Geo-spatial Visualization of Population Healthcare Data

Chao Tong

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

March 15, 2019

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

This thesis describes geo-spatial visualization of population healthcare data by a literature study and some practical research.

I first present a literature survey of narrative visualization including geo-spatial visualization. Throughout history, storytelling has been an effective way of conveying information and knowledge. In the field of visualization, storytelling is rapidly gaining momentum and evolving cutting-edge techniques that enhance understanding. Many communities have commented on the importance of storytelling in data visualization, and, in growing numbers, storytellers tend to be integrating complex visualizations into their narratives. We present a survey of storytelling literature in visualization, and present an overview of the common and important elements in storytelling visualization. We also describe the challenges in this field as well as a novel classification of the literature. Our classification scheme highlights the open and unsolved problems in this field as well as the more mature storytelling sub-fields. We can see that geo-space is relatively unexplored in this context. The benefits of our work offer a concise overview and a starting point into this rapidly evolving research trend, and provide a deeper understanding of this topic.

Then, we present a novel multivariate visualization that combing geo-spatial information with population healthcare data. The National healthcare Service (NHS) in the UK collects a massive amount of high-dimensional, region-centric data concerning individual healthcare units throughout Great Britain. It is challenging to visually couple the large number of multivariate attributes about each unit region together with the geo-spatial location of the clinical practices for visual exploration, analysis, and comparison. We present a novel multivariate visualization that we call a cartographic treemap, which attempts to combine the space-filling advantages of treemaps for the display of hierarchical, multivariate data together with the relative geo-spatial location of NHS practices in the form of a modified cartogram. It offers both

space filling and geospatial error metrics that provide the user with interactive control over the space-filling versus geographic error trade-off. The result is a visualization that offers users a more space efficient overview of the complex, multivariate healthcare data coupled with the relative geo-spatial location of each practice to enable and facilitate exploration, analysis, and comparison. We evaluate the two metrics and demonstrate the use of our approach on real, large high-dimensional NHS data and derive a number of multivariate narratives based on healthcare in the UK as a result. We then report the reaction of our software from two domain experts in health science.

While previous work focused on multivariate visualization combing geo-spatial data, we further extend the work by adding time-oriented data. Cartographic treemaps offer a way to explore and present hierarchical multi-variate data that combines the space-efficient advantages of treemaps for the display of hierarchical data together with relative geo-spatial location from maps in the form of a modified cartogram. They offer users a space-efficient overview of the complex, multi-variate data coupled with the relative geo-spatial location to enable and facilitate exploration, analysis, and comparison. In this chapter, we introduce time as an additional variate, in order to develop time-oriented cartographic treemaps. We design, implement and compare a range of visual layout options highlighting advantages and disadvantages of each. We apply the method to the study of UK-centric electronic health records data as a case study. We use the results to explore the trends and present a narrative of a range of health diagnoses in each UK health care region over multiple years exploiting both static and animated visual designs. We provide several examples and user options to evaluate the performance in exploration, analysis, and comparison. We also report the reaction of domain experts from health science.

Finally, we present a novel algorithm that enhances cartogram understanding and reduces error by adding features into it. Cartograms are very popular and useful for depicting data on a map. Dorling style and rectangular cartograms are very good for facilitating comparisons between unit areas: each unit area is represented by the same shape such as a circle or rectangle, and the uniformity in shapes facilitates comparative judgement. However, the layout of these more abstract shapes may also simultaneously reduce the map's legibility and increase error. When we integrate univariate data into a cartogram, its recognizability may be reduced. There is therefore a trade-off between information recognition and geo-information accuracy. And this is the inspiration of this part. We thus attempt to increase the map's recognizability and re-

duce error by introducing topological features into the cartographic map. The goal is to include topological geographic features such as a river in a Dorling-style or rectangular cartogram to make the visual layout more recognizable, increase map cognition and reduce geo-spatial error. We believe that compared to the standard Dorling and rectangular style cartogram, adding topological features provides familiar geo-spatial cues and flexibility to enhance the recognizability of a cartogram.

Acknowledgements

I would not be able to finish this thesis without the support of a number of people who have had a direct or indirect impact on my study for a PhD in Swansea. Having a retrospective view of this long journey, I am very thankful to all these people. Especially, I would like to thank my supervisor Dr. Robert S. Laramee for his endless and tireless guidance, support, care, patience and encouragement during my four years of study. I appreciate everything that I have learnt from him, including the amazing "minutes of meeting" and the "evil" coding conventions [1]. The help provided by him is not just about academic, but also about daily life, such as the optimism and positive attitude, and healthy living style. I would also thank my second supervisor Dr Daniel Archambault, who gave me great suggestions and advice about my research topic.

I am very thankful to the whole Department of Computer Science, especially the Visual and Interactive Computing Group. Thanks the whole group and the visible lunch program which aided sharing our knowledge, helping practice with presentations, and gave me valuable feedback. Especially, thanks to Richard Roberts, Liam McNabb, Dylan Rees for proofreading all of my submitted papers.

I also want to thanks all the co-authors of my publications: Rita Borgo , Kodzo Wegba, Aidong Lu, Yun Wang , Huamin Qu , Qiong Luo, Xiaojuan Ma, Jane Lyons, Angharad Walters, Damon Berridge and Daniel Thayer, Dave Greten, and Sean Walton.

My final words are dedicated to my family: my father Jianwei Tong and my mother Shuying Cao, for their unconditional love and support although I am so far away from them; my beloved wife, QiGao who has been taking good care of me with her gentleness and patience.

List of Publications

This thesis is based on the following publications:

1. Chao Tong, Roberts Richard, Borgo Rita, Laramee Robert S, Wegba Kodzo, Lu Aidong, Yun Wang, Huamin Qu, Qiong Luo, and Xiaojuan Ma. Storytelling and Visualization: A Survey. In Proceedings of the 9th International Conference on Information Visualization Theory and Applications (IVAPP) 2018. Funchal, Madira, Portugal, pages 212-224, 27-29 January 2018.

2. Chao Tong, Richard Roberts, Rita Borgo, Sean Walton, Robert S Laramee, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, and Xiaojuan Ma, Storytelling and Visualization: An Extended Survey, Information, Volume 9, Number 3, March 2018 pages 1-42

3. Chao Tong, Richard Roberts, Robert S Laramee, Daniel Thayer, and Damon Berridge. Cartographic Treemaps for the Visualization of Public Healthcare Data. In The Computer Graphics and Visual Computing (CGVC) Conference 2017. Manchester, UK, September, 2017.

4. Chao Tong, Liam McNabb, Robert S Laramee, Jane Lyons, Angharad M Walters, Daniel Thayer, and Damon Berridge. Time-Oriented Cartographic Treemaps for the Visualization of Public Healthcare Data. In The Computer Graphics and Visual Computing (CGVC) Conference 2017. Manchester, UK, September, 2017.

5. Chao Tong, Liam McNabb, and Robert S. Laramee, Cartograms with Topological Features, The Computer Graphics and Visual Computing (CGVC) Conference 2018, 12-14 September 2018, Swansea, UK

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivation

Contents

"Live as if you were to die tomorrow. Learn as if you were to live forever." -Mahatma Gandhi[1]

Data visualization is a general term that describes any effort to help people enhance their understanding of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in numeric on text-based data can be exposed and recognized easier with data visualization software [49]. Figure 1.1 represents a ubiquitous pattern of knowledge

[1] Mahatma Gandhi (1869-1948) was an Indian activist who was the leader of the Indian independence movement against British rule.

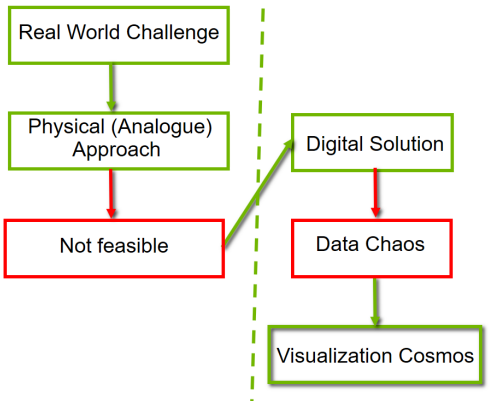


Figure 1.1: *A ubiquitous pattern of knowledge evolution [2].*

evolution that the collective digital society is experiencing. It consists of six basic constituents. It starts with a challenge or goal in the real world. The goal could be to build or optimize a design, like a car or computer. The start could be a challenge such as reaching a new level of understanding or observing a behavior or phenomenon rarely or never seen previously. The goal could be running a successful business and making a profit. We all have real world goals and challenges. We all have new understanding and knowledge we would like to obtain. We all have things we would like to build, create, and optimize.

When trying to build something we generally know that whatever it is, it can theoretically be built in the real-world. For example cars and structures can be built out of raw materials and components with the right tools. We also know that observations can be made, in general, by being in the right place at the right time, either personally or with recording equipment. Experiments can generally be conducted with the appropriate equipment. New levels of understanding can generally be obtained if enough people are employed to carry out of the task. This is what we called real-word solution.

However, when implementing a real-world solution, we often run into barriers. Cars and structures are extremely expensive to build and may also require a long term investment. Observations may be very expensive, very difficult, or even impossible. Some observations interfere with the very behavior or phenomena they are trying to study. Recording equipment may be too expensive or cause logistical problems. Equipment for experiments is generally very expensive. This is especially true if the equipment is specialized or for very small or very large scale investigations. Also, hiring people for new understanding may not be feasible due to

expense. A full-time research assistant costs 100K GBP per year under current funding agency full economic costing (FEC) requirements in the UK. Real-world solutions are generally very expensive or not feasible at all. Some real-world solutions are impossible.

It is because of the high cost of real-world solutions that collectively, as a society, we turn to digital solutions to address our challenges and goals. The dotted line in Figure 1.1 separates the real, physical, or analogue world on the left side from the digital world on the right. We all look to the digital world for the answers to our questions. "There must be an app for that." or "What app can be built to solve this problem?" is the collective thinking in this day in age. Society looks towards digital solutions for their real-world problems to deliver the user from the dilemma they may face. People believe that software is less-expensive to build than objects in the real world. The virtual world should be more feasible than the physical or analogue world. And this is true in many scenarios.

However, creating a digital solution to an analogue problem introduces new challenges. In particular, digital solutions including software, create massive amounts of data. The amount of data digital approaches generate is generally unbounded. Software and storage hardware are less and less expensive with time. Thus users collect, collect, and collect even more data. This is the point at which the knowledge evolution pipeline of Figure 1.1 becomes interesting. Large collections of complex data are not automatically useful. Extracting meaningful information, knowledge, and ultimately wisdom from large data collection is the main challenge facing the digital world today. The collection of essentially unbounded data is what we term data chaos. Collecting and archiving data without careful planning and well thought out information design quickly or slowly results in a chaotic data environment. Those who collect data are generally not yet aware of how difficult it is to then derive useful insight and knowledge from it.

On the other hand, the knowledge that visualization is a key technology to extract meaning from large data sets is rapidly spreading. This is one solution to the data chaos. In the early years of data visualization as a field, say the first 10 years, from 1987-1997, data visualization was considered very niche. Not many people knew about it nor knew of its existence. It is only since around the turn of the century that word started to spread. In the 2000s the first mainstream news stories including the phrase 'Data Visualization' were published. Nowadays, the field has come a long way from obscurity to breaking into the main stream. Its presence and importance as a field is starting to become understood. Word is spreading that a data visualization community exists and that this is a topic a student can study at university.

Figure 1.1 is the basic pattern of knowledge evolution [2]. The rest of the chapter provides concrete examples of these six stages from real-world challenges to the visualization cosmos. The focus is on the last two stages: from data chaos to the visualization cosmos.

1.1 The Universal Big Data Story (and Quandary)

We can find this pattern everywhere. It doesn't matter where we look. We can see in computational fluid dynamics. Physicists and astronomers are facing the challenges of big data. It's not possible to study all the stars and black holes physically. We see this pattern with marine biologists, biochemists, psychologists, sociologists, sport scientists, journalists, and those studying the humanities. We see this evolution with government councils, banks, call centers, retail websites, transportation. The list is virtually endless. You can experience this yourself as you collect your own photos. People like to collect things. This is another contributing factor to the data chaos. A person may not even have a goal to reach or a problem they are trying to solve. They just like to collect.

1.2 The Visual Cortex

Data visualization uses computer graphics to generate images of complex data sets. It's different from computer graphics."Computer graphics includes the creation, storage, and manipulation of model and images of objects. Computer graphics concerns the pictorial synthesis of real or imaginary objects from their computer-based models, whereas the related field of image processing treats the converse process: the analysis of scenes, or the reconstructing of models of 2D and 3D objects from their picture. "from the classic textbook "Introduction to Computer Graphics "by Foley et al, 2000. Visualization tries to generate images of reality. Visualization exploits our powerful visual system. We have several billion neurons dedicated to our visual processing and visual cortex [50].

The numbers of neurons are not very meaningful unless we put them into context. We have eight percent of the cortex dedicated touch and three percent dedicated to hearing. We have anywhere from 4 to 10 times of our cortex dedicated to visual processing than the other senses. It is advantageous to explore the visual processing power in our brains as opposed to the other senses. It's dedicated to processing color, motion, texture, and shape.

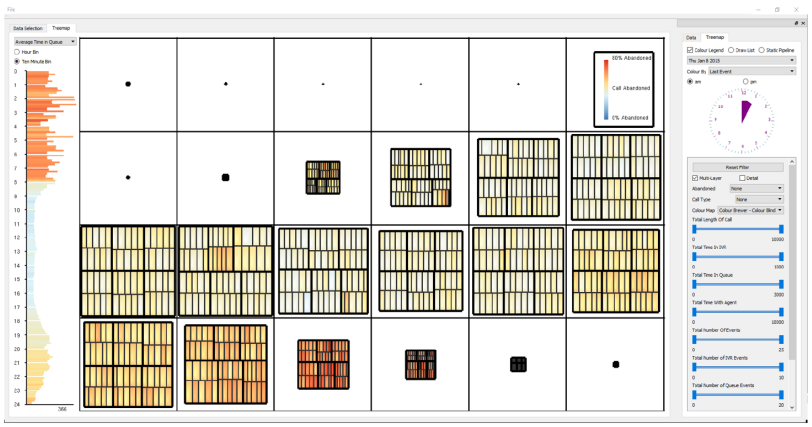


Figure 1.2: *Visualization of call center data. Image courtesy of Roberts et al. [3]*

1.3 Visualization Goals

Data visualization has some strengths and goals itself. One of the goals of data visualization is exploring data. This may be the case when the user does not know anything about their data set. They just want to find out what it looks like and its characteristics.

Users search for trends or patterns in the data. Exploration is for the user that's not very familiar with the dataset. Visualization is also good for analysis: to confirm or refute a hypothesis. An expert may have collected the data for a special purpose and would like to confirm or refute a hypothesis or answer a specific question. Visualization is also effective for presentation.

When our exploration and analysis is finished we can present the results to a wider audience. Visualization is also good for acceleration i.e. to speed up something such as a search process. This is often a decision making process or knowledge discovery process. We can see things that were otherwise impossible.

We try to explain the knowledge pattern and motivate the topic of visualization with several examples. The examples cover from business area to research area and from micro world to macro world.

1.4 Example: Visualization of Millions of Calls

Let's examine at this first example of this pattern of knowledge evolution. This is from a business context. One of Swansea University's industry collaborators is called QPC Ltd. They are an innovator in call center technology. Their goal is to understand call center behavior and to increase understanding of calls and all the activities that occur inside a call center. The call centers are staffed with many agents and the agents answer hundreds of thousands of calls every day. How can we increase our understanding of all those events and what is happening inside of a call center?

We theoretically could go down the analog or physical route. We could hire more people that stand and observe what's happening in the call center, and attempt to take notes to enhance understanding. Or maybe CCTV could be used to try to film everything that's going on. These analogue solutions will be very expensive and not very practical. The analog solution to hire more people for just observation is not practically feasible and will cost too much money.

So QPC Ltd chose the digital solution. They decided to implement an event database. The database logs all events in the call center: who called, when they call, how much time they spend navigating with menus inside the interactive voice recognition system (IVR), how long they spent in the queue before speaking to an agent, whether or not they abandon their call, which agent they spoke to, and how long they spoke to each agent etc. That digital solution in the form of a database stores of millions events everyday. A call center generates lots of activities. The UK employs over a million people in call centers or about five percent of its workforce are employed in call centers [51]. It's a large market.

How do we take the chaos of call center data and visualize it to make sense of it? We can use a treemap as one of the ways to visualize call center events. See Figure 1.2. The treemap is a hierarchical data structure. We start with an overview of the data and then zoom in down to different levels of detail. In this case, the size of the rectangles is initially mapped to call volume. The different hours start from midnight to midnight again. We can see when the call center opens and when the call volume increases and reaches its maximum at around lunchtime. Then it starts to descend again.

Color is mapped to the percentage of abandoned calls by default. We can notice call centers trying to avoid abandoned calls. We can observe a big increase in abandoned calls in the evening right after dinner around 7pm-8pm. The user can map the calls to different colors at different costs. They can also map the colors to different kinds of events for example

abandoned calls or successful calls.

We can also navigate the treemap. We can zoom in smoothly and see more details. We can zoom in to single hour and each rectangle represents a single call. We can visualize individual calls and how long they take. There is a call that lasted two hours. The unusual calls that last long time jump right out. Probably they spent a long time with an agent – a very dedicated agent spent a long time trying to solve a customer problem. The users can use a clock interface to smoothly zoom and navigate each hour. The software features a smooth zooming and panning operation and with the clock showing. The user does not get lost.

We can easily see which hours we are observing even when we zoom in. We can zoom in even further, one hour is broken up into 10-minute intervals and then those 10-minute intervals are broken up into single minute intervals. We also see a standard histogram on the left which represents the data and provides an overview. Each bar represents a 10-minute interval. Color can also be mapped to some data attribute chosen by the user in this case the average call length which we can see up in Figure 1.2. We can see, suddenly during, the evening average call length increases and we can see over the day the average call length increases throughout the day as an overall trend.

The treemap features a fine level of detail. Each rectangle can represent a single phone call and in this case how long each call lasted.

At the top level are not individual calls. Each rectangle represents an hour and then each hour is broken up into 10 minute blocks. So we have 6, 10 minute blocks and then each time in the block is broken up into individual minutes. This is an exciting project because this is the first time that QPC Ltd have ever seen overview of the call center activity in any way shape or form. As soon as we see the overview we can easily make observations about the call center volume about the increasing level of abandon calls. The average call length is also increasing as we examine the day.

We can filter calls using different sliders. This is the analytical part of the process. This is an example of focus and context visualization. See Figure 1.3. We focus on the calls that spend a longer time in a queue. We can focus on the inbound calls because call centers have inbound calls and outbound calls. These can be filtered by completed calls. We can combine filters in different ways.

We can click on an individual call and then obtain the most detailed level of information like how much time the caller spent in the IVR navigating menus, how much time they spent

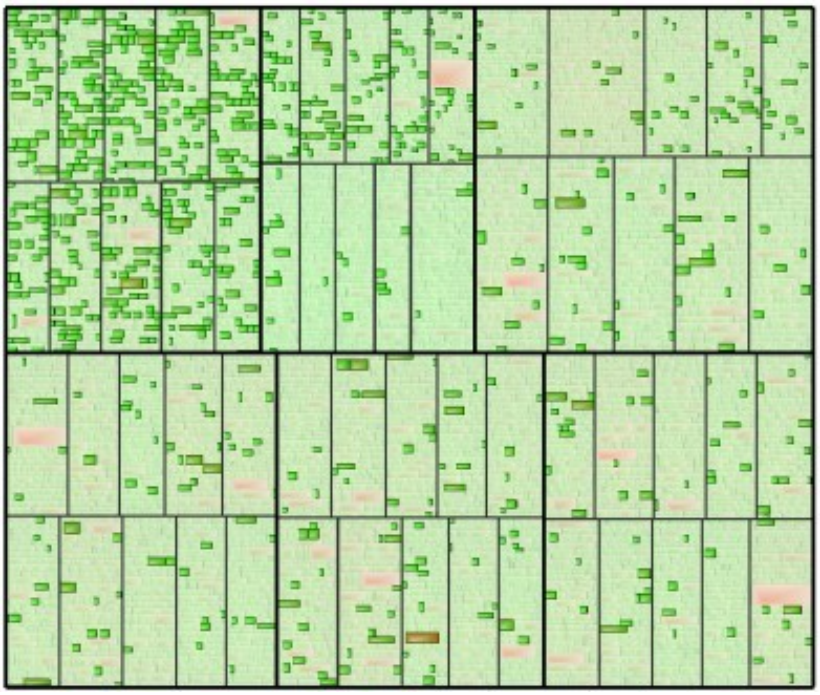


Figure 1.3: *Focus + context filtering feature of call center data. Image courtesy of Roberts et al. [3]*

queuing and how much time they spent talking to agents. We have two different queuing events, an agent event, a second agent event, back in the queue, back to another agent, back into the queue again. That is a complicated phone call. That is the lowest level of detail. We can also see the type of call in this case a consult call as it shows the number of events, one IVR event for queuing events and four different agent event.

One detailed view shows that the proportion each event as a unit proportion because sometimes the events disappear when they're too short for a traditional version.

In this example, the knowledge pattern start with real business world problem and cannot be solved with physical approach easily. So they collected data and analyzed data using visualization tools.

1.5 Example: Visualization of Sensor Data from Animal Movement

The next example is from marine biology. Marine biologists would like to understand marine wildlife and how marine wild life behaves. One of the challenges that they face is deep sea underwater diving. How do you study animals that dive deep underwater for hours or even days at a time? How is that possible? Theoretically the solution might be to follow the animal. That might be kind of an approach. But there are some problems with that. People cannot just dive a few kilometres underneath the water. They can try to build submarines or similar but to try to follow a cormorant or tortoise in a submarine is not a very practical solution. It's not feasible, very expensive, and the analog solution is one of those cases where the observation itself influences the behavior we are trying to study.

Marine biologists look to the digital world for a solution. They use sensor devices at Swansea University called a daily diary [4]. They actually capture the animals like a cormorant. They attach the digital sensor or maybe more than one digital sensor to the subject and then release it. See Figure 1.4. Then they recapture the sensor a few hours or a few days later. They remove it from the animal and they study the information that it collects about the local environment. It collects information on acceleration, local acceleration, local temperature, pressure, ultraviolet light, and a few other properties. Another challenge currently is that GPS does not work underwater at great depths. It's not possible to just plot a path naively in a dead reckoning fashion the same way we can for land animals.

However when the user get this data this is what it looks like (See Figure 1.4 right). This is a tiny little piece of what it looks like. They plot, for every attribute, magnitude versus time. Acceleration Magnitude is on the y-axis and time is on the x-axis. They claim they can infer animal behavior based on these wave patterns. They can look at a wave pattern and say that it looks like the animal is diving or the animal hunting.

But you can see that that's not easy. This is only a few seconds of data. If you plot the day's worth of data in this fashion, it will wrap around a building a few times. The acceleration has three components: *x, y, z*. These are three components decoupled. In reality they form a vector in 3-space.

The marine biologists asked us if we can drive visualizations that facilitate the understanding of marine wildlife behavior. We have a standard visualization coupled with a new visual-

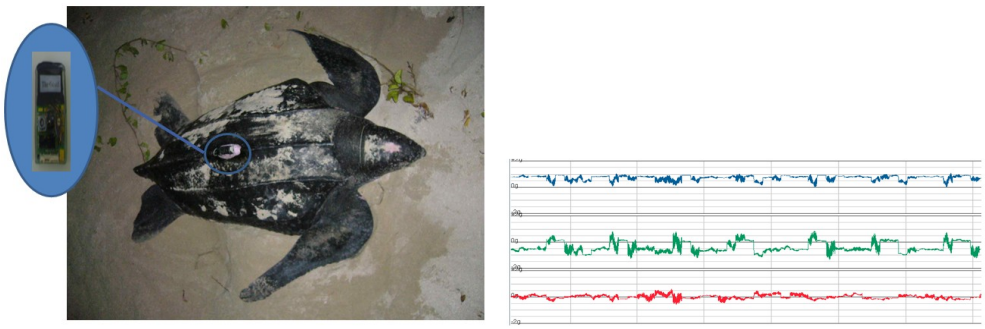


Figure 1.4: *Visualization of Sensor Data from Animal Movement. Image courtesy of Grundy et al. [4]*

ization. (See Figure 1.5.) In the new visual design we can see the geometry of the animals and how the animal is oriented immediately. What Grundy et al. did was reintegrate the *x, y, z* components of the acceleration and plot them in spherical space rather than time versus amplitude space. And they map the unit vectors onto a sphere and can immediately infer animal behavior. They can also map pressure to the radius. Figure 1.5 shows the animal swimming at the surface and then the pressure increases. Pressure mapped to radius represents diving behavior and the diving behavior is very easy to notice. Now that is visualized in spherical space we can observe swimming, hunting, searching behavior. This spherical space is interactive so that we can rotate, zoom, and pan at different angles.

Figure 1.6 presents a spherical histogram. The vectors are binned into unit rectangles and the more time an animal spends in a given posture at that orientation, the longer histogram bin. We can see the postures and the states that the animals spend a long time in. Rather than focusing on all of the time, the user chooses a special region and then the region is plotted up close in the left-hand corner. The user can cluster the vectors into different groups. (See Figure 1.7) Assigning each data point to a group that represents some interesting aspect of the animal behavior. The user can adjust the probability of any data sample belonging to one of the clusters. These are clusters of animal postures calculated using K-means clustering. Grundy et al. can represent clusters as spheres and then connect the spheres or the postures with edges that represent transitions from one orientation to another successively. We can observe the transitions between various states and postures. We can see the most popular or dominant states. That information pops up immediately.

In this example, the knowledge pattern starts with real research world problem which cannot be solved with ordinary physical approach. A digital solution is good way for collecting

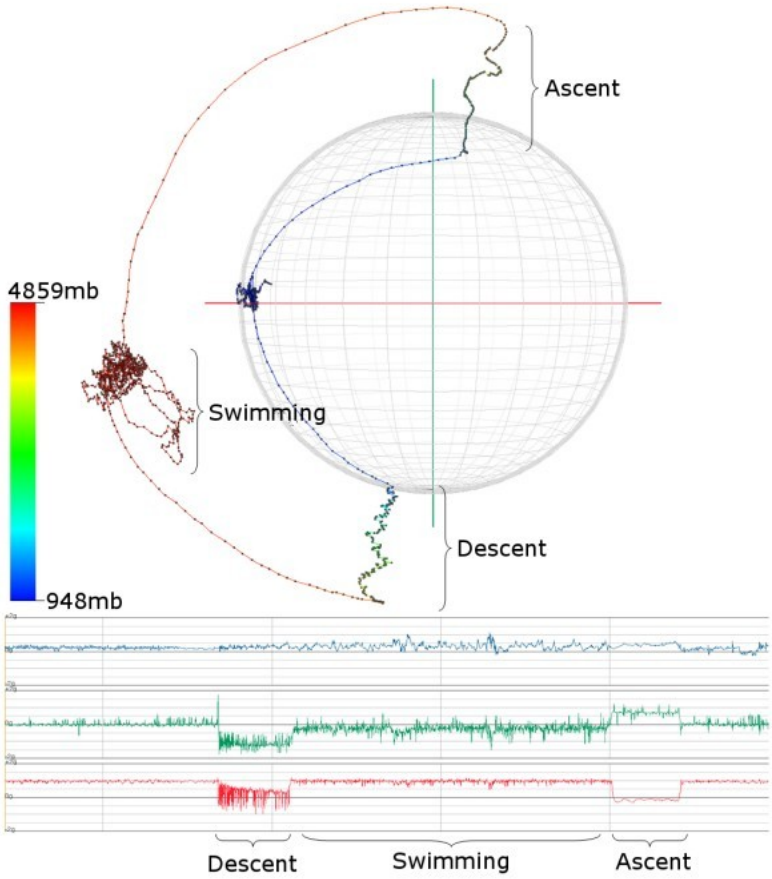


Figure 1.5: *Spherical visualization of sensor data coupled with standard visualization (bottom). Image courtesy of Grundy et al. [4]*

corresponding data, and visualization tools is developed for analyzing data and generating useful observations.

1.6 Example: Visualization of Molecular Dynamics Simulation Data

The goal here is to understand biology at the molecular level. There are analog approaches and solutions to this challenge. Biologists run experiments at the molecular level and try to understand behavior of molecules using experiments and nuclear magnetic resonance spectroscopy. These machines and experiments are very expensive.

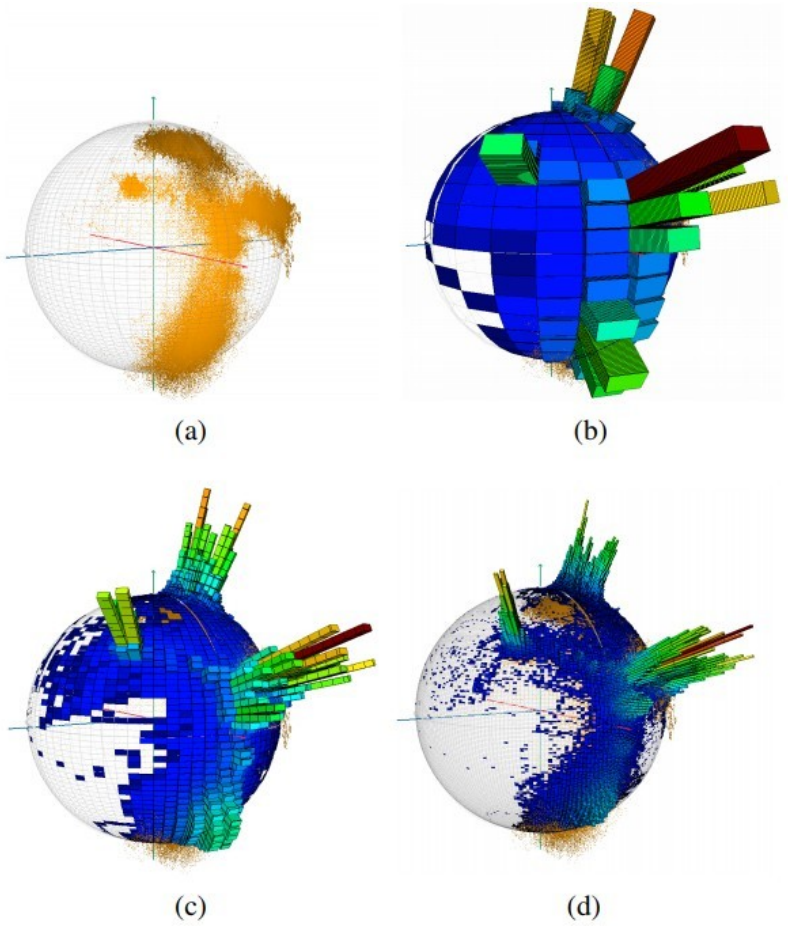


Figure 1.6: *Spherical histogram of sensor data. Image courtesy of Grundy et al. [4]*

The whole field of computational biology attempts to address this challenge in the digital world because it's much less expensive than the analog world. As with any simulation data all the simulation experts generate massive amounts of data. They try to use the latest high performance computing machines.

This is the interaction of lipids and proteins. See Figure 1.8. That's what this simulation data shows and Alharbi et al. [5] develop some visualization software to enhance understanding of this. These holes are protein and then the paths are lipid trajectories. See Figure 1.8. The computational biologists attempt to visualize the interaction between trajectories and the proteins.

Alharibi et al. are trying to develop visualizations to help computational biologist under-

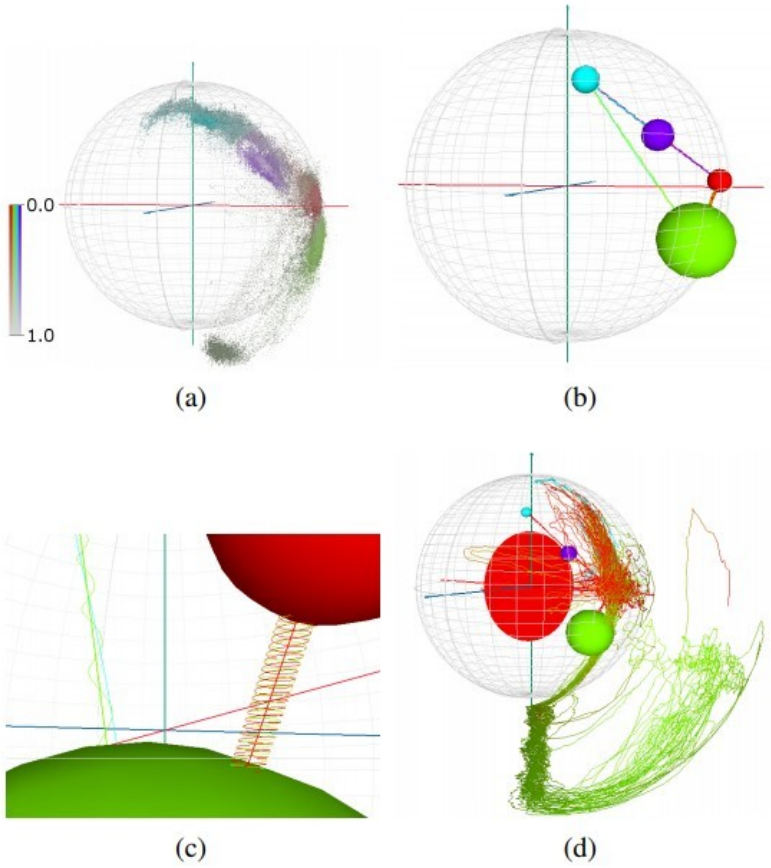


Figure 1.7: *Utilising data clustering methods of sensor data. Image courtesy of Grundy et al. [4]*

stand the data with a special focus, in this case, on path filtering. Given the massive number of trajectories hundreds of thousands or millions of trajectories over multiple time steps, is it possible to select a subset of those trajectories based on interesting properties that help the biologists understanding the behavior? Alharibi et al. develop tools for filtering and selection of these trajectories to try to understand behavior. One example is just changing the time step of the simulation or filtering the path by its length. They can focus on shorter paths or on longer paths. They can slide the filter over to long paths or the long trajectories.

The user can filter the paths based on other characteristics. They chose a few properties that they hope will be interesting for the computational biologists. One property is curvature. There are highly curved paths.

The atom trajectories are actually three dimensions but they're limited to a layer analogous to the biosphere such that the z dimension is relatively small compared to the x and y

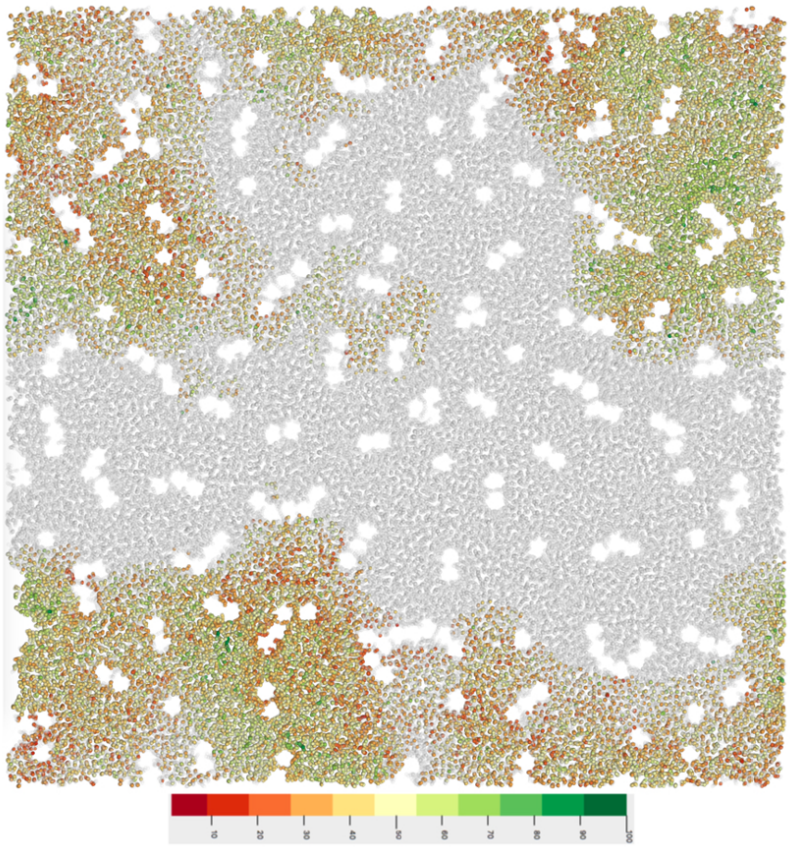


Figure 1.8: *Visualization of molecular dynamics simulation data. Image courtesy of Alharbi et al. [5]*

dimensions. They can visualize projected 2D space or the volumetric 3-space. The user can experiment with 2D versus 3D. The standard visualization packages for this are constrained to a two-dimensional plane and they're generally not interactive.

In this example, it focuses on real micro world challenge. It is too expensive to analyze the biology in molecular level by physical solution. Collecting data and developing visualization tools tend to be a better way for this problem, which fit our knowledge evolution pattern.

1.7 Example: Visualization of Public Healthcare Data

The last example is to analyse population healthcare. It starts with the real world challenge which is improving the health of the British population. The standard approach by the NHS to address this challenge is using hospitals and providing services through chemists, doctors

and other health related staff in UK. However, this approach is very costly. There are limits to the NHS budget. The NHS then try to collect as much data as possible to investigate whether they can optimise their services by analysing population healthcare data. That is the service of public healthcare in England [46] do. They switch from physical solution to a digital alternative inspired by budget constraints. The goal of this thesis is to help the domain experts make sense of the public healthcare data by using visualization. It fits the pattern in taking data chaos provided by Public Healthcare England and turning it into understanding.

1.8 Conclusion

This chapter presents a ubiquitous model of knowledge evolution witnessed at a collective level by a society deeply involved with the digital world. It presents a theory supported by a number of case studies ranging from the call center industry, to computational biology and to healthcare area. It sets the stage for data visualization as a vital technology to evolve our understanding of data and the world it describes to the next level. It will be exciting to witness how this model and pattern evolve over time.

1.9 Thesis Overview

This section contains summaries for the following main chapters. For continuity, these summaries are arranged to start with corresponding chapter headings.

Chapter 2: A Survey of Narrative Visualization Including Geo-space

Chapter 2 presents a literature survey of narrative visualization including geo-spatial visualization. Throughout history, storytelling has been an effective way of conveying information and knowledge. In the field of visualization, storytelling is rapidly gaining momentum and evolving cutting-edge techniques that enhance understanding. Many communities have commented on the importance of storytelling in data visualization. Storytellers tend to be integrating complex visualizations into their narratives in growing numbers. In this Chapter, we present a survey of storytelling literature in visualization and present an overview of the common and important elements in storytelling visualization. We also describe the challenges in this field as well as a novel classification of the literature. Our classification scheme highlights the open and unsolved problems in this field as well as the more mature storytelling sub-fields. We can see that geo-space is relatively unexplored in this context. The benefits offer a concise

overview and a starting point into this rapidly evolving research trend and provide a deeper understanding of this topic.

Chapter 3: Cartographic Treemaps for the Visualization of Healthcare Data

From Chapter 2 we learn that including geo-space can increase memorability and cognition. Chapter 3 presents a novel multivariate visualization combing geo-spatial information. The National healthcare Service (NHS) in the UK collects a massive amount of high-dimensional, region-centric data concerning individual healthcare units throughout Great Britain. It is challenging to visually couple the large number of multivariate attributes about each unit region together with the geo-spatial location of the clinical practices for visual exploration, analysis, and comparison. We present a novel multivariate visualization we call a cartographic treemap that attempts to combine the space-filling advantages of treemaps for the display of hierarchical, multivariate data together with the relative geo-spatial location of NHS practices in the form of a modified cartogram. It offers both space filling and geospatial error metrics that provide the user with interactive control over the space-filling versus geographic error trade-off. The result is a visualization that offers users a more space efficient overview of the complex, multivariate healthcare data coupled with the relative geo-spatial location of each practice to enable and facilitate exploration, analysis, and comparison. We evaluate the two metrics and demonstrate the use of our approach on real, large high-dimensional NHS data and derive a number of multivariate narratives based on healthcare in the UK as a result. We report the reaction of our software from two domain experts in health science.

Chapter 4: Time-Oriented Cartographic Treemaps

While the previous chapter focused on multivariate visualization combing geo-spatial data, in Chapter 4 we extend the work by adding time-oriented data. Cartographic treemaps offer a way to explore and present hierarchical multi-variate data that combines the space-efficient advantages of treemaps for the display of hierarchical data together with relative geo-spatial location from maps in the form of a modified cartogram. They offer users a space-efficient overview of the complex, multi-variate data coupled with the relative geo-spatial location to enable and facilitate exploration, analysis, and comparison. In this Chapter, we introduce time as an additional variate, in order to develop time-oriented cartographic treemaps. We design, implement and compare a range of visual layout options highlighting advantages and disadvantages of each. We apply the method to the study of UK-centric electronic health records data as a case study. We use the results to explore the trends and present a narrative

of a range of health diagnoses in each UK health care region over multiple years exploiting both static and animated visual designs. We provide several examples and user options to evaluate the performance in exploration, analysis, and comparison. We also report the reaction of domain experts from health science.

Chapter 5: Cartograms with Features

Chapter 5 presents a novel algorithm that enhances cartogram understanding and reduces error by adding features to a cartogram. Cartograms are very popular and useful for depicting data on a map. Dorling style and rectangular cartograms are very good for facilitating comparisons between unit areas. Each unit area is represented by the same shape such as a circle or rectangle, and the uniformity in shapes facilitates comparative judgment. However, the layout of these more abstract shapes may also simultaneously reduce the map's legibility and increase error. When we integrate univariate data into a cartogram, the recognizability of a cartogram may be reduced. There is a trade-off between information recognition and geo-information accuracy. This is the inspiration behind this Chapter. We thus attempt to increase the map's recognizability and reduce error by introducing topological features into the cartographic map. Our goal is to include topological geographic features such as a river in a Dorling-style or rectangular cartogram to make the visual layout more recognizable, increase map cognition and reduce geo-spatial error. We believe that compared to the standard Dorling and rectangular style cartogram, adding topological features provides familiar geo-spatial cues and familiarity to enhance the recognizability of a cartogram.

Chapter 6: Conculsion and future work

Chapter 6 concludes the thesis and provides potential future research directions based on our work.

Chapter 2

A Survey of Narrative Visualization Including Geo-space

Contents

"Study the past if you would define the future." -Confucius [1]

THis Chapter presents a literature survey of narrative visualization including geo-spatial visualization. Throughout history, storytelling has been an effective way of conveying information and knowledge. In the field of visualization, storytelling is rapidly gaining momentum and evolving cutting-edge techniques that enhance understanding. Many communities have commented on the importance of storytelling in data visualization. Storytellers tend to be integrating complex visualizations into their narratives in growing numbers. In this Chapter, we present a survey of storytelling literature in visualization and present an overview of the common and important elements in storytelling visualization. We also describe the challenges in this field as well as a novel classification of the literature on storytelling in visualization. Our classification scheme highlights the open and unsolved problems in this field as well as the more mature storytelling sub-fields. The benefits offer a concise overview and a starting point into this rapidly evolving research trend and provide a deeper understanding of this topic. This Chapter is based on the paper "Storytelling and Visualization: A Survey" [52] and "Storytelling and Visualization: An Extended Survey" [53].

[1] Confucius (551-479 BC) was a Chinese teacher, editor, politician, and philosopher of the Spring and Autumn period of Chinese history.

2.1 Introduction And Motivation

"We believe in the power of science, exploration, and storytelling to change the world" - Susan Goldberg, Editor in Chief of National Geographic Magazine, from "The Risks of Storytelling", October 2015 [54].

"In a world increasingly saturated with data and information, visualizations are a potent way to break through the clutter, tell your story, and persuade people to action" [55]. -Adam Singer, Clickz.com, "Data Visualization: Your Secret Weapon in Storytelling and Persuasion", October 2014.

Throughout history, storytelling has been an effective way of conveying information and knowledge [12]. In the field of visualization, storytelling is rapidly developing technique that enhance understanding. Many communities have commented on the importance of storytelling in data visualization [18]. Storytellers tend to be integrating complex visualizations into their narratives in growing numbers.

As contributions, we present a survey reviewing storytelling papers in visualization and present an overview of the common and important elements in storytelling visualization. We also describe the challenges in this field and present a novel classification of the literature on storytelling in visualization. Our classification highlights both mature and unsolved problems in this area. The benefit is a concise overview and valuable starting point into this rapidly growing and evolving research trend. Readers will also gain a deeper understanding of this rapidly evolving research direction.

2.1.1 Definition and Storytelling Elements

A story can be defined as "a narration of the events in the life of a person or the existence of a thing, or such events as a subject for narration" [56] or "a series of events that are or might be narrated" [57]. Storytelling is a popular concept that is used in many fields, such as media [18], education [58] and entertainment [59]. Storytelling is a technique used to present dynamic relationships between story nodes through interaction. According to Zipes [58], storytelling can involve animation and self-discovery, incorporating models, ethical principles, canons of literature, and social standards. In education, a storyteller can improve and strengthen the literacy of students. Also, the storyteller can engage audiences so they feel a desire to read, write, act, and draw. Audience members can learn to express themselves critically and imaginatively with techniques they may learn from the storyteller or teacher.

In the context of the visualization literature. Lee et al. [60] argue that "the community has been using the term 'storytelling' in a very broad way without a clear consensus or discussion on what a visual data story encompasses". They state that a visual data story includes a set of story pieces. Most of the story pieces are visualized to support one or more intended messages. Story pieces are presented with a meaningful order or connection between them to support the author's high level communication goal.

Furthermore no agreed definition of "visual data story" has yet emerged in the visualization literature [60]. For a full-length 6 page discussion on this topic, we refer the reader to Lee et al. [60].

For the purpose of this thesis, we define narrative visualization as a visual design that can be used to explain this result of visual exploration and analysis to a wider audience. This usually includes provenance information that can inform users as to how an observation was generated.

2.1.2 Classification of Literature and Challenges in Storytelling and Visualization

Although storytelling has been developing in other fields for years, storytelling is a relatively new subject in visualization. As such, it faces many challenges. In this survey we have extracted the fundamental characteristics of storytelling both as an entity and as a creative process. Our literature classification is based on the logical notions of *who are the main subjects involved in storytelling for visualization* (authoring tools and audience), *how are stories told* (narratives and transitions), *why can we use storytelling for visualization* (memorability and interpretation). From these characteristics we have then developed the following dimensions which are common to storytelling in visualization.

Authoring-Tools: Authorship addresses *who* creates the story and narrative. Authorship commonly refers to the state or fact of being the writer of a book, article, or document or the creator of a work of art [61] and its source or origin [62]. Central to this definition is the writer or author. Rodgers[63] defines an author as "an individual solely responsible for the creation of a unique body of work."

User-engagement: Engagement is about the audience and also concerns *why* we use storytelling. How can we ensure that the message comes across to the audience? Can we measure engagement?

Narratives: Narrative concerns *how* an author tells a story. Narrative structures include events and visualization of characters. Narrative visuals contain the transition between events. This entails, "Using a tool to visually analyze data and to generate visualizations via vector graphics or images for presentation," and then deciding "how to thread the representations into a compelling yet understandable sequence."[20]

Transitions: Transitions are about *how* authors may tell the story. Transitions seamlessly blend events within a story and are key to its flow. Successful transitions vary actions as little as possible to strengthen overall coherence. Transitions in visualization can be either dynamic or static.

Memorability: Memorability addresses *why* authors present data in the form of a story. Memorability is an important goal of storytelling. A good visualization technique draws the viewer's attention and increase a story's memorability [37].

Interpretation: Data interpretation refers to the process of critiquing and determining the significance of important data and information, such as survey results, experimental findings, observations or narrative reports.

When examined in the context of storytelling in visualization each dimension raises interesting questions: Are current storytelling platforms taking into account the role of the author and supporting the authorship process? What forms of narrative structures and visuals best apply to storytelling in visualization? Are static transitions or dynamic transitions more effective for storytelling in visualization? Can visualization increase the memorability of data information or knowledge? Does storytelling and visualization aid with data interpretation? What is the most effective way to engage an audience? Data preparation and enhancement is another challenge for which there is currently no literature. Thus we include it as a future research direction but not in our classification.

Starting from the logical notions of who, how, why, and these open questions we have chosen these dimensions to form the basis of our literature classification on storytelling in visualization. See Table 1. It is important to note that some papers address multiple topics in Table 1 and in our classification. We placed papers by what we determined to be the main focus of the paper. This is very useful for obtaining an overview. However some papers address more than one theme, e.g. authoring tools and narratives.

Table 2.1: Our classification of the storytelling literature. The y-axis categories fall into who-authoring-tools and user-engagement, how-narrative and transitions, why-memorability and interpretation. See section 2.1.2 for a complete description.

		Linear	User-directed/Interactive	Parallel	Random
Who	Authoring-Tools	Gershon et al. 2001[64] Lu and Shen, 2008[7] Cruz et al. 2011 [8]	Wohlfart, 2006 [9] Wohlfart et al. 2007[10] Lidal et al. 2012 [11] Lee et al. 2013[13] Lidal et al. 2013[12] Lundblad et al. 2013[14] Fulda et al. 2016[65] Amini et al. 2017 [66]	Eccles et al. 2007[15] Kuhn et al. 2012[16]	
	User Engagement		Figueiras, 2014 [25] Boy et al 2016 [67] Borkin et al,2016 [38]		
How	Narrative	Hullman et al. 2013 [19] Hullman et al. 2013 [20] Gao et al. 2014 [68] Amini et al. 2015 [69] Bach et al. 2016 [21]	Viegas et al. 2004[22] Hullman et al. 2011[23] Figueiras, 2014 [25] Figueiras, 2014 [24] Nguyen et al, 2014 [26] Satyanarayan et al. 2014 [70] Gratzl et al. 2016 [71]	Akashi et al. 2007[27] Fisher et al. 2008[28] Hullman et al. 2011[23] Bryan et al. 2017[72]	
	Static Transitions		Ferreira et al. 2013[29]	Robertson, 2008[30] Chen et al. 2012[31] Tanhashi et al. 2012[32] Liu et al. 2013[33] Ferreira et al. 2013[29]	
	Animated Transitions	Heer et al. 2007 [34] Liao et al. 2014 [73]	Bederson and Boltman, 1999[35] Akiba et al. 2010[36] Nagel et al. 2016[74]		
Why	Memorability	Bateman et al. 2010[37] Borkin et al, 2016[38]		Saket et al. 2015 [39]	
	Interpretation				

2.1.3 Classification of Literature: the Second Dimension

In addition, the literature is also classified by the ordering or sequence of events, which refers to the traversal the path viewer takes through the visualization. This dimension is adapted from Segal and Heer [18]. It forms our second categorization for Table 1. The classification includes:

Linear: A story sequence path in linear order is prescribed by the author.

User-directed path: The user selects a path among multiple alternatives or creates their own path. This is not pre-defined like it is in the case of linear.

Parallel: Several paths can be traversed or visualized at the same time. This can be linear or user-directed but is always parallel.

Random access or other: There is no prescribed path. There is currently no literature prescribing random order.

Table 2.2: An alternative classification of the storytelling literature based on scientific, information, and geo-spatial visualization. Geo-spatial is separated from scientific visualization because these two topics are historically always separted in the literature. Both mature areas and unsolved problems are apparent.

	Scientific Visualization	Information Visualization	Geo-spatial Visualization
Authoring Tools	Wohlfart, 2006 [9] Wohlfart et al. 2007[10] Lu and Shen, 2008[7]	Gershon et al. 2001[64] Cruz et al. 2011 [8] Kuhn et al. 2012[16] Lee et al. 2013[13] Fulda et al. 2016[65] Amini et al. 2017 [66]	Eccles et al. 2007[15] Lidal et al. 2012 [11] Lidal et al. 2013[12] Lundblad et al. 2013[14]
Narrative		Viegas et al. 2004[22] Akashi et al. 2007[27] Fisher et al. 2008[28] Segel and Heer, 2010[18] Hullman et al. 2011[23] Hullman et al. 2013 [19] Hullman et al. 2013 [20] Figueiras, 2014 [24] Figueiras, 2014 [25] Nguyen et al, 2014 [26] Amini et al. [69] Lee et al. 2015 [60] Bach et al. [21] Bryan et al. 2017[72] Gratzl et al. 2016 [71]	Gao et al. 2014[68] Satyanarayan et al. 2014[70]
Static Transitions		Robertson, 2008[30] Chen et al. 2012[31] Tanhashi et al. 2012[32] Liu et al. 2013[33]	Ferreira et al. 2013[29]
Animated Transitions	Akiba et al. 2010[36] Liao et al. 2014[73]	Bederson and Boltman, 1999[35] Heer et al. 2007 [34]	Nagel et al. 2016[74]
Memorability		Bateman et al. 2010[37] Borkin et al. 2013 [75]	Saket et al. 2015 [39]
Interpretation			
Engagement		Figueiras, 2014 [25] Mahyar et al.,2015 [17] Boy et al 2016 [67] Borkin et al,2016 [38]	

2.1.4 Literature Search Methodology

We search both the IEEE and ACM Digital libraries for the terms "storytelling", "narrative visualization", "memorability", "transitions in visualization", "user-engagement", and various combinations of these phrases. We focus primarily on the IEEE TVCG papers. We check the references of each paper and looked for related literature on storytelling. We also search the visualization publication data collection [76] for these major themes in visualization and storytelling. Google scholar is also used as part of our search methodology.

In summary, our literature search includes:

1. IEEE EXPLORE Digital Library
2. ACM Digital Library
3. Visualization publication data collection [76]
4. the annual EuroVis conference
5. the Eurographics Digital Library

Several other papers were discovered by looking at the related work section of the papers we found.

2.1.5 Survey Scope

The storytelling visualization papers summarized in this survey include the subjects of scientific visualization, information visualization, and geo-spatial visualization. In order to manage the scope of this survey, storytelling papers from other fields are not included, such as:

Virtual reality and augmented reality: For example, Santiago et al. [77] present "mogre-storytelling" as a solution to interactive storytelling. This tool provides different functionalities for creating and the customization of scenarios in 3D, enables the addition of 3D models from the Internet, and enables the creation of a virtual story using multimedia and storytelling elements.

Education: For example, Cropper et al. [78] address the extent of how scientific storytelling benefits our communication skills in the sciences, and the connections they establish with the information itself and others in their circle of influence.

Gaming: Alavesa et al. [79] describe the development of a small scale pervasive game which can take storytelling from camp-fire sites to modern urban environments.

Multi-media and Image Processing: For example, Chu et al. describe a system to transform any temporal image sequence to a comics-based storytelling visualization [80]. Correa and Ma present a narrative system to generate dynamic narrative from videos [81]. Image processing falls outside the scope of this survey. Video processing also falls outside the scope of the survey [69].

Language processing: Theune et al. [82] develop a story generation system. It can create story plots automatically based on the actions of intelligent agents living in a virtual story world. The derived plots are converted to natural language, and presented to the user by an embodied agent that makes use of text-to-speech.

There are other fields that study storytelling as well. In the next sections we describe the literature on storytelling in visualization. Our classification is presented in Table 2.1. An alternative classification is presented in Table 2.2. Figure 2.12 shows the visualization techniques used in storytelling for data visualization literature.

Ma et al. [6] state that a story that is well paced exhibits deliberate control over the rate at which plot points occur. They present a selection of scientific storytelling visualizations from NASA related work and describes various examples. The Scientific Visualization Studio (SVS) at NASA uses storytelling visualization to investigate observational data collected by instruments and sensors and make it more suitable for consumption by the public [83][84].

The science museum presents visualization to the public with complex and abstract geographic phenomena at extreme size scales for explanatory animations. The science museums provide further interpretation through labels, videos, and live demonstrations. See Figure 2.1[6].

Storytelling enables the user to interact with geographic data such as the Earth's climate or the collapse of a star by using a story model, such as story nodes or story transitions[36]. Ma et al. is based on previous scientific visualization work at NASA, based in the scientific research center and scientific museum and describe how visualization can be used to tell a good story, and tell it well. This is a topic that the scientific visualization research community paid little attention to at that time.

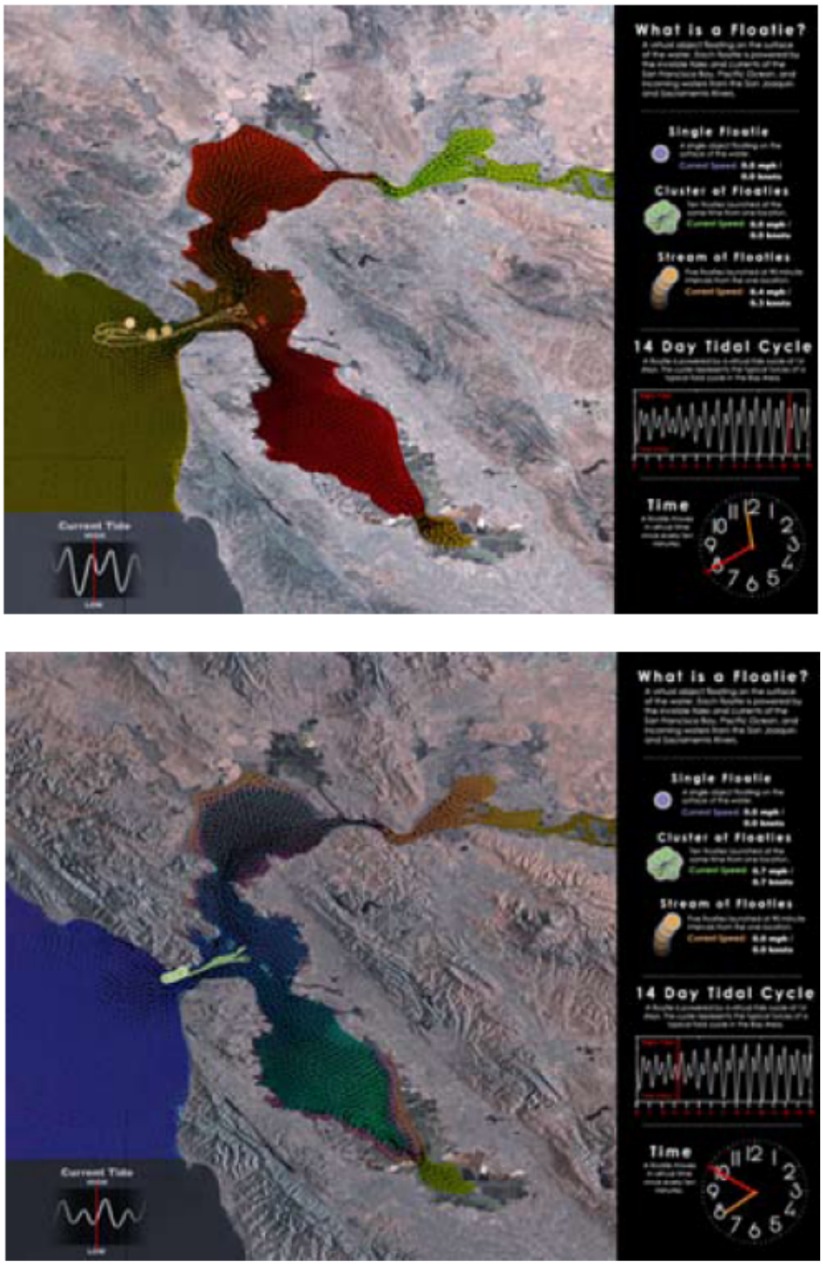


Figure 2.1: *Ma et al. show the interactive software used at the Exploratorium in San Francisco. The purpose of this software is to educate users on the process of how tides, currents and rivers combine in the estuary of San Francisco bay. A touch-screen is used to place floats into the virtual water so that the user can see the effects of the current on the float. Users can watch the effects of predicted tide and river flow cycles on the floats trajectory. Other contextual information is provided as an animation alongside the visualization [6]. Image courtesy of Ma et al. [6].*

2.2 Authoring-tools for storytelling and visualization

Authorship refers to writing or creating a book, article, or document, or the creator of a work of art according to The Oxford English dictionary[61] , especially with reference to an author, creator or producer [62]. For our purposes, we will adopt a definition of author described by Rodgers[63], "An author is best described as an individual solely responsible for the creation of a unique body of work." Hullman [20] et al. state, "Story creation involves sequential processes of context definition, information selection, modality selection, and choosing an order to effectively convey the intended narrative".

Presenting the findings of a qualitative study of undergraduate writers at The City University of New York, Hullman explores student perspectives on models of authorship, the relationships between these models and student experiences of authorship in different writing situations, and proposes the importance of distinguishing between the multiple models and definitions of authorship and the rhetorical contexts associated with each [63]. Rodger develops a qualitative study of 800 students on the definition of authorship and their rhetorical contexts over a one-hour interview. Students defined authors as "[people] who see writing as being beyond a hobby," and as a term that should be applied only to those individuals for whom writing is "something he or she has to do", "a career", or "an act that will lead to something being published."

All papers in this section focus on authoring-tools for storytelling. Wohlfart [9] creates new volume visualization stories for medical applications. Gershon [64] and Cruz [8] present general storytelling for information visualization. Kuhn [16], Lee [13] and Plowman [85] all develop unique creator tools for storytelling visualization.

It is important to note that our survey is not simply a list of papers. Individual papers are summarized according to a special methodology [86]. This process connects related papers together such that the connections and relationship to previously published literature is made clear.

2.2.1 Authoring-tools for Linear Storytelling

The literature in this sub-section focus on visual designs for authoring in a linear style that is prescribed, automatic, or semi-automatic (as opposed to interactive) or decided by the users. In other words, creators are provided with assets to formulate a linear story.

Gershon and Page state that storytelling enables visualization to reveal information as effectively and intuitively as if the viewer were watching a movie [64]. They introduce the concept of storytelling and presents advantages of storytelling.

One example presents a situation in which a number of enemy positions surround a school with children trapped inside as de facto hostages as the crossfire fills the space overhead and both sides move toward confrontation.Gershon and Page is based on previous work of Denning [87] and explain the usage of storytelling in information visualization.

Lu and Shen propose an approach to reduce the number of time steps that users required in order to visualize and understand the essential data features by selecting representative datasets. They design a flexible framework for quantifying data differences using multiple dissimilarity matrices [7]. A new visualization approach that filters data analysis results, which is achieved by measuring the degree of data similarity/difference and selecting important datasets that contain essential data features [7]. See Figure 2.2.

They interactively select representative datasets that include a significant portion of features of scientific data, whose data distribution requires more analysis than time sequence, reduces the amount of data to necessarily visualize and still keeps the essential data information. This can be used to improve the efficiency of time-varying data visualization [7].

An interactive storyboard is used to visualize and explore the overall content of time-varying datasets through composing an appropriate amount of information that can be efficiently understood by users [7].

Lu and Shen [7] is based on the previous work of time-vary visualization [88] and design a general method for comparing data dissimilarities. They do not require a dense sampling frequency to capture the object evolution and their work is not limited to specific feature models, such as geometry or interval volumes, and their attribute designs.

Storytelling, in the context of this article, deals with the core of information visualization by extracting relevant knowledge and enhancing its cognition [8]. Cruz et al. present generative storytelling as a conceptual framework for information storytelling. They create stories from data fabulas using computer graphics as a narrative medium. Data fabulas are a set of time-ordered events caused or experienced by actors [8].

A story is formed by characters. It involves the representation of the fabula's actors and the definition of a temporal structure. The engine transforming a fabula into story consists of two models. The event model creates a story timeline and an action model creates a set of actors

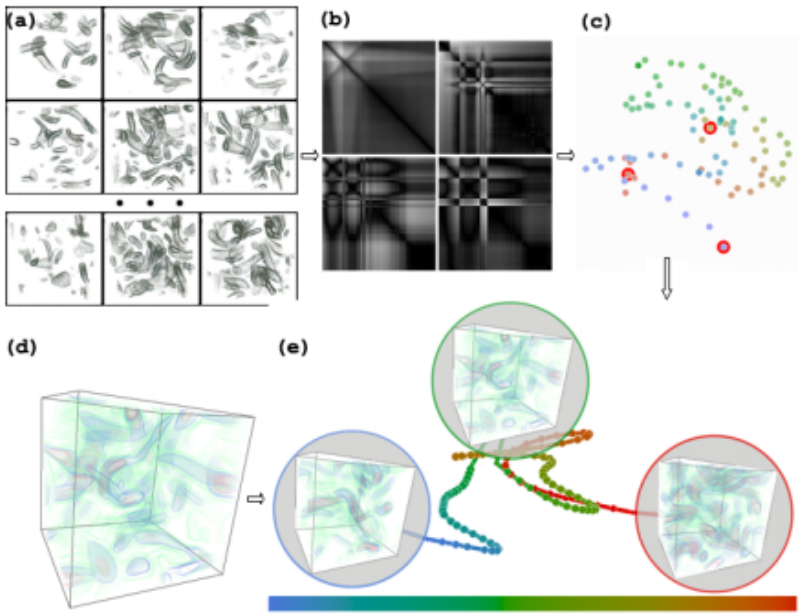


Figure 2.2: *This figure shows the system architecture from Lu and Shen. It integrates the information of data analysis and a single 3D data visualization method for users to explore and visualize overall time-varying data contents [7]. Image courtesy of Lu and shen [7].*

behaviours. For example an empire's decline visualizing western empire's decline in the 19th and 20th centuries. See Figure 2.3.

Cruz et al. is based on previous work of narrative theory[89] and presents generative storytelling as a conceptual framework for information storytelling.

2.2.2 Authoring-tools for User-directed and Interactive Storytelling

A large body of research has been carried out for authors wishing to create their own user-oriented or interactive stories. This literature focuses on interactive, user-driven authorship (as opposed to automatic or semi-automatic authorship). Storytelling is a relatively new form of interactive volume visualization presentation [9]. Wohlfart explores the usefulness of storytelling in the context of volume visualization. He presents a story telling model and divides the concept of volumetric storytelling into story authoring and storytelling constituents. He presents a volumetric storytelling prototype application, which is based on the RTVR (real time volume redering) Java library [90] for interactive volume rendering. See Figure 2.4. The storytelling model contains a range of hierarchy levels, in top-down order, which are: story

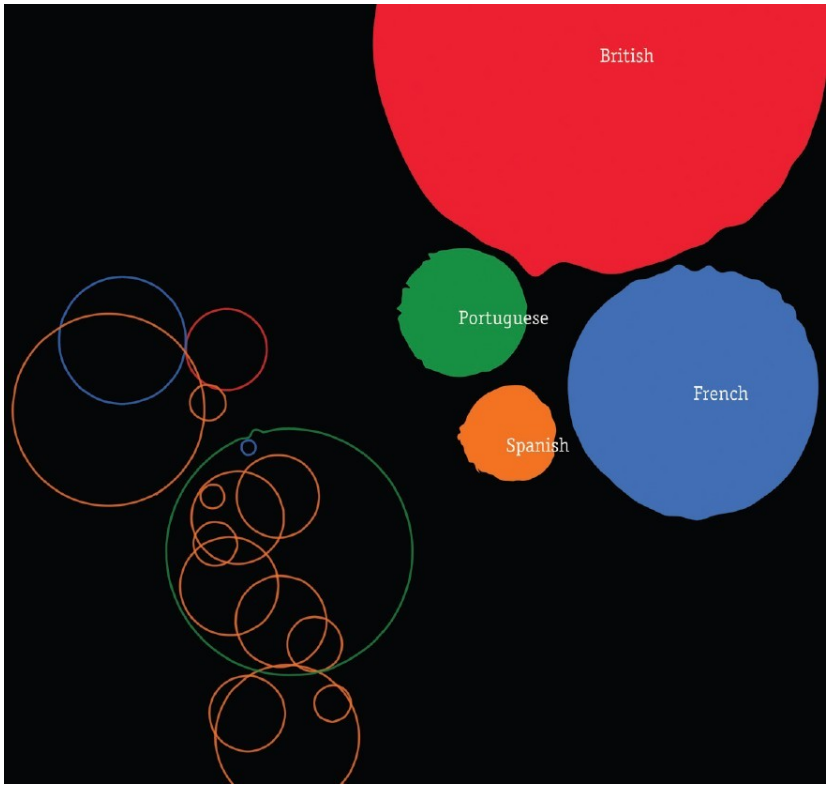


Figure 2.3: *Cruz et al. show the British hegemony and the newly independent South America in 1891. Each empire and independent territory is a circle whose area is proportional to that entity's land area. Former colonies are unfilled circles with rims in the corresponding empire's color [8]. Image courtesy of Cruz et al. [8].*

node, story transitions, story action group, story action atoms. See Figure 2.4. The story nodes form the corner marks of the story and store the state of the whole scene. Story nodes are connected by story transitions, each consisting of one or multiple story action groups. Each story action group stores the scene changes relative to its preceding action group (or story node) [9, 10].

The story authoring process contains two steps: a story recording process and a story editing process. The outcome of this recording process is a raw prototype of a story told through volume visualization. In the story editing step, this raw story is refined until the final story outline is reached [9, 10].

This process presents a volume visualization following the storytelling model. And the key feature is interaction, including viewing interaction, representing interaction and data interac-

tion [9, 10]. Wohlfart and Hauser also discuss the paradoxical integration of storytelling and interaction [10].

Figure 2.4 shows an image sequence taken from a sample linear volumetric story visualized with their prototype. The distinct story nodes refer to the key events in the story, which provide an overview first, then details on specific features in the dataset, and at the end a conclusion made by the story author. The necessary story transitions are represented as orange arrows from one story node to the next and are animated in the prototype application. The story consumer may take over some story parameters (e.g. camera angle) already during playback or at the end of the story to further investigate the dataset [9, 10].

This story guides the observers through the visualization, puts the contained visual representations into context with each other and finally introduces them to important features in the data [10]. See Figure 2.5. Wohlfart is based on previous work of volume visualization [91, 92, 93] and interactive visualization [94] and combine these concepts together to develop a storytelling model for volume visualization.

Geological storytelling is a novel graphical approach for capturing and visualizing the reasoning process that leads to a geological model [11] [12]. Lidal et al. present a sketch-based interface for rapid modelling and exploration of various geological scenarios. The authors present a concept that handles sketching processed over time and a novel approach for externalizing the mental reasoning process. The process can be presented and evaluated [11][12]. The geological storytelling model contains three main parts. See Figure 2.6.

The canvas is a sketch-based interface where the geologist can draw the geological story on a 2D seismic slice backdrop, utilizing a pen and paper interaction style. The StoryTree is a tree graph representation of all the geological stories, each with its own subtree of story nodes. Individual story nodes can be selected for editing in the canvas. One or more complete story trees can be selected for playback or comparative visualization in the InspectView. The InspectView serves two purposes. First, it is a view where a story can be played and evaluated. In addition, multiple stories can be played synchronized for a side-by-side visual comparison.

Lidal et al. is based on a previous storytelling model [9] for scientific visualization [6] and develops a storytelling model for geological visualization.

Lee et al. present SketchStory, a data-enabled digital white board to support real-time storytelling. It enables the presenter to stay focused on a story and interact with charts created during presentation. See Figure 2.7 [13].

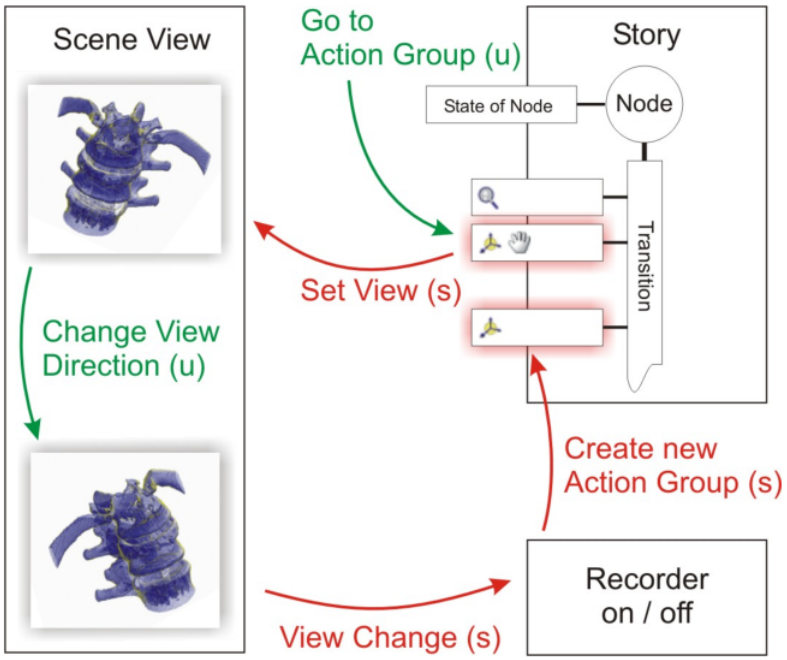


Figure 2.4: *The proposed method to author a story is to record the user's natural interaction with the visualization software. This image shows the process of the story creation by Wohlfart. Green annotations represent user interaction and red annotations refer to internal system processes. As soon as the software starts recording, a new story is created and all interactions are logged [9]. Image courtesy of Wohlfart [9].*

The data-based story is recorded in SketchStory as a sequence of charts in XML files. The charts are linked with specific sketch gestures. The presenter draws an example icon and then draws a sketch gesture for chart invocation. Sketchstory recognizes the gesture and creates the corresponding chart. Lee et al. is based on previous work for storytelling of information visualization [64, 18] and sketch-based interaction [95], and develops the SketchStory system to enhance storytelling in a presentation.

Storytelling is one of the most impactful ways to teach, learn, and persuade [14]. Lundblad and Jern present geovisual analytics software with integrated storytelling. It can be applied to large spatial-temporal and multivariate data through dynamic visual user interfaces.

Using a scatter plot matrix gives the analyst a good overview of all correlations between the selected indicators. The analyst can use the scatter plot matrix as an overview and then steer the scatter plot for interesting detailed combinations over time. See Figure 2.8[14]. The distribution plot presents a special visualization technique that displays the variation within

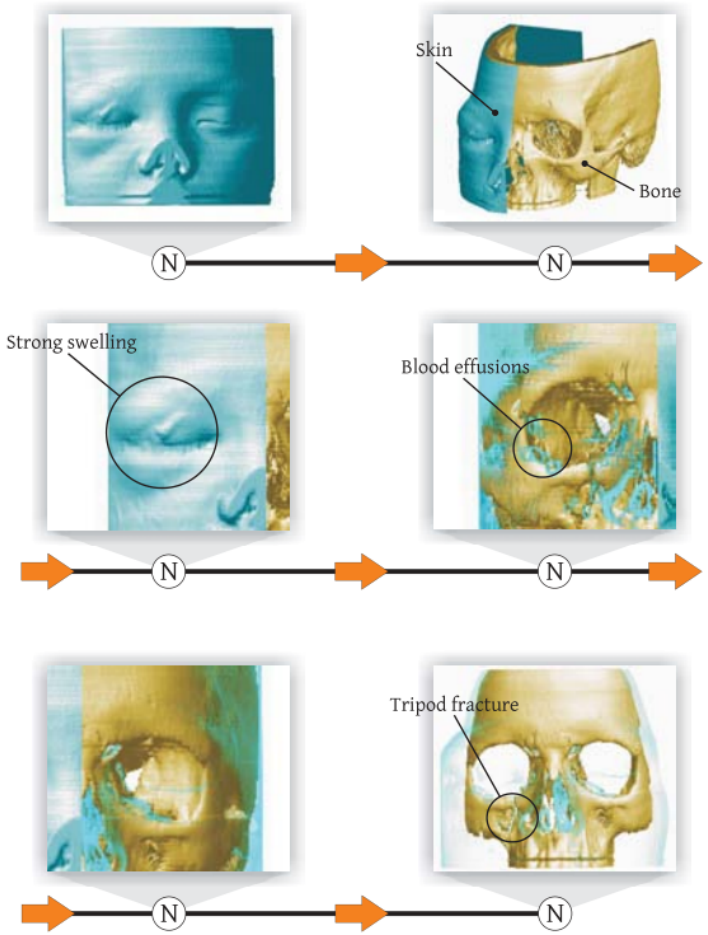


Figure 2.5: *The top two images show an overview of the CT scan data presented by Wohlfart and Hauser. A partial clipping reveals both the skin layer and bone layer, but shows the full set of data. The middle shows a zoomed view that isolates eye swelling in the image (left), and a filtered view that exposes some blood effusions in the swollen region. The bottom offers a comparison of the non-injured eye with the injured one and shows the cause of the swelling which is attributed to a tripod fracture just below the eye. This design offers the user a macro overview as to lay the foundations of a story background then narrows the scope to view the focal point of the image [10]. Image courtesy of Wohlfart [10].*

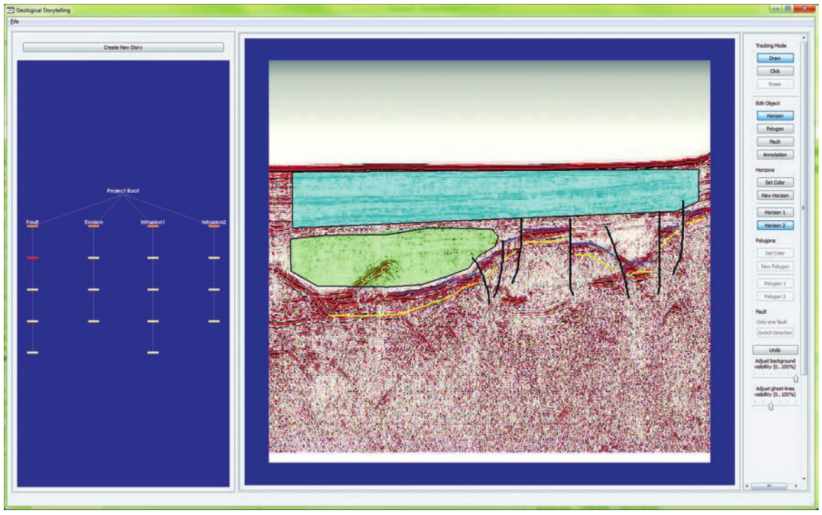


Figure 2.6: *Lidal et al. [11] [12] present a sketch-based interface for rapid modelling and exploration of various geological scenarios.The sketch-based interface is split into two windows. The Story Tree (left) which shows a tree graph representation of all the geological stories, and the Canvas (right) which shows the sketching interface which utilises a pen and paper interaction to record geological sedimentary data. A geological story is built using horizontal lines to separate different geological layers, vertical lines to show fault systems, and polygons for highlighting large sedimentary layers. The user can navigate through different geological stories with the story tree and then inspect the geological elements of that story. Image courtesy of Lidal et al. [11] .*

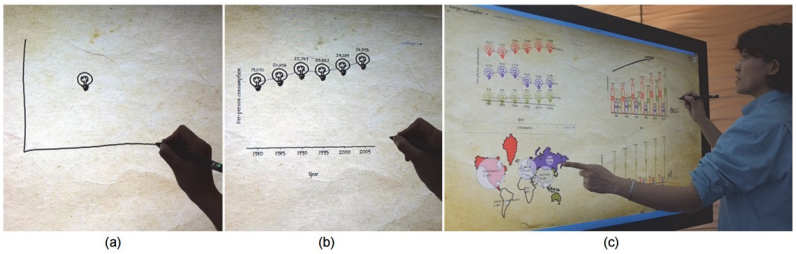


Figure 2.7: *Lee et al. show an example of SketchStory in information visualization presentation [13]. Image courtesy of Lee et al. [13].*

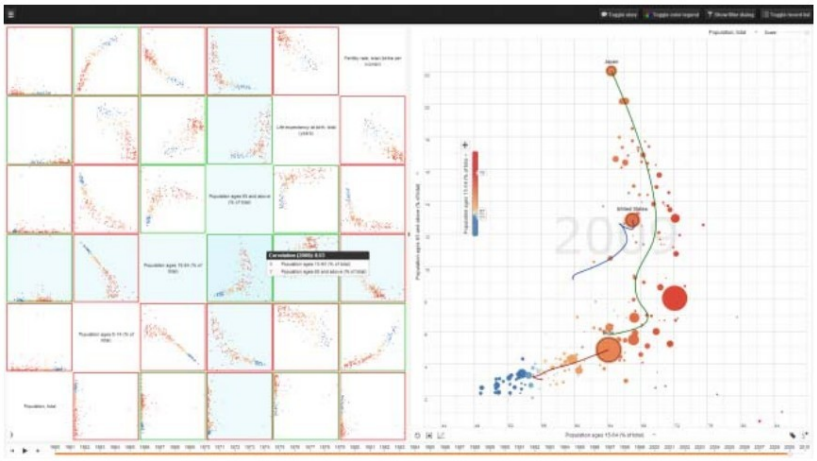


Figure 2.8: *Lundblad and Jern show Vislet aimed at a comparative visualization using linked Scatter Matrix and Scatter Plot to analyze national correlation between 6 indicators between 1960 and 2010 from the World Databank [14]. Image courtesy of Lundblad and Jern [14].*

individual European countries [14]. The Motala River map is visualized for different stories divided into different layers, such as a glyph layer, stream layer, polygon layer and background map layer. It shows the local and total water flow, and water path from source to ocean [14].

Lundblad and Jern is based on the previous work of the storytelling concept [64] and work of web-based geovisual tools, integrates storytelling with geovisual analytics software.

2.2.3 Authoring-tools for Parallel Storytelling

In this category of literature, authors create stories in parallel. In other words there may be multiple authors working in parallel i.e. simultaneously for the final outcome. This is opposed to a single author as in the previous subsection.

GeoTime events are recorded in *x, y, t*, coordinate space. This is used in observation analysis and can make a major contribution to a storytelling model. Eccles et al. presents the GeoTime stories prototype that combines a geo-spatial map with narrative events to produce a story framework. See Figure 2.9 [15]. This system provides functions for simple pattern detection in simultaneous movement activity. These functions look at possible interactions between people within the narrative, the speed at which they travel, and the type of location that they visit. Narrative text authoring enables the analyst to create and present stories found within the data. The story window displays this data as well as discovered patterns. The system

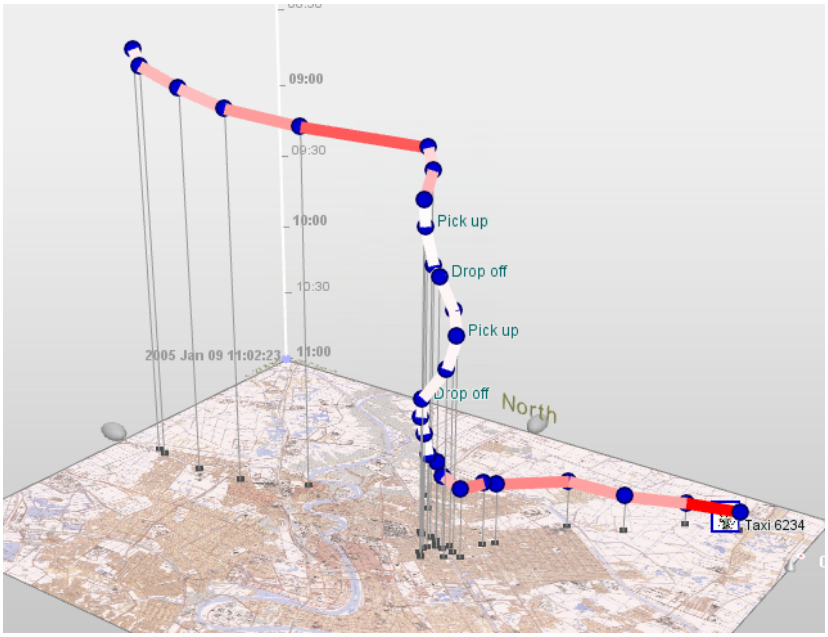


Figure 2.9: *Eccles et al. show a GeoTime visualisation instance. The L axis represented by height is temporal. X and Y axis represent the geospatial location Here you can see a taxi driver's route over the course of a few hours. Each pick up and drop off is labelled and the route is mapped on the X and Y axis using the map [15]. Image courtesy of Eccles et al. [15].*

enables multiple stories to connect together if they follow a linear flow. Also simultaneous narratives can be shown in a single image for a direct comparison.

This system uses a similar approach to Sense.us [96]. Instead of using a blog-type discussion workflow for adding text, Geotime is designed for authoring a single story and annotations are integrated into the data itself.

The CodeTimeline visualization by Kuhn and Stocker [16] enables developers who are new to a team to understand the history of the system they are working on. Designed to show a development team's tribal memory, the software offers a partial replacement for exhaustive documentation. See Figure 2.10.

A collaboration view presents visualizes code ownership and historical patterns in collaboration. A sourcecloud flow view presents a word cloud of added and removed vocabulary between software releases. Lifetime events can be added by users as a frame of reference in each of the visualizations. This method of linking also enables new users to learn more about the history of the software development. These events can include anything from email threads to pictures of the team during work.

Prior to Kuhn and Stocker, Ogawa [97, 98] presents "software evolution storylines" and "Code Swarm", which focus on the interactions between developers on projects but do not focus on telling a story about the software history. Codebook, a concept presented by Begel et al[99], outlines a social network that connects software engineers with their shared code base. It encourages interaction with their code and others, enabling a broader understanding of the project they share with other developers.

2.3 User Engagement

The literature in this category addresses an important but less developed research topic, namely user engagement. In other words, who do we engage with storytelling and how can we engage an audience?

Mahyar et al. [17] address how prior research in different domains define and measure user engagement. They discuss existing frameworks for engagement from other related fields and propose a taxonomy based on previous frameworks for information visualization.

They present five levels of user engagement in information visualization. See Figure 2.11.

1. Expose (Viewing): the user understands how to read and interact with the data.
2. Involve (Interacting): the user interacts with the visualization and manipulates the data.
3. Analyze (Finding trend): the user analyze the data, finds trends, and outliers.
4. Synthesize (Testing Hypotheses): the user is able to form and evaluate hypotheses.
5. Decide (Deriving Decisions): the user is able to make decisions and draw conclusions based on evaluations of different hypotheses.

2.3.1 User Engagement for User-directed visualization

The literature in this subsection focuses on interactive, user-driven visualization for user engagement. Engagement specifically focuses on each user's investment in the exploration of a visualization [67]. Boy et al. use low-level user interaction e.g. the number of interactions with a visualization that impact the display to quantify user engagement. They present the results of

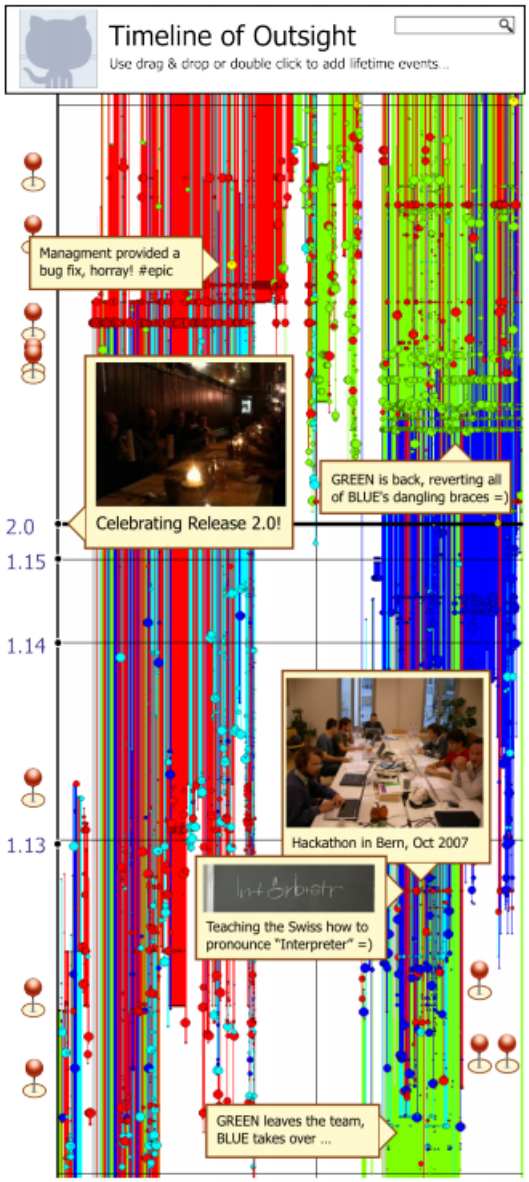


Figure 2.10: *Kuhn and Stocker show the CodeTimeline collaboration view. Colors denote different user contributions and each line represents the life of files in the code. Sticky notes are added so the users can learn the history of the code beyond the file evolution [16]. Image courtesy of Kuhn and Stocker [16].*

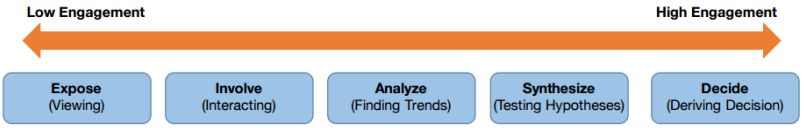


Figure 2.11: *Mahyar et al. present five levels of user engagement in information visualization [17].Image courtesy of Mahyar et al. [17].*

three web-based field experiments, and evaluate the impact of using initial narrative visualization techniques and storytelling on user-engagement with exploratory information visualizations. The main contribution of their work include: the design of three web-based experiments on user-engagement information visualizations. They hypothesize narrative elements should effectively engage the user in exploration of data and analysis the result.

Boy et al is based on previous work on narrative visualization [23] and user-centred metrics [100]. The negative outcome of their study clearly indicates that more future work is needed to investigate whether or not storytelling increases user engagement.

2.4 Narrative Visualization and Storytelling

Narrative structures include events and visualization of characters. An example narrative can be a simple interface that presents trends in keywords over time [28]. Narrative visuals contain the transition between events. It involves "using a tool to visually analyze data and to generate visualizations via vector graphics or images for presentation" to decide "how to thread the representations into a compelling yet understandable sequence"[20]. Plowman et al [85, 15] report that a narrative specifically refers to the macro-structure of a document in contrast to the term story which refers to both structure and content. This structuring of evidence, combined with the choice of appropriate rhetorical strategies, is referred to as "the art of storytelling" among literary scholars [85]. Research in narrative visualization points to visualization features that afford storytelling including guided emphasis and structures for reader-driven storytelling. It also includes the principles that govern effective structuring of transitions between consecutive visualizations in narrative presentations, and how different tactics for sequencing visualizations are combined into global strategies in formats like slideshow presentations. We separate transitions into their own section, section 5 for static transitions and section 6 for animated transitions, because of their importance.

A narrative can be seen as a macro-structure which creates global coherence, contributes to local coherence and aids recall through its network of causal links and signposting [85]. The focus of Plowman et al 's research is how students make sense of their learning with multimedia by constructing their own narratives in conjunction with the narrative guidance [85]. The design elements presented by the software constitute narrative guidance and can be a combination of features specific to interactive media, such as the need for clear navigational procedures, with features associated with traditional media, such as recognizable narrative and a clear relationship between tasks and the macro-narrative.

All papers in this section develop methods or structure on how to improve narrative storytelling visualization. Viegas et al. [22] present methods for improving data memorability. Fisher et al. [28] present ways for tracking narrative events over time. Segal and Heer [18] investigate the design of narrative visualizations and identify techniques for telling stories with data. Hullman et al. [23, 19, 20] design the structure of a visualization to present storytelling. Figueiras [24, 25] studies how to incorporate narrative elements as storytelling elements. Again, these papers may cover more than one topic in Table 1. The borders between categories are not 100% black & white. We place papers in the category reflective their main focus.

An overview of the visualization methods used in storytelling for visualization can be found in Figure 2.12. We include it in the section on narrative visualization since this is where the most research has been done. We can observe that most of the visualization designs used are familiar, such as color-coding, line chart, map and bar chart.

2.4.1 Narratives Visualization Summary

Segel and Heer state that storytelling is revealing stories with data and using visualization to function in place of written story [18]. The Oxford English Dictionary defines a narrative as "an account of a series of events, facts, etc., given in order and with the establishing of connections between them" [101]. Heer et al. investigate the design of narrative visualizations and identify techniques for telling stories with data graphics and challenges with the salient dimension of visual storytelling. They describe seven genres of narrative visualization: magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show, and video. See Figure 2.13. They also discuss directions for future reader-centric research [102].

In the New York Times visualization on steroid usage in sports, one larger image and

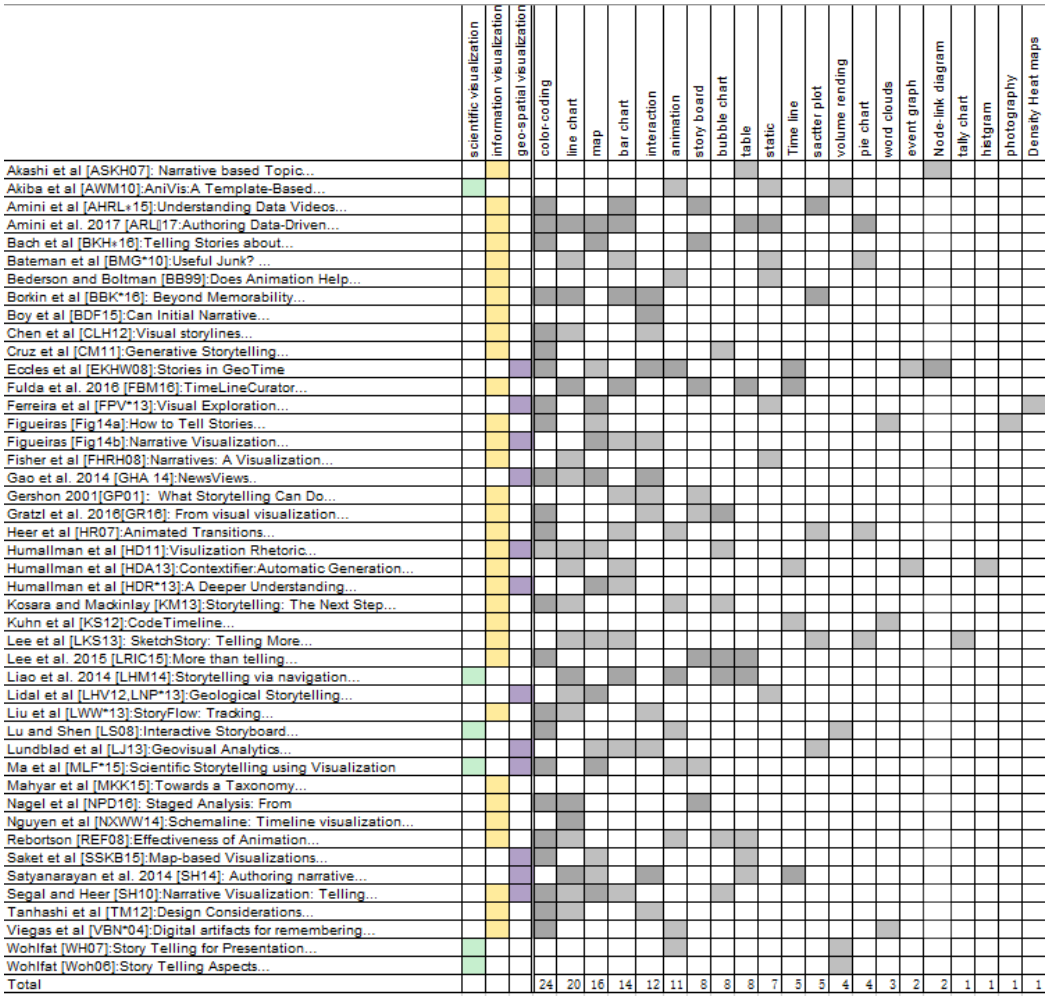


Figure 2.12: *A table summarizing the visualization techniques used in each storytelling paper. The papers are sorted alphabetically by the first author's surname.*

line chart are combined with small images, line charts, and bar charts to illustrate the usage of steroids status over 30 years. The visualization incorporates visual highlighting and connecting elements leading viewing order [103]. The year is mapped to the x axis, the amount of steroids is mapped to the y axis, and different colors represent different players.

In the New York Times visualization on budget forecast, a progress bar is used to describe the accuracy of past White House budgets predictions [104]. The time is mapped to x-axis, and budget situation is mapped to the y-axis.

The Afghanistan nation-building development project example is a interactive geographic visualization with details on-demand sliders that present the status of Afghanistan nation-

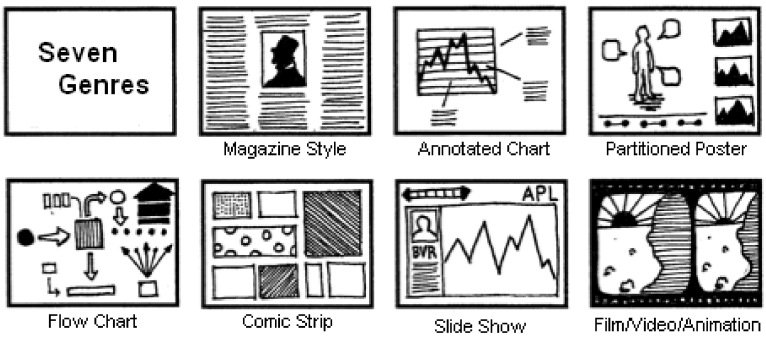


Figure 2.13: *The figure shows the seven genres of narrative visualization presented by Segal and Heer[18]. These vary in terms of the number of frames and the ordering of their visual elements. A video, for example has a strict ordering and high frame number, whereas a 'Magazine Style' poster may have a few frames in one image that are not strictly ordered. These genre elements dictate if a story is author-driven or reader-driven. Author-driven content uses a linear ordering of scenes and has no interactivity. Reader-driven content has no prescribed order to scenes and a high level of interactivity with the reader [18]. Image courtesy of Segal and Heer [18].*

building development projects [105]. Opium cultivation is mapped to the color, and countries are shown on the map. Time can be changed from 2005 to 2009 by dragging the control bar.

The Gapminder visualization uses animated bubble charts to show possible detrimental effects on a person's ability to follow trends [106]. Continent is mapped to color, region is mapped to each bubble, and size is mapped to bubble size, and position is mapped to average yearly income.

The Minnesota Employment Explorer shows how mouse-hover provides details-on-demand, double-clicking an industry triggers a drill-down into that sector while an animated transition updates the display to show sub-industry trends [107]. Color represents different industries, the x-axis represents the time, and the y-axis represents employment.

Segel and Heer is based on previous work of narrative structure, visual narratives, and storytelling with data visualization [102] and observes the storytelling potential of data visualization and drawn parallels to more traditional media. This paper identifies salient design dimensions, clarifies how narrative visualization differs from other storytelling forms and how these differences introduce both opportunities and pitfalls for its narrative potential.

2.4.2 Narrative Visualization for Linear Storytelling

The literature in this sub-section focus on narrative visualization using linear automatic or semi-automatic approaches (as opposed to interactive approaches). The research here involves tools and techniques with an emphasis on how stories are created.

Hullman et al. describe a system called contextifier, which automatically produces custom, annotated visualizations from a given article [19]. The system architecture contains four main sections. A news corpus consists of a large set of news articles. A query generator identifies the most-relevant company in the article. An annotation selection engine integrates selected features into an annotation. And the graph generator generates line graphs using annotations and series. The flow of information can be seen in Figure 2.14 [19].

Hullman et al. is based on previous work in storytelling in visualization [18] and Kandogan's automatic annotation analytics [108]. It develops a system that can automatically generate custom, annotated visualization from a news article of company. Hullman's work places more emphasis on providing background information or perspective on the data than Kandogan's [108].

Hullman et al. [20] outline how automatic sequencing (the order in which to present visualizations) can be approached in designing systems to help non-designers navigate structuring decisions in creating narrative visualizations. Their focus is on how linear, slideshow-style presentation can be optimized using knowledge of sequencing styles and strategies by incorporation.

Hullman et al. argue that analysts using narrated data presentations could be aided by tools for identifying effective sequences for visualizations. They conduct a qualitative analysis of the structural aspects of 42 examples of explicitly-guided professional narrative visualizations. One example is shown in Figure 2.14. They propose a graph-driven approach for finding effective sequences for narrative visualizations informed by their analysis, including defining data attributes for transitions, labelling, and maintaining consistency. The result suggests a need for more sophisticated global constraints than simply summing local transition costs to determine the best path through a graph of weighted visualization transitions. This paper is based on previous work of narrative sequencing[109] and narrative visualization[23, 18], and demonstrates that narrative sequencing can be systematically approached in visualization systems.

Amini et al [69] identify the high-level narrative structures found in professionally created data videos and identify their key components. They derive broader implications for the de-

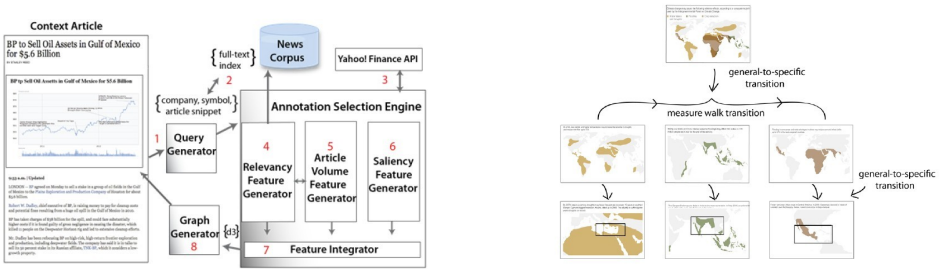


Figure 2.14: *Hullman et al. show the architecture of contextifier [19](left) and illustrate Parallelism in sequencing in the NYT Copenhagen[20](right). Image courtesy of Hullman et al. [19, 20].*

sign of an authoring tool to enable a wide audience to create data videos. Amini et al perform two studies to enhance understanding data videos. They conduct a qualitative analysis of 50 data videos from 8 reputable online sources, and observe that data video categories are also hierarchical and can be further decomposed into units: sequences that put forward different points contributing to a single category. They design a series of workshops to observe how professional storytellers create data video storyboards. They observe the creation process is non-linear and iterative. Amini et al is based on previous work on storytelling [64] and storytelling in information visualization [19].

Bach et al. [21] develop graph comics for data-driven storytelling to present and explain temporal changes in networks to an audience. Bach et al. present six steps to guide graph comics design. See Figure 2.15.

They first collect diagrams, comic literature, and pictures within comics to understand traditional comics structure. The second step is to find possible visual encodings that can represent graph objects, their properties, and the possible changes which they may undergo. They design principles that define when certain visual marks and their attributes can be used and when not. They exploit their design challenges and the structural principles to create comics. They contact two domain experts to collect external feedback and present a qualitative study to check if graphics comics are readable by a wider audience. Bach et al. is based on previous work on network exploration [110] and data-driven storytelling [64].

2.4.3 Narrative Visualization for User-Directed and Interactive Storytelling

The literature in this subsection focuses on interactive, user-driven narrative visualization (as opposed to automatic or semi-automatic). In other words, the papers focus on techniques that

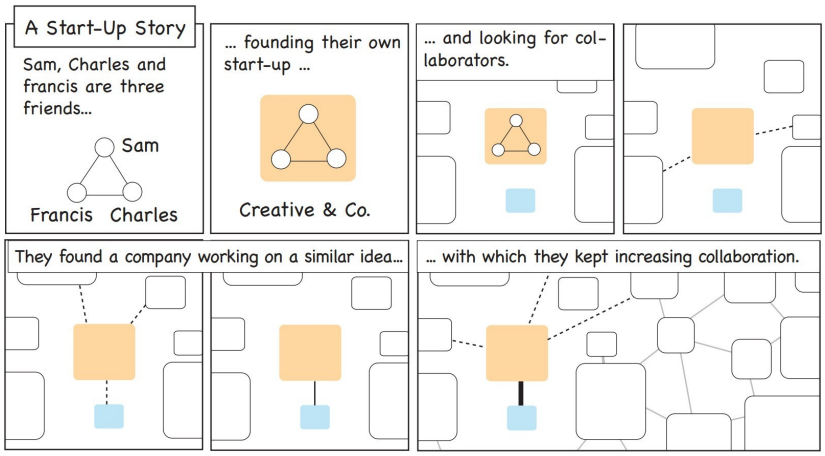


Figure 2.15: *Bach et al. present graph comics for data-driven storytelling [21]. Image courtesy of Bach et al. [21].*

enable users to create narratives interactively. Viegas et al. summarize two methods of visualizing email archives with the aim of improving memorability of the data. Both focus on the higher level patterns of the user′s email habits. The original goal was for these visualizations to uncover social patterns in the archive, but the resulting visualizations caused the user to be more reflective of the data as opposed to analytic. They look at data points and want to recall the story behind it, even share the visualization with friends. See Figure 2.16 [22].

For visualizing email activity, the two axes stand for time, and the dyadic relationship between user account holder and each human interaction. Pattern recognition includes interaction frequency, interaction rhythm, interaction balance, and archive size. The visualization interface includes two main panels; the calendar panel, showing email intensity, and the contacts panel, showing the names of the people being interacted with. When the user clicks on a day square in the calendar panel, the contact panel highlights the names of the people communicated with that day. A name can be clicked on in the contacts panel and each day where that person had corresponded will be highlighted in the calendar panel. The contact panel can be viewed as an animation transitioning through the year of data. The email header data is used to derive the social context of the communication. Five different relationship types are classified. This can be either directly between correspondents or through mutual recipients in group emails. The Social Network visualization looks at each message and evaluates the role of the user (through the email address used i.e. work, school or personal) and makes connections regarding the interaction accordingly. This data is visualized as an animation that evolves over time. Each

Figure 2.16: *Viegas et al. show the PostHistory visualisation. On the left is the calendar view, showing 365 squares to represent each day of the year (This image only shows data up until May). Size corresponds to the volume of email sent on that day. The colour highlights a specific recipient that has been selected in the contact panel (left). The contact panel shows all the contacts the user has been corresponding with over the year. A contact can be selected to highlight their interaction in the calendar view [22]. Image courtesy of Viegas et al. [22].*

second represents one day in the archive. A clustered word cloud is used to display the data. Previous visualizations of online social interaction data have been focused on unravelling the data from the researchers' perspective, whereas these visualizations are for the benefit of the user [111, 112].

Hullman and Diakopoulos state that narrative information visualizations are a style of visualization that often explores the interplay between aspects of both exploratory and communicative visualization [23]. This work contributes to information visualization design and theory by providing insight into the types and forms of given rhetorical techniques in narrative visualizations, and the interaction between those techniques and individual and community characteristics of end users. The authors study how rhetorical techniques are used in visualization. They then investigate the resulting effects of these techniques on user interpretation [23]. The authors collect 51 professional narrative visualizations e.g. from the New York Times and

BBC. Each visualization is "coded" using theory form semitics, statistics, decision theory, and media and communication studies. The visualizations are categorized according to a selection of rhetorics information access, provenance rhetoric, mapping rhetoric, procedural rhetoric, and linguistic rhetoric. Their work provides a taxonomy of specific information presentation manipulations used in narrative visualization. See Figure 2.17.

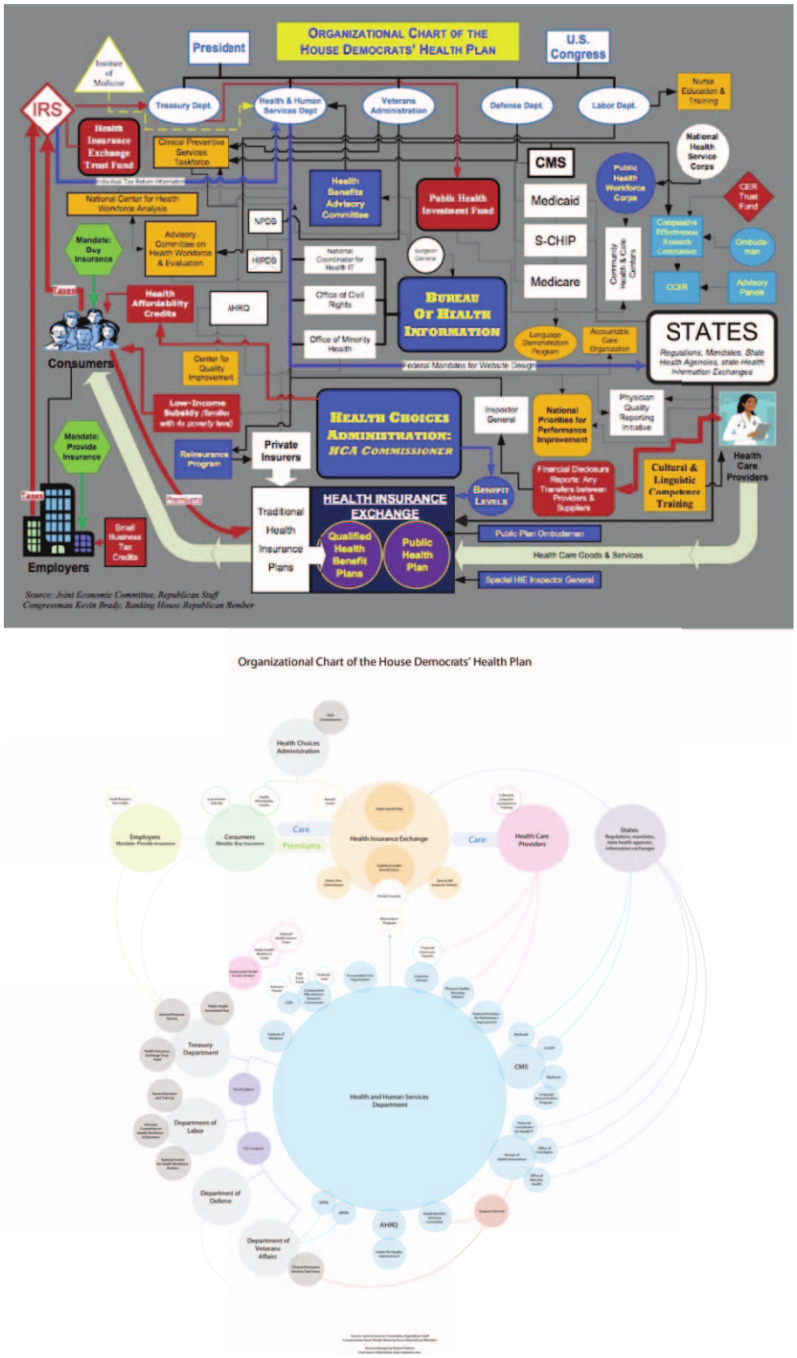
In the mapping America visualization example, The United Stated Census represents a nation wide attempt to provide an objective view of the demographic of the country. Ethnic group is mapped to color, samples are shown on a map and a single ellipse represents 200 people [113]. The poll visualization summarizes the accuracy of political poll predictions from several years and polling agencies in a small multiplies presentation of vertical line graph [114]. Colored bars representing the political parties are drawn to connect data points positioned on the y-axis according to the amount of time prior to the election and on the x-axis according to whether the predictions fell over (to the right) or under (to the left) a centred vertical line representing complete accuracy (or error of zero).

Hullman and Diakopoulos is based on the previous work of Segel and Heer [102] which makes an initial step towards highlighting how varying degrees of authorial intention and user interaction are achieved by general design components in narrative visualization. This work examines the design and end-user interpretation of narrative visualizations in order to deepen understanding of how common design techniques represent rhetorical strategies that make certain interpretations more probable.

A visualization with a narrative is set apart from a visualization without through both its structure and its content. A narrative-based visualization attempts to create a natural flow whereby the data has an obvious progression and therefore permits easier understanding and memorability [24].

Figueiras takes professionally produced visualizations as case studies to analyze how to incorporate narrative elements as storytelling elements. By presenting prototypes of storytelling in selected case studies, Figueiras presents a design study and model for narrative visualization by using storytelling techniques [24].

In the "How many households are like yours" example, users can choose the primary residents and secondary members of a household, then get the number and percentage of households. Figueiras [24] introduces short stories describing different kinds of families instead of having only one article about types of families. This technique engages the user with a focus



Organizational Chart of the House Democrats' Health Plan

Figure 2.17: *Hullman and Diakopoulos demonstrate how data can be window dressed to change the viewers opinion of it. These two images visualize the same data but each illustrator has different intended outcomes. The top image shows an unstructured, complicated graph of conflicting colors and shapes, clearly intended to confuse and obstruct the data, whereas the bottom lays the data out in a simple fashion using consistent shapes and colors [23]. Image courtesy of Hullman and Diakopoulos [23].*

on creating empathy.

"What does China censor online?" example is a tag cloud that only has a title and text shaped on a map of China. Figueiras [24] introduce a tooltips pop up when a user clicks on one region, which provides more detailed information. See Figure 2.18. Tooltips provide additional context in the form of text which help explain the possible reasons for censorship.

The "Death Penalty Statics, Country by Country" figure is a static visualization with different size of bubbles representing the number of death sentence rulings. Figueiras [24] designs an interaction such that when a user chooses a year, a graph displays the number of death sentences handed out that year, which provides extra temporal information and a redesign into a story.

The following Narrative Strategies are described:

1. Context: Providing context to a visualization enables the user to make sense of the data using additional information. Without a sufficient amount of context, less meaning can be derived from the data, whereas the addition of context gives the user more information to explore the data and begin to understand features found within it. This is made easier by the development of interactive visualizations and the ability for users to choose what layers of information they see.

2. Empathy: Although not often associated with information visualization, it has been found that emotive/empathetic visualizations are more memorable and more enjoyable for the user [115].

3. Time Narrative: Utilizing the temporal nature of data in visualization allows users to mentally map the data by adding a sense of story flow. This improves user memorability and aids in the understanding of the data [115].

Figueiras is based on previous work of storytelling [23][6][18] and narrative visualization [28], and develops a model to add storytelling in narrative visualization [24].

Storytelling aims to simplify concepts, create emotional connection, and provides capacity to help retain information [25]. Figueiras presents the results of a focus group study on collecting information on narrative elements. She then suggests strategies for storytelling in visualization [25].

Sixteen participants are asked to study 11 information visualizations of different types and different characteristics (interactive, non-interactive, introductory text, accompanying article,

Figure 2.18: *Figueiras shows a visualization of Chinese online censorship enhanced with storytelling. An interactive feature is added so that the user can click on an instance of censorship to learn more about it. This supplies context to the user and also may draw an empathetic response from the user [24]. Image courtesy of Figueiras [24].*

and audio narration). Then they are asked to rate visualizations in terms of comprehension, navigation, and likability, See Figure 2.19. The participants give high scores to all visualizations, particularly to interactive visualizations. The study suggests that a good storytelling visualization is well-structured and interactive with audience preferences. The results of the user study suggest that interactivity, the option of drilling-down, context, and a sense of relatability and importance for users to feel engaged.

Figueiras is based on previous work of narrative visualization [18], and storytelling visualization [115, 6]. The author uses a focus group to examine storytelling effects in information visualization and storytelling visualization.

Nguyen et al. [26] develop a new timeline visualization, SchemaLine, to gather, represent, and analyze information. They then use a preliminary study to evaluate its effectiveness. See Fig 2.20.

The system contribution includes: a visual design for an interactive timeline that groups notes into schema determined by the analyst; an algorithm to automatically generate a compact and aesthetically pleasing visualization of these schema on the timeline; and a set of fluid interactions with the timeline to support the sensemaking activities defined in the Data-Frame

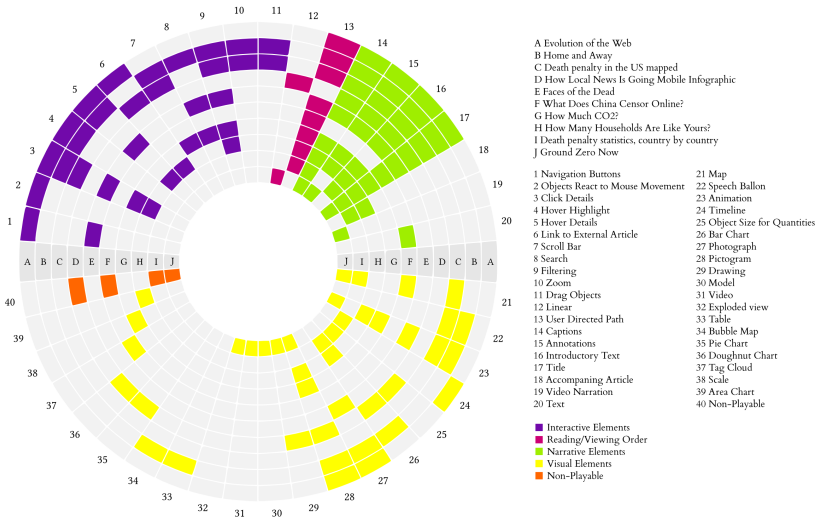


Figure 2.19: *Figueiras shows the visualizations used in the focus group study and the elements that compose them [25].Image courtesy of Figueiras et al. [25].*

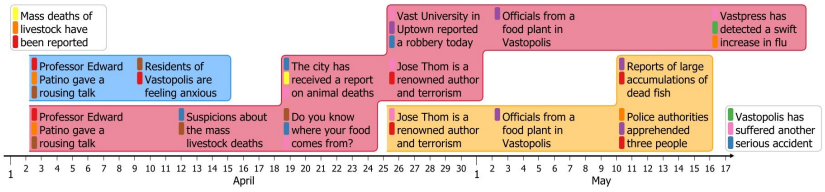


Figure 2.20: *Nguyen et al. present the SchemLine system [26]. Image courtesy of Nguyen et al. [26].*

model. Their work is based on previous work of timeline visualization [6, 32] and sensemarking with timeline [116].

2.4.4 Narrative Visualization for Storytelling in Parallel

In this category of literature, the structure of events is layed out in parallel. The research here focuses on tools and techniques that create multiple narratives at once, in other words simultaneously. These can be useful for groups.

Information visualization systems enable users to find patterns, relationships, and structures in data which may help users gain knowledge or confirm hypotheses [27]. The most basic element in a narrative is a character. An event occurs through the interaction of a set of characters. In this paradigm, a scene consists of a chunk of events, a story consists of a

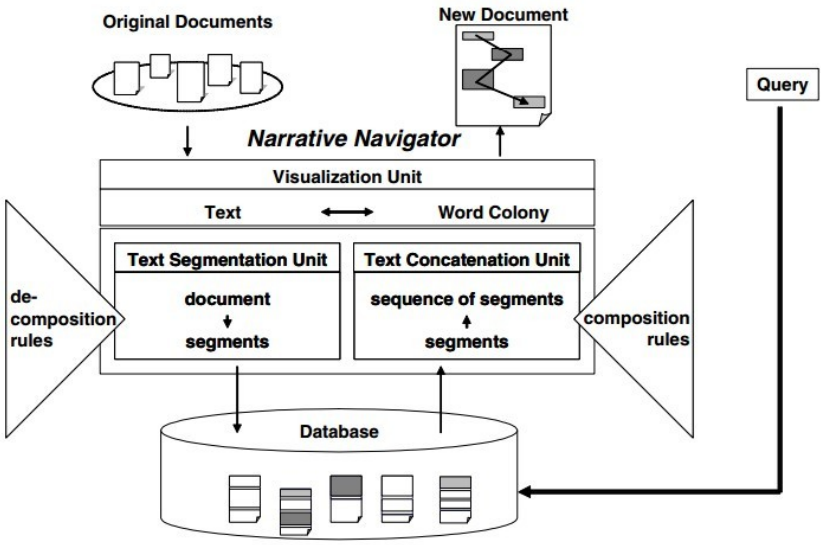


Figure 2.21: *This figure shows the architecture of Narrative Navigator [27].*

sequence of scenes, and a world model is made up of a set of stories. Akaishi et al. propose several methods for visualization of chronological data based on the narrative structure of a document [27]. Akaishi et al. map each narrative component (world model, story, scene, event, character onto elements of a document, set of stories, sequence of scenes, part of sentence, sets of terms). The system features a decomposition unit and a composition unit. A set of stories is stored in a database by the decomposition unit. In the database, each story is divided into scenes, forming a world model. Appropriate scenes are selected and used by the composition unit to compose a new story. When a user accesses the information, the software provides the results as a story. The story is presented in various ways.

The dependency relationship among terms forms a directed graph, called a Word Colony. In a Word Colony, interdependent terms are embedded into the same node. The strength of term dependence is mapped onto the distance between nodes of terms, and term attractiveness is mapped onto the size of node. To visualize this relationship, Akaishi et al. use a spring model graph, which is a visual overview of a document. Narrative navigator framework (NANA) represents the content of a document as a topic sequence and topic matrix. Topic sequence is regarded as the graphical plot of a document and topic matrix represents the relationships among several topic changes. Akaishi et al. support users' efforts to find desired parts of documents and to guess the context (plot) of the document.

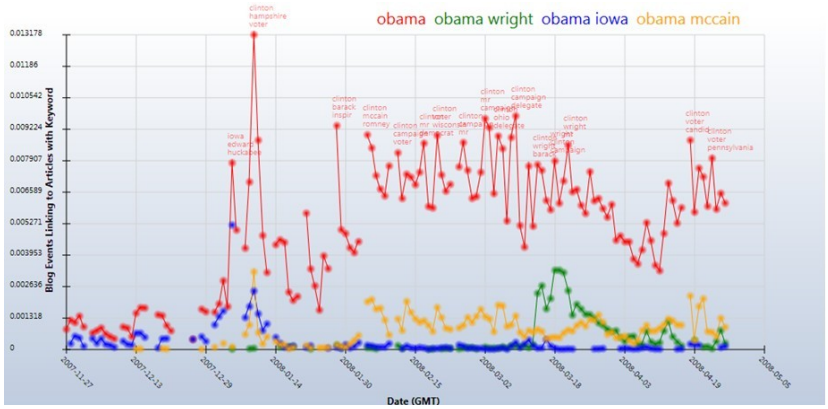


Figure 2.22: *Fisher et al. show daily references to four US presidential candidates from January 1 to March 26, 2008. Time passes along the x axis for each candidate; number of mentions of the term along the y axis [28]. Image courtesy of Fisher et al. [28].*

Narrative is a simple interface that straightforwardly presents trends in keywords over time [28]. Fisher et al. present narrative as a way of presenting temporally dynamic data. In this case, narratives help the user by tracking concepts found in news stories that change over time. Fisher et al. show how to piece together complex information and examine multiple variables, See Figure 2.22 [28].

The first step is based on a business analysis task to find trends and public relations. In this case study, the requirement is to find out how a topic has developed over time and to see the evolution of the latest and most interesting stories [28]. The system design includes data acquisition, temporal visualization, using other tools for correlation, understanding readership, and adding feature in narratives. The narratives project is based on Microsoft's Live Labs which provides real time data acquisition. Temporal visualization enables us see how a small group of words evolves over time relative to one another. By analyzing the form of correspondence and understanding readership, additional features can be added into the narrative project [28]. Fisher et al. is based on previous work in topic detection and tracking[117] [118], and temporal visualization [119], and presents narrative as a new technique in visualization [28].

2.5 Static Transitions in Storytelling for Visualization

A transition refers to the process or a period of changing from one state or condition to another according to the Oxford English Dictionary [120]. In the visualization literature, transitions

may be the focus of visualization and include both dynamic and static which are alternatives of presenting visualization. Static visualizations are those that do not rely on animation. Transitions may be considered part of narrative storytelling. However, we designate the literature here in its own category to reflect the importance of transitions and to keep related literature on this topic together. Several research papers focus on the transitions in storytelling. This is why they are separated into a special group.

In this section, the visual designs of transitions is generally static. The authors focus on presenting the trend of data along timelines. Robertson et al[30] evaluate three approaches of using bubble charts and attempts to discover which one works best for presentation and analysis. Tanahashi and Ma [32] presents a storyline visualization which consists of a series of lines, from left to right along the time-axis. Liu et al. [33] design a storyline visualization system, StoryFlow, to generate an aesthetically pleasing and legible storyline visualization. Ferreira et al[29] propose a method of visualizing a large amount of taxi data consisting of both spatial and temporal dimensions.

2.5.1 Static Transitions for User-directed and Interactive Storytelling

The literature in this subsection focuses on interactive user-driven transitions. The user creates static transitions interactively, i.e. using a process they have some control over(as opposed to automatically).

TaxiVis proposes a method of visualizing a large amount of taxi data consisting of both spatial and temporal dimensions. This approach examines trends over time as opposed to individual taxi trips, visualizing data from a day in length, up to a year. Seasonal events such as Thanksgiving and Christmas can be compared in a like-for-like fashion. See Figure 2.23 [29].

Time selection widget allows the user to change the time frame of the visualization. Maps server as the canvas for the visualization. A graph of the raw data with time plotted to the x axis and frequency of taxi trips on the y axis. To reduce clutter, a density heat map is used. This can either be as points on the map or averaging out the data within regions on the map.

Taxi behaviour is a popular focus of research. Among others, Veloso et al. explored patterns and trends in taxi ride data looking at the relationship between pick up and drop off points [121, 122]. Liao et al. developed a visual analytics system to error check GPS data streamed from taxis [123].

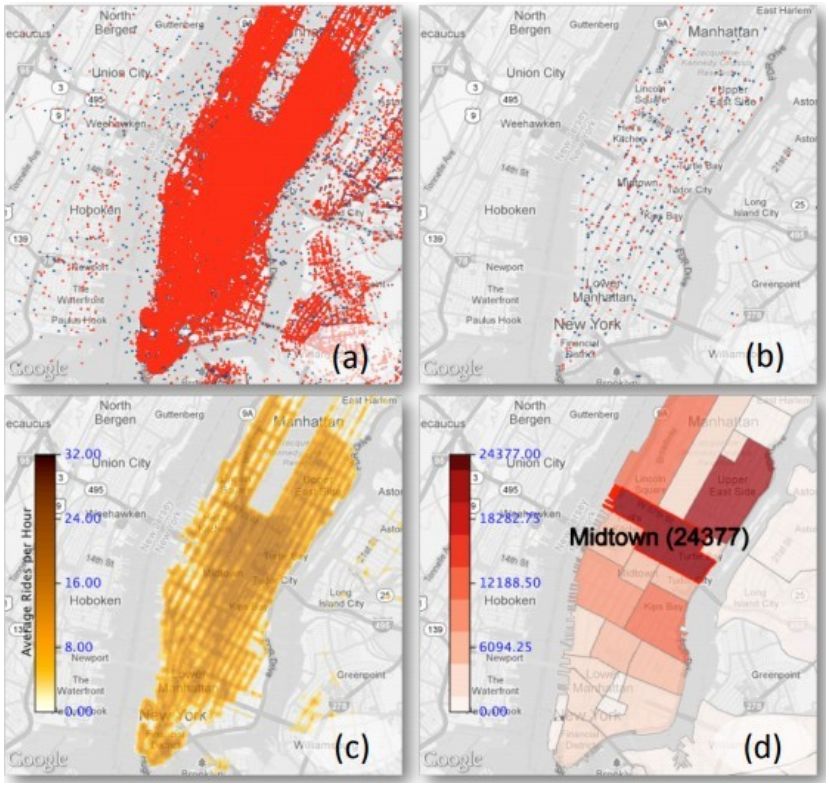


Figure 2.23: *The top-left image shows the trips rendered on the map. However the cluttered view can be reduced by employing a level-of-detail approach (top right) which takes a subsample based on the order in which the trips occurred. The bottom-left image shows a density heat map of the taxi trips whereas the bottom-right image averages out the data in each region to make a regional density heat map [29]. Image courtesy of Ferreira et al. [29].*

2.5.2 Static Transitions for Parallel Storytelling

In this category of literature, the static transitions are shown in parallel. In other words, many transitions can occur simultaneously. Robertson et al. define a trend in data as an observed general tendency. The most common way to see a trend in data is to plot a variable's change over time on a line chart or bar chart. If there is a general increase or decrease over time, this is perceived as a changing trend [30]. Robertson et al. propose two alternatives to animated bubble charts for visualizing trends in multiple dimensions and describes a user study that evaluates the three approaches for both presentation and analysis. In conclusion, Robertson et al. state that traces and small multiples work best for analysis [30].

The gapminder trendalyzer uses a bubble chart to show four dimensions of data, life ex-

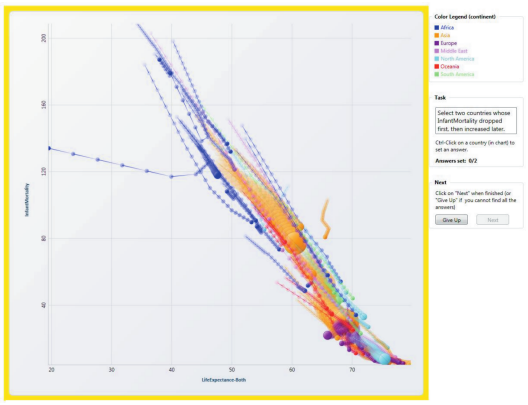


Figure 2.24: *Robertson et al. show the trace lines of the graph animation. The traces visualization shows bubbles at all x and y locations throughout the time frame. This is a conversion of an animation into a static image [30]. Image courtesy of Robertson et al. [30].*

pectancy is mapped to the x axis, infant mortality is mapped to the y axis, population is mapped to bubble size and continent is mapped to color [124]. We can see multiple parallel transitions in Figure2.24 as evolve over time.

An alternative multi-dimensional trend visualization provides the user with the ability to select particular bubbles such that the animation shows a trace line for the selected bubble as it progresses. See Figure 2.24 [125]. In a small multiples visualization, countries can be clustered based on position, size, and location. They are further grouped by continent and ordered alphabetically within each group [126]. Robertson et al. is based on earlier work by Tversky et al. [127] and Baudisch et al. [128]. Previous work is limited to small data set sizes (200 samples or less). Their work focuses on presentation rather than analysis and relies on animation to show trends over time.

Visual Storylines, by Chen et al. is designed to summarize video storylines in an image composition while preserving the style of the original videos [31]. Chen et al. present a new visual storylines method to assist viewers in understanding important video contents by revealing essential information about video story units and their relationships. [31]. The first step of the algorithm is to extract the storylines from a video sequence by segmenting a video into multiple sets of shot sequences and determining their relationships. See Figure 2.25. The second step is to visualize a movie sequence in a new type of static visualization by using a multi-level visual storyline approach, which selects and synthesizes important story segments according to their relationships in a storyline.

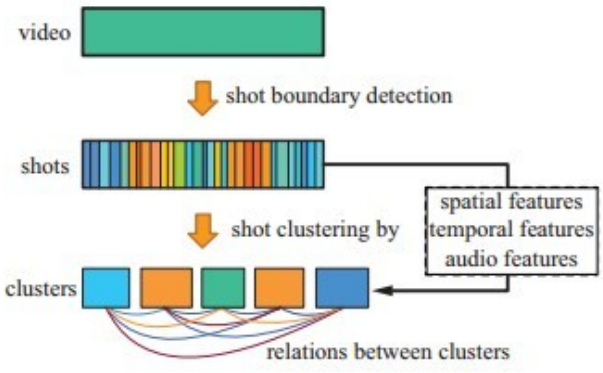


Figure 2.25: *Chen et al. presents video shot clustering algorithm combines both visual and audio features to generate a meaningful storyline [31]. Image courtesy of Chen et al. [31].*

Chen et al. is based on the work of video summarization [129] and first clusters video shots according to both visual and audio data to form semantic video segments.

Storyline visualization is a technique that portrays the temporal dynamics of social interactions by projecting the timeline of the interaction onto an axis [32]. Tanahashi and Ma present a parallel storyline visualization which consists of a series of lines, from left to right along the time-axis, that converge and diverge in the course of their paths [32]. Transitions are shown in parallel storylines in Figure 2.26. Algorithm overview is shown in Figure 2.26. The layout is based on a set of horizontal slots that divide the screen space along the y-axis. Each of these slots has the capacity to accommodate blocks of interaction sessions as long as they do not overlap in time [32]. Rearranging lines takes the slot-based layout of interaction sessions derived from a genome and determines the order of the line segments in each interaction session and its alignment in order to reduce unnecessary wiggles and crossovers [32]. In order to prevent such misleading effects, it is critical for the layout computation to include the removal of unnecessary white space to determine the final layout [32]. Tanahashi and Ma [32] is based on the idea of XKCD's hand-drawn illusion "Movie Narrative Charts" [130] and develops an algorithm for general storyline visualization.

Storyline visualizations, aim to illustrate the dynamic relationships between entities in a story [33]. Liu et al. design a storyline visualization system, StoryFlow, to generate an aesthetically pleasing and legible storyline visualization. It supports real-time user interaction, hierarchical relationships among entities, and the rendering of a large number of entity lines [33]. The layout pipeline consists of four steps: relationship tree generation, session/line or-

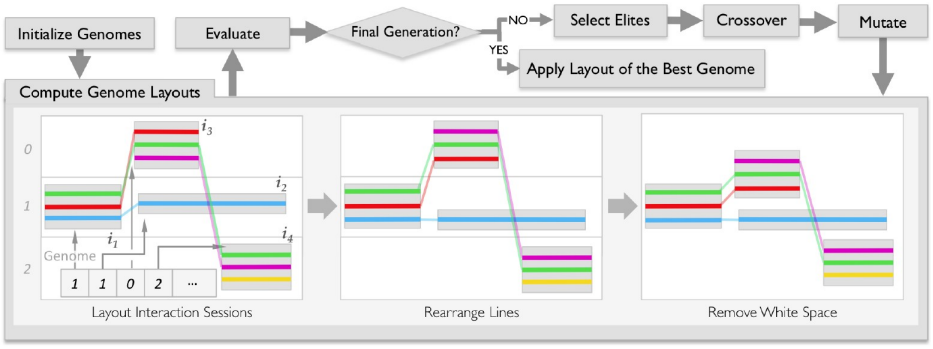


Figure 2.26: *Tanahashi and Ma present the overview algorithm of generating storyline visualizations [32]. Image courtesy of Tanahashi et al. [32].*

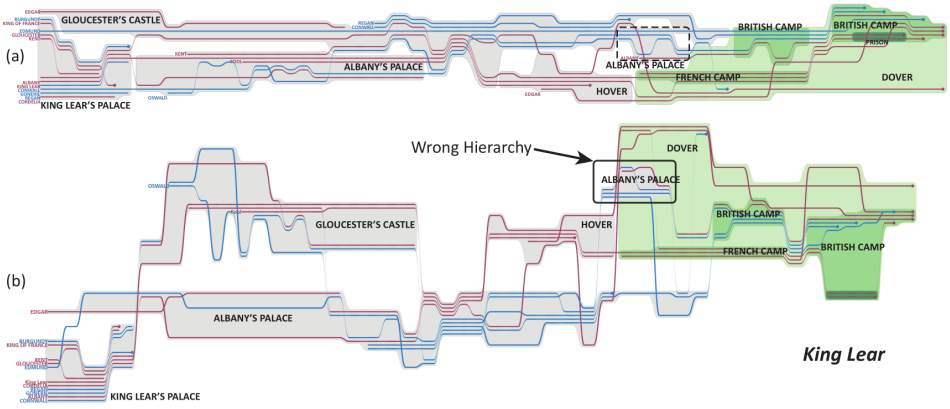


Figure 2.27: *Comparison of King Lear using both methods of layout; (a) - StoryFlow, (b) - previous method by Tanahashi and Ma [32]. The StoryFlow layout presented in this paper focuses on minimising white space and efficiently ordering the story lines to ensure the most concise visual representation of a story. Intersecting lines represent interaction between characters and major events in the story are labeled to add clarity to the visualization [33]. Image courtesy of Liu et al. [33].*

dering, session/line alignment, and layout compaction. In the first step, StoryFlow creates a set of dynamic relationship trees for different time frames, in which the relationship trees are used to order sessions and entity lines. Next, sessions/lines between successive time frames are aligned to maximize the number of straight lines in the layout. Finally, a quadratic optimization algorithm is performed to obtain a compact storyline layout. See Figure 2.27 [33].

Liu et al. is based on previous work of Tanahashi et al. [32]. Liu et al. add support for real-time interaction, hierarchical relationships, and a large number of entity lines.

2.6 Animated Transitions in Storytelling for Visualization

Gonzalez and Cleotilde define animation as a series of varying images presented dynamically according to user actions, in ways that help the user to perceive a continuous change over time and develop a more appropriate mental model of the task [131]. The results of their study show that decision making performance is highly contingent on the properties of the animation user interface such as image realism, transition smoothness, and interactivity style, and also sensitive to the task domain and the user's experience. Values of accuracy, time, ease of use, and enjoyability for the two types of images, transitions, and interactivity styles indicated that realistic images, gradual transitions, and parallel interactivity produced better decisions. Decision making accuracy, time, ease of use, and enjoyability in animated interfaces are influenced by the form of image representation, the transition effects, and the form of interactivity. This research supports the idea that to be an effective decision support tool, animation must be smooth, simple, interactive, and explicitly account for the appropriateness of the user's mental model of the task. Gonzalez and Cleotilde review selected empirical investigations from the literature in education, psychology, and HCI which suggest that animation may make interfaces easier, more enjoyable and understandable, and study the effect of animation on decision making [132].

2.6.1 Animated Transitions for Linear Storytelling

The literature in this sub-section focuses on animated transitions using automatic, or semi-automatic approaches(as opposed to interactive techniques to animated transitions).

Heer and Robertson investigate the effectiveness of animated transitions in traditional statistical data graphs, such as bar charts, pie charts, and scatter plots. A visualisation framework called DynaVis is created to test the effectiveness of animation on the user's preference and information retention. Graph animations are used to keep viewers engaged and to promote creative thinking about the data. See Figure 2.28 [34].

The software displays animated transitions of statistical data graphs. Sorting and filtering animation provide the user insight into the composition of the data. All transitions take place over a time frame rather than instantaneously so the user can see exactly how the visualisation has changed. Animations between different graph types are implemented by morphing the data from one shape and size to another. Statistically significant differences in user preference were found between static graphs and animated graphs. Animated transitions can improve

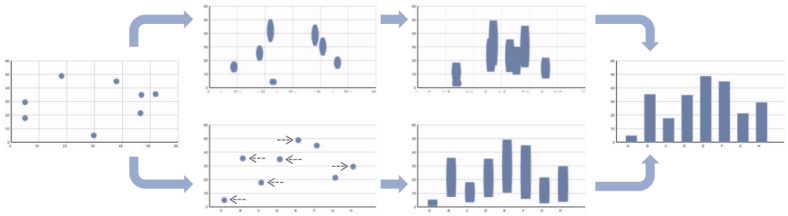


Figure 2.28: *Heer and Robertson show the process of transition for a scatter plot to a bar chart. The top path starts by stretching the points to size and then moving to the right location, whereas the bottom path moves the dots first, then resizes and reshapes them [34]. Image courtesy of Heer and Robertson [34].*

graphical perception. This is reflected in the findings of the user experiments testing recall and understanding. However, not all transition scenarios are found to be significantly different.

Heer and Robertson is based on the previous work of Bederson and Boltman [35] but builds upon it by testing different transitional events.

2.6.2 Animated Transitions for User-directed and Interactive Storytelling

The literature in this subsection focuses on interactive, user-driven transitions. The user or users create animated transitions interactively (as opposed to automatically as in the previous section). Bederson and Boltman examine how animating a viewpoint change in a spatial information system affects a user's ability to build a mental map of the information in the space. Based on a user-study involving a spatial map of a family tree, animation is found to improve subjects' ability to learn the spatial position of family members within the tree without a speed penalty [35].

Two different family trees of nine individuals are presented to two groups people with animation and without animation. The subjects were given three kinds of tasks; navigation of family trees, exploratory family trees, and reconstruction of family trees. The speed and accuracy of performance are recorded. In this experiment, there is a statistically significant improvement in accuracy of the reconstruction task over that of other tasks. Animation resulted in fewer task errors. See Figure 2.29.

Bederson and Boltman is based on Gonzalez [132] and Donskoy and Kaptelinin [133] which address the relationship between animation and users' understanding. Compared to previous work, Bederson and Boltman focus on animation of the viewpoint. The design of the

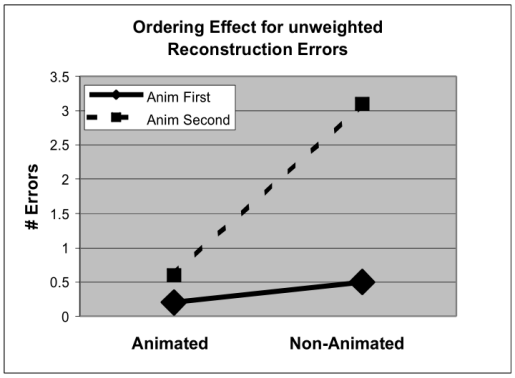


Figure 2.29: *Bederson and Boltman show the ordering effects when presenting an animated and non-animated graphic. If the animated graphic is shown first then there is little difference in recall error, however, if the animation graphic is shown second then the recall error is significantly higher for the non-animated graphic [35]. Image courtesy of Bederson and Boltman [35].*

experiment is to change from a single in-between frame to several in-between frames.

Akiba et al. introduce an animation tool named: AniVis for scientific visualization exploration and communication. This tool can turn the results of data exploration and visualization into animation content and the users can create a complex animation sequence by combining several simple effects [36].

Parameter-space blending operator creates intermediate frames between two instances of frames I_1 and I_2 by interpolating their respective parameters. If I_1 and I_2 do not overlap in time, they generate intermediate frames by interpolating the parameters of the last frame of I_1 and the first frame of I_2. Otherwise, they generate intermediate frames by interpolating the parameters of their corresponding frames [36].

An image-space blending operator creates the animation content between I_1 and I_2 by interpolating their respective image frames. Similarly to parameter-space blending, if I_1 and I_2 don't overlap in time, they generate intermediate frames by blending the last frame of I_1 and the first frame of I_2. The effect is that the last frame of I_1 gradually fades out as the first frame of I_2 gradually fades in. If I_1 and I_2 overlap, they generate intermediate frames by blending [36].

A playback operator lets users repeatedly loop through one or more consecutive instances of interest [36].

A MRI head data case study focuses on highlighting a brain tumor. The animation is

comprised of four pieces of dynamic content. The first is a spatial overview that rotates the volume data 360 degrees along the y-axis. The second piece is a spatial exploration in which the user customizes the view. The third is a parameter-space blending between a spatial exploration and a slicing, which reveals a tumor's inner structure. The parameter-space blending highlights a tumor by varying the opacity while zooming in on the region of interest. See Figure 2.30. A hurricane data case study has five components. The first is a caption showing the animation's content, blended with a spatial exploration that zooms in on the data. The second piece is a temporal exploration to show early time steps. The third is a variable overview that browses through three data attributes: vapor, wind speed, and cloud. The fourth piece is a temporal exploration to show later time steps. The fifth is a spatial exploration that zooms in on the hurricane's eye [36].

Akiba et al. is based on previous animation support [134] and an animation enhanced system [135] and develops template-based visualization tools for animation.

To explore the challenge of gradually moving from interest to insight, Nagel et al. [74] propose the term staged analysis. Invoking temporal and theatrical notions, they define staged analysis as a carefully choreographed process of breaking up a complex whole into its component parts and purposefully preparing the manner of their appearance. In the context of visualization, the concept of staging typically refers to animated transitions broken up to be more easily observed. They build on top of this notion of staging and extend it to a guided analysis process.

As we can see, the literature on transitions is spread amongst information and scientific visualization. Table 2.2 shows an alternative classification of the literature divided up into information, scientific, and geo-spatial visualization. We can see that most of the storytelling research focuses on information visualization.

2.7 Memorability for Storytelling and Visualization

Memory refers to the faculty by which things are remembered; the capacity for retaining, perpetuating, or reviving the thought of things past according to the Oxford English Dictionary [136]. Memorability is an important goal of storytelling. A good visual narrative technique engages the viewer's attention and increases a story's memorability [37].

All papers in this section evaluate the effects of visual narrative on memorability. Bateman et al. [37] explore the effects of embellishment on comprehension and memorability. Saket

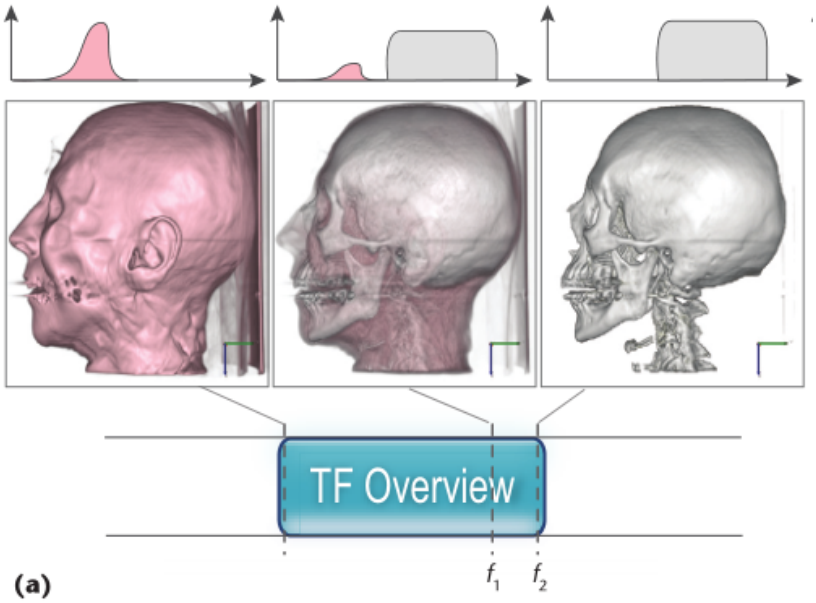


Figure 2.30: *Akiba et al. show the AniVis animation tool displaying MRI scan data. By blending the two layers of data together, a new layer of information is revealed (middle image) [36]. Image courtesy of Akiba et al. [36].*

et al. [39] illustrate that map-based visual narrative can improve accuracy of recalled data comparing with node-link visual narrative.

Borkin et al. [75] develop an online memorability study using over 2000 static visualizations that cover a large variety of visual narrative and determine which visual narrative types and attributes are more memorable. They investigate a domain at the interface between human cognition and visualization design.

A visual narrative taxonomy classifies static visualizations according to the underlying data structures, the visual encoding of the data, and the perceptual tasks enabled by these encodings. It features twelve main visual narrative categories and several popular sub-types for each category. Borkin et al. run memorability tests via Amazon's Mechanical Turk with 261 participants and gather memorability scores. The results in memorability comparison test demonstrates that there is memorability consistency with scenes, faces, and also visual narrative, thus memorability is a generic principle with possibly similar generic, abstract features. The result in visualization attribute tests illustrates that higher memorability scores were correlated with visual narrative containing pictograms, more color, low data-to-ink ratios, and high visual

densities.

Borkin et al. show that visualizations are intrinsically memorable with consistency across people. Visual narratives . with low data-to-ink ratios and high visual densities (i.e., more chart junk and "clutter") were more memorable than minimal, "clean" visualizations [75].

The literatures in this subsection indicates that maps increase memorability. This motivates our choice of using geo-spatial visualization in the remaining chapters of the thesis.

2.7.1 Memorability for Linear Visualization

The literature here shows and tests visual designs in linear order and focuses on memorability. Users are asked to compare the visual designs (e.g. standard bar charts) verses embellished bar charts. In other words, users are tested on their ability to recall one visual design at a time in linear fashion.

Bateman et al. examine whether embellishment is useful for comprehension and memorability of charts. Bateman et al. compare plain and embellished charts, and conclude that a user's accuracy in describing the embellished charts is no worse than for plain charts and that their recall after a two-to-three week gap is significantly better [37].

Fourteen embellished charts are selected from Nigel Holmes' book Designer's Guide to Creating Charts and Diagrams [137], and converted to plain charts. See Figure 2.31. Twenty participants are presented a chart on a slide, alternating between embellished and plain versions. Participants are required to perform two tasks (reading and describing task and recall task) after five-minutes and after 2-3 weeks. The eye-gaze and task performance of participants are recorded for analysis. This study shows that there is no significant difference between plain and embellished versions for interactive interpretation accuracy and recall accuracy after a five-minute gap, but after a long-term gap, recall of both topic and detail of chart (categories and trend) is significantly better for embellished charts. Participants saw the value in message more often in Holmes' charts than in the plain charts.

Previous studies have suggested that minor decoration in charts may not hamper interpretation [138], and work in psychology has shown that the use of imagery can affect memorability [139], but there is very little work that looks at how chart imagery can affect the way people view information charts.

Borkin et al. [38] present the first study incorporating eye-tracking as well as cognitive experimental techniques to investigate which elements of visualizations facilitate subsequent

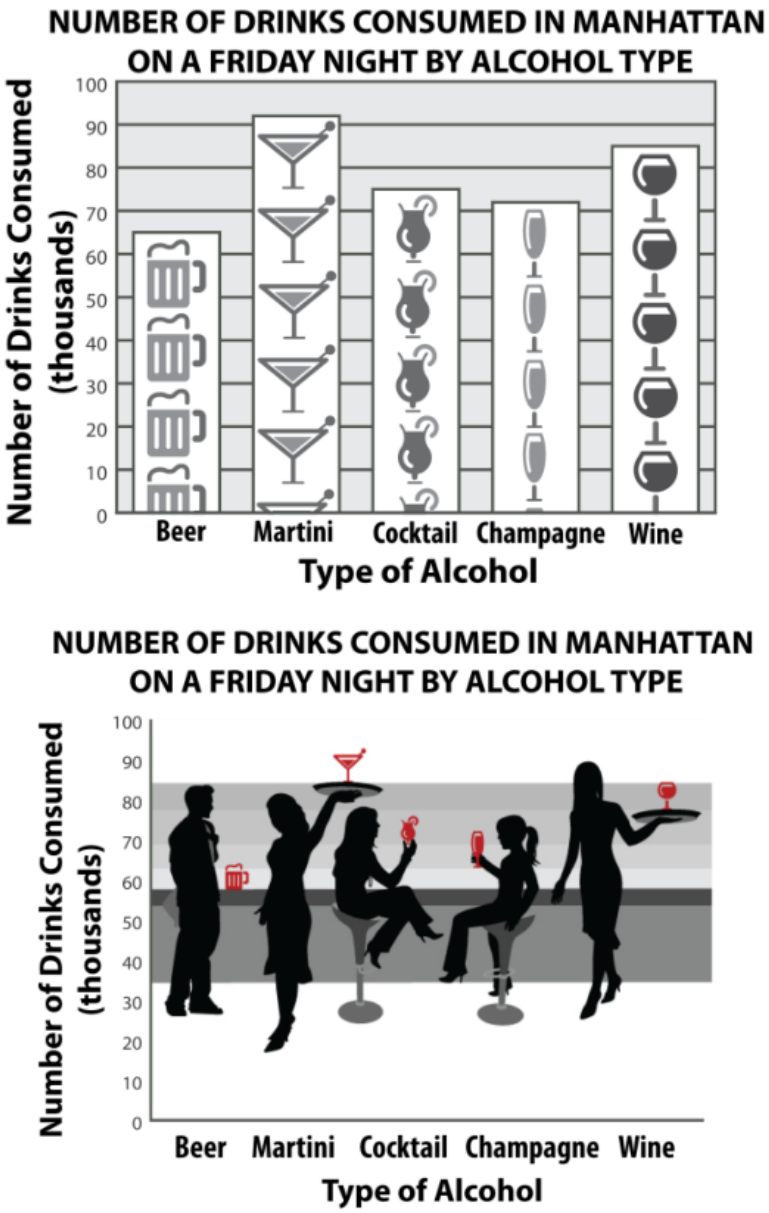


Figure 2.31: *Bateman et al. compare two different levels of graphical embellishment of the same data. The top graph is an embellished image but still retains the recognisable features of a bar chart. The bottom image replaces the bars with a silhouette of a person next to a drink where the height of the drink corresponds to the height of the original bar. This method also uses the addition of color to emphasize the data [37]. Image courtesy of Bateman et al. [37].*

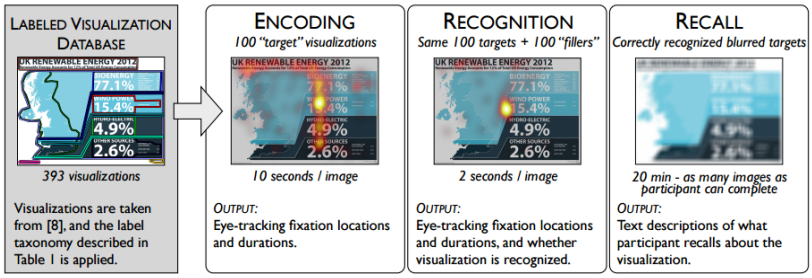


Figure 2.32: *Borkin et al. design three-phase experiment to evaluate viewer performance of recognition and recall [38]. Image courtesy of Borkin et al. [38].*

recognition and recall. They design a three-phase experiment (See Figure 2.32) and evaluate the performance of recognition and recall. The conclusion includes visualizations with more memorable content can be memorable 'at-a-glance.' Titles and text are key elements in a visualization and help recall the message. Pictograms do not hinder the memory or understanding of a visualization. Redundancy facilitates visualization recall and understanding.

Borkin et al. is based on previous work on perception and memorability of visualization [37] and eye-tracking evaluation visualization [140].

2.7.2 Memorability for Parallel Visualization

In this subsection, users are presented with a large number of relation data in parallel (as opposes to one at a time). And it focuses on memorability. Users are tested on their ability to process relationship data in parallel (all relationships simultaneously). This is distinct from memorability for linear visualization where recall focuses on one visual design at a time in linear order.

Saket et al. illustrate that different visualization designs can affect the recall accuracy of data being visualized. Compared to a node-link diagram, a map-based visual design is more effective [39].

Two datasets are examined. A book dataset (small) and LastFM dataset (large) are transformed into a node-link diagram and node-link group (map-based). See Figure 2.33. Three phrases are performed to examine the difference between node-link diagram and map-based visualization. In phase 1 participants examine two kinds of visual design without task with unlimited time. Phase 2 asks participant to study two kinds of visualization with six tasks in a required time. Phase 3 asks participants to recall what they read in phase 1 and 2, complete

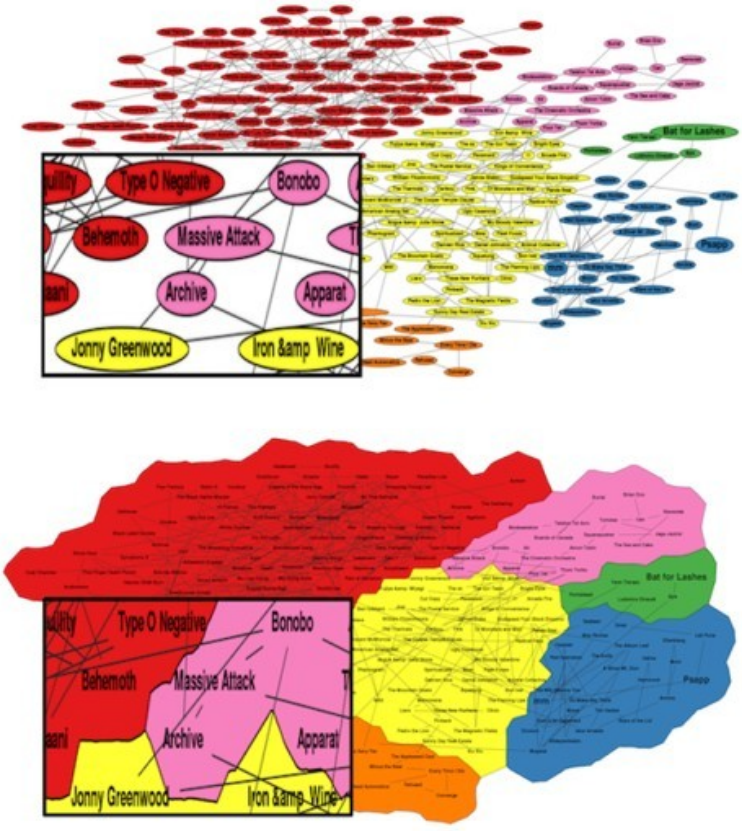


Figure 2.33: *Saket et al. show two visualization of the same data: node-link diagram and map-based diagram [39]. Image courtesy of Saket et al. [39].*

6 tasks similar to phase 2, and 3 new addition tasks [39]. The result of the experiment illustrates that recalling map-based diagrams is more accurate than recalling node-link diagrams, but no faster. The participants spent more time on map-based visualizations than node-link visualizations [39].

Saket et al. is based on previous work of visualization memorability [37] and a recalling experiment [141]

2.8 Discussion and Unsolved Problem

This chapter provides a novel up-to-date overview of narrative visualization. The most important recent literature is included and discussed. Since storytelling in visualization is a recently new subject, we expect an increase in research in the coming years. Moreover we believe it will evolve into a popular topic in the field of visualization.

By reviewing Table 1 and Table 2, we can see storytelling visualization focuses on information visualization more than scientific visualization, which conveys that more challenges are left unsolved in this field. However, by refining a storytelling model for scientific visualization [10], the implementation of storytelling in scientific visualization could increase in the future. We can also see that storytelling in visualization concentrates more on exploration than on presentation. Like Kosara and Mackinlay [115] state: "visualization techniques address the exploration and analysis of data more than presenting data".

In future work, there are many directions and unsolved problems. Narrative visualization will gain importance in data presentation and data exploration. Here is a summary of some unsolved problems in storytelling for visualization.

- It is clear that objective measures of user-engagement is a relatively unexplored area of research. Can we derive a mature classification of user engagement activities? Is user engagement something we can clearly define?

- Data preparation and enhancement: Virtually no one has addressed the challenge of data preparation and enhancement for storytelling. Moreover, is storytelling data best captured or derived from an existing data set or software system? Can a standard data file format be developed?

- Narrative visualization for scientific and geo-spatial visualization: Why has there been such an imbalance of research narrative visualization for information visualization but virtually none for scientific and geo-spatial visualization.

- Transitions for scientific visualization: The benefits of static transition versus dynamic transitions in visualization still remains relatively immature.

- Memorability for visualization: What are the key elements for making a memorable visualization? This is still an immature research direction.

- Animated transitions for geo-spatial visualization: Animated transitions for geo-spatial visualization remains an open research direction. This is surprising given the popularity and importance of geo-spatial visualization.
- Interpretation for scientific information, and geo-spatial visualization: Currently no papers to our knowledge focus on the topic of effective interpretation of stories, this topic remains largely unexplored.

The classification of literature, we present makes it clear that many future research directions remain open in storytelling and visualization.

Chapter 3

Cartographic Treemaps for Visualization of Healthcare Data

Contents

"Tell me and I forget, teach me and I may remember, involve me and I learn."-Benjamin Franklin[1]

[1] Benjamin Franklin (1705-1790) was an American polymath and one of the Founding Fathers of the United States.

This Chapter presents a novel multivariate visualization combining geo-spatial information. As we saw in the previous chapter, including geo-spatial information can increase memorability and cognition of information and data. The National healthcare Service (NHS) in the UK collects a massive amount of high-dimensional, region-centric data concerning individual healthcare units throughout Great Britain. It is challenging to visually couple the large number of multivariate attributes about each region unit together with the geo-spatial location of the clinical practices for visual exploration, analysis, and comparison. *We present a novel multivariate visualization we call a cartographic treemap that attempts to combine the space-filling advantages of treemaps for the display of hierarchical, multivariate data together with the relative geo-spatial location of NHS practices in the form of a modified cartogram.* It offers both space filling and geospatial error metrics that provide the user with interactive control over the space-filling versus geographic error trade-off. The result is a visualization that offers users a more space efficient overview of the complex, multivariate healthcare data coupled with the relative geo-spatial location of each practice to enable and facilitate exploration, analysis, and comparison. We evaluate the two metrics and demonstrate the use of our approach on real, large high-dimensional NHS data and derive a number of multivariate observations based on healthcare in the UK as a result. We report the reaction of our software from two domain experts in health science. This Chapter is based on the paper "Cartographic Treemaps for the Visualization of Public Health Care Data" [47].

3.1 Introduction

Coupling geo-space to the NHS data will facilitate understanding. Geo-spatial observations can be made and healthcare patterns can be coupled to their local population. However, multivariate geo-spatial visualization is an unsolved problem, thus novel visual designs are required. Because we are using UK map as our starting point, and its shape is narrow. So there are too much information crowded in very small area, especially for the London area. A large percentage of screen space is blank without showing any useful information. That's why we are using treemap, which is original space-filling technique, as our approach for displaying multi-variate and hierarchical information. We believe that our approach is more novel than a glyph-based approach.

The United Kingdom faces massive challenges with respect to providing the best healthcare via the National Health-care Service (NHS). In order to provide the best service, Public Health

England and the UK government collect years worth of region specific-health care data [46]. The public health profiles website [46] is used for publishing the latest national health care data in the UK. The data archive is designed to support GPs, clinical commissioning groups (CCGs), and local authorities to ensure that they provide and commission effective and appropriate health care services. However the size and complexity of the data creates challenges for deriving new knowledge and insight.

The NHS data includes a UK map divided into CCGs, which are groups of NHS practices. Each CCG contains the local population and high-dimensional health care data collected by the NHS, such as cardiovascular disease (CVD) diagnoses, indicators of respiratory health, mental health indicators, incidents of chronic obstructive pulmonary disease (COPD), kindey disease, as well as other diagnoses.

Our goal is to develop imagery that combines UK-centric geo-spatial information with high-dimensional NHS data in a unified framework. Moreover, we believe the principles apply equally well to other multivariate data sets of this kind. A hybrid visualization we call a Cartographic Treemap combines the geo-spatial properties of cartograms with the space filling properties of treemaps, inheriting advantages of both. We provide the user interactive control over the trade off between filling the most space, like a treemap, and geo-spatial error. Currently, visualizing multi-dimensional health care data based on CCGs is not possible because many CCGs cover the space of only a few pixels. Many CCGs are crowded into the London region, obstructing any geo-spatial visualization without a second magnified view. We propose a cartographic treemap to integrate a modified representation of the UK based on the geo-spatial information of CCG regions combined with a modified treemap to present the multivariate NHS data. Based on the output, we can generate a linear narrative visualization which try to engage the users with the data and increase user's memorability. The contributions of this chapter include:

- A new hybrid visualization, the Cartographic Treemap, combining geo-spatial information in the form of a modified cartogram with space-filling geometry for the visualization of high-dimensional data.
- A layout algorithm for rectangular cartographic treemaps: increasing region size incrementally and avoiding overlapping regions.

- A novel, interactive error metric and user options that trade-off screen space versus geo-spatial accuracy to facilitate user analysis.
- The novel application of our hybrid visualization to complex, real-world NHS data from the UK.

In order to achieve this, several challenges must be overcome. The first challenge is to derive an algorithm that can incorporate both the advantages of geo-spatial cartograms with those of space-filling treemaps. A second requirement is to preserve the local neighborhood relationships of CCG regions to maximize legibility. Another is to provide user-options to facilitate both exploration, analysis, and comparison of hierarchical, multivariate data.

Some very helpful survey papers provide an overview of health care research [142, 143, 144, 145]. However we would like to couple geo-spatial information with the health care data to increase understanding and discover patterns with respect to location.

Multi-variate data visualization: there are a number of survey papers that provide an overview of many multi-variate visualization techniques. We refer the reader to McNabb and Laramee for a comprehensive overview [145]. It summaries the visualization techniques used for multi-variate and hierarchical data, including treemaps, parallel coordinates, and glyphs. As we are going to make good use of the screen space, we choose treemap for presenting multi-variate information.

Geo-spatial related work falls into the areas of cartograms and spatially-ordered treemaps. we separate and review those two categories of previous paper here.

Cartographic visualization Cruz et al. [146] define a cartogram as "*a technique for displaying geographic information by resizing a map's regions according to a statistical parameter in a way that still preserves the map's recognizability*". They can display geo-spatial information and another data attribute (such as population or disease prevalence) in one visualization. Tobler [147] and Nursat and Kobourov [148] survey general cartograms. They present the development of value-by-area cartogram algorithms and performance in computer science.

Auber et al. [149] propose a layout method based on a geographic map metaphor, which facilitates the visualization and navigation of a hierarchy and preserves the order of the hierarchy's nodes.

Gastner and Newman [150] present a diffusion cartogram for constructing value-by-area cartograms, which provides a valuable tool for the presentation and analysis of geographic

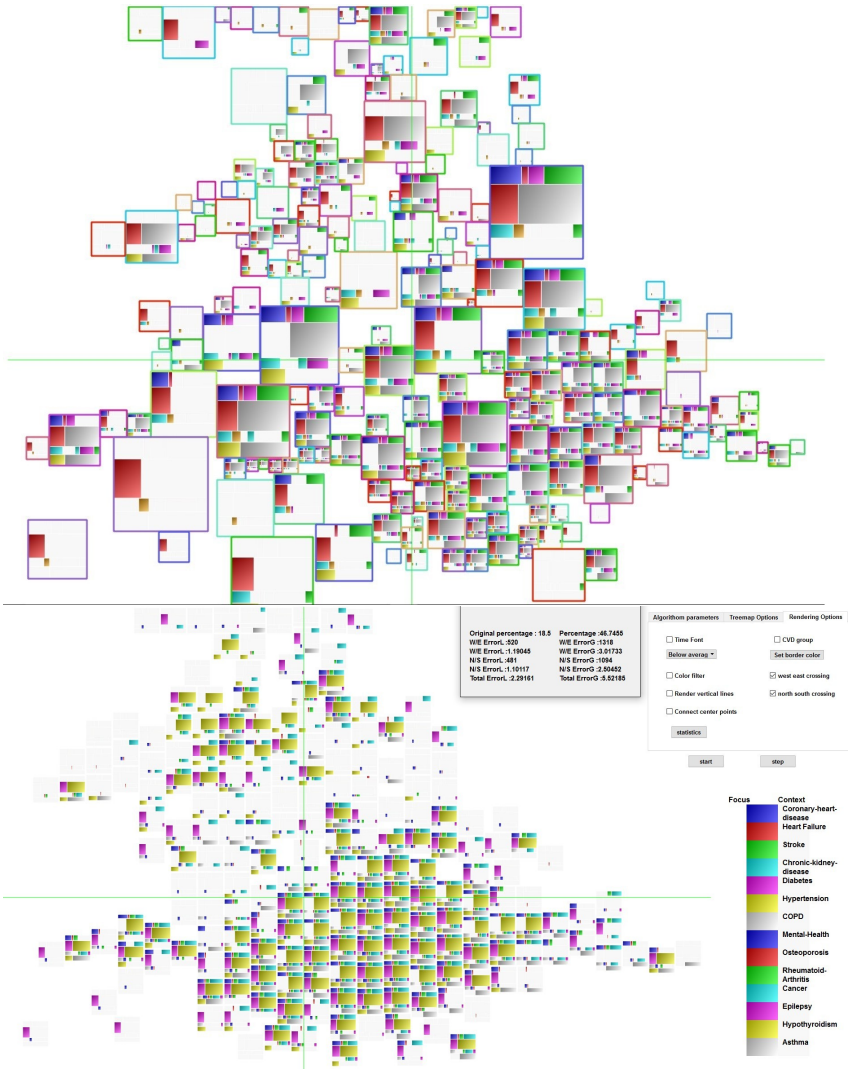


Figure 3.1: *This graph shows each region size proportional to its population with an added below average filter (top). The percentage of screen space occupied, s_0 = 41% and the local error, e_l = 3.5%, e_g=8.7% and uniform size output with a below average filter (bottom). s = 47%, e_l = 2.3%, and e_g = 5.5%. All the health care disorders that exhibit higher than average prevalence are filtered and shown as white context. Note how the London region is healthier with the exceptions of diabetes and mental health. This is an observation based on multiple variates that would be difficult to make otherwise.*

data. Keim et al. [151] develop a faster algorithm for cartograms. It enables display dynamic data with cartogram visualizations. These two algorithms are categorised as contiguous area cartograms. Their performance depends on the corresponding value in each area. If the value does not correspond to the area, the cartogram may be difficult to recognize.

Raisz [152] presents the rectangular cartogram, using rectangles instead of real area shapes. Dorling [153] presents the Dorling cartogram which uses circles instead of geographic area shape, similar to the modified cartogram we present. They are categorized as non-continuous area cartograms. They can display statistical information well, regardless of original shape of area, and preserve relative position. Van Kreveld and Speckmann [154] present the first algorithm for rectangular cartograms. They formalize region adjacencies in order to generate processable layouts that represent the positions of the geographic regions. It converts a rectangular cartogram to a contiguous area cartogram. Our modified cartogram does not fall into the category of continuous cartograms but resembles a cross between rectangular and Dorling cartograms [148]. Our algorithm can be considered as a modified space-filling rectangular cartogram with the addition of a hierarchical structure and multivariate data.

Heilman et al. [155] propose a novel visualization technique for geo-spatial datasets that approximates a rectangular partition of the rectangular display area into a number of map regions preserving important geo-spatial constraints. They use elongated rectangles to fill the space whereas we use uniform rectangles to fill the space such that regions can easily be compared with one another. Their work focuses on univariate, non-hierarchical data.

Panse et al. [156] combine a cartogram-based layout (global shape) with PixelMaps (local placement), obtaining benefits of both for improved exploration of dense geo-spatial data sets. Their work also focuses on univariate, non-hierarchical data.

Slingsby et al. [157] explore the effects of selecting alternative layouts in hierarchical displays that demonstrate multiple aspects of large multivariate data sets, including spatial and temporal characteristics. They demonstrate how layouts can be related, through animated transitions, to reduce the cognitive load associated with their reconfiguration whilst supporting the exploratory process. No metric for neighborhood preservation is described in this work.

Slingsby et al. [158] present rectangular hierarchical cartograms for mapping socio-economic data. They present a detailed map of 1.52 million UK unit postcodes in their spatial hierarchy, sized by population and coloured by the OAC category that most closely characterises the population. However, no algorithm for preserving geo-spatial information is provided. No metric for neighborhood preservation is described.

Alam et al. [159] present a set of seven quantitative measures (Average Cartographic Error, Maximum Cartographic Error, Adjacency Error, Angular Orientation Error, Hamming Distance, Average Aspect Ratio, Polygonal Complexity) to evaluate performance of cartograms

based on the accuracy of data and its readability. They compare previous cartogram algorithms based on statistical distortion, geography distortion and algorithm complexity and evaluate their performance with respect to different properties. Nursat and Kobourov [148] survey cartogram research in the field of visualization and present design guidelines as well as research challenges. They state that mapping multivariate data is still a challenge in cartogram research. In general, previous cartographic visualizations focus on flat, univariate data, whereas we process hierarchical, multivariate data.

Eppstein et al. [160] introduce a new approach to solve the association challenge for grid maps by formulating it as a point set matching problem. They present algorithms to compute such matchings and perform an experimental comparison that also includes a previous method to compute a grid map. Their work focuses on geo-spatial information and filling space. multivariate, hierarchical data is not considered.

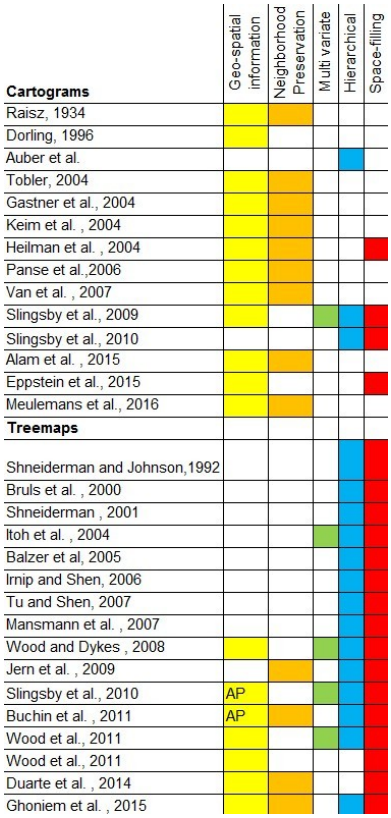
Meulemans et al. [161] design a comprehensive suite of metrics that capture properties of the layout used to arrange the small multiples for comparison (e.g. compactness and alignment) and the preservation of the original data (e.g. distance, topology and shape). Their work focuses on geo-spatial information and neighborhood preservation. Multivariate, hierarchical data is not considered.

We note that the visualizing multivariate data is one of the top future research challenges in the latest survey by Nursat and Kobourov [148]. Also cartograms, in general, are not space-filling and do not necessarily make the best use of screen space.

Treemaps First presented by Shneiderman and Johnson [162, 163, 164], the approach to building a treemap involves converting hierarchical data to a 2D space-filling region. The main challenge with building a treemap is the node packing algorithm that positions the leaf nodes.

Traditional Treemaps Variations on the traditional treemap enable data to retain its original order when visualised [165] or enable the viewer to maintain an understanding of the visualisation if the dataset is dynamic which traditionally results in nodes changing position [166]. The node appearance has also been researched to improve the aesthetic quality of the design and to reveal an insight into the hierarchical structure of the data within the node [167, 168, 169].

Further adaptions of the treemap have been created with a focus on aesthetics that place less emphasis on usability. Voronoi Treemaps do not use regular rectangle node shapes, but rather a many sided polygon consisting of curved and straight lines. The resulting imagery is



Cartograms	Geo-spatial information	Neighborhood Preservation	Multi variate	Hierarchical	Space-filling
Raisz, 1934					
Dorling, 1996					
Auber et al.					
Tobler, 2004					
Gastner et al., 2004					
Keim et al. , 2004					
Heilman et al. , 2004					
Panse et al.,2006					
Van et al. , 2007					
Slingsby et al., 2009					
Slingsby et al., 2010					
Alam et al. , 2015					
Eppstein et al., 2015					
Meulemans et al., 2016					
Treemaps					
Shneiderman and Johnson,1992					
Bruls et al. , 2000					
Shneiderman , 2001					
Itoh et al. , 2004					
Balzer et al, 2005					
Irnip and Shen, 2006					
Tu and Shen, 2007					
Mansmann et al. , 2007					
Wood and Dykes , 2008					
Jern et al. , 2009					
Slingsby et al., 2010	AP				
Buchin et al. , 2011	AP				
Wood et al., 2011					
Wood et al., 2011					
Duarte et al. , 2014					
Ghoniem et al. , 2015					

Figure 3.2: *This table shows characteristics of related work. It includes five visualization properties: geo-spatial information, neighborhood preservation, multivariate, hierarchical and space-filling. Geo-spatial information implicates whether a visualization conveys geographic information and AP in the column represents adjacency preservation only. Neighborhood preservation indicates an algorithm that features a distance metric to preserve neighborhood relationships. multivariate indicates the dimensionality of abstract data. Hierarchical indicates a type of hierarchical data and space-filling indicates how well the output visualization fills the screen. Cartographic treemaps feature all five properties.*

more impressive visually but may sacrifice accuracy and readability [167].

Geo-Spatial Treemaps Mansmann et al. [170] present HistoMaps for visual analysis of computer network traffic visualization with a case study showing that a geographic treemap can be used to gain more insight into these large data sets. However the visualization is essentially univariate (one scalar per level in the hierarchy). It is also not adjacency preserving.

Wood and Dykes [171] provide a squarified layout algorithm that exploits the two-dimensional arrangement of treemap nodes more effectively. It is suitable for the arrangement of data with a geographic component and can be used to create tessellated cartograms for geo-visualization. They convert a geographic distribution of French provinces to a spatial treemap layout and pre-

serve the corresponding geo-spatial relationships to some extent. However, they demonstrate that it is impossible to preserve local region adjacencies if nodes are constrained to a standard rectangle parent node. For example, a region map may only have one or two neighbors on a geographic map. We preserve geo-spatial relationships with less error by allowing gaps in screen space at the different levels of the data hierarchy.

Jern et al. [172] demonstrate and reflect upon the potential synergy between information and geo-visualization. They perform this through the use of a squarified treemap dynamically linked to a choropleth map to facilitate visualization of complex hierarchical social science data. It conveys the neighborhood relationships by using a second view.

Slingsby et al. [173] develop an OAC (Output Area Classifier) explorer that can interactively explore and evaluate census variables. There is no inherent information preserving the geo-spatial location of regions because a synthetic grid is used to sub-divide space. It is not possible to derive any information about the geography of the UK regions.

Buchin et al. [174] describe algorithms for transforming a rectangular layout without hierarchical structure, together with a clustering of the rectangles, into a spatial treemap that respects the clustering and also respects to the extent possible the adjacencies of the input layout. The work of Buchin et al. is similar to ours with few differences. First, they do not demonstrate their layout algorithm on a full geo-spatial map, e.g. the UK. Second, the space-filling requirement results in elongated rectangles that are difficult to compare. Third, the data is univariate.

Wood et al. [175] present Ballotmaps that using hierarchical spatially arranged graphics to represent two locations (geographical areas and spatial location of their names on the ballot paper) that affect candidates at very different scales but their work does not contain any neighborhood preservation algorithm.

Wood et al. [176] identify changes in travel behavior over space and time, aid station rebalancing and provide a framework for incorporating travel modeling and simulation by using flow maps. Their work focuses on univariate, non-hierarchical data.

Duarte et al. [177] propose a novel approach, called a Neighborhood Treemap (Nmap), that employs a slice-and-scale strategy where visual space is successively bisected in the horizontal or vertical directions. The bisections are scaled until one rectangle is defined per data element. Nmap achieves good space-filling visualization that couples related rectangles using a distance metric. However, the distance metric is not geo-spatial, it is also not a treemap of multivariate

data nor a hierarchical visualization.

Ghoniem et al. [178] present a weighted maps algorithm, which is a novel spatially dependent treemap. They present a quantitative evaluation of results and analyze of a number of metrics that are used to assess the quality of the resulting layouts. The work of Ghoniem et al. is similar to ours with some important differences. They place emphasis on evaluating adjacency relationships between nodes rather than geo-spatial positions. Requiring 100% space-filling results in higher geo-spatial error and elongated nodes. Also the data is not multivariate.

Treemaps: Geo-spatial information versus adjacency preservation: In general, the treemap layout algorithms attempt to reflect geo-spatial information implicitly through adjacency relationships between the nodes. As shown by Ghoniem et al. [178], this leads to high geo-spatial error, e.g. in the 40%-50%. It also leads to elongated rectangles which may be difficult to compare. It may be difficult to recognize the correspondence to the original geo-spatial map when looking at a treemap. In contrast, our algorithm emphasizes geo-spatial preservation with less emphasis on adjacency relationships. We give the user new interactive control over the amount of error and allow spaces and gaps to reduce geo-spatial error.

The work we present here differs from previous work in that it attempts to combine the space-filling, hierarchical characteristics of ordered space-filling treemaps together with the geo-spatial information conveyed by a cartogram. Table 3.2 compares the current work with the work presented here. No previous algorithm combines all five properties. Cartographic Treemaps convey geo-spatial information. They feature an error-driven distance metric between nodes and visualize multivariate hierarchical data. They also give the user interactive control over how much screen space is used.

3.2 NHS Data Description

The NHS data includes a UK map divided into CCGs, groups of NHS practices. See Figure 3.3. A standard map of the UK only covers about 18 % of screen space due to its awkward shape. Each CCG contains various categories of disease in prevalence value. Prevalence is the proportion of a population who have a specific medical diagnosis in a given time period, typically an illness, a condition, or a risk factor such as depression or smoking. Prevalence is a derived metric of the local population of each region. Prevalence is usually expressed as a percentage.

Typically this data is displayed using line charts, bar charts, and pie charts. The map provided by public health England is a standard UK map with 209 CCG regions. See Figure 3.3. The boundaries of CCG regions vary and are difficult for presenting high-dimensional data. The CCGs coupled directly to the geography do not make efficient use of space. The UK map itself only occupies 18% of screen space. For visualization purposes the CCG regions in London for example, crowd together and hamper our ability to visualize multi-dimensional data clearly. This will be true in the capital region of most countries and other densely populated areas. Other health care data, for example, the population distribution data is typically visualized using a single line chart showing the percentage of age groups distributed from 0-4 to 85+. Standard graphs show no connection with other health data attributes such as geo-spatial location and clinical diagnoses. This challenging data set is the inspiration behind cartographic treemaps. See the supplementary PDF for a description of the health disorders.

The specific data attributes of each CCG include:

- Local population, which contains the number of patients in 5 year intervals starting with age ranges from 0-4 to 80+.
- A Practice summary providing information on practice demography, deprivation, patient satisfaction and life expectancy estimates.
- Also included are estimated disease prevalence which includes prevalence estimates for cardiovascular disease (CVD), coronary heart disease (CHD), chronic obstructive pulmonary disease (COPD), hypertension and stroke.
- Coronary heart disease (CHD) contains the estimated prevalence value of CHD and heart failure, blood pressure readings and total measured cholesterol.
- CVD - Stroke and TIA contains estimated prevalence values of stroke, blood pressure reading and total measured cholesterol that relate to stroke and transient ischaemic attacks (TIA).
- CVD - Heart failure and atrial fibrillation contains estimated prevalence value of heart failure, atrial fibrillation, and estimated stroke risk.
- CVD - Risk factors for CVD contains prevalence of hypertension, obesity, smoking, ex-smoking.

Figure 3.3: *This graph shows the original 209 CCG regions (Clinical Commissioning Groups) provided by Public Health England [40].Only 18% of screen space is covered by a traditional map.*

- Diabetes contains estimated prevalence values of diabetes, hypertension, smoking and obesity.
- Mental Health contains estimated prevalence values of mental health, demetia and depression.
- Respiratory Disease contains estimated prevalence values of chronic obstructive pulmonary disease (COPD), asthma, smoking and ex-smoking.
- Chronic Kidney Disease (CKD) contains estimated prevalence value of chronic kidney disease, and reading of blood pressure.

- Musculoskeletal Conditions contains estimated prevalence values of osteoporosis and rheumatoid arthritis.
- Other Conditions which contains other QOF clinical indicators including cancer, epilepsy, learning disabilities, hypothyroidism, palliative cares and cervical screening.
- Secondary Care Use - Outpatients contains outpatient attendances, first outpatient appointment, and the value of general practitioner (GP) refers to various diseases per 1000 person.
- Secondary Care Use - and Inpatients contains indicators of hospital accident and emergency and inpatient use. It contains indicators for CHD, respiratory disease, diabetes, cancer, COPD and long-term neurological conditions.
- Child health contains demographic data and secondary indicators (such as attendances, Elective hospital admissions, Emergency hospital admissions) related to child health.

3.3 Cartographic Treemaps

This section describes the cartographic treemaps construction algorithm and interactive error control, starting with an overview. We choose cartograms because data is coupled to location. See Figure 3.4. The processing begins with reading the UK geo-spatial information and high-dimensional health care data. The algorithm is as follows:

(1) Compute region center points: We use the QGIS [41] tool to calculate the center points of each CCG region. The center points are the starting positions of the rectangular region nodes. (2) Update node size: We start with a unit square to represent each CCG region as a node in the cartographic treemap and increase the size of each node according to the user's chosen space-filling target or error constraint. (3) Update cartographic layout: During the region growing process, one region may shift adjacent neighboring regions to remove overlap and preserve relative position. When all regions reach their maximum size or the user-specified geo-spatial error is reached, the cartogram layout stops. We use the fast overlap removal algorithm [179, 180] incrementally for this process.

(4) Treemap node layout: After the cartographic node layout is completed, an ordered squarified treemap layout is used to present the multivariate health care data in each CCG

region, the lowest (finest) level in the treemap hierarchy. (5) Interactive user options: For further exploration, analysis and region comparison, several user options are designed to present the results focusing on different user requirements, such as modifying algorithm parameters, region selection for detail, modifying the color legend, and exploring the hierarchy.

1. Compute region center points: We use the QGIS [41] tool to calculate the center points of each CCG region. The center points are the starting positions of the rectangular region nodes.

2. Update node size: We start with a unit square to represent each CCG region as a node in the cartographic treemap and increase the size of each node according to the user's chosen space-filling target or error constraint.

3. Update cartographic layout: During the region growing process, one region may shift adjacent neighboring regions to remove overlap and preserve relative position. When all regions reach their maximum size or the user-specified geo-spatial error is reached, the cartogram layout stops. We use the fast overlap removal algorithm [179, 180] incrementally for this process.

4. Treemap node layout: After the cartographic node layout is completed, an ordered squarified treemap layout is used to present the multivariate health care data in each CCG region, the lowest (finest) level in the treemap hierarchy.

5. Interactive user options: For further exploration, analysis and region comparison, several user options are designed to present the results focusing on different user requirements, such as modifying algorithm parameters, region selection for detail, modifying the color legend, and exploring the hierarchy.

Computing Region Center Points QGIS (known as "Quantum GIS") is a cross-platform free, open-source desktop geographic information system (GIS) application that provides data visualization, editing, and analysis capabilities [41]. We use QGIS to calculate the centroid of each CCG region as seed points for our cartographic treemap algorithm. The **CentralPoint** calculation tool calculates the central point of each CCG region and stores them in a CSV file. See Figure 3.5. The algorithm for computing a centroid is from Boruke [181]. Consider the area made by n points from x_0, y_0 to $x_n - 1, y_n - 1$, the central point $c(x,y)$ is given by following

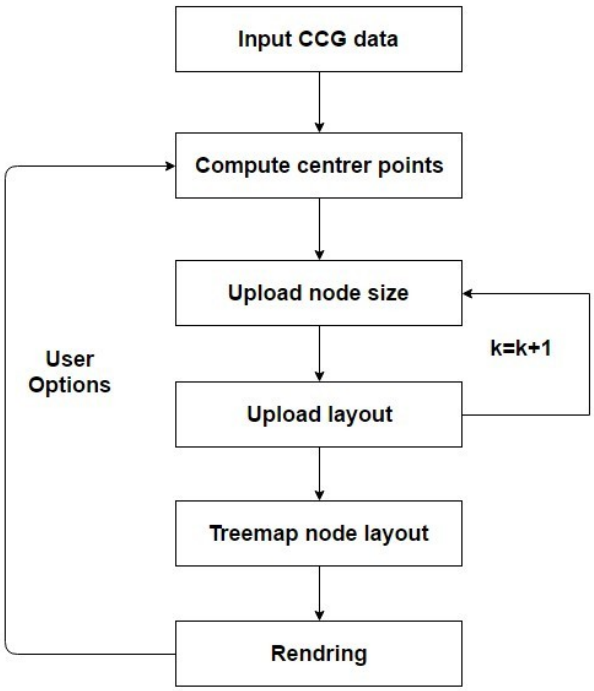


Figure 3.4: *This is the processing pipeline for producing cartographic treemaps. k is the counter used to gradually expand each region node during node layout.*

formula, and A is the polygon's signed area defined by

$$A = \frac{1}{2}\sum_{i=0}^{N-1}(x_i y_{i+1} - x_{i+1} y_i) \tag{3.1}$$

$$c_x = \frac{1}{6A}\sum_{i=0}^{N-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \tag{3.2}$$

$$c_y = \frac{1}{6A}\sum_{i=0}^{N-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \tag{3.3}$$

This formulation is provided by Boruke [181] for the computation of a closed 2D polygon centroid.

3.3.1 Updating Node Size

After calculating the center point of each CCG node, we initialize CCG nodes as unit squares on the cartographic treemap. The algorithm increases the size of each node to make the most efficient use of space. It terminates when the user-specified geo-spatial error or a target screen

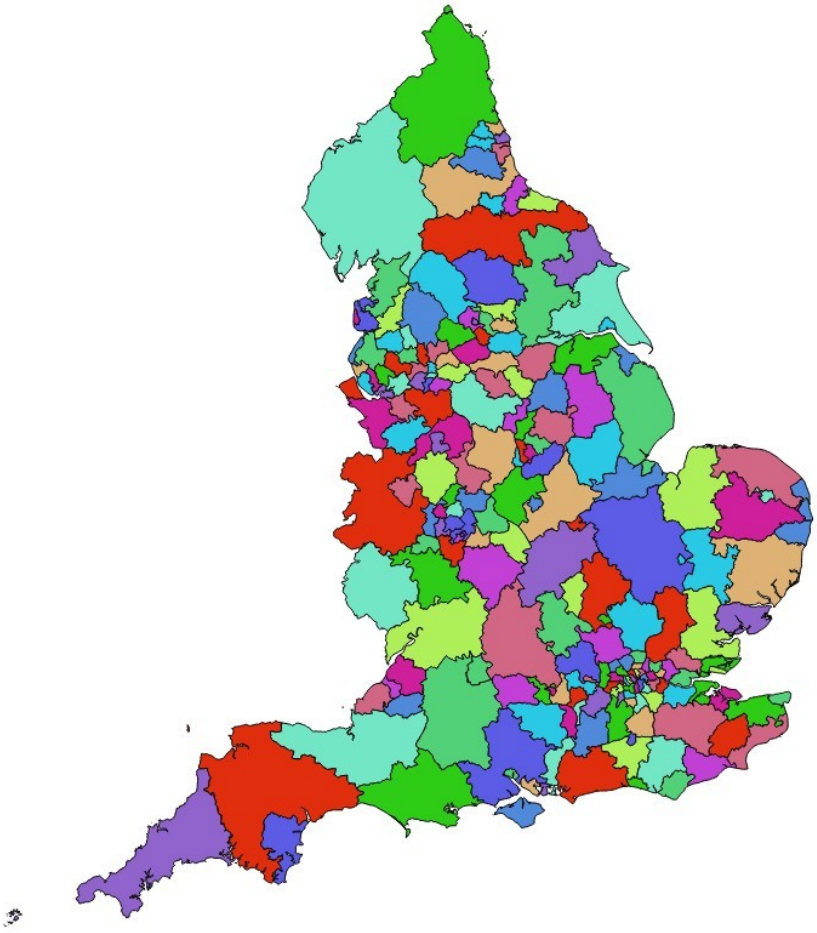


Figure 3.5: *This figure shows the original CCG map (top) filling 18% of screen space and the output with 60% space filling and 6.6% error(bottom). The QGIS color map is used[41].*

space percentage is reached. The algorithm can also increase the size of each node based on any property of the region (or proportion to a fixed maximum size region), e.g. the local population of the CCG like a traditional cartogram. Because we gradually increase the size of each CCG region node, the relative geo-spatial position between nodes is preserved. After the area of each square is increased by a small amount (1 pixel by default) some adjacent nodes may overlap. We then update the position of each node in the tree by running the fast node overlap removal algorithm[179, 180] described in the next section. We provide an animation to present the incremental processing from 1 pixel to maximum size. Slingsby et al. demonstrate the benefit of animation in this context [157].

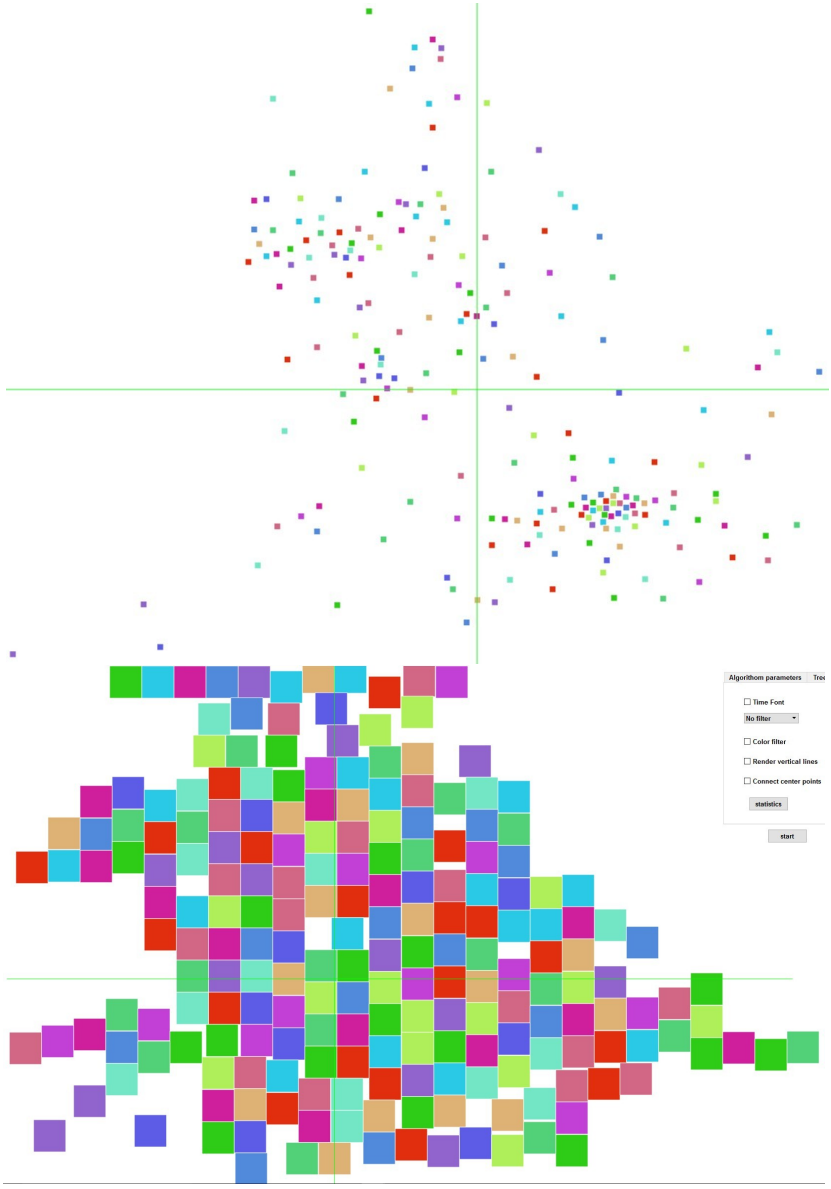


Figure 3.6: *This figure shows the resulting region node layout with 1% error (top) and the output with 60% space filling and 6.6% error(bottom). These use the QGIS color map [41].*

3.3.2 Updating Region Node Position

We use the fast node overlap removal algorithm presented by Dwyer et al. [179, 180] for removing overlap between neighboring region nodes. With this algorithm, the overlap is reduced in the quickest, most effective way. That means if a node, n, overlaps with its northern neighbor, n_n, running this algorithm shifts n south or its neighbor n_n north (similarly in the east-west orientation), the most effective way to remove overlap. By constraining the overlap to a small area, the relative position of adjacent nodes is preserved. If we increase all nodes to their maximum size before running the overlap removal algorithm, relative geo-spatial position of region nodes is not preserved either. The reason for this is when a node (n) is much smaller than its neighbor (n_n), it may lie completely inside its neighbor after its size has expanded to its maximum. In this case, it is faster to reduce overlap without preserving relative position.

The fast node overlap removal algorithm has two phases. In the first phase a number of constraints are applied that derive the separation distance between nodes. In the second phase, the solution is searched based on location as close as possible to the original node positions [179]. To address relative geo-spatial position preservation, we run the fast node overlap removal algorithm incrementally. In each pass, we increase the size of nodes by 1 unit and run the fast node overlap removal algorithm. In this way, the algorithm removes overlap and preserves relative position. The process is repeated until all nodes have reached their maximum size or a user specified error threshold is reached. (Some examples are shown below.) We can also animate the region growing process in order to increase the legibility of the visualization. Please see the accompanying video for a demonstration. Observing the evolution of each region provides benefit [157].

3.3.3 A Neighborhood Preservation Error Metric

We introduce a novel neighborhood preservation error metric that objectively quantifies how closely the relative geo-spatial positions of the resulting nodes correspond to their original positions. In other words, a west neighbor n_a should remain west of a given node after the layout is updated. Likewise for the east, north, and south directions. We consider it as an error when the relative geo-spatial position of the region center points cross. We use global error, e_g, to record any two center points crossing while we use local error, e_l, to record center points crossing when the distance between two center region points is less than a user specific threshold in Euclidean space, e.g. 20% of screen space.

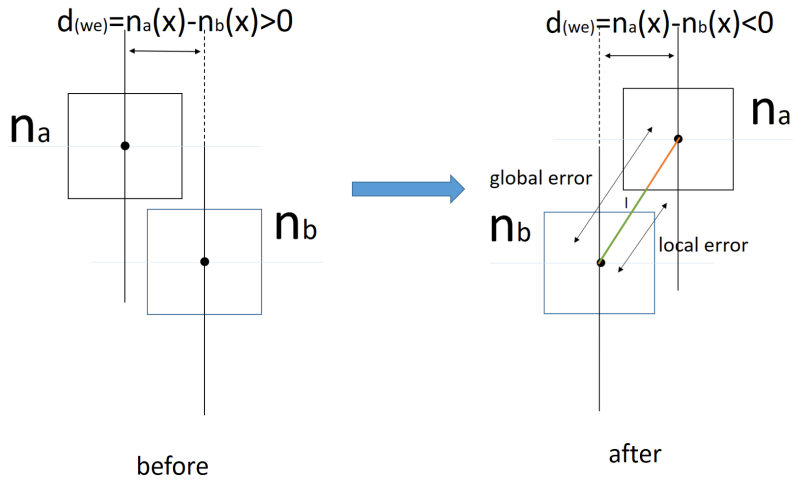


Figure 3.7: *The illustration of global and local error for neighborhood preservation. The error distance is decoupled into x(west-east) and y(north-south) components. The x components is illustrated here.*

As shown in Figure 3.7, we focus on the relative position of the center points of regions n_a and n_b. After looping through the layout algorithm, an error is counted if the longitudinal line of n_b crosses the longitudinal (along y) line of n_a. i.e. the longitudinal distance $d_{(we)}$=$n_a(x)$ -$n_b(x) > 0$ initially and $d_{(we)}$=$n_a(x)$ -$n_b(x) < 0$ after updating the node positions. That means the relative longitudinal positions of n_a and n_b are not preserved, thus we count this case as one error, similarly for the north-south orientation/position. If the total distance between the centers of n_a and n_b is less than a user specific distance, we consider this error as local error, e_l. We consider the worst-case scenario or maximum geo-spatial error when the whole map is flipped both latitudinally and longitudinally, similar to the worst case of bubble sort $O(n^2)$. Figure 3.8 shows an actual depiction of this error. We want to distinguish between local and global error because local error is more important in this context.

We consider the worst-case scenario when the center of every region node n crosses every other region node, $n-1$. We adopt the result that $n+(n-1)+(n-2)+...+1 = n(n+1)/2$. In our case n is 209, however node n cannot cross itself. Thus we use $n(n-1)/2$ as our worst case result. The worst-case number of crossings in our application is 21736. And all error can be expressed as a percentage of this total.

We do not claim that this is the best distance metric in all of the literature. Ghoniem et al. [178] and Nusrat and Kobourov [148] provide a comprehensive review and comparison of distance and error metrics for cartograms and spatial treemaps. In fact many of those could

be substituted here. We call this metric "novel" because this error metric is interactive as the user controls the level of error. For the first time the user controls the trade-off between filled screen space and relative error of geo-spatial position. Because we are focusing on preserving geo-spatial information with better screen space usage efficiency. This metric is directly fitted for our requirement. This metric is certainly not the only way to measure error. This is not main focus of this chapter, see Nusrat and Kobourov [148] for a survey of error metric.

3.3.4 Ordered Treemap Algorithm

After the size and position of each CCG region node is computed, a treemap node layout algorithm is used to visualize the non-spatial, multivariate health indicator data within each CCG. We require this data layed out consistently for each CCG region node to facilitate comparison between CCGs. Ordered treemap algorithms create rectangles in a visual order that matches the input order of the data. Bederson et al. [182] present two algorithms to display ordered treemaps: A Pivot treemap and the Strip treemap algorithm. Compared to the Pivot treemap algorithm, the Strip treemap results in a lower rectangular aspect ratio. This version is more squarified with a higher readability score. So we choose the Strip treemap algorithm to present data inside each individual CCG node. The Strip Treemap Algorithm of Bederson et al. [182] is described as below.
Input: Rectangle, r, to be subdivided into a list of items with area, l_1 to l_n.
Output: List of rectangles, r_1 to r_n

1. Scale the area of all the items on the input list so that the total area of the input equals that of the layout rectangle.

2. Create a new empty strip, the current strip.

3. Add the next rectangle l_i to the current strip, recomputing the height of the strip based on the area of all the rectangles within the strip as a percentage of the total layout area, and then recomputing the width of each rectangle.

4. If the average aspect ratio of the current strip has increased as a result of adding the rectangle, in step 3, remove the rectangle, pushing it back onto the list of rectangles to process and go to step 2. When the rectangle is removed from a strip, restore that strip to its previous state.

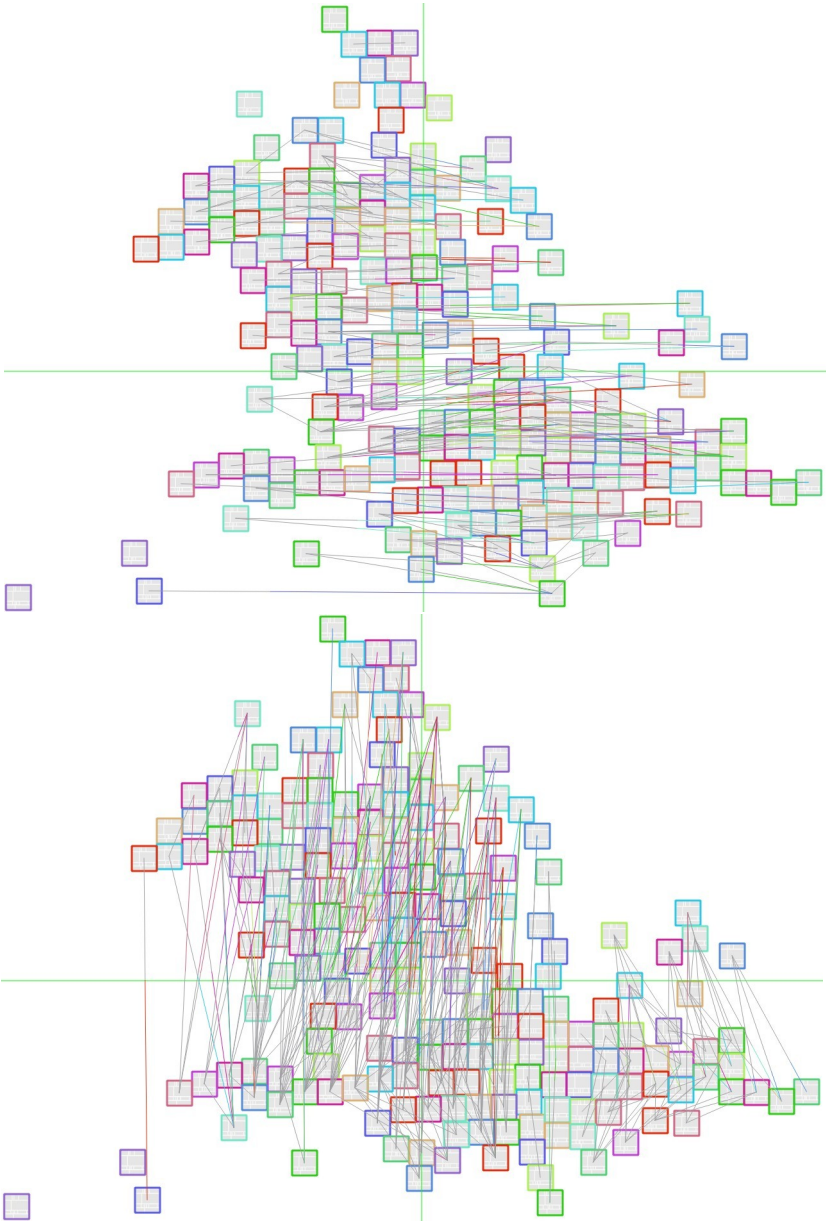


Figure 3.8: *Visualization of errors: Here we show what the geo-spatial error looks like. This figure shows error crossing edges in north and south orientation (top), in west and east orientation (bottom). The screen space-filling percentage, s, is 20% and e_l is 0.9%, and e_g is 1.8%.*

5. If all the rectangles have are processed, stop. Else, go to step 3.

3.3.5 Interactive User Options

For further exploration and analysis, several user options are available to explore and present the results focusing on different requirements such as filling the maximum space, specifying the local or global error, animating the node layout algorithm, modifying layout parameters, region selection for detail, modifying the color legend, and exploring the hierarchy.

Geo-spatial Error and CCG Region Node Size As our goal is to combine the geo-spatial properties of cartograms with the space filling properties of treemaps, the first user controlled parameter setting is the maximum geo-spatial error of the CCG regions. All CCG region sizes are uniform by default in order to facilitate comparison between regions. However, their size can also be proportional to the maximum sized region. The size of each CCG region can be mapped to the size of its local population or any health data indicator like a traditional cartogram. So we enable the user to set the maximum size of the region with the largest population and the other regions are adjusted relative to the maximum. As in Figure 3.9.

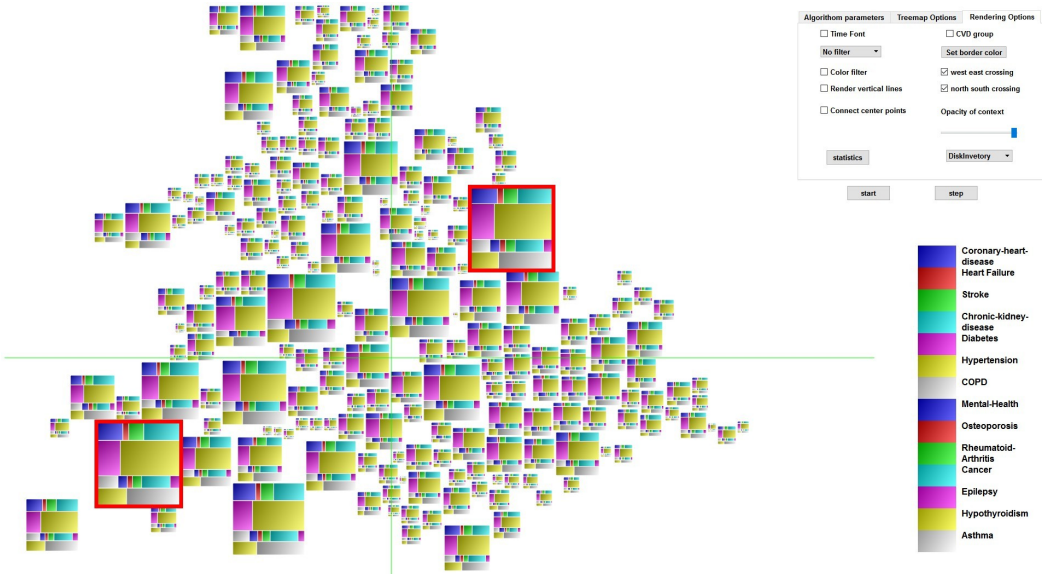


Figure 3.9: *Nodes proportional to CCG size. The screen space-filling percentage, s=36% and e_l=2.4%, e_g = 4.5. The two red outlines show the two biggest region nodes on the map: Cambridgeshire Peterborough and North East & West Devon. This is unexpected since we hypothesized the largest regions to be in London or Birmingham. This example uses color map from the Disk Inventory X tool [42].*

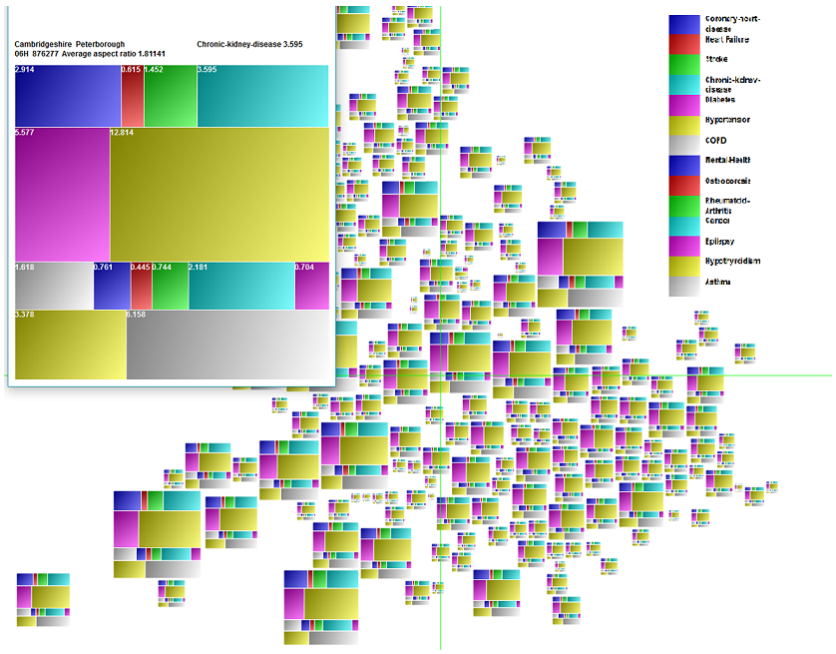


Figure 3.10: *This visualization shows the output of cartographic treemap with region size proportional to population, and with a details-on-demand window for one region node. s=30%, e_l=2.4% and e_g= 5.1%. The first three rectangles in each region node represent three CVD health disorders. Note the prevalence of hypertension and diabetes is very widespread the UK. This type of multivariate observation display itself clearly with this type of visualization.*

Node Size Increment and Animation In the cartographic treemap layout algorithm, the region size grows incrementally. As discussed in section 3.3.2, immediately increasing the node size to its maximum does not preserve the geo-spatial relationship between regions as well, while iterative increments take more time to generate the final result. So we provide a user option to explore an ideal size of area increase in a single layout algorithm pass. The increment size is set between 1 and 10 pixels. The layout takes more time when the increment size is small, but the accuracy of geo-spatial neighborhood relationships is increased. There is a trade-off between processing speed and accuracy of the geo-spatial relationship between nodes. A user option of animating the region node layout process is provided so the user can observe the correspondence between the original node position and the final visualization. Slingsby et al. [157] demonstrate the value of animation. The multi-pass layout algorithm is shown gradually from initial to final layout.

Uniform Size Regions A cartographic treemap node for a single region represents the

prevalence of various health disorders. As the size of each CCG region may be uniform or represent its population, the size of bottom level rectangles represents the proportion of the population with a particular health disorder in the respective region. We can get an overview of the prevalence of various diseases in CCG regions, as in Figure 3.10. However, as the population sometimes varies greatly among CCG regions, the size of bottom level rectangles may not be directly compared with other CCG region nodes. For example, a large population of heart failure in Oxfordshire CCG may not indicate heart failure there is relatively prevalent. The prevalence of heart failure in Oxfordshire is 0.51 which is lower than the average of 0.73. In order to facilitate direct comparison of health disorders across CCG region nodes, we provide a user option to generate uniform size region-level nodes set to true by default. In this way, the size of rectangles at the bottom level of the treemap hierarchy can be compared directly. As in Figure 3.11.

Difference Cartographic Treemap and Focus+Context To make the health care visualization clearer, we introduce a user option: a difference cartographic treemap. The size of each rectangle at the bottom level of the health care treemap does not represent the absolute prevalence value of each health disorder. Instead, it represents the difference from the average UK value. Using this option, we can emphasize how the prevalence of a specific health condition differs from the national average level and understand the conditions in a particular region. As in Figure 3.12. We also use a focus+context visualization incorporating a focus+context color map. The user may choose to focus on above average or below average values by user options. Focus attributes are then rendered in color while context rectangles are rendered in grey-scale. As in Figures 3.15 and 3.1.

Area Groups We introduce area groups to classify CCG regions into 27 area groups in the treemap hierarchy based on area code. This option creates a more space-filling cartogram and another hierarchy level in the treemap. It facilitates comparison of CCG regions health care data within their own CCG groups and enables exploration and analysis. As in Figure 3.13. It also results in a more space-filling layout with greater resemblance to a traditional treemap.

Details-on-Demand and on-mouse-over: For the finest (lowest) level of data detail in CCG regions or area groups, a details-on-demand feature is implemented. By hovering the mouse over or clicking on any region, a new window opens with a higher resolution treemap, providing the CCG code, CCG name and value of each health diagnosis category. As in Figure 3.10. To improve the appearance, we also add user options for various color maps and color

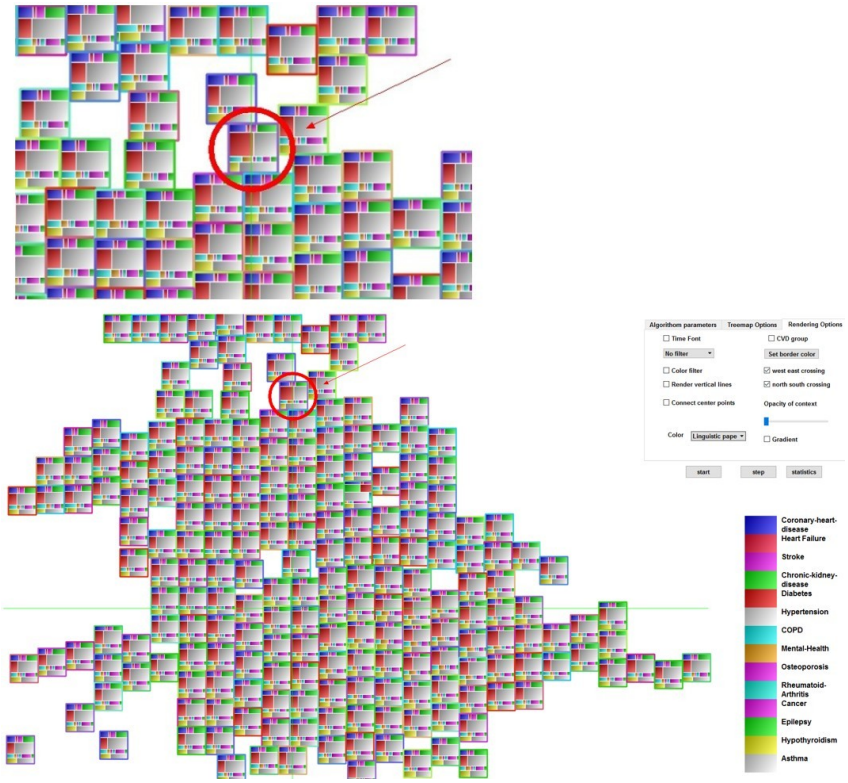


Figure 3.11: *This graph shows the output of cartographic treemap with uniform size region nodes. s=50% and e_l=2.4%, and e_g=5.8%. The region with the red circle (Bradford City) contains the largest purple rectangle which indicates the highest relative prevalence of diabetes in the UK. This example uses a published color-map from Setlur and Stone [43].*

gradient styles (See Figure in supplementary file). The color maps come from different sources; one is from the disk inventory X tool [42], the second one is from ColorBrewer [44], the third one is from Telea [45], the fourth is from QGIS [41], and the last one is from Setlur and Stone's paper [43]. As in Figure 3.11.

3.4 A Narrative of UK Population Healthcare Data

In this section, we present the results of our interactive metrics and derive a number of observations based on cartographic treemaps.

Accompanying Demonstration Video URL

https://vimeo.com/199637583

Evaluation of Space and Error Metrics To evaluate the performance of our algorithm,

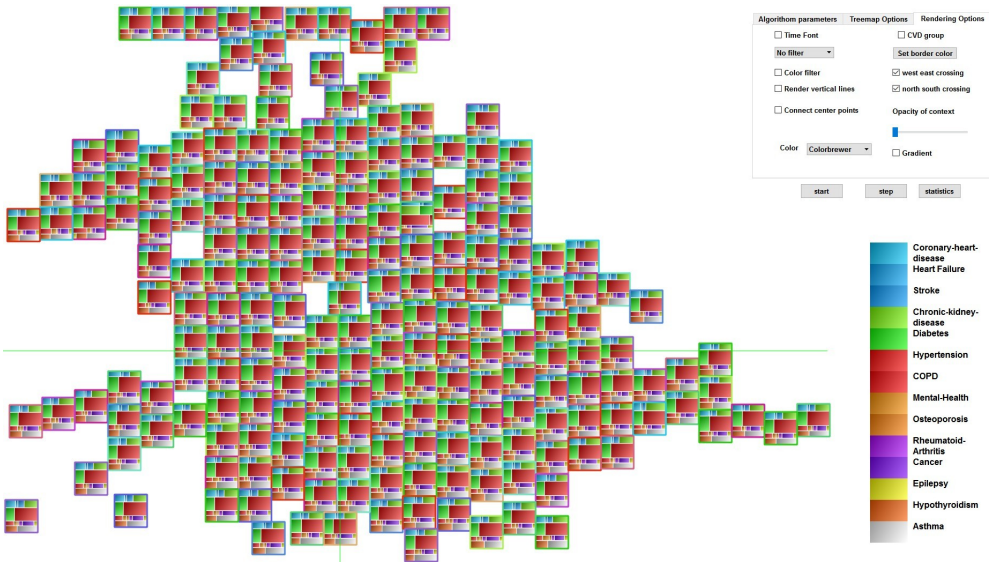


Figure 3.12: *This graph shows the cartographic treemap using average difference maps. s=50%, e_l=2.4%, and e_g=5.8%. The larger a bottom level rectangle is, the more it deviates from the UK average. This example uses a well-known color map from color-brewer [44].*

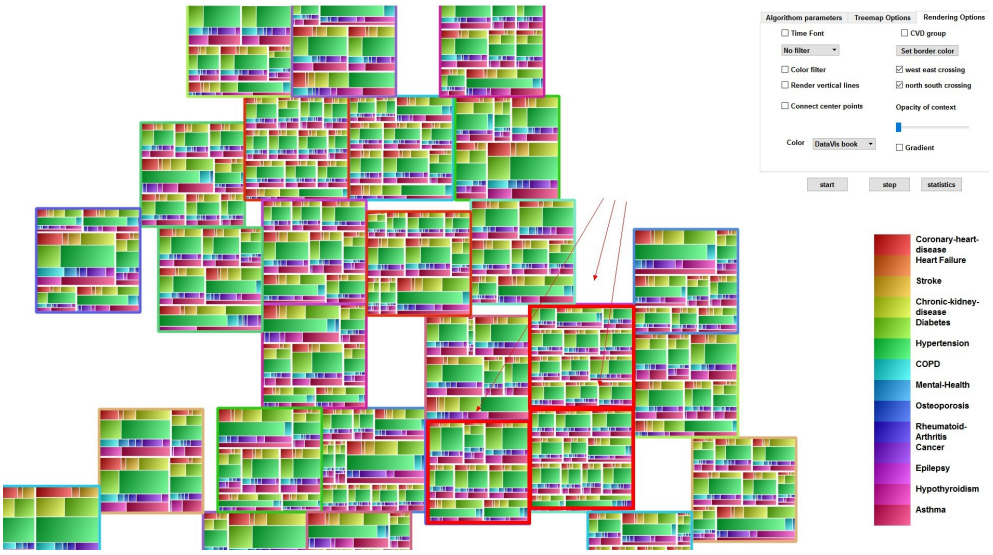


Figure 3.13: *This graph shows the cartographic treemap with 27 area groups. s= 70% and e_g = 5.2% . The regions in red highlights are London areas. This example uses Telea's color map [45].*

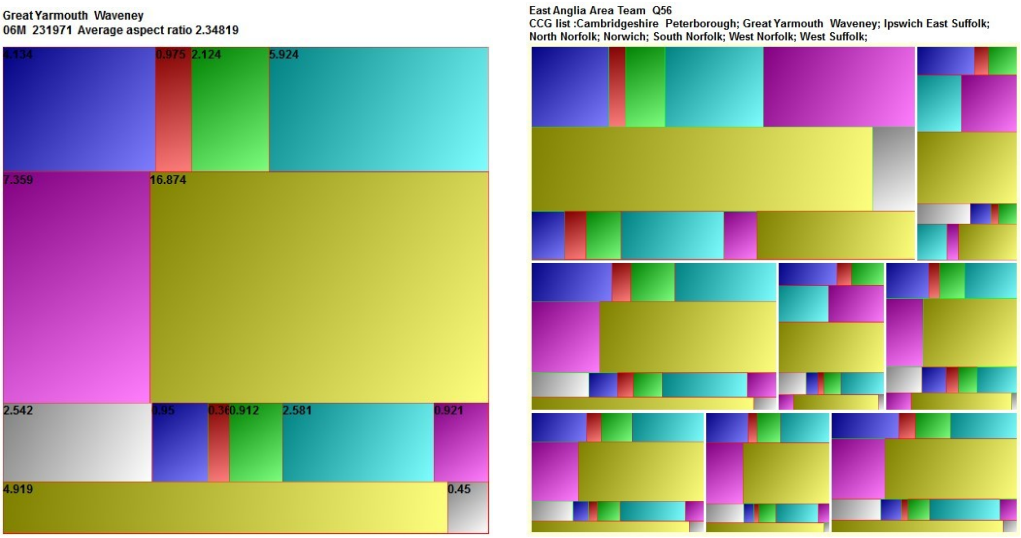


Figure 3.14: This figure shows the details-on-demand output map of one region (left) and detailed output of one area group (right).

Table 3.1: Neighborhood Preservation Metric

s	e_l	local error frequency	e_g	global error frequency
10%	0.4	164	0.7	293
20%	0.7	308	1.5	667
35%	0.9	409	2.5	1073
57%	1.1	476	3.1	1369
66%	1.2	524	3.6	1593

we measure the percentage of filled screen-space, s, versus the local and global geo-spatial error. As the original map is narrow, the space filled with respect to the screen is 18.5% and by using our algorithm the percentage of filled screen can reach up to 70%. The relationship between error and screen space filled is shown in Figure 3.17. We experimented with aspect ratios that are very common to commodity displays including 16:9, 4:3.

Based on the algorithm described in section 3.3.3, the local and global error is shown in Table 3.1 and Figure 3.7. It shows the connection between e_l, e_g and s. It presents percent space filled along with local and global percentage and frequency of center-axis crossings. We can see that e_l increases linearly with s occupied while e_g increases more rapidly. We can achieve 65% screen-space occupancy with only 1-4% global error.

Performance and Observation The algorithm requires less than a second to run (85ms-

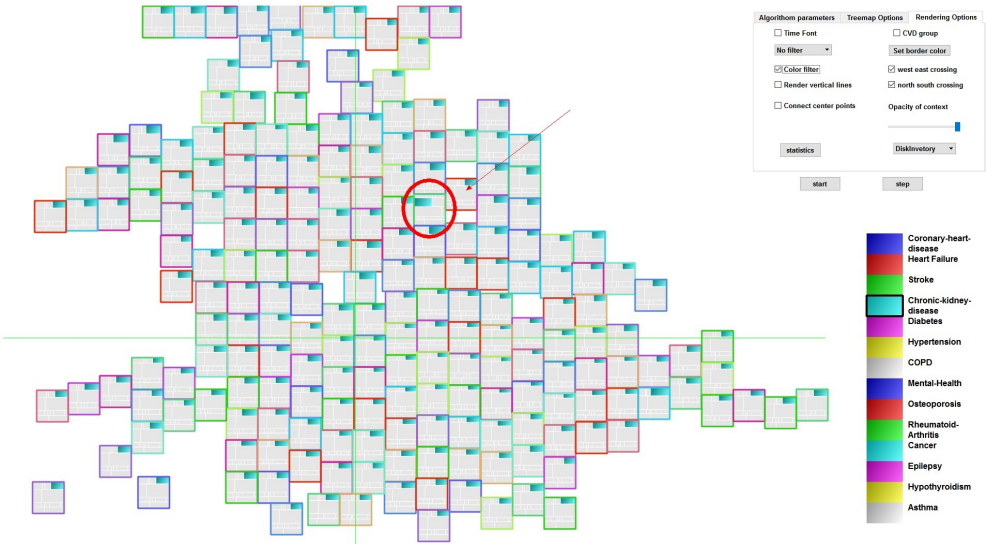


Figure 3.15: *A focus+context cartographic treemap visualization with uniform size regions. s=50%, e_l=2.4%, and e_g=5.8%. The data is mapped to two color scales: one for the focus data and the other for context. All the health care prevalence categories are shown as context except for user selected data attributes. The red circle shows the relatively largest rectangle in the map that represents the highest prevalence of Chronic-kidney-disease disorder in the UK (Nottingham North And East).*

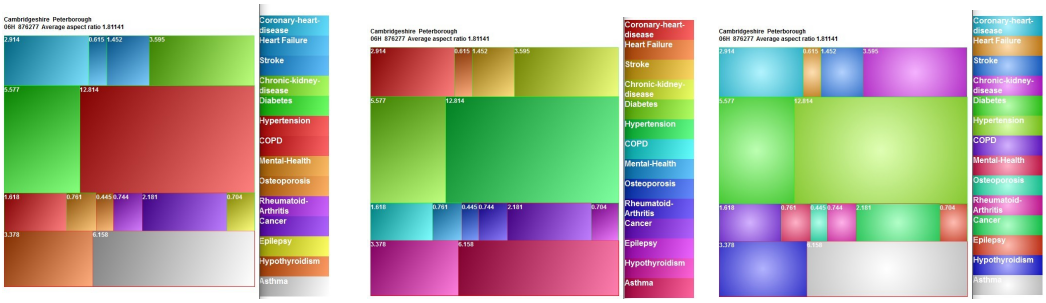


Figure 3.16: *This figure illustrates some different color and gradient mapping options. The color legend of the left treemap is from ColorBrewer [44] . The middle one is from Telea [45]. The right one is from QGIS [41] with an added color gradient.*

1000ms). The computer used to run this algorithm is a MSI desktop with i7-4770 CPU, 8g RAM, GeForce GTX 770 GPU and Windows OS. We slow it down for purposes of animation and user observation.

Based on the cartographic treemap visualization, several observations can be derived from the public health data. Several of these observations would be very difficult without the cartographic treemap.

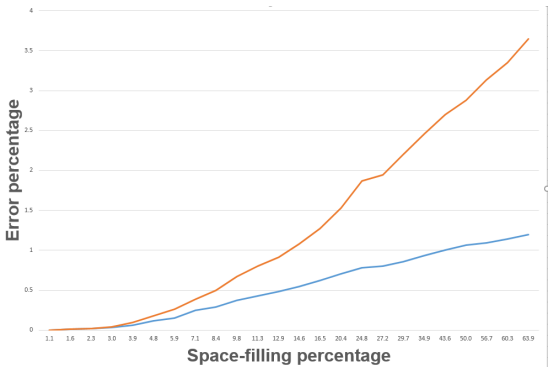


Figure 3.17: *This figure shows the relationship between percentage of both local and global error versus the amount of filled space. The red line shows the global error while the blue line indicates the local error.*

1. From the region node layout in Figure 3.13, we can see that the London area contains the most CCG group regions (32 in total) and the largest population.

2. The individual CCG regions with largest population are Cambridgeshire Peterborough and North East & West Devon. See Figure 3.9. This is not what we would expect but rather the largest populations in a London CCG.

3. Hypertension is most prevalent health disorder with the largest proportion throughout the UK. The second largest health disorder is Diabetes. See Figure 3.10. This is clear from an overview cartographic treemap.

4. Three kinds of CVD related disorders (Coronary-heart-disease, Heart Failure, Stroke) are prevalent throughout the UK, and coronary heart disease is the most common disorder in the CVD disorder group (a multivariate observation). See Figure 3.10.

5. From the uniform size nodes, the regions with a significantly higher prevalence health disorder can easily be observed. Bradford City has the relatively highest diabetes in the UK. See Figure 3.11. Also, we can find highest relative Chronic-kidney-disease disorder prevalence in the Nottingham North & East CCG. See Figure 3.15. And the highest relative mental health disorder prevalence is found in Islington.

6. Compared to the average value across all health disorders, regions in London are generally better than the average in most health categories with the exceptions of mental health and diabetes. See Figure 3.1. This is another multivariate observation.

7. CCGs closer to the coast have lower asthma. See Figure 3.1.

8. The Northwest regions are higher than average in most health disorders, such as, Cumbria and Northumberland. The values are higher than the average for a range of health disorders. For example, diabetes is more prevalent in Northern regions than Southern regions. This is shown in Figure 3.1. Cartographic Treemaps facilitate these kind of multivariate observations.

3.5 Health Science Domain Expert Feedback

This software is targeted at domain experts in healthcare analytic and not the general population. Therefore, no general user-study is performed. The domain experts are from the Medical School at Swansea University. One is professor and chair in applied statistics. And the other is a senior research officer at medical school of Swansea University. Domain expert 1: ”Data analysts are often required to analyse complex sets of spatial, multivariate, longitudinal, and event history public health data in order to answer research questions as part of major studies such as CORTEX, ELAStiC and the Carmarthenshire Housing Project. Cartographic treemaps facilitate the recognition of patterns within the data such as geographical clustering and temporal trends, as well as the identification of salient features including outliers and extreme values, thereby helping to complement machine learning and data mining techniques and to inform statistical modelling. This visualization will make a major contribution towards helping data analysts to achieve their research objectives. Therefore, we are delighted that this new technique will be utilised by data analysts in the Farr Institute @ CIPHER within Swansea University Medical School. We are confident that the cartographic treemaps will provide data analysts with the opportunity to gain additional deeper insights into their complex public health care data.”

Domain expert 2: ”Some of the biggest challenges of working with linked population health datasets relate to the sheer volume of the data: the scale is daunting in terms of the population sizes, and dimensionality. There are thousands of potentially interesting facts stored in various data sources. The depth and breadth of the data make it hard to see the big picture of what is going on in a population, as well as to sort through the noise to identify what information is relevant. These challenges are multiplied if the data is to be used directly in a clinical setting by people who are not expert analysts. Something that is necessary to derive maximum

benefit from available data resources. Visualization is a key technology to help users, both academic and clinical, make sense of the data. The cartographic treemap approach described here addresses our challenges by allowing a number of related variables to be presented simultaneously. Geography is often an important dimension in health research and service planning, and this technique allows data to be organized geospatially while transcending some of the limitations of traditional map-based visualizations. The ability to see geography, population sizes, and several health measures at the same time will help users get a much more accurate, at-a-glance understanding of the data and the population it represents. It has potential to aid research, particular in the hypothesis-generation phase; and it could be quite beneficial in the healthcare sector, supporting activities such as service planning."

3.6 Summary

This chapter presents a novel hybrid visualization, the Cartographic Treemap, combining geo-spatial information, a novel interactive neighborhood preservation metric, and space-efficient geometry for the interactive visualization of geo-spatial, and high-dimensional data. It combines the advantages of both cartograms and treemaps. We implement and demonstrate this visualization with a real-world high-dimensional health care data collected by NHS to support clinical commissioning groups (CCGs) and the health care service providers. Several interactive user options are available to explore and present the results focusing on different user requirements for further exploration, analysis and comparison. Also, we present several multivariate observations based on the cartographic treemap visualization and report feedback from two domain experts in health science.

This chapter fits the linear narrative and memorability column in Table 2.1, and geo-spatial, narrative and memorability column in Table 2.2. It use narrative visualization to link observations generated from the cartographic treemap, generate story board and present healthcare data to the audience. We aim to use storytelling and narrative visualization to increase the memorability of audience.

Future work includes investigating more optional color maps for high-dimensional data and a more in-depth user feedback study. Future work will include more attributes of NHS data in addition to population and health disorder prevalence, such as the number of practices per CCG, and rates of A&E admissions. More filtering options will also be introduced, such as by age range. Also the population and health disorder prevalence dynamics over time will

be presented in next chapter.

Chapter 4

Time-Oriented Cartographic Treemaps

Contents

"If I have seen further it is by standing on the shoulders of Giants."-Isaac Newton[1]

[1] Isaac Newton (1642-1726) was an English mathematician, astronomer, theologian, author and physicist who is widely recognised as one of the most influential scientists of all time, and a key figure in the scientific revolution.

While the previous chapter focus on multivariate visualization combining geo-spatial data, in this Chapter we extend the work by adding time-oriented data. Cartographic treemaps offer a way to explore and present hierarchical multi-variate data that combines the space-efficient advantages of treemaps for the display of hierarchical data together with relative geo-spatial location from maps in the form of a modified cartogram. They offer users a space-efficient overview of the complex, multi-variate data coupled with the relative geo-spatial location to enable and facilitate exploration, analysis, and comparison. In this chapter, we introduce time as an additional attribute, in order to develop time-oriented cartographic treemaps. We design, implement and compare a range of visual layout options highlighting advantages and disadvantage of each. We apply the method to the study of UK-centric electronic health records data as a case study. We use the results to explore the trends of a range of health diagnoses in each UK health care region over multiple years exploiting both static and animated visual designs. We provide several examples and user options to evaluate the performance in exploration, analysis, and comparison. We also report the reaction of domain experts from health science. This Chapter is based on paper "Time-oriented Cartographic Treemap for Visualization of Publice Health Care Data" [183].

4.1 Introduction

The Cartographic Treemap, combines geo-spatial information, a novel interactive neighborhood preservation metric, and space-filling geometry for the interactive visualization of geo-spatial, and high-dimensional data [47]. As a hybrid visualization, it combines the advantages of both cartograms and treemaps. We implement and demonstrate this visual design with real-world high-dimensional health care data collected by the NHS to support clinical commissioning groups (CCGs) and health care service providers.

In this chapter, we extend cartographic treemaps with time as an additional variate, in order to develop time-oriented cartographic treemaps. *Because the data is varying year-on-year, domain experts are very interested to see trends over time. In term of the outcome, domain experts are able to get an overview of the changes over years on several diseases. It will help the domain experts to figure out which places or which disease need more investment.* Based on a three year time span of health care data collecting by the NHS in the England, UK, we present and compare a range of visual design options highlighting advantages and disadvantages of each. We provide several user options to evaluate the performance in exploration, analysis, and

comparison based on a given set of prerequisites and user tasks. Also, we can generate linear narrative geo-spatial visualization from the observation of our visual design. It will help the user increase the memorability of the data set. The contributions of this chapter include:

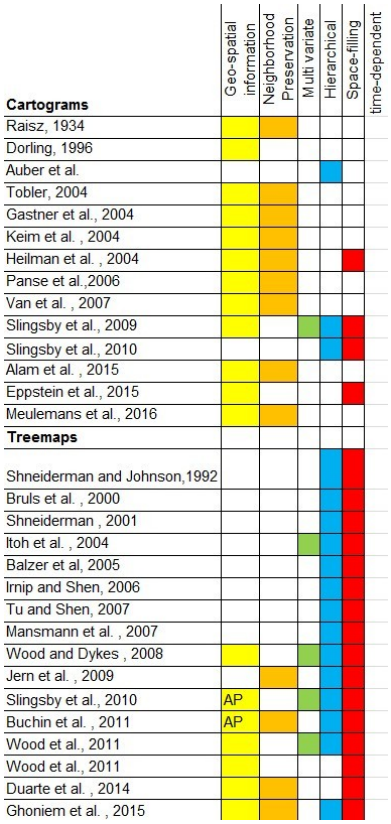
- A new time-oriented cartographic treemap that enables the user to explore hierarchical multi-variate data over a range of years.
- Both static and animated visual designs for cartographic treemaps: presenting the temporal trends of data.
- Interactive user-options that enable users to customize the visual layout.
- The application of our time based visualization to complex, real-world NHS data from England, UK.
- The reaction of domain experts from health science.

In order to achieve this, several challenges must be overcome. The first challenge is to develop several visual designs for incorporating time into cartographic treemaps. A second requirement is to compare the visual designs and present the relative advantages and disadvantages of each. Another is to provide user-options to facilitate both exploration, analysis, and comparison of time-dependent hierarchical, multi-variate UK-based health care data. This chapter extends the work of Tong et al. [47] by adding time as a variate.

The rest of the chapter is organized as follows. Section 4.2 presents a description of the time dependent UK-based NHS data. Section 4.3 presents several tasks and requirements for the visual design. Section 4.4 describes different visual designs and user options in exploration, analysis and comparison of time-dependent hierarchical, multi-variate data in a stand-alone application. Section 4.6 reports the reaction from health science domain experts. And the final section presents conclusions and future work within the field.

In previous work, we develop a layout algorithm for cartographic treemaps. We extend this to include time-variate data.

The work we present here differs from previous work in that it attempts to combine the space-filling, hierarchical characteristics of ordered space-filling treemaps together with the geo-spatial information conveyed by a cartogram. *It add time as a variate into the cartographic treemap. Domain experts are interested to see the trends over time as healthcare data*



Cartograms	Geo-spatial information	Neighborhood Preservation	Multi variate	Hierarchical	Space-filling	time-dependent
Raisz, 1934						
Dorling, 1996						
Auber et al.						
Tobler, 2004						
Gastner et al., 2004						
Keim et al. , 2004						
Heilman et al. , 2004						
Panse et al.,2006						
Van et al. , 2007						
Slingsby et al., 2009						
Slingsby et al., 2010						
Alam et al. , 2015						
Eppstein et al., 2015						
Meulemans et al., 2016						
Treemaps						
Shneiderman and Johnson,1992						
Bruls et al. , 2000						
Shneiderman , 2001						
Itoh et al. , 2004						
Balzer et al, 2005						
Irnip and Shen, 2006						
Tu and Shen, 2007						
Mansmann et al. , 2007						
Wood and Dykes , 2008						
Jern et al. , 2009						
Slingsby et al., 2010	AP					
Buchin et al. , 2011	AP					
Wood et al., 2011						
Wood et al., 2011						
Duarte et al. , 2014						
Ghoniem et al. , 2015						

Figure 4.1: *This table shows characteristics of related work. It includes six visualization properties: geo-spatial information, neighborhood preservation, multi-variate, hierarchical, space-filling and time-dependent. Geo-spatial information implicates whether a visualization conveys geographic information and AP in the column represents adjacency preservation only. Neighborhood preservation indicates a algorithm that features a distance metric to preserve neighborhood relationships. Multi-variate indicates the dimensionality of abstract data. Hierarchical indicates a type of hierarchical data. Space-filling indicates how well the output visualization fills the screen. And time-dependent indicates whether a visualization contain time as an attributes. Our time-dependent cartographic treemaps feature all six properties.*

is varying over years. Table 4.1 compares the current work with the work presented here. No previous algorithm combines all six properties. Especially, no other works contain a time variate. Time-dependent Cartographic Treemaps convey geo-spatial information. They feature an error-driven distance metric between nodes. They visualize multi-variate hierarchical data. They give the user interactive control over how much screen space is used. And they present time-dependent information in several visual designs.

McNabb et al. [145] summarized two survey papers of temporal visualization. Cottam

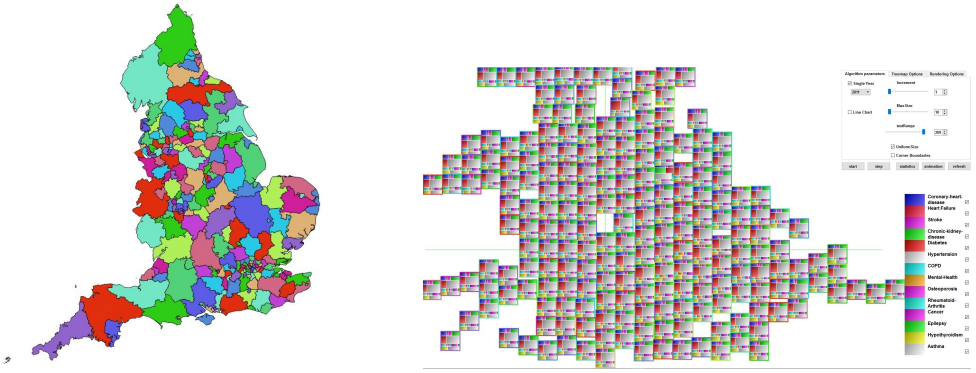


Figure 4.2: *The left map shows the original 209 CCG regions (Clinical Commissioning Groups) provided by Public Health England [46] (left). The original map only occupies 18% of screen space. The original visual design of cartographic treemap based on a single year (right) [47]. The cartographic treemap occupies 60% of screen space. This color map is from a published color-map from Setlur and Stone [43].*

et al. [184] dynamic visualizations as visualizations that change over time. They review the impact of dynamic data on Information Visualization, and how this data change can influence a visualization's discernability. Bach et al. [185] survey a variety of temporal data visualization techniques and discuss how their operations can be used with spacetime cubes in order to create a simple visualization from the 2D+time model.

4.2 Time-Oriented Public Health Care Data Description

We study open NHS health care data as a case study for time-oriented cartographic treemap visualization. The UK government collects yearly diagnoses of region specific health care data [46]. The public health profiles website [46] is used for publishing the latest national health care data in the England, UK. The data archive is designed to support GPs, clinical commissioning groups (CCGs), and local authorities to ensure that they provide and commission effective and appropriate health care services. See Figure 4.2. Typically this data is displayed using standard line charts, bar charts and pie charts. The standard visualizations do not feature any geo-spatial information. Also, time-related information is generally presented in isolation.

The dataset consists of 14 Excel files of around 10 Megabytes in total together with a CCG region map containing 209 regions (See Figure 3.3). There are more than 60,000 rows and an average of 100 columns in each file with three years data. We extract 14 health care disorders

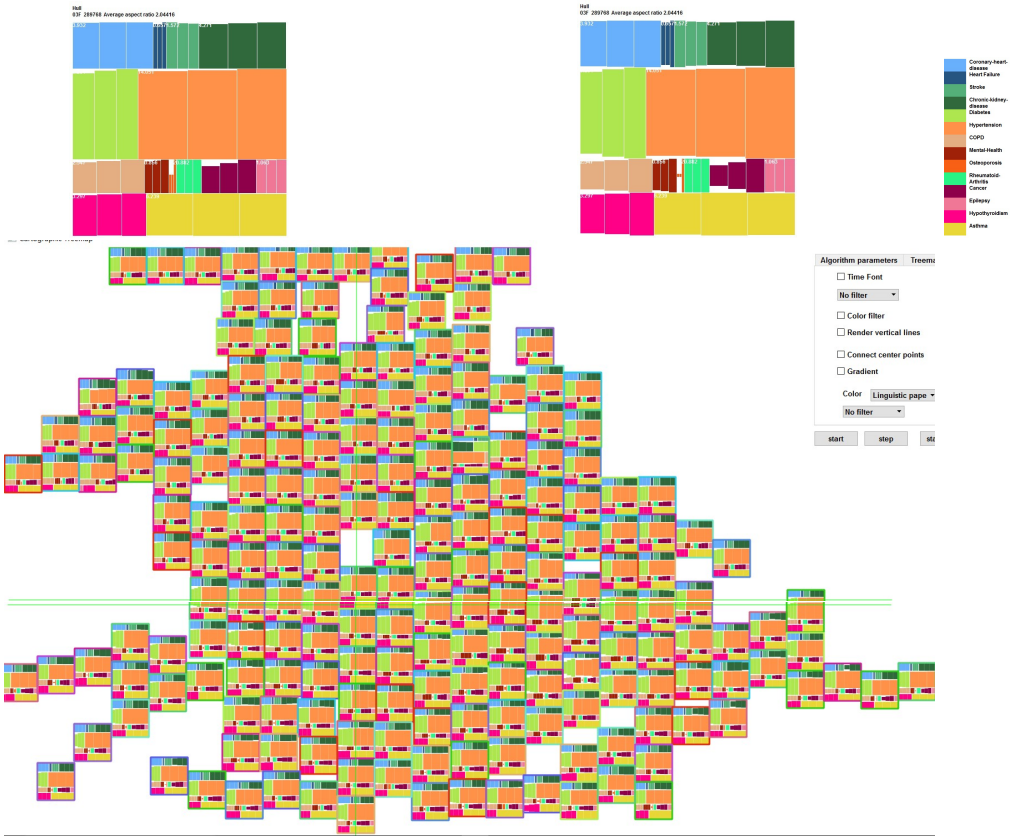


Figure 4.3: *This visualization shows the output of time-oriented cartographic treemaps with bar charts inside each health care variate, and with a details-on-demand window for one region node (top area of main map). It also shows the output of time-oriented cartographic treemaps with symmetric bar charts inside each health care variates (bottom half of UK cartogram), and with a details-on-demand window for one region node (top right). The three rectangles in each variates represent prevalence values over three years from 2011 to 2013. We observe that hypertension and diabetes are the most prevalent diagnoses over this time-period. The color map is derived from Colorgorical [48].*

over three years prevalence indicators 2011-2013 from the dataset and present the information in our time-oriented cartographic treemap system. ”The whole cartogram is resembles a treemap that represents a two-level hierarchy: geographical and various diagnoses in each box.”

Our goal is to combine hierarchical, multi-variate health care data with complex geo-spatial information using the cartographic treemap algorithm of Tong et al. [47] and add time-oriented trends in a unified visual design. The challenge is not only to show the overview of hierarchical, multi-variate health care data based on regional information, but also depict the temporal evolution trends of data inside each region. We use the NHS health care data from 2011 to 2013, and the NHS health care regions map as input.

4.3 Tasks and Requirements

The visual design of our application supports the following requirements and user tasks:

1. **T1**: To provide an overview, both temporal and spatial, of the prevalence rates for each diagnosis coupled with the geography.

2. **T2**: To provide selection and filtering options with a special focus on time-oriented trends, behaviors and patterns.

3. **T3**: To provide details on demand after exploration, filtering and selection have been performed.

These tasks mirror those outlined by Shneiderman[186] in 1996 and are customized for this particular setting.

4.4 Time-Oriented Cartographic Treemap

This section describes the visual designs we used to support tasks 1-3 by adding a time variate to previous cartographic treemaps. We use the previous cartographic treemap algorithm [47] for static data as our starting point and then implement several visual designs and user options for displaying time-oriented information in one visual system. The visual designs and user options are presented in the following subsections. First, we introduce time-oriented bar charts, symmetric bar charts, and gradient-oriented bar charts. We compare and discuss the relative

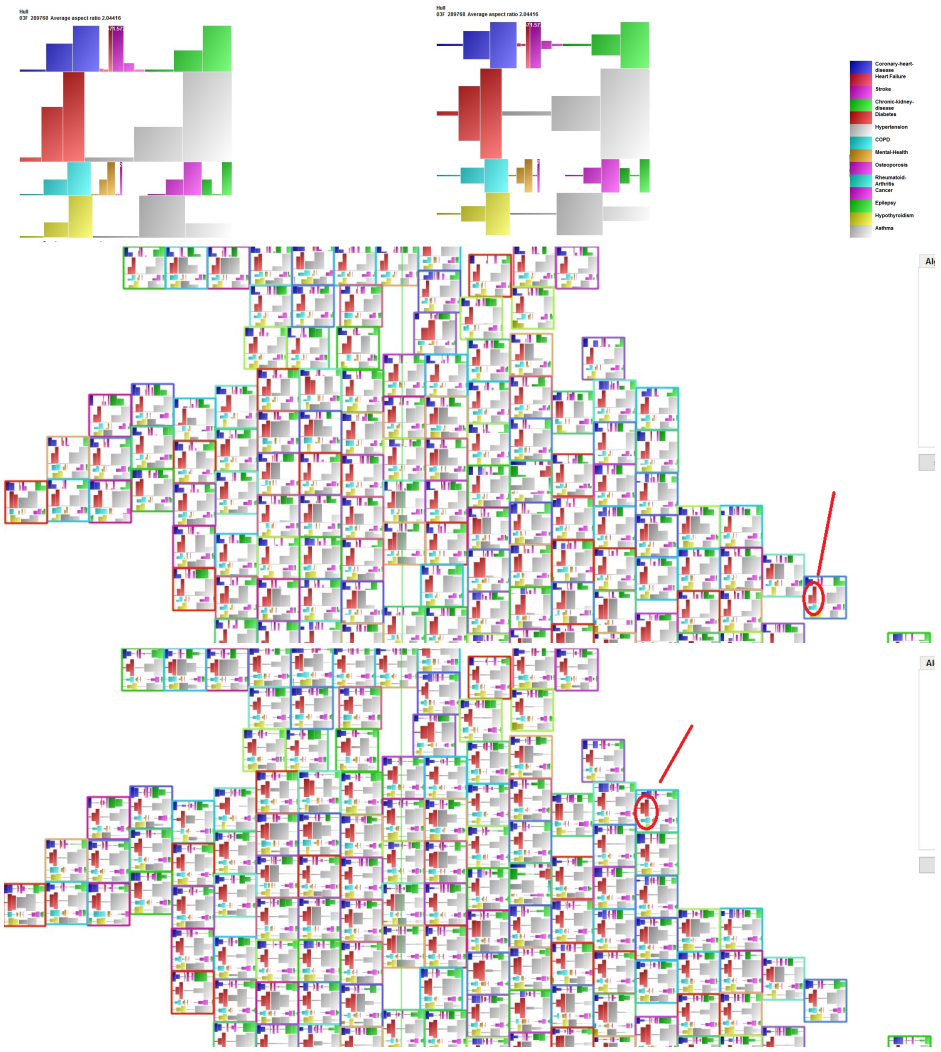


Figure 4.4: *This visualization shows the output of time-oriented cartographic treemaps with gradient-oriented bar charts (middle), and with a details-on-demand window for one region node (top left). It also shows the output of time-oriented cartographic treemap with the combinations of symmetric bar charts (bottom), and with a details-on-demand window for one region node (top right). Only the northern half of the UK is shown for presentation space purpose. The gradient-oriented bar charts really emphasize the increase in diabites over time. The visual design support task 1 and task 3.*

advantages and disadvantages of each. Then we add the option of animation, showing increasing versus decreasing diagnoses over time, we describe line charts and other user-options for further exploration including observations based on the visual designs. Finally, we develop an attribute selection option which enables the user to turn individual health care variates on or off.

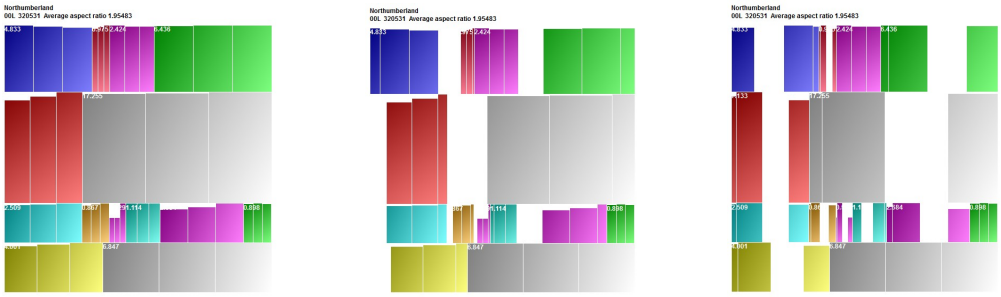


Figure 4.5: *This visualization shows three frames of the details-on-demond view using animation.*

4.4.1 Time-Oriented Bar Charts

One of good choices for mapping time to a visual primitive is using a bar chart. Bar charts are the most popular choice amongst authoring tools according to Lee et al. [187]. Bar chart is often used for retrieve value, and make comparisons. And bar chart is also an good combination with treemap and space-filling requirement. Each column can represent one year and one bar chart can represent the prevalence of each diagnosis. The bar chart is a traditional tool to visualize categorical data. We start off by using bar charts to display yearly data (2011-2013). Each bar chart fits inside the rectangular output of region node and treemap node from Figure 2 well. We integrate each bar chart into a single treemap node for displaying the temporal trend of each health care variate inside a single CCG region (See Figure 4.3). The result addresses task 1 by providing the user with an overview of the data.

The evolution of data over time is difficult to observe using standard bar charts, neither the bars nor the data vary in height very much. To make the difference between each bar more clear, we introduce **symmetric bars** as a modification to the traditional bar chart (See Figure 4.3). A symmetric bar chart varies the height of each bar from the top while raising the bottom of each bar by the same amount simultaneously. This emphasizes the differences between bars. A details-on-demand window for one region node showing a magnified view of the different style of bar charts is also provided. This supports task 3. By using two styles of bar charts, the time-oriented, hierarchical, multi-variate health care data 2011-2013 is presented in single visual design and an overview of yearly health care information can be derived from the output. The users can see both an overview of all regions and the details-on-demand for a single region. As we can observe from the result, hypertension is the most prevalent health disorder over the time-span with the largest proportion throughout the UK while the second most prevalent health

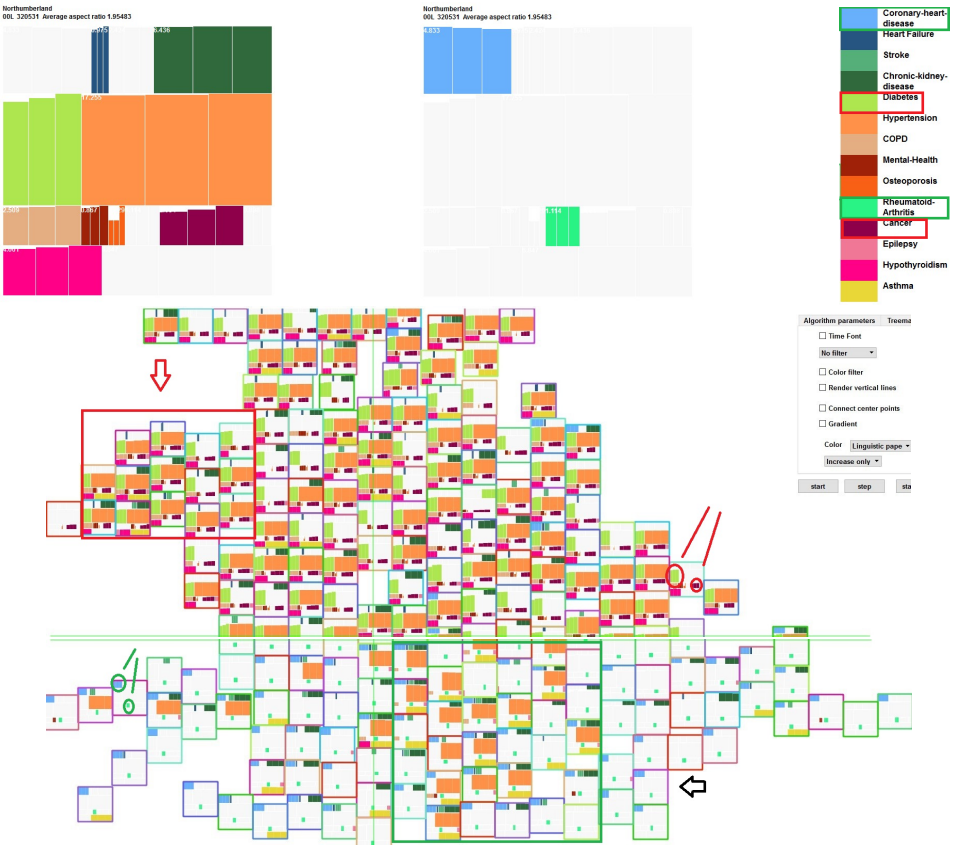


Figure 4.6: *This visualization shows the output of time-oriented cartographic treemaps with increasing only (top half) and decreasing only (bottom half) prevalance value filters to support task 2. Only the norhern half of the UK is displayed for increasing and southern half of the UK is displayed for decreasing values is shown for presentation space purposes. We can observe a region in the noth-east with a group of increasing health dignoses including storks, diabetes, rheumatoid, COPD, ostsoporosis, cancer, and hypothyoidism. Also the London region reports a decrease in hypertension. The color map is derived from Colorgorical [48].*

disorder during the time-period is diabetes. And both are generally increasing over time.

4.4.1.1 Gradient-Oriented Bar Charts

As the relative difference in height between bars over the three years is small, it is difficult to obtain a clear understanding of temporal trends inside each region from the previous visual design. We introduce a gradient-oriented version of the bar chart as a user option in order to highlight only the *changes* in prevalence rate during three years (See Figure 4.4). As opposed to the absolute values, in this version, the height of each bar represents the change between

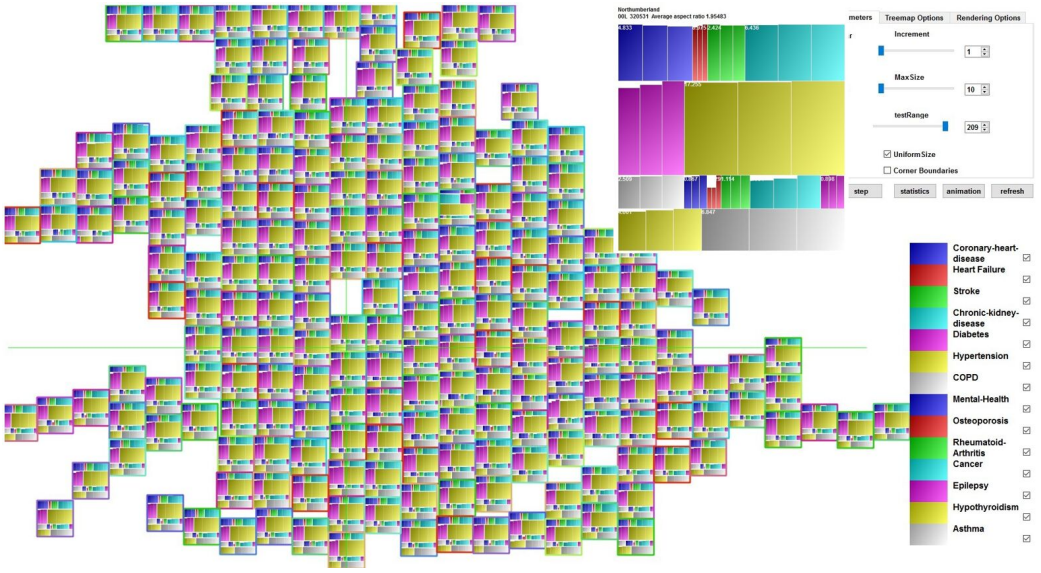


Figure 4.7: *This visualization shows the output of time-oriented cartographic treemap with bar charts inside each health care variates, and with a details-on-demand window for one region node. The three rectangles in each variates represent value of three years from 2011 to 2013.*

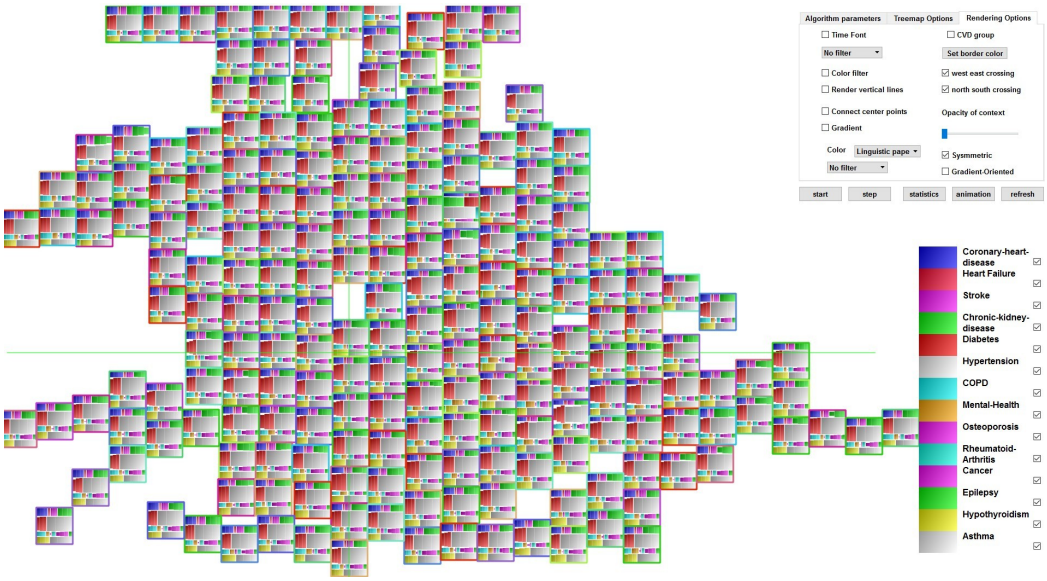


Figure 4.8: *This visualization shows the output of time-oriented cartographic treemap with sysmmetric bar charts inside each health care variates, and with a details-on-demand window for one region node. The three rectangles in each variates represent value of three years from 2011 to 2013.*

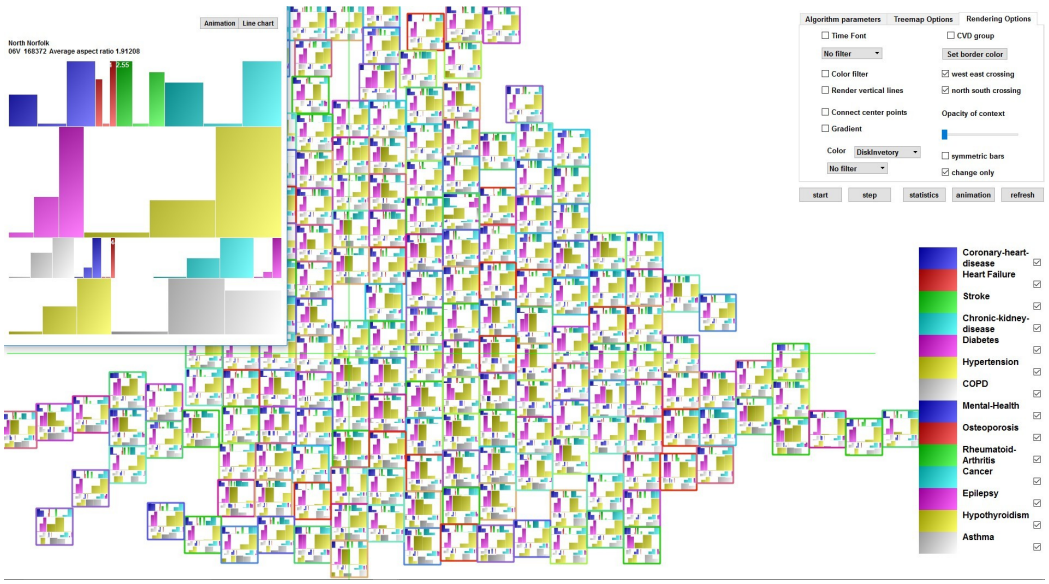


Figure 4.9: *This visualization shows the output of time-oriented cartographic treemap with change only display.*

minimum and maximum data values. Both the standard and symmetric bar charts can be used to depict the gradient information. The trends of increasing and decreasing diagnosis over time are depicted clearly from the gradient-oriented bar chars. The gradient-oriented bar charts really emphasize the increase of diabetes overtime. This supports task 1. However, with this design too much information is packed into a small area. Distinguishing increasing trends from decreasing trends is difficult. We introduce animation to further clarify the trends.

Symmetric, Gradient-Oriented Bar Charts Symmetric bar charts are also enabled in gradient-oriented user options to further highlight the difference between bars to reflect trends over time (See Figure 4.4). In this version, the changes in value over the three years are presented with heightened emphasis. An overview of trends for all regions and all health care variates can be obtained from this visual design. Because the changes in prevalence rates over time are exaggerated, the user is cautioned when interpreting the graphs. From gradient-oriented bar charts and symmetric gradient-oriented bar charts, the trend is increasing for the majority of health care diagnoses. From this visual design, we can observe that for a given CCG, e.g. Hull, all prevalence rates increase over time with the exception of asthma and stroke. This supports both tasks 1 and 3.

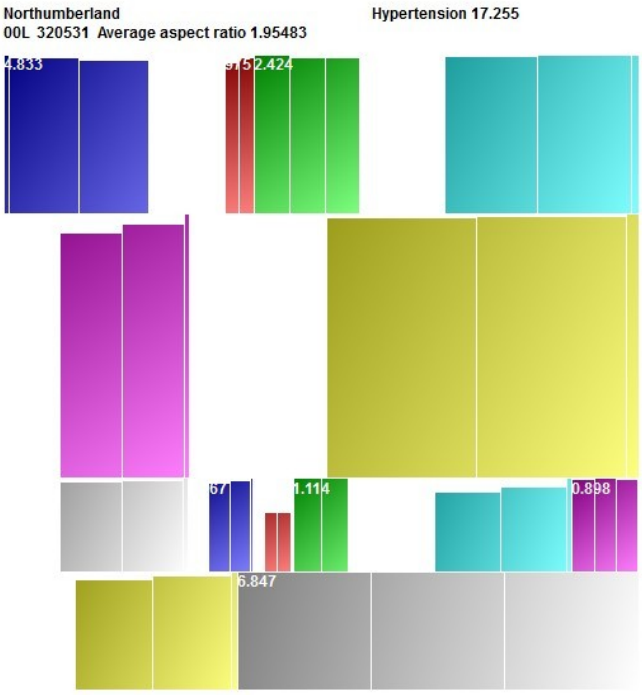


Figure 4.10: *This visualization shows the output of detail-on-demond view of animation.*

4.4.2 Animation

With bar charts, symmetric bar charts and gradient-oriented user options, the overview of time-oriented health care information is presented in various visual designs to support the domain expert user requirements. However, we can add another user option that distinguishes increasing trends from decreasing in the visual design display easily as an approach related to task 2. Thus we introduce an animation option to present increasing trends and decreasing values in different directions. See Figure 4.5, we animate the bars depicting increasing trends through translation from left-to-right. Decreasing trends are animated by translating the bars from right-to-left. A white gap is inserted between last and first year to ensure the users can decipher where the first bar is. From the animation, the trends of time-oriented values are emphasized even further. This supports task 1. In order to view the animation we encourage the reader to visit the video demonstration at `https://vimeo.com/223316576`.

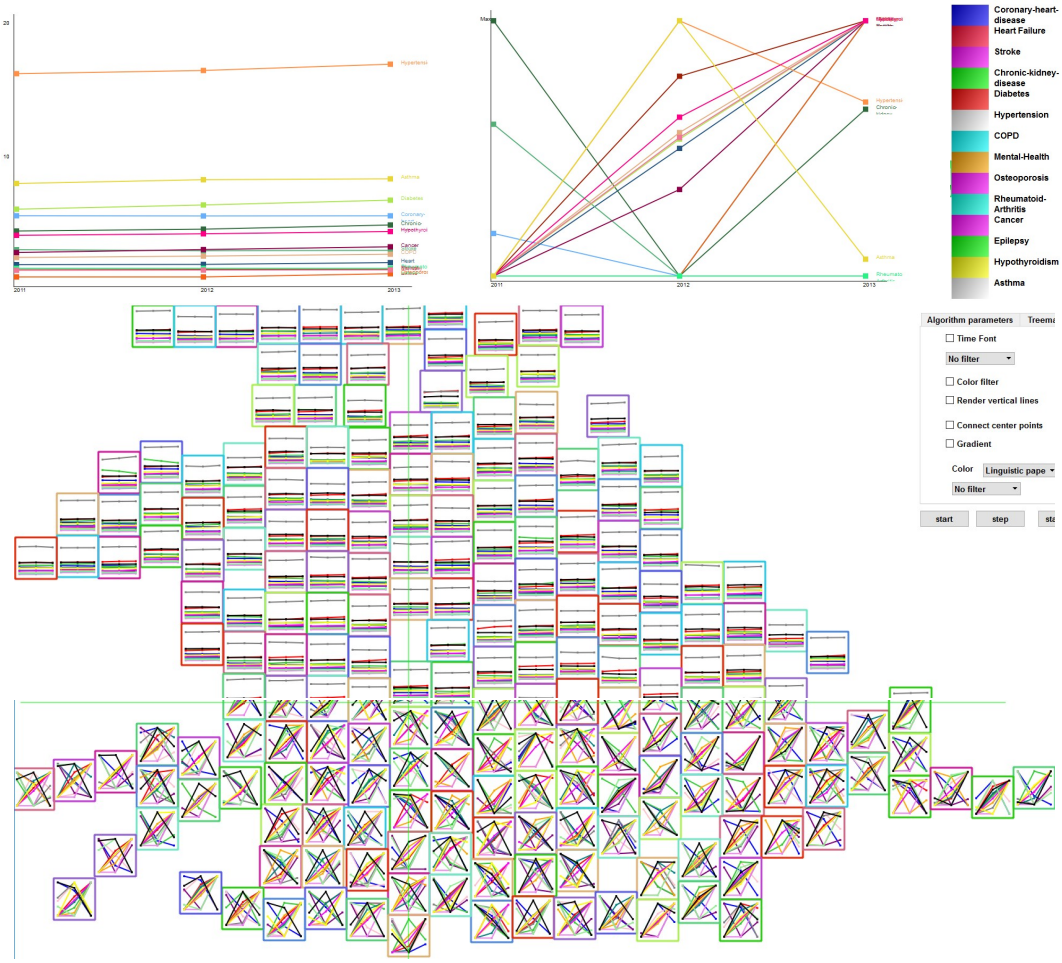


Figure 4.11: *This visualization shows the output of time-oriented cartographic treemaps with the line charts visual design user option (middle), and with a details-on-demand window for one region node (top left). It also shows the visual design with the gradient-oriented user option (bottom), and with a details-on-demand window for one region node (top right). Only the northern half of the UK and the southern half of the UK is shown for presentation space purposes.*

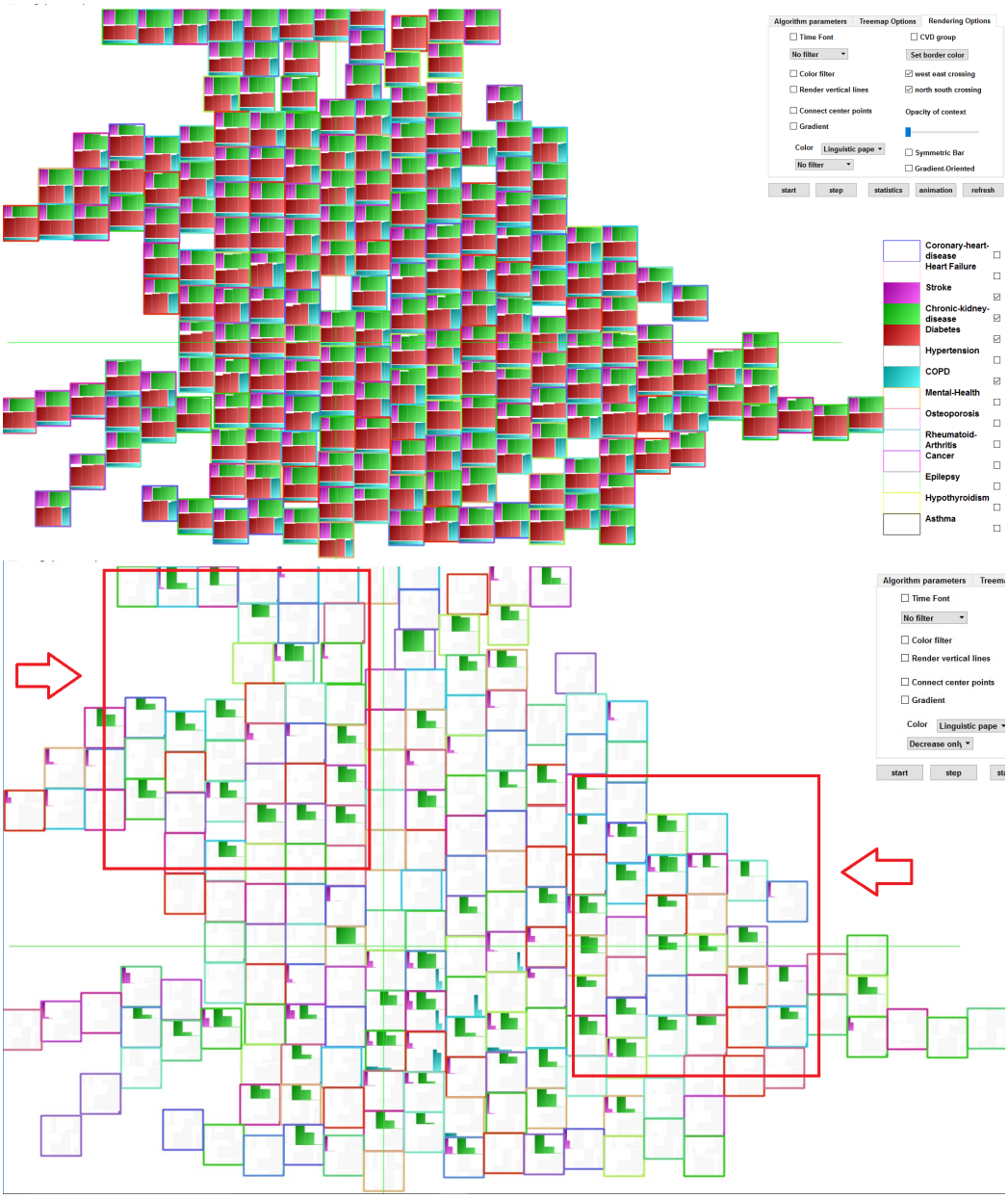


Figure 4.12: *This visualization shows the attributes selection user option to support task 2 with only four attibutes selected (top) and the decreasing only filter (bottom). We can observe that kidney disease is decreasing in the north west and the mid east of the UK.*

4.4.3 Filtering and Focus+Context Rendering

Even though we can obtain a direct overview of health care diagnosis trends from animation, animation requires video output to be observable. As an alternative, we implement filtering options based on increasing and decreasing prevalence rates combined with focus+context rendering options. Using these options, we can emphasize increasing and decreasing trends in the output visualization and support task 2. See Figure 4.6 and 4.13. From those figure, the user may choose to focus on increasing or decreasing diagnoses over time. Focus attributes are then rendered in color while context rectangles are rendered in grey-scale. And we may observe some useful patterns from the result. For example, most health care variates are increasing during 2011-2013, and Coronary-heart-disease is the most decreasing variate among 209 CCG regions except for the mid-east of England. Using animation and increasing and decreasing focus+context rendering user options, we can easily observe that coronary heart disease and rheumatoid-arthritis are the top two decreasing trends among CCG regions and approximately half of the hypertension diagnoses are decreasing too. The majority of diagnoses are increasing.

4.4.4 Line Charts

Bar charts are space-filling by nature and too many bars may crowd the display. Therefore we also experiment with line charts as an alternative visual design. We introduce line charts as a supplementary tool to simplify the time-oriented visualization. They also support task 1. By connecting a series of data points, line charts can present the trends of diagnoses occupying less visual design color and space. We implement line charts inside regions to replace the treemap layout (See Figure 4.11) as a user option. If we use standard line charts in a similar fashion as standard bar charts, it is difficult to observe trends. This is due to the very gradual change in diagnoses over time. Thus we incorporate a gradient-oriented version of the line chart as well. Gradient-oriented and details-on-demand user options are both provided for the line charts view. The user can filter and observe increasing and decreasing trends of all regions from overview and also focus on the details of a single region. As we can observe from line chart design, the increasing trends dominate diagnoses over time.

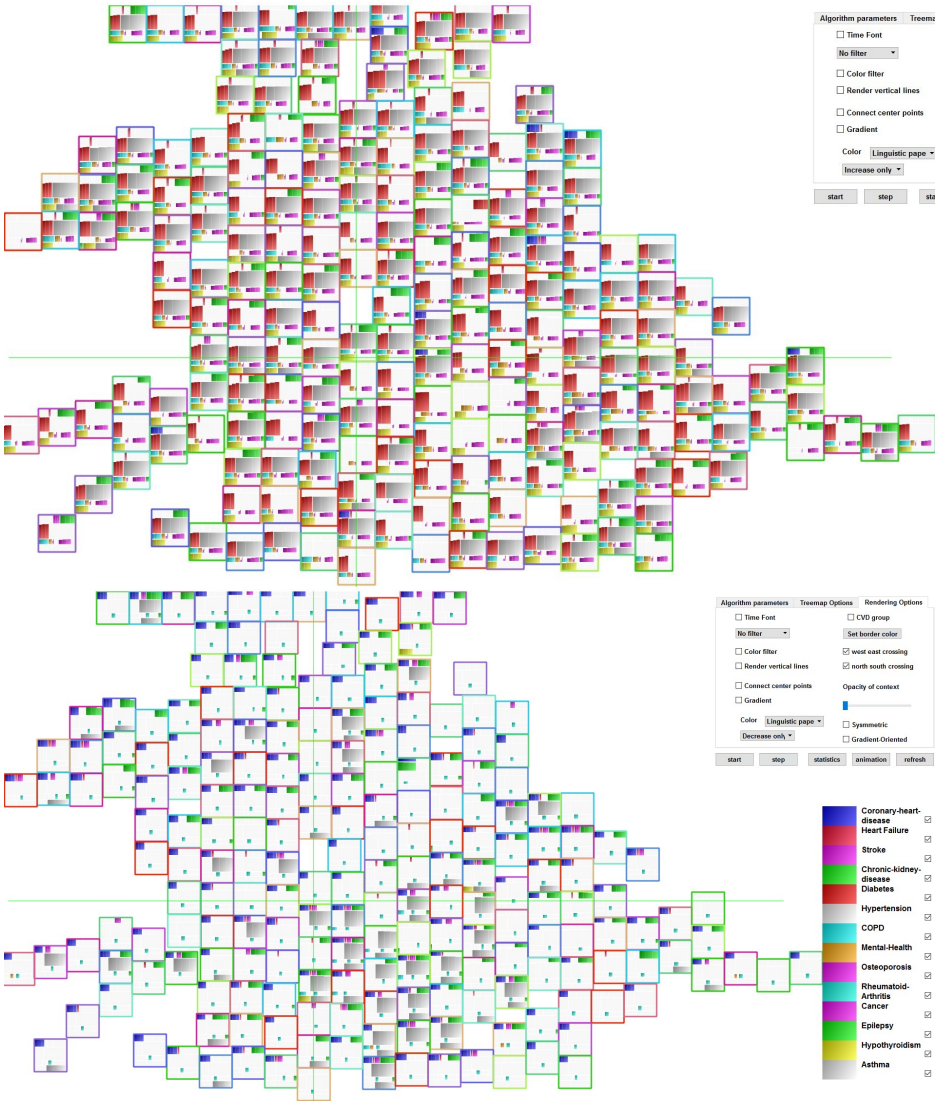


Figure 4.13: *This visualization shows the output of time-oriented cartographic treemaps with increasing only and decreasing only prevalance values filters. The selection user option is shown in focus, while other attributes are left as context information.*

4.4.5 Interactive User-options

For further exploration and analysis, several user options are available, to explore and present the results focusing on different requirements such as choosing individual years and attributes collectively, which support task 2.

Choosing Years To simplify the standard output of the time-oriented cartographic treemaps,

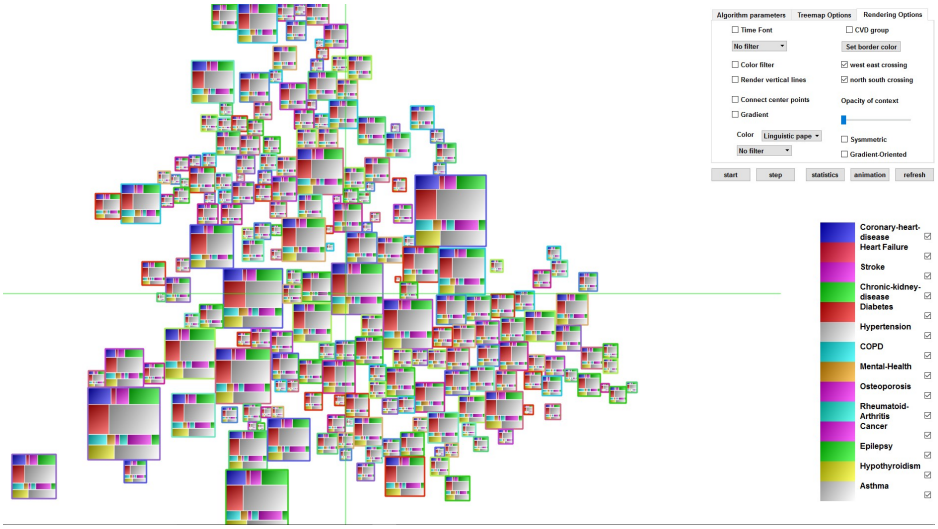


Figure 4.14: *This graph shows a single year with node size mapped to population. This color map is from a published color-map from Setlur and Stone [43].*

choosing individual years enables the user to focus on a single year of information rather than multiple years. The users can extract one year of information from the single year overview and switch between years and observe the differences over time.

The size of treemap nodes can mapped to the population of CCG regions. See Figure 4.14, choosing an individual year also enables the users to observe the changes to the population in 2011-2013.

Filtering Diagnoses For further simplifying the result and drawing the users attention to the information they require, we implement filtering attributes options. This enables the users to turn on and off specific attributes, and recompute the treemap layout with fewer attributes. In Figure 4.12, only four attributes are selected with an overview layout and details-on-demand output. The trends of only those four diagnosis in all CCG regions can be focussed on and observed more clearly. Figure 8 shows another important filtering option, depicting increasing or decreasing only prevalence values in a focus + context visual design style.

4.5 A Narrative of Time-oriented Population Healthcare Data

Based on the time-oriented cartographic treemap visualization, several observations can be derived from the public health care data.

1. Diabetes and hypertension are the most prevalent diagnoses over 2011-2013, as can be observed in figure 4.3.

2. Diabetes and cancer are increasing over time in most UK regions. See Figures 4.4 and 4.6.

3. Appropriately half of the CCGs exhibit increasing kidney disease over time while the other half exhibit kidney disease decreasing over time.

4. Coronary heart disease and Rheumatoid arthritis are decreasing over time in most UK regions. See Figure 4.6.

5. Kidney disease is decreasing in the north west and the mid east of the UK. See Figure 4.12.

6. A group of 11 connected CCGs in north west exhibit noticeable increase in Hypertension and diabetes. The CCGs regions are South Sefton, Liverpool, Blackpool, Southport and Formby, Knowsley, Fylde Wyre, St Helens, Halton, Bolton and Warrington. See Figure 4.6.

7. Hypertension is decreasing in the London area. The relevant CCG regions are Haringey, Islington, Wandsworth, Sutton, Herts Vallys, Richmond, Kingston, Surrey Downs, Brent, Hammersmith Fulham, Hounslow, North West Surrey, Guildford Waverley, Harrow and Ealing. See Figure 4.6.

These observations are consistent with our definition of storytelling in chapter 2. Because these serve to resort the results of exploration and analysis to a wider audience. It also fits the linear narrative and memorability column in Table 2.1, and geo-spatial, narrative and memorability column in Table 2.2.

4.6 Domain Expert Feedback from Health Science

This software is developed for a specific domain expert audience. Therefore we study their feedback rather than conducting a general user-study. A full general user-study would require a new thesis chapter in future work. The domain experts are from the Medical School at Swansea University. They professor and chair in applied statistics or senior research officer, or

research officer and data scientist or honorary research associate. The following is feedback directly from collaborators in health science. Time-oriented data, which are variously known as repeated, longitudinal or event history data, present analysts with a range of challenges. These issues become even more challenging when the data also vary spatially. The authors of this chapter have developed an eye-catching interactive tool with which data analysts may use animation (please see our later comments) to explore spatial and temporal trends in the values of one or more attributes, as well as to identify salient features such as outliers or extreme values.

We feel that potential users of this tool would require some guidance on using the various facilities, for example, filters to query the data, and exporting the equivalent numerical summaries into table or output format. Advice would also be welcomed on interpreting the visualizations in an efficient and effective manner. For example, the developers of this tool have implemented an algorithm that maximises the use of space by distorting the original shape of the outline of the area under scrutiny. Users will need to be advised on how best to avoid becoming disorientated by this particular feature of the tool. This guidance may need to vary depending on the user group, for example, data analysts compared to clinicians.

We envisage a wide range of possible applications for this tool. The authors of the current chapter have used animation to represent time. By using animation, the developers of this visualization tool have injected an element of dynamism into the analytical process, thereby enhancing the exploratory analysis of spatial longitudinal data.

Chapter 5

Cartograms with Features

Contents

"If you can improve yourself in a day, do so each day, forever building on improvement."-Zhu Xi[1]

[1] Zhu Xi (1130-1200) was a Chinese philosopher, politician, and writer of the Song dynasty.

This Chapter present a novel algorithm that introducing more geo-spatial information by adding features into cartogram. Cartograms are very popular and useful for depicting data on a map. Dorling style and rectangular cartograms are very good for facilitating comparisons between unit areas. Each unit area is represented by the same shape such as a circle or rectangle, and the uniformity in shapes facilitates comparative judgment. However, the layout of these more abstract shapes may also simultaneously reduce the map's legibility and increase error. When we integrate univariate data into a cartogram, the recognizability of cartogram may be reduced. There is a trade-off between information recognition and geo-information accuracy. This is the inspiration behind the work we present. We thus attempt to increase the map's recognizability and reduce error by introducing topological features into the cartographic map. Our goal is to include topological geographic features such as a river in a Dorling-style or rectangular cartogram to make the visual layout more recognizable, increase map cognition and reduce geo-spatial error. We believe that compared to the standard Dorling and rectangular style cartogram, adding topological features provides familiar geo-spatial cues and flexibility to enhance the recognizability of a cartogram. This Chapter is based on paper "Cartogram with Topological Features" [188]

5.1 Introduction and Motivation

Cartograms are a very popular and useful technique for depicting geo-spatial data. We summaries previous work on cartograms in previous chapter. It includes different type of cartograms and the main feature of them. None of those work introduced additional geo-spatial feature into the cartograms. (See Figures 3.2 and 4.1 for an overview of cartogram literature.) A cartogram can be defined as, "*a technique for displaying geographic information by resizing a map's regions according to a statistical parameter in a way that still preserves the map's recognizability.*" [146] According to Nusrat and Kobourov [148], cartograms can be categorized into four types: contiguous, non-contiguous, Dorling and rectangular. Dorling [153] style and rectangular cartograms are very good for facilitating comparisons between unit areas. Each unit area is represented by the same shape such as a circle or rectangle, and the uniformity in shapes facilitates comparative judgment. However, the layout of these more abstract shapes may also simultaneously reduce the map's legibility and increase error. When we integrate univariate data into a cartogram, the recognizability of cartogram may be reduced. There is a trade-off between information recognition and geo-information accuracy. This is the inspira-

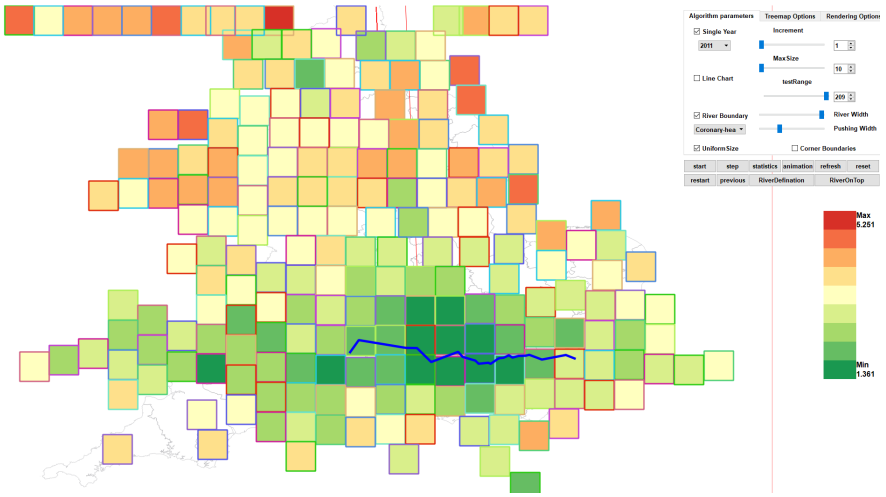


Figure 5.1: *A cartogram with the Thames river, featuring a wide push width, p_ε and a river width, r_ε of 10 pixels. Color is mapped to Coronary heart disease distribution in England.*

tion behind the work we present. We thus attempt to increase the map's recognizability and reduce error by introducing topological features into the cartographic map.

In this Chapter, we use the term topology slightly different from the traditional sense as with graphs and standard cartograms. In the context of graphs and cartograms, topology normally refers to the unit areas as nodes and the edge connections between them as their topology. In this work, we adapt the notion of topology from the flow visualization literature [189]. In flow topology, vector fields are divided up into different regions of flow behavior e.g. rotating flow versus linear flow. The edges or curves that separate and connect the different regions of the flow are referred to as the flow's topology. Hence, we introduce a (flow-inspired) topological feature into Dorling style and rectangular cartograms, i.e., a river. The river separates the unit areas into distinct regions and makes the cartogram more legible and reduces node layout error. This is analogous to a separatrix in the flow visualization literature [190], a standard feature in flow topology. A separatrix is a special type of streamline that connects two critical points and a curve that no flow crosses.

Our goal is to include topological geographic features such as a river in a Dorling-style or rectangular cartogram to make the visual layout more recognizable, increase map cognition and reduce geo-spatial error. We believe that compared to the standard Dorling and rectangular style cartogram, adding topological features provides familiar geo-spatial cues and flexibility to enhance the recognizability of a cartogram. For example, the Thames river can be converted

to topological feature on a cartogram. The regions on each side of the river on the real-world map are also often divided by this topological feature on a traditional map. This helps the user to locate corresponding region groups near the Thames river easily. We choose the Thames river base on the hypothesis that it is the most famous landmark or topological feature of the UK [191]. This is a test of the concept, and we can use other topological feature instead.

Our contributions include:

1. A new way to define and add real geographic features, such as a river, as a topological feature to a cartogram.

2. A novel cartographic layout algorithm that preserves a nearby region node's topological location with respect to a river.

3. The application of this cartogram design to real-world healthcare data provided by the NHS, England.

Previous chapters develop a Cartographic treemap to integrate a modified representation of the UK based on the geo-spatial information of CCG (Clinical commissioning group) regions combined with a modified treemap to present multivariate NHS data. They also present a metric to analyze the trade-off between space-filled and geo-spatial accuracy. To the best of our knowledge, the work we present here is the first of its kind to introduce topological features to Dorling-style and rectangular cartograms.

5.2 Adding Topological Features to Cartograms

This section describes the cartogram construction algorithm starting with an overview. The processing begins with reading the UK geo-spatial information. The algorithm summary is as follows:

1. Acquire and input selected river data.

2. Input geo-spatial data for cartogram generation.

3. Define a river approximation and add it to the geo-spatial data set that the cartogram is based on.

4. Compute unit area (or region) center points: We use the QGIS [41] tool to calculate the center points of each CCG (Clinical Commissioning Group) region. The center points are the starting positions of the rectangular region nodes.
5. Update node size: We start with a unit square to represent each CCG region as a node in the cartogram and gradually increase the size of each node.
6. Update cartographic layout: During the region growing process, regions may not cross topological features.
7. Test for river intersection. When a region intersects a topological feature. The layout algorithm returns this region to its previous position.
8. Remove overlap between nodes.
9. Test boundary conditions.
10. Render the cartogram with features.

Figure 5.2 shows an overview of the algorithm pipeline.

5.2.1 Input River Data

We search for and obtain the Thames river geo-spatial information from OpenStreetMap[192] for river shape, name and type. Then we use Overpass Turbo API[193] to build a query based

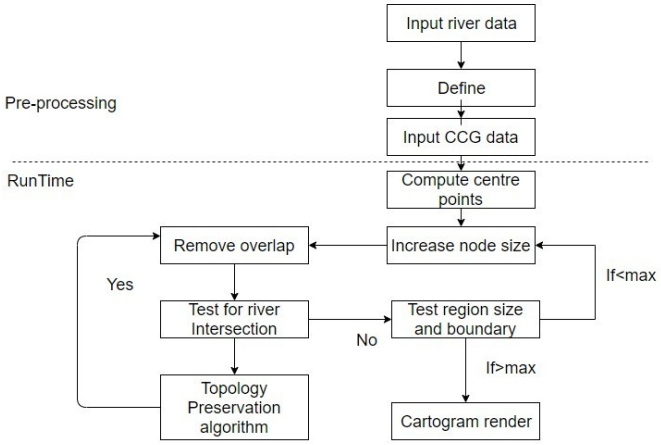


Figure 5.2: *The processing pipeline for producing a cartogram with topological features.*

on the search result and export the Thames river as a shapefile. We then input the Thames shapefile into QGIS[41] described in section 3.2 and combine it with a CCG map, and proceed to river approximation. The geo-spatial information is from a separate data source than our cartogram data for this special application which involves population healthcare prevalence values.

5.2.2 Input CCG Data

We study open NHS healthcare population data as a case study for our topology-based cartograms. The UK government collects yearly diagnoses of region-specific population healthcare data. The public health profiles website [46] is used for publishing the latest national healthcare data in England, UK. The data archive is designed to support GPs, clinical commissioning groups (CCGs), and local authorities to ensure that they provide and commission effective and appropriate healthcare services. Typically this data is displayed using standard line charts, bar charts and pie charts. The standard visualizations do not usually feature any geo-spatial information. The dataset consists of 14 Excel files of around 10 Megabytes in total together with a CCG region map containing 209 regions. There are more than 60,000 rows and an average of 100 columns in each file.

Our goal is to visualize this CCG data with cartograms, and make it more recognizable and reduce layout error by adding topological features. We use the NHS healthcare data and the NHS healthcare regions map as input.

5.2.3 River Definition and Approximation

Our goal is to include a topological geographic feature such as a river in a dorling-style cartogram to make the visual layout more recognizable, increase map cognition and reduce geo-spatial error introduced by the layout procedure. We approximate a river to match the cartogram style and to simplify testing for edge intersections (rather than using the original.) At first, a river approximation can be converted from geographic map positions to topological positions that separate the CCG regions on either side of the river. Our process for deriving a river approximation is as follows:

1. We overlay the river shapefile[192] and combine it with the CCG region map [40].

2. We identify the CCG regions on both sides along its full length.

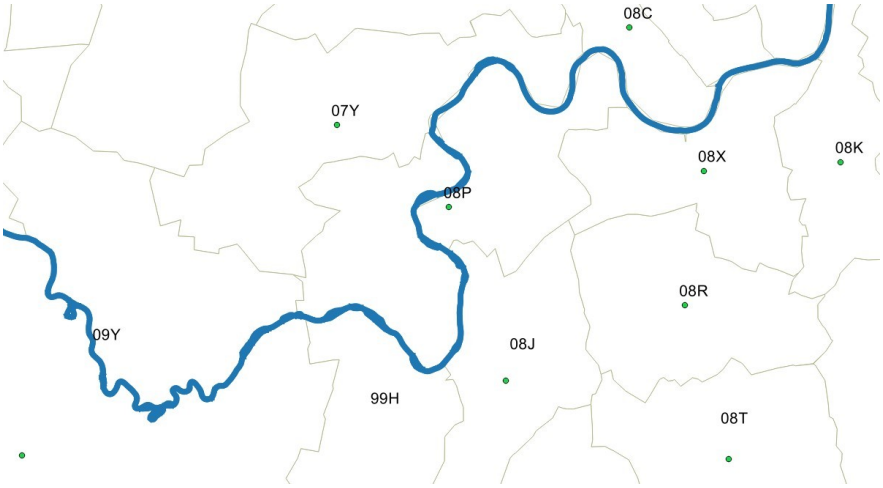


Figure 5.3: *This figure illustrates how we select pairs of CCG regions spanning the Thames from the QGIS file. Showing a subset of the Thames first shows CCG 08C and 08X on the border. We identify 08C and 08X as a corresponding pair for river definition and approximation. The river continues between 07Y and 08P, we identify 07Y and 08P as the next pair. If the river flows directly in the middle of one region, such as 08P, a nearby CCG region for this segment of the river is selected. In this case we add 07Y and 08J as a third pair.*

3. For each CCG region adjacent to the river, we couple corresponding CCGs on opposite sides of the river and save matching pairs of region nodes based on closest centroids. If the river flows directly in the middle of one region, such as 08P, a nearby CCG region for this segment of the river is selected. See Figure 5.3.

4. For each pair of corresponding CCG regions, we connect their centroids with an edge, $e(c_1, c_2)$.

5. We add the mid-point of each edge $e(c_1, c_2)$ as a vertex $v(r)$ on a polyline representing the river, $r(v_0, ..., v_n)$.

6. Connect all the derived river vertices, $v(r)_0$,...,$v(r)_n$ and add the river to the cartographic map. See Figures 5.3 and 5.4 for an illustration of this process.

5.2.4 Compute Region Center Points

To generate a Doring style or rectangular cartogram, we start with small rectangles to represent each CCG region, and increase the size of each rectangle until they gradually reach the max-

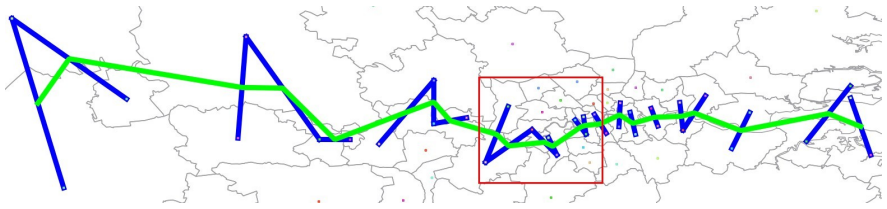


Figure 5.4: *This figure illustrates inserting the river polyline on the cartographic map by connecting all the derived river vertices. Green lines show edges connecting pairs of CCG regions, and the river line is formed by connecting the mid-point of all green lines. The red rectangle highlights the parts of detail view of Figure 1.*

imum size of which the user has control. Initially we input the original CCG map from NHS website[40] into QGIS [41] and use the centroid calculation tool to calculate the center points of each CCG region. We export the position of center points to a CSV file and use it as an input to our cartogram as the starting positions of the rectangular region nodes. Sample center points can be seen in Figures 3 and 4.

5.2.5 Update Node Size and Remove Overlap

We start with a unit square to represent each CCG region as a node in the cartogram and increase the size of each node gradually. During the region growing process, one region may shift adjacent neighboring regions in order to remove overlap and preserve relative position. We use the fast overlap removal algorithm [179, 180] incrementally for this process. When all regions reach their maximum size or one of the regions reaches the boundary of map, the cartogram layout stops. In each step we increase the size of all nodes by 1 pixel (if they have not reached their maximum size) to make them grow smoothly and gradually.

5.2.6 Test For River Intersection

We also incorporate a river intersection test algorithm to confirm that regions do not cross this topological feature. When the size of each region increases, we connect the current centroid with its previous centroid. If the connected edge intersects one of the river edges, we mark this region as an intersecting region. See algorithms 1 and 2. For the intersection test of two edges, we first test whether the bounding box of two segments intersect. If not, then the two segments do not intersect. Then we use a vector cross product to test whether two points of one segment

Algorithm 1 TestRiverIntersection

Input
CCGList : list of CCG regions
river: list of river edges
Local Variables
Edge : a line segment connecting current position and previous position of each CCG region
Output
intersectionList: a list of CCGs that intersect the rivers

```
 1: procedure TESTRIVERCROSSING (CCGList , river)
 2:     intersectionList
 3:     for index = 0; index < CCGList.size(); index++ do
 4:         Line Edge = Line(CCGList.at(index).getPosition(),
 5:             CCGList.at(index).getPreviousPosition())
 6:         if TestIntersection ( Edge, riverEdge) == TRUE then
 7:             intersectionList.append(CCGList.at(index))
 8:         endIf
 9:     endFor (index)
10:     return intersectionList
```

are on the opposite side of the other segment. If they are both true, the two segments intersect [194].

5.2.7 Topology Preservation Algorithm

After testing river intersection for each region, we select the crossing regions and move them back to their previous position (those positions are stored because a unit area's centroid is not allowed to cross a river edge). This backward transition may cause overlap between regions, so we rerun the overlap removal algorithm to remove overlap. In some cases, a region may shift repeatedly between the same two positions by the topology preservation algorithm and overlap removal algorithm, which we identify as a stalemate. When a stalemate occurs, we move the affected region back to its previous position again, and push all other overlapping regions in the reverse push region by same amount. Our first attempts at this algorithm used only a reverse push line segment. This alone was not enough to prevent a stalemate. Thus we introduced a reverse push region. See Figure 5.6. The width of the reverse push direction region, p_ε, may also be adjusted by user. See Algorithm 2 for a complete procedural description.

5.2.8 Test Region Size and Domain Boundaries

We use maximum size as a control for cartogram node size. The cartogram generation algorithm stops when all the region nodes reach a maximum size. Nodes are same size when the

Algorithm 2 LayoutNode

Input–
CCGList : list of CCG regions
river: list of river edges
Local variables
MaxSize : the maximum size of CCG regions set by user
crossing: TRUE if a CCG region crosses river
stalemate: TRUE if a CCG region is placed back in its previous position
reverseDirection: the reverse direction that lead neighboring CCG regions to stagnation

```
1:  procedure LAYOUTNODE ( CCGList , river )
2:      while CCGList.getMaxSize() < MaxSize do
3:          intersectionList = TestRiverIntersection( CCGList, river )
4:          if intersectionList.getSize() > 0 then
5:              intersectionList.saveCrossedPositions()
6:              intersectionList.setPositions (
7:                  intersectionList.getPreviousPositions())
8:              boolean stalemate = intersectionList.isStalemate ()
9:              if stalemate == TRUE then
10:                 intersectionList.deriveReverseDirections ()
11:                 // push neighboring CCGs
12:                 CCGList.counterMovement ()
13:             endIf  // Stalemate
14:         else  // No intersections
15:             CCGList.setSize ( CCGList.getSize()++ )
16:             CCGList.savePreviousPosition()
17:         endIf   // Intersections
18:         RemoveOverlap()
19:         UpdateLayout()
20:     endWhile
```

user chooses uniform size nodes or the maximum size node when user chooses unit area size mapped to area population.

The domain boundaries are set to ensure each node does not move outside the screen. When a region node reaches the north/south boundaries, it will be shifted to west/east direction which enables the cartogram generation algorithm to continue. When a region node reaches west/east boundaries, the algorithm terminates. This decision is made due to the elongated shape of the UK in the north/south orientation.

5.3 Results and Discussion

For a video demonstration of the algorithm and more results, please see the accompanying video https://vimeo.com/276194111. In this section, we present our cartograms with and without topological features. Figure 5.5 (top) shows the typical cartogram generated without

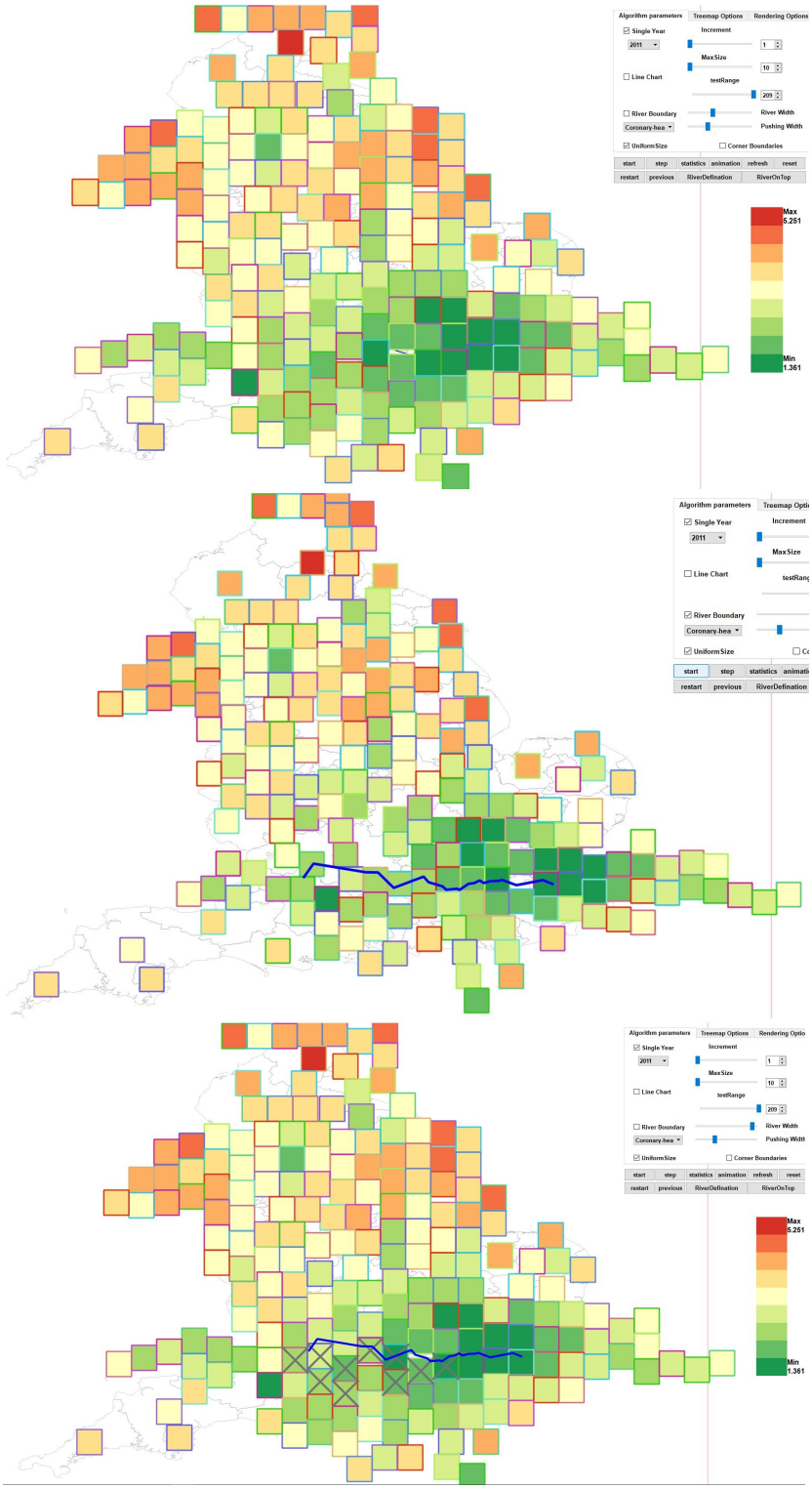


Figure 5.5: *The top shows the basic cartogram without topological features. The middle shows the cartogram with Thames river and a narrow push width, p_ε. The bottom shows regions marked with a gray cross are those that cross the Thames river if the topology is not preserved. Color is mapped to Coronary heart disease distribution in England.*

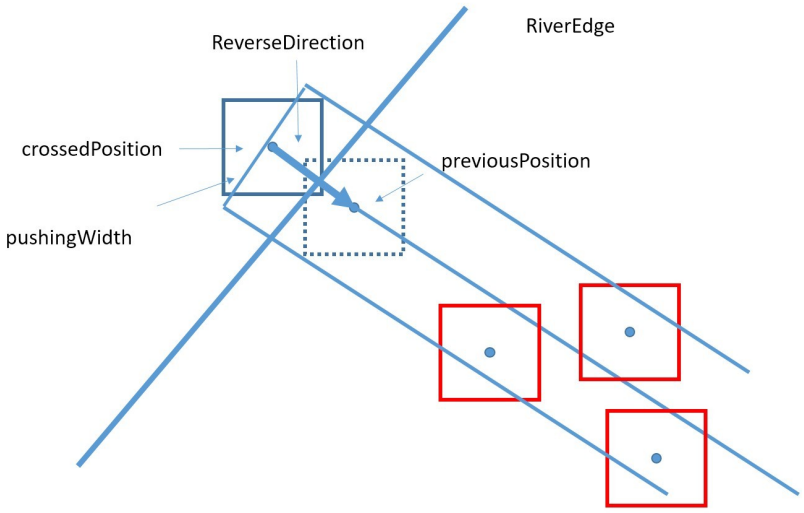


Figure 5.6: *This figure illustrates the configuration when a CCG region (blue) crosses the river and is placed back to its previous position (dashed outline). A reverse direction is derived used to push all neighboring CCG regions (red) in the reverse direction. The reverse direction is turned into a reverse region by introducing a width, p_ε.*

any topological features. Figure 5.5 (middle) presents the output of the cartogram with the Thames river topology. The reverse pushing width is set to p_ε = 40 % of a unit region width. Figure 5.8 shows the cartogram with Thames river and a larger pushing width p_ε =100 % of a unit region width. Figure 5.5 (bottom) shows the regions that cross the river topology if no intersection test is performed. In other words, topologically incorrect region nodes.

Also, instead of using uniform size, we can map the size of node to each regions' population. Figure 5.7 (top) shows the standard cartogram with size mapped to population. Figure 5.7 (middle) shows the cartogram with Thames river feature and mapped to population. Regions in Figure 5.7 (bottom) marked with a gray cross are those that cross the Thames river. If the topology is not preserved. There are 10 error regions crossing the river in the figure, which are Oxfordshire, Chiltern, Windsor Ascot & Maidenhead, Hillingdon, Slough, Ealing, Hammersmith Fulham, Hounslow, Richmond and Wandsworth.

We are using a diverging color map from dark green to dark red (from Colorbrewer [44]) to illustrate a user-chosen health care prevalence from minimum to maximum. In the figures we present, the color is mapped to coronary heart disease distribution in England. Users also can select any healthcare prevalence through a menu and generate different cartograms. We

also enable the user to select the maximum size of a region to control the trade-off between space-filled and map accuracy.

We also offer zooming and panning tools that help users to zoom close-up for details of the cartogram.

To increase the river recognizability, as a user option, we also present a river width setting, r_ε. A region is marked as crossing a river not just when its center point crosses the river polyline, but when the distance between its center point and any river polyline segments is less than a user specified distance. See Figure 5.1. For a supplementary video showing further results, please see the accompanying video.

5.4 Summary

This Chapter presents cartograms with topological features which increase the recognizability of a cartographic map and reduce layout error. We convert real geographic information such as a river into a topological feature on the cartogram using a definition and approximation algorithm. We then implement a Dorling style cartogram incorporating this topological feature constraint and display CCG population and healthcare data. Several interactive user options are available to explore and present the results focusing on different user requirements for further exploration, analysis and comparison. Future work includes investigating more topological features, e.g. including more major rivers and a more in-depth user feedback study. A detailed evaluation including a general user-study is future work and would require an additional thesis chapter. We believe the main novelty lied in the new concept and a sample implementation. We believe evaluation start with an objective error metric in this case.

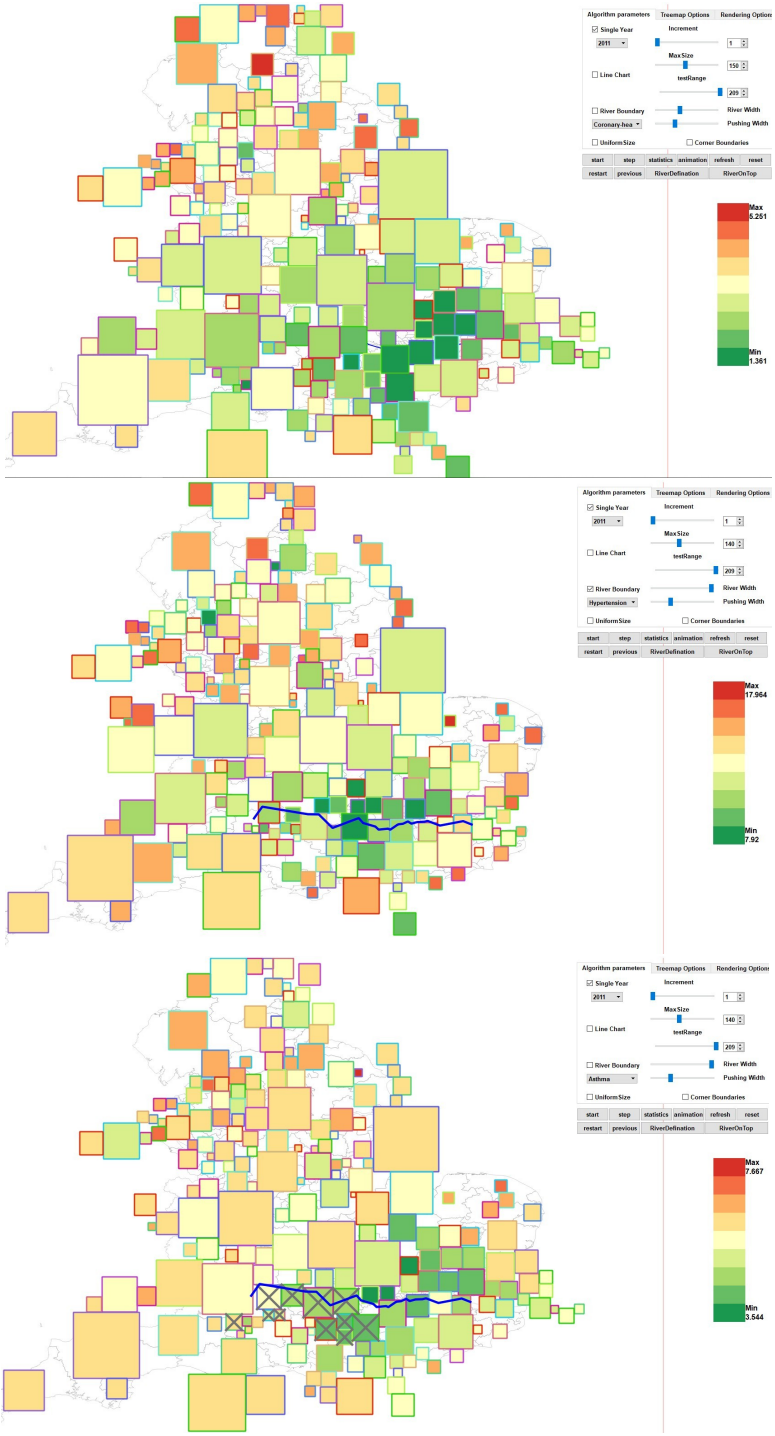


Figure 5.7: *The figure shows the cartogram with unit area size mapped to population without (top) and with (middle) the Thames topology feature. Regions marked with a grey cross are those that cross the Thames river if the topology is not preserved (bottom). The color is mapped to hypertension prevalence in England.*

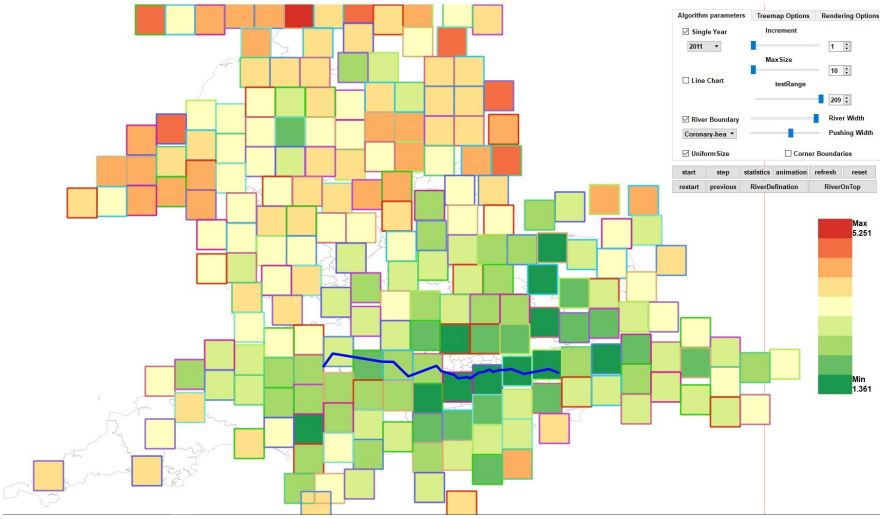


Figure 5.8: *This figure shows the cartogram with Thames river and a wide pushing width, p_ε = 100 %. Color is mapped to Coronary heart disease distribution in England.*

Chapter 6

Conclusion and Future Work

Contents

"Someday, with my sail piercing the clouds; I will mount the wind, break the waves, and traverse the vast, rolling sea."-Li Bai[1]

6.1 Conclusion

In this thesis, we work on geo-spatial visualization with population healthcare data. I describe a literature review of narrative visualization including geo-spatial visualization. It summaries the goals we want to reach in our research: increasing memorability and cognition. By using cartograms and treemaps, we are able to combine high-dimensional, mutlivariate data with corresponding geo-spatial information. The work we present differs from previous work in that it attempts to combine the space-filling, hierarchical characteristics of ordered space-filling treemaps together with the geo-spatial information conveyed by a cartogram (Chapter

[1] Li Bai (701-762), also known as Li Bo, Li Po and Li Taibai, was a Chinese poet acclaimed from his own day to the present as a genius and a romantic figure who took traditional poetic forms to new heights.

3). Additionally we include time as a variate into the cartographic treemaps (Chapter 4). Finally, we introduce topological features to Dorling-style and rectangular cartograms (Chapter 5). We strongly believe that our work provides a novel solution for visualizing this kind of complex data set.

To conclude this thesis, we again re-emphasise the main benefits and contributions of this thesis:

In Chapter 2, we present a literature survey of narrative visualization including geo-spatial visualization. We provide a novel up-to-date overview of storytelling in visualization, in which the most important recent literature is included and discussed. Since storytelling in visualization is a relatively new subject, we expect an increase in research in the coming years. Moreover we believe it will evolve into a popular topic in the field of visualization.

In Chapter 3, we present a novel hybrid visualization, the cartographic treemaps combining geo-spatial information, a novel interactive neighborhood preservation metric, and space-efficient geometry for the interactive visualization of geo-spatial, and high-dimensional data. It combines the advantages of both cartograms and treemaps. We go on to implement and demonstrate this visualization with a real-world high-dimensional health care data collected by NHS to support clinical commissioning groups (CCGs) and the health care service providers. Several interactive user options are available to explore and present the results focusing on different user requirements for further exploration, analysis and comparison. Also, we present several multivariate observations based on the cartographic treemap visualization and report feedback from two domain experts in health science.

In Chapter 4, we extend the cartographic treemaps presented in Chapter 3 by adding time-oriented data. In particular, we introduce a new time-oriented cartographic treemaps that enables the user to explore hierarchical, multi-variate data over a range of years. Both static and animated visual designs are used for cartographic treemaps to present the temporal trends of data. We also provide interactive user-options that enable users to customize the visual layout.

In Chapter 5, we present cartograms with topological features which increase the recognizability of a cartographic map and reduce layout error. We have done this by converting real geographic information such as a river into a topological feature on the cartogram using a definition and approximation algorithm. We then implement a Dorling style cartogram incorporating this topological feature constraint and displaying CCG population and healthcare data. Several interactive user options are available to explore and present the results focusing

on different user requirements for further exploration, analysis and comparison.

All my research implementation is based on C++ and the QT platform. We are using Git for version control. The geo-spatial information is converted by QGIS [41] from map files (shapefile or geojson) to CSV files. The CSV files contain all the useful information we need for further implementation, such as longitude and latitude of each region, region name, region code, and area code. We extract 14 main healthcare attributes from excel files provided on the NHS website [46]. There are many important classes for our software. The DataFile class is to read and convert all the modified input. The Region file is to store all the geo-spatial related information, such as region position, size and corresponding healthcare data. The ColorMap class is to implement different color legends from various sources. We use the Treemaps class to present treemaps inside a region node or to provide a detailed view. And we use the LayoutWidget class and Interface class to present the final output and enable user interaction. We also implement a "step" function to show each single frame of our visual design. This aids debugging as presented by Laramee in "using visualization to debug visualization"[195].

6.2 Future Work

My whole PhD session is working on geo-space with population healthcare data. We present an overview of narrative visualization, and develop three different cartograms systems focusing on multivariate data, time-oriented data and geo-spatial features. Looking ahead, there are several unsolved problems which need more investment in future work.

For survey of narrative visualization including geo-space

By reviewing Table 2.1 and Table 2.2, we can see that storytelling visualization focuses on information visualization more than scientific visualization, which conveys that more challenges are left unsolved in this field. However, by refining a storytelling model for scientific visualization [10], the implementation of storytelling in scientific visualization could increase in the future. We can also see that storytelling in visualization concentrates more on exploration than on presentation. Like Kosara and Mackinlay [115] state: "visualization techniques address the exploration and analysis of data more than presenting data".

In future work, there are many directions and unsolved problems. Storytelling will gain importance in data presentation and data exploration. Here is a summary of some unsolved problems in storytelling for visualization:

- It is clear that objective measures of user-engagement is a relatively unexplored area of research. Can we derive a mature classification of user engagement activities? Is user engagement something we can clearly define?

- Data preparation and enhancement: Virtually no one has addressed the challenge of data preparation and enhancement for storytelling. Moreover, is storytelling data best captured or derived from an existing data set or software system? Can a standard data file format be developed?

- Narrative visualization for scientific and geo-spatial visualization: Why has there been such an imbalance of research narrative visualization for information visualization but virtually none for scientific and geo-spatial visualization?

- Transitions for scientific visualization: The benefits of static transition versus dynamic transitions in visualization still remains relatively immature.

- Memorability for visualization: What are the key elements for making a memorable visualization? This is still an immature research direction.

- Animated transitions for geo-spatial visualization: Animated transitions for geo-spatial visualization remains an open research direction. This is surprising given the popularity and importance of geo-spatial visualization.

- Interpretation for scientific information, and geo-spatial visualization: Currently no papers to our knowledge focus on the topic of effective interpretation of stories; this topic remains largely unexplored.

The classification of literature we present makes it clear that many future research directions remain open in storytelling and visualization.

Cartographic treemaps

Future work includes investigating more optional color maps for high-dimensional data and a more in-depth user feedback study. Future work will include more attributes of NHS data in addition to population and health disorder prevalence, such as the number of practices per CCG, and rates of A&E admissions. More filtering options will also be introduced, such as filtering by age range. Also, a deeper user study on the performance of cartographic treemaps with more domain expert could be done in future work.

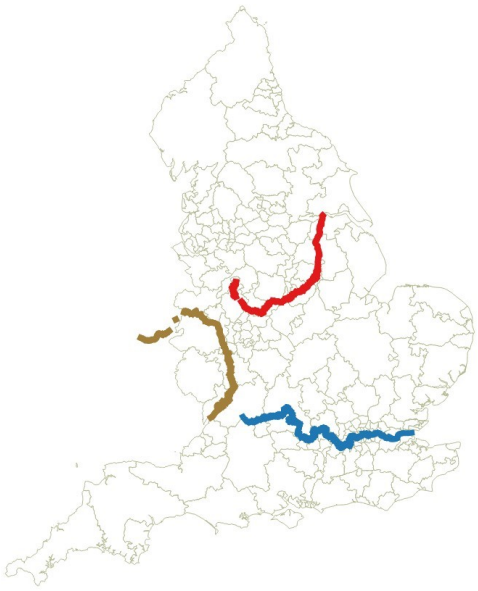


Figure 6.1: *This image shows a map of England with three main rivers, The River Severn(yellow), The River Thames(blue) and The River Trent(red).*

Time-oriented cartographic treemaps

Adding a longer period of time could be considered as future work. Assessing the utility of animation is also future work. A data analyst may wish to examine how prevalence rate varies by age group; in other words, to assess the degree of association between two attributes. A second categorical attribute such as age group could be accommodated readily within this tool by using clustered or stacked bar charts, pyramidal bar charts or heat maps.

Other possible extensions would involve the graphical representation of other types of attribute (e.g. histogram for a continuous measure or score variable) and combinations of different types of attribute (e.g. box and whisker plot to compare the distribution of a continuous measure or score variable between two or more age groups). Other issues that may arise in the analysis of longitudinal data include state dependence and the mover-stayer problem [196]. These issues could be explored by displaying a heat map within each tile in order to represent the matrix of transition probabilities at each location on the cartogram.

Cartograms with features

Future work includes investigating more topological features, e.g. including more major rivers, such as the River Severn and the River Trent. (See Figure 6.1). A more in-depth user feedback study is also a direction of future work. Do the topological features enhance the

recognizability of a cartogram? And how many topological features would be a good number to be implemented?

For the future work of this thesis, we also propose more research work on combining other visualization techniques with geo-spatial information and high dimensional dataset. A corresponding case study on presenting visualization output by storytelling techniques for user memorability and engagement is another direction for future work.

Bibliography

[1] R. S. Laramee, "Bob's Concise Coding Conventions (C3)," *Advances in Computer Science and Engineering (ACSE)*, vol. 4, no. 1, pp. 23–26, 2010.

[2] ——, "From Data Chaos to the Visualization Cosmos," in *KEYNOTE Talk at the 2nd International Conference on Big Data, Cloud and Applications*. Tetuan, Morocco, 29-30 March 2017.

[3] R. Roberts, C. Tong, R. Laramee, G. A. Smith, P. Brookes, and T. DCRUZE, "Interactive Analytical Treemaps for Visualisation of Call Centre Data," in *Proceedings of the Conference on Smart Tools and Applications in Computer Graphics*. Eurographics Association, 2016, pp. 109–117.

[4] E. Grundy, M. W. Jones, R. S. Laramee, R. P. Wilson, and E. L. Shepard, "Visualisation of sensor data from animal movement," in *Computer Graphics Forum*, vol. 28, no. 3. Wiley Online Library, 2009, pp. 815–822.

[5] N. Alharbi, R. S. Laramee, and M. Chavent, "Molpathfinder: interactive multi-dimensional path filtering of molecular dynamics simulation data," in *The Computer Graphics and Visual Computing (CGVC) Conference*, vol. 2016, 2016, pp. 9–16.

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

[7] A. Lu and H.-W. Shen, "Interactive storyboard for overall time-varying data visualization," in *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific*. IEEE, 2008, pp. 143–150.

[8] P. Cruz and P. Machado, "Generative Storytelling for Information Visualization," *IEEE computer graphics and applications*, no. 2, pp. 80–85, 2011.

[9] M. Wohlfahrt, "Story Telling Aspects in Medical Applications," in *Central European Seminar on Computer Graphics*, 2006.

[10] M. Wohlfahrt and H. Hauser, "Story Telling for Presentation in Volume Visualization," in *Proceedings of the 9th Joint Eurographics/IEEE VGTC conference on Visualization*. Eurographics Association, 2007, pp. 91–98.

[11] E. M. Lidal, H. Hauser, and I. Viola, "Geological Storytelling: Graphically Exploring and Communicating Geological Sketches," in *Proceedings of the International Symposium on Sketch-Based Interfaces and Modeling*. Eurographics Association, 2012, pp. 11–20.

[12] E. M. Lidal, M. Natali, D. Patel, H. Hauser, and I. Viola, "Geological Storytelling," *Computers & Graphics*, vol. 37, no. 5, pp. 445–459, 2013.

[13] B. Lee, R. H. Kazi, and G. Smith, "SketchStory: Telling More Engaging Stories with Data through Freeform Sketching," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2416–2425, 2013.

[14] P. Lundblad and M. Jern, "Geovisual Analytics and Storytelling Using HTML5," in *Information Visualisation (IV), 2013 17th International Conference*. IEEE, 2013, pp. 263–271.

[15] R. Eccles, T. Kapler, R. Harper, and W. Wright, "Stories in GeoTime," *Information Visualization*, vol. 7, no. 1, pp. 3–17, 2008.

[16] A. Kuhn and M. Stocker, "CodeTimeline: Storytelling with versioning data," in *Software Engineering (ICSE), 2012 34th International Conference on*. IEEE, 2012, pp. 1333–1336.

[17] N. Mahyar, S.-H. Kim, and B. C. Kwon, "Towards a Taxonomy for Evaluating User Engagement in Information Visualization," in *Workshop on Personal Visualization: Exploring Everyday Life*, vol. 3, 2015.

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

[19] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic Generation of Annotated Stock Visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2707–2716.

[20] J. Hullman, S. Drucker, N. H. Riche, B. Lee, D. Fisher, and E. Adar, "A Deeper Understanding of Sequence in Narrative Visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2406–2415, 2013.

[21] B. Bach, N. Kerracher, K. W. Hall, S. Carpendale, J. Kennedy, and N. H. Riche, "Telling Stories about Dynamic Networks with Graph Comics," in *Proceedings of the Conference on Human Factors in Information Systems (CHI)*. ACM, New York, United States, 2016.

[22] F. B. Viégas, D. Boyd, D. H. Nguyen, J. Potter, and J. Donath, "Digital Artifacts for Remembering and Storytelling: Posthistory and social network fragments," in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE, 2004, pp. 10–pp.

[23] J. Hullman and N. Diakopoulos, "Visualization Rhetoric: Framing Effects in Narrative Visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2231–2240, 2011.

[24] A. Figueiras, "Narrative Visualization: A Case Study of How to Incorporate Narrative Elements in Existing Visualizations," in *Information Visualisation (IV), 2014 18th International Conference on*. IEEE, 2014, pp. 46–52.

[25] ——, "How to Tell Stories Using Visualization," in *Information Visualisation (IV), 2014 18th International Conference on*. IEEE, 2014, pp. 18–18.

[26] P. H. Nguyen, K. Xu, R. Walker, and B. W. Wong, "Schemaline: Timeline visualization for sensemaking," in *2014 18th International Conference on Information Visualisation*. IEEE, 2014, pp. 225–233.

[27] M. Akaishi, K. Satoh, Y. Kato, and K. Hori, "Narrative Based Topic Visualization for Chronological Data," in *Information Visualization, 2007. IV'07. 11th International Conference*. IEEE, 2007, pp. 139–144.

[28] D. Fisher, A. Hoff, G. Robertson, and M. Hurst, "Narratives: A Visualization to Track Narrative Events as They Develop," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. IEEE, 2008, pp. 115–122.

[29] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration of Big Spatio-temporal Urban Data: A study of new york city taxi trips," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2149–2158, 2013.

[30] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko, "Effectiveness of Animation in Trend Visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1325–1332, 2008.

[31] T. Chen, A. Lu, and S.-M. Hu, "Visual storylines: Semantic visualization of movie sequence," *Computers & Graphics*, vol. 36, no. 4, pp. 241–249, 2012.

[32] Y. Tanahashi and K.-L. Ma, "Design Considerations for Optimizing Storyline Visualizations," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2679–2688, 2012.

[33] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu, "StoryFlow: Tracking the Evolution of Stories," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2436–2445, 2013.

[34] J. Heer and G. G. Robertson, "Animated Transitions in Statistical Data Graphics," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1240–1247, 2007.

[35] B. B. Bederson and A. Boltman, "Does Animation Help Users Build Mental Maps of Spatial Information?" in *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*. IEEE, 1999, pp. 28–35.

[36] H. Akiba, C. Wang, and K.-L. Ma, "Aniviz: A Template-based Animation Tool for Volume Visualization," *Computer Graphics and Applications, IEEE*, vol. 30, no. 5, pp. 61–71, 2010.

[37] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks, "Useful Junk?: The Effects of Visual Embellishment on Comprehension and Memorability of Charts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 2573–2582.

[38] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond Memorability: Visualization Recognition and Recall," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 519–528, 2016.

[39] B. Saket, C. Scheidegger, S. G. Kobourov, and K. Börner, "Map-based Visualizations Increase Recall Accuracy of Data," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 441–450.

[40] NHS, https://www.england.nhs.uk/resources/ccg-maps/, Accessed: June 2018.

[41] QGIS, http://www.qgis.org/en/site/, Accessed: June 2018.

[42] "Disk Inventory X," http://www.derlien.com/, Accessed: June 2018.

[43] V. Setlur and M. C. Stone, "A Linguistic Approach to Categorical Color Assignment for Data Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 698–707, 2016.

[44] "ColorBrewer," http://colorbrewer2.org/, Accessed: June 2018.

[45] A. C. Telea, *Data Visualization: Principles and Practice*. CRC Press, 2014.

[46] NHS, http://fingertips.phe.org.uk/profile/general-practice, Accessed: June 2018.

[47] C. Tong, R. Roberts, R. S. Laramee, D. Berridge, and D. Thayer, "Cartographic Treemaps for the Visualization of Public Health Care Data," *Proceedings of the Conference on Computer Graphics and Visual Computing (CGVC), 2017*, 2017.

[48] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss, "Colorgorical: Creating Discriminable and Preferable Color Palettes for Information Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 521–530, 2017.

[49] D. Visualization, https://searchbusinessanalytics.techtarget.com/definition/data-visualization, Accessed: June 2018.

[50] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[51] C. Centres, https://www.unison.org.uk/at-work/energy/key-issues/call-centres/, Accessed: June 2018.

[52] C. Tong, R. Richard, B. Rita, L. Robert S, W. Kodzo, L. Aidong, Y. Wang, H. Qu, Q. Luo, and X. Ma, "Storytelling and Visualization: A Survey," in *Proceedings of the 9th International Conference on Information Visualization Theory and Applications (IVAPP) 2018*. Funchal, Madira, Portugal, pages 212-224, 27-29 January 2018.

[53] C. Tong, R. Roberts, R. Borgo, S. Walton, R. S. Laramee, K. Wegba, A. Lu, Y. Wang, H. Qu, Q. Luo *et al.*, "Storytelling and visualization: An extended survey," *Information*, vol. 9, no. 3, p. 65, 2018.

[54] S. Goldberg, "The Risks Of Storytelling," in *National Geographic Magazine*, 2015.

[55] A. Singer, "Data Visualization: Your Secret Weapon in Storytelling and Persuasion," https://www.clickz.com/clickz/column/2378704/data-visualization/your-secret-weapon-in-storytelling-and-persuasion, 2014, Accessed: June 2018.

[56] D. Reference, http://dictionary.reference.com/browse/story, Accessed: June 2018.

[57] O. E. Dictionary, http://www.oed.com/view/Entry/190981?rskey=Wrp9f3&result=1, Accessed: June 2018.

[58] J. Zipes, *Creative storytelling: Building community/changing lives*. Routledge, 2013.

[59] J. Schell, *The Art of Game Design: A Book of Lenses*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[60] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale, "More Than Telling A Story: Transforming Data Into Visually Shared Stories," *IEEE computer graphics and applications*, vol. 35, no. 5, pp. 84–90, 2015.

[61] O. E. Dictionary, http://www.oxforddictionaries.com/definition/english/authorship, Accessed: June 2018.

[62] D. Reference, http://www.thefreedictionary.com/authorship, Accessed: June 2018.

[63] J. Rodgers, "Defining and Experiencing Authorship (s) in the Composition Classroom: Findings from a Qualitative Study of Undergraduate Writing Students at the City University of New York." *Journal of Basic Writing (CUNY)*, vol. 30, no. 1, pp. 130–155, 2011.

[64] N. Gershon and W. Page, "What Storytelling Can Do for Information Visualization," *Communications of the ACM*, vol. 44, no. 8, pp. 31–37, 2001.

[65] J. Fulda, M. Brehmel, and T. Munzner, "TimeLineCurator: Interactive Authoring of Visual Timelines from Unstructured Text," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 300–309, 2016.

[66] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani, "Authoring Data-Driven Videos with DataClips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 501–510, 2017.

[67] J. Boy, F. Detienne, and J.-D. Fekete, "Can Initial Narrative VisualizationTechniques and Storytelling help Engage Online-Users with Exploratory Information Visualizations?" 2015.

[68] T. Gao, J. R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos, "NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News," in *Proceedings of the 32nd annual ACM conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3005–3014.

[69] F. Amini, N. Henry Riche, B. Lee, C. Hurter, and P. Irani, "Understanding Data Videos: Looking at Narrative Visualization through the Cinematography Lens," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 1459–1468.

[70] A. Satyanarayan and J. Heer, "Authoring Narrative Visualizations with Ellipsis," in *Computer Graphics Forum*, vol. 33, no. 3. Wiley Online Library, 2014, pp. 361–370.

[71] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit, "From Visual Exploration to Storytelling and Back Again," in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 491–500.

[72] C. Bryan, K.-L. Ma, and J. Woodring, "Temporal Summary Images: An Approach to Narrative Visualization via Interactive Annotation Generation and Placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 511–520, 2017.

[73] I. Liao, W.-H. Hsu, and K.-L. Ma, "Storytelling via Navigation: A Novel Approach to Animation for Scientific Visualization," in *International Symposium on Smart Graphics*. Springer, 2014, pp. 1–14.

[74] T. Nagel, C. Pietsch, and M. Dörk, "Staged Analysis: From Evocative to Comparative Visualizations of Urban Mobility," in *Proceedings of the IEEE VIS Arts Program (VISAP'16)*, Baltimore, Maryland, October 2016, pp. 23–30.

[75] M. Borkin, A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, H. Pfister *et al.*, "What Makes a Visualization Memorable?" *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2306–2315, 2013.

[76] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko, "Visualization Publication Dataset," Dataset: http://vispubdata.org/, 2015, published Jun. 2015. [Online]. Available: http://vispubdata.org/

[77] A. D. Santiago, P. N. Sampaio, and L. R. Fernandes, "MOGRE-Storytelling: Interactive Creation of 3D Stories," in *Virtual and Augmented Reality (SVR), 2014 XVI Symposium on*. IEEE, 2014, pp. 190–199.

[78] A. Cropper, R. E. Luna, and E. L. Mclean, "Scientific Storytelling: From up in the clouds to down to earth A new approach to mentoring," in *Integrated STEM Education Conference (ISEC), 2015 IEEE*. IEEE, 2015, pp. 252–257.

[79] P. Alavesa and D. Zanni, "Combining storytelling tradition and pervasive gaming," in *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2013 5th International Conference on*. IEEE, 2013, pp. 1–4.

[80] W.-T. Chu, C.-H. Yu, and H.-H. Wang, "Optimized Comics-Based Storytelling for Temporal Image Sequences," *Multimedia, IEEE Transactions on*, vol. 17, no. 2, pp. 201–215, 2015.

[81] C. D. Correa and K.-L. Ma, "Dynamic Video Narratives," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, p. 88, 2010.

[82] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating Expressive Speech for Storytelling Applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1137–1144, 2006.

[83] N. science missions, http://science.nasa.gov/missions, oct 2014.

[84] N. scientific visualization studio, http://svs.gsfc.nasa.gov, oct 2014.

[85] L. Plowman, R. Luckin, D. Laurillard, M. Stratfold, and J. Taylor, "Designing Multimedia for Learning: Narrative Guidance and Narrative Construction," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 1999, pp. 310–317.

[86] Laramee and R. S, "How to read a visualization research paper: Extracting the essentials," *Computer Graphics and Applications, IEEE*, vol. 31, no. 3, pp. 78–82, 2011.

[87] S. Denning, *The Springboard: How Storytelling Ignites Action in Knowledge-era Organizations*. Routledge, 2001.

[88] C. D. Hansen and C. R. Johnson, *Visualization handbook*. Academic Press, 2011.

[89] M. Naratology, "Introduction to The Theory of Narrative," 1985.

[90] L. Mroz and H. Hauser, "RTVR: a Flexible Java Library for Interactive Volume Rendering," in *Proceedings of the conference on Visualization'01*. IEEE Computer Society, 2001, pp. 279–286.

[91] A. Gooch, B. Gooch, P. Shirley, and E. Cohen, "A Non-photorealistic Lighting Model for Automatic Technical Illustration," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 1998, pp. 447–452.

[92] R. B. Haber and D. A. McNabb, "Visualization Idioms: A Conceptual Model for Scientific Visualization Systems," *Visualization in scientific computing*, vol. 74, p. 93, 1990.

[93] I. Viola, "Importance-Driven Expressive Visualization," Ph.D. dissertation, Viola, 2005.

[94] M. Donald, "*Prcis* of Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition," *Behavioral and Brain Sciences*, vol. 16, no. 04, pp. 737–748, 1993.

[95] G. Li, X. Cao, S. Paolantonio, and F. Tian, "SketchComm: A Tool to Support Rich and Flexible Asynchronous Communication of Early Design Ideas," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 359–368.

[96] J. Heer, F. B. Viégas, and M. Wattenberg, "Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 1029–1038.

[97] M. Ogawa and K.-L. Ma, "Software Evolution Storylines," in *Proceedings of the 5th international symposium on Software visualization*. ACM, 2010, pp. 35–42.

[98] ——, "Code_swarm: A design study in organic software visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1097–1104, 2009.

[99] A. Begel, Y. P. Khoo, and T. Zimmermann, "Codebook: Discovering and Exploiting Relationships in Software Repositories," in *Software Engineering, 2010 ACM/IEEE 32nd International Conference on*, vol. 1. IEEE, 2010, pp. 125–134.

[100] D. Gotz and Z. Wen, "Behavior-driven Visualization Recommendation," in *Proceedings of the 14th international conference on Intelligent user interfaces*. ACM, 2009, pp. 315–324.

[101] O. E. Dictionary, https://en.oxforddictionaries.com/definition/narrative, Accessed: June 2018.

[102] J. Heer, F. B. Viégas, and M. Wattenberg, "Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 1029–1038.

[103] B. Bond, "Steroids or Not, the Pursuit is On," http://www.nytimes.com/2006/04/02/sports/20060402_BONDS_GRAPHIC.html, 2006.

[104] A. Cox., "Budget Forecasts VS. Reality." http://www.nytimes.com/interactive/2010/02/02/us/politics/20100201-budget-porcupine-graphic.html, 2010.

[105] S. T. M. Green, H. Warrell, S. Bernard, and M. Formentini, "Afghanistan: Behind the Front Line," http://www.gapminder.org/downloads/human-development-trends-2005/, Jan 2010.

[106] Gapminder, "Human Development Trends, 2005," http://www.gapminder.org/downloads/human-development-trends-2005/, 2005.

[107] A. H. J.Heer and M. Agrawala, "Minnesota Employment Explorer," http://minnesota.publicradio.org/projects/2008/07/16_minnesota_Slowdown, 2007.

[108] E. Kandogan, "Just-in-time Annotation of clusters, Outliers, and Trends in Point-based Data Visualizations," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012, pp. 73–82.

[109] J. B. Black and G. H. Bower, "Episodes as Chunks in Narrative Memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 18, no. 3, pp. 309–318, 1979.

[110] I. Herman, G. Melançon, and M. S. Marshall, "Graph Visualization and Navigation in Information Visualization: A survey," *IEEE Transactions on visualization and computer graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[111] D. Boyd, H.-Y. Lee, D. Ramage, and J. Donath, "Developing Legible Visualizations for Online Social Spaces," in *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*. IEEE, 2002, pp. 1060–1069.

[112] J. S. Donath, "Visual Who: Animating the affinities and activities of an electronic community," in *Proceedings of the third ACM international conference on Multimedia*. ACM, 1995, pp. 99–107.

[113] S. M.Bloch and A.McLean, "Mapping America: Every City, Every Block," http://projects.nytimes.com/census/2010/explorer, 2010.

[114] D.McCandless, "Poll Dancing: How Accurate are Poll Predictions," http://www.guardian.co.uk/news/datablog/2010/may/06/general-election-2010-opinion-polls_-information-beautiful#, Accessed: June 2018.

[115] R. Kosara and J. Mackinlay, "Storytelling: the Next Step for Visualization," *Computer*, no. 5, pp. 44–50, 2013.

[116] P. Pirolli and S. Card, "The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5, 2005, pp. 2–4.

[117] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing Tags over Time," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 2, p. 7, 2007.

[118] R. Swan and D. Jensen, "TimeMines: Constructing Timelines with Statistical Models of Word Usage," in *KDD-2000 Workshop on Text Mining*, 2000, pp. 73–80.

[119] J. J. Van Wijk and E. R. Van Selow, "Cluster and Calendar Based Visualization of Time Series Data," in *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*. IEEE, 1999, pp. 4–9.

[120] O. E. Dictionary, http://www.oxforddictionaries.com/definition/english/transition, Accessed: June 2018.

[121] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban Mobility Study Using Taxi Traces," in *Proceedings of the 2011 international workshop on Trajectory data mining and analysis*. ACM, 2011, pp. 23–30.

[122] M. Veloso, S. Phithakkitnukoon, C. Bento, N. Fonseca, and P. Olivier, "Exploratory Study of Urban Flow Using Taxi Traces," in *First Workshop on Pervasive Urban Applications (PURBA) in conjunction with Pervasive Computing, San Francisco, California, USA*, 2011.

[123] Z. Liao, Y. Yu, and B. Chen, "Anomaly Detection in GPS Data Based on Visual Analytics," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE, 2010, pp. 51–58.

[124] H. Rosling, "TED 2006," http://www.gapminder.org/video/talks/ted-2006--debunking-myth-about-the-third-world.html, 2006.

[125] ——, "TED 2007," http://www.gapminder.org/video/talks/ted-2007--the-seemingly-impossible-is-possible.html, 2007.

[126] E. Tufte, *Envisioning Information*. Graphics Press USA, 1990.

[127] B. Tversky, J. B. Morrison, and M. Betrancourt, "Animation: Can It Facilitate?" *International journal of human-computer studies*, vol. 57, no. 4, pp. 247–262, 2002.

[128] P. Baudisch, D. Tan, M. Collomb, D. Robbins, K. Hinckley, M. Agrawala, S. Zhao, and G. Ramos, "Phosphor: Explaining Transitions in the User Interface Using Afterglow Effects," in *Proceedings of the 19th annual ACM symposium on User interface software and technology*. ACM, 2006, pp. 169–178.

[129] I. Yahiaoui, B. Merialdo, and B. Huet, "Automatic video summarization," in *Proc. CBMIR Conf*, 2001.

[130] V. Ogievetsky., *Plotweaver xkcd/657 creation tool*, 2009, https://graphics.stanford.edu/wikis/cs448b-09-fall/FP-OgievetskyVadim.

[131] C. Gonzalez, "Animation in User Interface Design for Decision Making: a research framework and empirical analysis," 1995.

[132] ——, "Does Animation in User Interfaces Improve Decision Making?" in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1996, pp. 27–34.

[133] M. Donskoy and V. Kaptelinin, "Window Navigation with and without Animation: A Comparison of Scroll Bars, Zoom, and Fisheye View," in *CHI'97 extended abstracts on Human factors in computing systems*. ACM, 1997, pp. 279–280.

[134] H. Childs, E. Brugger, K. Bonnell, J. Meredith, M. Miller, B. Whitlock, and N. Max, "A Contract Based System for Large Data Visualization," in *Visualization, 2005. VIS 05. IEEE*. IEEE, 2005, pp. 191–198.

[135] C. Correa and D. Silver, "Dataset Traversal with Motion-controlled Transfer Functions," in *Visualization, 2005. VIS 05. IEEE*. IEEE, 2005, pp. 359–366.

[136] O. E. Dictionary, http://www.oxforddictionaries.com/definition/english/memory, Accessed: June 2018.

[137] N. Holmes, *Designer's Guide to Creating Charts and Diagrams*. Watson-Guptill, 1984.

[138] A. J. Blasio and A. M. Bisantz, "A Comparison of the Effects of Data ink Ratio on Performance with Dynamic Displays in A Monitoring Task," *International journal of industrial ergonomics*, vol. 30, no. 2, pp. 89–101, 2002.

[139] L. B. Gambrell and P. B. Jawitz, "Mental Imagery, Text Illustrations, and Children's Story Comprehension and Recall," *Reading Research Quarterly*, pp. 265–276, 1993.

[140] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-art of Visualization for Eye Tracking Data," in *Proceedings of EuroVis*, vol. 2014, 2014.

[141] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the Intrinsic Memorability of Images," in *Advances in Neural Information Processing Systems*, 2011, pp. 2429–2437.

[142] R. Kosara and S. Miksch, "Visualization Methods for Data Analysis and Planning in Medical Applications," *International Journal of Medical Informatics*, vol. 68, no. 1, pp. 141–153, 2002.

[143] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman *et al.*, "Interactive Information Visualization to Explore and Query Electronic Health Records," *Foundations and Trends in Human–Computer Interaction*, vol. 5, no. 3, pp. 207–298, 2013.

[144] V. L. West, D. Borland, and W. E. Hammond, "Innovative Information Visualization of Electronic Health Record Data: A Systematic Review," *Journal of the American Medical Informatics Association*, vol. 22, no. 2, pp. 330–339, 2015.

[145] L. McNabb and R. S. Laramee, "Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization," *Computer Graphics Forum*, vol. 36, no. 3, pp. 589–617, 2017.

[146] P. Cruz, A. Cruz, and P. Machado, "Contiguous Animated Edge-Based Cartograms for Traffic Visualization," *IEEE Computer Graphics and Applications*, vol. 35, no. 5, pp. 76–83, 2015.

[147] W. Tobler, "Thirty Five Years of Computer Cartograms," *ANNALS of the Association of American Geographers*, vol. 94, no. 1, pp. 58–73, 2004.

[148] S. Nusrat and S. Kobourov, "The State of the Art in Cartograms," in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 619–642.

[149] D. Auber, C. Huet, A. Lambert, A. Sallaberry, A. Saulnier, and B. Renoust, "Geographical Treemaps," *Technical Report*, 2011.

[150] M. T. Gastner and M. E. Newman, "Diffusion-based Method for Producing Density-equalizing Maps," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 20, pp. 7499–7504, 2004.

[151] D. A. Keim, S. C. North, and C. Panse, "Cartodraw: A Fast Algorithm for Generating Contiguous Cartograms," *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, no. 1, pp. 95–110, 2004.

[152] E. Raisz, "The Rectangular Statistical Cartogram," *Geographical Review*, pp. 292–296, 1934.

[153] D. Dorling, "Area Cartograms: Their Use and Creation," *The Map Reader: Theories of Mapping Practice and Cartographic Representation*, pp. 252–260, 2011.

[154] M. van Kreveld and B. Speckmann, "On Rectangular Cartograms," *Computational Geometry*, vol. 37, no. 3, pp. 175–187, 2007.

[155] R. Heilmann, D. A. Keim, C. Panse, and M. Sips, "Recmap: Rectangular Map Approximations," in *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004.* IEEE, 2004, pp. 33–40.

[156] C. Panse, M. Sips, D. Keim, and S. North, "Visualization of Geo-spatial Point Sets via Global Shape Transformation and Local Pixel Placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 749–756, 2006.

[157] A. Slingsby, J. Dykes, and J. Wood, "Configuring Hierarchical Layouts to Address Research Questions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 977–984, 2009.

[158] ——, "Rectangular Hierarchical Cartograms for Socio-economic Data," *Journal of Maps*, vol. 6, no. 1, pp. 330–345, 2010.

[159] M. Alam, S. G. Kobourov, S. Veeramoni *et al.*, "Quantitative Measures for Cartogram Generation Techniques," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 351–360.

[160] D. Eppstein, M. van Kreveld, B. Speckmann, and F. Staals, "Improved Grid Map Layout by Point Set Matching," *International Journal of Computational Geometry & Applications*, vol. 25, no. 02, pp. 101–122, 2015.

[161] W. Meulemans, J. Dykes, A. Slingsby, C. Turkay, and J. Wood, "Small Multiples with Gaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 381–390, 2017.

[162] B. Shneiderman, "Tree Visualization with Tree-maps: 2-D Space-Filling Approach," *ACM Transactions on graphics (TOG)*, vol. 11, no. 1, pp. 92–99, 1992.

[163] B. Johnson and B. Shneiderman, "Tree-maps: A Space-filling Approach to the Visualization of Hierarchical Information Structures," in *Proceedings of IEEE Visualization 1991*. IEEE Computer Society Press, 1991, pp. 284–291.

[164] B. Johnson, "TreeViz: Treemap Visualization of Hierarchically Structured Information," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1992, pp. 369–370.

[165] B. Shneiderman and M. Wattenberg, "Ordered Treemap Layouts," in *Proceedings of the IEEE Symposium on Information Visualization 2001*, vol. 73078, 2001.

[166] Y. Tu and H.-W. Shen, "Visualizing Changes of Hierarchical Data Using Treemaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1286–1293, 2007.

[167] M. Balzer, O. Deussen, and C. Lewerentz, "Voronoi Treemaps for the Visualization of Software Metrics," in *Proceedings of the 2005 ACM Symposium on Software Visualization*. ACM, 2005, pp. 165–172.

[168] J. J. Van Wijk and H. Van de Wetering, "Cushion Treemaps: Visualization of Hierarchical Information," in *IEEE Symposium on Information Visualization, 1999*. IEEE, 1999, pp. 73–78.

[169] P. Irani, D. Slonowsky, and P. Shajahan, "Human Perception of Structure in Shaded Space-filling Visualizations," *Information Visualization*, vol. 5, no. 1, pp. 47–61, 2006.

[170] F. Mansmann, D. A. Keim, S. C. North, B. Rexroad, and D. Sheleheda, "Visual analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1105–1112, 2007.

[171] J. Wood and J. Dykes, "Spatially Ordered Treemaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1348–1355, 2008.

[172] M. Jern, J. Rogstadius, and T. Astrom, "Treemaps and Choropleth Maps Applied to Regional Hierarchical Statistical Data," in *IEEE Information Visualisation 2009 Conference*. IEEE, 2009, pp. 403–410.

[173] A. Slingsby, J. Dykes, J. Wood, and R. Radburn, "OAC Explorer: Interactive Exploration and Comparison of Multivariate Socioeconomic Population Characteristics," *proceedings of GIS Research UK*, pp. 167–174, 2010.

[174] K. Buchin, D. Eppstein, M. Löffler, M. Nöllenburg, and R. I. Silveira, "Adjacency-Preserving Spatial Treemaps," in *Algorithms and Data Structures*. Springer, 2011, pp. 159–170.

[175] J. Wood, D. Badawood, J. Dykes, and A. Slingsby, "BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2384–2391, 2011.

[176] J. Wood, A. Slingsby, and J. Dykes, "Visualizing the Dynamics of London's Bicycle-hire Scheme," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 46, no. 4, pp. 239–251, 2011.

[177] F. S. Duarte, F. Sikansi, F. M. Fatore, S. G. Fadel, and F. V. Paulovich, "Nmap: A Novel Neighborhood Preservation Space-filling Algorithm," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2063–2071, 2014.

[178] M. Ghoniem, M. Cornil, B. Broeksema, M. Stefas, and B. Otjacques, "Weighted Maps: Treemap Visualization of Geolocated Quantitative Data," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 93 970G–93 970G.

[179] T. Dwyer, K. Marriott, and P. J. Stuckey, "Fast Node Overlap Removal," in *Graph Drawing*. Springer, 2006, pp. 153–164.

[180] ——, "Fast Node Overlap Removal Correction," in *Graph Drawing*. Springer, 2007, pp. 446–447.

[181] P. Bourke, "Calculating The Area and Centroid of A Polygon," 1988.

[182] B. B. Bederson, B. Shneiderman, and M. Wattenberg, "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 4, pp. 833–854, 2002.

[183] C. Tong, L. McNabb, R. S. Laramee, J. Lyons, A. Walters, D. Berridge, and D. Thayer, "Time-oriented Cartographic Treemap for Visualization of Publice Health Care Data," *Proceedings of the Conference on Computer Graphics and Visual Computing (CGVC), 2017*, 2017.

[184] J. A. Cottam, A. Lumsdaine, and C. Weaver, "Watch This: A Taxonomy for Dynamic Data Visualization," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2012, pp. 193–202.

[185] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale, "A Review of Temporal Data Visualizations Based on Space-time Cube Operations," in *Eurographics conference on visualization*, 2014.

[186] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in *IEEE Symposium on Visual Languages, 1996.* IEEE, 1996, pp. 336–343.

[187] S. Lee, S.-H. Kim, and B. C. Kwon, "Vlat: Development of a Visualization Literacy Assessment Test," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 551–560, 2017.

[188] C. Tong, L. McNabb, and R. S. Laramee, "Cartograms with Topological Features," *Technical Report*, 2018.

[189] R. S. Laramee, H. Hauser, L. Zhao, and F. H. Post, "Topology-based Flow Visualization, The State of The Art," in *Topology-based methods in visualization*. Springer, 2007, pp. 1–19.

[190] F. H. Post, B. Vrolijk, H. Hauser, R. S. Laramee, and H. Doleisch, "The State of The Art in Flow Visualisation: Feature Extraction and Tracking," in *Computer Graphics Forum*, vol. 22, no. 4. Wiley Online Library, 2003, pp. 775–792.

[191] E. River, https://englandexplore.com/england-rivers/, Accessed: June 2018.

[192] Openstreetmap, https://www.openstreetmap.org/, Accessed: June 2018.

[193] O. T. API, http://overpass-turbo.eu/, Accessed: June 2018.

[194] C. Ericson, *Real-time Collision Detection*. CRC Press, 2004.

[195] R. Laramee, "Using visualization to debug visualization software," *IEEE computer graphics and applications*, no. 6, pp. 67–73, 2009.

[196] R. Penn and D. Berridge, *Social Statistics-Four-Volume Set*. Sage Publications Ltd, 2010.