Hochschule für Angewandte Wissenschaften Hamburg
*Hamburg University of Applied Sciences*

# AW1 Ausarbeitung

**Truong Vinh Phan**

**Data Visualization**

*Fakultät Technik und Informatik*
*Department Informatik*

*Faculty of Engineering and Computer Science*
*Department of Computer Science*

# Contents

# List of Figures

# Data Visualization

Truong Vinh Phan

March 4, 2014

# 1 Motivation

Almost all people have to deal with more or less data as part of their daily routines or activities. This is especially true today, where those data come from different sources and in various forms, and almost two-third of them are in the form of electronic information. The main contributing sources for data include social networks and media, which have recently gain a lot of popularity and are becoming common for everyone, open government/transparency movements which allow governments and organizations to publicize all sorts of data, such as the Open Data Hamburg [5]. In fact, in 2012 it was estimated by IBM that 2.5 exabytes of new data were generated per daily basis [4], which leads to the term “information pollution/explosion” by some noted scholars.

To human, a visual representation is often more effective than written text. The visual representation help us to illustrate concepts and ideas, that, if expressed verbally, would be very difficult or even impossible to be fully understood. Edward Tufte maintains that “excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency” [25]. This is possible, as Colin Ware states in his study that the human vision, which is virtually a massive parallel processor made of the eye and the visual cortex of the brain, provides the highest-bandwidth channel of all the human senses. [29]

Perception (seeing) and cognition (understanding) are considered closely correlated. A general rule for the visual system is that when data is presented in certain ways, their patterns can be readily perceived, which would be otherwise difficult or impossible if presented in some other ways. Based on these knowledge, principles for displaying information are derived so that the important and informative patterns will stand out in a sea of data.

# 2 Foundations of Visual Representation

## 2.1 General Visualization Design Principles

As with all kind of designs, the most basic questions to ask when doing design work are the target (who), the purpose (why) and the method (how). A good design will convey information fully, while with a bad design, in which essential details are left out or distorted, the result is catastrophic.

- Show the data*:* Edward Tufte insisted on "drawing the viewer's attention to the sense and substance of data (using data graphics), and not to something else". [25]

- *Simplify:* Choose the most efficient visual representation to communicate the data and keep the graphic simple. The design is considered finished when no more details can be taken away without losing information. For small data sets, tables and dot plots are preferred over pop charts (bar charts, etc.) because of more conveyed information, according to William Cleveland. [9]

- *Reduce clutter:* Remove any unnecessary, redundant details (tick marks, grid lines, etc.) or decorations on the graphic.

- *Revise:* Just as the hard work of writing is rewriting, a good visual representation must constantly be revised for refinement and improvement.

- *Be truthful:* Edward Tufte emphasized that in order for the graphic to tell the truth, the "visual representation of the data must be consistent with the numerical representation" [25], as there are quite a few aspects that can distort the data, such as the aspect ratio or the scale.

Besides *visual hierarchy* and *visual flow*, the key thing to consider when designing visual layout is the *grouping of elements.* The Gestalt theory states that perception is influenced not only by the elements, but also by context [2]. Based on this theory, four Gestalt principles which are important in creating a focal point without cluttering the visual representation are *proximity*, *similarity*, *continuity* and *closure*, as depicted in Fig. 6.1

In addition to these principles, certain visual features, such as color, size, shape, orientation, etc., are defined by psychologists as pre-attentive attributes based on the fact that they can be processed in the brain concurrently and almost instantly with little mental effort, which provide more options for encoding data easily. According to Colin Ware, pre-attentive attributes can be basically categorized as *color*, *form*, *movement* and *spatial position*. [29]

Edward Tufte gave a classic example to illustrate the concept of pre-attentive attributes [25], which is illustrated in Fig. 6.2 and Fig. 6.3. Imagine that we have to find the relationships between those numbers in Fig. 6.2. This task sounds like quite difficult, unless we have a proper visual representation, which in this case, is the scatter-plot in Fig. 6.3. By utilizing those scatterplots, we can easily point out the extrema, groupings, trends, gaps or outliers (outstanding values) in the numbers.

## 2.2 History, Definition and Taxonomy of Visualization

Visual representation has a long history [11]. William Playfair is considered to have invented the line graph and bar chart of economic data in 1786, which first appeared in his Commercial and Political Atlas [28]. In 1858, the nurse Florence Nightingale published a polar-area diagram, or coxcombs, to show the causes of death in army during the Crimean war [19]. A very well-known graphic in the visualization field is the thematic map by Charles Minard in 1869 which depicted Napoleon's attempt to invade Russia and his ultimate defeat in 1812 [17]. Another notable visualization is the dot-map of Dr. John Snow which he used to identify the source of the cholera epidemic outbreak of 1854 in London. [23, 12]

Visual representations help us to understand data and therefore produce better information. According to Robert Spence, the process of visualizing data, in the context of a person who observes a visual representation of content, is defined as a cognitive activity with which people build internal representations of the environment around them. These representations are defined as mental models of data, from which human is able to expand on and understand the data [24]. It is rather an abstract concept. Though visualization process may be augmented with the use of visualization tools, which are growing in capability and complexity in recent years, it still remains a cognitive activity. [29, 24]

Spence also noted the three common uses for the term visualization, depending on various properties of the data [24]. For instance, if the data have a correspondence in physical space, it is considered as scientific visualization. In case of abstract data that don't necessarily have a spatial dimension, we speak of information visualization. Data visualization is a broader term which encompasses both information visualization and geographic visualization, which represents information on a map.

# 3 Common Visualization Techniques

## 3.1 Multivariate Data Sets

Multivariate data are collections of data in which many attributes change with respect to one or more independent attributes. While bivariate and trivariate data can be well represented by a traditional scatterplot on Cartesian axes, which is a very simple and intuitive visual form and works well when there are maximum two dependent attributes. However, the number of real-world situations in which one or two dependent attributes are involved is very limited. In fact, most practical problems require analyzing a rather high number of dependent attributes. [16]

Scatterplots can be adapted to multivariate problems to a certain extent by adding further visual element, like shape, dimension, color, etc., which allowing the mapping of more than two attributes and are called Extended Scatterplots. As an example, the extended scatterplot in Fig. 6.4 visualizes data from 174 nations to compare the level of wealth (GDP/capita, x-axis) and the state of health (life expectancy, y-axis). In addition, the graphic also maps the population of each nation (dimension of the visual element) and the continent to which each nation belongs (color of visual elements). Thus this scatterplot has been extended to map four attributes on the same graphics. Normally, extended scatterplots are effective for problems with up to seven varying attributes and not all problems and datasets can be used with scatterplots.

Other common techniques for dealing with multivariate data can be categorized into geometric techniques, pixel-oriented techniques and. iconic techniques. Geometric techniques map data onto a geometric space. One common technique that falls into this category is the Parallel Coordinates. This technique based on the idea pioneered by Alfred Inselberg of IBM in 1981, which defined a geometric space through an arbitrary number of axes, parallely arranged [14]. Other notable geometric techniques include Scatterplot Matrix, TableLens and Parallel Sets.

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

The basic idea of pixel-oriented techniques is to use each pixel of the screen as the base atomic unit to represent an element of the data set and mapping the data of the variable to the color of each pixel, thus maximizing the number of elements that can be represented. In addition, for multivariate data sets, data are grouped into specific areas of the screen according to their attributes called windows. Daniel Keim in his study has defined a number of factors to be considered when using pixel-oriented techniques, which include shape of the window, ordering of the windows, visual mapping (what to map), color mapping (how to map) and arrangement of the pixels. [15]

## 3.2 Network and Hierarchies

The basic idea in visualization of network data is to represent them as *graphs*, using a number of guidelines defined in [22]. Two common techniques to visualize simple network data are Concept Maps and Mind Maps, which are most widely used to describe ideas, situations or organizations in brainstorming sessions and educational environments. With complex network data, the main problem when using graph representations is scalability and extra optimization techniques, such as link reduction or minimum spanning trees (MST) has to be used in order to reduce the complexity of the graph. Other additional components such as geographic map can be used to visualize network topology, as appeared in [13, 10, 20]. A classic type of map used to represent transport networks is the tube map which is based on the design of Harry Beck in 1931.

Graphs can also be used to represent hierarchical data, highlighting the concept of *trees*, which consist of a root, parent nodes and child nodes. This is the basis for many hierarchical data representation techniques, such as the File System (files in folders and subfolders). In fact, tree is the dominant form of representing hierarchical data with many derivations developed to adapt to various kinds of visualization situation, such as the Cone Tree by Mackinlay et al., which is a 3-dimensional visualization form for hierarchies with large number of nodes [8], or the Botanical Tree by Wijk et al., which tries to solve the complexity problem of certain situations by imitating natural (botanical) trees [30], or the Treemap, a space-filling visualization algorithm by Shneiderman which aims to utilize all available space to represent hierarchical data using nested rectangles [21]. Newsmap by Marcos Weskamp is a well-known example that uses the treemap algorithm to displaying news from Google News, dimension of the rectangle to emphasize the importance of the news, color and color intensity to categorize the news and indicate the freshness of the news, respectively, as shown in Fig. 6.5.

## 3.3 Groups

Certain grouping algorithms, such as hierarchical agglomerative clustering, when applied to a data set, will produce groups of observations called a hierarchy of clusters. To visualize this hierarchy as a tree, a *dendrogram,* as shown in Fig. 6.6, is often used. A dendrogram allows not only the groups to be clearly represented, but also the similarity between clusters. There are normally two ways of representing a dendrogram, with the one on the right often used in molecular biology.

Other notable techniques for visualizing groups include Decision Tree, which is generated using a group of decisions based on certain attributes and Cluster Image Map, which is essentially a dendrogram combined with a heat map to display complex, high density data.

# 4 Interactions through Dynamic Techniques

All visualization techniques will eventually face scalability problems, when the data set grows to a certain complexity, if there are no interactions between the user and the visualization tool. Interactions can help solving the scalability problem by allowing the user to modify the input data, change the visual mapping or manipulate the generated views. This is valuable in explorative visualization in that relationships or patterns may be uncover that will otherwise remain hidden in static view. Interactive visual representations may fall into one of three categories, as defined in [16]:

- *Static representations:* No interaction is available, only a single, static view is generated. Fig. 6.7 shows an example of the static representation of the poem "Herr von Ribbeck auf Ribbeck im Havelland", which is part of the visualization project led by Boris Müller et al. for the Poetry on the Road annual international literature festival in Bremen, Germany. The idea is to assign a numerical value to every single letter of a word in the poem, adding them together to get a number that represents the word, then arrange all the words on a circular path with the diameter depends on the length of the poem. The aim of the project is to put hundreds of poems together to produce an artistic and aesthetic graphic, rather than to convey information.

- *Manipulable representations:* Allow users to interact with the generated view through zooming, panning, rotation, etc., with the visualization of the phone data of Malte Spitz, a German Green party politician, by Zeit Online magazine at http://www.zeit.de/datenschutz/m
spitz-data-retention as a good example.

- *Transformable representations:* Enable users to affect the pre-processing phase, e.g. modify the input data through data filtering, with the result of changing or modifying the generated view. This is the highest form of interactive visual representations. As an example, the visualization project GED VIZ at http://viz.ged-project.de/ enable the analysis of the relations between various nations based on different criteria, with each criterion will modify the generated view.

Interactions, however, may add up to the complexity of the visualization tool, and as a consequence, reduce the usability of it from the user's point of view. Thus, an important goal of interaction design is to reduce interaction [18], using a variety of methods.

- *Data brushing:* also known as slicing, is a technique that displays the same piece of data simultaneously in multiple different views. The selected data is displayed throughout all graphics that are linked to each other.

- *Nearness selection:* Howard Wainer proposed this technique to highlight all data points within a specified distance starting from a specified start position (optional). [27]

- *Sorting, rearranging, searching and filtering:* These techniques provide more interactions to users which will further enhance data exploration. However, they must be carefully designed in order not to negatively affect the usability of the graphic.

# 5 Confirmatory and Exploratory Visualization

## 5.1 Exploratory Data Analysis

Is the application that seeks to uncover hidden information, knowledge, trends and patterns within the data set. Visual representations and the ability of analysis via visual perception by human cognitive system are of great advantage to this application, with Bertin defines it as "the visual means to resolve logical problems" .[7]

As an example, Fig. 6.9 shows the distribution of breast cancer cases among women in Germany in 2009. With the color coding, we can recognize the geographical areas with fewer (light red), and with higher number (dark red) of cases. It is noticeable how the number of cases (and possibly death) are found predominantly in the western states. Germany's epidemiological cancer registry association (GEKID) states that the number of cases in eastern states are 20-30 percent lower than in western states and this phenomenon is in some way associated with the lifestyles of the female population [26]. The visualization does not provide an explanation but it may be used as a suggestion for researchers to carry out more extensive epidemiological studies in those areas.

## 5.2 Confirmatory Data Analysis

Another use for visualization is to carry out confirmatory analysis. By representing the structural relationships between series of data visually, the graphic can be used to confirm hypotheses on the data. Fig. 6.8 illustrates this concept by comparing the values of the German stock market index (DAX) to those of American Dow Jones over the course of a year. The visualization makes it clear that the two indices correlate to each other following a rather similar trend in the rising and falling phases, which can be formulated using complex mathematical formulas. But by using the latter method, it would certainly be far less intuitive and expressive for the majority of audiences.

There are certain attributes associated to each of these two approaches to data visualization, as compared and shown in Fig. 6.10. Modern approach to journalism, such as data journalism, can benefit the most from exploratory visualization, which enables journalists to tell complex stories through engaging infographics.

# 6 Conclusion

Data visualization is considered a compelling approach to communicate the subtleties and complexities of the hidden information in data sets range from large to immense (big data). There are several general principles concerning a good visual representation design with their foundations partly based on studies in cognitive psychology, such as the Gestalt principles, and partly on graphics design to guide the layout and data encoding.

There is a variety of visual representations, such as graphs, trees, plots, etc., as well as visualization tools, each of them tailored for a specific category of data, like univariate, multivariate, hierarchies, groups, etc. Finally, dynamic techniques found in software user interfaces, e.g. data brushing, add the "Controller" component, as in the Model-View-Controller (MVC) software pattern, into the visualization process, as shown in Fig. 6.11.

With the data set acting as the Model, the generated graphic (or visual representation) acting as the View, the Controller component extends the reach of visualization for exploratory data analysis by enhancing it with interactions. Such a case is modern data journalism, which is a combination of traditional journalism and the ability to leverage the sheer scale and range of available digital information to tell compelling stories. Visualization software tailored for data journalism are used more and more intensively to find connections between millions of documents, as demonstrated by Hans Rosling in his impressive TED talks on visualizing world stats with Gapminder [3], or the talks by David McCandless which show the importance of clear design in distilling big numbers [1]. This is the one area of data visualization application that has more or less involvement in the big data field and has been seeing much research efforts, development as well as enjoying rapid adoption in recent years, into which I plan to put my focus.

# Appendix: List of Figures

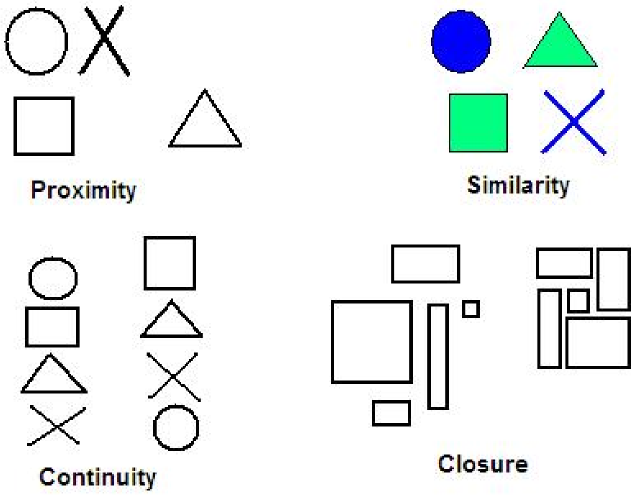


Figure 6.1: Four important Gestalt principles.

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

Figure 6.2: Attentive processing (numbers) is slow, requires much conscious effort. Source: [25]

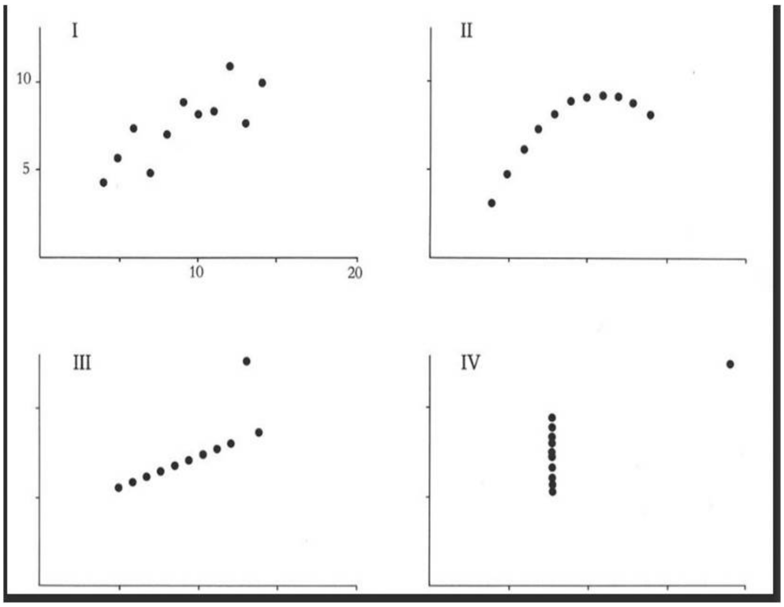


Figure 6.3: Attentive processing (numbers) is slow, requires much conscious effort. Source: [25]

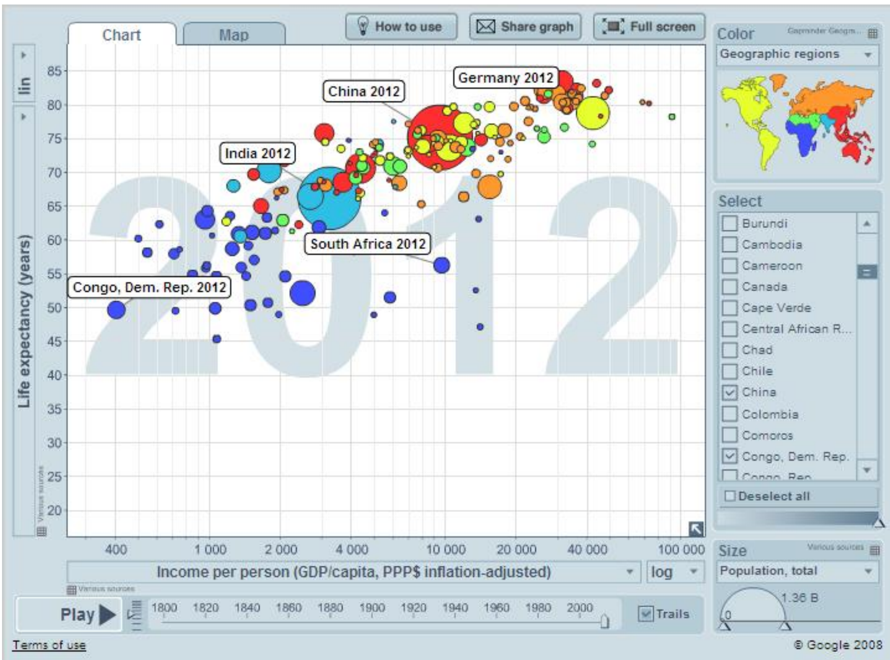


Figure 6.4: Extended scatterplot visualizing four attributes. Source: http://www.gapminder.org


Figure 6.5: Newsmap, taken on February 19, 2014.
Source: http://marumushi.com/projects/newsmap.

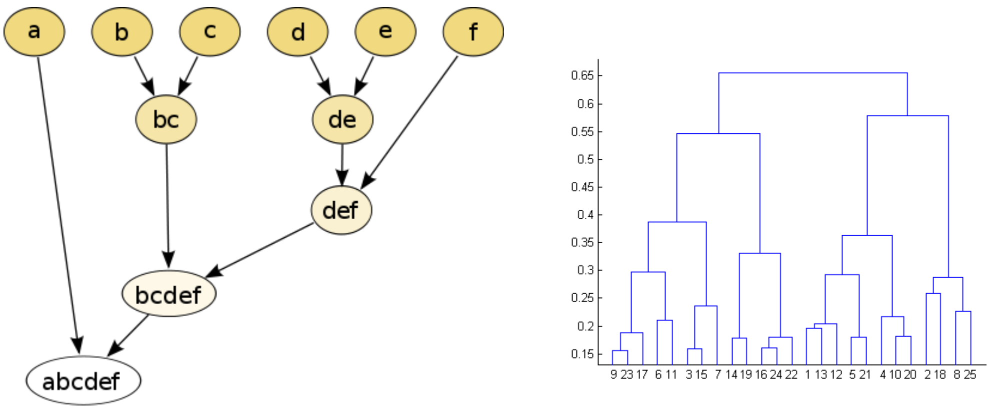


Figure 6.6: Dendrograms.Source: Wikipedia, Mathworks.

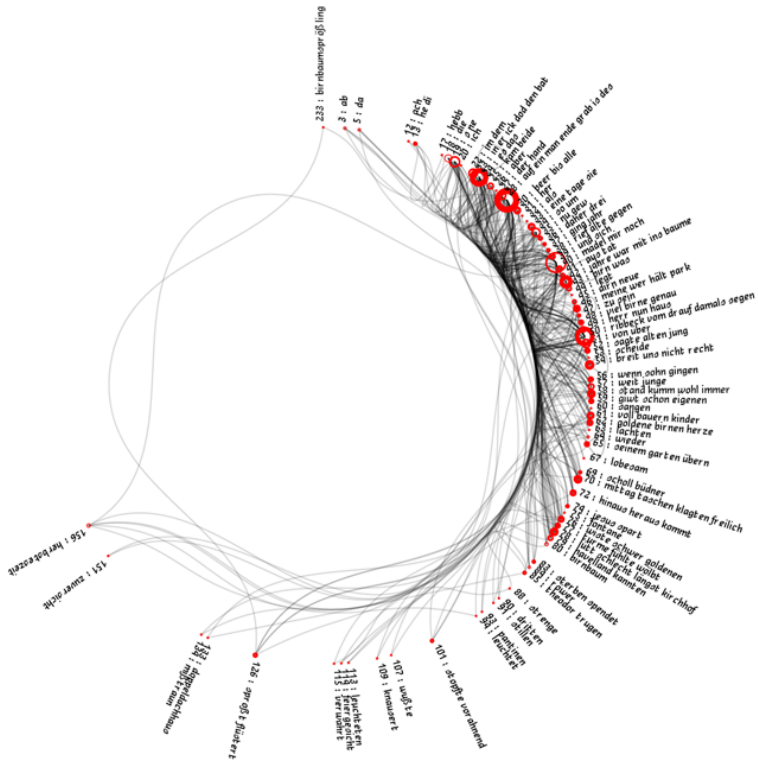

Figure 6.7: Code Poetry by Boris Müller et al.
Source: http://www.esono.com/boris/projects/poetry06/.

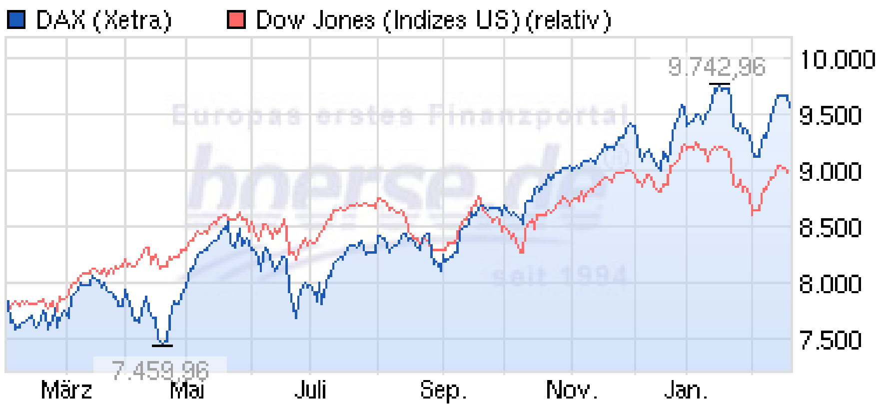


Figure 6.8: DAX (Deutscher Aktien IndeX) compares to U.S. Dow Jones.
Source: http://www.boerse.de/chartsignale/DAX/DE0008469008.

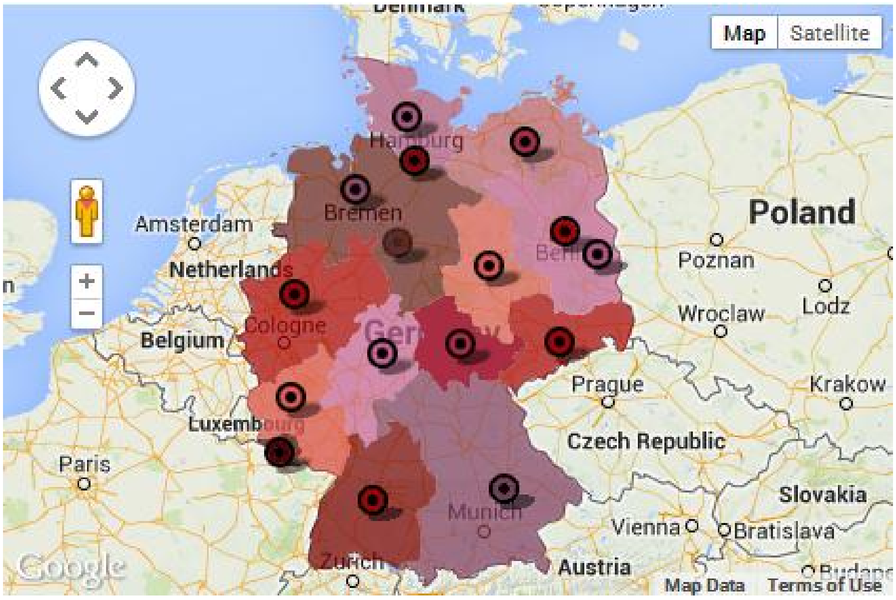


Figure 6.9: Map of Germany showing the distribution of cancer cases in the female population in 2012, subdivided into states.
Source: [26].

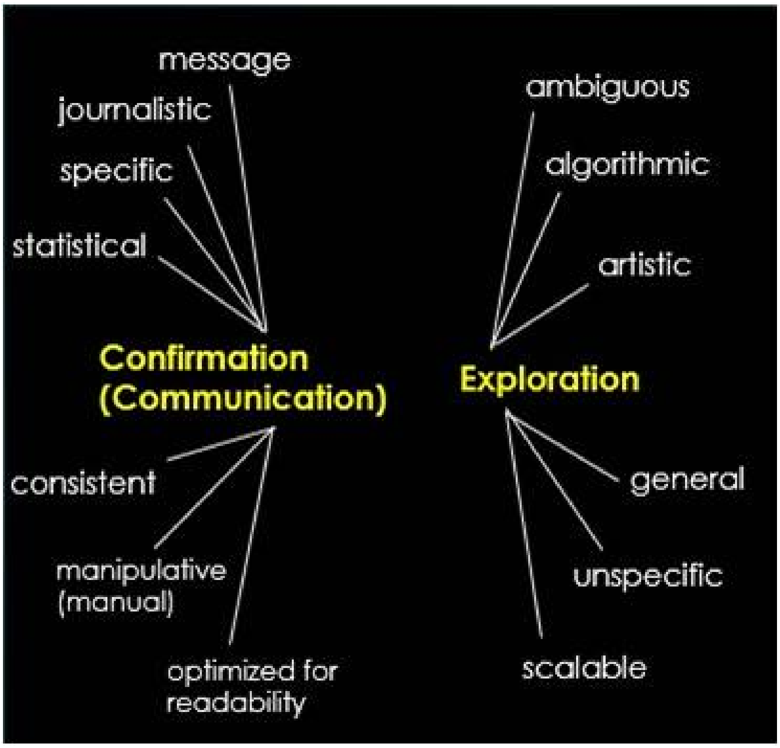


Figure 6.10: Confirmatory vs. exploratory visualization.

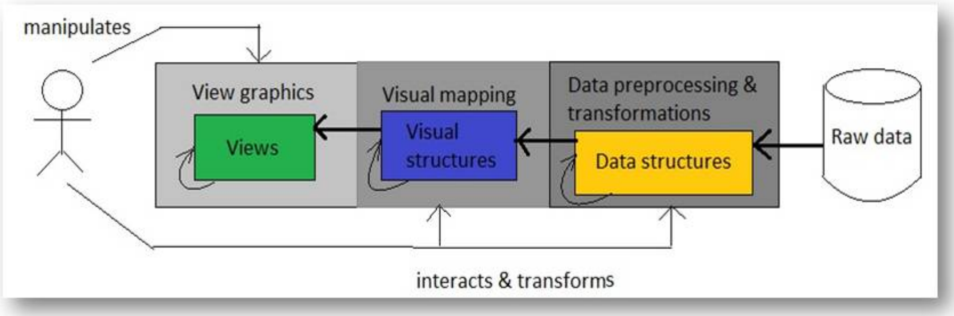


Figure 6.11: Visualization workflow as analog to MVC software pattern.

# Bibliography


[1] : *David Mc. Candless: The Beauty of Data Visualization*. 2014. – URL http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

[2] : *Gestalt Principles- Scholarpedia*. Februar 2014. – URL http://www.scholarpedia.org/article/Gestalt_principles

[3] : *Hans Rosling: The Best Stats You've Ever Seen*. 2014. – URL http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

[4] : *IBM - What is big data?* Februar 2014. – URL http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[5] : *Open Data Portal Hamburg*. Februar 2014. – URL http://daten.hamburg.de

[6] : *What is Data Journalism: The Data Journalism Handbook*. 2014. – URL http://datajournalismhandbook.org/1.0/en/introduction_0.html

[7] Bertin, J.: *Graphics and Graphic Information Processing*. Berlin : Walter de Gruyter, 1981

[8] Card, S. ; Mackinlay, J. ; Robertson, G.: Cone Trees: Animated 3D Visualizations of Hierarchical Information, 1991 (Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology), S. 189–194

[9] Cleveland, W. S.: *The Elements of Graphing Data*. 2nd. New Jersey : Hobart Press, Oktober 1994

[10] Dodge, M. ; Kitchin, R.: *Atlas of Cyberspace*. Harlow : Pearson Education, 2002

[11] Friendly, M.: *Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*. 2009. – URL http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf

[12] Gilbert, E. W.: Pioneer Maps of Health and Disease in England. In: *Geographical Journal, 124:172-183* (1958)

[13] He, T. ; Eick, S. ; Cox, K.: 3D Geographic Network Displays, ACM SIGMOD Record, 1996, S. 25(4):50–54

[14] INSELBERG, A.: N-dimensional Graphics Part I: Lines and Hyperplanes / IBM Scientific Center. 1981. – Forschungsbericht

[15] KEIM, D.: Designing Pixel-Oriented Visualization Techniques: Theory and Applications, IEEE Transactions of Visualization and Computer Graphics, 2000

[16] MAZZA, R.: *Introduction to Information Visualization*. 1st. Springer, 2009

[17] MINARD, C. J.: *Carte figurative des pertes successives en hommes de l'armée qu'Annibal conduisit d'Espagne en Italie en traversant les Gaules (selon Polybe). Carte figurative des pertes successives en hommes de l'armée française dans la campagne de Russie, 1812-1813*. 1869. – URL http://www.math.yorku.ca/SCS/Gallery/images/minard/orig.gif

[18] MYATT, G. J. ; JOHNSON, W. P.: *Making Sense of Data, Part II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. New Jersey : John Wiley and Sons, 2009

[19] NIGHTINGALE, F.: *Mortality of the British Army*. London: Harrison and Sons, 1857

[20] PATTERSON, R. ; COX, D.: Visualization Study of the NSFNet / National Center for Supercomputing Applications (NCSA). 1994. – Forschungsbericht

[21] SHNEIDERMAN, B.: Tree Visualization with Treemaps: a 2D Space-Filling Approach, 1991 (ACM Transactions on Graphics), S. 92–99

[22] SHNEIDERMAN, B. ; CARD, K. S. ; MACKINLAY, J. D.: *Readings in Information Visualization: Using Vision to Think*. San Francisco : Morgan Kaufmann, 1999

[23] SNOW, J.: *On the Mode of Communication of Cholera*. 1855

[24] SPENCE, R.: *Information Visualization*. Addison-Wesley, 2001

[25] TUFTE, E. R.: *The Visual Display of Quantitative Information*. 2nd. Cheshire : Graphics Press, Mai 2001

[26] VONHOFF, A.: *Krebsatlas: Hier ist das Tumorrisiko am größten*. 2014. – URL http://www.focus.de/gesundheit/ratgeber/krebs/symptome/tid-25423/krebs-in-deutschland-krebsatlas-hier-ist-das-risiko-am-groessten_aid_731982.html

[27] WAINER, H.: *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. New Jersey : Princeton University Press, 2005

[28] WAINER, H. ; SPENCE, I.: *The Commercial and Political Atlas and Statistical Breviary*. Cambridge University Press, 2005

[29] Ware, C.: *Information Visualization: Perception for Design.* 3rd. San Francisco : Morgan Kaufmann, Juni 2012

[30] Wijk, J. J. v. ; Wetering, H. van d. ; Kleiberg, E.: Botanical Visualization of Huge Hierarchies, 2001 (Proceedings IEEE Symposium on Information Visualization (InfoVis)), S. 87–94