majority involvement status (i.e., a LN level is included in the graph if 67% of the patients share that involvement). The miniature consensus graphs are a variation of the previously described graph representations: solid and outlined nodes are consensus nodes, affected in more than 67% of the patients in that cluster, while square marks indicate nodes affected in less than 67% of the patients in that cluster. Unilateral involvement is shown by a single consensus graph, while bilateral involvement is shown by two stacked miniature graphs, one for each side of the head and neck. We note that the miniature consensus graphs do not provide a complete descriptor of cluster membership.

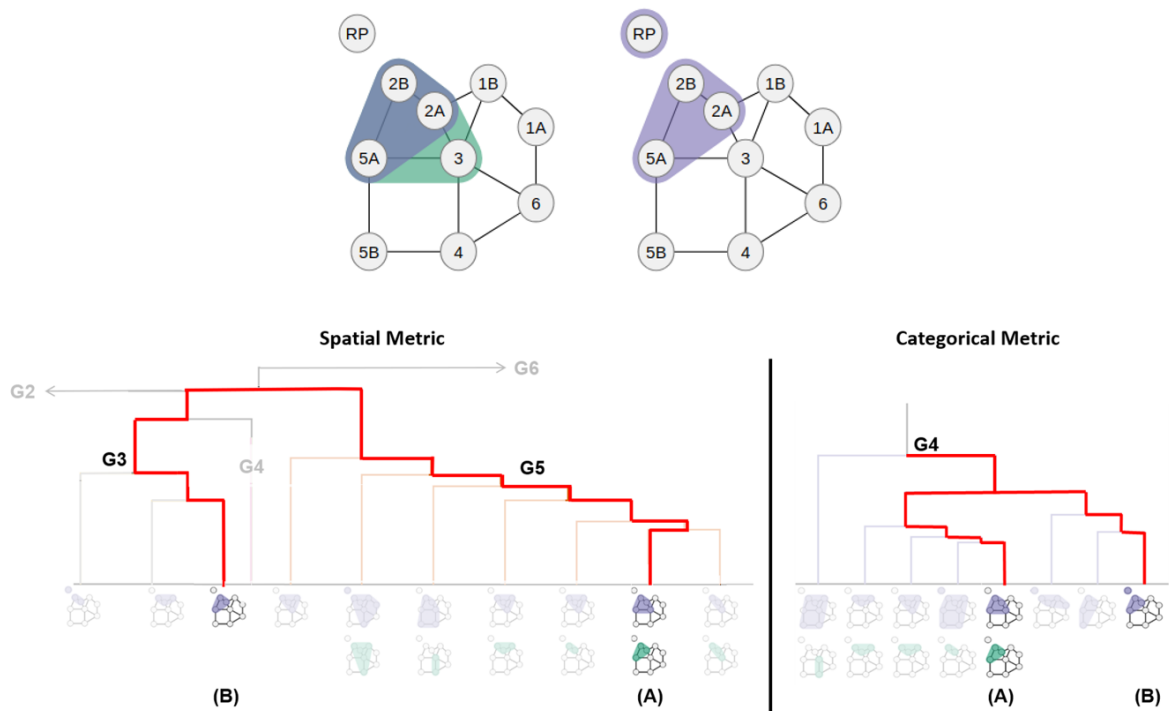


Figure 21. Two subjects with different groupings based on the similarity measure. Patient A (top left) possesses a bilateral nodal spread with LN level 3 involvement while Patient B (top right) only possesses a unilateral nodal spread with LN level 3 involvement. Because the spatial measure uses the geometrically different nodal involvement, it separates Patient A and B into the two main clusters, G3 and G5 (bottom left). In contrast, the categorical measure combines the two patients under the same main cluster, G4 (bottom right).

4.3 Results

4.3.1 Spatial vs Categorical Node Patient Categorization

Our approach was able to successfully discriminate patients based on spatial involvement in cases where the categorical approach failed. For example, the spatial measure was able to discriminate between patients with bilateral spread and patients with unilateral node involvement by placing them into

separate cohorts. The spatial measure also discriminated between RP node involvement versus no involvement, regardless of pattern spread complexity. Consequently, this approach allowed for binning of patients in cohorts that were deemed by clinicians and end-users (collaborators CDF, HE, BE) significantly more informative than categorical binning.

Figure 21 shows a representative example of the value of spatial-measure. Shown are two patients that have drastically geometrically different LN level involvements. Using $k = 6$ clusters, these patients are erroneously binned together under the categorical measure (Figure 21, bottom right), while our spatial approach successfully discriminates between them (Figure 21, bottom left). In particular, Patient A possesses a bilateral lymphatic nodal spread as well as a LN level 3 involvement. Involvement of level 3 implies potential radiation dose to laryngeal structures and is thus a potentially meaningful correlate of radiation-associated sequelae [114]. Likewise, RP node positivity discriminates dose to superior pharyngeal constrictor which is atypical and has the potential for specific toxicity discrimination [115]. In the clinicians' assessment, these are important distinctions, given prior data that shows differential swallowing toxicity as a function of superior pharyngeal constrictor (SPC) versus cricopharyngeus muscles [116; 117].

In the spatial measure, Patient A was also clustered together with other patients that have node 3 involvement, while Patient B was clustered together with no other patients that have node 3 involvement. Conversely, Patient B was primarily clustered together with patients with RP involvement (67% with RP involvement), while Patient A was not (16% with RP involvement). However, the RP partitioning may have been more related to the bilaterality of nodal involvement.

TABLE V. Toxicity Outcome Distributions of the Spatial-Metric Groups

<i>Group</i>	<i>Patient count</i>	Feeding Tube Placement	Aspiration
		<i>Patients with outcome (%)</i>	<i>Patients with outcome (N%)</i>
G2	174	31 (17.9%)	28 (16.1%)
G3	28	3 (10.7%)	4 (14.3%)
G4	77	21 (27.3%)	14 (18.1%)
G5	51	17 (33.3%)	21 (41.2%)

4.3.2 Domain Expert Feedback

Qualitative feedback from repeated evaluation with our collaborating clinicians emphasized the usefulness of this approach. When presented with the informational dendrogram (Figure 22, one clinician stated that he felt confident he could take the visualization back to his clinic that day and use it when describing the potential outcome risks alongside proposed treatment plans to his patients. In addition to comparing patients of the cohort, the clinicians also identified several patients whose LN levels had been previously mislabeled in the dataset due to segmentation or data processing pipeline errors.

During the evaluation process, the clinicians noted that it is common practice to delineate patient groups based on bilateral involvements and the nodal spread between LN levels 2 and 3. Of the two approaches to group patients based on their lymphatic nodal spread, the clinicians felt that the spatial similarity measure, which inherently separated patients between uni- and bilateral involvements as well as the LN level 2 and 3 nodal spread, most closely represented what is expected in a clinical setting.

4.3.3 Hierarchical Clustering Analysis

Figure 22 displays the informational dendrogram resulting from patient binning using the spatial measure. In the dendrogram, clusters of highly similar patients are represented along the x-axis using

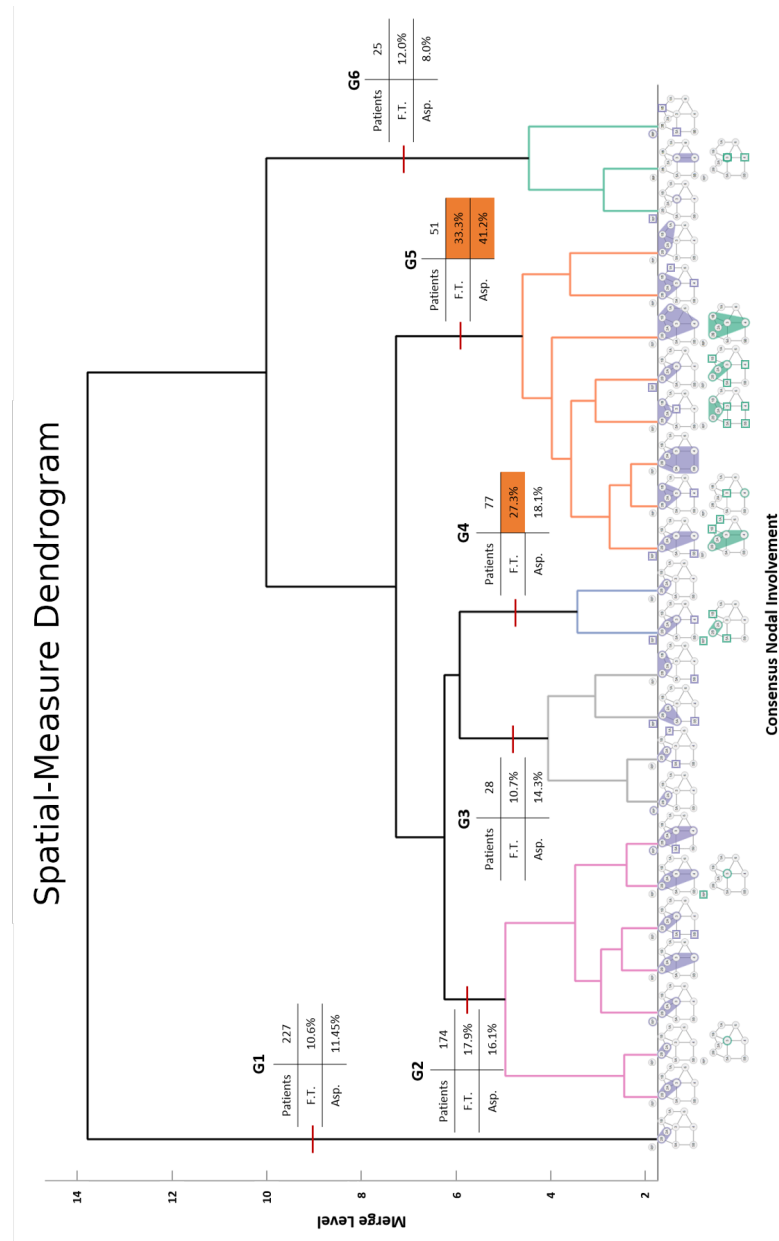


Figure 22. A dendrogram showing the spatial similarity measure $k=6$ patient groupings and corresponding statistics. The six groups are indicated by arrows. In this dendrogram, a clear distinction between bilateral and unilateral nodal spread can be seen between groups G3 and G4, as well a divide between patients with and without LN level 3 involvement (G4 and G4). The consensus involvement (67%) of each group can be identified along the x-axis of the of the dendrogram.

the visual representation defined in Section 4.2.4 D to capture the consensus nodal involvement of the cluster.

In the dendrogram (Figure 22), we identified two distinct groups by focusing on the nodal involvement across the x-axis of each group. First, the cut that separated groups G2-G4 from G5, near merge level 7, also partitioned the cohort according to the involvement laterality: groups G1-G3 consisted of patients with unilateral involvement, groups G4 and G5 of patients with bilateral involvement, and group G6 of patients with unique (singular to the cohort) nodal involvement. Next, the cut that separated groups G3 and G4, near merge level 5, also discriminated based on LN level 3 involvement, creating another clear distinction between groups with (G2, G4, G5) and without (G1, G3) the involved lymph node.

In contrast, the involvement status of LN level 3 occurred throughout each of the six groups generated through the non-spatial/categorical approach. Furthermore, four of the six groups generated through the categorical approach contained patients with bilateral involvement. Therefore, the categorical approach fails to capture a meaningful demarcation between LN level 2 and level 3 involvement, as well as patterns of bilateral involvement.

4.3.3.1 Measure Agreement

In terms of agreement between the spatial and categorical approaches, we identified two identical groups between the spatial- and categorical-approach clusterings (G1 and G6). While these two groups represent 43% (252 patients) of the cohort, the consensus nodal involvements in each are also the simplest patterns in the cohort. For example, all 227 patients in both G1 groups possess a unilateral LN level 2 involvement. G6 groups together all the 25 unique LN level involvement patients in the cohort. Out-

side of these two groups, the categorical-approach did not have the discriminatory value of the spatial approach advocated in this chapter.

After removing the two groups G1 and G6 from each of the clusterings, the computed Rand index between the spatial and the categorical results was a similarity measure of 0.55. This value indicates that outside of the two groups G1 and G6 of simple patterns, the two approaches are significantly dissimilar in terms of how they group the patients within the cohort.

4.3.4 Statistical Analysis Results

Statistical significance is reported assuming a level of $p < 0.05$, based on the occurrence of the toxicity symptoms within the groups. Table V shows the toxicity outcomes distributions of the four spatial-measure groups with the highest incidence rates. In terms of the toxicological outcomes, there was a significant difference in the rate of feeding tube (FT) placement among the $k = 6$ spatial-measure groups ($p < 0.01$). The measure was able to identify two cohorts (G4, G5) that had almost double the outcome incidence compared to the other four (G1-G3, G6). G4 and G5 had FT placement rates of 27.3% and 33.3%, respectively, while G3-G6 had rates less than or equal to 17.9%. Additionally, the spatial measure identified one group (G5) with more than double the aspiration rate (41.2%) compared to the other five groups.

4.3.5 Performance

We performed all computation on a 4.0 GHz Quad Core i7 machine with 32G of RAM. The average run-time to compute the similarity on the cohort of 582 patients was approximately 90 seconds per similarity measure. The hierarchical clustering and statistical analysis averaged 45 seconds to partition

the patients into groups, compute the Chi-squared and Fisher’s exact test, and output the statistics and dendrogram per measure.

4.4 Discussion

Our analysis of results and the domain expert feedback support our claim that spatial correlates can provide insight into therapy strategies where treatment depends on the spatial patterns of disease, such as intensity-modulated radiation therapy for HNSCC. The spatial method we introduce captures and ranks patients correctly and more clinically accurately compared to the categorical approach. Furthermore, we have shown that when combined with hierarchical clustering, our novel graph-based similarity measure partitions an OPC patient cohort into clinically meaningful groups. In particular, we have shown that our spatial approach can capture groups of patients more susceptible to dysphagia toxicity (aspiration and feeding tube) based on the pattern of nodal involvement.

In terms of limitations, our similarity measure captures but a few of the many features that can be used in therapy response-driven decisions and predictive outcome models. While toxicity is heavily predicated on the relationship between the spatial location of involvement and the administered radiation dose, many therapy outcomes and side-effects result from other non-spatial features. A direction of future research, while beyond the scope of this work, would be to combine our spatial similarity scores with other relevant non-spatial features, such as T-Category and patient age [118], to create a more semantically meaningful view of the patient regarding treatment response and survival. Likewise, our approach notes but does not explicitly incorporate into the similarity measure, the tumor location with respect to the lymph-structures (which is typically upstream in the head and neck). Other clinical

applications may feature higher variability in the tumor location, and in those cases, the location of the tumor may need to be explicitly incorporated into the similarity measure.

Next, we note that our evaluation was limited to one moderately sized cohort of patients. Many of these patients were referrals whose data was collected outside of the treatment facility. As a result, a significant amount of time spent working with this cohort was spent cleansing the data of malformed classifications. Furthermore, our expert feedback was limited to radiation oncology clinicians who were all members of the same clinical lab. Last but not least, our approach is constructed around a 2D graph representation that takes advantage of the symmetry about one of the principal axes of the structural model. While this approach is ideal for domains where symmetry is inherently built into the model (e.g., symmetry about the head and neck), it may also be easily extended to non-symmetric situations. In contrast, extending this approach to situations where 3D location is important would require modifications to the underlying graph representations and similarity measure.

4.5 Conclusion

In conclusion, we have introduced and evaluated a novel methodology to compare head and neck cancer patients based on their spatial patterns of LN involvement. Our approach demonstrates how the spatial location and neighborhood of the head and neck LN levels can be abstracted to a 2D topological representation, which can then be used to quantify similarity within a cohort of patients based on their extracted spatial attributes. This work also contributes two visual representations that provide clinicians with response-based correlates within the ranked cohort. Statistical analysis and expert feedback indicate that our spatial approach can be useful in clinical settings. Furthermore, we show that our spatial approach provides superior patient similarity and groupings in terms of clinical relevance when com-

pared to the categorical approach. The presented methodology may find application beyond the 2D head and neck lymph node analysis in other domains that feature topological structures.

Few, if any, studies have attempted to use spatial-similarity techniques to compare post-diagnosis patients and “close the gap between mere data and useful knowledge, as desired in current Precision Medicine” [104]. Moving forward, we aim to integrate our proposed measure into a risk-prediction model. We believe that when applied to spatially-driven diseases such as OPC, approaches such as ours can play a vital role in fulfilling precision medicine’s goal of maximizing the effectiveness of each patient’s treatment through customized care [119].

CHAPTER 5

“DETAILS-FIRST, SHOW CONTEXT, OVERVIEW LAST”:

SUPPORTING EXPLORATION OF VISCOUS FINGERS IN LARGE-SCALE ENSEMBLE SIMULATIONS

This chapter was originally published in the IEEE Transactions on Visualization and Computer Graphics (TVCG) Journal © in 2018 [3]. This version has been edited to be consistent with the rest of the dissertation. In particular, relevant sections that appeared in the original manuscript were relocated to the Introduction and Discussion chapters of this dissertation to help motivate this work and explain its contributions. Coauthors on this work include Andrew Burks (AB), Cassiano Sugiyama (CS), Jonathan Komperda (JK), and G. Elisabeta Marai (GEM). The contributions from each author included: AB and CS co-designed and developed the FingerFinder web-application, including the underlying visual encodings and algorithms, as National Science Foundation REUs under the direction of GEM and me; AB also designed the layout of the merge tree to reduce cluttering between finger tracks; JC served as the mechanical engineering expert for this project, provided theoretical and qualitative feedback on the Details-first model, and provided support with the software testing; GEM conceived this project and its theoretical framework, and directed the design, implementation and testing of the FingerFinder tool. My (TL) contributions to this work included the conception of the merge tree for the tracking of the features over time as well as contributions to the design and implementation of the FingerFinder tool.

Additionally, I worked with GEM and JC on developing the theoretical underpinnings of the Details-first approach. I am the first author on this work.

This chapter rounds out our investigation by examining a situation in computational fluid dynamics (CFD) where scientists often benefit more from design strategies that provide the ability to first access details instead of an overview of the dataset. In these situations, experts often have an excellent mental overview of their data [8]. As a result, these experts often directly explore spatial features of interest in the early stages of their workflows; overviews, when employed at all, tend to exist to the latter stages of the workflow and are often in the form of summary statistics. These workflows are different from those we have encountered in the previously described domains. Up to this point, we have presented design strategies generally based on providing the expert with some form of awareness of their data by harnessing the Overview-first or Search-first strategies, or a hybrid combination of the two. However, in domains such as CFD, these two models do not present the target users with the detailed information that they wish to explore from the onset of their analyses.

To address these requirements, we present an alternative model to the Overview-first and Search-first mantras. This alternate approach can be defined as “Details-first, show context, overview last,” and supports situations where the user workflow is oriented along the spatial or spatiotemporal feature analysis. Using a practical example in the CFD domain as a vehicle to drive our analysis, we demonstrate how our Details-first approach can be used to inform design strategies in large-scale, spatial data collaborations. Specifically, we discuss how these design strategies – from the novel abstract representations to their visual arrangement within the interface – led to the instantiation of our Details-first approach.

5.1 Introduction

As we have seen earlier in this dissertation, a common goal in visualization is the design of techniques that provide both overview visualizations and support for feature exploration. Overviews can help the user find regions where further investigation in more detail might be productive [5]. Spatial features are, in turn, at the very core of most engineering and biomedical visualization endeavors, from vortices in flow simulations to bonding sites in protein structures.

While many such visualization designs follow the information seeking mantra: “Overview first, zoom and filter, then details on demand” [6], there are situations where providing an initial overview is not relevant or practical for users, while providing details is paramount. For instance, in a wide class of problems, including the problem illustrated in this chapter, details do not have a precise definition, and their identification relies on internalized knowledge in the domain expert’s head. As further argued by van Ham and Perer [7] in their alternative “Search, show context, expand on demand” mantra for large graphs, there is also a significant class of scientific users who are not interested in global patterns in the data, but have specific questions about one or several specific data points. As a practical example, an astronomer who studies a class of quasars is typically not interested in an overview of the entire observable universe [1]. A step further, in computational fluid dynamics (CFD), domain scientists often work on the same problem for months, and have a good mental overview of the underlying data [8]. From an information theory perspective, Chen et al. [8] argue briefly that in such a case, having the direct ability to reach a detailed view (details-first) would reduce the cost of step-by-step zoom operations. Nevertheless, visualization textbooks only report the Overview-first mantra and the Search-first mantra [5].

Other arguments against first presenting global overviews to users are of a more practical nature. As illustrated in this chapter, overviews may be derived from imprecisely-defined details and thus may not be readily available. In the case of large-scale multidimensional datasets, creating an overview may also not be feasible, in particular when a large dataset is being maintained at a centralized location, and transferring it to multiple client machines is not an option [7; 9; 10]. Last but not least, in some scientific problems, for instance in simulation ensemble visualization [11], the problem overview is not one spatial dataset, but a collection of datasets, whose summarization in an overview is not necessarily clear to the domain expert.

This chapter provides theoretical and practical evidence to support an alternative approach to the two established design mantras, Overview-first and Search-first. This alternative can be defined as “Details-first, show context, overview last,” and supports situations where the main user workflow is oriented along spatial or spatiotemporal feature analysis, while the problem overview can only take the form of a summary. In this model, the analysis starts with the spatial feature(s) of interest, with the help of a computational back-end that can help identify and track those features over space and time. The detail features are then used to automatically filter the feature-context in space and time, while controlling the complexity of the visualization. Last, detail-derived calculations are used to summarize and compare collections of features and potentially datasets, presenting a summarization overview to the user.

We construct this alternative model with the help of scientific workflow theory [20] and of a practical example in the CFD domain, the exploration of viscous fingers in large-scale ensemble simulations [120]. Viscous fingers are areas of high concentration formed when a higher density fluid (e.g., oil) is poured into a lower density fluid (e.g., water); the fingering process is nondeterministic, and can

lead to instability. To study this process, multiple stochastic simulations with non-deterministic properties must be executed, resulting in a simulation “ensemble.” In turn, these simulation ensembles are nearly impossible to analyze computationally, due to the large number of parameters involved and the ill-defined nature of both the analysis process and the finger structures themselves.

Using this problem as a vehicle, the Details-first model allows domain scientists to explore a total volume of data approaching half a billion multi-dimensional data points through an interactive web-based application. The contributions of this work are:

- A Details-first, show context, overview last model for the exploration of large-scale spatial data;
- A constructive instantiation of this model, using scientific workflow theory and the problem of viscous finger exploration; the instantiation constructs methods for identifying and tracking finger structures over time, for filtering the spatiotemporal context of the computed features, and for supporting overview summarization;
- An evaluation of this model on a large-scale dataset, including feedback from CFD researchers on both the instantiation and the underlying theoretical model;
- A discussion of the merits, applicability and limitations of this approach, and of its fit with existing models.

5.2 Background and Related Work

We begin our discussion by highlighting representative work on detail identification for spatiotemporal CFD visualization; we follow with a summary of representative work that uses CFD details to

visually filter data, and of work in ensemble visualization. For a review of supporting paradigms and terminology in visualization design, refer back to Chapter 1.2.

5.2.1 Spatial Features as Details

In the context of information theory and CFD, Chen et al. [8] directly relate details to spatial features. Obermaier and Peikert [121] further note that the concept of feature in scientific visualization is derived from its definition in computer vision [122], where it describes a salient feature of an image, such as an edge or a ridge. For example, features in flow visualization include vortices, shock waves, isosurfaces, separation lines, and statistical features.

5.2.2 Features and Soft-knowledge

Obermeier and Peikert [121] note that in the ideal case, features have a precise mathematical definition which does not depend on any “tuning” parameters. In contrast, other feature definitions involve a parameter and “require a visualization system where parameters can be controlled by the user.” [121]. Similarly, Weber and Hauser [121] define features as data subsets of interest to the user, sometimes “due to prior knowledge.” Last, Chen and Golan [123] introduce “soft information, knowledge, and priors” in the context of information theory in visualization, to capture known theories, intuition, belief, and meta-knowledge about a system.

In our work, details denote spatial features. Following Chen and Golan, *soft-knowledge features* denote those spatial features whose definition involves one or more parameters controlled by the user. The “soft” qualifier refers to the fact that this type of knowledge is difficult to capture and represent computationally.

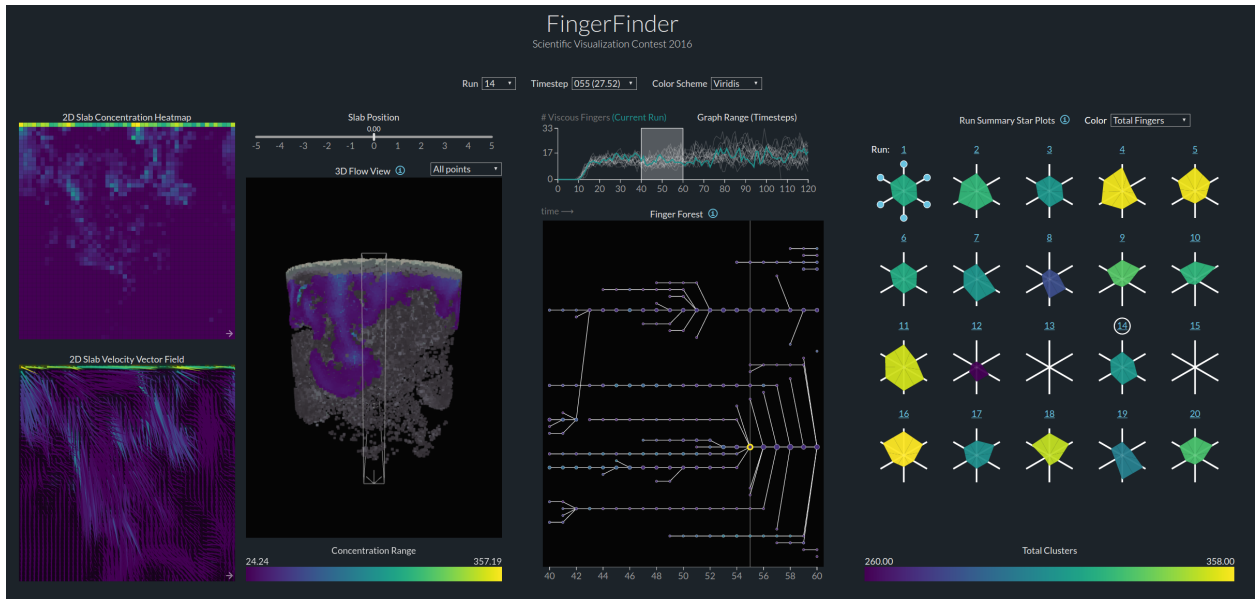


Figure 23. The Details-first, show context, overview last model supports the interactive, web-based exploration of ensemble simulations. From left to right: detail and spatial-context panel comprising two 2D slices and a 3D flow view; a temporal-context panel comprising a time chart and a finger forest; overview panel showing a small multiple of Kiviat diagrams. Linked interaction and a computational back-bone allow users to identify fingers and track their evolution over time, and to analyze the data at multiple levels of detail.

5.2.3 CFD Visualization

5.2.3.1 Feature Extraction

A common practice in the visual analysis of CFD spatiotemporal relationships is the detection, extraction, and exploration of features of interest over time [124; 125; 126]. Oftentimes, these approaches require the presence of clearly defined features in isomorphic structures, and are not directly relevant to our illustrative example: finger structures are soft-knowledge features. Favelier et al. [127] and Lukasczyk et al. [128] use an adaptation of Shepard's kernel method [129] to identify such features

based on concentrations. Both these works rely on user-defined thresholds. In our recent work, neural networks have been trained to identify shock features based on descriptors such as the strain tensor and schlieren value at each timestep [10]. In a similar machine learning approach, Maries et al. [9] utilize K-means clustering to group and label points in areas of interest based on the velocity stress and strain tensor. Our finger identification method resembles Maries et al.’s [9] approach in that we define features based on groups of points with similar salt concentrations. However, we threshold the feature groups based on local-proximity and point concentration.

5.2.3.2 Feature Tracking

Most methods extract soft-knowledge features from each timestep separately and track how they progress over time [130; 131; 132]. These methods rely on the temporal and/or spatial coherence of attributes and location of the feature as it moves through time and space. Other methods [133; 134; 89; 135] use a contour-based, merge-tree ideology to enable tracking of regions of interest in combustion simulations. Our finger tracking solution builds upon the combined success of these spatiotemporal feature graphs.

5.2.3.3 Feature-based Filtering

CFD data is multivariate and dense, causing visual occlusion even at modest scales. In consequence, the body of work that uses details for filtering flow data is enormous. Here we report only on the works most relevant to our approach, where the features do not have a pre-defined formula for extraction. Multiple coordinate views (MCV) have been deployed simultaneously to explore multivariate data and identify potential regions of interest [18]. These approaches harness linking-and-brushing techniques [136] to select and filter features between the multiple views. Furthermore, many of these

approaches follow a focus+context style, where a general view or physical view is brushed to uncover specific features [137; 136; 11].

However, this approach is difficult in the case of temporal features—users have to mentally integrate multiple samples across timesteps to understand the feature over time [138]. In our work, the data is automatically filtered based on the finger structures we extract.

5.2.3.4 Ensemble Visualization

Multiple simulations are often used to quantify and mitigate uncertainty in models that contain stochastic effects [139]. The resulting multiple simulation runs (collectively termed “a simulation ensemble”) are often large, multivariate, multi-valued and time-varying, and have been described as “awkward” [140] and difficult to visualize [141; 142]. Many ensemble visualizations aim to reduce complexity by presenting basic derived statistics such as the mean and standard deviation of observed properties [139; 11]. Alone, these techniques can capture ensemble variations between runs and provide strong indications of overall ensemble behavior. However, unlike our work, they may not capture more nuanced changes across time-steps or among ensemble members, and do not attempt to display user-defined spatial features.

Basic visual abstractions such as line charts, quartile charts, and histograms are commonly used in ensemble visualization to encode statistical parameters [143], as well as reduced spatial aggregate views [128; 144] to display specific attributes at a specific time and location. To facilitate further exploration of ensemble members across space and time, these aggregate views are linked to range-based representations [145]. These representations may include colored overlays, multidimensional scaling

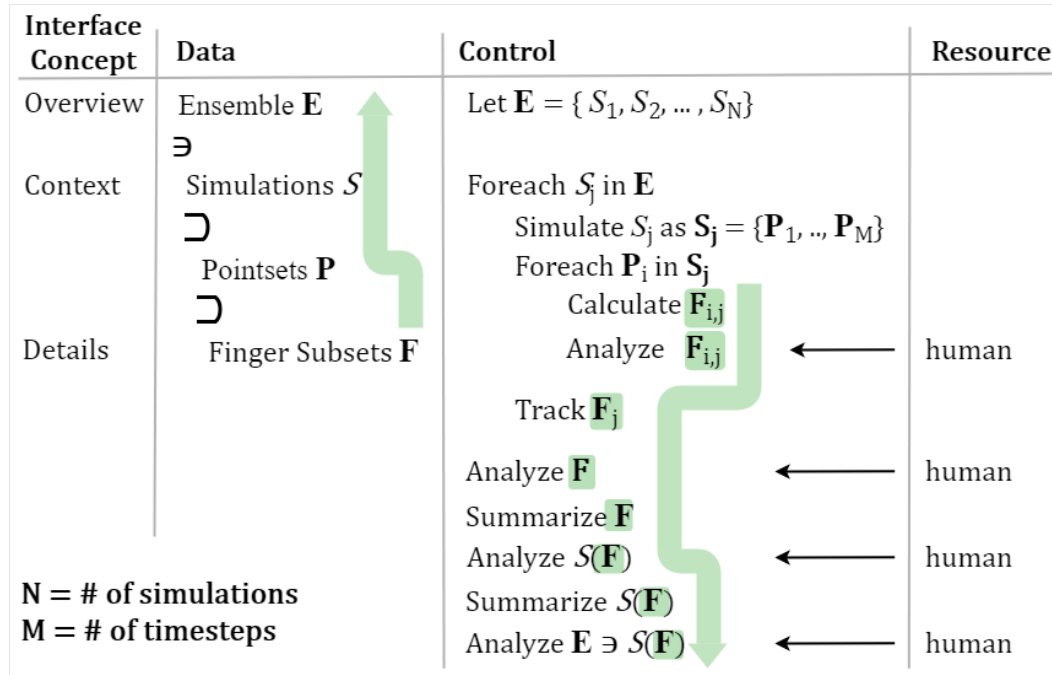


Figure 24. Workflow decomposition of the finger calculation and exploration process along the main axes of scientific workflows: data, control, and (human) resource components. An additional column maps the data elements to the design components corresponding to overview, context and details. The Resource column only shows the steps where humans are involved; the remaining steps are computational. Note how the details \mathbf{F} travel down the control flow, and up the data and the interface elements.

projections [146], and various types of tracking graphs [128; 133; 134]. We use similar encodings for several of the features we compute.

5.3 Model Instantiation

5.3.1 Constructive Example and Workflow Analysis

We illustrate the Details-first approach on a constructive example from the IEEE VIS 2016 Scientific Visualization Contest [120]. The visualization design (Figure 23) was created in close collaboration with

a CFD researcher (collaborator JK) with over seven years of experience in turbulent flow computational research.

5.3.1.1 Data and Tasks

The contest problem is centered on the spatiotemporal exploration of viscous fingers in large-scale ensemble simulations. One of the datasets provided is a simulation ensemble containing multiple stochastic simulation runs. Each simulation run in the ensemble captures the diffusion of an infinite salt source placed at the top of a cylinder filled with pure water. Over time, the higher-density salt diffuses into the water, forming structures known as viscous fingers. Each simulation is run using a Finite Pointset Method (FPM) approach with 250,000 points at the lowest resolution, and over 120 timesteps per simulation. In this ensemble, “the behavior of so-called viscous fingers is of primary interest. The six-dimensional nature and size of the data is the main challenge for visualization. Effective browsing, summarization, and data reduction strategies are needed to obtain meaningful insight into the data” [120]. The simulation ensemble cannot be analyzed purely computationally, due to the large number of parameters involved and the ill-defined nature of both the analysis process and the finger structures themselves.

5.3.1.2 Model Perspective

From a model perspective, the finger structures are defined based on soft-knowledge on the user side. Finger structures are typically visualized and identified via the use of concentration thresholds and contours. In this approach, a threshold is specified, and the structures are identified at the interface where concentrations are greater than or less than the threshold. This approach, which is not an exact formula for finger structure, is based on the knowledge that a) the features have higher concentration

than surrounding areas, and that b) the features form blobs that are similar in shape to fingers. In other words, the finger structures are features that depend on the expert's soft-knowledge.

The second aspect weighing into the model perspective is that the details are here the finger structures and their evolution over time. The context is likely the physical volume around the spatial features, respectively the features' behavior over time across simulations. The overview, in turn, can be considered at two levels: 1) the spatial overview of all simulations, respectively 2) a summarization of the simulations. The spatial overview (a plain upright cylinder with 250K points) poses clutter and rendering time challenges, and its overall structure is also familiar to the domain experts. The summarization overview, in contrast, will likely be encoded by a visual abstraction unfamiliar to a CFD expert.

5.3.1.3 Scientific Workflow Analysis

Given these observations, let us consider the problem from a workflow perspective. In particular, the finger calculation and exploration process can be decomposed along the main axes of scientific workflow theory [20]: data, control, and (human) resource components. In scientific workflow theory, *data* captures the information that is required during the execution of a workflow; *control-flow* describes the set of steps that make up the process and the way in which the thread of execution is routed between them; *resource* identifies the people and facilities that actually carry out the process. Figure 24 captures the data, control and resource elements for this problem, with an additional column mapping the data elements to the design components corresponding to overview, context and details.

This decomposition captures a number of traits of this workflow. First, the spatial features (i.e., details, highlighted in green in Figure 24) are central to the entire process. Simulation summaries are a function of the finger characteristics, and the ensemble is a function of the simulation summaries, and

thus also a function of the finger characteristics. In other words, the context is a function of details ($S(\mathbf{F})$), and the overview is a function of details ($\mathbf{E} \sim (S(\mathbf{F}))$).

Second, a human is involved in all the analysis steps. Because finger structures are identified empirically, human input is necessary at that stage. Human input is necessary when selecting the set of measures used to characterize the finger structures. A human is further necessary when analyzing a simulation and extracting the measures that characterize the simulation in terms of its fingers, and when analyzing the entire ensemble.

In the following sections, we describe briefly the computational and human-input steps in this example, along with the visual encodings for each output, and then the resulting visualization design.

5.4 Finger Segmentation and Spatial Context Calculation

The description of the segmentation step captures the close interplay between human input and the feature identification process. In order to identify features within the data using the definition of a viscous finger (a contiguous area of “high” concentration), we run a custom clustering algorithm on the data. Along with determining the finger structures, this process simultaneously allows us to calculate the spatial context of the fingers.

Because the simulation data is mesh-free, and provided in the form of a seven-dimensional point cloud (point position, velocity and concentration), the first step was to construct an adjacency network that captures the local neighborhood of each point. Next, a simple clustering algorithm was run on this adjacency network, grouping together those points within the network which had a high concentration of above $\mu + \sigma/k$, where μ and σ are the mean and standard deviation of concentration for that timestep, respectively, and $k = 7$ was an empirical value determined through visual analysis. The clustering

algorithm iteratively connects the nodes of the graph into clusters, based on the relative position of each point to its neighbors. For each point, the heuristic polls the cluster to which the neighboring points belong. If only one neighbor belongs to a cluster, the heuristic adds the point to that cluster. However, if the neighbors of the current point all belong to a different cluster, the heuristic combines those clusters and adds the current point to it.

Using the concentration heuristic alone can lead to all clusters near the saline top (which is a constant source of high saline concentration) being grouped together. To circumvent this artifact, the algorithm ignores points within 0.5 units of the top of the cylinder. In particular, the CFD expert (JK) noted that ignoring points immediately near the boundary condition is logical and acceptable because, by any definition of a finger, a constant source would satisfy the finger concentration condition. These empirical thresholds for the clustering can be calibrated, however the data will need to be reprocessed to perform the clustering again with the new thresholds.

The final clusters that result from this algorithm form the viscous fingers for that timestep. The algorithm assigns each cluster a unique cluster identifier. By keeping track of both cluster identifiers and point IDs, we can track the points whether or not they appear in one of the clusters as they move through time.

5.4.1 Finger Visualization

Finger structures and their spatial context are visually represented using 3D & 2D views (Figure 25).

5.4.1.1 3D View and Context

Finger structures can be inspected in a 3D Flow View. Users can select the specific simulation and timestep to view. The cylindrical 3D view provides the context of the simulation domain, with the saline

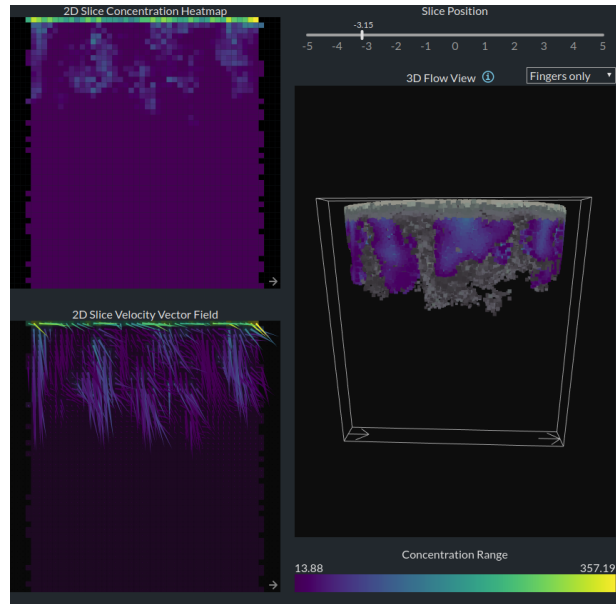


Figure 25. Detail and spatial context. The 3D Flow View (right) provides the spatial context of finger structures. A vertical slab can be used to analyze finger structures in detail using 2D Views (left). This snapshot captures the formation of two large finger structures.

layer displayed at the top. Each point color is mapped to the concentration of that point. To provide further context, we display points considered by the finger clustering algorithm in gray (i.e., points of higher concentration).

5.4.1.2 Vertical Slab and 2D View

To alleviate cluttering, a 3D vertical slab (cutting plane with depth) is used to carve out a subset of points for in-depth analysis. The slab points can be analyzed through linked views which show the concentration heatmap and velocity vector field of the data contained in the slab. We chose the slab representation, as opposed to a plane, because fingers are not restricted to 2D; the slab can be used to capture the finger depth along the cylinder cross diameter as well. The linked 2D views aggregate over

the slab width the average value of the concentration and velocity, and help analyze in detail the finger concentration and velocity. Fingers may be viewed by moving the slab onto a specific viscous finger, which in turn allows the user to view the concentration heatmap and velocity vector field of the slab containing the finger. All viscous fingers are displayed by default. Specific viscous fingers can be also interactively highlighted in the 3D Flow view through the Finger Forest view described further below. Selecting a specific node of a tree highlights the finger in color (mapped to concentration).

5.4.2 Finger Properties and Analysis

To locate each cluster in the next timestep, several properties of each cluster are calculated and used, with input from the domain expert. From the finger segmentation output, this approach produced for each finger an attribute set which includes:

- the number of points
- the total concentration
- the concentration-weighted average position of points
- the concentration-weighted average velocity of points
- the average magnitude of velocity
- the concentration-weighted average magnitude of velocity
- an axis-aligned bounding box around the feature

The first property is an average position of the points within the cluster, weighted by the concentration of the points. Second, the average velocity of the points in the cluster is calculated, also weighted by the concentration of the points. To find the cluster nearest to another in a different time step, the

centers of concentration of the clusters are used, which are both corrected for the difference in time by adjusting the coordinates using the average velocity of the each cluster. The output of these algorithms are multiple clusters of points for each timestep, as well as the information linking these clusters to each other across multiple timesteps. The results of the data preprocessing are used for feature tracking, as well as in the simulation summarization.

From a model perspective, notice how extracting the details relies on soft knowledge on the user side; and how the domain expert input is essential to extracting the measures to characterize the features and their context.

5.5 Finger Tracking and Temporal Context Calculation

To track the fingers' evolution across a simulation, we run a two-stage algorithm on the finger clusters that were identified in the previous step. This process allows us to determine the temporal context of the finger evolution. This temporal context captures the appearance, dissipation, merging, and splitting of fingers.

The two-stage algorithm is based on the tracking graph algorithm proposed by Bremer et al. [133]. The algorithm first uses the size, position, concentration, and average velocity of the viscous fingers to label each cluster in each timestep with an ID, unique to each viscous finger over the course of the simulation; in other words, fingers within a single timestep can not share an ID, but fingers between timesteps can. In the second stage, these IDs are used to index the fingers into an adjacency list per timestep. A grouping procedure is then run on each pair of consecutive timesteps, constructing relationships between the fingers, over time. For each of the $M-1$ pairs of consecutive timesteps, the algorithm iterates over the two lists, comparing both finger properties and IDs. If the procedure finds a match

between both properties and IDs, then the corresponding finger persisted between the timesteps and the two adjacency list entries for that ID are linked. Similarly, if the procedure finds a match between properties but not IDs, then the finger in the earlier timestep has merged into the finger in the later timestep, and the two adjacency list entries are connected. Finally, the procedure treats all unmatched nodes as either having split or dissipated, depending on whether the unmatched node is present in the later or prior timestep, respectively.

The trees output by the algorithm capture the evolution of each viscous finger throughout a simulation. We assign each tree the same ID as the viscous finger that is mapped by the tree root. By binding the finger structures to the trees, we can track the spatial features as they evolve. The linked IDs also allow us to select a node in the tree interactively and highlight that specific viscous finger in the 3D Flow view and 2D feature views.

5.5.1 Simulation Analysis

To analyze the finger evolution over time, we turn again to input from the domain expert. We will use two of the finger properties previously derived, as well as an additional parameter. These properties are: the number of points in each finger, the average concentration of the points in each finger, and now also the total number of fingers in the simulation.

5.5.1.1 Temporal Context Visualization and Filtering

The temporal context calculated in this step is shown in a temporal-encoding panel. The panel contains one horizontal, time-aligned tree for each finger in the simulation, as well as a Time Chart which can be used to control the temporal context shown.

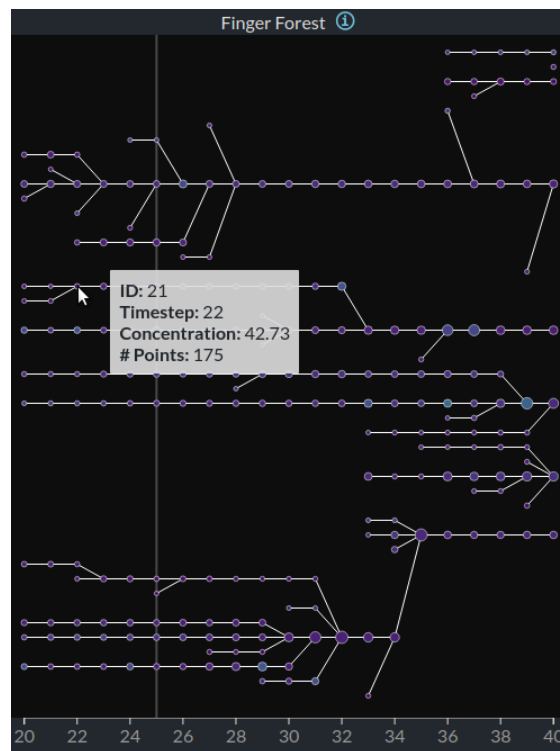


Figure 26. Temporal context visualization. A Finger Forest shows one horizontal, time-aligned tree for each finger in the simulation. Each node in a tree represents one viscous finger as that timestep. Nodes are colored by the average concentration of the points in the finger, and the radius of each node is scaled by the number of points in the finger. The trees may merge or split according to the finger evolution over time. A vertical bar indicates the current timestep.

5.5.1.2 Time Chart View

The Time Chart can be used to select the range of timesteps to be graphed in the Finger Forest. The number of fingers in each timestep is graphed for all simulations, with the graph for the current simulation highlighted in color.

5.5.1.3 Finger Forest View

The Finger Forest (Figure 26) displays a set of horizontal, time-aligned trees that encode the evolution of fingers in a simulation, over the time interval selected in the Time Chart. Each node in a tree represents one viscous finger as that timestep, similar to the graphs of Bremer et al. [133]. The nodes are colored by the average concentration of the points in the finger, and the radius of each node is scaled by the number of points in the finger. The trees may merge or split according to the finger evolution over time. A vertical bar indicates the current timestep.

In order to minimize edge-crosses, we balance the trees using a heuristic similar to Widanagamaachchi et al.'s [134; 89]. The heuristic begins with the fingers in the last timestep and recursively enumerates and sorts the children of each finger based on the latest timestep in which that finger appears. For each enumerated finger, the heuristic then splits the nodes into two groups and positions them above and below the parent so that the oldest fingers are closest to the parent, and the most recent ones are furthest from the parent.

We note that both rendering a spatial overview and rendering a complete temporal overview would be impractical in this setting. A spatial overview of the entire information space would be affected by cluttering and rendering constraints. Similarly, a complete temporal overview would be affected by rendering constraints (computation time, minimal node size for visibility, minimizing edge crossings). From a model perspective, filtering by spatial and temporal context helps control visual complexity; these contexts are derived based on detail calculations.

5.5.2 Simulation Summarization and Ensemble Analysis

The last stage of the control-flow in our workflow decomposition (Figure 24) seeks to summarize the properties of the simulations that form the ensemble. These properties are derived from the finger properties, with input from the human expert. One of these properties characterizes the simulation as a whole; five additional properties are computed for each timestep, and averaged over the duration of the simulation:

- the total number of unique fingers over the entire simulation
- the number of fingers in each timestep
- the average concentration of fingers in each timestep
- the average concentration of points in viscous fingers in each timestep
- the average finger speed (points' average magnitude of velocity) in each timestep
- the number of merges (not including fingers which disappear) in each timestep

5.5.2.1 Ensemble Analysis

The simulation properties are summarized in a small-multiple overview panel (Figure 27). The panel comprises one Kiviat diagram [147] per simulation. Kiviats are a graphical method of displaying multivariate data in the form of a two-dimensional chart, in which three or more quantitative variables are represented on axes starting from the same point. Unlike most radial plots, which tend to capture temporal sequences, the Kiviat relative position and angle of the axes is typically uninformative. Kiviat are equivalent to a parallel coordinates plot (PCP) in polar coordinates, and are seldom effective when

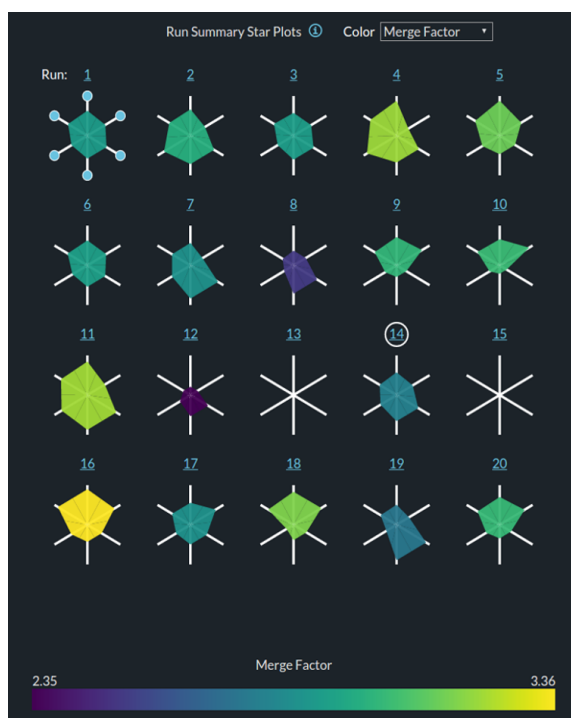


Figure 27. The Kiviats diagram panel captures similarity between members of the ensemble. The Kiviats map their axes to six computed properties of each simulation. The color can be mapped to different properties, for example the total number of fingers of each simulation. Note the similarity (diagram shape and color) between simulations 1, 3, 6, and 14. Simulation 12 stands out as an outlier. Simulations 13 and 15 are empty (no content at the 250K resolution).

more than two Kiviats are overlaid [148]. However, due to their closed polygon shape, which is a pre-attentive feature, Kiviats are particularly effective in small multiple form [113]. The axis ordering is not an issue, because each Kiviats uses the same axis ordering across the small multiple, resulting in similar polygon shapes for similar simulations.

Each Kiviats axis is mapped to one of the simulation properties. Hovering over each Kiviats axis shows how each property was computed. The Kiviats are further color-mapped to a simulation property

selected by the domain expert, for example the total number of fingers in each simulation. In Figure 23 right, note the similarity (diagram shape and color) between simulations 1, 3, 6, and 14. Simulation 12 stands out as an outlier. Simulations 13 and 15 are empty (no content at the 250K resolution). Through this small-multiple panel summarization, simulation properties can be compared between ensemble members. From a model perspective, these properties were also derived from detail calculations.

5.5.3 Design and Implementation

The model instantiation was developed through a parallel prototyping approach, which included 1) exploring encodings and potential properties, 2) evaluation with a CFC expert and revising properties, and 3) discarding a variety of measures as well as encodings (including parallel coordinate plots and scatterplots). The work benefited from repeated evaluation with and feedback from the CFD expert.

Figure 28 shows three iterations through the design process; the final design is shown in Figure 23. Given that CFD experts were unlikely to be familiar with abstract representations of ensemble simulations, the original top-level design for the application adopted a multiple coordinated views approach. The approach has been shown to support visual scaffolding [149]—helping domain experts build from familiar visual representations towards unfamiliar representations. Within this approach, the design then tried to follow, left-to-right, an Overview-first, Filter, Details-on-Demand paradigm (Figure 28 top and middle). Multiple cycles with the domain expert made it clear that, linked-views or not, their analysis always started with the finger structures, i.e., the details. The Details view was also the interface area where the domain expert spent most time. As in an Overview-first paradigm, subsequent analysis steps switched repeatedly between details, context, and overview.

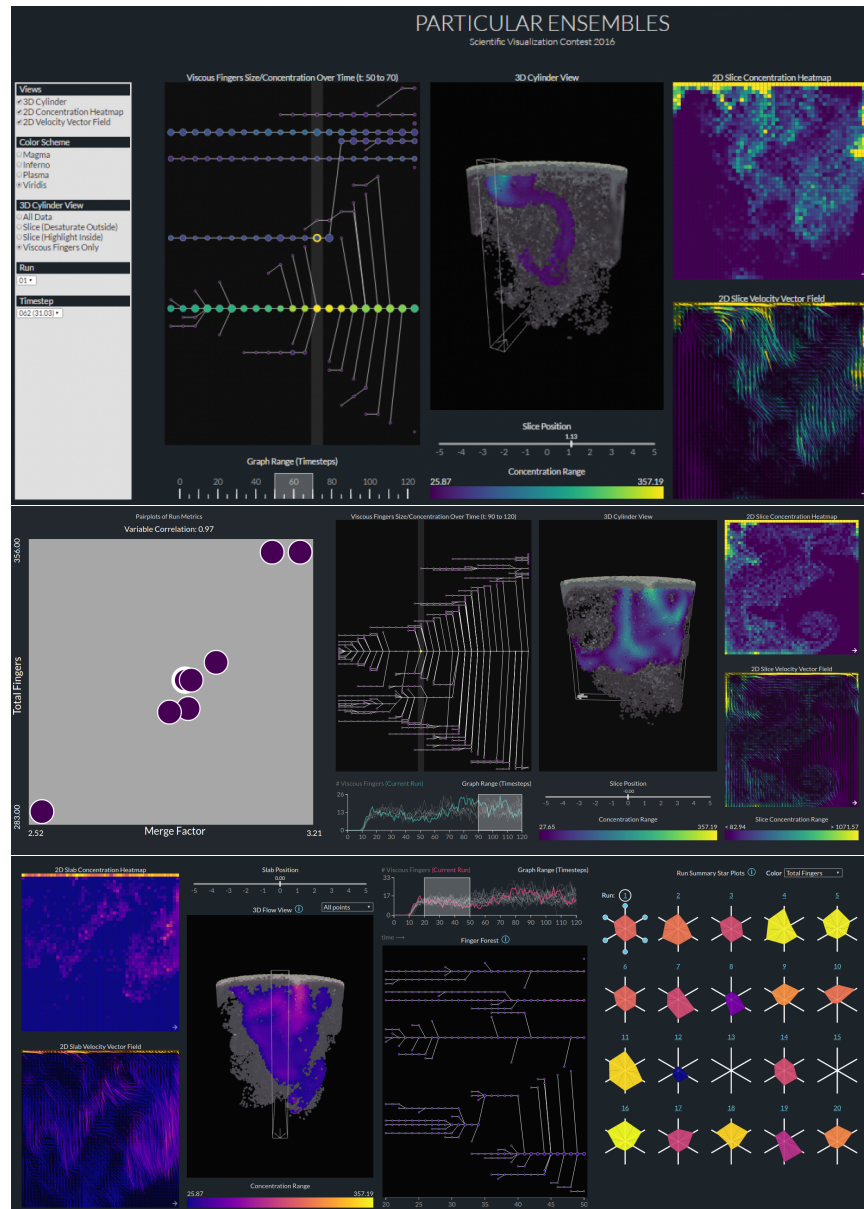


Figure 28. Three snapshots through the design process. Top: Filter/Context-First, Details-Last. Middle: Overview-First, Zoom and Filter/Context, Details-on-Demand. Bottom: Details-first, Show Context, Emphasized Overview (through bold, eye-catching colors). The final design (de-emphasized overview) is shown in Figure 23.

Following a workflow decomposition along scientific workflow theory (5.3.1.3), a Details-first design emerged (Figure 28, bottom), which, unsurprisingly, turned out to be successful. A last attempt to emphasize the overview through an eye-catching color-scheme (Figure 28, bottom, Kiviat panel) still failed to produce a single expert workflow that would lead with the overview, when evaluated with a small group of CFD researchers. In the final design (Figure 23), the color scheme for the overview is de-emphasized, completing the “Details-first, show context, overview last” model instantiation.

In this instantiation, the detail, context, and overview are tied together through brushing, linking and filtering. Specific viscous fingers can be highlighted in the 3D Flow View by selecting a node in the Forest Tree. If the selected viscous finger is from a different timestep than those currently displayed, the corresponding timestep is loaded into the 3D Flow View, and the selected finger is highlighted. Selecting a finger from the Finger Forest also updates the 2D views by automatically moving the slab over the selected finger.

To implement this web-based instantiation, we used the D3 and Three.js JavaScript libraries. The data provided by the contest website was downloaded as individual simulation packs, at its lower 250K resolution due to the size limitation of the web-based platform. However, the web-based platform allowed us to prototype between design iterations quickly and to provide our collaborators (JK) with access to results (cross-platform, no installation required, etc.). We did not process and visualize every resolution of the data due to limitations in the amount of data a web-browser can import and render while still maintaining semi-realtime interaction rates. These limitations are a product of the environment we chose for development and, potentially, would not be a issue in a higher-performance setting. The

analyses reported below were performed at interactive rates (2 fps under Google Chrome, Windows 10, 8GB, Intel i7-3537u @ 2.9 GHz).

5.6 Evaluation

As noted in Chapter 1, a model or theory can be acceptably supported by as little as one to a few concrete examples coming from the experience of one to a few authors [25; 26; 27; 22]. In this work, in addition to the constructive example, we present as supporting evidence two scenarios performed by our CFD expert collaborator, using the model instantiation. We further report instantiation feedback from senior CFD researchers, and theoretical model feedback from the CFD community. The evaluation is rounded by considering further evidence from reports in the visualization literature.

The scenarios below were completed online through web-based exploration of a total volume of data approaching half a billion seven-dimensional data points. The two analyses were conducted by the domain expert using a 18 panel tiled display wall at 21.9 feet by 6.6 feet and 6144 by 2304 pixels; the application used the full height and 2/3 length of the tiled display. The visualization researchers took detailed notes. The usage of interface components (detail, context, overview) was noted based on both the expert's discourse and the physical motion cues as the expert walked from one interface area to another. The observed wall-display usage was consistent with the expert's observed interface usage on a regular display.

5.6.1 Domain Expert Scenarios

5.6.1.1 Exploring Finger Formation

In a first phase, we used the system repeatedly, over several weeks, to identify, define and refine the finger structures, and to derive the relevant characteristics used to generate the context and sum-

marization overview. In this second phase, we investigate viscous finger formation throughout the first simulation run in the ensemble. After loading the run, the investigation begins with the Detail panel, where we notice the appearance of several fingers around timestep 25. Moving the 2D slab from side-to-side, we examine the salt concentration and velocity in detail. A large finger (Figure 25) catches our interest: it appears larger and with higher concentration than others.

To get a better sense of the spatial and temporal context of this finger snapshot, we rotate the 3D Flow cylinder to center the finger in the slab, and also examine the temporal context panel. We notice a downward spike in the finger count between timesteps 20 and 40 in the Time Chart, so we center the *Finger Forest* over that range and advance the 3D Flow View to timestep 25.

We suspect that the decrease in fingers resulted from a few of the fingers merging together to form larger ones, so our analysis moves back from the 3D Viewer to the Finger Forest. As suspected, we notice that many of the nodes between timestep 20 and 40 have merged to form larger nodes. For example (Figure 26), fingers 7 and 21 merge into a single larger node at timestep 22, shown close to the top of the view.

Another spike in the node count is around timestep 95, so we change the range of the Finger Forest to center around that timestep. We again notice that many of the smaller fingers begin to merge into one into one much larger finger by timestep 100. However, we also notice that many smaller fingers begin to form close to timestep 105, which appears to have the most new fingers. Sure enough, we observe in the 3D Flow View that while many of the fingers began to merge at the bottom of the cylinder, many new fingers began to form near the salt source.

5.6.1.2 Similar Simulation Analysis

In the second study we investigate two similar simulation runs. Again, we use the application URL to load the first run and visit the detail finger structures. Noticing low finger formation between timesteps 40 and 60, we center the context Finger Forest using the Time Chart. We observe that timestep 44 appears to have the highest finger formation in the selected range, so we next load that timestep into the 3D Flow View. We hypothesize that the decrease in finger count was caused by large fingers breaking apart to spawn smaller structures. Sure enough, the 3D Flow View displays larger fingers near the edges of the cylinder, with smaller fingers above and below. The Finger Forest further validates this inference, showing us that the majority of the structures in the range have both average high point counts and high concentrations.

Intuitively, we suspect that simulations with similar global properties might exhibit similar merge behavior, so we move to the diagrams in the overview panel. We begin to investigate the other simulations by hovering over the individual axes of the first Kiviat diagram. Starting at 12 o'clock and rotating clockwise, we observe the six computed summary statistics for average finger velocity, average finger density, total finger, fingers per timestep, merge factor, dissipation factor, average finger concentration, and average finger point concentration, respectively. The interface allows us to change the colormap of the diagrams based on one of the derived statistics. Changing the map to "Merge Factor," we observe the color and shape of runs 1, 3, 6 and 14 are similar (Figure 27), indicating that their runs are similar in regards to their derived values.

We decide to investigate simulation 14 in more detail, so we select its Kiviat diagram to load the data into the other views. To our surprise, we notice that the merge tree of run 14 differs from the

previous run, over a similar range. Moving back to the 3D Flow View, we notice significantly more finger structures than in run 1, many of which are smaller in size and still forming at the top of the cylinder. These differences indicate that despite having similar global properties, the fingers of run 14 and run 1 do not follow the same structural formation and evolution.

5.6.2 Domain Expert Feedback

5.6.2.1 Instantiation Expert Feedback

We have collected feedback on this instantiation from two CFD researchers who worked directly with the online system, and two small groups of researchers (5 to 6 participants) who were given demonstrations of the work. One of the groups specialized in advanced computing at a national research laboratory, and included two CFD researchers; the second group consisted of three domain experts and two visualization researchers, as part of the SciVis Contest 2016 [120]. The feedback, reported below, was enthusiastic.

The first CFD researcher (collaborator JK) exclaimed repeatedly “I want this!” (for exploring supersonic and hypersonic flows and turbulent combustion), in particular with respect to the Details-first and temporal context exploration capabilities of the system. He noted that “Oftentimes in CFD we are *details first* because we are already familiar with the simulation and wish to investigate specific features in the data” and then: “When I say *details first* I mean that we look at specific regions or quantities. We are often interested in specific things happening at specific locations or specific times. A [summarization] overview without physical context lacks specificity and therefore is hard to extract meaning from, so we often perform [such] overviews at a later stage.” “It is often very impractical to create an overview of the data as well. Seeing many, or all variables at many (or even some) times is extremely costly in real

world datasets – for example, 10.5TB.” The expert noted that “This type of visualization can be used to investigate underlying physics of the temporal evolution of features of interest. It has applications to a wide range of CFD problems, notably vortex pairing and turbulent mixing.”

The second CFD researcher is a senior investigator who studies computationally turbulence in the aerodynamics of aircraft. His research involves running multiple dynamic simulations with soft-knowledge spatial features. He noted that standard CFD visualization systems (Paraview [150] or VisIt [151]) are frequently employed in a typical CFD workflow to identify simple areas of interest (“Details-first”). Sometimes those features are then used offline to summarize multiple outcomes. However, that summarization is usually in the context of simulations that can be “easily summarized in terms of mean and standard deviation values while discarding lower-level features.” In contrast, our instantiation “enabled analysis at multiple scales,” allowed repeated refining of soft-knowledge features “within their original spatial setting” and the fluid reuse of those “spatially-derived characteristics to summarize multiple outcomes,” well beyond state-of-the-art capabilities. The researcher was keen to have a similar system for his work.

Similar supportive feedback was collected from the larger groups. CFD experts were particularly excited about the smooth coupling of spatial feature characteristics to the temporal context (“extremely intuitive”) and to the summarized overview. The spatial-feature based summarization was “more powerful than anything else [they] had seen.” As in the reports above, group members spent most time operating in the finger detail space, where they were “immediately able to extract meaning to see the formation of fingers,” and used the temporal context and overview mainly for navigation in the finger space. They expressed repeatedly interest in similar analysis tools for their research projects, which also

study features based on soft-knowledge (“[the feature is] hard to define, but if you see it, you recognize it immediately”). Last, we quote feedback from the SciVis Contest contribution [152]: “extremely impressive due to the very well thought-out visualization design;” “clearly superior in visualization design,” “very good and well-crafted,” “in particular, the presentation of the ensemble [...], as well as the layout and linking of all views to facilitate interactive exploration, by far exceed all other submissions.” This feedback attests to the value of our instantiation as a powerful tool for CFD analysis.

5.6.2.2 Theoretical Model Feedback

Our theoretical model sparked equal interest in the CFD community. After clarifying the visualization terminology, our CFD collaborator (JK) engaged in numerous background readings and conversations with other domain experts. Particularly intrigued by the “overview” concept, he set out to find examples of overview usage in standard CFD visual workflows, as employed by a group of nine CFD researchers: two doctoral researchers who use routinely Paraview, an industry researcher and two doctoral researchers who use routinely VisIt, one doctoral researcher who uses routinely ANSYS [153], and a postdoctoral researcher and two senior researchers who are familiar with a variety of platforms. Through short discussions and observations, he sought to establish what software they use, what kind of plots they make, how do they use them, what is the first thing they do, and where, if at all, they use spatial or summarization overviews. He found out that no expert used spatial overviews in their everyday work. Summarization overviews were used, when necessary, last. He then compacted his findings in a common workflow description, best described as: 1) Details first (narrow down what is present); 2) Create filters, expressions, statistics within context; 3) Create a summarization overview of features (describe behavior of features as a whole over entire dataset); 4) Find something of interest then return

to 1) Details and Repeat. In the group’s assessment, much of this workflow stems from the fact that, very similar to the finger instantiation, a number of the physical phenomena they are investigating do not have concrete definitions. These CFD phenomena (e.g., turbulence or reattachment length) typically require a skilled user in order to be visually identified, separated, and investigated. As a result, it becomes difficult to draw conclusions from an overview first, when they “do not know exactly what is present in the data.”

5.7 Discussion

5.7.1 Model Summary

This work is not a general critique of the “Overview first” mantra, but of its sometimes inappropriate application, without careful consideration of user and data workflows. At the same time, while instantiations of our alternative model are particularly common in flow visualization, they are in no way specific to the CFD domain: “details-first” approaches also exist, anecdotally, in biology [154] and in journalism [155].

The alternative “Details-first, show context, overview last” model we advocate supports situations where the main user activities are oriented along (spatial) feature analysis. The model specifically applies to situations where the features are defined through soft-knowledge on the user side, and those features drive both the relevant context for the exploration process and the calculation of the summarization overview.

From a wider analytical perspective, the model applies to domain expert workflows that start with an in-depth exploration of one model or simulation, then seek to extrapolate or generalize the findings to a collection of models. In such workflows, including in forensic analysis, users may wish to start with

the features of interest, in particular when those features are ill-defined and need repeated refinement. The relevance of user-driven refinement in our model is in agreement with Doleisch et al.’s observation: “for interactive analysis, in many cases, the *question of what actually is (or is not) considered to be a feature refers back to the user*: depending on what parts of the data the user (at an instance of time) is most interested in, features are specified accordingly.” [18]. Our model enhances this observation and frames it in a “details-first” paradigm.

Formally, our model further emphasizes the importance of providing the spatial and temporal context of features when they have an inherent spatial structure (3D or Cartesian coordinates). This model aspect is also in agreement with observations in the literature: “[Feature localization] is usually *provided in the context of simulation data*, that has some spatial context.” [18], and with feedback from our CFD collaborator (JK: “CFD/ensemble features are not meaningful outside of their context.”).

Our model instantiation shows how a computational back-end can help identify and track features over space and time, and use those details to automatically filter the spatial and the temporal context. The “show context” step of the model has the triple benefit of 1) helping anchor the features in space and time; 2) reducing visual clutter by controlling complexity of the visualization; and 3) improving rendering times for large scale datasets, in particular in online, platform-agnostic, web-based environments.

Last, this model extends the use of spatial details into the calculation of summarization overviews. In our model instantiation, extracted spatial features and calculations over those features are used to summarize and compare simulation ensembles.

5.7.2 Relationship to Other Models and Theories

Similar to the van Ham and Perer approach [7], the Details-first model signals a set of limitations of the Shneiderman mantra [6]. In contrast to the van Ham and Perer mantra, the present model emphasizes the importance of Details (not Search for a particular item) for a class of problems, and the relevance of user input in specifying and refining those details. In a further departure from the van Ham and Perer approach, where overviews are circumvented as being both impractical and not relevant under specific circumstances, our model handles situations where summarization overviews are necessary. In particular, our model extends and provides a frame for the use of details into the calculation of such overviews.

The Details-first model further relates to Chen et al.'s Information Theory framework [8]. Our model encompasses their anecdotal observation that, in particular in flow visualization, the Shneiderman mantra can be suboptimal when the user is already intimately familiar with the overview. Beyond agreeing with their observation, this work highlights: 1) the tight connection that exists between the user “knowledge in the head” [15] and the very definition of spatial features; and 2) how those details can propagate into the construction of filtering operations, and then into the construction of summarization overviews.

The model's “overview last” aspect may also be related to the principle of visual scaffolding [149], captured by the domain experts' typical resistance to unfamiliar visual encodings.

5.8 Conclusion

In conclusion, this work introduces and documents an alternative “Details-first, show context, overview last” approach to visualization design. The approach supports situations where the user ac-

tivities are oriented along (spatial) feature analysis. This work further highlights the tight connection that can exist between user input and the definition of spatial features, and then how those details can propagate into the construction of filtering operations, and then into the construction of overviews. A model instantiation demonstrates the effectiveness of this approach with an online web-based exploration of a total volume of data approaching half a billion seven-dimensional data points. The approach is supported by endorsements from CFD domain experts.

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 Discussion

This dissertation first investigates the use of two well-established paradigms (Overview-first and Search-first) in the context of developing design strategies for S&E spatial data problems. Across three S&E domains, this work demonstrates how the Overview-first and Search-first paradigms can facilitate analysis by directly applying a paradigm to the problem at hand (Chapter 2), combining two or more of the paradigms together to form a hybrid-design approach (Chapter 3), or abstracting the spatial data to work within these known design space of a paradigm (Chapter 4).

However, overviews (that are not used merely as context) are conspicuously rare in the scientific visualization literature. This observation is not surprising. In 2016, Chen et al. [8] were the first to note that in “many scenarios, we often observe that an experienced viewer may find [overview first and details on demand] frustrating, as the viewer knows exactly where the interesting part of a detailed representation is. For example, in flow simulation, scientists work on the same problem for months.” Their anecdotal observation is reflected in a vast number of additional works in scientific visualization that support explicitly spatial feature exploration, and display the rest of the information primarily for context (e.g., [18; 137; 136; 11; 156; 157; 158; 159]).

In these situations, the main expert activities are centered around the analysis of spatial features that are defined through soft-knowledge on the user side. As of result of the imprecise nature of the features

they wish to analyze, these details of interest can neither be searched against or aggregated prior the expert's specification; in short, neither the Overview-first nor Search-first paradigms can be used to develop design strategies when user-specified details are the primary component driving the analysis. Therefore, in a fourth S&E visualization collaboration involving spatial data (Chapter 5), we deviate from the common notion that analysis should start with some form of an overview or a search operation and present an alternative model – Details-first – that establishes user-specified details as the primary component driving the analysis.

Overall, while we have demonstrated the success of these design strategies across four interdisciplinary collaborations involving spatial data, these four corroborating examples of success do not constitute proof that our presented strategies can be extended to work for all other spatial data problems. However, this shortcoming was not unforeseen as it is related to the tight coupling between design strategies and the domain characterizations of spatial data problems. Karl Popper, a 20th-century scientific theorist, best summarized the problem of anecdotal evidence when he stated “it is impossible to prove a methodology on cited experience alone” [160].

Instead, this dissertation suggests that S&E collaborations involving spatial data be approached by first identifying commonalities that exist among successful design strategies. As we have seen throughout this work, one potential approach would be to rely on scientific workflow theory to provide insight into the ethnographic differences between the various problem spaces. This decomposition can lead to design insights such as how multi-views can be utilized to facilitate multiple user workflows. In summary, the adoption of a particular design strategy in these collaborations should take into account the

benefits, limitations, and possible co-existence of each approach, with careful consideration of the data, knowledge base, and workflows of the collaborating experts.

6.2 Conclusion

This dissertation investigates three challenges behind developing design strategies for spatial data problems in S&E collaborations. These challenges include teasing out the underlying requirements of the domain scientist when developing design strategies, evaluating their success, and extrapolating generalizable knowledge from the strategies of successful collaborations. Specifically, this dissertation examines how the data and task abstractions, workflow processes, and user expertise affect the decisions behind the design strategies of four S&E spatial data problems. In three of these collaborations, we investigate the use of the Overview-first and Search-first paradigms in the context of spatial data problems and whether they can be appropriated to fit within the target user workflows. In a fourth collaboration where an overview of the data is neither relevant nor practical, we introduce an alternative strategy that establishes user-specified details as the primary component driving the analysis. This dissertation demonstrates the merits of the deployed design strategies through the development and deployment of four integrated visualization applications and evaluates their successes through case studies (Chapter 2,3 and 5), statistical analysis (Chapter 4), and expert feedback (Chapter 2-5).

Specifically, the contributions of this work include:

- descriptions of the tasks and data associated with specific problems related to spectroscopic analysis of galaxies in observational astronomy, functional mutations in protein families in molecular biology, lymph node metastasis in radiation oncology, and viscous finger evolution in mechanical

engineering. These four analyses characterized the spatial data problems of specific situations and identified the workflows of the experts working to solve them. Through the analysis of each domain problem, this dissertation demonstrated the complexities involved with interdisciplinary research and the necessity of working directly with domain scientists and their data. The resulting strategies demonstrate how the domain characterizations can be used as a basis on which spatial data design strategies are built and serve as a reference for researchers who are working on endeavors which are similarly characterized to the four specific domain situations described in this dissertation;

- several novel visual representations and techniques for spatial data. Throughout the collaborations in observational astronomy (Chapter 2), molecular biology (Chapter 3), and radiation oncology (Chapter 4), this dissertation presented novel visual abstractions for spatial data that were motivated by the established Overview-first and Search-first strategies. In each collaboration, this work demonstrated how these two strategies could be used to facilitate spatial data analysis in the experts' workflows to tackle their specific research problems within each domain situation. In our collaboration with observational astronomers working on spectroscopic analysis (Chapter 2), we demonstrated how combining the query-based, Search-first approach with a novel, pixel-based representation can assist observational astronomers in identifying trends in large collections of spatial observations. Next, this dissertation demonstrated how the Overview-first and Search-first paradigms could be combined to create a hybrid design strategy by incorporating both a 3D structure representation of a target protein and a novel visual abstraction of its closest family members to accommodate both molecular biology and bioinformatics workflows (Chapter 3). Finally, this

work described a strategy for abstracting the structural information of the head and neck lymph node regions to build a spatial measure that captures similarity within a cohort of cancer patients, and how the resulting similarity measure can be used with the Search-first paradigm to rank and query for similar patients based on their spatial correlates;

- a novel alternative design approach, Details-first, to the Overview-first and Search-first design paradigms. Based on a collaboration in mechanical engineering (Chapter 5), we instantiated this new approach for situations where providing the user with an overview is either impractical or infeasible and demonstrated that the workflows in these situations often rely on the experts' *soft knowledge* of their data. In these situations, we showed how the resulting design strategies can be centered around features (details) in situations where providing an overview of the data is neither relevant nor practical to the user; if necessary, this awareness could be presented at the end stages of the analysis in the form of summary statistics;
- the design, implementation, deployment, and evaluations of four integrated systems. We developed complete, end-to-end solutions that include novel visual representations and techniques for spatial data problems. Each of these collaborations was performed in close collaboration with domain experts and detailed how their expertise was invaluable when developing these the strategies behind the presented successes. These collaborations spanned the design process from domain characterization to evaluation using Munzner's nested model of visual design and evaluation [22]. Finally, we demonstrated how the three paradigms – Overview-first, Search-first, and Details-first – can integrated into the experts' workflows to solve real-world problems in four different spatial data domains;

- a discussion of the merits, applicability, and limitations of the design strategies as they were applied to each of the domain problems. At the end of each chapter, we reflected on the successfulness of our designs in the context of the spatial data problem the collaboration aimed to solve. In doing so, we detailed how our designs helped lessen the problems of the expert, potential future applications that our strategies might benefit, and areas in our design that require further research.

In terms of limitations, the different paradigms used in the strategies presented in this work have complementary strengths and limitations. However, because the limitations of overview-based approaches has been well documented outside of this work [7; 28; 12], we will focus primarily on the Details-first model.

First and foremost, the presented Details-first model may not be necessary when the analysis can be conveniently broken into two separate processes; for example, feature detection and simple statistical summarization. This model also does not apply: to situations where overviews are irrelevant (use Search instead, or default to Details-first, no overview); when user prior knowledge is not relevant, when global changes are likely, or when each search starts from scratch (use Overview-first instead); or when the features are well-defined and computable, or not at the very core of the user activity (a variety of other approaches apply, including pure computation).

Next, although multi-views have the potential to relieve workflow-related design constraints, we note that the model principle still applies, in the designer-assigned color scheme, size and location of overviews and context views in the overall design. However, extensions of this paradigm to single-view, reduced screen space settings, may be particularly limiting, considering the complementary benefits of

summarization overviews. A step further, the process of overview summarization itself may miss an unexpected global change.

Finally, this dissertation, by itself, does not fully capture the importance of the layout strategies in designing visualization applications and the interplay shared with the different interrelated processes within the experts' workflows. While we experimented with different layout strategies when instantiating the Details-first approach (Chapter 5) – and to a small extent, in the design of the FixingTIM interface (Chapter 3) – visual layout strategies were not explicitly addressed in the design of the Astroshef framework (Chapter 2) and spatial neighborhood method (Chapter 4).

In terms of future work, our observational evaluation in much of this dissertation draws on a left-to-right, multi-view instantiation, designed and evaluated with small groups of experts, several from the same labs. Such multi-view instantiations take advantage of the complementarity of multiple representations, and also have the potential to facilitate multiple user workflows [149; 113]. In practice, we have not observed domain expert analysis workflows that did not lead with the details view. A formal user study to analyze the likelihood of different mantras would be interesting, although beyond the scope of this chapter. Any such study should take particular care in the participant selection, given the central soft-knowledge aspects of our model, and the limited availability of domain experts.

Additionally, this dissertation only briefly examined strategies for situations where the experts' workflows contain a mixture of spatial and non-spatial data attributes. Therefore, one avenue of potential future research would be to determine how these data integration efforts affect the design choices of the collaboration; for example, adding additional non-spatial features (e.g., age, T-Category, etc.) to the spatial measure presented in Chapter 4. Additionally, this work touched upon the influence of a

workflow decomposition using scientific workflow theory in regards to the layout strategy used in the Detail-first approach (Chapter 5). The anecdotal evidence presented in the instantiation of this approach suggests that the layout design, not just the visual encodings, can play an important role in the success of a design strategy. The role of layout in design strategies can be further expanded to examine its interplay with experts' workflows.

In conclusion, this dissertation examines design strategies for visualization in the context of four problem-driven, spatial data collaborations in S&E domains. Using these collaborations as a vehicle to ground our research, we demonstrated the benefit of visual design strategies for solving real-world, spatial data problems through the development, implementation, deployment, and evaluation of four end-to-end visualization applications. We hope that these analyses may also serve as a starting point for researchers who are interested in or are currently working with similar spatial data problems as those described in this dissertation.

APPENDICES

Appendix A

SUPPLEMENTAL MATERIAL LINKS

The following are the links to various videos detailing the applications found in this dissertation:

Two videos of the Astroshelf application in Chapter 2:

IEEE LDAV 2012 Paper Teaser: <http://visualizlab.org/results/videos/Luciani-2012-PAZ.mp4>

IEEE TVCG 2014 Supplement: <https://vimeo.com/user97518538/review/331163268/335eed5e10>

A video showcasing the FixingTIM application in Chapter 3:

BMC Proceedings 2014 Supplement: <https://vimeo.com/84635382>

A video describing the Details-first model instantiation in Chapter 5:

IEEE SciVis 2018 Supplement: https://www.youtube.com/watch?v=q-KM_oRgxk&feature=youtu.be

Appendix B

PERMISSION FOR REUSE

The following presents written permission from the journal's/publisher's website outlining their copyright policies.

BMC Proceedings

Copyright

- Copyright on any open access article in a journal published by BioMed Central is retained by the author(s).
- Authors grant BioMed Central a [license](#) to publish the article and identify itself as the original publisher.
- Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.
- The [Creative Commons Attribution License 4.0](#) formalizes these and other terms and conditions of publishing articles.

Luciani, T., Wenskovitch, J., Chen, K., Koes, D., Travers, T., and Marai, G.: FixingTIM: Interactive Exploration of Sequence and Structural Data to Identify Functional Mutations in Protein Families. BMC Proc., 2014.

Appendix B (Continued)

IEEE

Full-Text Article

If you are using the entire IEEE copyright owned article, the following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]

Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.

In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

Luciani, T., Cherinka, B., Oliphant, D., Myers, S., Wood-Vasey, W.M. Labrinidis, A., and Marai, G.E.: Large-Scale Overlays and Trends: Visually Mining, Panning and Zooming the Observable Universe. IEEE Trans. Vis. Comp. Graph. (TVCG), pages 1-14, 2014. © IEEE. Reprinted, with permission, from authors.

Luciani, T., Burks, A., Sugiyama, C., Komperda, J., and Marai, G.E.: Details-first, Show Context, Overview last: Supporting Exploration of Viscous Fingers in Large-Scale Ensemble Simulations. IEEE Trans. Vis. C pages 1225–1235, 2018. © IEEE. Reprinted, with permission, from authors.

Appendix C

ADDITIONAL ACKNOWLEDGEMENTS

Chapter 2: Large-Scale Overlays and Trends: Visually Mining, Panning and Zooming the Observable Universe

This research was funded through US National Science Foundation (NSF)-OIA-1028162, NSF-IIS-0952720, and NSF GRFP-2012144824. They gratefully acknowledge the enthusiastic help and unwavering support of our collaborators Jeffrey Newman and Arthur Kosowski: no one could wish for better collaborators. They thank the AEGIS, SKA and LSST astronomy communities, the anonymous reviewers and the Pitt VisLab for feedback and interesting discussions; Rebecca Hachey, Ruhsary Rexit, Boyu Sun, and the cs1699 students for contributions to prototype codes; and the ADMT lab for occasional testing and feedback. Further acknowledgments to Noel Gorelick and Jeremy Brewer, for generously sharing their Google Sky development experience.

Chapter 3: Interactive Exploration of Sequence and Structural Data to Identify Functional Mutations in Protein Families

This work has been supported by grant NSF-IIS-0952720 and by the NSF Graduate Research Fellowship program. Many thanks to Adrian Maries, Xinghua Lu, Lujia Chen and the VisLab group for feedback and useful discussions. We gratefully acknowledge the dataset provided by Drs. Magliery and Sullivan at The Ohio State University for the purposes of the BioVis 2013 Contest

Appendix C (Continued)*Chapter 5: Details-First, Show Context, Overview Last: Supporting Exploration of Viscous Fingers in Large-Scale Ensemble Simulations*

We thank Min Chen (University of Oxford) and Roberto Paoli (UIC and Argonne National Laboratory) for their shared insights, the anonymous reviewers for their useful feedback, the US National Science Foundation (NSF) Graduate Research Fellowship program for supporting Tim, the NSF REU program for supporting Andrew, and the Brazilian Scientific Mobility Program for supporting Cassiano's summer at EVL. We thank Dagstuhl seminar 18041 and seminar 18161 for cross-pollination of ideas. We thank the SciVis 2016 Contest organizers and the San Diego Supercomputing Center for sharing their datasets and for their feedback. We further thank our colleagues at EVL, in the UIC Mashayek lab, and at Argonne National Laboratory for their support and help. This work was supported in part by US federal agencies, through awards NSF CNS-1625941, NSF CAREER IIS-1541277, NIH NCI-R01CA214825, NIH NCI-R01CA2251, and NIH NLM-R01LM012527.

CITED LITERATURE

1. T. Luciani, B. Cherinka, D. Oliphant, S. Myers, A. Wood-Vasey, W.M. Labrinidis, and G. Marai. Large-Scale Overlays and Trends: Visually Mining, Panning and Zooming the Observable Universe. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1–14, 2014.
2. T. Luciani, J. Wenskovitch, K. Chen, D. Koes, T. Travers, and G. Marai. Fixingtim: interactive exploration of sequence and structural data to identify functional mutations in protein families. *BMC Proc.*, 2014.
3. 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. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1225–1235, 2018.
4. A. J. G. Hey, S. Tansley, and K. M. Tolle. *Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
5. T. Munzner. *Visualization analysis and design*. AK Peters visualization series. CRC Press, 2014.
6. B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. IEEE Symp. Vis. Lang.*, pp. 336–. 1996.
7. F. van Ham and A. Perer. “Search, Show Context, Expand on Demand”: Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 953–960, 2009.
8. M. Chen, M. Feixas, I. Viola, A. Bardera, H. Shen, and M. Sbert, eds. *Visualization and Information Theory*. A K Peters/CRC Press, 2016.
9. A. Maries, T. Luciani, P. H. Pisciueneri, M. B. Nik, S. L. Yilmaz, P. Givi, and G. E. Marai. A Clustering Method for Identifying Regions of Interest in Turbulent Combustion Tensor Fields, pp. 323–338. Springer, 2015.
10. M. Monfort, T. Luciani, J. Komperda, B. Ziebart, F. Mashayek, and G. E. Marai. A Deep Learning Approach to Identifying Shock Locations in Turbulent Combustion Tensor Fields, pp. 375–392. Springer, 2017.

11. K. Potter, A. Wilson, B. P-t, C. Williams, C. Doutriaux, P. V, and C. Johnson. Ensemble-Vis: A Framework for the Statistical Visualization of Ensemble Data. *In IEEE Proc. Work. Know. Disc. Climate Data: Pred., Extr.*, pp. 233–240. 2009.
12. K. Hornbaek and M. Hertzum. The notion of overview in information visualization. *Int. J. Human-Comp. Studies*, pp. 509–525, 2011.
13. R. Spence. *Information Visualization: Design for Interaction*. Prentice-Hall, Inc., second edn., 2007.
14. E. Tufte. *Envisioning Information*. Graphics Press, 1990.
15. D. A. Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, Inc., 2013.
16. S. Greene, G. Marchionini, C. Plaisant, and B. Shneiderman. Previews and Overviews in Digital Libraries: Designing Surrogates to Support Visual Information Seeking. *J. Am. Soc. Inf. Sci. (JASIST)*, pp. 380–393, 2000.
17. S. K. Card, J. D. Mackinlay, and B. Shneiderman, eds. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publ. Inc., 1999.
18. H. Doleisch, M. Gasser, and H. Hauser. Interactive Feature Specification for Focus+Context Visualization of Complex Simulation Data. *In VisSym*, pp. 239–248. 2003.
19. G. W. Furnas. A Fisheye Follow-up: Further Reflections on Focus + Context. *In Proc. SIGCHI Conf. Human Fact. Comp. Sys.*, pp. 999–1008. 2006.
20. N. Russell, W. M. van der Aalst, and A. H. M. ter Hofstede. *Workflow Patterns: The Definitive Guide*. MIT Press, 2016.
21. V. Curcin and M. Ghanem. Scientific workflow systems - can one size fit all? *In Cairo Int. Biomed. Eng. Conf.*, pp. 1–9. 2009.
22. T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 921–928, 2009.
23. IEEE Sci. Vis. Topics and Paper Types. <https://ieevis.org/year/2018/info/call-participation/scivis-paper-types>, 2018.

24. IEEE Information Visualizations Topics and Paper Types. <https://ieevis.org/year/2018/info/call-participation/infovis-paper-types>, 2018.
25. D. Lloyd and J. Dykes. Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2498–2507, 2011.
26. A. J. Pretorius and J. J. Van Wijk. What does the user want to see?: What do the data want to be? *In Proc. IEEE Symp. Info. Vis. (InfoVis)*, pp. 153–166. 2009.
27. M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2431–2440, 2012.
28. B. Craft and P. Cairns. Beyond Guidelines: What Can We Learn from the Visual Information Seeking Mantra? *In Proc. IEEE Symp. Info. Vis. (InfoVis)*, pp. 110–118. 2005.
29. T. Luciani, B. Cherinka, S. Myers, B. Sun, W. Wood-Vassey, A. Labrinidis, and G. Marai. Panning and Zooming the Observable Universe with Prefix-Matching Indices and Pixel-Based Overlays. *IEEE Large-Scale Data Anal. Vis. Symp. (LDAV)*, pp. 1–8, 2012.
30. M. Davis, P. Guhathakurta, N. P. Konidaris, J. A. Newman, M. L. N. Ashby, *et al.* The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets. *Astrophysical J. Letters*, pp. L1–L6, 2007.
31. R. H. Becker, R. L. White, and D. J. Helfand. The VLA’s FIRST Survey. *In* D. R. Crabtree, R. J. Hanisch, and J. Barnes, eds., *Astronomical Data Analysis Software and Systems III*, pp. 165–. 1994.
32. Google Sky. <http://www.us-vo.org/>.
33. National Virtual Observatory. <http://www.google.com/sky/>.
34. WorldWide Telescope. <http://www.worldwidetelescope.org/>.
35. Sloan Digital Sky Survey. <https://skyserver.sdss.org>.
36. NASA/IPAC Infrared Science Archive. <https://irsa.ipac.caltech.edu>.
37. SIMBAD Astronomical Database. <http://simbad.u-strasbg.fr/simbad/>.

38. C. Y. Ip and A. Varshney. Saliency-Assisted Navigation of Very Large Landscape Images. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1737–1746, 2011.
39. J. Kopf, M. Uyttendaele, O. Deussen, and M. Cohen. Capturing and viewing gigapixel images. *Proc. SIGGRAPH Conf. Comp. Graph.*, 2007.
40. R. Machiraju, J. E. Fowler, D. Thompson, B. Soni, and W. Schroeder. *EVITA — Efficient Visualization and Interrogation of Tera-Scale Data*, pp. 257–279. Springer, 2001.
41. G. W. Furnas and B. B. Bederson. Space-scale diagrams: understanding multiscale interfaces. *In Proc. SIGCHI Conf. Human Factors Comp. Sys.*, pp. 234–241. 1995.
42. B. B. Bederson and J. D. Hollan. Pad++: a zooming graphical interface for exploring alternate interface physics. *In Proc. 7th Ann. ACM Symp. User Inter. Soft. Tech.*, pp. 17–26. 1994.
43. G. Furnas. Generalized fisheye views. *Procs. SIGCHI Conf. Human Factors Comp. Sys.*, pp. 16–23, 1986.
44. H. Lieberman. Powers of Ten Thousand: Navigating in Large Information Spaces. *In Proc. 7th ACM Symp. User Int. Soft. Tech.*, pp. 15–16. 1994.
45. C. Page. Indexing the Sky. <https://www.star.le.ac.uk/cgp/ag/skyindex>.
46. K. Chodorow and M. Dirolf. *MongoDB: The Definitive Guide*. O'Reilly Media, 2010.
47. N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J. Fekete. Zame: Interactive large-scale graph visualization. *In Proc. IEEE Pacific Vis. Symp. (PacificVIS)*, pp. 215–222. 2008.
48. C. Appert, O. Chapuis, and E. Pietriga. High-precision magnification lenses. *In Proc. SIGCHI Conf. Human Factors in Comp. Sys.*, pp. 273–282. 2010.
49. W. Javed, S. Ghani, and N. Elmqvist. Gravnav: using a gravity model for multi-scale navigation. *In Proc. Int. Working Conf. Advanced Visual Interfaces*, pp. 217–224. 2012.
50. W. Javed, S. Ghani, and N. Elmqvist. Polyzoom: multiscale and multifocus exploration in 2d visual spaces. *Proc. SIGCHI Conf. Human Factors Comp. Sys.*, 2012.
51. C. Weigle, W. Emigh, G. Liu, R. Ii, J. Enns, and C. Healey. Oriented Sliver Textures: A Technique for Local Value Estimation of Multiple Scalar Fields, 2000.

52. D. Fisher. Hotmap: Looking at geographical attention. *In Proc. IEEE Symp. Info. Vis. (InfoVis)*, 2007.
53. A. Bokinsky. *Multivariate data visualization with data-driven spots*. Ph.D. thesis, Univ. n. Carolina, Chapel Hill, 2003.
54. P. McLachlan, T. Munzner, E. Koutsofios, and S. North. LiveRAC: Interactive Visual Exploration of System Management Time-series Data. *In Proc. SIGCHI Conf. Human Factors in Comp. Sys.*, pp. 1483–1492. 2008.
55. P. Craig and J. B. Kennedy. Coordinated graph and scatter-plot views for the visual exploration of microarray time-series data. *In Proc. IEEE Symp. Info. Vis. (InfoVis)*, pp. 173–180. 2003.
56. D. Albers, C. N. Dewey, and M. Gleicher. Sequence Surveyor: Leveraging Overview for Scalable Genomic Alignment Visualization. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2392–2401, 2011.
57. J. J. Van Wijk and E. R. Van Selow. Cluster and Calendar Based Visualization of Time Series Data. *In Proc. IEEE Symp. Info. Vis. (InfoVis)*, pp. 4–. 1999.
58. D. Keim. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 59–78, 2000.
59. J. C. Jacob, D. S. Katz, G. B. Berriman, J. C. Good, A. C. Laity, *et al.* Montage; a grid portal and software toolkit for science; grade astronomical image mosaicking. *Int. J. Comp. Sci. Eng.*, pp. 73–87, 2009.
60. E. W. Greisen and M. R. Calabretta. Representations of world coordinates in FITS. *Astronomy & Astrophysics*, pp. 1061–1075, 2002.
61. M. R. Calabretta and E. W. Greisen. Representations of celestial coordinates in FITS. *Astronomy & Astrophysics*, pp. 1077–1122, 2002.
62. X. Fan. Evolution of high-redshift quasars. *New Astronomy Rev.*, pp. 665–671, 2006.
63. C. G. Kotanyi and R. D. Ekers. Radio galaxies with dust lanes. *Astronomy & Astrophysics*, 1979.
64. S. S. Shabala, Y.-S. Ting, S. Kaviraj, C. Lintott, R. M. Crockett, *et al.* Galaxy Zoo: Dust lane early-type galaxies are tracers of recent, gas-rich minor mergers. *ArXiv*, 2011.

65. C. Moellenhoff, E. Hummel, and R. Bender. Optical and radio morphology of elliptical dust-lane galaxies - comparison between ccd images and vla maps. *Astronomy & Astrophysics*, pp. 35–48, 1992.
66. W. Zhong, G. Altum, R. Harrison, P. Tai, and Y. Pan. Mining protein sequence motifs representing common 3d structures. *In Proc. IEEE Comp. Sys. Bioinform. Conf. - Workshops*, pp. 215–216. 2005.
67. N. Chen, T. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, *et al.* WormBase: A Comprehensive Data Resource for *Caenorhabditis* Biology and Genomics. *Nucleic Acids Res.*, pp. 383–389, 2005.
68. S. Montgomery, T. Astakhova, M. Bilenky, E. Birney, T. Fu, *et al.* Sockeye: A 3D Environment for Comparative Genomics. *Genome Res.*, pp. 956–962, 2004.
69. W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, *et al.* Human Genome Browser, UCSC. *Genome Res.*, pp. 996–1006, 2002.
70. W. C. Ray, R. W. Rumpf, B. Sullivan, N. Callahan, T. Magliery, *et al.* Understanding the sequence requirements of protein families: insights from the biovis 2013 contests. *BMC Proc.*, pp. S1–, 2014.
71. B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, pp. e118–, 2011.
72. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Ann. Rev. Biophysics*, p. 289316, 2008.
73. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, *et al.* Protein Data Bank. *Nucleic Acids Res.*, pp. 235–242, 2000.
74. The Uniprot Consortium. Update on activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, pp. D43–D47, 2013.
75. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. Madden. BLAST+: Architecture and Applications. *BMC Bioinform.*, 2009.
76. N. Eswar, M. Marti-Renom, B. Webb, M. Madhusudhan, D. Eramian, *et al.* Comparative protein structure modeling with modeller. *In Current Protocols in Bioinformatics*. John Wiley & Sons Inc., 2006.

77. Schrödinger, LLC. The PyMOL molecular graphics system, v. 1.3r1, 2010.
78. C. A. Brewer. Color Brewer: A web tool for selecting color or maps, 2009.
79. U. Pieper, N. Eswar, H. Braberg, M. Madhusudhan, F. Davis, *et al.* MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res. - Database issue*, pp. D347–354, 2009.
80. A. Waterhouse, J. Procter, D. Martin, M. Clamp, and G. Barton. Jalview Version 2 a multiple sequence alignment editor and analysis workbench, 2009.
81. National Cancer Institute. Oral Complications of Chemotherapy and Head/Neck Radiation. URL <https://www.cancer.gov/about-cancer/treatment/side-effects/mouth-thr>
82. D. Brandizzi, M. Gandolfo, M. Lucia Velazco, R. Luis Cabrini, and H. Lanfranchi. Clinical features and evolution of oral cancer: A study of 274 cases in Buenos Aires, Argentina. *Medicina oral, patologica oral y cirugia bucal*, pp. E544–E548, 2008.
83. B. W. Stewart, H. Greim, D. Shuker, and T. Kauppinen. Defence of IARC Monographs. *Lancet*, pp. 1300–, 2003.
84. L. Zhang, A. S Garden, J. Lo, K. Kian Ang, and *et al.* Multiple regions-of-interest analysis of setup uncertainties for head-and-neck cancer radiotherapy. *Int. J. Rad. Onc., Bio., & Phys.*, pp. 1559–1569, 2006.
85. A. Nakata, K. Tateoka, K. Fujimoto, Y. Saito, and *et al.* The Reproducibility of Patient Setup for Head and Neck Cancers Treated with Image-Guided and Intensity-Modulated Radiation Therapies Using Thermoplastic Immobilization Device. *Int. J. Med, Phys., Clin. Eng. & Rad. Onc.*, pp. 117–124, 2013.
86. J. Timar, O. Csuka, E. Remenr, G. Rpssy, and M. Ksler. Progression of head and neck squamous cell cancer. *Cancer Metastasis Rev.*, pp. 107–127, 2005.
87. B. Cartmill, P. Cornwell, E. Ward, W. Davidson, R. Nund, *et al.* Emerging understanding of dosimetric factors impacting on dysphagia and nutrition following radiotherapy for oropharyngeal cancer. *Head & Neck*, pp. 1211–1219, 2013.
88. V. Gregoire, K. Ang, W. Budach, C. Grau, *et al.* Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiotherapy & Oncology*, pp. 172–181, 2014.

89. W. Widanagamaachchi, P. Klacansky, H. Kolla, A. Bhagatwala, J. Chen, V. Pascucci, and P. T. Bremer. Tracking features in embedded surfaces: Understanding extinction in turbulent combustion. *In Proc. IEEE 5th Symp. on Large Data Anal. Vis. (LDAV)*, pp. 9–16. 2015.
90. 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. J.*, pp. 1–22, 2014.
91. C.-C. Teng, L. Shapiro, I. J. Kalet, C. Rutter, and R. Nurani. Head and neck cancer patient similarity based on anatomical structural geometry. *In Proc. IEEE Int. Symp. Biomed. Imag.*, pp. 1140–1143. 2007.
92. B. Yener, C. Gunduz, and S. H. Gultekin. The cell graphs of cancer. *Bioinform.*, pp. i145–i151, 2004.
93. M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE Reviews Biomed. Eng.*, pp. 147–171, 2009.
94. E. G. M. Petrakis and C. Faloutsos. Similarity searching in medical image databases. *IEEE Trans. Knowl. Data Eng. (TKDE)*, pp. 435–447, 1997.
95. 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 Digital Imaging*, pp. 1025–1039, 2013.
96. T. T. Tanimoto. *An elementary mathematical theory of classification and prediction*. Tin. Business Machines Corp., 1958.
97. D. Bajusz, A. Rácz, and K. Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics*, pp. 20–, 2015.
98. S. Klinger and J. Austin. Weighted superstructures for chemical similarity searching. *In Proc. 9th Joint Conf. Info. Sci.* 2006.
99. X. Yan, P. S. Yu, and J. Han. Substructure similarity search in graph databases. *In Proc. ACM SIGMOD Int. Conf. Mana. Data*, pp. 766–777. 2005.
100. A. Tomovi, P. Janii, and V. Keelj. n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Comp. Methods & Programs Biomed.*, pp. 137–153, 2006.

101. B. King. Step-wise clustering procedures. *J. Amer. Stat. Assoc.*, pp. 86–101, 1967.
102. F. Murtagh. A Survey of Recent Advances in Hierarchical Clustering Algorithms. *Comp. J.*, pp. 354–359, 1983.
103. A. Dong, N. Honnorat, B. Gaonkar, and C. Davatzikos. CHIMERA: Clustering of Heterogeneous Disease Effects via Distribution Matching of Imaging Patterns. *IEEE Trans. on Med. Im. (TMI)*, pp. 612–621, 2016.
104. J. Bruse, M. Zuluaga, A. Khushnood, K. Mcleod, *et al.* Image Data: Metrics Analysis for Hierarchical Clustering Applied to Healthy and Pathological Aortic Arches. *IEEE Trans. Biomed. Eng. (TBME)*, pp. 2373–2383, 2017.
105. P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Passessing the accuracy of prediction algorithms for classification: An overview. *Oxford Bioinform.*, pp. 412–24, 2000.
106. R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bulletin*, pp. 1409–1438, 1958.
107. M. Inc. Matlab and statistics toolbox release 2018a.
108. O. Z. Maimon and L. Rokach. *Data mining and knowledge discovery handbook*. Springer, 2010.
109. C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge Univ. Press, 2018.
110. W. Rand. Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc.*, pp. 846–850, 1971.
111. R. A. Fisher. *Statistical methods for research workers*. Kalpaz, 2017.
112. M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2301–2309, 2011.
113. G. Marai, C. Ma, A. Burks, F. Pellolio, G. Canahuate, D. Vock, A. Mohamed, and C. Fuller. Precision Risk Analysis of Cancer Therapy with Interactive Nomograms and Survival Plots. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1–13, 2018.
114. T. Rancati, M. Schwarz, A. Allen, F. Feng, A. Popovtzer, *et al.* Radiation Dose-Volume Effects in the Larynx and Pharynx. *Int. J. Rad. Onc., Bio, & Phys.*, pp. S64–S69, 2010.

115. C. R. Spencer, H. A. Gay, B. H. Haughey, *et al.* Eliminating radiotherapy to the contralateral retropharyngeal and high level ii lymph nodes in head and neck squamous cell carcinoma is safe and improves quality of life. *Cancer*, pp. 3994–4002, 2014.
116. T. Dale, K. Hutcheson, A. Mohamed, J. S. Lewin, *et al.* Beyond mean pharyngeal constrictor dose for beam path toxicity in non-target swallowing muscles: Dosevolume correlates of chronic radiation-associated dysphagia (rad) after oropharyngeal intensity modulated radiotherapy. *Radiotherapy & Oncology*, pp. 304–314, 2016.
117. M. Kamal, A. S. Mohamed, S. Volpe, and et al. Radiotherapy dosevolume parameters predict videofluoroscopy-detected dysphagia per DIGEST after IMRT for oropharyngeal cancer: Results of a prospective registry. *Radiotherapy & Oncology*, pp. 442–451, 2018.
118. A. S. R. Mohamed, B. P. Hobbs, and K. A. e. a. Hutcheson. Dose-volume correlates of mandibular osteoradionecrosis in Oropharynx cancer patients receiving intensity-modulated radiotherapy: Results from a case-matched comparison. *Radiotherapy & Oncology*, pp. 232–239, 2017.
119. G. H. Fernald, E. Capriotti, K. J. Karczewski, R. Daneshjou, and R. B. Altman. Bioinformatics challenges for personalized medicine. *Bioinform.*, pp. 1741–1748, 2011.
120. IEEE VIS Scientific Visualization Contest. <https://www.uni-kl.de/scviscontest>, 2016.
121. G. H. Weber and H. Hauser. *Interactive Visual Exploration and Analysis*. Springer, 2014.
122. J. Canny. A computational approach to edge detection. *In Readings in Computer Vision*, pp. 184–203. Elsevier, 1987.
123. M. Chen and A. Golan. What may visualization processes optimize? *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2619–2632, 2016.
124. S. Mehta, S. Parthasarathy, and R. Machiraju. Visual Exploration of Spatio-temporal Relationships for Scientific Data. *In IEEE Symp. Visual Anal. Sci. Tech.*, pp. 11–18. 2006.
125. N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data*. Springer, 2005.
126. W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing Time-oriented data-A Systematic View. *Comp. Graph.*, pp. 401–409, 2007.

127. G. Favelier, C. Gueunet, and J. Tierny. Visualizing Ensembles of Viscous Fingers. *In IEEE Sci. Vis. Contest*, pp. 1–18. 2016.
128. J. Lukasczyk, G. Aldrich, M. Steptoe, G. Favelier, C. Gueunet, *et al.* Viscous Fingering: A Topological Visual Analytic Approach. *In Physical Modeling for Virtual Manufacturing Systems and Processes*, pp. 9–19. 2017.
129. D. Shepard. A Two-dimensional Interpolation Function for Irregularly-spaced Data. *In Proc. 23rd ACM Nat. Conf.*, pp. 517–524. 1968.
130. G. Ji. Feature Tracking and Viewing for Time-Varying Data Sets, 2006.
131. R. Samtaney, D. Silver, N. Zabusky, and J. Cao. Visualizing Features and Tracking Their Evolution. *Comp.*, pp. 20–27, 1994.
132. D. Silver and X. Wang. Tracking and visualizing turbulent 3D features. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 129–141, 1997.
133. P. T. Bremer, G. Weber, V. Pascucci, M. Day, and J. Bell. Analyzing and Tracking Burning Structures in Lean Premixed Hydrogen Flames. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 248–260, 2010.
134. W. Widanagamaachchi, C. Christensen, V. Pascucci, and P.-T. Bremer. Interactive exploration of large-scale time-varying data using dynamic tracking graphs. *In IEEE Symp. Large Data Anal. Vis. (LDAV)*, pp. 9–17. 2012.
135. H. Guo, C. L. Phillips, T. Peterka, D. Karpeyev, and A. Glatz. Extracting, Tracking, and Visualizing Magnetic Flux Vortices in 3D Complex-Valued Superconductor Simulation Data. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 827–836, 2016.
136. H. Hauser. *Generalizing focus+context visualization*, pp. 305–327. Springer, 2005.
137. J. Kehrer and H. Hauser. Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 495–513, 2013.
138. P. Muigg, J. Kehrer, S. Oeltze, H. Piringer, H. Doleisch, B. Preim, and H. Hauser. A Four-level Focus+Context Approach to Interactive Visual Analysis of Temporal Features in Large Scientific Data. *In Eurographics / IEEE - VGTC Conf. Vis.*, pp. 775–782. 2008.

139. K. Potter, A. Wilson, B. P-T, C. Williams, C. Doutriaux, P. V, and C. Johnson. Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated ViSUS-CDAT systems. *J. Phys.: Conf. S.*, 2009.
140. A. T. Wilson and K. C. Potter. Toward Visual Analysis of Ensemble Data Sets. *In IEEE Proc. Work. Ultrascale Vis.*, pp. 48–53. 2009.
141. C. Johnson and A. Sanderson. A Next Step: Visualizing Errors and Uncertainty. *IEEE Comp. Graph. Appl. (CGA)*, pp. 6–10, 2003.
142. K. Potter, P. Rosen, and C. Johnson. *From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches*, pp. 226–249. Springer, 2012.
143. T. Höllt, A. Magdy, P. Zhan, G. Chen, G. Gopalakrishnan, v. Hoteit, C. D. Hansen, and M. Hadwiger. Ovis: A Framework for Visual Analysis of Ocean Forecast Ensembles. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1114–1126, 2014.
144. I. Demir, C. Dick, and R. Westermann. Multi-Charts for Comparative 3D Ensemble Visualization. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2694–2703, 2014.
145. L. Hao, C. Healey, and S. A. Bass. Effective Visualization of Temporal Ensembles. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 787–796, 2015.
146. S. Biswas, K. W. Bowyer, and P. J. Flynn. Multidimensional Scaling for Matching Low-Resolution Face Images. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, pp. 2019–2030, 2012.
147. K. W. Kolence and P. J. Kiviat. Software Unit Profiles & Kiviat Figures. *SIGMETRICS Perform. Eval. Rev.*, pp. 2–12, 1973.
148. A. Maries, N. Mays, M. O. Hunt, K. Wong, W. Layton, G. Marai, *et al.* Grace: A Visual Comparison framework for integrated spatial and non-spatial geriatric data. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 2916–2925, 2013.
149. G. Marai. Visual Scaffolding in Integrated Spatial and Nonspatial Visual Analysis. *In Proc. 6th Int. EuroVis Workshop Vis. Anal.*, pp. 1–5. 2015.
150. J. Ahrens, B. Geveci, and C. Law. Paraview: An end-user tool for large data visualization. *Visualization Handbook*, 2005.

151. H. Childs, E. Brugger, B. Whitlock, J. Meredith, S. Ahern, D. Pugmire, *et al.* VisIt: An End-User Tool For Visualizing and Analyzing Very Large Data. *In High Performance Visualization – Enabling Extreme-Scale Scientific Insight*, pp. 357–372. Chapman and Hall (CRC), 2012.
152. A. Burks, C. Sugiyama, T. Luciani, J. Komperda, and G. Marai. Interactive Exploration and Tracking of Viscous Fingers in Large-Scale Ensemble Simulations. IEEE Scientific Visualization Contest 2016, 2016.
153. J. A. Swanson and G. DeSalvo. ANSYS-Engineering analysis system user’s manual. *Swanson Analysis Systems, Inc., Elizabeth, Pa*, 1989.
154. M. Chapman, M. Lawrence, J. Keats, K. Cibulskis, C. Sougnez, *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature*, pp. 467–72, 2011.
155. N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale. *Data-driven Storytelling*. CRC Press, 2018.
156. A. Sherbondy, D. Akers, R. Mackenzie, R. Dougherty, and B. Wandell. Exploring connectivity of the brain’s white matter with dynamic queries. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 419–430, 2005.
157. 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.
158. S. Zhang, c. Demiralp, and D. H. Laidlaw. Visualizing Diffusion Tensor MR Images Using Streamtubes and Streamsurfaces. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 454–462, 2003.
159. J. Caban, A. Joshi, and P. Rheingans. Texture-based feature tracking for effective time-varying data visualization. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1472–1479, 2007.
160. B. Gower. *Scientific Method: A Historical and Philosophical Introduction*. Taylor & Francis, 2012.

VITA

Timothy Basil Luciani
tlucia2@uic.edu

Education

University of Illinois at Chicago, Electronic Visualization Laboratory

Chicago, IL

PH.D. IN COMPUTER SCIENCE, EMPH. DATA VISUALIZATION

Expected May/June 2019

- Thesis topic: *Problem-Driven Design Strategies for Scientific Data Visualization*
- Cumulative GPA: 3.9

Dietrich School of Arts and Sciences, University of Pittsburgh

Pittsburgh, PA

PH.D. IN COMPUTER SCIENCE

January 2012 - April 2014

- Focus on real-time GPGPU rendering and large-scale data for interdisciplinary visualizations and applications
- Transferred to University of Illinois at Chicago
- Cumulative GPA: 3.688

Dietrich School of Arts and Sciences, University of Pittsburgh

Pittsburgh, PA

B.S. IN COMPUTER SCIENCE

August 2008 - December 2011

- Emphasis: Mathematics, Physics
- Graduated Cum Laude
- Cumulative GPA: 3.5 (in major)

Experience

University of Illinois at Chicago, Electronic Visualization Lab

Chicago, IL

GRADUATE RESEARCH ASSISTANT

August 2017 - Current

- Investigated patient cohort similarity based on spatial descriptors.

National Science Foundation

Chicago, IL

GRADUATE RESEARCH FELLOW

August 2015 - August 2017

- Continued research in large-scale data visualization for interdisciplinary domains.

General Dynamics - Mission Systems | Viz

Pittsburgh, PA

LEVEL 2 SOFTWARE ENGINEER

July 2014 - May 2017

- Co-authored software for emergency response coordinators to manage resources in real-time in both times of crisis and routine operation
- Architected the next-generation, in-house charting and visualization framework.

National Science Foundation

Pittsburgh, PA

GRADUATE RESEARCH FELLOW

January 2012 - July 2014

- Continued research in large-scale data visualization for interdisciplinary domains.

University of Pittsburgh

Pittsburgh, PA

UNDERGRADUATE RESEARCH ASSISTANT

May 2011 - December 2011

- Worked with researchers in Astronomy and Physics disciplines to develop tools for visualizing Large-scale data
- Worked with new web-technologies such as WebGL and HTML5
- Built upon existing code bases using CUDA/OpenCL to create faster visualizations.

Honors & Awards

2017	IEEE Visual Analytics Science and Technology (VAST) Challenge, MC2	IEEE Vis Conference	Phoenix, AZ
2017	IEEE Visual Analytics Science and Technology (VAST) Challenge, MC3	IEEE Vis Conference	Phoenix, AZ
2016	Student Volunteer of the Year Award	IEEE Vis Conference	Baltimore, MD
2016	Honorable Mention	IEEE Vis Conference: VGTC VPG Data Visualization Contest	Baltimore, MD
2016	Cover art of JIST January/February 2016 issue	Journal of Imaging Science and Technology	
2013	Data Contest Visualization Award	IEEE BioVis Conference Data Contest	Atlanta, GA
2012	Best-Paper Runner-Up	IEEE Large-Scale Data Analysis and Visualization Conference	Seattle, WA
2012	National Science Foundation Graduate Research Fellowship Program Recipient	NSF	
2012	Winner	University of Pittsburgh, CS Dept. Digital Media Contest	Pittsburgh, PA

Publications

BOOK CHAPTERS

- B3** M. Monfort, T. Luciani, J. Komperda, B. Ziebart, F. Mashayek, G.E. Marai, "Deep learning features of interest from turbulent combustion tensor fields", Modeling, Analysis, and Visualization of Anisotropy. 2017.
- B2** G. E. Marai, T. Luciani, A. Maries, S.L. Yilmaz, M.B. Nik, "Visual Descriptors for Dense Tensor Fields in Computational Turbulent Combustion: A Case Study", Journal of Imaging Science and Technology, vol 60, no 1, Jan. 1, 2016
- B1** A. Maries, T. Luciani, P.H. Piscuneri, M.B. Nik, S.L. Yilmaz, P. Givi, G.E. Marai, "A Clustering Method for Identifying Regions of Interest in Turbulent Combustion Tensor Fields", Visualization and Processing of Higher Order Descriptors for Multi-Valued Data. Editors: Ingrid Hotz and Thomas Schultz, Springer, pp. 1–18, 2015.

JOURNAL PUBLICATIONS

- J5** T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D.M. Vock, C.D. Fuller, G.E. Marai, "A Spatial Neighborhood Method for Computing Lymph Node Carcinoma Similarity In Precision Medicines", IEEE Transactions on Biomedical Engineering. Under Review.
- J4** T. Luciani, A. Burks, C. Sugiyama, J. Komperda, G.E. Marai, "Details-First, Show Context, Overview Last: Supporting Exploration of Viscous Fingers in Large-Scale Ensemble Simulations", IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 01, pp. 1–11, Jan. 2019.
- J3** C. Ma, T. Luciani, A. Terebus, J. Liang, and G. E. Marai. "PRODIGEN: Visualizing the Probability Landscape of Stochastic Gene Regulatory Networks in State and Time Space." BMC Bioinformatics. Feb. 2017. *(Presented at BioVis 2016)*
- J2** T. Luciani, J. Wenskovitch, K. Chen, D. Koes, T. Travers, G.E. Marai. "FixingTIM: FixingTIM: Interactive Exploration of Sequence and Structural Data to Identify Functional Mutations in Protein Families" BMC Bioinformatics, Aug. 2014.
- J1** T. Luciani, B. Cherinka, D. Oliphant, S. Myers, W.M. Wood-Vasey, A. Labrinidis, G.E. Marai. "Large-Scale Overlays and Trends: Visually Mining, Panning and Zooming the Observable Universe", IEEE Transactions on Visualization and Computer Graphics, pp. 1–12, July 2014.

CONFERENCE PUBLICATIONS

- C6** A. Wentzel, P. Hanula, T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D. Vock, C.D. Fuller, G.E. Marai. "Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration". IEEE Scientific Visualization Conference, Vancouver, BC, CA, Oct. 2019. Under Review
- C5** T. Luciani, A. Burks, C. Sugiyama, J. Komperda, G.E. Marai, "Details-First, Show Context, Overview Last: Supporting Exploration of Viscous Fingers in Large-Scale Ensemble Simulations", IEEE Transactions on Visualization and Computer Graphics, pp. 1–10, Oct. 2018. *(cross-listed as J5 above)*
- C5** C. Ma, T. Luciani, A. Terebus, J. Liang, and G. E. Marai. "PRODIGEN: Visualizing the Probability Landscape of Stochastic Gene Regulatory Networks in State and Time Space," pp 1–13, IEEE BioVis 2016. *(cross-listed as J3 above)*
- C4** D. McNamara, J. Tapia, C. Ma, T. Luciani, A. Burks, J. Trelles, and G. E. Marai. "Spatial Analysis of Employee Safety Using Organizable Event Quiltmaps". In Proceedings of the IEEE VIS 2016 Workshop on Temporal and Sequential Event Analysis, Baltimore, MD, USA, Oct. 2016.

- C3** J. Wenskovitch, T. Luciani, K. Chen, G.E. Marai. "FixingTIM: Identifying Functional Mutations in Protein Families through the Interactive Exploration of Sequence and Structural Data", IEEE BioVis 2013 Data Competition, pp. 1–4, Oct. 2013. **Data Contest Visualization Award.** (*Invited to J2*).
- C2** T. Luciani, S. Myers, B. Sun, B. Cherinka, W.M. Wood-Vasey, A. Labrinidis, G.E. Marai. "Panning and Zooming the Observable Universe with Prefix-Matching Indices and Pixel-Based Overlays", IEEE Large-scale Data Analysis and Visualization Symposium, pp. 1-8, Oct. 2012. **Best-Paper Runner-Up Award.** (*expanded into J1*).
- C1** P. Neophytou, R. Gheorghiu, R. Hachey, T. Luciani, B. Sun, A. Labrinidis, G.E. Marai, P.K. Chrysanthis. "AstroShelf: Understanding the Universe through Scalable Navigation of a Galaxy of Annotations", SIGMOD 2012 Demonstrations Comp.

PEER-REVIEWED CONFERENCE SHORT PAPERS, ABSTRACTS AND SYSTEM DEMONSTRATIONS

- P10** T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D. M. Vock, C.D. Fuller, G.E. Marai. "Correlating Toxicity Outcomes with Spatial Patterns of Lymph Node Metastasis for Oropharyngeal Cancer Patients". American Society for Radiation Oncology, Chicago, IL, USA. Sept. 2019.
- P9** Castor, J. Borowicz, A. Burks, M. Thomas, T. Luciani, G.E. Marai, "MC2 - Mining Factory Pollution Data through a Spatial-Nonspatial Flow Approach", IEEE Visual Analytics Science and Technology (VAST) Challenge 2017 Proceedings, pp. 1-2, 2017. **VAST Challenge Honorable Mention (MC2)** in competition with 56 submissions from teams in academia, industry, and government.
- P8** V. Mahida, B. Kupiec, A. Burks, T. Luciani, G.E. Marai. "MC3 - A Web-Based Interactive Image Explorer for Temporal Analysis of Satellite Images", IEEE Visual Analytics Science and Technology (VAST) Challenge 2017 Proceedings, pp. 1-2, 2017. **VAST Challenge Honorable Mention (MC3)** in competition with 56 submissions from teams in academia, industry, and government.
- P7** A. Wentzel, P. Hanula, T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D. M. Vock, C.D. Fuller, G.E. Marai. "Cohort-Based Spatial Similarity can Predict Radiotherapy Dose Distribution". American Society for Radiation Oncology, Chicago, IL, USA. Sept. 2019.
- P6** T. Luciani, J. Trelles, C. Ma, A. Burks, M. Thomas, K. Bharadwaj, S. Singh, P. Hanula, L. Di, G.E. Marai. "Multi-scale Voronoi-based ACT Assessment ". IEEE VGTC VPG International Data-Visualization Contest, Baltimore, MD, USA. **Honorable Mention.** Oct. 2016.
- P5** T. Luciani, C. Ma, J. Trelles, and G. E. Marai. "Developing a Data-Driven Wiki of Spatial-Nonspatial Integration Tools". In Proceedings of the IEEE VIS 2016 Workshop on Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization (C4PGV), Baltimore, MD, USA, Oct. 2016.
- P4** A. Burks, C. Sugiyama, T. Luciani, J. Komperda, G. E. Marai. "Interactive Exploration and Tracking of Viscous Fingers in Large-Scale Ensemble Simulations." IEEE Scientific Visualization Contest, 2016.
- P3** T. Luciani, A. Maries, M. Nik, S.L. Yilmaz, "Visualization of Tensor Quantities Used in Computational Turbulent Combustion", 66 Annual Meeting of the APS Division of Fluid Dynamics, Nov., 2013.
- P2** T. Luciani, A. Maries, H. Tran, M. Nik, S.L. Yilmaz, G.E. Marai, "A Novel Method for Tracking Tensor-based Regions of Interest in Large-Scale, Spatially-Dense Turbulent Combustion Data", IEEE Visualization 2012, Poster Abstracts with System Demonstration, pp. 1-2, Oct. 2012.
- P1** T. Luciani, R. Hachey, D.Q. Oliphant, B.A. Cherinka, G.E. Marai. "Pixel-based Overlays for Navigating a Galaxy of Observations". IEEE Visualization 2011 Large-Scale Data Analysis and Visualization Symposium Poster Compendium, Oct. 2011.

Invited Presentations

A Deep Learning Approach to Identifying Shock Locations in Turbulent Combustion Tensor Fields

Dagstuhl, Germany

DAGSTUHL VISUALIZATION AND PROCESSING OF ANISOTROPY IN IMAGING, GEOMETRY, AND ASTRONOMY

Oct. 2018

- Presented proof-of-concept work on deep learning approaches in computational fluid dynamics

Developing a Data-Driven Wiki of Spatial-Nonspatial Integration Tools

Baltimore, MD

VISUALIZATION OF TENSOR QUANTITIES USED IN COMPUTATIONAL TURBULENT COMBUSTION

Oct. 2016

- Presented current efforts at organizing our survey into a public electronic repository

6th Annual Meeting of the APS Division of Fluid Dynamics

Pittsburgh, PA

VISUALIZATION OF TENSOR QUANTITIES USED IN COMPUTATIONAL TURBULENT COMBUSTION

Nov. 2013

- Presented past research on flow visualization techniques

Allegheny Observatory Public Lecture Series

Pittsburgh, PA

PANNING AND ZOOMING THE OBSERVABLE UNIVERSE WITH PREFIX-MATCHING INDICES AND PIXEL-BASED OVERLAY

July 2013

- Presented current astronomy research on visual trends in spectral data

Technology Leadership Initiative Workshop

Pittsburgh, PA

INTRODUCTION TO ANIMATION AND VIDEO GAMES TUTORIAL

May 2013

- Taught Technology Leadership Initiative Workshop (TLIW) to 20 high school students

IEEE Large-scale Data Analysis and Visualization (LDAV) Conference

Seattle, WA

PAPER TRACK

Oct. 2012

- Presented paper entry (C3) at the annual conference

Pittsburgh Science and Technology Academy

Pittsburgh, PA

SCITECH SCIENCE FORUM

Jan. 2012

- Presented research in data visualization to high school students to promote interest in CS

All-Wavelength Extended Groth Strip International Survey (AEGIS)

Pittsburgh, PA

PITTSBURGH CONFERENCE

June 2011

- Presented astronomy research to AEGIS community for feedback during their annual conference

Committees

2019 **Chair**, IEEE VIS Student Volunteer Program

Vancouver, BC, CA

2018 **Chair**, IEEE VIS Student Volunteer Program

Berlin, Germany

2017 **Day Captain**, IEEE VIS Student Volunteer Program

Phoenix, AZ

2016 **Day Captain**, IEEE VIS Student Volunteer Program

Baltimore, MD

2013 **Vice-President**, University of Pittsburgh, Graduate Student Organization

Pittsburgh, PA