# Investigating the visualization process. Development of an integrated framework

**Master Thesis**

A thesis submitted for the

Master of Science in Web Science

**Anna Presnyakova**

Immatrikulationsnummer: 214100933

E-Mail: apresnyakova@uni-koblenz.de

Faculty 4: Computer Science

Institute for Information System Research

**University of Koblenz-Landau**

**Germany**

Supervisors:

Prof. Dr. Susan Williams

Florian Schwade

**Koblenz, September 2017**

## Erklärung

Ich versichere,

dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Mit der Einstellung dieser Arbeit in die Bibliothek bin ich <nicht> einverstanden. Der Veröffentlichung dieser Arbeit im Internet stimme ich <nicht> zu.

**Anna Presnyakova**

Koblenz, September 2017

## Acknowledgments

I would like to express my special thanks of graduate to my supervisors Prof. Dr. Williams and Florian Schwade who gave me this opportunity to write this thesis on the interesting topic for me. I highly appreciate the valuable feedback and all guidance during all project time which helped me to keep moving and improve this work.

Also, I want to express my gratitude to my parents, family, and friends. Especially, I am grateful to my uncle without whom this study journey would have never started. Special thanks to my partner, for endless support and belief in me without which this project would not have been finished.

## Abstract

The extensive literature in the data visualization field indicates that the process of creating efficient data visualizations requires the data designer to have a large set of skills from different fields (such as computer science, user experience, and business expertise). However, there is a lack of guidance about the visualization process itself. This thesis aims to investigate the different processes for creating data visualizations and develop an integrated framework to guide the process of creating data visualizations that enable the user to create more useful and usable data visualizations. Firstly, existing frameworks in the literature will be identified, analyzed and compared. During this analysis, eight views of the visualization process are developed. These views represent the set of activities which should be done in the visualization process. Then, a preliminary integrated framework is developed based on an analysis of these findings. This new integrated framework is tested in the field of Social Collaboration Analytics on an example from the UniConnect platform. Lastly, the integrated framework is refined and improved based on the results of testing with the help of diagrams, visualizations and textual description. The results show that the visualization process is not a waterfall type. It is the iterative methodology with the certain phases of work, demonstrating how to address the eight views with different levels of stakeholder involvement. The findings are the basis for a visualization process which can be used in future work to develop the fully functional methodology.

## Contents

## List of figures



## List of tables



# 1 Introduction

## 1.1 Problem statement and Motivation

Due to the rapid growth of social networking, internet of things and data management, modern society is producing a vast amount of data every day. According to the IBM estimate, society generates 2.5 quintillion bytes of data every day ("IBM - What is big data?," n.d.). This overwhelming amount of data has valuable information, which can help to guide and manage activities in businesses, governments, and personal lives. To derive value out of this information, a data scientist must find ways to explore and communicate the message meaningfully.

The purpose of visualization is to support the understanding of data based on the human visual system, which gives the ability to see patterns, show trends and identify outliers. Well-made data visualizations can make the cognitive process faster, improve comprehension and decision-making (Patterson et al., 2014). A powerful visualization will engage more diverse audiences and delivers the message or story more efficiently. The challenge is to create effective and engaging visualizations appropriate to the particular data set by combining existing data visualization techniques.

Additionally, data visualization is quite a broad and complex field. Segel and Heer note that visualization designers are "melding the skills of computer science, statistics, artistic design, and storytelling." (Segel & Heer, 2010). A significant amount of research has already been conducted in this field. However, each research study has focused on a specific aspect of the data visualization field. It could be chart drawing, visual encoding, human perception, interactivity, tool investigation, narrative construct and many other different aspects. Data visualization can be a challenging process. A lot of questions must be answered by the data scientist during the visualization process, such as:

- How to filter data?
- Which business requirement has to be considered?
- Which visual encoding has to be used for certain data type?
- How to represent multivariable data?
- How to represent a particular type of relationship?
- Which chart to use?
- How to highlight the crucial point?
- Should be interactivity or dashboards used?

The data analyst can easily become lost in the vast amount of the literature and decisions which must be made. Thus, a common methodology is needed to guide the user through the data visualization process in a more efficient way. During literature review, several frameworks (Bertini, Tatu, & Keim, 2011; Chen & Jänicke, 2010; Fry, 2004; Liu, Cui, Wu, & Liu, 2014; Munzner, 2009) have been identified that were developed for this aim. All frameworks show the process of taking some data and making it useful with the help of data visualization. However, they cover different things to guide the user, and each of them has a slightly different focus. Although they have some similarities and also some differences, all of them provide useful information for the data visualization process. Nevertheless, different approaches might be controversial and confuse the user.

### 1.2 Structure of the Thesis

The thesis is organized as follows.

Chapter 2 explains the state of the art of the data visualization field, by examining the connected literature. This chapter aims to familiarize the reader with the subject of the data visualization. It provides an understanding of the concept of the data visualization and explains the difference between the data visualization and infographics. Then, this chapter explains the purposes and reasons to create the data visualization. Moreover, it gives the understanding of the chart and its components. Lastly, it provides with the application of the data visualization.

Chapter 3 presents the first phase of the research. It has the identification of the existing frameworks for the data visualization and analysis of them. The analysis of the existing frameworks introduces the building components of frameworks and views on them. Later, this chapter represents the comparative analysis of the frameworks and the strength and limitations of each framework.

Chapter 4 develops the preliminary framework based on the finding in Chapter 3. It represents the detailed description of eight views and their diagrams. It gives the first insight into the interconnection of the views. Lastly, the Integrated Visualization Analysis and Design Framework is developed by providing the diagram of the visualization process.

Chapter 5 represents the last phase of the research. Firstly, it evaluates the preliminary framework, in order to understand its problems. The evaluation is done with the help of the workshops between the clients and the data designer. The chapter contains the results and summarization of each workshop in a form workshop journal. Lastly, the Integrated Visualization Analysis and Design Framework is improved based on the findings of workshops.

Chapter 6 concludes and summarizes the results and findings of the research, furthermore provides the suggestions for the future work.

### 1.3 Research design

#### 1.3.1 Research aim, objectives and questions

The aim of this research is to understand and investigate the different processes for creating data visualizations and develop an integrated framework to guide the process of creating data visualizations, that enable the user to create more useful and usable data visualizations.

In order to achieve the research aim, the following research objective (RO) have been proposed based on the preliminary literature review. Each research objective has one or more research questions (RQ), which represented below as well.

| RO 1 | **To identify existing frameworks (processes/pipeline) for creating data visualizations, then to analyze them, to understand their key elements and the strengths and limitations of different approaches.** |
|---|---|
| | |
| **RQ 1.1** | What frameworks exist to guide the process for creating data visualizations and how can they be classified? |
| **RQ 1.2** | What are the different components/elements of each of the identified frameworks? What is the function and purpose of each component? |
| **RQ 1.3** | What are the strengths and limitations of existing frameworks? |

| RO 2 | **To synthesize the elements from existing frameworks develop a preliminary version of an integrated framework for data visualization.** |
|---|---|
| | |
| **RQ 2.1** | How can the components identified in Objective 1 be integrated into a more comprehensive framework? |
| **RQ 2.2** | What tasks should be included in each component? |

| RO 3 | **To test and validate the preliminary framework in the context of Social Collaboration Analytics** |
|---|---|
| | |
| **RQ 3.1** | To which extent can the preliminary framework support visualization in the context of Social Collaboration Analytics? |

| RO 4 | **If necessary, refine the framework in light of the testing phase.** |
|---|---|
| | |
| **RQ 4.1** | What new elements should be added to improve the framework? |

As a research methodology, the design science methodology was chosen because the key focus of this methodology is the contribution of the new knowledge (Kuechler & Vaishnavi, 2007), which is appropriate for this research.

The data source for the first and second research objectives is the literature. The data source for the third and fourth objectives is the results of the workshops between the client and the data designer.

The results of the research will be represented with a help of diagrams, visualizations and textual descriptions.

At the end of this work, the Integrated Visualization Analysis and Design Framework will be presented.

### 1.3.2 The Design Science Research Methodology

This section represents the short description of the design science research model. This model is adapted from a computable design process model developed by Takeda et al. (1990). Design science research cycle is displayed in the Fig. 1.5.1. It has five main stages of research:

- **Awareness of Problem** comes from the literature review in the certain research field or the industry. The connected field to the initial research topic can also bring the new insight into the research problem. The research proposal might be a result of this phase.
- **Suggestion** for the problem is proposed based on existing knowledge in the problem area. For this, new functionalities are discussed based on present and new elements. In order to recognize possible suggestions, the researchers will have to do a literature review, and they must be creative.
- **Development** phase is responsible for developing the solution based on results from the previous step. Development technique depends on the research problem. It is the most important phase of the design science research.
- **Evaluation** phase is testing the developed solution according to the criteria that must be identified in the research proposal on the awareness phase. Then, the proposed solution should be improved based on the result of the evaluation.
- **The conclusion** summarizes the research finding.

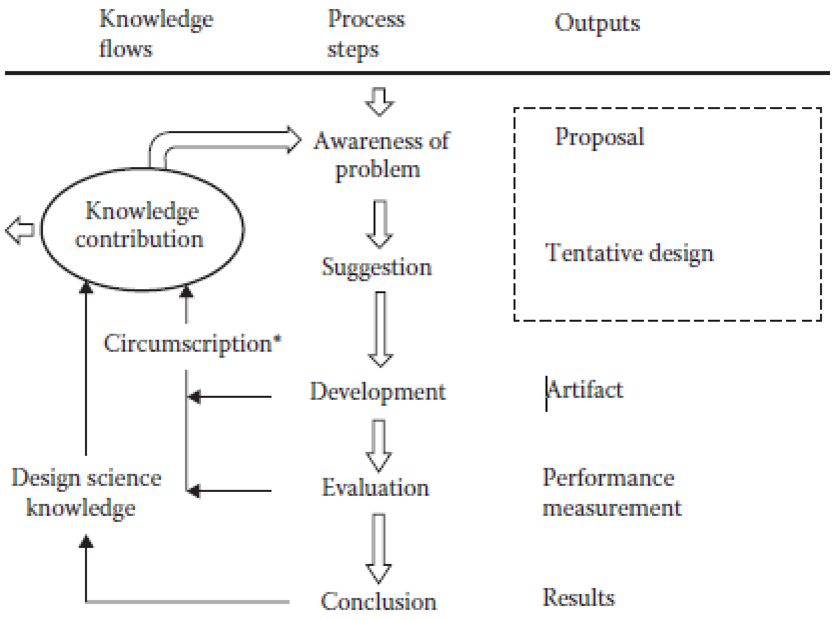



Fig. 1.3.1: Design Science Research cycle (Kuechler & Vaishnavi, 2007, p. 406)

### 1.3.3 Research Design Steps

This section shows the mapping from the Design Science Research to the thesis's research objectives and questions. The mapping is displayed in the Fig. 1.5.2: Research Design.

In the beginning, the literature review is done to be aware of the problem in the data visualization field. During the literature review, it can be seen that the data visualization requires an extensive set of skills from the data designer in order to visualize the data efficiently. However, no clear guidance on the visualization process exists.

The suggestion phase of the design science research corresponds to the first research objective; the existed frameworks need to be identified and analyzed to see their strength and limitations.

The development phase of the design science research corresponds to the second research objective, where the researcher develops a solution for the problem suggested on the previous stage based on the analysis. In this case, it is the development of the preliminary version of the integrated framework for the visualization.

The evaluation phase of the design science research corresponds to the third and fourth objectives in the research. The preliminary framework needs to be tested to evaluate the solution. Then the preliminary framework must be improved based on the testing results.

Lastly, the conclusion phase of the design science research relates to the conclusion and the research findings.

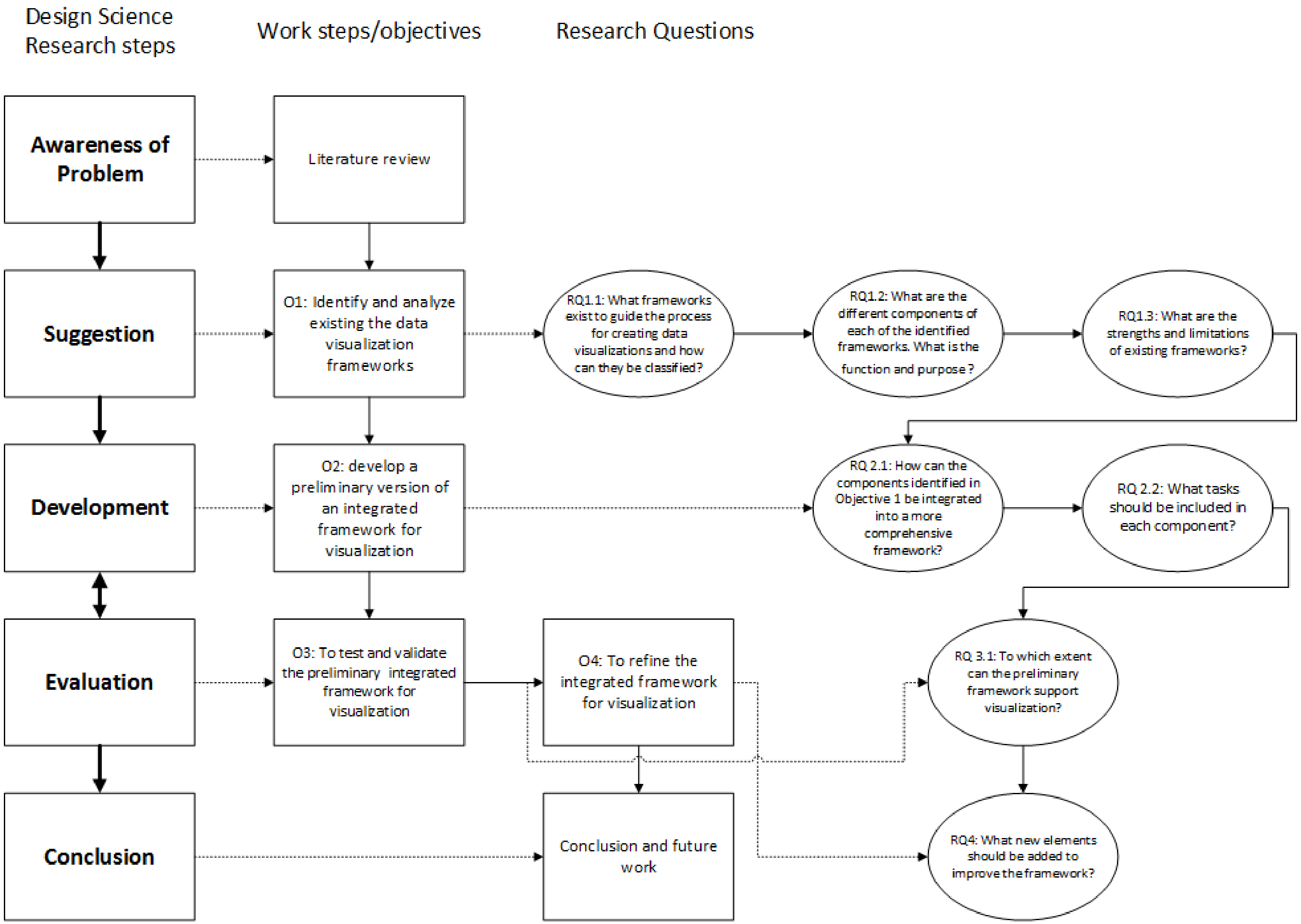



Fig. 1.3.2: Research Design

# 2 State of the Art in Data Visualization

## 2.1 What is data visualization?

As a large amount of data available, the visualization part is taking that data and figuring out the ways to translate it into the visuals that relatively easily understood. The reader can read lines, shapes, and color instead of rows and columns of data. Data visualization is about turning numbers into pictures and stories; it allows a person to explore and understand the data in a different way. Visuals are very fast by a human in processing. The good data visualization can bring the unexpected insights into the data (Knaflic, 2015).

Data visualization looks for the best way to represent underlying data visually using things (like color, size, shape) to convey some information or insight to the audience or reader, using storytelling and narrative elements; in other words, this is a process of mapping values to visual content (Murray, 2012).

## 2.2 Data Visualization versus Infographics

Data Visualization and Infographics are both used to represent the message behind the data. However, they are not the same and often confused with meaning. Some people can use the terms interchangeably. The distinction between infographics and data visualization is based on both form and origin. Infographics are mainly drawn as a sketch in the programs such as Adobe Illustrator, after some changes, it is needed to be adapted manually. Also, it has many design decorations and normally not suited to work with a large amount of data. At the same time, the data visualization is drawn with programming tools or the data analytics software, such as Tableau. The same algorithm for the data visualization can be used for different data sets and a large amount of data. Additionally, the data design normally is not suited for decorations (Iliinsky & Steele, 2011). The differences between them are summarized in the Table 2.2.1.

Table 2.2.1: Differences between Data Visualization and Infographics (the table conducted based on (Iliinsky & Steele, 2011))

| Aspect | Infographics | Data visualization |
|---|---|---|
| How is it drawn | manually drawn | algorithmically drawn |
| Data Domain | attach to specific data | can be used for different datasets |
| Data size | relatively data-poor | relatively data-rich |
| Visual component | aesthetically rich | aesthetically poor |

## 2.3 Purposes of creating Data Visualization

In general, researchers define two main categories of purposes of the data visualization: exploratory and explanatory. While they serve two distinct purposes, some techniques and approaches might be appropriate for one category but are not good for another. For this reason, it is important to distinguish them. The Exploratory Data Analysis could be perceived as a conversation between a data analytics and the data. While Data Visualization is a conversation between an audience and the data (Iliinsky & Steele, 2011).

### 2.3.1 Exploratory

Exploratory data visualization is used by a visual designer on the stage when he/she has many data but does not know anything about it. The basic chart can give the data designer the first awareness of the data. The aim is to understand what data is, not to try to prove something. It helps to identify some features about the data, like lines, patterns, trends, curves or outliers. This category of visualization normally happens in the data analysis phase and helps to find the story behind the data. It can be done with the aid of basic plots, scatter plots, histograms and so on (Iliinsky & Steele, 2011).

### 2.3.2 Explanatory

Explanatory data visualization is used when a visual designer already knows the story behind the data, and he/she tries to tell this story to his/her audience. On this stage, the data origins are known to the data designer. Therefore he/she should make the visualization, which highlights his message. The data designer should make the editorial decisions at this stage, like to choose an appropriate visual type, to eliminate the distracting data from the message, to draw audience attention to the relevant information, to represent the narrative. These factors have the main aim to represent the focused data set (Iliinsky & Steele, 2011).

### 2.3.3 Hybrids: Exploratory Explanation

This category is a combination of two approaches. Firstly, a data designer chooses the certain data set, represent it on the first, then allows the reader explore the further details and make some conclusion on his own. It is done through the interactive interface where a user can select the subset of the data selecting some parameters. The result can be that the reader discovers some new aspects of the data, which the author had not noticed before (Iliinsky & Steele, 2011).

## 2.4 Reason to create graph

### 2.4.1 Human vision

The key reason lies in the immense power of people's visual processing system. Danish physicist Tor Norretrander converted the bandwidth of human senses to computer terms (Fig. 2.4.1) to help us understand the power of human vision system. Sight takes the majority of the frame. In fact, it is said that sight can process information up to the speed of computer Networks or an Ethernet cable (Koch et al., 2006; Nørretranders, 1999)

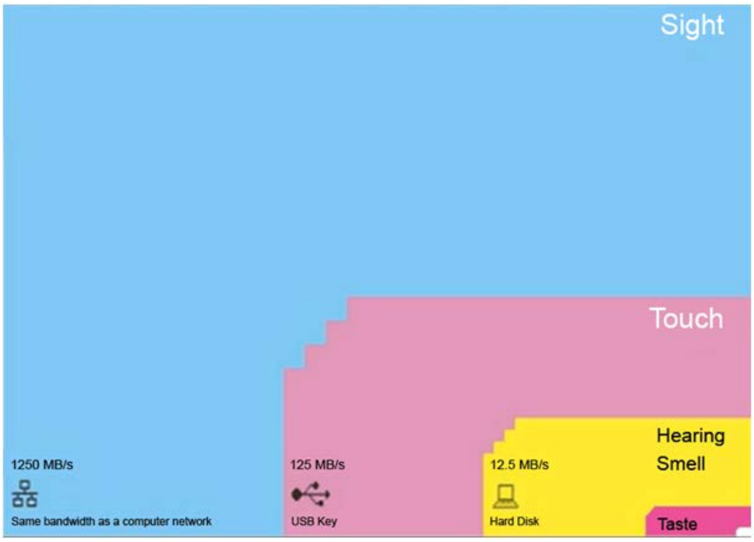



Fig. 2.4.1: The bandwidth of the human's senses (Nørretranders, 1999, p. 124)

### 2.4.2 Anscombe's Quartet

Anscombe's Quartet (Fig. 2.4.2) was introduced by the statistician Francis Anscombe in 1973. It represents four datasets from 11 points each (Table 2.4.1). Every data set has the same summary statistics such as mean, variance, correlation coefficient and the line of best fit (Table 2.4.2). A person would expect that since these data sets have the same mean, variance, correlation coefficient and the line of best fit, that they look very similar when visualized. However, these datasets are different. It can be seen in the Fig. 2.4.2 that the effect of curvature and outliers have drastically thrown off the summary statistic. Anscombe's Quartet demonstrates how important it is always to plot the data rather relying only on summary statistics (Anscombe, 1973).

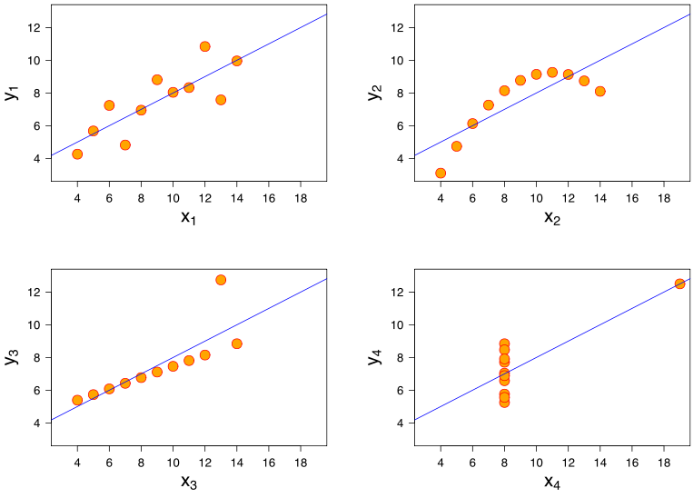


Fig. 2.4.2: Anscombe's Quartet Visualization

Table 2.4.1: Anscombe's Quartet datasets (Anscombe, 1973, p. 19)

| 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 |

Table 2.4.2: Anscombe's Quartet summary statistics (Anscombe, 1973, p. 19)

| Property | Value | Accuracy |
|---|---|---|
| Mean of x | 9 | exact |
| Sample variance of x | 11 | exact |
| Mean of y | 7.50 | to 2 decimal places |
| Sample variance of y | 4.125 | plus/minus 0.003 |
| Correlation between x and y | 0.816 | to 3 decimal places |
| Linear regression line | y = 3.00 + 0.500x | to 2 and 3 decimal places, respectively |
| Coefficient of determination of the linear regression | 0.67 | to 2 decimal places |

## 2.5 Charts

Charts are composed of different visual encodings, each of them most effectively conveys a specific aspect of the underlying data. A chart can represent tabular numeric data, functions or some kinds of qualitative structure and provides different information (Wilkinson, 2005).

While there might be infinite numbers of ways to create charts through various types of visual encoding, in practice, there are limited numbers of combinations that make sense and a smaller number that people use in practice. Part of this is due to the constraints and rules of graphical representation, but also it comes from the fact that it is easier to understand familiar patterns. While each type might be slightly better for slight differences in the data and what a presenter is trying to communicate, in practice, he/she will get by with only a few.

To summarize, chart type is simply a set of visual encodings applied to data types and combined with some relationship between those data. It is visualized in the formula above (Few, 2004b):

$$ChartType = VisualEncoding + DataType + DataRelationship$$

### 2.5.1 Data types

Before the process of visualization is started, data designer needs to understand data and data types. Almost all the information can be turned into the data and after it can be visualized. However, the way to visualize a data strongly depends on the data type. Most of the data types can be categorized into three basic types (Few, 2004b):

- Quantitative: any variables that have exact numbers. Quantitative data can be discrete or continuous;
    - Continuous: it has a meaning of measurement, such as weight;

  - Discrete: it has a meaning of count, such as number of people;
- Categorical: represent characteristics;
  - Nominal is a way of labeling or categorizing the data, such as color;
  - Ordinal data: this data type is a mix numerical and categorical data. It is categories, but they have some orders or ranking such as the size in words small, big, large and extra-large;
- Time series: it is a collection of observations collected at regular intervals over the period, for example, the average temperature over the year.

In order to show the example of data types and visual encodings, a well-known visualization called "200 years that changed the world" from Gapminder was taken (Fig. 2.5.1). This visualization was made based on the global health data and intends to show how the human wealth was changing over two hundred years. It shows the correlation between income per person and life expectancy, in many different countries.

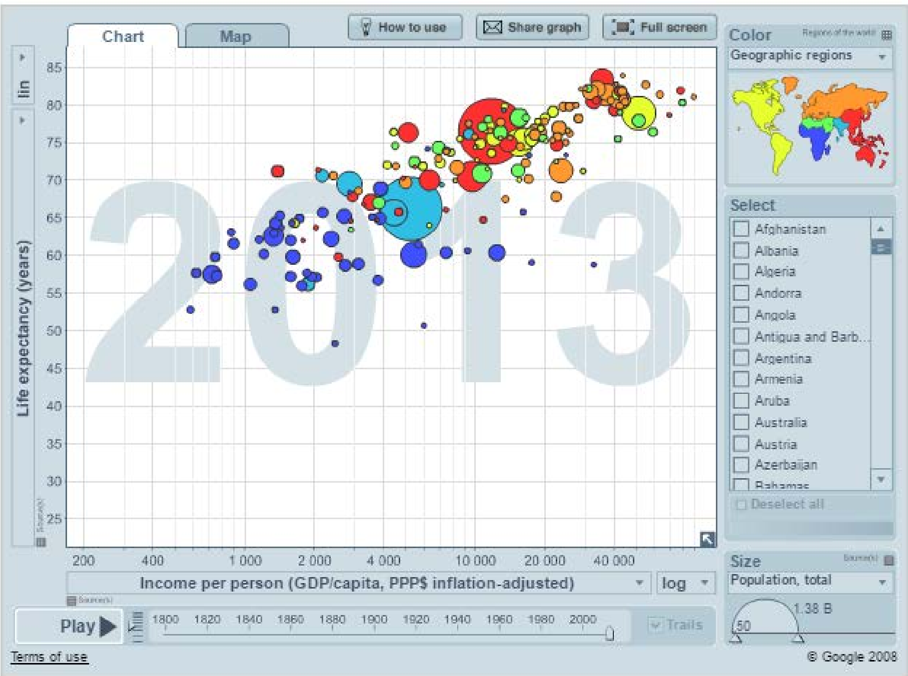



Fig. 2.5.1: The screenshot visualization from Gapminder ("Gapminder World," n.d.)

The various data types can be found in this visualization. First, the quantitative data, which includes variables, like life expectancy, income per person, total population and year. In this example, the total population would be considered a discrete variable, while income per person would be a continuous variable.

The second data type, which appears in this visualization, is categorical data. Nominal data is a way of labeling or categorizing the data into groups. In this case, the countries of the world are grouped into regions. The America is colored yellow, Africa is colored blue. Categorical data may take on numerical values, but these numerical values will not necessarily carry any mathematical meaning.

Next, ordered data is similar to categorical data, except that the categories are ordered or ranked in some particular way (this data type does not represent on this visualization). One example would be if the designer takes the populations of each country and then splits them into bins. For example, it could be taken four groups of countries, where each of the countries falls into a population bin if their population is between 0 and 10 million, 10 and 100 million, 100 and 300 million or greater than 500 million.

Another example would be a class of difficulty, such as beginner, intermediate and advanced. Those three types of classes would be a way that we could label the classes, and they have a natural order in increasing difficulty. It is important to know for a data scientist that each of these data types can be encoded visually.

In the next section, it will be discussed how appropriate choices for visual encoding can be made based on the different data types.

### 2.5.2 Visual Encoding

In other words, visual encoding is mapping from data to the display elements. There are two main categories of visual encodings: spatial (which use an element's position on a graph to encode variables) and retinal (which use a visual component in order to encode variables, such as form or color). The detailed explanation is described below (Abela, 2009; Few, 2004b, 2012; Heer, Bostock, & Ogievetsky, 2010).

The position is considered a planar variable because it locates points in the space. It is one of the most prominent visual encodings in data visualization. This visual encoding a viewer can perceive with great accuracy. The position is good to encode two variables (x- and y-axes).

In the case, when a designer needs to visualize higher dimensions of the data (for example, more than two variables), it could be done along with a z-axis. However, humans' eyes poorly perceive 3D models, as it is harder to make quantitative comparisons between points. For example, it is not clear which of three points (Fig. 2.5.2: Example of 3D model) has a greater y-value or z-value. For this reason, there is a need for other techniques to encode the third variable for the data visualization. The retinal variables can be used to encode additional variables for the data set. Size is an example of a retinal variable; it is particularly suitable for ordered data.

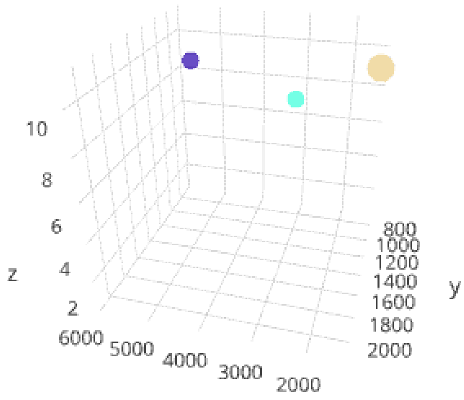



Fig. 2.5.2: Example of 3D model

The retinal variables are size, orientation, color saturation, color hue, shape, and texture. Size, orientation, color saturation (Fig. 2.5.3) is especially effective for order data. However, it might be difficult to perceive quantitative differences using these visual encodings. The perceived value of two tons of red, for example, light red and very light red, is not obvious. Other retinal variables such as color hue, shape and texture are suitable for encoding nominal variables.

Ilinsky and Steele conduct the helpful table that summarizes the ideas described in the sections 2.5.1 and 2.5.2 (Iliinsky & Steele, 2011). This table shows the examples of the visual encoding and addresses the issue of choosing the certain visual encoding in respect to the data type.

| Example | Encoding | Ordered | Useful values | Quantitative | Ordinal | Categorical | Relational |
|---|---|---|---|---|---|---|---|
| | position, placement | yes | infinite | Good | Good | Good | Good |
| 1, 2, 3; A, B, C | text labels | optional alpha or num | infinite | Good | Good | Good | Good |
| | length | yes | many | Good | Good | | |
| | size, area | yes | many | Good | Good | | |
| | angle | yes | medium | Good | Good | | |
| | pattern density | yes | few | Good | Good | | |
| | weight, boldness | yes | few | | Good | | |
| | saturation, brightness | yes | few | | Good | | |
| | color | no | few (<20) | | | Good | |
| | shape, icon | no | medium | | | Good | |
| | pattern texture | no | medium | | | Good | |
| | enclosure, connection | no | infinite | | | Good | Good |
| | line pattern | no | few | | | | Good |
| | line endings | no | few | | | | Good |
| | line weight | yes | few | | Good | | |

Fig. 2.5.3: Use of the visual encoding according to the data type (Iliinsky & Steele, 2011, p. 28)

In order to demonstrate the use of different visual encoding, the Fig. 1.5.1 is used. Firstly, life expectancy is encoded visually along the y-axis. Secondly, spending per person is encoded along the x-axis. It means that countries that are located in the top-right corner of the graph had a high income per person and high life expectancy. The opposite will be true for the countries in the lower-level corner. Thirdly, to encode the geographical region the author used colors. This choice of color makes it easier to compare income and life expectancy across developing and developed countries. Lastly, the size of the circle was used to encode the population size of a country. It helps the reader immediately see the countries with high population. However, the most interesting point of this visualization is the fact how the authors represented the time in this visualization. In this case, time is animated through the individual frames. The combination of these visuals encoding and animated time allows showing the change in life expectancy, income per person and population over time, which makes this visualization engaging.

### 2.5.3 Featured Relationships

Choosing the chart type is important to pay attention to the relationship between variables. According to Few, there are eight types of quantitative relationship depending on how one type of value related to another type of value (Few, 2004b, 2012):

- Time Series. This is measurements, which made with constant frequency in given period. It is represented by the line chart where, usually, the x-axis is time, the y-axis is measurement, for example, daily sales.
- Nominal Comparison. It is a simple comparison of categories of one or more measurement. For example, it is a comparison of total sales in different continents.
- Ranking. This is categorical subdivisions with certain measurement or ranking, for instance, the sales over countries order by the rank according to sales.
- Part-to-Whole. This is also categorical subdivisions with measurement but with respect to the ratio as a whole. For example, it could be the sales in countries, but the sale represents the percentage of current country from the total sale over all countries
- Deviation. This is categorical subdivisions with measurement compared to a reference measure; it shows the differences between the certain measurement and the reference measurement. For example, the country sales comparing with average sales over all countries. A bar chart could represent it, x-axis the average sale (the reference measurement), if the bar of the particular country above the x-axis, it the means this country has the sale above the mean and vice versa.
- Distribution. It sums the values in the certain interval from lowest to highest. For example, it could be the distribution of sales per week, where the interval is one day of the week and sum of the values is the amount of sales per day (or total sales per day).
- Correlation. It allows comparing two sets of measurements and seeing if these measurements depend on each other.  For example, we can check if the total sale in companies correlates with the number of employees in firms. The positive correlation means the total sales increased with the number of the employees in companies, as a general trend.
- Geospatial. This type displays the measurements on the map to see their location. It allows understanding geographical tendencies.

### 2.5.4 Visualization types

A large number of chart types exist. The authors categorize visualization types differently. However, several authors agree on the categorization of visualization types according to the data type they are supposed to demonstrate (Heer et al., 2010; Shneiderman, 1996; Zoss, 2017). This taxonomy includes five main categories:

- Planar or geospatial categories of charts suit to represent data which is connected to geographical location (cartogram, dot distribution map, proportional symbol map);
- Temporal category of the visualization is suitable for representing the data which is time dependent (such as timeline, time series, Gantt chart, Sankey diagram, polar diagram);
- Multidimensional category is aimed to visualize the data with more than two variables (tree map, bubble chart, parallel coordinates, radar chart, waterfall chart);

- Tree category is for visualize hierarchical data (dendrogram, radial tree, hyperbolic tree, treemap, icicle chart);
- Network category is for representation node and connections between them (matrix, node-link diagram, dependency graph).

It has to be noticed here, that this list is not exhaustive and other chart types and visualization types exist. The aim here is to give the reader the main overview of what exists. Due to the limitation of this work, not all of them can be introduced.

## 2.6 Application

Data visualization has various application domains, due to the fact that it helps to improve the understanding of complex data. Thus, the data visualization has application in all fields, which require analysis of a large amount of data. Keim highlights several domains (Keim et al., 2008; Keim, Kohlhammer, & Ellis, 2010):

- Engineering domain. The engineers can measure and analyze the performance with the help of data visualization. It can be used in the prototype development phase, experimental test series, customers' feedback. Analyzing the received data in these cases can help to make the process faster.
- Financial Analysis. Financial information systems have a large amount of heterogeneous data which changes over the time. The data visualization helps to see trends, extremums over the time and make predictions for this data (for example trading data).
- Socio-economic. Modern society is a complex system; it depends on a lot of factors in fields such as politics, economics, culture, and demography. The data visualization helps to analyze all these factors visually and see trends.
- Public Safety & Security. In this domain, an analyst needs to monitor huge amounts of complex information streams; the data visualization can contribute to seeing a correlation between some factors, which influence on the safety and security.
- Environment and Climate change. This domain very often needs to analyze the large log of data over the time. The data visualization can address this issue and help to find the patterns in this data (for example, change of animal population over a year).
- Biology and medicine. It can help to analyze a large amount of data in the bioinformatics field.
- Physics and astronomy also require analysis of a large amount of data.

# 3 Frameworks

Due to a significant amount of literature in the field of data visualization, some authors are trying to create pipelines for the data visualization process. The aim is to give a general overview. However, it can be seen immediately that all of them are slightly different from each other. The aim of this chapter is to identify these frameworks in the literature, analyze them to find their strength and limitations, and lastly to compare them in order to understand if a new approach is needed.

As there is a lot of data out of there, businesses, researchers, universities and other stakeholders want to analyze them to gain some knowledge, value, money or make decisions based on this analysis. The clear way to do so is to visualize this data. However, there are many ways to visualize data.

During the literature review, it was found that several visualization pipelines or frameworks have been developed to guide a user to visualize data in a way that is more efficient. These models (frameworks) cover different things to guide the user. These models (frameworks) represent the process of taking some data and making it useful. Each framework has a slightly different focus. There are some similarities, but there are also some differences. However, all of them provide some useful information.

This chapter provides a detailed description of each framework. The preliminary analysis and comparison will be introduced.

## 3.1 Related works and identification of existing frameworks

In the literature analysis, five main frameworks were discovered:

- Framework 1: Visualization Pipeline. Research paper "A survey on information visualization: recent advances and challenges" by S. Liu, W. Cui, Y. Wu (Liu et al., 2014);
- Framework 2: Data Visualization Process. The doctoral thesis "Computational Information Design" written by B. Fry (Fry, 2004);
- Framework 3: The Nested model for Visualization. Research paper "A Nested Process Model for Visualization Design and Validation" written by T.Munzner (Munzner, 2009);
- Framework 4: An information-theoretic Framework for Visualization. Research paper "An information-theoretic framework for visualization" written by M. Chen and H. Jänicke (Chen & Jänicke, 2010);
- Framework 5: Quality Metrics Pipeline. Research paper "Quality metrics in high-dimensional data visualization: An overview and systemization" written by E. Bertini, A. Tatu and D. Keim (Bertini et al., 2011).

For the sake of more convenient reference system, throughout this work, the name of the framework in some cases will be omitted, the only enumeration will be used, for instance 'Framework 1 uses …'.

In the beginning, we need to understand every framework separately. In the next five sections, each of the listed above frameworks will be introduced and described, providing the background of the research papers. Also for each of the frameworks, an analysis of its component has been undertaken and a table summarizing the components produced. Components can be perceived as a building block (function or activity). From, the first glance it can be seen that some frameworks have similar components; however, the whole in-depth comparison will be presented afterward.

### 3.1.1 Framework 1: Visualization Pipeline

Liu wrote the article "A survey on information visualization" in 2014 (Liu et al., 2014), where he provides a summary and analysis of state of the art on data visualization. The research is a global overview of this topic, which groups together the separate fields and areas of it.  In addition, Liu demonstrates an overview of the Visualization Pipeline model (Fig. 3.1.1).

This framework consists of five main components: (1) *data transformation and data analysis*, (2) *filtering*, (3) *mapping*, (4) *rendering* and (5) *UI controls*. He shows for every component what is the input and output, and how components interact with each other. The input to this model is data collection, which might be unstructured. The main aim of the Data Transformation and Analysis component is to extract structured data from the input data. Data mining techniques can be used in this stage, such as categorization or clustering methods. The purpose is to adopt data and find the structure for next visualization. Next, transformation operations are performed on this data such as remove noise, interpolating missing values, correcting errors in the data. The output of this component is the structured data. Then, it is sent to the filtering component, which automatically or semi-automatically chooses the data to visualize (focus data). The mapping component maps the focus data to geometric data, for example, to geometric forms and their attributes (lines, dots, size, position, color). Then, this data is rendered to image data. On the last component UI controls, a user interacts with image data, in order to explore and understand the data from different sides (Liu et al., 2014).

Table 3.1.1: Components of Framework 1: Visualization Pipeline

| **Components** |
|---|
| • Data Transformation<br>• Data Analysis<br>• Data Filtering<br>• Visual Mapping<br>• Rendering<br>• Ui Controls |

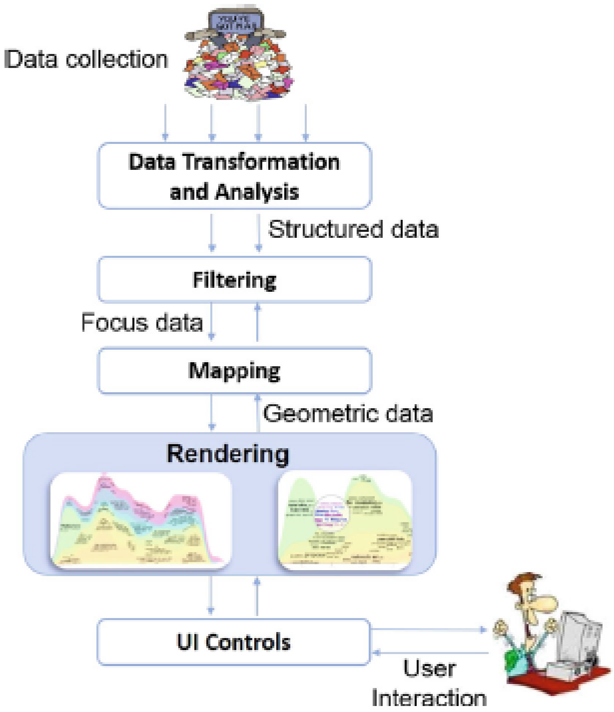



Fig. 3.1.1: Data Visualization Pipeline (Liu et al., 2014, p. 1374)

### 3.1.2 Framework 2: Data Visualization Process

Ben Fry received his Ph.D. degree from the Aesthetics + Computation Group at the MIT Media Laboratory in 2004. The topic of his doctoral thesis is called “Computational information design” (Fry, 2004). The focus of his work is the process of understanding information through the combination of fields, such as computer science, computer graphic, data visualization, and statistics. For the first time, he introduced in his thesis so-called the “Data visualization process” (Fig. 3.1.2). In 2008, he wrote the book “Visualizing Data”(Fry, 2008). The first chapter in the book briefly explains the data visualization process, mentioned in the thesis. The rest of the book discusses the displaying of data in the different visual techniques (such as mapping, trees, time series and so on) and practical advice for Acquiring and Parsing Data steps.

The main motivation for Fry’s doctoral thesis(Fry, 2004) was the fact that the ability to collect, store and manage data is increased, however, the human ability to understand this data remains the same. As a solution to this problem, Fry suggests bringing together the individual fields (particularly computer science, data mining, statistics and mathematics, graphic design, human-computer interaction, information visualization) as part of an integrated process.


Fig. 3.1.2: Data visualization process (Fry, 2004, p. 13)

This Data visualization process consists of seven components, and it introduces fields to which each component belongs. Further description of every component is listed below:

1. *Acquire* – The first component takes into consideration how the data is obtained and how they stored
2. *Parse* – Parse component is transforming raw data into a useful piece of data. It might include pre-filtering (unpacking, decrypting, finding appropriate offset ), parse to valuable structures for a program, later ordering into categories
3. *Filter* – Removing all but the data of interest. This component filters data to receive only useful data for the analysis. It might be needed if the initial data set is too big or a part of data set is not relevant to the certain message.
4. *Mine* – This component helps to see patterns or mathematical context in the data. Statistical (min/max, normalize, standard deviation, sorting, count unique instances) or data mining methods (dimensional measures and transformation, clustering and classification) are used in this step.
5. *Represent* – In this part of the process the data set is represented on a basis for the first time, such as a bar graph, list, or tree. It helps to explore the data.
6. *Refine* – The aim on this component is to refine the basic representation using visual design techniques in order to place attention of the audience to the intended message.
7. *Interact* – The right interaction method helps to convey the message and increase the user's engagement with visualization and the story behind the data.

Table 3.1.2: Components of Framework 2: Data Visualization Process

| **Components** |
|---|
| • Acquiring Data<br>• Data Parsing<br>• Data Filtering<br>• Data Mining<br>• Representing<br>• View Refinement<br>• Interaction Representation |

According to the author, every component goes in the order listed above. However, he points out that it is an iterative process. Depending on a case, it might be needed to come back to any component if later on, a visual designer understands objectives better based on received information. As an example he introduced in his book the project for a zip code numbering system, which illustrates the seven steps, he described. In Fig. 3.1.3, he shows how the following stage component can influence on earlier stage component in this project. With that Fry stresses the importance of tackling the project as a whole (Fry, 2008).

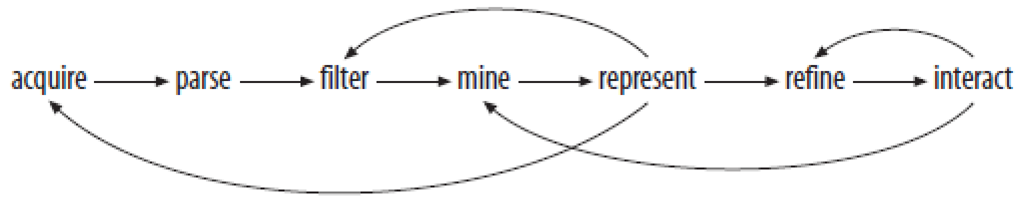



Fig. 3.1.3: Interaction between seven components, example (Fry, 2008).

### 3.1.3 Framework 3: The Nested model for Visualization

Munzner proposed a four nested layers model for the visualization design (Munzner, 2009), also validation with four layers (Fig. 3.1.4). She highlights in her work that the output from a level above is the input for the level below. This work focuses on how to evaluate visualization systems. The main contribution of this framework is to guide the choice of evaluation methodology, which is suitable for validation of each of the design choices made during the visualization process.

The top-level component is 'domain problem characterization'. At this component, a visual designer must learn about a specific domain of the data set, about a user of this domain and specific domain's vocabulary. Although the eliciting requirement might be time-consuming and not easy, this part is necessary for the success of model and visualization at the end of the process. The output of this component is often a set of questions, which a data visualization should answer. The second layer of this model called 'data/operation abstraction design'. In this stage, a visual designer should transfer problems and the vocabulary domain into the abstraction of the computer science vocabulary. The output of this component is the description of necessary data types and operations in order to perform the next component of visual encoding. By operations, it is meant the generic tasks performed with data, such as find minimum and maximum, determine relationships, to understand cause and effect, find noise and correlation, characterize distribution. On the third level, the visual encoding and interactions are performed. The author does not split visual encoding and interaction into the different levels because she considers them interdependent. The inner component of this model is algorithm design, which executes the visual encoding and interaction designs. (Munzner, 2009)

Table 3.1.3: Components of Framework 3: The Nested model for Visualization

| **Components** |
|---|
| • Domain Problem Characterization<br>• Data design<br>• Data Type Abstraction<br>• Visual Encoding<br>• Interaction Algorithm Design |

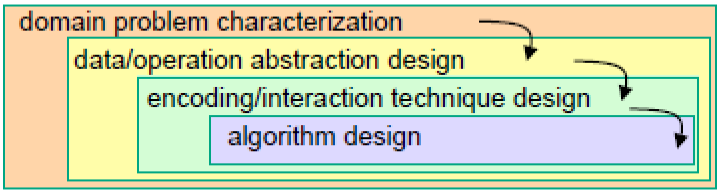



Fig. 3.1.4: The Nested model for Visualization (Munzner, 2009, p. 2)

Fig. 3.1.5 outlines the Munzer's ideas about the threats and validation approaches, which might be needed, on each level mentioned before. Some of the validation for the top level might need the results of the lower level. Thus, it was differentiated two kinds of validation approaches: immediate and downstream. The red line in the Figure shows the connection between a threat and its downstream validation. The direct approaches can be performed before this component, and lower level components were finished. However, they can guarantee only partial success. The downstream is harder to execute, due to the fact that all lower level influences on this validation. However, they are necessary for the good outcome of the model.

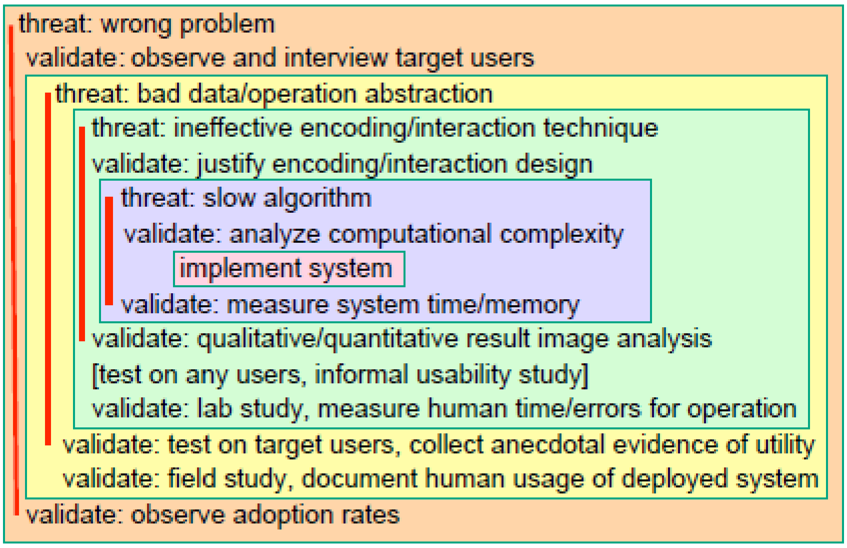



Fig. 3.1.5: Threats and validation in the nested model (Munzner, 2009, p. 3)

### 3.1.4 Framework 4: An information-theoretic Framework for Visualization

Chen and Jänicke developed An information-theoretic Framework for Visualization (Chen & Jänicke, 2010). Their framework is based on the connection between visualization and the applications of information theory. Their study proves that a large number of facts in visualization can be explained by information theory, as visualization's aim is to communicate information with visual encoding. Additionally, that data visualization is needed for data compression, information discovery process; a viewer does not necessarily receive the message; quality of a visualization can be measured by probabilistic theory. All of that shows the conceptual connection between the visualization and information theory because of the both fields code information visually and communicating it.

Furthermore, the authors demonstrate the similarity between the communication pipeline and data visualization pipeline, as well as; they show the distinctions between them. The data visualization mainly focused on the cognitive process of data understanding and knowledge discovery, but communication focuses on delivering the message through the communication channel. (Chen & Jänicke, 2010)

On this model (Fig. 3.1.6), Chen and Jänicke compare the visualization pipeline to a general communication system. It contains three main parts (it is underlined in the figure): the visual encoder, the visual channel and the visual decoder. On the first stage, the visual decoder receives raw data from the source and *filter* them. After he received the information about the dataset, he *maps* data to visual code. Next, the visual encoder *renders* the received geometry forms and labels into an image. Then the image is *displayed* in the visual channel. The optical transmission happens with the optical signal.  A user views the received optical signal. Human vision transforms this optical signal into an image. After that, a human brain can perceive it. During the *perception* phase, it receives certain information out of the perception stage. On the *cognition* stage, the user transforms the received information into knowledge, which is final destination stage on this model (Chen & Jänicke, 2010).

Table 3.1.4: Components of Framework 4: An information-theoretic Framework for Visualization

| **Components** |
|---|
| • Data Filtering<br>• Visual Mapping<br>• Rendering<br>• Displaying<br>• Perception<br>• Cognition |

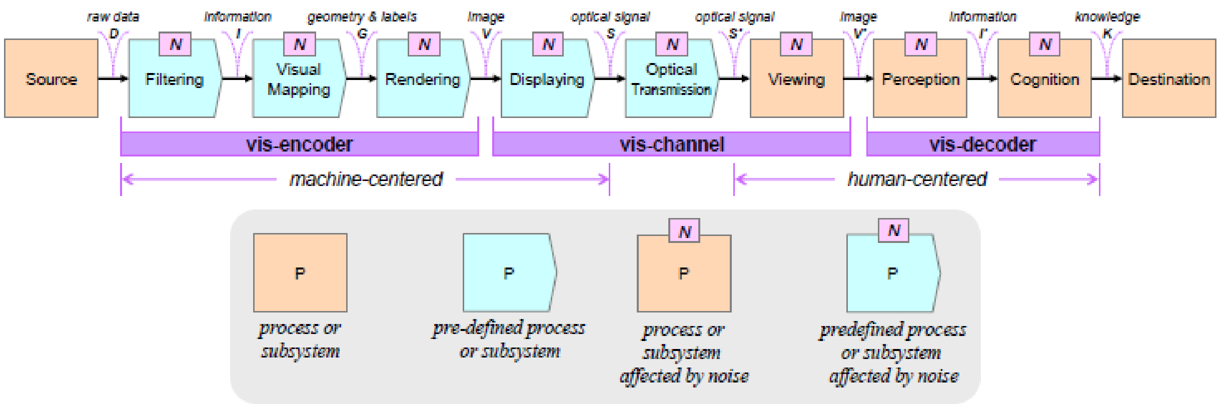



Fig. 3.1.6: An information-theoretic Framework for Visualization (Chen & Jänicke, 2010, p. 2)

### 3.1.5 Framework 5: Quality Metrics Pipeline

In the paper "Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization", Bertini presents a systematic approach to methods for using quality metrics for high-dimensional data in order to help to find the meaningful patterns in visualization. These quality metrics allows the user to concentrate on the most engaged visualizations. In this paper, Berini represents a model that illustrates the process of transformation data visualizations through the series steps (Fig. 3.1.7). This model includes Quality Metrics. It calculated at any step of the visualization pipeline that takes into account properties which are useful for finding important information about the data (Bertini et al., 2011).

This model is focused on Quality metrics of data visualization. It consists of three main components: Data transformation, Visual Structure, View Transformations. First, the source data is transformed to the required format. On this component, the alternative data subset is generated with the help of operations such as feature selection, sampling, and aggregation. The visual mapping component maps transformed data into visual structures. The visual designer generates the ordering. View transformation renders visual structures into pixels. The quality metrics module influences all stages and helps to see the alternatives. At the data transformation stage, it can contribute to generating different data projections or data samplings. At the visual mapping, it uses different types of visual mapping techniques or generates different orderings to see alternatives. At the view transformation, the model displays the modifications of the current view, like highlighting different specific points. The user influences all stages of this pipeline. He sets the quality metrics parameters and explores the result of rendering (Bertini et al., 2011).

Table 3.1.5: Components of Framework 5: Quality Metrics Pipeline

| **Components** |
| --- |
| • Data Transformation<br>• Visual Mapping<br>• View Transformation<br>• Rendering<br>• Feedback loop<br>• Quality-Metric module |

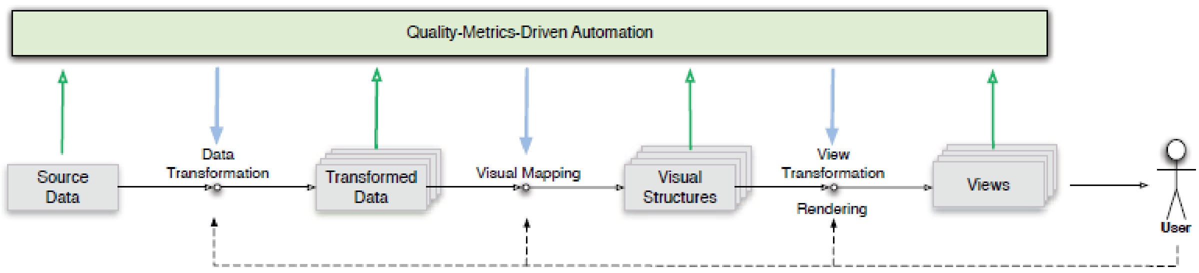



Fig. 3.1.7: Quality metrics pipeline (Bertini et al., 2011, p. 2206)

## 3.2 Analysis of existing frameworks

Ahead of developing a new framework, we should have a clear understanding if there is a necessity to do so. The reason for developing a new framework must be clear too, in order to avoid mistakes which were made by the prior frameworks.

### 3.2.1 Components

Before the following analysis, it should be understood the general summary of elements, which exist on all of introduced before frameworks. These elements are called 'component' for the sake of the common terminology in this thesis. After careful analysis and placing all these frameworks in one place, the list of existing components is made. All the components are contained in the table below (Table 3.2.1). Components from this table, exist in at least one of the previously discussed frameworks.

Table 3.2.1: List of components(modules) observed in the analysis of existing frameworks

| Component | Frameworks | Description |
|---|---|---|
| Domain Problem Characterization | Framework 3 | The understanding of a goal and motivation for data visualization. Analyze the problem which has to be solved. Try to avoid too much information. Domain data knowledge should be understood (Munzner, 2009). |
| Acquiring Data/ Data collection | Framework 2, Framework 1 | Data collection process. The data can be provided by other or find. The main source types are search engines, data from the Application Programming Interface, Universities, Topical data (geography, sport, world, government, and politics), social media. |
| Data Parsing | Framework 2 | The analysis of the strings should be done and brake it on parts. Data formats: delimited text, JSON, XML. Ex. Parse HTML file. (Fry, 2008) |
| Data Transformation | Framework 1, Framework 5 | Common operations on transforming data. It includes: feature selection, projection, aggregation, sampling, removes noise, interpolates missing values, correct error in measurements. |
| Data Analysis | Framework 1 | Preliminary analysis of data set. Normally, it is statistical analysis, such as find max, min, median, normalize, variance, standard deviation, sorting, distance metrics. |
| Data Filtering | Framework 1, Framework 2, Framework 4 | Selection of relevant sub-dataset to visualize |
| Data Mining | Framework 2 | Data mining helps to discover patterns in data. Some methods for mining data are principal component analysis, clustering, classification, searching, scoring methods (Fry, 2004). |
| Data Type Abstraction | Framework 3 | Understanding the type of data(categorical, ordered, nominal) and relationship between variables (Munzner, 2009) |
| Visual Mapping/ Visual Encoding/ Representing | Framework 1, Framework 2, Framework 3, Framework 4, Framework 5 | Core stage for visualization process. Data values are mapped to geometric forms (size, color, direction) |
| Rendering | Framework 1, Framework 4, Framework 5 | Render geometric forms to an image. Many quality metrics applied in the image space (the pixels which generated in the visual process) (Liu et al., 2014) |
| View Transformation/ View Refinement | Framework 5, Framework 2 | Refinement and improvement of the first image. It helps to make it clearer and more visually engaging (Fry, 2008) |
| Displaying | Framework 4 | The representation component, whose roots lie in the field of computer graphics, concerns the mapping from data to representation and how that representation is rendered on display. |
| Interaction / UI Controls | Framework 2, Framework 3, Framework 1 | The interaction component involves the dialog between the user and the system as the user explores the data set to uncover insights. The interaction component's roots lie in the area of human-computer interaction (HCI). (Yi, Kang, Stasko, & Jacko, 2007) |

| | | |
|---|---|---|
| Feedback loop | Framework 5 | The feedback loop from the user. He should be engaged with the process and influence on all stages of the pipeline (Bertini et al., 2011) providing feedback to evaluate and improve the visualization. |
| Quality-Metric module | Framework 5 | The metrics calculated which allows measuring the quality of received information (Bertini et al., 2011). |
| Algorithm Design | Framework 3 | The process of creating an algorithm, which implements the visual encoding and interaction (Munzner, 2009). |
| Perception | Framework 4 | Understanding the process how humans percept visual information. |
| Cognition | Framework 4 | This component is about understanding how received information transforms to knowledge by humans. |

This table is the first step into the frameworks analysis. It gives the first introduction into components from which frameworks consist and their short descriptions. These components could be considered as building blocks or functions for the frameworks. However, the straight list of components does not give a comprehensive understanding of frameworks. The next step shows which frameworks contain which components.

### 3.2.2 Comparison of Frameworks

Even though a full comparison has not been made yet, it can be seen in section 3.2.1 that there are many different framework components. The next task is to perform some kind of comparative analysis to provide a general overview of all five frameworks. This overview will help to see a clear picture of all frameworks and to help identify which problems and limitations these five frameworks have overall.

In order to perform this analysis, the comparison table (Table 3.2.2) is introduced. This table will help to see which components existed on every framework. It also will help with every particular component, for instance, if some of the components are included in every framework or another way around if some component is supported by only one or two frameworks.

In this table, the rows are components, and the columns are the framework. The intersection of the row and the column shows the existence of the certain component in the framework. Color on the intersection of row and column indicates the existence

The notations for the Comparison of Frameworks table is shown below:

| | |
|---|---|
| (red) | Does not exist in the model |
| (yellow) | Exists in the model with some modification (the modification is noted with a number and described below the table under corresponded number) |
| (green) | Exists in the model |

Table 3.2.2: Comparison of Frameworks

| | Framework | | | | |
|---|---|---|---|---|---|
| **Component** | Framework 1: Visualization Pipeline | Framework 2: Data Visualization Process | Framework 3: Nested model for Visualization | Framework 4: An information-theoretic Framework for Visualization | Framework 5: Quality metrics pipeline |
| Domain problem characterization | | | | | |
| Acquiring Data/ Data collection | (1) | | | | |
| Data Parsing | | | | | |
| Data Transformation | (2) | | (4) | | |
| Data Analysis | (3) | | (5) | | |
| Data Filtering | | | (6) | | |
| Data Mining | | | | | |
| Visual mapping/ Visual encoding/ Represent | | | (8) | | |
| Rendering | | | | | (10) |
| View Transformation / View Refinement | | | | | (11) |
| Representation / Displaying | | | | | |
| Interaction / Ui Controls | | | (9) | | |
| Feedback loop | | | | | |
| Quality-Metric module | | | | | |
| Algorithm Design | | | | | |
| Perception | | | | | |
| Cognition | | | | | |

As it might be seen from the table (Table 3.2.2: Comparison of Frameworks) below, certain frameworks have different approaches to the certain components. These modifications are indicated with a yellow color and a specific number. The comments to the numbers are listed below:

- (1): In the Framework 1: Visualization Pipeline, Data Collection component does not occur as the exact function of this model, it is more as the input of the model. The author looks on this component more as a prerequisite to the framework.
- (2) and (3): Data Transformation and Data analysis components are combined into one in the Framework 1: Visualization Pipeline.
- (4), (5) and (6): All the data operation combined into one stage called 'Data design' in the Framework 3: The Nested model for Visualization. After considering the description in the paper (Munzner, 2009), it appears that under this stage, the author meant the components Data Transformation, Data Analysis, and Data Filtering. This approach is not clear from the framework and needs to be revised.
- (7): Data Type Abstraction is merged with the Data Design stage on one-step in the Framework 3: The Nested model for Visualization. Doubtless, these two components are interconnected with each other. Nevertheless, the approach of merging them into one component might cause certain confusion and problems while applying this framework.
- (8) and (9): Additionally, the Framework 3: The Nested model for Visualization merges these two components Encoding Component and Interaction component. The argument is similar to the previous point. Also, the fact that interaction is used not in all visualization supports this argument of the separation of these two components.
- (10) and (11): Framework 5: Quality Metrics Pipeline

The Table 3.2.2: Comparison of Frameworks turns out to be a very powerful visualization tool for current analysis. It shows the general overview clearly. It can be seen already on this stage of analysis that there is a need for the more comprehensive and integrated framework. Also, several issues were noticed in the existing framework. However, going further into the analysis phase, the components will be grouped by areas that they perform in.

### 3.2.3 Views

Previously in the section 3.2.1, many different components were identified. This list of components is not very useful for the future work and analysis. However, the grouping of them into categories will help to structure the contained information, in order to see higher layers out of these frameworks. This approach is closed to Keim's work which discusses the areas related to the visual analytics (Keim et al., 2008); he mentions there the areas such as visualization, data management, perception and cognition, human-computer interaction, evaluation. We also see from this that there is no comprehensive integrated framework, it could be helpful to add a higher level of abstraction.

It can be seen that there are some certain components, which close to each other in one way or another. Such as, for example, data transformation and data analysis are both responsible for working with data.

The components will be grouped into several categories according to the functions that they perform. In a sense, this group reflects the perspective from which we look from in the Data visualization pipeline. Thus, in this thesis, they will be called Views. The list of views and a corresponding list of components which belong to this view is listed in the Table 3.2.3.

Firstly, the group of the components, which is connected with data management, stands out very clearly. This group of operations should be done before any of the visualization components. The Data Collection component is responsible for collecting the data before performing following functions under the data set. Data Parsing and Data Transformation helps to put the data into a form appropriate for analysis. The Data Analysis component performs first preliminary or explorative analysis on the dataset. The Data Mining helps to mine the data and discover the useful patterns in the data. Next, the data filtering component selects the relevant subset of data for future visualization.

The second major group of components is a group that is more related to the visual part such as visual mapping, View Transformation, Rendering and Display. Visual mapping components guide the user to map data values to geometric forms of graphs.

From Framework 4: An information-theoretic Framework for Visualization, two components Perception and Cognition can be seen, as this framework based on Information Theory and understanding of information. These two components can form a third group, which we call Human Cognition view. The Perception components are responsible for the understanding human psychology for the perception of visual codes, such as color or size. The Cognition component helps to understand how the human extracts knowledge from graphs and visual pictures. The designer should always take into account this group because at the end the visualization is made for people who will read the message from the visual picture which the designer will give them. If the user cannot understand the message or understands it in a wrong way, all the work done before will be useless.

The fourth group of components is Quality Metrics module and Feedback loop, which is checking if we receive a good outcome during the visualization. Quality Metrics is about proving that certain standards of quality exist, during each visualization step (such as correctness of the problem, well-designed data abstraction, effective encoding). The feedback loop is directed to receiving feedback from users from every stage of the process, in order to improve our visualization.

Additionally, there are three components which stand alone out of all groups previously discussed. These are Domain Problem Characterization, Interaction and Algorithm Design. The first component is Domain Problem Characterization. This component is responsible for goal and motivation of the data visualization. It has to be understood very clearly at the start of the process because it influences on the success of the whole process. Surprisingly, this component appears only in one of the five frameworks. Although, it might be seen as the Business group which should be researched further in more details, like what aspect of business should be considered during the visualization process (for example, requirements of the project and limitations). The second component is Interaction. In this stage, the user and data visualization can communicate in a certain sense. With interaction techniques, the user can explore the dataset further and discover new insights on the visualization. This feature is not possible on a static picture, which makes this component especially interesting. The third one is the Algorithm design component. This component is responsible for the code to carry out the visualization and interaction decisions. The main issue of this component is that the code should not run too slowly.

Table 3.2.3: Components views

| **Views** | **Components** | **Frameworks** |
|---|---|---|
| Business View | • Domain Problem Characterization | Framework 3 |
| Data Management View | • Acquiring Data/ Data collection,<br>• Data Parsing,<br>• Data Transformation,<br>• Data Analysis,<br>• Data Filtering,<br>• Data mining | All frameworks, but limited<br>Framework 2 supports most of the components from this view<br>Framework 4 and Framework 5 have only one component of this view |
| Visual View | • Visual Mapping/ Visual encoding,<br>• View Transformation/View Refinement,<br>• Rendering,<br>• Representation /Displaying | All frameworks, but limited |
| Programming View | • Algorithm Design | Framework 1 |
| Interactivity | • Interaction/ UI Controls | Framework 1<br>Framework 2<br>Framework 3 (with modification) |
| Human Cognition View | • Perception,<br>• Cognition | Framework 4 |
| Quality View | • Quality-Metric module,<br>• Feedback loop | Framework 5 |

In the next section, based on this analysis a summary analysis of strength and weakness of the existing frameworks will be derived.

### 3.2.4 Strength and limitations

As it was seen from the chapters above, all the frameworks have their specific aspects. To review these different aspects of the frameworks, the analysis of frameworks table (Table 3.2.4) is presented. It brings together and summarizes all the issues discussed above in the chapter.

Table 3.2.4: Analysis of frameworks

| Name of Framework | Strengths | Limitations |
|---|---|---|
| Framework 1: Visualization Pipeline | This framework is a general overview of the visualization process. | The framework is quite straightforward and oversimplified.<br>It does not cover the business view at all |
| Framework 2: Data Visualization Process | The model created by Fry is common in the literature. It focuses not on the visualization as on the whole data process. For each component, Fry mentions to which computer science area this component belongs. | A lot of views have not been covered such as business, psychological, programming and quality views. |
| Framework 3: Nested model for Visualization | The specific strength of this model is that the layers are nested into each other. Each level is based on the previous ones. The problem which occurs on the first level might influence at the final stage of the process. If the problem is not defined correctly at the beginning, it will affect the whole visualization process. | The layers 'data/operation abstraction design' and 'encoding/interaction technique design' are overgeneralized. These are four different aspects which should be considered differently because they have their specific aspects to explore. |
| Framework 4: An information-theoretic Framework for Visualization | The interesting point of this framework is that the author compares the visualization pipeline with a general communication pipeline. The author highlights which part of the pipeline is encoding and which is decoding and how this signal is transferred in the visualization pipeline. Also, he shows which part performed with the machine and which with a human. None of other frameworks distinguishes the actors and encoder/decoder as an aspect in their framework. Additionally, it is the only framework which takes into account the psychology view speaking about human cognition and perception | This framework has strict linear order. Thus the model is not flexible.<br>It does not cover the business view at all and only slightly discusses the data view with the Data filtering component. |
| Framework 5: Quality metrics pipeline | This framework is based on the general visualization pipeline. However, it has an additional layer called quality metrics base automation. This layer receives the information from all the stages of the pipeline and influences on the processes. User influence component goes along to quality metric automation and provides feedback as well. It is only one framework which takes into account quality metrics and feedback loop. | This framework covers many views but only with one component per each view.<br>Additionally, it does not cover the business view at all |

The above table shows that all frameworks have a specific focus. As it might be seen, there is no consistency between the existing frameworks. It emphasizes one more time that a common, more comprehensive approach is needed.

### 3.2.5 Iterative approach

During the analysis, it was noticed that the identified frameworks have different approaches to the order of components. Additionally, the issue of how to go from one component to another is addressed differently as well. Several frameworks allow going back and forth from one component to another. In this thesis, this aspect of frameworks is called Iteration. Almost all frameworks look differently on order and iteration, which might create some misunderstandings.

In order to structure these findings, the Table 3.2.5 was introduced. It illustrates the ways that existing frameworks address these two issues.

Table 3.2.5: Order and Iteration in Framework

| Framework | Order and Iteration |
|---|---|
| Framework 1: Visualization Pipeline | This framework supports some iteration between components. This fact is illustrated in the Fig. 3.1.1 with the arrows between components in the opposite directions. Nonetheless, this question is not discussed in the paper (Liu et al., 2014) |
| Framework 2: Data Visualization Process | The process goes in the order listed in the framework, but with the author's comment that iteration from one component to another is possible, however, this aspect is not implemented in the final framework |
| Framework 3: The Nested model for Visualization | As the model is nested, the author states that output of top level has an influence on the levels under it and all through the model until the bottom level |
| Framework 4: An information-theoretic Framework for Visualization | Strict linear order |
| Framework 5: Quality Metrics Pipeline | Strict linear order |

As is seen from the table above, there is no one single approach to this question of order and iteration. This fact emphasized that this should be developed and standardized.

## 3.3 Findings and Discussions

During analysis, it was shown that each framework has its specific points. These should be considered in developing a new framework. These points are introduced below.

Firstly, Framework 1: Visualization Pipeline has the stage, which is called "Data Transformation and Analysis". According to author's description, it performs "[…] a data reduction technique is applied first. For unstructured data, some data mining techniques such as clustering or categorization can be adapted to extract related structure data for visualization. With the structured data, this module then removes noise by applying a smoothing filter […]" (Liu et al., 2014, p. 1374). Thus, it can be seen that author groups into this step several components, such as Data Transformation, Data Mining, Data Analysis and also data cleaning techniques.

Also, this framework has the UI controls component but not Interactivity component. However, the author describes this component as follows: “Users can then interact with the generated image data through various UI controls to explore and understand the data from multiple perspectives” (Liu et al., 2014, p. 1374). Therefore, Interactivity component and UI control component can be considered the same, and this is a question of terminology.

Secondly, in Framework 3: The Nested model for Visualization, the Data view is quite blurry. The author mentions this stage, but the description is rather short with no detail only references to other works.

Lastly, the main problem with Framework 4: An information-theoretic Framework for Visualization is that there is no detailed description of the framework itself in the paper (Yi et al., 2007). Also, the user will not find the processes in each component and explanation about what this component supposed to do. As a result, the users will have difficulties in applying this framework into practice.

Additionally, it was noticed there are several general problems or inconsistency in these frameworks:

1. Several components are hard to differentiate. There is a thin line between them.
2. Some components on the different frameworks can have the same name but essentially mean a different thing.
3. In addition, it could be another way around. Components have the same name, but the author of one framework defines this component differently from the author of another framework.
4. There is a lack of description of the framework and its components. In several cases, it is difficult to see what the author means by one or another component.
5. Four out of five frameworks do not consider the acquiring component; these frameworks assume that they already have the data.
6. Some views are supported only by one framework. These are the Business view (Framework 3: The Nested model for Visualization); the Human Cognition view (Framework 4: An information-theoretic Framework for Visualization); the Programming view (Framework 3: The Nested model for Visualization); Quality view (Framework 5: Quality Metrics Pipeline)

## 3.4 Conclusion and summary

In this chapter, five frameworks were identified and analyzed. As shown, they have a variety of components. The mapping of the components indicates that there is no single framework, which is very integrative. After careful analysis, all of the frameworks turn out to be very fragmented. In fact, they cover different things. The guidance which they provide (or the depth) in many cases is quite shallow. All the frameworks describe rather briefly which work has to be done in every component. However, none of the frameworks shows the process of how to perform that work. Additionally, the view approach, made in the chapter 3.2.3, illustrates that there is no single framework, which covers all the views listed in this chapter. To conclude, the analysis made in this chapter supports the argument that a more detailed and integrative framework should be developed.

# 4 Developing the Integrated Visualization Analysis and Design Framework

## 4.1 Introduction

Based on the analysis conducted in the previous chapter, it is seen that the existing frameworks for data visualization are quite fragmented and all have some limitations. According to the design science research methodology, the next step is to develop a solution for the problem, which was found and analyzed in the previous chapter. For this reason, in this chapter, a new integrative framework will be developed. The integrative framework should combine the findings from the detailed framework analysis, taking the best features of the five existing frameworks and avoiding their weaknesses.

We need to find the systematic process for creating this data visualization, which might help to manage this process well. For instance, the user should be sure if he is using the right data. The aspect of motivation for visualization has to be clear during the whole process. The visualization allows a user to draw some useful conclusions to address the business needs identified at the start. The main aim is that user can systematically prepare the data to deliver the message or provide something that he/she can use help him/her to understand something.

As a matter of scope, this thesis will not introduce the in-depth description of each component. Since the aim of this thesis is to provide the general methodology which will guide the user the during the visualization process. Several components (such as data mining or visualization evaluation) represent the separate vast field of computer science. For instance, a lot of literature and scientific research exist on the data mining topic, so all the aspects cannot be observed in this work. Nevertheless, the intention here is to show a general overview of activities which can exist in the data visualization field. Also, the aim is to give the direction and the overview to this process, that the data designer could look up the literature necessary for his/her case for the particular case.

During this chapter, the new integrative framework will be proposed. To do so, firstly, the missing components which were founded in the literature will be addressed. For this reason, the next step will be to update the Table 3.2.3: Components views; it will include new components. Secondly, each view will be analyzed and described in more detail based on the existing frameworks and the critical literature review. Each view will have a specific diagram to summarize the main concepts of this view. Finally, all the views will be combined in the whole process, which will be visualized, described and discussed.

## 4.2 Missing components

During the identification and analysis phase, it was noticed that some of the aspects of data visualization were discussed in the literature review, but none of the frameworks mentioned them. The Table 4.2.1: Proposed additional components to introduce the new components, which will be integrated into the integrative framework. These are three new components: Animation, Narrative construct, and Tool choice. They are described shortly in the table below.

Table 4.2.1: Proposed additional components

| Component | Description |
|---|---|
| Animation | Animation helps to highlight or emphasize some points in data visualization. |
| Narrative construct | Methods to convey the message to the audience |
| Tool choice | Choice of a tool for the data visualization |

## 4.3 Updated Views

In the previous section, we added a new component. To keep development stage consistent with the previously made analysis, each new component should be assigned to the appropriate view. For this reason, the next step is to update the Table 3.2.3: Components views. The new components view table is the Table 4.3.1.

Firstly, the Animation component is assigned to the interactivity view because it adds some action to the data visualization. Secondly, the new Tool Choice component is assigned to the Programming view, since the tool choice has an influence on the algorithm and its performance. Lastly, Narrative construct is introduced as a new view, since it out of the scope of existing views.

Table 4.3.1: Updated Components views

| Views | Components |
|---|---|
| Business View | • Domain Problem Characterization |
| Data Management View | • Acquiring Data/ Data collection,<br>• Data Parsing,<br>• Data Transformation,<br>• Data Analysis,<br>• Data Filtering,<br>• Data Mining |
| Visual View | • Visual Mapping/ Visual encoding,<br>• View Transformation/View Refinement,<br>• Rendering,<br>• Representation /Displaying |
| Interactivity | • Interaction/ UI Controls,<br>• Animation |
| Programming View | • Algorithm Design,<br>• Tool Choice |
| Human Cognition View | • Perception,<br>• Cognition |
| Quality View | • Quality-Metric module,<br>• Feedback loop |

| | |
|---|---|
| Narrative View | • Storytelling<br>• Narrative genre<br>• Narrative style |

### 4.4 Description of views

In this section, each view will be discussed in more detail. Every subsection introduces the particular view, and the view representation diagram is at the end of each subsection. Most of the views are displayed as a set of components which this view has, but several views (such as Business and Cognition view) are a set of concepts or ideas which have to be considered when performing this view. It is done for the reason that these views have components which are in separately connected with each other and sometimes it is hard to differentiate them. For example, Cognition and Perception component in Human Cognition view interact between each other, perception strongly influences on the human's cognition; understanding fundamentals can help to leverage it with the help of right perception cues. The analogy for representation of a view of a set of concepts is mindmap technique.

The symbols on the diagrams are: normal rectangle denotes a component; round corner rectangle – a view; snip corner rectangle – a concept or idea (Fig. 4.4.1).

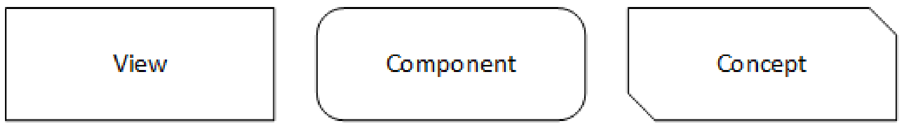



Fig. 4.4.1: Notation for the diagrams in section 4.4

#### 4.4.1 Business Understanding view

Due to the reason that the Business view is not supported by existing frameworks. This view requires elaboration.

At the start, it will be discussed using Munzer's approach on this aspect as it is only one framework which supports this view. She calls this component Domain problem characterization; each domain has a specific aspect and vocabulary which should be well understood by the data designer without any assumptions. Also, the elicitation of system requirements should be done at this stage. The author summaries it in the design study.

Next, the several approaches to prepare for the data visualization process are taken from the literature. For example, regarding the business domain, Marcus suggests to clarify and answer several questions before visualization (Marcus, Feng, & Maletic, 2003):

- Tasks – why is the visualization needed? What tasks should be done?
- Audience – who will read the data visualization? What is the knowledge of the audience about the topic of this data set?

- Medium – where will the visualization be posted?
- Target – what is the data source to represent?

Knaflic highlights that for the successful data visualization, it is important to know the context (Knaflic, 2015). She summarizes this in three words "Who? What? How?".

- Who? It is the audience whom the data designer is trying to communicate the message. The best practice is to narrow down the audience as much as possible.
- What? What should the audience know or do after reading the visualization? Why should the audience be interesting in this data at all?
- How? How can the data be used to convey the point of the visualization? The data should be support for the story or message, and all the non-supporting data is better to leave out.

As Yau writes the data designers "have to know the who, what, when, where, why, and how—the metadata, or the data about the data" (Yau, 2013). In this issue, he suggests to examine next questions:

- Who: The source of the data and its level of trust.
- How: The way the data was collected.
- What: The subject of the data
- When: The time period of the collected data
- Where: The place there was collected. For example, Twitter and Facebook data might have specific aspects.
- Why: The reason for collecting this data and sanity check for bias.

To summarize all said above, there are some business and project aspects which should be considered for the data visualization process. Firstly, it is the motivation for the data visualization, which function the visualization should perform. Secondly, it is all aspects regarding the project, such as the project objectives, requirements, and limitations. Thirdly, the domain knowledge should be clear, as the data designer should become familiar with the topic of this data set. For example, if the data is about the financial situation in the USA for the last decade, the data analyst should get familiar with economic terms, the economic situation in this country and perhaps some significant historical events, which might influence the financial situation in the country. Fourthly, the data designer must understand the future audience who will read the visualization and the media where (source and type of medium) the data visualization will be posted. For example, two visualizations for the web and printed newspaper will be different; similarly, whether it will be published on the news web portal or the corporate website of the company, or perhaps it could be only for internal analysis. Additionally, all information about how data was collected should be known (such as time, a place where the data was collected, and sanity check for the data bias). Finally, the goal must be clear and accurate, what should be achieved with the advantage of the data visualization.

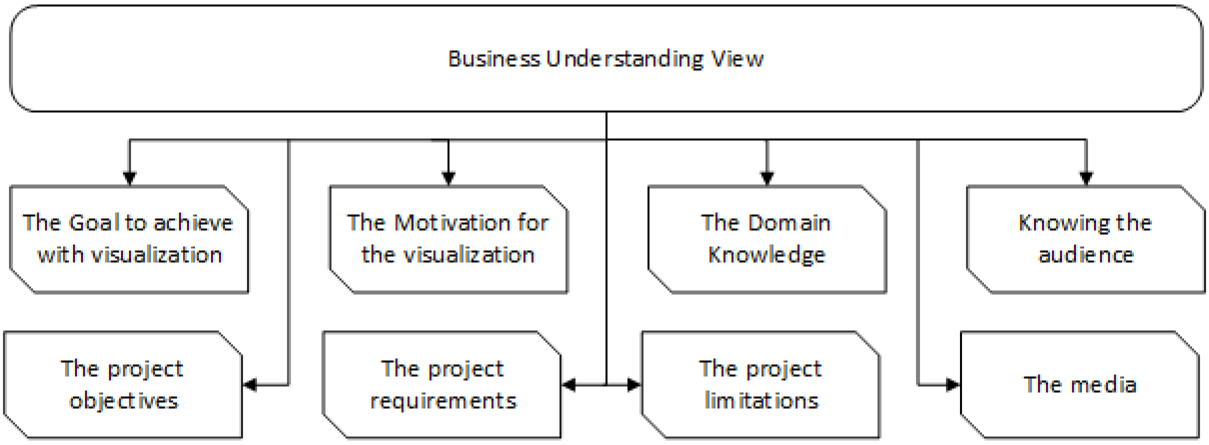



Fig. 4.4.2: Business Understanding Component

### 4.4.2 Data Management view

Data Management view is a second layer of the integrative framework. It includes various components for the data processing before the actual step visualization. Obviously, the data visualization is inseparably linked with Data Operation view, although most of the existing frameworks just briefly mention this view and miss a lot of components of this view. The goal in this section is to bring all the components in one view and to give the first introduction to each component.

**Data Collection**

The first step which has to be performed in the Data Management view is data collection because the data is essential for any visualization. Data collection is the process of gathering data, one of the most straightforward components, another term used for this component is Acquiring Data. Data can be provided from the experts in the certain fields or collected by the data designer. The main sources to acquire the data yourself are search engines, universities, data applications (API of Amazon, Wikipedia, Freebase and so on), topical data (such as geography, sports, the world, government websites). In any case, the data designer must know the context of the data collection and specific aspects this data source has, for instance, the time and place it was collected, if it is poll then who answered.  Also, the possible limitations of the collected data, for example, the certain time frame.

**Data Parsing**

The parsing step involves breaking the raw data into known structures that can be manipulated by software. The result of this step is a set of data structures in memory. That might be parsing structured data, textual data or binary data. Some of the formats are easy to parse (such as RSS feed, XML, TSV, CSV, HTML, JSON and so on) or it could be in the binary format.  The ways of parsing can be divided into categories: text; Text Markup Languages; compressed data; binary data formats. Many tools provide the possibility to parse the data (Fry, 2008).

**Data Transformation**

Common operations on transforming data. It includes feature selection, projection, aggregation, sampling, removing noise, interpolating missing values, correct errors in measurements.

**Data Analysis**

Data analysis gives the first idea about data. Most of the times it is performed with the help of basic math and statistics such as max/min, normalize, finding variance and standard deviation. Depends on the aim, quite often the situation is to count unique instances, calculate distance metrics, sorting numeric values. Building the chart is recommended too. All that gives the first insight into the dataset, to see trends, outliers or other interesting points.

**Data Filtering**

Data filtering component prepares and selects a relevant or interesting subset of the data on which the data designer will perform all the next operation. Sometimes the dataset can be too large or have too much not very interesting information; this step allows to select what is needed (Fry, 2004; Liu et al., 2014).

**Data Mining**

Data mining is whole separate data science field which cannot be investigated in this thesis in depth because due to research limitation. However, Fry suggests two categories of data mining which is particularly useful for data visualization process. The first category is Dimensional Measures and Transformations. It contains methods for managing multidimensional data sets, such as principal components analysis, multidimensional scaling, Fourier transform. These methods detect the most interesting and relevant dimensions in multivariable data sets, which very important in case if we have a significant amount of data. The second category is “Classification, Scoring, & Search“. After the data fit the model and dimensionality is found, this step can be used to find the interesting groups. Clustering methods detect groups of elements similar to each other; dimensional reduction reduces and simplifies multidimensional data; scoring methods determents interesting aspects in the data; search and optimization methods allow to search and optimize the data (Fry, 2004).

The representation of the Data management view is displayed in Fig. 4.4.3. The arrows in the figure show that this view is highly interactive. The user can go from one step to another and back. Also, he/she can skip some steps; it is highly dependent on the data set, the goal and the question which the data visualization has to answer. For example, after performing the data filtering, the user can understand that he/she has not enough data and needs to acquire the data again.

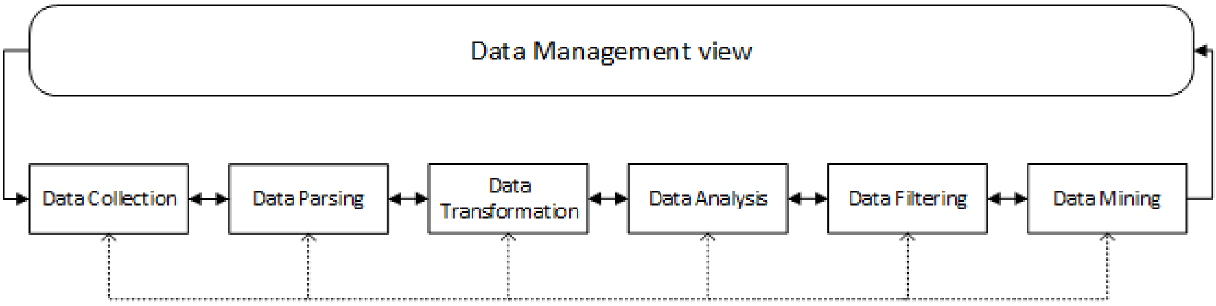



Fig. 4.4.3: The representation of Data Management view

### 4.4.3 Visual view

The visual view is the root part of the data visualization process, and it is directly connected with the future visualization

**Visual Mapping/ Visual encoding**

Visual mapping component transforms the data into geometric forms (charts, bars, lines and so on) and their attributes (color position, size and so on)

The design of visual encodings receives high attention in the information and data visualization literature. Many books and papers discuss the problem of creating charts. Most of the authors recommend certain categories depending on the type of data: Spatial Data, Networks and Graphs, Hierarchies and Trees, Time series data, categorical data, Qualitative, Quantitative,

The chart type is based on the few aspects: the relationship between the variables, the data type of these variables, dimension the data has and also visual encoding. The graphical representation of this idea is shown in the Fig. 4.4.4: Visual Mapping component. To choose the right graph will help existing Graph Selection Matrix, tables and chart suggestion diagrams which were developed by the prominent researchers in this field (Abela, 2009; Few, 2004b, 2012; Heer et al., 2010). Overall chart layout also plays the big role in a representation such as axes, grid, textual annotation.

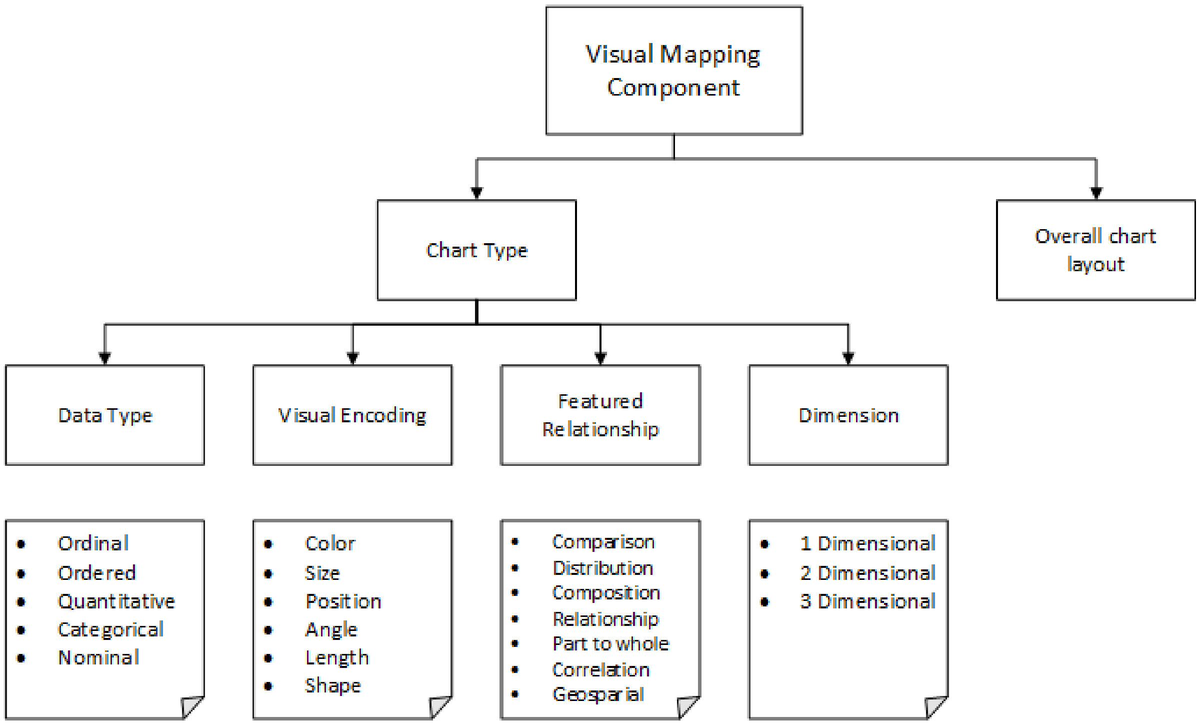



Fig. 4.4.4: Visual Mapping component

**View Transformation/View Refinement**

This component introduces the iterative loop on the work with the visual picture. Few points out for this stage: “it is not sufficient to simply assign a representation to a data set that it fits” (Fry, 2004, p. 110). Each dataset can be represented in several ways, but the task here is to highlight the most interesting features out of the data set by refining the basic representation, which most of the time serve the explanatory goal.

As the most of the time, the data which should be communicated is complex; the data designer should check if there is the better way to represent this data. For instance, the questions which have to be answered here: Is it clear enough? Is it confusing? Is there any clutter? The data designer’s task is to find the best suited the data representation to communicate the goal most efficiently (Fry, 2004).

**Rendering**

Rendering transforms geometric forms into the image with special graphical properties that turn these forms into pixels. The Rendering component shows the importance of the image; the later success of the future visualization will depend on it, and it should satisfy the quality standards (Bertini et al., 2011; Liu et al., 2014).

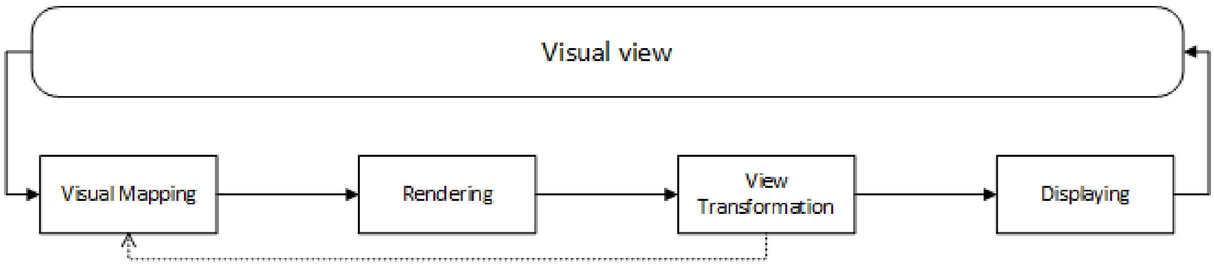



Fig. 4.4.5: The representation of Visual view

### 4.4.4 Interactivity view

Interactivity view is not an obligatory view, but it is recommended. Often data visualization is represented as a static image. However, the interactivity view brings a lot more to the visualization. It can significantly improve comprehension allowing a user to explore the visualization in depth. Thus the user can investigate new insights in the data.

**Interaction**

Certain types of taxonomies on information visualization interaction techniques exist in the literature (Heer, Shneiderman, & Park, 2012; Lee, Isenberg, Riche, & Carpendale, 2012; Yi et al., 2007). They are often intersection with each other because they are based on the tasks and activities, which are supposed to be performed on the visualization.

Yi introduces the most prominent approach. His framework on Interaction techniques in Information visualization identifies seven categories. They are more general and based on a user's intentions to do certain actions with the visualization (Yi et al., 2007):

- Selecting. Allows the user to mark interesting points in the data set, and observe if the representation has changed.
- Exploration. These techniques allow a user to select the part of the introduced data set. These techniques are beneficial in case of the big data set. A user cannot observe a significant amount of data points one at the time due to certain limitations, such as screen, perception, and cognition.
- Reconfiguration techniques change the spatial arrangement of visual presentation, for example reordering data points according to certain criteria. It very often helps to see a characteristic that was not noticed before.
- Encoding techniques empower a user to change the visual encoding to certain data elements, which give the opportunity to see the data in a new representation and to see new insights.
- Filtering techniques allow for choosing a subset out of the data according to certain criteria. For example, filtering population of a country by regions, which allows seeing certain tendencies based on the regions.

- Connecting techniques is used for two aims. The first aim is to emphasize the relationship between represented data elements. The second one is to show the hidden data elements which connected to the specific elements. For example, in case of the country map with hover over a given region can show where people immigrate.
- Abstraction techniques change the level of abstraction by showing more or fewer details. It helps to see a contextual overview. For example, possible to observe the subtree in a treemap visualizations.

A general overview of discussion in Interactive visualization is displayed in the Fig. 4.4.6 . The data visualization community separates two types of interaction: low-level (between the user and the software) and high-level (between the user and the data). The goals in the low-level interaction are show patterns, trends, relationship and the tasks are sort, compute values. (Pike, Stasko, Chang, & O'Connell, 2009). For high-level interaction, the goal is to provide a better understanding; the tasks are reconfigured, encode, connect.

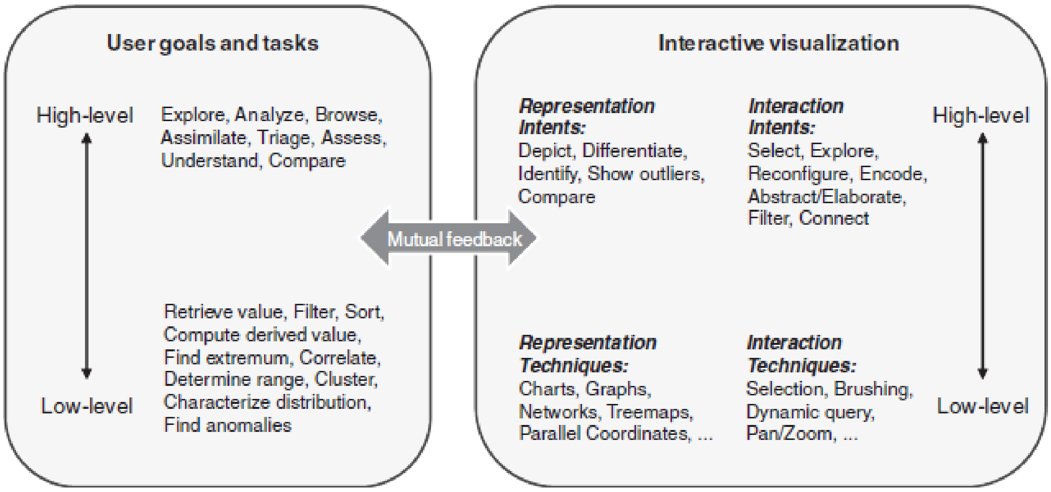



Fig. 4.4.6: General overview of discussion on Interactive Visualization (Pike et al., 2009, p. 266)

**Animation**

The animation in visualization is useful for visualizing a different kind of changes. Robertson and other point out that it is suitable to show: "1) transitions of data from one state to another, 2) transitions between one view and another, 3) illustration of how something works, and 4) trends" (Robertson, Fernandez, Fisher, Lee, & Stasko, 2008, p. 1328). Also, it can bring new perspectives to the time-oriented visualizations (Aigner, Miksch, Schumann, & Tominski, 2011).

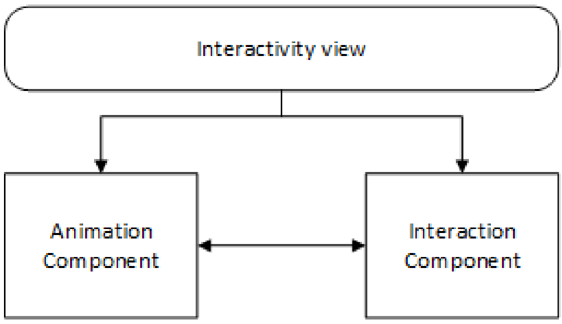



Fig. 4.4.7: The representation of Interactivity view

### 4.4.5 Programming View

Programming view includes Algorithm design and Tool choice components. These two components are put together for the reason that they are interconnected. To explain this fact better, we will start from describing Tool choice.

One of the questions, which is raised during the visualization process, is which tool to use for data visualization. Many tools exist which provides the opportunity to make the data visualization. Nathan Yan in his book "Visualize This" discusses this issue in great detail. He separates three main types of tools: out-of-the-box tools; programming tools; illustration tools. Each type has its advantages and trade-offs (Yau, 2011).

Out-of-the-box tools work by drag-and-drop principle. They are easy to start with, the data designer does not need to learn to work with them. Drawbacks are the loss of some flexibility and lack of functions. However, they are perfect for quick and easy explorative analysis. Examples of this type of tools are Microsoft Excel, Google Spreadsheets, Tableau.

Programming tools require programming knowledge of the certain language, such as R, Python, JavaScript. These tools give flexibility, an ability to adapt to the different types of the dataset and do more with it. However, the main disadvantage is the data designer needs time to learn a new programming language.

Illustration tools (such as Adobe Illustrator, Inkscape, Corel Draw) are not meant to work with data or analyze it. Nevertheless, these type of tools is good for making a visualization more appealing before presenting it to the audience.

A well-implemented algorithm implements a task faster and more efficiently. There are quite a lot of literature is written in this topic, which will help to get familiar with it. Although Algorithm component relates more to computer science field than visualization, a data designer should keep with the question in mind, if he/she uses programming tools. The reason is that a data designer quite often works with a big amount of data, which require a fast algorithm to process.

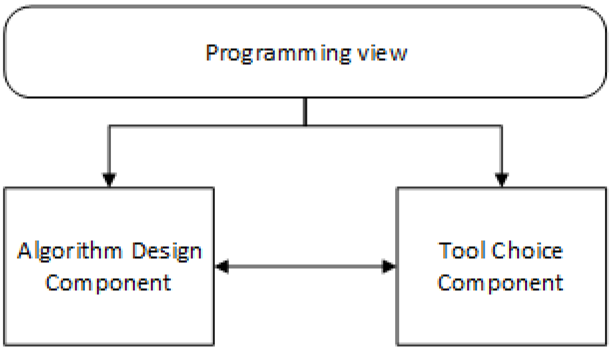



Fig. 4.4.8: Programming View

### 4.4.6 Human Cognition View

The human Cognition view consists of two components. Understanding the human perception and cognition process is crucial in data visualization process because at the end human will read the visualization and read the message, which the designer was trying to convey in this visualization. The designer should never suppose that his/her point of view on visualization is exactly correct because he/she can have a subjective view to the citation that he/she is more familiar with data and message. Thus, the designer should be familiar with the general aspect of human perception and cognition.

Patterson developed a human cognition framework for information visualization. This framework helps to understand how the cognitive process goes to the reasoning process (Patterson et al., 2014). Fig. 4.4.9 shows the information flow in human cognition process. This process starts with visual stimulus, then information goes to attentional capture and encoding. After this, it flows between working memory, long-term memory and pattern recognition. At the last step, a decision is made, and human gives some response to it. Thus, the data visualization is the stimulus in a sense which the reader will process next, and the visual designer should understand some aspects of encoding and memory.

There is three types of memory: iconic memory, short-term memory, long-term memory. Iconic memory is very fast, the information stayed there for a millisecond and passed to short-term memory, and it is set pre-attentive attributes, which are very powerful visualization tool. Short-term memory is limited (four chunks of visual information), so it is not recommended to create the graph with a lot of different colors, shapes, and size. After short-term memory, the information either forgotten or passed to long-term memory, which is formed during all life experience. The aim of visualization is to focus audience attention and stimulate work of long-term memory.

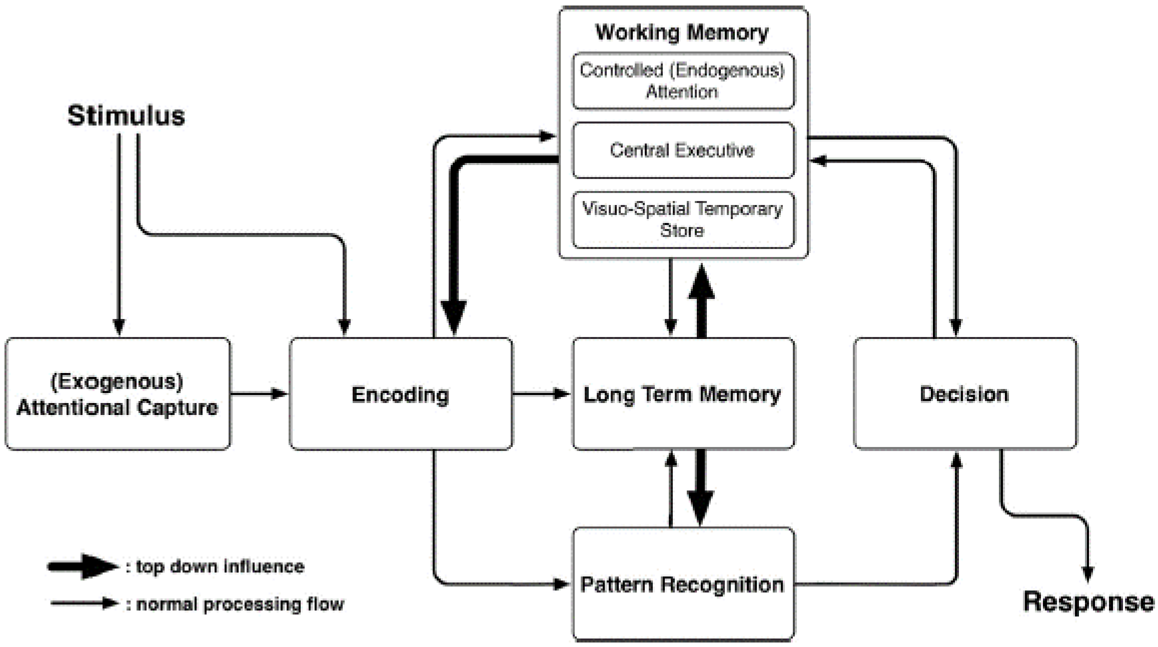



Fig. 4.4.9: Human cognition based on a dual-process framework for reasoning and decision-making (Patterson et al., 2014, p. 45)

However, the main achievement of Patterson's work is that he represents six leverage points that can be used by visualization designers in order to influence on human cognition process (Patterson et al., 2014):

(1) exogenous attention – use main cues (like color) to drives human attention;
(2) endogenous attention – use the right organization to minimize distraction;
(3) chunking – group cues to minimize overload of working memory as it has limited capacity;
(4) reasoning with mental models – provide strong retrieval cues to structure knowledge in long-term memory;
(5) analogical reasoning – structure the information to help faster retrieve the knowledge
(6) implicit learning – develop training regimes for implicitly learning about statistical regularities within a visualization.

Knaflic selects the four most important aspects which allow focusing user attention on the certain points of the data visualization: pre-attentive attributes, human perception of colors, human perception of sizes, position (Knaflic, 2015). A data designer should know how to apply them well during the data visualization process.

**Pre-attentive attributes**

As mentioned above, the fundamental attribute of human sight is that it perceives objects as a combination of simple visual attributes. Humans have attentive and pre-attentive perception. "Attentive" perception requires conscious efforts. In contrast, the basic visual attributes, which are perceived without conscious effort, are called pre-attentive attributes. Pre-attentive processing is very fast and can perceive large amounts of these simple visual attributes (Few, 2004a; Ware, 2010). This fact of human's perception should be used by the data designer to highlight the certain element among a big amount of the same objects. List of pre-attentive attributes displayed in the Fig. 4.4.10.

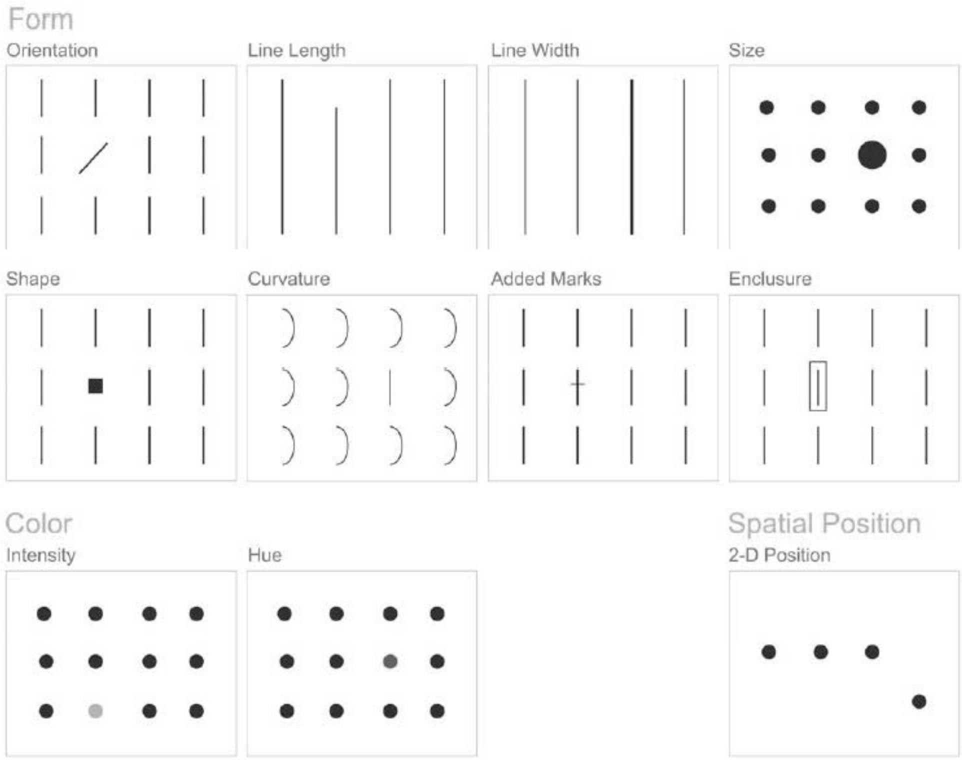



Fig. 4.4.10: List of the pre-attentive attributes (Few, 2004a, p. 5)

Some of these pre-attentive attributes are stronger than others. As many visual encodings exist, the data designer should decide which of them is more valuable. Cleveland conducts an experiment (Cleveland & McGill, 1985), the aim was to understand the ability of a reader to judge the accuracy of different type of visual encoding. He finds out that the reader can distinguish the difference better in position in common scale and the least accurate is color. The results of this experiment are provided in the Table 4.4.1.

Table 4.4.1: Ordering visual encoding by accuracy (Cleveland & McGill, 1985, p. 830)

| **Rank** | **Aspect Judged** |
|---|---|
| 1 | Position along a common scale |
| 2 | Position on identical but nonaligned scales |
| 3 | Length |
| 4 | Angle |
| 5 | Slope |

| | |
|---|---|
| 6 | Area |
| 7 | Volume<br>Density<br>Color Saturation |
| 8 | Color Hue |

**Using Color**

A data designer can use color as a powerful leverage in order to convey the message, or he can ruin the whole visualization only by misusing colors. Using the color in visualization might seem easy at first but it quite tricky task. It shows the important reason to know basic rules for using color in the data visualization.

Few in his work “Practical Rules for Using Color in Charts” (Few, 2008) discusses several main rules for using colors. Firstly, he says that people perceive the colors in the surrounding context. For this reason, the background should be consistent, for example, two objects with the same color can be perceived differently if the background made as gradient color. Also, in case if a text is present in a visualization, the color of background should contrast with the color of letters. Secondly, he suggests using color meaningfully and with restraint. It means that color should be used if it serves a specific communication goal. Additionally, different colors should be used if they correspond to differences of meaning in the data. Thirdly, he suggests using color palettes, which will serve for particular purposes, such soft and natural color for the main information and bright colors to highlight information. Fourthly, he says that for each component should be used appropriate color such as for axis and lines should be used gray color, for background – white color, for bars and data points, a hue of medium intensity. Lastly, he offers to avoid light, shadows, and combination of red and green colors in the same display, in order to help color-blind people to distinguish colors (Few, 2008).

Iliinsky and Steele also note that color has a strong impact on reader’s thought processes and cognitive load; therefore, colors should be used carefully and correspond with the message that the designer is trying to convey. Additionally, they highlight that color is not naturally ordered. Thus it should not be used for ranking or ordering data (Iliinsky & Steele, 2011).

Katz also gives the guidance for choosing effective colors (Katz, 2012). Firstly, she suggests using colors that are distinguishable from each other. Second, colors should be simple to remember and name. Lastly, the context of the certain color is essential depending on the visualization’s topic, such as red color for caution sign, green for nature. It helps to make associations and remember better.

**Using Size**

As a color, size can be the powerful visualization feature, but also has to use very carefully. Iliinsky and Steele recommend using size in order to draw a user’s attention to the central point. However, they say that human brain is good in comparing the areas in situations when the only length is different between them. If the width and length are both different, the reader cannot judge the differences in size (Iliinsky

& Steele, 2011) accurately. For this reason, Few advises to avoid the pie charts; they can be used to compare few parts as a whole, only in cases when there is no need for high precision (Few, 2007).

**Using position**

The data designer should be careful how to position the elements on the page. Most of the people scan some visual image from the top left corner, go to the upper-right corner, next to the bottom left corner and last the bottom-right corner. It looks like zigzag motion (letter Z, Fig. 4.4.11). Thus, it is better to put the most important information at the top left corner if possible. Also, the rule of thumb, scale range should go from negative to positive, from left to right, because this is the way how the most of the people used to perceive (Knaflic, 2015).

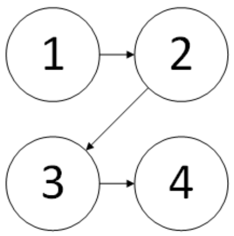



Fig. 4.4.11: Order of perceiving information on a screen or page

The overview on the Human Cognition view is displayed in the Fig. 4.4.12 below.

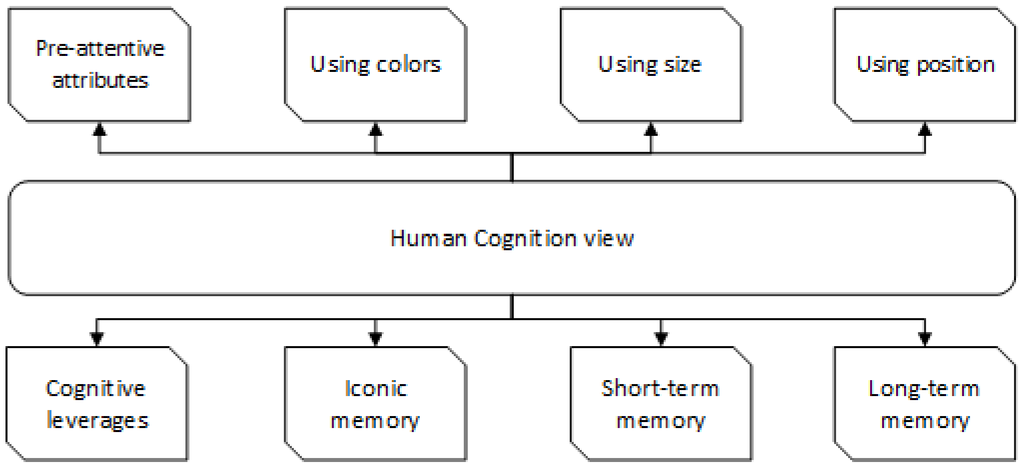



Fig. 4.4.12: The representation of Human Cognition view

### 4.4.7 Quality View

The quality view allows measuring the effectiveness of the data visualization. Also, it helps to predict the success of the visualizations based on objective metrics. It allows to compare different visualization methods, improve the visualization, and check its effectiveness.

The evaluation methods for the data visualization can be split into two main categories: Data-oriented and Task oriented.

Data-oriented methods focus on data sets and visualization techniques, determine how well works a specific visualization technique.

One of the prominent taxonomies on evaluation methods is conducted by Zhu (Zhu, 2007):

1. Data-oriented
   a) Qualitative vs. quantitative evaluation
      a. Qualitative
         - Expert review: heuristic rules. For example, line plot can only be used to express a continuous function
         - Popularity vote
         - Observing users
         - User interviews and questionnaires
      b. Quantitative
         - Scoring points (counting a number of features user can identify in certain time)
         - Task completion time (ask user to perform certain task based on the visualization, the criteria is the time user completed the task)
         - Counting the errors (the same as previous, the criteria is a number of mistakes user has done)
         - Measuring information content
            - Brath's measure of visualization complexity
            - Ward's paper of measuring the information loss
            - Tufte's data-ink ratio
   b) Subjective vs. objective evaluation
   c) User-study based vs. non-user-study based
2. Task-oriented
   a. Shneiderman's seven high-level tasks (Shneiderman, 1996)
   b. Freitas' tasks (Freitas et al., 2002)
   c. Tasks by Pillat et al. (Pillat, Valiati, & Freitas, 2005)

Freitas et al. (Completeness, Spatial organization, Information coding, State transition) tasks are the mixture of qualitative and quantitative evaluation criteria, also the mix of objective and subjective evaluations. The paper does not provide details on how these criteria can be measured.

The disadvantage of the Qualitative metrics is that they can be imprecise. The evaluation based on the heuristic rules can be subjective and biased.

The more novel approach in the visualization evaluation is to focus not on the evaluation methods rather on the evaluation scenarios. This method was developed by Isenberg et al. and Lam et al. by making the extensive literature review in Information visualization field and used over 800 research papers

(Isenberg, Isenberg, Chen, Sedlmair, & Möller, 2013; Lam et al., 2012). These scenarios are different in their goals, research questions and methods. The summary of this approach is displayed in the Table 4.4.2. It has two categories the process and the visualization. In the process category, the aim is to understand the process and the function played by the visualization. The design choice is tested in the visualization category.

Table 4.4.2: Evaluation scenarios (Table conducted by me based on (Isenberg et al., 2013; Lam et al., 2012))

| Category | Name of Scenarios | Short Description | Examples |
|---|---|---|---|
| Process | Understanding Environments and Work Practices (UWP) | The aim is to understand domain expert's analysis practices, the focus here is current practices | Field Observation<br>Interviews<br>Laboratory Observation |
| | Visual Data Analysis and Reasoning (VDAR) | The aim is to understand domain expert's analysis practices, the focus here is the value of an introduced tool | Case Studies<br>Interviews<br>Laboratory Observation<br>Controlled Experiment |
| | Evaluating Communication Through Visualization (CTV) | It checks if and how the visualization deliver the message | Interviews<br>Field Observation<br>Controlled Experiment |
| | Evaluating Collaborative Data Analysis (CDA) | It tests if the tool supports the collaborative analysis and teamwork | Heuristic Evaluation<br>Log Analysis<br>Field or Laboratory Observation |
| Visualization | User Performance (UP) | In this scenario, it is evaluated objective metrics of user performance, such as time completion. | Controlled experiments<br>Field Logs |
| | User Experience (UE) | It tries to understand the subjective user reaction on the visualization | Informal Evaluation<br>Usability Test<br>Field Observation<br>Laboratory Questionnaire |
| | Algorithm Performance (AP) | It tests the generated output qualitatively and studies the performance of visualization algorithm. | Visualization Quality Assessment<br>Algorithmic Performance |
| | Qualitative Result Inspection (QRI) | It addresses the quality of the result, such as an image or encoding quality. | Comparative and isolated result inspections |

### 4.4.8 Narrative construct view

Narrative construct view was not represented in the analyzed frameworks; however, this is an essential aspect for the data visualization. The reason for this is the purpose of data visualization, which is to tell the story or to deliver the message to the audience. Although this view is based on the traditional principals of constructing a story, which takes root from ancient times, it has certain specifics in the data visualization domain. This view focused on the ways to tell a message or story to the audience in a more efficient way.

**Storytelling with Data**

The way of constructing the story has not changed for many years and is also applicable for the data visualizations. The story should have the three main stages (Knaflic, 2015):

- The beginning, tell the plot. The data designer should introduce the context of the visualization and its message. Also, he/she should explain why this should be interesting for the audience.
- The middle, twist. The main problem and message should be discussed on this part, also a possible solution to the problem.
- The end, call to action. On this stage, it should be clear what the audience should do with new information or knowledge, and which actions they must do.

**Narratives**

Narrative helps to organize the story or even multiple stories in the data visualization.

Seven narrative genres for data visualization could be underlined based on the visual layout and the way the data organized in the visualization. These genres are magazine style, annotated chart, partitioned poster, flowchart, comic strip, slide show, film/video/animation (Fig. 4.4.13). These genres are established on two aspects. The first is the number of frames, which visualization contains (different visual scenes, which involved time and/or space). The second aspect is the ordering of its visual elements. The genres do not have to be mutually restricted so they can be nested into one another. Depends on the story type or data, a certain genre can work differently in a certain context (Segel & Heer, 2010).

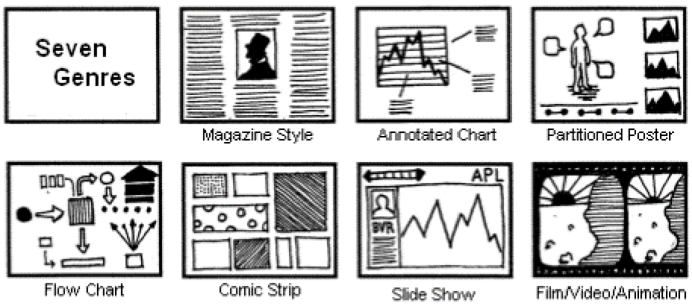



Fig. 4.4.13: Narrative Genres for Data Visualization (Segel & Heer, 2010, p. 7)

**Narrative structures**

The traditional approach to storytelling can also be applied to the narrative visualization. However, the age of new mediums, particularly computers and Internet, have given life to new narrative structures that can often be complex and concentrate on more details. Segel describes two types of Narrative Structures (Segel & Heer, 2010) depending on who directs the story:

- Author-Driven. An author decides how direct user through a visualization. These visualizations are typically linear and have predetermined narratives
- Reader-Driven. The reader picks up the message that he/she needs them self. This type allows the viewer to interact with the data directly to craft their own data story.

**Combination of Author-Driven and Reader-Driven styles**

In their paper Segel and Heer (Segel & Heer, 2010) also describe three ways to combine author- and reader-driven narratives (Fig. 4.4.14):

- Interactive Slideshow. Small narratives in every slide, a reader can observe other data before moving to the next slide
- Drill-Down Story. The general theme is introduced, and every instance can reveal sub-story.
- Martini Glass Structure. The author tells the story, after that, a reader can observe facts which he/she likes. It has two stage. Firstly, viewers start at the story which author gives them (the martini glass stem) continuing the single path. Then, the visualization opens and allows the viewer to explore different details within the data and find their narrative (reader-driven, the mouth of a martini glass).

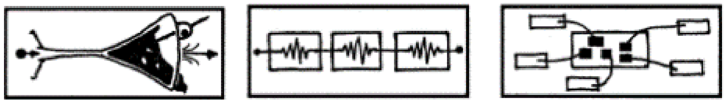


Fig. 4.4.14: Combination of Author-Driven and Reader-Driven styles (Segel & Heer, 2010, p. 7)

All in all, it is important for the data designer to know the narrative types in the data visualization. Firstly, from the journalist point of view, the data visualization has to have the story connected to it. Secondly, the data designer should decide the genre of the data visualization. The genre is kind of layout for placing elements on the page and way it action (it could be static in the Magazine Style genre, or interactive in the Slide Show genre). Lastly, the narrative structure should be chosen: the author direct the reader in the narrative or the author give the visualization and user decide himself what point to explore. The general overview of Narrative Construct view is shown in the Fig. 4.4.15.

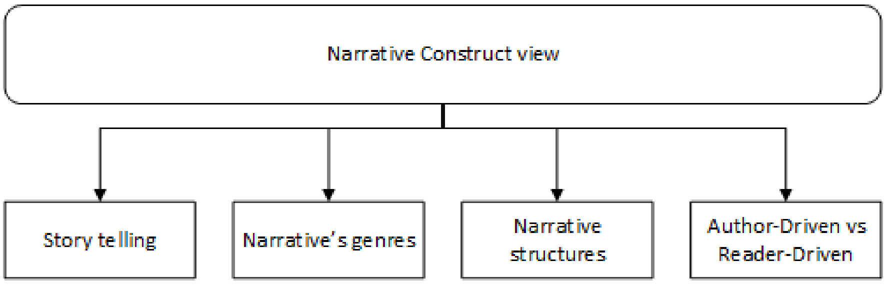



Fig. 4.4.15: The representation of Narrative construct view

## 4.5 Conceptual representation of the views

Several components and views are tightly coupled, and they have the dependencies. The different views are linked together in some ways. To understand the interconnection between them is important because it will help to organize the integrative framework in a way that is more efficient. As was noticed during the framework development, some views are interconnected with each other:

1. Interactivity view and Programming view

   The reason of connection with the Programming view, particularly Tool choice component, is that possible to add interaction to the data visualization depends on the tool. To add interactivity to the data visualization, the data designer should choose the tools, which based on the HTML, JavaScript, and CSS. The business intelligence tools, like Tableau, allows interacting with the data visualization but only inside the tool; it is not possible to bring to the audience the interactive visualization. However, they are suitable for data analysis and some presentation (for example, financial report). Overall, it depends on the goal of the data visualization.

2. Programming view and Data Management view
   The data management operation (such as collect data, filter data, transform data and so on) cannot be done without programming tool.
3. Interactivity view and Cognition view
   Well-designed interaction and animation can assist the cognition process by providing the better understanding for the user (Heer et al., 2012; Lee et al., 2012; Pike et al., 2009; Robertson et al., 2008; Yi et al., 2007).
4. Narrative construct view and business view
   Representation of the message of the visualization depends on the business requirement and the motivation of the data visualization. Such questions as which layout to use, which media to use, how to direct the user in the visualization will be designed base on the
5. Narrative construct view and Interactivity view
   In the section 4.4.8, the narrative genres are introduced. Some of them require the interaction to use them, for example, slideshow. However, the static visualization might be as efficient as

the interactive visualization.  Thus, the decision to use the interaction or no is connected with the method to convey the message.

6. Narrative construct view and Cognition view
   The way the story was constructed has a  strong influence on the user's understanding of the final message (Segel & Heer, 2010).
7. The Data Management influences on the Visual display
   The visuals depend on the data types. For example, the representation of time-series data and network data is completely different, most common chart for the time series data is the line chart, as for the network data is the node-link diagram. In addition, the visual encodings vary on the data type, for instance, for categorical data, it could be shaped; for ordered data, it could be color hue.

Conceptual diagram (Fig. 4.5.1) meant to show how views interconnected between each other. This diagram was inspired by the way mathematicians show the intersections between two or more sets or inclusion. The color does not have a particular meaning; different colors were used with the aim to separate them from each other. This diagram shows the areas, how they are overlaps with each other and what is linkages between the areas. The size tries to demonstrate an amount of effort should be sent to some view. However, it is more about relative efforts or relative importance, and it does not have scientific meaning behind it.


Fig. 4.5.1: Conceptual diagram of the views

## 4.6 The Integrated Visualization Analysis and Design Framework

After careful analysis of each view, we can start to put the views together in the general framework. The general overview of the new integrative framework is displayed in the Fig. 4.6.2 and notations to it in the Fig. 4.6.1.

All views are split into the three categories. These categories are based on the idea how the view is applied and what kind of activities are done in the view:

- Conceptual views. This category is idiomatic; the result could be some document discussion on this view, for example, requirement document (for Business view) or pen-and-paper sketch (for human cognition view);
- Execution views. This category is about implementation discussed concepts in the conceptual phases. It could be filtering the dataset with the help of tool (on Data Management view) or implement detail-on-demand feature (for Interactivity view);
- Design views. This category is a prototyping phase; it is a combination of requirements, limitation.

As the Visual view is the iterative stage, which can be changed few times during the process, it was decided to split it into three substages according to the progress in the visualization. This approach will help to organize the process in the temporal order. These stages are listed below:

1. Visual view: EDA visualization. It is explorative visualization, where the data designer just uses the visualization to get familiar with a data set.
2. Visual view: Draft Visualization. The data designer has already decided which visualization type he will use. In this stage, he/she starts to think about a better way to represent it for the audience.
3. Visual view: Final Visualization. Before this stage, all solutions about design are made; this is stage take care of the layout of the final visualization. For example, the data designer might decide to use several visualizations and put it on the dashboard.

The Business view goes on the top of all other views because it determines the start and motivation for the whole process and has to be followed as a guideline during each view. It plays the role of "light-house" during this process and all the views directed by it in some sense. It also goes as a threat through the whole process; we need to keep in mind the aim of the visualization during the process. The arrows from the Business view to other view are not displayed to keep the diagram clear. For example, Business view influences even on Programming view, because a client can require using a certain tool.

After, the data designer has analyzed the business need, requirements and goals he can start to work with data: gather the data, filter and analyze it, it is shown on the diagram with the arrow, which follows from the Business view and to the Data Management view.

Then, the data designer can start to visualize the data in order to explore it, for example, to see some trends or outliers. Data management and Visual view: EDA visualization can interact with each other, the user has some insights in the data set after the visualization and work with the data set again, for example, to filter it.

After the data designer has the first understanding of his/her data set and the way it could be represented, he/she can continue with the draft visualization taking into account the Human Cognition view and its principles.

If it is necessary, Interactivity view is applied taking into consideration Human Cognition and Narrative view. As it was mentioned before, the interaction can help to improve the human understanding and convey the message or story.

Programming view is used during Data Management view, Visualization view and Interactivity view as it helps to implement these views. The gray arrows in the diagram show how where the programming view is used.

At the end of the process, the data designer can perform an evaluation of the visualization. The final visualization goes as a feedback loop to the business view, where it could be reconsidered.

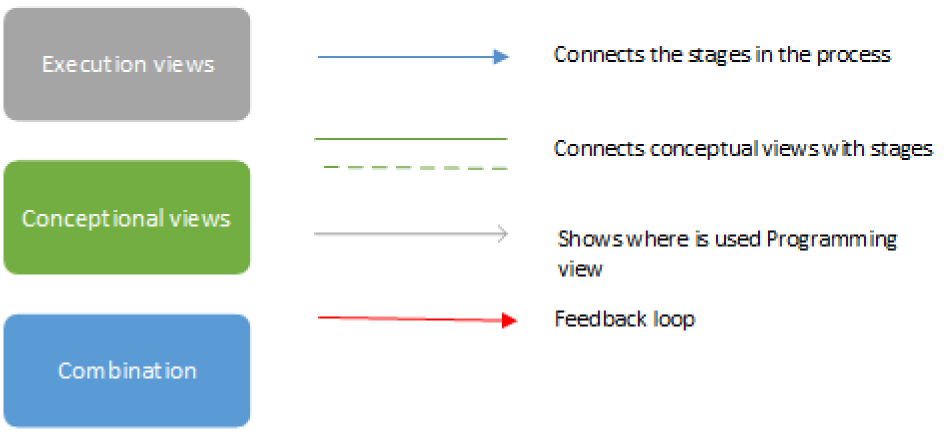



Fig. 4.6.1: Notations for the diagram in the Fig. 4.6.2

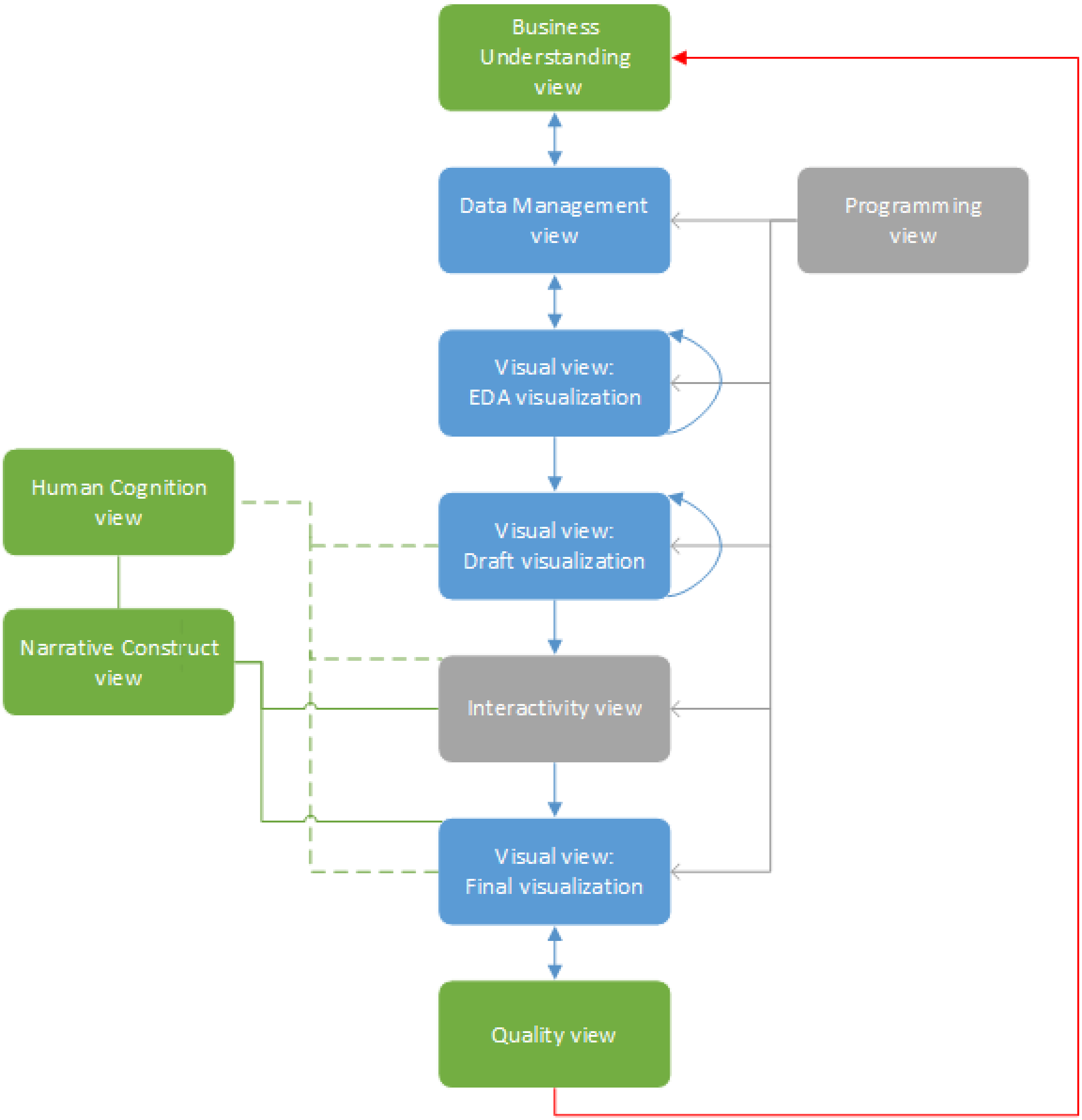



Fig. 4.6.2: Representation of the Integrative Framework.

## 4.7 Discussions

Some controversial issues aroused, during the developing phase of the Integrative framework. The first issue, it is how to keep the process in more temporal order. Some of the views as Human cognition view and Narrative construct view are not possible to put in strict temporal order. They represent themselves the idiomatic views, which are responsible for requirements, limitations, best practices and so on. For this reason, the data designer should use them as certain guidelines during some stages of the process. Programming view is also an arguable moment in the Integrative framework from the point of temporal order. This view can be seen not as a certain stage in the process rather as the tool to perform these stages. It helps to implement Data Management view, Visual view, and Interactivity view.

The second issue, it was noticed that certain phases of work exist in the process. For example, the design phase, which is responsible for the final visualization and picture, or the implementing phase, which is includes programming tasks in different views. However, it was difficult to identify them in this stage of work; more analysis and evaluation are needed to explore this fact.

In addition, the Integrative framework was done based on the previous frameworks analysis and the literature review. This approach is theoretical and gives the good overview of the activities in the data visualization process. Some views or components introduce themselves the separate research field and have an extensive amount of literature inside, so it is hard to observe all the issues related to them. Thus, it might have hidden problems which will arise in practice. Some of the components might be unnecessary; other ones might be missing or incomplete.

## 4.8 Conclusion

The preliminary Integrated framework was developed in this Chapter. The development phase gives the understanding that the views are interconnected, and they communicate with each other in certain phases of work. The evaluation should bring more insight into this issue.

Overall, it is essential to perform the evaluation phase. It will give the understanding of how the integrative framework functions on real case example. The evaluation phase will give the clear picture of applicability of the framework and understanding of its problem, which later allows modifying and improving the framework. In addition, the issue connected with the phase of work will be exanimated in the evaluation stage. In the next chapter, the evaluation of each stage and view of the integrative framework will be done.

# 5 Evaluation of the Integrated Visualization Analysis and Design Framework

In the previous chapter, the Integrative framework was proposed. However, some issues arose during the development phase, such as the temporal order of the process or identifying the phases of the framework. According to the design science research methodology, the next step is to evaluate the solution, which was proposed in the previous chapter, in order to test if the proposed solution applies to real life and to identify its problems. For this reason, the integrative framework will be tested and evaluated in this chapter. The aim of this part of the research is to understand what problems the integrative framework has and how well it works in the real case example. The Integrative framework will be tested using a case of Social Collaboration Analytics using the UniConnect data logs.

The job is not to deconstruct each of these views and list all methods and techniques, which should be used in them. Rather, it is to give more information about what has to be done in general in these views. The result of the analysis should be a clearer understanding of several issues, such as the problems mentioned above; the necessity for more iterations; comprehensibility (is it understandable), completeness (is something more still required?); complexity (should something be removed); limitations.

## 5.1 Evaluation approach

The testing phase simulates the business case between a data designer and the clients; the client has some problem or question about their business, he/she has the data which should solve his/her problem, and he/she requests to analyze it visually from a data designer. In this case, the data designer is the author of this thesis (Anna Presnyakova); the clients are the supervisors of the thesis (Prof. Dr. Susan Williams and Florian Schwade).

The testing phase will be organized as a workshop session. The group will go through each stage of the Integrative Framework from two points of view.

Firstly, the group goes through each stage, as the designer and clients. For each stage, the group meets in the lab with a whiteboard. The data designer keeps the journal of the workshop to summarize what happened on each stage.

Secondly, the group organizes the feedback session, try to reflect and gather the impression what happened in the current cycle. The aim is to understand what were the constraints and limitations, what was missing, what did not make sense, how useful is this stage, how important is it. Reflecting helps to understand what happened during the testing.

Lastly, suggestions for improvement are identified, and the integrative framework is modified according to the results of the testing phase.

## 5.2 Workshops' Journal

The workshops' journal is aimed to capture the evaluation of the Integrative framework (Section 4.6) and developed views (Section 4.4). In order to do this, the data designer should track two quite different sides of the evaluation phase. First, it is to track the information regarding the client's project. This is the information that will help to data designer in the process of the actual data visualization. Second, it tracks the reflection on testing phase of the Integrative framework. It helps to understand what the process better, its elements and components.

### 5.2.1 Workshop 1

The aim of the first workshop is to evaluate the Business Understanding view developed in the section 4.4.1. The objectives for this are to test out a business view; run through the questions that were represented in the literature; then to reflect on these issues and if they were adequate to the problem. The list of the questions for the Business Understanding view is displayed in the Appendix B. Questions for the evaluation Business Understanding view.

**Received Information**

UniConnect is a domain for the proposed visualization. It is research and teaching platform; it is social collaboration platform. The main entity of the platform is the community where the students and researchers exchange the information used in teaching and research. The community has a name, description, status, members, files, wiki, blog, forums, and bookmarks. The members of a community can communicate with the help of blogs and forums; also they can exchange files and bookmarks. The manager can post some information regarding the community in the description box.

The UniConnect platform has two main types of communities: research communities and teaching communities. Mainly, research communities are active for a longer time, because they are updated during the research project lifetime (for example, Research group FG EIM). These type of communities might be active for several years until the research group or project are closed or no longer exist. However, the teaching communities are active for shorter periods of time because they attach to a certain semester (for example, Digital Communication WT15/16). Thus, once the semester is over, these communities are not used anymore.

The clients want a visualization of the communities in UniConnect, which shows them which communities are active and which are inactive. A desirable feature is to see the period of how long the community was inactive. The objective of this visualization is to manage the platform better; it is a sort of housekeeping task for the platform. It should give a clear understanding of what is happening on the platform. More advanced questions, which should be answered, are: What type of communities are inactive? Where are they? Why these communities inactive?

The audience for the data visualization is the management team of the platform. The clients are the audience. Thus, their knowledge about the topic is high, but knowledge about the actual outcome of the visualization is low. They know about the topic, but they do not know about the findings. They would like to know about how many active or inactive communities, which ones are inactive and type of them (teaching or research).

The clients see the visualization as kind of dashboard; activity is distinguished by the color; the period of inactivity is shown in layout with dimension. The visualization should be in online media, but it might be that they want to print it, the usage is mostly internal. They will interpret it and present it; there is no teaching element. The visualization is a snapshot of communities' activity at this point in time.

The system collects the data, and every action is traced. The period of the collected data is twelve months. It contains information about communities mostly in Germany.

To sum up, the main motivation for the clients is the ability to manage inactive communities. The clients want to perform some activities with those inactive communities based on this visualization. It might be three main future actions: to delete inactive communities, to archive, or to set read-only. The client's focus is inactive communities. Thus the visualization should give the names of inactive communities.

**Reflection on the workshop**

The testing process showed that the list of questions was helpful and unhelpful at the same time. The questions should be elaborated on; the clients were not prepared or able to give a clear answer. Thus, the questions should be completed in advance of the meeting with the designer. Also, they were repetitive, but this is a good thing, the group should go through this process of clarifying the purpose. The iterative process helps to hold the picture of what the clients want.

The Business Understanding view is fundamental because it is setting the parameters for the visualization, which is a key thing. This view helps us to focus on what things must to be done. It shows the process we have to go through and the different elements or aspects of the problem. It should not contain the real detailed data level but rather the business level of this data. However, it also has a downside that the group might end up answering the same questions repeatedly.

The testing showed that right at the start of Business Understanding view, a bit of the Data Management view is required. The workshop proves that they are connected. However, first of all, the objectives must be clear; then the data can be observed; lastly, the project can be discussed in a detailed way. However, Business Understanding view cannot be discussed without the data. These two views are linked to each other, which is shown on a diagram, going back and forth between views. The group had to revisit these views, so it is going to be repetitive.

The testing also reveals that views are overlapping. In terms of data gathering, it shows that the phases of work are needed rather than steps.

- Project phase (This is about awareness and planning: what are you doing? What data do you have? Where is the data coming from?);
- Design phase (What are your objectives? How do you meet those objectives?);
- Developing phase;
- Quality phase.

As it was pointed out above, the Business Understanding view has the different phases as well; they need to be highlighted. They are shaded with the color of a group in Fig. 5.2.1 (3, audience and media are part of the requirements). Goal, motivation, domain, project objective group together.

In the end, it is shown that there would not be the perfect process. It will be the process, which has the methodology designed to make it work. The project bubbles and milestones must be defined. Nevertheless, the goals might change as the different visualizations produced might tell more about business but the visualization process cannot be started without the certain base of the project knowledge.

The main finding of this workshop is that the Business Understanding view in a certain way is similar to Software requirement phase in software engineering. Software requirement has two stages elicitation of requirement and its specification. It was understood that in Business Understanding, first, we need to collect the information about the project then to specify it. The group should have a clear specification of the requirements, the clear picture of the project and the clear understanding of the data source before going into details.

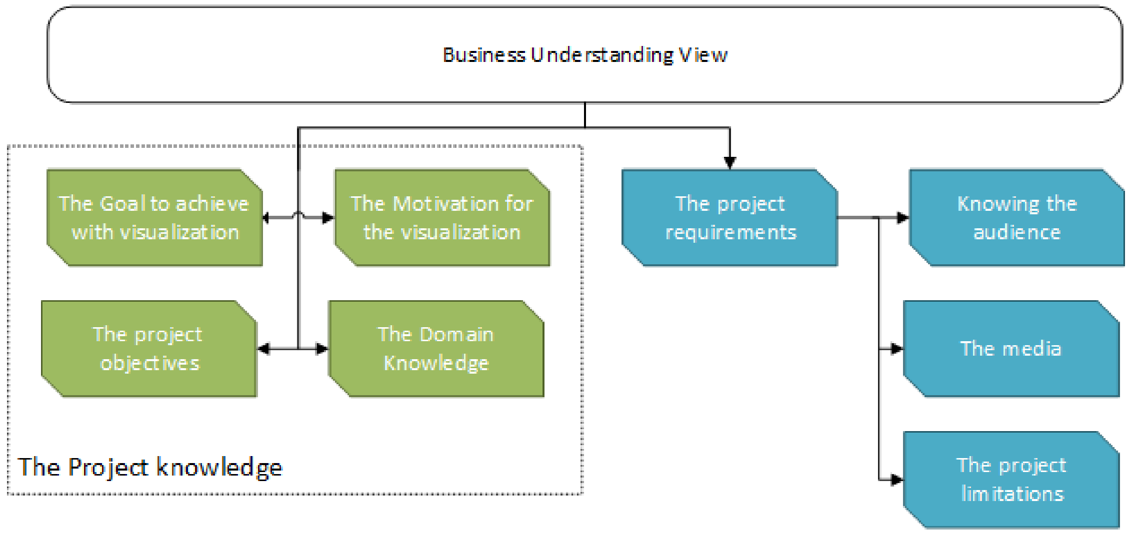



Fig. 5.2.1: Updated Business Understanding view

### 5.2.2 Workshop 2

The second workshop aims to test the Data Management view. The objective of the workshop is to identify which components of the Data Management view (from the section 4.4.2) appear in this particular case. In this particular case, the client provided the data set to the data designer. For this reason, the goal of the second workshop is to discuss which actions were performed on this data set and to understand what kind of analysis the client would like to do.

**Received Information**

Data Management view and its components, which were used in this project, are represented in the Fig. 5.3.2. From the second workshop information was gathered regarding of data management steps in the project:

- The data was collected by the system of UniConnect and provided to the data designer by the clients.
- Data parsing. All data is in the human-readable format. However, there was parsing from the database to CSV format in order to perform future visualization.
- Data analysis. Firstly, it the amount of active and inactive communities might be analyzed. This gives the first picture in analyzing communities in UniConnect by providing the ratio of active to inactive communities in the UniConnect. Secondly, an optional analysis is to see activity rate of the certain community (more active, less active). For example, it might occur that one community has 1000 events for last month and another one has 200 events for the same period. This analysis also will increase the understanding of the activity tendencies. Also, it is interesting to investigate an average amount of the events in communities.
- Data filtering. In general, the data was taken from the database of UniConnect as it was in the system log. However, only one table out of the database is used for the main data visualization. This table has the log of activities, which happened for last 12 month. In this table, the type of activities, which happened outside of the community, are filtered out by the clients.
- No data transformation was performed. Also, the project does not imply any data mining operations.

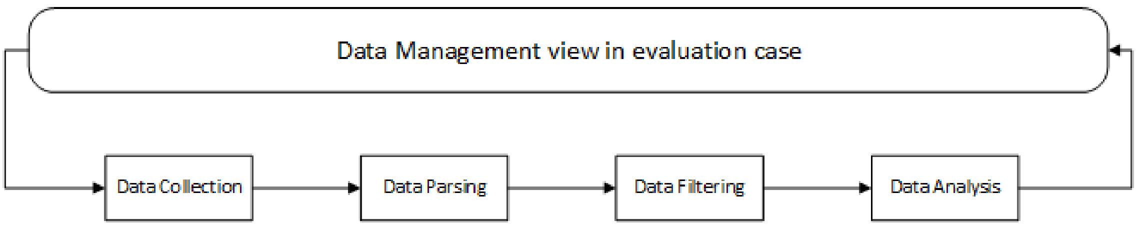



Fig. 5.2.2: Data Management view in evaluation case of UniConnect Communities' Activity

**Reflection on the workshop**

During this workshop, firstly, the way the clients collected the data was discussed. Then, what the clients performed to parsing and filtering before providing the data to the data designer was discussed. Lastly, what type of the analysis the clients would like to be done was discussed. Interesting to note, this brought some ideas and insights about what the future visualization might look.

In general, it showed that there is not a lot of data management processes in this project. As it was discussed in the section 4.4.2, the components of the Data Management view are interactive between each other and not all of them must appear in the visualization process. The workshop showed that this was the case in the project, as Data Transformation and Data Mining components are not needed in this certain project.

Overall, the second workshop did not reveal any fundamentally new insights into the Integrative framework.

### 5.2.3 Workshop 3

The aim of workshop 3 is to evaluate the Visual view: EDA. The objective is to make first mockups of the future visualization. First, the data designer represents his/her ideas for the visualization in the form of the mockup. Second, the clients give their feedback on the visualization, the missing or unnecessary elements of this visualization. Lastly, the group discussed the insights, which this workshop is revealed, on the framework.

**Received Information**

The data designer prepared two sketches (Fig. 5.3.3) for the visualization, which meet the goal of the project. During the workshop discussion, the Fig. 5.3.3 (b) was chosen as a draft for the future visualization as it meets the main objective of the future visualization.

The clients want to see a bit more information about the inactive communities. One of the interesting aspect, which would be nice to have, is the box plot of a number of communities which became inactive in the certain month and also a cumulative graph of a number of inactive communities, to see how many communities became inactive at a certain point in time. Additionally, the visualization should show the total number of active and inactive communities.

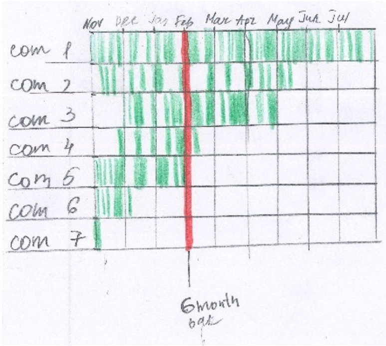
It was discussed that the type of community is not possible to identify in a programming way. One option is to take the smaller sample of the data for University Koblenz-Landau. In this case, two visualizations could be done: one is analyzing entire data set; other is analyzing a subset of data for the university (certain action could be done on the subset, like entering the type of community manually).

Distribution of active and inactive communities

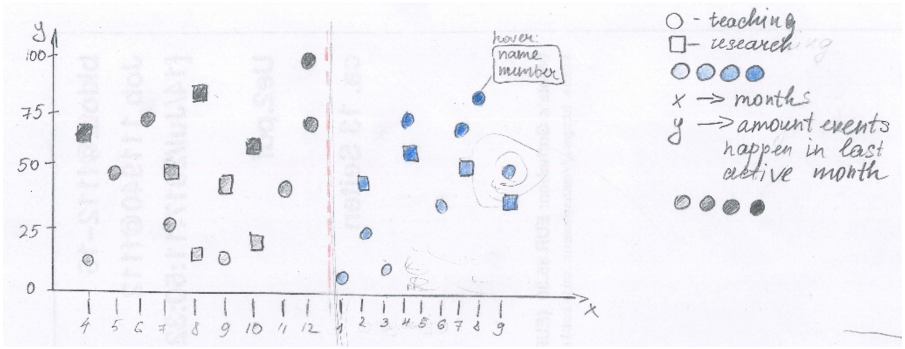
Other variables, which could contribute to the understanding of the inactive communities, are:

- The size of the communities. Is it big or small community, which became inactive?
- Age of community. For example, the certain community had few events after the creation data and became inactive. Alternatively, this community was running for years and suddenly it ended.
- The department. For example, the marketing department set up 100 communities, but they are not active. The engineering department set up 10 communities but they are active.

All in all, the decision was to start with the main basic visualization where the important thing is to see active or inactive communities. Anything else (such as the size of the community, the type of community, and a number of events in last active month) add value and would be potentially useful to see, but these are an optional aspect of the future visualization. The visualization should not be too complicated because the main objective is to see active and inactive communities.




(a)




(b)

Fig. 5.2.3: The proposed mockups for the visualization

**Reflection on the workshop**

The third workshop has revealed a few findings. Firstly, the group should pay attention to the data and metadata regard to the visualization. The situation had happened that the group wanted the certain visualization, but it was not achievable because of a lack of some metadata.

Secondly, the workshop revealed the certain degree of client's involvement in the different views. The clients are heavily involved in the informing views. The clients are not involved in Human Cognition view, though they are involved in Visual view that one is connected with Human Cognition view. The clients do not participate in Programming view, which is a supportive instantiating view, the way of doing it.

Thirdly, it showed how the Visual view engaged with other views. In this stage, Programming view has light engagement, because the data is known and it gives the idea of what needs to be done. However, as the process goes further down, Programming view becomes more dominant, Business Understanding view becomes less dominant or less changeable because the focus is known (What we want to do? How we are going to achieve it?).

Lastly, the workshop shows that it is still analysis, design, and prototyping at this moment of the testing. It is answering the questions: What the clients really want? Do the clients have the data to do this? The group might collect the data again in a different way.

It is worth to notice that the first version of the sketch was made during the discussion of Business Understanding view.

During the testing, the group mainly worked on the Visual Mapping components, Display component and a bit of Transforming component.

After this workshop, the refined version of the integrative framework was made.

### 5.2.4 Workshop 4

The aim of the fourth workshop is to evaluate the Visual view: Draft Visualization. The objective is to discuss a second version of the mockup visualization (which is based on the first version with improvements from the third workshop). First, the data designer represents his/her ideas for the visualization in an improved copy of the mockup. Second, the clients give their feedback on the second version of the mockup; the last version of the mockup has been done based on the discussion on this workshop. Lastly, the group discussed the insights, which this workshop is revealed, on the framework.

**Received Information**

For this workshop, the first version of the sketch (Fig. 5.3.3) was refined. Then the group discussed the second version of the mockup (Fig. 5.3.4).

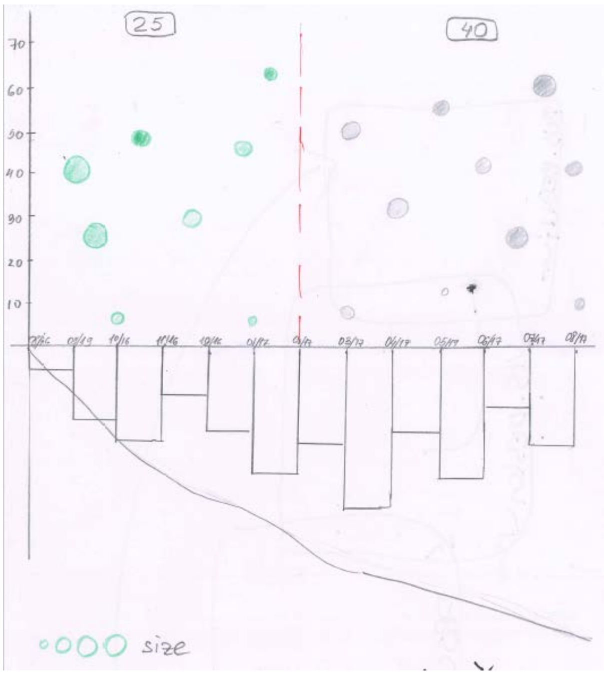



Fig. 5.2.4: The second version of the mockups for the visualization

It was decided to leave out the cumulative graph and to keep the scatter plot and bar chart, which helps to focus on activity in the communities. The bar chart shows the total number of communities which were last active in particular period. It might show the trends when the communities become inactive.

The scatter plot illustrates the spread of communities at the time. If the user clicks on the bar, it shows the detailed information. The size of community indicated with the size of the dot. It will help to relate, for example, if it is a small community with high events, or it is a big one with small events.

The Inactive/active community is the flexible concept at the moment. For this reason, it was decided to use interaction feature to drag the inactivity threshold. For example, the line could be put on six months back or 9 months back. Put the line manually makes sense because it became less specific to this project's definition of inactivity. For example, if the threshold is applied to a different platform, it might be seen that the distribution between active and inactive communities is different and the inactive threshold is slightly different as well.

Two colors will be used (one color to note the active communities, other – inactive ones) but no shading colors will be used. When the user moves the threshold line, the dots change the color to indicate which communities active or inactive according to the threshold, which was chosen. The dots from the left side of the line are inactive; from the right side are active communities.

The second part of the visualization is the detailed table of inactive communities. It has fields such as community id, name, created date, a number of users, last updated. The Table has the information about the inactive communities in the period. It is updated when the user clicks on the bar or moving the threshold line.

All of that forms the interactive dashboard which assists on the main goal of the project to analyze the inactive communities. The final mockup for the visualization is represented in Fig. 5.3.5.

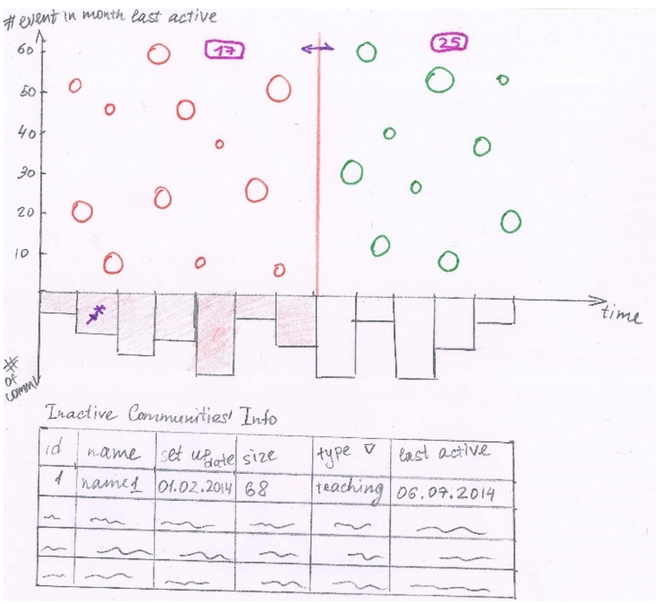



Fig. 5.2.5: The final version of the mockups for the visualization

**Reflection on the workshop**

The workshop showed the views are not steps, rather they are things, which the group has to take into account to lead the process, they are not sequential. The group does not first finish Visual view then Interactivity view. They are overlapping, and the aim is to understand, how they tightly coupled. At the end it might be not activities, they are not necessarily the process.

This workshop proved one more time the presence of phase of works. They are listed below:

1. Business Requirement phase (Business Understanding view, Data management view, Visual view: EDA visualization). Business phase is the business requirements for the visualization. It includes the data, because if these requirements cannot be met without the data. The overlap with next phase is on Visual view: EDA because a draft for the visualization could be done during the first discussion of Business requirements. We are going from business requirement to visual design.
2. Design phase (or prototyping phase). This phase is a hand-and-paper mockup of the visualization. This phase is quite time-consuming if the group creates the new visualization, the different possible and meaningful ways to present this data have to be considered. It brings back to the requirements. It could happen that the requirements are not completely precise enough, the group provisionally gains them, but the group might understand some more information is needed.

   When the group is in Design phase making several mockups for the visualization, it is a lot of iteration, which includes interactivity, the usability, design principles, Human Cognition view. It is not good to decide on those questions at the end of the process. The groups look on these connections, but it does not make the final decision about the color and how it meets cognitive retirements. At this point, the final version of the visualization is not produced. However, it is about drafting what might be possible; this is a refinement phase in order to get in the best visualization.

   The mockup can change several times, so it is not worthy of the efforts programming the visualization, which completely changes in the next meeting. Thus, before the visualization is programmed, the group has to think about the final mockup of the visualization.
3. Instantiating phase. Eventually, closer to the final mockup, the group can move from the paper-based view to the computer view and start to program the visualization. It will be highly iterative.
4. Evaluating phase. This phase includes final Visual view and Quality view connected with the feedback loop to the Business Understanding view. The aim is to review and test the requirements.

Interactivity view should be interlinked with Visual view on the diagram. The lines between Programming view and other views need the refinement. It has the solid line to Data Management view because the data should be collected and processed with the help of the computer tool. The line is dashed to the first draft of the visualization because Programming view is not used while Prototyping phase. The line is solid to the next Visual view because it is need thinking about how the mockup will be programmed. Programming view has a bold line to Interactivity view and Visual view: Final visualization, because these views are not possible without Programming view, also they are in Instantiating phase.

In conclusion, Phases of activities are highly iterative, and all these parts are closely coupled. The process appears to be not a waterfall flow, but it is more an iteration cycle. However, we do move from the sequence in the phase of works. All the refinements, which were proposed during this and previous workshops, are displayed in Appendix A.

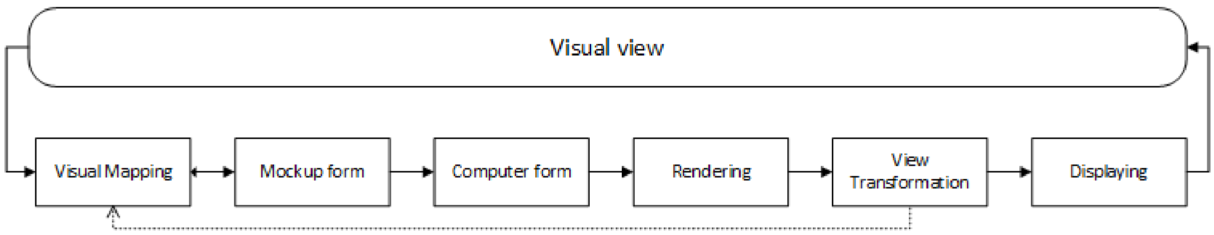



Fig. 5.2.6: Refined Visual view

### 5.2.5 Workshop 5

The aim of the fifth workshop is to evaluate the Interactivity and Programming views. The first objective is to discuss in the more precise way interactivity techniques, which might be used in the project. The second objective is to examine the parts of Programming view and activities it includes. Lastly, the group discussed the insights, which this workshop is revealed on the framework and the views.

**Received Information**

The visualization has the select of the threshold as the interactivity feature. Also, it could be select of time frame, pan across a long time frame drag the picture right/left, zoom in/out, if UniConnect has thousands of communities, it will help to see clusters.

**Reflection on the workshop: Interactivity view**

Interactivity view should be formed as a list of things, which needed to be asked in Interaction component. It could be a checklist for preparing visualization. For example, do you need to select? To go through the list and think would it be helpful in our case.

Interactivity view is another stage of discussion with the user about the visualization. The group, first of all, got the basic idea about the visualization. Then, it is another cycle, where the group needs to say what additional functionality or additional capabilities is needed (for example, the possibility of zoom in). Some of that might came out already in the previous discussions because the user already might say we want to do this. Thus, this is the view but not necessary the step in the process.

This view fairly straightforward. It is another set of questions about the visualization itself. It will be linked to the Programming view, at this stage, we would like to decide we want something animated, highly functional in zooming, in panning. When we have to go back and think the tool we are using, like Excel would not be good.

Interactivity view has two phases inside: mockup and computer-based. The view has the meaning in the process, and it is very closely attached to the design phase.

**Reflection on the workshop: Programming view**

Programming view is driven by the expertise of the programmer. Current programming view has two things inside: Tool Choice component and Algorithm Design component. The Tool choice component has to be first in the order as it defined the algorithm.

The workshop showed that another aspect of the view is missing. This level is the way how it is connected with other views. Programming is making sure that the visualization can be made. Interactivity, Narrative Construct view, Visual view and Data Management view only can be brought together with Programming view. The group can discuss what is needed, but it does not come all together until Programming view. For a reason said above, it was decided to rename this view into Instantiating view. This name reflects better the idea of the view.

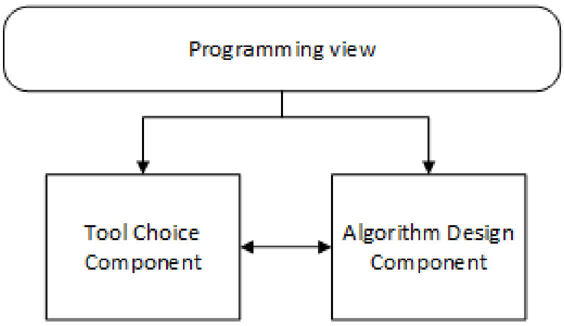



Fig. 5.2.7: Refined Programming view

**Reflection on the framework**

The original goal was to have the pipeline for the data visualization process. As the workshops showed, it is not that at all, it is a very iterative process. The group reached the point where these different necessary views what were identified (they are different ways, the user has to think about and deal with when he/she is creating a visualization). Unimportant views were not found, the group found the views which are more complicated as it seemed on the analysis phase. For instance, the actual Visual view has phases to it. It has seen that the views come together at the different phases of work. For instance, Interactivity view is very closely connected to Visual view. However, the interactivity does not play a role at the start, the group has thought about the data, and how to represent it, the interactivity came in the second cycle.

The views take the different importance during the phases of work, going from no visualization to fully programmed visualization:

- Business Understanding view is important at the start and the end. At the begging, the data designer should think about the objective, requirements of the project and how to represent it with the visualization. At the middle of the work, Business Understanding view has less role because it becomes fixed. In the end, the data visualization should be checked on the subject of meeting those objectives and requirements.
- Human Cognition view is not important at the start. It comes at the beginning of Visual view and when it gets more important. In the end, it is becoming less important because the group is programming the visualization and it became fixed. The group has already know how to visualize and do not need to know about Visual view.
- Programming view starts with Data Management view when it becomes more important once we go to computer-based form visualization.
- The data view is important at the start; then it became less important because we decided what we need and use this data.

It was noticed two different dimensions. The first dimension is activities, which we have to do. They are views, certain perspective. The second one includes phases moving from the requirement to design, to finalize, to evaluate. The methodology is the process of how to going through those views. The process cannot be done before the understanding of the views.

It was also decided to rename the sub-steps of Visual view because the current names not completely reflect the ideas what the group has done in the workshops (Visual view: EDA visualization - Visual view: Visual Mapping; Visual view: Draft Visualization - Visual view: Mockup Form; Visual view: Final Visualization - Visual view: Final computer-based form).

### 5.2.6 Workshop 6

The aim of the last workshop was to evaluate the last three views: Human Cognition view, Narrative Construct view and Quality view. The objectives are to see how these views coordinate in the visualization process. Additionally, this view has to reveal the final insights in the Integrative Framework.

**Reflection on the workshop: Human Cognition view**

Cognition view is an information or aspects, which have to be thought in Visual view. This view might come in question form in the methodology.

After discussion Business Understanding view with the clients, the designer might go back and show the visualization which he/she has done, and the client says the feedback. For example, the designer recommends using certain colors because they are more accessible to default. Alternatively, the data designer might present the other type of chart from what the clients initially wanted, because it is the better way of presenting this type of information.

The section 4.4.6 tells about the different cognition aspects and how to use them in the visualization. The clients cannot say a lot on this view. The designer can say the things, which should be taken into account. Many cognition issues are already addressed in predefined visualization types, for example, the default colors are usually set, they are good for colorblind people; the choice of scale is usually that something is reasonable.

For the project of analyzing UniConnect communities, the default starting scale might be in certain point, the user can move left or right or zoom in and out, but the starting point should be the most meaningful point. If this project had ten years of data, the user could not understand, so the data designer should come to some default point.

In conclusion, Human cognition factor is driven from the designer side but with the requirements in mind from the user.

**Reflection on the workshop: Narrative Construct view**

Narrative Construct view might come in the different stages of the visualization process. In this particular case, it was discussed in Visual view (Mockup form) and Business Understanding view as well. The data designer can ask clients the question: Are you trying to tell the story or do you want to leave it open to exploration? This view is also expert driven.

**Reflection on the workshop: Quality view**

The workshops revealed that Quality view has two levels. The first level is about the quality of the visualization to display; it is similar to usability testing. The group assessing if the visualization is complete; if it conveys the right amount of information. It can be done using heuristic evaluation methods or quantitative measure methods (like a scoring point). For example, the data designer asks the people how many communities were last active in May. He/she sets them a task, if they found out the answer, the visualization works. If it took them 15 minutes to find out, the visualization is not working well. This level is about evaluating the visualization itself.

The second level is about the evaluation if the visualization meets the business requirements (matching requirements). The questions should be answered here: Are all requirements met? If some of them are not met, why? They might be not being met because the group had no data, or it is not possible to visualize on particular way the group wanted? Is it delivering the information we wanted?

The Business Understanding view was missing from the pipelines from the literature. Therefore, it is missing throughout the whole process and particularly at the end, when the group meets the requirements.

**Reflection on the framework**

Firstly, we just see the views and their relationships to each other. Rather have them all on one diagram, decided to split them up.

The business people are involved in the business phase, it covers the whole thing but it most important at the beginning and the end. In the middle, the designer is working with the programmer to get the thing. The designer has a significant role, he/she involved right at the start gathering the requirements because of the designer expertise, but the designer role is more towards the middle. They are the different perspectives. The business phase is what do the clients want, the design phase: how do the experts design it, and programming phase how do we make it happen. Why do we want it? What is the best visualization and how we achieve it? What technology is available and how can we be used to achieve these three things? It is just the additional perspective on these views.

## 5.3 The Revised Integrated Visualization Analysis and Design Framework

This section represents the revised version of the Integrating framework, summarizing the information described in the workshop sections.

Firstly, the frames, which show the phase of work, were added to the diagram

- Business Requirement phase – the orange frame;
- Design phase – the dark red frame;
- Instantiating phase – the violet frame;
- Evaluating phase – it includes Visual view: Final computer-based visualization; Business Understanding view, Quality view and the feedback loop (it does not display with the frame due to the clutter on the diagram).

Secondly, the several views were renamed on the diagram, because they do not convey the meaning correctly. After renaming, they are:

- Visual view: EDA visualization - Visual view: Visual Mapping;
- Visual view: Draft Visualization - Visual view: Mockup Form;
- Visual view: Final Visualization - Visual view: Final computer-based visualization;
- Programming view – Instantiating view.

The arrows between Instantiating view and other views are refined:

- The arrow to Data Management view is solid and means the computer tool is required to collect and process the data.
- The arrow to Visual view: Visual Mapping is dashed. It means that there is no need for the computer tool during the prototyping phase.
- The line is solid to Visual view: Mockup Form because it is need thinking about how the mockup will be programmed.
- The arrow to Interactivity view and Visual view: Final computer-based visualization is bold because these views are in Instantiating phase.

The final diagram of Integrated Visualization Analysis and Design Framework represented on Fig. 5.3.1.

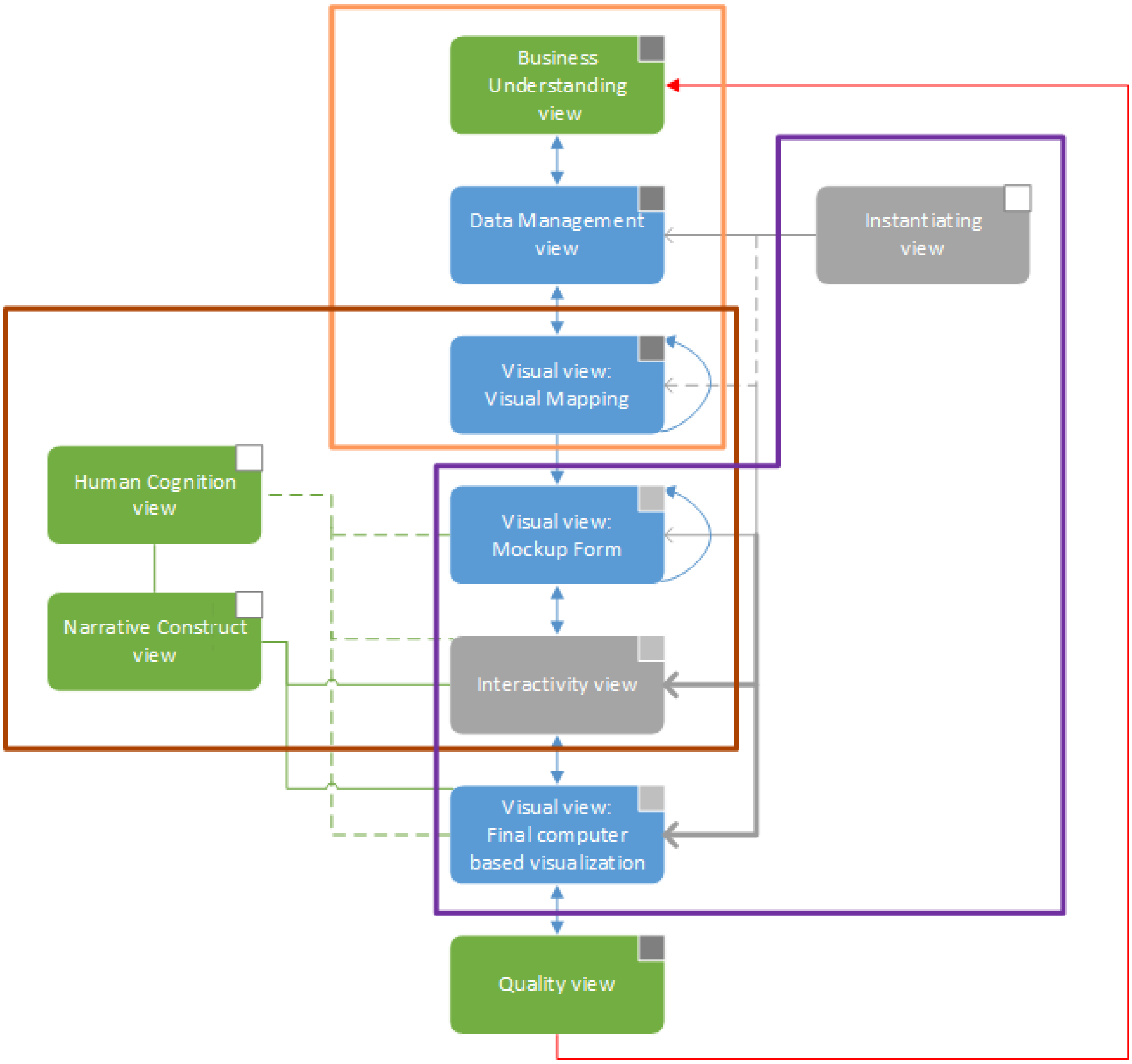



Fig. 5.3.1: Final Version of Integrated Visualization Analysis and Design Framework

Additionally, the decision was to represent the framework with four separate diagrams each to the separate phase, in order to avoid the clutter on the diagram.

The first phase is Business Requirements phase (Fig. 5.3.1). The business expert and the clients discussed the project concept, the requirements, also the data and its metadata. It is important for the future work to discuss the data in this phase in order to meet all the objectives. It will help to avoid the situation when the client wants the certain visualization, but there is no necessary data to represent it. The client or the business expert might suggest the first sketch of the visualization on this phase.

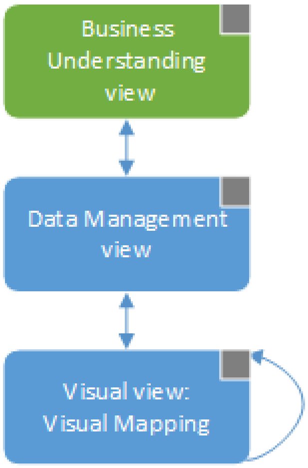



Fig. 5.3.2: Business Requirement phase

The next phase is Design phase (Fig. 5.3.2). The mock-up for the future visualization is developed in this phase. The designer plays the main role in this phase. He/she develops several mockups for the visualization based on the requirements and first sketch from Business phase. During producing the mockups, the data designer takes into consideration the cognition and the narrative principles. The client plays less role during this phase. The client might give the feedback to the designer mock-up, but the designer is an expert of this phase. The client and the designer might discuss the first ideas for interaction in the visualization. The result of this phase is the final mock-up of the visualization, which the data designer represents to the programmer.

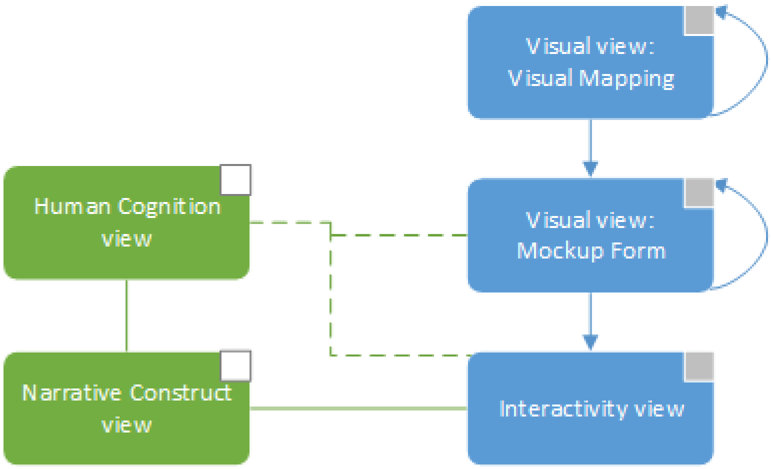



Fig. 5.3.3: Design phase

In the instantiating phase, the programmer plays the main role. At the begging of the process, the programmer might collect the data and process them. He/she takes mockup from the data designer and reproduce it in the computer-based form. The programmer might propose different interactivity features. The result of this phase is the fully computer-based visualization.

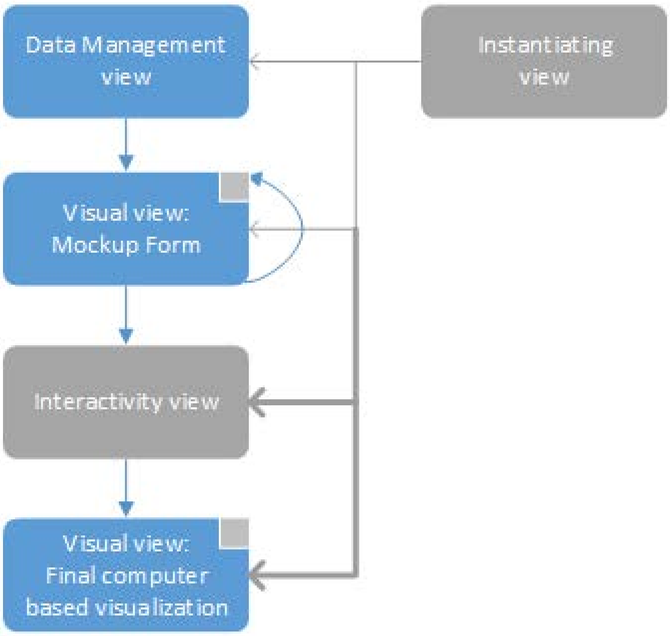



Fig. 5.3.4: Instantiating phase

The last phase is Evaluation phase. On this phase, the business expert and the client interact with each other again. The final visualization should be evaluated on two levels. The first level is to check if it conveys the information well. The second level is to check if it is meeting the business requirements.

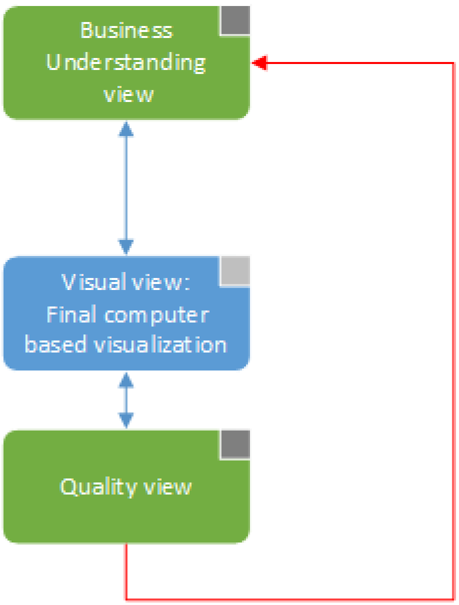



Fig. 5.3.5: Evaluating phase

## 5.4 Conclusion

The main outcome of the evaluation phase is that it is demonstrated that the visualization process cannot be shown as waterfall pipeline, and it represents itself the certain cycle of activities. The developed views show these activities. The views provide the information, examples of different activities or considerations, which need to be thought about during the visualization process. However, the cycle has the certain phases of work rather than steps, which are responsible for the different dimensions of the visualization process. These phases are Business requirements, Design, Instantiating and Evaluation. Business requirement phase is in charge of the gathering requirement from the client; Design phase – for the prototyping of the visualization; Instantiating phase – for making computer-based visualization; Evaluation phase – for assessing a quality of the visualization project from two sides: a quality of the visualization itself and matching requirements defined by the client at the initial stage of the visualization process.

All in all, each view is important, each takes different importance at different phases of work. The views are tightly coupled, so it is not a pipeline going from one view to the next view. Thus, they are coming with groups may be one view dominating, and more works need to be done in the future to develop the methodology.

Additionally, the evaluation revealed that the different stakeholders are involved in the process. Such as, the business expert influence more during Business requirement phase; the designer influences in the process during Design phase; the programmer is more important during Instantiating phase; the clients communicate with the business expert in Business Requirements phase.

In conclusion, the evaluation phase demonstrated that the thesis has the raw material to build the proper methodology. However, more appliance work must be done to do so (testing with 6-7 visualizations). The views are important; they each bring something in the visualization (Section 4.4). However, the framework is not stepwise process, and it is highly iterative, sometimes considering two to four views in the same phase of methodology.

# 6 Conclusion and future work

In the beginning, the addressed problem was to develop the process or pipeline to guide the process of creating a data visualization. At the first stage of the work, five data visualization frameworks from the literature were analyzed. Framework 1: Visualization Pipeline focuses on the data and oversimplified. Framework 2: Data Visualization Process focuses on the data and does not cover over aspects of the visualization process. Framework 3: The Nested model for Visualization is the only one which has a focus on the business part of the visualization. Framework 4: An information-theoretic Framework for Visualization focuses on the visualization as part of the transferring information process. Framework 5: Quality Metrics Pipeline focuses only on quality metrics. Even from the first glance on these frameworks, it is seen that all of them has a different focus.

However, the analysis phase helped to identify the building components of these frameworks. The components are responsible for performing certain stack of actions in the visualization process. The analysis clearly showed the frameworks have some commonalities between them. Such as that, all framework has Visual Mapping component (as it is the core step in the visualization process). Nevertheless, many components exist in one to three frameworks.

Additionally, careful analysis of the components showed eight different views of the visualization process. These views are stacks of components, and they represent the certain dimension of visualization. For instance, Data Management view is responsible for the data and includes the components, such as Data filtering, Data Mining, Data Collection, Data Transforming and so on. The analysis of the views shows that none of the frameworks support all views.

Therefore, the analyzed frameworks appeared to be incomplete and fragmented; they had no detailed description, their depth is shallow. Because the frameworks have different focuses, none of the frameworks shows the process in all the aspects of the data visualization. Thus, the framework analysis supports the argument that a more detailed and integrative framework should be developed.

Then, in the development phase, these eight views were examined in the greater depth. Each of these views brings certain information into the visualization process. The preliminary version of the visualization process was build based on the eight views. The framework was built to a certain extent, but the development phase showed that it is problematic to build the temporal order framework because the views are tightly coupled with each other. In order to build the process, we need to understand fully how the views are related to each other, how they overlap.

Additionally, the process of building these views into the framework showed the different phases of work in the process. The views can be involved in the several phases and take different importance on each phase. Also, the views can communicate between each other depending on the phase.

In deepening the analysis, examining these views in more depth, we identified cycling or recursive nature of the process, stakeholder involvement, phases. It proves that we do not have the perfect stepwise process because the views bring the different things, which are tightly coupled to each other. However, these views provide quite a lot of information, which have to be taken into account when we do the visualization, they provide the information examples of different activities or considerations in the visualization process. The cyclic nature means it is not a pipeline or waterfall type of method, it is very iterative and finding the best solution takes many iterations. Different stakeholders involved during this process, also they take different importance at the different time, and this shows in the phases different activities happen.

As a result, this thesis has developed the raw material, which could help to make better visualizations by providing information about the views. Thus, the checklists or guidelines for the visualization process could be created based on this work, in order to improve the visualization by providing the guidance to the users. The achieved outcomes are valuable. Also, it explains the reason why some of the literature frameworks are noticeably superficial. They were oversimplified because it is considerably complicated to build the process. All their focus on the narrow thing.

In the end, this work gives the better idea of the process and the richer idea of how the visualization process is functioning. However, we do not have clear guidance and methodology. The methodology can be developed based on this research. If we had tried to build the methodology from existent frameworks, we would have lost many details about the components and views.

## Bibliography

Abela. (2009). Chart Suggestions — A Thought-Starter. Retrieved from http://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf

Aigner, Miksch, Schumann, & Tominski. (2011). *Visualization of Time-Oriented Data*. (1st Edition, Ed.). Springer.

Anscombe, F. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Bertini, E., Tatu, A., & Keim, D. (2011). Quality metrics in high-dimensional data visualization: An overview and systemization. *IEEE Transactions on Visualization and Computer Graphics*, *17*(6), 2203–2212.

Chen, M., & Jänicke, H. (2010). An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1206–1215.

Cleveland, W. S., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science*, *229*(4716), 828–833. https://doi.org/10.1126/science.229.4716.828

Few. (2004a). Data Presentation : Tapping the power of visual perception. *Intelligent Enterprise*, *13*(7), 33.

Few. (2007). Save the Pies for Dessert. *Visual Business Intelligence Newsletter*, 1–14.

Few. (2008). Practical Rules for Using Color in Charts. *Visual Business Intelligence Newsletter*, 1–13.

Few, S. (2004b). EENIE , MEENIE , MINIE , MOE : Selecting the Right Graph for Your Message. *Intelligent Enterprise*, *7*(14), 35.

Few, S. (2012). Designing Effective Tables and Graphs. Retrieved from www.PerceptualEdge.com

Freitas, C. M. D. S., Luzzardi, P. R. G., Cava, R. A., Winckler, M., Pimenta, M. S., & Nedel, L. P. (2002). On Evaluating Information Visualization Techniques. *Computer*, *7*, 7–8. https://doi.org/10.1145/1556262.1556326

Fry, B. (2004). Computational Information Design. *SciencesNew York*, *Ph. D*(May 1997), 170.

Fry, B. (2008). *Visualizing Data*. *American Scientist* (Vol. 92).

Gapminder World. (n.d.). Retrieved June 15, 2017, from http://www.gapminder.org/world/

Heer, J., Bostock, M., & Ogievetsky, V. (2010). VIISUALIZATION A Tour through the Visualization Zoo A survey of powerful visualization techniques , from the obvious to the obscure. *Communications of the ACM*, *53*(5), 59–67. https://doi.org/10.1145/1743546

Heer, J., Shneiderman, B., & Park, C. (2012). A taxonomy of tools that support the fluent and flexible use of visualizations. *Interactive Dynamics for Visual Analysis*, *10*, 1–26. https://doi.org/10.1145/2133416.2146416

IBM - What is big data? (n.d.). Retrieved May 11, 2017, from https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Iliinsky, N. P. N., & Steele, J. (2011). *Designing data visualizations*. O'Reilly.

Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., & Möller, T. (2013). A Systematic Review on the Practice of Evaluating Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2818–2827. https://doi.org/10.1109/TVCG.2013.126

Katz, J. (2012). *Designing information: Human factors and common sense in information design*. John Wiley & Sons.

Keim, D., Andrienko, G., Fekete, J., Carsten, G., Melan, G., Keim, D., … Carsten, G. (2008). Visual

Analytics : Definition , Process and Challenges. *Information Visualization - Human-Centered Issues and Perspectives*, 154–175. https://doi.org/10.1007/978-3-540-70956-5_7

Keim, D., Kohlhammer, J., & Ellis, G. (2010). *Mastering the Information Age: Solving Problems with Visual Analytics*.

Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business*.

Koch, K., Mclean, J., Segev, R., Freed, M. A., Berry Ii, M. J., Balasubramanian, V., & Sterling, P. (2006). Report How Much the Eye Tells the Brain. *Current Biology*, *16*, 1428–1434. Retrieved from http://www.cell.com/current-biology/pdf/S0960-9822(06)01639-3.pdf

Kuechler, W., & Vaishnavi, V. (2007). A Framework for Theory Development in Design Science Research : Multiple Perspectives Science Research : Multiple Perspectives. *Journal of the Association for Information*, *13*(6), 395–423.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Lam, H., Bertini, E., … Lam, H. (2012). Empirical Studies in Information Visualization : Seven Scenarios To cite this version : Empirical Studies in Information Visualization : Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics (2012)*.

Lee, B., Isenberg, P., Riche, N. H., & Carpendale, S. (2012). Beyond mouse and keyboard: Expanding design considerations for information visualization interactions. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2689–2698. https://doi.org/10.1109/TVCG.2012.204

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization: recent advances and challenges. *Visual Computer*, *30*(12), 1373–1393.

Marcus, A., Feng, L., & Maletic, J. I. (2003). 3D Representations for Software Visualization. *In Proceedings of the 2003 ACM Symposium on Software Visualization*, 27–ff. Retrieved from http://www.cs.kent.edu/~jmaletic/papers/softvis03.pdf

Munzner, T. (2009). A Nested Process Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921–928.

Murray, S. (2012). *Interactive Data Visualization for the Web*. O'Reilly Media.

Nørretranders, T. (1999). *The user illusion : cutting consciousness down to size*. Penguin Books.

Patterson, R. E., Blaha, L. M., Grinstein, G. G., Liggett, K. K., Kaveney, D. E., Sheldon, K. C., … Moore, J. A. (2014). A human cognition framework for information visualization. *Computers & Graphics*, *42*(August), 42–58.

Pike, W. A., Stasko, J., Chang, R., & O'Connell, T. A. (2009). The Science of Interaction. *Information Visualization*, *8*(4), 263–274. https://doi.org/10.1057/ivs.2009.22

Pillat, R. M., Valiati, E. R. A., & Freitas, C. M. D. S. (2005). Experimental study on evaluation of multidimensional information visualization techniques. *Proceedings of the 2005 Latin American Conference on Human-Computer Interaction - CLIHC '05*, (October), 20–30. https://doi.org/10.1145/1111360.1111363

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of Animation in Trend Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1325–1332. Retrieved from http://ieeexplore.ieee.org/document/4658146/

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. https://doi.org/10.1109/VL.1996.545307

Ware, C. (2010). *Visual thinking: For design*. Morgan Kaufmann.

Wilkinson, L. (2005). *The grammar of graphics*. Springer.

Yau, N. (2011). *Visualize this : the Flowing Data guide to design, visualization, and statistics*. Wiley Pub. Retrieved from https://books.google.de/books/about/Visualize_This.html?id=CB9XRIv9oigC&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false

Yau, N. (2013). *Data Points: Visualization That Means Something*.

Yi, J. S., Kang, Y. A., Stasko, J., & Jacko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1224–31. https://doi.org/10.1109/TVCG.2007.70515

Zhu, Y. (2007). Measuring effective data visualization. *Advances in Visual Computing*, 652–661.

Zoss, A. (2017). Introduction to Data Visualization: Visualization Types. Retrieved June 19, 2017, from http://guides.library.duke.edu/datavis/vis_types

## Appendix A. The diagram's versions of the Integrated Visualization Analysis and Design Framework

The table below represents the record of changes for the Framework diagram. It gives the information about the diagram's version, short description, changes, which were done from the previous version, and the phase of work when this particular version was done.

| **Version Number** | **Figure** | **Description** | **Changes** | **When it is done** |
|---|---|---|---|---|
| Version 1 | Fig. A.5.5.1 | The first version in the Developing phase. | The first draft. | On the developing phase, after creating and analyzing all views (3.2.3). |
| Version 2 | | The second and last version in the Developing phase. | Changed representation of the views in the diagram;<br>Added the color of the block diagram to show the existence of the different phases of the process. | After discussion of the first version of the diagram |
| Version 3.1 | Fig. A.5.5.2 | The first improvement in the Evaluation phase. | Involvement of the client in the process. | After discussion on Workshop 3 |
| Version 3.2 | Fig. A.5.5.3 | The second improvement in the Evaluation phase. | Added frames for the phases of work;<br>Links between Programming view and other views were changed;<br>Links between Interactivity view and other were changed. | After discussion on Workshop 4 |
| Version 3.3 | Fig. A.5.5.4 | The third improvement in the Evaluation phase. | Corrected Frames for the phases of work;<br>Renamed Programming view and sub-steps of the Visual view. | After discussion on Workshop 5 |

The Fig. A.5.4.1 was the first version of the integrative framework. During the discussion with supervisors, it was noticed that there is a need in a temporal diagram, in order to perform the testing phase in a way that is more efficient. This is a reason why in the final version, Visual view was split into three sub-steps. In addition, it was found that the certain phases exist in this diagram, such as conceptual phase (Narrative and Human Cognition view, they are more about the requirements, limitations) and execution phase (Programming and Interactivity view; they are about implementation). Based on this discussion was born the idea of Conceptual diagram (Fig. 4.5.1) and the idea to use the colors on the final diagram (Fig. 4.6.2) in order to show the perspective of the certain view. Also, it was noticed that Programming view helps to implement several other views, such as Data Management, Visual and Interactivity view.

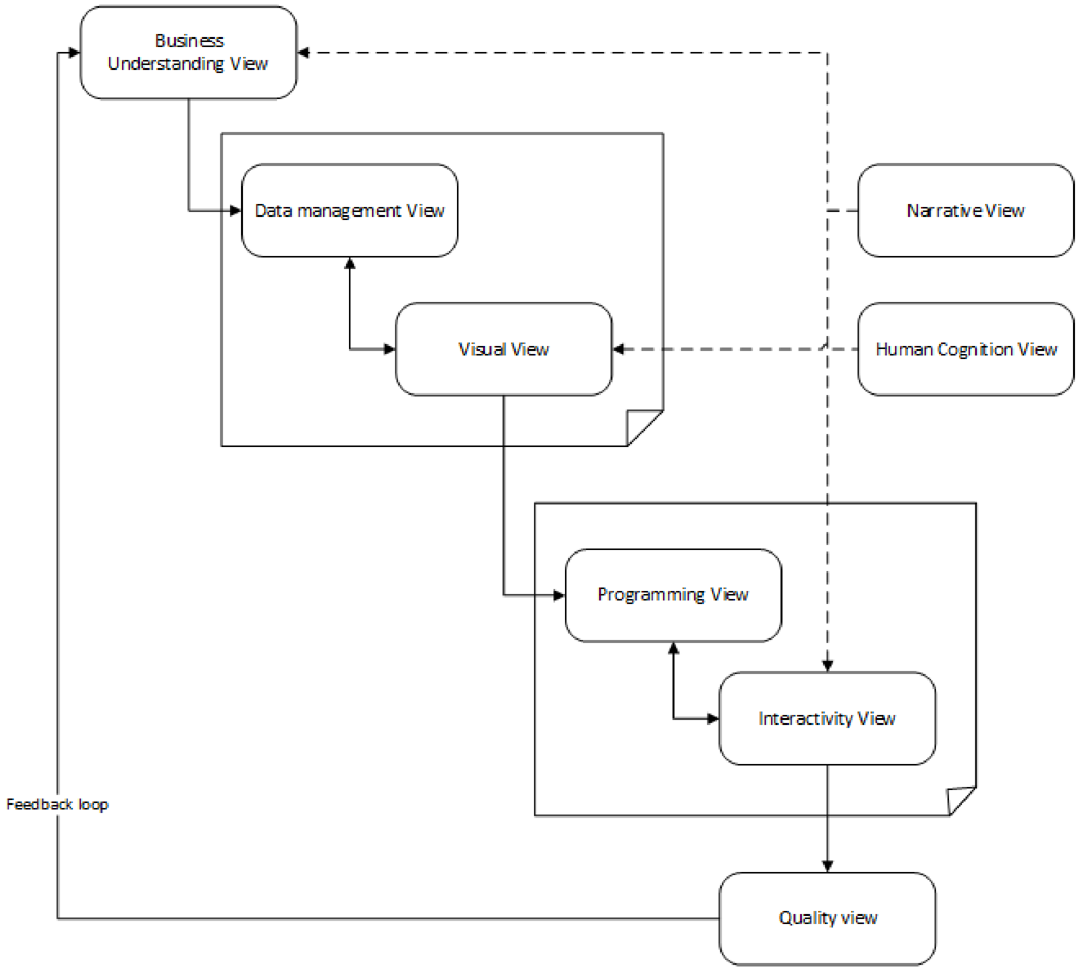



Fig. A.5.4.1: Version 1

In the Fig. A.5.4.2, it was added the discussed concept of the client's involvement in the process. The darker the square the heavily client involved; the lighter – less involved.

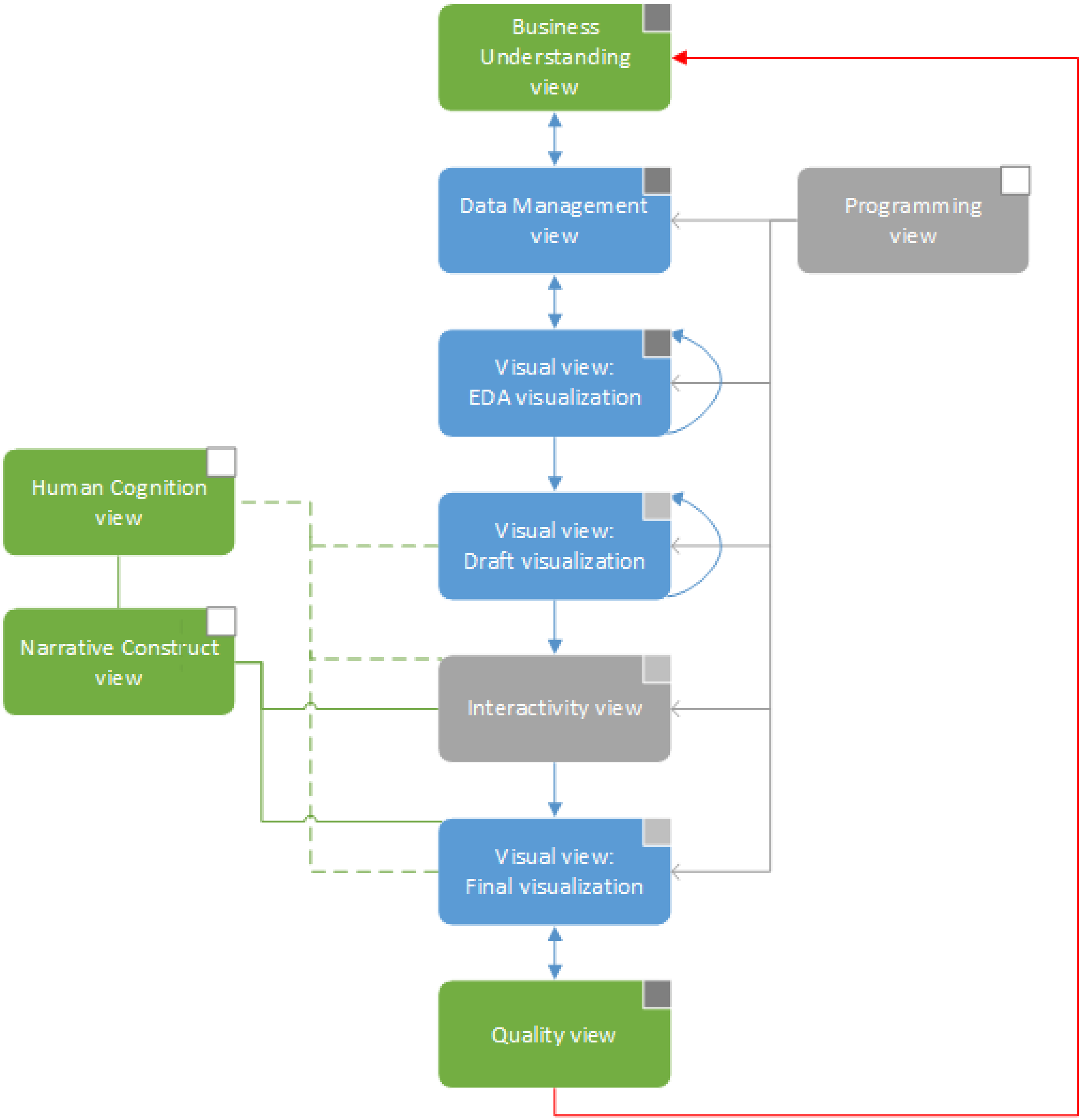



Fig. A.5.4.2: Version 3.1

In the Fig. A.5.4.3, the phases of work were added. Also, the arrows between programming view and other were changed.

Interactivity view is interlinked with Visual view on the diagram.

The lines between Programming view and other views were refined. It has the solid line to Data Management view because the data should be collected and processed with the help of the computer tool. The line is dashed to Visual View: EDA visualization because Programming view is not used during Prototyping phase. The line is solid to the next Visual view because it is need thinking about how the mockup will be programmed. Programming view has a bold line to Interactivity view and Visual view: Final visualization, because these views are not possible without Programming view, also they are in Instantiating phase.

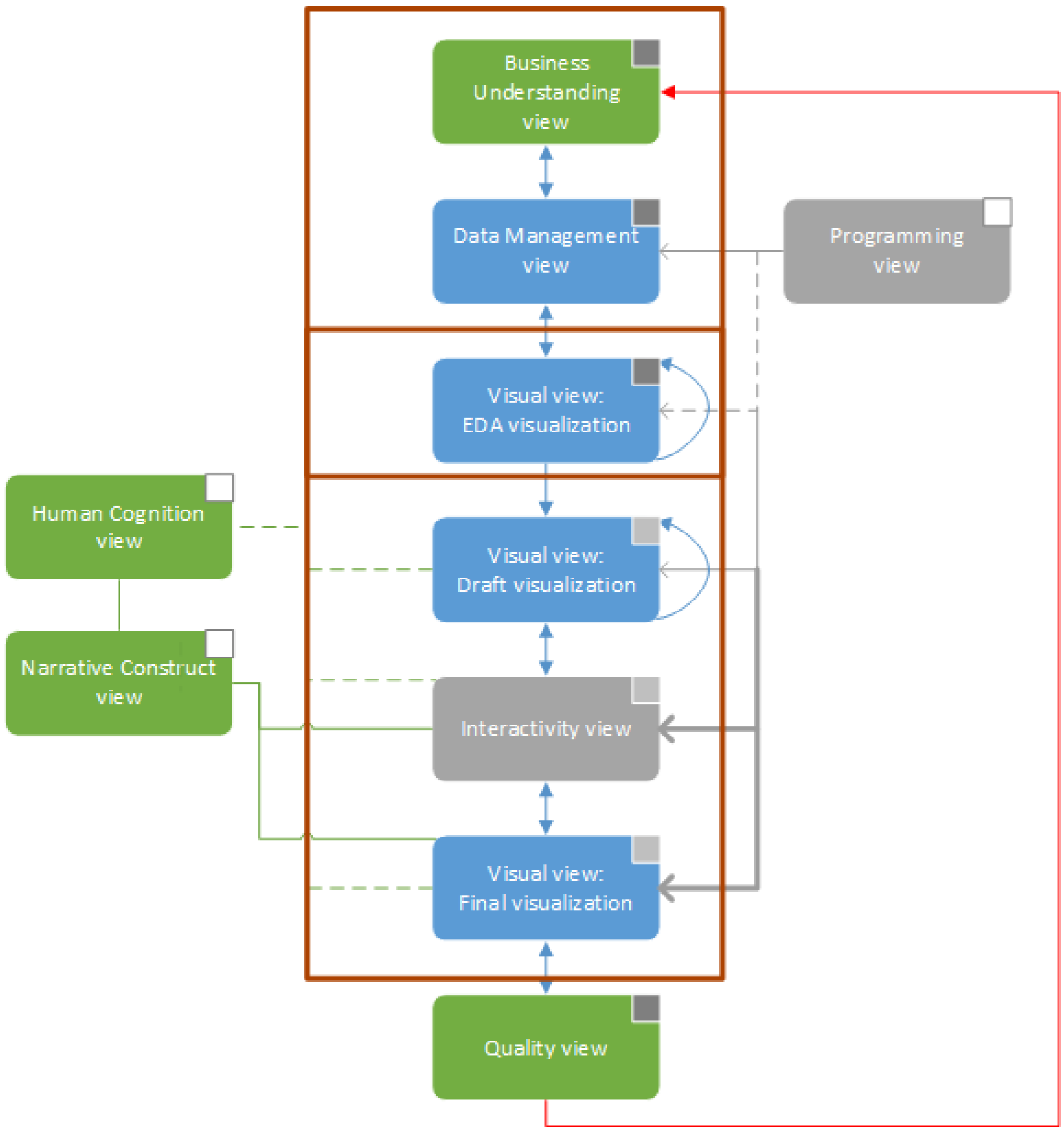



Fig. A.5.4.3: Version 3.2

The Fig. A.5.4.4 is the last version. The several views were renamed in this diagram, such as Visual view: EDA visualization - Visual view: Visual Mapping; Visual view: Draft Visualization - Visual view: Mockup Form; Visual view: Final Visualization - Visual view: Final computer-based visualization; Programming view – Instantiating view. The frames for phases of work were corrected on this diagram: Business Requirement phase – orange frame; Design phase – dark red frame; Instantiating phase – violet frame; Evaluating phase – it includes Visual view: Final computer-based visualization; Business Understanding view, Quality view and the feedback loop (it does not display with the frame due to the clutter on the diagram).

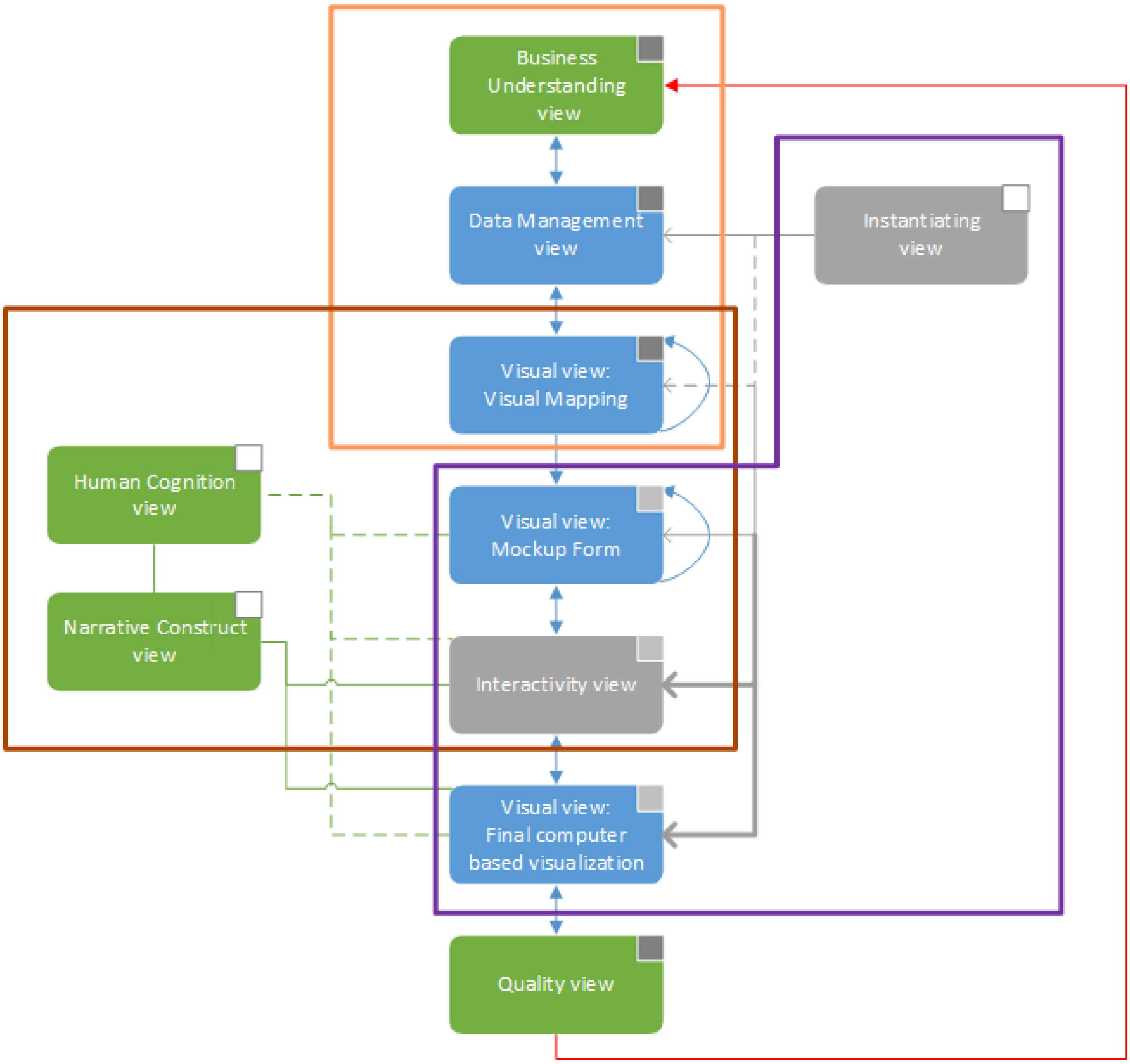



Fig. A.5.4.4: Version 3.3

# Appendix B. Questions for evaluation Business Understanding view

Project related questions (Marcus et al., 2003):

- Motivation – why is the visualization needed?
- What questions should be answered?
- Goal - What tasks should be done?
- What data should be analyzed?
- What data we have to do so?
- What are the objectives of the project? Outcomes? What do you want to achieve?

Domain related questions:

- What is the domain of the analyzed data?
- Is there a previous analysis done?
- Is any specific aspect the data designer should know?
- Social collaboration analytics
- UniConnect

Audience related questions (Knaflic, 2015; Marcus et al., 2003):

- Who is the future reader for the data visualization? The best practice is to narrow down the audience as much as possible.
- Why should the audience be interesting in this data at all?
- What is the knowledge of the audience about the topic of this data set?
- What should the audience know or do after reading the visualization?

Media related questions (Marcus et al., 2003):

- Where will the visualization be posted? (online or printed, internal or external use)
- Will be the personal presentation on the created visualization?

Limitation and specific requirement:

- Is there the time limitations?
- Is there any tool requirements?
- Is there any design requirements? (such as color, layout)

Data related questions (Yau, 2013):

- Who: The source of the data and its level of trust.
- How: The way the data was collected.
- What: The subject of the data.
- When: The time period of the collected data
- Where: The place there was collected.
- Why: The reason for collecting this data and sanity check for bias.