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

Oluwakemi Ola[1], Kamran Sedig[1]

1. INSIGHT Lab, Western University, Canada

Abstract

Public health (PH) data can generally be characterized as big data. The efficient and effective use of this data determines the extent to which PH stakeholders can sufficiently address societal health concerns as they engage in a variety of work activities. As stakeholders interact with data, they engage in various cognitive activities such as analytical reasoning, decision-making, interpreting, and problem solving. Performing these activities with big data is a challenge for the unaided mind as stakeholders encounter obstacles relating to the data's volume, variety, velocity, and veracity. Such being the case, computer-based information tools are needed to support PH stakeholders. Unfortunately, while existing computational tools are beneficial in addressing certain work activities, they fall short in supporting cognitive activities that involve working with large, heterogeneous, and complex bodies of data. This paper presents visual analytics (VA) tools, a nascent category of computational tools that integrate data analytics with interactive visualizations, to facilitate the performance of cognitive activities involving big data. Historically, PH has lagged behind other sectors in embracing new computational technology. In this paper, we discuss the role that VA tools can play in addressing the challenges presented by big data. In doing so, we demonstrate the potential benefit of incorporating VA tools into PH practice, in addition to highlighting the need for further systematic and focused research.

Keywords: big data, analytics, public health informatics, visual analytics, analytical reasoning, interactive visualizations, human-information interaction

Introduction

Data and information are both currency and product within the field of public health (PH) [1]. 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 [2], may be unreliable, encoded in a variety of formats [3], [4], and can be volatile (i.e., changing, and available only for a limited amount of time) [5], 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.

1

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 [6]. 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 [6] 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 [7]– [8]. The efficient and effective use of data determines the extent to which PH stakeholders can sufficiently address the health concerns of the community [1], [9]. 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. Discussion with area doctors reveals that local hospitals have confirmed 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 [10]. Furthermore, as stakeholders interact with data they encounter a number of obstacles relating to its volume, variety, velocity, and veracity [2]– [5], [11]– [12]. 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 [13], and interactive visualization tools such as Malaria Atlas Project [14]. 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 [15].

Public health tends to lag behind other sectors in the adoption of new technology (for examples, see England et al.'s [16] examination of PH's slow rate of information technology adoption, and Shortliffe's [17] comparison of healthcare with other sectors). The recent emergence of a

2

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 [18], [19]. This is beneficial for data-intensive fields [19] such as PH, finance, insurance, sales, and climatology, to name a few. While many fields, such as finance and sales [20], 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 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.

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.

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. [5], 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 [21]. Irrespective of their area of expertise and sub-field of application, stakeholders must interact with data to perform a myriad of work activities.

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 [11], [22], [23], [12], great variety [2], [24], [25], high velocity [5], and low veracity [9], [23], [26]. These four features of PH data are

3

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 [6], 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 [27]. 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 [22], [28]. The seamless incorporation of user-guided analysis techniques into computational tools is crucial in facilitating the systematic use of data.

PH Activities

PH stakeholders engage in a variety of work activities in an effort to improve and ensure the health of the community [5], [24], [21]. 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 [29]. 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 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 [30]. 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 [6]. 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 [31]. As knowledge workers, PH stakeholders engage in a myriad of cognitive activities including analytical reasoning, decision-making, sensemaking, and problem solving.

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 [27]. 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 [18], [32]. Although analytical reasoning is a structured and disciplined process, the aforementioned tasks typically occur in an

iterative and non-linear fashion [27]. 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 [33], [34]. Analytical reasoning can be viewed as a transformative process in which new information, knowledge, and insight are derived from given data [27], [35]. In some situations, this new information, knowledge, or insight serves as the basis for decision-making and problem solving [33]. 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 [26], [36], 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 [7], [37], [38]. 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 [8], [26]. The success of this approach is contingent on PH stakeholders being able to effectively interact with and use data [22].

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) [6]. 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 [39], Glenberg and Langston [40], and Card et al. [41]. 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 [41]. 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 [42], to the recent use of atlases for mapping the risk of malaria in Africa [43],

they put the brunt of the information-processing load (i.e., analytical reasoning and decision making) on the cognitive resources of their users [6], [27], 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 [27], reducing, and possibly bridging, the gap between the internal (mental) representations of the user and the external (visual) representations of the tool [27], [44], [45]. Interactive visual representations can offer users flexibility, support convergent and divergent thinking, and accommodate the user's perceptual and cognitive needs [18], [27]. Furthermore, interactive representations allow stakeholders to control which subset of data is visually displayed while still having access to data latent in the system [27] [46], which 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 [27], [46], 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 [39], [47], and even constrain and limit stakeholders as they engage in a particular task [45], [47], [48]. Therefore, PH stakeholders stand to benefit from tools that allow users to manipulate visual representations, a capability made possible through interaction.

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 [27], [45]. 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 [27]. 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 [27]. 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 [44]. 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 stakeholder's analytical reasoning tasks.

Visual Analytic Tools

VA is sometimes defined as the *"science of analytical reasoning facilitated by interactive visual interfaces"* [18]. 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 [18], [19]. 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 [49]. Current data analytics tools that assist PH stakeholders in analyzing data include Stata [13] and EpiInfo [50]. 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 [15]. 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 [41], while others portray physical data such as the human body and molecules [51]. Current PH interactive visualization tools include Malaria Atlas Project [14] and Spatio-Temporal Epidemiological Modeller [52]. While beneficial, these types of tools prove inadequate when faced with problems requiring advanced computational analysis and big data [15].

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 [15]. 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 [44], which are described below.

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 [49]. In the data transformation stage, the pre-processed data is

7

converted into a form that is more conducive to data analysis. This stage includes tasks such as data normalization and aggregation [49].
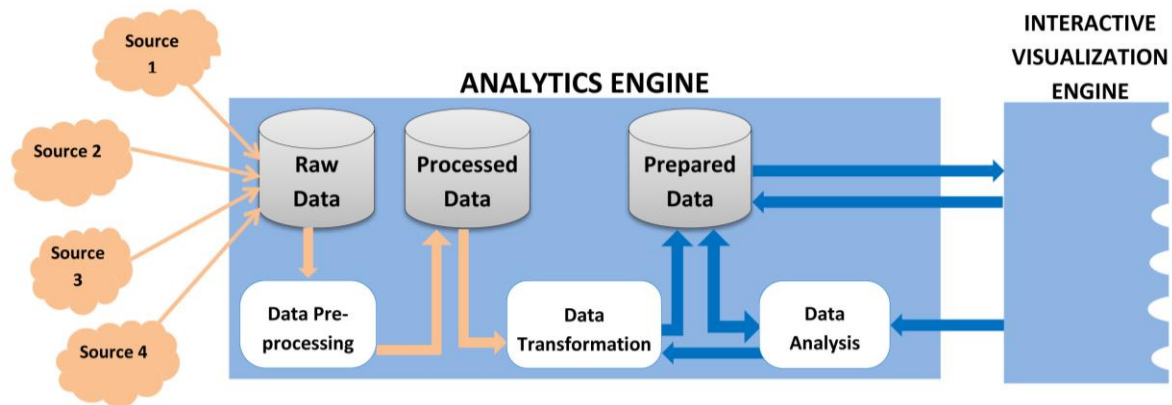

Figure 1: The analytics engine component of VA tools

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) [53]–[54]. 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.

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 [18], [19]. 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.
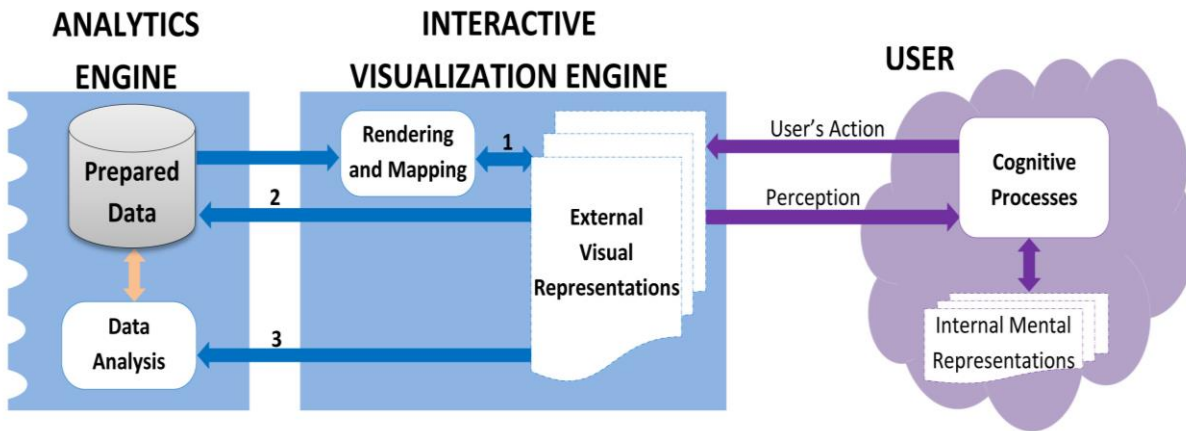
Figure 2: Interactive visualization engine component in VA tools

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 [27]. 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 [55], [56]. 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) [27]. 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.
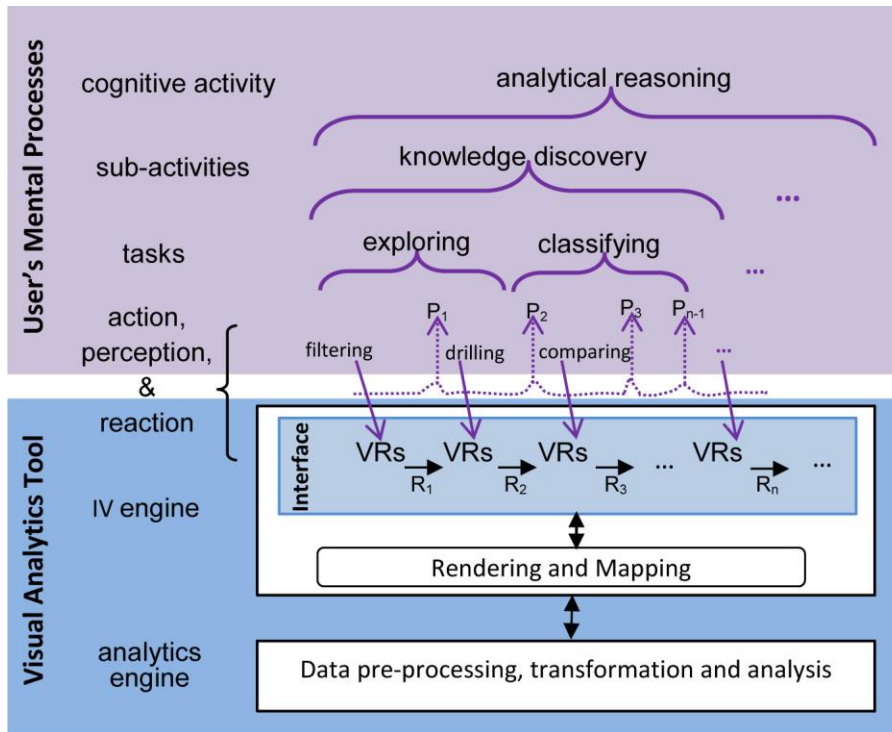
9

Figure 3: The hierarchical structure of analytical reasoning emerging from lower level processes, adapted from [27]. 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).

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 [27], [47], [55]. 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 [27], [57]. 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 [48]. 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 [58]). As users perform analytical reasoning tasks, they seek to harmonize and coordinate their internal representations and the tool's external representations [48], [59]. When processing data in such a dynamic manner, a cognitive coupling is formed between the user and the tool [27], [60]. The strength of the coupling between the user's internal representations and

10

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 [27]. 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 [27]). 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 [44], [57]). 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 [56]). These four 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.

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 visualizations. 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 [5]. 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.

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.

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 [11], [22], [23], [12]. Because of this deficiency, stakeholders find themselves

11

spending more time wading through data, and less time actually addressing the health concerns of their community. As discussed in [2], *"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.

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 [3] 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) [4], [14]. 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 [5]. 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.

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 [9], [23], [26]. 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

12

can facilitate understanding nurses' practices during night shift, or traffic data of the intersection in front of a hospital.

Current application of VA tools in PH

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.

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 [38]. 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 [61]. Furthermore, nSpace provides an environment that supports the generation of hypotheses and evaluation of relevant evidence [61]. Epidemic intelligence involves the early identification, assessment, and verification of potential public health hazards [62] 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 [63], and to gauge adverse public reaction to certain drugs [64]. 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.

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 [65] and SpotFire [66] 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 [67]. 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 [68] and exists in situations that have complex dynamics and interdependencies [69]. 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 [70]. VA tools modeling scientific uncertainty in policy simulations can provide policy makers with more information on possible outcomes.

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 [38]. 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 [71] 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 [71]. A similar tool is Epinome, which, in addition to allowing epidemiologists to explore outbreaks, tracks the user's interactions for post analysis [72].

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 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

14

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 25th, 50th, and 75th 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.
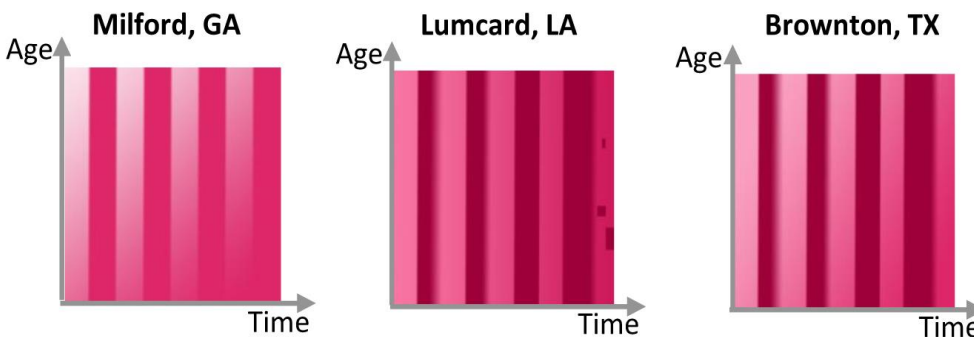

Figure 4: Image plots of WNV cases for the 3 selected cities from 2008 - 2013

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.
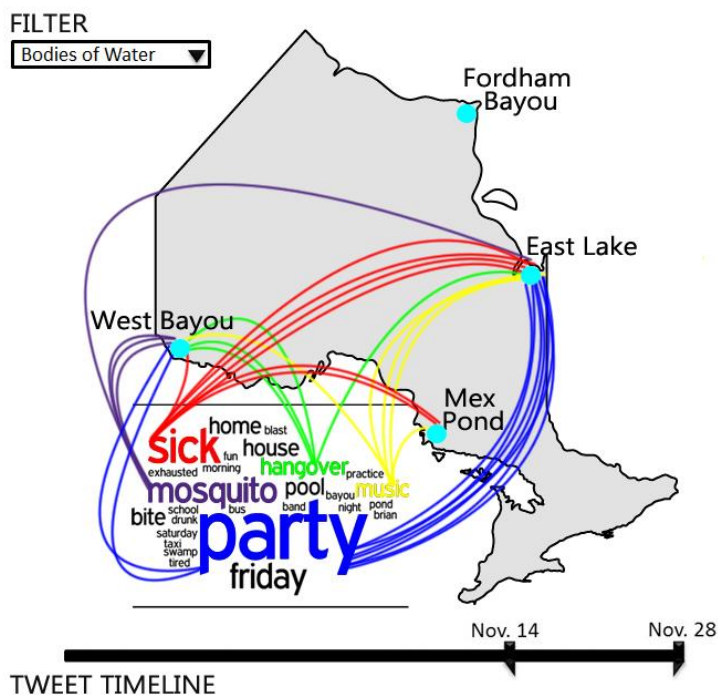

Figure 5: Visual representation depicting spatial relationships between most frequent words in tweets and local bodies of water in Lumcard.

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,

16

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.

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.

References

1. O'Carroll PW. "Introduction to Public Health Informatics," in in *Public Health Informatics and Information Systems*, vol. 6, no. 3, P. W. O'Carroll, W. A. Yasnoff, M. E. Ward, L. H. Ripp, and E. L. Martin, Eds. New York, NY, USA: Springer-Verlag, 2003, pp. 3–15.

2. Revere D, Turner AM, Madhavan A, Rambo N, Bugni PF, et al. 2007. Understanding the information needs of public health practitioners: A literature review to inform design of an interactive digital knowledge management system. *J Biomed Inform*. 40(4), 410-21.

3. Rambo N. 2000. Information Needs and Uses of the Public Health Workforce -- Washington, 1997-1998. *MMWR Morb Mortal Wkly Rep*. 49(6), 118-20. PubMed

4. Turner AM, Liddy ED, Bradley J, Wheatley JA. 2005. Modeling public health interventions for improved access to the gray literature. *J Med Libr Assoc*. 93(4), 487-94. PubMed

5. O'Carroll PW, Cahn MA, Auston I, Selden CR. 1998. Information needs in public health and health policy: results of recent studies. *J Urban Health*. 75(4), 785-93. PubMed http://dx.doi.org/10.1007/BF02344508

6. Sedig K, Parsons P, Dittmer M, Ola O. 2012. Beyond information access: Support for complex cognitive activities in public health informatics tools. *Online J Public Health Inform*. 4(3), e2. PubMed http://dx.doi.org/10.5210/ojphi.v4i3.4270

7. Brownson RC, Gurney J, Land G. 1999. Evidence-Based Decision Making in Public Health. *J Public Health Manag Pract*. 5(5), 86-97. PubMed http://dx.doi.org/10.1097/00124784-199909000-00012

8. Brownson RC, Fielding JE, Maylahn CA. 2009. Evidence-Based Public Health: A Fundamental Concept for Public Health Practice. *Annu Rev Public Health*. 30, 175-201. PubMed http://dx.doi.org/10.1146/annurev.publhealth.031308.100134

9. Reeder B, Revere D, Hills RA, Baseman JG, Lober WB. 2012. Public Health Practice within a Health Information Exchange: Information Needs and Barriers to Disease Surveillance. *Online J Public Health Inform*. 4(3), e4. PubMed http://dx.doi.org/10.5210/ojphi.v4i3.4277

10. Goddard M, Mowat D, Corbett C, Neudorf C, Raina P, et al. 2004. The Impacts of Knowledge Management and Information Technology Advances on Public Health Decision-Making in 2010. *Health Informatics J*. 10(2), 111-20. http://dx.doi.org/10.1177/1460458204042233

11. Higgins JW, Strange K, Scarr J, Pennock M, Barr V, et al. 2011. 'It's a Feel. That's What a Lot of Our Evidence Would Consist of ': Public Health Practitioners' Perspectives on Evidence. *Eval Health Prof*. 34(3), 278-96. PubMed http://dx.doi.org/10.1177/0163278710393954

12. Turner AM, Stavri Z, Revere D, Altamore R. 2008. From the ground up: information needs of nurses in a rural public health department in Oregon. *J Med Libr Assoc*. 96(4), 335-42. PubMed http://dx.doi.org/10.3163/1536-5050.96.4.008

13. StataCorp. "Stata Staistical Software: Release 12." StataCorp LP, College Station, TX, 2009.

14. Guerra CA, Hay SI, Lucioparedes LS, Gikandi PW, Tatem AJ, et al. 2007. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malar J*. 6(17). PubMed

15. Keim DA, Mansmann F, Thomas J. 2009. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explor. Newsl.* 11(2), 5-8. http://dx.doi.org/10.1145/1809400.1809403

16. England I, Stewart D, Walker S. 2000. Information technology adoption in health care: when organisations and technology collide. *Aust Health Rev*. 23(3), 176-85. PubMed http://dx.doi.org/10.1071/AH000176

17. Shortliffe EH. 2005. Strategic action in health information technology: why the obvious has taken so long. *Health Aff (Millwood)*. 24(5), 1222-33. PubMed http://dx.doi.org/10.1377/hlthaff.24.5.1222

18. Thomas J, Cook K. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos, CA, United States: IEEE Computer Society, 2005.

19. Keim DA, Kohlhammer J, Ellis G, Mansmann F. *Mastering The Information Age-Solving Problems with Visual Analytics*. Eurographics Association, 2010.

18

20. K. Schlegel, R. L. Sallam, Y. Daniel, and J. Tapadinhas, "Magic quadrant for business intelligence platforms.," 2013.

21. Committee on Educating Public Health Professionals for the 21st Century, *Who Will Keep the Public Healthy? Educating Public Health Professionals for the 21st Century*. Washington, D.C.: The National Academies Press, 2003, pp. 4 – 11.

22. K. Keough, "The Third Amyot Lecture. How science informs the decisions of government.," *Can. J. Public Heal. Rev. Can. santé publique*, 93(2), 104–8, Jan. 2002.

23. LaPelle NR, Luckmann R, Simpson EH, Martin ER. 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. PubMed http://dx.doi.org/10.1186/1471-2458-6-89

24. Rambo N. 1998. Information resources for public health practice. *J Urban Health*. 75(4), 807-25. PubMed http://dx.doi.org/10.1007/BF02344510

25. Cohen BB, Franklin S, West JK. 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. *J Public Health Manag Pract*. 12(2), 155-60. PubMed http://dx.doi.org/10.1097/00124784-200603000-00007

26. L. Kiefer, J. Frank, E. Di Ruggiero, M. Dobbins, D. Manuel, P. R. Gully, and D. Mowat, "Fostering evidence-based decision-making in Canada: examining the need for a Canadian population and public health evidence centre and research network.," *Can. J. Public Heal. Rev. Can. santé publique*, 96(3),11–40, Jan. 2005.

27. Sedig K, Parsons P. 2013. "Interaction design for complex cognitive activities with visual representations: A pattern-based approach," *AIS Trans. Hum Comput Interact*. 5(2), 84-133.

28. Berner ES, Moss J. 2005. Informatics challenges for the impending patient information explosion. *J Am Med Inform Assoc*. 12(6), 614-17. PubMed http://dx.doi.org/10.1197/jamia.M1873

29. Rambo N, Zenan JS, Alpi KM, Burroughs CM, Cahn MA, et al. 2001. Public Health Outreach Forum: lessons learned. *Bull Med Libr Assoc*. 89(4), 403-06. PubMed

30. National Research Council. *The Future of Public Health*. Washington, D.C.: The National Academies Press, 1988.

31. Reinhardt W, Schmidt B, Sloep P, Drachsler H. 2011. Knowledge Worker Roles and Actions-Results of Two Empirical Studies. *Knowl Process Manage*. 18(3), 150-74. http://dx.doi.org/10.1002/kpm.378

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

33. Leighton J. "Defining and describing reason," in in *The Nature of Reasoning*, J. Leighton and R. Sternberg, Eds. New York: Cambridge University Press, 2004, pp. 3 – 11.

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

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

36. D. L. Mowat and J. Hockin, "Building capacity in evidence-based public health practice (Editorial)," *Can J Public Heal.*, 93(1). Jan. 2002.

37. Baltussen R, Niessen L. 2006. Priority setting of health interventions: the need for multi-criteria decision analysis. *Cost Eff Resour Alloc*. 4(1). PubMed http://dx.doi.org/10.1186/1478-7547-4-14

38. National Research Council. *The Future of Public Health*. Washington, D.C.: The National Academies Press, 1988.

39. Larkin J, Simon H. 1987. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cogn Sci*. 11(1), 65-100. http://dx.doi.org/10.1111/j.1551-6708.1987.tb00863.x

40. Glenberg AM, Langston WE. 1992. Comprehension of illustrated text: Pictures help to build mental models. *J Mem Lang*. 31(2), 129-51. http://dx.doi.org/10.1016/0749-596X(92)90008-L

41. Card SK, Mackinlay JD, Shneiderman B. *Readings in Information Visualization: Using Vision to Think*. San Francisco, Calif: Morgan Kaufmann Publishers, 1999.

42. Snow J. *On the mode of communication of cholera*, 2nd ed. London: Churchill, 1855.

43. Le Sueur D, Binka F, Lengeler C, De Savigny D, Snow B, et al. 1997. An atlas of malaria in Africa. *Afr Health*. 19(2), 23-24. PubMed

44. Sedig K, Parsons PC, Babanski A. 2012. Towards a characterization of interactivity in visual analytics. *J. Multimed. Process.* 3(1), 12-28.

45. Parsons P, Sedig K. 2013. Adjustable Properties of Visual Representations: Improving the Quality of Human-Information Interaction. *J Am Soc Inf Sci Technol*. (In press).

46. Sedig K. "Interactive mathematical visualizations: Frameworks, tools, and studies," *Trends Interact. Vis.*, pp. 343–363, 2009.

47. Zhang J, Norman D. 1994. Representations in Distributed Cognitive Tasks. *Cogn Sci*. 18(1), 87-122. http://dx.doi.org/10.1207/s15516709cog1801_3

48. Zhang J. 2001. External representations in complex information processing tasks. *Encycl. Libr. Inf. Sci.* 68(31), 164-80.

49. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)*, 3rd ed. Morgan Kaufmann, 2011.

50. Centers for Disease Control and Prevention. "Epi Info. " Centers for Disease Control and Prevention, Atlanta, 2012.

51. Mackinlay JD. 2000. Opportunities for information visualization. *IEEE Comput Graph Appl*. 20(1), 22-23. http://dx.doi.org/10.1109/38.814540

52. Ford DA, Kaufman JH, Eiron I. 2006. An extensible spatial and temporal epidemiological modelling system. *Int J Health Geogr*. 5(4), 1-6. PubMed

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

54. Soille P. *Morphological Image Analysis: Principles and Applications*, 2nd ed. Heidelberg: Springer-Verlag, 2003.

55. Hollan J, Hutchins E, Kirsh D. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Interact.* 7(2), 174-96. http://dx.doi.org/10.1145/353485.353487

56. Parsons P, Sedig K. "Distribution of Information Processing while Performing Complex Cognitive Activities with Visualization Tools (In Print)," in in *Handbook of Human Centric Visualization*, T. Huang, Ed. Springer, 2013, pp. 639–715.

57. Sedig K, Parsons PC, Dittmer M, Haworth R. "Human-centered interactivity of visualization tools: Micro- and macro-level considerations (In Print)," in in *Handbook of Human Centric Visualization*, T. Huang, Ed. Springer, 2013, pp. 717–743.

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

59. Liu Z, Stasko J. 2010. Mental models, visual reasoning and interaction in information visualization: a top-down perspective. *IEEE Trans Vis Comput Graph*. 16(6), 999-1008. PubMed http://dx.doi.org/10.1109/TVCG.2010.177

60. Brey P. 2005. The Epistemology and Ontology of Human-Computer Interaction. *Minds Mach*. 15(3–4), 383-98. http://dx.doi.org/10.1007/s11023-005-9003-1

61. Proulx P, Tandon S, Bodnar A, Schroh D, Harper R, et al. "Avian Flu Case Study with nSpace and GeoTime," in *2006 IEEE Symposium On Visual Analytics And Technology*, 2006, pp. 27–34.

62. Paquet C, Coulombier D, Kaiser R, Ciotti M. 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro Surveill*. 11(12), 212-14. PubMed

63. Google. "Google Flu Trends. Available at: http://www.google.org/flutrends.".

64. Bian J, Topaloglu U, Yu F. "Towards large-scale twitter mining for drug-related adverse events," in *Proceedings of the 2012 international workshop on Smart health and wellbeing - SHB '12*, 2012, pp. 25 – 32.

65. Tableau Software. "Tableau." Seattle, WA, 2013.

66. TIBCO Spotfire Software. "Spotfire." Somerville, MA, 2013.

67. Sims JN, Isokpehi RD, Cooper GA, Bass MP, Brown SD, et al. 2011. Visual analytics of surveillance data on foodborne vibriosis, United States, 1973-2010. *Environ. Health Insights*. 5, 71-85. PubMed

68. Schabas R. 2002. Is public health ethical? (Editorial). *Can J Public Health*. 93(2), 98-99. PubMed

69. Kramer RA, Dickinson KL, Anderson RM, Fowler VG, Miranda ML, et al. 2009. Using decision analysis to improve malaria control policy making. *Health Policy*. 92(2–3), 133-40. PubMed http://dx.doi.org/10.1016/j.healthpol.2009.02.011

70. D'Alessandro U, Buttiëns H. 2001. History and importance of antimalarial drug resistance. *Trop Med Int Health*. 6(11), 845-48. PubMed http://dx.doi.org/10.1046/j.1365-3156.2001.00819.x

71. Maciejewski R, Livengood P, Rudolph S, Collins TF, Ebert DS, et al. 2011. A pandemic influenza modeling and visualization tool. *J Vis Lang Comput*. 22(4), 268-78. http://dx.doi.org/10.1016/j.jvlc.2011.04.002

72. Livnat Y, Rhyne T-M, Samore M. 2012. Epinome: A Visual-Analytics Workbench for Epidemiology Data. *IEEE Comput Graph Appl*. 32(2), 89-95. http://dx.doi.org/10.1109/MCG.2012.31

73. Bates MJ. 2005. Information and knowledge : an evolutionary framework for information science. *Inf. Res.* 10(4), 1-24.

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

[2] For an in-depth discussion on the differences between data and information see [44] and [73].

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