Integrating Explicit Knowledge in the Visual Analytics Process
Knowledge-assisted Visual Analytics Methods for Time-oriented Data

Markus Wagner
IC\M/T-Institute of Creative\Media/Technologies, St. Poelten University of Applied Sciences, Matthias Corvinus-Straße 15, St. Poelten, Austria
Faculty of Informatics, Vienna University of Technology, Erzherzog-Johann-Platz 1/180, A-1040 Vienna, Austria

ABSTRACT

In this paper, I describe my thesis project for the integration of explicit knowledge from domain experts into the visual analytics process. As base for the implementation of the research project, I will follow the nested model for visualization design and validation. Additionally, I use a problem-driven approach to study knowledge-assisted visualization systems for time-oriented data in the context of real world problems. At first, my research will focus on the IT-security domain where I analyze the needs of malware analysts to support them during their work. Therefore I have currently prepared a problem characterization and abstraction to understand the needs of the domain experts to gain more insight into their workflow. Based on that findings, I am currently working on the design and the implementation of a prototype. Next, I will evaluate these visual analytics methods and finally I will test the generalizability of the knowledge-assisted visual analytics methods in a second domain.

Keywords. Visual Analytics, Implicit Knowledge, Explicit Knowledge, Problem-driven Research, Time-oriented Data, Knowledge-assisted Visualization.

1 RESEARCH PROBLEM

Visual analytics, "*the science of analytical reasoning facilitated by interactive visual interfaces*" (Thomas and Cook, 2005, p. 4), is a comparably young research field. A major tenet of visual analytics is that analytical reasoning is not a routine activity that can be automated completely (Wegner, 1997). Instead it depends heavily on analysts' initiative and domain experience. Visual interfaces, especially Information Visualizations (InfoVis), are high bandwidth gateways for perception of structures, patterns, or connections hidden in the data. Interaction is "*the heart*" of InfoVis (Spence, 2006, p. 136) and allows the analytical reasoning process to be flexible and react to unexpected insights. Furthermore, visual analytics involves automated analysis methods, which perform computational activities on potentially large volumes of data and thus complement human cognition.

When analysts solve real world problems, they have large volumes of complex and heterogeneous data at their disposal. On the one hand time-oriented data (see Section 1.1) is of particular importance due to its central role in many analysis contexts and tasks and on the other hand the distinct characteristics of the dimension time make distinct methods necessary. By externalization and storing of the implicit knowledge, it will be made available as explicit knowledge (see Section 1.2). In addition to sophisticated analysis methods, implicit and tacit knowledge about the data, the domain or prior experience are often required to make sense of this data and not get overwhelmed. In this work I examine *how the visual analytics process can benefit from explicit knowledge of analysts*. This will help to develop more effective environments for gaining insights – the ability to specify, model and make use of auxiliary information about data and domain specifics. In addition to the raw data they will help to better select, tailor, and adjust appropriate methods for visualization, interaction, and automated analysis. Potential application domains benefiting from this are healthcare, biotechnology, urban- and cyberinfrastructures, environmental science and many more.

The main goal of this thesis is to develop knowledge-assisted visualization and interaction methods (see Section 1.3) that make use of explicit knowledge to improve these methods in a context-specific manner. This reflects intricate problems which are recognized by the visual analytics community as important research challenges (Pike et al., 2009).

To be effective, visual analytics needs to provide 'precise' data, "*which is immediate, relevant and un-*

Wagner M..
Integrating Explicit Knowledge in the Visual Analytics Process - Knowledge-assisted Visual Analytics Methods for Time-oriented Data.

derstandable to individual users, groups, or communities of interest" (Kielman et al., 2009, p. 240). For example analysts might have hunches, which sources they believe to be trustable, which results appear plausible and which insights they deem relevant. By externalizing this knowledge and using it, analysts can avoid cognitive overload and use visualization and automated analysis methods more effectively. They can avoid reinventing the wheel, when they repeat analysis on a different dataset, a year later, or through a different technique. They can keep track of interpretations and analysis steps, communicate with co-analysts, and document results for insight provenance. Leading visualization researchers have repeatedly called for the integration of knowledge with visualization. Chen (2005) lists 'prior knowledge' as one of ten unsolved InfoVis problems. He argues that InfoVis systems need to be adaptive for accumulated knowledge of users, especially domain knowledge needed to interpret results. In their discussion of the 'science of interaction', Pike et al. (2009) point out that visual analytics tools have only underdeveloped abilities to represent and reason with human knowledge. Therefore, they declare 'knowledge-based interfaces' as one of seven research challenges for the next years.

1.1 Time-oriented Data

Visual exploration and analytical reasoning with time-oriented data are common and important for numerous application scenarios, e.g., in healthcare (Combi et al., 2010), business (Lammarsch et al., 2009), and security (Fischer et al., 2012; Saxe et al., 2012). Furthermore, time and time-oriented data have distinct characteristics that make it worthwhile to treat it as a separate data type (Shneiderman, 1996; Andrienko and Andrienko, 2005; Aigner et al., 2011). Explicit knowledge may model the relevance of data items in respect to zoom levels and recommend summarization techniques depending on task(s) and domain(s).

When dealing with time, we commonly interpret it with a calendar and its time units are essential for reasoning about time. However, these calendars have complex structures. In the Gregorian calendar the duration of a month varies between 28 and 31 days and weeks overlap with months and years. Furthermore, available data may be measured at different levels of temporal precision. Some patterns in time-oriented data may emerge when a cyclic structure of time is assumed, for example, traffic volume by time of day, temperature by season.

In other cases, an analyst will need to balance such effects to understand long term trends. An analyst may be interested to compare developments in the data that do not cover the same portion of time. For such comparisons, they are interested in relative time to some sentinel events. Therefore they would align patient data by the beginning of a specific therapy, and show all events one day after the beginning (Wang et al., 2008; Rind et al., 2011).

1.2 Explicit Knowledge

Computerized representations of interests and domain knowledge will be referred to as 'explicit knowledge'. As there are many competing definitions of 'knowledge' in scientific discourse, the definition of the community of knowledge-assisted visualization is:

> "*Knowledge: Data that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or the transcripts of some knowledge acquired by human beings.*" (Chen et al., 2009, p. 13)

In this work mainly the second part of this definition is of importance. Wang et al. (2009) further distinguish between explicit knowledge that "*can be processed by a computer, transmitted electronically, or stored in a database*" while tacit knowledge "*is personal and specialized and can only be extracted by human*". In this thesis, the focus will be to investigate how explicit knowledge can be used to support interactive visualization (knowledge-assisted visualization). The specification of the users knowledge will not be a part of this work.

1.3 Knowledge-assisted Visualization

There are numerous ways to optimize visualization and interaction methods based on explicit knowledge. For example choosing variables for scatter plot axes, zooming to an area of interest instead of the viewport center, highlighting data items in a different color, or drawing reference lines in the background of a plot. Such optimizations can be applied to most aspects of the visualization and developing a general framework instead of scenario-specific solutions is a challenging task (Tominski, 2011).

The visual analytics of data is an explorative process. If there is given a dataset, the user needs to decide, which visualization method(s) he wants to use for the data exploration. The objectives of knowledge-assisted visualizations include the sharing of explicit knowledge (domain knowledge) from different users. Thus, it reduces the stress on users for appropriate knowledge about complex visualization techniques (Chen and Hagen, 2010).

For example, explicit knowledge can be used to summarize and abstract a dataset. These summarizations and abstractions will form another dataset, which can be visualized through a wide range of existing visualization and interaction methods. Typically this abstraction process reduces the size of the dataset significantly. However, analysts also need to access the input dataset and switching between visualizations of both datasets should be facilitated by techniques like semantic zoom (Perlin and Fox, 1993) or brushing and linking (Becker and Cleveland, 1987). The wide ranging potential of utilizing explicit knowledge has already been demonstrated in recent research (Chen and Hagen, 2010). Despite this, most current visualization systems do not take advantage of explicit knowledge captured from domain experts.

2 OUTLINE OF OBJECTIVES

In this project, the overall aim is to develop knowledge-assisted visual analytics methods to gain insights effectively from time-oriented datasets (see Section 1.1). In these methods, explicit knowledge is treated as externally given, and the focus will be on how to best integrate them into the process to improve sense-making.

Knowledge-assisted visualization and interaction methods (see Section 1.3) will be developed to explore time-oriented datasets. I hypothesize that explicit knowledge (see Section 1.2) will afford for more effective analytical reasoning processes (e.g., through semi-automated visualization) and prevent data interpretation errors. Finally, all developed methods need to undergo evaluation. Scenarios will be identified with target users, tasks, and datasets that act as testbeds. Designs and prototypes will be iteratively evaluated and refined. Based on these aims this work investigates the following research questions:

- **Main Question:** How can the visual analytics process benefit from explicit knowledge of analysts?
- **Sub Question:** How can explicit knowledge be visually represented effectively in an visual analytics system?
- **Sub Question:** Is it possible to generalize the interaction with knowledge-assisted visualization methods for different application scenarios?
- **Sub Question:** How can analysts during the exploration of a large amount of data benefit from knowledge-assisted visual analytics methods?

The developed methods of this thesis will primarily deal with time-oriented data, but in future work they will also be applicable for other datasets.

3 STATE OF THE ART

The permanent growth of methods and parameters which are available for data visualization can be confusing for novice users and even for domain experts. Another problem is that the extensive know-how is not stored in a central place because it is separated in sub-communities (Nam et al., 2009; Mistelbauer et al., 2012). Knowledge-assisted visualizations (KAV) are a fast increasing field which uses direct integrated expert knowledge to produce effective data visualizations. Most of the KAV systems concentrate on the integration of specific domain knowledge which can only be used for exactly these analysis tasks. Additionally it is important that the users become aware of the different methods which are needed for the data exploration and interaction but not all methods are usable or effective for the different data types to gain the expected results (Wang et al., 2009; Mistelbauer et al., 2012). Existing data visualization systems need a manual specification for each data attribute of possible visualizations. This is also significant for data which are available as linked open data and systems which represent the data as graphs with objects and weighted edges with labels (Cammarano et al., 2007). It is important to differentiate between automatic visualization systems (AVS) and automated visualization systems. Automatic visualization systems make independent decisions about the visualization activities. The automated visualization system is a programming system for the automated generation of diagrams, graphics and visualizations. In general it is necessary that the flow of an automate visualization system works like an expert would perform it (Wills and Wilkinson, 2010).

Cammarano et al. (2007) described in their paper the automatization of the data integration and the automatic mapping of data attributes to visual attributes. This workflow was described as the "*schema matching problem*" (Cammarano et al., 2007). It includes the automated finding of ways in the data model for each needed visualization attribute based on visualization templates. The used data model equals the Resource-Description-Framework (RDF). Each subject-predicate-object triple of the RDF model corresponds to the edge which connects a subject with an object. Based on the provided experiments the authors showed that the needed data could be identified frequently enough that the system could be used as an exploration tool. This way it saves the user from schema-heterogeneity.

Falconer et al. (2009) treated the generation of adapted visualizations based on ontological datasets

and the specification of ontological mappings. The usability of this approach was demonstrated by the use of the ontology-mapping-tool *COGZ* in this paper, whereat ontological mappings would be translated into software transformation rules. With this transformations, the domain specific data are converted in a way to fit to a model which describes the visualization. To perform the mappings, the authors developed a rule description library based on Atlas Transformation Language (ATL) (Jouault and Kurtev, 2006). With this library they converted the specific source data into target mappings. The tests of the system showed that the system performed an automated mapping of the data, whereby the user was assisted greatly in his work.

Gilson et al. (2008) described in their paper the automated generation of visualizations from domain specific data of the web. Therefore, they described a general system pipeline which combines ontological mappings and probabilistic argumentation techniques. In the first step, they mapped a website into a domain ontology which stores the semantics of the specific subject domains (e.g. music charts). Subsequently they mapped it to one or more visual-representation-ontologies whereby each contains the semantic of a visualization technique (e.g. treemap). To guarantee the mapping between the two ontologies, they introduced a semantic-bridge-ontology which specifies the suitability of each ontology. Based on this approach, they implemented a prototype with the name *SemViz*. For the tests of the system, they used the data of popular music websites without having prior knowledge about the pages.

Mackinlay et al. (2007) introduced in their paper the tool *Show Me* which is an integrated set of interface commands and standard values which automatically integrate data presentations into the tool *Tableau*. The key aspect of Tableau is VizQL (visualization query language) which would be used by Show Me to generate automated presentations in a view table. One of the major aspects is the usability of the tool which has to support the flow of visual analytics. This includes the automated selection of marking techniques, commands to combine individual fields to one view and some commands to generate views from multiple fields. The APT system by Mackinlay (1986) forms the basis for the automated design of graphical representations of relational information. The authors implemented Bertins semiology of graphics as algebraic operations (Bertin, 1983) and used them for the search of effective presentations for the information.

Wills and Wilkinson (2010) described the data viewer tool *AutoVis* which reacts on content (text, relational tables, hierarchies, streams, images) and presents the containing information in an appropriate form (e.g. like an expert will do it). The design is based on the grammar of graphics and the logic is based on statistical analysis. This automatic visualization system was developed to provide a first look on the data until the modeling and analysis are finished. AutoVis was designed to protect the researchers ignoring missing data, outliers, miscodings and other anomalies which injure the statistical adoption or the validity of the models. The design of this system contains some unique features: a spare interface, a graphics generator, a statistical analysis to protect users from false conclusions and pattern recognition.

Tominski (2011) described in his paper a new approach for event-based visualizations which contains three fundamental stages. First, the event specification is to generate event types which are interesting as a visualization for the users. This translates the user interests in an understandable representation for the computer, where they should be formulated for the user as easy as possible. The second stage specified where the interests of the users intersects with the data. This detection must be kept as general as possible so that it is applicable to a large number of event types. The basic task is to assess encapsulated the conditions of event types. The aim of the third step is to integrate the detected event instances in visual representations (which reflect the interests of users). The event representation has great influence on the extent to which the event-based visualization closes the gap for world view. This general model allows the use of many different visualizations and the specific data-driven events focused on relational data visualizations of today.

Kadlec et al. (2010) described in their paper that scientists are using seismic 3D data for 30 years to explore the earth crust by the interpretation of seismic data which needs a lot of expert knowledge. But it is possible to use the knowledge of experts in order to facilitate the segmentation of the geological features. To reduce the need for knowledge of seismic data and attributes, this new method uses surfaces which are growing in surfaces of known geologic characteristics. The result is a knowledge-assisted visualization and segmentation system that allows non-expert users a fast segmentation of geological features in complex data collections. The process begins with the exploration of seismic datasets using 2D slices. This 3D volume are searched interactively for possible interesting features. The elements are rendered and the user receives a feedback on the quality of the segmented features. If the system indicates a link to a

non-feature, the user has the ability to repair this. This approach transferred the expert knowledge very fast and reliable for non-expert users. This way the analysis quality of non-expert users increases similar to those of experts.

Nam et al. (2009) described that domain specific know-how is separated in sub-communities. To overcome this problem they started to store visualization expertises and methods in combination with possible datasets. An important aspect is to edit newly generated datasets with the existing expert knowledge from a database. Therefore, they used several levels of granularity to use the knowledge of the database correctly. Thus, they described the first step of a framework specifically in relation to the data categorization and classification by using a set of feature vectors. The usability of the framework was demonstrated on four medical datasets (knee, chest and head 2x) in a 2D application. They calculated for every dataset feature points in a local density histogram and described them as low-level feature vectors. These would be used to prepare high-level-models of the data objects. Furthermore, they want to support a general framework for classification tasks by indexing a database knowledge for knowledge-assisted visualization systems (KAV).

Wang et al. (2009) differentiated between two types of knowledge (implicit and explicit) and defined four conversion processes between them (internalization, externalization, cooperation and combination) which were included in knowledge-assisted visualizations. They showed the applications of these four processes their roles and utilities in real-life scenarios using a visual analysis system for the Department of Transportation. The authors assume that the analysts can learn more through the interaction between implicit and explicit knowledge through the use of interactive visualization tools. As a further distinction between implicit and explicit knowledge in knowledge-assisted visualization, the following is stated by the authors:

- "*Explicit knowledge is different from data or information.*"
- "*Tacit knowledge can only result from human cognitive processing (reasoning).*"
- "*Explicit knowledge exists in data, and is independent from the user or his tacit knowledge.*"
- "*Explicit and tacit knowledge are related and can be connected through the use of interactive visualization tools.*" (Wang et al., 2009, p. 2)

Upon connection of the system to an ontological knowledge source, the visual analytics system enables the user an interactive access to the expertise of the expert. Thus, this visualization system showed that the four knowledge conversion processes are possible for the design of knowledge-assisted visualization.

Mistelbauer et al. (2012) described in their paper a knowledge-assisted system for medical data visualization (*Smart Super Views*). This system has been tested in the medical domain and expert feedback was obtained. The Smart Super Views system contains three major steps: In the first step the information from different sources will be collected and merged. In the second step, the user decides where a region of interest (ROI) is located in the data and which visualization technique should be used. In the third step, the user interacts with the provided visualization and starts with a detailed inspection of the data. In contrast to other systems where the user himself has to select the visualization, this system will support the user in his decisions. The rule specification module of the system defines the connection between the input data and the output visualization. To model these connections there will be used *if-then* clauses, which were specified by domain experts. Additionally, these clauses were stored in a user-readable form in a file.

3.1 Discussion

The automation of the data integration and the automatic mapping of data attributes to visual attributes is discussed in many papers (e.g. (Cammarano et al., 2007; Falconer et al., 2009; Gilson et al., 2008; Mackinlay et al., 2007; Wills and Wilkinson, 2010; Kadlec et al., 2010; Mistelbauer et al., 2012)). The generation of adapted visualizations which are based on ontological datasets and the specification of ontological mappings are treated by Falconer et al. (2009). A similar approach was also followed by Gilson et al. (2008). They described a general system pipeline which combines ontology mapping and probabilistic reasoning techniques. The approach of Gilson et al. (2008) is described by the automated generation of visualizations of domain-specific data from the web. In contrast, Falconer et al. (2009) used the *COGZ* tool for their approach which converts ontological mappings in software transformation rules so that it describes a model which fits the visualization. Cammarano et al. (2007) describes a similar process as "*schema matching problem*". It describes finding ways in the data model for each required visualization attribute based on visualization templates. In the end, most of the automated data mappings for visualizations try to perform in similar ways. Gilson et al. (2008) maps the semantic data to visual-representation-ontologys, each part contains the semantics of a visualization (e.g. treemaps). A

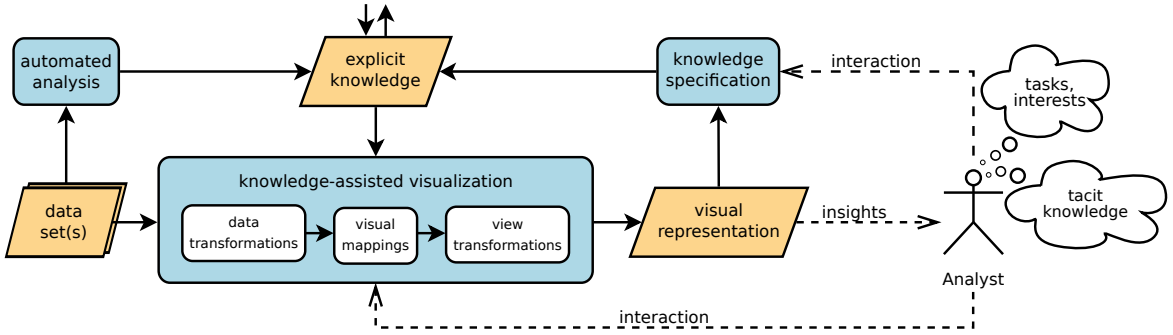

Figure 1: This image shows the integration of explicit knowledge for knowledge-assisted visualizations in the visual analytics process.

slightly different approach by Mackinlay et al. (2007) has a set of interactive commands, defaults, automated data integration and presentations to accomplish the automated data presentation in *Tableau*. Due to the automatic selection of markers, commands and combination of individual fields to a view, the user is able to rapidly and easily create visualizations by the use of the tool *Show Me*. Furthermore, the tool *AutoVis* was implemented by Wills and Wilkinson (2010) to take a first look at data which has to be visualized. For this, the system used statistical analysis for modeling the visualizations. Thus, the user should be prevented of ignoring missing data, outliers, missing codes and other anomalies. The protection (e.g. (Cammarano et al., 2007)) or the support of the users during their work (e.g. (Falconer et al., 2009; Mackinlay et al., 2007; Tominski, 2011; Kadlec et al., 2010; Mistelbauer et al., 2012)) is one of the main foci of this papers.

The *event-based model* by Tominski (2011) permits the applicability for many different visualizations which are divided into three stages. A stepwise subdivision is also used by Gilson et al. (2008) for the required mapping instances and Mistelbauer et al. (2012) used a stepwise subdivision for the three processing steps of the *Smart Super Views*. The three essential steps for a knowledge-assisted visualization tool according to Mistelbauer et al. (2012) are: first to collect and merge the data; second, to determine the region of interest (ROI) in the data by the user; third, the interaction of users with the generated visualization. The automated generation of visualizations respectively the assigning of the data to predefined visualization templates is also carried out in other papers, which were presented in this state of the art report, on similar ways. In some papers it is also described that the knowledge of experts is distributed. Therefore, it is important to develop knowledge-assisted visualization systems to make the knowledge of experts available for the users (e.g. (Kadlec et al., 2010; Nam et al., 2009; Wang et al., 2009)). Usually, the knowledge of experts is stored in files (e.g. (Mistelbauer et al., 2012)),using RDF (e.g. (Cammarano et al., 2007)) or in a knowledge database (e.g. (Nam et al., 2009)).

Based on these findings, it can be seen that most of the papers treat the storing or the availability of explicit knowledge. Additionally, most of the currently implemented knowledge assisted visualization systems are concentrated on the integration of specific domain knowledge which could only be used for precise analysis task. Even automated generations of visualizations are described for example by Mackinlay et al. (2007), but knowledge-assisted visualizations methods in combination with visual analytics are not clearly addressed in those papers. Thus it is clear that a lot of space for future research is available in the field of knowledge-assisted visualizations in combination with visual analytics, especially in the generalization of knowledge-assisted visualization methods.

4 METHODOLOGY

In this section, the plan how to apply and study *knowledge-assisted visualizations* in combination with visual analytics methods will be presented. *Explicit knowledge* of domain experts will be used to support users during the analysis of time-oriented data.

By the use of knowledge-assisted visualizations, the available datasets will be turned into interactive

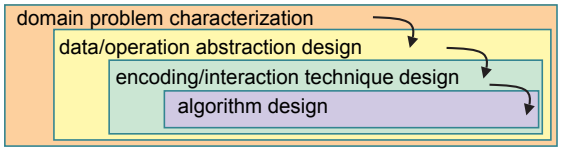

Figure 2: The 4 levels of the *nested model for visualization design and validation* by Munzner (2009).

and visual representations (see Figure 1). Thus, explicit knowledge will be used to achieve effective representations in terms of the analysis' tasks. The visualization process can be described using the reference model of Card and Card (1999) or the data state model of Chi and Riedl (1998). Both descriptions relate to the *internalization* of the model of Wang et al. (2009).

Throughout this project, I will follow the well-known *nested model for visualization design and validation* as proposed by Munzner (2009) (see Figure 2). This unified approach splits visualization design into 4 levels in combination with corresponding evaluation methods to evaluate the results at each level. Starting from the top, the levels of the nested model for visualization design and validation are:

- **Domain Problem and Data Characterization:** On this level, the goal is to understand the problem domain, the users' tasks and their goals.
- **Operation and Data Type Abstraction:** Within the abstraction level, domain specific vocabulary (problems and data) will be mapped to a more generic description which fits to the vocabulary of computer scientists (visualization community).
- **Visual Encoding and Interaction Design:** In the third level, the visual encoding of the data and the interaction methods for the data exploration will be designed.
- **Algorithm Design:** Designing of the implementation of the visual encoding and interaction methods.

Since these are nested levels, the output of the upstream level which is situated above, is the input of the downstream level which is situated below. Considering it is current practice, visual analytics was defined by Keim et al. as the "*[combination] of automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making in the basis of very large and complex datasets*" (Keim et al., 2010, p. 7). In general the nested model for visualization design and validation does not include automated analysis explicitly, but it can be conceptualized on the abstraction level where the data transformation takes place. This thesis will focus on knowledge-assisted visualizations for visual analytics to develop novel visual encoding and interaction methods for time-oriented data.

For the research I will follow a problem-driven approach to study knowledge-assisted visualization systems for time-oriented data in the context of real world problems. At first my research will focus on the IT-security domain. More specifically, I will analyze the needs of malware analysts in relation to their work on behavior-based malware pattern analysis Dornhackl et al. (2014). Therefore, I will design knowledge-assisted visual analytics methods and implement a software prototype as proof of concept to test the designed methods. After this, the system will be tested in the context of a second domain to be specified.

To ensure a knowledgeable research I will start with a problem characterization and abstraction based on the design study methodology of Sedlmair et al. (2012), which brings me into the first level (domain problem and data characterization) of the nested model. From there, I will work inwards along Munzner's nested model for visualization design and validation. To perform the problem characterization and abstraction, I will follow a threefold qualitative research approach which consists of a systematic literature research, a focus group (Lazar et al., 2010, p. 192) and semi-structured interviews (Lazar et al., 2010, p. 184) with domain experts. Based on the results of the threefold approach, I will use the *design triangle* as proposed by Miksch and Aigner (2014) to analyze the data, the users and the tasks which fits to the second level of Munzner's model (operation and data type abstraction).

In the following steps, I will start with the visualization and interaction design followed by the algorithm design and implementation based on a user centered design process (Sharp et al., 2007). Therefore I will produce sketches, followed by screen prototypes and functional prototypes (Kulyk et al., 2007, p. 50). This way I will fulfill the third (visual encoding and interaction design) and the fourth (algorithm design) level of Munzner's nested model. During these steps, focus group members will be included in the design and implementation process to get feedback about the design and the functionality of the knowledge-assisted visual analytics system. Thus it will be possible to improve the design and the handling of the designed knowledge-assisted visualization methods.

Additionally, user studies will be performed with predefined datasets to evaluate the usability (Cooper et al., 2007, p. 70) of the new knowledge-assisted visualization methods based on the implemented visual analytics system.

After the performed user studies, which will be based on the first real world problem (behavior-based malware pattern analysis) of the knowledge-assisted visualization methods are completed, I will start to test their applicability on a second real world problem. Therefore, I will adapt/extend the knowledge-assisted visualization methods, if it is necessary, and I will repeat the previously described research process in the required extent.

5 EXPECTED OUTCOME

The goal of this thesis is to show how the visual analytics process can benefit from the use of knowledge-assisted visual analytics methods. To achieve this, implicit knowledge of the domain-specific analysis experts will be stored as explicit knowledge (e.g. in a database). This explicit knowledge will be used to support users during their workflow in a context-specific manner (e.g. behavior-based malware pattern analysis) to achieve their goals. Thus, that knowledge-assisted visualization methods will support the generation of more effective visual analytics environments to gain more insights and achieve better quality results compared to current methods.

In addition to the raw data, knowledge-assisted visual analytics methods will help to better select, tailor, and adjust appropriate methods for visual representation, interaction, automated analysis and prevent data interpretation errors. By externalizing the domain-specific expert knowledge and using it, analysts can avoid cognitive overload and use visualization and automated analysis methods more effectively. This way, analysts can avoid reinventing the wheel, when they repeat analysis on a different dataset, a year later, or using a different technique. Thus, they can concentrate on the important steps of interpretations and analysis, communicate with co-analysts, and document results for insight provenance.

Furthermore, the tested knowledge-assisted visualization methods will be generalized and applied to different domains. Based on this, generalizations and the results of the interviews and user studies (see Section 4), I will propose general design guidelines for future knowledge-assisted visual analytics environments to support the community. Additionally, it will be demonstrated how similar knowledge-assisted visualization methods can be used for different domains.

6 STAGE OF THE RESEARCH

I started with my doctoral studies on March 01, 2014. Currently I have developed a problem characterization and abstraction for the field of malware pattern analysis which I presented at *ACM VizSec14* (Wagner et al., 2014). This paper deals with a user-study based on a threefold research approach which includes a literature research, focus group meetings and semi-structured interviews. In relation to the combined findings of the user-study, there was developed a data-users-tasks analysis, based on the design triangle by Miksch and Aigner (2014), to analyze and abstract the results for other domains.

According to the findings of the user-study and the data-users-tasks abstraction, I developed different mockups and work scenarios for the prototype which will be supported by knowledge-assisted visual analytics methods. Additionally, I started with the initial work for the implementation of a screen prototype of the visual analytics system and the background data handling process.

In addition to the performed research steps, we presented the KAVA-Time project at the European Researchers' Night 2014 in Vienna to the general public.

6.1 Next Steps

In the following steps, I will finish the design and implementation of the prototype system to test the new designed knowledge-assisted visual analytics methods. For the design and implementation of the prototype, I will follow the paradigm of user centered design process (Sharp et al., 2007) in cooperation with a focus group (Lazar et al., 2010, p. 192). After these, I will perform a user-study (Lazar et al., 2010) with predefined datasets to test the new knowledge-assisted visual analytics methods. Additionally, I will test the applicability on a second domain, whereby I will adapt/extend the knowledge-assisted visual analytics methods, if it is necessary. For all this steps, I will follow the nested model for visualization design and validation by Munzner (2009).

ACKNOWLEDGEMENT

This work was supported by the Austrian Science Fund (FWF) via the KAVA-Time project no. P25489. Many thanks to my mentor Wolfgang Aigner and my colleague Alexander Rind for their feedback to my manuscript.

REFERENCES

Aigner, W., Miksch, S., Schumann, H., and Tominski, C. (2011). *Visualization of Time-Oriented Data*. Springer, London.

Andrienko, N. and Andrienko, G. (2005). *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, Berlin, New York.

Becker, R. A. and Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, 29(2):127–142.

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

Cammarano, M., Dong, X., Chan, B., Klingner, J., Talbot, J., Halevy, A., and Hanrahan, P. (2007). Visualization of heterogeneous data. *TVCG*, 13(6):1200–1207.

Card, S. and Card, M. (1999). *Readings in Information Visualisation. Using Vision to Think.: Using Vision to Think*. Morgan Kaufman Publ Inc, San Francisco, Calif.

Chen, C. (2005). Top 10 unsolved information visualization problems. *CG&A*, 25(4):12–16.

Chen, M., Ebert, D., Hagen, H., Laramee, R., Van Liere, R., Ma, K.-L., Ribarsky, W., Scheuermann, G., and Silver, D. (2009). Data, information, and knowledge in visualization. *CG&A*, 29(1):12–19.

Chen, M. and Hagen, H. (2010). Guest editors' introduction: Knowledge-assisted visualization. *CG&A*, 30(1):15–16.

Chi, E. H.-H. and Riedl, J. (1998). An operator interaction framework for visualization systems. In *IEEE Symposium on Information Visualization, 1998. Proceedings*, pages 63–70.

Combi, C., Keravnou-Papailiou, E., and Shahar, Y. (2010). *Temporal Information Systems in Medicine*. Springer, New York.

Cooper, A., Reimann, R., and Cronin, D. (2007). *About Face 3: The Essentials of Interaction Design*. Wiley, Indianapolis, IN, 3rd edition.

Dornhackl, H., Kadletz, K., Luh, R., and Tavolato, P. (2014). Malicious behavior patterns. In *IEEE 8th International Symposium on Service Oriented System Engineering*, pages 384–389.

Falconer, S., Bull, R., Grammel, L., and Storey, M. (2009). Creating visualizations through ontology mapping. In *CISIS*, pages 688–693.

Fischer, F., Mansmann, F., and Keim, D. A. (2012). Real-time visual analytics for event data streams. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 801–806. ACM.

Gilson, O., Silva, N., Grant, P., and Chen, M. (2008). From web data to visualization via ontology mapping. *Computer Graphics Forum*, 27(3):959–966.

Jouault, F. and Kurtev, I. (2006). Transforming models with ATL. In Bruel, J.-M., editor, *Satellite Events at the MoDELS 2005 Conference*, number 3844 in Lecture Notes in Computer Science, pages 128–138. Springer Berlin Heidelberg.

Kadlec, B., Tufo, H., and Dorn, G. (2010). Knowledge-assisted visualization and segmentation of geologic features. *CG&A*, 30(1):30–39.

Keim, D., Kohlhammer, J., Ellis, G., and Mansmann, F., editors (2010). *Mastering the information age: solving problems with visual analytics*. Eurographics Association, Goslar.

Kielman, J., Thomas, J., and May, R. (2009). Foundations and frontiers in visual analytics. *Information Visualization*, 8(4):239–246.

Kulyk, O., Kosara, R., Urquiza, J., and Wassink, I. (2007). Human-centered aspects. In Kerren, A., Ebert, A., and Meyer, J., editors, *Human-Centered Visualization Environments*, number 4417 in Lecture Notes in Computer Science, pages 13–75. Springer, Berlin.

Lammarsch, T., Aigner, W., Bertone, A., Gartner, J., Mayr, E., Miksch, S., and Smuc, M. (2009). Hierarchical temporal patterns and interactive aggregated views for pixel-based visualizations. In *Information Visualisation, 2009 13th International Conference*, pages 44–50.

Lazar, J., Feng, J. H., and Hochheiser, H. (2010). *Research Methods in Human-Computer Interaction*. Wiley, Chichester, West Sussex, U.K, 1 edition.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141.

Mackinlay, J., Hanrahan, P., and Stolte, C. (2007). Show me: Automatic presentation for visual analysis. *TVCG*, 13(6):1137–1144.

Miksch, S. and Aigner, W. (2014). A matter of time: Applying a data-users-tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics, Special Section on Visual Analytics*, 38:286–290.

Mistelbauer, G., Bouzari, H., Schernthaner, R., Baclija, I., Kochl, A., Bruckner, S., Sramek, M., and Groller, M. (2012). Smart super views: A knowledge-assisted interface for medical visualization. In *VAST*, pages 163–172.

Munzner, T. (2009). A nested model for visualization design and validation. *TVCG*, 15(6):921–928.

Nam, J. E., Maurer, M., and Mueller, K. (2009). A high-dimensional feature clustering approach to support knowledge-assisted visualization. *Computers & Graphics*, 33(5):607–615.

Perlin, K. and Fox, D. (1993). Pad: An alternative approach to the computer interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 57–64, New York. ACM.

Pike, W. A., Stasko, J., Chang, R., and O'Connell, T. A. (2009). The science of interaction. *Information Visualization*, 8(4):263–274.

Rind, A., Aigner, W., Miksch, S., Wiltner, S., Pohl, M., Turic, T., and Drexler, F. (2011). Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation. In Holzinger, A. and Simonic, K.-M., editors, *Information Quality in e-Health*, number 7058 in LNCS, pages 301–320. Springer, Berlin.

Saxe, J., Mentis, D., and Greamo, C. (2012). Visualization of shared system call sequence relationships in large malware corpora. In *International Workshop on Visualization for Cyber Security*, VizSec '12, pages 33–40. ACM.

Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *TVCG*, 18(12):2431–2440.

Sharp, H., Rogers, Y., and Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, Chichester ; Hoboken, NJ, 2. edition.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages. Proceedings*, pages 336–343.

Spence, R. (2006). *Information Visualization: Design for Interaction*. Prentice Hall, New York, 2nd rev. edition.

Thomas, J. J. and Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr. Published: Paperback.

Tominski, C. (2011). Event-based concepts for user-driven visualization. *Information Visualization*, 10(1):65–81.

Wagner, M., Aigner, W., Rind, A., Dornhackl, H., Kadletz, K., Luh, R., and Tavolato, P. (2014). Problem characterization and abstraction for visual analytics in behavior-based malware pattern analysis. In *International Workshop on Visualization for Cyber Security*. ACM.

Wang, T. D., Plaisant, C., Quinn, A. J., Stanchak, R., Murphy, S., and Shneiderman, B. (2008). Aligning temporal data by sentinel events: Discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 457–466, New York. ACM.

Wang, X., Jeong, D. H., Dou, W., Lee, S.-W., Ribarsky, W., and Chang, R. (2009). Defining and applying knowledge conversion processes to a visual analytics system. *Computers & Graphics*, 33(5):616–623.

Wegner, P. (1997). Why interaction is more powerful than algorithms. *Commun. ACM*, 40(5):80–91.

Wills, G. and Wilkinson, L. (2010). AutoVis: Automatic visualization. *Information Visualization*, 9(1):47–69.