Visual Analytics

Perspectives on the Field of Interactive Visualization

MAGNUS JÄNDEL (ED.), PETTER BIVALL, PETER HAMMAR,
RONNIE JOHANSSON, FARZAD KAMRANI, MATTIAS J. QUAS

Total population (1000 inhabitants)
11,445.1
Population aged 0
Population aged 15-64 (% of total...
Population aged 65+ (% of total p...
26.6

FOI-R--4200--SE
ISSN 1650-1942 January 2016

FOI-R--4200--SE

Magnus Jändel (Ed.), Petter Bivall,
Peter Hammar, Ronnie Johansson,
Farzad Kamrani, Mattias J. Quas

Visual Analytics

Perspectives on the Field of Interactive Visualization

Bild/Cover: Petter Bivall

Titel	Visuell dataanalys
Title	Visual Analytics
Rapportnr/Report no	FOI-R--4200--SE
Månad/Month	February
Utgivningsår/Year	2016
Antal sidor/Pages	90
ISSN	1650-1942
Kund/Customer	Intern
Forskningsområde	1. Beslutsstödssystem och informationsfusion
FoT-område	Avskanning av forskningsfronten
Projektnr/Project no	I354201
Godkänd av/Approved by	Lars Höstbeck
Ansvarig avdelning	IAS

Sammanfattning

I inledningen i *kapitel 1* definierar vi *visuell dataanalys* (VA) som vetenskapen om att stödja mänskligt tänkande och situationsuppfattning med hjälp av visuella representationer där interaktion är en viktig del av analysprocessen. Dessutom noterar vi att VA är en vetenskap som är i gränslandet mellan beteendevetenskaper och datavetenskap. Vi betonar att VA är beroende av människor för att uppnå integrerad kunskap, för upptäckter och insikter. Den kreativa process som stöds av visualiseringar är normalt inte automatisk utan förlitar sig på mänsklig förståelse, upptäckter och insikter. Det behövs därför träning i att tolka representationer och använda visualiseringsverktyg för att uppnå insikter och situationsförståelse. Målgruppen för denna rapport är FOI-forskare som behöver veta mer om visualisering.

Visualiseringsverktyg och visuella representationer är i många fall djupt rotade i arbetsprocesser i organisationen och har utvecklats organiskt när människor lär sig att använda VA-verktyg för att utföra och kommunicera sitt arbete. Nya medarbetare behöver utbildning i att förstå och använda de visualiseringar som används i organisationen.

Metoderna för visuell dataanalys kan användas för, men är inte beroende av, de stora samlingar av data i vissa organisationer och företag som stereotypt benämns "Big Data".

Kapitel 2 undersöker de senaste tre årens utveckling inom visuella representationer och interaktionstekniker. Det visar att kända riktlinjer från de mest ärevördiga experterna i visuell dataanalys inte alltid vilar på vetenskaplig grund och i minst ett rapporterat fall står i konflikt med den senaste psykologiska forskningen.

Kapitel 3 handlar om interaktiv visualisering av flerdimensionella data. Många tillämpningar lider av "dimensionsförbannelsen", vilket innebär att antalet alternativ att presentera visuellt är överväldigande stort. Detta är oftast på grund av en så kallad "kombinatorisk explosion", där varje nytt ja- eller nej-alternativ som läggs till beslutsutrymmet fördubblar antalet beslutsvägar att analysera. Kapitel 3 presenterar två huvudsakliga metoder för att hantera detta i visualisering: 1) behandla alla dimensioner lika i syfte att utforska beslutsrymden utan förutfattade meningar, och 2) att minska antalet dimensioner genom att tillämpa matematiska metoder såsom exempelvis principalkomponentanalys (PCA) i syfte att göra viktiga val synliga bland röran av alternativ. För vart och ett av dessa metodologiska grenar granskas flera olika VA-metoder.

I många ledningssituationer inom affärsvärlden, produktion, logistik och militärväsendet kämpar ledare med att hantera ett mycket stort antal alternativ. Det är ofta svårt att göra en fullständig uppsättning handlingsalternativ begripliga för ledare under sådana omständigheter. Helst bör visuella representationer användas för att orientera chefen i den täta undervegetationen av

handlingsmöjligheter och möjliga utfall. *Kapitel 4* handlar om hur man löser denna typ av problem genom visuell dataanalys och vi finner att forskarsamhället har lagt liten vikt vid denna problemtyp. Trots detta definieras i kapitel 4 problemet med att visuellt representera flera alternativ och pekar på möjliga lösningar och forskningsinriktningar.

Data fångas från interaktion i användargränssnitt, genereras av simuleringar eller genom mätningar och är därför aldrig en exakt bild av verkligheten. Data är osäkra på grund av användarfel, orealistiska simuleringsmodeller och mätfel. Denna osäkerhet skulle kunna göra visuella representationer av data vilseledande. *Kapitel 5* behandlar osäkerhet i VA med fokus på visualisering av osäkerhet och de metoder vi behöver för att göra användaren medveten om osäkerheter i underliggande data – och i de visuella representationerna. Problemet är att osäkerheten är en extra dimension att visualisera i de uppgifter som ofta redan är tyngda av flerdimensionella data, så som beskrivs i kapitel 3.

Vi går härefter mer i detalj in på ett illustrativt exempel på en praktisk tillämpning av VA. Numerisk strömningsmekanik (CFD) är en grundläggande fysikalisk vetenskap som är viktig i processer så som fordonskonstruktion, konstruktion av jetmotorer eller fartygsskrov. Strömningsmekanik är också avgörande för många försvars- och säkerhetstillämpningar, inklusive förståelse för spridning av farliga ämnen och för undervattenstillämpningar. *Kapitel 6* handlar om hur visuell dataanalys används i verksamhet baserad på CFD och visar därmed hur VA fungerar i verklig forskning. Det bör påpekas att VA är djupt integrerat i forskarnas arbetsprocess och särskilt i efterbehandlingen där data från en CFD-simulering analyseras och jämförs med experimentella data i syfte att förstå resultatet. Simuleringsresultat visualiseras i syfte att hjälpa forskarna till bättre insikter som kan leda till nya lösningar på det aktuella problemet. Kapitel 6 diskuterar vidare hur många organisationer har utvecklat sina egna visualiseringsverktyg och representationer och nämner hur viktig programvara med öppen källkod ofta är.

Författarnas sammanfattande synpunkter på visuell dataanalys finns i *kapitel 7*, där vi påpekar att intelligens och kreativitet i VA- processen kommer från människor och att teknik så som virtuell verklighet kommer att vara viktig för utvecklingen av VA. Vidare beskriver vi tillämpningar av VA för försvarstillämpningar. I sektion 7.4 visar vi att läsaren med fördel kan betrakta varje kapitel i denna rapport som grönt ektoplasma som väller ut ur högdimensionella sprickor i väggarna i det stolta bygget som är visuell dataanalys.

Nyckelord: Visuell dataanalys, Visualisering, Interaktiv visualisering, Datarepresentationer, Informationsvisualisering, Vetenskaplig visualisering, Användargränssnitt, Interaktionsteknik, Interaktionsdesign, Numerisk Strömningsmekanik, Simulering, Stora datamängder, Osäkerhetsvisualisering, Flerdimensionella data, Dimensionell reduktion, Projektionsmetoder, Flervalsvisualisering, Beslutsstöd, Externa representationer, Fysiska datarepresentationer, Haptik, Kraftåtermatning, Fysiska användargränssnitt.

Summary

In the Introduction in *chapter 1,* we define Visual Analytics (VA) as *the science of supporting human reasoning and sense-making via visual representations in which interaction is an essential part of the analysis process*. Furthermore, we note that Visual Analytics as a Science is in the borderland between the Behavioral sciences and Computer science. We emphasize that Visual Analytics depends on humans for integrating knowledge, for discovery and insights. The creative processes that are enabled by visualizations are typically not automated but rely on human individuals or teams, trained in both visual literacy and the use of visualization tools, to achieve insights and sense-making. The target audience of this report is FOI-researchers in need of knowledge about visualization.

Visualization tools and representations are in many cases deeply engrained in the work processes of the organization, and have evolved organically as people learn to use Visual Analytics tools for performing and communicating their work. New team members need training in understanding the visualizations that are used in the organization.

Visual Analytics processes can use, but are not dependent on, the vast collections of data in some organizations and businesses that stereotypically are referred to as “Big Data”.

Chapter 2 surveys the last three years developments in visual representations and interaction techniques. It demonstrates that age-old wisdom stemming from the most venerable authorities in Visual Analytics is not always based on scientific evidence, and in at least one reported case stands in conflict with recent psychology research.

Chapter 3 is about Interactive visualization of multidimensional data. Many applications suffer from the “curse of dimensionality” which means that the number of options to investigate is overwhelmingly large. This is typically due to a so called “combinatorial explosion”, caused by each new yes/no choice added in exploring a decision space doubles the number of options to analyze. Chapter 3 reviews two main methods for handling this in visualization: 1) treating all dimensions equally for the purpose of exploration and 2) reducing the number of dimensions by applying mathematical methods such as principal component analysis (PCA) for the purpose of making the most important choices visible among the clutter of options. For each of these main methodological branches several data processing and visualization options are reviewed.

Many management situations including business, production, logistics and battle management include a very large number of options. It is often hard to make the full set of options comprehensible to managers in such circumstances. Ideally, Visual Analytics should be employed to guide the manager in the dense forest of

action opportunities and possible outcomes. *Chapter 4 Effective visualization of multiple options* addresses this challenge, finding that the Visual Analytics research community has given it little emphasis. In spite of this, Chapter 4 defines the problem of visually representing multiple options and points to feasible solutions and research directions.

Data is captured from interactions with user interfaces, generated by simulations or by measurements. Data is never a precise reflection of reality but is fraught with uncertainty, which could make visual representations of the data misleading. *Chapter 5* handles Uncertainty in Visual Analytics focusing on *visualization of uncertainty* which is the methodology that we need for making the user aware of uncertainties in the underlying data. Chapter 5 reviews appropriate visualization methods for showing the degree of uncertainty in the data, and also reflects on uncertainty in the representations. The level of uncertainty constitutes an extra dimension in the often already multidimensional information, thereby requiring the use of methods for visualization of multidimensional data, as described in Chapter 3.

Computational Fluid Dynamics (CFD) is a basic Physical Science underlying much of the design of vehicles, jet engines or ship hulls. CFD is also crucial for defense and security applications, including understanding contaminant flows and underwater applications. Our *chapter 6* on Visual Analytics in CFD illustrates how VA is used in applied research. It is pointed out that Visual Analytics is deeply integrated in the work processes, particularly in the post-processing stage where data from a CFD-simulation is analyzed and compared to experimental data for the purpose of understanding the result. Simulation output is visualized for stimulating human insight. Furthermore, chapter 6 discusses how many organizations have developed in-house visualization tools and representations and mentions the important role of open-source software.

The authors’ final opinions on Visual Analytics are provided in *Chapter 7 Discussion and Conclusions* in which we point out that intelligence and creativity in the VA-process is supplied by humans and that 3D Virtual Reality displays will be important for the evolution of VA. Furthermore, we describe applications of VA for defense. In section 7.4 we argue that the reader advantageously can view each chapter in this report as green ectoplasm seeping out from high-dimensional cracks in the shiny walls of the edifice of Visual Analytics.

Keywords: Visual Analytics, Visualization, Interactive Visualization, Data Representations, Information Visualization, Scientific Visualization, User Interfaces, Interaction Techniques, Interaction Design, Computational Fluid Dynamics, CFD, Simulation, Big Data, Visualization of Uncertainty, Multidimensional Data, Dimensional reduction, Projection Methods, Visualization of Multiple Options, Decision Support, External Representations, Data Physicalization, Haptics, Force Feedback, Tangible User Interfaces.

Contents

1 Introduction

One-year-old Emile Ouamouno died in December 2013. Shortly after, his sister, mother and grandmother also died. This marked the onset of the recent Ebola epidemic in Africa. Local health authorities and the World Health Organization (WHO) thereafter faced the task of tracking and containing the epidemic that eventually claimed over 10.000 lives.

Visualization of databases as in Figure 1 proved to be an essential instrument for understanding and containing the disease. By generating and interacting with

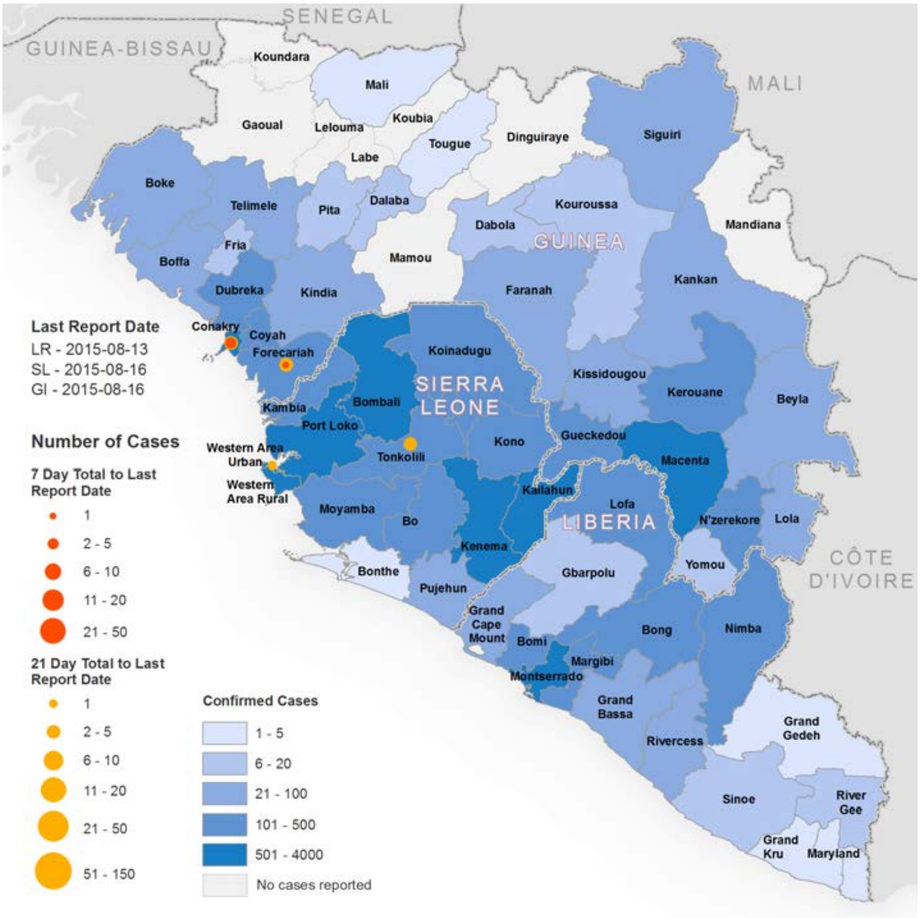


Figure 1. Distribution of new and total confirmed Ebola cases, 19 August 2015. From the World Health Organization website, http://www.who.int/csr/disease/ebola/maps/en/

visual displays of the outbreak, decision makers could make sense of the situation and focus containment and health-care resources on the proper targets.

"… outbreaks rarely have only local or regional consequences in our highly interconnected and interdependent world," claims one of the most prominent leaders of the containment effort, Dr, Margaret Chan of WHO.

The science behind the tools used by experts, researchers and decision-makers to visualize, analyze, understand and communicate vast and/or complex data is the topic of this report.

This section defines where Visual Analytics (VA) belongs in the taxonomies of applications and research fields.

Thomas and Cook (2005) define Visual Analytics as *the science of analytical reasoning facilitated by interactive visual interfaces*. Applications of Visual Analytics (see Figure 2) are often partitioned as information visualization or scientific visualization where the former is about visualization of abstract facts about e.g. organizations, economics, communication and software structures.

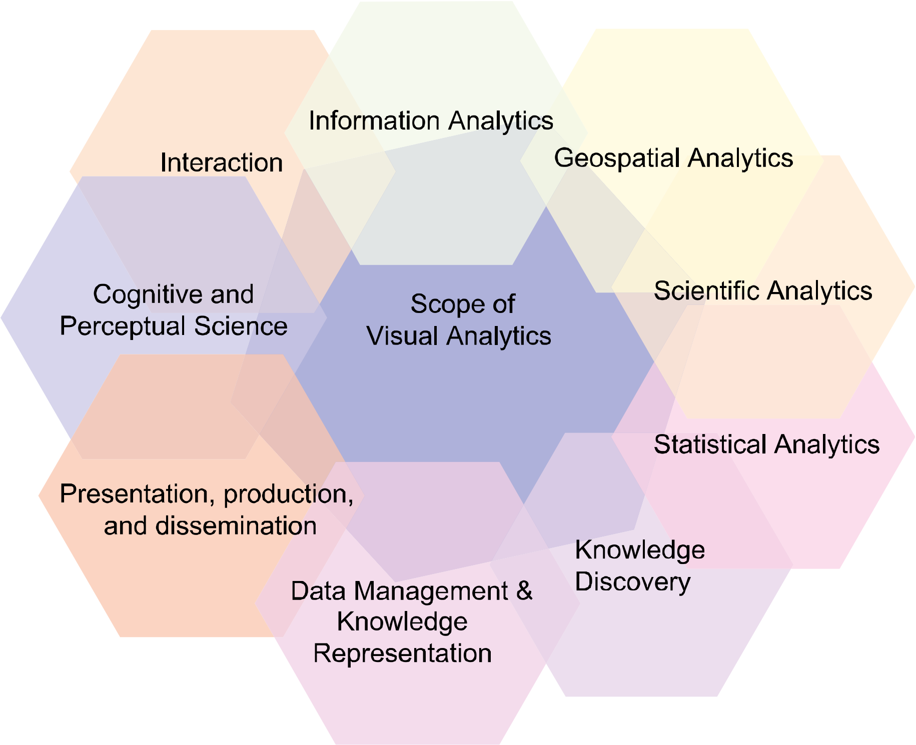

Information Analytics
Interaction
Geospatial Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Scientific Analytics
Statistical Analytics
Presentation, production, and dissemination
Knowledge Discovery
Data Management & Knowledge Representation

Figure 2. Further details about the application scope of Visual Analytics. Adaptation from figure in (Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008).

Scientific visualization is taken to be about physical phenomena that can be displayed in space-time dimensions as for example hydrodynamic flows.

In Figure 3 we furthermore define how Visual Analytics fits into the taxonomy of sciences. Since Visual Analytics is about supporting human reasoning it overlaps with behavioral sciences including Psychology, Neuroscience and in particular the field of Perception research. In chapter 2 we exemplify how psychology lab studies are essential for confirming or falsifying VA methods. In addition, Visual Analytics is to a great extent a part of Computer science including subfields such as computer graphics and user interface design. It shares this position in-between behavioral science and computer science with the fields of Human-Machine Interfaces (HMI) and Human-Machine Interaction (HMI).

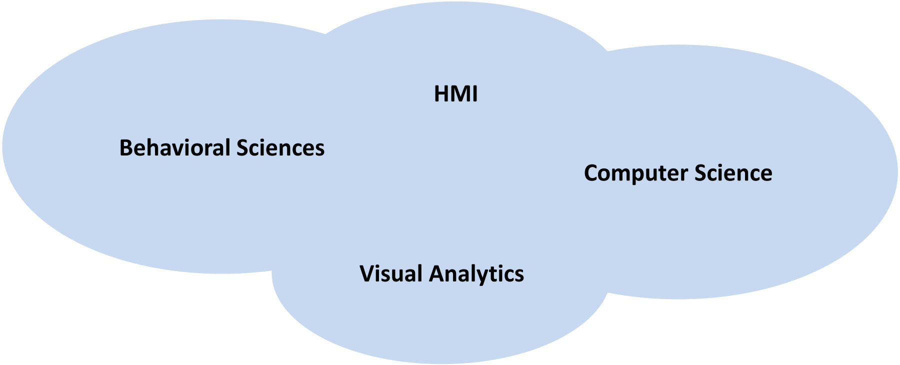


Figure 3. The research landscape of visual analytics.

1.1 Our Interpretation of Visual Analytics

Given the extensive scope of Visual Analytics and the opportunities for broadening the range of the field we have selected to follow Thomas and Cook (2005) and restrict our study *to support of human reasoning and sense-making via visual representations in which interaction[1] is an essential part of the analysis process.*

[1] By interaction we mean human-computer interaction in computer science roughly as described in http://en.wikipedia.org/wiki/Human%E2%80%93computer_interaction

1.2 Broader Definitions of Visual Analytics

Visual analytics can and have been generalized to include other sensory faculties than vision thus encompassing human-machine interaction via hearing, haptics (tactility and proprioception[2]), olfaction and nociception[3]. Visual representations could also support reasoning even if no interaction is employed as a part of the process.

1.3 The Process of Visual Analytics

Figure 4 shows an outline of the pipeline for Visual Analytics. If expanded the process can be described as consisting of the following steps:

1) Compiling data guided by the present knowledge of the situation.
2) Transformation of the data to a standard format in preparation for visualization.
3) The user defines filters for describing essential aspects of the data. In the Ebola case a filter could be a parametrized data structure representing the health status of a province or a country as related to the known phases describing how an epidemic infection spreads.
4) Data mining for building content that will support the visualization. This step typically follows procedures from statistical analysis or machine learning.
5) Feeding the visualization interface with filtered and processed information.
6) User interaction in which knowledge is extracted by the user. The next phase of hypothesis formation and visualization is guided by insights gained by the user.
7) Repetition of steps 1-6 for the next round of data transformation, model-building, visualization and user interaction.
8) Further user interaction supported by refinement of filtering parameters and possibly by further compilation of data based on new insights gained in the process.

[2] Sensory perception of bodily position, such as relative joint movements.
[3] The brain’s processing of pain and other sensory data related to harm to own body

The interactive analysis process presumes a fairly quick response of the system even in cases where the user interaction may trigger very complex database queries.

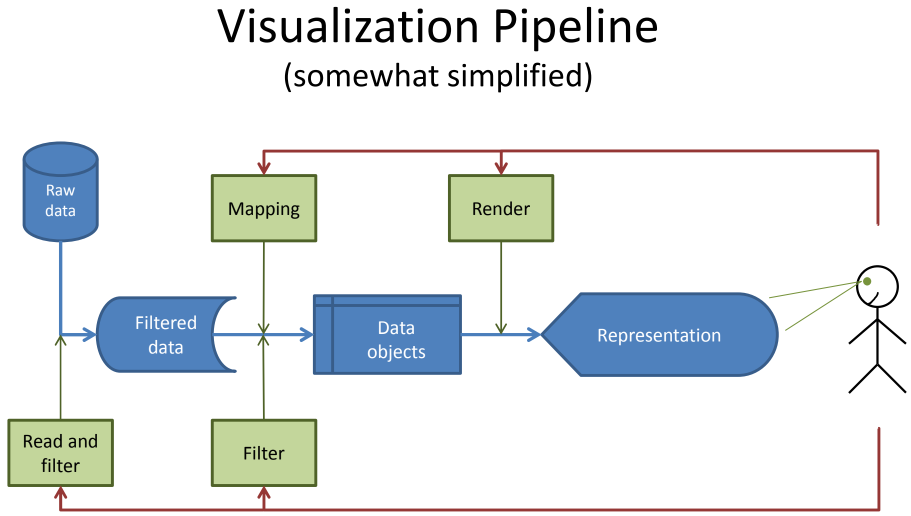


Figure 4. The visualization pipeline as presented in most visualization courses.

1.4 Visual Literacy

When working with visualizations it is important to learn how to read them – to develop a visual literacy. The visual literacy has to be considered in design of visual representations, and these considerations go beyond the mere perceptual, stretching into cultural and subjective notions of meaning (Byrne, Angus, & Wiles, 2016). The issue of how novices make sense of unfamiliar visualizations has been studied by Lee et al. (2016) who managed to map the process to five major cognitive activities. Furthermore, Ruchikachorn and Mueller (2015) showed that visual literacy about representations can be improved by visual morphing. In their work they presented the user with a view of a familiar representation that morphed into an unfamiliar, thereby making it possible to see an analogy between the two.

In addition to visual literacy, spatial ability also has a very large impact on the viewer's ability to comprehend, extract meaning and learn from visual representations (Ottley, et al., 2016; Bivall, 2010). Both visual literacy and spatial ability are topics of research in their own right. However, dwelling deeper into visual literacy or the field of spatial ability is beyond the scope of this report.

1.5 Mapping the Field of Visualization

Traditionally the field of visualization has been divided into two major areas: *Scientific visualization* and *Information visualization*. Representations where the dimensions of the representation have a natural mapping to the dimension of the measurement or simulation, most frequently three or four dimensions, have normally been sorted into scientific visualization. Examples are medical images such as computer tomography or magnetic resonance imaging data, or representations of fluid dynamics such as those described in chapter 6. Representations of abstract, sometimes multi-variate data, numerical or categorical data, where there does not necessarily exist a natural mapping to the dimensions of the representation, end up under the umbrella of information visualization.

This dichotomy has been prevalent but is becoming obsolete. The field of visualization has matured, and the types of data as well as representations have become more mixed and methods of data analysis are more often shared across the former boundary. In this report we unify the fields by using the term Visual Analytics.

Another dichotomy could be applied to the field of visualization through a divide based on the different usages of presentation/communication and data analysis. One indication of the power of visualization for communication is the increasing number of infographics present in almost every magazine or newspaper. External representations can sometimes clarify rather complex relations or rapidly put information into a context, such as information related to populations rendered on maps in different shapes and forms.

While the power of visualization as a means for communication is widely recognized, visualization for data analysis on the other hand, despite its potential, is left untouched by many scientists outside the field of visualization. Alternatively, as touched upon in parts of this report, the concept of visual data analysis is over-exploited as the remedy to all corporate business analysis needs.

It is important to note that the Visual Analytics process normally is not automated and driven by intelligent software. The intelligence, creativity and insights must come from humans who are trained in using the appropriate software tools and the various visualization options and that are well versed in the domain under study. Designers of underlying models can also be expected to encode human knowledge and visualization traditions in the software that generates visual representations.

Sometimes teamwork takes the form of a staff preparing briefings for managers using visual representations. The analytics team must understand the needs of the decision makers that they are serving and guiding. The ability to foresee the best decision and present visualizations that guide the decision maker in the right direction is an essential skill for the analysts but can also be construed as a threat against the independence and integrity of the decision makers. The precise boundary between adept visualization and skillful manipulation is difficult to draw.

Visual analytics is often described in a language that seems to imply that Visual Analytics is a subject who acts, achieves and thinks. For example:

"Visual analytics integrates the analytic capabilities of the computer and the abilities of the human analyst, thus allowing novel discoveries and empowering individuals to take control of the analytical process. Visual analytics sheds light on unexpected and hidden insights, which may lead to beneficial and profitable innovation."[4]

The reader of this text could easily be lead to believe that there is some process or software that provides automated intelligence although in reality human workers must integrate, discover and provide insights. There is no automated AI that performs the analysis and presents results via a visual interface. Humans must provide creativity and insights while using the data processing tools.

Kohlhammer et al. (2011) is much clearer about the important role of human labor and thinking in the Visual Analytics process:

"visual analytics is the creation of tools and techniques to enable people to:

- *Synthesize information and derive insight from massive, dynamic, ambiguous, often conflicting data.*
- *Detect the expected and discover the unexpected.*
- *Provide timely, defensible, and understandable assessments.*
- *Communicate these assessments effectively for action"*

This is also in line with the definition by Card, Mackinlay and Shneiderman (1999) that visualization is about: *"the use of computer-supported, interactive, visual representations of data to amplify cognition"*.

[4] www.**vismaster**.eu/wp-content/uploads/.../**VisMaster**-book-lowres.pdf *preface, first paragraph.*

Both individuals and organizations wishing to make use of Visual Analytics can only make limited progress by simply downloading or acquiring a software toolkit. In addition they must be prepared to:

- Train themselves or others in Visual Analytics.
- Set up a Visual Analytics process related to their domain of interest.
- Integrate the Visual Analytics process in relevant internal processes including management and decision-making processes.

Visual Analytics is often described as an essential tool for making sense of the huge piles of data that accumulate everywhere in the modern world. Advocates of Visual Analytics speak about organizations' problems of making use of *Big Data* and how we all suffer from information overload. Visual Analytics is promoted as the remedy for these problems. In our view, the loudest advocates of VA promise more than they can deliver as they are driven, most likely, by commercial interests or the fierce hunt for research funding.

Through this report we promote the use of Visual Analytics but prefer also to inform and increase awareness of both its risks and opportunities. In summary, VA has great potential to increase the understanding of data and VA can be a valuable tool for analysis. Furthermore, visualization is a very powerful way to communicate. At the same time there are no free lunches, meaning that to fully exploit the power of VA the visualization tools are likely to require customization to fit the particular problem at the hand of the analyst. Caution is also needed to ensure that representations are not selected that only show the aspects of the data that the researcher/user already knows are present, thereby preventing discovery of new and unknown features in the data.

Hence we conclude that the Visual Analytics process and tools typically are firmly bound to and integrated with the domain of the application at hand. The tools and techniques for visualization in Computational Fluid Dynamics as described in chapter 6 would for example be quite useless for financial analytics and vice versa. Most organizations invent their own tools based on generic visualization tools or otherwise acquire a set of tools that meet the visualization needs of the organization. An important part of the on-the-job-training is to learn how to read and generate the visualizations in use by the organization and, as researcher, to also learn to identify the visualization needs and generate the tools required.

Selecting the appropriate representations does require knowledge about what type of data and/or features each representation can convey, which, in turn, require that the representation has been evaluated. Forsell, together with Johansson or Cooper (2010; 2010; 2012) have conducted thorough evaluations of their solutions for parallel coordinates plots, and also presented general methods for how to evaluate visualization tools. Their work has been an important contribution to the visualization community as many visualization techniques have been presented but frequently not proven through evaluation. It is not until recent years that evaluation has become a major topic in the field of visualization, nevertheless the questioning of both novel and established representations can be required. For example, the work of (Rubio-Sanchez, Raya, Diaz, & Sanchez, 2016) shows that the common way of using two representations (called RadViz and Star Coordinates) can actually introduce distortions with respect to the data.

1.6 Aim and Outline of the Report

The present study is the result of a field survey project aiming at increasing the knowledge of visual analytics at FOI. The target audience of this report is researchers in need of knowledge about visualization. The intention is not to provide a complete overview of the field of visual analytics, nonetheless we postulate that the report summarizes a substantial part of the available methods in the field, and thus also serves as a good starting point for readers with low prior knowledge of visual analytics.

At this point we want to explain the choice of subjects in the chapters that follows. We have focused on aspects that can be issues of concern in VA-applications. Although these aspects are present in academic visualization research, they are often lacking in overviews, especially by VA sales people. Hence the chapters discuss, uncertainty in data, issues related to the “curse of dimensionality”, the evidential basis of VA methods, multi-option decision-making and how VA is used in real-life research. We think that the reader is best served by getting information about subjects that are typically excluded from VA reviews and sales presentations. For a researcher in need of analysis tools or for someone considering making use of VA in their organization this should be of greater value than yet another uncritical overview and homage to the advantages of visualization.

2 Visual Representations and Interaction Techniques

Translating data into a visual representation that helps the user understand and explore the data content is one of the main challenges in visualization. In visual analytics the interaction with such representations is another important aspect, often vital to the exploratory process. The foremost aim for this chapter of the present report is to survey the most recent developments in visual data representations and interaction techniques, mainly within the last three years, to find novel techniques that create modes of visualization and interaction that go beyond scatter plots and drop-down selection boxes. Some old examples are provided for completeness and to put the reported findings into a context.

This chapter is also written as an attempt to inspire those looking for new ways to visualize their data. To this end, given the aim to inspire, all publications referenced have not been scrutinized to the level of a peer-review, meaning that publications providing interesting examples of visualizations have been included even if there might be issues of concern, such as lacking evaluation.

A secondary question for this survey, next to finding research of novel data representations, is to qualitatively assess how inventive the visualization community is when it comes to evolving data representations. Where is the visualization research focus put when striving to improve the ability to reach conclusions about information/data at hand? Is it: 1) To use existing visualization techniques such as tree maps, scatter plots, matrices etc. and how they can be combined, or, 2) To put emphasis on developing entirely new visualization techniques? Likely it is a trade-off between risk and award as the latter probably has a greater potential to become ground-breaking, but is harder to invent and certainly comes with a risk for failure.

2.1 Making a Visualization Appealing and Memorable

Edward Tufte is a recognized visualization researcher and has written some of the seminal books in the field, for example (Tufte, 1990; 2001). Many of the guidelines presented by Tufte are taught as good rules of practice in the majority of visualization courses around the World and cover, for example, how to avoid clutter or how to use glyphs and color. Although some of his guidelines are not proven effective through research, they often make sense intuitively, which is the likely reason for the wide acceptance in the visualization community.

One of the guidelines promoted by Tufte is the principle to avoid *chart junk*, that is, to avoid decorations, excessive annotations and similar visual elements. If not avoided, it is argued that such abundance of graphical elements might distract the reader from the data that should be shown as clearly as possible. At the same time it can be argued, from an intuitive view, that a more engaging visualization is also more effective as the user becomes more involved in understanding the data it represents. The use of chart junk is discussed by Borkin et al. (2013) in the context of existing psychology lab research supporting the traditional view, and other research showing improved retention from increased chart junk. This is relevant to the research of Borkin et al. (2013) as they focus on finding out what makes a visualization memorable.

The aim of the research presented by Borkin et al. (2013) is to investigate what factors make some visualizations *intrinsically* more memorable than others. They make a parallel to previous research on images of natural scenes, where some images have been found to be more memorable, independent of an individual's bias. Borkin et al. (2013) acknowledge that memorability does not necessarily have anything to do with comprehensibility of a visualization, but argue that it is a step towards finding out what makes visualizations engaging and/or effective. It should be noted that the experiments conducted include static images and not the type of interactive data analysis tools that are the main focus for this report.

Very briefly summarized, Borkin et al. (2013) identified properties that make some visualizations intrinsically more memorable. Perhaps not surprising, visualizations including photographs or other images of human recognizable objects were more memorable than those without. In addition, a higher number of colors increased memorability, especially compared to black and white. Furthermore, and very interesting when relating to the predominant views on chart junk, high visual density and low *data-to-ink ratios* increased memorability. Similar to chart junk, the data-to-ink ratio has been considered a measure of how effective a representation is, promoting minimal use of graphical elements (minimal amount of ink) to represent a dataset. In addition, despite our training to read bar charts and line graphs, such representations are also less memorable than tree-views or grid/matrix representations.

What are the implications of the study by Borkin et al. (2013)? The facts that comprehensibility was not part of the study, and that the focus of the report was on *static* representations rather than the *interactive* analysis tools, makes it difficult to answer the question. Even so, if designers of tools for Visual Analytics want the representations to be memorable, perhaps the findings indicate that pictograms such as icons should be used instead of text when an exchange is possible. Also, instead of using shades of a single color a more extensive palette could be applied, and the designers should perhaps not be too afraid of adding supporting elements to the representations simply by the argument of keeping visual density low and the data-to-ink ratio high. However, caution has to be taken both regarding the properties of colors and how they are perceived, and regarding cognitive aspects of load and attention. Further studies on these topics are required, and it would be interesting to see how such research could guide invention of novel data representations and creative means for interaction.

2.2 Fundamental Types of Representations in Visual Analytics

Browsing through visualization textbooks and other resources, it soon becomes clear that almost all modern tools for visual data analysis share a common set of representations, forming what can be seen as the base-line of data representations. This base-line is often extended to fit the intended application of each tool, a specialization normally required to make VA reach its full potential for a specific application or type of data. However, for many purposes, especially when beginning to use Visual Analytics, a little goes a long way. Except for the universal line charts and bar charts, what representations that constitute the common base-line set is not established, but a few are very common and therefore presented here as examples.

- Parallel Coordinates Plot (PCP), described in section 3.1.3 and Figure 23**,** is a major player in Visual Analytics and a part of every serious VA toolkit. A properly designed PCP makes it possible to rapidly see trends in the data, detect correlations between variables and filter data.

- Scatter Plot, described in 3.1.1 and seen in Figure 5, can combine location, color and point size, making it possible to map four dimensions of the data into one plot, and to see trends and detect outliers. With a 3D scatter plot, as seen in Figure 15, the plot can show five dimensions, although requiring more interactivity to explore the data.

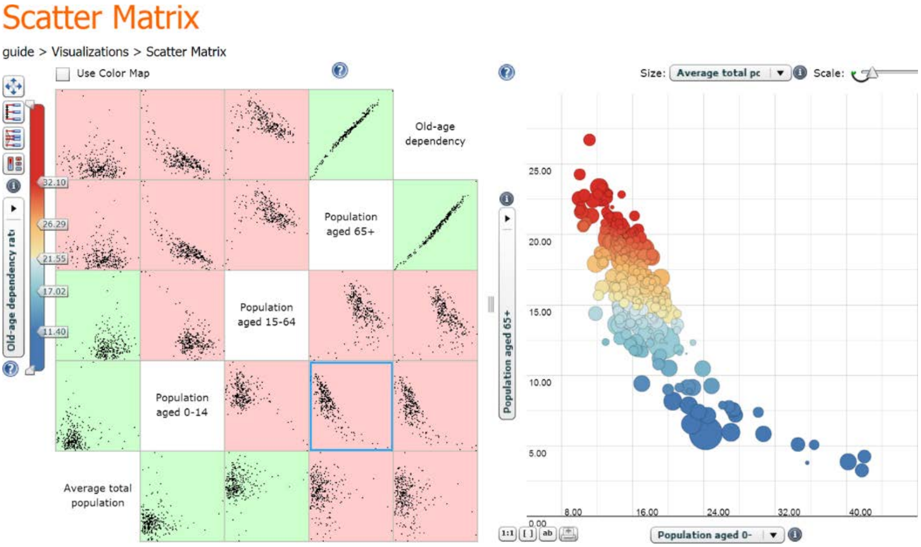


Figure 5. An example of a scatter plot matrix (left) and a scatter plot (right) providing a detailed view of the selected matrix element.

- Scatter Plot Matrix. By setting up a matrix with scatter plots for different variables it becomes possible to quickly analyze correlations between multiple variable pairs. The scatter plot matrix is often rendered using two variables for each matrix element, showing the relation between those two variables, ignoring the rest, see Figure 5.
- Table Lens. A way to graphically represent the values in each cell in a table, see Figure 6.
- Tree Map. Shows the data in a hierarchical manner where both structure and the size of each level is represented through size and/or color, see Figure 7.

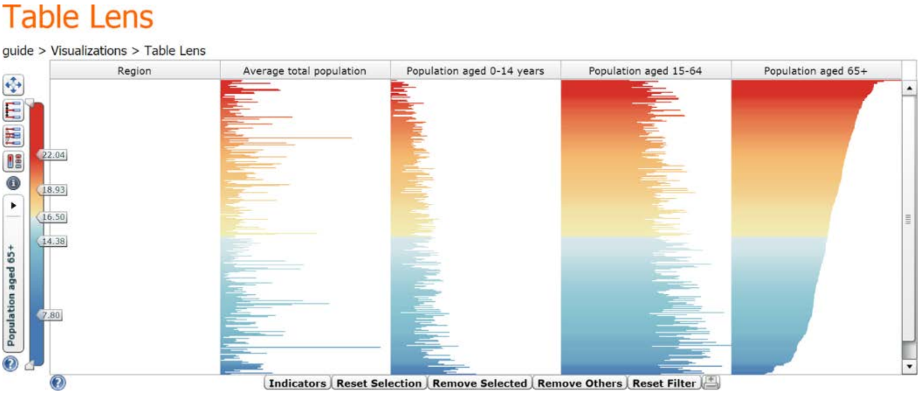


Figure 6. Example of a table lens representation.

It is not necessary to go to the very latest in visualization research to find creative data representations. The gallery for the D3 JavaScript library (Bostock, 2015) provides many examples of innovative ways to visualize data, some of which are shown in Figure 8 and Figure 9. The D3 library is one of several JavaScript-based libraries or toolkits that have emerged to enable use of interactive (or static) visualization on websites. Both commercial and free libraries are available, some provide specific visual data representations such as a 3D scatter plot, others are more generic for rendering of graphical components and handling of data structures. As previously pointed out, achieving the most efficient

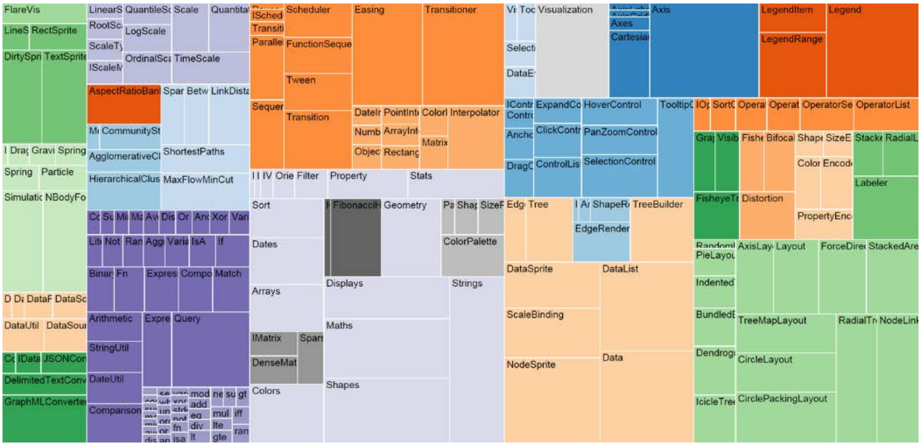


Figure 7. Tree Map picture. Generated by a D3-example found at http://bl.ocks.org/mbostock/raw/4063582/, accessed through the D3 gallery at https://github.com/mbostock/d3/wiki/Gallery.

visualization tools for a specific application normally requires customization of software. The JavaScript libraries can be recommended as a starting point, at least for prototyping. With a wide range of online resources and user communities, the libraries can generally be considered to have a fairly low entry threshold for someone with programming skills.

2.3 Representations Beyond the Basics

In VA, as in most fields of science, the corpus of knowledge grows slowly by small increments on previous work. Most representations found in the survey presented in this chapter are small but inventive extensions to their more basic counterparts, or clever combinations of existing techniques that enable novel ways to perform VA.

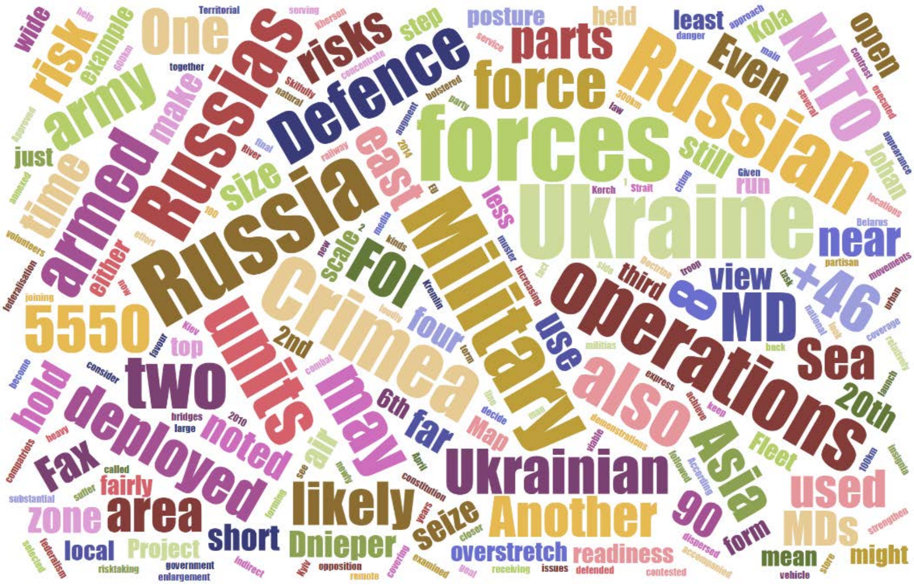


Figure 8. A word cloud based on a FOI report (Norberg & Westerlund, 2014). The size of a word represents how frequently it appears in the text relative to the other words.

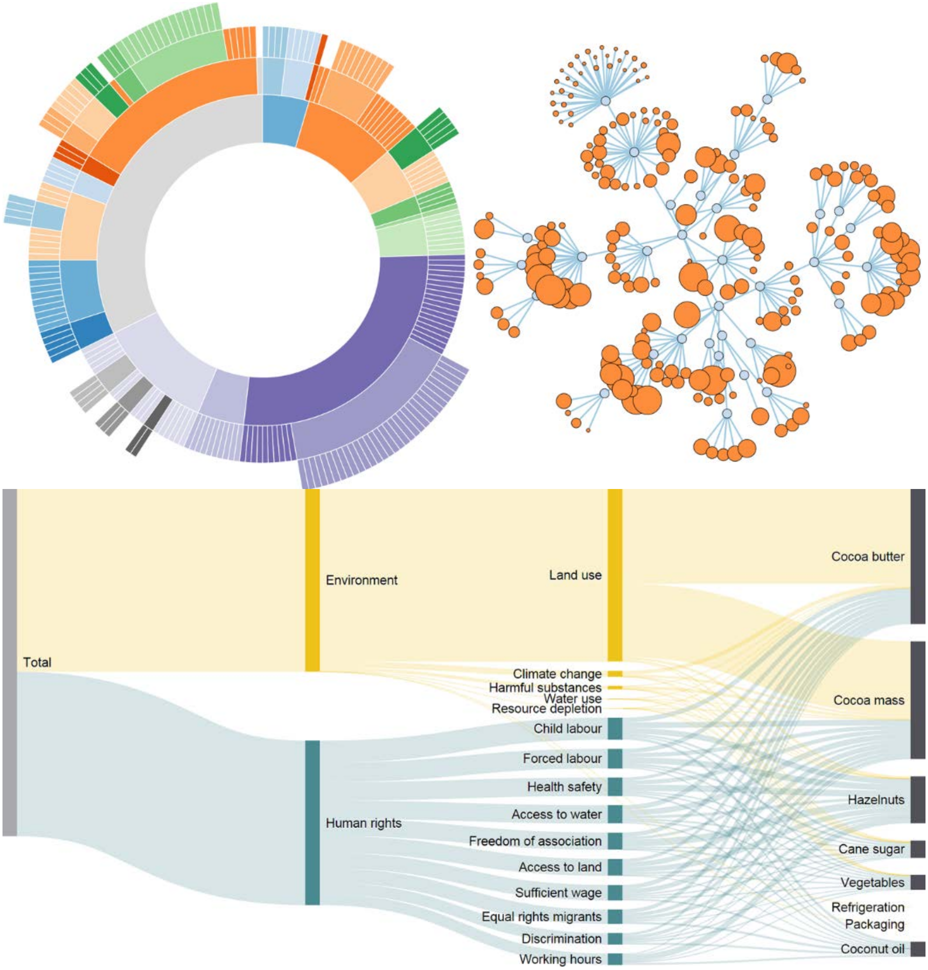


Figure 9. Examples of representations created with the D3 library, all different representations of hierarchical structures. Upper left: Sunburst, similar to a tree map but with a radial layout. Upper right: Force tree. Lower: A Sankey diagram.

The fact that the parallel coordinates plot and the scatter plot are brought up in three chapters of this report (see sections 3.1.1, 3.1.3 and 5.3.8 apart from this chapter) is an indicator of the roles PCP and scatter plots play in Visual Analytics. Therefore, the overview of recently published research will start with variants on those two representations. Johansson and Forsell (2016) present a survey on research with user-centered evaluation of parallel coordinate plots, a publication which also includes a wide range of different PCP examples, although those are left for the interested reader.

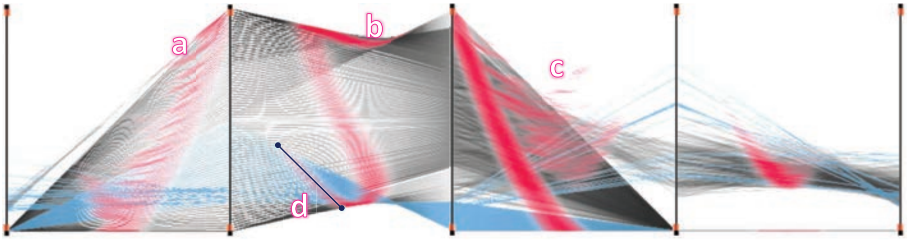


Figure 10. The OPCPs (red). (a) Visual enhancement of small patterns between the first two dimensions of the data, i.e., small structures obstructed by a strong pattern. (b) Facilitated identification of distinct patterns between the second and third data dimension. (c) Improved readability of outliers, i.e., low density information areas, in the representation. (d) Efficient and accurate selection (blue) of a specific data structure, using the proposed O-Brushing technique (dark line). © 2016 IEEE. Reprinted with permission, from Raidou et al. (2016).

One development of PCP is the Orientation-enhanced Parallel Coordinates (OPCPs) presented in the work by Raidou, Eisemann, Breeuwer, Eisemann and Vilanova (2016). The OPCPs was developed to improve pattern discernibility in cluttered PCPs and at the same time enhance outliers that might otherwise be undetectable in areas with sparse data, see Figure 10. They also present a customized interaction technique described in more detail in section 2.4. Increased clarity in the OPCPs is achieved by enhancing parts of each PCP line with respect to its slope.

Although inventive, the PCP version presented in the work by Zhou, Xu, Ming and Qu (2014) does not seem as promising as the OPCPs, at least not for large datasets. In their PCP the lines between the axes are made from text, making the lines act as their own labels. Despite applying a technique to minimize overlaps between lines by using different line curvature, which seems to be a good feature, the text soon becomes unreadable and clutter appears even worse than for the standard PCP.

Looking instead at scatter plots the work by Chen et al. (2014) addresses the relevant problem of over-drawing in scatter plots. This can be an issue when working with a single class of data, and potentially even worse when simultaneously presenting data of multiple classes in the same scatter plot. If using color to separate classes in a conventional scatter plot, occlusion can occur due to data from different classes being drawn on top of each other, see Figure 11. Manual reordering is required to see patterns from all classes, although such work is tedious and it can still be difficult to compare the patterns. The approach proposed by Chen et al. (2014) is to process the data and plot the scatter plot in a manner that keeps the point-distribution and relative density between the data

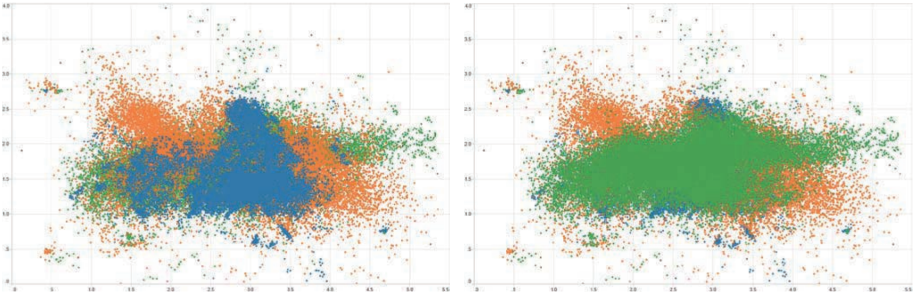


Figure 11. Manually changing drawing orders for the conventional multi-class scatter plot can reveal otherwise occluded patterns in the data. © 2014 IEEE. Reprinted with permission, from Chen et al. (2014).

classes, see Figure 12. The claim is that quantitative analysis can still be performed efficiently, although it would be interesting to study how the quantitative properties are conveyed in the processed version as compared to viewing the original data. However, it might not be possible to perform such a comparison on multiple classes of data.

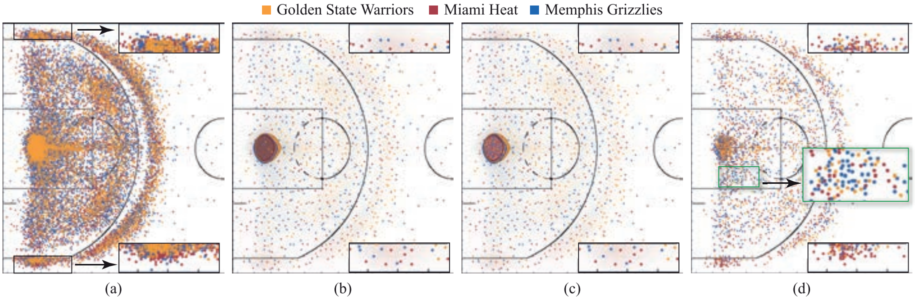


Figure 12. Figure from Chen et al. (2014). Different scatter plot techniques showing basketball (NBA) shooting data for three teams. (a) A conventional scatter plot. (b) Using the Splatterplot technique. (c) Using color weaving to enhance the perception of the data classes in dense regions. (d) The proposed method which should preserve the point distributions and the relative density orders among classes, and possible to use for quantitative analysis. © 2016 IEEE. Reprinted with permission.

It can be noted that Kay and Heer (2016) show that the scatter plot outperforms several other representations in the ability to convey correlation, both positive and negative. The result is also very consistent between the individuals participating in their study.

If the aim is to represent not only the same type of data for multiple classes, but also to include multiple attributes to be visualized simultaneously, another set of challenges arise. Cheng and Mueller (2016) approach those challenges by calculating similarity and correlation data to visualize the relationships between both data and attributes. This relates to the issues discussed in chapter 3, although here we focus more on the representation, the visual Data Context Map. To explain how the representation can be applied we provide an adaptation of the use case presented by Cheng and Mueller (2016), to be read while viewing Figure 13.

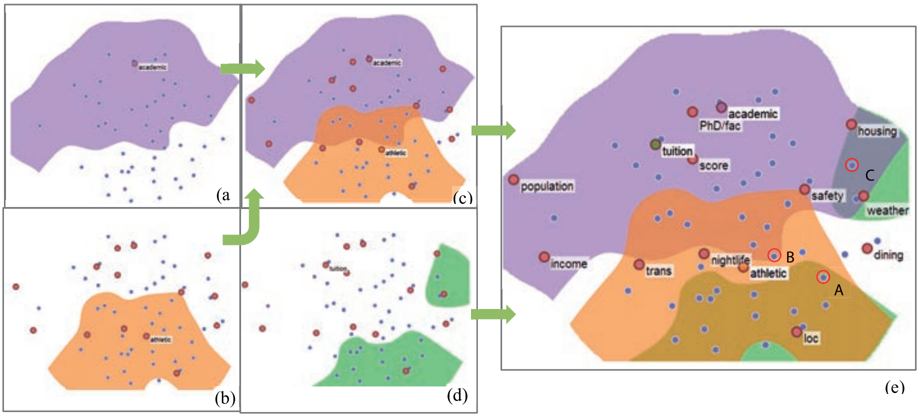
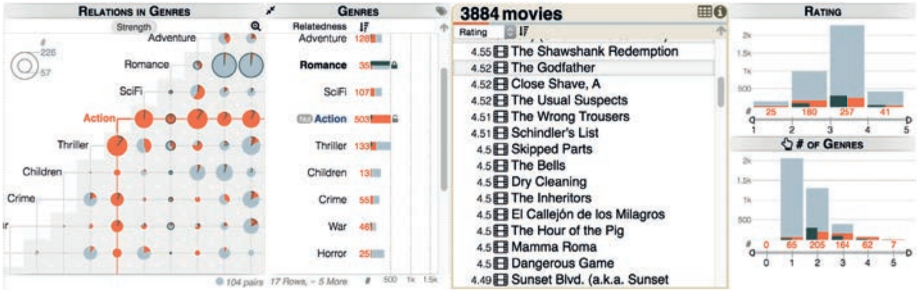


Figure 13. Example of the Data Context Map showing university data. (a) good academics region (>9). (b) good athletic region (>9). (c) combined region by (a) and (b). (d) low tuition region(<$18,000). (e) combined region by region (c) and (d). © 2016 IEEE. Reprinted with permission, from Cheng and Mueller (2016).

Tom is looking for a university and investigates his university data as he aims for a school that has high athletics (>9), high academics (>9), but low tuition (<$18,000). He begins by generating the decision boundaries based on these three criteria, shown in Figure 13 (a)-(d), after which he merges and gets (e). Unfortunately, there is no university that can satisfy all three criteria at the same time, easily detectable as the three regions do not overlap simultaneously. In the figure the red points with labels represent the attributes, while the small blue points represent the universities. Universities that locate close to a given attribute node have high values for the attribute, and correspondingly, those that locate far away have a low value. This leads to the process of selecting candidates in the areas where two regions overlap, and as close as possible to the attribute node from the third region, these are marked with A, B, and C in Figure 13 (e).

Interestingly, the final choice is up to the user, Tom, demonstrating the high degree of influence from the human in the loop in Visual Analytics.



4.55	The Shawshank Redemption
4.52	The Godfather
4.52	Close Shave, A
4.52	The Usual Suspects
4.51	The Wrong Trousers
4.51	Schindler's List
4.5	Skipped Parts
4.5	The Bells
4.5	Dry Cleaning
4.5	The Inheritors
4.5	El Callejón de los Milagros
4.5	The Hour of the Pig
4.5	Mamma Roma
4.5	Dangerous Game
4.49	Sunset Blvd. (a.k.a. Sunset

Figure 14. Interface for exploration of a movie dataset with multiple genres (sets) and ratings using AggreSet. Aggregate histograms are used for set-list and set-degrees, whereas the aggregate matrix (left) is used for set-pair intersections. The gray distributions visualize the number of elements per aggregate. © 2016 IEEE. Reprinted with permission, from Yalçin et al. (2016).

Visually the Data Context Map developed by Cheng and Mueller (2016) looks like a scatter plot, and while sharing a lot of features the Data Context Map still brings novelty into the representation. However, not all problems require development of new visual representations. Sometimes solutions are found by using inventive and creative combinations of existing methods for visualization and/or data processing. This is the case for AggreSet by Yalçin, Elmqvist and Bederson (2016) where a combined set of visual representations is applied to allow for exploration of sets and their relations using multi-valued attributes, such as genres per movie or courses per student, see Figure 14. Data in the sets are aggregated with the aim to allow for easier investigations of relations between the sets. One argument for the presented approach is to achieve consistency in user interface design, thereby avoiding separate ways of interaction when exploring sets as compared to exploration of non-set data.

Events that occur over time is a topic for study in different areas, making VA tools with support for exploration of time stamped data potentially applicable in many fields of research. Although challenging, techniques for visual analysis of time-stamped data have been

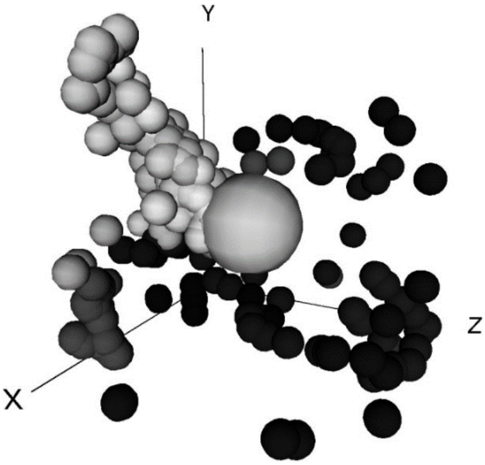


Figure 15. A 3D-scatterplot generated from time-stamped user interaction log files. Time is mapped to the color density (black to white) and position is based on the user's location in virtual space. From Bivall (2010).

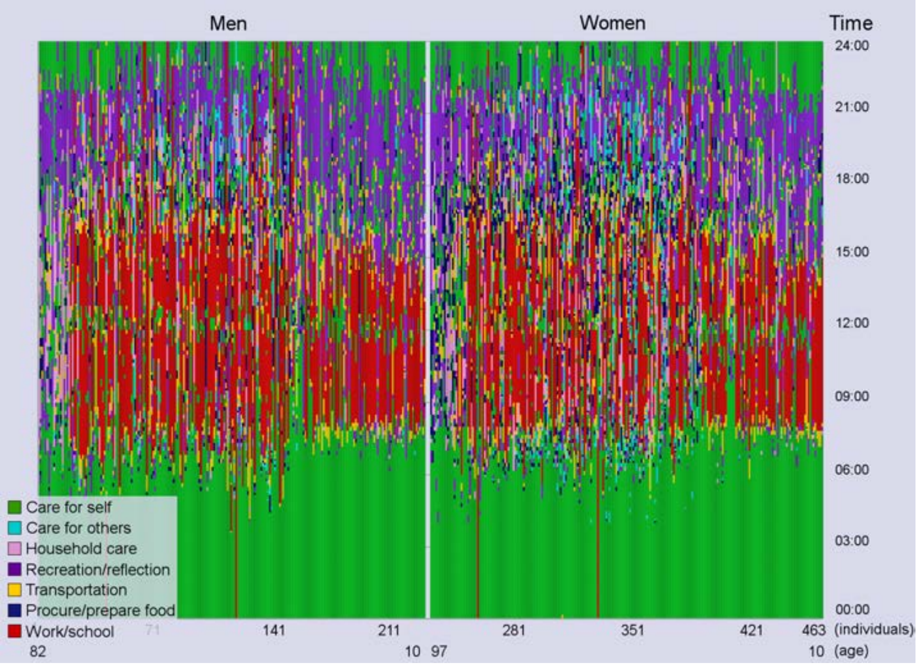
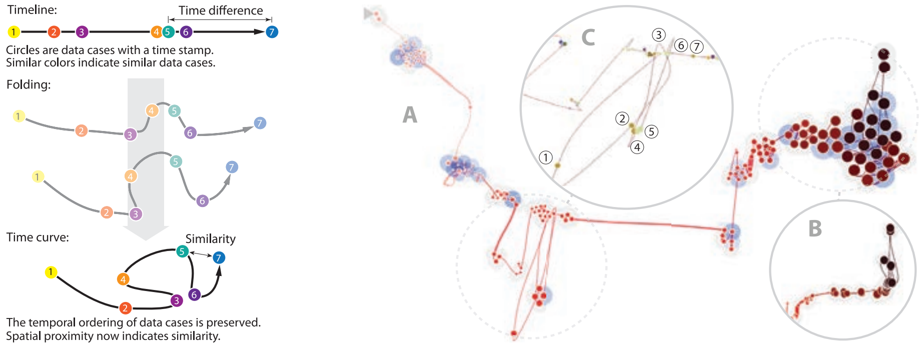


Figure 16. Representation showing the distribution of categorized activities as they occur during a day for 463 individuals. Used under permission from Katerina Vrotsou.

developed, see for example the straight forward 3D-scatter plot in Figure 15, and Figure 16 from the more extensive event sequence visualization work by Vrotsou (2010) in the field of time geography.

Time Curves, shown in Figure 17, is another representation considering the time-domain. It is developed by Bach et al. (2016) for analysis of temporal events and their similarity. The method is fairly generic in that it can be applied to any data with time-stamped entries, as long as a similarity metric can be determined to compare the entries. One example presented by Bach et al. is the evolvement of an article on Wikipedia, where similarity metrics can be calculated based on how much the text has changed between entries.


Timeline:
Time difference
1
2
3
4
5
6
7
Circles are data cases with a time stamp.
Similar colors indicate similar data cases.
Folding:
Time curve:
Similarity
The temporal ordering of data cases is preserved.
Spatial proximity now indicates similarity.
A
B
C

(a) Folding time

(b) History of the Wikipedia article on Palestine

Figure 17. (a) The principle of the Time Curve representation. (b) A specific example showing the Time Curve for an article from Wikipedia. © 2016 IEEE. Reprinted with permission, from Bach et al. (2016).

2.4 Interaction Techniques

Interactivity is a cornerstone in the process of Visual Analytics, and interactivity requires some form of interaction technique to put the human in the loop (see section 1.3). In this section, we provide examples of interaction techniques such as different ways to perform selections or navigating and comparing data. In addition to the normal modes of interaction we also report on some examples that go beyond keyboard and mouse.

Selection might seem rudimentary; however, the operation of selection must be put into the context of the data and the visual representation. For researchers working with point- or particle-based data, such as some astronomical measurements, the selection and segmentation of structures are tedious tasks. This is partly due to the size of the data, and partly due to its complexity with occlusion of structures. Visually the situation can be compared to the multi-class scatter plot shown in Figure 11, with the complications that the data is in 3D and has not yet been assigned its classes. To support the selection process Yu, Efstathiou, Isenberg and Isenberg (2016) developed the Context-Aware Selection Techniques (CAST) for analysis of large particle datasets. CAST consists of lasso-style selection, line drawing selection and point-click selection, and corresponding supporting algorithms, see Figure 18. The selection algorithms can help in selecting the appropriate parts of the point cloud, including partially

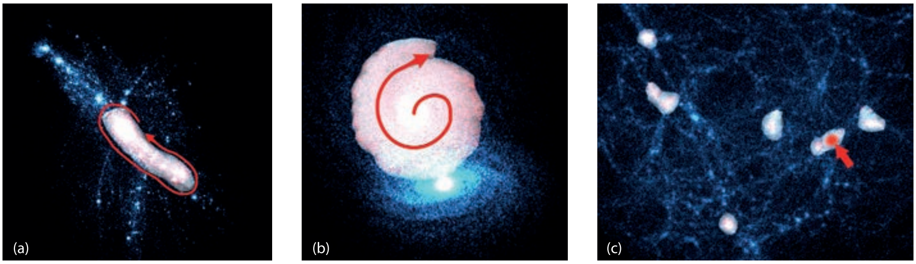


Figure 18. Selection techniques from Yu et al. (2016). (a) Lasso-style selection (named SpaceCast). (b) Selection by line-drawing (named TraceCast). (c) Selection by point-click (named PointCast). © 2016 IEEE. Reprinted with permission.

obscured structures. Yu et al. (2016) show that the CAST methods are both effective in that the selection goals are achieved, and efficient as they are faster than the other tested techniques.

Although the CAST methods might work for selections in some situations with 3D-scatter plots, they are not likely to work with PCPs. The current state of the art in selection (or brushing) for PCPs is reported by Raidou et al. (2016) and is shown in Figure 19. In their work with OPCPs (see section 2.3) they also developed Orientation-enhanced Brushing (O-Brushing), where the selection tools utilize additional data generated for the OPCPs. It is claimed that using O-Brushing reduces the need for user interaction, making selection more efficient, but their tests also show that the use of OPCPs requires more training compared to the normal PCP.

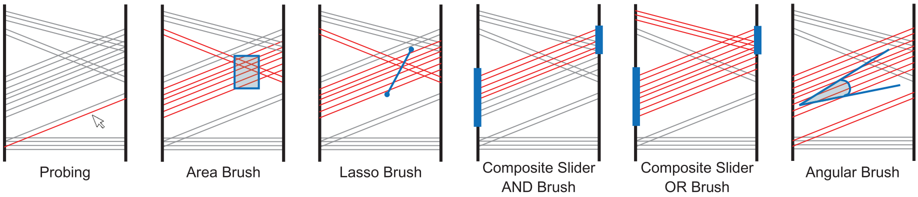


Figure 19. Brushing (selection) techniques for interaction with a PCP, from Raidou et al. (2016). Red denotes the resulting selections in each case, blue denotes the operation. © 2016 IEEE. Reprinted with permission.

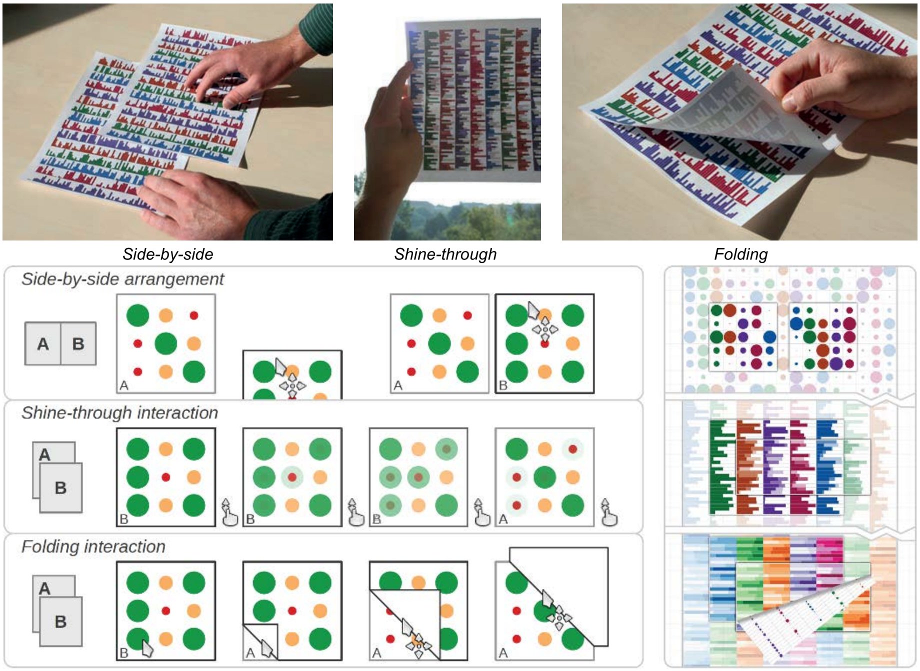


Figure 20. The physical (top) and virtual (bottom) versions of the interaction techniques proposed in the work by Tominski, Forsell and Johansson (2012). © 2012 IEEE. Reprinted with permission.

A common approach in graphical user interfaces is to create interaction metaphors based on a real world tool or process. One example is the looking glass metaphor often used for zooming. Mimicking real world interaction behavior was the foundation for the work by Tominski, Forsell and Johansson (2012) when they designed representations and interaction techniques. By studying how a type of data used to be examined when printed on paper they created corresponding ways to interact with the representation, and also explored some additional features becoming available in the virtual form. The real-world paper interaction techniques and illustrations of their virtual counterparts are presented in Figure 20.

2.5 Physical Representations and Interfaces

Today's human-computer interactions, which are beyond the visual and keyboard/mouse interface, have become possible partly because of the development and drop in price of technology such as touch-screens, sensors, and equipment like actuators and other hardware providing force feedback. These developments have enabled implementation of more physical interaction and representations at a reasonable price, even in public environments such as museums.

The magnifying glass analogy is a fairly common technique when adhering to the principle of providing the user with simultaneous focus and context (or details and overview). Using this analogy, an overview of the data can be examined by moving the magnifying glass across the representation, thereby getting a view that zooms in on a part of the data and providing details.

In the exhibit *Plankton Populations*[5] at the Exploratorium in San Francisco, the virtual magnifying glass is moved (back) into the physical world to present an alternative way of exploring data, while maintaining the principle of focus and context (see Figure 21). The exhibit uses a table-top screen showing a global map onto which the distributions of planktons in the sea are rendered as colored areas and streams. A visitor can explore the details of a particular area of the sea by moving a physical magnifying glass over the table-top screen. Tracking of the magnifying glass is applied and a schematic representation of different planktons is rendered within the area of the magnifying glass, with the ratio of different plankton types in the rendering retrieved from the original data.

In contrast to the common information visualization technique of "simply" zooming in on the data, using the same representation as in the overview, the plankton example shows the potential benefit of presenting the zoomed data using a representation that differs from the representation of the overview. The

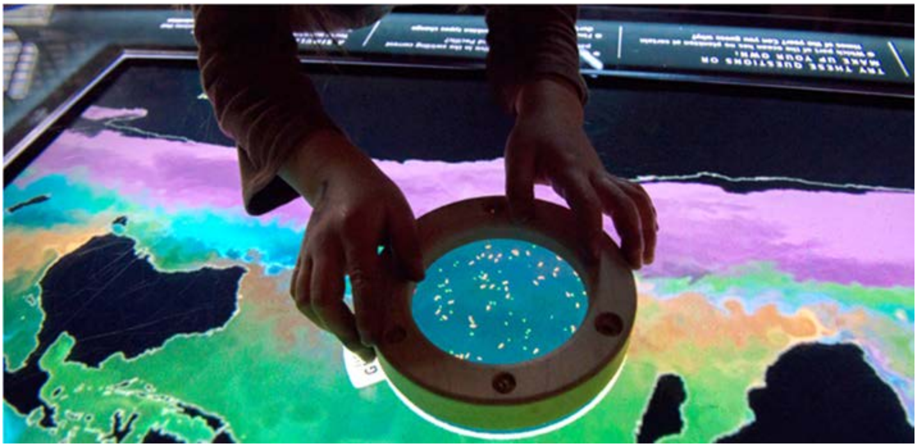


Figure 21. The magnifying glass used as a physical user interface in an exhibit at the Exploratorium in San Francisco. The magnified view shows the amount and type of plankton underneath. Used under permission, photo by Amy Snyder/Exploratorium.

[5] http://www.exploratorium.edu/visit/east-gallery/plankton-population

example provided here is intended for a public exhibition and the perception of the exact amounts of different planktons was not crucial, whereas a more research-focused data representation might have to provide more accurate quantitative measures.

It is not only the user interface that can be made physical, the data can also be presented/rendered in physical form. This is sometimes referred to as *data physicalization*. A definition of data physicalization is suggested by Jansen et al. (2015) as: *A data physicalization (or simply physicalization) is a physical artifact whose geometry or material properties encode data.* This is closely related to the research field of *tangible user interfaces*, although there seems to be a distinction between the data representation (the physicalization) and the user interface. However, sometimes these are hard to separate if the physical data representation is also manipulated.

Many physicalizations seem to be fairly rudimentary ways of representing data, such as strings stretched over nails or pieces of plastic, like the case of the work in progress reported by Stusak and Aslan (2014). At the same time there are examples of technically advanced physicalization research, like development of interactive physical displays. Some such displays, lacking a better word, are constructed using a grid/matrix of dynamically moving bars, where the bars form a surface by being elevated to a height that is dependent on the data being represented. The elevated bar becomes a physical pixel. These physical displays are often combined with an image projected onto the formed surface, thereby enhancing the representation of the data or acting as a support to the interaction interface. These techniques are interesting from a research perspective, both as tangible interaction interfaces and as data representations, although the implications to visual/physical analytics remain to be determined.

In summary, if being inventive the use of simple artifacts, such as a magnifying glass, and well-designed tracking, can lead to very efficient and intuitive ways of physically interacting with a data representation.

3 Interactive Visualization of Multidimensional Data

Visualizing and interacting with large data sets can be challenging from two perspectives, visualizing the high dimensionality of each data element, and presenting and working with the large number of data elements. This chapter will summarize a few strategies and viewpoints on this subject, focusing on the high dimensionality.

In a military context, comparisons of plans and strategies could be examples of multidimensional problems. Complex multidimensional problems also arise in procurement and administration.

There are generally two approaches for dealing with high dimensionality. One approach is to treat all dimensions uniformly. This is useful for exploration tasks, where the analyst has no prior knowledge of the data, such as knowing what is more important and what is less important. The other approach is to reduce the number of dimensions, which is done by projecting the data onto a subspace of lower dimensionality. Methods using this approach can identify important dimensions, but the dimensionality reduction will always lead to a loss of information. In the case of treating all dimensions equally, data is generally not lost, but the vast amount of information can nonetheless hide the information that is sought for.

Making visualization interactive adds an extra constraint to the problem: the computational cost for producing visualization must be affordable within the timeframe that an analyst/observer can wait for an update of the presentation. Dealing with large data sets, puts limitations on what methods are practically useful.

3.1 Treating Dimensions Equally

Aiming at treating all dimensions of a high-dimensional data set uniformly can be seen as one general approach to visualizing the data. Without prior knowledge of what dimensions are more interesting or important for the analysis at hand, a manual selection of dimensions to visualize cannot be done. Many techniques are designed to help finding relevant dimensions, as well as identifying correlations between the dimensions.

This section does not intend to give an exhaustive portrayal of available methods, but rather to describe a few in order to provide an impression of the techniques. This will be exemplified by four methods: (i) scatter plots, (ii) parallel axes, (iii) tabular visualization, and (iv) iconic techniques.

3.1.1 Scatter Plots

One of the most well used ways of presenting numerical data is a scatter plot, where two axes (x and y) are drawn perpendicular to each other, and a data element is depicted as a point with a location corresponding to the value the data takes on each axis. By adding interaction such as rotation of view or holographic techniques, even a third axis/dimension can be added. The application of this technique to higher dimensional data is done by constructing the whole set of scatter plots corresponding to each possible pair of dimensions. The pairwise plots are often collected in a scatter plot matrix, where each dimension is assigned both a row and a column, and each matrix element (except the diagonal) is a 2D scatter plot (Schubert & Hinshaw, 2011).

3.1.2 Tabular Visualization

A straightforward way of gathering data is in the form of a matrix, such as to have each data element presented on a row with the respective data value for each dimension in different columns. As a visualization technique, called tabular visualization, the data value in each matrix cell is displayed in some graphical form, for example a rectangle where the size maps to the numerical value of the cell.

The main limitation to tabular visualization is the ability to simultaneously view a huge matrix. Interaction is a way of approaching this. The overall idea of interacting with such a matrix graphical representation is to filter and rearrange items and part of the matrix to aid visual recognition of patterns in the data. A problem here is that the number of possible rearrangements, such as permutations of dimensions and elements is intractably large (the product of the factorials of number of dimensions and data items), thus requiring computer supported automation in order to be manageable. (Siirtola, 2007), and references therein, discuss this further.

3.1.3 Parallel Coordinates

In parallel coordinates plots (PCP), each dimension is visualized by a vertical axis, see Figure 22 and Figure 23. All axes are of the same size and placed in parallel, and the values of each dimension are mapped onto the size of its axis. Lines drawn in the PCP intersect each axis at a point corresponding to the value, or category, of the particular dimension the axis represents. The resulting plot with all data points will thus consist of a large number of lines. This makes it possible to detect patterns and analyze correlations between variables. Although fairly complex to describe in words, the principle is more clearly outlined in Figure 22. The figure does not include the interactive elements of a well-designed PCP which include, among other things, the possibility to filter the data along each axis, selection of individual lines or sets of lines, and rearrangement of the axis order to investigate correlations.

Dealer Cost	Engine Size (l)	Cyl	HP	City MPG
39014	3.5	6	225	18
17878	2	4	170	21

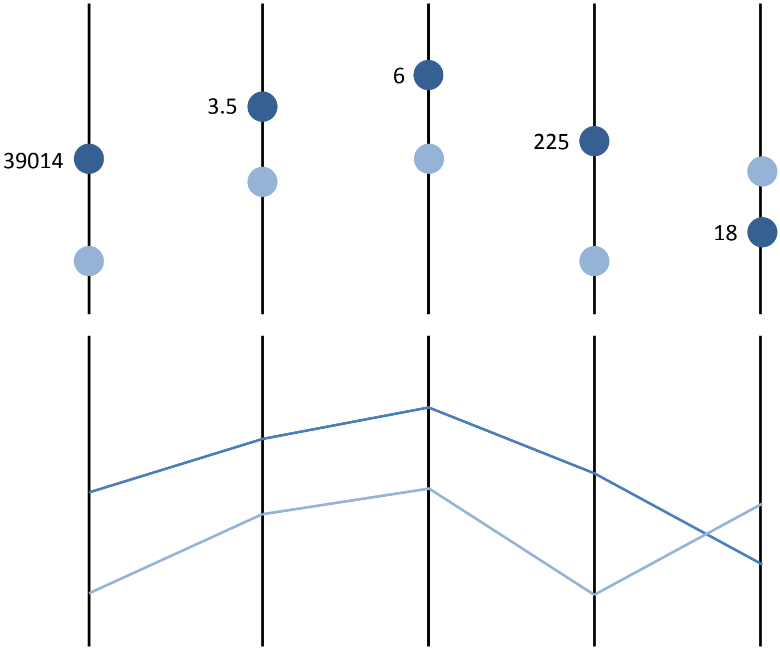


Figure 22. Construction of a Parallel Coordinates Plot from tabular data.

Parallel coordinates suffer from being hard to read if there are many data elements, where the lines from the data elements clutter and also prevent the ability to follow a single line (comprehend a single data element). Siirtola (2007) discusses a few redemptions to this. As such the method is better suited for observing trends rather than seeing details. Analysis is mainly based on identifying clusters of lines that share one or several features; when a large number of data points fall within a short range in one of the axes, this can be spotted as an increased density of lines, or even seen as a bottleneck in the diagram. Correlations between dimensions are identified by observing how such densities vary between the axes. Such direct comparisons between data points or for the whole data set are easy for two adjacent axes, but given a high dimensional data set the ordering can be crucial to get an understanding of the data. Interaction is a remedy for this, by allowing the user to change the ordering of the axes, although the analysis is still limited by the user's ability to assess a potentially huge number of permutations. Another important way of interaction with parallel coordinate plots is accomplished by sorting and filtering techniques such as displaying only data elements within a certain range on one or more axes.

A lot of visualization research has gone into the development of different PCP versions aimed at enabling analysis of large, sometimes time-variate, data sets. In Sweden the work by Jimmy Johansson (2008; 2014) has been aimed at improving usability of Parallel Coordinates and inventing ways to increase clarity of the data represented, including the use of three dimensional PCP.

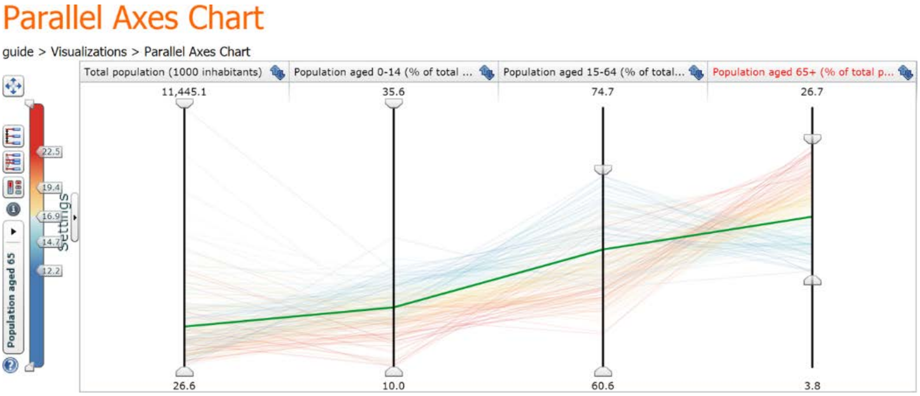

Parallel Axes Chart
guide > Visualizations > Parallel Axes Chart
Total population (1000 inhabitants)
Population aged 0-14 (% of total ...
Population aged 15-64 (% of total...
Population aged 65+ (% of total p...
11,445.1
35.6
74.7
26.7
26.6
10.0
60.6
3.8
22.5
19.4
16.9
14.7
12.2
Settings
Population aged 65

Figure 23. Screen shot showing a population data PCP with one line selected, handles for filtering the data and a color legend.

3.1.4 Iconic Techniques, Glyphs

With icons techniques, small images often also called glyphs, are used for representing data, so that the visual features (shape, size, color, texture, etc.) of the icons reflect features in the data displayed. Each glyph corresponds to a data point, or for example an average of a set of data points. The idea is that differences and similarities in the underlying data should become visible by comparing the glyphs. Chernoff faces are one of the most common examples of glyphs, where data is encoded in size and shape of features of a stylized face, the idea stemming from the observation that humans in general are good at recognizing faces. Another iconic technique is star plots, where axes are laid out with equal angular distance in all directions radiating from an origin, and the each data point is represented by one such star where a line is connecting the values on each axis. Glyphs can be effective for identifying data points that differ from the majority (Borgo, et al., 2013), (Schubert & Hinshaw, 2011) and (Siirtola, 2007).

3.2 Projection Methods

This section presents a few ways of visualizing high-dimensional data by reducing the number of dimensions.

3.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a non-parametric method to visualize high dimensional data sets. The assumption is that the data set is expressed in a number at least partly correlated variables. Variation in data along the different dimensions are assembled and the output is a number of new dimensions that are uncorrelated. The data set is thus linearly transformed into the base that best describes the data set. The goal is to reduce the dimensionality to a number that can be visualized (2 or 3), by displaying only the dimensions that captures the highest variation in data (see Figure 24). Dimensions that have a low variance carry little information as to distinguish the data elements from each other, which motivates not visualizing these. The best description in terms of PCA is one that sums up as much variance as possible in a small number of components. Once the new base is found and the problem visualized, there remains the task of understanding the new dimensions in terms of the old dimensions (Schubert & Hinshaw, 2011).

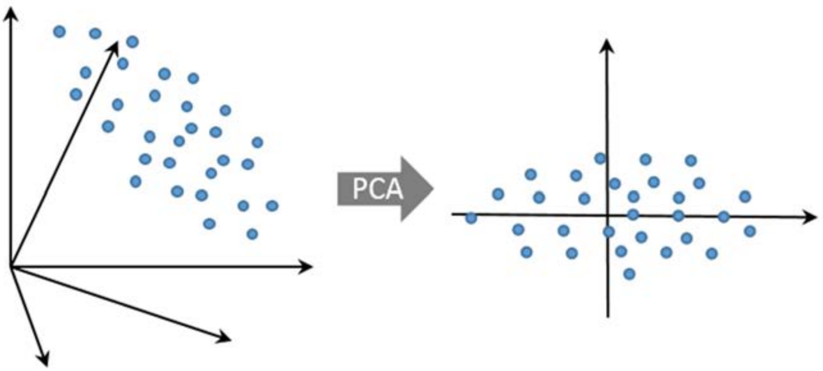


Figure 24. PCA. Left: A high dimensional data set. Right: The dataset visualized by the two first principal components.

3.2.2 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) method uses some a-priori information regarding the data set, such as a division of the data into separate classes. The idea is then to maximize the separability between items belonging to the different classes, and uses similar transformation techniques as PCA. LDA is mostly used for data classification. As such, the starting point is a dataset where each data item is labelled with a class. LDA aims at finding a transformation such that the difference in variance between the classes (or, alternatively the overall variance) and the variance within each class is maximized. Classification of new data elements is done by comparing, for each class, the (Euclidean or other) distance between the transformation of the new data element and the mean of the class (Dzemyda, Kurasova, & Žilinskas, 2013). LDA is illustrated in Figure 25.

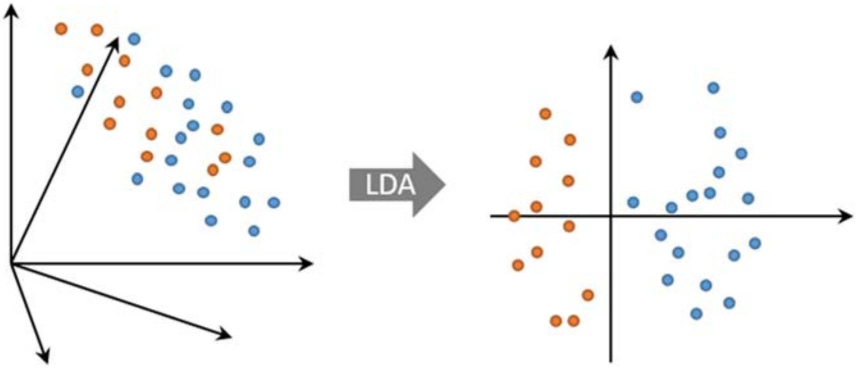


Figure 25. LDA. Left: A high dimensional data set belonging to two beforehand known classes. Right: The dataset visualized by LDA.

3.2.3 Manifold Based Visualization

A lot of data pertaining to the real world tends to lie on a manifold of low dimension as compared to the full dimensionality of the data. In a manifold the local structure is well described by the Euclidean space, although the overall structure is not. For example the planet can locally be mapped on a 2D map with high accuracy although on a global scale that does not hold. Visualization principles based on manifolds aims at keeping neighboring relationships such that elements that are close in the full dimensional space (on the manifold) are also close in the dimensionality reduced space. See (Dzemyda, Kurasova, & Žilinskas, 2013) for references. Isometric Feature Mapping and Locally Linear Embedding are examples of manifold based visualizations. Unfolding of a manifold is illustrated in Figure 26.

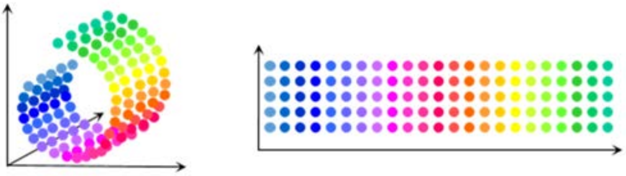
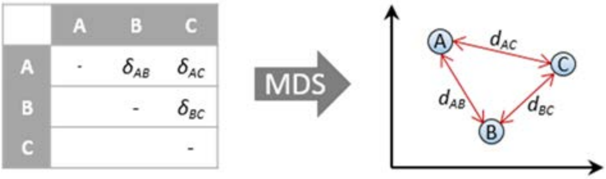


Figure 26. Manifold based visualization. Left: data points on a manifold in 3D. Right: 2D projection of the data.

3.2.4 Multidimensional Scaling

Multidimensional Scaling (MDS) is one of the most employed techniques in multidimensional data visualization. MDS is based on the concept of retaining the pairwise proximities between items, so that the distances between data points in a low dimension space are as close as possible to the distances between the points in the full dimension space (see Figure 27). The data objects to be visualized need not be points in a multidimensional space, what is needed is a measure of pairwise similarity/dissimilarity between the objects. A common case is however data points in a multidimensional space, where the dissimilarity is defined as a distance measure in that space.

The transformation is done by minimizing a stress function, a weighted linear or nonlinear sum of all deviations between dissimilarities or distances (for example Euclidian) in the high-dimensional space. The task is to optimize the positions of the images of the data points on the projection plane such that the stress function is minimized. The dissimilarities of objects in the full dimensional space can be of arbitrary type, the MDS method only requires a scalar dissimilarity measure. Multidimensional scaling is thus cast as an optimization problem. Dzemyda, Kurasova and Žilinskas (2013) present a number of optimization algorithms as well as thorough discussion of other technical issues of the method. France and Carroll (2011) also present a review of MDS techniques.



	A	B	C
A	-	δ_{AB}	δ_{AC}
B		-	δ_{BC}
C			-

Figure 27. MDS. Left: Pairwise distance measures *δ* for objects A, B and C. Right: MDS mapping, differences between (Euclidean) distances *d* in the projection space and distances *δ* in the high-dimensional space.

A general problem is that there are often many local minima, and certain problems gets different interpretations depending on which local minimum is visualized. Finding a global minimum is often very difficult. The high dimensionality of the problem (number of data points and dimensions in the projection space) can make it computationally intractable, and global optimization problems have high algorithmic complexity. The differentiability of the stress function might be a theoretical hindrance.

By using different distance measures, the projected image will differ. The distance measure used in the projection space is in general of greater importance than the distance measure chosen for the full dimensionality space (Dzemyda, Kurasova, & Žilinskas, 2013). Although the images are different, thus opening up for different interpretation of the data visualized, it is difficult to say what image gives the best picture for the analysis task at hand.

The quality of the projection can be measured by the smallest relative error (calculated for example as the square root of a normalized stress function). The dimensionality of the projection space influences the visualization error, and higher dimensionality gives a lower error, with a significant decrease in error going from one dimension to two dimensions, and again from two to three dimensions. Going to higher dimensionality than three further decreases the error, but it is unclear how such a projection space could be visualized. The ability of having stereo screens interactively displaying three-dimensional visualizations is a clear benefit of Visual Analytics as compared to printed images on paper.

3.2.5 Artificial Neural Networks Applied to MDS

This section discusses the application of Artificial Neural Networks (ANN) to the task of visualization of multidimensional data. A problem with Multidimensional Scaling is that the computational complexity scales quadratically with number of data points, thus limiting the usability in interactive visualization since the visualization tends to take longer time than a user is willing to wait. To overcome this problem a lot of research has been devoted to alternative optimization techniques, of which Artificial Neural Networks is one of the most studied. An advantage of the ANN approaches is that these algorithms scale at most linearly, thus enabling interactive analysis.

Artificial Neural Networks are computational constructs inspired by biological neurons and networks of neurons. Outside the scope discussed here, they are employed in machine learning applications such as clustering, classification and function approximation. The function of an artificial neuron is to map a multidimensional input signal to a unidimensional output signal. Typically this is done by forming a linear weighted sum of the input and then apply some nonlinear function (the simplest being a threshold) to the attained sum in order to form an output signal. Letting a number of artificial neurons work in parallel creates a multidimensional output space (of arbitrary dimensionality, independent on the input dimensionality), and such a many-to-many transformation is called a layer of neuron. Several layers can then be stacked on top of each other to form an artificial neural network.

Once the structure of the ANN is chosen (which is not a straight forward task), the task of constructing the mapping from the input space (input signals) to the output space (desired output signals) is a matter of finding the weights in the summations. In the application to visualization of multidimensional data, the input signals is the original data to be visualized, each dimension being one input signal, and the output signals of the network is the two dimensions of the projection. The basic principle of finding the weights is called learning, and is an iterative step where the network is applied to each data point and the weights are incrementally updated according to the error in output as compared to a wanted output. In the case where the wanted output for each input data point is known the learning is labelled supervised. There are also techniques for an ANN to learn a meaningful mapping also in the case when there is no predefined correct output. This is called unsupervised learning. Once the learning is done, the network can be used to map unknown input to the output space.

Since the network is adopted to the data presented in the learning phase, there is an issue of overfitting: if the training (learning) is done too well the network can learn the individual data points and thus lose the ability to generalize to produce meaningful output also for data points not used in the training. If there is labelled data, for example classifications, supervised learning can be applied to a network in order to create a mapping that projects input data into clusters in the output plane corresponding to the classes.

The idea of using neural networks for visualization of multidimensional data is to design a network that takes a data point in the high dimensional space as input and outputs a coordinate in the projection space. A benefit of ANN approaches is that the computational cost is very low once the training of the network is done, as opposed to many other multidimensional scaling methods where the setup has to be regenerated for every new data point. Thus, it is applicable to huge data sets. There is a vast number of different techniques employing ANN to create projections for visualization. Here, we mention just a few to give a taste.

Auto-associative neural networks. In auto-associative neural networks, also called autoencoders, a mapping to the projection space is created without any labels for the data used in the learning. The principle is to place a layer with the same dimensionality as the projection space in the middle of the network, and same dimension for the output as the input, see Figure 28. The network is then trained with supervised learning using the same data point as output target as the input. The projection is found in the layer in the middle. (Dzemyda, Kurasova, & Žilinskas, 2013)

Neuro scale. Neuro scale employs radial basis functions as the nonlinear part of the network, and optimizes the network according to errors between explicitly calculated pairwise distance measures in input and output spaces to achieve a mapping where adjacent points are kept together in the projection. (Dzemyda, Kurasova, & Žilinskas, 2013)

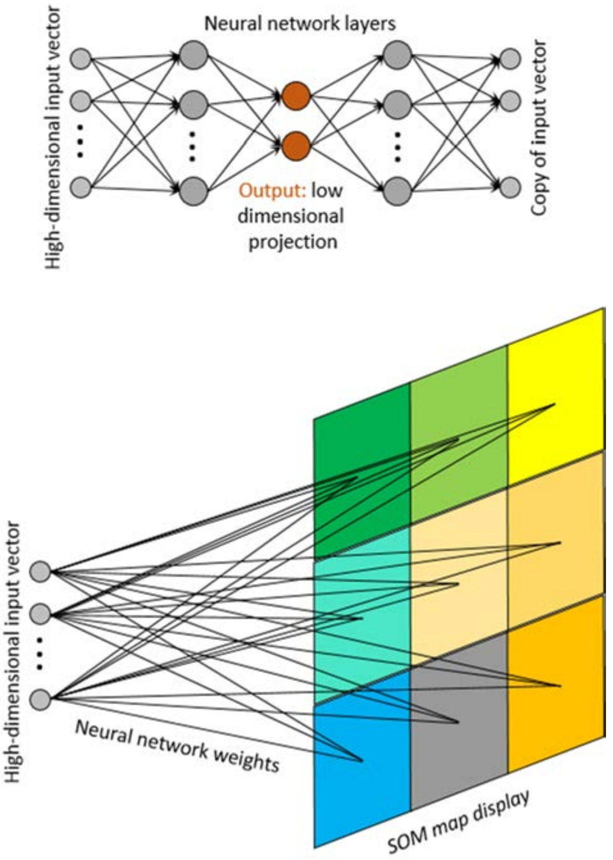

Neural network layers
High-dimensional input vector
Output: low dimensional projection
Copy of input vector
High-dimensional input vector
Neural network weights
SOM map display

Figure 28. Above: Auto-associative neural network. A symmetric mapping is optimized by neural network training such that a low dimensional (2D) image is obtained in the middle of the training network. The training aims at replicating the input vector (to the very right) into the output vector (to the very right), requiring that the low dimension bottleneck captures the relevant information. Below: Self-organizing maps. Weights in a neural network determines the connections between a high-dimensional data input vector and an output positioned on a grid forming a 2D visualization.

Self-Organizing Maps, SOM. Self-organizing maps is an unsupervised learning neural network approach that is used to cluster the data element, and subsequently present them on a 2D-grid. The output from the neural network is laid out in a grid structure (“map”), so that each neuron corresponds to one location on the output grid. It uses a local topological preserving model that places similar data in neighboring clusters, by employing a connectivity between neurons that are nearby on the grid (Flexer, 2001). SOM is illustrated in Figure 28.

As a general remark, it is worth noting that the projection techniques presented here are in most cases not mainly developed and used for data visualization, but rather for other data processing and analysis purposes such as clustering. As such the majority of the research regarding these techniques deals with algorithmic and technical issues rather than exploring the applicability to visualization.

4 Effective Visualization of Multiple Options

In various situations, users are considering a large number of options concerning a system of interest (e.g. a business, production line, battle space, or vehicle). We consider options to be possible actions at hand (e.g., business transactions, or troop movements) or alternative state hypotheses consistent with stored uncertain data.[6] In the former case, possible decisions under consideration may be combinations of actions or system parameter values, and the latter set of hypotheses may result from multiple alternative associations between uncertain observational data and hypotheses. To limit the complexity of this problem in this discussion, we in most cases assume that the underlying data and set of options are static during a presentation session. Hence, for instance, dealing with dynamically changing data and re-evaluation of options is thus beyond the scope of this chapter.

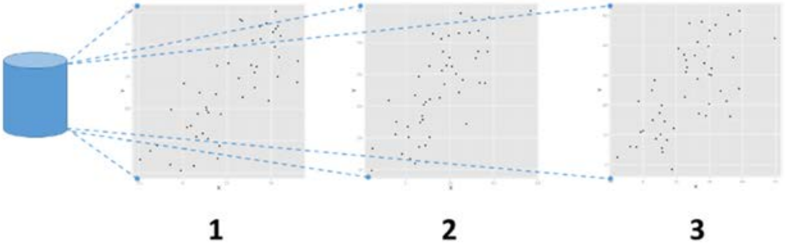


Figure 29. Three alternative certain plots (options) of the same underlying uncertain data

One main feature of this visualization problem is that it does not concern “ordinary” data, such as collected from sensors, or a catalogue of employee data. Instead, if the options represent actions, the visualization is primarily dependent on the available set of actions and their parameters, indirectly affected by underlying system data. Compared to actions, state hypothesis options are similar to “ordinary” data but transformations or expansion thereof. A simple example is that of probabilistic databases (Suciu, Olteanu, Ré, & Koch, 2011), (Johansson, Nilsson, & Pelzer, 2014) where a database of uncertain data is interpreted to be consistent with a number of alternative certain (i.e. standard) databases. Hence, if the represented data in the probabilistic database is a set of uncertain points, the

[6] To the latter case, we can also add candidate machine learning models as options.

multiple options visualization might be a set of alternative point clouds, as illustrated in Figure 29 where three certain (but alternative) plots are shown.[7]

The illustration in Figure 29 serves primarily to describe to the reader of this report how an uncertain dataset may relate to certain ones (not as a visualization for a user). However, the same illustration could possibly be used to visualize the (uncertain) data to a user.

Another example of state hypotheses is state estimation in the battle field (including positions and states of enemy resources). In this case, observed uncertain data can be associated in different ways leading to alternative state estimates.[8]

When the number of action parameters or available data for data association grows, the number of options grows exponentially. Hence, another feature, and challenge, of the visualization problem of multiple options is when the number of options to present is too many. By “too many”, we mean that the full set of options is hard (if not impossible) to visualize in one view for a human user to comprehend effectively.

There are various ways interaction enters into multiple option visualization, most pertinent perhaps being “navigation” by “zooming in” on subsets of the large set of options, or selecting evaluation metrics. Additionally, related interaction involves creating an overview over the set of options by aggregation or filtering, and, in some cases, changing the type of admissible options (e.g., pruning the set of hypotheses, or increasing or decreasing the set of action parameters).

In the remaining sections of this chapter, we further discuss this Visual Analytics problem and variations thereof.

4.1 Work Process

As far as we know, there is no well-established multiple options subfield (as coined and outlined in this chapter) of Visual Analytics, and hence no dedicated literature. A sub problem focusing on options as machine learning models is identified and briefly addressed in (Johnston, 2002). There, evaluating multiple different models (i.e. different model parameters) with respect to training data is discussed.

[7] Probabilistic databases contain uncertain data, but can instead be interpreted as a probabilistic uncertainty over certain (standard) databases. One advantage with “expanding” uncertain data to alternative certain datasets is that standard analysis methods can typically be applied to each of the latter.

[8] These two views on options, i.e. as actions and hypotheses, may be integrated in a common system by using the set of alternative state estimates to evaluate or rank alternative actions.

Since we lack previous work to relate to, we first attempt to characterize the problem by its pertinent features and then address the different parts by providing own ideas and references to previously developed visualization techniques.

4.2 Characterization

Already in the introduction, we described two features of visualizing multiple options: i) options is a special kind of data, and ii) “too many” options to visualize. We here dive further into the details of these two features.

Options as a special kind of data means that we need to consider how to represent and visualize actions as that, rather than underlying raw data. This is further discussed in section 4.3 below.

Managing too much information to visualize is not an unknown problem to the visualization community. In section 4.4, we discuss ideas about how to address that problem for multiple options.

Naturally, following from the “too many” feature, user interaction is useful to facilitate navigation of the set of options, including selection of option evaluation metrics. There is also an opportunity for the user to contribute with its expertise to the processing of options. In the case of a growing set of system state hypotheses, the user may be allowed to “prune” unreasonable hypotheses to simplify the further processing of state hypotheses. How the computer and human can collaborate concerning multiple options is discussed in section 4.5.

4.3 Option Representation

One important issue to solve is how to represent options appropriately. By representation, we mean the type of *data* and *meta-data* of individual options and its type of *manifestation* in the presentation interface (the presentation modality we focus on in this chapter is only visual). For instance, if the options are actions, their representation should, as a minimum, convey an accurate understanding of the range of actions available to the user, preferably combined with expected impact of actions. As an example, in Figure 30, we imagine a single action parameter controls the speed of a production line. The visual manifestation of the speed parameter control is a *turn dial*, which in the current discussion is a depiction on a screen, but could as well be a physical dial instead. The parameter has a few possible values (i.e., data) indicated by the small lines that extend from the black circle (representing increasing speed in a clock-wise order). A part of the manifestation is also the meta-data including the current state of the production line (i.e. the estimated risk of failure, shown on the left), and performance values associated with each action value (i.e., updated estimated risk, and estimated productivity increase, shown on the right).

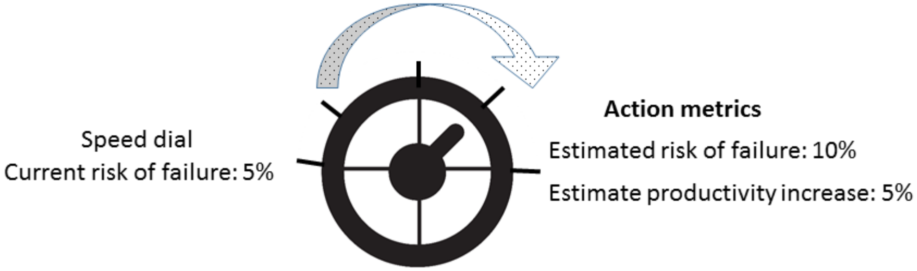

Speed dial
Current risk of failure: 5%
Action metrics
Estimated risk of failure: 10%
Estimate productivity increase: 5%

Figure 30. A single action parameter (speed of a hypothetical production line) with a discrete number of possible values. The illustration is based on a figure from `http://openclipart.org`.

The actual representation to select, is of course heavily dependent on the application in question, but we can discern a few common issues:

1. If options are actions, the degrees of available freedom should be clearly conveyed to the user. One thing to consider is that the set of feasible actions may change dynamically depending on context with each new presentation session. The action parameters that correspond to physical actuators may be truthfully visualized. In other cases, appropriate metaphors capture the set of options efficiently, e.g. the turn dial in Figure 30 which here is discrete but could in principle capture an infinite number of action values.

2. Meta-data of options could opportunistically be taken into consideration to guide the visualization (e.g., by ordering options based on evaluation metrics). An example is provided in the text following this list.

3. Options are typically similar (even overlapping in structure sometimes). For instance, complex actions involving multiple parameters, may have the same values for some parameters, differing only in a few parameters. In that case, some kind of simplified succinct representation of the action options might be possible. For instance, consider a gardening system where one parameter concerns regulating a water sprinkler. In a decision situation where there is rain, the water sprinkler parameter can be excluded from the visualization. A further example is provided in the text following this list.

Concerning issue 2 above, as an example, if meta-data such as the expected performance of candidate actions is measured, or the likelihood of state hypotheses, the options could be ordered to promote the most valuable ones and suppress the others.

In cases where options are evaluated with respect to multiple metrics (e.g. low cost and low risk of failure, similar to the example in Figure 30), visualizing multiple options is largely equivalent to visualizing multidimensional data as discussed in chapter 3.

Figure 31 provides an example with a set of options (each denoted by a ‘*’) plotted with respect to the two aforementioned metrics. For each metric, an action which leads to a low value is preferred. If these two metrics are the only basis for decision, the visualization should primarily concern the options belonging to the so called Pareto front, as for each of the remaining options, they should be suppressed as there is at least one option belonging to the front which is better in all metrics.

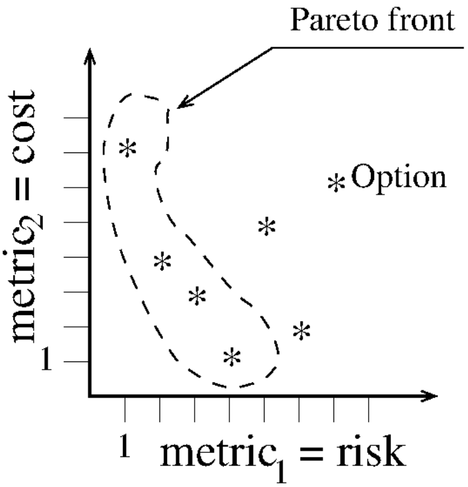


Figure 31. The values of a number of options (‘*’) are evaluated and plotted with respect to two metrics. The visualization can focus solely on visualizing the options in the Pareto-optimal set.

Concerning issue 3 above, occasionally, strong structural similarity between options correlates with their value and options can be grouped based on similar value to facilitate aggregated views of options. In Figure 32, a set of alternative sequences of associations are considered. As many of the sequences are overlapping, the sequences (i.e. options) can be collapsed into a compact tree-representation of options, where each path (from root node to leaf) is an option.

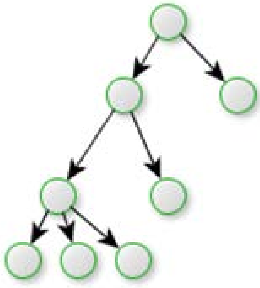


Figure 32. A tree representation of alternative options. Each path in the tree (from root node to leaf node) represents an option. (Published with courtesy of Andreas Horndahl.)

4.4 Dealing with Multiple Options

In section 4.3, we discussed the visualization of a single option. Here, we further discuss the visualization of the whole set of options at hand.

First consider the special case where there is no option to select, instead the focus is to describe the whole set of options. As a concrete example, imagine a set of likelihood-weighted state hypotheses concerning a target state. The set can be integrated to create a single estimate of the target state, if that is what the application requires. One could also extract extreme values (such as upper and lower bounds) on the target state. Here, various types of descriptive data mining techniques such as clustering, and rule extraction can be applied (see for instance Hand et al. (2001)).

In many applications, however, a summary or aggregate of the set of options is not requested, but the user is required to select one or a subset of options from the set. Then the challenge appears to visualize the individual options, and sometimes they are too many to fit the presentation interface.

A few obvious means to reduce the set of visualized options are to exploit

1. meta-data to rank the options and only display the highest ranked ones;
2. structural similarities between options.

With respect to 1, evaluation of options can be considered to be meta-data; hypotheses can be ordered (prioritized) according to likelihood, machine learning models according to how well they fit data, and actions according to expected utility. In the previous section, we gave an example of how multiple metrics can be used to select which options to show (i.e., the Pareto set in Figure 31).

With respect to 2, we also gave an example in the previous section of how structural similarity between sequences of associations can be presented succinctly by a tree (Figure 32).

What to do then if, after reduction of the set of options, the set is still too large to visualize immediately? We have not found any scientific publications on the subject, but practitioners share their personal experience. A compilation of two of those sources (Bertini, 2011; Scheidegger, 2015) are in the following list:

- *Sampling* – selecting random subset of the full set, whose size is suitable for the presentation interface

- *Filtering/segmentation* – similar to sampling, but the selection is not random but based on a set of rules (for instance only selecting options with high likelihood or high expected utility, or only studying the subset where one parameter is fixed such as only data on males)

- *Aggregation* – summarize subsets of the options, for instance finding cluster centers. For instance, one might cluster actions with similar performance, let the user choose a cluster, and then select an option randomly from the cluster to implement.

- *Interaction* – unlike the methods above, the interaction approach does not lose data and instead lets the user navigate the full set of options by, for instance, restricting allowable values, parameter ranges (“zooming in”), metrics to consider, alternative visualization (e.g., 2-D or 3-D), etc. Make sure that the user knows the context of the current view (e.g. the current range selection).

It might seem disappointing to have to resort to the approximation methods of sampling, filtering, and aggregation mentioned above. However, if there is a large number of options, a large subset of options may, in practice, be roughly equally good. So a lengthy manual search for a best option can sometimes be replaced by a (hopefully negligibly) suboptimal but fast one.

4.5 Interaction and Collaboration

Interaction is an important part of visualization of multiple options. Already in the previous section, we discussed interaction as a means to manage the option selection from a huge set of options. Furthermore, in some cases, the effectiveness of the visualization can be improved by prompting the user to apply its expert knowledge to suppress unlikely or otherwise unwanted options.

A typical example is where options are state hypotheses (based on sensor to target associations) and the number grows over time. In this case, for computational efficiency, it is important to prune the growing tree of possible hypotheses. Hence, if the presentation interface provides user input on more or less likely hypotheses, this can be important feedback to the algorithm that manages the set of options.

An additional example of computer-human collaboration is described in (Karami & Johansson, 2014). Here, a user is considering different sensor control options to acquire new relevant information to support an intelligence analysis problem at hand. At the bottom of the image is a computer which ranks configuration options based on the evidence collected by sensors. However, the evidence does not cover all aspects that could be included when taking a decision about configuration option. The user has important experience with, for instance, cost and risks with different configuration options. The computer's ranking acts as input to a multi-attribute decision making (MADM) module which also integrates the human's inputs. The result is a ranking of the options which considers both the human and computer aspects.

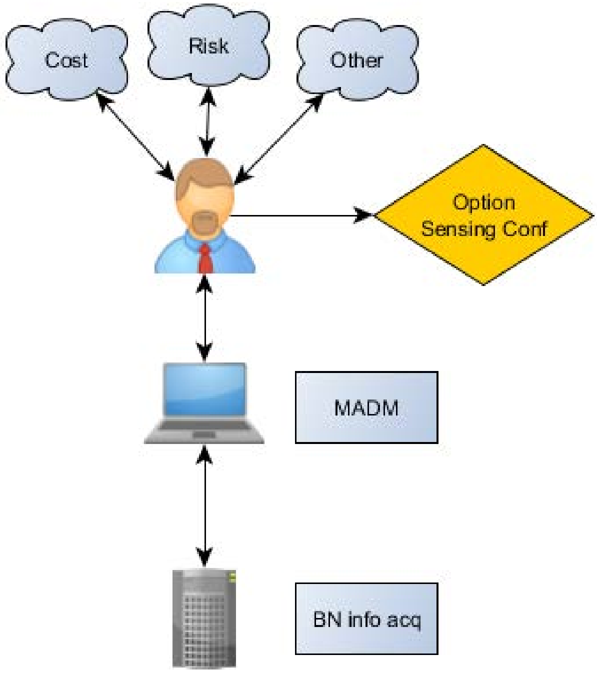


Figure 33. A system which supports a user to make decision about sensor configuration options.

4.6 Summary and Discussion

In this chapter, we outlined a Visual Analytics problem which involves presenting a large number of options to a user (for instance system action configurations, state hypotheses, or machine learning models). A concrete example is provided in section 7.3.2.

Our study did not discover a research subfield dedicated to the visualization of multiple options. Instead we approach the issue from different perspectives: representation (section 4.3), multiplicity (section 4.4), and human-computer collaboration (section 4.5), and describe how current result can contribute to deal with these perspectives.

The most important findings of this study are:

1. Options to visualize are not the same as the underlying data (which is what is visualized in most applications), instead options merely reflect the underlying data and meta-data. This fact can either simplify visualization or make it more complex. For instance, if the number of actions is fixed, the number of data points in the underlying dataset will have little impact on the visualization of the fixed set of options.
2. There are basically two different ways to deal with the visualization of too many options: i) “destructive” simplification (sample, filter, or aggregate) or ii) let the user iteratively zoom in and navigate through the set of options.

5 Uncertainty in Visual Analytics

This chapter is mainly influenced by the reviews performed by Brodlie, Allendes Osorio and Lopes (2012) and Bonneau, et al. (2014).

All data are intrinsically associated with uncertainty, ambiguity and imprecision. Yet, most visualization techniques do not reflect this fact and assume that the displayed data are exact. Too often, the uncertainty remain overlooked in visualization, mostly due to difficulties in applying existing visualization approaches, increasing visual complexity of addition of uncertainty, and the lack of obvious visualization techniques (Bonneau, et al., 2014). Even though, *error bars* and *boxplots* are frequently used in representation of uncertainty in scientific publications, in other contexts, usually the concept of uncertainty in representation of data is omitted, especially when visualizations are used in decision making. For instance, contour maps rarely incorporate any notion of uncertainty, furthermore, the very crispness of a contour line conveys the impression of confidence that is frankly an illusion (Brodlie, Allendes Osorio, & Lopes, 2012). However, in recent years, the awareness of the uncertainty problem within the visualization community has grown (e.g. see recommendations by Thomas and Cook (2005)), and many traditional methods have been extended to embrace the concept of uncertainty.

5.1 Sources of Uncertainty

Visual analytics is a means through which scientists and decision makers, investigate, evaluate and explore available or simulated data in order to identify patterns and generate hypotheses. Uncertainty in Visual Analytics may refer to the lack of certainty in all different stages of this process and be originated from different sources, ranged from uncertainty observed in sampled data, uncertainty measures generated by models or simulations, and uncertainty introduced by the data processing or visualization process (Bonneau, et al., 2014). We follow the taxonomy presented by Brodlie, Allendes Osorio and Lopes (2012) and distinguish between the two broad classes of uncertainty: *visualization of uncertainty* and *uncertainty of visualization*.

While visualization of uncertainty deals with the problem of depicting the uncertainty of data, the uncertainty of visualization considers how much uncertainty is added to the data as they are processed through the visualization pipeline. To understand types and sources of uncertainty consider the visualization reference model presented by Haber and McNabb (1990) in Figure 34, adapted by Brodlie, Allendes Osorio and Lopes (2012). The data from measurement or simulation are pre-processed and filtered. This step often includes approximation and interpolation. The filtered data is then passed through mapping stage, that is, some visualization algorithm that produces geometrical objects. In the final stage, the geometry is rendered to an image, which is presented to the user. All these stages encompass uncertainty. While visualization of uncertainty concerns with uncertainty in data itself, uncertainty of visualization refers to the uncertainty that is added to the data through the visualization pipeline.

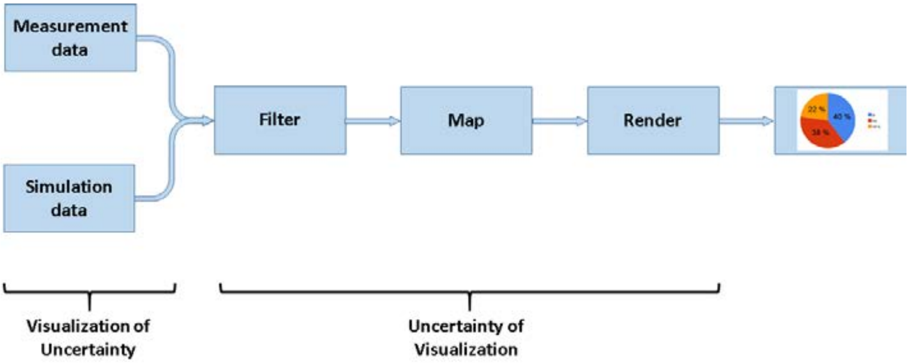


Figure 34. Visualization of uncertainty and uncertainty of visualization (Brodlie, Allendes Osorio, & Lopes, 2012).

5.1.1 Visualization of Uncertainty

The data, either measured in experiments or being result of the simulation, are seldom exact and are associated with some uncertainty. The visualization of uncertainty focuses on how to visualize the uncertainty, which is associated with the data.

5.1.2 Uncertainty of Visualization

Several operation during the filtering stage may increase the uncertainty of the data. For instance, approximation and rounding of the data, or interpolation to infer incomplete data. The mapping process involves some visualization algorithm that produces a geometrical object. These algorithms may introduce or increase the uncertainty of the data, for example, numerical solutions of equations, or approximation of curved surfaces by polygons. Rendering stage involves discretization, which may obscure information whenever the resolution of the output image is lower than the resolution of the data. As the data are processed through the visualization pipeline, these uncertainties are accumulated and may be amplified. Thus, it is of utmost importance that visualization methods have proper facilities to manage and represent uncertainty.

5.2 Approaches to Visualization of Uncertainty

One of the main reasons that makes visualization of the uncertainty a difficult topic, is that one needs at least an extra dimension to visualize it. For instance, consider the simple example of plotting a zero-dimensional point in Figure 35. If the value of a data point in the 2-dimeonsional plan is exact for example $(x, y) = (3{,}5)$, it can be illustrated using a point marker having no dimensions. However, if there is uncertainty in the y-value (e.g.$y = 5 \pm 0.5$), we need an extra dimension to visualize the uncertainty as an error bar (the marker becomes a line). The dimension of the data is unchanged and still zero, but the extra dimension is required in the visualization to display uncertainty. In the same manner, isolines and isosurfaces become areas and volumes in the presence of uncertainty. Adding a new dimension will especially be difficult for $3D$ and higher dimensions, where we already have trouble with visualization for the exact values.

Incorporating uncertainty, implies increasing the complexity of the visualization. One way to avoid cluttered visualization, is to let the user interactively choose between adding and removing an overlay indicating the uncertainty (e.g. standard deviation) to the crisp representation of the data.

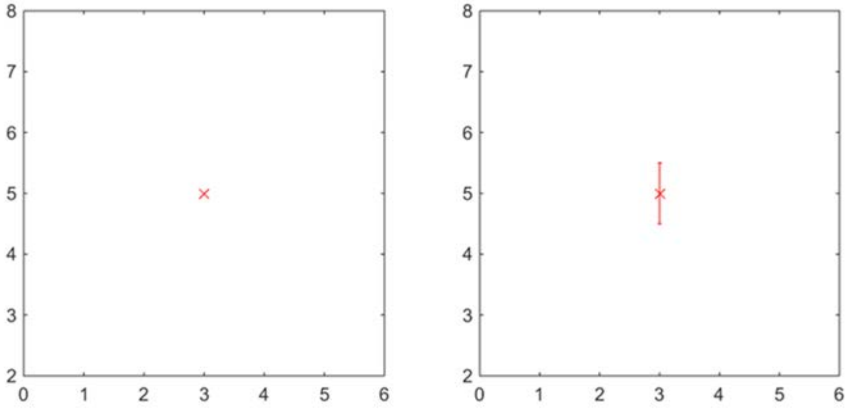


Figure 35. Uncertainty increases the dimension of a point from 0 to 1: point becomes a line.

When it is not possible to add more spatial dimensions, we have to create a visual ‘dimension’ using different means. Examples of such solutions are the following:

- **Juxtaposition:** providing the visualization of uncertainty in a separate pictures, for example plotting standard deviation alongside a mean value plot.
- **Animation:** using time as an extra dimension, for example by displaying a sequence of possible instances of a model.
- **Overly:** Superimposing the visualization of uncertainty over the normal visualization, for example overlaying a counter map of standard deviation on top of a heat map of the mean values.
- **Color:** using the hue, saturation and the value of colors as an extra dimension to encode uncertainty.

5.3 Methods for Visualization of Uncertainty

In the following some of the efforts used in visualization of uncertainty are presented. The sample is far from exhaustive and serves only to demonstrate the variety of the used methods. Interested readers are referred to more comprehensive reviews of this topic, e.g. (Pang, Wittenbrink, & Lodh, 1996) and (Brodlie, Allendes Osorio, & Lopes, 2012).

5.3.1 Points in 2D or 3D

In visualization of a collection of points in a 3D space, for example positions of astronomical objects in the universe, which are described by their distance (from earth) and equatorial coordinates, right ascension (RA) and declination (Dec), the positional uncertainty of the objects can be decomposed into two components: a radial component along the sight (from earth to the object) and a spherical coordinate (RA/Dec). This positional uncertainty can be visualized by plotting a line segment (error bars) along the line of sight from earth, ignoring the uncertainty in RA/Dec, since the uncertainty in distance is usually of a larger orders of magnitude than the uncertainty in RA/Dec. The uncertainty can instead be visualized using ellipsoids centered on the objects, to visualize all uncertainty terms simultaneously, if the RA/Dec component is not negligible (Li, Fu, Li, & Hanson, 2007).

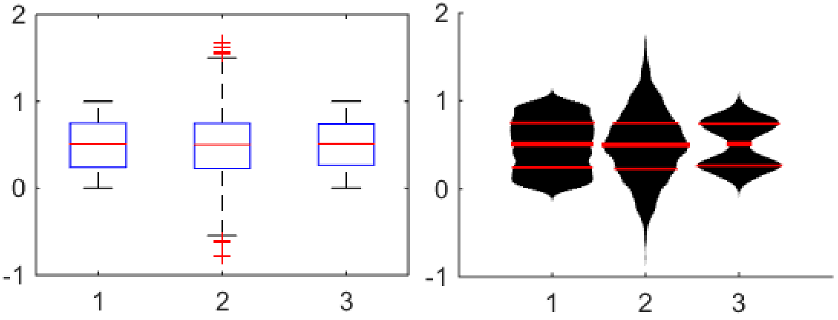


Figure 36. At the left, boxplots visualizing uncertainty of 3 datasets randomly generated, at the right violin plots for the same data. While the interquartile range is shown by a blue rectangle in boxplots, in the violin plots the median and the interquartile range are marked by red lines.

5.3.2 Boxplots

For visualizing uncertainty of the value of a data point (or several data points) that is presented by different observations of a scalar variable, the standard method is the boxplot (Tukey, 1977), which depicts upper and lower bounds, upper and lower quartiles, median and possible outliers. The upper and lower bounds may represent different values, for example, minimum and maximum of the data, or one standard deviation above and below the mean of the data. Different extensions have been suggested to include higher-order statistics such as skewness, kurtosis and tailing into the boxplot (Potter, Kniss, Riesenfeld, & Johnson, 2010). Violin plots (Hintze & Nelson, 1998) add the information available from probability density of the data at different values (density shape) to the boxplot. Figure 36 shows boxplots and violin plots for 3 randomly generated uncertain datasets. While the boxplots for the first and the third datasets seem almost identical, the violin plots for the same datasets reveal that they are distributed quite differently.

5.3.3 Two-dimensional Graphs

One of the most common visualization objects is two-dimensional graph. The uncertainty of the underlying points that constitute the graph can be added by different means. For instance, one can choose to add error bars to the data point markers or use size or color of the point markers to encode the uncertainty. For a continuous graph, the graph itself can be color coded using an uncertainty color map. Figure 37 illustrates two-dimensional graphs visualizing uncertainty with size and colors.

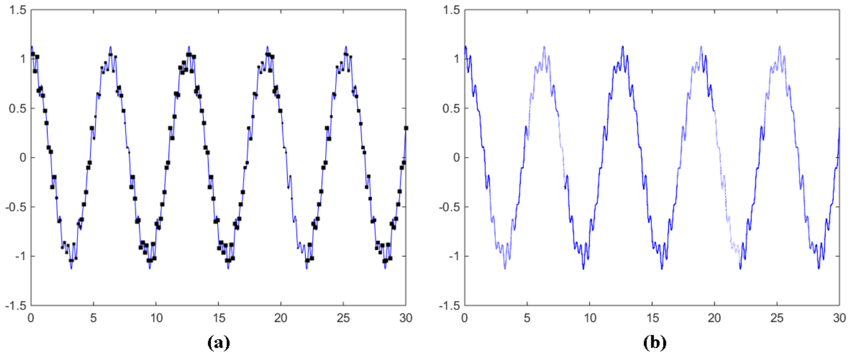


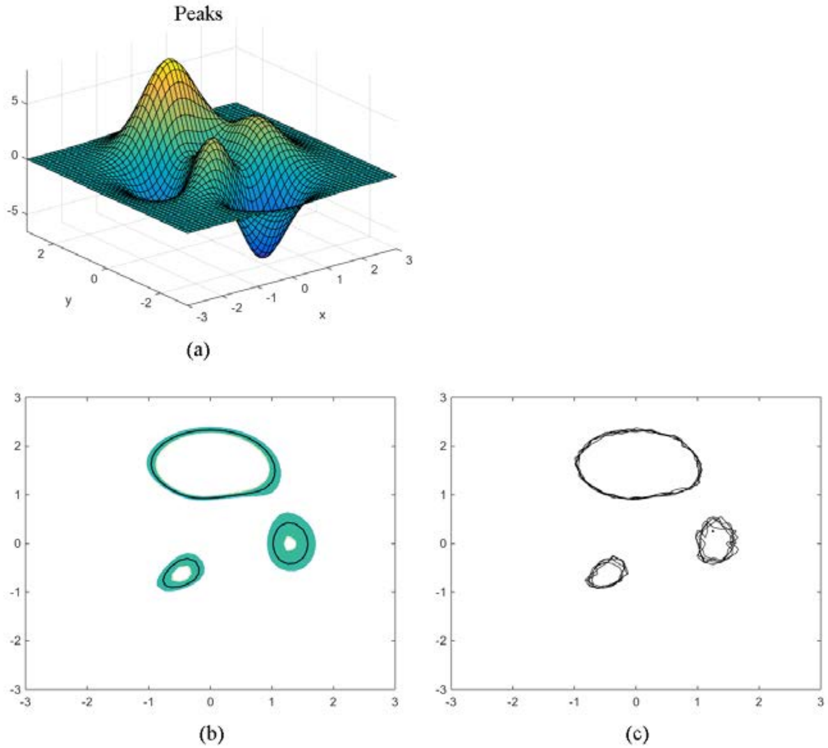
Figure 37. Two-dimensional graphs visualizing uncertainty. In (a), the uncertainty is added by the point marker size (the larger, the more certain), and in (b) the uncertainty of the continuous model is visualized by different saturation of the graph color (the sharper, the more certain).

5.3.4 Three-dimensional Surface Plots

In three-dimensional surface plots, akin to the two-dimensional graphs, adding glyphs at data points on the surface with varying size and color, error bars, and color mapping of the surface are frequently used. Another common method is adding time dimension and animation effects (Ehlschlaeger, Shortridge, & Goodchild, 1997). Uncertainty in the data is transmitted from the spatial to the temporal domain, and visual vibrations are used to indicate the level of imprecision at visualized data points (Brown, 2004).

5.3.5 Contour Lines

Visualization of uncertainty in contour lines can be distinguished into two categories. In the first category, the uncertainty in isolines is produced as a result of the uncertainty in the constant value (evaluation). In the second category, the uncertainty is in the space of the independent variable. The first is called *value uncertainty*, and the second *positional uncertainty* (Brodlie, Allendes Osorio, & Lopes, 2012). Value uncertainty is visualized by a crisp isoline depicting the mean value and an overlay that indicates the uncertainty of the value, for example by a standard deviation. Positional uncertainty is usually visualized by using a spaghetti plot, in which an isoline is drawn for each model in an ensemble (see Figure 38). There are other methods that do not fall into either category; for a thorough discussion of value uncertainty, positional uncertainty and other approaches see (Brodlie, Allendes Osorio, & Lopes, 2012) and references therein, especially (Sanyal, et al., 2010; Potter, et al., 2009; Juang, Chen, & Lee, 2004; Allendes Osorio & Brodlie, 2008; Pöthkow, Weber, & Hege, 2011).



(a)

(b)

(c)

Figure 38. Uncertainty in contour lines. (a) The Matlab's "peaks" function. (b) Visualization of value of uncertainty; points on the peaks function having a value 3 ± 05. (c) Positional uncertainty; spaghetti plot for isolines for 5 instances of the peaks function with added noise.

5.3.6 Heatmaps

A heatmap is a color mapping used for visualizing a scalar field, which associates a scalar value (e.g. temperature) to each point of a two-dimensional space. Different methods have been suggested to embrace uncertainty in heatmaps, among others, providing heatmaps of mean and standard deviation and adding whiteness to uncertain areas. For more examples and discussion over these methods see (Love, Pang, & Kao, 2005; Hengl, 2003; Cedilnik & Rheingans, 2000; Coninx, Bonneau, Droulez, & Thibault, 2011).

5.3.7 Isosurface

An isosurface is a three-dimensional equivalent of the two-dimensional isoline (contour line) that is a surface representing points in a three-dimensional space having a constant value (e.g. temperature, pressure). In the presence of the uncertainty the situation becomes difficult, since the three space dimensions are used for visualizing the surface. Similar to an isoline, two types of uncertainties may exist: visualization of value uncertainty and positional uncertainty. For value uncertainty, the isosurface of the mean value alongside with an indication of the uncertainty in data (using either color or glyphs) can be used, for example see (Johnson & Sanderson, 2003; Rhodes, Laramee, & Bergeron, 2003; Newman & Lee, 2004; Grigoryan & Rheingans, 2004). An illustration of an isosurface with value uncertainty is found in Figure 39. For positional uncertainty of an isosurface see (Pöthkow & Hege, 2010; Pöthkow, Weber, & Hege, 2011). Both value and positional uncertainties for isosurfaces are studied by Zehnera, Watanabea & Kolditz (2010) and Love, Pang & Kao (2005).

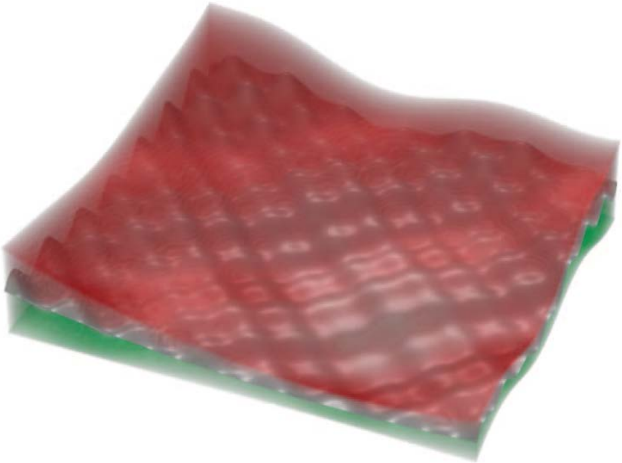


Figure 39. Isosurface of a synthetic data set indicating value uncertainty around the isovalue (Johnson & Sanderson, 2003). © 2003 IEEE. Reprinted with permission.

5.3.8 Scatter Plot and Parallel Coordinates

Multivariate data without specific dependency on space, time or other physical values (which is common in scientific visualization), are often visualized using *scatter plot (matrices)* or *parallel coordinates*. Xie, Huang, Ward & Rundensteiner (2006) and Feng, Kwock, Lee & Taylor (2010) augment scatter plots and parallel coordinates plots to incorporate uncertainty and integrate them with existing multivariate analysis techniques. However, generalizing these methods to cases where the multivariate data depend on some variables (e.g. space and time) is challenging. Example of methods that address challenges in unsteady multi-field visualization can be found in (Jänicke, Wiebel, Scheuermann, & Kollmann, 2007).

5.3.9 Arrow Plots

In visualizing an overall picture of a flow vector field, a common approach is to position glyphs at grid points (or superimposed on contour lines) to convey properties such as direction or velocity. An example is the *arrow plot* where the direction in which the arrow points is the direction of the vector and the length of the arrow is its magnitude. To incorporate uncertainty Wittenbrink, Pang & Lodha (2006) and Zuk, Downton, Gray, Carpendale & Liang (2008) use varying arrow shapes to indicate uncertainty in bearing and magnitude of the vectors in steady flows.

5.4 Uncertainty of Visualization

Even if the data is precise, the visualization process itself may generate uncertainty. Two main sources for introducing errors are: (i) filtering stage, while we create a model of the data by interpolating the available data, and (ii) mapping and rendering stage, when we represent the model by a graphical object.

The uncertainty introduced to the representation can be of such magnitude that scientists' efforts to generate data using higher order approximation appear meaningless. Methods such as ray tracing, which are able to pass higher order data through the visualization pipeline usually suffer from poor performance and depend heavily on high computational power. However, Nelson & Kirby (2006) and Nelson, Haimes & Kirby (2011) show that for high accuracy, direct ray tracing of high order finite element is superior to marching cubes (one of the most used algorithms in computer graphics) for drawing isosurfaces.

5.5 Conclusion

As challenging it is to incorporate the uncertainty in visualization, as easy it is to ignore it. However, the uncertainty in the ground truth does not disappear just because we overlook it.

Fred Brook in his keynote speech at the IEEE Visualization '93 conference reminded the audience the obligation of truthfulness in scientific presentations and stated "Scientific visualization surpasses all other computer graphics in the pre-eminent obligation for truthfulness in what it conveys" (Brooks, 1993). Increasing the expressive power of visualization techniques has not rendered these recommendations obsolete, and we still rely on the integrity of the visualization scientist. It is the responsibility of the scientist to create an honest visual representation of the data and provide the user with an indication of how reliable the representation is (Brodlie, Allendes Osorio, & Lopes, 2012).

6 Visual Analytics in Computational Fluid Dynamics

Visual Analytics plays an important role when managing search and exploration of data during analysis work. Due to ever increasing capabilities in data harvesting for analysis purpose and decision making, the methodology becomes interesting for more and more applications. The ability to interact with the data sets through data extraction/reduction and visualization is often vital for the understanding of complex problems. For the intelligence gathering done by the security agencies, VA is imperative since the security agencies around the world nowadays work more and more with analysis of huge disparate data streams (Thomas & Cook, 2005). Other fields such as general law enforcement, infrastructure protection and financial fraud analytics are today using tailored methods to deal with large quantities of data, see e.g. (Kielman, Thomas, & May, 2009).

The scope of VA is, as argued in (Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008), to employ intelligent algorithms and user interfaces in order for the human to visualize and directly interact with the information; to take the analysis process to the next level. “A science of analytical reasoning facilitated by interactive visual interfaces” as proposed in (Thomas & Cook, 2005). In (Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008), a formal model of this process is presented with the key concepts explained. As mentioned in e.g. (Simoff, Böhlen, & Mazeika, 2008) visual data mining by use of more sophisticated software and procedures that moves the analyst in close contact with the data sets and provides the ability to interactively explore the data is not, however, being widely enough used. Several reasons for this exist. In (Simoff, Böhlen, & Mazeika, 2008) it is argued that, amongst other things, the field is new and that, in order to deploy the methods, the analyst needs to be proficient in both data mining and visualization. Furthermore, the focus is very often on the results of the analysis rather than the justification of the analysis procedure as such.

Within the field of Computational Fluid Dynamics (CFD), Visual Analytics plays a highly important role and has done so for many years. In fact, the ability of being able to, sometimes on the fly, visualize and interact with the data is of fundamental importance for the understanding of fluid flow. This is due to the complex nature of flow physics that makes it extremely hard to predict intuitively. In Figure 40, examples of visualization of different types of fluid flow are shown.

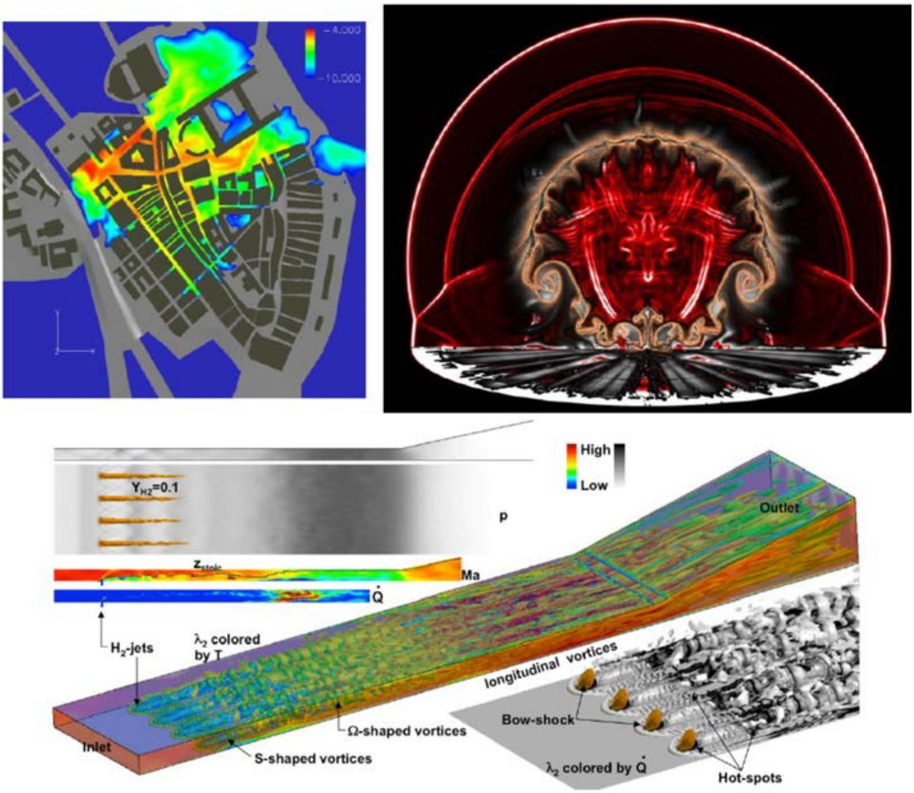


Figure 40. Examples of Computational Fluid Dynamics visualization. Upper left: Contaminant transport in Old Town, Stockholm (Parmhed, Svennberg, Burman, & Thaning, 2004). Iso-surface colored with contaminant mass fraction. Upper right: Explosives blast close to ground (Fedina, 2014). Post explosion after burning with burning coal particles. Shock waves reflecting on ground. Lower: Vorticity visualization from fuel injection nozzle in HYSHOT Scram-jet engine (Chapuis, et al., 2013). Iso-surface of vorticity colored by temperature.

The engineering terms used within the CFD community to describe these activities are *pre*- and *post processing* rather than Visual Analytics. When working with CFD, many of the operations conducted in order to setup the problem and to get a better understanding of the governing flow physics are based on VA. Even the analysis step including the pure number crunching of the numerical approximations of the governing physics equations, the *solve* or *solution* phase, is of interest both to monitor and analyze visually. Since the simulation is often performed on large High Performance Computing (HPC) systems, both flow physics variables, numerical convergence related variables as well as more computer related variables such as storage capacity/usage and compute cluster performance are targeted. In Figure 41, a general schematic of the typical analysis workflow is shown (Fureby, 2015).

The procedures developed to analyze different parts of the flow are highly tailored by the analyst. There exist numerous, readily available tools, for example statistics toolboxes, graph plotters, visualization environments, movie making software etc., but, in general, a significant part of the overview of the analysis toolbox is created, set up and maintained by the analyst. Thus, there is always a risk of inventing the wheel over and over again. In Figure 42, typical analysis output are shown: the numerical solutions are compared with experimental results.

The CFD software vendor list ranges from large software companies listed on the stock market, to small companies and groups usually distributing their free software through the GNU licensing format. Common for many of these programs is that they are targeted specifically for CFD application. Most of these are sold or distributed as pre-processing or post-processing tools rather than VA tools. Today, many of the larger companies provide software that are, in essence, VA tools tailored for the CFD application, with all the functionality within the native, fully-integrated product suite.

For a VA environment also working as a wrapper tool, a portal that merges output from different software, the utilization for CFD work should be high since many analysts work with software of different origin for pre- and post processing as well as for the numerical solver. There should thus exist a natural interest in a platform to filter, control and display data from these different sources. Furthermore, due to the ever increasing challenge with handling extremely large data sets, the structure and analysis framework making it possible to extract specific data fast is becoming more important.

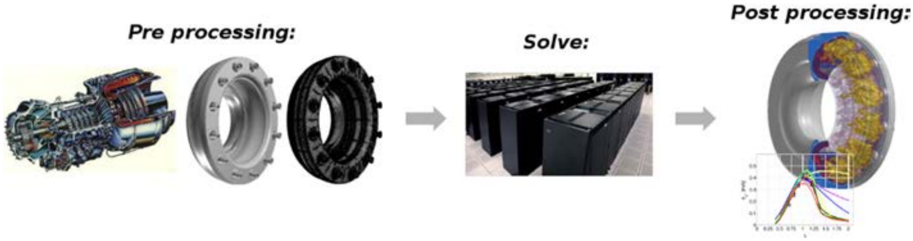


Figure 41. CFD analysis work flow. Problem formulation and setup in the pre-processing stage with a CAD representation of the physical domain and the subsequent generation of the computational grid used by the numerical method. This is followed by number crunching with help of physics models and numerical methods on large HPC compute resources in the solve phase. For the converged solution, visualization and interpretation of simulation final output is done during the post processing stage (Fureby, 2015).

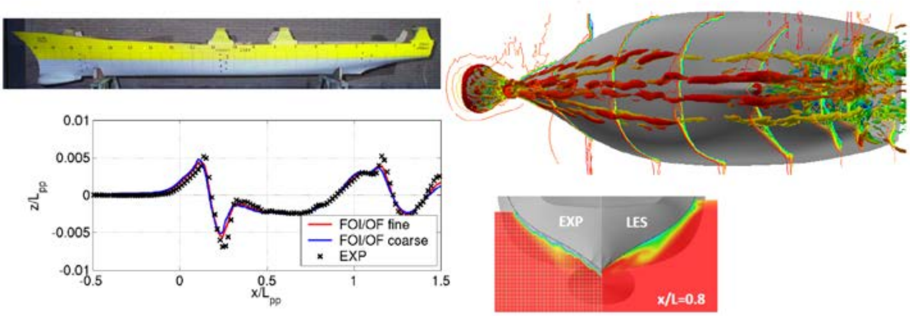


Figure 42. Surface ship hydrodynamics. Transient simulation including water surface and surface waves. Detailed comparison versus experimental data with respect to wave height, velocities, forces and moments (Svennberg, Fureby, Liefvendahl, & Alin, 2011). Upper left: Model of ship hull. Upper right: Instantaneous vorticity on bottom side of hull colored by velocity. Velocity contour plots in cross sections along the hull showing boundary layer profile. Lower left: Simulation vs. experiment. Surface capturing. Lower right: Cross-sectional comparison of experimental and simulated axial velocity fields.

Within the CFD community there is a large difference with respect to what type of analyzes are performed. The community ranges from mechanical engineers working close to production, to research groups. In production industry the lead time could be short which means that the methods and software used are designed to produce results fast. Within the research community, the focus could be to improve and use the state-of-the-art.

As mentioned earlier, many types of data is of interest to monitor and analyze during a typical CFD analysis. Primary solved physical variables (velocity, pressure, density, temperature etc.), numerical convergence data but also more computer related data such as storage consumption, cluster performance and job queueing. Complete integration of such data is often not available within typical CFD dedicated software and complete automation of this may be hard to obtain. A tool that makes this work fast and easy should, however, be most valuable. Furthermore, increased capability to access data for comparison is also in general of interest. This data could come from previous simulations, from experiments or elsewhere. Using a more advanced, dynamic, VA platform could improve overall quality with better macroscopic view of the project and thus, improved analysis capability.

As argued in (Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008) user acceptability and development of a thorough understanding of what these VA tools can do remains a challenge. Although many techniques have been presented, they are still not used widely outside the VA community. For the specific CFD community, VA tools are already well developed and used. Furthermore, fluid dynamics specialists are often skilled in both data mining and visualization. There is thus a good chance of future VA tools developed within the body of other research environments to be deployed and used with CFD as well. The computational fluid dynamics community would therefore most likely benefit a lot from the fact that VA is becoming widely used. To give a few references to more in-depth material, in e.g. (Hansen & Johnson, 2005), common techniques used to visualize fluid flow are presented. Streamlines, particle tracking, volume rendering etc. are discussed as well as a several suggestions on how to visualize vorticity for time resolved data. Several good references on visualization are also available in chapter 5. At present, there are several research and science centers focused on VA, e.g. the groups Information Visualization, Interactive Visualization, and Medical Visualization at Linköping University (http://www.itn.liu.se/mit/research?l=en), Visual Arena Research in Gothenburg (http://visualisering.gu.se/) and the HPCViz (www.kth.se/en/csc/forskning/hpc-viz) at the Royal Institute of Technology, KTH to name a few residing in Sweden.

7 Discussion and Conclusions

7.1 Reflections

Visual Analytics is a process that is heavily dependent on a *human in the loop*, as can be noted in Figure 4 where the human is controlling the processing occurring in the visualization pipeline. Expert knowledge and well-designed human-computer interaction are therefore required to fully exploit the power of VA. However, using freely available visualization tools is a low threshold entry point for anyone who wishes to perform an initial visual data analysis, which has the potential to lead to new insights.

As large quantity data analysis and management is becoming a more thoroughly integrated part of many disciplines, e.g. business, environmental monitoring, security, disaster and emergency management, software analytics and engineering physics to name a few (Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008), the methods and tools for VA are now rapidly developing as well. Related research fields may, provided that this is migrated across and implemented properly, have a huge advantage of these developments. This should definitely be true within the field of CFD, where the increased attention for VA in other research areas should contribute to the overall development of new visualization techniques for fluid dynamics purposes as well.

Visual analytics is a powerful tool for understanding data, however, it is a double-edged sword. At the same time that it provides the ability to come to conclusions quickly, it may tempt to bypass the requirements for data analysts and lead to unreliable and inaccurate decisions. Despite huge progresses in techniques and methods in visualizations, many of challenges that were stated by the pioneers in the field remain the same, one of the most crucial being "visualization and scientific truth". Maybe as relevant it is to discuss how VA can lead us to new insights and understanding the data, equally vital it is to ask "how can we avoid misleading our viewers?" (Brooks, 1993).

A trend in VA is to move the analysis and control from the data analyst to the decision maker (who, with the help of VA can "understand" the data independently of a trained and experienced analyst). However, we argue that this might be an unsafe move, since a trained analyst have greater knowledge of interpreting data, for example in terms of understanding statistics and limitations, and thus there is a risk that the data will be interpreted incorrectly.

Especially in a commercial context the Visual Analytics tools are too often presented as *intelligent* assistants to the decision-maker. Tools that seemingly automatically process and adapt the visual content to the needs of the user. See for example http://vis.pnnl.gov/ “We take into consideration the application of human judgment to make the best possible use of incomplete, inconsistent, and potentially deceptive information in the face of rapidly changing situations”.

If VA is used wisely the visual analytic tools can help decision-makers to make sense of a situation, and based on that take correct decisions. However, it should be emphasized that focus is on humans involved in the decision process, and humans choosing how to view and analyze the data.

Proponents of Visual Analytics often claim that the Visual Analytics process is data-centric i.e. discovers hypotheses directly from data. As with the claims of intelligent VA tools, such statements must be handled carefully. A hypothesis can be formed based on what is found in the data during the VA process, but if the data is also used to confirm the hypothesis there is a prevalent risk of self-fulfillment as the data might confirm itself. In addition, there might be concerns with multiple layers of theory since hypotheses are also needed for transforming sensor signals to ‘data’, be it sensors measuring physical phenomena or, even worse, sensors in the form of complex algorithms scanning vast amounts of texts or financial data.

Two main fallacies can occur in attempts of working with discoveries of hypotheses directly from data, be it through VA or through correlation-finding algorithms for Big Data. These are 1) that causality cannot be concluded from correlations and 2) sampling bias. A hypothesis that is useful for real-world decisions must say something about causation e.g. Ebola is spread from fruit bats to humans. This means presumably that the prevalence of Ebola infections in fruit bats and in humans are correlated although the hypothesis cannot be derived from the correlation. However, it can be falsified by lack of correlation in the data. Likewise, stock market prices are correlated to the spread of Ebola but it would be wrong to conclude that Ebola cases are caused by stock market losses since it in general is impossible to derive causality from correlations. It is beneficial to view hypotheses as human conjectures to be pruned by comparing with data. Sampling bias makes it dangerous to assume that Twitter streams represent attitudes and feelings of a nation. Twitter users in the U.S. are for example predominantly young, urbanized and black which means that Twitter data is an indicator of the opinions and sentiments of a significant U.S. sub-population but not the entire population.

7.2 Research Perspectives

7.2.1 3D Virtual-reality

Based on the success and performance of the Oculus Rift[9] Virtual Reality Helmet we believe that 3D virtual-reality environments will be a key aspect in future implementations of Visual Analytics. The Oculus Rift is a head-mounted display that following a Kickstarter campaign was developed by Oculus VR. The user is mobile while wearing the headset and get the experience of being immersed in a fully 3D virtual world which makes gaming the main application of the technology. Facebook acquired Oculus VR 2014. Samsung released an Oculus Rift clone 2015 marketed as Gear VR[10]. Because of the gaming market supporting the development of advanced VR headsets, we think that VR helmets will be widely available at a reasonable price in the near future and that it therefore will be quite feasible to use VR headsets for Visual Analytics. This opportunity will drive Visual Analytics research in a new direction.

7.2.2 Semantic Hashing with Deep Learning

Deep Learning is a comparatively new machine learning technology with a potential for approaching human-like intelligence. Using deep layers of autoencoders[11] it is possible to perform dimensional reduction of input data. This technique, called “semantic hashing”, has been used for indexing and retrieving documents in semantically relevant classes. We think that such semantic hashing will be widely used in Visual Analytics.

7.2.3 Decision Support

Although the purpose of decision support is ultimately to support decisions or actions that have a positive impact on some operation of interest to the user, current research on visualization for decision support seems to reach no further than just to accurately present information. For instance, in a research paper focusing on the “visualization-based decision support” (Sauter, Mudigonda, Subramanian, & Creely, 2011), the most stressing issue seems to be the selection of a suitable visualization tool to aid decision making, without considering the following step of how to present the actual decision or action options. In chapter 4, we address that “missing part” of the decision process, that of selection of options, which could be part of the decision support (and sometimes “should be

[9] http://www.ign.com/wikis/oculus-rift

[10] https://en.wikipedia.org/wiki/Samsung_Gear_VR

[11] https://en.wikipedia.org/wiki/Autoencoder

part of" due to the need for automatic management of a large number of time-varying options). At FOI, in the SBFP project (described in section 7.3.2 below), an effort is now made to complement the pure visualization of data with the commander's operational action options.

7.3 Relevance for Defense

This section describes options for using Visual Analytics in military contexts with concrete application in some ongoing FOI projects.

7.3.1 ONSIM

Intelligent On-line Simulation Support for Operational Battle Management (ONSIM) is an ongoing FOI-project aiming at creating a new model for leading and managing air combat. The key idea is to use advanced simulation to predict the outcome of possible moves combined with AI for managing the simulation resources and compiling reports for military leaders. Commanders use a graphical interface to sketch plans such as for example to commit all available air defense resources to thwart an ongoing attack. The AI maps out possible implementations of the sketch and starts a set of simulations for analyzing possible long-run outcomes and analyzes conclusions that can be drawn from the simulation results. Firm conclusions are presented for the commander via the graphical interface. A possible overall conclusion might be that the ongoing attack will be defeated but the second attack wave will break through because of depleted defense resources. We can view the simulations as means for collating a vast database of possible events and evolutions in the ongoing air battle. The graphical user interface supports a Visual Analytics process in which the commander interacts with the system by suggesting possible plans and gets feedback as intelligent digests of simulation results. The inbuilt AI handles the detailed mechanics of initializing and starting simulations. From the point of view of the commander and his/her staff, operational battle management is transformed to a Visual Analytics process where people suggest plans and machines propose likely outcomes. One of the partners of ONSIM is Thales who provides graphical tools for enhancing situational awareness in air combat[12] with potential for integrating the ONSIM analytical tools.

[12] https://www.thalesgroup.com/sites/default/files/asset/document/WebS%C2%B2AT.pdf

7.3.2 Simulation-based Operation Planning

The *Simulation-based operation* planning project (SBFP) is focused on using simulation-based methods for analysis of commander questions concerning military operation planning. These methods are used to evaluate alternative military scenarios and plans, involving resources used for military operations.

The thousands of detailed simulations cover a range of input parameter values and output a number of performance values. These are collected and statistics are visualized to the commander in a multitude of ways. An example is shown in Figure 43, where, briefly, the importance of various simulation parameters on the beneficial result is shown (Schubert, Johansson, & Hörling, 2015).

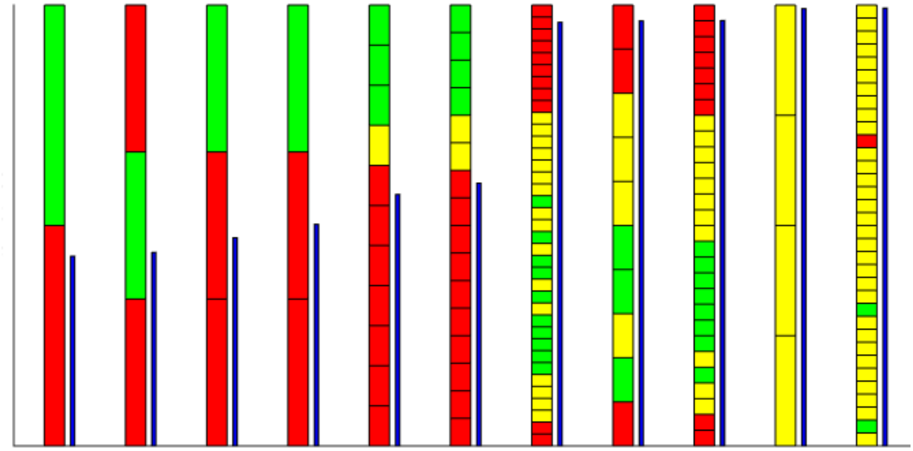


Figure 43. A visualization to compare the statistics of some parameters based on a large number of simulations.

7.3.3 Uncertainty in Terrain Analysis

Terrain Analysis for Simulation Applications is an ongoing FOI project that among others develops a geo-analysis software library that can be integrated in different simulation environments with the focus on path planning and 3D line-of-sight analysis. This project employs a method for line-of-sight analysis that does not give an answer of the binary type (0 = no sight, 1 = free sight), but incorporates uncertainty in the height data in the analysis process and provides the probability of the free sight (a value between 0 and 1), see (Tolt, Follo, Hedström, & Härje, 2014).

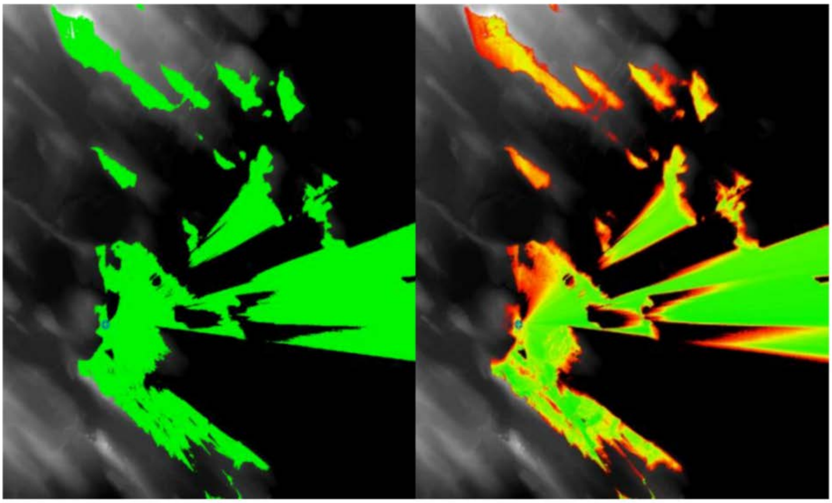


Figure 44. Left: traditional line-of-sight analysis. Right: Probability based line-of-sight analysis that takes into the account the uncertainty in the data.

Real-time line-of-sight analysis has many applications in military operational planning, for instance in urban terrains. Introducing uncertainty to the analysis and permitting the user to interact with the visualized result (e.g. changing probability measures) provides a far more advantageous decision support tool to the planner.

7.3.4 Decision Support with Multiple Options

Information Interoperability and Intelligence Interoperability by Statistics, Agents, Reasoning and Semantics (IN-4-STARS2.0) is a European defense agency (EDA) project[13] involving FOI and partners from Estonia and the Netherlands. IN-4-STARS2.0 deals with various aspects of secure interoperability and processing of heterogeneous information in a networked information exchange infrastructure for large-scale intelligence analysis.

The FOI part of the project primarily consists of supporting an intelligence analyst, who is monitoring the collection of information from distributed heterogeneous information sources (including sensor data and social media information), information management, and inference for threat estimation. Based on the processed information, the analyst should be prepared to respond to

[13] EDA project no. B 0983 IAP4 GP (IN-4-STARS)

questions about the security status of a particular region of interest. To aid the analyst, he/she has a decision support tool which presents threat levels, with the help of a library of inference models (in our case Bayesian networks).

The collected information is typically highly uncertain and sometimes it is even difficult to know which potential threat it concerns, resulting in multiple alternative interpretations of the data (Johansson, Horndahl, & Rosell, 2015). The analyst’s need for Visual Analytics is hence strongly related to the discussion in chapter 4 about visualization of multiple options.

In Figure 45, an initial prototype of an analyst’s interface is sketched. On the far right is a tree structure, which concisely depicts all alternative data interpretations, i.e. each branch in the tree represents a particular sequence of interpretations of the collected data so far. In this case, each branch also represents an instantiation of nodes in a Bayesian network-based threat model and updated probabilities on all involved variables. As the number of branches (interpretations) grows exponentially with each new collected data, the tree also allows the user an opportunity to use its expertise to cut unlikely branches.

To the left of the tree are three boxes entitled Avg, Max and Min. Max and Min are the two alternative interpretations that yield the highest and lowest value on the analyst’s threat variable of interest, 80% and 10% respectively, in the example. The instantiated Bayesian networks for each case is also shown. Avg shows a mean threat probability over the whole set of interpretations.

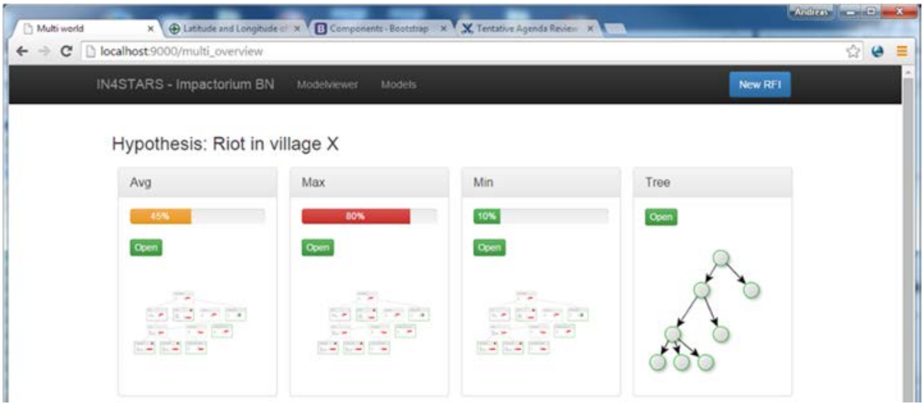


Figure 45. Prototype visualization of multiple options as alternative interpretations of data (illustration by Andreas Horndahl).

7.4 Relation to the Teenage Mutant Ninja Turtles Universe

Some readers might miss a logical flow in the chapter structure of this report. We could for example have organized chapters according to a process pipeline or systematically covering the relation to various neighboring scientific domains. However, each author has written a chapter that typically focus on some problematic point of Visual Analytics. One could argue that the chapters in this report follow a different dimension as compared to the process pipeline in Figure 4. This is correct but best described by comparing to the Teenage Mutant Ninja Turtles Universe (TMNT). In TMNT there is a normal world that works more or less according to normal logic and reason as well as an extra dimension X from which evil monsters and green slime[14] oozes out. The mission of the Turtles is to contain such breakouts from dimension X. Each chapter in this report can hence be viewed as some problematic issue discharging from the otherwise beautifully organized and marketed edifice of Visual Analytics. The authors have brought out their best moves to restore order to it all.

[14] http://tmnt2012series.wikia.com/wiki/Mutagen_Ooze

8 Acknowledgements

The authors wish to thank Ekaterina Fedina, Prof. Christer Fureby, Prof. Mattias Liefvendahl, Dr. Oskar Parmhed and Andreas Horndahl at FOI for providing us with valuable illustrations. We also extend our gratitude to Dr. Katerina Vrotsou, and to Dr. Jennifer Frazier at the Exploratorium in San Francisco for allowing the use of their images.

We also wish to Björn Thuresson at KTH for his comments on our work.

We finally acknowledge our use of the image service openclipart.[15]

[15] https://openclipart.org/

9 Bibliography

Allendes Osorio, R. S., & Brodlie, K. W. (2008). Contouring with uncertainty. In I. Lim, & W. Tang (Eds.), *Theory and Practice of Computer Graphics* (pp. 1-7). The Eurographics Association.

Bach, B., Shi, C., Heulot, N., Madhyastha, T., Grabowski, T., & Dragicevic, P. (2016). Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 559-568.

Bertini, E. (2011). *How do you visualize too much data?* Retrieved Nov 3, 2015, from Fell in love with data: http://fellinlovewithdata.com/guides/how-do-you-visualize-too-much-data

Bivall, P. (2010). *Touching the essence of life: haptic virtual proteins for learning.* Norrköping, Sweden: Department of Science and Technology, Linköping University.

Bonneau, G.-P., Hege, H.-C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., & Schultz, T. (2014). Overview and state-of-the-Art of uncertainty visualization. In C. D. Hansen, M. Chen, C. R. Johnson, A. E. Kaufman, & H. Hagen (Eds.), *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization* (pp. 3-27). Springer.

Borgo, R., Kehrer, J., Chung, D., Maguire, E., Laramee, R., Hauser, H., . . . Chen, M. (2013). Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. (M. Sbert, & L. Szimay-Kalos, Eds.) *Eurographics*, 1-25.

Borkin, M., Vo, A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2306-2315.

Bostock, M. (2015). *Data-Driven Documents*. Retrieved Nov 5, 2015, from D3js: http://d3js.org/

Brodlie, K., Allendes Osorio, R., & Lopes, A. (2012). A Review of Uncertainty in Data Visualization. In J. Dill, R. Earnshaw, D. Kasik, J. Vince, & P. C. Wong, *Expanding the Frontiers of Visual Analytics and Visualization* (pp. 81-109). Springer-Verlag London Limited.

Brooks, J. F. (1993). Keynote address: A vision for visualization. In G. M. Nielson, & D. Bergeron (Ed.), *Proceedings of 4th IEEE Visualization Conference* (p. 2). San Jose, California: IEEE Computer Science Press.

Brown, R. (2004). Animated visual vibrations as an uncertainty visualisation technique. *Proceedings of the International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia* (pp. 84-89). Singapore: ACM Press.

Byrne, L., Angus, D., & Wiles, J. (2016). Acquired Codes of Meaning in Data Visualization and Infographics: Beyond Perceptual Primitives. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 509-518.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think.* San Francisco: Morgan Kaufmann Publishers Inc.

Cedilnik, A., & Rheingans, P. (2000). Procedural Annotation of Uncertain Information. *Proceedings of the 11th IEEE Visualization Conference* (pp. 77-84). IEEE Computer Society Press.

Chapuis, M., Fedina, E., Fureby, C., Hannemann, K., Karl, S., & Martinez Schramm, J. (2013). A computational study of the HyShot II combustor performance. *Proceedings of the Combustion Institute, 34*(2), 2101–2109.

Chen, H., Chen, W., Mei, H., Liu, Z., Zhou, K., Chen, W., . . . Ma, K.-L. (2014). Visual Abstraction and Exploration of Multi-class Scatterplots. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1683-1692.

Cheng, S., & Mueller, K. (2016). The Data Context Map: Fusing Data and Attributes into a Unified Display. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 121-130.

Coninx, A., Bonneau, G.-P., Droulez, J., & Thibault, G. (2011). Visualization of uncertain scalar data fields using color scales and perceptually adapted noise. *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV)* (pp. 59-66). Toulouse, France: ACM.

Dzemyda, G., Kurasova, O., & Žilinskas, J. (2013). *Multidimensional Data Visualization: Methods and Applications.* New York: Springer.

Ehlschlaeger, C. R., Shortridge, A. M., & Goodchild, M. F. (1997). Visualizing spatial data uncertainty using animation. *Computers and Geosciences, 23*(4), 387-395.

Fedina, E. (2014). *TNT/Aluminum Afterburning in Air Blasts.* Stockholm: FOI - Swedish Defence Research Agency.

Feng, D., Kwock, L., Lee, Y., & Taylor, R. M. (2010). Matching visual saliency to confidence in plots of uncertain data. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 980-989.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis, 5*, 373-384.

Forsell, C. (2010, July). A Guide to Scientific Evaluation in Information Visualization. *Proceedings of the 14th International Conference on Information Visualisation* (pp. 162-169). IEEE.

Forsell, C., & Cooper, M. (2012). A Guide to Reporting Scientific Evaluation in Visualization. *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 608-611). New York, NY, USA: ACM.

Forsell, C., & Johansson, J. (2010). An Heuristic Set for Evaluation in Information Visualization. *Proceedings of the International Conference on Advanced Visual Interfaces* (pp. 199-206). New York, NY, USA: ACM.

France, S. L., & Carroll, J. D. (2011). Two-Way Multidimensional Scaling: A Review. *IEEE Transactions on Systems, Man and Cybernetics, 41*(5), 644-661.

Fureby, C. (2015, Sept 20). Private communication. *FOI - Swedish Defence Research Agency*. Stockholm, Sweden.

Grigoryan, G., & Rheingans, P. (2004). Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics, 10*(5), 564-573.

Haber, R. B., & McNabb, D. A. (1990). Visualization idioms: a conceptual model for scientific visualization systems. In B. Shriver, G. M. Nielson, & L. J. Rosenblum (Eds.), *Visualization in scientific computing* (pp. 74-93). IEEE.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* The MIT Press.

Hansen, C. D., & Johnson, C. R. (2005). *The visualization Handbook.* Elsevier.

Hengl, T. (2003). Visualisation of uncertainty using the HSI colour model: computations with colours. *Proceedings of the 7th International Conference on GeoComputation*, (pp. 8-17).

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plotdensity trace synergism. *The American Statistician*, *52*(2), 181-184.

Jansen, Y., Dragicevic, P., Isenberg, P., Alexander, J., Karnik, A., Kildal, J., . . . Hornbæk, K. (2015). Opportunities and Challenges for Data Physicalization. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3227-3236). New York, NY, USA: ACM.

Johansson, J. (2008). *Efficient Information Visualization of Multivariate and Time-Varying Data (Ph.D. dissertation).* Linköping University, Department of Science and Technology.

Johansson, J., & Forsell, C. (2016). Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 579-588.

Johansson, J., Forsell, C., & Cooper, M. (2014). On the usability of three-dimensional display in parallel coordinates: Evaluating the efficiency of identifying two-dimensional relationships. *Information Visualization, 13*(1), 29-41.

Johansson, R., Horndahl, A., & Rosell, M. (2015). A Data Association Framework for General Information Fusion. *Proceedings of the International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2015).* San Diego, USA.

Johansson, R., Nilsson, S., & Pelzer, B. (2014). *Uncertain databases: Scanning the research frontier.* Stockholm: FOI - Swedish Defence Research Agency.

Johnson, C. R., & Sanderson, A. (2003). A Next Step: Visualizing Errors and Uncertainty. *Computer Graphics and Applications, IEEE, 23*(5), 6-10.

Johnston, W. (2002). Model visualization. In U. Fayyad, G. G. Grinstein, & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery* (pp. 223-227). Academic Press.

Juang, K.-W., Chen, Y.-S., & Lee, D.-Y. (2004). Using sequential indicator simulation to assess the uncertainty of delineating heavy-metal contaminated soils. *Environmental Pollution, 27*(2), 229-238.

Jänicke, H., Wiebel, A., Scheuermann, G., & Kollmann, W. (2007). Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1384-1391.

Karami, A., & Johansson, R. (2014). Utilization of multi attribute decision making techniques to integrate automatic and manual ranking of options. *Journal of Information Science and Engineering, 30*(2), pp. 519-534.

Kay, M., & Heer, J. (2016). Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics , 22*(1), 469-478.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. In S. J. Simoff, M. H. Böhlen, & A. Mazeika (Eds.), *Visual Data Mining, Theory* (pp. 76-90). Springer Link.

Kielman, J., Thomas, J., & May, R. (2009). Foundations and frontiers of visual analytics. *Information Visualization, 8*(4), 239-246.

Kohlhammer, J., Keim, D., Pohl, M., Santucci, G., & Andrienko, G. (2011). Solving problems with visual Analytics. *Proceedings of the 2nd European Future Technologies Conference and Exhibition, 7,* pp. 117-120.

Lee, S., Kim, S., Hung, Y., Lam, H., Kang, Y., & Yi, J. (2016). How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 499-508.

Li, H., Fu, C.-W., Li, Y., & Hanson, A. J. (2007). Visualizing large-scale uncertainty in astrophysical data. *IEEE Transactions on Visualization and Computer Graphics,* 1640-1647.

Love, A. L., Pang, A., & Kao, D. L. (2005). Visualizing spatial multivalue data. *IEEE Computer Graphics and Applications, 25*(3), 69-79.

Nelson, B., & Kirby, R. M. (2006). Ray-tracing polymorphic multidomain spectral/hp elements for isosurface rendering. *IEEE Transactions on Visualization and Computer Graphics, 12*(1), 114–126.

Nelson, B., Haimes, R., & Kirby, R. M. (2011). GPU-based interactive cut-surface extraction from high-order finite element fields. *IEEE Transactions on Visualization and Computer Graphics, 17*(12), 1803–1811.

Newman, T. S., & Lee, W. (2004). On visualizing uncertainty in volumetric data: techniques and their evaluation. *Journal of Visual Languages & Computing, 15*(6), 463–491.

Norberg, J., & Westerlund, F. (2014). *Russia and Ukraine: Military-strategic options, and possible risks, for Moscow.* FOI - Swedish Defence Research Agency.

Ottley, A., Peck, E., Harrison, L., Afergan, D., Ziemkiewicz, C., Taylor, H., . . . Chang, R. (2016). Improving Bayesian Reasoning: The Effects of

Phrasing, Visualization, and Spatial Ability. *IEEE Transactions on Visualization and Computer Graphics , 22*(1), 529-538.

Pang, A. T., Wittenbrink, C. M., & Lodh, S. K. (1996). Approaches to uncertainty visualization. *The Visual Computer, 13*(8), 370-390.

Parmhed, O., Svennberg, U., Burman, J., & Thaning, L. (2004). *Large Eddy Simulation of Urban Dispersion.* Stockholm: FOI - Swedish Defence Research Agency.

Potter, K., Kniss, J., Riesenfeld, R., & Johnson, C. R. (2010). Visualizing summary statistics and uncertainty. *Eurographics/IEEE-VGTC Symposium on Visualization,* 823-832.

Potter, K., Wilson, A., Bremer, P.-T., Williams, D., Doutriaux, C., Pascucci, V., & Johnson, C. R. (2009). Ensemble-Vis: a framework for the statistical visualization of ensemble data. *Proceedings of the 2009 IEEE international conference on data mining workshops* (pp. 233-240). IEEE Computer Society.

Pöthkow, K., & Hege, H.-C. (2010). Positional uncertainty of isocontours: condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*.

Pöthkow, K., Weber, B., & Hege, H.-C. (2011). Probabilistic marching cubes. *IEEE Symposium on Visualization, 30*(3).

Raidou, R., Eisemann, M., Breeuwer, M., Eisemann, E., & Vilanova, A. (2016). Orientation-Enhanced Parallel Coordinate Plots. *IEEE Transactions on Visualization and Computer Graphics , 22*(1), 589-598.

Rhodes, P. J., Laramee, R. S., & Bergeron, R. D. (2003). Uncertainty visualization methods in isosurface rendering. In M. Chover, H. Hagen, & D. Tost (Ed.), *Proceedings of Eurographics.* The Eurographics Association.

Rubio-Sanchez, M., Raya, L., Diaz, F., & Sanchez, A. (2016). A comparative study between RadViz and Star Coordinates. *IEEE Transactions on Visualization and Computer Graphics , 22*(1), 619-628.

Ruchikachorn, P., & Mueller, K. (2015). Learning Visualizations by Analogy: Promoting Visual Literacy through Visualization Morphing. *IEEE Transactions on Visualization and Computer Graphics , 21*(9), 1028-1044.

Sanyal, J., Zhang, S., Dyer, J., Mercer, A., Amburn, P., & Moorhead, R. J. (2010). Noodles: a tool for vsualization of numerical weather model.

IEEE Transactions on Visualization and Computer Graphics, 16(6), 1421-1430.

Sauter, V. L., Mudigonda, S., Subramanian, A., & Creely, R. (2011). Visualization-based decision support systems: An example of regional relationship data. *International journal of Decsion Support Systems, 3*(1), pp. 1-20.

Scheidegger, C. (2015). *Data Visualization Principles: interaction, filtering and aggregation.* Retrieved Nov 13, 2015, from http://cscheid.net/courses/spr15/cs444/lectures/Interaction.pdf

Schubert, J., & Hinshaw, F. (2011). *Data Farming En omvärldsanalys (in Swedish).* Stockholm: FOI - Swedish Defence Research Agency.

Schubert, J., Johansson, R., & Hörling, P. (2015). Skewed distribution analysis in simulation-based operation planning. *Proceedings of the 9th Operations Research and Analysis Conference,* (pp. 1-13). Ottobrunn, Germany.

Siirtola, H. (2007). *Interactive Visualization of Multidimensional Data.* Tampere: University of Tampere.

Simoff, S. J., Böhlen, M. H., & Mazeika, A. (2008). *Visual Data Mining: theory, techniques and tools for visual analytics.* Springer-Verlag.

Stusak, S., & Aslan, A. (2014). Beyond Physical Bar Charts: An Exploration of Designing Physical Visualizations. *Proceedings of the Extended Abstracts of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1381-1386). New York, NY, USA: ACM.

Suciu, D., Olteanu, D., Ré, C., & Koch, C. (2011). *Probabilistic databases.* Morgan & Claypool.

Svennberg, U., Fureby, C., Liefvendahl, M., & Alin, N. (2011). Large Eddy Simulation of flow past the DTMB 5415 surface combatant hull with and without bilge keels. In L. Eça, E. Oñate, J. García, T. Kvamsdal, & P. Bergan (Ed.), *5th International Conference on Computational Methods in Marine Engineering.*

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the Path - The Research and Development Agenda for Visual Analytics.* (J. J. Thomas, & K. A. Cook, Eds.) Washington, DC: IEEE Computer Society Press.

Tolt, G., Follo, P., Hedström, J., & Härje, T. (2014). *Omvärldsmodellering: slutrapport 2014 [Synthetic Environment Modeling: final report 2014].* Stockholm: FOI - Swedish Defence Research Agency.

Tominski, C., Forsell, C., & Johansson, J. (2012). Interaction Support for Visual Comparison Inspired by Natural Behavior. *IEEE Transactions on Visualization and Computer Graphics , 18*(12), 2719-2728.

Tufte, E. R. (1990). *Envisioning Information.* (E. R. Tufte, Ed.) Cheshire, Connecticut, USA: Graphic Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). (E. R. Tufte, Ed.) Cheshire, Connecticut, USA: Graphic Press.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison Wesley.

Wittenbrink, C. M., Pang, A. T., & Lodha, S. K. (2006). Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics, 2*(3), 266-279.

Vrotsou, K. (2010). *Everyday mining: exploring sequences in event-based data.* Norrköping: Department of Science and Technology, Linköping University.

Xie, Z., Huang, S., Ward, M. O., & Rundensteiner, E. A. (2006). Exploratory visualization of multivariate data with variable quality. *Proceedings of the IEEE Symposium on Visual Analytics Science & Technology,* (pp. 183-190).

Yalcin, M., Elmqvist, N., & Bederson, B. (2016). AggreSet: Rich and Scalable Set Exploration using Visualizations of Element Aggregations. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 688-697.

Yu, L., Efstathiou, K., Isenberg, P., & Isenberg, T. (2016). CAST: Effective and Efficient User Interaction for Context-Aware Selection in 3D Particle Clouds. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 886-895.

Zehnera, B., Watanabea, N., & Kolditz, O. (2010). Visualization of gridded scalar data with uncertainty in geosciences. *Computers & Geosciences, 36*(10), 1268-1275.

Zhou, H., Xu, P., Ming, Z., & Qu, H. (2014). Parallel Coordinates with Data Labels. *Proceedings of the 7th International Symposium on Visual Information Communication and Interaction* (pp. 49-57). New York, NY, USA: ACM.

Zuk, T., Downton, J., Gray, D., Carpendale, S., & Liang, D. J. (2008). Exploration of uncertainty in bidirectional vector fields. *Proceedings of SPIE-IS&T Conference on Electronic Imaging.*

FOI, Swedish Defence Research Agency, is a mainly assignment-funded agency under the Ministry of Defence. The core activities are research, method and technology development, as well as studies conducted in the interests of Swedish defence and the safety and security of society. The organisation employs approximately 1000 personnel of whom about 800 are scientists. This makes FOI Sweden's largest research institute. FOI gives its customers access to leading-edge expertise in a large number of fields such as security policy studies, defence and security related analyses, the assessment of various types of threat, systems for control and management of crises, protection against and management of hazardous substances, IT security and the potential offered by new sensors.

FOI
Swedish Defence Research Agency
SE-164 90 Stockholm

Phone: +46 8 555 030 00
Fax: +46 8 555 031 00

www.foi.se