How Our Intentions are Seen: Effective Communication in Data Storytelling

By

KESHAV DASU
B.S. University of California Riverside 2015

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Dr. Kwan-Liu Ma , Chair

Dr. Lace Padilla

Dr. Josh McCoy

Committee in Charge

2023

To my loyal companion Jessie . . .

my unacknowledged co-author.

CONTENTS

ABSTRACT

How Our Intentions are Seen: Effective Communication in Data Storytelling

Data visualization can be defined as the visual communication of information. As data visualization designers or storytellers, we go through many steps in order to communicate the intended insights in data to an audience. We begin with the data and some goal or intent. We process this data using a variety of methods to discover the relationships contained in the data. Often we find that this analysis by itself, is not enough for us to realize our intentions or our goals. To get closer to our intentions, we have to visualize the data and the underlying relationships. Through the pairing of analysis, interaction, and visualization, we are able to manipulate data into a language imbued with meanings more than what is literally presented. Visualizations impact our perceptions of the data and the underlying insights or narratives. We are able to see causal relationships, uncertainty, and evidence to support or disprove hypotheses.

The complexity of the conversations brought out through data visualizations has steadily increased. Infographics and interactive visualizations are now commonly employed for conveying information to the general public (e.g., as seen in NYT, Reuters, The Pudding, etc.). One important barometer for the success of visualization is whether the intents of the communicator(s) were faithfully conveyed. This intention can be of many forms such as to persuade, to educate, to inform, or even to entertain. A second important factor is that the visualization must be consumable and enjoyable. Finally, it must preserve scientific integrity.

The processes of constructing and displaying visualizations have been widely studied by our community. However, because of the lack of consistency in this literature, there is a growing acknowledgment of a need for frameworks and methodologies for classifying and formalizing the communicative component of visualization. The widespread usage and increasing complexity of data visualizations has increased the importance of delving deeper into this domain of research. In my dissertation research, which consists of four components, I seek to leverage lessons from storytelling and linguistics to offer ideas on how the visualization community and designers can be more precise, consistent, engaging, and impactful in their visual communication of information, particularly when communicating to the public.

An important consideration when designing a visualization tool for non-experts to understand and explore complex data is aesthetics. Visual metaphors provide desirable aesthetics to viewers by employing familiar representations. In the first component, I review several data stories of mine that utilized different visual metaphors. I present different techniques for producing visual metaphors and discuss their strengths and limitations. In the second component of this research, I present a retrospective analysis of a complex visualization that we developed for a science museum. We identified a set of considerations that must be taken into account while designing narrative visualization that seeks to convey complex scientific information in an informal learning environment such as the museum. This work also identified directions for further research into storytelling structures as they apply to data visualization. The third component is concerned with data characters. In data stories, numerous methods have been identified for constructing and presenting a plot. I posit that there is an opportunity to expand how we think and create the visual elements that present the story. Stories are brought to life by characters; often they are what make a story captivating, enjoyable, memorable, and facilitate following the plot till the end. In an effort to improve how data stories can be produced and be more accessible to larger audiences, I offer guidance on the design of data stories from a character-oriented perspective. Finally, I propose to deepen the understanding of how our intentions as designers affect the visualizations we create by translating linguistic practices and techniques into communicative visualization. I then seek to take these methods and concepts and apply them to a system for automated storytelling. This work enables the integration of theories and frameworks from linguistics and storytelling into visualization and grows our theoretical toolset for studying visualization. Additionally, it illustrates how to apply these introduced frameworks when creating data visualizations with communicative intent. The frameworks and design solutions presented in this dissertation are expected to be applicable to data storytellers or visualization designers who seek to be more effective in communicating their intended messages to an audience.

Meanings encapsulated within a visualization are no longer always straightforward and direct, or unambiguous.

ACKNOWLEDGMENTS

This dissertation is dedicated to those who supported me through the difficult journey that is becoming an independent researcher. There are many times during this voyage I would find myself lost and doubting my ability or choices, and in those moments certain individuals would be there to help me rekindle my love of research, guide me back on the path when I would stray, and remind me of the joy research has to offer when you share it with others.

First, I would like to thank my research advisor, Professor Kwan-Liu Ma who consistently supported me throughout all my research endeavors. There were many times others may have dissuaded me from the research path I chose, however, he was never one of them. I owe a great portion of this journey to him, as I was able to explore so many different research domains outside of data visualization and work on translating concepts and frameworks back into this field we love. With him as my advisor, I was motivated to keep pushing through the obstacles and challenges presented to me and through him, I was afforded many exciting opportunities and collaborations to grow as a researcher.

I want to thank all my dissertation proposal and dissertation committee members, Professor. Michael Neff, Professor. Lace Padilla, Professor. Josh McCoy, and Professor. Hao-Chuan Wang. Many of the chapters and content in this dissertation have been greatly influenced by their feedback and comments.

I want to thank my dad, Professor Sriram Dasu, for being patient with me, giving me a goal to chase, and inspiring me to even attempt to be a researcher. I thank my mom, Rama Dasu, for everything I don't think there are enough words to enumerate how much you have done and continue to do for me. I thank my friends and colleagues at UC Davis & VIDI for being there with me and for all the memories we shared. Thank you Senthil Chandrasegaran for all the coffee conversations, motivation, and advice. Thank you, Yun-Hsin Kuo, for entertaining my ideas and helping me take them to completion. Special thanks to John Chan, Meghann Ma, David Bauer, Mike Brevard, Spencer Russell Botticelli-Smith, and Vivek Pandrangi.

Chapter 1
Introduction

The modern world works with large volumes of information that require interpretation and presentation in a wide range of settings. Frequently, the same information is presented to individuals that come from different domains and have different backgrounds and goals. Information can often be very abstract and intangible, which can lead to difficulties in communication. Data visualization makes use of visual representations of complex and often large datasets to reveal hidden insights about both known and unknown phenomena, showcase findings, and share insights with broad audiences. The potential for data visualization is vast. Visualization leverage benefits from multiple domains, including storytelling [1, 2, 3], semiotics [4], linguistics [5], and design [6]. It pairs multiple forms of communication and learning such as tactile, written, audio, and visual, and uses all of these to explore new findings and communicate information to broader audiences. As a result, visualizations are widely employed in industry and educational institutions. We also see visualizations in news, social media, personal applications, and entertainment. A wide majority of news outlets have their own personal data visualization groups and most of their articles contain some sort of interactive visualization. The beauty of visualization lies in its ability to make the intangible tangible, the invisible visible, and the inaccessible accessible.

As society became more comfortable with and used to "simple" visualizations and learned how to interface with them, the overall visual literacy of society increased. People naturally develop increased familiarity with interpreting and understanding visualization through repeated use and exposure. This increased literacy and ubiquity of data visualization have led to more nuances and complexity in how meaning is conveyed. This can be seen in the recent uses of network, sankey, and even some bespoke visualizations from news media (e.g., NYT [7]). These visualizations are layering text, audio, graphics, and touch to communicate their messages. Designers will continue to adapt to their audience's increased ability and are likely to increase the complexity of visualizations.

Today, the role of visualization is no different than that of language; we use language to express what is being observed to us. The goal of visualization, like ordinary speech, often goes beyond presenting facts to achieve actions or outcomes. In any case, once data is visualized it can alter how the original context is perceived, (e.g., visualizing the uncertainty in data [8, 9]). There is an ongoing discussion on design as communication [10, 11, 12, 13] and there is a body of work [10, 12, 13] that gives credence to viewing visual design as communication.

As our field of communicative visualization grows and becomes a mainstay of society for presenting and sharing data, it becomes increasingly important for us to have frameworks and tools to understand this evolving language.

1.1 Challenges in Communication

The potential for communication created by the versatility of data visualization techniques creates many challenges for this field. It inherits many of the communication challenges and limitations of traditional languages. Communication challenges for data visualization have been well documented by our community. We need infrastructure to assess and understand the relationships among the visual elements and the meaning they convey to different audiences. For one, we need to understand different intentions and how these intentions are conveyed by the design. There is also an opportunity to draw from other fields that address communication.

Treating visualization as a language has been considered, although exploring the value of this association and what it affords is limited. Purchase et al. [14] have explicitly made these connections, and briefly describe the use of linguistic theory, namely pragmatics, as a means to provide an over-arching framework for information visualization. They comment on the relationship between visualization and language and discuss how information visualization should rely on a multitude of theories rather than a singular theory. Hullman and Diakopoulos [15] study the application of linguistic-based rhetoric in narrative visualizations. Several others have presented theoretical visualization frameworks [16, 17, 18, 19] and implicitly imply that visualization is a language. They elegantly demonstrate how applying frameworks from spaces such as distributed cognition, information theory, an algebraic basis, or conceptual metaphor theory can contribute to the betterment and improved understanding of the use of visualization.

A vocabulary is the body of words used in a language. If we are to claim visualization is a language then its vocabulary would be visual encodings. Grammar-based visual encodings, as well as declarative languages [20, 21, 22, 23, 24, 25], arose out of a need to fluidly and precisely articulate a set of intents for communicating data. These works provide formalized approaches for describing tables, charts, graphs, maps, and tables and give credence to treating visualization as a language.

Although significant progress has been made in developing frameworks, much work remains to be done. The accessibility of complex information to wide audiences is an ongoing challenge. Some of the basic challenges and needs to address in communicative visualization are: Accessibility and the audience, keeping track and the consistency of the message, and the accuracy between the intended and interpreted meanings of our visualizations. It is important to find methods and factors to bridge misunderstandings or misconceptions into insights and gain knowledge. For many non-domain experts or the general public, it is difficult to interpret certain data and build their own useful insights. Therefore, we need to provide proper methods to share these data with the public. We need to further our understanding of both what devices are available and how can be attuned for a communicative task as well as learn about how settings and external factors influence our visualization design.

Data visualization is a vast and growing field. In this thesis, I am mainly concerned with communicative visualization and in particular data storytelling. Communicative visualization represents the majority of the visualizations to which an average person is exposed. Data storytelling [2, 26, 1, 27, 28, 15], as the name suggests, leverages storytelling techniques in combination with data, images, and interactions.

Additionally, within data storytelling, in our work, we are interested in data-driven, visual storytelling, particularly the characters that bring them to life. Data storytellers want to create rich experiences that evoke an emotional response, draw the audience in, and leave them with something to remember. Stories are brought to life by characters; often they are what make a story captivating, enjoyable, memorable, and facilitate following the plot till the end. However, within stories keeping track of all our story points and their relations can become arduous as we layer more and more information. It, therefore, becomes important to put more consideration into how we design our visual narrative elements and characters. That is we must define what a data character is, how we design characters in data storytelling, and how this design offers consistency to the author when designing the story as well as the audience when consuming the story.

1.2 Overview of Contributions

In this dissertation, I look to contribute the following. In Chapter 2, I begin by reviewing what has been achieved in the landscape of data storytelling and narrative visualization. How have current works addressed the needs of the audience and authors, what are the available devices and techniques for communicating information, and what is the value of such efforts?

In Chapter 3 I examine the application of such efforts through three data stories that were accepted in the Pacific VIS data storytelling contest. Each of those stories applies different techniques, methods, datasets, and goals to satisfy a communicative goal. One of those contest entries is unpacked further as it pertains to visual metaphors. We look at the value visual metaphors afford us in communicating and a method for communicating co-occurrence to the general public.

The subsequent chapter, Chapter 4, focuses on the role audience and the setting have and the lessons learned from developing an exhibit for the public. This work is a reflection on the development of an interactive data visualization for museum visitors. It demonstrates how the museum context can affect the presentation of a story and identifies additional considerations that should be factored in. This work raises some questions about current approaches for developing stories and suggests new research directions. One direction is the role of a through-line and characters in data storytelling.

Within Chapter 5, I describe how to frame data using lessons from visual and written storytelling in terms of characters and offer the following; an analysis of the form and properties of characters in 160 data stories, a framework for data characters, and a design space for developing characters named *character-oriented design*. A key concept is that all visual entities shown to an audience must have some narrative goal or intent.

Chapter 6 focuses on designer intent. I offer a conceptual framework named VisActs, describing intent in communicative visualization. This work translates frameworks and techniques from linguistics, specifically from speech act theory, and contextualizes them for a data visualization community. By understanding intentions, how they can be expressed, and giving examples of their nuances our communication of information can be more precise. Additionally, this work illustrates how by understanding how designer intent maps to visible design choices we can potentially develop more sophisticated ML models for generating interactive visualization interfaces based on user prompts.

In Chapter 7 I develop a system named NOVA, News Outlet Visual Assessment, for assessing user beliefs of mainstream media outlets. This system leverages some of the concepts from Chapters 5 and 6 as we develop a central character to reflect the audience and their beliefs as they move through the system. We conduct a user study to demonstrate the potential of visualizing one's beliefs and contrasting that against the data. We find some evidence that this approach can be effective in scaffolding the public to assess if their beliefs align with what the data shows.

Lastly, I conclude this dissertation by summarizing my contributions and thoughts on the use of interactive data-driven visualization for effective communication.

Chapter 2
Visual Data Storytelling

In this section, we walk through key storytelling terms and contextualize them for data storytelling. Storytelling structure and paradigms [29, 30, 31] have been ever present and are constantly evolving with new mediums and formats. However, we are interested in data stories and determining what applies to storytelling in general and how it can be translated and used for our purposes as *data* storytellers. Data stories can be presented in a variety of ways, such as infographics, comics, videos, virtual experiences, and so on.

People have shared and told stories for ages [32] and consequently have postulated the rules and structures for effective storytelling. When considering where to focus on influences for data storytelling, we draw on elements from both written [29, 30] and visual media [33]. In *Aspects of the Novel* by E. M. Forster [29], he analyzed the common aspects that all English-language novels share: *story*, *people*, *plot*, *fantasy*, *prophecy*, *pattern*, and *rhythm*. He describes a **story** as "a narrative of events arranged in their time sequence" and the **plot** as "also a narrative of events, with the emphasis falling on causality". Naturally, we look to see how these principles translate into data stories. With written and visual media, authors typically have more freedom and flexibility when creating their stories.

Data stories, however, tend to have less flexibility in that they often are constrained by "non-alterable non-fiction" [34]. Still, from existing literature [34, 35], we can surmise that common themes for both storytelling and data storytelling are: *characters*, *plot*, *theme*, *setting*, and *conflict*. The goal of most **data stories** is to reach a wide or targeted audience through the

Storytelling Terminology	Description	Contextualized for Data Storytelling (DS)
Story or Narrative	Consists of many subsystems working together (e.g., characters, plot, and theme). It is an account of events arranged in their time sequence [29].	A combination of visualized findings or messages with connections such as temporal or causal relations [34, 36].
Theme	A recurring idea. A story can have many themes.	These appear as concepts paired with an intention. (i.e., to inform, persuade, entertain, comfort, explain, or terrorise [26]).
Through-Line	A single theme that runs from the start to the end of a story. It interweaves the roles of the characters with the plot.	One primary concept and intention that drives the story and motivates the characters' actions [26].
Plot	A combination of events and how those events are revealed. The arrangement (or the sequence) of events in the story.	A causal relationship [37] between a set of events, depicted by the behaviors and interactions [38] of data-driven visual elements.
Event or Story Piece	Atomic element of a story. It may focus on one character's status or behavior, or the relationships among multiple characters.	"Story pieces" [37] that are derived from data, provided by data analysts and domain experts, e.g., data facts and human insights.
Genre	Classification and organization of works into categories.	Magazine, annotated chart [39, 40, 41], partitioned poster, flow chart, comic strip [28], slide show [42], and film/video/animation [27, 43, 44]
Setting	A cluster of actual states of affairs or various events where the story takes place.	The devices and location where the story is presented [37] and the various scientific domains [38] where the insights were derived from.
Storytelling	To both share and provide an experience to an audience. To give the audience a form of knowledge [30, 33] that is both emotional and entertaining.	This is the same, with an additional condition that the story must be based on the data.
Audience	The targeted group of people who will receive the story	This is the same in data storytelling.

Table 2.1: Terminology with definitions contextualized for data storytelling. These terms and their mappings were derived from a breadth of visual and written literature [30, 33, 45, 29], corroborated with the narrative and data storytelling literature [1, 38, 46, 37, 34, 26].

presentation of visualized findings or messages. The role of a storyteller [30] is to summarize all the narrative events such that the audience perceives this as a self-contained story.

Within visualization, there is a large body of work that utilizes storytelling. This chapter offers background on traditional written and visual storytelling and compares it with the current literature for narrative visualization and data storytelling. Much of the work in this manuscript leverages concepts from storytelling from both the traditional and the visualization communities. This chapter provides a table of storytelling terminology and definitions based on literature that is mapped into data storytelling as well.

2.0.1 Terminology: Elements of a Story

The terminology for data storytelling as it relates to storytelling has some differences. We define a set of storytelling terms for consistency and clarity in Table 2.1. However, even within data stories, there are many varying explanations [38, 46, 37, 35, 34] of what *through-line, plot*, and *character* mean. Therefore, we provide a deeper explanation of those in this subsection.

Through-Line. When we look at the storytelling process [30, 47, 33] and data storytelling process [37], both require ascertaining what the author wants to offer the audience. As authors of data stories, we want the audience to feel something [34] as they consume our story and leave with something [26] when they are finished. It is important for us to identify the core message that we want the audience to come away with. For example, a data story could seek to persuade the audience to take action against climate change by communicating the impact microplastics have on biodiversity. Alternatively, the story could intend to explain the flaws of a misconception, all types of plastic have the same effect on the environment. In both scenarios, the authors may communicate the life cycle of microplastics to the audience, while their intention influences the narrative.

This message is often referred to as the theme of a story, or a recurring idea. A story can have many themes (e.g., biodiversity) but a through-line [33] is the *theme* (e.g., the impact of microplastics) that runs from the start to the end of a story. It is what helps keep the story on track and helps drive it from start to finish. A through-line interweaves the roles of the characters with the plot. It helps ensure consistency and continuity for the entire story.

Plot. A plot is a description of a set of events with a purpose [30]. Each event that transpires is causally connected and each event is essential to the overall story. As Poe [48] finds with his theory of "unity of effect", every element of a story should help create a single emotional impact. Namely, every element in our story must tie in with the through-line.

Character. A character is an entity that influences either itself or others. A story typically tracks what an entity wants [31], what the entity will do to get it, and what cost the entity will have to pay along the way. What pulls the character along is "desire". A character will take action in pursuit of a desire while learning new information about this desire. This new information causes a change in the course of actions and hence influences the story. A character in pursuit

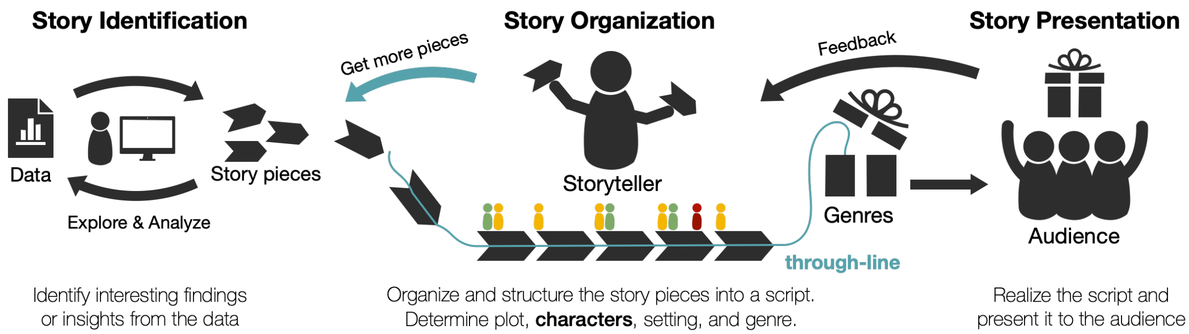


Figure 2.1: The role of a data story author can be partitioned into three stages: story identification, organization, and presentation. In the identification stage, the key task is to acquire the information and story pieces of what we intend to convey. A key task in story organization is to determine what **main**, **supporting**, **antagonist** characters, etc. are and their relations to the plot. Namely, what visual elements or elements unify our story pieces and how are these pieces related? Lastly, the presentation stage focuses on the implementation details of communicating the story (i.e., transitions, annotations, devices, etc.)

of a desire always encounters a struggle, which may cause a change in the character itself. The task of the storyteller is to present a change in a character or illustrate why that change did not occur. The task of this work is to show how this same process can be applied to data stories.

2.1 Storytelling in Visualization

Within visualization, there is a large body of work [37, 2] that utilizes and documents storytelling with data. Data storytelling and narrative visualization are two notable branches. Recently, the literature contained in these branches was further organized into three groups [2]. The works are sorted based on whether they (1) *address who are the main subjects involved* (authoring tools and audience), (2) *assess how stories are told* (narratives and transitions), or (3) consider why storytelling is effective for visualization (memorability and interpretation). This work seeks to contribute to *how stories are told* by systematically investigating and identifying distinguishable features of characters in data stories. The general storytelling process [30, 33] considers both the plot and the characters. We find little has been done to address the design of data characters, as there is a current focus on creating and conveying the data story plot. In this section, we review the notion of a data character and how it has been previously considered.

2.1.1 Data Storytelling Process

To better see the relationship between the storyteller and the data storyteller, we organize data storytelling literature into three stages, as shown in Figure 2.1, based on the process of the data storyteller [37] — identification, organization, and presentation of the story. At the beginning of developing a story, the data storyteller may not know what to discuss or share yet. A set of events (or story pieces [37]), such as data facts or human insights, should be identified beforehand so that the data storyteller can figure out what to share.

Merely having a set of story pieces to report does not give a story. Through the relationships among story pieces (e.g., causality), the data storyteller can organize the story pieces into a story plot for presentation. However, if the storyteller is unclear on the relationships or why the subject matter is worth presenting, the audience may feel similarly when consuming the story. The organization stage is where data storytellers select story pieces, sequence them into a story plot, and leave with a structured outline for the story. Several frameworks offer systematic ways to sequence a story plot for various genres [26, 1, 27, 28, 15]. Here, we review the relevant literature on story plots from two aspects: the narrative structure and the communicative goal.

There is a popular set of three narrative structures [1] for data storytelling: the drill-down, the martini glass, and the interactive slideshow, which have been heavily discussed in this space. Recently, Yang et al. [27] provides guidelines for applying Freytag's Pyramid, a narrative structure that has been widely used in film and literature, to data stories. These structures are similar to what Truby describes as the*story movement* in the space of written storytelling [30]. Truby describes a set of narrative structure patterns that storytellers draw from nature to connect elements in a sequence (e.g., linear, meanderings, spiral, and branching). For data stories, the narrative structure addresses the story movement, yet the pace of this movement is often influenced by the audience's interactions [1]. For example, the drill-down structure can present multiple plot branches and even different stories together, thus enabling the audience to explore the story. Meanwhile, the alternatives often present a single linear narrative, which is easier for the author to maintain the plot consistency, and therefore has been the focus of automated reasoning [49, 50]. In terms of the communicative goal, Adegboyega and Heravi [26] identify seven types of story plots from data stories. Their work provides a great jumping-off point for story creation. In section 5.4, we expand on their work and describe generically how data characters may weave within these seven story plots.

It is important to note that the genre [1] and the setting [37] of a story should also be settled at this organization stage. Determining the genre (or genres) of the story can help data storytellers narrow down the choices of the ideal story plot and narrative structure, aligning with their intentions better. Segel and Heer [1] identify a set of genres for data stories: magazine, annotated chart [39, 40, 41], partitioned poster, flow chart, comic strip [28], slide show [42], and film/video/animation [27, 43, 44, 51]. Akin to other storytelling fields, these genres can be paired with each other. Many of these genres have been further developed, and methodologies and frameworks for using them have been offered. Refer to Tong et al. [2] for more details.

During the organization stage, the data storytellers may begin assessing which visual elements will best achieve their communicative goals and support advancing the story. Understanding the relevancy of each visual element to the story theme often can help the data storytellers locate important story pieces [15]. Irrelevant story pieces in a story can become multiple story sub-plots. Without proper curation, the story likely encounters issues at the presentation stage, such as the audience struggling with comprehending the story and losing interest [52, 53]. This motivates our work to investigate how visual elements can be modeled as data characters. The characterization of story pieces with data characters may help reinforce the through-line and construct the story plot, supporting the data storytellers in delivering a clear message to the audience with their stories.

2.1.2 Characters in Data Stories

Prior research has addressed character-based [54, 55] or character-driven storytelling [56, 57]. Cai et al. [56] address the importance of character- and plot-driven storytelling and offer a hybrid system involving both approaches. Cavazza et al. [57, 54, 55] focus on character-driven stories and propose an engine that models character behaviors. Their work has a clear focus on automatic narrative generation, where characters in all of these works are presented as either human or virtual agents. These works focus on improving stories that either are rooted in traditional storytelling (i.e., not data-driven and allow for more creative freedoms of the author) or express the story content using virtual humanoid agents. However, data stories often are not expressed using human or virtual agents but rather through abstract representations. Data stories are also often rooted in non-alterable non-fiction, constraining the storytelling process. Our work contends that we should view and think of abstract visual representations as data characters, similar to how prior research applies techniques to virtual agents.

To provide recent examples of what we consider an effective data character, we look to the popular storyline visualization [58, 59, 60]. This visualization depicts an abstract line encoding that represents the progress of an entity during a temporal period. We can view this encoding as a character, drawing the attention of an audience. Furthermore, it can be used as a through-line and an entry point into understanding the change and growth of a single data-driven entity. Storyline visualizations contain multiple storylines, thus multiple characters, illustrating their relationships. Other examples include visualizations with properties that make them more identifiable or personable, such as anthropographics [61, 62]. These representations were shown to be effective in creating an emotional connection with underlying content [63, 62], while other evidence revealed their limitations of eliciting a specific emotion from an audience [64]. These works motivate a need for the data storytelling community to translate devices and structures from visual storytelling (i.e., film literature). We feel data characters may serve as a stepping stone to developing and eliciting an emotional connection with the audience, yet to be investigated in future work.

A data character can be the visual elements that are "performing" the data story [38]. There are some direct references [38, 46, 65] on the role a data character assumes; however, a story can not exist without characters. When designing data characters, we are looking for visual entities that relate to the theme of the story and advance the story plot. A data character could naturally take on rules of visual encoding, the well-established combination of marks and channels with their respective mappings to data. A data character should consider properties that delve deeper into the communicative effort and that help the audience bridge the gap between science and the story. Thus, a foundation for the properties of a data character can be derived from works that address visualization design [46, 61, 62, 66, 65] and visual metaphors [4, 67, 6, 68, 69, 70, 71]. There is limited research that addresses the role of a data character in data storytelling. In Chapter 5, we provide a framework to classify data characters and their behaviors.

Chapter 3
Visual Metaphors for Data Storytelling

This chapter includes content from: **Dasu, K.**, Fujiwara, T., & Ma, K. L. (2018, October). An organic visual metaphor for public understanding of conditional co-occurrences. In 2018 IEEE Scientific Visualization Conference (SciVis) (pp. 1–5). IEEE.

3.1 Introduction

In the prior chapter, I presented an overview of data storytelling processes and storytelling terminology. In this chapter, I apply storytelling techniques to convey scientific content that is often complex and unfamiliar to the general public. I present three of my data visualizations, where each one deals with a topic of varying familiarity to the public and applies different storytelling techniques. This chapter concludes with explanations of the design and methodology of a novel visual metaphor for explaining co-occurrence data to the lay audience.

3.2 Data Stories

I developed three data stories to better understand the application of storytelling techniques in data visualizations. The development process of each story revealed several insights and lessons that were influential in the development of subsequent stories. I begin with a story about the United States' 2016 presidential election.

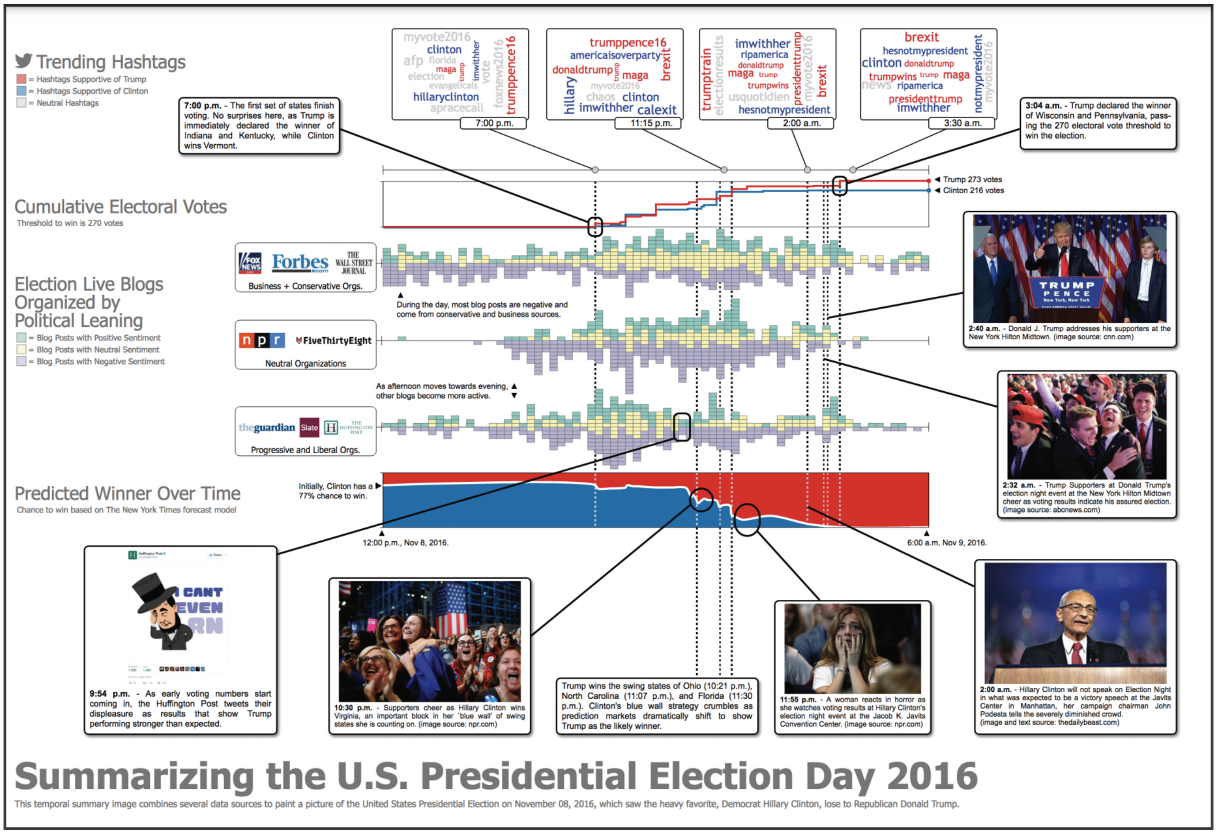


Figure 3.1: Infographic summarizing the 2016 United States presidential election. This work was presented in Seoul, Korea as part of the PacificVis 2017 storytelling contest, and those abroad were able to successfully understand the emotional swing various political blogs had as the election night progressed.

3.2.1 Summarizing the U.S. Presidential Election Day 2016

This story's plot is fairly accessible as the pre-requisite domain knowledge is not complex. Namely, the content does not require the user to have deep existing prior knowledge on the subject, in this case, American politics, in order to understand the story and follow along. A focus of this infographic was the portrayal of the sentiment of several political blogs about the election as it progressed throughout the day. The creation of this data story was facilitated by using Temporal Summary Images [72] (TSIs). We used this framework, which combines small multiples with linked time-dependent visualizations and overlaid annotations, to build a data-driven infographic of the U.S. presidential election day—November 08, 2016. The result uses several online data sources: hashtags extracted from over twelve million tweets, electoral college results, predicted winner forecasting and blog posts scraped from the election live blogs of several news organizations. Blog sources are grouped by political bias and sentiment analysis has been applied to their posts. Annotations provide additional context about the data streams and highlight important temporal events throughout the night. By combining automatic data generation and layout with manual refinement and visual design, TSIs allowed for a streamlined design in producing presentation-quality graphics.

This data story was developed within the genre of an annotated chart, as described in Chapter 2. The narrative experience for the audience was completely driven by our goals as authors; an author-driven story. That is, the audience for this story has no control or influence in controlling what type of content or experience they can take away from this data story. Some drawbacks of this approach are that it is static, and the audience can not engage or interact further in the pursuit of a deeper understanding. Our story was effective in its presentation and in summarizing the events of the election, however, in some ways, it remains a shallow experience for a user. Through this story, I was able to grow my understanding of how to apply techniques in a visualization to direct attention [15] as well as principles of laying content out such that it is palatable for the audience. One main takeaway from this experience was the value of audience participation and interaction with data stories. If we had developed an interactive system that initially presents itself as this infographic and then supported audience interaction for detailed content it would have been more effective in engaging and retaining their attention. From this work, I felt that developing data stories without audience interaction leaves it wanting, and sought to change that in the next story.

3.2.2 Learning About Disease Associations in Taiwan

In the next data story, we offered an author-driven narrative in the beginning that was followed by free interactive exploration by the audience. We applied the martini glass structure [1]for the narrative. We used this approach to analyze data from Taiwan's National Health Insurance Research Database (NHIRDB). The data is stratified into ten age groups, and for each group, the likelihood of a disease co-occurring with each of the other diseases is presented. We depict disease associations among the studied subjects through a visualization that allows us to answer questions, such as "If I have heart failure what else am I susceptible to?" and "How do these associations evolve with age?" Our visualization is modeled after a salmonella bacteria

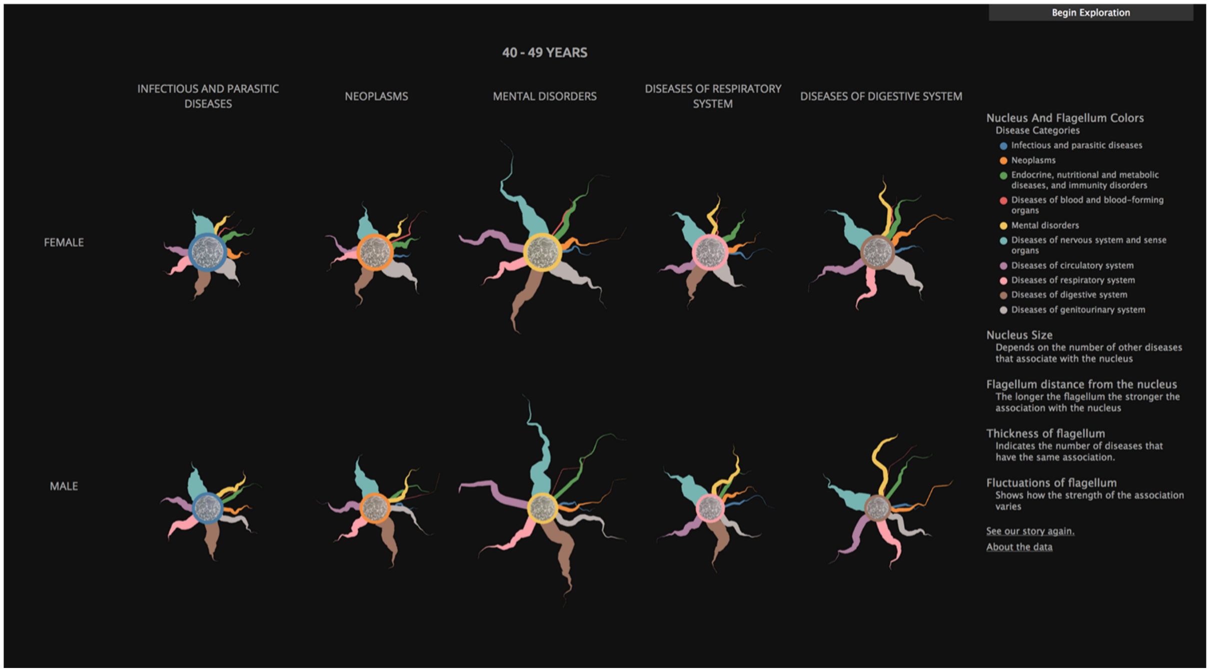


Figure 3.2: Disease associations in male and female 30–39 year olds. The visualization shows that mental disorders have a strong and varied association for each disease category from the length and fluctuation of the flagella. On the other hand, infectious and parasitic diseases do not; they have a larger amount of diseases with associations than mental diseases. This is indicated by the size of the capsule. Furthermore, we can see that diseases of the digestive system have stronger associations to mental disorders (indicated by the longer length of the orange flagellum) when compared to that of other disease categories.

cell. The capsule size denotes the number of diseases that are associated with a focal disease category. Each flagellum represents a disease category. The strength of the association with the focal disease category is conveyed by the length of the flagellum. Diseases further away from the center have a stronger association. The thickness of the flagellum at any distance from the center represents the number of diseases, in that category, that have the same degree of association.

There were some concerns about whether the metaphor of a salmonella bacteria would make it easier for the audience to infer the purpose of the visualization.

This story was presented at the PacificVis 2018 storytelling contest in Kobe, Japan, and was given an honorable mention. Those at the conference who interacted with our story would change disease categories. Through these interactions, they were able to deepen their understanding of the data that binds to the metaphor. Without such an interaction it is doubtful that

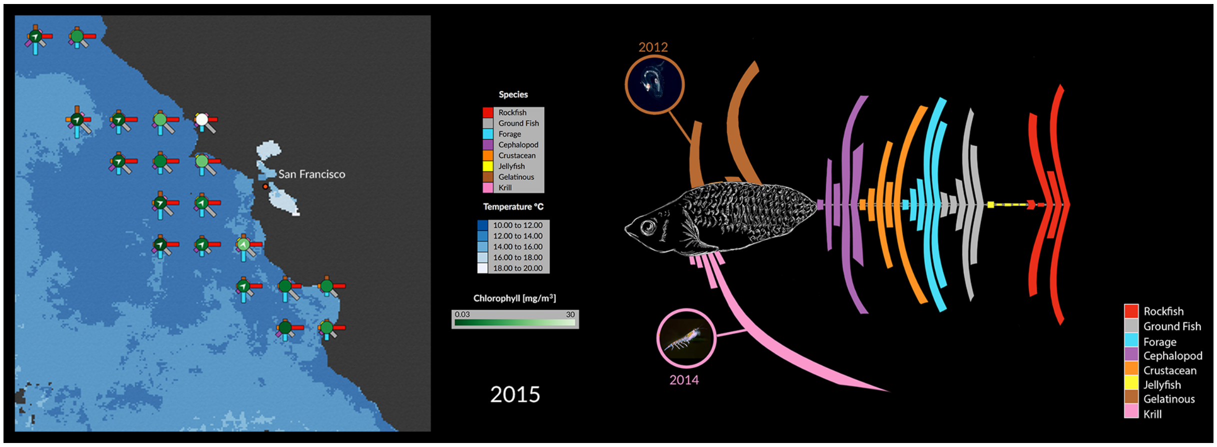


Figure 3.3: Species diversity, species abundance, wind direction, and chlorophyll associations in the San Francisco Bay Area for 2015. The visualization on the left shows that the interaction of a warm water mass with high chlorophyll levels led to an increase in species diversity. On the right, these changes show how the population of gelatinous plankton increased while the number of krill decreased, showing a shift in the food web.

at a glance the audience would be able to make sense of this type of visual representation. The visual metaphor was effective in conveying to the public that this content was about disease. Overall, we found the use of an author-driven beginning combined with an interaction that was paced by the audience, enabled the audience to understand the visual representation and infer the purpose of the visualization. In section 3.2 we will provide a detailed discussion of this visualization.

3.2.3 Periodic Temperature Effects on Biodiversity

After studying the genres of annotated charts and an interactive system, the next genre of data story I wanted to explore was a data video. From the previous story, we found that interactive systems encourage audience participation and help build a relationship with the story. With data videos come a new set of storytelling devices to apply in communicating a story, primarily audio. Additionally, with a video, we have greater control over the type of content the audience sees. In contrast to infographics, we have dynamism with transitioning from scene to scene and access to more visual cues to direct attention to the important content of the data visualization. However, we do lose the unique and paced exploration audiences would have with interactive visualization. Instead, audiences can have an experience that the author believes is the “best” way to present the content.

This last story used data provided by the National Oceanic and Atmospheric Administration (NOAA). NOAA collects near real-time data on the oceans and atmosphere. We analyzed this data in the region spanning from Monterey Bay up to the San Francisco Bay Area in the United States from 2009 to 2015 and created a visualization to help illustrate the relationship between climate and biodiversity to the public. We designed a glyph, inspired by Slingsby's work, to represent chlorophyll volume, wind direction, and species diversity. Our glyph is applied to a tilemap of the region, where the tile color is sea surface temperature. The implementation of the glyph and tilemap was done using Three and exported as images. Using Photoshop we manually illustrated species abundance data. The combination of our glyph and our tile map allows us to better see how these variables influence one another and shed light on how climate affects biodiversity.

This story was presented as a data video, so each scene utilized an audio narrative paired with supplemental visuals to help contextualize and frame the story to the audience. For example, the video opens up with ocean waves transitioning into a scene that zooms in from the globe to the specific region of the ocean where the data was sampled from. From there we transition into the data visualization and show how temperature over time affected the ecosystem. We also provided a visual metaphor to show the differences over the year of the fish population growth. The feedback for this work was the glyph and small multiples were useful for people to get a sense of how changes in temperature can affect these ocean ecosystems.

From these stories, one of the techniques that stood out as effective in communicating data was using some form of a visual metaphor. To learn more about the strengths of a visual metaphor and explore the processes of developing one we had further. We extend the work on the co-occurrence metaphor.

3.3 An Organic Visual Metaphor for Public Understanding of Conditional Co-occurrences

I will now provide details of the data-driven visual metaphor from the data story that focused on disease associations. In many domains, conditional co-occurrence—the likelihood of other events given that one event has occurred—is used for understanding associated events and making decisions. For example, market researchers are interested in customer purchase patterns (e.g., the likelihood of a customer purchasing product A, given that he/she has purchased product B). The retail industry uses co-occurrences of purchases between different product categories to manage inventory and pricing decisions [73, 74, 75]. Also, microbiologists have an interest in observing the co-occurrence of various plankton and bacteria populations to understand the coexistence within biological communities [76].

In health care, conditional co-occurrence of diseases/disorders (or comorbidities [77]) shows the presence of one or more additional diseases/disorders co-occurring with the primary disease/disorder. Therefore, co-occurrence relates to the progression of diseases and mortality risks [78]. The understanding of co-occurrence drives many critical clinical decisions, such as choice of drug therapies, surgical procedures, etc [79, 80].

However, learning about co-occurrence is not limited to domain experts. For example, in healthcare, the public, including patients, could understand the potential consequences of some diseases on their quality of life by understanding the co-occurrence of diseases. It is reasonable to assume a person who is diabetic will be interested in knowing they have a high likelihood of becoming visually impaired. By understanding the co-occurrence of diseases, we can make better choices to improve our personal health [81]. As a result, this improves our health literacy and leads directly to improved health outcomes [82]. However, the collected co-occurrence data by researchers is often complex [83, 84, 85]. Thus, for non-domain experts, or the general public, it is difficult to interpret this data and build their own useful insights. Therefore, we need to provide a proper method to share co-occurrence data with the public.

In an effort to improve public understanding of co-occurrence, we introduce an organic visual metaphor, which is designed not only to provide a summary of complicated co-occurrence data but also accessibility to the public [86, 70]. This metaphor is an extended version of our previous work for the PacificVis Visual Storytelling Contest [87]. In this chapter, we describe our metaphor design and the algorithm for generating the visualization. We showcase our metaphor model with an interactive application, through which the user can review the co-occurrence of diseases and the influences of demographic factors (e.g., age and sex) on the co-occurrence. We also present two examples that illustrate the utility of our model.

3.3.1 Background and Related Work

In order to make the public understand conditional co-occurrence data more easily, we consider two major design criteria in our visualization: 1) effectively summarizing large, complex data and 2) utilizing a visual metaphor. First, summarizing data effectively is important to avoid overwhelming viewers. To accomplish this, we utilize the hierarchical nature of many commonly occurring data sets. Hierarchical data permits us to bundle elements that belong to the same group [88] For example, in market research, conditional co-occurrences of the purchase of carrots or onions with the purchase of cheese or yogurts can be summarized as co-occurrence of dairy product purchase with the purchase of vegetables. Also, in order to clarify the direction of the conditional relationship of co-occurrence, we use an egocentric visualization. Second, as described above, visualizing data with familiar representations for the viewers, including non-experts, can help better understand and motivate the exploration of complex information.

3.3.2 Design Considerations

An important consideration when designing a visualization tool for non-experts to understand and explore complex data is aesthetics. If certain aesthetics are well-embedded in a visualization, it is easier for the viewer to process the information presented. For example, in graph drawing, Purchase [89] found that minimizing edge crossing has a strong effect on human understanding. Cawthon and Moore [90] performed a study to determine the effect of aesthetics on usability in data visualization. They found that people are less likely to abandon a task when they perceive the visualization to have high quality. This implies that people are likely to be more patient when visuals are aesthetically appealing.

Visual metaphors provide a desirable aesthetic to the viewers by employing a more familiar representation. For example, Wang et al. [70] studied how users react to different personal visualization designs. They found that illustrative designs are the most well-balanced for relaying information and motivating data explorations. Sallaberry et al. [69] employed a biological metaphor to effectively illustrate multiple relationships in a hierarchical dataset. Fung et al. [91] developed an egocentric visualization based a tree metaphor. In a later design study [86], they compared their organic visual metaphor against a set of visualizations and found that their visually attractive approach was the best for at-a-glance characterization of the data. While evaluating the value of the visual metaphors, Risch [4] argued that the users understand visualizations through image schemas–structural patterns established in their childhood. By employing familiar visual metaphors, we enable the user to interpret the data in terms of existing internal schemas. That is, visual metaphors help viewers translate abstract or complex concepts into a form that is more understandable.

In order to make the public understand conditional co-occurrence data more easily, we consider two major design criteria in our visualization: 1) effectively summarizing large, complex data and 2) utilizing a visual metaphor. First, summarizing data effectively is important to avoid overwhelming viewers. To accomplish this, we utilize the hierarchical nature of many commonly occurring data sets. Hierarchical data permits us to bundle elements that belong to the same group [88] For example, in market research, conditional co-occurrences of purchase of carrots or onions with purchase of cheese or yogurts can be summarized as a co-occurrence of dairy product purchase with purchase of vegetables. Also, in order to clarify the direction of the conditional relationship of co-occurrence, we use an egocentric visualization. Second, as described above, visualizing data with familiar representations for the viewers, including non-experts, can help better understand and motivate the exploration of complex information.

3.3.3 Visualization Design

Based on the design criteria, we derive our organic metaphor model and algorithm to generate the visualization.

3.3.3.1 Organic Metaphor Model

Before visualizing conditional co-occurrence, we need to obtain a metric. Here, as one concrete example, we use an association measure q which was used in previous studies in the biomedical science field [92, 93]. However, we can use a different metric based on the needs of the application. q is calculated as follows:

$$q = \frac{P(X|Y)}{P(X)} = \frac{P(X \cap Y)}{P(X)P(Y)} = \frac{N_{X \cap Y}/N}{(N_X/N)\cdot(N_Y/N)} \; (0 \leq q \leq \infty) \tag{3.1}$$

where X and Y are two different items (e.g., diseases), $P(\cdot)$ denotes the probability, N is the population size of a given subject group (e.g., patients), N_X and N_Y are numbers of individuals who have at least one occurrence of item X and Y over a given period (e.g., in five years)

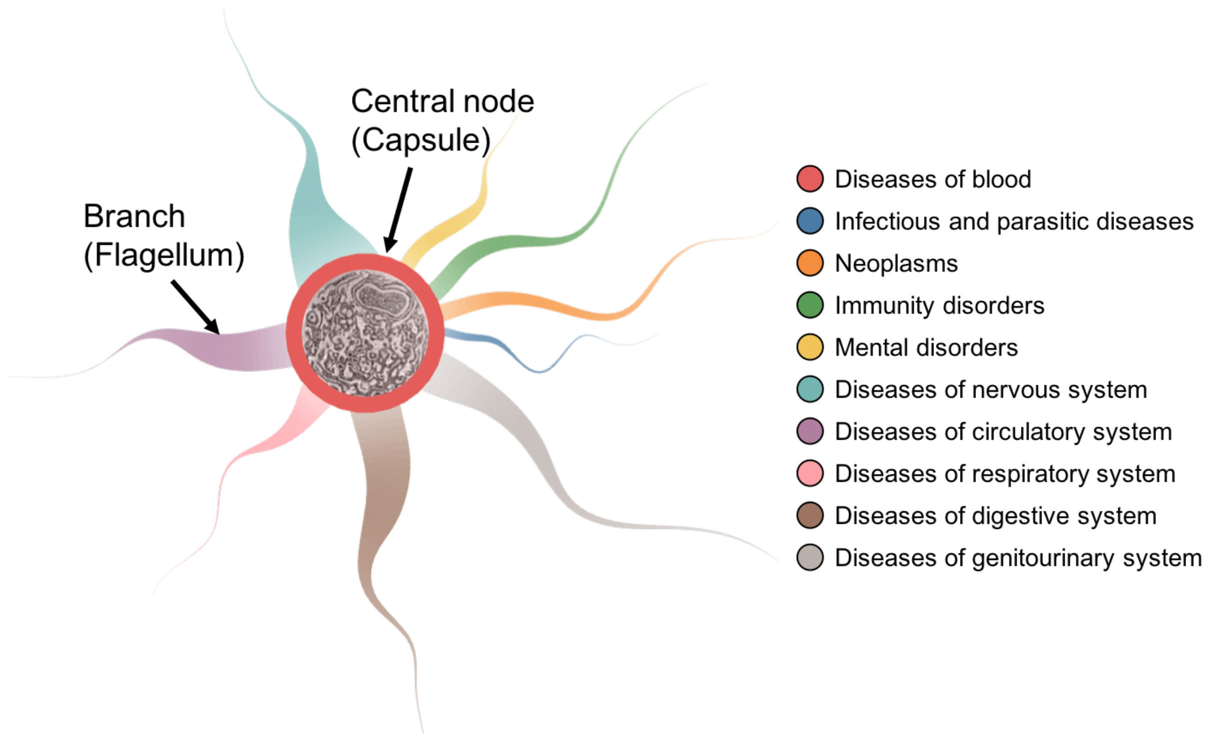


Figure 3.4: A bacterium metaphor generated with our model. The central node and branches mirror the bacterium's capsule and flagella.

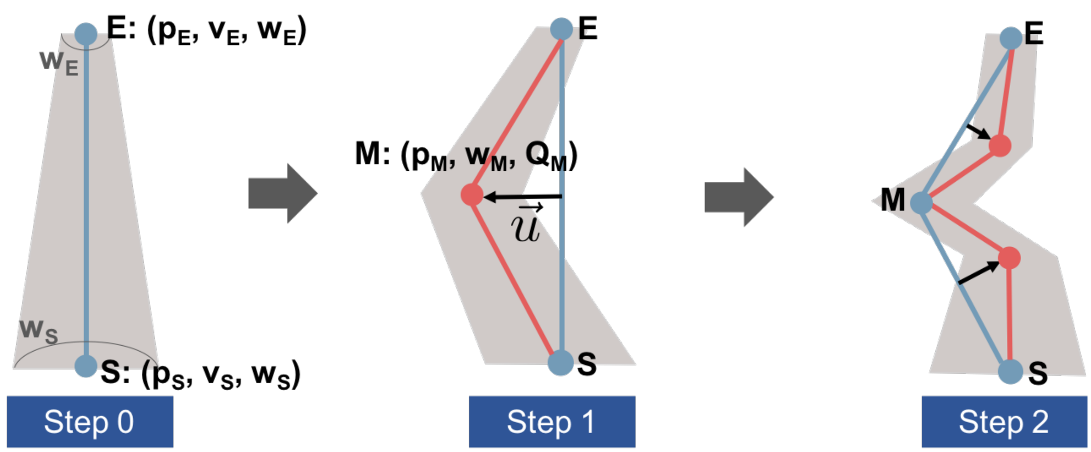


Figure 3.5: Recursive steps for generating a flagellum with L-system. Gray backgrounds show the generated flagellum. Blue and red lines are segments before and after applied L-system respectively.

respectively, and $N_{X \cap Y}$ is a number of individuals who have had occurrences of both X and Y in the same period. The value q measures how strongly the occurrence of X is impacted by the occurrence of Y. When $q < 1$, $q = 1$, or $q > 1$, it means there either is a negative, neutral, or positive association between X and Y, respectively. To specify, when $q > 1$, X and Y are more strongly associated as q increases. For example, when q is a large value, the occurrence of X is significantly increased given that Y has occurred. We consider only the positive associations (i.e., $q > 1$) and use $log(q)$ as a value v of the co-occurrence metric for the purpose of this work.

For the visualization, we developed an organic metaphor model. Figure 3.4 shows an example of visualizing the co-occurrence of diseases using our model. The visualization not only summarizes the necessary information in an egocentric way but also motivates users to explore the data [86, 70]. In order to reflect the negative connotation of disease associations, our design employs a bacterium metaphor. The central node and branches correspond to the bacterium's capsule and flagella. Our organic metaphor visualizes places the focal category (e.g., diseases of the circulatory system) in the capsule, while other related categories are placed in the branches or flagella. The colors used for the central node and branches represent the respective category of each item. To further refine our model, various statistical measures related to v in the co-occurrence metric are also visually encoded: 1) the size of the central node denotes the total number of items where $v > 0$ (i.e., $q > 1$, positively associated); 2) the length of the branch represents the value of v for that respective item; 3) the thickness of a branch represents the number of items with the same v at a given point; 4) the fluctuation of a branch conveys variation in v across different items in the same category corresponding to that branch.

The bacterium model helps users understand the visualized results by tapping into prior knowledge about organic matter. For example, using the bacterium metaphor, the viewer compares multiple bacteria generated across different groups (e.g., younger and older females), as shown in Figure 3.8. A bacterium that has longer branches would denote that the disease category placed in the central node has stronger effects on occurrences of the other diseases (e.g., an older female has a higher chance to have some neoplasms after a mental disorder).

3.3.3.2 Generation Algorithm

In order to generate an organic appearance for each branch, our model employs L-systems [94, 95], which recursively create botanical shapes, lightnings, etc. Figure 3.5 shows the first two steps of the recursion of our algorithm. Our algorithm starts from two points S and E. S is the point at $v = 0$ and E is the point where the maximum value of v for that branch is attained. The algorithm recursively decides the placement and values of the middle point M. Let p_i, v_i, w_i be a 2D coordinate, a value of v, and a width of a branch at point i (S, M, or E) respectively. p_M, v_M, w_M are calculated with

$$p_M = (p_S + p_E)/2 + \vec{u} \tag{3.2}$$

$$v_M = (v_S + v_E)/2 \tag{3.3}$$

$$w_M = \alpha N_{\leq v_M} \tag{3.4}$$

where $\vec{u} = c\beta\sigma\vec{n}$, and c is either 1 or -1 depending on whether the number of recursions at that point is an even or odd number. σ is the standard deviation of v of items (e.g., diseases) where $v_S \leq v \leq v_E$, $\vec{n}$ is the unit normal vector of $(p_S - p_E)$, $N_{\leq v_M}$ is the number of items whose $v \leq v_M$. α and β are parameters used for controlling the width and fluctuation of a branch respectively. After the first step above, the algorithm applies the same rules to pairs of points S and M and M and E. By calling these steps recursively with an indicated repeat count, we can obtain a polyline.

The parameter β amplifies the fluctuation and emphasizes the differences in the standard deviations among items. However, a large β can cause an overlap of branches. Thus, we choose β where it emphasizes the differences, yet avoids overlaps. We describe how to choose β as follows. We assume that the y-coordinates of points S and E are 0, as shown in Figure 3.6. Let r and L be the radius of the capsule and the length of the segment from S to E. As seen in Figure 3.6, when a branch is between the straight lines represented with $y = (w_S/2r)x$ and $y = -(w_S/2r)x$, the branch does not overlap with other branches. To fulfill this condition, the y-coordinate of the orange point in Figure 3.6 must be less than or equal to the y-coordinate of

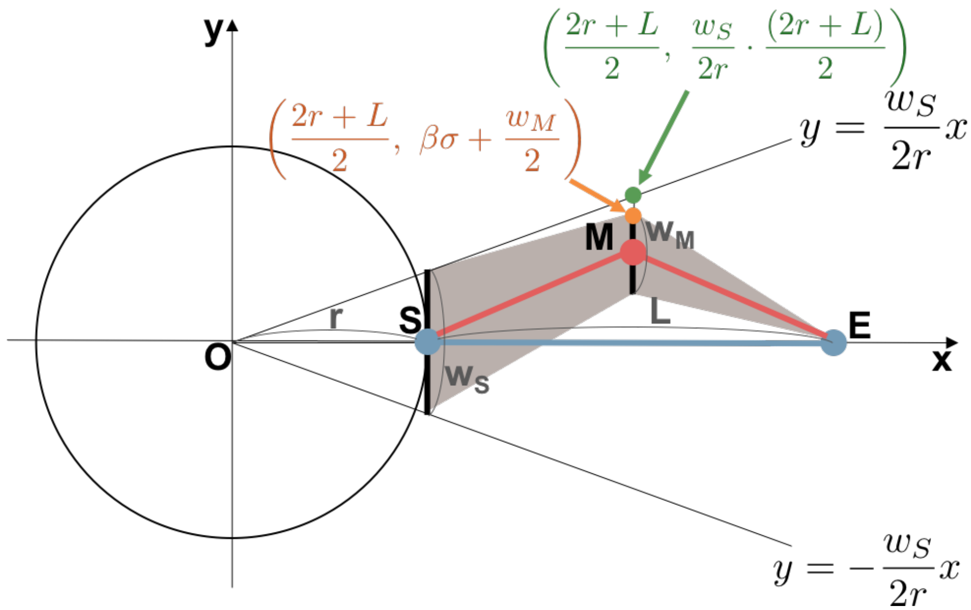


Figure 3.6: The position of a branch where current β avoids overlapping with other branches.

the green point in Figure 3.6. Therefore,

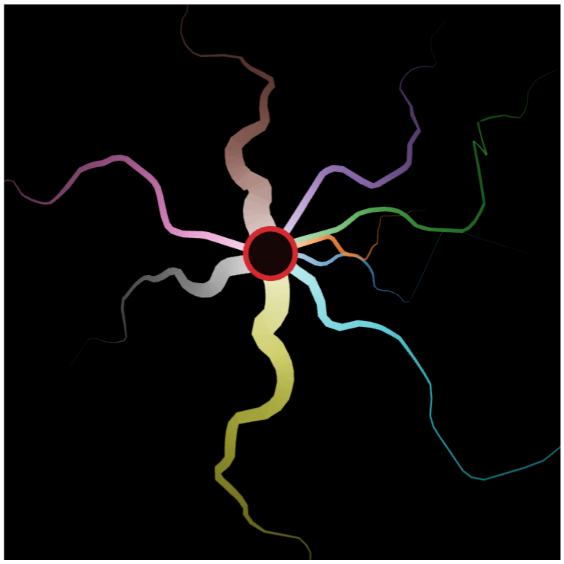
$$\beta\sigma + \frac{w_M}{2} \leq \frac{w_S}{2r} \cdot \frac{2r+L}{2} \quad (r > 0) \tag{3.5}$$

$$\therefore \begin{cases} \beta \leq \frac{1}{2\sigma}\left(w_S + \frac{w_S L}{2r} - w_M\right) & (\sigma > 0) \\ \beta \leq \infty & (\sigma = 0) \end{cases} \tag{3.6}$$

Equation 3.6 considers only one branch. Therefore, we need to generate the condition indicated in Equation 3.6 for each branch. Then, we can choose β to fulfill all the conditions.

Next, in order to generate a smoother shape, we apply the cubic spline and the monotonic cubic spline [96] interpolations to the positions and the branch width, respectively. We generate branches for all the categories of items with these steps and arrange them around the central node. Also, we use an organic texture in the central node.

We use the bacterium metaphor in Figure 3.4 as an example of our organic metaphor. However, we can produce other kinds of metaphors by changing the parameters in the algorithm, texture, etc. For example, in 3.7a, we create a lightning metaphor with longer branches, large β, higher recursion counts, and no texture on the central node. In 3.7b, we create a flower metaphor with shorter branches, small β, and a texture representing disk flowers. While we use



(a) A lightning metaphor

(b) A flower metaphor

Figure 3.7: Examples of visual metaphors generated with our model.

the bacteria metaphor to leverage the negative feelings towards co-occurrence of diseases. The lightning or flower metaphor would be useful to imply a neutral or positive impression.

3.3.4 Application Example and Evaluation

We test our model on a dataset from the Taiwan National Health Insurance Database [97] from the period of 2000 to 2002, which contains information from 782 million outpatients. The dataset includes disease codes based on ICD-9-CM [98], disease names, their disease categories (e.g., diseases of the digestive system), the values N, N_X, N_Y, $N_{X \cap Y}$ in Equation 3.1, and demographics of patient groups (age groups and sexes). The disease categories provided by ICD-9-CM have a hierarchical structure. For example, diseases of the respiratory system have subcategories of acute respiratory infections, pneumonia, influenza, etc. We calculate value q with Equation 3.1 for each pair of diseases for each patient group (e.g., 20–29-year-old females). We showcase the usage of our model with an interactive application for understanding the conditional co-occurrence of diseases. We also employ two cases to demonstrate how our model can be used to learn about associations.

3.3.4.1 Application Example

Figure 3.8 shows our application for exploring the co-occurrence of diseases. The application consists of two views (Figure 3.8A and C) and a control panel (Figure 3.8B). Figure 3.8A Employs our bacterium metaphors placed in a grid structure. With the control panel (Figure 3.8B), the viewer can adjust what metrics the x and y dimensions of the grid can be. For example, in Figure 3.8, while males and females are selected for y dimension, age-groups of 20–29 and 30–39 years are selected for x dimension. The control panel also allows the viewer to set a disease category shown as the central node (e.g., mental disorders are selected in Figure 3.8). A metaphor in each grid shows co-occurrence visualization related to the selected combination. For instance, the metaphor placed on the top left shows conditional co-occurrence from mental disorders to other disease categories of 20–29 years old females. In addition, we also visualize a small tree-map in Figure 3.8A to help compare the total number of diseases in each branch. Also, when the mouse hovers over the tree-map, the system shows the corresponding total number in a pop-up box. The application also allows the viewer to select the central node or branch by mouse clicking (e.g., a selected branch is colored with high-saturated red in Figure 3.8). Then, the application shows the number of diseases in each disease subcategory in the selected branch as a bar chart, as shown in Figure 3.8C. The bar chart arranges the subcategories for the x direction and plots the numbers of diseases as y-coordinates. By hovering a mouse, the viewer can see the name of the selected subcategory and its number of diseases, as shown at the bottom of Figure 3.8C.

Our application is web-based and developed with JavaScript, D3.js[1] for drawing graphs, Three.js[2] for drawing bacterium metaphors. To avoid running our algorithm while using the application, we preprocessed the dataset to prepare polygons of the metaphors in advance. The source code for this polygon generation is available online[3]. The preprocessing takes around 80 ms for one metaphor with 3 recursion counts on a 3.1 GHz Intel Core i7 processor with 16 GB memory.

[1]D3.js, `https://d3js.org/`, accessed: 2018-7-29

[2]Three.js, `https://threejs.org/`, accessed: 2018-7-29

[3]The source code for the polygon generation, `https://github.com/takanori-fujiwara/organic-visual-metaphor`

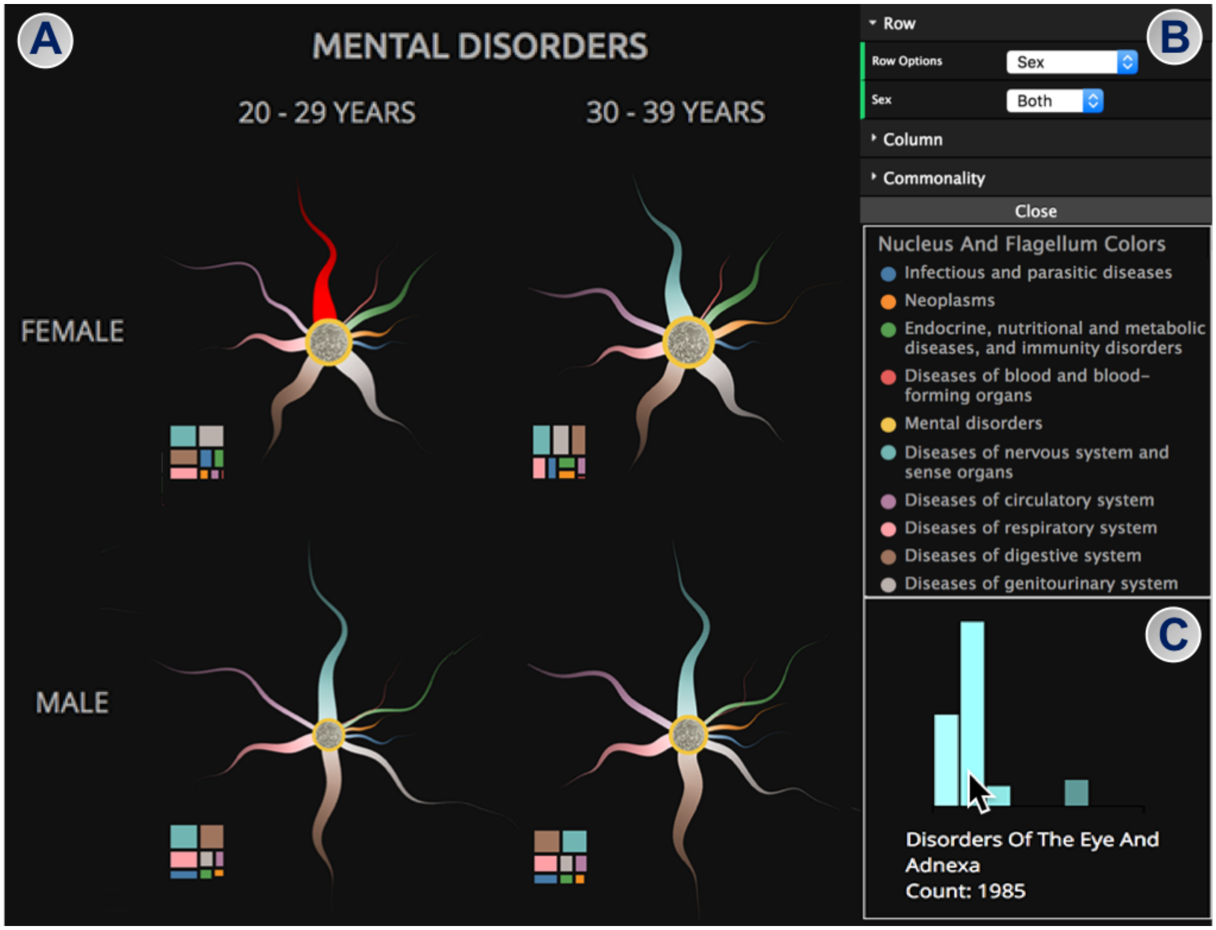


Figure 3.8: The user interface of our application example.

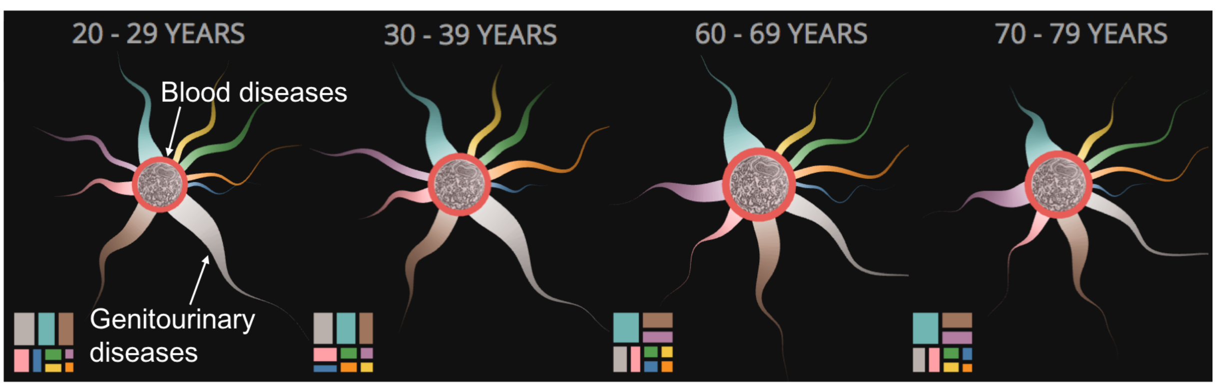
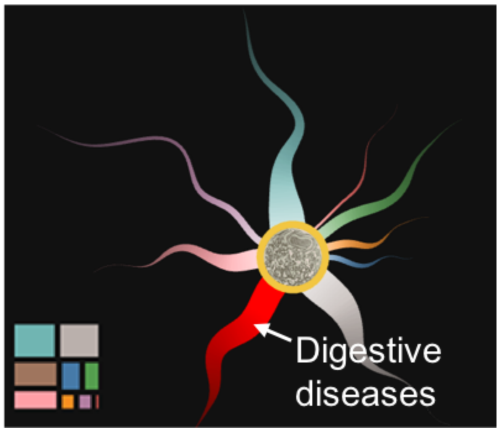


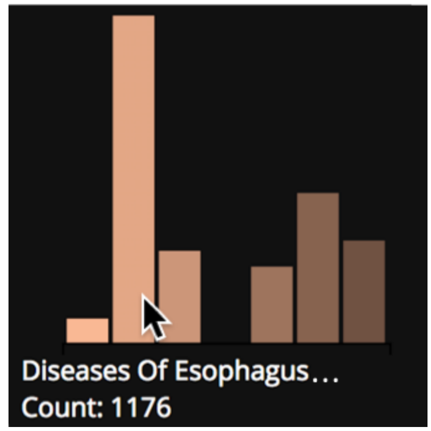
Figure 3.9: The co-occurrences with the blood diseases in female groups.

3.3.4.2 Case Study 1: Aging Impact on the Co-occurrences

For our first case, we observe the co-occurrences from blood diseases to others. First, we select the blood diseases as the central node and visualize two age ranges of female adults (20–29 and 30–39 ages), as shown in the left side of Figure 3.9. At a glance, we can see that the



(a)



(b)

Figure 3.10: Exploring the diseases' subcategories from the co-occurrence of the female 20–29 groups visualized in Figure 3.8. In (a), the branch for digestive diseases is selected. Its subcategories are visualized in (b).

genitourinary diseases (gray branches) stand out the most. When looking at the corresponding tree maps, we confirm that genitourinary diseases are reported the most for females of these two age groups. We then explore how aging impacts the co-occurrences of diseases with blood diseases in females. We add the female age groups of 60–69 and 70–79, as shown in Figure 3.9. We find that the genitourinary branches are much thinner and shorter when compared with the age groups of 20–29 and 30–39, indicating the co-occurrence of genitourinary diseases decreases in the older age groups. This finding is also supported by medical studies about this relationship between age and genitourinary diseases in females. At first, younger females are at a higher risk for anemia and autoimmune disorders [99, 100]. In addition, females in their 20–30's have a higher risk to suffer from sexually transmitted infections, including UTI and STD [101, 102].

3.3.4.3 Case Study 2: Co-occurrences of Mental Disorders

In this second example, we explore the co-occurrence of mental disorders in males and females. We select mental disorders to be the central node of our bacteria. We then include males and females in the age ranges of 10–19 and 20–29. The visualized results are shown in Figure 3.8. We can see that four bacterium metaphors have varying sizes and lengths in the central nodes and branches.

First, we examine a single bacterium for males of the age group 20–29. The flagellum for both the nervous system (the branch colored light blue) and digestive disorders (brown) stand out. Comparing the male 20–29 groups to the female 20–29 groups, we notice that the central node size for the female is larger. This indicates that this female group has more reported cases of other diseases co-occurring with mental disorders than this male group. We also notice that, unlike the male group, genitourinary diseases (gray) co-occurred more often as indicated by their wider width. Also, with all four bacteria, we can see that the central node size increases as age increases in both females and males.

As stated earlier, the length of a branch depends on the value of co-occurrence with the central node. Branch thickness is indicative of the number of diseases that have the same association. Therefore, in all four bacteria, we see that mental disorders (represented by the central nodes) and digestive diseases (brown branches) have a strong association. This finding is reasonable from our everyday experience of stomach pains caused by our anxiety.

We can also show more detailed information on the co-occurrence of diseases by selecting each branch. For example, in Figure 3.8A, we select the branch for nervous system diseases for the female age group of 20–29, as highlighted with high-saturated red. Then, in Figure 3.8C, the system visualizes the number of diseases included in the subcategories of the nervous system diseases. As shown in Figure 3.8C, when hovering the mouse over the longest bar, we can see that the dominant group is "disorders of the eye and adnexa". Similarly, as shown in 3.10a, we select the branch for digestive diseases and find that the subcategory for "diseases of esophagus, stomach, and duodenum" in 3.10b is also highly associated with mental disorders. Pulling all these details together, we can ask why patients with mental disorders also report either diseases that fall under "diseases of esophagus, stomach, and duodenum", or "disorders of the eye and adnexa". Our application enables us to ask such questions, which could be the starting point of additional inquiry or research. Interestingly, some researchers have found relationships between irritable bowel syndrome and anxiety [103] as well as evidence showing those who go blind also tend to experience anxiety or forms of depression [104].

3.3.5 User Feedback from a Visual Storytelling Contest

We received positive feedback from all judges in IEEE PacificVis Visual Storytelling Contest informing us they liked the original visualization and the visual metaphor introduced. The comments stated: *'I enjoyed the use of the visual metaphor', 'design and data well organized'*. We received positive feedback about our visualization from all judges in IEEE PacificVis Visual Storytelling Contest although the demonstrated system [87] was the state before improving the visual metaphor and functionalities. The contest informed us that "all judges liked the original visualization and the visual metaphor introduced". Also, the positive comments from conference attendees stated: "I enjoyed the use of the visual metaphor", "design and data well organized", "The metaphor looks very gross and off-putting which suits the subject matter", and "visualization design is really effective". Some attendees had some suggestions or concerns about the design stating: "Would this work for other data sets?", "It is hard to see the smaller branches since they are covered", "I would like to see more information about specific diseases", and "Animating the fluctuations in the branches could reinforce the metaphor and attract more people". We have approached most of these concerns in this work. For example, we have demonstrated how we can generate a proper metaphor for the different types of data, as shown in Figure 3.7. We have resolved the overlapping problem as discussed in subsubsection 3.3.3.2. Our application integrates additional visualization and interactions to see more detailed information, as described in subsubsection 3.3.4.1.

Some attendees had some concerns about the design stating: *"I am unsure if this approach is truly effective", "It is hard to see the smaller branches since they are covered", "Would this work for other data sets?", and "Are you embedding too many dimensions into the branch?"*

3.3.6 Conclusion and Future Work

The use of a visual metaphor is effective when presenting complex information to the public. It provides a stepping stone for them to engage and learn more about an unfamiliar and complex topic. We have introduced an organic metaphor model to assist the public in interpreting conditional co-occurrences. Through a set of case studies, we demonstrated how we can uncover interesting findings with our application example, using our model.

The intent of our visual metaphor is to provide the general public with a method on how to view conditional co-occurrence data. We consider that traditional methods either focus on one disease and its co-occurrences or present all disease associations. As demonstrated in the case studies, our model provides an overview of these associations and can be used to explore a variety of relationships when placed within a system. However, our model would not scale well if there are many item categories. Also, in both case studies, we visualize only a limited number of groups as branches. We utilized the hierarchical categories of the dataset in order to summarize co-occurrence information into a small number of categories. Adding animation to our organic metaphor is one interesting future direction. In order to better attract the public, we could employ effective organic animation.

For future iterations, we plan to conduct user studies to evaluate our metaphor in more depth. We want to evaluate the impact of attractiveness on both user engagement for data exploration and abandonment rate for complex tasks. Also, we would like to verify this effect on different types of audiences: domain experts, higher learners, and the general public.

Chapter 4
Data Storytelling in Public Settings

This chapter includes content from: **Dasu, K.**, Ma, K. L., Ma, J., & Frazier, J. (2020). Sea of genes: A reflection on visualising metagenomic data for museums. IEEE Transactions on Visualization and Computer Graphics, 27(2), 935–945.

4.1 Introduction

The previous work I conducted on creating both data stories and designing a visual metaphor for a public audience delivered the stories in a controlled setting. This setting was that of either a web page, a video, or a poster. It was controlled in the sense that the story expected a single audience member to receive the story and as an author, I could ensure the audience all took the same entry point to consume the story. However, there are public settings where it is not as predictable for how many audience members you have and when during your data story they join in. I examined the process of designing an exhibit to communicate scientific findings from a complex dataset and unfamiliar domain to the public in a science museum. This work was a collaborative effort between UC Davis, the Exploratorium, and Stamen Design. Our exhibit sought to communicate new lessons based on scientific findings from the domain of metagenomics. This multi-user exhibit had three goals: (1) to inform the public about microbial communities and their daily cycles; (2) to link microbes' activity to the concept of gene expression; (3) and to highlight scientists' use of gene expression data to understand the role of microbes. To address these three goals, we derived visualization designs with three corresponding stories, each corresponding to a goal. We present three successive rounds of design and evaluation of our attempts to convey these goals. We could successfully present one story but had limited success with our second and third goals. This work presents a detailed account of an attempt to explain tightly coupled relationships through storytelling and animation in a multi-user, informal learning environment to a public with varying prior knowledge on the domain and identify lessons for future design.

4.2 Visualising Data for Museums

Visualizations are increasingly central to the practice of science. They are used across a range of scientific disciplines to analyze phenomena, such as changes in microbiomes and shifts in climate. There have been several large-scale efforts to develop scientific and information visualizations for the public: the National Oceanic and Atmospheric Administration's (NOAA's) Science on a Sphere presents earth systems datasets such as tsunamis, climate models, and sea surface temperature on a large spherical display for aquariums and museums [105]; DeepTree [106] visualizes evolutionary data for exploration on a tabletop interface in natural history museums; MacroScope [107] ports a range of visualizations into a large interactive display for a wide range of academic and museum settings; and Living Liquid [108, 109] created interactive visualizations for a hands-on museum environment. Each of these projects, as well as many others [110, 111, 112], have contributed to our understanding of the opportunities and limitations of visualizations in museum settings. However, these projects visualized concepts such as currents, weather, evolutionary trees, and migration paths that the public has familiarity with.

This chapter examines the challenges of creating a museum exhibit from a complex dataset from an emerging and unfamiliar field: metagenomics. Metagenomics, the characterization of all the genetic data in a sample, is revolutionizing our understanding of microbes. Researchers use these data to determine what species are present, what functions they perform, how these functions change over time, and infer how microbes interact [113]. Metagenomics is one of the primary ways researchers study microbes. Microbes play a central role in almost all aspects of life on earth [113]. Ocean microbes use energy from the sun to produce half the oxygen we breathe and drive our climate; soil microbes impact the food we eat; and scientists are beginning

Figure 4.1: Museum visitors using the Sea of Genes exhibit in the Life Sciences Gallery of the Exploratorium in San Francisco. (Photo is provided by Jennifer Frazier.)

to understand the complex and critical roles billions of microbes living in our bodies have on our health [113]. Despite the significance of metagenomics to scientific research, few efforts have introduced these data to the public through interactive visualization. Instead, exhibits [114, 115, 116] rely on electron micrographs and graphics of microbes.

Even though the visualization field has explored narrative elements as a strategy for engaging users with complex data [1, 117, 118], there has been limited work on visualization exhibits that present complex content from an unfamiliar domain in a museum setting. Unfamiliar domain, in this context, refers to the targeted audience having little prior knowledge of the domain. We examine the application of narrative visualization strategies and animation effects to reduce complexity and create familiarity when presenting scientific findings through a museum exhibit. The exhibit is called *Sea of Genes*. First, we provide detailed documentation of the design process for developing such an exhibit and the challenges we faced compared to our prior exhibit design experiences. We then address the limitations and constraints of designing a visualization exhibit in an informal learning environment and point out directions for research in this space. This work presents a detailed account of an attempt to explain tightly-coupled relationships through storytelling and animation in a multi-user, informal learning environment to the public with varying prior knowledge of the domain. We discuss takeaways and provide guidance for studying science museum exhibits, which we believe is especially valuable to both the field of metagenomics and other scientific domains.

4.3 Background

This section provides a brief background on the use of storytelling in the setting of a museum as well as prior research on the use of animation for learning.

4.3.1 Data Storytelling in Museums

Extensive work has been done on the application of narrative devices and visualization of complex data, see Chapter 2. There has also been research on the use of animation for teaching unfamiliar concepts. With *Sea of Genes* we paired principles from both educational animation and narrative visualization to develop an exhibit that could successfully present content from both a complex dataset in an unfamiliar domain. This work expands on the current literature on how the museum setting, especially a highly interactive one, affects the presentation of a narrative.

Segel and Heer [1] provide an overview of how visual elements have been employed in traditional media such as comics, books, and films to tell stories. Their focus is on the role of graphical elements and interactivity in maintaining continuity in the flow of the narrative. They identify author-driven and reader-driven as two polar extremes of visualization. We contextualize these two terms for the museum setting as designer-driven and visitor-driven. In a designer-driven approach, the story is linear and the visitor has no control of the narrative. It presents to the visitor a fixed sequence of events with which they can interact. In a "pure" visitor-driven approach, there is no predefined narrative. Instead, there is no fixed sequence of events, and the visitor would select and order events to create a narrative. By blending these two extremes in our own design, we seek to retain continuity and provide visitors the freedom to explore.

Narratives have been widely applied in history and art museums to help visitors make personal connections to an object or a collection [119, 120]. A study [121] on the roles narratives play in interactive science exhibits found enhancing exhibits with personal stories improved the exhibit experience for visitors and helped create personal connections to the content. However, adding stories seemed to reduce the visitors' physical interactions and explorations with the exhibit. Similarly, a study [122] examining the use of narrative introduction to describe the dataset visualized in an exhibit, found it did not improve data exploration. Further study of narrative applied to interactive visualization is needed, examining its applicability and effectiveness in communicating complex and unfamiliar content. We identified a set of related stories, which we could present in layers.

4.3.2 Animation for Learners

Research on how animation affects learning has gone through two eras of consideration. In the first era (1990s), researchers studied the impact animation has on learning by evaluating it next to static graphics [123, 124, 125]. These studies report inconsistent or inconclusive findings on the effects of animation on learning. In particular, although Schnotz and Grzondziel [126] found animation performed better, it had an interactive component [127] confounding the results. Tversky and Morrison [123] were highly skeptical that animations could be effective for conveying complex systems. They suggest two principles to note as conditions for an animation to be effective: Apprehension and Congruence. The *Apprehension* principle states "the structure and content of the external representation should be readily and accurately perceived." A drawback of animation is the perceptual and cognitive limitation of processing a changing visualization, e.g. complex interactions may occur too quickly to be understood. The *Congruence* principle states "the structure and content of the external representation should correspond to the desired structure and content of the internal representation." In principle, animation should be effective for expressing changes. Most animations violate these principles. People conceive a dynamic process as a sequence of steps, thus violating the Congruence principle. In order for an animation to be effective Tversky [123] believes animations must explain rather than simply show.

Rather than comparing animations' effectiveness to static graphics, recent studies focus on understanding the cognitive processes involved in processing dynamic visualizations and identifying the steps leading to comprehension [128, 129]. Berney and Betrancourt [130] conduct a meta-analysis on animation for learning and section the factors into three main groups: (a)

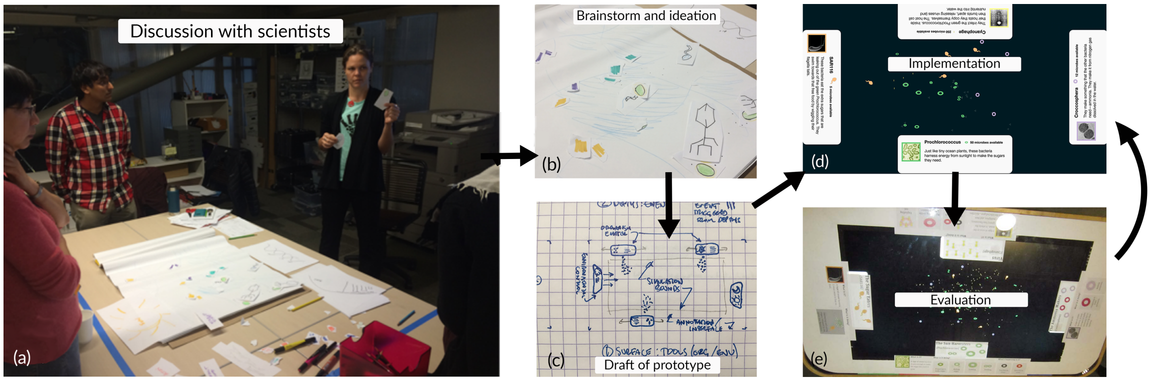


Figure 4.2: Design process (a) Photo capturing a discussion from the initial brainstorming session with a scientist from C-MORE. (b) Sketches of microbe behavior inferred from genomic data at the brainstorm. (c) At the end of the brainstorm, we adopted a sketch of Prototype 1 produced by Stamen Design. (d) Prototype 1 still. (e) Prototype 1 on the floor of the Exploratorium during evaluation.

specific to the learners, (b) specific to the instructional material, and (c) specific to the learning context. Studies [131, 132] that address group (a), to which museum visitors belong, found varying prior knowledge requires varying presentation forms to achieve a learning task. Therefore, with *Sea of Genes* we need to consider other ways to enforce the animation, which we detail in Section 5. Chanlin [131] found animation enhanced both novice and experienced learners' learning. Specifically, for novices, it helped facilitate the learning of descriptive facts. Berney and Betrancourt form a hypothesis based on Ploetzner and Lowe's [133] work that "well-designed" expository animations contain all the elements needed to draw learners' attention to the right place at the right time. That is, our animation should facilitate directing visitors' attention to each point when necessary.

4.4 Exhibit: Sea of Genes

Sea of Genes is a multi-user interactive visualization exhibit at the Exploratorium, a science museum. The exhibit is within the Living Systems Gallery of the museum, and uses a combination of animation techniques and narrative elements to communicate three key stories found in a metagenomic dataset of marine microbes:

- Microbial interactions occur in a predictable daily rhythm.
- Genes turn on and off according to a daily rhythm.
- Scientists collect data about the genes of microbes.

This exhibit was created through an interdisciplinary collaboration that brought together expertise in academic and commercial visualization practices, scientific research, and museum design and evaluation. The Exploratorium led the collaboration and provided a curator, project manager, writer, graphic design, exhibit designer, learning researcher, and evaluator. A visualization group provided a graduate student and professor to provide expertise in visualization and HCI research to assist the exhibit design. The University of Hawaii at Manoa provided a lead data scientist and marine microbiologist who provided datasets and content expertise. Stamen Design provided a digital graphic designer and visualization designer to provide expertise in public-facing commercial visualization design and public installation.

4.4.1 The Dataset

The data used for *Sea of Genes* were collected and analyzed by oceanographers affiliated with the Center for Microbial Oceanography: Research and Education (C-MORE) at the University of Hawai'i at Manoa and the Monterey Bay Aquarium Research Institute (MBARI). A full description of the data collection and analysis methods was published in a series of articles [134, 135, 136] during 2014–2017.

The 2014–2015 samples were collected using an Environmental Sampling Processor (ESP) [134, 135], a free-drifting sampling device that collects environmental and genomic data at specified times in the ocean, in this case, every 4 hours for 3 days. The 2017 samples were collected every 4 hours for 4 days using Niskin bottles [136] deployed from a research vessel. Planktonic microbial assemblages were collected by passing seawater through a 0.22 μm pore-sized filter, preserved in RNA later, and stored at -80°C within 24 hours of retrieval from the instrument. RNA was extracted, cDNA was generated, and Illumina sequencing [136] was performed. Metatranscriptome reads were mapped to ortholog clusters of proteins constructed from the phylogenetic groups of interest. Function was assigned by KEGG Orthology annotation. Read count tables were normalized to total read count, with the threshold set to achieve R2 value > 0.8 using the R packages igraph and WGCNA [137]. These count tables contained information about daily patterns in microbes such as time of collection, taxonomic assignment, gene function and expression levels, and the peak time of expression. From this, C-MORE and MBARI scientists were able to infer which microbes were present, what functions they performed, and when those functions occurred over the course of a 24-hour period. They provided us access to these data sets and assisted in interpretation, (Figure 4.2a).

4.4.2 Hardware and Implementation

For *Sea of Genes* we decided to use an interactive touch table, which has been shown to encourage collaboration and attract attention [138, 139, 140]. A feature of the museum context is the ability to support social experiences [141]. Hinrichs et al.'s [142] findings suggest using a large interactive display gives the visualization a presence within an exhibition. These displays allow people to enjoy and participate from a distance and decide whether to engage further. In our previous project, we used the Multitaction object-tracking table [108], which attracted and engaged visitors with the visualization. We took advantage of the social context of the museum by using a larger 3M 65" touch-table at 4k resolution as our exhibit display to accommodate either 6 visitors interacting all around it or 3 visitors from one side.

To support an iterative development cycle, *Sea of Genes* is web-based and written in ECMAScript 6, JavaScript 6. JavaScript is lightweight and is suited for rapid prototyping. Each microbe had its own custom sprite and a set of animated behaviors derived from their transcripts. The transcripts were functionally annotated, and patterns of sequential function were grouped into high-level categories (e.g., genes involved in preparing a cell to divide, then genes involved in the actual division). The time of expression for each transcript was determined by the time of day the sample was collected and the normalized amount of transcript in the sample. Details on the visualization process are expanded upon in Section 5. We provided a configuration file the museum staff can edit, allowing them to modify parameters such as the number of microbes and length of time for the 24-hour period to cycle. To package for distribution we used the open source software Electron by github.io [143]. For our own development, we deployed the exhibit on OSX architecture.

4.5 Design Considerations and Evaluation Methods

Interpreting metagenomic data requires understanding microbes, their genes, and gene expression. To create an experience around this complex dataset, we worked with C-MORE scientists to (1) synthesize their research into a narrative comprised of three related stories, and (2) apply techniques from established narrative frameworks to layer the following three stories into a cohesive narrative:

S1. Microbial interactions occur in a predictable daily rhythm. The first story conveyed to visitors that microbes form communities similar to larger organisms. The interactions and functions these microbes perform during the day and night differ.

S2. Genes turn on and off according to a daily rhythm. The second story focused on how the microbial interactions and actions in the first story are a result of the gene expressions of specific genes, which control microbial function.

S3. Scientists collect data about the genes of microbes to make sense of the temporal patterns in microbial functions. The final story was scientists collected data and identified the expressed genes responsible for microbial function.

These stories were the synthesis of the C-MORE scientists' research and, taken together, could provide the public with an understanding of how marine microbes are studied and what they do. From discussion with the scientists we found S1 and S2 were closely related to one another, with S2 explaining the molecular underpinnings of, or genetic expression for, the behavior in S1. S3 further elaborates that scientists study S2 to make sense of the temporal patterns in microbial interaction captured in S1. In short, we needed to show the public (S1) microbes have interactions that occur in a predictable daily rhythm which (S2) are the result of gene expressions, and (S3) scientists analyze these gene expression data to identify temporal patterns. How we present these stories is constrained by the considerations of an informal learning environment in an interactive science museum.

4.5.1 Museum Considerations

Museums are informal learning environments referred to as "designed environments" in which exhibits are developed to help structure visitor experiences, in line with institutional goals and values [144]. In addition to facilitating visitor engagement and comprehension of complex datasets, the team needed to ensure the exhibit design considered the informal learning context. The following considerations were identified and informed by our collaborators at the Exploratorium, which we used to constrain and guide our design process.

C1. Free-choice learning environment. As with other types of informal learning environments, the experiences in museums, as compared to the formal setting of the classroom, are motivated and guided by personal interests rather than compulsory requirements [121, 119]. This is often referred to as a ”free-choice” learning environment. In such an environment visitors may not encounter or even choose to attend to our exhibit. The exhibit design must consider methods to attract and retain visitor attention. As free-choice learning environments [144], museums employ a variety of techniques to attract and sustain visitors interests and engagement at exhibits. For example, DeepTree [145], an interactive visualization of the tree of life had strategically placed features which invited attention and used the interactive table to encourage collaboration. Likewise, the interactive plankton visualization in the Living Liquid project [108], used an animated visualization paired with a tangible interface to captivate visitors' interest and serve as a gateway for exploration of plankton patterns.

C2. Public comprehension. The audience we design for is the visiting public, who are typically not domain experts. Since our museum attracts a diverse audience we do not know where, on the spectrum of novice to expert, our visitor's prior knowledge is. Furthermore, the open space layout of the exhibits means there are no guarantees a visitor will come to the *Sea of Genes* exhibit with the prerequisite knowledge learned from a prior exhibit [146]. Therefore, we cannot assume familiarity with the underlying dataset or domain itself. Similarly, we cannot assume representations which experts use for interpreting the data will translate to the public [145]. However, we should be mindful to not trivialize the experience to exclude experts or people who want to explore the content matter deeply.

C3. Readily decipherable. When designing an exhibit in a science museum there is a need for fast decoding and ready interpretation of the visualization [147]. In the Exploratorium's *Traits of Life* exhibit collection holding times at a single exhibit ranged from 12 to 149 seconds [148]. In other words, visitors have a short dwell time at exhibits and within this time they need to decode what is visually presented. Our design should thus accelerate this decoding process.

C4. Support Multi-user Interaction. Exhibit design must allow for multiple visitors to view or interact. This comes in both the need to support social groups who frequent the museum and facilitate collaborative learning [149]. There are also logistical reasons for multi-user exhibits, such as preventing queues and facilitating visitor movement in the overall exhibit space, and providing more visitors access to an exhibit. Designing for multiple users has several implications. Because we cannot assume a visitor will come to the exhibit in its initial state, the design should ensure a visitor can interpret and interact with the exhibit regardless of the state of the visualization. Furthermore, a visitor's interaction with the exhibit must not adversely affect another visitor's experience. Ideally, there are supports to encourage visitors to share their thoughts with each other and come to a common understanding of their shared exhibit experience.

4.5.2 Evaluation Process

Formative evaluation is an integral part of the iterative exhibit development process at the Exploratorium (Figure 4.2 e). Depending on the complexity of the exhibit, development may entail several rounds of prototyping and evaluation, with each successive round testing prototypes with modifications informed by visitor feedback and behavior data collected through evaluation. For *Sea of Genes* we conducted three successive rounds of prototyping and evaluation.

4.6 Visualization Exhibit Design

Three iterations were designed and tested, each adding on one story from **S1–3**. During each iteration, every design choice was guided by our considerations, **C1–4**. The following discussion is organized according to key design decisions made during our iterative development and evaluation process.

4.6.1 Constructing the Stories

The first step we took, guided by Lee et al.'s [37] approach, was to spend time exploring the data and extracting data excerpts to use and support **S1 and S2**, as described in Section 4. A study was conducted earlier at the Exploratorium to examine prior knowledge and interests in marine microbes and metagenomics. A large majority (96%) of the 136 visitors interviewed described microbes by a role they believed microbes played, while few used scientific taxonomic classifications [150]. Consequently, we decided to focus on functional roles. To identify familiar functions from the dataset we referred to the Next Generation Science Standards (NGSS) [151]

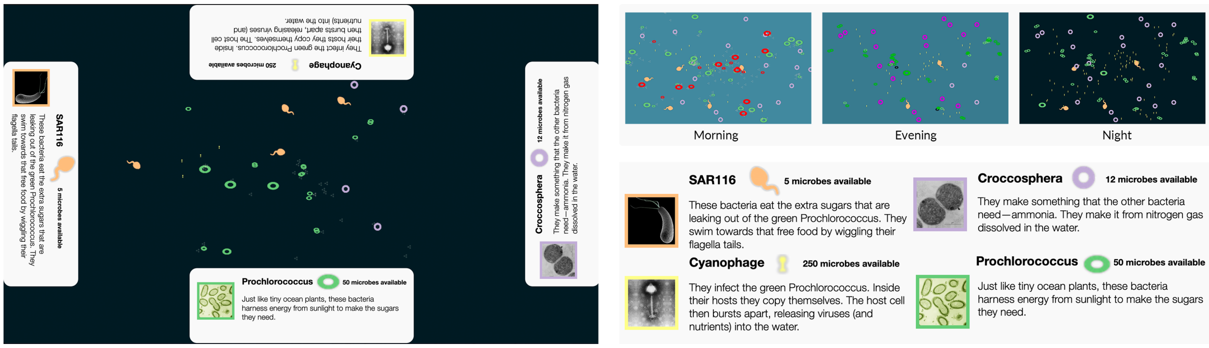


Figure 4.3: (a) Sea of Genes Prototype 1 visualized four main microbial characters depicted as 4 unique icons. This visualization was tested on a large tabletop display. (b) Stills of the central animation at different times of the day. (c) Cards providing information for each microbe.

and consulted with our partners at the University of Hawaii (UH). The science standards specify science concepts taught at each grade level and are used to guide the design of educational experiences. The Exploratorium often designs for the middle school level; however, we found that most of the functions were not covered until high school. With this in mind, we consulted with our partners, who suggested selecting microbes from their dataset based on their roles in a microbial ecosystem and could demonstrate **S1–2**. Selecting microbes based on roles rather than taxonomic classification would assist visitor familiarity **C2**. Four microbes were chosen for the first prototype (Figure 4.2). We selected phototrophs, *Prochlorococcus*, which draw energy from the sun, heterotrophs, *SAR116*, which consume other forms of energy like sugar, photo-heterotrophs, *Crocosphaera*, which draw energy from the sun and eat other forms of energy, and viruses. Viruses are not in NGSS; so, we relied on another prior study [152] that found 71% of teens knew viruses caused infections, and 79% recognized images of the type of virus used in the Hawaii dataset. Next, we chose microbial functions to animate for each microbe. We selected functions from the data that had a strong daily pattern and were familiar to museum visitors [150], selecting functions behind photosynthesis and cell division, which are concepts that are encountered in U.S. middle schools according to NGSS.

Prior studies [150, 152] suggest visitors believed microbes had a larger role in our ecosystems. However, we needed to determine which story to center the design around. Our previous study [150] found 95% of visitors knew microbes lived in the ocean, and although 28% were initially surprised that microbes have genetic material, a majority (71%) when told this fact found it reasonable and believable. **S2 and S3** required explaining to the public the link between microbes and genetics. Prior work [153] found the general public had a limited understanding of basic genetic terms and concepts, suggesting that visitors would have difficulty with **S2** and hence **S3**. For this reason, the team decided that the main feature of the exhibit should be **S1**, an animation of microbial behavior with familiar descriptions.

4.6.2 Prototype 1

Our scientific partners at UH helped identify which stories could be told from their data. One of the stories within the dataset was that microbes have functions that are on a daily cycle. Based on a previous study [150], we focused on this daily cycle of microbial functions, which we predicted may give visitors a more familiar entry point in to the metagenomics data.

We collaborated with our scientific partners at the UH to create a visualization with somewhat familiar representations to visitors **(C2)** and to simplify the complexity of the data **(C3)**. From our discussions, we created a model of microbial interactions that could best tell **S1** and offer some familiarity for visitors. The model simulates a 24-hour period showing the functions each microbe performed during this time. The objective of Prototype 1 was to see if visitors could follow **S1**. If they were able to do so then we would try to layer in **S2** and finally **S3**.

We chose to have a central animation based on prior work [131] indicating that animations could be effective for conveying these concepts to novices, **C2**. However, animations, when designed for teaching those with varying domain knowledge, require varying the presentation forms to be effective in achieving a desired learning task [131, 132]. We decided to layer the three related stories in an exhibit, we sought to present each story within a form, starting with **S1** as an animation. We chose **S1** to be the focus of the animation since of the three stories it could have the most familiarity with visitors [150] and, as the first story, it is the foundation upon which the other stories are built. This animation would serve as our entry point [154], and be the centerpiece to attract visitors to the exhibit. This would also provide us with the opportunity to determine if a "well-designed" expository animation actually contains all the elements needed to draw the learners' attention to the right place at the right time [130] allowing visitors to quickly decode (**C3**) and understand **S1**.

With Prototype 1 our intent was to have a minimalist animation. Our animation (Figure 4.3a) was driven by our curated model of microbial interaction and portrayed microbes as icons interacting with one another. This animation had three main elements: (1) Four main microbial characters (Figure 4.3c), (2) A background that transitioned from light to dark blue (Figure 4.3b) and back over a set period of 45 seconds (**C3**) to illustrate a 24 hour period, and (3) Control panels that described each of the four microbes (Figure 4.3a). The control panels were static and only provided textual information about each microbe. We included a simple visitor interaction of tapping on the control panel to inject viruses into the pool. Our goal was to communicate **S1** using an animation that showed the functions of microbes during a 24-hour period informed by metagenomic data.

The evaluation for Prototype 1 sought to determine if visitors could interpret this first story from the animation. The exhibit was placed near other exhibits that focused on microorganisms. One was an exhibit on the microbes that live in the termite gut, with a live microscope view of these microbes. The other was a Winogradsky panel that shows microbial diversity and discusses energy production. While there was not a designed exhibition, this context seemed to best support the content of *Sea of Genes*.

To recruit participants for the evaluation, an evaluator stood near the exhibit and approached every third person as they walked passed a predetermined imaginary line near the exhibit. If the systematically selected visitor was with a group, the whole group was invited to participate as well. Consenting visitors were asked to use the exhibit. Because it was an early prototype and not all of the labels or the touch interactivity was implemented, the evaluator verbally described these aspects to the participants:

Evaluator: *This exhibit shows how microscopic life behaves in the ocean. The water changes color from dark, for the night, to light blue, for the day. You can release different organisms into the water to see what they do. The touch is not working yet, so just let me know which organism you want, and I'll release them for you.*

When participants indicated that they were done, the evaluator asked the visitor who interacted the most within their group a set of questions designed to gauge usability and comprehension. In this evaluation, we talked with a total of 38 visitors[1] over the course of four days.

[1]The demographic breakdown of the evaluation participants was: 18 adults, 12 teenagers, and 8 children, with 20 females and 18 males.

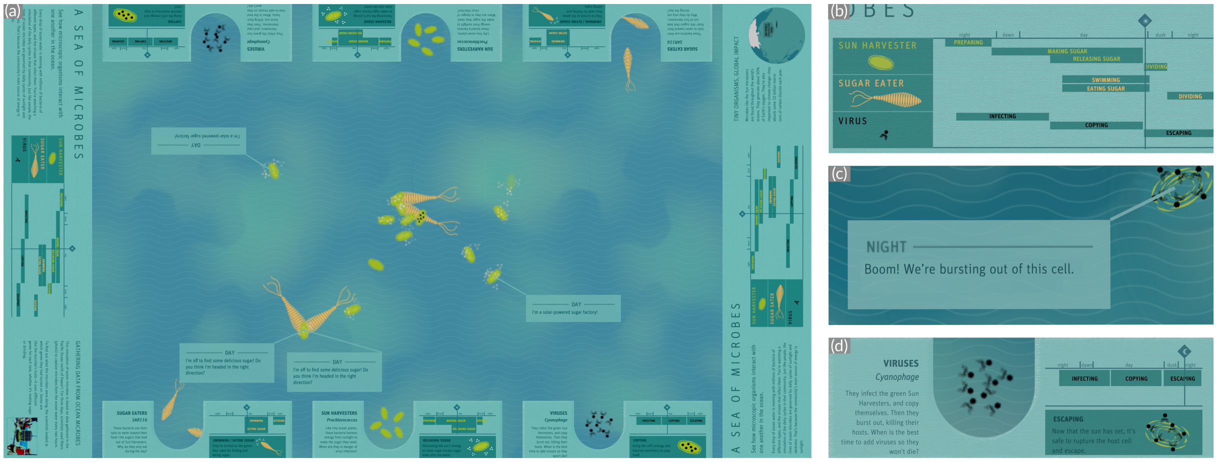


Figure 4.4: (a) Sea of genes Prototype 2's control panel and interpretive label. Panels are placed on two sides of the table to allow for more visitors to interact. Microbe annotations orient to the corresponding direction based on their position. (b) The timeline chart depicts an overview of what each microbe is currently doing and will do in the animation. (c) Annotation describes the microbe's current behavior. (d) Control panel describing the microbe and a timeline chart showing its current active behavior.

When we asked visitors what they found interesting, over 55% of them talked about the interactions between microbes in general, with a minority (6 out of 38 visitors) mentioning a specific interaction shown in the animation, for example:

Visitor03: *[It was interesting] infection of Cyanophage to Prochlorococcus, and SAR116 eating. How it [infected Prochlorococcus] burst and a bunch came out.*

Visitor37: *One thing was SAR116 and Prochlorococcus working together [was interesting]. [I] noticed the SAR116 were going by Prochlorococcus and right after the Prochlorococcus were making sugar.*

However, a majority (58%) of the visitors found parts of the animation confusing. There were multiple reasons visitors had difficulty interpreting this animation. First, a few visitors complained that in the animation, the microbes' small size and thus poor resolution made it difficult to distinguish one type from another, e.g.:

Visitor01: *Some stuff was really small, so you couldn't see what was happening*

Visitor17: *These [Prochlorococcus] in particular appear to move randomly. A better graphic representation would be helpful.*

Second, visitors could not decipher parts of the animation to make sense of microbial behavior, e.g.,:

Visitor24: *No idea what those guys [Crocosphaera] do.*

Visitor29: *Didn't understand what SAR116 were doing.*

This was particularly the case when microbes appeared and disappeared as they were born and died, e.g.:

Visitor01: *Some would disappear, especially this one [SAR116]. It was hard to see why they disappeared.*

A smaller number (29%) of visitors noticed differences between day and night, even though the evaluator described the transition at the beginning of their exhibit use.

Visitor38: *The light and dark. Seeing the difference between what's there and what thrives in the light versus dark.*

Our evaluation of this prototype showed visitors were interested in the interactions between microbes. A few noticed behaviors such as infection and eating. The majority, however, could only glean that there is a microbial community but may not have discerned the specifics of the behaviors or relationships. This may have been due to the presence of too many unique animated elements; Pylyshyn and Storm [155] showed people can only track up to 5 independent moving targets accurately. Rather than notice the individual unique animations between microbes, the excessive amount of moving elements may have led to it being processed as one entity. Thus, we hypothesize this prototype fell to the Gestalt principle of Common Fate [156], which states humans perceive visual stimuli that move in the same speed or direction as parts of a single stimulus. Furthermore, processing both this visual information and decoding what it means may have distracted visitors from paying attention to the background color change. We needed to improve how we portrayed our microbes to make clear the interactions of interest and focus visitor attention. Yet, this evaluation indicated that the animation was able to convey that microbes interact with one another, enough that 55% of visitors talked about it, the first aspect of **S1**; however, it drew too much attention, resulting in few visitors seeing or discussing

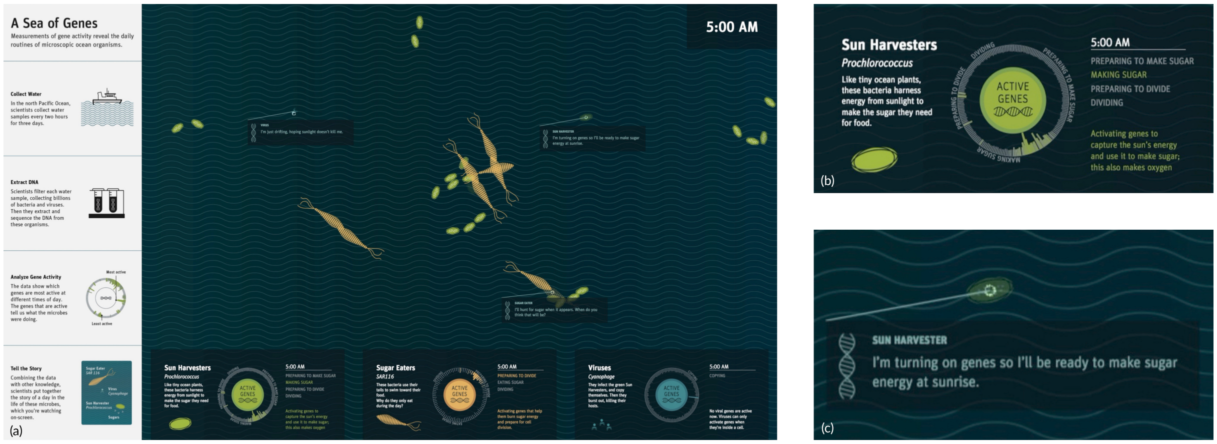


Figure 4.5: (a) Final design of the Sea of Genes exhibit. (b) Legend contains information about Sun Harvesters and in the center is the activity gene showing the genes responsible for making sugar is being expressed. (c) Annotation showing what the Sun Harvester is doing.

the daily aspect. Based on the evaluation data findings, it was clear that we had to improve our animation to better support visitors' interpretation of the microbial functions and their daily rhythm.

4.6.3 Prototype 2

For the next version, we wanted to (1) improve our animation presenting **S1** by making it easier for visitors to interpret and (2) layer on **S2** by visually communicating that gene expression going on and off is what drives the microbial functions seen in **S1**. Furthermore, for this version, we elected a more designer-driven approach for presenting the narrative. That is rather than let the visitors independently navigate the visualization and discover stories on their own, we sought to have more control in actively guiding visitors to the stories. Because these changes would introduce more information to decode, we had to carefully revise the animation to convey the additional information without overwhelming the visitors [36, 130]. To accomplish these tasks we focused on the following elements of the exhibit.

- Appearance of microbes and their behavior **(C1 and C3)**. Visual designers worked closely with the UH scientists and exhibit specialists to define how microbes and their behavior would appear in the exhibit (Figure 4.6).
- Control panel for interaction and interpretation **(C2 and C4)**. Central to interactivity was a control panel with a "well" of microbes that visitors could drag into the exhibit. The

control panel also described the creatures and a timeline chart that tracked the timing of activities seen throughout the day (Figure 4.4a).

- Annotations: New text was added to focus visitors on relevant information and reduce time finding what to observe **(C3)**. These annotations also conveyed that microbial behavior and relationships follow a daily pattern (Figure 4.4b).

To improve the interpretation of each story, we sought to reduce noise and confusion by both lowering the number of microbes and improving the quality of assets and animations [15, 123]. Some visitors who saw Prototype 1 reported having trouble identifying the microbes. Therefore, the microbe community was reduced from 4 to 3 and the assets of each microbe were changed to be more realistic compared to the previous version, **(C3)**. The simulated ocean animation now only contained three microbial types: SAR116, Cyanophage, and Prochlorococcus as shown in Figure 4.6. We created non-scientific names for these microbial characters to reinforce their functional role in the ocean ecosystem and provide some familiarity to our visitors [157, 158, 159]. "Sun Harvesters" was the name given to Prochlorococcus, a microbe that makes energy from the sun. "Sugar Eaters" was the name given to SAR116, a microbe that lives on sugars produced by other microbes. "The virus" was the name given to cyanophage, a virus that infects Prochlorococcus. To reduce the amount of information visitors needed to process, we also showed a fewer number of microbes in the overall animation.

The previous prototype relied heavily on visitor participation and engagement; specifically, they had to invest time in interpreting and navigating the animation to discover **S1**. This dependency we formed on visitor participation conflicts with **C1** so we chose to pivot in a different direction. In this prototype, we wanted to assert our narrative and lessen the time it takes to do so **(C3)**. We added annotations to direct visitors to the stories **(C3)**. Annotations have been used effectively in several studies of information visualizations [15, 160, 161, 162] to add information, convey meaning, show data provenance, represent uncertainty, and highlight points of interest for users. Annotations also strengthen the narrative by drawing attention to aspects of the story we want to tell **(C3)**. Our annotations would pop up and highlight a microbial action (Figure 4.4b) delivering a characterized message of what microbial function was occurring as well as reinforcing when it occurred. This reinforcement aligns with the theories [131, 132] about the learning benefits of multiple forms of representation. The characterized message was also designed to both anthropomorphize the microbe and highlight key interactions. In marketing, anthropomorphism has been shown to have a positive and significant influence on personal value [163]. By providing human-like characteristics to the messages we theorized that visitors would engage more and process the narrative quicker **(C3)**. These annotations would be triggered when an observable function occurred. Only one would be triggered at a time to not overwhelm the visitors **(C3 and C4)** but to guide them through the story.

Lastly, we updated the control panels (Figure 4.4a) by simplifying the text and providing a timeline chart to highlight the temporal aspect of microbial behavior. This version included all microbes in the simulation and not just viruses. The timeline was included to enable visitors to see the entire daily cycle for each type of microbe and provide context for what was occurring in the animation. An indicator, synced to the internal animation clock, would slide across the timeline to both reinforce the time and highlight what function each microbe was performing **(C2 and C3)**. Although animation implicitly illustrates time [129], we needed to convey to the visitors the repetition of similar behaviors during the 24-hour period.

The evaluation was conducted with 21 museum visitors recruited near the exhibit prototype[2], following an evaluation protocol similar to that of Prototype 1. However, in this evaluation, the evaluator did not describe anything about the exhibit since all the exhibit labels and touch interactivity were implemented. Instead, visitors were invited to use the prototype however they saw fit.

This evaluation found that visitors noticed the microbial interactions; when asked what they thought the exhibit was trying to show, 71% mentioned microbial interactions, **S1**, with 86% of visitors mentioning at least one microbial interaction when asked what they saw in the exhibit. These findings suggest that a majority of the visitors understood aspects of the first layer of the story: Microbes interact with each other, **S1**. But, they continued to struggle with noticing the daily cycle: Close to half of the visitors said that it was difficult to distinguish between day and night in the animation, and only one-third of the visitors mentioned a specific temporal pattern in microbial activities. This was despite the addition of the timeline chart and emphasizing

[2] There were 15 adults, 5 teenagers, and 2 child participants. Nine were male, and 12 were female.

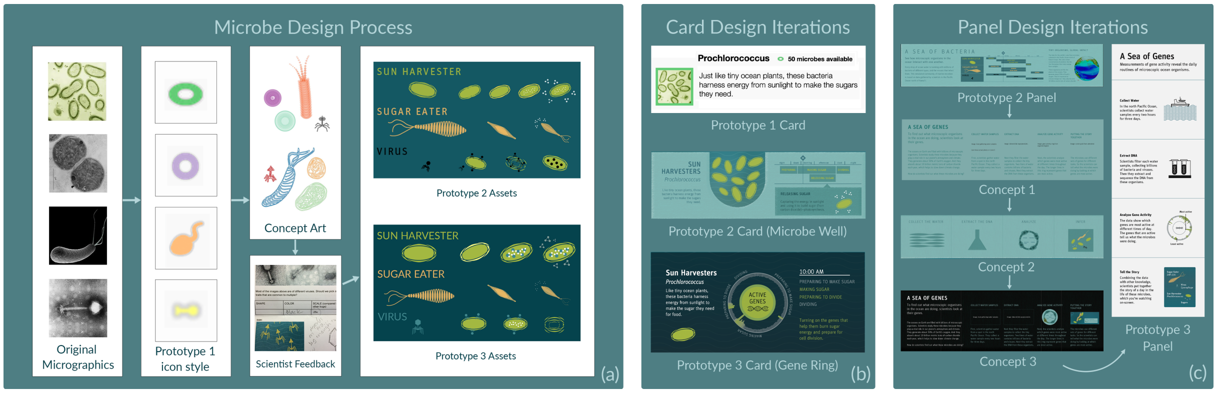


Figure 4.6: (a) The design process of transforming the microbes into the characters of our stories over all prototypes. Initially, we used icons to represent each, in the further iterations we used sprites. The final image depicts SAR116, Cyanophages, and Prochlorococcus in their final iterated state. (b) The card design changes between Prototype 1 to Prototype 3. (c) Various design iterations of the side panel from Prototype 2 to Prototype 3.

temporal patterns in both the control and the label.

The third layer of the story, **S3**, was only partially conveyed. Although 62% reported thinking that the animations were based on real data, a third of that group thought what they saw in the animated ocean was a representation of what researchers see. For example:

> Visitor08: *They went on a boat and collected it in a bucket and put it in a petri dish and put it under a microscope and looked at it.*

These results suggest that the first story, depicted through the central animation, was communicated clearer due to the additions of anthropomorphized annotations and asset improvements. We suspect that annotations helped visitors decode more readily, made the unfamiliar more familiar, and drew their attention to the salient parts of the visualization. Visitors no longer needed to decipher what the role of "procholoroccus" was and instead could observe the "Sun harvester" explain simply what it was performing.

For our next version, we sought to further improve how we convey **S2 and S3**. Specifically, we sought to help visitors track temporal changes such as noticing microbes perform different abilities over a period, which many did not readily notice. And, we needed to better highlight the underlying meta-genomics and meta-transcriptomics data, **S2**.

4.6.4 Prototype 3

For the final iteration of *Sea of Genes*, we focused efforts on sharpening the communication of **S2** and **S3** by emphasizing connections to the underlying metagenomic data. Our evaluation of Prototype 2 showed that although the prototype had fewer distractions relative to Prototype 1, **S2** and **S3** were largely unnoticed. The changes in Prototype 2 made the exhibit more effective in communicating **S1**.

We changed the orientation of the exhibit and reduced the number of panels around the table to focus attention on the animation. We hypothesized that having a single orientation for the exhibit would make it easier for visitors to decode the visualization **(C3)**. Fixing the orientation appears to contradict our **C4**, supporting multi-user interaction. However, an evaluation of a similar tabletop visualization, Plankton Populations [108], found visitors tended to use one side even though it supported multi-orientation use. In line with **C2**, we moved the label to the left side of the exhibit and removed all complex graphic elements [157, 158, 159]. The new label (Figure 4.5a) told **S3**, expressing how the data was collected and how the representations in the visualization were linked in four steps: (1) Collect Water; (2) Extract DNA; (3) Analyze Gene Activity; (4) Tell the story.

To emphasize that the animation is based on genetic data, **S2**, we made the following additions: A large title that included the word genes (A Sea of Genes) (Figure 4.5a), the legends and microbe annotations were adapted to refer to genes (Figure 4.5b), and the iconic DNA helix was added to the middle of the gene activity wheel and the annotation boxes (Figure 4.5c). We modified the design of the timeline chart to emphasize **S2** and passively support **S1**. The new representation would display the underlying genomic data as a radial histogram (Figure 4.5b). We considered using a circos visualization, as they have been used by the New York Times to supplement stories on metagenomics and comparative genomics [164], however, it would only increase decode time (violating **C3**) and we could not assume our visitors would be familiar with that representation **(C2)**. Rather, we designed a visualization where each gene related to a particular behavior was grouped around the circumference of the circle (all of the photosynthesis-related genes under "Preparing to Make Sugar"). Each gene had an activity range and was displayed as a dynamic histogram; the length of the bar was determined by the normalized amount of transcript in the sample. So, through the course of 24 hours, different areas of the circle would have waves of gene activity, similar to an equalizer.

Part of the success of conveying **S1** in the previous prototype may have been a result of the interactivity we provided visitors, i.e., the ability to add microbes to the animation. We chose to shift the story the interaction emphasized from **S1** to **S2**. Rather than adding a microbe, tapping on the revamped card (Figure 4.5b) would highlight all microbes of that type and show their gene activity, Figure 4.5c.

Prototype 3 was evaluated as part of a larger summative evaluation conducted by an external museum evaluation group, Inverness Research. This evaluation sought to find the key understandings that visitors came away with from their interaction with the exhibit. While conducting the study, Inverness found that visitors were not spending sufficient time at the exhibit and therefore would not be able to answer the questions in their exit interview. So, they focused their efforts on mediated interviews, in which visiting groups were recruited to interact with the exhibit and answer questions immediately afterward. A total of 13 mediated interviews were conducted. Visitors' time with the exhibit ranged from 2 to 89 seconds, with 53% staying 10 seconds or less (30 out of the 56 observations). The interview results reported 2 of the 13 groups understood that scientists collected data while on boats and that something about what they collected is represented on the screen of the table. However, there was no clear evidence that visitors understood that the data represented was gene expression. Furthermore, the mediated interviews indicated that visitors could not figure out what to do or where to begin, and were often confused. For example, seven visitors said they were not sure where to start or what to do:

Group A: *Honestly, its way over my head. I'm interested in what's happening, I just don't know what to do and I can't understand it. I guess there are Sun Harvesters and Sugar Eaters? There is a lot of empty space. I keep waiting for something to happen. My instinct is to ask how do I make it work so I can learn something? But I can't make it work. So there are three types of genes?*

Despite our efforts to design the exhibit to emphasize **S2 and S3**, visitors were still extremely confused by what was presented. The additions of annotations, improving the exhibit, changing to a single orientation, clearer animations, better assets, and fewer visual elements only seemed to help bring visitor attention to **S1**. One reason, we believe, is usability problems made it difficult to convey the last two stories. That is, visitors did not understand what their role was with the exhibit and did not feel they could participate with it. The time spent figuring out their role **(C3)** resulted in visitors leaving the exhibit before learning anything deeper than aspects of **S1**.

4.7 Discussion

This project was a collaborative effort between several parties in an effort to develop a functional exhibit for deployment. With real-world collaborative projects, time and resources are often limited, which adds significant constraints. We worked with exhibit designers, who have a deep understanding of conveying scientific information to the public. Our scientific partners at the University of Hawaii ensured the accuracy and fidelity of the data and their representation. A design and data-visualization firm from the industry both expedited the process and offered its own insights into design and development. Every prototype was tested with real targeted users. The entirety of the project itself was built over the course of 2016–18. From our significant evaluation effort over the iterative design process of developing *Sea of Genes*, we identify relevant usability issues and areas of future work.

With this work, we introduced a set of considerations that should be addressed when designing narrative visualizations for an informal learning environment. We believe applying these considerations, as we have shown in this chapter, can support future designers attempting to visualize complex data for museum environments. At the time of development, we looked over literature [108, 109, 106, 1, 66] for relevant techniques to apply to develop a successful exhibit under our considerations; **(C1)** free-choice learning environment, **(C2)** public comprehension, **(C3)** readily decipherable, and **(C4)** multi-user friendly. We found limited research that met such an intersection and applied what we could, which gave us some success. With more time and resources we would have rigorously examined each technique in narrative visualization under these considerations. However, in a real word setting of developing an exhibit where time and staffing are constrained, such analysis was not possible. We studied available narrative frameworks to identify concepts relevant to our work. We now present a discussion on the intersection of narrative visualization and museum design.

Storytelling Frameworks: Segel and Heer [1] present a set of structures for balancing "designer"-driven vs "visitor"-driven narratives. Stolper et al. [66] provide an updated discussion of narrative visualization strategies with a focus on systems with an "author"-driven predefined narrative. These structures have been applied and shown to be effective in a variety of situations. However, these structures need to be viewed and evaluated under our considerations **C1–4**. Testing these structures and developing frameworks and methodologies that cater to "visitor-driven" would aid the museum community. Most analysis has been done on settings (e.g. online journalism) where the user does not have distractions or free choice. Visitors need to be engaged, and with other settings like the Exploratorium, exhibits need to support interaction and multiple users. There is a clear need to aid exhibit designers through further exploration and research into this space.

Here, we examine three structures: Martini Glass, Interactive Presentation, and Drill-down for our museum exhibit design considerations.

The Martini Glass structure for narrative visualization allows for directing visitor attention explicitly to a set of points before releasing them with an understanding to make inferences for themselves. However, with **C1** and evidenced by Boy et al. [52], it is difficult to assume they will get past the "designer"-driven direct messaging and reach the exploration part. A study [122] conducted at the Exploratorium examined if a narrative introduction could better contextualize the exhibit and found it had no real advantage over not-including it. This introduction was a slideshow presenting where the dataset came from and its scientific significance, similar to **S3**. Furthermore, under **C4** if the exhibit is in the exploration state, then new visitors are missing key information that allows them to truly participate, excluding them from this experience.

Next, the Interactive Presentation structure allows for an individual to progress through the story when they are ready to do so and allows them to repeat steps. This structure, however, does not allow for multiple people to follow along **(C4)**, in the sense of allowing any visitor to step forward or backward, which could disrupt others' experience. This constraint could be addressed by transforming the presentation into a looped animation. The loop could also allow visitors to follow along back to points where they missed, ideally allowing for understanding at their pace. However, even with using an animated loop, as the animation advances new information is continuously presented. While visitors have started understanding a scene the animation has progressed, introducing new material to decode, leading to either confusion or frustration with the exhibit, as seen with Prototype 3.

The Drill-Down approach appears to have the most promise but there is still a constraint on what you can train a visitor to do and expect **C2**. The structure . However, there are several challenges with this structure when attempting to construct an exhibit experience with intuitive interactions that can be received by the majority of visitors. There is a delicate balance of presenting either new visualizations or materials in the sub-views without overwhelming a visitor with content. Then there is the additional challenge of ensuring that the exhibit is accessible to multiple visitors (i.e., if one visitor is drilling-down, it will not interfere with another visitor's experience). Furthermore, depending on the relationships between stories this structure may struggle since it requires a central story to reach all other sub-stories. Treating stories as graphs, and documenting what kind of graphs each of these structures can present under **C1-4** is a direction that could merit great value for presenting in such conditions. How it handles such stories as *Sea of Genes* with strong parallel themes between stories, **S1 and S2**, is an open question.

Stories as graphs: Narratives are predominantly linear and are most effective in conveying a single perspective. According to Spiro and Jehng [165], *"linearity of media is not a problem when the subject matter being taught is well structured and fairly simple. However, as content increases in complexity and ill-structuredness, increasingly greater amounts of important information are lost with linear approaches."* Can we create visual narratives that permit multiple perspectives and allow different narrative flows? As Hullman and Diakopolus [15] point out conveying a point of view requires careful over-emphasis. It is well known that multiple perspectives are needed to learn complex topics [166]. This study also calls for research on narrative visualization. It highlights the importance of considering the motivation, target audience, and narrative structures when developing exhibits. We do not have good frameworks for classifying story complexity in a manner that can inform visualizations. A simple narrative has one causal pathway and is unidirectional. How do we characterize structures that are more complex? A taxonomy may allow us to develop visualization techniques. Effective storytelling is a subject of interest to a diverse group of researchers in social sciences, computer science, and biological sciences. A taxonomy may allow us to map findings from these disparate domains and develop theories and guidelines. One approach could be to use graph theory. Here milestones, events, or information would be nodes, and connections or flows represent edges. A simple story is a unidirectional planar graph with no branches. Let us call this a basic graph. In our narrative, we had a more general network. The activities of an individual microbe (e.g., sun-harvester preparing sugar) are close to a basic graph. The activities of these microbes interacting (e.g., "preparing sugar" & "eating sugar") create a more general graph. Since these connections occurred in parallel and all share the same time dynamics we have a directed graph that represents **S1**. The role of genes, **S2**, however, changes the structure of the graph. We could view genomics as nested information. Embedded in each node corresponding to an event (e.g., "preparing sugar"), there were genomic data (e.g., time of expression & amount of transcript). In our visualization, we present the embedded data in a narrative that was occurring in parallel in a separate space. That is, we presented genomic data in a dynamic histogram on the bottom of the screen separate from the animation in the center while both update in parallel. Our limited success in effectively connecting these two stories for the visitors highlights the need to consider other visualization techniques for these graphs. In short, we contend that there is a need for a richer taxonomy.

4.8 Conclusion and Future Work

Reflecting on this endeavor we find there is space for further research at the intersection of storytelling, data visualization, and informal learning. The current storytelling structures are effective in many situations; however, delving deeper into the union of exhibit design and narrative visualization could extend the current structures, introduce "visitor"-driven methodologies, or offer adjustments and considerations to "author"-driven methodologies.

The process of communicating scientific findings as multiple stories visually in an informal learning environment brought many challenges. We need a better understanding of how to construct readily decipherable visual abstractions of a complex science, while maintaining scientific authenticity and accessibility to the public. If the abstraction was too simple they didn't understand or notice the science, when the science was emphasized they were confused. Communicating multiple related stories is another challenge. We need to ensure the underlying connections between each story are reflected visually. In our final iteration, each story was embedded in a unique encoding; the community of microbes **S1**, a radial histogram **S2**, and a side panel **S3**. A visual link between these three was not explicit enough to be received by visitors. We had some implicit clues, such as when a visitor touches a microbe a small radial histogram inside the microbe appears that correlates to the larger one. At the time of development, designing a visual link was not considered. Our entry point into the exhibit was the community of microbes. Yet, from the entry point to the two ancillary visualizations, there isn't an explicit visual cue for a visitor to follow. We lacked in our design a visual encoding that functions as a through-line for our stories. In other words, there should be a visual encoding to reflect a common theme or consistent element within our stories. Perhaps such an encoding would help visitors continue to the other stories beyond the entry point. This requires the stories are not disjoint and have a minimum of one factor in common.

We reviewed our process for designing *Sea of Genes*. From our reflection, we believe there is clear value in sharing experiences and lessons learned. Overall this study supports retrospective analysis of design work in new cross-disciplinary domains even if the desired goals were not met. Theories are typically shown to work in their documented space; however, there is value in reporting how these theories behave when tested outside of these documented spaces. We hope our extensive case study will stimulate additional research in approaches for visualizing complex data from unfamiliar domains for the public to explore in physical settings including museums and visitors centers.

Chapter 5
Character Oriented Design for Visual Storytelling

From my past work pairing storytelling and data visualization, I sought to delve deeper into the relationship between the storyteller's intentions and the story by exploring the role characters play in data storytelling. When telling a data story, an author has an intention they seek to convey to an audience. This intention can be of many forms such as to persuade, to educate, to inform, or even to entertain. In addition to expressing their intention, the story plot must balance being consumable and enjoyable while preserving scientific integrity. In data stories, numerous methods have been identified for constructing and presenting a plot. However, there is an opportunity to expand how we think and create the visual elements that present the story. Stories are brought to life by characters; often they are what make a story captivating, enjoyable, memorable, and facilitate following the plot until the end. Through the analysis of 160 existing data stories, we systematically investigate and identify distinguishable features of characters in data stories, and we illustrate how they feed into the broader concept of "character-oriented design". We identify the roles and visual representations data characters assume as well as the types of relationships these roles have with one another. We identify characteristics of antagonists as well as define conflict in data stories. We find the need for an identifiable central character that the audience latches on to in order to follow the narrative and identify their visual representations. We then illustrate "character-oriented design" by showing how to develop data characters with common data story plots. With this work, we present a framework for data characters derived from our analysis; we then offer our extension to the data storytelling process using character-oriented design.

5.1 Introduction

Information, at times, can be abstract and intangible, which may lead to difficulties in communication. The beauty of visualization is captured in its ability to make the intangible tangible, the invisible visible, and the inaccessible accessible. Through visualization, we can utilize visual representations to embody complex and often large datasets, reveal hidden insights about both known and unknown phenomena, and afford a means to showcase findings as well as share insights with broader audiences. We, as data storytellers, are concerned with presenting these findings to large audiences. Stories and visual storytelling have been shared and consumed by our earliest ancestors. Some of the earliest forms of visual storytelling [32] played a role in communicating where rich sources of food can be located or where to avoid dangerous beasts. In visualization, we have utilized storytelling for a variety of communicative needs since it is effective for engagement [167], memorability [168, 169], and showing casualty [29].

As data storytellers, we play a role in capturing and sharing the wonder we see in data with others. In our stories, we are challenged to emphasize the scientific insights of our content and simultaneously engross [34] the audience with our narrative. The challenge of ensuring our content is both consumable and enjoyable while preserving scientific integrity constrains our story design. These constraints can result in the audience having a difficult time understanding [53, 52] insights, topic relevancy, or where in the story to focus.

In data stories, numerous methods have been identified for constructing and presenting a plot. A story plot [29] is a narrative of events, with the emphasis falling on causality. The data storytelling process [37] can be viewed as three stages — identification, organization, and presentation. Typically, the first step resolves in the accumulation of a set of events (i.e., "story pieces"). These pieces are often the insights derived from either the collaborative efforts of data analysts and domain experts or the automation leveraged by statistics [170, 171]. The collection of events is guided by the shared intent of the author and analysts, which is the intention they seek to convey to the audience. This intention can take on many forms [26] (e.g., to inform, to educate, to entertain, or to explore) and centers the story. Next, in the organization stage, several narrative frameworks [37, 1, 66, 27] can assist us in sequencing these events into a cohesive story plot. During this stage, we need to ascertain several properties about these events, namely their relationship to one another and their ordering. We should end up having a structured outline of what we want to convey and the sequence in which to present them. Lastly, we have the presentation stage, where we give the look and feel to the story. There are many methodologies [49, 15, 172] at our disposal for tailoring our story for the target audience. However, it is within the presentation stage that there is an opportunity to expand how we view and design the visual elements that act out our story plots.

In our work, we are interested in data-driven, visual storytelling, particularly the characters that bring them to life. Data storytellers want to create rich experiences that evoke an emotional response, draw the audience in, and leave them with something to remember. Stories can be brought to life by characters; often they make a story captivating, enjoyable, memorable, and facilitate following the plot until the end. In other media, characters are often used as a bridge for the audience to cross into an unfamiliar and perhaps complex new worlds [30, 31, 29]. For example, in *Star Wars: Episode IV – A New Hope*, Luke Skywalker, the protagonist, is the bridge that leads the audience into the *Star Wars* world. Throughout the story, the audience learns more about the setting and the rules of this world (e.g., force) through his behaviors, rather than from a list of terminology and definitions. Through the lens of characters, the audience can gain an understanding of a world without prior knowledge. We are inspired to investigate the possibility of applying characters to convey scientific insights in data stories. A deeper understanding of data characters could address open data storytelling opportunities [34].

This work seeks to address the following — what a data character is, how we classify characters in data storytelling, and a space for how we can develop a data character and apply it within a story. Through the analysis of 160 existing data stories, we present a framework for data characters, where we identify three fundamental character roles: **main**, **supporting**, and **antagonist** characters. With this character-oriented design space, we further investigate how "conflicts" are contextualized in data storytelling, as they drive the narrative and can elevate a telling of a story [30, 29]. Notably, we find that designing an identifiable central character can support the author in aligning the story pieces with their intentions, arranging the sequence with a consistent message, and delivering this message to the target audience.

We consider our primary contributions are:

- a framework [1] for data characters that extend to the data storytelling process;
- a summary of storytelling terminology derived from a variety of storytelling and visualization literature as well as an assessment of data characters in the current literature; and
- case studies in various data story genres to demonstrate the applicability of our design space.

5.2 Data Character Framework

In this section, we describe the development of our framework for data characters that were derived through discussions with storytelling experts and from the analysis of existing data stories. In subsection 5.2.2, we present our framework for how we classify data characters and the behaviors we identified. We then describe a space for how we can develop a data character and apply it within a story. We refer to this space as *character-oriented design*, to be elaborated in section 5.3 and section 5.4.

5.2.1 Framework Derivation

To better understand the nature of data characters, the forms that they assume, and their behavior in data storytelling, we created a corpus of 160 data stories and analyzed these stories using a codebook we developed. The corpus was made by merging several other corpora [173, 63, 174, 175] and updated to include newer stories. The merged corpus size was reduced based on repeats and any stories that were no longer accessible (i.e., required Adobe Flash Player or went offline). We further filtered down the set to primarily visual stories. In these stories, the content and core messages are communicated via data-driven visualizations, animations, or videos. We excluded those heavily dependent on multiple paragraphs of text, where the visualizations served as annotations or figures rather than the driving narrative force. Newer stories were

[1] https://chaorientdesignds.github.io/

sourced from either accredited data storytelling sites (e.g., Bloomberg) or the Information is Beautiful awards [176]. We also added data stories that addressed minority domains in the merged corpus.

To develop the codebook, which we used to analyze our data story corpus, we consulted with experts in literary media and data storytelling, including published authors. Through multiple open-ended consultations and discussions, we gained an understanding of their process of character creation and character development. These insights laid the foundation for our codebook, including identifying a fundamental set of character roles and their effects on the narrative. We then shared the codebook with two separate groups of data visualization and data storytelling experts, whose feedback helped further translate and contextualize these insights for data storytelling. In each session, we exchanged our findings and translations from the literary consultations. Each session took approximately 2–3 hours. This process enabled us to broaden the dimensions of the codebook. For example, we include a new type, real-world photos, for identifiable central characters, to be illustrated in subsection 5.2.3. We also expand the types of **antagonist**, to be elaborated in subsection 5.2.5. The visualization experts and data story practitioners validated the finalized codebook. Using the codebook, two researchers independently coded the corpus. They met for three sessions to compare and discuss any mismatches until reaching a consensus. The finalized codebook is presented on our website [2] along with our corpus.

From our analysis, we derive that the role that visual elements and visualizations serve in data storytelling falls into two states: (1) *given an existing visualization, how can we adapt it to emphasize and explain a finding* [170] and (2) *given a finding, how can we visualize it* [177]. Both states share the communication goal of data storytelling; however, they present two different starting points for what we view as characters. The former has an existing character that will be developed and altered to depict the plot, whereas the latter starts from scratch. This work delves deeper into the former. We investigate how the 160 data stories would be framed with data characters. We find that the constant involvement of a data character in all the story pieces can support the audience in seeing and tracking continuity in a story. Future work may inves-

[2]https://chaorientdesignds.github.io/

tigate concrete strategies for maintaining and presenting the connections between story pieces (i.e., through-line) in a story, including how to elicit an emotional connection to the stories.

5.2.2 Character Roles & Behaviors

In written storytelling [30, 31], there are many character roles in addition to the main character, such as a deuteragonist or a love interest.

In the simplest form of a story, there would be a protagonist, a character that drives the plot forward, and an antagonist, a character that stands in direct opposition to the protagonist [30, 31].

For simplicity, in this work, we only focus on just three essential roles: **main**, **supporting**, and **antagonist** characters.

It is documented that a main character need not be the protagonist, as often there are other characters that can advance the story, even the antagonist is included [31]. In a data story, the protagonist could be the audience, as they interact with data stories and can be the ones that drive the story plot forward, whereas the main character remains a visual element in the story. Future work may investigate the nuances and interplay of characters and potentially data character roles.

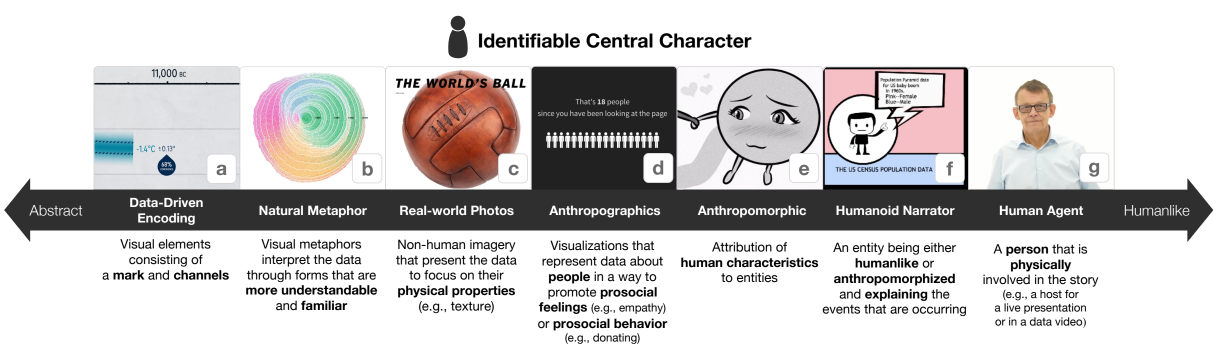


Figure 5.1: Identifiable Central Character (ICC). In character-oriented visual data storytelling, we require a character that is central to the story and can be visually identified by the audience. From our analysis, we found this character's visual depiction can range from an abstract representation to a person. (a) a visual encoding such as a line to present the temperature [178] (b) Tree rings [71] showing the immigration (c) Soccer ball photos to show the texture progression in panels [179] (d) the anthropographic people [180] denote the loss to COVID-19 (e) a character that represents productivity from a data comic [181] (f) a humanlike character explaining the content [65] (g) Professor Hans Rosling [182] starring in a data video.

5.2.3 Main Character

The core of our data characters is a concept. While concepts are often intangible ideas, let alone visual, this cannot be the case for the **main character (MC)**. After all, data storytelling is a visual medium. The **MC** is the device that the audience can rely on to make sense of and contextualize what is transpiring; therefore, the **MC** should be visually present in the story. More precisely, we need a visual element that can be identified by the audience as the center of the story, to which we refer as an **identifiable central character (ICC)**. The ICC becomes the vehicle that visually navigates the audience through the story plot to the conclusion. After reviewing our corpus of data stories, we find the visual representation of the **MC** (i.e., potential ICCs) ranges from an abstract representation to a real person. As shown in Figure 5.1, we present the common forms we identified from our analysis.

The ICC can be any visual encoding, visualization, or set of visual elements, so long as it is central to communicating the core message.

It is up to the author on what visual representation will suit their narrative needs the best. For example, in the data story about the U.S. immigration from Cruz et al. [71], the ICC is a natural metaphor, as shown in Figure 5.1b. The story begins with the authors revealing this data-driven natural metaphor without any context; the audience thus may not be able to understand the inherent meaning.

When the story begins, our **MC** starts in the “ordinary world” or an expected state [31]. In this initial state, the audience should be primed with the relevant context and understanding to decode what is visually presented. The introduction of the character must get across the relevant context, and background information as well as decoding information to relate the visual element to what they mean. Furthermore, it sets up the motivation and purpose of this character. It should also introduce doubts, the uncertainty, and begin to challenge the character in its purpose. This context drives the data character’s actions and sets up expectations for the audience. For example, the U.S. immigration story opens up by introducing a large **tree ring** with a description of how it represents U.S. immigration. However, the explanation for what the colors mean is initially left unanswered, prompting the audience to scroll onward to learn more about what they are seeing.

The behaviors of a character are driven by its desire to reach a personal goal [31]. The desire is determined by the author by factoring in their intent, the through-line, and the story pieces. The lens of the audience into a story is through the **MC**. The **MC** has a **desire** that it seeks throughout the course of the story. For simplicity, the other two characters either aid the **MC** in attaining its desire or are in opposition. Those in opposition will bring the **MC** to change and evolve throughout the story, and the audience will learn what those changes mean. In the case of Figure 5.1b, the **MC** may desire to *inform* the audience about population growth and demographic changes over the years in the United States. As the audience scrolls onward, the authors introduce more characters in the story, illustrated in the following subsection.

Although it is ideal for a **MC** to also be the sole identifiable central character, ICC, there are exceptions. With other forms of storytelling, technically, there can be identifiable characters that are central to the story (e.g., *The Lord of the Rings*). It is important to be cautious when having multiple ICCs in a story, as they will compete for the attention of the audience. This may result in issues with the message of the data storyteller being delivered effectively. Introducing more characters will increase the complexity of the story, but depending on the communicative goal, it can be effective in communicating the intended message.

5.2.4 Supporting Characters

A **supporting character (SC)** brings out dimensions of the **MC** and helps push it towards its desire. **SCs** must have some relationships with either the **MC** or the **antagonist**, but do not require a relationship among themselves, as shown in Figure 5.5. For simplicity and the sake of clear story illustration, we recommend not to have a **SC** that supports both the **MC** and the **antagonist**; however, it is possible (e.g., the betrayer in storytelling). The **SC** needs not actively work towards supporting the **MC** to pursue its desire; rather, it should never intentionally impede the **MC**. We found **SCs** in data stories are often tasked with providing missing context, extra information, and even alternative representations of the data.

Continuing with the story of U.S. immigration, we left off with the story introducing the **MC**, a data-driven natural metaphor of a **tree ring**. As the story advances, the story authors introduce other characters that offer *extra information* and *alternate representations* to help the audience gain deeper insights and understanding about the **tree ring**. As shown in Figure 5.2,

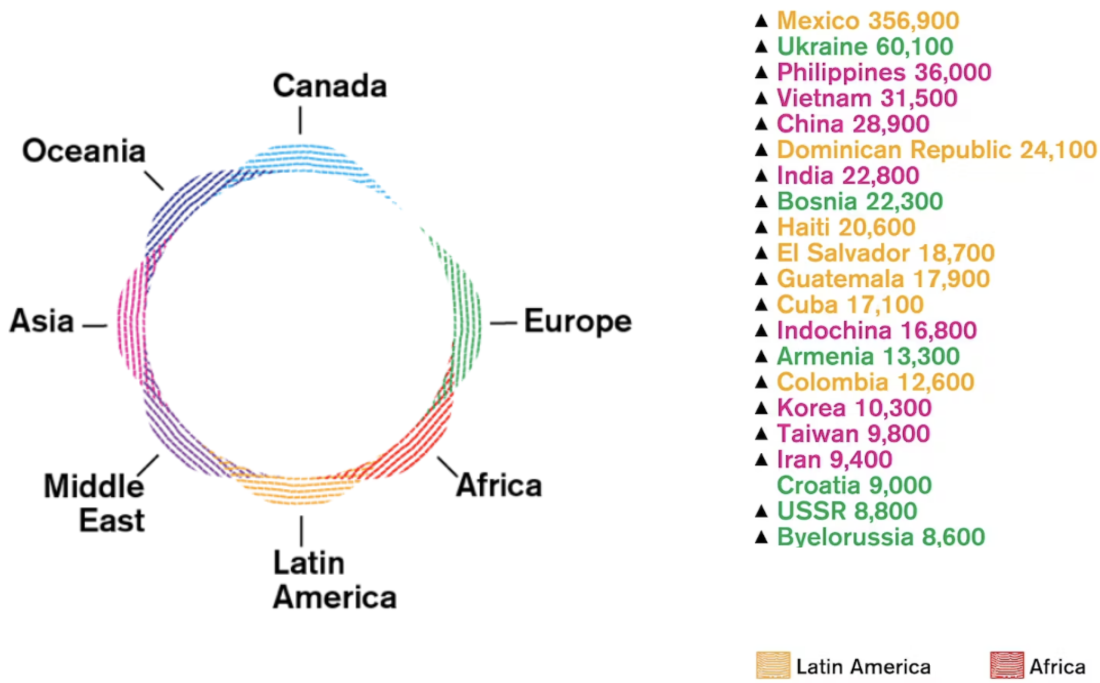


Figure 5.2: An example of two **SCs**

. The **MC** is a natural metaphor illustrating the U.S. immigration [71]. The **SCs** of this story include a legend and additional details of where immigrants are emigrating from. These **SCs** provide extra information for the audience to better understand the **MC**.

we see a **legend** that decodes the visual encodings, allowing the audience to understand what the color means and infer more insights about the **tree ring**. The **tree ring** is also accompanied by a **list of locations**, providing extra information on the geographic origins of individuals immigrating to the U.S. These devices help the audience understand and interpret our **MC** as the story progresses.

5.2.5 Antagonistic Character / Force

The difference between an **antagonistic force (AF)** and an **antagonist character (AC)** is that an **AF** is an ethereal presence that exists but is not seen directly, whereas an **AC** is visually present in the story. We found that when the **antagonist** appears as a visual character in data stories, it is often the **MC** driving the story forward. As shown in Figure 5.3-(d), there is a depiction of an **AC** in the data story “Out of Sight” about the U.S. drone strikes in Pakistan. The **AC** in this story are the grey lines that symbolize individual U.S. drone strikes.

If there is no **antagonist**, the **MC** will achieve its desire unimpeded. In the context of visual data storytelling, the **antagonist** needs to be present to stop the **MC** from pursuing its desire. *How can a visualization have a "villain" or "antagonistic force", and what does that imply?* In our framework, the **AF** can represent *misunderstandings* and *misconceptions*. A misconception is a mistaken belief or having the wrong idea. For example, it is a misconception to hold the belief that the earth is flat. A misunderstanding is having different interpretations of a meaning. As an example, thinking that a rainbow color map implies weather data would be a misunderstanding. Throughout the story, these **AFs** or the **AC** prevent the **MC** from achieving its resolution. In the context of the U.S. immigration data story, the antagonist is not visually present and therefore is an **antagonistic force**. The force manifests as misconceptions being hurled at the **MC** about a lack of context and understanding of where people immigrated from.

From our analysis, we find the **antagonist** in data stories often is a force. We further identify three forms that these **AFs** assume to represent misconceptions and misunderstandings — *lack of value*, *lack of context*, and *lack of trust*.

Lack of value. The **AF** creates external conflicts with the characters in the form of questions, such as "why is the design meaningful?" or "why do microtubules matter?", driving the story plot to address these questions. In a data story about colorized math equations [183], as shown in Figure 5.3-(a), the **AF** is constantly questioning the value of the design and its usability. As a result, the story attempts to motivate and demonstrate the design is effective.

Lack of context. The characters face external conflicts due to missing information, leading to a misunderstanding or a complete misconception. Often this kind of conflict is resolved by showing the "scale" of a phenomenon (e.g., the scale of loss, the scale of gun violence, or the scale of climate change). For instance, as shown in Figure 5.3-(b), the data story "Pace of Death" addresses the **AF** by depicting the number of people who passed away, in a given time interval, due to COVID-19.

Lack of trust. The **AF** introduces internal conflicts where the characters must prove or refute claims on their integrity (e.g., data authenticity, uncertainty, or credibility). In our analysis, we find this can be seen in the form of uncertainty visualizations or visualizations that delve into how the "black box" of a machine learning model behaves.

In all these instances, the **antagonist** is constantly attempting to challenge the **main character** and introducing conflicts to prevent the **MC** from reaching its desire.

5.2.6 Conflict & Tension

The rationales behind developing and understanding data characters include (1) structuring story content via characters helps filter out irrelevant information and leaves the data storytellers only with the content serving their communicative goal and (2) it unlocks the device of *conflicts*, akin to the debate in the context of science-based works. The difference between an explanation and a story is a conflict. Here, we illustrate more on internal and external conflicts.

Why do we want conflicts in a science-based story? Wouldn't that obfuscate, if not detract from, the messages we want to get across? The conflict is a device that helps the audience understand the motivation of the story and can lead to an appreciation for the endeavor or relevancy of what is being communicated. In other forms of storytelling, the conflict is used as a means to bring the audience into the story and to form an emotional connection to the characters, such that the audience cares or want the character to succeed in fulfilling its desire. From our analysis, we find the conflict assumes two forms, *internal* and *external* conflicts. Internal conflict exists when a character struggles with their own opposing desires or beliefs. It happens within the character and drives its development. External conflict sets a character against something (e.g., the **antagonist**) or someone (e.g., the reader) beyond its control. External forces stand in the way of a character's motivations and create tension as it tries to reach its goal. A majority of the data stories that we review tend to contain external conflicts where the story characters are fighting external forces to either refute claims or provide the missing context to remove a misconception or misunderstanding. The story resolves once the conflicts are addressed.

5.2.7 Data Comic – "Something's wrong"

To better understand data characters, their roles, and the conflict, we will go through a data comic and address the questions as follows.

- Who is the **MC**?
- Does the story have ICC? If so, what kind of ICC it is?
- Who are the **SCs**?

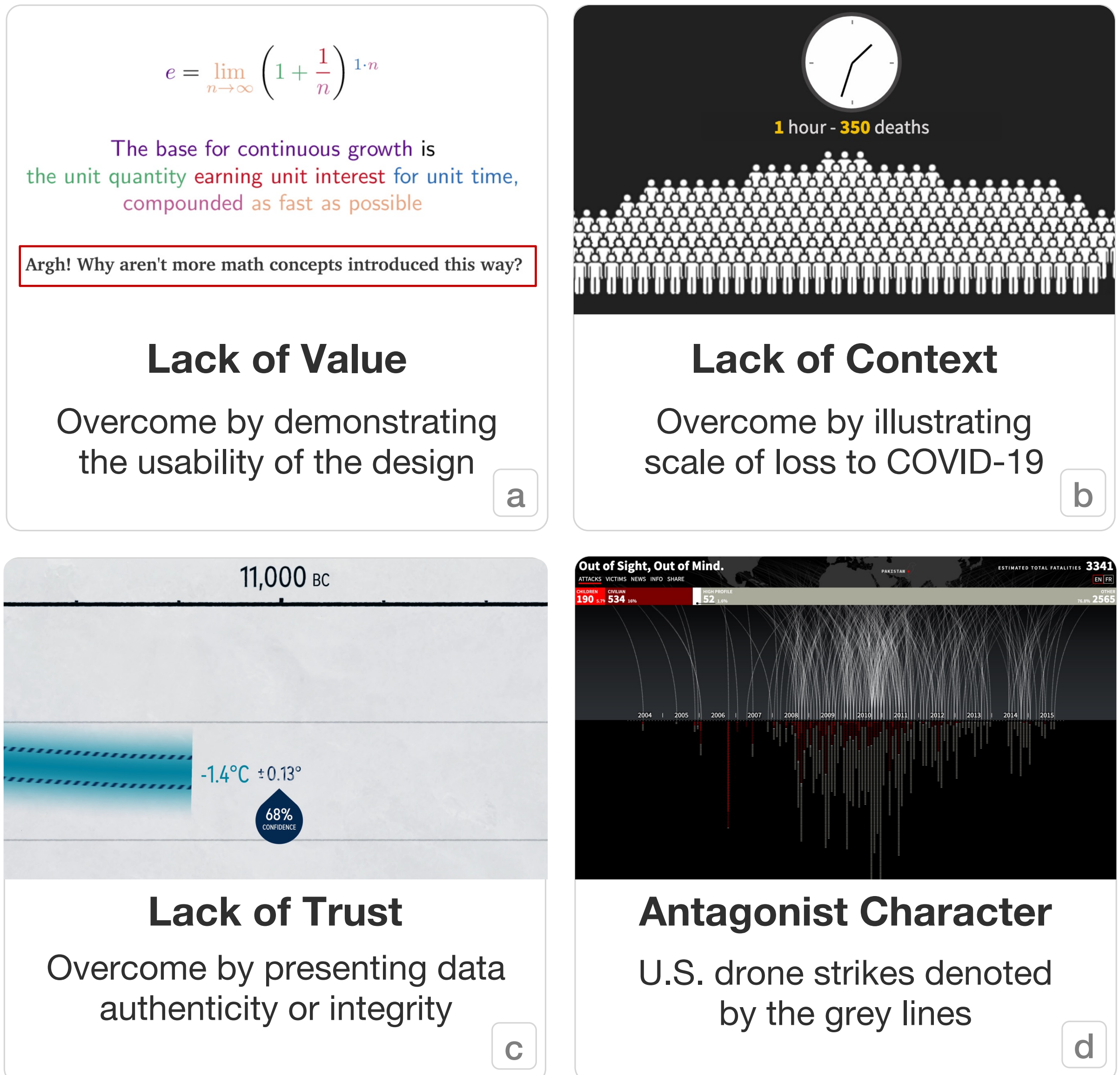


Figure 5.3: Examples of **Antagonist** Forces (a–c) and Character (d) in DS. (a) The story about colorized math equations [183] argues the design **value**. (b) The story "Pace of Death" [180] provides the **context** for how many have died from COVID-19. (c) The video "Degrees of Uncertainty" [178] attempts to demonstrate the **integrity** of the data analysis for climate change. (d) The story "Out of Sight, Out of Mind" [184] depicts the antagonist as drone strikes.

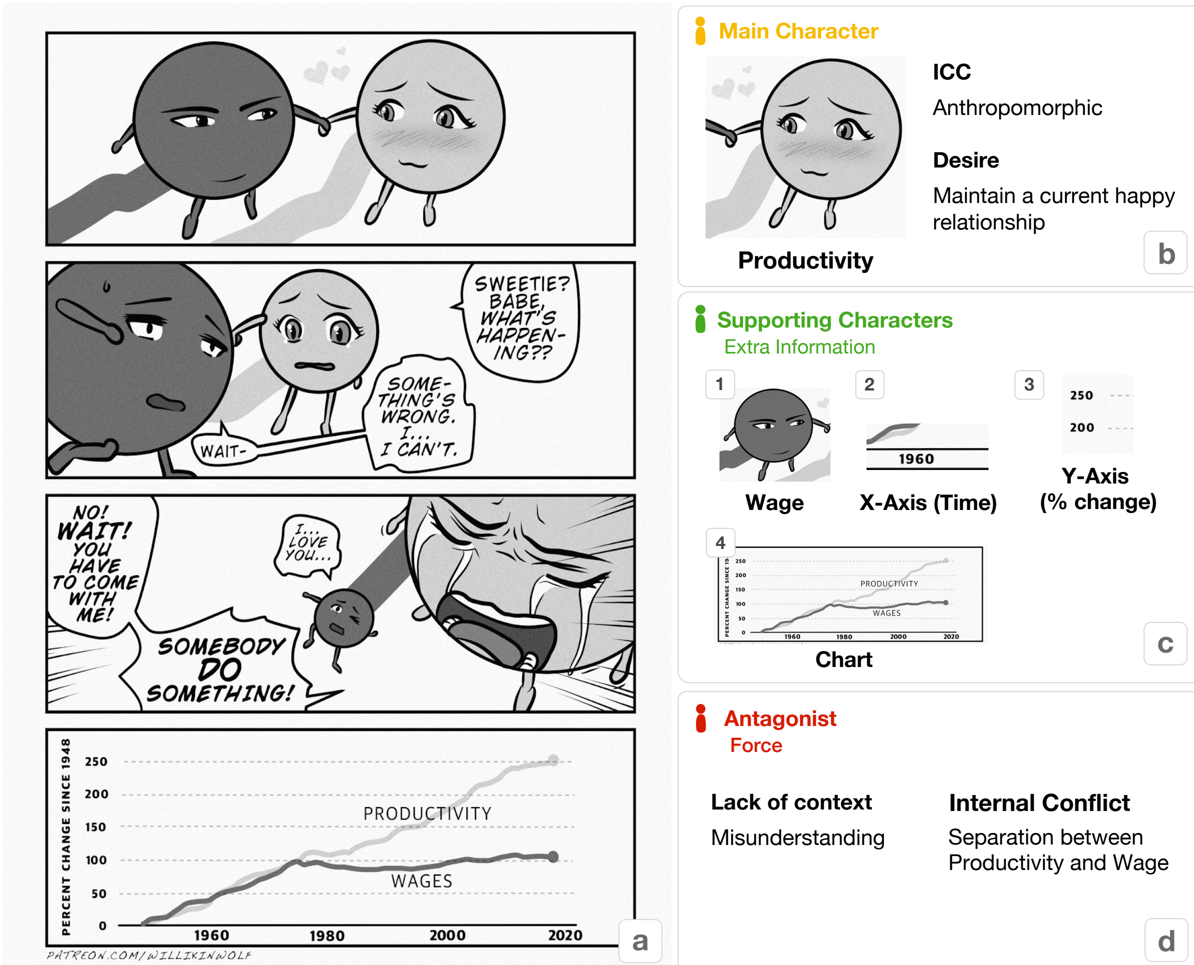


Figure 5.4: The data comic "Something's wrong", created by Willikin Woolf [181] and included in the

data comic gallery [185]. (a) Original data comic (b) The **MC** – **productivity** (c) The **SCs** – (c1) **wage**, (c2) **x-axis** with the year, (c3) **y-axis** with percentage changes, and (c4) the line **chart** with the same colors as **productivity** and **wage**. (d) The antagonist is an unaddressed force (**AF**), i.e., the reason behind the separation.

- What is the **antagonist**? How does it introduce the conflict?
- What is the relationship between the plot and the characters?
- What is the main theme, i.e., the through-line?

Data comic [28], inspired by the visual language of comics, is a rising and popular genre for presenting information effectively. We will analyze a data comic *Something's wrong* by Willikin Woolf [181], featured on the data comic gallery curated by Bach et al. [185]. This story depicts the relationship between two identifiable comic characters. The **MC** is likely the "productivity" character, as shown in Figure 5.4-(b). This is because from Figure 5.4-(a) panel 3, we see the panel focuses more on the despondent **productivity**, rather than the "wage" character. We may infer that the desire of **productivity** is to maintain the current happy relationship with "wage", as seen in Figure 5.4-(a) panel 1. The visual representation of the **MC** (i.e., ICC) is anthropomorphism.

The **SCs** are wage, the x- and y-axes, and the chart, indicated in Figure 5.4-(c1) to (c4). As shown in Figure 5.4-(a) panels 1-3, **wage** seems eager to stay with **productivity**. Meanwhile, the **axes** and the **chart** are only introduced in the last panel. The **axes** help us understand more about **wage** and **productivity** for providing the additional context of their percentage change over time, from similar trends to drastic separation. This is because the colors of both characters correspond to the line colors in the **chart**.

The **antagonist**, in this story, would be the reason behind this separation. Visually absent in the story, this **AF** remains unaddressed or unexplained either, at least on this page of the comic. The **AF** introduces the conflict to **productivity** in the form of a separation from **wage**. The through-line between all these events is to explain the relationship between **wage** and **productivity**. A possible theme could be to persuade the audience into thinking that these two characters should not drift apart and have a linear relationship.

5.3 Character Oriented Design Space

From our framework and analysis of data stories, we have unpacked a data character, identified common representations of data characters, defined properties of basic data character roles, and contextualized conflict in data storytelling. In this section, we describe how data characters, as defined in this work, can be developed as well as how they could fit into the broader data storytelling space. When we refer to this space as *character-oriented design*, we intend for it to be used as a guide for helping data storytellers in developing or identifying characters in their stories.

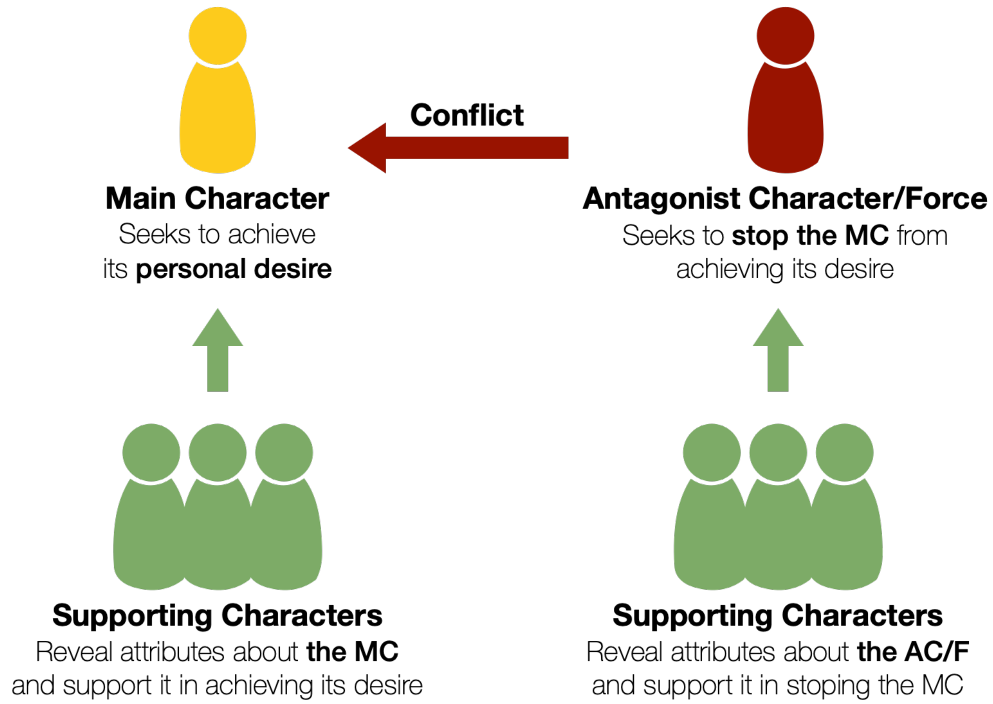


Figure 5.5: Character Web. The relationships between **main**, **supporting**, and **antagonist** characters. The main character (MC) has a desire and is the focus of the story. The antagonist force (AF) tries to prevent the MC from achieving its desire. Supporting characters (SC) reveal dimensions about either the MC or AC/F.

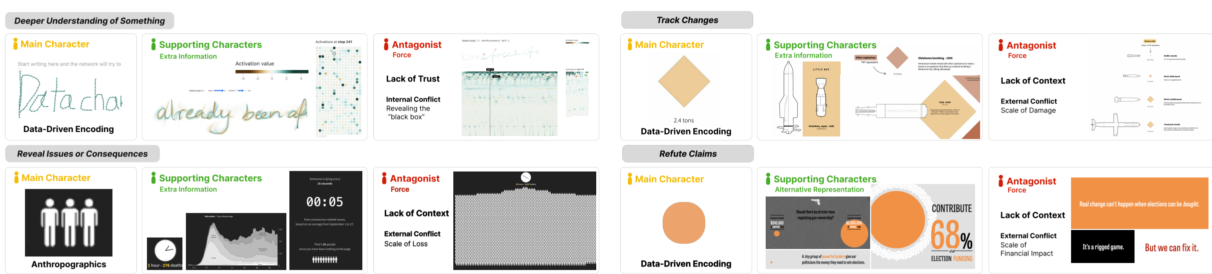


Figure 5.6: Plot Types and Data Characters. For four of the five plot types, we illustrate through existing data stories how data characters appear and behave. The first column represents the **main character** followed by **supporting** and **antagonistic forces**. Image sources: [186, 187, 180, 188].

5.3.1 Character Motivation

To begin, we first outline our process for character creation, which was also informed by the aforementioned discussions with experts. We consider a **character** as the lens for our audience to view the abstract, which often is a complex concept, we intend to convey. How a character is defined determines both the presentation of the data and the story focus. A data character is a derivation of a character, thus it must adhere to some core storytelling principles: (1) a data character must serve a narrative goal and (2) a data character must have a **desire**. We define **desire** as the rationale that motivates a character pursuing a goal. By assigning a desire to a data character, we should ensure the behaviors of the character are associated and align with this base desire, throughout the data story. That is, the character is consistent in its behaviors as it pertains to achieving this desire. Contextualized for data storytelling, this desire is the communication of some scientific concept, linked to an existing dataset with varying intentions. We offer the following steps as a suggestion for developing a data character and what their **desire** could be.

1. Identify the meta-concept (through-line).
2. Distill the concept into smaller more distinguishable aspects.
3. Relate these aspects to the data (i.e., via examples or non-examples).
4. Rank which aspects best illustrate the concept.
5. Select the visual representation.

To not lose sight of the original intention of the story, and to track whether our characters are in line with their desires, it is good to make use of a through-line (Table 2.1). As discussed in subsection 2.1.1, data stories can be composed of many plots and subplots, steering the original narrative in many directions. A through-line in data storytelling would be the core concept that ties these subplots together, providing guidance and relating back to the original narrative. Identifying a through-line requires the assessment of the story pieces for what common theme or over-encompassing plot could link these pieces all together. If we have five disjoint story pieces, then the through line has to be the commonality among the pieces. Alternatively, if we cannot find a common link between these pieces, we should then consider re-evaluating what we seek to convey. In our case, for data storytelling, this commonality is a scientific concept that each individual piece relates to. We can utilize the seven story plot types [26] along with the seven genres [1] to identify a central theme, through-line, for our story. As for the situations where certain story pieces address disparate concepts, one of them must be prioritized. The remaining groups should either be omitted [15] or be part of another story.

Premise & Through-Line: When organizing a data story, the first step is to identify what is the through-line between our story pieces. Namely, *what is the meta-concept that ties the messages we want to convey?* If we know our through-line, we may understand how our characters relate to the plot. Then, we can create a premise, or a small road map, of the entire story. A premise is the simplest expression of the story being told and presents a sense of the **MC** as well as the outcome of the story.

Possibilities: Once we have an understanding of the premise and through-line, we should identify what is possible in this premise. What types of genres do we have access to, what structures can we make use of, and what setting will it take place in?

Challenges + Problems: Identifying the unique challenges and problems ties to conveying the content of our data story. These problems could be the audience lacking familiarity with what is being expressed, the subject matter is too abstract, there are difficulties simplifying the science, and so on. We need to note what the main misunderstanding or misconceptions about the story content are.

After answering these points, we have developed a foundation for our **main character** and **supporting characters**. The next challenge is determining what should be our **main character** and how to develop this from our through-line and **antagonist**.

5.3.2 Character Creation

Naively, a data character inherits the properties of a data-driven visual element It would likely contain a set of attributes and behaviors that relate to explaining the concept. A character can be a lens into a meta-concept or scientific domain. Initially, a character represents an idea in a world full of misconceptions, misunderstandings, and uncertainty. The role of a character in the story is often to explore its relationship to this perceived world and attempt to overcome misconceptions, asserting its place. Contextualized for communicating scientific information, a data character would be born from data or namely analytics. As discussed in subsection 5.2.3, the motivation of a character is its desire, which would be the core message to get across. Thus, it is important to understand what concepts represent a data character. Depending on the complexity of the core concept, we often may rely on multiple characters to best convey it. This is due to concepts having many dimensions that should be viewed and expressed with varying lenses. The concept we prioritize and give the focal lens would be the basis of a **main**

character. Collections of smaller aspects that exemplify and illustrate concepts as a collection would result in **supporting characters**. Aspects from the data that may challenge the concept (i.e., contradictions) can lead to potential **antagonists**.

We suggest the best starting point is the meta concept that contains the smaller sub-concepts, i.e., the through-line. Once we identify the meta concept, we may create sub-groups of the findings to express more distinguishable aspects of the meta concept. We then can investigate how each sub-group illustrates the aspects of the meta concept. Precisely, we inspect whether each sub-group serves a direct example, or perhaps a non-example (i.e., contradiction), of the meta concept. These relations allow us to rank each sub-group based on their capability of demonstrating these aspects. Some sub-groups are stand-alone, some aspects may show details of phenomena in others, while some may introduce doubt or uncertainty. This gives us a basis for the **main**, **supporting**, and **antagonistic** characters. This can also be the point where data storytellers begin to design the visual elements that will portray these characters.

5.4 Characters in Story Plots

From our analysis of the corpus, we identify several patterns in how characters often behave in the stories. In this section, we present four concrete story examples to describe how the relationships of the character roles, as shown in Figure 5.5, could be contextualized in the story plots. Here, we first introduce the initial seven story plots for data stories, identified by Ojo and Heravi [26]:

1. Refute claims.
2. Reveal unintended consequences.
3. Reveal information of personal interest.
4. Enable deeper understanding of a phenomenon.
5. Reveal anomalies and deficiencies in systems.
6. Reveal information about an entity in increasing levels of detail.
7. Track changes in systems.

We may consider “phenomenon”, “entity”, and “system” as exchangeable main characters, depending on the communicative goal.

Consequently, we collapse the second and the fifth plot types into *reveal anomalies, issues, or unintended consequences*. We merge the fourth and the sixth plot types into *deeper understanding of something*, as these two story plots explore a character (e.g. phenomenon or entity) in depth. By factoring in the character roles, we integrate these seven story plots into the following five types:

1. Refute claims.
2. Reveal anomalies, issues, or unintended consequences.
3. Reveal information of personal interest.
4. Deeper understanding of something.
5. Track changes.

In the remaining section, we describe how data characters may be woven into each of these story plots, including a generic through-line, the desire of **MC**, **AF**, and a possible premise of the story plot. We focus on **AF** as we find it to be the majority of **AF/C** in our corpus, whereas **AC** is rather intuitive due to its visibility. While these examples do not represent the only way to construct such a story plot, they illustrate how we may pair characters and story plots together to organize the story.

Refute claims. The through-line can be to persuade the audience into believing the inverse of the claim. The **MC** desires only to prove the inverse of the claim. The **AF** may take the form of misconceptions or misunderstandings that support the claim. Consequently, the conflict, when resolved by the **MC**, helps refute the claim. *Premise:* the **MC** heads towards the state, where the world is opposite of the claim. The **AF** causes conflicts to prevent the **MC** from reaching its desire. By overcoming the conflict, the **MC** fulfills its desire and refutes the claim.

For example, we examine the story, *Money Wins Elections*, as shown in Figure 5.6. The through-line is a claim that "money wins elections", with other concepts addressing corruption in the U.S. government.

The **MC** is a **point mark** that represents a vote, where its size encodes financial investments. We may infer the claim the **MC** seeks to refute is that "all votes are equal" or that "nothing can decide election outcomes", as the **MC** desires to persuade the audience that money can buy election results. The **AF** is the misconception that causes external conflicts. The **MC** resolve the conflict by demonstrating that the scale of financial investment can influence the election outcome.

Reveal anomalies, issues, and unintended consequences. The through-line is theme-dependent and affected by the disposition of the author towards the consequences or unexpected events. The desire of the **MC** is to maintain the world state as expected. The **AF** introduces an action or event that creates conflict with the world **MC** expected. While the **MC** addresses this conflict, the result causes the **MC** to deviate from the direction it expected.

Referring to Figure 5.6, we can see an instance of this type of story, *The Pace of Death*. The through-line, in this case, is COVID-19, with a sub-concept of mortality rate. The **MC** is a person depicted by an anthropographic **icon**. There are several **SCs** providing extra information about the person, such as a **clock** illustrating how many people pass away after a period of time. The **MC** is initially under the impression that the mortality rate is not high. As the story advances, the **AF** reveals what the true mortality rate looks like. One of the **SCs** is a **timer** that reveals the scale of loss that has been incurred while the audience consumes this story; thus, we may infer the primary intent of the author is to terrorize the audience, in the sense of making them aware of the severity of COVID-19.

Reveal information of personal interest. The through-line is often to inform the audience about some information. The desire of **MC** is to bring attention to the personal interest, who is the author in this case [26], whereas the **AF** is to dismiss the attention. In this story plot, the **MC** tries to illustrate and bring attention to a topic. The **AF** introduces conflicts by casting doubt on why the topic is interesting. By overcoming the conflict, the **MC** persists in showing why the topic is interesting.

Deeper understanding of something. The through-line is to explain something (e.g., a phenomenon) to the audience. The desire of **MC** is to explain the mechanisms behind the topic as simply as possible, whereas the **AF** are misunderstandings about this topic. The **MC** attempts to explain to the audience, however, the **AF** introduces hurdles based on misunderstandings. To overcome the conflict, the **MC** must unpack the content further, until the author feels the story goal is achieved.

In the story, *Four Experiments in Handwriting with a Neural Network*, the **MC** is the **hand-**

written phrase of the user, as shown in Figure 5.6. The **MC** seeks to better understand how a neural network learns to the handwriting style. The through-line is about neural networks with sub-concepts in handwriting and generative models. The **AF** comes in the disguise of an internal conflict, the black box nature of neural networks. To overcome this conflict, the **MC** unpacks the black box, the layers of the neural network.

Track changes. The through-line is often to inform the audience of changes in a character (e.g., an entity or a system). The is the same for the desire of **MC**. However, **AF** presents obstacles to MC in the attempt of causing changes. In this story plot, the **MC** is trying to stay "unchanged" but is presented with obstacles, resulting in conflicts. To overcome these conflicts, the MC must change.

This story plot can be seen in the data story, *How powerful was the Beirut blast?*, as shown in Figure 5.6. The through-line is the devastation that an explosion causes. The **MC** is a data-driven **diamond mark** that represents an explosion measured in TNT equivalent (e.g., it starts at 0.01 tons). As the story progresses, the **mark** grows larger and larger, and the narrative changes to convey how different blasts compare. The **AF** for this story appears as a lack of context, a misconception of the magnitude of how devastating various bombs or explosive accidents are. The **MC** overcomes this conflict by comparing these incidents.

5.5 Discussion

The goal of our work is to illustrate how characters can be utilized to frame abstractions and communicate insights or findings through the story plot. Through our framework, we have identified specific features, relationships, and roles of data characters. From these findings, we describe a space for developing data characters and applying them in data stories, which we refer to as a character-oriented design space. Within this space, we hope to motivate data storytellers to view visual elements as characters with narrative goals, rather than data-bound abstractions to be explained. By treating visual elements as characters, we organize visual element(s) into those that are driving the story (**MC**), supporting the story (**SC**), and those contribute via contradiction (**AC/F**). We further investigate how data characters weave into representative story plots for data stories. The scope of this work is to offer a framework for understanding data characters and a design space to serve as the foundation for developing data characters. Our work introduces new considerations for the data storyteller: (1) they must develop the characters to best tell their story, (2) the roles (i.e., **MC**, **SC**, **AC/F**) for their characters, and (3) the number of characters needed to tell their story. In this section, we discuss the audience as data characters and character roles that are specific to data stories.

Audience as a character. There are many character roles outside the ones discussed in this work, and stories can utilize these roles in a myriad of ways that we did not address.

When data stories are interactive, it becomes more "reader-driven" as the audience is now a part of the story, often the protagonist [1]. A protagonist is defined as the character that moves the story plot forward. In this role, the audience is not always visually represented as a character, they may control the **main character** and even at times could be an **antagonistic force**, but may not be an ICC themselves. For example, the audience may play the role of the "devil's advocate" towards either the **main character** or the **antagonist**. In some instances, the audience can be visible and represented as an avatar or other ICC depiction, as shown in Figure 5.1. The audience can also serve as secondary key characters, known as a deuteragonist [30, 31], they can play a role akin to either the **main character**, **antagonist**, or even a neutral agent.

Data story specific roles. A potential direction for future work would be identifying character roles unique to data storytelling. A method that could be applied to identify such roles is character archetypes, as shown in Table 5.1, which are templates for generic characters based on certain types of behaviors and patterns. This device can give the storyteller guidance on the specific nature and behaviors that the **MC**, **SC**, or **AC/F** will take on. Archetypes differ in other storytelling media as they focus on types of people and the human condition. In novels and screenplays [30, 31, 33], some examples of archetypes include: father, queen, mentor, warrior, and lover. Often these archetypes come with strengths, inherent weaknesses, and understandable relationships to help build a story. However, often it is not a focus of data stories to discuss or explore the human condition.

From our analysis, we find a potential starting place for character archetypes would be systems, anomalies, entities, and phenomena. The properties of the data can suggest some behaviors. For example, hierarchy suggests depth, temporality may imply change, and spatiality

Character Terminology	Description
Character	A visual entity that influences itself and others and serves a narrative purpose.
Desire	The rationale that motivates a character's actions towards a goal
Conflict	Arises when a character while pursuing their desire faces an obstacle. The pressure that is applied to the MC forces change.
Archetypes	Patterns within an entity; the behaviors they exhibit that are essential in how they interact with others.
Main Character (MC)	**MC** has the central problem and drives the action in an attempt to solve the problem. It has a personal desire that it seeks to attain.
Supporting Character (SC)	This character role complements either the **MC** or **AC**. They provide a means for the audience to see more depth about the **MC** or **AC**. The desire of **SC** can align with either the AC or MC, but will not go against it.
Antagonist Character or Force (AC/F)	**AC** seeks to prevent the **MC** from reaching their desire. It causes conflict with **MC**. The story plays out when the conflict is resolved. It can be a character or an external force.

Table 5.1: Character-specific storytelling terminology. These terms and their mappings were derived from a breadth of visual and written storytelling literature [30, 33, 45, 29, 31].

could imply closeness or bonds. We may extend this thinking by taking data types into account, such as nominal, categorical, numerical, and their pairings. Multiple numerical datasets could contrast with one another. We discuss three potential data character archetypes; *overview*, *parental*, and *cluster*.

The archetype of the *overview* would be a character that knows the broad picture of the relationships between other characters, but need not give an opinion, similar to an observer. An example can be a visual analytic system for managing production lines. With stories centered around prediction, we can pull out a *parental* archetype, where the parent has multiple attributes presented as a formula to describe their child. The story could be interested in the parent-child relationship, what is the relationship's strength, how the child affects the parent, how external negative factors that the child faces also impact the parent, etc. This sort of archetype could be useful for explainable AI. Another potential character archetype would be *clusters*. A pairing of numerical and categorical can result in clusters. They are commonly used in visualization to denote relationships and close associations. The properties of a cluster can reveal commonalities, uniqueness, and how they are affected by the change (e.g., does the group stay or split).

While there are many character roles identified in other forms of storytelling, it is unclear how relevant these are to data stories. This work lays the foundation by providing a fundamental set of data character roles (i.e., **MC**, **SC**, **AC/F**), investigates how they weave into data stories, and discusses how the space of data characters could be expanded. We hope this work presents an opportunity within the space of data characters to extend past the roles identified in this work.

5.6 Conclusion & Future Work

Our goal with storytelling is to engage the audience while preserving the scientific integrity of the content. A data story should provide an entry point for the audience, a cohesive plot, and a cast of characters to lead the audience along to where the storyteller intends. As storytellers, we want the audience to create an emotional connection with the story and leave with the intended message. By exploring the role data characters play, we believe characters present the pathway to this goal.

We review 160 data stories and identify features of data characters, the roles they assume, types of antagonists in data stories, types of conflicts for data stories, and the relationships data characters have among one another, and offer a framework for data characters and design space for developing characters and applying them in a data story. From the perspective of a data storyteller, we show the role of characters and their importance in data storytelling with consideration of where character design should occur. We introduce the idea of an identifiable central character (ICC) as a device that data storytellers can use to select their **MC**, illustrated through our case studies, the relationships between the **MC**, **SC**, and **AC/F** in the context of data stories, and provide an outline for how characters can be woven with five types of data story plots.

For future work, we suggest exploring general patterns in common data stories in terms of broadening character roles. We believe, by providing a discussion on the relationships between characters and data-driven visual entities, we can achieve a more consistent language among designers. Stories require a through-line to connect all the story pieces together cohesively. However, these connections should remain clear to the audience. This would require some identifiable central character or characters for the audience to contextualize the presented information and latch on to as the story unfolds. The relationship between the plot and the story is not complete without the characters. The effective weaving of the two gives us the story. Our desire with this manuscript is for the readers to view and think in terms of characters when creating either a data story, narrative, or explanatory visualizations.

Chapter 6
VisActs Describing Intent in Communicative Visualization

In my previous chapters, I worked on different methods and frameworks for representing information and communicating scientific content to an audience through data visualizations as well as through data storytelling. In this chapter, I focus on the communication between the designer and the audience, namely, I am interested in developing frameworks for closely studying how a visualization designer's intent manifests in their design and visualizations. I developed a framework for describing designer intent in data visualization by translating relevant techniques from linguistics.

Data visualization can be defined as the visual communication of information. One important barometer for the success of visualization is whether the intents of the communicator(s) are faithfully conveyed. The processes of constructing and displaying visualizations have been widely studied by our community. However, due to the lack of consistency in this literature, there is a growing acknowledgment of a need for frameworks and methodologies for classifying and formalizing the communicative component of visualization. This work focuses on intent and introduces how this concept in communicative visualization mirrors concepts in linguistics. We construct a mapping between the two spaces that enables us to leverage relevant frameworks to apply to visualization. We describe this translation as using the philosophy of language as a base for explaining communication in visualization. Furthermore, we illustrate the benefits and point out several prospective research directions.

6.1 Introduction

Data visualization is a vast and growing field. In this chapter, we focus on the subspace of communicative visualization, a space that is concerned with the explanatory side of data visualization and is often what the average person is exposed to. In this space, the communicative goals can range depending on the designer's intentions and audience [189, 26, 190]. The role data visualizations play in communication varies where some designers use them to supplement their written and spoken messages, whereas others recognize them as an entirely effective mode of conveying the message [1, 66, 191]. Consequently, we are seeing wide usage of data visualization in industry and academia to communicate increasingly diverse and sophisticated messages. The complexity and diversity of interactive data visualization usage suggest that we could benefit from looking at it as a rich language. Developing frameworks from this perspective could then allow us to glean insights from naturally occurring experiments in practice and enable research that can guide future practice. Given the growing sophistication of visual communication, there is value in exploring the relevance to data visualization of frameworks developed by linguists and language philosophers. Some initial frameworks to examine are speech act and discourse theory; speech act because it distinguishes between the what is 'said', the intentions of the speaker, and the consequences of what is said in how it is processed by listening. We consider discourse theory as it examines how our communication is shaped by external structures.

Here the focus is on intent. Designers are tasked with creating and evaluating visualizations for targeted audiences. These audiences have varying motivations to engage with the presentation and with differing levels of prior knowledge of the subject matter. Designers have a wide array of intents. These intents range from journalists attempting to *inform* readers, teachers trying to *explain* concepts, scientists attempting to *discover* relationships among variables, policymakers hoping to *persuade* the public about the rationale for a decision, a blogger seeking to *evoke* strong emotions and activists hoping to get volunteers to *act*. How can we classify these intents in a manner that advances our ability to visualize data? Classifying intent is a prerequisite for determining if a visualization adequately satisfies the communicative intent of designers. Thus, to build and evaluate communicative visualizations we need a refined and principled language for describing communicative intent.

Recent work in communicative visualization by Adar and Lee [5] tackles the question of "how do we formally describe communicative intent in visualizations?" Their work offers an initial taxonomy for intents and enables an additional discussion on how to communicate information using visualization. Others have also identified the importance of intent. For example, Schoenlein et al. [192] note a central problem in visual communication is understanding how people infer meaning from visual features. All this points to a need to assess and understand if our communicative goals as designers are being correctly imprinted in our visual features as intended. We posit that when considering the question of "how can we formalize intents" with regards to visualization, we can draw from the philosophy of language, particularly speech act theory [193, 194, 195, 196, 197, 198].

Our work aims to link the field of visualization to the field of linguistics and demonstrates how doing so offers a broader perspective of the space and introduces new opportunities for research that can facilitate design. We illustrate the connection between these spaces by explaining the link between a sub-space of visualization, i.e., communicative visualization, to a sub-space of linguistics, i.e., speech act theory. We show how this relationship can help grow our understanding of communicative visualization. The insights and formalization developed there can guide us in developing a formal language for intent in visualization.

6.2 Data Visualization as a Language

There is an ongoing discussion on design as communication [10, 11, 12, 13] and there is a body of work [10, 12, 13] that gives credence to treating visual design as communication. In this work, we engage with this ongoing discussion and identify the implications and research directions that emerge from considering visualization design as a language.

Visualizations share many commonalities with language as it is also used to express what we observe to others. The goal of visualization, like ordinary speech, often goes beyond presenting facts to achieving actions or outcomes. In any case, how data is visualized can alter how the original context is perceived (e.g., visualizing the uncertainty in data [8, 9]).

Treating visualization as a language has been considered, although exploring the value of this association and what it affords is limited. Purchase et al. [14] have explicitly made these connections, and briefly describe the use of linguistic theory, namely pragmatics, to provide an over-arching framework for information visualization. They comment on the relationship between visualization and language and discuss how information visualization should rely on a multitude of theories rather than a singular theory. Hullman and Diakopoulos [15] study the application of linguistic-based rhetoric in narrative visualizations. Several others have presented theoretical visualization frameworks [16, 17, 18, 19] and implicitly imply that visualization is a language. They elegantly demonstrate how applying frameworks from spaces such as distributed cognition, information theory, an algebraic basis, or conceptual metaphor theory can contribute to the betterment and improved understanding of the use of visualization.

A vocabulary is the body of words used in a language. If we are to claim visualization is a language, then its vocabulary would be visual encodings. This association has been observed by Wilkinson [199], who identified general rules that govern the creation and presentation of data graphics and presented a structure within which these rules might be operationalized efficiently. He supposed that if the grammar is successful, it should be possible to reduce any data visualization problem into a graphic utilizing the rules outlined. Grammar-based visual encodings, as well as declarative languages [20, 21, 22, 23, 24, 25], arose out of a need to fluidly and precisely articulate a set of intents for communicating data. These works provide formalized approaches for describing charts, graphs, maps, and tables and give credence to treating visualization as a language.

In summary, researchers have recognized that visualization is a language and that it would benefit from formalizing the relationships to languages. If we are to treat the totality of visualization as a language and apply linguistic frameworks, we would have common ground for discussion and understanding of the properties of expressing visualizations, thereby facilitating the development of the field.

We present an approach for translating relevant theoretical frameworks from the space of linguistics into visualization. We develop a mapping between a subspace of visualization and linguistics *to illustrate* the potential for more work in this direction and immediate benefits. Our focus is on the intent of the designer. We propose a theoretical structure for both describing and assessing the intents of visualization designers and the respective consequences.

The motives of the designer of a visualization – to achieve actions or outcomes – and the impact of the visualization on perceptions, whether intended or not, should be considered while developing theoretical frameworks for studying visualizations. A framework to capture how we design interactive visualizations and their effects can be obtained by developments in speech act theory. Speech act theory, a sub-field of pragmatics and linguistics, studies how words are used to both present information as well as carry out actions. We describe a mapping of speech act theory into visualization and offer a theory on visualization acts, *VisActs*. This framework will be linguistically motivated using the foundation of speech act theory but made relevant for the space of visualization. That is, it must account for fundamentally how visual elements are processed and interpreted, which delves into semiotic theory. Furthermore, it must take into account both the conventional content and the underlying principles of data visualizations. Finally, such a theory should also offer testable predictions about the kinds of *VisActs* performed across the space of visualization. Particularly, it should offer the ability to assess how our intents manifest within visualization and their respective consequences.

Next, we delve into intent in visualization. Subsequently, we explain speech act theory and how it relates to communicative visualization. This is immediately followed by our introduction of *VisActs*, a translation of speech act theory contextualized for visualization researchers. We ground the relevance and application of this translation through a series of examples, followed by a discussion about our mapping.

6.3 Communicative Intent in Visualization

Communicative visualizations are created for a broad audience and represent the majority of visualizations that the public encounters. As stated earlier, communicative visualization occurs in a range of settings including journalism, education, museums, and public policy discussions. The audience differs in terms of their backgrounds, familiarity with the subject, and initial level of interest in the topic. This is in sharp contrast to visualizations that are designed for analysts

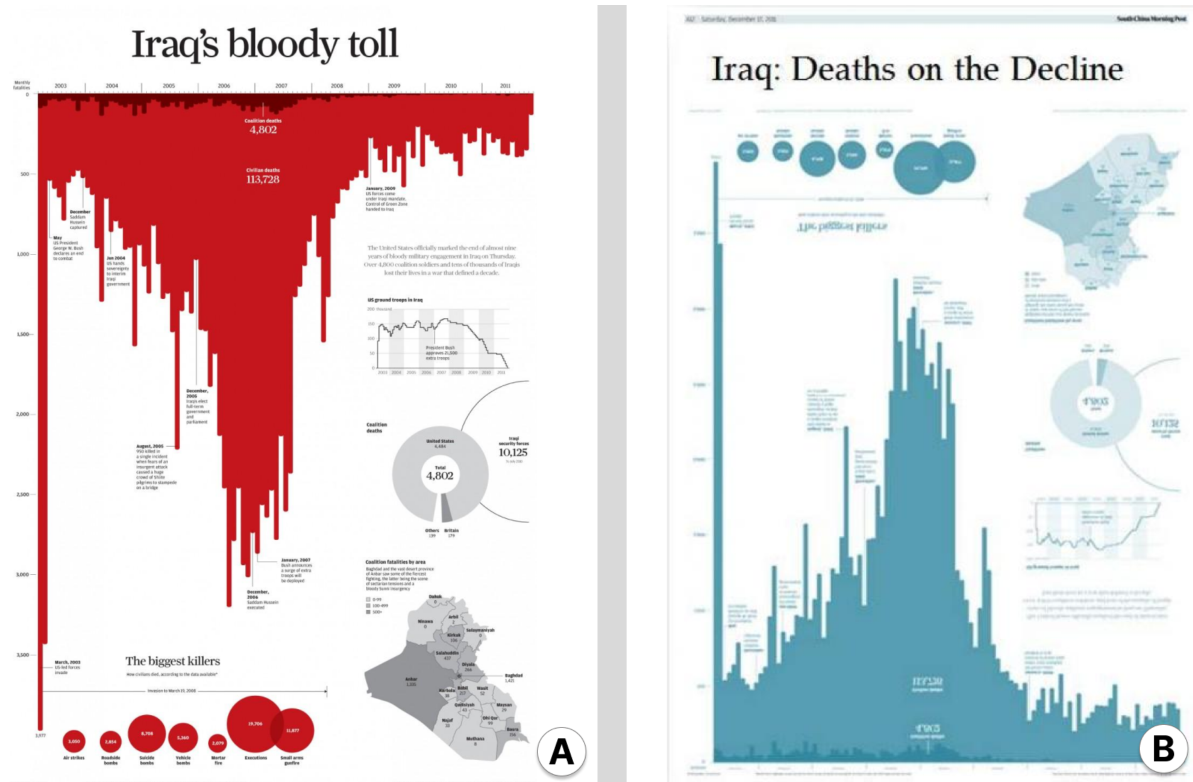


Figure 6.1: (A) Depicts a visualization, by Simon Scarr [200], with a possible intention to persuade the audience how devastating the Iraq War was. (B) This is a revision of the same visualization by Andy Cotgreave [201]. He adjusts the design so it is potentially received as more "neutral."

or domain experts, where the designer has an understanding of their audience's prior knowledge and has some assurance that the expert will use or attempt to use the visualization tools. Furthermore, the designer's intent is to provide visualizations that facilitate a set of specific tasks described by the experts. The diversity in the audience for communicative visualization makes it important to understand intent.

To start, we can consider intent as what the designer, or one who puts forth information, would like to be conveyed and communicated. The intent here would closely parallel the intent of a speaker in routine life. For example, if a child asks her mother while eating soup "is there any salt?", her intent could be to request some salt. However, what she stated may also be a query about the availability of salt. As this example illustrates, the intent of the speaker may not be perceived by the recipient and the desired outcome fails to occur.

Similarly, let us consider a visualization designer who creates a chart about war. One intent of such a visualization could be to terrorize [26] the audience in action against the war, for example, in Figure 6.1a the designer could produce a visualization that appears very dramatic and possibly violent. However, the same content, data, and chat with a different intent can be visualized differently as seen in Figure 6.1b where we can see a more "neutral" design with a possible intent to inform the public about the war.

In both situations, we want the communicative intentions to be received accurately. However, the diversity of the audience may make the intent of the speaker or the designer different from what is perceived by the receiver. The similarity in the communication challenges in these domains stems from the nature of the audience. Hence it seems fruitful to leverage the findings in speech act theory to inform design practices in communicative visualization. In the space of visualizations, intent has not been formally defined. In the following, we will offer some formalism. At a very basic level, it is useful to consider intent from two perspectives, user-intents and designer-intents.

6.3.1 User Intent

Dimara and Perin [202], when characterizing interaction in visualization, offer a definition of the term *data-oriented intent*. In their work, they describe interaction as a goal-oriented activity with a data-oriented intent. They survey the literature that describes intents from the perspective of a user [203, 204, 189, 205, 206, 191]. They find that the visualization literature classifies *intent* from the perspective of a user either as a goal, task, or problem. User intent could be to explore high-level data, gain insights, or gain multiple perspectives of data. The intent of a user can also be to collect and correct data, identify insights, or make decisions. In this literature, the intent has been described and identified at a low operation level (e.g., altering representations and collecting data) as well as at a higher level (e.g., information foraging, sense-making, and knowledge creation).

As designers, we tend to remove ourselves and our own intentions and in a manner treat ourselves as an outside entity constructing an interface from the perspective of a user to satisfy the user's intentions. Our research field has identified a variety of ways to adapt [207] our visualizations and systems to the intentions of the user. Designers spend a lot of time describing user intentions in terms of workflows and strategies and curating systems accordingly. A goal as a designer is to create interfaces that enable users to effortlessly express their intentions through the data.

6.3.2 Designer Intent

In the spaces of narrative visualization and data storytelling, there are many papers [66, 1, 122, 52, 37, 53] that provide frameworks and methodologies for communicating narratives. Although these papers do not explicitly define or identify the designer's intent, they subsume a diffuse concept of intent.

Bako et al. [190] assessed how designers utilize and apply data visualization examples by determining example usefulness, curation practices, and design fixation. Their work gives us methods for capturing the designer's intent as they begin the process of developing their visualizations. Often designers may not be able to articulate the form of what they intend to communicate. Examples are an effective way to express and collage together this form. Another type of designer intent that has been explicitly identified is artistic intent. Artistic intent often disregards functionality, making some works unintentionally incomprehensible. Lau and Moore [208] offer a conceptual model for incorporating these intentions formally. Intent could also have social motivations such as coordination, collaboration, or presentation to an audience.

Recently, Adar and Lee [5, 209] put forth a definition and a taxonomy for expressing intents in communicative visualization. To our best knowledge, their work is the only attempt to provide a formal classification of **intents** that are broadly applicable. They proposed a cognitive taxonomy in which they frame intents to be of the structure, "The viewer will [verb] [noun]." Verbs are selected from a specified set of cognitive constructs; nouns are from a set of specified knowledge dimensions. Their primary claim is that a good language for describing intents is the language of learning objectives. They assert that the advantages of using learning objectives are: "(1) being capable of describing objectives regardless of what the viewer wants; (2) allowing for a designer to create specific tests to validate if a visualization leads to a viewer achieving an objective; (3) finding out if the objectives are achieved both when a viewer is looking at a visualization and when it is taken away" [5]. A limitation of their work is that it restricts the intent of the designer to educate the audience. On the other hand, this is the only paper that provides some formalization of the designer's intent.

We seek to add to the discussion of designer intent by providing an alternative perspective for viewing intent in visualization and demonstrating how we can assess and analyze such designer intent with this perspective.

6.3.3 Challenges with Intentions

Our intentions can manifest in many forms in data visualization, especially as our communicative goals evolve and become more nuanced. Through examination of research [5, 210, 26, 192] that addresses the various types of intentions in communicative visualization, we highlight the following set of intentions to illustrate these forms; however, similar to spoken word, they are not limited to this set.

1. **Inform**: the intention is to have the audience be *aware* of a concept.
2. **Educate**: the intention is to have the audience *understand* a concept.
3. **Emote**: the intention is to *illicit* an emotional response from the audience. (enjoy, anger, sadness, etc.) from the presentation of the concept.
4. **Provoke**: the intention is to get the audience to *react* or *take action* to the concept presented.
5. **Discovery**: you are *obscuring* information on purpose so that people work for it and through that work, they *gain* some insight that can only be gained through this process.

It is known to be challenging, with absolute certainty, to derive an individual's original intentions behind an action [193] and likewise a data visualization. However, through other contexts, structures, and cues, it is possible to infer a close approximation of their intent. Linguistics has spent time studying both pragmatic and semantic structures in language as a means to accurately gauge intent, which has attracted interest from law practitioners as well as the NLP (Natural Language Processing) community. We hope frameworks for analyzing data visualizations, such as this one, can help the rapidly developing communicative visualization subspace and its relationship with ML4VIS.

6.4 Speech Act Theory Fundamentals and Terms

Table 6.1, contains the terminologies that we translate and contextualize for data visualization. This section will review our translation process and provide additional information on the ter-

Speech Act Theory Taxonomy	Description	Translation into Visualization
Fundamental Concepts		**A theoretical framework for describing visualization designer's intents.**
Locutionary Act	The utterance of a phrase. *What is heard.*	To show data. *What is shared.*
Phatic Act	An utterance of words which has meaning	The selection of data. "Data Act"
Propositional Act	The act of expressing the proposition, the content	Expression of data via analysis."Analytic Act"
Sentence Type	The type of sentence (e.g. declarative, exclamatory, etc.) has an impact on the force of an utterance.	The visualization type (i.e. informative, instructive, narrative, explorative, and subjective) has an effect.
Illocutionary Act	The utterance of a phrase with an intention. *What is intended.*	The design of visualization with an intention. *What is seen.*
Perlocutionary Act	Effect utterance had on the listener. *The consequence.*	The effect a visualization has on the viewer.*What is understood.*
Context	A cluster of actual states of affairs or various events related to the utterance.	The objects or entities which surround a focal event and provide resources for its appropriate interpretation.
Convention	Societal rules and norms govern countless behaviors.	Visualization design abides by these as well.
Illocutionary Force	The speaker's intention behind the utterance	The designer's design rationale behind their visualization. (Section 5.4)
Illocutionary Point (IP)	The point or purpose of a type of illocution	
Assertive Point	Convey information. The utterance that informs how things are. Either true or false.	To visually state, claim, or suggest something is the case.
Commissive Point	Make a commitment.	The guarantees of what a visualization will offer and abide by (data authenticity).
Directive Point	Attempts by the speaker to get the hearer to do something.	Engaging or motivating the viewer to do something via the visualization.
Declarative Point	Create a new state. Utterances that change the world by representing it as being changed.	Data transitions or transformation as well as predictive visualizations.
Expressive Point	Reveal the speaker's attitude or emotion towards a particular proposition.	Revealing personal bias or sharing personal opinions through visualization.
Degree of strength of IP	These points can be achieved with different degrees of strength.	Degree of the design's effort to convey an IP through the visualization.
Mode of achievement of IP	The various means a speaker utilizes to achieve the IP of an utterance.	The means a designer employs to communicate the IP of the visualization.
Propositional Content Conditions	A limitation on the nature of the state of affairs for an IP.	Each IP has conditions that need to be met for the illocution to register.
Preparatory Conditions	A state of affairs that is presupposed is a necessary condition for the non-defective employment of the force.	Assumptions the designer makes about a viewer when employing a particular force.
Sincerity Conditions	The psychological state of the speaker concerning the IP.	The designer and the viewer take the visualization and all its content as intentional.
Degree of strength of sincerity conditions	The strength of the psychological state the speaker commits to when employing an IP.	

Table 6.1: Key Information on speech act theory concepts, applications, and meaning in visualization.

minology.

The field of speech act theory examines how words are not only used to convey information but also to carry out actions. Many philosophers and linguists study speech act theory to gain insights and a better understanding of how we communicate. A speech act can be described as something that is expressed by an individual that not only offers some information but also performs an action.

The initial foundation of speech act theory was introduced by J.L Austin [193] and the theory has since been developed and expanded by several other scholars [196, 195, 198]. Austin introduced the terms *locutionary*, *illocutionary*, and *perlocutionary* acts. Where locutionary act is the utterance of the phrase, illocutionary is what was meant or intended by the utterance, and perlocutionary act is the effect the utterance has upon the listener. These terms of locutionary, illocutionary, and perlocutionary can, respectively, be thought of as: what is being put forth, how is it being put forth, and what does putting it forth achieve?

6.4.1 Forces

Classical speech act theory [193, 195, 211, 196] introduces the idea that our utterances, words with some meaning that we put forth, have a variety of forces. Grice [196] introduced a concept of speaker meaning. Whereby a speaker attempts to get the audience to believe something by relying on the audience to take the speaker's intention as a reason for belief. Grice finds that in order for a speaker's meaning to occur, the speaker must first intend to produce an effect on an audience and also intend that this very intention be recognized by that audience. Next, the speaker must also intend this effect on the audience to be produced at least in part by their recognition of the speaker's intention. Speech act theory recognizes [212, 213, 198] that an illocutionary force contains the speaker's intent. Namely, illocutionary force is the intended message a speaker assigns to a sentence they utter. Searle and Vanderveken [213] assert that the force in speech is comprised of 7 parts; illocutionary point (IP), degree of strength of the IP, mode of achievement, propositional content conditions, preparatory conditions, sincerity conditions (SC), strength of SC. The illocutionary point can be of the following forms: assertive, commissive, directive, declarative, and expressive.

Neo-Gricean theories modify Grice's principles to some extent. From these modifications, we are given relevance theories [214] as well as modifications to forces allowing for more focus on the speaker's intention. In this work, we use the Neo-Gricean analysis [215, 216] as a basis for our mapping between communication in visualization and speech act theory. Mapping these forces into visualization requires careful consideration of what is consistent and what additionally needs to be factored in. Murray and Starr [198] proposes that the force of an utterance is its communicative function. They examined Grice's definition of communicative intention [196] and found that it did not consider how signals are used to coordinate communications. Although Murray and Starr [198] state that the approach we adopt does not address how agents use signals to coordinate, in the context of visualization we fill this gap using semiotic theory. As visualization designers, we make use of visual signals which is explained by semiotic theory. An important takeaway of semiotics [217, 218, 219] is how social conventions influence meaning. In other words, the force of an utterance is contextual and subject to conventions.

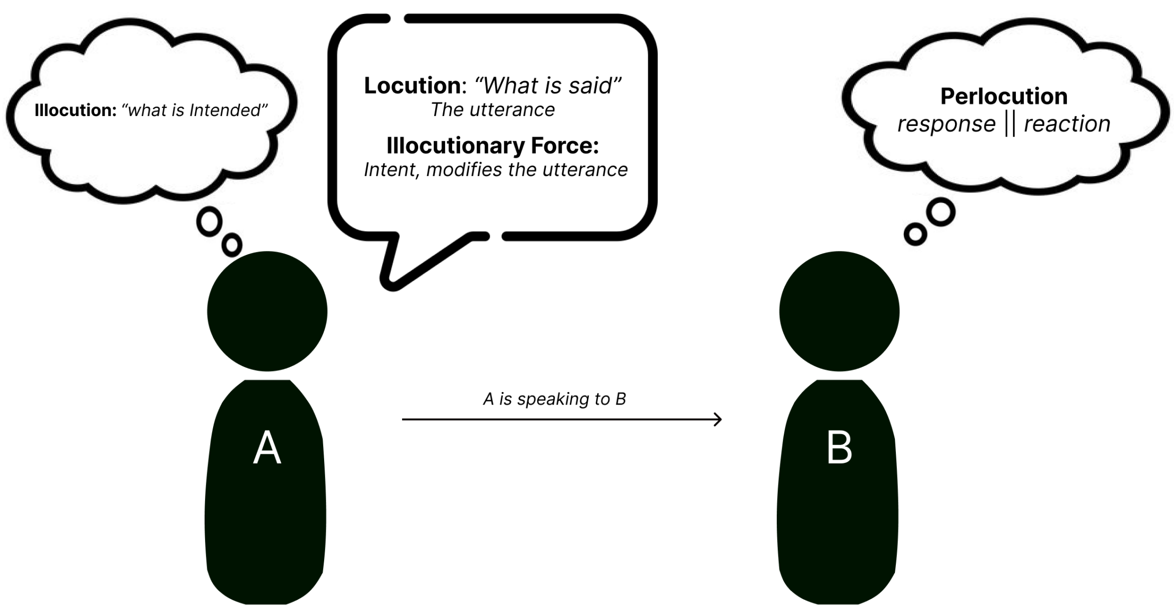


Figure 6.2: Illustration of a speech act. Person A utters a phrase, a locution, to person B. This locution is what person B will hear. Person A simultaneously also performs an illocutionary act by applying a force with their intention. How person B responds or what they do after processing what Person A has said would be classified as a perlocution.

6.4.2 Speech Act Example: Alice & Bob

To provide a clear example of what is a speech act and what can one do with it let us observe a conversation between Alice and Bob Fig.6.2.

> Alice: *"Would it be too much trouble for me to ask you to hand me the salt?"*

Alice utters a sentence to Bob that is asking two questions concurrently. The first question is if Bob is capable of passing the salt, while the second is an actual request. The *locutionary act* in this case is what was said, the literal sentence. The *illocutionary act* is what Alice means by uttering the sentence and the *illocutionary force* is the intent. Specifically, Alice's intention was a request for Bob to give her salt, she then issued an utterance with the illocutionary force of command to Bob. If Bob, successfully processes this utterance and its force and proceeds to acquire and hand Alice the salt then he has performed the *perlocutionary act*. In order for Bob to identify the illocutionary force and corresponding act Bob must either passively or actively factor in relevant contextual information or social conventions and practices to help determine what Alice's intents are.

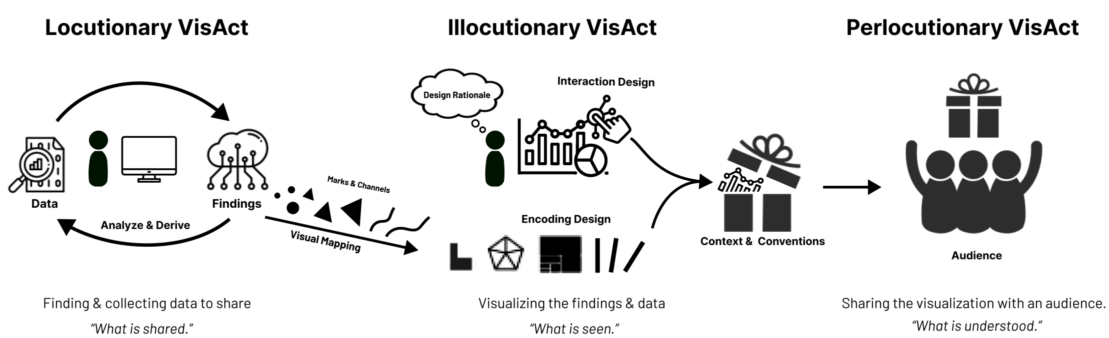


Figure 6.3: VisActs consists of three core actions; Locutionary, Illocutionary, and Perlocutionary acts. Through each of these actions, the designer imbues their intent into the visualization with the hope the audience understands and interprets the message as intended.

6.5 VisActs

The speech act was developed for understanding and examining how we communicate, specifically with words and speech, however, it can be extended past words for visual communication. Visualization is becoming a dominant mode of communication. As visualization becomes more complex for expressing and communicating data it will inherit the challenges of languages. At one level it has the ability to express more precisely but concurrently opens itself up to more ambiguity and multiple ways to be interpreted, which can have a variety of implications.

6.5.1 Framework

With VisActs, we borrow some concepts from linguistics to use as a foundation and then proceed to contextualize and translate how these structures apply in data visualization, specifically the communicative side of data visualization. In Speech Act Theory, we can use three structures to frame intents and their effects; locutionary, illocutionary, and perlocutionary speech acts. With this framework, we offer varying levels of depth for the analysis of designer intent, as seen in, Table 6.1. Furthermore, we focus on a retrospective analysis of designer intentions in visualizations and particularly data stories to illustrate VisActs.

To begin our translation we first contextualize each of these for visualization. A locutionary VisAct is the *data* or finding we present to the targeted user or audience. An illocutionary VisAct is the *visual representation* or *encoding* this finding or takes assumes when presented to the target user or audience. The illocutionary force or VisForce, is the *design rationale* for the representation or encoding. Lastly, the perlocutionary VisForce represents the *evaluation* of the encoding design after the audience has viewed and processed it. Through the perlocutionary, VisForce the designer gains an understanding of if their intended outcomes were met, that is the audience decoded the encodings and understood the findings or data presented as intended by the designer.

With this framing, we have separated stages of visualization design into several bins. In the first bin, locutionary VisAct, we focus on isolating what is the specific or derived data to convey to the audience. This bin is not concerned with *how* this data is visually represented or modified but focused on the semantic *what* part of the data is being shared. It is in the illocutionary VisAct and its accompanying VisForce, that we can begin teasing and understanding how the design impacts the communication of the data. There are several means through which a designer can transform the visualization to reflect their intentions. The two categories this work will focus on are encoding and interaction design. However, how we communicate, design, and interpret data-driven visual content is also affected by societal conventions and other contextual information.

The goal of VisActs is to provide an alternative means to assess here how are intentions shape *visually* the data we are communicating as well as better infer a designer's original intentions for producing a visualization.

6.5.2 Locutionary VisAct

For the purpose of this work, we are only concerned with data that has been identified to be shared. VisActs does not consider data whose content is largely unknown and expected to be explored. The Locutionary VisAct made by the designer is the process of selecting data, tasks, and initial analysis methods (i.e., data cleaning). As these choices reflect part of the designer's intentions. For example, in data storytelling, this is the process of identifying the "story pieces" to be communicated [37]. The data selection and modification affect the visualization design, as it may constrain what visualization options if any [191, 220, 221]. For example, hierarchy suggests depth, temporality may imply change, and spatiality could imply closeness or bonds. Thus, we may be visualizing data as a treemap, flows, or possibly on a map. By taking data types into account, such as nominal, categorical, numerical, and their pairings we can begin to define a space of what representations are available to fulfill our communicative goals.

6.5.3 Illocutionary VisAct

The illocutionary VisAct is the process of designing a visualization from the data to then be shared with an audience. This visualization may be interactive but must be data-driven. Similar to speech acts, we are not concerned with visualizations that have no data or "'meaning" associated with them. The design of both interactive and non-interactive data visualization is heavily influenced by the designer and their choices. Based on how the designer intends to communicate this data (i.e., to educate or persuade) may influence their design rationale. The design influence appears in (1) the encoding design and (2) the interaction design.

Encoding Design. How we design visualizations, in terms of binding the data to the visual elements, greatly impacts how the data are perceived and understood by the audience [222, 220, 221, 223]. Certain visualization design choices may elicit emotional responses from the audience, which can also help better communicate the designer's intentions to the audience [63].

Interaction Design. Interaction design as it pertains to data visualization is a heavily docu-

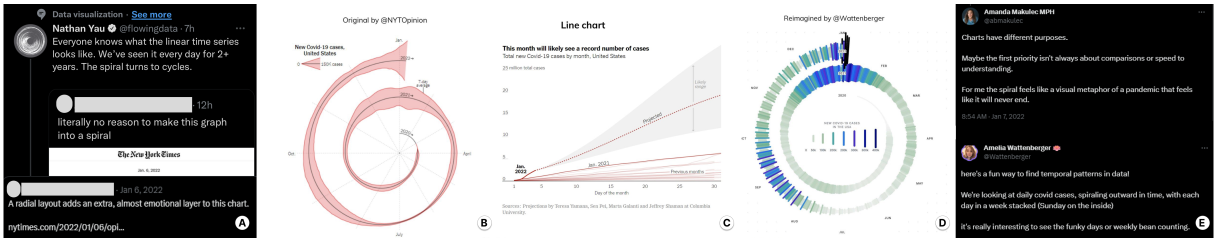


Figure 6.4: Jeffrey Shaman wrote an article [225] predicting when the Omicron variant of Covid would peak with an accompanying visualization. The visualization sparked an online Twitter debate (A), with some finding the design ineffective and better than a line chart (B). Amelia Wattenberger re-imagined the design with a different intent (C), and Amanda Makulec (Executive Directory of DataVizSociety) wrote a thread[226] on how visualization design can address different needs or intents (D).

mented [202, 224, 189, 205, 206]. From these works we can surmise the following: (1) interaction design effects visually what is seen by the audience, (2) interaction design influences how the audience perceives the data, (3) and interaction design impacts audience engagement with the data. The designer's choice of which interactions are available to the audience can steer the audience toward the designer's goal.

6.5.4 Perlocutionary VisAct

A perlocutionary VisAct is performed by the audience. This action informs the designer whether or not their desired outcome has transpired. If the outcome was to *call to act* over climate change by signing a linked petition and they received a signature that would be a success. However, if the outcome was to *educate* museum visitors on metagenomics [53] via an interactive system and the majority of visitors failed to understand the system, then it was unsuccessful. The granularity of success and failure, much like with evaluation is up to the designer to classify. In the context of data visualization and research, this stage is evaluating what was understood but the audience and how that aligns with what the designer intended to transpire [227, 220, 221].

6.5.5 Convention

Social conventions are rules and norms that govern countless behaviors we all engage with every day. Bicchieri [228] defines conventions as descriptive norms that have endured the test of time. She states that if one's main objective is to coordinate with others, and the right mutual expectations are present, people will follow whatever convention is in place. In visualization design adhering to and following conventions is necessary for effective communication. For example, color norms differ based on culture, knowledge, and context. Rainbow-colored maps are typically associated with temperature variations [229]. If a designer applies a rainbow color map in a geospatial visualization to depict crop yields then many viewers may not properly decode the visualization.

6.5.6 Context

In communicative visualization, several works [230, 231, 191, 227] identified challenges in the interpretation of data-driven visualizations and how different contexts affect this interpretation. Mukherjee et al. [230] proposed the use of semantic discriminability theory, a general framework for understanding conditions determining when people can infer meaning from perceptual features. There is a large body of linguistics research [193, 195, 215, 214, 232] showing how context influences the meaning of an utterance. Sbisà [232] proposes contexts are continuously shifting, but at each moment of interaction it is possible to evaluate the performed act against the context. This literature suggests context can be classified along the following dimensions: (1) Given vs constructed context, (2) limited vs unlimited context, and (3) context change.

Given vs. constructed context: In a given context, the context is set once the event starts and is not mutable going forward. For example, many narrative visualizations [53, 1, 37] or analytical systems predetermine or have a fixed context. Whereas in a constructed context the context of an interactional event is created by its participants as the interaction proceeds. One form of this in visualization could be highly interactive and collaborative visualizations that function off of user inputs. These visualization evolve and change based on these interactions. A different example of this can be seen in Figure 6.4 where the context of a public forum influences the designer's intent. This begins with Jefferey Shaman creating a visualization and it is shared on a public forum. The public became invested in whether the design is effective or not, how can it be improved, and what is the intent of this visualization. In response to the visualization, others were created with a different intent. As shown in Figure 6.4d, Amelia Wattenberger attempted to improve on the original, Figure 6.4b, with some believing she did. The constructed context in this scenario is that initially, the context of the visualization was to forecast the omicron virus for a period of time, however as more individuals debated the effectiveness of the visualization the new visualizations produced gained a constructed context of attempting to provide an improved design and convey the original message.

Limited vs. unlimited context: When is acquiring information to interpret what is occurring no longer necessary? Is the context finite or something that needs to be continuously provided? Context, in speech act theory, has been considered a bounded resource that only includes 'what is needed' for [233] interpretation or evaluation. Conversely, there is an argument that context is ever-changing and that there is no end to the details one might want or need. Searle [211] views context as indefinitely extensible and potentially all-inclusive. That is every speech act has meaning only against a set of background assumptions. These assumptions are indefinite and the number of processes for the analysis and development of an idea are endless. Other views [197] find context as always extensible but delimited. They believe that context is needed as background information (or what the speaker believes is, or intends to treat as background information) and is delimited on every occasion by the presuppositions a speaker happens to make. Additionally, actions typically involve results, such as bringing about a new state, referencing the past, or substituting a new state for an older one. The objective or cognitive nature of context affects the action. After an action or event occurs its content is added to the participant's presuppositions, and therefore to the cognitive context. For example, in dashboards with linked views a user may filter on a subset of the data altering the chart. An accompanying view may re-render to reflect this filter and reveal new insights on the subset reflecting that initial interaction. This change is an implicit communication to the viewer that these two views are linked and the data in one is influencing the other. The discussion of limited vs unlimited context is ongoing in speech act theory. However, the distinctions and points made, as well as points made for future works, directly apply to visualization. For example, Mantri et al. [227] examine how a variety of additional contexts impact the interpretation of communicative visualizations and synthesis of information that comes from consuming cumulative discoveries. They discuss how science and journalism often present their content as an ongoing discussion and evolution rather than a finality (i.e., discoveries that build on, contradict, contextualize, or correct prior findings).

These considerations of context clearly arise in visualization and have been defined implicitly by this classification. Several frameworks [37, 234, 53, 222] have discussed context and its

VisActs Terminology	Description
Locutionary VisAct	To show data. *What is shared.*
Data Act	The curation & selection of a dataset(s)
Analytic Act	Expression of the data through analysis.
Data Type	The type of data (e.g., temporal, spatial, quantitative, discrete, etc.).
Illocutionary VisAct	To visualize the data. *What is seen.*
Image Act	The production of an image.
Semiotic Act	The expression of data through signs & codes.
Encoding Act	Visual encodings mapped to data.
Visualization Type	The type of visualization (i.e., informative, instructive, narrative, explorative, and subjective).
VisForce	The designer's rationale.
Perlocutionary VisAct	The effect the visualization has on the audience. *What is understood.*

Table 6.2: VisAct terminology. These terms and their mappings were derived from a breadth of linguistics and data visualization research.

influences on visualization design. For example, they describe external context as (1) an understanding of the target audience's needs and prior knowledge, (2) the type of device or medium the visualization will be expressed through, (3) and the physical setting. Context's effect on inferring and interpreting visualizations has also been examined [222, 191, 227, 231, 230]. Padilla et al. [222] identify how after viewing the encoding, the audience mentally searches long-term memory for knowledge and context relevant to interpreting the visualization. Furthermore, context, as it pertains to visualization, has many influences on the design and the designer's intent. Consequently, subtle changes in intent can be reflected and seen in the visualization, Figure 6.4. With VisActs, we provide a framework to facilitate studying at a granular level how intents influence the design.

6.6 VisActs: Application to Visualization

To ground the value of viewing visualization as a language and applying the speech act theory we provide a set of examples. The first example uses VisActs asses an NYT visualization. These examples use VisAct at a granular level to study the intention from the perspective of two archetypes commonly observed in communicative visualization: storyteller and educator. As a disclaimer, we can not know for certain the original designer's intent. However, we can tease out to an extent what it plausibly could be based on design decisions and available documentation.

6.6.1 Example: Storyteller

The storyteller is concerned with expressing data as a narrative. Here visualization is a means to engage an audience with the data and the visualizations are carefully sequenced to illustrate causality. We apply *VisActs* to study a recent narrative visualization piece, *How the Virus Got Out* [7].

This narrative piece is composed of several visualizations, five of which are shown in Fig. 6.5. The story starts with Fig. 6.5a, a visualization that promises that the authors will explain visually *How the Virus Got Out* of China. It also asserts that the pandemic started in China.

Let us first focus on what is being shared, not inferred from the visualization. This is defined as the locution or locutionary act. When contextualized as a *VisAct*, the locutionary act and the locution describe what the underlying content is, namely the data. Here, the **locutionary VisAct** is the data and analysis used to understand the spread of the virus. The *data act* is the selection and curation of the datasets that represent people, their movements, and infection data. As mentioned in the article [7], their data came from Baidu, two Chinese telecoms, Fred Hutchinson Cancer Research Center, the University of Washington, and the Johns Hopkins Center The *analytic act* are the estimations and relevant methods applied to the dataset to help bring out the findings to then be visualized. Lastly, the *data types* are spatio-temporal data.

The visualizations are the illocutionary **VisAct**. The **image act**, consists of the low-level visual elements. The viewer would see a web page composed of color, shapes, and text (e.g. the marks and channels). The **encoding acts** determine how marks and channels should be paired and how they are bound to the underlying data. In this example, the data contains temporal information about people's movements, location, and estimates of the percentage of the population with COVID. Individuals are represented as points, the position of a point connotes a geo-spatial location at an instance in time, and the color denotes whether an individual has COVID. The meaning provided by the encoding act is supplemented by the **semiotic act**. The semiotic act constructs the relationships between the image and other factors such as culture and society. The grouping of shapes and text together is seen as a map of China and neighboring countries. This visualization uses projections and map aesthetics that most of the populace has familiarity with from map-based applications. Therefore, the movement of points across this map is associated with transit. In Western cultures red color has negative connotations. In this case, red is used to symbolize those infected with the virus. Although the piece presents facts, because of its emphasis on temporal flow and implied causality, the **type of visualization** is a narrative. It is important to note that the type of visualization influences the meaning it will convey.

The **VisForce** is the intention underlying Fig. 6.5a. One intention here is a promise to the readers *to educate* how the virus spread. A *VisForce* is comprised of the seven properties described in Section 5.4. To understand the *VisForces* at play let us first identify the set of **illocutionary points** at work. Fig. 6.5a has a commissive point and an assertive point. It asserts that the virus started in China and it promises to provide a justification for this assertion. The **degree of strength** of the promise is moderate, as we have to infer the promise. The mode of achievement of the *VisForces* is through the sequence of steps used to build the visualization. It starts with the text "China", which slides and snaps onto a map of China. Then they add a set of red dots that quickly transitions to streams of red dots flowing out of China. The felicity conditions for these points are the assumptions by the designers that they have the information and that the reader will benefit from the visualization.

Throughout the story, the visualizations make several assertive points about the spread of COVID, where it originated, and facts about specific days. In Fig. 6.5c, the designers present a set of points depicting the number of reported cases in December. The *VisForces* consists of two assertive points and an expressive point. The illocutionary *VisAct* of a small cluster of red points asserts that only a few dozen cases were known to the doctors. The second assertion was that the true number of infected was closer to a thousand. The corresponding illocutionary *VisAct* is a larger cluster. The expressive point was the emphasis the authors placed on the difference between the two assertions. The illocutionary *VisAct* to achieve the expressive point is the animation that grows the volume of the dots. Its degree of strength is high.

In Fig. 6.5d, volumes of varying sizes of red points are shown on a map. The designer assumes that the size of the cluster will be interpreted as the size of the infected population by the viewer. We can argue that Fig. 6.5c introduces the context necessary for this interpretation. As we have stated earlier the *VisForce* depends on the context, which can change or evolve. The visualization in Fig 6.5e opens with a declarative point. There is a state change in the *VisAct*. The overall semantic state (image act, encoding act, and semiotic act) of the visualization has changed. The visual narrative transitions from a geo-spatial visualization to a scale-free representation that mimics a subway map. This enables the designers to make an assertive point of Wuhan's centrality in the pandemic.

In addition to internal contexts, there are external contexts such as social conventions or norms that can passively or directly influence the *VisForce*. For a viewer to recognize the *VisForces* in this story they must be (1) aware a pandemic occurred resulting in a global lockdown, (2) familiar with reading and interpreting maps, and (3) able to understand English. Also, as we have seen, *context change* contributes to the intended meaning. Information necessary for interpreting visual encodings is presented and then applied in different settings. Lastly, there are some **conventions** this visualization takes advantage of. Namely, it uses the popular scrollytelling design and assumes those viewing the page understand the convention of scrolling to reveal new content. The **sincerity** of the designers is seen on the final page of the story where they provide notes talking about limitations of what is presented as well as sources for the data and statements made.

To recap, we have organized the mapping of this narrative into two sections. The first section, **locutionary acts** addresses what the designers put forth and what it is we see. In the second part, we focus on **illocutionary acts**. We identify (infer) the intents and examine how they are expressing their intentions. This is addressed by delving into the *VisForces*, the illocutionary points, modes of achievement, and the context. In this example, the designers use several assertive illocutionary *VisForces* to convey to the viewer "How the Virus Got Out". We also identified and discussed declarative, expressive, and commissive points.

Finally, let us look at the **perlocutionary act**. This third and final component of *VisActs* addresses the consequences of presenting the visualization to the viewer. The perlocutionary act captures the effect presenting the visualization had and assesses if the viewer understood the

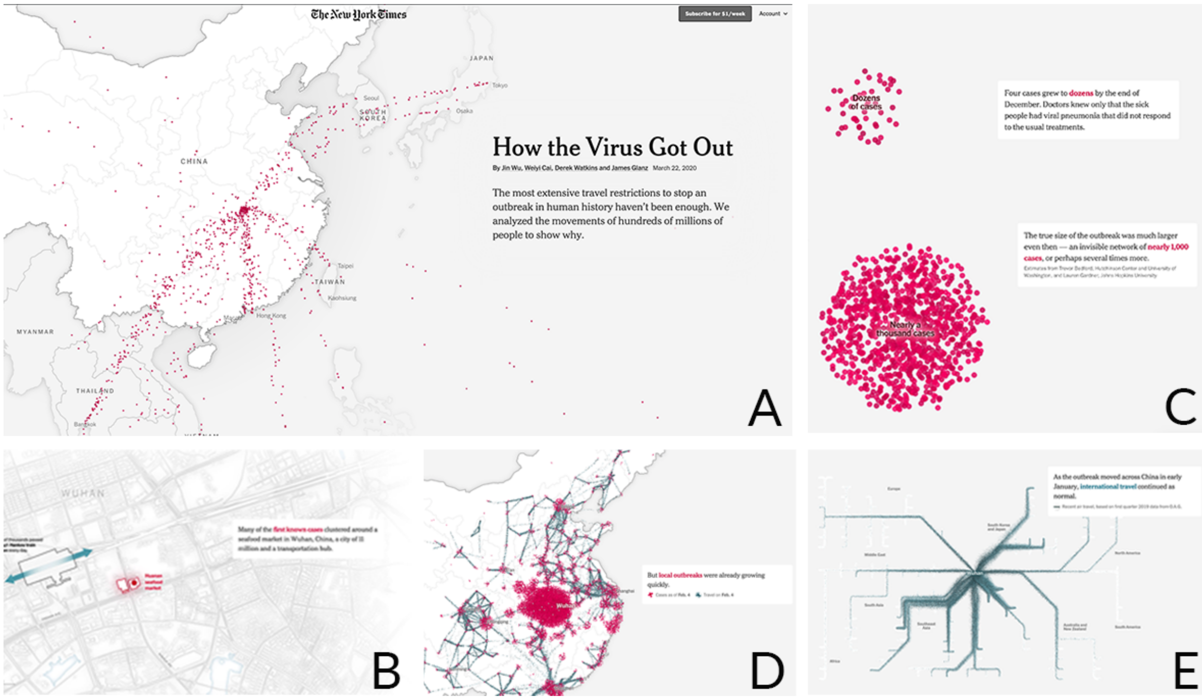


Figure 6.5: *"How the Virus Got Out"* [7]. (a) Title page, (b) map of Wuhan, China, (c) visualization of what was assumed to be the number of COVID cases in December compared to what it actually was, (d) overview of the virus spread, (e) scale-free representation of the world.

designer's intent. In our field, we conduct user studies and evaluations to determine this. To ascertain if the visualization was successful in communicating the intent of asserting the factors that led to the virus spreading across the world we would need to interview viewers.

6.6.2 Example: Educator

A common use of communicative visualization is to teach or to inform. An educator uses visualization to simplify complex data to explain concepts. This can take place in a formal setting such as a classroom or in an informal setting like a museum. We examine the museum exhibit DeepTree [235, 236] using our framework.

Here, the **locutionary VisAct** is the phylogenetic tree of life. The *data act* is the phylogenetic dataset and the corresponding timelines. The *analytic act* could be any data formatting or association of the phylogenetic tree with the temporal information. Lastly, the *data types* are temporal and image data.

DeepTree's **illocutionary act**, as seen in Fig. 6.6a, has a tree-like structure supporting a set of images. The **image act** of Fig. 6.6a consists of point and line marks. DeepTree's **encoding**

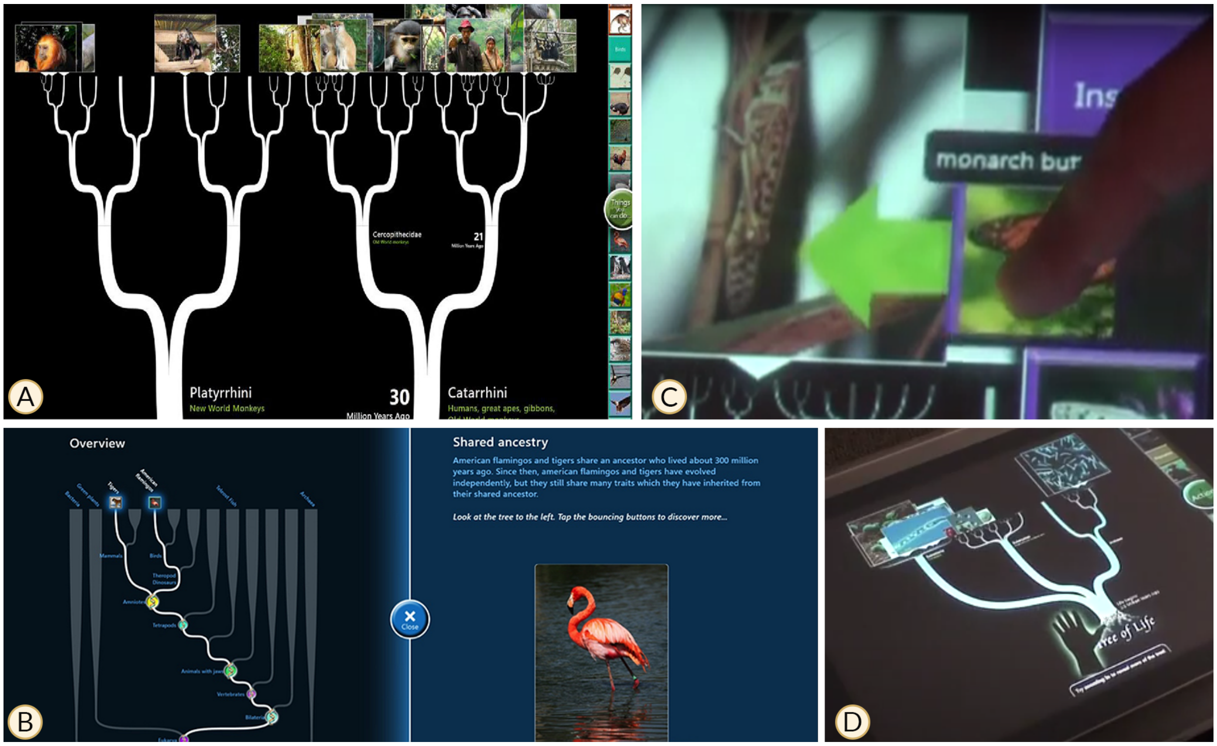


Figure 6.6: (a) DeepTree [235] exhibit overview (b) simplified view (c) selection interaction (d) zoom and pan interaction. (permission pending)

act maps the visual elements to the underlying data. As a result, what is seen in Fig. 6.6a is a phylogenetic tree dataset that consists of images and text. The designers create a tree visualization where leaves are images of species. The visualization is composed of a line mark and takes advantage of size and position channels to convey the dataset. Images on the tree depict species along with a label with their common name. The **semiotic act** provides a tree metaphor, where the trunk symbolizes the universal common ancestor and the branches the children. The ”branches” of this tree represent splits in species and portray the phylogenetic classification. This **type of visualization** is primarily explorative and is supplemented with some informative elements.

As with the prior example, to understand the illocution, we identify the *VisForces*. This visualization makes use of four of the five illocutionary points as described in Table 6.1. It has assertive, directive, commissive, and declarative points. We will focus on the directive and commissive points.

Figure6.6a is the entry point for this visualization. The *VisForces* here include commisive and directive points. The *VisAct* promises to inform the visitor about the Tree of Life. The mode of achievement is a collection of images of different species and animations indicating relationships among them. The directive point is to get the viewers to drag their hands across the screen. The mode of achievement is an animation of a hand dragging downward to instruct viewers how to engage with the application, Fig. 6.6d. This visualization has several other directive points that are achieved through different modes such as tapping a button, downward dragging to traverse down the tree, pushing upward to move up the tree, flicking to quickly move through the tree, single-touch pan, multi-touch zoom, pushing to select, and dragging an element onto a destination. Each directive point is achieved by using techniques such as highlighting, animating, or annotating visual elements to cue the viewer to interact. The degree of strength for a directive point depends on the visual emphasis placed on that technique.

External factors and conventions also influence the directive points. The museum setting and use of an interactive touch-screen table to display the visualization add to the *VisForce*.

The side panel in Fig. 6.6a has a commissive point. The promise here is to inform the viewer of the location of the tree of each species portrayed in the side panel. In DeepTree when a user selects an image of a species from the side-panel, Fig. 6.6a, the image jumps to its position in the tree. If an image is pressed a graphical element, an arrow, appears showing the viewer where to slide (Fig. 6.6c). The directive point here is to get the viewer to slide the image. This directive point is weaker than the earlier directive point for getting a viewer to drag their hand onto the table.

Let us next examine the **propositional content condition** for directive points. These are the designers' beliefs that the viewer will perform an action they request. In DeepTree, the designers believe that by animating an element to grow and shrink, adding a highlight around it, and having a text annotation above it saying "learn more" the viewer will tap on it. The **preparatory conditions** for all directive points assume that the viewer is able to perform the suggested actions. The **sincerity condition** of these directive points is that the designer wants the viewer to perform the actions. The degree of strength for the sincerity condition is the importance of these actions to the designer. In DeepTree it is very important for the designers that the viewers pan and navigate the tree. This action is crucial and is evidenced by the visual emphasis placed on this *VisAct*.

The viewer can also press a button that takes them to a new simplified view of the tree. The illocutionary *VisAct* of this visualization is a tree of life contextualized to the selected species, Fig. 6.6b. The *VisForces* here have a declarative and commissive point. The declarative point is the transition from the earlier state, Fig. 6.6a, to a simplified visualization, Fig. 6.6b. This declarative point's mode of achievement is an animated transition.

The commissive point the simplified view makes is that it promises the viewer that the designer will return them back to the original state, Fig. 6.6a. The **mode of achievement** for this commissive point is a button. The **propositional content condition** for this commissive point is that the designer will fulfill the commitment. That is, upon a viewer pressing the "X" button, seen in Fig. 6.6b, the simplified view will disappear and "close" and they will be returned back to the overview, Fig. 6.6a. The **preparatory conditions** for the commissive point is that the designer is able to complete this promise within their design. The **sincerity condition** of this commissive point is that designer, and therefore the visualization intends to satisfy the promise.

Briefly, there are many assertive made in both the simplified and overview visualizations. These assertive points state facts about species and their ancestry. The modes of achievement the designers selected to express their assertive points include dialog/pop-up boxes, annotations, and color to highlight relationships between species.

So far, we have gone over some of the *VisForces* present in this exhibit to illustrate how to use our framework and the structure it provides. Namely, we showed that there are assertive, declarative, commissive, and directive points and thus those respective *VisForces*. We walked through the properties of some of these forces and gave examples (i.e. we described eight directive *VisForces*, degree of strength, conditions, and mode of achievement). However, we also have to account for how social conventions and external contexts influence the visualization design and its overall meaning.

DeepTree relies on external contexts and conventions present in an informal learning environment; specifically a museum and the considerations and conventions [53] that it comes with. Furthermore, DeepTree relies on its viewers to have familiarity with touch-based devices [235] (e.g., iPads and iPhones).

Lastly, the perlocutionary *VisAct* addresses the reaction the viewers had upon seeing the visualization. It can be used to determine if the designer was successful in conveying their intended meaning. The designers of DeepTree documented their evaluation [235] and from it we can see that their directive and a comissive *VisForces* were understood by the viewer. For example, both the commissive force of a promise to the viewer that by tapping the find button something will occur in the future and the directive force of a request to the user to tap and drag an image off the image reel were successful. Viewers would tap on the button to find a species, signifying the viewer believes a promise will be fulfilled. Additionally, they were then presented with a slot to place an image. They inferred the directive point and dragged an image onto the slot. After doing so, the promise made by the designer is fulfilled as the visualization "flies" through the tree via animation to where the species in the image is located. This *VisAct* had "emotional" perlocutionary response in the viewers, where the designers documented responses such as "wow, this is big" or "woah".

6.7 Discussion

With VisActs, we present a conceptual framework for inferring designer intent. This framework is not a finality but rather a foundation to be iterated and expanded upon. There is a growing need to infer and assess designer intent, as well as grow our understanding of which types of design decisions illustrate an intent (i.e., negative emotions can be conveyed via a variety of design choices [63]). With this chapter, we provide a translation and contextualization of frameworks from linguistics to communicative visualization and offer a framework for inferring intent in interactive data-driven visualizations. In this section, we discuss the immediate and prospective values and directions that VisActs can lead.

Biases Assessment: One use of VisActs is to tease out design decisions that could reflect the designer's biases. For example, in Figure 6.7a we see a bar chart communicating *"America's economic growth in the 21st century"*. This chart at a glance shows the tallest bar to be for the year 2021. Upon closer review, however, Figure 6.7b, we see that the y–axis has a mistake. This mistake extends the 2021 bar to be slightly more exaggerated in comparison to the other years. This mistake was shared for several hours on Twitter before a correction was made. In that span, the community responded with many complaints and comments over the chart. There

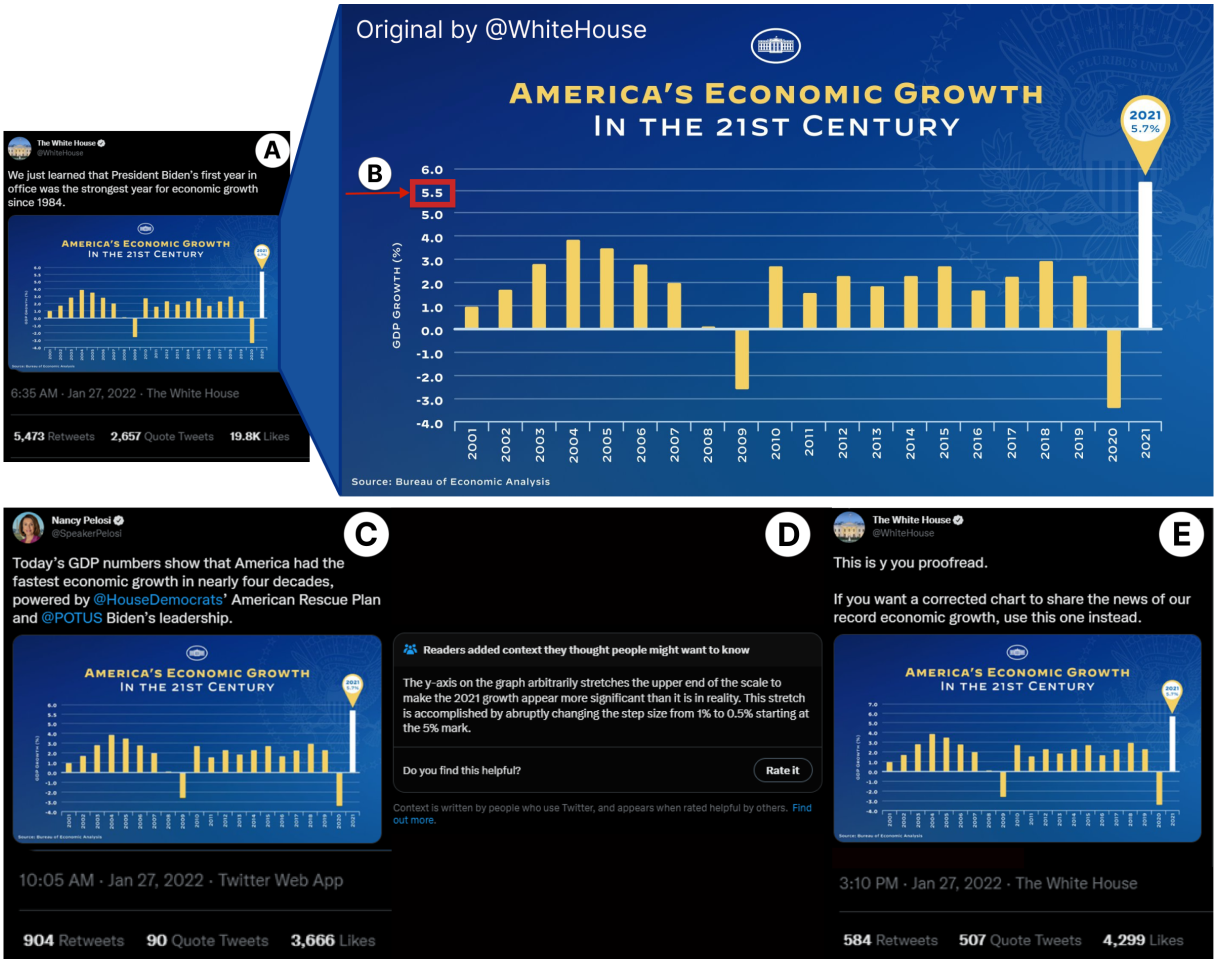


Figure 6.7: (a) Tweet by the @WhiteHouse account conveying the economic growth of the current administration [237]. (b) The y–axis label increments by a half-step rather than a whole point as it was previously. (c) Nancy Pelosi shares this figure with the error to her following [238] (8.1 million followers). (d) The Twitter community flags the discrepancy and many replies and threads are made questioning the integrity of the data. (e) @WhiteHouse updates the figure stating it was a *proofreading* issue.

were claims that it was intentionally added as a means to persuade that the 2021 administration is very effective. The public would attempt to evidence these claims by suggesting the mistake only occurred at a point on the y-axis that would only affect the 2021 bar and nothing else. Namely, the public assessed the encoding design to infer a plausible designer intent.

It is impossible to say with absolute certainty whether this case was a genuine mistake or an intentional design choice. However, there is a need to provide structures to assess and better debate the biases in charts that circulate in the wild. These visualizations are seen by many and affect how society trusts data and science. One possible extension of VisActs is to apply the framework to a corpus of data visualizations and create associations between design and possible intents. Through developing richer classification models, we could develop faster or more accurate linters that could identify these discrepancies and provide community notes, Figure 6.7d.

ML4Vis application. Bako et al. [190] call attention to how visualization examples can be used as inputs to author visualizations as a direction for future work. To help achieve this goal, we need to really understand and expand on which aspects of these examples correlate to what the designer intends to do. This could be task-based as well; however, it is possible that task-based may not be granular enough to effectively capture the needs of designers and the specific design elements they may be interested in. Expanding on classification models VisActs has an application in the ML4Vis space. Principles and frameworks from speech act and discourse theory have been studied and leveraged in NLP. Similarly, VisActs can be utilized in future works for automatically generating visualizations and their design based on designer inputs. By offering a framework to infer the designer's intentions based on design rationale and assess which features could contribute to particular intentions, we can better train machine learning models on data visualizations to build associations on which visual features correspond to particular intentions.

As with Midjourney and DallE prompts, VisActs can be a stepping stone to developing applications that allow users to provide their data to visualize and enter prompts to tailor the visualization to their needs. In order to achieve such automation we need to develop rich associations between designer intents and corresponding design choices. Through VisActs, it is possible to develop these associations. This framework can be applied at a granular level, as seen in the examples for Storyteller and Educator.

6.8 Conclusion

This work takes the view that visualization is a language and can therefore benefit from applying frameworks and theories from linguistics to systematically understand and analyze how we communicate through data visualization. We provide a translation of a sub-field of linguistics and offer our framework VisActs. We then use examples applying our framework to illustrate its potential application to our field. This translation affords us a means to deconstruct the language of visualization, identify low-level communicative components, and learn how these components individually and collectively accomplish the communicative goal of the visualization. Our detailed examples demonstrate how these concepts can be used to examine designer intents and describe the forces at play.

This is an initial mapping of the two spaces and future work can tighten this association and build upon its structure. We believe that our work gives credence to the relevance of linguistics frameworks for the study of visualization and supports continued efforts in translating other frameworks and theories into our domain. We hope our work enables the future integration of theories and frameworks from linguistics into visualization and grows our framework for studying visualization design intent.

Chapter 7

NOVA: A Visual Interface for Assessing Polarizing Media Coverage

Utilizing the frameworks I developed in Chapters 5 & 6, I wanted to apply them in an interactive visualization system for the public. The focus of this system would be on capturing the audience's belief about a topic and enabling them the ability to assess and explore their own belief through data visualizations.

Within the United States, the majority of the populace receives their news online. U.S mainstream media outlets either generate or influence the news consumed by U.S citizens. Many of these citizens have their own personal beliefs about these outlets and question the fairness of their reporting. We offer an interactive visualization system to assess mainstream media's coverage of a topic and compare an individual's preconception to the data. We gathered $\sim$25k articles from the span of 2020-2022 from six mainstream media outlets. Our system design emphasizes transparency and facilitates users' ability to evaluate their beliefs on media outlets. To evaluate our system, we present usage scenarios alongside a user study with a qualitative analysis of user exploration strategies for personal belief assessment. We report our observations from this study and discuss future work and challenges of developing tools for the public to assess media outlet coverage. From our study, we find our system is effective in retrospective assessment, allowing users to verify if their biases against media outlets are justified or should be reconsidered.

7.1 Introduction

In the current age, written news media is a combination of content that is either automatically generated by news algorithms [239] or long thematic pieces curated by journalists to provide deeper analysis and thorough explanations to the public. The public typically receives this content from mainstream media websites or via social media [240] and often gravitates to outlets that align with their personal biases. Simultaneously, media bias and its effects are widely documented [241, 240, 242, 243], while regular news consumers are typically not fully aware of the severity and scope of bias [243]. There are many methods and different stages of the journalistic process to inject bias into a news article. The manner in which an article is written and framed [244] can strongly impact people's opinions and perspectives on issues. Although many online articles adhere to journalistic integrity and journalists greatly limit introducing their own biases, these biases can still permeate the writing. Thus, with this work, we focus on the public's assessment of online written content generated by mainstream media outlets. Namely, we offer a system to both support the layperson's assessment of mainstream media reporting and verify if the sentiment of an outlet's coverage aligns with their own preconceptions. However, there remains a challenge of how to analyze the data to create and visualize an individual's lens for how they perceive the biases for each media outlet and contrast it without the lens.

Recent developments in Natural Language Processing techniques in computer science have sparked an increased effort in rethinking how we analyze media bias. Researchers from social science and psychology have been calling for an interdisciplinary approach with computer science to better assess news media content [244]. Previous works have applied machine learning models on annotated datasets to identify certain types of media bias [245, 246, 247, 242]. However, that approach is susceptible to performance issues [248, 249]. The difficulty of building a high-quality bias dataset is a factor behind those issues. This difficulty is due to readers' being unable to agree on what's biased due to factors like background knowledge, political ideology, or even the Hostile Media Effect (HME) in psychology [243].

An alternative method to study bias effects is to use sentiment analysis. Several works have shown the viability of combining sentiment and named entity to identify framing characteristics [250, 251, 252, 253]. We extract and aggregate named entities (i.e., person, location, organization) mentioned by the news media and visualize the aggregated sentiment score. At the same time, previous works have reported that "the interpretations of the text can be multiple and they depend on the personal background knowledge, culture, social class, religion, etc. as far as what is normal (expected) and what is not are concerned" [250]. Other works [244, 254, 243] suggest that people actively or passively isolate themselves in a "filter bubble" or an "echo chamber". A key insight from these works is that individuals who deepen their understanding of how their own perspective was formed, via exploration, are able to reveal their own echo chambers.

As mentioned previously, we are interested in exploring the relationship between one's personal beliefs about news media outlets and how those outlets actually behave. We take into account the key findings in other fields and introduce an alternative approach to explore and assess if one's perception of a media outlet aligns with how that outlet actually reported on a topic. To achieve this, we introduce a system for News Outlet Visual Assessment (NOVA). *NOVA is a visual interface designed to facilitate the generation of hypotheses on mainstream media's coverage of an entity, assess the rationale for the sentiment of articles covering the given entity, and contrast media outlets' coverage of an entity to one's prior expectations.* Our target audience is a subset of the general public, we cater to individuals who either encounter or actively seek out news media articles and content (e.g., NYT, CNN, etc) fairly regularly (i.e., at least once a week). We walk through an example workflow to demonstrate how NOVA can be used to assess media outlet coverage as it relates to one's personal beliefs. Additionally, we present a usage scenario alongside a user study with a qualitative analysis of user exploration strategies for personal belief assessment.

7.2 Related Works

Our work lies at an intersection of text document analysis, as it pertains to news-related content, and the assessment of an individual's belief or preferences and illustrating those effects. One aspect of our system is to provide a platform for individuals to assess if their personal beliefs or predispositions to various outlets are aligned with the data. In this section, we review visual analytic systems with a focus on news media. The systems also apply natural language processing (NLP) as well as other text visualization-based techniques. Additionally, we discuss personal

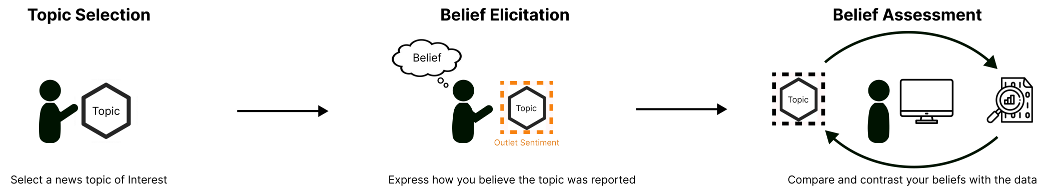


Figure 7.1: NOVA is a three-stage process of (1) selecting a topic of interest, (2) eliciting a user's belief by allowing them to construct a visual representation for a news outlet sentiment when reporting on the selected topic of interest, and (3) an assessment of their belief contrasted with the data by allowing the user to explore the articles that may conflict with their beliefs.

belief visualization systems.

7.2.1 Media Bias

In other domains, such as sociology, it is common to study the detailed effects of media bias. These methods and procedures are achieved either via manual efforts or by making use of automation methods and computer science techniques [244]. Research in this domain documents how media bias can be attributed to several factors and how it occurs at different stages during the media creation pipeline. With NOVA, we narrow our scope of media content to written media, specifically web-based articles.

The framing effect on an individual reader is a result of long-term isolation in informational reality. Without prior knowledge about the political stances of a media outlet, a particular article can be perceived as neutral or objective with only a slight slant. These slants have been shown to bias a reader's perception of an article and influence how they interpret or recall the article's content. Ecker et al. conducted a study [255] comparing the effects of news headlines with different slants on people's ability to recall and interpret the corresponding articles. They found that misleading headline framing impaired the reader's memory of factual articles and their inferential reasoning for opinion-based articles. How media frames their content can help create an isolated informational reality (e.g., an echo chamber) limiting people's exposure to different opinions that contradict their own. From the literature [244] we learn that word choice and labeling are key in revealing possible slanted news coverage.

Additionally, when reviewing large volumes of text-based data we do not want to overwhelm our users with information initially. The use of text summarization and sentiment analysis has been previously demonstrated as effective [256, 252, 257, 251] for abstracting large volumes of news data into a space where a user can get a sense of general themes. Abbar et al. [251] demonstrate how entities (i.e., Person, Organization) and sentiment can be utilized for recommending to a user what to read next. These works, however, apply techniques to assess domain expert research questions about a specified topic with limited interactivity or visualization efforts. The design of NOVA is intended to be accessible to the general public, and enable them to utilize the methods described above to assess how media outlets cover a given topic. Namely, NOVA should provide general audiences access to these techniques such that they are able to form their own questions and either prove or disprove them.

7.2.2 News Visual Analytic Systems

When reviewing visual analytic systems, with a focus on evaluating news media news, we examine those who support domain experts in their meta-analysis as well as those who offer the public a means to see more than what is written.

The COVID-19 pandemic itself produced many research works assessing how information was disseminated and various phenomena that happened in that period of history. One such work by Zhang et al. [258] examined the visualizations produced during that time and assessed how visualizations were used in a crisis situation to disseminate information. Kong et al. [259] studied how visualization titles can introduce subtle slants, which could bias a viewer's perception of the content. The purpose of NOVA is to assess the written media content produced during COVID-19 from mainstream U.S. media outlets, rather than the visualization content these outlets may also have put out. However, we are also interested in studying if an individual's perception of biases of these new outlets may have aligned with the content the outlets produced. One method of visualizing these biases is to use sentiment analysis.

There are several visual analytic systems examining news articles that primarily use sentiment [256]. However, a common problem [257, 260, 261, 262] with using sentiment analysis is when the analysis is used to denote a summary of the overall sentiment of a collection of articles, polarizing collections and neutral collections are indistinguishable. For example, it is a very common case to group a collection of related articles as either a topic or event and assigns an overall sentiment to this single entity. However, for polarizing topics or events, the impact of positive articles and negative articles that are contained typically negate each other. Thus, a topic that is polarizing contains articles that are overtly positive and overtly negative on the same subject, and a topic that is neutral is indistinguishable as they may resolve to the same sentiment score. Ilyas et al. [257] use the mean sentiment score of daily tweets to summarize a daily sentiment. Based on the sentiment score fluctuation some extremely negative or positive topics are identifiable, however, topics that have both a high volume of positive tweets and negative tweets are hidden as a result of the mean operation. Hamborg et al. [260] use a weighted sum rather than a mean to summarize the sentiment score of an individual, which results in polarizing topics being indistinguishable from neutral scores. Other works [261, 262] use the ratio of positive and negative scores, yet they also succumb to the same problem.

With NOVA, we address this problem by introducing a two-dimensional sentiment score for topics. Recognizing that an overly complex sentiment score could confuse the general audiences, our design focuses on presenting this content in an interpretable manner that is easy for people to comprehend.

7.2.3 Personal Belief Visualization

There can be multiple differences between individuals' interpretations of media, be it of news, infographics, etc., that stem from their personal background [250]. Hardy et al. [263] found that these multiple differences when observed by news media resulted in the media producing polarizing articles and visualizations, enhancing the overall polarizing effect. Whereas, Westfall et al. [264] found people typically overestimate, perceptually, the polarization. Particularly, their perceived attitudes are often more extreme than their actual attitudes toward the "other" political party. With NOVA we seek to provide another dimension as a platform to allow individuals to assess their perceived attitudes against a media outlet when presented with what content these outlets produced via real data.

To achieve this, our design needs to consider the implications of several findings from visualization research that addressed communication bias [249, 254, 15, 265]. These works demonstrate the effectiveness of letting an individual discover their personal beliefs via personalized data-driven visualizations. Koonchanok et al. [266] offers a system to explore belief elicitation, where people can test their beliefs against data-driven visualizations. They found when people are pressed to test their beliefs against data they tend to be more deliberate in their analysis. With NOVA, we apply lessons from personal belief visualization in the form of providing a means for people to inject their personal beliefs into the system and see how that adjusts the visualizations. We focus on supporting news media coverage assessment while keeping the users conscious about their analytical decisions as it relates to their personal beliefs.

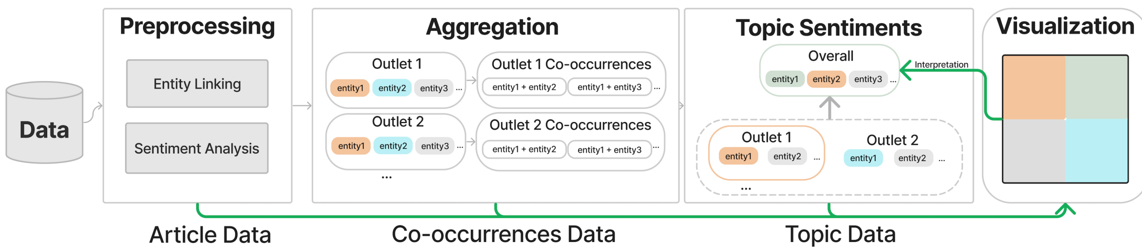


Figure 7.2: The data transformation process NOVA. Collected news articles were preprocessed with entity linking and sentiment analysis. Then articles are aggregated by entities and further aggregated by co-occurrences to represent topics. Sentiment scores are generated with descriptive statistics for each topic. The preprocessed article data, co-occurrences data, and topic sentiment data are all requested from the frontend. Through user interaction, the aggregated sentiment type of each entity is categorized as neutral, positive, negative, or mixed. Green lines indicate data communication between the backend and the frontend.

7.3 NOVA Design Considerations

With NOVA we have two main objectives that inform our design considerations. The first is, NOVA is intended for general audiences to freely assess mainstream media coverage on a variety of topics. In this work, we are focused on topics around COVID-19; however, NOVA is generalizable to any news topic. Secondly, NOVA should serve as a platform for facilitating the assessment of an individual's personal beliefs about these outlets. We derive four primary considerations (**DCs**) to help shape our design to meet the aforementioned objectives.

- **DC1: Support sense making.** News media coverage can be diverse and overwhelming to process, especially for the general public. One cannot assume the average person always has adequate context for what is presented. Thus, NOVA must ensure all visual content is easy to interpret and the public has adequate context. Additionally, the design must refrain from overwhelming the public with unnecessary details, as they are non–experts and have free–choice in participation.

- **DC2: Limit External Bias.** An objective of NOVA is to allow people to assess if their personal belief aligns with what outlets actually reported. To ensure any revelations are authentic we must ensure NOVA does not introduce any external bias which could confound people's assessment.
- **DC3: Encourage Personal Belief Elicitation.** The design of NOVA should encourage people to adjust the system parameters and visualizations reflecting their personal beliefs as described in 7.2.3. Additionally, the visualization design should reflect how their personal beliefs may differ from the data and what the implication of the change means.
- **DC4: Separate hypothesis generation and evaluation.** The design of NOVA should facilitate the generation of hypotheses and the ability to evaluate them within our system. Since we are accommodating personal beliefs into our system we need to separate the two stages. Where in the first stage based on their personal beliefs individuals can form hypotheses and conjectures on how an outlet may report, in the second stage, it is clear their hypotheses are being tested against data that is uninfluenced by both the designer and users.

The purpose of **DC1** is to ensure NOVA is accessible to the general public and easy to understand and interpret. Thus, the design of NOVA is modular and allows complexity to be built up according to an individual's request. We must ensure not to overwhelm our audience and maintain their interest during their assessment of the content and make it as effortless as possible for them to understand and make sense of what they see. From the existing literature, see Section 2, we see the effectiveness of text summarization and sentiment analysis on news articles as a good method to onboard our audience into the data.

To further support **DC1**, we leverage narrative visualization and data storytelling techniques into our design as a way to passively help our audience make sense of the data. From narrative visualization, we utilize techniques to sequence our audience through the content and give them control of the pace by which information is presented to them. Our design utilizes an interactive presentation structure [1] as well as a blend of author and reader–driven [66] strategies to retain our audience's attention as they assess their hypothesis.

If NOVA is to achieve its objective as a platform that can allow people to assess their personal beliefs, then NOVA itself must not introduce or greatly limit the introduction of any additional biases. One form of assessment we anticipate users to look for is examining patterns in how outlets frame a topic. The framing [267] of an outlet is represented by the articles they publish. If different patterns of framing emerge between different subsets of articles, users can hypothesize that there is bias involved. If a user catches this by using NOVA and it either affirms or disproves a personal belief then that is a success of our system. However, prior works [268, 15] illustrate how these framing biases can be introduced by human labeling or hard-coded parameters. That is, our design might introduce an external bias that could lead to a false assessment. Thus, to keep in line with **DC2** our design of NOVA needs to restrict human-generated bias in both data processing and visualization design.

To address our second objective of NOVA, the design needs to allow people to easily express and communicate their personal beliefs and have that reflected in the visualization once expressed. Thus with **DC3**, NOVA should encourage people to enter adjustments to parameters based on their personal beliefs. However, we should avoid encouraging users to exploit or even enhance their own biases leading them to report possibly false narratives extracted from NOVA. To account for this, **DC4**, it should be clear the differences between the hypothesis generation and evaluation process. Therefore, after users form a hypothesis based on their personal beliefs they understand they are later testing it against data that is agnostic of those beliefs.

7.4 Data Processing

Our project has been collecting news articles since January 2020 and is currently still collecting any COVID-19-related articles from the aforementioned outlets. The data is stored in a Postgres database and hosted on an AWS EC2 server. For each article, the source, title, content, time published, and URL are recorded in the database. The data transformation process is depicted in Figure 7.2. The pre-processing and aggregation stages are pre-computed and hosted on our server. The frontend takes input from user interaction to interpret and conduct a final transformation to visualize the data. We perform entity linking and sentiment analysis on each article in the preprocessing stage.

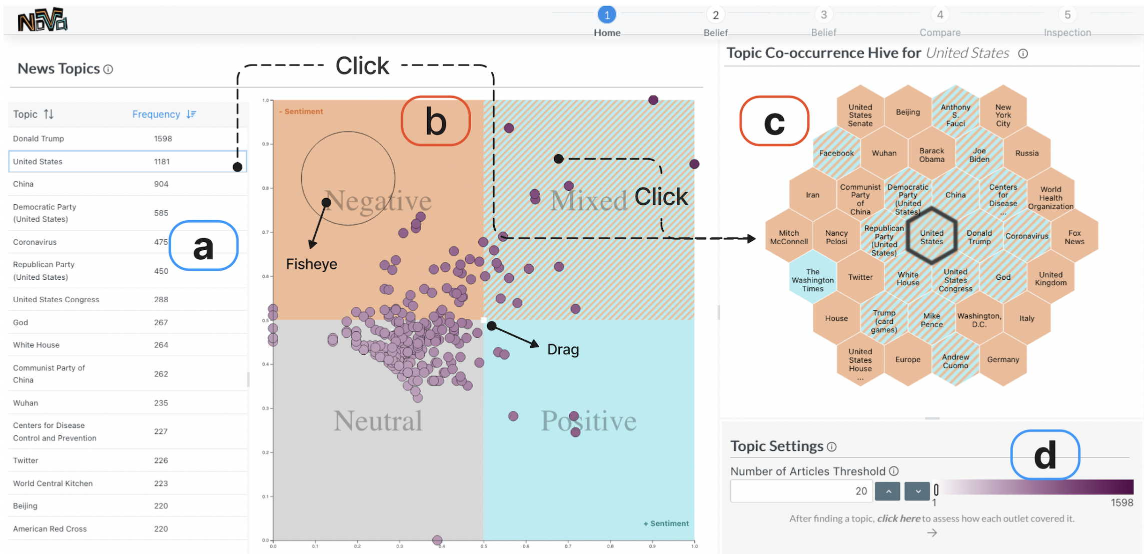


Figure 7.3: A screenshot of the Topic Selection stage. The user is presented with visualizations of the extracted topics. (a) An interactive table showing a list of entities and their number of mentioned articles. (b) The topic scatter plot shows two-dimensional sentiment for each topic. The region is divided into four categories to separate polarizing (mixed), positive, negative, and neutral topics. The segmentation point can be dragged to adjust by the user. (c) Clicking any row in (a) or dot in (b) triggers the visualization of frequently co-occurring topics with the selected topic in (c) using a topic co-occurrences hive. (d) Utility panel providing a filter on an entity based on article frequency and a color scale on article frequency.

Entity Linking. An *Entity* (or *Named Entity*) is a word or phrase mentioned in the unstructured text that references real-world objects, such as a person, an organization, or a location. *Named Entity Recognition* extracts mentions in text and assigns a *type* to each mention. *Entity Linking* further assigns a unique identifier (e.g.Wikipedia page ID or any URI) for each extracted mention, therefore providing a "link" to different mentions of the same entity. For example, "Donald Trump", "Trump", and "President Trump" are different mentions referring to the same person, which can be recognized by an entity linking model.

NOVA uses the extracted and linked entities to aggregate news articles. Each aggregated result represents a subset of articles related to a similar topic. Previous works generated *topics* by incorporating topic modeling techniques. We use entities to represent topics rather than running topic modeling for two reasons: (1) Current topic modeling techniques are not suitable for our targeted tasks. For supervised topic modeling, there is no dataset with proper labeling that contains a universal set of topics touched by news media. For unsupervised topic modeling techniques such as LDA [269], the results are incomprehensible for our target audiences, **DC1**. (2) On the other hand, named entities have the advantage of being self-explanatory and the potential to provide a rich context for the topic if different entities are frequently co-occurring.

We used the Radboud Entity Linker (REL) model proposed in [270] for our entity linking task. REL detects and assigns a Wikipedia ID to each entity, allowing us to link entities mentioned in different articles. The REL model is able to extract entities at document-level, but to incorporate the subsequent sentence-level target-dependent sentiment analysis, we ran REL on sentence-level. We first split each news article into sentences and assign a sentence index to each sentence. Then we extract entities from each sentence using the REL model and record the sentence index if an entity is identified, along with the extracted entities. This information is useful for identifying sentiments targeting an entity, which will be described next. In addition, the REL result contains the start offset and text length of each entity mentioned, which is used in the *Article Reviewer* described in subsubsection 7.5.1.4 to annotate extracted entities.

Sentiment Analysis. The sentiment analysis result is core to the identification of media bias in NOVA; therefore, the extracted sentiment needs to be both accurate and interpretative. We use NewsSentiment [271], a target-dependent sentiment classification model trained for news articles to meet **DC2**. NewsSentiment runs on top of the named entity linking result described in section 7.4 and assigns a sentiment type (positive, negative, or neutral) to each entity in the sentence, describing how the sentence is targeting the entity. Note that if more than one entity appears in a sentence, each entity sentiment is assigned independently to guarantee accuracy. To do that, for each pair of *(sentence, entity)*, we use the offset information in the REL result to separate the sentence into three parts: *(left, entity, right)*. The NewsSentiment model then takes the triple as input and classifies it as positive, negative, or neutral.

To ensure simplicity and therefore interpretability, we further aggregate the sentence-level results into document-level. For each news article, we run NewsSentiment on each sentence in the article that mentions at least one entity and classifies the sentence as being positive, negative, or neutral towards the mentioned entity. This sentiment result is independent for each entity in the sentence. Then for each entity mentioned in an article, we take $max(\#pos\_sentences, \#neg\_sentences, \#neu\_se$ as the target-dependent document-level sentiment.

7.5 NOVA: Interface & Visualization Design

As mentioned earlier, NOVA has two objectives; (1) to allow general audiences to freely assess mainstream media coverage on a variety of topics and (2) to serve as a platform to facilitate the assessment of one's personal beliefs toward these outlets. To address the first objective we take advantage of narrative visualization techniques and strategies to pace out the assessment and allow users to have more control over their process. To support the second objective NOVA facilitates belief elicitation by allowing users to express their beliefs about the six outlets as well as adjust certain parameters of the application, which are also reflected in our visualization design.

7.5.1 Interface Design

NOVA's interface design stems from the interactive presentation narrative structure [1], where we have 4 stages: *Topic Selection*, *Belief Elicitation*, *Outlet Comparison*, and *Article Reviewer*. The design intends to provide a loose structure to help our users in their assessment of news outlets, **DC1**. Additionally, the stages are separated to initially facilitate hypothesis generation and conclude with hypothesis evaluation, **DC4**. To support user engagement and ease of use, while users focus on hypothesis generation and verification, we keep track of certain decisions and interactions users make in previous stages and carry them over onto later stages to help lessen the cognitive burden. We incorporate modals and annotations to help direct user attention and to provide context for what they are assessing. Throughout NOVA, our system offers a means for users to express their personal beliefs, if they choose to do so, **DC3**. The system, in the first three stages, adjusts the visualizations to reflect their beliefs allowing them to form a hypothesis based on their existing beliefs and then evaluate them in the final stage.

7.5.1.1 Stage One: Topic Selection

The first stage, Topic Selection, is focused on easing users into the system and allowing them to get a sense (**DC1**) of the data and pick a topic to delve into. As shown in Figure 7.3-b, the Topic Scatter uses the same encoding described in subsubsection 7.5.2.1 on subsets of articles related to a specific topic.

The generation of sentiment score $s = (score_{pos}, score_{neg})$ of each topic is described later.

Users can manually set the segmentation point by dragging the segmentation controller to any position determined by their own observations or background knowledge. After the adjustments, they can select any topic that appeals to them. For example, a user may be curious and select a topic in the *mixed* region as mixed implies that the topic is polarizing, having both a high amount of positive articles and a high amount of negative articles. In Figure 7.3-d, we provide filtering functionality where one can set a threshold for the minimum number of articles associated with a topic to ensure they only browse topics with adequate coverage.

After selecting a topic, the *Topic Co-occurrence Hive*, described in subsubsection 7.5.2.2, is rendered. This visualization depicts other topics that frequently co-occur in articles with the selected topic, as shown in Figure 7.3-c. In stage one, the hive facilitates making sense (**DC1**) of this collection of articles and helps identify interesting topics for further assessment. From the hive, users can quickly speculate what aspects of the selected topic were covered by the media and which had more attention. The sentiment type, denoted by color, for each hexagon, can be influenced based on personal belief, **DC3**, by dragging the segment point controller in the *Topic Scatterplot*.

In this stage, our design presents many opportunities for users to inject their personal beliefs into NOVA. This can be achieved by changing the segmentation rules such that the hive reflects their own personal "information cocoon", based on their background knowledge of the topic. This cocoon will then be unpacked in stage two, *Outlet Comparison*.

7.5.1.2 Stage Two: Belief Elicitation

The Belief Elicitation stage takes user input on perceived media outlet coverage. Five hives are presented to the user and they are asked to choose any of the hives that most closely match what he/she believes the outlet would cover the shown topics. This input is used in the next stage to generate a user-perceived version of media outlet coverage. To not overwhelm and bore the user with six questions, we randomly select two of the six outlets, therefore reducing to only two selections. After two such selections, we use linear interpolation/extrapolation to predict the remaining four outlet covers. Specifically, for each outlet, we generate five hexes: one corresponding to the user-controlled segmentation point Figure 7.4, and another four for each corner of a square. The side length of the square is a hyperparameter of the system, which

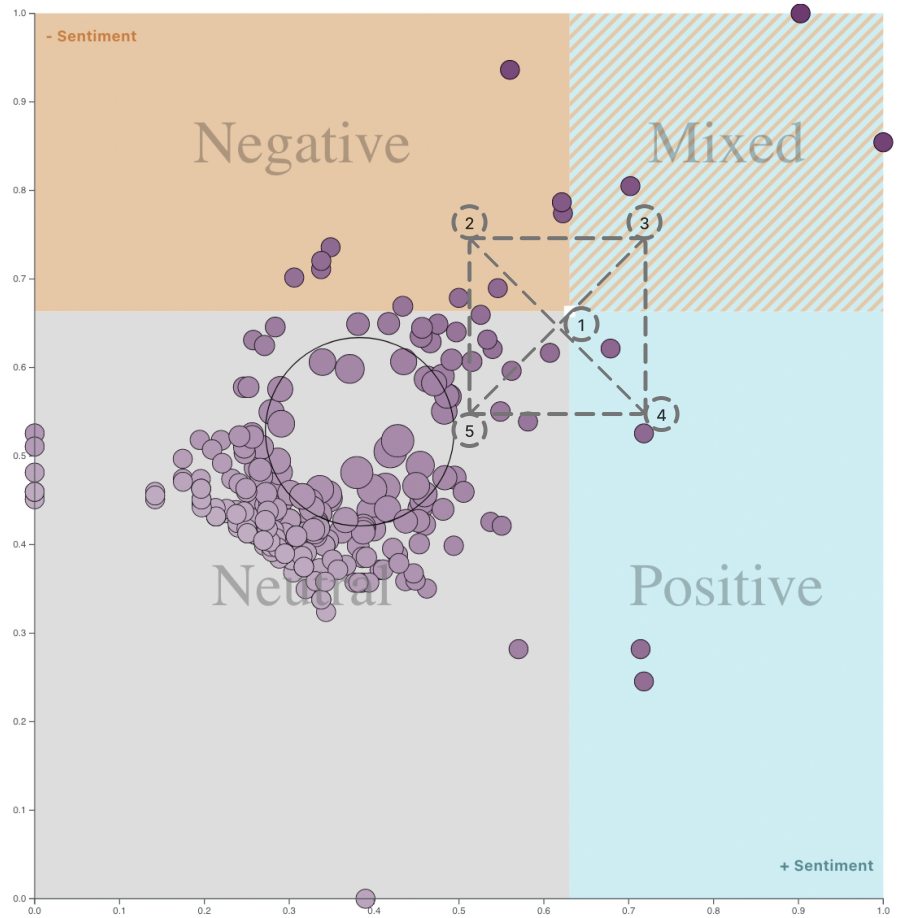


Figure 7.4: Belief hex generation. A square is generated where the center (1) is the user-controlled segmentation point. The hexes are made using the segmentation point corresponding to each corner (2–5). The circle represents the fish-eye lens that is drawn at the mouse cursor allowing for easier topic selection.

we set to 0.3 based on preliminary user studies. The square corners are clamped on the borders of the scatterplot. By this definition, the user selection of the hive then assigns a coordinate on the scatterplot to the surveyed outlet. The system collects two such coordinates for a random selection of two outlets and uses linear interpolation/extrapolation to automatically generate a coordinate for all other four outlets. The interpolation/extrapolation is based on the media bias chart in AllSides [1]. The result is then used to generate a predicted version of user perception on all six outlets in the next stage.

[1]https://www.allsides.com/media-bias/media-bias-chart

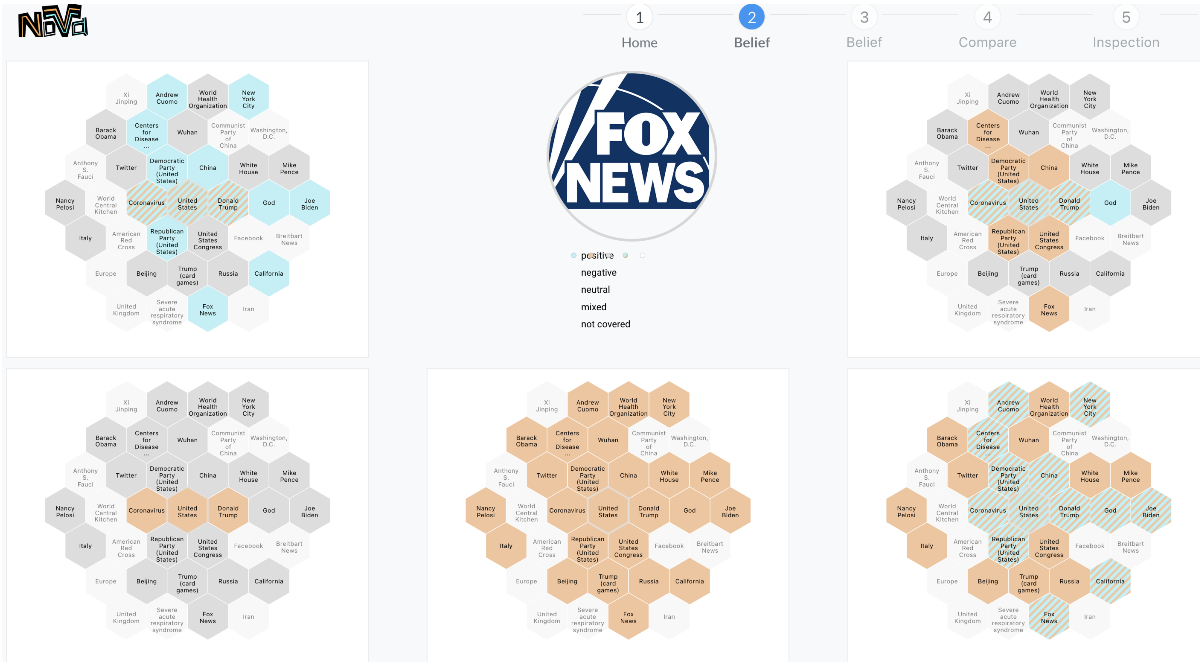


Figure 7.5: Belief Elicitation. The user is presented with two subsequent survey questions, each with five generated hives and a random outlet. They are prompted to select any of the generated hives that align closest with their perception of how the outlet would cover the selected topic. The input is then linear interpolated/extrapolated to generate a user-perceived version of outlet coverage on the selected topic.

7.5.1.3 Stage Three: Outlet Comparison

In this stage, the design focus shifts facilitating hypothesis generation (**DC4**). A depiction of each news outlet's coverage of a selected topic is visualized via a *Topic Co-occurrence Grid*, as shown in Figure 7.6-a. The grid presents multiple hives, one for each outlet in juxtaposing positions with aligned co-occurring topics. The system automatically selects the most frequently co-occurring topics combining all outlets and choosing the top k(k=36 in Figure 7.6) ones. If an outlet did not cover a specific topic, the corresponding hexagon is left faded out. All co-occurring topics have the same relative position in the hive for easy cross-comparison and pattern recognition. For example, in Figure 7.6-a, we can see how the selected topic of *Donald Trump* for **New York Times** and **Breitbart** has more negative cells compared to the other outlets. At the same time, *Donald Trump* is surrounded by all negative cells in **Washington Post**, but *Donald Trump* himself is mixed. Observations like these can lead to questions as to why NYT and Breitbart are different, or why Donald Trump is not negative when all co-occurring

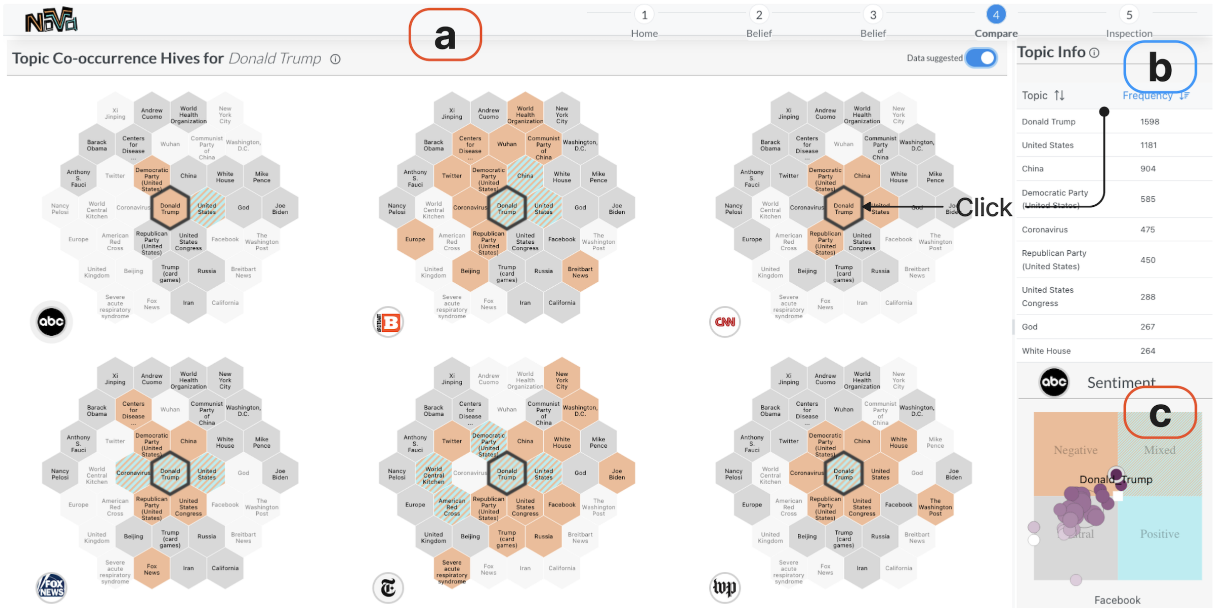


Figure 7.6: The Outlet Comparison stage provides an overview visualization of news outlet coverage. (a) A grid layout showing the topic co-occurrences hive for each media outlet. The topics all have fixed positions for easy cross-comparison. A toggler on the top-right corner allows users to switch between their perception and the data-suggested version of the hives. (b) A sidebar with an interactive table. Users can click on the rows to change the center of hives. (c) Sentiment Scatterplot showing the sentiment scores for co-occurring entities of the selected outlet.

topics are negative. A sentiment scatter plot in Figure 7.6-c provides additional information about the co-occurring topics. Once the users seek to unpack the details behind an outlet's perceived behavior they can click the "Inspection" button to continue on.

7.5.1.4 Stage Four: Article Reviewer

The Article Reviewer stage focuses on hypothesis assessment, **DC4**. In this stage, we present *Article View*(Figure 7.7-a), which depicts to users the articles associated with the initially selected topic from stage one and a co-occurring topic selected from stage three. The upper half of Article View(Figure 7.7-a) lists the article's headlines, classified as positive or negative with keywords extracted as described in section 7.4. We place the list of positive and negative articles in juxtaposition with each other to make it easier to compare. Selecting a headline will load the *Article Reviewer*(Figure 7.7-b), showing the full content of the selected article with selected entities highlighted. The *Notes Panel*(Figure 7.7-c) is positioned next to the *Article Reviewer Panel* with any and all of their prior observations.

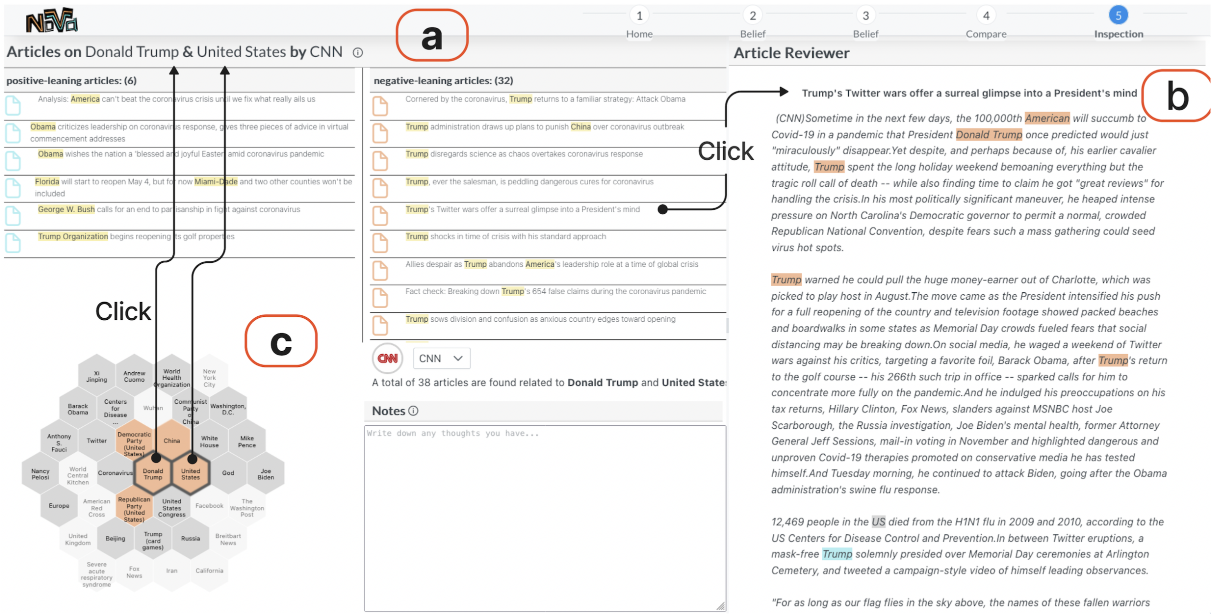


Figure 7.7: Article Reviewer stage. (a) Article Panel showing positive and negative articles on the selected topic in separate columns. (b) depicts the annotated content of the selected article. (c) The hive renders content from the selected media outlet allowing the functionality to switch topics or outlets. Notes Panel for documenting insights.

The *Outlet Coverage* view (Figure 7.7-c) also serves as a reminder of users' previous line of inquiry up to this point. They can use this hive to switch to a different co-occurring topic or switch to a different media outlet and filter, which in turn loads different articles to browse. At this point, users either have found evidence via the articles that help address their initial question and may feel done, or they may seek to return to stage one and start up a new line of inquiry. To return to the start, they can use the browser back button or click on the *Steps* in the top right corner. All observations and selections are saved and they are free to cycle back and forth with NOVA till they are satisfied.

7.5.2 Visualization Design

A challenge in visually assessing the coverage of media outlets is to present polarizing topics separately from neutral ones. Through our data analysis, we found a way to distinguish these from one another; however, we needed a simple and interpretative way to visualize these results. With our *Sentiment Scatter Plot*, we address this issue by visualizing a 2D sentiment of a topic where each quadrant implication is easy for our users to understand. Additionally, we needed a simple visualization that could express coverage differences as well as depict common ground

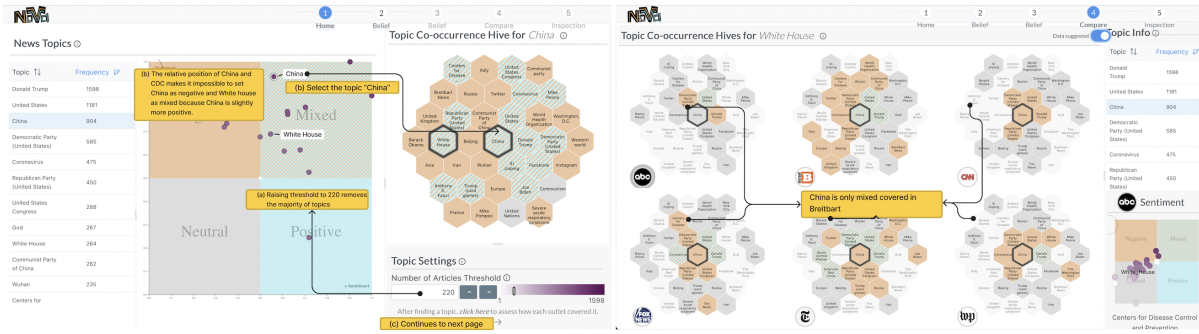


Figure 7.8: The interactions observed during the example workflow. Through the interactions indicated by (a), (b), and (c), the user selects China as an interesting topic and then goes to the next stage. In Outlet Comparison Stage, the user then discovers an interesting pattern and goes to Article Reviewer Stage to review the articles published by Breitbart.

among media outlets; consequently, we designed and visualized *Topic Co-occurrences Hive* that works in tandem with *Sentiment Scatter Plot*.

7.5.2.1 Sentiment Scatter Plot

Currently, state-of-the-art sentiment analysis techniques are not mature enough to be conducted beyond document-level. However, the assessment of coverage of any topic stems from a collection of articles. Basic sentiment analysis techniques are ineffective when applied directly to a topic or when used to summarize a collection of articles. Researchers typically avoid going beyond document-level or use descriptive statistics to measure the sentiment of a collection of articles. A problem with using descriptive statistics (i.e., mean) is that the results for the following two cases are often indistinguishable: (1) when most articles are neutral and (2) when articles have equally high positive and negative scores. In the second case, the scores negate each other and the results often look similar to the mean of the first case. Thus, a topic that contains coverage with articles having both high positive scores and high negative scores may be misrepresented as neutral. Such topics in actuality are under mixed coverage or polarizing and can be potentially interesting to study further. Therefore, it is critical that these cases are separated.

To achieve that with NOVA, we treat positive scores and negative scores as two independent variables associated with the coverage of the topic and plot each topic as a dot on a scatter plot. We use a scatter plot as the general public is quite familiar with this representation and understand how to interpret and parse information from it. The positive score and negative score of a topic's coverage are descriptive statistics of the associated positive and negative articles, respectively. We use min-max normalization on the count of articles for the descriptive statistic to normalize the scores into [0, 1]. Using a min-max normalization makes sense because the absolute sentiment scores of each topic give no information when analyzing coverage bias. Instead, relative scores are more likely to reveal coverage bias. As explained in Equation 7.1, we calculate $score_t$ for each topic $t \in T$, where T represents the collection of topics extracted from the data set as described in section 7.4. Then we plot each topic using $score_{pos}$ as x-axis and $score_{neg}$ as y-axis. We further encode $count(articles)$ with a logarithmic color scale to assist users to understand the underlying statistics.

$$
\begin{aligned}
&\forall t \in T, \quad score_t = (score_{pos}, score_{neg}), \\
&score_{pos} = min\_max\_norm(count(positive\ articles)) \\
&score_{neg} = min\_max\_norm(count(negative\ articles))
\end{aligned}
\tag{7.1}
$$

To help users understand the meaning of a polarizing topic as well as assess the overall coverage of the topics, we use a *segmentation point* to explicitly divide the scatter plot into four regions: *neutral, positive, negative, and mixed*. Topics that fall into each region are automatically classified respectively. The segmentation point controller can be moved by users to change the segmentation rule. Providing the segmentation point controller for users adheres to **DC2** and **DC3**. Previous works commonly used hard-coded thresholds (e.g., 0.33) for dividing positive, negative, and neutral scores. The hard-coded threshold would be problematic in our case since our sentiment analysis model is not fine-tuned on a media bias data set considering **DC2**. The consequence of not fine-tuning is two-fold. The sentiment analysis result distribution will not be centered around 0. Instead, the distribution is likely to be shifted or have "bias" under the context of "bias and variance" in machine learning terminology. On the other hand, as we can see from Figure 7.3-a, the model captures an important characteristic of news media: news articles tend to be negatively toned. A hard-coded threshold may lower trust from users if they find a topic being classified in a way they don't necessarily agree with. Providing the controller for users to change the threshold value not only solves this problem but also reflects **DC3** encouraging personal belief elicitation, in that users can adjust the threshold based on their background knowledge of the topics. Finally, we incorporate fish-eye interaction to reduce cluttering, which is a user-controlled focus point for indicating which part of the scatterplot is to be zoomed. A circle around the cursor indicates the area to be zoomed. When the cursor moves towards a particular node, the node will be attracted along the moving direction to enable easy clicking.

7.5.2.2 Topic Co-occurrences Hive

To help users make sense of the articles contained within a topic we visualize these relationships as a Topic Co-occurrence Hive. We utilize a hive metaphor to represent “information cocoons”, the notion that we are more willing to accept new information that supports our beliefs, and this, in turn, can isolate some of us in a separate informational reality. A hive shows a deeper level of detail by visualizing the context of a selected topic (**DC1**). The center cell in the hive represents a topic selected from the *Sentiment Scatter Plot*. All the cells that surround the center represent frequently co-occurring topics with the initially selected topic. This representation enables users to speculate what context or events the news outlets were or are covering as it relates to their selected topic. The color of each cell is assigned by the segmentation rule specified in *Sentiment Scatter Plot*. Thus, if a user changed the segmentation rule, the result is reflected in the hive as well. The position of each hexagon is not strictly encoded; however, the closer to the center hexagon, the more frequently that topic co-occurs with the selected topic.

As shown in Figure 7.6, each outlet has its own “hive” for a specific topic with diverse coverage differences. We encourage users to emerge from their “information cocoons” and learn about what other information cocoons look like. The Topic Co-occurrences Hive is used across all stages of the system, and at each stage, it is used differently to facilitate different tasks. The Topic Selection stage mainly serves to provide background and context of a selected topic. In the Outlet Comparison stage, we create hives for each news outlet to compare the differences in coverage and see common grounds.

7.6 Usage Scenario: Coverage of China

Let us walk through an example scenario of how one could use NOVA Figure 7.8. We begin with our user opening up NOVA in her web browser and viewing the Topic Selection stage. After closing the modal, the stage only shows the Topic table and the Sentiment Scatter on the left-hand side, with a text prompt on the right-hand side asking the user to hover over the news topics. She notices a majority of topics are clustered in the lower left region labeled as neutral. However, she is interested in topics that are well covered and raises the article number threshold to 220. After the adjustment, she recognizes several topics, which she personally associates to be either negative or polarizing from her memory of that time. For example, she hovers over the topics of *Donald Trump* and the *White House* and sees that they are considered mixed, which aligns with her personal view of the time as *Donald Trump* was a polarizing figure. She then clicks on *China* because she can't recall why China would be classified under mixed, as she recalls a lot of negativity around *China* during the beginning of COVID-19 and believes that it should fall under the *Negative* region. She then adjusts the segmentation controller such that the topic categorization fits her recollection. However, she quickly realizes that perhaps her perception of China may be skewed. One indication of this is if China is in the negative region, then the White House can not be in the mixed region. This is because according to the data, China receives more positivity than the White House. This observation conflicts with the user's perception of how China was covered by the mainstream media so she wants to verify if this is indeed true. She focuses her attention on the topic co-occurrence hive, which loads in after selecting China, to assess the validity of China being labeled mixed coverage and not just negative. The hive shows other topics that co-occur frequently with articles that refer to China. She is unsurprised to learn that *Wuhan* and the *Communist Party of China* co-occur frequently and negatively, but *Donald Trump* being mixed instead of negative is something she did not anticipate.

However, she is still curious as to why there is consistently mixed sentiment about China across all six outlets, so she continues to the next stage. After navigating to the next stage she is asked to select one of the five presented hives that align closest with what she believes NYT coverage for China would have been. She selects a hive based on her prior understanding of media outlets and their coverage. This process repeats again on the next page for a different outlet and then she is directed to the Outlet Comparison stage.

In the Outlet Comparison stage, she sees a hive of *China* for each outlet, which is generated based on her selection in the previous stage. The overall prediction matches her expectations, but when she switched to the true reporting version of the six hives, she found that some parts of her perceived version does not align with the reported version. For example, she expects that Breitbart would be negative on China, but the data suggests that Breitbart is actually mixed on China, implying Breitbart covers China more positively than she expects. She then navigates to the Article Reviewer Stage to review what Breitbart has published about *China*.

Upon arriving in the Article Reviewer stage, she reviews positive-leaning articles on China. Upon close inspection, she finds that the sentiment analysis model is misclassifying *Chinese coronavirus* or *China Virus* as referring to *China* rather than coronavirus or COVID-19. From her own recollection, she associates the word Chinese coronavirus as negative against China and notices it is frequently used by Breitbart. She believes that these articles are biased in terms of word choice, suggesting that China is responsible for the outbreak. This close inspection confirms her perception of the coverage of China and avoids her from being misled by the errors of the sentiment analysis model. The user then notes down her observations (Figure 7.9). In addition, she writes down two questions after finding this error. The first question is *How is the COVID-19 situation actually being covered by the mainstream media?* The second question is *How is China actually being covered by other media outlets?*

To address the first question, she goes back to the Outlet Comparison stage and uses the topic table to switch to coronavirus. This time, she finds most outlets are neutral or negative towards coronavirus except Fox News. This implies that Fox News is more positive on coronavirus coverage than other outlets. She clicks Fox News and goes to the Article Reviewer to find articles. She finds Fox News has significantly more positive-learning articles than other outlets. In particular, these articles depict a positive situation of the COVID-19 outbreak, stating the outbreak is under control and praising Donald Trump for his handling of the situation. According to her personal experience of that time and how the pandemic turns out, she believes that these articles are biased. She then writes her findings for the first question.

As for her second question, she goes back to the Outlet Comparison Stage and switches the main topic to China. She finds that New York Times, a media outlet that she believes to be fair and neutral, is also quite negative about China. She navigates to the Article Reviewer Stage and again finds that New York Times was publishing quite a lot of negative articles on

China. In one particular article talking about the propaganda pushed by China, the article first described the propaganda and how it does not reflect reality, which she believes to be fair and neutral. However, the article then goes on to reference unverified stories and unfounded claims to fight back, which she believes have become biased. Through NOVA, our user was able to freely explore and investigate both her own biases as well as assess media outlet reporting.

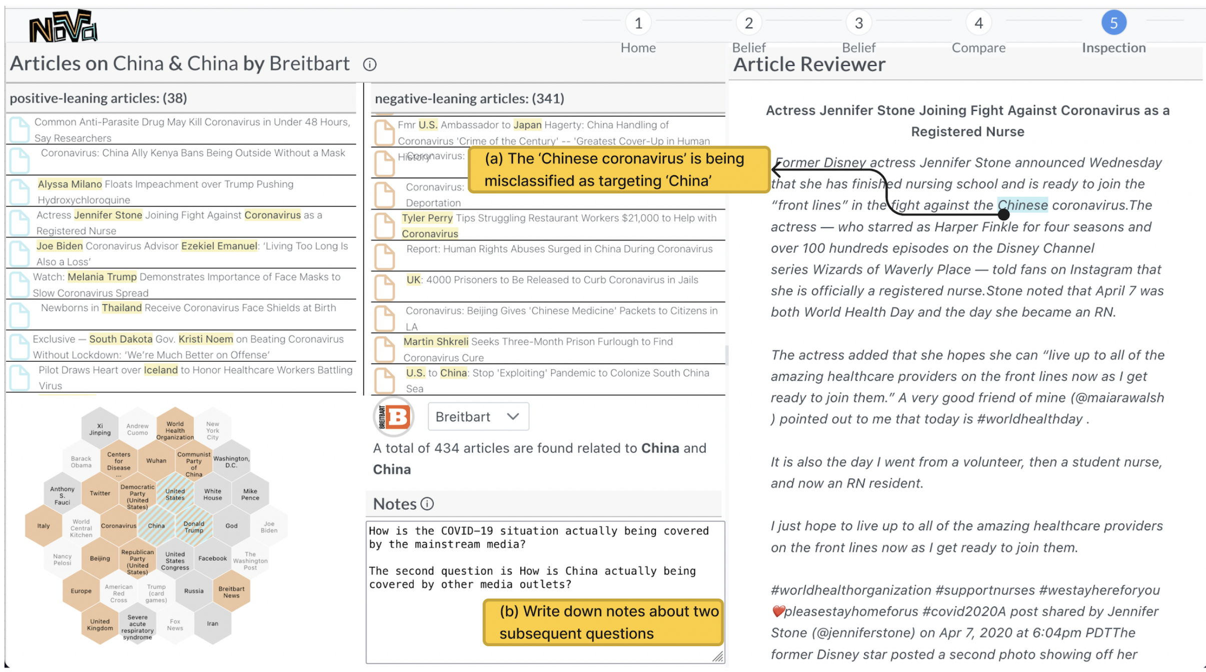


Figure 7.9: A user discovers an error produced by the sentiment analysis model and formulates two subsequent questions. The user documents the questions in the notes panel and continues their investigation within the Article Reviewer.

7.7 Evaluation

To evaluate NOVA, we performed a user study to measure the usefulness and generalizability of NOVA. We report our qualitative analysis results based on recorded videos and survey answers. Participants were given 2 tasks. Task 1 requires participants to find a reported topic of interest using NOVA and write down their expectations for how it was covered and the rationale. Task 2 requires participants to assess their expectations using NOVA, by comparing differences in sentiment between their perception and the actual reporting while writing down any insights. The goal of Task 1 is to evaluate the effectiveness of NOVA in capturing users' beliefs on mainstream media outlet coverage for a selected topic, while the goal of Task 2 is to ascertain if NOVA is effective in facilitating the assessment of their beliefs toward media outlets. We performed thematic coding with user survey results and categorized user motivations to derive what motivates user interactions in NOVA and what aspects can be further considered when designing visualization systems for personal belief assessment.

In our study, we recruited 12 people[2] to use NOVA. Before each session, each participant would fill out a survey of their U.S. news consumption as well as provide demographic information. From the initial survey, 75% of our participants actively seek out news content, with a majority (66.7%) reading the news more than once a day. Most of our participants access their news via social media (75%) and/or an online subscription (50%). All our participants had at a minimum read an article or seen an article headline within 2 weeks of taking the study from at least one of the 6 U.S-based media outlets we gathered articles from. After filling out the initial survey and tutorial on NOVA, they would then be directed to our application in a Google Chrome browser.

NOVA Facilitates Topic Exploration from Personal Experience: While using NOVA, participants were able to apply their personal experiences about news outlets and prior context to investigate why a topic was assigned a sentiment. Many participants (8/12) tended to select topics that evoked a memory rather than sentiment or volume of articles. A majority of our participants (10/12) derived an insight or were surprised about media outlet coverage.

Participant01: *Not sure if this should be under positive-leaning?*

Participant07: *[NOVA] is easy to navigate and look at topics reported among 6 new sources. Color coding helped me see what was negative or positive about the reporting. I was able to verify that negative or positive leaning with the article reviewer.*

Participant09: *The New York Times isn't as fair as I thought [their] results were consistently negative.*

Additionally, several (9/12) found NOVA was fairly low-effort to use and identify these insights. However, a few did find it challenging to contextualize.

[2]The demographic breakdown of the evaluation participants was: 75% between ages 20–29, 17% between ages 30-39, and 8% above 60, with 6 females and 6 males.

Participant01 *Not too hard to do what I want to do with [NOVA], a lot of the views explain what they do.*

Participant04 *I never felt like I was out of control, however at times I felt I was spending a lot of time putting [the big picture] all together.*

NOVA Helps Users Develop Various Exploration Strategies: Participants had small differences in their workflows through NOVA. A subset (5/12) was drawn to the polarizing topics to gain context as to why it was polarizing and the differences between coverage. Whereas others (6/12) would seek out outlets they had a predisposition towards to verify if that expectation aligned.

Participant08: *Since the presentation was very neutral it encouraged me to explore outlets that I usually don't agree with politically, which lead to some interesting insights and somewhat reduced my bias towards them.*

Participant06: *I liked the breakdown of the various articles into positive and negative and subject matter. Illustrating the individual outlets for each subject was also insightful.*

Participant02 *I would want to use this tool scaled up to international news outlet comparison*

NOVA Supports the Generation of Questions on Reporting: Most participants (10/12) formulated questions about media reporting or questioned why their prior expectations were misaligned. Half (6/12) were able to either validate their expectation of a media outlet or reveal they were biased about the outlet.

Participant05: *[I was] surprised that media impressions I had were not borne out by data*

Participant08: *[NOVA] confirmed some [of my] biases but revealed some unexpectedly good coverage of outlets that I usually do not consume.*

Participant04: *The visual representations of the [topics] in the bee-hive along with the adjustments when comparing news outlets opened my perspective on my own biases.*

From our evaluation, one immediate direction for future work is to improve the transparency of the sentiment model to our users, as well as allow them to flag sentiments they disagree with.

However, most of our participants were able to derive insights and assess their beliefs about mainstream media coverage.

7.8 Limitations

NOVA allows general audiences to freely assess media coverage while facilitating the assessment of one's personal beliefs toward media outlets. However, NOVA does have some limitations. First, using entities to represent topics, although provides the benefit of not requiring a labeled dataset and being self-explanatory, is still limited in terms of accuracy. Some irrelevant articles still occur in Article Reviewer stage. Also as a result of lacking a human-level performance model that can extract the main characters or events, we did not incorporate the use of temporal information provided by our dataset. This is because we can not maintain a consistent thread of storyline evolution from a collection of articles without a human-level performance model; especially, when our target audiences are the general public, any inaccuracy or inconsistency could lead to confusion. Second, the performance of our system is limited by the accuracy of our sentiment analysis model. Since we do not allow users to change the result of sentiment classification, a misclassified article is potentially misleading. Fortunately, the modular design of our system makes it compatible with any improvement in sentiment analysis. At last, we intend to provide an additional functionality on facilitating the analysis of user-marked articles in the future. We believe this functionality will benefit our users by supporting self-assessment of their personal beliefs.

7.9 Conclusion

We present NOVA a visual assessment interface to support two main objectives. The first is NOVA should be accessible for general audiences to freely assess mainstream media coverage on a variety of topics. Secondly, we designed NOVA such that it could also serve as a platform for facilitating the assessment of an individual's personal beliefs for these outlets. To demonstrate and evaluate NOVA we used a subset of our larger dataset that contains articles written during the beginning of the COVID-19 Pandemic in the United States (Feb–Jun 2020). To help reduce the mental demand and overload of information to our users, NOVA utilizes narrative visualization techniques and structures to help sequence users through the rich content and reduce some of this demand. We designed several visualizations, such as the Co-Occurrence Hive and the Sentiment Scatter to be readily decipherable and accessible to general audiences when used to assess media outlet coverage.

From our evaluation, we found NOVA was effective in enabling users in assessing their personal beliefs about news outlets. Additionally, we found our Co-Occurrence Hive visualization was very effective. Participants were able to quickly make sense of, DC1, the context behind their selected topic, and were able to then frame and assess the outlets. However, half of our participants felt the mental effort to work with NOVA was quite high. When asked to elaborate further, some expressed that using NOVA is simple but contextualizing the topics with their memory of what happened was a lot to keep track of. For future work, we plan to add more provenance improvements to better assist users in keeping track of the current line of inquiry and offer more context to help lower the overall mental effort. NOVA's design is flexible to be easily extended or adapted for other scenarios. From our evaluation exit survey, several participants made note of how useful NOVA could be for research in sociology as well as other domains that deal with document-level data. Many participants, during their workflows, would either question and evaluate the sentiment score or scrutinized several articles until concluding the effectiveness of the sentiment model. This suggests that NOVA's design may be effective if repurposed to focus on NLP model assessment and training, as well as the dynamic improvement of models via user input.

Chapter 8
Conclusion

With this dissertation, I offer a set of frameworks to improve our understanding of how we communicate through data visualizations as well as new approaches for communicating through data visualization. This work began with the hope of documenting and understanding how data visualizations can be better utilized in precisely communicating what we as authors, designers, storytellers, teachers, etc., intend. For this purpose, I delved into domains of linguistics, philosophy of language, communications, journalism, media bias, sociology, psychology, marketing, visual and literary storytelling, semiotics, and of course data visualization.

There is a growing need to share our knowledge of interpreting and communicating through visualizations. As mentioned, data visualization is ubiquitous in our daily lives. They appear in some form on almost everything from appliances due to IoT, to our media via news or video games, it is deeply embedded in our personal health tracking and healthcare. Overall, we utilize them in all facets of employment, managing our finances, and creative pursuits. As we become more familiar with these representations their complexity and nuance will grow. This is already happening. For example, when I started working in this field in 2016 on the *Sea of Genes* collaboration with the Exploratorium, at one stage during our brainstorms we tossed around the idea of communicating the metagenomic relationships using a Sankey Diagram as the base. The veteran museum curators, who have designed many complex interactive exhibits both physical and digital, felt from their prior works and experience that the average visual literacy of the general visiting public was not high enough to decode a Sankey Diagram before the fatigue would set in. So we pivoted our design to be catered to something more familiar and approachable to attract these visitors and retain their attention. At the time of writing this, it is 2023, and visual literacy in at least the United States has improved dramatically. Mainstream outlets are sharing fairly complex charts and even bespoke visualization to convey information to the general public. These charts are even being shared and discussed on public forums like Reddit and Twitter. It is an inspiring time for those who have an affinity and love for all things data visualization.

My current work and ongoing work are centered on the idea and goal that there exists a medium or language that can be used to communicate and share all complex findings and topics with anyone from any background. This current work is an attempt to use interactive data-driven visualizations as a means to create this bridge between complex unfamiliar concepts and some audience. I have used data-driven visual metaphors and data stories to convey fairly complex concepts to lay audiences. From these works, I learned the challenges of designing visual metaphors and the misunderstandings or misinterpretations that can arise. With data stories, I developed a framework, Character-Oriented Design, to identify data characters, and I presented a design space for developing these characters and how they can be applied in the storytelling process. Working closely with stories and the concepts of through-line and the larger goal of ensuring the core messages reach the audience, I became interested in better understanding designer intent. I sought to identify which design elements in our interactive visualizations are reflecting our overall intentions and how. Data visualizations are often very interactive and used to communicate information. One means to understand this communication is to view the design elements in terms of what greater task it serves. I was interested in understanding how the lower-level elements serve to accomplish the designer's intent. As there are many different designs that can accomplish a task; however, the design can be influenced and visually altered based on the underlying intentions of the designer. With VisActs, I offer my approach to inferring these intentions in the design. A broader goal with VisActs is to apply it to a corpus of visualizations to be used as inputs to automate the authoring and ideating of new visualizations. Throughout my work, I have learned that it is necessary to understand our intentions and be able to assess and track how we express them to really understand why our communicative goal either fails or succeeds. With NOVA, I developed a system to visually capture the user's beliefs and contrast them with the data. The goal is to facilitate an assessment of whether their beliefs align with the data. NOVA's design applied both my VisActs work and Character-Oriented Design as I needed a central character to guide the users through this as well as ensure the overall design remained neutral.

To conclude, I believe as visualizations become more commonplace, there will be a need to study the deeper nuances and communicative effects that lay within data visualizations. I hope this work serves as a base or start for (1) how versatile data visualizations can be for communication, (2) a guide for how to use data stories to communicate, and (3) a start for analyzing the presence and implications that designer's communicative intent have on their data visualizations.

References

[1] E. Segel and J. Heer, "Narrative visualization: Telling stories with data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1139–1148, 2010.

[2] C. Tong, R. Roberts, R. Borgo, S. Walton, R. S. Laramee, K. Wegba, A. Lu, Y. Wang, H. Qu, Q. Luo, and X. Ma, "Storytelling and visualization: An extended survey," *Information*, vol. 9, no. 3, 2018.

[3] W. Willett, B. A. Aseniero, S. Carpendale, P. Dragicevic, Y. Jansen, L. Oehlberg, and P. Isenberg, "Superpowers as inspiration for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 2021, 2021.

[4] J. S. Risch, "On the role of metaphor in information visualization," *arXiv preprint arXiv:0809.0884*, 2008.

[5] E. Adar and E. Lee, "Communicative visualizations as a learning problem," *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[6] Y.-N. Li, D.-J. Li, and K. Zhang, "Metaphoric transfer effect in information visualization using glyphs," in *Proceedings of the International Symposium on Visual Information Communication and Interaction*. New York, NY: ACM, 2015, pp. 121–130.

[7] J. Wu, W. Cai, D. Watkins, and J. Glanz, "How the virus got out," https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html, 2020, new York Times.

[8] M. Greis, A. Joshi, K. Singer, A. Schmidt, and T. Machulla, "Uncertainty visualization influences how humans aggregate discrepant information," in *Proceedings of the Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[9] J. Hullman, "Why authors don't visualize uncertainty," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 130–139, 2019.

[10] N. Crilly, D. Good, D. Matravers, and P. J. Clarkson, "Design as communication: exploring the validity and utility of relating intention to interpretation," *Design Studies*, vol. 29, no. 5, pp. 425–457, 2008.

[11] K. Krippendorff, "On the essential contexts of artifacts or on the proposition that" design is making sense (of things)"," *Design issues*, vol. 5, no. 2, pp. 9–39, 1989.

[12] M. Barnard, *Graphic design as communication*. Routledge, 2013.

[13] D. George, "From analysis to design: Visual communication in the teaching of writing," *College composition and communication*, pp. 11–39, 2002.

[14] H. C. Purchase, N. Andrienko, T. J. Jankun-Kelly, and M. Ward, "Theoretical foundations of information visualization," in *Information Visualization*, 2008, pp. 46–64.

[15] J. Hullman and N. Diakopoulos, "Visualization rhetoric: Framing effects in narrative visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2231–2240, 2011.

[16] Z. Liu, N. Nersessian, and J. Stasko, "Distributed cognition as a theoretical framework for information visualization," *IEEE transactions on visualization and computer graphics*, vol. 14, no. 6, pp. 1173–1180, 2008.

[17] M. Chen and H. Jäenicke, "An information-theoretic framework for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1206–1215, 2010.

[18] G. Kindlmann and C. Scheidegger, "An algebraic process for visualization design," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2181–2190, 2014.

[19] P. C. Parsons, "Conceptual metaphor theory as a foundation for communicative visualization design," in *IEEE VIS Workshop on Visualization for Communication (VisComm 2018)*, 2018.

[20] J. Mackinlay, "Automating the design of graphical presentations of relational information," *Acm Transactions On Graphics (Tog)*, vol. 5, no. 2, pp. 110–141, 1986.

[21] P. Hanrahan, "Vizql: a language for query, analysis and visualization," in *Proceedings of the International Conference on Management of Data*, 2006, pp. 721–721.

[22] J. Heer and M. Bostock, "Declarative language design for interactive visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1149–1156, 2010.

[23] 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, 2011.

[24] J. K. Li and K.-L. Ma, "P5: Portable progressive parallel processing pipelines for interactive data analysis and visualization," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1151–1160, 2019.

[25] 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, 2016.

[26] A. Ojo and B. Heravi, "Patterns in award winning data storytelling: Story types, enabling tools and competences," *Digital journalism*, vol. 6, no. 6, pp. 693–718, 2018.

[27] L. Yang, X. Xu, X. Lan, Z. Liu, S. Guo, Y. Shi, H. Qu, and N. Cao, "A design space for applying the freytag's pyramid structure to data stories," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 922–932, 2022.

[28] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, and N. Henry Riche, "Design patterns for data comics," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York, NY: IEEE, 2018, pp. 1–12.

[29] E. M. Forster, *Aspects of the Novel*. RosettaBooks, 2010.

[30] J. Truby, *The anatomy of story: 22 steps to becoming a master storyteller*. Farrar, Straus and Giroux, 2008.

[31] J. Campbell, *The hero with a thousand faces*. New World Library, 2008, vol. 17.

[32] C. Fritz and G. Tosello, "The hidden meaning of forms: methods of recording paleolithic parietal art," *Journal of archaeological method and theory*, vol. 14, no. 1, pp. 48–80, 2007.

[33] S. Field, *Screenplay: The foundations of screenwriting*. Delta, 2005.

[34] P. Isenberg, B. Lee, H. Qu, and M. Cordeil, "Immersive visual data stories," in *Immersive Analytics*. New York, NY: Springer, 2018, pp. 165–184.

[35] R. Bran, "Message in a bottle telling stories in a digital world," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 1790–1793, 2010.

[36] N. Gershon and W. Page, "What storytelling can do for information visualization," *Communications of the ACM*, vol. 44, no. 8, pp. 31–37, 2001.

[37] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale, "More than telling a story: Transforming data into visually shared stories," *IEEE Computer Graphics and Applications*, vol. 35, no. 5, pp. 84–90, 2015.

[38] K.-L. Ma, I. Liao, J. Frazier, H. Hauser, and H.-N. Kostis, "Scientific storytelling using visualization," *IEEE Computer Graphics and Applications*, vol. 32, no. 1, pp. 12–19, 2012.

[39] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: automatic generation of annotated stock visualizations," in *Proceedings of the Conference on Human Factors in Computing Systems*. New York, NY: ACM, 2013, pp. 2707–2716.

[40] C. Bryan, K.-L. Ma, and J. Woodring, "Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 511–520, 2016.

[41] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe, "Chartaccent: Annotation for data-driven storytelling," in *IEEE Pacific Visualization Symposium*. New York, NY: IEEE, 2017, pp. 230–239.

[42] Q. Wang, Z. Li, S. Fu, W. Cui, and H. Qu, "Narvis: Authoring narrative slideshows for introducing data visualization designs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 779–788, 2019.

[43] F. Amini, N. Henry Riche, B. Lee, C. Hurter, and P. Irani, "Understanding data videos: Looking at narrative visualization through the cinematography lens," in *Proceedings of Conference on Human Factors in Computing Systems*. New York, NY: ACM, 2015, pp. 1459–1468.

[44] J. D. Bradbury and R. E. Guadagno, "Documentary narrative visualization: Features and modes of documentary film in narrative visualization," *Information Visualization*, vol. 19, no. 4, pp. 339–352, 2020.

[45] E. J. Fink, *Dramatic story structure: A primer for screenwriters*. Routledge, 2014.

[46] Y. Shi, C. Bryan, S. Bhamidipati, Y. Zhao, Y. Zhang, and K.-L. Ma, "Meetingvis: Visual narratives to assist in recalling meeting context and content," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 6, pp. 1918–1929, 2018.

[47] E. Coats, "Pixar's 22 rules of storytelling," https://www.aerogrammestudio.com/2013/03/07/pixars-22-rules-of-storytelling/, accessed: 2022-02-22.

[48] D. Shen, "Edgar Allan Poe's Aesthetic Theory, the Insanity Debate, and the Ethically Oriented Dynamics of ""The Tell-Tale Heart""," *Nineteenth-Century Literature*, vol. 63, no. 3, pp. 321–345, 12 2008.

[49] J. Hullman, S. Drucker, N. H. Riche, B. Lee, D. Fisher, and E. Adar, "A deeper understanding of sequence in narrative visualization," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2406–2415, 2013.

[50] Y. Kim, K. Wongsuphasawat, J. Hullman, and J. Heer, "Graphscape: A model for automated reasoning about visualization similarity and sequencing," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York, NY: ACM, 2017, pp. 2628–2638.

[51] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani, "Authoring data-driven videos with dataclips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 501–510, 2016.

[52] J. Boy, F. Detienne, and J.-D. Fekete, "Storytelling in information visualizations: Does it engage users to explore data?" in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY: ACM, 2015, pp. 1449–1458.

[53] K. Dasu, K.-L. Ma, J. Ma, and J. Frazier, "Sea of genes: A reflection on visualising metagenomic data for museums," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 935–945, 2020.

[54] M. Cavazza, F. Charles, and S. J. Mead, "narrative representations and causality in character-based interactive storytelling," *Proceedings of CAST*, pp. 139–142, 2001.

[55] ——, "Characters in search of an author: Ai-based virtual storytelling," in *International Conference on Virtual Storytelling*. Springer, 2001, pp. 145–154.

[56] Y. Cai, C. Miao, A.-H. Tan, and Z. Shen, "A hybrid of plot-based and character-based interactive storytelling," in *Technologies for E-Learning and Digital Entertainment*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 260–273.

[57] F. Charles, S. J. Mead, and M. Cavazza, "Character-driven story generation in interactive storytelling," in *Proceedings International Conference on Virtual Systems and Multimedia*. IEEE, 2001, pp. 609–615.

[58] Y. Tanahashi and K.-L. Ma, "Design considerations for optimizing storyline visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2679–2688, 2012.

[59] R. Gove, "Automatic narrative summarization for visualizing cyber security logs and incident reports," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1182–1190, 2021.

[60] T. Baumgartl, M. Petzold, M. Wunderlich, M. Hohn, D. Archambault, M. Lieser, A. Dalpke, S. Scheithauer, M. Marschollek, V. M. Eichel, N. T. Mutters, H. Consortium, and T. V. Landesberger, "In search of patient zero: Visual analytics of pathogen transmission pathways in hospitals," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 711–721, 2021.

[61] L. Morais, Y. Jansen, N. Andrade, and P. Dragicevic, "Showing data about people: A design space of anthropographics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 3, pp. 1661–1679, 2022.

[62] J. Boy, A. V. Pandey, J. Emerson, M. Satterthwaite, O. Nov, and E. Bertini, "Showing people behind data: Does anthropomorphizing visualizations elicit more empathy for human rights data?" in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York, NY: ACM, 2017, pp. 5462–5474.

[63] X. Lan, Y. Wu, Y. Shi, Q. Chen, and N. Cao, "Negative emotions, positive outcomes? exploring the communication of negativity in serious data stories," in *Proceedings of the Conference on Human Factors in Computing Systems*, 2022.

[64] J. Liem, C. Perin, and J. Wood, "Structure and empathy in visual data storytelling: Evaluating their influence on attitude," in *Computer Graphics Forum*. Wiley Online Library, 2020, pp. 277–289.

[65] Z. Zhao, R. Marr, and N. Elmqvist, "Data comics: Sequential art for data-driven storytelling," Univ. of Maryland, Tech. Rep., 2015.

[66] C. D. Stolper, B. Lee, N. H. Riche, and J. Stasko, "Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories," Microsoft Research, Washington, USA, Tech. Rep., 2016.

[67] S. Petridis and L. B. Chilton, "Human errors in interpreting visual metaphor," in *Proceedings of the Conference on Creativity and Cognition*, 2019, pp. 187–197.

[68] Z. Cai, Y.-N. Li, X. S. Zheng, and K. Zhang, "Applying feature integration theory to glyph-based information visualization," in *IEEE Pacific Visualization Symposium*. New York, NY: IEEE, 2015, pp. 99–103.

[69] A. Sallaberry, Y.-c. Fu, H.-C. Ho, and K.-L. Ma, "Contact trees: Network visualization beyond nodes and edges," *PLOS ONE*, vol. 11, no. 1, pp. 1–23, 01 2016.

[70] S. Wang, Y. Tanahashi, N. Leaf, and K.-L. Ma, "Design and effects of personal visualizations," *IEEE Computer Graphics and Applications*, vol. 35, no. 4, pp. 82–93, 2015.

[71] P. M. Cruz, J. Wihbey, A. Ghael, S. Costa, R. Chao, and F. Shibuya, "Process of simulating tree rings for immigration in the us," *IEEE VIS Arts Program Annotated Projects*, 2018.

[72] C. Bryan, K.-L. Ma, and J. Woodring, "Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 511–520, 2017.

[73] G. Russell and A. Petersen, "Analysis of cross category dependence in market basket selection," *Journal of Retailing*, vol. 76, no. 3, pp. 367–392, 2000.

[74] P. Chintagunta and S. Haldar, "Investigating purchase timing behavior in two related product categories," *Journal of Marketing Research*, vol. 35, no. 1, pp. 43–53, 1998.

[75] J. Chung and V. R. Rao, "A general choice model for bundles with multiple-category products: Application to market segmentation and optimal pricing for bundles," *Journal of Marketing Research*, vol. 40, no. 2, pp. 115–130, 2003. [Online]. Available: https://doi.org/10.1509/jmkr.40.2.115.19230

[76] R. J. Williams, A. Howe, and K. S. Hofmockel, "Demonstrating microbial co-occurrence pattern analyses within and between ecosystems," *Frontiers in Microbiology*, vol. 5, p. 358, 2014.

[77] A. R. Feinstein, "The pre-therapeutic classification of co-morbidity in chronic disease," *Journal of Chronic Diseases*, vol. 23, no. 7, pp. 455–468, 1970.

[78] M. Divo, C. Cote, J. P. de Torres, C. Casanova, J. M. Marin, V. Pinto-Plata, J. Zulueta, C. Cabrera, J. Zagaceta, G. Hunninghake *et al.*, "Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease," *American Journal of Respiratory and Critical Care Medicine*, vol. 186, no. 2, pp. 155–161, 2012.

[79] T. M. Koppie, A. M. Serio, A. J. Vickers, K. Vora, G. Dalbagni, S. M. Donat, H. W. Herr, and B. H. Bochner, "Age-adjusted charlson comorbidity score is associated with treatment decisions and clinical outcomes for patients undergoing radical cystectomy for bladder cancer," *Cancer*, vol. 112, no. 11, pp. 2384–2392, 2008.

[80] C. Zhou, Y.-L. Wu, G. Chen, J. Feng, X.-Q. Liu, C. Wang, S. Zhang, J. Wang, S. Zhou, S. Ren *et al.*, "Erlotinib versus chemotherapy as first-line treatment for patients with advanced egfr mutation-positive non-small-cell lung cancer (optimal, ctong-0802): a multicentre, open-label, randomised, phase 3 study," *The Lancet Oncology*, vol. 12, no. 8, pp. 735–742, 2011.

[81] E. Karni and D. Schmeidler, "Self-preservation as a foundation of rational behavior under risk," *Journal of Economic Behavior and Organization*, vol. 7, no. 1, pp. 71 – 81, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167268186900223

[82] G. A. Kaplan, S. A. Everson, and J. W. Lynch, "The contribution of social and behavioral research to an understanding of the distribution of disease: a multilevel approach," in *Smedley and Syme. Promoting Health: Intervention Strategies from Social and Behavioral Research*. National Academy Press, 2000, pp. 31–55.

[83] D. Constanza, R. Carlos, G. Lara, M. Nieves, P. Gemma, V. Sergi, T. Rosa, R. J. A., and C. Miguel, "Attention deficit hyperactivity disorder in cocaine-dependent adults: A psychiatric comorbidity analysis," *The American Journal on Addictions*, vol. 22, no. 5, pp. 466–473, 2013.

[84] J. F. Tellez-Zenteno, S. B. Patten, N. Jetté, J. Williams, and S. Wiebe, "Psychiatric comorbidity in epilepsy: A population-based analysis," *Epilepsia*, vol. 48, no. 12, pp. 2336–2344, 2007.

[85] F. G. Davis, B. J. McCarthy, S. Freels, V. Kupelian, and M. L. Bondy, "The conditional probability of survival of patients with primary malignant brain tumors," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 85, no. 2, pp. 485–491, 1999.

[86] T.-L. Fung, J.-K. Chou, and K.-L. Ma, "A design study of personal bibliographic data visualization," in *Pacific Visualization Symposium*. New York, NY: IEEE, 2016, pp. 244–248.

[87] K. Dasu, S. Bae, T. Fujiwara, and K.-L. Ma, "Learning about disease associations in Taiwan," IEEE PacificVis 2018 Visual Storytelling Contest, https://k-dasu.github.io, accessed: 2018-6-11.

[88] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE Transactions on visualization and computer graphics*, vol. 12, no. 5, pp. 741–748, 2006.

[89] H. C. Purchase, "Metrics for graph drawing aesthetics," *Journal of Visual Languages & Computing*, vol. 13, no. 5, pp. 501–516, 2002.

[90] N. Cawthon and A. V. Moere, "The effect of aesthetic on the usability of data visualization," in *Information Visualization*. IEEE, 2007, pp. 637–648.

[91] T. L. Fung and K.-L. Ma, "Visual characterization of personal bibliographic data using a botanical tree design," in *IEEE VIS Workshop on Personal Visualization: Exploring Data in Everyday Life*, vol. 15. New York, NY: IEEE, 2015.

[92] M. Moldovan, R. Enikeev, S. Syed-Abdul, P. A. Nguyen, Y.-C. Chang, and Y. C. Li, "Disease universe: Visualisation of population-wide disease-wide associations," *Advances in Systems Science and Applications*, vol. 14, no. 2, pp. 144–158, 2014.

[93] N. K. Dubey, S. Syed-Abdul, P. A. Nguyen, R. Dubey, U. Iqbal, Y.-C. Li, W.-H. Chen, and W.-P. Deng, "Association between anxiety state and mitral valve disorders: A taiwanese population-wide observational study," *Computer Methods and Programs in Biomedicine*, vol. 132, pp. 57–61, 2016.

[94] G. Rozenberg and A. Salomaa, *The mathematical theory of L systems*. Academic press, 1980, vol. 90.

[95] C. LaPointe and D. Stiert, "Volume lightning rendering and generation using l-systems," *Advanced Computer Graphics*, 2009.

[96] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.

[97] National Health Research Institutes, "National Health Insurance Research Database, Taiwan," https://nhird.nhri.org.tw/en/, accessed: 2018-7-29.

[98] CMS and NCHS, "The international classification of diseases: 9th revision, clinical modification: ICD-9-CM," 1991.

[99] B. J. Cohen and Y. Gibor, "Anemia and menstrual blood loss." *Obstetrical and Gynecological Survey*, vol. 35, no. 10, pp. 597–618, 1980.

[100] K. Rubtsova, P. Marrack, and A. V. Rubtsov, "Sexual dimorphism in autoimmunity," *The Journal of Clinical Investigation*, vol. 125, no. 6, pp. 2187–2193, 2015.

[101] C. M. Kunin, "Urinary tract infections in females," *Clinical Infectious Diseases*, vol. 18, no. 1, pp. 1–10, 1994.

[102] W. C. Miller, C. A. Ford, M. Morris, M. S. Handcock, J. L. Schmitz, M. M. Hobbs, M. S. Cohen, K. M. Harris, and J. R. Udry, "Prevalence of chlamydial and gonococcal infections among young adults in the united states," *JAMA*, vol. 291, no. 18, pp. 2229–2236, 2004.

[103] M. D. Fossey and R. B. Lydiard, "Anxiety and the gastrointestinal system." *Psychiatric Medicine*, vol. 8, no. 3, pp. 175–186, 1990.

[104] M. M. Moschos, "Physiology and psychology of vision and its disorders: a review," *Medical Hypothesis, Discovery and Innovation in Ophthalmology*, vol. 3, no. 3, p. 83, 2014.

[105] "Science On a Sphere," https://sos.noaa.gov/What_is_SOS/, National Oceanic and Atmospheric Administration (NOAA).

[106] F. Block, M. S. Horn, B. C. Phillips, J. Diamond, E. M. Evans, and C. Shen, "The DeepTree exhibit: Visualizing the tree of life to facilitate informal learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2789–2798, 2012.

[107] D. Smit, T. Grah, M. Murer, V. van Rheden, and M. Tscheligi, "Macroscope: First-person perspective in physical scale models," in *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*. New York, NY: ACM, 2018, pp. 253–259.

[108] J. Ma, I. Liao, K.-L. Ma, and J. Frazier, "Living Liquid: Design and evaluation of an exploratory visualization tool for museum visitors," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2799–2808, 2012.

[109] C.-H. Hsueh, J. Chu, K.-L. Ma, J. Ma, and J. Frazier, "Fostering comparisons: Designing an interactive exhibit that visualizes marine animal behaviors," in *IEEE Pacific Visualization Symposium*. New York, NY: IEEE, 2016, pp. 259–263.

[110] T. Geller, "Interactive tabletop exhibits in museums and galleries," *IEEE Computer Graphics and Applications*, vol. 26, no. 5, pp. 6–11, 2006.

[111] H. Schmidt, U. Hinrichs, A. Dunning, and S. Carpendale, "memory [en]code Building a Collective Memory within a Tabletop Installation," in *Computational Aesthetics in Graphics, Visualization, and Imaging*. Aire-la-Ville, Switzerland: The Eurographics Association, 2007.

[112] U. Hinrichs and S. Carpendale, "Gestures in the wild: Studying multi-touch gesture sequences on interactive tabletop exhibits," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2011, p. 3023–3032.

[113] N. R. Council, *The new science of metagenomics: Revealing the secrets of our microbial planet*. National Academies Press, 2007.

[114] S. Perkins and R. DeSalle, "The secret world inside you," https://www.amnh.org/exhibitions/the-secret-world-inside-you/, american Museum of Natural History.

[115] "Invisible you-the human microbiome exhibition," http://www.edenproject.com/visit/whats-here/invisible-you-the-human-microbiome-exhibition, eden project.

[116] "Zoo in you," https://omsi.edu/exhibitions/zoo-in-you/, oregon Museum of Science and Industry (OMSI).

[117] Y.-Y. Chan and H. Qu, "Finavistory: Using narrative visualization to explain social and economic relationships in financial news," in *International Conference on Big Data and Smart Computing*. New York, NY: IEEE, 2016, pp. 32–39.

[118] C. Bryan, K.-L. Ma, and J. Woodring, "Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 511–520, 2017.

[119] L. Bedford, "Storytelling: The real work of museums," *Curator: The Museum Journal*, vol. 44, no. 1, pp. 27–34, 2001.

[120] J. Rounds, "Storytelling in science exhibits," *Exhibitionist*, vol. 21, no. 2, pp. 40–43, 2002.

[121] S. Allen, *Finding significance*. Exploratorium, 2004.

[122] J. Ma, "Engaging museum visitors with scientific data through visualization: A comparison of three strategies," in *Annual Meeting of the American Educational Research Association*. Washington, DC: The American Educational Research Association, 2013, pp. 1–18.

[123] B. Tversky, J. B. Morrison, and M. Betrancourt, "Animation: Can it facilitate?" *International Journal of Human-Computer Studies*, vol. 57, no. 4, pp. 247–262, 2002.

[124] M. Hegarty, S. Kriz, and C. Cate, "The roles of mental animations and external animations in understanding mechanical systems," *Cognition and Instruction*, vol. 21, no. 4, pp. 209–249, 2003.

[125] R. Moreno and R. Mayer, "Interactive multimodal learning environments," *Educational Psychology Review*, vol. 19, no. 3, pp. 309–326, 2007.

[126] W. Schnotz, J. Böckheler, and H. Grzondziel, "Individual and co-operative learning with interactive animated pictures," *European Journal of Psychology of Education*, vol. 14, no. 2, pp. 245–265, 1999.

[127] E. L. Ferguson and M. Hegarty, “Learning with real machines or diagrams: Application of knowledge to real-world problems,” *Cognition and Instruction*, vol. 13, no. 1, pp. 129–160, 1995.

[128] J. B. Morrison, B. Tversky, and M. Betrancourt, “Animation: Does it facilitate learning,” Association for the Advancement of Artificial Intelligence, Palo Alto, California, Tech. Rep. SS-00-04, 2000.

[129] S. Ainsworth, “How do animations influence learning,” in *Recent innovations in educational technology that facilitate student learning*. Charlotte, NC: Information Age Publishing, 2008, pp. 37–67.

[130] S. Berney and M. Bétrancourt, “Does animation enhance learning? a meta-analysis,” *Computers & Education*, vol. 101, pp. 150–167, 2016.

[131] L.-J. ChanLin, “Animation to teach students of different knowledge levels,” *Journal of Instructional Psychology*, vol. 25, no. 3, pp. 166–175, 1998.

[132] S. Kalyuga, “Relative effectiveness of animated and static diagrams: An effect of learner prior knowledge,” *Computers in Human Behavior*, vol. 24, no. 3, pp. 852–861, 2008.

[133] R. Ploetzner and R. Lowe, “A systematic characterisation of expository animations,” *Computers in Human Behavior*, vol. 28, no. 3, pp. 781–794, 2012.

[134] E. A. Ottesen, C. R. Young, S. M. Gifford, J. M. Eppley, R. Marin, S. C. Schuster, C. A. Scholin, and E. F. DeLong, “Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages,” *Science*, vol. 345, no. 6193, pp. 207–212, 2014.

[135] F. O. Aylward, J. M. Eppley, J. M. Smith, F. P. Chavez, C. A. Scholin, and E. F. DeLong, “Microbial community transcriptional networks are conserved in three domains at ocean basin scales,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 17, pp. 5443–5448, 2015.

[136] F. O. Aylward, D. Boeuf, D. R. Mende, E. M. Wood-Charlson, A. Vislova, J. M. Eppley, A. E. Romano, and E. F. DeLong, “Diel cycling and long-term persistence of viruses in the ocean’s euphotic zone,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 43, pp. 11 446–11 451, 2017.

[137] P. Langfelder and S. Horvath, “Wgcna: An r package for weighted correlation network analysis,” *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[138] E. Hornecker, ““I don’t understand it either, but it is cool”–Visitor interactions with a multi-touch table in a museum,” in *IEEE International Workshop on Horizontal Interactive Human Computer Systems*. New York, NY: IEEE, 2008, pp. 113–120.

[139] Y. Rogers and S. Lindley, “Collaborating around vertical and horizontal large interactive displays: which way is best?” *Interacting with Computers*, vol. 16, no. 6, pp. 1133–1152, 09 2004.

[140] B. Potvin, C. Swindells, M. Tory, and M.-A. Storey, “Comparing horizontal and vertical surfaces for a collaborative design task,” *Advances in Human-Computer Interaction*, vol. 2012, no. 6, p. 10 pages, 2012.

[141] J. H. Falk and L. D. Dierking, *Learning from museums: Visitor experiences and the making of meaning*. Altamira Press, 2000.

[142] U. Hinrichs, H. Schmidt, and S. Carpendale, "EMDialog: Bringing information visualization into the museum," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1181–1188, 2008.

[143] "Electron — Build cross platform desktop apps with JavaScript, HTML, and CSS," https://electronjs.org/.

[144] National Research Council, *Learning science in informal environments: People, places, and pursuits*. National Academies Press, 2009.

[145] P. Davis, M. Horn, L. Schrementi, F. Block, B. Phillips, E. M. Evans, J. Diamond, and C. Shen, "Going deep: Supporting collaborative exploration of evolution in natural history museums," in *Proceedings of the Conference on Computer Supported Collaborative Learning*, vol. 1. Madison, WI: International Society of the Learning Sciences, 2013, pp. 153–160.

[146] P. Bell, B. Lewenstein, A. W. Shouse, M. A. Feder *et al.*, *Learning science in informal environments: People, places, and pursuits*. National Academies Press Washington, DC, 2009, vol. 140.

[147] J. Ma, K.-L. Ma, and J. Frazier, "Decoding a complex visualization in a science museum–an empirical study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 472–481, 2019.

[148] G. Hein, "Traits of life: A collection of life science exhibits," Exploratorium, San Francisco, CA, Tech. Rep., March 2003.

[149] J. H. Falk and L. D. Dierking, *The museum experience revisited*. Routledge, 2016.

[150] J. Ma, "Visitors' prior knowledge and interests in marine microbes and metagenomics," Exploratorium, San Francisco, CA, Tech. Rep., jan 2011, "https://www.exploratorium.edu/sites/default/files/pdfs/visitors-prior-marine.pdf".

[151] NGSS, *Next generation science standards: For states, by states*. The National Academies Press, Washington, DC, 2013.

[152] A. N. Spiegel, J. McQuillan, P. Halpin, C. Matuk, and J. Diamond, "Engaging teenagers with science through comics," *Research in Science Education*, vol. 43, no. 6, pp. 2309–2326, 2013.

[153] A. D. Lanie, T. E. Jayaratne, J. P. Sheldon, S. L. Kardia, E. S. Anderson, M. Feldbaum, and E. M. Petty, "Exploring the public understanding of basic genetic concepts," *Journal of Genetic Counseling*, vol. 13, no. 4, pp. 305–320, 2004.

[154] E. Hornecker, P. Marshall, and Y. Rogers, "From entry to access: how shareability comes about," in *Proceedings of the Conference on Designing Pleasurable Products and Interfaces*. New York, NY: ACM, 2007, pp. 328–342.

[155] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism," *Spatial Vision*, vol. 3, no. 3, pp. 179–197, 1988.

[156] S. E. Palmer, *Vision science: Photons to phenomenology*. MIT press, 1999.

[157] R. E. Mayer and R. Moreno, "A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory." *Journal of Educational Psychology*, vol. 90, no. 2, pp. 312–320, 1998.

[158] R. Moreno, R. E. Mayer, H. A. Spires, and J. C. Lester, "The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?" *Cognition and Instruction*, vol. 19, no. 2, pp. 177–213, 2001.

[159] P. Ginns, "Meta-analysis of the modality effect," *Learning and Instruction*, vol. 15, no. 4, pp. 313–331, 2005.

[160] J. Heer, F. B. Viégas, and M. Wattenberg, "Voyagers and voyeurs: Supporting asynchronous collaborative information visualization," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY: ACM, 2007, pp. 1029–1038.

[161] P. Isenberg and S. Carpendale, "Interactive tree comparison for co-located collaborative information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1232–1239, 2007.

[162] D. P. Groth and K. Streefkerk, "Provenance and annotation for visual exploration systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1500–1510, 2006.

[163] P. A. Rauschnabel and A. C. Ahuvia, "You're so lovable: Anthropomorphism and brand love," *Journal of Brand Management*, vol. 21, no. 5, pp. 372–395, 2014.

[164] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[165] R. Spiro and J. Jehng, "Cognitive flexibility, random access instruction and hypertext: Theory and technology for the nonlinear and multi-dimensional traversal of complex subject matter," *Cognition, Education, and Multimedia: Exploring Ideas in High Technology*, pp. 163–205, 1990.

[166] E. Einsiedel, "The challenges of translating genomic knowledge," *Clinical Genetics*, vol. 70, no. 5, pp. 433–437, 2006.

[167] J.-A. Kang, S. Hong, and G. T. Hubbard, "The role of storytelling in advertising: Consumer emotion, narrative engagement level, and word-of-mouth intention," *Journal of Consumer Behaviour*, vol. 19, no. 1, pp. 47–56, 2020.

[168] H. Ç. Sarıca and Y. K. Usluel, "The effect of digital storytelling on visual memory and writing skills," *Computers & Education*, vol. 94, pp. 298–309, 2016.

[169] N. M. Dudukovic, E. J. Marsh, and B. Tversky, "Telling a story or telling it straight: The effects of entertaining versus accurate retellings on memory," *Applied Cognitive Psychology*, vol. 18, no. 2, pp. 125–143, 2004.

[170] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko, "Augmenting visualizations with interactive data facts to facilitate interpretation and communication," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 672–681, 2018.

[171] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang, "Datashot: Automatic generation of fact sheets from tabular data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 895–905, 2020.

[172] J. Heer and G. Robertson, "Animated transitions in statistical data graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1240–1247, 2007.

[173] F. Hohman, M. Conlen, J. Heer, and D. H. P. Chau, "Communicating with interactive articles," *Distill*, vol. 5, no. 9, p. e28, 2020.

[174] S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer, "Visual narrative flow: Exploring factors shaping data visualization story reading experiences," *Computer Graphics Forum*, vol. 36, no. 3, pp. 377–387, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13195

[175] L. Yang, X. Xu, X. Lan, Z. Liu, S. Guo, Y. Shi, H. Qu, and N. Cao, "A design space for applying the freytag's pyramid structure to data stories," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 922–932, 2022.

[176] Data Visualization Society, "Information Is Beautiful Awards," https://www.informationisbeautifulawards.com/, accessed: 2023-06-29.

[177] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao, "Calliope: Automatic visual data story generation from a spreadsheet," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 453–463, 2020.

[178] N. Halloran, "How sure are climate scientists, really?" https://www.youtube.com/watch?v=R7FAAfK78_M, 2021.

[179] T. N. Y. Times, "The world's ball," https://www.nytimes.com/interactive/2014/06/13/sports/worldcup/world-cup-balls.html, 2014.

[180] S. Scarr, "Covid-19: The pace of death," https://www.reuters.com/graphics/HEALTH-CORONAVIRUS/DEATHS/xlbpgobgapq/index.html, Sep 2020.

[181] W. Woolf, https://twitter.com/WillikinWolf/status/1176006515968241665, accessed: 2022-02-22.

[182] GAPMINDER.ORG, https://www.gapminder.org/answers/will-saving-poor-children-lead-to-overpopulation/, accessed: 2022-02-22.

[183] K. Azad, "Colorized math equations," https://betterexplained.com/articles/colorized-math-equations/, 2019.

[184] P. Interactive, "Out of sight, out of mind."

[185] B. Bach, https://datacomics.github.io/, accessed: 2022-02-22.

[186] Distill, "Four experiments in handwriting with a neural network," https://distill.pub/2016/handwriting/, 2016.

[187] Reuters, "How powerful was the beirut blast?" https://graphics.reuters.com/LEBANON-SECURITY/BLAST/yzdpxnmqbpx/index.html, 2020.

[188] T. Chu, "Money wins elections," http://letsfreecongress.org/, 2014.

[189] B. Lee, P. Isenberg, N. H. Riche, and S. Carpendale, "Beyond mouse and keyboard: Expanding design considerations for information visualization interactions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2689–2698, 2012.

[190] H. K. Bako, X. Liu, L. Battle, and Z. Liu, "Understanding how designers find and use data visualization examples," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1048–1058, 2023.

[191] G. J. Quadri and P. Rosen, "A survey of perception-based visualization studies by task," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 5026–5048, 2022.

[192] M. A. Schoenlein, J. Campos, K. J. Lande, L. Lessard, and K. B. Schloss, "Unifying effects of direct and relational associations for visual communication," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 385–395, 2023.

[193] J. L. Austin, *How to do things with words*. Oxford university press, 1975, vol. 88.

[194] J. R. Searle, "Austin on locutionary and illocutionary acts," *The philosophical review*, vol. 77, no. 4, pp. 405–424, 1968.

[195] J. R. Searle and J. R. Searle, *Speech acts: An essay in the philosophy of language*. Cambridge university press, 1969, vol. 626.

[196] H. P. Grice, "Meaning," *The philosophical review*, vol. 66, no. 3, pp. 377–388, 1957.

[197] R. C. Stalnaker, "Context and content: Essays on intentionality in speech and thought," 1999.

[198] S. E. Murray and W. B. Starr, "Force and conversational states," *New work on speech acts*, pp. 202–236, 2018.

[199] L. Wilkinson, "The grammar of graphics," in *Handbook of computational statistics*. Springer, 2012, pp. 375–414.

[200] Simon Scarr, "Iraq's bloody toll," https://www.scmp.com/infographics/article/1284683/iraqs-bloody-toll, accessed: 2023-06-29.

[201] Andy Cotgreave, "Should you trust a data visualisation," https://gravyanecdote.com/uncategorized/should-you-trust-a-data-visualisation/, accessed: 2023-06-29.

[202] E. Dimara and C. Perin, "What is interaction for data visualization?" *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 119–129, 2019.

[203] C. Abras, D. Maloney-Krichmar, J. Preece *et al.*, "User-centered design," *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, vol. 37, no. 4, pp. 445–456, 2004.

[204] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," *Information Visualization*, vol. 8, no. 1, pp. 42–55, 2009.

[205] H.-N. Liang, P. C. Parsons, H.-C. Wu, and K. Sedig, "An exploratory study of interactivity in visualization tools:'flow'of interaction," *Journal of Interactive Learning Research*, vol. 21, no. 1, pp. 5–45, 2010.

[206] Z. Liu and J. Stasko, "Mental models, visual reasoning and interaction in information visualization: A top-down perspective," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 999–1008, 2010.

[207] K. Gadhave, J. Görtler, Z. Cutler, C. Nobre, O. Deussen, M. Meyer, J. Phillips, and A. Lex, "Predicting intent behind selections in scatterplot visualizations," *Information Visualization*, vol. 20, no. 4, p. 207–228, August 2021. [Online]. Available: https://doi.org/10.1177/14738716211038604

[208] A. Lau and A. V. Moere, "Towards a model of information aesthetics in information visualization," in *2007 11th International Conference Information Visualization (IV'07)*. IEEE, 2007, pp. 87–92.

[209] E. Lee-Robbins and E. Adar, "Affective learning objectives for communicative visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1–11, 2023.

[210] ——, "Affective learning objectives for communicative visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1–11, 2022.

[211] J. R. Searle, *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1985.

[212] L. J. Cohen, "Do illocutionary forces exist?" *The Philosophical Quarterly (1950-)*, vol. 14, no. 55, pp. 118–137, 1964.

[213] J. R. Searle and D. Vanderveken, "Speech acts and illocutionary logic," in *Logic, thought and action*. Springer, 1985, pp. 109–132.

[214] D. Sperber and D. Wilson, *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA, 1986, vol. 142.

[215] K. Bach and R. M. Harnish, *Linguistic communication and speech acts*. MIT Press, 1979.

[216] P. R. Cohen and C. R. Perrault, "Elements of a plan-based theory of speech acts," *Cognitive science*, vol. 3, no. 3, pp. 177–212, 1979.

[217] C. S. Peirce, *Peirce on signs: Writings on semiotic*. UNC Press Books, 1991.

[218] J. D. Johansen and S. E. Larsen, *Signs in use: An introduction to semiotics*. Psychology Press, 2002.

[219] B. Curtin, "Semiotics and visual representation," *Semantic Scholar*, 2009.

[220] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, "The science of visual data communication: What works," *Psychological Science in the public interest*, vol. 22, no. 3, pp. 110–161, 2021.

[221] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[222] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, "Decision making with visualizations: a cognitive framework across disciplines," *Cognitive research: principles and implications*, vol. 3, no. 1, pp. 1–25, 2018.

[223] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*. Esri Press, 2010.

[224] E. Hoque, V. Setlur, M. Tory, and I. Dykeman, "Applying pragmatics principles for interaction with visual analytics," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 309–318, 2017.

[225] Jeffrey Shaman, "Here's When We Expect Omicron to Peak," https://www.nytimes.com/2022/01/06/opinion/omicron-covid-us.html, accessed: 2022-04-18.

[226] Amanda Makulec, "@Wattenberger's chart twitter thread," https://twitter.com/abmakulec/status/1479496579040034822, accessed: 2023-07-01.

[227] P. Mantri, H. Subramonyam, A. L. Michal, and C. Xiong, "How do viewers synthesize conflicting information from data visualizations?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1005–1015, 2023.

[228] C. Bicchieri, *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

[229] J. Ma, "Visitors' interpretations of images of the nanoscale," *Nanoscale Informal Science Education Network, July*, 2008.

[230] K. Mukherjee, B. Yin, B. E. Sherman, L. Lessard, and K. B. Schloss, "Context matters: A theory of semantic discriminability for perceptual encoding systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 697–706, 2022.

[231] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang, "Mapping color to meaning in colormap data visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 810–819, 2019.

[232] M. Sbisà, "Speech acts in context," *Language & Communication*, vol. 22, no. 4, pp. 421–436, 2002.

[233] D. Kaplan, "Afterthoughts," 1989.

[234] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K.-L. Ma, and H. Hagen, "Collaborative visualization: Definition, challenges, and research agenda," *Information Visualization*, vol. 10, no. 4, pp. 310–326, 2011.

[235] F. Block, M. S. Horn, B. C. Phillips, J. Diamond, E. M. Evans, and C. Shen, “The deeptree exhibit: Visualizing the tree of life to facilitate informal learning,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2789–2798, 2012.

[236] P. Davis, M. Horn, F. Block, B. Phillips, E. M. Evans, J. Diamond, and C. Shen, ““whoa! we’re going deep in the trees!”: Patterns of collaboration around an interactive information visualization exhibit,” *International Journal of Computer-Supported Collaborative Learning*, vol. 10, no. 1, pp. 53–76, 2015.

[237] WhiteHouse, “America’s Economic Growth in the 21st Century,” https://twitter.com/WhiteHouse/status/1486709480351952901, accessed: 2023-07-01.

[238] ——, “Pelosi Sharing WhiteHouse Chart via Twitter,” https://twitter.com/SpeakerPelosi/status/1486762298752573442, accessed: 2023-07-01.

[239] N. Diakopoulos, *Automating the news*. Harvard University Press, 2019.

[240] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media,” *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[241] S. Wolton, “Are biased media bad for democracy?” *American Journal of Political Science*, vol. 63, no. 3, pp. 548–562, 2019.

[242] W. Chen, K. A. Khatib, H. Wachsmuth, and B. Stein, “Analyzing political bias and unfairness in news articles at different levels of granularity,” *CoRR*, vol. abs/2010.10652, 2020. [Online]. Available: https://arxiv.org/abs/2010.10652

[243] T. Spinde, F. Hamborg, K. Donnay, A. Becerra, and B. Gipp, “Enabling news consumers to view and understand biased news coverage: a study on the perception and visualization of media bias,” in *Proceedings of the ACM/IEEE joint conference on digital libraries*, 2020, pp. 389–392.

[244] F. Hamborg, K. Donnay, and B. Gipp, “Automated identification of media bias in news articles: an interdisciplinary literature review,” *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 391–415, 2019.

[245] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa, “Neural media bias detection using distant supervision with babe-bias annotations by experts,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1166–1177.

[246] C. Hube and B. Fetahu, “Neural based statement classification for biased language,” in *Proceedings of the ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 195–203.

[247] ——, “Detecting biased statements in wikipedia,” in *Companion proceedings of the Web Conference*. New York, NY: ACM, 2018, pp. 1779–1786.

[248] A. Dallmann, F. Lemmerich, D. Zoller, and A. Hotho, “Media bias in german online newspapers,” in *Proceedings of the ACM Conference on Hypertext & Social Media*. New York, NY: ACM, 2015, pp. 133–137.

[249] T. Spinde, C. Kreuter, W. Gaissmaier, F. Hamborg, B. Gipp, and H. Giese, "Do you think it's biased? how to ask for the perception of media bias," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021, pp. 61–69.

[250] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *arXiv preprint arXiv:1309.6202*, vol. abs/1309.6202, 2013.

[251] S. Abbar, S. Amer-Yahia, P. Indyk, and S. Mahabadi, "Real-time recommendation of diverse related articles," in *Proceedings of WWW'13*, 2013, p. 1–12.

[252] S. Park, M. Ko, J. Kim, Y. Liu, and J. Song, "The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns." New York, NY, USA: Association for Computing Machinery, 2011, p. 113–122.

[253] J. M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. P. de Vries, "Rel: An entity linker standing on the shoulders of giants," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, p. 2197–2200.

[254] N. Gillani, A. Yuan, M. Saveski, S. Vosoughi, and D. Roy, "Me, my echo chamber, and i: introspection on social media polarization," in *Proceedings of the ACM World Wide Web Conference*, 2018, pp. 823–831.

[255] U. K. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai, "The effects of subtle misinformation in news headlines." *Journal of experimental psychology: applied*, vol. 20, no. 4, p. 323, 2014.

[256] S. Urologin, "Sentiment analysis, visualization and classification of summarized news articles: a novel approach," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, 2018.

[257] S. H. W. Ilyas, Z. T. Soomro, A. Anwar, H. Shahzad, and U. Yaqub, "Analyzing brexit's impact using sentiment analysis and topic modeling on twitter discussion," in *The 21st Annual International Conference on Digital Government Research*, 2020, p. 1–6.

[258] Y. Zhang, Y. Sun, L. Padilla, S. Barua, E. Bertini, and A. G. Parker, "Mapping the landscape of covid-19 crisis visualizations," in *Proceedings of the Conference on Human Factors in Computing Systems*, 2021.

[259] H.-K. Kong, Z. Liu, and K. Karahalios, "Frames and slants in titles of visualizations on controversial topics," in *Proceedings of the Conference on Human Factors in Computing Systems*, 2018, p. 1–12.

[260] F. Hamborg, K. Heinser, A. Zhukova, K. Donnay, and B. Gipp, "Newsalyze: Effective communication of person-targeting biases in news articles," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021, pp. 130–139.

[261] P. Saleiro, S. Amir, M. Silva, and C. Soares, "Popmine: Tracking political opinion on the web," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. IEEE, 2015, pp. 1521–1526.

[262] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Proceedings of Symposium on Visual Analytics Science and Technology*, 2010, pp. 115–122.

[263] B. B. Hardy, "Jittery gauges: Combating the polarizing effect of political data visualizations through uncertainty," Ph.D. dissertation, Brigham Young University, 2017.

[264] J. Westfall, L. Van Boven, J. R. Chambers, and C. M. Judd, "Perceiving political polarization in the united states: Party identity strength and attitude extremity exacerbate the perceived partisan divide," *Perspectives on Psychological Science*, vol. 10, no. 2, pp. 145–158, 2015.

[265] H. Elhamdadi, A. Gaba, Y.-S. Kim, and C. Xiong, "How do we measure trust in visual data communication?" in *2022 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, 2022, pp. 85–92.

[266] R. Koonchanok, P. Baser, A. Sikharam, N. K. Raveendranath, and K. Reda, "Data prophecy: Exploring the effects of belief elicitation in visual analytics," 2021.

[267] T. Spinde, F. Hamborg, K. Donnay, A. Becerra, and B. Gipp, "Enabling news consumers to view and understand biased news coverage: A study on the perception and visualization of media bias," ser. JCDL '20. ACM, 2020, p. 389–392.

[268] I. Alieva, "How american media framed 2016 presidential election using data visualization: The case study of the new york times and the washington post," *Journalism Practice*, vol. 17, no. 4, pp. 814–840, 2023.

[269] Y.-S. Chen, L.-H. Chen, and Y. Takama, "Proposal of lda-based sentiment visualization of hotel reviews," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 687–693.

[270] J. M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. P. de Vries, "Rel: An entity linker standing on the shoulders of giants," in *Proceedings of the International Conference on Research and Development in Information Retrieval*, 2020.

[271] F. Hamborg and K. Donnay, "Newsmtsc: (multi-)target-dependent sentiment classification in news articles," in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Apr. 2021.