Defragging Datavisualizations

Adding new insights on the analysis of the external physical form of datavisualizations to the field of Information Visualization

Joëlle Dirksz | 3684814 | University of Utrecht

Master Thesis: Defragging Datavisualizations: Adding new insights on the analysis of the external physical form of datavisualizations to the field of Information Visualization.

Joëlle Dirksz
3684818

Tutor: dr. Imar de Vries
Second reader: dr. Michiel de Lange
University of Utrecht

August 15th, 2016

Abstract

The field of Information Visualization or datavisualization is related to a wide variety of other disciplines (like computer science and statistical analysis), but because of this it lacks a clear theoretical base upon which scholars and researchers can build their research. This makes describing and comparing achievements within the field hard to validate and defend. Since datavisualization is a process that involves the computer as well as the human, both sides should receive attention within Information Visualization theory. By drawing on theory on the external physical form of datavisualizations and images like that of Lev Manovich (2011), the abstraction/figuration distinction and Peirce's Sign Theory (1974 [1931-1958]), this thesis will explore in what way these theoretical approaches can help to further shape the theory and tools needed for the analysis and interpretation of datavisualizations within the field of Information Visualization.

Keywords: Information Visualization, datavisualization, theoretical framework.

Contents

1. Introduction

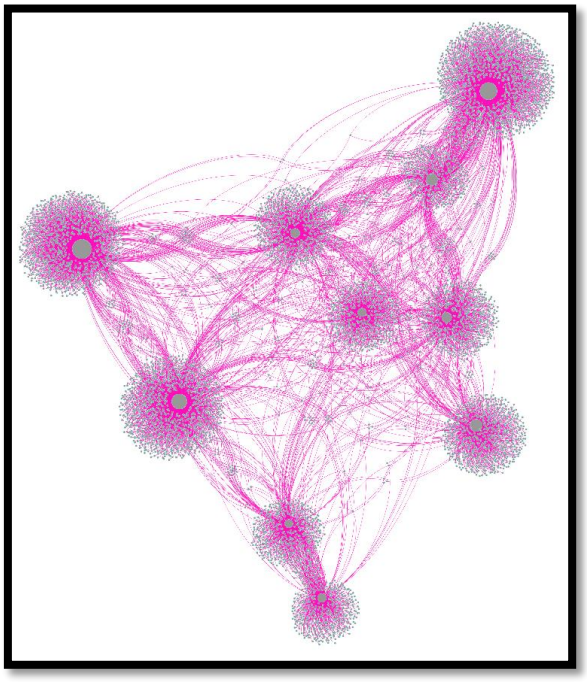

Image 1: A network visualization[1]

When you look at the network visualization just above this text, what do you see? What do you think it means? How do you interpret what you see and what you think it means into a meaningful understanding of the datavisualization? And why is it important for scholars to understand these datavisualizations anyway? Like many others I believe that the (big) data revolution is affecting the way we work, learn and think (Mayer-Schönberger and Cukier 2013). The amount of data we produce every day is astounding. "Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today, has been created in the last two years alone." (IBM). A significant part of this revolution is the way data is transformed into visual patterns that we as humans can understand and interpret. This is called the visualization of data, or datavisualization. Without this visualization of data, understanding data can become very difficult for us humans. More data leads to more visualizations and we now find ourselves surrounded by a diversity of datavisualizations from which we extract all sorts of information. For example, a visualization showing us information about the trees in your city (and who presumably owns them)[2], how healthy you ate today, who's tweets are trending on Twitter

[1] Twitter datavisualization that shows how a group of ten Twitter users (the big clusters) are connected to eachother (the lines) based on their Followers. Visualization created by author in cooperation with Big Fellows B.V. and the department VTH (Vergunningen, Toezicht and Handhaving – Licenses, Supervision and Enforcement) Gemeente Utrecht.

[2] These visualizations are part of a Dutch app called "Bomenapp": http://www.bomenapp.nl/.

or when to leave the main road because of a traffic jam on the road ahead. Data alone will not directly show you which tree does not have an owner yet, this is why we rely on the visualization to explain us that the tree you wanted to adopt has already been claimed by your neighbor. So if datavisualizations are important to fully understand the data that we use in our daily lives, why don't we have proper, up-to-date academic tools to analyze and interpret these upcoming datavisualizations from an academic perspective? How should scholars that use datavisualizations as research objects describe and compare them to other datavisualizations? Which standards, theories or principles should we use?

In "Theoretical Foundations of Information Visualization" (Purchase et al. 2008), the need for a more grounded theory for Information Visualization also becomes apparent. In their article Purchase et al. focus mostly on the tools used in datavisualization and state that "There is much unease in the community as to the lack of theoretical basis for the many impressive and useful tools that are designed, implemented and evaluated by Information Visualization researchers" (Purchase et al. 2008, 46). And that "the absence of a framework for Information Visualization makes the significance of achievements in this area difficult to describe, validate and defend" (Purchase et al. 2008, 46). For this reason they propose a framework to both evaluate and predict users' insight or understanding of visualization and their use of it. One of the issues described is how to evaluate a datavisualization for its information content. As one of the authors, Matthew Ward uses communication theory to look at the information content of the external physical form of a datavisualization (an activity usually performed by the reader). He proposes several methods for the analysis of information type, information content and information loss during the process of datavisualization. He concludes that there is a need to define measures of information within the datavisualization process to better evaluate information loss during the datavisualization process for example (2008, 58). The more information is transferred from the dataset to the datavisualization, the higher the information content.

Daniel Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler also voice this need for a clear Information Visualization theory in their article on the challenges in visual data analysis (2006). Their answer to the fast growing amount of data and the incapability of data analysis to keep up with these enormous amounts, very much focusses on determining the correct method to visualize data and making sure the outcome best represents the underlying data and will be most understandable for us humans. They call this approach *Visual Analytics*, which can be seen as a holistic view on data visualization (Keim et al. 2006, 10). "Visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis" (Keim et al. 2006, 10). Interesting about visual analytics, is that it pays specific attention to the human factor

within datavisualization. This is something that is very uncommon in the computer science field when it comes to datavisualization. But Keim et al. state that interpretability of datavisualizations is still one of visual analytics biggest challenges (2006, 12). Since the primary purpose of datavisualizations is to turn raw data into an understandable image for humans. Keim et al. argue that the holistic approach of visual analytics is an answer to this.

The author that has dedicated al lot of his research to shaping the Information Visualization field, is Lev Manovich. Lev Manovich is an author, scientist and Professor at The Graduate Center, the City University of New York and the founder and director of the Software Studies Initiative[3] and the only author within datavisualization that has given specific attention to characterizing the differences between datavisualization types based on the interpretation of their appearance. He has written a number of articles on datavisualization that consider both the technological perspective as well as the interpretative human perspective of datavisualization (2013). In his work on the exploration of datavisualization categories (2011), he compares how different datavisualizations make use of the principle of space and the principle of reduction, two principles that have been used to create datavisualizations for several hundred years (Manovich 2011, 38; Friendly and Denis 2009, 15). In short, the principle of space refers to the favoring of the spatial variable in the creation of datavisualizations (Manovich 2011, 39). The principle of reduction refers to the reducing of data to be able to show otherwise unseen patterns or connections (2011, 38). In chapter 2 these principles will be discussed into more detail.

Where Matthew Ward used communication theory to analyze datavisualizations for their information content, in 'Defragging Datavisualizations' I will take a first step in providing the Information Visualization field with academic tools to understand the external physical form of modern/emergent datavisualizations from a semiotics perspective. This will provide scholars and researches with the academic resources to analyze the meaning of the visual elements in datavisualizations. The work of Lev Manovich will play a significant role within this research and his work on the categorization of datavisualization can be seen as one of the main motives to get involved in the research on the appearance of datavisualizations done in this thesis.

1.1 What is a datavisualization?

The definition of what a datavisualization actually is, is not a simple one. But since they form the center of this research, an explanation of what is considered a datavisualization in this thesis is in place. It is mainly because the visualization of data as a research method is used in many different fields and does not belong to just one (Strecker 2012, 2), that makes its definition so hard to grasp.

[3] More on the Software Studies initiative can be found at lab.softwarestudies.com.

The multi-disciplinary usage of the visualization of data, resulted in not just one definition for the term 'datavisualization'. In this thesis the definition of a datavisualization will be the same as that of Michael Friendly and Daniel Dennis (2009), namely the "visual representation of 'data', defined as information which has been abstracted in some schematic form, including attributes or variables for the units of information as a visual result of data analysis, where complex data is made more accessible, understandable and usable by visualizing it" (Friendly and Denis 2009, 2). Notably the word 'understandable' is an important part of this definition, because it suggests that a datavisualization is understandable and that the data itself is not. Which is a statement I also agree with.[4]

To illustrate the diversity in datavisualizations today, in images two, three and four there are three different types of datavisualization shown. Image two is a network visualization, image three is a timepiece graph and image four is a data-map.

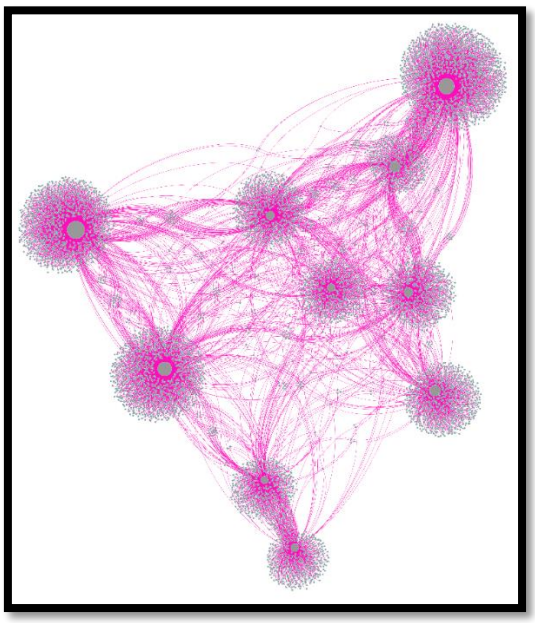

Image 2: A Twitter datavisualization that shows how a group of ten Twitter users (the big clusters) are connected to eachother (the lines) based on their Followers.[5]

[4] I will leave the discussion on what can be considered to be information to others, because it would open up an entire debate on its own. Not meaning I do not acknowledge the importance of it, but simply because it would steer the focus away from the research objective.

[5] Visualization created by author in cooperation with Big Fellows B.V. and the department VTH (Vergunningen, Toezicht and Handhaving – Licenses, Supervision and Enforcement) Gemeente Utrecht.

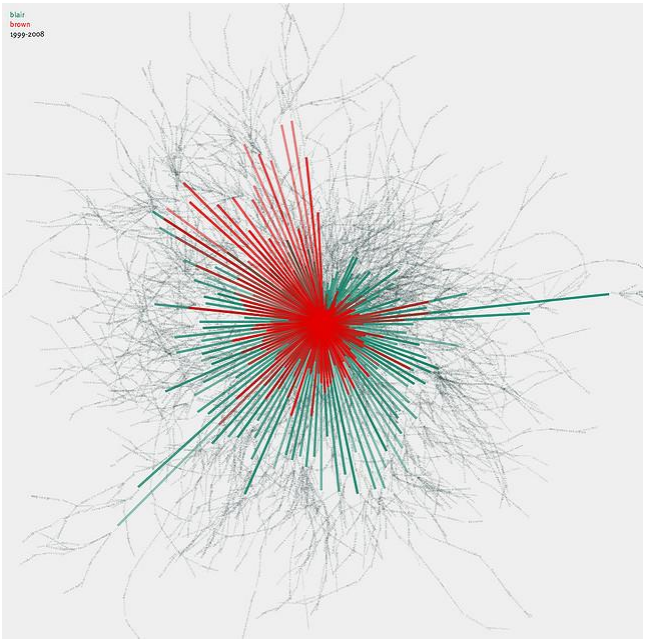

Image 3: A timepiece graph visualization of the mentions of Tony Blair and Grodon Brown in The Gardian from 1999 until 2008[6].

Image 4: Traffic intensity map of the city Assen, created by using the data collected by Dat.Mobility[7] trough the Sensor City project. The scale on the right shows the colors that match the intensity levels.[8]

[6] Image can be found at http://www.flickr.com/photos/blprnt/3347062507/in/photostream/

[7] Dat.Mobiliy is a relatively new company that that officially made its start in January 2014 and has worked closely with the organization of the Sensor City project to create traffic data-maps and mobility tools. They are specialized in solving mobility related issues using data and creating custom made tools. More information about Dat.Mobility or the Sensor City mobility project see: www.dat.nl.

[8] Image provided by Dat.Mobility.

These three images are all considered to be datavisualizations even though they do not necessarily share any visual similarities. For example, image two and image three have some visual similarities, because they are both abstract and use lines/bars to shown some kind of relation between data. However they use different kinds of data, use it in an entirely different way and they both portray very different information. The network visualization (image two) shows the relationship between Twitter users and the number of Followers[9] they have shown by the connections between the dots. The timepiece graph shows the number of mentions of 'Tony Blair' and 'Gordon Brown' in the newspaper 'The Guardian' from 1999 until 2008. The third datavisualization is even more different than other two, since its visual structure is completely different. This data-map shows the current state of traffic intensity in Assen and also predicts future traffic flows based on the data that it collects using several sensors.

The similarity between the three images above, exemplifies the need to understand the differences between datavisualizations. Almost any visualization that is made by using data can be called a datavisualization, despite of the fact that they might not have any visual elements in common. The objective of this thesis is to add insight on the analysis of the external physical form of datavisualizations to the field of Information Visualization and its theoretical framework by drawing on the principles of reduction and space and theory from semiotics

1.2 Research questions

In this thesis I will explore the principles of reduction and space in relation to emergent datavisualizations and additionally provide a new principle that can be used to better analyze and situate these datavisualizations. Therefore the main research question is: *How can the perspective of semiotics help to further understand the external physical form of emergent datavisualizations and help to continue building the Information Visualization theoretical framework?*

In order to answer the main research question, two guiding sub questions are asked. The first sub question is: *What information can(not) be extracted from emergent datavisualizations by using the already known principles of reduction and space?* This question will be discussed in chapter two by exploring how the principles of reduction and space relate to emergent datavisualizations. The answer to this question will also reveal to which extent the old principles are presumably unable to interpret emergent datavisualizations. The second research question will build on the first one by exploring how *Peirce's Sign Theory* (1974 [1931-1958]) and the *Abstraction/Figuration distinction* (The Art Story Foundation 2014) can help to build an additional useful principle for the analysis of emergent

[9] Twitter followers are users on the social network site Twitter that have connected to other users so that they can follow the tweets (messages) that are posted by these users.

datavisualizations. Therefore the second sub question is: *What other information can be extracted from emergent datavisualizations by reaching out to the Abstraction and Figuration distinction and Peirce's Sign Theory?* This question will be discussed in chapter 3 and will lead to the introduction of an additional datavisualization principle.

1.3 Method

Where Purchase et al. (2006) were most interested in measuring the information content in datavisualization based on communication theory (53), I am interested in understanding the actual meaning of the information in the datavisualization. By drawing on the principles of reduction and space as well as theory from semiotics (abstraction/figuration theory and Peirce Sign Theory) this research will provide and exploration of the type of information that can be extracted from emergent datavisualizations. Whether this approach will lead to something that can be measured or turned into a quantified method to comparing the information content of datavisualizations (as done in the research of Purchase et al. 2006) is not included in the scope and left to future research.

To explore the principles of reduction and space in relation to the analysis of emergent datavisualizations, I will be conducting exploratory research on both principles to reveal their analytical abilities and limits when analyzing emergent datavisualizations. In this section a brief explanation of the method I have used for this research will be given as an introduction to the further execution of this method in the following chapters. As beautifully described in Robbert A. Stebbins 'Exploratory research in the social sciences', "Exploratory research is about putting one's self deliberately in a place -again and again- where discovery is possible and broad" (Stebbins 2001, vi). In this thesis the discovery of the usefulness of the two acknowledged principles of datavisualization will be put to the test by relating them to emergent datavisualizations. This exploration will result in knowing what information the two principles can extract from emergent datavisualizations and will reveal any shortcomings that the principles of reduction and space might have.

A qualitative textual analysis of a specific case study will be done in order to exemplify what information the principles of reduction and space and new principle can extract from emergent datavisualizations. According to Robert Yin, social sciences scholar, using a case study is a good way to do an in-depth exploration of a specific case (Yin 2014, 4). As stated by Bonnie Brennen (2012), a textual analysis is not just the analysis of an actual 'text'. In the context of a textual analysis, a text is more than just a composition of letters, words and sentences. A text can refer to any cultural artifact that has impacted our society (Brennen 2012, 193). In this thesis a datavisualization as a result of data analysis is considered to be a text, a cultural artifact that visualizes certain patterns. Also, the scope of this type of analysis is not determined by a set method, but by the used theories that guide it (199).

Consequently, the three principles discussed in this thesis are used to guide the textual analysis, which also results in discussing the research case in both chapters two (where the old principles are discussed) and three (where the new principle is introduced).

The case will consist out of traffic data-maps from the Sensor City project in Assen[10]. The Sensor City project was an experiment to better manage the dynamic traffic patterns in Assen using a variety of self-collected traffic data (2014 (Noordegraaf, Jonkers and de Kruijf 2014). In order to conduct the textual analysis of this case study some screenshots were provided by Dat.Mobility[11], a partner of the Sensor City project who has used the generated data to create traffic data-maps. Traffic-data maps include many characteristics of emergent datavisualizations because they work with dynamic, always changing and up to date data that is visualized directly onto a digital geographical map[12]. This makes the Sensor City traffic-data maps an excellent example of an emergent datavisualization. Because of the structure of this thesis, the analysis of the case will follow directly after the discussion of the principles of reduction and space (chapter two) and the new principle of abstraction and figuration (chapter three). Meaning that the case study is divided into two sections that are discussed in both chapter two and chapter three. Using the case study in this way helps to better exemplify the way each principle would relate to an emergent datavisualizations, since a direct link to an emergent datavisualization can be made in the case study following the discussed principle. Also by analyzing the case in this way, I was able to use the insights from the analysis in chapter two (the analysis of the case study using the principles of reduction and space) for the shaping of the new principle in chapter three. This enables the new principle to fill in some of the shortcomings of the principles of reduction and space, rather than being a standalone principle that has no relation to these shortcomings. This method therefore makes it more likely that the proposed principle will serve as a connected addition to the old principles.

1.4 Theoretical framework

In this thesis the main theory that is used to form the additional principle is derived from the analysis of emergent datavisualizations using the principles of reduction and space done in chapter two. Because of this, the principles of reduction and space provide the theoretical starting point of the exploration. The theories used to shape the new principle of abstraction and figuration have their

[10] Additional information on the Sensor City can be found at www.sensorcity.nl.
[11] Dat.Mobiliy is a company that that officially made its start in January 2014 and has worked closely with the organization of the Sensor City project to create traffic data-maps and mobility tools. They are specialized in solving mobility related issues using data and creating custom made tools. More information about Dat.Mobility or the Sensor City mobility project see www.dat.nl.
[12] In this thesis it is not possible to show a dynamic datavisualization, the dynamic properties of the datavisualizations will still be taken into consideration in the analysis. Because of the format in which this thesis is written, screenshots have been used to illustrate the case.

roots in the field of semiotics, namely the abstraction/figuration distinction and Peirce's Sign Theory (1974 [1931-1958]). In addition to this, the necessity of an image having to be objective for the shown information to be true, will be questioned by relating objectivity to the interpretation of emergent datavisualizations. Therefore the theory from Lorraine Daston (1992), Peter Galison (1992, 2010), Danah Boyd and Kate Crawford (2012) on the notion of objectivity has also helped to shape how we can interpret emergent datavisualizations.

This theoretical framework will provide a brief overview of the principles of reduction and space, the abstraction/figuration distinction, Peirce's Sign Theory and how this relates to objectivity.

1.4.1. The principles of reduction and space

Both the principle of reduction and the principle of space have been acknowledged principles for the creation and analysis of datavisualizations since the early 1800's, when the visualization of data was booming for the first time (Manovich 2011, 39; Friendly and Denis 2009, 15). According to Michael Friendly and Daniel J. Denis all "modern"[13] forms of data display were invented between 1800 and 1850 (2009, 15). These are the traditional datavisualizations that are well known to most educated people, like the bar or line graph, pie chart or the scatterplot. In short, the principle of reduction refers to the reduction of the amount of data in order to visualize otherwise unseen patterns or connections (Manovich 2011, 38). The principle of space refers to the favoring of the spatial variable for the display of significant information in the creation of datavisualizations (2011, 39). The often used visualization categories: *information visualization*, *scientific visualization* and *information design*, are based on the principles of reduction and space (Manovich 2011, 37; Friendly and Denis 2009, 2). Meaning that the principles of reduction and space form the foundation of how we distinguish between datavisualizations when using this categorization. In chapter two I will go into more detail on this categorization, since it has its roots in the principles of reduction and space and therefore is a helpful tool in the analysis.

In his article 'What is datavisualisation?' (2011), professor Lev Manovich actively questions the sufficiency of the old principles in relation to emergent datavisualizations, since they are approximately 200 years old. In his exploration he concludes that the principle of reduction and the principle of space do not account for every visualization produced in the last 300 years (2011, 41). However, "they are sufficient to separate infovis[14] (at least as it was commonly practiced until now)

[13] The word "modern" is put between double quotation marks since according to Friendly and Denis modern datavisualiztions are considered to be "bar and pie charts, histograms, line graphs and time-series plots, contour plots, and so forth" (Friendly and Denis 2009, 15). Today, emergent datavisualizations can be much more complex in both their visual appearance as in their underlying calculations, making it possible to create other forms of datavisualizations than the "modern" datavisualizations mentioned by Friendly and Denis.

[14] Infovis is an abbreviation of information visualization which in this context can also be called datavisualization.

from other techniques and technologies for visual representation: maps, engraving, drawing, oil painting, photography, film, video, radar, MRI, infrared spectroscopy, etc. They give infovis its unique identity – the identity which remained remarkably consistent for almost 300 years (i.e. until the 1990s)." (Manovich 2011, 41). Manovich explicitly remarks that both principles have been sufficient to separate information visualization from other types of visual representation until the 1990's. This is because from the 1990's onwards, the creation of more complex datavisualizations was made possible by using computers, which consequently blurred the lines that used to separate information visualization from other types of visual representations of data. The work of Manovich therefore implies that the two old principles used to analyze and interpret datavisualizations have become insufficient for the analysis of emergent datavisualizations that are digitally created for example (like data-maps). In chapter two I will further elaborate on these principles and Manovich his view on them.

1.4.2. The Abstraction/Figuration distinction and Peirce's Sign Theory

As will become apparent in chapter two, in addition to analyzing the relation between the visual representation of the data and the data itself using the principles of reduction and space, semiotics can be used to explore the relation between the visual elements in the datavisualization and the actual meaning of these visual elements. Therefore providing an in-depth understanding of many visual elements in emergent datavisualizations that were not present in more traditional forms of datavisualization like the bar and line graphs. This in-depth understanding cannot be achieved using just the two known principles, which is why semiotics can provide some new insights into emergent datavisualizations.

Within semiotics there are two directions that help to shape the proposed principle of abstraction and figuration. The first is the abstraction/figuration distinction (The Art Story Foundation 2014) and the second is Charles Peirce's Sign Theory (1974 [1931-1958]). The proposed principle is used to explore how datavisualizations relate to being abstract, figurative or both. It is concerned with the relation between the datavisualization and the objects or ideas that it represents. It will also help to identify the different layers and elements of a datavisualization. These individual elements will be further analyzed using Peirce's Sign Theory to understand what they represent and how they relate to the entire image. Peirce's theory of the sign consists out of many different theories, concepts and ideas on the relation between the sign, object and the interpretant which Peirce himself rethought a number of times (1974 [1931-1958], 243, 263). For the creation of the principle of abstraction and figuration, the

relationship between the sign (or the representamen) and its object (the actual thing that the sign is representing) is will be the main focus in this research.[15]

1.4.3. Objectivity

When introducing the principle of abstraction and figuration I will also discuss its relation to objectivity and rekindle the discussion on the necessity of (emergent) datavisualizations having to be objective or not. Here the work of Lorraine Daston (1992) and Peter Galison (1992, 2010) on the changing notion of objectivity and the work of Danah Boyd and Kate Crawford (2012) on the interpretation of data in relation to objectivity will be discussed. The necessity of objectivity might seem a different discussion, but as will become apparent it is strongly tied to the abstract and figurative elements in an image. The reason for incorporating the objectivity of an image, is because it appeared to be a logical next step in the analysis of an emergent datavisualizations when the emphasis is put on understanding the meaning of individual elements of the image and the image as a whole. Which is what the new principle of abstraction and figuration does. The objectivity of an image like a datavisualization, is narrowly tied to its visual appearance and thus to the interpretation and representation of the image. In this exploration, objectivity can be viewed as a critical note tied to the principle of abstraction and figuration.

1.5 Chapter overview

The following two chapters will be used to answer the main research question and the two sub questions. In chapter two the first sub question will be discussed by critically looking at the principles of reduction and space in comparison to emergent datavisualizations. This chapter will also introduce and analyze the case in relation to the principles of reduction and space using the three datavisualizations categories also discussed by Lev Manovich (2011) and a textual analysis. In chapter three the second sub question will be discussed and a new principle will be proposed: the principle of abstraction and figuration. In this chapter the case will be analyzed again but now in relation to the newly proposed principle of abstraction and figuration. Chapter four is the final chapter of this thesis where the answers to the research questions and the limitations of the research method will be discussed.

2. The principles of reduction and space put into context

The principle of reduction and the principle of space are two key principles in datavisualization that have been used in this field since the eighteenth century (Manovich 2011, 38; Friendly and Denis

[15] In this thesis, the way in which the interpretant (the interpretative effect on the person) is established when a person is viewing a datavisualization will only be seen from a theoretical point of view. Meaning that no test subjects were involved in this research.

2009, 15). Even though more modern techniques have become available to create datavisualizations[16], today these principles are still the only two key principles available to analyze and interpret datavisualizations. This raises the question if they are still sufficient as the only two principles of datavisualization. In this chapter the first sub question is discussed: *What information can(not) be extracted from emergent datavisualizations by using the already known principles of reduction and space?* I will critically reflect on these principles by putting them in relation to emergent datavisualizations and exemplify in what way these principles are possibly insufficient to analyze emergent datavisualizations.

2.1 The Principle of Reduction

The first key principle is the principle of reduction, which is used to show that certain information can only be seen when reducing the amount of data to expose patterns or connections (Manovich 2011, 38). Datavisualization uses graphical primitives such as lines, points and shapes to represent certain relations of values of objects, people, stock prices, etcetera. Image five shows a simple bar graph in which the relationship between the estimated amount of data being produced in the entire world and the timespan in which this happens can be seen.

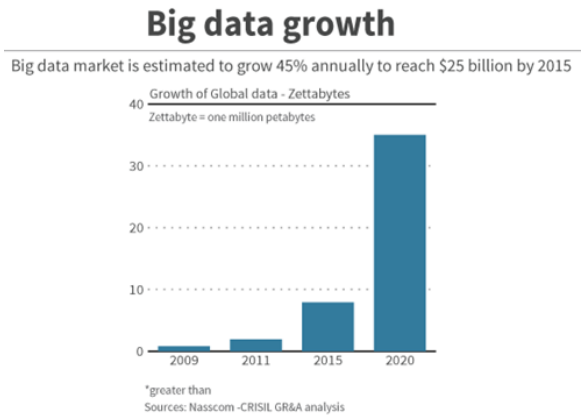

Image 5. A bar graph that shows the growth and expected growth of the amount of data produced in the world.[17]

Reducing the visibility of larger amounts of data to these absolute primitives enables the revealing of structures and patterns that would otherwise remain unseen, creating new insights from the same data. Reducing an amount of data to these primitives does come with the consequence of extreme schematization (2011, 38). In the case of extreme schematization as little as one percent of rich data is visible in the visualization, which means that 99 percent of the data is invisible to the viewer of the

[16] For example computer generated datavisualizations that use an algorithm to calculate what the datavisualization will eventually look like.
[17] Image 4 can be found at http://www.datamation.com/applications/how-vertical-markets-will-drive-big-data.html.

datavisualization. However as can be seen in image five, it would not be possible to see important connections without the reduction of data. Interestingly, Lev Manovich (2011) has tried to break down this presumed fact by creating a new type of datavisualization that works without reduction of data or extreme schematization. Manovich calls this visualization process *direct visualization* or *media visualization* (41), because these visualizations show 100 percent of the used data. In this type of datavisualization there is no reduction of data, meaning that all the original data is visible in one visualization. Several examples made by Manovich can be found in Manovich's article 'Media Visualization: Visual Techniques for Exploring Large Media Collections' (2011a). This article also further explains the methods used to create media visualizations. An example of a frequently used media visualization is a tag or word cloud[18] (see image 6), a datavisualizations in which the most used words of a certain website, weblog or any other text are shown in proportion to the number of times they were mentioned (tagged).

Image 6: A word cloud made from the words in Lev Manovich's article 'What is visualization?'.

Bigger words in the tag cloud are mentioned or tagged more than words that are smaller. Tag clouds can be seen as a form of direct visualization since this type of datavisualization is visually build out of the data it uses. Manovich states that "rather than representing text, images, videos or other media through new visual signs points or rectangles, direct visualization builds new representations out of the original media. Images remain images; text remains text" (Manovich 2011, 41). This suggests that the principle of reduction is important in emergent datavisualizations (for example see the network visualization in image one or two), but that this principle is not a requisite in the creation of every datavisualization. Therefore emergent datavisualizations do not always need to comply to the principle of reduction to be able to show new patterns of connections.

[18] A more detailed explanation of a tag or word cloud can be found at http://nl.wikipedia.org/wiki/Tag_cloud

2.2 The Principle of Space

The second principle is the principle of space. Lev Manovich states that a common aspect between datavisualizations is that they favor the spatial dimension over other visual dimensions like color, shape or shading (Manovich 2011, 39).

> "I believe that the majority of information visualization practices from the second part of the eighteenth century until today follow the same principle – reserving spatial arrangement (we can call it 'layout') for the most important dimensions of the data, and using other visual variables for remaining dimensions." (Manovich 2011, 39).

Examples of datavisualizations types that favor the spatial dimension are line graphs, bar graphs and network visualizations (images one, two and five). The other non-spatial variables could for example function as group labels, which add readability since the colors can be easier to use that the actual description of the group in words, but this does not necessarily add new information (Manovich 2011, 40).

The favoring of the spatial variable has been the standard for many years and remains popular in datavisualization today (39). However, it is not a requisite to always favor the spatial variable as the dominant variable in order to create a datavisualization. Modern day technology has led to datavisualizations in which the non-spatial dimensions are not just used as categorical dimensions that function as labels (40-41). In the heat map seen in image seven, the important information is visualized by the use of color (hue) and not by the spatial variable. The more intense the concentration of a certain value (in this case the popularity of locations with tourists), the brighter the color.

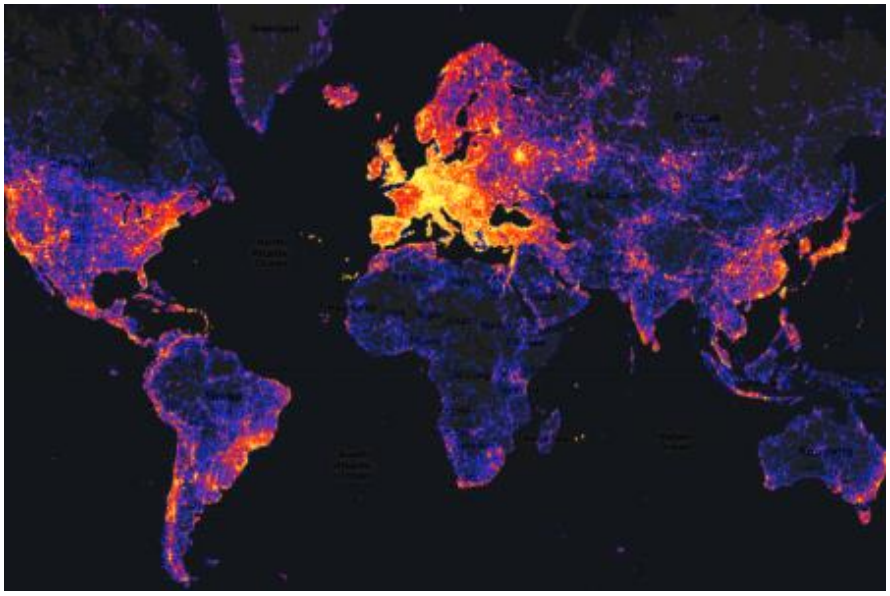

Image 7: A heat map of the most popular tourist locations. Brighter areas represent more popular locations.[19]

In this visualization, the level of popularity is the most important piece of information and it is represented by color or hue. The spatial dimension in this visualization is the map in the background, which is important to understand the visualization but it is not used to portray the actual information from the underlying data (the level of popularity of tourist locations). This is done by the different colors and the level of hue. This heat-map shows that other non-spatial dimensions can be favored as the dominant variable to show important information also. On the other hand, the heat-map visualizations would not be able to show us the information about popular tourist locations without the spatial dimension serving as the background. In this case a non-spatial dimension can be used to show information, but to some extent the visualization still relies on the spatial dimension to frame the information.

2.3 Case study: The datavisualization categorization in relation to traffic data-maps

Another way to explore if the principles of reduction and space are still sufficient to analyze more recent (emergent) datavisualizations, is by comparing these datavisualizations to the three datavisualization types that are based on the principles of reduction and space (Manovich 2011, 37; Friendly and Denis 2009, 2). These categories are information visualization, scientific visualizations and information design. In this section the traffic data-maps from the Sensor City will be used as a case and will be put in relation to this categorization to illustrate how emergent datavisualizations would relate to these categories and which problems would occur when doing so.

2.3.1 The traffic data-maps from the Sensor City in Assen

The traffic data-maps originate from the Sensor City project, which is an experiment to better manage the dynamic traffic patterns in Assen with the help of traffic data. The experiment attempted to measure different kinds of data in the city of Assen using approximately 200 sensors in the city center and numerous others located in mobile phones, navigational systems and tablets (Noordegraaf, Jonkers, de Kruijf 2014, 5). The data from these sensors was used to improve mobility during busy events and help the development of other technologies by sharing the collected data. The latter causing many organizations (public and private) to partner up with the Sensor City to get access to this sensor data. The creator of the traffic data-maps, Dat.Mobility, is one of them. The traffic data-maps from the Sensor City are used as an example in both this chapter and in chapter three. Traffic data-maps like those from the Sensor City, are good examples of emergent datavisualizations because they have characteristics that are not found in more traditional datavisualizations. For example, traffic data-

[19] Image can be found at http://www.digidrift.com/tourist-heat-map/

maps work with dynamic, always changing and up to date data that is visualized continuously onto a digital geographical map. This means that the data that is visualized needs modern technology to be able to show this near to real-time traffic information. This is an example of a feature that is not present in the more traditional datavisualizations, but is present in certain emergent datavisualizations.

As stated in the introduction, Bonnie Brennen mentions that a textual analysis is the analysis of a 'text', but this does not mean that this has to be an actual 'text'. In the context of a textual analysis, a text is more than just a composition of letters, words and sentences. A text can refer to any cultural artifact that has impacted our society (Brennen 193). Brennen states that "when we do textual analysis, we evaluate the many meanings found in texts" (Brennen 2012, 194). In this case, the traffic data-maps are seen as 'texts', which are analyzed to understand the different meanings and elements within the text. German sociologist and critical theorist Siegfried Kracauer maintained that "the goal of textual analysis (which he initially called qualitative content analysis) was to bring out the entire range of potential meanings in texts" (Brennen 2012, 194). "In textual analysis, researchers' theoretical perspectives can inform the type of textual analysis they use, as well as the types of questions that they ask" (Brennen 2012, 197). In this research the textual analysis is shaped by the explorative research of the (in)sufficiency of the principles of reduction and space, and the analytical properties of the newly introduces principle of abstraction and figuration.

2.3.2 Information visualization, scientific visualizations and information design

Manovich points out that "for some researchers, information visualization is distinct from scientific visualization in that the latter uses numerical data while the former uses non-numeric data such as text and networks of relations (Chen 2005). Personally, I am not sure that this distinction holds in practice" (Manovich 2011, 37). Meaning that these three categories are not meant to be set or determining, but that they have to be looked at as fluent and guiding (2011, 37). Within different disciplines, different meanings of the visualization types are used, making it even more valid to look at these categories as fluid instead of set. For this reason I will not try to "fit" emergent datavisualizations like traffic data-maps into one of these categories, but rather use these categories to explore how different elements of these categories do, or do not apply to them. This will result in a better understanding of the working of principles of reduction and space.

The best way to elaborate on the three visualization categories, is to point out their differences and similarities. The first distinction can be made between information visualization and information design (2011, 38). The most important difference between these two, is that the latter works with information and the former works with data. Lev Manovich states that the data that information design uses is already structured and therefore fixed to having a certain form (the data on its own can

already be seen as information). The goal in information design is to visualize this already existing structure and not to create it. Most infographics[20] are good examples of this. Here the already existing information is being represented by images and it is not obtained by creating these images with data. Information visualization (which can be considered as actual datavisualization) works with data that still needs to be visualized to reveal otherwise unseen patterns and structures (Manovich 2011, 38; Grinstein &Ward 2002, 21). So here the structure is determined by the data. A network visualization is a good example of this, since the structure is determined by the data and the information is based on the appearance of that structure.

In the case of the Sensor City, the type of datavisualization (a data-map) that helps to improve insight into mobility patterns is also based on data. As can be seen in image eight, the structure that emerges when the data gets visualized is dependent on the infrastructure of the city and the location of the sensors.

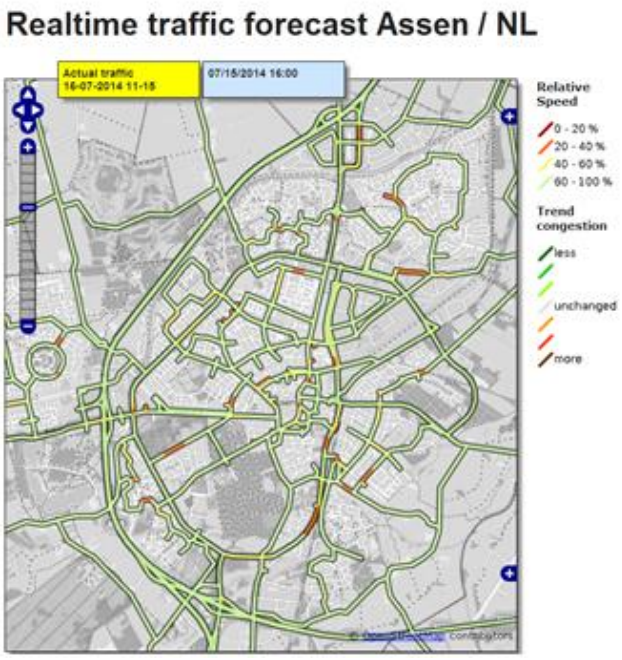

Image 8. This map portrays the real time current speed of the vehicles passing through Assen.[21]

This means that the data cannot change the geographical appearance of the map (the structure). Also, the data cannot be considered information unless it is layered onto the map. It is necessary to visualize the traffic data in order to obtain information from it, meaning that the information is not

[20] Examples and explanations of infographics can be found here: www.frankwatching.com/dossiers/infographics/.
[21] Image provided by Dat.Mobility.

present before the visualization process, which excludes traffic data-maps from being a form of information design.

The distinction between information visualization and scientific visualization is quite similar to the distinction between information visualization and information design, because in both cases the main difference is the á priori determined layout of the visualization. The difference lies in the way scientific visualizations do not visualize a layout or a structure, they use other visual dimensions than the spatial one to visualize data onto an already exiting layout, but the layout is still determining to the visualization. “Since the layout of such visualizations is already fixed and cannot be arbitrarily manipulated, colour and/or other non-spatial parameters are used instead to show new information” (Manovich 2011, 41). In information visualization, the other visual dimensions are mostly used to categorize or group information, but the spatial dimension remains the most important in the visualization process. According to Manovich this is the case when there is no á priori layout or structure in the data (Manovich 2011, 41), which makes the construction of the spatial dimension part of the visualizing process. So the structure of the visualization is determined by the data. This is different from scientific visualization, because the structure in this type of datavisualization is already present before the visualization process takes place.

In the case of the traffic data-maps, they would not be placed within the category of information visualization, but in that of scientific visualization. Information visualization specifically uses data to create a previously undetermined structure, but traffic data-maps have an already predetermined layout, namely the unchanging geographical background.

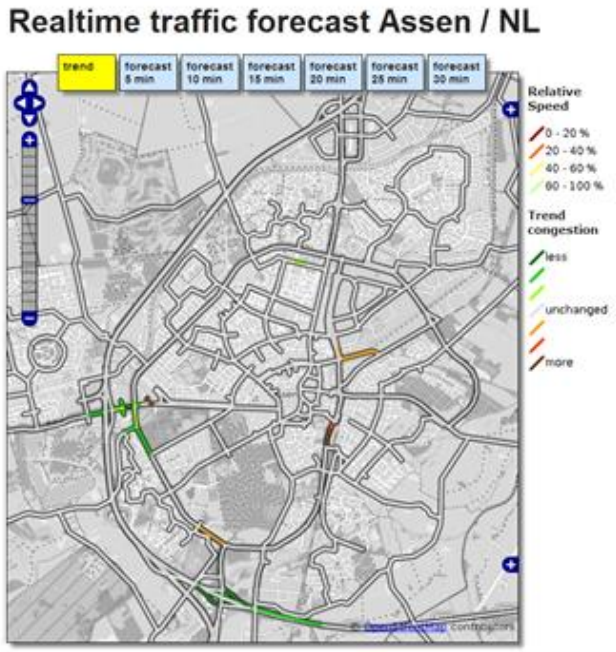

Image 9. A traffic data-map that shows a prediction of the locations where traffic speed will increase/decrease in the next half hour. [22]

The geographical background binds the other data to this structure, which means that traffic data-maps already have a structure prior to the visualization process (see image eight and nine). Data is used to enrich the map with information on mobility patterns, but won't change the geographical background. In the case of the traffic data-maps, color is more important than the spatial variable since it is the color that enables new information to be shows. The spatial variable is a guiding factor in the visualization, which is also important but is not able to show new information on its own. The principle of space therefore does not apply to traffic data-maps. With that said, the traffic data-maps from the Sensor City show the most resemblance to a form of scientific visualization.

When comparing the three categories discussed above, the fact that these categories are based on the principles of reduction and space becomes ever more evident. Every category uses a form of reduction to create a visualization and both information visualization and information design seem to favor the spatial variable when presenting the data in the visualization. Scientific visualization does not use the spatial variable to show new information, but the spatial variable is still determining in the datavisualization since it cannot be changed. In the case of the traffic data-maps from the sensor city, the reduction of data is a fact. It is needed to create insight into mobility patterns that consist out of data from many different sources (sensors, mobile phones, etcetera). This principle is therefore valid in the case of the traffic data-maps. But the principle of space is not, since the spatial variable is not dominant in the visualization for showing the important information. An interesting conclusion that can be drawn from this, is that the distinctions between visualization types are mostly based on the appearance and use of the spatial variable (the principle of space). This variable has been dominant during the major part of the datavisualization history, but as Manovich argues, is starting to be replaced by non-spatial variables like color and intensity (like the heat map shown in image seven) (2011, 40-41). In the traffic data-maps this lessening of the dominance of the spatial variable also becomes apparent, since color is the main variable that shows new information in this datavisualization.

2.4 Limitations of the principles of reduction and space

The exploration of the principles of reduction and space and analysis of the case using the datavisualization categories, reveals that both principles have some limitations when seen in relation to emergent datavisualizations. The principle of reduction is still useful but other methods can create datavisualizations that ensure no loss of original data. Manovich's direct or media visualization is an example of this (2011, 41). The principle of space is still important but no longer the only variable that

[22] Image provided by Dat.Mobility.

can show new information. Emergent datavisualizations do not necessarily just use the spatial variable to show important connections, contrasts, etcetera.

Because of emergent datavisualizations we are able to create visual meaning through datavisualization in more ways than before, making us realize that not just the methods to visualize data can change, but the way we interpret the visualizations can too. One could say that emergent datavisualizations have enabled us to extract a different kind of meaning from all datavisualizations, one that is based on the interpretation of the image itself. This type of information might have been less relevant in traditional datavisualizations, since datavisualizations like bar graphs do not need a lot of extra (visual) context to be understood (although there are exceptions[23]). They rely heavily on the principles of reduction and space and do not need the extra (visual) context to become understandable. Now that more visualizations leave the traditional forms and methods of datavisualization, context is added to make the image understandable. Because of this I am not convinced that the principles of reduction and space have to be replaced (in contradiction to Manovich his argument), but that they need additional academic principles to help give more meaning to the datavisualizations that lean much less on the principles of reduction and space.

Another outcome of this exploration therefore is the lack the ability to expand the analysis into the interpretation of the created datavisualization. By using just the two principles as a guideline, no attention was paid to the meaning of the created image. The only relation that the principles of reduction and space can explain, is the one between the data and the visualization. The relation between the visualization and the interpretation remains nonexistent. It would be useful to better understand how the visualization influences our interpretation of this information. For example, do we extract information from an abstract image differently than we would a figurative image, and how does this affect the meaning of the visualization? Having some knowledge about the way a datavisualization is interpreted based on its visual appearance, could therefore be helpful to understand how information from the datavisualization is extracted by viewers.

To explore and understand the visual, interpretable elements in emergent datavisualizations, it is useful to 'dissect' the image into different layers so every layer can be evaluated separately and in relation to each other. Within the humanities the field of semiotics provides useful theory that can help to break up the datavisualization into different layers, which is why semiotics and in particular Peirce's Sign Theory and the abstraction/figuration distinction will form the base of the proposed principle of abstraction and figuration. In addition to semiotics, the idea of an image being objective

[23] For example, this statistical map that portrays the losses suffered by Napoleons army in the Russian campaign of 1812, needs additional information to understand the meaning of the different elements in the visualizations. https://www.edwardtufte.com/tufte/posters

will be used to gain a deeper understanding of the relation between abstract and figurative objects in emergent datavisualizations, rekindling the discussion on the need for datavisualizations to be objective. In the following chapter I will introduce this new principle as a reaction to the lacking ability to interpret the visual appearance of emergent datavisualizations using only the two known principles. Additionally, I will mark the importance of objectivity in relation to the interpretation of datavisualizations when using this new principle.

3. A new principle to analyze emergent datavisualizations

In chapter two it has become apparent that the principles of reduction and space have some limitations in relation to the interpretation of emergent datavisualizations. This chapter will explore how semiotics can help to analyze the visual appearance of emergent datavisualizations by answering the second sub question: *What other information can be extracted from emergent datavisualizations by reaching out to the Abstraction and Figuration distinction and Peirce's Sign Theory?* In reaction to the limitation of the principles of reduction and space, the principle of abstraction and figuration is proposed by using the abstraction/figuration distinction (The Art Story Foundation 2014) as a starting point and Charles Peirce's theory of the sign (1974 [1931-1958], 243, 263) to further explore the abstraction/figuration distinction. Secondly, the reason why objectivity is related to this new principle will be discussed.

Semiotics is often used to analyze 'texts', which includes anything that can be seen as a cultural artifact (Brennen 2012, 193). The semiotic view helps to interpret and understand any text in relation to its cultural context. The overarching distinction of abstraction and figuration is well-known within art history (The Art Story Foundation 2014), but can also be relevant when it comes to analyzing datavisualizations like data-maps. By using the principle of abstraction and figuration, a datavisualization can be placed within the abstraction/figuration distinction by determining if the representation of data is abstract or figurative. Whereas abstract art can be seen as "relating to or denoting art that does not attempt to represent external reality, but rather seeks to achieve its effect using shapes, colours, and textures" (Oxford online dictionary), figurative art represents "forms that are recognizably derived from life" (Oxford online dictionary). Some datavisualizations might be a complete abstract representation, like the datavisualization in image ten. And others could be almost completely figurative, like the datavisualization in image eleven. Being able to determine if a datavisualization is abstract or figurative, enables the visualization to be further divided into elements.

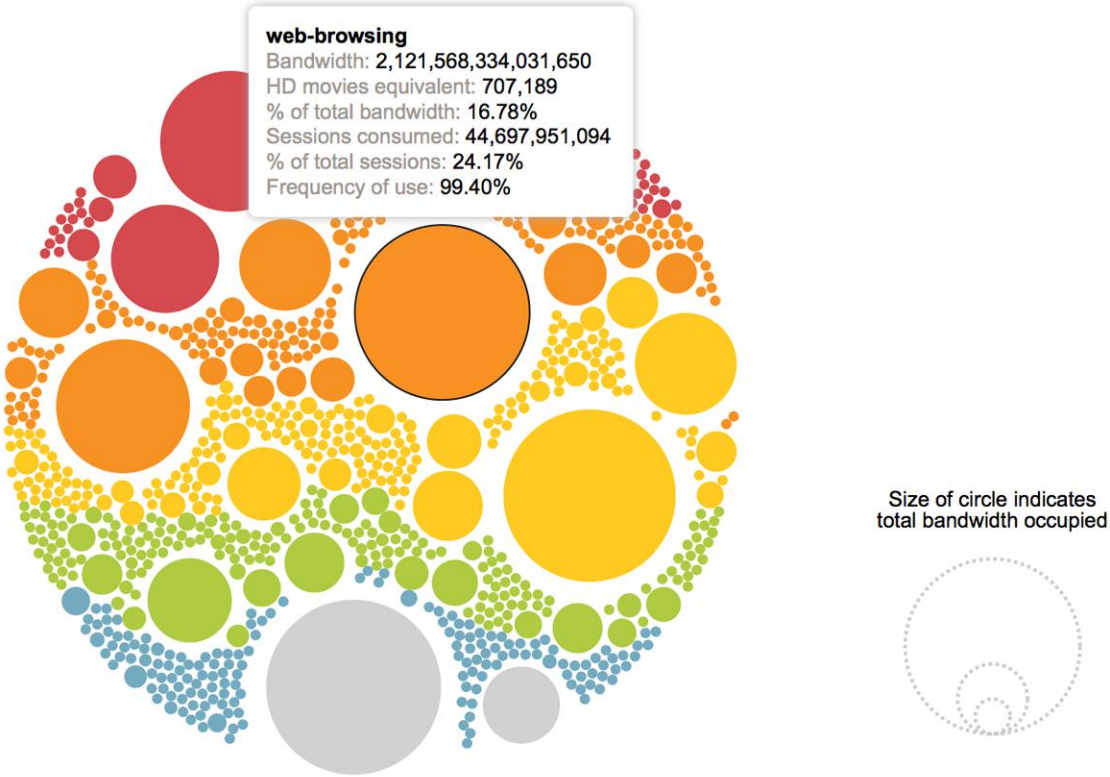

Image 10. A datavisualization showing the amount of web browsing by the relative sizes of the circles.[24]

The example in image ten shows the amount of web browsing by the relative size of the circles, which is an abstract representation of web browsing since this does not visually compare to web browsing in any way. The only way to know that this datavisualization is meant to show the amount of web browsing in relation to others, is by reading the information provided in or around the datavisualization. Since there is no visual link to this information, the datavisualization can be considered an abstract representation of the underlying data.

An example of a datavisualization that has mostly figurative elements is a highly visual datavisualization (Dirksz 2013, 3), like the one in image eleven.

[24] Image can be found at http://www.threestory.com/images/blog/pan_viz1.png.

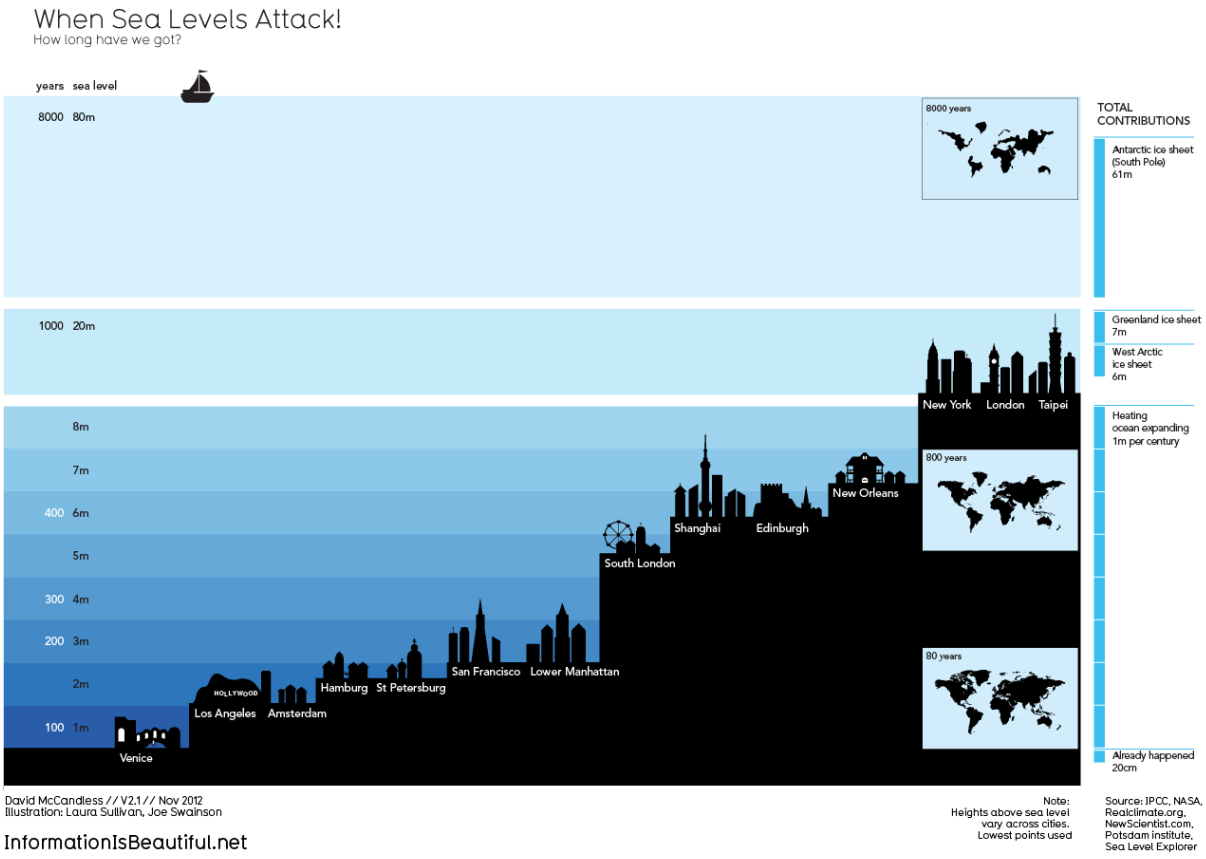

Image 11. A highly visual datavisualization titled 'When sea levels attack' visualizing the consequences of rising sea levels for the planet and in particular for a selection of cities shown on top of the bars. The vertical axis shows the sea level in meters and the years it will take until that level is reached. The horizontal axis shows some of the cities that will be flooded when a certain sea level is reached.[25]

In this datavisualization the represented objects are directly shown as the actual visual form of those objects, but incorporated into a more traditional bar-graph design. Because the datavisualization is visually referring to its objects by representing their 'real-life' form, this datavisualizations consists out of mostly figurative elements that need less explaining to understand what they represent.
By knowing a datavisualization is abstract, figurative or both, the relations between the represented and the object that is supposed to be represented can be further explored using Peirce's theory of the sign (1974 [1931-1958]).

Peirce's theory of the sign consist out of many different theories, concepts and ideas on the relation between the sign, the object and the *interpretant*. When defining different characteristics between datavisualizations, the relationship between the sign (or the *representamen*) and its object (the actual thing that the sign is representing) is the most relevant in relation to the analysis of the visual

[25] 'When sea levels attack' made by Laura Sullivan and Joe Swainson. Visualization available at http://www.informationisbeautiful.net/visualizations/when-sea-levels-attack-2/.

appearance of emergent datavisualizations. Peirce further divides the sign within the sign-object relation (the representation) into the icon, the index and the symbol, and states that each of them have a different relation to the object (the actual "thing" that the sign represents). In the relationship of the sign with its object, an icon is a sign that refers to its object by resembling it (like a photograph) (Peirce 1958, 99). The essential aspect of the relation between an icon and its object is their similarity. An index refers to its object via a causal link between the sign and its object. Smoke for example, is an index of fire. The object affects the sign. Finally, symbols refer to their object by virtue of law, religion, convention, etcetera. There is no similarity or causal link suggested in the relation between the sign and the object. For example consider the word 'tree'. Only when someone knows what this composition of shapes we call letters means, he or she will connect the word 'tree' to the actual object, the tree (Eco 1984, 26-39). If this knowledge is not present, then there is no other link between the sign 'tree' and the object 'tree' that could make this connection apparent.

Using Peirce's theory of the sign allows for a more in depth exploration of the previously determined abstract of figurative datavisualization. For example, the elements in the highly visual datavisualization in images eleven and twelve can be further analyzed using the theory of the sign.

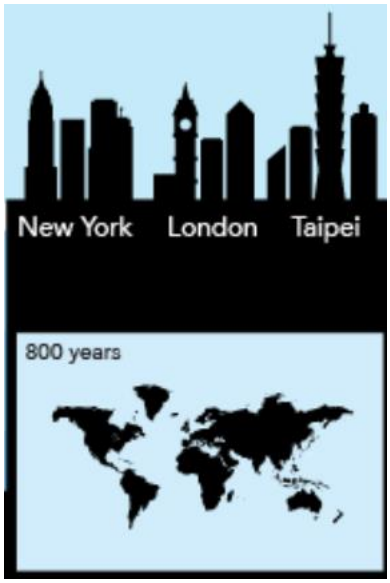

Image 12. Section of the 'When sea levels attack' visualization showing the world map and some cities.

In this datavisualization there is not just one type of relationship between the sign and object. When looking at the visualization as a whole in image eleven, it seems to contain both iconic as indexical signs. The height of the bars for example are determined by the data, which suggests a causal link between them and making this element have a sign-object relation that is considered an index. The world map in image twelve can be considered an icon, because it resembles the actual map of the world. The buildings on top of the bars are also similar to the way we would picture a building from

afar and the boat on the top left is a similar shape as the boats we know from real life. Another function of the boat is that it helps us understand the blue color underneath it as water[26]. By using the principle of abstraction and figuration, it becomes apparent that a datavisualization can be seen as a layered image that consists out of different element. Understanding how these different elements relate to the objects they represent in the datavisualizations, is not possible using just the principles of reduction and space. The principle of abstraction and figuration therefore creates valuable new insights for the interpretation of datavisualizations.

Concluding this section, in addition to the principles of reduction and space where reducing the amount of data and the favoring of the spatial variable are key, the principle of abstraction and figuration can be used to explore the relation between a datavisualization and the objects it represents. Since the principle of reduction and space lays the emphasis on the relation between the datavisualization and its representation or meaning, it could a valuable addition to the principles that are already in place. Especially when the interpretation of a datavisualization has to involve more than just the relation between the data and the visualizations, the principle of abstraction and figuration can help to analyze if the manner of visualization effects the way we interpret its meaning.

3.1 Objective datavisualizations?

A question that follows from the introduction of the principle of abstraction and figuration, is one concerned with the objectivity of datavisualizations. If a datavisualization is a way of structuring and transforming data in such a way that we can acquire some sort of information from this data, does the visualization therefore need to be an objective reflection of the analyzed data? Is that even possible? To further explore the this relation, this section will be dedicated to discussing if it is important to take objectivity into account when analyzing datavisualizations. Firstly it is important to mention is that objectivity means different things within different disciplines and therefore needs to be explained prior to relating it to the interpretation of datavisualizations. For this reason, the work of Lorraine Daston (1992) and Peter Galison (1992, 2010) on the changing notion of objectivity and the work of Danah Boyd and Kate Crawford (2012) on the interpretation of data in relation to objectivity will be discussed first. By discussing their work it will become clear why the term objectivity is hard to grasp and it will clarify the validity for asking in what way objectivity plays a role in the analysis of datavisualizations using the proposed principle of abstraction and figuration.

Looking back on the history of scientists striving for objective representations and theories, there are several types of objectivity that have influenced the way we think about the term today. In Lorraine

[26] Of course we have to take into account that these meanings are strongly influenced by one's cultural background.

Daston and Peter Galison's article 'The Image of Objectivity' (1992) these different types of objectivity are described as different elements within objectivity. Daston and Galison claim that the (in 1992) modern way of thinking about objectivity doesn't seem to integrate these elements into a single (if layered) concept, but tends to mix them together.

> "This layering accounts for the hopelessly but interestingly confused present usage of the term *objectivity*, which can be applied to everything from empirical reliability to procedural correctness to emotional detachment." (Daston and Galison 1992, 82).

So what are these layers and how do they contribute to understanding what objectivity means today? In 'The Objective Image' (2010) Galison gives a brief overview on the history of the objectivity of scientific images like graphs and representations of natural phenomena, to elaborate on why objectivity has become a somewhat confusing term. He states that at the beginning of the 1820's the common procedure to create an objective view of any object, was through general depiction (also known as 'typus'). This meant that the object of study had to be depicted in a way that is true to nature (Galison 2010, 8-10; Daston & Galison 1992, 81). Around the same time, mechanical objectivity was also being used to obtain an objective view. Mechanical objectivity is obtained when "nature could be transferred to the page without intervention or interpretation" (Galison 2010, 8). In other words, the scientist had to let nature speak for itself and could not alter anything to make the object look better or more understandable. The transition from nature to paper had to be as direct as possible. The last form of objectivity that Galison describes is judgmental objectivity, which begins to displace mechanical objectivity around the 1920's. Judgmental objectivity is based on the "eye of the expert" (2010, 9) and allows him or her to "alter, interpret and select figures" (9) in order to create a clear representation of the object. As a result, objectivity has become a mixture between these described types of objectivity, which is the reason why Daston and Gallison state that the term has become confusing (1992, 82). In this thesis, I will consider objectivity as a reflection of 'nature', in the sense that nature is an original form that can be considered as the unaltered truth. In the case of datavisualizations, the raw data prior to the analyses can be seen as a direct reflection of the measured truth. This still leaves the question if datavisualizations can be considered objective because the raw data are transformed into a visual representation of data.

Two researchers that have been exploring the relation between datavisualizations and objectivity, are media and social scholars Danah Boyd and Kate Crawford.[27] In 'Critical Questions for Big Data' (2012),

[27] Danah Boyd is a Principal Researcher at Microsoft research, founder of Data and Society Research Institute, Visiting Professor at New York University's Interactive Telecommunications Program and a faculty affiliate at Harvard's Berkman Center. Kate Crawford is also a Principal Researcher at Microsoft Research, a Visiting

Boyd and Crawford discuss some critical notes about Big Data research. One of the arguments that Boyd and Crawford make is that data does not speak for itself but that it needs to be analyzed and interpreted by humans to become information (666). An important statement considering that datavisualizations are usually the result of structured data. They also state that "interpretation is at the center of data analysis"(668). The data is not self-explanatory which also suggests that data research a subjective process, instead of an objective one (667). The process of collecting, analyzing and interpreting data is mostly initiated by humans, meaning that there is always a chance of human error or choice to leave certain data out of the visualization. It makes the process of data analysis a very delicate one that could potentially provide wrong insights when not processed correctly (670). This is something to consider when analyzing datavisualizations, since the possibility of human error can easily affect the appearance of the datavisualization and the conclusions that are drawn from it.

So when discussing the objectivity of datavisualizations, eventually a controversy arises. On the one hand, we trust datavisualizations be a readable reflection of the underlying data. But on the other hand, since almost every datavisualization can be considered to be the result of the combination of raw data[28], an algorithm and the interpretation of the data scientist, finding a purely objective (as in: true to nature) datavisualization that can also be considered informative is very hard in my opinion. In his work 'Visual Explanations: Images and Quantities, Evidence and Narrative' (1997), Edward R. Tufte emphasizes that the principle of design must replicate the principles of thought, so that "the act of arranging information becomes an act of insight" (Tufte 1997, 9). Stating that the structuring of data is "an act of insight" (1997, 9) instead of the result of an algorithm, is very contradictive to a supposedly objective outcome of datavisualization, which further contributes to the friction between objectivity, the interpretation and ending up with a reliable and truthful datavisualization.

During the visualization process a lot of choices are made in order to create a meaningful and understandable datavisualization. Less human interference could therefore result in a more objective datavisualization, but could make the outcomes less understandable for humans, show incorrect information or connections, or show something that is understandable but does not make sense in relation to what one wants to know about the data. With this in mind, it seems that human interpretation is important in creating a meaningful datavisualization, but it also enables room for human error and the choice to manipulate the data. Therefore a datavisualization is not necessarily objective just because it is created out of objective data. The data itself data might be objective, but

Professor at the MIT Center for Civic Media, a Senior Fellow at NYU's Information Law Institute, and an Associate Professor at the University of New South Wales.

[28] With raw data I mean data that is collected from a certain source and after this, not yet structured or manipulated in any way through human intervention.

the visualization of that data usually is not, because the data needs some form of manipulation to create an understandable datavisualization.

3.2 Case study: Using the principle of abstraction and figuration to analyze the traffic data-maps from the Sensor City

To briefly exemplify how the proposed principle of abstraction and figuration relates to emergent datavisualizations, this principle will be used to analyze the traffic data-maps from the Sensor City. This section will explore how traffic data-maps relate to being abstract, figurative or both by breaking up the image into elements and opening up the discussion on datavisualizations as being layered images. These elements are placed within Peirce's theory of the sign (1974 [1931-1958]), which will help to identify the relationship between the appearance of each element (the sign) and what it represents (the object). By looking at the traffic-data maps as texts from a semiotic perspective, I am able to critically reflect on what they represent and how they represent it (Gripsrud 2006, 39-40).

When trying to place traffic data-maps within the abstraction/figuration distinction (The Art Story Foundation 2014), it seems that on the one hand the traffic data-maps are mostly abstract, since the image is not a direct reflection derived from 'life' and use shapes and colors to create its informative effect (see image thirteen).

Image 13: Traffic intensity map of the city Assen, created by using the data collected by Dat.Mobility trough the Sensor City project. The scale on the right shows the colors that match the intensity levels.[29]

[29] Image provided by Dat.Mobility.

On the other hand, even though a traffic data-map does not resemble anything we can see in 'life', it does represent a geographical landscape, something we can easily relate to as being true. Traffic data-maps are abstract since they are not a reflection of a realistic object, but can also be considered figurative since they do represent something realistic. This twofold of abstraction and figuration can be considered a necessity when it comes to traffic data-maps, since they can only hold information when the data is layered on a recognizable geographical background. Therefore this background has to directly represent a certain geographical location for the data to become information on traffic intensity for example (see image thirteen). Traffic data-maps can thus be seen as layered images consisting out of both a figurative and an abstract layer.

When trying to place traffic data-maps within Peirce's theory of the sign (1974 [1931-1958]), it becomes clear that analyzing the visualization as a whole becomes challenging since traffic data-maps both have figurative and abstract layers in them. The figurative layer can be seen as the map on the background, which directly refers to a certain location in real life, and the abstract layer is the visualized data since it is an abstract representation of a certain traffic situation (see image fourteen).

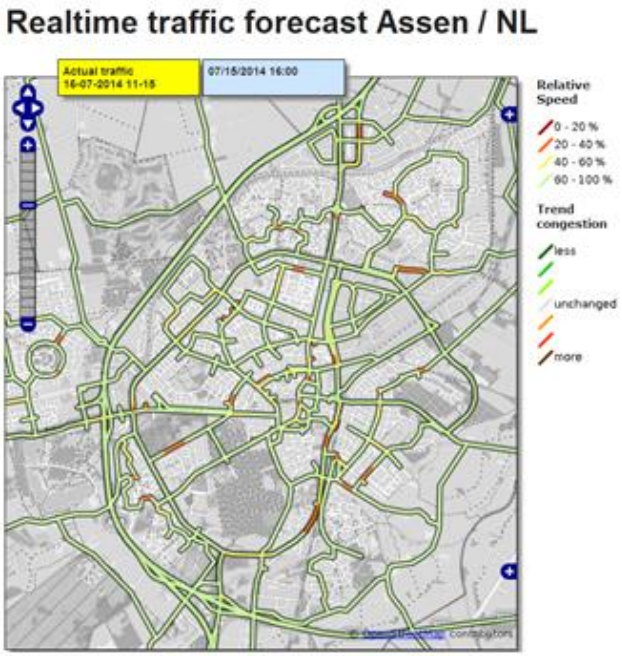

Image 14. This map portrays the real time current speed of the vehicles passing through Assen.[30]

To be able to understand the relation between the object and the sign in traffic data-maps, dividing the image into two layers becomes inevitable. In terms of Peirce's theory of the sign (1974 [1931-1958]), the layer that shows the background map can be seen as an icon, since it is the most figurative

[30] Image provided by Dat.Mobility.

part of the image and refers to a geographical location that we recognize from 'life'. The geographical map resembles the appearance of the actual geographical location, which makes this part of the visualization an icon. The added layer of data (the colors in the specific areas) is an abstract representation of the intensity of traffic flows, but have no figurative link to resemble actual traffic. The added data layer can therefore be seen as a symbol for traffic, since it represents a constructed link between the data and the visualization that needs some form of explanation in order to understand.

In conclusion, by using the principle of abstraction and figuration it has become apparent that traffic data-maps are neither completely abstract nor completely figurative, but are a combination of these two. When looking at the relationship between the sign and the object, the two layers both have a different connection to the objects they represent. The figurative map in the background resembles a certain geographical location, so in terms of Peirce's sign theory (1974 [1931-1958]) this layer has an iconic relationship with the object it represents. In contrast, the layer of data is abstract and only connects to its object through a certain convention, one that the maker of the datavisualizations gives it. So the data-layer is a symbolic representation of the traffic flow it represents. Another conclusion that can be drawn from analyzing traffic data-maps as being a layered image, is that even though these two layers seem to be very different and have different properties, they are only informative when seen as one merged image. The background cannot show traffic information without the data and the data layer cannot show information on traffic flows without the map in the background to refer to the geographical location. This also reflects back on the principle of space, which is the favoring of the spatial variable. In this case the spatial variable is not showing the new information, since this is done by color, but is used as a guiding factor to understand what the colors mean. The principle of space is therefore still useful, just not in terms of the spatial variable being dominant but playing a different role in the datavisualization.

3.2.1 What about objectivity?

A question that follows the exploration of how datavisualizations represent the objects they refer to, is concerned with the necessity of the image having to be objective for the information shown to be a true reflection of the data. Since traffic data-maps show the traffic flow on the streets of a certain location that is not a complete figurative reflection of the actual event, the question to what level objectivity is a requisite in the creation of these images therefore becomes valid. The fact that the background map is recognizable as a geographical representation of a place does not mean that it is objective, it means that the location is recognizably represented.

When looking at the two layers of traffic data-maps, the figurative background facilitates the connection between abstract data and the object that it represents (which is traffic flow) (see image

fifteen). Secondly, it makes abstract data less abstract because it enables the viewer to connect the correct interpretation to the data. The data is shown through the use of different colors, but in order to know what these colors mean, the creator of the datavisualization has to assign a certain meaning to each color. Also, in the creation of the datavisualization the creator has to choose which data will be shown on the map and which data will be left unseen. In order to distinguish between these types of information, the creator has to determine which data gets used where. In image fifteen the information is the predicted traffic speed, but in image fourteen this is the real-time traffic speed. The visualization is based on data collected from the sensors, but created by using the correct interpretation skills of the data scientist.

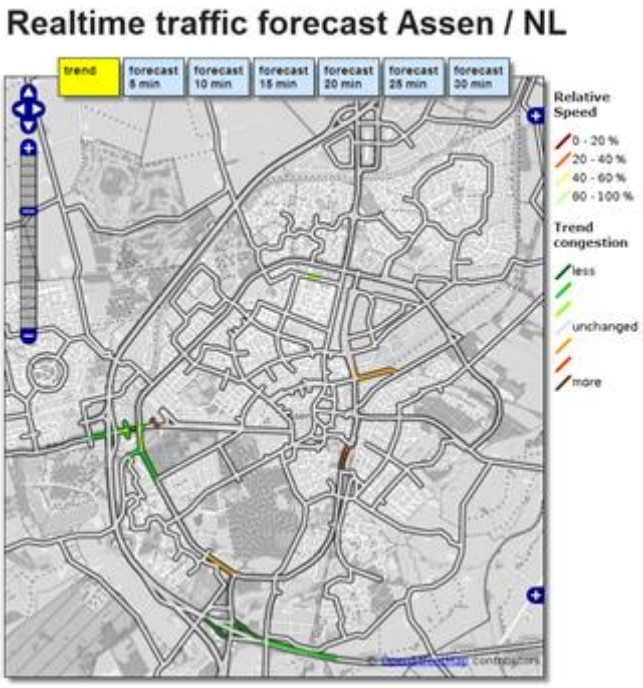

Image 15. A traffic data-map that shows a prediction of the locations where traffic speed will increase/decrease in the next half hour. [31]

When creating a datavisualization, the creator always has some sort of influence on how the data will be interpreted because of the choices that he or she makes in preparing and visualizing the data. So a datavisualization does not have to be objective to be meaningful or true, it has to be a true representation of the trend that it shows (Galison 2010, 8).

The conclusion that can be drawn from the controversy between datavisualizations and objectivity, is that objectivity does not seem to be a necessity for datavisualizations to reflect a truthful

[31] Image provided by Dat.Mobility.

visualizations of the underlying data. Since a direct, objective representation of data would make the creation of a datavisualizations that is anything other than the raw data itself impossible. The choices made in the visualization process are necessary to create the datavisualization and will inevitably be visible in the visualization. This will result in viewers forming a certain interpretation so they can extract certain conclusions from the data that would be impossible to do without the data being visualized. In other words, human intervention and interpretation is necessary for the creation of a correct and meaningful datavisualization. So when analyzing a datavisualization, knowing the effects of the choices made in the visualization process on the outcome of the datavisualization, have to be taken into account when conclusions are drawn based on the meaning of that datavisualization. Having some knowledge of the visualization process could therefore be of great help when trying to understand and interpret the meaning of the visualization.

4. Conclusion and Discussion

The objective of this thesis was to add insight on the analysis of the external physical form of datavisualizations to the field of Information Visualization and its theoretical framework by drawing on the principles of reduction and space and theory from semiotics. The main research question of this thesis was: *How can the perspective of semiotics help to further understand the external physical form of emergent datavisualizations and help to continue building the Information Visualization theoretical framework?*

I have answered the main research question by answering two guiding sub questions. The first sub question was: *What information can(not) be extracted from emergent datavisualizations by using the already known principles of reduction and space?* This question has been answered in chapter two by discussing the principles of reduction and space and relating them to modern datavisualizations in a case study of the traffic data-maps from the Sensor City. From the textual analysis of the traffic data-maps in relation to the datavisualization categories, it became clear that especially the use of the spatial variable as the dominant variable seemed to be untrue for traffic data-maps. In the traffic data-maps from the Sensor City, the main variable to show the important information was color and not the spatial dimension. Also, today there are other methods that ensure no loss of original data, which means that those datavisualizations do not follow the principle of reduction. Manovich's direct or media visualization is an example of this (2011, 41). In conclusion it has become apparent that the principle of space is still important but no longer dominant in emergent datavisualizations like traffic data-maps. The upcoming importance of other non-spatial dimension has played a significant role in the complex datavisualizations that are possible today, since many of them incorporate more than just the spatial dimension to show important connections, contrasts, etcetera. Unlike Manovich's view, from the analysis it has become clear that the principles of reduction and space are still useful, but

cannot extract meaning or give meaning to the datavisualization as an image. Not meaning that they can be replaced. However, the analysis of datavisualizations would certainly benefit from an approach to also analyze the meaning of the datavisualization which will otherwise remain unseen.

The second sub question was: *What other information can be extracted from emergent datavisualizations by reaching out to the Abstraction and Figuration distinction and Peirce's Sign Theory?* This question was answered in chapter three by using theory from semiotics and Peirce's Sign Theory to propose the principle of abstraction and figuration. This principle is useful when exploring the relation between a datavisualization and the objects it represents. By using this principle, it can be determined if a datavisualization is abstract or figurative and to further explore the relationships between the visually represented data and the objects that it represents. The datavisualization can be interpreted as a whole, but the principle can also be used to extract meaning from the different elements within the datavisualization. An important addition to the principle of abstraction and figuration is concerned with the relation between the datavisualization and objectivity. Since during the visualization process many human choices are made in order to create a meaningful and understandable datavisualization, which are helpful but also create the possibility for human error. So when analyzing a datavisualization, knowing the effects of the choices made in the visualization process on the outcome of the datavisualization, have to be taken into account when conclusions are drawn based on the meaning of that datavisualization.

In conclusion the way to interpret and understand the external physical form of emergent datavisualizations, is not by replacing the old principles of reduction and space (as suggested by Manovich), but by defining clear academic tools to interpret them. This conclusion is similar to that of Purchase et al.: "Investigating theoretical approaches used in other disciplines, and their relation to Information Visualization, is an obvious way forward, and can provide a useful way for researchers in the area to present, discuss and validate their ideas" (Purchase et al. 2006, 62). Just like the outcomes of this research, Purchase et al. concluded that there is a lack of solid theoretical analyses in Information Visualization and that exploratory research like this will help to form the theoretical framework for the field (2006, 62). By going beyond the principles that are already there, it has become clear that much more can be said about the visual appearance of datavisualizations than just the relation between the datavisualization and the data when it comes to the external physical form of emergent datavisualizations. They hold more information than just the insights that are gained from the reduction of data or principle of space. This could mean that also the older or more traditional datavisualizations can hold interesting new insights when the principle of abstraction and figuration is used to analyze their visual appearance. Meaning that the introduced additional principle of

abstraction and figuration is not exclusively for the analysis of the newer kinds of datavisualizations, but could be used to analyze many datavisualization types. Emergent datavisualizations have therefore both created the need for more academic tools for the analysis of these divers datavisualizations, as well as opened our eyes to the narrow scope we used to analyze datavisualizations by only considering the principles of reduction and space in analysis.

Lastly, a few notes have to made on the execution of this research. Because I used one case to exemplify the proposed principle, generalizing the outcomes should not be one without additional research and reproducing the analysis becomes difficult. This research is meant to be a first step in creating more principles to understand emergent datavisualizations to add to the theoretical framework of Information Visualization. The choice to use a textual analysis was made because it is currently the most suitable method to analyze the external physical form of datavisualizations, since this enables research that is focused on the meaning of the research object. Also, the influence of culture on the appearance of different types of datavisualizations has not been taken into account in this particular study, making the possible effect that culture has on the interpretation of datavisualization an interesting subject for future research. A more urgent call for future research is to explore other theories that can be helpful to include in the analysis of datavisualizations, so that a clear theoretical framework for Information Visualization can be defined.

5. Bibliography

Boyd, Danah & Kate Crawford. 2012. Critical Questions for Big Data. In *Information, Communication & Society* 15 (5): 662 – 679. Routledge.

Brennen, Bonnie S. 2012. *Qualitative Research Methods for Media Studies.* New York: Taylor & Francis Routledge.

Chen, C. 2005. Top 10 Unsolved Information Visualization Problems. Proceedings of *IEEE Computer Graphics and Applications* 25 (4) (July–August): 12-16.

Daston, Lorraine and Galison, Peter. 1992. The Image of Objectivity. *Representations* 0 (40): 81-128.

Dirksz, Joëlle. 2013. *Highly Visual Datavisualizations: the defining and analysis of a new type of datavisualization.* Unpublished paper from Software Studies course. University of Utrecht, department of the Humanities.

Edward Segel and Jeffrey Heer. 2010. Narrative Visualization: Telling Stories with Data. *Visualization and Computer Graphics* 16 (6): 1139-1148.

Eco, Umberto. 1984. *Semiotics and the Philosophy of Language*. Bloomington, IN: Indiana University Press.

Friendly, Michael, L. and Daniel J. Denis. 2009. Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization. *Seeing Science: Today American Association for the Advancement of Science*. August 24th. http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf.

Galison, Peter. 2010. *The Objective Image*. Oration in the Occasion of Accepting the Treaty or Utrecht Chair and Utrecht University.

Grinstein, Georges G., and Matthew O. Ward. 2002. Introduction to data visualization. *Information Visualization in Data Mining and Knowledge Discovery* (1): 21-45.

Gripsrud, Jostein. 2006. Semiotics: Signs, Codes and Cultures. In *Analyzing Media Texts,* eds. Marie Gillespie and Jason Toynbee, 9-41. Maidenhead, UK: Open University Press.

IBM. 2011. *Stepping up to the challenge: CMO insights from the Global C-suite Study.* http://public.dhe.ibm.com/common/ssi/ecm/gb/en/gbe03593usen/GBE03593USEN.PDF

Keim, Daniel A., Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler. 2006. Challenges in visual data analysis. 2006. Proceedings of *Information Visualization* (IEEE) IV: 9–16.

Laney, Douglas. 2012. *The Importance of Big Data: a Definition*. Gartner. Retrieved June 21st 2012.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Manovich, Lev. 2011. What is Visualisation? *Visual Studies* 26 (1): 36-49.

Manovich, Lev. 2011. *Media Visualization: Visual Techniques for Exploring Large Media Collections.* San Diego: University of California, Visual Arts Department. http://manovich.net/content/04-projects/069-media-visualization-visual-techniques-for-exploring-large-media-collections/66-article-2011.pdf.

Manovich, Lev. 2013. The Algorithms of our Lives. *The Chronicle of Higher Education* 60 (16): 10-13.

Noordegraaf Vonk, Diana, Jonkers, Eline and Janiek de Kruijff. 2014. *Eindrapport Sensorcity Mobility.* Delft, NL: TNO.

Peirce, Charles S.. 1974 [1931–1958]. *Collected Papers*. Cambridge, MA: Harvard University Press.

Purchase, Helen C., Natalia Andrienko, T.J. Jankun-Kelly and Matthew Ward. 2008. Theoretical foundations of information visualization. In *Information visualization: Human-centered issues and perspective*, eds. by A. Kerren, J.T. Stasko, J. Fekete and C. North. Lecture notes in computer science, vol. 4950, 46–64. Berlin, Heidelberg: Springer-Verlag.

Stebbins, Robert A. 2001. *Exploratory research in the social sciences*. Vol. 48: Sage Publications.

Strecker, Jaqueline. 2012. Data Visualization in Review: Summary. *Evaluating IDRC Results: Communicating Research for Influence*. http://idl-bnc.idrc.ca/dspace/bitstream/10625/49286/1/IDL-49286.pdf.

The Art Story Foundation. 2014. Abstract vs. Figurative Art. *The Art Story*. http://www.theartstory.org/definition-abstract-vs-figurative-art.htm.

Tufte, Edward R. and Elizabeth Weise Moeller. 1997. *Visual explanations: images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Yin, Robert K. 2014. *Case Study Research: Design and Methods* (5th edition). London, UK: Sage Publications.

5.1 Images

Image 1 and 2: A Twitter networkvisualization. Created by author.
Created by Joëlle Dirksz in collaboration with Big Fellows B.V and the department VTH (Vergunningen, Toezicht and Handhaving – Licenses, Supervision and Enforcement) Gemeente Utrecht.

Image 3: A timepiece graph visualization of the mentions of Tony Blair and Grodon Brown in The Gardian from 1999 until 2008.
http://www.flickr.com/photos/blprnt/3347062507/in/photostream/.

Image 4: Traffic intensity map of the city Assen.
Created by Dat.Mobility.

Image 5: A bar graph that shows the growth and expected growth of the amount of data produced in the world. http://www.datamation.com/applications/how-vertical-markets-will-drive-big-data.html.

Image 6: A word cloud made by the author from the words in Lev Manovich's article 'What is visualization?'.

Image 7: An interactive and searchable heat map of the most popular tourist locations on the planet based on the number of "panaramio"[32] photo's per location.
http://www.sightsmap.com/.

Image 8: A datavisualization showing the amount of web browsing by the relative sizes of the circles. http://www.threestory.com/images/blog/pan_viz1.png.

Image 9: A highly visual datavisualization titled 'When sea levels attack' visualizing the consequences of rising sea levels for the planet and in particular for a selection of cities shown on top of the bars. Made by Laura Sullivan and Joe Swainson.
http://www.informationisbeautiful.net/visualizations/when-sea-levels-attack-2/.

Image 10: Section of the 'When sea levels attack' visualization showing the world map and some cities. http://www.informationisbeautiful.net/visualizations/when-sea-levels-attack-2/.

[32] Panaramio is a platform where travelers can share their vacation pictures: http://www.panoramio.com/.

Image 11: A traffic data-map that portrays the real time current speed of the vehicles passing through Assen. Created by Dat.Mobility

Image 12: A traffic data-map that shows a prediction of the locations where traffic speed will increase/decrease in the next half hour. Created by Dat.Mobility.

Image 13: Traffic intensity map of the city Assen. Created by Dat.Mobility.

Image 14: A traffic data-map that portrays the real time current speed of the vehicles passing through Assen. Created by Dat.Mobility.

Image 15: A traffic data-map that shows a prediction of the locations where traffic speed will increase/decrease in the next half hour. Created by Dat.Mobility.

5.2 Other references

Article on the five cases proposed by IBM that find solutions using Big Data: http://www-01.ibm.com/software/data/bigdata/use-cases.html.

Article on the variety of definitions of Big Data in different contexts by Gil Press of *Forbes*: http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#694325c821a9 .

Bol.com creating a user centered website with data: http://www.marketingfacts.nl/berichten/big-data-successen-voor-het-voetlicht-op-data-donderdag.

Bomenapp: http://www.bomenapp.nl/.

Dat.Mobility and the Sensor City mobility project: www.dat.nl.

Digital weather map example: www.buienradar.nl.

IBM Quote on big data: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Infographic examples: http://www.frankwatching.com/dossiers/infographics/.

Oxford online dictionary: http://www.oxforddictionaries.com/.

Statistical map of Napoleon journey in the Russian campaign of 1812: https://www.edwardtufte.com/tufte/posters.

Tag cloud example and explanation: http://nl.wikipedia.org/wiki/Tag_cloud.

Utrecht Data School: www.dataschool.nl.

Waterakkers interactive map: http://new-waterakkers.azurewebsites.net/Waterakkers/Content/helo_stuw.