Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterarbeit

Truong Vinh Phan

Immersive Data Visualization and Storytelling based on 3D | Virtual Reality Platform: a Study of Feasibility, Efficiency, and Usability

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Truong Vinh Phan

Immersive Data Visualization and Storytelling based on 3D | Virtual Reality Platform: a Study of Feasibility, Efficiency, and Usability

Masterarbeit eingereicht im Rahmen der Masterprüfung
im Studiengang Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Philipp Jenke

Abgegeben am October 7, 2016

Truong Vinh Phan

Thema der Masterarbeit

Immersive Datenvisualisierung und Storytelling, die auf 3D bzw. virtueller Realität-Plattform basiert: eine Studie der Machbarkeit, Effizienz und Usability.

Stichworte

immersive Datenvisualisierung, 3D, visueller Data-Mining, virtuelle Realität, Open-Data, Big-Data, UX, Userbefragung

Kurzzusammenfassung

Seit der Datenexplosion dank der Open-Data- bzw. Transparenz-Bewegung sind Datenanalyse und -exploration eine zwar interessanter aber immer schwieriger Herausforderung, nicht nur für die Informationstechnik und Informatik sondern auch für unsere allgemeine Gesellschaft, geworden. Wegen der Arbeitsweise des menschlichen Gehirns ist Visualisierung eine der ersten Go-to Methoden, um komplexe Datensätze verständlich, anschaulich und zugänglich zu machen. Diese Arbeit untersucht aus Sicht der IT die Rolle der Visualisierung in Datenanalyse und Datenjournalismus sowie die Machbarkeit und Effizienz einer neuen Visualisierungsmöglichkeit in virtueller Welt, ermöglicht durch die rapiden Fortschritte in Virtual-Reality Technologien.

Truong Vinh Phan

Title of the paper

Immersive Data Visualization and Storytelling based on 3D | Virtual Reality Platform: a Study of Feasibility, Efficiency and Usability

Keywords

immersive data visualization, 3D, visual data mining, virtual reality, storytelling, open data, big data, UX, user study

Abstract

As data is being generated and flowing into modern society in mass quantity, exploring and analyzing these vast data volumes becomes an increasingly difficult challenge. The ability of human reasoning is greatly limited in traditional, mainstream data mining techniques. On the other hand, in data visualization, computational analysis only has a minor role. Visual data mining is a new approach which combines traditional mining techniques with information visualization in exploring large data sets. This thesis aims to study the role of information visualization in visual data mining and storytelling as well as examine how new types of data representation, especially with recent advancements in virtual reality technologies, can be used in combination with traditional visualization techniques to give data exploration and knowledge discovery a more immersive experience and how this approach can be stacked against conventional, purely two-dimensional visualization.

Contents

1 Introduction

1.1 Background

Knowledge transfer and discovery has always been a key part in the evolvement process of any civilization. Data has since long become a new type of raw material. And just like any other raw materials, it can be produced, processed, transformed and used to produce other materials and products. As ever more aspects of our daily life become connected in the webbed environments of urban landscapes, the sheer amount of information that is generated and consumed collects into massive databases and is set to bypass the zettabyte threshold by the end of 2016, according to Cisco's Virtual Networking Index (VNI) report (Cisco, 2016).

To human, a visual representation is often more effective than written text. Visual representations help us illustrate concepts and ideas — that if expressed verbally would be very difficult or even impossible to be fully understood, just as Edward Tufte once stated that "excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency" (Tufte, 2001). Colin Ware found out in his study that the human vision – a massive parallel processor made of the eye and the visual cortex of the brain, provides the highest-bandwidth channel of all the human senses (Ware, 2012a). Perception (seeing) and cognition (understanding) are considered closely correlated. A well-designed visual system works on the principle that when data is presented in certain ways, their patterns can be easily perceived, which would be otherwise difficult or impossible if presented in some other ways. Therefore, a good visual representation of data (i.e. data visualization) plays a key role in knowledge discovery and transfer, as well as data analysis.

Meanwhile, the transparency movement is gaining momentum, materialized through the popularity of open government data, which has its root dated back since the beginning of Web 2.0 (Tauberer, 2014). What this means for the field of information technology is a new horizon in data analysis. Using open data sets in combination with a wide variety of visualization software packages and tool sets – many of which were built upon the source code made available through another initiative called *Open Source*, great infographics and visualizations were produced to provide a great deal of new insights into various disciplines that had never before been found or touched.

A notable example is the use of the Global Positioning System (GPS), which became ubiquitous after the United States government made GPS signals readily available for civilian use from the year 2000 onward, in combination with open data feeds from various government bureaus, e.g., with U.S. Census Bureau on the nation's roads and the U.S. Geological Survey's satellite imagery and terrain data to create maps. One of the early applications in the modern open data movement were crime maps based on local police data. Adrian Holovaty's *ChicagoCrime.com* in 2005 was one of the first Google Maps mashups, and its successor

– Holovaty's EveryBlock, helped in jump-starting the open data movement (Tauberer, 2014). Among the journalists, the open data movement took its form under a new principle called data journalism, which utilizes open data feeds and interactive visualization to tell lively stories and produce credible reports.

1.2 Motivation and Goal of this Thesis

With the big data explosion in full swing, data is being resourcefully aggregated across multiple industries for various purposes, from business intelligence to military and scientific applications. Growing proportionally with the volume of data is its inherent complexity and cumbersomeness as well as the number of data dimensions, like Chavez had emphasized in his IEEE presentation on Virtual Reality (VR) and Visualization (Chavez, 2014), which has elevated the analysis task to a serious challenge — sometimes goes beyond today's capability. One of the ongoing challenges of information visualization is to utilize and combine latest technology with the natural but powerful capabilities of human cognition to extract knowledge and advantage from the information. With data represented in traditional 2D formats, ranging from raw data tables to different types of chart | graph | plot (e.g., pie chart, scatter plot, etc.), there is a limit to how much information or actionable insights we can actually take out and use for making decisions, planning, targeting a specific group of interest, etc. Business intelligence software packages try to address this shortcoming by automatically providing insights and highlight standouts. Furthermore, we have data mining solutions which are designed to help uncover hidden trends in the data. Still, if we solely rely on them, there is still a whole layer of possible trends and knowledge buried under the mountains of data, yet to be uncovered.

Information visualization has been a topic of discussion over the years with many new, inspiring ideas and concepts coming up. David McCandless presented in his 2010 TED Talk "The Beauty of Data Visualization" great examples of how data can be transformed into a whole new landscape, explorable with human eyes, by presenting it visually in combination with some aesthetics elements (McCandless, 2010a). As we approach a new era of cheap and powerful computing power that sees fast-paced advancements in mobile and web technologies as well as 3D imagery, a whole new world of possibilities is opened for the field of data visualization. 3D imagery in Google Maps is one of the great examples for this, as well as *zSpace*, an interactive 3D virtual holographic platform that can be used to build visualization solutions aim to analyze big data (Chavez, 2014). Physically visualizing data further empowers human cognitive processes and enables us to see what we might not be able to see with normal 2D landscape. VR is one of the few platforms that have been getting rapid developments in recent years. Unlike traditional user interfaces, using VR we can create a simulated environment that places the user inside an immersive experience. Instead of looking at a flat, 2D screen, users are "immersed" and able to interact with 3D worlds. That makes VR a prominent candidate as a medium for a new way of storytelling, because in VR we create worlds and invite users to experience them.

Taking advantage of a variety of widely available VR viewers and the maturity of VR technology, the main goal of this thesis is to design a 3D visualization prototype in VR environment that aims to visualize time-series and chronologically ordered data, and use it in combination with open data to study whether immersive visualization in general might be proven better in term of user experience and knowledge discovery/delivery than conventional “flat-screen”, 2D counterparts. Among the aspects that will be taken into consideration in this study are feasibility, efficiency and usability. For this purpose, two variants of the newly developed visualization prototype will be implemented — an interactive 3D visualization with raw data tables and traditional 2D charts on desktop platform, and a VR variant of the same prototype to work on the Google Cardboard platform. The visualizations will be using education data sets made available by the Integrated Post-secondary Education System (IPEDS) [1]of the U.S. National Center for Education Statistics (NCES) and the open data portal of the United States government[2]. The study will conclude with a small user study to evaluate user experience based on impressions and feedback. Thus, we can roughly have a basic understanding of how the new immersive approach might be appeal and beneficial to a wide spectrum of users, from end users and decision makers to journalists.

1.3 Restrictions

The visualization prototype presented and demonstrated in this thesis makes use of open education data sets published by the Integrated Post-secondary Education Data System of the U.S. National Center for Education Statistics, also made available on the U.S. government open data portal[3]. Since these data sets belong to the public domain, Chapter 2 will be touching upon the topic of Big Data and Open Data and discuss their fundamentals, with restriction to this domain only. Data from other domains, e.g., private sectors, IoT, ubiquitous computing, etc., comprises a very broad subject that stretches across multiple scientific disciplines (Dumbill, 2012), and therefore is beyond the scope of this thesis. The main focus of this thesis is on multi-variate data visualization, its techniques and tool sets as well as the basics of evaluation methods for information visualization, concretely through a small-scale user study of the visualization prototype developed within the scope of this thesis, with respect to user experience (UX) and efficiency.

Finally, it should be noted that there is a limit as to how in-depth the topics in this thesis will be covered. Going thoroughly through each of the topics and their relevant aspects is not possible due to the limited scope of this thesis. Therefore, at a minimum, only the

[1] http://nces.ed.gov/ipeds/datacenter/Default.aspx

[2] https://www.data.gov/

[3] https://data.gov

most fundamentals of each topic will be covered and discussed. Readers who wish to delve further into any specific topic or seek a broader and in-depth discussion are advised to refer to relevant literatures as well as online resources.

1.4 Structure of this thesis

Chapter 2 touches upon the topic of Big Data and Open Data. This includes a general introduction as well as the fundamentals with basic terms and definitions relevant to said topics. It is followed by a discussion on the potentials of Big Data / Open Data, which also provides a brief overview of the current situation of the open data movement in Germany and abroad, especially in the United States – since this thesis uses open data from the U.S. government, as well as relevant developments and challenges in the said field.

In Chapter 3, a selective number of important techniques and technologies that power many open data platforms will be introduced and discussed. These techniques and technologies are behind many open data / big data platforms and help make the data accessible to a broader audience, including for civilian and non-civilian use, individuals as well as organizations. They include but not limited to big data-related technologies that for example, pave the way for easier access to huge structured and unstructured data sets, the platforms upon which open data infrastructure is built and software packages that power open data portals.

Chapter 4 focuses on the practice of extracting knowledge / insights from large data sets. It will touch upon the topic of data-mining / knowledge discovery and thus, includes an introductory overview of the terms, concepts as well as a brief discussion of some typical data-mining algorithms and processes. This chapter serves as a basic introduction into big data analysis and interpretation.

Following Chapter 4, Chapter 5 delves into another way of making sense of large data sets, and is also the main focus of this thesis: *visualization*. Starting with a usual introduction of visualization, Chapter 5 goes on to provide a brief history of information visualization, from the very early use of data visualization to acquire knowledge (Nightingale (1857), Snow (1855)) to the recent, modern developments of the field, as well as its utilization in data journalism. It continues with a discussion on the role that visualization plays in solving the big data analysis challenge, as well as the basic scientific grounds behind it. The chapter then goes into several important and most fundamental techniques in visualizing data, including networks / hierarchies, groups, interactions, etc., and also presents some typical algorithms that utilize these techniques. Chapter 5 also gives a brief introduction into a process known as visual data-mining and concludes by examining an example software tool call V-Miner that is specifically geared towards visualization of multi-variate data sets.

Chapter 6 goes deeper into the focus of this thesis and presents a new prototypical visualization approach, designed to visualize timelines and chronologically ordered data and make use of 3D technology, which is called the *StreamViz*. It starts by discussing the concepts and ideas around the StreamViz, as well as defining requirements for the visualization prototype. The chapter goes on by presenting some early sketches and prototypes for these concepts and ideas, then describes the data acquisition and implementation process of the StreamViz demos through a concrete use case. The final part of this chapter gives some thoughts on the end results of the visualization prototype, as well as outlines a few possible ideas to evaluate the StreamViz. It concludes with a detailed description of the chosen methodology of the evaluation process, as well as an in-depth analysis and discussion of the assessment results.

Chapter 7 serves as the concluding chapter of this thesis, offering a few final thoughts and observations on what had been learned, as well as providing a quick summary of the whole thesis. The chapter concludes by outlining a few ideas and prospects for possible future work on the topics presented in the thesis.

2 Big Data and Open Data

2.1 Introduction

The topic of big data | open data serves as one of the foundations for this thesis. This chapter offers a quick overview of the big data landscape, provides an example of how governments open their data, and reports on the current status of the open data | transparency movement. Section 2.2 introduces some basic terms and concepts around big data | open data and the transparency movement, both in general and with regard to visualization. It then goes on describing the potentials of big data | open data and what kind of benefits the transparency movement could bring in Section 2.3. The recent developments of the open data ecology in the United States, Germany and around the world as outlined in Section 2.4 help provide the reader with a sense of how relevant this topic is in the evolvement of today's society. And like every other movements and developments, there also exist various challenges to the open data ecology, and those will be discussed in Section 2.5 of this chapter.

2.2 Terms and Concepts

2.2.1 Data, information and knowledge

Data, information and knowledge are terms used in almost all disciplines out there such as psychology, medical sciences, epistemology, military, etc., most extensively and relevant in computer science and engineering with different and competing definitions. In some cases, the use of these three terms is not consistent and even conflicting. Generally, the terms *data* and *information* are used interchangeably, i.e., data processing and information processing can be considered more or less the same. In specific contexts, such as from a system perspective, data is understood as the bits and bytes stored on or transmitted through a digital medium. Literatures such as (Cleve and Lämmel, 2014) define data as a series of symbols with corresponding syntax, is a basis material to be processed by IT-systems and differentiate between these three terms in such a way that when the data has a meaning, it will then become information. For instance, Cisco's Virtual Networking Index (VNI) 2016 report for annual global IP traffic (Cisco, 2016) emphasizes that by the end of 2016, the amount of data generated globally will pass the zettabyte threshold. According to this definition, if one zettabyte is the data, then its context meaning of "annual global IP traffic" turns it into information. Now still according to this definition, if by using the data collected over a period of time, a rule to forecast the amount of annual globally generated data can be derived, then this will become knowledge.

In the visualization field, the terms data, information, and knowledge are often used in an interrelated context and indicate different levels of abstraction, understanding and truthfulness (Chen et al., 2009). In literatures concerning visualization, we often find phrases such as "the primary objective in data visualization is to gain insight into an information space", or "information visualization is part of data mining and knowledge discovery process" (Fayyad et al., 2002). It is suggested in (Chen et al., 2009) that these three terms may serve as both input and output of a visualization process.

Data can be classified into *structured, semi-structured* and *unstructured* category. Unstructured data represents about 80% of all data, and can be understood as data from which it is (very) difficult to extract knowledge using any kind of automated processing or algorithms. It often contains text and multimedia content that is very difficult to fit into a relational database. Examples include emails, word documents, images, audio | video files, PDFs, etc. Unstructured data is mostly machine-generated (e.g., satellite images, scientific data, photos and video, etc.), but also comes from human (e.g., texts, website content, social media data, etc.). Semi-structured data is data that, although not resides in relational database, does possess to some extent organizational properties such that it is easier to analyze. Examples are CSV, XML or JSON documents. NoSQL databases can also be considered semi-structured. Most semi-structured data can be stored in relational databases after some processing, and they represent about 5-10% of all data. Structured data is data that has clear structures, types and order. Example of structured data is all data that can be stored in a relational database, in tables with rows, and columns. They often have relational key and can be mapped to pre-designed fields. Structured data is the easiest to process but like semi-structured data, it only makes up around 5-10% of all data.

In perceptual and cognitive space, Eliot first mentioned a popular model for classifying human understanding called Data-Information-Knowledge-Wisdom hierarchy (DIKW), which agrees that data, information, and knowledge are three distinct concepts and different from each

Category	*Definition*
Data	Symbols
Information	Data that is processed to be useful, providing answers to “who”, “what”, “where”, and “when” questions
Knowledge	Application of data and information, providing answers to “how” question.

Table 1: Russell Ackoff’s definitions of data, information, and knowledge in perceptual and cognitive space. Source: Chen et al. (2009)

other (Sharma, 2008). Table 1 shows definition of data, information, and knowledge in perceptual and cognitive space according to (Ackoff, 1989). In computational space, however, Chen et al. (2009) suggested that data is an overloaded term, and can be referred to both information and knowledge as a special form of data. These definitions are shown in Table 2.

Category	Definition
Data	Computerized representations of models and attributes of real or simulated entities
Information	Data that represents the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings
Knowledge	Data that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or the transcripts of some knowledge acquired by human beings

Table 2: Definitions of data, information, and knowledge in computational space, according to Chen et al. (2009)

2.2.2 Big Data and Open Data

Although Big Data is a relatively new concept, it has already had various different definitions. One of the earliest definitions of the concept — back when it began to gain momentum in the year 2000, was from industry analyst Doug Laney in his 2001 report (Laney, 2001), as a combination of the three Vs:

- Volume: big data is massive in volume, being collected from a wide variety of sources: business transactions, social media, sensors, machine-to-machine, etc., resulting in the challenge of data storage.
- Velocity: data flows in at an unprecedented speed due to technological advanced

equipment: RFID tags, sensors, smart metering, etc., resulting a challenge for near-real time data processing.

- Variety: data comes in all possible types of formats, from structured to semi-structured and unstructured, resulting in a need for effective data cleaning and transformation prior to processing.

The transparency movement dated back in the 1990s with the Transparency International founded in 1993 to fight corruption. Since then, many other initiatives for openness have followed suit, among which are the open government initiative and the open government data concept.

The The Open Knowledge Foundation (2014) defines the concept of *open knowledge* as:

> Any content, information or data that people are free to use, re-use and redistribute — without any legal, technological, or social restriction.

According to this definition, open data and content are the building blocks for open knowledge. They become open knowledge when they are useful, usable, used — and therefore have several requirements, which are partly based on the Open Source Definition. To be called *open,* the data must be available and made accessible (possibly through an internet resource for download) in a convenient and ready-for-modification form, and at a reasonable cost (or completely free of charge). Moreover, the data must be provided in a machine-readable format and with a license that allows reuse and redistribution, possibly together with other data. Furthermore, the data is meant for everyone's use (*universal participation*). There must not be any discrimination or restrictions against any fields, organizations, or individuals, including permission for commercial use.

Open data comes from a wide variety of sources, including both public and private sectors (government, NGOs, education institutions, businesses, etc.), and covers most major fields, such as economics, finance, healthcare, education, labor, social science, technology, agriculture, development, and environment.

2.2.3 Open Government and Open Government Data

The idea of an open-to-public-scrutiny government dates back to as early as the 18th century in Europe. The United States passed its Freedom of Information Act in 1966. Following that, similar or equivalent laws were also passed in Europe and other countries around the world with Germany and India in 2005. The Memorandum on Transparency and Open Government[4], brought to life in 2009 by the Obama Administration, gives a clear vision of what an open government should look like. It emphasizes on government transparency (promotion of government accountability), government participation (promotion of public engagement),

[4] https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_fy2009/m09-12.pdf

and government collaboration (promotion of collaboration between government, NGOs, businesses and private sector individuals).

The first and foremost requirement to enable government transparency is to provide the public with information about government | public bodies activities. This information is to be disclosed in open data formats and without any restrictions on reuse (Open Government Standards, 2016). This brings up the concept of *open government data*, which is defined by the Working Group on Open Government Data at the Open Knowledge Foundation as (Open Government Data, 2016a):

> Data produced or commissioned by government or government controlled entities.
>
> — and —
>
> Data which is open as defined in the Open Definition – that is, it can be freely used, reused and redistributed by anyone.

and should adhere to eight principles according to Carl Malamud (Open Government Data, 2016b): complete, primary, timely, accessible, machine-processable, non-discriminatory, non-proprietary, and license-free. Furthermore, the following properties for open government data are recommended: online | downloadable | free, promote analysis (ACM recommendation), safe file formats, public input/review, has provenance, web-searchable, possess global unique identifiers (GUIDs), linked open data (semantics) (Tauberer, 2012).

The work in this thesis is wholly based on open government data, from the education sector. And because today, the open government data concept is often used interchangeably with the open data concept (Dietrich, 2011a), the term *open data* in this thesis will henceforth be used to refer to open government data.

2.3 Potentials of Big Data | Open Data

Open data represents a huge resource for government, public sector as well as both organizations and individuals alike. It brings enormous value for the economy as well as benefits for the society. It is already possible to name a few of the areas where open data is being used to creating value, including transparency and democratic control, improving/creating new private products/services, more efficient and effective government services, knowledge discovery, etc. Projects like the Finnish visualization "*Tax Tree*"[5] are being set up around the world to help improve transparency by tracking government activities, tax spending, etc. (Web) services like Germany's "*mapnificient*"[6] which takes into account various inputs to help people find places to live, *"vervuilingsalarm.nl*"[7] in the Netherlands to measure air quality, are great examples of how open data is used to power self-empowering projects.

[5] http://www.hri.fi/2years/3-apps4finland.html

[6] http://www.mapnificent.net/

[7] HTTP://ourservices.eu/?q=taxonomy/term/43&page=1

For businesses and the economy, open data represents a annual market worth tens of billions of Euros in the EU (Open Data Handbook, 2016), with a good example being Google Translate using a vast amount of documents in all European languages to train its translation algorithm. The governments and NGOs of course, also benefit from the data they open through public participation and collaboration on analyzing open data sets — government services and procedures are being improved and become more efficient. The concept of "*civic hacking*" continues to take form, a notable example being the non-profit Sunlight Foundation in the U.S., whose goal is to improve government transparency and accountability through the use of technology. Lastly, open data often contains untapped potentials, hidden knowledge, and insights. As these knowledge and potentials are being uncovered through mining processes, new fields of application and possibilities are created which all contribute towards better public infrastructures and improving life quality.

2.4 Current Developments and Challenges of Open Data

The development and implementation of the transparency concept and the principles of open data are gaining momentum around the world. Still, they are not happening equally in all countries and in different areas of the society. As of 2015, the Center for the Development of Information and Communication Technologies in Spain had already implemented some 280 open data projects around the world (CTIC, 2015). In the U.S., the open data trend only began to truly take off in 2009 after President Obama's Open Government Directive , with numerous conferences on transparency already taken place and apps being developed. Today, the official open data portal of the U.S. government houses more than 183,946 data sets from a wide variety of fields, from agriculture to public safety (Data.gov, 2016). Similarly, in the U.K., data is also being opened through the government's official open data portal[8]. These projects have paved the way for similar open data projects in other countries like Japan[9], Australia[10], and Canada[11], etc. (Forsterleitner and Gegenhuber, 2011).

Other countries, most notably European countries, have had a fairly early start with open data, partly due to historical factor as well as early society's understanding and recognition of transparency concept (Wikipedia, 2016). In Germany, as of 2015, the situation of open government data and transparency looks much positive, as it is being implemented on a nation-wide scale — from federal down to regional and municipal level. The fact that various statistical data sets are already being made available on the Federal Statistical Office data portal[12] is one of the good examples of transparency at federal level.

[8] https://data.gov.uk

[9] http://www.data.go.jp/

[10] http://data.gov.au/

[11] http://open.canada.ca/en

[12] https://www.destatis.de/EN/Homepage.html

At regional level, various states and cities either already issued laws on transparency, for example in the state of Bremen, data is being opened obligatorily based on the state-wide Freedom of Information Act (Freie Hansestadt Bremen, 2016). In the state of Hamburg, the Freedom of Information Act has since 2012 been superseded by the transparency laws that require government agencies to open even more data to the public. This also includes institutions and companies that work on government and public projects. Since 2014 Hamburg has already pioneered its own open data portal[13], which offers around 10,000 data sets and counting as of 2016.

Although the concept and principles of open data have been widely recognized and slowly adopted around the world, there still exist challenges. Due to differences in political and law system, the interpretation and practical implementation of those concepts and principles also vary and are sometimes very difficult. In Germany for example, the fact that various phases in the flow like data gathering, preparing / processing and making data accessible are distributed and the responsibility of various bodies has a profound impact on open data (Dietrich, 2011b). The data is being opened in a not fully standardized manner, with data format varies between inappropriate and proprietary which leads to poor machine-readability and greatly limits usability (e.g., due to web- | mobile-incompatible). Another possible concern raised by Dietrich (2011b) is the differences and incompatibility in licensing, which further hinder reuse and shareability of the data. He also named a more complex challenge – the heterogeneity of the vocabulary and classification system used to define the data semantically. This phenomenon can be understood as being caused by the lack of a set of global standards and it happens at various levels of data aggregation and evaluation.

2.5 Conclusions

Though the open data movement only truly gained momentum a decade back, it has gone a long way since its early initial stage as only a concept, and have been seeing great, continuous developments and rapid rate of adoption in recent years. Of the remaining challenges, the most important one is to develop a decentralized approach to the open data processes and workflows, which is based on open global standards that unify data formats and licensing, so that data reuse and opening process will not be limited. In order for this to succeed, more laws on transparency / freedom of information must be issued or changed to require government bodies and agencies to actively participate in the program, besides an active participation from the people to further promote this concept. This is not only the case in Germany, but also in various countries around the world that still face those same challenges.

[13] http://transparenz.hamburg.de/

3 Big Data | Open Data Techniques and Technologies

3.1 Introduction

In order to publish data and make it open, there are several rules and principles that need to be adhered to, as already introduced in Chapter 2. The main goal is to make data accessible, ready to be reused without any limitation. This chapter introduces the technical infrastructures needed to power such a platform that can be used to house open data and expose it to the public. Section 3.2 discusses the technologies needed to store and process large data sets, and also relates them to the big data context with an emphasis on the storage and processing.

The interfaces to an open data platform also play an important role in fulfilling the high-accessibility requirement, because they provide the means for different groups of users to access and use the data. It is therefore of great importance that open data platforms and software packages provide highly-optimized APIs, made ready for this purpose. Section 3.2 also gives a quick overview of some of the popular open data software packages and platforms, then goes on to introduce the concept of "*Linked Data*" in Section 3.3 as a method to expose, share, and connect pieces of data, information, and knowledge on the semantic web – which enhances usability and re-usability of open data.

3.2 Technologies and Platforms

As already introduced in Chapter 2, Big Data is a term often seen in the Open Data ecology, because the data that is being produced and opened mostly fulfills the three-Vs of Big Data. These data sets are typically unstructured, massive in volume, and contain valuable insights | knowledge (Klein et al., 2013). To effectively process and mine these data sets, there exist special storage and processing technologies that are specifically designed with Big Data in mind.

3.2.1 NoSQL databases

NoSQL (Not-Only SQL) databases are designed with a non-relational concept in mind and therefore typically target big, unstructured data sets. These database systems are designed to fulfill use-cases where traditional relational database systems fail or operate very inefficiently, typically when data needs to be created and inserted much more frequently than it needs to be modified (Klein et al., 2013). NoSQL databases are divided into four categories:

- *Key-Value Stores*: the main concept behind key-value stores is the use of a hash table containing unique keys and pointers to particular items of data. It is the simplest to implement but is inefficient for partial queries (update of part of a value). The *in-memory* variant offers advantages in performance, while the *on-disk* allows for more reliable data persistence. Example systems include Redis, Voldemort, Amazon SimpleDB, Oracle BDB.

- *Column-Family Stores*: these are designed to store and process very large amount of data distributed over many nodes. The keys points to multiple columns, which in turn are arranged by column family. Notable examples are Apache Cassandra and HBase.
- *Document Stores*: similar to key-value stores and inspired by Lotus Notes, the model of these systems is basically versioned documents that are in turn collections of other key-value collections. These semi-structured documents are typically stored in formats like JSON and the system allows nested values associated with each key to enhance query efficiency. Well-known examples include MongoDB and CouchDB.
- *Graph Databases*: a flexible graph model is used instead of tables of rows, columns and a rigid structure of SQL, which again allows distributed scaling across multiple nodes.

3.2.2 Big Data platforms, e.g., Apache Hadoop

This section discusses Big Data platforms through the example of Apache Hadoop, one of the most well-known open source framework, which is specially designed to handle Big Data distributed storage and processing, and run on clusters of computing nodes. Hadoop is a collective of several frameworks built around Big Data, which together form an ecosystem. These software packages can be deployed on top of Hadoop or run alongside it. At its core, Hadoop consists of two main components: a data storage module called *Hadoop Distributed File System* (HDFS), a data processing module called *MapReduce* — inspired from the same concepts by Google — and a resource management platform to manage computing resources on clusters called *Hadoop YARN*. Other notable software packages in the ecosystem include Apache Hive, Apache HBase, Apache Spark, etc. Each of which provides special services for Big Data like distributed data storage; warehousing infrastructure for data summarization, query and analysis; scheduling, distribution, and coordination of jobs; managing and supervising of nodes and clusters; collecting, aggregating, and moving large amount of log data; etc.

Hadoop works by splitting data into large blocks and distributing them across computing nodes in a cluster, then transfer packaged code to these nodes to process their own data in parallel. This approach is implemented based on the MapReduce programming model for processing and generating large data sets using distributed, parallel algorithm on a cluster. The *Map* procedure performs sorting and filtering, while the *Reduce* method performs aggregation | summary operation. The framework automatically orchestrates processing by marshaling distributed nodes, managing communication and data transfer between them, and providing redundancy as well as fault tolerance. MapReduce is therefore an implementation of the *divide and conquer* paradigm.

3.2.3 Available software packages for Open Data portals

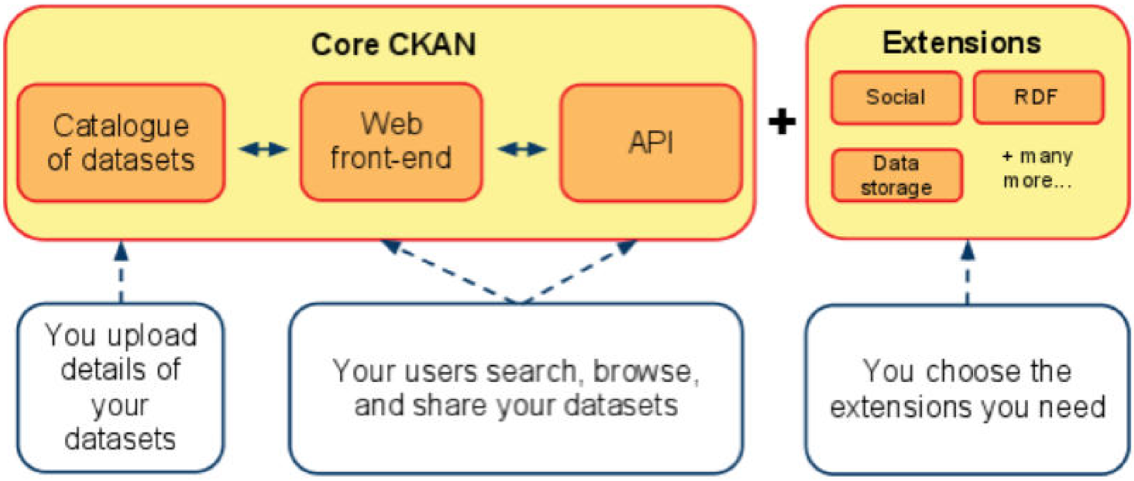

Core CKAN
Catalogue of datasets
Web front-end
API
Extensions
Social
RDF
Data storage
+ many more...
You upload details of your datasets
Your users search, browse, and share your datasets
You choose the extensions you need

Figure 1: The CKAN's architecture. Source: Open Knowledge Foundation - CKAN Team

Traditional content management systems (CMS) are not normally suitable for Open Data scenarios, because they are designed for different use cases with different requirements in mind. Currently, there are only a few well-known software packages on the market — either free under Open Source license or commercial based — that fulfill higher requirements that Open Data scenarios often demand. *Socrata Open Data Portal*[14] is an example of the commercial package that powers the open data portal of the World Bank, San Francisco, and New York among others.

Among the free and open-source based packages is the highly popular *Comprehensive Knowledge Archive Network* (CKAN)[15], which is a project from the Open Data Foundation itself. It currently powers approximately 151 portals around the world (CKAN, 2016), ranging from those of public organizations and government agencies (including *Data.gov* of the U.S. and *Data.gov.uk* of the U.K.) to those of businesses that want to join the Open Data ecology. The architecture of CKAN is modular, highly extendable, and the platform works similar to a CMS but is used to manage and publish collections of data instead of pages and entries (Figure 1). Other than that, CKAN offers all utility functions such as search, and an API to access published data programmatically. At its core are three main components: a data catalog that holds all the data sets; a comprehensive user interface for sysadmin, organizations,

[14] https://socrata.com/products/open-data/
[15] http://www.ckan.org/

data publishers to access and manage the data catalog, and for end users to browse, search, filter, and preview all the data sets in the catalog; and a RESTful API that allows access to all the platform's functionalities programmatically. The CKAN platform architecture is extensible through many extensions that allow tailoring of the platform to specific needs and use cases.

The basic unit of data in CKAN is the data set. A data set contains metadata (title, name, data, publisher, etc.) about itself and resources (a.k.a. raw data), which can be in a variety of formats like spreadsheets, PDFs, images, etc. The resources (or raw data) are stored in either the *FileStore* or *DataStore*, with the latter is an extension to the former, to allow for an *ad-hoc database* to store structured data from CKAN resources and provides data preview as well as an API to manage and update the data *on-site* (without having to re-upload).

CKAN resources can be automatically previewed, explored, and visualized, depend on the data type. This is enabled using Recline.js JavaScript library, which is also developed by the Open Data Foundation. Data sets from other CKAN portals can also be "*harvested*" — a data import process from one portal into another. Data sets can be accessed either through the graphical web interface or via the RESTful API, which allows other apps to make use of the data offered by the CKAN portal.

Although a data set may contain big resources with millions of data entries, there are currently only very few, if not zero, CKAN portals that hold catalogs of millions of data sets. The main database engine that CKAN utilizes to manage data sets is PostgreSQL, which, comparing to NoSQL databases, is not ideal to handle great amount of unstructured data. However, the Apache Solr enterprise search platform provides good scalability as the data in the CKAN platform grows.

3.3 Linked Data | Open Data

To reuse or extract knowledge | insights from data efficiently, *relationships* among data must be formed and made available. *Linked Data*[16] is an approach to structured data publication, in such a way that it is interlinked and allow for semantic queries, better inter-operable data exchange, and reevaluation process. Wikipedia (2016) defines Linked Data as:

> "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."

As outlined by Tim Berners-Lee — one of the pioneers of Semantic Web — in his note (Berners-Lee, 2009), Linked Data should satisfy four principles:

1. Use URIs for identifiers.
2. Use HTTP URIs so that resource can be looked up (interpreted, de-referenced).

[16] http://linkeddata.org/

3. Provide useful meta information about an identifier when it is looked up using standards like RDF, SPARQL, etc.
4. Refer to other resources by their HTTP URIs for web publishing.

The concept of Linked Data revolves around the Semantic Web, which is a Web of Data (of dates, titles, properties, etc.), or any other data we might think of. Semantic Web offers a variety of technologies, such as RDF, OWL, SPARQL, etc., to query data and extract inferences using vocabularies, etc. The first requirement for the Web of Data is that the data itself must be in a standard format, accessible, and manageable by these technologies. The second requirement is that the data must be interlinked and relationships between them must also be available. The W3C provides a set of technologies, such as RDF, R2RML, RIF, SPARQL, etc., to achieve and create Linked Data, as well as to set up endpoints to query the data more efficiently.

Linked Data is what powers the Semantic Web, which enables large scale integration of, and reasoning on, data on the web. A good example of large linked data set is DBPedia[17], which makes the content of Wikipedia available in RDF and contains links to other data sets, e.g., Geonames, on the web. This enables applications to make use of knowledge from a variety of data sets, thus produce better value and user experience.

Linked Open Data is essentially Linked Data with an additional rule besides the four principles mentioned above: the data must be open content. Tim Berbers-Lee also suggested a 5-star rating system for different levels of data openness, from the most basic and arbitrary data with open license ($\star$); data that is structured or machine-readable, such as Excel spreadsheets ($\star\star$); data that is in open, non-proprietary formats such as CSV ($\star\star\star$); data that possesses URIs to allow for linking—RDF standards complied—($\star\star\star\star$); and finally, data that links to other data to provide contexts—linked RDF— ($\star\star\star\star\star$).

3.4 Conclusions

As new standards for the Open Data ecology are being slowly developed and set in place, we are seeing more and more open data sets on the Web as well as practical applications that make use of them.

The challenges that pertain to Big Data also apply to the Open Data ecology as well, because at its core, open data is often *"big"* in volume. Fortunately, techniques and technologies specific to Big Data can also be used by Open Data publishers and developers to tackle these challenges.

Practical implementation of Open Data is made even easier and more accessible today with an open source platform developed by the Open Knowledge Foundation called CKAN, which

[17] http://wiki.dbpedia.org/

offers a solid architecture and infrastructure specially designed for publishing, managing, and opening data sets, as well as an API for application developers to build a useful ecology around these data sets.

The Linked Data concept lies at the heart of the Semantic Web, which is an approach that adds more value to the existing open data sets and makes them interlinked, so that more potentials for knowledge discovery are opened up. Although there exist only relatively few large linked data sets in practice, with a variety of standards already in place, creating and publishing Linked Data is becoming more accessible.

4 Making Sense of Data

4.1 Introduction

One of the core values of Big Data is that it potentially has valuable insights | knowledge hidden in the vast amount of raw data, and it is these knowledge and insights that are of most interest to the user. Businesses use them to enhance sales, improve performance and customer relationship, etc. Data journalists use raw data and knowledge derived from it to support their reports and articles, to provoke thoughts and debates, and to tell stories.

To extract possible insights, trends, and knowledge from large amount of data, traditional and common statistical techniques are often not enough and effective anymore. That is where the concept of *data mining* becomes relevant. Data mining is a process of extracting and discovering patterns, which involves multiple steps such as analysis, database management, data pre-processing, post-processing of discovered structures, and *visualization*, among others; and is the analysis step of the Knowledge Discovery in Databases process (KDD) (Fayyad et al., 1996). This chapter introduces fundamental terms and concepts, as well as common techniques and workflows of the knowledge extraction process

4.2 Data Mining and Knowledge Discovery

Knowledge Discovery in Databases, or KDD, is an interdisciplinary field which intersects a variety of sub-fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence, data visualization, and high performance computing, among others. The goal of KDD is formally defined by Fayyad et al. (1996) as the non-trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. It is considered a multi-step process which includes data preparation, selection, cleaning, etc., and encompasses activities such as data storage and access, scaling algorithms to massive data sets, result interpretation, and visualization among others, as well as strive to improve the level of efficiency of these activities. Still according to Fayyad et al. (1996), there is a clear distinction between KDD and data mining in that data mining is a particular sub-process of KDD — a step that consists of applying data analysis and discovery algorithms automatically or semi-automatically that produces a particular enumeration of patterns | models over the data . The KDD process as a whole, ensures useful knowledge is derived from the data, as opposed to blind application of data mining, which could lead to false and invalid patterns. This section however, focuses on the various aspects of data mining including its applications, in which the term *data mining* will be used synonymously to the KDD process. Data mining today has application across industries and businesses among other fields e.g., marketing, finance (including investment, banking, e-commerce), fraud detection, manufacturing, advertising, sports, etc., with varied goals — to target customers more effectively, or to increase production efficiency, minimizing risks and maximizing revenues — for instance.

In the following section, the basic flow of a KDD / data mining process will be briefly discussed.

4.3 Basic Flow of a KDD Process

4.3.1 KDD: a seven-step process

As with any other process, defining a clear goal is considered the very first and most important step. As for requirements, the application domain and prior relevant knowledge must be understood. The following flow of the KDD process is based on the research paper of Fayyad et al. (1996). After the goals of the KDD process are clearly defined, the next step is to create a target data set:

1. *Selection of a target data set*: the target data set is the data set upon which the KDD process should run. Depending on the predefined goals, sometimes it is sufficient to just focus on a subset of variables or data samples of the whole data set.

2. *Data cleaning and pre-processing*: activities include removing noise, which consists of false data and missing data, etc. among others. If the data set contains a lot of noise, strategies must be devised to model and account for it. E.g., how to treat missing data fields, handle time-sequence data, and account for known changes.

3. *Data transformation*: depending on the selected mining algorithm, data is transformed into appropriate database scheme and data reduction | projection is applied to find invariant representations for the data based on the pre-defined goals. For example keywords and character encoding are assigned to textual attributes, or dimensionality reduction operation with for example, classification strategies, value intervals, etc., is carried out to reduce the number of variables in consideration.

4. *Method selection, model and hypothesis formation*: depending on the pre-defined goals, appropriate mining methods will be selected. Among the most common methods are classification, regression, summarization, clustering, etc. Then using exploratory analysis, appropriate models and parameters will be selected and the mining methods will be matched with the overall criteria of the KDD process.
5. *Mining data for patterns*: the selected methods and algorithms will be run against a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering to search for interesting patterns. By going through all the previous steps, the user provides significant aid to the mining method, helping it run faster and more efficiently.
6. *Interpretation | evaluation of mined patterns*: based on pre-defined criteria, the mined patterns will be interpreted and evaluated. If necessary, step 1 through 6 will be re-run with appropriate adjustments. A few possible criteria to determine whether the mined patterns are of interest might be, for example, *validity*, *uniqueness*, *usefulness* and *comprehensibility*. This step can also involve visualization of the results (and the models), as well as of the raw data set given the models.
7. *Acting on the discovered knowledge | insights*: a few typical actions include incorporation of new knowledge into other systems, documenting and reporting the newly found knowledge to interested parties, as well as double-checking and resolving any possible conflicts between the new knowledge and the current beliefs | hypothesis | knowledge.

The KDD process can be iterative, at the step-level: each step can be repeated multiple times until the process is ready to move on to the next step. Figure 2 illustrates all the steps introduced above.

4.3.2 The role of visualization in the KDD process

Visualizing the mined patterns helps the user in the process of understanding the knowledge and any possible relationships to the data, as emphasized by Cleve and Lämmel (2014): “A good visualization is essential for the success of any data mining projects”. Cleveland (1993) also supported this view by emphasizing:

> “Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones.”

Visualizing the results also help the user verify the validity of the mined patterns, and later transfer or present the found knowledge to other users effectively. How the user interpret the mined patterns also depends on other factors as well, for example domain knowledge, but visualization plays a decisive cognitive role in this interpretation. There are a few other

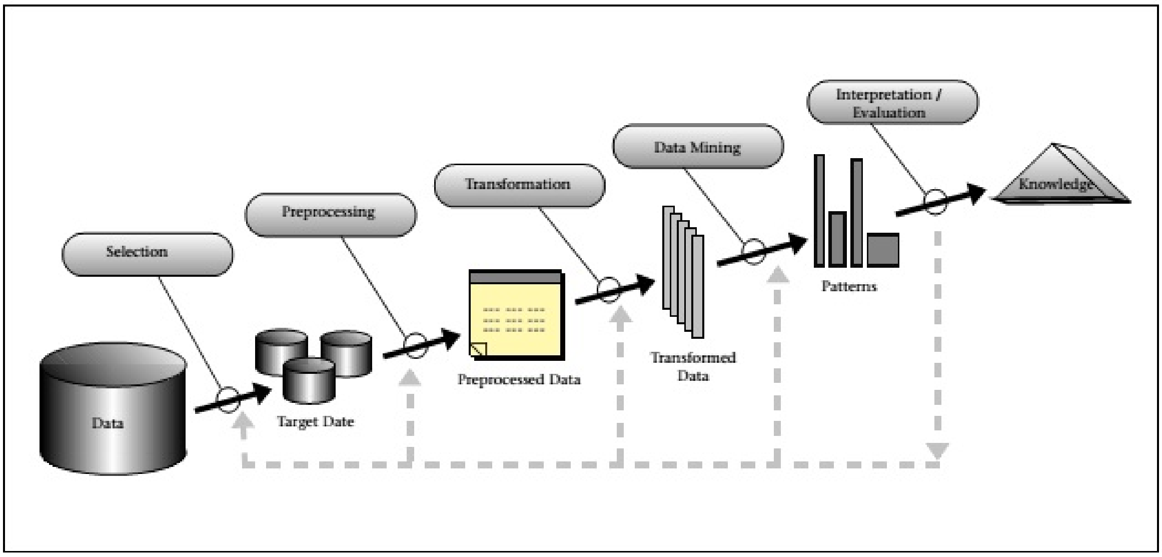


Figure 2: An overview of the steps that compose the KDD process. Source: Fayyad et al. (1996)

important disciplines in which visualization is the foundation, for instance Visual Analytics / -Data Mining or exploratory Information Visualization.

The field of data visualization will be discussed in detail in the next chapter, since it is the focus of this thesis.

4.4 Other Mining Processes and Applications

4.4.1 Text mining

Text mining, or text analytics, is the process of deriving high-quality information from text. Possible use cases include grouping | summarization of text or documents based on similarity or categorization of text based on topics, text clustering and concept | entity extraction, etc.

The flow is similar to that of the KDD process. Text is inherently classified as unstructured data, so that the extraction of information — or information retrieval — must be carried out as the first step, after that advanced statistical methods or natural language processing will be applied to identify named entities. This *named entity recognition* step is to extract named text features such as people, organizations, place names, stock ticker symbols, certain abbreviations, and so on. The other steps in the process involve *recognition of pattern identified entities* like phone numbers, email addresses, etc.; *co-reference* — identification of phrases and terms referring to the same object; re*lationship, fact and event extraction*; *sentiment analysis* — extracting terms referring to emotion, mood, etc.; *quantitative text analysis*, etc.

4.4.2 Web mining

Web mining is a concept denoting the application of data mining techniques to discover patterns from data originated from the World Wide Web. Depending on what type of data is being mined, web mining is typically categorized into *web content mining* — extraction of high-quality information and knowledge from the content of the web, including but not limited to: text, video, images, URLs, etc.; *web structure mining* — using graph theory to analyze the nodes and structure of web sites; and *web usage mining*.

Web usage mining is the mining of web usage data, typically web server data (server logs), application server data, or application level data (events, etc.), to discover interesting usage patterns with the goal is to improve serving web applications.

4.5 Conclusions

This chapter gave an overview of how patterns, insights and knowledge can be extracted from huge amount of data through a seven-step process called Knowledge Discovery in Databases, in which data mining is a core step and forms the foundation of the KDD process. This is part of the greater effort to understand and interpret large quantities of data.

Data mining is a complex process and requires sufficient domain knowledge from the user in order to interpret the results. For the mining methods and algorithms to perform efficiently, the data must undergo a series of pre-processing steps, cleaning and transforming it when and where necessary. Visualization plays a notable and important role in the interpretation and validation of the results. It helps highlight possible relationships between the results and existing hypothesis, aids in the understanding and validation of the mined patterns for correctness and usefulness, as well as simplifying the knowledge transfer and presentation process.

Data mining is inherently an iterative process. The reason is because there is no guarantee that new or useful patterns can be found. In case no new or useful pattern is found, or the mined patterns are determined to be wrong in the validation process, the whole KDD process can be run again with adjusted parameters until there is at least a useful result.

Data mining can be applied on a variety of data types, and so become sub-concepts like Web Mining and Text Mining relevant depending on the data used. Typical goals and tasks include grouping of data records (*cluster analysis*), detecting unusual data records (*anomaly detection*), and extracting dependencies and relationships (*association rule mining*), among others. As such, it has application in many fields, from business and finance to industries and artificial intelligence (e.g., decision support system). Data mining also raises concerns about privacy, ethics, and copyright, but those are beyond the scope of this thesis.

In the next chapter, data visualization will be discussed in detail to highlight the important role it plays in various fields and processes, including data mining and knowledge discovery.

5 Data | Information Visualization

5.1 Introduction

Today we all have to more or less work with data as part of our daily work and activities. Data comes from various sources, in various forms, and almost two-third of them are in the form of electronic information (digital). The main contributing sources for digital data include social networks and media — which have recently seen a huge surge in popularity and usage — and open government data, which is discussed in Chapter 2, among others. Already in the year 2012, it was estimated by IBM that 2.5 exabytes of new data were generated per daily basis (IBM, 2014). This phenomenon helps coin the term *information pollution | explosion* in some research papers.

To human, a visual representation is often more effective than written text. It helps illustrate concepts and ideas, that, if expressed verbally, would be very difficult or even impossible to be fully understood. Data visualization can be thought of as a modern equivalent of visual communication. It revolves mainly around the creation and study of the visual representation of data — defined by Friendly (2009a) as "information that has been abstracted in some schematic form, including attributes or variables for the units of information". One of the primary goals of data visualization is to communicate information clearly and efficiently using various visual forms such as statistical | information graphics, plots, charts, etc., with data points being encoded by different visual elements like dots, lines, bars, size, colors, etc. A good and carefully crafted visualization helps users analyze and reason about complex data by making it more accessible, understandable, and usable.

In this chapter, data visualization will be discussed in detail. Starting in Section 5.2 is a brief summary of the long history of human visual representation, which dated back from as early as 200 B.C., based on the work of Friendly (2006). This is followed by the basics of data visualization such as key principles, terms, and definitions. Section 5.3 continues with a discussion on some most common visual representations typically found in charts, graphs, etc., then dives deeper into this topic in Section 5.4 by introducing a few visualization techniques designed specifically to visualize data sets with higher number of variables (more than two). This chapter concludes with Section 5.5 touching the topic of *visual data mining,* which is data mining with an emphasis on visualization, through an examination of a visual mining tool called V-Miner.

5.2 Data Visualization Fundamentals

5.2.1 A brief history of visualization

5.2.1.1 Prior to 17th century

The earliest seeds of visualization have long been existing in the form of geometric diagrams, layouts of stars and celestial bodies, and maps. Ancient Egyptians pioneered the use of positional units similar to latitude and longitude in their map making attempts as early as 200 B.C. Tufte (1983) reproduced a 10th century multiple time-series graph depicting the position changes of seven prominent heavenly bodies over space and time. As described by Funkhouser (1936) and shown in Figure 3, the y-axis represents the inclination of the planetary orbits and the x-axis shows time as thirty intervals. This graph is considered one of the earliest representations of quantitative information. (Oresme 1482; 1968) suggested in his 14th century work the concept of plotting a theoretical function and the logical relation between tabular and plotted values. The 16th century saw many important developments including triangulation and methods to determine accurately mapping locations, along with ideas for capturing image and the first modern cartographic atlas.

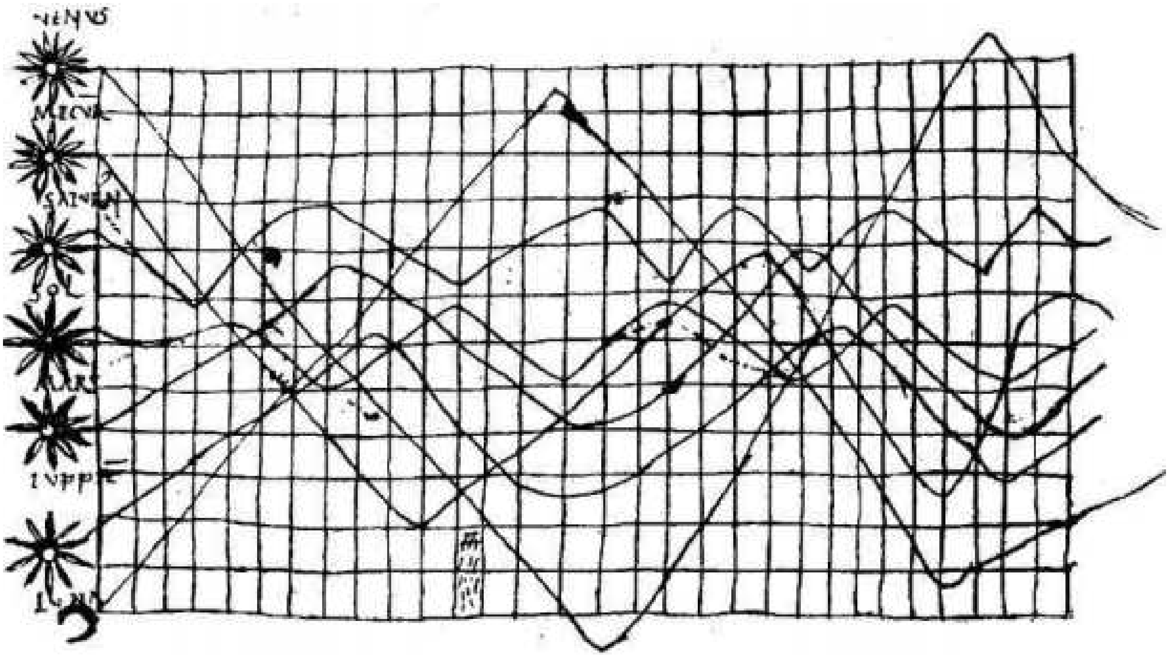


Figure 3: Planetary movements shown as cyclic inclination over time. Source: Funkhouser (1936, p. 261)

5.2.1.2 During 17th century

New theories and practical application rose sharply including analytic geometry and coordinate systems, probability theory, and demographic statistics, among others. Tufte (1983) coined the principle of "*small multiples*" based on an idea introduced in a visualization by Scheiner around 1630 — aimed to show the changing configurations of sunspots over time — shown in Figure 4. Also claimed by Tufte to be the first visual representation of statistical data was a 1644 graphic by astronomer M. F. van Langren, which shows estimated differences in longitude between Toledo and Rome. C. Huygens made the first graph of a continuous distribution function in 1669, and by the mid 1680s the first bi-variate plot derived from empirical data was already there. And so by the end of this century the foundations of visualization and visual thinking had already been laid: development of graphical methods, collected real data, theories and hypothesis, and concepts for visual representation.

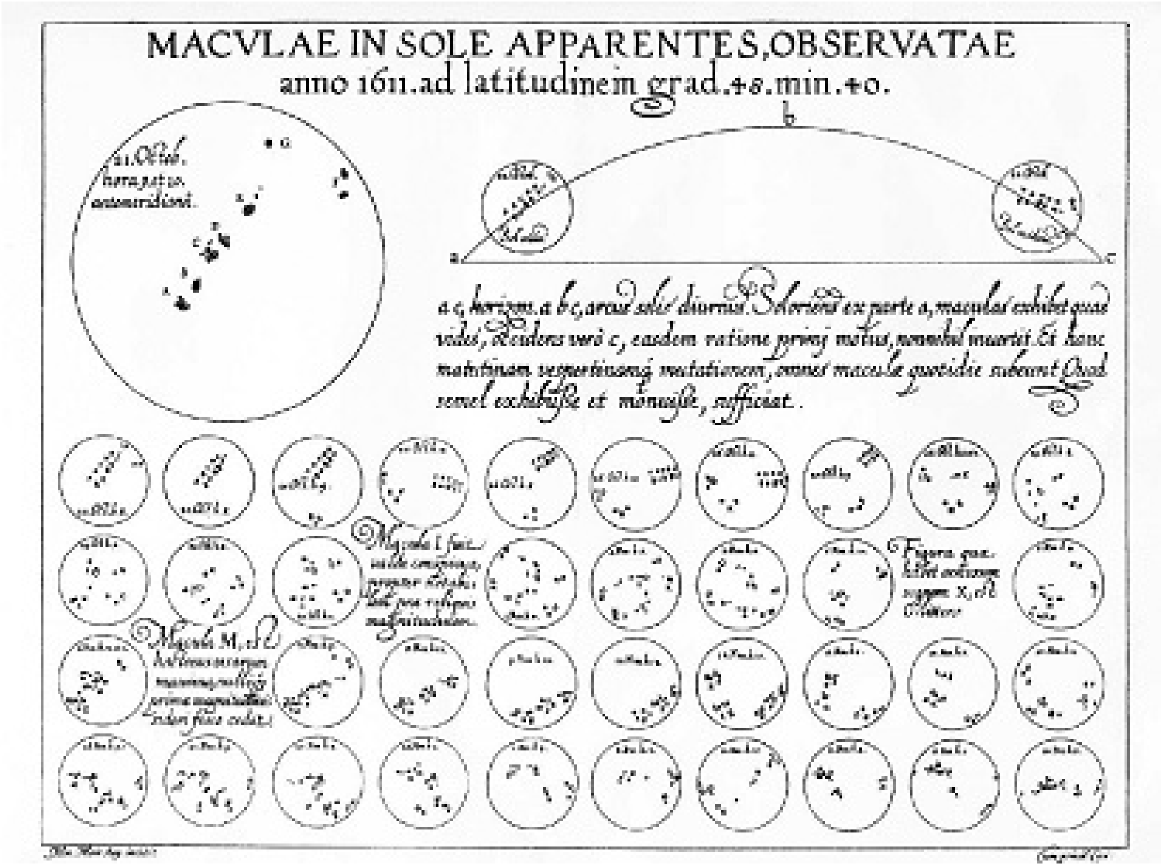

MACVLAE IN SOLE APPARENTES, OBSERVATAE
anno 1611. ad latitudinem grad. 48. min. 40.
a
b
c

Figure 4: Visualization of Scheiner's 1626 recordings of the changes in sunspots over time. Source: Scheiner (1630)

5.2.1.3 During 18th century

The 18th century saw further developments in cartography (isolines and contours, thematic mapping of physical quantities), abstract and functions graphs. Empirical data was also being collected more widely and systematically, including economic and political data, and with it rose the need for more novel visual forms to represent the data. Contour maps and topographic maps were also introduced by Buache (1752) and du Carla-Boniface (1782).

J. Barbeu-Dubourg first implemented the concept of *timelines* (cartes chronologiques) in an annotated chart of all history on a 16.5 meter scroll (Ferguson, 1991). In 1786 W. Playfair invented the first line graph and bar chart (Playfair, 1786) — one of the most widely used visual forms today — then later the pie chart and circle graph in (Playfair, 1801). An example is shown in Figure 5, which was a combination of various visual elements including circles, pies, and lines. Also shown in Figure 6 was an important milestone in visualization, represented by Playfair's time-series graph, depicting the price of wheat, weekly wages, and reigning monarch using three timelines of over 250 year time span. By the end of this century, although graphing in scientific applications had seen much utility, the practice remained uncommon until some thirty years later, partly due to the lack of data.

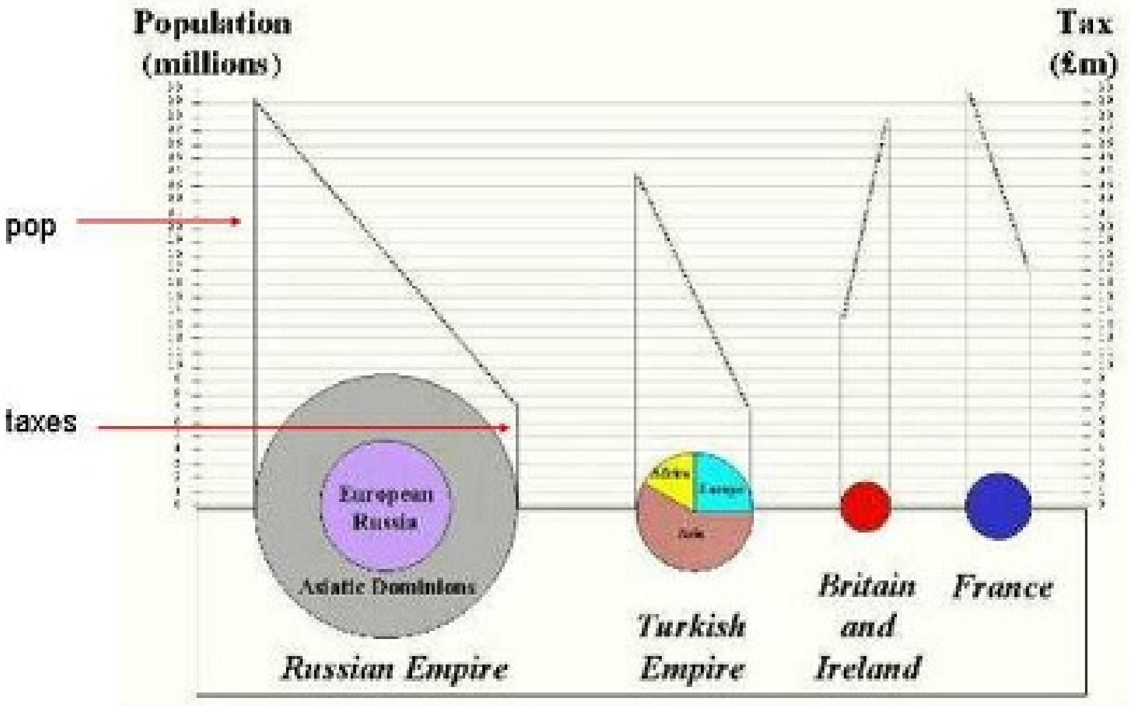


Figure 5: A redrawn of Playfair's 1801 pie-circle-line chart, comparing population and taxes in different nations. Source: Friendly (2006)

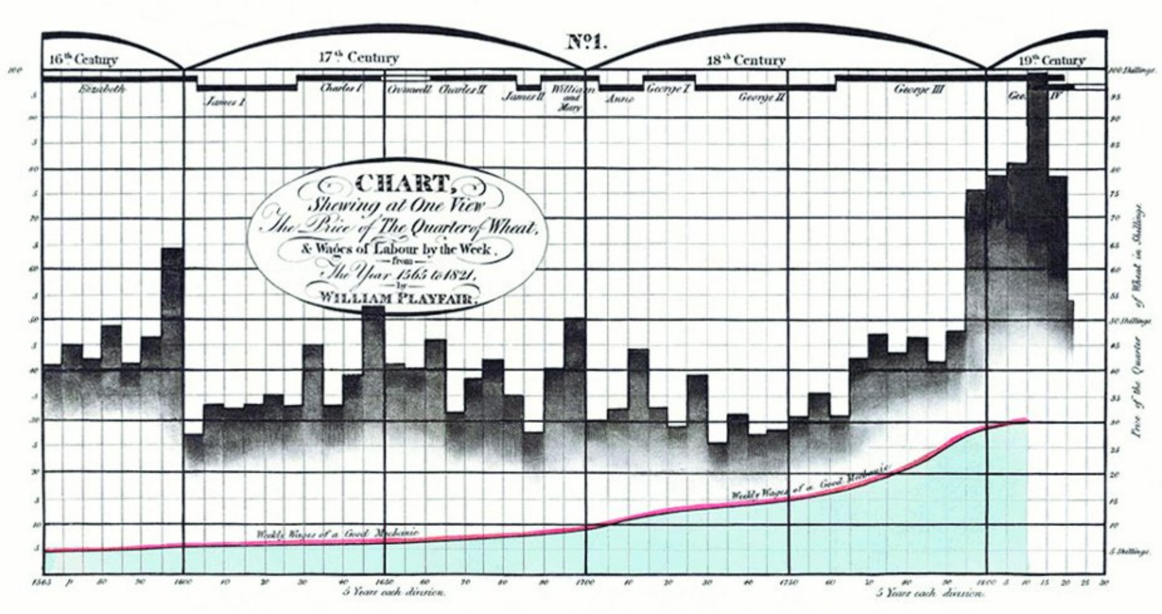


Figure 6: Playfair's time series graph of wheat prices, wages and ruling monarch, first published in Playfair (1821). Source: Tufte (1983, p. 34)

5.2.1.4 During 19th century

The first half of the 19th century saw the booming of statistical graphics and thematic mapping, with all of the modern visual forms had already been invented, including bar- and pie charts, histograms, scatterplots, and time-series plots among others. W. Smith pioneered geological cartography by introducing the first geological map in 1801 (Smith, 1815), then C. Dupin invented the use of continuous shadings (white to black) to depict distribution and degree of France's illiteracy in the 1820s (Dupin, 1826). A significant development in data collecting activities was in 1825 when France instituted the first centralized, nation-wide system for crime reporting. The year 1831 saw the first outbreak of Asiatic cholera in Great Britain with over 52,000 fatal cases, followed by subsequent outbreaks of 1848-1849 and 1853-1854. The cause was discovered and the location narrowed down in 1855 by Dr. John Snow with his famed dot map[18] (Snow, 1855), shown in Figure 7. The first cholera disease map, however, is attributed to Dr. R. Baker in 1833 (Baker, 1833) showing the severe outbreak of 1832 in Leeds (Great Britain), although it did not result in an impressive discovery like that of Dr. Snow. Other noted graphical inventions were made by C. J. Minard around 1830-1850, with an example being an early progenitor of the mosaic plot, as described in (Friendly, 1994).

[18] http://www.math.yorku.ca/SCS/Gallery/images/snow4.jpg

The second half of the century saw rapid growth of visualization with greatly improved aesthetics and innovations in graphics and thematic cartography, and is referred to as the Golden Age of statistical graphics. Attempts were made to break through the boundary of the flatland (2D world). Notable of those are Zeugner (1869) of Germany and later Perozzo (1880) of Italy with the construction of 3D surface plots of population data[19]. In 1861, Minard developed the use of divided circle diagrams on maps and later the flow map, which uses flow lines on maps with their widths proportional to quantity variables. A popular example is Minard's graphic depicting the destruction of Napoleon's army, described in (Tufte, 2001) as the "best graphic ever produced", which was able to encode six data variables in two-dimensional format including the number of troops, distance, temperature, latitude | longitude, direction of travel, and location relative to specific dates — shown in Figure 8. Another notable form of graphic called coxcombs (polar area charts) was invented by F. Nightingale to show the causes of mortality during the Crimean war.

Although there were much innovation and development during this so-called Golden Age of Visualization, the use of graphical representations had not really taken off due to the high cost of production of such graphics.

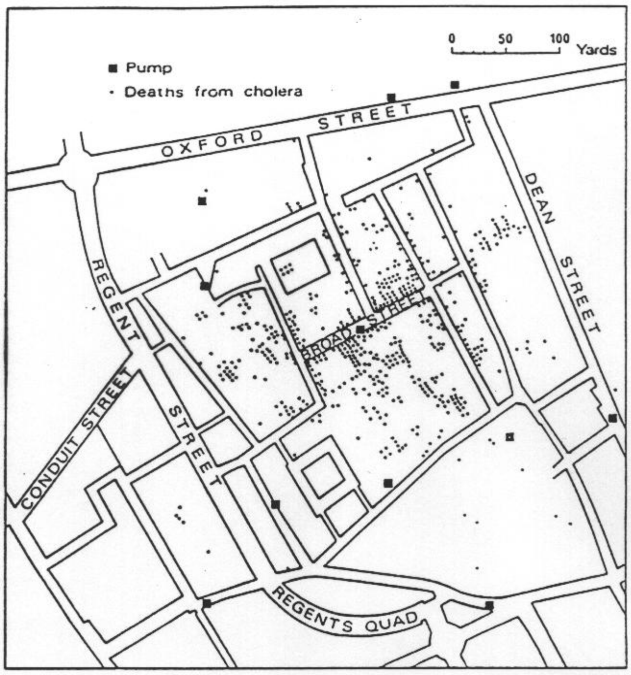


Figure 7: John Snow's 1854 dot map showing cholera deaths of Soho. Source: Snow (1855)

[19] Image: http://math.yorku.ca/SCS/Gallery/images/stereo2.jpg

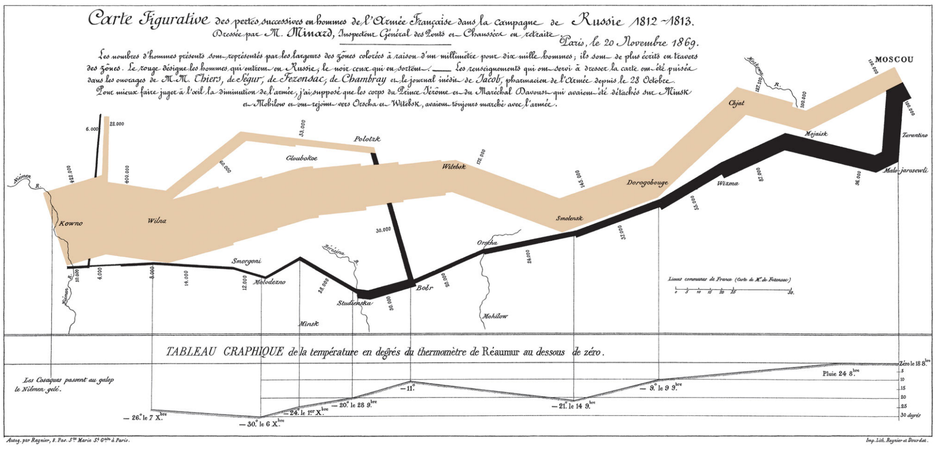

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Niémen R.
Kowno
Wilna
Polotzk
Gloubokoe
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Wizma
Orscha
Mohilow
Bobr
Studienska
Minsk
Molodezno
Smorgoni
Bérézina R.
Moskowa R.
422.000
400.000
6.000
22.000
60.000
33.000
175.000
145.000
127.100
100.000
100.000
96.000
87.000
55.000
37.000
24.000
20.000
50.000
30.000
28.000
12.000
14.000
8.000
4.000
10.000
Lieues communes de France (Carte de M. de Fezensac)
0 5 10 15 20 25 50.
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24. le 1er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

Figure 8: Minard's map of Napolelon campaign of 1812. Source: Wikipedia

5.2.1.5 During 20th century

During the first half of the 20th century, the enthusiasm for visualization, which was seen in the late 19th century, had been mostly supplanted by the rise of quantification and formal, statistical models. Statistical graphics became mainstream and had entered textbooks, was standard use in science, government, and commerce (Friendly, 2006). Experiments were carried out to compare the efficacy of various graphics forms (Eells, 1926; von Huhn, 1927; Washburne, 1927), as well as standards and rules for graphic presentation were being established. Also, new ideas and methods for multi-dimensional data came out to boost the process of going beyond 2D plane. The next generation of data visualization was around the corner with the development of the machinery of modern statistical methodology and the advent of computational power and display devices.

Data visualization was on the rise again in the mid 1960s with three important developments:

- J. Tukey invented new, simple and effective visual representations under the rubric of *Exploratory Data Analysis*, including stem-leaf plots, boxplots, two-way table displays, among others, and published in (Tukey, 1977).
- The book "Semiologie Graphique", published by Bertin (1967), contains the organization of the visual and perceptual elements of graphics according to features and relations in data. Around the same time, J. P. Benzecri introduced an exploratory and graphical approach to multi-dimensional data, which offered an alternative and visually-based view of statistics.
- With the introduction of *FORTRAN* — the first high-level programming language — in 1957, computer processing of statistical data and graphics construction had quickly taken over hand-drawn methods by 1960s. Interactive applications and high-resolution

graphics were also developed but were not in common use until later.

By the 1970s, interactive systems for 2D and 3D graphics were already in existence. During the last quarter of the 20th century, data visualization had already become a mature, vibrant, and multi-disciplinary research field. New technologies, computing power and software tools also kept a good pace in development. A few noteworthy themes include the development of highly interactive computing system for statistics, new methods and techniques for interactive visual data analysis (e.g., linking and brushing (Becker and Cleveland, 1987), selection, etc.), new techniques for multi-dimensional data visualization (e.g., the grand tour (Asimov, 1985), scatterplot matrix (Tukey and Tukey, 1981), parallel coordinates (Wegman, 1990), etc.), and increased focus on cognitive and perceptual aspects of data display, among others.

The above advances in visualization were possible thanks to the equally-rapid advancements in technologies. Some of these include large-scale statistical and graphics software packages (e.g., SAS, R, Lisp-Stat, etc.), extensions of classical statistical modeling to wider domains and huge leaps in computing power, data storage technologies, and network speed. Data dimensionality still remained a main concern, although generalizations of projecting a high-dimensional data set to lower dimensional views based on older reduction techniques had been introduced. By the mid 1980s, techniques for visualizing high-dimensional data were already developed such as mosaic plot (Hartigan and Kleiner, 1981), fourfold display (Fienberg, 1975), etc., of which the mosaic plot is the most useful and widely implemented, according to (Friendly 1994; 1999).

The true potential for handling multi-dimensional data, however, was proven to be a combination of interactivity and dynamic graphic display, allowing for direct manipulation of graphical objects and associated properties with a notable example being *PRIM-9* — the first such system for projecting, 3D rotating, isolating, and masking data of up to nine dimensions and developed by Fishkeller et al. (1974). From 1990 onward, there were already systems that unified some of the above ideas and concepts to form dynamic and interactive graphical systems that provided data analysis and manipulation in coherent and extensible computing environments (e.g., Lisp-Stat, ViSta). Young et al. (2006) presented in their book a detail description of the ideas and concepts behind today modern interactive graphics.

5.2.2 General visualization design principles

As with all types of designs, the most basic questions to ask when designing visualizations are the *target* (who), the *purpose* (why) and the *method* (how). A good design will convey information fully, while with a bad design, in which essential details are left out or distorted, the result will be catastrophic. The following are a few main principles:

- *Show the data*: focus on drawing the viewer's attention to the sense and substance of data (using data graphics), and not to something else, as advised by Tufte (2001).

- *Simplify*: choose the most efficient visual representation to communicate the data and keep the graphic simple. The design is considered finished when no more details can be taken away without losing information. For small data sets, tables and dot plots are preferred over pop charts (e.g., bar charts, etc.) because of more conveyed information, according to Cleveland (1994).
- *Reduce clutter*: remove any unnecessary, redundant details (e.g., tick marks, grid lines, etc.) or decorations on the graphic.
- *Revise*: just as the hard work of writing is rewriting, a good visual representation must constantly be revised for refinement and improvement.
- *Be truthful*: Tufte (2001) emphasized that in order for the graphic to tell the truth, the visual representation of the data must be consistent with the numerical representation as there are quite a few aspects that can distort the data, such as the aspect ratio or the scale.

Besides visual hierarchy and visual flow, the main aspect to consider when designing visual layout is the grouping of elements. As stated by the Gestalt principles, perception is influenced not only by the elements, but also by context (Scholarpedia, 2016). Based on this theory, four Gestalt principles which are important in creating a focal point without cluttering the visual representation are *proximity*, *similarity*, *continuity,* and *closure*, as depicted in Figure 9.

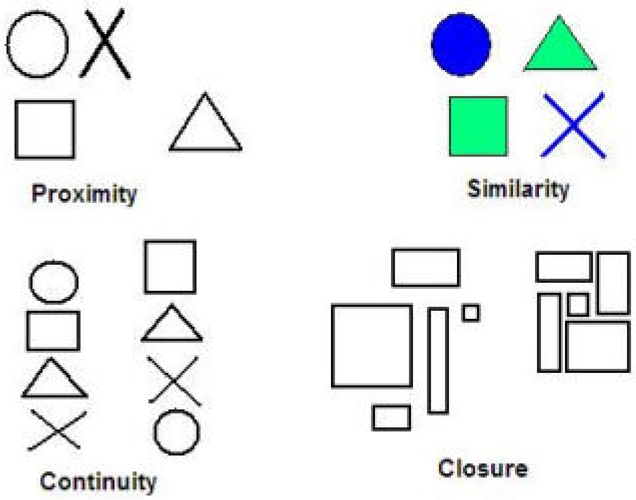
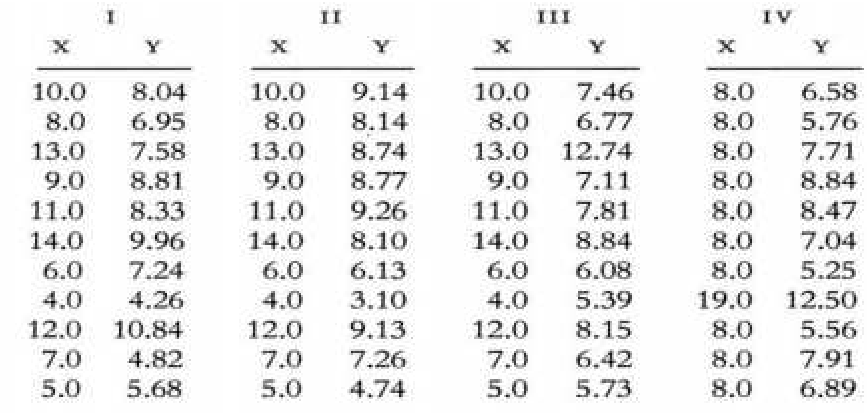


Figure 9: Four important Gestalt principles.

In addition to these principles, certain visual features such as color, size, shape, orientation, etc., which are called pre-attentive attributes by psychologists based on the fact that they can be processed in the brain concurrently and almost instantly with little mental effort, provide more options for encoding data easily. According to Ware (2012a), pre-attentive attributes can be categorized into *color*, *form*, *movement,* and *spatial position*. A classic example that illustrates the concept of pre-attentive attributes is described in (Tufte, 2001). Imagine one needs to find the relationships between the numbers in Figure 10. This task appears to be quite difficult because it requires much conscious effort. But if we have a proper visual representation, in this case the scatterplots shown in Figure 11, we can easily point out the extrema, groupings, trends, gaps, or outliers (outstanding values) in the numbers.



I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Figure 10: Attentive processing of numbers is slow and requires much conscious effort. Source: Tufte (2001)

5.2.3 Key terms and taxonomy of visualization

Visual forms help us understand the data and thus deliver information and knowledge. Robert Spence defined the process of visualizing data, in the context of a person who observes a visual representation of content, as a cognitive activity with which people build internal representations of the environment around them (Spence, 2001). These representations are defined as mental models of data, from which human is able to expand on and understand the data. It is considered an abstract concept. Although visualization process might be augmented with the use of visualization tools, which have been growing in capability and complexity in recent years, it still remains a cognitive activity (Ware, 2012b; Spence, 2001).

Spence (2001) also differentiates three common uses for the term visualization, depending on what type of data is being visualized. For example, if the data has a correspondence in physical space, it is considered as *scientific visualization*. In case of abstract data that does not necessarily have a spatial dimension, we speak of information visualization. Data visualization is a broader term which encompasses both information visualization and geographic visualization, which places visual representations of information on a map.

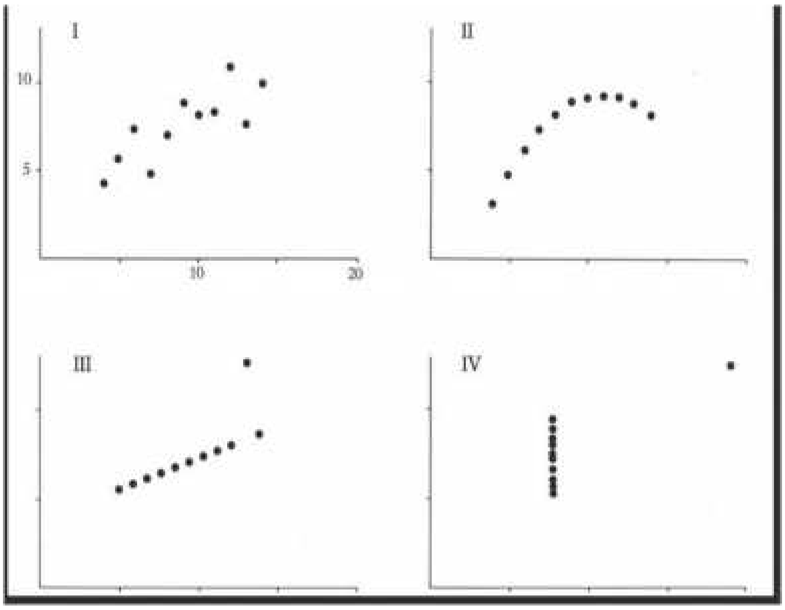


Figure 11: Proper visual representation speeds up the cognitive process greatly. Source: Tufte (2001)

5.2.4 Confirmatory and exploratory visualization

5.2.4.1 Confirmatory visualization

Visualization might be used to carry out confirmatory data analysis. By representing the structural relationship between series of data visually, the graphic can be used to confirm hypothesis on the data (statistical hypothesis testing). Figure 12 illustrates this concept by comparing the values of the German stock market index (DAX) to those of American Dow Jones over the course of a year. The visualization makes it clear that the two indices correlate to each other following a rather similar trend in the rising and falling phases, which can be formulated using complex mathematical formulas. But with the latter method, it would certainly be far less intuitive and expressive for the majority of audiences.

5.2.4.2 Exploratory visualization

...is the application that seeks to uncover hidden information, knowledge, trends, and patterns within the data set and can be seen as akin to visual data mining. Instead of testing hypothesis on the data, the data will be used to suggest hypothesis to test. Visual representations and human's ability of analyzing via visual perception using cognitive system are

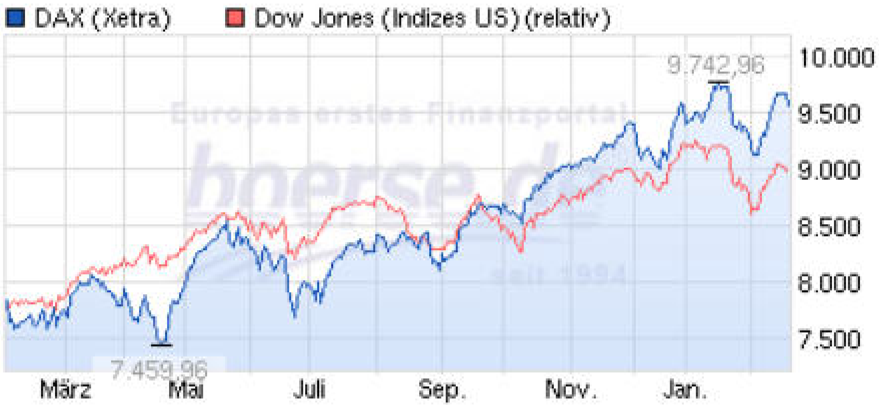


Figure 12: DAX (Deutscher Aktien IndeX) compared to U.S. Dow Jones. Source: Boerse.de

of great advantage to this application, with Bertin (1981) defined it as the "visual means to resolve logical problems".

As an example, Fig. 6.9 shows the distribution of breast cancer cases among women in Germany in 2009. With the color codes, we can recognize the geographical areas with fewer (light red), and with higher (dark red), number of cases. It is noticeable how the number of cases (with possible deaths) are found predominantly in the western states. Germany's Epidemiological Cancer Registry Association (GEKID) states that the number of cases in eastern states are 20-30 percent lower than in western states, and this phenomenon is in some way associated with the lifestyles of the female population (von Hoff, 2009). The visualization is not used to confirm any hypothesis and also does not provide an explanation but instead, the user is left to draw their own conclusion and the result might be used as a suggestion for researchers to carry out more extensive epidemiological studies in those areas.

Figure 14 shows a few typical attributes associated with each of the above two approaches to data visualization. Fields like data journalism can also benefit from exploratory visualization, which enables journalists to tell complex stories through engaging infographics while encouraging viewers to explore context information and supportive data.

5.3 Interactivity with Dynamic Techniques

Normal visual forms and visualization techniques will eventually face scalability problems when the data set grows to a certain volume and complexity. Facilitating interaction between the user, the data, and the visualization tool is one common approach to this challenge, which can help solving the scalability problem by allowing the user to modify the input data, change the visual mapping, or manipulate the generated views. This is valuable in exploratory visualization in that interesting relationships or patterns might be uncovered that will otherwise

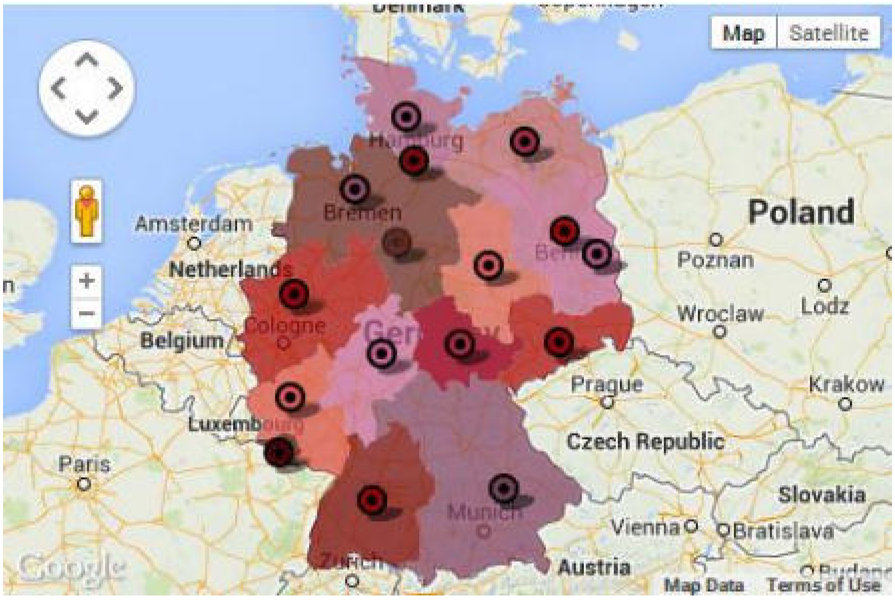


Figure 13: Distribution of cancer cases in the German female population in 2009 per state. Source: von Hoff (2009)

remain hidden in static view. Interactive visual representations may fall into one of the three categories, as defined in (Mazza, 2009):

- *Static representation*: no interaction is available, only a single and static view is generated. Figure 15 shows an example of such static representation in the form of a graphic that depicts the poem *Herr von Ribbeck auf Ribbeck im Havelland*. The graphic is part of a visualization project developed by Boris Müller[20] for Germany's Poetry on the Road annual international literature festival. The main idea is to assign a numerical value to every single letter of a word in the poem, adding them together to get a number that represents the word, then arrange all the words on a circular path with the diameter depends on the length of the poem. The aim of the project is to put hundreds of poems together to produce an artistic and aesthetic graphic, rather than to convey information. Thus, this static visualization is more of creative, artistic, and aesthetic value than statistical value.

- *Manipulable representation*: allows the user to interact with the generated view through basic interactions such as zooming, panning, rotating, etc. As an example, the Zeit Online magazine put together a visualization of the position data taken from the mobile phone of Malte Spitz[21] — a German politician — that allows the user to rapidly step through the timeline to see his movements and phone calls during this time window.

[20] http://www.esono.com/boris/projects/poetry06/

[21] http://www.zeit.de/datenschutz/maltespitz-data-retention

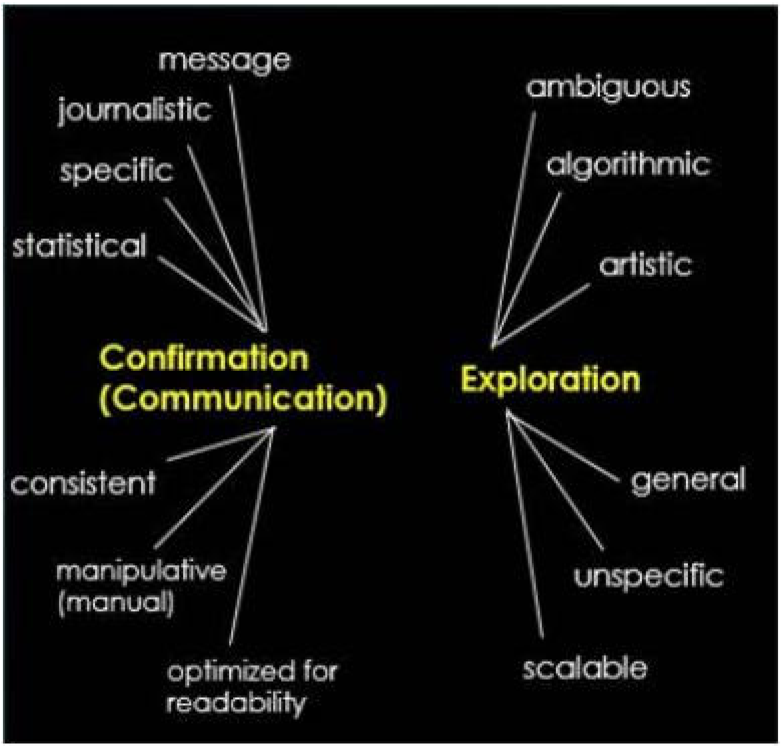


Figure 14: Confirmatory and exploratory visualization — typical attributes.

- *Transformable representation*: enables the user to affect the pre-processing phase, e.g., modify the input data through data filtering, with the result of changing or modifying the generated view. This is the highest form of interactive visual representation. An example is the visualization project *GED VIZ*[22], which allows the user to explore and compare different aspects between various nations. By choosing different filter criteria, the view will change to reflect the user's selection.

Interactivity, however, might add up to the complexity of the visualization and as a consequence, reduces its usability from the user's point of view. Thus, an important goal of interaction design is to reduce excessive interactivity (Myatt and Johnson, 2009). A few common methods for interactivity include:

- *Data brushing*: also known as slicing, is a technique that displays the same piece of data simultaneously in multiple different views. The selected data is displayed throughout all graphics that are linked to each other.
- *Nearness selection*: H. Wainer proposed this technique to highlight all data points within a specified distance starting from a specified start position (optional) (Wainer, 2005).

[22] http://viz.gedproject.de/

- *Sorting, rearranging, searching and filtering*: these techniques provide more interactivity to the user which would further promote data exploration. However, they must be designed carefully so as not to negatively affect the usability of the graphic.

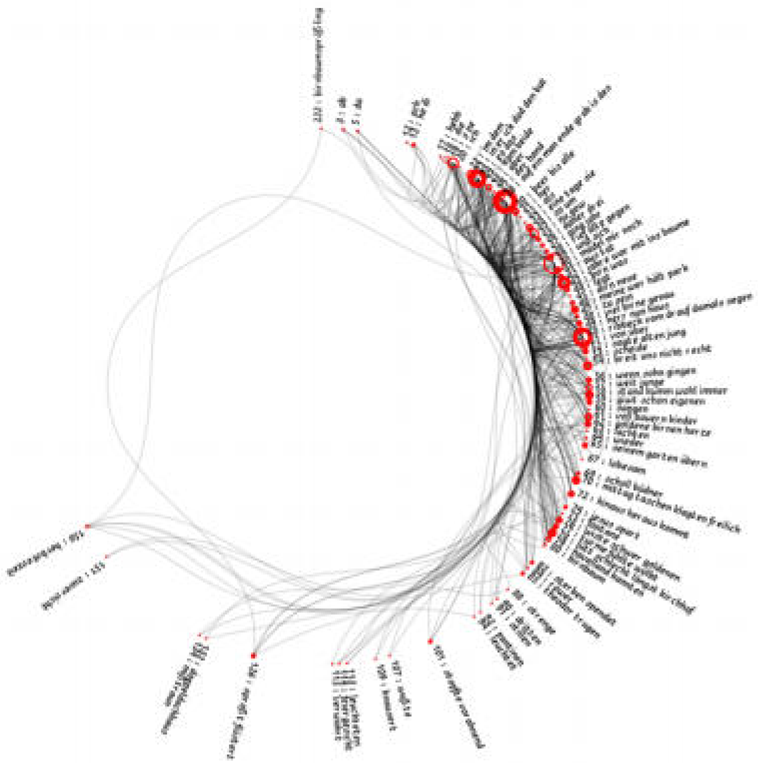


Figure 15: Code poetry by Mueller et al. Source: Boris Mueller's portfolio

5.4 Visualization of Multi-variate Data

5.4.1 Introduction

Multi-variate data are collections of data in which many attributes change with respect to one or more independent attributes. While bi-variate and tri-variate data can be well represented by a traditional scatterplot on Cartesian axes, which is a very simple and intuitive visual form, and works well when there are maximum two dependent attributes. However, the number of real-world situations in which only one or two dependent attributes are involved is very limited. In fact, most practical problems require analyzing a rather high number of dependent attributes (Mazza, 2009).

5.4.2 Common techniques

Scatterplots can be adapted to multi-variate problems to a certain extent by adding further visual elements, such as shape, dimension, color, etc., which allows for mapping of more than two attributes and are called *extended scatterplots*. As an example, the extended scatterplot shown in Figure 16 visualizes data from 174 nations to compare the level of wealth (GDP per capita, x-axis) and the state of health (life expectancy, y-axis). In addition, the graphic also maps the population of each nation (dimension of the visual element) and the continent to which each nation belongs (color of visual element). Thus this scatterplot has been extended to map four data attributes on the same graphic. Normally, extended scatterplots are effective for problems with up to seven variables and not all problems and data sets can be visualized with scatterplot.

Other common techniques for visualizing multi-variate data can be categorized into *geometric techniques*, *pixel-oriented techniques,* and. *iconic techniques*. Geometric techniques map data onto a geometric space. An example is the *Parallel Coordinates*. This technique was developed based on the idea by Inselberg (1981), which defined a geometric space through an arbitrary number of axes, arranged in parallel. Other common geometric techniques include *Scatterplot Matrix*, *TableLens,* and *Parallel Sets*.

Iconic techniques leverage geometric properties such as color, shape, size, etc., of a figure or glyph, by assigning each variable to a feature of the geometric glyph and mapping data to the properties of each feature. *Star Plots* and *Chernoff Faces* are two notable techniques that belong to this category.

The basic idea of pixel-oriented techniques is to use each pixel of the screen as the base atomic unit to represent an element of the data set and map the variable to the color of each pixel, thus maximizing the number of elements that can be represented. In addition, for multivariate data sets, data is grouped into specific areas of the screen according to their attributes called windows. Daniel Keim defines a number of factors to be considered when using pixel-oriented techniques, which include shape of the windows, ordering of the windows, visual mapping (what to map), color mapping (how to map), and arrangement of the pixels (Keim, 2000).

5.4.3 Group data

Certain grouping algorithms, such as *hierarchical agglomerative clustering*, when applied to a data set, will produce groups of observations called a hierarchy of clusters. To visualize this hierarchy as a tree, a *dendrogram* — as shown in Figure 17 — is commonly used. A dendrogram allows not only the groups to be visually represented, but also the similarity between clusters. There are commonly two ways of representing a dendrogram, with the one

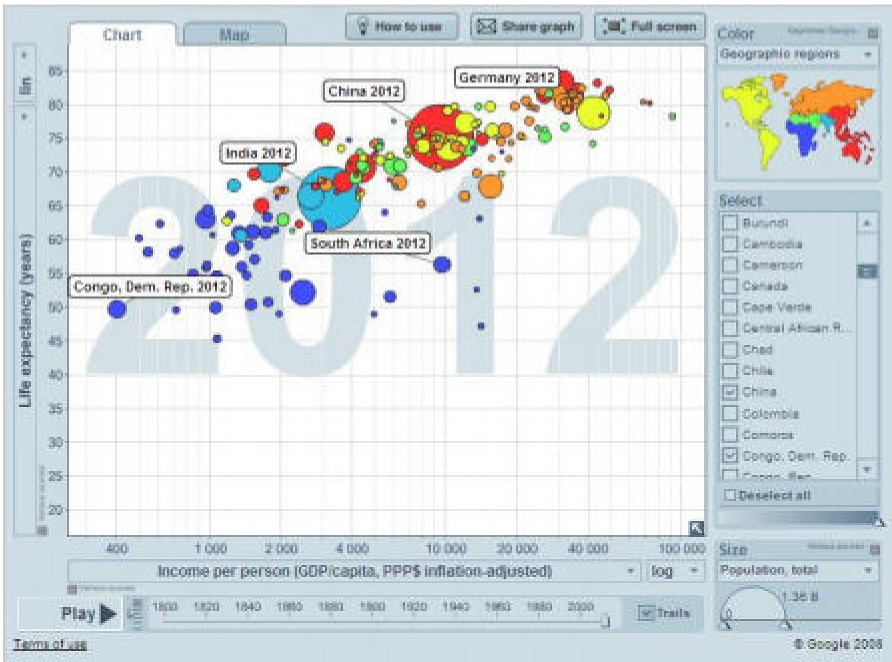


Figure 16: Extended scatterplot encoding four data variables. Source: Gapminder

on the right in Figure 17 often used in molecular biology. Other notable techniques for visualizing groups include *Decision Tree*, which is generated using a group of decisions based on certain attributes and *Cluster Image Map*, which is essentially a dendrogram combined with a heat map to display complex, high density data.

5.4.4 Network and hierarchical data

The basic idea for visualizing network data is to represent them as *graphs*, following a number of guidelines defined in (Shneiderman et al., 1999). Two common techniques to visualize simple network data are *Concept Maps* and *Mind Maps*, which are most widely used to describe ideas, situations or organizations in brainstorming sessions and educational environments. With complex network data, the main problem when using graph representation is scalability, thus extra optimization techniques such as link reduction or minimum spanning trees (MST) have to be applied in order to reduce the complexity of the graph. Other additional components such as geographic map can be used to visualize network topology, as appeared in (Dodge and Kitchin, 2002; He et al., 1996; Patterson and Cox, 1994). A classic type of map commonly used to represent transport networks is the tube map that is derived from the concept of H. Beck in 1931. Graphs can also be used to represent hierarchical data following the tree paradigm, which consists of a root, parent nodes, and child nodes. This is the basis for many hierarchical data representation techniques, such as the *File System*

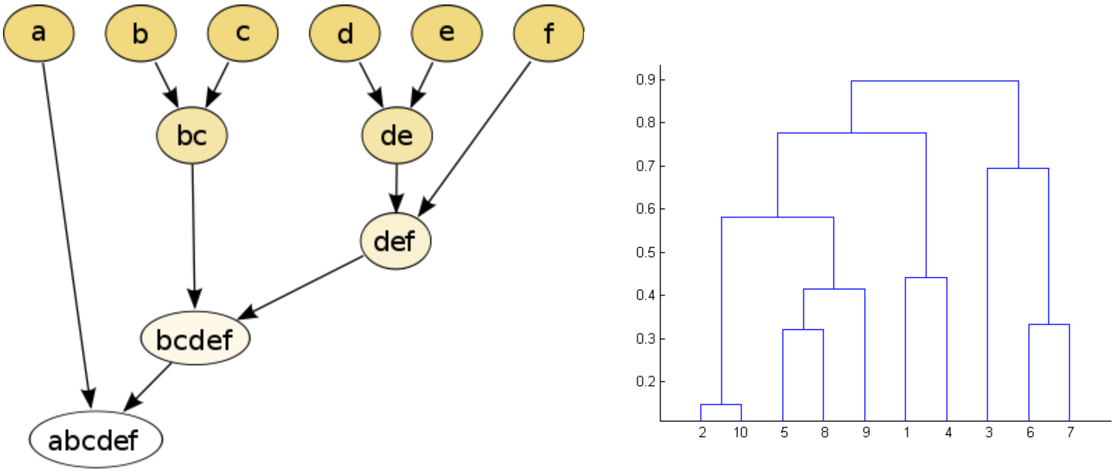


Figure 17: Dendrograms. Source: Wikipedia and MATLAB

(files in folders and sub-folders). In fact, tree is the dominant form of representing hierarchical data with many derivations developed to adapt to various visualization situations, such as the *Cone Tree* by Card et al. (1991), which is a three-dimensional visualization form for hierarchies with large number of nodes, or the *Botanical Tree* by von Wijk et al. (2001), which tries to solve the complexity problem of certain situations by imitating natural (botanical) trees, or the *Treemap*, a space-filling visualization algorithm originally developed by Shneiderman (1991) which aims to utilize all available space to represent hierarchical data using nested rectangles. Newsmap[23] by M. Weskamp is a well-known example that uses treemapping to display information from Google News — dimension of the rectangle signifies the importance of the news while color and color intensity show news categories and indicate the freshness of the news, respectively.

The Treemap algorithm along with Trellis Display for visualizing high-dimensional data and Linked Views for visual exploration will be discussed in detail below.

5.4.5 Tree-mapping: a space-filling method to visualize hierarchical information structures

5.4.5.1 Introduction

The treemap algorithm maps hierarchical information to a 2D rectangular display in a space-filling fashion. It utilizes 100% available space and provides interactive control to specify structural (e.g., depth bounds) and content (display properties e.g., color mappings) information. Display space is partitioned into rectangular bounding boxes and can be allocated

[23] http://marumushi.com/projects/newsmap

proportionally to the importance of the information. The drawing of nodes and display size can be interactively controlled by the user, thus allowing for hierarchical structures with large number of nodes to be displayed and manipulated even within a fixed display space.

5.4.5.2 The treemap method

The main goal is to display the tree structure in a visually appealing way while fully utilizing available space, which is quite difficult to achieve. Interactive elements enable users to control both structural and content representation, e.g., visual properties, to maximize the utility of the visualization.

By controlling the partitioning of the display space, structural information can be adjusted to best fit the task. The representation of treemap displays is similar to that of *quad-trees* and *k-D trees*, with the key difference lies in the direction of the transformation. Treemap represents hierarchical structures on 2D displays, as opposed to quad-trees, which creates hierarchical structures to store 2D images (Samet, 1989). A weight, which may be single or combined domain properties, e.g., disk usage, is assigned to each node to determine the size of the node's bounding box and can be viewed as a measure of importance (Furnas, 1986). The following relationships should always hold:

1. If *Node1* is an ancestor of *Node2*, then *Node1*'s bounding box is either equal to, or completely encloses *Node2*'s bounding box.
2. The bounding boxes of two nodes intersect if and only if one node is an ancestor of the other.
3. The size of a node's bounding box is strictly proportional to its weight.
4. The weight of a parent node is greater than or equal to the sum of the weights of its children.

Structural information is implicitly presented, but can also be explicitly indicated by nesting, which enables direct selection of all nodes (internal and leaf). The disadvantage thereof is the reduction of displayable tree size (Travers, 1986). While non-nested displays can only enable direct selection for leaf nodes, a pop-up display can provide further information and facilities. And although non-nested displays also cannot represent internal nodes in degenerate linear sub-paths, such paths rarely occur.

The process of mapping content information to the display can be manipulated through a variety of display properties, which determine how a node is drawn within its bounding box. Aside from primary visual properties such as color, shape, texture, etc., other domain dependent properties may also exist, which result in a rich set of mapping possibilities between content information and display properties (Ding and Mateti, 1990).

Interactive elements are essential and critical, since the number of variables that can be coded in the tree is limited and there is also an upper bound on the complexity of graphical representation for human perception . Dynamic feedback is available through the use of pop-ups which show information about the current node.

5.4.5.3 The treemap algorithm

The treemap method consists of two algorithms. The first one is the drawing algorithm (Figure 46), used to draw a series of nested boxes representing the tree structure. With this algorithm, a treemap can be drawn during one pre-order pass through the tree in *O(n)* time under the condition that node properties (e.g., weight, name) have already been pre-computed and assigned. The second algorithm is the tracking algorithm (Figure 47) which enables the path to a node containing a given point in a display to be determined using only a simple descent.

5.4.5.4 Conclusions

The space-filling approach to visualize hierarchical structures, that treemapping is a typical example, has great potentials for exploratory visualization of large data sets. First, it enables effective representation of large hierarchies in a limited space, which is often the challenge when visualizing large amount of data using today standard displays. It also has good prospects for future extensions, such as an alternate scheme for structural partitioning, for visual display of various content information — including numeric and non-numeric. Treemap visualization could be further enhanced with dynamic views, e.g., animated time slices and more advanced operations on elements of the hierarchy besides the standard ones, such as zooming, panning, selecting, searching, etc. Because large data sets often contain some sorts of hierarchy, this visualization technique can be applied in a wide variety of applications, e.g., in financial portfolios.

5.4.6 Trellis displays: an approach to high-dimensional data visualization

5.4.6.1 Introduction

Trellis display is introduced by Becker et al. (1996) as an approach to visualize multi-variate data. Unlike mosaicplots which employ a recursive layout, trellis displays use a *grid-like* (lattice) layout to arrange conditioned plots onto panels. The use of the same scales in all panel plots enables plot comparison across rows and columns.

A single trellis display can theoretically hold up to seven variables. Five of them need to be categorical, with up to three can be used as conditioning variables to form rows, columns, and pages. The other two can be continuous. Up to two variables—called axis variables—can be plotted in a panel plot. All the panel plots share the same scale. Rows, columns, and pages of the trellis display are created using up to three conditioning categorical variables.

Shingling is a concept which was introduced with trellis displays and is defined as the process of converting a continuous variable into a discrete one by splitting a continuous variable into (overlapping) intervals. Overlapping intervals may lead to various data representations in a trellis display. Figure 18 shows an example of a more complex trellis display — a cars data set being plotted using scatterplots of *Gas Mileage* (MPG) vs. *Weight* (axis variables). The grid is formed by the conditioning variables *Car Type* (x-axis) and *Drive* (y-axis). The adjunct variable *Number of Cylinders* is color-coded in the scatterplots.

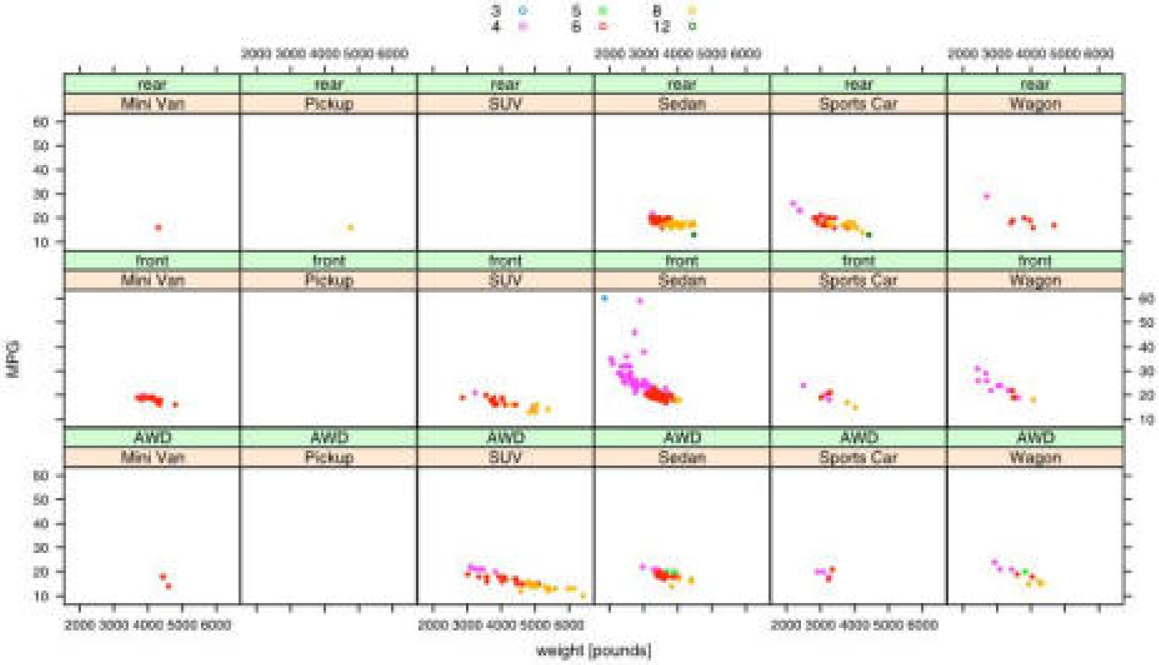


Figure 18: A trellis display with five variables. Source: Chen et al. (2008)

5.4.6.2 Trellis displays with interactive elements

Based on the fact that the conditional framework and a single view in a panel of a trellis display can be regarded as static snapshots of interactive graphics and highlighted part of the panel plot graphics for the conditioned subgroup, respectively, trellis displays can be extended with interactive elements. Figure 19 (left) shows some examples of possible interactions in an interactive session, including selecting a specific subgroup in the left mosaic plot or moving the brush (selection indicator) along one or two axes of the plot. This is where the concept of shingle variables can be useful, as the selection from the brush can be regarded as an interval of a shingle variable. The shingling process divides a continuous variable into intervals, which correspond to the snapshots of the continuous brushing process, as illustrated in Figure 19 (right).

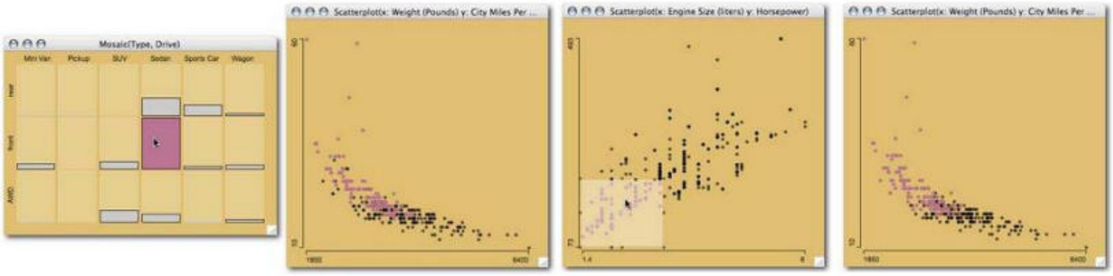

Mosaic(Type, Drive)
Mini Van
Pickup
SUV
Sedan
Sports Car
Wagon
Rear
Front
AWD
Scatterplot(x: Weight (Pounds) y: City Miles Per ...
1850
6400
Scatterplot(x: Engine Size (liters) y: Horsepower)
1.4
8
Scatterplot(x: Weight (Pounds) y: City Miles Per ...
1850
6400

Figure 19: Selecting / brushing a subgroup / scatterplot on the left-side view highlights the subgroup / values in the panel plot on the right. Source: Chen et al. (2008)

5.4.6.3 Conclusions

Trellis displays are most suitable for continuous axis variables, categorical conditioning | adjunct variables and are intended to be an approach to visualize multi-variate data sets. Despite the advantage of a flat learning curve and the ability to add model information to the plots, current trellis display implementations do not offer any interactions. However with the use of interactive linked graphics, e.g., linking panel plots to bar charts | mosaicplots of the conditioning variables, brushing over shingle variables, etc., trellis displays can offer to some degree exploratory data analysis.

5.4.7 Linked views for visual data exploration

5.4.7.1 Introduction

One of the most common challenges in visualization is the physical dimensional limitation of the presentation device, whether it be paper or computer screen, visualization is limited in a 2D space, or *flatland*. To address this problem, there are commonly four approaches:

- Use of a virtual reality | pseudo 3D environment in a 3D setting to portray higher dimensional data, which is the main focus of this thesis.
- Projection of high-dimensional data onto 2D coordinate system using data reduction methods.

- Use of a non-orthogonal coordinate system, such as parallel coordinate.
- Linking of multiple low dimensional displays, which is the basic idea behind Linked Views.

This idea is not new, with the use of identical plot symbols and colors to indicate same cases across multiple displays in the development of static displays as mentioned in (Tufte, 2001) and (Diaconis and Friedman, 1983) and first implemented in (McDonald, 1982) to connect two scatterplots. The most widely used implementation of linked views is the *scatterplot brushing*, including linking in both scatterplots and scatterplot matrices, as promoted in (Becker et al., 1987).

The main benefits of using linked views with regard to exploratory data analysis are the simplicity of underlying graphical displays, speed and flexibility in portraying various data aspects. Another advantage of linked views is the applicability to complex data structures, such as geographically referenced data in the context of spatial data exploration, as discussed in (Anselin, 1999; Wills, 1992; Roberts, 2004). Linked views is mainly applied in statistical exploration of data sets, to address issues such as finding unusual behaviors, detecting relationships, patterns, etc.

5.4.7.2 Linking schemes and structures

The principal behind linked views is the sharing and exchanging information between plots. To achieve this, first, a linking mechanism is needed to establish a relationship between the plots, then two questions need to be answered: what information is shared and how?. A separation of data displays in their components, as proposed by Wilhelm (2005) set the foundation to create a wide variety of linking schemes and structures.

According to Wilhelm (2005), a display D is made of a frame F, a type with a set of graphical elements G and a set of scale S$_G$, a model X with scale S$_X$, and a sample population Ω. Thus, the data part is the pair $((X, S_X), \Omega)$ and the pair $(F, (G, S_G))$ is the plotting part. According to the above definition, it is theoretically possible to define a linking structure as a set of relations among any two components of the displays. In practice, however, only the relations between identical layers of the display are of relevance. Thus, possible linking schemes between active display D$_1$ and passive display D$_2$ are as depicted in Figure 20.

From the separation of data display in components, linking schemes are separated into four types:

- *Linking sample populations*: defined as a mapping $m : \Omega_1 \to \Omega_2$ in which elements of sample population space Ω_1 are mapped to some elements of space Ω_2. There are three common types of sample population linking: *identity linking* (empirical linking : $id : \Omega \to \Omega$), *hierarchical linking* ($m : \Omega_1 \to \Omega_2$ with filtration), and *neighborhood /*

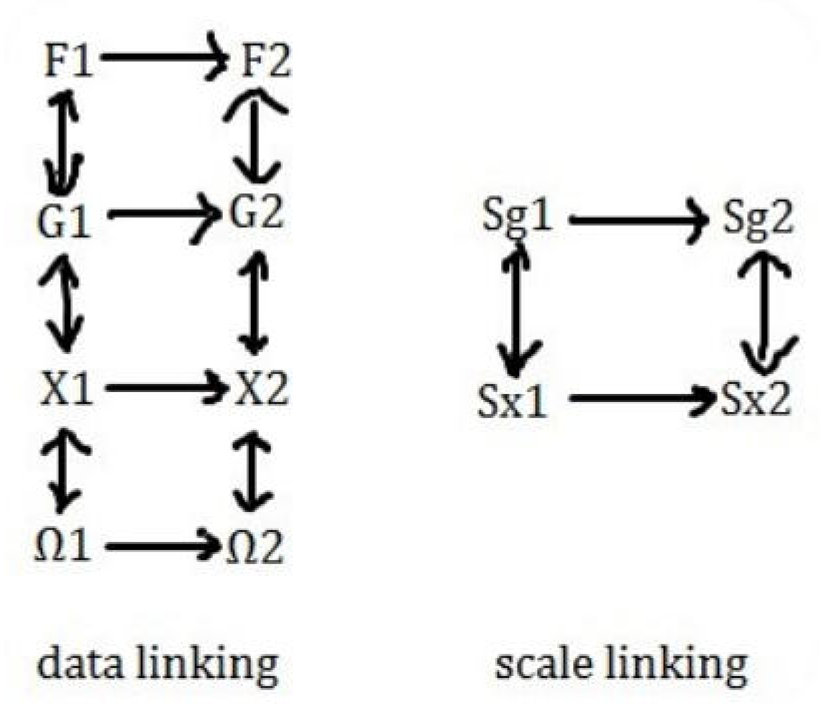


Figure 20: Possible linking schemes between sender plot D_1 and receiver plot D_2

distance linking (for geographical data, linking relation depends on definition of neighborhood or distance).

- *Linking models*: models describe precisely the amount of information to be visualized. For example, the histogram of a quantitative variable is based on the categorization model. Linking models can be further categorized into *type linking* and *scale linking*, with scale linking being the more common type and most widely implemented in the form of sliders for dynamic queries, which was discussed in details in (Shneiderman, 1994) and (Shneiderman, 1997). Linking observations is restricted to the variables used in the model, as illustrated in Figure 21. Young et al. gave a fairly thorough introduction and proposals for linking observations in (Young et al., 1993).
- *Linking types*: the type layer covers most visible components in a graphical display and aims to represent the model as well as possible. Due to this close connection, congruities at the type level are typically the result of linked models. Direct link between type levels without model linking is uncommon, except for color and size, which are attributes that can be linked regardless of model linking. It is often required to link type information to properly compare between various plots.
- *Linking frames*: frames control the shape and size of the plot window. Linking frames is important for the accurate comparison of graphical displays and to achieve a screen

space-saving layout.

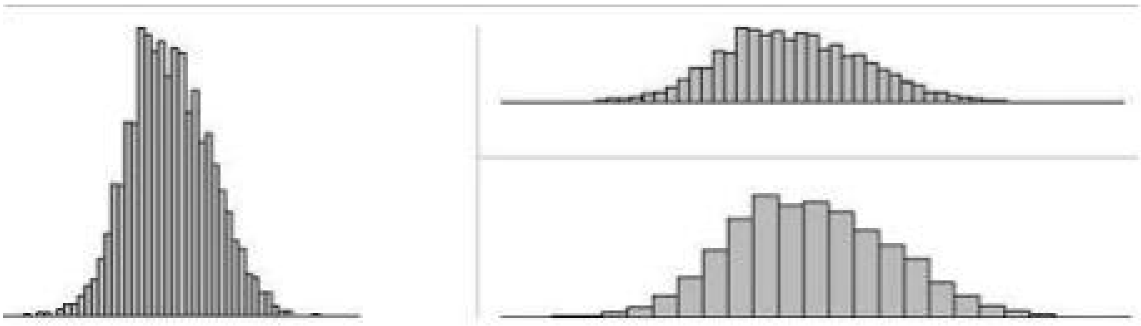


Figure 21: Three histograms of the same variable. The two plots on the right side have same frame size but different scales. The top-right plot has the same scale as the left plot. Source: Chen et al. (2008)

5.4.7.3 Implementation strategies for linked views

Information sharing is the back-end mechanism that drives the linked views paradigm. Information sharing occurs in various circumstances, such as in an interactive session with the user making changes to a plot while exploring and investigating the data. This scenario raises the question of where the information should go and how it can be optimal represented. According to Robert et al. (2000), three different strategies for implementing linked views can be distinguished:

- *Replacement strategy*: while this strategy can be applicable for plot parameters, it proved to be not useful for subsetting and conditioning approach because of the replacement of old information, except for the case in which each observation has its individual plot symbols. Even then, the inability to compare different versions of the plot makes this strategy inappropriate for exploratory data analysis, where it is essential to keep track of changing scenarios and different plot versions. As discussed in (Roberts, 2004), implementing a history system similar to those in geo-visualization systems to help keep track of plot changes is very helpful.
- *Overlaying*: while this strategy is typical for comparing two conditional distributions, it creates two problems: one is the basic restriction in the freedom of parameter choice for the selected subset, because the parameters are inherited from the original plot, the other is the problem of overplotting | occlusion, in which part of the original display is hidden by the overlaid plot. This problem is mostly irrelevant for area-based displays and scatterplots but plays an important roles in complex plots such as boxplots. Figure 22 shows the overlaying strategy in a scenario of a histogram (left) being linked to a

bar chart (right). The selection in the bar chart is propagated to the histogram and overlaid on the original plot. Plot parameters are inherited.

- *Repetition*: the third strategy is to multiply the displays, with each display represents a different view of the data and all are presented to the user at the same time. The advantage is that the user gets a comprehensive picture of the data, a fairly complete overview which enables easy observation of the impact of parameter changes or user interactions. The downside is that the overview might become complex by various changing and adapted views, therefore a mechanism to keep track of various changes and user interactions as well as an effective system to arrange the displays on computer screen are needed. Juxtaposition is an example form of the repetition strategy that works very effectively for subsetting scenarios.

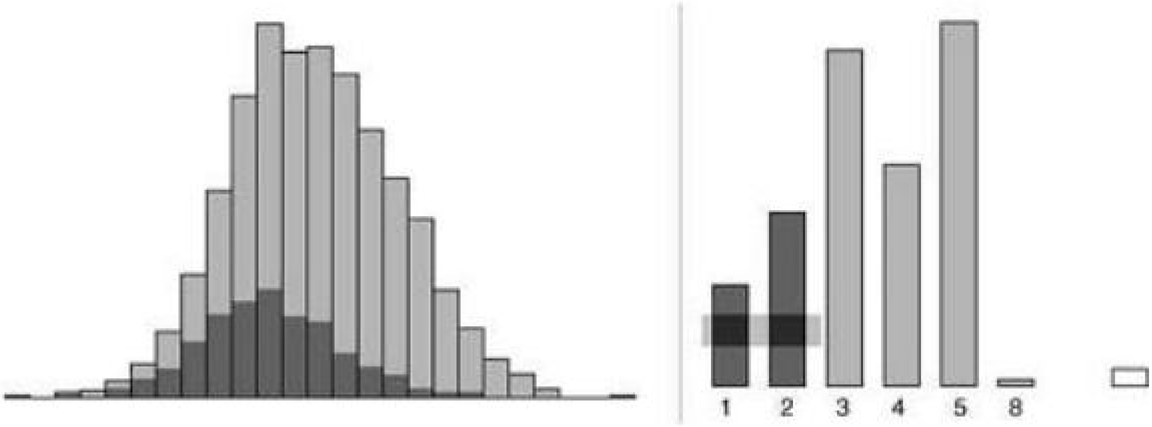


Figure 22: Overlaying strategy in linked plots. Source: Chen et al. (2008)

5.4.7.4 Special forms of linked views

More complex forms of linking such as *m-to-1* in hierarchical linking poses a few challenges. Take for example, two levels of a hierarchy: a macro level (e.g., a set of counties | states), and a micro level (e.g., a set of cities | towns). A partial selection of some cities | towns would be represented best by partial highlighting. The problem arises when the macro level is represented by non-regular shapes that cannot be subdivided properly. A general approach to this problem would be to use different color intensities to fill the according graphical elements, which is recommended for non-regular shaped graphical elements as well as other shapes for its easiness of decoding

5.4.7.5 Conclusions

Linking multiple simple 2D views by establishing relationships between plots that show different aspects of related data enables the user to explore and understand structures and patterns of more comprehensive data sets. This concept is essential in the field of visual data mining and provides the required *human-computer interaction* (HCI) to understand hidden structures and patterns. The linking procedures work best with complex data sets, as in the case of big data, which have very large number of observations and variables (high-dimensional), a mixture of variable types, as well as possibly incomplete and missing values. Generalization of linking and the use of a same scale ensure consistency in data views and comparisons of visual displays.

5.5 Visual Data Mining with V-Miner

5.5.1 Introduction

V-Miner (Visual Miner) is a multi-variable visualization tool proposed by Zhao et al. (1994) and described as being “designed for mining product design and test data”. The tool was developed and deployed at Motorola and aims to discover useful or actionable knowledge from mobile phone testing data. The uncovered knowledge might be provided as feedback to design engineers, who will in turn use it to improve both the product design and the product development process. Thus, the design cycle of new products can be shortened.

5.5.2 Product design process and available data

A typical design process for consumer products typically involves three stages: first, engineers will design various aspects of the product based on previous designs, new specifications, guidelines, etc., then electrical and software aspects will be finalized and prototypes are built. Finally, various tests will be carried out to determine if the product meets all the requirements. If not, the whole cycle will be repeated with modified design.

For existing product platforms, engineers have a good understanding of what works, what does not and possible solutions. For new product platforms, the amount of understanding is much less thus lead to more lengthy design cycles. V-Miner is developed to mine for useful knowledge from electrical test data that can be used to guide decision making in design cycles. After each design change, engineers will measure all electrical test variables, which consist of more than a hundred. The result measurement data is then mined. Each variable is numeric and has an upper and lower limit, which should not be exceeded to be accepted, as well as an ideal value called target value. Figure shows a sample test data. From the data, a few aspects are of interest such as (significant) changes in variable values after design changes and cause, stable variable values, patterns of values / changes / failures, etc.

Using traditional rule mining systems yield a few significant setbacks such as large number of rules generated due to large number of variables or only a subset of patterns are found but not all interesting ones, as in the case when a decision tree approach is used. V-Miner's visualization is based on an extended version of parallel coordinates called *enhanced parallel coordinates,* which also shows trend figures and enables querying by approximate matching.

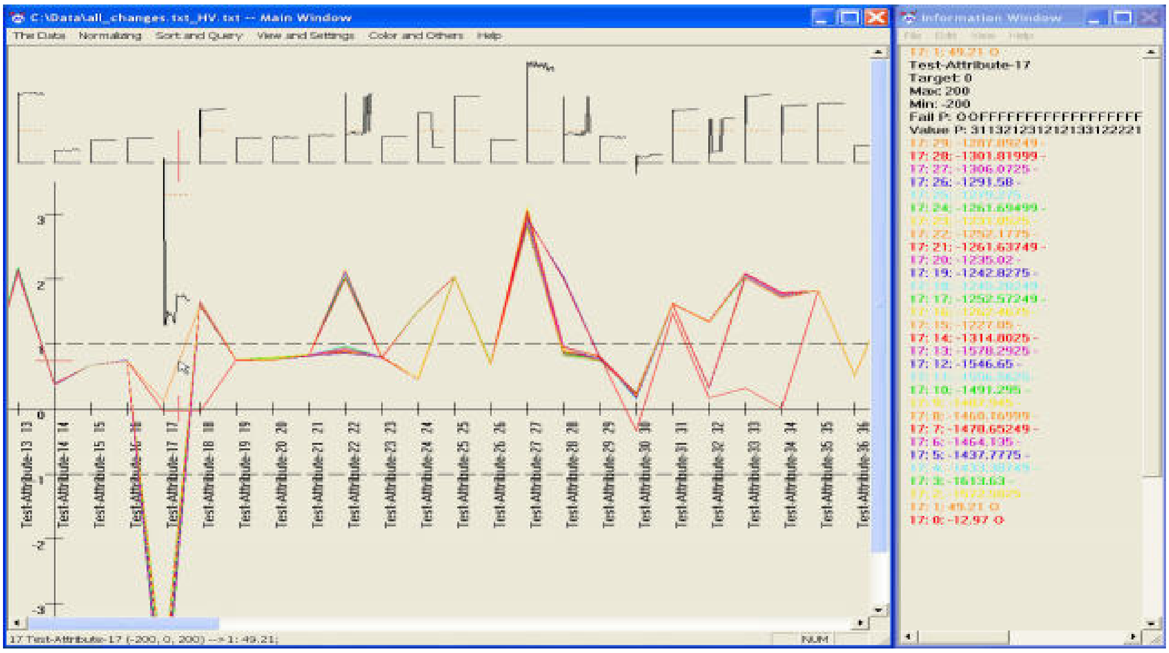
5.5.3 Typical usage scenario

After the data is normalized and loaded into V-Miner, the user is shown the visualization like the one in Figure 23 (with example attribute names). The left main window contains the visualization as well as possible user interactions. Test variables are shown on the x-axis with name. The y-axis shows the normalized value of each variable after every design change. A secondary window on the right displays detailed information as mouse cursor is being moved over points in the visualization. V-Miner visualizes data from different designs encoded using different colors and matches the color to the detailed information in the right window. A trend figure is drawn on top of the screen for each test variable, which shows the correlation of design changes and value changes of each variable. Similar change patterns should lead to similar figures. Dashed lines are shown on the y-axis at Y = 1 | -1 to easily identify the outstanders. Finally the querying function allows variable sorting so that interesting patterns and facts can be viewed quickly.

After data is loaded into the tool, the user can easily identify the outstanders (values that fall out of range, which is between -1 and 1), understand the similarity in the behavior of variable changes from the trend figures — therefore will be able to derive correlation among the variables in a sequence of changes, as well as identify the stable variables.

5.5.4 Conclusions

V-Miner is a visualization tool used to mine mobile phone test data with aim to aid design process. V-Miner extends parallel coordinate technique with trend figures — summarization of sequence-dependent trends in the data, and interactive features such as rearranging and grouping of variable. It also has a querying mechanism to allow the user to specify patterns to be mined. V-Miner still lacks the ability to analyze time-series data and guide product testing, which could be of potential interest.



```
17 Test-Attribute-17 (-200, 0, 200) --> 1: 49.21;
```

```
17: 1: 49.21 O
Test-Attribute-17
Target: 0
Max: 200
Min: -200
Fail P: OOFFFFFFFFFFFFFFFFFFF
Value P: 311321231212133122221
17: 29: -1287.89249 -
17: 28: -1301.81999 -
17: 27: -1306.0725 -
17: 26: -1291.58 -
17: 25: -1270.275 -
17: 24: -1261.69499 -
17: 23: -1231.0525 -
17: 22: -1252.1775 -
17: 21: -1261.63749 -
17: 20: -1235.02 -
17: 19: -1242.8275 -
17: 18: -1240.20249 -
17: 17: -1252.57249 -
17: 16: -1262.4675 -
17: 15: -1227.05 -
17: 14: -1314.8025 -
17: 13: -1578.2925 -
17: 12: -1546.65 -
17: 11: -1506.5625 -
17: 10: -1491.295 -
17: 9: -1467.945 -
17: 8: -1460.16999 -
17: 7: -1478.65249 -
17: 6: -1464.135 -
17: 5: -1437.7775 -
17: 4: -1433.38749 -
17: 3: -1613.63 -
17: 2: -1572.0625 -
17: 1: 49.21 O
17: 0: -12.97 O
```

Figure 23: V-Miner: initial display of the test data. Source: Zhao et al. (1994)

5.6 Summary

This chapter delved into the topic of *data visualization*, starting with a brief review of data visualization throughout history from pre-17th century until today. Visualization is considered a compelling approach to communicate the subtleties and complexities of the hidden information in data sets with amount ranging from small to large (big data). There are a few guidelines | principles to a good visual design and a wide variety of visual forms, such as graphs, trees, plots, etc., each suitable for different data types and scenarios. With the data acting as the model and the visualization itself as the view — as in the *model-view-controller* paradigm — adding interactivity via dynamic techniques, e.g., brushing, linking, etc., means adding the controller component to the visualization, thus making data exploration more viable.

Section 5.4 introduced a few visualization techniques designed with multi-variate data in mind. Treemap method, trellis displays, and linked views are attempts to address multivariate data visualization challenges by adding interactive component to the visualization besides utilizing other visual components and managing the distribution / partitioning of data points across multiple displays as well as linking them together. With today's data having become too large and often have too short a lifespan, more problems and challenges surface. The above visualization techniques have been around for quite a long time, therefore did not take into account one problem of big data: *data quality*, e.g., imperfection, defects, distortions, gaps, etc. in data points. An effective visualization system must therefore make users aware of the quality of the data by explicitly conveying data quality attributes, besides data content. The data itself must also undergo data cleansing processes before being visualized.

As more powerful displays — including mobile displays — becoming widely accessible and more common, visualization systems should make good use of display resources. Research and development are already underway to address new challenges associated with large amount of data, for instance Zaixian et al. (2006) discussed the data quality issue in multi-variate data visualization. A system designed to deliver visualization to mobile displays through web-based OLAP is described in (Tim et al., 2011), and finally in (Beyer et al., 2013), Johanna et al. described a system for interactive exploration of petascale volume biological data, taking into account the incompleteness of data and scalability .

6 The StreamViz: an Approach to Visualize Multi-variate Data on 3D Platform

6.1 Introduction

Data representation on a two-dimensional plane is still the predominant format of visualizations today. While most data can be well represented in tables and a variety of charts, the 2D environment limits the amount of data variables and properties that can be encoded, thus limit the ability of knowledge delivery and discovery. A 2010 TED talk by David McCandless[24] showcased how aesthetics factor can make two-dimensional visualizations more attractive, but interactive three-dimensional visualization is the next step in the evolution of data representation, and has potentials in the Big Data landscape as 3D technologies and -displays keep maturing. This chapter presents the *Stream Visualization Prototype (StreamViz)* — an approach to exploratory and interactive visualization in three-dimensional and Virtual Reality (VR) environment for time-series data. The basic concept is that time-series data can be represented in a stream-like flow form, and with the addition of a third dimension, more data variables and attributes can be encoded and meaningfully visualized. This concept is based on an idea in the Ice Bucket Challenge Visualization.[25]

Section 6.2 talks about the process taken when designing the StreamViz, including various considerations and requirements. Then, concept sketches, ideas, and requirements will be presented and discussed in Section 6.3. Section 6.4 presents two implemented varieties of the StreamViz prototype on two different settings, using open data sets: one is a three-dimensional implementation on two-dimensional plane (traditional display, 3D-over-2D), the other is immersive three-dimensional visualization in VR environment (Google Cardboard),

[24] http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en

[25] https://www.youtube.com/watch?v=qTEchen97rQ

which will be called *StreamViz VR.* These two implementations will be presented in consideration of various aspects including data exploration, usability, and efficiency, as well as the challenges surfaced. The chapter will conclude with a discussion about possible methods to evaluate the visualization with regard to the aforementioned aspects.

6.1.1 What is VR?

Virtual Reality (VR), also called *immersive multimedia* or *computer-simulated reality* is defined by Wikipedia[26] as

> “a computer technology that replicates an environment, real or imagined, and simulates a user’s physical presence and environment to allow for user interaction. Virtual realities artificially create sensory experience, which can include sight, touch, hearing, and smell.”

In short, it is the use of computer technology to create a simulated environment. Unlike traditional user interfaces, VR places the user inside an immersive experience. Instead of viewing a flat screen in front of them, the user is “immersed” and able to interact with 3D worlds. That makes it a perfect medium for a new way of storytelling, since the content author creates worlds and invite the user to experience them. This means that in VR the user is not only an observer but also a participant in the story with an option to influence the story itself.

6.1.2 Types of VR

Virtual Reality comes in different forms. Computers could generate images (CGIs) and display live images from the physical or real world. Then Heads Up Displays (HUDs), or Heads Mounted Displays (HMDs) can superimpose CGIs onto the real world . This type of setting is often referred to as Mixed or Augmented Reality.

There is 360° video technology — also known as *3DVR* and *Stereoscopic VR* — that uses multiple cameras to capture the image from 360 degrees. Technically, a standard 360° video is just a flat equi-rectangular video displayed on a sphere, akin to the face of a world map on a globe, but with VR, the user’s head / view is on the inside of the globe looking at the inner surface. As the user moves, the tracking mechanism on the VR device tracks the head motion, giving the user the feeling like they are inside the scene (immersed).

360° video is usually augmented with stereoscopic 3D which adds another level of immersion by adding depth between the foreground and background. With stereoscopic 3D in VR, that depth information has to be overlaid and mapped to the sphere. Because of parallax between cameras, this can be tough to achieve. There are anomalies often occur in

[26] https://en.wikipedia.org/wiki/Virtual_reality

badly designed VR experience, which makes it sometimes uncomfortable to watch or even promotes headaches, eye strain, and motion sickness.

The goal of storytelling in VR is to immerse the viewer in the created world, then serve them with the story. Being immersed in the virtual world, the user will live the story and probably be able to find it more convincing, engaging, and comprehensible, especially when combined with interactions.

6.2 Approach and Risk Assessment

6.2.1 General structure

The application (the visualization as a whole) should consist of four components following the 4x4 model for knowledge content, devised by Bill Shander[27] (Figure 24). These include visualization, story telling, interactivity and shareability. The first component — visualization, describes content that is short, succinct and direct, usually reflecting ideas. The purpose of this content is to engage audience and grab attention. The next level of content is story telling — content that is longer and a bit more in-depth — which is a progression from the previous level and further explains the ideas. It should tell a compelling story in order to have good relatability and thus is the most difficult to create. The interactivity component dictates that content should give the audience a true in-depth knowledge of the topic | ideas presented in the previous components and can be backed by raw data. This kind of content is not common but most powerful, the audience can adjust and filter the data to their interest.

6.2.2 Set a theme for the story

A news headline, a tweet, or a thirty-second video could all be the sources for the first inspiration for the story. Theme can varies from sports to politics, to healthcare and finance, or to climate and education, which will be filtered based on interest and domain knowledge of the content author. The main question in this step is “What topic will the story have?”.

The implementations for the StreamViz will set *education* as the main topic for the story. Concretely, the data will revolve around post-secondary education (e.g., bachelor and master) in all institutions across the United States with an emphasis on graduation rates and number of enrollments.

[27] http://inspiredm.com/winning-knowledge-content/

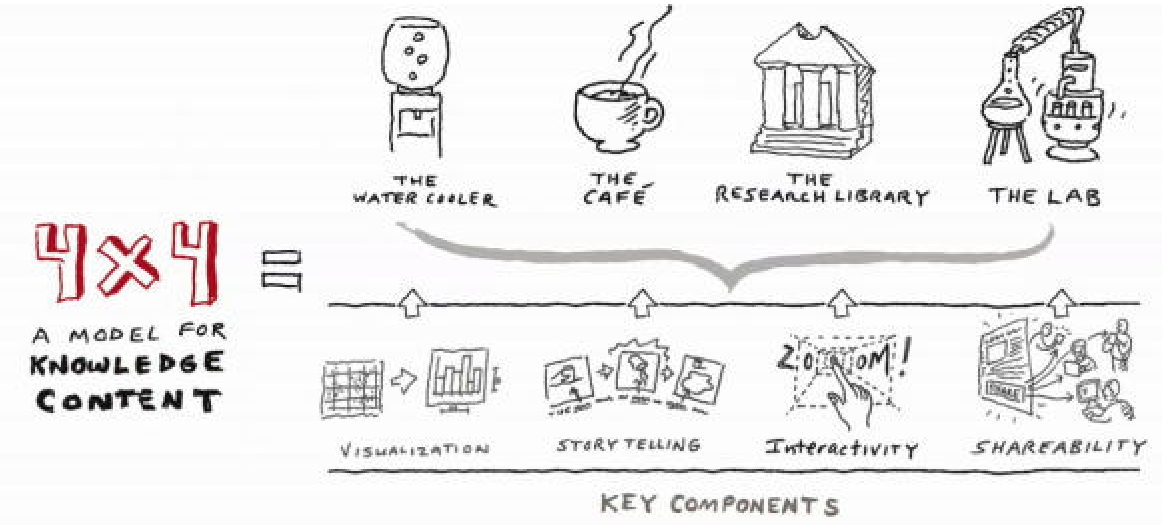


Figure 24: The 4x4 model for winning knowledge content. Source: Bill Shander

6.2.3 Channel the audience

As in any communications, whether it is a website, a video, or a blog post, visualization also requires knowing the audience to adjust the presentation to them. Among other things, the following points outline the most important aspects that we need to understand about our audience:

- Culture: has an effect on language, perspective, context and color, etc.,. What looks odd in one culture might be completely normal in another one. Another important factor that varies among cultures is the narrative context. For example, a visualization about hockey statistics for people in northern countries, with whom hockey is a more familiar sport, certainly needs less context and is more comprehensible than if it is for people in other parts of the world. The question is “does our audience know the underlying story of what we are talking about?”. The demo implementation of the StreamViz in this thesis targets general audience, with predominantly audience from Germany expected.
- Level of expertise: affects the amount of context, the type of used language, etc., The language of the story might contain more lingo and less context or vice versa depending on the level of audience’s expertise in that topic. More background information might need to be provided and the story might be shallower with less details for general audience. Expected audience for the StreamViz consists of predominantly decision makers e.g., managers as well as corporate employees, team members, etc., thus moderate level of knowledge of the subject is required. The demo implementation of the StreamViz targets general academic audience, who might possesses good

background knowledge of the education field but eventually be beginner | novice in the field of VR.

- Consumption context | channel: the environment, in which the story is going to be published, also affects the approach to visualization. If it is a more serious environment, a higher standard of excellence will be required, with more statistical integrity. Otherwise, a less details-oriented approach and lower journalistic standards might be acceptable, which the StreamViz implementations will adopt.
- Accessibility: the required level of accessibility is also an important factor that affects color, contrast, font size, among other things. The visualization project is going to target mainly sighted people, so it might not need much effort to ensure a good level of accessibility. However, color blindness is an important issue that needs to be addressed if a high level of accessibility is aimed for. Roughly 8% of men and 0.5% of women have color vision deficiency[28], with the most common form being Deuteranopia (red-green color deficiency). There are tools to help mitigate this problem, e.g., the Color Blindness Simulator enables the simulation of the color perception of various color-blindness forms[29]. The StreamViz implementations target sighted people and therefore will not take into account color blindness issue.
- True believers | skeptics among the audience is an important factor to consider. Understand the skepticism and argument of the audience against the story and the data behind it helps reduce bias, which might affect the credibility of the visualization and the story. Whether the visualization project aims to change minds, to convince people, or only to provide “facts” also has an effect on the level of interactivity and detail.
- Action | reaction of the target audience needs to be planned to design better outcomes. Questions to be asked would be “do we need our audience to perform a specific action (e.g., share on social media, answer a poll question, etc.) after seeing the visualization and learning of our story?” and “does our design lead toward this outcome?”. Since StreamViz is a web application, it has high shareability, which is a desired action for a “fact providing” visualization. Besides that, no further action from the audience is planned. As for user’s reaction, the desired outcome would be that the audience would find the three-dimensional setting | immersive experience more engaging, intuitive and easy to grab knowledge from than traditional two-dimensional setting.

6.2.4 Data acquiring and preparing

Data visualization itself is the end artifact, after a multi-step process — including finding reliable data sources, formatting and cleaning the data, and finding the story it tells. Sourcing

[28] http://www.color-blindness.com/2006/04/28/colorblind-population/

[29] http://www.color-blindness.com/2010/03/16/red-green-color-blindness/

a large and interesting data set in the age of Open Data is easy. A few example sources include:

- Open government data, including city-specific and political data from governmental and regional data portals such as Data.gov, Socrata, Transparenzportal Hamburg, DeStatis, etc.
- Data aggregators house data from various sources, which help finding category-specific data easier. Example: Programmable Web, Infochimps, Google Public data explorer, etc.
- Social | news data: using APIs provided by social and news sites such as Instagram, Foursquare, Twitter, Facebook, The New York Times, The Guardian, etc. It is possible to access and explore data on each particular platform (news feeds, articles, etc.)

The StreamViz prototype implementations use educational data sets — published by the U.S. goverment open data portal[30]*,* in the Education section. The data taken is part of the Integrated Post-secondary Education Data System and sampled within a period of ten years, from 2004-14 (IPEDS 2004-14) which is introduced as:

> "a study that was part of the Integrated Post-secondary Education Data System (IPEDS) program; program data is available since 1980 at *http://nces.ed.gov/ipeds/*. IPEDS (http://nces.ed.gov/ipeds/) was a web-based system designed to collect basic data from all post-secondary institutions in the United States and the other jurisdictions. Key statistics produced from IPEDS allowed the National Center for Education Statistics (NCES) to describe the size of one of the nation's largest enterprises–post-secondary education– in terms of students enrolled, degrees and other awards earned, dollars expended, and staff employed."[31]

For each year, the IPEDS will conduct a series of surveys in all educational institutions across the United States. The StreamViz demos will use data from the following surveys: *Institutional Characteristics, 12-Month Enrollment* and *Completions.* For each data set except the Institutional Characteristics, the data attributes of interest are among others Total Completions for male | female (*CTOTALM* | *CTOTALW*) and bachelor | master (*AWLEVEL*), Total Completions for male and female | bachelor and master (*CTOTALT*), *CIPCODE*, *UNITID*, etc. Some variables are aggregated attributes, calculated from other variables by the StreamViz application.

Big data sets almost always contain errors, such as false or missing values | characters and often not come in the right format to be parsed and processed. Therefore data adjustments, conversions and cleaning need to be done. IPEDS data sets come in CSV format and contain inconsistencies e.g., between attribute names due to changes in naming schemes during the

[30] *https://www.data.gov*

[31] https://catalog.data.gov/data set/200506-integrated-postsecondary-education-data-system

ten-year time window. Besides, data from Institutional Characteristics also contains inconsistencies due to missing / excessive institutions, varying with each year. As such, all data sets must be brought back to the same consistency level. Using basic tools and spreadsheet application (Microsoft Excel) with advanced functions like Pivot Table, extra data parameters can be easily calculated and missing values cleaned up. These steps are inevitable and part of the data exploration process, and risk consuming a large amount of time due to data complexity.

The Institutional Characteristics data set contain approximately 7660 data values, representing the same number of surveyed educational institutions across the United States. All other data sets contain values for each of these institutions, sometimes with redundancy and thus have approximately at least 150,000 data values each.

Understanding the data is the next step of the process, which helps reduce errors and increase accuracy. Basic mathematics and statistics knowledge is applied to calculate additional parameters, such as mean, median, actual | rank indices, percentile, etc. IPEDS data sets come each with their own *meta data sets* which contain important information such as variable list, explanation, possible value ranges, sample size and sampling methodology, etc. These pieces of information are valuable in evaluating the quality and reliability of the data. Being able to establish accurate relationships (correlation | causation) between data points and sum the data in a few main ideas / headlines help avoid making false claims and delivering false knowledge to the audience. Using Excel, quick visualizations can be made which would help obtain a comprehensive first-look of the data and establish hypotheses.

6.2.5 Define the narrative for story telling

Interactive visualizations are not necessarily meant to be consumed in a linear way, and thus should not control how the audience processes the information. Instead, the StreamViz demo aims to create a story taking advantage of the nature of time-series data, with a narrative process. The goal is to encourage but not forcing the audience to walk through the information in a linear, progressive way while exploring the data at the same time using sorting and filtering mechanisms. The basic structure of the story should include a *beginning* (headlines, introduction), a *middle* (call-outs, main ideas / theses, data, details) and an *end* (conclusion, data sources, follow-ups). Imagery and metaphors should be applied if possible to increase relatability to complex data facts, and giving the audience deeper impressions and better comprehension. This is not vital and compulsory but a helpful addition to the visualization.

6.2.6 Experiment with visual designs and elements

Before going into actual design work, it is important to first experiment with different visual designs and elements with mock-ups, e.g., using wireframes and sketches. The advantages they offer include speed, flexibility, and scale, all of which are vital to get to ideas and iterate on things quickly without having to know how to implement them technically or their feasibility. Typical visual elements to consider include:

- Illustration and iconography: used to capture attention, reinforce themes | linear story telling structure and make content more relatable, therefore must be content relevant and theme-based. Imagery should be uniform and clear, as to not obstruct the reading of data values and content. Risks: difficulties in graphics designing | sourcing and overuse of imagery.
- Typography: is also used to capture attention, emphasize content and can change perception and understanding of the audience. Depending on the type (axes, legends, labels, infographics, call-outs, etc.), different typeface, font weight, etc., will be applied but should remain uniform across the project. Risks: difficulty in maintaining balance between accuracy, readability, story telling, data granularity and aesthetics might lead to false perception | knowledge.
- Position, size, shape, color and contrast: are five main elements to show variance in the data and create distinction among objects. Color and contrast are useful to create emphasis and highlights, but might pose a challenge for visually impaired audience.
- Scales: have a big impact on perception and must be selected carefully to reflect accurately the relationships in the data. Bias could lead to choosing the wrong scales, thus delivering false impressions and knowledge to the audience.
- The right paradigm: depends on the number of variables, the type of data (hierarchical, network, geographical, etc.) and the required level of aesthetics and uniqueness, choices must be made between various visual paradigms to represent the data. Be it basic graphs, charts and maps, or something new, creative and innovative, or a combination of these which is the chosen approach for the StreamViz demo, the balance between accuracy, readability and aesthetics must be maintained, which could be a difficult and risky task.

Actual visual design | element experiments for the StreamViz will be presented more concretely and in-depth in form of concept wireframes and sketches in the next section.

6.2.7 Select the right technologies for implementation

Interactive visualization requires technical implementation. There are a wide variety of technologies for creating visualization with different features and benefits. The most important criteria for picking the right combination are outlined below:

- Platform vision: whether the visualization project is a short-term or long-term one affects the choice of platform. In case of the StreamViz demo, it is a short-term one and thus does not require reusability. Therefore the chosen front-end platform should offer simplicity and speed (web application with modern front-end technologies such as HTML5, JavaScript, CSS3, etc.). Otherwise, the StreamViz generally can be implemented with a more complex, scalable and modular back-end platform to offer reusability and robustness (bringing in more server-side and Big Data technologies such as NoSQL DBs, PHP, NodeJS, etc.).
- Audience: can be categorized into tech-savvy and general, less tech-savvy people. For modern, techno-driven audience, implementing using modern technologies should not pose any challenges. Otherwise, device compatibility could be an issue. For instance, Flash technology is not compatible with iOS devices, or older versions of various browsers do not play well with some modern web technologies such as SVG, WebGL, WebVR, etc.,. In case of a broader, mixed audience, cross-browser | platform technologies can be used with fall-back mechanisms (browser | platform detection, alternatives, etc.). The StreamViz demo will assume a tech-savvy audience and thus will not take into account browser | device compatibility.
- Visual | conceptual goals: the complexity of the project from a visual standpoint also plays a role in technology choice. Out-of-the-box software only offer limited features and visualization capabilities. Complex visual shapes and ideas require more technical and versatile platforms. Risks: time cost to learn the required technologies.

The StreamViz requires knowledge in 3D | VR technologies such as WebGL, Unity, WebVR, etc. As such, more time is needed to take the project from initial design stage to actual implementation, including the technology learning time.

6.2.8 Share, study and assess results

As with any visualization project, shareability is a desired goal, so that user feedback can be gathered and used to improve the next iteration of any further visualization project. One of the possibilities to achieve this is through a mini user survey with a limited questionnaire of about five questions. Sample size might be limited to approximately five to ten participants. The goal of the survey is to gather user feedback on usability and user experience of the visualization. Because the survey is small scale in nature, there is always a risk of bias, so that the results have to be interpreted carefully.

6.3 Early Concept Specifications and Design Sketches

6.3.1 First concept: with e-Commerce data

The StreamViz is an approach to interactive and exploratory visualization which exploits new visualization paradigms, such as 3D graphics, VR, and animations, to allow for richer user experiences when visualizing time-series data. The original concept for the StreamViz is that through the addition of a third dimension, more data variables may be visualized, therefore giving the user a more comprehensive view of the data without having to split the visualization into multiple linked charts. This concept is developed with e-commerce (e.g., online shop) data in mind (Figure 48). The goal is to visualize the time span of shop orders, from the moment a customer gave up an order until the moment that order is fulfilled, together with all the sale developments during the lifespan of that order. Those events could be email correspondence between shop system / employees and the customer, including purchase confirmation | invoice, payment receipt, delivery confirmation | tracking information, etc. In the initial concept, the StreamViz was also envisioned to visualize data from product marketing campaigns, enabling the shop operators to track performances of those campaigns easier. It might visualize posts | mentions | tweets on social media as well as user reactions (like / dislike) about a product.

In this use case, data from the sales data warehouse should be visualized. This data warehouse has a star schema and aggregate sales data from many database tables, from order details to customer details and product details. The main requirement is that the user should be able to read out key facts — such as sales performance over a specified time period, from the visualization. Because the data warehouse can be potentially huge, the data set is expected to be big that can scale up to millions of data rows. This presents a challenge to the usability of the visualization. One possible approach to this challenge would be to incorporate multiple views into the visualization that implements various levels of aggregation.

The visual elements should be predominantly 3D objects, with the exception of some user interaction elements such as menus, buttons and pop-ups. The application should be web-based and follows the client-server paradigm. The StreamViz is designed for time-series data and as such, the data must be able to be ordered chronologically.

6.3.1.1 Aggregated view

Aggregate all data points (e.g., purchase orders) in a specified timeline and visualize them (Figure 25). The x-axis shows the timeline with each tick representing a time unit (e.g., a month or a year) in that time period. The y-axis' values show sales performance, represented by revenue. The Stream has a cylindrical form (a tube), and is segregated into different clickable areas according to the ticks on the timeline. Various filters can be applied to the visualization for better usability, including data type and data set filter. A filter for meta data (e.g., store location, manufacturer, etc.) can be useful to reduce clutter in the stream. A typical use case for this view is that the user wants to view the details on the performance of a particular time window — such as within a month — so the user clicks on the part of the stream corresponding to that time window to go into Detail View.

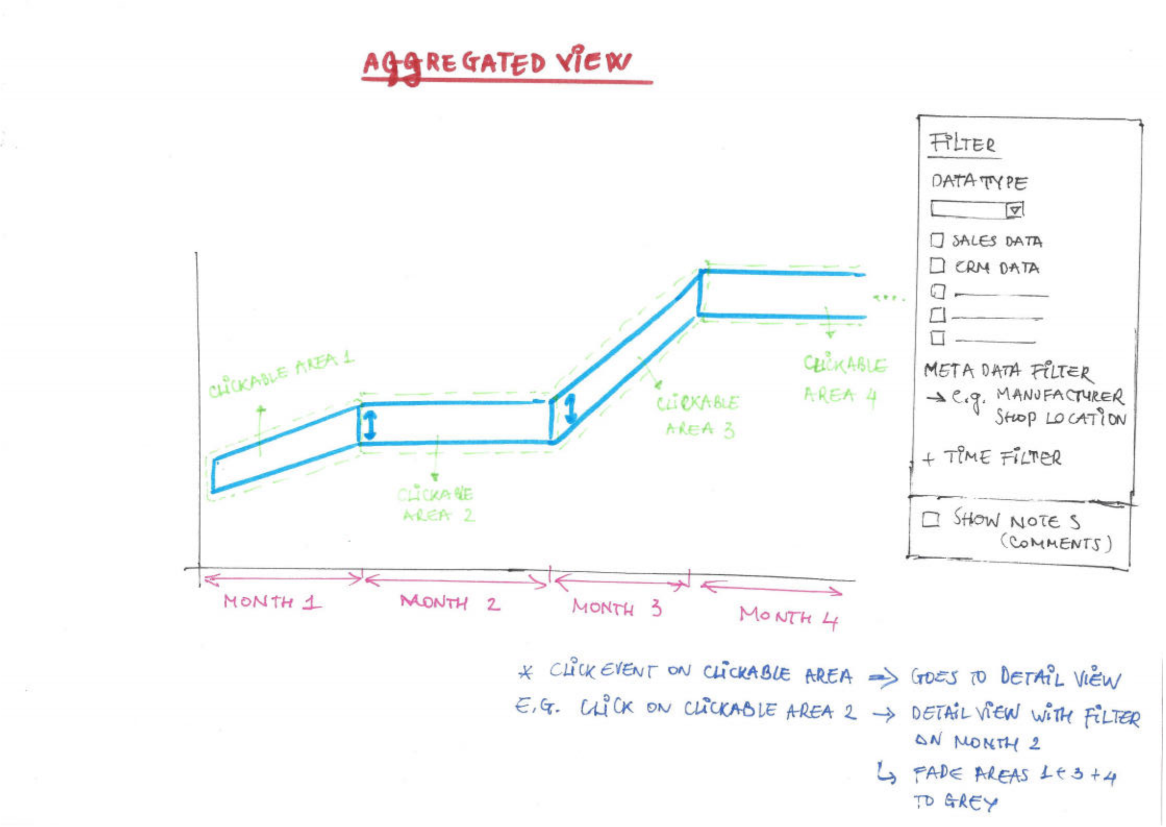


Figure 25: The StreamViz concept – Aggregated View.

6.3.1.2 Detail view

The Detail View shows more details for data points in the selected time window (Figure 26). The corresponding part of the stream should be put into focus and the other parts grayed out. In this view, all individual data points are clickable, which will take the user into Context View. Statistical indicators like average, min, max can be overlaid and toggled.

6.3.1.3 Context view

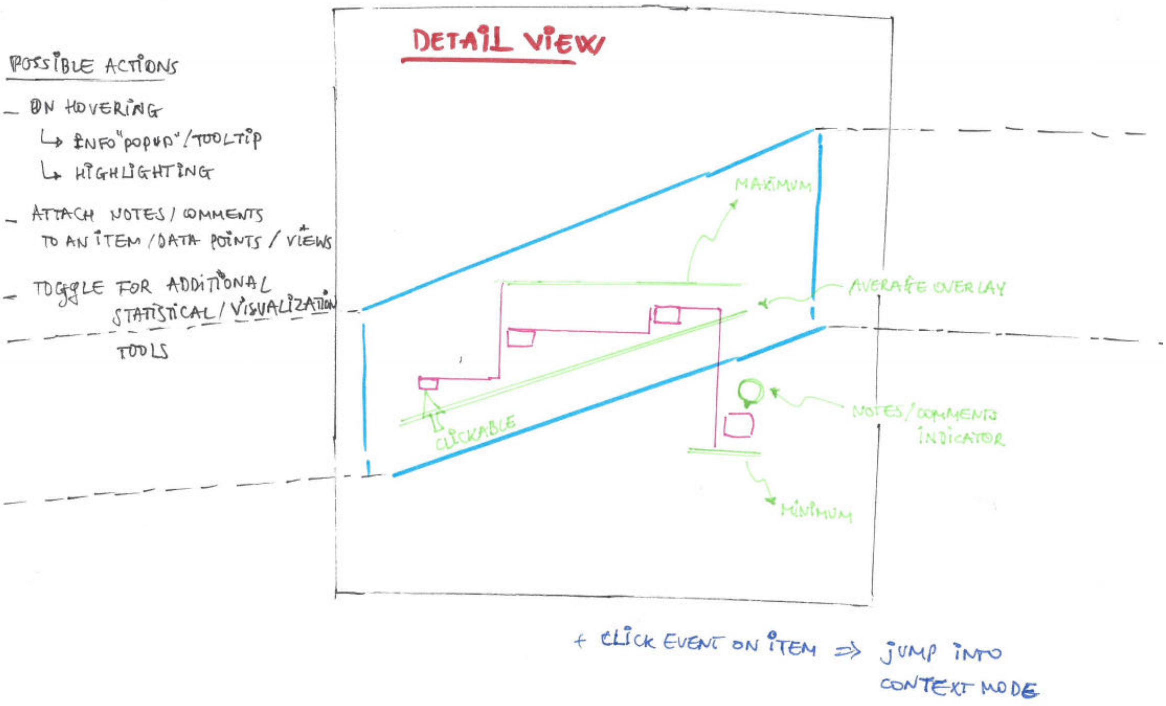

DETAIL VIEW
POSSIBLE ACTIONS
– ON HOVERING
↳ INFO "POPUP" / TOOLTIP
↳ HIGHLIGHTING
– ATTACH NOTES / COMMENTS TO AN ITEM / DATA POINTS / VIEWS
– TOGGLE FOR ADDITIONAL STATISTICAL / VISUALIZATION TOOLS
MAXIMUM
AVERAGE OVERLAY
NOTES / COMMENTS INDICATOR
CLICKABLE
MINIMUM
+ CLICK EVENT ON ITEM ⇒ JUMP INTO CONTEXT MODE

Figure 26: The StreamViz concept – Detail View.

In context mode, each individual data point (e.g., a purchase order progress) is shown together with all meta data (contextual information) associated with that data point. For example, relevant marketing data (e.g., Google AdWords data) can be linked to a purchase order based on date and content. Correspondence (e.g., emails) during the purchase process can also be linked to the timeline, together with tracking data and customer feedback. Additional information pertaining to order items can be aggregated and shown based on user interaction (e.g., mouse hover, mouse clicks, etc.). (Figure 27)

6.3.1.4 Adding notes | comments

The user should also be able to add comments / notes throughout the visualization. Comments | notes can be stored using standard techniques, including in traditional SQL or NoSQL databases. Figure 28 shows an example of table structures for storing comments | notes.

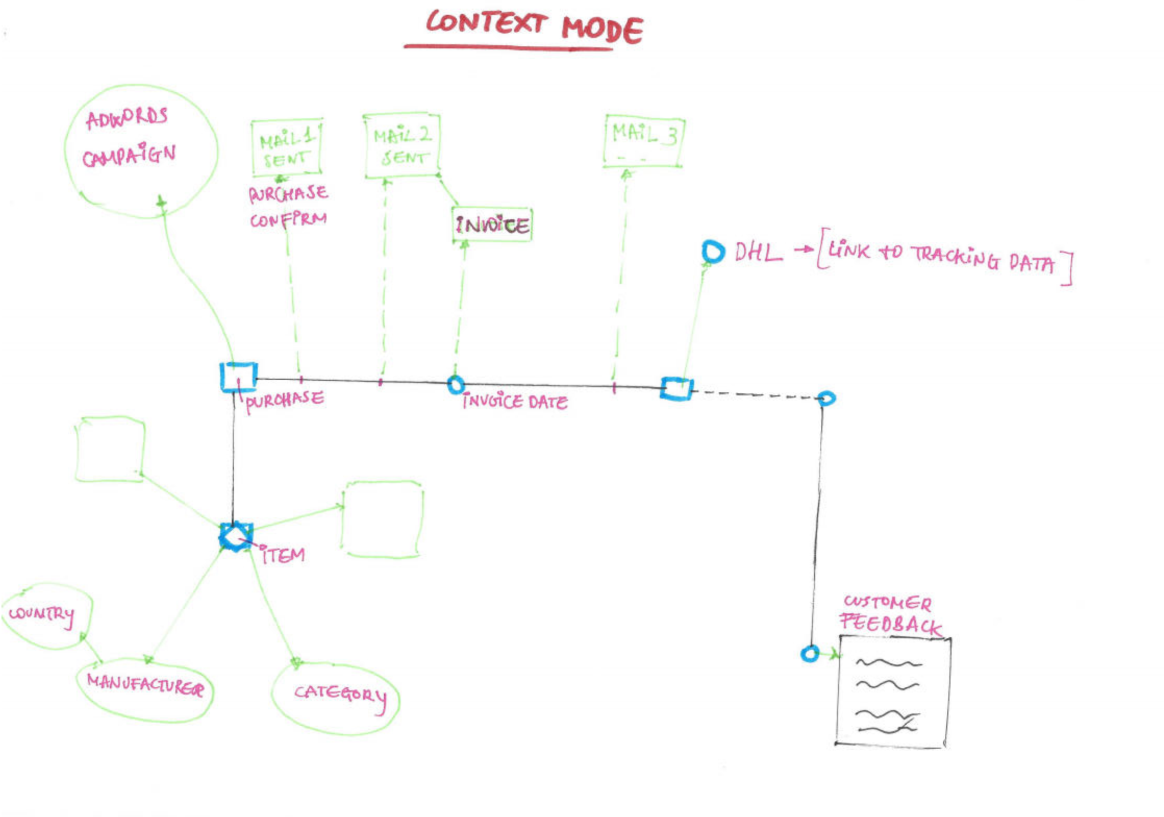
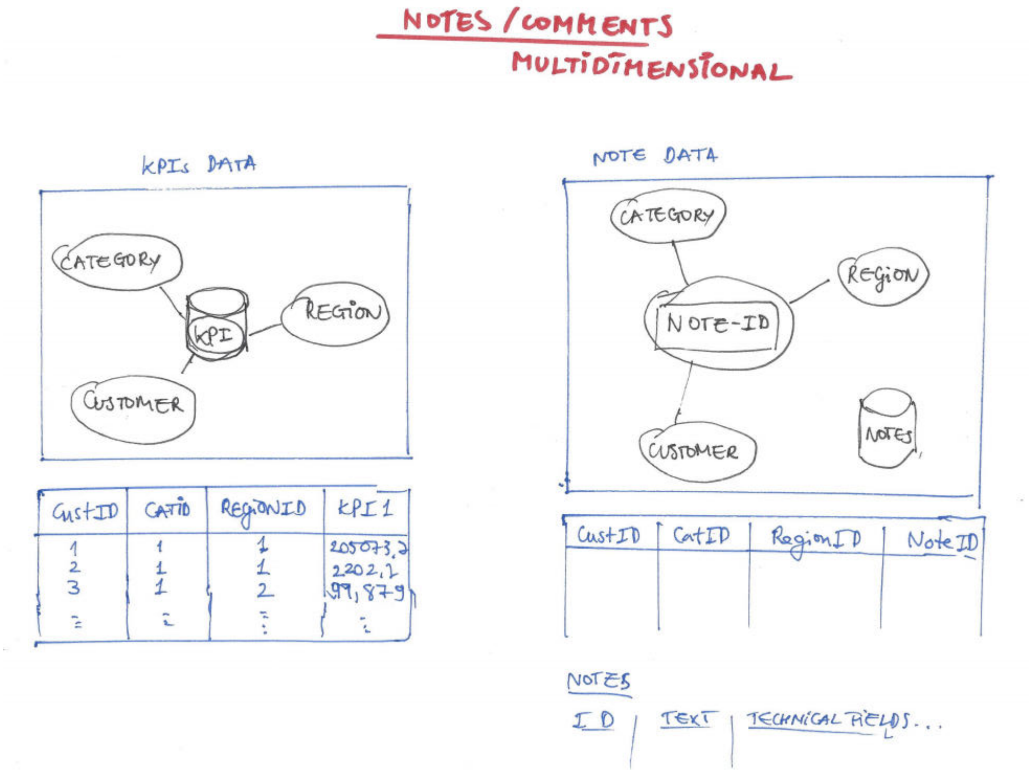


Figure 27: The StreamViz concept – Context View.

6.3.2 Second concept: Visualization of the development of refugee camps over time

There were a few challenges for the demo implementation of the first concept of the StreamViz. First, real world sales databases are quite difficult to acquire. Second, the visualization might become overly complex such that special domain knowledge (e.g., in finance) would be required for the user to evaluate the visualization. In such cases it could eventually drastically reduce user experience and thus introduce bias into the end result.

In this second concept, the StreamViz would be used to visualize the development of refugee camps in Germany (or possibly on a regional scope). With the recent influx of migrants into the country, more and more camps are being built around cities to try to meet the ever-increasing demand for accommodation. This visualization aims to study this flow of migrants and the ability of the government to cope with it by visualizing the a camp's capacity and the actual occupation over time. Real data might possibly be acquired through open sources | government agency contacts. For a proof-of-concept, dummy data might also be used in place of real data. The benefit of this concept is that the visualization is greatly simplified, thus no special domain knowledge is required and the feedback will remain as objective and



CustID	CATID	REGIONID	KPI 1
1	1	1	205073,2
2	1	1	2202,1
3	1	2	99,879
:	:	:	:

CustID	CatID	RegionID	NoteID

NOTES

ID	TEXT	TECHNICAL FIELDS...

Figure 28: The StreamViz concept – Conceptual data structure for comments / notes.

unbiased as possible.

The visualization could be split into different views and aggregation levels. The camps are put on a map according to their geographical locations. The displayed map will vary based on different aggregation levels (e.g., country map with states, state map with cities, city map, etc.). The timeline is limited to a two-year window with each tick represents a month. The camp's planned capacity is represented by outer 3D cylinders. The camp's actual occupation is represented by nested (also 3D) cylinders. Colors are used to encode the ratio between planned capacity and actual occupation, as well as different camps.

6.3.2.1 Front view (2D)

The front view of the visualization is basically a 2D map with various circles — representations of camps. The location of the circles on the map corresponds to the geographical location of the camps (latitude | longitude). Each camp is represented by two nested circles, with the inner one showing the actual occupation and the outer one showing the planned capacity. More details are shown on labels by mouse hovering over the camps. (Figure 29)

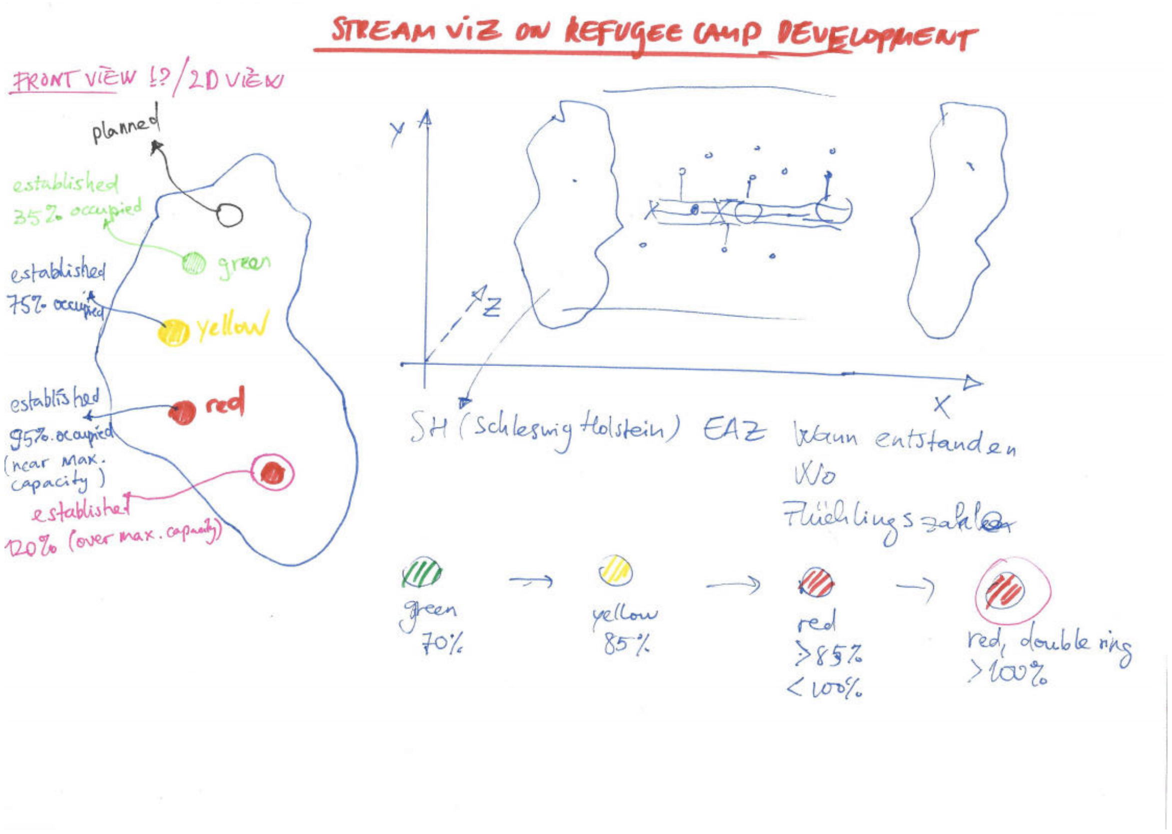


Figure 29: The StreamViz concept – Front View of the Refugee camp developments visualization.

6.3.2.2 Side view (3D)

The stream representing the camp's developments over time is shown as 3D nested tubes with changing radiuses denoting the difference between planned capacity and actual occupation. More details and meta data — legends, raw data, etc. — can be shown around the 3D scene and on labels on mouse hovering. The timeline ticks are spread along the z-axis at the base of the map in one month-step and will rotate as the user change view perspectives. The parts of the stream can be clicked on to show more details about a particular camp at a particular time window. (Figure 30)

6.3.2.3 Data point connection variants

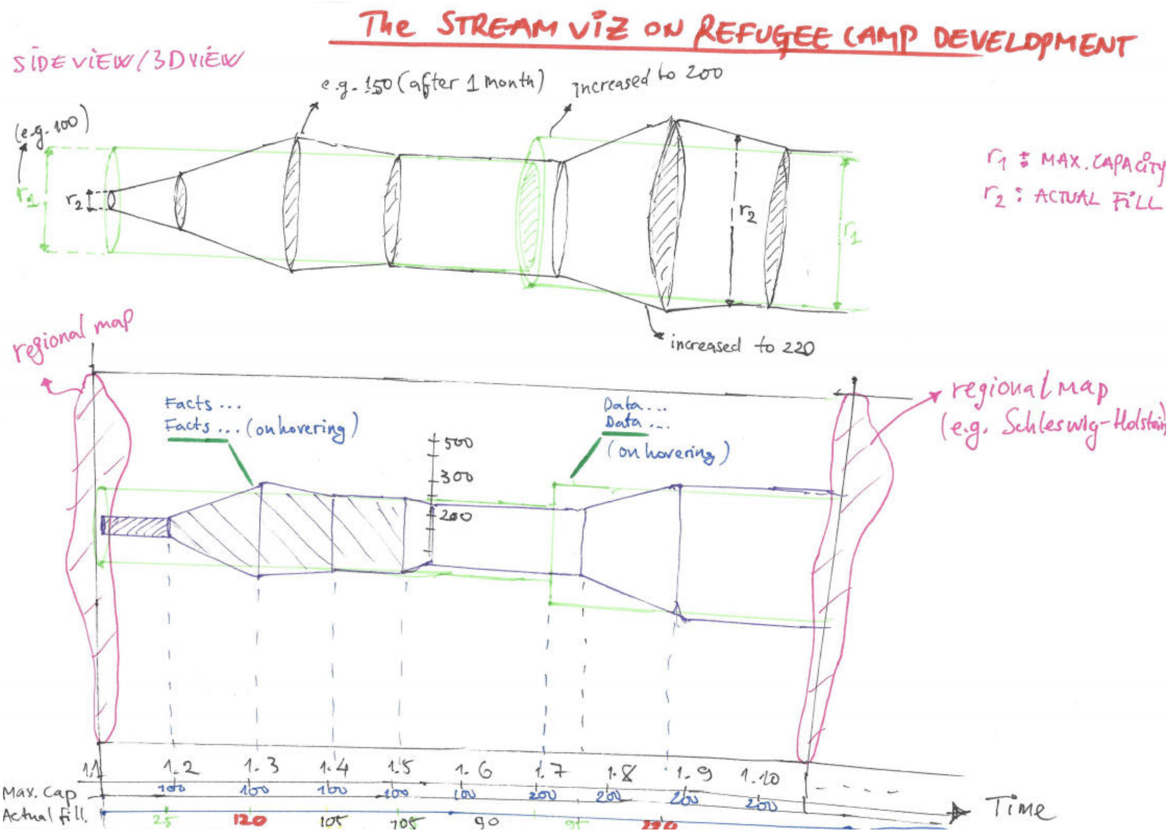


Figure 30: The StreamViz concept – Side View of the Refugee camp developments visualization.

There are two options for data point connection. The *point-specific* connection projects the tube's radius onto the next plane consistently, i.e., the tube size stays the same between two planes. Whereas with the *flow-specific* connection, the tube size on the other end will gradually increase or decrease to match the radius of the next cylinder on the next plane. The latter approach has the advantage of having more aesthetics and transfer immediate knowledge (e.g., of the deviation in the camp sizes between two time windows, thus slightly improve usability. The disadvantage being that it is very difficult to implement other features such as showing detail camp information for a specific time window when the user clicks on the relevant tube. (Figure 31)

Figure 32 shows a sketch of the rendering of the flow-specific 3D cylinders.

6.3.2.4 Implementation with dummy data

The visualization is implemented mainly in JavaScript using Three.js, which is a 3D JavaScript

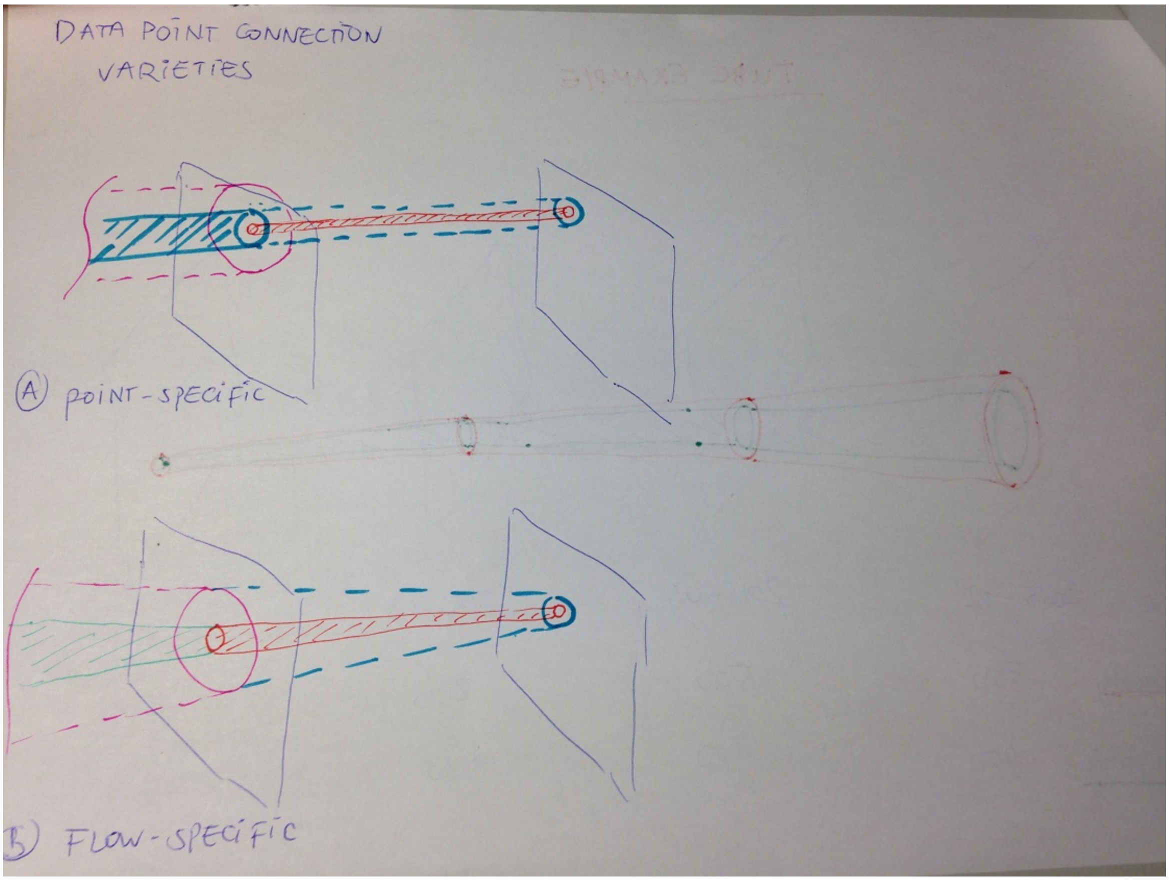
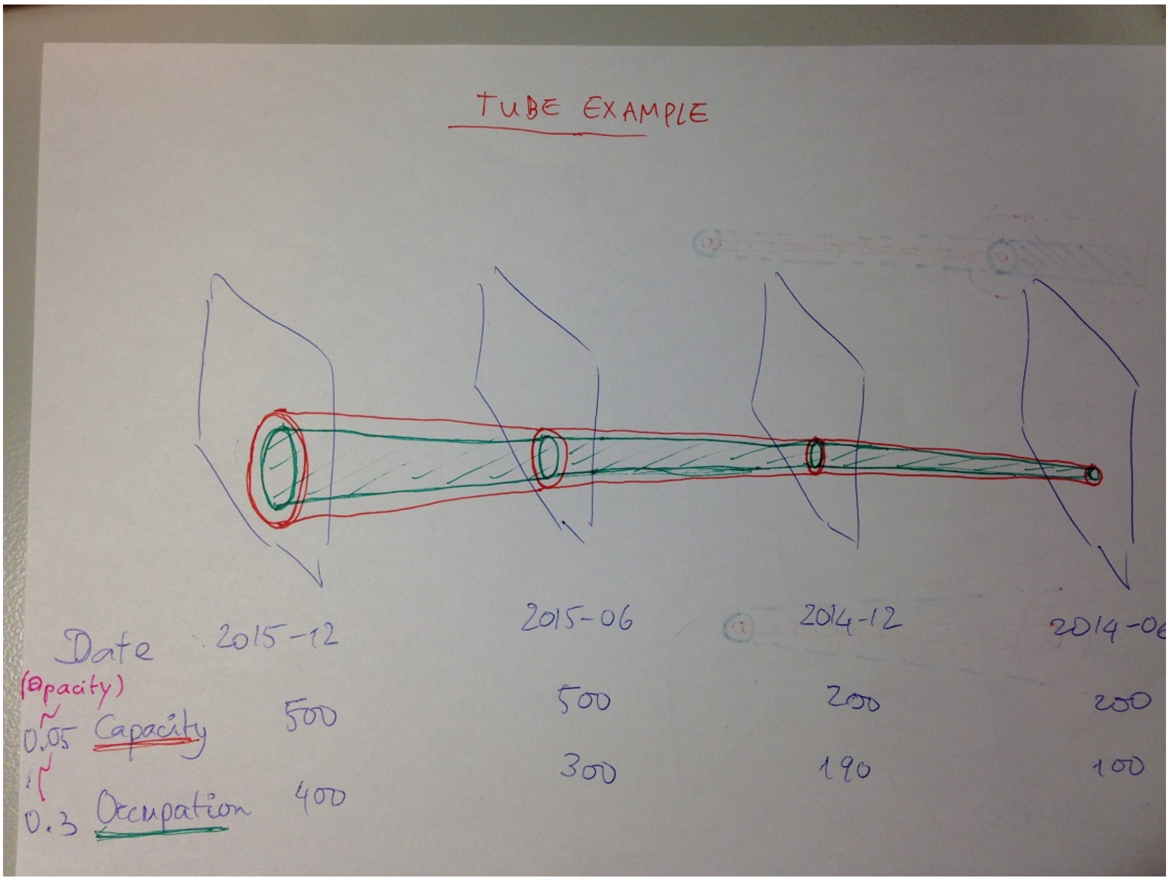


Figure 31: Data point connection variants.

library that allows constructing and rendering of complex 3D scenes and objects and also offer support for various controls (e.g., OrbitControls, DeviceOrientationControls, etc.). The rendering is done on client-side in the browser using WebGL. Initially, the geographical coordinates (latitude, longitude) of the administrative divisions (e.g., city, state, country, etc.) are queried against Google Maps API for the map images. This method is proven to be more difficult than using LeafletJS framework with OpenStreetMap API. Using this approach, the coordinate pair with a specific zoom level is sent to OpenStreetMap API via LeafletJS, then converted into a static image using Mapbox's Leaflet-Image library. Because this process is time consuming, the result map images might be stored in a database or cached into the client browser's LocalStorage for future retrieval, which will significantly reduce the overhead and thus the overall rendering time of the visualization.

Figure 33 and 34 show early renders of the StreamViz. The outer cylinders are color-coded according to the camps while the inner cylinders are traffic light color-coded based on the ratio between the camp's capacity and occupation (e.g., red indicates the camp was overcrowded for that month, yellow indicates near-full occupation, etc.). A legend containing a



Date	2015-12	2015-06	2014-12	2014-0
(Opacity) 0.05 Capacity	500	500	200	200
0.3 Occupation	400	300	190	100

Figure 32: The StreamViz: 3D cylinders – flow-specific.

list of all camps is shown in the top right corner. In the top left corner is a selectable list for all possible administrative divisions (e.g., states, cities, etc.)

6.4 Third Concept: StreamViz Prototype with Education Open Data

Because at the time this thesis is written, the migrant flow phenomenon in Germany has just only begun and the development of the refugee camps are still in its early stages (maximum one and a half year since the surge), acquiring sufficient real-world data for the visualization attempt is not possible (a 10-year period would be optimal). Therefore the third concept of the StreamViz is built upon real open data sets, namely education data sets from the U.S. Open Data portal. For this concept, two demos are built — one using the same technologies and settings as the previous concept, and the other one will be using Virtual Reality environment setups.

The data sets are cleaned, structured and loaded into a MySQL database server. The

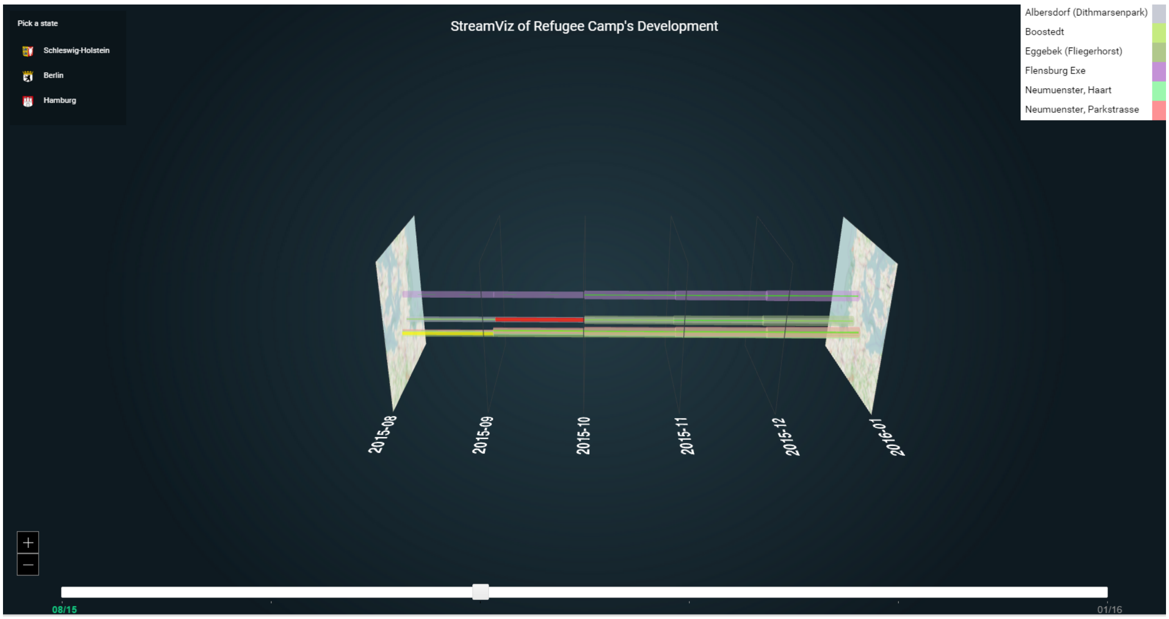


Figure 33: Early side view render of the StreamViz - refugee camp visualization.

StreamViz will then perform different queries on these data based on different options and filter settings, and calculate the needed variables to visualize.

6.4.1 Demo A: Visualization of U.S. IPEDS Graduation Data Set (3D-on-2D)

6.4.1.1 Side view

The IPEDS post-secondary education data sets contain data for 50 states of the U.S. and its five major territories, from 2004 to 2014. In the upper right corner, the user is presented with a few options to filter and control the visualization. Some of which are: visualization of different ratio combinations such as the completion (graduation) rate for Bachelor / Master degree, completion rates among males | females, for either Bachelor or Master degree. The user can also filter based on a particular state | territory by using the check list on the right. Each state / territory is color-coded and correspond to the color of the associated checkbox. The slider and zoom controls at the bottom allow the user to walk through the timeline with ease. For each time window, the user can view detail information by hovering over the cylinders. (Figure 35)

6.4.1.2 Front view

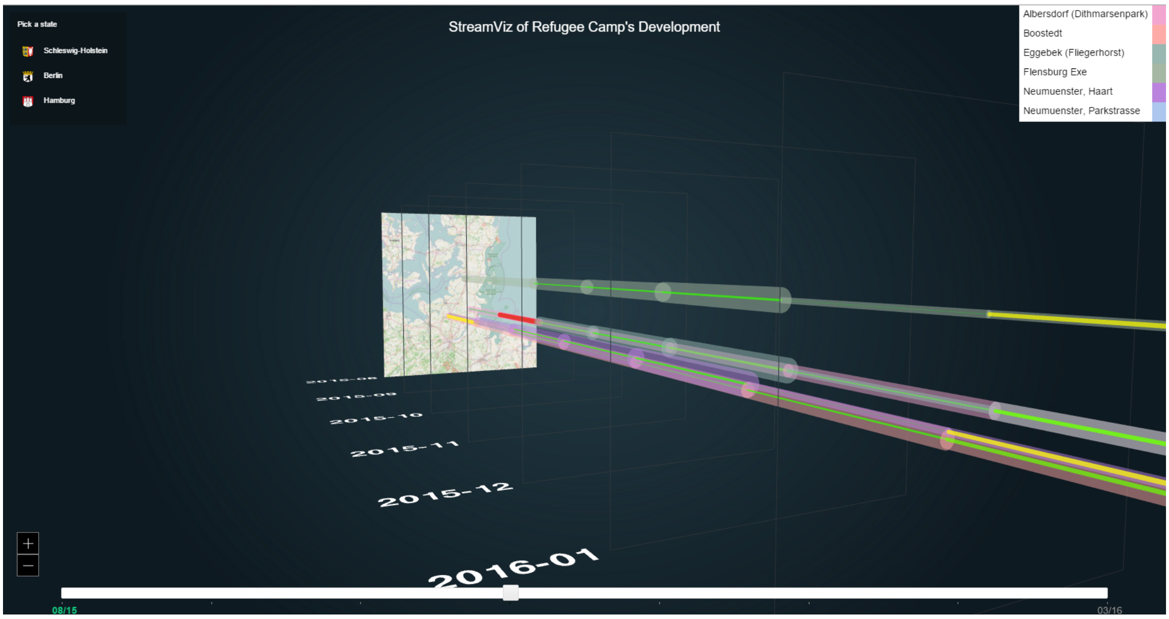


Figure 34: Early rotating view render of the StreamViz - refugee camp visualization – point-specific.

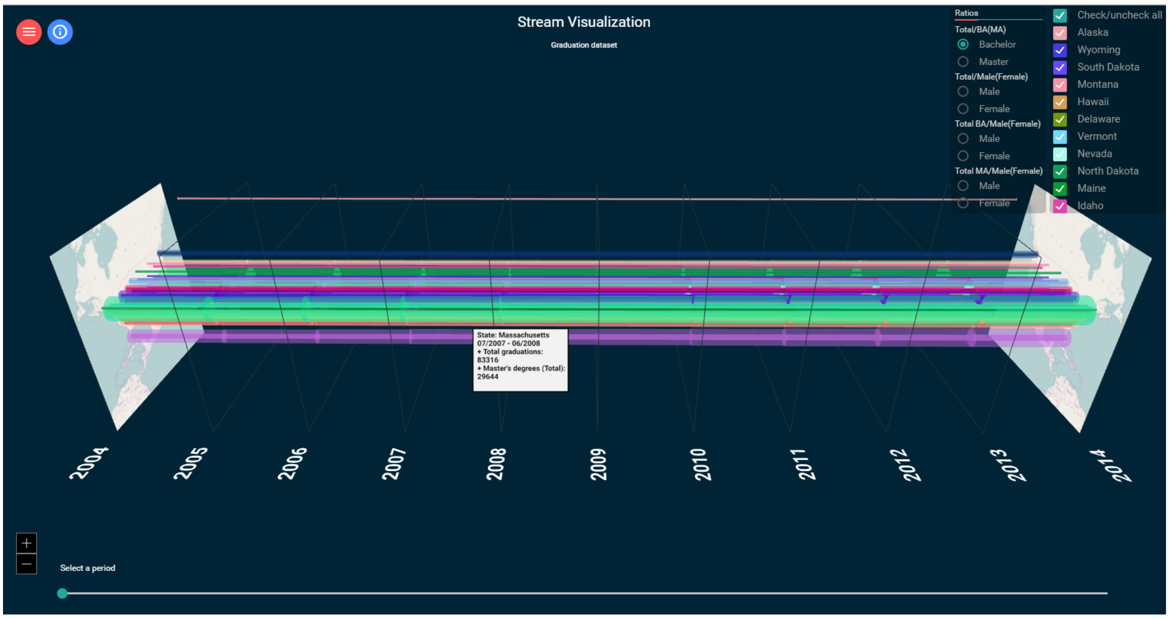
Looking at the Front View, the user will be presented with the U.S. map with nested circles representing different states and territories (Figure 36). The outer rings (and outer cylinders) show the divisor — total values of a ratio (e.g., total completions of BA and MA), the inner ones (and inner cylinders) indicate the dividend — selected variable (e.g., BA completions). The circles and cylinders are color-coded according to each state | territory.

6.4.1.3 Information-assisted visualization

A recent assortment of visualization techniques allows visualizing complex features in data by relying on information abstracted from the data, which coined the term “*information-assisted visualization*” (Chen et al., 2009). In this setup, the user is provided with a second visualization pipeline, which typically presents information about the input data sets among other useful information. The process is illustrated in Figure 37. These techniques might be helpful in providing a bridge between scientific- and information visualization, with increasing data set size and complexity.

The visualization provides the user with an option to browse through all raw data sets. There are a total of ten data sets corresponding to the ten year period and a dictionary data set with directory information about all education institutions. (Figure 38)

For each state, in each year, the visualization also provides additional information on the



Total/BA(MA)
● Bachelor
○ Master
Total/Male(Female)
○ Male
○ Female
Total BA/Male(Female)
○ Male
○ Female
Total MA/Male(Female)
○ Male
○ Female

☑ Check/uncheck all
☑ Alaska
☑ Wyoming
☑ South Dakota
☑ Montana
☑ Hawaii
☑ Delaware
☑ Vermont
☑ Nevada
☑ North Dakota
☑ Maine
☑ Idaho

State: Massachusetts
07/2007 - 06/2008
+ Total graduations:
83316
+ Master's degrees (Total):
29644

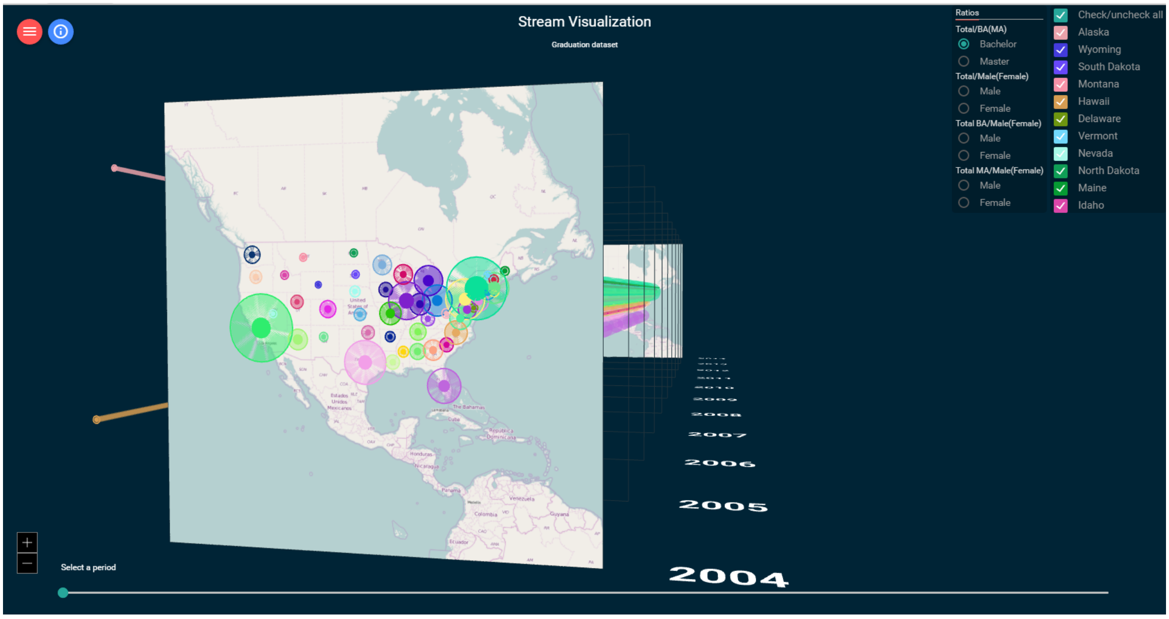
Figure 35: The StreamViz demo – IPEDS education data visualization, Side View.

data sets through various chart combinations, each accompanied by raw data tables for better checkup and validation. This is akin to a drill-down interaction with the user being able to drill down to a 1-year window, on an individual institution level, for a particular state, in a particular year. Example use cases include: view total BA | MA | male | female completions of the state New York in the year 2012, or view total completions of each individual education institution in the state of Massachusetts from 2004 through 2014, etc. (Figure 39)

In this demo, two separated data sets are used. The first one is the completion data set that reflects graduation / degree completion rate among education institutions, the other reflects the amount of enrollments for different academic degrees. The main menu allows switching between these two data sets as well as viewing raw data. (Figure 40)

6.4.2 Demo B: Visualization of U.S. IPEDS Graduation Data Set (3D in VR)

Using only open data sets from the Completion survey of the IPEDS (2004 - 2014), in this second demo of the StreamViz, Virtual Reality technology is used to give the user an immersive experience and to facilitate storytelling. The StreamViz VR makes use of the new, experimental WebVR JavaScript API in combination with Three.js' VRControls to provide access to Virtual Reality devices (e.g., Google Cardboard). The main view of the visualization is a *3D stacked bar chart* that shows the completions of Bachelor's and Master's degree of 50 U.S. states and territories. The bottom, solid color bars represent completions of Master's degree while the translucent bars stacked above represent completions of Bachelor's



Ratios

Total/BA(MA)
● Bachelor
○ Master

Total/Male(Female)
○ Male
○ Female

Total BA/Male(Female)
○ Male
○ Female

Total MA/Male(Female)
○ Male
○ Female

☑ Check/uncheck all
☑ Alaska
☑ Wyoming
☑ South Dakota
☑ Montana
☑ Hawaii
☑ Delaware
☑ Vermont
☑ Nevada
☑ North Dakota
☑ Maine
☑ Idaho

Figure 36: The StreamViz demo – IPEDS education data visualization, Front View.

degree. An info-window on the top right corner of the user's field of view provides assistive information such as state name, type of value and the actual value of the bar that the user sets their focus on. (Figure 41)

By focusing on any bar of the main bar chart for a specific amount of time, the user will be taken to a second visualization (the Stream) that shows the completion ratio between males and females of the particular degree (BA | MA) in the particular state associated with the selected bar for the same time window (2004 - 2014). Using the same settings as the previous StreamViz demo, the outer cylinder represents value for Female completions while the inner cylinder shows value for Male completions. The user can see instantly, for example in Figure 42, that the number of Female completions of Bachelor's degree in the state of New York is far greater than that of Male by comparing the diameter of the cylinders and also can read concrete values from the info-window on the top right corner. The color of the outer cylinder corresponds to the color of the selected bar and indicates the selected state. The color of the inner cylinder is automatically calculated to be a complement color of the outer cylinder's color. The cylinders and thus the timeline rise upwards (in ascending fashion over the user's head) in the initial animation.

The menu allows the user to slowly ride along the cylinder to the top while exploring the data the cylinders reflect, and pause at a few important milestones and get assisted with some narrations that offer a quick digest of the data up to that point (option *Normal Ride*). Or the user can choose to quickly ascend to the top (the end of the timeline) without having to stop along the way and read the narrations (option *Quick Ride to Top*). The option *Show All States*

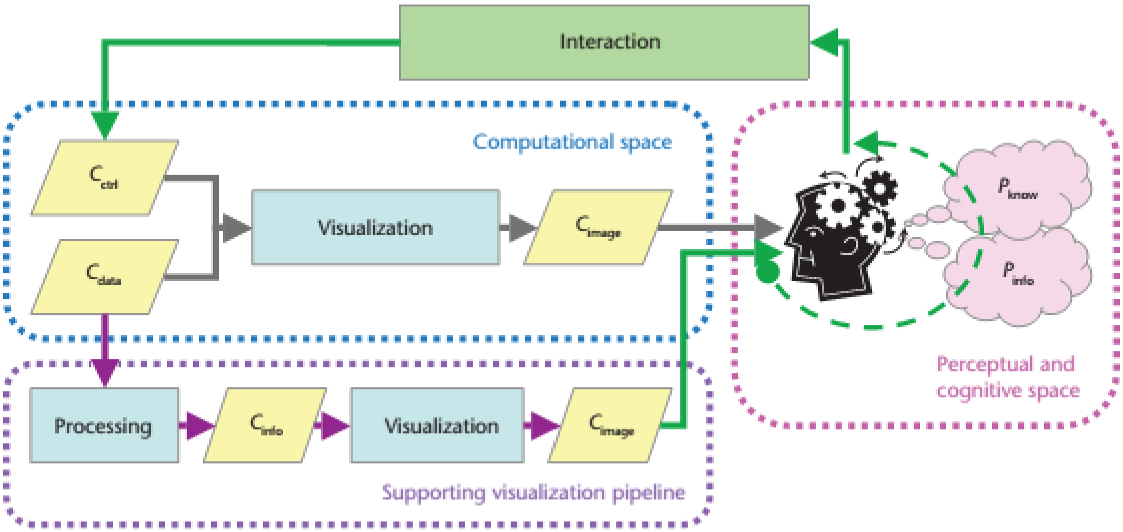


Figure 37: Typical flow of information-assisted visualization, with an additional pipeline displays information about the input data. Source: Chen et al. (2009)

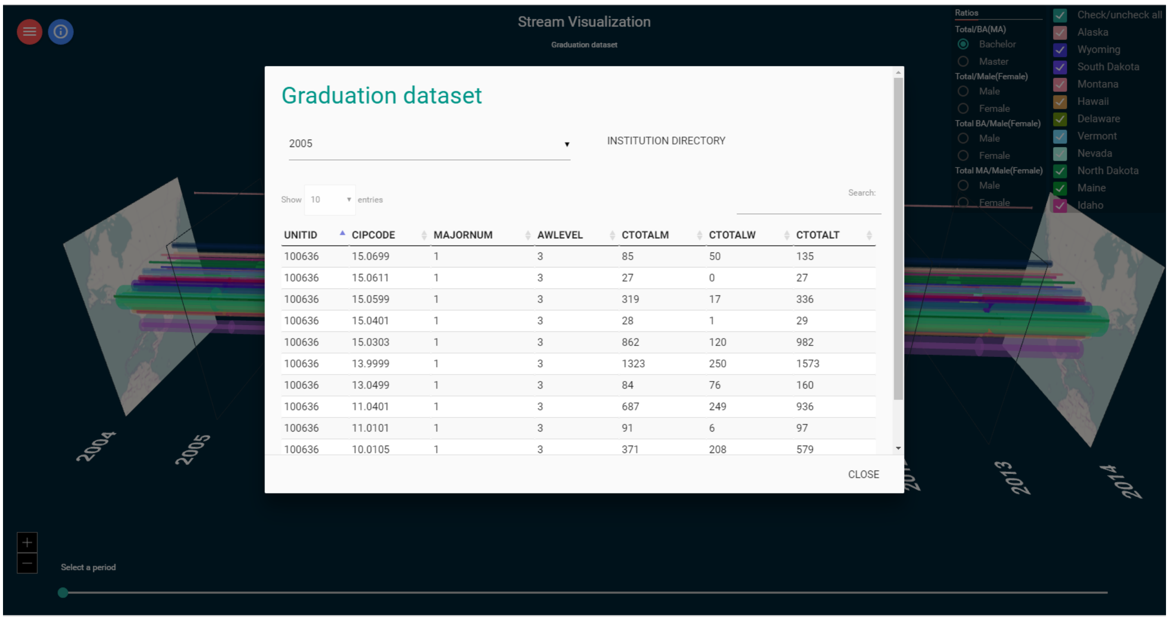
allows the user to view all the cylinders (representing all states) at once. Lastly, the user can go back to the main bar chart at any time. (Figure 43)

If the option Normal Ride is selected, the platform slowly rises to take the user to the top, stopping at a few important milestones along the way to present the user with narrations that summarize the data up to that point in a few main ideas. (Figure 44)

6.4.3 Challenges

As with any visualization projects, challenges during the design consideration and implementation phase are inevitable. There are difficulties working in a 3D environment. Unlike in traditional 2D-plane, the positioning and distribution of objects in 3D space is significantly more complicated. In the case of refugee camp visualization demo, if the number of camps is large and there are no big differences in their geographical distribution then the visualization must be broken down into several views, otherwise object overlapping issue will cause a huge negative impact on usability and user experience. For instance, a country view should only show aggregation (e.g., aggregation of camps) for its next smaller administrative division (e.g., states), and a state should show aggregation for its cities, and so on. This approach also promotes user interaction.

According to design, the 3D cylinders are nested into each other to emphasize the proportion between the planned capacity of a camp and the actual occupation. In order to maintain an acceptable level of usability, transparency must be employed to allow better view through



Graduation dataset

2005 ▾ INSTITUTION DIRECTORY

Show 10 ▾ entries Search:

UNITID	CIPCODE	MAJORNUM	AWLEVEL	CTOTALM	CTOTALW	CTOTALT
100636	15.0699	1	3	85	50	135
100636	15.0611	1	3	27	0	27
100636	15.0599	1	3	319	17	336
100636	15.0401	1	3	28	1	29
100636	15.0303	1	3	862	120	982
100636	13.9999	1	3	1323	250	1573
100636	13.0499	1	3	84	76	160
100636	11.0401	1	3	687	249	936
100636	11.0101	1	3	91	6	97
100636	10.0105	1	3	371	208	579

CLOSE

Figure 38: The StreamViz demo – IPEDS education data visualization, raw data browser.

these cylinders. However, overuse of transparency is known to have caused different problems, one of which is non-deterministic rendering, when being applied to multiple nested objects.

Another challenge is the placement of the camps onto a 2D map geographically accurate. With the geographical coordinates (latitude, longitude) of the camps known, the translation into pixel-coordinates for a 2D display can be done easily. However, since the 3D world in Three.js implements a different coordinate system, the translation from geographical coordinates into 3D world units needs a different approach. This approach requires the geographical bounds of the map to be known, and a transformation coefficient — the ratio between the geographical coordinates and the 3D coordinates — be calculated.

The actual capacity and occupation of a camp must also be mapped to the radius of the cylinder because the real value could theoretically go up to as high as tens of thousands. The scaling parameters must be correctly determined so that the proportions between the mapping and mapped objects are retained and reflected accurately in the 3D scene. There are also usability considerations, one of which is whether to allow the user to freely explore the 3D space with the mouse via the *OrbitControls*, or to limit that ability and instead provide the user with a fixed set of on-screen controls, e.g., tilt, pan, zoom, rotation, etc. The latter option would allow for a more streamlined experience, whereas the former option allows for a more natural and immersive experience. If the user is allowed to change the camera view freely to explore the 3D scene, all objects in the visualization should also be able to adapt to camera changes to retain all important information within the user's field of view,

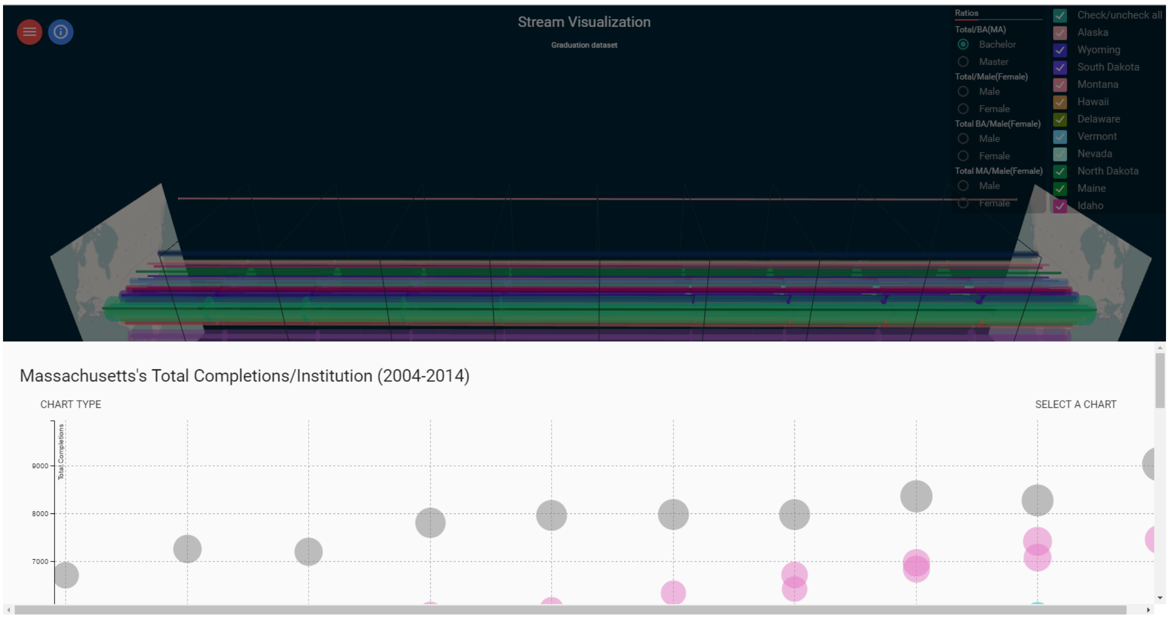


Figure 39: Additional chart combinations and raw data tables offer drill-down information for each individual state, in a particular year, for a particular education institution.

in such a manner that key information is not distorted. This proves to be far more labor-intensive to implement for a 3D world. To make navigation easier, a slider is provided to help the user move along the visualization's timeline. Performance is also a potential problem when the timeline or the number of objects in a 3D scene grows too large. Animations can help delivering a better user experience, although it requires considerably more effort to implement in a 3D environment, also has its own set of difficulties, and might impact usability negatively if overused. To keep the visualization dynamic, the maps are queried from the *OpenStreetMap* mapping service using geographical coordinates with a fixed zoom level, then transformed into a static image. These map images do not always fit the visualization's context, are user-friendly enough, and also contain a lot of unnecessary details. These could be a factor in reducing user experience. An approach to this problem is to customize the generated map images using libraries like *Mapbox*, but this would generate more overheads, besides the ones already coming from the dynamic queries for the camp's geographical coordinates based on their addresses.

Other use cases and demos of the StreamViz also pose similar challenges. In addition to the issues discussed above, there are also Virtual Reality-specific challenges. On traditional 2D platforms, the user can explore the environment using mouse and keyboard with interactions such as zooming, panning, dragging, etc. In VR environment, the user is limited to head movements. Although the Cardboard has a physical "switch" to act as an interaction mechanism, its utilization is greatly limited. Thus the first challenge is the need for a more flexible

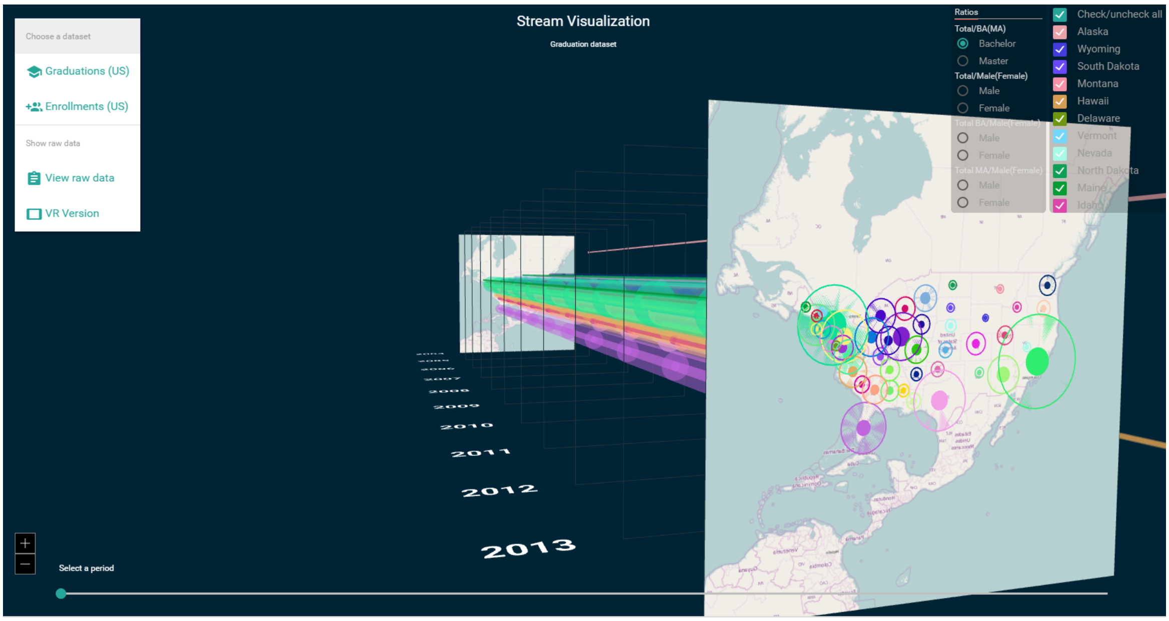


Figure 40: Main menu allows switching of data set and viewing raw data.

mechanism to facilitate user interaction with the VR world. To approach this challenge, the StreamViz enables the user to interact with the visualization by gazing at a specific element for a fixed amount of time. To achieve this, a pseudo mouse cursor is created and set at the center of the user's field of view. To select an action, the user only has to move the head to focus the "cursor" on a specific action | element until the success feedback registers (via sound and visuals). The main challenge however, is to adapt the StreamViz design to the Virtual Reality environment. Since the user is limited to head movements, they can only look around the environment, but are unable to move around. Therefore the visual design needs to be engaging to attract user's attention and animation is utilized to move the user around pre-defined paths should they choose to do so. A general risk that pertains to every VR environments that needs to be acknowledged is the VR sickness | cybersickness[32] that is potentially nausea, vomiting, and drowsiness-inducing if the VR experience is not carefully designed. Therefore, the StreamViz keeps animations to only what are essential and also positions most of the main visual elements directly in front of the user in order to limit the cause of this sickness.

The nature of time series data make it an excellent choice for storytelling. The StreamViz utilizes a basic idea: to allow the user to move along the timeline and view short narrations (e.g., data summarization) along the way. Storytelling in VR environment is much more engaging and immersive than on 2D platforms, and thus would be able to make up for the lack of rich user interactions that pertain to traditional environments.

[32] https://en.wikipedia.org/wiki/Virtual_reality_sickness

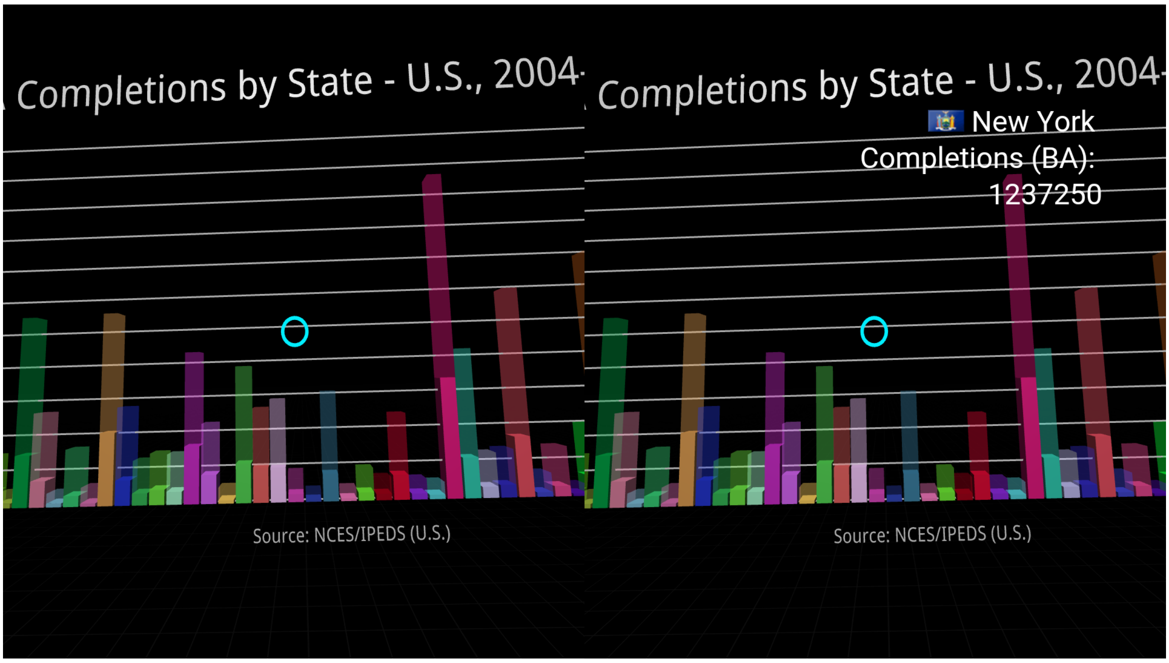

Completions by State - U.S., 2004
Completions by State - U.S., 2004
New York
Completions (BA):
1237250
Source: NCES/IPEDS (U.S.)
Source: NCES/IPEDS (U.S.)

Figure 41: The StreamViz VR – Main View.

6.4.4 Conclusions

The StreamViz is a visualization prototype which aims at multivariate data visualization. By utilizing the third dimension, more variables can be meaningfully encoded. The design process adheres to the 4x4 model for knowledge content to ensure its content is engaging to a broader audience with different levels of interest. To achieve this, it follows a workflow with a set of specific steps, from content researching to selecting, designing, and crafting visual elements. Virtual Reality is an emerging platform that allows for an excellent immersive user experience if carefully designed, thus make it a good medium for storytelling. The StreamViz attempts to further enhance user's engagement and knowledge transfer effectiveness by adopting the VR environment combining with basic storytelling. Each setting contain its own challenges and issues that need to be addressed to ensure an effective visualization. The VR environment has some downsides and limitations which either do not exist or could be easily addressed in traditional 2D platforms but might cause huge negative impact on user experience, including inducing cybersickness, in VR environment if not carefully designed.

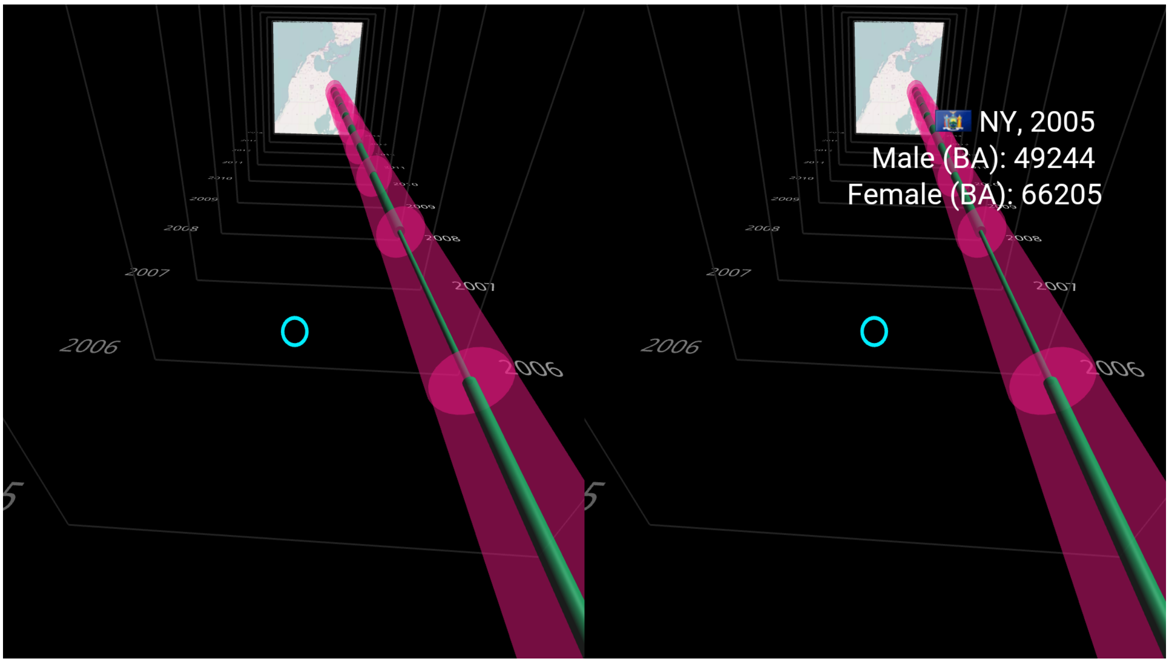


Figure 42: The StreamViz VR – the timeline, outer and inner cylinders.

6.5 Testing and Evaluation of Visualizations

6.5.1 Introduction

Two foundational problems of visualization mentioned in (Zhu, 2007) are the definition of the effectiveness of visualization and the method to measure it. Still according to Zhu (2007), the term "*effective visualization*" despite being used extensively in literature, does not have a consistent and standard definition. There are a few existing definitions of effectiveness of visualization and each has its own limitations. Some researchers follow the data-centric approach which suggests that effectiveness depends largely on the level of correspondence between the visualization and the data it represents. Tufte (1983) proposes that an effective visualization should maximize the data / ink ratio, i.e. be packed with as much data as possible. This view is challenged by Kosslyn (1985) for the reason that there is no empirical evidence that this approach leads to more accurate interpretations or better task efficiency. Other researchers take the task-centric view as guideline and state that a visualization must be designed with a specific task in mind and that it is considered effective if it improves task efficiency. Examples are (Casner, 1991), (Bertin, 1983), (Nowell et al., 2002), and (Amar and Stasko, 2005)

So far, results from studies in psychology and HCI seem to support the task-centric approach. Some psychological studies have even suggested that task complexity has certain impact on

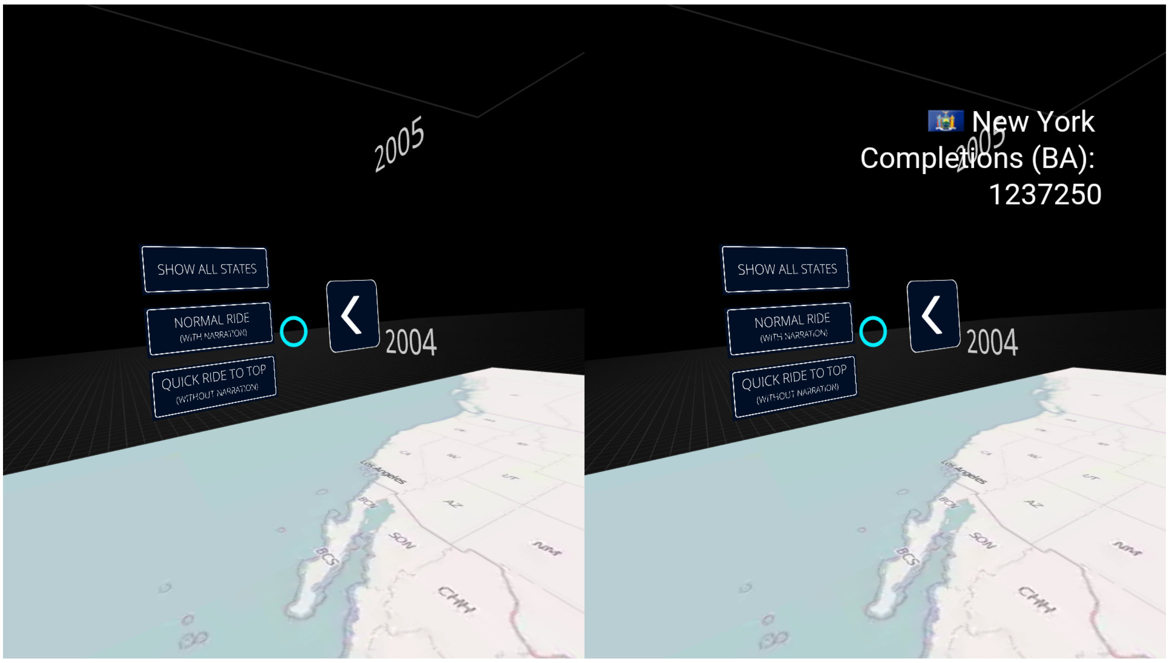

2005
SHOW ALL STATES
NORMAL RIDE (WITH NARRATION)
QUICK RIDE TO TOP (WITHOUT NARRATION)
2004
New York
Completions (BA):
1237250

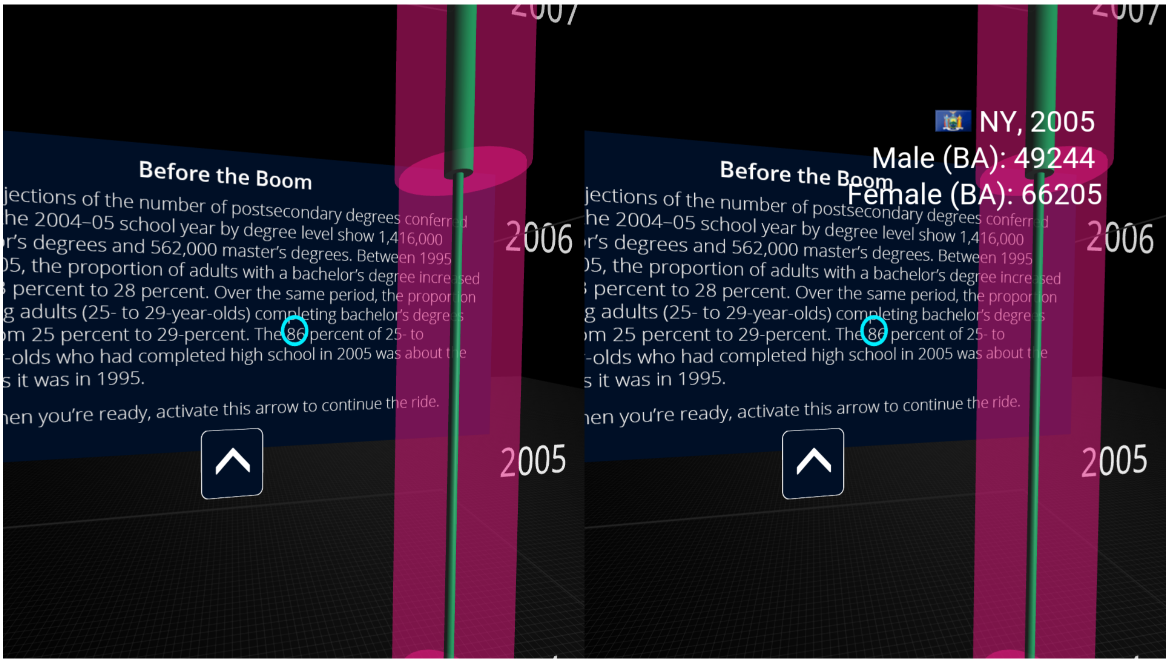
Figure 43: The StreamViz VR - Main menu.

the effectiveness of visualization, and that it also depends on the interaction between the external and internal representation of data (Scaife and Rogers, 1996). Tversky et al. (2006) defines the Principle of Congruence — structure and content of a visualization and of the desired mental representation should be in correspondence, and the Principle of Apprehension — structure and content of a visualization should be readily and accurately perceived and understood, but does not take into account the influence of task. Domain knowledge, memory capacity, experience with visualization techniques, and skills like explanatory and reasoning are also contributing factors to the user's perception of visualization effectiveness.

As for how to measure the effectiveness of a visualization, there are two main methods — heuristic evaluation and user studies, as stated in (Zhu, 2007). The heuristic method concerns with the evaluation of visualization designs by experts, based on a set of rules and principles, and has certain advantages as well as weaknesses. For the scope of this thesis, the StreamViz prototypes will be evaluated using the user-study approach.

6.5.2 Definition and measurements of an effective visualization

There are a few common measures of effectiveness in the user study approach, such as task completion time, error rate, user satisfaction | user experience, and task efficiency. Each has its own strengths and drawbacks, as discussed in (Zhu, 2007). In that paper, the effectiveness of visualizations is also defined in term of three principles:



Before the Boom

jections of the number of postsecondary degrees conferr d
he 2004–05 school year by degree level show 1,416,000
r's degrees and 562,000 master's degrees. Between 1995
)5, the proportion of adults with a bachelor's degree incre sed
percent to 28 percent. Over the same period, the propor on
g adults (25- to 29-year-olds) completing bachelor's degre s
m 25 percent to 29-percent. The 86 percent of 25- to
-olds who had completed high school in 2005 was about t e
s it was in 1995.

hen you're ready, activate this arrow to continue the ride.

NY, 2005
Male (BA): 49244
Female (BA): 66205

Figure 44: The StreamViz VR - Narrations.

1. Principle of accuracy: an effective visualization is one whose attributes of visual elements must match the attributes of data items, and its structure must also match that of the data set. This principle defines the relationship between the data and its visualization. Proposed steps to measure accuracy include definition of a taxonomy of visualization techniques and domain specific data analysis.
2. Principle of utility: an effective visualization is one that enables users to achieve the goal of specific tasks. This principle defines the relationship between the visualization and its tasks. Among the methods of utility measurement mentioned in the paper are domain task analysis and establishment of an annotated benchmark database.
3. Principle of efficiency: an effective visualization is one that reduces cognitive load for a specific task — more effective in delivering knowledge, over non-visual representations. This principle defines the relationship between the visualization and its users. Among the proposed steps to measure efficiency are task completion time benchmarking, eye movement tracking and querying for user's subjective opinions.

Qualitative and quantitative measures of effectiveness still according to (Zhu, 2007) are summarized in Table 3

Merčun (2014) points out three types of evaluation that are common to empirical | user studies — evaluating user performance, user experience and the visualization algorithm itself. That paper also gives an in-depth discussion about the important role user experience (UX)

	Quantitative measurements	Qualitative measurements
Accuracy	Number of interpretation errors.	Interview, observation, expert | novice comparison.
Utility	Number of achieved benchmark goals. Frequency of usage of the visualization by users to conduct a task.	Interview, observation, expert | novice comparison.
Efficiency	Task completion time. Eye movements. Learning curve.	Visualization complexity analysis, interview, observation, expert | novice comparison.

Table 3: Quantitative and qualitative measures of effectiveness. Source: Zhu (2007)

plays in the success of a system | service and different constructs with which UX is measured, for example flow, aesthetics, emotion, etc., and raises the question of how to measure these constructs as well as how to define UX dimensions.

6.5.3 The reaction card method

Developed at Microsoft in 2002 to address the limitations of standard feedback mechanisms such as Likert scale, the product reaction card method, a.k.a. desirability toolkit, provides "a way for users to tell the story of their experience, choosing the words that have meaning to them as triggers to express their feelings – negative or positive – about their experience" (Benedek and Miner, 2002). It can be used as a baseline for comparison or along with other instruments to tap into UX.

The reaction card method is typically deployed at the end of a testing session and involves participants picking out from a set of cards, which contains various adjectives, those that most accurately reflect their experience with the product | system and when necessary, also commenting on their choices. The original paper defines a set of 118 cards in total but only a limited subset of it is typically used in various studies (Merčun, 2014). Barnum and Palmer (2010) reported a few metrics used to analyze the results of the reaction card method in their experiments such as the number of positive cards, the ratio between chosen positive and negative cards, mapping and clustering the selections, etc.

6.5.4 Study goals and design

The two variants — 3D-on-2D and Virtual Reality, of the StreamViz will be compared and evaluated in term of UX | usability using the reaction card method in order to examine user's perceptions and experience with individual variants of the visualization. The questionnaire mainly focuses on the VR variant and identifies a set of 5-6 visualization scenarios. Each scenario represents a task | interaction and requires interactive exploration to answer the questions. In addition to age range and gender, the users were also asked to specify their level of experience with the VR environment which consists of the following choices: novice, beginner, competent, proficient and expert. Each of the six study participants will work on the questionnaire individually in separate sessions.

The reaction card method is modified using only a limited set of cards, which consists of 29 out of the original 118 adjectives (14 positive and 15 negative – Table 4), which aim to describe interactions with visual designs. After completing the specified task in a scenario in a given prototype variant, the participant will be asked to choose any number of reaction cards that are listed in the questionnaire in a random order. At the end of each scenario, the participant is given an opportunity to provide qualitative feedback on that scenario as well as to elaborate more on the design variant they tested.

time consuming	unfriendly	useless
frustrating	clumsy	hard to use
quick to understand	inefficient	difficult to understand
informative	unappealing	efficient
innovative	organized	useful
complex	deficient	appealing
easy to use	transparent	fun
advanced	convenient	interesting
illogical	attractive	intimidating
opaque	logical	

Table 4: The set of reaction cards used in the study.

Merčun (2014) states that different designs elicits different kinds of UX, i.e. while some designs might be more prevalent in term of usefulness and organization, others are more popular in term of appeal and interest. Therefore, the reaction cards can be categorized into five dimensions, as suggested in the paper. The proposed dimensions include: perceived ease of use, perceived usefulness, perceived efficiency, appeal, and engagement. Thus, the 29 cards used in this study can be grouped as follow:

- Perceived ease of use: frustrating, complex, hard to use, easy to use, unfriendly, intuitive.
- Perceived usefulness: useless, useful, informative, deficient.
- Perceived efficiency: quick to understand, difficult to understand, inefficient, efficient, time consuming, logical, illogical.

- Appeal: appealing, unappealing, clumsy, convenient, fun, organized, attractive.
- Engagement: interesting, innovative, opaque, advanced, transparent, intimidating.

Of the six scenarios | tasks in the questionnaire, five are related to the Virtual Reality variant. They are defined based on various interactions | use cases including:

1. Using the 3D stacked bar chart to find out information, for example, the number of bachelor's degree completions in the state of California from 2005-2014.
2. Drilling into a bar to transit to the 3D cylinder | stream at the beginning of the timeline.
3. Taking a normal ride (with narrations) to the end of the timeline while trying to find out information, for example the number of male | female bachelor's degree graduations of California in 2009.
4. Taking a quick ride (without narrations) to the end of the timeline.
5. Switching between showing the 3D cylinders | streams for all 50 states and only current selected state, and going back to the beginning.
6. Overall impression with the visualization in VR as a whole.
7. Playing around with the 3D-on-2D variant and comparing the VR variant to it.

For each of those scenarios, the user will be asked to carry out the task first, then express their impressions based on the reaction cards organized into 5 dimensions, then provided with a chance to comment | give feedback in words. An excerpt from the survey questionnaire is shown in Figure 53.

In addition to measuring the UX based on reaction cards, the 3D-on-2D and VR variant prototypes will also be tested for rendering performance on desktop PC and mobile devices. To stress-test rendering performance, the 3D-on-2D prototype is loaded with a thousand-month long timeline in order to force the WebGL engine to render a large number of 3D objects. The latest versions of Chrome (52.0.2743.116) and Firefox (47.0.1) browser at the time this thesis is written, running on a mid-range desktop PC, are utilized. The VR variant will be stress-tested in the same manner, running on the latest version of Chrome (52.0.2743.98) mobile browser on a Google Nexus 5 smartphone by rendering the 3D cylinders | streams for all data.

6.5.5 Assessing and analyzing results

Performance test's results for the 3D-on-2D variant shows some minor lag issues while moving along the timeline. Delays on rendering time are negligible.

The overall performance of the VR variant was smooth except some minor lags in the cursor animation and potential of an overheat issue if used for an extended period of time. Performance took a considerable negative impact when the mobile browser had to render a large number of 3D objects and the device was overheated quickly. For more capable devices however, performance is expected to vastly improve. The VR variant performed and rendered best on the Google Chrome browser, but encountered various rendering issues on other browsers such as mobile Safari, due to the fact that the WebVR standard being still in early stages, and thus not uniformly supported cross-browsers.

To assess the results of the UX test, we shall first begin with the analysis of individual adjectives | cards and the frequency with which they have been selected. The sample pool consists of 66.7% of the 18-29, and 33.3% of the 30-39 age group, and is divided 50/50 between the novice and beginner level of familiarity to the Virtual Reality environment. To get a quick overview of the impressions at first glance, the frequencies of the selected cards are visualized using word clouds to present the quantitative outcome, as shown in Figure 49 and Figure 50. The colors of the words reveal that the initial impressions of the 3D tube of the VR variant and the guided tour along the timeline with narrations received more negative responses than other parts of the visualization, and the size of the word tell us that the 3D stacked bar chart, the 3D tube, the slow timeline tour and the overall VR visualization were seen as interesting, informative, logical, and useful. The 3D bar chart was appraised as quick to understand, the showing of all the tubes at once was seen as complex and clumsy, while the VR variant was assessed as more innovative and intuitive than the 3D-on-2D counterpart. It is also easy to recognize that the quick tour without narrations was appraised with the most positive impressions.

While word clouds provide a quick overview and enable rapid assessment, they fail to communicate results in a clear manner that would allow for a more in-depth comparison. Classical charts would enable for a more precise observation of the differences between selected cards within an individual visualization or in comparison to another variant. Figure 45 shows the selection of impression cards for the VR variant in comparison to the 3D-on-2D variant. The chart tells us that participants in the study recognized the VR variant as significantly more fun, innovative and intuitive, besides being easier to use and more logical. On the other hand, it is also possible to see that more than half of the participants also thinks it is somewhat more time-consuming, while a few opinions show it is more clumsy and somewhat useless.

Although the assessment of the selection of individual cards is the more common approach and in most cases, sufficient for result analysis, some other types of exploration can also be quite useful and informative. For example the overall number of selected cards for individual tasks can be an indicator of the user's attitude toward that part of the visualization and also how engaging it is. The figures in Table 5 make it clear that the VR visualization is most engaging as a whole, while the participants had the most impressions in Task 1 and 3.

The ratio of positive and negative cards as shown in Table 6 even better reflects the overall results of the UX test as Task 1, 3, 4 and 6 gathered the highest rates of positive impressions, while Task 2 and 5 had a significantly higher percentage of negative cards. The fact that the

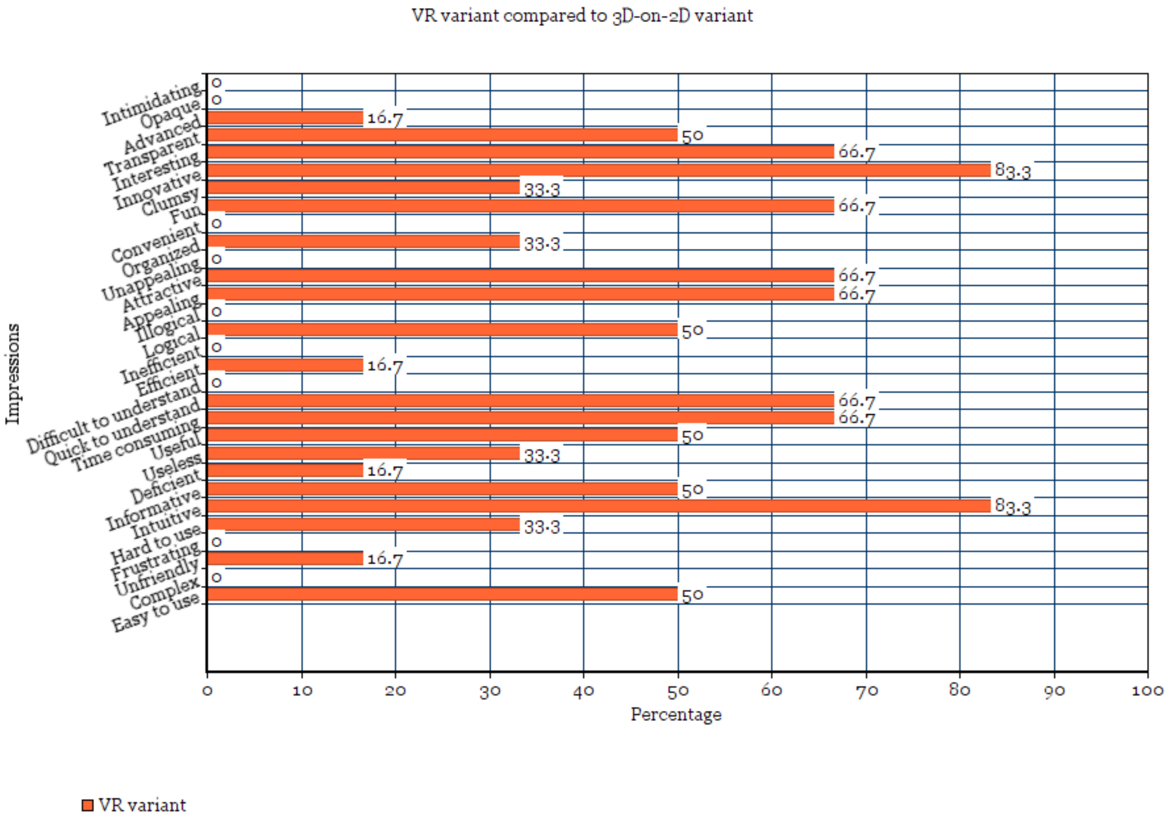


Figure 45: Reaction cards selected for the VR variant in comparison to the 3D-on-2D variant.

overall impression task received over 80% of positive cards therefore tells us that the VR prototype of the visualization design evoked also positive UX.

We also need to assess the results in term of the five dimensions into which the reaction cards are categorized. The radar chart provides the most suitable visualization tool for this type of analysis. For graphical presentation, positive and negative impressions are separated a and to simplify the analysis, it is sufficient to only visualize positive aspects, as negative cards basically form a mirror image of their positive counterparts. Figure 51 maps the values collected from the reaction cards to the five dimensions. This presentation provides us with a higher overview of the UX test results, for example, it can be observed that Task 5 (showing all 3D tubes) scored relatively low on ease of use and usefulness, but better in term of engagement and appeal. This chart also gives us a good idea of which parts of the visualization prototype scored an overall lower UX, based on the values achieved on individual dimensions.

Figure 52 reveals another interesting aspect in term of the shape of the graph. The two charts show that while Task 1, 2, 5 and the overall visualization created a more balanced

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
66	57	65	58	54	73

Table 5: Total number of selected cards per task.

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
+	88%	75.44%	86.15%	100%	68.52%	84.93%
-	12%	24.56%	13.85%	0%	31.48%	15.07%

Table 6: Ratio of selected positive and negative cards.

shape and thus hint at a more all-round UX, Task 3 and 4 formed a skewed shape which indicates that in those tasks, the five UX dimensions are unevenly supported.

Lastly, the radar charts also enable the calculation of their surface areas, the result of which gives a mathematical expression to the overall UX. This method is defined in (Mosley and Mayer, 1999) and known as Surface Measure of Overall Performance (SMOP) and typically used in benchmarking analysis. Using the formula:

$$((P_1 * P_2) + (P_2 * P_3) + (P_3 * P_4) + \cdots + (P_n * P_1)) * \sin(360/n)/2$$

where P is the data point on the axis of the radar chart and n is the number of axes, the surface area of the radar graphs can be calculated and shown in Table 7. The results of the surface area calculation clearly indicate that the parts of the visualization corresponding to Task 1 and Task 4, by this measure, provide the best UX.

Task 1	Task 2	Task 3	Task 4	Task 5
318.6	175.5	299.6	318.6	132.2

Table 7: Surface area of radar graphs.

6.5.6 Discussion and conclusions

By using the reaction cards method, user experience can be more effectively analyzed and assessed. By mapping the cards to different dimensions, even more value is added to this method which potentially enables a statistical comparison of visualization designs by individual dimensions. More research and consideration, however, is needed to correctly identify the necessary sets of adjectives and dimensions to be used.

While the reaction cards method can be used just as a tool to gain a deeper understanding of how the user feels about a visualization and why, it is possible to use more quantitative analysis to get more tangible results regarding UX, thus might lead us to a better understanding of UX in the context of information visualization.

Although three-dimensional visualization, especially in combination with a Virtual Reality environment, allows for a more immersive and engaging user experience for data exploration, it requires capable hardware and up-to-date technologies in order to achieve a decent and desired performance. Basically, there is a need for the improvement of multi-core GPU processors as well as upgrading the address bus throughput between CPU and GPU or even replacement for wireless transfer computations on cluster system. Currently, with the Google Cardboard VR viewer, the experience is mostly limited to moving the view around the final image and thus navigation inside the model seems to be another influential issue. For a richer and more immersive user experience with more sophisticated interactions, high-end (thus more expensive) hardware, such as the HTC Vive or Oculus Rift are required. Scaling is another one of the main challenges from the Big Data visualization point of view, which is relevant in multidimensional systems where there is a need to delve into a piece of information in order to obtain some specific value or knowledge. Unfortunately, it is very difficult to approach this issue from a static point of view. Likewise, integration with motion detection hardware (e.g., wearables) would highly increase user experience and usability for such visualization system. It might be a key to the interaction with visualized data in a more native way.

In general, the StreamViz VR prototype of the StreamViz managed to deliver a rich user experience by offering a high degree of immersion and a good level of attractiveness | engagement, not to mention the fun factor, for the participants. This fact is reflected in the written feedback with opinions stating that the VR environment was cool and the idea was great. Besides the positive comments, the written feedback also helps with identifying the aspects that need to be fixed | improved, together with the ones being already identified based on the result assessment from the reaction cards. These include:

- Changing the position of the infobox to fit better into the user's field of view.
- More specific information on the axes of the stacked bar chart.
- More information for the infobox (e.g., extrema, etc.).
- Possibility for the user to change the speed during the tour. Also possibility of speed changing (e.g., slowing down, small pauses, etc.) automatically when approaching places of interest (e.g., extrema, etc.) on the stream to catch user's attention.
- Possibility for "moving back and forth, a.k.a. zooming" interactions for better overview.
- Whole diagram should be visible on screen to keep an overview | track of the tour progress.
- The "showing all 3D tubes" functionality might be superfluous, and thus not necessary to include. Besides, it is clumsy and might cause frustration | performance degradation.
- For beginner, it might be difficult to get started without any instructions. Potentially more time consuming to extract information from the 3D variant than with conventional 2D visualizations. Learning curve for beginner might also be steeper.

- Presentation and readability of data could be improved. More possibilities for more types of comparison (e.g., education situation between states, within a state, etc.).
- Controls in the environment are still clumsy | not so efficient. Potential to cause nausea | headache.
- More questions about what kind of information should be visualized should be raised and studied so that the visualization would represent more meaningful information and provide answers to those questions.

While the “excitement” and “fun” factor are almost always present for the beginner | novice group of end users, other groups of users (proficient | expert), either in the field of VR or in the field of statistics, might find the experience less engaging or have more critics. Due to the limited scope of this thesis, the sample pool consists of mostly of participants from the first two groups. Therefore broader studies might be needed in the future to further identify the advantages | shortcomings of the visualization prototypes.

7 Epilogue

7.1 Summary

In this thesis, we were given an overview of the fundamentals as well as current developments of the Open Data ecology, and also a study of how new advancements in computer graphics and Virtual Reality technologies might help innovating the way Big Data visualization is perceived by the end users. It brought into discussion a variety of aspects and disciplines that all contribute to the big context (open data, big data analysis and information visualization) from a technical point of view. Starting with an introduction to the concept of open data, we learned about its provisioning, analysis, and the extraction of knowledge and insights out of it, in term of public sector data. The opening of various data sets starting at the governmental and public agency level, which had led to the concept of open government data, means more transparency and potentially high economic values. We were also informed of the current developments and implementations of different open data projects of different governments and agencies domestically and internationally, and had come to the conclusion that data openness has been coming ever closer to becoming a norm in the operations of public sectors today, despite the many existing challenges and issues.

In the following sections, we talked about the technical infrastructure as well as technologies needed to open our own data, making it available to the public and the importance of making the right choice of them (e.g., databases, platforms, etc.). The requirements for such platform were also vaguely outlined based on the examination of one of the most popular and widely used open data solutions (CKAN). We were also introduced to the concept of Linked Data

— which purpose is to facilitate better exposing, sharing, and connecting pieces of data and information — as well the closely related concept of Semantic Web. Next, we learned of how data mining processes are utilized to analyze and extract useful information out of data sets by first describing the details of a typical data mining workflow — concretely, the steps of the Knowledge Discovery in Databases (KDD) process. Then we also made a brief mention of a few other important mining processes and applications (e.g., web mining), as well as how visualization might fit in the whole process.

In the subsequent chapter, we shifted focus on data visualization, starting with a summarized history of it. Then we learned of some key design principles, terms and definitions, taxonomy, as well as a differentiation between confirmatory and exploratory visualization. In the following sections, we were introduced to various dynamic visualization techniques to facilitate user interaction, as well as a few common methods and algorithms used to visualize multi-variate data, such as the TreeMap method and Trellis displays. We concluded the chapter with a close examination of a multi-variate data visualization tool called V-Miner, including its functionalities and a typical use case.

We then started looking into the question of how recent advancements in web- and Virtual Reality technologies might help innovating the way multi-variate data visualization is done in the semi-final chapter. To answer this question concretely, we chose to design a new visualization prototype aiming at visualizing time-series data that we called the StreamViz., and made a few demos based on open education data and 3D web technologies in Virtual Reality environment. In the subsequent sections, we reviewed the original concepts and ideas behind the StreamViz's designs, as well as introduced a few early implementation attempts. We also detailed the workflow taken when designing the StreamViz, which consists of various steps starting from identifying the visualization model to picking the right technologies. We then went into the details of the design and implementation of two variants of the StreamViz (i.e., 3D-on-2D and VR) and explained how we approached the assessment and testing of these new visualization prototypes in the following section. We took a short review into current researches regarding evaluation methods for information visualizations, then described the approach we adopted for the survey in this thesis (i.e., in the form of a mini user survey and functional testings). Concretely, we described how the user study was designed to evaluate UX and based on what method. Finally, we provided a detailed analysis of the results from the assessment.

To summarize, as with today's computing power and rapid advancements in VR and AR, its is possible to design new types of information visualizations that would take user experience to a whole new level. New types of VR/AR-based experience definitely gives the user a great level of fun and excitement, since those technologies are still relatively new and have yet to become mainstream and common. But beyond the initial "wow" factor, it is how the user experience is designed that decides how well the user is willling to embrace this new form of visualization. If we aim for the "wow" factor, we might achieve short term success but would ruin the whole system in the long run, because while everyone can develop for AR/VR, it is difficult to design a good user experience. Currently, Virtual Reality is often used as a new medium for storytelling to immerse the user into the story. This might be perceived as a huge leap for the field of data journalism, but as for corporate and enterprise environments, conventional 2D visualizations still dominate since VR is still perceived by some as nothing more than a gimmick and that it does not really bring any real benefits and advantages. Based on assessment results and user feedback, although the VR variant of the StreamViz has more potential and a greater fun factor than the 3D-over-2D counterpart, it is still to a certain extent confusing to the perception of the end users and greatly limited in term of user interaction due to the limitations of the Cardboard platform.

7.2 Future Work

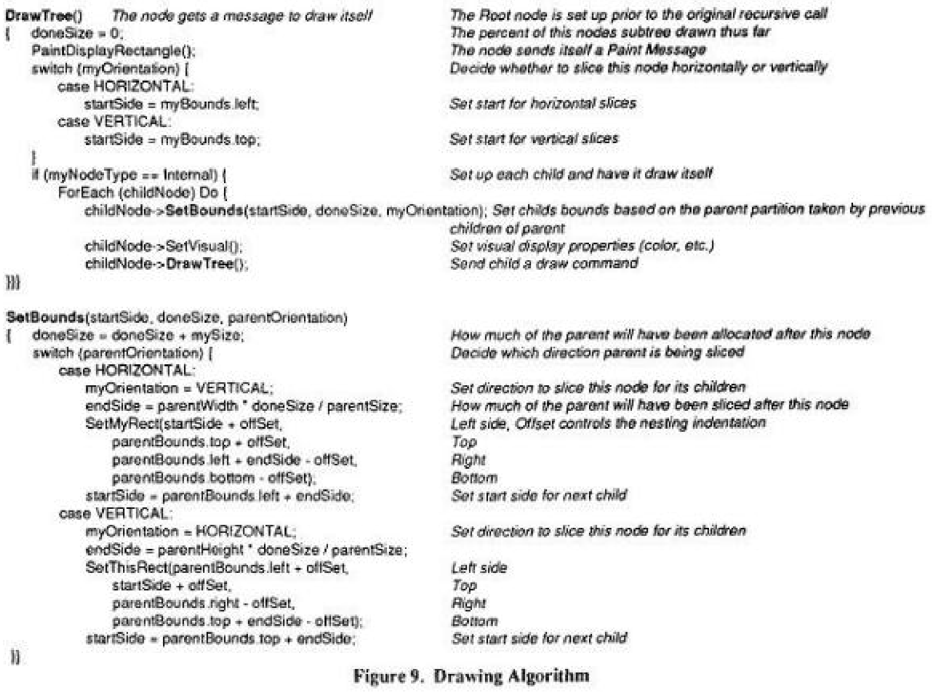
This thesis sets focus on the study of feasibility, efficiency and usability of a new visualization prototype which is based on 3D web technologies and Virtual Reality. It uses open data as part of the visualization and as such, also dedicates the first few chapters to the study of the context and technologies revolving around Open Data. Nevertheless, due to the limited scope, only open government data was observed and actually used in all the implementations. At this point, it is possible to further delve into the Open Data ecology and study how data from the private sectors might be made open as well as its relationship to the public sector data and the impact it has on various fields | the economy | the society, etc. It is also possible to study the Open Data topic with special regard to the field of data journalism and how modern technologies (e.g., web | database technologies, computer graphics, etc.) might be integrated and facilitated to provide better tool sets for this field.

Another possibility for further studies lies in the topic of data mining | KDD. It might be worthwhile to study how effective different mining processes can be applied to various forms of open data (from both sectors) to extract valuable insights and knowledge. For those who have interest in the topic of information visualization, ideas for new forms of visualization tailored to various types of data can be further studied, designed, and implemented. Or existing visualization techniques (e.g., the ones mentioned in this thesis) can be further studied to improve and extend their usage scope and efficiency.

Lastly, for those who have interest in VR/AR and/or the StreamViz, there exists a wide range of possibilities to either further improve the StreamViz in term of UX and/or functionality based on the assessment results and user feedback presented in this thesis, or in term of HCI, study how data visualization in VR/AR environment can be best designed to achieve higher level of user engagement and interaction, possibly using higher-end hardware (e.g., the Vive or Oculus Rift), and how to best integrate VR/AR into the context of visualizing corporate | enterprise data.

A Appendix

A.1 The treemap drawing and tracking algorithm

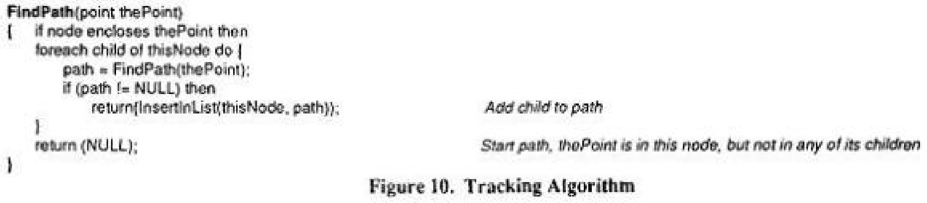


```
DrawTree()      The node gets a message to draw itself         The Root node is set up prior to the original recursive call
{   doneSize = 0;                                              The percent of this nodes subtree drawn thus far
    PaintDisplayRectangle();                                   The node sends itself a Paint Message
    switch (myOrientation) {                                   Decide whether to slice this node horizontally or vertically
        case HORIZONTAL:
            startSide = myBounds.left;                         Set start for horizontal slices
        case VERTICAL:
            startSide = myBounds.top;                          Set start for vertical slices
    }
    if (myNodeType == Internal) {                              Set up each child and have it draw itself
        ForEach (childNode) Do {
            childNode->SetBounds(startSide, doneSize, myOrientation); Set childs bounds based on the parent partition taken by previous
                                                               children of parent
            childNode->SetVisual();                            Set visual display properties (color, etc.)
            childNode->DrawTree();                             Send child a draw command
}}}

SetBounds(startSide, doneSize, parentOrientation)
{   doneSize = doneSize + mySize;                              How much of the parent will have been allocated after this node
    switch (parentOrientation) {                               Decide which direction parent is being sliced
        case HORIZONTAL:
            myOrientation = VERTICAL;                          Set direction to slice this node for its children
            endSide = parentWidth * doneSize / parentSize;     How much of the parent will have been sliced after this node
            SetMyRect(startSide + offSet,                      Left side, Offset controls the nesting indentation
                parentBounds.top + offSet,                     Top
                parentBounds.left + endSide - offSet,          Right
                parentBounds.bottom - offSet);                 Bottom
            startSide = parentBounds.left + endSide;           Set start side for next child
        case VERTICAL:
            myOrientation = HORIZONTAL;                        Set direction to slice this node for its children
            endSide = parentHeight * doneSize / parentSize;
            SetThisRect(parentBounds.left + offSet,            Left side
                startSide + offSet,                            Top
                parentBounds.right - offSet,                   Right
                parentBounds.top + endSide - offSet);          Bottom
            startSide = parentBounds.top + endSide;            Set start side for next child
}}
```

Figure 9. Drawing Algorithm

Figure 46: The treemap drawing algorithm. Source: Shneiderman (1991)



```
FindPath(point thePoint)
{   if node encloses thePoint then
    foreach child of thisNode do {
        path = FindPath(thePoint);
        if (path != NULL) then
            return(InsertInList(thisNode, path));          Add child to path
    }
    return (NULL);                                         Start path, thePoint is in this node, but not in any of its children
}
```

Figure 10. Tracking Algorithm

Figure 47: The treemap tracking algorithm. Source: Shneiderman (1991)

A.2 The StreamViz - initial concept with e-commerce data

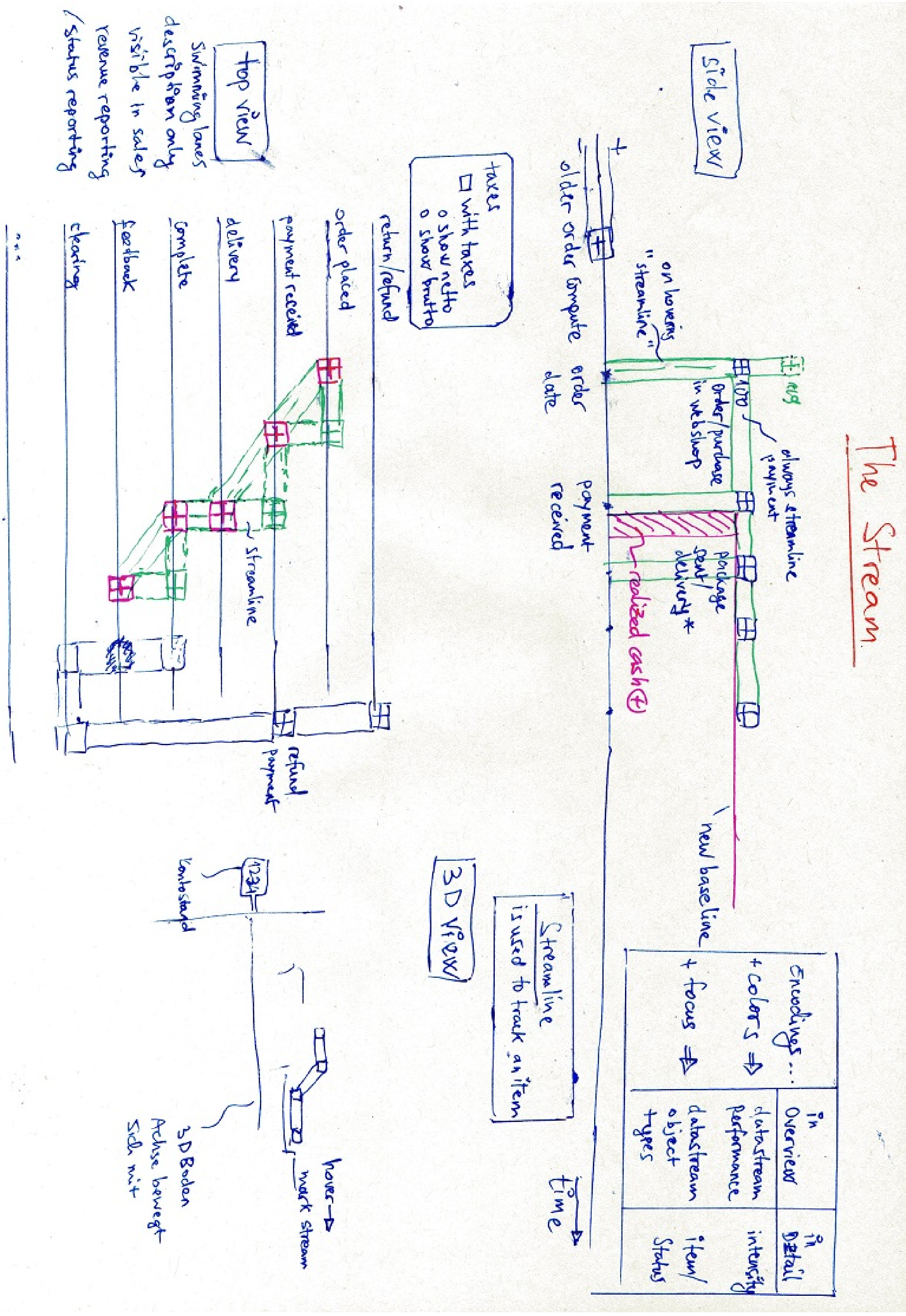


Figure 48: The StreamViz - initial concept with e-commerce data.

A.3 StreamViz evaluation - word clouds for individual tasks

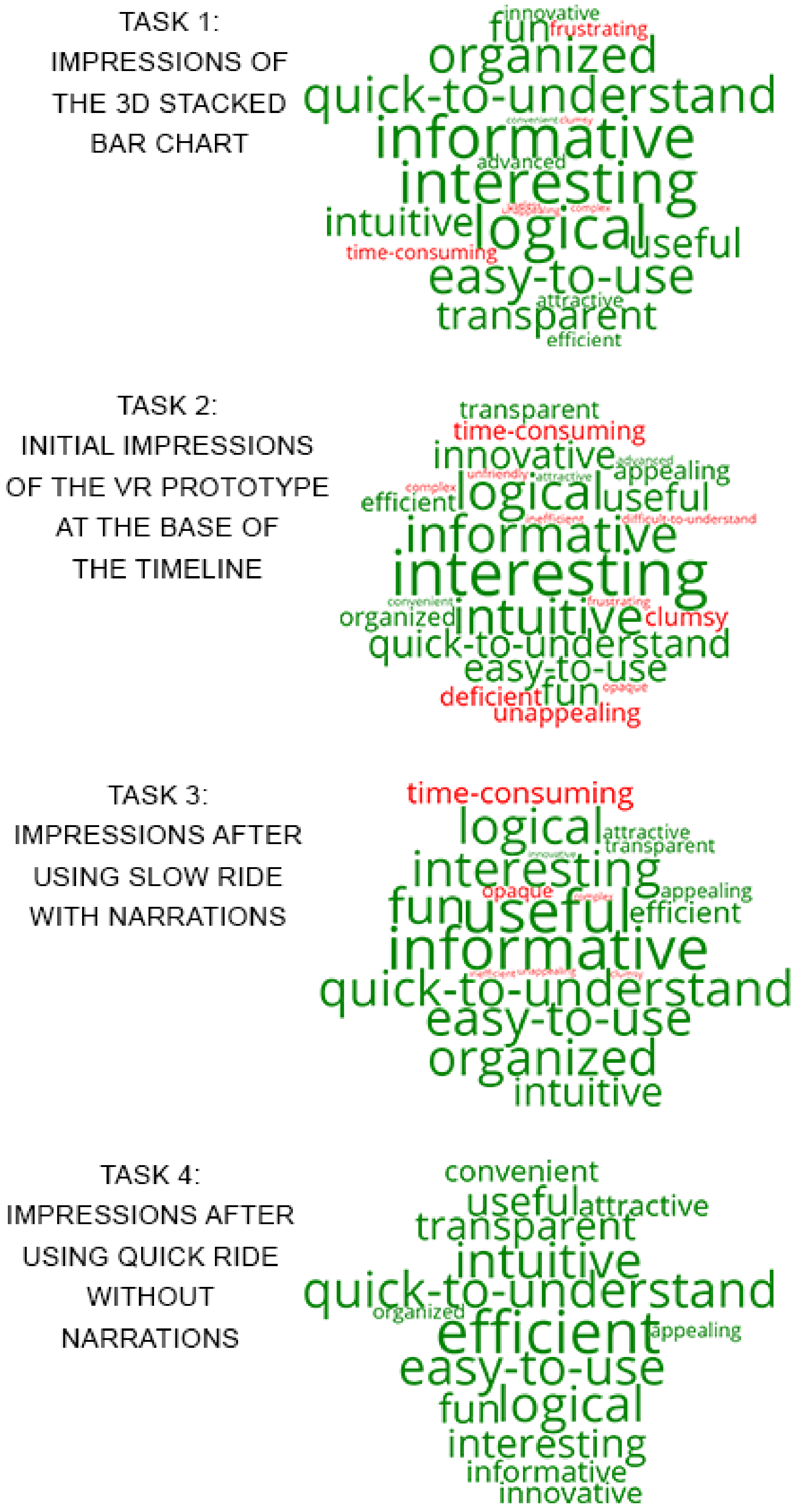


Figure 49: Word clouds visualizing impressions of the first 4 tasks.

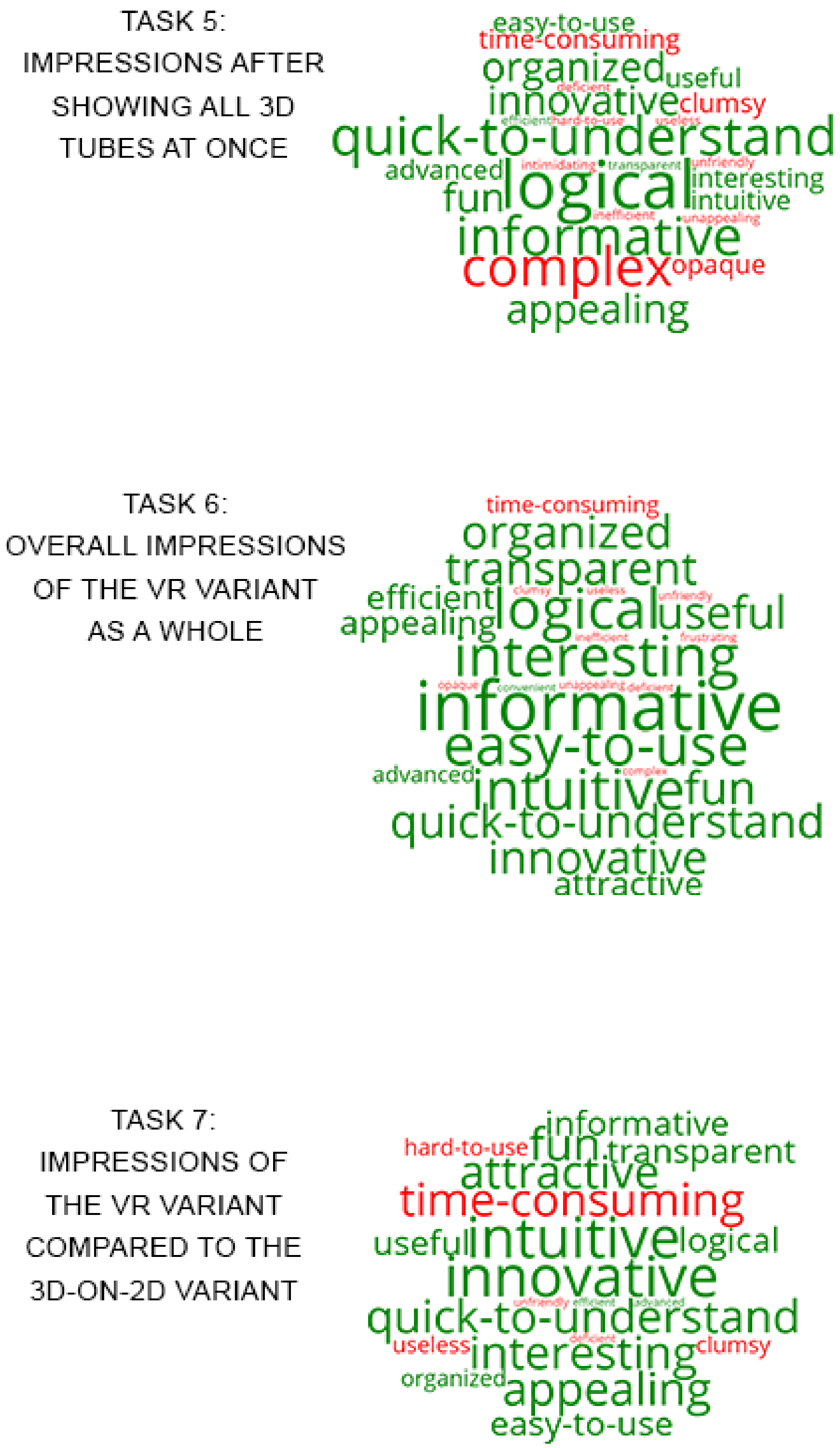


Figure 50: Word clouds visualizing impressions of the last 3 tasks.

A.4 StreamViz evaluation – radar charts comparing tasks on 5 dimensions

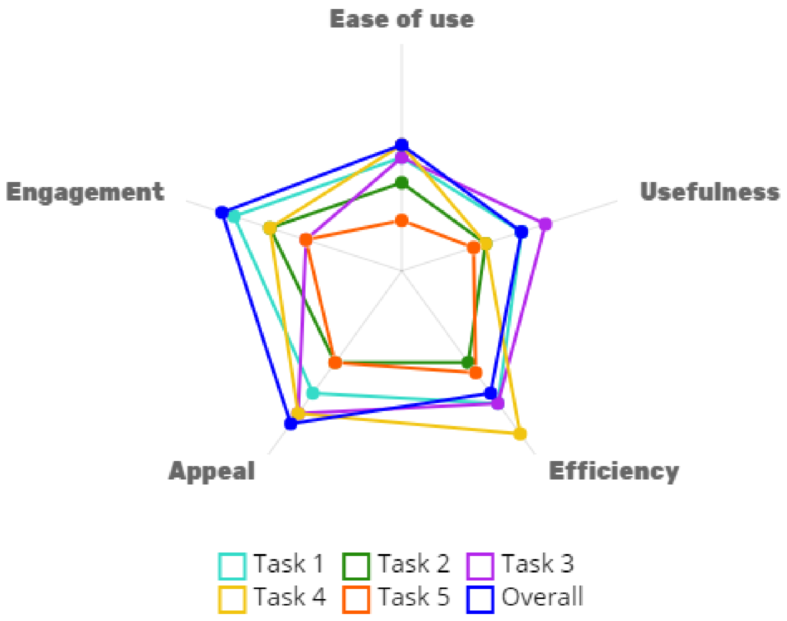


Figure 51: Radar chart comparing the impressions of the 5 tasks and the overall impressions of the VR prototype.

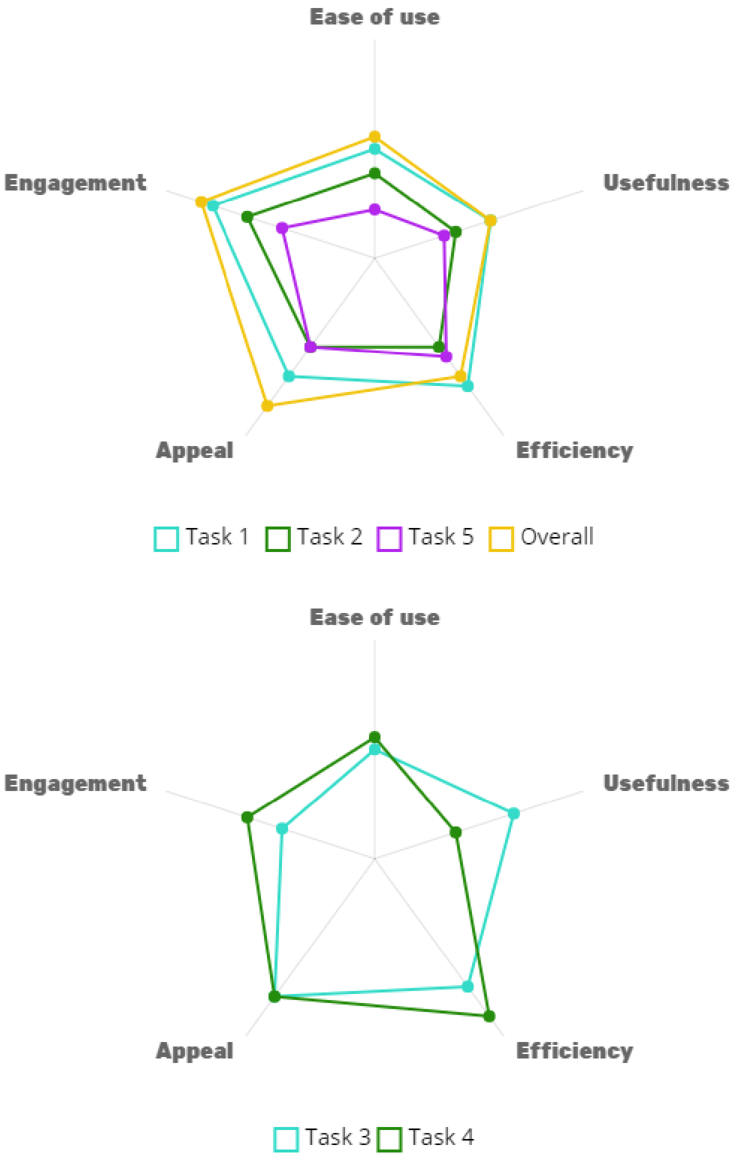
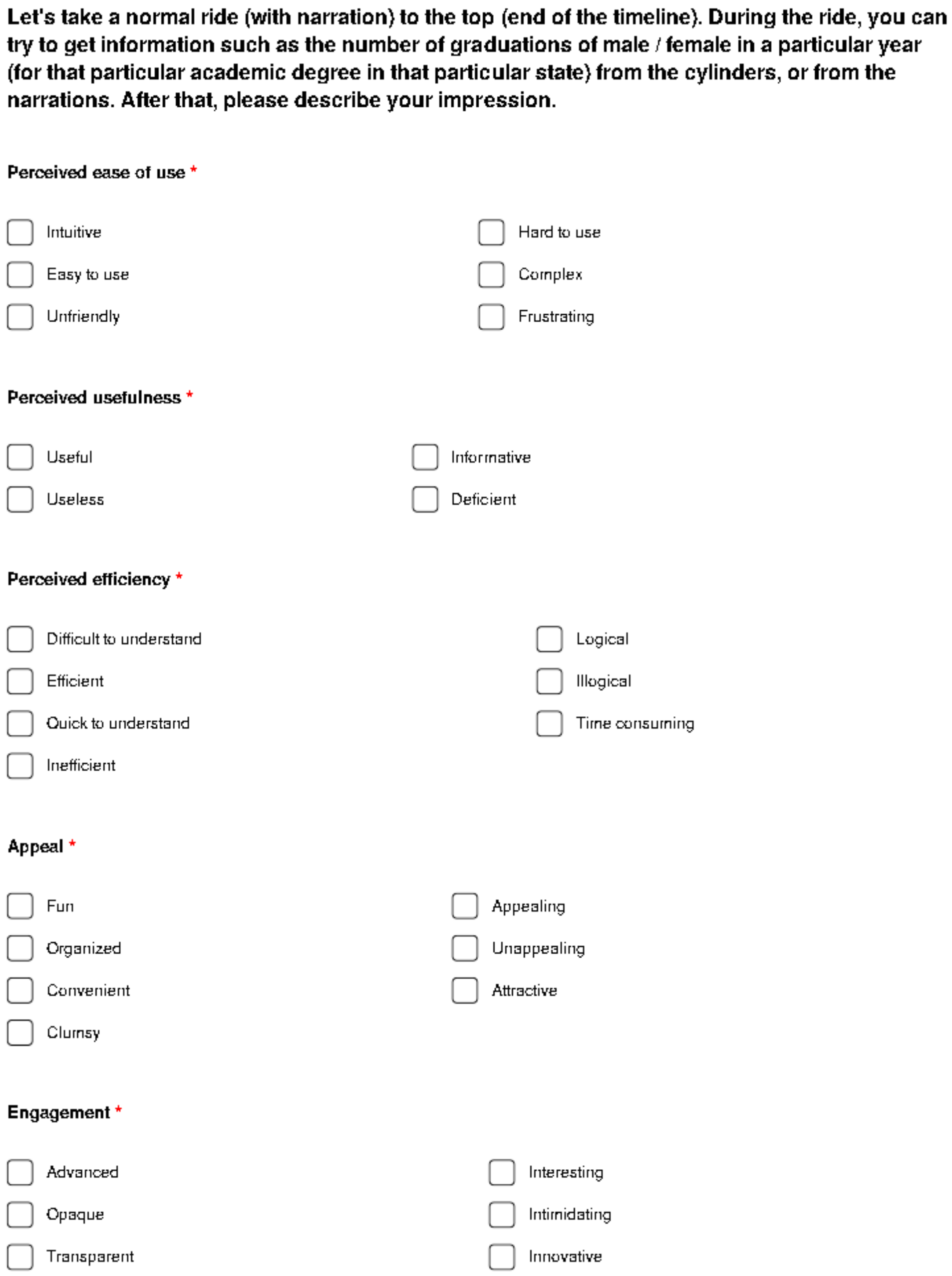


Figure 52: Different shapes of radar charts.

A.5 StreamViz evaluation – Excerpt from the survey's questionnaire



Let's take a normal ride (with narration) to the top (end of the timeline). During the ride, you can try to get information such as the number of graduations of male / female in a particular year (for that particular academic degree in that particular state) from the cylinders, or from the narrations. After that, please describe your impression.

Perceived ease of use *

□ Intuitive
□ Easy to use
□ Unfriendly
□ Hard to use
□ Complex
□ Frustrating

Perceived usefulness *

□ Useful
□ Useless
□ Informative
□ Deficient

Perceived efficiency *

□ Difficult to understand
□ Efficient
□ Quick to understand
□ Inefficient
□ Logical
□ Illogical
□ Time consuming

Appeal *

□ Fun
□ Organized
□ Convenient
□ Clumsy
□ Appealing
□ Unappealing
□ Attractive

Engagement *

□ Advanced
□ Opaque
□ Transparent
□ Interesting
□ Intimidating
□ Innovative

Figure 53: Excerpt from the user study's questionnaire.

References

Cisco. (2016) The zettabyte era — trends and analysis. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html [Visited on: 2016-08-08] 1, 6

E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed. Graphics Press, 2001. 1, 27, 29, 30, 31, 32, 44

C. Ware, *Information Visualization: Perception for Design*, 3rd ed. Morgan Kaufmann, 2012. 1, 30

J. Tauberer, *Open Government Data: The Book*, 2nd ed. Amazon Digital Services, 2014. 1, 2

D. Chavez. (2014) Is 3d visualization the next step for big data? [Online]. Available: http://computer.ieeesiliconvalley.org/wp-content/uploads/sites/2/2014/07/IEEE-CS-Big-Data-VR-15.pdf [Visited on: 2016-07-02] 2

D. McCandless. (2010) Ted: The beauty of data visualization. [Online]. Available: http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en [Visited on: 2016-06-05] 2

E. Dumbill, *Planning for Big Data - a CIO's Handbook to the Changing Data Landscape*, 1st ed. O'Reilly Media, 2012. 3

F. Nightingale, *Mortality of the British Army*. Harrison and Sons, 1857. 4

J. Snow, "On the mode of communication of cholera," 1855. 4, 26, 27

J. Cleve and U. Lämmel, *Data Mining*. De Gruyter Studium, 2014. 6, 19

M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, K. L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver, "Data, information, and knowledge in visualization," *IEEE Computer Graphics and Applications*, vol. 29, no. 1, pp. 12–19, Jan 2009. 6, 7, 70, 73

U. Fayyad, G. Grinstein, and A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2002. 6

N. Sharma. (2008) The origin of data information knowledge wisdom hierarchy. [Online]. Available: https://www.researchgate.net/publication/292335202_The_Origin_of_Data_Information_Knowledge_Wisdom_DIKW_Hierarchy [Visited on: 2016-06-01] 7

R. L. Ackoff, "Data, information, and knowledge in visualization," *Journal of Applied Systems Analysis*, vol. 16, pp. 3–9, 1989. 7

D. Laney. (2001) 3d data management: Controlling data volume, velocity, and variety. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [Visited on: 2016-07-17] 7

The Open Knowledge Foundation. (2014) What is open? [Online]. Available: https://okfn.org/opendata/ [Visited on: 2016-07-20] 8

Open Government Standards. (2016) The open government standards. [Online]. Available: http://opengovstandards.org/ [Visited on: 2016-07-20] 9

Open Government Data. (2016) What is open government data. [Online]. Available: http://opengovernmentdata.org/ [Visited on: 2016-07-20] 9

——. (2016) The annotated 8 principles of open government data. [Online]. Available: https://opengovdata.org/ [Visited on: 2016-07-21] 9

J. Tauberer. (2012) A brief legal history of open government data. [Online]. Available: https://opengovdata.io/2014/legal-history/ [Visited on: 2016-07-22] 9

D. Dietrich. (2011) Open data - was sind offene daten? [Online]. Available: http://www.bpb.de/gesellschaft/medien/opendata/64055/was-sind-offene-daten [Visited on: 2016-07-22] 9

Open Data Handbook. (2016) Why open data? [Online]. Available: http://opendatahandbook.org/guide/en/why-open-data/ [Visited on: 2016-07-23] 10

CTIC. (2015) Public dataset catalogs faceted browser. [Online]. Available: http://datos.fundacionctic.org/sandbox/catalog/faceted/ [Visited on: 2016-07-24] 10

Data.gov. (2016) Federal agency participation. [Online]. Available: https://www.data.gov/metrics [Visited on: 2016-07-24] 10

C. Forsterleitner and T. Gegenhuber, "Lasst die daten frei! open government als kommunale herausforderung und chance," *Freiheit vor Ort: Handbuch kommunale Netzpolitik*, pp. 233–266, 2011. 10

Wikipedia, "Open government," 2016. [Online]. Available: https://en.wikipedia.org/wiki/Open_government [Visited on: 2016-08-08] 10

Freie Hansestadt Bremen. (2016) Transparenzportal bremen. [Online]. Available: http://transparenz.bremen.de/sixcms/detail.php?gsid=bremen02.c.734.de [Visited on: 2016-07-24] 11

D. Dietrich. (2011) Open data - offene daten in deutschland. [Online]. Available: http://www.bpb.de/gesellschaft/medien/opendata/64061/offene-daten-in-deutschland [Visited on: 2016-07-22] 11

D. Klein, P. Tran-Gia, and M. Hartmann, "Big data," *Informatik-Spektrum*, vol. 36, no. 3, pp. 319–323, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00287-013-0702-3 12

Open Knowledge Foundation - CKAN Team. (2012) Ckan data hub. [Online]. Available: https://commondatastorage.googleapis.com/ckannet-storage/2012-02-13T201110/CKAN_Overview.pdf [Visited on: 2016-07-02] 14

CKAN. (2016) Ckan instances around the world. [Online]. Available: http://ckan.org/instances/ [Visited on: 2016-07-27] 14

Wikipedia. (2016) Linked data. [Online]. Available: https://en.wikipedia.org/wiki/Linked_data [Visited on: 2016-07-27] 15

T. Berners-Lee. (2009) Linked data - design issues. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html [Visited on: 2016-07-27] 15

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. (1996) From data mining to knowledge discovery in databases. [Online]. Available: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf [Visited on: 2016-08-05] 17, 18, 20

W. S. Cleveland, *Visualizing Data*. Hobart Press, 1993. 19

IBM. (2014) What is big data. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.htmll [Visited on: 2016-07-27] 22

M. Friendly. (2009) Milestones in the history of thematic cartography, statistical graphics, and data visualization. [Online]. Available: http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf [Visited on: 2016-07-29] 22

——, "A brief history of data visualization," in *Handbook of Computational Statistics: Data Visualization*, C. Chen, W. Härdle, and A. Unwin, Eds. Heidelberg: Springer-Verlag, 2006, vol. III, (In press). 22, 25, 28

E. R. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 1983. 23, 24, 26, 78

H. G. Funkhouser, "A note on a tenth century graph," *Osiris*, pp. 260–262, 1936. 23

N. Oresme, *Tractatus de Latitudinibus Formarum*. Padova, 1482. 23

——, *Nicole Oresme and the Medieval Geometry of Qualities and Motions: A Treatise on the Uniformity and Difformity Known as Tractatus de Configrationibus Qualitatum et Motuum*. Madison WI: University of Wisconsin Press, 1968. 23

C. Scheiner, "Rosa ursina sive sol ex admirando facularum & macularum suarum phoenomeno varius," 1630. 24

P. Buache, *Essai de geographie physique*. Memoires de L'Academie Royale des Sciences, 1752. 25

M. du Carla-Boniface, "Expression des nivellements; ou, methode nouvelle pour marquer sur les cartes terrestres et marines les hauteurs et les configurations du terrain," *From the Depths to the Heights, Surveying and Mapping*, vol. 30, p. 396, 1782. 25

S. Ferguson, *The 1753 carte chronographique of Jacques Barbeu-Dubourg*. Princeton University Library Chronicle, 1991, vol. 52. 25

W. Playfair, “Commercial and political atlas: Representing, by copper-plate charts, the progress of the commerce, revenues, expenditure, and debts of england, during the whole of the eighteenth century,” *The Commercial and Political Atlas and Statistical Breviary*, 1786. 25

——, “Statistical breviary; shewing, on a principle entirely new, the resources of every state and kingdom in europe,” *The Commercial and Political Atlas and Statistical Breviary*, 1801. 25

——, “Letter on our agricultural distresses, their causes and remedies; accompanied with tables and copperplate charts shewing and comparing the prices of wheat, bread and labour, from 1565 to 1821,” 1821. 26

W. Smith, “A delineation of the strata of england and wales, with part of scotland; exhibiting the collieries and mines, the marshes and fenlands originally overflowed by the sea, and the varieties of soil according to the substrata, illustrated by the most descriptive names.” *BL: Maps 1180*, 1815. 26

C. Dupin, “Carte figurativ de l’instruction populaire de la france,” *BNF: Ge C 6588*, p. 300, 1826. 26

R. Baker, “Report of the leeds board of health,” *BL: 10347.ee.17*, 1833. 26

M. Friendly, “Mosaic displays for multi-way contingency tables,” *Journal of the American Statistical Association*, vol. 89, pp. 190–200, 1994. 26, 29

G. Zeugner, “Abhandlungen aus der mathematischen statistik,” *Leipzig BL: 8529.f.12*, 1869. 27

L. Perozzo, “Della rappresentazione graphica di una collettivita di individui nella successione del tempo,” *Annali di Statistica BL:S.22*, vol. 12, pp. 1–16, 1880. 27

W. C. Eells, “The relative merits or circles and bars for representing component parts,” *Journal of the American Statistical Association*, vol. 21, pp. 119–132, 1926. 28

R. von Huhn, “A discussion of the eells’ experiment,” *Journal of the American Statistical Association*, vol. 22, pp. 31–36, 1927. 28

J. N. Washburne, “An experimental study of various graphic, tabular and textual methods of presenting quantitative material,” *Journal of Educational Psychology*, vol. 18, pp. 361–376, 465–476, 1927. 28

J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977. 28

J. Bertin, *Semiologie Graphique: Les diagrammes, les reseaux, les cartes*. Gauthier-Villars, 1967. 28

R. A. Becker and W. S. Cleveland, “Brushing scatterplots,” *Technometrics*, vol. 29, pp. 127–142, 1987. 29

D. Asimov, “Grand tour,” *SIAM Journal of Scientific and Statistical Computing*, vol. 6(1), pp. 128–143, 1985. 29

P. A. Tukey and J. W. Tukey, “Graphical display of data sets in 3 or more dimensions,” *Interpreting Multivariate Data*, 1981. 29

E. J. Wegman, “Hyperdimensional data analysis using parallel coordinates,” *Journal of the American Statistical Association*, vol. 85(411), pp. 664–675, 1990. 29

J. A. Hartigan and B. Kleiner, “Mosaics for contingency tables,” *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pp. 268–273, 1981. 29

S. E. Fienberg, “Perspective canada as a social report,” 1975, unpublished paper. 29

M. Friendly, “Extending mosaic displays: Marginal, conditional, and partial views of categorical data,” *Journal of Computational and Graphical Statistics*, vol. 8(3), pp. 373–395, 1999. 29

M. A. Fishkeller, J. H. Friedman, and J. W. Tukey, “Prim-9: An interactive multidimensional data display and analysis system,” 1974. 29

F. W. Young, P. Valero-Mora, and M. Friendly, *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*. Wiley, 2006. 29

W. S. Cleveland, *The Elements of Graphing Data*. Hobart Press, 1994. 30

Scholarpedia. (2016) Gestalt principles. [Online]. Available: http://www.scholarpedia.org/article/Gestalt_principles [Visited on: 2016-07-02] 30

R. Spence, *Information Visualization*. Addison-Wesley, 2001. 31

C. Ware, *Information Visualization: Perception for Design*, 3rd ed. Morgan Kaufmann, 2012. 31

J. Bertin, *Graphics and Graphic Information Processing*. Walter de Gruyter, 1981. 33

A. von Hoff. (2009) Krebsatlas: Hier ist das tumorrisiko am grössten. [Online]. Available: http://www.focus.de/gesundheit/ratgeber/krebs/symptome/tid-25423/krebs-in-deutschland-krebsatlas-hier-ist-das-risiko-am-groessten_aid_731982.html [Visited on: 2016-07-02] 33, 34

R. Mazza, *Introduction to Information Visualization*. Springer, 2009. 34, 36

G. J. Myatt and W. P. Johnson, *Making Sense of Data, Part II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. Wiley and Sons, 2009. 35

H. Wainer, *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton University Press, 2005. 35

A. Inselberg, “N-dimensional graphics part i: Lines and hyperplanes,” 1981. 37

D. Keim, “Designing pixel-oriented visualization techniques: Theory and applications,” *IEEE Transactions of Visualization and Computer Graphics*, 2000. 37

B. Shneiderman, K. S. Card, and J. D. Mackinlay, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999. 38

M. Dodge and R. Kitchin, *Atlas of Cyberspace*. Pearson Education, 2002. 38

T. He, S. Eick, and K. Cox, “3d geographic network displays,” *ACM SIGMOD Record*, vol. 25(4), pp. 50–54, 1996. 38

R. Patterson and D. Cox, “Visualization study of the nsfnet,” 1994. 38

S. Card, J. Mackinlay, and G. Robertson, “Cone trees: Animated 3d visualizations of hierarchical information,” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology*, pp. 189–194, 1991. 39

J. J. von Wijk, H. van der Wetering, and E. Kleiberg, “Botanical visualization of huge hierarchies,” *Proceedings IEEE Symposium on Information Visualization*, pp. 87–94, 2001. 39

B. Shneiderman, “Tree visualization with treemaps: a 2d space-filling approach,” *ACM Transactions on Graphics*, pp. 92–99, 1991. 39, 91

H. Samet, *Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1989. 40

G. W. Furnas, *Generalized Fisheye Views*, 1986. 40

M. Travers, *A Visual Representation for Knowledge Structures*. ACM Hypertext ’89 Proceedings, Implementations and Interfaces, 1986. 40

C. Ding and P. Mateti, *A framework for the automated drawing of data structure diagrams*, 1990. 40

R. Becker, W. Cleveland, and M. Shyu, *The visual design and control of trellis displays*, 1996, vol. 1. 41

C.-H. Chen, W. Härdle, and A. Unwin, *Handbook of Data Visualization*. Berlin-Heidelberg: Springer-Verlag, 2008. 42, 43, 46, 47

P. N. Diaconis and J. H. Friedman, “M and n plots in recent advances in statistics,” 1983. 44

J. A. McDonald, *Interactive Graphics for Data Analysis*, 1982. 44

R. A. Becker, W. S. Cleveland, and A. R. Wilks, “Dynamic graphics for data analysis,” *Statistical Science 2*, 1987. 44

L. Anselin, “Interactive techniques and exploratory spatial data analysis,” *Geographical Information Systems: Principles, Techniques, Management and Applications*, 1999. 44

G. Wills, “Spatial data: Exploring and modelling via distance-based and interactive graphics methods,” 1992. 44

J. Roberts, “Exploratory visualization with multiple linked views,” 2004. 44, 46

A. F. X. Wilhelm, “Interactive statistical graphics: The paradigm of linked views,” *Handbook of Statistics*, vol. 24, 2005. 44

B. Shneiderman, “Dynamic queries for visual information seeking,” 1994. 45

——. (1997) Information visualization: White paper. [Online]. Available: http://www.cs.umd.edu/hcil/members/bshneiderman/ivwp.html [Visited on: 2016-08-02] 45

E. W. Young, R. A. Faldowski, and M. M. McFarlane, “Multivariate statistical visualization,” *Handbook of Statistics*, vol. 9, 1993. 45

J. C. Robert, R. Knight, M. Gibbins, and N. Patel, “Multiple window visualization on the web using vrml and the eai,” 2000. 46

K. Zhao, B. Liu, T. Tirpak, and A. Schaler, “V-miner: Using enhanced parallel coordinates to mine product design and test data,” 1994. 48, 50

X. Zaixian, H. Shiping, O. W. Matthew, and A. R. Elke, “Exploratory visualization of multivariate data with variable quality,” *IEEE Symposium on Visual Analytics Science and Technology*, 2006. 51

H. Tim, W.-S. Luk, and P. Stephen, “Data visualization on web-based olap,” *ACM*, 2011. 51

J. Beyer, M. Hadwiger, A. Al-awami, W.-K. Jeong, N. Kasthuri, J. W. Lichtman, and H. Pfister, “Exploring the connectome: Petascale volume visualization of microscopy data streams,” *IEEE 33: Computer Graphics and Applications*, vol. 4, pp. 50–61, 2013. 51

Y. Zhu, “Measuring effective data visualization,” in *Proceedings of the 3rd International Conference on Advances in Visual Computing - Volume Part II*, ser. ISVC’07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 652–661. [Online]. Available: http://dl.acm.org/citation.cfm?id=1779090.1779161 78, 79, 80, 81

S. Kosslyn, “Graphics and human information processing: A review of five books,” vol. 80, 1985, pp. 499–512. 78

S. Casner, “A task-analytic approach to the automated design of graphic presentation,” vol. 10, 1991, pp. 111–151. 78

J. Bertin, *Semiology of Graphics*. University of Wisconsin Press, 1983. 78

L. Nowell, R. Schulman, and D. Hix, “Graphical encoding for information visualization: An empirical study,” in *Proceedings of the IEEE Symposium on Information Visualization*, 2002. 78

R. Amar and J. Stasko, “Knowledge precepts for design and evaluation of information visualizations,” vol. 11, 2005, pp. 432–442. 78

M. Scaife and Y. Rogers, “External cognition: how do graphical representations work?” vol. 45, 1996, pp. 185–213. 79

B. Tversky, M. Agrawala, J. Heiser, P. Lee, P. Hanrahan, D. Phan, C. Stolte, and M.-P. Daniel, "Cognitive design principles for automated generation of visualizations." Lawrence Erlbaum Associates, 2006. 79

T. Merčun, "Evaluation of information visualization techniques: Analysing user experience with reaction cards," in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, ser. BELIV '14. New York, NY, USA: ACM, 2014, pp. 103–109. [Online]. Available: http://doi.acm.org/10.1145/2669557.2669565 80, 81, 82

J. Benedek and T. Miner, "Measuring desirability: New methods for evaluting desirability in a usability lab settings," 2002. 81

C. M. Barnum and L. A. Palmer, "More than a feeling: Understanding the desirability factor in user experience," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '10. New York, NY, USA: ACM, 2010, pp. 4703–4716. [Online]. Available: http://doi.acm.org/10.1145/1753846.1754217 81

H. Mosley and A. Mayer, *Benchmarking national labour market performance : a radar chart approach*, 1999, vol. 99-202. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-128578 86

G. S. Owen, G. Domik, T. M. Rhyne, K. W. Brodlie, and B. S. Santos. (2008) Definitions and rationale for visualization. [Online]. Available: www.siggraph.org/education/materials/HyperVis/visgoals/visgoal2.htm [Visited on: 2016-06-05]

A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, pp. 69–91, 1985.

M. Friendly. (2009) Milestones in the history of thematic cartography, statistical graphics, anad data visualization. [Online]. Available: http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf [Visited on: 2016-07-02]

J. Roberts, "Exploratory geovisualization," 2005.

D. McCandless, "The beauty of data visualization," *TED Talk*, 2010. [Online]. Available: http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en [Visited on: 2016-07-30]

V. Ikonomou, "Open data basierte digitale narrative strukturen," 2015.

Hiermit versichere ich, dass ich die vorliegende Arbeit im Sinne der Prüfungsordnung ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, October 7, 2016 Truong Vinh Phan