# ABSTRACT


Title of dissertation: SITUATED ANALYTICS FOR DATA SCIENTISTS

Andrea Batch, Doctor of Philosophy, 2022

Dissertation directed by: Professor Niklas Elmqvist
College of Information Studies



Much of Mark Weiser's vision of "ubiquitous computing" has come to fruition: We live in a world of interfaces that connect us with systems, devices, and people wherever we are. However, those of us in jobs that involve analyzing data and developing software find ourselves tied to environments that limit when and where we may conduct our work; it is ungainly and awkward to pull out a laptop during a stroll through a park, for example, but difficult to write a program on one's phone. In this dissertation, I discuss the current state of data visualization in data science and analysis workflows, the emerging domains of immersive and situated analytics, and how immersive and situated implementations and visualization techniques can be used to support data science. I will then describe the results of several years of my own empirical work with data scientists and other analytical professionals, particularly (though not exclusively) those employed with the U.S. Department of Commerce. These results, as they relate to visualization and visual analytics design based on user task performance, observations by the researcher and participants, and evaluation of observational data collected during user sessions, represent the first thread of research I will discuss in this dissertation. I will demonstrate how they might act as the guiding basis for my implementation of immersive and situated analytics systems and techniques.

As a data scientist and economist myself, I am naturally inclined to want to use high-frequency observational data to the end of realizing a research goal; indeed, a large part of my research contributions—and a second "thread" of research to be presented in this dissertation—have been around interpreting user behavior using real-time data collected during user sessions. I argue that the relationship between immersive analytics and data science can and should be reciprocal: While immersive implementations can support data science work, methods borrowed from data science are particularly well-suited for supporting the evaluation of the embodied interactions common in immersive and situated environments. I make this argument based on both the ease and importance of collecting spatial data from user sessions from the sensors required for immersive systems to function that I have experienced during the course of my own empirical work with data scientists. As part of this thread of research working from this perspective, this dissertation will introduce a framework for interpreting user session data that I evaluate with user experience researchers working in the tech industry.

Finally, this dissertation will present a synthesis of these two threads of research. I combine the design guidelines I derive from my empirical work with machine learning and signal processing techniques to interpret user behavior in real time in Wizualization, a mid-air gesture and speech-based augmented reality visual analytics system.

# SITUATED ANALYTICS FOR DATA SCIENTISTS

by

Andrea Batch

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Niklas Elmqvist, Chair/Advisor
Professor Kimbal Marriott
Professor Eun Kyoung Choe
Professor Vanessa Frias-Martinez
Professor Jordan Boyd-Graber

# Dedication

To my wife, Chloe; my mentor, Niklas; my siblings, Jenny, Aaron, & Jonathan; and my grandparents (Kapper and Julca alike), whose lives have been an inspiration to my own.

# Acknowledgments

There are many people who have played tremendously important roles in my life throughout the process of my doctoral research without whom I could never have arrived at this point. This dissertation would not have been at all possible without the presence of two people in particular: My advisor, guide, and lifeline from the beginning to the end of my doctorate, Niklas Elmqvist; and my wife, Chloe Batch, without whom I would surely be dead in a ditch somewhere, and whose artist's eye and—in no less than two instances—illustrations have contributed greatly to my body of published work.

I must also give special thanks to Vanessa Frias-Martinez and Eun Kyoung Choe, who have been on my committee since before I had a committee and who have given me invaluable direction throughout my time at Maryland. I will be eternally indebted to Catherine Plaisant for using her emerita time to guide my work. I am equally indebted to Kim Marriott for sacrificing his late-night hours to offer his guidance from the other side of the planet. I thank Jordan Boyd-Graber, as well, for his expert guidance despite his many commitments. I am also truly grateful to the Assistant Chief Economist at the U.S. Bureau of Economic Analysis, Abe Dunn, and to the Chief Economist, Dennis Fixler, and to the BEA as an agency, for their flexibility, support, and direction throughout this long process.

When I think about getting down to the gritty technical details and odd hours spent working on software (and sometimes hardware) engineering, it will forever be my colleagues, coauthors, and comrades Biswaksen Patnaik, Pete Butcher, Sungbok Shin, Max Cordeil, Andrew Cunningham, Sigfried Gold, Hanuma Teja Maddali, Kyungjun Lee, Yipeng Ji, Sebastian Hubenschmid, Jonathan Wieland, Daniel Fink, Johannes Zagermann, and Julia Liu who will come to mind. And of course what use is engineering without vision and wisdom? Tim Dwyer, Bruce H. Thomas, Harald Reiterer, Panos Ritsos, Jian Zhao, Mingming Fan, and Moses Akazue all plotted the intellectual routes I have followed. Finally, my dear

friends Sydney Hodges; Andrew Bossi; Bryan Seaborne Reid IV, Esquire; and Erica and John Henderson: Without our regular chats and D&D sessions, I would likely have become an even more eccentric hermit by the end of this thing than the mildly eccentric semi-hermit that I have indeed become.

Table of Contents

# Part I

# Part 1: Overview

# Chapter 1: Introduction

Sometimes, we want to keep working even when we're not at our desks. Or maybe we simply want a nice change of scenery to do the work we were already planning on doing in front of a screen in a stuffy office. This has never been more true than it is now, with many of us in a professional climate that has pivoted dramatically toward remote work following the rise of the COVID-19 pandemic. However, programming and data analysis are hard to do using a phone, leaving remote workers just as stuck with the desktop or laptop as their core working environment as they were prior to the global outbreaks beginning in 2020. There is a world beyond the mouse, keyboard, and monitor that is currently being delved to its depths by a small but growing population within the visualization research community which concerns itself with immersive environments—settings that surround the user with virtual representations of their work, either bringing the data to their world or bringing them into the imagined and abstract world of their data—but the average analytical professional has yet to see the fruits of all the labor performed by these researchers. Within this immersive analytics community is the domain of situated analytics, and therein lies the means through which the analyst can free themselves of the desk.

But this begs the questions: Should they? What do they stand to gain, beyond simply changing their surroundings? We want to build immersive and situated visual analytics systems that fit the needs of the average data analyst based on their real-world workflow. But first, we need to establish that there is justification in doing so–not just because we want to take our analysis on the road and pull ourselves away from the desk, but because there

are real benefits in terms of task performance, work enjoyment, and other elements that the analyst takes into consideration when deciding where and how to do their analyses. I will demonstrate through both a review of existing work (Section 2.2) and my own original work (Chapter 4) that immersive visual analytics systems do offer real benefits over traditional environments.

## 1.1 Context

In his 90-minute-long "Mother of All Demos" given at the 1968 Fall Joint Computer Conference in San Francisco, Douglas Engelbart introduced, among many other elements of modern computing, the computer mouse.[1] The manipulation of windows (also introduced by Engelbart in the same demo) and other interface components (several of which were also introduced at the same time) via the mouse cursor would change the course of computer navigation and define what would long be considered modern personal computing. The reduction of the barrier between user and virtual objects represented a dramatic shift toward direct manipulation [217].

In a 1993 magazine article on the concept of ubiquitous computing that he had introduced two years prior, Mark Weiser said that "[the] best user interface is the self-effacing one, the one that you don't even notice" [247, p. 71]. In other words, one of the central assumptions of both ubiquitous computing and of direct manipulation is that the thinner the perceived boundary (i.e., the interface) between the user and their objectives, the better. Touchscreens further reduce this barrier. While the invention of the touchscreen in 1946 predates the invention of the computer mouse by about two decades, touchscreens at first suffered a bad reputation among both human-computer interaction (HCI) research community and the computing device industry until the introduction of three redeeming features by researchers

[1] Video of the Mother of All Demos at the timestamp of the demonstration of the computer mouse is available at `https://www.youtube.com/watch?v=yJDv-zdhzMY&t=1888s`.

at the University of Maryland (UMD) Human-Computer Interaction Lab (HCIL): the "lift-off strategy" of waiting for touch input to end before triggering an event (1988) [190], touchscreen switches and sliding toggles (1990 and 1991) [189],[2] and a comparative analysis of high-precision touchscreens (1991) [211].

In 2013, Elmqvist and Irani [66] narrowed Weiser's vision specifically to the area of data analysis, coining the term "ubiquitous analytics:" The use of embedded and mobile networked devices for data analysis, and ideally for analyses that exploit so-called "big data." This vision of ubiquitous analytics has been hampered, however, by the fact that serious programming and data analysis remain difficult to do using a phone, and portable devices like laptops may offer a complete working environment for analysts, but they are not truly mobile devices. The following year, Roberts et al. [198] extended Elmqvist and Irani's vision to include immersive displays, such as virtual reality head-mounted displays (HMD) or multisensory displays that convey data through touch, sound, scent, or taste instead as well as visually.

It is only now, nearly a decade later, that HMDs are increasingly affordable and capable of rendering high-quality environments and the visual analytics research community is shifting its focus more intently toward the immersive through the domains of immersive analytics (IA) and situated analytics (SA). As a new field, IA presents a number of "grand challenges", several of which have been outlined by Ens et al. [71]; many of these challenges revolve around SA, the use-cases and evaluation of IA, collaboration, and interaction. One such challenge in IA that does not fit into these categories but does mesh well with the focus on large datasets and computer vision modeling seen in early work in ubiquitous analytics is the combination of human and computer intelligence. Computer intelligence—specifically, computer vision—was also identified by Roberts et al. [198] as an "enabling technology"

[2]Materials, including papers and video of a demonstration of the touchscreen switches and toggles, can be found at `https://www.cs.umd.edu/hcil/touchscreens`.

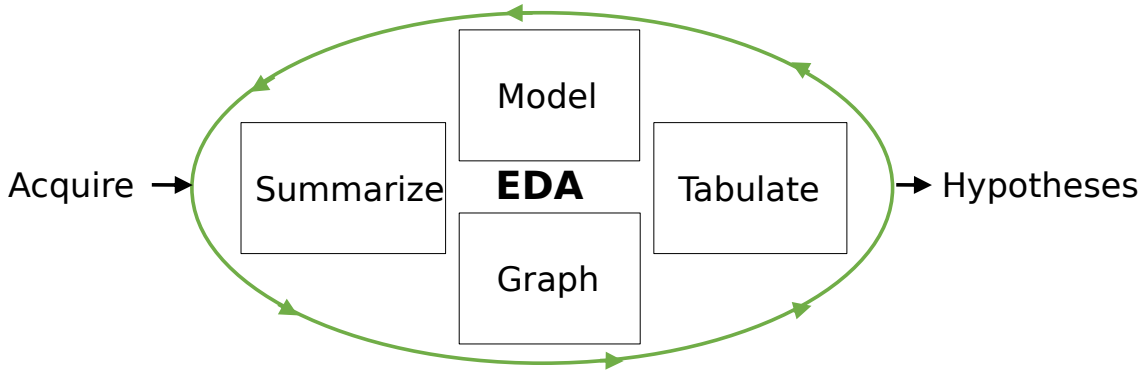



Figure 1.1: Exploratory data analysis elements and flow.

for ubiquitous visualization displays.

## 1.2 Terminology

Below I define some of the commonly-used terms throughout this dissertation.

**Immersive Analytics (IA)**: A domain of visualization research covering techniques and systems in which the user is spatially co-located with virtual representations of abstract data.

**Situated Analytics (SA)**: IA that deals specifically with mixed reality (MR) views of data and the situational context of the user.

**Exploratory Data Analysis (EDA)**: The steps after acquiring a dataset during which the analyst tabulates, summarizes, and creates visual representations of data to form hypotheses (Figure 1.1).

**Grammar of Graphics**: Any declarative language and set of rules that is composed of a set of building blocks for specifying visual representations of data.

**Machine Learning (ML)**: The use of computer algorithms that "improve" (typically in accuracy given a set of target outputs) based on repeated exposure to data.

**Computer Vision (CV)**: The use of computer algorithms, typically machine learning algorithms, for processing and interpreting image and video data.

**Human-Computer Interaction (HCI)**: A field of research focusing on all points at which the human and the computer meet.

**Contextual Design**: A method of software and system design that involves inserting the researcher into the context in which the target user goes about their daily lives.

**Contextual Inquiry**: The data collection methods of contextual design involving the observation of, and communication with, participants.

## 1.3 Research Problems, Questions, and Objectives

The grand challenges described by Ens et al. [71] aside, Elmqvist and Irani's [66] vision of ubiquitous analytics and the emergence of the domains of situated and immersive analytics demonstrate a great deal of interest in data analysis beyond the monitor and keyboard on the part of the visualization research community. I, too, believe that the future of visual analytics is in immersive, situated systems for research, development, and data analysis. Yet researchers in IA and SA face a problem: **We do not see data scientists flocking to either mobile or immersive environments to do their daily work.** This observation—that data scientists and similar analytical workers are not making use of immersive technology for their work—begs the problem central to this dissertation: If there is so much to be gained by adopting a display system that is both mobile or situated and immersive for serious data analysis, **what are the design requirements for real data analysts to voluntarily do real work in such a system?**

Augmented reality HMDs of today are, I argue, in a similar market position to the one touchscreens were in following the publications by UMD HCIL noted in Section 1.1: There is a great deal of interest in the research communities to which they are relevant, but the industry and consumer markets for them have not yet caught up to the research. The HoloLens 2, is listed at the time I am writing as being available to purchase at the hefty—but

representative for AR HMDs—price tag of $3,500 (USD); AR HMDs are close to being, but are not quite yet, poised to become a common household appliance. In conjunction with the price, one thing holding the AR HMD back from ubiquity is the lack of a so-called "killer app." I believe that such an app may be found in analytical work; no device but the AR HMD offers as thin an interface between the user and their environment, nor an interface that can be so seamlessly integrated with the tools of the trade and workflows currently common among analytical workers.

In addition to my beliefs about immersive and situated analytics as the future of visualization research, I also argue that systems in this domain are particularly well-suited for machine learning pipelines designed to support user interaction evaluation as part of the system design and development process. I make this argument based on the self-evident observation that immersive and situated systems require necessity of a level of detail in spatial tracking of users that is neither needed nor typical in traditional displays. Another research question also arises from both ubiquitous and immersive analytics literature in this context: **How can machine learning and related data modeling algorithms be applied to user data to support immersive and situated analytics?**

At its highest level, my work here revolves around the objective of creating a full-featured R&D and data analysis workspace beyond the 2D screen that is based on real domain expert work processes with the extension of the user's view and experience based on their inputs and activities as a focal center of my system's design. These inputs can be any data observable during their user sessions from keyboard activity to camera or sensor tracking data. In other words, while my published work to date is sadly lacking the influence to prove my thesis by virtue of having created the killer app that makes AR HMDs become an item found in every household, **my research objective is to make full use of multi-modal user inputs (e.g., combining spoken word with hand gesture with direct manipulation of selected objects) in IA and SA environments, which should in turn be designed for easy integration with**

**existing data science workflows.**

## 1.4 Thesis Statement

*My dissertation focuses on combining design parameters based on observations of data science workflows with models for interpreting real-time user interaction data to produce an improved system and grammar of graphics for situated analytics.*

## 1.5 Relevance and Contributions

Included in my work here is material from the following published and still under review research papers, with my role relative to the author authors described following each item within the list below:

1. ***Andrea Batch*** *and Niklas Elmqvist. The interactive visualization gap in initial exploratory data analysis. IEEE Transactions on Visualization and Computer Graphics, 24(1):278–287, Jan. 2018.*

   In this work, my first publication—conducted mainly during the summer prior to my first doctoral-level classes—I played the role of the embedded researcher, conducting user sessions with participants across several agencies within the U.S. Department of Commerce and performing a qualitative evaluation of their work and how visualization fits into the initial iterations of their exploratory analysis process.

2. ***Andrea Batch****, Kyungjun Lee, Hanuma Teja Maddali, and Niklas Elmqvist. Gesture and action discovery for evaluating virtual environments with semi-supervised segmentation of telemetry records. In Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality, Piscataway, NJ, USA, 2018. IEEE.*

   In this publication, I, Kyungjun Lee, and Hanuma Teja Maddali split the tasks of

developing the machine learning pipeline for identifying and clustering novel gestures, with my focus being on combining pre-trained joint recognition models with the computational statistics algorithms used to segment and cluster joints, Kyungjun focusing on constructing the conputer vision models used to validate the clusters, which I also evaluated, and Teja sharing in both areas of application as well as in constructing the architecture for our pipeline.

3. ***Andrea Batch**, Andrew Cunningham, Maxime Cordeil, Niklas Elmqvist, Tim Dwyer, Bruce H. Thomas, and Kim Marriott. There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics. IEEE Transactions on Visualization and Computer Graphics, 26(1):536–546, 2020.*

   In this publication, I once again took up the mantle of the embedded researcher, this time exclusively at the U.S. Bureau of Economic Analysis, where I lead the data collection and analysis; I conducted all user sessions with our expert participants during the course of this study, and all authors contributed to evaluation of the sessions. I also made some minor software development contributions in extending ImAxes, an implementation introduced by my coauthors Maxime Cordeil, Andrew Cunningham, Tim Dwyer, Bruce H. Thomas, and Kim Marriott in their prior work [52].

4. ***Andrea Batch**, Yipeng Ji, Jian Zhao, Mingming Fan, and Niklas Elmqvist. uxSense: Supporting user experience evaluation using visualization and computer vision. Pending review for publication.*

   In this work, I developed the computational statistics and computer vision pipeline back-end, and co-developed the front end with Yipeng Ji; Yipeng also worked on the audio processing filters of our system. I also conducted the remote user studies and evaluated the data we collected.

5. ***Andrea Batch**, Sungbok Shin, Julia Liu, Peter W.S. Butcher, Panagiotis Ritsos, and*

*Niklas Elmqvist. The world is your Holodeck: View management for situated visualization. Pending review for publication.*

In this evaluation study, I lead the software development of the implementations we used to test our view management techniques, with Sungbok Shin, Julia Liu, and Pete Butcher sharing the responsibility as co-developers. Pete and I evenly split the task of conducting our user sessions, which I then evaluated.

6. ***Andrea Batch**, Peter W.S. Butcher, Panagiotis Ritsos, and Niklas Elmqvist. Wizualization: A "Hard Magic" WebXR Visualization System. Pending review for publication.*

   In this implementation, Pete Butcher and I split the roles of developing software for our Wizualization ecosystem, with my work including our rendering system (Wizualization), signaling server (Weave), and code notebook (Spellbook), while Pete constructed our grammar of graphics (Optomancy).

There have also been several peripherally-related publications that I played roles in which are relevant enough to cite in this dissertation, but which were either not so integral to the scope of this dissertation to merit inclusion as content or for which my role was not central enough to justify its inclusion here, including:

1. *Biswaksen Patnaik, **Andrea Batch**, and Niklas Elmqvist. Information olfactation: Harnessing scent to convey data. IEEE Transactions on Visualization and Computer Graphics, 25(1):726–736, 2018.*

   In this publication, Biswaksen Patnaik engineered the hardware and I developed the software to construct an olfactory display system able to swap modes between both desktop and VR; I also evaluated a selection of work from across several disciplines, including HCI, cognitive science, and neurology, and used it as the grounding for our theoretical model of information olfactation.

2. ***Andrea Batch**, Biswaksen Patnaik, Moses Akazue, and Niklas Elmqvist. Scents and sensibility: Evaluating information olfactation. In Proceedings of the ACM Conference on Human Factors in Computing Systems, page 1–14, New York, NY, USA, 2020. ACM.*

   Biswaksen and I both reprised our respective roles in this publication, I as lead software developer and he as lead hardware engineer, with Moses Akazue joining the project as a leading expert on thermal display engineering. Biswaksen conducted the user studies in person for this study, and all co-authors contributed to evaluation.

3. *Sebastian Hubenschmid, Jonathan Wieland, Daniel Immanuel Fink, **Andrea Batch**, Johannes Zagermann, Niklas Elmqvist, and Harald Reiterer. ReLive: Bridging in-situ and ex-situ visual analytics for analyzing mixed reality user studies. In Proceedings of the ACM Conference on Human Factors in Computing Systems, New York, NY, USA, 2022. ACM.*

   While this publication is relevant to this dissertation, my contributions in this work were not significant enough to justify its inclusion here; my role initially revolved around developing the 2D web-based interface for the ReLive system, but as my other responsibilities and projects piled up, most of the development tasks for the 2D interface were offloaded onto my very gracious co-authors, especially Sebastian Hubenschmid, Jonathan Wieland, and Daniel Fink, who handled it with aplomb. I also contributed anonymized data from one of my prior studies as use cases, and conducted a small number of the user studies with expert visualization researchers for evaluation of the ReLive system.

The publication history of my work in this area does, I hope, indicate that my contributions have been deemed relevant, interesting, and important enough by the visualization and HCI communities to have a place in the discussion of immersive analytics, user evaluation,

and the intersection of both specializations. Indeed, the primary two contributions of this dissertation revolve around this intersection, and they are as follows:

**First**, I construct a design space from several years of evaluating a) the data analysis processes of data science and domain experts in their places of work, and b) user studies of immersive and situated systems and techniques. I also describe in detail the methods, results, and various implementations that I have contributed to as noted in the lists above.

**Second**, I introduce Wizualization, a novel situated analytics implementation designed to be practical for data scientist and domain expert use based on the first two contributions of my work: Practical, in that it fits the requirements of the user's workflow and the value it adds to the user's work is worth the inconvenience of routinely using a HMD. While not every aspect of prior systems featuring machine learning pipelines that I have contributed to—most notably uxSense, an HCI evaluation visual analytics tool featuring a pipeline and supporting the detection of a priori events of interest in user data for the purpose of evaluating unanticipated interactions with the interface—have been integrated with the Wizualization system, many of the methods I have picked up for working with multi-modal action and interaction data have been integrated with deterministic models as part of the extensible nature of Wizualization and the components we have developed as part of its ecosystem.

## 1.6 A Positionality Statement

The data science work process is central to my thesis, and I also want to give you an idea of what my own work process has looked like, so I've tried to represent it here. I see myself first and foremost as a software developer or engineer: I like to build things that turn data into insight, and that's what I came to Maryland in the hopes of getting better at. That's what the largest part of the process flow I have included in Figure 1.2, "Iterative

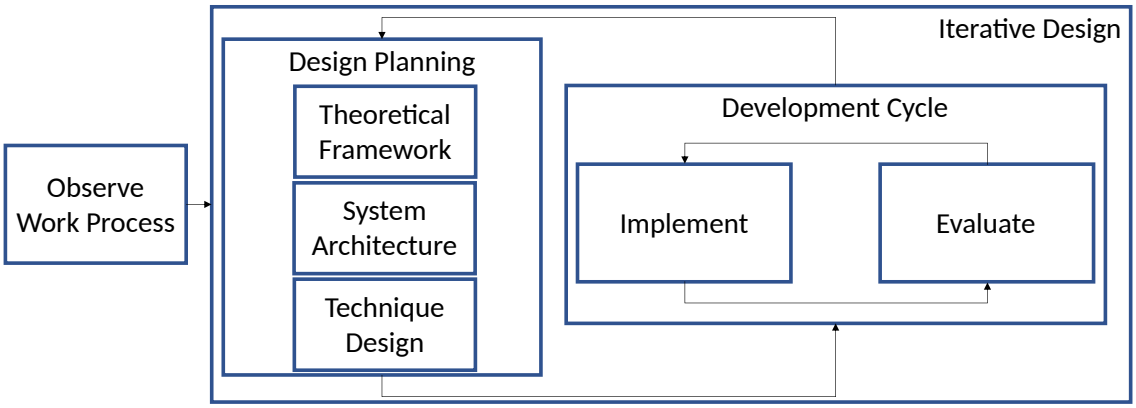



Figure 1.2: I consider myself to be, above all else, a software architect at heart; given the focus in my work on data science processes, I feel it appropriate to also describe the general cycles of my research and development process.

Design," is all about. But before we can do that, I first needed to understand the ways that my target users already make sense of data, which is where the first step, observing their work processes, on the left of my iterative design cycle, comes into play. This is why so much of my work has revolved around the qualitative evaluation of my target user group.

However, despite this reliance on qualitative work, consider this as a design perspective: We believe tools should be designed to be deployed into uncontrolled, messy, settings, where context matters and the task flow is be determined by the user. Data collection out in the wild is going to be limited to events and entities in the system and the sensors available on the user's hardware that the system is able to access. As an economist by training and trade, my instinct is to use this observational data to detect patterns in user behavior. This was our perspective in the algorithm we introduced in our work on gesture and action discovery in Chapter 6. We wanted to take this design perspective to its logical conclusion–take it a step further to ease the qualitative evaluation process, which we believe we demonstrate in our work on developing uxSense, presented in Chapter 7.

All the work in this dissertation involving participants has been approved by the University of Maryland, College Park Institutional Review Board.

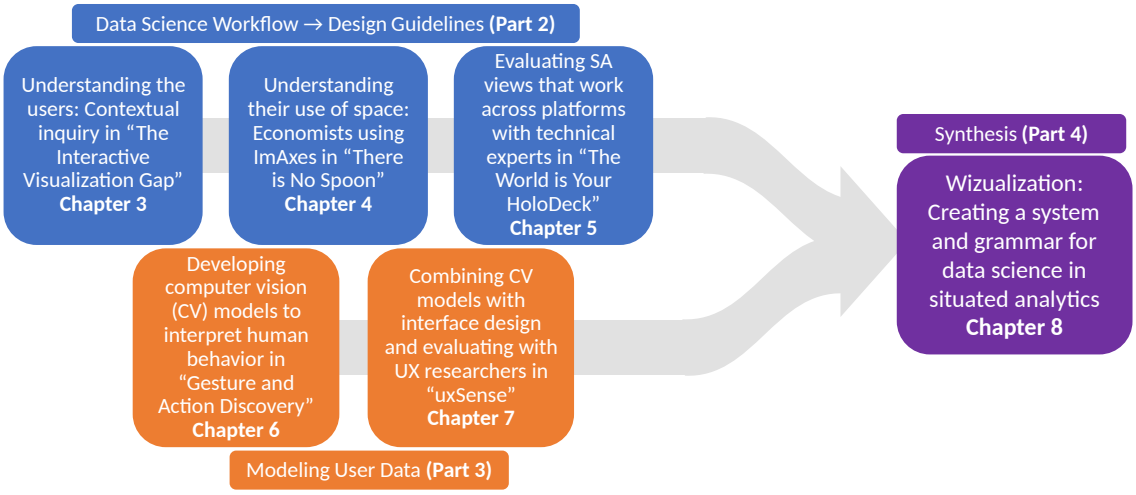



Figure 1.3: Two threads of research, synthesized in the implementation of Wizualization and the components of its ecosystem. The Part in this dissertation that each thread corresponds to and the Chapter in this dissertation that each study corresponds to are shown in bold.

## 1.7 The Structure of this Dissertation

The structure of this dissertation is split into four parts: Part 1, which begins with this chapter, presents a top-level roadmap to my work, including the related work particularly in visualization, human-computer interaction, and computer vision. My work can be viewed as falling along two threads of research, as shown in in Figure 1.3; Part 2 deals with the first thread, in which I seek to gain an understanding of the user's data analysis work process and use that to construct design guidelines. Part 3 deals with the second thread, in which I expand on the perspective I have just introduced in Section 1.6 regarding the use of observational data collected during user sessions for identifying patterns in their interactions and experience through the use of statistical and machine learning models. Part 4 synthesizes these two threads of research via the implementation of a system, Wizualization; as I said, I see myself first as a software developer, so what better way to present the synthesis of my work than through the implementation of software?

In the second chapter of Part 1 detailing related work, *Chapter 2*, I begin with the

most important part of any analytical system: The human user. The users that my work has targeted have all been analytical experts of some stripe: Data scientists, economists, UX researchers, and others following similar pursuits. Section 2.1 briefly details prior work investigating the tasks, tools, and processes that such users are often concerned with. Section 2.2 makes what may at first seem a jarring pivot into the discussion of how physical space and the objects in them can support cognition; however, it is this relationship between space and analysis that has motivated my work, so please bear with me. Section 2.3 briefly discusses some of the more influential grammars of graphics for data visualization; the reason yet another seemingly abrupt shift in direction will become evident to the reader by the end of Section 2.2.5. Section 2.4 covers a broad selection of qualitative and quantitative methods for evaluating the points at which human meets computer that I believe are most germane to the methods that I myself have applied.

Again, the chapters of Part 2 detail my work along that first thread of research in Figure 1.3; each of these chapters presents the reader with my methods, the study results, and the discussion of those results which lays out any design findings that arose from our study. In the first chapter of Part 2, *Chapter 3*, I discuss my exclusively qualitative methods in Section 3.1, my results in Section 3.2, and the design ramifications of our findings in Section 3.3. The second, Chapter 4, describes a multi-phase study that involved a mixture of qualitative and quantitative methods, described in Section 4.1, selected specifically for immersive analytics systems; the results of our study are discussed in Section 4.2, and the design implications are discussed in Section 4.3. We narrow our focus further to a study of view management for situated visualization in Chapter 5; much of the focus of this chapter is on the design space itself–namely, properties and challenges of SA view management (Section 5.1) and how those properties and challenges relate to the techniques we opted to evaluate (Section 5.2). We discuss the methods we used in Section 5.4 and our results in Section 5.5, and the design implications in Section 5.6.

The chapters of Part 3 detail my work along the second thread of research in Figure 1.3. Chapter 6 presents a lens through which models for detecting and classifying spatial and multi-modal human behavior might be seen as tools to interpret the time users spend in a system, with IA systems in particular in mind, as they involve a fuller use of the user's body and the space around them; in this chapter, we introduce and evaluate a pipeline for detecting novel actions based on user pose data extracted from video without prior knowledge of what those actions may be. The methods we used for this pipeline, described in Section 6.1, are exclusively quantitative, and covers the use of machine learning and statistical models for evaluating observational data. Chapter 7 continues this train of thought by introducing uxSense, a client/server system that implements our framework for extracting user behavior from video; an overview of our mixed methods for evaluating uxSense is provided in Section 7.3.

Part 4 synthesizes the threads of research discussed in Parts 2 and 3. The first chapter of Part 4, *Chapter 8* introduces Wizualization, its grammar of graphics (Optomancy), its signaling server (Weave), and the code notebook we built to demonstrate ecosystem (Spellbook). Chapter 9 summarizes the limitations of my work, which I view as falling into three categories: Methodological (in the case of qualitative work, which practically necessitates small user samples), technical, and ethical. Finally, this dissertation concludes with *Chapter 10*. I summarize my contributions in Section 10.1, while Section 10.2 discusses some suggestions for future directions in research.

# Chapter 2: Related Work

In order to understand the design space of situated visualization and analysis tools, I must first discuss the design issues and problems that affect the wider world of immersive analytics (IA). For my application in particular, I must also first understand the workflow for which situated analytics is perhaps most relevant.

## 2.1 Data Science Workflows: Batch's Visualization Gap

Digital tools are critical to data science and analytics workflows, and current practice spans data analysis tools such as R[1], Pandas [157][2], and SAS[3]; database systems such as MySQL[4], MongoDB[5]; data warehousing services such as Amazon Redshift[6]; and machine learning libraries such as scikit-learn [186][7] and TensorFlow [1][8]. While there is no formally standardized workflow or process that fits every data scientist, and every professional tends to establish their own, a common process typically consists of the following general stages [4, 18, 119]:

1. *Discovery:* Formulating an interesting question and determining the data necessary to

[1] http://r-project.org/
[2] http://pandas.pydata.org/
[3] http://www.sas.com/
[4] http://www.mysql.com/
[5] http://www.mongodb.com/
[6] https://aws.amazon.com/redshift/
[7] http://scikit-learn.org/
[8] https://www.tensorflow.org/

answer it;

2. *Acquisition:* Locating, organizing, and preparing data so that it is accessible to the chosen analysis environment;[9]

3. *Exploration:* Investigating and analyzing the dataset in order to collect insights and understand the data;

4. *Modeling:* Building, fitting, and validating a model that can explain the dataset and the observed phenomena; and

5. *Communication:* Disseminating the results to stakeholders in reports, presentations, and charts.

Static visualization is commonly used in the communication phase of data science workflows, and data scientists sometimes use them as part of the analysis as well [87, 119].

John Tukey's notion of exploratory data analysis [237] is firmly entwined with visual methods. Four years ago, Niklas Elmqvist and I began an exploration of how visualization fit into data scientists' exploratory analysis process, and at that time, interactive visualization was generally not a standard component of this workflow [18]. Spreadsheet software chart generation and tools that extend data tables, such as Tableau [228], along with Python and R libraries such as ggplot2 [252], were used for a variety of static visualization techniques in a format easily accessible and usable by data scientists during the course of my study. While examples of visualization researchers developing techniques using environments popular with data scientists do exist [187], they were not commonplace.

The response to my findings by the visualization community has shifted the needle forward on closing this gap. Shortly before my work was published, Satyanarayan *et al.* [205] had already begun to address the gap by introducing a high-level grammar of

[9] Also often called "ETL," meaning "extract, transform, and load."

graphics, Vega-Lite, which presents a set of standardized linguistic rules for producing interactive data visualizations using a concise JSON format for data to be represented by the grammar. Since my publication on the visualization gap, further inquiry has been conducted to characterize the analytical processes of data scientists and domain experts, and how visualization research can better support their work. Milani *et al.* [160], for example, look evaluate the early parts of the data science workflow, creating profiles of data pre-processing activities. In their retrospective analysis, Crisan *et al.* [54] break the data science workflow into four higher-order processes and fourteen lower-order processes, with each **higher-order** process containing one *lower-order* process for which visualization was a key component: *profiling* (**preparation**), *interpretation* (**analysis**), *monitoring* (**deployment**), and *dissemination* (**communication**). New libraries and systems have also been introduced which combat the visualization gap, including Vistrates [11], and an implementation that has perhaps held most true to my recommendations, B2 [259], a Jupyter notebook that brings interactive visualization to the tools that are already common for exploratory analysis among data scientists.

## 2.2 "Space to Think"

Managing and navigating space, virtual or physical alike, has always been central to human cognition. As Norman holds, "*it is things that make us smart"* [175], and according to distributed [107], embodied [213], and extended [47] forms of cognition, this very much includes physical space. In seminal work from cognitive psychology, Kirsh [125] demonstrated that humans tend to offload cognitive tasks in physical space to simplify choice, perception, and internal computation. But how many of these ideas translate to digital space on a computer screen?

Kirsh and Maglio [126] showed that screen space can support internal computation in

so-called *epistemic* actions—actions that serve no other purpose than to facilitate thought—in the video game Tetris. Similar effects have also been observed for recall through *spatial memory*: using the Data Mountain [199], where digital objects are arranged on the face of a pseudo-3D "mountain," participants were able to find previously placed website icons significantly faster than when using a conventional bookmark display. This harnessing of spatial memory is also similar to users leveraging *physical navigation* [14] in large and immersive displays with persistent locations of objects, thus allowing muscle memory and proprioception to replace some of the mental effort involved in spatial navigation.

In particular, having access to large visual spaces has been shown to be useful for analytical tasks. For example, screen space can be organized into complex structures such as lists, stacks, heaps, and composites [216], thus reducing the need for mental models. Tan *et al.* [231] compared analytical task performance between monitors and wall displays, and showed that a physically large display yields significant improvements due to the increased immersion and presence, which biased participants to adopt an egocentric view of the data. Reda *et al.* [194] built on such findings to study the impact of physical size on actual visual exploration of data, and found consistent effects where more pixels yielded more discoveries and insights. Finally, Andrews *et al.* [5] (to whom I owe the title of this section) directly addressed strategies for spatial arrangement of documents on a large, tiled 2D display in a visual analytics task. Their observations unearthed several interesting phenomena, such as the support of external memory, the structuring of the space using grouping and layout, and the high degree of integration between process, representation, and data that the large display space scaffolded.

One of the mechanisms by which IA supports human cognition is by simulating the experience of space and self: Through presence, immersion, and embodiment. *Presence* is the subjective psychological experience of being in a virtual or remote space, and *immersion* is the objective characteristics of the technology used to present the space [123]. The sense

of *embodiment* refers to the sensations that accrue while being inside, having, and controlling a body in VR. A common method of measuring presence is with questionnaires [15, 138, 210, 222, 241, 257]. These studies make a distinction between immersion and presence, where immersion is a necessary (but not sufficient) condition for the experience of presence in a VR interface [83, 222, 257]. Another necessary condition is *involvement* (or attention): the internal processes and external conditions influencing the user's ability to focus on stimuli in the environment [257]. Clearly, while immersion is tied to the technology used to deliver the virtual environment, presence is a more holistic property that is harder to pin down. Witmer and Singer argue, backed by other foundational research on the subject, that immersion and presence are determined by factors influencing the user's sense of control, realism, sensory feedback/stimulation, and distraction [138, 214, 257].

### 2.2.1 Direct Manipulation and Sketching in Visualization

Beyond games, one of the original applications of direct manipulation (as discussed in Section 1.1) was in visualization [217]. As a case in point, the seminal dynamic queries method minimizes indirection and reduces barriers between users and visual representations [2]. Building on the direct manipulation idea, Elmqvist et al. [68] proposed the notion of *fluid interaction*, arguing that "interaction in visualization is the catalyst for the user's dialogue with the data, and, ultimately, the user's actual understanding and insight into these data." They define a fluid interface as being one that: (1) promotes flow (a mental state of complete immersion in an activity) [55], (2) supports direct manipulation [217], and (3) minimizes Norman's gulfs of interaction [177] (i.e., the difference between the user's intended action and the actions afforded by the system).

Sketch-based systems are an example of fluidity in interaction. In SketchStory, Lee et al. [136] enable users to give ad-hoc data presentations by authoring visualizations on the

fly using sketch-based pen strokes. ScribbleQuery [174] applies touch-based sketching to brushing and selection in parallel coordinate plots. In Visualization-by-Sketching [209], Schroeder and Keefe pivot from analytical users to artistic users as their target population in implementing a system that augments the digital work of artists and designers with real data.

The advent of consumer-ready immersive displays reduces the barrier further still, with the most common modes of interaction in immersive analytical environments being virtual hands and virtual ray pointers [242]. *Scientific sketching* [120] allows for free-form 3D sketching to support data analytics in an immersive environment; the approach has since been adapted to multiple applications, including 3D fluid flow, collaborative analysis, and paleontology [178]. With that said, the use of sketching in immersive systems is not new; in fact, work in the area of direct manipulation was already exploring this mode of input for creating 3D scenes in 1996 with the SKETCH system [263]. However, this and similar work of that era focused on simulated objects or free-form design [109] rather than on data visualization. The increasing affordability of immersive head-mounted displays have given rise to new, more specialized fields of research, most relevantly that of visualization revolving around immersive and situated analytics.

### 2.2.2 Immersive Analytics

AR, along with virtual reality (VR) and mixed reality (MR)—immersive display and input technologies on the reality-virtuality continuum [161]—have long been used for visualizing physically embedded data [128, 131, 197, 239]. Recently, this has been extended to include more abstract data using *immersive analytics* [43, 65, 154]. IA is a visualization (VIS) framework and research focus on environments in which the user is spatially co-located with virtual representations of abstract data [43]. According to Dwyer *et al.* [65], *"Immersive Analytics is the use of engaging, embodied analysis tools to support data understanding*

*and decision making."* This notion fits well with a definition of AR by Mackay [146], which describes an environment augmented by interactive, networked objects. Mackay's viewpoint also approaches Weiser's [246] Ubiquitous Computing [16], and in that regard is well aligned with the notions of ubiquitous [66], immersive, and situated analytics [154] explored in this work.

Several IA applications have emerged that leverage the presence and engagement of VR. Simpson *et al.* [221] proposed an IA tool to explore climate economy models by leveraging spatial understanding from immersion on 2D multidimensional representations. The open-source ImAxes system [52] introduced the concept of an embodied axis to enable users to quickly build multidimensional visualizations in VR using natural interactions. FiberClay [106] uses an immersive approach for exploring large-scale spatial trajectory data in 3D, and the system was informally evaluated with air traffic controllers. However, none of these systems involved formal studies on how experts use the available 3D space, or how they might use immersive systems in day-to-day data analysis. Patnaik *et al.* [183] introduced the design space of information olfactation—the use of scent to convey abstract information—and implemented viScent as proof of concept. Butscher *et al.* [38] proposed the ART tool for collaborative AR parallel-coordinate-plot viewing with tabletop touch-input and performed an informal group-based walkthrough evaluation of the system with expert users exploring immersion, presence, spatial layout, and engagement.

The premise of IA is that the immersive setting will yield a richer and more embodied data analysis experience than traditional means. IA has been touted to decrease level of indirection, allow more natural input mechanisms, and the free-form space of a 3D virtual environment, which enables intelligent space usage [5, 125]. However, navigation and orientation issues arise in immersive environments in general, and immersive visualization tools are no exception. Problems like occlusion [67], depth and distance perception [60], as well as interaction with said distant objects [249] remain challenging issues in 'vanilla'

VR/MR and, consequently, in IA. Last but not least, discomfort, motion sickness, and other ocular and non-ocular symptoms of HMD use, well-explored in the VR domain [114, Pp. 159–221], are worthy of consideration in any immersive implementation.

There are still few studies that test these factors for IA, but empirical grounding for IA and its expanding design space has begun to emerge from both the virtual and mixed reality (VR/MR, collectively denoted XR) [250], and VIS communities [9, 17], including examples of multimodal data representation—systems that present data to the user using senses other than vision—such as the work in olfactation that I have conducted alongside Biswaksen Patnaik, Moses Akazue, and Niklas Elmqvist [21]. In a related study, Steed *et al.* [225] evaluated these factors with Samsung Gear VR and Google Cardboard, and they found tangible evidence of aspects of presence and immersion being measurable in this setting. Mottelson and Hornbæk [163] conducted a similar field-deployed evaluation with cardboard VR devices, comparing the results to a laboratory study. Their findings are consistent with those of Steed *et al.*, yet also indicate that performance is impacted by the quality of the VR technology and the internal validity of the study. However, because the discipline is relatively new, the problems involved in designing IA environments have not been thoroughly defined or validated. In Batch *et al.* [17], Andrew Cunningham, Maxime Cordeil, Niklas Elmqvist, Tim Dwyer, Bruce H. Thomas, Kim Marriott, and I conducted an evaluation of economists exploring multivariate temporal data "in the wild;" few other recent studies exist that study VR "in the wild," and even fewer exist for multidimensional data visualization.

### 2.2.3 Situated Analytics

Situated analytics (SA) [69] is a subspace of IA that deals specifically with MR views of information that visually link virtual and physical objects of interest, registering spatial

locations for abstract information and supporting analytical interactions. SA has strong links to early research efforts on mobile, wearable—and often cumbersome—AR, such as the Touring Machine [76], which provided information, as labels that were situated near various buildings of a university campus. Since then, the hardware requirements have become significantly less cumbersome, and the application areas and task requirements of SA systems have also diversified. However, despite the growing trend of emergent SA systems and use cases, there are still technical vulnerabilities—such as inaccurate GPS sensors [110, 251]—and intrinsic challenges—such as occlusion and distortion [45]—that adversely affect user experience and task performance [45].

While IA and SA are relatively new areas of research, there are a multitude of existing implementations that could be called "immersive analytics" systems, many of them also situated. Some systems address universally challenging issues, such as labelling [89, 149], efficient highlighting [70], impact of real-world background on visualization perception [203], and synergy of HMDs and handhelds [132]. Yet most SA systems tend to be use-case-specific.

The advent of light, hand-held, multi-functional devices (e.g., smartphones, tablets etc.) has made AR accessible to a larger audience. This, consequently, has aided the emergence of SA systems for the general public, such as for tourism [42], sports [140], entertainment [13] and shopping [69]. In many research and analytical disciplines, the work space is out "in the field" rather than an office or laboratory setting. Indeed, the decision-making processes and situational awareness of manufacturing, construction [24], agricultural [270], and utilities employees who work in the field [207] is an example of an enterprise-facing domain for applications of *in situ* analysis. In this space, Whitlock, Wu, and Szafir conduct a design probe involving expert users from five such disciplines to evaluate the needs and challenges of existing situated analytical systems for data analysis and collection, and demonstrate their resulting design recommendations via their implementation, FieldView [251]. Last but not

least, medicine and medical imaging has been a popular application domain for MR and has has yielded important techniques, such as the cutting plane implemented in this work, which originates in the interactive slicing of brain imaging data [99].

For a more comprehensive study of techniques, Zollman *et al.* [272] present a taxonomy for visualization in AR, based on many such examples, and use it to extend the traditional data visualization pipeline to situated implementations. They identify six recurring design dimensions in their AR visualization taxonomy—purpose (for using AR), visibility (vs. occluded or out of view), depth cues, abstraction (for reducing data complexity), filtering (to a subset of observations to reduce clutter), and compositing (the method for modifying the user's non-augmented view of reality)—and the domains thereof. While my work may be comparable to Zollman *et al.* [272] in that my conclusions are based on a compendium of existing implementations, I focus more narrowly on the view management end of the visualization pipeline.

Most situated analytics systems seek to take advantage of the user being embedded in the same space as data with spatial attributes by presenting them with an immersive integration—i.e., taking up the user's field of vision. If we accept Mackay's definition of AR described in Section 2.2.2, then the means and methods by which objects of interest are networked together and joined to the user's view of reality are inseparable from the experience itself. While those means include commercially-developed platforms such as Unity, Apple's ARKit and Google's ARCore, they also include open source solutions developed and maintained by individuals and research and development communities, such as the Unity-based DXR [219] and IATK [51] toolkits, and the Web-based framework VRIA [37].

### 2.2.4 Interaction and Situated Analytics

Interaction has a central role in visualization despite typically receiving much less attention than visual aspects [260], and this is equally true for immersive and situated visualization [156]. Nevertheless, enabled by technological advances in contemporary immersive technologies, recent efforts have explored novel interaction techniques and device synergies in an SA context [35]. For example, Bach et al. [9] assessed the effectiveness of direct tangible interaction with 3D holograms. They compared the use of the Microsoft HoloLens with fiducial markers in an AR visualization setting to a handheld and a desktop-based setup.

The notion of analyzing user interactions in MR spaces has also received attention. MRAT [171] is a MR toolkit that allows the visualization of usage data of interaction techniques in MR, providing mechanisms for interaction tracking, task definition and evaluation mechanisms, and visual inspection tools with in-situ visualizations. Likewise, Büchel et al. [36] present MIRIA, a toolkit designed to support in-situ visual analysis of spatial temporal interaction data in mixed reality and multi-display environments. MIRIA provides mechanism to depict and analyze movement of users and tracked devices, interaction events as well as to identify issues such as tracking problems or obstructions from physical objects. Flex-ER [144] is a web-based environment that enables users to design, run and share investigations in MR, supporting different platforms and interfaces via a JSON specification of interactions and tasks. In a similar theme of cross-device analysis of MR, ReLiVe [104], a mixed-immersion tool, combines an IA in-situ view with a synchronized visual analytics ex-situ desktop view.

Cross-device synergies have also been exploited towards enhancing the analytical process within SA environments. Butscher et al. [38] investigate synergies between tabletops and AR-enabled HMDs to visualize and manipulate 3D parallel coordinate plots. Hubenschmid

et al. [105] explore similar synergies with tablets, and interactions via touch and voice commands. Reipschläger et al. [195] does this for AR HMDs and touchscreens. Finally, Langner et al. present MARVIS [132], a framework enabling the combination of mobile devices and AR HMDs in AR-based analytical setting. MARVIS allows the depiction of 3D visualizations above and between devices through the HMD. However, these are all one-off designs for specific displays and devices.

Of particular interest to my work are the interaction affordances of toolkits designed for building immersive experiences. DXR [219], a Unity-based IA/SA toolkit, supports multi-visualization workspaces, with interactions (toggles, filters etc.), specified in JSON specification. Interactive elements include tooltips, view manipulation, and configuration controls, and its grammar can be extended to work with other modalities such as tangible, direct manipulation, gesture, and speech input. IATK [51] defines a high-level interaction model that provides filtering, brushing, linking and details on demand functionalities, harnessing GPU power to optimise performance. Building on IATK, RagRug [82] uses data streams from Internet-of-Things devices in SA. RagRug portrays the potential of cross-device connectivity with a visualization pipeline that combines IoT devices, data acquisition via MQTT, Node-RED for filtering, and IATK for visual encoding and rendering in MR. Finally, VRIA [37], although predominately designed to work in VR, also works in AR settings, largely thanks to the ongoing development of the open WebXR specification. Beyond its grammar, which is discussed in the next section, a central aspect is that VRIA is built with web technologies, an approach I take in this work as well.

Interestingly, Besançon et al. [28] point out that new interaction techniques for exploring, filtering, selecting, or manipulating 3D data are often published in non-visualization venues, and thus remain unnoticed by visualization researchers. They also note that leveraging sensing technologies and adapting 3D interaction techniques from other contexts has significant potential for positive impact on 3D visualization. Nevertheless, structured mechanisms for

creating and defining interaction affordances in SA environments, especially for voice and gesture input (such as my contribution in this paper), is much less explored.

### 2.2.5 Mid-Air Gestures and Speech for Visualization

Combining speech and gestures as input mechanisms for controlling virtual objects have long been a vision for the future of computing [95]. In 1980, for example, Bolt [32] combined the two for a large-screen display system, "Put-That-There," allowing a user to combine spoken commands with pointing to populate a room-sized space with 3D objects. However, speech with gesture as input is much less well-explored in visualization and visual analytics systems.

The visualization *of* gestures [78, 113, 117] is a more prevalent theme than the use of gestures *for* visualization. One exception is Proxemic Lenses [10], where collaborators use explicit gestures and implicit body language to interact with large-scale data displays. Another is DA-TU [96], which features a tablet-based multi-finger gestural vocabulary for interacting with objects in a large database. A 2018 AVI workshop [137] highlighted this shortcoming in exploration of input modalities in contemporary visualization research. Even following this workshop, however, the use of mid-air gestures for immersive analytics research has remained uncommon, with one notable exception in Filho et al.'s evaluation of an immersive space-time cube [81].

Considerably more work has been conducted in the area of speech as an input mode for visualization for traditional displays [56, 262], large screens [10], and immersive environments [12]. The Natural Language for Data Visualization (NL4DV) system [170] is noteworthy in that it integrates contemporary visualization tools with multimodal user input and popular analytical tools and workflows. Even the subject of combining speech and touch—not gesture, but touchscreen interaction—has been addressed in visualization

literature [202], with the results confirming that multimodal input is preferred over single modes of input for either speech or touch. However, I argue that *speech and mid-air gestures* have not been used in combination for immersive analytics in the existing literature; this is one of my goals in this paper.

## 2.3 Visualization Grammars and Beyond

Visualization grammars, first introduced by Leland Wilkinson as the eponymous Grammar of Graphics [254] in 1999, provide combinatorial building blocks for specifying visual representations using a concise declarative language. This approach is radically different from the standard chart template galleries used in tools such as Microsoft Excel. Wilkinson's grammar was quickly adopted by the visualization and statistics communities. Hadley Wickham's ggplot2 [253] operationalizes the grammar in the R language. Vega [206] is a low-level explanatory specification expressed as JSON and rendered on the web using SVG or Canvas; upon it is built the higher-level Vega-Lite [205], which facilitates rapidly building interactive visualization without the full complexity of the Vega backend. Several special-purpose visualization grammars have since evolved, including Atom [182] for unit visualizations, Cicero [124] for responsive web-based visualizations, and PGoG [191], a probabilistic extension to Wilkinson's original Grammar of Graphics.

Visualization grammars can also serve as low-level specification backends for higher-level visualization environments. This allows point-and-click interaction rather than textual specification. For example, Tableau (originally published as Polaris [228]) is built on the underlying VisQL grammar. Similarly, the Lyra 2 [273] interactive visualization environment generates Vega or Vega-Lite specifications as output.

Finally, there exists some visualization grammars designed specifically for immersive, situated, and ubiquitous analytics. ImAxes [52] can be said to be one such system, enabling a

VR user to freely combine axes to author various multidimensional visualizations in 3D using direct manipulation. DXR [219] uses a JSON specification language similar to Vega-Lite to author 3D visualizations for immersive analytics in the Unity engine, even providing an interface for modifying the representations while in the immersive environment. VRIA [37] supports a similar Vega-like JSON specification, but is entirely implemented using open web technologies such as A-Frame[10], React, and D3.js rather than Unity. Compared to these existing offerings, in this paper I propose a visualization environment for ubiquitous and immersive analytics based on mid-air gesture and speech interaction. Similar to VRIA, my approach is built on open web technologies rather than a proprietary graphics engine. To the best of my knowledge, this work is the first to allow users to author visualization grammar specifications in 3D mixed reality using such a direct manipulation method.

## 2.4 Evaluation

Given the need for further evaluation and empirical work in IA in general and SA specifically, I must take a broader view of evaluation in the HCI and VIS communities. In the domain of HCI and data visualization, some approaches to understanding factors influencing users' experiences and needs for systems involve constructing personas–representational archetypes of "typical" users and their daily lives [102]. This often involves qualitative and ethnographic methods in which the researcher tracks, records, and interprets the users' daily activities in collaboration with the participant, reaching a shared understanding of the user's thought processes through interview and activity [102, 215], but approaches using observational data, statistical models, and machine learning are becoming increasingly common.

[10] `https://aframe.io/`

### 2.4.1 Cooperative and Contextual Inquiry for Visualization

In their seminal paper, Wixon *et al.* [258] introduce "contextual design" as a systems development method in which the researcher partners with the user at the user's place of work to "develop a shared understanding" of the user's activities, and they define *contextual inquiry* as the first part of the broader process. Specifically, contextual inquiry is the data collection step of the field research element of the contextual design method, and it emphasizes four essential principles: (1) the *context* of the activity being performed by the user, (2) the *partnership* between the researcher and the participant, (3) the spoken verification that the investigator's *interpretation* of the activity matches the user's, and (4) the *focus* of the study as central to the approach taken by the interviewer [29, 102]. The most typical application of contextual inquiry is in the form of a *contextual interview*, which begins in the user's actual work environment as a traditional interview regarding the user's recollections of their work activities, and, within fifteen minutes, is transitioned o an activity in which the participant conducts their work while the researcher watches and takes a participatory role by sharing and summarizing their understanding of the user's work [29, 102].

Cooperative inquiry is a qualitative evaluation method based on an iterative cycle of three primary steps: contextual inquiry, participatory design, and technology immersion [63]. Contextual inquiry is the data collection process in which the researcher and participant form a partnership to reach a shared understanding of the user's experience as part of a broader design study [29, 258]. In my "visualization gap" study [18], I employed contextual inquiry to understand data scientist workflows and their relationship to interactive visualization through in-depth interview sessions.

In participatory design, the user partners with the researcher to continuously develop new prototypes for the implementation. One method that I particularly draw from participatory

design is to embed a researcher with both the users and designers of the system to act as a *values lever*: A link between user and researcher team who is responsible for translating user requests into technical specifications [215]. On an operational level, this is similar to the *pair analytics* approach proposed by Arias-Hernandez *et al.* [6], where a visual analytics expert "drives" the system while a domain expert gives directions.

While alternatives to the qualitative methods for developing personas may be applicable in certain cases (e.g., where mouse events are the most notable method for human-computer interaction) [264], this approach is more difficult to apply in design for the sciences beyond simply categorizing event sequence structures [143]. Field survey methods are still a popular approach for determining the direction of design targeting scientific users [148, 192], including data scientists [23, 119].

Kandel *et al.* [119] conducted what might be considered a contextual interview study similar to my own in that they analyze data scientists' self-reported work processes, and attempted to interview participants at their place of work in as many cases as possible. They propose three main archetypes that data scientists may be classed into: *Hackers*, who build processes chaining together multiple programming languages of different types (analytical, scripting, and database languages, for example) and who use visualization in a variety of environments; *Scripters*, who perform most of their analysis in an analytical environment (e.g., R) and perform the most complex statistical modeling of the types but who do not perform their own ETL; and *Application Users* who performed most or all of their work in an application such as Excel or SPSS and, like Scripters, relied on others (namely, their organizations' IT departments) for ETL. The appropriateness of contextual inquiry for analytical professions in more contemporary research is further evidenced by the recent, complete contextual design study of data scientists [181] conducted by IBM, a notable employer of data scientists.

### 2.4.2 Quantitative Measures in Task Performance, Use of Space & Time

Quantitative analysis of interfaces in the context of data visualization typically revolves around task performance metrics given a type of data [218]. Task performance studies often measure accuracy, correctness, speed, and other measures of how well and how easily the user is able to interpret different types of data: Namely, quantitative, ordinal, and nominal data [226]. Empirical work on graphical perception, from early seminal work by Cleveland and McGill [49], Mackinlay [147], and Bertin [27], to more recent work in visual perception [180, 250] or—in the case of multisensory IA studies—other senses [21], often attempts to determine an internal ranking between visualization techniques or sensory channels and these three data types.

Events within the interface, such as mouse activity [3], may be used to develop "data-driven personas" [265] for specific types of users; platforms for crowdsourcing experiments such as Amazon Mechanical Turk [127] make the creation of these types of personas more manageable at larger scales. With the rise of IA and SA, approaches analogous to this are becoming more popular as a means of characterizing study participants' navigation through space, view arrangement, and time use [17, 22, 204]. These types of studies, whether early or more recent, are often accompanied by qualitative evaluation to provide nuance and identify patterns that were not captured by the quantitative data collected during the study [17, 59, 238].

### 2.4.3 Characterizing User Behavior with Machine Learning

In the machine learning (ML) community, there has been more than a decade's worth of literature exploring methods for action classification [133, 173], motion and path prediction[145], eye tracking [129], and gesture detection [162]. While there have been a few position papers [39] and more serious studies [98] advocating for a closer relationship between the

HCI and machine intelligence communities, the current body of literature on the subject is surprisingly sparse. If a trained ML model can identify individuals' emotions, moods, and expressions [73, 143, 230, 261, 267] and can accurately predict whether a basketball player is good or bad [26], why is there so little work identifying whether a user's experience has been positive or negative, or their task performance good or bad? This section will review a range of applications in computer vision (CV) literature to human behavior, and then provide an overview of HCI work that does make use of CV methods.

#### 2.4.3.1 Deep Learning Models for Human Poses

Considerable prior research has explored the topic of human pose estimation using deep learning in single-person single camera [271], multi-person single camera, and multi-person multi-camera settings [40, 86]. Recent work includes top-down approaches using a 2-stage pipeline with a CNN for frame-level pose prediction followed by a matching algorithm to efficiently link the predictions to specific people[40, 86, 223, 244]. The CNN itself can use a 3D mask as in Girdhar *et al.* [86] to incorporate temporal data for more robust prediction. In my project, I use the pretrained OpenPose model [40] to jointly detect human body, hand, and facial keypoints (in total 135 keypoints) on single frames.

Walker *et al.* [243] tried to address the video forecasting problem by taking advantage of the strengths of Variational Autoencoders (VAEs) and GANS. Instead of solving this forecasting problem directly in the pixel-level space, this paper projects the problem into the human-pose space through the human-pose estimation. In Batch *et al.* [20], my motivation was similar in that I also try to address the action classification problem in the human-pose space, instead of classifying actions directly from videos.

#### 2.4.3.2 Behavioral Coding

Coding behavioral events in video is common research practise in HCI and other fields, often those related to the social sciences [196]. It is largely performed in three steps. First, a coding scheme that describes the categories of actions has to be created via a bottom-up, top-down [245], or a hybrid approach. In a bottom-up approach the themes for the actions emerge from the data itself and are agreed upon by the coders after watching and rewatching of the videos. In a top-down approach, labels emerge from the theoretical literature on human gestures. The second step would be to train some number of coders which takes an amount of time proportional to the complexity of the videos. The final step is to actually label the videos and ensure that the coders are able to label videos in a consistent way which is measured by an agreement metric such as Cohen's Kappa [101]. The codebook might be rewritten in iterations during this process.

Several existing tools have been built to support the video coding process, particularly to help with coder training and video labeling in a systematic way. For example, ANVIL, Datavyu [232], VACA [34], and VCode [92]. There have also been systems in the past that have leveraged crowdworkers instead in the codebook creation and video labeling process [134]. In our system [20], Kyungjun Lee, Hanuma Teja Maddali, Niklas Elmqvist, and I implemented a hybrid approach in which an unsupervised clustering mechanism grouped actions in the data by a measure of similarity related to change in pose. A human in the loop then used knowledge of relevant theoretical models to select potential AOI, either though expected actions or outlier detection. The action detection and label assignment process in my pipeline, however, was completely automated via an action classification model.

#### 2.4.3.3 Unsupervised Video Summarization

Summarization models are probably closer to my objective than any other, but my target is the narrow context of HCI researchers discovering new actions based on user interactions in systems using 3-dimensional body motion and gestures, and reducing the computational cost of model training is a high priority. Mahasseni *et al.* [151] take what might be considered the most contemporary approach to detecting events in video for summarization by using generative adversarial networks (GANs) to detect keyframes—frames marking the end or beginning of transitions in motion—in high-resolution video. In their model, the generative network (summarizer) creates a summary of a longer video in order to trick the discriminator, and the discriminator network is trained to discriminate between the summarizer and the human-summarized video. They use the SumMe dataset, which has short, human-made summaries for a corresponding set of longer videos (1 to 6 minutes in length) [91].

The use of keyframes itself is not a new idea. In fact, as an alternative approach to detecting keyframes, the study that originated the SumMe benchmark dataset used by Mahasseni *et al.* [151], Gygli *et al.* [91] draw from video editing theory in proposing *Superframe segmentation*, a technique that cuts video into arbitrary segments and then shifts the the cuts to neighboring frames with the least motion, as part of a video summarization pipeline. Following segmentation, they evaluate numerous other features of the video—including attention, color, contrast, edge distribution, and object detection (people and landmarks)—and then calculate an "interestingness" score. The interestingness of a segment must meet a predefined threshold in order to be cut into the output of the model, which concatenates the most interesting segments of the video into a short summary.

#### 2.4.3.4 Machine Learning in HCI

Video and audio recordings tend to be a nearly-ubiquitous form of data to capture and analyze during user study sessions. HCI and visualization research communities have already begun to make advancements that shift away from video and audio recordings being an intractable media format, cheap to capture but expensive to analyze in evaluation studies for HCI and visualization, toward the use of video and audio inputs as a revealed behavior dataset that is time-cost cheap and therefore scalable for the analysis of large user populations. Relevant examples include discovering speech patterns [75], identifying gesture [152, 193, 234] and gaze [164, 266], classifying user emotion and facial expression [94, 159, 229], and detecting characteristics of the user, such as gender [240][11], by constructing and implementing neural network architecture. The visualization community has also made contributions to the toolkit of methods used in evaluating user video, logs, transcripts, and other qualitative data [44], as well as user gesture analysis [122]. Systems for visualizing and analyzing visual and semantic features of cinematic films in the context of film studies been implemented, for example, in VIAN [93], which represents information about average frame color to the user, who can then manually segment the video with semantic annotations. Kurzhals *et al.* [130] introduce a system that uses the text of movie scripts to assign semantic labels to frames, which is graphically represented to the user along with motion and other visual frame information in an interactive dashboard that affords user annotation. Pavel *et al.* [184] present a system for automatically segmenting and summarizing lecture recordings and append them with crowdsourced transcripts. QuickCut [236] is a system for fast video editing and annotation that allows audio annotations corresponding to timestamped clip segments to be quickly transcribed, semantically matched, and cut together. Leake *et al.* [135] create a system for automatically generating audio-video slideshows using text and

[11] I note that this approach, like many similar projects conceived with little thought to their sociotechnical impact, are a highly questionable practice.

imagery from written articles.

However, these systems either generate read-only output to be consumed (rather than analyzed) by the user, or they require semantic information that is derived either manually by the user, or via existing scripts or metadata that contains semantic information beyond that which is contained in the video and audio data itself. In the scenario I envision, the visualization or HCI researcher can simply add a video and/or audio recording setup, or turn on the onboard camera of a test computer and mobile or wearable device, to collect additional semantic information about the user's speech and physical actions while participating in the user study. The video and audio footage can then be easily and quickly analyzed using off-the-shelf models, resulting in different data streams (e.g., pitch, speech rate, gaze direction, hand posture, and the output from semantic models) synchronized to the rest of the study telemetrics. All of these metrics can then complement task performance data collected in a user study, to reveal deeper insights about the evaluated system and/or targeted users. The user can then annotate the session recording with their thoughts as they conduct their analysis.

As ML continues demonstrating its potential, qualitative researchers are becoming increasingly interested in adopting ML into their analysis flows. However, they often face challenges when incorporating ML into their analysis. First, although traditional classification and clustering ML methods are helpful for generating additional labels to inform analysis, these labels alone are often not sufficient for addressing human-center research problems. Instead, human-centered researchers need to leverage their skills to make sense of the ML-generated labels to gain a deeper and more nuanced understanding of the data. Second, many ML methods require a significant amount of data for optimize parameters and thus have limited accuracy when dealing with small-scale yet rich-in-meaning human-behavior data. Such challenges have inspired researchers to investigate ways, such as interactive visualizations, to better integrate ML into qualitative researchers'

analysis workflow.

One line of research is to support qualitative coding, which is a powerful yet labor-intensive method. Felix *et al.* [77] designed a visual data analysis tool that integrates unsupervised learning methods to provide suggestions to help researchers progressively code a large corpus of texts. Another challenge that qualitative researchers often face is to resolve conflicts among researchers when analyzing qualitative data. Drouhard *et al.* [62] designed a tool, Aeonium, that identifies potential conflicts in codes created by different coders using ML and highlights the conflicts to facilitate coders to spot their disagreements and resolve conflicts efficiently.

Another line of research is to support the analysis of user interaction data to uncover users' intentions and reasoning processes. Both low-level user inputs (e.g., mouse clicks,drags, key presses [88, 200]) and high-level graphical structures of user interactions [97] are captured and visualized to help researchers make sense of their analytic activity. Moreover, eye-tracking data (e.g., scanning trajectory, Area of interest) have also been visualized to help researchers analyze users' interactions and even predict users' intents [31, 220]. Furthermore, researchers have investigated manually recorded provenance (e.g., user-generated annotations) and developed visual interfaces to uncover hidden sense-making patterns [268, 269]. In addition to using proxy data (e.g., mouse events, eye-tracking data) and manual provenance (e.g., user-generated annotations), researchers have recently begun to investigate think-aloud data, which are generated by asking users to verbalize their thought processes while working on a task, to better understand their hidden thinking process. Think-aloud data have been used to understand analysts' reasoning processes [61, 142] as well as users' interactions [74]. $VA^2$ visualizes think-aloud, interaction, and eye movement data to facilitate the analysis of multiple concurrent evaluation results [30]. Recently, Fan *et al.* built an ML model that predicts usability problems of think-aloud sessions based on users' speech and verbalization patterns, and further designed VisTA to visualize ML's predictions as

well as speech related features on synchronized timelines [75]. In addition, using advanced analytical technologies, several researchers developed systems that detect users' moods and facial expressions to facilitate user experience evaluation [57, 167, 224, 230]. Inspired by much of this prior work, in Batch *et al.* [19], Yipeng Ji, Mingming Fan, Jian Zhao, Niklas Elmqvist and I extended this line of research by considering a wider range of modalities of data extracted from video and audio footage, that are indicative of users' experiences (speech rate, transcripts, gaze direction, facial expressions, semantic actions), to create a more comprehensive visual analytical tool to better support the analysis of users' behaviors.

# Part II

# Part 2: Data Science Workflow to Design Guidelines

# Chapter 3: The Interactive Visualization Gap

We conducted an investigation of how data scientists engage in the early stages of exploratory data analysis (EDA) with an eye toward how visualization, specifically, fits into that process. In this chapter, I describe the methods we used, our results, and the design recommendations we construct based on our findings.

## 3.1 Contextual Inquiry with Data Scientists

In Batch *et al.* [18], we conducted our study as a contextual inquiry [102], where we first interviewed participants to establish their everyday work practice. However, our study deviated slightly from standard contextual inquiry protocols in that we then asked participants to solve specific problems that we provided (instead of using their own datasets). These problems were based on (1) *artifacts* used throughout the participants' work process, including code, databases, spreadsheets, methods documentation, and checklists; (2) on our prior knowledge of data science workflows; and (3) on user feedback gathered during beta testing of an R library developed to aid in the extract, transform, and load (ETL) processing of data from a major producer of economic statistical indicators.

Our motivation for the modification was that we already have a reasonable understanding of current data science practice (e.g., as described by Anderson [4] and Kandel et al. [119]), the practices of our participants based on their organizational artifacts and their feedback, and we were more interested in directing participants towards specific tasks to elicit a better

understanding of the initial exploratory stages of the data analysis process. We believe that inferences about these stages would be difficult to make if participants were instead asked only to walk through routine data product maintenance procedures or to give a verbal explanation of already completed projects. By controlling the tasks and problems to work on, we hoped to eliminate some of the wide variation in tools and approaches that individual analysts may exhibit.

### 3.1.1 Participants

We recruited eight data scientists and economists from several federal agencies in Washington, D.C., USA to participate in our experiment. Five of the participants were male and three were female, their ages ranged from 26 to 50 (mean age: 35.5), and they all had normal or corrected-to-normal vision (self-reported). Six participants had earned masters degrees in quantitative fields, one had started—but not finished—a Ph.D. program in economics, and the remaining participant was in the process of earning a masters degree in economics. The participants' experience in their fields ranged from 4 years to 20 years (self-reported). Participants were screened to be experts in data analysis; all participants reported routinely using data management and analysis operations in their daily work and had several years of experience working on this type of duties. Four of these participants had developed or contributed to the development of interactive data visualization projects.

Once screened, participants self-selected in response to emailed requests for their involvement in our study. The self-selection and small sample size must be acknowledged as a limitation to how representative this study may be, but is not uncommon in field studies involving the entry of researchers into the personal or professional environments of the participants [48, 108, 158]. Similarly, the sample was selected based on their employment with, and roles within, federal agencies, which must also taken into consideration with

respect generalizing based on our results.

### 3.1.2 Apparatus and Locale

All inquiry sessions were conducted in the workplace of the participant and using their everyday computing environment to ensure their familiarity and comfort during the study. The exact computing platform, hardware setup, and data analysis software thus varied significantly between participants. Because of this difference, screen recording tools varied across two organizations; one organization had a preexisting screen recording utility and security settings prevented the use of external screen recording software, and the other participants used a free screen recording application. All participants used pencils and paper provided by the researchers for the sketching activity.

### 3.1.3 Procedure

A single inquiry session consisted of the study administrator arriving at the participant's workplace, collecting informed consent, and then giving a brief background of the study. Significantly, at **no time**—either in recruitment or during the introduction of the session—did the administrator mention the visualization theme of our study. The reason for this omission was to avoid priming and potentially biasing participants with regards to their use of visualization. The rest of the study then consisted of four primary steps:

1. A preliminary interview regarding the participant's work processes and tools used in their work (10 to 15 minutes);

2. A data analysis activity designed to mimic a standard data science workflow [4] (approximately 1 hour);

3. A formative design activity during which the participants were asked to sketch visualizations appropriate to tasks in the preceding analysis activity (20 to 30 minutes); and

4. A final semi-structured interview on visualization in the context of the participant's workflow (10 to 15 minutes).

Each session lasted approximately two hours. After finishing a session, the administrator summarized the participant's findings, asked for clarifications or corrections, and answered any remaining questions.

### 3.1.4 Problem Set

Each participant was asked to pick one of the four questions below to answer using real, public data by the end of the Stage 2 within one hour of making their selection (see the Appendix for more details):

1. "How has the rate of a specific type of crime changed over the last few years?"

    - **Optional:** "What might be causing this change?"

2. "Tell me something interesting about the careers or personal finances (e.g., income, spending habits, or employment) of a particular group of people compared to (an)other group(s)."

    - **Optional:** "Suggest an explanation for your observations."

3. "When and where has a number of major catastrophic events occurred? Do they share anything in common with events you didn't expect to exhibit similar characteristics"

    - **Optional 1:** "How frequently and how long after the fact did people talk about/reported on these events?"

- **Optional 2:** "What was the weather like in the area of the event before and afterward?"

4. "What's been going on with gasoline for the past few decades? Tell me as many things about it as you can."

As noted at the beginning of the methods section, these questions were based mainly on artifacts used throughout the participants' work process (code commentary, spreadsheet notes, process documentation, and so on). Questions were made fairly open-ended so that analysts could use their experience to not only determine how they would answer it, but also to decide what constitutes a satisfactory solution.

### 3.1.5 Data Collection and Analysis

Participant voices and on-screen activities were recorded during each session, and some participants drew sketches which were retained by the researchers. Furthermore, the test administrator took extensive notes of observations as well as discussions with the participants during the session. These transcripts and notes form the primary data collected from the study.

We followed a basic qualitative interview analysis method when extracting insights from these transcripts. We first listened through the audio recordings in their entirety to form a general understanding of the themes and topics of the discussion. We then used the interviews to start coding these themes and topics. While we did not use a formal Grounded Theory approach, we did apply an open-coding scheme and regularly stopped to calibrate and merge codes as needed.

## 3.2 Contextual Inquiry Results

In Batch *et al.*, we reported our results for each of the four different stages of the evaluation: (1) preliminary interview, (2) data analysis using a problem set, (3) formative sketching, and (4) final post-experiment interview.

### 3.2.1 Stage 1: Pre-experiment Interview

With one exception, all participants described their work procedures to largely occur within the context of existing information systems and data structures.

#### 3.2.1.1 Self-Reported Workflows

The work processes reported by all participants began at the point understanding the problem or issue they were addressing in their analyses. Participants all moved on to describing the sources of their data, and all participants described a central component to their work being to join or infer relationships between series across different data stores. Three participants noted that the most frustrating part of their work process is often these first two stages when it required communication with data providers. In describing the methods used, all analysts described a need to extract data from an external source and transform it for use with statistical programming languages (R, FAME, and Python).

Participants described using models of varying complexity in their typical work process; most notably, they mentioned statistical language processing and other information matching and retrieval methods, as well as hierarchical and relational structures. Three participants reported the end of their workflow as generally being the communication of their findings, with the remainder reporting archival as the final stage. Five participants reported recent work projects ending in the completion and deployment of tools for data manipulation or

analysis; the remaining three conducted their analyses using existing tools.

#### 3.2.1.2 Work Focus

All participants had recently (within the last year) conducted independent analytical or development projects for which they were the lead or sole contributor. One participant described his work as consisting of running projects that primarily start from scratch. This participant recently developed a search method for large, unstructured, and highly technical text data that had been accruing for roughly forty years.

The four remaining participants reported that the primary focus of their work was in the context of an existing information system. Three of these had made lasting and substantive methods contributions to the body of data science or analytical systems within their current agencies: one had built a user interface for querying agency databases; another had restructured a complex, hierarchical data structure; the third had constructed a revision analysis tool referencing a node aggregation structure.

#### 3.2.1.3 Self-Reported Tool Use: Revisiting Kandel's Archetypes

In some ways, the results from the study by Kandel et al. [119] are similar to ours (e.g., finding appropriate data, ETL, and integrating datasets from several sources took up a large share of many of the analysts' time). However, in contrast to the findings that lead them to propose their three archetypes, interview question responses from the participants in our study indicate that they invariably straddled the "Hacker" and "Scripter" role; not one of them relied on others within their organization for data ETL (although some reported receiving data from external providers under contract as part of a wider process that involved conducting their own ETL). Perhaps even more importantly, all of our respondents reported performing the bulk of their analyses in a scripting or analytical language and had used

multiple languages on the job. This difference may, admittedly, be a result of our small sample size, but it may also be an indicator that their third archetype, the "Application User," has become passé in analytical professions. Alternatively, it may mean that we have not yet reached a tool maturity where this archetype can become dominant.

In our study, one participant reported mainly using Python, and noted that the SciPy, NumPy, multiprocessing, and glob libraries were essential for recent work, but that a number of additional libraries made their work easier, with the "ujson" library being among their most favored. This participant also made a note of recent work made use of the Python interface for the Stanford Network Analysis Project (SNAP). Four participants reported using R, but only two of these reported using it regularly on the job. Four participants reported developing interactive visualizations using Plot.ly, Leaflet, and $D^3$ [33], among other tools, at least once in the past. Three of these also reported using JavaScript/HTML/CSS infrequently on the job to communicate output from statistical models to colleagues. These same three participants further reported having used Python, but this was mainly used for personal projects (e.g., combining the use of an API of a financial newspaper, a string pattern recognition algorithm, and a text-to-speech function in order to find and produce audio summaries of news related to their interests which they could no longer find the time to read through manually). Five participants reported used Excel and the time-series database and programming environment FAME ("Forecasting Analysis and Modeling Environment") as the primary environment for analysis on the job.[1] For all of these participants, FAME was described as the environment used most heavily for analysis, whereas Excel was described as being used mainly for the purpose of viewing data and communicating analysis results to others.

[1]FAME is a time-series database with many easily accessible APIs and a domain-specific programming language.

Table 3.1: Participant time use and static visualization rate by task types. Participants spent by far the most time in discovering the appropriate dataset to use in answering their selected question. "Static Visualization Rate" in this context refers to the percentage of participations who created static visualizations during their activity.

| **Task** | **Average Time** | **Static Visualization Rate** |
|---|---|---|
| Discovery | 37 minutes | 50.0% |
| Data ETL | 9 minutes | 0.0% |
| Exploration | 14 minutes | 62.5% |

### 3.2.2 Stage 2: Problem Set

Of the eight participants, two partly answered the question asked in the problem set to their own satisfaction, and the remaining six participants fully answered the question. In all cases, the main stage that participants found impediments to their progress was in the "Discovery" stage. Interactive visualization was not implemented at any stage of the problem set activity, but static visualization was used by a majority of participants (Table 3.2.2).

Several participants used interactive visualizations built by others regarding the data they were considering using to answer the problem. We also observed that all participants using programming environments either received syntax error messages or had minor difficulties reshaping the data which required minutes to resolve.

#### 3.2.2.1 Summary of Tools and Visualizations Used

During the activity, one participant used Python without an IDE, three participants used R in RStudio, and five used Excel. For direct manipulation and analysis of the data, three participants *only* used Excel, and two participants *only* used R in RStudio. Of the participants who stated during the interview section that their primary analytical environment was FAME, if any visualization was produced during their session, both the visualization and the analysis itself were done using Excel. None of the participants in this study used any visualization

tools outside of those built into their analytical environments. All participants used the "look at the data" (or "show me the numbers" [79]) approach as primary means of verifying the relevance and completeness of the data prior to communication stage (i.e., looking at the data in whatever format it was stored). The two most experienced users in this study did not use visualization at any stage of the problem set.

#### 3.2.2.2 Discovery

The discovery stage was by far the most time-intensive activity for all participants during the approximately 1-hour-long problem set activity, taking participants on average **37 minutes** to complete. Of this time spent in discovery,

- An average of approximately 22 minutes was spent *reading reference material* (excluding metadata) to find potential causal factors, and to explore statistical methods including syntactical options within analytical environments. The participants referred to a combination of news, academic, and data science blog articles to assist with this stage of their process. Three participants mainly referenced articles, two of whom read online tutorials (e.g., R cookbook), StackOverflow, and R help documentation; of these, one also referred to API documentation and metadata, and the other participant mainly referenced financial news, academic articles, and statistical reports from government agencies. The third of these participants mainly referenced popular press articles and data science blog posts. Two participants made a point of referring to visualizations produced by others in their readings.

- An average of approximately 15.25 minutes was spent *referencing site or API metadata and conducting searches* as a means to find the location of the correct data. One participant spent the large majority of the discovery stage searching and exploring site metadata, and virtually no time reviewing other reference material. No visual

representation of the reference metadata was referenced or created by any of the participants.

All participants exclusively selected government data; one used local government data for crime statistics, while all others used federal government data.

#### 3.2.2.3 Acquisition and Transformation

None of the participants used visualization during this stage. The average amount of time spent on data acquisition (ETL) was approximately **9 minutes**.

- *Data extraction and loading* took, on average, approximately 2.25 minutes, which was skewed upward by a participant who needed to extract several large datasets from a site, and skewed downward by a participant who extracted the data using an API request that took only the amount of time required to write the request function (approximately 10 seconds). One participant used a REST API, and the remaining three exclusively used site download tools.

- Once it was loaded into the analytical environment, *transforming* the data to prepare it for modeling took slightly longer for participants across all environments, taking an average of approximately 7.75 minutes. This process was lengthier in cases where the structure of the source data being used in the model was more complex, and in cases where the data was being manipulated using a programming language, and was skewed downward where Excel was used with minimal transformation.

#### 3.2.2.4 Exploration, Modeling, and Communication

This process took, on average, approximately **14 minutes**. The most complex model attempted was a basic linear regression model. One participant attempted a categorical

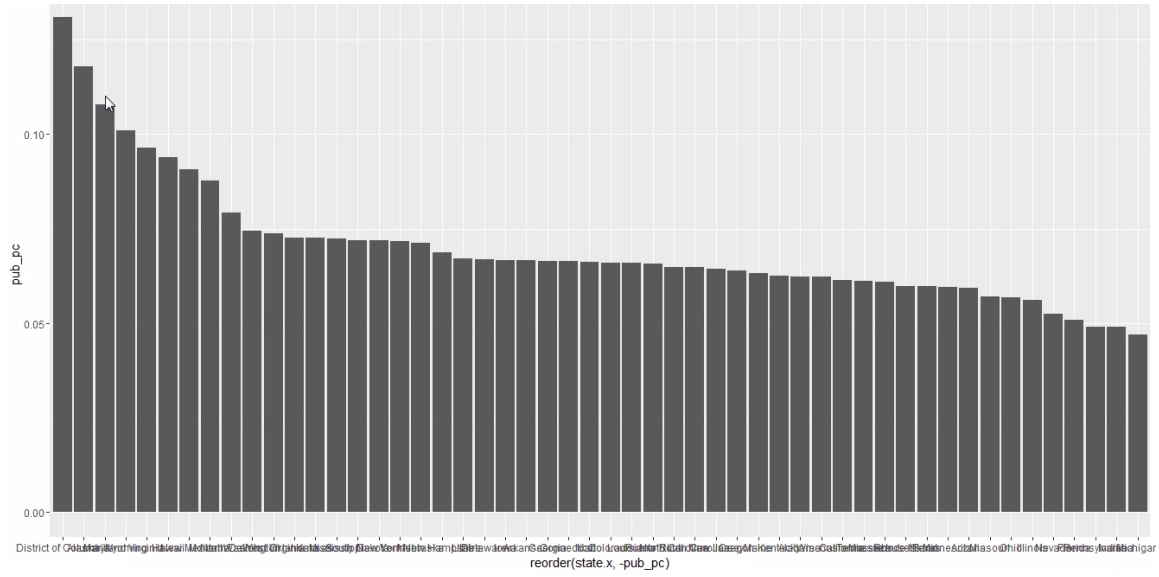



Figure 3.1: One user produced a column chart with U.S. Census Bureau data in RStudio using the ggplot2 library

parent/child aggregation hierarchy, but was unable to finish the analysis. The participant using this hierarchy did not use visualization at any stage of the problem set activity. Of the remaining participants, one examined a cross-section of ratios across geographic categories; this participant produced a column chart comparing public sector employment rates against private sector employment rates by state using ggplot2 (Figure 3.1). This participant also expressed a desire to create a grid of faceted bar charts (also using ggplot2), but decided against it because it would take too long.

One participant examined the rate of change of two potentially related time series with different units of measurement, and produced a line chart comparing the series scaled to different axes to explore the potentially causal relationship. This participant was the only one who used a chart to inform the later stages of analysis, first charting one series and using that information to search the time period of interest, and was the only participant to perform comparative data analysis. The remaining participant examined the rate of change in a single series and produced a line chart representation of the series. All charts used

or produced during this activity were static. All participants who used visualization for exploration used the same charts as part of the communication of their findings.

### 3.2.3 Stage 3: Sketching

As in other studies [46], we opted to a sketching activity to allow for the creation of visualization in instances which may otherwise have been constrained by either technological barriers or the time limitations of our interview sessions. The most common theme in participant sketches of potentially helpful visualizations during this stage was that most participants viewed a table as the *most* beneficial visual aid. Only four of them drew a chart, and in one of these cases, it was mainly as an afterthought. All participants focused on the work involved in data discovery as the most difficult element of the activity, including participants who were already familiar with the source of the data they selected. All participants were most strongly interested in methods for multistage search-and-filter interface design; all participants included either drop-down menus or search bars (or both) in their sketches. Three participants also included tables in their sketches; two of these sketches contained lists of potential data sources, the third contained the data itself (Figure 3.2). One participant expressed interest in a related-data search and discovery tool inside the RStudio IDE.

Of the participants whose sketches extended beyond search-and-filter methods for data discovery, one drew a bar chart representation of a hierarchical time series and expressed an interest in better illustrating the hierarchy. Another participant expressed a desire to represent auto-regression models of the series used during the problem set activity, and noted that it would have been easier for them to do using Stata. A third participant, who we consider to have the most experience in developing interactive visualizations within the study cohort, incorporated interactive elements within his sketch as a small window which

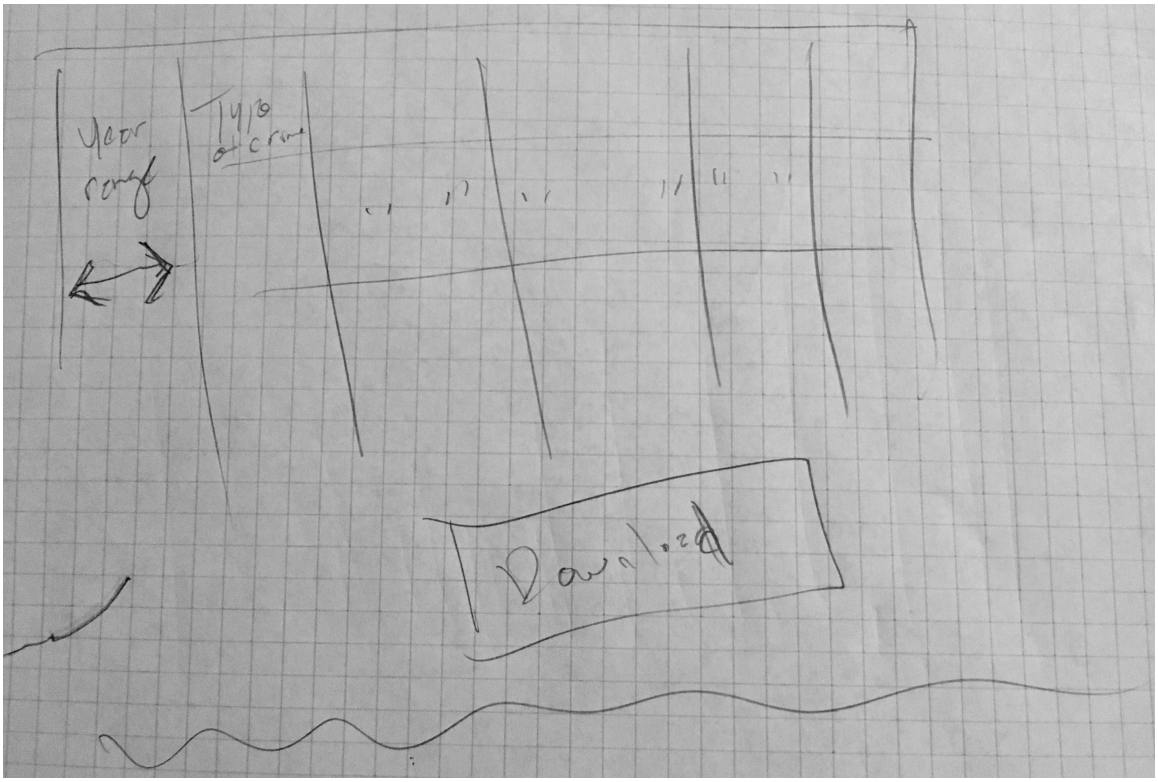



Figure 3.2: The fifth participant simply sketched a data table and, not without sarcasm, added a "download" button.

appears on mouse-over (i.e., a tooltip) with details regarding data linked to a visual object within the view (Figure 3.3).

### 3.2.4 Stage 4: Post-experiment Interview

When asked about reasons for not using visualization, three recurring themes arose during the post-experiment interviews: (1) Visualization was too time-consuming to be worth their effort, (2) numeric data provided more detail in many instances than visualization could, and (3) visualization was just not needed.

#### 3.2.4.1 Not Enough Time

Five of the eight participants stated that visualization is important, but that they did not have time to do it often. One participant said that only one of their projects, not a routine part of their work process, involved visualization in order to check the accuracy of predictive models. This participant said that building visualization into their typical workflow was difficult due to time constraints. When asked about the tools they typically use for visualization, they responded that use of Excel was most common, but that they have used Stata, Eviews, and R for visualization as well in their free time or as a student. Regarding ggplot2, one participant remarked: "The syntax just doesn't feel right[...] to come up with one beautiful graph, if I put it in a nice block format, it would be like fifteen additional lines. To me, that seems superfluous. I also don't like this syntax—using 'plus' signs between each line. R's syntax is more functional—traditional functions have commas, all within the same parens; I understand that maybe the philosophy is that you have to be explicit about [features...] but that seems like overkill." We found this emblematic of the guidelines we propose: It is not enough to build tools for interactive visualization, or even to port them to the researcher's environment—we must also make it syntactically familiar,

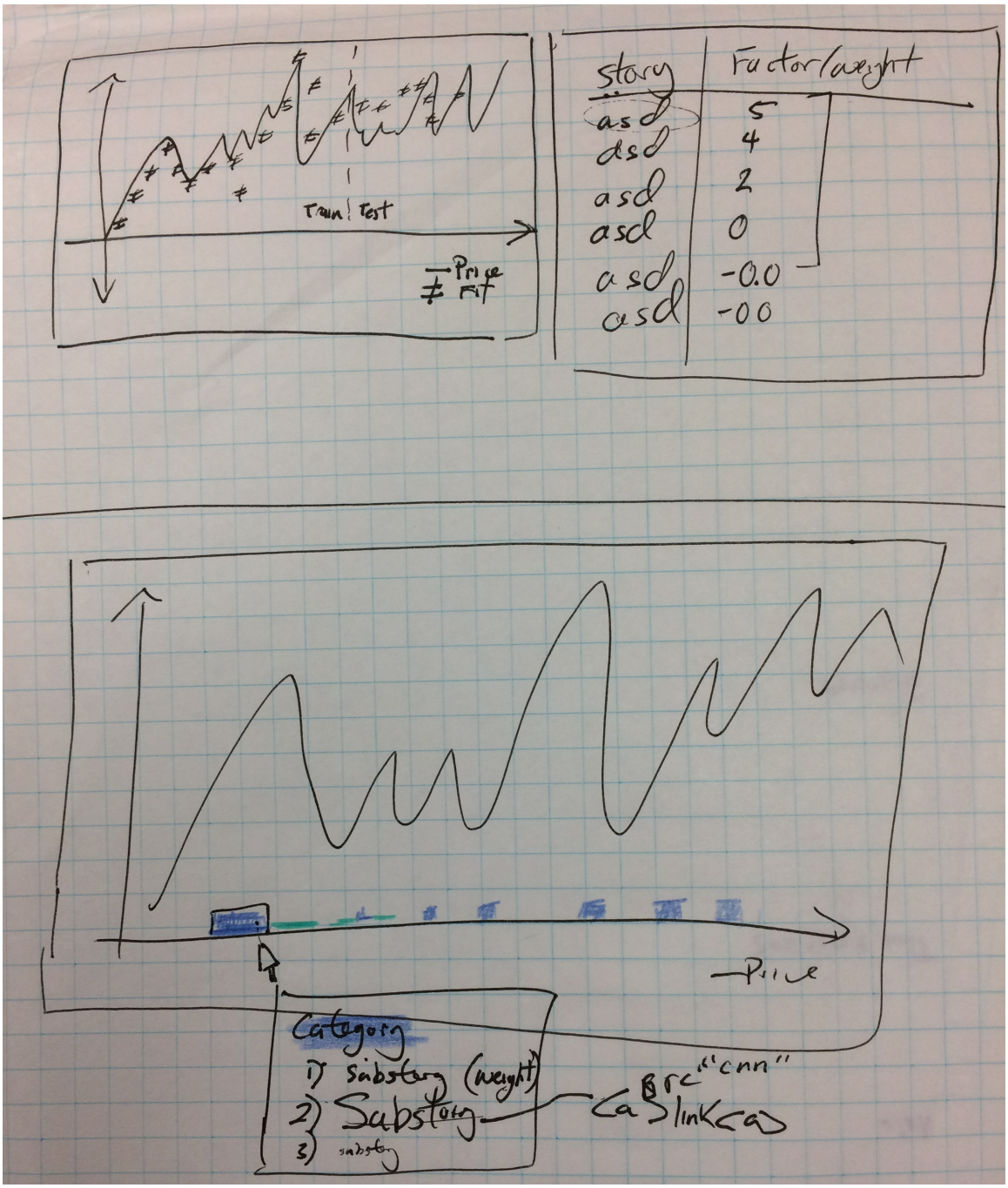



Figure 3.3: Interactivity appeared only once in our study, in a sketch; this indicates that the desire to build interactive views is present within the data science community, but the costs of using the tools outweigh the need during initial exploration.

concise, and convenient to use within that environment.

#### 3.2.4.2 Show me the Numbers!

One participant said that they occasionally use a line graph to track rates of change, but that they typically just look at the numeric representation of a time series when checking for volatility or revisions, as they find it clearer and more accurate than the line chart. This participant noted, however, that representing thresholds or other important characteristics by changing the color of the number or background was helpful.

#### 3.2.4.3 Visualization is Unnecessary

Five participants noted that the data was straightforward enough that there was not a strong need to visualize it, and one of these, along with one other participant, noted that familiarity with the conceptual context of the data coupled with a quick examination of the numeric data was sufficient for their purposes. One data scientist stated that they virtually never used visualization except to communicate their findings with others, and during the post-activity interview, noted that the exception to this was in cases where data was either structurally complex (e.g., representing networks), or when it was intrinsically spatial.

## 3.3 A Discussion of the Visualization Gap

In Batch *et al.* [18], we analyze the results of a contextual inquiry conducted with eight data scientists and economists using qualitative methods and derived both expected and surprising findings. More specifically, our results confirmed that visualization was primarily seen as a communication tool among professional analysts and that few of our participants ever use visual representations of their data in the middle of an analysis process. The reason stated for this was that visualization tools are generally seen as endpoints in the process

in that they (a) are separate from the computational tools that data scientists typically use (R, Matlab, SPSS, JMP, among others), (b) require extensive data wrangling [118] to use, and (c) provide poor functionality for exporting insights, operations, and filters used in the visualization. We did note that our participants have a quite pragmatic view of the use of visual representations; visualization is just yet another tool, and they claim no intrinsic bias against its use if it provides clear utility. This represents a promising opportunity for the visualization field provided that our tools can be better integrated in data science workflows.

Our results highlighted the quandary of visualization in professional data science: visualizations—including static visualizations—were rarely seen as obligatory or even useful components of the initial analytical process, and were instead relegated to the final checking and dissemination stages of the process. In other words, a dynamic visual representation was considered a good tool for communicating results with a lay audience, but was not considered vital when trying to understand which results to communicate in the first place. Based on our conversations with the data scientists involved in our contextual inquiry, we outlined a few measures that the visualization field should focus on:

- For visualization scientists collaborating with data scientists, **use the same programming environments and syntax that they do and build visualization elements into "data discovery" libraries**, creating or tying together data ETL tools that can be used in a non-interruptive step within the analytical environment to facilitate sensemaking. Sensemaking is often described as a cognitive skill requiring human intervention [80, 90, 115, 201], and libraries within statistical environments are nothing if not artifacts of data scientists' efforts to simplify that process for their peers.

- **Conduct user experience (UX) design sessions with data scientists** to investigate ways to soothe the frustration of errors and data foraging. All of our participating data scientists noted that the user experience of their most commonly used tools left

much to be desired. Unfortunately, given their small population size and because of the haphazard and highly personal data science process, not enough attention has been spent on this topic.

- **The verdict on data tables: Not bad**. Participants of this study gravitated toward the data table format as their visual representation of choice, and every single participant viewed the data in a tabular format. Those using Excel, which links chart creation with table views, were able to more quickly and successfully visualize their data; however, many of these users expressed a degree of embarrassment at resorting to Excel. Those using R or Python either did not attempt to visualize, or found the syntax to be inconvenient. Bridging visualization and data science may require visualization researchers spending more time on augmenting basic representations such as tables with additional functionality rather than designing entirely novel visual representations.

- **Design self-contained, visualization components** that can integrate into the command-line interfaces that data scientists routinely use while still allowing for full-fledged interaction (zooming and panning, filtering, details-on-demand, etc) [218]. The syntax of calling the components must match that of the target environment; for instance, calling visualizations using single-line functions with parenthetical variables and specifications was a feature more than one of our respondents mentioned finding desirable. Furthermore, these visualization components should be first-class members of the analytical process so that actions and transformations interactively performed in the component can be exported and passed on to the next component in the sequence.

- **Education, not evangelization** is what is primarily needed to improve visualization adoption within data science, including providing easily accessible galleries of useful visualization techniques based on data type and tasks, giving examples of best prac-

tices, and finding allies within the data science community who can evangelize on our behalf.

# Chapter 4: Evaluating Performance, Space Use, and Presence in Immersive Analytics

Having evaluated the exploratory data analysis process of data scientists in their typical work environment, we wanted to pivot toward studying exploratory data analysis in immmersive environments. To do this, we conducted a multi-phase user study with economists and statisticians using an existing IA tool, ImAxes [52].

## 4.1 Mixed Methods for Immersive Evaluation: Supervised and In-the-Wild

In Batch *et al.* [17], our study involved four main phases (Figure 4.6): a pilot study (P), a formative "in-the-wild" phase (F), and two in-depth phases (S1+S2).

### 4.1.1 Setting and Participant Pool

All phases of the study were conducted at a U.S. federal agency where one of the authors was embedded. The participant pool for all experiments thus consisted of data scientists, economic analysts, and economists employed, interning, or contracting at this agency. Overall, the education level was high among our participant pool, with all participants having advanced degrees in economics (collectively, 6 master's degree and 3 Ph.D.s), statistics/mathematics (1 master's, 2 Ph.D.s), public policy (2 master's), political science (1 master's, 1 Ph.D.) or similar domains. Participants in our individual in-depth experiments were screened to be experts in data analysis; they routinely used data management and

analysis operations daily and had several years of experience working in this duty.

### 4.1.2 Apparatus

All studies were conducted in a small office of approximately $10 \times 10$ feet ($3 \times 3$ meters) dedicated to this study. The computing equipment was a personal computer equipped with a Nvidia GeForce GTX 1060 (6GB) GPU, Intel Xeon E5-2620 v3 (2.40GHz) CPU, and 16GB RAM, and running Microsoft Windows 10. The rig was equipped with an HTC Vive VR system, including a head-mounted display (HMD), two base stations, and a monitor that enabled the experimenter to observe the viewpoint of the HMD. The ImAxes application was built using Unity 5.6.5f1. Additional evaluation of video and telemetry data was conducted using a PC equipped with an EVGA GeForce GTX 1080 SC (8 GB) GPU, Intel Core i7-7700 CPU (3.60GHz, 4 cores), and 24 GB RAM, also running Windows 10.

### 4.1.3 Data Collection

Here we review the data collection methods employed across all studies. We use the identifiers P (pilot), F (formative), S1, and S2 (summative 1 and 2) to match collection methods to specific phases:

**Demographics Survey (P, S1, S2):** We began our sessions by introducing the study and gathering demographics information. Specifically, we used a written survey to inquire about their past use of VR, their past use of visualizations, gaming experience, and their professional and academic experience.

**Telemetry Recordings (P, F, S1, S2):** The software was instrumented to record controller and headset tracking data over time. The system also recorded specific interactions, such as grabbing and manipulating axes, creating visualizations, and selecting data.

**Video Recordings (P, F, S1, S2):** Two Raspberry Pi Zeros with 8MP Pi cameras and with MotionEyeOS served as webcams set up to capture the interaction space whenever the software was active. One Raspberry Pi was positioned at chest height directly in front of the user's starting position, and the second was positioned in a top corner of the room, a location it shared with one of the Vive's base stations.

**Screen Recordings (S2):** Screen activities were captured using the Windows built-in screen recorder from the game bar. These showed the virtual environment from the participant's viewpoint.

**Audio Recordings (S1, S2):** We recorded participant think-aloud utterances during in-depth sessions using a mobile device.

**Exit Interview (S1, S2):** We ended sessions with a survey and an open-ended interview; answers were recorded and transcribed.

### 4.1.4 Common Procedure

Users signed in to use the device. Users were only permitted to access the system if they were formal participants of the study who had signed the consent form. Participants were verbally informed that their activities would be recorded even if the researcher was not present during their use of the implementation. At the end of the study, participants were asked to complete an exit interview and a survey. The procedure for S1 and S2 in particular is given in detail in the study preregistration: `https://osf.io/53e7n/`

### 4.1.5 Data Analysis

Our collected data was analyzed with several common methods across the different phases. Here we describe these methods in detail.

#### 4.1.5.1 Visualization of Spatial Activity

The tracked 3D telemetry data over time provides important insights in how participants move around (physical navigation), interact with 3D objects (axes and visualizations), and arrange their space. To best analyze and present this data, we aggregate movement data over time into a projected 2D grid of the space. We use a top-down view to study physical navigation as well as spatial arrangements of views and axes (heatmap), and a side view to explore interaction heights (histogram).

#### 4.1.5.2 Replaying Participant Sessions

By combining telemetry data and interaction logs, we are able to replay individual participant sessions. This allows us to understand the participant's view of the analytical space at any point in time. This ability to replay sessions is useful for understanding dynamic behavior and to recreate the arrangement of the space at different times.

### 4.1.6 Formative: Pilot and "In the Wild" Studies

Deploying a novel technical intervention in a new environment typically requires careful customization [212]. Prior to actually evaluating the utility of IA for economic analysis, we thus conducted a month-long formative study that included a pilot study (2 weeks) and an "in-the-wild" deployment (3 weeks). We opted to use the ImAxes platform [52] for immersive multidimensional visualization as our starting point; see the next section for

details.

An added benefit of this formative approach is that it allowed us to continuously iterate on the design based on results from the user sessions as they occurred throughout the duration of the study. Participants were updated on notable changes to the system as they occurred and were asked to engage in additional tutorial, challenge, exploration, and interview activities following each major change to the system.

**The ImAxes System** ImAxes [53] is an IA system based on the concept of *embodied axes* to let users build data views in a 3D virtual environment. Each axis corresponds to a dimension in a multivariate dataset. Users define visualizations by positioning axes in the 3D space, a spatial grammar producing specific visualizations based on their layout.

The basic operations consist of combining two or three orthogonal axes, which produces 2D or 3D scatterplots, respectively. Axes arranged in parallel to each other yield a parallel coordinates plot [111]. More advanced operations consist of stacking axes at the extremities of the axes of an existing scatterplot, which extends 2D and 3D scatterplots to scatterplot matrices. ImAxes also uses the proximity between visualizations to create linked 2D and 3D scatterplots.

**Pilot Study** During the initial pilot study, we invited 6 participants to use the ImAxes platform for hour-long individual sessions with ImAxes left unmodified from its previous incarnation apart from the inclusion of an embedded tutorial. The dataset used during the pilot was the classic cars dataset [64]. The purpose of the pilot study was to: 1) identify the new features to add to ImAxes, 2) calibrate our data collection mechanisms, and 3) determine the datasets participants wanted to view.

**"In-the-Wild" Study** After having established a working baseline system, we launched an "in-the-wild" formative study where the equipment stood available for a full three weeks

for anyone to use at their own discretion. The author embedded at the agency advertised the study via the agency intranet, encouraging interested volunteers to bring their own datasets to explore. The room was kept unlocked, basic documentation was made available in the room, and the software configured to allow new participants to sign in and load their own data. However, while no experimenter was present during these sessions, IRB regulations required us to collect signatures of informed consent from volunteer participants. This allowed us to record video and telemetry whenever the equipment was in use. Similar to the pilot, the purpose of this study was to collect data for how to customize the system for an economist audience.

A total of six participants were engaged in this formative study (all provided signatures of informed consent; no unauthorized person used the tool). They logged a total of 3.8 hours of use in ImAxes during this phase. Figure 4.6 outlines the significant findings from our review of the logged data: this includes several observations that lead to refinements, as well as direct feature requests by the participants.

Throughout the three weeks of the formative study, we rolled out new features as soon as they were implemented, essentially using the field deployment as a "living laboratory."

### 4.1.7 Improvements to ImAxes

The original ImAxes system lacked many features necessary for an economics setting, including some that aid users regardless of domain background. We thus extended the system with additional features to support general use improvements to visual exploration and analysis of data based on feedback from economists. Below we list the main features added (labels refer to Figure 4.6).

▸ **DR1: Tooltip (Details-on-demand).** We implemented details-on-demand as a tooltip for 2D and 3D visualizations using a pointer metaphor (Figure 4.7). By pressing a button

on the controller and pointing in the direction of a 2D visualization, the data values of the nearest point are shown in a 2.5D box with a leader attached to the point. To obtain details-on-demand in a 3D scatterplot, a small pointer sphere is attached to the VR controller that can probe the nearest values.

➤ **DR2: Time-series data.** The original ImAxes supported only scatterplots and parallel coordinates plots. Since many of our users wanted to explore time-series data, we added line graphs as well.

➤ **DR3: Visualization design menu.** We added a simple menu control panel attached to the VR controller. This allows users to remap data dimensions to axes, create and bind a gradient colour to a continuous variable, and map the size of the points or lines to a data attribute.

➤ **DR5: Add mountain range backdrop.** Participants disliked the space's flat, sharp horizon and featureless terrain, causing us to add a mountainous landscape in the distance.

➤ **DR6: Axis selection.** Vanilla ImAxes used a *shelf metaphor* for selecting axes, where data axes were arranged in rows like books on a bookshelf.[1] While this metaphor is easy to understand, it does not scale with the number of axes and requires a large amount of locomotion (walking or teleporting). Our first solution, a "Rolodex" (DR4), was poorly received. Instead we implemented a rotational menu based on a "Lazy Susan" metaphor. The menu can be rotated like a Lazy Susan via the controller touchpad. An axis is selected by pulling it out of the Lazy Susan menu. Thus, in the end, there was no shelf.

➤ **DR7: Grouped selection.** Based on user feedback, we implemented a group selection mechanism that allows the user to move a linked group of visualizations instead of a single one. This enables the user to arrange visualizations around them without breaking links.

[1] We need axes, lots of axes. (`https://youtu.be/5oZi-wYarDs`)

### 4.1.8 Summative: Case Studies in Economics

To understand the utility of IA for professional analysts and data scientists, we conducted a contextual inquiry using our ImAxes tool in case studies involving participants from one of several bureaus of the U.S. federal government. This part of the study was split into two phases: Summative 1 (S1) and Summative 2 (S2). participants during S1 used a version of ImAxes that was slightly different in S2 (Figure 4.6). We report on both below, highlighting differences when needed. Unfortunately, a software error precluded collection of axis position data from S1. Other data was collected from both groups.

Table 4.1: Phases and datasets for summative participants.

| # | Job Title | Yrs exp | Education | Phase* | Dataset† |
|---|---|---|---|---|---|
| 1 | Economist | 12 | M.A., Economics | S1 | D1+D2 |
| 2 | Economist | 2 | M.A., Economics | S1 | D3+D4 |
| 3 | Economist | 9 | Ph.D., Economics | S1 | D1+D5 |
| 4 | Economist | 6 | M.A., Economics | S1 | D2 |
| 5 | Economist | 4 | M.A., Economics | S1 | D6 |
| 6 | Econ spec. | 8 | M.A., Int. Business | S1 | D6 |
| 7 | Economist | 2 | M.A., Econ/Public Policy | F, S2 | D7 |
| 8 | Economist | 5 | M.A., Public Policy | S2 | D8 |
| 9 | Statistician | 3 | M.S., Statistics/Math | P, S2 | D7 |
| 10 | Economist | 9 | Ph.D., Economics | S2 | D8 |
| 11 | Economist | 13 | M.A., Economics | S2 | D8 |
| 12 | Economist | 3 | Ph.D., Economics | S2 | D8 |

* P = pilot, F = formative ("in the wild"), S1/2 = summative 1/2

†Dataset labels in Table 4.3.

### 4.1.9 Participants

We recruited twelve participants (six in each phase) with expertise in economics, statistics, and data science. The participants were all employees at a U.S. federal agency with job

descriptions that include data analysis, all with 2–12 years of experience ($M = 6.21, SD = 3.93$) and graduate degrees in economics or related fields (Table 4.1). They had significant experience in using data analysis tools in their daily work (Table 4.2). While outside the scope of this study, the typical workflow in government and industry data analysis is described by Batch and Elmqvist [18] and Kandel et al. [119]. Six participants had used VR previously, and five participants routinely played video games (1+ hr/wk).

Table 4.2: Count and context of participant use of specific tools.

| **Environment/Language** | **Ever** | **Work** |
|---|---|---|
| Graphic analytical env. (e.g., Tableau, Excel) | 12 | 11 |
| Statistical lang. (e.g., Stata, R, SAS, Julia) | 11 | 10 |
| DBMS (e.g., SQL, PostGres, dBase) | 8 | 6 |
| Econometric DBMS (e.g., FAME, Aremos) | 5 | 5 |
| Markdown/doc-creation (e.g., HTML, LaTeX) | 6 | 4 |
| Object-oriented lang. (e.g., Python, JS) | 7 | 3 |
| Imperative lang. (e.g., FORTRAN, Pascal) | 1 | 1 |

### 4.1.10 Procedure

Our study consisted of several stages: preparation, tutorial, exploration, presentation, and post-session interview.

**Preparation:** Before participants even appeared at the study session, we asked them to send suitable datasets (Table 4.3) that we could integrate into ImAxes prior to the study. Some of these datasets caused us to make changes to the tool itself, such as the axis selection metaphor, as described in Section 4.1.7:DR6, which we implemented to accommodate a larger number of variables than practical in the shelf layout.

Table 4.3: Datasets used by summative participants.

| # | Dataset Name |
|---|---|
| D1 | Compensation by State and Industry |
| D2 | Nominal PCE by State and Industry |
| D3 | International Trade: Services |
| D4 | U.S. Military Spending |
| D5 | BLS Consumer Price Index |
| D6 | National PCE Price Indexes |
| D7 | Blended Health Care Satellite Accnt/Capita Exp. Index |
| D8 | Nominal PCE by State |

**Instructional Tutorial (10 mins):** We began by familiarizing the users with ImAxes via a tutorial embedded in the system. Pre-recorded interactions were played, and the user was prompted to follow along to learn how to use the tool. During this stage, the affordances of a single axis were exhaustively demonstrated before moving on to two axes, then three, and finally SPLOMs and parallel coordinate plots. After each feature was demonstrated, we asked participants to use that feature in a sample dataset. Before finishing the training, participants were encouraged to freely explore the sample dataset while verbalizing their thought process using a think-aloud protocol.

**Exploration (30 mins):** Now participants were set free to explore their own dataset on their own. Exploration was structured as a sequence of iterations, each no less than five minutes, and started with giving the participant the option of introducing a new dataset if desired. For each iteration, the researcher prompted the participant to maintain the think-aloud protocol, and would gently inquire about their motivations throughout the duration. The goal of each iteration was to generate at least one insight and corresponding visualization. Participants were told that they would be expected to present their findings, and were regularly updated on remaining time.

**Presentation (30 mins):** Finally, the participant was asked narrate their findings as if they were presenting their analysis to an external party (the experimenter). The participant was reminded that the experimenter could see what they saw on a monitor, and was asked to create at least one distinct visualization for each point in their narrative. They could use speech, gestures, and ImAxes itself to tell their story.

**Post-Session Interview:** Immediately after the exploration activity, the researcher and participant engaged in a brief, semi-structured interview and survey to (a) validate the researcher's understanding of the user's motivations for their actions during the exploration activity, and (b) Evaluate the user's sentiment regarding the existing iteration of the implementation, including features that they felt were lacking.

## 4.2 Findings in IA with Economists

Table 4.1 reviews the participants and their datasets. Below we discuss a representative use case derived from the experiment. We then present the performance and subjective results.

### 4.2.1 Representative Use Case

The following scenario is a pastiche based on our observations of participants as they explored and presented insights from their macroeconomic data. It is not a description of a single user session; rather, it is a collection of real observations from multiple sessions organized into a representative, narrative summary. In other words, unlike the scenario in the introduction, it is not fictional; these events all happened. The scenario begins with our economist "Sasha" loading their regional personal consumer expenditures dataset into ImAxes. Sasha has just found this dataset from a public source and wants to explore the

2007–2009 Great Recession's effect on trends in consumer expenditures.

SASHA dons their VR headset, launches ImAxes, and grabs three axes from the Lazy Susan. They build a 3D scatterplot of TimePeriod $\times$ Goods $\times$ GeoFips (states) by first holding TimePeriod and Goods orthogonal to each other, then placing GeoFips orthogonal to the scatterplot's origin. They orient the visualization so that they are looking down the temporal axis, leveraging the depth perception afforded by VR to provide a view of the states where the goods have trended higher over time. Using this view, they activate the details on demand using the controllers for these states to obtain numeric values of points along the axes.

Sasha then flips the view so that they are looking at TimePeriod from the side, and points out the general upward trend in total consumer spending for all commodities over all time periods except the Great Recession around 2009. They create a 2D scatterplot of gasoline expenditures over time, noting that the trend is less stationary (i.e., has greater variance over time) in that particular commodity than in others.

Sasha creates a 3D scatterplot of Food Services $\times$ TimePeriod $\times$ Off Premises Food and Drink. Grabbing another Time Period axis, they switch from a 3D scatterplot to two separate 2D scatterplots, which they stack on top of each other. They observe that there is a switch from spending on restaurants (Food Services) toward spending more on groceries (Off Premises Food and Drink) during the Great Recession.

Once they have constructed all of charts they intend to discuss with their colleagues, they arrange them in the space in a linear order from left to right roughly corresponding to the narrative order they plan on following, a little like a museum or gallery of artifacts. As they discuss each point, starting with the most aggregate commodity bundles and drilling down into more detailed commodities, they dynamically interact with the visualization with one or both hands, shifting it for a different viewing angle with one hand and calling the tooltip with the other hand to give their expert audience the detail they would otherwise demand. When they are done discussing the points related to one visualization, they walk to the right to begin their next talking point, until they have run through all of the economic trends they wish to discuss.

### 4.2.2 Explore Stage

Participants spent between 4 and 10 minutes ($M$ =4:33, $SD$ =1:50) in the explore stage. All participants would begin the stage by facing the Lazy Susan within arm's reach, and would rotate it until they found an axis they recognized from which they could start exploring the data. Participants would then often rotate their body away from the Lazy Susan to create a work space by building basic 2D and 3D scatterplots. Figure 4.1 shows that most participants stayed in one place and arranged views egocentrically (**E1**). However, none utilized the full 360° space.

This behavior of recycling the views and axes in their workspace instead of physically moving to a new workspace also supports prediction **E1.2** (participants would arrange their views within easy reach).

To examine prediction **E1.1**—that participants would arrange views at roughly chest

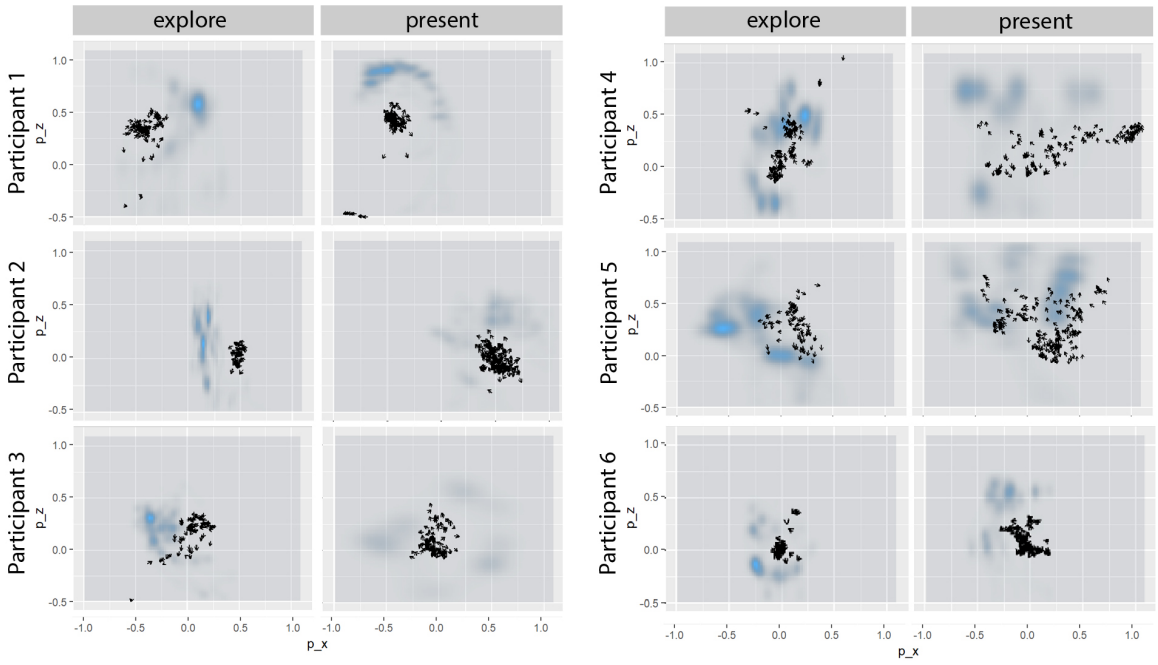



Figure 4.1: Heatmaps of axis interaction in S2 (top-down). Participant position and view direction is represented by a direction arrow.

level—we studied interaction patterns w.r.t. height. Since our tracking data only includes headset position, we estimate chest height to be approximately 30 cm below this position. Figure 4.2 shows a histogram of these relative interactions. Interaction above eye level often occurred when building scatterplot matrices, highlighting a physical limitation of the ImAxes systems (that a user must be able to reach the ends of a scatterplot matrix). While this limitation is somewhat mitigated by design refinement **DR7** (grouping mechanism), these observations suggest that the issue is still present.

Participants would often discard axes and visualizations while exploring the data, maintaining only one to two visualizations at a time (supporting **E2**). Essentially, participants were recycling their views and continuously cleaning their space. Furthermore, we observed that certain types of visualization would be more transient than others. Notably, linked visualizations, whether between two axis or between an axis and a scatterplot, were created

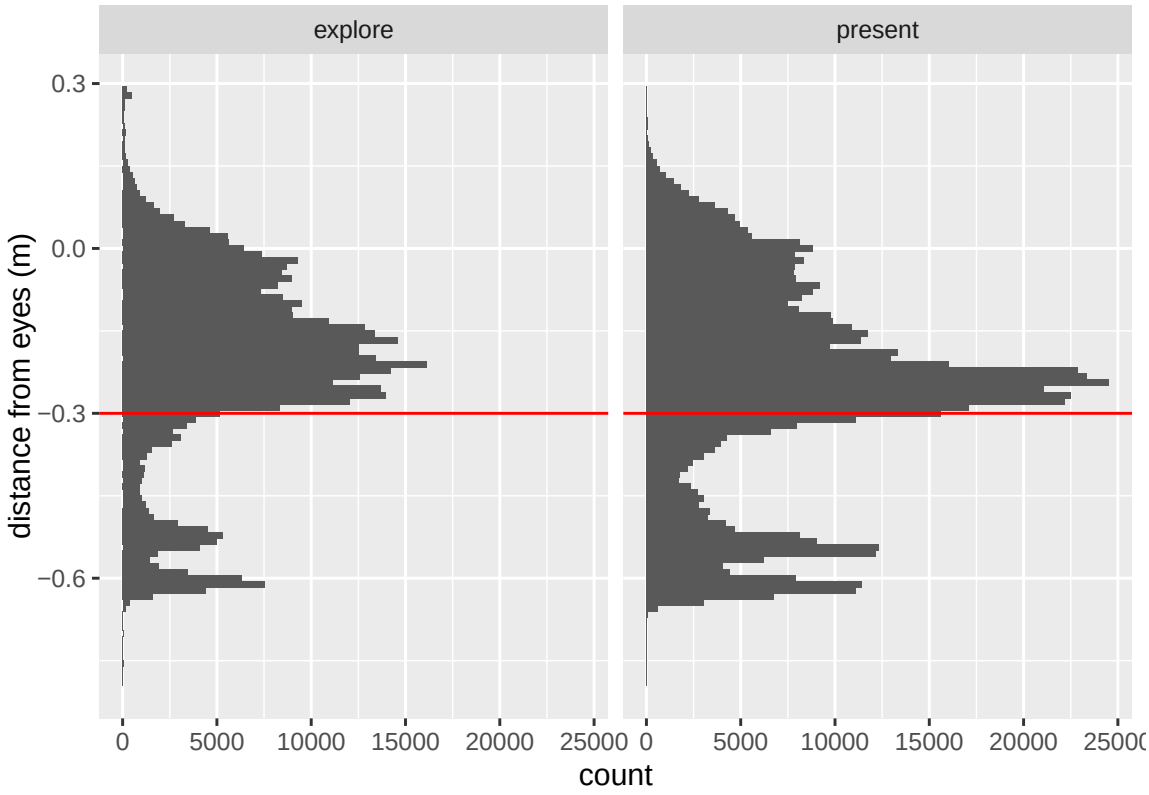



Figure 4.2: Histogram showing the vertical distance of participant interactions with axes relative to their eye level. Eye level is at 0, and the approximate chest level is represented by the red line.


Figure 4.3: Macroeconomics analysis in the ImAxes immersive analytics tool [52]. (Photo by Samuel Zeller on Unsplash.)

and used more than any other type of visualization, but the majority existed for less than five seconds.

### 4.2.3 Presentation Stage

Participants spent between 7 and 11 minutes each ($M$ =6:30, $SD$ =2:30) during the presentation stage. Most participants chose to organize their views in either a linear or semi-circular layout. For example, Participants 4 and 5 placed a series of visualizations in a left to right "narrative order" (Figure 4.1). This somewhat supports prediction **P1** (participants will arrange the views in an exocentric way). However, as can be seen in the "present" columns of Figure 4.1, these arrangements were not strictly exocentric, but remained egocentric (undermining **P1**). We belatedly realized that since the experimenter—the intended audience of the presentation—viewed the 3D space through the eyes of the participant, there was no incentive for the participant to organize the space in an exocentric fashion. However, we did find support for views being arranged in chronological order (**P1.1**); Figure 4.3 shows snapshots of several final view layouts.

We predicted (**P2**) that participants would build more complex visualizations during the

Table 4.4: Count of view creations per participant, split into exploration **(E)** and presentation **(P)** stages.

| | 2D Scatterplot | | 3D Scatterplot | | SPLOM | | Link | |
|---|---|---|---|---|---|---|---|---|
| Participant | **E** | **P** | **E** | **P** | **E** | **P** | **E** | **P** |
| 1 | 8 | 17 | 10 | 2 | 29 | - | 111 | 113 |
| 2 | 9 | 33 | - | 12 | 5 | 5 | 116 | 201 |
| 3 | 27 | 74 | 6 | 16 | - | 26 | 103 | 4136 |
| 4 | 7 | 43 | 1 | 11 | 8 | 18 | 123 | 516 |
| 5 | 5 | 58 | 2 | 15 | 6 | 22 | 31 | 195 |
| 6 | 3 | 49 | - | 23 | 2 | 35 | 527 | 2866 |

presentation stage as they would spend time to carefully craft a meaningful visualization. This is also supported by our data; Table 4.4 indicates that most scatterplot matrices were used in the presentation stage. All participants except Participant 1 created a scatterplot matrices while preparing for presentation stage; however, only Participant 3 actually used a scatterplot matrix when presenting their data. Five of the six participants explored the data using parallel coordinate plots. However, it is worth noting that during the presentation stage, only Participant 3 used a parallel coordinate plot.

### 4.2.4 All Stages

We captured view creation events for 2D and 3D scatterplots, SPLOMs, and linked views. These events are summarized in Table 4.4. Contrary to prediction **A1** (that participants would avoid complex visualizations), all participants (except P3) experimented with creating scatterplot matrices during exploration. The majority of these scatterplot matrices involved adding a third axis to an existing 2D scatterplot in order to see the relationship between two variables and on a common axis (such as a time-series axis).

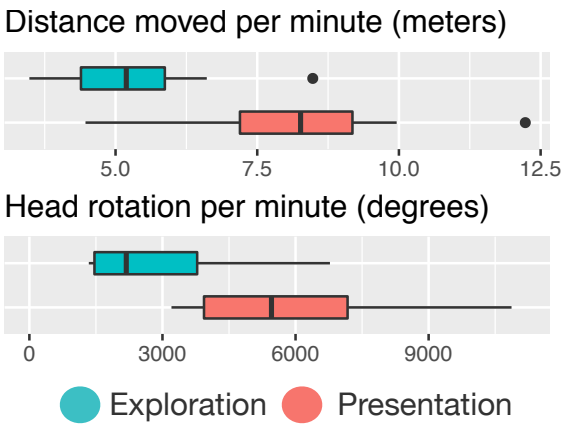



Figure 4.4: Movement frequency distributions by study phase

All participants except P2 and P6 created 3D scatterplots during the explore stage. However, all participants used 3D scatterplots during their presentation stage. Notably, P2 and P6 used 3D scatterplots exclusively during the presentation stage, and P5 used three 3D scatterplots and a 2D scatterplot during the presentation stage. This result ran counter to prediction **A1.1** (participants would avoid creating 3D visualizations). One participant commented that they felt they may as well use 3D scatterplots and other kinds of visualizations as they were exploring data in VR, saying "I wanted to create more graphs of different types, [especially for] my presentation."

We found only weak support for **A2**; during the explore stage, participants would merely choose the nearest open space for creating new views, i.e., not using an organizing principle. Only in the explore stage were they more conscious of structuring the space; more specifically, as noted in our observations supporting **P1.1**, chronology was a common such organizing principle (also partially supporting **A2.1**). We also noted that many undertook a "curation" stage where they would select views that should be included in the presentation, and move them to a designated area.

When considering **A2.2**, we expected participants to minimize their walking, relying instead on rotating their viewpoint. We found that participants walked less during

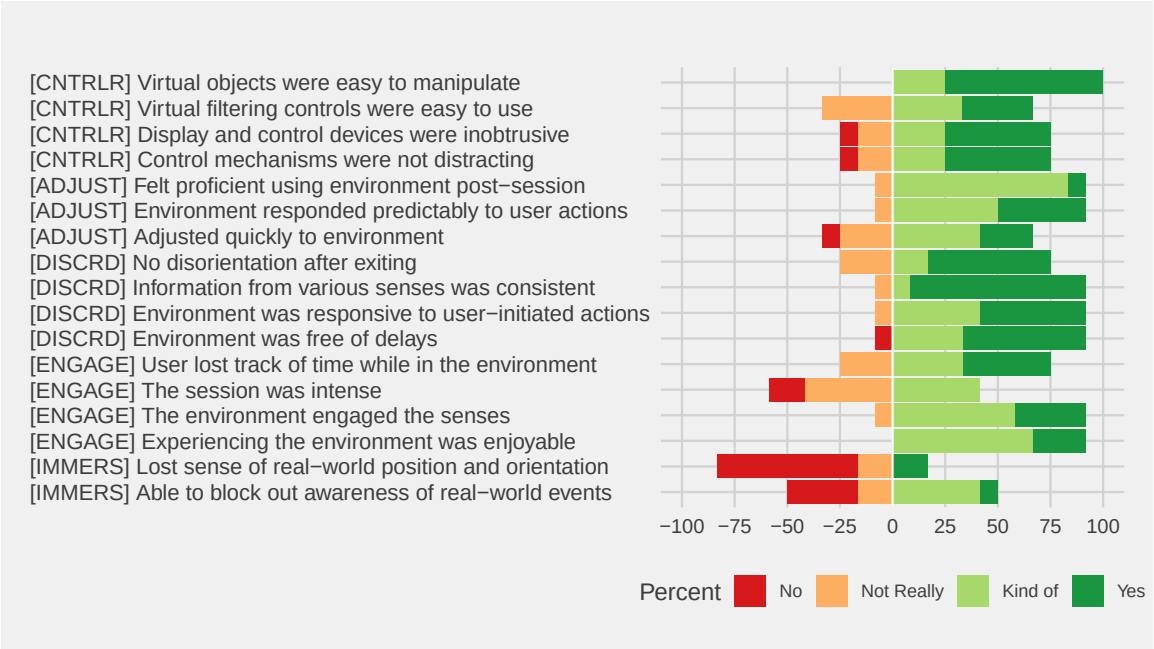



Figure 4.5: Subjective ratings from exit survey. Subfactors include: controller ease of use [CNTRLR], adjustment to environment [ADJUST], perceived immersion [IMMERS], user engagement [ENGAGE], and avoidance of sensory discord [DISCRD].

the explore stage compared to the presentation stage. We ran a paired sample t-test to compare the movement per minute of the explore and present stages. The present stage ($M = 8.10$m, $SD = 2.18$m) had significantly more movement than the explore stage ($M = 5.41$m, $SD = 1.8$m); $t(5) = -3.456, p = 0.018$ (see chart). One participant commented that "When I start thinking of myself as a visual focal point rather than thinking of myself as being surrounded by [vertical] boards, viewing the environment became easier and I felt comfortable using more of the space." We did not find a significant difference in head rotations per minute between the present stage ($M = 6671.86, SD = 5397.94$) and the explore stage ($M = 2571.38, SD = 1321.33$); $t(5) = -2.277, p = 0.072$.

### 4.2.5 Self-reported Perceptual and Cognitive Effects

Even if ImAxes depicts an abstract data analysis setting, it is subject to the same strengths and weaknesses as a general virtual environment; Figure 4.5 shows self-reported perceptual and cognitive effects similar to typical such environments (**A3**). According to the figure, participants reported high scores for perceived engagement (**A3.1**), rating the experience as enjoyable and engaging the senses.

We expected participants to report a high level of presence (**A3.2**) using the system. Supporting this prediction, the survey responses (Figure 4.5) show that participants felt that they were interacting in a natural environment (100% described the environment as being realistic and generally feeling natural, 83% felt moving around was natural). One user described the experience as being somewhat like "*being in the Mojave Desert.*" Several reported that they lost track of time while in the environment. However, full presence may not always be ideal; two thirds of participants reported that their exploration of the data was not intense, and several users pointed out that the sound of the researcher's voice improved their sense of orientation in the real room.

As for increased fatigue level (**A3.3**), we were not able to find support for this prediction. In fact, as our discussion for **A2.2** shows, physical navigation actually *increased* for the presentation stage (which followed exploration), suggesting that fatigue was not a factor. Furthermore, the nausea level was low, which is another indication that participants were not fatigued by the end of the study. We made several predictions related to the challenges that an immersive VR system could potentially introduce. We predicted that participants would suffer from reduced text legibility (**A3.4**). However, the reported Likert scores for the ability to examine closely and obtain details from objects in the environment was high (Figure 4.5), which undermines this prediction. We also expected that participants would suffer from the impreciseness of the VR wand controllers (**A3.5**). Again to the contrary, the Likert scores

indicate that actually participants were able to effectively interact with the 3D virtual objects in ImAxes. By and large, one of the things the participants stated liking most about the environment was that visualizations were very fast and easy (or intuitive) to create relative to their traditional 2D working environment. However, there was a more even split in the participants in regards to wearing the physical VR headset, with one participant reporting "*I don't really like wearing a headset. It's cool to look at things in 3D, but it doesn't really add enough value. However, I also don't usually create visualizations in general during my analyses.*"

Finally, we made two predictions in regards to VR experience; that participants with a lack of VR experience would encounter significant navigation issues (**A4**), and that gaming experience would mitigate a lack of VR experience (**A4.1**). Based on differences in survey responses for users with VR experience versus those without, we find support for the first of these predictions, but not for the second. In fact, participants without VR experience who regularly play computer games for more than an hour in a week reported having more difficulty with the controls and had a more difficult time examining objects in the environment than those who were not regular gamers.

### 4.2.6 Qualitative Feedback

Beyond interaction and visualization requests, participants provided several insightful comments. Whiteboard analogies were commonplace: "*It feels like I'm surrounded by whiteboards,*" said one user; another, after arranging axes around himself in a semi-cylinder shape, described it as feeling like a "*wraparound whiteboard.*"

When asked how ImAxes compared to their traditional desktop display, participants had a range of responses; 58.3% reported feeling more engaged in the problem while in ImAxes than while in their traditional environment (**A3.2**). The most common draw participants

felt to the environment was that creating visualizations was easier and faster in ImAxes than in their typical environments. In general, participants said they might be able to use ImAxes for preparing presentations, reports, and video communications, or for exploratory analysis data validation. One participant responded that they could use it to detect errors during the monthly multi-stage process of reviewing economic indicator estimates prior to publication. Another said that their indicator estimation process involves multilateral aggregation for price indices, and ImAxes could be useful for exploratory analysis during that process. One user noted that export for 2D display presentation would be particularly helpful for the purpose of creating reports.

Several participants said that a major barrier to wanting to use ImAxes on the job is that VR is inconvenient for the purpose of the type of work they perform, which typically involves programming and switching between multiple environments. Said one participant, "*VR seems more oriented toward real-time demonstrations, which is great, but that's not useful for [the participant's] analytical process, which involves long periods of exploration and evaluation switching between tabular views, charts, modeling, and programming.*"

While we were able to implement some changes between our formative phase and our summative phase, there were some changes that were not practical to implement during the span of this study; some of these might be considered applicable for general use, while others are more economics domain-specific. One participant, who was not interested in using ImAxes on the job, said "*it would be sick [sic] if I could click something and see the full hierarchy of categories in the data.*" The absence of this feature wound up being the primary reason for their recalcitrance. Other features this particular participant wanted to see included the ability to run regressions, a group-by mechanism, extra-grammatical filtering mechanisms for building views, and simple computational tasks. Like this participant, several other participants during the summative and earlier phases of the study suggested the inclusion of matrix and column-wise algebraic operations. A number of participants

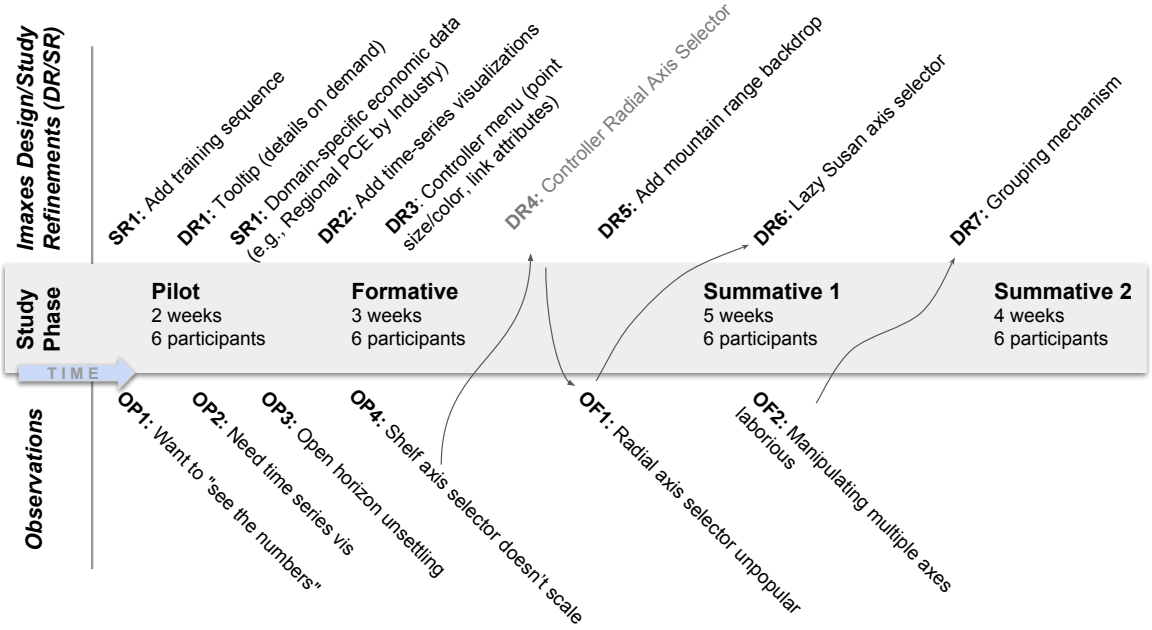



Figure 4.6: Process, timeline, refinements, and observations for our study of ImAxes [17]. DR4 shown in grey, as it was replaced based on user feedback.

also requested linear modeling operations and views of multicollinearity, which they noted as being particularly relevant for hedonic modeling. Another common request was that we extend ImAxes as a tool specializing in outlier detection. Economists typically evaluate time series; while our addition of a line mark connecting scatterplot points was one change we did implement to accommodate this activity, participants regularly reused time period axes, and the option of having a convenient "favorite axes" quick-access area was requested by multiple users. Finally, one participant strongly suggested the addition of a Markov Chain Monte Carlo simulation, stating that it is "what everyone is doing now" in econometric modeling.

### IA with Economists: An Iterative Development Vignette

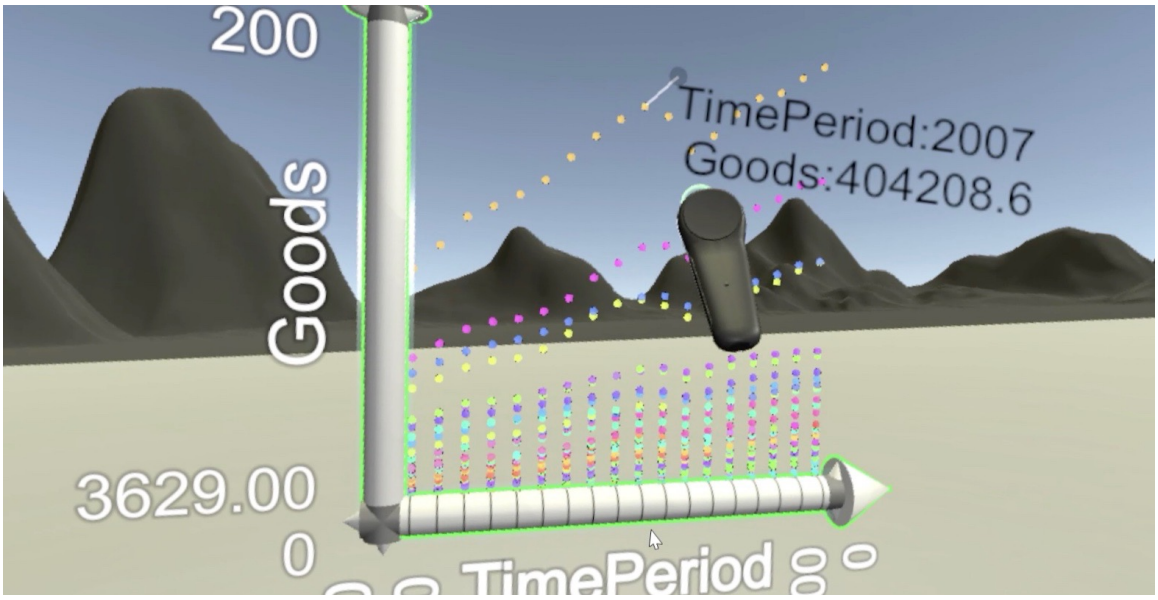



Figure 4.7: Tooltip providing details-on-demand for data items.

The original ImAxes system lacked many features necessary for an economics setting, including some that aid users regardless of domain background. We thus extended the system with additional features to support general use improvements to visual exploration and analysis of data based on feedback from economists. Below we list the main features added (labels refer to Figure 4.6).

**DR1: Tooltip (Details-on-demand).** We implemented details-on-demand as a tooltip for 2D and 3D visualizations using a pointer metaphor (Figure 4.7). By pressing a button on the controller and pointing in the direction of a 2D visualization, the data values of the nearest point are shown in a 2.5D box with a leader attached to the point. To obtain details-on-demand in a 3D scatterplot, a small pointer sphere is attached to the VR controller that can probe the nearest values.

**DR2: Time-series data.** The original ImAxes supported only scatterplots and parallel coordinates plots. Since many of our users wanted to explore time-series data, we added line graphs as well.

**DR3: Visualization design menu.** We added a simple menu control panel attached to the VR controller. This allows users to remap data dimensions to axes, create and bind a gradient colour to a continuous variable, and map the size of the points or lines to a data attribute.

**DR5: Add mountain range backdrop.** Participants disliked the space's flat, sharp horizon and featureless terrain, causing us to add a mountainous landscape in the distance.

**DR6: Axis selection.** Vanilla ImAxes used a *shelf metaphor* for selecting axes, where data axes were arranged in rows like books on a bookshelf.[2] While this metaphor is easy to understand, it does not scale with the number of axes and requires a large amount of locomotion (walking or teleporting). Our first solution, a "Rolodex" (DR4), was poorly received. Instead we implemented a rotational menu based on a "Lazy Susan" metaphor. The menu can be rotated like a Lazy Susan via the controller touchpad. An axis is selected by pulling it out of the Lazy Susan menu. Thus, in the end, there was no shelf.

**DR7: Grouped selection.** Based on user feedback, we implemented a group selection mechanism that allows the user to move a linked group of visualizations

[2]We need axes, lots of axes. (`https://youtu.be/5oZi-wYarDs`)

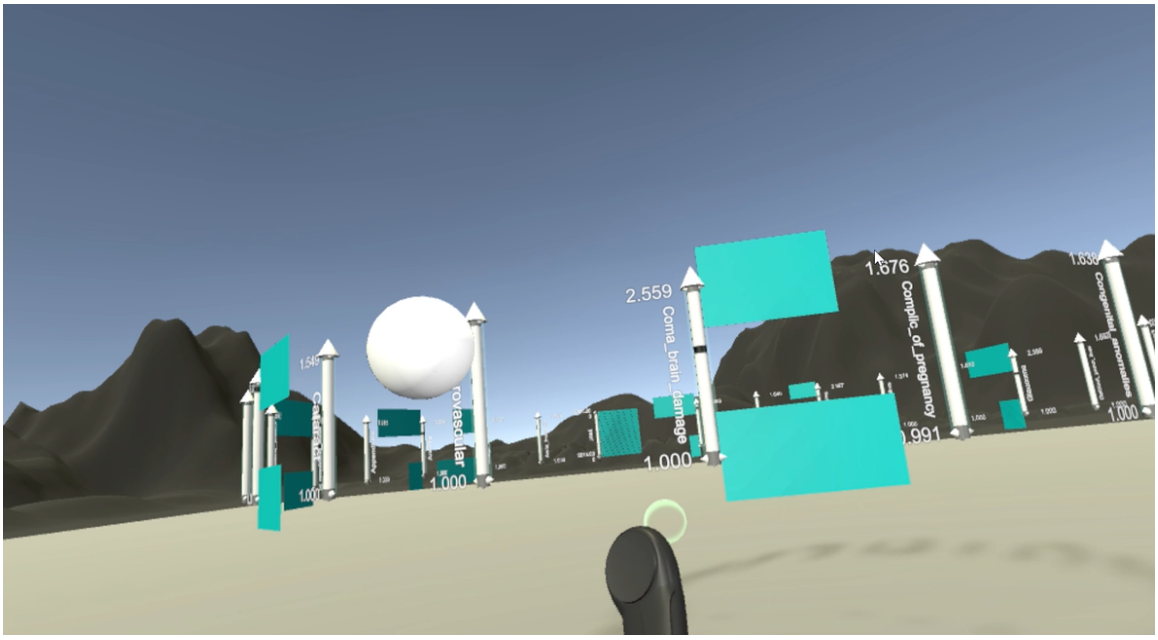



Figure 4.8: Lazy Susan menu implemented for the study. Participants spin the menu by rotating their thumb on the controller touchpad.

instead of a single one. This enables the user to arrange visualizations around them without breaking links.

## 4.3 A Discussion of Our Predictions Versus Our Results

In Batch *et al.* [17], we reported on a design study investigating the use of IA [154], specifically Virtual Reality (VR), for professional economic analysis in a U.S. federal agency. Inspired by Sedlmair's design study methodology [212], this overall study consisted of multiple phases:

1. A *design stage* where we collected requirements using contextual inquiry methodology [29] and improved an existing immersive VR system for multidimensional data analysis—ImAxes [52]—to support macroeconomics data;

2. A *formative "in-the-wild" deployment* of the prototype application in a communal space, which lead to multiple incremental insights and improvements of the prototype; and

3. An *in-depth mixed methods study* (preregistered) involving professional economic analysts exploring their own datasets in our immersive economics environment, and then presenting their findings to the experiment administrator.

The results from these studies include observations, video and audio recordings, interaction logs, and subjective interview plus survey feedback from the participants. In particular, we report on the use and organization of space to support analysis and presentation, barriers against effective use of immersive environments for data analysis, and the impact of immersion on navigation and orientation in 3D. Even more specifically, our predictions—organized into the stage of our study they refer to, exploration (*E*), presentation (*P*), and all (*A*)—were as follows:

**E1** Participants will arrange the views egocentrically around themselves. *Motivation:* For individual work, it is more efficient to use local space around yourself.

**Result:** **We found evidence supporting this prediction**.

**E1.1** Participants will tend to arrange their views at chest level. *Motivation:* Participants have no specific VR training, and will thus likely not utilize the 3D environment to the fullest.

**Result:** **We found mixed or inconclusive evidence about this prediction**.

**E1.2** Participants will arrange their views within easy reach of the center of the space. *Motivation:* The small space that the study is conducted in will not permit significant physical navigation.

**Result:** **We found evidence supporting this prediction**.

**E2** Participants will build many ephemeral visualizations that they quickly discard. *Motivation:* ImAxes supports exploration by creating transient and new visualizations through brushing.

**Result:** **We found evidence supporting this prediction**.

**P1** Participants will arrange the views in an exocentric way. *Motivation:* During presentation, it makes sense to more carefully arrange the views, e.g., in a gallery or sequence.

**Result:** **We found mixed or inconclusive evidence about this prediction**.

**P1.1** Participants will arrange the views in a chronological order w.r.t. to their presentation. *Motivation:* The intelligent use of physical space can help streamline a narrative.

**Result:** **We found evidence supporting this prediction**.

**P2** Participants will build more complex visualizations in the presentation stage than the explore stage. *Motivation:* Presentation involves creating a linear, coherent, and comprehensive narrative. Care can thus be spent on crafting complex visualizations. *Note:* This prediction was not part of our preregistration.

**Result:** **We found evidence supporting this prediction**.

**A1** Participants will prefer basic visual representations (scatterplots, linegraphs, maps), and avoid more complex ones (parallel coordinates, scatterplot matrices). *Motivation:* These more complex representations are not commonplace in real-world data analysis.

**Result:** **We found evidence against this prediction**.

**A1.1** Participants will avoid using 3D representations (such as 3D scatterplots or surfaces). *Motivation:* Our participants have no VR training and are accustomed to 2D displays in their work.

**Result:** **We found evidence against this prediction**.

**A2** Participants will utilize the physical space to structure their work. *Motivation:* Physical space can be used to support specific tasks, e.g., to simplify choice, perception, and computation [125].

**Result:** **We found evidence supporting this prediction**.

**A2.1** Participants will group views in space based on their logical relationships. *Motivation:* Views that belong together should be grouped in physical proximity.

**Result:** **We found evidence supporting this prediction**.

**A2.2** Participants prefer interacting with objects at a near distance than those at a far distance. *Motivation:* Near objects require no physical navigation to access, and ImAxes does not support a reliable distance interaction technique.

**Result:** **We found evidence supporting this prediction**.

**A3** Participants will report typical perceptual and cognitive effects of VR on their performance and perception. *Motivation:* Even if ImAxes depicts an abstract data analysis setting, it is subject to the same strengths and weaknesses as other VR applications.

**Result:** **We found mixed or inconclusive evidence about this prediction**.

**A3.1** Participants will report a high level of engagement. *Motivation:* VR is commonly associated with high engagement because of realism and low indirection.

**Result:** **We found evidence supporting this prediction**.

**A3.2** Participants will report a high level of presence. *Motivation:* VR is commonly associated with high presence because of the low indirection, natural interaction, proprioception, and the perception of physical space.

**Result:** **We found evidence supporting this prediction**.

**A3.3** Participants will report fatigue from physical navigation and interaction. *Motivation:* The use of gross body motor controls to navigate the virtual environment

and interact with its objects will require significant effort by the participants.

**Result:** **We found mixed or inconclusive evidence about this prediction**.

**A3.4** Participants will suffer from reduced legibility of text in the 3D environment. *Motivation:* HMDs have a significantly lower resolution than typical monitors, and labels in ImAxes are 3D and thus subject to distance and orientation concerns.

**Result:** **We found evidence against this prediction**.

**A3.5** Participants will suffer from the challenge of using VR wands to interact with virtual 3D objects. *Motivation:* While more direct than using a mouse and keyboard, the HTC Vive controllers still do not allow for hand and finger interaction.

**Result:** **We found evidence against this prediction**.

**A4** Participants will encounter significant navigation and interaction hurdles due to a lack of VR expertise. *Motivation:* Our participant pool has no specific VR training, and will thus be challenged by 3D navigation and interaction concerns.

**Result:** **We found mixed or inconclusive evidence about this prediction**.

**A4.1** Participants with 3D computer gaming experience will be less hindered by lack of VR training. *Motivation:* 3D gaming experience will help people interact more efficiently.

**Result:** **We found mixed or inconclusive evidence about this prediction**.

We were surprised to find that many of our predictions—particularly those anticipating negative effects of VR use—found no support in the collected data. For example, we noticed few effects of fatigue (**A3.3**), legibility was not a clear concern (**A3.4**), and even participants with little gaming and/or VR experience were able to use our tool efficiently (**A3.5**). Some of these findings can be easily explained—e.g., that the lack of an exocentric layout likely

happens because presenters actually view the environment through the eyes of the participant (**P1**)—but others are more unexpected. Most of the time, while aptly highlighting our lack of knowledge, these contrary results are actually in favor of IA; for example, participants did actually use advanced visualizations (**A1**), not merely sticking to scatterplots.

On the surface, this finding disappointingly does not extend to the intelligent use of space (**A2**), as participants in the explore stage merely picked the closest free space to put new visualizations. However, when viewed through the sensemaking loop [188], this makes more sense as one of its early phases involves placing potentially relevant information in a so-called "shoebox." When foraging for information, analysts typically do not have time to worry about structure, similar to how the purpose of sketching for artists is to generate new ideas rather than fixate on existing ones. Only in a secondary curation step in our study would participants evaluate these views and organize them into a designated area in the environment (the "evidence file" in the sensemaking loop). Incidentally, Andrews [5] noted similar observations, referring to them as "evidence marshalling," often having the same chronological organizing principle (**A2.1**) as our findings.

Still, it is clear that participants did not use the full available 3D space of the analysis environment to its full potential. Telemetry data and video recordings showed participants mostly stayed in place and merely used the space directly in front of them. We did see participants making better use of the space in the presentation stage. Reasons for this may include the cramped confines of our experimental space, the interactions needed to move visualizations, and no automatic layout control. This points to the need for system support, such as constraints and organization frameworks, to help users organize their spaces as has previously been done for 2D GUI tools [141, 169].

That many of our—in retrospect pessimistic—predictions about the drawbacks of the virtual environment were not supported is worth unpacking. One reason may be the high presence and engagement levels reported, leaving participants willing to simply overlook

minor usability concerns. The novelty factor may also be working in our favor and produce goodwill towards the tool. Finally, perhaps the natural interaction metaphors in ImAxes simply aided participants in quickly learning the system and exploring their data. In fact, we were surprised by the level of interest in using ImAxes in the workplace. There were a number of domain-specific tasks and requests which we were unable to fully accommodate in the scope of this study—MCMC simulation, linear modeling and views of multicollinearity for hedonics, a suite of features for outlier detection—as well as the more generally-applicable requests for matrix algebraic operations, quick-access-axes for "favorite" axes (like time series period indices), and integration with hierarchical views of the data. While these changes were not practical to implement in this study, we view them as low-hanging fruit for future work extending immersive implementations either for general analytical tasks or more specialized economic analysis ones.

# Chapter 5: View Management for Situated Visualization

Advances in mobile and wearable display interfaces, positional sensing, and computer graphics have fueled recent efforts of displaying data *in situ*. Current research themes such as immersive [43], ubiquitous [66], and situated analytics [233] (IA/UA/SA) explore a world where contextual data is readily available at the fingertips of the user, anywhere and anytime. Particularly exciting is the topic of *situated visualization*, where data relevant to a location is visualized directly in said physical location [69, 219, 248]. However, there are several challenges in making such situated visualization practically useful, such as the intrinsic VR/MR challenges of registration mechanisms, power consumption, device ergonomics, etc.; a recent paper surveyed the grand challenges of immersive analytics research [72]. One such challenge is *view management* of situated visualizations: effectively viewing (and interacting with) the situated visualizations that co-inhabit the user's physical space in an AR environment.

A *situated visualization* is a data visualization — often 3D volumetric in nature — that has been embedded into an AR or MR environment [8] to support situated [208], ubiquitous [66], and immersive analytics [154]. However, because of the 3D nature of the environment, making efficient use of such situated visualizations requires significantly more overhead, navigation, and layout than for traditional visualizations drawn on a normal 2D screen. Bell *et al.* [25] define *view management* for AR and VR as "maintaining visual constraints on the projections of objects on the view plane, such as locating related objects near each other, or preventing objects from occluding each other." Drawing on this definition,

we refer to *view management for situated visualization* as optimizing the user's view of the visualizations — both on an individual as well as a collective level — in a IA/SA environment.

We present an analysis of the challenges of view management for situated visualizations in MR, enumerating concerns such as physical distance and reach, orientation and legibility, and depth and occlusion. These challenges apply to both the components of a single situated visualization, as well as to multiple visualizations that exist in the same physical environment. Based on this analysis, we revisit existing techniques from the domain of computer graphics and visualization to propose a set of interaction, layout, and presentation techniques that are designed to mitigate these aforementioned challenges. More specifically, we investigate the following techniques:

1. a *shadowbox* that enables eliminating effects of perspective foreshortening and occlusion in 3D visualizations, including an *unfolding* interaction for transforming a 3D visualization into individual orthographic 2D views;
2. a *cutting plane* interaction for accessing occluded elements of a 3D visualization;
3. a *world in miniature (WIM)* technique for overviewing and accessing multiple visualizations in a 3D situated analytics environment;
4. a *summoning* interaction for bringing distant visualizations to the user and arranging them in an accessible grid (*dispelling* returns them to their original locations); and
5. a *data tour* for guiding the user through a 3D situated analytics environment to visit all visualizations of interest.

While each of our techniques are derived from existing work, we claim that their combination as well as their application to situated visualization in MR, presented in this paper, is novel. Furthermore, to validate the work, we present a practical implementation

of these ideas inside a novel situated visualization environment implemented using the web-based VRIA [37] framework. We also report on findings from a remote user study where 12 participants used their smartphones to perform situated analytics tasks using our proposed techniques. While these results highlight many of the typical challenges associated with mixed reality and sensemaking, we also found convincing evidence supporting our suite of view management techniques.

## 5.1 Properties and Challenges in Situated Visualization View Management

While there are significant and valid concerns with using 3D visualizations in the first place [168], these are mostly moot when discussing SA on MR devices. In such settings, the user is by definition active in the real world using hand-held or HMD technologies, and thus representing visualizations in 3D is inevitable even if the visualizations themselves are not 3D. To understand the underlying challenges of view management for situated visualization, let us first enumerate the basic properties of a visualization inhabiting a SA environment [25]:

- **Position:** The visualization's location in the environment;
- **Size:** Its geometric size in relation to the rest of the world;
- **Transparency:** The opacity of the visualization, which also incorporates its general geometry (i.e., some visualizations such as a 3D scatterplot are more sparse than a volume rendering);
- **Priority:** A relative priority for each individual visualization (potentially whether a visualization is selected or not);
- **Orientation:** The visualization's 3D rotation;
- **Distance:** Its distance from the viewer and other visualizations;

- **Area of interest:** The area (often a 3D volume) from which to optimally view the visualization; and
- **Spatial relation to the surrounding world:** The visualization's relation to real objects co-inhabiting the physical world.

The above list is by no means a minimal one, as some properties are derivatives of others (distance vs. position, for example). However, it is useful to distinguish each of these properties individually, as they all give rise to specific challenges. We outline these below; again, we make no effort to streamline these challenges (e.g., visibility subsuming occlusion), but instead list them individually because they add reasoning power to our argument. Furthermore, some of these challenges apply within a single visualization (e.g., occlusion within the points in a 3D bar chart), whereas others apply for multiple visualizations (e.g., overview for all of the visualizations in an environment), and some apply to both (occlusion between marks, as well as occlusion between two visualizations).

**Visibility.** Maintaining visibility of a situated visualization in the user's field of view is a fundamental challenge [25]. Many situated visualizations are just that, situated in a specific location in the physical world, which means that they easily fall outside the user's vision, either by being too far away or above, below, or behind the user. In such situations, the visualizations cannot be moved to always be visible, and other mechanisms must be employed to make the user aware of their existence and location. This challenge is compounded when multiple visualizations are jostling for space on the user's field of vision.

**Occlusion.** The three-dimensional nature of SA environments means that a geometric object can be hidden by other objects even if they do not intersect in 3D space [67]; the problem is further exacerbated when they do intersect. This fundamental challenge affects

both marks within a single visualization, such as a cluster of marks in a 3D scatterplot occluding an outlier on the far end, as well as between multiple visualizations, such as a 3D volume occluding a barchart in the distance.

**Overview.** Overview is a central aspect of data visualizations [218], but gaining an overview of all of the visualizations in a SA environment is particularly challenging because of the 3D nature of the space. This is not merely about the ability to access and read an individual visualization, but being aware of its existence in the first place; a visualization that is fully occluded by other visualizations, outside the user's field of view, or too far away to see, will inevitably not be included in the overview. This means that many of the below challenges contribute to the overall Overview challenge.

**Perspective Foreshortening.** A more subtle aspect of the 3D environment is the impact of perspective foreshortening due to visualizations being at different distances from the viewer. Perspective foreshortening arises from the non-linear 3D perspective, essentially making nearer items disproportionately larger than more distant ones. Besides having an impact on the Occlusion challenge, it also makes it difficult to compare between two visualization marks at different distances, such as two different bars in a 3D bar chart.

**Legibility.** A particular concern for situated visualizations that are not rotation invariant or not always facing the user, such as a billboard,[1] is legibility, particularly of text. In such situations, the slanted or rotated view of the visualization makes reading more difficult or even impossible. In addition, similar legibility concerns arise when a visualization is far away from the viewer, making graphical features in general—and text in particular—too small to distinguish.

[1] In 3D computer graphics, a "billboard" is a 3D object that is drawn to always be facing the viewer either along just the vertical axis (like a sign spinning around its post to face the user), or both vertical and horizontal axes.

**Physical Navigation.** When a situated visualization of interest is too far away to be legible or manipulated, one (or both) of either the visualization or the user will generally need to move. When this task falls on the user, such as when the visualization cannot be moved from a specific geographic position or real-world object, this translates to the user physically having to navigate to the object of interest. Unlike in dedicated VR spaces, such as open labs or even CAVEs, such navigation can be particularly tricky in a physical environment filled with slippery or uneven surfaces, physical barriers, and other people, as well as when using interaction devices such as wands, gamepads, or touch surfaces to manipulate virtual objects in the environment.

**Physical Reach.** Even when physical navigation is not needed, many situated visualizations require interaction that involve the physical reach of the user. In fact, sometimes a situated visualization can be located in a position that is not physically accessible to a person moving around in the real world, and which they cannot reach.

**Temporal and Spatial Continuity.** Finally, as observed by Bell *et al.* [25], given view management strategies that minimize the above challenges, it is important to maintain continuity over time and in space so that objects and visualizations do not "jump around" due to discontinuous layouts that are calculated independently from frame to frame. Thus, objects should move smoothly over time and space.

## 5.2 Prototyping Situated View Management

In Batch *et al.* [22], Sungbok Shin, Julia Liu, Peter Butcher, Panagiotis Ritsos, Niklas Elmqvist and I used the design space described in Section 2.2.2 as a generative lens for designing situated views. For each technique below, we enumerated its properties, how the technique modifies the properties of the situated visualizations, and which of the challenges

the technique addresses (see Table 5.1 for a summary).

Table 5.1: Challenges addressed by each technique.

| Challenge \ Technique | WIM | Summon & Dispel | ShadowBox | Cutting Planes | Data Tour |
|---|---|---|---|---|---|
| Visibility | ✔ | ✔ | ✔ | ✔ | ✔ |
| Occlusion | ✔ | ✔ | ✔ | ✔ | ✔ |
| Overview | ✔ | ✔ | ✔ | | |
| Perspective Foreshortening | | | ✔ | | |
| Legibility | | | ✔ | | ✔ |
| Physical Navigation | ◑ | ✔ | | | ✔ |
| Physical Reach | ◑ | ✔ | | ✔ | ✔ |
| Temporal & Spatial Continuity | ◑ | ✔ | | | ◑ |

WIM Summon & Dispel ShadowBox Cutting Planes Data Tour

◑ partially or indirectly addressed

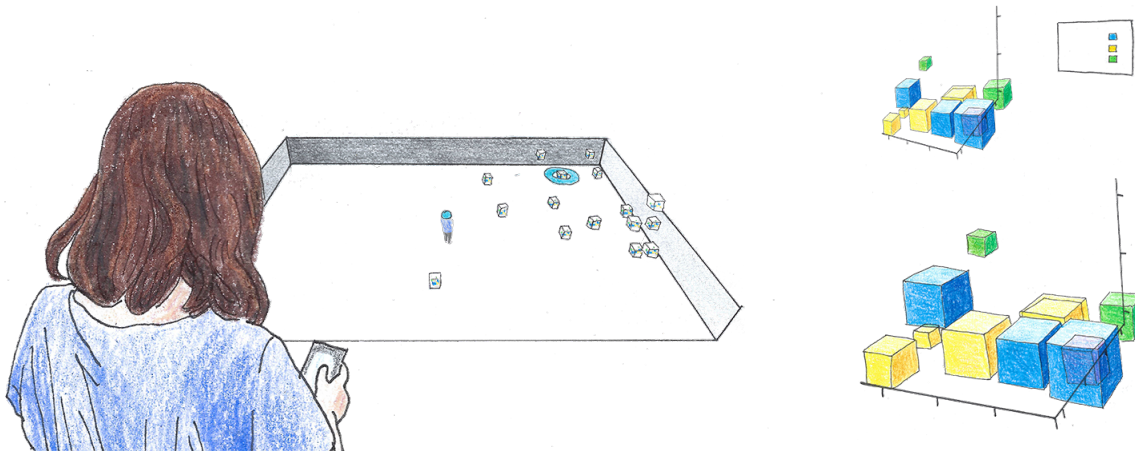


Figure 5.1: Sketch of an example world-in-miniature scenario.

### 5.2.1 World-in-Miniature

**Synopsis:** A World-In-Miniature [227] (WIM) is a miniaturized view of the environment that is controlled by the user, allowing them to see their surroundings from any direction and distance.

We apply the basic WIM approach to a situated visualization environment where the WIM is instantiated by the user and is represented by a box containing a miniature represen-

tation of all virtual features in the scene (including the user). The WIM omits real-world features, but may include contextual information such as a mesh of the landscape as detected by the device, or a map tile layer based on the user's GPS coordinates. Finally, the WIM, as an object in the user's view itself, has its own properties, in addition to affecting the properties of the situated visualizations.

**Properties affected:**

- *Position:* Situated visualizations can be moved by dragging them in the WIM. Their positions within the WIM also reflect their relative positions in the world.
- *Size:* The WIM duplicates all situated visualizations at a significantly smaller scale within a space.
- *Transparency:* Making WIM elements semi-transparent allows the user to see the real world as well as spot virtual elements that are occluded by other virtual elements.
- *Orientation:* The orientation a situated visualization in the WIM should reflect that of its true orientation in the world.
- *Distance:* The relative distance between the user and situated objects is represented accurately by the WIM.
- *Spatial relation to the surrounding world:* The WIM allows for decoupling situated visualizations from the real world. The WIM itself is non-situated.

**Challenges Addressed:** The WIM technique directly addresses *Overview*, *Visibility* and *Occlusion* by creating miniature copies of all objects in the scene, including those not visible to the user, and by giving the user the freedom to rotate the scene to discover and access hidden content. Furthermore, this also eliminates the need for *Physical*

*Navigation* and *Physical Reach*, as the miniature allows for easy navigation and access. However, the technique does not address *Perspective Foreshortening* and *Legibility*, but can be designed to respect *Temporal and Spatial Continuity* by synchronizing the positions of the miniature visualizations with those of their true situated counterparts.

### 5.2.2 Summon and Dispel

**Synopsis:** The Summoning interaction brings all points of interest to the user's position, whereas Dispelling returns them back.

Summoning can be applied to all situated objects, or only to those fitting a specific criteria, such as a special data type, visualization, or spatial area. We opt for displaying the summoned objects using a "shelf" layout, with the points of interest arranged in a grid in front of the user. Interactive or aggregated alternatives can also be considered, such as a "Lazy Susan" or carousel-style layout [17].

**Properties affected:**

- *Position:* Summoning arranges situated visualizations in a neat layout that is readily visible to the user; dispelling reverses this operation.
- *Distance:* Summoning drastically reduces the distance between a situated visualization and the viewer.
- *Area of interest:* Each visualization will be placed where it can be optimally viewed and manipulated.
- *Spatial relation to the surrounding world:* Summoning eliminates a virtual object's relation to the spatial world; dispelling restores this mapping.

**Challenges Addressed:** Summoning is primarily designed to minimize *Physical Navigation* and facilitate *Physical Reach* by essentially bringing the spatial visualization content to the user rather than having the user travel to them. As a secondary effect, this also reduces the *Visibility* and *Occlusion* challenges and provides improved *Overview* of the virtual space. By enabling dispelling the objects to their original locations, the technique also supports *Temporal and Spatial Continuity*.

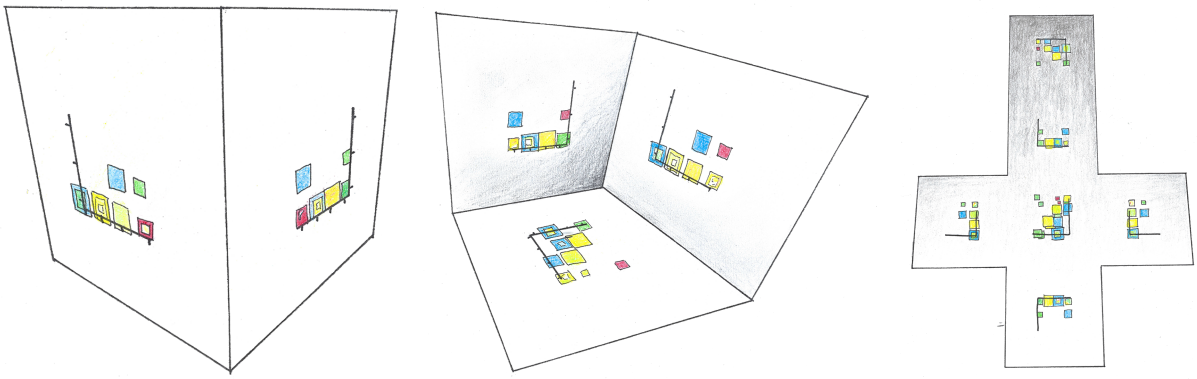
### 5.2.3 Shadowbox

**Synopsis:** The Shadowbox puts a given 3D object (situated visualization) inside a virtual "display case" represented as a 3D box, with 2D orthogonal projections of the object on each of its faces, and support for unfolding the box.

In this way, the Shadowbox is similar to ExoVis [235] but presents the user with either an exterior or interior a view of the box and enables hiding the 3D object to mitigate occlusion. We also propose "unfolding" the sides of the box to align multiple 2D projections in one plane, allowing the user to view all projections at once (Figure 5.2c).

**Properties affected:**

- *Area of interest:* The Shadowbox provides optimal views of a situated visualization along each of the primary axes.
- *Spatial relation to the surrounding world:* The Shadowbox has the possibly problematic side-effect that it isolates and separates the situated visualization being examined from the rest of the world.



(a) Exterior projection. (b) Interior projection. (c) Unfolding.

Figure 5.2: The Shadowbox technique, including its (a) exterior and (b) interior projection modes, as well as the (c) unfolding interaction.

**Challenges Addressed:** The Shadowbox was primarily designed to manage the *Perspective Foreshortening* typical in 3D environments by using an orthographic projection for each of the planes, thus enabling exact visual comparison for, e.g., the bars in a 3D barchart. However, it can also help aid *Visibility* and *Occlusion* as well as support *Overview* by providing a structured view of the 3D object. This axis alignment can also facilitate *Legibility*.

### 5.2.4 Cutting Planes

**Synopsis:** A Cutting Plane is an interactive filtering and view-flattening mechanism that takes a 2D slice of the 3D object from a position and orientation determined by the user's manipulation of a plane [99].

Like the WIM technique, cutting planes are a classic approach that have stood the test of time. Here we discuss how it affects the properties of a situated visualization:

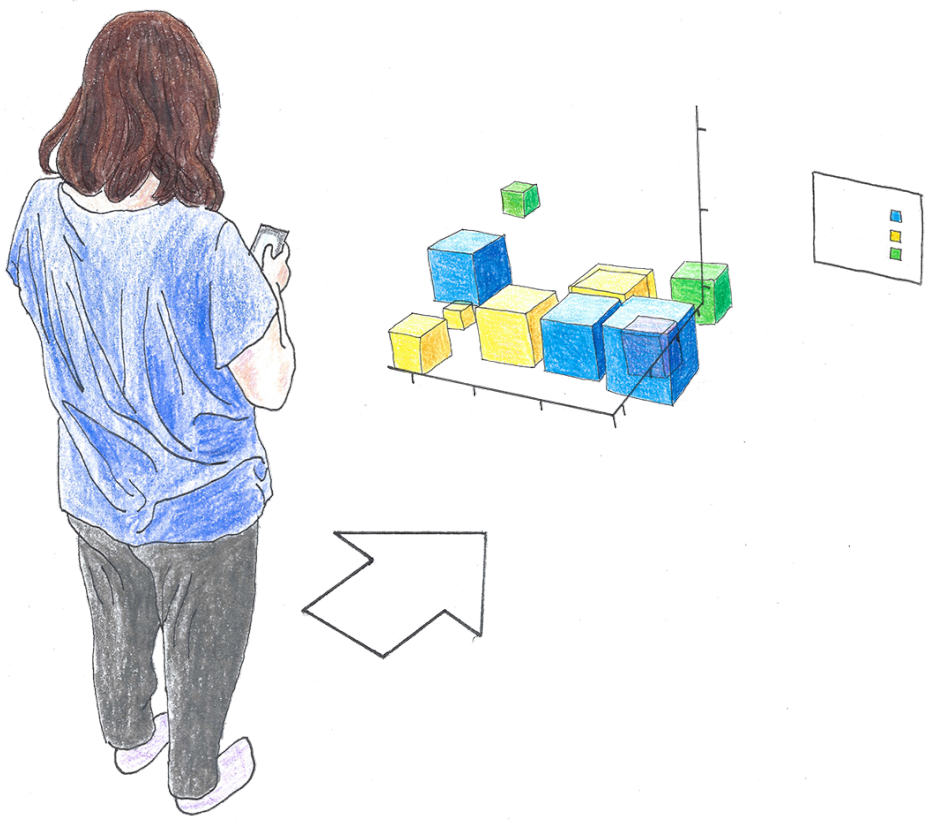


Figure 5.3: A user is guided to a location of interest by an arrow overlay positioned on the ground in front of her.

**Properties affected:**

- *Transparency:* A cutting plane essentially renders the cut part of the situated visualization fully transparent, enabling easy visual inspection and access to its interior.

**Challenges Addressed:** The primary design rationale for cutting planes is to optimize *Visibility* and to manage *Occlusion* arising from the 3D object obscuring itself (such as in a 3D scatterplot). It also facilitates *Physical Reach* of areas that would otherwise be inaccessible to the user.

### 5.2.5 Data Tour

**Synopsis:** A Data Tour is a guided walk through all points of interest in an environment.

The complexity of the algorithm for placing navigation cues may vary. A simplified set of steps for implementing a data tour are as follows:

1. Select all nodes not already visited from the user.
2. Render a visual guide in the user's field of view to the currently nearest point of interest.
3. Once the user comes in close proximity to the point of interest, add it to the list of visited points and repeat from Step 1.

A more sophisticated implementation would implement actual wayfinding into the system, thus taking advantage of street-level maps and blueprints to guide the user on the most optimal path to a target.

**Properties affected:**

- *Position:* Visualizations are not moved; instead, the user visits the visualizations through physical navigation.
- *Distance:* The Data Tour tries to bring each point of interest within optimal distance to the user over the entire tour.
- *Spatial relation to the surrounding world:* Significantly, the Data Tour preserves the spatial location and mapping of each situated visualization to the physical world.

**Challenges Addressed:** The Data Tour facilitates accessing situated visualizations to avoid *Occlusion* and support *Visibility* by guiding the user's *Physical Navigation*. This also means that the *Legibility* and *Physical Reach* challenges can be addressed by guiding the user to the objects. In particular, the Data Tour provides *Temporal and Spatial Continuity* since it does not alter the environment at all.

## 5.3 Motivating Scenario

Sydney is investigating quarterly productivity measures in her manufacturing company. She has determined that it would be informative to visit the shop floor of a plant that has shown unusual fluctuations in productivity over the last two months. While on the shop floor, it becomes apparent that she needs to collectively evaluate a handful of production input and intermediate output records that had not been on their radar prior to her visit and thus is not present in any of the prepared views she has on hand.

Rather than traversing the company's intranet looking for the data, Sydney uses her mobile device to view the measures *in situ* using space itself as an index. She creates an assortment of 3D visualizations of the records that are summoned and laid out in a view centered on her current location. The visualizations form a shared virtual workspace visible only to her, wrapping her in an mixed reality environment of contextual data that is anchored to specific machines and processes on the factory floor. Noticing an unexpected pattern in one such visualization, Sydney first runs a cutting plane through one of the 3D charts, and then instantiates a shadowbox around it, which she unfolds to add precision to her view of relationships between multiple variables in the chart.

Based on these findings, Sydney is able to pinpoint a part of the process that is most concerning, select it, then dispel the visualizations to their situated locations. Following

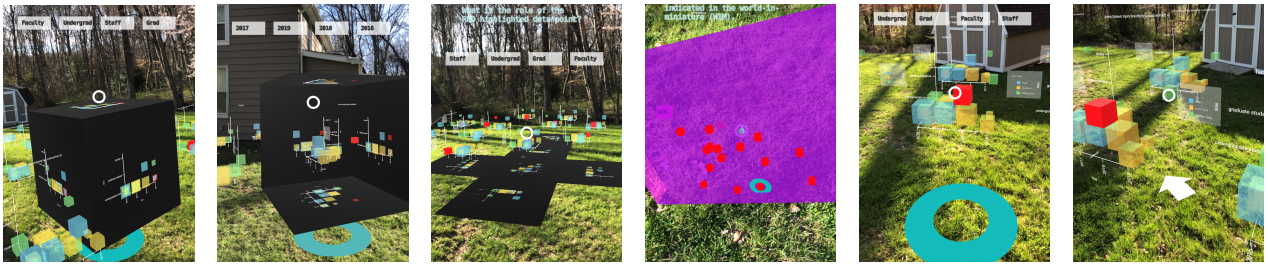
a guidance arrow indicating the direction to the workstation most relevant to the process, she quickly arrives at the failing location in physical space. After an inspection of the workstation, Sydney finds that a component of a machine at the workstation used for producing an intermediate output is suffering from a faulty component.

### 5.3.1 Situated Analytics Implementation Details

We have implemented all techniques described in this section as a unified system demonstrating their synthesis using `three.js`, VRIA [37], and AFrame-React. VRIA was used to create "staged" visualizations—the initial collection of 3D visualizations instantiated upon loading the page. To evaluate our implementation, we set up a MERN stack; the backend of our system is described in Section 5.4.2.

Our motivation for using the Web as a development ecosystem was two-fold. Firstly, we believe that the Web—and in particular the mobile Web—is the most ubiquitous, collaborative, and platform-independent way to build and share information [198]. Unlike game engine-based systems, there is no need to download bespoke applications or executables, whereas the outputs can be easily integrated in other Web-based applications [37]. These characteristics make the Web an excellent platform for situated visualizations as interconnected hypermedia, whether in MR such as those presented in this work, or not. In addition, standardization advances within the Web ecosystem, such as WebXR or WebRTC, provide interoperability capabilities, such as those implied by Mackay [146]. This make mobile MR-based SA possible. When these capabilities are coupled with faster communication, such as 5G, the opportunities for providing interconnected, data-driven information, *in situ*, as enhancements of our physical world, increase significantly.

Secondly, the reliance on such Web technologies and tools, besides providing a familiar and versatile development ecosystem, enabled the collaborative development and real-time



(a) Exterior projection. (b) Interior projection. (c) Unfolding. (d) World-in-miniature. (e) 3D VRIA chart with marker. (f) Guided tour arrow.

Figure 5.4: The user study implementation, featuring 3D VRIA figures (e,f) and a Shadowbox with exterior and interior projection modes (a and b), as well as the unfolding technique (c), for the analysis stage. The WIM (d), target marker (e), and guided tour (f) were all used for navigation.

inspection of MR prototypes across two continents, much like one would do with 'traditional' 2D visualizations, with outputs accessible through a WebXR-compatible browser. Due to VRIA's WebXR support, a single codebase can be experienced with immediacy, in a variety of devices, without prior knowledge of the underlying hardware. Finally, it enabled us to carry out an evaluation, described in the next session, which respected the social distancing rules many of us have to abide to, albeit with some limitations on the platforms examined.

## 5.4 Evaluating Situated Analytics

In Batch *et al.* [22], our study's goal was to gauge a selection of the presented view management techniques, described in Section 5.2. Rather than testing hypotheses in a confirmatory experiment, we conducted an exploratory study and evaluated the participants' use of time and space, the correctness of their responses to basic analysis tasks, and self-reported measures of task-related user experience factors. COVID-19 pandemic restrictions forced us to perform all testing remotely. This meant that we had to conduct our evaluation using standard consumer-level hardware (smartphones) rather than the MR HMDs (HoloLens 2) we had initially planned to use. This also impacted our choice of which techniques to assess. More specifically, we did not evaluate the summon and dispel technique in part because it is suited for larger, possibly outdoor spaces, rather than the indoor spaces we

anticipated most would use.

### 5.4.1 Participants

We recruited a convenience sample of 12 participants (hereafter referred to as "users") with professional backgrounds in user experience, design, and interface or systems development, maintenance, and engineering. All users were screened to possess AR-compatible Android mobile devices; at the time of our writing this paper, the only WebXR viewer available on Apple iPhones, Mozilla's WebXR Viewer, was no longer being maintained, and its final version suffers from major performance issues with one of the libraries upon which our implementation depends.

We opted to filter our users to those with professional experience in these domains because they are the target user that motivates this work, as described in the scenario in Section 5.3. However, professional domain experts are difficult to recruit with monetary incentive alone. Thus, we argue that our approach of collecting a convenience sample of individuals within our personal and professional networks is a necessary one in this instance. Due to the requirements by which we filtered our users (i.e., those with professional experience in technical domains who already possess compatible mobile devices), we argue that the relatively small sample represents an adequate contribution.

### 5.4.2 Apparatus and Data

As noted in Section 5.4.1, we screened out users who did not possess mobile devices compatible with the libraries upon which our implementation depends. These libraries include A-Frame and `three.js`. We set up an Express NodeJS server API endpoint that posts to a MongoDB database to log user session events (navigation, multiple choice question responses), position and orientation during their experiment sessions. When the scene is

| Step | Task Type | Variable | Technique | POI | Correct Answer | Incorrect Options |
|---|---|---|---|---|---|---|
| 1 | Navigation | lat/longitude | | 10 | POI 10 | |
| 2 | Analysis | injury severity (0-5) | | 10 | 3 | 1, 2, 4 |
| 3 | Analysis | weather | | 10 | Clear | Snow, Sleet, Rain |
| 4 | Analysis | year | | 10 | 2016 | 2017, 2018, 2019 |
| 5 | Analysis | role | | 10 | Undergrad | Faculty, Grad, Staff |
| 6 | Navigation | lat/longitude | | 11 | POI 11 | |
| 7 | Analysis | injury severity (0-5) | | 11 | 4 | 1, 2, 3 |
| 8 | Analysis | weather | | 11 | Sleet | Snow, Clear, Rain |
| 9 | Analysis | year | | 11 | 2018 | 2016, 2017, 2019 |
| 10 | Analysis | role | | 11 | Undergrad | Faculty, Grad, Staff |
| 11 | Navigation | lat/longitude | | 1 | POI 1 | |
| 12 | Analysis | injury severity (0-5) | | 1 | 0 | 1, 2, 3 |
| 13 | Analysis | weather | | 1 | Sleet | Snow, Clear, Rain |
| 14 | Analysis | year | | 1 | 2019 | 2016, 2017, 2018 |
| 15 | Analysis | role | | 1 | Undergrad | Faculty, Grad, Staff |
| 16 | Navigation | lat/longitude | | 9 | POI 9 | |
| 17 | Analysis | injury severity (0-5) | | 9 | 1 | 0, 2, 3 |
| 18 | Analysis | weather | | 9 | Clear | Snow, Sleet, Rain |
| 19 | Analysis | year | | 9 | 2019 | 2016, 2017, 2018 |
| 20 | Analysis | role | | 9 | Grad | Faculty, Undergrad, Staff |
| 21 | Navigation | lat/longitude | | 3 | POI 3 | |
| 22 | Analysis | injury severity (0-5) | | 3 | 2 | 1, 3, 4 |
| 23 | Analysis | weather | | 3 | Snow | Clear, Sleet, Rain |
| 24 | Analysis | year | | 3 | 2019 | 2016, 2017, 2018 |
| 25 | Analysis | role | | 3 | Undergrad | Faculty, Grad, Staff |
| 26 | Navigation | lat/longitude | | 4 | POI 4 | |
| 27 | Analysis | injury severity (0-5) | | 4 | 2 | 1, 3, 4 |
| 28 | Analysis | weather | | 4 | Snow | Clear, Sleet, Rain |
| 29 | Analysis | year | | 4 | 2018 | 2016, 2017, 2019 |
| 30 | Analysis | role | | 4 | Grad | Faculty, Undergrad, Staff |

Hovering Marker (Control) WIM Data Tour VRIA Chart (Control) ShadowBox (Folded, Interior Projection) ShadowBox (Folded, Inverted/Exterior Projection) ShadowBox (Unfolded)

Table 5.2: Sequence of tasks.

initialized, an ID is assigned to the session and posted to the server from the client via the API. Each one-second interval after the session is posted, the client interface posts the user's scene camera position and rotation along with their session ID, and the server updates the database collection with these features, the server timestamp, and the user's IP. Whenever the user answers a question, the client interface posts the user's answer, the possible answers (with the correct answer indicated), the POI and task it corresponds to. These questions all correspond to features of a synthetic dataset comprised of spatially-tagged falling accident events on an institution's campus with features of the time and setting in which the accident took place (weather, season, year), of the victim (gender, role within the institution), and of

the accident itself (injury severity).

User experience survey data was collected via Google Forms following the session task completion experiment. The survey structure was based heavily on the NASA Task Load Index (TLX)[2]. All questions from the NASA TLX were asked for each technique individually, and the survey ended with two ranked voting questions—one for the techniques used for navigation and one for the techniques used for analysis—and three open-ended text entry questions about the user's experience. NASA TLX questions and scales were used verbatim in our survey, with one exception: We flipped the scale of the question "How successful were you in accomplishing what you were asked to do?" prior to conducting our formal study after several pilot users reported misinterpreting the scale's order for that question specifically. Qualitative data was also collected in the form of researcher notes, video and audio recording, and transcripts from the session.

### 5.4.3 Experimental Design and Procedures

We conducted user sessions using video conferencing on Zoom lasting approximately 30 minutes to 1 hour during which the user was given a summary description of what to expect during the session. The researchers verbally confirmed that the user was aware that video and audio recording would be collected throughout the session; the users were also informed once the recording begins. Users were asked to stand in the middle of the space they would be using during the session, and to visit a randomly-generated URL for the experiment implementation using the Chrome browser on their mobile device. The experiment implementation prompted the user to find their way to a point-of-interest (POI) using one of three techniques represented in AR in their living space through their mobile device, and once they had arrived at their destination, to answer a multiple choice question

[2] `https://humansystems.arc.nasa.gov/groups/tlx/`

using either a shadowbox technique or the VRIA figure alone. These tasks and the order in which they were encountered by users are described in Section 5.4.4.

Once the users completed this series of sequences, they were redirected to a page informing them that the study was complete and providing them with a link to the survey. The users were asked to complete their survey while they were on the video call with the researcher present and to discuss any final thoughts about the session that they did not feel were captured by the survey. After the user completed the survey and discussed any additional feedback about the techniques they wished to provide, the video and audio recording were halted and the session was ended. All user sessions were conducted within the span of four days.

While the techniques we have discussed thus far applicable to HMD users located in public spaces with real-world objects or places that are linked to virtual objects in a MR view, there are two major constraints that have resulted in our use of mobile users in their home environments in which the "situatedness" of objects in the view is only a simulated one. First, the relative unavailability of consumer-facing MR HMD devices makes it highly unlikely that recruitment of individuals would be possible–particularly if those users must also be professionals willing to spend an hour of their time completing a user session. Second, the state of pandemic lockdown at the time that this study was conducted made in-person sessions impossible due to safety and ethical concerns, as well as organizational policy. Consequently, we were forced to limit the study to mobile users and simulated a situated view, and did not link the virtual objects with real-world ones.

### 5.4.4 Tasks

We asked users to perform a series of navigation tasks (see Figures 5.4e, 5.4d, and 5.4f), each of which initialized a sequence of analysis tasks in which the user was required to an-

swer multiple choice questions (see Figures 5.4e, 5.4b, 5.4a, and 5.4c) about an observation in the synthetic dataset described in Section 5.4.2. The sequence and permutations of tasks requested of the user are detailed in Table 5.2.

The techniques used for navigation tasks include a guided data tour using an arrow on the floor (Figure 5.3), the WIM (Figure 5.1), and a control condition in which a blue ring marker was rendered directly beneath the target POI. When the user's viewport reached a position within 0.5 meters of the POI, the user was prompted by the implementation with the first of a series of three multiple choice questions about variables represented via the position of square marks relative to three axes, each question referencing a different variable from the synthetic dataset. This sequence—navigate, then answer three multiple choice questions—was repeated six times.

The multiple choice questions during the first four repetitions referred to only one analysis technique per repetition. The final two repetitions cycled through the analytical techniques, with a different technique being applied for each question. For the first two repetitions, the guided tour arrow (Figures 5.3, 5.4f) was used for the navigation task; the following two repetitions used the WIM (Figures 5.1, 5.4d); the final two questions used the control condition of a blue ring beneath the target point of interest (Figure 5.4e). The first of the four analysis conditions used was the control condition of a 3D VRIA chart (Figure 5.4e). This sequence was followed by another sequence using the exterior wall projection view of the Shadowbox (Figures 5.2a, 5.4a), followed by the interior wall projection (Figures 5.2b, 5.4b), followed by the unfolded view (Figures 5.2c and 5.4c).

Our rationale for selecting the techniques we chose for the tasks—and for omitting the summon and dispel, and cutting plane techniques—is as follows: To begin with, we opted to omit the summon and dispel technique for either navigation or analysis tasks, because it negates the "situatedness" of the visualizations by clustering them all in front of the viewer. Another reason that we omitted the summon and dispel technique, as mentioned at

the beginning of Section 5.4, is that it is more appropriate to the larger public or outdoor spaces we initially envisioned our study taking place in, not the smaller living spaces the pandemic forced us to conduct our study in. The WIM and guided data tour both address or partly address the challenges of physical navigation, physical reach, and spatial/temporal continuity (Table 5.1), and so they meet the criteria of being appropriate for navigation tasks. Our navigation control condition—the use of a hovering ring beneath the navigation target—met the criteria for indicating that a point in space was a target of interest, but we did not feel that it was a good candidate for inclusion in our list of techniques by itself (Section 5.2), because it is more a standalone mark than a full-fledged technique. We chose the three states of the Shadowbox as our sole test condition, omitting the cutting plane, mainly for two reasons: First, the Shadowbox is arguably a novel technique contributed by this work, while cutting planes are not. Second, we believed that the introduction of a fifth analysis task condition with an interactive user input mechanism would push the task load for our users from reasonable to onerous. Finally, we did not evaluate cutting planes in part because none of our abstract datasets benefited from this technique; it is most suited for volumetric 3D representations.

In practice, we found that the time required by users to complete the tasks did indeed tend to hit near the maximum amount of time they were willing to commit for several users. The control condition for analytical tasks of having the user respond to questions using a VRIA figure, rather than a shadowbox, represents what we believe to be a reasonable default of using the targets of the Shadowboxes' projections and removing the Shadowboxes.

## 5.5 SA View Management Findings

In Batch *et al.* [22], each user was exposed to each technique multiple times, and there was a clear discrepancy between task performance during the tutorial attempt relative to

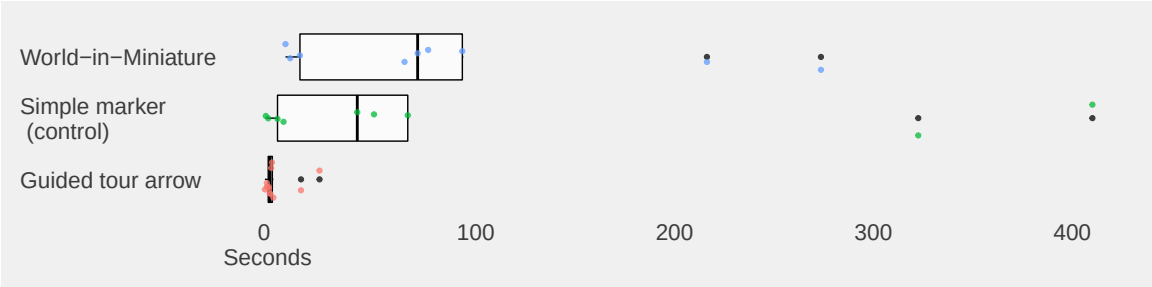



Figure 5.5: Distribution of each user's average task completion time for navigation tasks (wayfinding to target point of interest) by technique, excluding the first navigation task using each technique. The white area of the box plots begin on the left at the 25th percentile, are split at the 50th percentile, and end at the 75th percentile. The whiskers extend from the 25th and 75th percentile hinges to the farthest observation within 1.5 times the inter-quartile range of either hinge.

all following task completion attempts, as the users were still acclimating to the technique during the tutorial stage. For this reason, we opted to evaluate only observations after the first tutorial attempt per technique (i.e., to exclude the first attempts).

Despite the guided tour arrow technique being users' very first method for navigation in the AR scene, users were able to locate the target POI significantly faster than they were using the other two techniques (Figure 5.5). Users also reported feeling most confident in their success at the navigation task when they were using the guided tour in the NASA TLX, and generally did not feel stressed, under time pressure, or as if they had to physically or mentally overexert themselves while using the technique (Figure 5.10). Users took significantly longer finding their way to the target POI using the WIM; like the guided tour, the task completion time matches the users' responses to the NASA TLX, where they reported feeling least successful and generally quite negatively about the WIM's application to navigation tasks.

User question response correctness was slightly superior while using the Shadowbox's folded view with 2D charts projected onto the interior walls of the box (Figure 5.2b) relative to 3D charts, although the inverted, exterior projection had a long right tail, with several users correctly answering all questions using the exterior projection. The unfolded Shadowbox

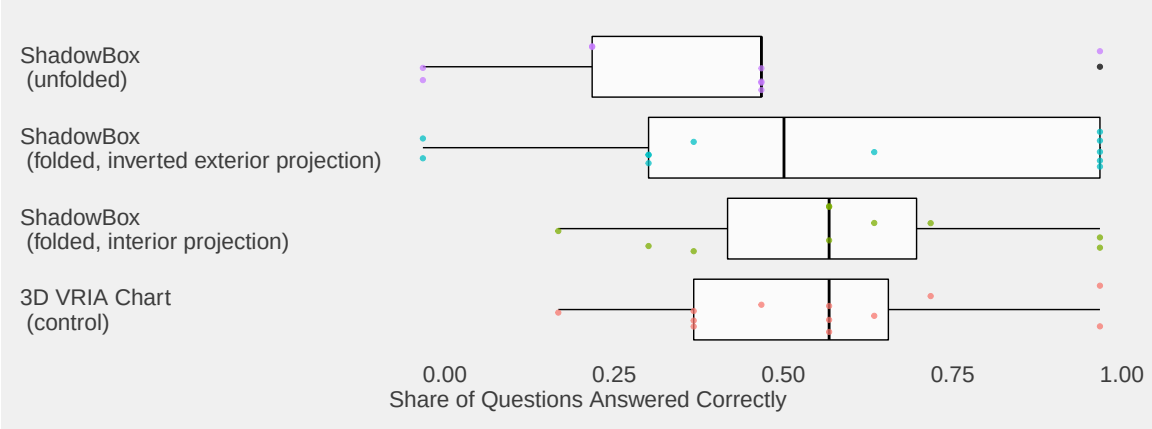



Figure 5.6: Distribution of average correctness for each user by technique, excluding the first attempted question response using each technique.

performed poorly relative to other techniques, with users answering fewer questions correctly using this view than other views.

Despite its poor performance in correctness, the unfolded view did see a faster response time than all other analysis techniques (Figure 5.7), followed by the interior projection and the exterior-wall projection views (roughly tied, with the interior projection having slightly faster mean times but a long tail of slower responses), while 3D view responses took significantly longer.

Participants' use of space is summarized in Figure 5.8. A disclaimer to these results should be provided immediately for the sake of transparency: During three user sessions, the positions of the users became untrackable, resulting in "null" values being recorded for a small portion of the experiment near its end; this issue had not been encountered during the pilots, and we were unable to trace the cause. All users were using different models of mobile devices, so the hardware cannot be pinpointed as the root cause of the matter.

Broadly, users preferred viewing the POIs with their viewport at a height between 1.2 and 1.5 meters. The exception to this is most notably the unfolded Shadowbox, which saw users' point of view angle diverge dramatically from that of all other techniques; they raised their

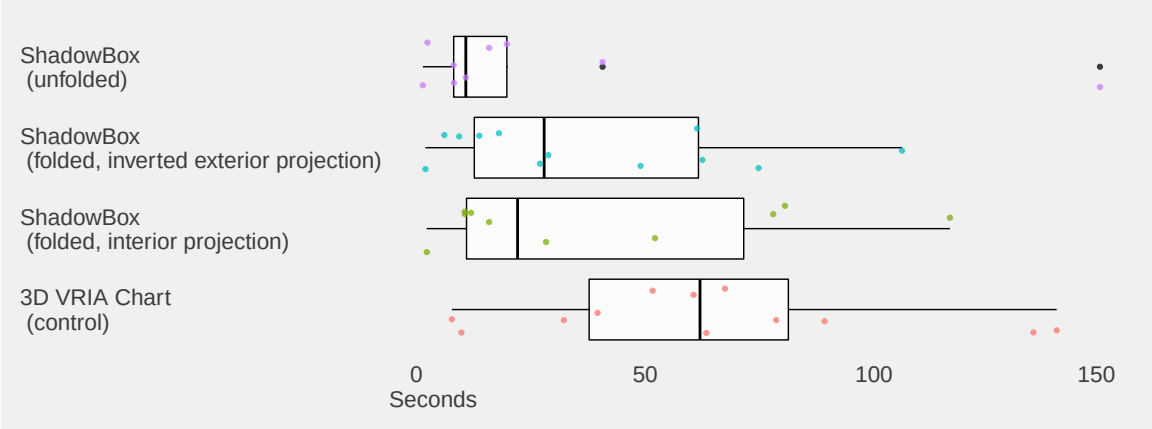



Figure 5.7: Distribution of each user's average task completion time for analysis tasks (multiple choice question answering) by technique, excluding the first attempted question response using each technique.

devices higher while answering questions using the unfolded Shadowbox than they did with other devices. They also tended to view the unfolded Shadowbox at a greater distance; this was particularly true during their first sequence of questions using the technique, and then participants tended to move closer to the center of the POI for later attempts at interpreting dataset values using this technique. In conjunction with the generally poor survey rating (Figure 5.11) and correctness (Figure 5.6) associated with the unfolded Shadowbox, along with user feedback, we must conclude that the reason for this was that users suffered some difficulty in actually viewing the panels of the unfolded Shadowbox; they also encountered minor occlusion issues as the remaining 3D VRIA objects were left in the view during this stage of the experiment.

Conversely, users tended to prefer viewing interior wall projection Shadowboxes at a closer distance than the other techniques, and moved closer to the POI during the first set of questions using the technique, but then spent more time farther away from the POI during the final question using this technique. In fact, this pattern of dramatically changing the distance and then reverting to a distance more similar to the one observed during the first

sequence appears in all techniques except for the unfolded Shadowbox.

When asked to rank their preferred techniques for navigation, users responded resoundingly against the WIM (Figure 5.9a), with the target marker coming in slightly behind the guided tour as most preferred. When asked to rank their preferred techniques for analysis, users preferred the 3D chart (Figure 5.9b), and—somewhat surprisingly—gave the interior projection view of the Shadowbox the fewest votes.

The measure of "general effort for completion" shown in Figures 5.10 and 5.11 correspond to the NASA TLX question *"How hard did you have to work to accomplish your level of performance?"*[3] Users reported finding the general amount of effort required to achieve the level of performance they did during navigation sequences of the sessions to be greatest for the guided tour. However, upon reflection and review of the transcripts and observational data discussed above, this appears to be a result of a common misunderstanding of the scale for this question, as several users who explicitly mentioned finding the WIM difficult to use and/or finding the guided tour easy to use rated the WIM as requiring less effort to achieve their level of performance than they rated the guided tour as requiring. In light of this, we must also disregard the results for this question as applied to the analysis task techniques. However, we have opted to include these results in this paper for the sake of transparency.

Despite of the possible misinterpretation of the question regarding general effort for completion, the remaining patterns in navigation tasks largely match observations, feedback, and results; users reported that the guided tour left them feeling less stressed and irritated, less pressed for time, and required less physical or mental exertion than the other navigation techniques, although it did see competition from the ring marker control condition. One user volunteered that they found the guided tour superior for finding their way to the target POI, but the ring marker did a better job of helping them pick the right POI when multiple POIs

[3]Note: We did not rephrase this question, or any other NASA TLX question; the labels are shortened only for representation in Figures 5.10 and 5.11.

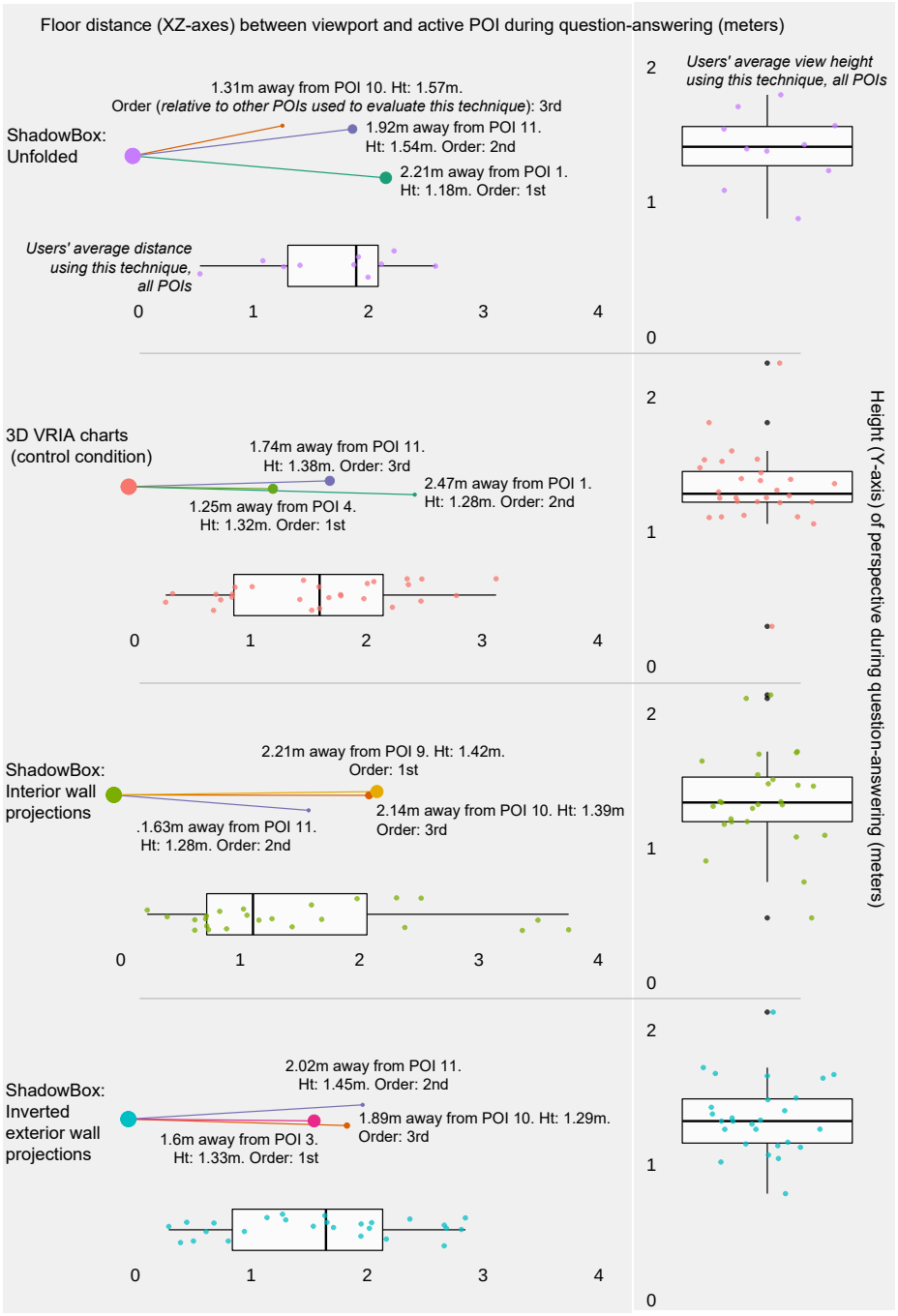
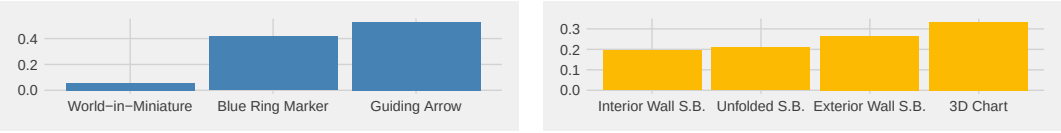



Figure 5.8: Distributions of user means in viewport heights, distance from the active point of interest (POI) during the question-answering period for each analysis technique, with mean distances by POI. Mean user-POI distances, heights, and the order in which the user interacted with each POI annotated by mark indicating distance. These results exclude the first attempted question response for each technique.




(a) Navigation technique preferences.

(b) Analysis technique preferences.

Figure 5.9: Technique ranked vote results.

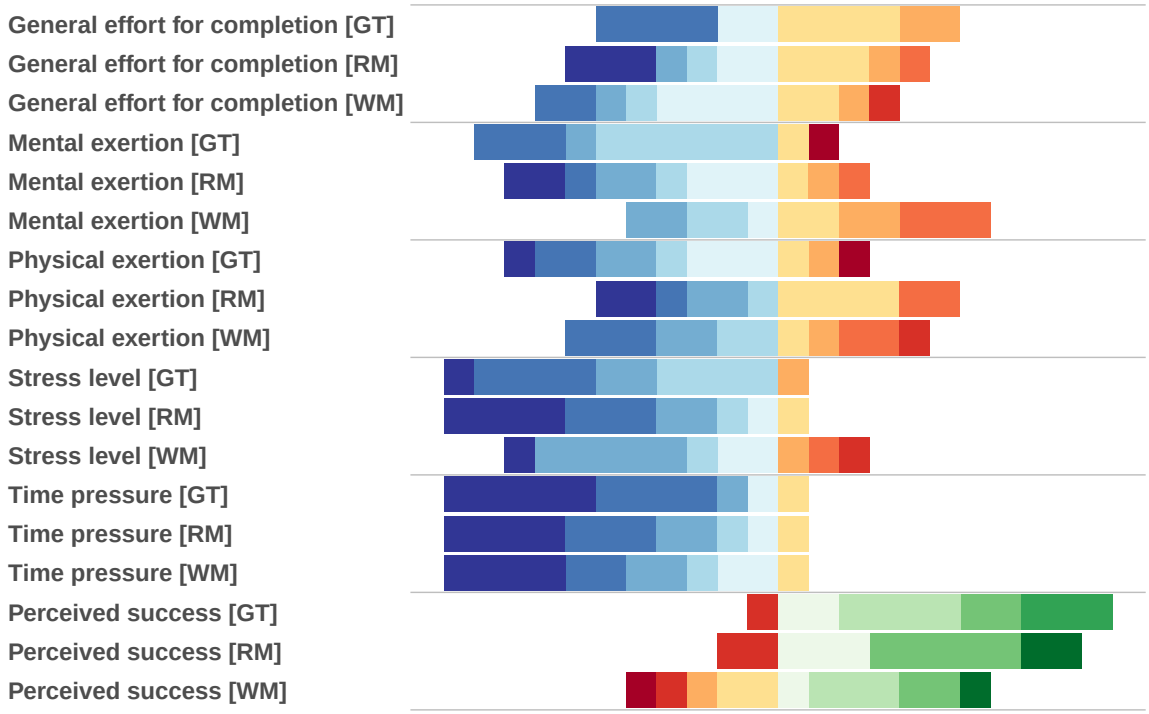



Figure 5.10: Survey response measures for navigation tasks based on NASA TLX. Yellow/red spectrum is bad, blue or green is good, color intensity represents response intensity, and bar length indicates the share of users in each category. Abbreviations: [GT]: Guided Tour; [RM]: Ring Marker (control condition); [WM]: World-in-Miniature. Note that the labels have been shortened from the phrasing used in the survey itself to improve the legibility of this figure

were situated near each other. On the other hand, other users did find the ring markers easy to pick out or fun to hunt down. The general consensus among users was that the WIM was not very effective for navigation, but that a large part of the problem was the mechanism by which the users controlled its rotation (namely, via the orientation of the phone). Users reported feeling the most successful in their task completion while using the guided tour, and the accuracy of this sentiment is strongly reflected in the task completion times shown in Figure 5.5.

As with the navigation task survey responses, we must disregard the responses measuring "general effort for completion" in light of conflicting interpretation of the question by users. In general, users seemed to feel most comfortable with the 3D VRIA chart visualizations acting as our control condition, which they reported as requiring less mental exertion, and inflicting less stress and irritation upon them relative to other techniques. The 3D VRIA charts performed generally well in the users' survey responses for most categories, and they reported feeling most successful using this technique—despite answering more questions correctly using the interior wall projection view of the Shadowbox, and taking longer to complete their tasks using the 3D chart than using any other technique. This may be a result of the view itself being somewhat more common relative to the Shadowbox views. Despite this, users did report that the 3D charts required more physical exertion than any other technique, while the interior wall view of the Shadowbox required the least. They also reported that the interior wall projection made them feel least pressed for time. The unfolded Shadowbox was received generally negatively for reasons noted above.

One aspect of the use of situated visualizations that appeared during our user sessions but has not been discussed thus far is that of fun and enjoyment. In response to open-ended questions about their experiences, four users described the ability to move around the scene as "fun" without being prompted, saying that *"*[it] *was fun to navigate"*, *"I like* [being able to move and look at charts] *from all around—such fun!"*, *"hunting around for*

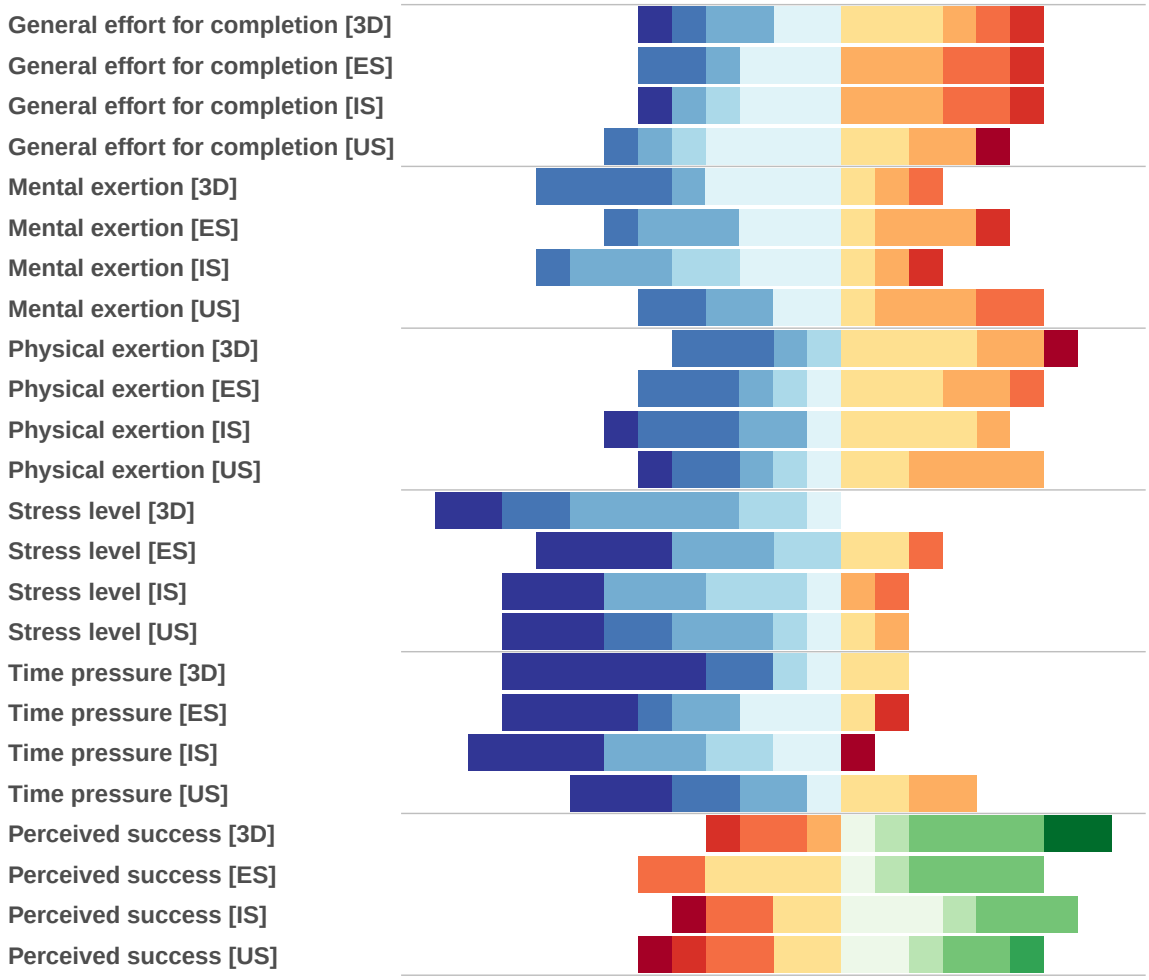



Figure 5.11: Survey response measures for analysis tasks based on NASA TLX. Yellow/red spectrum is bad, blue or green is good, color intensity represents response intensity, and bar length indicates the share of users in each category. Abbreviations: [3D]: 3D charts (control condition); [ES]: [IS]: 2D projection onto Interior walls of Shadowbox; 2D projection onto Exterior walls of Shadowbox; [US]: Unfolded Shadowbox. As with Figure 5.10, the labels have been shortened for this figure for legibility.

[the target ring marker] *actually added some fun to the session."*, and *"AR mode was fun!"* Several other users who did not mention the factor of "funness" in the survey did express similar sentiments verbally during the user sessions, and in two cases the user's onlooking partner made note of how fun it was for them to watch the session in progress despite not participating themselves.

## 5.6 A Discussion of Post-Experiment Thoughts

In Batch *et al.* [22], we presented an analysis of the challenges of view management for situated visualizations, enumerating concerns such as physical distance and reach, orientation and legibility, and depth and occlusion. These challenges apply to both the components of a single situated visualization, as well as to multiple visualizations that exist in the same physical environment. Based on this analysis, we propose a set of interaction, layout, and presentation techniques that are designed to mitigate these challenges.

Our results indicate that the interior projection view of the Shadowbox outperforms other techniques when the three factors of completion time, correctness of interpreting data, and user experience are all taken into consideration. It is closely followed by the 3D VRIA charts in correctness, slightly outperformed by 3D charts in the survey, and, like every other technique we have evaluated, beaten outright in task completion time by the unfolded Shadowbox. Our results also indicate that the guided tour arrow outperforms the other techniques when completion time and user experience are taken into consideration, given that it received the most favorable ratings in most user navigation time was significantly faster using this technique than the other two evaluated in this study. The WIM did not perform well for user navigation, but part of the blame for this result lays with the rotation mechanism in our implementation. Our results also indicated, anecdotally, that users find data exploration fun when it's in AR.

Beyond these broad results, there were a number of details in our results that were conflicting. One such detail was in the participant use of space. We suspect that there may be several of the following factors at play in explaining the pattern of the differences in users' view distance during the second task discussed in Section 5.5. The difference from first to second sequence may be the result of users becoming more competent with the system by the second sequence, correcting view distance issues encountered during the first sequence. The difference from second to third sequence may be a result of the user, now a seasoned expert, feeling free to take a leisurely stroll around the scene. Given the potential for confounding factors, we examined the data for outliers, but after reproducing Figure 5.8 several times with one potential outlier user omitted each time, we found that the results did not significantly diverge from Figure 5.8, which includes all users.

There is always the possibility of measurement error as an explanation; if that is true in this case, it may be the result of some of the limitations of the WebXR utilities used in the design of our implementation, which represent the current state of the art in AR-in-the-browser applications. When the user initializes an AR session, the scene camera no longer shares user position. As a workaround, we attached a `three.js` Object3D to the camera, and the world position of this object could be used to derive the user's position. However, upon review in VR views of the scene, the coordinates logged by the Object3D did not always perfectly correspond to those of the camera.

A different issue related to remote nature of the study was user interpretation of subjective survey measures that could possibly have been avoided if the researchers and users had been sharing a room, as it may have been easier to notice the discrepancy between user responses relative to user remarks and researcher observations. The brief window of time during which user studies were conducted, and lack of opportunity for a chance observation, precluded a thorough review of survey responses as they were completed. Future remote research in WebXR should mind such pitfalls in data collection, but we are enthusiastic about its

potential.

# Part III

# Part 3: Models for Interpreting User Session Data

# Chapter 6: Gesture and Action Discovery

With the exception of models such as GOMS, Fitts' law, and the Steering Law, there are few theoretical models in human-computer interaction (HCI) sufficiently sophisticated to enable formal verification. For this reason, validating a new technique, innovation, or system in HCI often leaves empirical evaluation as the only available recourse. It is not uncommon that such empirical evaluation reduces to systematically observing and coding video recordings of users engaging with the interactive system. This is particularly true for virtual reality (VR) systems, where the interaction to be evaluated may involve the user's whole body. However, such coding is generally costly, time-consuming, and prone to inconsistencies [134]. Furthermore, it often requires multiple coders agreeing on and calibrating a common code book. Finally, the very nature of this process injects subjective biases that make the experiment results difficult to reproduce [172].

To address this issue, we propose a semi-automated computer vision system for behavioral coding of videos that will make the process more robust and scalable. Existing systems and action recognition studies tend to focus on actions that are familiar and meaningful in a larger range of contexts such as walking, running, eating, opening the fridge. These studies are able to take advantage of large amounts of publicly-available video data, but these datasets are not typically applicable to usability testing of a novel VR system. In VR, users perform specialized actions to interact with objects in the virtual environment using a custom interface. The actions may take the form of moving parts of the body in a non-generalizable way such as swinging arms diagonally, or tilting their head. These actions

do not necessarily map perfectly to real-world scenarios that may be assigned semantic labels in existing video datasets.

In our scenario, researchers might be more interested in outlier actions and want to manually filter out certain actions indicative of bugs in the system in unlabeled videos. An automated system that segments videos into potential actions of interest (AOI) and indicates these in unlabeled videos would make the process faster, more scalable, and easier to reproduce. Our approach also draws on additional data—such as depth, audio, tracked marker positions, etc—synchronized with video for efficient identification of AOIs. We believe that this approach is well-suited to most contemporary VR devices, which rely on sensors to detect the positions of the user's controllers and head-mounted display (HMD). In a user study involving the collection of video data, this ground truth can be used for video segmentation.

Our pipeline can be used by HCI researchers who can collect video and telemetry data capturing users during sessions in which the user explores the virtual environment. After all user data has been collected, telemetry data is segmented, and then segments are clustered into micro-gesture classes, using a set of statistical methods described in Section 6.1. Video data is temporally labeled with gesture codes based on telemetry segmentation, and predicting these gestures becomes the training/testing target of our neural network architecture. While our results leave some room for improvement, we believe that—given the lack of a semantic ground truth—our model performs with reasonably high accuracy relative to the current state of the art in action discovery.

## 6.1 Observational Data Modeling

In Batch *et al.* [20], we proposed a novel pipeline for semi-supervised behavioral coding of videos of users testing a device or interface, with an eye toward human-computer

interaction evaluation for virtual reality. Our system applied existing statistical techniques for time-series classification, including e-divisive change point detection and "Symbolic Aggregate approXimation" (SAX) with agglomerative hierarchical clustering, to 3D pose telemetry data. These techniques create classes of short segments of single-person video data–short actions of potential interest called "micro-gestures." A long short-term memory (LSTM) layer then learns these micro-gestures from pose features generated purely from video via a pre-trained OpenPose convolutional neural network (CNN) to predict their occurrence in unlabeled test videos. We present and discuss the results from testing our system on the single user pose videos of the CMU Panoptic Dataset.

Figure 6.1 shows the overall pipeline for our system. The videos in our dataset had synchronized 3D pose data available that was used in the training phase. The output of this part of the pipeline was a list of pseudo-ground-truth labels for a selection of "micro-gestures" detected in this 3D data which act as our AOI.

1. **Clustering Phase:** The synchronized 3D pose data is converted to features indicating temporal variance using e-divisive change-point detection followed by SAX edit distance matrix transformation. A hierarchical clustering method is then used to group these features into clusters that indicate similar video segments. A researcher can then be presented with an interface that displays the identified video segment groups to check for qualitative similarity and a set of potential AOI.

2. **Training Phase:** The AOI video segments act as training data for the following LSTM network. The input frames are converted into pose feature vectors using a pre-trained CNN and the output of the LSTM is an action label for this input frame.

3. **Testing Phase:** In the testing phase we only use the unlabeled video to predict an action label for every frame. The ground truth for these is the output from the hierarchical clustering.

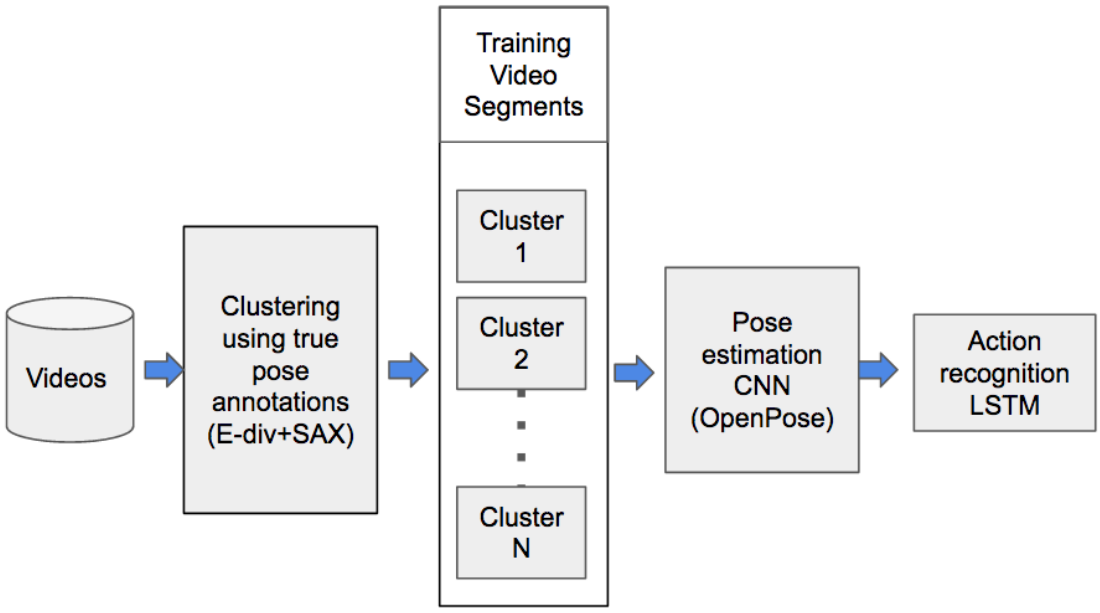



Figure 6.1: Our overall pipeline.

### 6.1.1 Statistical methods

We exploit the spatio-temporal continuity of the human body by using sensor data tracking human joint positions to temporally segment full-length video into short (less than 15 seconds) micro-gestures. Before beginning the process, we select eleven angles ($\theta$) between 15 joints from the CMU Panoptic dataset (Figure 6.2).

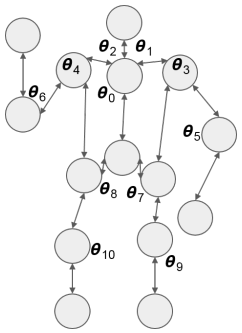



Figure 6.2: Subject joint angles.

#### 6.1.1.1 Joint Angle Segmentation

Matteson and James [155] originate the e-divisive method for detecting changes in the mean of multivariate time series: Estimated temporal divergence measure for any two joints $\theta^X, \theta^Y \varepsilon\ \mathbb{R}^d$ is given by Equation 6.1,

$$\widehat{\mathscr{E}}(\theta_n^X, \theta_m^Y; \alpha) = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |\theta_n^X - \theta_m^Y| - \gamma(\theta_n^X; \alpha) - \gamma(\theta_m^Y; \alpha) \tag{6.1}$$

where $\alpha$ is some value between 0 and 2; following James and Matteson [112], we use $\alpha = 1$. $\gamma$ is given by Equation 6.2.

$$\gamma(\theta_n; \alpha) = \binom{n}{2}^{-1} \sum_{1 \leq i < k \leq n} |\theta_i - \theta_k|^\alpha \tag{6.2}$$

This gives us a measure of scaled empirical divergence between joint angles, $\widehat{\mathscr{Q}}(\theta_n^X, \theta_m^Y; \alpha)$, described in Equation 6.3.

$$\widehat{\mathscr{Q}}(\theta_n^X, \theta_m^Y; \alpha) = \frac{mn}{m+n} \widehat{\mathscr{E}}(\theta_n^X, \theta_m^Y; \alpha) \tag{6.3}$$

Then the locations of change points ($\tau$) can be estimated as shown in Equation 6.4.

$$(\widehat{\tau}, \widehat{\kappa}) = \underset{\widehat{\tau}, \widehat{\kappa}}{\operatorname{argmax}}\ \widehat{\mathscr{Q}}(\theta_n^X, \theta_m^Y; \alpha) \tag{6.4}$$

In Figure 6.3, these change points are demonstrated as red vertical lines laid over the time series representation of $\theta_1$ from Figure 6.2. This can be generalized to any number of variables.

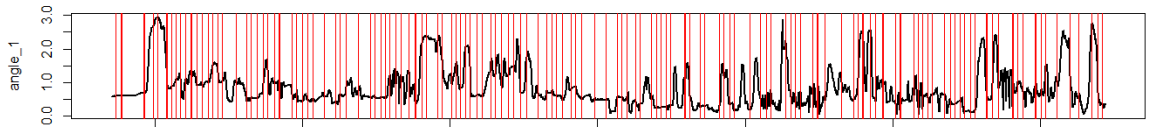



Figure 6.3: Example of joint angle time series segmented based on change point estimates ($\hat{\tau}_k$).

#### 6.1.1.2 Symbolic Representation

Once the time series representation of the human joint angles have been segmented based on change points, we transform the segments into strings to derive an edit distance matrix of the symbolic representations of each segment, reduced to a constant number of observations (or "characters"). Lin et al. [139] introduce this method of string representation for time series clustering, known as SAX, as being performed in the following steps:

1. Normalize the segment around a mean of zero.

2. Piecewise Aggregate Approximation (PAA) [121]: Convert a series $x$ of length $n$ to series $\bar{x}$ of $N$ dimensions as shown in Equation 6.5.

$$\bar{x}_i = \frac{N}{n} \sum_{j=\frac{N}{n}(i-1)+1}^{\frac{N}{n}i} x_j \tag{6.5}$$

   This effectively stretches or squashes all of the micro-gesture joint angle segments to the same length.

3. Symbolic representation: Given a series segment converted to a specified length $N$ with normal distribution around zero, assign letters to values along the segment (Figure 6.4).

4. Derive the edit distance matrix. We use Levenshtein edit distances: The lowest number of transformations—character insertion, substitution, or deletion—required to turn

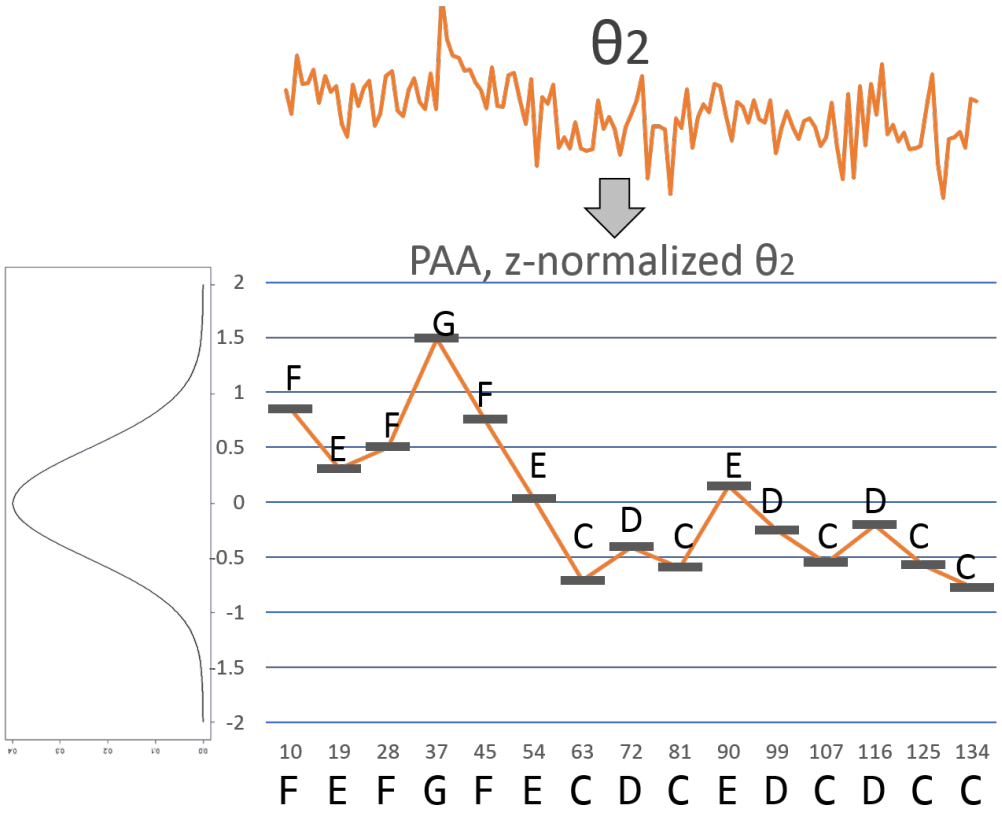



Figure 6.4: SAX: PAA-converted, z-normalized, character representation of $\theta_2$ for a single segment determined by change points estimated by e-divisive.

one word into another.

We modify this approach somewhat by padding our segments with their surrounding neighbors: Each micro-gesture $\mathscr{G}_t$ is grouped with the segments $\mathscr{G}_{t-1}$ and $\mathscr{G}_{t+1}$ prior to applying the first step of the above process. At the end of the clustering process described in Section 6.1.1.3, the micro-gesture label is only assigned to segment $\mathscr{G}_t$.

### 6.1.1.3 Semi-Supervised Clustering

We create clusters from the edit distance matrix created from the process described in Section 6.1.1.2 by applying a fast iterative agglomerative clustering algorithm implemented

by Müllner [166]: Individual nodes are grouped together with their nearest neighbor, values are stored for the group, and the next nearest neighbors are grouped together, iterating until all nodes are in one group. For this paper, the number of clusters is determined by the frame rate and the length of the video: $k = \mathscr{T}/\text{fps}^2$, where $\mathscr{T}$ is the total number of frames in the tracking data; in this case, $\bar{k} = 166$.

Figure 6.5 represents the clusters used in this paper, given arbitrary labels $\mathscr{G}_{1,2,\ldots,166}$ (not shown). The human-in-the-loop comes into play at this stage in two ways: First, the researcher must determine $k$; second, the researcher must select a subset of videos to train. In application, the HCI researcher could, at this point, view samples of all micro-gestures by cluster, select the groups that constitute their AOI to be trained on, and walk away from the model.

For this paper, we simply select the ten most frequently observed micro-gestures across all videos. Prior to initiating model training, it is also worth noting that extra frames are at the end of a segment

Once the researcher chooses a subset of gesture classes, a randomly selected sample of 70% of all micro-gesture segments with a label in the set of AOI is marked for training, the remaining 30% is marked for testing, and a proportionally equivalent random selection of micro-gestures without labels in the chosen set are selected for training and testing as well. We take this approach because our experiment involves comparing models trained with a no-gesture class in the target vector (a detector) against models trained which simply does not count no-gesture frames in its target vector. This approach is described in greater detail in Section 6.2.

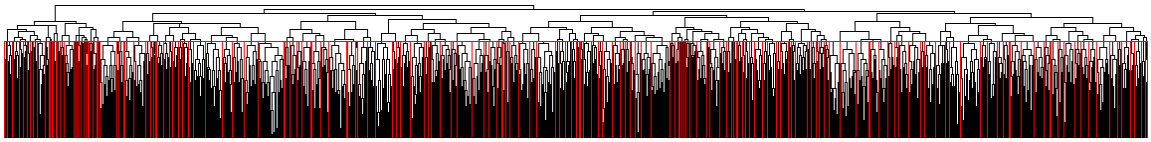


Figure 6.5: fastcluster method, with hierarchical tree cut at depth of $k = 166$.

### 6.1.2 Deep Learning Network

Given the labels generated by the statistical methods discussed in Section 6.1.1, we train a deep learning model to recognize an action from videos. A naive approach would be to recognize an action by giving a raw video directly to the model. This approach, however, would make this action classification problem intractable, as it induces huge computations costs to solve the problem in a pixel-level space. Instead, we first project this classification problem into a human-pose-level space by getting human pose data through a CNN and then use a LSTM network architecture [100] to recognize an action given a sequence of human poses, which is similar to prior work [243]. As noted by Walker et al. [243], projecting this problem into a human-pose space would reduce the problem complexity as it reduces the computation costs in the deep learning model. Our deep learning architecture is shown in Figure 6.6.

#### 6.1.2.1 OpenPose CNN

To obtain human pose data from an image, we utilize the OpenPose CNN model [40]. In this model, a human pose is estimated by first finding human joints and part affinity fields (PAFs) from an image and then combining the joints using the PAFs information. The PAFs data consists of vectors that contain the connection information between the joints. With PAFs, the CNN model can correctly estimate an appropriate edge only between relevant joints and thus generate a human pose skeleton data. This model is publicly available on

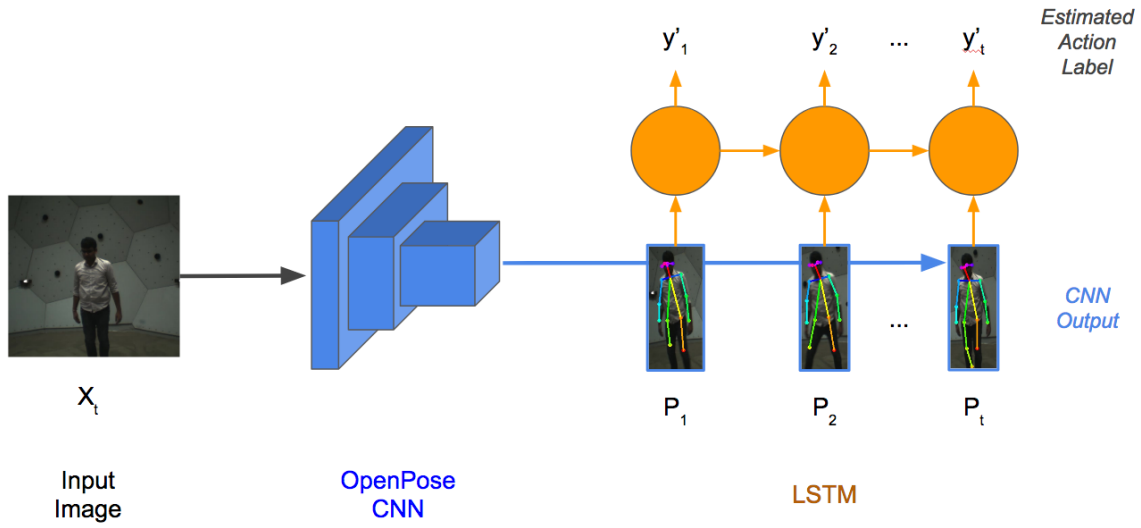



Figure 6.6: Our deep learning architecture: Input images are given to the pre-trained OpenPose CNN, the images and joint positions are given to a LSTM unit trained to predict the labels assigned (as described in Section 6.1.1), and the LSTM remembers and forgets input and output information from previous frames.

GitHub [1]. As shown in Figure 6.6, we first pass an image frame into the OpenPose CNN model to collect its human pose data. This human pose data is then used in the following recurrent neural network, LSTM, to perform action classification.

#### 6.1.2.2 LSTM

Based on human pose data generated by the OpenPose CNN model, our recurrent neural network LSTM [100] is used to estimate its action label. As the LSTM architecture is used to remember important values over arbitrary time intervals and forget other values, we force the model to learn only the features that can be useful for action classification during training. Since we are focusing on an action classification problem, we place the softmax layer on top of the output of the LSTM network to obtain a label inference. During training, the cross entropy loss function (Equation 6.6) is used to enforce the model to encode in the

[1] `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

hidden states any features relevant to action estimation.

$$L = -\sum_{i} y_i \log(y_i') \tag{6.6}$$

## 6.2 Experiments in Computer Vision for Pose Grouping

To test our pose detection pipeline, we use a publicly-available 3D pose dataset as a benchmark, and evaluate constant, adaptive (without a no-gesture class), and adaptive (with a no-gesture class) training results.

### 6.2.1 Dataset

For our experiments we used the CMU Panoptic Pose Dataset [116] consisting of videos collected from simulated social settings in a massively multiview environment. The dataset contains data from 480 VGA cameras (25 FPS, 640x480), 31 HD cameras, and 10 RGB-D sensors. We chose this dataset because of our initial focus on using video data to evaluate immersive (specifically, VR) environments; we wanted to use data that focused on 3D pose of the video subjects, and we wanted to be able to get consistent outcomes regardless of the point of view of the camera. As we were working on this, there was no publicly available video dataset of users wearing VR head-mounted displays that also had synchronized position coordinates. We had to use some alternative dataset, and unfortunately, it meant choosing the 3D position sensor data over VR equipment being present in RGB video data. Because the Panoptic dataset has both 3D joint position ground truth labels and perspectives of the same session sequence from multiple points of view, it made the optimal choice for our purpose. We opted to do use an existing dataset rather than create our own because our focus was mainly on constructing the pipeline, and we felt that constructing a system for collecting video data from multiple perspectives, getting participants with whom to collect

the data, and then labeling the dataset was out of scope for this work.

### 6.2.2 Methods

We trained our system using VGA videos split into 5,603 segments of varying temporal length, along with the segments' associated synchronized 3D pose data for 15 joints, using a single GPU (EVGA GeForce GTX 1080 SC). The target output is a vector representing the labels generated by the combined $e-divisive \rightarrow SAX \rightarrow fastcluster$ technique described in Section 6.1.1.

For the deep learning model, we use the OpenPose CNN model [40] and a single-layer LSTM network [100] consisting of 256 hidden units. We compare results from three different approaches:

- A **constant detector network** that attempts to discern between no-gesture frames and frames with *any* AOI and then classify them using a constant learning rate of 0.0001.
- An **adaptive detector network** that, like the previous network, targets a ground truth that includes a no-gesture class using a learning rate starting at 0.001 that linearly decreases each epoch by 0.000018 until it reaches 0.00001.
- An **adaptive AOI classification network** without a no-gesture class that uses the same linearly-decreasing learning rate as the adaptive detector network.

We use the Adam optimizer to train the LSTM network, taking the two different approaches to learning rates noted above. As mentioned in the previous section, the cross entropy loss is used to calculate loss between the ground-truth labels and the estimated labels during the LSTM training. As the OpenPose CNN well estimates human poses from videos in the target dataset, we do not train this CNN model — the performance of the OpenPose

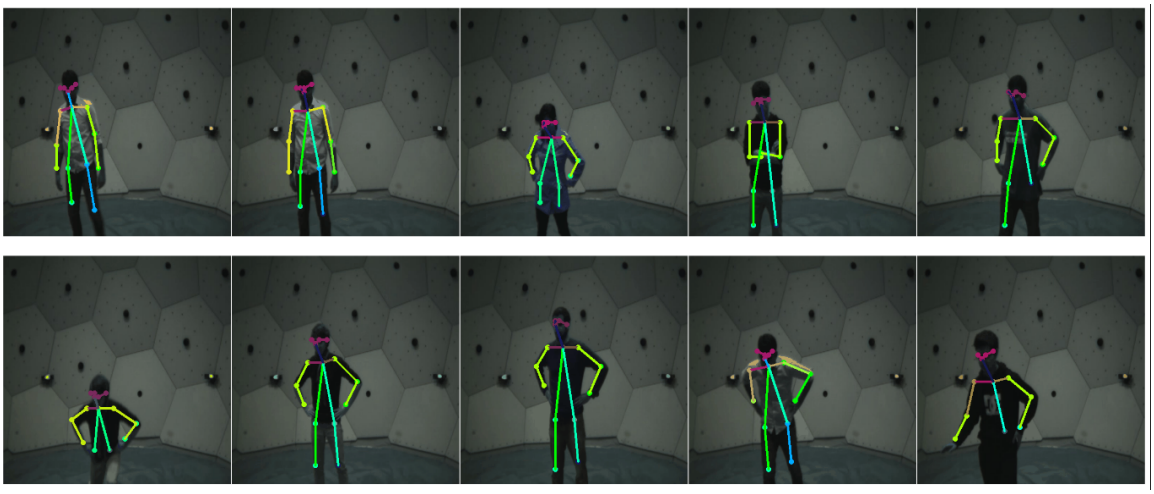


Figure 6.7: Skeleton overlaid onto frames by the pre-trained OpenPose network within a group clustered by statistical methods described in Section 6.1.1.

CNN model is evaluated in the next section. The OpenPose CNN is simply utilized to generate human pose data during the training and testing phase of the LSTM network

### 6.2.3 Qualitative Results

To evaluate the performance of the OpenPose CNN model, we first tested the pose model with some of the action videos. As shown in Figure 6.7, the model successfully estimates human poses from the videos. Based on this observation, we decided not to train the CNN model, but simply use it to generate human pose data for the following recurrent neural network to simplify the problem; in other words, the action classification problem is projected into the human pose space, which size is smaller than that of the pixel space.

Our qualitative results intuitively demonstrate both a shortcoming and a nice feature of our implementation: Because of our agglomerative hierarchical edit distance clustering, gestures are grouped together into what may accurately be described as “like classes,” but there is a trade-off between the amount of data available within a class and the similarity any one micro-gesture has across all of the other micro-gestures it is classed with. As such,

Figure 6.8: Constant learning rate detector model output of predictions for gesture label shown in Figure 6.7. Gestures have been qualitatively subgrouped and shown in bounding colors by the authors to highlight apparent similarity, with one micro-gesture in the lower left corner being visually similar to both the green and the purple subgroups. Conversely, the center micro-gesture shows little similarity to any other micro-gesture in the class.

we see what could arguably be defined as multiple classes that overlap somewhat in both our modeled input and in our test output (Figure 6.8).

### 6.2.4 Quantitative Results

Our results indicate that the trained model performs best with a constant learning rate of .0001, and is slightly better at detection than classification.

Table 6.1 shows the highest accuracy scores of all approaches used, that of the detector using a constant learning rate of 0.0001. The most surprising result of our study is also shown in Table 6.1: While training accuracy continues to rise and training loss falls at each epoch (Figure 6.9), the test results from our first epoch are greater than or equal to those of any further training epochs.

The shrinking learning rate detector performs poorly relative to the previous approach, and quickly overfits by epoch 30, as demonstrated in Table 6.2 and Figure 6.10. However, a

Table 6.1: Constant detector with a "no-gesture class" included. For all tables, best-performing accuracy scores are shown in bold.

| Epochs | Training Accuracy | Training Loss | Test Accuracy |
|---|---|---|---|
| 1 | 0.385990 | 1.969203 | **0.366118** |
| 5 | 0.391999 | 1.943364 | **0.366118** |
| 10 | 0.385557 | 1.896348 | 0.355147 |
| 15 | 0.391457 | 1.831509 | 0.339201 |
| 20 | 0.418742 | 1.796008 | 0.344304 |
| 25 | 0.427891 | 1.762800 | 0.339074 |

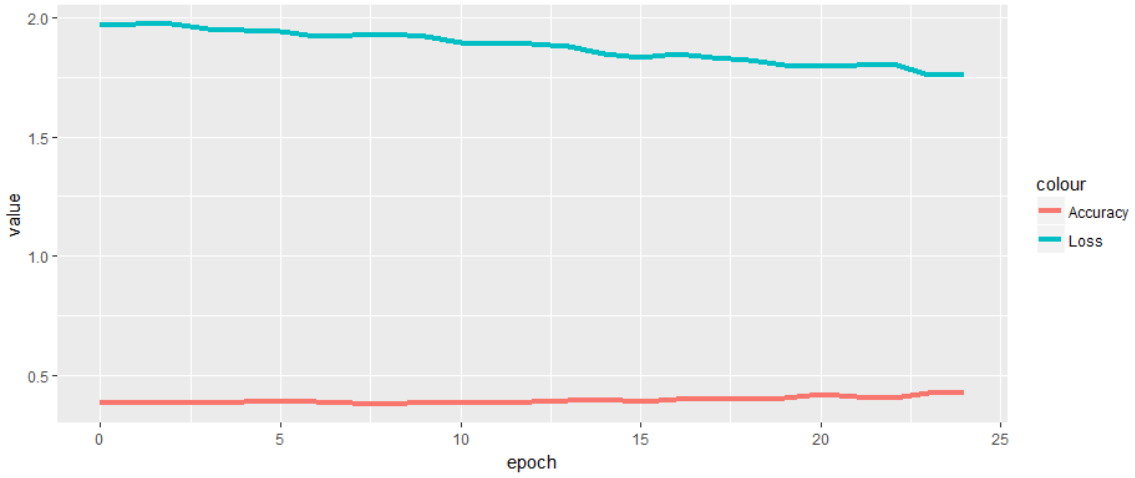



Figure 6.9: Training loss and accuracy for our detector with constant learning rate .0001.

Table 6.2: Adaptive detector with a "no-gesture class" included.

| Epochs | Training Accuracy | Training Loss | Test Accuracy |
|---|---|---|---|
| 5 | 0.505847 | 1.563875 | 0.067457 |
| 10 | 0.513859 | 1.559077 | 0.067571 |
| 15 | 0.514508 | 1.560636 | 0.087170 |
| 20 | 0.515808 | 1.558239 | 0.073952 |
| 25 | 0.517757 | 1.557487 | **0.326231** |
| 30 | 0.517757 | 1.557487 | 0.063241 |

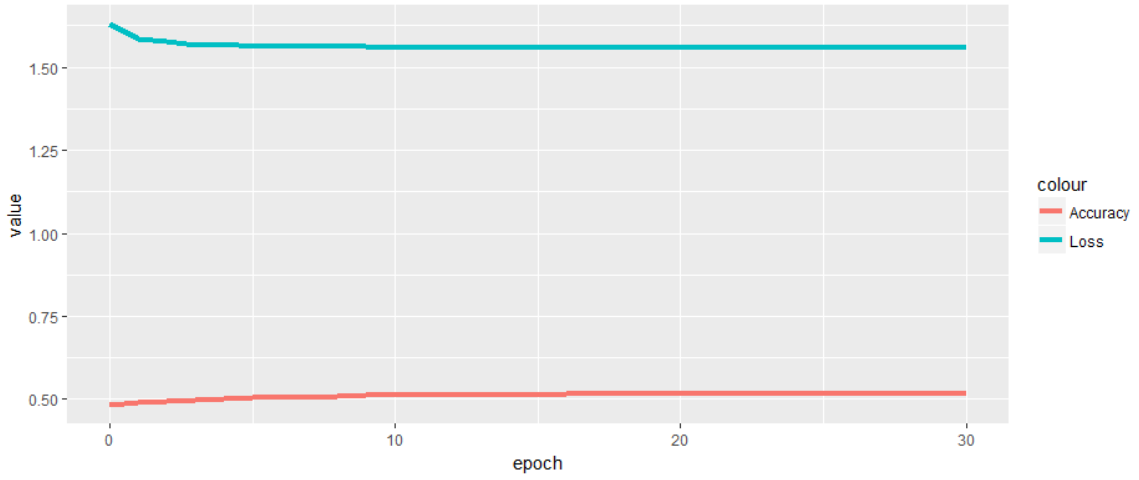



Figure 6.10: Training loss and accuracy for our detector with adaptive (linearly diminishing) learning rate.

Table 6.3: Adaptive classifier model without a "no-gesture" class.

| Epochs | Training Accuracy | Training Loss | Test Accuracy |
|---|---|---|---|
| 1 | 0.575123 | 0.983921 | 0.086918 |
| 50 | 0.639233 | 0.979427 | **0.147600** |
| 105 | 0.639233 | 0.979359 | 0.119579 |
| 160 | 0.641106 | 0.979615 | 0.106015 |

surprising result is highlighted in Table 6.2: Surrounded by poorly-performing epochs at all other tested stages of training, epoch 25 shows an accuracy score comparable to those seen in Table 6.1. Comparable results were seen with multiple tests of the trained model stored from epoch 25 of the shrinking-rate detection model.

For classification using a shrinking learning rate, the optimal number of epochs appears to fall in the range $[2, 105)$–probably closer to a value in the middle of the range, if the trend in test accuracy is smooth, or in the range $[2, 30)$, if our adaptive detector's results can be generalized. The classification model performed more poorly than the detector, but due to resource and time constraints, a constant classification model was not included in this study.

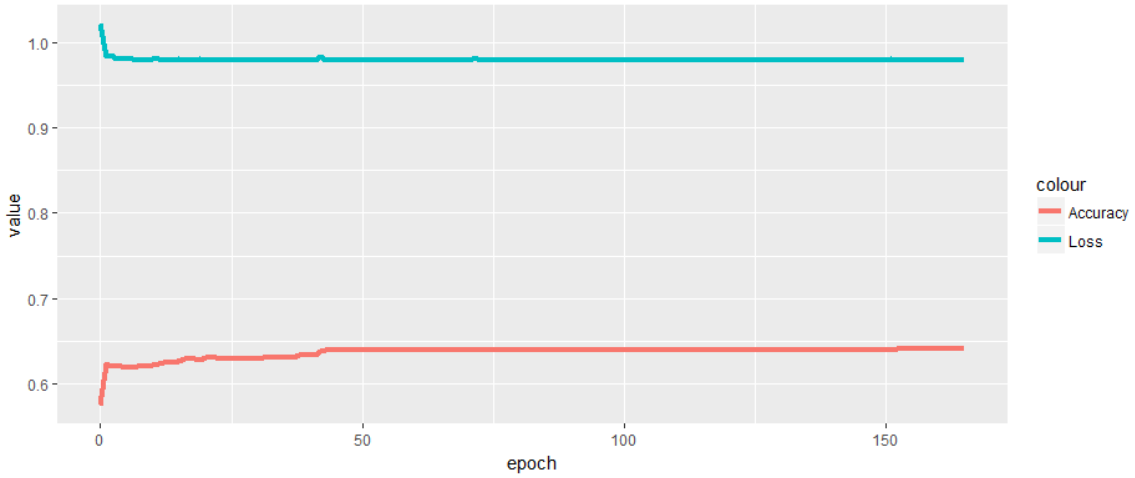



Figure 6.11: Training loss and accuracy for our classifier with learning rate starting at 0.001, decreasing linearly to .00001 over 500 epochs (note: only 160 epochs were trained during this study).

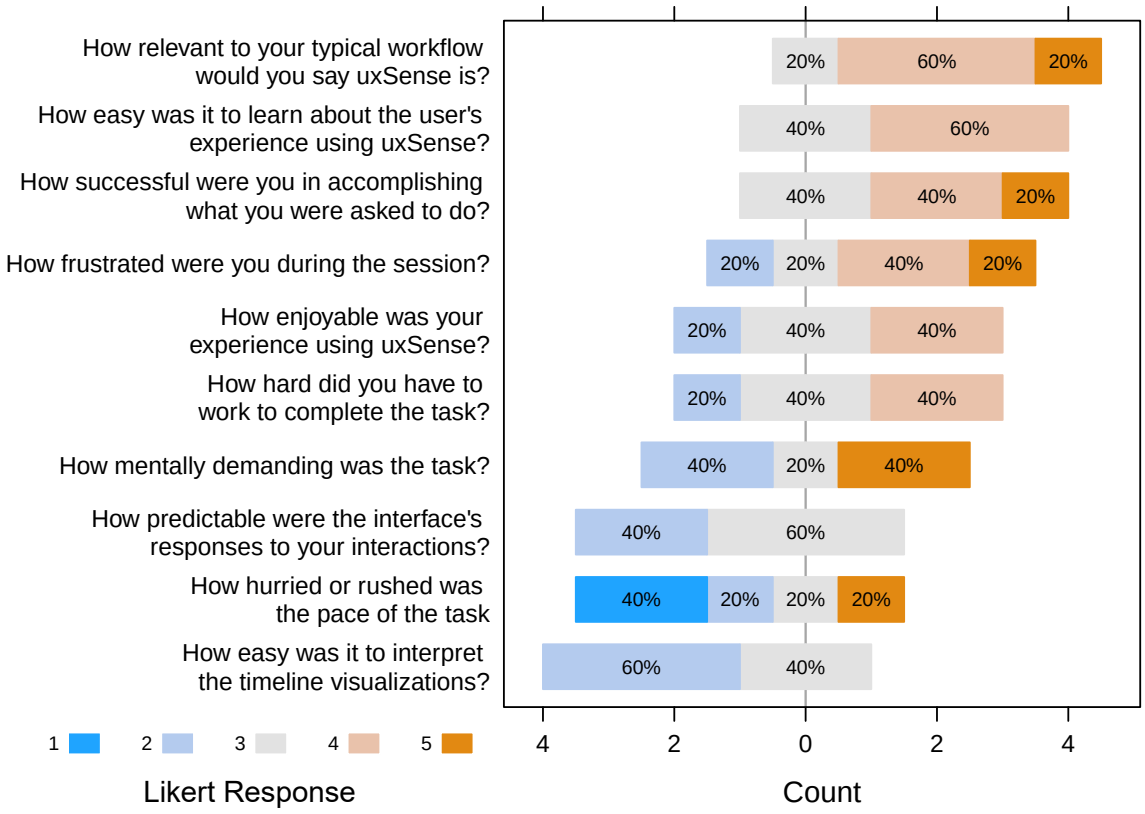



Figure 6.12: Likert responses to user experience survey.

## 6.3 A Discussion of Our Pipeline Results

We have presented our pipeline for discovering novel human actions from video and telemetry data using both statistical methods for time series analysis and deep learning networks for video analysis. Using the e-divisive and SAX methods, we grouped AOI into several classes. Given these pseudo action classes, we used a sample of frames from the Panoptic dataset to train a simple LSTM network. By training the recurrent neural network, we were expecting the deep learning model to detect and classify human actions from videos in which a human is an active agent.

Relative to architecture featuring existing semantic labels, our model did not perform overwhelmingly well, with the best observed test accuracy (0.366118) resulting from a model used simply to detect whether there is *any* AOI in the view that was trained for only one to five epochs. It is possible that the accuracy of our approach may be improved in future implementations by modifying hyper-parameters in the LSTM network: More epochs, larger or smaller learning rates, and so on. Furthermore, creating additional layers in the LSTM network has the potential for improving future iterations of this work; for example, a stacked LSTM network has multiple hidden LSTM layers, which allows for greater model complexity. Finally, future work should thoroughly review the statistical methods used and their outputs (action clusters) to ensure that they generate valid datasets.

Our results do not appear extremely successful relative to existing video action classification work using pre-existing labels. Beyond the reasonable allowances that may be made for an architecture that creates novel gesture labels as a pseudo-ground-truth, we believe that the low accuracy score is largely due to three reasons. First, hyper-parameter configurations have not been fully explored and optimized. In our deep learning model, there are various hyper-parameters–epochs, learning rates, and the hidden state size. Although finding effective hyper-parameters for a deep learning network is notoriously known to be

difficult, it is essential to try with various combinations of hyper-parameters. We found that our adaptive learning rate, which we had hoped would improve our test accuracy, had the opposite effect, and overfit the model early in the training process.

Second, the portion of the background class ("NA") in the training and testing dataset should be carefully evaluated. In early stages of our experiments, we used actions labelled as "NA," which means that the actions do not belong to any of our predefined clusters, and classify them as one of the action classes—a catch-all "no-gesture class" in our target ground truth. Our expectation was that the NA label would make training difficult for our deep learning mode due to the fact that the actions in the NA class may not have consistent features in the human pose data. To bypass the effects of having the NA class in the dataset, we spent most of our training time on a model that does not consider these actions as a class; instead, they are simply represented as a non-clustered action (no class), and the target ground truth for the associated frame is a zero vector. Our results show, however, that the detector network has significantly higher test accuracy–even the adaptive detector has at least one epoch outperforming any of the observed test accuracy scores of the classifier network. However, we trained our classifier network with an adaptive learning rate, which appears to perform poorly relative to the constant rate.

Finally, but most importantly, the statistical methods may not generate what might be considered "valid" action clusters. After segmenting the videos into action clusters generated by the statistical methods, we manually reviewed all of the actions within a certain cluster (a certain class). Throughout this review process, we observed that segments within the single cluster do not always all appear to have the same gestures; rather, they may contain dissimilar or semi-similar gesture sub-classes that are "bridged" by one or two video segments featuring gestures that are somewhat similar to two or more other sub-classes in the cluster (Figure 6.8). Thus, our training and testing datasets may not be valid. This implies that our deep learning model may not learn features from action videos appropriately,

so our statistical methods warrant additional examination in future work.

In spite of somewhat limited results, we believe that this approach is relevant for HCI researchers as a means of reducing the time cost of coding user interactions. In particular, we believe this to be true for HCI research involving implementations within a VR environment. Most contemporary consumer-ready VR devices feature sensors for detecting object position and orientation. The VR experience is three-dimensional, and involves full-body gestures. Our pipeline is applicable for inferring three-dimensional points on the human body in relatively small video datasets with telemetry data, but without existing semantic labels. Video data of this nature could be collected during a user study in a typical academic research lab environment. Our methods are feasible on hardware that is a minimum system requirement for most contemporary VR HMDs: A single, yet reasonably powerful GPU. As such, in the context of HCI research, future implementations of the method we present should focus more narrowly on the specific problem space of evaluating novel VR applications.

# Chapter 7: UX Evaluation using Visualization and Computer Vision

Having developed a model for computer-supported detection of important moments in video data using human pose detection, we wanted to return to our original goal in using learning models: Supporting evaluation of user sessions. To do this, we combine models for video and audio data with interactive visualization and introduce a visual analytics system, uxSense, that presents the UX analyst or researcher with time series data outputs from said models.

## 7.1 Framework: Extracting Behavior from Video

Our framework revolves around the notion of using entirely automatic methods in CV, ML, and signal processing to present semantic information to the UX researcher in a visual interface that supports user session video evaluation and report generation. We discuss our design constraints, data model, and applications below.

We consider the following aspects of the overall design space:

D1 **Video-based data:** While it is certainly possible to combine our sensing mechanisms with data from clickstreams, IR motion tracking, and sociometric badges, our scope here is currently limited to extracting behavior from video files, including live imagery and its associated audio.

D2 **Multiple concurrent streams:** Instead of trying to derive every conceivable feature of a given video in a single pass, we anticipate individual modules generating specific

data streams from the same footage. Specific modules could include tracking the user's head orientation, physical location, activity level, speech, etc.

D3 **Physical navigation:** A crucial feature to be extracted is the user's location over time in a physical space. The accuracy of this measure is likely lower than specialized forms of tracking, but we still anticipate achieving reasonable performance.

D4 **Gross motor interaction:** While fine motor interactions, such as which button on a touch interface the user tapped, are challenging to track due to low camera resolution and occlusion, gross motor interactions (i.e., limbs, torso, head, and facial landmarks) are relatively easy to track with high accuracy.

D5 **Facial expressions:** The ability to track facial expression—potentially curtailed by occlusion due to movement or head-mounted displays partially obscuring the user's face—may yield an insight into the user's emotional state.

D6 **Audio features from video:** User session video typically includes audio that captures dialogue between the researcher and the participant, as well as think-aloud or pair analytics procedures that involve the participant verbally externalizing their sense-making process. A system for evaluation should take advantage of this information by generating a transcript from speech, visually representing features of the audio signal, and using it in computer vision models for predicting semantic activity.

D7 **Annotation and report generation:** UX researcher evaluation workflows often involve annotation of key moments in user sessions. These annotations should play a central role in report generation, as they are one of the main products of such reviews.

| Name | Filter Description | Data Stream | Model |
|---|---|---|---|
| VideoPose3D | 3D user position | Pose time series | Pavllo et al. [185] |
| E-Divisive | Joint angle intervals | Frame segments | Batch et al. [20], Matteson et al. [155] |
| Kinetics-I3D | Action classification | Action P(X=x) | Carreira et al. [41] |
| face-classification | Emotion classification | Emotion P(X=x) | Arriaga et al. [7] |
| VisTA | Audio/speech transcript | Text, rate, pitch | Fan et al. [75] |

Table 7.1: List of the current filters implemented in uxSense.

### 7.1.1 Data Model

Due to the computational time costs of predicting semantic features of video data, we anticipate a data model consisting of uploaded video files as input, with server-side computation processing occurring asynchronously over a brief period of time before model output is accessible to the client. However, the design ideal would be in minimizing the latency between video input to model output. For this reason, we have opted to use real-time implementations (e.g., using a real-time emotion classification framework [7]) where doing so would not heavily compromise the accuracy of our output. Table 7.1 provides a cross-reference of the models (or "filters") involved in our present pipeline.

### 7.1.2 Practical Considerations

Gross motor movement is typically a slower process than fine motor movement, and 3D pose prediction and action prediction are the most computationally costly parts of our pipeline; as such, we downsample the frames used in both pose detection and action classification to 1 FPS. To compensate for this, we rely on change-point detection in joint angles: 3D pose coordinates are predicted [185] as an intermediate input, and are used to calculate joint angles [20], which are in turn used to segment the video into parts [20, 155]. to be classified by an action classification network. The predicted labels and probabilities assigned

by the action classification network represent a discrete event stream, pose coordinates are a continuous value stream.

The facial emotion classification network, which relies on finer features of the video subject's face, is calculated concurrently and asynchronously relative to the 3D pose and action prediction over fixed-width intervals using a rolling window. Because of this reliance on finer features, we use 30 FPS video input for emotion classification. However, this does not greatly affect model performance, since we have chosen a real-time model [7] for deriving the emotion labels.

Concurrently, the audio feature of the video input is used to derive transcripts, speech rate, and pitch [75]. The transcript output from this part of the model is the only client-facing output which enters the uxSense interface not as a timeline, but as a transcript table that coincides with video playback and can be used to enable video captions.

Model output from all filters is presented to the user in the form of timelines. These timelines are linked to annotations what may be called a "human-in-the-loop" stage of the pipeline; as the user evaluates the video, their annotations are always made to specific timelines. Following this human-in-the-loop evaluation that constitutes the primary user interaction with uxSense, the final output of the pipeline is generated in the form of report figures with key components of the annotations, which we call "annotlettes," as described in Section 7.2.1.4 and demonstrated in the vignettes in Section 7.4.4. The transcript also appears again at the final stage of the process as part of the micro-document output.

### 7.1.3 Applications

We currently envision two primary applications for our framework:

- **Evaluation:** Understanding user behavior while engaging with complex applications is a key aspect of evaluation for both scientific as well as commercial settings. In

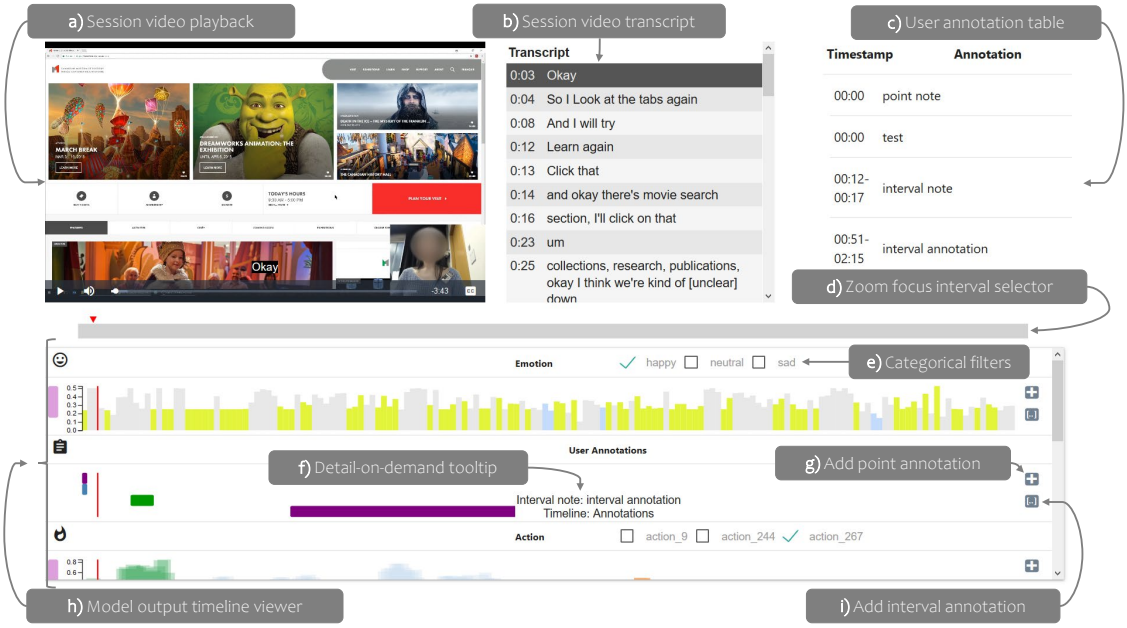



Figure 7.1: Schematic overview of the analysis interface in the uxSense web-based client (view reflects tutorial video). **a)** *Video playback*: View user session video, with or without captions. **b)** *Session transcript*: View timestamped transcript of speech from video, and navigate video by clicking on line of text. **c)** *User annotation Table*: View the text and timestamp of all annotations made by the user. **d)** *Zoom focus*: Select, zoom, and pan whole extent of the video. Red arrow marker indicates current video time, while brushed region shows zoom extent in context of video duration. **e)** *Categorical filters*: When selected, non-selected elements of the view are shown with low opacity. **f)** *Details-on-demand*: Mouseover to get details of observation in model output at given time. **g)** *Point annotation* and **i)** *Interval annotation*: Add an annotation corresponding to given timeline for either the video's current time (**g**) or the brushed interval range (**i**) with (**d**). **h)** *Model output timeline viewer*: The UI element that represents the output of all models is described in Section 7.1.1, as well as the user annotation timeline.

particular, this mechanism can complement the need for manual coding of user actions using video footage. For now, this is our primary interest in the framework, and we propose below a web-based evaluation framework that allows for browsing, comparing, and annotating multiple concurrent data streams extracted from the video footage.

- **Non-traditional interaction:** If the behavior extraction pipeline can be run on live footage in a real-time fashion, it would be plausible to feed back the data streams to the application itself. This would, in turn, allow for letting the application react to

the extracted behavior, such as having the interface respond to the user's gaze, their location, or even their mood.

## 7.2 System Infrastructures

In order to account for a breadth of contextual information, the Wizualization system must exploit computer vision methods in recognizing relevant semantic cues of the user's setting. In order to expand the context in which the analytical user's work process is practical, Wizualization must create situated views that allow the user to not only extend their view of software, literature, their writing, their notes, and other tools and resources in the place they routinely work, but also in the places that they do not. Wizualization should also support the user's workflow using senses beyond vision, such as sound or even smell. There are many requirements for supporting all of these features within a single system; this section will cover existing implementations that we have completed that individually afford these features, with the ultimate goal of synthesizing techniques based on and related to this work within a single system.

### 7.2.1 uxSense: Computer Vision for HCI

uxSense is a client/server system that implements our framework for extracting user behavior from video (Section 7.1) with a computational and storage backend and a web-based interactive frontend. In broad terms, uxSense supports qualitative user experience by using a variety of temporal visualization techniques for continuous (time series) and discrete (event) timelines representing *feature extraction filters* (or just filters)–the products of our models (Table 7.1)–to highlight potential points of interest. Because many video analytics algorithms require significant computational power and processing time to complete, uxSense is based on an asynchronous steering workflow where the user can shut down the interface while the


Figure 7.2: Our evaluation pipeline of the prototype uxSense system for extracting user behavior from video footage using deep learning to support in-depth and advanced analysis of participant performance in user studies. We use uxSense to evaluate user sessions in which professional UX researchers use uxSense to evaluate a sample Tableau user session. **a)** sample user session with commercial visual analytics tool (Tableau Public); **b)** sample user session video and audio streams; **c)** server compute of sample user session data: Video pose-estimate-based temporal segmentation, video emotion and action classification, speech detection and audio signal processing; **d)** evaluation user sessions with professional UX researchers using uxSense interface with sample session video, model output; **e)** UX researcher user session video and audio streams; **f)** server compute of video and audio models using professional UX researcher session data; **g)** authors' evaluation of professional UX researcher user sessions using uxSense.

video is processed on the server. Results are streamed back to the client as it is made available. The tool provides an interactive visual analysis interface for viewing structured data streams, comparing across different participants for the same experiment, and generating reports.

In this section, we describe the various components of the uxSense system, including the overall workflow, feature extraction filters, the analysis interface and its data streams, and the report generation.

#### 7.2.1.1 Overall Workflow

The main uxSense workflow is entirely asynchronous; any step of the following (all conducted using the web-based interface) process can be visited at any point (Figure 7.2):

(1) **Upload** one (or more) video or audio recordings to the server.

(2) **Start** asynchronous computation of the available list of filters.

(3) **Monitor** computation progress in real time, or

(4) **Close down** the interface while computation continues.

(5) **Analyze** the data as it becomes available.

(6) **Generate** reports from data analysis.

All of these steps are associated with a specific project. The backend constantly runs computations while there are still recordings to process and filters that have not been executed. Uploading new footage will schedule new computation for the unprocessed video. Results are streamed dynamically to the analysis interface as it becomes available.

#### 7.2.1.2 Feature Extraction Filters

The basic building block of uxSense is the *feature extraction filter*, an algorithm that is capable of processing sequential video data $v_i$ and generating a corresponding sequence of extracted features $d_i$ (or *data streams*), e.g. $f_{filter} : (v_1, \ldots, v_t) \rightarrow (d_1, \ldots, d_t)$. Video frames consist of both imagery and audio, and extracted features are derived from any of these (or both). For example, a typical feature extraction filter may track the position of a person in the frame over time, the direction of their gaze, and their perceived fatigue. Some features are continuous, such as the user's head direction, whereas others are discrete, such as time intervals when a person is pointing or forming another gesture. In addition to the sequential frame data, filters also maintain summary and aggregate data relevant to the tracked feature, such as the user's cumulative movement, their activity level, individual gestures, etc.

Our prototype uxSense implementation provides an initial library of feature extraction filters (see Table 7.1). Many practical computer vision models track multiple features at once, such as the user's pose and a semantic label overlaid on top of the video playback. However, the uxSense design philosophy is to provide feature extraction filters such that

the user can rearrange, hide, and reveal streams based on what they deem most important, relevant, or revealing for their analysis. This facilitates a more semantically meaningful configuration of which filters to include based on the research question and data being evaluated.

#### 7.2.1.3 Analysis Interface

The uxSense analysis interface provides a mechanism for viewing and comparing feature data streams for one or multiple video recordings in a setup with parallel and time-synchronized *tracks* akin to video editing software. Figure 7.1 shows a schematic overview of this interface. The three main elements of the analysis interface are the footage view, the text view, and the timelines (or tracks). A common *time indicator* on the *track* shows and governs which frame of the current footage is being viewed. The *footage view* shows that frame from the currently selected recording. The *text view* displays timestamped textual information about the video from the transcript and from the user's own annotations, and can be clicked to navigate to different points of interest throughout the video.

Each data stream is visualized in the timeline view depending on its data type (in the order as shown from top to bottom in Figure 7.3):

- **Action Predictions:** A plot of discrete events (using hues) based on action labels assigned to video segments over time, with prediction probability represented with rectangle height. Segments are based on change points detected in human joint angles [20] from 3D position predictions [185] using the E-Divisive procedure [155].
- **Emotion Prediction**: Facial expression emotion classification labels and prediction probabilities [7].
- **Speech Rate**: Speech rate is calculated using the word frequency over fixed time intervals using speech-to-text model output [75].

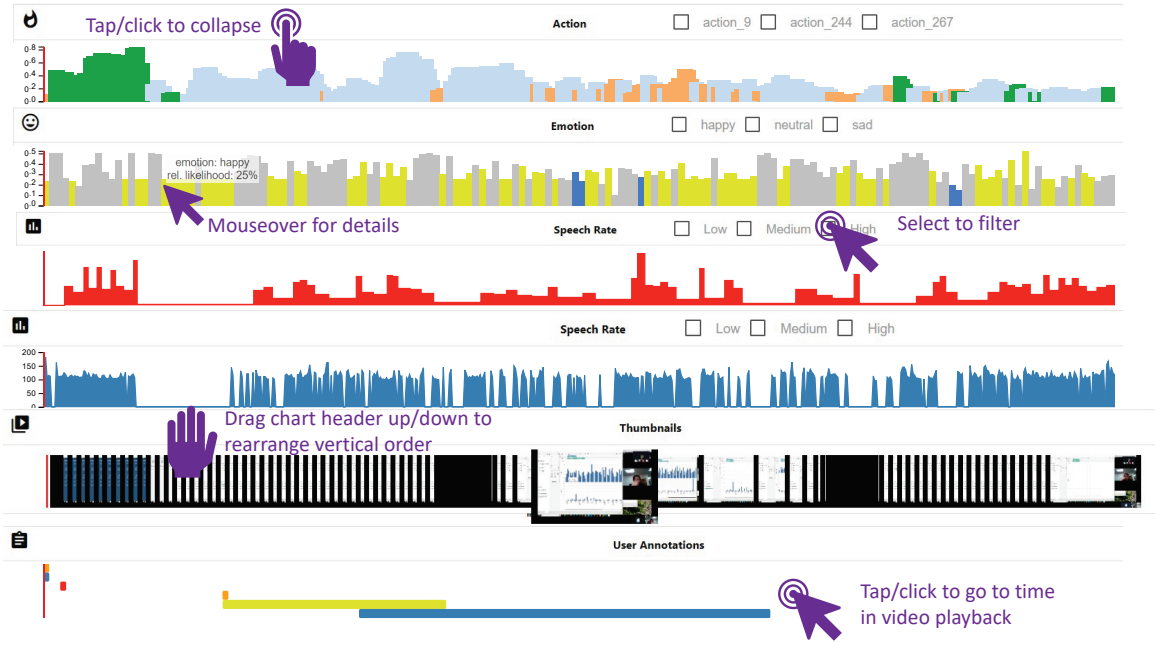



Figure 7.3: Timeline view acts as both a video scrubber to select current time and an interactive representation of the filter output (and the user's own annotations). Mousing over an element shows text details of the timeline at the current frame. Filtering using the checkboxes on the header highlights all observations meeting the filter criteria by making all other observations semi-transparent. Clicking on the timeline navigates the video to the selected timestamp.

- **Pitch**: Audio signal pitch (see Section 7.2.1.5).
- **Thumbnails**: Thumbnails with the moused-over video timestamp frame shown in relief larger than the others, which are dynamically repositioned using Cartesian fisheye distortion.
- **Annotations**: The user's own annotations are visualized as a step function, with step colors signifying the annotation's timeline, and mouseover details showing the annotation and name of the corresponding timeline.

Beyond the time-marker based track view of each feature data stream, the user can zoom in on a point of interest by brushing the focus interval selection bar (Figure 7.4). Once zoomed, the timelines can be dragged left to right to pan through the video and all of the

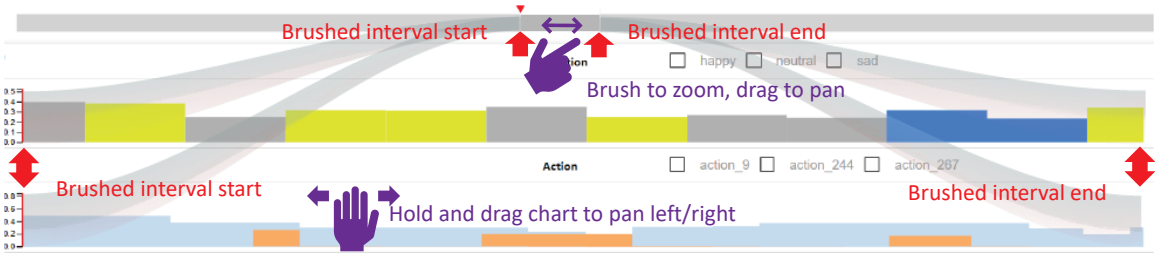



Figure 7.4: The focus-brushing feature selects an interval of the video, zooms all timelines, and allows the user to drag either the timelines or the selected rectangle to pan through the video and model output visualizations.

timelines.

#### 7.2.1.4 Annotlettes: Micro-Report Generation with uxSense

The final destination of uxSense—the concluding step in our workflow (Section 7.2.1.1)—is to generate reports and figures that can be used to report on user behavior in an academic paper, internal memo, or industry whitepaper. The report generation functionality of uxSense creates a small vector graphic document for each individual annotation created by the user that we have named "annotlettes" (Figure 7.5) that link the relevant timeline, transcript, and user annotation for each of the notes the user has created in the analysis interface.

- **Timeline Chunk:** A zoomed view of the annotated timeline (Figure 7.5d) is used to represent the model output or annotation timeline for the selected time period; and
- **Transcript Snippet:** A static version of the transcript (Figure 7.5b) during the selected time period; and
- **Annotation:** The user's annotation text, with the metadata associated with it formatted as a header (Figure 7.5a&c).

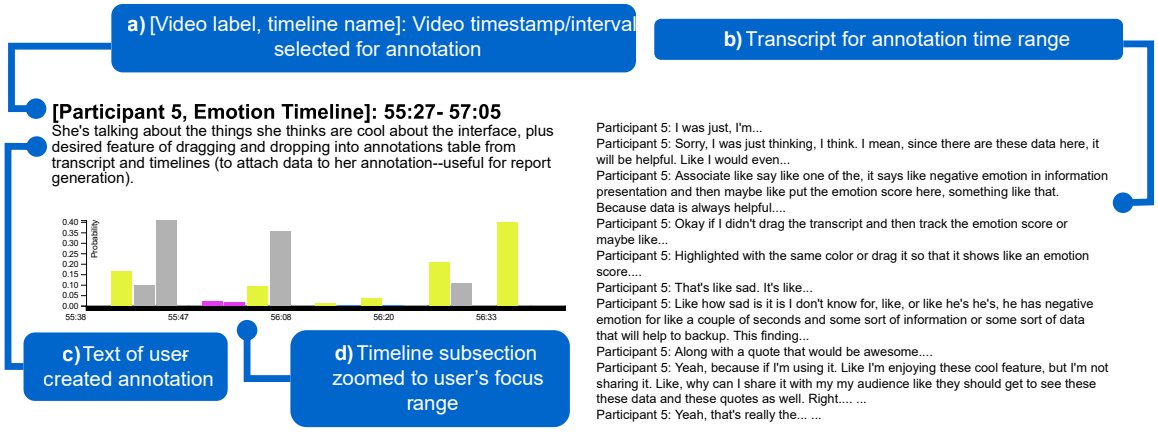



Figure 7.5: An annotated example of an annotlette generated by uxSense during our own evaluation of user sessions in which UX professionals used uxSense. **a)** Metadata regarding the annotation. **b)** Transcript from the period selected for annotation. **c)** The annotation text created by the user. **d)** A zoomed view of the timeline associated with the annotation for the period selected for annotation.

#### 7.2.1.5 Implementation Notes

We implemented uxSense as a Node.js[1] application with several of the server-side components implemented in Python and R that are spawned from the Node.js process. Individual filters used a combination of freely available model implementations; see Table 7.1 for details. For audio analysis, we used Praat[2] to extract audio features; we also use the Google Cloud Speech-to-Text API[3] to transcribe audio to text. The web client was implemented in JavaScript, HTML5, and CSS. We used HTML5 video and canvas for the video playback, and D3.js [33] for the visualization components. The interactive transcription is made by using the videojs-transcript plugin[4].

Since some models provide multiple features in the same computation, we used a server-side caching scheme where different filters that rely on the same model could look up previously computed data instead of rerunning the same analysis from scratch. For example,

[1] `https://nodejs.org/`
[2] `www.praat.org`
[3] `https://cloud.google.com/speech-to-text`
[4] `https://github.com/walsh9/videojs-transcript`

for a head-tracking filter that uses a full-body 3D tracking model to merely extract the user's gaze direction, the recovered full 3D skeleton of the user would be stored in a local file. If another filter was introduced that relied on the same model (e.g., determining the user's position in 3D space), that filter could merely look up the previously stored data instead of rerunning the same model.

Our framework is Open Source and can be accessed on GitHub at [anonymized link]. Furthermore, as we are keen to make this technology available to other researchers working in this area, we also distribute prebuilt software packages on the uxSense website at [anonymized link] to facilitate dissemination.

## 7.3 Expert UX Designer Review of CV for HCI

Our user study involved several professional UX researchers from several large tech companies in the United States. The study evaluation data analysis process was not only a means for evaluating our participants' responses, it was also an opportunity for us to eat our own dog food, so to speak, and use uxSense itself to analyze our users' evaluation of user session footage using uxSense. Figure 7.2 shows the study workflow.

Prior to running actual tasks, we piloted the study with a single HCI master's student who was not familiar with the system. This pilot session enabled us to identify and troubleshoot issues with the software, as well as refine the testing protocol.

### 7.3.1 Participants

All participants were women in possession who held the job title "UX Designer," one of whom occupied a senior UX designer position. Two participants had 5 years of professional experience, two had 2-3 years of professional experience, and one had 0.5 years of professional experience (in addition to relevant graduate school experience). Two participants held

Ph.D.s in Information Science, one held a Ph.D. in HCI, one participant held a Master's degree in Product Innovation, and one participant held a Bachelor's degree in Media and Advertising.

When asked about their typical evaluation process, participants responded with the following descriptions:

*"I normally video/audio record the session, take notes during the session and sometime collect data about the tasks we ask people to do during the session. User sessions are typically semi-structured with some tasks and open feedback. I normally code the data based on themes in spreadsheets and also create video clips that demonstrate the key themes."*

*"There are different UX research methods and they are conducted differently. The most common ones are usability test, interview, survey. In usability test, I mark the success of each task, quantify some of the useful actions, such as user errors, user habits, user preferences and quotes. I write down their pain points and extract themes from there."*

*"[I]nterview, survey, observation, usability testing."*

*"[U]sers' behavioral and perception data, could be task-driven, or self-reported response about their workflow, interview, survey, diary study."*

Participants reported frequently using the following categories of tools for evaluating user data:

- **Video conference platforms**: Google Hangouts, Zoom, and GoToMeeting.
- **Spreadsheet applications**: Microsoft Excel and Google Sheets.
- **Programming languages and extensions** for working with quantitative data and for NLP: R, Python, and internal tools.
- **UX session and survey data collection tools**: Validately, Qualtrics, and QuickTimer.

### 7.3.2 Apparatus

uxSense was hosted on an institutional server running Linux (Ubuntu version 16.04.1) with an Intel Xeon CPU E5-1650 v3 (3.50GHz). Because our system's backend pipeline involves a processing time longer than real-time and depends on GPU support for efficient model output, the model output data accessed by our users was pre-generated prior to the beginning of their session on a laptop running Windows version 10.0.18362 with an Intel Core i7-8750H CPU (2.2GHz) and a Nvidia GeForce GTX 1060 GPU (8GB RAM). Users were able to access the system remotely using their own devices.

User interactions with the system and video playback activities were asynchronously posted to the server. Session video and audio activity was recorded using Zoom's Enterprise cloud service, which also generated transcripts of the user sessions that were evaluated both outside and within uxSense. Video of the users' faces and audio of their voices during the session was recorded using their own systems' webcams and microphones. The model output for the video and data collected during the session was generated using the same laptop that was used to pre-generate data for the user sessions.

### 7.3.3 Tasks and Procedures

Participant activity involved a user session 60-75 minutes long followed by a Google Forms survey that they were asked to fill out at their earliest convenience. At the end of the user session, all participants were paid US$50/hour for their time, with this post-session survey being compensated for in advance as a half-hour's worth of work.

Participants were asked to perform an open-ended exploration of the uxSense interface features in a pre-activity training phase. Once they were done with their exploration, the researcher informed the participant of any features not discovered during exploration, and answered their questions about the system. During the second stage of the user session, the

participants were tasked with identifying problems with the user's experience in an 11.67 minute-long video of a novice Tableau user exploring a dataset they hadn't seen before. They were told that they would need to give a brief summary of their observations at the end of the session. After this activity ended, participants were paid via Venmo and emailed a link to the post-session survey. All participants who completed the session also completed the post-session survey.

Participants were asked to think aloud during all stages of the session and spoke frequently of issues related to their own experience during the session. After the session activities, they were asked about their general thoughts, and they summarized their experience and offered suggestions on how the interface may be modified to improve the user experience. Finally, after the session, they were asked to complete a post-test survey with both Likert scale questions and open-ended text response questions. All participants completed this survey less than 1 week after their session.

## 7.4 Computer Vision for HCI User Study Results

in Batch *et al.* [19], we originally recruited six UX professionals to participate in our study. However, the third session failed due to incompatibility problems with the participant's browser, as well as unexpected latency due to the geographical distance between the server and the client coupled with large file size and poor compression. For this reason, that participant was unable to complete their session. We revised the software based on these experiences and successfully conducted the study with the five participants reported below (numbered 1-6, omitting participant 3).

### 7.4.1 Think-Aloud Transcripts

As reported in Section 7.3, we asked participants to follow a think-aloud protocol during their session. We used Zoom to automatically transcribe participant utterances, analyzed them, and report our findings below.

In general participants felt that taking notes and marking their corresponding time at the same is demanding. However, the multiple timeline interface can relatively reduce some effort. For example, participants used the annotation timeline to help them keep track of and organize the notes that they could refer to during their analysis. Also, they used the emotion timeline to better understand the user's behavior; for example, P4 commented that *"I found that those positive emotions are related to excitement that he experienced when he [the pictured user] found something really interesting."* Although participants found the user's pitch information was indicative of excitement, they felt that the current visualization of pitch information did not help them quickly spot the moments of unusual pitches. Moreover, participants felt that having access to multiple timelines during their first pass of analysis was overwhelming, because there was too much information to attend to and they wanted to focus on the video. Instead, P2 felt the timelines might *"be more helpful in the second round of analysis."* Lastly, participants hoped to be able to configure the layout of different panels, so that they could temporarily hide panels that are non-essential to their analysis at hand. They also requested a synchronized video transcript view with the timelines.

### 7.4.2 User Experience Survey

Expert responses to the Likert-scale user experience survey questions described are shown in Figure 6.12. Participants also gave open-ended responses to survey questions; their responses are reported in Table 7.2. Participants generally found uxSense to be relevant to their typical workflow and allowed them to easily learn about the user's experience,

| Feature | Summary of Participant Responses in Open-Ended Survey Question |
|---|---|
| Video | • Participants universally reported wanting to be able to speed up or slow down video playback speed.<br>• Video was reported as the primary focus for all sessions.<br>• Resizeable video was suggested by all participants. |
| Transcript | • Transcript should be exportable<br>• Two users found transcript to be most important (after video) |
| Annotations | • Three participants recommended that the annotations table be exportable.<br>• Two participants suggested that the annotation table needs stronger visual feedback when updated.<br>• Time constraints inhibited use of the annotation feature; a typical evaluation of 10 minutes of user session video takes >30 minutes.<br>• The annotation button should be a single button. |
| Timelines (All) | • Vertical sort dragging affordance was not intuitively implied by interface.<br>• Difficult to divert attention to timelines on first watchthrough of video. Two participants said that they would look at it more on a second pass of video.<br>• Timeline brushing to focus and zoom almost went overlooked, and panning with a zoomed interval was confusing at first. |
| Timelines (Action & Emotion) | • Emotion and Action timelines were deemed most relevant to the evaluation workflow by three participants.<br>• One participant reported a lack of trust in the action and emotion model output, while two participants reported a sense of trust in the model output. Two participants said they could see clear links between emotion labels assigned by the model and their observations in the video and transcript.<br>• Participants found the arbitrary numbered action IDs we assigned to action classification model output confusing without the semantic association of a type of action. |
| Timelines: Pitch & Speechrate | • One participant found pitch timeline informative when used in conjunction with emotion timeline, but general feedback was that speech rate and pitch were marginally redundant and not relevant to their workflow. |
| Timelines: Annotation & Thumbnails | • Annotation timeline was deemed helpful to workflow once understood, but was described as confusing at first.<br>• Thumbnails were deemed too small to be very useful; one participant said thumbnails would make more sense embedded directly in video playback. |
| General/ Multiple-Feature Feedback | • Two participants would have liked a stronger link between transcript and annotations via interaction for embedding transcript lines in annotations table or vice-versa.<br>• Explicit labeling of interface features and data variable descriptions was suggested by all participants.<br>• Two participants said universal design guidelines could be better applied.<br>• In light of video being focus of sessions, the cognitive load and information overload of viewing many features at once was commented upon by two participants.<br>• One participant said "[adding] a free-form note taking section for researcher to take initial notes and maybe allow them to organize them when do more post analyses." |

Table 7.2: Summary of open-ended survey responses.

sometimes revealing points of interest in the video that they may have otherwise missed. On the other hand, they also found that it sometimes responded in unexpected ways and made them feel frustrated, and that the model output timelines could be difficult to interpret.

Their feedback in response to our open-ended survey questions (Table 7.2) generally suggested design changes that were more in line with universal design guidelines in layout and interaction (e.g., mindfulness of information overload, improved visual feedback to user content creation, and better descriptive labeling of features). Video playback speed control (0.5x speed, 1.5x speed, 2x speed), 10-second skip-forward/skip-back buttons, and video view resizing were the most commonly requested features, as the video was reported to be the central focus for the participants. Most participants also strongly desired a transcript or annotation table export feature, and several participants saw value in visually linking the annotations with the transcripts and/or the selected point or interval on the timelines by embedding two or all three in either the transcript or the annotations table. It was the combination of these final two points in their feedback that motivated us in our design of the post-study implementation of the “annotlettes” feature shown in Sections 7.2.1.4 and 7.4.4.

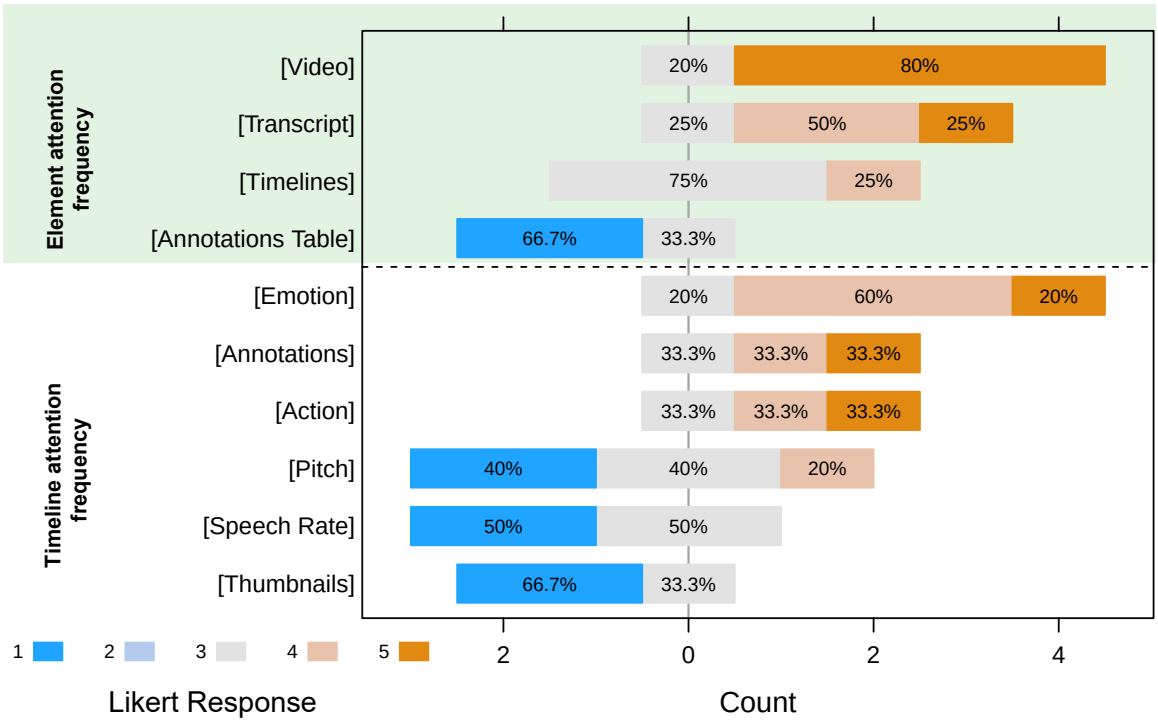



Figure 7.6: Likert responses to time-attention questions on post-session survey.

### 7.4.3 Time Use: Observed and Self-Reported

We analyzed event logs, survey responses, and qualitative feedback from user sessions to create a descriptive flow report of their interactions using uxSense. Given the participants' heavy use of the video playback feature, we studied their navigation of the video in detail (Figure 7.7). All participants spent a similar share of their time in undirected exploration of features of the interface (shown in red) relative to time spent in the UX evaluation stage of the session. However, their navigation of the video shows very different viewing patterns during evaluation. Participants 2 and 4 opted to watch the video the whole way through in a largely linear way before looping back to a small number of points of interest that they spent longer stretches of time on. Participants 1 and 6 start off by skipping around the video in a way that generally progresses forward before looping back and taking a second pass

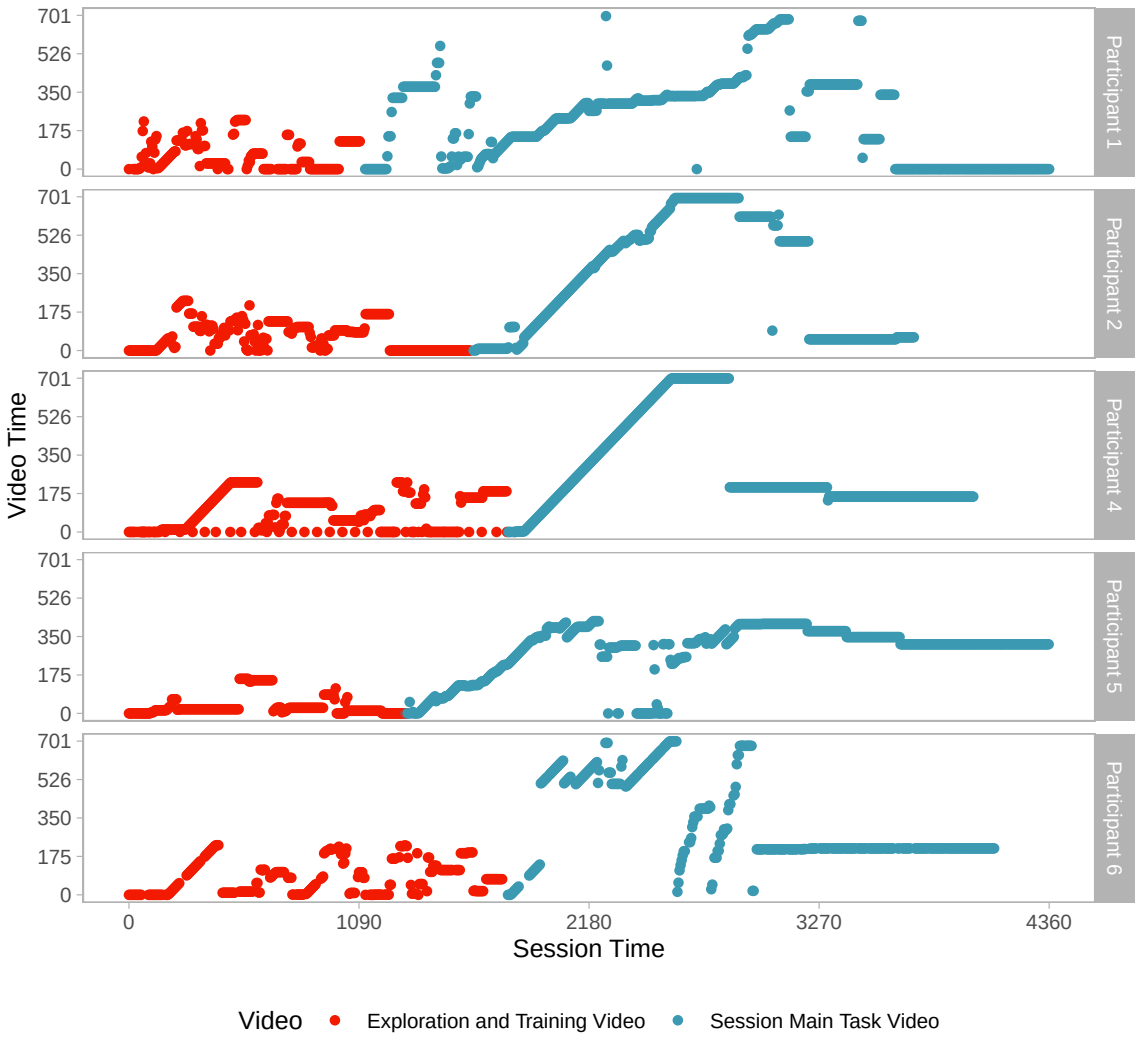



Figure 7.7: Participant video playback activity logs.

that involves frequent short-range backtracking. Participant 5 performs a lot of short-range backtracking on her first pass, then has a period of frequent skipping back and forth between several intervals in the video that she found most informative before spending longer spans of time examining key points in the video.

Participants were asked to assess where their own attention fell most frequently during their user study as part of the post-session survey (Figure 7.6). As we saw in Section 7.4.2, nearly all participants reported spending most of their time examining the video playback, and most reported that the transcript was nearly as central to their evaluation process using uxSense.

In order to add depth to our inferences about the participants' use of time during the sessions, we asked them about their own assessment of how and why they distributed their attention during the session. Participants' self-reported attention time use was recorded on a Likert scale (Figure 7.6).

### 7.4.4 Eating Our Own Dogfood: uxSense in Three Vignettes

Based on feedback gathered during the user sessions, the authors extended uxSense to produce annotlettes (as described in Section 7.2.1.4). After the sessions in which expert UX researchers used uxSense to evaluate a sample user session, we used uxSense to evaluate our sessions with them, and then generated annotlettes with our own annotations. uxSense is a tool intended for more detailed reporting than can be contained in a single section, so we demonstrate this feature three example vignettes from our evaluation of our user sessions.

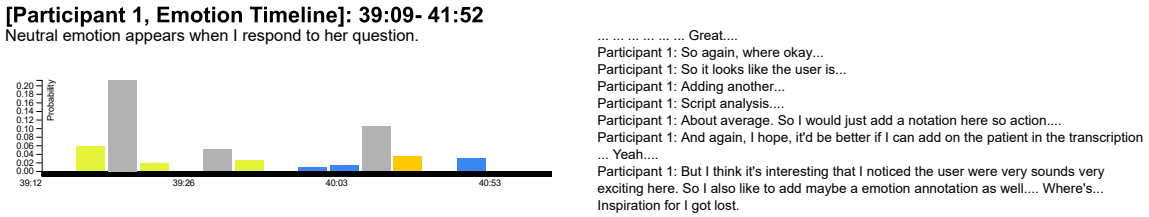



Figure 7.8: Researcher made an annotation with the observation the appearance of a high-confidence prediction of a neutral facial expression when they responded to the participant's question.

WHILE exploring the video of Participant 1's user session in uxSense, the researcher notices that there were a small number of high-probability "neutral" facial expression predictions from the model output that appear as peaks in the emotion timeline. Using uxSense's focus+context interval selection feature, she adds an interval note (Figure 7.8) highlighting what she observes about that annotation. Then, navigating to the next high-probability neutral expression in an entirely different phase of the study, she notices that there is another question-answering dialogue between the researcher and the participant about the system, and adds an annotation for that time interval of video as well (Figure 7.9). After generating uxSense annotlettes, she makes the further observation the high-probability neutral expressions tend to following a low-probability happy expression label. Using this pattern observation, she is able to quickly identify other instances of question-answering during the user session.

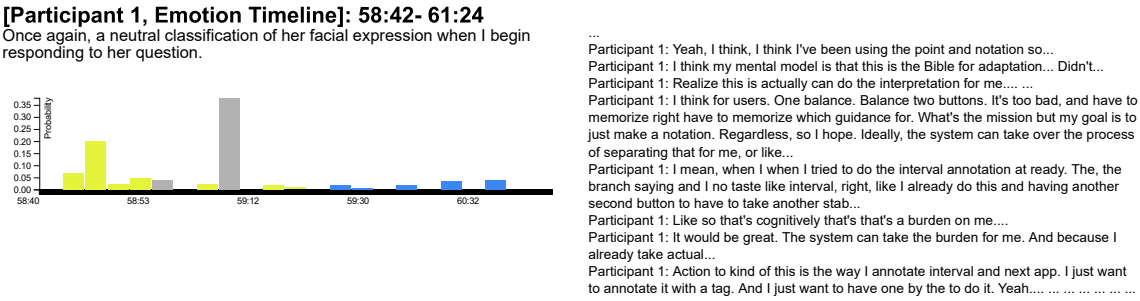



Figure 7.9: The researcher makes another annotation with the observation of an association between researcher question-answering and the appearance of a high-probability neutral expression.

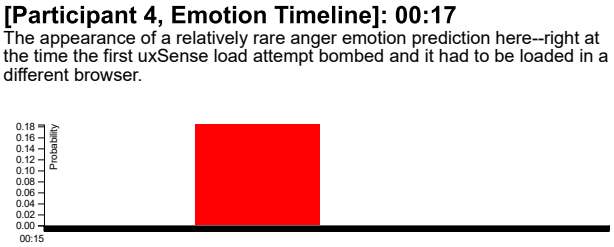



Figure 7.10: The first appearance of an "angry" emotion classification label; here it is associated with the inconvenience of an unanticipated system error that had yet to be worked out. There is no associated transcription because there was no dialogue directly at this single point in time.

AFTER working her way through user session videos until she reaches the fourth participant, the researcher notices the appearance of the rare "angry" emotion label classification twice for this participant's session. She uses the point annotation feature to make a note of what happens in the video associated with the emotion timeline. The bright red emotion indicator appeared when the user ran into a compatibility issue with the browser she was using, and the system failed to load properly (Figure 7.10). The researcher seeks out another appearance of the uncommon emotion, again in the same participant's session; this time, it corresponds to an equally uncommon evaluation

of a user session: Disapproval toward the user's analytical process. She creates another point note describing her observation (Figure 7.11).

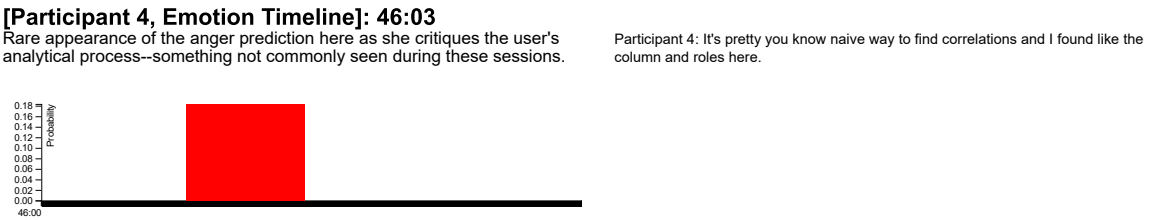



Figure 7.11: Another appearance of the rare "angry" emotion classification label, again within the same participant's timeline. This time it is associated with the equally rare criticism of the user's analytical choice.

MOVING on with her evaluation of the user sessions to the final participant, the researcher sees that the appearance of action_290 is something of an outlier in the actions timeline. Going to that point in the timeline reveals that the participant has said "I wish I could see this more clearly, because right now it's really small" in reference to the video playback viewer. The researcher makes a note of it without looking too closely at the transcript (Figure 7.12). This particular action is uncommon in this participant's action timeline, so she navigates to its next appearance. Once again, she sees that the participant is expressing her frustration with the small size of the video playback viewer, and she makes a note of it (Figure 7.13). Satisfied for the moment, she generates annotlettes for the annotations she has made thus far, and inserts them as figures in her report. Upon inspection of the annotlettes she has created, the researcher is now able to see the semantically-loaded action labels that she tagged during her first pass at evaluating her video dataset. She finds that action_290 is "shaking head;" she amends her annotation with this new information. She also sees upon closer review that the actions timeline picked

out something that she may have missed without it: In the interval selected in the first annotation (Figure 7.12), the participant speaks quietly enough that the transcript failed to capture her complaint about the video playback viewer size.

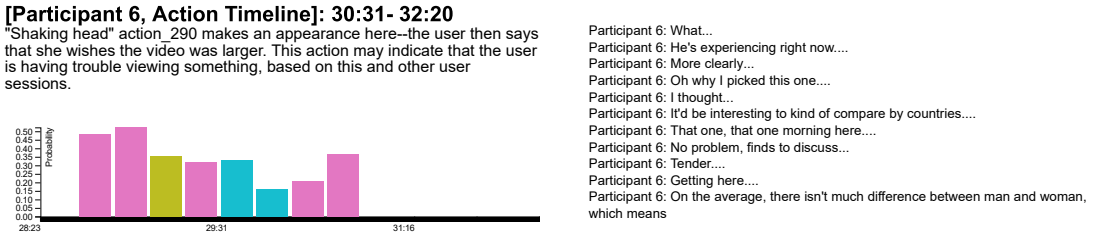



Figure 7.12: The researcher observes the appearance of action_290 during a moment in the session in which the participant is expressing frustration at being unable to resize the video playback viewer, but the transcript has not captured her sentiments.

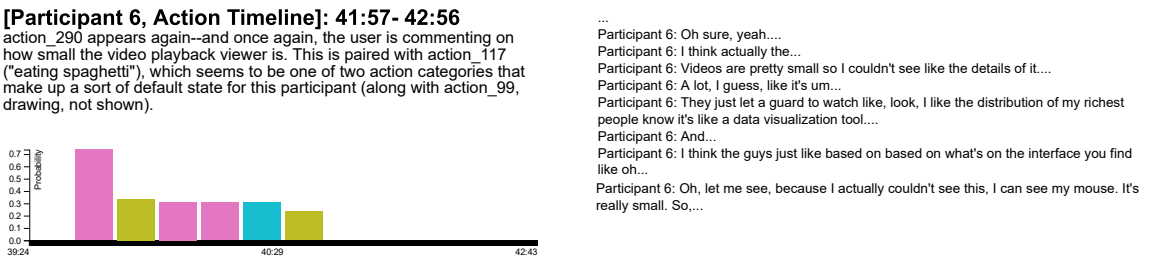



Figure 7.13: The researcher observes another appearance of action_290 capturing the participant's complaint about video size, the same design concern featured in Figure 7.12.

## 7.5 A Discussion of the Vision uxSense Represents and our Evaluation Findings

In Batch *et al.* [19], we proposed a vision and a tool for extracting features of human behavior from video and audio footage using ML tecniques to support UX and usability

professionals in their analysis of user session data using interactive visualization. This vision is bolstered by a literature review from the domains of pattern recognition, computer vision, and machine intelligence, which all point to the feasibility of this vision. Our prototype system—UXSENSE—validated the vision by providing a web-based interface to a computational backend that asynchronously runs a range of *filters* to extract different types of data from one or several time-synced video streams uploaded by the user as *data streams*. Filters in the uxSense system are designed as plugins that can be added or removed at run-time, each responsible for extracting one or more types of data, such as spoken language, gaze direction, position in a 3D space, arm gestures, orientation, hand postures, and even facial expressions. Data streams recovered from each filter is shown in the visual interface as parallel time tracks, similar to the track-centric approach with a horizontal timeline used by video editing software. Finally, the tool provides robust annotation features where the user can select events as well as intervals across the timelines and add notes for future analysis, collaborative iteration and review, and report generation.

To prove the feasibility of our vision, we present results from an expert review involving five professional UX designers working at several large tech companies on the U.S. West Coast. We asked these participants to use uxSense to analyze a think-aloud usability session performed on the Tableau [228] visualization environment. In order to further showcase the utility of uxSense, we used the system itself to evaluate data streams recorded during these expert review sessions. Our findings show positive promise for our vision of ML to facilitate usability and UX testing, as well as for our uxSense prototype implementation. In particular, while one of our participants wanted to know more about the accuracy of the models before she would trust it, the general sentiment among them was that the idea use of such preprocessing filters could significantly ease their own daily work processes, or at least help them identify relevant points of interest. While we did not evaluate this functionality in our expert review, our implementation included several utilities that can be used to label

actions that a person performs, their gaze direction, and their 3D posture.

# Part IV

# Part 4: Synthesis, Limitations, and Research Vision

# Chapter 8: Implementation: Wizualization, Optomancy, Weave, and Spellbook

"Any sufficiently advanced technology is indistinguishable from magic."
– Arthur C. Clarke, *Profiles of the Future* (1973 revision)

In 2007, prolific fantasy author Brandon Sanderson introduced the concept of "hard magic"—magic systems that have rules explicitly described by the author—as well as *Sanderson's First Law of Magics: "An author's ability to solve conflict with magic is* ***directly proportional*** *to how well the reader understands said magic."*[1] Analogously, a visualization system's ability to support a user in mapping their data to a visual encoding is directly proportional to how well the user understands the system's grammar of graphics—which steps that must be taken to create what visual marks and how to modify its visual channels. Might a grammar of magic lend itself to a more intuitive adoption of a visualization system? In this paper, building on the idea of superpowers as inspiration for data visualization [255], I propose an approach to visualization authoring based on such a hard magic system.

Modern interface technologies, such as virtual, augmented and mixed realities (VR/MR/AR, collectively called e*X*tended Reality, XR), offer many opportunities for depicting data that is specific, timely, and suitable for the user's context, location, and task at hand [198]. The domains of immersive and situated analytics (IA/SA) [154, 233] explore the use of such immersive technologies in analytical settings, often encapsulating broader research on

[1] Sanderson's First Law, Brandon Sanderson (2007).

post-WIMP (windows, icons, menus, points) analytical environments, away from traditional desktop settings [198]. Notions such as ubiquitous analytics [66], embedded [256], and tangible [50], and multisensory visualization [21] represent novel ways to interact with data using spatial 3D interfaces [28], cross-device synergies [84], and multiple sensory modalities [156].

What if visualization authoring in XR could be cast as the invocation of spells through a combination of mid-air gestures, voice commands, and data bindings? After all, these input modalities bear striking similarity to the classic notion of magic: conjuring up new objects—colorful visualizations responding to touch, voice, and gaze—out of thin air. Towards this vision, I take a page from the book of Garcia and Ginat, who propose to "demystify computing with magic" by teaching computer science as practical magic tricks described in a step-by-step manner [85]. Similarly, performing magic invocations can be an engaging and useful mental model for users to author visualizations for situated analytics. These ideas are also compatible with recent movements within immersive and situated visualization [248] toward modeless grammars of graphics for immersive analytics (IA) [52], such as virtual hand and raycaster interactions for manipulating data in IA systems [242]. Indeed, using the whimsical metaphor of magic systems for an immersive grammar of graphics seems such a natural design choice that one book on magic system design in games includes a chapter on "magical grammar" [103].

I call this overall idea of applying a magic metaphor to visualization authoring mixed reality WIZUALIZATION, and propose the following specific contributions in this chapter:

- **Optomancy**: A grammar of graphics (GoG) of immersive visualization and interactions—the first GoG, to the best of my knowledge, that is structured to include multimodal user interactions—for situated analytics environments;
- **Spellbook**: An augmented reality code notebook and dictionary of mid-air gestures

and spoken commands (spells);

- **Arcane Focuses** and **Weave**: Arcane focuses facilitate the verbal and touchscreen system control interfaces (e.g., smartphones), while Weave propagates the signals between them and the HMD and gesture control interface; and

- **Wizualization**: A rendering system tying together Optomancy, Spellbook, Arcane Focuses, and Weave for enabling speech and gesture-driven data visualization and analysis on the current generation of mixed reality hardware.

I have made all of the subsystems that constitute my contribution to eponymous repositories under the GitHub organization page at `https://github.com/Wizualization`.

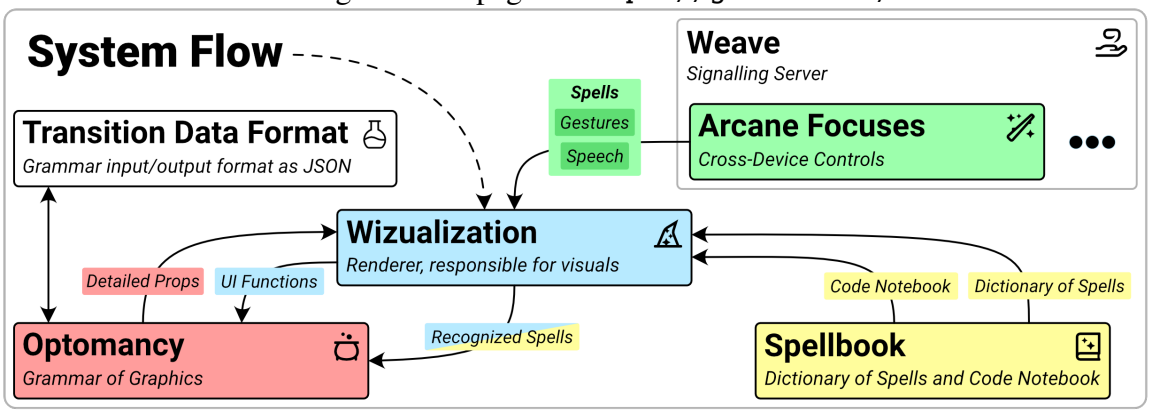



Figure 8.1: **System overview.** The main system components of Wizualization, including their connections and data flows.

## 8.1 Design of the Wizualization ( ) Rendering System

Consider: What are the core elements of a magic system?

In its most rudimentary definition, magic represents the transformation of the spellcaster's will into action through mechanics that are not immediately apparent to the onlooker. The Arthur C. Clarke quote above reflects the role of perception in defining what does and does not constitute magic. In this sense, magic represents abstraction: Once the viewer,

reader, or user has an understanding of the hidden details of a "magic" system, it is no longer magic. We imagine the spell-caster as having an understanding of the language and application of the magic system that they are exploiting—for without it, how could they cast their spells?—but we do not assume that they understand the source of their powers to its core, nor the processes through which their use of said system turns their desires into realities.

As with my abstraction analogy, literary analysis of magic typically deals more with magic as a metaphor and less with the mechanics of the system in the works of fiction being evaluated. Magic in games, however, deals more directly with these details—out of necessity, as the rules of the game must be explicitly written so they can acted upon.

Wizualization is the overall name for my magic-inspired visualization authoring environment for mixed reality. In the context of this system, I will be using the following terms throughout the paper:

- A **spell** is a recorded sequence of user inputs—in my implementation, this includes sequences of words, sequences of fingertip position data, and QR tags—that have been assigned a unique ID by Weave and have been associated with a set of transformations defined within the Optomancy grammar.

- **Casting** a spell is when the user evokes a method or series of methods—analogous to a recorded macro—that are associated with a stored input sequence that is the closest match to the action they have just performed.

- **Crafting** a spell is when the user performs an action that will be tracked and stored as an input sequence to be compared to future spell-casting events. The user first specifies the input sequence to be used to evoke the spell, and then records the methods that the spell will call when evoked.

### 8.1.1 System Overview and Specifications

Wizualization is the client application layer for the implementations I will describe in this paper. It written in typescript and was built in React[2] and depends on react-three-fiber[3], a React wrapper over three.js.[4] It also depends on the WebXR API[5] for hand pose, the Web Speech API[6] for speech recognition, and the AR.js library for object recognition.[7] Wizualization translates and renders the specifications generated by our grammar, Optomancy, which it imports directly as a Node.js package; it also imports my code notebook, Spellbook, directly as a Node.js package and component library, and renders the UI components it returns. It communicates across devices using Weave, my signaling server.

### 8.1.2 Cross-Virtuality: Arcane Focuses (🪄) and Weave (🕸)

In many fantasy settings, magic objects of two common types often appear: (1) *enchanted items* or artifacts that can do specific things regardless of the ability of the user; and (2) *arcane tools* or "focuses," such as staves and wands, that allow for amplification or a more detailed control of magic and are dependent on the ability of the user. J.R.R. Tolkien, whose "soft magic" systems are often used as a counterexample to Brandon Sanderson's "hard magic," demonstrates possibly the best-known example of an enchanted item with his One Ring, although his writing about Middle Earth are full of examples of enchanted items (and very few examples of Arcane Focuses, with a notable exception in the staves of the *Istari*—wizards, such as Gandalf and Saruman). On the other hand, two examples of Arcane Focuses in fiction include The Doctor's sonic screwdriver in *Doctor Who*, and the bells used

[2] `https://reactjs.org`
[3] `https://docs.pmnd.rs/react-three-fiber`
[4] `https://threejs.org`
[5] `https://immersive-web.github.io/webxr`
[6] `https://wicg.github.io/speech-api`
[7] `https://ar-js-org.github.io/AR.js-Docs`

by the *Abhorsen* to lay the undead to rest in Garth Nix's *Old Kingdom* series; both examples depend on the ability of the wielder in order to be effective, but support the wielder doing things that they could not do without the tool. The *sa'angreal* in Robert Jordan's *Wheel of Time* series serve a similar purpose, allowing the wielder to safely draw more of the One Power to amplify their abilities. While these are less commonplace examples of Arcane Focuses than, say, a wand, so is my own; Arcane Focuses are often wielded in the hand or hands, and as such, for my own implementation of magic, I have opted to designate the user's mobile device as an Arcane Focus.

Our Node.js signaling server, Weave, acts as a relay linking all devices—or Arcane Focuses—inhabiting a given room via socket.io[8] connections. The signaling server can be layered over a database of existing spells and Optomancy specifications, and in the current version of my implementations, its two primary roles are:

1. **Device linking:** Wizualization runs on client devices and passes spoken commands and gestures, as well as recognized spell IDs, to Weave, which are used to connect all devices in the same room (i.e., connected via endpoints with room IDs); and

2. **Workspace management:** In addition to the rooming system connecting devices, it is also the mechanism by which workspaces are organized. Stored lists of spells, sequences of interactions, and Optomancy specification files are linked to a given room, which determines the initial layout and are updated based on the user's interactions. Each spell is assigned a unique ID by Weave.

### 8.1.3 Indirect User Input Interpretation

In constructing our grammar, we considered common elements of magic systems: In the tabletop role-playing game *Dungeons & Dragons* (D&D), for example, spells all have some

[8] `https://socket.io`

permutation of verbal, somatic (i.e., gestural), or material (e.g., a pearl costing 100 gold, a feather, or a piece of cured leather) components. In fact, these three components of casting magic spells often appear in fantasy fiction, myths, and lore. Depictions of spellcasting often involve the waving of hands, the recitation of vocal commands or incantations, and the use of mysterious powders, odd bits and bobs, and dried pieces of beasts (mythological or otherwise).

Wizualization uses existing web APIs for the detection of both the words spoken by the user and the user's hand position and rotation, and then applies deterministic algorithms to select the gesture, word, or combination of the two in order to identify the closest match to the spell that the user is attempting to evoke. The JSON configuration for the grammar, to be further detailed in Section 8.2, is updated dynamically as the user interacts with the system.

#### 8.1.3.1 Verbal Components (Spoken Commands)

Wizualization uses the Web Speech API to recognize the user's spoken commands. For the purpose of recognizing which spell the user is attempting to cast, I use the Levenshtein edit distance: The existing spell with the fewest transformations—character insertion, substitution, or deletion—needed to turn the string of characters representing that spell's spoken words into the one that the user is attempting to cast. Because the Microsoft HoloLens 2 did not support the use of the Web Speech API in the Edge browser at the time of my system design, I opted to offload verbal input to the user's handheld Arcane Focus. The Arcane Focus transmits spells casted or crafted by the user to Weave, which then relays it to all other devices connected to the room.

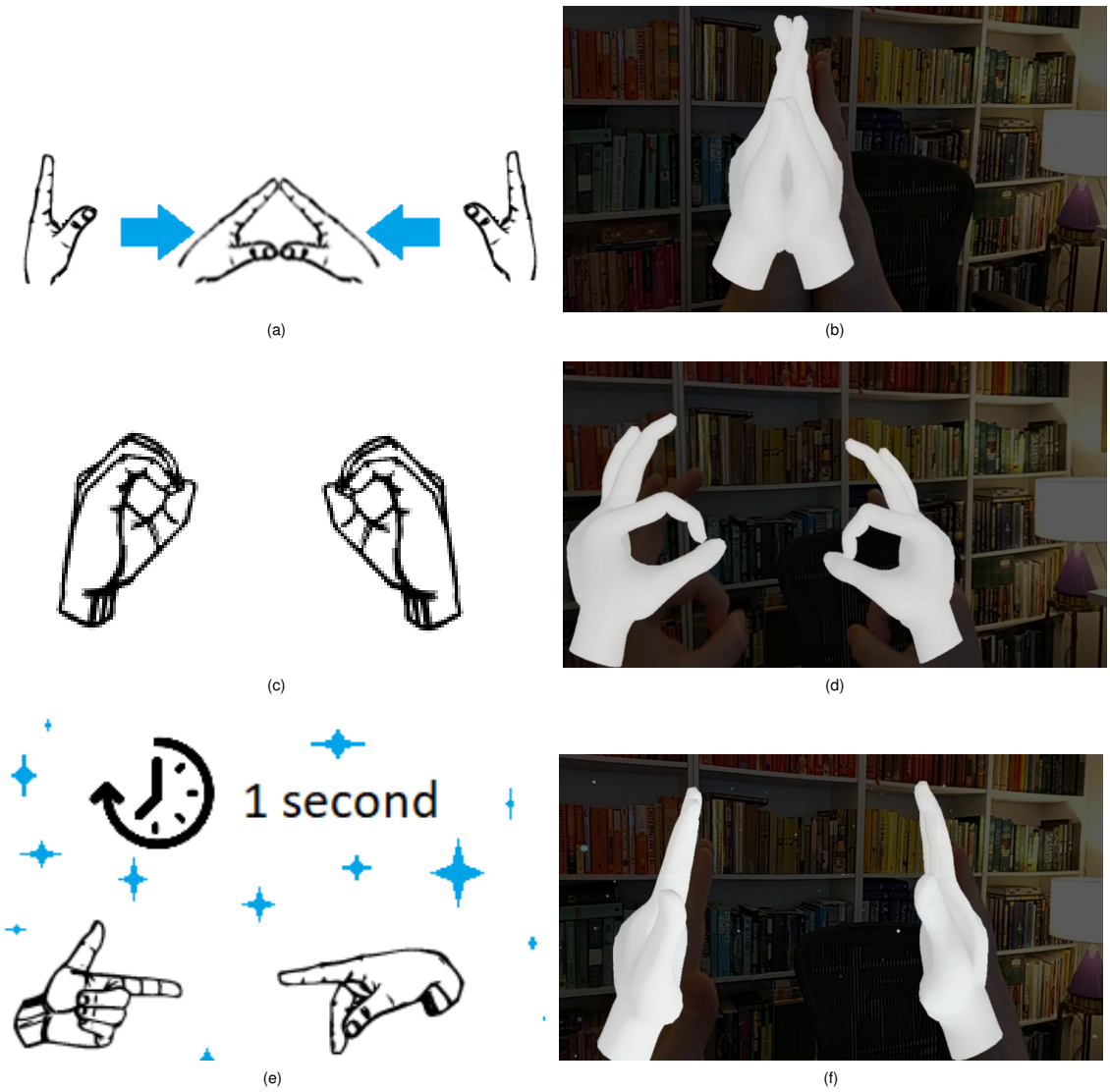



Figure 8.2: **Gesture recognition.** The clapping gesture [(a) and(a)]—or any gesture that involves touching all opposite-hand fingertips recognized in the view—initializes a spell-casting period, while the double-pinch gesture [(c) and(d)] initializes a gesture recording period that will allow for macro creation. When the gesture period begins after a 1-second delay, the user is presented with a sparkling aura around their hands. Attribution for hand illustrations in (a), (c), and (e): American Sign Language alphabet, laid out by Darren Stone, derived from the Gallaudet-TT font. Attribution for clock icon used in (e): Time by Wayne Middleton from NounProject.com

#### 8.1.3.2 Somatic Components (Gestures)

Howard gives fifteen examples of games from 2001 through 2012 using gestural commands (beginning and ending with Peter Molyneux: Starting with *Black & White* and finishing with *Fable: The Journey*) [103, p. 78], and has argued that words and gestures have historically been the preferred input modes for spell-casting in games.[9]

The mid-air gestures are represented as JSON objects with unique IDs assigned by Weave; the sequence of fingertip positions recognized by the WebXR device API is stored as an array of objects every tenth frame, which I found to be sufficient for my purposes. In conjunction with the verbal component, a spell sent to and returned from Weave takes the following structure, with fingertip positions between the first position of the first fingertip (the left hand thumb) at the start of recording and the last position of the last fingertip (the right hand pinky) at the end of recording omitted for brevity (example in Listing 8.1).

Listing 8.1: **Gesture data.** Example of the Wizualization gesture data storage format.

```
{
    "key": "a07ff089-2ca2-1341-cc58-74509f1d8577",
    "gesture": [
      {
        "thumb0": {
          "x": 0.23995332419872284,
          "y": 1.5958237648010254,
          "z": -0.08416889607906342
        },
        [...other fingertips for first frame] },
        [...remaining frames],
        { [...other fingertips for last frame]
        "pinky1": {
          "x": 0.40444329380989075,
          "y": 1.5946118831634521,
          "z": 0.18247440457344055
        }
      }
    ],
    "words": "line"
}
```

[9] Magick Systems in Theory and Practice, Installment 2: Word and Gesture as Input Methods in Gaming History, *Jeff Howard* (2010)

To interpret the user's gestures, Wizualization uses an implementation of the Dynamic Time Warping (DTW) algorithm [165]. Specifically, I use time series of forward vectors from the user's headset to all fingertips recognized by the WebXR API via the navigator.xr canvas context[10] as inputs to the algorithm, and selects the gesture with the minimum distance value output by the algorithm.

#### 8.1.3.3 Material Components ("Enchanted Items")

So what of material components? Games will very often limit or eliminate the use of material components in their magic systems. While material components (as defined in *D&D*) may not play as significant a role in game magic systems, that does not mean that there is no place for magical materials in the form of usable or consumable objects. Because material components do play such a limited role relative to words and gestures in many contemporary interactive environments (namely, games [103]) that involve magic systems and spell-casting, I define Optomancy's "material components" as effectively being closer to the concept of enchanted items as defined in Section 8.1.2. As an homage to *D&D*, I opt to refer to these spell components as material components, despite their closer resemblance to enchanted items.

Optomancy supports the use of material components by treating virtual and real objects as triggers for UI methods that support situating visualizations in the real space around the user. Material components can trigger events in the same way that gestures or verbal commands might, but can, optionally, be used to position and orient the results of the method they trigger. Material components exist partly outside of the loop shared by the other two types of components. There is no back-and-forth communication between the devices in the room via Weave for material components during sessions: Their lifecycle exists within

[10] See `https://immersive-web.github.io/webxr/explainer.html` for an explainer on the API.

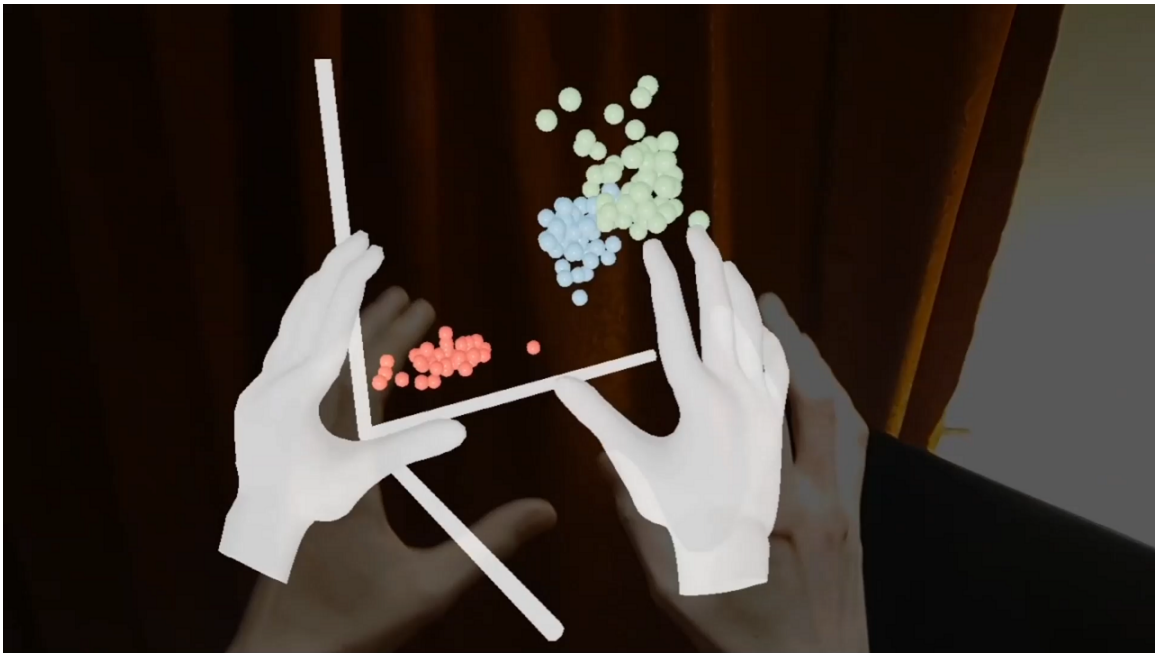


Figure 8.3: **Visualization authoring.** 3D scatterplot generated by the Wizualization renderer using an Optomancy specification file.

the HMD and the grammar, with the exception of object recognition features and 3D object models can be stored via Weave and loaded by new HMD connections.

I opted to keep this relatively simple for the time being, with an eye toward future extension both by users and in later iterations of the Wizualization rendering system. An enchanted item specification with a "barcodeValue" parameter for use with the AR.js three.js extension class, THREEx.ArMarkerControls[11], can be bound to a combination of variables to render a figure. This barcodeValue is sent to Weave as a parameter of a JSON object that is treated differently from the verbal and somatic components, since it has a natural spatial mapping that can be used to situate virtual objects in the scene.

[11] `https://ar-js-org.github.io/AR.js-Docs/marker-based`

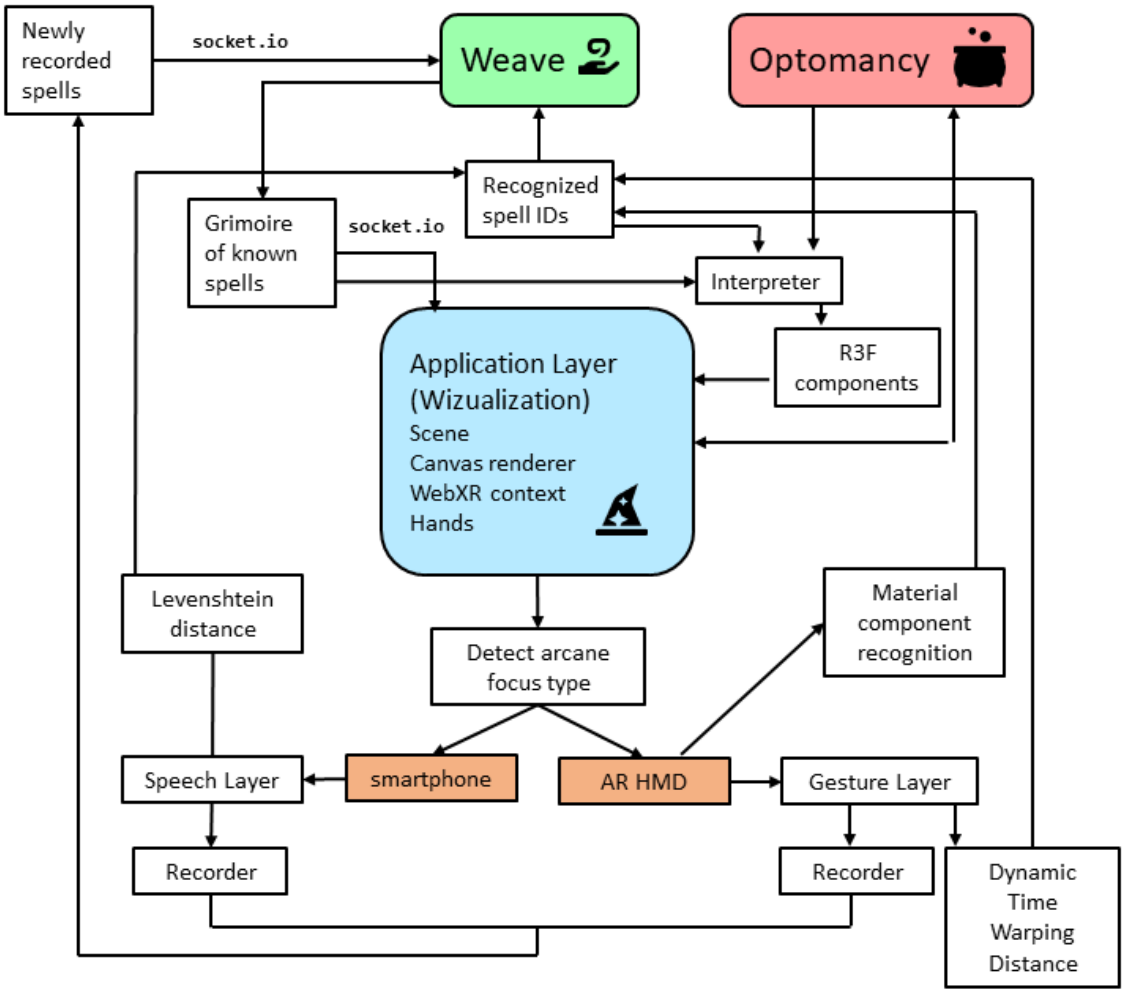



Figure 8.4: Wizualization renderer system overview

## 8.2 Optomancy: The Grammar of Wizualization

"At its heart, a magic system is a language." – Jeff Howard [103, p. 21]

At the core of my visualization system is Optomancy, a grammar of graphics for immersive and situated analytics. It enables a user to convert spells crafted with speech and gestures into visualizations in XR. Spells are the central input mechanism to author visualizations with Wizualization. Each time a spell is cast, Wizualization makes a call

to Optomancy's API, using an approach not too dissimilar to that of other visualization grammars including ggplot2 and Vega-Lite's JavaScript API. The difference being that Wizualization is responsible for making those calls, and not the user.

Just like the other modules of Wizualization, Optomancy was designed as a plug-in module, existing separately from the rest of the codebase. This modular design means that Optomancy is system agnostic, and could potentially integrate with different renderers and vizualization authoring systems. In fact, such systems would not necessarily need to use gestures or speech as inputs.

Wizualization imports Optomancy directly as a Node.js package, which exposes the (`Optomancy`) class. When a user is beginning a new Wizualization session, the only required upfront setup is to supply a tabular dataset in either CSV, TSV, JSON or text format, which is passed through to `Optomancy`'s constructor.

The `optomancy` class uses the data selected for the workspace by the user and a *transition format* (see Section 8.2.2) that is iteratively updated based on the user's interactions with the system. A layering method is then applied: `handleSpell`, which will compute domains, ranges, and scales for all encoding channels; this method is rerun to add new visualization layers on each new cast input. The the transition format specification for each visualization object is generated alongside a rich definition format which exports a visualization title, views, mark, encoding, and scales which are subsequently passed to the renderer.

Optomancy abstracts new layer additions to visualizations in the user's workspace to cast events via the (`cast`) method that takes a single spell ID as an input, as well as an optional list of operands. Upon a user spell cast event, this method takes the spell ID, adds a new layer to a persistent specification object, returns a snapshot of the transition file format, and a rich technical description of the visualization state which is exported to the renderer.

Figure 8.4 presents a detailed overview of how spells are recorded, recognised, and passed to the grammar and renderer.

### 8.2.1 Interactions and Spell Chaining: Macro Recording as Spellcrafting

Another consideration for our grammar was just how the user might use the grammar to iteratively simplify their own workflows. Even without this iteration, IA systems have been praised by users as accelerating the creation of graphical representations of data relative to their typical workflow [17].

Suppose, for example, that a user finds themselves repeating the same sequence of tasks to create a visualization:

1. Select two quantitative variables and join them into a 2D scatterplot;
2. Change the mark type from point to line;
3. Add a third variable and create a 3D scatterplot;
4. Clone the 3D scatterplot.

What if the user could reduce this into a single gesture or spoken word? Once a spell has been crafted by the user in Wizualization, it can not only be cast at will, but can also be folded into future user-crafted spells. The iterative layering nature of Optomancy supports the reduction of analytical task components into increasingly abstracted method chains. Thus, a sequence like the above can indeed be reduced to a single phrase or wave of the hand.

### 8.2.2 Grammar Transition Data Format and Cast List

Optomancy produces a specification file that can be interpreted by the Wizualization for rendering. The specification file can also be modified directly by the user and submitted as an input to both the Optomancy grammar and the Wizualization rendering system.

Wizualization sessions can be stored and resumed at a later point in one of two ways. The first is via Optomancy's transition format. The second is via a list of cast spells. Wizualization's transition format is an intermediate grammar definition file format described in a declarative JSON format, similar in many ways to the grammar specification formats of VRIA, DXR and Vega-Lite. An example of this format is shown in Listing 8.2. Wizualization also exports a list of cast spells, allowing a user to produce a visualization entirely from their spell history.

Listing 8.2 shows an example grammar transition format specification for a 3D scatterplot.

Listing 8.2: **3D scatterplot.** Optomancy grammar transition format specification of a 3D scatterplot for the classic `iris` dataset.

```
{
  "title": "The Iris Flower Dataset",
  "data": { "url": "../assets/iris.json" },
  "mark": "point",
  "encoding": {
    "x": {
      "field": "petalWidth",
    },
    "y": {
      "field": "petalLength",
    },
    "z": {
      "field": "sepalWidth",
    },
    "color": {
      "field": "species",
    }
  }
}
```

## 8.3 Spellbook: A Mixed Reality Code Notebook

Spellbook—hosted on `optomancy.com`—acts as both a proof of concept and a demonstration of Wizualization and Optomancy. Using Spellbook, an analyst can use a collection

of predefined gestures or sequences of spoken words to create visualizations of their data. They may also choose to craft their own spells—i.e., define their own methods, as one might record a macro—using gestures or spoken words.

### 8.3.1 Compendium of Primitives

Wizualization and Optomancy allow the user to record and translate gestures, speech, and materials into config files for custom workflows. I have pre-recorded and loaded a collection of the following gestural and spoken primitives into Spellbook. For my gestural primitives, I use conceptually-related words from American Sign Language (ASL). For my spoken primitives, I use the English corollaries to the ASL word.

The Spellbook primitives are as follows:

- **Workspace Creation**, triggered by the ASL word for "workshop", the workspace creation command initializes a new room.

  Weave's rooming system differentiates between workspaces, so the user can return to previous workspaces and Wizualization will re-render them.

- **View Type Selection**, triggered by the ASL word for "type" (i.e., "kind"), acts as a shortcut for swapping linked variables' view types (e.g., from SPLOM to 3D scatterplots or parallel coordinate plots) if the user does not wish to directly manipulate the view to achieve the same effect.

- **Object Grouping**, triggered by the ASL word for "group", initializes a free selection sequence that allows the user to select multiple objects in the scene for shared application of grammar translations.

- **Mark Type Selection** modifies the type of mark used for a selected visualization or group of visualizations with one of the following options:

- *Points* (Figure 8.5(a)-(b)) selected using the ASL word for "dot" (i.e., a small, round mark);
- *Columns* (Figure 8.5(c)-(d)) selected using the ASL word for "columns" (as in the architectural building structures);
- *Bars* (Figure 8.5(e)-(f)) selected using the ASL word for "bar" (i.e., "rod"); and
- *Lines* (Figure 8.5(g)-(h)) selected using the ASL word for "line".

For the sake of brevity, I will limit the ASL words represented via Figure 8.5 within this section to the ones used to select marks.

### 8.3.2 Linked Blocks

While traditional code notebooks, such as Jupyter Notebook, typically present blocks of code in a scrollable, top-down page, Spellbook presents the user with floating windows, each of which are visually linked via a curve with the block that precedes it. The Spellbook blocks—or "pages", as I have taken to calling them—represent the workflow of the user. In the present version of Spellbook, each page represents either a gesture or spoken command of the user, or an element of the Optomancy config generated by the user's interactions. Pages can be dragged and arranged anywhere in the space surrounding the user (Figure 8.6(b)).

The dashed curve linking each page of Spellbook (Figure 8.6(b)) is animated such that line segments flow from each earlier step of the user's work steps into the subsequent step. As such, Spellbook represents the workspace's analysis history. While considerable prior work has been conducted on the structuring and visualization of user workflows—for example, work by Maguire et al. [150], who implement and evaluate the Java tool *AutoMacron* designed to do exactly that—I opt to limit the scope of Spellbook by restricting it to being a lightweight component library with linear representation of interactions afforded by Wizualization and the Optomancy grammar.

## 8.4 Postmortem Discussion of Wizualization

My approach in Chapter 8 was built on the fantastical and somewhat whimsical concept of magic, and it is certainly worth asking how far this metaphor should be taken and whether I have taken it too far. After all, many applications of data visualization and analytics are critical, serious, and a long shot from the whimsical, and it could be argued that casting spells and conjuring charts lacks the appropriate gravitas for such applications. I even suspect that some readers of these words may frown at the apparent disrespect implied by applying such fantastical concepts to a respectable, rational, and reductivist scientific field such as data visualization. To such critics I hasten to note that no such disrespect is intended. Magic is now mainstream. The fantastical has shed its disreputable and lurid character of the 1940s and 50s. Brandon Sanderson, who coined the term "hard magic" that I introduce at the beginning of Chapter 8, is a bestselling mainstream author whose books have sold in 21 million copies; 150 million copies of the *The Lord of the Rings* has been sold, making it the third most bestselling book of fiction of all time. Regardless of whether you stan superheroes or not, the fact that the Marvel Cinematic Universe and its ilk currently dominates cinema is without question.

Rather, I base our rationale on the fact that metaphors have power, both in terms of knowledge transfer and familiarity, but also for engaging and motivating the user. As a case in point, the desktop metaphor for personal computing—with icons, windows, files, folders, and trashcans—has persisted and empowered users ever since Engelbart's "Mother of All Demos" in 1968, and Kay's Xerox Alto personal computer from 1970. In addition, the supernatural and the fantastical—that which is beyond the visible and observable—are extraordinarily fascinating to humans, and serve as perfect catalysts for immersive analytics. This notion even has support in visualization literature given the recent Best Paper at IEEE InfoVis 2021, where Willet et al. [255] introduce the idea of superpowers as inspiration and

motivation for specific data visualization features and representations.

A more valid question would be to ask what limitations may be introduced by basing our work on a magic metaphor. Are there specific features, transformations, or visualizations that are impractical in such a system? One such potential limitation is that 3D visualizations in general are plagued by several challenges, such as occlusion, perspective foreshortening, legibility, reach, and the need for 3D navigation. However, this challenge is not unique to our system, and, besides, many of these problems can be mitigated [153]. Another, more relevant, limitation is the discoverability and usability of both voice commands as well as gestures. As noted by Norman [176], "*a pure gestural system makes it difficult to discover the set of possibilities*" that the system provides. The same is true of spoken commands. We attempt to address this in the paper by both providing visual guides, as well as basing most of our gestures on American Sign Language as well as common spoken words, respectively.[12] However, addressing discoverability in "natural interfaces" is not our focus with this paper, and I do not claim that our mitigation strategies are optimal.

Of course, Chapter 8 is primarily an engineering contribution and presents no empirical evidence on its own supporting any of my claims about the utility of mixed reality visualization authoring. Rather, my value proposition is qualitative: that visualization authoring in the field and on-the-go is only consistent with simple touchscreen interaction, mid-air gestures, and verbal commands in mixed reality, and that magic is as effective a metaphor as any (and more engaging than most) to guide users in this task. Fastidiously typing visualization specifications in JSON, or even selecting visualization templates or building blocks from a touchscreen menu, is not. While I agree that it would be productive to conduct empirical user studies to guide specific design choices or validate the utility of specific features, I firmly believe that the overall utility of our Wizualization system is sound and unassailable.

[12] Happily, no *Expecto Patronum!* for us.

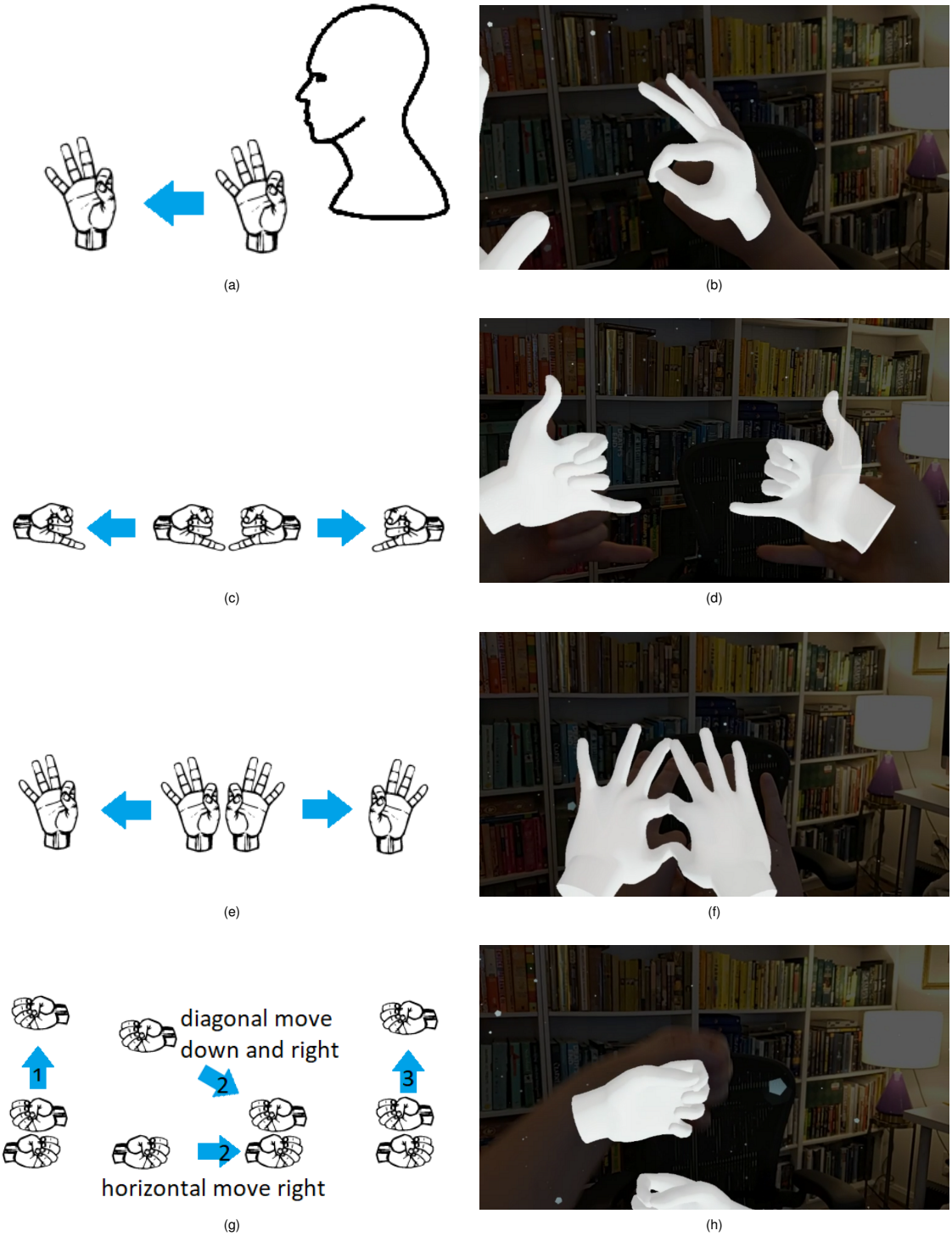
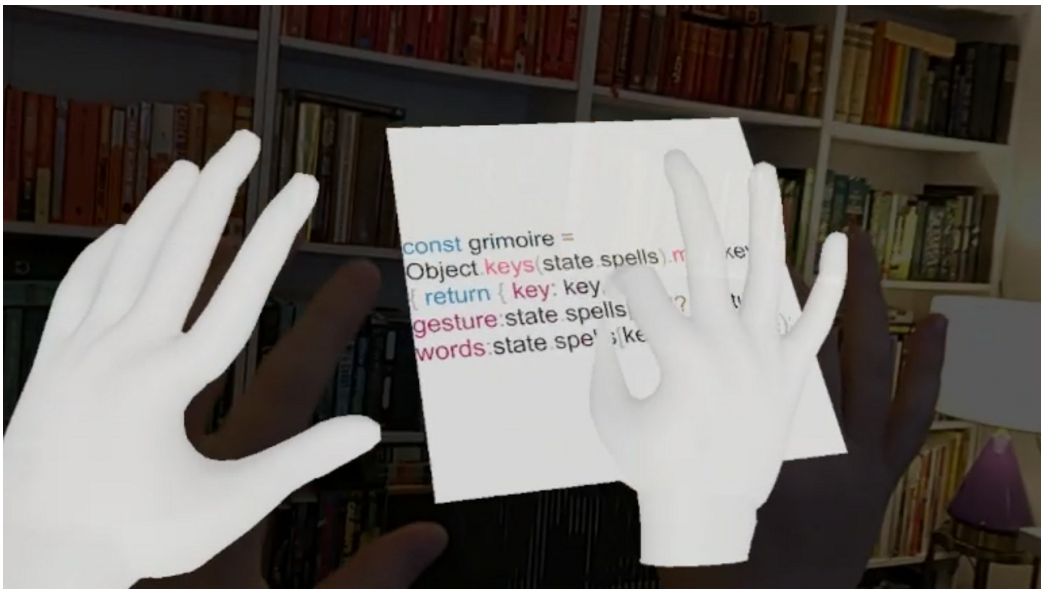



Figure 8.5: **Mark selection gestures.** From left to right, then top to bottom: ASL words for point [(a) and (b)], line [(c) and (d)], bar [(e) and (f)], and column [(g) and (h)]. Attribution for all hand illustrations in (a), (c), (e), and (g): American Sign Language alphabet, laid out by Darren Stone, derived from the Gallaudet-TT font. Attribution for head illustration in (a): Male Profile by Chris Homan from NounProject.com.




(a)


(b)

Figure 8.6: **Spellbook manipulation.** Spellbook's blocks can be directly grabbed and moved (a). All blocks are linked in a temporal sequence based on the actions performed by the user in evaluating the data (b).

# Chapter 9: Limitations

I view the limitations to my work as falling into three main categories: Methodological (in the case of qualitative work, which practically necessitates small user samples), technical, and ethical.

## 9.1 Limitations in Small-Sample Qualitative Work

Our contextual inquiry study [18] was not intended to be representative of all data scientists, so we must be careful about how our findings can be generalized. While our participants were all professionals who engaged in data-driven analysis on a daily basis, we were only able to recruit eight individuals to our study. However, data scientists are generally a protected population that are difficult to engage in studies such as ours. In other words, to our knowledge, our work is the first extended contextual inquiry to study this particular population of information professionals for the express purpose of understanding when and how they use interactive visualization, particularly in their initial exploratory analysis. For this reason, our findings provide at least an initial understanding of this type of visualization for data science from a human-computer interaction and visual analytics perspective.

Furthermore, all of our participants were employees of the U.S. federal government, which may also bias the type and scale of analysis projects they perform. It is possible that data analysts from industry, or even from outside the United States, may have a different outlook, process, or dataset scale and type. This may have an impact on how widely our

results can be applied. However, from our informal discussion with our participants, we are under the impression that the data science process—while far from standardized—looks similar across both government and industry. Our participants were all well-versed in the tools and software that data scientists use, and did not appear to be artificially constrained—in terms of budget, philosophy, or expertise—by the government agencies they worked for. As for scale, U.S. federal agencies remain one of the top clients for big data [179].

We did not anticipate that the majority of the second activity would, for all participants, be spent searching and reading—either about the data or about methods. In other words, participants spent a large portion of the sensemaking process in the early discovery phase even before the data was extracted, transformed, and loaded, and from participant responses, it is likely that they formed much of their intuitions about the data during this early stage. That much of data analysis is spent diagnosing, cleaning, and transforming a dataset prior to starting the actual analysis process has already been recognized as a major challenge [118]. In fact, Dasu and Johnson [58] estimate that up to 80% of the development time is spent on data cleaning. However, an interesting secondary finding from our results is that some of the sensemaking may already be happening during this discovery stage.

Another corollary from our study is that data scientists' actual work processes have left them, as users, to sit at a desk using a keyboard and mouse to navigate largely GUI-free lines of characters, both in discovering external data and for the purposes of syntactical error management. While such command-line interfaces are often powerful and effective for expert users, they make integration with interactive visual representations challenging. Tools such as R and Matlab do provide dedicated rendering systems to produce visualization windows, but these are more or less static and do not let the user interact with them in a meaningful way. RStudio extends R with, among other features, a viewer which interprets HTML/CSS/Javascript, and a tabular view panel for data.frame class objects.

In our uxSense [19] user study, to give another example, our evaluation was done with

a very small set of professional UX researchers from tech companies. Furthermore, our study used only a single sample user study session video for the evaluation activity. We acknowledge that it is unlikely that this video session would represent the potential variations of a wide variety of products and end-users. More work is needed for a representative measure of Wizualization's impact on UX researcher analysis over a broad range of users and products.

## 9.2 Technical Limitations

The engineering work required to construct a web-based situated analytics system with computer vision and other machine learning models that are updated based on user input is quite onerous. As such, technical issues may arise as early as the evaluation stage of developing and evaluating systems and techniques under one or both of these frameworks, or in broader IA work. These technical issues may affect data collection, interface performance, or timeliness in model output.

For example, in Batch *et al.* [22], when the user initialized an AR session, the scene camera no longer shares user position. As a workaround, we attached a `three.js` Object3D to the camera, and the world position of this object could be used to derive the user's position. However, upon review in VR views of the scene, the coordinates logged by the Object3D did not always perfectly correspond to those of the camera.

Evaluating the use of space in immersive environments is a relatively new focus in the literature, and results are not always clear-cut. For example, beyond our broad results in Batch *et al.* [22], there were a number of details in our results that were confounding. One such detail was in the participants' use of space. We suspect that there may be several of the following factors at play in explaining the pattern of the differences in participants' view distance during the second task discussed in Section 5.5. The difference from first to second

sequence may be the result of users becoming more competent with the system by the second sequence, correcting view distance issues encountered during the first sequence. The difference from second to third sequence may be a result of the user, now a seasoned expert, feeling free to take a leisurely stroll around the scene. Given the potential for confounding factors, we examined the data for outliers, but after reproducing Figure 5.8 several times with one potential outlier user omitted each time, we found that the results did not significantly diverge from Figure 5.8, which includes all users. Future remote research in WebXR should mind such pitfalls in data collection, but we are enthusiastic about its potential.

## 9.3 Ethical Limitations

In today's surveillance society, an increasing portion of our lives takes place inside a camera viewfinder. It seems as if every week uncovers some new horror of how digital video can be abused to threaten the privacy, security, and even safety of people just trying to live their lives. Thus, it could well be argued that a system such as uxSense is misguided in that it builds on fundamentally problematic ideas about recording human participants, and could even facilitate future abuses in the same vein. Indeed, while our current prototype does not include a facial recognition component, we could easily see the utility of having such a filter for when a system is deployed in the field and the researchers want to track how specific people use the system over time.

Rather than merely disavow any future such events as beyond our control, let us here acknowledge that this is a possible outcome. We ourselves commit to safeguarding our own use of the approach and prototype system so that the recordings are not distributed or used for other purposes than for what participants gave informed consent to.

We also note that researchers already collect plenty of video recordings of their participants, much of which is only lightly analyzed and then archived. These videos will obviously

identify the participants. Due to the prohibitive size of video, we suspect that much of this data resides on unprotected network or external drives. However, a fundamental feature of our approach is to process high-bandwidth video data to extract key data streams from the footage. These extracted data streams are refined and precise; the position of a person in three dimensions, their fatigue level, or the direction of their gaze. After all relevant data has been extracted, the original video can be deleted, thereby saving storage space but also eliminating identifiable likenesses of participants. Thus, it could be argued that our approach may actually serve to improve participant privacy, as it allows researchers to safely discard video while retaining deidentified data.

Nonetheless, the ethical considerations of data collection with such a system would require that video and audio data either be processed on the client side to return anonymized model output, or protected with secure system protocols. Since the computational requirements of computer vision modeling would have a significant impact on system performance if performed on the client side, I have opted to implement a secure authentication feature in my system.

# Chapter 10: Conclusions

In planning for future work, we are informed by our prior observations about the world. Nowhere is this more true than in research. This section summarizes my key findings, the implementations that have resulted from my work, and the limitations of my work, along with my plans and predictions for the future of SA and the modeling of multi-modal user input.

## 10.1 Questions Answered and Objectives Met

In Section 1.3, I asked how machine learning and related approaches for modeling user data can be further used to support immersive and situated analytics. The answer to this question was a rather circuitous one; along the way, we implemented uxSense (Section 7.2.1), a system for the general purpose of analyzing user experience, and recruited expert UX researchers to conduct our own analysis of uxSense (Sections 7.3 and 7.4). As part of the backend of that system, we developed an algorithmic pipeline for segmenting and clustering user gestural activity (Section 6.1), which we validated using a combination of a convolutional neural network and a recurrent neural network LSTM in our architecture for action classification (Sections 6.1.2 and 6.2). While the modeling of user behavior was more limited in our evaluation of economists (Section 4.1) in the VR implementation ImAxes [52], we found that visualization of their use of space (Section 4.2) was integral to our analysis of their behavior and our predictions (as discussed in Section 4.3).

Also in Section 1.3 I asked why, if there is much to be gained by adopting mobile, IA, or SA systems for serious data analysis, data scientists are not doing so. I clarified this as a question revolving around design requirements, specifically the design requirements for real data analysts, for such systems. To understand this, we conducted multiple studies with real data analysts and researchers, including the study of uxSense with professional UX researchers noted above (and in Sections 7.3 and 7.4). Our other studies involved a contextual inquiry in which we merely observed how data scientists and similar analysts go about exploratory analysis (Sections 3.1 and 3.2), a mixed methods study of how economists might go about using VR for exploring and presenting data (Sections 4.1 and 4.2), and another mixed methods study of how technical professionals perform question-answering tasks under different situated analytics view management techniques (Sections 5.4 and 5.5). These studies culminated in what amount to a collection of design recommendations, laid out in Sections 3.3 through 5.6 of the discussion in Chapter 9, for visual analytics tools and techniques for analytical professionals, particularly with regard to either (a) SA or IA or (b) the use of output from ML models to support qualitative analysis. These recommendations correspond to the design spaces discussed in Chapter 7, and present answers to a number of the design concerns discussed therein.

Ultimately, the answer to both of these questions can be boiled down to a straightforward, obvious one: Let the professionals do what they already know how to do using the tools they know how to use, but help them save their precious time and don't force them to dedicate mental resources to remembering esoteric aspects of the toolkit you are handing them. This is where an AR HMD has a distinct advantage over a VR HMD: The user does not have to change what they see around them; much like Wickham's ggplot [253], AR analytical systems simply add an additional layer over what the user already has in front of them.

Finally, again in Section 1.3, I stated my research objective as being the implementation of a SA system that makes use of multi-modal action and interaction data, and the result of

this objective was the second of the contributions noted in Section 1.5. With Wizualization, I try to put the recommendations for design discussed in the first four sections of Chapter 9 to use. In Chapter 8 of this paper, I have presented the product of this work in WIZUALIZATION, a mixed reality system for visualization authoring based on mid-air gestures and speech dressed up in a classic magic metaphor. Using this system, a user can create data visualizations in 3D space using a combination of hand gestures, verbal commands, and data bindings using an Augmented Reality HMD such as a Microsoft HoloLens 2 for seamless immersive analytics anywhere and anytime. I continue the magic metaphor to the components that make up the overall system: (i) *Optomancy* is the grammar of graphics for immersive visualization and interaction; (ii) *Spellbook* is the collection of gestures and verbal commands for invoking various actions (spells); (iii) *Arcane Focuses* integrate "magical artifacts" (smartphones) into the system and *Weave* propagates signals between the physical components; and (iv) *Wizualization* is the rendering system tying together all of these components on current-generation mixed reality HMDs. I have described each of these different components in depth and explained their interactions and roles in the overall system. This being an engineering contribution, my validation in the paper is the demonstration of how these different components come together in a single research prototype showcasing them in action. I direct the reader to `http://www.optomancy.com/` to test the system on their own given the appropriate hardware.

## 10.2 Future Work

I see significant potential for future work in the area of SA for data analysis professionals and researchers alike. Beyond the laboratory and field studies discussed earlier in this paper, there is opportunity for new visual representations, new mid-air interaction techniques, and new analytics methods that are specialized for the mobile analytics setting. Many

additional questions, however, remain. What is the equivalent of a dashboard in a mobile 3D workspace? How do you help users discover and remember gestures and commands? And how do you overcome challenges intrinsic to 3D such as occlusion, perspective foreshortening, and legibility? It is my sincere hope that the Wizualization ecosystem becomes used widely enough by developers and analysts to warrant ongoing iteration, maintenance, and field testing, but even if it does not, these are all questions I intend to pursue myself.

While the Wizualization ecosystem is built around multi-modal user inputs, another area of future work that I see room for significant growth is in multisensory analytical systems. In our work in information olfactation [21, 183], we introduce a theoretical framework and evaluation thereof for conveying features of data through smell. Under the design space of IA, I hope to see work on conveying data through non-visual senses in general continue to expand.

Finally, our work in using human pose and speech data—which we could call signals from the body—for evaluating user sessions in uxSense and then as multi-modal input in Wizualization shows that signals from the user can support both evaluation by the HCI researcher and system interaction and extension by the visual analytics system user. It is not too much of a leap to argue that the use of signals from the brain in brain-machine interfaces are another type of signal that could be explored in much greater depth in the field of visualization and visual analytics in general and in the domains of IA and SA more specifically.

# Addendum

Wizard's Tenth Rule: "Willfully turning aside from the truth is treason to one's self."

– Terry Goodkind, *Phantom* (2006)

"Ninety percent of most magic merely consists of knowing one extra fact."

– Terry Pratchett, *Night Watch* (2002)

"Do not meddle in the affairs of wizards, for they are subtle and quick to anger."

– J.R.R. Tolkien, *The Fellowship of the Ring* (1954)

# Bibliography


[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A system for large-scale machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[2] C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic queries for information exploration: An implementation and evaluation," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1992. doi: 10.1145/142750.143054 pp. 619–626.

[3] A. A. E. Ahmed and I. Traore, "A new biometric technology based on mouse dynamics," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, no. 3, pp. 165–179, 2007. doi: 10.1109/TDSC.2007.70207

[4] P. Anderson, J. Bowring, R. McCauley, G. Pothering, and C. Starr, "An undergraduate degree in data science: Curriculum and a decade of implementation experience," in *Proceedings of the ACM Symposium on Computer Science Education*, 2014. doi: 10.1145/2538862.2538936 pp. 145–150.

[5] C. Andrews, A. Endert, and C. North, "Space to think: large high-resolution displays for sensemaking," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2010. doi: 10.1145/1753326.1753336 pp. 55–64.

[6] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, "Pair analytics: Capturing reasoning processes in collaborative visual analytics," in *Proceedings of the Hawaii International Conference on System Sciences*. IEEE, 2011. doi: 10.1109/HICSS.2011.339 pp. 1–10.

[7] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *Computing Research Repository (CoRR)*, vol. abs/1710.07557, 2017. [Online]. Available: http://arxiv.org/abs/1710.07557

[8] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997. doi: 10.1162/pres.1997.6.4.355

[9] B. Bach, R. Sicat, J. Beyer, M. Cordeil, and H. Pfister, "The hologram in my hand: How effective is interactive exploration of 3D visualizations in immersive tangible augmented reality?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 457–467, 01 2018. doi: 10.1109/TVCG.2017.2745941

[10] S. K. Badam, F. Amini, N. Elmqvist, and P. Irani, "Supporting visual exploration for multiple users in large display environments," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/VAST.2016.7883506 pp. 1–10.

[11] S. K. Badam, A. Mathisen, R. Rädle, C. N. Klokmose, and N. Elmqvist, "Vistrates: A component model for ubiquitous analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 586–596, 2019. doi: 10.1109/TVCG.2018.2865144

[12] S. K. Badam, A. Srinivasan, N. Elmqvist, and J. Stasko, "Affordances of input modalities for visual data exploration in immersive environments," in *Proceedings of the IEEE VIS Workshop on Immersive Analytics*, 2017.

[13] M. Balduini, I. Celino, D. Dell'Aglio, E. Della Valle, Y. Huang, T. Lee, S.-H. Kim, and V. Tresp, "Bottari: An augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams," *Journal of Web Semantics*, vol. 16, pp. 33 – 41, 2012. doi: 10.1016/j.websem.2012.06.004

[14] R. Ball, C. North, and D. A. Bowman, "Move to improve: promoting physical navigation to increase user performance with large displays," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2007. doi: 10.1145/1240624.1240656 pp. 191–200.

[15] D. Banakou, R. Groten, and M. Slater, "Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12 846–12 851, 2013. doi: 10.1073/pnas.1306779110

[16] E. Barba, B. MacIntyre, and E. Mynatt, "Here We Are! Where Are We? Locating Mixed Reality in The Age of the Smartphone," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 929 –936, April 2012. doi: 10.1109/JPROC.2011.2182070

[17] A. Batch, A. Cunningham, M. Cordeil, N. Elmqvist, T. Dwyer, B. H. Thomas, and K. Marriott, "There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics," *IEEE Transactions*

*on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 536–546, 2020. doi: 10.1109/TVCG.2019.2934803

[18] A. Batch and N. Elmqvist, "The interactive visualization gap in initial exploratory data analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 278–287, Jan. 2018. doi: 10.1109/TVCG.2017.2743990

[19] A. Batch, Y. Ji, J. Zhao, M. Fan, and N. Elmqvist, "uxSense: Supporting user experience evaluation using visualization and computer vision," in *IEEE Transactions on Visualization and Computer Graphics*. Piscataway, NJ, USA: IEEE, 2022 (pending review).

[20] A. Batch, K. Lee, H. T. Maddali, and N. Elmqvist, "Gesture and action discovery for evaluating virtual environments with semi-supervised segmentation of telemetry records," in *Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality*. Piscataway, NJ, USA: IEEE, 2018. doi: 10.1109/AIVR.2018.00009

[21] A. Batch, B. Patnaik, M. Akazue, and N. Elmqvist, "Scents and sensibility: Evaluating information olfactation," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2020. doi: 10.1145/3313831.3376733. ISBN 9781450367080 p. 1–14.

[22] A. Batch, S. Shin, J. Liu, P. W. Butcher, P. Ritsos, and N. Elmqvist, "The world is your Holodeck: View management for situated visualization," in *IEEE Transactions on Visualization and Computer Graphics*. Piscataway, NJ, USA: IEEE, 2022 (pending review).

[23] A. Begel and T. Zimmermann, "Analyze this! 145 questions for data scientists in software engineering," in *Proceedings of the ACM International Conference on Software Engineering*, 2014. doi: 10.1145/2568225.2568233 pp. 12–23.

[24] A. H. Behzadan and V. R. Kamat, "Visualization of construction graphics in outdoor augmented reality," in *Proceedings of the Winter Simulation Conference*. Piscataway, NJ, USA: IEEE, 2005. doi: 10.1109/WSC.2005.1574469 p. 7.

[25] B. Bell, S. Feiner, and T. Höllerer, "View management for virtual and augmented reality," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2001. doi: 10.1145/502348.502363 pp. 101–110.

[26] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am I a baller? Basketball performance assessment from first-person videos," in *Proceedings of the IEEE International Conference on Computer Vision*. Piscataway, NJ, USA: IEEE, 2017. doi: 10.1109/ICCV.2017.239 pp. 2196–2204.

[27] J. Bertin, *Sémiologie graphique*. Paris, France: Mouton/Gauthier-Villars, 1967. ISBN 978-1589482616

[28] L. Besançon, A. Ynnerman, D. F. Keefe, L. Yu, and T. Isenberg, "The state of the art of spatial interfaces for 3D visualization," *Computer Graphics Forum*, vol. 40, no. 1, pp. 293–326, 2021. doi: 10.1111/cgf.14189

[29] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems*, ser. Interactive Technologies. Elsevier Science, 1997. ISBN 9780080503042

[30] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl, "VA2: A visual analytics approach for evaluating visual analytics applications," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 61–70, 2016. doi: 10.1109/TVCG.2015.2467871

[31] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "Visualization of eye tracking data: A taxonomy and survey," *Computer Graphics Forum*, vol. 36, no. 8, pp. 260–284, 2017. doi: 10.1111/cgf.13079

[32] R. A. Bolt, ""Put-That-There": Voice and gesture at the graphics interface," in *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: ACM, 1980. doi: 10.1145/800250.807503 p. 262–270.

[33] M. Bostock, V. Ogievetsky, and J. Heer, "$D^3$: Data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec 2011. doi: 10.1109/TVCG.2011.185

[34] B. Burr, "VACA: A tool for qualitative video analysis," in *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. ACM, 2006. doi: 10.1145/1125451.1125580. ISBN 1-59593-298-4 pp. 622–627.

[35] W. Büschel, J. Chen, R. Dachselt, S. Drucker, T. Dwyer, C. Görg, T. Isenberg, A. Kerren, C. North, and W. Stuerzlinger, "Interaction for immersive analytics," in *Immersive Analytics*, K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas, Eds. Cham: Springer International Publishing, 2018, pp. 95–138. ISBN 978-3-030-01388-2

[36] W. Büschel, A. Lehmann, and R. Dachselt, "MIRIA: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2021. doi: 10.1145/3411764.3445651

[37] P. W. S. Butcher, N. W. John, and P. D. Ritsos, "VRIA: A web-based framework for creating immersive analytics experiences," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 7, pp. 3213–3225, 2021. doi: 10.1109/TVCG.2020.2965109

[38] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer, "Clusters, trends, and outliers: How immersive technologies can facilitate the collaborative analysis of multidimensional data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2018. doi: 10.1145/3173574.3173664. ISBN 9781450356206 pp. 90:1—-90:12.

[39] E. Cambria, G. Huang, L. L. C. Kasun, H. Zhou, C. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. C. M. Leung, L. Feng, Y. Ong, M. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B. Oh, J. Jeon, K. Toh, A. B. J. Teoh, J. Kim, H. Yu, Y. Chen, and J. Liu, "Extreme learning machines [trends & controversies]," *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 30–59, 2013. doi: 10.1109/MIS.2013.140

[40] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. doi: 10.1109/CVPR.2017.143 pp. 1302–1310.

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2017. doi: 10.1109/CVPR.2017.502 pp. 4724–4733.

[42] M. Cavallo, G. A. Rhodes, and A. G. Forbes, "Riverwalk: Incorporating historical photographs in public outdoor augmented reality experiences," in *Adjunct Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/ISMAR-Adjunct.2016.0068 pp. 160–165.

[43] T. Chandler, M. Cordeil, T. Czauderna, T. Dwyer, J. Glowacki, C. Goncu, M. Klapperstueck, K. Klein, F. Schreiber, and E. Wilson, "Immersive analytics," in *Proceedings of the International Symposium on Big Data Visual Analytics*. Piscataway, NJ, USA: IEEE, Sep 2015. doi: 10.1109/BDVA.2015.7314296 pp. 1–8.

[44] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Peppler, N. Elmqvist, and K. Ramani, "VizScribe: A visual analytics approach to understand designer behavior," *International Journal of Human-Computer Studies*, vol. 100, pp. 66 – 80, 2017. doi: 10.1016/j.ijhcs.2016.12.007

[45] Z. Chen, Y. Su, Y. Wang, Q. Wang, H. Qu, and Y. Wu, "MARVisT: authoring glyph-based visualization in mobile augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 8, pp. 2645–2658, 2019. doi: 10.1109/TVCG.2019.2892415

[46] M. Cherubini, G. Venolia, R. DeLine, and A. J. Ko, "Let's go to the whiteboard: How and why software developers use drawings," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY,

USA: ACM, 2007. doi: 10.1145/1240624.1240714. ISBN 978-1-59593-593-9 pp. 557–566.

[47] A. Clark and D. Chalmers, "The extended mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998. doi: 10.1093/analys/58.1.7

[48] T. Clegg, E. Bonsignore, J. Yip, H. Gelderblom, A. Kuhn, T. Valenstein, B. Lewittes, and A. Druin, "Technology for promoting scientific practice and personal meaning in life-relevant learning," in *Proceedings of the 11th International Conference on Interaction Design and Children*. ACM, 2012, pp. 152–161.

[49] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, Sep. 1984. doi: 10.2307/2288400

[50] M. Cordeil, B. Bach, Y. Li, E. Wilson, and T. Dwyer, "Design space for spatio-data coordination: Tangible interaction devices for immersive information visualisation," in *Proceedings of the IEEE Pacific Symposium on Visualization*. Piscataway, NJ, USA: IEEE, 2017. doi: 10.1109/PACIFICVIS.2017.8031578 pp. 46–50.

[51] M. Cordeil, A. Cunningham, B. Bach, C. Hurter, B. H. Thomas, K. Marriott, and T. Dwyer, "IATK: an immersive analytics toolkit," in *Proceedings of the IEEE Conference on Virtual Reality*. Piscataway, NJ, USA: IEEE, 2019. doi: 10.1109/VR.2019.8797978 pp. 200–209.

[52] M. Cordeil, A. Cunningham, T. Dwyer, B. H. Thomas, and K. Marriott, "ImAxes: Immersive axes as embodied affordances for interactive multivariate data visualisation," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2017. doi: 10.1145/3126594.3126613. ISBN 978-1-4503-4981-9 pp. 71–83.

[53] M. Cordeil, T. Dwyer, K. Klein, B. Laha, K. Marriott, and B. H. Thomas, "Immersive collaborative analysis of network connectivity: CAVE-style or head-mounted display?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 441–450, 2017. doi: 10.1109/TVCG.2016.2599107

[54] A. Crisan, B. Fiore-Gartland, and M. Tory, "Passing the data baton : A retrospective analysis on data science work and workers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1860–1870, 2021. doi: 10.1109/TVCG.2020.3030340

[55] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. New York, NY, USA: Harper Collins, 1991.

[56] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J.-G. Lou, and D. Zhang, "Text-to-Viz: Automatic generation of infographics from proportion-related natural language statements," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 906–916, 2020. doi: 10.1109/TVCG.2019.2934785

[57] R. Y. da Silva Franco, R. Santos do Amor Divino Lima, M. Paixão, C. G. Resque dos Santos, B. Serique Meiguins *et al.*, "Uxmood—a sentiment analysis and information visualization tool to support the evaluation of usability and user experience," *Information*, vol. 10, no. 12, p. 366, 2019.

[58] T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*. Wiley, 2003.

[59] C. Demiralp, C. D. Jackson, D. B. Karelitz, S. Zhang, and D. H. Laidlaw, "Cave and fishtank virtual-reality displays: A qualitative and quantitative comparison," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 3, pp. 323–330, 2006. doi: 10.1109/TVCG.2006.42

[60] C. Diaz, M. Walker, D. A. Szafir, and D. Szafir, "Designing for depth perceptions in augmented reality," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, 2017. doi: 10.1109/ISMAR.2017.28 pp. 111–122.

[61] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang, "Recovering reasoning processes from user interactions," *IEEE Computer Graphics and Applications*, vol. 29, no. 3, pp. 52–61, 2009. doi: 10.1109/MCG.2009.49

[62] M. Drouhard, N.-C. Chen, J. Suh, R. Kocielnik, V. Pena-Araya, K. Cen, X. Zheng, and C. R. Aragon, "Aeonium: Visual analytics to support collaborative qualitative coding," in *2017 IEEE Pacific Visualization Symposium (PacificVis)*. Seoul, South Korea: IEEE, 2017. doi: 10.1109/PACIFICVIS.2017.8031598. ISSN 2165-8773 pp. 220–229.

[63] A. Druin, "Cooperative inquiry: Developing new technologies for children with children," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1999. doi: 10.1145/302979.303166 pp. 592–599.

[64] D. Dua and C. Graff, "UCI machine learning repository," http://archive.ics.uci.edu/ml, 2017.

[65] T. Dwyer, K. Marriott, T. Isenberg, K. Klein, N. Riche, F. Schreiber, W. Stuerzlinger, and B. H. Thomas, "Immersive analytics: An introduction," in *Immersive Analytics*, ser. Lecture Notes in Computer Science. Springer, 2018, vol. 11190, pp. 1–23. ISBN 978-3-030-01388-2

[66] N. Elmqvist and P. Irani, "Ubiquitous analytics: Interacting with big data anywhere, anytime," *IEEE Computer*, vol. 46, no. 4, pp. 86–89, 2013. doi: 10.1109/MC.2013.147

[67] N. Elmqvist and P. Tsigas, "A taxonomy of 3D occlusion management for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 5, pp. 1095–1109, 2008. doi: 10.1109/TVCG.2008.59

[68] N. Elmqvist, A. Vande Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, "Fluid interaction for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 327–340, Oct. 2011. doi: 10.1177/1473871611413180

[69] N. A. M. ElSayed, B. H. Thomas, K. Marriott, J. Piantadosi, and R. T. Smith, "Situated analytics: Demonstrating immersive analytical tools with augmented reality," *Journal of Visual Languages & Computing*, vol. 36, pp. 13–23, 2016. doi: 10.1016/j.jvlc.2016.07.006

[70] N. A. ElSayed, R. T. Smith, and B. H. Thomas, "HORUS EYE: See the invisible bird and snake vision for augmented reality information visualization," in *Adjunct Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/ISMAR-Adjunct.2016.0077 pp. 203–208.

[71] B. Ens, B. Bach, M. Cordeil, U. Engelke, M. Serrano, W. Willett, A. Prouzeau, C. Anthes, W. Büschel, C. Dunne, T. Dwyer, J. Grubert, J. H. Haga, N. Kirshenbaum, D. Kobayashi, T. Lin, M. Olaosebikan, F. Pointecker, D. Saffo, N. Saquib, D. Schmalstieg, D. A. Szafir, M. Whitlock, and Y. Yang, "Grand challenges in immersive analytics," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: ACM, 2021. doi: 10.1145/3411764.3446866. ISBN 9781450380966

[72] ——, "Grand challenges in immersive analytics," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2021. doi: 10.1145/3411764.3446866 pp. 459:1–459:17.

[73] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/CVPR.2016.600 pp. 5562–5570.

[74] M. Fan, S. Shi, and K. N. Truong, "Practices and challenges of using think-aloud protocols in industry: An international survey." *Journal of Usability Studies*, vol. 15, no. 2, pp. 85–102, 2020.

[75] M. Fan, K. Wu, J. Zhao, Y. Li, W. Wei, and K. N. Truong, "VisTA: Integrating machine intelligence with visualization to support the investigation of think-aloud sessions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 343–352, 2020. doi: 10.1109/TVCG.2019.2934797

[76] S. Feiner, B. MacIntyre, H. Tobias, and A. Webster, "A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment," in *Proceedings of the International Symposium on Wearable Computers*. Piscataway, NJ, USA: IEEE, Oct. 1997. doi: 10.1007/BF01682023 pp. 74–81.

[77] C. Felix, A. Dasgupta, and E. Bertini, "The exploratory labeling assistant: Mixed-initiative label curation with large document collections," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '18. New York, NY, USA: Association for Computing Machinery, 2018. doi: 10.1145/3242587.3242596. ISBN 9781450359481 p. 153–164.

[78] K. Fennedy, J. Hartmann, Q. Roy, S. T. Perrault, and D. Vogel, "OctoPocus in VR: Using a dynamic guide for 3D mid-air gestures in virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 12, pp. 4425–4438, 2021. doi: 10.1109/TVCG.2021.3101854

[79] S. Few, *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press, 2004.

[80] ——, "Data sensemaking requires time and attention," *Visual Business Intelligence*, June 2015. [Online]. Available: https://www.perceptualedge.com/blog/?p=2052

[81] J. A. W. Filho, W. Stuerzlinger, and L. Nedel, "Evaluating an immersive space-time cube geovisualization for intuitive trajectory data exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 514–524, 2020. doi: 10.1109/TVCG.2019.2934415

[82] P. Fleck, A. Sousa Calepso, S. Hubenschmid, M. Sedlmair, and D. Schmalstieg, "RagRug: A toolkit for situated analytics," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2022. doi: 10.1109/TVCG.2022.3157058

[83] G. Fontaine, "The experience of a sense of presence in intercultural and international encounters," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 4, pp. 482–490, Jan. 1992. doi: 10.1162/pres.1992.1.4.482

[84] B. Fröhler, C. Anthes, F. Pointecker, J. Friedl, D. Schwajda, A. Riegler, S. Tripathi, C. Holzmann, M. Brunner, H. Jodlbauer, H.-C. Jetter, and C. Heinzl, "A survey on cross-virtuality analytics," *Computer Graphics Forum*, vol. 41, no. 1, pp. 465–494, 2022. doi: 10.1111/cgf.14447

[85] D. D. Garcia and D. Ginat, "DeMystifying computing with magic," in *Proceedings of the ACM Symposium on Computer Science Education*. New York, NY, USA: ACM, 2012. doi: 10.1145/2157136.2157164 pp. 83–84.

[86] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-Track: Efficient pose estimation in videos," *CoRR*, vol. abs/1712.09184, 2017. [Online]. Available: http://arxiv.org/abs/1712.09184

[87] S. Goodwin, C. Mears, T. Dwyer, M. G. de la Banda, G. Tack, and M. Wallace, "What do constraint programming users want to see? exploring the role of visualisation in profiling of models and search," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 281–290, Jan 2017. doi: 10.1109/TVCG.2016.2598545

[88] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," in *2008 IEEE Symposium on Visual Analytics Science and Technology*. Piscataway, NJ, USA: IEEE, 2008. doi: 10.1109/VAST.2008.4677365 pp. 123–130.

[89] R. Grasset, T. Langlotz, D. Kalkofen, M. Tatzgern, and D. Schmalstieg, "Image-driven view management for augmented reality browsers," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2012.

[90] G. Grolemund and H. Wickham, "A cognitive interpretation of data analysis," *International Statistical Review*, vol. 82, no. 2, pp. 184–204, 2014. doi: 10.1111/insr.12028

[91] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proceedings of the IEEE European Conference on Computer Vision*. Piscataway, NJ, USA: IEEE, 2014. doi: 10.1007/978-3-319-10584-0˙33. ISBN 978-3-319-10583-3 pp. 505–520.

[92] J. Hagedorn, J. M. Hailpern, and K. Karahalios, "VCode and VData: Illustrating a new framework for supporting the video annotation workflow," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*, 2008, pp. 317–321.

[93] G. Halter, R. Ballester-Ripoll, B. Flueckiger, and R. Pajarola, "VIAN: A visual annotation tool for film analysis," *Computer Graphics Forum*, vol. 38, no. 3, pp. 119–129, 2019. doi: 10.1111/cgf.13676

[94] K. Harezlak, P. Kasprowski, and M. Stasch, "Idiosyncratic repeatability of calibration errors during eye tracker calibration," in *Proceedings of the International Conference on Human System Interactions*. Piscataway, NJ, USA: IEEE, 2014. doi: 10.1109/HSI.2014.6860455 pp. 95–100.

[95] A. G. Hauptmann and P. McAvinney, "Gestures with speech for graphic manipulation," *International Journal for Man-Machine Studies*, vol. 38, no. 2, pp. 231–249, 1993. doi: 10.1006/imms.1993.1011

[96] S. He, M. L. Huang, and L. Zhu, "Natural user interface design in DA-TU: An interactive clustered data visualization system," in *Proceedings of the International Conference on Information Visualisation*. Piscataway, NJ, USA: IEEE, 2015. doi: 10.1109/iV.2015.25 pp. 83–88.

[97] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1189–1196, 2008. doi: 10.1109/TVCG.2008.137

[98] S. Heng and D. Yunfeng, "Research on cooperative control of human-computer interaction tools with high recognition rate based on neural network," in *Proceedings of the IEEE International Conference on Virtual Reality and Visualization*. Piscataway, NJ, USA: IEEE, 2014. doi: 10.1109/ICVRV.2014.6 pp. 350–354.

[99] K. Hinckley, R. Pausch, J. C. Goble, and N. F. Kassell, "Passive real-world interface props for neurosurgical visualization," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1994. doi: 10.1145/191666.191821. ISBN 0897916506 pp. 452—458.

[100] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[101] H. Holle and R. Rein, "The modified cohen's kappa: Calculating interrater agreement for segmentation and annotation," *Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour*, pp. 261–275, Jan. 2013.

[102] K. Holtzblatt and H. Beyer, *Contextual Design: Evolved*, ser. Synthesis Lectures on Human-Centered Informatics. San Rafael, CA, USA: Morgan & Claypool Publishers, 2014. ISBN 9781627055598

[103] J. Howard, *Game Magic: A Designer's Guide to Magic Systems in Theory and Practice*. CRC Press, 2014. ISBN 9781466567870

[104] S. Hubenschmid, J. Wieland, D. I. Fink, A. Batch, J. Zagermann, N. Elmqvist, and H. Reiterer, "ReLive: Bridging in-situ and ex-situ visual analytics for analyzing mixed reality user studies," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2022. doi: 10.1145/3491102.3517550

[105] S. Hubenschmid, J. Zagermann, S. Butscher, and H. Reiterer, "STREAM: Exploring the combination of spatially-aware tablets with augmented reality head-mounted displays for immersive analytics," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2021. doi: 10.1145/3411764.3445298

[106] C. Hurter, N. Riche, S. Drucker, M. Cordeil, R. Alligier, and R. Vuillemot, "FiberClay: Sculpting three dimensional trajectories to reveal structural insights," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 704–714, 2019. doi: 10.1109/TVCG.2018.2865191

[107] E. Hutchins, *Cognition in the Wild*. Cambridge, MA, USA: MIT Press, 1995. ISBN 9780262581462

[108] H. Hutchinson, W. Mackay, B. Westerlund, B. B. Bederson, A. Druin, C. Plaisant, M. Beaudouin-Lafon, S. Conversy, H. Evans, H. Hansen *et al.*, "Technology probes: inspiring design for and with families," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 17–24.

[109] T. Igarashi, S. Matsuoka, and H. Tanaka, "Teddy: A sketching interface for 3D freeform design," in *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: ACM, 1999. doi: 10.1145/311535.311602 p. 409–416.

[110] H. Imottesjo and J.-H. Kain, "The Urban CoBuilder – a mobile augmented reality tool for crowd-sourced simulation of emergent urban development patterns: Requirements, prototyping and assessment," *Computers, Environment and Urban Systems*, vol. 71, pp. 120–130, 2018. doi: 10.1016/j.compenvurbsys.2018.05.003

[111] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multidimensional geometry," in *Proceedings of the IEEE Conference on Visualization*. Piscataway, NJ, USA: IEEE, 1990. doi: 10.1109/VISUAL.1990.146402 pp. 361–378.

[112] N. A. James and D. S. Matteson, "ecp: An R package for nonparametric multiple change point analysis of multivariate data," *Journal of Statistical Software*, vol. 62, no. i07, 2015. [Online]. Available: https://arxiv.org/abs/1309.3295

[113] S. Jang, N. Elmqvist, and K. Ramani, "MotionFlow: Visual abstraction and aggregation of sequential patterns in human motion tracking data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 21–30, 2016. doi: 10.1109/TVCG.2015.2468292

[114] J. Jerald, *The VR Book: Human-Centered Design for Virtual Reality*. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2015. ISBN 9781970001129

[115] S. Jolaoso, R. Burtner, and A. Endert, "Toward a deeper understanding of data analysis, sensemaking, and signature discovery," in *Proceedings of INTERACT*, 2015. doi: 10.1007/978-3-319-22668-2˙36 pp. 463–478.

[116] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. C. Nabbe, I. A. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic Studio: A massively multiview system for social interaction capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. doi: 10.1109/TPAMI.2017.2782743. [Online]. Available: http://arxiv.org/abs/1612.03153

[117] S. Kallio, J. Kela, J. Mäntyjärvi, and J. Plomp, "Visualization of hand gestures for pervasive computing environments," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*. New York, NY, USA: ACM, 2006. doi: 10.1145/1133265.1133363 p. 480–483.

[118] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling," *Information Visualization*, vol. 10, no. 4, pp. 271–288, 2011.

[119] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, "Enterprise data analysis and visualization: An interview study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2917–2926, Dec 2012. doi: 10.1109/TVCG.2012.219

[120] D. F. Keefe, D. A. Feliz, J. Miles, F. Drury, S. Swartz, and D. H. Laidlaw, "Scientific sketching for collaborative VR visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 4, pp. 835–847, 2008. doi: 10.1109/TVCG.2008.31

[121] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2000, pp. 285–289.

[122] C. Kerdvibulvech and H. Saito, "Vision-based detection of guitar players' fingertips without markers," in *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*. Piscataway, NJ, USA: IEEE, 2007. doi: 10.1109/CGIV.2007.88 pp. 419–428.

[123] K. Kilteni, R. Groten, and M. Slater, "The sense of embodiment in virtual reality," *Presence: Teleoperators and Virtual Environments*, vol. 21, no. 4, pp. 373–387, 2012. doi: 10.1162/PRES˙a˙00124

[124] C. Kim, P. Chiu, and Y. Tjahjadi, "A web-based remote assistance system with gravity-aware 3D hand gesture visualization," in *Proceedings of the ACM Conference on Interactive Surfaces and Spaces*. New York, NY, USA: ACM, 2019. doi: 10.1145/3343055.3360742 p. 319–322.

[125] D. Kirsh, "The intelligent use of space," *Artificial Intelligence*, vol. 73, no. 1–2, pp. 31–68, 1995. doi: 10.1016/0004-3702(94)00017-U

[126] D. Kirsh and P. Maglio, "On distinguishing epistemic from pragmatic action," *Cognitive Science*, vol. 18, no. 4, pp. 513–549, 1994. doi: 10.1207/s15516709cog1804˙1

[127] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2008. doi: 10.1145/1357054.1357127 pp. 453–456.

[128] D. Koller, P. Lindstrom, W. Ribarsky, L. F. Hodges, N. Faust, and G. Turner, "Virtual GIS: a real-time 3D geographic information system," in *Proceedings of the IEEE Conference on Visualization*. Piscataway, NJ, USA: IEEE, 1995. doi: 10.1109/VISUAL.1995.480800 pp. 94–100.

[129] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/CVPR.2016.239 pp. 2176–2184.

[130] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf, "Visual movie analytics," *Trans. Multi.*, vol. 18, no. 11, p. 2149–2160, Nov. 2016. doi: 10.1109/TMM.2016.2614184

[131] B. Laha, D. A. Bowman, and J. J. Socha, "Effects of VR system fidelity on analyzing isosurface visualization of volume datasets," *IEEE Transactions on Visualization & Computer Graphics*, vol. 20, no. 4, pp. 513–522, Apr. 2014. doi: 10.1109/TVCG.2014.20

[132] R. Langner, M. Satkowski, W. Büschel, and R. Dachselt, "MARVIS: Combining mobile devices and augmented reality for visual data analysis," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 5 2021. doi: 10.1145/3411764.3445593. ISBN 978-1-4503-8096-6/21/05

[133] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2008. doi: 10.1109/CVPR.2008.4587756. ISSN 1063-6919 pp. 1–8.

[134] W. S. Lasecki, M. Gordon, D. Koutra, M. F. Jung, S. P. Dow, and J. P. Bigham, "Glance: Rapidly coding behavioral video with the crowd," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2014. doi: 10.1145/2642918.2647367. ISBN 978-1-4503-3069-5 pp. 551–562.

[135] M. Leake, H. V. Shin, J. O. Kim, and M. Agrawala, "Generating audio-visual slideshows from text articles using word concreteness," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020. doi: 10.1145/3313831.3376519. ISBN 9781450367080 p. 1–11.

[136] B. Lee, R. H. Kazi, and G. Smith, "SketchStory: Telling more engaging stories with data through freeform sketching," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2416–2425, 2013. doi: 10.1109/TVCG.2013.191

[137] B. Lee, A. Srinivasan, J. Stasko, M. Tory, and V. Setlur, "Multimodal interaction for data visualization," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*. New York, NY, USA: ACM, 2018. doi: 10.1145/3206505.3206602

[138] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff, "A cross-media presence questionnaire: The ITC-Sense of Presence Inventory," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 282–297, 2001. doi: 10.1162/105474601300343612

[139] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107–144, 2007. doi: 10.1007/s10618-007-0064-z

[140] T. Lin, Y. Yang, J. Beyer, and H. Pfister, "SportsXR—-Immersive analytics in sports," in *Proceedings of the ACM CHI Workshop on Immersive Analytics*, 2020. doi: 10.48550/arXiv.2004.08010

[141] M. A. Linton, J. M. Vlissides, and P. R. Calder, "Composing user interfaces with InterViews," *Computer*, vol. 22, no. 2, pp. 8–22, 1989. doi: 10.1109/2.19829

[142] H. R. Lipford, F. Stukes, W. Dou, M. E. Hawkins, and R. Chang, "Helping users recall their reasoning process," in *IEEE Symposium on Visual Analytics Science and Technology*. Piscataway, NJ, USA: IEEE, 2010. doi: 10.1109/VAST.2010.5653598 pp. 187–194.

[143] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2014. doi: 10.1109/CVPR.2014.226 pp. 1749–1756.

[144] M. J. Lobo, C. Hurter, and P. Irani, "Flex-ER: A platform to evaluate interaction techniques for immersive visualizations," *Proceedings of the ACM Conference on Human Factors in Computing Systems*, vol. 4, no. ISS, 2020. doi: 10.1145/3427323

[145] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2017. doi: 10.1109/CVPR.2017.493 pp. 4636–4644.

[146] W. E. Mackay, "Augmented reality: Linking real and virtual worlds: A new paradigm for interacting with computers," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*. New York, NY, USA: ACM, 1998. doi: 10.1145/948496.948498 pp. 13–21.

[147] J. D. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Transactions on Graphics*, vol. 5, no. 2, pp. 110–141, 1986. doi: 10.1145/22949.22950

[148] K. Madhavan, N. Elmqvist, M. Vorvoreanu, X. Chen, Y. Wong, H. Xian, Z. Dong, and A. Johri, "DIA2: Web-based cyberinfrastructure for visual analysis of funding portfolios," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1823–1832, Dec 2014. doi: 10.1109/TVCG.2014.2346747

[149] J. B. Madsen, M. Tatzqern, C. B. Madsen, D. Schmalstieg, and D. Kalkofen, "Temporal coherence strategies for augmented reality labeling," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 4, pp. 1415–1423, 2016. doi: 10.1109/TVCG.2016.2518318

[150] E. Maguire, P. Rocca-Serra, S.-A. Sansone, J. Davies, and M. Chen, "Visual compression of workflow visualizations with automated detection of macro motifs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2576–2585, 2013. doi: 10.1109/TVCG.2013.225

[151] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. doi: 10.1109/CVPR.2017.318 pp. 202–211.

[152] G. Marqués and K. Basterretxea, "Efficient algorithms for accelerometer-based wearable hand gesture recognition systems," in *Proceedings of the IEEE International Conference on Embedded and Ubiquitous Computing*. Piscataway, NJ, USA: IEEE, 2015. doi: 10.1109/EUC.2015.25 pp. 132–139.

[153] K. Marriott, J. Chen, M. Hlawatsch, T. Itoh, M. A. Nacenta, G. Reina, and W. Stuerzlinger, "Immersive analytics: Time to reconsider the value of 3d for information visualisation," in *Immersive Analytics*, ser. Lecture Notes in Computer Science, K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas, Eds. Springer, 2018, vol. 11190, pp. 25–55. ISBN 978-3-030-01388-2

[154] K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas, Eds., *Immersive Analytics*, ser. Lecture Notes in Computer Science. New York, NY, USA: Springer International Publishing, 2018, vol. 11190. ISBN 978-3-030-01388-2

[155] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 334–345, 2014. doi: 10.1080/01621459.2013.849605

[156] J. McCormack, B. Roberts, Jonathan C.and Bach, C. D. S. Freitas, T. Itoh, C. Hurter, and K. Marriott, "Multisensory immersive analytics," in *Immersive Analytics*, K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas, Eds. Cham: Springer International Publishing, 2018, pp. 57–94. ISBN 978-3-030-01388-2

[157] W. McKinney, "Data structures for statistical computing in Python," in *Proceedings of the Python in Science Conference*, 2010, pp. 51–56.

[158] P. McLachlan, T. Munzner, E. Koutsofios, and S. North, "Liverac: Interactive visual exploration of system management time-series data," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1357054.1357286. ISBN 978-1-60558-011-1 pp. 1483–1492.

[159] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 916–929, 2016. doi: 10.1109/TCYB.2015.2418092

[160] A. M. P. Milani, F. V. Paulovich, and I. H. Manssour, "Visualization in the preprocessing phase: Getting insights from enterprise professionals," *Information Visualization*, vol. 19, no. 4, pp. 273–287, 2020. doi: 10.1177/1473871619896101

[161] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented reality: A class of displays on the reality-virtuality continuum," in *Proceedings of Telemanipulator and Telepresence Technologies.*, vol. 2351. SPIE, 1995. doi: 10.1117/12.197321 pp. 282–292.

[162] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/CVPR.2016.456 pp. 4207–4215.

[163] A. Mottelson and K. Hornbæk, "Virtual reality studies outside the laboratory," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. New York, NY, USA: ACM, 2017. doi: 10.1145/3139131.3139141 pp. 9:1–9:10.

[164] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015. doi: 10.1109/TMM.2015.2482819

[165] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-74048-3˙4 pp. 69–84.

[166] D. Müllner, "fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python," *Journal of Statistical Software*, vol. 53, no. i09, 2013.

[167] K. M. Munim, I. Islam, M. Khatun, M. M. Karim, and M. N. Islam, "Towards developing a tool for UX evaluation using facial expression," in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, IEEE. Piscataway, NJ, USA: IEEE, 2017. doi: 10.1109/EICT.2017.8275227 pp. 1–6.

[168] T. Munzner, *Visualization Analysis and Design*. Boca Raton, FL, USA: CRC Press, 2014. ISBN 9781466508910. [Online]. Available: http://www.cs.ubc.ca/%7Etmm/vadbook/

[169] B. A. Myers, D. A. Giuse, R. B. Dannenberg, B. Vander Zanden, D. S. Kosbie, E. Pervin, A. Mickish, and P. Marchal, "Garnet: Comprehensive support for graphical, highly interactive user interfaces," *Computer*, vol. 23, no. 11, pp. 71–85, Nov 1990. doi: 10.1109/2.60882

[170] A. Narechania, A. Srinivasan, and J. Stasko, "NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 369–379, 2021. doi: 10.1109/TVCG.2020.3030378

[171] M. Nebeling, M. Speicher, X. Wang, S. Rajaram, B. D. Hall, Z. Xie, A. R. E. Raistrick, M. Aebersold, E. G. Happ, J. Wang, Y. Sun, L. Zhang, L. E. Ramsier, and R. Kulkarni, "MRAT: The mixed reality analytics toolkit," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2020. doi: 10.1145/3313831.3376330 p. 1–12.

[172] L. K. Nelson, "Computational grounded theory: A methodological framework," *Sociological Methods & Research*, vol. 0, no. 0, p. 0049124117729703, 0. doi: 10.1177/0049124117729703

[173] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2007. doi: 10.1109/CVPR.2007.383132. ISSN 1063-6919 pp. 1–8.

[174] M. Nielsen, N. Elmqvist, and K. Grønbæk, "Scribble query: fluid touch brushing for multivariate data visualization," in *Proceedings of the Australian Conference on Computer-Human Interaction*. New York, NY, USA: ACM, 2016. doi: 10.1145/3010915.3010951 pp. 381–390.

[175] D. A. Norman, "Cognition in the head and in the world: An introduction to the special issue on situated action," *Cognitive Science*, vol. 17, no. 1, pp. 1–6, 1993. doi: 10.1207/s15516709cog1701˙1

[176] ——, "Natural user interfaces are not natural," *Interactions*, vol. 17, no. 3, pp. 6–10, 2010. doi: 10.1145/1744161.1744163

[177] D. A. Norman and S. W. Draper, Eds., *User Centered System Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1986. ISBN 978-0898598728

[178] J. Novotny, J. Tveite, M. L. Turner, S. Gatesy, F. Drury, P. Falkingham, and D. H. Laidlaw, "Developing virtual reality visualizations for unsteady flow analysis of dinosaur track formation using scientific sketching," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 5, pp. 2145–2154, 2019. doi: 10.1109/TVCG.2019.2898796

[179] W. H. O. of Science and T. Policy, "National big data research and development initiative," March 2012.

[180] B. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri, "Face to face: Evaluating visual comparison," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 861–871, 2019. doi: 10.1109/TVCG.2018.2864884

[181] Z. Padgett and E. Davidovits, *Finding the user in data science*. http://datascience.ibm.com/blog/finding-the-user-in-data-science/: IBM Data Science Experience Blog, June 2016.

[182] D. Park, S. M. Drucker, R. Fernandez, and N. Elmqvist, "Atom: A grammar for unit visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 12, pp. 3032–3043, 2018. doi: 10.1109/TVCG.2017.2785807

[183] B. Patnaik, A. Batch, and N. Elmqvist, "Information olfactation: Harnessing scent to convey data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 726–736, 2018. doi: 10.1109/TVCG.2018.2865237

[184] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala, "Video digests: A browsable, skimmable format for informational lecture videos," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, ser. UIST '14. New York, NY, USA: Association for Computing Machinery, 2014. doi: 10.1145/2642918.2647400. ISBN 9781450330695 p. 573–582.

[185] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2019, pp. 7753–7762.

[186] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[187] V. G. Pinto, L. Stanisic, A. Legrand, L. M. Schnorr, S. Thibault, and V. Danjean, "Analyzing dynamic task-based applications on hybrid platforms: An agile scripting approach," in *Proceedings of the Workshop on Visual Performance Analysis*, Nov 2016. doi: 10.1109/VPA.2016.008 pp. 17–24.

[188] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of the International Conference on Intelligence Analysis*, vol. 5, 2005, pp. 2–4.

[189] C. Plaisant and B. Shneiderman, "Scheduling home control devices: design issues and usability evaluation of four touchscreen interfaces," *International Journal of Man-Machine Studies*, vol. 36, no. 3, pp. 375–393, 1992. doi: 10.1016/0020-7373(92)90040-R

[190] R. L. Potter, L. J. Weldon, and B. Shneiderman, "Improving the accuracy of touch screens: an experimental evaluation of three strategies," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1988. doi: 10.1145/57167.57171 pp. 27–32.

[191] X. Pu and M. Kay, "A probabilistic grammar of graphics," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2020. doi: 10.1145/3313831.3376466 pp. 1–13.

[192] D. Pugmire, J. Kress, J. Choi, S. Klasky, T. Kurc, R. M. Churchill, M. Wolf, G. Eisenhower, H. Childs, K. Wu, A. Sim, J. Gu, and J. Low, "Visualization and analysis for near-real-time decision making in distributed workflows," in *Proceedings of the IEEE Parallel and Distributed Processing Symposium Workshops*, May 2016. doi: 10.1109/IPDPSW.2016.175 pp. 1007–1013.

[193] Ramakant, N.-e.-K. Shaik, and L. Veerapalli, "Sign language recognition through fusion of 5DT data glove and camera based information," in *Proceedings of the IEEE International Advance Computing Conference*. Piscataway, NJ, USA: IEEE, 2015. doi: 10.1109/IADCC.2015.7154785 pp. 639–643.

[194] K. Reda, A. E. Johnson, M. E. Papka, and J. Leigh, "Effects of display size and resolution on user behavior and insight acquisition in visual exploration," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2015. doi: 10.1145/2702123.2702406 pp. 2759–2768.

[195] P. Reipschläger, S. Engert, and R. Dachselt, "Augmented displays: Seamlessly extending interactive surfaces with head-mounted augmented reality," in *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2020. doi: 10.1145/3334480.3383138 p. 1–4.

[196] H. T. Reis and C. M. Judd, *Handbook of Research Methods in Social and Personality Psychology*, 2nd ed. Cambridge University Press, 2014. ISBN 978-1107600751

[197] W. Ribarsky, J. Bolter, A. O. den Bosch, and R. Van Teylingen, "Visualization and analysis using virtual reality," *IEEE Computer Graphics and Applications*, vol. 14, no. 1, pp. 10–12, 1994. doi: 10.1109/38.250911

[198] J. C. Roberts, P. D. Ritsos, S. K. Badam, D. Brodbeck, J. Kennedy, and N. Elmqvist, "Visualization beyond the desktop – the next big thing," *IEEE Computer Graphics and Applications*, vol. 34, no. 6, pp. 26–34, Nov. 2014. doi: 10.1109/MCG.2014.82

[199] G. Robertson, M. Czerwinski, K. Larson, Robbins, D. C., D. Thiel, and M. van Dantzich, "Data mountain: Using spatial memory for document management," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 1998. doi: 10.1145/288392.288596 pp. 153–162.

[200] A. C. Robinson and C. Weaver, "Re-visualization: Interactive visualization of the process of visual analysis," in *Workshop on Visualization, Analytics & Spatial Decision Support at the GIScience conference*. Bern, Switzerland: International Cartographic Association, 2006.

[201] D. Russell, M. Stefik, P. Pirolli, and S. Card, "The cost structure of sensemaking," in *Proceedings of the INTERACT and ACM Conferences on Human factors in computing systems*, 1993. doi: 10.1145/169059.169209 pp. 269–276.

[202] A. Saktheeswaran, A. Srinivasan, and J. Stasko, "Touch? Speech? or touch and speech? Investigating multimodal interaction for visual network exploration and analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 6, pp. 2168–2179, 2020. doi: 10.1109/TVCG.2020.2970512

[203] M. Satkowski and R. Dachselt, "Investigating the impact of real-world environments on the perception of 2d visualizations in augmented reality," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 05 2021. doi: 10.1145/3411764.3445330. ISBN 978-1-4503-8096-6/21/05

[204] K. A. Satriadi, B. Ens, M. Cordeil, T. Czauderna, and B. Jenny, "Maps around me: 3d multiview layouts in immersive spaces," *Proceedings of the ACM Conference on Human Factors in Computing Systems*, vol. 4, no. ISS, Nov. 2020. doi: 10.1145/3427329

[205] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, "Vega-Lite: A grammar of interactive graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 341–350, 2017. doi: 10.1109/TVCG.2016.2599030

[206] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer, "Reactive Vega: A streaming dataflow architecture for declarative interactive visualization," *IEEE Transactions*

*on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 659–668, 2016. doi: 10.1109/TVCG.2015.2467091

[207] G. Schall, E. Mendez, E. Kruijff, E. Veas, S. Junghanns, B. Reitinger, and D. Schmalstieg, "Handheld augmented reality for underground infrastructure visualization," *Personal Ubiquitous Computing*, vol. 13, no. 4, p. 281–291, 2009. doi: 10.1007/s00779-008-0204-5

[208] D. Schmalstieg and T. Höllerer, *Augmented Reality: Principles and Practice*. Boston, MA, USA: Addison-Wesley, 2016. ISBN 978-0321883575

[209] D. Schroeder and D. F. Keefe, "Visualization-by-Sketching: An artist's interface for creating multivariate time-varying data visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 877–885, 2016. doi: 10.1109/TVCG.2015.2467153

[210] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 266–281, Jun. 2001. doi: 10.1162/105474601300343603

[211] A. Sears, C. Plaisant, and B. Shneiderman, *A New Era for High Precision Touchscreens*. USA: Ablex Publishing Corp., 1993, p. 1–33. ISBN 0893917516

[212] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012. doi: 10.1109/TVCG.2012.213

[213] L. Shapiro, *Embodied Cognition*. New York, NY, USA: Routledge, 2011. ISBN 978-1138746992

[214] T. B. Sheridan, "Musings on telepresence and virtual presence," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 1, pp. 120–126, 1992. doi: 10.1162/pres.1992.1.1.120

[215] K. Shilton, "This is an intervention: Foregrounding and operationalizing ethics during technology design," in *Emerging Pervasive Information and Communication Technologies: Ethical Challenges, Opportunities and Safeguards*, K. D. Pimple, Ed. Dordrecht, The Netherlands: Springer, May 2014, pp. 177–192.

[216] F. M. Shipman III, C. C. Marshall, and T. P. Moran, "Finding and using implicit structure in human-organized spatial layouts of information," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1995. doi: 10.1145/223904.223949 pp. 346–353.

[217] B. Shneiderman, "Direct manipulation: A step beyond programming languages," *Computer*, vol. 16, no. 8, pp. 57–69, 1983. doi: 10.1109/MC.1983.1654471

[218] ——, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the IEEE Symposium on Visual Languages*. Piscataway, NJ, USA: IEEE, 1996. doi: 10.1016/b978-155860915-0/50046-9 pp. 336–343.

[219] R. Sicat, J. Li, J. Choi, M. Cordeil, W.-K. Jeong, B. Bach, and H. Pfister, "DXR: A toolkit for building immersive data visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 715–725, 2019. doi: 10.1109/TVCG.2018.2865152

[220] N. Silva, T. Schreck, E. Veas, V. Sabol, E. Eggeling, and D. W. Fellner, "Leveraging eye-gaze and time-series features to predict user interests and build a recommendation model for visual analysis," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. New York, NY, USA: ACM, 2018. doi: 10.1145/3204493.3204546 pp. 1–9.

[221] M. Simpson, J. O. Wallgrün, A. Klippel, L. Yang, G. Garner, K. Keller, D. Oprean, and S. Bansal, "Immersive analytics for multi-objective dynamic integrated climate-economy (DICE) models," in *Proceedings of the ACM Companion on Interactive Surfaces and Spaces*. New York, NY, USA: ACM, 2016. doi: 10.1145/3009939.3009955 pp. 99–105.

[222] M. Slater, M. Usoh, and A. Steed, "Depth of presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 3, no. 2, pp. 130–144, Jan. 1994. doi: 10.1162/pres.1994.3.2.130

[223] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. doi: 10.1109/CVPR.2017.590. ISSN 1063-6919 pp. 5563–5572.

[224] J. Staiano, M. Menéndez, A. Battocchi, A. De Angeli, and N. Sebe, "UX_Mate: from facial expressions to UX evaluation," in *Proceedings of the Designing Interactive Systems Conference*. New York, NY, USA: ACM, 2012. doi: 10.1145/2317956.2318068 pp. 741–750.

[225] A. Steed, S. Friston, M. M. López, J. Drummond, Y. Pan, and D. Swapp, "An 'in the wild' experiment on presence and embodiment using consumer virtual reality equipment," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 4, pp. 1406–1414, Apr. 2016. doi: 10.1109/TVCG.2016.2518135

[226] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946. doi: 10.1126/science.103.2684.677

[227] R. Stoakley, M. J. Conway, and R. Pausch, "Virtual reality on a WIM: Interactive worlds in miniature," in *Proceedings of the ACM Conference on Human Factors in*

*Computing Systems*. New York, NY, USA: ACM, 1995. doi: 10.1145/223904.223938. ISBN 0201847051 pp. 265—272.

[228] C. Stolte, D. Tang, and P. Hanrahan, "Polaris: A system for query, analysis, and visualization of multidimensional relational databases," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 52–65, 2002. doi: 10.1109/INFVIS.2000.885086

[229] P. Suja, K. V. P. Kumar, and S. Tripathi, "Dynamic facial emotion recognition from 4D video sequences," in *Proceedings of the International Conference on Contemporary Computing*. Piscataway, NJ, USA: IEEE, 2015. doi: 10.1109/IC3.2015.7346705 pp. 348–353.

[230] C. T. Tan, S. Bakkes, and Y. Pisan, "Inferring player experiences using facial expressions analysis," in *Proceedings of the 2014 Conference on Interactive Entertainment*. New York, NY, USA: ACM, 2014. doi: 10.1145/2677758.2677765 pp. 1–8.

[231] D. S. Tan, D. Gergle, P. G. Scupelli, and R. Pausch, "With similar visual angles, larger displays improve performance on spatial tasks," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2003. doi: 10.1145/642611.642650 pp. 217–224.

[232] D. Team, "Datavyu: A video coding tool," *Databrary Project, New York University*, 2014. [Online]. Available: http://www.datavyu.org/

[233] B. H. Thomas, G. F. Welch, P. Dragicevic, N. Elmqvist, P. Irani, Y. Jansen, D. Schmalstieg, A. Tabard, N. A. M. ElSayed, R. T. Smith, and W. Willett, "Situated analytics," in *Immersive Analytics*, ser. Lecture Notes in Computer Science, vol. 11190. New York, NY, USA: Springer, 2018. doi: 10.1007/978-3-030-01388-2˙7 pp. 185–220.

[234] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics*, vol. 33, no. 5, pp. 169:1–169:10, 2014. doi: 10.1145/2629500

[235] M. Tory, M. Atkins, A. Kirkpatrick, M. Nicolaou, and G.-Z. Yang, "Eyegaze analysis of displays with combined 2D and 3D views," in *Proceedings of the IEEE Conference on Information Visualization*. Piscataway, NJ, USA: IEEE, 01 2005. doi: 10.1109/VISUAL.2005.1532837. ISBN 0-7803-9462-3 pp. 519–526.

[236] A. Truong, F. Berthouzoz, W. Li, and M. Agrawala, "Quickcut: An interactive tool for editing narrated video," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, ser. UIST '16. New York, NY, USA: Association for Computing Machinery, 2016. doi: 10.1145/2984511.2984569. ISBN 9781450341899 p. 497–507.

[237] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977. ISBN 978-0201076165

[238] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, "Evaluation of a context-aware voice interface for ambient assisted living: Qualitative user study vs. quantitative system evaluation," *ACM Transactions in Accessible Computing*, vol. 7, no. 2, May 2015. doi: 10.1145/2738047

[239] A. van Dam, D. H. Laidlaw, and R. M. Simpson, "Experiments in immersive Virtual Reality for scientific visualization," *Computers & Graphics*, vol. 26, no. 4, pp. 535–555, 2002. doi: 10.1016/S0097-8493(02)00113-9

[240] J. van de Wolfshaar, M. F. Karaaba, and M. A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *Proceedings of the IEEE Symposium Series on Computational Intelligence*. Piscataway, NJ, USA: IEEE, 2015. doi: 10.1109/SSCI.2015.37 pp. 188–195.

[241] J. Vasconcelos-Raposo, M. Bessa, M. Melo, L. Barbosa, R. Rodrigues, C. M. Teixeira, L. Cabral, and A. A. Sousa, "Adaptation and validation of the Igroup presence questionnaire (IPQ) in a Portuguese sample," *Presence: Teleoperators and Virtual Environments*, vol. 25, no. 3, pp. 191–203, 2016. doi: 10.1162/PRES˙a˙00261

[242] J. Wagner, W. Stuerzlinger, and L. Nedel, "Comparing and combining virtual hand and virtual ray pointer interactions for data manipulation in immersive analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2513–2523, 2021. doi: 10.1109/TVCG.2021.3067759

[243] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3352–3361.

[244] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[245] L. Weingart, M. Olekalns, and P. Smith, "Quantitative coding of negotiation behavior," *International Negotiation*, vol. 9, no. 3, pp. 441–456, 2004. doi: 10.1163/1571806053498805

[246] M. Weiser, "The computer for the 21st Century," *Scientific American*, vol. 265, no. 3, pp. 94–104, Sep 1991. doi: 10.1145/329124.329126

[247] ——, "Hot topics-ubiquitous computing," *Computer*, vol. 26, no. 10, pp. 71–72, 1993. doi: 10.1109/2.237456

[248] S. White and S. Feiner, "Sitelens: Situated visualization techniques for urban site visits," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2009. doi: 10.1145/1518701.1518871 p. 1117–1120.

[249] M. Whitlock, E. Harnner, J. R. Brubaker, S. Kane, and D. A. Szafir, "Interacting with distant objects in augmented reality," in *Proceedings of the ACM IEEE Symposium on 3D User Interfaces*, 2018. doi: 10.1109/VR.2018.8446381 pp. 41–48.

[250] M. Whitlock, S. Smart, and D. A. Szafir, "Graphical perception for immersive analytics," in *Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces*. Piscataway, NJ, USA: IEEE, 2020. doi: 10.1109/VR46266.2020.00084 pp. 616–625.

[251] M. Whitlock, K. Wu, and D. A. Szafir, "Designing for mobile and immersive visual analytics in the field," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 503–513, 2020. doi: 10.1109/TVCG.2019.2934282

[252] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2009. ISBN 978-0-387-98141-3

[253] ——, "A layered grammar of graphics," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 3–28, 2010. doi: 10.1198/jcgs.2009.07098

[254] L. Wilkinson, *The Grammar of Graphics*. Berlin, Heidelberg: Springer-Verlag, 1999. ISBN 0387987746

[255] W. Willett, B. A. Aseniero, S. Carpendale, P. Dragicevic, Y. Jansen, L. Oehlberg, and P. Isenberg, "Perception! Immersion! Empowerment! Superpowers as inspiration for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 22–32, 2022. doi: 10.1109/TVCG.2021.3114844

[256] W. Willett, Y. Jansen, and P. Dragicevic, "Embedded Data Representations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 461–470, Jan. 2017. doi: 10.1109/TVCG.2016.2598608

[257] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 3, pp. 225–240, 1998. doi: 10.1162/105474698565686

[258] D. Wixon, K. Holtzblatt, and S. Knox, "Contextual design: An emergent view of system design," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1990. doi: 10.1145/97243.97304 pp. 329–336.

[259] Y. Wu, J. M. Hellerstein, and A. Satyanarayan, "B2: Bridging code and interactive visualization in computational notebooks," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, ser. UIST '20. New York, NY, USA: Association for Computing Machinery, 2020. doi: 10.1145/3379337.3415851. ISBN 9781450375146 pp. 152—-165.

[260] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, 2007. doi: 10.1109/TVCG.2007.70515

[261] J. Yi, X. Mao, L. Chen, Y. Xue, and A. Compare, "Facial expression recognition considering individual differences in facial structure and texture," *IET Computer Vision*, vol. 8, no. 5, pp. 429–440, 2014. doi: 10.1049/iet-cvi.2013.0171

[262] B. Yu and C. T. Silva, "FlowSense: A natural language interface for visual data exploration within a dataflow system," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1–11, 2020. doi: 10.1109/TVCG.2019.2934668

[263] R. C. Zeleznik, K. P. Herndon, and J. F. Hughes, "SKETCH: An interface for sketching 3D scenes," in *Proceedings of ACM Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: ACM, 1996. doi: 10.1145/237170.237238 p. 163–170.

[264] H. Zhang, Y. Hou, D. Qu, and Q. Liu, "Correlation visualization of time-varying patterns for multi-variable data," *IEEE Access*, vol. 4, pp. 4669–4677, 2016. doi: 10.1109/ACCESS.2016.2601339

[265] X. Zhang, H.-F. Brown, and A. Shankar, "Data-driven personas: Constructing archetypal users with clickstreams and user telemetry," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2016. doi: 10.1145/2858036.2858523 pp. 5350–5359.

[266] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, Jun. 2015. doi: 10.1109/CVPR.2015.7299081 pp. 4511–4520.

[267] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/CVPR.2016.374 pp. 3438–3446.

[268] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan, "Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 261–270, 2017. doi: 10.1109/TVCG.2016.2598543

[269] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan, "Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 340–350, 2017. doi: 10.1109/TVCG.2017.2745279

[270] M. Zheng and A. G. Campbell, "Location-based augmented reality in-situ visualization applied for agricultural fieldwork navigation," in *Adjunct Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. Piscataway, NJ, USA: IEEE, 2019. doi: 10.1109/ISMAR-Adjunct.2019.00039 pp. 93–97.

[271] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016. doi: 10.1109/CVPR.2016.537 pp. 4966–4975.

[272] S. Zollmann, R. Grasset, T. Langlotz, W. H. Lo, S. Mori, and H. Regenbrecht, "Visualization techniques in augmented reality: A taxonomy, methods and patterns," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 9, pp. 3808–3825, 2021. doi: 10.1109/TVCG.2020.2986247

[273] J. Zong, D. Barnwal, R. Neogy, and A. Satyanarayan, "Lyra 2: Designing interactive visualizations by demonstration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 304–314, 2021. doi: 10.1109/TVCG.2020.3030367