Beyond simple charts: Design of visualizations for big health data

Oluwakemi Ola[1], Kamran Sedig[1]

1. Insight Lab, Western University, Canada

Abstract

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. Interactive visualizations have the potential to amplify big data's utilization. Visualizations can be used to support a variety of tasks, such as tracking the geographic distribution of diseases, analyzing the prevalence of disease, triaging medical records, predicting outbreaks, and discovering at-risk populations. Currently, many health visualization tools use simple charts, such as bar charts and scatter plots, that only represent few facets of data. These tools, while beneficial for simple perceptual and cognitive tasks, are ineffective when dealing with more complex sensemaking tasks that involve exploration of various facets and elements of big data simultaneously. There is need for sophisticated and elaborate visualizations that encode many facets of data and support human-data interaction with big data and more complex tasks. When not approached systematically, design of such visualizations is labor-intensive, and the resulting designs may not facilitate big-data-driven tasks. Conceptual frameworks that guide the design of visualizations for big data can make the design process more manageable and result in more effective visualizations. In this paper, we demonstrate how a framework-based approach can help designers create novel, elaborate, non-trivial visualizations for big health data. We present four visualizations that are components of a larger tool for making sense of large-scale public health data.

Keywords: big health data, visualization design, health-related tasks, multifaceted data, human-data interaction, sensemaking, public health, healthcare

Correspondence: oola@uwo.ca, sedig@uwo.ca

DOI: 10.5210/ojphi.v8i3.7100

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 [1]. 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 [1–4].

Interactive visualizations[1] 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 [5]. 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[2] rely on Microsoft Office and off-the-shelf business intelligence tools to perform their data-driven tasks [6]. Majority of these tools use simple visualizations, such as scatter plots, heat maps, bar charts, choropleth maps, and radar charts [6, 7]. These visualizations typically only represent one or two facets of the data (e.g., attributes, relationships) [8–11]. When working with big data, there is the challenge of needing to analyze non-explicit and unknown relationships among the data elements as well [12, 13]. 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 [7, 13].

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 [7, 13, 14]. 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 [15]. 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 [16]. Researchers have tried to organize the plethora of existing visualization techniques to give structure to the selection and design process [17, 18]. 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 [6, 19, 20]. 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 [15, 16, 21]. 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 [21].

To this end, Sedig and Parsons [22] 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 public health data that we have implemented. Finally, Section 5 concludes the paper.

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.

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 [23, 24]. 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) [23, 25, 26]. 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 [23]. 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 [27–29]. Data that exhibits one or more of these qualities of big data presents processing challenges for health-related human-data interaction tasks.

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 [6, 24]. 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 [30]. 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 [31, 32]. 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 [22].

Visualizations for big data tasks

Visualizations[3] 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 [22]. 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 [5, 16]. However, the extent to which a visualization supports a task depends on how the data, and how much of it, is encoded [33, 34]. 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 [6]. 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 [35]. 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-term memory can become overloaded [36]. As data increases, this approach has been shown to result in an inaccurate understanding of trends [37]. Another approach represents facets in multiple visualizations distributed through space. Data dashboards, typical in business intelligence tools, employ this approach [38]. Though beneficial, as visualizations crowd the dashboard, users are forced to mentally combine representations to perform tasks [39]. 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 [15, 16, 21].

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 [40–43]. 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 [7]. Non-trivial visualizations that effectively encode data items can play a prominent role in how people use big data and interact with it [13, 14, 42]. In the context of health-related visualizations, in a systematic and comprehensive review, Carroll et al. [6] 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 the health domain, where it has been noted that visualization design is not as advanced as in other domains [4]. 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.

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 [22, 42]. Part of the challenge of developing tools for big data is determining how to structure or organize data items within visualizations [44]. 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.

Description of Patterns

Sedig and Parsons define a pattern as a regularity in some dimension [22]. 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[4]. 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.
- **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.

Pattern Blending and Syntax

Sedig and Parsons note that "Designers can blend different patterns to devise representational structures that have different organizational affordances" [22]. 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.

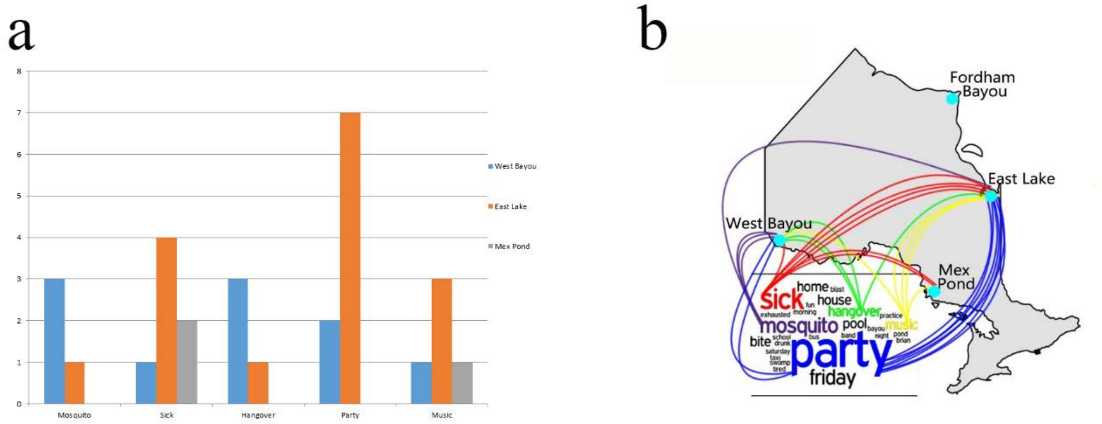

Figure 1: (a) Grouped bar chart (b) Alternative visualization for making sense of tweets.

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.

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[5] 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) [45]. 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 [42]. 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[6] across 187 countries.

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 69. 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 [46, 47]. 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.

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 (or location) 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.

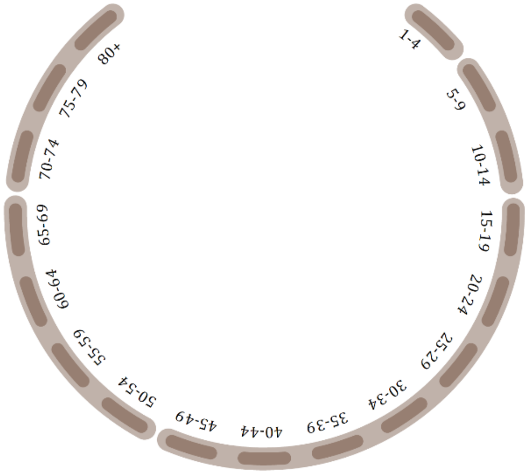

Figure 2: Demography sub-visualization for age groups.

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 facilitate 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 facilitates 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[7] 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.

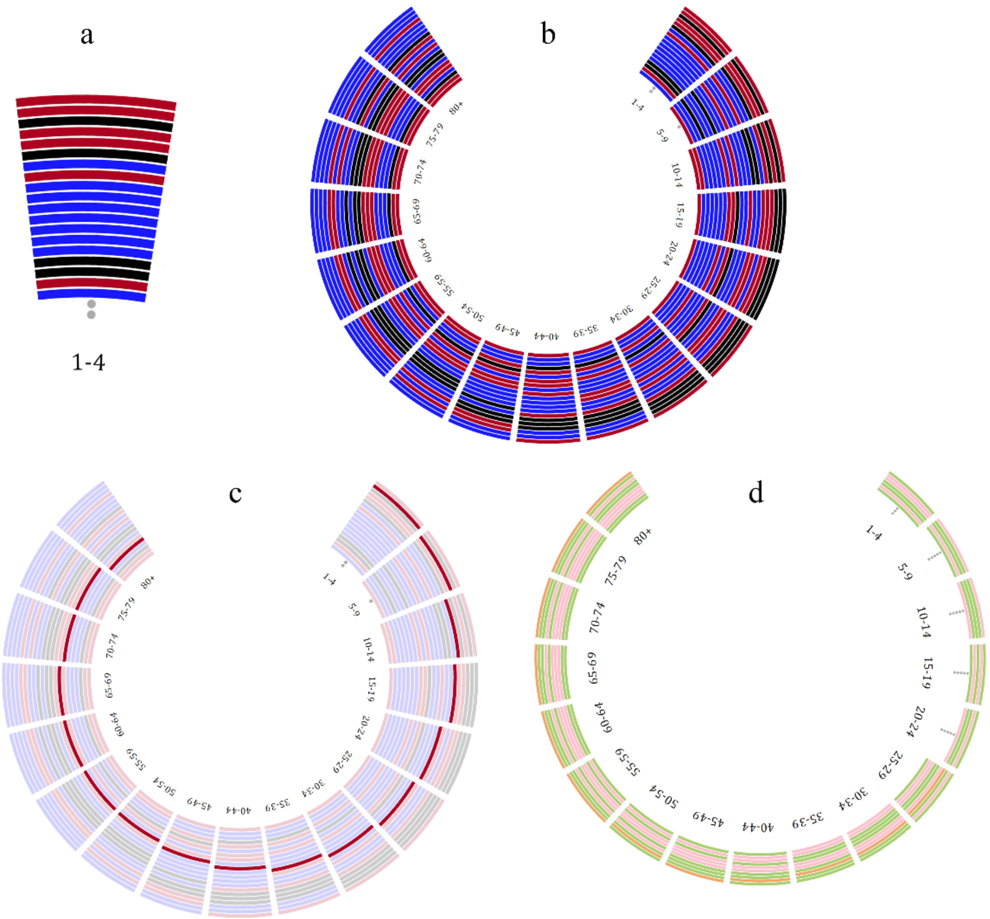

Figure 3: (a) visualization of cause clusters for children 1-4 years old (b) cause clusters ranking sub-visualization for all age groups (c) cause clusters sub-visualization with the neglected tropical diseases and malaria cluster emphasized (d) risk clusters sub-visualizations for all age groups.

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.

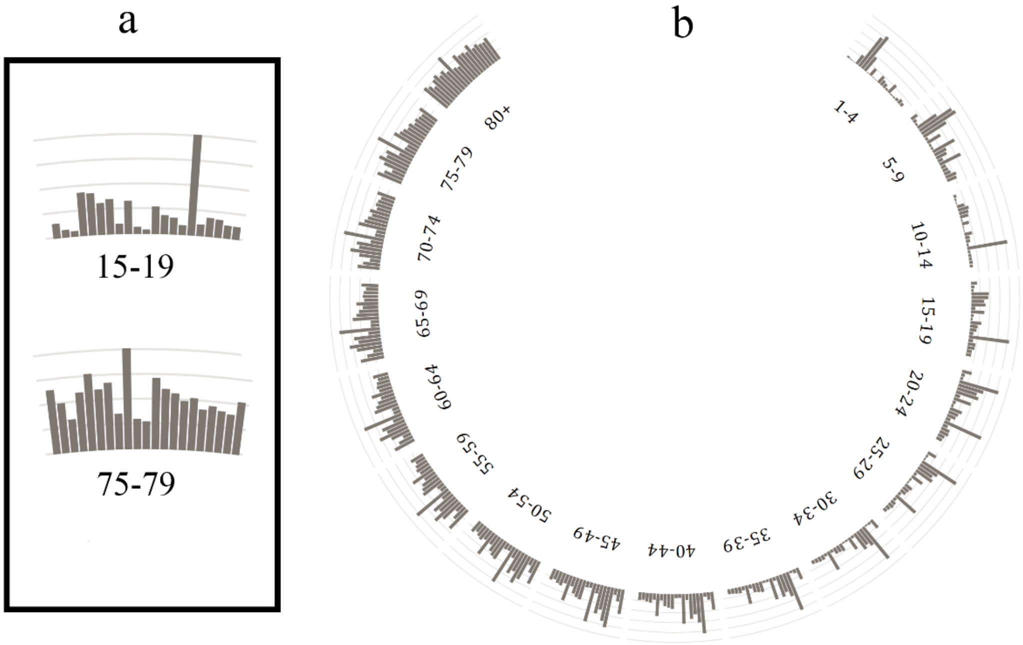

Figure 4: (a) [Token•List•Coordinate]-based bar charts for age groups 15-19 and 75-79 (b) demography sub-visualization for locations.

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.

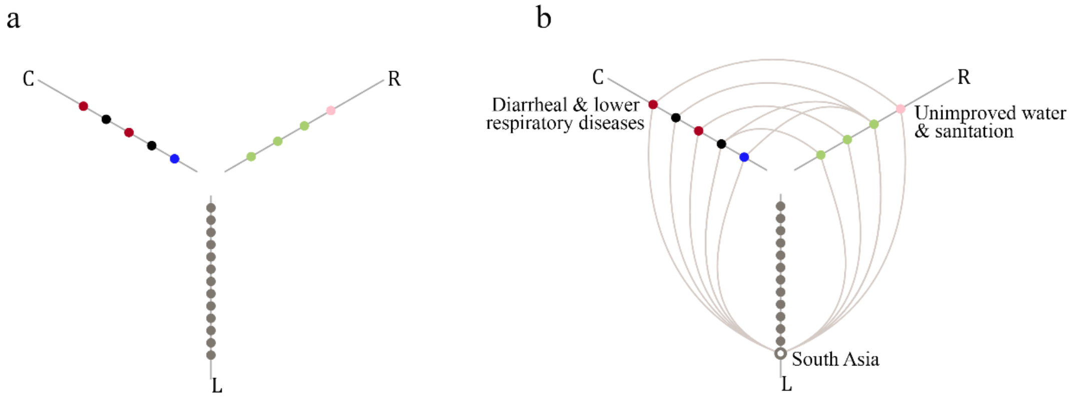

Figure 5: (a) Coordinate axes for cause, risk, and location clusters (b) Demography sub-visualization for relationships between cause, risk, and location clusters for individuals between the age of 5 and 14.

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 850 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.

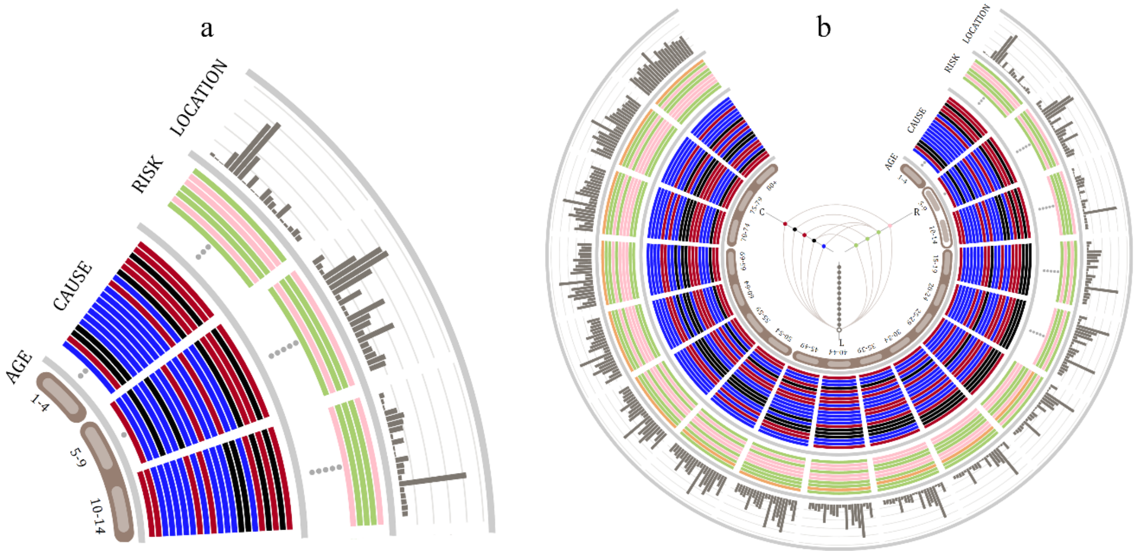

Figure 6: (a) Enlarged partial view of the first four sub-visualizations for demography (b) Overall visualization for demography based on [Stack•Track•Token•Group•Link•List•Coordinate].

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.

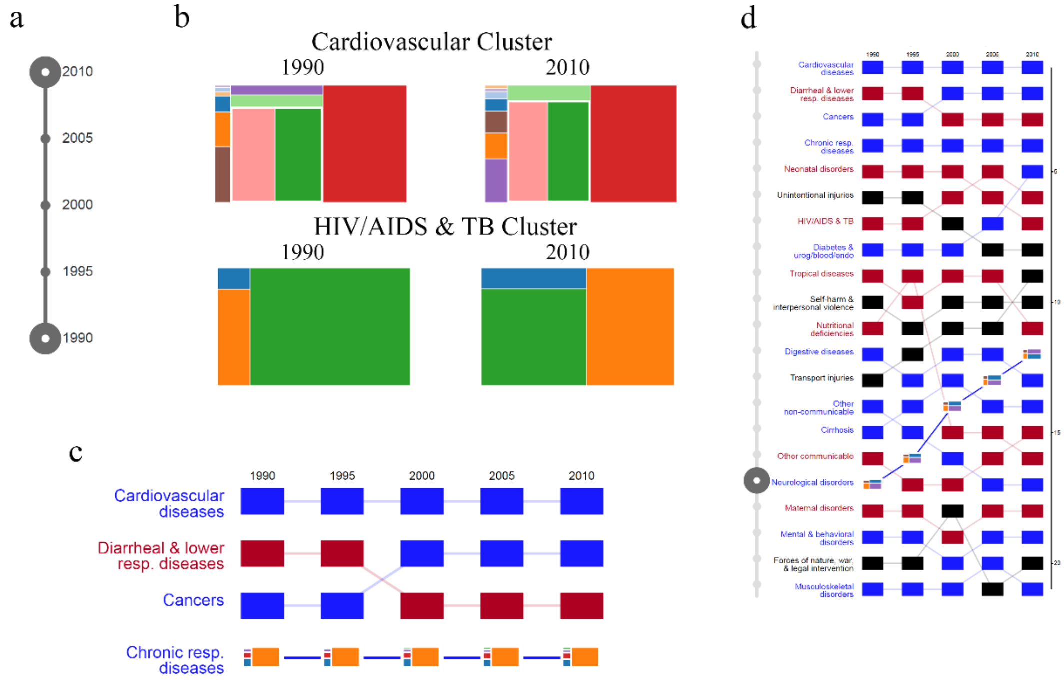

Figure 7: (a) [Token•Coordinate]-based representation for years (b) hierarchical visualization for cardiovascular diseases and HIV/AIDS & tuberculosis clusters (c) top portion of cluster-specific mortality ranking (d) Chronology sub-visualization for cause cluster-specific mortality.

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.

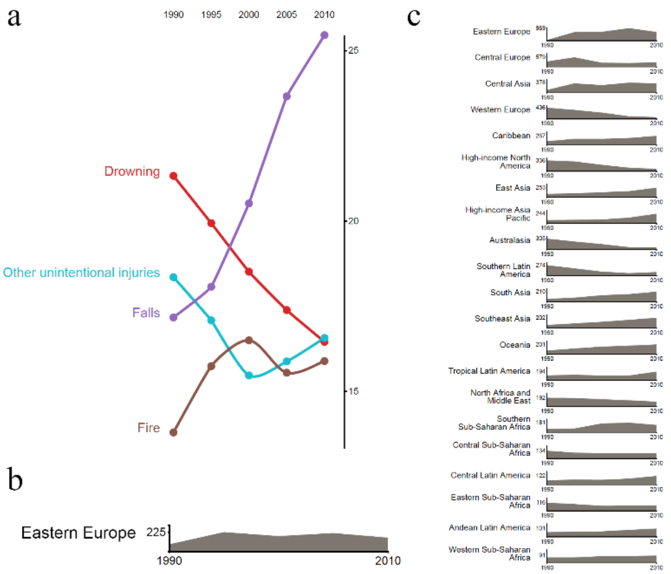

Figure 8: (a) Portion of chronology sub-visualization for cause proportion (b) Area chart for Eastern Europe for the cancer cluster (c) Region cluster-specific mortality for cardiovascular disease cluster.

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 Caribbean and the clusters 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.

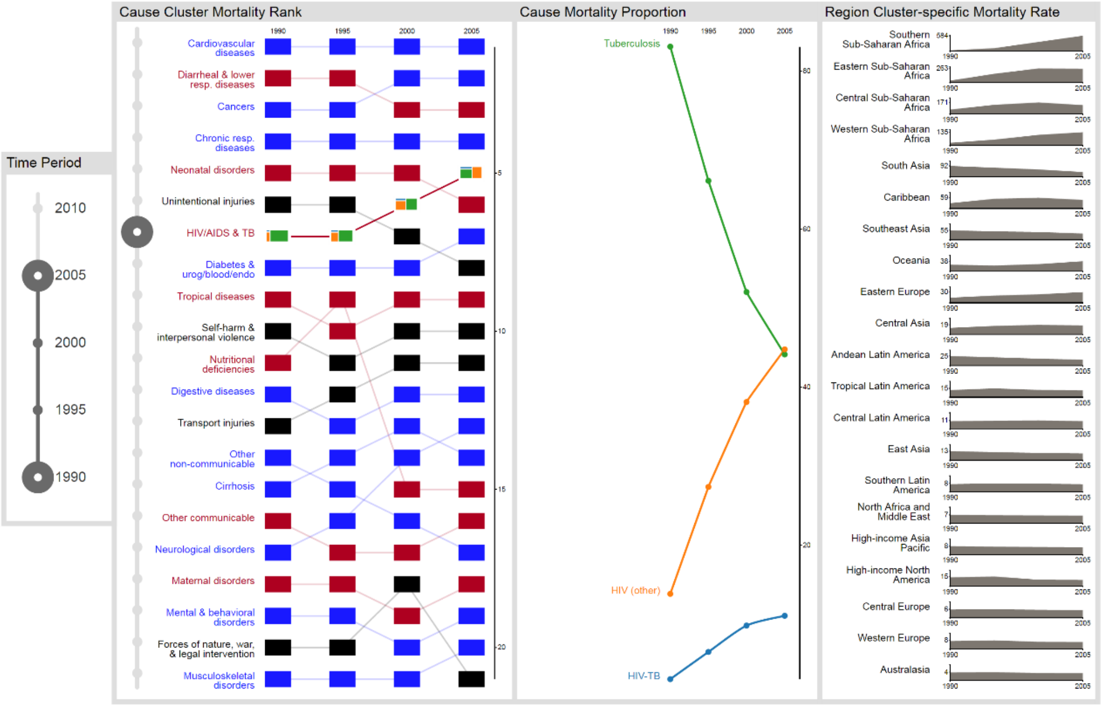

Figure 9: Overall [Fusion•Coordinate•Token•Hierarchy•Cell•Link•Group]-based visualization for chronology.

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 modes. 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.

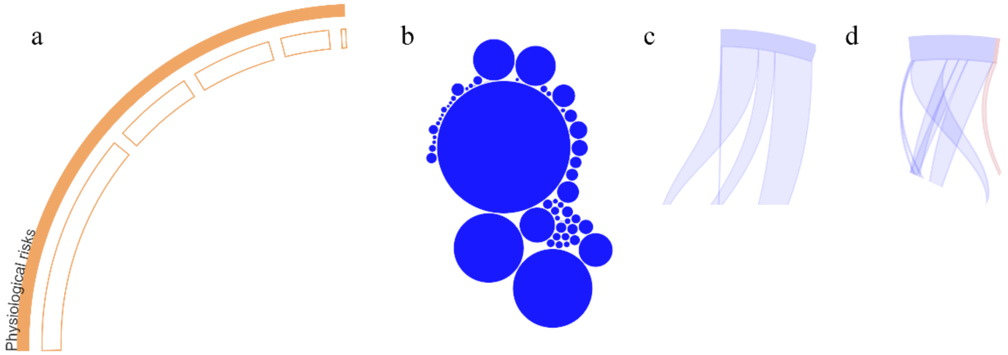

Figure 10: (a) Hierarchical structure of the physiological risk cluster (b) Representation of non-communicable disease group by individual causes (c) Diet low in fruit risk visual element (d) High fasting plasma glucose visual element.

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, 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 the number of deaths. 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[48]; one noticeable difference is that hierarchy 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.

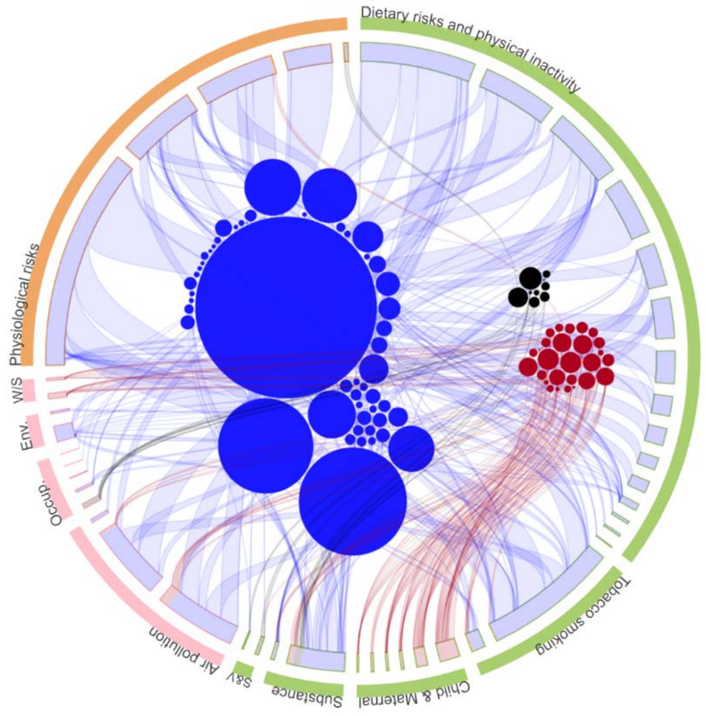

Figure 11: Geography sub-visualization for cause-risk relationships at a global level from a cause-centric point of view.

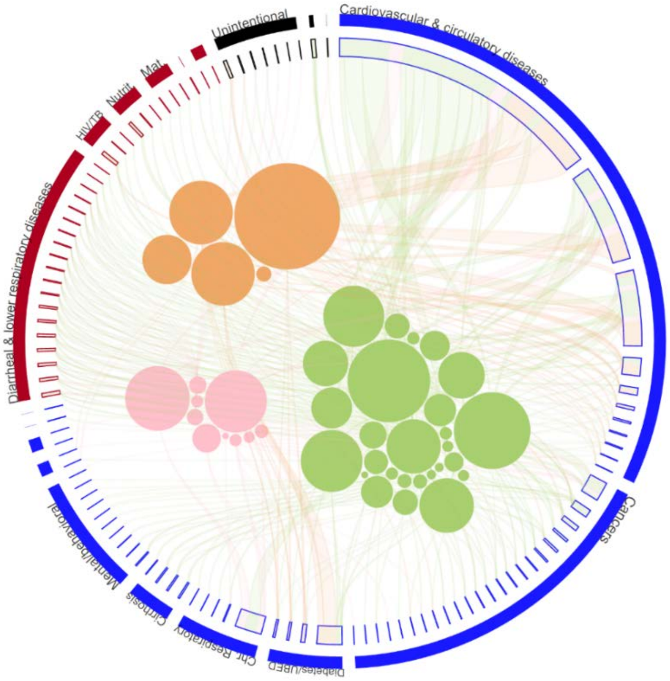

Figure 12: Geography sub-visualization for cause-risk relationships at a global level from a risk-centric point of view.

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.

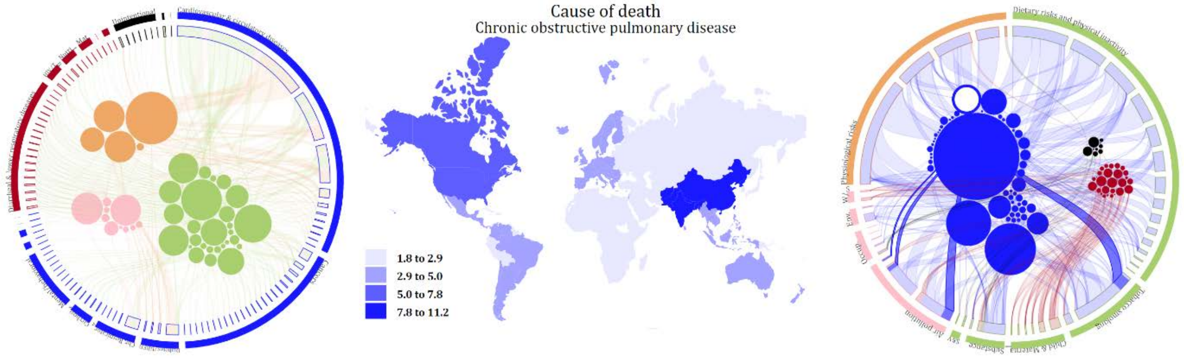

Figure 13: First three geography sub-visualizations, the impact of chronic obstructive pulmonary disease is depicted in the map-based visualization.

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 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.

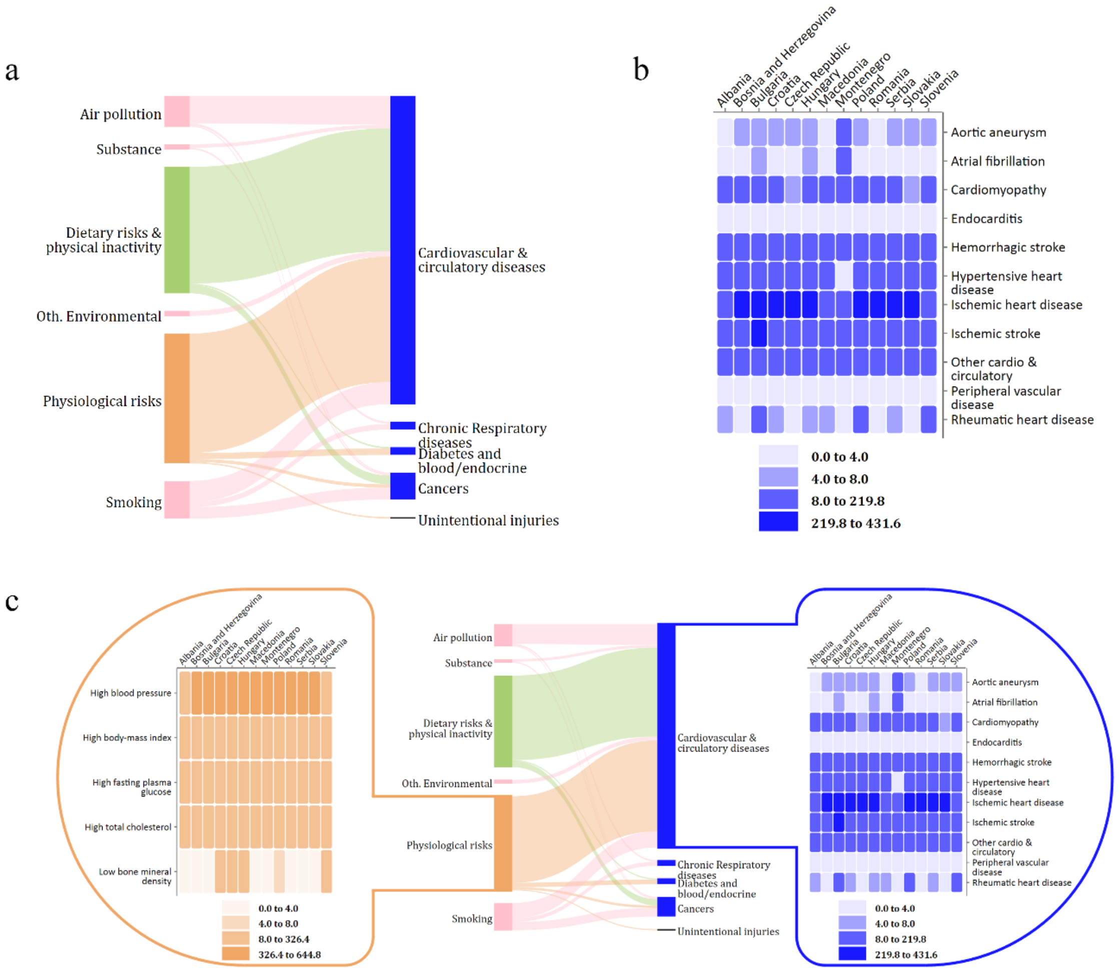

Figure 14: (a) Cause-risk cluster level relationships sub-visualization (b) Visualization of cardiovascular diseases for central European countries (c) Fourth major sub-visualization for geography which combines cause-risk cluster level relationships and risk/cause specific distribution for central Europe.

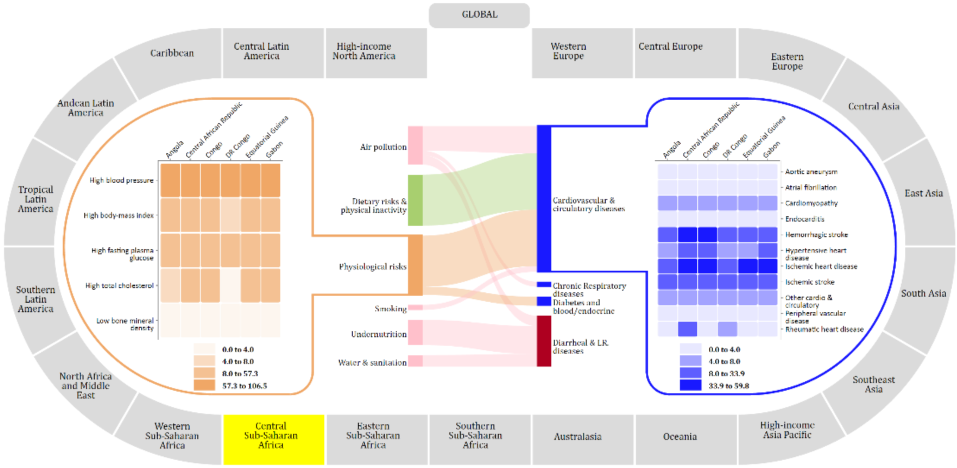

Figure 15: Geography sub-visualization for a country cluster.

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 high blood pressure is presented, as well as the relationships between causes and risk factors for the central sub-Saharan African cluster.

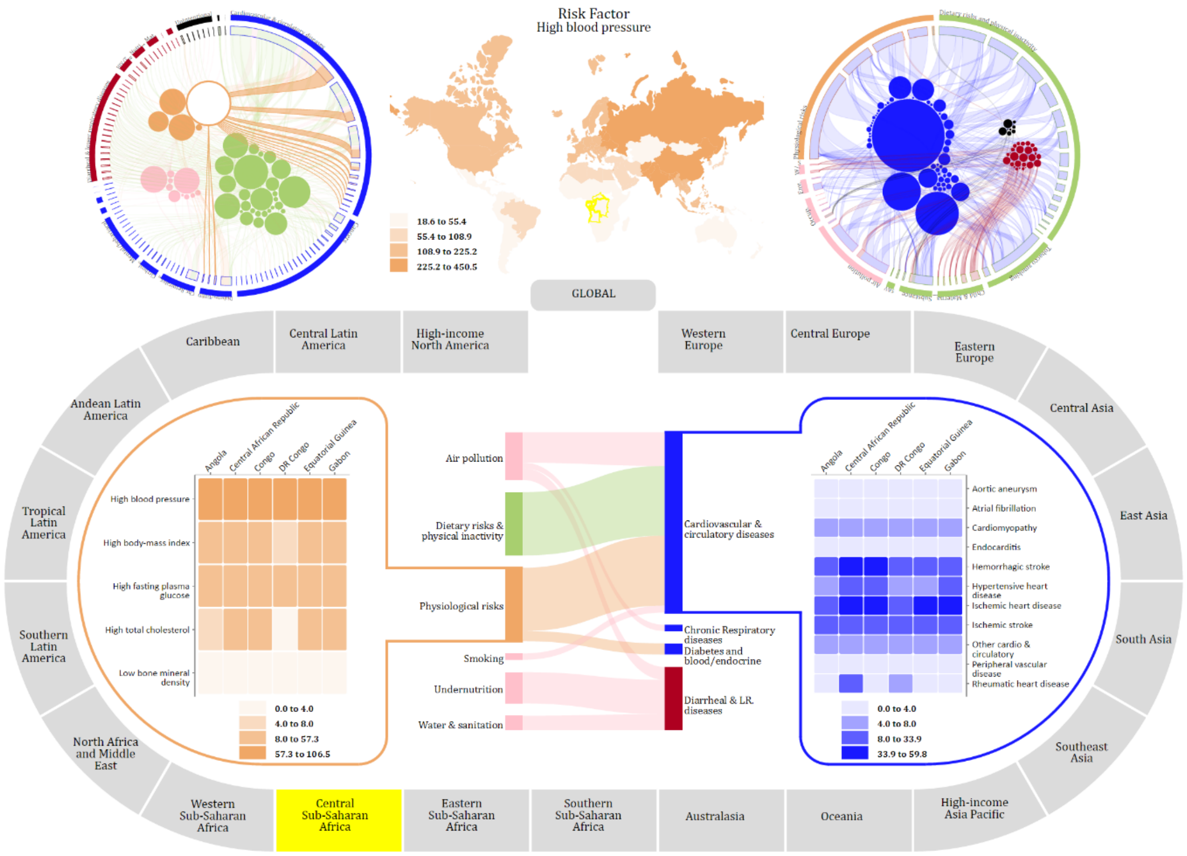

Figure 16: Overall [Branch•Token•Coordinate•List•Group•Spectrum•Area•Cell]-based visualization for geography.

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.

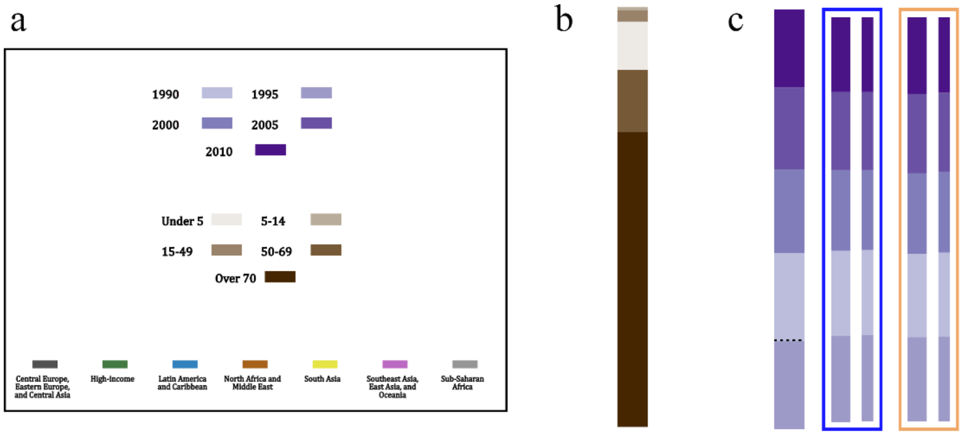

Figure 17: (a) Legends for overview visualization (b) Mortality by age group sub-visualization (c) [Stack•Token]-based representations for year-based mortality.

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.

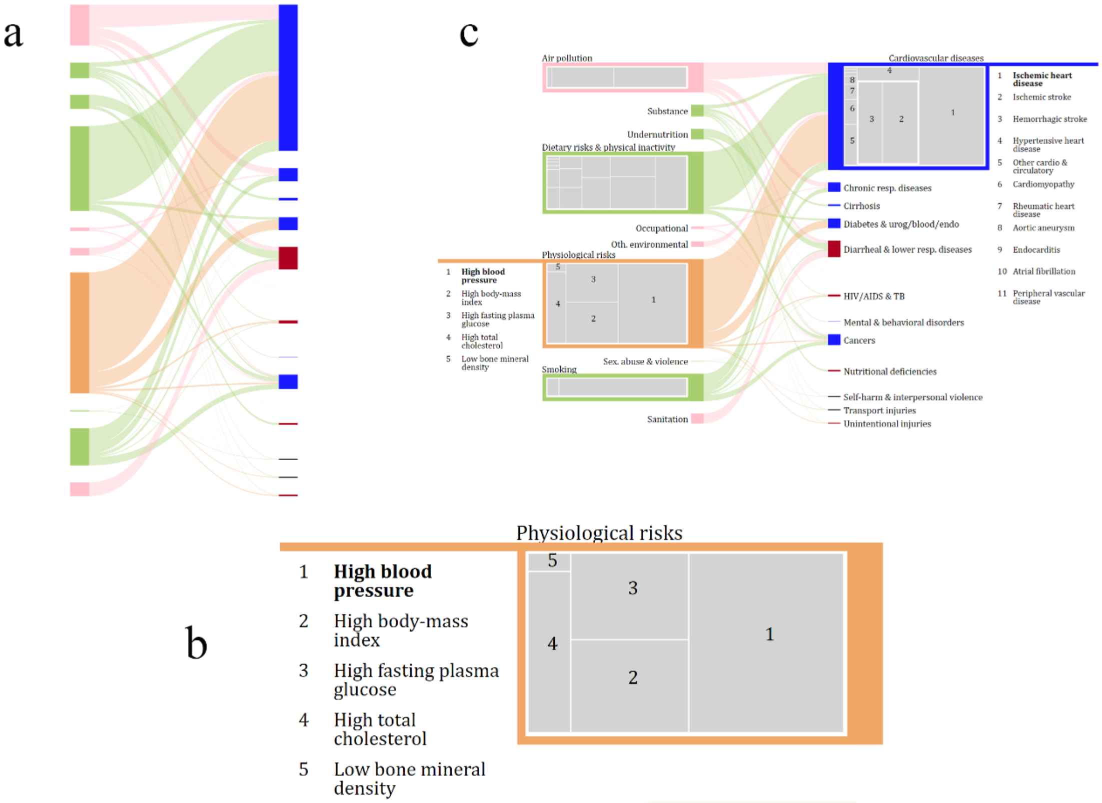

Figure 18: (a) [Branch•Token•Group]-based sub-visualization that shows the prevalent cause-risk cluster relationships at a global level in 2010 for all age groups (b) Physiological risks hierarchy and prevalence sub-visualization (c) Overview sub-visualization for cluster relationships and inter-cluster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

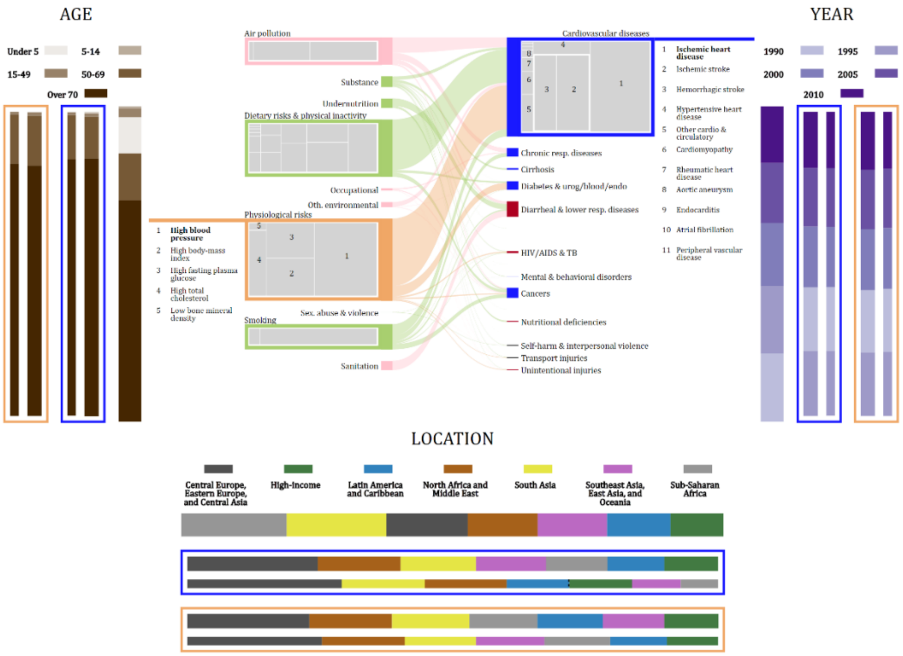

Figure 19: Overall overview [Branch•Token•Group•Cell•Hierarchy•Stack]-based visualization.

Conclusion

The health field 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 the health sector 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, thus facilitating the development of visualizations that model the depth and multifaceted intricacies of big data.

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.

References

1. Groves P, Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: Accelerating value and innovation [Internet]. 2013 [cited 2016 May 16]. Available from: http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

2. Sullivan F and. Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations. 2011.

3. Dhar V. Big Data and Predictive Analytics in Health Care. Big Data [Internet]. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2014 Sep 15 [cited 2015 Jul 16];2(3):113–6. Available from: http://online.liebertpub.com/doi/full/10.1089/big.2014.1525

4. Shneiderman B, Plaisant C, Hesse BW. Improving Healthcare with Interactive Visualization. Computer (Long Beach Calif) [Internet]. 2013 May [cited 2014 Jun 25];46(5):58–66. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6415893

5. Ola O, Buchel O, Sedig K. 2016. Exploring the Spread of Zika: Using Interactive Visualizations to Control Vector-Borne Diseases. *Int J Dis Control Contain Sustain.* 1(1), 47-68. http://dx.doi.org/10.4018/IJDCCS.2016010104

6. Carroll LN, Au AP, Detwiler LT, Fu T-C, Painter IS, et al. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform [Internet]. 2014 Apr 16 [cited 2014 May 23];51C:287–98. Available from: http://www.sciencedirect.com/science/article/pii/S1532046414000914

7. Zhang L, Stoffel A, Behrisch M, Mittelstadt S, Schreck T, et al. Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST) [Internet]. IEEE; 2012 [cited 2015 Jan 13]. p. 173–82. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6400554

8. Rind A, Wang T, Aigner W, Miksch S, Wongsuphasawat K, et al. Interactive Information Visualization to Explore and Query Electronic Health Records. Found Trends Human-Computer Interact [Internet]. 2013;5(3):207–98. Available from: http://cgis.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2010-19

9. Kosara R, Miksch S. Visualization methods for data analysis and planning in medical applications. In: International Journal of Medical Informatics. 2002. p. 141–53.

10. Aimone AM, Perumal N, Cole DC. 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. Int J Health Geogr [Internet]. BioMed Central; 2013 Jan 10 [cited 2016 May 22];12(1):1. Available from: http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-12-1

11. Faisal S, Blandford A, Potts HWW. Making sense of personal health information: challenges for information visualization. Health Informatics J [Internet]. 2013 Sep 1 [cited 2014 May 27];19(3):198–217. Available from: http://jhi.sagepub.com/content/19/3/198.short

12. Endert A, Hossain MS, Ramakrishnan N, North C, Fiaux P, et al. 2014. The human is the loop: new directions for visual analytics. *J Intell Inf Syst*. http://dx.doi.org/10.1007/s10844-014-0304-9

13. Cybulski JL, Keller S, Nguyen L, Saundage D. Creative problem solving in digital space using visual analytics. Comput Human Behav [Internet]. 2013 [cited 2014 Jan 27]; Available from: http://www.sciencedirect.com/science/article/pii/S0747563213004111

14. Gotz D, Borland D. Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. IEEE Comput Graph Appl [Internet]. 2016 May [cited 2016 May 11];36(3):90–6. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7466736

15. Sedig K, Parsons PC, Dittmer M, Haworth R. Human-centered interactivity of visualization tools: Micro- and macro-level considerations. In: Huang T, editor. Handbook of Human Centric Visualization. Springer; 2013. p. 717–43.

16. Thomas J, Cook K. Illuminating the Path: The Research and Development Agenda for Visual Analytics. Thomas JJ, Cook KA, editors. Los Alamitos, CA: IEEE Computer Society; 2005.

17. Heer J, Bostock M, Ogievetsky V. A Tour through the Visualization Zoo. ACM Queue [Internet]. ACM; 2010 May 1 [cited 2014 Jul 22];8(5):20. Available from: http://dl.acm.org/ft_gateway.cfm?id=1805128&type=html

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

19. Turner AM, Stavri Z, Revere D, Altamore R. From the ground up: information needs of nurses in a rural public health department in Oregon. J Med Libr Assoc [Internet]. 2008 Oct [cited 2012 Nov 8];96(4):335–42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2568844&tool=pmcentrez&rendertype=abstract

20. Folorunso O, Shawn Ogunseye O. Challenges in the adoption of visualization system: a survey. Chen M, editor. Kybernetes [Internet]. 2008 Oct 17 [cited 2016 May 23];37(9/10):1530–41. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-54949133359&partnerID=tZOtx3y1

21. Purchase HC, Andrienko N, Jankun-Kelly T, Ward M. Theoretical foundations of information visualization. Kerren A, Stasko JT, Fekete J-D, North C, editors. Inf Vis Human-Centered Issues Perspect [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008;4950:46–64. Available from: http://link.springer.com/chapter/10.1007/978-3-540-70956-5_3

22. Sedig K, Parsons P. Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework. Synth Lect Vis [Internet]. Morgan & Claypool Publishers; 2016 Apr 18 [cited 2016 Apr 24];4(1):1–185. Available from: http://www.morganclaypool.com/doi/abs/10.2200/S00685ED1V01Y201512VIS005

23. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data [Internet]. Springer; 2014 Jun 24 [cited 2015 Apr 27];1(1):2. Available from: http://www.journalofbigdata.com/content/1/1/2

24. Ola O, Sedig K. The challenge of big data in public health: an opportunity for visual analytics. Online J Public Health Inform [Internet]. 2014 Jan [cited 2014 Jun 2];5(3):223. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3959916&tool=pmcentrez&rendertype=abstract

25. Revere D, Turner AM, Madhavan A, Rambo N, Bugni PF, et al. Understanding the information needs of public health practitioners: A literature review to inform design of an interactive digital knowledge management system. Public Heal Informatics [Internet]. 2007;40(4):410–21. Available from: http://www.sciencedirect.com/science/article/pii/S1532046407000020

26. Fuller S. Tracking the Global Express: new tools addressing disease threats across the world. Epidemiology [Internet]. 2010 Nov [cited 2014 Jun 2];21(6):769–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20924231

27. Kiefer L, Frank J, Di Ruggiero E, Dobbins M, Manuel D, et al. 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 [Internet]. 2005 Jan [cited 2013 Jan 18];96(3):I1-40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15913085

28. LaPelle NR, Luckmann R, Simpson EH, Martin ER. Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. BMC Public Health [Internet]. 2006 Jan [cited 2012 Aug 30];6(1):89–101. Available from: http://www.biomedcentral.com/1471-2458/6/89

29. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, et al. 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 [Internet]. 2012 Dec 15 [cited 2014 Jul 10];380(9859):2095–128. Available from: http://www.thelancet.com/journals/a/article/PIIS0140-6736(12)61728-0/fulltext

30. Gazmararian JA, Curran JW, Parker RM, Bernhardt JM, DeBuono BA. Public health literacy in America: an ethical imperative. Am J Prev Med [Internet]. 2005 May [cited 2015 Jul 31];28(3):317–22. Available from: http://www.sciencedirect.com/science/article/pii/S0749379704003368

31. Rind A, Aigner W, Wagner M, Miksch S, Lammarsch T. Task Cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation. Inf Vis [Internet]. 2015 Dec 27 [cited 2016 May 16];1473871615621602-. Available from: http://ivi.sagepub.com/content/early/2015/12/23/1473871615621602

32. Sedig K, Parsons P. 2013. Interaction design for cognitive activity support tools: A pattern-based taxonomy. *AIS Trans Human-Computer Interact.* 5(2), 84-133.

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

34. Parsons P, Sedig K. Adjustable Properties of Visual Representations: Improving the Quality of Human-Information Interaction. J Assoc Inf Sci Technol [Internet]. 2013 Feb 15 [cited 2014 Jan 16];65(3):455–82. Available from: http://doi.wiley.com/10.1002/asi.23002

35. Gapminder [Internet]. Available from: http://www.gapminder.org/

36. Sedig K, Rowhani S, Liang H-N. 2005. Designing interfaces that support formation of cognitive maps of transitional processes: an empirical study. *Interact Comput*. 17(4), 419-52. http://dx.doi.org/10.1016/j.intcom.2005.02.002

37. Robertson G, Fernandez R, Fisher D, Lee B, Stasko J. Effectiveness of animation in trend visualization. IEEE Trans Vis Comput Graph [Internet]. 2008 Jan [cited 2014 Jul 21];14(6):1325–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18988980

38. Al-Hajj S, Pike I, Riecke B, Fisher B. Visual Analytics for Public Health: Supporting Knowledge Construction and Decision-Making. In: 2013 46th Hawaii International Conference on System Sciences [Internet]. 2013 [cited 2013 Apr 26]. p. 2416–23. Available from: http://ieeexplore.ieee.org.proxy2.lib.uwo.ca/xpl/articleDetails.jsp?reload=true&arnumber=6480137&contentType=Conference+Publications

39. Wang Baldonado MQ, Woodruff A, Kuchinsky A. Guidelines for using multiple views in information visualization. In: Proceedings of the working conference on Advanced visual interfaces - AVI ’00 [Internet]. New York, New York, USA: ACM Press; 2000 [cited 2014 Aug 28]. p. 110–9. Available from: http://dl.acm.org/citation.cfm?id=345513.345271

40. Fan W, Bifet A. Mining big data. ACM SIGKDD Explor Newsl [Internet]. ACM; 2013 Apr 30 [cited 2016 Jul 25];14(2):1. Available from: http://dl.acm.org/citation.cfm?doid=2481244.2481246

41. Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, et al. Big data and its technical challenges. Commun ACM [Internet]. 2014 Jul 1 [cited 2016 Jul 25];57(7):86–94. Available from: http://dl.acm.org/citation.cfm?doid=2622628.2611567

42. Heer J, Kandel S. Interactive analysis of big data. XRDS Crossroads, ACM Mag Students [Internet]. ACM; 2012 Sep 1 [cited 2016 Jul 25];19(1):50. Available from: http://dl.acm.org/citation.cfm?doid=2331042.2331058

43. Gorodov EY, Gubarev VV. Analytical Review of Data Visualization Methods in Application to Big Data. J Electr Comput Eng [Internet]. Hindawi Publishing Corp.; 2013 [cited 2016 Jul 25];2013:1–7. Available from: http://www.hindawi.com/journals/jece/2013/969458/

44. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, et al. Big data, bigger dilemmas: A critical review. J Assoc Inf Sci Technol [Internet]. 2015 Aug [cited 2016 Jul 25];66(8):1523–45. Available from: http://doi.wiley.com/10.1002/asi.23294

45. Institute for Health Metrics and Evaluation. Global Burden of Disease [Internet]. 2013. Available from: http://www.healthdata.org/gbd

46. Pirolli P, Card SK. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: 2005 Internaltional Conference on Intelligence Analysis. 2005. p. 6 pp.

47. Bodnar J. Making sense of massive data by hypothesis testing. Int Conf Intell Anal [Internet]. 2005 [cited 2014 May 30]; Available from: https://www.e-education.psu.edu/drupal6/files/sgam/Making Sense of Massive Data by Hypothesis Testing.pdf

48. Vizuly. Halo [Internet]. 2014. Available from: http://vizuly.io/product/halo/

Footnotes

[1] In the rest of this article, the terms ‘interactive visualization’ and ‘visualization’ are used interchangeably.

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

[3] Henceforth, the term ‘visualizations’ refers to both static as well as interactive visualizations.

[4] 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*[22].

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

[6] 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.

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