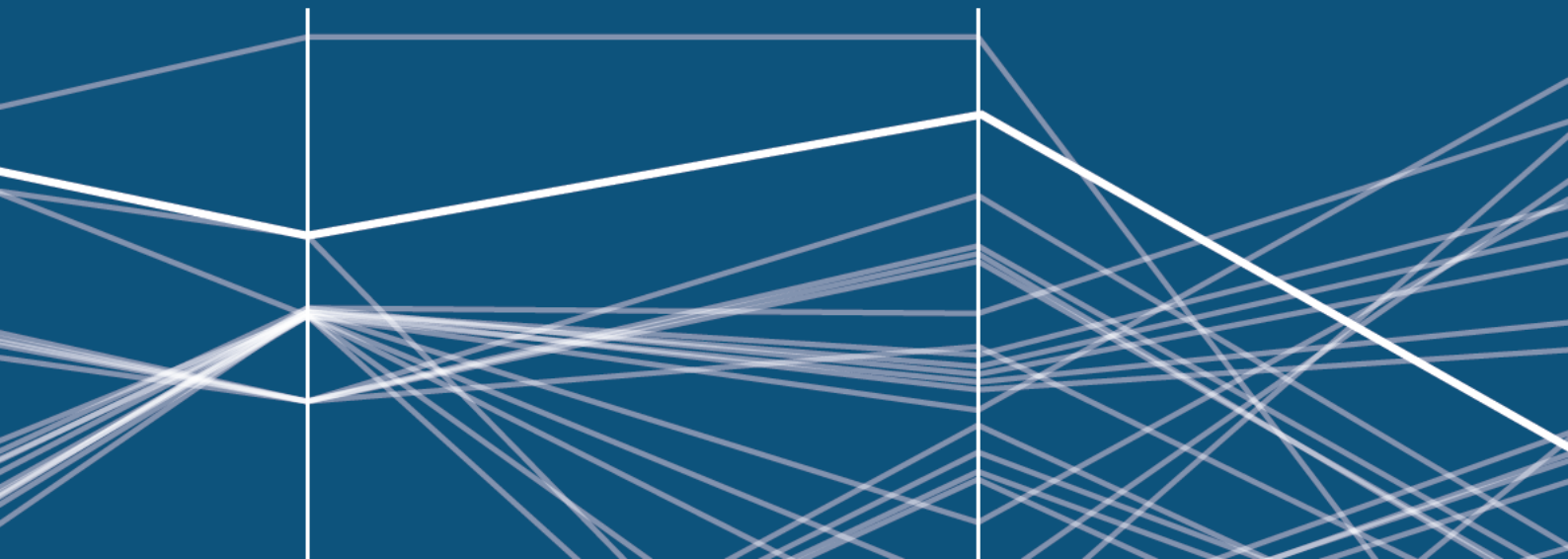


Visual fraud analysis

Supporting visual communication between data and analyst

Anne M. Postma

Graduation Report
MSc Science Education and Communication
Delft University of Technology & Fox-IT
April 2014



Visual fraud analysis

Supporting visual communication between data and analyst

Anne M. Postma

1327542

April 2014

Graduation Report

MSc Science Education and Communication

Contact

anne@annepostma.nl

06 497 24 323

Supervisors

Drs. Caroline Wehrmann, Science Education and Communication
Dr. Maarten Wijntjes, Industrial Design Engineering
Dr. Ir. Steven Flipse, Science Education and Communication

Company

Fox-IT BV.

Ir. Johannes de Vries

TU Delft Delft University of Technology
Science Education and Communication

FOX IT
FOR A MORE SECURE SOCIETY

Preface

Working with large datasets and using them to get a more informed view on the world around us, is something that I love to do both in my studies as in my spare time. So when I learned that Fox-IT was looking for a graduate student to explore the possibilities of visualizing large streams of online-banking data to assist fraud analysis, I did not hesitate to take the opportunity. Visualizing data that is both complex and has a social importance, is for me the (graduation) project of my dreams.

My studies have been diverse. It started with a bachelor in Industrial Design Engineering that combined both technology and creativity, where my interest in the discipline of Information Visualization emerged. The Science Communication department allowed me to combine Science Communication with a self-proposed Information Visualization curriculum. This included courses from the faculty of Computer Science, Industrial Design Engineering, and Technology, Policy and Management.

This project incorporated all of these different fields of science into one research project. It is therefore neither a pure Science Communication, nor a pure Design, nor a pure Information Visualization report; it includes facets from all of these schools. During the study it was therefore not uncommon to face my own information overload in the form of results from literature research, note taking, (user-test) results, and the urge to understand all aspects that the different fields have to offer. This made me also understand the broad scope of this subject.

I want to express my gratitude to Fox-IT for giving me the opportunity to work in a very inspiring company with many smart, creative, fun and technology-oriented people. Especially, I want to thank Johannes de Vries for giving me a place in the DetACT team, supporting me and taking the time to read my many reports. I also like to thank the DetACT analysts and the rest of the DetACT department, you inspired me greatly and I hope your valuable input is reflected in the proof of concept design.

This project would not have been possible without the support of Caroline Wehrmann who encouraged my interest in Information Visualization within Science Communication from the beginning of my masters until the very end. Thank you for always making time for me and giving me the confidence to follow my academic interests.

Thank you Maarten Wijntjes for showing me creative ways of visualizing and for your endless enthusiasm towards the project. You motivated my enthusiasm and made me curious for new and creative solutions concerning visualization and perception.

I also received generous support from Jack van Wijk from the Eindhoven University of Technology, whose feedback was very important and valuable during the designing of the proof of concept. Your great work for the Information Visualization community and your roots in Industrial Design Engineering inspires me to want to follow a similar path.

Special thanks to Steven Flipse, Maarten van der Sanden and the Science Communication Department of the Delft University of Technology for their support and inspiration on completing this master.

Finally, I want to thank my friends and family for supporting me throughout this period – especially Marc for the many hours of gathering thoughts and my parents for their endless motivation – and reminding me to foremost enjoy it. You made my final months as a student exceptional.

How to read this report

The report starts with an introduction on the need of visualization in chapter 1. Chapter 2 continues with the project description of this study. Chapter 3 explains the methods used in this study. In chapter 4 and 5 the results of the study are presented in the form of an Orientation and Iteration phase. Furthermore, I conclude on the final results and answer the research questions in chapter 6. Lastly, in chapter 7 I discuss the results' limitations, give recommendations for future work and sketch implications for the field of Science Communication.

This report is structured in a way that it is to be read from beginning to end to understand the full extent of the research. However, the reader could also read each chapter separately to only be informed on that part of the study. A reader interested in Design could be mostly interested in the designing of prototypes (chapter 5), while an Information Visualization specialist might want to see the proof of concept and its functionalities (chapter 6.1). A person mostly interested in the value of this project for Science Communication could go to the chapter which explains a model on how visualization can aid an analysts cognitive processes (chapter 6.2) or how I believe visualization could benefit from Science Communication and vice versa (chapter 7.3).

I hope you enjoy reading this thesis report!

Anne Postma, April 2014

Executive summary

This is a research graduation project for the master program Science Communication at TU Delft, conducted at the company Fox-IT as a research internship from September 2013 to April 2014. The goal of this report is to show how visual communication design can increase the efficiency to gather insights from large and complex datasets.

In the current information age the Internet allows us to gather and analyze more data than ever. We are therefore facing information overload and do not have the necessary means to cope with all that data. Analyzing data can easily lead to a persons' cognitive limits.

The case of fraud analysis is an extreme example where gathering more insights from large datasets is crucial as the relevant information within the dataset is very scarce (i.e. few frauds within a large dataset of non-fraudulent cases). To prevent frauds from happening in the future, fraud analysts need to design rules to indicate different types of fraud.

Using visualization as a means to more easily see patterns and relations in data has proven to help greatly in the past. The phrase, I see what you mean, shows that seeing and understanding can be synonymous. With visualization one could relieve the load of reading the data, which means that the analysts will have more cognitive power to think about patterns or relations within the dataset.

The goal of this project is therefore to present data analysts with the right visual elements in the form of a visual support tool for support fraud analysis. **The research question is:**

Which elements should an optimal visual support tool have to improve the efficiency of analysing large data streams for fraud analysis?

The research is conducted in the form of Design-Based Research where theory and practice are integrated. Analysis of practical problems is done in collaboration with online-banking fraud analysts from Fox-IT. Solutions to these problems are informed by theory from the fields of Semiotics, Technical Communication, Visual analytics, and Visual Perception. The solutions, i.e. prototypes, are designed iteratively throughout the study.

The results of this study include:

- A task analysis of fraud analysts,
- A list of relevant theories from the different research fields,
- A database of visualization and interaction techniques,
- Prototype designs,
- User feedback on the prototypes formulated in the form of design requirements,
- Reflection on the prototypes based on theory.

This report presents answers to the research questions in the form of a proof of concept of a visual support tool for the fraud analysts at Fox-IT, a visualization model for designing a visual support tool for analysts in general, and by answering sub research questions.

The proof of concept allows analysts to not only confirm or reject their hypotheses on frauds but it also gives the options for exploring alternative hypotheses. This is done by visually showing different levels within the data (i.e. overview on variables, specific relation between variables,

distribution of one variable) and by offering useful and fluent interactions to form rules to indicate frauds. The analyst thereby sacrifices specific details of individual data elements to get better insights into the whole dataset and can thereby create more sophisticated rules to indicate fraud. The author recommends implementing computer-aided suggestions to highlight interesting variables within the large dataset. Further research on the long term implications of this proof of concept have to be done to fully grasp the added value of the visual support tool on data analysis *and* to further improve on design details.

The visualization model encompasses the scope from low-level perceptual processes to high-level cognitive processes and explains these from different perspectives (i.e. the analysts, their interaction with the visual tool, design elements and the visual communication designer). The model offers guidelines on *how* other visual support tools could be designed and therefore offers procedural power. Additionally the model offers explanations *why* such a design could aid visual analysis and thereby shows descriptive power.

This study provides an example on how to integrate Science Communication theory (specifically Semiotics and Technical Communication) in a data visualization project. The added value of communication is to gain better insights into the analytical processes of the analysts and to use that insight to enhance the visual tool design. The author recommends to bring together different specialists to work on more integrated projects in order to connect the world of visualization and communication. Design-Based Research could provide the right methodology.

During this project the researcher worked with confidential information that cannot be publicly disclosed. Therefore some information used during the study was withheld from this report. The reports during this study were always checked by the Fox-IT supervisor and research methodology was confirmed by the supervisors from the Delft University of Technology.

Table of Contents

Preface **3**

Executive summary **5**

List of Figures **10**

List of Tables **12**

Glossary **13**

1 Introduction **15**

1.1 Information age 15

1.2 Fraud analysis 15

1.3 Online banking fraud 16

1.4 This research report 17

2 Project description **19**

2.1 Problem statement 20

2.2 Project goal 21

2.3 Research fields 22

2.4 Research questions 25

3 Methods **27**

3.1 Design Based Research 27

3.2 Mixed research methods 31

4 Orientation **35**

4.1 Fraud analysis in practice 36

4.2 Relevant theory 45

4.3 Known visualization and interaction techniques 53

5 Iteration **59**

5.1 Prototyping 60

5.2 User feedback 71

5.3 Reflection on theory and practice 80

6 Conclusion **85**

6.1 Proof of concept 86

6.2 Model for visualization 91

6.3 Answers to the research questions 95

7 Discussion 99
7.1 Visual support tool 99
7.2 Visualization model 101
7.3 Science Communication 104
7.4 What would I do differently? 107

References 108

Appendices 115

List of Figures 118

List of Tables 121

A. Detailed descriptions of the methodology 123

B. Theories on data analysis tasks 142

C. Detailed descriptions of theories 146

D. Data types and descriptive statistics 171

E. Visualization techniques 173

F. Visualization table 186

G. Screenshots third prototype 188

H. Recommendations for proof of concept 191

J. User test protocol first round 192

K. User test protocol second round 195

L. User test protocol third round 198

M. Influence of sample size on visuals of third prototype 201

N. Measurements of theories for reflection 202

O. Method of reflection and structuring 212

P. Reflection and structuring of theories 214

Q. Theory evaluation table after first prototypes 228

R. Theory evaluation table after second prototypes (1/2) 232

S. Theory evaluating table after second prototypes (2/2) 237

T. Theory evaluating after third prototype 241

List of Figures

Figure 1: Structure of this research project based on the context, problem statement, research goal, question, fields and methods used. 19

Figure 2: Anscombe's quartet. Four different datasets showing the same descriptive statistics on the left (mean, variance, correlation, linear regression) but show very different behaviour when visualized (on the right) 21

Figure 3: Two bar charts showing the same data still they tell a different story. The difference is that the left bar chart starts at zero while the right starts at 2000, making the difference seem more extreme. 21

Figure 4: Venn diagram to show all research field have overlap with each other, see Figure 13 for more details how they overlap see. 23

Figure 5: The research questions linked to research phases that are based on Design-Based Research characteristics. 25

Figure 6: Research phases based on Design-Based Research characteristics. Each phase is linked to the research questions it helped answer. 29

Figure 7: Methodologies (left) linked to the research phases(middle) and the research questions (right). The colour of the linking lines reflect each research phase so that the relevance and overview of links between methodology, research phase and research questions can be easily seen. Note that the last chapter (answers to the research questions) is based upon all research done before (represented with a vertical line to the previous chapters). 33

Figure 8: Relational dataset with each data element as a point in a n-dimensional space. 36

Figure 9: A general flowchart of fraud analysis based on Cognitive Task Analysis of DetACT analysts 39

Figure 10: Table of different outcomes of a rule based on the rule analysis of a session and the real description of that session. Correct outcomes are green and false outcomes are orange. 39

Figure 11: Diagram of abstractions of the task analysis into data questions 43

Figure 12: Cost-of-knowledge characteristic function, showing how visualization could raise the curve and provide access to more data elements in less time. 49

Figure 13: Overlap between fields, based on theories found during this study. Position of the theories is based on the journal of origin and what other fields it relates to. Areas with low density of theories shows lack of relevant theories found in literature. 51

Figure 14: The difference is scope for each round of prototyping. The blue triangle represents the scope of the analysts' workflow and the yellow triangle the visual encoding options that were considered. 60

Figure 15: First round of prototype design examples. The data used to generate these charts came from the DetACT analysts. 62

Figure 16: The four different visualization techniques that were most successful during the first round of user tests and literature reflection. Different variation examples are shown on the second or third row of images. 63

Figure 17: (next page) Second round of prototype design examples. Each bold heading references to an interaction category and each grey heading to a sub category mentioned in Table 11. To show the interaction technique on the static page, the author put different screenshots of the visualizations next or below one another. For example the top left image shows the visualization with firstly the linear scale and secondly the logarithmic scale. 64

Figure 18: Combining three different highlighting functions in the third prototype. The fraud type is highlighted in red, a selection of sessions is shown in yellow and the hovering of the mouse over a data element is shown in blue. For an introduction to the most distinctive functions and screens of the third prototype, see Figure 19. 68

Figure 19: Different states of the third prototype. The prototype is presented in a web browser and allows interaction to create the different states. Selecting needs the action of the user to click on or drag (i.e. brush) over specific variables or sessions (see image 1,2,4 and 6). Hovering means that the mouse only needs to be at the same location of that visual element (see the 5th image). Making something means the user needs to select a button so that the tool will show a specific visualization (see image 3). For more screenshots on other states see Appendix G on page 189... 69

Figure 20: Examples of optimization of diagram axes. Images made by the author, based on recommendations of Tufte (2001) 70

Figure 21: Start screen of visual support tool proof of concept 86

Figure 22: Change scales or sort variables in the visual support tool 87

Figure 23: Proof of concept where four variables are selected in the overview and two variables for the detailed visualizations. Three sessions are under inspection and a selection is made on two variables to form a rule to indicate these sessions. 88

Figure 24: Larger mouse area for hovering highlights 88

Figure 25: Adding a heat map or arranging the dots of the scatter plot differently makes it more apparent when a cluster of dots are overlapping. 89

Figure 26: Visualization model. On the horizontal axes are the different relevant elements for the visual support tool. The vertical axes represents the level (high/low) of perception/cognition. 93

List of Tables

Table 1: Summary of the research fields Technical Communication, Semiotics, Visual Analytics and Visual Perception and their relevance to the research. 24

Table 2: Overview of methods used during this research, a short summary of each and its relevance to the case. Table continues on next page. 31

Table 3: Fraud analysis workflow compared to literature 41

Table 4: Task actions reformulated into data questions linked to data types 43

Table 5: Relevant theories found on combining visual and textual signs for communicating 45

Table 6: Relevant theories found on communicating data visualizations with clarity 47

Table 7: Relevant theories found on how people perceive visuals 48

Table 8: Relevant theories found on how people reason with visualizations 49

Table 9: List of relevant theories found during literature research and which 52

Table 10: Visualization techniques categories used during this research 54

Table 11: Interaction technique categories used during this research 57

Table 12: Different prototypes rounds explained with their goal and means 61

Glossary

Fraud	The wrongful or criminal deception intended to result in financial or personal gain. In the case of online banking fraud, it is the criminals intention to gain monetary means by influencing online banking transactions in their favour.
Rule	A rule in this report refers to a rule to indicate fraud so it can be prevented in the future. Therefore, a rule in this context is a set of explicit values on variables in the data that can predict whether or not an online banking session is fraudulent.
Cognition	The mental action or process of acquiring knowledge and understanding through thought, experience, and the senses.
Iteration	The repetition of a process. In the case of this research it meant repeating the process of reflecting on requirements, designing prototypes, evaluating them with user tests and theory and repeat this cycle. The author did this three times.
Design-Based Research (DBR)	A systematic but flexible methodology aimed to improve educational practices. Researchers do this through iterative analysis, design, development, and implementation. These are based on collaboration among researchers and practitioners in real-world set- tings and lead to contextually-sensitive design principles and theories. (Wang & Hannafin, 2005)

1 Introduction

The goal of this report is to show how visual communication design can increase the ability to gather insights from large and complex datasets. In the current information age the internet allows us to gather and analyse more data than ever. Fraud analysts are taking advantage of the large amount of data people create (online) to prevent frauds from happening in the future. The case of fraud analysis shows an extreme example where gathering more insights from large datasets is crucial as the relevant information within the dataset is very little (i.e. few frauds within a large dataset of "normal" cases). Visualization can offer the opportunity to gather more insights about the data under inspection more easily.

1.1 Information age

After the digital revolution there was an increase of digital and online products. Now, in 2014, with more people with an internet connection than ever, a large amount of data is generated. Cisco imagines the internet to be more than a zettabyte (1.000.000.000.000.000.000.000 (10^{21}) bytes) by 2015 (The Guardian, 2011).

We are facing information overload and do not have the necessary means to cope with all that data. Some people consider this a new problem. However, Anaïs Saint-Jude (Lift Conference, 2012) explains that each era claims to have information overload. She continues by saying that this is part of the human condition. "Human beings are endlessly curious and will always be creating information, therefore will always need to create new and innovative more efficient ways to store, collect, summarize and search" (sic). In other words, information overload is the force that can spark innovation as new needs arise.

Innovations like the quantified self (i.e. self-monitoring of own data to improve one's productivity or gain knowledge about oneself) or personalized online education have filled the needs of dealing with the abundance of (online) data. The general public has become more aware of their data and how they can use it to their advantage. But they also learned about the negative aspects like (digital) privacy, identity theft and spamming that could make them and their data very attractive to criminals.

1.2 Fraud analysis

With cybercrime, criminals take advantage of modern telecommunication networks, like the internet and mobile phones. They use viruses, malware, Denial-of-Service attacks or other digital means with the goal of damaging a reputation (e.g. of a company), inflicting pain (e.g. to an individual) or are after unlawful gain (e.g. of money). The latter, unlawful gain, is defined as digital fraud.

Fraud investigation teams try to find and prevent fraud from happening. One could think of digital fraud as insurance fraud, online traffic fraud, identity theft or credit card fraud. Fraud teams analyse the digital tracks the criminals leave behind by means of data analysis.

A characteristic of fraud analysis is that there is often a large dataset with a lot of variables to research and only a small number of frauds that the analysts know of. The analysts will therefore look for small clues that could identify anomalies and normal and abnormal behaviour. Because the dataset is large and the clues are small, analysing the data can often reach the analysts' cognitive limits. Therefore, aiding this process of data analysis is becoming very important in maintaining a company's analytics.

Using visualization as a mean to see patterns and relations in data more easily has proven to help greatly in the past.

> *The beauty of data visualizations is that they exploit the quickness of the human perceptual processing system. They enable users to see the structure of large amounts of data in one view and to identify outliers, aberrations, trends, and affinities at a glance* (Mirel, 1998)

This research aims to design a proof of concept on how visualizations could support data analysts with fraud analysis to gather more or better insights. In this report the author researches the analysts' old workflow and designed and sought visualization techniques that it could support. Based on the research the author has designed a new workflow that includes *visual* data analysis.

1.3 Online banking fraud

The most visible and viable targets for digital fraud are large retail banks (Deloitte, 2013). In 2012 in the Netherlands, 93% of the Dutch population was already connected to the internet (Worldbank, sd) which means the Netherlands is in fourth place of the highest percentage of internet users. Here are some figures that show the size of online banking in the Netherlands: four out of five Dutch citizens (between the age of 16-75 year) perform online banking, 93% of the transactions happen online which add up to around three billion transactions with a total worth of 3200 billion euro (Nederlandse Vereniging van Banken, n.d.). With that the Netherlands has the most online banking penetration of Europe together with Finland (Statistics Netherlands, 2012). Therefore, the Dutch online banking system is very attractive to criminals.

To prevent this, banks have their fraud detection teams that keep up with current technologies and stop fraudulent transactions while they are happening. The bank's fraud detection teams try to find clues in the stream of *internet data* during a transaction to indicate whether or not it is a fraudulent one. However, the criminals keep up with technology as well and try to hide their digital footprint as much as possible. The challenge to find clues of fraudulent transactions becomes harder as the clues of fraud within the data become weaker.

For example, bank frauds could be phishing attacks, where criminals try to acquire information such as usernames and passwords by impersonating a bank. But the criminal could also use malware (malicious software) like Trojans or tamper with the internet communication between victim and bank. They could use so called mules to transfer the money or use *social engineering techniques* to convince a victim to transfer money.

Banks lose money on these and new types of fraudulent transactions every year. Currently the amount of money lost has been kept relatively low. According to the Dutch Banking Association (2013) 34.8 million euro was lost on online banking fraud in 2012 which is 0.001% of the yearly transaction turnover. The total amount of banking fraud (not only *online* banking) in 2013 decreased 60%

compared to 2012 to 33.3 million euro (Nederlandse Vereniging van Banken, 2014). Still the Dutch Banking Association does not see a decrease of fraud attempts which emphasizes the need to keep working on fraud detection and fraud analysis.

Fox-IT is an IT security company and expert in the field of information security, cybercrime, and digital forensics. Fox-IT aims to make technical and innovative solutions with regards to security to improve the security of critical IT networks and systems.

At Fox-IT, DetACT is one of the products that Fox-IT can offer banks to monitor bank transactions. DetACT is developed to increase productivity of bank fraud-teams by presenting the internet data in a user-friendly interface. The system can also help to stop attacks in real-time by alerting the bank. Furthermore, DetACT offers services to help the banks with specific questions and keep the system updated for new types of frauds with the expertise of data analysts.

The data analysis of DetACT analysts show similar signs of the more universal fraud analysis: a small number of frauds in a very large dataset with numerous amounts of variables. Therefore, this research uses the case of fraud analysis at Fox-IT to show the value visual data analysis could give to other fraud analysts. The difference between frauds DetACT handles and other types of frauds is that it concerns online banking. Insurance fraud-teams could therefore handle insurance claims and its corresponding data, while online banking fraud-teams handle online banking transactions and its data.

Because of the sensitive nature of online banking fraud, the author needed to work with classified information (e.g. workflow, dataset of the analysts). Because the added value of this report is valuable for fraud analysis in general, the research is not dependent on that classified information and it is therefore left out of the report. The author used the classified information for specific detailing and the confirming of hypotheses on how visualization could best support the analysts. The classified information was reviewed by the Fox-IT supervisor.

1.4 This research report

Firstly, the author expands on the project description of the research by formulating the problem statement, project goal, relevant research fields included in the research, and research questions. Secondly, the research method of Design-Based Research and corresponding mixed research methods are explained. Thirdly, the author describes the results of the orientation phase of the research by giving an inside view of fraud analysis in practice, defining relevant theory, and show relevant work on visualization and interaction techniques. Fourthly, the results of the iterative cycles during the prototyping of a visual support tool are presented. Fifthly, the conclusion is given in the form of an answer to the research questions, a proof of concept for the visual support tool, and a visualization model for applying visualization techniques in other cases. Finally, the author discusses the results and findings and suggests future work on this subject within the field of Science Communication.

Preface 3

Executive summary 5

List of Figures 10

List of Tables 12

Glossary 13

1 Introduction 15

2 Project description 19

2.1 Problem statement 20

2.2 Project goal 21

2.3 Research fields 21

2.4 Research questions 25

3 Methods 27

4 Orientation 35

5 Iteration 59

6 Conclusion 85

7 Discussion 99

References 108

Appendices 115

2 Project description

This chapter will give an overview of this research project based on its context, which was presented in the introduction, the research problem, goal, question, relevant research fields and the method used during research (see Figure 1). The problem of the current ways of fraud analysis is that it often reaches the analysts' cognitive limits (chapter 2.1). This research aims to design a proof of concept for a visual support tool for fraud analysts which aims to relieve that cognitive load (chapter 2.2). The analysts would then be able to extract weaker signals to prevent more complex fraudulent transactions from happening in the future. Knowledge from four research fields, namely Semiotics, Technical Communication, Visual Analytics, and Visual Perception, was used to achieve this goal (chapter 2.3). The main research question of this project is "Which elements *should an optimal visual support tool have to improve the efficiency of analysing large data streams for fraud analysis?"* (chapter 2.4). The methods the author used to answer this question can be found in chapter 3.

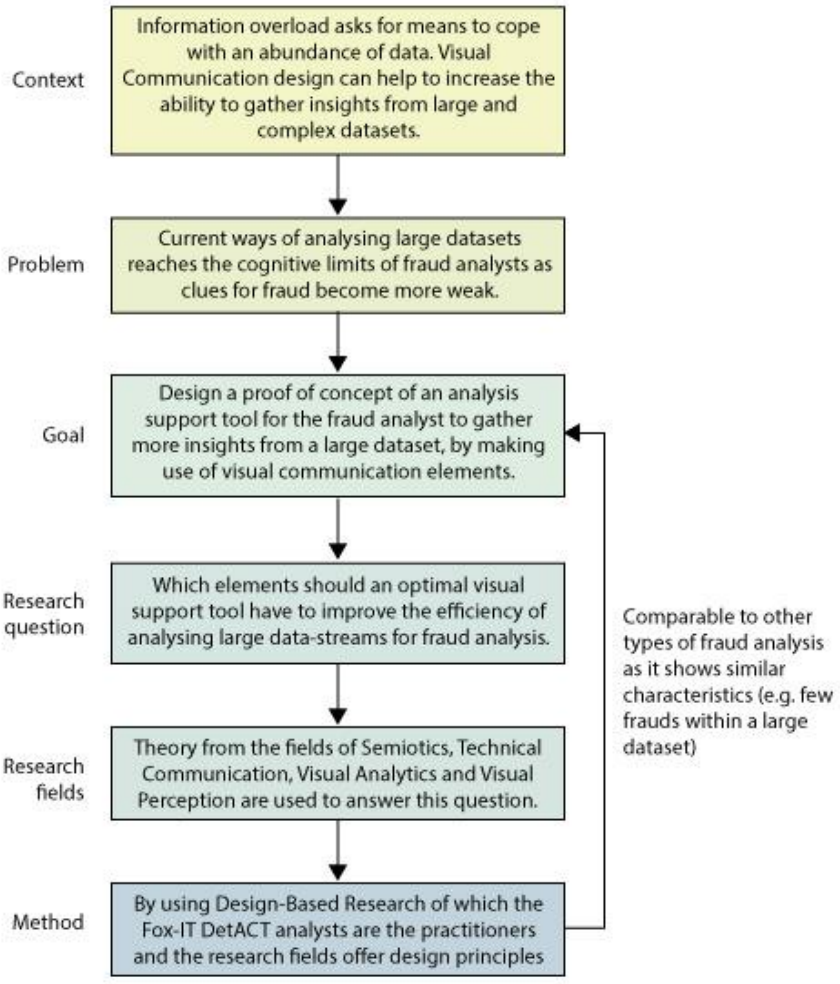


Figure 1: Structure of this research project based on the context, problem statement, research goal, question, fields and methods used.

2.1 Problem statement

The fraud analysts need to analyse a large dataset with many data elements for which they have put a system in place. In that system rules are placed by the analysts to annotate “strange” behaviour that could indicate fraud. When those rules are triggered a message is sent to the banks’ fraud detection team. To keep the system updated the analysts continuously search for signals that could indicate fraud.

The system is designed to analyse individual data elements within the dataset in the case of DetACT these are online banking sessions. The analysts’ main work consists of looking at these individual banking sessions and comparing them. However, inspecting individual sessions can only go so far. They often need to compare, scan, and conclude on multiple sessions at once which asks for a lot of cognitive load. Distracting an analyst during mid-analysis could make him lose his train of thought.

Additionally, a lot of their work is based on experience. An analyst could reason *“Have I seen this often or is it new to me, and if it is new it might be worth researching if other sessions show similar behaviour”*. When a new fraud type occurs it is hard to pinpoint what new rule could indicate it as a fraudulent session.

the clues to identify a fraud become smaller and more precise because the criminals get smarter in hiding their tracks,. Often the analysts is searching for weak signals: a signal that, if implemented as a rule, could cause a lot of the non-fraudulent sessions to be flagged as fraudulent. A weak signal is often not sufficient but a combination of weak signals - or the finding of a strong signal - in the dataset is. A lot of the work of the analyst consists therefore of finding the (weak) signals for *new* frauds.

The problem is that the current workflow of the fraud analysts often reaches the cognitive limits of the analysts when new complex frauds with weak signals occur.

2.2 Project goal

The fraud analysts at DetACT put great effort into using new means of analysis, like statistical analysis. Using statistical analysis provides more options to find weaker signals, however it also gives the analysts more data to analyse, i.e. more cognitive load. Relieving the load of reading the data will mean that the analysts will have more time to think about patterns or relations within the dataset.

Within the field of statistics, scholars make great use of visualizations to actually *see* if there is a pattern in a dataset. It is therefore not surprising to use visual data analysis for fraud analysis. A great example on how visualization can benefit statistics is Anscombe's quartet (Tufte, 2001) which shows that four datasets with similar descriptive statistics can show very different behaviour when visualizing them (see Figure 2).

Anscombe's Quartet: Raw data

	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
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
variance	10.0	4.1	10.0	4.1	10.0	4.1	10.0	4.1
correlation	0.816		0.816		0.816		0.816	
linear regression	y=3+0.5x		y=3+0.5x		y=3+0.5x		y=3+0.5x	

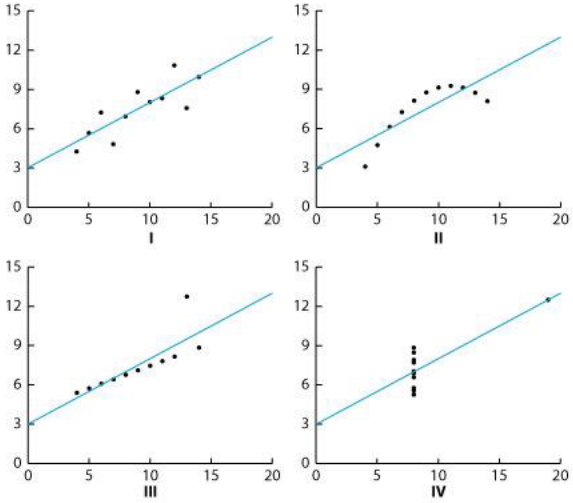


Figure 2: Anscombe's quartet. Four different datasets showing the same descriptive statistics on the left (mean, variance, correlation, linear regression) but show very different behaviour when visualized (on the right).

However, there are also examples that show how you can lie with statistics or visualization. If wrongly implemented, visualization could actually harm the analysis process (see Figure 3). The goal of this project is therefore presenting the analysts with the right visual elements to support fraud analysis. This will be done in the form of a visual support tool.

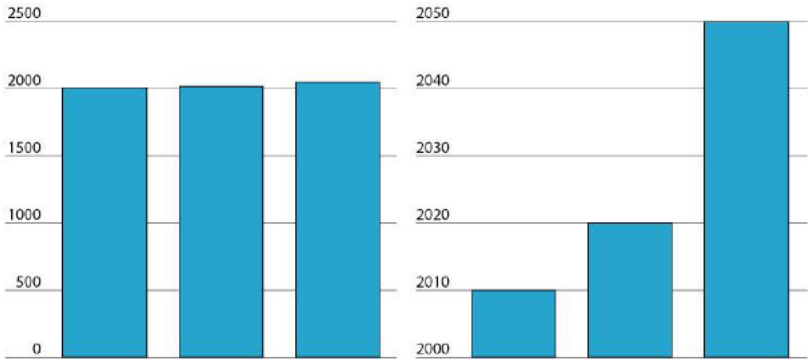


Figure 3: Two bar charts showing the same data still they tell a different story. The difference is that the left bar chart starts at zero while the right starts at 2000, making the difference seem more extreme.

2.3 Research fields

The field of Science Communication is relatively young, its main focus has always been to facilitate theory or models on communication between the scientific community and the general public. Even though Science Communicators have been communicating complex issues, communicating raw data is not usual within the field.

This problem can be viewed from different perspectives. First of all, it starts with the data. The field of statistics has allowed people to calculate summarizing numbers from large or small quantities of data and has built a lot of the ground work of the field of data visualization.

The visualization community within the field of Computer Science has brought visualization to another level. Implementing new visualization techniques with multiple interactive possibilities has allowed users to explore data more in-depth. These researchers have also been very interested in the scalability of data both within visualization techniques (i.e. what if we have 1 million instead of 100 data points) as its effect on algorithm design (e.g. aimed response time for rendering or interaction).

It is very important to take into account the user who needs to work with these visualizations. Interaction designers or the field of human-computer interaction have been researching how interaction between people and digital appliances influences behaviour. In this case, the interaction with the computer to gather more sophisticated information to prevent fraud could influence an analysts' analysis workflow greatly. While analysing, the analysts go through a cognitive process which involves processes like perception and attention. The phrase "I see what you mean" captures this perfectly in a sense that seeing and understanding can be synonymous.

2.3.1 Choice of relevant research fields

I would define Science communication as communication to make a difficult subject understandable for a broader audience. I believe data visualization strives towards the same result. Visualization scholars want to make complex datasets or (too) large datasets, that are not easy to comprehend, understandable for a wider audience than solely the data experts. Allowing the analysts, or even non-analysts, to find patterns and the right indicators to use for fraud analysis, is - to me - a way of *communicating* the complex raw data to a recipient. The way in which this is done should be carefully chosen and will rely on knowledge from various relevant research fields as well as to knowledge from Science Communication.

This research will focus on how to design the visual communication process between raw data and a user. Therefore the choice was made to include the fields of Semiotics and Technical Communication, both closely related to the field of Science Communication, Visual Analytics and Visual Perception.

The Semiotic tradition is the study of signs. It investigates how a message comes to have meaning for a person. It has a long research tradition and practitioners from a variety of fields, like mathematics, philosophy, logic. Within the world of semiotics, a sign could be anything from verbal to auditory to gestural and visual. In this project the specific theories on the visual signs are most relevant.

The *Technical Communication* field is aiming to make a receivers life easier and productive with the right information (Society for Technical Communication, n.d.). Technical Communication, like Science Communication, is still relatively young, broad, and has not taken on its definite form and therefore is still in the process of formulating its theoretical basis. According to the Society for

Technical Communication (n.d.) Technical Communication consists of communicating about technical or specialized topics by providing instructions about how to do something by using technology. Most common work of technical communication is document design and specifically user manuals but Technical Communicators can also be usability experts or interface designers.

The field of *Visual Analytics* originates from the field of Computer Science and specifically from information visualization. Visual Analytics is a very new research field and has been founded to contribute to *the science of analytical reasoning facilitated by interactive visual interfaces* (Thomas & Cook, 2005). It involves a lot of new techniques and case studies on how data visualization can benefit an analyst. Their research agenda from 2005 is still relevant today and the work on the influence of visualization on analytical reasoning continuous to be important.

Finally, the field of *Visual Perception* presents theories on how cognitive processes allow a user to perceive visual stimuli and how they extract shapes or colours that could help them analyse a visualization. According to Ware (2004) visual perception can have different perspectives, from the school of art, computer graphics, semiotics, psychology and neuroscience.

The overlap is large. The work on perceptual stimuli has much to do with how semioticians suggests signs are perceived by a recipient. The attention processes of visual perception interact with a person's analytical reasoning which is researched in the field of Visual Analytics. Both Technical Communication and Visual Analytics talk about usability and interface design. And all fields have different perspectives on how visual communication should or could benefit a user.

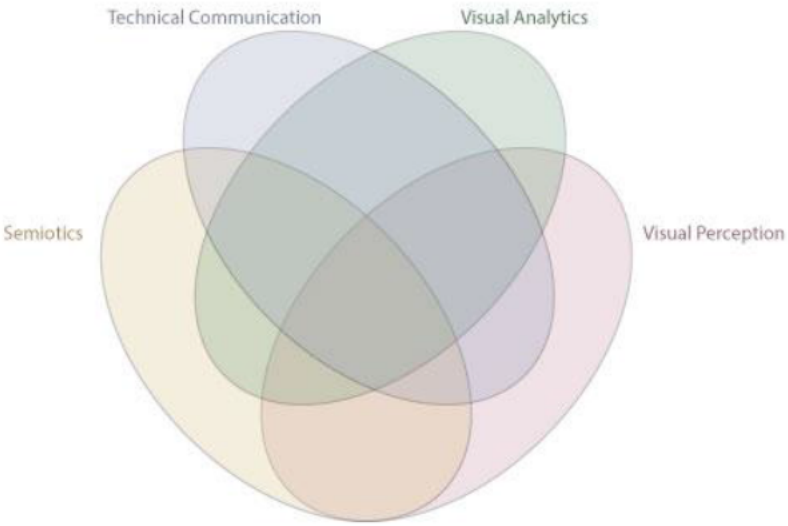


Figure 4: Venn diagram to show all research field have overlap with each other, see Figure 13 for more details how they overlap see.

Because this project is done within the department of Science Communication this research hopes to get a clearer view on how Science Communication knowledge, i.e. Technical Communication and Semiotics, complements this case and what design aspects need support from other fields.

The choice for the additional research fields was done to grasp the full extent of the case. Visual Analytics showed both the numerous cases done within the field of Computer Science but with a specific focus on analytics. Visual Perception was chosen as it is of a very different level than the other fields but with a lot of relevant theory on how a user perceives visuals on the bases of numerous visualization theories.

The field of Statistics was not within the scope of this project although the supervisor on behalf of the company is an expert on using statistics within fraud analysis. Also interface and interaction design is not specifically researched, although the author has a BSc in Industrial Design Engineering from the Delft University of Technology which included interaction and interface design. Additionally, one of the

supervisors is from the department Human Information Communication Design of the Industrial Design Engineering faculty is involved in research on perception and visual communication design. It needs to be noted that the focus of this study is not on designing the best interface (including subjects like graphical style) but on designing the best visual tool *elements* that could be implemented in an interface. Table 1 summarizes the different research fields included in this study and its relevance to this case according to the author.

Table 1: Summary of the research fields Technical Communication, Semiotics, Visual Analytics and Visual Perception and their relevance to the research.

Field	Summary	Relevancy to case	Relevant topics
Technical Communication (TC)	Any form of communication that is: About technical or specialized topics, or done by using technology, or providing instructions about how to do something. Technical communicators provide the right information in the right way to make the receivers life become easier and more productive. (Society for Technical Communication, n.d.)	The visual support tool aims to improve the efficiency of analysing large data streams. The technical communication field has a similar aim when designing their documents and research in this area could therefore benefit this project.	- Document design - Text supported by visuals - Usability
Semiotics (Sem)	Semiotics, or the study of signs, researches how a sign becomes to have meaning. A sign can be any form of sensory stimulus (e.g. auditory, visual, tactile) and is "something that stands to somebody for something in some respect or capacity" (Littlejohn, sd) (Moriarty, 1996)	In this project the plan is to change the predominantly textual data stream in a better comprehendible visual one. Semiotics compares sign systems (codes) and has theories on how someone makes sense of different sign systems.	- Classification of sign systems based on interpretation and rhetoric - Reasoning on the meaning of signs
Visual Analytics (VA)	Visual analytics, "The science of analytical reasoning facilitated by interactive visual interfaces" (Thomas & Cook, 2005), is a field that resides within computer science. However it includes a broader view (e.g. cognition, communication, interaction) on how interfaces can support analysis of large data specifically to improve homeland security.	As this projects aims to support analysts during their analysis process, Visual Analytics is a very relevant field. They have done similar work within security visualization and have done research on how analytical reasoning works within data analysis in general.	- Analytical reasoning - Discourse between data and analyst - Visualization as external memory
Visual Perception (VP)	Perceiving and thinking are intertwined and influence our reasoning processes (Arnheim, 1969). Visual perception includes both perspectives, and offers theories on how they work together as cognitive processes.	Analysts will need to be able to perceive the visualizations and to gather information from it. Visual perception can help guide with theories on e.g. information processes but could also provide more concrete guidelines (e.g. colour guidelines).	- Acts of perception (i.e. bottom-up and top-down) - Visual processing theories

2.4 Research questions

Analysing data by inspection of textual output is effective for current fraud analysis. But the increasingly big challenge of finding weak signals forces analysts to use other solutions that make better use of the cognitive power of the analysts. A solution is to help the analyst extract weaker signals from their dataset by allowing them to visualize the data and make use of statistical analysis. This research aims to find the right presentation of visuals in the form of a tool that best supports the analysts. Theory from the fields of Semiotics, Technical Communication, Visual Analytics and Visual Perception are used to provide existing principles on this subject. A visualization model is made to make the results available for similar cases but with different context. The model is based on the theoretical information and insights about their use in practice.

The following research question was leading for this project: **Which elements should an optimal visual support tool have to improve the efficiency of analysing large data streams for fraud analysis?**

Elements refers to interface design elements, this can be graphs or interface functionalities but also colour choices or instructions. This is deliberately chosen to be very broad to entail the whole scope of the visual support tool. The *Visual Support Tool* is a name the author gave the tool that should be able to support the analysts in their analysis process and contains the aforementioned *elements*. The *efficiency* of analysing refers to the time and steps it takes to come to and verify hypotheses. This is a qualitative research, so the author does not give quantitative results but compares feedback from the analysts on their current workflow and one which includes a visual support tool.

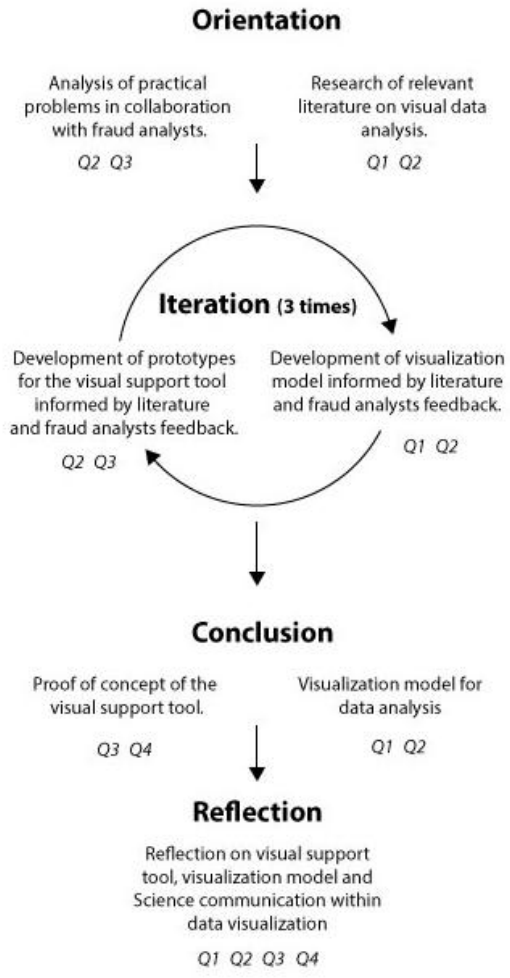


Figure 5: The research questions linked to research phases that are based on Design-Based Research characteristics.

The following sub questions were formulated to help answer the main research question:

Q1. How can theory from Semiotics, Technical Communication, Visual Analytics and Visual Perception be integrated into a visualization model?

Q2. How can the visualization model help in designing a visual support tool?

Q3. Which support tool design elements can support the fraud analysts?

Q4. To what extent does the designed support tool influence the data analysis workflow?

To answer these questions the author made use of Design-Based Research methodology (see chapter 3 for more details and Figure 5 for a schematic diagram). The research questions were answered based on research done throughout the project and were concluded and reflected upon in the final phases (see chapters 6 and 7).

Preface 3

Executive summary 5

List of Figures 10

List of Tables 12

Glossary 13

1 Introduction 15

2 Project description 19

3 Methods 27

3.1 Design Based Research 27

3.2 Mixed research methods 31

4 Orientation 35

5 Iteration 59

6 Conclusion 85

7 Discussion 99

References 108

Appendices 115

3 Methods

The overall methodology followed during this research is Design Based Research (chapter 3.1). It offers a way to combine theory and practice to develop practical products with a theoretical foundation. During the different phases, i.e. orientation, iteration, conclusion and reflection, mixed research methods were used; Cognitive Task Analysis, User testing, Literature research, Design methodology, Field observations and Storyboards (chapter 3.2 and Appendix A). The theoretical results of this research are incorporated into a visualization model (see chapter 6.2) to answer research questions 1 and 2 (how can theory be integrated and how can it help in designing a visual support tool?). The practical results are presented in the form of a proof of concept (see chapter 6.1) which helps to answer research question 3 and 4 (which design elements can support the analyst and to what extent does it influence the analysis workflow?).

3.1 Design Based Research

This research aims at combining theory from different fields for a case study on visual fraud analysis, Methodology that is uniquely aimed at integrating theory and practice is Design-Based research (DBR).

Design-Based Research was originally proposed as an extension of educational research methods. It embraced the fundamentally applied nature of educational research by combining theoretical research with design experimentation. Using interventions, or prototypes, to refine rich, theory-based innovations in realistic settings (Design-Based Research Collective, sd).

The Science Communication department of the Delft University of Technology acknowledges the potential added value of DBR for communication because education and communication research take place in similar applied contexts. DBR is therefore a specific research topic within the Science Communication department and more research on the subject could contribute to a more solid DBR methodology.

Wang and Hannafin (2005) define DBR as a systematic but flexible methodology aimed to improve educational practices. They do this through iterative analysis, design, development, and implementation. These are based on collaboration among researchers and practitioners in real-world set- tings and lead to contextually-sensitive design principles and theories.

This form of research resembles other types of research such as *research through design* (Zimmerman, et al., 2007) or *design inclusive research* (Horváth, 2008). DBR, research through design and design inclusive research all aim at researching practical problems and offer design solutions to change the current state into a preferred state. The author chose to use DBR methodology, versus research through design or design inclusive research, as this type of research is emerging in the field of science communication and adding to this work could contribute to a better understanding of DBR.

Design Based Research consists of four phases (Amiel & Reeves, 2008): First, the analysis of practical problems. Secondly, solutions informed by existing design principles are developed. Third, iterative steps to test and refine these solutions in practice are done. And finally the researcher reflects upon these solutions to produce (new) design principles and enhance the solution implementation. These phases ultimately lead to a strong connection between practice and theory.

The Design-Based Research Collective (2003) describe five characteristics of DBR:

1. Central goals of designing and developing theories are intertwined
2. Development and research take place though continuous cycles of design, enactment, analysis, and redesign
3. Research on designs must lead to sharable theories
4. Research must account for how design functions in authentic settings
5. Relies on methods that can document and connect processes of enactment to outcomes of interest.

Wang and Hannafin (2005) described similar but more concrete and practical characteristics; 1) pragmatic, 2) grounded in theory, 3) interactive iterative and flexible 4) integrative (mixed research methods) 5) contextual.

A challenge during DBR is that the researcher is both advocate and critic (The Design-Based Research Collective, 2003). It is therefore recommended to often triangulate with multiple sources and kinds of data. Because it is multifaceted it is important to ensure knowledge claims by having more methodological development to enhance rigor (The Design-Based Research Collective, 2003). Additionally, replication of complex interventions is largely impossible and causality can be difficult to decipher (The Design-Based Research Collective, 2003).

An example of a DBR project that relates to this research is Boticki et al. (2013). Boticki et al. designed a user-interface for collaboratively learning mathematics and Chinese in primary school. They used DBR as they did not want to hold variables constant in real-life settings. They could implement interventions so that they could observe and evaluate its result and optimize the design. They did three cycles of evaluation. There is no detailed explanation about methodologies used. They only describe the overarching concept of DBR to use theory to inform design decisions and use real scenarios. Another related DBR project is Starcic et al. (2013), who designed a tangible user interface for geometry teaching in a classroom. Their research was done over a two-year period and included three iterations. The first cycle consisted of collaborative design by using interventions for the practitioners. The precise design methodology or evaluation approaches of these projects are not described in great detail.

The methodological descriptions the author could take as guidance was that:

- Mixed integrated research methods should be used.
- Theory should be continually researched and tested in collaboration with practice and with experts on the subject as an iterative process with sufficient iterations.

Unfortunately, there was not enough precise DBR methodology descriptions found. The author therefore took the four phases of DBR (Amiel & Reeves, 2008) as a template and used design and evaluation methodology known from the field of industrial design engineering, Visual Analytics, and Information Visualization to give more precise guidance.

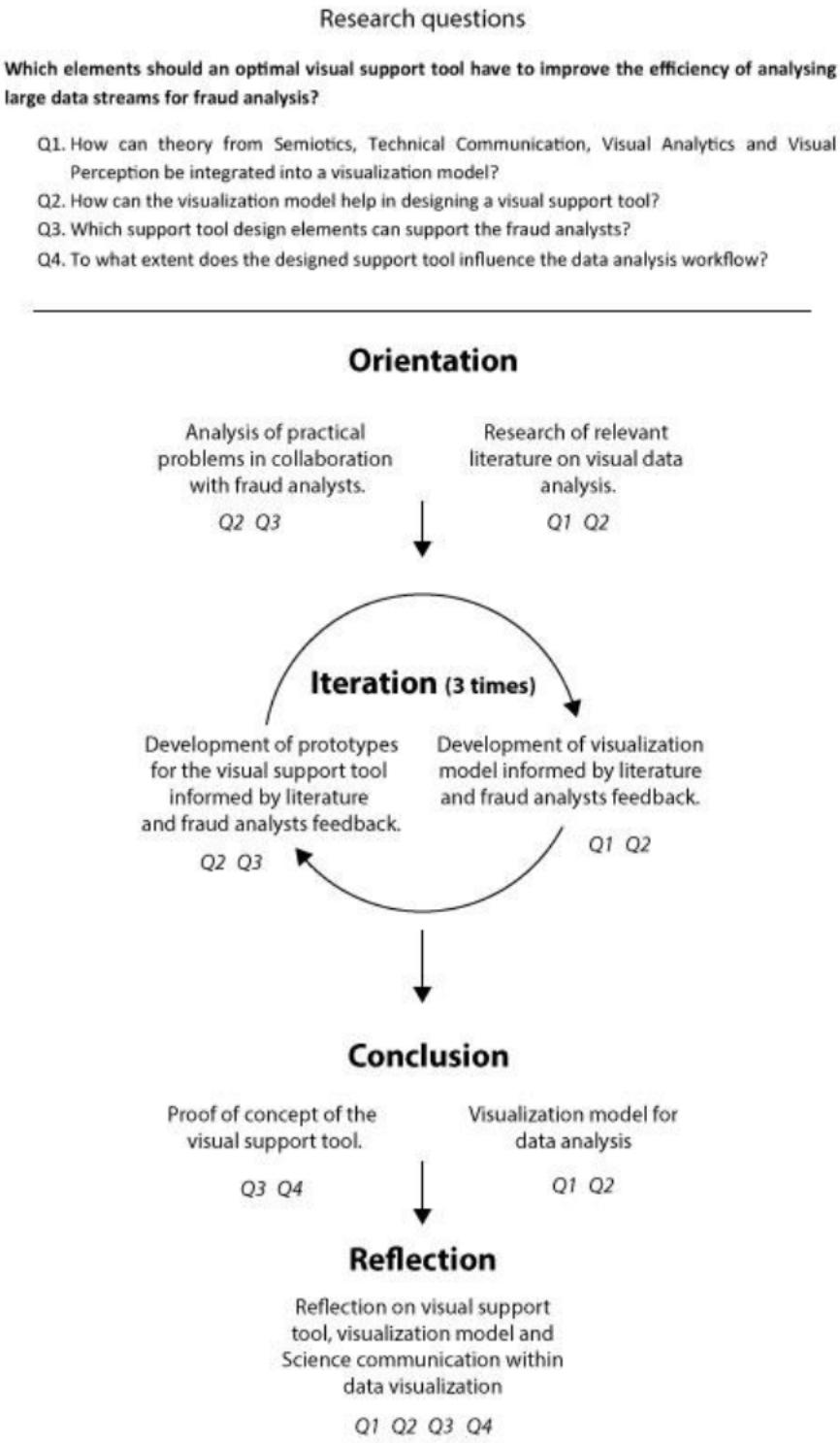


Figure 6: Research phases based on Design-Based Research characteristics. Each phase is linked to the research questions it helped answer.

Figure 6 shows how DBR was used in this research and how it relates to the sub-questions of this research. The phases, that are also leading in this report, are *Orientation* of the domain, *Iteration* of prototypes and finally the *Conclusion* of the Proof of Concept and Visualization model and its *Reflection*. During all phases, but in particular during the iteration phase, the author had frequent contact with the fraud analysts.

Each iteration included different prototype designs which were tested to gather more insights about an optimal design for a visual support tool. A lot of insights for future steps also occurred outside of the user tests. They were based on day-to-day field observations of analysts but also came from the design of the prototypes and the design-problems that occurred (e.g. realistic scenarios, visualization techniques limitations). These relevant insights were also documented and used during the design of the final proof of concept.

For the fraud analysts it was often their first experience with visual analysis. Additional subject-expert were consulted during the research. Professors from the graphics departments of both the Delft University of Technology and the Eindhoven University of Technology were asked to give feedback during different phases of the iterative prototyping rounds.

The author made the choice for three rounds at the beginning of the research. This number was a compromise as time did not allow for more rounds and the author considered two rounds insufficient to include all aspects of a visual support tool design (e.g. visualization, interaction) as each iteration and its insights do not provide final designs but just takes a step in the right direction.

The results of each iteration helped the author further to a better understanding of the design requirements but a final (or next) design still needs to be evaluated. The proof of concept shows the possibilities of visualization for fraud analysis in the best possible ways according to the evaluation of each round. Improving on this is still recommended by means of further prototyping, evaluation and refining.

"The most direct way to find out what someone knows is to ask them"

- Cooke, 1999 as cited in Clark et al., 2008 p. 5

3.2 Mixed research methods

Design Based Research is the overarching methodology under which all other methods used can be grouped. The following table provides an overview of the other methods used, for more details on the methods see Appendix A on page 123.

Table 2: Overview of methods used during this research, a short summary of each and its relevance to the case. Table continues on next page.

Method	Short description	Relevance to the case
Cognitive Task Analysis (CTA) including: - **Document analysis** - **Unstructured interviews** - **Observations** - **Conclude on tasks in four iterations with the analysts**	CTA tries to map the cognitive processes next to the tasks themselves. It includes a family of methods of which the author chose the relevant ones for the case. It took place in the work environment of the fraud analysts and also involved colleagues of the analysts. The observations were done in the end so that any discrepancies between the findings during document analysis, the unstructured interviews and actual observations could be identified. For more detail see Appendix A.1 on page 123.	Technical Communication often mentions the importance of task analysis for better document design. Cognitive mapping techniques (Siau & Tan, 2006) and formulation of task techniques (Albers & Mazur, 2003) were found but did not include specifics on how to get to such mappings or formulations. CTA has proven to be valuable in the field of Visual Analytics to gather information on tasks performed during visual data analysis.
Literature research	A systematic review of relevant literature for designing a visual support tool. The author chose peer-reviewed key journals of the four research fields and closely related peer-reviewed journals. Keywords and filters were used to find relevant papers with literature search engines. Additionally the author took important and relevant references from the relevant papers to further enhance the batch of literature. Finally the author also used books that are referred to in the papers and that are considered to be key publications in the different research fields. For more detail see Appendix A.2 on page 127.	Because Design-Based Research aims towards a strong connection between practice and theory. A systematic review hopes to include the most relevant theories for this case.
Design methodology including: - **Basic design cycle** - **Sketching** - **Prototyping** - **Design models** - **Mind mapping**	To design the different prototypes, different methods were used as each helped the author during different phases. The design cycle helped the author to keep track and to structure design thinking. Secondly, sketching helped to explore the parameters of each design problem. Prototyping is recommended both in DBR as design to facilitate quick learning cycles of the design problems and start working towards a solution. Design models from the field of information visualizations were particularly helpful. Finally, mind mapping was used by the author to structure thoughts on the project to help conclude on design decisions. For more detail see Appendix A.3 on page 131.	For Design-Based Research designing solutions, i.e. interventions, is at the core of the research process. Still, DBR does not provide concrete design methodologies. These design methodologies were chosen on the one hand intuitively by the author, because of the background in the field of Industrial Design Engineering, and on the other hand more specifically because of its relevance to the fuzzy aspects of getting insights into the design problem.

Method	Short description	Relevance to the case
User tests	The user tests were done to gather user feedback on the different prototypes. Some researchers could describe this as justification of the design encoding (Munzner, 2009) but it also helps to gather more information on the actual design problem. Evaluation techniques (i.e. User Experience, User Performance and Visual Data Analysis and Reasoning evaluation techniques) from the field of Visual Analytics were used as guidance. For more detail see Appendix A.4 on page 139.	The iterative nature of Design-Based Research helps to arrive at better solutions. During DBR close contact with the practitioners is the key element. To structure this contact, the author chose to do user tests guided by evaluation techniques from the field of Visual Analytics as they offered the most work on evaluating visual support tools.
Field observations	During the research the author worked at the same department as the fraud analysts. The author's role can be formulates as a non-participant overt observer. The author was non-participating, as fraud analysis requires great expert knowledge, and overt because the analysts were aware of the author's presence and know the goal of the research. Recording was done by note-taking as other means were not allowed because of security constraints. Most of the observing was done non obtrusive. For more detail see Appendix A.5 on page 140.	Design-Based Research emphasizes the need to work closely and collaboratively with the practitioners. To gather more domain knowledge than the user tests could give, field observations span a longer period and offer a more objective view on the day-to-day fraud analysis work.
Storyboard	To communicate the proof of concept in an understandable way, the author made use of storyboards. These can be either static or animated. Because the author chose to present the interaction of the tool next to its static design, the author made use of animated images. For more detail see Appendix A.6 on page 141.	It is mostly relevant for the author for communicating the proof of concept to practitioners and use for final presentation research. But because the introduction of an visual support tool needs to be done well, it is still mentioned here.

During the research the author used note-taking to document results of the user-tests and observations as recording was not possible because of the sensitive nature of the information. That is also why not all of the raw results can be found in this report or in its appendices. The results have been made available for Fox-IT.

These methods were used during different phases of the research. Figure 7 shows each method and during which phase it was used. The colours represent each phase (e.g. orientation is green) making it apparent that literature research was done often and during each phase while Design methodology is used during the iteration phase and specifically during prototyping.

Additionally the phases are linked to the research questions allow for deduction of the relevance of each methodology for the research. For example the Cognitive Task Analysis (CTA) helps to provide answers to research questions 2, 3 and 4 as the green line goes from CTA to the fraud analysis in practice chapter which leads to research question 2, 3 and 4 (see Figure 7). Note that the answers to the research questions (last chapter of the conclusion) are linked to all research results (represented in the vertical line which links it to the previous chapters).

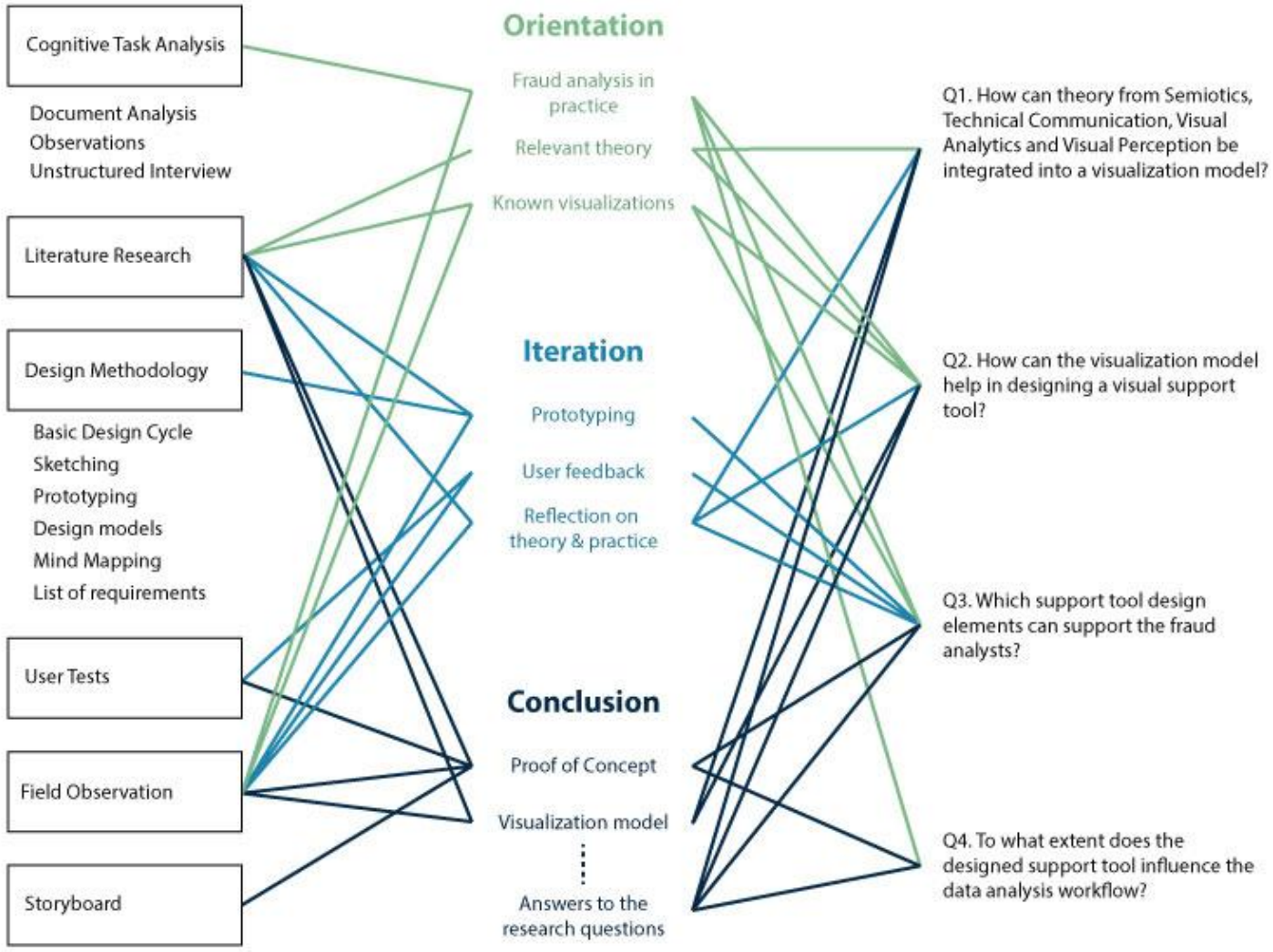


Figure 7: Methodologies (left) linked to the research phases(middle) and the research questions (right). The colour of the linking lines reflect each research phase so that the relevance and overview of links between methodology, research phase and research questions can be easily seen. Note that the last chapter (answers to the research questions) is based upon all research done before (represented with a vertical line to the previous chapters).

Preface 3

Executive summary 5

List of Figures 10

List of Tables 12

Glossary 13

1 Introduction 15

2 Project description 19

3 Methods 27

4 Orientation 35

4.1 Fraud analysis in practice 36

4.1.1 Dataset 36

4.1.2 Analysts 37

4.1.3 Analysis system 37

4.1.4 Workflow & tasks 38

4.1.5 Data questions 43

4.2 Relevant theory 45

4.2.1 Combining visual and textual signs to communicate data 45

4.2.2 Data visualizations communicated with clarity 47

4.2.3 Perception based on previous knowledge and the perceptual system 48

4.2.4 Reasoning with visualizations to form hypotheses 49

4.2.5 Overlap between research fields 50

4.2.6 List of theories 52

4.3 Known visualization and interaction techniques 53

4.3.1 Visualization techniques 53

4.3.2 Interaction techniques 56

5 Iteration 59

6 Conclusion 85

7 Discussion 99

References 108

Appendices 115

4 Orientation

The first phase of this research was orientation. It included the analysis of practical problems experienced by fraud analysts and the existing theoretical principles that could provide solutions for these problems. The analysis of practical problems helped to answer the research questions 2 and 3 (how can the model help in designing a visual support tool and which support tool design elements can support the fraud analysts?). The analysis of existing principles gave answers to research questions 1 and 2 (how can theory be integrated in a visualization model and how can the model help in designing a visual support tool?).

The analysis of the practical problems was done based on the case of DetACT fraud analysts at Fox-IT and its data, analysts, analysis system, workflow and data questions (chapter 4.1). The findings were compared to data analysis described in literature to know to what extent the Fox-IT findings are comparable to data analysis in general and what is recommended for a visual support tool design for this type of analysis. For the theoretical principles four research fields, Semiotics, Technical Communication, Visual Analytics and Visual Perception are included (chapter 4.2). Additionally the author formed a database of existing visualization and interaction techniques based on similar work and input from the visualization community (chapter 4.3). The knowledge gained during the orientation phase helped to design the interventions, i.e. prototypes, during the iteration phase (see chapter 5).

"To visualize data, you must understand what it is, what it represents in the real world, and in what context you should interpret it in."

- Nathan Yau, 2013 p. 42

4.1 Fraud analysis in practice

The description of fraud analysis in practice has different facets. First, the large *dataset* on which fraud analysis is done (chapter 4.1.1). Secondly, the *analyst* that performs the analysis to find and prevent fraud (chapter 4.1.2). Thirdly, the *analysis system* in which the analysts are able to perform their analysis (chapter 4.1.3). Fourthly, the analysis *workflow* with specific tasks that are characterizing fraud analysis (chapter 4.1.4). Finally, the author abstracted specific *data questions* of the analysts that can be used for designing the visual support tool (chapter 4.1.5). The results of this analysis helps to answer research question 2, 3 and 4 (how can a visualization model be of help in designing a support tool, which design elements support the analysts, and to what extent do they influence the analysis workflow?).

4.1.1 Dataset

Fraud analysts often work with datasets that grow over time. Network monitoring analysts continue to receive data on network activities, insurance fraud analysts receive insurance claims and online banking analysts receive online banking sessions. These elements can be described as data elements within a bigger dataset.

There are different types of datasets and an analyst in each domain could deal with different ones (e.g. relational, hierarchical, network). In this particular case the analysts worked with the most common, relational database which resembles a two-dimensional table. Each row represents a data element (in this case a banking session) and each column represents a variable (i.e. a property about the data element).

Shneiderman (1996), in his famous paper about task-by-data taxonomy, characterized datasets as being either 1-dimensional, 2-dimensional, 3-dimensional, temporal, multi-dimensional, tree or network. For more explanation on these types of data see Appendix D on page 171. In this case of online banking fraud, the analysts work with a multi-dimensional dataset. In a multi-dimensional data set data elements can be described as points in a *n*-dimensional space. *n* being the number of variables within the data (see Figure 8).

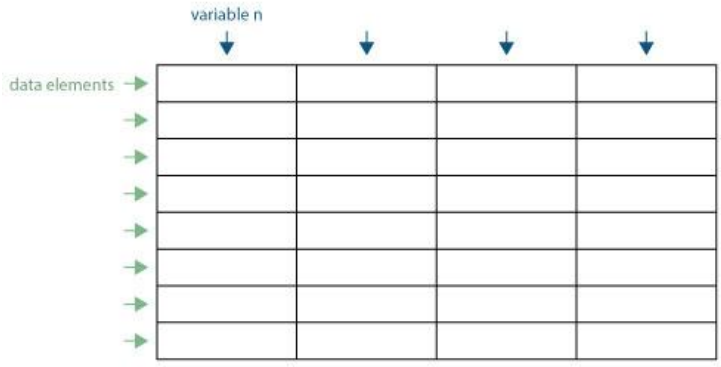


Figure 8: Relational dataset with each data element as a point in a n-dimensional space.

The different variables within the data can have different levels of measurements and characteristics. For more explanation on these types of variables see Appendix D on page 171. Based on the Cognitive Task Analysis (CTA) the different variables in this case can be described as:

- Categorical variables (nominal, ordinal or binary) (e.g. a web browser name)
- Numerical (ratio, discrete or continuous) (e.g. the time it took for a session to complete)

The design of the data table influences the way variables are formulated. Each variable that describes a data element needs to be captured in one data cell. This means that variables that change during a session need to be described differently. For example a change in the web browser used during a session cannot be described as one field with the mentioning of the different web browsers. Instead the analysts needs to think about the characteristics of this change (e.g. how often did it change, what was the first web browser, which web browser occurred most often) and make a variable of each characteristic.

This makes the already large dataset even larger to analyse. But it also allows more precise measurements and offers abilities to spot weak signals for fraud analysts based on more detailed variables.

4.1.2 Analysts

The analysts can be seen as detectives that try to prevent fraudulent sessions in the future. During analysis, the analysts are very concentrated on what they are doing. Disturbing them mid-analysis can make them lose their train of thought and they sometimes need to revisit what they were searching for (e.g. which session they were working with). They work individually on the classification of sessions and the making of new identification rules. Though they collaboratively decide, also with the bank, whether or not a new rule should be implemented.

A lot of expert knowledge is currently necessary to do fraud analysis. This can be knowledge on previous fraud types, how a normal session would look like or what the common values are for a specific variable (e.g. common web browsers). They therefore share a lot of information with each other and each analyst could have his specific expertise on a fraud or on the dataset.

In general the analysts can be working on real-time data but this is not the case when the analysts are analysing the dataset for identification rules. Real-time data is used for the real-time stopping of fraud which is done by the system or - less common - by people. During the identification of rules it is less pressing to react quickly. In general the analysts take their time to thoroughly inspect the data, although a fraud could be getting out of hand or a current rule could be generating too many false signals (i.e. flagging sessions incorrectly as frauds) which make a quick investigation process more necessary.

4.1.3 Analysis system

Each system could be storing different data. In the case of online banking fraud the system evaluates online banking sessions. Each session in the system now consists of data variables in a textual form with some indicator signs to structure them (e.g. bullet lists, headings).

As explained earlier, the system has rules in place to signal the bank when a transaction could be fraudulent. The analysis of these rules on each session is shown in each session report in the interface. A rule is mentioned by name when it was triggered (e.g. a web browser changes mid-session). The *names* of each feature of a session, rule or other relevant information is reflecting the use of that variable. So a rule which alerts on the changing of a web browser could be called “web browser change”. This makes it easier for the analysts to know the semantic meaning of each variable as remembering cryptic names would make fraud analysis even more difficult.

The defining of fraudulent sessions within the dataset can result from various sources of information, often reflecting the complexity of the fraud. These sessions are used for further investigation and the aim is to make identification rules for these sessions. Each fraud team could have a different focus on which sessions they would like to investigate. For example, when little fraud occurs a fraud team could be focused on finding out why certain sessions still show anomalous behaviour. When a new fraud type occurs those sessions are most relevant to investigate and the analysts will want to design a new rule for this fraud.

4.1.4 Workflow & tasks

To increase the understanding of what fraud analysts do, Cognitive Task Analysis was performed by the author (see Appendix A.1). The result was a more clear view of the analysts task. A general workflow of fraud analysis is presented in the flowchart in Figure 9. *[1]

The analysts could jump in at all items in the flowchart. The most common being either to start gathering information on fraudulent sessions and use that batch of sessions for analysis (top left), to check currently active rules and determine their efficiency and if needed to try and optimize, or make a new rule (top right), or to already have an hypothesis on an interesting variable for a rule (middle).

During the hypothesizing on fraud (second item top left) the analysis questions are relatively easy. For example, when did a value change over the past weeks or is specific data in the session report. This is because the analysts already know what fraud types have occurred in the past and what fraud types are active.

During in-depth research on the interesting variables it is very common for the analyst to have a lot of data in front of him to sensibly conclude from the multiple variables among different data elements.

[1] * A red asterisk means the author cannot disclose more detailed information on the subject.

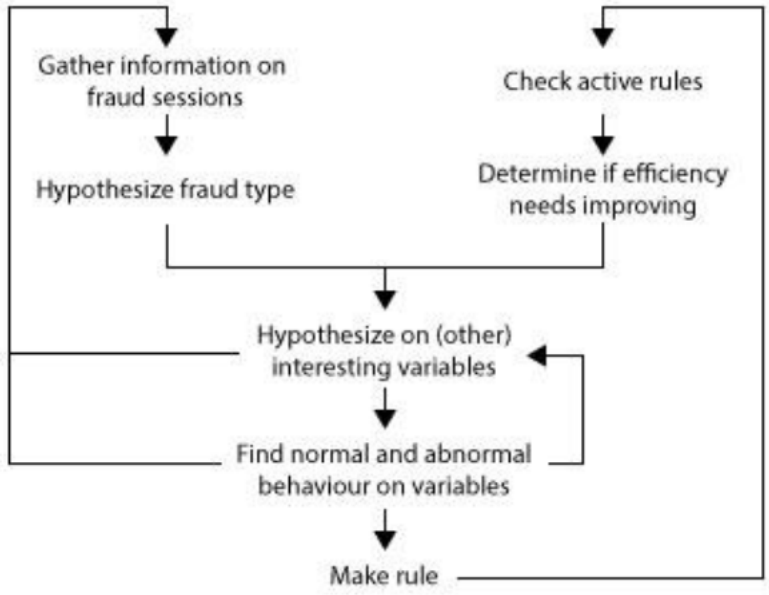


Figure 9: A general flowchart of fraud analysis based on Cognitive Task Analysis of DetACT analysts.

The main criterion for an efficient rule is the number of correctly or incorrectly flagged sessions. If a rule is formed to predict which sessions could be fraudulent session, the rule could display different results when it is activated (see Figure 10). The rule could correctly indicate sessions as fraud (*True positive*) and disregard sessions that are not fraudulent (*True negatives*). However, false results could also occur. When a rule incorrectly indicates a session to be a fraud while it is not it is called a *False positive* (Type I error). When the rule incorrectly indicates a session as non-fraudulent it is called a *False negative* (Type II error). A weak signal or rule is one that generates a lot of false positives and an ineffective one is one that generates a lot of false negatives.

The issue of having weaker signals, with a lot of false positive results, makes it more difficult to find abnormal behaviour and relevant variables. For this the analysts need more sophisticated data analysis techniques. For example descriptive statistics, see Appendix D on page 171, are relevant to find distributions and relations within the dataset and can help to identify outliers. The analysts can dig deeper into details of sessions over a larger sum of sessions to find normal and abnormal behaviour more easily. Using the conjunction of more variables could also specify frauds more accurately.

		Reality	
		The session is fraudulent	The session is not fraudulent
Analysis result of rules	The session is flagged as fraudulent	True Positive	False Positive
	The session is not flagged as fraudulent	False Negative	True Negative

Figure 10: Table of different outcomes of a rule based on the rule analysis of a session and the real description of that session. Correct outcomes are green and false outcomes are orange.

Another method for fraud analysis could consist of Data Mining. This term is often overused and misused as increasingly more people are forced to work with these techniques to handle their datasets. The author uses the definition as given by the New Oxford American Dictionary:

> "The practice of examining large pre-existing databases in order to generate new information"

This examination can be done in different ways and with different techniques. For example clustering, the grouping together of data elements that have similar values on specific variables, is a common approach. Another example is Anomaly Detection which is striving towards identifying data elements that do not show an expected pattern, i.e. do not show normal behaviour. Or regression, that tries to find the relationships among variables and aims at formulating a regression model that can describe the dataset.

These methods can help the analysts in an automated way to find areas within the dataset that could be interesting for rule-making. Fekete (2008) compared visualization to automated and confirmatory methods (statistics and data mining). He claims that there is no competition between them, he says:

> *When a model is known in advance or expected, using statistics is the right method. When a dataset becomes too large to be visualized directly, automating some analysis is required. When exploring a dataset in search of insights, information visualization should be used, possibly in conjunction with data mining techniques if the dataset is too large. Furthermore, combining data mining with visualization is the central issue of Visual Analytics.*

Because the analysts need to find more and better insights in a very large dataset, visualization with additional data mining techniques is recommended.

There are tools available that can help to visually explore datasets, though these are not used by the fraud analysts in this case. For example analysts can make use of software like Tableau public or JMP (SAS). This kind of software allows the user to view the data from different angles and gives great visual flexibility. However, they allow much more functionalities and options than the analyst will need. This is great for the general user, but can distract during fraud analysis.

Another important requirement of a visual support tool is that it should be able to fit into the current analysis system of the analysts, in this case the DetACT interface. Software like Tableau and JMP are not sufficiently flexible since they cannot be used as a part of an existing interface.

The visualization community often uses JavaScript libraries. JavaScript is a computer scripting language and is most commonly used within web browsers. Libraries are handy as they offer a lot of pre-programmed methods but still allow for a lot of freedom to code yourself. Common libraries are D3.js & Highcharts and also Tableau, mentioned earlier, has a JavaScript API. The downside, which is an upside for some designers, is that there is often a lot to be programmed and it does not always have a standard visualization available like in Tableau.

The important step is therefore to figure out which visualization and interaction techniques are necessary and to evaluate if the current tools could be used for this task or if (JavaScript) libraries are needed when it concerns a specific design.

Workflow compared to literature

Literature from the different fields have described visual data analysis tasks in different models. Relevant models for this research are (for a more detailed description of each task theory, see appendix B on page 142):

- The Sensemaking process (Pirolli & Card, 2005): the transformation of information by certain leverage points in different loops (i.e. foraging loop where one collects data, sensemaking loop where a mental model is formed and an overarching reality/policy loop)
- Categories of use (D'Amico & Kocka, 2005): people use visual data presentation for monitoring, inspecting, exploring, forecasting and communicating.
- Logical reasoning (Pike, et al., 2009): abduction, deduction and induction take place to construct knowledge
- Human Cognition model (Green, et al., 2008): both human and computer bring strengths to a mixed-initiative visualization tool

The goal of comparing these models to the current data analysis workflow is to take into account the recommendations of these models when designing a visual support tool. Additionally it helps to categorize the fraud analysis process at DetACT which helps focus the designing process and shows how it compares to other types of analyses.

Table 3: Fraud analysis workflow compared to literature

Fraud analysis workflow	Sensemaking process (see appendix B.1)	Categories of use (see appendix B.2)	Logical reasoning (see appendix B.3)	Human Cognition Model (HCM) (see appendix B.4)
Gather information on fraud sessions	Foraging loop	-	Abduction, induction, deduction	Discovery
Hypothesize fraud types	Sensemaking loop	Inspecting and exploring	Abduction	New knowledge creation
Check active rules	Foraging loop	Inspecting		Discovery
Determine efficiency of rules	Foraging loop	Inspecting	Deduction	Discovery
Hypothesize on (other) interesting variables	Foraging loop and sensemaking loop, reality/policy loop	Exploring	Induction	Discovery
Find normal and abnormal behaviour	Foraging loop, sensemaking loop, reality/policy loop	Exploring	Abduction, induction, deduction	Discovery
Alter or make new rule	-	Forecasting	Abduction, induction, deduction	New knowledge creation

The main iterative loops of the fraud analysis workflow is best described as exploring according to the categories of D'Amico and Kocka (2005). They emphasize with their categories of use that no single data presentation may be generically suitable for each purpose. Ribarsky et al (2009), from the Human Cognition Model, agree and advise to design multiple windows, each focusing on one aspect of the problem but all including 'balanced interaction' to not hamper the 'rhythm of reasoning'.

Pirolli and Card (2005) have identified critical points during the sensemaking process, which is crucial when the analysts need to form hypotheses based on the cost-limitations of humans. These cost-limitations is something the fraud analysts are facing as well. Ribarsky et al (2009) advise to include the HCM model during the foraging loop of the sensemaking process to partly solve this. They identify three classes of objects: data tasks, artefacts and the sensemaking stages. During specific data tasks, the computer could automatically show relevant artefacts and thereby support the according sensemaking stage(s).

Pirolli and Card (2005) continue by identifying other critical human aspects during the second loop, namely the analyst's span of attention, their ability to generate alternative hypotheses, and whether they are affected by confirmation bias (i.e. the tendency to interpret new evidence as confirmation of one's existing beliefs or theories). These aspects could be categorized as lack of creativity during the analysis process. That might be why Green et al (2008) criticized the sensemaking process as it does not include an explanation of spontaneous insight or the cognitive processes involved.

Abduction helps to describe creativity during an analyst's reasoning processes. Different scholars have specifically identified abduction, "educated guessing towards a hypothesis", as a close description on how visualization influences the reasoning process (Moriarty, 1996; Lundberg, 2000; Kolko, 2010; Hoffmann, 2007a). Additionally, abduction asks for more creativity of the observer on hypotheses generation. Supporting abductive reasoning could help the analyst to stay creative during hypotheses generation. Still, deduction and induction should follow to identify if a hypothesis is actually true or if it is likely true (Manning, 2002; Pike, et al., 2009; Liu, et al., 2010).

Workflow compared to other analyses

Fraud analysis of DetACT analysts does not include the monitoring of sessions. It is therefore different from real-time analysis of anomalies (e.g. with network monitoring). According to the data scientist at DetACT their analysis workflow is comparable to other types of fraud analysis. This becomes clear as the analysis models from literature describe the analysis process at DetACT very well, though not as specifically in tasks as in the flowchart of Figure 9.

The author thinks that the dataset influences the manner of *visually* analysing. For example analysing networks (e.g. of people) will use different visualization techniques (e.g. this person or group of people could be criminals) than analysing an n-dimensional table. However, both analyses can be described with the flowchart in Figure 9 (e.g. instead of the word "sessions", "people" or "relationships of people" could be used).

4.1.5 Data questions

To have a more clear description of tasks that a visual support tool should support the author described the tasks from the task analysis (see Figure 9) in more detail to find specific actions the analysts do. Some of the tasks occur in one's head (e.g. hypothesize or communicate) rather than in an analysis system. These tasks are important to take into account for context but do not provide specific information for specific actions or data questions.

The actions per different task have overlap in the sense that they support performing the task in a similar manner. Those actions were bundled together and reformulated in data questions and related data types (see Figure 11 and Table 4).*

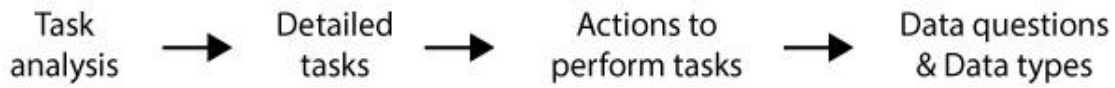


Figure 11: Diagram of abstractions of the task analysis into data questions

Table 4: Task actions reformulated into data questions linked to data types

Actions	Reformulated data questions	Data types
Show variable difference/similarity with norm on that variable over sessions	1. What variable in this session is different from other sessions 2. How does a specific session score in a multivariate dataset	- Nominal or ordinal - Quantitative
Show (cor)relations of two or more variables	3. How do two or more variables relate	- Nominal variables - Quantitative variables - Combination of nominal and quantitative variables
Show change of pattern over time	4. Does a certain variable change over time	- Nominal or ordinal - Quantitative
Show (cor)relations over time	5. Does a relationship of variables change over time	- Nominal variables - Quantitative variables - Combination of nominal and quantitative variables

During each iteration of prototypes the author iterated on these data questions to check if they are correct or if they needed to be edited (see next page for the change in data questions after each iteration). In general the operation abstraction is **to find abnormalities and relations or dependencies between variables *and* find the norm for certain fraud types.** When the data questions are answered by a visual support tool, that information could be used almost directly for rule-making.

First iteration on data questions:

- Can I find normal & abnormal behaviour in the multivariate dataset?
- How do two or more variables relate?
- Is a session within a cluster describing normal behaviour?

(Temporal data questions are left out as these were not relevant during fraud analysis because the analysts need current data)

Second iteration on data questions which are confirmed in the last iteration:

- Which variable(s) are interesting to inspect?
- Can I find normal & abnormal behaviour in the multivariate dataset?
- Where in the multivariate dataset can I find the fraudulent session(s)?
- Is a fraudulent session within a cluster describing normal behaviour?
- What combination of variables should a future rule be based on?

(The questions are more focussed on rule-making)

Each data question is also associated with the *levels of measurements* of the data. Bertin argues that there are as many types of questions as variables within information. But within each type there are three levels of reading: Elementary, intermediate or overall level (Bertin, 1983) that categorize all possible questions. An elementary level reading could be what is the temperature on the fourth of July, the intermediate level could be what was the average temperature in the whole of July and the overall level could be what is the trend in temperature over the years. According to Bertin, when more graphics are needed the efficiency of the graphic goes down.

Of course these abstractions of data questions were formulated based on the "current" non-visual workflow. A visual support tool could lead to a new workflow with other types of tasks. Still the data questions and data types will serve as input for visualizations as these extend the realm of investigating textual data. They are formulated abstractly than specific actions and could be used as input *from* the old workflow *for* the new workflow that could include the use of a visual support tool.

Data questions as practical problems

The practical problems that the analysts have during their fraud analysis (first phase of DBR) can be described as finding answers to the data questions posed earlier. Now they find answers to the data questions with limited visual aid, in which their current hypotheses are either confirmed or denied by numbers but do not offer room for new or nuanced hypotheses. This form of analysis also reaches the analysts cognitive limits. A visual support tool could offer better insights in the data and offer better answers to these data questions in a way that relieves cognitive strain and thereby solving the analysts' practical problems.

4.2 Relevant theory

A selection of journals, papers and books on the four research fields (Technical Communication, Semiotics, Visual Analytics and visual perception) was analysed. During the literature research, all found papers were scanned for theories relevant to the visualization of the fraud analysis case. These theories are used during the design of the visual support tool but also to build a more generic visualization model for visual detection. The results in this chapter help to answer research question 1 and 2 (how can theory be integrated in a model and how can that model help in designing a visual support tool?).

In this chapter the different theories are written in italic. For more explanation on literature research methodology see Appendix A.2 and for a more detailed description on theories see Appendix C.1 on page 149. Because the theories found have substantial overlap, each paragraph represents a theme of the literature research (chapter 4.2.1 to 4.2.4). The table at the beginning of each subject shows the field that contributed to relevant theories on the subject. Note that some theories originate from journals closely related to the four research fields, these theories are listed in the column they are most related too. When multiple fields describe the theory, the theory is mentioned in all corresponding columns. The overlap of theories from different research fields is presented in chapter 4.2.5 and the final selection of theories to include in this Design-Based Research is listed in chapter 4.2.6.

Note that not all of the four research fields (Technical Communication, Semiotics, Visual Analytics and visual perception) have concrete theories, let alone, relevant theories. The field of Technical communication even has an ongoing debate about the need for more theories. Still the goal was to have theories from all fields. The author therefore continued literature research on each field until there were no more relevant theories found.

4.2.1 Combining visual and textual signs to communicate data

Table 5: Relevant theories found on combining visual and textual signs for communicating

Semiotics	Technical Communication	Visual Analytics	Visual Perception
- Codes - Ten classes of sign	- Information designer - Visual Rhetoric		

In most communication fields the verbal language is predominant. Even Semiotics, the study of signs, is mostly concerned with the verbal sign than other types of signs. Moriarty (1996) believes the reason for this is that communication is primarily language based and has limitations as an explanatory scheme of nonverbal communication.

Within the field of semiotics two scholars have been most influential: Ferdinand de Saussure and Charles Sanders Peirce. In this research the works of Peirce will be of more interest as he had a more broad definition of a sign as "something that stands to somebody for something in some respect or capacity" (Moriarty, 1996). Hereby, Peirce opened up the discussion for other sign languages like auditory, tactile and visual signs. In comparison to the textual signs, these signs are more subject to interpretation which could demand more responsibility from the sender as a communication designer.

Signs are almost never presented by themselves. They have a context and are often presented with other signs or can be composed of multiple signs, whether or not from the same sign language. The connection between those signs are *codes*.

"Codes are a system of signs that are held together by paradigmatic and syntagmatic relations." – Sebeok, 2001

Recently, Technical communicators have been paying more attention to visual combined with textual document design (Johnson-Sheehan & Baehr, 2001; Harrison, 2003; Doumont, 2004; Amare & Manning, 2007; Desnoyers, 2011). Johnston (2012) believes this is caused by the icon-dominated digital media of the 20th century.

A lot of the cases on data visualization have been done in the field of computer science. William Hart-Davidson (2001) believes that the technical communicator should be getting a bigger role within Information Technology than he has now. He could be the so-called gardener of information to make it better accessible (e.g. by weeding out unnecessary data). Karen Schriver (2003) describes such a person as an *Information designers*, someone who "brings together words and images in ways that enable people to understand, take action, or make decisions". Information design becomes more important when the information gets larger or more complex.

Visual and verbal rhetoric have nevertheless been treated as two separate units (Manning & Amare, 2006), or even with visuals as being subordinate to text (Portewig, 2004). In contrary, Amare and Manning (2006) believe that document readability and graphics readability are both governed by the same (rhetorical) rules. They claim text and graphics are both visually configured information and that they both need to align with audience needs and authorial goals. The ability of a sender to use the right type of sign for his message has been called *visual rhetoric*:

> *Visual rhetoric has been defined as "the ability of the writer to achieve the purpose of a document through visual communication, at any level: for example, through the choice of a typeface (Courier, Helvetica), of graphic cues (bullets, lines, icons), of textual arrangement (lists, flowcharts, trees), of data displays (a pie chart, line graph), even of the color, shape, and size of the page" – Kostelnick, 1989 as cited in Markel, 1998 p. 47*

Amare & Manning (2006) based their work on the work of Semiotician Peirce who identified *ten classes of sign* types (including both text and graphics) within a single sign system. Amare & Manning used that sign system to develop principles, the most important being that both raw text and raw images must be diagrams to be informative. That is why an aerial photo is insufficient to show you the way without adding more diagrammatic information about the roads or where you currently are and where you want to go. Similarly a large text with directions without punctuation, a bulleted list or a map will also be failing to give you quick answers.

4.2.2 Data visualizations communicated with clarity

Table 6: Relevant theories found on communicating data visualizations with clarity

Semiotics	Technical Communication	Visual Analytics	Visual Perception
	- Visual Rhetoric Ethics	- Visual Rhetoric ethics - Data Ink Ratio - Lie Factor - Visual Channels of accuracy	- Colour display guidelines - Integral and separable visual dimensions

Because visualizations for data analysis are utilitarian in nature, the rhetoric of data visualization is to present data as objectively and with as much clarity as possible (Kostelnick, 2008). Kostelnick says "Perhaps in no other visual realm than data design is the notion of clarity more critical or more contested". He claims visual rhetoric approaches were initiated by the perception and cognitive science of data design but when they were seeking for an optimal design the concern for *visual rhetoric ethics* came up. The moral importance of clarity became most apparent with Huff and Geis's "How to Lie with Statistics" which demonstrated how a designer can manipulate charts for their own ends.

Visual Rhetoric Ethics "... concerns the time wasting inefficiency of images in conveying information that would have been conveyed more effectively ..." – Manning & Amare, 2006 p. 203

An advocate for clear visualizations is Tufte (2001), a prominent figure in contemporary data design, who developed the *Data Ink Ratio* and the *Lie Factor.* He also presented the following maxim on graphical excellence:

> *Graphical excellence is the well-designed presentation of interesting data- a matter of substance of statistics, and of design. Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency. Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space. Graphical excellence is nearly always multivariate. And graphical excellence requires telling the truth about the data.- Tufte, 2001 p. 51*

Perceptual theories help to promote visual clarity in design based on what people can perceive well and accurately (i.e. *colour display guidelines* (Ware, 2004; Ware, 2008; Yau, 2013), *integral and separable visual dimensions* (Ware, 2004), *visual channels of accuracy* (Mackinlay, 1986; Munzner, 2009)).

Ware (2004) explains that images are better for spatial structures and relations while the more detailed words, textual signs, are better for representing procedural information and abstract verbal concept. Some say a picture is worth a thousand words but Ware explains that the greatest advantage of words over graphical communication is that it is ubiquitous. "it is by far the most complete and widely shared system of symbols" (Doumont, 2004; Ware, 2004). So only when there is a clear advantage, visuals should be used.

Dual coding, using multimedia like both words and visuals, could lead to an even better understanding. Ware (2004) says that both visual and verbal representation should be actively constructed together and to make links between them. The links can be static (e.g. diagrams) or dynamic (animation or spoken). He continues by saying that images next to text enhance recall on the information.

4.2.3 Perception based on previous knowledge and the perceptual system

Table 7: Relevant theories found on how people perceive visuals

Semiotics	Technical Communication	Visual Analytics	Visual Perception
- Three functions of a graphic	- Visual Literacy	- External memory - Six ways visuals amplify cognition - Cost-of-knowledge characteristic function - Three functions of a graphic	- Gestalt principles - Feature Integration theory - Two waves of neural activity - External memory

When it comes to interpreting the meaning of visual signs Technical Communicators, Semiotic and Visual Analytics practitioners have identified the importance of a reader's *visual literacy* level. Visual literacy could be formed based on social conventions, like verbal literacy is, but is also led by the human capabilities of sight.

Visual literacy may be defined as a "basic system of grammar for learning, recognizing, making, and understanding visual messages that are negotiable by all audiences" – Brasseur, 1997 as cited in Mirel, 1998 p. 493

The field of visual perception has different theories on how visual elements are registered. Visual elements are perceived as a whole unitary object (*Gestalt principles*) or in parallel with separate features that can be identified as objects with focused attention (*Feature Integration theory*) (Wertheimer, n.d.; Treisman & Gelade, 1980; Ware, 2004). Both the cognitive and perceptual processes are important for understanding the meaning of a sign. Visual Perception scholars have described the *two waves of neural activity* that occur when our eye settles on a point of interest as an information-driven (top-down) wave and an attention-driven (bottom-up) wave.

Visualizations have therefore a great cognitive relation with its viewer, it can be described as a form of *external memory* that can amplify cognition. O'Regan (1992) claims the whole world can be seen as external memory and that "seeing" it involves active processes of probing the world. O'Regan explains this is also the reason why people have the subjective impression of seeing well despite the poor quality of the visual apparatus (e.g. blind spots, displacements or differences in resolutions). Munzner (2009) explains that external memory in visualization allows people to offload cognition to the perceptual system by using carefully designed images. Card, Mackinlay and Shneiderman (1999) have identified *six ways in which visuals amplify cognition*:

- Increasing memory and processing resources available
- Reducing search for information
- Enhancing the recognition of patterns
- Enabling perceptual inference operations
- Using perceptual attention mechanisms for monitoring
- Encoding info in a manipulable medium

The *Cost-of-knowledge characteristic function* (Card, et al., 1999) can help us to understand the cost structure of visualizations that aid foraging. On the x-axis the cost in time, and on the y-axis the nr of items accessible is plotted (see Figure 12). The point of a foraging visualizations is to raise this curve and be able to see more in less time. (Card, et al., 1999) (Pirolli & Card, 2005).

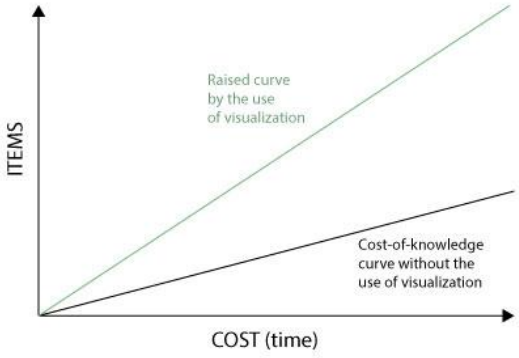


Figure 12: Cost-of-knowledge characteristic function, showing how visualization could raise the curve and provide access to more data elements in less time.

Bertin (1983) in his semiology of graphics distinguishes *three functions of a graphic* representation that can still hold today (i.e. recording, communicating, processing). According to Bertin (1983), information with three components or less should be able to be coded into a visualization that can perform all three functions while information with more than three components should be constructed according to the intended function as described above. Visual efficiency of a graphic is inversely proportional to the number of images necessary for the perception of the data to be done.

4.2.4 Reasoning with visualizations to form hypotheses

Table 8: Relevant theories found on how people reason with visualizations

Semiotics	Technical Communication	Visual Analytics	Visual Perception
- Diagrammatic reasoning - Diagrammatic reasoning with multiple diagrams	- Diagrammatic reasoning with multiple diagrams	- Analytical discourse - Information seeking mantra - Three classes of action - Internal model of visualization exploration process - Principles of effective animation - Situational awareness - Van Wijk visualization model - Reference model for visualization	- Visual Problem Solving Processes

When talking about information visualizations and how they allow to offload cognition in external memory, the use of diagrams is very common. Kim & Hahn (1997) say the diagram has very powerful impact on the perceptual and conceptual processes and is sometimes better than a textual representation. Peirce's *Diagrammatic reasoning*, and *diagrammatic reasoning with multiple diagrams* (Kim & Hahn, 1997), is helpful for explaining what reasoning process happens when people use diagrams to explain complex problems.

Diagrammatic reasoning: "To facilitate individual or social thinking processes in situations that are too complex to be coped with exclusively by internal cognitive means." (Hoffmann, 2007ap.5)

Diagrammatic reasoning seems very related to what the Visual Analytics scholars call *analytical discourse* between data and analyst. "Analytical work on data consists of mixed-initiative dialogues (between an analyst and the data) as mediated by visual representation of data on the interface." (Cai, 2007) (Thomas & Cook, 2005). There are different interaction theories that help to facilitate different dialogues (*information seeking mantra* (Shneiderman, 1996)*, three classes of action , internal model of visualization exploration process, principles of effective animation, situational awareness)*. The field of visual analytics distinguishes itself from other information science and visualization fields as it has a unique emphasis on the analytical discourse (Cai, 2007).

This again emphasizes the important relation between visualization and the user, mentioned by all research fields, and the responsibility of the visualization designer to optimally support the interplay between them. Both the information visualization field and the visual perception field provide models of how the user interacts with data visualizations (*Van Wijk visualization model, Visual Problem Solving Processes, Reference Model for Visualization*).

4.2.5 Overlap between research fields

Designing a visual support tool, allowing communication between fraud analyst and their data, could benefit from theories or principles of all perspectives of this issue. In Figure 13 The overlap between the different perspectives is been made visual.

Important to note here is that the theories found in the journals relating to the four specific research fields are placed in the circles of that field (e.g. Information Visualization subjects are placed in the Visual Analytics circle). The positioning of each theory is not only based on the journal of the paper but the author also inferred how the theory relates to other research fields. An example of this reasoning is: Analytic discourse is specifically mentioned in Visual Analytics, but the concept is very related to Technical Communication that wants to promote a well designed relation between a document and a reader.

The author became aware of the lack of true Semiotic and Technical Communication theories. The author tried to find more theories to balance the distribution of theories better (is visually shown in the lack of theories on the left and the abundance on the right) but unfortunately was not successful in finding additional theories during the study.

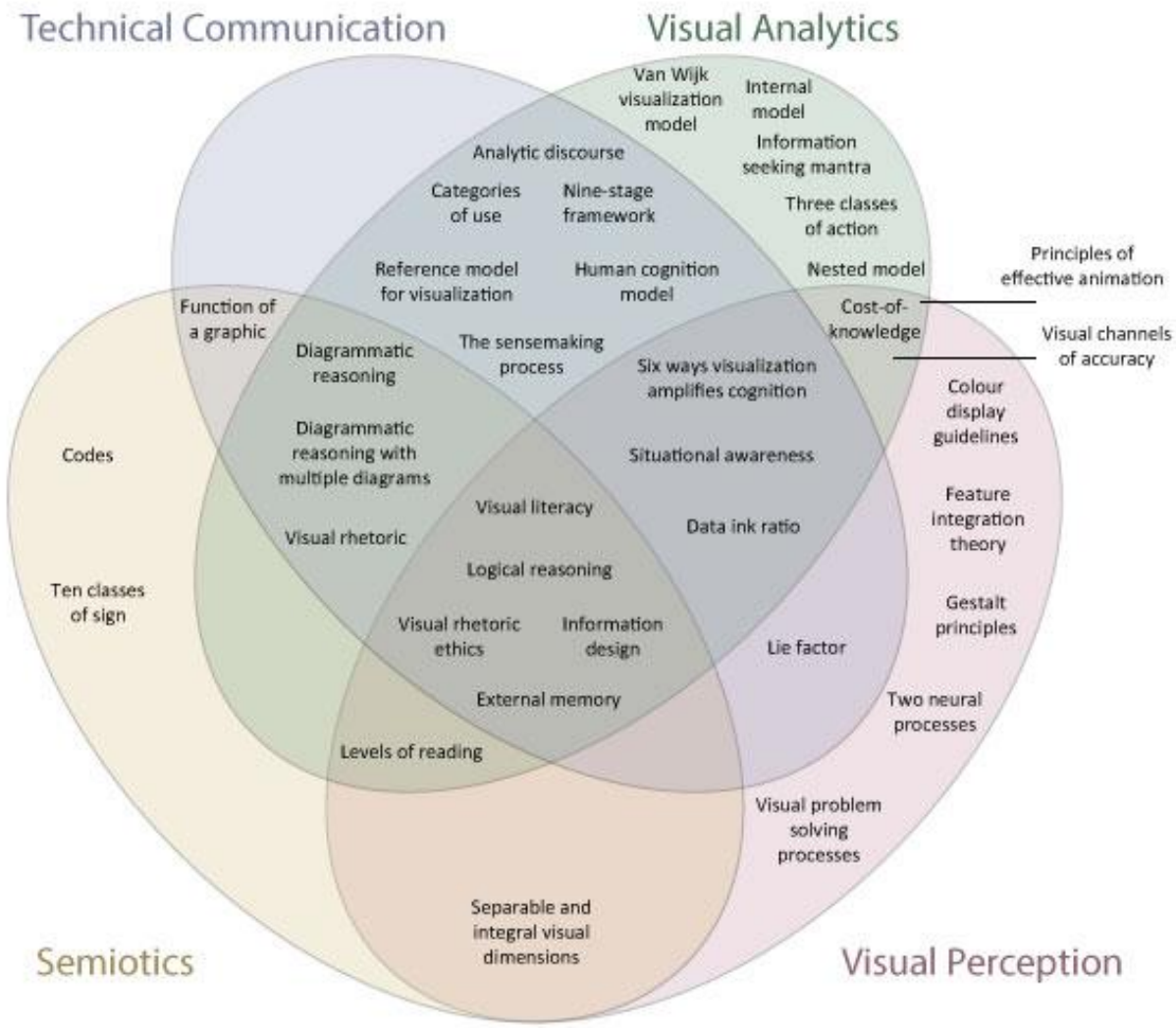


Figure 13: Overlap between fields, based on theories found during this study. Position of the theories is based on the journal of origin and what other fields it relates to. Areas with low density of theories shows lack of relevant theories found in literature.

4.2.6 List of theories

Based on the literature research (for methodology see Appendix A.2) a list of relevant theories or concepts was made. This list was iterated during the study as the literature research continued throughout the study. Still, most of the relevant theories were found before the iterative prototyping phase began. Table 9 lists the theories that were used for the continuation of the research to reflect upon prototype designs. A small summary of each theory and its relevance to the case and a more detailed description can be found in Appendix C on page 146.

Table 9: List of relevant theories found during literature research and which are used for reflection on the prototypes during this research

1. Analytical discourse
2. Categories of use.
3. Codes
4. Colour display guidelines
5. Cost-of-knowledge characteristic function
6. Data ink ratio
7. Diagrammatic reasoning
8. Diagrammatic reasoning with multiple diagrams
9. External memory
10. Feature integration theory
11. Function of a graphic
12. Gestalt principles
13. Human cognition model
14. Information design
15. Information seeking mantra.
16. Internal model of visualization exploration process
17. Levels of reading
18. Lie factor
19. Logical reasoning
20. Nested model for visualization design and validation
21. Nine-stage framework
22. Principles of effective animation
23. Reference model for visualization
24. Separable and integral visual dimensions
25. Situational awareness
26. Six ways visualization amplifies cognition
27. Ten classes of sign
28. The sensemaking process
29. Three classes of action
30. Two neural processes: bottom-up and top-down
31. Van Wijk visualization model
32. Visual channels of accuracy.
33. Visual literacy
34. Visual problem solving processes
35. Visual rhetoric
36. Visual rhetoric ethics

The list of theories was used for reflection on each round of prototypes to simultaneously improve the design and generate better insights about the use of the theories in practice. This will both improve the visual support design as well as to help to form a visualization model on how to apply visualization in other cases.

Note that the theories used during the reflection on analysis tasks (i.e. the sensemaking process, categories of use, logical reasoning and the human cognition model), on data questions (i.e. levels of reading) and design methods from the field of information visualization (i.e. nested model for visualization design and validation and the nine-stage framework) are also included as the author believed them to be relevant for use in a general visualization model and as reflection on the iteration cycles.

4.3 Known visualization and interaction techniques

Designing is a creative, intuitive process (van Boeijen, et al., 2013). The designer, in this case the author, uses own insights during the study to create designs and iterate on them. However, the author tried to structure this by making use of a visualization and interaction database. The results of this database helped to give answer to research questions 2 and 3 (how can a visualization model help in designing a visual support tool and which design elements can support the analysts?).

4.3.1 Visualization techniques

Making a complete visualization database and using this for visual tools is something visualization experts have been trying to do from the beginning of the field. Bertin (1983) does so with his semiology of graphics and computer scientists try to automate what visualization fits what goal so a designer would not be necessary anymore. Technical communication experts also try to make a taxonomy which they believe could increase visual literacy (Desnoyers, 2011).

They have not yet reached a consensus on what would be the best taxonomy. This database therefore does not claim to be complete, but it consists of most general visualization types to support the author on designing a proof of concept for the visual support tool.

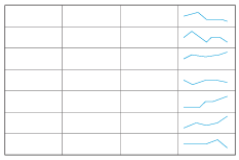
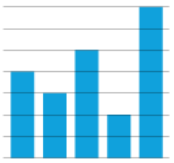
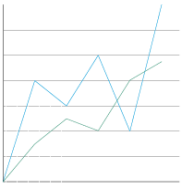
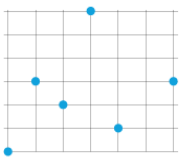
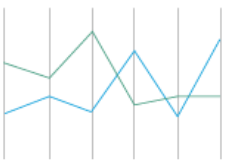
A common pitfall in visualization design is to include too few solutions (Sedlmair, et al., 2012), therefore the database was aimed to be big enough to get the most different possibilities.

The visualization database was filled with visualizations techniques based on:

- Visualizations found in similar work that describe a visual support tool
- Visualizations found on the internet (e.g. blogs, online newspapers)
- Inspiration gained during the time spent at Fox-IT
- Own experience

Only known visualization designs were taken into account, because there is a practical issue of the diversity of the data which prevents a design focus for a novel visualization technique. Additionally, “less is more” and simple is often better. Experts in the field of visualization have noted that a simple design often does the trick. Only when it does not, a visualization designer could have a reason to search for more complex and integrated visualizations. This concept can be supported by the idea that a visualization should match a person’s visual literacy level (see Chapter 4.2.3 or Appendix C.1 on page 149). Simple visualizations are often the ones they are most familiar with and are used often, and could therefore be easiest and best to read. The categories are presented with examples in Table 10, the full database can be found in Appendix E on page 173.

Table 10: Visualization techniques categories used during this research

Visualization technique category	Visualization example
Visual additions to tables. To have both the familiar data table but use some of the benefits of visual cues, a table could have little linecharts (sparklines), bars or coloured rows or cells. For a more detailed description see Appendix E.1 on page 174.	
Quantities The distribution of quantities can be visualized by encoding the number into a visual element. This can be length (either horizontal or vertical), like in the bar chart in the example on the right, area (e.g. pie chart) or colour (e.g. heatmap). For a more detailed description see Appendix E.2 on page 175.	
Change Change (e.g. over time) can be visualized by connecting each subsequent value with a line and plotting these values either horizontally or vertically on an axes. Area between the 0 value and the plotted value can also be filled with colour to resemble the quantity but because they are connected by a line mostly resemble change. For a more detailed description see Appendix E.3 on page 177.	
Points, dots or scatters A cartesian map is used to plot the values of each data elements on different variables. The differences within this category is the number of variables the visualization can encode (2, 3 or more in a matrix), how the values are plotted (with dots or shapes) and they offer means to plot valuable statistical information. For a more detailed description see Appendix E.4 on page 179.	
Relations, networks and graphs Relational or network data is most commonly visualized by giving each data element a node and each relationship between nodes a line or use open space to reflect distance. These types of node-link diagrams can have different lay-outs (e.g. circular, force-directed, in columns or linear and connected with arclines). The elements could also be put in a matrix which can show a colour to reflect a relation between the element in the row and column (e.g. co-occurences matrix). For a more detailed description see Appendix E.5 on page 180.	
Hierarchies Similar to the relational or network data is hierarchical data, the exception is that in hierarchical data the elements can have enumerous links to children nodes but only one to a parent node (except for the root node). The nodes can again be connected with lines and have different lay-outs. Hierarchical data can also be visualized by areas that have nested the areas of its children inside or adjecent to the parent area. Again these visualizations can have different lay-outs (circular, tree-like or square). For a more detailed description see Appendix E.6 on page 181.	
Patterns To show patterns between data elements on multiple variables, each value score on a variable can be plotted on an axes. Each value score of the same element should then be connected with a line to see each data elements scores on all variables. The visualizations differ with the amount of variables they can visually allow and if the data elements are plotted on the same axes (like the parallel coordinates in the example on the right) or that each data element has its own axes (e.g. glyph). For a more detailed description see Appendix E.7 on page 182.	

Spatial
Spatial data can be plotted on their location on a map. The variables of each data element, i.e. a location, can be plotted on top of the map (like in the example on the right) or by colouring or putting a pattern inside of the geographical area. The downside of the latter is that big areas could seem more important while they do not necesarilly need to be. Another option to solve this is by removing the map, using only symbols (e.g. circle), but keep them at their corresponding location (e.g. cartogram). For a more detailed description see Appendix E.8 on page 183.



Diagrams
A diagram is a very broad description for all visualizations that try to simplify data. The characteristic is often that it has less data elements encoded than the other visualizations techniques. For example it could show one measurement in relation to the average of a dataset (e.g. gauge), show relationships between actions (e.g. flowchart), show relations between items (e.g. venn diagram) or make a simplified illustration (e.g. infographic). For a more detailed description see Appendix E.9 on page 184.



Physical
3 dimensional data (e.g. about a geographic surface) can be visualized by rendering it in a 3-dimensional space (e.g. surface area, 3D area). For a more detailed description see Appendix E.10 on page 184.



Each visualization could also include visualizations by using different lay-outs (e.g. grouped visualizations, small multiples, matrix). It can also have attributes like number of colours or the previously mentioned theory on data-ink ratio.

There are numerous techniques to optimize a visualization, for example the “Multi-scale banking to 45 degrees” (Heer & Agrawala, 2006). This technique tries to optimize line charts so that it is visually more easy to see trends in the data. With banking to 45 degrees the aspect ratio (the proportional relationship between the x and y-axis) of a chart is computed so that the average absolute orientation (i.e. angle) of line segments in the chart is equal to 45 degrees. Heer and Agrawala defined a technique that combines banking to 45 degrees with spectral analysis to generate charts with different aspect ratios that could identify different trends.

Visualizations like bar charts are sometimes made in 3D either by design choice or to plot multiple charts after one another in a 3-dimensional space. In general 3-dimensional charts are frowned upon by the visualization community as they often conceal or distort visual elements which makes it less easy to read (e.g. multiple rows of bar charts in front of one another in 3D, hides bars that are placed in the background).

4.3.2 Interaction techniques

In an ideal situation every visual element is an active element within the visualization (Ware, 2004). Still the elements should not distract because of unexpected interaction or because the response time of the interaction is too slow. Both Ware (2004) and Ribarsky et al (2009) emphasize that the train of thought should not be interrupted. When the user needs to switch attention to the interface itself the effect can be hampering the thinking process.

The following categorization is based on different taxonomies of the visualization community and what was found in literature (Ware, 2004) (Gotz & Zhou, 2009) (Heer & Shneiderman, 2012) (Demšar, 2007) (Ferreira de Oliveira & Levkowitz, 2003) (Gleicher, et al., 2011) of which the Information seeking mantra (Shneiderman, 1996) also made the list of theories. Again, the taxonomy does not claim to be complete. However, it combines the knowledge of others into a new version of an interaction taxonomy as each taxonomy had a (slight) difference in focus which could all be relevant for this case.

"The goal of the design cycle is satisfying rather than optimizing: while there is usually not a best solution, there are many good and ok ones. The problem of a small consideration space is the higher probability of only considering ok solutions and missing a good one." – Sedlmair, et al., 2012.

Table 11: Interaction technique categories used during this research

Interaction technique category	Interaction techniques
Data manipulation To get new or more information and gather evidence together.	**Data selection** - Query the data - Filter the data **Generation of visualization** - Derive values and models - Sort - Filter
View interaction The main exploration and navigation tools for the visualization. The user can look for content, manipulate the graphic, adjust his browser and relate items to one another.	**Visual exploration** - Change visual encoding o Different scales o Change visualization technique - Add to the visual encoding o Juxtaposition o Superposition o Folding (Tominsky, et al., 2012) o Redundant coding (e.g. numerical labels) o Add summary statistics o Reference structures (e.g. grid lines) o Highlights (e.g. outlines) **Data exploration** - Change data o Drag and drop variables o Filtering o Select value or a range o Merge o Sort - Split Selecting data o Brushing - More data (change range) o Zooming ▪ Loop ▪ Simple zooming ▪ Table lens o Panning ▪ Scroll ▪ Panning - Specific data (details-on-demand) o Hover o Tooltip o Reference line o Add point o Circle around mouse for attention
Meta interaction These actions involve problem solving and insight documentation. The user can form and document hypotheses to evaluate his initial idea about the data.	**History** - Record actions o Delete o Edit o Redo o Undo o Revisit - Provenance of items **Visual Insight annotation** - Annotate - Bookmark **Knowledge insight annotation** - Annotate o Modify o Remove **Extract** - Share Guide

Preface 3

Executive summary 5

List of Figures 10

List of Tables 12

Glossary 13

1 Introduction 15

2 Project description 19

3 Methods 27

4 Orientation 35

5 Iteration 59

5.1 Prototyping 60

5.1.1 Focus of each prototype round 60

5.1.2 Prototype designs 61

5.2 User feedback 71

5.2.1 User test protocols 71

5.2.2 Design requirements based on user tests 73

5.3 Reflection on theory and practice 80

5.3.1 Reflection methods 80

5.3.2 Reflection insights 80

6 Conclusion 85

7 Discussion 99

References 108

Appendices 115

5 Iteration

To design the best possible proof of concept for visualization during the fraud analysis processes, iterative steps were taken according to DBR methodology. Each iteration is characterized by a phase of prototype design (chapter 5.1), user feedback on that design (chapter 5.2) and reflection on the design based on both theory and practice (chapter 5.3). Each phase contributes to answering research question 3 (which design elements can support the analysts?). The reflection on the design based on both theory and practice also helps to answer research question 1 and 2 (how can theory be integrated into a visualization model and how can the model help in designing a visual support tool?).

The following paragraphs each describe the process and results of the different phases of this iteration cycles.

The iteration rounds helped to design a visual support tool that both helps the fraud analyst during data analysis as it is grounded on theory from the four research fields (Semiotics, Technical Communication, Visual Analytics, and Visual Perception). The conclusions, based on these iteration rounds, led to a proof of concept for the visual support tool and the visualization model (see chapter 6).

5.1 Prototyping

When prototyping, the goal is to gain insights that can be used to make informed design decisions for the next prototypes and ultimately the proof of concept. During the designing of these prototypes the author made use of different methodologies (see chapter 3.2). The prototype results helped to answer research question 3 (which design elements can support the analysts).

5.1.1 Focus of each prototype round

There were three rounds of iterations done to capture the extent of the research problem. The prototype design solutions for the first iteration were based on the Cognitive Task analysis, the first literature recommendations and design problems that came up during the design process of the prototypes. Because the scope for the design was therefore still large, the author made use of paper prototypes to make the designs relatively quick. The goal was to find out which of those techniques had the greatest likelihood to work well with the data and for the analysts.

The prototypes for the next two iterations zoomed in more on specific design requirements and encoding options and discarded other options based on results from the previous round. This was done because of time constraint, there was no time to research the full extent, including interaction, of each visualization and because it helped to focus the design process and the design requirements.

On the other hand, each iteration was also zoomed out when it came to the scope of the prototypes within the workflow of an analyst (see also Figure 14). The first prototypes were static images that could benefit a small set of data questions. The second prototypes included interaction techniques and therefore extended the scope to more complex data questions but also a more useful product for the analyst. The third and last prototype included design choices on how the tool could influence the complete workflow of the analyst.

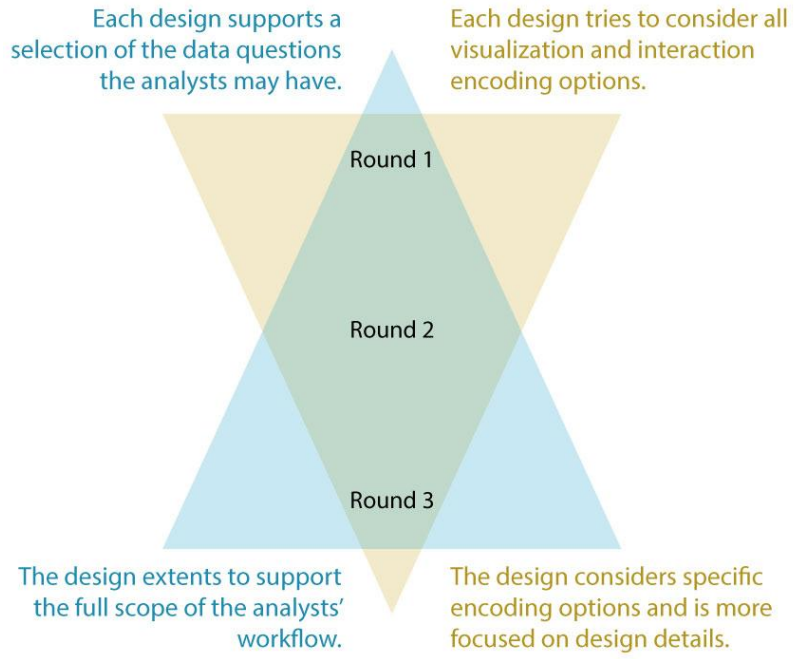


Figure 14: The difference is scope for each round of prototyping. The blue triangle represents the scope of the analysts' workflow and the yellow triangle the visual encoding options that were considered.

Note that for this research new visualization techniques were left out of the scope as it is often recommended to keep it simple. This will coincide with the concept of visual literacy as a simple visualization will be more likely to be understood than a complex one. Additionally, the author believed there was no need for an extremely complex new visualization technique except if it were a result from the research. Table 12 shows the difference in focus of each iteration.

Table 12: Different prototypes rounds explained with their goal and means

Iteration Round	Goal	Means
1: paper prototypes	The goal was to see how the analysts respond to a more visual form of the data. Foremost, to research which visualization design has the greatest likelihood to work well *with* the data and *for* the analysts.	- Building blocks for a visual session design with paper prototypes - Different visualization designs for in-depth data analysis with paper prototypes
2: interactive prototypes	The visualizations found best in iteration round one are used and made to include interaction. The goal was to see what combination of visualizations and interaction techniques could benefit the analysts most.	- Different visualization designs for in-depth data analysis - Different interaction techniques for in-depth data analysis - Different highlighting techniques for emphasizing specific values in visualization designs.
3: workflow prototypes	The final round consisted of the design results from the previous rounds and translating this in a workflow design. This prototype is tested to find its limitations and recommendations for further development and for a final proof of concept design	- A prototype including most of the use a visualization could have for in-depth data analysis

5.1.2 Prototype designs

Each prototype round offered better insights into the data questions, visualization and interaction techniques the visual support tool should be able to support. To link visualization easily to the data questions posed by the fraud analysts, the visualizations were put in a table. The visualizations that were not relevant because the data was not present within the dataset (e.g. spatial) were left out.

A real dataset from the analysts was used to design the visualizations to make it both familiar to the analyst and make them see how it can benefit them, but also because it is recommended by all methodologies and literature.

First prototype round

During the design process it became clear that some visualizations become unreadable when combined with the data and were therefore rejected as design possibility (see Appendix F on page 186 for more explanation why visualization techniques were excluded from the first prototype round). Still as many visualization techniques as possible were used as that allowed the author to test those techniques and confirm or reject them before the second round of prototypes. The author could infer which visualizations could be most successful (based on design guidelines, similar work and experience) but tried to keep an open mind. The author was also curious if there were any surprising results.

The author added interface elements from the fraud analysts' current interface to make the prototype designs more familiar to the analyst. The colours used were as neutral as possible and the author tried to stick with the minimalistic design principles of Tufte (e.g. data ink ratio and lie factor)

Based on the data questions (see Chapter 4.1.5), visualization database (see Chapter 4.3), recommendations from literature, and methodologies (see Chapter 4.2 and Appendix A.3) the author designed the first batch of paper prototypes (examples in Figure 15).

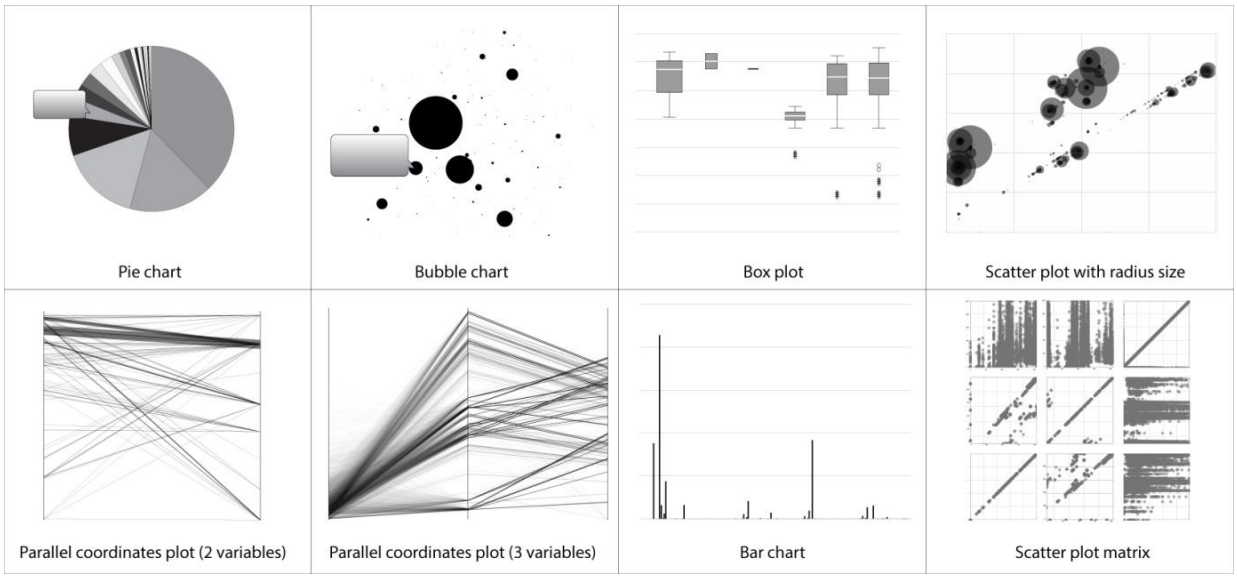


Figure 15: First round of prototype design examples. The data used to generate these charts came from the DetACT analysts.

The results of the user tests and reflection on the first prototypes were that the heat tables, bar charts, scatter plots and parallel coordinates plot were found to most likely succeed well with the data and the analysts (see Figure 16). These were therefore used as visualization techniques for the second round of prototypes. These visualization techniques still showed design possibilities (e.g. a bar chart could also be made into a grouped, stacked or small multiples bar chart) (see the second and third row in Figure 16).

The heat table colours rows based on a value (e.g. occurrences), it creates a more saturated area the higher the value of that area. The bar charts encode a value by the length of a bar. The scatter plots plot data elements in a two-dimensional chart. The scatter plot with heat map calculates the amount of data elements in a specific area and colours that area more saturated when there are more data elements. The parallel coordinates plot plots each dimension on a vertical axes next to one another. A data elements values on those variables are connected with a line. The thickness of a line could have additional encoding (e.g. occurrences).

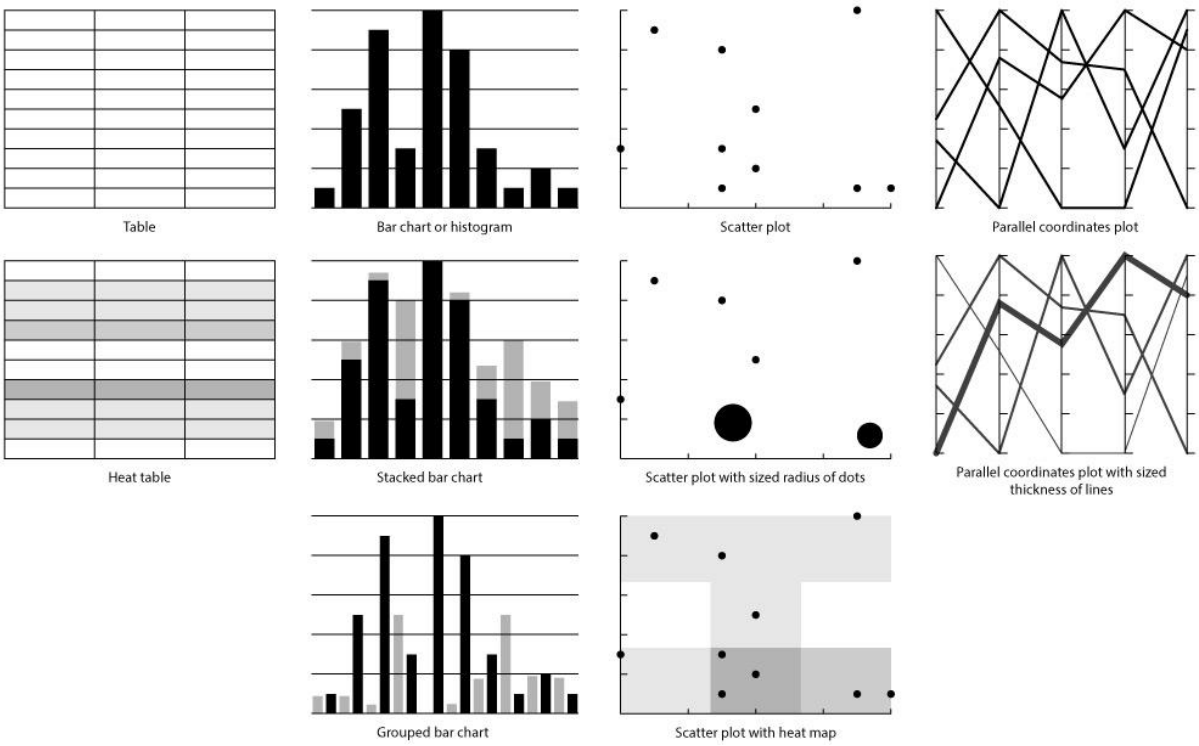


Figure 16: The four different visualization techniques that were most successful during the first round of user tests and literature reflection. Different variation examples are shown on the second or third row of images.

Second prototype round

During the first prototypes the aim was to use real data, but it became clear that using real scenarios are just as important. The author was, during the first prototypes, less aware of the semantic meanings of each variable and that some variables are not realistic to compare. The analysts were therefore asked to provide a realistic dataset *and* an accompanying scenario and story.

The author applied an iterative design process during this design by using Pretorius and Van Wijk's (2009) two mutually enforcing questions: *What does the user want to see? What does the data want to be?* When a visualization designer switches between these two perspectives, he can address design issues that could have been hard to find otherwise. Pretorius and Van Wijk (2009) explain that information visualization is as much about understanding a problem as it is about finding solutions. In this case what does the user want to see was reflected in the use of data questions and what does the data want to be is guided by what literature and previous projects suggest on visualization types.

The *levels of reading* theory, which suggests the best visualization should be able to show elementary, intermediate and overall levels of reading, was also proved helpful. But because D'Amico and Kocka (2005) and Ribarsky et al (2009) also claim that multiple diagrams could be necessary to show all perspectives of the data, it made sense to study how many and which diagrams would work best for the analysts. The author therefore decided to group all visualizations for each scenario together in one HTML page so that the analysts could interact with all and decide which one they found most useful.

Because providing overview in the data is key, the author revisited multivariate visualization techniques recommended by other scholars (e.g. heat map, parallel coordinates plot, small multiples, matrix, complex scatter plots, glyphs, tree maps, networks). Still the parallel coordinates plot was the only one that was able, in a logical and most simple way, to plot all individual sessions.

The author made use of JavaScript (and the JavaScript library d3.js and jQuery), HTML and CSS because it allowed a lot of interaction options. These tools are often used in the field of information visualization. The interaction techniques applied were limited to the authors abilities of programming within a short amount of time. This did not necessarily mean that a limited batch of techniques could be tested as the data manipulation and meta interaction (based on the interaction Table 11) could be applied in the next prototype round. The view interactions were most relevant during this round, within these categories only the change range (e.g. zooming and panning) was not applied as this became too slow when applying it to the prototypes.

Additionally, all visualizations could have all interaction techniques available. However, this would make the prototypes more complex and was probably not necessary. The author therefore made sure that each interaction technique was used at least once in the different visualizations and scenarios. One trick that helped the author to make the visualizations quicker was pre-processing the data (e.g. already counting occurrences for distribution visualizations) also the library d3.js and its community offer a great amount of examples to work with and were used during the designing.

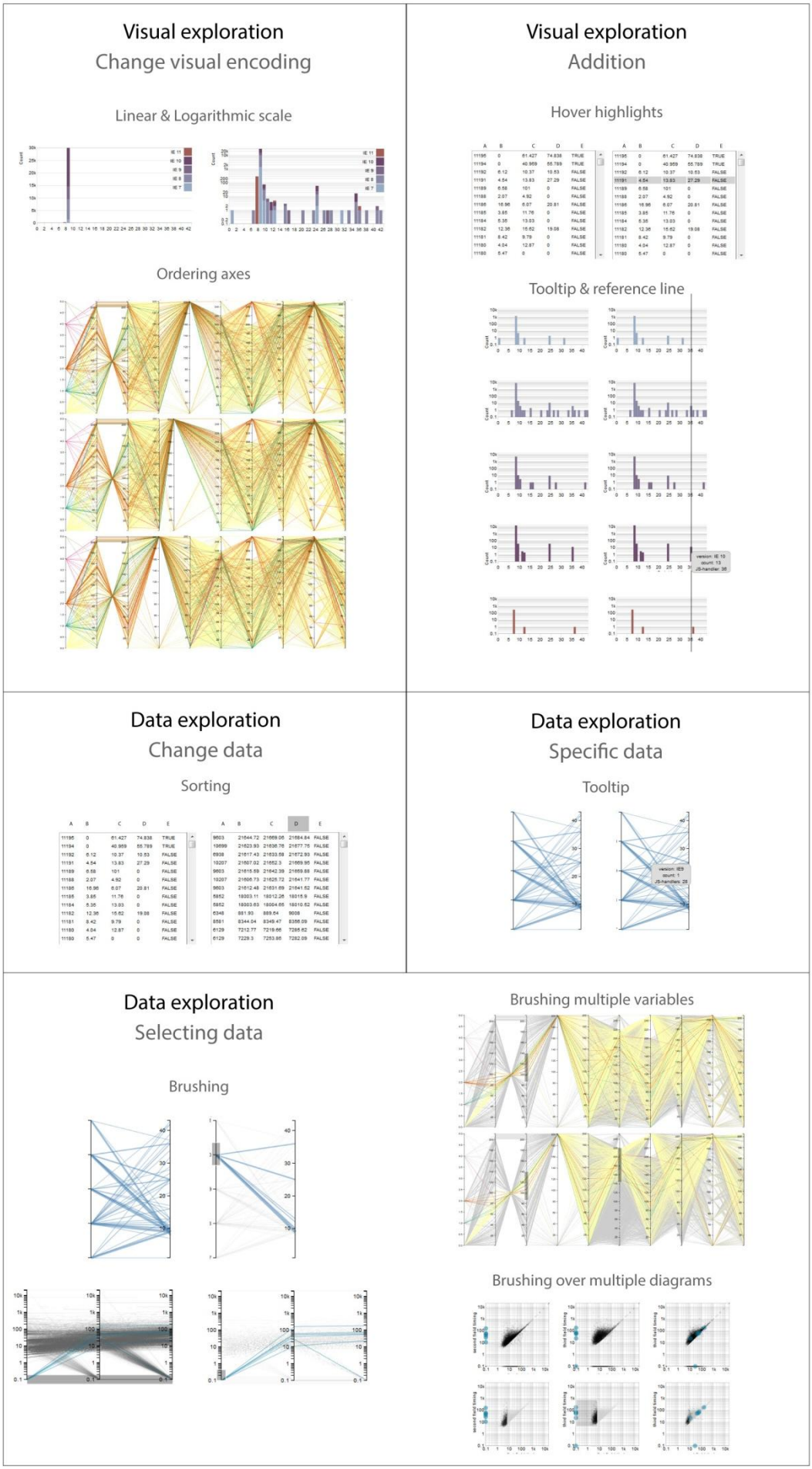
The colour and style used in the prototypes were based on the examples taken from d3.js and the authors intuition on neutral colours versus the ability to see different groups within the data (if present). For example, showing grey coloured groups is less clear than coloured groups therefore the colours of the d3.js example were copied.

When the visualization or interaction techniques became complex the author wrote descriptions with the visualizations to allow the analyst to use this as information during the user tests.

The author also made a separate batch of prototypes in which the designs of the scenarios data elements were highlighted in different manners. The different highlighting techniques were based on the theories *feature integration theory*, *visual channels of accuracy*, *gestalt principles* and other recommendations from the field of perception. The highlighting techniques that worked with the visual encoding were applied (e.g. a dot in a scatter plot does not allow visual cues on areas like patterns as it does not have a large area).

Based on the data questions (Chapter 4.1.5), visualization and interaction database (Chapter 4.3), recommendations from literature and methodologies (Chapter 4.2 and Appendix A.3), and user feedback (Chapter 5.2) the author designed the second batch of prototype. Examples can be found in Figure 17.

Figure 17: (next page) Second round of prototype design examples. Each bold heading references to an interaction category and each grey heading to a sub category mentioned in Table 11. To show the interaction technique on the static page, the author put different screenshots of the visualizations next or below one another. For example the top left image shows the visualization with firstly the linear scale and secondly the logarithmic scale.


Visual exploration
Change visual encoding
Linear & Logarithmic scale
Ordering axes
Visual exploration
Addition
Hover highlights
Tooltip & reference line
Data exploration
Change data
Sorting
Data exploration
Specific data
Tooltip
Data exploration
Selecting data
Brushing
Brushing multiple variables
Brushing over multiple diagrams

From the second prototype the following visualization and interaction techniques worked best for the analysts and their data questions (see also Figure 16 and Figure 17 for examples on the visualization and interaction techniques mentioned here):

- A table with possible additions like heat. This will show the elementary and intermediate level of the data
- A simple bar chart to show distributions for which a smart solutions has to be thought of when a variable consists of different subgroups
- A scatter plot with a heat map to show relations between 2 variables
- A parallel coordinates plot to show a multitude of variables which should allow the analyst to see which variable is interesting for further inspection
- Changing of the scale of the axes was used often
- Sorting the data in the table was used often
- Brushing needs the be done more intuitively or even done by the computer instead of a user
- Tooltips (with links to the individual sessions)

These were again chosen based on a combination of user and data requirements and theory reflection but also on the evaluation if the users were able to make correct conclusions. The analysts could sometimes make mistakes reading the visualizations (e.g. this large circle could mean it is highlighting an interesting anomaly? While it shows the occurrences and is therefore not anomalous). The author had more experience in "reading" visualizations *and* designed the prototypes and could therefore assess whether or not the analysts made correct conclusions.

Third prototype round

Even though the main visualization and interaction encoding were chosen, the author still had freedom to design different solutions for the third and final prototype. The choices for this prototype were made on the cumulated knowledge gathered on the domain and encoding possibilities.

The theory of Bertin on *Levels of reading* (Bertin, 1983) claims that each visualization should be able to answer questions on an elementary, intermediate and overall level. When this is not possible multiple visualizations could be used. Because the dataset is very large and complex, the author believes using only one visualization that could do all would be insufficient. In this case this would be the parallel coordinates plot. However, this visualization shows too little detail. This led the author to make use of the results in the second prototype round - which visualizations were useful for the analysts - to come to a conclusion. These included both the parallel coordinates plot but also the relevant, less dimensional visualizations (i.e. bar chart when it concerned distribution and a scatter plot for relation). The author therefore considered the recommendations from *diagrammatical reasoning with multiple diagrams* to apply perceptual and conceptual integration between visualizations with visual and contextual cues.

Visualization is mostly useful when *exploring* the data without knowing in advance what you are looking for. This is also what the analysts were doing most in their workflow (see Table 3 on page 41). To support this, the author therefore took inspiration of Bret Victor's Ladder of abstraction (2011). Victor believes that support exploring you should move between levels of abstraction. An example Victor gives is when you move to a new city you could learn the territory by walking around or use a map, but using both together is most effective. The different visualizations for the visual support tool should also resemble different levels of abstraction, combining or showing data variables or focussing in on one. Victor's work shows how to move fluidly up and down the ladder of abstraction with the

help of smart interaction. His work was used as inspiration that the visualizations of the support tool should also be as adjustable and interactive as possible.

Victor continues by saying that each system has the same anatomy: Independent variables, a structure and data. The structure is that which we control and adjust. In this visual support tool the structure is the rules that the analyst needs to make. Victor gives examples how this can be done with interaction. The author used this as inspiration for the rule-making aspect of the prototype. It should be possible for a user of the prototype to design and adjust selections and sub selections and see, when this selection is used as a rule, what influence this has on the number of false positives and other statistics.

The author considered visual ways of showing rule-statistics, however these were less clear. The author agrees with Ware (2004) and Doumont (2004) that textual signs are sometimes better than visuals when no ubiquity is allowed. The statistics are therefore shown in a separate space within the prototype, not hidden behind tooltips as it needs to be visible at all time as it should greatly influences rule-making.

During the design process the author made sure not to get lost in design details but to keep in mind the goal: supporting the data analysis workflow. The author used the earlier mentioned CTA (chapter 4.1.4) analysis for this.

The dataset used for this prototype was a new one for which data between a specific timeframe was taken. Because this dataset was not available during the beginning of the third prototype round, and to allow changes to the dataset, the author made sure the prototype was designed in such a way that it was independent on the dataset elements and variables. For this, the author used a so-called configuration file to adjust specifications when needed (e.g. dataset file to be loaded into the tool).

During the designing of the prototype the different data types needed to be handled differently (e.g. the nominal data and the quantitative data needed to be plot differently as the nominal data did not have a numerical value so the *textual* values need to be mapped to a *numerical* position). But there were also other data exceptions. For example variables with missing data also needed to be plotted. Within this specific case, missing values cannot be excluded from the sample (as some analyses do). Missing values could be an outlier, i.e. abnormal behaviour, in the dataset and therefore make a good rule to stop fraud. The author therefore showed these values on the parallel coordinates plot on the top part of the axes (see for example the cluster of lines going to the top of the fourth axis of the parallel coordinates plot in Figure 19). The author did not include the missing values in the scatter plot or bar charts because of time constraint.

Based on the results of the highlighting prototypes, the author chose colour as the highlighting technique of sessions. Important lessons learned for highlighting in the parallel coordinates plot and the scatter plot are that the highlighted visual elements need to be moved to the front of the visualization (as they are otherwise overlapped) and need to be less opaque than the non-highlighted elements. Furthermore, it is suggested not to use too much colours and that differences in grey are well perceived by a user. The author therefore kept the rest of the tool in neutral grey tones. The colours used for highlighting are based on the guidelines from literature (e.g. enough luminance contrast).

Additionally, in accordance with the field of Technical Communication, the author made sure there were instructions on each visualization and interaction technique available via an information button.

Based on the data questions (see Chapter 4.1.5), visualization and interaction database (see Chapter 4.3), recommendations from literature and methodologies (see Chapter 4.2 and Appendix A.3), and user feedback (see Chapter 5.2) of the second iteration round the author designed the third prototype, see Figure 18 and Figure 19. For more screenshots on the different states of the third prototype, see Appendix G on page 188.

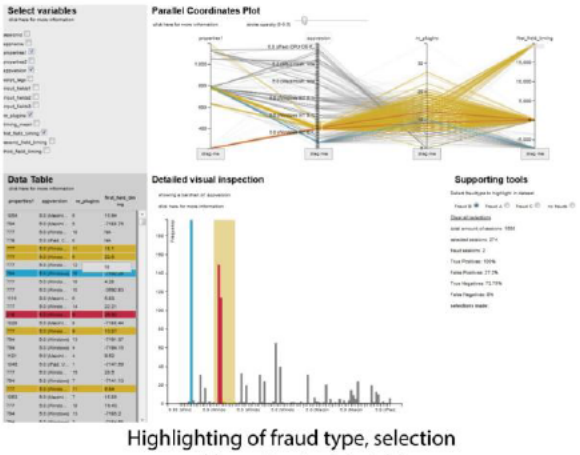


Figure 18: Combining three different highlighting functions in the third prototype. The fraud type is highlighted in red, a selection of sessions is shown in yellow and the hovering of the mouse over a data element is shown in blue. For an introduction to the most distinctive functions and screens of the third prototype, see Figure 19.

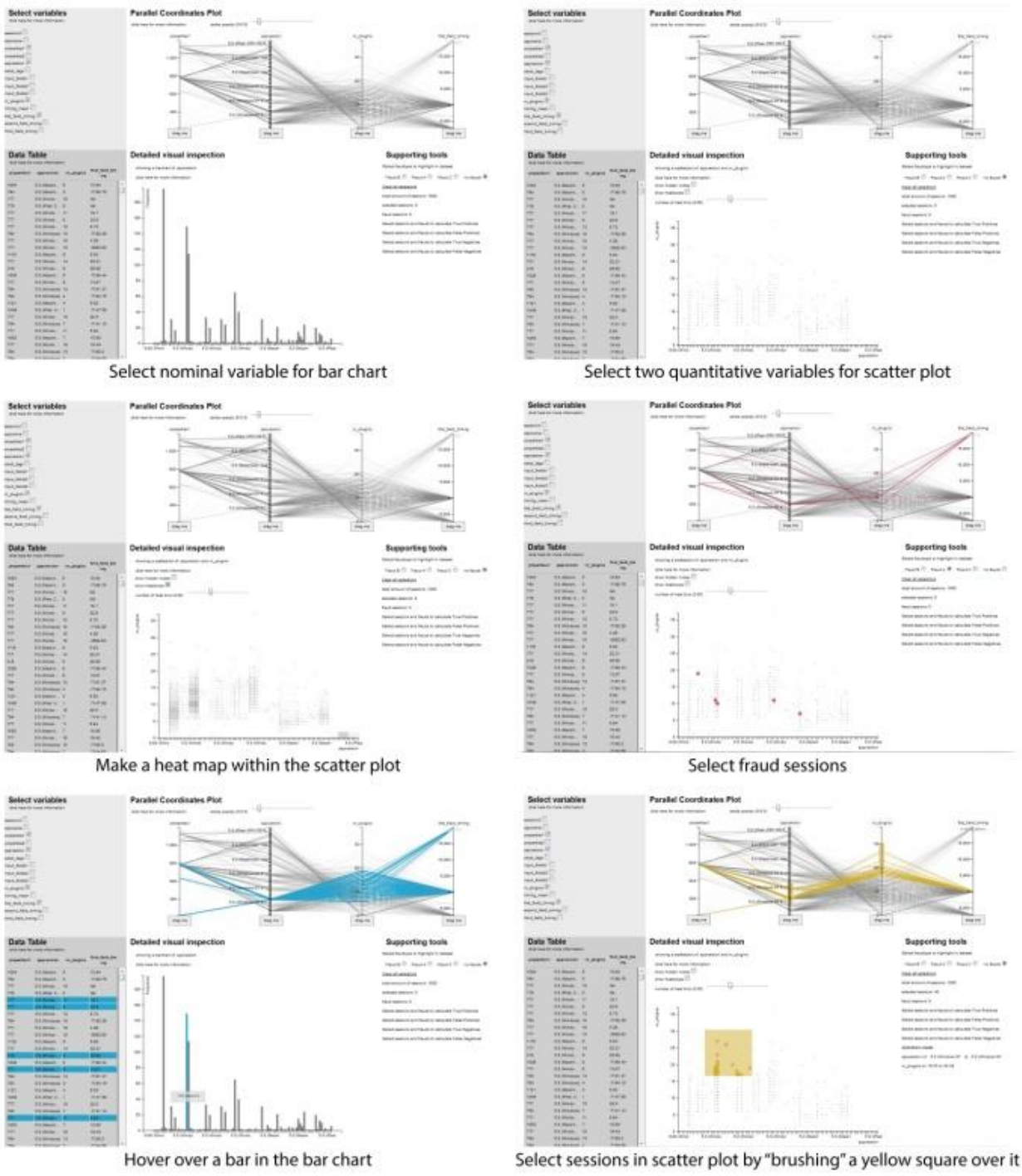


Figure 19: Different states of the third prototype. The prototype is presented in a web browser and allows interaction to create the different states. Selecting needs the action of the user to click on or drag (i.e. brush) over specific variables or sessions (see image 1,2,4 and 6). Hovering means that the mouse only needs to be at the same location of that visual element (see the 5th image). Making something means the user needs to select a button so that the tool will show a specific visualization (see image 3). For more screenshots on other states see Appendix G on page 188.

Because of the limited time and programming skills of the author, there were still techniques unexplored in the prototype rounds. The author deducted their use for the final proof of concept based on the user feedback.

Unfortunately it was not possible to implement all minimalistic guidelines and examples of Tufte (2001). For example the axes could be redesigned to be read more easily, occupy less space (and attention) just enough to let the user be able to read the values of the data points, or that they add distribution information to the visualization (see Figure 20).

What is most notable here is that the axes in the parallel coordinates are hard to optimize in design. For the parallel coordinates there are also optimizations options for the lay-out (e.g. automated ordering) of the axes for example the work of Dasgutpa and Kosara (2010) where they took into account the minimum number of crossings, the maximum or minimum angle at crossing, optimizes for parallel lines and used inversions of axes.

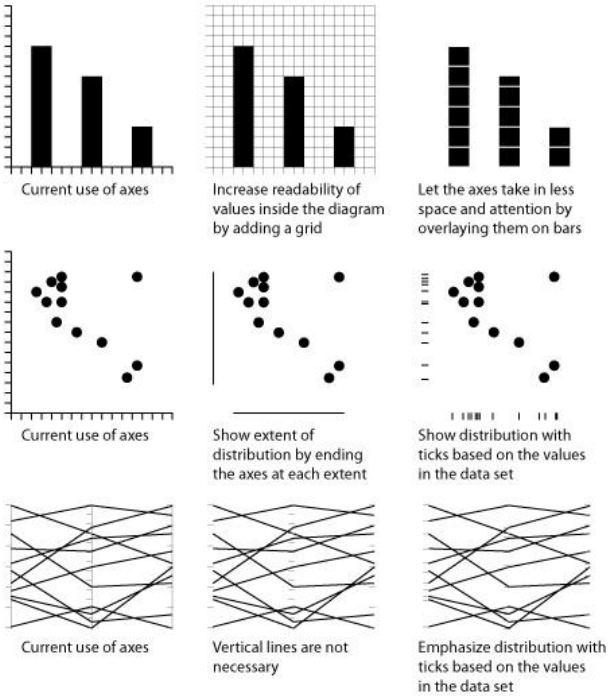


Figure 20: Examples of optimization of diagram axes. Images made by the author, based on recommendations of Tufte (2001)

Another optimization could be the bar width of the individual bars in a bar charts. In fraud analyses there are often variables that could have long “tails” in their distribution. This could be caused by a outliers with extreme values. This would make the bars very small and hard to see. An option is to make the bar width stretch over the last values. This will still show a bar that is easy to spot, but also lets the user see that it spread over a large value range. Different bar widths have been applied in histograms before.

During the user testing of the last iteration the author made a list of all techniques that were not implemented because of time constraints. Based on the user feedback and literature reflection, some of those techniques are recommendations for the proof of concept (see Appendix H). Note that during the making of the prototypes the author became very aware of the technical challenges some interaction techniques pose. For the final proof of concept the author did not take these limitations into account as a professional programmer should be able to program these techniques more easily.

5.2 User feedback

For each iteration round user testing was done to gather user feedback. A test protocol was written for each user test (chapter 5.2.1). The results of the user tests were recorded by taking notes and taking pictures or screenshots. They were later evaluated for conclusions in the form of design requirements (chapter 5.2.2). Each of the different prototypes had a different focus (see Table 12 on page 61). The user feedback and its formulation into design requirements helped to answer research question 3 (which design elements can support the analysts?).

5.2.1 User test protocols

The first user tests were relatively structured and took around 1.5 hours per analyst. The goal was to introduce a visual DetACT to the analysts and see how they would respond to this. Which visualization could benefit them most? Which visualization do they all understand best? The protocol helped to structure the tests, but there was room left for valuable information that could arise outside of this structure. The analysts had to perform specific tasks including ranking visualizations and therefore part of the result was quantitative. However, most of the useful information originated from the qualitative results in the form of responses on questions or remarks during the tasks.

"Number fetishism leads usability studies astray by focusing on statistical analyses that are often false, biased, misleading, or overly narrow. Better to emphasize insights and qualitative research." (Nielsen, 2004)

The user tests were set-up in such a way that it started relatively familiar for the analysts with their currently used session reports. They were asked to create their ultimate session report based on different building blocks: the current design of sections, the current sections with added highlighting and textual information and the current sections with added visualizations when possible. This allowed them to start thinking towards a more visual way of handling their data.

After this, the different data questions and their visualization options with fictitious fraud types were given to the participant and asked to be ranked on preference. Last, some final questions were asked about the representativeness of these prototypes and how they could see visualization within their workflow.

The full User test set-up of round one can be found in Appendix G on page 188.

While designing the second prototypes, more explicit scenarios were used. For this all possible visualizations were put in an interactive HTML page. Research questions for the following prototypes were formed based on the first user tests and questions that came up during reflection on theories. They are:

1. What interaction is needed to support the analysts during their analysis
2. How many diagrams should be shown together and is that preferable over one diagram.
3. How to highlight fraud sessions without distorting the visual coding of a visualization
4. What added information do the analysts need when exploring the data visually

As the qualitative information proved to be most valuable, the second user tests were set-up to find mostly that. They were semi-structured to including an introduction and a main question that the analyst needed to answer. But most of all they were asked to respond to the prototype and "play" with it. The author also asked specific questions on each visualization and interaction technique if the analyst did not mention it himself. These tests were done in-between work of the analysts and did not take more than half an hour per test per scenario. These tests resembled evaluations in the group of User experience (Lam, et al., 2012) which aims to find out what the target user thinks of the visualization.

The highlighting prototypes were more structured, the analyst needed to find specific data points in a visualization and were timed on how long it took. This resembled evaluations on User Performance which aims to find the best technique as measured by human performance.

The user-tests were distributed over the weeks. Some tests were done right after one another and some individually. This also differed per analysts. In this way development of the prototypes and user testing could be done parallel. The user tests for highlighting is done during the same time span. This allowed more in-between feedback to prevent time spend on making visualizations that are not useful and it is in line with the design based research approach where there is close communication between researcher and practitioner.

The full User test set-up of round two can be found in Appendix J on page 195.

For the last prototype the whole workflow of rule-making was taken into account. It was put in an interface-like tool. Research questions for the following prototypes were formed based on the second user tests and questions that came up during reflection on theories. They are:

1. Is the overview visualization understandable?
2. Does selecting a range for input for a rule work?
3. Are the analyst able to determine interesting variables?
4. Does multiple highlighting in colour work?
5. Does this prototype encourage exploring?
6. Are the analysts given confidence with the statistics?
7. Do the analysts need more statistical values?

The user-test set up was inspired by the approaches of evaluation of Visual Data Analysis and Reasoning (VDAR) which evaluates if a tool can provide actionable and relevant information for a specific domain.

The user tests took around 75 minutes and were done closely after one another. The full User test set-up of round three can be found in Appendix K on page 198.

It is interesting to note that within the computer science visualization field the number of evaluations done on the process and roles of visualizations is relatively low (Lam, et al., 2012). Implementing evaluation based on the role of the visualizations within analysis can give "high practical value beyond specific individual tools and can benefit both researchers and practitioners in all areas of visualization." and including this within this research offers a more holistic view of user experience. The other evaluation types are less relevant to this project as they focus more on subjects like algorithm design, or how tools should be implemented during day to day work.

5.2.2 Design requirements based on user tests

Based on the evaluation of the prototypes, a list of requirements was made to help guide the design process of the proof of concept. Note that obliging to these requirements does not necessarily mean a foolproof product as each design should again be evaluated to reflect on its success.

Each requirement is justified by information from various sources (e.g. literature, field observation and user tests). The user feedback is crucial and influences the requirements a lot. This is the reason why the list of requirements is given in this chapter. Note that the raw results and details of the user tests are available for Fox-IT but because of potential sensitive information are not disclosed in this report.

The time available for this study does not allow to fully develop all visualization and interaction techniques. The requirements concerning these techniques therefore do not claim that other techniques would not be successful in a visual support tool, but this filtering was necessary because of the limits of this study and because the proof of concept was made for this case specifically. Still, when a similar dataset for similar fraud analysis needs to be made, these visualizations could work very well.

The final list of requirements for the visual support tool is given below and is categorized in workflow, visual encoding, interaction encoding, highlight encoding, algorithm design and data. Note that the author made use of the MoSCoW (Must have, Should have, Could have, Would have) technique to assign a level of importance to the requirements.

Workflow

1. *Must* support rule-making.
2. *Should* provide a link between visual data elements and the original data-elements.
3. *Should* allow easy exploring of the data.
4. Instructions *should* be made possible.
5. *Could* allow annotation and communication means.

Visual Encoding

6. Visualizations *must* be readable.
7. Visualizations *should* be understood.
8. *Should* be able to change visualization encoding based on data (combinations).

Highlight encoding

9. Highlighting *should* be done with colour and obey perceptual guidelines.

Interaction encoding

10. *Must* include linking.
11. *Must* take into account hovering and clicking.
12. Interaction *should* be intuitive and well understood.
13. *Should* allow optimization of visual encoding.

Algorithm design

14. *Must* be able to handle the relevant data types.
15. *Should* have a representative sample.

The following paragraphs explain these requirements in more detail.

Workflow

1. Must support rule-making

The main aim during fraud analysis is making sense of the fraud and its characteristics. The rule is a form of identifying this fraud in the past but also in the future, so it can be prevented. The use of a visual diagram is very useful in the reasoning on an appropriate rule (i.e. *diagrammatic reasoning*). Adding interactive elements for the forming of this rule would increase the possibilities of *analytical discourse* for the analyst and could positively influence *abductive reasoning*.

The author believes that this process should be enhanced with computer-aided discovery (Green, et al., 2008) by suggesting variables or even complete rules. Automatic solutions will never fully take over the role of the analysts, as analysts are more easily aware of the semantic meaning of variables, while an automatic solution could mistake statistical importance for real importance. Small computer aided additions could also be implemented for dynamic querying and auto completion of search terms.

For the selection of rules, the visuals should be able to make a "global" selection. But more precise rule selection should also be possible. The author suggests a separate selection tool for this.

The final output of the visual support tool should be a description of a new rule and its efficiency in the form of false positives. The false positive ratio can therefore be optimized for the situation or for client preferences.

> "I could filter according to my needs, but it would be nice to select exact values of variables"
>
> – DetACT Analyst during user testing of the third prototype

2. Should provide a link between visual data elements and the original data-elements

It could be that this is mostly useful at first implementation of the tool, when the analysts are still used to their old workflow. The author believes there are also upsides to linking it to the original data elements, in this case session reports, in later stages of use.

> "I would want to see more of this specific session to better form my hypothesis"
>
> – DetACT Analyst during user testing of the third prototype

Analysts often want to confirm hypotheses. The *sensemaking processes* also shows the top-down processes where the analyst wants to re-evaluate and search for support. The original data-elements could be of use to the *deductive reasoning* of hypotheses. Showing original data elements also functions as a details-on-demand interaction technique (3rd item in the *information seeking mantra*). There could be different ways of showing the data element (e.g. via a hyperlink or show a preview of the session).

Additional research can be done on the linking of data elements with other similar data elements (search by example is mentioned by the *Human Cognition model*). Doing this would help to get a more precise view (e.g. absolute numbers) on how many false positives a rule could give when implemented.

Adding a detailed data table was used in the prototypes but it became apparent that they easily become unreadable when a large selection of variables is made. It also distracted the analyst from visual thinking about analysis problems and could therefore distract during visual analysis. The author recommends trying if the analysts can do without.

3. Should allow easy exploring of the data

Exploring is what characterizes the fraud analysis process best. Exploring the data could lead to more options for *abductive reasoning* that could be necessary when the clues become weaker which is often the case in fraud analysis. It also allows more back and forth interaction between the analyst and the visual results, positive for *analytic discourse* and *diagrammatic reasoning*. Ware (2004) also recommends that this process should not be hampered which means the interaction should be done fluently and with great clarity because of *visual rhetoric ethics*. More detailed recommendations in the interaction encoding requirements.

4. Instructions should be made possible

Even though the analysts should be working with intuitive tools, instructions are always necessary as reference but also when the analyst still needs to figure out all the functionalities.

> "I like the tool, but I might need a mini-course on how to get the most out of it. I might not have been using all interaction possibilities"
>
> – DetACT Analyst during user testing of the third prototype

According to different Technical Communication guidelines (i.e. *information design*) the textual instructions should be enhanced with visual diagrams. The author believes the instructions could be integrated in the visual support tool by means of opaque previews or with small pop-ups that occur during specific interactions (the latter is also recommended by Green for automation of computer aided discovery (Green, et al., 2008)).

This is especially recommended for the visual rule-making as this is new for the analysts and could therefore need some support, because it is crucial to get the best out of visual fraud analysis.

5. Could allow annotation and communication means

The analysts still needed annotations to write down important variables. The author believes they were not aware that they could do the same visually (e.g. by selection important sessions) and that they will become more skilled the longer they use the tool. However, to communicate their findings to other analysts or clients, annotation could be used as means of communication. The annotations and supporting visualizations should then also be able to be exportable in easy to handle file formats (e.g. .doc/.pdf/.jpeg) or in a permalink to the specific page which includes the exact view at that time.

> "I'd like to have a permalink to a page with all currently selected options, so I can share the screens that led me to a conclusion with my colleagues"
>
> – DetACT Analyst during user testing of the third prototype

Visual Encoding

6. Visualizations must be readable

Readability of the labels and other textual elements in the tool is important. But being able to "read" the visual elements is also important as otherwise the eye could miss important information during its rapid-eye-movement scanning (*visual problem solving*). It is also important, as the *cost-of-knowledge* would still be very high as the analyst would need to concentrate more and could not use the tool as proper *external memory*.

For text, a large enough font size and contrast with the background is important. For the visual element the contrast with the background, be it the real background or the overlapping of other

elements, is most important. For example, several analysts had mistaken a line to be the axes of a diagram. This can be avoided by using proper contrast and difference in visual coding between the two.

Additionally, bars in histograms or bar charts can become too small when they contain little data elements. But these bars could be the most relevant ones as they contain values that are not that common. The author recommends changing the bar-width for these values so the bar-height will be large enough to see.

To make potential messy visualization, e.g. parallel coordinates plot, better readable there are techniques on how to update the lay-out of such visualizations. For the parallel coordinates there are methods of re-ordering or inverting the axes based on crossings, angles of crossings and parallelism. Other options are the reducing of the opacity of the visual elements. Still the designer must be aware that during analysis the analyst might want to change the opacity as he might want to see distributions within clusters or wants to see outliers more easier.

7. Visualizations should be understood

> "I first found the Parallel coordinates plot confusing, but here it actually shows me a clear view on the data"
>
> – DetACT Analyst during user testing of the second prototype

Analysts are welcoming visualizations and are willing to learn if necessary. But simple visualizations are often more intuitively understood and require less *visual literacy* skills. They need less initialization costs, which, according to the *Van Wijk model* a good influence on the value of the visualization. Using simple visualizations makes it more easy for the visualization designer to convey the right *visual rhetoric* (e.g. this is a distribution is clearer in a bar chart than in a bubble chart).

The author recognized the following visualizations as easily understood by the analysts: the data table (with possible additions) for familiarity of the data, the histogram or bar chart to see distributions and the scatter plot for relations between two variables. The parallel coordinates plot is a good visualization for the multivariate data but might need some additional instruction to make all functionalities clear.

8. Should be able to change visualization encoding based on data (combinations)

To show different *levels of reading* the data and to adjust the tool to specific *visual problem solving* strategies the visualizations should provide some flexibility. The analysts will need different tools for different types of hypotheses (e.g. distribution or relation related) the tool should be able to reflect this. The author recommends using the parallel coordinates plot as overview visualization with added bar charts or scatter plots to show more details on 1 or 2 variables. The parallel coordinates should always be visible to connect the data together. For this an additional data table could be added.

> During the round of the second prototype the analysts were asked to argument their hypotheses with specific visualizations. The visualizations chosen depended on their hypotheses (e.g. one analyst chose the bar chart while another the scatter plot)

Highlight encoding

9. Highlighting should be done with colour and obey perceptual guidelines

Colour is mostly used for highlighting and is one of the best alternatives to any other form of highlighting according to Ware (2004). The author confirmed it in the prototypes and recommends using coloured highlighting. However, the designer needs to adhere to *perceptual guidelines* like enough luminance and saturation contrast and not too many different colours. It is important to keep the highlighting encoding consistent over visualisations (see also requirement 10) as this could help to get a clearer view on what is highlighted exactly.

Overlap of highlighted elements should be avoided as they become less clear. These elements should therefore move to the front of the visualization. Additionally, because the dataset is very large, the visual encodings where each data point is plotted (e.g. scatter plot) will likely have added low opacity levels to their visual elements. When these elements are highlighted they should increase the opacity of these elements.

The designer should take into account how to highlight aggregated elements in the visualizations (e.g. bars in a histogram). Highlighting the complete element feels most intuitive but could also be misinterpreted. An analyst is not able to know if they contain the whole bar or just a slice of the bar. The author recommends using both the stroke of the bar as highlighting and add a coloured bar as overlay to represent the number of sessions that is highlighted.

During the first round of prototypes some analysts expressed their concerns on the readability of the highlighting and how they should read it. During the later rounds the analysts did not have problems reading the highlights which could be explained because of the added interaction that linked highlighted elements together over different visualization encodings.

Interaction encoding

10. Must include linking

Linking diagrams together with visual and contextual cues is recommended for *diagrammatic reasoning with multiple diagrams*. Preferable each data element should be linked and this link should be made visible by means of hovering or selection. The author believes the additional linking between a data table and the visualizations could help analysts better understand the visualizations and how they work.

11. Must take into account hovering and clicking

"For this pie-chart, how would I be able click on one of the small elements of the chart?"

- DetACT analyst during user testing of the first prototype

During the user tests the importance of selecting individual visual elements became clearer. The visual elements should be easy to select for rule-making which is easy to achieve, but hovering over small lines or dots is nearly impossible. The author believes a more friendly hovering interaction can be achieved when the mouse or the visual elements have a larger area that triggers hovering. This is similar to hovering over a bar in a bar chart, as it will highlight all data points in that area over all visualizations.

12. Interaction should be intuitive and well understood

In contrast with the visualization techniques that allow the analyst to learn it, interaction should be understandable. The author believes that "interaction literacy" is something that most people still need to grasp better. Using less intuitive interaction could therefore be less forgiven by the user.

Using different scales, allow sorting, and tooltips (for *details-on-demand*) are well understood interaction techniques. Additional helpful interactions are brushing and zooming but need to be well implemented by making use of current interface conventions. The reordering of axes is something that could help. But because it changed the visuals greatly it could hamper the rhythm of reasoning of the analyst as he needs to rebuild his image of the data when the axes are re-ordered. To prevent this, animation could be very helpful in connecting the dots between the previous and current view. Though used for the re-ordering of the parallel coordinates plot could still be complicated for users.

Specific recommendations on which interaction *must* be in a general visual support tool cannot be provided yet from this study. To give recommendations the designer should try different techniques to see what works well and could use different interaction taxonomies for inspiration (as the author did in this study).

13. Should allow optimization of visual encoding

Each dataset is different, in size but also in distribution of data points. This could hinder some visual problem solving strategies as for example the opacity for lines in a parallel coordinates plot is helpful for a normally distributed dataset but not when most elements overlap each other. Allowing the user to change the opacity is therefore recommended as it tries to optimize visualizations for better perception which is the first step in *situational awareness* to grasp what the dataset is about.

In projects on similar cases these adjustments are linked to the distribution of a variable making it possible for the user to adjust the encoding for different parts in the distribution instead of for the whole dataset. However, the author believes that this is for the experienced users only as this could also lead to misreading visual information when the user forgets his adjustments.

> "This visualization could become a mess with a lot of variables and sessions, it would then be hard to see an individual line"
>
> - DetACT analyst during user testing of the first prototype

Algorithm design

14. Must be able to handle the relevant data types

> "The visualisation of missing values is problematic. It could lead to wrong conclusions"
>
> – DetACT Analyst during user testing of the third prototype

It seems like an obvious requirement but sometimes this can be overlooked. In this case missing values are important and should not be left out of the dataset, which is sometimes done in other fields of analysis. The visualization but also the algorithm should be able to handle these values. Based on the user test the author proposes not to plot missing values as an additional fictional value but use a different way of coding them. Based on examples of Theus and Urbanek (2009), the author suggests using a different coloured bar to represent the amount of missing values in a bar chart. In the scatter plot the dots can

be plotted on the outside of the axes (instead of inside the grid) and in the parallel coordinates plot an additional small axes could be placed above each axes, representing missing values on that variable.

The author believes that making the tool as independent of the dataset specifications as possible the more it is made into a robust tool. This way the analysts could also use the tool when the data changes.

15. Should have a representative sample

The third prototype was able to handle a dataset with 1000 data elements with 14 variables. When using larger samples the tool became too slow and multiple experts emphasize the need for a quick reaction time (e.g. (Ware, 2004)). For the final tool the analysts should be able to handle a sample that is more representative for the complete dataset (5000-10000 data elements according to the data scientist). The author believes this is possible with the right optimizations.

Still, when optimization is hard to do, the designer should study how much influence a different sample size has visually. When the same conclusions can be made with a smaller sample then it is representative *enough*. The author tested the differences in sample for the third prototype. The conclusions can be found in Appendix L on page 201.

Another solution could be not to plot each data element but only plot clusters. Or the analysts could have the option to change the samples to see whether or not the same False Positive ratio will occur in another sample dataset.

5.3 Reflection on theory and practice

For each iteration round the author reflected upon the prototype designs with the help of theory. The reflection method is shortly described in chapter 5.3.1 and the insights that the reflection gave are explained in chapter 5.3.2. For a more detailed description of the different theories see Appendix C. The reflection on the prototypes by making use of theories helped to answer research question 1,2 and 3 (How can theory be integrated into a visualization model, how can that model help in designing a visual support tool and which design elements can support the analysts?).

5.3.1 Reflection methods

After each round of iteration a reflection round on theories was done. The insights that the author gained were used to structure the theories into a model. Each iteration includes a description of a new structure within the theories and a visual model of this structure, including the previous categorizations. This way the model is shaped and re-shaped until a final model is formed.

This description makes it seem like the researcher takes a passive approach in this, the opposite is true. The researcher needs to "design" each iteration of the model to take into account all requirements of categorization or findings found during reflection. The researcher did this visually by making use of Adobe Illustrator Creative Cloud (CC). This way the model elements can be easily shuffled or re-grouped and categorization in the form of colour or containment is easily done (or removed).

For more explanation on the method of reflection and structuring, see Appendix N.

5.3.2 Reflection insights

The reflection of the 36 theories found in literature on the different prototypes was done with means of a big table. The tables for each prototype round can be found in the Appendices P on page 228 (after 1st round), Appendix Q on page 232, Appendix R on page 237 (after 2nd round), and S on page 241 (after 3rd round). After each reflection the author tried to make sense of what was learned about these theories by listing the insights and structuring the theories into a diagram (see Appendix O).

Because the raw results of this reflection are too detailed to mention all, this chapter will give the main insights. Both the reflection and the structuring insights helped to build the visualization model. The insights are mentioned per iteration round to show how the insights grow after each round of reflection (first very related to the research fields and categories of theories and later on how they coincide). Note that the theories are written in italic.

After first round

The research fields

It became clear that the field of visual perception does not describe a visualization communication function. The field of semiotics does not provide theories that asses different prototype designs. It could therefore be wiser to use other fields for more concrete guidance. For example the field of Visual Analytics which is the only field in this research scope to offer interaction theories.

Relation between theory and practical function of the theory

It seems that the more abstract theories are, the more they describe the communication function of a visualization. These theories are higher level in a sense that they do not assess individual elements of the tool but more the concept of the tool. Therefore, it could be that these theories are most valuable to link to the communication function a designer has in mind for a visual support tool.

The *visual channels* do not always seem corresponding to what the author found during the user tests. For example the dot chart is coded with location and should therefore be perceived accurately, but during the user tests the analysts thought they could miss the small dots and therefore read the visualization less well.

Visual literacy shows that more familiar graphics are easily negotiable by users and *diagrammatic reasoning* becomes more important when a diagram shows more variables.

The *ten classes of sign* describe all visualizations very similarly, the only differences the visualizations could have is if they have an informative icon (e.g. diagram) or signalling indices (e.g. arrows). How the *code* theory could group different visualization techniques seems to have to do with the amount of variables to display and what lay-out (e.g. Cartesian) it uses.

Relationships between theories

The author expected some sort of relation between the *gestalt principles* and the *feature interaction theory* because they both describe perceptual processes, however there does not seem to be one. The two theories of Bertin (1983) (i.e. *levels of reading and function of a graphic*) seem very related which is logical as Bertin claims a visualization should be able to allow all levels of reading and if this is not possible, one should consider the function of a graphic to allow for specific ways of reading.

The author found different types of theories, some described interaction, analysed tasks, the communication function or the visual encoding while others gave a judgement on the visualizations or help to align the design process.

After second round

Summarizing theories

There are some theories (i.e. *visual rhetoric ethics, two neural processes, codes*) that are overarching and do not have precise measurements of their own, the author therefore formulated them based on what the originator of the theory described in his paper or book. The author thinks that these theories might be summarizing the more lower level theories (i.e. *ten classes of sign, gestalt principles, feature integration theory, visual channels of accuracy, visual rhetoric ethics, data ink ratio, lie factor*) and could use those as more precise measurements.

Relationships between theories

There are theories that seem to have a relation with each other. For example, the more need for *diagrammatic reasoning*, the less visualization becomes negotiable according to *visual literacy*. Which makes sense as diagrams with a lot of information could become complex and less easy to understand. *Diagrammatic reasoning* also seems to be more valuable when the diagram has interaction options as doing this in one's head becomes more difficult.

The negotiability of *visual literacy* seems to be related to whether or not there are *separable or integral visual dimensions* used and if the intersection of *bottom-up and top-down neural processes* occurs easily.

More clarity on theories

The earlier found judgement theories do not seem to be based on the interaction encoding used but more on the visualization type. This might indicate that there are not enough theories on interaction. There are some interactions, like scaling or aggregating values, that cause *the lie factor* to increase because it is distorting the visuals.

It also became apparent that the visualization describing theories (i.e. *gestalt principles, feature integration theory, function of a graphic, levels of reading*) can distinguish well between the chosen visualizations in the prototypes. However during the highlighting reflection the theories of *gestalt and feature integration theory* still do not show a strong relationship (as was noticed during the first iteration round).

The author found more distinct categories in the batch of theories. Some evaluate workflow, or the visualization while others only describe it. There are also theories that help explain results from the user tests.

After third round

Interaction theories

There are no real low-level theories on interaction found which made it hard to compare the interaction techniques individually. The *principles of effective animation* can be considered as one but there was not a lot of animation in the third prototype round.

There are also theories that hint towards interaction from a workflow perspective (e.g. *reference model for visualization*) but do not offer specifics. Interaction taxonomies offer more guidance (i.e. *three classes of action, information seeking mantra*) and showed that there are some functions not implemented in the tool (data editing, annotation, history from the three classes of action or extraction of the outcomes). The author consciously decided which interaction types the author wanted to test in the limited amount of time of this study.

Discourse between data and analyst

It seems that the reflection on *analytic discourse*, which describes mixed-initiatives between data and analysts, asks the analyst to start this discourse by choosing variables. When the analysts did this, visualizations appear after which the mixed-initiatives start. The first initiation of the analyst is therefore based on his intuition, which could be described as their inductive hypotheses (according to *logical reasoning theory*). The more creative abductive reasoning could be continued based on the mixed-initiatives. This is because abductive reasoning needs to perceive something first before it could start to form hypotheses. However, when more computer-aided suggestions are done (a big aspect in the *human cognition model*), the mixed-initiatives could start right away.

The prototype does not offer monitoring functionalities (one of the *categories of use*). Forecasting and communicating is also not the focus and is not very much supported. This is also apparent when the author reflected on the *six ways visualization can amplify cognition*, the only one not supported is the monitoring of data.

Perception

All visual encodings are accurately perceived according to *visual channels of accuracy* and therefore support the *visual problem solving processes*. Also, all coding is made to be *separately* perceived so the user can do different visual queries (and not *integral*). It also seems there are different levels of *gestalt principles* (e.g. proximity) present, firstly the perception of the different areas (e.g. data table, overview visualization) but also the components within an area (e.g. instruction, diagram) and within a graphic (e.g. data elements).

More use could have been made of Tufte's optimization suggestions (2001)to optimize the *visual rhetoric ethics*. However, when *data ink ratio* would be followed to the maximum, 'clickability' of visual elements could be hindered. The colours chosen by the author were optimized according to the *colour guidelines* but might not be aesthetically pleasing.

Reasoning

The linking of the different diagrams for *reasoning with multiple diagrams* is done visually (with highlighting) and contextually (by offering statistics of selections). It seems as if the author thereby made a new *code* for the support tool by adding a syntagmatic relation between the visualizations.

Based on *information design* the analysts need to understand it, be able to form new hypotheses and conclude on rule-making. However, the comprehension phase of *situational awareness* is sometimes difficult and they may need to learn how to comprehend visual information (note that the other phases of perception and projection are more easily done).

Without comprehension it is difficult to visually schematize rules or patterns(*sensemaking process*). They sometimes stick to schematizing in their head or note-taking instead of doing it visually. The analysts were able to understand the *visual rhetoric* of the author but were sometimes reluctant to try it. While in other areas they showed that their visual *literacy level* increased, for example with the understanding of the meaning and coding of the parallel coordinates plot.

With the visual support tool, the analysts got to read the data at a *different level.* Without the tool they mostly had elementary levels of reading and with visualization they can also read intermediate and overview levels. They were also offered to see more items in less time (*cost-of-knowledge function*). The author thinks that the amount of insights in less time is more important to increase efficiency of analysis. With visualizations it is harder to retrieve the individual data element, but this allows more insights on the whole dataset instead of only one session.

The analysts could stop their analysis and still remember where they left it. This shows how it could be an *external memory* and does not require to store it in one's head. The *function of the graphic* is therefore mainly to record the data (and not communicate or process it). The analysts show different strategies according to the *internal model of visualization exploration processes* (that is also described by the theory) but they all end up visiting the same steps, one way or another.

Costs

The main cost would be in the initialization phase (*Van Wijk visualization model*). However, this subject was out of the scope of this study. The implementation and deploy phases of the *nine-stage model* helped on forming recommendation on these phases.

To see the implications of this insights into a visualization model go to chapter 6.2.

Preface 3

Executive summary 5

List of Figures 10

List of Tables 12

Glossary 13

1 Introduction 15

2 Project description 19

3 Methods 27

4 Orientation 35

5 Iteration 59

6 Conclusion 85

6.1 Proof of concept 86

6.1.1 Proof of concept design 86

6.1.2 Proof of concept elements not implemented 90

6.1.3 Proof of concept in context of the analyst's workflow 90

6.2 Model for visualization 91

6.3 Answers to the research questions 95

6.3.1 Theory integrated into a visualization model 95

6.3.2 Visualization model in action 96

6.3.3 Visual support tool design elements 96

6.3.4 Visual support tool and its influence on the analysts workflow 97

6.3.5 Answer to the main research question 97

7 Discussion 99

References 108

Appendices 115

6 Conclusion

In this chapter the author firstly presents the proof of concept, which is designed based on the Design-Based Research methodology, for visual fraud analysis (see Chapter 6.1). The main aim is to support exploring in an intuitive manner and support the analysts in rule-making for fraud types. Secondly, the author introduces a more general visualization model that could benefit other visual communication designers who want to make a visual support tool (see Chapter 6.3). The model is based on the reflection of theories on the different prototype rounds and the evaluation of the author about their usefulness. Finally, the author answers the sub research questions and the main research question posed in Chapter 2 (see Chapter 6.3). The research goal was met with the more physical results in the form of the proof of concept and a visualization model (see chapter 2.2) and helped to answer respectively research questions 3 and 4 (which design elements of a visual tool support fraud analysts and to what extent do they influence their workflow?) and research questions 1 and 2 (how can theory be integrated into a visualization model and how can this model help in designing a visual support tool?). The last paragraph of this chapter specifically answers all the research questions of which the answers were found during the making of the physical results (i.e. proof of concept and visualization model?).

6.1 Proof of concept

The proof of concept is designed based on three iterations of prototyping (see chapter 5.1), user testing (see chapter 5.2) and reflection of relevant theories on the prototypes (see chapter 5.3). It gives an answer to the main research question as it shows a more efficient way of analysing large datasets versus textual analysis. The visualization model in chapter 6.3 explains how the theories could support the designing of a visual support tool in other cases. The results of this chapter helped to answer research questions 3 and 4 (which design elements of a visual tool support fraud analysts and to what extent do they influence their workflow)

6.1.1 Proof of concept design

When the visual support tool is started the analyst should have sufficient instructions even though there are no variables selected yet. The main instructions are to drag variables to different areas for which it will be transformed into a visual encoding.



Figure 21: Start screen of visual support tool proof of concept

When the analyst will hover over different areas for a long time without starting his analysis, he will get an instructional screen explaining with text and small visual illustrations or animations what he should do. When the analyst already started to work with the tool and has forgotten the instructions he should be able to retrieve them by right-clicking the area. Still, the author believes the instructions are less necessary once the analyst has started the analysis process.

The selection of variables can be done by dragging them to visualization area(s) (these areas show opaque previews, see Figure 21). To select statistically interesting variables the analyst could reference the small histograms can be found besides each variable in the variable list (e.g. an uninteresting variable will have evenly distributed values without many outliers, this will show three equally sized bars).

When a variable is dragged to a visualization location, the variable name label will show up in a row (it resembles a sort of breadcrumbs lay-out which is common in user interfaces). This also gives an ability to re-order them. When the analyst drags the labels away from the visualization location the variable will be removed from the visualization. When the user hovers over the labels or the bar chart axes, they get additional options of scaling or sorting, see an example in Figure 22. When scales are changed the visualizations will be animated to their new state as this makes it possible for the analyst to link the before and after visual encodings together (without losing track of data elements).

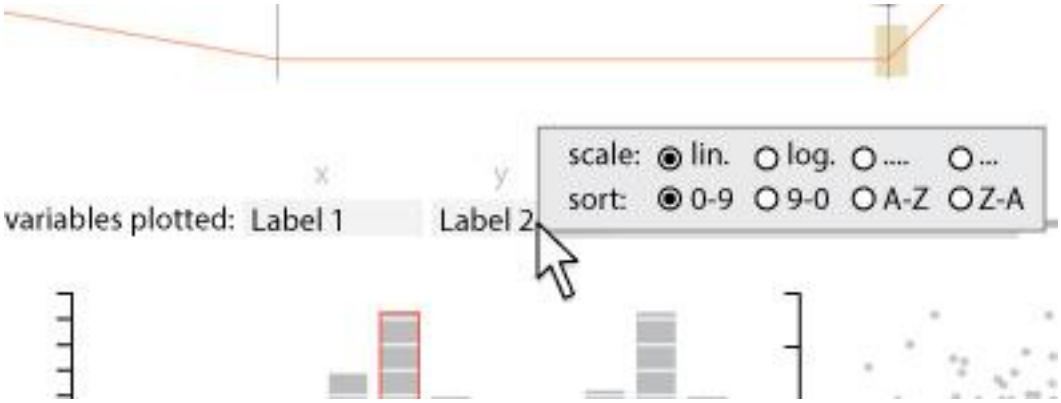


Figure 22: Change scales or sort variables in the visual support tool

On the top part of the tool, the parallel coordinates plot will show an overview of all variables selected. In the bottom area the analyst can make a scatter plot and two corresponding bar charts. When only one variable is selected, only a bar chart is shown. This way the analyst is able to see different levels within the data by zooming in on specific variables (maximum of two) in the bar charts, see their relation in the scatter plot and see its relation with the dataset in the parallel coordinates plot.

Each variable added to the detailed visualizations (at the bottom) will also be added to the overview visualization (on the top). When more than two variables are added to the detailed visualization a pop-up appears “three variables are not possible for a scatter plot. Do you want to add the variable to the overview visualization or exchange it for one of the variables already used in the scatter plot?”.



Figure 23: Proof of concept where four variables are selected in the overview and two variables for the detailed visualizations. Three sessions are under inspection and a selection is made on two variables to form a rule to indicate these sessions.

When a variable is selected from the variable list, they are moved to the top (above the search field) so the analyst can easily see which variables he is analysing. When sessions are selected to analyse for rule-making, statistics are applied on the variables and it will cause statistically relevant variables to be moved to the top of the search list.

All data elements are linked between all visualizations. When the mouse hovers over a data element in one visualization, it is highlighted in the other. For visualizations with small elements (i.e. scatter plot and parallel coordinates plot) the mouse has a larger area for hovering as it would otherwise become difficult to hover one specific element.

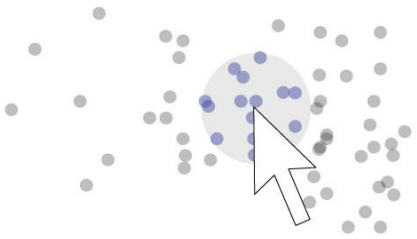


Figure 24: Larger mouse area for hovering highlights

For rule making, the analyst can fill in query operators for each variable. The selection will become apparent in the visualizations. The analyst can hereby adjust the selections either in the query operators or directly in the visualizations. The analyst can also start by “brushing” a selection (i.e. drag a square over the visualization to select, of which the start and end point of the dragging represent two corners of the square) in the visualization and adjust this again in a similar manner.

The visualizations have an optimization option by making it available to change the opacity of visual elements. The scatter plot has additional tools as it has possibilities to add a heat map (to see the distribution of dots more clearly when dots are overlapping) or move dots based on how much

they are present at a certain location. See also Figure 25. The analyst could indirectly deduct the same information when selecting this area with a brush movement and evaluating if the number of sessions selected is large.

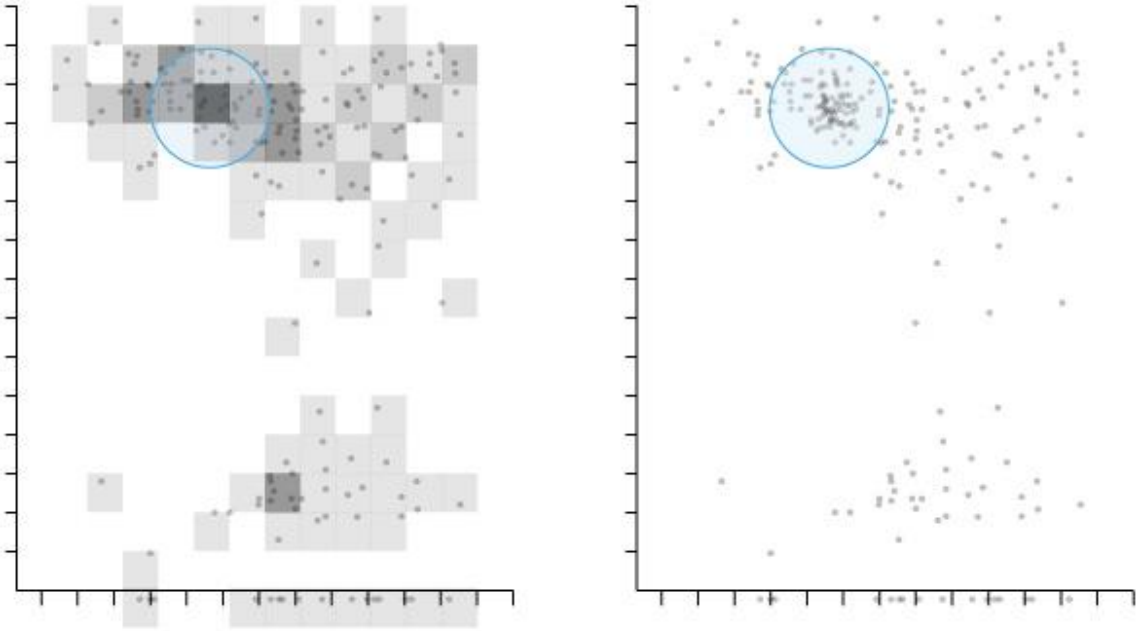


Figure 25: Adding a heat map or arranging the dots of the scatter plot differently makes it more apparent when a cluster of dots are overlapping.

An analyst could also load a current rule into the visual support tool and re-evaluating the selection and how many sessions it selects. When he also highlights corresponding fraud sessions he can evaluate the false positives it generates. The support tool will then also suggest other statistically relevant variables (by moving them to the top of the list, as explained earlier), this could allow the analyst to optimize rules.

When the analyst wants to communicate his findings to others he can export the visualizations made and will get an option to annotate them.

6.1.2 Proof of concept elements not implemented

The author did not implement zooming options. This is done deliberately as the author believes this will cause the analyst to forget the context of the data elements and make false conclusions. To find hidden data the other interaction options should be sufficient (e.g. changing of scales, adding a heat map). Also, the selection of *precise* rules can be made with the query operator fields and therefore does not need to happen visually.

Adding annotation could make the screen more cluttered. More research would be needed to find if there is great added value for this. Literature suggests implementing annotation (Thomas & Kielman, 2006; Chen, et al., 2009; Dou, et al., 2009), but during the user tests only one analyst mentions it could be a possibility. Another analyst used note-taking to write down sub conclusions during analysis. The author believes he could have done the same actions by (de-)selecting or ordering variables in the parallel coordinates plot.

The recording of actions, and the ability to go back and undo actions, has been done in other visualization tools. The author believes this is less useful here as the analysis will be cumulative. Undoing actions will mean going back in the analysis process. When a variable, that was discarded earlier, could become interesting again the analyst could simple add it back to the visualization. Saving states could be done by exporting the visualizations.

The author did not add a function to change the sample, however this could be done by simply adding a button.

6.1.3 Proof of concept in context of the analyst's workflow

The visual support tool can be used for all main ways of fraud analysis (see Figure 9 on page 39), namely to inspect interesting sessions, to analyse interesting variables or to evaluate current rules. It is very important that the interaction and selection of variables happens fluently and does not hamper the reasoning process.

The author did not implement specific links with individual sessions, because the added value of this over a more statistically focused analysis is not clear yet. Still the author believes this could be implemented with tooltips. It should also be possible to go from an individual interesting session to the visual support tool, this could be done with a hyperlink. The same should be possible for any rule-making pages.

Important design considerations were made not only based on what is perceptually right, but also how to introduce visualization to the more textually focused analysts. The author hopes an intuitive interface will help to make the transition easier because it is perceptually easy to understand. However the author believes the instructions at the beginning, and possible throughout the analysis process, are of great importance as well. How to design these instructions was aided with Technical Communication research.

6.2 Model for visualization

The model for visualization offers guidance to other visual communication designers that aim to build a visual support tool. The author believes this can be generalized to visual analysis in general and not solely for fraud analysis. The results of this chapter helped to answer research questions 1 and 2 (how can theory be integrated into a visualization model and how can this model help in designing a visual support tool?).

Based on the reflection and structuring of the theories the author tried to capture the relevance of each theory in a visualization model (see fold-out page Figure 26 on page 93). The model tries to encompass the whole spectrum of the visual support tool, its context in the analytical process and the design aspects. The model can therefore be used as a reference during the designing of a visual support tool but also to gain general knowledge on how a visual support tool can benefit analysts.

To explain the model, the author will use this research case as an example.

The author, in this case the designer (mentioned in the bottom right of Figure 26) will have different decisions to make. A great design model that could help design and validate the tool is the *model for visualization design and validation* of Tamara Munzner (2009). Here domain problems, data/operation abstractions and possible visual and interaction encodings are designed. The visualization model for designers (when the fold-out page is closed and one sees only the design and designer column) can be used to guide the design process from a highly cognitive level (where the analysts form hypotheses) to a low perceptual level (where the analysts perceive the visualizations) (see most left column).

On a slightly lower cognitive level, where the analyst will actually interact with the designed tool, the designer needs to make decisions based on a *Visual Rhetoric* and *Information design* vision. What are the elements to incorporate in the visual support tool design to benefit the analysts? This is related to any information gathered in the highest level of cognition. Finally, the designer needs to consider *Visual Rhetoric Ethics*. How could the tool *best* support the analysts?

For these questions the designer could take guidance from the theories mentioned at the left (design) column in Figure 26. For example for the domain abstraction, the designer can make use the *categories of visualization use* and reflect on the aspects in the *Human Cognition Model* that could be relevant. When designing the tool the designer should strive for an increase in *Cost-of-Knowledge* (more data elements in less time). Additional information on *codes* and *external memory* could help to find alternatives that suit the purpose.

For the designing of interactions, i.e. communication between tool and analyst, the designer could use techniques from different interaction taxonomies. Added knowledge on how the visual support tool could *amplify cognition* and on the *classes of sign* that are present could give more insight into the communication process that the communication designer wants to design.

With respect to Visual Rhetoric Ethics the designer could take guidance from multiple theories. The *visual channels of accuracy* explains the visual encoding that works best with a specific data type, or if the designer wants the analyst to find information with *separable or integral visual dimensions*. The

designer should adhere to certain reasonable guidelines and could get more insight into how visual elements are perceived by the analyst.

In general the author believes the fold-in model could be sufficient information for a communication designer that needs to design a visual support tool. However, the theories that could explain the aforementioned guidelines, concepts and theories could be found at the left side of Figure 26 (i.e. Cognitive processes and Analyst column) when you fold the page out. The added value of this for a designer is that he or she could become more aware of her influences on the analytical discourse processes between analyst, visual support tool and data.

In the middle (cognitive processes) column, theories are mentioned that explain how the perceptual elements, and interactions with these elements, are translated towards the higher cognitive level. On the top cognitive level this is shown with the loop of exploration and increasing of knowledge.

This loop of increasing knowledge with exploration also links the cognitive processes to the analyst as an individual. His knowledge will influence the process of interaction and exploration and the information he is able to extract. This is made more concrete with the *internal model of visualization exploration process*, the analysts' *visual literacy level*, and his *levels of reading* the diagrams in the visual support tool.

The designer could use the background information - of the columns analyst and cognitive processes - to evaluate certain parameters. For example, if the support tool does not work, it could be because their internal model of visualization exploration does not match the design, or that comprehension into a mental model based on the visual input does not occur. The model does not offer a precise guideline on how to solve these problems (except for the heuristics for design at the right side of the model). But knowing why the problem occurs is the first step in solving it. Furthermore, a communication designer could have other communication theories that could help with certain cognitive problems.

HIGH-LEVEL cognition

Analytical discourse

Projection (into the future)

Comprehension (into a mental model)

Logical reasoning: Abduction, Induction, Deduction

Knowledge increase

Knowledge loop

if not fully comprehended continue exploration

Knowledge

Visual exploration

Data transformations

Sensemaking loop

Problem solving strategy

Visual query construction

Pattern finding loop

Eye movement control loop

Intrasaccadic image scanning loop

Foraging loop

Observing results of diagram

Experimenting with diagram

Constructing diagram

Cost-of-knowledge

Categories of use: Monitoring, Inspecting, Exploring, Forecasting, Communicating

HCM computer-aided discovery

Codes

External memory

Nested model for visualization design and validation:
- Domain Problem Characterization
- Data/operation abstraction design
- Encoding/interaction technique design
- Algorithm design

HIGH-LEVEL cognition

Analytical discourse

Interacting with the visual support tool

Visual literacy

"A basic system of grammar for learning, recognizing, making, and understanding visual messages that are negotiable by all audiences"

Visual mapping transformations (i.e. change visual encoding)

Internal model of visualization exploration process

Three classes of action: exploration, insight, meta

Information seeking mantra

Function of graph: Recording, Communicating, Processing

Six ways visualization amplifies cognition

Ten classes of sign

Visual rhetoric

"The ability of the writer to achieve the purpose of a document through visual communication at any level"

Information design

"bring together words and images in ways that enable people to understand, take action, or make decisions".

Interacting with the visual support tool

Attention processes

Levels of reading:
- Elementary
- Intermediate
- Overall

View transformations (i.e. adjust visual elements like scales)

Data ink ratio

(low) Lie factor

Visual channels of accuracy

Gestalt principles

Visual Rhetoric Ethics

"... it concerns the time wasting inefficiency of images in conveying information that would have been conveyed more effectively ..."

Attention processes

Low level perceptual features

LOW-LEVEL perception

View transformations (i.e. adjust visual elements like opacity)

Diagrammatic reasoning with multiple diagrams (perceptual & contextual integration)

Principles of effective animation

Colour guidelines

Separable and integral visual dimensions

Feature integration theory

Low level perceptual features

LOW-LEVEL perception

Exploration actions

Output influencing exploration actions exploration actions influencing output

Embedded processes; information and attention driven

ANALYST

COGNITIVE PROCESSES

DESIGN

DESIGNER

Figure 26 : Visualization model. On the horizontal axes are the different relevant elements for the visual support tool. The vertical axes represents the level (high/low) of perception/cognition.

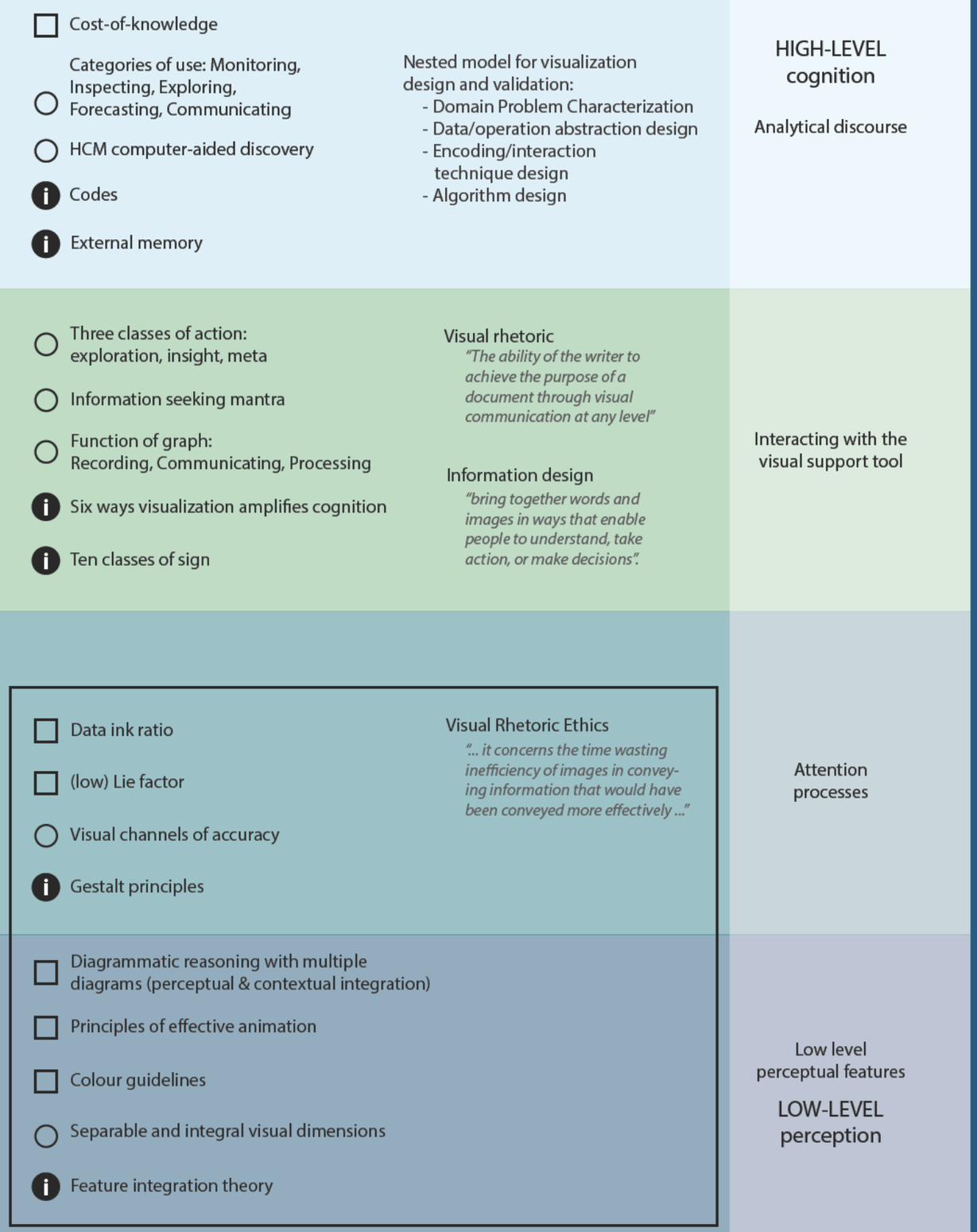


Figure 26 : Visualization model. On the horizontal axes are the different relevant elements for the visual support tool. The vertical axes represents the level (high/low) of perception/cognition.

Visualization model

What is this model?

A model for designing visual support tools for data analysis. This model is a product from the research towards visual fraud analysis to support communication between data and analyst. It includes theories from the fields of Semiotics, Technical Communication, Visual Analytics and Visual Perception.

When to use this model?

A designer could use this model at the start of a design process, to find out why problems occur with a current visual support tool design, or to evaluate a visual support tool design with the goal of improving it.

How to read this model?

The right part of the model (with the Design and Designer column titles on the bottom) is specifically for designing a new visual support tool for visual analysis. The columns represent the design, with guidelines on design elements, and the designer, with his vision on what the design should be able to do. The rows describe the levels of cognition and perception. On the top is the highest level where analytical discourse will occur, below that the analyst interacts with the visual support tool including both cognitive decisions as perceptual influences. The third level describes the attention processes where the analyst is actively searching for specific visual cues which on the lowest level should be met with perceptual features.

In each column different theories are given that help with design choices or requirement formulations. The theories that are referred to can be found in more detail in the rest of this report. The icons in the design column have different meanings:

- ☐ Means the designer should check whether this aspect is met in the visual support tool design.
- ○ Represents that the designer should make choices on those aspects.
- ⓘ Represents more specific background information on visual elements that could offer alternative designs.

For who is this model?

For visual communication designers who want to design a new visual support tool for analysis. The target audience in mind preferably has a background in the field of Science Communication.

Why this model?

Because most models found in literature focus on specific aspects for designing a visual support tool. For example only on perceptual processes, interaction, offer directions on evaluation or offer guidelines. This model aims to offer all these perspectives by integrating theory from literate and by referring to relevant guidelines from literature. It thereby offers both procedural power (*how* can a visual support tool be designed) as descriptive power (*why* does such a design aid visual analysis).

The left part of the model (with the Analyst and Cognitive processes column titles on the bottom) is for getting a better understanding of the analyst and his understanding and interaction with a visual support tool. The same rows are applied as on the design part of the model, one could thereby link design aspects with its background information on cognitive processes of the analyst. The theories again refer to theories mentioned in more detail in this report.

The left side of the model can also be used when problems occur with the implementation or adaption of the visual support tool by a target audience. Namely, there could be misalignments found with the cognitive processes.

For example when a visualization is not fully comprehended (top part of the model) the analyst will continue to go in a loop to find more visual information to make a mental model of the data. However, he would never be able to increase his knowledge without comprehending it. Or even earlier in the loop an analysts' visual literacy level (analyst column) might not be high enough to understand the diagrams that are constructed in the tool.

Another example could be that the visual queries made cannot be met by the pattern finding loop (cognitive processes column). The user could use view transformations to improve this, but when these are not available a user could not solve his visual analysis problems.

6.3 Answers to the research questions

The problem was that the current way of analysing large datasets reaches the cognitive limits of the analysts as clues become weaker. The goal therefore was to design a proof of concept of a visual analysis support tool for these analysts. The research questions were formulated as follows:

Which elements should an optimal visual support tool have to improve the efficiency of analysing large data streams for fraud analysis?

Q1. How can theory from Semiotics, Technical Communication, Visual Analytics and Visual Perception be integrated into a visualization model?
Q2. How can the visualization model help in designing a visual support tool?
Q3. Which support tool design elements can support the fraud analysts?
Q4. To what extent does the designed support tool influence the data analysis workflow?

The proof of concept and a visualization model that could aid designing a support tool in other analysis situations can be found in chapters 6.1 and 6.2. In this chapter the author reflects on the research questions.

6.3.1 Theory integrated into a visualization model

The first research question was *How can theory from Semiotics, Technical Communication, Visual Analytics and Visual Perception be integrated into a visualization model?*

The research fields were chosen based on what fields that are related to science communication are relevant (i.e. Semiotics and Technical Communication) for this research problem. Additionally the fields of Visual Analytics has done similar work and Visual Perception helps to explain the perceptual effects of visualization.

The author did not find a lot of relevant Semiotics and Technical Communication theories, see Figure 13 on page 51. The theories found in Semiotic literature often resemble concepts of which scholars of other fields made use of, to build more concrete theories or explanations (e.g. diagrammatic reasoning). Technical Communication theories found offered a way of understanding the analysts and its cognitive relation with visualizations or explain how the rhetoric of the designer could influence this.

The field of Visual Analytics alone proved too limited of a scope, that is why the author included other Information Visualization literature, but with a specific focus on analytics. The great abundance of theories found proved to be valuable on different levels (e.g. rhetorical, design, interaction, task analysis and sensemaking).

One of the reasons for including Visual Perception was because of its difference in detail compared to the others research fields (i.e. focused on specific design aspects of visual elements not only higher level concepts on how it could look). It is therefore not surprising that these theories help to understand low level perceptual aspects and to understand how a visual support tool should be designed.

The fields of Technical Communication, Information Visualization (which includes Visual Analytics) and Visual Perception would have been sufficient to grasp all relevant aspects necessary for building a visualization model. Still the author decided to also include Semiotic theories as they offer a different - more philosophical - perspective on thinking about visualization.

The overlap between theories is large both in what they strive to explain and in the constructs they use. This helped to bring the theories together and construct a coherent model from the found theories instead of from separate building blocks of these theories. The final model design can be found in chapter 6.2.

6.3.2 Visualization model in action

The second research question was *How can the visualization model help in designing a visual support tool?*

At the start of the study the plan was to *first* build a theoretical framework and *then* start to iterate. During the study it became apparent that more practical knowledge on the theories was necessary to build a model. This led to the visualization model also being iteratively refined next to the proof of concept.

The final visualization model was therefore not used by the author as it was an end product of this study. However, the reflections during each round suggests ways of using theories or models during the design. For example some theories offered encoding alternatives while others helped to evaluate the encoding.

The author believes that part of the model could help the designer with building a visual support tool while the other part could give a better understanding on how the design could influence the cognitive processes of the analyst. The design-part of the model helps to formulate a designer's rhetoric and design vision and gives guidelines on how to design visual elements. The descriptive theories, on the cognitive processes and the analyst, could be used to *understand* these design guidelines. When the design shows problems of perception or analyst adaptation of the tool, the designer could use these theories to understand why this is and make more informed decisions on improving the design.

How the visualization model could have helped the author during this study when it was known from the start, is described in chapter 6.2.

6.3.3 Visual support tool design elements

The third research question was *Which support tool design elements can support the fraud analysts?*

The main answer can be found in the chapter that shows a proof of concept of a visual support tool (See chapter 6.1). However, the author also tries to answer this question more broadly here.

Firstly, the visual encodings possible depends greatly on the data at hand. Visualizing a network of relations or banking sessions have great influence on the design elements that could be used. For example network visualizations were of no use in this study.

Secondly, the visualization and interaction encoding - and its implementation - also depends on the analyst (i.e. user). For example, other analysts might not want missing values visualized or rule-making abilities visualized, while these are very important for the fraud analysts at DetACT.

The two mutually enforcing questions of Pretorius and Van Wijk (2009) (i.e. *What does the user want to see? What does the data want to be?*) seem to describe this issue very well.

Thirdly, the author recognizes the importance of the rhetoric of the designer. In theory a designer could design a very complex and less intuitive tool if their goal is to make it more difficult for analysts. Still the rhetoric of the designer is often linked to what the user wants to see. However, sometimes a user does not know what he wants to see and the designers needs to figure this out themselves.

There could therefore be no ultimate answer to this research question, however the proof of concept offers *an* answer to this question. Still, one design element is probably necessary to be included in all future support tools, namely computer aided discovery. This is a very technical design element and the specific (algorithm) making of this was beyond the scope of this study. The author suggests how computer-aided discovery could help in the proof of concept design in chapter 6.1.

6.3.4 Visual support tool and its influence on the analysts workflow

The fourth and last research question was *To what extent does the designed support tool influence the data analysis workflow?*

The big influence that a visual support tool could give is that the analysts get "more value for their hypothesis". Currently, they either confirm or deny their hypotheses with their textual analysis. With a visual tool they can see nuances and get more information on how to adjust or change their hypothesis including multiple variables.

It also gives the analyst more time to explore instead of reading the data. It allows more cognitive power to be used for valuable insights on fraud. They also do not need to gather all the relevant data; the tool would provide a direct link with the dataset and its relevant variable.

They do not necessarily test rules by implementing them either as this testing could be done visually in the tool (e.g. show the false positive ratio by selecting a rule). If this is linked to rule-making directly, it could also allow for easy adjustments when it is necessary.

However, to find out all influences of a visual support tool the tool needs to be tested longer and be used by the analyst during day-to-day work. This could also help solve any usability aspects or other design elements that need optimizing.

6.3.5 Answer to the main research question

The main research question was *Which elements should an optimal visual support tool have to improve the efficiency of analysing large data streams for fraud analysis?*

The answer to this question is not clear-cut. There can be multiple designs based on the analysis, data, analysts and rhetoric of a designer. The proof of concept (in chapter 6.1) offers one of the answers that could be given to this question. To design other visual support tools in other analysis situations the visualization model (chapter 6.2) based on theory from different fields will be useful. To find out the precise influence of a visual support tool, it needs to be tested extensively and for a long period of time.

Preface 3

Executive summary 5

List of Figures 10

List of Tables 12

Glossary 13

1 Introduction 15

2 Project description 19

3 Methods 27

4 Orientation 35

5 Iteration 59

6 Conclusion 85

7 Discussion 99

7.1 Visual support tool 99

7.1.1 Implementation 100

7.1.2 Existing solutions 100

7.1.3 Compared to other types of fraud analysis 100

7.2 Visualization model 101

7.2.1 Influence of the author 101

7.2.2 Number of theories included 101

7.2.3 Compared to other visualization models 102

7.2.4 Future work for the visualization model 103

7.3 Science Communication 104

7.3.1 Science Communication is important in data visualization 104

7.3.2 Research methodology reflection 105

7.3.3 Future of Science Communication and data visualization 106

7.4 What would I do differently? 107

References 108

Appendices 115

7 Discussion

The author reflects upon her research results (see chapter 5 and 6) that helped to design a proof of concept for a visual support tool for fraud analysis (see chapter 7.1) and the visualization model for designing other visual support tools (see chapter 7.2). Additionally, a reflection on how the work affects the field of Science Communication and how Science Communication could affect visualization design (see chapter 7.3). Finally, the author describes some aspects which could have been done differently (see chapter 7.4).

7.1 Visual support tool

The visual support tool has proven that a visual way of fraud analysis is feasible. It has an added value over textual analysis as it gives the ability to *see* a solution within the data, rather than reading it.

In textual analysis, when the analysts analyse the dataset, they make hypotheses based on intuition “this seems out of the ordinary”. They can support their hypotheses by testing their rules on the database and rely on those numbers “this generates 10.000 false positives a day, which is not acceptable”. The analyst is not offered guidance on how to change his hypotheses or how big an influence a slightly different selection would make. It could also occur that the analyst is not searching for alternative hypotheses and is only looking to confirm his intuition (i.e. confirmation bias). In short, he does not generate more insight in the data than he already has, his hypothesis is either confirmed or rejected.

Visualization helps the analyst to easily change perspective or combine perspectives. Each variable selection will show a slight subsection of the complete dataset. Making it easy to combine and change these slices, which makes it easier and maybe even more fun to browse through large sets of variables in search of the one that could help him best. The cost of getting more into the data as a whole is that one cannot immediately extract very detailed information of the individual element. This is very different than what the analysts are used to with their textual datasets.

The making of rules is becoming more detailed as fraud becomes more complex and show weaker signals. The author believes that this is important to be aware of during the design of a visual analysis tool and include as much detailed information as possible. Still, if the signals become weaker and the rules will need to be more detailed, will visual analysis that lacks precise numbers suffice? The author believes it will.

The analysts can make more *complex* rules because they can *combine* variables and their weak signals. Off course it is easier to see distinct abnormalities, which could help with easier fraud types with stronger signals. This leaves analysis time for the more complex fraud types. The author believes that the way to cause that the less distinct abnormalities also grab the analysts’ attention is to combine visualization with computer-aided discovery from data mining techniques. This is recommended in literature when the dataset becomes too large, which is definitely the case in fraud analysis.

7.1.1 Implementation

It is important to not hamper the rhythm of reasoning, still a lot of preparation will be needed for the implementation. For example a system should be put in place to gather all data variables needed. Additionally, the analysts need to keep up their labelling of data variables and elements, either automatically or by hand. This is necessary to not question the truthfulness of the visual elements that represent them. Imagine a variable to be ambiguously labelled, the analyst could misinterpret and distrust the visual results.

It is also very important to let an expert on data mining help in building the statistical rules. To make it all work together without "hampering the rhythm of reasoning", the author suggests a professional programmer to develop a final prototype to see if it is feasible. This could be a challenge as it takes smart programming and great computation power. But because computers and their technology are being made faster and more powerful the author believes that it will be possible. If not now, than very soon.

7.1.2 Existing solutions

A question asked earlier in this report was whether or not the analysts could make use of already existing tools. The author is aware of multiple tools that could aid in similar visual analysis. Still the parallel coordinates plot, which is key for generating an overview in this proof of concept, is not popular enough to be part of these software packages. Which comes a bit as a surprise since within data visualization they are used more and more. Overview visualizations on network or hierarchical data are available.

More statistically focused software (e.g. (The R Foundation, n.d.)) often requires a large knowledge on statistics, software functionalities or programming skills and do not offer easy integration of visualizations or overviews. They are therefore less accessible for non-expert users. As the author tried to make the tool as intuitive as possible the author believes it is also more accessible for non-expert analysts.

The possibilities for integrating existing products with the current analysis system are often non-existent which makes it less easy to implement them in the analysts workflow. The author is not aware of commercial solutions that have both the visual functionalities (overview and detailed views) and data mining functionalities for non-network data and also have the option to be integrated.

7.1.3 Compared to other types of fraud analysis

The proof of concept could be used for other fraud analysis, though not all fraud analysis have a need for rule-making. Some fraud analysts will need to find the criminal who is behind the fraud or figure out how to improve the infrastructure of the system to prevent fraud. Still they will need to indicate what is happening with the fraud to figure out the answers to these questions. In this tool this can be specifically formulated in the form of a rule, but does not necessarily need to happen.

The author had the ability to use a dataset of another department within Fox-IT as input for the tool. Here the goal was also not rule-making but seeing why two variables do not align. Because this is about relations both the parallel coordinates plot (overview visualization) and the scatter plot proved to be helpful.

The author does recognize the limitations of this design for monitoring real-time data. An analyst has to select variables for inspection, so constant monitoring of the whole dataset is not possible.

7.2 Visualization model

The model hopes to give guidance to other visual communication designers. Still intersemiotic translation between any signs other than the verbal is very subjective. When another designer would use this model he could therefore end up with very different results from the author. Imagine if the designer had no knowledge of the parallel coordinates plot. Would he still arrive at the same conclusions? The author likes to think he would, as other elements in the model emphasize aspects as overview and exploration, but these terms could mean different things for different people from different backgrounds.

7.2.1 Influence of the author

The visualization model is based on a selection of theories. What if other theories were found? Would the author arrive at a similar model? The influence of the author can also be found in the reflection on the theories on the prototypes. The author developed measurements of each theory chosen based on what the originator dictated or, if this was not present, what the author inferred the originator would have meant with the theory.

The author tried to combine the theories found in a more complete view on this analytical process. During the literature research the author did not find such a comprehensive view, combining heuristics, low-level perceptual understanding and high cognitive processes. Still, the author believes a lot of specialists from all research fields could add to the model to make its foundation and linking between theories more clear, even though the author did not find of some sorts.

7.2.2 Number of theories included

With respect to theories from the field of Technical Communication and Semiotics, the author was surprised by the lack of information visualization within Technical Communication. Information Visualization is, when designed properly, in essence a way of “providing the right information, in the right way, at the right time to make someone’s life easier and more productive“ (description of Technical Communication by (Society for Technical Communication, n.d.)).

The field of Visual Analytics offered surprisingly little theories or methodologies concerning analytics. Most of the found papers still originated from the Information visualization community. The field of Visual Analytics did show a lot of case studies. The field of Visual Perception offered more concrete guidance on design and evaluation. Both Visual Perception, Technical communication and Semiotics offered background information on why certain design choices could be made.

Even though the author benefitted from the fact that the author had different perspectives on the problem, the author now knows that it could be better to choose fewer theories. Get deeper knowledge into the ones chosen and use those as focus within the project. That way you would get better insights into the relevance of those theories, whether it is relevant or not. The downside is that you do not have a fallback when your theories do not apply.

By focusing in on fewer theories the author might have been able to grasp the construct of each theory better. The final steps towards the visualization model were hard steps to make, as the author did not only wanted them to be separate building blocks, but wanted them to be integrated. How do they overlap, connect, relate. To get to this point the author had to take a step back and intuitively

reflect on what was learned from those theories. Still when the author would have had more time to make the measurements of the theories more detailed, the result could have been made more easily.

Choosing fewer theories would also have helped the author with phases of research information overload. The author had a hard time trying to grasp the expert knowledge on tasks of the analysts. Also there was a phase of acceptance that there was no time to fully research all theories as thoroughly as the author would have liked. The author still aimed for it, which sometimes resulted in loosing oneself in the details.

7.2.3 Compared to other visualization models

The author had several visualization models that were used in her theory reflection namely: The nested model, Van Wijk visualization model, reference model and visual problem solving. Each had a different perspective. The model proposed by the author, made use of knowledge of each model but also additional theories for the purpose of building a framework that could stretch over different levels and research fields. The difference with other models is that those do not propose guidance on why to design the visualizations in a certain way. The author, and likely another communication designer, needed more insight into what visualizations effect the user in what way and how this can be used for a visual support tool than current models offer.

The theories on cognitive processes concerning visualizations (i.e. diagrammatic reasoning) or the more perceptual processes (i.e. two neural processes) allowed the author better insight into the effect of the visualization design on the analyst. Implementing these in the model could therefore improve the question *why* certain designs work better than others.

During the research Meyer et al. (2013) published their Nested Blocks and Guidelines model (NBGM) , which extends the four-level nested model. The author recognizes the importance of the model for structuring design decisions, however, it still does not offer more insight into why certain visualizations could work better than others and depends on the designer to find his own guidelines for design.

The visualization model proposed by the author has both procedural power in a way that it can guide a communication designer in a similar situation, as descriptive power in that it should help to explain how this comes to be.

The NBGM uses Beaudouin-Lafon's three ways to assess the power of a model: descriptive (ability to describe existing interfaces), evaluative(ability to asses alternatives) and generative(ability to help create new designs). This model can be considered descriptive in a sense that the design column could describe other designs, however it is more helpful as evaluative and generative power. The author believes especially the background information (left columns) help to generate insight into the *why* of the model.

The NBGM is most helpful to describe other visualizations but does provide a precise answer as to why their model works. Besides from the work referenced to the nested blocks model itself (original version) there is no argumentation why the blocks are as they are. They are most likely formed from the researcher's great experience and insights into visualization design. The visualization model presented in this research hopes to explain why certain theories to design a visual support tool (right part of the model) work as they do.

7.2.4 Future work for the visualization model

The author recognizes that when new theories appear this model could become outdated. Still the theories chosen now will likely still be relevant and offer an introduction into the different aspects during the designing of a visual support tool. The author therefore emphasizes each visualization communication designer, or other scholars, to add to this model with other theories that might be more focused on one specific aspect. In particular the author hopes more interaction theories (e.g. moving elements) concerning perception would be helpful to complete the picture.

The author believes the model could change with the addition of new insights but the structure of low level, high level and different actors within visualization design will stay.

7.3 Science Communication

At the beginning of this project the question was asked, whether or not this project would fit in with science communication. The author and the supervisors did not question this, but were not able to articulate easily why it is or should it be part of Science communication. During the research the author became more convinced that this is definitely the case.

7.3.1 Science Communication is important in data visualization

For starters, the sender is not passive in the designing of a visual support tool. In common Communication models there is a *sender* who sends a *message* to a *receiver*. The first question that came up concerning this research is that the message could be the visual support tool, the receiver is the analyst, but who is the sender? And does he have an active role? It seemed in the beginning as if making an objective and clear tool does not allow rhetoric but the opposite is true. Views on Visual rhetoric, i.e. achieve the purpose with the design, and visual rhetoric ethics, i.e. do not waste time of the user, proof how important a communication designer can be for information visualization.

Secondly, the field of communication has left it up to scholars of the field of Computer Science to design communication and interaction processes between humans and computers. Data visualization is a great example of their work. Still, critics have mentioned their lack in usability and user testing. Because the field of communication has more knowledge on the user and its cognitive processes the author would fit in well to solve usability problems.

Based on the author's experiences at the MSc Science Education and Communication at the Delft University of Technology, the author thinks most science communication methodologies and theories stick to traditional communication means where the sender has almost direct contact with the receiver and where the receiver has less options to interact (e.g. in written form, oral presentations, movies or commercials). There are two-way communication theories but often the sender is physically present so the receiver could interact with him or her personally. However there are other possibilities where a science communicator could have a virtual presence. This could be considered as the sender taking a passive role, but it is just as important. Smart products (e.g. Smartphones) are becoming a bigger role in our life and how a user interacts and communicates with these seemingly inanimate objects could be enhanced with knowledge of the Science Communication field.

The author was not able to get a lot of relevant theoretical support from the field of communication. The final result therefore relied mostly on heuristics, guidelines or theories from the fields of visualization and perception. Still the author believes diagrammatic reasoning, originally proposed by Semiotician C.S. Peirce, could be revived and used as a basis for visual communication theories. The field of semiotics seems outdated in a sense that it does not provide a lot of studies on how (visual) signs on computers are interpreted. On the other hand, the field of Technical Communication tries to design solutions for these modern problems. Still, they are more textual than visual focused for which the author believes they are leaving a gap in their research.

To continue working on visual communication and visual analysis support in the field of science communication, scholars of different fields should work together. For example, in this research the author has a background in Industrial Design Engineering, the supervisors come from the fields of Science Communication, Perception and Statistics and the author consulted experts from the field of

information visualization. By bringing together knowledge from different fields a more clear view on the problem could be formed.

So can Computer Science scholars be too technologically focused, this is reflected in the low number of evaluation papers focused on visualization technique (user experience, performance and algorithm) and the lack of papers on the data analysis process evaluation. But they offer a lot of knowledge on how to solve difficult case studies with visualization techniques. They could enhance their work to ask other scholars (e.g. science communication, perception) to aid in building a visualization tool that actually helps the user with his data analysis.

Science Communication at the Delft University of Technology strives to research and design communication processes and products that are both founded on research but are also practical. The author therefore believes, in a world with more data than ever, science communication should enter the world of “big data”. Because of the lack of experience in the field they could collaborate with Information Visualization scholars or even enhance the team of Visual Analytics! The author recommends including the classical theories on signs and diagrammatic reasoning. Even though the examples of diagrammatic reasoning given often refer back to Plato or other classical figures, diagrammatic reasoning could still be a very relevant concept within information visualization that tries to define the analytical process of analysts. In that sense, Information Visualization scholars should also seek the company of (Science) Communicators to enhance research in the Visual Analytics field.

7.3.2 Research methodology reflection

The main research method used was Design-Based Research (DBR) where the researcher brings together practitioners and existing principles from literature to develop a theory-based solution to practical problems. Because publications on DBR did not offer very practical guidance as to how to bring practice and theory together the author used the basic steps (i.e. analysis, iteration and proposition of new principles and its reflection) as a template. This is reflected in the chapters of this report (i.e. analysis step in the orientation chapter, iteration in the iteration chapter and in the conclusion a proof of concept and a new model as principle is presented and reflected upon in the discussion). The author chose other methodologies as practical guidance and chose to do three iterative cycles to gather the most feedback but keep it realistic for the time available.

The author made use of Cognitive Task Analysis, from the field of Visual Analytics, did systematic literature research, applied design methodology from the field of industrial design engineering and information visualization, performed field observations and made storyboards to communicate results. The relevance of these methodologies is proven in the way it all helped to contribute to answers on the research questions. Specifically the systematic literature research and design methodology helped guide the research.

The downside to design research, and also Design-Based research, is that replication of the research is largely impossible (The Design-Based Research Collective, 2003). Could you therefore say the research is unreliable? The author believes not. This proof of concept and visualization model offers a solution and has never stated that it is the only solution.

The aim of this research was to support the analysts with the best possible design elements in the form of a visual support tool. The initial design elements chosen or the execution of these

elements in the prototypes could have greatly influence the results. However, the author tried to keep the styling as minimum as possible (e.g. neutral colours) and tried to structure the visualization and interaction encoding choices by making use of a database that was made based on similar work and knowledge from the visualization community. Additionally, the Cognitive Task Analysis helped to triangulate the task analysis which was used as initial formulation of practical problems.

As stated before the results offer *a* solution to the research problem and offered valid answers to the research questions. Still both research question 2 (how van the visualization model help in designing a visual support tool) and research question 4 (to what extent does the designed support tool influence the data analysis workflow) could have benefitted from a longer study to answer these questions to the full extent.

The literature study performed showed the author the limited scope of the research fields chosen at the start of the research. Semiotics, Technical Communication and Visual Analytics did not offer a great number of theoretical principles. The author therefore broadened the search to include journals from other, related fields as well (e.g. Information Visualization journals relate to the field of Visual Analytics as it resides in Information Visualization). There was one field the author wished she would have included in the research and that is interaction related fields (e.g. Human Computer Interaction). There were not a lot of low-level perceptual theories on interaction and those fields could have offered more insight into interaction design.

Design Based Research is still making its entry in the field of science communication. The author hopes more methodology for DBR will become available when more scholars practice it. For now the author was very happy to benefit from methodology from other relating fields (i.e. design, Information visualization). The author therefore hopes this research could also contribute as an example of how to use mixed research methods for DBR.

There was a noticeable tension between communication and design research. The author incorporated a lot of theories during the reflection on the prototypes but a lot did not reside from the field of communication and are more design-focused (This is also visible in the “design”-side of the visualization model in **Error! Reference source not found.**). There also needed to be a balance between spending time on building prototypes, i.e. interventions for Design-Based Research, and actually testing them and using that results to improve on existing principles. The author hoped to have found that balance by giving a lot of attention to the theoretical value of this project. However as probably most designers want, the author would have liked to continue testing and improving the prototype.

7.3.3 Future of Science Communication and data visualization

More often there will be situations where a receiver needs to handle raw data, be it companies that need to analyse their sales data or is it an individual that wants to know which political party he or she needs to vote for. Science Communicators have the ability to communicate scientific and complex subjects to a broad audience very well. A lot of technical fields, e.g. economics, statistics, mathematics, have already done work on making sense of an abundance of data, however they do not always offer the right solutions to communicate this to a target audience (either in an interactive tool or with a static image). When Science Communicators would enter the market of big data they could have a key role in designing a user-friendly but also an appropriate solution to information overload of this century.

7.4 What would I do differently?

During the research the author stumbled upon practical issues (e.g. no recording of sensitive user test information) that the author would like to have done differently but was not possible because of confidentiality issues. But there are also other steps that the author might have done differently in retrospect. The three iterations proved very necessary to both zoom in to a solution and grasp the broad scope of a visual support tool design. However, the importance of real scenarios, next to real data, became apparent during the first iteration. Additionally, this knowledge could have resulted in the including of other visualization techniques, as most of the discarding of visualization techniques was done during the first round.

The author also wished to have had more possibilities to quickly develop prototypes. This could have contributed to more options for user testing or allowing more experimentation with visualization and interaction encodings.

In retrospect the author recommends a more specific view on the case. The author used in total 36 theories for reflection and the making of a model. This could be limited to 10 more valuable theories and gave a better focus and allowed more in-depth understanding of each theory. Still, the author does not regret the broad approach as this allowed her to revisit all facets of visual communication and gave her great insight into the fact that there is still a lot to research and to win when it comes to information visualization and its influence on analytics.

The author would have also liked to research more relevant design and evaluation methodologies *outside* of the visualization community. It is not surprising that those are the most relevant methods, but trying out methodologies of the other fields could have led to more insights in what those research field could offer for a similar visual analysis case. Now, a communication designer either has to do research into this or follow the visualization model the author proposes.

Overall the author believes the research contributed to a proof of concept on visual data analysis, showing that it is possible but also feasible from an analytical perspective. The author hoped to have found a way to motivate the analysts by using intuitive designs but also showing, with the help of instructions, how it can benefit them. For the Science Communication community the author hopes to have given an example on how one could apply Design-Based Research and emphasized, like others have, the need for a more solid methodological foundation. Additionally, the author hopes the visualization model could help communication designers in similar cases while still giving pointers where the model could be improved. Thereby, the author offers visual analysis options for analysing large datasets to improve the efficiency of data analysis.

References

Aish, G., n.d.. *Using data visualization to find insights in data - The Data Journalism Handbook.* [Online]
Available at: http://datajournalismhandbook.org/1.0/en/understanding_data_7.html
[Accessed 16 September 2013].

Albers, M. J. & Mazur, B., 2003. *Content & Complexity, Information Design in Technical Communication.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Amare, N. & Manning, A., 2007. The Language of Visuals: Text + Graphics = Visual Rhetoric. *IEEE Transactions on Professional Communication,* 50(1), pp. 57-70.

Amiel, T. & Reeves, T., 2008. Design-Based Research and Educational Technology: Rethinking and the Research Agenda. *Educational Technology & Society,* 11(4), pp. 29-40.

Arnheim, R., 1969. *Visual Thinking.* Berkeley and Los Angeles: University of California Press.

Bakker, A., 2007. Diagrammatic Reasoning and Hypostatic Abstraction in Statistics Education. *Semiotica,* Volume 164, pp. 9-29.

Bertin, J., 1983. *Semiology of Graphics, diagrams networks maps.* Redlands, California: Esri Press.

Boeriis, M. & Holsanova, J., 2012. Tracking visual segmentation: connecting semiotic and cognitive perspectives. *Visual communication,* 11(3), pp. 259-281.

Borkin, M. A., Vo, A., Bylinskii, Z. & Isola, P., 2013. What makes a visualization memorable. *IEEE Transactions on Visualization and Computer Graphics,* 19(12), pp. 2306-2315.

Boticki, I., Wong, L. H. & Looi, C.-K., 2013. Designing technology for content independent collaborative mobile learning. *IEEE transactions on learning technologies,* 6(1).

Cai, G., 2007. Formalizing Analytical Discourse in Visual Analytics. *IEEE Symposium on Visual Analytics Science and Technology, VAST'07,* p. 217–218.

Card, S. K., Mackinlay, J. D. & Shneiderman, B., 1999. *Readings in Information Visualization Using Vision to Think.* San Francisco, California: Morgan Kaufmann Publishers, Inc..

Chen, Y., Yang, J. & Ribarsky, W., 2009. *Toward Effective Insight management in Visual Analytics systems.* Beijing, IEEE Pacific Visualization Symposium. PacificVis '09., pp. 49-56.

Clark, R. E. et al., 2008. Cognitive Task Analysis. In: *Handbook of research on educational communications and technology (3rd edition).* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 577-594.

Cleveland, W. S. & McGill, R., 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association,* 79(387), pp. 531-554.

Crandall, B. W. & Hoffman, R. R., 2013. Cognitive Task Analysis. In: *The Oxford Handbook of Cognitive Engineering.* New York: Oxford University Press, pp. 229-239.

D'Amico, A. & Kocka, M., 2005. Information assurance visualizations for specific stages of situational awareness and intended uses: lessons learned. *IEEE Workshop on Visualization for Computer Security (VizSec 05).*

Dasgupta, A. & Kosara, R., 2010. Paragnostics: Screen-Space Metrics for Parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics,* 16(6), pp. 1017-1026.

Deloitte, 2013. *uk-ers-digital-fraud.pdf.* [Online]
Available at: http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Services/ERS/uk-ers-digital-fraud.pdf
[Accessed 06 March 2014].

Demšar, U., 2007. Investigating visual exploration of geospatial data: an exploratory usability experiment for visual data mining. *Computers, environment and urban systems,* 31(5), pp. 551-571.

Design-Based Research Collective, n.d. *Design-Based Research Collective.* [Online]
Available at: http://www.designbasedresearch.org/dbr.html
[Accessed 11 February 2014].

Desnoyers, L., 2011. Toward a Taxonomy of Visuals in Science Communication. *Technical Communication,* 58(2), pp. 119-134.

Doumont, J.-L., 2004. Verbal versus visual: A word is worth a thousand pictures, too. *Technical Communication,* 34(1), pp. 31-42.

Doumont, J.-L., 2009. *Trees, Maps and Theorems: Effective Communication for Rational Minds.* Belgium: Pricipiæ.

Dou, W., Jeong, D. H., Stukes, F. & Ribarsky, W., 2009. Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications,* 29(3), pp. 52-61.

Dubinsky, J., Paretti, M. C. & Armstrong, M., 2001. *Technical Writing: Document Design: Charts and graphs.* [Online]
Available at: http://wiz.cath.vt.edu/tw/TechnicalWriting/docdesign/charts.htm
[Accessed 20 December 2013].

Erbacher, R. F. et al., 2010. A multi-phase network situational awareness cognitive task analysis. *Information visualization,* 9(3), pp. 204-219.

Fekete, J.-D., Wijk, J. J., Stasko, J. T. & North, C., 2008. The value of information visualization. *Information Visualization,* pp. 1-18.

Ferreira de Oliveira, M. C. & Levkowitz, H., 2003. From visual data exploration to visual data mining: a survey. *IEEE transactions on visualization and computer graphics,* 9(3), pp. 378-394.

Fisher, D., Popov, I., Drucker, S. M. & Schraefel, M. C., 2012. Trust me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster. *Proceedings of the 2012 Conference on Human Factors in Computing Systems (CHI 2012).*

Gleicher, M. et al., 2011. Visual Comparison for Information Visualization. *Information Visualization,* 10(4), pp. 289-309.

Goodall, J. R., 2008. *Introduction to visualization for computer security.* s.l., Springer Berlin Heidelberg, pp. 1-17.

Gotz, D. & Zhou, M. X., 2009. Characterizing users'visual analytic activity for insight provenance. *Information Visualization,* 8(1), pp. 42-55.

Green, T. M., Ribarsky, W. & Fisher, B., 2008. Visual analytics for complex concepts using a human cognition model. *IEEE Symposium on Visual Analytics Science and Technology,* pp. 91-98.

Harrison, C., 2003. Visual social semiotics: Understanding how still images make meaning. *Technical Communication,* 50(1), pp. 46-60.

Hart-Davidson, W., 2001. On writing, technical communication, and information technology: the core competencies of technical communication. *Technical communication,* 48(2), pp. 145-155.

Heer, J. & Agrawala, M., 2006. Multi-scale banking to 45 degrees. *IEEE transactions on Visualization and Computer Graphics,* 12(5), pp. 1-8.

Heer, J., Bostock, M. & Ogievetsky, V., 2010. A tour through the visualization zoo. *Communications of the ACM,* 53(6), pp. 56-67.

Heer, J. & Robertson, G. G., 2007. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics,* 13(6), pp. 1240-1247.

Heer, J. & Shneiderman, B., 2012. Interactive Dynamics for Visual Analysis. *Communications of the ACM,* 55(4), pp. 45-54.

Hoffmann, M. H., 2007a. Cognitive conditions of diagrammatic reasoning. *Semiotica-Special Issue on Peircean Diagrammatical Logic.*

Hoffmann, M. H., 2007b. Seeing problems, seeing solutions. Abduction and diagrammatic reasoning in a theory of scientific discovery. In: *Abduction and the process of scientific discovery.* s.l.:s.n., pp. 213-236.

Horváth, I., 2008. Differences between 'research in design context' and 'design inclusive research' in the domain of industrial design engineering. *Journal of Design Research,* 7(1), pp. 61-83.

infodesignpatterns.com, n.d.. *patterns | infodesignpatterns.com.* [Online]
Available at: http://www.niceone.org/infodesignpatterns/index.php5#/patterns.php5
[Accessed 16 September 2013].

Johnson-Sheehan, R. & Baehr, C., 2001. Visual-spatial thinking in hypertexts. *Technical Communication,* 48(1), pp. 22-30.

Johnson-Sheehan, R. & Baehr, C., 2001. Visual-spatial thinking in hypertexts. *Technical Communication,* 48(1), pp. 22-30.

Johnston, R. R., 2012. Graphic trinities: languages, literature, and words-in-pictures in Shaun Tan's the arrival. *Visual Communication,* 11(4), pp. 421-441.

Kelleher, C. & Wagener, T., 2011. Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software,* Volume 26, pp. 822-827.

Kim, J. & Hahn, J., 1997. *Reasoning with Multiple Diagrams: Focusing on the cognitive integration process.* s.l., s.n., pp. 376-381.

Kim, J., Hahn, J. & Hahn, H., 2000. How do we understand a system with (so) many diagrams? Cognitive Integration Processes in Diagrammatic Reasoning. *Information Systems Research,* 11(3), pp. 284-303.

Kolko, J., 2010. Abductive thinking and sensemaking: The drivers of design synthesis. *Design Issues,* 26(1), pp. 15-28.

Kostelnick, C., 2008. The visual rhetoric of data displays: The conundrum of clarity*. *IEEE Transactions on Professional Communication,* 51(1), pp. 116-130.

Lam, H. et al., 2012. Empirical Studies in Information Visualization: seven scenarios. *IEEE Transactions on Visualization and Computer Graphics,* 18(9), pp. 1520-1536.

Lift Conference, 2012. *Perspective on information overload by Anaïs Saint-Jude - Lift Conference.* [Online]
Available at: http://videos.liftconference.com/video/4622173/perspective-on-information
[Accessed 30 12 2013].

Littlejohn, D. S. W., 2011. *Theories of Human Communication.* 10 ed. s.l.:Waveland press.

Liu, Z., Lee, B., Kandula, S. & Mahajan, R., 2010. Netclinig: Interactive visualization to enhance automated fault diagnosis in enterprise networks. *IEEE Symposium onVisual Analytics Science and Technology (VAST),* pp. 131-138.

Lundberg, G. C., 2000. Made sense and remembered sense: Sensemaking through abduction. *Journal of Economic Psychology,* 21(6), pp. 691-709.

Mackinlay, J., 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics,* 5(2), pp. 110-141.

Manning, A., 2002. The theory and practice of thinking outside the box: interfacing abstractions with practical communication puzzles. *IEEE International Professional Communication Conference,* pp. 544-551.

Manning, A. & Amare, N., 2006. Visual-rhetoric ethics: Beyond accuracy and injury. *Technical Communication,* 53(2), pp. 195-211.

Markel, M., 1998. Testing visual-based modules for teaching writing. *Technical Communnication,* 45(1), pp. 47-76.

Meyer, M., Sedlmair, M., Quinan, P. S. & Munzner, T., 2013. The nested blocks and guidelines model. *Journal of Information Visualization, Special Issue on Evaluation.*

Mirel, B., 1998. Visualizations for Data exploration and Analysis: A Critical Review of Usability Research. *Technical Communication,* pp. 491-509.

Moriarty, S. E., 1996. Abduction: A theory of visual interpretation. *Communication Theory,* 6(2), pp. 167-187.

Moszkowicz, J., 2011. Gestalt and Graphic Design: an Exploration of the Humanistic and Therapeutic effects of Visual Organization. *Design Issues 27,* Volume 4, pp. 56-67.

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

Munzner, T., 2009. Visualization. In: *Fundamentals of Computer Graphics, third edition.* s.l.:CRC Press, pp. 675-707.

Murch, G. M., 1973. *Visual and auditory perception.* USA: The Bobbs-Merrill Company.

Nederlandse Vereniging van Banken, 2013. *scherpe daling fraude internetbankieren - Nederlandse Vereniging van Banken.* [Online]
Available at: http://www.nvb.nl/nieuws/2013/1812/scherpe-daling-fraude-internetbankieren.html
[Accessed 06 March 2014].

Nederlandse Vereniging van Banken, 2014. *Fraude betalingsverkeer opnieuw flink gedaald - Nederlandse Vereniging van Banken.* [Online]
Available at: http://www.nvb.nl/nieuws/2014/2636/fraude-betalingsverkeer-opnieuw-flink-gedaald.html
[Accessed 11 March 2014].

Nederlandse Vereniging van Banken, n.d.. *Internetbankieren: Wat is de omvang van internetbankieren in Nederland? - Nederlandse Vereniging van Banken.* [Online]
Available at: http://www.nvb.nl/veelgestelde-vragen/1335/internetbankieren-wat-is-de-omvang-van-internetbankieren-in-nederland.html
[Accessed 20 March 2014].

Neurath, O., 1936. *International Picture Language: the first rules of isotype.* London: Kegan Paul, Trench, Trubner & co.

Nielsen, J., 2004. *risks of quantitative studies.* [Online]
Available at: http://www.nngroup.com/articles/risks-of-quantitative-studies/
[Accessed 10 December 2013].

O'Regan, J. K., 1992. Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology,* 46(3), pp. 461-488.

Pike, W. A., Stasko, J., Chang, R. & O'Connel, T. A., 2009. The science of interaction. *Information Visualization,* 8(4), pp. 263-274.

Pirolli, P. & Card, S., 2005. The sensemaking process and leverage points for analysts technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis,* Volume 5, pp. 2-4.

Portewig, T. C., 2004. Making sense of the visual in technical communication: A visual literacy approach to pedagogy. *Journal of technical writing and communication,* 34(1), pp. 31-42.

Pretorius, A. J. & van Wijk, J. J., 2009. What does the user want to see? What does the data want to be?. *Information visualization,* 8(3), pp. 153-166.

Ribarsky, W., Fisher, B. & Pottenger, W. M., 2009. Science of Analytical reasoning. *Information Visualization,* 8(4), pp. 254-262.

Sagepub, n.d.. *Microsoft Word - Call for papers_2013 - Special_CPF_2013.pdf.* [Online]
Available at: http://jbt.sagepub.com/site/includefiles/Special_CFP_2013.pdf
[Accessed 10 December 2013].

Schriver, K., 2003. Foreword. In: *Content and Complexity, Information Design in Technical Communication.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc., pp. ix-xii.

Sebeok, T. A., 2001. *Signs, introduction to Semiotics, second edition.* Toronto: University of Toronto Press Incorporated.

Sedlmair, M., Meyer, M. & Munzner, T., 2012. Design Study Methodlogy: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis),* 18(12), pp. 2431-2440.

Shneiderman, B., 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. *IEEE symposium on Visual Languages,* pp. 336-343.

Shneiderman, B., 2007. Creativity Support Tools. *Communications of the ACM,* 50(12), pp. 20-32.

Siau, K. & Tan, X., 2006. Cognitive mapping techniques for user-database interaction. *IEEE Transactions on Professional Communication,* 49(2), pp. 96-108.

Society for Technical Communication, 2013. *Sessions at Society for Technical Communication SUMMIT 2013 about Data Visualization.* [Online]
Available at: http://lanyrd.com/2013/society-for-technical-communication-summit/on/data-visualization/
[Accessed 26 December 2013].

Society for Technical Communication, n.d.. *Defining Technical Communication.* [Online]
Available at: http://www.stc.org/about-stc/the-profession-all-about-technical-communication/defining-tc
[Accessed 2 September 2013].

Somov, Y. G., 2012. Codes, heterogeneities, and structures: Visual information and visual art. *Semiotica,* Volume 192, pp. 219-233.

Starcic, A. I., Cotic, M. & Zajc, M., 2013. Design-based research on the ude of tangible user interface for geometry teaching in an inclusive classroom. *British journal of educational technology,* 44(5), pp. 729-744.

Statistics Netherlands, 2012. *CBS - Nederland in Europese top met internetbankieren.* [Online] Available at: http://www.cbs.nl/nl-NL/menu/themas/vrije-tijd-cultuur/publicaties/artikelen/archief/2012/2012-3662-wm.htm [Accessed 11 February 2014].

Stroupe, C., 2000. Visualizing English: Recognizing the hybrid literacy and verbal authorship on the web. *College English,* 62(5), pp. 607-632.

The Design-Based Research Collective, 2003. Design-Based Research: An emerging paradigm for educational enquire. *Educational Researcher,* 32(1), pp. 5-8.

The Guardian, 2011. *What's a zettabyte? By 2015, the internet will know, says Cisco | Technology | theguardian.com.* [Online] Available at: http://www.theguardian.com/technology/blog/2011/jun/29/zettabyte-data-internet-cisco. [Accessed 30 December 2013].

The R Foundation, n.d.. *The R Project for Statistical Computing.* [Online] Available at: http://www.r-project.org/ [Accessed 15 December 2013].

Theus, M. & Urbanek, S., 2009. *Interactive Graphics for Data Analysis.* Boca Ration, Fl.: CRC Press, Taylor & Francis Group.

Thomas, J. J. & Cook, K. A., 2005. *Illuminating the Path, The Research and Development Agenda for Visual Analytics,* s.l.: National Visualization and Analytics Center.

Thomas, J. & Kielman, J., 2006. Challenges for Visual Analytics. *Information Visualization,* 8(4), pp. 1-6.

Tominsky, C., Forsell, C. & Johansson, J., 2012. Interaction support for visual comparison inspired by natural behavior. *IEEE transactions on visualization and computer graphics,* 18(12), pp. 2719-2728.

Treisman, A. M. & Gelade, G., 1980. A Feature-integration theory of attention. *Cognitive psychology,* 12(1), pp. 97-136.

Tufte, E. R., 2001. *The Visual Display of Quantitative Information, second edition.* Cheshire, Connecticut: Graphics Press.

van Boeijen, A., Daalhuizen, J., Zijlstra, J. & van der Schoor, R., 2013. *Delft Design Guide.* 1 ed. s.l.:BIS publishers.

van Wijk, J. J., 2005. *The Value of Visualizations.* Minneapolis, MN, USA, IEEE, pp. 79-86.

Victor, B., 2011. *Up and down the ladder of abstraction.* [Online] Available at: http://worrydream.com/LadderOfAbstraction/ [Accessed 28 November 2013].

Visualizing.org, n.d.. *A taxonomy of data visualization | visualizing.org.* [Online] Available at: http://visualizing.org/stories/taxonomy-data-visualization [Accessed 16 September 2013].

Wang, F. & Hannafin, M. J., 2005. Design-based research and technology-enhanced learning environments. *Educational technology research and development,* 53(4), pp. 5-23.

Ware, C., 2004. *Information visualization.* 2nd ed. San Francisco, CA: Morgan Kaufmann Publishers.

Ware, C., 2008. *Visual Thinking for Design.* Burlington, MA: Morgan Kaufmann Publishers.

Wertheimer, M., n.d. *Laws if Organization in Perceptual Forms.* [Online]
Available at: http://psychclassics.yorku.ca/Wertheimer/Forms/forms.htm
[Accessed 10 September 2013].

Worldbank, n.d. *Internet users (per 100 people) | Data | Table.* [Online]
Available at:
http://data.worldbank.org/indicator/IT.NET.USER.P2?order=wbapi_data_value_2012+wbapi_data_value+wbapi_data_value-last&sort=desc
[Accessed 11 2 2014].

Yates, K. A. & Clark, R. E., 2012. Cognitive task analysis. In: *International Guide to Student Achievement.* s.l.:s.n.

Yau, N., 2013. *Data points.* Indianapolic, Indiana: John Wiley & Sons.

Zimmerman, J., Forlizzi, J. & Evenson, S., 2007. *Research through design as a method for interaction design research in HCI.* s.l., Human-Computer Interaction Insitute.

"I *see* what you mean"

