9-1-2017 10:30 AM

# The Design of Interactive Visualizations and Analytics for Public Health Data

Oluwakemi Ola
*The University of Western Ontario*

Supervisor
Dr. Kamran Sedig
*The University of Western Ontario*

Graduate Program in Computer Science
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

# Abstract

Public health data plays a critical role in ensuring the health of the populace. Professionals use data as they engage in efforts to improve and protect the health of communities. For the public, data influences their ability to make health-related decisions. Health literacy, which is the ability of an individual to access, understand, and apply health data, is a key determinant of health. At present, people seeking to use public health data are confronted with a myriad of challenges some of which relate to the nature and structure of the data. Interactive visualizations are a category of computational tools that can support individuals as they seek to use public health data. With interactive visualizations, individuals can access underlying data, change how data is represented, manipulate various visual elements, and in certain tools control and perform analytic tasks. That being said, currently, in public health, simple visualizations, which fail to effectively support the exploration of large sets of data, are predominantly used. The goal of this dissertation is to demonstrate the benefit of sophisticated interactive visualizations and analytics. As improperly designed visualizations can negatively impact users' discourse with data, there is a need for frameworks to help designers think systematically about design issues. Furthermore, there is a need to demonstrate how such frameworks can be utilized. This dissertation includes a process by which designers can create health visualizations. Using this process, five novel visualizations were designed to facilitate making sense of public health data. Three studies were conducted with the visualizations. The first study explores how computational models can be used to make sense of the discourse of health on a social media platform. The second study investigates the use of instructional materials to improve visualization literacy. Visualization literacy is important because even when visualizations are designed properly, there still exists a gap between how a tool works and users' perceptions of how the tool should work. The last study examines the efficacy of visualizations to improve health literacy. Overall then, this dissertation provides designers with a deeper understanding of how to systematically design health visualizations.

## Keywords

Visualization; Health-Related Tasks; Human-Data Interaction; Sensemaking; Visualization Literacy; Health Literacy; Interaction; Visual Representation; Analytics; Public Health Informatics; Visual Analytics; Human-Computer Interaction

# Co-Authorship Statement

**Chapter 1** is my own original work in introducing the dissertation and explaining connections between chapters. **Chapters 2, 3, 4, 6, and 7** were a collaborative effort with my supervisor, Kamran Sedig. **Chapter 5**, which focuses on the design of a visualization for vector-borne diseases, was a collaborative effort with my supervisor, but it also included the help of a colleague, Olha Buchel, who worked on the implementation of the tool and assisted in the conceptualization of the case study. **Chapter 8** was written by me, to summarize the dissertation and outline future areas of research.

With the exception of the tool described in Chapter 5, I was primarily responsible for the conceptualization, design, and implementation of the visualizations presented in this dissertation. The data for the visual analytic study presented in Chapter 6 was collected by another graduate student working under the supervision of Dr. Sedig. For this chapter, I was responsible for the design of the study, the analysis of the data, and the implementation of the computational models. For the studies presented in Chapter 7, I designed them and collected and analyzed the data.

# Acknowledgments

I would like to thank my supervisor, Dr. K. for having faith in my ability to learn. For your patience, guidance, and mentorship, I am truly grateful.

I would also like to thank the past and present members of INSIGHT lab. The crew: Doc, Robert, Didandeh, Olha, "The laughters", Demelody, Viet, and Anthony. Thank you all for the brainstorming sessions, reviews, comments, and suggestions. I would also like to thank the ladies in the office, Dianne, Janice, Cheryl, and Laura (by extension). Thanks for always answering my questions and being a constant support. My appreciation goes out to Jeff and Art, the systems crew, for supplying the computational support for me to run the studies.

I would also like to thank my Canadian family. Mommy Gwen, Joe, Nellyfield, Joyetti, Chadleyy, J, JJ, Isabelle, Mateo, Trevor, Vanessa, Sister Hortensiii, and Brother Neville. Thank you for your love, food, support, rides, prayers, jokes, and advice. Shout out to my Indy family as well, Pastor T., Aunty Dr. Folu, ToluwaniTiny, and DamilareDaviddd. I couldn't have done this without your support. I love you all so very much. The skills I learned in Indy on how to persevere have been invaluable during this process. My North London church family, thank you for being present. To all my friends over the years and members of my extended family who have played a role, thank you very much.

I am who am I because of my parents who have travelled this road before me. My father, my father, the only father I have, I love you Paps, thanks for loving and believing in your daughters. My mother, who is an angel, thanks for your consistency in praying me through this work and your endurance at Andrews. To my sisters, Oluwatostos, Oluwabusbus, and Oluwajojo, thanks for living life without fear and encouraging me to do the same.

Nyasha Mudzengi, the love of my life, thanks for always being sunshine and a reflection of God's love.

Finally, I would like to give all honor and glory to God for life, love, grace, and peace.

# Table of Contents

# List of Tables

# List of Figures

# List of Appendices

# Chapter 1

# 1 Introduction

## 1.1 Motivation

Public health data plays a critical role in ensuring the health of the populace. For health professionals, data influences every aspect of their mandate (O'Carroll, 2003). From an assessment standpoint, data is needed to investigate and analyze the causation of health issues. From a policy development perspective, data plays the crucial role of helping professionals prioritize and determine which issues need to be addressed. From an assurance viewpoint, timely data is required for the management of resources and for educating the public. For the general public, health data influences their ability to make sound decisions. An individual's ability to access, read, and understand health information is a public health imperative (Gazmararian, Curran, Parker, Bernhardt, & DeBuono, 2005; Kickbusch, Pelikan, Apfel, & Tsouros, 2013; Sørensen et al., 2012). This ability has been termed "health literacy" and is a key determinant of an individual's health. According to the American Medical Association, health literacy is a stronger predictor of a person's health than age, income, employment status, education level, and race ("Health literacy: report of the Council on Scientific Affairs. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association.," 1999).

While the role of public health data is indisputable, laypeople and professionals seeking to use the data are confronted with a myriad of challenges some of which relate to the nature and structure of the data. Public health data originates from a wide variety of sources, is encoded in different formats, and is aggregated at different levels (Gotz & Borland, 2016; Herland, Khoshgoftaar, & Wald, 2014; E. Liu, Zhao, Wei, Roumeliotis, & Kaldoudi, 2016; Ola & Sedig, 2014; Shneiderman, Plaisant, & Hesse, 2013). The public health informatics community recognizes the need for data to be presented in ways in which individuals can work with it effectively (Higgins et al., 2011; Keough, 2002; LaPelle, Luckmann, Simpson, & Martin, 2006; Turner, Stavri, Revere, & Altamore,

2008). For centuries, visualizations have been used to facilitate public health tasks involving data. For instance, in 1792, Finke produced a world map of diseases and six years later Seaman used spot maps to trace yellow fever cases in New York (Barrett, 2000; Stevenson, 1965). In the mid-19th century, John Snow challenged the theory of cholera being an airborne disease by plotting the spread of the disease as it relates to the Broad Street water pump (Snow, 1855). Around the same time, Florence Nightingale used the coxcomb representation to visualize patient data and educate the Crown on sanitation-related deaths of soldiers during the Crimean War (B. Cohen, 1984). From outbreak detection to health promotion, these examples highlight the varied use of static visualizations.

However, the rate at which data is currently being generated has reduced the effectiveness of past visualization approaches to support the tasks in which individuals engage (Cybulski, Keller, Nguyen, & Saundage, 2013; L. Zhang et al., 2012). On the one hand, simple visualizations (e.g., bar charts, line plots), which only encode one or two attributes of data items, limit the ability of users to analyze non-explicit and unknown relationships (Cybulski et al., 2013; Endert, Hossain, et al., 2014). There is a need for visualizations that encode multiple aspects of the data at the same time. On the other hand, static visualizations, which require that all data items be encoded at once can overwhelm the cognitive resources of individuals (Kirsh, 2013; Pike, Stasko, Chang, & O′Connell, 2009; Tominski, 2015). Making a visualization interactive, facilitates the gradual disclosure of data and allows users to control how data is shown and in what quantities. When dealing with large sets of data, which is typically the case in public health, interaction has been shown to be effective in aiding analysts to explore and understand large, multivariate datasets (Torres, Eicher-Miller, Boushey, Ebert, & Maciejewski, 2012).

In certain situations, providing users with elaborate and interactive visualizations is still not sufficient to support their tasks. For example, an epidemiologist may need to perform statistical analysis to understand the spread of Chikungunya across the Caribbean. Allowing the epidemiologist to visualize summary statistics is beneficial as research notes that the complexity of the mathematical models is hard for users to understand

without aid (Robert Spence, 2007). This coupling of analytics with visualization is the focus of the nascent field of visual analytics. Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces (Thomas & Cook, 2005). Visual analytics tools are comprised of the analytics engine which stores, transforms, and performs computational analysis of the data and interactive visualizations which encode the data in a visual format that the user can then work with (Ola & Sedig, 2014).

Before we can effectively design visual analytics tools for public health, there is a need to have a deeper understanding of the individual components and how best to develop them. Researchers have called for the development of visualization tools that allow users to access underlying data, change how data is represented, identify patterns and trends, analyze data, and perform a wide variety of tasks (Bhowmick, Griffin, MacEachren, Kluhsman, & Lengerich, 2008; Cybulski et al., 2013; Endert, Ribarsky, Turkay, Wong, & Nabney, 2017; Fisher, DeLine, Czerwinski, & Drucker, 2012; Gotz & Borland, 2016; Katsis, Koulouris, Papakonstantinou, & Patrick, 2017; Pretorius, Khan, & Errington, 2016; L. Zhang et al., 2012). The design of such visualizations is a non-trivial endeavor that requires designers to take into consideration the structure of the data, users' tasks, and human factors. Part of the challenge of developing visualizations is determining how to organize and encode data items and how best to support users' tasks. Currently, there is confusion and lack of direction over how to create effective health visualizations (Carroll et al., 2014; Folorunso & Ogunseye, 2008; Turner et al., 2008).

As improperly designed visualizations can end up negatively impacting users' discourse with data (Kirsh, 2009) there is a need for frameworks to help designers think systematically about design issues (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008; Sedig, Parsons, Dittmer, & Haworth, 2013; Thomas & Cook, 2005). Furthermore, there is a need to demonstrate how such frameworks can be utilized. Even when visualizations are designed properly, there still exists a gap between how the tool was designed and users' perceptions of how the tool should work (Norman, 2013). This is particularly the case with novel interactive visualizations. Borner et al. highlight the need for instruction so that individuals are better equipped to understand novel visualizations (Borner, Maltese, Balliet, & Heimlich, 2016). The goal of this research is to demonstrate how

visualizations and analytics can be designed for public health, explore how instructional materials can help individuals to learn to use visualizations, and demonstrate how visualizations can impact health literacy efforts.

## 1.2 Structure of this dissertation

The rest of this dissertation is broken into 7 chapters as follows:

In **Chapter 2**, we present the data-based challenges in public health and make a case for the use of visual analytics tools. This chapter also provides background for the dissertation and briefly discusses the field of public health and the components of visual analytics tools.

In **Chapter 3**, we describe the use of visualizations in public health and discuss the need for visualizations that effectively model the complexity of the data. In this chapter, we also describe a framework that can aid in the design of elaborate visualizations, and apply the framework to the design of four novel visualizations that are part of a tool for making sense of public health data.

In **Chapter 4**, we discuss interaction and its role in improving the discourse between users and public health data. We also present a process for designing interaction that is based on users' task and use three scenarios to highlight the efficacy of our approach.

In **Chapter 5**, we demonstrate how interactive visualizations can support public health stakeholders' decision-making tasks. In particular, we present a visualization tool we created that can support control efforts related to the recent Zika outbreak in Brazil. This chapter also demonstrates how complex statistical measures can be incorporated into visualization tools.

In **Chapter 6**, we present a visual analytic study that explores the discourse of health issues on Twitter. We describe how computational models can be used to assess the theme of tweets, as well as determine the type of user sending the tweet.

In **Chapter 7**, we present the results of two studies we conducted with the visualizations we designed. The first study explores visualization literacy and how individuals learn to use unfamiliar and non-typical visualizations. We also investigate the effect of instructional materials on improving an individual's ability to learn how data is encoded and how to interact with the visualization. In the second study, we examine the ability of visualizations to improve health literacy.

In **Chapter 8**, we draw some conclusions from the research reported in the preceding chapters, discuss the contributions of this research to the wider scientific community, and highlight some areas of future research.

Finally, readers should keep in mind that the chapters of this dissertation can be read sequentially or individually. Chapters 2, 3, and 5 have been published; chapter 7 has been accepted; and chapters 4 and 6 will soon be submitted. The dissertation is written in an integrated article format, so chapters 2 through 7 are self-contained.

# Chapter 2

# 2 The Challenge of Big Data in Public Health: An Opportunity for Visual Analytics

This chapter has been published as O. Ola and K. Sedig, "The challenge of big data in public health: an opportunity for visual analytics.," Online J. Public Health Inform., vol. 5, no. 3, pp. 1-21, Jan. 2014.

Please note that the format has been changed to match the format of the dissertation. Figure numbers mentioned herein are relative to the chapter number. For instance, "Figure 1" corresponds to Figure 2-1. Additionally, when the term "paper" or "article" is used, it refers to this particular chapter.

## 2.1 Introduction

Data and information are both currency and product within the field of public health (PH) (O'Carroll, 2003). PH data is often highly complex because of its high volume, its various sources, its velocity of generation, and sometimes the low degree of veracity of the sources from which it originates. PH data is gathered from heterogeneous sources (Revere et al., 2007), may be unreliable, encoded in a variety of formats (Rambo, 2000; Turner, Liddy, Bradley, & Wheatley, 2005), and can be volatile (i.e., changing, and available only for a limited amount of time) (O'Carroll, Cahn, Auston, & Selden, 1998), all characteristics attributed to big data. These characteristics of PH data pose a challenge to the PH workforce in terms of whether and how effectively the data is used.

The PH workforce is comprised of people trained in a variety of disciplines with daily duties necessitating the extraction of information and construction of knowledge from the mass of available data. In this paper, we refer to any individual seeking to use PH data in a professional capacity as a stakeholder. As stakeholders interact with data, they engage in various cognitive activities such as analytical reasoning, interpreting, decision-making, planning, and problem solving (Sedig, Parsons, Dittmer, & Ola, 2012). Performing these activities with data can involve complex cognition and can pose cognitive challenges for

the unaided mind. Thus, computer-based information systems and tools may be needed to support the activities in which PH stakeholders engage.

In the context of PH, access to data does not necessarily guarantee that the data will be used well—i.e., that cognitive activities will be performed in an effective manner (see (Sedig, Parsons, Dittmer, et al., 2012) for more discussion of this issue). Additionally, the PH community acknowledges that decisions and policies are often made in an ad hoc fashion devoid of evidence (Baltussen & Niessen, 2006; Brownson, Fielding, & Maylahn, 2009; Brownson, Gurney, & Land, 1999). The efficient and effective use of data determines the extent to which PH stakeholders can sufficiently address the health concerns of the community (O'Carroll, 2003; Reeder, Revere, Hills, Baseman, & Lober, 2012). Consider the following scenario in the fictional town of Lumcard, Louisiana, which demonstrates the critical role of data in addressing public health issues.

The day before Thanksgiving, the director of Lumcard's health department receives an alert showing an unusually high incidence of complaints of diarrhea, vomiting, high fever, and sore throat. Discussions with area doctors reveal that local hospitals have confirmed the diagnosis of West Nile Virus (WNV) in a high number of patients; this helps the director dismiss his first assumption that a food poisoning outbreak exists in Lumcard. The regional epidemiologist is made aware of the situation and immediately begins to investigate this unseasonable occurrence. The epidemiologist not only needs to be able to access data, but also must compare health records, filter out irrelevant data, examine environmental influences, and identify relationships among various factors. In addition, she will need to develop, test, and discard hypotheses about the cause of WNV and collaborate with other PH stakeholders in order to determine how best to ensure the health of the citizens of Lumcard. While having access to data is critical, the Lumcard PH team's success in addressing the potential health hazard is largely dependent on their ability to effectively use the available data in their reasoning, sensemaking, decision-making, and planning activities.

Under different time constraints, PH stakeholders must perform a myriad of activities, which ultimately have health, social, political, economic, and ethical implications for the

community (Goddard et al., 2004). Furthermore, as stakeholders interact with data they encounter a number of obstacles relating to its volume, variety, velocity, and veracity (B. B. Cohen, Franklin, & West, 2006; Higgins et al., 2011; Keough, 2002; Kiefer et al., 2005; LaPelle et al., 2006; O'Carroll et al., 1998; Rambo, 1998, 2000; Revere et al., 2007; Turner et al., 2008, 2005). Over the course of the last 20 years, computational tools and systems have been developed to support the work activities of PH stakeholders. Current tools include data analytics tools such as Stata (StataCorp, 2009), and interactive visualization tools such as Malaria Atlas Project (Guerra et al., 2007). While these types of tools are beneficial in addressing certain work activities of PH stakeholders, they fall short in supporting cognitive activities that involve the use and working with large, heterogeneous, and complex bodies of data (Keim, Mansmann, & Thomas, 2009).

Public health tends to lag behind other sectors in the adoption of new technology (for examples, see England et al.'s (England, Stewart, & Walker, 2000) examination of PH's slow rate of information technology adoption, and Shortliffe's (Shortliffe, 2005) comparison of healthcare with other sectors). The recent emergence of a category of computational tools known as visual analytics (VA) tools is no exception. These tools are intended to alleviate some of the shortcomings of the aforementioned tools with regard to the complexity of data and support of the visuo-analytical reasoning of their human users[1]. VA tools combine interactive visual representations with advanced analytics techniques to synthesize, analyze, and facilitate visuo-analytical reasoning and other high-level cognitive activities involving data (Keim, Kohlhammer, Ellis, & Mansmann, 2010; Thomas & Cook, 2005). This is beneficial for data-intensive fields (Keim et al., 2010) such as PH, finance, insurance, sales, and climatology, to name a few. While many fields, such as finance and sales (Schlegel, Sallam, Daniel, & Tapadinhas, 2013), have seen widespread adoption of VA tools, PH has not. In this paper, we discuss the role that VA tools can play in assisting PH stakeholders to perform cognitive activities involving big data. We focus on analytical reasoning as an activity that plays an important role in many other activities. Through a synthesis of research across multiple

[1] In this paper, the terms user means "human user" and is used interchangeably with the term stakeholder.

fields including cognitive science, data mining, human-computer interaction, and informatics, we explicate the benefits of VA tools in addressing the challenge that big data poses to PH practice.

The rest of this paper is organized as follows. Section 2 discusses foundational concepts—i.e., PH data and information, analytical reasoning, visual representations, and human-information interaction. Section 3 describes VA tools, their components, and how they facilitate analytical reasoning. Section 4 discusses the benefits and role of VA tools in PH and highlights current tools in use. Through the use of a hypothetical scenario, Section 5 further explicates the usefulness of VA tools. Finally, Section 6 provides a summary and briefly outlines limitations and some future areas of investigation.

## 2.2 Background

This section presents necessary background concepts and terminology used in this paper. In order to address the health concerns of the community, PH stakeholders interact with data to perform a variety of work activities. We depict the needs that VA tools must address for PH stakeholders, by describing the data they interact with, the nature of their work activities, and the analytical reasoning tasks in which they engage. Furthermore, a VA tool's interface influences the stakeholder's ability to access data and perform visuo-analytical reasoning. Therefore, we explain two major components of the interface — namely: visual representations and interactions.

### 2.2.1 PH Stakeholders

The workforce charged with safeguarding and improving the health of the community through a population focus, characterized in this paper as PH stakeholders, is highly varied. As discussed by O'Carroll et al. (O'Carroll et al., 1998), the PH workforce may be more diverse than any other group of health professionals. PH stakeholders come from a diverse set of backgrounds and are trained in a myriad of disciplines (Committee on Educating Public Health Professionals for the 21st Century, 2003). Irrespective of their area of expertise and sub-field of application, stakeholders must interact with data to perform a myriad of work activities.

### 2.2.2 PH Data and Information

To frame our discussion, we characterize data as digitally stored, sensed changes in the environment, and information as processed, organized, and/or analyzed data that depicts its relationships[2]. PH data can be described by its high volume (Higgins et al., 2011; Keough, 2002; LaPelle et al., 2006; Turner et al., 2008), great variety (B. B. Cohen et al., 2006; Rambo, 1998; Revere et al., 2007), high velocity (O'Carroll et al., 1998), and low veracity (Kiefer et al., 2005; LaPelle et al., 2006; Reeder et al., 2012). These four features of PH data are typical characteristics of big data. As a result, PH data is big data. While synthesis of and access to PH data has been a focus of the PH informatics literature (Sedig, Parsons, Dittmer, et al., 2012), the use of data by stakeholders to create information, particularly as mediated by computational tools, presents a growing challenge. Computationally-mediated reasoning requires not only the ability to access relevant data, but the ability to control how data is structured, combined, displayed, and interacted with (Sedig & Parsons, 2013). In addition, stakeholders must be presented with representations that accurately communicate what is known or unknown, the impact of actions, relationships that exist, and extent of uncertainty and risk that are involved during analysis (Berner & Moss, 2005; Keough, 2002). The seamless incorporation of user-guided analysis techniques into computational tools is crucial in facilitating the systematic use of data.

### 2.2.3 PH Activities

PH stakeholders engage in a variety of work activities in an effort to improve and ensure the health of the community (Committee on Educating Public Health Professionals for the 21st Century, 2003; O'Carroll et al., 1998; Rambo, 1998). These activities vary by work group (e.g., epidemiologist or nutritionist), by level within a work group (e.g., state, local, federal), and by function (Rambo et al., 2001). In the United States, these work activities have been grouped by the Institute of Medicine (IOM) into three core functions—namely: 1) *Assessment*, which includes investigating and analyzing the

---

[2] For an in-depth discussion on the differences between data and information see (Sedig, Parsons, & Babanski, 2012) and (Bates, 2005).

occurrence and causation of health problems and hazards; 2) *Policy Development*, which includes priority setting, advocacy, and development of policies; and 3) *Assurance*, which includes managing resources and informing and educating the public about health issues and services (National Research Council, 1988). In this paper, we use the IOM core functions classification to group PH work activities. Regardless of the core function with which the PH stakeholder is tasked, PH work activities are a form of knowledge work (Sedig, Parsons, Dittmer, et al., 2012). In other words, at a basic level, PH stakeholders are knowledge workers—that is, most of their work is performing information-dependent cognitive activities. Knowledge work activities are non-routine and require a combination of convergent, divergent, and creative thinking in order to be completed (Reinhardt, Schmidt, Sloep, & Drachsler, 2011). As knowledge workers, PH stakeholders engage in a myriad of cognitive activities including analytical reasoning, decision-making, sensemaking, and problem solving.

### 2.2.4 Analytical Reasoning

While a comprehensive discussion of high-level cognitive activities is beyond the scope of this paper, to fully appreciate the utility of VA tools, we examine PH stakeholders' cognitive processes as they work with data. To this end, we focus on analytical reasoning and discuss some of its characteristics, explain how it facilitates other high-level cognitive activities, and briefly highlight its impact on PH work activities.

Analytical reasoning is based on a rational, logical analysis and evaluation of data and information and encompasses different kinds of reasoning such as inductive, deductive, and analogical reasoning (Sedig & Parsons, 2013). An inference or conclusion is reached based on the systematic analysis of data. As an activity, analytical reasoning emerges from the completion of lower-level tasks. Some of the tasks include, but are not limited to, identifying relationships among pieces of data, asserting and testing key assumptions, testing biases, assessing alternatives, developing hypotheses, and supporting conclusions with adequate evidence (Heuer, 1999; Thomas & Cook, 2005). Although analytical reasoning is a structured and disciplined process, the aforementioned tasks typically occur in an iterative and non-linear fashion (Sedig & Parsons, 2013). In other words, the

order in which low-level tasks occur is not fixed, but varies according to the cognitive needs and overall goals of the stakeholder.

Analytical reasoning seldom occurs in a vacuum, but instead may occur concurrently with other cognitive activities. In particular, analytical reasoning facilitates problem solving and decision-making (Green & Maciejewski, 2013; Leighton, 2004). Analytical reasoning can be viewed as a transformative process in which new information, knowledge, and insight are derived from given data (Gilhooly, 2004; Sedig & Parsons, 2013). In some situations, this new information, knowledge, or insight serves as the basis for decision-making and problem solving (Leighton, 2004). To illustrate the interconnectedness of analytical reasoning, decision-making, and problem solving, consider further the situation in Lumcard: the epidemiologist, engaged in analytical reasoning, concludes that there is a direct correlation between temperature and incidences of WNV in the city. In addition, from her analysis, she is able to narrow down the list of possible mosquito breeding sites to two local bodies of water. Subsequently, the epidemiologist and health director make the decision to restrict access of the residents to local bodies of water, and also send out an environmental health scientist to collect samples to determine the mosquito infestation levels at the shortlisted locations.

Due to the complex, dynamic, and interdependent nature of public health issues, a faulty decision or policy can have a negative impact that may not be immediately recognizable. Analytical reasoning provides the basis for decisions, plans, and policies and should, therefore, not be overlooked. While the PH community recognizes that information should be used to inform policy-making and program development (Kiefer et al., 2005; Mowat & Hockin, 2002), the reality is that decisions and policies are often made in an ad hoc fashion, mostly based on gut feelings, short-term goals, and/or information satisficing (Baltussen & Niessen, 2006; Brownson et al., 1999; National Research Council, 1988). For this reason, there has been a push to move stakeholders closer to adopting evidence-based approaches in PH practice. This approach advocates the systematic use of information and application of scientific reasoning principles in a contextualized manner while making decisions and creating policies (Brownson et al., 2009; Kiefer et al., 2005).

The success of this approach is contingent on PH stakeholders being able to effectively interact with and use data (Keough, 2002).

### 2.2.5 Visual Representations

When reasoning is mediated by VA tools, data is made accessible to the user of the tool through external visualizations—i.e., visual representations. Therefore, it is necessary to discuss the benefits of visual representations and their effect on stakeholders' activities. Visual representations encode data items using visual marks (e.g., lines, dots, shapes) and combine and integrate these into more complex structural forms (e.g., scatter plots, heat maps, bar charts) (Sedig, Parsons, Dittmer, et al., 2012). These representations seek to capitalize on the human visuoperceptual system, which is specifically suited to rapid processing of data and recognition of visual patterns. The benefits of such representations have been discussed by researchers including Larkin and Simon (Larkin & Simon, 1987), Glenberg and Langston (Glenberg & Langston, 1992), and Card et al. (Card, Mackinlay, & Shneiderman, 1999). According to Card et al., visual representations can amplify cognition by increasing the memory and processing resources available to users, reducing the search for information, enhancing the detection of patterns, enabling perceptual inference operations, and encoding information in a *manipulable* medium (Card et al., 1999). The manipulability of a medium is an important factor. While static representations have been historically used by PH stakeholders, from John Snow's use of a map to reason about a cholera outbreak in 1850 (Snow, 1855), to the recent use of atlases for mapping the risk of malaria in Africa (Le Sueur et al., 1997), they put the brunt of the information-processing load (i.e., analytical reasoning and decision-making) on the cognitive resources of users (Sedig & Parsons, 2013; Sedig, Parsons, Dittmer, et al., 2012), hence negatively affecting their usability.

Computers, on the other hand, allow visual representations to be interactive and dynamically manipulable. This allows information processing to be shared between the user and the tool (Sedig & Parsons, 2013), reducing, and possibly bridging, the gap between the internal (mental) representations of the user and the external (visual) representations of the tool (Parsons & Sedig, 2013a; Sedig & Parsons, 2013; Sedig, Parsons, & Babanski, 2012). Interactive visual representations can offer users flexibility,

support convergent and divergent thinking, and accommodate the users' perceptual and cognitive needs (Sedig & Parsons, 2013; Thomas & Cook, 2005). Furthermore, interactive representations allow stakeholders to control which subset of data is visually displayed while still having access to data latent in the system (Sedig, 2009; Sedig & Parsons, 2013). This is important for fields like PH where large amounts of data cannot be visualized all at once. In addition, interactive visual representations allow stakeholders to choose how things are represented (Sedig, 2009; Sedig & Parsons, 2013), which has an effect on the reasoning tasks in which stakeholders engage. Researchers in cognitive science have demonstrated that different representational forms can impact how cognitive activities are performed (Larkin & Simon, 1987; J. Zhang & Norman, 1994), and even constrain and limit stakeholders as they engage in a particular task (Parsons & Sedig, 2013a; J. Zhang, 2001; J. Zhang & Norman, 1994). Therefore, PH stakeholders stand to benefit from tools that allow users to manipulate visual representations, a capability made possible through interaction.

### 2.2.6 Human-Information Interaction

Through interaction, the user of a VA tool is able to control, not only the form or content of the visual representation, but also the entire dialogue with information (Parsons & Sedig, 2013a; Sedig & Parsons, 2013). Interaction moderates the discourse between information and the user and can be conceptualized at different levels. In this paper, we describe interaction in terms of the actions the user performs on the interface of the tool, the consequent changes and reactions in the visual representations, and the user's perceptions of changes to the representations (Sedig & Parsons, 2013). In the context of VA tools, by performing actions on the visual representations, the user is able to reach into the database and operate upon data. Examples of such actions include filtering, annotating, drilling, selecting, and comparing (Sedig & Parsons, 2013). In response, the reactions visible through changes in the visual representations (i.e., on the interface) ensure that the discourse is not one-sided. Equally important are the reactions that are not visually perceptible that occur within the VA tool (Sedig, Parsons, & Babanski, 2012). The user's perceptions of changes to visual representations complete the interaction loop. Together, actions, reactions, and perceptions promote the back-and-forth dialogue

between the user and the represented information. The sequence in which actions are performed is sometimes at the discretion of the user. This is beneficial in fields such as PH where software designers are not privy to how various subsets of data will be used in analysis by the stakeholder. The user-guided sequencing of actions and discourse with information is critical in VA tools that function to facilitate PH stakeholders' analytical reasoning tasks.

## 2.3 Visual Analytic Tools

VA is sometimes defined as the *"science of analytical reasoning facilitated by interactive visual interfaces"* (Thomas & Cook, 2005). VA tools combine data analytics and interactive visualizations to support users' reasoning, and create an environment in which users engage in a more involved discourse with data and information (Keim et al., 2010; Thomas & Cook, 2005). Prior to the development of VA, various groups of computational tools sought to address the information-based needs of professionals. In PH, two groups are data analytics and interactive visualization tools. This section highlights the limitations of these two groups of tools, describes the components of VA tools, and explains how analytical reasoning can be performed using VA tools.

Data analysis or analytics tools incorporate techniques and algorithms from a variety of fields including statistics (e.g., mean and correlation), data mining (e.g., classification and clustering), and machine learning (e.g., artificial neural network and support vector machines) to facilitate the discovery and understanding of patterns in data (Han, Kamber, & Pei, 2011). Current data analytics tools that assist PH stakeholders in analyzing data include Stata (StataCorp, 2009) and EpiInfo (Centers for Disease Control and Prevention, 2012). While the aforementioned standalone data analytics tools are capable of processing massive amounts of data, they neither deal with noisy and highly heterogeneous data efficiently, nor are capable of handling ill-defined problems that require human judgment (Keim et al., 2009). Because these tools take over the analysis process and mostly hide the intermediary steps, stakeholders can only be minimally in control of or involved in the analytical reasoning process.

Complementing data analytics tools, interactive visualization tools represent data in a visual form, allow users to control the flow of data, and let them customize representations to cater to their cognitive and contextual needs. Some interactive visualization tools focus on visualizing abstract, nonphysical data such as text and statistical data (Card et al., 1999), while others portray physical data such as the human body and molecules (Mackinlay, 2000). Current PH interactive visualization tools include Malaria Atlas Project (Guerra et al., 2007) and Spatio-Temporal Epidemiological Modeller (Ford, Kaufman, & Eiron, 2006). While beneficial, these types of tools prove inadequate when faced with problems requiring advanced computational analysis and big data (Keim et al., 2009).

### 2.3.1 Components of VA Tools

While data analytics tools with advanced automated analysis and interactive visualization tools aided by human judgment are advantageous in certain situations, their respective limitations create a void, and it is only through VA tools that some of today's most pressing data analysis problems can be addressed (Keim et al., 2009). VA tools fuse the strengths of both sets of tools to create an environment in which the user engages in a more involved discourse with data. This process is not simply an internal automated analysis with an external visual representation displayed at its completion. Instead, it is an integrated human-information dialogue in which data processing is distributed between the user and the main components of the tool—described in this paper as the analytics engine and interactive visualization engine (Sedig, Parsons, & Babanski, 2012), which are described below.

### 2.3.2 Analytics Engine

Human cognition displays several limitations when confronted with mental tasks that are data-intensive (i.e., they involve the use of bodies of data that are too large or too complex), and as a result computational tools can be used to support such tasks. The analytics engine in VA tools is intended for this purpose. It stores, transforms, and performs computational analysis on data. This process, as shown in Figure 1, is subdivided into three main stages: 1) data pre-processing, 2) data transformation, and 3)

data analysis. In the pre-processing stage, data retrieved from a variety of sources is automatically processed. Common tasks in this stage include data cleaning, integration, fusion, and synthesis (Han et al., 2011). In the data transformation stage, the pre-processed data is converted into a form that is more conducive to data analysis. This stage includes tasks such as data normalization and aggregation (Han et al., 2011).

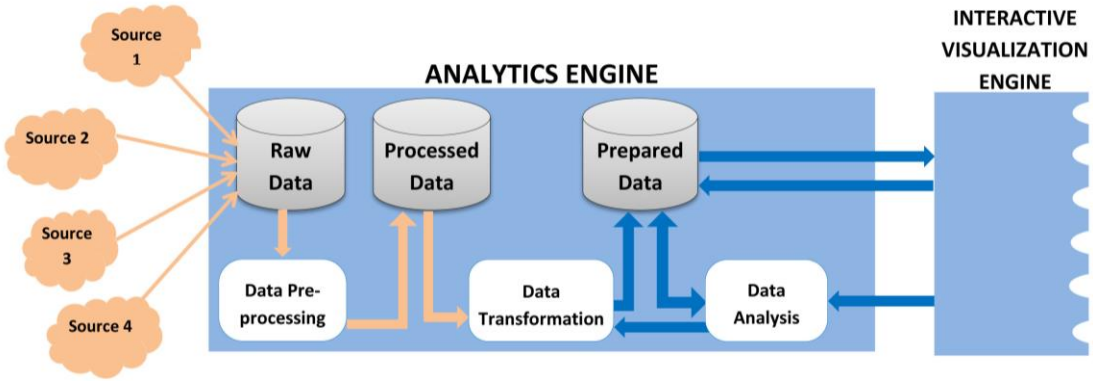



**Figure 2-1: The analytics engine component of VA tools**

Finally, the data analysis stage involves the discovery of patterns and allows for the extraction of valuable information. While historically computational tools have focused on the analysis of one form of data, VA tools overcome this limitation and can analyze and discover patterns in multiple forms of data (e.g., text, video, geo-spatial, etc.) together in order to create information. This is done by drawing on the tasks and techniques that originate from a myriad of fields including statistics (e.g., standard deviation, correlation analysis), machine learning (e.g., classification, clustering, dimension reduction), textual analysis (e.g., document summarization, concept extraction), image analysis (e.g., image segmentation, object recognition), video analysis (e.g., motion detection), and geo-spatial analysis (e.g., surface analysis, locational analysis) (Alpaydin, 2009; Fairclough, 2003; Smith, Goodchild, & Longley, 2006; Soille, 2003; Weisi et al., 2011). In some VA tools, computational analysis is not a system-controlled process but a user-controlled one. The blue arrows in Figure 1 are indicative of the extent of the user's involvement in the analysis process. This process is a sophisticated discourse that goes beyond simplistic interaction to deep user-guided

analysis of data. The interactive visualization engine allows the user to access and control the flow and analysis of data.

### 2.3.3 Interactive Visualization Engine

In VA tools, the interactive visualization engine is composed of the rendering and mapping component that takes analyzed data and creates interactive visual representations (i.e., information). Interactive visual representations allow the user to access, restructure, analyze, and modify amount and form of displayed information (Keim et al., 2010; Thomas & Cook, 2005). The user's actions can impact the discourse in many ways, three of which are shown in Figure 2. Firstly, as shown by blue arrow 1, the user can change how the visualized information is encoded, as, for instance, by replacing a pie chart with a bar graph. Secondly, as depicted by blue arrow 2, the user can change the subset of information displayed. Thirdly, as depicted by blue arrow 3, the user has the ability to guide the analysis process by selecting and ordering how data analysis tasks occur. This in turn sets off a chain of internal reactions resulting in the execution of additional data processing tasks previously shown in Figure 1.

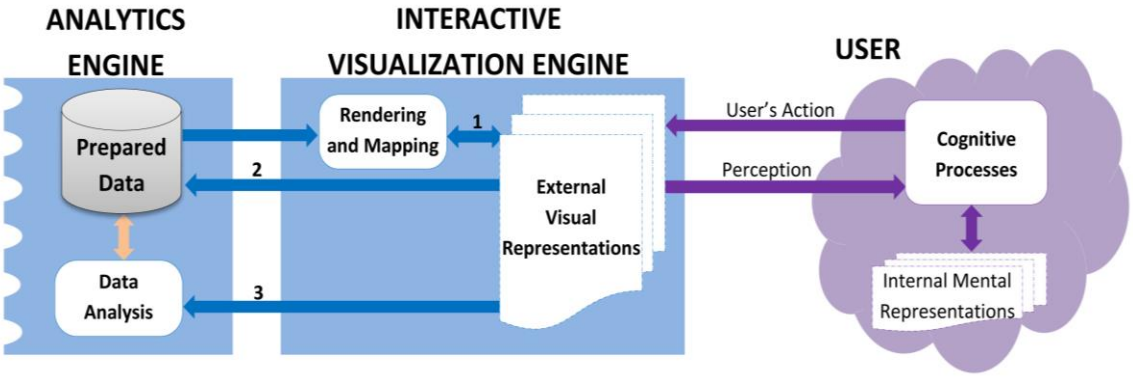



**Figure 2-2: Interactive visualization engine component in VA tools**

### 2.3.4 Discourse Mediation by VA Tools

In order to understand how the application of VA tools facilitates analytical reasoning in PH contexts, it is necessary to explicate the human-information discourse that occurs when PH stakeholders use VA tools. Analytical reasoning emerges from the

collaboration between the user and the tool (Sedig & Parsons, 2013). Consequently, the internal cognitive processes of the user and the components of the analytics and interactive visualization engines are all involved in the predominantly user-controlled dialogue with information (Hollan, Hutchins, & Kirsh, 2000; Parsons & Sedig, 2013b). As shown in Figure 3, as the user performs actions on the interface, the VA tool's visible reactions are communicated by changes in the representations, which the user can perceive.

Analytical reasoning can be conceptualized as the top level of a hierarchical structure of processes. When mediated by VA tools, analytical reasoning can be broken down into sub-activities (e.g., knowledge discovery, sensemaking), which emerge from tasks (i.e., goal-oriented behaviors such as exploring, organizing). These tasks can also be broken down into sub-tasks, which in turn emerge from the completion of lower level actions performed on the tool (e.g., filtering, annotating) (Sedig & Parsons, 2013). For instance, as shown in Figure 3, the epidemiologist engaged in analytical reasoning about the origin of WNV might first need to *discover new knowledge* about the situation in Lumcard. In order to do this, she might first need to complete the task of *exploring* the redacted health records of confirmed cases. At this point, it is possible she might choose to *filter* out unconfirmed cases, *drill* down into the demographic characteristics of confirmed cases, and then *compare* the attributes (e.g., age, ethnicity, gender etc.) to determine if a correlation exists. Thus, analytical reasoning emerges over time through a back-and-forth cyclic chain of actions, reactions, and perceptions.

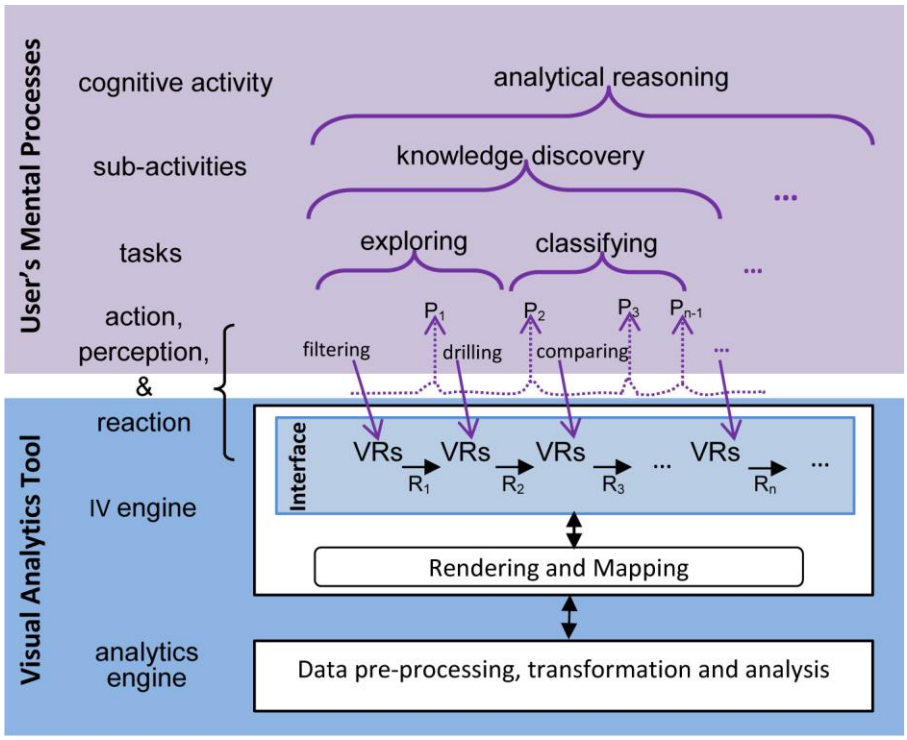



**Figure 2-3: The hierarchical structure of analytical reasoning emerging from lower level processes, adapted from (Sedig & Parsons, 2013). Where visual representations are depicted as VRs, perceptions as $P_x$, and reactions as $R_x$ (where x stands for 1, 2, 3, and n-1)**

### 2.3.5 Factors Affecting Quality of Discourse

Recent theories of cognition suggest that cognitive processes do not take place solely within an individual's head, but are distributed across social relationships, the material environment, and time (Hollan et al., 2000; Sedig & Parsons, 2013; J. Zhang & Norman, 1994). In other words, analytical reasoning, formerly conceived as a cognitive activity that occurs exclusively in the brain of the PH stakeholder, can in fact be distributed across computational tools and other PH stakeholders. As a result, in the context of VA tools, a joint cognitive system is formed between the user and the tool (Sedig & Parsons, 2013; Sedig et al., 2013). VA tools therefore play an important role in—and depending on their design can either enhance or impede—the human-information discourse. Some factors affecting the quality of the discourse are: how information is encoded in visual representations, how seamless the coordination is between the user's internal

representations and the tool's external visual representations, and how information processing is distributed between the components of the joint cognitive system (i.e., user, analytics engine, and interactive visualizations).

In VA tools, external representations not only convey information, but also guide, constrain, and even determine cognitive behavior of the user (J Zhang, 2001). The manner in which interactive visual representations are designed is an important consideration as research has shown that external representations should be appropriate for the task in which the user is engaged (for an in-depth discussion see (Parsons & Sedig, 2014)). As users perform analytical reasoning tasks, they seek to harmonize and coordinate their internal representations and the tool's external representations (Z. Liu & Stasko, 2010; J. Zhang, 2001). When processing data in such a dynamic manner, a cognitive coupling is formed between the user and the tool (Brey, 2005; Sedig & Parsons, 2013). The strength of the coupling between the user's internal representations and the tool's external representations is dependent upon a number of factors, including what actions are made available to the user and the quality of these actions (Sedig & Parsons, 2013). In most situations, interactions should allow the user to select which subset of information to display, to manipulate external representations, and to choose which analysis techniques to perform so that s/he is able to complete the task at hand (for an in-depth discussion see (Sedig & Parsons, 2013)). Another consideration relating to the discourse is the quality of interaction (i.e., interactivity) that emerges through the use of VA tools. This consideration is important because research suggests that the quality of interactions has important cognitive effects (for an in-depth discussion see (Sedig, Parsons, & Babanski, 2012; Sedig et al., 2013)). As information processing is distributed across the joint cognitive system, properly designed VA tools must take into consideration the strengths and limitations of the components of the system when distributing the requisite load of information processing in any given context (for an in-depth discussion see (Parsons & Sedig, 2013b)). These three considerations, among others, affect the ability of tools to facilitate reasoning and as a result VA tools must not be viewed as a silver bullet to alleviate all the problems facing stakeholders, as the efficacy of the human-information discourse in these tools depends on how well and human-centered their design is.

## 2.4 Benefits of Visual Analytic Tools in Public Health

VA tools are advantageous to numerous fields, including PH, because they combine the benefits of both data analytics and interactive visualization tools. In PH, conclusions or inferences drawn may need to be conveyed to different groups of stakeholders including legislators, hospital directors, or community group leaders who were not involved in the analysis process (O'Carroll et al., 1998). Information, therefore, must be conveyed in a manner commensurate with the cognitive and contextual needs of the PH workforce. Because VA tools allow users to participate in the data analysis process, and give them partial control over the system's behavior, these tools can provide the flexibility to accommodate the needs of this diverse workforce. This is beneficial to PH in a number of ways, four of which are described. Firstly, through interactive visual representations, stakeholders are able to select the most appropriate visual form from a pre-defined set to perform the task at hand. Secondly, through interaction, stakeholders are able to control their dialogue with information. This process as previously discussed is not a linear one, and VA tools support the unstructured, non-linear process of thinking and data exploration in which PH stakeholders typically engage. Thirdly, VA tools can automatically generate tailored reports for different groups of stakeholders. Finally, VA tools can also adjust and scaffold tasks in order to accommodate the cognitive needs of novice and learned stakeholders alike. The rest of this section is divided into two parts; the first describes how VA tools can address the challenge of big data in PH, while the second highlights current VA tools that can support PH stakeholders' analytical reasoning tasks.

### 2.4.1 Utility of VA Tools in Addressing Challenges of Big Data in PH

While interacting with PH data, stakeholders encounter challenges relating to the volume, variety, velocity, and veracity of data. VA tools have accounted for and are addressing these challenges.

### 2.4.2 Data Volume

PH stakeholders are overwhelmed with massive amounts of data on a regular basis, and the PH informatics community has yet to sufficiently address the need for data to be presented in a more tractable form (Higgins et al., 2011; Keough, 2002; LaPelle et al., 2006; Turner et al., 2008). Because of this deficiency, stakeholders find themselves spending more time wading through data, and less time actually addressing the health concerns of their community. As discussed in (Revere et al., 2007), "*data set 'overload'—the consequence of increasingly large data sets generated by surveys and other data collection tools—has forced many epidemiologists to become data managers, making it more difficult to analyze data from a variety of sources in order to detect disease outbreaks at an early stage."* The user-controlled environment that VA tools provide allows the stakeholder to guide the analytics engine on how to manage and analyze data. As a result, the user is still cognizant of the characteristics of the data but cedes its processing to the tool. Through the division of information processing labor, VA tools relieve stakeholders of the tedious task of managing and analyzing obscure and intractable patterns in data. Additionally, through interaction, the user is able to control the flow of data and access latent data as needed.

### 2.4.3 Data Variety and Velocity

The great variety and high velocity of PH data can impede stakeholders' reasoning. In regards to its variety, PH data is stored in different formats such as numerical, textual, geospatial, and multimedia (Rambo, 2000) and ranges from structured (e.g., health indicators survey data), to unstructured, which in its original state can only be meaningfully interpreted by the human mind (e.g., free-form paragraph in a policy brief or tweets about medical symptoms) (Guerra et al., 2007; Turner et al., 2005). In terms of its velocity, PH data is updated at varying time frames and in some situations is made available for a transient period of time (O'Carroll et al., 1998). VA tools do not merely synthesize federated data originating from a variety of sources. Through the analytics engine, stakeholders can also process various forms and structures of data, and with the interactive visualization engine, these different forms of data can be presented in a manner that is conducive to reasoning. For example, in the WNV scenario, VA tools can

help the epidemiologist reason more efficiently with tweets related to health, relevant redacted EHR, related national health policy documents, and parish climate data, without having to worry about the original form or source of the information.

### 2.4.4 Data Veracity

As PH policies and decisions have implications that affect the very fabric of society, the veracity of PH data cannot be overemphasized. PH data is often incomplete and inaccurate (Kiefer et al., 2005; LaPelle et al., 2006; Reeder et al., 2012). As a result, stakeholders are faced with the challenge of dealing with incomplete and discrepant data during reasoning. While some of these challenges require a more efficient health information exchange system, in comparison to data analytics and interactive visualization tools, VA tools are more equipped to support stakeholders. Through the inclusion of models that describe scientific uncertainty and visual representations that highlight outliers and anomalies, stakeholders are able to better understand the integrity of the data and the ramifications of possible decisions. Furthermore, as humans are better able to use incomplete data to make decisions (in comparison with computers), tools that allow for a user-guided analysis process enable users to incorporate their previous knowledge into reasoning tasks.

VA tools not only address the challenges arising from existing data repositories, but have the potential to enable the use of new sources of data (such as edge data) into PH practice. Edge data, which refers to peripheral data that exists in the immediate, surrounding environment, can provide significant information on health events and their impact—example of these include water utility data that can help make sense of how cholera spread within a city, cell tower data can facilitate understanding nurses' practices during night shift, or traffic data of the intersection in front of a hospital.

## 2.5 Current Application of Visual Analytic Tools in Public Health

Even though PH has been slow to adopt VA tools, other fields, including finance and sales, have aggressively incorporated these tools into their practice. This section highlights current VA tools both within and beyond the field of PH, and how these tools

can facilitate the work activities with which PH stakeholders are charged. It is subdivided based on the core functions of PH.

### 2.5.1 Health Assessment

Work activities in this area include investigating the occurrence of health issues, analyzing the origins and contributing factors to health hazards, and identifying health trends (National Research Council, 1988). Stakeholders engaged in these activities seek answers to a myriad of questions including what causes disease or injury, what current risks are, what trends exist, and who is at risk. As analytical reasoning emerges from the human-information discourse mediated by VA tools, the user is able to address these questions by applying a variety of analysis techniques. In addition, VA tools can provide an environment in which hypotheses can be systematically developed, supported, or refuted. One such VA tool that does this is nSpace which allows stakeholders to rapidly scan and triage thousands of search results in one display (Proulx et al., 2006). Furthermore, nSpace provides an environment that supports the generation of hypotheses and evaluation of relevant evidence (Proulx et al., 2006). Epidemic intelligence involves the early identification, assessment, and verification of potential public health hazards (Paquet, Coulombier, Kaiser, & Ciotti, 2006) and is essential to safeguarding the health of the community. To this end, there has been an increase in the use of social media data to gain insight into the condition of populations, as, for example, garnering information from Twitter to estimate flu activity faster than traditional systems (Carneiro & Mylonakis, 2009), and to gauge adverse public reaction to certain drugs (Bian, Topaloglu, & Yu, 2012). Epidemic intelligence stands to benefit from advances in textual analysis techniques, which, when incorporated into the analytics engine of VA tools, can support PH stakeholders' analytical reasoning tasks.

### 2.5.2 Policy Development

PH work activities in this area include prioritizing criteria, finding corroborating evidence, comparing possible policy options, and selecting the best option. By incorporating decision analysis frameworks, VA tools can help PH stakeholders explore the complex implications of various policy options in an interactive fashion thus

facilitating the use of evidence in policy development. Commercial VA tools such as Tableau (Tableau Software, 2013) and SpotFire (TIBCO Spotfire Software, 2013) are being used by business and finance professionals to create interactive dashboards useful for PH stakeholders. One such application in PH involves the analysis of foodborne vibriosis in the United States (Sims et al., 2011). These visualizations present various options and potential outcomes to enable decision makers to select a course of action. VA tools also allow stakeholders to rapidly access and search through available research from relevant studies. Uncertainty is inherent in policy making (Schabas, 2002) and exists in situations that have complex dynamics and interdependencies (Kramer et al., 2009). For instance, because of the rapid spread of chloroquine-resistant vectors in East Africa, it is difficult to predict the effectiveness of malaria policy in that region (D'Alessandro & Buttiëns, 2001). VA tools modeling scientific uncertainty in policy simulations can provide policy makers with more information on possible outcomes.

### 2.5.3 Assurance

In PH, after health issues have been identified and analyzed, and after policy has been developed, it falls on those stakeholders involved with assurance to ensure public awareness of preventative measures and access to health services. This includes work activities such as enforcing health laws and policies, communicating with the public, managing health resources, educating the health workforce, and evaluating the effectiveness and accessibility of health services (National Research Council, 1988). Stakeholders engaged in these activities inquire into the services being delivered, the impact programs have, the capacity of PH stakeholders to deal with outbreaks, and the supply of resources for potential epidemics. Once again, commercial VA tools such as Tableau can be utilized to ensure health resources are managed and dispensed properly. VA tools create an environment that incorporates predictive models to support PH stakeholders' reasoning. Panviz (Maciejewski et al., 2011) is an interactive predictive decision support environment that allows stakeholders to explore epidemic models and understand the effect certain response measures could have on the spread of an epidemic. It has also been used to educate Indiana PH stakeholders in designing optimal response strategies (Maciejewski et al., 2011). A similar tool is Epinome, which, in addition to

allowing epidemiologists to explore outbreaks, tracks users' interactions for post analysis (Livnat, Rhyne, & Samore, 2012).

## 2.6 Hypothetical Scenario

VA tools can meet specific challenges facing PH stakeholders as they reason with big data. We will illustrate, through the Lumcard scenario, how VA tools can potentially support analytical reasoning tasks of PH stakeholders. In this section, we demonstrate how stakeholders are able to control the flow of data, choose representations that are applicable to the task, and use various interactions to perform analytical reasoning tasks. In the process, we show how analytical reasoning emerges from lower level interactions and how information processing is distributed between the user and the tool.

In our scenario, we focus on the analytical reasoning tasks of a regional epidemiologist, who, in late November, received a phone call from the Lumcard health director about a potential outbreak of WNV. To investigate the situation, the epidemiologist will need access to various forms of data including: 1) surveillance data which includes tweets from Twitter and redacted EHR from local hospitals[3]; 2) geographical data which includes the landscape of the city, and places of high volume interaction including relevant environmental places (e.g., lakes, other local bodies of water, schools, and hospitals); 3) weather data for the city, state, and nation spanning the last ten years; 4) public epidemiological data on WNV for recent years; and 5) journal articles relating to the emergence of vector borne diseases in North America, to name a few. Possible cognitive sub-activities may include sensemaking to determine if there is in fact a WNV outbreak in Lumcard and knowledge discovery to determine the origins of this unseasonable occurrence.

In order to *make sense* of the situation in Lumcard, the epidemiologist will engage in a variety of analytical reasoning tasks that may include exploring the demographic attributes of confirmed WNV cases, comparing the situation in Lumcard to the rest of the

[3] In this scenario, the epidemiologist has access to a centralized database which stores EHRs of patients with WNV from area hospitals.

nation, triaging documents to discover relevant literature on WNV, and gathering evidence to present to other PH stakeholders. Using EpiProbe (i.e., a hypothetical VA tool) to *explore* the available data, the epidemiologist examines the EHRs to discover collective properties of confirmed cases. With the use of visual representations, the epidemiologist immediately notices a disproportionate number of adolescents (i.e., individuals between the ages of 10 – 20) with the disease. She annotates the visual representation and saves it in the evidence box.

Next, to *contrast* the situation in Lumcard with cities across the USA, the epidemiologist visualizes the total number of confirmed cases by parish for the current year. Then she interacts with the visualization to arrange cities based on the number of confirmed WNV cases. EpiProbe calculates the $25^{th}$, $50^{th}$, and $75^{th}$ percentile of cases, and further groups the cities accordingly. Using the box-plot graph the epidemiologist assesses Lumcard is one of three cities with an unusually high number of cases that are in fact outliers. She concludes from this exploration that there is indeed an outbreak in Lumcard. Next, she decides to triage the literature to find PH documents for the three cities over the last year. Her initial query results in over 125 articles. She instructs EpiProbe to narrow the list by displaying only articles with the word *'mosquito'*. This reduces the list to seven articles, and EpiProbe provides a short narrative of each of the articles by using the summarization textual analysis technique. Each of the articles indicates higher infestation of mosquito in the three cities as a result of climate change. With the annotate feature in EpiProbe, the epidemiologist writes a brief summary of her thoughts and adds these articles to the evidence box. She then directs the tool to calculate and visualize the least mean square for the dependent variable (i.e., number of confirmed cases) and the independent variable (i.e., temperature) for the three cities. She observes a positive correlation between temperature and confirmed cases in all of the cities. The corresponding scatter plot is added to the evidence box as well.

In order to *discover knowledge* about the cause of the outbreak, she returns to her initial observation of the prevalence of the disease among adolescents. At this point in time, she decides to *compare* the age, time, and number of cases in the three cities. She *selects* the image plot representation shown in Figure 4 and immediately notices three things. The

first is the cyclic nature of the cases; the second is that over the course of the last 5 years all three cities have seen a steady increase in cases. This observation provides further proof that climate change has had an impact on the prevalence of the disease. The third detail she observes is that Lumcard is the only city with a high percentage of adolescents with the disease as shown by the three small rectangular shapes in the last column. She describes her findings and adds snapshots of the visual representations as evidence.

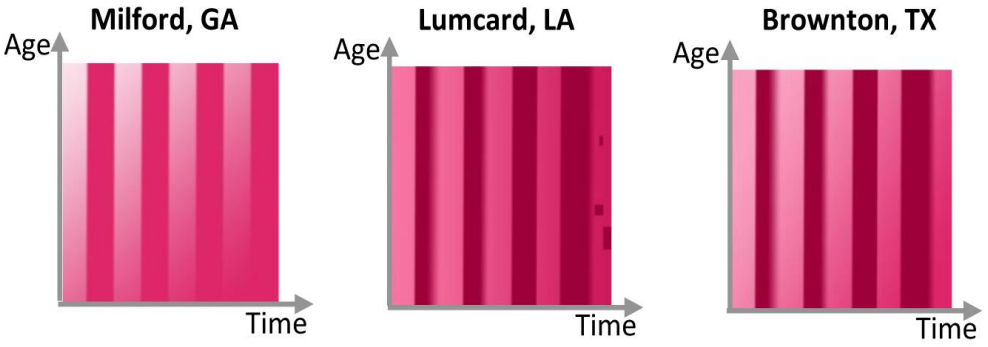



**Figure 2-4: Image plots of WNV cases for the 3 selected cities from 2008 – 2013**

WNV is transmitted by mosquitoes to humans. Thus, using a textual analysis technique, the epidemiologist *filters* and *searches* for relationships to determine if there is any correlation between the tweets by Lumcard adolescents and any references to mosquitoes or local bodies of water. As depicted in Figure 5, after a series of additional tasks, she identifies a subset of tweets referencing two parties at East Lake and West Bayou two weeks prior. At this point in time, she accesses GIS coordinates for the location and instructs the local environmental scientist to collect samples from these locations. Further instruction is given to place a warning at the sites until further investigation is completed.

As the scenario shows, all of the aforementioned tasks can be completed with the VA tool, EpiProbe. The epidemiologist is able to explore EHRs, contrast cases across the country, pay closer attention to cities that defied the norm, perform statistical analysis to determine correlation, use textual analysis techniques to search through published journal articles, and ultimately make sense of twitter data to find a possible lake location potentially contributing to the outbreak amongst adolescents. Additionally, the epidemiologist is able to delegate computationally intensive tasks to the tool, such as

searching through existing literature and finding relationships between tweets. This delegation of tasks then allows the epidemiologist to focus on the overall task of determining the causation of the outbreak in Lumcard. The benefit of VA tools is portrayed in this simple example illustrating their critical application in the PH field.

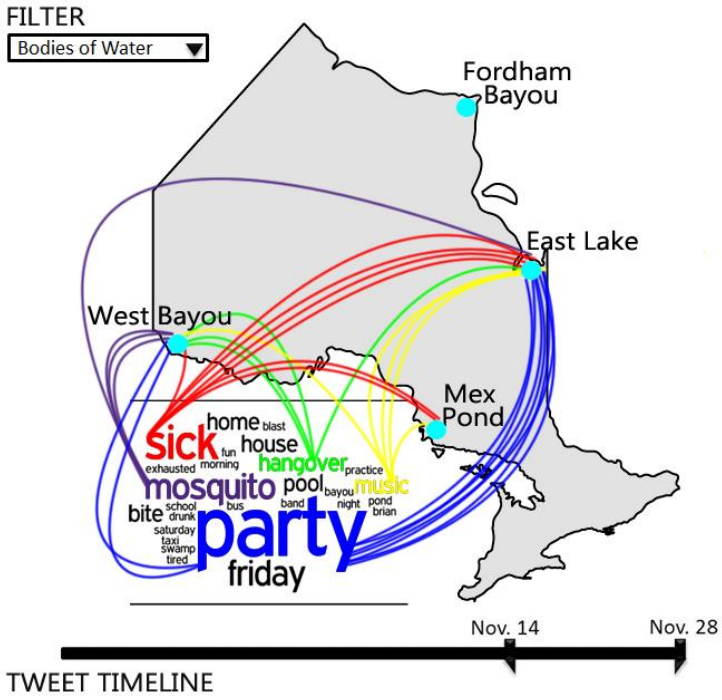



**Figure 2-5: Visual representation depicting spatial relationships between most frequent words in tweets and local bodies of water in Lumcard**

## 2.7 Summary and Conclusion

The success of an evidence-based approach to PH practice is contingent on stakeholders being able to efficiently use and reason with synthesized, federated sets of big data. As such, computational tools that support analytical reasoning can be beneficial to PH stakeholders. Through an examination of Visual Analytics (VA) tools and a discussion of challenges facing PH stakeholders, this paper has shown how VA tools can address the big data concerns of PH stakeholders.

Through the combination of interactive visual representations and advanced data analysis algorithms, VA tools create a user-guided environment in which PH stakeholders can

interact and reason with data. The data analytics engine in VA tools allows for the storage, synthesis, and analysis of, as well as the discovery of patterns within, different forms of digitally stored data. Seeking to exploit the human visuoperceptual system, in an effort to enhance cognition, the interactive visualization engine creates representations that display information in a predominantly non-textual manner. These interactive visual representations allow the user to access and control the flow and analysis of data. As cognitive processes do not take place solely in the brain of the PH stakeholder, VA tools, that allow the user to access, structure, analyze, and modify the amount and form of displayed information, help to bridge the gap between the internal representations of the user and the external representations, thus facilitating analytical reasoning.

We have shown that VA tools can facilitate collaboration, coordinate internal representations with external representations, and efficiently provide comprehensible assessments to stakeholders. Furthermore, VA tools provide flexibility which allows for the customization of the tool to cater to the cognitive, perceptual, and contextual needs of the diverse PH workforce, and ultimately facilitates stakeholder reasoning and decision-making. The features of VA tools make them suitable to address the challenge of big data in PH that arise from the data's high volume, great variety, high velocity, and low veracity. Because of these reasons, as well as the existing evidence of the success and proliferation of VA tools in other domains, we conclude that the use of VA tools can be advantageous in PH, where stakeholders must use big data to address the concerns of the populace.

## 2.8 Limitations

VA, being an area of research in its infancy, still faces major challenges in understanding how to develop sophisticated tools. While current VA tools for PH show promising initial results, more research is needed to develop this promise into tried and tested solutions. The quality of the human-information discourse that is facilitated by VA tools is dependent upon numerous factors including the integration of different sources of data, the distribution of information processing load among components of the joint cognitive system, the design of visual representations, and the operationalization of interaction techniques to support the mental tasks of stakeholders. The manner in which the

aforementioned factors are considered in developing VA tools will partly determine the utility of these tools in addressing the challenge of big data in PH. There is need for systematic and focused research within the context of PH. In other papers, we have developed preliminary frameworks to guide the development of VA tools. Further research in these areas will contribute to the development of VA tools that effectively support PH stakeholders as they interact and reason with ever more divergent, dynamic, and complex bodies of data.

# Chapter 3

# 3 Beyond Simple Charts: Design of Visualizations for Big Health Data

This chapter has been published as O. Ola and K. Sedig, “Beyond simple charts: Design of visualizations for big health data,” Online J. Public Health Inform., vol. 8, no. 3, Dec. 2016.

Please note that the format has been changed to match the format of the dissertation. Figure numbers mentioned herein are relative to the chapter number. For instance, “Figure 1” corresponds to Figure 3-1. Additionally, when the term “paper” or “article” is used, it refers to this particular chapter.

## 3.1 Introduction

Technological advances have resulted in increased data collection, digitization, and storage across many fields. In health, this data includes population surveys, electronic medical records, genomic sequencing data, gene microarrays, and social media posts on ailments. Health data is often big data due to its high volume, low veracity, great variety, and high velocity. Big health data has the potential to improve productivity, eliminate waste, and support a broad range of tasks related to disease surveillance, patient care, research, and population health management. For instance, it has been estimated that using big data in the United States can save the healthcare industry $300 billion dollars a year (Groves, Kayyali, Knott, & Van Kuiken, 2013). However, big data’s impact is contingent on the availability of tools that can help derive meaning from it. To date, health lags behind other fields (e.g., finance and business) in the development of computational tools for big data (Dhar, 2014; *Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations*, 2011; Groves et al., 2013; Shneiderman et al., 2013).

Interactive visualizations[4] are a category of computational tools that store and process data, represent it visually, and allow for its interactive exploration. They have the potential to amplify big data's utilization. With interactive visualizations, individuals can access underlying data, change how data is represented, manipulate various visual elements, and in certain tools control analysis tasks (Ola, Buchel, & Sedig, 2016). Visualizations can be used in health to support a variety of tasks, some of which include: tracking the geographic distribution of diseases, developing health policies, analyzing the prevalence of disease, triaging medical records, predicting outbreaks, and discovering at-risk populations. Currently, many health professionals[5] rely on Microsoft Office and off-the-shelf business intelligence tools to perform their data-driven tasks (Carroll et al., 2014). Majority of these tools use simple visualizations, such as scatter plots, heat maps, bar charts, choropleth maps, and radar charts (Carroll et al., 2014; L. Zhang et al., 2012). These visualizations typically only represent one or two facets of the data (e.g., attributes, relationships) (Aimone, Perumal, & Cole, 2013; Faisal, Blandford, & Potts, 2013; Kosara & Miksch, 2002; Rind et al., 2013). When working with big data, there is the challenge of needing to analyze non-explicit and unknown relationships among the data elements as well (Cybulski et al., 2013; Endert, Hossain, et al., 2014). To address this challenge, users also need to be able to explore various data elements and facets simultaneously. Such being the case, having access to only one or two data elements at a time is not sufficient. Users need to be able to perform related tasks and see many facets and elements of data at the same time so they can quickly perceive patterns, develop insights, and create and discard hypotheses. Consequently, existing simple and chart-like visualizations are not effective at supporting tasks involving large, complex health datasets (Cybulski et al., 2013; L. Zhang et al., 2012).

Rapid rise in health data necessitates creation of visualizations that encode multiple facets of data simultaneously to support complex health-related tasks. In recent years the need for advanced visualizations that address the challenges of big data has been highlighted

---

[4] In the rest of this article, the terms 'interactive visualization' and 'visualization' are used interchangeably.

[5] In this article, we use the term 'users' to refer to all individuals, both professionals and laypeople, who use visualization tools.

(Cybulski et al., 2013; Gotz & Borland, 2016; L. Zhang et al., 2012). In addition to the need for such visualizations, close attention must be paid to how they are designed as bad design can have unintended negative consequences (Sedig et al., 2013). The design of visualizations for big data is a labor-intensive process and requires an understanding of the data, user's tasks, cognitive and perceptual considerations, and, of course, visualization techniques and their utility. In their seminal work on visual analytics, Thomas and Cook note that we need new methods to simplify the development process of visualizations for big data (Thomas & Cook, 2005). Researchers have tried to organize the plethora of existing visualization techniques to give structure to the selection and design process (Aigner, Miksch, Schumann, & Tominski, 2011; Heer, Bostock, & Ogievetsky, 2010). These classifications are helpful in the selection of some familiar visualizations; however, the emergence of big data and its attendant tasks ask for new frameworks, ones that go beyond classification and are more robust and flexible.

There is confusion, lack of direction, and shortage of guidelines about how to create effective visualizations for health data (Carroll et al., 2014; Folorunso & Ogunseye, 2008; Turner et al., 2008). Given the high stakes in health, be it in education or outbreak detection, it is essential for these visualizations to be designed systematically—hence, the need for framework-based approaches to the design of health data visualizations. Even though a framework should support systematic design, it must not constrain creativity; furthermore, it must allow designers to come up with novel and elaborate visualizations that capture and encode the complexity of new data (Purchase et al., 2008; Sedig et al., 2013; Thomas & Cook, 2005). A design framework should integrate relevant concepts from multiple fields, be theory-driven, be conceptually sound, bring much-needed structure to the design and evaluation process, and provide a common and consistent vocabulary to design thinking. Without the structured design thinking provided by a framework, design of visualizations can take on an ad hoc approach without much systematicity (Purchase et al., 2008).

To this end, Sedig and Parsons (Sedig & Parsons, 2016) have recently developed a comprehensive framework for the design of visualizations for human-information interaction. This framework includes a pattern language. This language provides

designers with 14 patterns for mapping data to visual structures and a simple grammar-like syntax for blending these patterns. The goal of this framework is to enable designers to create elaborate and sophisticated visualizations in a systematic, principled manner with interactive task possibilities at the foreground of design thinking. The *purpose of this paper* is to demonstrate how designers of health visualization tools can go beyond simple chart-like visualizations and design novel visualizations for big health data. We use the pattern language and apply it to large public health data to illustrate how elaborate and complex visualizations for health-driven tasks can be created in a systematic way.

The rest of the paper is organized as follows. Section 2 provides the terminological and conceptual background of the paper. Section 3 presents elements of Sedig and Parsons' framework used to develop our visualizations. Section 4 details the design of four non-trivial visualizations for global health data that we have implemented. Finally, Section 5 concludes the paper.

## 3.2 Background

In this section, we first describe public health data and the tasks in which professionals engage. Then we describe visualizations and highlight how they are and can be used to support big data tasks.

### 3.2.1 Big Data in Public Health

Data collected from the population or on the population is used to assess the health of communities, develop policies, manage resources, and educate the public about health issues (Herland et al., 2014; Ola & Sedig, 2014). In this paper, we use the term data item to refer to any entity, property, or relationship within a dataset such as a database record, tweet text, image, document, geolocation, or property. Public health data is voluminous, gathered either by traditional (e.g., hospitals) or non-traditional means (e.g., social media or sensors), and is stored in different formats (e.g., geospatial, textual, numerical) (Fuller, 2010; Herland et al., 2014; Revere et al., 2007). This data is often aggregated at various levels of granularity. For instance, public health datasets may be aggregated by geographic or demographic attributes, and, as a result, one dataset may portray cancer patients by income level, while another dataset may focus on the country in which people

live. In addition, public health data is created at different time intervals (Herland et al., 2014). For instance, population survey data may be collected once a year, while surveillance data is updated hourly. This varying velocity of data impacts how it is stored and processed. Furthermore, the accuracy and completeness of public health data vary across countries and organizations (Kiefer et al., 2005; LaPelle et al., 2006; Lozano et al., 2012). Data that exhibits one or more of these qualities of big data presents processing challenges for health-related human-data interaction tasks.

### 3.2.2 Public Health Tasks

Professionals and laypeople use and interact with public health data for a variety of reasons. Professionals, charged with improving and protecting the health of the community, use this data to detect disease clusters, predict outbreaks, identify risk factors, prepare intervention procedures, evaluate strategies, educate the community, and analyze the occurrence and causation of health problems (Carroll et al., 2014; Ola & Sedig, 2014). At the same time, the general public may need such data to understand health risks, recognize biases in health information, vote on environmental issues, and make decisions about their lifestyle (Gazmararian et al., 2005). Irrespective of background, many people are in need of public health data to perform health-related tasks. As need for exploring various facets of big data grows, human-data interaction tasks become more complex. For instance, to make sense of the global spread of the Zika virus, a college student may choose to *browse* through trending tweets with the hashtag #zikavirus, *rank* tweets based on their reputability, and then *triage* new articles linked to reputable tweets. These tasks are inter-related, hierarchical in nature, and emerge from the completion of smaller tasks (Rind, Aigner, Wagner, Miksch, & Lammarsch, 2015; Sedig & Parsons, 2013). To perform the task of ranking tweets, the user may first *filter* tweets by Twitter handles to focus on tweets from health organizations, and then *arrange* the remaining tweets by the number of retweets. As these tasks are typically co-occurring, non-routine, and performed in a non-linear fashion, there is need for tools that support the convergent and divergent processes in which users engage. In the context of big data, interactive visualizations can significantly enhance the completion of such tasks (Sedig & Parsons, 2016).

### 3.2.3 Visualizations for Big Data Tasks

Visualizations[6] are composed of basic visual marks (e.g., dot, point, line) organized into different structures. These visual marks have certain properties (e.g., size, color, angle, texture) that are used to encode data items (Sedig & Parsons, 2016). Visualizations can support users' tasks by synthesizing and integrating data from various sources, reducing the search for information, and enhancing the discovery of patterns, trends, correlations, and outliers (Ola et al., 2016; Thomas & Cook, 2005). However, the extent to which a visualization supports a task depends on how the data, and how much of it, is encoded (Parsons & Sedig, 2013a; J. Zhang & Norman, 1994). In this context, users' discourse with data is through the visual representations. As a result, the visual form in which the data is presented can either enhance or hinder tasks. For example, consider a situation in which a user needs to make sense of the geographical spread of Chikungunya across islands in the Caribbean; using a bar chart to represent the number of cases in each island may not be as effective as using a map. As big data tasks seldom occur in isolation, there is need for visualizations that not only encode data effectively, but also support inter-related tasks and allow users to explore various facets of the data simultaneously.

Currently, simple visualizations found in Microsoft Office and off-the-shelf business intelligence tools are typically used by health professionals (Carroll et al., 2014). Simple visualizations usually only encode one or two aspects of the data. For instance, a bar chart, heat map, or pie chart that shows the mortality rates in sub-Saharan countries is a simple visualization. While such visualizations are beneficial for simple tasks, they are less effective for more complex tasks. One approach to using simple visualizations for multifaceted data is representing facets in a single visualization and using animation to show other aspects of the data. A well-known example is Gapminder Trendalyzer, which shows trends in multivariate data ("Gapminder," n.d.). While this approach may be beneficial for narrative tasks, it is not always effective for analytical tasks. Because animated visualizations are temporal and substitutive, as one representation replaces another in time, users need to recall previous states of the visualization and their short-

[6] Henceforth, the term 'visualizations' refers to both static as well as interactive visualizations.

term memory can become overloaded (Sedig, Rowhani, & Liang, 2005). As data increases, this approach has been shown to result in an inaccurate understanding of trends (Robertson, Fernandez, Fisher, Lee, & Stasko, 2008). Another approach represents facets in multiple visualizations distributed through space. Data dashboards, typical in business intelligence tools, employ this approach (Al-Hajj, Pike, Riecke, & Fisher, 2013). Though beneficial, as visualizations crowd the dashboard, users are forced to mentally combine representations to perform tasks (Wang Baldonado, Woodruff, & Kuchinsky, 2000). Furthermore, data dashboards typically organize visualizations in a tabular manner, regardless of how the data are related. When the external organization of information does not represent the data appropriately, the internal mental process of users may be negatively impacted (Purchase et al., 2008; Sedig et al., 2013; Thomas & Cook, 2005).

There is a need to move beyond simple visualizations to more sophisticated visualizations that encode multiple aspects and/or layers of the data within the same space. Researchers have noted that the very nature of big data and its associated tasks require the development of novel visualizations that help with pattern identification and analysis of large and complex data (Fan & Bifet, 2013; Gorodov & Gubarev, 2013; Heer & Kandel, 2012; Jagadish et al., 2014). A review of off-the-self business intelligence tools suggests that these tools tend to focus on simple visualizations, with limited capability for handling large complex data (L. Zhang et al., 2012). Non-trivial visualizations that effectively encode data items can play a prominent role in how people use big data and interact with it (Cybulski et al., 2013; Gotz & Borland, 2016; Heer & Kandel, 2012). In the context of health-related visualizations, in a systematic and comprehensive review, Carroll et al. (Carroll et al., 2014) suggest that there is a need for tools that represent large multivariate datasets that have multiple levels, various relationships, and/or layers of patterns. However, the ability of visualizations to facilitate big data tasks is contingent on their proper design, which is often challenging. This challenge is particularly amplified in healthcare, where it has been noted that visualization design is not as advanced as in other disciplines (Shneiderman et al., 2013). In the next section, we present a pattern language that we believe can help designers with systematic, yet creative and flexible, design of non-trivial visualizations for big data in health.

## 3.3 Pattern Language

When dealing with simple tasks, such as ranking diseases solely based on their mortality rate, designing a visualization is straightforward. However, as tasks become more complex (i.e., require the completion of subtasks) and the nature of the data more varied, design of visualizations becomes less apparent (Heer & Kandel, 2012; Sedig & Parsons, 2016). Part of the challenge of developing tools for big data is determining how to structure or organize data items within visualizations (Ekbia et al., 2015). As the external organization of information affects users as they perform tasks, there is a need for frameworks to help structure the design of elaborate visualizations. Sedig and Parsons have proposed a comprehensive design framework composed of conceptual elements including a pattern language, design process, and spaces. In this paper, we focus on their pattern language and describe how it can support the development of elaborate visualizations for big data. The pattern language consists of 14 abstract patterns and a syntax for describing how patterns are blended. These patterns are described next.

### 3.3.1 Descriptions of Patterns

Sedig and Parsons define a pattern as a regularity in some dimension (Sedig & Parsons, 2016). Their goal was to identify patterns that help organize information items by mapping them to visual structures. The patterns operate at an abstract level and are independent of any particular technology, platform, or domain. As a result, they can be used across domains to help create novel visualizations[7]. The 14 patterns are described next:

- **Area:** used to map data items onto visualizations in such a way that their boundary, shape, region, and/or area are encoded.
- **Branch:** used to map data items onto visualizations and organize them in a branched and/or subdivided fashion.
- **Cell:** used to map data items onto visualizations and organize them by segmenting, compartmentalizing, or containing them within cell-like structures.
- **Coordinate:** used to map data items onto visualizations and organize them with respect to a frame of reference.

[7] For an in-depth discussion on the identification and naming process of the patterns, the reader can consult the book: *Design of Visualizations for Human-Information Interaction* (Sedig & Parsons, 2016).

- **Cycle:** used to map data items onto visualizations and organize them in a circular, wheel-like, rotational, spiral, and/or cyclical fashion.
- **Fusion:** used to map multiple data items onto a single visualization in a continuous fashion, such that the items are integrated and fused together.
- **Group:** used to map data items onto visualizations and organize them by congregating them close to each other.
- **Hierarchy:** used to map data items onto visualizations and organize them in a hierarchical, multi-level, pyramid-like fashion, where higher levels are superior to or contain and encompass lower level items.
- **Link:** used to map data items onto visualizations and organize them by connecting them together using paths, routes, lines, or other similar structures.
- **List:** used to map data items onto visualizations and organize them by placing in a sequential, successive fashion.
- **Spectrum:** used to map data items onto visualizations and organize them in a spectral fashion. Often instantiated using multiple saturation or luminance values of a particular hue, or using multiple hues or textures.
- **Stack:** used to map data items onto visualizations and organize them by placing on top of one another in a piled or stacked fashion; visualizations are often placed on top of one another such that they are touching or are very close together.
- **Token:** used to map one or more data items onto a visualization that can be regarded as a unit, whether in atomic form or composite form made of discrete parts.
- **Track:** used to map data items onto visualizations and organize them in a lane-, stripe-, and/or track-like fashion.

The patterns are divided into three groups: (1) *primary*, (2) *substrate*, and (3) *relational*. The first category, primary, consists of the Token and Fusion patterns. These two patterns often act as primary building blocks for creating visualizations. The second category, substrate, consists of the Area, Cell, Coordinate, and Track patterns. These four patterns are often used for designing underlying structures in/on which other representations are placed. The third category, relational, consists of the Branch, Cycle, Group, Hierarchy, Link, List, Spectrum, and Stack patterns. These patterns are often used for creating structures that encode relationships, variations, and/or movements among data items.

The patterns in the language are not concrete structures and, as a result, must be instantiated as visual structures. For example, the Token pattern—which is used to map data items onto a single unitized visual representation—may be instantiated as a dot to represent each cause of death, or a square to represent the incidence rate of breast cancer

in developing nations. Any pattern can be instantiated using many different structures. This flexibility promotes creativity and supports the creation of a diversity of visualizations. Every visualization is an instantiation of one or more blended patterns and is not the pattern itself. Next, we discuss how patterns can be blended.

### 3.3.2 Pattern Blending and Syntax

Sedig and Parsons note that "Designers can blend different patterns to devise representational structures that have different organizational affordances" (Sedig & Parsons, 2016). In other words, instances of different patterns can be blended to create sophisticated visualizations that are beneficial for showing different aspects and features of the data. For example, to communicate the grouping of risk factors that contribute to a disease, designers can blend the Token and Group patterns together. When instantiated, the blending of these two patterns results in a visualization that conveys both the uniqueness and classification of each risk factor. The pattern language employs a simple syntax to represent the blending of different patterns. The syntax has three elements:

- 14 codes (e.g., TK for Token and CR for coordinate) to denote the different patterns; however, in this paper we use the pattern words and eschew the codes to promote comprehension;
- the symbol "•" to denote a blending, where blended patterns appear together in square brackets "[]"; and
- the symbol "∈" to denote that a visualization or representational structure "is derived from," "is based on," "instantiates," or "is an instance of" a blending.

For example, the expression *V*∈[Token•Hierarchy•Cell] signifies that the visualization, *V*, is derived from the blending of the three patterns. It is important to note that the ordering of blended patterns does not affect the instantiated visualizations. Instances of patterns can be blended in various ways to support users' tasks, including: nesting (i.e., placing one inside of another), overlapping and layering (i.e., placing one on top of another), and placing them side by side. One strength of this framework is that through pattern blending, complex data structures can be modeled and then instantiated.

To illustrate how patterns can be blended and instantiated, consider a designer charged with creating a visualization for making sense of tweets from individuals infected with a rare vector-borne disease. Typically, in public health, either a heatmap or bar chart is

used to visualize this. The grouped bar chart depicted in Figure 1a is a [Group•Coordinate•Token]-based visualization that encodes the number of tweets with specific keywords in each area of interest. With the pattern language, the designer can create a new visualization by first choosing which aspects of the data to organize. For instance, she can decide to communicate the uniqueness of each tweet and geographical location (Token) and the spatial distribution of the location (Area). She can also convey the relationship between each site and the tweet (Link) as well as the geographical group to which each tweet belongs in a single visualization (Group). At this stage, the designer proceeds to create a [Token•Link•Area•Group]-based visualization. She instantiates the Token pattern for geographic locations using a circle and keywords with text. She instantiates the Area pattern using a map-like structure, the Link pattern with lines from each keyword to geographic location, and the Group pattern is encoded using color. The resulting visualization based on the blending is depicted in Figure 1b. With this visualization, the user of the tool gets a more comprehensive representation of the space and then can generate hypotheses of how the disease spreads. We use this example not to suggest that this is a good visualization, but, instead, to demonstrate the flexibility and creativity afforded by blending patterns to map different aspects of the data to an integrated visual structure. Indeed, the representation on the left may be better suited for the simple task of comparing the occurrence of keywords in tweets.

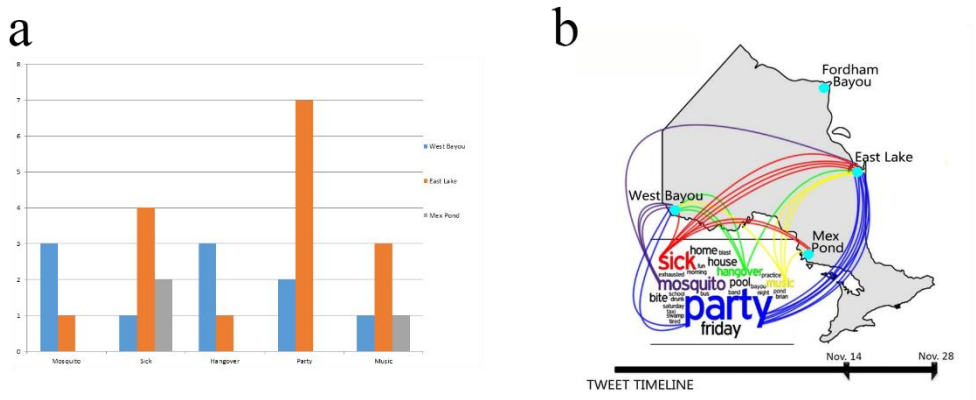



**Figure 3-1: (a) Grouped bar chart (b) Alternative visualization for making sense of tweets**

Instead of dealing with thousands of visualization techniques, by using the pattern language, based on the organizational structures that they want to convey, designers can select which patterns to blend and then design a visualization. The abstract nature of the patterns allows for flexibility and creativity as the same blending can result in different instantiations. In the next section, we demonstrate the utility of the pattern language to help designers of health visualization tools convey more data in a systematic manner.

## 3.4 Systematic Design of Visualizations for Big Public Health Data

The four visualizations presented in this section are part of a tool designed to facilitate making sense of the global burden of disease through an analysis of causes and risk factors[8] associated with mortality across the world. First, we present a high-level overview of the overall activity of sensemaking and the datasets used and then delve into the design of each visualization.

The whole of public health data relevant to understanding cause and risks attributed to mortality across the world is diverse. As data collection and access vary within each continent, and the quality of collected data is not easily verifiable, we utilize standardized data from the Institute for Health Evaluation and Metrics (IHME) (Institute for Health Metrics and Evaluation, 2013). This data includes a large number of attributes and has been gathered from various sources. The level of complexity of the data requires that it be analyzed at many levels of granularity. While the size of the data is not in the terabytes, the highly varied nature of this data is a characteristic of big data (Heer & Kandel, 2012). When combined, the datasets include over 12 million records that present mortality estimates for 57 risk factors and 235 causes of death that fall into 17 age groups[9] across 187 countries.

[8] For the remainder of the paper, we will use the terms risks and risk factors interchangeably.

[9] While IHME data includes 20 different age groups, we only use 17 of them, as the mortality estimates for the three age groups representing children under the age of 1 is not available for all datasets.

This data is further aggregated at the level of clusters. We use the term cluster to refer to an intermediary level of grouping. For example, the cardiovascular cluster of causes includes ischemic heart disease, hypertensive heart disease, cardiomyopathy, hemorrhagic stroke, and other diseases. There are 21 cause-clusters which are further classified into three main groups: 1) non-communicable, 2) injury-based, and 3) communicable, maternal, neonatal, and nutritional. There are ten risk clusters which are categorized into three groups: 1) behavioral, 2) metabolic, and 3) environmental and occupational. From a geographical perspective, mortality rates have also been aggregated at the level of geographical clusters and regions. There are 21 clusters (e.g., western sub-Saharan Africa, southeast Asia) and seven regions (e.g., Asia, Europe). Age-distributed mortality is also aggregated into five main age groups: under 5, 5-14, 15-49, 50-69, and over 70. Some datasets provide estimates for specific years (e.g., 1990, 2010, and 2013), while others span timeframes (e.g., 2000-2010 and 1970-2010). In general, to make sense of data, users perform a variety of tasks, including searching and filtering data; organizing, categorizing, and examining relevant data; developing, proving, and discarding hypotheses; and integrating data into mental models (Bodnar, 2005; Pirolli & Card, 2005). Providing users with means to explore data through different perspectives is beneficial to sensemaking. In the following subsections, we first present visualizations that explore the burden of disease from three perspectives: demography, chronology, and geography and then conclude the section with an overview visualization.

### 3.4.1 Demography Visualization

Demography is the study of human populations with respect to various subjects, including birth and death rate, socioeconomic status, and age and sex distributions. To make sense of the burden of disease, we focus on age-specific death rates for different causes and risk factors across regions of the world. The datasets include estimates of overall mortality, cause cluster-specific mortality, and mortality attributed to risk clusters for different age groups across geographical regions. As opposed to using the seven regions of the world, we use the country clusters created by IHME so that users can explore demographic trends at a lower level of granularity. Data that approximates death resulting from specific risks at the level of cause-clusters is also utilized. Users'

sensemaking tasks include: identifying different age groups and understanding how they are classified; identifying and distinguishing cause and risk clusters by their groupings; exploring the distribution of death across age groups; and comparing mortality for specific age groups across geographical regions. Ranking the clusters for specific age groups and comparing trends across age groups are additional relevant tasks. To describe the visualization that supports these tasks, we will discuss the five sub-visualizations that represent age groups, cause-clusters, risk clusters, country clusters, and relationships of mortality across these facets. This approach of describing a visualization by the sub-visualizations that support its main tasks will be used for the chronology, geography, and overview visualizations as well.

We organize our data according to age to emphasize demography. We want users to be able to locate each unique age group in the visualization; for this, we use the Token pattern and instantiate it as an oval-like shape. Each oval represents a unique age group (1-4, 5-9, 10-14, etc.). To support exploration, we arrange age groups in a sequential fashion using both the List and Coordinate patterns. A polar coordinate system on which oval shapes are placed next to each other instantiates [List•Coordinate]. To support users' understanding of the larger categories to which age groups belong, we organize the age groups by placing them close to each other and contained in a larger oval shape, thus instantiating the Group pattern. Figure 2 shows the [List•Coordinate•Group•Token]-based sub-visualization. This visualization supports locating age groups and recognizing how the groups are combined into larger groups.

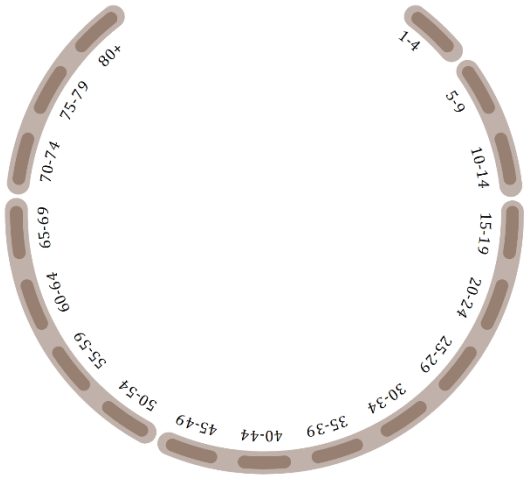



**Figure 3-2: Demography sub-visualization for age groups**

For each age group, users need to explore how different cause-clusters contribute to mortality. To do this, they will need to identify cause-clusters and their groupings, rank clusters for each age group, and assess trends across age groups for specific clusters. To support these tasks, we first use the Token pattern to organize each cluster, instantiated as an arc. Certain age groups do not have all cause-clusters; these data items are encoded using gray circles (see Figure 3a). To emphasize each cluster's group, we also organize clusters with the Group pattern, which is instantiated using color. For the cause groups, we use *blue*, *red*, and *black* for non-*communicable*, *communicable*, and *injury* clusters respectively. This instantiation of [Token•Group] is used in other visualizations, and, henceforth, we will not describe it in detail. To support comparison, we utilize the Stack pattern. Arcs are placed on top of each other to denote co-occurrence for the age group as well as their rank. Clusters are stacked in order of their rank, with the cluster that accounts for the most deaths at the top. Figure 3a shows the instantiation of [Stack•Group•Token] used to represent the ranking of cause-clusters for 1- to 4-year-old children. As depicted, there are two cause-clusters that do not contribute to death, and the highest ranking cluster falls under the communicable disease group. To encode the ranking for all age groups, we use the same polar coordinate structure (i.e., an instantiation of [List•Coordinate]) from the first sub-visualization. The main difference between the two sub-visualizations is that, instead of an oval-like shape, we use the [Stack•Group•Token]-based visualization. The resulting

[Stack•Group•Token•List•Coordinate]-based visualization is shown in Figure 3b. This sub-visualization supports locating cause-clusters and understanding the ranking of clusters for each age group, as well as trends across age groups. For instance, as depicted in Figure 3b, users can observe that for the last three age groups (i.e., individuals >=70), the three deadliest cause-clusters are within the non-communicable group (as denoted by the three blue arcs at the top of the last three segments in the visualization). Users can also observe how for younger age groups (i.e., 1-14 years) the highest ranked cluster falls under communicable diseases, which is expected as this group includes neonatal disorders. Figure 3c depicts an alternative configuration of the visualization in Figure 3b. In this mode, the neglected tropical diseases and malaria cluster has been selected[10] so users can observe trends across age groups. Figure 3d portrays the risk cluster sub-visualization, which is organized in a similar fashion; the main difference is the colors used to encode risk groups. Light shades of *orange*, *green*, and *pink* are used for *metabolic*, *behavioral*, and *environmental and occupational* risk groups respectively. Users may notice that not all risk factors contribute to mortality in younger individuals. In particular, metabolic risk clusters (which are encoded as orange arcs) do not contribute to death for individuals under the age of 25.

[10] As the focus of this article is on visualization design, we do not go into the details of the interactive features of this tool.

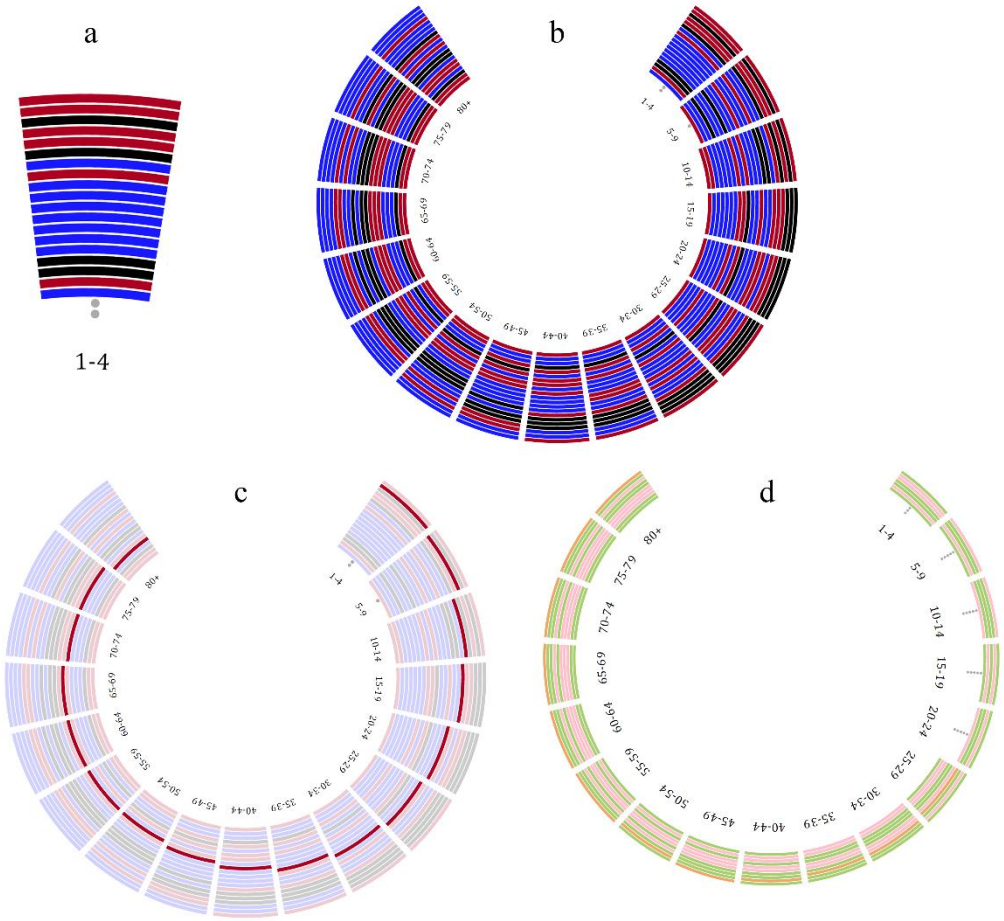



**Figure 3-3: (a) visualization of cause-clusters for children 1-4 years old (b) cause-clusters ranking sub-visualization for all age groups (c) cause-clusters sub-visualization with the neglected tropical diseases and malaria cluster emphasized (d) risk clusters sub-visualizations for all age groups**

To enable interpretation of age-specific mortality for country clusters we want users to be able to compare mortality rates across regions for a specific age range, and so we use the Coordinate pattern. For each age group, the scale is different so as to emphasize trends across country clusters as opposed to across all ages. We use the List pattern and place regions side by side in a successive fashion. The locations are ordered left to right by their region starting with the region with the highest mortality rate for all ages and ending with the lowest. The ordering of regions is as follows: Europe, sub-Saharan Africa, high-income North America, Pacific, Asia, Latin America and the Caribbean, and North Africa

and the Middle East. Using the List pattern in this manner supports comparison within regions. Figure 4a shows the [Token•List•Coordinate]-based bar charts for age groups 15-19 and 75-79. Similar to the previous three sub-visualizations, we use an instantiation of [List•Coordinate] to organize mortality for all age groups. The resulting [Token•List•Coordinate]-based sub-visualization is shown in Figure 4b. Users can observe that for younger ages mortality varies widely across country clusters as opposed to older age groups where mortality is relatively consistent. In this sub-visualization we use [List•Coordinate] in different ways. One instantiation is the 2D bar chart, while the other is at a higher level of granularity and orders the bar charts (for all age groups) on a polar coordinate system. This flexibility in how designers instantiate pattern blendings is one of the strengths of the pattern language.

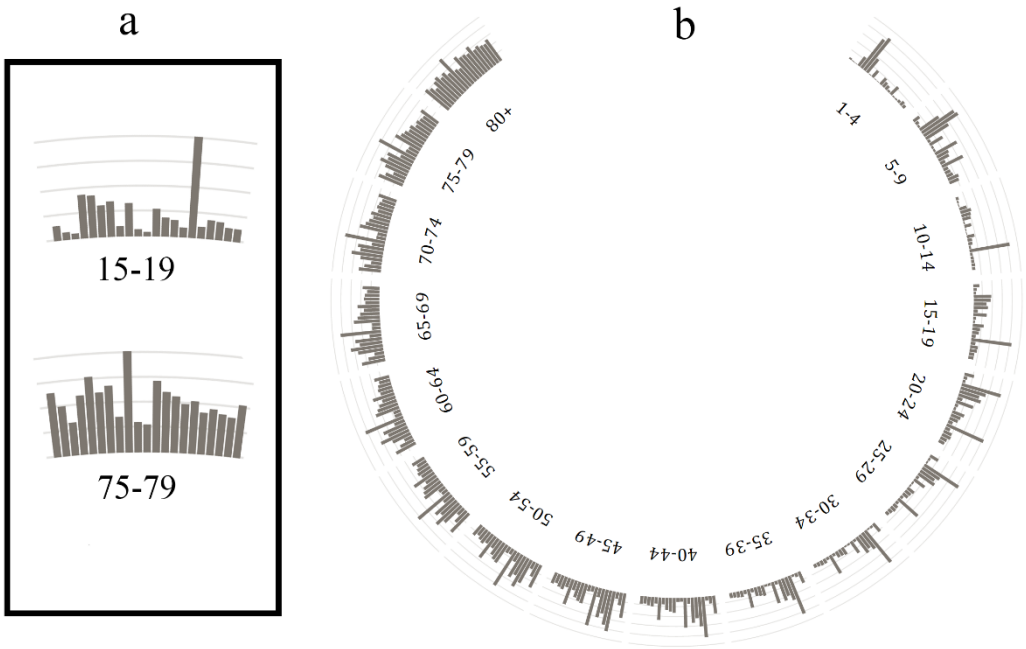



**Figure 3-4: (a) [Token•List•Coordinate]-based bar charts for age groups 15-19 and 75-79 (b) demography sub-visualization for locations**

In addition to understanding mortality for each aspect of the data (i.e., country clusters, cause-clusters, and risk clusters), it is also of benefit to explore relationships among the different aspects. [Group•Token] is instantiated as color-coded circles to represent each

cluster. We use the Coordinate pattern to organize aspects as three axis-like structures and the List pattern to organize the different clusters in each aspect as shown in Figure 5a. The clusters in each aspect are arranged by rank with the cluster with the highest aggregated mortality rate at the top. As the number of relationships is large, we only encode relationships that fall above the third quantile (i.e., top 25%). To show the presence of a relationship between aspects, we use the Link pattern, encoded as a curved line. The resulting [Coordinate•List•Group•Token•Link]-based visualization is shown in Figure 5b. As depicted, the south Asia country cluster is selected, and from this visualization users can surmise that addressing the issue of water and sanitation in south Asia will significantly impact death from diarrheal and lower respiratory diseases for people between the ages of 5 and 14.

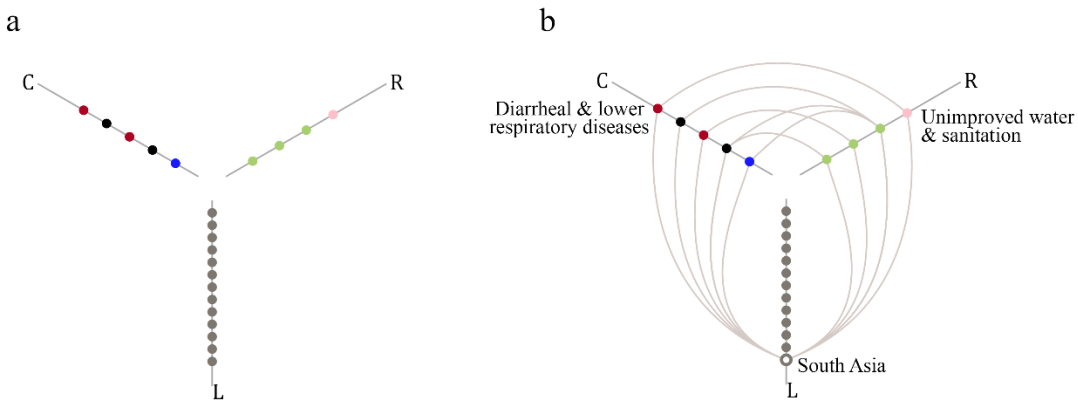



**Figure 3-5: (a) Coordinate axes for cause, risk, and location clusters (b) Demography sub-visualization for relationships between cause, risk, and location clusters for individuals between the age of 5 and 14**

Each of the five sub-visualizations discussed above represents one aspect of the demographical distribution of mortality. One design intention is to facilitate the exploration of cause, risk, and country clusters independently of each other as well as simultaneously. To organize the sub-visualizations to support this task, we use the Track pattern which places the visualizations in a lane or track-like fashion. With this pattern, we can highlight the individual nature of each sub-visualization. As four of the sub-visualizations use the same polar coordinate system to organize data items, we also use

the Stack pattern to show relationships across the four sub-visualizations. A [Stack•Track]-based structure is used to organize the sub-visualizations as depicted in Figure 6a. The inmost lane encodes the age clusters; placed on top of that is the cause visualization, then the risk visualization, and finally the location visualization is the outermost lane. The fifth sub-visualization is put in the center as shown in Figure 6b. Organized in this manner, we can convey both the uniqueness of each sub-visualization while at the same time show co-occurrence of common age groups across visualizations. It is important to note that the instantiation of [Stack•Track] is not at the same level as previous pattern blendings; here we are using the pattern language to organize sub-visualizations as opposed to individual data items. Figure 6b shows the [Stack•Track•Token•Group•Link•List•Coordinate]-based visualization for demography. The visualization provides a dense lens through which the data can be explored; its initial configuration encodes over 820 data items. Each of which serves as a selector to reveal latent data. With this visualization, users can perform a series of inter-related tasks that facilitate making sense of the demographical distribution of mortality. Through interaction, users can increase or decrease the amount of data that is visible.

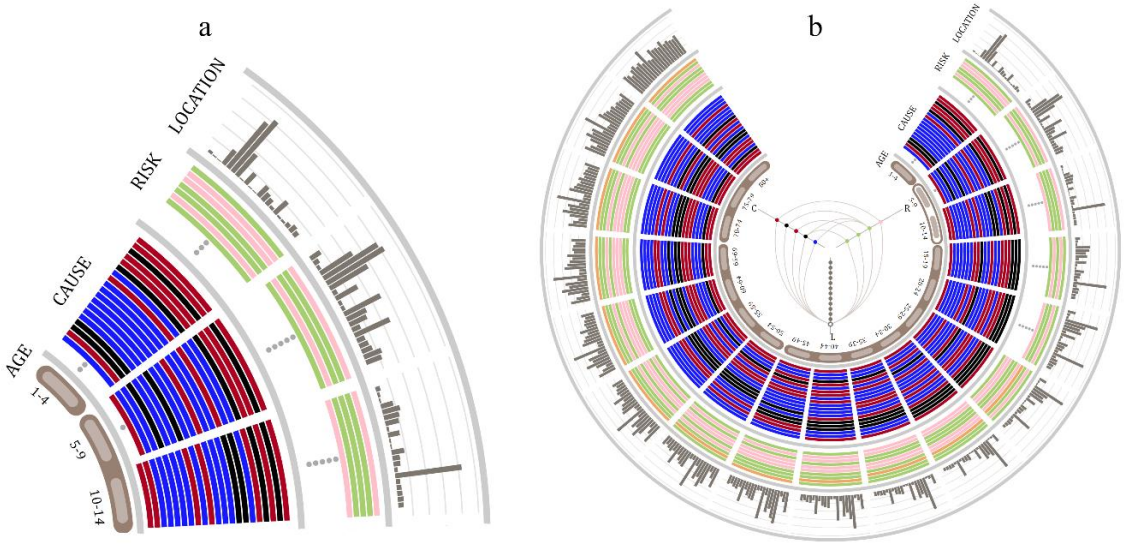



**Figure 3-6: (a) Enlarged partial view of the first four sub-visualizations for demography (b) Overall visualization for demography based on [Stack•Track•Token•Group•Link•List•Coordinate]**

### 3.4.2 Chronology Visualization

Chronology is concerned with the arrangement of events in order of their temporal occurrence. Here we describe a visualization that allows users to explore temporal trends in mortality. We utilize datasets that provide rates for cause and cause cluster-specific mortality in 5-year increments between 1990 and 2010. To make sense of temporal trends of mortality, users' tasks include recognizing time intervals, identifying causes and clusters that contribute to mortality at a global level and making sense of how different regions of the world are affected by specific groups of diseases. We will discuss the design of the visualization by focusing on sub-visualizations that address cause cluster-specific trends and cause-specific trends at a global level and cluster-specific trends for different geographical areas.

First, users need to be able to identify the major points in time (i.e., 1990, 1995, 2000, 2005, 2010). For this we use an instantiation of [Token•Coordinate] to convey the uniqueness of each year across a scaled structure (see Figure 7a). This representation is used to control the three chronology sub-visualizations. The first sub-visualization focuses on cluster-specific mortality. We use [Token•Group] to encode each cause-cluster so that users can identify clusters and the group to which they belong. Clusters are composed of causes with varying prevalence. For example, in 1990, the chronic respiratory diseases cluster consisted of five causes including chronic obstructive pulmonary disease (COPD), asthma, and pneumoconiosis. COPD accounted for over 60% of all the deaths attributed to this cluster. Because we want to convey the distribution of causes that make up a cluster, we use the Cell pattern. As the hierarchical structure of the cluster is also of importance, we also use the Hierarchy pattern. We instantiate a blending of [Token•Cell•Hierarchy] to convey both the hierarchical structure and proportion of items within each cluster. Figure 7b depicts the cardiovascular diseases and HIV/AIDS & tuberculosis clusters for 1990 and 2010. One notable observation is that from 1990 to 2010 the proportion of deaths from tuberculosis (i.e., green rectangle in HIV/AIDS & tuberculosis cluster) decreased.

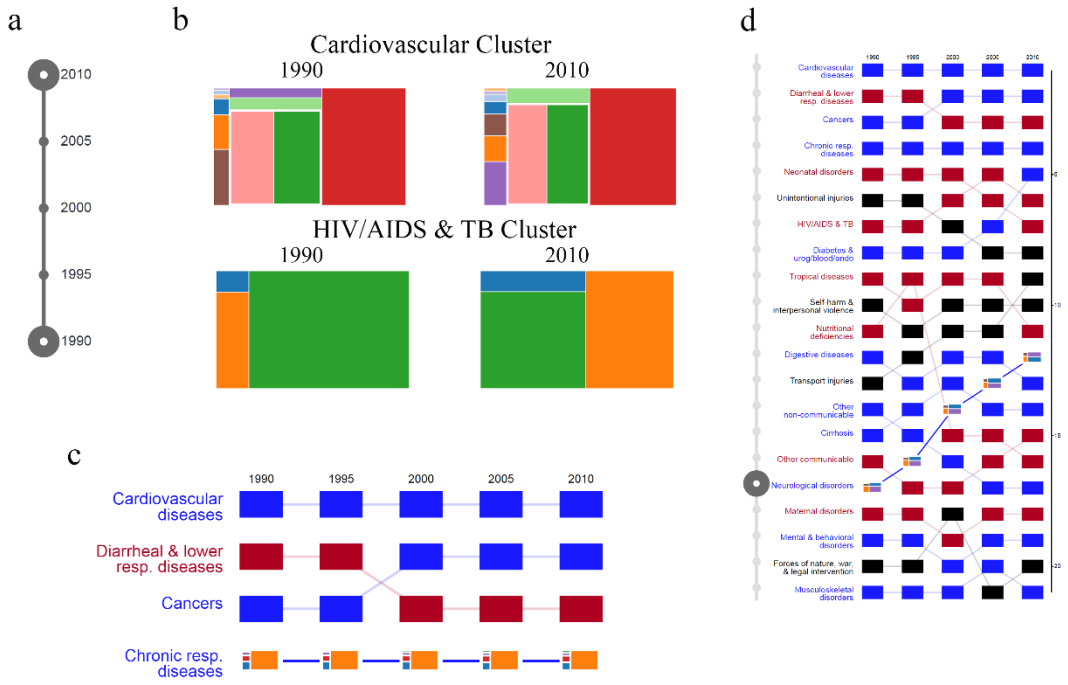



**Figure 3-7: (a) [Token•Coordinate]-based representation for years (b) hierarchical visualization for cardiovascular diseases and HIV/AIDS & tuberculosis clusters (c) top portion of cluster-specific mortality ranking (d) Chronology sub-visualization for cause cluster-specific mortality**

To represent a temporal change for each cluster, we utilize the Link pattern, instantiated as a colored line between represented clusters. Figure 7c shows the top four clusters with the hierarchical structure of chronic respiratory diseases exposed. Clusters are ranked based on their percentage of the overall global mortality. To convey the ranking, we use the Coordinate pattern, instantiated using a 2D coordinate system. The horizontal dimension represents years, and the vertical dimension represents rank from 1 – 21 with the axis reversed so that one is at the top. Each cluster sub-visualization is positioned using this frame of reference. The resulting [Token•Hierarchy•Cell•Link•Coordinate•Group]-based sub-visualization, depicted in Figure 7d, conveys cluster-specific mortality ranking at a global level. One observation is that the top four clusters have remained the same with a change in position between cancers and diarrheal and lower respiratory diseases in 2000. Upon closer examination of neurological disorders, one notices that it has risen from position 17 to 12, thus accounting for more deaths. Furthermore, within the neurological cluster, the proportion

of deaths from Alzheimer's disease (i.e., light blue rectangle) is significant and has grown since 1990.

The second sub-visualization supports the exploration of temporal trends for cause-specific mortality rates within each cluster. Similar to the previous sub-visualization's design, we use the Token, Group, Link, and Coordinate patterns to organize data items. [Token•Group] is instantiated as colored circles for each cause at a point in time. The temporal relationship for a cause is encoded using a curved line (i.e., Link pattern). A 2D coordinate system where the horizontal dimension is for years and the vertical dimension is for proportion is utilized. A portion of the resulting [Link•Coordinate•Token•Group]-based visualization for the unintentional injuries cluster is shown in Figure 8a. As depicted, the percentage of deaths by drowning has decreased, while the percentage of deaths from falls increased. The colors used to encode each cause are the same ones used in the first sub-visualization, thus allowing users to make a connection between the visualizations.

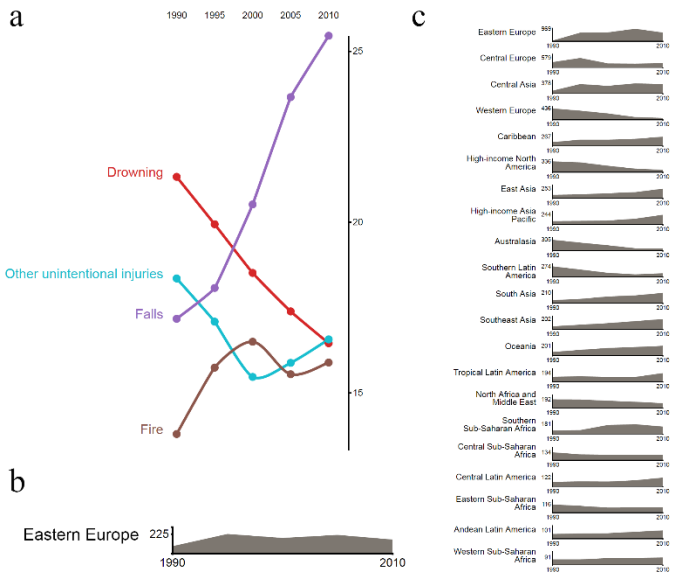



**Figure 3-8: (a) Portion of chronology sub-visualization for cause proportion (b) Area chart for Eastern Europe for the cancer cluster (c) Region cluster-specific mortality for cardiovascular disease cluster**

The first two sub-visualizations support making sense of cluster- and cause-specific mortality at a global level. The final sub-visualization for chronology focuses on

temporal trends for different geographical regions. For each country cluster, we want to communicate continuous mortality patterns, and so we select the Fusion pattern. By using the Fusion pattern instead of the Token pattern, users will be able to understand overall trends for each region as opposed to distinct values for each year. To facilitate comparison, we use the Coordinate pattern and blend it with the Fusion pattern to derive an area chart as shown in Figure 8b. The representation also includes instantiations of the Token pattern for country cluster names and values on the x- and y-axes. The [Fusion•Coordinate•Token]-based area chart depicted in Figure 8b shows the mortality rate for cancers for eastern Europe. To facilitate comparison of death rates for clusters of geographical areas, we use the Coordinate and List Patterns. This blending is instantiated by ordering the area charts by their 2010 mortality rate in descending order. The resulting [Fusion•Coordinate•Token•List]-based visualization for cardiovascular diseases from 1990-2010 is shown in Figure 8c. Each area chart's y-axis is independent of the others. As designers, we choose to use separate scales so that users can identify trends for specific regions. If the same scales were used for all country clusters, the mortality rates for southern sub-Saharan Africa would appear constant because the difference between 146 and 181 is hard to perceive when put on a scale between 0 and 969 (i.e., the highest mortality rate for eastern Europe).

As each of the sub-visualizations supports one part of the overall task, we organize them in a way that conveys separation of information as well as membership. For this, we used the Cell pattern instantiated as compartments in which each sub-visualization is placed. The overall [Fusion•Coordinate•Token•Hierarchy•Cell•Link•Group]-based visualization, shown in Figure 9, facilitates the exploration of mortality from a temporal perspective. In its current configuration, users can make sense of mortality trends from 1990-2005. One observation is that at a global level deaths from nutritional deficiencies have dropped from a high position of 9 in 1995 to 15 in 2005. When the HIV/AIDS & tuberculosis cluster is selected, one can notice that tuberculosis has decreased significantly in proportion while HIV/AIDS causes of death have increased. In the last panel, users can observe that HIV/AIDS & tuberculosis mortality rates have increased for the Carribean and the regions in sub-Saharan Africa.

Using the pattern language, we are able to analyze the tasks and select patterns to organize data items. These patterns are then blended to create sub-visualizations which are arranged in a manner such that users can perform multiple co-related cognitive tasks. It is worth mentioning that each sub-visualization instantiates the Coordinate pattern in a different manner. This flexibility that the pattern language provides supports designer creativity, while allowing designers to structure the design process.

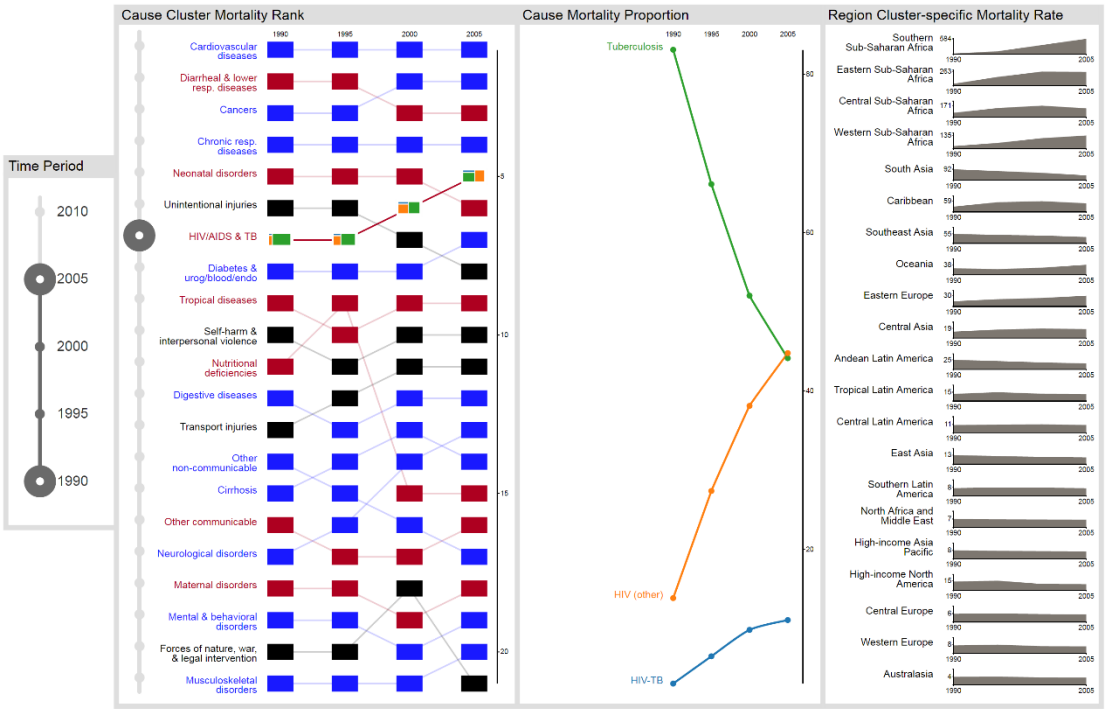



**Figure 3-9: Overall [Fusion•Coordinate•Token•Hierarchy•Cell•Link•Group]-based visualization for chronology**

## 3.4.3 Geography Visualization

The next visualization we present facilitates the exploration of mortality from a geographic perspective. We utilize data that includes cause-specific and risk-specific death rates. The data is aggregated at various levels of granularity. For cause of death, the levels are individual causes and their clusters; for risk factors, the levels are risk factors and their clusters; for geography the levels are countries, clusters of countries, and global. We also use data that quantifies the burden of disease attributable to each risk for each cause of death, thus focusing on the relationship between causes and risks. One starting

point for making sense of the geographic distribution of death is examining the relationship between causes and risk factors at a global level. By supporting this task, users can identify causes and risks of interest and then choose to explore their impact on different geographical regions of the world. As the number of causes and risk factors is large, this approach can help guide exploration. Other tasks include assessing the variability of mortality across the globe for specific causes and risks, exploring the prevalent causes of death and risks for each country cluster, and comparing the distribution of cause-specific and risk-specific mortality across countries.

For users to learn about causes and risks that contribute to death at a global level, they will need to perform a series of tasks. These tasks include identifying the major entities (i.e., causes and risks), exploring the hierarchical structure of entities, ranking entities based on mortality rates, and assessing relationships between entities at different levels of granularity. As the relationship between causes and risks can be explored from a cause or risk-centric point of view, we opt to design a visualization that can be configured to support both options. The organization of data items is similar for both modes, and so we will discuss the design of the cause-centric visualization and provide a screenshot of the risk-centric visualization.

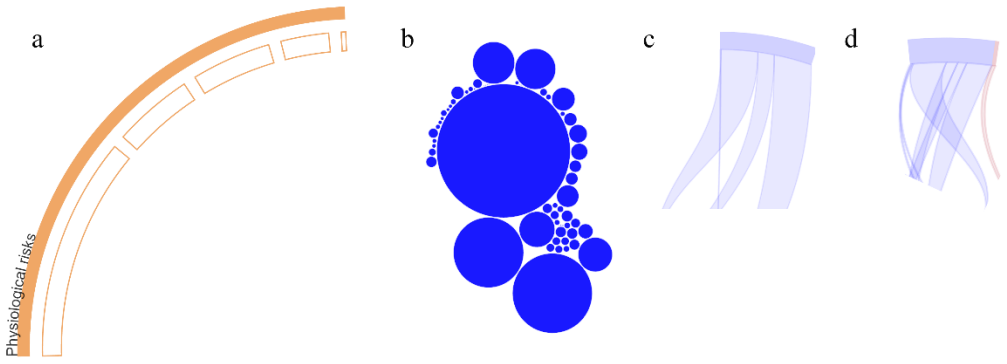



**Figure 3-10: (a) Hierarchical structure of the physiological risk cluster (b) Representation of non-communicable disease group by individual causes (c) Diet low in fruit risk visual element (d) High fasting plasma glucose visual element**

As with previous visualizations, we use a [Token•Group]-based representation to support the identification of each risk, cluster, and the group to which it belongs. We use colored arcs, where size encodes mortality rate, to instantiate this blending (see Figure 10a).

Because we want to convey the hierarchical structure and severity of risk factors we use [List•Hierarchy]. The outer arcs represent risk clusters, while the inner arcs represent the risk factors. Combined, the two tiers of arcs convey the structure of risk factors and instantiate the Hierarchy pattern. Within each tier, we use the List pattern to organize entities by their mortality rate so that users can rank entities within each cluster. For instance, Figure 10a shows the five risks that make up the physiological risk cluster arranged by mortality rate. Next, we want to convey the mortality of each cause and the group to which it belongs and so we a [Token•Group]-based representation. Where Group is instantiated with color and position and the Token pattern is instantiated as a circle for each cause. Once again, we use size to denote severity. Figure 10b shows the non-communicable disease group. Next, the relationship between risks and causes needs to be encoded. We use the Branch pattern to convey how a risk factor can contribute to multiple causes of death. Figure 10c shows an instantiation of the [Token•Branch]-based representation for the risk factor, a diet low in fruit. The top portion represents mortality attributed to the specific risk for all causes and the lower portion is composed of smaller branches each representing mortality for a specific cause. Color is used to encode the group to which the cause belongs. For instance, Figure 10d shows the instantiation for high fasting plasma glucose; this risk factor is connected to seven causes of death, one of which belongs to the communicable group as indicated by the red link.

Figure 11 shows the resulting sub-visualization when the above elements are combined. The [List•Hierarchy•Token•Group•Branch]-based visualization is a variation of a visualization developed by Vizuly (Vizuly, 2014); one noticeable difference is that the hierarchy of risks is encoded. With this sub-visualization, users can rank and explore the hierarchical makeup of risk factors. For instance, within the behavioral group, smoking is attributed to more deaths than child and maternal undernutrition, and within the smoking cluster, there are two risk factors. Regarding relationships between causes and risk factors, users can explore and notice that communicable diseases (i.e., red circles) are predominately not linked to dietary and physical inactivity risk factors. In this mode, it is challenging to rank causes of death. Figure 12 shows the risk-centric visualization. The arcs are used to encode the hierarchy and prevalence of causes, while the circles encode

risk factors. For each of the cause-clusters, users can explore the constituent causes, their ranking, as well as their relationship to risk factors.

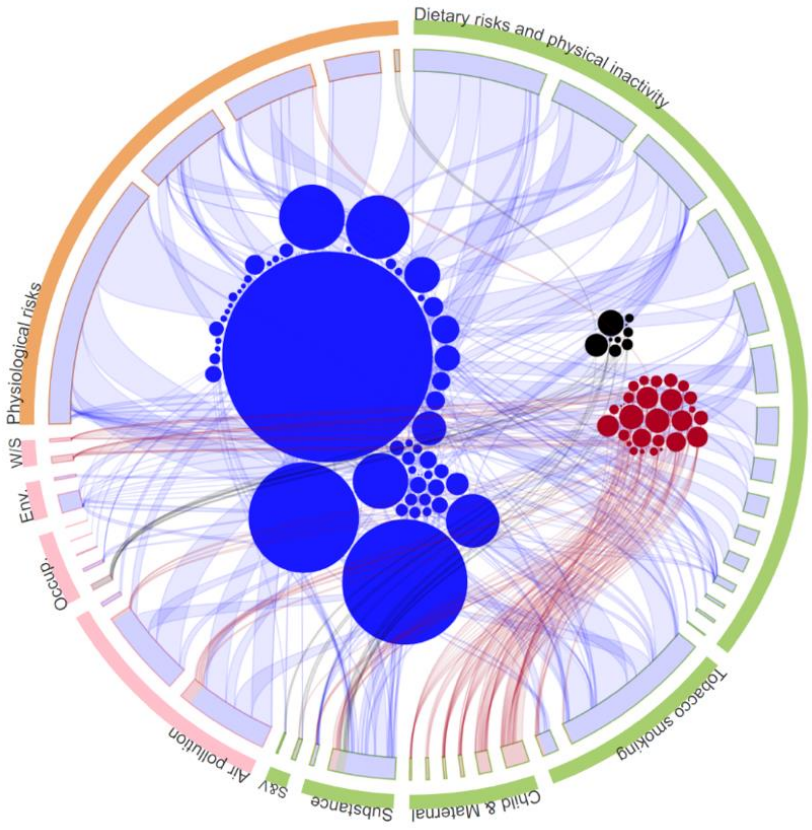



**Figure 3-11: Geography sub-visualization for cause-risk relationships at a global level from a cause-centric point of view**

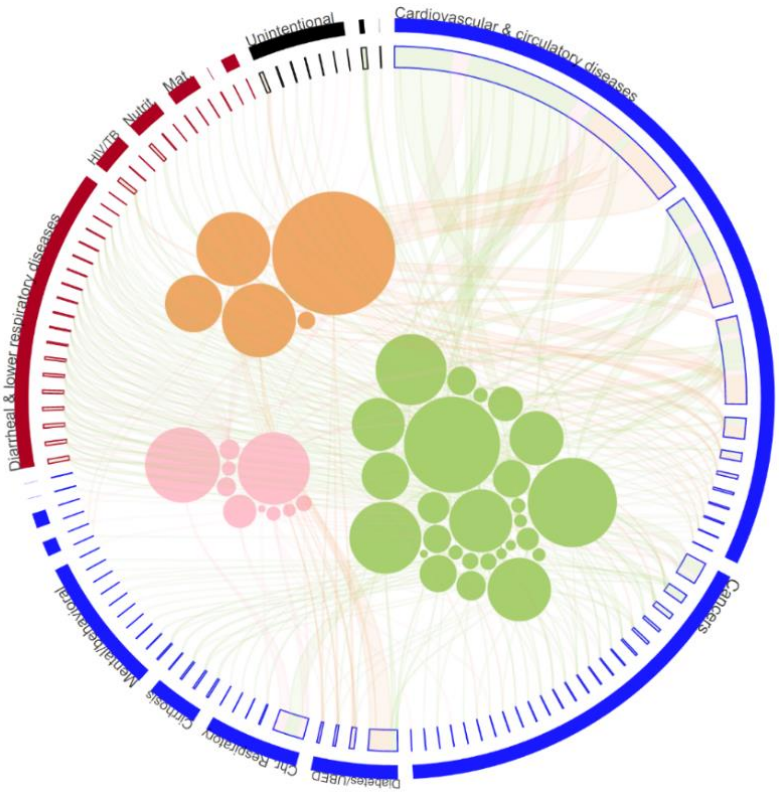



**Figure 3-12: Geography sub-visualization for cause-risk relationships at a global level from a risk-centric point of view**

The third sub-visualization facilitates the exploration of cause- and risk-specific mortality for different regions of the world. When designing the demography and chronology visualizations, we represented geographical entities without encoding their spatial dimensions. As the goal here is to present mortality through the lens of geography, we organize geographical entities by their spatial attributes. To do this, we use the Area pattern and instantiate it with a map demarcated at the level of country clusters. Making sense of mortality across 187 countries may seem tedious, and so we first present the data at the level of the country clusters and then provide users the ability to compare mortality rates within a cluster. As users need to investigate the variability of death across the globe we use the Spectrum Pattern. We use color saturation to instantiate this pattern, the darker the color, the higher the mortality rate. A [Token•Spectrum]-based legend is also created to facilitate comprehension of different saturation values. The resulting [Spectrum•Area•Token]-based visualization is placed between the first two sub-visualizations as shown in Figure 13. As depicted, users can make sense of the global

distribution of mortality, as well as, explore how selected risks and/or causes affect different country clusters. In Figure 13, the impact of chronic obstructive pulmonary disease is depicted.

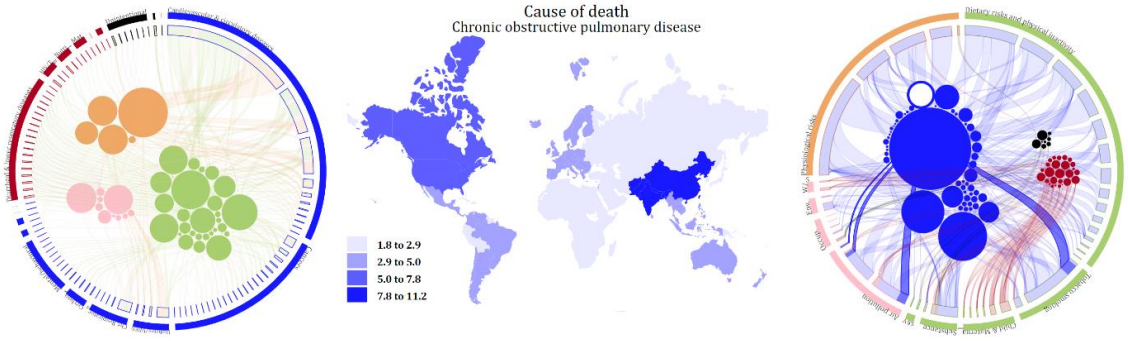



**Figure 3-13: First three geography sub-visualizations, the impact of chronic obstructive pulmonary disease is depicted in the map-based visualization**

The next task we facilitate is exploring relationships between the cause and risk clusters for a particular country cluster. The sub-visualization uses data at the level of clusters (i.e., country, risk, and cause). To support this task, we use the Token, Group, and Branch patterns. The Token pattern is instantiated with a rectangle and discrete name (e.g., neonatal disorders), while color is used to instantiate the Group pattern. The size of the rectangle encodes death rate. Since we want to show how risks contribute to different causes of death, we use the Branch pattern. We instantiate this pattern as links that emerge from risk clusters and go to cause-clusters. Figure 14a depicts the [Branch•Token•Group]-based visualization that shows the prevalent relationships for the central Europe country cluster. After gaining an understanding of cluster relationships, users may want to make sense of mortality at the level of cause, risk, and country. To support comparison at this lower level of granularity, we design the fifth sub-visualization. To convey each country's risk or cause mortality, we use [Spectrum•Token] depicted as colored squares. These data items are organized using [Coordinate•List] where the horizontal axis is used for countries and the vertical axis is used for causes or risks. The resulting [Spectrum•Token•Coordinate•List]-based visualization shown in Figure 14b depicts the distribution of mortality for causes in the cardiovascular diseases cluster. We use the Cell Pattern instantiated as a boundary

structure to blend the two sub-visualizations (Figure 14a and b). Figure 14c shows the resulting [Spectrum•Coordinate•List•Branch•Token•Group•Cell]-based visualization for central Europe.

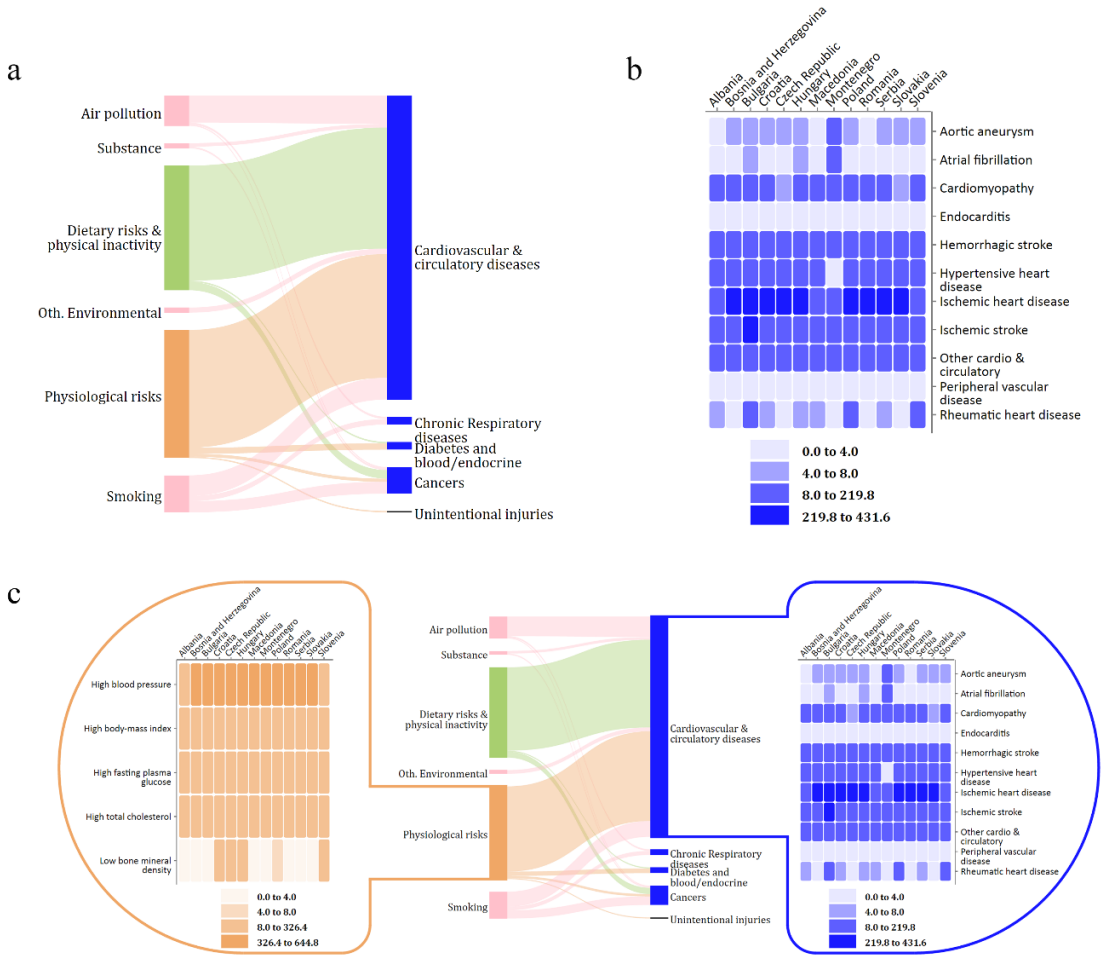



**Figure 3-14 (a) Cause-risk cluster level relationships sub-visualization (b) Visualization of cardiovascular diseases for central European countries (c) Fourth major sub-visualization for geography which combines cause-risk cluster level relationships and risk/cause specific distribution for central Europe**

Figure 14c depicts the cause-risk relationship for one country cluster; but it is important that users be able to explore the relationships for other geographical regions as well. While using a map is beneficial for exploration, it pre-supposes that individuals know what the country clusters are and where they are located. To address this assumption, we instantiate [Cell•Token] with arcs and text that encode the 21 country clusters by name (see Figure 15). Itemizing each country cluster is beneficial for two reasons. First, it

provides a clear way for users to identify geographical areas regardless of their prior background, second, it helps users learn where geographical clusters are located when linked to the map in Figure 13. We use the Cell pattern to organize the sub-visualizations as shown in Figure 15. As depicted, central sub-Saharan Africa has been selected, and the cardiovascular disease cluster and physiological risk cluster have been expanded.

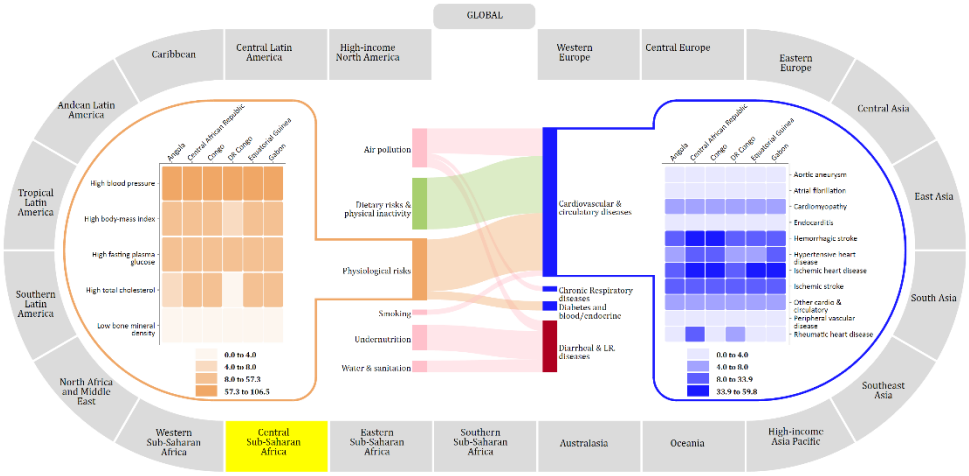



**Figure 3-15: Geography sub-visualization for a country cluster**

We combine the sub-visualizations for global, country cluster, and country-level mortality as depicted in Figure 16. The [Branch•Token•Coordinate•List•Group•Spectrum•Area•Cell]-based visualization supports understanding the geographical distribution of mortality at multiple levels of granularity. As illustrated, the geographical distribution of deaths attributed to a diet high in sodium is presented, as well as the relationships between causes and risk factors for the central sub-Saharan African cluster.

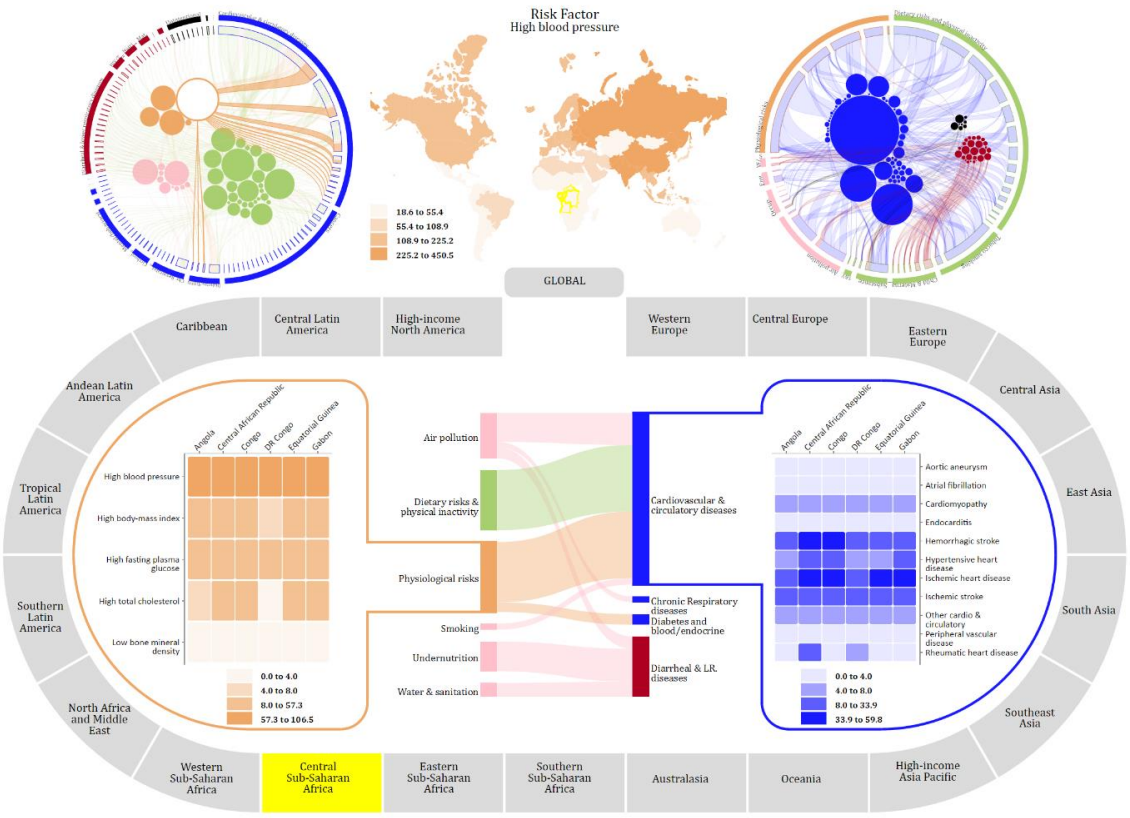



**Figure 3-16: Overall [Branch•Token•Coordinate•List•Group•Spectrum•Area•Cell]-based visualization for geography**

### 3.4.4 Overview Visualization

The last visualization provides a high-level summary of mortality trends for different age groups and geographical regions, at various points in time. With this visualization, users can assess overall and cause-specific mortality, the burden of death attributed to each risk factor, as well as relationships that may exist between specific causes and risk factors. In addition, users need to be able to understand the major data item groups and how they relate to each other at a high-level. As the number of data items is sizable, it is beneficial to provide landmarks that will support exploration. To provide an overview of the burden of disease, we utilize datasets aggregated at the highest level of granularity for geography (i.e., seven geographical regions) and demography (i.e., five main age groups) in 1990, 1995, 2000, 2005, and 2010. Users' sensemaking tasks include: identifying different age groups, geographical regions, and years; assessing the distribution of mortality from each perspective; exploring the relationship between cause and risk clusters; and

understanding the structure of clusters. We will describe the overall visualization by discussing sub-visualizations that support the above four tasks.

To support the identification of age groups, regions, and years we need to map data items in a manner that conveys uniqueness, and so we use the Token pattern. Each data item is encoded as a rectangle with a textual label as shown in Figure 17a. We use color to distinguish data items, shades of purple are used for years, shades of brown for age groups, and seven unique colors for geographical regions. After users can identify and select age groups, years, and regions of interest to explore, it helps to understand how the selected items contribute to the burden of disease. For instance, users may want to determine what age group contributes the most to mortality in Asia. To support this task of assessing proportion, we design a second sub-visualization. Because our goal is to allow users to compare data items that have similar features we use the Stack pattern. The [Stack•Token]-based visualization in Figure 17b shows the percentage of overall deaths for each age group. The size of each rectangle represents the percentage of total mortality for each age group. The rectangles are stacked in descending order with the highest proportion at the bottom. By organizing data items by position, users can compare items without relying solely on the size of the rectangle. When two items have the same proportion, we use a black dashed line between them to denote equality. Since we want users to be able to contrast patterns at different levels, we create instantiations of the [Stack•Token]-based representation for mortality, cluster-specific mortality, and cause- and risk-specific mortality. The Group pattern is instantiated as a bounding box. Figure 17c shows the year-related global mortality proportions for all individuals over the age of 5. We have sub-visualizations similar to Figure 17c for demography and geography thus enabling users to explore the proportion of mortality at three different levels for all three perspectives.

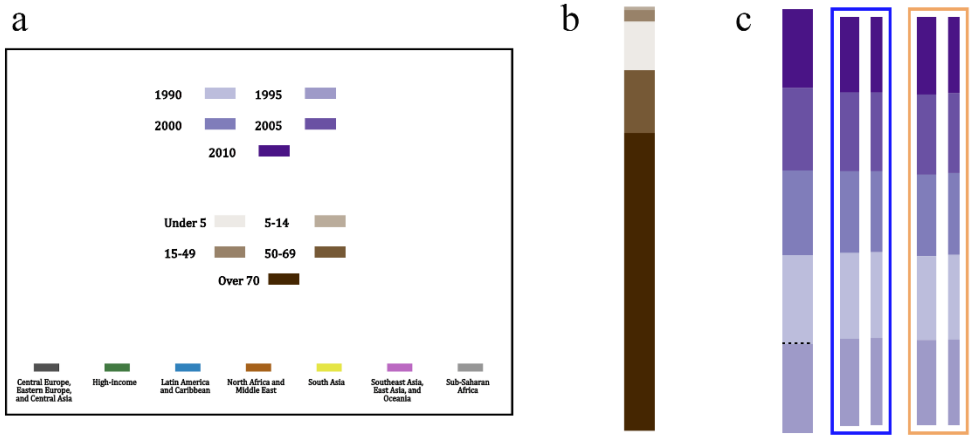



**Figure 3-17: (a) Legends for overview visualization (b) Mortality by age group sub-visualization (c) [Stack•Token]-based representations for year-based mortality**

The third sub-visualization focuses on the relationships that exist between cause and risk clusters. We use an instantiation of [Branch•Token•Group] similar to Figure 14a to convey the group to which each cluster belongs, the relationship between risks and cause-clusters and the uniqueness of each cluster. Figure 18a depicts the [Branch•Token•Group]-based sub-visualization that shows the prevalent cause-risk cluster relationships at a global level in 2010 for all age groups. The last task focuses on understanding the structure of clusters. Because users need to understand the causes that are most prevalent within each cluster, we use a [Token•Cell•Hierarchy]-based visualization similar to the one in Figure 7b. Figure 18b depicts the [Token•Cell•Hierarchy]-based sub-visualization for the physiological risk cluster. We use a Token-based textual notation to label each rectangle and provide the names of each risk factor. The labeling of each rectangle is in ascending order such that 1 represents the risk or cause that has the largest proportion. Since we want users to understand the hierarchy and burden of disease for each cluster, we use the Cell pattern to blend the [Token•Cell•Hierarchy]-based sub-visualization with the [Branch•Token•Group]-based sub-visualization. The resulting [Branch•Token•Group•Cell•Hierarchy]-based sub-visualization is shown in Figure 18c. By default, the structure of clusters with smaller mortality rates are not shown, but can be explored through interaction.

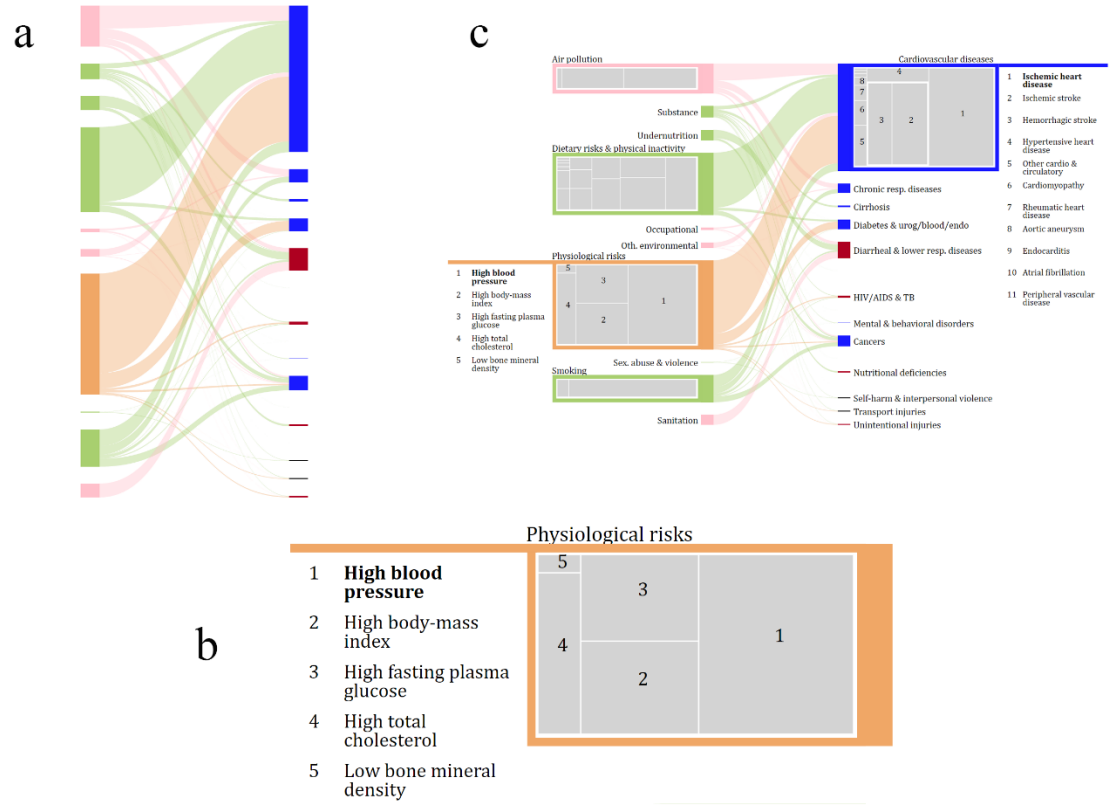



**Figure 3-18: (a) [Branch•Token•Group]-based sub-visualization that shows the prevalent cause-risk cluster relationships at a global level in 2010 for all age groups (b) Physiological risks hierarchy and prevalence sub-visualization (c) Overview sub-visualization for cluster relationships and inter-cluster hierarchy**

Since our goal is to allow users to perform all four tasks with the same visualization, we blend the sub-visualizations using the Cell pattern. The overall overview visualization is shown in Figure 19. This [Branch•Token•Group•Cell•Hierarchy•Stack]-based visualization allows users to explore mortality at three different levels for demography, chronology, and geography. In addition, users can examine the relationship between cause and risk clusters, and make sense of the structure of each cluster.

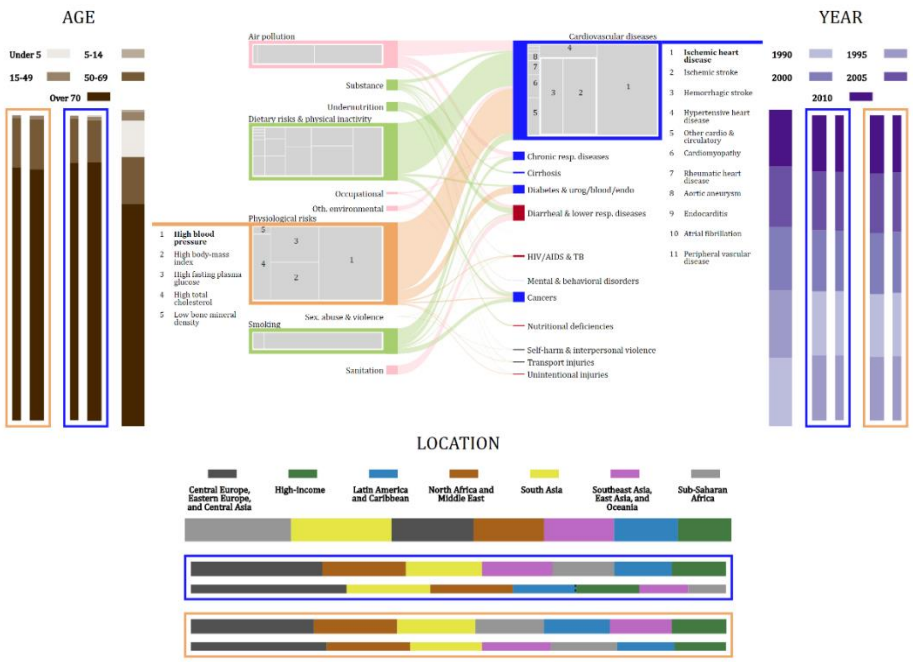



**Figure 3-19: Overall overview [Branch•Token•Group•Cell•Hierarchy•Stack]-based visualization**

## 3.5 Conclusion

The field of healthcare is being inundated with massive amounts of data. In addition to its size, health data is generated at varying rates, collected from heterogeneous sources, and has different levels of veracity. These qualities of health data can negatively impact users' mental processes and increase their cognitive load as they interact with the data. As the ability for big data to revolutionize how healthcare is conducted is contingent on the effective use of this data, there is need for tools that can support users as they engage in a variety of tasks. Interactive visualizations can play a critical role in harnessing the potential of big data. These tools mediate users' discourse with data and, as a result, the manner in which they represent data can either support or impede human-data interaction. When dealing with big data tasks, providing users with the ability to interact with multiple facets of the data is important. Currently, many health visualization tools use simple charts that typically represent only one or two facets of the data thus limiting users' interaction with other facets. Simple charts cannot represent the complexity of big data; they fail to support multifaceted tasks effectively. Therefore, there is a need for

sophisticated visualizations that encode many data elements simultaneously and allow users to perceive patterns and develop insights quickly.

At present, there is a lack of direction about how to create effective visualizations for big data. We contend that the design of visualizations cannot be left to ad hoc processes without the use of frameworks. There is a critical need for support structures, such as conceptual frameworks, that enable the design of visualization tools for big data. This is especially true in the health sector, where previous suites of computational tools have not been well received for a variety of reasons. Frameworks can help designers create elaborate and sophisticated visualizations in a systematic manner with interactive task possibilities at the foreground of design thinking. This is important as human-data interaction is guided by the tasks users seek to complete. Furthermore, conceptual frameworks allow designers to have an awareness of the cognitive implications of design choices, while at the same time facilitating systematic design thinking. Sedig and Parsons have developed a framework which includes a pattern language.

In this paper, we demonstrate how the pattern language can be useful when creating sophisticated visualizations. Through a description of four novel visualizations, we have explicated how the pattern language supports design creativity and flexibility. For instance, the chronology visualization instantiated the coordinate pattern in three different ways to facilitate making sense of mortality at different levels of granularity. The demography visualization provided a concrete example of how designers can structure and encode data items to support tasks. As the external organization of information affects how users perform tasks, clear thinking about how to structure multifaceted data is of particular importance. The multifaceted nature of big data tasks requires users to perform inter-related tasks. Elaborate visualizations designed in a systematic fashion can support these tasks. For instance, with the geography visualization, users can understand the cause-risk relationships at a global level, explore the impact of a specific cause in different regions of the world, and understand how a specific risk factor impacts countries in a region. In conclusion, if we are to support complex health-related tasks effectively, our design thinking needs to be research-based and systematic, this facilitating the

development of visualizations that model the depth and multifaceted intricacies of big data.

## 3.6 Limitations

The work we have presented is part of a larger research plan aimed at developing tools to make sense of big health data. The visualizations we developed use reputable data as opposed to the full spectrum of data collected by local and international organizations. As a result, we did not address issues related to the quality of data. Future work should include the incorporation of other sources and types of data, including real-time data. In this paper, we have focused on the visual representation of data, but the manner in which the tool provides users with control over tasks is another important factor that influences human-data interaction. When dealing with big data, users cannot simply look at the data and understand it; additionally, they must be able to interact with it and change its form as they perform inter-related tasks. As interaction promotes the gradual unfoldment of data within a visualization, it is important to explore how interactions can be incorporated in such tools to support users' tasks better. Furthermore, for the domain to fully embrace sophisticated visualizations for big data, there is a need for studies that evaluate the impact of visualizations to better understand how they improve users' discourse with data.

# Chapter 4

# 4 Discourse with Health Data: Design of Human-Data Interaction

To be submitted to Multimodal Technologies and Interaction Journal.

Please note that the format has been changed to match the format of the dissertation. Figure numbers mentioned herein are relative to the chapter number. For instance, "Figure 1" corresponds to Figure 4-1. Additionally, when the term "paper" or "article" is used, it refers to this particular chapter.

## 4.1 Introduction

The massive influx of data has the potential to revolutionize population health efforts and enhance personalized medicine. Health data can help reduce healthcare costs, support early detection of diseases, improve insurance fraud detection, manage population health, and facilitate identification of epidemics or at-risk groups in society (Gotz & Borland, 2016; Kruse, Goswamy, Raval, & Marawi, 2016; Luo, Wu, Gopukumar, & Zhao, 2016; White, 2014). While health data presents rich opportunities, the health community has historically been slow to leverage the data (White, 2014). This is in part due to the nature of the data. Health data is relatively large, comes from a variety of sources, is generated at different velocities, is largely unstructured, and sometimes is erroneous or incomplete (Kruse et al., 2016; Raghupathi & Raghupathi, 2014; Viceconti, Hunter, & Hose, 2015). These characteristics make it difficult for users to effectively work with the data.

Interactive visualizations, when properly designed, can provide a way to analyze and explore large sets of data. Interactive visualizations can help maintain context during exploration, support the identification of patterns, and facilitate a wide variety of tasks in which individuals engage in (Fisher et al., 2012; Pretorius et al., 2016). When involved in data-driven efforts, the tasks that users perform are mostly non-routine, exploratory, and inter-related (Bikakis & Athens, 2016; Fisher et al., 2012; Katsis et al., 2017; Pike et al., 2009). Users need to be able to interact with data seamlessly to complete these tasks.

They need to be able to ask questions and receive a response. As data is accessible through the visually-perceptible interface of the tool, the ability of users to complete tasks is partially dependent on how effectively the visualization mediates the discourse. This back-and-forth between users and the tool is made possible by interaction.

Through interaction, users can become active participants in the analysis of data. For example, interaction can allow users to gradually retrieve or display data. This progressive unfoldment of data is critical, as encoding only one aspect of the data in a visualization, or encoding too much data, strains the cognitive resources of users (Kaufman, Kannampallil, & Patel, 2015). Allowing users to reveal data gradually within a visualization has been shown to be effective in aiding analysts in exploring and understanding large, multivariate datasets (Torres et al., 2012). To date, much of the interaction available in visualizations allow users to perform simple visual representation manipulations and selection of model choices (Ko et al., 2016; Pretorius et al., 2016). This is unfortunate, as researchers note that in addition to allowing users to control the flow of information, visualizations need to allow users to view data from different perspectives (e.g., changing the visual representation form) or add their inferences (e.g., annotating data items) (Bikakis & Athens, 2016; Heer, 2013). The more ways users can interact with the data, the more involved their discourse is with the data, and the more effective their analysis will be (Che, Safran, & Peng, 2013; Endert et al., 2017; Pike et al., 2009; Tominski, 2015).

Researchers have highlighted the need for a deeper understanding of interaction (Aigner, 2011; Dou et al., 2012; Endert, Chang, North, & Zhou, 2015; Heer, 2013; Murray et al., 2012; Sedig & Parsons, 2013; Tominski, 2015). In the health field, designers seeking to create interactive visualizations are bereft of guidance (Carroll et al., 2014; Folorunso & Ogunseye, 2008; Turner et al., 2008). Because visualizations are often designed for different domains, the research on how to properly design interaction is fragmented across disciplines. There is a need for theoretical structures that can help designers systematically create interactive visualizations. Frameworks that bring together concepts from multiple fields, provide a consistent vocabulary, and have structure while at the same time allowing for designer creativity may be of benefit (Purchase et al., 2008; Sedig

et al., 2013; Thomas & Cook, 2005). Sedig et al. have developed a comprehensive framework concerned with the different aspects of human-data discourse mediated by visualizations tools. In a previous paper, we have demonstrated how designers of health visualizations can use elements of the framework to create sophisticated visualizations (Ola & Sedig, 2016). The *purpose of this paper* is to demonstrate how to design interaction so that users can engage with data in a more meaningful discourse.

To this end, the rest of the paper is organized as follows. Section 2 provides necessary terminological and conceptual background. Section 3 presents elements of the framework relevant to designing interaction. Section 4 details the design process that we developed based on the framework. Section 5 presents three scenarios that highlight how users can utilize the visualizations in an interactive manner to learn about global health trends. Section 6 concludes the paper.

## 4.2 Background

### 4.2.1 Data-driven Tasks

The field of health historically has generated massive amounts of data. As far back as the 1980s, the widening gap between data collection and usage has been discussed (Wurman, 1989). As humans, our ability to solve problems does not solely rely on the collection and storage of data, but in our ability to use the data to complete tasks (Sedig, Parsons, Dittmer, et al., 2012). We conceptualize health tasks as any set of goal-oriented behaviors involving the use of health data. As such, our discussion of health tasks is not limited to one specific task but encompasses a variety of tasks in which individuals engage. This includes, but is not restricted to, health professionals using data to diagnose patients, ascertain the cause of disease, and determine if there is an outbreak, and laypeople using data to understand their treatment options, explore risk factors that relate to a disease, and seek support from an online community.

In the context of interactive visualizations, tasks can be thought of as having three aspects: cognitive, visual, and interactive (Sedig & Parsons, 2016). Cognitive tasks are conscious and deliberate mental processes such as generating hypotheses, comparing them to existing mental structures, and constructing analogies (Parsons & Sedig, 2013b).

Visual tasks are behaviors carried out by our visuoperceptual system as we look at visualizations (Sedig & Parsons, 2016). For instance, consider the scenario in which an individual is using a choropleth map to understand the distribution of HIV/AIDS in South Asia. Some of the visual tasks may include locating Bhutan and perceiving which nation has the highest mortality rate. Interactive tasks require users to manipulate the visual representation. For instance, in the example above, the user may need to *rank* nations based on mortality rate, *identify* the nation with the lowest transmission rate, and *assess* countries to determine those that would benefit most from external aid. In this paper, our discussion centers on how to support interactive tasks.

At a fundamental level, tasks are emergent in nature, co-occurring, and can be performed in an iterative and ill-defined manner (Knauff & Wolf, 2010; Sedig & Parsons, 2013). In many situations completing tasks in a straightforward progression is unlikely and may be impossible (Bikakis & Athens, 2016; Pike et al., 2009). As users interact with data, new questions arise that may change which tasks need to be performed, as well as the order in which they are executed. It is important to allow users to not only complete a single task but engage in a series of tasks in the manner of their choosing (Sedig & Parsons, 2013). In the next section, we briefly discuss how visualizations support users' tasks.

### 4.2.2 Visualization Tools

In this paper, we use the term visualization to refer to computational tools that represent data primarily in a visual format and allow users to manipulate how the data is shown. Visualizations can extend the capacity of individuals to use complete tasks (Sedig & Parsons, 2016; Shneiderman et al., 2013). When using visualizations to mediate a user's discourse with data, there is a partnership in which data processing is shared between the tool and the user (Sedig, Parsons, & Babanski, 2012). For instance, a doctor who needs to diagnose a patient may first observe the patient's symptoms. Next, she may use a visualization to view the summary of the patient's medical history before asking for certain tests to be done. This partnership between the user and the tool allows for the computational strengths of the tool to be used in conjunction with human abilities.

As data is accessible through the visually-perceptible interface of the tool, the user's ability to complete tasks is partially dependent on how effectively the tool encodes data (Dou et al., 2010; J. Zhang, 2001). Even when the visual elements of the tool are properly designed, there still exists a perceptual and cognitive distance between the internal and external realms (Kirsh, 2009; Z. Liu & Stasko, 2010). In other words, a gap exists between the user's internal representations and the tool's external representations. Hence, part of the process involves users coordinating these distinct representational forms. Interaction allows users to harmonize and coordinate their mental representations with the external visual representations (Kirsh, 2009; Z. Liu & Stasko, 2010; Sedig & Parsons, 2013; Ziemkiewicz & Kosara, 2007). In the next section, we discuss the role of interaction.

### 4.2.3 Interaction

When discussing visualization tools, interaction can be conceptualized as the actions users perform and the consequent reactions that occur via the tool's interface (Parsons & Sedig, 2013b). Interaction is critical to human-data discourse as it allows users to engage in the process of testing assertions, assumptions, and hypotheses (Endert, North, Chang, & Zhou, 2014). The ability of users to pose questions and get answers from the data is made possible by interaction (Pike et al., 2009). Also, interaction can strengthen the partnership between users and the tool. This is important as humans have an irreplaceable role in the analysis process. For example, even when using advanced statistical techniques, human judgment plays a vital role in outlier analysis tasks (Cao, Lin, Gotz, & Du, 2017). Through interaction, the analysis of data can be user-directed and this is beneficial for several reasons. First, it promotes a seamless flow of data and reduces the cognitive load of users (Kaufman et al., 2015). Second, it allows for the incorporation of the users' knowledge in the analysis process (G. Andrienko et al., 2007; Parsons & Sedig, 2013b; Thomas & Cook, 2005; Tominski, 2015). Furthermore, interaction allows users to adjust features of the tool to suit their cognitive, perceptual, and contextual needs, thus better supporting their exploration experience (Bikakis & Athens, 2016; Kamel Boulos, Viangteeravat, Anyanwu, Ra Nagisetty, & Kuscu, 2011).

If we are to use visualizations to capitalize on the potential of collected health data, interaction must allow users to reach into the dataset and perform various tasks. As human judgment is at the center of successful data analysis the more ways humans can control their discourse with data the better their analysis will be (Fisher et al., 2012; Heer, 2013; Pike et al., 2009). Pike et al. note that "this manipulative aspect is crucial; the more ways users can 'hold' the data; the more insight will accumulate" (Pike et al., 2009). In a survey of visualizations for web-linked data, researchers note that visualizations need to have interactions that provide users with the ability to customize the exploration experience based on their preferences and the problem requirements (Bikakis & Athens, 2016). To this end, users need to be able to view data from different perspectives (i.e., changing the visual representation form), select latent data (i.e., filtering or drilling into the data), or add their inferences (i.e., annotating data items).

There is a need for tools that allows humans to have more control in the analysis and exploration of data (Bikakis & Athens, 2016; Elmqvist et al., 2011; Endert et al., 2015; Fisher et al., 2012; Greitzer, Noonan, & Franklin, 2011; Ko et al., 2016). To date, much of the interactive tasks that are supported by visualization tools allow users to perform simple visual representation manipulations (e.g., panning and zooming) and selection of model choices (e.g., selecting Naïve-Bayes or Support Vector Machine technique). In a recent survey of biomedical visualizations, it was observed that tools typically offer interactions like rotating and zooming, but provide limited support for querying and other more advanced interactive tasks (Pretorius et al., 2016). Ko et al. conducted a survey of visualizations for financial data and noted that most visualizations failed to support tasks such as exploring, annotating, and linking (Ko et al., 2016). In a survey of malware visualizations tools, it was noted that most tools did not have interactions that allowed users to incorporate their knowledge sufficiently in the analysis process (Wagner et al., 2015). Creating visualizations that effectively support users' interaction with data is not a trivial task. The process requires a proper understanding of interaction. In the next section, we present conceptual constructs that can help systematize the design of interaction.

## 4.3 Elements of Theoretical Framework

### 4.3.1 Conceptualization of the Human-Data Discourse

In the framework developed by Sedig and Parsons, the human-data discourse mediated by visualization tools is characterized at four levels of granularity: events, actions, tasks, and activities (shown in Figure 1). Events are physical occurrences that users perform on the visualization. Examples include clicking, touching, and swiping. As users complete these events, epistemic actions emerge. Epistemic actions are actions taken to transform the world to facilitate mental information-processing needs (Sedig & Parsons, 2013). In other words, actions alter the visualization in a manner that supports mental processes. Let us consider a situation in which a data analyst needs to assess health trends for individuals over the age of 70. The analyst may choose to *filter* the data; to do this, he may *click* on a visual item or *swipe* the screen to reveal a sub-menu that he then *clicks* on to *filter* the data. Through a combination of events, the analyst can perform the action of filtering. Table 1 includes a subset of the actions identified in the framework (Sedig & Parsons, 2013).

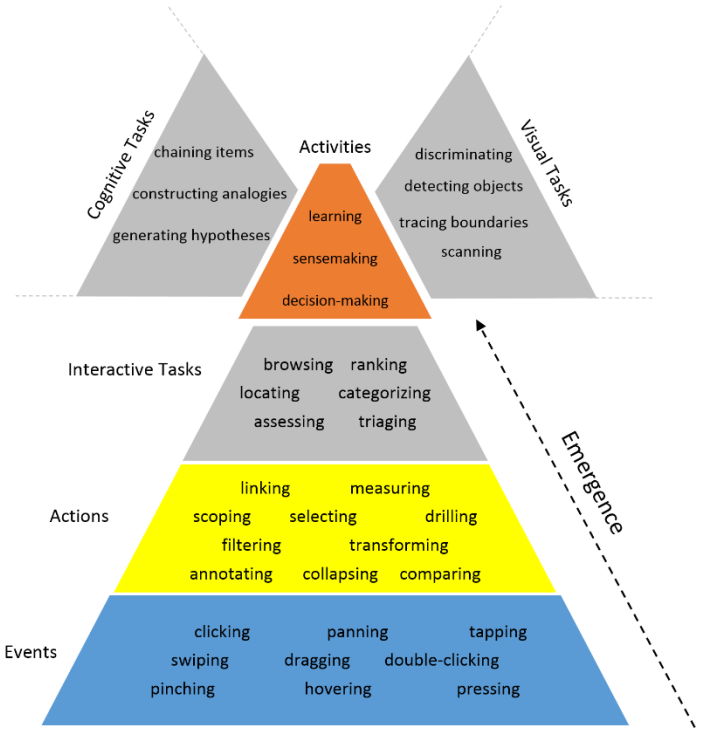



**Figure 4-1: Conceptualization of the human-data discourse**

At the next level is interactive tasks. Interactive tasks are goal-oriented behaviors that emerge from the completion of actions. For example, to complete the task of *triaging* with a visualization, an ER nurse may need to *arrange* patient records based on the severity of symptoms, and then *annotate* each record to assign a priority level. The performance of tasks leads to the emergence of activities. Activities (e.g., decision-making, analytical reasoning, problem solving) are made up of not only interactive tasks, but visual and cognitive tasks as well. For instance, for an epidemiologist to decide that an epidemic of West Nile virus exists, she may have to engage in the cognitive task of testing a hypothesis, the visual task of observing the spread of the disease on the visualization, and the interactive task of categorizing the severity of the disease in each country. While our discussion is focused on interactive tasks, it is worth mentioning that cognitive and visual tasks can also be characterized at multiple levels of granularity.

The conceptualization of the human-data discourse as a multi-leveled phenomenon is of benefit because it helps designers structure the design process (Tominski, 2015). Let us

imagine a doctor who needs to diagnose a patient. How does one create a visualization that supports diagnosis? First, designers can break down the activity into a series of tasks that doctors typically perform with data. Next, for each task, designers can select epistemic actions that facilitate the data-based mental processes of physicians. For instance, for doctors to *assess* the patient, they may need to be able to *filter* out extraneous data, *select* relevant medical information, and *compare* current physiological data to previous data. Once actions have been determined, designers can then decide how to best operationalize them with events.

**Table 4-1: Some of the epistemic actions from (Sedig & Parsons, 2013)**

| Action | Characterization: acting upon visualizations to … |
|---|---|
| Annotating | augment them with additional visual marks and coding schemes, as personal meta-information |
| Arranging | change their ordering, either spatially or temporally |
| Blending | fuse them together such that they become one indivisible, single, new visualization |
| Comparing | determine degree of similarity or difference between them |
| Drilling | bring out, make available, and display interior, deep information |
| Filtering | display a subset of their elements according to certain criteria quantify |
| Navigating | Move on, through, and/or around them |
| Searching | seek out the existence of or locate position of specific items, relationships, or structures |
| Selecting | focus on or choose them, either as an individual or as a group |
| Translating | convert them into alternative informationally- or conceptually-equivalent forms |
| Collapsing/ Expanding | fold in or compact them, or conversely, fold them out or make them diffuse assemble |
| Linking/ Unlinking | establish a relationship or association between them, or conversely, dissociate them and disconnect their relationships |

### 4.3.2 Quality of Interaction

The manner in which interaction is operationalized contributes to the quality of users' discourse with data and thus is an important consideration for designers (Elmqvist et al., 2011; Sedig, Parsons, Liang, & Morey, 2016). For instance, one visualization might allow users to perform a series of actions that changes the subset of data visualized, while another may only allow users to change aesthetic qualities of the visualization such as size and color. While both visualizations are interactive, the difference is in the quality of the interaction. The mere presence of interaction does not equate to effectiveness. The

distinction between interaction and the quality of interaction is important because if the quality of interaction is inadequate, the ability of users to complete tasks will be negatively impacted (Sedig et al., 2013). In our subsequent discussions, we will refer to the quality of interaction as interactivity. As the exploration and analysis of data require the performance of inter-related tasks, it is important to examine interactivity at the level of actions. At this level, interactivity is concerned with how the combination and chaining of individual actions affect and facilitate tasks. Sedig et al. have identified some factors that influence interactivity at the level of actions. In this paper, we focus on two of those factors: complementarity and flexibility (Sedig et al., 2013).

**Complementarity** is concerned with how well actions work with and supplement each other. It is important to provide users with actions that, when performed in conjunction, lead to the emergence of a task. Studies suggest that complementary actions can contribute towards the completion of sensemaking tasks (Groth & Streefkerk, 2006; Sedig et al., 2016; Siirtola & Räihä, 2006; Wang, Wongsuphasawat, Plaisant, & Shneiderman, 2011). For example, in a study on making sense of 4D mathematical structures, the authors note that providing complementary actions can enhance users' discourse with the data (Sedig et al., 2016). Furthermore, complementary actions support flexibility by increasing the ways in which users can complete a specific task (Tominski, 2015). For each task, designers should consider which actions should be used in conjunction. For instance, let us examine the task of triaging data; designers can allow users to filter the data, and select and annotate an encoded data item for further analysis. Thus, the actions of filtering, annotating, and selecting should be operationalized.

**Flexibility** is concerned with the degree to which users can adjust properties of the interface to suit their preferences, characteristics, and goals. This factor is of great importance in health, as past computational tools that adopted a one-size-fits-all approach failed to sufficiently support the diverse user groups and their needs (Berner & Moss, 2005; Turner et al., 2005, 2008). One way of making a visualization flexible is by allowing users to adjust properties of the visualization. Sedig and Parsons have identified essential properties of visualizations that influence cognitive and visual processes (Parsons & Sedig, 2013a). As activities emerge from the completion of cognitive, visual,

and interactive tasks, it is important for us to consider how manipulating the visualization (i.e., performing interactive tasks) facilitates cognitive and visual tasks. Some of the adjustable properties include:

- Appearance: aesthetic features (e.g., color and texture) by which information items are encoded in a visualization
- Complexity: degree to which encoded data items exhibit elaborateness and intricacy in terms of their quantity and interrelationships in a visualization
- Configuration: manner of arrangement, organization, and ordering of data items that are encoded in a visualization
- Density: degree to which data items are encoded compactly in a visualization
- Interiority: degree to which data items are latent and remain hidden below the surface of a visualization, but are potentially accessible and encodable
- Type: form of a visualization in which data items are encoded

In this section, we have highlighted two elements of the framework that can influence interaction design. In the next section, we present a design process based on these elements.

## 4.4 Systematic Design of Interactions

Here we present a process for designing interaction health visualizations. We first explicate the design process and then illustrate the process with an example.

### 4.4.1 Design Process

The process has four main stages: analyzing data and tasks, mapping tasks to actions, linking actions to adjustable properties in the visualization, and operationalizing actions with events. Figure 2 depicts the major stages.

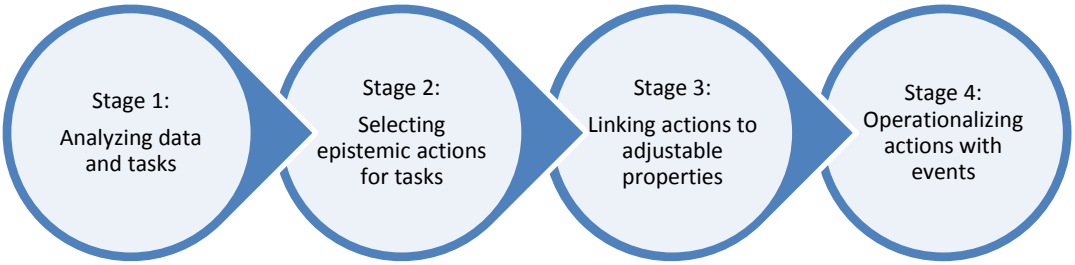



**Figure 4-2: Interaction design process for visualizations**

The first stage is concerned with understanding the data and users' tasks. At this stage, designers need to consider the sources of data, how often the data is updated, as well as the properties, relationships, and typology of data items. Task analysis requires an explication of users' intended activities and tasks. In this stage, we are primarily concerned with determining what the users' goals and intentions are in relation to the data. For more information on how to analyze data and tasks, the interested reader can consult (Munzner, 2014; Sedig & Parsons, 2016; Tominski, 2015; Ward, Grinstein, & Keim, 2015).

The second stage involves selecting actions for each task. In this stage, designers need to consider how users will manipulate the data via the interface of the tool. While it may seem that having every possible action operationalized is a good idea, research indicates that too many actions may result in a high time consumption and increased cognitive demand, thus negatively impacting users ability to complete tasks (Lam, 2008; van Wijk, 2006). At this stage, it is beneficial to itemize the actions that will contribute to the completion of each task and then check to see whether the selected actions are appropriate for the users and the context in which the tool will be used.

The third stage is concerned with linking actions to properties in the visualization that can be adjusted. As previously mentioned, interaction is composed of the action of the user and the reaction that takes place in the tool. The reaction is evident in the change in

certain properties of the visualization. The manipulation of the properties can influence cognitive and perceptual processes, thereby strengthening the bond between the user and the tool (Parsons & Sedig, 2013a). Designers need to determine which properties of the visualization would need to be adjusted so that an action can be performed.

The last stage involves the operationalization of actions with lower-level events. In this stage, designers need to consider how to present each action so that users know it is available and how to activate it (i.e., use it). One consideration at this level is the number of events necessary to complete an action. For instance, if a user needs to drill, does he first click the visual item, drag it into a bin, and then click a button? Or can he just click the item and drilling occurs? Another consideration at this level pertains to when the reaction occurs. Should it occur immediately or be delayed? While in this paper we do not discuss interactivity at the level of events, it is an important aspect that impacts the human-data discourse. For more information on interactivity at the level of events see, (Sedig et al., 2013) and (Sedig, Haworth, & Corridore, 2015).

The design process outlined above is an iterative one. The process can be carried out multiple times for each visualization or sub-visualization that exists in the tool. Typically, the design of interaction happens in conjunction with the design of the visualization, thus providing designers with maximum flexibility. However, even with previously designed static visualizations, designers can engage in the design process to make them interactive.

### 4.4.2 Illustration

We have created a visualization tool to facilitate making sense of the global burden of disease. This tool includes visualizations that allow users to explore the demographical, geographical and chronological distribution of mortality. In this section, we illustrate how the design process helped systematize the design of interaction for the demography visualization.

The Institute for Health Evaluation and Metrics (IHME) aggregated the data used in our tool (Lozano et al., 2012). The datasets include over 12 million records that present estimates of mortality for causes, risk factors, cause-clusters and risk-clusters. We use the

term cluster to refer to an intermediary level of grouping. For example, the physiological risk-cluster includes the following risk factors: high blood pressure, high body-mass index, high fasting plasma glucose, high total cholesterol, and low bone mineral density. The datasets include 235 causes of death that are grouped into 21 cause-clusters which are further aggregated into three main groups: 1) non-communicable, 2) injury-based, and 3) communicable, maternal, neonatal, and nutritional. The datasets also include estimates for 57 risk factors which are grouped into 10 risk-clusters and further categorized into three groups: 1) behavioral, 2) metabolic, and 3) environmental and occupational. From a geographical perspective, mortality rates are aggregated at the level of regions or location clusters. From a geographic standpoint, the datasets include estimates for 187 countries which belong to 21 regions (e.g., eastern Europe, southern sub-Saharan Africa, and tropical Latin America). In terms of age groups, mortality is aggregated into 17 age groups and also at a higher level into five main age groups: 1-4, 5-14, 15-49, 50-69, and over 70. For more information on how the data was collected and aggregated, refer to (Lozano et al., 2012).

The demography visualization shows the distribution of mortality by age group. In its initial configuration, the visualization, shown in Figure 3a, encodes over 800 data items. The risk-clusters and cause-clusters are encoded as arcs. Each visual item also encodes the group to which the cluster belongs. For the cause groups, we use *blue*, *red*, and *black* for non-*communicable*, *communicable*, and *injury* clusters, respectively. For the risk groups, we use light shades of *orange*, *green*, and *pink* for *metabolic*, *behavioral*, and *environmental and occupational* risk groups, respectively. The location clusters are encoded as *grey* bars. The clusters are ranked and arranged according to their mortality rate per 100,000 people. For the cause and risk aspects, the cluster with the highest rank is on top, while location clusters are arranged in descending order from left to right. Figure 3b shows an enlarged portion of the visualization. For certain age groups, not all risk- or cause-clusters contribute to mortality; these clusters are encoded as light grey circles. Through observation, users are able to learn about how mortality affects different age groups. That being said, the visualization is densely packed, and without interaction, it will be difficult for users to perform various tasks.

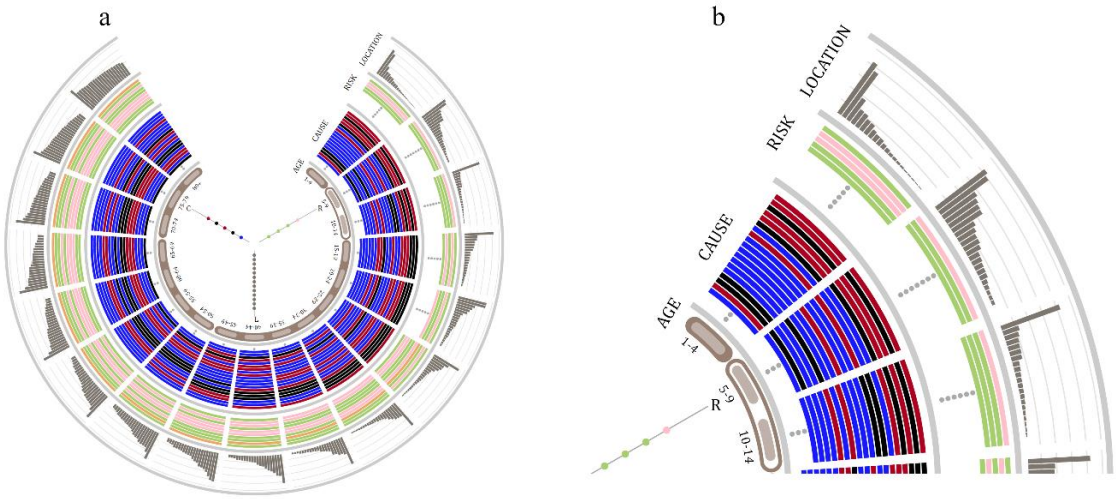



**Figure 4-3: (a) Overall visualization for demography (b) Enlarged partial demography visualization**

Being that the visualization has already been designed, we do not delve into an analysis of the data in this section. For more information on the data, consult the research that focuses on the design of the visualization (Ola & Sedig, 2016). Next, we must analyze the activities and tasks of users. The overall activity that the visualization supports is sensemaking. Sensemaking typically involves users performing a variety of tasks including searching and filtering data; organizing, categorizing, and examining relevant data; developing, proving, and discarding hypotheses; and integrating data into mental models (Parsons & Sedig, 2013b). More specifically, to make sense of the demographic distribution of mortality, users need to be able to identify the ranking of clusters, explore mortality rates and ranking across age groups for different aspects (i.e., cause, risk, or location clusters), assess mortality across geographical regions, explore age-specific trends, examine the distribution of mortality at lower levels of granularity, and investigate relationships that exist across aspects.

To facilitate the completion of these tasks, we need to determine the epistemic actions from which the tasks emerge. For users to be able to identify a cluster's ranking, we will need to enable users to select each cluster that is represented. To facilitate exploration users need to be able to select, search for, and filter clusters. In order to assess mortality for specific age groups, users need to be able to select, filter, and compare data items. In

the visualization, mortality is depicted at the level of clusters. To allow users to investigate the distribution of mortality at lower levels of granularity (e.g., country); users need to be able to retrieve data that is latent in the system. To do this, they will need to be able to drill. To allow users to explore relationships that exist, they need to be able to link and unlink items.

Before we operationalize the itemized actions, it is important to consider if there are additional actions that may work in tandem with the chosen actions. For example, our visualization has many data items encoded in its default configuration and as a result identifying a specific item may become tedious. In order to support identification, it would be beneficial to allow users to reduce the amount of data visualized at a point in time. Collapsing and expanding are two actions which enable users to control the number of items visualized. Allowing users to collapse and expand portions of the visualization can help reduce the burden that high degrees of density may cause. In addition, it may be advantageous to allow users to change the visualization in order to assess the properties of data items. Different types of visualizations have different benefits and limitations for communicating information (Munzner, 2014), and as a result changing the visualization may support users in understanding the various aspects of the data. Translating is the action that allows users to convert visualizations to an alternative informationally- or conceptually- equivalent representation. A similar process of determining complementary actions was carried out for the other tasks and no additional actions were identified. At the end of stage two, the final list of actions is selecting, searching, filtering, drilling, comparing, linking, unlinking, collapsing, expanding, and translating.

Now that we have our list of actions we need to determine which properties of the visualization will be manipulated. In other words, we need to determine the reaction (i.e., how the visualization will change). Selecting is concerned with focusing on an item or group of items, searching is concerned with seeking out specific items or relationships, and filtering is concerned with displaying a subset of elements that meet specific criteria. For these three actions, changing the aesthetic features (e.g., color, texture, saturation) can facilitate and increase the speed of identification of visual items. For example, in Figure 4a, the risk track has been selected and the saturation of the cause and location

visual items has been altered. In Figure 4b, the cause track has been filtered and only the visual items for nutritional deficiencies are emphasized.

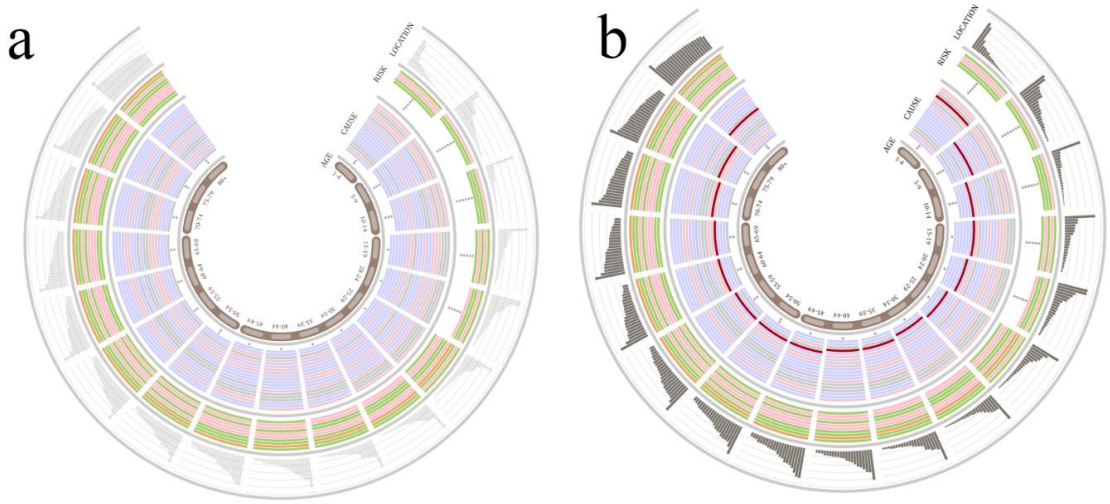



**Figure 4-4: (a) Risk track emphasized (b) Nutritional deficiencies cluster emphasized in the cause track**

Linking allows users to establish a relationship or association between items. To support linking and unlinking, we need to allow users to alter the complexity of the visualization. Complexity is an adjustable property concerned with the quantity and relationships between data items in the visual representation. Research indicates that a significant burden is placed on the mental faculties of users when the complexity is not suitable for the task at hand (Moody, 2007). If users were not able to manipulate the configuration, the sub-visualization would look like Figure 5a. Alternatively, the approach we take is to have the data items shown without any relationships (see Figure 5b) and allow users to select which relationships to explore (see Figure 5c).

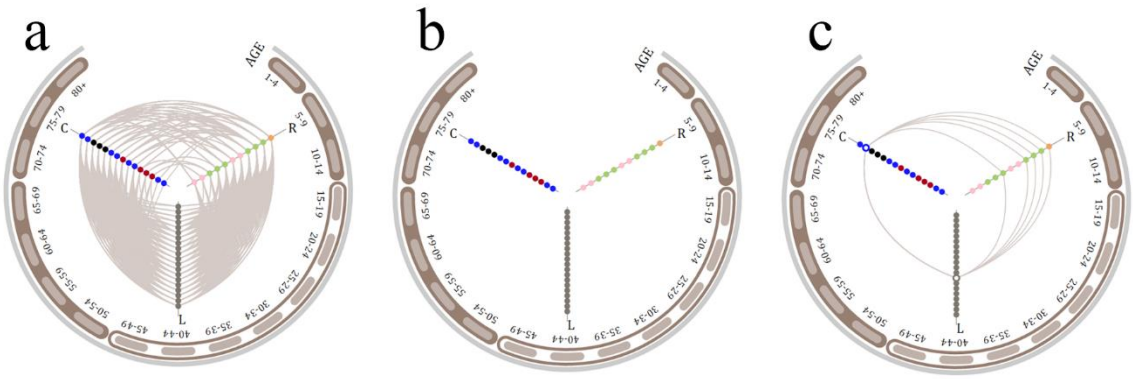



**Figure 4-5: (a-c) Different states of the demography visualization with complexity adjusted**

Comparing is concerned with determining the degree of similarity or difference between items. It is worth mentioning that the word *comparing* can be used to refer to a visual, cognitive, or interactive task. In this paper, we use the word *comparing* in the interaction realm but at the level of actions and not tasks. In other words, comparing refers to the ability of users to select two or more visual items so that the tool will emphasize differences and similarities. In this context, in addition to selecting the item that will be compared, users may need to change the arrangement, organization, or ordering of items (i.e., altering the configuration). Translating, is concerned with converting the visualization into an alternative form. To facilitate this action, we will allow users to change the visualization's type.

Next, we need to determine which property needs to be adjusted so that users can collapse and expand segments of the visualization. The main idea behind collapsing and expanding is changing the amount of data visualized at a specific point in time. Density is concerned with the degree to which items are compactly encoded in a visualization. Research indicates that when the density is too high, perceptual tasks such as locating and extracting pertinent information becomes difficult (Rosenholtz, Li, & Nakano, 2007). That being said, sometimes it is beneficial to have a high level of density as it may allow users to obtain a high-level overview of the data (Hornbæk & Hertzum, 2011). By giving users control over the number of data items encoded, they are able to control the density

to suit their needs. Figure 6(a – c) shows three possible states of the demography visualization when density is varied.

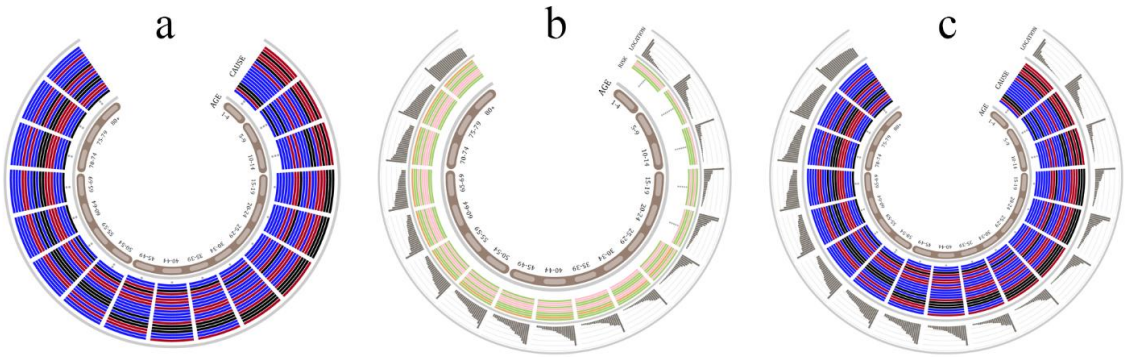



**Figure 4-6: (a) Collapsed location and risk tracks (b) Collapsed cause track (c) Collapsed risk track**

Drilling is concerned with revealing data that is not currently visualized. To do this, we need to allow users to change the degree to which data items are latent and remain hidden in the visualization. In the default arrangement of the visualization, users can obtain an overview of the demographic distribution of mortality, but if they wanted to learn about mortality rates for children living in Eastern European countries, they would need to access data that is not currently visualized. By adjusting the interiority of the visualization, we enable users to access this data, thus controlling the flow of information.

Now that we have determined the adjustable properties, in the last stage we focus on the events the users will perform on the visualization to prompt the change. To ensure consistency in how users interact with the tool we opted to use the clicking event. In addition to clicking visual items, buttons were used to indicate the presence of actions that required multiple events. For example, to filter cause-clusters, users will first need to click on the filter button for cause-clusters, so that different cluster options are presented, and then click on the cluster they wish to focus on.

## 4.5 Scenarios

The three visualizations presented in this section are part of a tool designed to facilitate making sense of the global burden of disease. The first visualization is the demography visualization from the preceding section, while the other two visualizations focus on the geographical and temporal distribution of mortality. For each visualization, we present a scenario to illustrate how through interaction users can engage in a meaningful discourse with data to learn about global health trends.

### 4.5.1 Demography Visualization

Let us consider a college student who is interested in knowing the causes of death for young people. The student may start by learning the rank of different cause-clusters. To focus specifically on causes, he collapses the risk and location parts of the visualization as shown in Figure 7a. At this point, he observes that for individuals between the ages of 15 and 34, injury-related causes of death are highly ranked (i.e., for each age group in the range, black-colored arcs are in the top 5 positions). To explore the demographical distribution, he may choose to filter by cause-cluster. Figure 7b depicts the state of the visualization when the self-harm and interpersonal violence cluster is filtered. He continues this process until he understands the ranking of clusters. One trend he observes is that mortality from neglected tropical diseases and malaria decreases as one gets older. He also observes an opposite trend for cardiovascular and circulatory diseases. To obtain a better understanding of what causes make up the self-harm and interpersonal violence cluster, he drills and then explores the mortality rates, as shown in Figure 7c. At this point, he can assess death rates and may notice that self-harm has a higher mortality rate than assault by firearm for individuals between the ages of 15 and 19. By clicking on each arc, he notices that the same trend applies to the other young adult age groups (i.e., 20 – 24, 25 – 29, 30 – 34). This dispels a previous notion he had that assault by weapon was the primary cause of death on a global scale for young people. At this point, he collapses the cause part, expands the risk part and engages in a similar exploration of risk factors. Next, he chooses to explore mortality across geographical regions. By filtering, he learns that the location cluster with the highest mortality rate for young people is southern sub-Saharan Africa. He focuses specifically on the age group 25 – 29 and

notices that for this age group, in addition to the injury-related cause-clusters, there is a communicable cluster (i.e., red colored arc) that is highly ranked. He drills to retrieve latent data that relates to the causes that make up this cluster. As shown in Figure 7d, tuberculosis (TB) and two types of HIV/AIDS make up this cluster. At this point, he explores the relationship between cause, risk and location clusters. By drilling and linking visual items, he notes that there are five location clusters with a strong relationship between HIV/AIDS & TB and the physiological risk-cluster. The student can continue to interact with the data and learn more about mortality for different age groups.

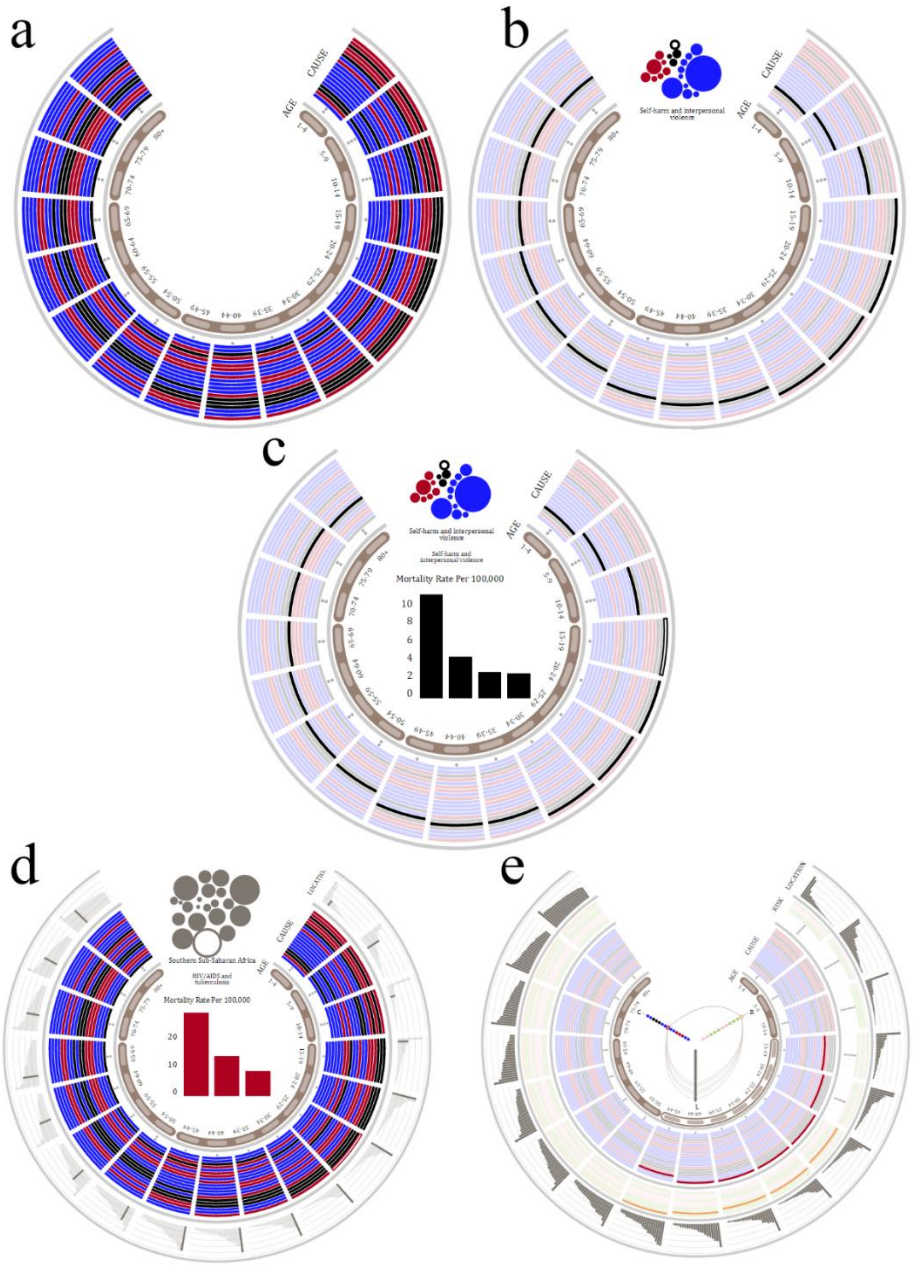



**Figure 4-7: (a-e) Screenshots of the demography visualization**

## 4.5.2 Geography Visualization

The next visualization supports making sense of the burden of disease from a geographic perspective. One of the datasets we use quantifies mortality as attributed to each risk for each cause of death, thus focusing on the relationship between causes and risks.

Additional datasets utilized include the global, regional, and country-level estimates of mortality for causes, cause-clusters, risk-clusters, and risks.

Depicted in Figure 8, the top half of the visualization details relationships between risks and causes at a global level, as well as the geographical distribution of mortality for a selected risk or cause across the 21 regions of the world. The circular sub-visualization on the left shows the relationship between causes and risks. Each cause is encoded as an arc, while each risk factor is encoded as a circle. Risk factors are colored and clustered together to emphasize their grouping. For instance, the largest orange circle represents high blood pressure (HBP), the color orange signifies that HBP is a member of the metabolic group. Causes and their clusters are arranged circularly based on their group and are similarly colored. The links between the arcs and the circles represent the attributed mortality between a cause and risk factor. The circular sub-visualization on the right shows the same information but in a different manner. In this visualization, the causes are encoded as circles while the risk factors are encoded as arcs. For instance, the largest blue circle represents ischemic heart disease, while the longest green arc represents the cluster dietary risks and physical inactivity. One reason why both representations are shown is that they emphasize different aspects of the data and thus may facilitate different tasks. The choropleth map in between them shows the geographical distribution of death for a selected cause, risk, or cluster across regions of the world. In its default configuration, the bottom half of the visualization is comprised of four main elements. The first is the circular track, which is divided into 21 segments each representing a region of the world. The second part is the flow diagram, which is in the center of the track. The flow diagram shows the relationships between cause- and risk-clusters for a specific region. The last two parts are heatmaps which are located on either side of the flow diagram. The heatmaps show the mortality rates for countries for each cause or risk in the selected clusters.

To design interaction, we analyzed the tasks in which users may engage. These tasks include examining mortality or the relationship between causes and risk factors at different levels of granularity, focusing on a region, assessing the variability of mortality, exploring prevalent causes and risks for regions, and discovering the similarities and

differences of the burden of disease between two regions, or between countries in a region. To support these tasks we have operationalized the following actions, selecting, drilling, filtering, searching, arranging, translating, and comparing. Similar to the demography visualization, to activate or initiate an action, users can either click on a visual item or for compound interactions (i.e., interactions that involve multiple steps), click on the appropriate button.

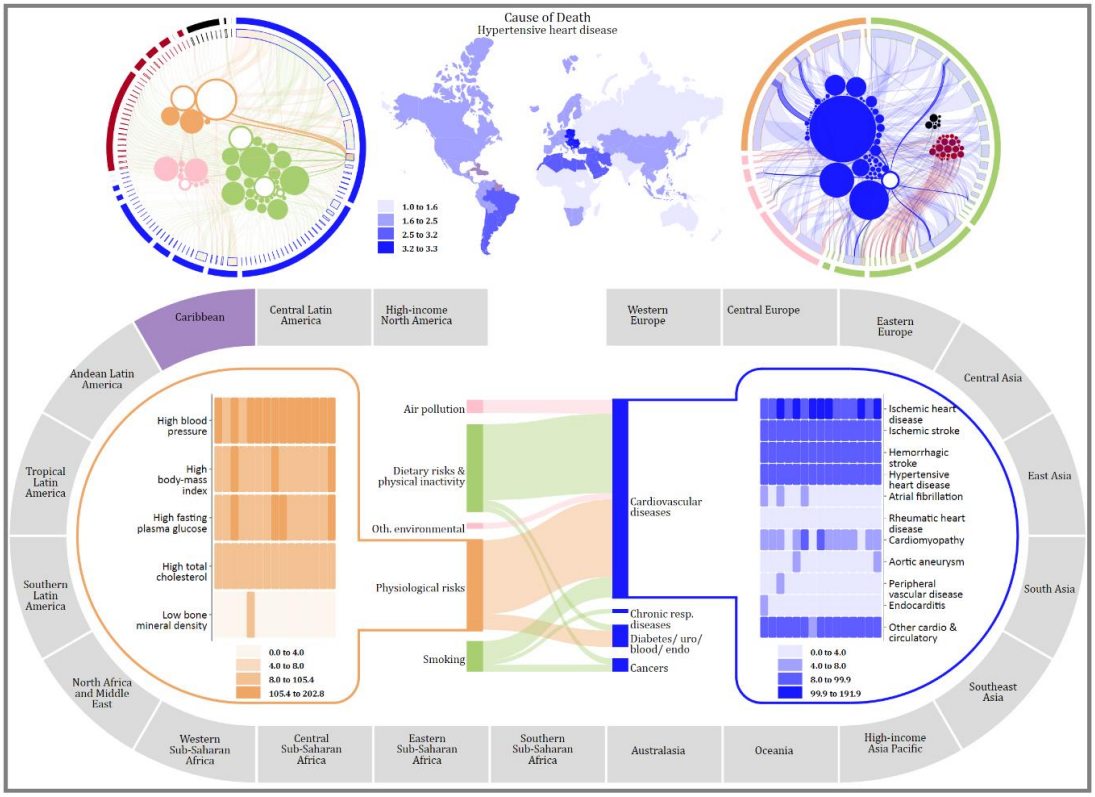



**Figure 4-8: Geography visualization with hypertensive heart disease and the Caribbean selected**

Let us imagine a situation in which an employee of a non-governmental agency needs to develop a proposal that reduces mortality by tackling risk factors in high-risk regions of the world. Using the circular sub-visualizations, she can select a risk factor or cluster to determine the distribution across the regions of the world. An alternative approach may be to search for a specific risk factor using the searching capabilities of the tool. Figure 9a shows the top half of the visualization when alcohol use has been selected.

Unfortunately, while this view allows her to see how alcohol affects different regions, it is not effective in determining whether eastern Europe has a higher rate than east Asia. To address this issue, she switches the representation to a bar chart as shown in Figure 9b and observes that eastern Europe has the highest rate. Next, she drills to obtain a more detailed view of the relationships between causes and risk factors for eastern Europe as shown in the bottom half of Figure 9b. As she also wants to understand how alcohol use affects each country in the region, she drills to display country-level data related to substance abuse. The heatmap in the lower left portion of Figure 9b depicts how both alcohol and drug use affect the nations in eastern Europe. By filtering, she can ascertain the countries in eastern Europe that have a high mortality from alcohol use (i.e., Belarus, Russia, and Ukraine). The employee repeats this process of searching for risk factors, determining the region that has the highest mortality rate, and exploring the distribution at the level of countries until she has a better understanding of which nations can benefit from intervention measures directed at certain risk factors.

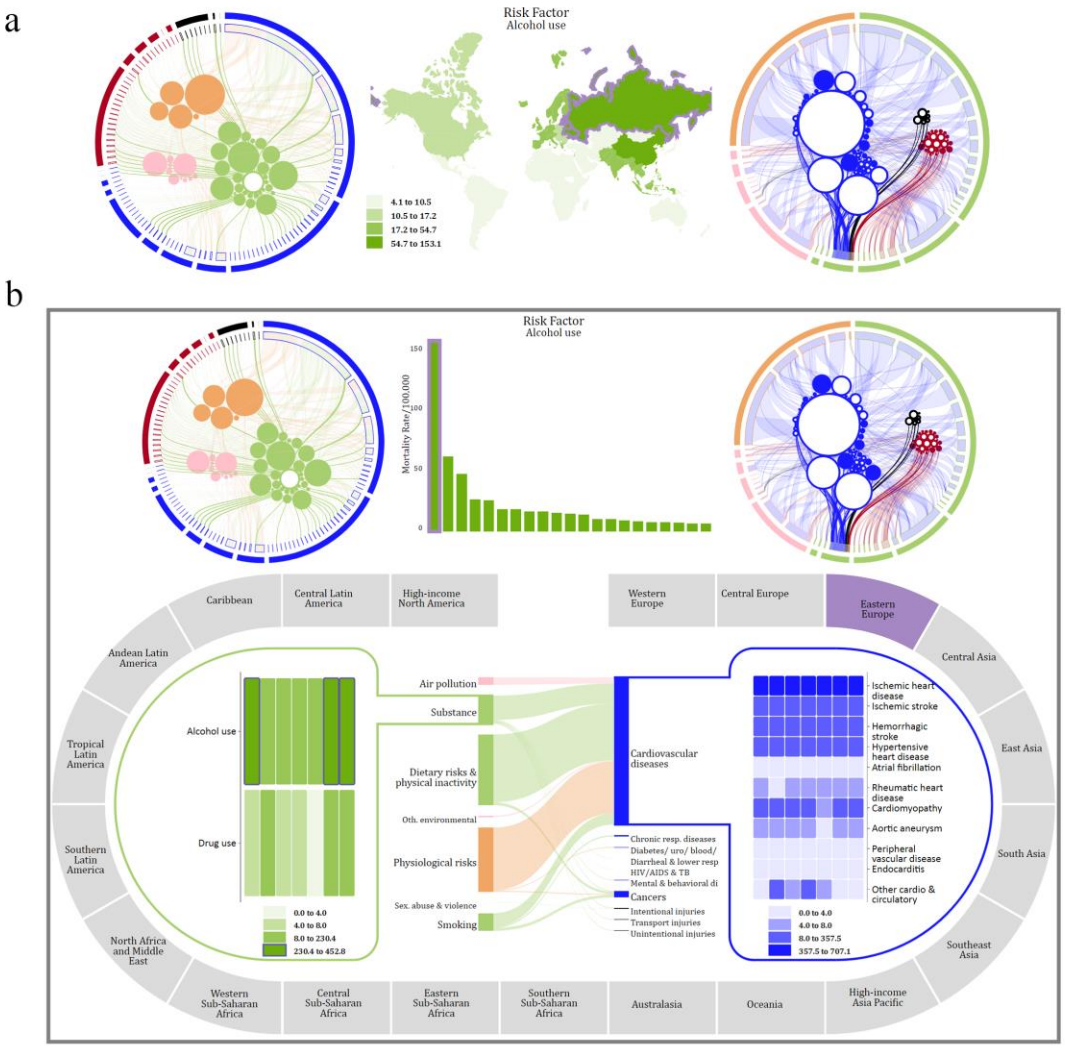
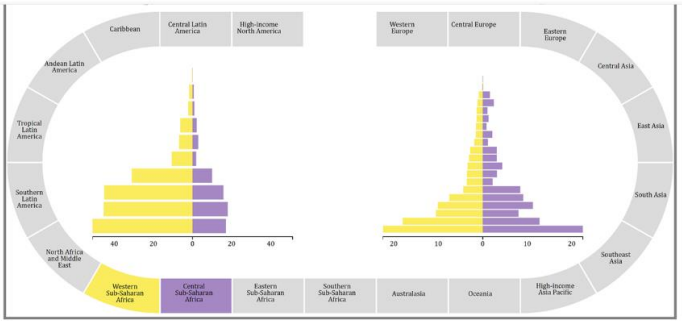



**Figure 4-9: (a-b) Screenshots of geography visualization with alcohol use and eastern Europe selected**

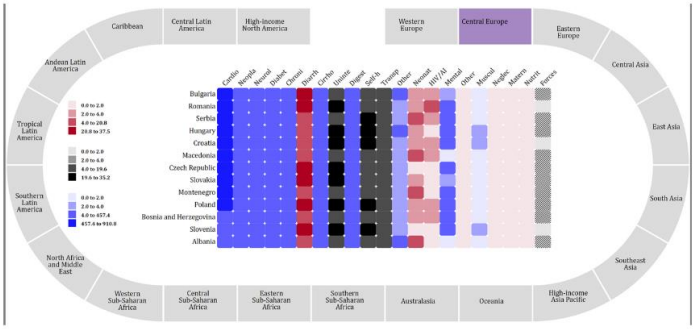
Next, to assess the similarities in mortality between geographical areas, she compares. Figure 10a shows the bottom of the visualization when she is comparing the regions of western and central sub-Saharan Africa. Next, she decides to contrast the mortality rates for cause-clusters across countries in central Europe. To identify the cause-cluster that has the highest mortality rate for Bulgaria, she can choose to re-arrange the heatmap as shown in Figure 10b. An alternate and time-consuming approach would be to select each one of the clusters, and mentally keep track of the cluster with the highest mortality rate.

Next, she re-arranges the heatmap by cause-cluster (as opposed to country) because she is interested in determining which country has the highest rate of death from diarrheal diseases. While she may be deviating slightly from her original task of understanding the impact certain risk factors have, the interactive nature of the tool supports this divergence. Next, she chooses to investigate on a global level which regions are most affected by diarrheal diseases and so she returns to the top half of the visualization and selects that cluster (top of Figure 10c). At this point, she notices that diarrheal diseases significantly impact regions in Africa and that Oceania is also impacted. She decides to take a break and will pick up her exploration by examining the risk factors that significantly contribute to death in countries in Oceania. In this scenario, the professional was able to navigate between data aggregated at different levels. She started with exploring the cause-risk relationship at a global level, then moved on to exploring how risk factors affect different regions of the world, and ended her session by learning about the impact of certain diseases for specific nations.

a




b




c

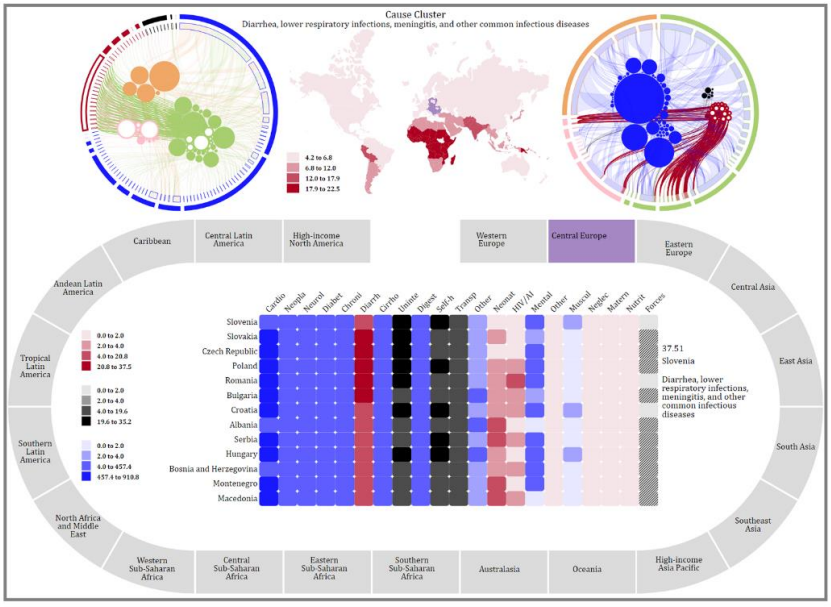



**Figure 4-10: (a-c) Screenshots of geography visualization that emphasize comparison**

### 4.5.3 Chronology Visualization

The last visualization we present facilitates the understanding of temporal trends of mortality. The data utilized includes estimates of mortality at regional and global levels for each cause and cause-cluster. The estimates are at five different points in time: 1990, 1995, 2000, 2005, and 2010. The visualization has three main parts (see Figure 11). The first part (i.e., the left panel) presents the ranking for cause-clusters at a global level. Each rectangle represents a cause-cluster for a specific year. We use color to represent the group to which each cluster belongs, and the position of the rectangle represents the cluster's rank. The scale on the left side shows the rank values from 1 to 21, with 1 representing the cluster with the highest mortality rate. For instance, one can notice that the cardiovascular diseases cluster is consistently ranked number 1, while the nutritional deficiencies cluster starts at position 11, goes up to position 9 and drops to position 16. Embedded within each rectangle is the hierarchical and proportional make-up of the cluster. Each cluster is comprised of several causes, each with varying prevalence. For example, the HIV/AIDS & TB cluster is made up of three causes: Tuberculosis, HIV from TB, and HIV diseases resulting in other unspecified diseases. In 1990 tuberculosis accounted for over 80% of the deaths in this cluster, but by 2010, tuberculosis accounted for less than 50%. The proportion of each cause for a cluster is also depicted in the second panel. This sub-visualization uses a multi-line chart to depict the proportion of mortality for causes.

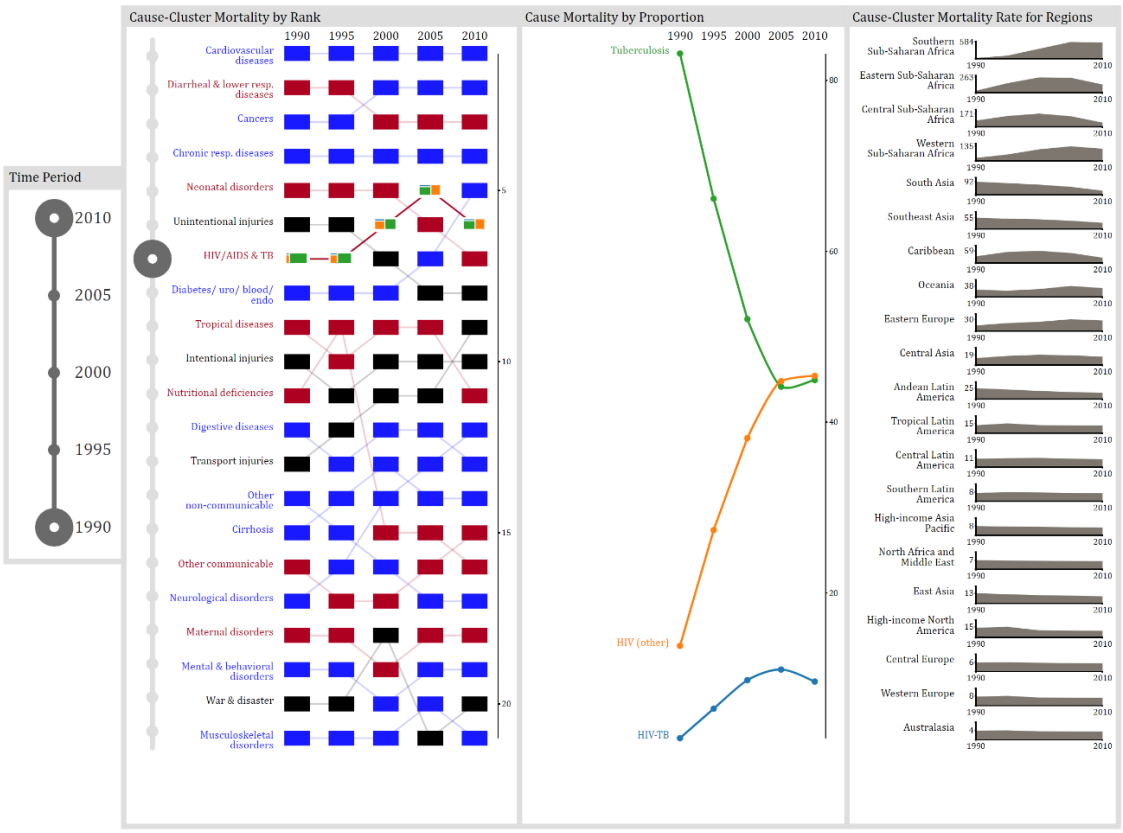



**Figure 4-11: Chronology visualization**

The third part of the visualization uses area charts to depict the temporal distribution of a selected cluster for each region of the world. For example, one can observe that for South Asia, death from HIV & TB has decreased over the 20-year period. The area charts are arranged according to their mortality rate, with the region with the highest mortality rate at the top. To design interaction, we once again considered the tasks that users would perform with the data. These tasks include assessing trends for cluster-specific mortality at a global and regional level, comparing cause and cause-cluster ranks, exploring temporal trends within a cluster on a global scale, and comparing rates between geographical regions. To support these tasks, we have operationalized the following actions: selecting, drilling, filtering, arranging, collapsing, expanding, and comparing.

For this last scenario, let us imagine a student, enrolled in a global health course, who has an assignment that requires him to answer the following questions:

1. Between 1990 and 2010, which cause-cluster increased in rank the highest?

2. Over the 20-year period, what was the highest rank for liver cirrhosis?
3. Between 1990 and 2005, which digestive diseases significantly decreased in proportion?
4. Which region of the world had the lowest mortality rate from cardiovascular and circulatory diseases between 1995 and 2005?

After reading over the questions, the student recognizes that the first three questions do not require regional-level data, and so he collapses the third panel. Through selecting each cluster, he observes that the neurological disorders cluster increased from position 17 to 12 over the 20-year period. He selects this cluster (see Figure 12a), writes down the answer, and then proceeds to the second question. To determine the highest rank for the liver cirrhosis cluster, he changes the time range and then selects the cluster as shown in Figure 12b. He notes that the highest position for liver cirrhosis was 13 in 2010. To discover the digestive disease that significantly decreased in proportion between 1990 and 2005, he changes the time range, selects the cluster and then focuses his attention on the cause mortality panel. Next, he performs a visual search and determines that peptic ulcer significantly decreased in proportion (see Figure 12c). To answer the last question, he collapses the cause panel and expands the region panel as shown in Figure 12d, and also changes the time frame so that only data between 1995 and 2005 is visualized. Next, he selects the cardiovascular and circulatory disease cluster and notices that Eastern Sub-Saharan Africa has the lowest mortality rate during the specified time frame.

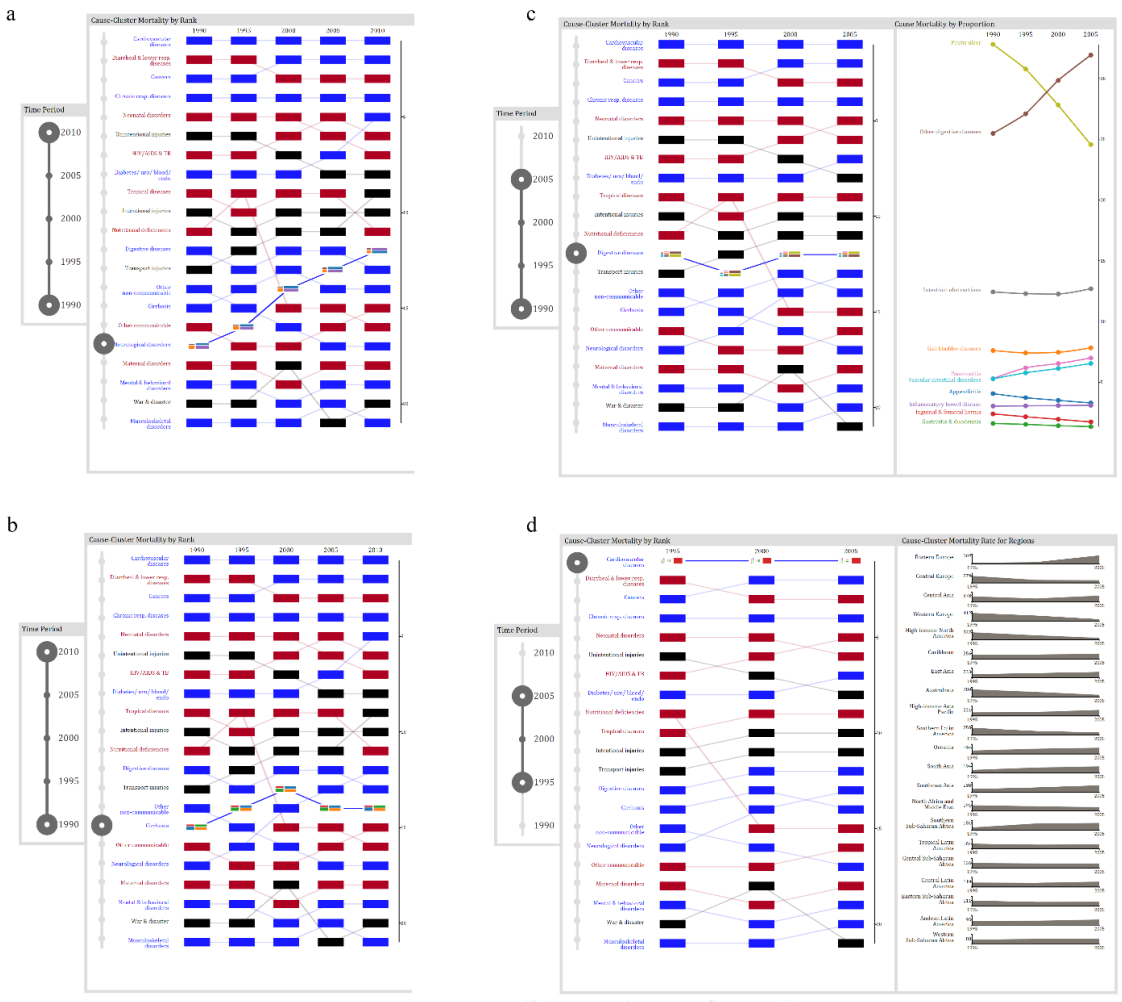



**Figure 4-12: (a-d) Chronology visualization screenshots**

The scenarios presented in this section briefly highlight some of the ways in which users can interact with the underlying data. In the first and second scenarios, we demonstrated how exploring the data in an open-ended fashion may occur. In the third scenario, we focused on how users can use the visualization to address specific questions. We have shown that when interaction is designed in a systematic fashion; users can perform a variety of tasks. The way users perform tasks is dependent on the complementary actions made available to them and the way in which the visualization reacts to their prompting.

## 4.6 Conclusion

Data has the potential to impact health care efforts positively. However, in the past, the health community has been slow to leverage the data and capitalize on the opportunities it presents. This is partially due to the complexity of the data and the challenges it presents for individuals trying to complete tasks. Interactive visualizations can play a role in supporting data-driven health tasks. While research indicates that visualizations allow users to interact with the data, to date, recent surveys suggest that much of the interaction in visualizations allows for simple manipulations.

In this paper, we contend that when dealing with large sets of data, users need to be able to interact with it and change its form so that they can perform tasks effectively. Designing interaction is a non-trivial issue and efforts to design it in an ad hoc manner may lead to tools that inadvertently constrain users' ability to complete tasks. There is a need for conceptual structures to help systematize the design process. In this work, we have demonstrated how elements of a theoretical framework contribute to structuring the design process for interaction. We have presented a design process and illustrated how it could be used. We have demonstrated the utility of designing interaction in a structured fashion. In the scenarios, we show how users can seamlessly perform inter-related tasks.

It is our intent to help raise awareness of the potential of interactive visualizations to support data-driven health tasks. Our future work will include user studies that detail the exploration experience. Although this research uses global health data, we anticipate that the design process presented may be beneficial in supporting other datasets in healthcare as well as other domains. Furthermore, systematically designing interaction has applications for large datasets, where leveraging expert knowledge is critical.

# Chapter 5

# 5 Exploring the Spread of Zika: Using Interactive Visualizations to Control Vector-Borne Diseases

This chapter has been published as O. Ola, O. Buchel, and K. Sedig, "Exploring the Spread of Zika: Using Interactive Visualizations to Control Vector-Borne Diseases," Int. J. Dis. Control Contain. Sustain., vol. 1, no. 1, pp. 47–68, 2016.



Please note that the format has been changed to match the format of the dissertation. Figure numbers mentioned herein are relative to the chapter number. For instance, "Figure 1" corresponds to Figure 5-1. Additionally, when the term "paper" or "article" is used, it refers to this particular chapter.

## 5.1 Introduction

Vector-borne diseases (VBDs), such as zika, malaria, and dengue fever, do not respect geopolitical boundaries, as evidenced by their prevalent spread across the globe. For example, for the first time in history, Chikungunya, a disease endemic in African and South Asian countries, is now present in Caribbean nations (Charrel, Leparc-Goffart, Gallian, & de Lamballerie, 2014). Vector-borne diseases are a major public health threat which results in over 750 thousand deaths each year (World Health Organization, 2012). VBDs increase health inequities, put a strain on health services, and negatively impact development and economic growth (Campbell-Lendrum et al., 2015; World Health Organization, 2012). In full awareness of the consequences of VBDs, public health (PH) stakeholders[i] have implemented various preventive, control, and treatment measures. While substantial progress has been made, there is still much more that can and should be done.

To control VBDs, PH stakeholders must make sense of the epidemiological and entomological data, analyze the local determinants of the disease, compare possible vector-control methods, predict morbidity levels so as to ensure sufficient supply of treatment measures, and perform various other decision-making tasks. While engaged in these tasks, PH stakeholders interact with data[ii]. This data has high volume, has extensive variety, and, in some situations, has low veracity (Eisen & Eisen, 2011; World Health Organization, 2012). These factors all contribute to the complex situation in which PH stakeholders operate to address VBDs. In addition to the challenges data presents, the multivariate nature of VBD poses additional obstacles to PH stakeholders. These challenges include understanding the complicated dynamics, interdependencies, and uncertainties that arise from various control strategies over time, and the impact of human-environment interaction on vector populations (Kramer et al., 2009). As with all infectious diseases, time plays a crucial role; the early detection of VBD outbreaks is essential to their control. When dealing with VBDs, the stakes are high, the challenges are immense, and a timely response is paramount. Therefore, computational tools that support the decision-making tasks of PH stakeholders are much needed. Fortunately, technological advances can dramatically change our capacity to predict, prevent, and control VBDs.

Interactive visualization tools (henceforth simply referred to as visualization tools without the adjective ‘interactive’) are a group of computational tools that has gained prominence in several disciplines over the last 20 years. These tools use interactive visual representations to convey information and support decision-making tasks by allowing users to control the flow of information, customize visual representations, and, in certain cases, perform other analytical tasks (Parsons & Sedig, 2014). Visual representations encode abstract or concrete information (e.g., geographic, scientific, or health data) in a visual form, and can be static or interactive. From the time Seaman used spot maps to study yellow fever in 1796 to the use of atlases to make sense of the endemicity of malaria in recent times, static visual representations have been used by PH stakeholders (Stevenson, 1965; Le Sueur et al., 1997). Though useful, static representations do not effectively support data-intensive tasks in which stakeholders engage. Visualization tools,

on the other hand, employ the use of interactive visual representations and, as a result, are better equipped to support the decision-making tasks of stakeholders.

Since VBDs decision-making requires PH stakeholders to reason with heterogeneous data, visualization tools can play a major role. The effective and efficient use of data determines the extent to which PH stakeholders can sufficiently address VBDs (Reeder et al., 2012; Thomsen et al., 2016). Therefore, tools that allow users to interact with information systematically can support the decision-making process. Visualization tools include Spatio-Temporal Epidemiological Modeller (Ford et al., 2006), Panviz (Maciejewski et al., 2011), and Google Flu Trends (Carneiro & Mylonakis, 2009). An awareness of the potential of these tools is needed so as to facilitate their incorporation into PH practice. To this end, this article will focus on how visualization tools can support PH stakeholders make better decisions when dealing with VBDs.

The rest of this article is organized as follows. Section 2 presents visualization tools through a discussion of their characteristics and functionalities. Section 3 highlights challenges facing stakeholders as they engage in decision-making tasks and explains how visualization tools can address them. Section 4 presents a visualization tool we developed to support making sense of the recent Zika outbreak in Brazil. Section 5 provides a summary.

## 5.2 Interactive Visualization Tools

With visualizations, users can control how much and the visual form in which information is represented; here, we discuss visual representations, the role of interaction, and how visualization tools can support decision-making tasks. In this section, we also briefly discuss the functionality of various visualizations tools.

### 5.2.1 Visual Representations

Visual representations encode data in forms that are visually perceptible to the stakeholder. Using visual marks (e.g., lines and dots) to compose more complex structural forms (e.g., maps and pie charts), these representations encode concrete or abstract information that can be geographic, scientific, or mathematical in nature (Sedig,

Parsons, Dittmer, et al., 2012). In VBD control, visual representations have been used to support various tasks. For instance, visual representations were used for analyzing the prevalance of dengue distribution in Malaysia (Aziz et al., 2012) and exploring the interrelationships among diseases and international air service routes (Huang, Das, Qiu, & Tatem, 2012).The widespread use of visual representations is based on research showing that they capitalize on the strength of the human perceptual system and are superior to textual representations for certain tasks (Kirsh, 2009; Parsons & Sedig, 2014). However, while visual representations can enhance the decision-making tasks of stakeholders, in some situations an improper representation can hinder the task in which the user is engaged (Parsons & Sedig, 2014). Therefore it is important that visual representations be appropriate for the task at hand (Sedig & Parsons, 2016).

In this article, we make a distinction between static and interactive visual representations. Though static visual representations are useful, nonetheless, due to not being dynamically manipulable (i.e., interactive) by those who use them, they put the brunt of the information-processing load (i.e., decision-making tasks) on the cognitive resources of their users (Sedig & Parsons, 2013). As a result, their effectiveness in data-intensive decision-making tasks is limited. Visualization tools, on the other hand, employ the use of interactive visual representations and, as a result, are better equipped to support the decision-making tasks of users. Empirical evidence from human-computer interaction literature shows that interactive visual representations can support decision-making tasks better than static representations. Specifically, when compared to static visual representations, interactive visual representations can significantly improve reasoning about mathematical, geospatial, medical, and unstructured textual data, to name a few (Ahonen-Rainio & Kraak, 2005; Carroll et al., 2014; Liang & Sedig, 2010; Robinson, MacEachren, & Roth, 2011). Furthermore, interactive visual representations can foster serendipitous discoveries, creativity, and hypothesis generation (Guo, Gahegan, MacEachren, & Zhou, 2005; Koua & Kraak, 2004).

### 5.2.2 Interaction

In the context of visualization tools, interaction allows users to manipulate visual representations by controlling the form and content of the visual representations. We use

the characterization of (Sedig & Parsons, 2013) – small actions initiated by users on visual representations, subsequent responses in the visual representations, and the user's perceptions of the changes – to describe interaction. For example, as shown in Figure 1, a user selects a subset of information for further analysis; in response, the visualization tool reacts to this prompt and highlights the selected information in a different color and the user can observe the changes on the visual representation. As different representational forms can have different effects on the user, the ability to change how data is represented aids in the completion of the specific decision-making task in which the stakeholder is engaged. Not only does interaction allow the user to manipulate visual representations, but it also plays a greater role of mediating the entire discourse between the user and information. The cyclic process of the user's action on the interface, the tool's corresponding reaction visible in the visual representations, and the user's perception of their changes promotes the discourse between the user and information.

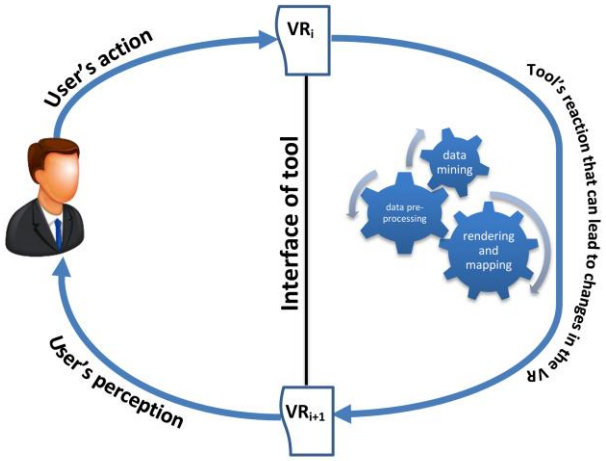



**Figure 5-1: Depicting how the user and tool interact, where $VR_i$ represents visual representation and $VR_{i+1}$ represents the altered representation**

As interaction plays such a pivotal role in facilitating the discourse, the quality of interaction must not be overlooked. Currently, there exist computational tools that boast of being interactive but only allow the user to alter the size of current representations (e.g., zooming, panning). In this article, we focus on interaction that allows the user to not only change the size of a visual representation but also engage in a more meaningful

discourse with data. In this context, interaction lets the user filter data, select analysis algorithms and techniques, drill deeper to examine latent information, transform the current representation, categorize data based on certain characteristics, compare scenarios, and perform a host of other tasks. All of these actions allow for a more involved discourse that promotes the distribution of decision-making tasks between the user and the tool.

### 5.2.3 Facilitating Decision-Making Tasks

Visualization tools can support decision-making tasks in which stakeholders engage (e.g., exploring, organizing, hypothesis generation, and comparing). When using such tools, tasks do not occur solely in the mind of the stakeholder, but instead can be distributed between the user and the tool (Sedig & Parsons, 2013). In other words, the user and the tool both work together to complete the task. For instance, an epidemiologist charged with determining the spread of dengue in a community might choose to delegate the computational sub-task of finding hotspots to the visualization tool. From the changes in the visual representation, the epidemiologist generates a hypothesis about how the disease has spread. In this scenario, both the tool and the epidemiologist collaborate to determine the origins of the outbreak. While using visualization tools, the stakeholder's ability to effectively address health concerns and hazards are enhanced. It is important to note that visualization tools do not take over the analysis process but instead support stakeholders as they engage in various tasks. As VBD problems are inherently ill-defined, this user-guided discourse is beneficial. Research has shown that tools that exclude the user's knowledge and focus on automated processing are ill-equipped to effectively support stakeholders (G. Andrienko et al., 2007; Ola & Sedig, 2014; Thomas & Cook, 2005). Humans are better able to reason about ill-defined problems with incomplete information. Through interaction and distributed task processing, visualization tools allow for the synergetic work in which both the human and the tool cooperate in a manner that capitalizes on the strength of each (G. Andrienko et al., 2007). Hence, tools that create a joint cognitive system with the user are essential for insightful thinking (Parsons & Sedig, 2013b). These tools accept stakeholder's background knowledge, support flexible thinking, and distribute the load of analysis. The extent to which tasks are aided by the

tool differs and depends on the visualization tool that is being used. In addition, the user's ability to control how and when decision-making tasks are performed also varies with the tool.

### 5.2.4 Diverse Functionality of Visualization Tools

Visualization tools include information visualization, visual analytics, and geographic visualization tools. Information visualization tools incorporate interaction techniques and visual representations to create an environment that supports the storage and exploration of abstract data. These tools allow users to control what information is represented, as well as the form in which it is represented. Furthermore, information visualization tools may include basic analysis algorithms that allow users to gain a deeper understanding of the data. Visual analytics tools, on the other hand, go a step further and support decision-making tasks that are not effectively addressed through the use of information visualization tools. Through the use of advanced storage and processing algorithms, visual analytics tools allow for the synthesis and analysis of heterogeneous data in ill-defined problems. Furthermore, visual analytics tools seek to incorporate the user's knowledge into the decision-making process by providing users with greater control over the discourse with information. For an in-depth discussion on the differences between information visualization and visual analytics tools, the reader can refer to (Keim et al., 2008, 2009).

Geographical information systems are systems which keep track of events, activities, other phenomena, and where they all happen or exist (Longley, Goodchild, Maguire, & Rhind, 2005). In this article, we distinguish between GIS tools that use static visual representations and those that use interactive ones. Though beneficial, not only do static maps require users to bear the brunt of decision-making tasks, but also these maps come with some degree of subjectivity. The issue of subjectivity arises because "maps are never fully formed and their work is never complete" (Kitchin & Dodge, 2007, p. 331). As a result, they represent a snapshot of reality from the view of the cartographer (Davies, Fabrikant, & Hegarty, 2013) which is not always objective and therefore can result in bias that influences decision-making tasks. Furthermore, GIS tools with limited actions (e.g., having only zooming and panning) are not well-suited for making informed

decisions in a spatial context (N. Andrienko & Andrienko, 2003). These limitations have led researchers to recognize the need for tools that support the use of interactive and dynamically alterable thematic maps which can facilitate "visual thinking" about spatially referenced data (N. Andrienko & Andrienko, 2003). From this point on when we refer to GIS tools as geographic visualization tools, we are focused on GIS tools with interactive capabilities that go beyond simple alterations of the representation's size.

Geovisualization tools facilitate visual exploration, analysis, synthesis, and presentation of geospatial data (Kraak, 2006). These tools use interactive maps that reduce subjectivity and facilitate exploratory visual analysis rather than the pre-defined mapping common in older GIS tools (Andreinko, Jern, Dykes, Fabrikant, & Weaver, 2007). Spatial decision-making is an ill-defined process and, as a result, automated and pre-defined mapping methods are inadequate in addressing stakeholders' activities. Through interaction, stakeholders can perform a myriad of actions including selecting, navigating, filtering, and animating, to name a few (Buchel & Sedig, 2014). Furthermore, stakeholders can manipulate information, and the geovisualization environment becomes a place where multiple decision-making tasks can take place.

## 5.3 The Role of Interactive Visualization Tools in Addressing Challenges Facing Stakeholders

With an understanding that VBD control and eradication is dependent on a variety of factors, the World Health Organization has advocated for the adoption of an integrated vector management approach to decision-making. This approach seeks to improve the efficacy, cost-effectiveness, ecological soundness, and sustainability of vector disease control (World Health Organization, 2012). This approach is predicated upon a systematic and rational analysis of data, incorporation of control/treatment methods based on knowledge of influencing factors, development of policies that use a range of interventions, and interdisciplinary collaboration that spans both public and private organizations. The multidimensional nature of VBD presents challenges that limit stakeholders' ability to effectively and efficiently control, prevent, and treat these diseases.

The use of visualization tools in decision-making can be advantageous to PH stakeholders. Through interaction, these tools allow for the distribution of decision-making tasks between the user and the tool, thus providing a synergetic environment in which both the tool's processing capacity and the human's ability to deal with ill-defined problems is used to its fullest (G. Andrienko et al., 2007). Visualization tools facilitate collaboration and support the evolutionary and iterative process of decision-making (Sedig & Parsons, 2013; Thomas & Cook, 2005). Additionally, these tools help users extend their problem solving abilities through processing massive amounts of data quickly, providing timely and comprehensible assessments, discovering trends, patterns, correlations, and outliers, and reducing the search for information (Gotz & Borland, 2016; Ola & Sedig, 2014; Robinson et al., 2011). As tools with interactive visual representations, these tools enable latent information to be made explicit when desired, coordinate internal representations with external representations, provide explicit encodings of information that facilitate discussion, increase the memory and processing resources available to users, and convey visually difficult statistical algorithms in a comprehensible manner (Kirsh, 2009; Sedig, 2001). The benefits above are not an exhaustive list; rather they provide a cross-sectional view. The remainder of this section describes how visualization tools can address specific challenges relating to the nature of the data, human-environment variables, and the dynamics of the disease and highlights current tools.

### 5.3.1 Multifaceted Data

As data plays a pivotal role in all stages of VBD control, stakeholders' ability to systematically access, use, interact with, and analyze collected data is crucial (Thomsen et al., 2016). Some of the facets of VBD data can be described in terms of its volume, variety, and veracity. VBD data is voluminous, originates from a myriad of sources, and is often collected independently by discipline (Guerra et al., 2007). This data includes medical records collected at hospitals, environmental data generated by remote sensing surveillance systems, mosquito migration patterns collected by vector ecologists, genomic data from biological databases, policy briefs from government legislature, and mosquito net distribution data from non-governmental agencies or local health

departments. In regards to its variety, VBD data is stored in different formats (e.g., numerical, textual, and geospatial), and ranges from structured (e.g., malaria indicator survey data) to unstructured[iii] forms (e.g., free-form paragraphs in a policy brief or tweets about medical symptoms) (Eisen & Eisen, 2007; Kelly, Tanner, Vallely, & Clements, 2012; World Health Organization, 2012). In addition, the low veracity (i.e., accuracy and completeness) of the data further complicates stakeholders' decision-making tasks. Low veracity can result from partial or erroneous reporting at the point of data collection and temporary gaps in transmission from surveillance systems (Mandl et al., 2004; Wilkins, Nsubuga, Mendlein, Mercer, & Pappaioanou, 2008). The nature of VBD data presents challenges to stakeholders who engage in decision-making tasks.

While visualization tools cannot change the characteristics of data that make it challenging to use (i.e., high volume, various sources, and low veracity), these tools can support the systematic use of such data. Visualization tools can facilitate the storage, transformation, and analysis of data to help stakeholders arrive at conclusions, form new knowledge, and make sense of the data. Furthermore, through interaction, stakeholders can control the amount of information presented at one time, thus dealing with the challenge of information overload. For instance, Koua & Kraak (2004) developed a geovisualization tool that supports the visual data mining and exploration of health statistics and survey data. Previously the size of this data would present a challenge to stakeholders, but through the use of self-organizing maps and artificial neural networks, stakeholders engage in exploration to gain insights into the patterns, trends, and appropriate underlying distributions inherent in the data. Epinome is another tool that addresses concerns of multiple streams of heterogeneous data that stakeholders may use in decision-making tasks. This tool facilitates the detection, monitoring, exploration, and discovery of infectious disease (Livnat et al., 2012). Another beneficial tool, MaGnET, facilitates the exploration of genomics data related to the malaria parasite, *Plasmodium falciparum* (Sharman & Gerloff, 2013). Through the use of multiple linked views, the user can examine features of groups of genes across different datasets, explore networks of proteins, query the database to retrieve a subset of information, and perform other decision-making tasks. Another tool seeking to address the data challenge facing stakeholders is WHO's HealthMap (Freifeld, Mandl, Reis, & Brownstein, 2008). This

tool gathers data from disparate web-accessible sources, including unstructured data, and presents on an interactive map alerts of various diseases, including Zika, Dengue, Malaria, Chagas, and Lyme disease.

### 5.3.2 Human-Environment Interactions

VBD control requires an understanding of the interaction among various components including humans, environment, and vector populations (Campbell-Lendrum et al., 2015; Chareonviriyaphap et al., 2013). As the distribution and prevalence of VBDs are strongly influenced by ecological conditions in the natural environment, an understanding of environmental factors is crucial for controlling the vector population (Tabachnick, 2010; World Health Organization, 2014). Human activities have impacted the environment and have led to the emergence and resurgence of many vector-borne diseases (Kilpatrick & Randolph, 2012; LaDeau, Allan, Leisnham, & Levy, 2015; Weaver, 2013). Therefore, in addition to understanding vector populations, it is also important to consider the social and behavioral aspects of humans that impact the environment (Heggenhougen, Hackethal, & Vivek, 2003). Changing migratory patterns, globalization, rapid urbanization, and irrigation are examples of human behaviors that can exacerbate or lead to deforestation, increased salination, and a host of other environmental issues that have been linked to influencing the vector populations in an area (Chan, 2014; Kraemer, Sinka, Duda, Mylne, Shearer, Brady, et al., 2015; Weaver, 2013). For instance, urbanization has contributed to the spread of dengue and chikungunya in recent years (Kraemer, Sinka, Duda, Mylne, Shearer, Barker, et al., 2015). An understanding of the complex human-environment interactions provides a unique opportunity to combat the spread of VBDs (Hartemink, Vanwambeke, Purse, Gilbert, & Van Dyck, 2015). Unfortunately, current disease prevention and control interventions that consider the human-environment connections tend to be the exception more than the rule (Boischio, Sánchez, Orosz, & Charron, 2009). Therefore, tools that can facilitate an understanding of human-environment interactions would be of great assistance to the decision-making tasks of stakeholders.

Visualization tools can include models that promote the understanding of human-environment interactions. Research shows that imported cases of malaria belong to

networks of people with similar travel patterns (Koita et al., 2013). These human migratory patterns have an impact on VBD control measures. VBD-Air is a web-based visualization tool that examines the role of human migration via air travel and helps stakeholders make sense of how such migratory patterns influence the transmission and spread of VBDs (Huang et al., 2012). With VBD-Air, stakeholders can explore the interrelationships among modeled distributions of diseases. In addition, such tools can incorporate models (such as the three Dengue models evaluated by Nakhapakorn & Tripathi (2005)) to facilitate the understanding of how climatic factors such as rainfall and humidity affect incidences of dengue. Such models can be used to forecast the number of future dengue incidents based on current environmental factors. While such models do not consider the effect of social factors on vector populations, they integrate physio-environmental factors (e.g., land use and cover) that result from human-environment interaction and thus can serve as a starting point on which further models can be built. Chang et al. (2009) report another tool that uses Google Earth as part of a dengue surveillance program. This tool visualizes locations of interest related to larval infestation (e.g., location of tire dumps and large areas of standing water) to allow stakeholders engaged in decision-making tasks to prioritize specific high-risk neighborhoods based on ecological factors.

### 5.3.3 Changing Disease Dynamics

The locality, uncertainty, and interdependencies of VBD dynamics further complicate stakeholders' decision-making tasks. The variability inherent in VBD dynamics makes it difficult to ascertain future ramifications and the effectiveness of current policy decisions on a community (Kramer et al., 2009; Mendis et al., 2009). The locality of disease dynamics requires the contextual understanding and comparison of intervention strategies because not all strategies are effective for a particular community. For instance, chloroquine was the major treatment for malaria in West Africa in the 1990s, but East African nations had to change their antimalarial drug policies mid-season because of the rapid spread of a chloroquine-resistant parasite (D'Alessandro & Buttiëns, 2001). This example shows how one treatment measure fails to address the concerns of another environment. Furthermore, the uncertainties of disease dynamics within a community

also present an additional challenge as effective strategies today might prove ineffective in the future within the same community (Bloland, 2001). For instance, insecticide resistance levels have risen over the years, and there is emerging evidence that such resistance is negatively impacting malaria control efforts (Mnzava et al., 2015; World Health Organization, 2015). In situations where control of the vector population is the primary mode of prevention, the use of ineffective insecticides can potentially have a disastrous effect on communities. The mutation of the mosquito, the resistance of the virus to treatment, and the differing impacts on local communities create dynamic challenges for the defining of long-term policies. Computational tools that can introduce scientific uncertainty into models, thus providing bounds for ill-defined problems and creating environments to explore choices for different communities and time frames, can help stakeholders address this challenge.

Visualization tools allow for the incorporation of dynamic decision analysis models that can help stakeholders explore the possible impacts of choices over different time frames as well as the various communities based on their ecological characteristics. One such model described by Luz, Vanni, Medlock, Paltiel, & Galvani (2011) focuses on simulating dengue transmission to facilitate the understanding of the evolution of insecticide resistance and immunity in the human population. Another tool, STEM, uses mathematical models of disease to simulate the development, evolution, and transmission of a disease in both space and time (Ford et al., 2006). This tool accommodates the creation of models specific to a population and disease strains. Developed by IBM, STEM is an open source tool and has tutorials on how stakeholders can create and use models in their decision-making tasks. A customized visualization tool that specifically addresses rift valley fever dynamics is Panviz. This tool supports epidemic modeling and response evaluation for decision-making (Maciejewski et al., 2011). It does not automatically determine the best solution, but instead allows stakeholders to understand the effect various measures would have on disease control so that they can determine the best set of optimal control measures. Furthermore, Panviz facilitates the exploration of temporal decisions as well, allowing stakeholders to examine the role time plays in combating the disease. Visualization tools allow stakeholders to develop possible control

measures and gauge their impact on the disease before they are deployed in the community.

## 5.4 Case Study: Zika Outbreak in Brazil

This section demonstrates how visualization tools can support decision-making tasks of PH stakeholders engaged in VBD control. In particular, we focus on the recent Zika outbreak in Brazil. Zika—a vector-borne disease transmitted among humans by *Aedes* mosquito species—was first detected in Brazil in May 2015. Since then, the Zika virus (ZIKV) spread rapidly around the region. In the last year, there has also been a surge in the number of infants born with microcephaly. Microcephaly is a rare neurological condition associated with incomplete brain development. The prevalence of the disease led WHO to declare that the Brazilian cluster of microcephaly constitutes a Public Health Emergency of International Concern (World Health Organization, 2016). Our goal is not to show a causal relationship between Zika infection during pregnancy and microcephaly, but to support PH stakeholders in understanding the spatiotemporal characteristics of both diseases so as to improve control efforts. To develop control responses, PH stakeholders engage in a variety of tasks. One of these tasks involves the exploration of disease frequencies across and within geographical regions so that the parties concerned can determine which communities are most at risk. As diseases do not respect regional boundaries, it is important to explore the prevalence of diseases at multiple levels of aggregation.

To this end, we utilize a subset of the Dryad data package (N. Faria et al., 2016) that contains records of suspected ZIKV and microcephaly cases in Brazil for the year 2015. The findings based on this dataset are published in (N. R. Faria et al., 2016). This dataset includes incidence of ZIKV cases and passively reported microcephaly cases per 100,000 people in each federal state. It is important to note that because microcephaly is being passively reported the dataset is not as complete as the ZIKV dataset. Faria et al. (2016) aggregated ZIKV and microcephaly data by provinces and showed disease frequencies on two regional maps. In the Dryad dataset, however, the data are referenced at a more fine-grained level (i.e., municipalities) which implies that the dataset has a latent spatial structure with random effects which cannot be made explicit when the data are

aggregated by large regional units. Understanding the intricacies of spatial data structures is critical for making sense of the spatial relationships among diseases and environmental factors.

Using spatial data to develop control efforts is not a trivial matter. ZIKV and microcephaly have spatially non-stationary processes—processes which exhibit significant spatial variation. Static visualizations of geographical summary (GW) statistics have been used in some studies to explore non-stationary processes (Homan et al., 2016; Matthews & Yang, 2012). GW models have been found to provide insights into spatial targeting of intervention and control programs against disease outbreaks (Y. Liu et al., 2011). GW summary statistics (e.g., means, standard deviation, skewness) aid stakeholders in exploring the complexities of non-stationary processes in great detail as they describe each locational fragmentation and variance with multiple statistics. Consequently, they are beneficial in developing plans, policies, intervention procedures, and decision-making strategies to mitigate the adverse effects of spatial variability (Brunsdon, Fotheringham, & Charlton, 2002; Harris, Clarke, Juggins, Brunsdon, & Charlton, 2014). The visualization we designed utilizes GW statistics as opposed to using standard simple means and choropleth maps. We use GW summary statistics because diseases are distributed not across an "average" space but full of variations; hence statistical techniques must account for different forms of spatial heterogeneity or non-stationarity (Goodchild, 2004; Lu, Harris, Gollini, Charlton, & Brunsdon, 2011). The four GW statistics we use are briefly described below.

- **GW means** are computed by weighting each observation in the dataset according to its proximity to a summary point. The closest points are given higher weights than remote points. After a certain threshold the points' weight is reduced to 0. Because of geographic weighting, the means do not vary as much as the raw counts of disease incidents. GW means are especially useful when the numbers of samples taken at different places vary as in the case with ZIKV in Brazil.

- **GW standard deviation** is beneficial in emphasizing areas of high variability. They can also help detect transitional zones between low and high incidence areas.

Knowing where transitional zones are located is essential for managing and controlling epidemics, as transitional zones are affected first when a virus spreads.

- **GW skewness** determines the symmetry of the spatial distribution. In statistics, skewness plays a major role in hypothesis testing and analysis of variance as it determines whether data are normally distributed and whether local skewness departs from global. Additionally, skewness plays an important role in modeling (Bekaert & Harvey, 2002; Harvey, 2000).

- **GW coefficient of variation (or difference)** considers the degree of variation in proportion to the changing mean. A constant coefficient of variation across all locations implies that the proportion of local variability of incidents is fixed and there are no spatial non-stationary processes (Fotheringham, Brunsdon, & Charlton, 2002). Knowing the coefficient of variation is beneficial to stakeholders as it helps measure the dynamicity of the outbreak/virus.

In addition to GW summary statistics, we also use Standard Deviational Ellipse—a model for describing spatial distribution. This model measures the disease concentration or dispersion around the mean center (Mitchell, 2005). If the data are normally distributed, one standard deviation ellipse covers approximately 68% of incidents. The mean center of the ellipse is a point representative of the center of all disease cases. The ellipse is an example of global statistics which shows a general spatial trend. Calculation of the ellipse generates one set of results, representing one set of relationships, which are assumed to apply equally to the entire region under investigation. These results characterize an "average" type of phenomenon behavior. The uniqueness of ellipse is in showing the directionality of the spatial trend (specifically, it indicates the angle of rotation). A change in directionality may be indicative of changes in disease vectors (Khan, 1992). Next, we describe the visualization and then with the use of a scenario demonstrate how the tool may benefit PH stakeholders.

### 5.4.1 Visualization Description

The subset of the data we utilized from the Dryad dataset includes 50 consecutive epidemiological weeks. We preprocessed the data by filtering out locations that had no ZIKV or microcephaly cases. As a result of this filtering, the dataset was reduced to 380 locations. These sites were geocoded using Google Geocoding API. We use a map-based visualization so that stakeholders can explore cases of ZIKV and microcephaly across municipalities. The visualization was created using Mapbox API mashup and developed in JavaScript. The mashup utilizes spgwr R package (Bivand & Yu, 2009) for producing GW summaries and OpenCPU package (Ooms, 2014) for reading outputs from R packages in JavaScript. We also utilized the jQuery and D3.js libraries for implementing additional graphs and sliders. Figure 2 shows the default view of the visualization. The interactive visualization has some elements namely, layered-map, control elements, streamgraph, and description panel. The layered map as shown in Figure 2 depicts the GW means for ZIKV across the municipalities in Brazil. Each marker on the map represents values of local means, standard deviations, skewness, and coefficient of variation depending on what the map shows. We use gradient color (from yellow to red) to denote the intensity of values in ascending order.

Layered on the map is an ellipse which represents the standard deviational ellipses**.** Below the map is a streamgraph that allows users to compare frequencies of ZIKV and microcephaly over time. A streamgraph is a stacked area graph displayed around a central axis. As depicted, the yellow stream represents the flow of ZIKV over time, while the red stream represents microcephaly. With the provided controls, users can interact with the information. On the left of the map are two drop-down lists that allow users to choose between ZIKV and microcephaly, as well as, select a statistical measure. Below the drop-downs are four sliders which enable users to filter the data based on the statistical measures. With filtering, users can identify communities, make sense of spatial relationships within communities, track the evolution of spatial relationships within communities, and prioritize communities for future intervention efforts. In addition, sliders are useful for understanding how individual summary measures (means, standard deviations, skews, and differences) work as many novice users may not know the

mathematical models behind GW summary measures. Moreover, the sliders allow users to ask questions about outliers and unique, atypical properties of municipalities (e.g., high means but low standard deviations).

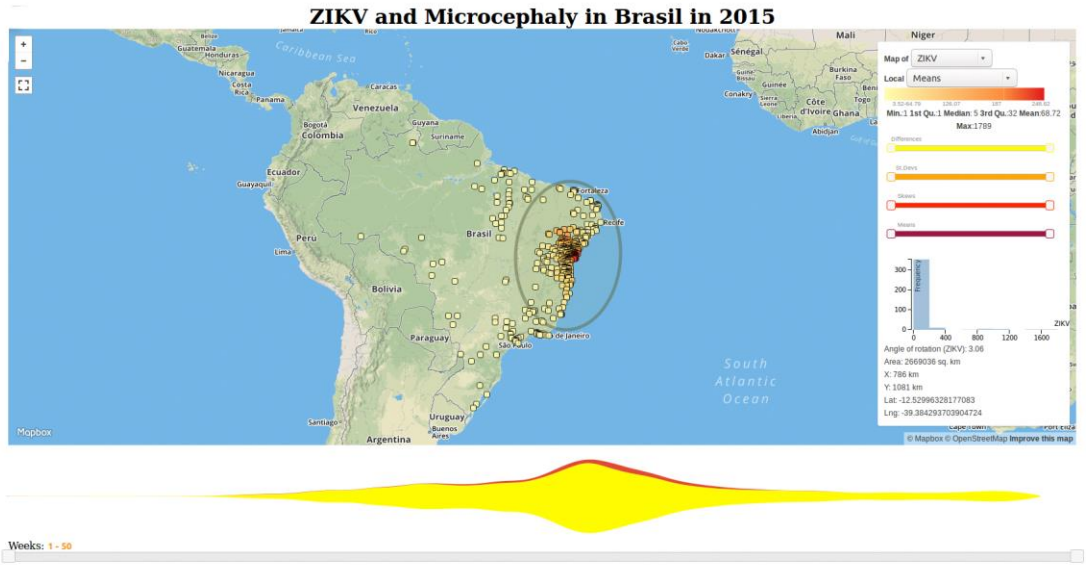



**Figure 5-2: Default screenshot of the visualization tool**

Below the GW sliders is a histogram that shows how frequencies of disease incidents are distributed globally. This summary is important because it gives users a sense of the largest and smallest values in the dataset and complements spatial descriptions provided by standard deviational ellipse and GW summary. The statistics under the histogram are descriptions of the properties of the standard deviational ellipse. They help users better understand changes in ellipses over time (e.g., whether the angle or radii have changed). Below the streamgraph is a timeline filter that enables users to make temporal selections. Interaction with the timeline allows stakeholders to inspect the evolution of non-stationary processes over time. Changes in time ranges affect the representations of the map and the legend (which are recalculated upon each adjustment). During the exploration, stakeholders may notice that ZIKV and microcephaly are non-stationary, not only spatially but also temporally. For example, through interaction, we identified that a cluster near Salvador, BA, with high means disappeared during the last ten weeks of the year, while a cluster near Petrofina, PE, became more prominent. Such observations may

lead to hypotheses about causes of such spatial transformations and help stakeholders assess whether preventive measures work. Changing ranges on the timeline also affects the representation of ellipses on the map. In spatial triaging, users often use snapshots of ellipses at discrete points in time. While this may be beneficial, it fails to effectively support understanding how ellipses evolve or change as they transition from one snapshot to another (for example, whether they rotate, shrink or expand). Animating the ellipse's transformation can reveal the center of the outbreak as well as lead to questions about how vector-borne diseases spread.

### 5.4.2 Scenario

As stakeholders interact with the visualization tool, they can make observations that are critical to prioritizing communities. Here, we present a brief scenario describing the usage of the visualization tool. The default representation of the tool is the GW means of ZIKV (Figure 2), at this point, the user may notice that incidents of ZIKV vary across administrative boundaries as well as within them. By selecting microcephaly, the user observes the same pattern for the congenital disease. As the user explores ZIKV, he may perceive that within each administrative region, several sub-regions emerge from data. For example, Figure 3 shows different sub-regions within the administrative boundaries of Bahia province in Brazil. As the user filters the means, with the sliders, he notices three clusters. Means with high values are grouped together in the area near Salvador, BA. They form a cluster with the red centroid and orange markers. Another orange cluster is located in the northwest, near Petrolina, PE. This cluster is not entirely in Bahia; it crosses the administrative boundary of Pernambuco which would not have been noticeable if data were aggregated by administrative boundaries. The red and orange clusters are hotspots of ZIKV. Through interacting with the GW means representation of ZIKV cases, the user determines these clusters and possibly develops a hypothesis about breeding sites of mosquitoes that carry ZIKV.

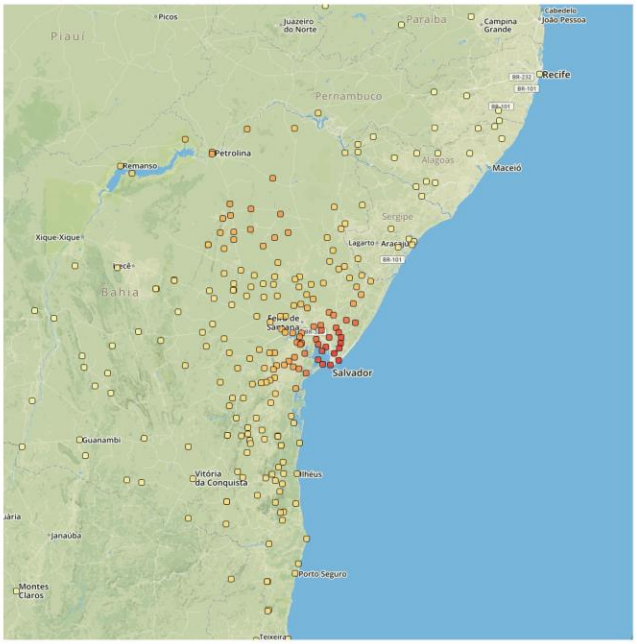



**Figure 5-3: The GM means representation shows that incidents of ZIKV vary not only across administrative boundaries but also within**

Next, the user may decide to examine the GW standard deviation to ascertain if there are areas of high variability. Areas of high variability may exist because of a variety of socio-economic and/or environmental factors. The GW standard deviation also highlights transitional zones between low and high incidence areas. Figure 4 shows the orange cluster near Petrolina, PE; this is the area that has the highest standard deviations. The user may notice that in addition to Petrolina, Salvador, BA, also has high standard deviations. These two clusters may not be of usual interest to the user because their means are also high. However, the next cluster (emphasized in Figure 4 by the circle) may be of interest because it includes yellow markers located south from Salvador, BA. By using the available sliders, the user identifies zones between low and high incidence areas (i.e., transitional zone). While this area does not have means as high as the other two clusters, based on its GW standard deviations it can be considered an area of interest.

Because the user needs to understand the variability of the diseases, he selects Differences from the dropdown for ZIKV and then for microcephaly. Both ZIKV and microcephaly have local variability in areas remote from the clusters with high means.

Microcephaly's coefficient of variation falls within the range -2.1 – 0.8, and ZIKV's falls within the range -5.4 – 2.1. If the coefficient of variation is constant across all locations, it implies that the degree of variability as a percentage is fixed across all locations (Fotheringham et al., 2002), even though the absolute values of means and standard deviations may fluctuate more. However, that is not the case here. While a majority of the areas have a coefficient of variation close to 2 (i.e., red-colored markers in Figure 5), there are areas where high means of microcephaly are observed in combination with low differences (i.e., yellow markers). This lets him know that there is spatial variability. Spatial variability is important because it suggests that further analysis should be done with geostatistical methods and not classical statistics to better understand the spread of the disease.

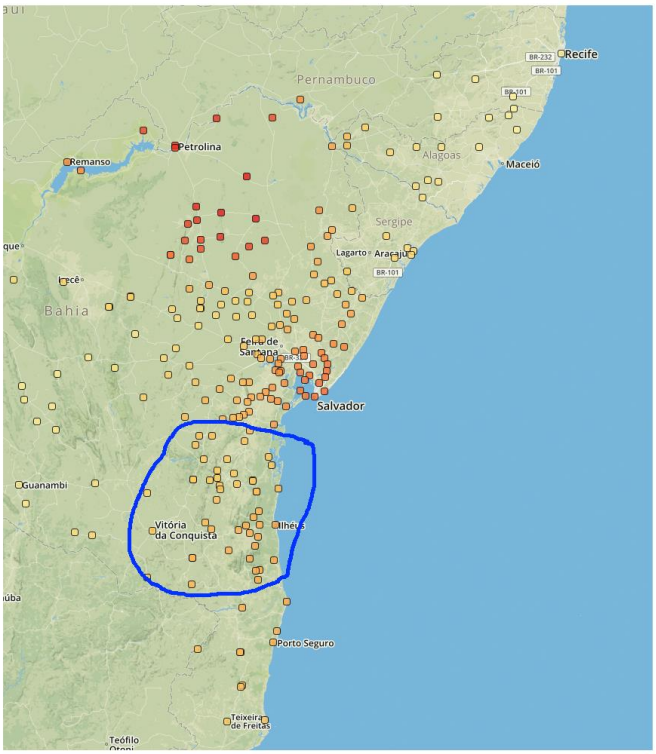



**Figure 5-4: The GW standard deviation representation for ZIKV shows that the clusters near Petrofina, PE, and Salvador, BA, have high standard deviations and the annotated area may be a transitional zone**

As the user continues his exploration, he observes that the ellipses for ZIKV and microcephaly look different, despite significant overlaps. While both ellipses enclose 95% of disease incidents, the angles of rotation and radii from east to west and north to south in ellipses are different as shown in Figure 6. The ZIKV ellipse covers a highly concentrated cluster along the South Atlantic belt.  The microcephaly ellipse covers almost the entire territory of Brazil. The area of microcephaly ellipse is larger because the distribution of microcephaly is more homogenous and is more spread out than the distribution of ZIKV. The overlap suggests that diseases are co-occurring.


**Figure 5-5: The GW coefficient of variation representation for microcephaly**

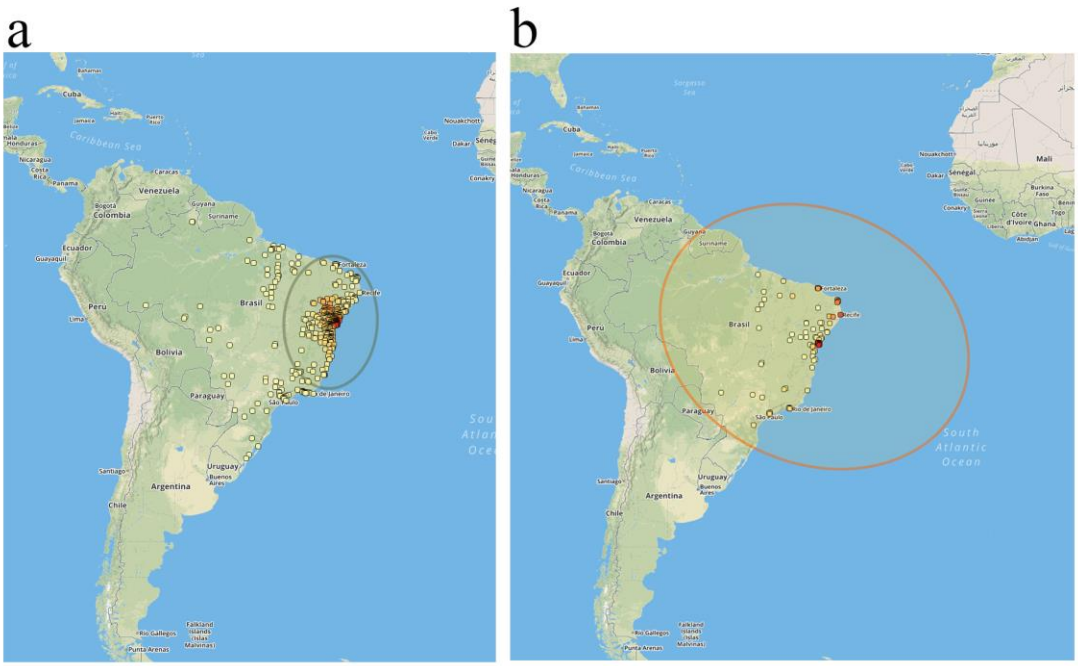



**Figure 5-6: (a) ZIKV standard devotional ellipse; (b) Microcephaly standard deviational ellipse**

In this brief scenario, we have demonstrated how an interactive visualization tool can support making sense of the spatial-temporal distribution of ZIKV and microcephaly. Users can explore different time frames, select specific statistical measures of interest, filter out extraneous data, compare the distribution of ZIKV and microcephaly, as well as, examine the disease dispersion around the mean center. Based on these interactions, users can make sense of disease prevalence, prioritize communities, and develop hypotheses that can be followed upon later.

## 5.5 Summary

PH stakeholders engage in a variety of decision-making tasks as they seek to control, prevent, and treat vector-borne diseases. While involved in these tasks, PH stakeholders encounter a variety of challenges arising from the nature of VBD data, the impact of human-environment interactions on vector, and the changing disease dynamics. As a result, computational tools that can support decision-making tasks are greatly needed. Through a discussion of interactive visualization tools, this article has demonstrated how such tools can address some of the challenges facing stakeholders.

Visualization tools use interactive visual representations to communicate information and support decision-making tasks by allowing users to control the flow of information, to customize visual representations, and, in certain tools, to perform analytical reasoning tasks. Capitalizing on the strengths of the human visuoperceptual system, visual representations can enhance the stakeholders' ability to perform numerous decision-making tasks. In particular, interactive visual representations have been shown to be more effective than static ones in reasoning activities in which stakeholders engage. Through interaction, stakeholders engage in meaningful discourse with information. Stakeholders can filter, drill, compare, categorize, and perform other actions that promote the distribution of decision-making tasks between the user and the tool. Decision-making tasks are user-guided and, as a result, stakeholders can incorporate their knowledge into the decision-making process. This synergetic discourse between the user and the visualization tool ameliorates challenges previously discussed.

Visualization tools can process massive amounts of data. They provide timely and comprehensible assessment and help with the discovery trends, patterns, correlations, and outliers. In addition, visualization tools facilitate the storage, transformation, and analysis of data, thus supporting the systematic use and conversion of data into actionable information. These tools allow for collaboration and customization so as to support a diverse group of stakeholders. Visualization tools can include models that promote the understanding of human-environment interactions that can negatively impact intervention, treatment, and control measures. Furthermore, these tools can increase the memory and processing resources available to users and communicate visually difficult information such as the complex dynamics of VBDs.

In this paper, we also presented a visualization designed to support decision-making related to the recent Zika outbreak in Brazil. As stakeholders develop control responses, they may need to explore the incidence rates so as to prioritize communities for intervention. The interactive visualization we created uses a map-based representation layered with geographically weighted statistical measures to describe non-stationary processes common in vector-borne diseases. The GW statistics we used describe datasets similar to standard descriptive statistics; however, by using a map-based representation,

they show patterns that have well-interpretable meaning in epidemiology. With a brief scenario, we demonstrated how interaction allows users to explore different time frames, examine specific statistical measures, and filter out extraneous data.

In conclusion, visualization tools can help PH stakeholders engaged in decision-making tasks dealing with VBDs. These tools should not be viewed as a cure-all for the challenges facing PH stakeholders. The effectiveness of visualization tools in PH practice is dependent upon their proper design. This calls for more research in this area. In this article, we have presented the utility of visualization tools through a discussion of their characteristics, examination of challenges facing stakeholders, and presentation of existing tools. As VBDs are multifaceted, complex, and changing constantly, what has been presented here can serve as a starting point for researchers in this area. Research into tools must evolve so that stakeholders can effectively treat, prevent, control, and, eventually, eradicate these diseases.

# Chapter 6

# 6 Understanding the Discussion of Health Issues on Twitter: A Visual Analytic Study

To be submitted to Health Informatics Journal.

Please note that the format has been changed to match the format of the dissertation. Figure numbers mentioned herein are relative to the chapter number. For instance, "Figure 1" corresponds to Figure 6-1. Additionally, when the term "paper" or "article" is used, it refers to this particular chapter.

## 6.1 Introduction

People gather health information from diverse mediums, including social media. Using social media allows individuals to explore conversations occurring outside of the traditional health space in a rapid fashion (Cole-Lewis et al., 2015; Park, Rodgers, & Stemmle, 2013). Twitter is one of the largest social media platforms with over 317 million active accounts as of January 2017 (Aslam, 2017). This platform allows users to post short comments (i.e., tweets) that contain 140 characters or less. Tweets may also contain pictures, videos, or links to webpages. Users can like, retweet (i.e., repost a tweet), and reply to tweets. Unregistered users can only read tweets. The unrestricted access to opinions and large user base has made Twitter a source for the collection and dissemination of information for various domains including health (Gurman & Clark, 2016; Hughes, 2016).

Currently, health organizations are using Twitter to promote healthy lifestyle choices, identify disease outbreaks, explore human behavior, and assess the public's perception of health issues (Charles-Smith et al., 2015; Finfgeld-Connett, 2015; Park et al., 2013; Salathé & Khandelwal, 2011; Weeg et al., 2015). Various health organizations use Twitter for health promotion. The Department of Health and Human Services in the United States is one such organization that uses Twitter to provide the public with actionable health information (Osborne, 2012). A study on three health organizations

observed that the organizations used Twitter to inform people about services, educational programs, and events; solicit for readers to take action; inform people about health risks; and encourage them to receive preventative screening or modify their lifestyles (Park, Reber, & Chon, 2016). In addition to health organizations, individuals, news organizations, businesses, interest groups, and a host of other entities discuss health on Twitter.

On any given day, over 500 million tweets are posted (Aslam, 2017). The sheer number of tweets present challenges for the public as they seek to use Twitter to improve their knowledge on a wide variety of health issues. Observational studies on specific health issues on Twitter shows an abundance of both formal and informal conversations taking place. While following a health organization's Twitter account may be beneficial for learning about a specific health hazard, for individuals who want to obtain a high-level understanding of the social discourse on a wide variety of health issues, challenges abound. Currently, it is difficult for users to understand the overall sentiment on a health issue, the types of users involved in the discourse, and what they are tweeting about. The brevity of the message can result in its true meaning being distorted and possibly taken out of context (Chou, Hunt, Beckjord, Moser, & Hesse, 2009; Kamel Boulos, 2013). In addition, the quality of the information is highly variable and the identity of the tweeter (i.e., who is tweeting), which is an important clue in assessing information credibility, is not always known (Kamel Boulos, 2013; Schein, Wilson, & Keelan, 2010). For Twitter to be an effective tool for health promotion, people need to be equipped to understand and appraise health information on the platform (Sørensen, 2017). A high-level understanding can help address misinformation and equip individuals with a better mental structure to assess how health issues are discussed. In addition to supporting the information-seeking tasks of the public, an analysis of the health discourse on Twitter benefits health professionals and social scientists. It provides them with a lens through which they can better understand the public's perception of health issues and determine how best to utilize Twitter for health promotion (Ghosh & Guha, 2013; Korda & Itani, 2013).

Manual content annotation and computational models have been used to analyze the discourse of health on Twitter. Studies that utilize manual content analysis have looked at health issues such as swine flu (Chew & Eysenbach, 2010), dental pain (Heaivilin, Gerbert, Page, & Gibbs, 2011), concussions (Sullivan et al., 2012), breast cancer (Thackeray, Burton, Giraud-Carrier, Rollins, & Draper, 2013), and marijuana usage (Krauss, Grucza, Bierut, & Cavazos-Rehg, 2016). These studies involve content analysis of a small set of tweets (e.g., 1,000 to 10,000). Manual content analysis studies are typically time-consuming because they require the manual coding of tweets by individuals. On the other hand, computational models have been employed to analyze large samples of Twitter data in a timely manner. Some of the work has focused on sentiment analysis. Sentiment analysis is concerned with the use of natural language processing and computational linguistics to identify and extract subjective information, such as opinion, sentiment, evaluations, attitudes, and emotion from written language (Cao & Cui, 2016). Salathé and Khandelwal (2011) applied sentiment analysis to understand the perception of the H1N1 vaccine on Twitter. Myslin et al., meanwhile, used machine learning classifiers to detect sentiment and relevance for tobacco-related tweets (2013). In addition to sentiment, Cole-Lewis et al. used machine learning techniques to classify tweets based on user description, genre, theme, and relevance to the topic of e-cigarettes (2015). Existing research has focused predominantly on understanding one or two health topics on Twitter.

Our goal is to build on this research and provide insight into a variety of health issues through a visual analytic study. Visual analytics enhances the understanding of data by combining computational models with interactive visualizations (May, Hanrahan, Keim, Shneiderman, & Card, 2010; Ola & Sedig, 2014). Our study is meant to demonstrate how machine learning techniques and visualizations can be used to analyze and make sense of the discussion of health on Twitter. To this end, we retrieved over half a million health-related tweets, and randomly selected a sample of 3000 on which we conducted manual content analysis. We used the sample to create models that classified tweets based on their content and user category. These models were then applied to the larger tweet dataset. Finally, we created a visualization that allows us to explore the discourse of health issues on the social-media platform. In this paper, we report our findings and

discuss implications. The rest of the paper is organized as follows. Section 2 presents the research methods. Section 3 discusses the results. The final section, Section 4, presents general conclusions.

## 6.2 Research Methods

In this study, supervised machine learning was used to build classification models that predict the theme of a tweet and a user category for the person who posted the tweet. For our analysis, we do not include re-tweets, as we are more concerned about what is being said about certain health issues as opposed to its frequency or popularity. In addition to the tweet, Twitter allows developers to access relevant metadata about the user who posted the tweet. User information includes username, description of the account, the number of followers, the number of people the user is following, and the number of tweets the user has posted (Twitter, 2007). In this section, we describe how the data was collected and processed.

### 6.2.1 Data Collection

In the past, hashtags and search terms have been used to retrieve health-related tweets (Palomino, Taylor, Göker, Isaacs, & Warber, 2016; Paul & Dredze, 2014; Symplur, 2010). We opted to use search terms. Our initial list of search terms is comprised of causes of death identified by the Institute for Health Metrics and Evaluation (IHME) (Lozano et al., 2012). We utilized these causes as search terms primarily because this work is part of a larger research plan to facilitate sensemaking of health data and we wanted to ensure consistent terminology. IHME classifies causes into 21 cause-clusters which are aggregated into three main groups: 1) non-communicable, 2) injury-based, and 3) communicable, maternal, neonatal, and nutritional.

To get a better understanding of the ability of these terms to provide relevant tweets we collected a sample of over 50,000 tweets in December of 2015. We utilized Tweepy—a Twitter application programming interface—to search for and retrieve the tweets (Tweepy, 2009). In an iterative fashion, for each search term, we retrieved up to 200 recent tweets to determine whether the search terms predominately retrieved health-related tweets. In certain situations, search terms were combined or shortened to improve

results. For instance, our initial list included cirrhosis of the liver secondary to hepatitis B, cirrhosis of the liver secondary to hepatitis C, and cirrhosis of the liver secondary to alcohol use. These three causes were combined and the search term used was liver cirrhosis. Another search term that was adjusted was exposure to forces of nature; it was expanded to include earthquake deaths, tsunami deaths, flood deaths, and hurricane deaths. Appendix 1 includes the full final list of the 117 search terms used. Over a 1-month period between March 17, 2016 and April 17, 2016 we retrieved tweets using the search terms. The total number of unique English language tweets retrieved during this period was 535,973. The tweets were stored in a MongoDB database.

### 6.2.2 Analysis

#### 6.2.2.1 Sentiment Analysis

Similar to existing research practice we measured sentiment as being either negative, positive, or neutral (Cole-Lewis et al., 2015; Palomino et al., 2016; Salathé & Khandelwal, 2011). In our research, we utilized AlchemyAPI's sentiment analysis tool to assign polarity and sentiment value to our tweets. We selected AlchemyAPI because at the time it was one of the leading free sentiment analysis tools with a high accuracy rate (Meehan, Lunney, Curran, & McCaughey, 2013; Saif, He, & Alani, 2012; Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015). For example, in a previous study, sentiment analysis by AlchemyAPI was compared to manual testing on 5370 tweets and the accuracy rate was 86.01% (Meehan et al., 2013). When we used the product, AlchemyAPI was a text mining platform that extracted metadata such as concepts, keywords, categories, sentiment, and relations from text-based documents. The company has since been acquired by IBM, and its functionality incorporated into Watson Natural Language Understanding Service (Devarajan, 2017). For a text fragment, AlchemyAPI returns a sentiment category and score. The sentiment score is in the range (-1, 1) and expresses the strength of the sentiment. The category is based on the score value. For a score less than 0, the category is negative, for a score over 0, the category is positive, and for a score of 0, the category is neutral. Table 1 includes some of the tweets and the corresponding sentiment score and category it was assigned.

**Table 6-1: Sample of AlchemyAPI sentiment analysis**

| Tweet | Score | Category |
|---|---|---|
| Involved lymph nodes in HPV positive oropharyngeal cancer Regional control is preserved after dose de excavated | 0.0000 | neutral |
| Ambulance came In hospital with trial flutter On t like this | -0.2296 | negative |
| a treat of an occasional cup of coffee won t give you diabetes you'd have to have a lot of sugary coffee to be at risk him V xx | 0.4292 | positive |
| hi guns doing a skyline for prostate cancer can i get a shout out please birmingham fan cheers | 0.0000 | neutral |
| I think I'm donna die of drug overcome I've taking so many pills and my headache still won't go away | -0.9750 | negative |
| Share the love via CandyGram amp support to feed people affected by HIV AIDS valentinesday | 0.4615 | Positive |

#### 6.2.2.2 Manual Annotation

To obtain a better understanding of who was tweeting and the content of each tweet we performed content analysis on 500 tweets that were randomly selected from the corpus. Based on previous research (Cole-Lewis et al., 2015) and our analysis, five content themes and six categories of users were established. The five identified content themes are as follows:

- Educational: post about relevant health-related news, factoid, resource, research, or public health announcement. Tweet that contains general health information, research, or information to raise awareness on a health issue. For example,
    - "Brain cancer two essential among acids might hold key to better outcome cancer News"
    - "Preparation and Characterization of Irinotecan Loaded Cross Linked Bovine Serum Albumin Heads for River Cancer"
- Fundraising: post that seeks to raise funds or solicit money or services for a health organization, cause, or individual needing medical treatment. For example,
    - "That dollar goes to the Measles and Rubella Initiative to buy a vaccine for a child against Measles and Rubella"

    - “LETS SAVE A LIFE Baron has suffered with Throat cancer for 5 years and lung cancer for eyes Your contribution matters”

- Personal: post in which the user is giving an opinion on a health issue, reporting on their own personal health status, or asking health-related questions. For example,
    - “His bronchitis has my chest feeling heavyyyyyy”
    - “I am wheeling like an old man with asthma after a joy Thank you of”
    - “Migraine all day yet again Time to go see a Neurologist”
- Promotional: post promoting or advertising a for-profit health event or product. For example,
    - “Find out how you can prevent and reverse diabetes won The At Real Good Health Summit”
    - “Out And You The Ultimate Out Diet and Cookbook with recipe to get you started on a proper diet”
    - “Or Lane Vishnubala will be teaching our coming Of Obesity and Diabetes Specialist Instructor course”
- Unrelated: post that contains search terms but is unrelated to health. For example,
    - “I feel like I am drowning without your looooooveeeeeeeeee”
    - “Nationalism is an infantile disease It is the measles of mankind”

The user categories are as follows:

- Businesses: for-profit organizations, e.g., retailers, pharmaceutical companies, fitness companies.
- Celebrities: famous people in pop culture, politics, sports and news media.

- Interest Groups: unofficial organizations for specific health interests, e.g., school groups, health food groups, anti-vaccination groups.

- Media: reputable news source such as New York Times, Washington Post, Wall Street Journal, Associated Press and reputable journals that publish health research.

- Official Agencies: government agencies and large non-government health agencies, e.g., National Institutes of Health, Centers for Disease Control and Prevention, American Heart Association.

- Public: general public that does not fall into one of the aforementioned categories.

Four analysts independently coded a subset of 3000 tweets. The classification data are presented in Table 2. The predominant user category is public with 2264 tweets which account for 75.5% of the total number of tweets. For the themes, the most predominant theme is education with 45.7%. Of the coded tweets, 74.3% of the tweets were found to be health-related tweets. In the next section, we describe how classification models were built.

**Table 6-2: Categorization of tweets by user and content**

| User | | Content | |
|---|---|---|---|
| Category | Frequency | Theme | Frequency |
| Public | 2264 (75.5%) | Educational | 1370 (45.7%) |
| Interest Groups | 227 (7.6%) | Personal | 770 (25.7%) |
| Media | 227 (7.6%) | Unrelated | 761 (25.3%) |
| Businesses | 215 (7.2%) | Promotional | 66 (2.2%) |
| Celebrities | 40 (1.3%) | Fundraising | 33 (1.1%) |
| Official Agencies | 27 (0.9%) | | |

### 6.2.2.3 Model Construction

Our models were constructed with the Scikit Learn library (version 0.17.1) for Python (version 3.5.2). We used the Bag of Words approach to extract numerical features from text content. The Bag of Words approach is comprised of three main parts. The first is tokenization, which involves splitting each document (i.e., tweet or text) into words based on whitespace and punctuation. Next, the occurrences of each word are counted and

stored in a matrix. The third part of the strategy involves normalizing and weighting the occurrences. Normalization is important because when dealing with a large corpus, common words like ‘a’ and ‘the’ which frequently appear typically convey little meaningful information about the content of the document. Re-weighting was done with the term frequency-inverse document frequency transform (tf-idf), which helps to measure how important a word is to a document in a collection by taking into consideration the number of times a word appears in a document and the frequency of the word across the entire corpus (Silge & Robinson, 2017). In the following subsections, we discuss how models were constructed for the user category and the content themes.

#### 6.2.2.3.1 User Category

As previous research points to the benefits of using Support Vector Machines for short text (e.g., tweets), we utilized this technique (Cole-Lewis et al., 2015; Myslín et al., 2013). Variations of the classification technique were used based on the following attributes:

- **User description:** user-provided string that describes their account (e.g., “United Nations Development Programme helps empower lives & build resilient nations. To learn more, follow @ASteiner & visit: http://www.undp.org”).

- **User verified:** indicates whether the account has been deemed authentic by Twitter. Twitter authenticates an account so that the public is aware that the account holder’s identity has been verified. This is typically done for individuals in the entertainment, government, religious, news, business, or sports industries.

- **User screen name:** unique user name or handle name that is used to identify the tweeter, typically preceded by the @ symbol in tweets (e.g., @UNDP, @WHO, @UNICEF).

- **Influence score:** this attribute helps determine how influential an account is on Twitter. Past research notes that influence is not solely based on the number of people that follow you on Twitter but is also affected by the number of people you follow (Anger & Kittl, 2011). The score is calculated by dividing the number

of followers by the number of people that the account followers. For instance, for @UNDP the number of followers is 1.13 million while the following is 4656. The influence score is 242.70.

Table 3 shows the average accuracy rate for 100 runs for four different models. Accuracy rate is defined as the percentage of observations that were correctly classified in the test dataset. 80% of the coded data was used to train the model while 20% was used to test the model. The experiment was run 100 times for each of the models created. The model with the highest accuracy rate was Model A1, which used the user description alone. Subsequent models that incorporate the username, influence score, and verified status of the account resulted in lower accuracy rates.

**Table 6-3: Accuracy rate for user category model construction**

| Model | Average Accuracy Rate (%) |
|---|---|
| A1: description | 86.86 |
| B1: description and screen name | 79.83 |
| C1: description, name, and influence score | 79.84 |
| D1: description, name, influence score, and verified | 79.75 |

### 6.2.2.3.2 Tweet Theme

Machine learning models were built for the tweet theme based on the tweet text and user verified status. We used a Bag-of-Words approach and Support Vector Machine technique for our models. The first model uses the tweet, the second model uses the tweet text as well as the number of reserved news words (e.g., newspaper, news, official), the third model uses the tweet and the verification status of the tweeter's account, and the last model uses the tweet, the verification status, and the number of reserved news keywords. Table 4 shows the average accuracy rate for the tweet themes for the four models.

**Table 6-4: Accuracy rate for tweet theme model construction**

| Model | Average Accuracy Rate (%) |
|---|---|
| A2: tweet | 80.99 |
| B2: tweet and count of reserved keywords | 81.09 |
| C2: tweet and user verification status | 81.14 |
| D2: tweet, count or reserved keywords and user verification status | 81.44 |

Based on the experimental analysis of model construction, we used Model A1 and Model D2 to classify the entire tweet corpus. 24% of the tweets were classified as unrelated and were removed. In the next section, we discuss the results of the remaining tweets.

## 6.3 Results

A total of 416,900 tweets remained in our corpus after unrelated tweets were removed. These tweets represent 117 different causes that contribute to mortality. Each tweet also has a sentiment score and type, category for the user who sent the tweet, and content theme. In this section, we first present a brief overview of the results, describe the design of a visualization we created to facilitate making sense of the discourse of health on Twitter, and then highlight results for certain cause-clusters.

Table 5 shows the frequency of tweets categorized by sentiment, theme, and user group. 73% of the tweets were deemed negative, while 27% of the tweets were either positive or neutral. Similar to the manually coded data, the majority of tweets in our corpus were tweeted by the general public (83.5%). The tweets by the media and official agencies made up less than 5% of corpus. This is important to note because for the general public, they may assume that a significant portion of the information they are reading is from reputable sources, which is not the case. In terms of the content, 66% of the tweets were educational tweets, while personal themed tweets made up 33% of the corpus. Combined, fundraising and promotional tweets were less than 1 percent.

**Table 6-5: Frequency for sentiment, theme, and user categories**

| Sentiment | Frequency (%) | Theme | Frequency (%) | User | Frequency (%) |
|---|---|---|---|---|---|
| negative | 72.85 | educational | 65.99 | businesses | 4.98 |
| neutral | 14.47 | fundraising | 0.16 | celebrities | 0.01 |
| positive | 12.68 | personal | 33.62 | interest groups | 6.71 |
| | | promotional | 0.23 | media | 4.73 |
| | | | | official agencies | 0.04 |
| | | | | public | 83.52 |

The visualization described in this section is part of a tool that allows users to explore causes and risk factors from multiple perspectives, including geography, demography, and chronology (Ola & Sedig, 2016). In addition to the metrics previously discussed, our visualization includes prevalent words (i.e., non-search terms that frequently appear in the corpus) and the net sentiment rate for causes as well as for clusters of causes. In the context of tweets, net sentiment rate is defined as the subtraction of the number of negative tweets from the number of positive tweets divided by the total number of tweets.

$$Net\ Sentiment\ Rate = \frac{number\ of\ positive\ tweets - number\ of\ negative\ tweets}{total\ number\ of\ tweets}$$

The visualization, depicted in Figure 1, has three main parts. The first part is comprised of circular arcs that frame the rest of the visualization. These arcs represent the top 50 words across the entire corpus. The size and location of each arc depict its prevalence. The larger the arc, the more times it appeared in the corpus. By hovering over the arc (i.e., a word), the number of occurrences appears. The arcs are arranged from left to right in descending order based on prevalence. As shown, the words get, health, like, women, may, type, and new are frequent words in the corpus.

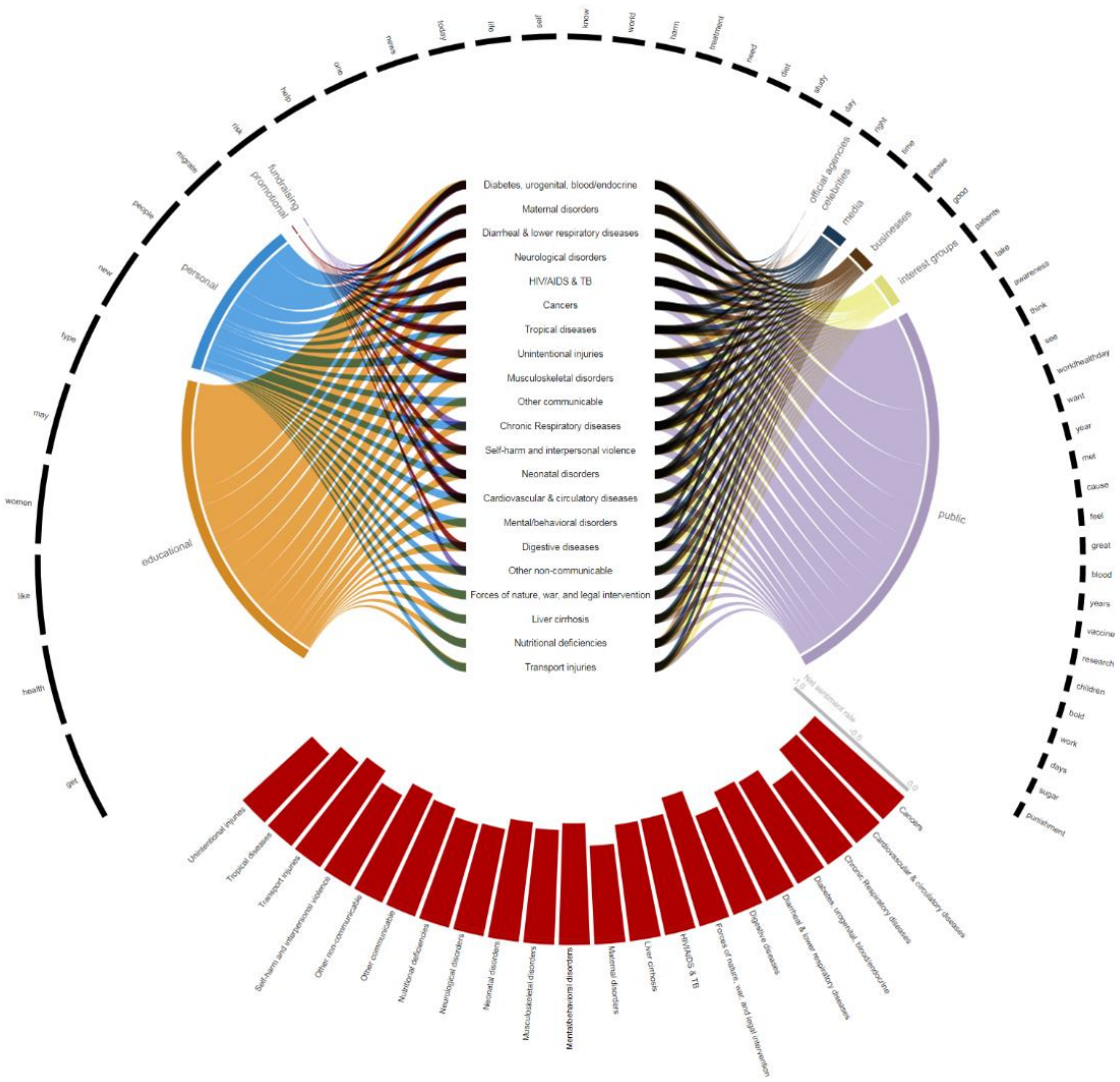



**Figure 6-1: Default configuration of the sentiment visualization**

Some of the screenshots used in the figures only include partial representations of the entire visualization; this is done so as to aid in the reading of the textual content in the visualization. The central portion of the visualization (see Figure 1) depicts the breakdown of tweets by cause-clusters, user category, and tweet theme. In the center of the visualization is a list of the 21 cause-clusters arranged in descending order according to the number of tweets. The diabetes, urogenital, blood/endocrine cluster has the most number of tweets in the corpus, while the transport injuries cluster has the least. On the left side of the cluster list is a sub-visualization of the tweets by content themes. The links

that branch out of each theme represent the presence of tweets for a cause-cluster. For instance, when the promotional theme is selected, users of the tool can observe that there are 13 links (see Figure 2). This is because in our tweet corpus, there are only promotional-related tweets from causes in 13 clusters. The clusters that do not have promotional-themed tweets are greyed out.

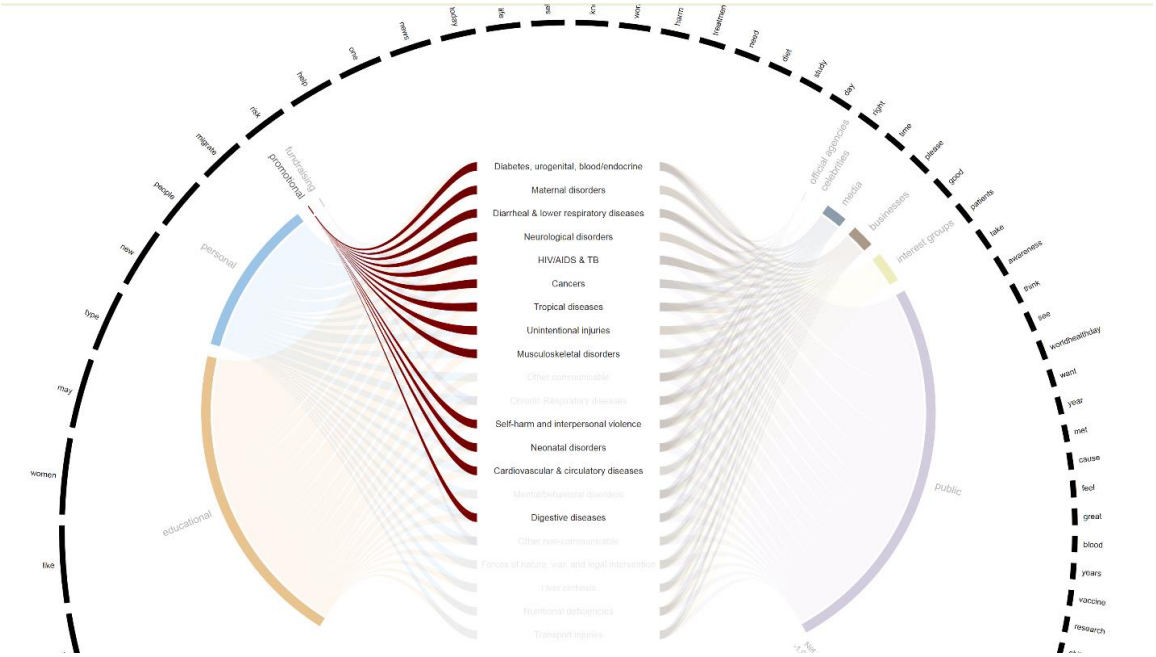



**Figure 6-2: Screenshot of sentiment visualization with promotional theme selected**

The right sub-visualization which shows the breakdown of tweets by user categories is encoded in a similar fashion. For instance, Figure 3 shows the state of the visualization when the celebrities user category is selected. It is worth mentioning that the content themes and user categories are arranged based on the number of tweets. In other words, we use both size and location to encode quantity so that users do not have to strain to determine which group is bigger. For example, for the user categories (see Figure 1), the media (4.73%) and businesses (4.98%) arcs appear to be the same size but because the arcs are ordered, users of the tool can deduce that the businesses category has more tweets.


**Figure 6-3: Screenshot of sentiment visualization with the celebrities user category selected**

The lower portion of the visualization has two alternating views. The first view is shown in Figure 1 and it depicts the net sentiment rate for cause-clusters. The second sub-visualization depicts sentiment for the causes that make up a specific cluster. This sub-visualization contains curved heatmaps and is divided into two parts. The first part shows the breakdown of sentiment by the user categories and the second part by the theme of the tweets. The sections of the heatmap are encoded with color, where red is used to indicate negative polarity, green for positive, and grey is used to depict the absence of data. For instance, as shown in Figure 4 when the cardiovascular & circulatory diseases cluster is selected, the visualization shows that there are no tweets from official agencies or celebrities for all the causes that make up the cluster. In addition, the atrial flutter, hemorrhagic stroke, cardiomyopathy, and peripheral arterial disease causes have a net sentiment score that is positive for certain themes and user categories. With this visualization, users can explore the sentiment for different causes and cause-clusters, learn about the different user groups that tweet and also get a sense of what those tweets are about.

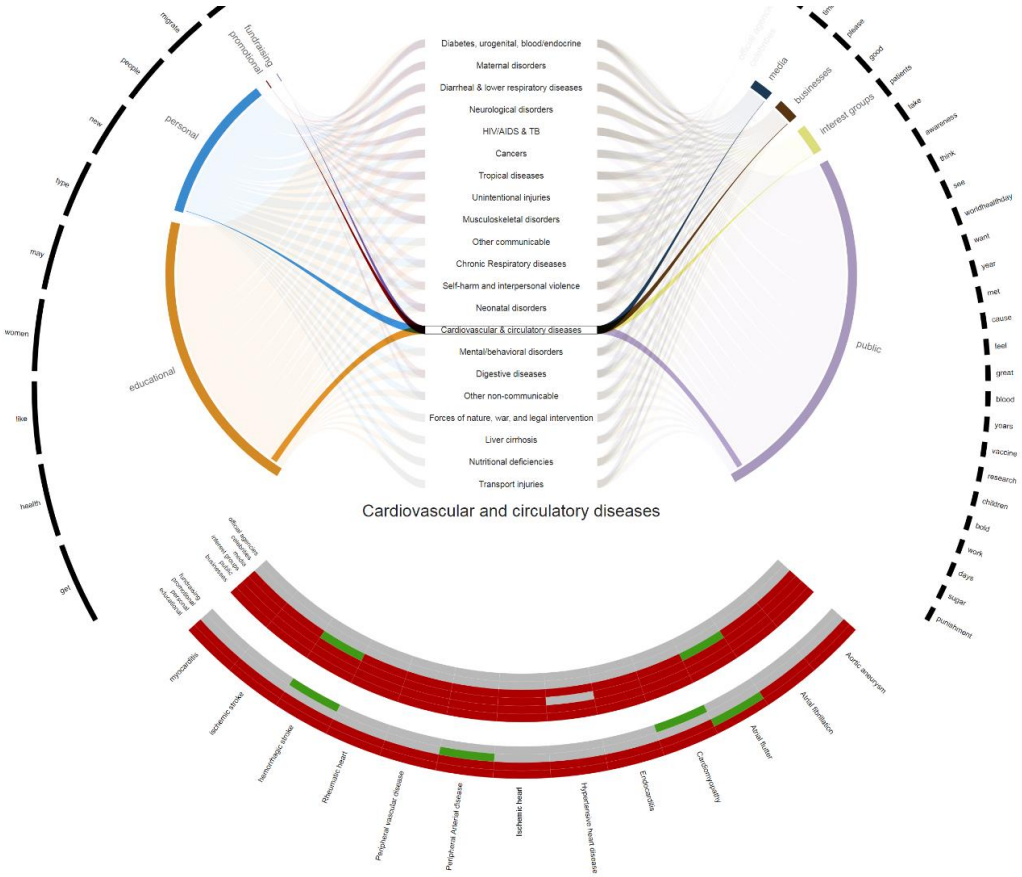



**Figure 6-4: Screenshot of sentiment visualization with the cardiovascular & circulatory diseases cluster selected**

Now that the visualization has been described let us take a close look on how it aids in the understanding of the discourse on health issues. In particular, we will focus on the HIV/AIDS&TB, mental and behavioral disorders, and neglected tropical diseases clusters. Figure 5a depicts the breakdown of tweets for the HIV/AIDS&TB cluster by user category and content theme. This cluster is one of the few clusters in which tweets on all four content themes are present in the corpus. In addition, all user categories are tweeting on at least one cause in this cluster. Figure 5b depicts the sentiment across the various categories. With this sub-visualization, one is able to notice that the tweet corpus does not include any tweets from celebrities on tuberculosis, but the discussion on HIV/AIDS includes all user groups. Another observation is that for promotional and fundraising tweets, the sentiment is positive for both HIV/AIDS and tuberculosis. It may

seem intuitive that promotional and fundraising tweets are more positive, but the same pattern is not observed for other cause-clusters.

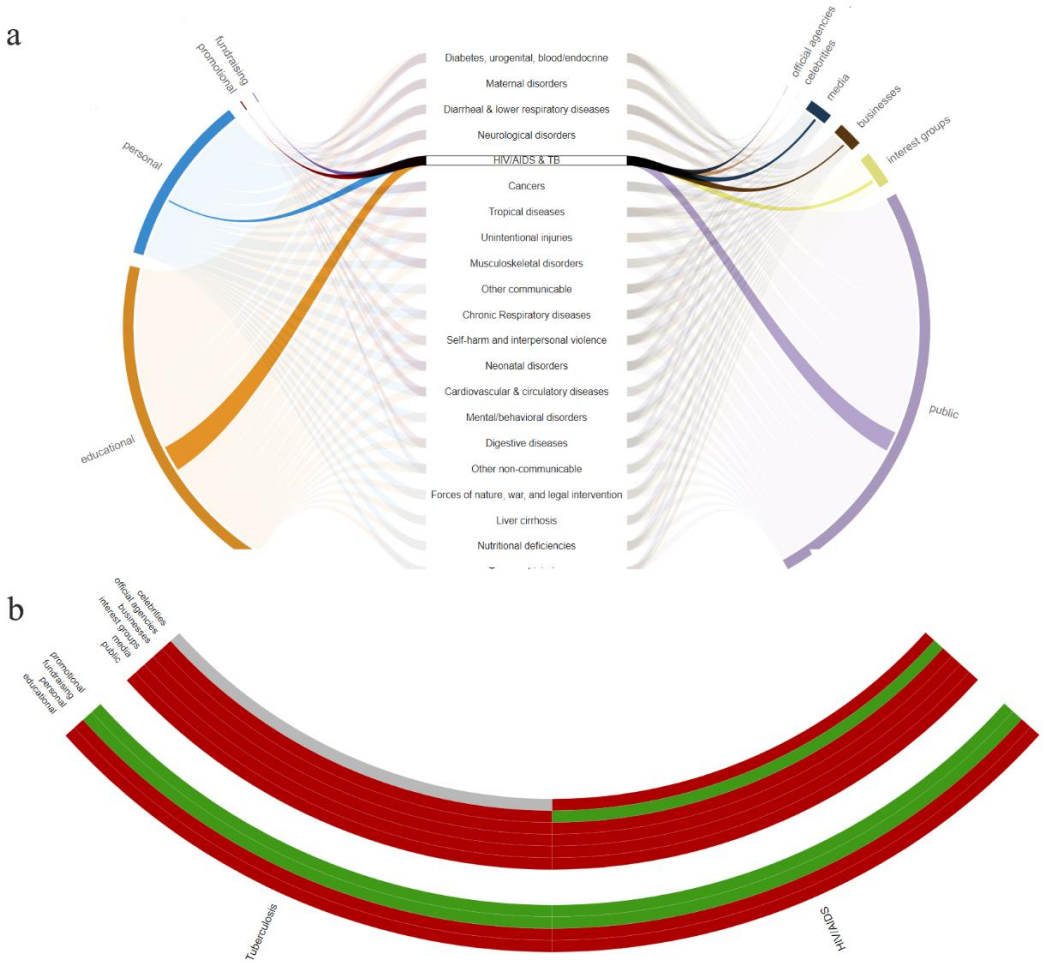



**Figure 6-5: (a-b) Screenshots of sentiment visualization with the HIV/AIDS & TB cluster selected**

For the mental and behavioral cause-cluster, the tweets in the corpus do not include fundraising- and promotional-themed tweets. Furthermore, official agencies are not tweeting on alcohol use disorders, but they are tweeting on drug use disorders. Figure 6a shows the lower portion of the visualization when the mental and behavioral cause-cluster is selected. Another observation worth highlighting is the positive net sentiment of alcohol use tweets and the negative sentiment of drug use tweets by personal accounts. This finding corroborates a recent content analysis study that noted a preference for

marijuana use over drinking alcohol (Krauss et al., 2016). The discussion on tropical diseases such as malaria, dengue, ebola, and chikungunya is highly varied. Figure 6b depicts the net sentiment rate for tropical diseases. The sentiment for the discussion of Ebola is mostly positive. This may seem erroneous, given the 2014-15 outbreak that resulted in thousands of deaths. But it is important to remember that our corpus includes twitter chatter from March – April in 2016. This coincides with a statement released by the World Health Organization in which the public emergency alert raised because of the outbreak was terminated (WHO, 2016). Our study provides a cross-sectional analysis of the discussion on Twitter for a broad range of health issues for a limited time frame. Next steps would be to provide real-time analysis that includes historical data so that users can understand the discussion of health issues and how it changes over time.

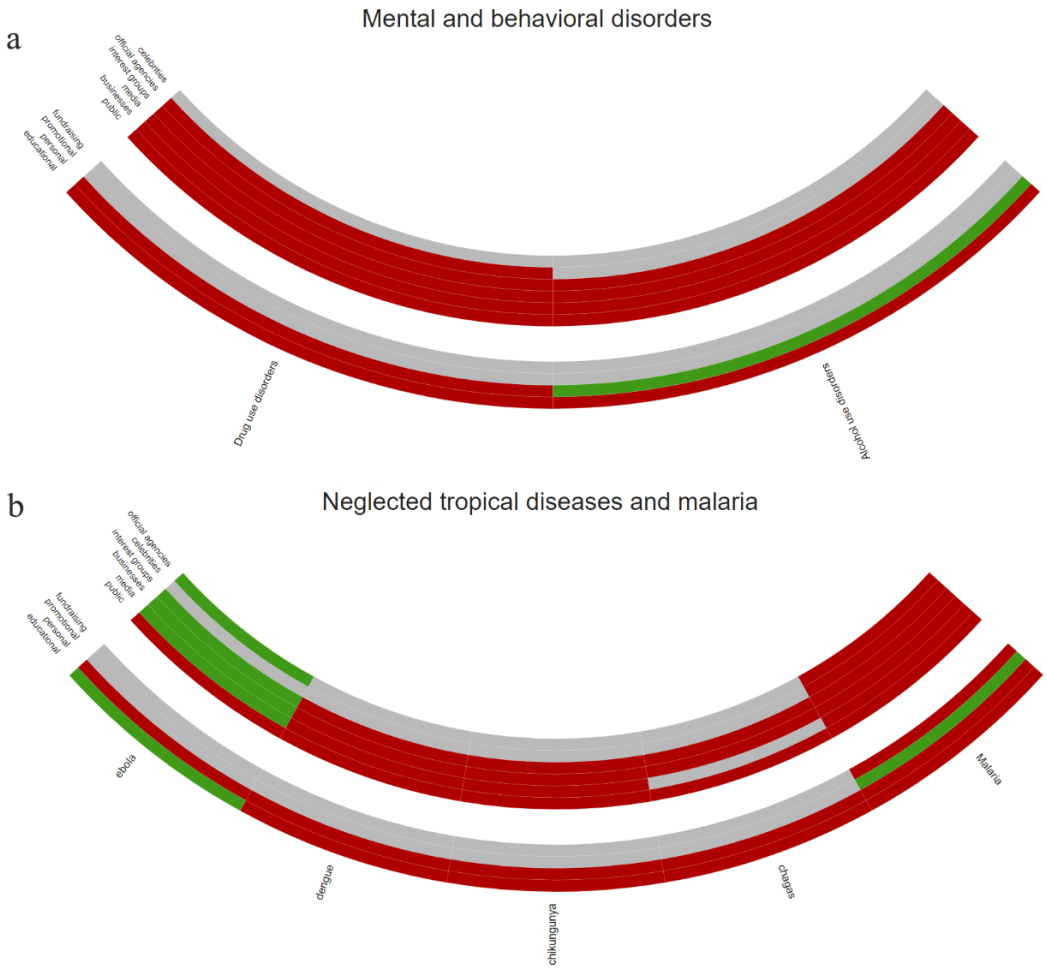



**Figure 6-6: (a-b) Screenshots of sentiment visualization with the mental & behavioral cluster and the neglected tropical diseases cluster selected**

## 6.4 Discussion and Conclusion

This paper has presented a visual analytic study that contributes to the growing body of literature on understanding how health issues are portrayed on social media platforms. The main contributions of the study are a demonstration of how supervised machine learning methods can be combined with interactive visualizations to make sense of the discourse of health issues on Twitter. In this research, we analyzed over half a million tweets based on 117 unique search terms. Although we tried to apply as much rigor as possible, certain limitations exist. First, we only utilized one month of data in 2016 which may have resulted in certain health issues being oversampled and others being

undersampled. Future studies can examine the discourse for longer periods. Secondly, we only retrieved English-language tweets. As a result, our findings cannot be generalized to other languages. Despite this limitation, we did not specify a geographical location, and consequently, our analysis may be relevant in countries in which English is widely used. It is worth mentioning that our analysis is of the discourse on Twitter and as Twitter is not widely used across all demographics, our study cannot be generalized to be a true reflection of the entire public discourse on health issues. Lastly, our constructed classification models are based on manual content analysis, which may be subject to bias.

Despite the limitations mentioned above, findings emerge from this study. The online-discourse on health topics is largely mediated by the public. This indicates that Twitter is a platform that can be used for health promotion as it is currently being used predominantly by the public to discuss health issues. The discourse is largely on educational materials such as information on treatments and news reports on health ailments. For health professionals and policy makers, the fact that the public plays a significant role and that the majority of the content is educational in nature may present challenges for health promotion efforts. The recent use of Twitter to spread misinformation on yellow fever and Ebola outbreaks across the globe highlights this issue (Ortiz-Martínez & Jiménez-Arcia, 2017; Oyeyemi, Gabarron, & Wynn, 2014). More research is needed to determine the influence (i.e., reach of tweets) for different types of users. Official health and news agencies which typically provide reputable data are largely underrepresented in the discussion. While efforts exist to use social media platforms for health education, our research highlights that there is still more work to be done. Though educational tweets make up 66% of the tweet corpus, most of these tweets come from the general public and not reputable health organizations. These findings corroborate research that health organizations are yet to effectively use Twitter to educate or engage in dialogue with the general public (Gurman & Clark, 2016). Overall, we expect that our findings can improve the ethical use of Twitter data by equipping individuals with the ability to weigh information based on the reputability of the source of the tweet. In addition, the computational model for classifying themes emphasizes the ability to use machine learning techniques to understand the content of tweets for a wide range of health issues.

In conclusion, this work is an important step in understanding the health discourse on Twitter. Findings from this study highlight the need for future studies to understand the reach of content by various user groups. Our work demonstrates the efficacy of a visual analytic approach to making sense of social media data. Furthermore, it provides a foundation on which further research that involves real-time analysis of Twitter data can be built upon. It also provides the general public with a way to understand which topics are being discussed and by whom, which has implications for health literacy. Furthermore, this research provides a reference point for public health officials engaged in using social media to promote health policies.

# Chapter 7

# 7 Health literacy for the General Public: Making a Case for Non-trivial Visualizations

This chapter has been accepted to the Informatics journal.

Please note that the format has been changed to match the format of the dissertation. Figure numbers mentioned herein are relative to the chapter number. For instance, "Figure 1" corresponds to Figure 7-1. Additionally, when the term "paper" or "article" is used, it refers to this particular chapter.

## 7.1 Introduction and Rationale

Health literacy can be defined as an individual's ability to make health decisions based on a sound analysis of relevant data. Over the last few decades, health literacy has garnered attention across the world. This in part is due to research that suggests that health literacy is a key determinant of health. For instance, according to the American Medical Association, health literacy is a stronger predictor of a person's health than age, income, employment status, education level, or race ("Health literacy: report of the Council on Scientific Affairs. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association.," 1999). A survey conducted across eight European countries notes that individuals with lower levels of health literacy tend to have worse health (Sørensen et al., 2015). In addition to the health implications, low health literacy has financial implications for individuals as well as governments (Kickbusch et al., 2013; Rasu, Bawa, Suminski, Snella, & Warady, 2015; Vernon, Trujillo, Rosenbaum, & Debuono, 2007).

Health literacy is multifaceted and encompasses a person's ability to access, understand, process, and apply health information relevant to disease prevention, healthcare, and health promotion (Sørensen et al., 2012). Disease prevention is an important aspect of public health (National Research Council, 1988). In 2000, 35% of deaths in the United States were linked to tobacco and alcohol use, poor diet, and physical inactivity (Mokdad,

Marks, Stroup, & Gerberding, 2004) . On a global scale, 10% of mortality is attributed to physical inactivity and dietary risk factors (Lim et al., 2012). From a disease prevention standpoint, individuals with low health literacy have been shown to make poor health choices, engage in risky behavior, and have low self-management (Kickbusch et al., 2013). Though professionals are charged with educating the public about health risks, hazards, and issues, there is need for personal empowerment as well (Lemire, Sicotte, & Paré, 2008; Nutbeam, 2000; Schulz & Nakamoto, 2013). Improving health literacy is a nontrivial endeavor. Currently, individuals seeking to access and understand health data are confronted with a myriad of data-related challenges. For instance, health data is often voluminous and originates from heterogeneous sources (Gotz & Borland, 2016; Herland et al., 2014; E. Liu et al., 2016; Ola & Sedig, 2014; Shneiderman et al., 2013). As a result, people find themselves having to engage in a time-consuming traversal of multiple websites to access relevant data. In addition to access, presenting data to individuals in a dense and understandable fashion is crucial to improving health literacy for the public (E. Liu et al., 2016). Given the scale and complexity of the data related to disease prevention, visualizations have the potential to play a crucial role.

Interactive visualizations predominately represent data in a visual format and allow users to manipulate how the data is shown. Simple visualizations such as bar charts, scatter plots, and pie charts have been used extensively over the last two centuries in the health domain. However, as the size of data increases, there is a need for visualizations that can mirror the complexity of the data and facilitate its understanding without straining the cognitive resources of users (Ola & Sedig, 2016). While the development of elaborate non-trivial visualizations has increased in recent years, research on instructional materials for visualizations is sparse (Lee et al., 2015; Ruchikachorn & Mueller, 2015; Tanahashi, Leaf, & Ma, 2016). As users' understanding of the tool influences their ability to use the tool to complete tasks effectively, more research on visualization literacy—which is the ability of users to interpret and extract information from visualizations—is necessary. Borner et al. highlight the need for instruction so that individuals are better equipped to understand novel visualizations (Borner et al., 2016). While some may avoid using non-typical visualizations because of their complexity, it is important to investigate if with training individuals can learn to use such visualizations. Therefore, before we can

explore the use of non-typical visualizations for health literacy, it is important to first examine visualization literacy.

The purpose of this paper is twofold. First, to present research that investigates the ability of individuals to learn to use elaborate interactive visualizations. Second, to examine the ability of non-trivial visualizations to improve health literacy. To this end, we have created a visualization tool, HealthConfection, that allows individuals to make sense of the causes and risk factors that contribute to mortality across the world. Using this tool, we have conducted two user studies. The results from the first study, which is for visualization literacy, informs the second study that investigates health literacy. In this paper, we report our findings and discuss the implications for the visualization and health communities. The rest of the paper is organized as follows. Section 2 provides some conceptual and terminological background. Section 3 describes the visualization tool that we have created. Section 4 presents the research methodology and results from the visualization literacy study. Section 5 presents the health literacy study that we conducted. The final section, Section 6, presents the general conclusions.

## 7.2 Background

### 7.2.1 Health Literacy

Health literacy is concerned with the ability of an individual to access, read, and understand health information, and act based on that information (Sørensen et al., 2012). Health literacy is a public health imperative (Gazmararian et al., 2005; Kickbusch et al., 2013). Studies indicate that individuals with low health literacy are at a greater risk of long-term and life-limiting health conditions, as well as earlier mortality (Berkman et al., 2011; Bostock & Steptoe, 2012). Individuals with low health literacy are less likely to be able to make sense of information related to clinical issues, risk factors, and social and physical determinants of health. In addition to the individual repercussions, low health literacy increases healthcare utilization and expenditure (Rasu et al., 2015). A 2007 report estimates that the cost of low health literacy to the U.S. economy was between $106 and $238 billion each year (Vernon et al., 2007). Advancing health literacy may also lead to

more equity and sustainability of changes in public health (Rowlands, Shaw, Jaswal, Smith, & Harpham, 2017; Sørensen et al., 2012).

In this paper, we focus on disease prevention. From a health literacy standpoint, individuals need to be able to access, understand, and interpret information on risk factors for health (Sørensen et al., 2012). Disease prevention data is sourced from hospital records, demographic and health surveys, mortality reports, and research studies. Even after the data has been aggregated, individuals typically need to traverse multiple text-based tables to find information. To understand the causes that lead to mortality and the implications of certain risk factors is an exploratory process, in which individuals need to be able to ask questions, get answers, and observe trends. In other words, they need to be able to interact with the data seamlessly. While videos and infographics have been beneficial in helping to improve health literacy (Occa & Suggs, 2016), when it comes to large sets of data there is a need for tools that allow users to control the flow of data and how data is represented.

### 7.2.2 Visualizations for Health Literacy

Visualizations, otherwise known as visual representations, have been used in varying capacities to help promote the understanding of health data. In the mid-19$^{th}$ century, Florence Nightingale used the coxcomb to visualize patient data and educate the Crown on sanitation related deaths of soldiers during the Crimean War (B. Cohen, 1984). Visualizations have evolved in complexity both with respect to how data is represented and how users can interact with the data. On one hand, simple visualizations, such as bar charts and scatter plots, are being replaced with visualizations that allow users to encode multiple aspects of the data simultaneously (Ola & Sedig, 2016). On the other hand, static visualizations are being replaced with interactive ones that allow users to control how and what data is shown at a specific point in time. In this section, we highlight some of the recent work aimed at providing the public with an accessible manner to make sense of health data.

HealthMap provides a comprehensive view of the current global state of infectious diseases by bringing together disparate data sources (Freifeld et al., 2008). Health GeoJunction extracts textual information from scientific literature, PH reports, and news reports to support the discovery of relationships between documents (MacEachren, Stryker, Turton, & Pezanowski, 2010). Weave is a web-based analysis and visualization environment that has been used to facilitate the exploration of breast and ovarian cancer data (Purushe, Grinstein, Smrtic, & Lyons, 2011). Community Health Map allows users to explore and compare the health-care indicators across counties in the United States (Sopan et al., 2012). Zhao et al. (Zhao et al., 2013) integrate ringmaps into the InstantAtlas software environment to explore complex socio-spatial patterns of cardiovascular disease in New Zealand. Their tool supports the exploration of cardiovascular disease at multiple levels of granularity. Liu et al. (E. Liu et al., 2016) have developed a tool that allows patients to visualize data from PubMed on cardiorenal disease and its comorbidities as well as patient data from wearable sensors.

While existing research has advanced the use of visualization tools to make sense of health data, most tools typically focus on a specific disease or viewpoint. For instance, the tool by Zhao et al. focuses solely on cardiovascular diseases. Similarly, HealthMap supports heterogeneous data sources, but only for one group of diseases—infectious diseases. One notable exception is the suite of visualizations created by the Institute for Health Metrics and Evaluation (Institute for Health Metrics and Evaluation, 2013). Our visualization prototype, HealthConfection, which will be described in Section 3, builds on existing research and seeks to advance the use of visualizations for health literacy.

### 7.2.3 Visualization Literacy

Visualization literacy has been defined as the ability and skill to read, interpret, and extract information from visualizations (Lee, Kim, & Kwon, 2017). How people learn to use a visualization can influence their ability to understand the underlying data and complete tasks with the tool (Tory & Möller, 2004). A study that involved 273 participants and 20 common visualizations provides strong evidence that a very high

proportion of adults and youth have low visualization literacy (Borner et al., 2016). Although users can improve visualization literacy through trial-and-error processes, past research indicates that sometimes when a faulty conceptualization of a visualization is formed users tend not to revise that conceptualization (Lee et al., 2015). If users do not know how to properly use a visualization, they are less likely to use it and may abandon the information-seeking tasks entirely if they become frustrated. To support information-seeking behavior, it is necessary to provide users with tools that support, rather than hinder, their tasks.

More work on empowering individuals to understand visualizations is needed (Borner et al., 2016). The visualization community recognizes this and is taking steps to improve visualization literacy within the general public. Recent efforts to improve visualization literacy investigate how instructional materials should be designed (Alper, Riche, Chevalier, Boy, & Sezgin, 2017; Kwon & Lee, 2016; Ruchikachorn & Mueller, 2015; Tanahashi et al., 2016). Ruchikachorn and Mueller demonstrated that by morphing visualizations from the familiar to the unfamiliar, participants could learn new representational forms (Ruchikachorn & Mueller, 2015). Alper et al. (Alper et al., 2017) have developed an online platform for children in grades K to 4 to learn about pictographs and bar charts. Tanahashi et al. (Tanahashi et al., 2016) investigated the top-down and bottom-up teaching methods, and active or passive learning types for the scatter plot, graph, storyline, and treemap. In general, they observed that participants who used the instructional materials that utilized the top-down teaching method and catered to active learning showed the greatest improvement in the test segment. Kwon and Lee further studied active learning strategies. Using the parallel coordinates visualization and three tutorials types: static, video, and interactive, they observed that participants with the interactive and video tutorials outperformed participants with static or no tutorials (Kwon & Lee, 2016). Some of the studies mentioned above have focused on simple visualizations, while others have investigated visualization literacy for static visualizations. Our research builds on this foundation and explores the impact of video tutorials for complex, sophisticated interactive visualizations.

## 7.3 HealthConfection

HealthConfection is a visualization tool that allows users to explore and make sense of the risk factors and the causes of mortality. The tool incorporates selected datasets aggregated by IHME (Institute for Health Metrics and Evaluation, 2013). The datasets include over 12 million records that estimate the 57 risk factors and over 235 causes that lead to death. Part of the challenge when working with large datasets is determining how users will explore the data. In visualizations, providing an overview is beneficial. When properly designed, overviews can provide users with an immediate appreciation for the size and extent of the data space, and support the navigation and exploration of the data space (Hornbæk & Hertzum, 2011). Previous visualization tools have shown the importance of providing users with a high-level overview of the data (Purushe et al., 2011; Zhao et al., 2013). In addition to creating an overview visualization, we have also developed visualizations that emphasize four different perspectives through which users improve their health literacy: demography, geography, chronology, and sentiment.

When working with multiple visualizations, it is important to provide users with consistent structures and navigational cues and anchors (Hornbæk & Hertzum, 2011; R Spence, 2014). As users navigate a data-centered tool, they find themselves confronted with familiar questions, including *where am I? where can I go? and how do I get there?* Visual metaphors can help to provide consistent structures. When users internalize visual metaphors, they can navigate visualizations effectively (Ziemkiewicz & Kosara, 2007). One technique to organize several representations is to use the visual confection metaphor. A visual confection is an assembly of visual representations, juxtaposed to tell a story, present visual comparisons, and show relationships and transitions (Tufte, 1997). Confections focus on the organization of representations through compartments, which can then be used to zoom in on visual elements. The consistent structure and navigation allow users always to be aware of their current location. Based on the Gestalt principle of symmetry, one viable technique for juxtaposing visual confections is to have a central representation around which other representations are arranged (Gadanidis, Sedig, & Liang, 2004). Placing a representation at the center implies that the representations surrounding it are conceptually related to it (Gadanidis et al., 2004). The central

representation, then, is where users begin their exploration of *the story of the data*. Figure 1 shows the visual organization of our tool.

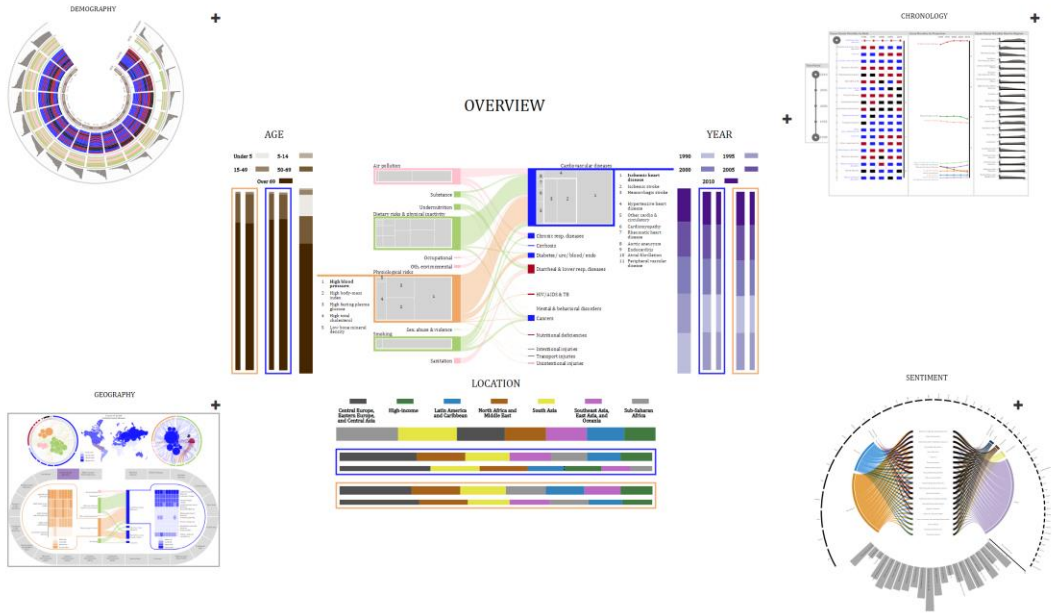



**Figure 7-1: HealthConfection visualization tool**

HealthConfection provides cues that allow users to explore health data from different perspectives while at the same time minimizing visual discontinuity. By interacting with the '+' anchor to the right of each compartment, users can explore a perspective, control which visualization is in the center, watch the tutorial, and hide other visualizations. The *Overview* visualization in Figure 1 shows the relationships between causes of death and risk factors at a global level and allows users to select specific age groups, geographic locations, or points in time for investigation. The surrounding compartments allow users to explore *the story of the data* from the four perspectives. In the IHME datasets, causes and risk factors are grouped at the level of clusters and groups. For causes, there are 21 clusters and three groups: communicable, non-communicable and injury. For risk factors, there are ten clusters and three groups: metabolic, behavioral, and environmental and occupational risks. In our visualizations, we use a consistent color coding to emphasize the hierarchical structure of causes and risks. Non-*communicable*, *communicable*, and

*injury* causes are encoded with *blue*, *red*, and *black* respectively. For the risk groups, we use light shades of *orange*, *green*, and *pink* for *metabolic*, *behavioral*, and *environmental and occupational* risk groups, respectively.

The *Demography* visualization allows users to explore which risks and causes affect different age groups. It also ranks the regions of the world based on their mortality rate for each age group. The visualization, enlarged in Figure 2a, has five main components, four of which are arranged as tracks. The innermost track represents the age groups at which the data is aggregated (e.g. 1-4, 50-54). The second track depicts the ranking of cause-clusters for each age group. Clusters are arranged in descending order, with the cause-cluster with the highest rank on the outside. The third track depicts the ranking of risk-clusters. The gray circles in the cause or risk tracks depict clusters that do not contribute to mortality for the age group. The last track shows the ranking of location clusters. Risk, cause, and location clusters are ranked and arranged according to their mortality rate per 100,000 people. The sub-visualization placed in the center of the tracks depicts the relationship between causes and risks for specific locations for a specific age group. The Demography visualization is a dense visualization that encodes over 800 data items in its initial configuration. Through interaction, users can control the amount of data shown and perform a variety of tasks. For instance, users can filter to understand how a risk-cluster affects different age groups. Users can also search for a specific cluster and then drill to get more information on the causes or risk factors that make up that cluster.

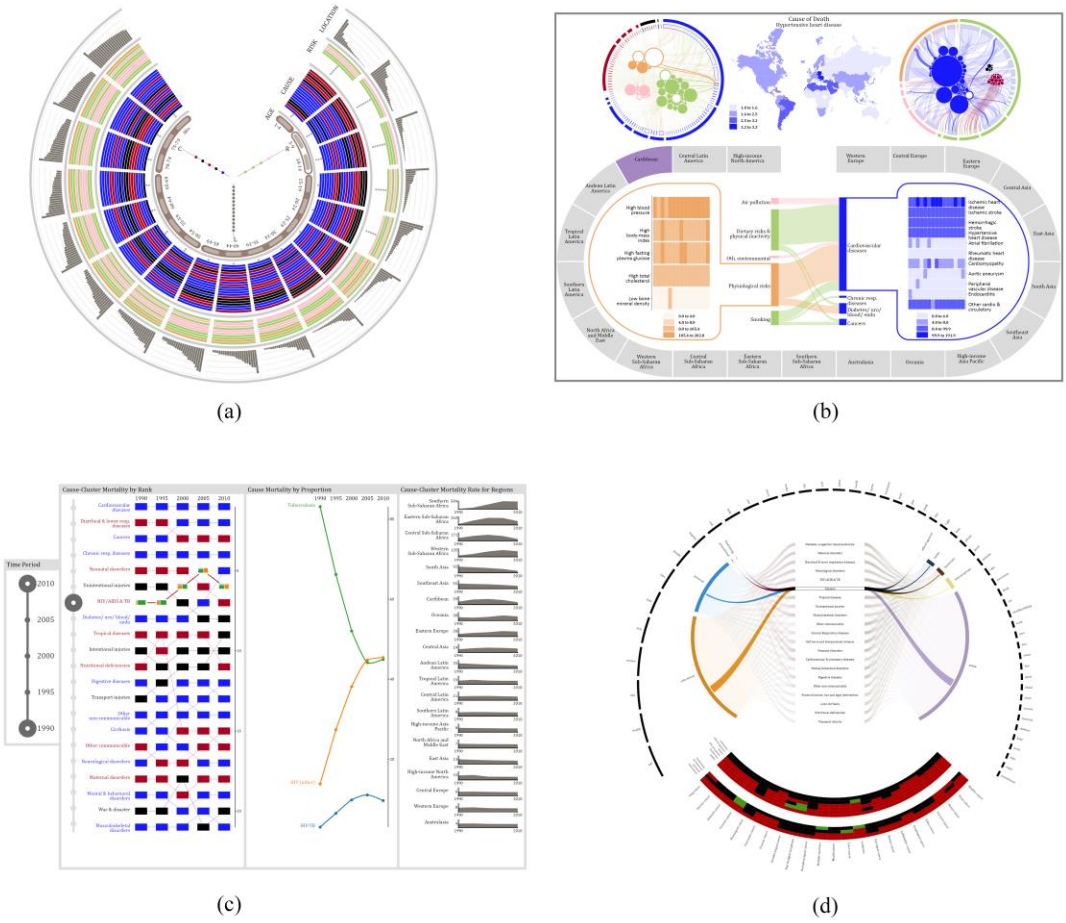



**Figure 7-2: (a) Demography visualization; (b) Geography visualization; (c) Chronology visualization; (d) Sentiment visualization**

The *Geography* visualization (Figure 2b) allows users to explore the relationships between causes and risk factors at three levels of granularity: global, regional, country. The top half of the visualization encodes the relationship between risk factors and causes at a global level and the regional distribution of mortality for a selected cause or risk factor. The circular sub-visualizations on either side of the map show the same relationships but from different perspectives. The left one shows risk factors as circles and the causes related to them as arcs, while the sub-visualization on the right shows causes as circles and risk factors as arcs. The map shows how a selected risk or cause affects different regions of the world. The bottom half of the visualization allows users to explore the cause-risk relationship for a specific region of the world. The oval track is comprised of 21 visual elements each representing a region. By selecting a region, cause, and risk related mortality rates are shown as heatmaps, for the countries in the region.

Connecting the risk and cause heatmap portions of the visualization are links that emphasize the relationship between cause-clusters and risk-clusters for that specific region. By interacting with the Geography visualization, users can determine the regions of the world that are most affected by a cause, cause-cluster, risk, or risk-cluster. They can also compare the impact that certain diseases have on countries and make sense of the relationship between causes and risk factors at multiple levels of granularity.

The *Chronology* visualization (Figure 2c) allows users to explore how mortality has changed over time. This visualization has two main controls and three panels. The first control allows users to filter data by selecting a specific time period. The second control is part of the first panel and allows users to select a cause-cluster for further examination. The first panel depicts the ranking of cause-clusters at a global level over the specified time frame. Each cause-cluster is arranged based on its rank for a specific year and links are drawn between each year's placement to help users understand the temporal trend. The second panel depicts the proportion of mortality for causes in a selected cluster. The third panel portrays the temporal distribution of cause-cluster specific mortality for each region of the world. With interaction, users can determine which cause-cluster results in the highest mortality at a global level and explore how mortality has changed over time. The *Sentiment* visualization (Figure 2d) allows users to explore the public's perception of different health hazards. This visualization uses Twitter data (data not from IHME) that includes over half a million health-related tweets. Using machine learning models, we classified each tweet by its user category and subject theme. The circular arcs at the top of the visualization represent the top 50 words for the dataset. The middle portion depicts the categorization of tweets by user groups and tweet themes. In its initial configuration (see Figure 1), the bottom of the sentiment visualization shows the sentiment rate for cause-clusters. Users can drill to retrieve additional information for a selected cause-cluster. For instance, in Figure 2d, when cancer is selected, the curved heatmaps depict the sentiment for each cause in the cluster for each user group and tweet theme.

Interaction plays a crucial role in the exploration of data. To facilitate the understanding of health patterns and trends, each visualization has different interactions such as filtering, drilling, selecting, searching, and comparing that are operationalized in a

consistent manner. For an in-depth discussion of how the visualizations were designed, the interested reader is directed to (Ola & Sedig, 2016).

## 7.4 Visualization Literacy Study

### 7.4.1 Research Methodology

Ethics approval for this study was granted by the University of Western Ontario (Appendix 2). To investigate how instructional material influences individuals as they seek to make sense of non-typical visualizations, we utilized the Demography and Geography visualizations from HealthConfection (see Figure 2a and 2b). We selected these two visualizations as the testbed because they include novel and unfamiliar sub-visualizations. For each visualization, we used two versions in our study, one that had a video tutorial and one that did not include the tutorial. The video tutorials were hosted on Youtube (https://youtu.be/HaR7sRfaVtY and https://youtu.be/HwKF9Cbozpo).

#### 7.4.1.1 Participants

A total of 33 participants were recruited from a university in Canada. All of the participants had to be at least 18 years of age and registered students. Participants also needed to be able to use a mouse, keyboard, and computer without any assistance. To recruit participants, we visited first- and second-year class sessions and presented a five-minute summary on the study and allowed students to sign up or send emails to indicate that they desired to participate. Posters and flyers were also posted on university boards. All of the participants were volunteers, and none had seen or used the visualizations before.

#### 7.4.1.2 Procedure

The experiment was conducted in the following steps. After providing consent, each participant was randomly assigned to either the control or the treatment group. Next, we provided a general introduction to the study. The participant then completed a short demographics form. Following this, the participant was given access to the Demography visualization and allowed to explore it. If the participant was a part of the treatment group, they watched the tutorial (i.e., short 5-minute video) and then explored the tool for

an additional 5-minutes. If the participant was a part of the control group, they did not receive the tutorial but were given an equal amount of time to familiarize themselves with the tool (i.e., a total time of 10 minutes regardless of the group). Next, the participant was given access to the online question set for demography and instructed to use the visualization to complete the question set in 25 minutes. At the end of the timeframe, the participant could take a short break. Similar to the first part, the participant explored the Geography visualization, was provided access to the second question set, and instructed to use the visualization to complete the question set. Following this, the participant was given an experience questionnaire to self-report their experience of using the visualizations. Finally, the participant was asked to fill out a form to indicate whether they would like to participate in the interview session. The entire procedure took approximately 90 minutes.

Of those who did not object to being interviewed, some participants were invited to participate in an interview session. During the interview session, after signing the consent form, the participant was asked questions to elaborate on their previously written responses. Also, they were shown the other version of the two visualizations and asked a series of questions. The interview session was audio-recorded. The entire procedure for a participant in this session took approximately 30 minutes.

#### 7.4.1.3 Sources of Data

Four sources of data were used in the study: (1) achievement results and confidence scores obtained from the statistical analysis of the scores on the two question sets; (2) demographics forms; (3) experience questionnaires; and, (4) interview transcripts, obtained from the audio recording during the interview sessions.

Instead of paper and pencil tests, online tests were used to keep track of the overall time spent by each participant. The purpose of the tests was to provide a comparative measure to ascertain participants' understanding of the visualizations. The questions were multiple-choice and fill-in-the-blank type questions. The questions were designed to assess an individual's understanding of how data was encoded and how to interact with the visualization. Some questions required users to perform one sub-task. For example,

for the Geography visualization users were asked, within the environmental and occupational risk group, which risk factor contributes to the most deaths worldwide? To answer this question, participants had to use one of the circular sub-visualizations to identify the largest risk that belonged to the specified group. Other questions required users to perform multiple sub-tasks. For instance, for the Demography visualization, to answer the question, what country in sub-Saharan Africa has the highest mortality rate for individuals between the ages of 35 and 39, participants had to perform three sub-tasks. They had to identify or search for all of the regions of sub-Saharan Africa for the age group. Next, they needed to select each region and then drill to determine which country had the highest mortality rate. In addition to answering the questions, for each question, participants were asked to rank their confidence in the correctness of their answer on a 7-point Likert scale. The demographics form included questions relating to participants' age, major, and gender. The form also asked questions about participants' previous use of, and exposure to, visualizations. The experience questionnaire was used to collect quantitative and qualitative data detailing participants' opinions of the visualizations and tutorials (the treatment group only). The purpose of the interviews was to provide further information about the responses on the experience questionnaire and to help provide a deeper understanding of quantitative data. During the interview, participants from the control group viewed the tutorials and were asked for their opinions. Audio recordings were made of all the interviews. The recordings were later transcribed by the investigators.

#### 7.4.1.4 Hypotheses

The visualization literacy study attempted to test the following two hypotheses.

- Hypothesis I: Instructional materials (i.e., short, minimalist video tutorials) will improve participants' understanding of non-typical visualizations. The group with instructional materials will outperform the control group. Performance will be measured using two main indicators (1) question set scores and (2) self-reported confidence scores.

- Hypothesis II: This was the null hypothesis of the study: the performance of the two groups would be the same.

### 7.4.2 Results

To provide a clearer picture of the participants, prior to a discussion of the results, we present a summary of some data gathered from the demographics forms. The participants came from a wide range of departments, including creative writing, health science, urban development, medical science, kinesiology, music, computer science, biology, geography, women's studies, economics, actuarial science, psychology, media studies, linguistics, and library information science. On a 7-point Likert scale, participants were asked to measure their use of typical and non-typical visualizations on a weekly basis. 64% of the participants in both groups reported using typical visualizations at least occasionally. On the other hand, 64% of participants in both groups reported using non-typical visualizations rarely, very rarely, or never. Table 1 shows a summary of demographic information of the participants by their group.

**Table 7-1: Summary of participant demographics for visualization literacy study[1]**

| Gender | | | | | | | |
|---|---|---|---|---|---|---|---|
| Group | Male | | | Female | | | |
| Control | 6 | | | 11 | | | |
| Treatment | 7 | | | 9 | | | |
| Program Level | | | | | | | |
| | Undergraduate | | | Graduate | | | |
| Control | 15 | | | 2 | | | |
| Treatment | 14 | | | 2 | | | |
| Use of Typical Visualizations | | | | | | | |
| | Always | Very Frequently | Frequently | Occasionally | Rarely | Very Rarely | Never |
| Control | 1 | 3 | 0 | 7 | 4 | 1 | 0 |
| Treatment | 2 | 0 | 4 | 3 | 4 | 1 | 0 |
| Use of Non-Typical Visualizations | | | | | | | |
| | Always | Very Frequently | Frequently | Occasionally | Rarely | Very Rarely | Never |
| Control | 0 | 0 | 0 | 2 | 5 | 6 | 3 |
| Treatment | 0 | 1 | 1 | 3 | 2 | 3 | 4 |

[1] Two participants PT08 and PT10 declined to answer the questions relating to their use of visualizations.

The rest of this section is divided into two subsections. In the first section, we present an analysis of the quantitative results. In the second section, we present an analysis of the qualitative data gathered from the experience questionnaires and during the interview sessions.

#### 7.4.2.1 Analysis of Quantitative Results

In this section, we first discuss the scoring of the test. Next, we present an analysis of the overall score, and in the last section, we present an in-depth analysis of the quantitative data for each visualization separately.

*Scoring of the tests*

A simple scheme was utilized; and, questions were awarded points based on the number of sub-questions. The first seven questions on the Demography visualization were awarded one point each, while the last three questions were awarded four points each because they each included four sub-questions. For the geography test, the first eight questions were awarded one point each, while the last two questions were awarded four points each. Skipped or incomplete questions were awarded a mark of zero. The points were added and then converted to a percentage. For the scoring of confidence, the Likert scale values were converted to numerical numbers (i.e., 7 - strongly agree, 1 - strongly disagree). The values were then added and converted to percentages. The overall achievement and confidence score is the average of the Demography and Geography visualization score for each participant.

*Overall*

Statistical analysis was conducted on the achievement and confidence scores. Table 2 shows the descriptive statistical summary. The treatment group generally performed better than the control group. The mean difference for the achievement score between the groups is 13.4%. Figure 3 shows the box plot of the overall achievement scores per group. The mean difference for the self-reported confidence is 9.5%.

**Table 7-2: Overall descriptive statistical summary for the visualization literacy study**

| | Question Set | | Confidence Level | |
|---|---|---|---|---|
| | Control | Treatment | Control | Treatment |
| Mean | 57.31 | 70.68 | 75.92 | 85.38 |
| Standard Error | 4.16 | 4.16 | 2.85 | 2.43 |
| Median | 56.41 | 69.33 | 80.18 | 85 |
| Mode | 48.68 | 88.49 | 82.14 | N/A |
| Standard Deviation | 17.17 | 16.66 | 11.74 | 9.72 |

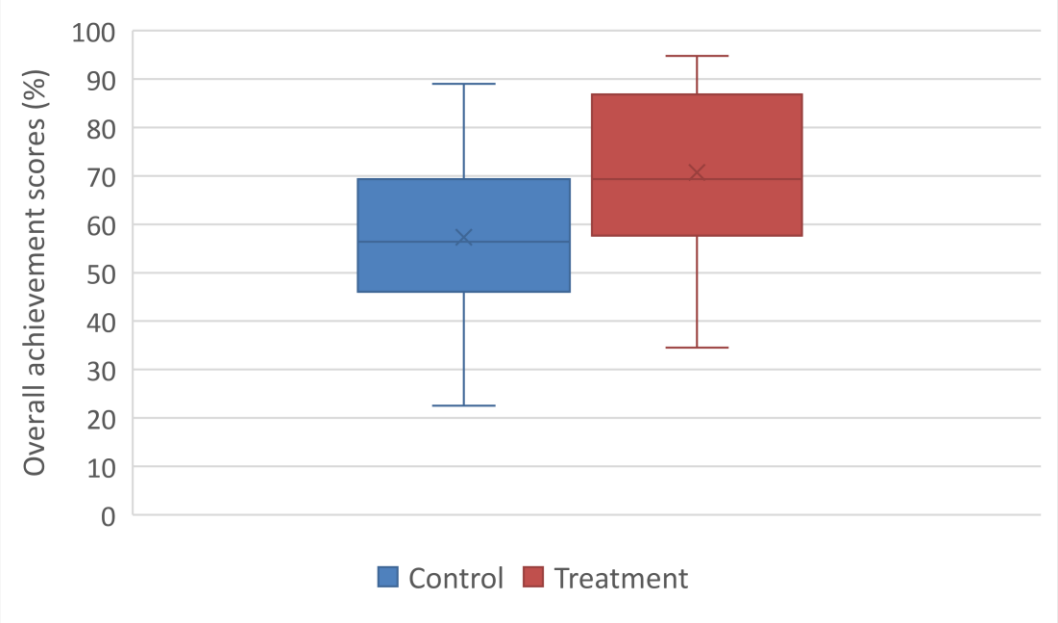



**Figure 7-3: Box plot of the overall achievement scores for the control and treatment groups**

To examine whether there is any statistical significance to using the tutorial a one-way analysis of variance (ANOVA) test was performed. The analysis results are depicted in Table 3. We found that participants who used the tutorial performed significantly better on the question sets than participants in the control group, $F(1, 32) = 5.15$, $p < .05$. A one-way ANOVA statistical test was performed on the confidence levels as well. The results indicate the difference in confidence scores is significant, $F(1,32) = 6.31$, $p < .05$. Based on the ANOVA tests and the descriptive statistical analysis of the achievement and

confidence scores, Hypothesis I can be accepted and the null hypothesis can be rejected for the study.

**Table 7-3: One-way variance analysis test for the visualization literacy study**

| | Question Set | Confidence |
|---|---|---|
| F(1,32) | 5.15 | 6.31 |
| Fcrit | 4.16 | 4.16 |
| p-value | 0.030 | 0.017 |

Further analysis was performed to assess if there was any difference in literacy for each visualization by splitting the results into two categories based on the visualizations.

*Analysis by visualization*

Statistical analysis was conducted on the question set and confidence scores for each visualization. Table 4 shows the descriptive statistical summary. The treatment group generally performed better than the control group. For the Demography visualization, the mean difference for the question set was 17.4%, while the difference for the confidence was 10.7%. For the Geography visualization, the difference in means is smaller, 9.4% for the question set scores and 8.2% for confidence. These results seem to suggest that the tutorial was more effective for the first visualization (i.e., Demography) than for the second visualization.

**Table 7-4: Descriptive statistical summary by visualization**

| | Demography Visualization | | | | Geography Visualization | | | |
|---|---|---|---|---|---|---|---|---|
| | Question Set | | Confidence Level | | Question Set | | Confidence Level | |
| | Control | Treatment | Control | Treatment | Control | Treatment | Control | Treatment |
| Mean | 51.39 | 68.77 | 71.77 | 82.50 | 63.24 | 72.66 | 80.06 | 88.26 |
| Standard Error | 4.59 | 4.69 | 3.02 | 3.18 | 5.41 | 4.49 | 3.05 | 2.74 |
| Median | 47.37 | 68.42 | 71.43 | 83.98 | 62.50 | 78.13 | 80.00 | 90.71 |
| Mode | 47.37 | 89.47 | 66.67 | 85.71 | 81.25 | 87.50 | 92.86 | 98.57 |
| Standard Deviation | 18.93 | 18.76 | 12.45 | 12.72 | 22.30 | 17.95 | 12.56 | 10.95 |

ANOVA tests were performed to examine whether there is any statistical significance to using the tutorial for each visualization. The analysis results are depicted in Table 5. The results show that using the tutorial improved participants' performance on the question set for the Demography visualization ($p < .05$) but not for the Geography visualization ($p > .05$). Regarding the confidence level, the same is observed, the Demography visualization results indicate statistical significance to using the tutorial visualization while the Geography visualization results are not statistically significant.

**Table 7-5: One-way variance analysis test by visualization**

| | Demography | | | Geography | |
|---|---|---|---|---|---|
| | **Question Set** | **Confidence** | | **Question Set** | **Confidence** |
| F(1,32) | 7.01 | 5.99 | | 1.77 | 3.97 |
| Fcrit | 4.16 | 4.16 | | 4.16 | 4.16 |
| p-value | 0.013 | 0.020 | | 0.193 | 0.055 |

Although this would require further investigation, we believe that the lack of statistical significance for the Geography visualization is in part due to transferable skills learned while participants interacted with the first visualization. Participants were already familiar with the structure of the underlying data (i.e., risks, causes, cause-clusters, and groups) and how to interact with visualizations based on their previous exposure to the Demography visualization. Even the participants without tutorials had engaged in a trial-and-error process that impacted their understanding. During the interview session, we asked participants follow-up questions to further explore this difference.

### 7.4.2.2 Analysis of Qualitative Results

As the above quantitative analysis shows, using the tutorials improved participants' achievement and confidence scores. In this section, we present the analysis of the qualitative data to get a better understanding of the experience of participants and the effect of the tutorials. These results include a combination of responses from the experience questionnaire and comments during the interview sessions. Participants are referred to by their number and their group, participants in the control group are referred to as PC<#>, while those in the treatment group are referred to as PT<#>.

*Effect of the Tutorial*

On the experience questionnaire, participants were asked to speak to the effect of the tutorial on their ability to complete the question set. Some of the comments are:

- "I had never seen a Demography visualization before so the video introduced it to me and taught me how to use it. The video, although short, really explained how to use the visualization and made it clear where to find the things I needed to find." (PT06)

- "Simply looking at the 2 circles was a bit offputting; with the tutorial it was made clear what the purpose was. I was immediately confused about the lines; however tutorial cleared that up." (PT10)

- "At first glance, the geography visual is intimidating and the tutorial breaks it down nicely." (PT12)

- "Being told how to interpret complex diagrams is very helpful when presented with a wide array of options/buttons to click. Being told what things meant and how to find them was very helpful." (PT13)

- "The speaker was slow and she made it concise and short to understand. The demography visual was immediately intimidating but tutorial cleared confusion." (PT14)

- "Without any instructions on how the data is organized, it is difficult to get the hang of it yourself without spending lots of time." (PT15)

During the interview members of the treatment group were shown the version without the tutorial and asked if it would have been more or less difficult to use. PT07 said, "Extremely more difficult. I felt that I had difficulty even after having seen the tutorial, so I worry that without it I wouldn't have managed to be slightly confident for the tasks". PT13 had a similar opinion: "It would have been so much more difficult because the amount of data that you are trying to show to somebody. I'm sure I could have figured it out, but it would have taken me at least an hour to figure it out without the aid of the tutorial."

*Strategy for making sense without aid*

When asked how their strategy for using the visualization to complete the question set would have differed without the tutorial, PT07 said, “I can’t imagine how, but I want to say yes. Because for most of the questions I had an idea of where to start because I knew basically how most of the visualization worked, so I think I had a starting point. It would have been much more random guessing at the start of each question until I found something that answered the question and then I would have tried to figure it out from there.”

This observation is similar to the responses of the participants in the control group who participated in the interview. When they were asked how they learned to use the visualization, they said:

- “My process was just to click around until something happened and then try to understand what happened. I was able to figure out the second one because of the color scheme; the reserved colors help me to know that they were related.” (PC06)

- “Explore and understand it step by step. So, I break it down and go through the different sections to try and understand how they work together. It is kinda of funny. I didn’t notice the legend on the side until I had already gone through it and figured out what the categories meant on my own.” (PC10)

- “I started looking at the headings and just stared at it for a while. I did not realize that you could click or interact with it. And then when I started looking at the questions and answering them it started to make sense. Then I saw the + sign at the top and all the other things that started popping up.” (PC12)

*Experience of participants without the tutorial*

Four participants from the control group participated in the interview session, where each of them was shown the tutorial and then asked a series of questions. In terms of interacting with the visualizations, three of the four interviewed participants were unaware of many interaction options that existed for the Demography visualization. PC10

said, “I did not know that [the menu with five different interaction options] was there. I didn’t know how to use the hive plot. I just put low for the answer because I did not know what to do”. After watching the tutorial, PC12 said “WHAT!!! I did not see that, no!!! I knew there was more but wasn’t sure how to get to it. There are so many things!!”. PC06 said, “Oh wow. This would have been beneficial to helping me use the tool.” When asked how using the tutorial might have impacted their exploration, they said:

- “It would have potentially helped me to find the other elements a little more easily. The things in the tutorial where things I figured out along the way. Where I struggled was combining different parts to find the answer. Narrowing down to the region or a specific country within an age group for a cause or risk. I think it would have helped me to skim off that part of figuring it out.” (PC02)

- “I would have been more purposeful in my interaction. I wouldn’t have had to click randomly to see the connection.” (PC06)

- “It would have helped me to feel more secure in the knowledge and my understanding of it. I think that in terms of which one was highest or lowest that was definitely something that I had to poke around with to figure it out. To figure out which was highest or lowest, when I clicked on it, I would compare the actual numbers. Understanding how the interactions work, that was something I was iffy on, so that would have been something that the tutorial would have helped with.” (PC10)

- “It would have made it better for me to figure things out. It would have changed my strategy. Cause I would know where to look for things because at first it was going to try and see what pops up and one of the things that I assumed that the causes at the top were the highest but I wasn’t sure if it was that way.” (PC12)

When asked about why their performance and confidence level improved when they used the second visualization (i.e., Geography). PC02 said, “For the first one, there were moments when I didn’t understand the differences between causes and risk factors. And maybe that is why the geography was more intuitive because I had a better understanding of geography”. PC12 mentioned, “because I already knew that I had to click on things to

get more information. For the first one, I did not know what I can click and do, and what would pop up. For the second one, I knew that there was more. So that kinda like made it easier." PC06 said, "Anxiety and nervousness, I was more calm during the second one. I did not know how to interact with the first one".

Our study reveals that the participants' achievement and confidence scores increased with the use of the short, minimalist video tutorials. The qualitative data further underscores the benefits of instructional materials, especially when time is a factor. In addition, we observed that once participants have used a visualization, there is a transfer of skills to other visualizations. Some researchers believe that we should only use simple, chart-like visualizations and have argued against the use of elaborate visualizations. This study shows that even when individuals have a low exposure to complex visualizations, the majority of participants reported using non-typical visualizations rarely or less than that, they were able to increase their visualization literacy through focused exploration. This study helps to emphasize the benefits of video tutorials and the ability of humans to learn to use non-trivial visualizations. Now that we have evidence indicating that individuals can properly use non-typical visualizations, in the next section, we investigate HealthConfection's ability to improve health literacy.

## 7.5 Health Literacy Study

### 7.5.1 Research Methodology

In this section, we describe the research methodology to investigate the ability of visualizations to improve health literacy. Ethics approval for this study was granted by the University of Western Ontario (Appendix 3). Once again, our study tool was HealthConfection. Participants used the Geography, Demography, Chronology, and Overview visualizations and had access to the respective tutorials. The Sentiment visualization was not included in the study because the public's opinion on health issues is not an aspect of disease prevention.

#### 7.5.1.1 Participants

A total of 28 participants were recruited from a university in Canada. All of the participants had to be at least 18 years of age and registered undergraduate or graduate students. Participants also needed to be able to use a mouse, keyboard, and computer without any assistance. To recruit participants, we visited first- and second-year classes and presented a five-minute summary on the study and allowed students to sign up or to send emails to indicate whether they desired to participate in the study. Posters and flyers were also posted on university boards. All of the participants were volunteers. None of the participants had seen or used the tool before.

#### 7.5.1.2 Procedure

The experiment involved two sessions: a test session and an interview session. For the test session, the participants were randomly assigned to either the control or the treatment group, and were given the appropriate consent form. After obtaining consent, each participant completed a short demographics form. Next, for those in the control group, they were administered the health literacy quiz. This concluded their participation in the study. For a participant in the treatment group, they were given a brief overview of the tool and then given a task sheet to complete. The task sheet was designed to facilitate a guided exploration through each visualization. Upon completion of the tasks, the participant could take a short break. Next, the quiz was administered and then the participant was given an experience questionnaire to self-report their experience of using the tool. Lastly, the participant was asked to fill out a form to indicate whether they would like to be interviewed. The entire procedure for a participant in the treatment group was approximately 100 minutes, while for a participant in the control group it was approximately 25 minutes.

Of those who consented to being interviewed, some participants were invited to participate in an interview session. During the interview session, the participant was asked a series of questions. The entire procedure for a participant in this session took approximately 25 minutes.

#### 7.5.1.3 Tasks

Participants in the treatment group were asked to complete a series of tasks. Being that data space is large (i.e., over 12 million records), an unguided exploration by participants would result in different concepts being learned. The tasks were intended to provide participants with pre-determined goals to facilitate the learning of specific health concepts within the limited duration of the study. For each visualization, participants were asked to complete five tasks. Most tasks required users to perform a combination of sub-tasks and to interpret how the data was encoded. For instance, for the Geography visualization, participants were asked to determine which regions of the world are severely impacted by a diet low in fruits. This task can be completed in multiple ways. One way would involve, first locating, and then selecting the diet low in fruits risk factor from one of the two circular sub-visualizations in the top half of the Geography visualization. Next, a participant could use the map's legend to select the regions that fall between the third and fourth quartiles. If a participant is unfamiliar with the regions highlighted, then he/she could select each region to determine its name. Participants were not told which steps to take. Instead, they were given the tasks and instructed to use the tool to complete them. Table 6 includes a sampling of tasks assigned. As users performed the assigned tasks, they were able to gradually explore *the story of the data* and discover different trends that exist.

**Table 7-6: Sample tasks for health literacy study**

| Visualization | Task |
|---|---|
| Geography | At a global level, what are the risk factors that contribute to death from tuberculosis? |
| Chronology | Which cause-clusters significantly increased in rank between 1990 and 2010? |
| Demography | For which age groups, is dietary risk factor and physical inactivity the highest ranked risk-cluster that contributes to death? |
| Overview | Which cancer results in the highest number of deaths for adults in sub-Saharan Africa? |

#### 7.5.1.4 Sources of Data

Four sources of data were used in the study: (1) achievement results obtained from the statistical analysis of the quiz scores; (2) demographics forms; (3) experience

questionnaires; and, (4) interview transcripts, obtained from the audio recording during the interview sessions. A paper-and-pencil quiz was constructed. The purpose of the quiz was to ascertain participants' global health literacy. The quiz contained 20 multiple-choice questions, which were based on the exploration tasks. The demographics form included questions relating to participants' age, major, and gender. The form also asked questions about participants' interest and exposure to global health concepts, as well as their previous use of and exposure to visualizations. The experience questionnaire, which was only for the treatment group, was used to collect quantitative and qualitative data detailing participants' opinions of the visualizations. On the questionnaire, we surveyed seven questions regarding HealthConfection on the 7-point Likert scale: 1) Engagement; 2) Fun; 3) Ease of use; 4) Ease of learning; 5) Enjoyability; 6) Benefit to health literacy; and, 7) Layout of the visualization. During the interview sessions, participants in the treatment group were asked to expound on some of their responses on the experience questionnaire and provide detailed feedback on the efficacy of the tool. The investigators transcribed the audio recordings of the interviews.

#### 7.5.1.5 Hypotheses

The health literacy study attempted to test the following two hypotheses.

- Hypothesis III: The developed visualization tool improves health literacy. The group that uses the tool will outperform the control group on the quiz. Performance will be measured by achievement scores.

- Hypothesis IV: This was the null hypothesis of the study: the performance of the two groups would be the same.

### 7.5.2 Results

Before a discussion of the results, we present a summary of some data gathered from the demographics forms. The participants were from diverse departments including biology, computer science, psychology, kinesiology, political science, chemistry, biochemistry, linguistics, occupational therapy, management and organizational studies, urban development, electrical engineering, and library information science. On a 7-point Likert

scale, participants were asked to measure their use of non-typical visualizations on a weekly basis. 70% of the participants in both groups use non-typical visualizations rarely, very rarely, or never. More than half of the participants in both groups mentioned that they had been exposed to global health in a formal school setting. Table 7 shows a summary of demographic information of the participants by their group.

**Table 7-7: Summary of participant demographics for health literacy study**

| Gender | | | | | | | |
|---|---|---|---|---|---|---|---|
| Group | Male | | | Female | | | |
| Control | 5 | | | 9 | | | |
| Treatment | 5 | | | 9 | | | |
| Program Level | | | | | | | |
| | Undergraduate | | | Graduate | | | |
| Control | 8 | | | 6 | | | |
| Treatment | 9 | | | 5 | | | |
| Use of Non-Typical Visualizations | | | | | | | |
| | Always | Very Frequently | Frequently | Occasionally | Rarely | Very Rarely | Never |
| Control | 1 | 0 | 0 | 0 | 4 | 4 | 5 |
| Treatment | 1 | 0 | 1 | 2 | 3 | 1 | 6 |
| Exposure to global health concepts in school | | | | | | | |
| | Strongly agree | Agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Disagree | Strongly disagree |
| Control | 1 | 3 | 4 | 0 | 4 | 1 | 1 |
| Treatment | 1 | 3 | 5 | 1 | 3 | 1 | 0 |

### 7.5.2.1 Quiz Results

Each question on the global health literacy quiz was awarded one point. Skipped or incomplete questions were awarded a mark of zero. The points were then converted to percentage. The participants in the treatment group achieved a higher score than those in the control group. Table 8 shows the descriptive statistical summary by group and Figure 4 shows the box plot of the overall scores. To determine if the effect of HealthConfection to improve health literacy is statistically significant, we applied a one-way ANOVA test on the quiz scores. The results of these analyses, $F= 195.40$, $F_{crit} = 13.74$, $p = 1.33 \times 10^{13}$

thus $p < 0.001$, confirm our third hypothesis that the visualization tool improves health literacy.

**Table 7-8: Descriptive summary of quiz scores**

| | Treatment | Control |
|---|---|---|
| Mean | 78.93 | 21.07 |
| Standard Error | 3.44 | 2.30 |
| Median | 82.50 | 20.00 |
| Mode | 60.00 | 15.00 |
| Standard Deviation | 12.89 | 8.59 |

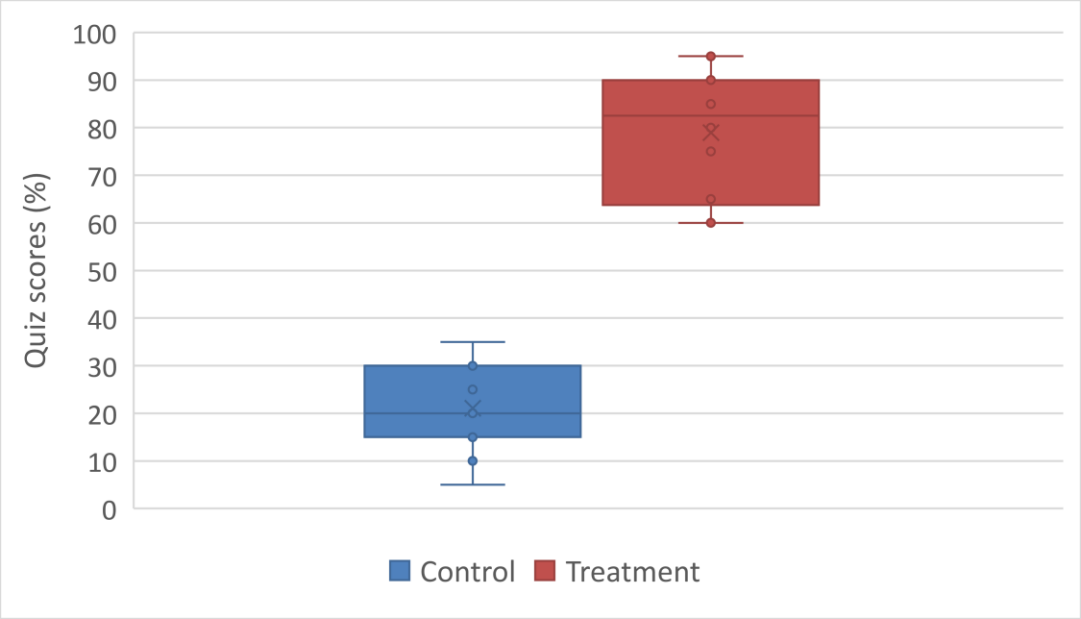



**Figure 7-4: Box plot quiz scores for the treatment and control groups**

### 7.5.2.2 Experience Questionnaire and Interview Feedback

In this sub-section, we examine the quantitative and qualitative feedback received on the experience questionnaire and during the interview sessions. On the experience questionnaire, participants in the treatment group were surveyed to ascertain their experience with HealthConfection. Some of the questions were related to the layout of the visualizations and how engaging, fun to use, easy to use, easy to learn, and enjoyable to use the tool was. We also asked them to state whether they thought the tool improved

their understanding of global health concepts. A 7-point Likert scale was utilized and Figure 5 details the responses. In addition to these questions, participants also provided written comments on their experience. Three of the 14 participants took part in the interview session. Participants are referred to as PT<#>, where # represents their identification number.

Generally, the participants' responses were positive. Thirteen of the 14 participants agreed or strongly agreed with the statements relating to engagement, fun, and enjoyability. In the comments section, PT02 wrote "Super cool! I was very mesmerized by the entire program. Very interactive and fun to play around on. Good for visual learners. Elegant presentation of a mind-boggling amount of information". PT13 wrote "Really neat! I think it could be really helpful for those who aren't as mathematically inclined or those who learn visually". In terms of ease of use, one participant was ambivalent, while the majority of participants somewhat agreed with the statement. During the interview, both PT13 and PT07 alluded to having to return to the tutorials during their completion of the tasks because they were not sure how to use the tool properly. When asked about the layout of the visualizations, five of the participants strongly agreed that it was beneficial for navigation, while six agreed and three somewhat agreed. During the interview, PT13 mentioned that the benefit of having the layout is that you see everything together and know everything that is being offered. PT07 liked the layout and said, "it is like a mind map that improves navigation." This sentiment was echoed by PT05 who said "The layout was beneficial for me; you can see how things are related. It is easier to move through because they are all close together. This idea of being able to move through the visualization and thus navigate through the data is beneficial for exploration.

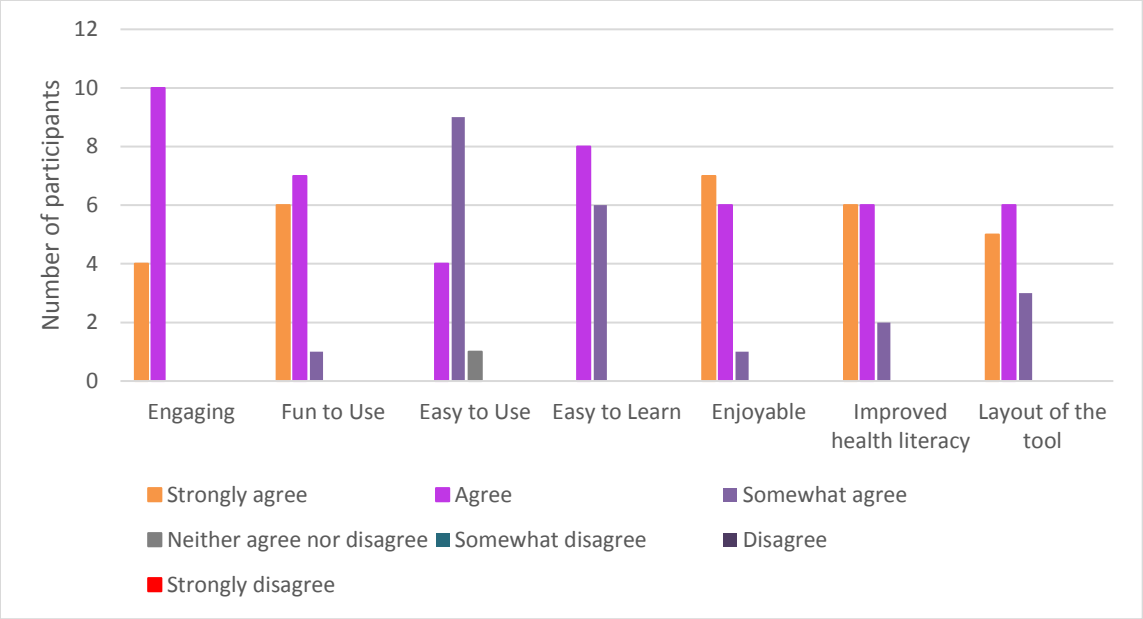



**Figure 7-5: Responses on health literacy experience questionnaire**

On the questionnaire, regarding health literacy, 6 participants strongly agreed that the tool was beneficial, six agreed, and two somewhat agreed with the statement. In the comments section of the question, PT08 wrote, "Very informative and really fun to play around with and learn about global health. As someone who does not know very much about global health, I really enjoyed using this tool to learn about this topic". PT11 commented "Very impressive; I wish they used these in class, it would really help the students learn better especially for health scientists". During the interview session PT05 who had his/her interview a week after the first session commented on the memorability of the data, "I still remember some of the information, like it was about my country, I was like Oh I didn't know that. I would love to use it again". Both PT05 and PT13 highlighted that, as self-described people who are not good at mathematics or who do not like reading, the way information was represented was very beneficial for their learning. PT13 said that the tool made the data accessible to people who are not mathematically inclined while PT05 said, "I'm the type of person who doesn't like reading information. A tool like this that you can go and eliminate data is easy. Better than Google. The information was very direct; you don't have to go through a lot of reading to find it. I think it is really cool".

In this study, we investigated whether non-trivial visualizations can be used as health literacy tools. Our results were statistically significant and indicate that visualizations can be used to improve the general public's understanding of health patterns and trends. An analysis of the qualitative data emphasizes the positive response of participants regarding HealthConfection as a health literacy tool.

## 7.6 Discussion and Conclusion

This paper has presented two multi-method empirical studies: the first investigated the use of short video tutorials to improve visualization literacy and the second investigated the use of visualizations to improve health literacy. The testbed for both investigations was a visualization tool that we created, HealthConfection. This tool uses aggregated datasets of global health data.

The first study evaluated the effect of video tutorials on visualization literacy. The study showed that even without support structures, participants could learn how to use two sophisticated, non-trivial visualizations. In particular, participants with the tutorials achieved higher scores than those without instructional materials, indicating that the video tutorials improved participants' understanding of the Geography and Demography visualizations. This study and its results have certain limitations. First, the participants are all university students who are not an accurate representation of the general public. Second, interviewees may have wanted to please the interviewer by providing desirable answers. Despite these limitations, the research can lead to a few general conclusions that have implications for the use of visualizations. First, our results can be generalized to other elaborate and unfamiliar visualizations; we believe that short, minimalist video tutorials can help to improve the public's ability to use such visualizations. Second, contrary to our expectations, participants without the tutorial could make sense of aspects of the visualizations. These results suggest that if given time, the general public can make sense of and learn how data is encoded and how to interact with novel, non-typical visualizations, even though they are complex and unfamiliar. That being said, it is possible that the closed nature of the questions served as an unconscious tutorial. Further research is needed to design more advanced and open-ended questions to better ascertain such knowledge. Insights from this study also indicate that visualization knowledge is

transferable and people with limited exposure to or experience with visualizations can transfer their learned knowledge of certain visualizations to other visualizations. More research is needed to understand this phenomenon.

The second study investigated the use of non-typical visualizations to improve health literacy. The study showed that during an hour of goal-oriented exploration, the participants were able to improve their understanding of global health trends. Some limitations of the study include the sample size and the fact that students are not representative of the general public. Another limitation was that the control group did not have any exposure to the data within the visualization tool before taking the quiz. Future research should compare the use of visualizations to the use of existing repositories of data, including reports and search engines. In spite of these limitations, the study has implications for health literacy. The findings of this research demonstrate that non-trivial visualizations can be used to improve health literacy. In situations where individuals are motivated to learn, visualizations that initially may seem complex can be learned with short video tutorials. While in the past, typical visualizations, such as column charts and line charts, have been advocated for because of their simplicity, our research implies that more complex visual representation forms can be used to improve health literacy. Furthermore, the research suggests that when confronted with large amounts of data, visualizations that allow users to disclose information gradually are beneficial. The ease of use highlighted by users and their quiz scores underscore this point. The interactive nature of visualizations is also important. The research suggests that providing users with diverse interactions allows them to perform various tasks. Also, when exploration is a key task of users, the layout of visualizations may impact their ability to navigate. While we did not test the impact of different layouts, both the comments of users and the quantitative data suggest that providing users with a single interface that provides an overview, clear and consistent structures, and navigational cues is beneficial. With these visualizations, participants were able to engage in an exploration of *the story of the data*. The findings of this research imply that visualizations can be used to empower the general public to learn about disease prevention in an engaging format. Overall, we expect that our findings on using tutorials to improve visualization literacy and non-typical visualizations to improve health literacy could be generalized to other

visualizations and other domains where large repositories of data need to be made available to the public in an accessible manner. Ultimately, we hope that our work serves as encouragement to those seeking to advance health literacy.

# Chapter 8

# 8 Conclusion

This dissertation has touched upon several aspects relating to the design of interactive visualizations and analytics for public health data. Chapter 2 presented a broad overview of the data-related challenges facing the public health community. Chapters 3 and 4 focused on the design of visual representations and interaction. Chapter 5 demonstrated how analytics could be incorporated into visualizations. Chapters 6 and 7 presented studies relating to mining twitter data to understand the discourse of health issues on the platform and improving visualization and health literacy. This chapter concludes the dissertation and is divided into three main sections: (1) a summary of each chapter and some of its contributions; (2) general conclusions on the contribution of this research to the wider scientific literature; and (3) areas of future research.

## 8.1 Dissertation Summary

Visual analytics tools and public health. In Chapter 2, we discussed the challenges facing the public health community, described visual analytics tools, and discussed the role that visual analytics tools can play in addressing the challenges. In doing so, we demonstrated the potential benefit of incorporating visual analytics tools into public health practice and highlighted the need for further systematic and focused research.

Visualization design. In Chapter 3, we presented how frameworks can guide designers in the creation of non-trivial visualizations. In this chapter, we made a case for the development and use of sophisticated visualizations that can represent the complexity of datasets by allowing multiple facets of the data to be encoded simultaneously. We also presented how the patterns in the data can help drive the design of visualizations in a systematic fashion. To make this point, we provided detailed explanations on how four non-typical visualizations were designed.

Interaction design. In Chapter 4, we presented a design process for interaction based on elements of a framework. We also demonstrated the significance of understanding users'

tasks and how tasks influence the design of interaction. We also discussed the limitations of interaction provided by existing visualization tools and presented a case for interaction that allows users to have a more meaningful discourse with data. This chapter ended with scenarios that highlighted the efficacy of a task-based approach to interaction design.

Coupling visualization and analytics. In Chapter 5, we demonstrated how to combine analytics and visual representations to support the work of public health stakeholders. In particular, we showed how statistical measures that are typically difficult for the average person to comprehend could be understood with visualizations. In this chapter, we discussed a visualization tool we created to facilitate making sense of the spread of Zika and showed, with a case study, the efficacy of visualizations to confront emerging health issues.

Visual analytic study. In Chapter 6, we conducted a visual analytic study to explore the discourse of health on Twitter. We reported on a process for conducting visual analytic studies and demonstrated how analytic models could be constructed to provide more insight on Twitter chatter. In this chapter, we also discussed the design of a visualization that will allow individuals to learn more about how health is being discussed on Twitter and demonstrated the role of visualizations in analysis.

Visualization and health literacy studies. In Chapter 7, we presented the results of two research studies that we conducted with visualizations we developed. The first study explored visualization literacy. In particular, we explored how people go about learning to use non-trivial visualizations and the impact of instructional materials. One of the key findings of this research is that short minimalist video tutorials can improve visualization literacy. In the second study, we demonstrated that during an hour of goal-oriented exploration, participants were able to improve their understanding of global health trends.

## 8.2 General Contributions

As described in Chapter 1, the broad concern of this research surrounds the design of interactive visualizations and analytics for the domain of public health. Currently, there is a paucity of research in this area and our goal was to help bridge the gap between

theoretical concepts and practice. One contribution of this dissertation is the explication of the design process for visualizations and interaction for the public health domain and emphasizing the need for systematic design approaches. Using this dissertation, designers can approach the creation of health visualizations with a deeper understanding of the nuances of how the process unfolds.

Another contribution of this dissertation is the visualization tool, HealthConfection, that is discussed in chapters 3, 4, 6, and 7. The five interactive visualizations that we designed are elaborate and sophisticated visualizations that can be implemented and used across the globe to improve health literacy. In addition, these visualizations can be reconfigured to work with different bodies of data. In particular, the Demography visualization can be explored as a new visualization technique.

Other contributions emerge from the studies we conducted. Future studies can be modeled after the three studies presented in Chapter 6 and 7 to further explore how to improve visualization and health literacy and how to understand the discourse of health on social media platforms. While there has been an increased interest in the development of visualizations, research on how people make sense of visualizations has not been explored as much. In addition, most of the existing research on visualization literacy has focused on simple and/or static visualizations. Our work moves the visualization field forward by exploring visualization literacy for non-typical elaborate interactive visualizations. In the past, there has been reticence on using elaborate visualizations partially because of their perceived complexity, our research has challenged this notion and has opened the door for more use and research on advanced visualizations in the health domain.

Finally, our research contributes to health literacy efforts. There is a need to empower individuals to seek and understand data. This dissertation provides the public health domain with evidence of the efficacy of visualizations for health literacy. This research has implications for other fields inundated with massive amounts of data that need to be made accessible and understandable for the general public.

## 8.3 Future Work

The work presented in this dissertation lays the foundation for the design and use of visual analytics tools in public health. While we focused primarily on how visualizations are designed and how analytics can be incorporated into visualization tools, more research is needed to explore how to best couple the two distinct components together. Research on how information processing should be distributed between the components and how to externalize analytic processes is needed. In addition, as our work focuses on interaction at the level of actions, more research on understanding how interaction at the level of events influences the completion of tasks is needed.

From a visualization literacy standpoint, more research is needed to better understand how people make sense of sophisticated visualizations. One line of research can investigate the transference of skills. In our study, we noticed that individuals performed better on the question set for the second visualization. Empirical studies can explore what factors contributed to this increase. Another line of research relates to support structures for different demographics. Our study utilized students between the ages of 18 and 35 as participants, additional studies can help ascertain the effectiveness of instructional materials for different age groups.

This dissertation introduces the use of elaborate visualizations for health literacy. Our study was laboratory based, in that we required participants to use the visualizations for a fixed time in a location in which distractions were at a minimum. Studies that evaluate the use of visualizations in classrooms, public areas, and more realistic settings will contribute to a better understanding of the efficacy of such tools. Furthermore, studies in other countries can help make HealthConfection a tool that is beneficial to individuals around the globe.

# References

Ahonen-Rainio, P., & Kraak, M.-J. (2005). Deciding on Fitness for Use: Evaluating the Utility of Sample Maps as an Element of Geospatial Metadata. *Cartography and Geographic Information Science*, *32*(2), 101–112.

Aigner, W. (2011). Understanding the role and value of interaction: First steps. In S. Miksch & G. Santucci (Eds.), *International Workshop on Visual Analytics*.

Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). *Visualization of Time-Oriented Data*. (J. Karat & J. Vanderdonckt, Eds.)Human-Computer Interaction Series. London: Springer London.

Aimone, A. M., Perumal, N., & Cole, D. C. (2013). A systematic review of the application and utility of geographical information systems for exploring disease-disease relationships in paediatric global health research: the case of anaemia and malaria. *International journal of health geographics*, *12*(1), 1. BioMed Central.

Al-Hajj, S., Pike, I., Riecke, B., & Fisher, B. (2013). Visual Analytics for Public Health: Supporting Knowledge Construction and Decision-Making. *2013 46th Hawaii International Conference on System Sciences* (pp. 2416–2423).

Alpaydin, E. (2009). *Introduction to Machine Learning (Adaptive Computation and Machine Learning series)* (2nd ed.). The MIT Press.

Alper, B., Riche, N. H., Chevalier, F., Boy, J., & Sezgin, M. (2017). Visualization Literacy at Elementary School. *Proceeding of the Conference on Human Factors in Computing Systems (CHI)*.

Andreinko, G., Jern, M., Dykes, J., Fabrikant, S., & Weaver, C. (2007). Geovisualization and synergies from InfoVis and Visual Analytics. *2007 11th International Conference Information Visualization (IV '07)* (pp. 485–488). IEEE.

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D. A., Kraak, M. - J., MacEachren, A., & Wrobel, S. (2007). Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, *21*(8), 839–857. Taylor & Francis, Inc.

Andrienko, N., & Andrienko, G. (2003). Informed spatial decisions through coordinated views. *Information Visualization*, *2*(4), 270–285. Palgrave Macmillan.

Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11* (pp. 1–4). New York, New York, USA: ACM Press.

Aslam, S. (2017). Twitter by the Numbers: Stats, Demographics & Fun Facts. Retrieved July 17, 2017, from https://www.omnicoreagency.com/twitter-statistics/

Aziz, S., Ngui, R., Lim, Y. A. L., Sholehah, I., Nur Farhana, J., Azizan, A. S., & Wan Yusoff, W. S. (2012). Spatial pattern of 2009 dengue distribution in Kuala Lumpur using GIS application. *Tropical biomedicine*, *29*(1), 113–20.

Baltussen, R., & Niessen, L. (2006). Priority setting of health interventions: the need for

multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation*, *4*(1).

Barrett, F. A. (2000). Finke's 1792 map of human diseases: the first world disease map? *Social Science & Medicine*, *50*(7–8), 915–921.

Bates, M. J. (2005). Information and knowledge : an evolutionary framework for information science. *Information Research*, *10*(4), 1–24.

Bekaert, G., & Harvey, C. R. (2002). Research in emerging markets finance: Looking to the future. *Emerging Markets Review*, *3*(4), 429–448.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., Viera, A., Crotty, K., Holland, A., et al. (2011). Health literacy interventions and outcomes: an updated systematic review. *Evidence report/technology assessment*, (199), 1–941.

Berner, E. S., & Moss, J. (2005). Informatics Challenges for the Impending Patient Information Explosion. *Journal of the American Medical Informatics Association*, *12*(6), 614–617. Elsevier Science.

Bhowmick, T., Griffin, A. L., MacEachren, A. M., Kluhsman, B. C., & Lengerich, E. J. (2008). Informing geospatial toolset design: understanding the process of cancer data exploration and analysis. *Health & place*, *14*(3), 576–607.

Bian, J., Topaloglu, U., & Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. *Proceedings of the 2012 international workshop on Smart health and wellbeing - SHB '12* (pp. 25–32). New York, NY, USA: ACM Press.

Bikakis, N., & Athens, N. T. U. (2016). Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art - Semantic Scholar.

Bivand, R., & Yu, D. (2009). spgwr: Geographically Weighted Regression. Retrieved from https://cran.r-project.org/package=spgwr

Bloland, P. B. (2001). Drug resistance in malaria. Geneva: World Health Organization. Retrieved from http://www.who.int/csr/resources/publications/drugresist/malaria.pdf

Bodnar, J. (2005). Making sense of massive data by hypothesis testing. *International Conference on Intelligence Analysis*.

Boischio, A., Sánchez, A., Orosz, Z., & Charron, D. (2009). Health and sustainable development: challenges and opportunities of ecosystem approaches in the prevention and control of dengue and Chagas disease. *Cadernos de Saúde Pública*, *25*, S149–S154. Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz.

Borner, K., Maltese, A., Balliet, R. N., & Heimlich, J. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, *15*(3), 198–213. SAGE Publications.

Bostock, S., & Steptoe, A. (2012). Association between low functional health literacy and mortality in older adults: longitudinal cohort study. *BMJ*, *344*.

Brey, P. (2005). The Epistemology and Ontology of Human-Computer Interaction. *Minds and Machines*, *15*(3–4), 383–398.

Brownson, R. C., Fielding, J. E., & Maylahn, C. A. (2009). Evidence-Based Public Health: A Fundamental Concept for Public Health Practice. *Annual Review of Public*

*Health*, *30*, 175–201.

Brownson, R. C., Gurney, J., & Land, G. (1999). Evidence-Based Decision Making in Public Health. *Journal of Public Health Management and Practice*, *5*(5), 86–97.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (2002). Geographically weighted summary statistics — a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, *26*(6), 501–524.

Buchel, O., & Sedig, K. (2014). Making sense of document collections with map-based visualisations: Role of interaction. *Information Research*, *19*(3).

Campbell-Lendrum, D., Manga, L., Bagayoko, M., Sommerfeld, J., WHO, Lozano, R., Murray, C., et al. (2015). Climate change and vector-borne diseases: what are the implications for public health research and policy? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *370*(1665), 2095–2128. The Royal Society.

Cao, N., & Cui, W. (2016). Visualizing Sentiments and Emotions. *Introduction to Text Visualization* (pp. 103–114). Paris: Atlantis Press.

Cao, N., Lin, Y.-R., Gotz, D., & Du, F. (2017). Z-Glyph: Visualizing outliers in multivariate data. *Information Visualization*, 1–19.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. *The Morgan Kaufmann series in interactive technologies*. San Francisco, Calif: Morgan Kaufmann Publishers.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *49*(10), 1557–64.

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T.-C., Painter, I. S., & Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of biomedical informatics*, *51C*, 287–298.

Centers for Disease Control and Prevention. (2012). Epi Info. Atlanta: Centers for Disease Control and Prevention. Retrieved from http://wwwn.cdc.gov/epiinfo/7/index.htm

Chan, M. (2014). Ebola Virus Disease in West Africa — No Early End to the Outbreak. *New England Journal of Medicine*, *371*(13), 1183–1185. Massachusetts Medical Society.

Chareonviriyaphap, T., Bangs, M. J., Suwonkerd, W., Kongmee, M., Corbel, V., & Ngoen-Klan, R. (2013). Review of insecticide resistance and behavioral avoidance of vectors of human diseases in Thailand. *Parasites & vectors*, *6*(1), 280.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H. Y., Olsen, J. M., Pavlin, J. A., et al. (2015). Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. *PloS one*, *10*(10), e0139701.

Charrel, R., Leparc-Goffart, I., Gallian, P., & de Lamballerie, X. (2014). Globalization of

Chikungunya: 10 years to invade the world. *Clinical microbiology and infection*, *20*(7), 662–663.

Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. *International Conference on Database Systems for Advanced Applications* (pp. 1–15). Springer Berlin Heidelberg.

Chew, C., & Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. (M. Sampson, Ed.)*PLoS ONE*, *5*(11), e14118. Public Library of Science.

Chou, W. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the United States: implications for health communication. *Journal of medical Internet research*, *11*(4), e48. Journal of Medical Internet Research.

Cohen, B. (1984). Florence Nightingale. *Scientific American*, *250*(3), 128–137.

Cohen, B. B., Franklin, S., & West, J. K. (2006). Perspectives on the Massachusetts Community Health Information Profile (MassCHIP): developing an online data query system to target a variety of user needs and capabilities. *Journal of Public Health Management and Practice : JPHMP*, *12*(2), 155–60.

Cole-Lewis, H., Varghese, A., Sanders, A., Schwarz, M., Pugatch, J., & Augustson, E. (2015). Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. *Journal of medical Internet research*, *17*(8), e208.

Committee on Educating Public Health Professionals for the 21st Century. (2003). *Who Will Keep the Public Healthy? Educating Public Health Professionals for the 21st Century*. (K. Gebbie, L. Rosenstock, & L. M. Hernandez, Eds.). Washington, D.C.: The National Academies Press.

Cybulski, J. L., Keller, S., Nguyen, L., & Saundage, D. (2013). Creative problem solving in digital space using visual analytics. *Computers in Human Behavior*.

D'Alessandro, U., & Buttiëns, H. (2001). History and importance of antimalarial drug resistance. *Tropical Medicine and International Health*, *6*(11), 845–8.

Davies, C., Fabrikant, S. I., & Hegarty, M. (2013). Towards Empirically Verified Cartographic Displays. In J. Szalma, M. Scerbo, P. Hancock, R. Parasuraman, & R. Hoffman (Eds.), *Cambridge Handbook of Applied Perception Research*. Cambridge, U.K.: Cambridge University Press.

Devarajan, D. (2017). Retirement of AlchemyAPI service. Retrieved July 17, 2017, from https://www.ibm.com/blogs/bluemix/2017/03/bye-bye-alchemyapi/

Dhar, V. (2014). Big Data and Predictive Analytics in Health Care. *Big Data*, *2*(3), 113–116. Mary Ann Liebert, Inc.

Dou, W., Ziemkiewicz, C., Harrison, L., Jeong, D. H., Ribarsky, W., Wang, X., & Chang, R. (2012). Toward a Deeper Understanding of the Relationship between Interaction Constraints and Visual Isomorphs. *Information Visualization*, *11*(3), 222–236. SAGE Publications.

Dou, W., Ziemkiewicz, C., Harrison, L., Jeong, D. H., Ryan, R., Ribarsky, W., Wang, X., et al. (2010). Comparing different levels of interaction constraints for deriving visual problem isomorphs. *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (pp. 195–202). IEEE.

*Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations*. (2011). . Mountain View,CA.

Eisen, L., & Eisen, R. J. (2007). Need for improved methods to collect and present spatial epidemiologic data for vectorborne diseases. *Emerging infectious diseases*, *13*(12), 1816–20.

Eisen, L., & Eisen, R. J. (2011). Using geographic information systems and decision support systems for the prediction, prevention, and control of vector-borne diseases. *Annual review of entomology*, *56*, 41–61.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., et al. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, *66*(8), 1523–1545.

Elmqvist, N., Moere, a. V., Jetter, H.-C., Cernea, D., Reiterer, H., & Jankun-Kelly, T. (2011). Fluid interaction for information visualization. *Information Visualization*, *10*(4), 327–340.

Endert, A., Chang, R., North, C., & Zhou, M. (2015). Semantic Interaction: Coupling Cognition and Computation through Usable Interactive Analytics. *IEEE Computer Graphics and Applications*, *35*(4).

Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., & Andrews, C. (2014). The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*.

Endert, A., North, C., Chang, R., & Zhou, M. (2014). Toward Usable Interactive Analytics: Coupling Cognition and Computation. *KDD 2014 Workshop on Interactive Data Exploration and Analytics (IDEA)*.

Endert, A., Ribarsky, W., Turkay, C., Wong, B. L. W., & Nabney, I. (2017). The State of the Art in Integrating Machine Learning into Visual, *0*(0), 1–28.

England, I., Stewart, D., & Walker, S. (2000). Information technology adoption in health care: when organisations and technology collide. *Australian Health Review*, *23*(3), 176–85.

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research*. Routledge.

Faisal, S., Blandford, A., & Potts, H. W. W. (2013). Making sense of personal health information: challenges for information visualization. *Health informatics journal*, *19*(3), 198–217.

Fan, W., & Bifet, A. (2013). Mining big data. *ACM SIGKDD Explorations Newsletter*, *14*(2), 1–5. ACM.

Faria, N., Azevedo, R., Kraemer, M., Souza, R., Cunha, M., Hill, S., Thezé, J., et al.

(2016, March 24). Data from: Zika Virus in the Americas: early epidemiological and genetic findings. *Science*. Dryad Data Repository.

Faria, N. R., Azevedo, R. do S. da S., Kraemer, M. U. G., Souza, R., Cunha, M. S., Hill, S. C., Thézé, J., et al. (2016). Zika virus in the Americas: Early epidemiological and genetic findings. *Science (New York, N.Y.)*, *352*(6283), 345–9. American Association for the Advancement of Science.

Finfgeld-Connett, D. (2015). Twitter and Health Science Research. *Western journal of nursing research*, *37*(10), 1269–83.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. *Interactions*, *19*, 50.

Folorunso, O., & Ogunseye, O. S. (2008). Challenges in the adoption of visualization system: a survey. (M. Chen, Ed.)*Kybernetes*, *37*(9/10), 1530–1541.

Ford, D. A., Kaufman, J. H., & Eiron, I. (2006). An extensible spatial and temporal epidemiological modelling system. *International Journal of Health Geographics*, *5*(4), 1–6.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association : JAMIA*, *15*(2), 150–7. BMJ Publishing Group Ltd.

Fuller, S. (2010). Tracking the Global Express: new tools addressing disease threats across the world. *Epidemiology (Cambridge, Mass.)*, *21*(6), 769–71.

Gadanidis, G., Sedig, K., & Liang, H.-N. (2004). Designing online mathematical investigation. *Journal of Computers in Mathematics and Science Teaching*, *23*(3), 275–298.

Gapminder. (n.d.). . Retrieved from http://www.gapminder.org/

Gazmararian, J. A., Curran, J. W., Parker, R. M., Bernhardt, J. M., & DeBuono, B. A. (2005). Public health literacy in America: an ethical imperative. *American journal of preventive medicine*, *28*(3), 317–22.

Ghosh, D. (Debs), & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90–102. Taylor & Francis.

Gilhooly, K. (2004). Working memory and reasoning. In J. Leighton & R. Sternberg (Eds.), *The Nature of Reasoning* (pp. 49–77). New York: Cambridge University Press.

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language*, *31*(2), 129–151.

Goddard, M., Mowat, D., Corbett, C., Neudorf, C., Raina, P., & Sahai, V. (2004). The Impacts of Knowledge Management and Information Technology Advances on

Public Health Decision-Making in 2010. *Health Informatics Journal*, *10*(2), 111–120.

Goodchild, M. F. (2004). The Validity and Usefulness of Laws in Geographic Information Science and Geography. *Annals of the Association of American Geographers*, *94*(2), 300–303. Blackwell Publishing.

Gorodov, E. Y., & Gubarev, V. V. (2013). Analytical Review of Data Visualization Methods in Application to Big Data. *Journal of Electrical and Computer Engineering*, *2013*, 1–7. Hindawi Publishing Corp.

Gotz, D., & Borland, D. (2016). Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. *IEEE Computer Graphics and Applications*, *36*(3), 90–96.

Green, T. M., & Maciejewski, R. (2013). A Role for Reasoning in Visual Analytics. *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 1495–1504).

Greitzer, F. L., Noonan, C. F., & Franklin, L. R. (2011). *Cognitive Foundations for Visual Analytics*.

Groth, D. P., & Streefkerk, K. (2006). Provenance and Annotation for Visual Exploration Systems. *IEEE Transactions on Visualization and Computer Graphics*, *12*(6), 1500–1510.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). *The big-data revolution in US health care: Accelerating value and innovation*. Retrieved May 16, 2016, from http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

Guerra, C. A., Hay, S. I., Lucioparedes, L. S., Gikandi, P. W., Tatem, A. J., Noor, A. M., & Snow, R. W. (2007). Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malaria Journal*, *6*(17).

Guo, D., Gahegan, M., MacEachren, A. M., & Zhou, B. (2005). Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*, *32*(2), 113–132.

Gurman, T. A., & Clark, T. (2016). #ec: Findings and implications from a quantitative content analysis of tweets about emergency contraception. *Digital Health*, *2*, 2055207615625035. SAGE Publications.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Harris, P., Clarke, A., Juggins, S., Brunsdon, C., & Charlton, M. (2014). Geographically weighted methods and their use in network re-designs for environmental monitoring. *Stochastic Environmental Research and Risk Assessment*, *28*(7), 1869–1887. Springer Berlin Heidelberg.

Hartemink, N., Vanwambeke, S. O., Purse, B. V., Gilbert, M., & Van Dyck, H. (2015). Towards a resource-based habitat approach for spatial modelling of vector-borne disease risks. *Biological Reviews*, *90*(4), 1151–1162. Blackwell Publishing Ltd.

Harvey, C. R. (2000). The Drivers of Expected Returns in International Markets. *SSRN Electronic Journal*. Retrieved September 18, 2016, from http://www.ssrn.com/abstract=795385

Heaivilin, N., Gerbert, B., Page, J. E., & Gibbs, J. L. (2011). Public health surveillance of dental pain via Twitter. *Journal of dental research*, *90*(9), 1047–51. SAGE Publications.

Health literacy: report of the Council on Scientific Affairs. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association. (1999). *JAMA*, *281*(6), e2188–e2188.

Heer, J. (2013). Interactive Visualization of Big Data. *O'Reilly Strata*. Retrieved February 22, 2017, from http://radar.oreilly.com/2013/12/interactive-visualization-of-big-data.html

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A Tour through the Visualization Zoo. *ACM Queue*, *8*(5), 20. ACM.

Heer, J., & Kandel, S. (2012). Interactive analysis of big data. *XRDS: Crossroads, The ACM Magazine for Students*, *19*(1), 50. ACM.

Heggenhougen, H. K., Hackethal, V., & Vivek, P. (2003). *The behavioural and social aspects of malaria and its control: an introduction and annotated bibliography.* Geneva: World Health Organization.

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal Of Big Data*, *1*(1), 2. Springer.

Heuer, R. (1999). *Psychology of Intelligence Analysis*. Washington, DC: Central Intelligence Agency, Center for the Study of Intelligence.

Higgins, J. W., Strange, K., Scarr, J., Pennock, M., Barr, V., Yew, A., Drummond, J., et al. (2011). "It's a Feel. That's What a Lot of Our Evidence Would Consist of ": Public Health Practitioners' Perspectives on Evidence. *Evaluation & the Health Professions*, *34*(3), 278–296. SAGE Publications.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196.

Homan, T., Maire, N., Hiscox, A., Di Pasquale, A., Kiche, I., Onoka, K., Mweresa, C., et al. (2016). Spatially variable risk factors for malaria in a geographically heterogeneous landscape, western Kenya: an explorative study. *Malaria Journal*, *15*(1), 1. BioMed Central.

Hornbæk, K., & Hertzum, M. (2011). The notion of overview in information visualization. *International Journal of Human Computer Studies*, *69*(7–8), 509–525.

Huang, Z., Das, A., Qiu, Y., & Tatem, A. J. (2012). Web-based GIS: the vector-borne disease airline importation risk (VBD-AIR) tool. *International journal of health geographics*, *11*(1), 33.

Hughes, E. (2016). Can Twitter improve your health? An analysis of alcohol

consumption guidelines on Twitter. *Health Information & Libraries Journal*, *33*(1), 77–81.

Institute for Health Metrics and Evaluation. (2013). Global Burden of Disease. Retrieved from http://www.healthdata.org/gbd

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94.

Kamel Boulos, M. N. (2013). Social media and mobile health. In I. Kickbusch, J. M. Pelikan, F. Apfel, & A. D. Tsouros (Eds.), *Health literacy: the solid facts* (pp. 63–67). Copenhagen: WHO Regional Office for Europe.

Kamel Boulos, M. N., Viangteeravat, T., Anyanwu, M. N., Ra Nagisetty, V., & Kuscu, E. (2011). Web GIS in practice IX: a demonstration of geospatial visual analytics using Microsoft Live Labs Pivot technology and WHO mortality data. *International journal of health geographics*, *10*, 19.

Katsis, Y., Koulouris, N., Papakonstantinou, Y., & Patrick, K. (2017). Assisting Discovery in Public Health. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics - HILDA'17* (pp. 1–6). New York, New York, USA: ACM Press.

Kaufman, D. R., Kannampallil, T. G., & Patel, V. L. (2015). Cognition and Human Computer Interaction in Health and Biomedicine (pp. 9–34). Springer International Publishing.

Keim, D. A., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In A. Kerren, J. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization* (pp. 154–175). Berlin Heidelberg: Springer-Verlag.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering The Information Age-Solving Problems with Visual Analytics*. Eurographics Association.

Keim, D. A., Mansmann, F., & Thomas, J. (2009). Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, *11*(2), 5–8. ACM.

Kelly, G. C., Tanner, M., Vallely, A., & Clements, A. (2012). Malaria elimination: moving forward with spatial decision support systems. *Trends in Parasitology*, *28*(7), 297–304.

Keough, K. (2002). The Third Amyot Lecture. How science informs the decisions of government. *Canadian Journal of Public Health. Revue canadienne de santé publique*, *93*(2), 104–8.

Khan, A. A. (1992). An integrated approach to measuring potential spatial access to health care services. *Socio-Economic Planning Sciences*, *26*(4), 275–287. Pergamon.

Kickbusch, I., Pelikan, J., Apfel, F., & Tsouros, A. (2013). Health literacy: the solid facts. *Copenhagen: WHO Regional Office for …*, 7–8.

Kiefer, L., Frank, J., Di Ruggiero, E., Dobbins, M., Manuel, D., Gully, P. R., & Mowat, D. (2005). Fostering evidence-based decision-making in Canada: examining the need for a Canadian population and public health evidence centre and research network. *Canadian Journal of Public Health. Revue canadienne de santé publique*, *96*(3), I1-40.

Kilpatrick, A. M., & Randolph, S. E. (2012). Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *The Lancet*, *380*(9857), 1946–1955.

Kirsh, D. (2009). Interaction, External Representation and Sense Making. *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1103–1108).

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction*, *20*(1).

Kitchin, R., & Dodge, M. (2007). Rethinking maps. *Progress in Human Geography*, *31*(3), 331–344.

Knauff, M., & Wolf, A. G. (2010). Complex cognition: The science of human reasoning, problem-solving, and decision-making. *Cognitive Processing*, *11*(2), 99–102.

Ko, S., Cho, I., Afzal, S., Yau, C., Chae, J., Malik, A., Beck, K., et al. (2016). A Survey on Visual Analysis Approaches for Financial Data. *Computer Graphics Forum*, *35*(3), 599–617.

Koita, K., Novotny, J., Kunene, S., Zulu, Z., Ntshalintshali, N., Gandhi, M., & Gosling, R. (2013). Targeting imported malaria through social networks: a potential strategy for malaria elimination in Swaziland. *Malaria journal*, *12*(1), 219.

Korda, H., & Itani, Z. (2013). Harnessing Social Media for Health Promotion and Behavior Change. *Health Promotion Practice*, *14*(1), 15–23. SAGE PublicationsSage CA: Los Angeles, CA.

Kosara, R., & Miksch, S. (2002). Visualization methods for data analysis and planning in medical applications. *International Journal of Medical Informatics* (Vol. 68, pp. 141–153).

Koua, E. L., & Kraak, M.-J. (2004). Geovisualization to support the exploration of large health and demographic survey data. *International journal of health geographics*, *3*(1), 12. BioMed Central.

Kraak, M.-J. (2006). Visualization viewpoints: beyond geovisualization. *IEEE Computer Graphics and Applications*, *26*(4), 6–9.

Kraemer, M. U. G., Sinka, M. E., Duda, K. A., Mylne, A. Q. N., Shearer, F. M., Barker, C. M., Moore, C. G., et al. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife*, *4*, e08347. eLife Sciences Publications Limited.

Kraemer, M. U. G., Sinka, M. E., Duda, K. A., Mylne, A., Shearer, F. M., Brady, O. J., Messina, J. P., et al. (2015). The global compendium of Aedes aegypti and Ae. albopictus occurrence. *Scientific Data*, *2*. Nature Publishing Group.

Kramer, R. A., Dickinson, K. L., Anderson, R. M., Fowler, V. G., Miranda, M. L.,

Mutero, C. M., Saterson, K. A., et al. (2009). Using decision analysis to improve malaria control policy making. *Health Policy*, *92*(2–3), 133–140.

Krauss, M. J., Grucza, R. A., Bierut, L. J., & Cavazos-Rehg, P. A. (2016). "Get drunk. Smoke weed. Have fun." *American Journal of Health Promotion*, *31*(3).

Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016). Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR medical informatics*, *4*(4), e38. JMIR Medical Informatics.

Kwon, B. C., & Lee, B. (2016). A Comparative Evaluation on Online Learning Approaches Using Parallel Coordinate Visualization. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 993–997. New York, New York, USA: ACM Press.

LaDeau, S. L., Allan, B. F., Leisnham, P. T., & Levy, M. Z. (2015). The ecological foundations of transmission potential and vector-borne disease in urban landscapes. (K. Evans, Ed.)*Functional Ecology*, *29*(7), 889–901.

Lam, H. (2008). A framework of interaction costs in information visualization. *IEEE transactions on visualization and computer graphics*, *14*(6), 1149–56. IEEE.

LaPelle, N. R., Luckmann, R., Simpson, E. H., & Martin, E. R. (2006). Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. *BMC public health*, *6*(1), 89–101.

Larkin, J., & Simon, H. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, *11*(1), 65–100.

Lee, S., Kim, S.-H., Hung, Y.-H., Lam, H., Kang, Y., & Yi, J. (2015). How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 1–1.

Lee, S., Kim, S.-H., & Kwon, B. C. (2017). VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 551–560.

Leighton, J. (2004). Defining and describing reason. In J. Leighton & R. Sternberg (Eds.), *The Nature of Reasoning* (pp. 3–11). New York: Cambridge University Press.

Lemire, M., Sicotte, C., & Paré, G. (2008). Internet use and the logics of personal empowerment in health. *Health Policy*, *88*(1), 130–140.

Liang, H.-N. H.-N., & Sedig, K. (2010). Role of interaction in enhancing the epistemic utility of 3D mathematical visualizations. *International Journal of Computers for Mathematical Learning*, *15*(3), 191–224.

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, *380*(9859), 2224–60.

Liu, E., Zhao, Y., Wei, H., Roumeliotis, S., & Kaldoudi, E. (2016). Navigating Health Literacy Using Interactive Data Visualisation. *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 44–50). IEEE.

Liu, Y., Jiang, S., Liu, Y., Wang, R., Li, X., Yuan, Z., Wang, L., et al. (2011). Spatial epidemiology and spatial ecology study of worldwide drug-resistant tuberculosis. *International Journal of Health Geographics*, *10*(1). BioMed Central.

Liu, Z., & Stasko, J. (2010). Mental models, visual reasoning and interaction in information visualization: a top-down perspective. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 999–1008.

Livnat, Y., Rhyne, T.-M., & Samore, M. (2012). Epinome: A Visual-Analytics Workbench for Epidemiology Data. *IEEE Computer Graphics and Applications*, *32*(2), 89–95.

Longley, P. A., Goodchild, M., Maguire, D. J., & Rhind, D. W. (2005). *Geographic Information Systems and Science* (2nd ed.). New York: Wiley.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, *380*(9859), 2095–128.

Lu, B., Harris, P., Gollini, I., Charlton, M., & Brunsdon, C. (2011). *Introducing the GWmodel R and python packages for modelling spatial heterogeneity*.

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, 1.

Luz, P. M., Vanni, T., Medlock, J., Paltiel, A. D., & Galvani, A. P. (2011). Dengue vector control strategies in an urban setting: an economic modelling assessment. *The Lancet*, *377*(9778), 1673–1680.

MacEachren, A. M., Stryker, M. S., Turton, I. J., & Pezanowski, S. (2010). HEALTH GeoJunction: place-time-concept browsing of health publications. *International Journal of Health Geographics*, *9*(1), 23.

Maciejewski, R., Livengood, P., Rudolph, S., Collins, T. F., Ebert, D. S., Brigantic, R. T., Corley, C. D., et al. (2011). A pandemic influenza modeling and visualization tool. *Journal of Visual Languages & Computing*, *22*(4), 268–278.

Mackinlay, J. D. (2000). Opportunities for information visualization. *IEEE Computer Graphics and Applications*, *20*(1), 22–23.

Mandl, K. D., Overhage, J. M., Wagner, M. M., Lober, W. B., Sebastiani, P., Mostashari, F., Pavlin, J. A., et al. (2004). Implementing syndromic surveillance: a practical guide informed by the early experience. *Journal of the American Medical Informatics Association : JAMIA*, *11*(2), 141–50.

Matthews, S. A., & Yang, T.-C. (2012). Mapping the results of local statistics. *Demographic Research*, *26*, 151–166.

May, R., Hanrahan, P., Keim, D. A., Shneiderman, B., & Card, S. K. (2010). The state of

visual analytics: Views on what visual analytics is and where it is going. *2010 IEEE Symposium on Visual Analytics Science and Technology* (pp. 257–259). IEEE.

Meehan, K., Lunney, T., Curran, K., & McCaughey, A. (2013). Context-aware intelligent recommendation system for tourism. *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (pp. 328–331). IEEE.

Mendis, K., Rietveld, A., Warsame, M., Bosman, A., Greenwood, B., & Wernsdorfer, W. H. (2009). From malaria control to eradication: The WHO perspective. *Tropical medicine & international health : TM & IH*, *14*(7), 802–9.

Mitchell, A. (2005). *The ESRI guide to GIS analysis*. Redlands, CA: ESRI.

Mnzava, A. P., Knox, T. B., Temu, E. A., Trett, A., Fornadel, C., Hemingway, J., Renshaw, M., et al. (2015). Implementation of the global plan for insecticide resistance management in malaria vectors: progress, challenges and the way forward. *Malaria Journal*, *14*(1), 173. BioMed Central.

Mokdad, A. H., Marks, J. S., Stroup, D. F., & Gerberding, J. L. (2004). Actual Causes of Death in the United States, 2000. *JAMA*, *291*(10), 1238–1245. American Medical Association.

Moody, D. (2007). What Makes a Good Diagram? Improving the Cognitive Effectiveness of Diagrams in IS Development. *Advances in Information Systems Development* (pp. 481–492). Boston, MA: Springer US.

Mowat, D. L., & Hockin, J. (2002). Building capacity in evidence-based public health practice. *Can J Public Health*, *93*(1), 19–20.

Munzner, T. (2014). *Visualization Analysis and Design*. New York: A K Peters/CRC Press.

Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, *380*(9859), 2197–223.

Myslín, M., Zhu, S.-H., Chapman, W., & Conway, M. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, *15*(8), e174. Journal of Medical Internet Research.

Nakhapakorn, K., & Tripathi, N. K. (2005). An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. *International journal of health geographics*, *4*, 1–13.

National Research Council. (1988). *The Future of Public Health*. Washington, D.C., USA: The National Academies Press. Retrieved April 11, 2013, from http://www.nap.edu/catalog.php?record_id=1091

Norman, D. A. (2013). *The design of everyday things*. Basic Books.

Nutbeam, D. (2000). Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health*

*Promotion International*, *15*(3), 259–267.

O'Carroll, P. W. (2003). Introduction to Public Health Informatics. In P. W. O'Carroll, W. A. Yasnoff, M. E. Ward, L. H. Ripp, & E. L. Martin (Eds.), *Public Health Informatics and Information Systems*, Health Informatics (Vol. 6, pp. 3–15). New York, NY, USA: Springer-Verlag.

O'Carroll, P. W., Cahn, M. A., Auston, I., & Selden, C. R. (1998). Information needs in public health and health policy: results of recent studies. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, *75*(4), 785–793.

Occa, A., & Suggs, L. S. (2016). Communicating Breast Cancer Screening With Young Women: An Experimental Test of Didactic and Narrative Messages Using Video and Infographics. *Journal of Health Communication*, *21*(1), 1–11. 2016.

Ola, O., Buchel, O., & Sedig, K. (2016). Exploring the Spread of Zika: Using Interactive Visualizations to Control Vector-Borne Diseases. *International Journal of Disease Control and Containment for Sustainability*, *1*(1), 47–68.

Ola, O., & Sedig, K. (2014). The challenge of big data in public health: an opportunity for visual analytics. *Online journal of public health informatics*, *5*(3), 1–21.

Ola, O., & Sedig, K. (2016). Beyond simple charts : Design of visualizations for big health data. *Online Journal of Public Health Informatics*, *8*(3).

Ooms, J. (2014). The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. Retrieved September 18, 2016, from http://arxiv.org/abs/1406.4806

Ortiz-Martínez, Y., & Jiménez-Arcia, L. F. (2017). Yellow fever outbreaks and Twitter: Rumors and misinformation. *American Journal of Infection Control*, *45*(7), 816–817.

Osborne, H. (2012). Using Twitter and Other Social Media to Communicate About Health Literacy (HLOL #80). Retrieved July 18, 2017, from http://healthliteracy.com/2012/07/10/using-twitter-and-other-social-media-to-communicate-about-health-literacy-hlol-80/

Oyeyemi, S. O., Gabarron, E., & Wynn, R. (2014). Ebola, Twitter, and misinformation: a dangerous combination? *British Medical Journal*, *349*(October), g6178.

Palomino, M., Taylor, T., Göker, A., Isaacs, J., & Warber, S. (2016). The Online Dissemination of Nature-Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder". *International journal of environmental research and public health*, *13*(1), 142. Multidisciplinary Digital Publishing Institute.

Paquet, C., Coulombier, D., Kaiser, R., & Ciotti, M. (2006). Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, *11*(12), 212–4.

Park, H., Reber, B. H., & Chon, M.-G. (2016). Tweeting as Health Communication: Health Organizations' Use of Twitter for Health Promotion and Public Engagement.

*Journal of Health Communication*, *21*(2), 188–198. Routledge.

Park, H., Rodgers, S., & Stemmle, J. (2013). Analyzing Health Organizations' Use of Twitter for Promoting Health Literacy. *Journal of Health Communication*, *18*(4), 410–425.

Parsons, P., & Sedig, K. (2013a). Adjustable Properties of Visual Representations: Improving the Quality of Human-Information Interaction. *Journal of the Association for Information Science and Technology*, *65*(3), 455–482.

Parsons, P., & Sedig, K. (2013b). Distribution of Information Processing while Performing Complex Cognitive Activities with Visualization Tools. In T. Huang (Ed.), *Handbook of Human Centric Visualization* (pp. 639–715). New York: Springer.

Parsons, P., & Sedig, K. (2014). Common Visualizations: Their Cognitive Utility. In W. Huang (Ed.), *Handbook of Human Centric Visualization* (pp. 671–691). New York, NY: Springer New York.

Paul, M. J., & Dredze, M. (2014). Discovering Health Topics in Social Media Using Topic Models. (R. Lambiotte, Ed.)*PLoS ONE*, *9*(8), e103408. Public Library of Science.

Pike, W. a, Stasko, J., Chang, R., & O'Connell, T. A. (2009). The science of interaction. *Information Visualization*, *8*(4), 263–274.

Pirolli, P., & Card, S. K. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *2005 Internaltional Conference on Intelligence Analysis* (p. 6 pp.).

Pretorius, A. J., Khan, I. A., & Errington, R. J. (2016). A Survey of Visualization for Live Cell Imaging. *Computer Graphics Forum*, *36*(1), 46–63.

Proulx, P., Tandon, S., Bodnar, A., Schroh, D., Harper, R., & Wright, W. (2006). Avian Flu Case Study with nSpace and GeoTime. *2006 IEEE Symposium On Visual Analytics And Technology* (pp. 27–34). IEEE.

Purchase, H. C., Andrienko, N., Jankun-Kelly, T., & Ward, M. (2008). Theoretical foundations of information visualization. (A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North, Eds.)*Information Visualization. Human-Centered Issues and Perspectives*, Lecture Notes in Computer Science, *4950*, 46–64. Berlin, Heidelberg: Springer Berlin Heidelberg.

Purushe, S., Grinstein, G., Smrtic, M. B., & Lyons, H. (2011). Interactive Animated Visualizations of Breast, Ovarian Cancer and Other Health Indicator Data Using Weave, an Interactive Web-based Analysis and Visualization Environment. *2011 15th International Conference on Information Visualisation* (pp. 247–252). IEEE.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. BioMed Central Ltd.

Rambo, N. (1998). Information resources for public health practice. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, *75*(4), 807–25.

Rambo, N. (2000). Information Needs and Uses of the Public Health Workforce -- Washington, 1997-1998. *Morbidity and Mortality Weekly Report*, *49*(6), 118–120.

Rambo, N., Zenan, J. S., Alpi, K. M., Burroughs, C. M., Cahn, M. A., & Rankin, J. (2001). Public Health Outreach Forum: lessons learned. *Bulletin of the Medical Library Association*, *89*(4), 403–6.

Rasu, R. S., Bawa, W. A., Suminski, R., Snella, K., & Warady, B. (2015). Health Literacy Impact on National Healthcare Utilization and Expenditure. *International Journal of Health Policy and Management*, *4*(11), 747–755.

Reeder, B., Revere, D., Hills, R. A., Baseman, J. G., & Lober, W. B. (2012). Public Health Practice within a Health Information Exchange: Information Needs and Barriers to Disease Surveillance. *Online Journal of Public Health Informatics*, *4*(3).

Reinhardt, W., Schmidt, B., Sloep, P., & Drachsler, H. (2011). Knowledge Worker Roles and Actions-Results of Two Empirical Studies. *Knowledge and Process Management*, *18*(3), 150–174.

Revere, D., Turner, A. M., Madhavan, A., Rambo, N., Bugni, P. F., Kimball, A., & Fuller, S. S. (2007). Understanding the information needs of public health practitioners: A literature review to inform design of an interactive digital knowledge management system. *Public Health Informatics*, *40*(4), 410–421.

Rind, A., Aigner, W., Wagner, M., Miksch, S., & Lammarsch, T. (2015). Task Cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation. *Information Visualization*, 1473871615621602-.

Rind A., Wang T., Aigner W., Miksch S., Wongsuphasawat K., Plaisant C., & Shneiderman, B. (2013). Interactive Information Visualization to Explore and Query Electronic Health Records. *Foundations and Trends in Human-Computer Interaction*, *5*(3), 207–298.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE transactions on visualization and computer graphics*, *14*(6), 1325–32.

Robinson, A. C., MacEachren, A. M., & Roth, R. E. (2011). Designing a web-based learning portal for geographic visualization and analysis in public health. *Health informatics journal*, *17*(3), 191–208.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2), 1–22. The Association for Research in Vision and Ophthalmology.

Rowlands, G., Shaw, A., Jaswal, S., Smith, S., & Harpham, T. (2017). Health literacy and the social determinants of health: a qualitative model from adult learners. *Health Promotion International*, *32*(1), 130–138.

Ruchikachorn, P., & Mueller, K. (2015). Learning Visualizations by Analogy: Promoting Visual Literacy through Visualization Morphing. *IEEE Transactions on Visualization and Computer Graphics*, *21*(9), 1028–1044.

Saif, H., He, Y., & Alani, H. (2012). Semantic Sentiment Analysis of Twitter. *International Semantic Web Conference* (pp. 508–524). Springer, Berlin,

Heidelberg.

Salathé, M., & Khandelwal, S. (2011). Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. (L. A. Meyers, Ed.)*PLoS Computational Biology*, *7*(10), e1002199. Oxford University Press.

Schabas, R. (2002). Is public health ethical? *Can J Public Health*, *93*(2), 98–99.

Schein, R., Wilson, K., & Keelan, J. (2010). *Literature review on effectivess of the use of social media: A report for Peel Public Health*.

Schlegel, K., Sallam, R. L., Daniel, Y., & Tapadinhas, J. (2013). *Magic quadrant for business intelligence platforms.* Retrieved from http://www.gartner.com/technology/reprints.do?id=1-1DZLPEP&ct=130207&st=sb

Schulz, P. J., & Nakamoto, K. (2013). Health literacy and patient empowerment in health communication: The importance of separating conjoined twins. *Patient Education and Counseling*, *90*(1), 4–11.

Sedig, K. (2001, October 16). A knowledge Management Support Tool: Visualizing Design Activities. *Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI*.

Sedig, K. (2009). Interactive mathematical visualizations: Frameworks, tools, and studies. In E. Zudilova-Seinstra, T. Adriaansen, & R. van Liere (Eds.), *Trends in Interactive Visualization* (pp. 343–363). London: Springer-Verlag.

Sedig, K., Haworth, R., & Corridore, M. (2015). Investigating variations in gameplay: Cognitive implications. *International Journal of Computer Games Technology*, *2015*. Hindawi Publishing Corporation.

Sedig, K., & Parsons, P. (2013). Interaction design for cognitive activity support tools: A pattern-based taxonomy. *AIS Transactions on Human-Computer Interaction*, *5*(2), 84–133.

Sedig, K., & Parsons, P. (2016). Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework. *Synthesis Lectures on Visualization*, *4*(1), 1–185. Morgan & Claypool Publishers.

Sedig, K., Parsons, P. C., & Babanski, A. (2012). Towards a characterization of interactivity in visual analytics. *Journal of Multimedia Processing*, *3*(1), 12–28.

Sedig, K., Parsons, P. C., Dittmer, M., & Haworth, R. (2013). Human-centered interactivity of visualization tools: Micro- and macro-level considerations. In T. Huang (Ed.), *Handbook of Human Centric Visualization* (pp. 717–743). Springer.

Sedig, K., Parsons, P., Dittmer, M., & Ola, O. (2012). Beyond information access: Support for complex cognitive activities in public health informatics tools. *Online Journal of Public Health Informatics*, *4*(3).

Sedig, K., Parsons, P., Liang, H., & Morey, J. (2016). Supporting Sensemaking of Complex Objects with Visualizations: Visibility and Complementarity of Interactions. *Informatics*, *3*(4), 20.

Sedig, K., Rowhani, S., & Liang, H.-N. (2005). Designing interfaces that support formation of cognitive maps of transitional processes: an empirical study. *Interacting with Computers*, *17*(4), 419–452.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38.

Sharman, J. L., & Gerloff, D. L. (2013). MaGnET: Malaria Genome Exploration Tool. *Bioinformatics*, *29*(18), 2350–2.

Shneiderman, B., Plaisant, C., & Hesse, B. W. (2013). Improving Healthcare with Interactive Visualization. *Computer*, *46*(5), 58–66.

Shortliffe, E. H. (2005). Strategic action in health information technology: why the obvious has taken so long. *Health Affairs (Project Hope)*, *24*(5), 1222–33.

Siirtola, H., & Räihä, K.-J. (2006). Interacting with parallel coordinates. *Interacting with Computers*, *18*(6), 1278–1309. Oxford University Press.

Silge, J., & Robinson, D. (2017). Term Frequency and Inverse Document Frequency (tf-idf) Using Tidy Data Principles. Retrieved July 19, 2017, from https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html

Sims, J. N., Isokpehi, R. D., Cooper, G. A., Bass, M. P., Brown, S. D., St John, A. L., Gulig, P. A., et al. (2011). Visual analytics of surveillance data on foodborne vibriosis, United States, 1973-2010. *Environmental health insights*, *5*, 71–85.

Smith, M. J. de, Goodchild, M. F., & Longley, P. A. (2006). *Geospatial Analysis*. Troubador Publishing Ltd.

Snow, J. (1855). *On the mode of communication of cholera* (2nd ed.). London: Churchill.

Soille, P. (2003). *Morphological Image Analysis: Principles and Applications* (2nd ed.). Heidelberg: Springer-Verlag.

Sopan, A., Noh, A. S.-I., Karol, S., Rosenfeld, P., Lee, G., & Shneiderman, B. (2012). Community Health Map: A geospatial and multivariate data visualization tool for public health datasets. *Government Information Quarterly*, *29*(2), 223–234.

Sørensen, K. (2017). The Need for "Health Twitteracy" in a Postfactual World. *HLRP: Health Literacy Research and Practice*, *1*(2), e86–e89.

Sørensen, K., Van den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z., & Brand, H. (2012). Health literacy and public health: a systematic review and integration of definitions and models. *BMC public health*, *12*(1), 80.

Sørensen, K., Pelikan, J. M., Röthlin, F., Ganahl, K., Slonska, Z., Doyle, G., Fullam, J., et al. (2015). Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *European journal of public health*, *25*(6), 1053–8. Oxford University Press.

Spence, R. (2007). *Information Visualization: Design for Interaction* (2nd ed., Vol. 2). Harlow ;; New York: Pearson Prentice Hall.

Spence, R. (2014). *Information Visualization: An Introduction* (3rd ed.). Heidelberg:

Springer.

StataCorp. (2009). Stata Staistical Software: Release 12. College Station, TX: StataCorp LP.

Stevenson, L. G. (1965). Putting disease on the map. The early use of spot maps in the study of yellow fever. *Journal of the history of medicine and allied sciences*, *20*, 226–61.

Le Sueur, D., Binka, F., Lengeler, C., De Savigny, D., Snow, B., Teuscher, T., & Toure, Y. (1997). An atlas of malaria in Africa. *Africa health*, *19*(2), 23–4.

Sullivan, S. J., Schneiders, A. G., Cheang, C.-W., Kitto, E., Lee, H., Redhead, J., Ward, S., et al. (2012). "What"s happening?' A content analysis of concussion-related traffic on Twitter. *British journal of sports medicine*, *46*(4), 258–63.

Symplur. (2010). Healthcare Hashtag Project. Retrieved July 19, 2017, from https://www.symplur.com/healthcare-hashtags/

Tabachnick, W. J. (2010). Challenges in predicting climate and environmental effects on vector-borne disease episystems in a changing world. *The Journal of experimental biology*, *213*(6), 946–54.

Tableau Software. (2013). Tableau. Seattle, WA. Retrieved from http://www.tableausoftware.com/

Tanahashi, Y., Leaf, N., & Ma, K.-L. (2016). A Study On Designing Effective Introductory Materials for Information Visualization. *Computer Graphics Forum*, *35*(7), 117–126.

Thackeray, R., Burton, S. H., Giraud-Carrier, C., Rollins, S., & Draper, C. R. (2013). Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. *BMC cancer*, *13*(1), 508. BioMed Central Ltd.

Thomas, J., & Cook, K. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. (J. Thomas & K. A. Cook, Eds.). Los Alamitos, CA: IEEE Computer Society.

Thomsen, E. K., Deb, R. M., Dunkley, S., Coleman, M., Foster, G., Orlans, M., & Coleman, M. (2016). Enhancing Decision Support for Vector-Borne Disease Control Programs—The Disease Data Management System. (M. A. Johansson, Ed.)*PLoS Neglected Tropical Diseases*, *10*(2), e0004342. Public Library of Science.

TIBCO Spotfire Software. (2013). Spotfire. Somerville, MA. Retrieved from http://spotfire.tibco.com/

Tominski, C. (2015). *Interaction for Visualization*. *Synthesis Lectures on Visualization* (Vol. 3).

Torres, S. O. S. O., Eicher-Miller, H., Boushey, C., Ebert, D., & Maciejewski, R. (2012). Applied Visual Analytics for Exploring the National Health and Nutrition Examination Survey. *45th Hawaii International Conference on System Sciences* (pp. 1855–1863). IEEE.

Tory, M., & Möller, T. (2004). Human factors in visualization research. *IEEE*

*Transactions on Visualization and Computer Graphics*, *10*(1), 72–84. IEEE Educational Activities Department.

Tufte, E. R. (1997). *Visual explanations: images and quantities, evidence and narrative.* Cheshire, CT: Graphics Press.

Turner, A. M., Liddy, E. D., Bradley, J., & Wheatley, J. A. (2005). Modeling public health interventions for improved access to the gray literature. *Journal of the Medical Library Association : JMLA*, *93*(4), 487–94.

Turner, A. M., Stavri, Z., Revere, D., & Altamore, R. (2008). From the ground up: information needs of nurses in a rural public health department in Oregon. *Journal of the Medical Library Association : JMLA*, *96*(4), 335–42.

Tweepy. (2009). Tweepy. Retrieved July 18, 2017, from http://www.tweepy.org/

Twitter. (2007). Twitter Developer Documentation. Retrieved July 19, 2017, from https://dev.twitter.com/rest/public

Vernon, J. a, Trujillo, A., Rosenbaum, S., & Debuono, B. (2007). *Low Health Literacy: Implications for National Health Policy*. *World Education*.

Viceconti, M., Hunter, P., & Hose, R. (2015). Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1209–1215.

Vizuly. (2014). Halo. Retrieved from http://vizuly.io/product/halo/

Wagner, M., Fischer, F., Luh, R., Haberson, A., Rind, A., Keim, D. A., & Aigner, W. (2015). A Survey of Visualization Systems for Malware Analysis. *EuroVis*.

Wang, T. D., Wongsuphasawat, K., Plaisant, C., & Shneiderman, B. (2011). Extracting Insights from Electronic Health Records: Case Studies, a Visual Analytics Process Model, and Design Recommendations. *Journal of Medical Systems*, *35*(5), 1135–1152. Springer US.

Wang Baldonado, M. Q., Woodruff, A., & Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. *Proceedings of the working conference on Advanced visual interfaces - AVI '00* (pp. 110–119). New York, New York, USA: ACM Press.

Ward, M., Grinstein, G. G., & Keim, D. (2015). *Interactive data visualization : foundations, techniques, and applications* (2nd ed.).

Weaver, S. C. (2013). Urbanization and geographic expansion of zoonotic arboviral diseases: mechanisms and potential strategies for prevention. *Trends in Microbiology*.

Weeg, C., Schwartz, H. A., Hill, S., Merchant, R. M., Arango, C., & Ungar, L. (2015). Using Twitter to Measure Public Discussion of Diseases: A Case Study. *JMIR Public Health and Surveillance*, *1*(1), e6. JMIR Public Health and Surveillance.

Weisi, L., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E., & Wang, H. (Eds.). (2011). *Multimedia Analysis, Processing and Communications: Studies in Computational Intelligence*. Springer.

White, S. (2014). A review of big data in health care: challenges and opportunities. *Open Access Bioinformatics*, *Volume 6*, 13. Dove Press.

WHO. (2016). Statement on the 9th meeting of the IHR Emergency Committee regarding the Ebola outbreak in West Africa. Retrieved July 25, 2017, from https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html

van Wijk, J. J. (2006). Views on visualization. *IEEE transactions on visualization and computer graphics*, *12*(4), 421–432.

Wilkins, K., Nsubuga, P., Mendlein, J., Mercer, D., & Pappaioanou, M. (2008). The Data for Decision Making project: assessment of surveillance systems in developing countries to improve access to public health information. *Public Health*, *122*(9), 914–922.

World Health Organization. (2012). *Handbook for Integrated Vector Management*. Geneva: World Health Organization. Retrieved November 19, 2013, from http://www.who.int/heli/risks/vectors/vector/en/

World Health Organization. (2014). *A global brief on vector-borne diseases*. *WHO*. Geneva, Switzerland: World Health Organization.

World Health Organization. (2015). *World Malaria Report*. Geneva, Switzerland.

World Health Organization. (2016). *WHO Director-General summarizes the outcome of the Emergency Committee regarding clusters of microcephaly and Guillain-Barré syndrome*. *WHO*. Geneva, Switzerland: World Health Organization. Retrieved September 19, 2016, from http://www.who.int/mediacentre/news/statements/2016/emergency-committee-zika-microcephaly/en/

Wurman, R. S. (1989). *Information Anxiety* (1st ed.). Doubleday.

Zhang, J. (2001). External representations in complex information processing tasks. *Encyclopedia of library and information science*, *68*(31), 164–180. Citeseer.

Zhang, J., & Norman, D. (1994). Representations in Distributed Cognitive Tasks. *Cognitive Science*, *18*(1), 87–122.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., et al. (2012). Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 173–182). IEEE.

Zhao, J., Exeter, D., Moss, L., Hanham, G., Riddell, T., & Wells, S. (2013). Incorporating ringmaps into interactive web mapping for enhanced understanding of cardiovascular disease. *2013 21st International Conference on Geoinformatics* (pp. 1–6). IEEE.

Ziemkiewicz, C., & Kosara, R. (2007). The shaping of information by visual metaphors. *IEEE transactions on visualization and computer graphics*, *14*(6), 1269–1276.

# Appendices

## Appendix 1: List of search terms for visual analytic study

| | | |
|---|---|---|
| abortion | hemorrhagic stroke | pancreatic cancer |
| alcohol use disorders | hepatitis | pancreatitis |
| alzheimer | hiv/aids | paralytic ileus |
| aortic aneurysm | hurricane death | parkinsons disease |
| asthma | hypertensive heart disease | peptic ulcer |
| atrial fibrillation | influenza | peripheral arterial disease |
| atrial flutter | interpersonal violence | peripheral vascular disease |
| bile duct disease | intestinal ischemic syndrome | pharyngeal cancer |
| biliary tract cancer | intestinal obstruction | pneumoconiosis |
| bladder cancer | iron-deficiency anemia | pneumonia |
| brain cancer | ischemic heart | poisonings |
| breast cancer | ischemic stroke | pregnancy hypertensive |
| bronchitis | kidney cancer | preterm birth complications |
| cardiomyopathy | kidney disease | prostate cancer |
| cervical cancer | laryngeal cancer | protein-energy malnutrition |
| chagas | leukemia | pulmonary sarcoidosis |
| chikungunya | liver cancer | rheumatic heart |
| chronic obstructive pulmonary disease | liver cirrhosis | rheumatoid arthritis |
| colon cancer | low back pain | road injury |
| congenital anomalies | lung cancer | self-harm |
| dengue | malaria | sepsis |
| diabetes | male infertility | skin disease |
| diarrhea diseases | maternal hemorrhage | skin melanoma |
| diffuse parenchymal lung disease | measles | stds |
| drowning | medical treatment adverse effect | stomach cancer |
| drug overdose | meningitis | subcutaneous disease |
| earthquake death | migraine | syphilis |

| | | |
|---|---|---|
| ebola | mouth cancer | tetanus |
| encephalitis | multiple myeloma | tornado death |
| endocarditis | multiple sclerosis | trachea cancer |
| epilepsy | myocarditis | transport injury |
| esophageal cancer | nasopharyngeal cancer | tsunami death |
| falls | neck pain | tuberculosis |
| fire death | neonatal encephalopathy | typhoid |
| gall bladder | nervous system cancer | typhoon death |
| gallbladder cancer | non-hodgkin lymphoma | urinary disease |
| glomerulonephritis | oropharyngeal cancer | urinary organ cancer |
| gout | osteoarthritis | uterine cancer |
| heat death | ovarian cancer | whooping cough |

# Appendix 2: Ethics approval for visualization literacy study

Western Research

Research Ethics

**Western University Non-Medical Research Ethics Board**
**NMREB Delegated Initial Approval Notice**

**Principal Investigator**: Kamran Sedig
**Department & Institution**: Science\Computer Science,Western University

**NMREB File Number**: 108944
**Study Title**: Visual Literacy Research

**NMREB Initial Approval Date**: February 15, 2017
**NMREB Expiry Date**: February 15, 2018

**Documents Approved and/or Received for Information**:

| Document Name | Comments | Version Date |
|---|---|---|
| Western University Protocol | | 2017/02/08 |
| Letter of Information & Consent | | 2017/02/08 |
| Letter of Information & Consent | | 2017/02/08 |
| Other | Follow Up Form | 2017/02/08 |
| Other | Debriefing Letter | 2017/02/08 |
| Instruments | Demography Visualization Questions | 2017/01/18 |
| Instruments | Geography Visualization Questions | 2017/01/18 |
| Other | Introduction Script: script that give instructions of how participants will proceed through the exploration session | 2017/01/18 |
| Instruments | Experience Questionnaire B: questionnaire for the treatment group. | 2017/01/18 |
| Recruitment Items | Information Card | 2017/01/18 |
| Instruments | Experience Questionnaire A: questionnaire for control group. | 2017/01/18 |
| Recruitment Items | In Class Recruitment: script of in-class recruitment speech. | 2017/01/18 |
| Other | Interview Script | 2017/01/18 |
| Instruments | Demographics Form: form to gather demographic information from participants. | 2017/01/18 |
| Recruitment Items | Interview Invite Email: Email to be sent to participants to invite them to an interview. | 2017/01/18 |
| Recruitment Items | Recruitment Email: This is the email that will be sent to faculty at Western asking for permission to stop by their class to give an announcement about the study. | 2017/01/18 |
| Recruitment Items | Poster | 2017/01/18 |

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the above named study, as of the NMREB Initial Approval Date noted above.

NMREB approval for this study remains valid until the NMREB Expiry Date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario.

Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

[redacted]

**Western University, Research,** Support Services Bldg., Rm. 5150
London, ON, Canada N6G 1G9 t. 519.661.3036 f. 519.850.2466 www.uwo.ca/research/ethics

# Appendix 3: Ethics approval for health literacy study

Western Research

**Western University Non-Medical Research Ethics Board**
**NMREB Delegated Initial Approval Notice**

**Principal Investigator**: Kamran Sedig
**Department & Institution**: Science\Computer Science,Western University

**NMREB File Number**: 108994
**Study Title**: Global Health Literacy Research

**NMREB Initial Approval Date**: March 27, 2017
**NMREB Expiry Date**: March 27, 2018

**Documents Approved and/or Received for Information**:

| Document Name | Comments | Version Date |
|---|---|---|
| Letter of Information & Consent | Letter of Information and Consent II for Treatment Group | 2017/03/23 |
| Letter of Information & Consent | Letter of Information and Consent I for Treatment Group | 2017/03/23 |
| Advertisement | Recruitment Email | 2017/01/31 |
| Other | Interview Script | 2017/01/31 |
| Instruments | Experience Questionnaire | 2017/01/31 |
| Other | Follow Up Form to determine if participants wants to be invited for an interview | 2017/01/31 |
| Instruments | Demographics Form | 2017/01/31 |
| Instruments | Global Health Literacy Quiz | 2017/01/31 |
| Instruments | Task Sheet | 2017/01/31 |
| Western University Protocol | | 2017/03/16 |
| Recruitment Items | Poster | 2017/03/16 |
| Recruitment Items | Information Card | 2017/03/16 |
| Recruitment Items | In Class Recruitment | 2017/03/16 |
| Letter of Information & Consent | Control | 2017/03/16 |

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the above named study, as of the NMREB Initial Approval Date noted above.

NMREB approval for this study remains valid until the NMREB Expiry Date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario.

Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

[redacted]

**Western University, Research,** Support Services Bldg., Ste. 5150
London, ON, Canada N6G 1G9 t. 519.661.2161 f. 519.661.3907 www.westernu.ca/research

# Curriculum Vitae

**Name:** Oluwakemi Ola

**Post-secondary Education and Degrees:**

The University of Western Ontario
London, Ontario, Canada
2012-2017 Ph.D.

University of Houston
Houston, Texas, USA
2008-2010 M.Sc.

Trine University
Angola, Indiana, USA
2002-2005 B.Sc.

Southwestern Michigan College
Dowagiac, Michigan, USA
2000-2002 A.Sc.

**Honours and Awards:**

Association of PH Epidemiologists in Ontario Student Award
2015

UWORCS Research Presentation Winner: HCI Track
2014, 2015

Western's 3-Minute Thesis People's Choice Award
2014

Western Graduate Research Scholarship
2012-2015

University of Houston Computer Science Scholarship
2009-2010

**Related Work Experience**

<u>Limited Duties Lecturer, Computer Science Department</u>
The University of Western Ontario
Fundamentals of Computer Science (2016 – 17)
Data Analytics: Principles and Tools (2016 – 17)
Exploring the Landscape of Science (2017)

Senior Product Development Advisor
Public Health Ontario
2015 – 16

Teaching Assistant, Computer Science Department
The University of Western Ontario
2012 – 2015

**Publications:**

O. Ola and K. Sedig, "Beyond simple charts: Design of visualizations for big health data," Online J. Public Health Inform., vol. 8, no. 3, Dec. 2016.

O. Ola, O. Buchel, and K. Sedig, "Exploring the Spread of Zika: Using Interactive Visualizations to Control Vector-Borne Diseases," Int. J. Dis. Control Contain. Sustain., vol. 1, no. 1, pp. 47–68, 2016.

O. Ola, O. Buchel, and K. Sedig, "Interactive Visualizations as 'Decision Support Tools' in Developing Nations: The Case of Vector-Borne Diseases," in Transforming Public Health in Developing Nations, M. Sheikh, M. Aziza, and H. Mowafa, Eds. IGI Global, 2015.

O. Ola and K. Sedig, "The challenge of big data in public health: an opportunity for visual analytics.," Online J. Public Health Inform., vol. 5, no. 3, p. 223, Jan. 2014.

K. Sedig, P. Parsons, M. Dittmer, and O. Ola, "Beyond information access: Support for complex cognitive activities in public health informatics tools.," Online J. Public Health Inform., vol. 4, no. 3, Dec. 2012.

**Professional Talks:**

Invited Talks

The Challenge of Big Data in Public Health: The Case for Visual Analytics

Informatics Team Training, Public Health Ontario. Presented: February 2016

Interactive Visualizations for the Effective Communication of Public Health

PHO Rounds: Epidemiology, Public Health Ontario. Presented: September 2015

Visual Analytics Tools to Support Epidemiologists

London Middlesex Health Unit. Presented: January 2014

Conference Talks

Adapting lesson delivery to accommodate changing classroom sizes
2017 Western Conference on Science Education (WCSE)

Western's Integrated Science Program: A Response to Complex Global Problems
2017 WCSE. Co-Presented with WiSC Teaching Team

Reflections From First Time Blended Programming Instructors – Teaching Challenges and Lessons Learned
2015 WCSE. Co-Presented with Laura K. Reid

Visual Analytics for the Effective Communication of Public Health Information
2015 Association of Public Health Epidemiologists in Ontario Conference

---

[i] Here we refer to any individual seeking to use PH information in a professional capacity as a PH stakeholder.

[ii] Here we refer to data as digitally stored, sensed changes in the environment.

[iii] Unstructured data requires additional processing for it to be interpreted by the computer.