Aalto University
School of Science

Department of Information and Computer Science

Paula Järvinen

A DATA MODEL BASED APPROACH FOR VISUAL ANALYTICS OF MONITORING DATA

Thesis submitted in partial fulfillment of the requirements for the degree of Licentiate of Science in Technology.
Espoo, April 10, 2013
Supervisor: Professor Samuel Kaski
Instructor: Docent Kai Puolamäki
Inspector: Dr. Matti Gröhn

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of licentiate thesis

Author Paula Järvinen	
Title of thesis A data model based approach for visual analytics of monitoring data	
Department Department of Information and Computer Science	
Field of research Information technology	
Supervising professor Samuel Kaski	**Code of professorship** T-61
Thesis advisor(s) Kai Puolamäki	
Thesis examiner(s) Matti Gröhn	
Number of pages 135 +20	**Language** English
Date of submission for examination 13.03.2013	

Abstract

Data modelling is a well-established method in software engineering. This work explores its use in the emerging field of visual analytics. Visual analytics is a recent approach to finding knowledge from data masses. It combines the strengths of automatic data processing and the visual perception and analysis capabilities of the human user. The approach has its roots in information visualization and data analysis, in which the use of data models is not common practice.

The backbone of this work is the domain data model. The model incorporates the main concepts of a given domain, which remain similar regardless of the application, but which can be tuned for visualization and analysis purposes. The work proposes three uses for data models. The first is the construction of visual analytics applications in the domain. The second is supporting reasoning with the help of metadata. The third is using the data model as an approach to visualize large data spaces.

The study focuses on the analysis of monitoring data, which is nowadays collected in vast amounts and from a wide variety of fields. The approach is evaluated using two cases from different applications in the monitoring data domain: analysing the eating and exercise habits of dieting people, and studying the energy efficiency and indoor conditions of buildings. In addition to the approach and the evaluation cases, the work introduces visual analytics, data modelling and monitoring data, and discusses the evaluation of visual analytics. The multi-discipline research area of visual analytics is represented in the form of a framework constructed as a part of this work.

The results suggest that data modelling is a useful method in visual analytics. A domain model approach can save effort in constructing new visual analytics applications. Supporting reasoning and browsing data with the help of the data model would be especially useful for users who are not so familiar with data analysis, but know the application domain well. Combining the data model approach with descriptive visualizations can bring powerful tools for analysing data.

Keywords Visual analytics, information visualization, data modelling, monitoring data

Lisensiaatin tutkimuksen tiivistelmä

Tekijä Paula Järvinen		
Työn nimi Tietomallien hyödyntäminen monitorointidatan visuaalisessa data-analyysissä		
Laitos Tietojenkäsittelytieteen laitos		
Tutkimusala Informaatiotekniikka		
Vastuuprofessori Samuel Kaski	**Professuurikoodi** T-61	
Työn ohjaajat Kai Puolamäki		
Työn tarkastajat Matti Gröhn		
Jätetty tarkastettavaksi 13.03.2013	**Sivumäärä** 135 +20	**Kieli** Englanti

Tiivistelmä

Tiedon mallinnus on vakiintunut ohjelmistotekniikan menetelmä. Työssä tutkitaan sen soveltamismahdollisuuksia visuaaliseen data-analyysiin. Visuaalinen data-analyysi on tuore tutkimusalue, jonka tavoitteena tietämyksen esiin saaminen tietomassoista. Siinä pyritään hyödyntämään ihmisen visuaalisia kykyjä ja päättelytaitoja yhdessä automaattisten analyysimenetelmien kanssa. Tiedon mallinnuksen käyttäminen visuaalisessa data-analyysissä ei ole vakiintunutta.

Työn perustana on sovellusaluekohtainen tietomalli. Malli sisältää sovellusalueen keskeiset käsitteet, jotka pysyvät samoina yksittäisestä sovelluksesta toiseen. Lisäksi malli huomioi visualisointi- ja analyysimenetelmien tarpeet. Työssä ehdotetaan kolmea käyttötapaa tietomallille. Ensimmäinen on analyysisovellusten tuottaminen malliin perustuvan alustan avulla, toinen on käyttäjän päättelyn tukeminen malliin talletetun metadatan perusteella ja kolmas laajojen tietoaineistojen selailu mallista generoidun näkymän avulla.

Työssä keskitytään monitorointidataan, jota kerätään nykyisin yhä useammilta alueilta. Lähestymistapaa arvioidaan soveltamalla sitä kahteen monitorointisovellukseen: ihmisten ruokailu- ja liikuntatapojen analysointiin sekä rakennusten sisäilmaolosuhteiden ja energiatehokkuuden arviointiin. Lisäksi työssä esitetään laaja katsaus visuaaliseen data-analyysiin, tiedon mallinnukseen, monitorointidataan sekä visuaalisen data-analyysin käytettävyyden arviointiin.

Tulokset antavat olettaa, että sovellusaluekohtaisesta tietomallista on hyötyä visuaalisessa data-analyysissä. Malliin perustuva alusta voi nopeuttaa uusien sovellusten tuottamista. Mallin tarjoama päättelytuki ja aineiston selailunäkymä hyödyttävät erityisesti kohdealueen asiantuntijoita, jotka eivät ole data-analyysin tuntijoita. Yhdistämällä lähestymistavan tarjoamiin mahdollisuuksiin kohdetta kuvaavia visualisointeja, voidaan saada aikaa tehokkaita työkaluja tiedon analysointiin.

Avainsanat visuaalinen data-analyysi, tiedon visualisointi, tiedon mallinnus, monitorointi

Acknowledgements

I would like to thank my advisor, Docent Kai Puolamäki, for the valuable advice and support he has given me in the writing of this thesis. I would also like to thank my advisor Professor Samuel Kaski for leading me through the final stages of the process. Thanks also to nutritionist Leena Metsäranta for valuable discussions, and to all other participants of the user evaluations. My deepest thanks go to my colleagues at VTT, especially Sanni, and my friends and family for their understanding and support. I am also grateful to my employer VTT for the financial contribution.

Contents

1 Introduction

Problem

Huge amounts of measurement data are collected in databases. In industry, sensors monitor the state of processes and machinery, while building automation systems store building information, environmental meters collect measurement data from the environment, and enterprises track the actions and behaviour of consumers. This data is expected to contain important knowledge that can be used to improve processes and support decision making. Utilising this collected data is not, however, a straightforward task. Analysing the data requires special skills, methods and tools that are usually unfamiliar to the data users or decision makers. Similarly, specialists of these tools and methods may lack the expertise in the application domain that is needed to effectively interpret the results. In addition, analysis tools require data to be available in simple files when, in reality, this is seldom the case. Instead, data normally exists in disparate forms and storages, ranging from unstructured data to operational databases.

Figure 1 illustrates the phases of data analysis and the problems related to them. Data coming in different forms from several sources requires a considerable amount of pre-processing; the relevant data must be recognized and extracted from the data sources, and noisy data must be cleaned. In operational databases the interesting and important features can be hidden behind complex structures and codes. Variables also often exist in non-comparable units that have to be harmonized. Analysis and visualization methods require specific data formats, requiring the data to be transformed accordingly. The sheer variety of data and analysis methods can be confusing to the analyst. Even a simple operational database can contain dozens of data entities, and each entity dozens of attributes. From this, the analyst has to find the most meaningful properties and attribute combinations, as well as the best methods to analyse them.

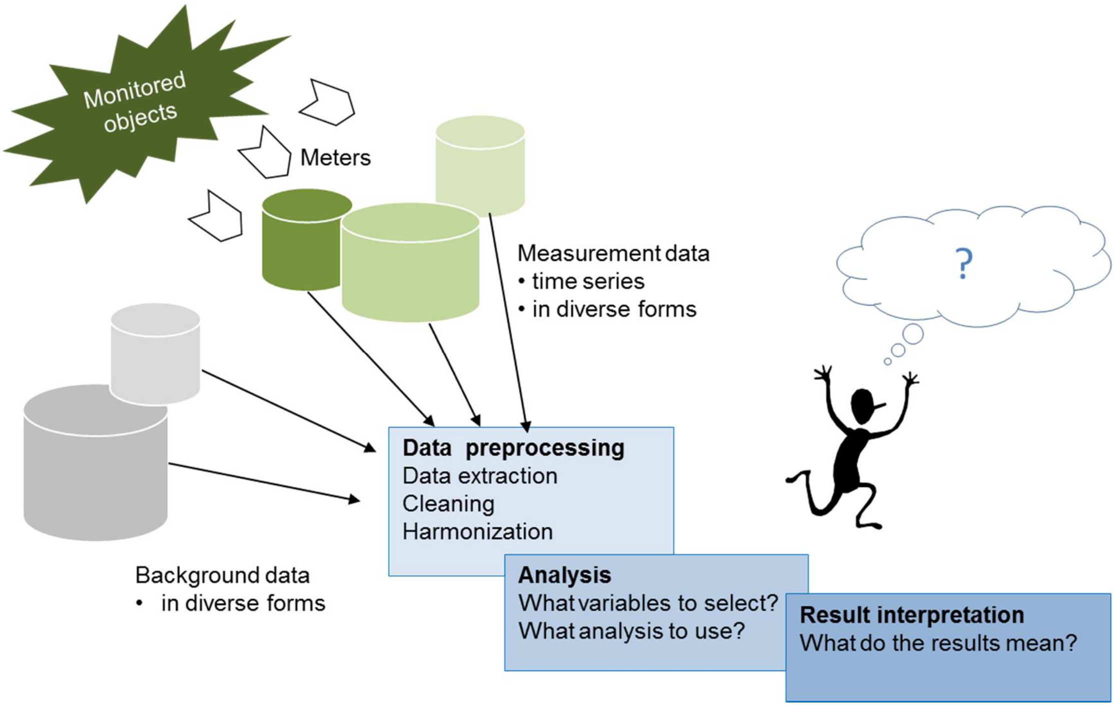


Figure 1. Phases and problems of data analysis.

Visual analytics

Visual analytics is a recent approach to analysing accumulated data resources. It promises easy-to-use tools that combine interactive visualizations and data analysis techniques. The approach is based on the utilisation of human skills to interpret visual representations in combination with automatic data analysis. Visual analytics research and tool development have emerged in the recent years, with visual analytics conferences and a dedicated journal, and tools developed by research laboratories as well as by commercial companies. These tools vary from application-specific solutions to products of established business analytics vendors (SAP, IBM, ORACLE, SAS).

Data models in visual analytics

This work examines how data models could be utilized in visual analytics. Data modelling is an established method in software engineering and has many uses in the software development process. A data model defines the key concepts, their relationships, attributes and data types of an application domain. The model forms a template for storing data and constructing the software. The model is also a good means of communication between system developers, users and business people.

The use of data models in visual analytics is not common practice. This work proposes using a domain data model in visual analytics. Domain means here a field or scope of knowledge or activity, such as monitoring of objects, marketing or bioinformatics, comprising a variety of applications. The domain model defines the concepts of a given data domain. It is based on the idea that the key concepts and the user tasks inside a data domain stay similar, regardless of the application. The model is constructed in such a form that data visualization and analysis can be easily applied. To keep the work within reasonable limits the study focuses on the monitoring data.

Three uses of the data model are suggested. The first is the construction of visual analytics applications. The data model can be used as an architectural model for a platform of visual analytics applications. The platform contains a database that is constructed based on the model, a library of predefined visualization and analysis methods that are compatible with the model, and a general-purpose domain-specific user interface. At the simplest level a new application can be constructed by merely loading the application data onto the database and making the predefined analysis and visualization methods available through the user interface. There is no need to build case-by-case solutions for each application in the domain.

The second use of the model is to support reasoning. Metadata can be added to the data model to guide the user in reasoning by suggesting visualization and analysis methods that are most applicable in different situations. This kind of support could be especially useful for users that are familiar with the application domain but not with data analysis.

The third use is to represent the data contents in the user interface in the form of a data model. An application data model, generated from the database, is shown to the user as an aid to navigate the data jungle. This offers an approach to the focus–context problem of information visualization. The application model can also be used as a building block of more elaborate user interfaces showing descriptive models of the objects of interest, such as maps or building models. Linking analysis results with descriptive visualization could help users in interpreting the results.

The research questions addressed in this work are:

- Are data models useful in visual analytics?
- What kind of model is useful?
- What kind of visual analytic tool can be constructed based on the suggested model?
- Does this approach help users find useful knowledge from the data?
- How straightforward is it to apply the approach to a new case in the domain?
- Can the results be generalized to other domains?
- What are the advantages and disadvantages of this approach compared to other kinds of visual analytics solutions?

These questions are examined by constructing a domain model for monitoring data, defining a visual analytics tool based on the model, and applying the suggested approach to two application cases: analysing the eating and exercise habits of dieting people (case HyperFit), and studying the energy efficiency and indoor conditions of buildings (case MMEA). The objective in the HyperFit case is to examine how users are able to reason and gain insight using this kind of tool. The MMEA case focuses on determining how well the concept can be applied to other kinds of data, and how to construct an application-specific user interface based on the model.

Contents of this work

The work starts with an introduction to visual analytics (Chapter 2), presenting the concept, the state of the art of research and the commercial markets for visual analytics. The building blocks of this multi-discipline research area are introduced with the help of a framework (introduced in Chapter 3) that was constructed in this work, in order to structure and study the building blocks of visual analytics applications. The evaluation methods of visual analytics are discussed in Chapter 4. Next, the two other key elements of this work are introduced: data modelling in Chapter 5 and monitoring data in Chapter 6. Chapter 7 presents the contributions of this work by introducing the suggested domain model and Chapter 8 the concept for a visual analytics tool based on the model. Chapter 9 presents experiments conducted in monitoring dieting people (case HyperFit) and Chapter 10 the experiments on energy and indoor conditions of buildings (case MMEA). Finally, Chapter 11 presents the summary and conclusions.

2 Visual analytics

Visual Analytics is a recent approach to finding knowledge from data masses, defined as “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook, 2005). It provides visual tools for finding insight from complex, conflicting and dynamic data to support analytical reasoning and decision making. The basic idea of visual analytics is to unite the strengths of automatic data analysis and the visual perception and analysis capabilities of the human user. Humans can easily recognize patterns, colour, shape, orientation and spatial position, detect changes and movement, and identify specific areas and items from visual presentations. Humans are good at reasoning and generating problem-solving heuristics (Card, Mackinlay and Shneiderman, 1999). Computers, on the other hand, have superior working memory and unlimited information processing capacity without cognitive biases (Green, Ribarsky and, Fisher 2008). The goal is to create systems that utilize human strengths while providing external aids to compensate for human weaknesses (Thomas and Cook, 2005). Visual analytics is especially focused on situations where huge amounts of data and the complexity of the problem make automatic reasoning impossible without human interaction.

Visual analytics is a multi-disciplinary field combining methodologies from several research areas. Figure 2 illustrates the scope of visual analytics (Keim et al, 2006). It merges different branches of analytics, data management, knowledge representation and human factors into an analytical reasoning process. Production, presentation and dissemination serve to summarize the results of analytical efforts and deliver them to the intended audience.

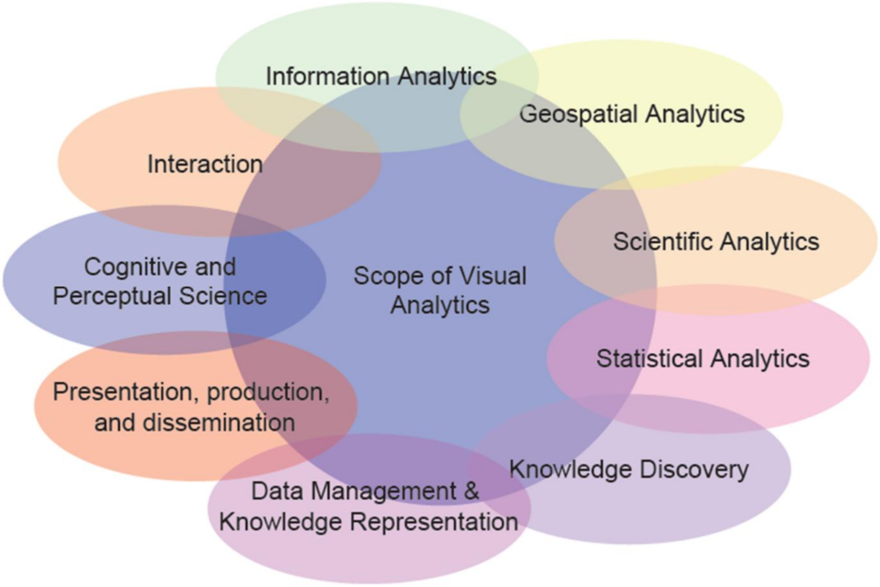

Information Analytics
Geospatial Analytics
Interaction
Scientific Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Statistical Analytics
Presentation, production, and dissemination
Knowledge Discovery
Data Management & Knowledge Representation

Figure 2. Scope of visual analytics (Keim et al, 2006)**.**

Visual analytics tools

Visual analytics produces tools to support analytical reasoning. A visual analytic tool processes the raw data and shows the information in the form of abstract and interconnected visualizations. Users can then look for patterns, trends, anomalies, similarities and other relevant "nuggets of information" from the visualizations. Users can launch analysis, browse and navigate visualizations and highlight and select important areas for further study. The analysis can be a collaborative process, shared among analysts working on related problems (Thomas and Cook, 2005).

Thomas and Cook list the tasks involved in visual analytics as including (1) information gathering, (2) re-representation of the information in a form that aids analysis, (3) development of insight through the manipulation of this presentation, and (4) creation of some knowledge product or direct action based on the knowledge insight, repeated in an overall cycle. They give an example of a car shopper who first gathers information from magazines, the Internet and dealers and then creates a table of attributes. The data is manipulated until the shopper gets insight and finally makes a purchase decision.

Figure 3[1] shows a vision of a visual analytics tool. It combines data coming from different sources and in different formats. The tool supports data exploration from different perspectives and levels. The user can look for trends and patterns, similarities and differences from a variety of visualizations. The visualizations are interlinked and equipped with interactive direct manipulation techniques.

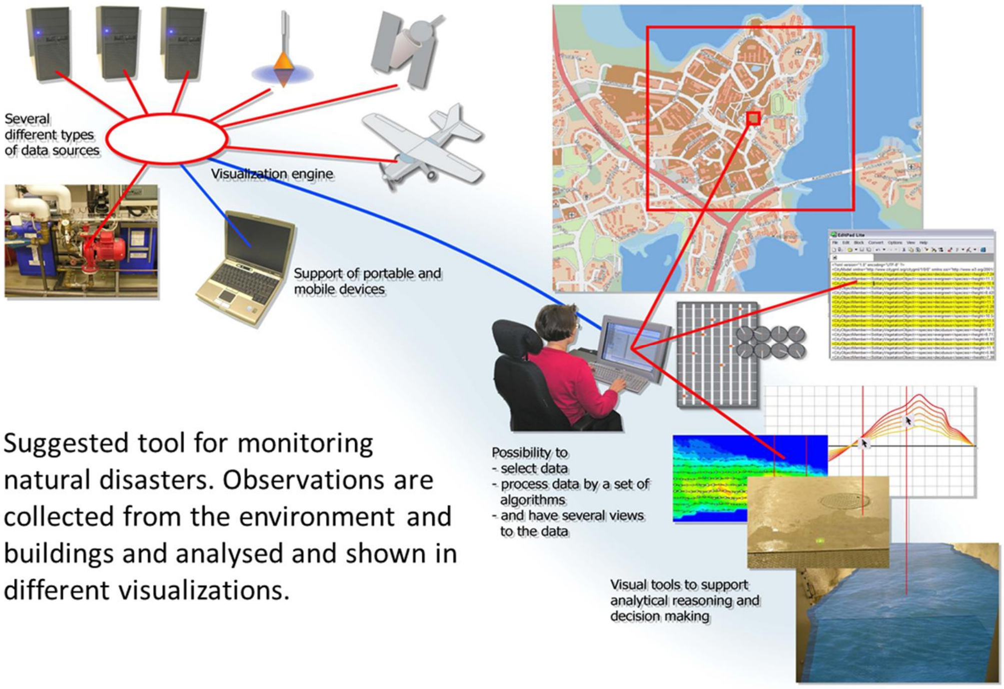

Several different types of data sources
Visualization engine
Support of portable and mobile devices
Suggested tool for monitoring natural disasters. Observations are collected from the environment and buildings and analysed and shown in different visualizations.
Possibility to
- select data
- process data by a set of algorithms
- and have several views to the data
Visual tools to support analytical reasoning and decision making

Figure 3. A vision of a visual analytics tool.

[1] VTT internal material, illustration by Hannu Kuukkanen

2.1 State of the art of research

Visual analytics has its origins in US national security. The US Department of Homeland Security (DHS) started a research initiative on visual analytics for homeland security. The "National Visualization and Analytics Centre" (NVAC), founded in 2004, coordinates these research efforts. The agenda for the US visual analytics research programme is laid out in the book "Illuminating the Path" (Thomas and Cook, 2005), which defines visual analytics and describes the research challenges. The agenda calls for new means to analyse the disparate, conflicting and dynamic information that is central to identifying and preventing threats and disasters. Essential in the analysis process is human judgement in making decisions from data in rapidly changing situations. The agenda names these fundamentally new solutions as visual analytics.

The agenda covers appropriate methods and technologies, needs for technology and research, and the steps for putting research into practice. The addressed methods and technologies include analytical reasoning, visual representations, interaction techniques, data representations and transformations, and communicating results. Scalability, privacy and security, collaboration and sharing information across organizations and agencies are highlighted. Although the report focuses on homeland security, such as analysing terrorist and border threats and emergencies, it suggests that the new capabilities have an impact on all areas where understanding complex and dynamic information is important, ranging from business to scientific research.

Idea and challenges

While the abovementioned agenda leaves the concept of visual analytics at a fairly general level, several publications have since appeared introducing the idea and challenges of visual analytics in more detail. The first is the article "A Visual Analytics Agenda", published in IEEE Computer Graphics and Applications (Thomas and Cook, 2006). The article summarises the agenda and underscores that visual analytics is not limited just to safety and security, but has broader applicability (Johnson et al, 2006). To achieve the goals of the agenda there is a need to put effort into studies that integrate visualization technology, data analysis, and perception and cognition sciences.

The report "NIH/NSF Visualization Research Challenges" by Johnson et al (2006) complements the agenda and introduces similar goals, but from the point of view of visualization. The report discusses the broad spectrum of application domains that can benefit from visualizations, including health, science and engineering, and outlines a data exploration process supported by visualization, adapted from van Wijk (2005). The report gives visualization research recommendations and evaluates the state of the field. In it, Johnson et al emphasize the integration of information visualization with other disciplines; human perception and cognition, data exploration and user interaction. They also call for new visualization techniques, novel interaction metaphors, and methods for measuring and evaluating the value of visualizations.

Daniel Keim et al have published a variety of articles introducing the concept. "Challenges in Visual analytics" (Keim et al, 2006) discusses visual analytics and its scope based on the visual analytics agenda. The article defines key visual analytics problems, including the need for integrated data sources and human expertise in finding solutions, and lists research challenges and solutions for the field. Scalability challenges arise from the large amounts of data involved. Tools for filtering, aggregation and data reduction as well as scalable visualizations are required. Data streams, heterogeneous data fusion, preprocessing, and data quality also need to be taken into account. The key challenges identified include supporting reasoning by providing semantics for analysis and decision-making, integrating technologies with existing systems, and evaluation. The article introduces the visual analytics mantra: "Analyze first Show the important, Zoom, Filter, and analyze further, Detail on demand". It is a modification of information seeking mantra by Shneiderman (1996): "Overview first, zoom/filter, Details on demand".

In "Visual Analytics: definition, process, challenges" Keim et al (2008) present similar definitions and technological challenges to the previous article. A quick tour of the state of the art of components is given, covering visualization, data management, data analysis, perception, cognition, human computer interaction, infrastructure and evaluation. The visual analytics process is introduced, modified from the suggestions by van Wijk (2005) and Johnson et al (2006). The article "Visual analytics: Scope and challenges" (Keim et al, 2008b) is similar in content to the previous articles, although application domains are also introduced. These include physics and astronomy, business, environmental monitoring, disaster and emergency management, security, software analytics, biology, medicine and health, engineering, and personal information management. Similarly, "Visual analytics, combining automated discovery with interactive visualizations" (Keim et al, 2008a) repeats the earlier content and introduces gives a text analysis application of visual analytics.

Icke presents the developments in visual analytics with the help of a framework called the "multifaceted overview" (Icke 2009), which divides the elements of visual analytics into system, user and human-machine collaboration aspects. The system aspects include data analytical tasks and visualization types; the user aspects include expertise and collaboration; and the human-machine collaboration aspects include interactivity and utility. The article focuses mainly on the system aspects, such as data types and visualization types, and covers the other aspects lightly. Icke does not consider human reasoning, insight, perception and cognition at all.

The book "Mastering the Information Age" (Keim et al, 2010) is the outcome of a two-year project called VisMaster CA. It was a coordination action funded by the European Commission during 2008–2010. The book introduces the topic of visual analytics, presents a brief history, and introduces application areas, the visual analytics process, the building blocks of visual analytics, and scientific challenges and recommendations. The book covers the components of visual analytics, including data management, data mining, space and time, infrastructure, perception and cognitive aspects, and evaluation of visual analytics. For each of these components, the state of the art, specific challenges and opportunities with respect to visual analytics research are discussed.

A technical report by VTT[2] (Järvinen et al, 2009) studied visual analytics in order to clarify the concept and to examine its industrial implications. A demonstration tool was developed and a roadmap for industrial and consumer applications.

[2] Technical Research Centre of Finland

Journals

Special journal issues on visual analytics have been published since 2004. The earliest is in IEEE Computer Graphics and Applications September/October 2004. The introductory article by Wong and Thomas (2004) defines visual analytics as "an outgrowth of scientific and information visualization that combines the art of human intuition and the science of mathematical deduction to directly perceive patterns and derive knowledge and insight from them". The term visual analytics was probably used for the first time here. Wong and Thomas consider the formation of abstract visual metaphors in combination with a human-information discourse to be essential to the visual analytics concept. The objective is to enable detection of the expected and discovery of the unexpected within massive dynamically changing information spaces. According to Wong and Thomas, visual analytics integrates knowledge management, statistical analysis, cognitive science, and decision science. They suggest that visual analytics suits almost all fields, being driven by critical needs in biology and security. The journal issue includes six articles as representatives of visual analytics, covering document visualizations, flaws and intruders in computer networks systems, analysis of geospatial data point sets, finding trading patterns in stock market data, studying ocean bottoms, and analysis of social behaviour in web environments.

IEEE Computer graphics and Applications published another special issue on visual analytics in 2007. The introductory article "Discovering the unexpected" (Cook, Earnshaw and Stasko, 2007) outlines a visualization timeline. It shows the key developments in visualization leading to visual analytics. The timeline starts from 1952 when A.S. Douglas received a PhD in human computer interaction using a CRT display and Edsac 1 computer[3], and ends with the first visual analytics IEEE symposium in 2006. The article emphasises the role of human interaction, using models to cope with the large scale and diversity of data, minimizing cognitive load with the right kinds of visualizations, and a seamless sense-making process. The other articles in the issue cover computer network defence, earthquake simulations, tools for understanding set members, and tools and methods to assist sense-making.

Since the publication of the Visual Analytics Agenda in 2006, visual analytics articles have appeared in a variety of journals. Among them the most frequent publishers are IEEE Transactions on Visualization and Computer Graphics, and Information Visualization by SAGE Journals. In 2012 visual analytics received its own journal, the International Journal of Visual Analytics for Advanced Information Management (IJVAAIM) by the Information Resources Management Association[4]. The journal's stated mission is to "collect the latest research essential to the emerging field of scientific and information visualization analytics".

Conferences

Visual analytics has also gained a strong presence in international conferences. The first symposium on Visual Analytics Science and Technology (VAST) was held by IEEE in 2006. Chung Wong et al (2007) introduce the highlights of this conference in the first visual analytics themed issue of the Information Visualization Journal. The issue covers topics such as wormhole detection in wireless networks, exploration of

[3] http://www.pongstory.com/1952.htm (accessed 1 October 2012)
[4] http://www.irma-international.org/ (accessed 1 October 2012)

network data, productive reading techniques, collaborative decision making, linguistic algorithms, exploration of patters of hotel visits, and pixel-based visualization optimization. Visual analytics contests arranged in conjunction with the VAST conferences since 2008 have had an important impact on research, bringing novel ideas and solutions to the field. In addition to VAST, the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery (VAKD) has also since been held.

2.2 Commercial markets

To examine the current state of the visual analytics market, a web survey was conducted. The survey searched for products using the criteria "visual analytics", "visual data mining" or "data mining with visualizations". Tools producing only static visual reports, basic statistical tools such as Excel, and analysis packages such as R and Matlab, were omitted from the survey. A list of studied products is presented in Appendix A.

The search resulted in 55 software products which were grouped into six categories, as shown in Figure 4: (1) visual analytics tools, (2) tools by major software houses, (3) analysis and visualization environments, (4) libraries, (5) application-oriented tools, and (6) technology-oriented tools.

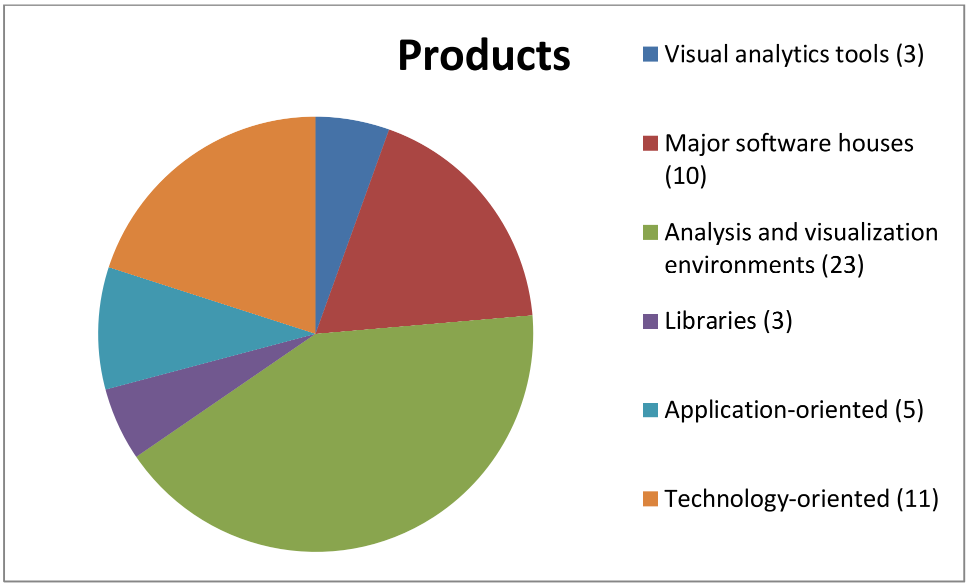


Figure 4. Visual analytics product categories.

Three tools defined themselves as visual analytics tools: Data clarity suite[5] by Visual Analytics Inc., Nuix Visual Analytics[6] by Nuix, and PV-WAVE[7] by Rogue Wave. Of the major software houses, SAP, IBM, ORACLE, SAS, Microsoft and Information Builders had added visual analytics to their offering, usually by buying an independent visual analytics product and integrating it as part of their services. The biggest group (22 products) were general-purpose data analysis and visualization environments that contained features characteristic to visual analytics, such as data

[5] http://www.visualanalytics.com (accessed 1 October 2012)
[6] http://www.nuix.com/visual-analytics (accessed 1 October 2012)
[7] http://www.roguewave.com/products/pv-wave-family.aspx (accessed 1 October 2012)

analysis combined with interactive visualizations. Three libraries provided open source software for data mining and visualization tasks. The application-oriented tools (5 products) focused on specific application areas, including learning, software analysis, business, and biosciences. The technology-oriented tools (11 products) offered data analysis focused on a specific data type (text, geospatial data, network data), analysis method (Bayesian modelling, neural networks) or visualization technique (3D, 4D).

The majority of the tools were provided by business consultants. Three were open source products. Most of the tools were "general-purpose" suggesting application areas covering almost every aspect of human activity: business, marketing, banking and finance, safety and security, healthcare, medical diagnostics, science, industry, government, politics, media, communications, transportation, logistics, consumer applications and education.

Another source for the market survey consisted of reports from market analysis companies. Three companies were included: Frost & Sullivan[8], ReportLinker[9] and Gartner Group[10]. All recognized the new field and made reference to recommended application areas. Frost & Sullivan mention visual analytics and business intelligence with reference to "what's hot", highlighting the use of visual analytics in network security and medical analysis. ReportLinker recognized genomics, market research, education and training, medicine, the energy sector and mobile phones as potential application domains. Gartner Group refers to visual analytics as a "hot topic" in business intelligence and market research, security, social analytics and the big data market. It recommends extending the use of visual analytics and mentions products from SAS and Tibco Software. Gartner positioned data visualization on a strongly upward curve.

[8] http://www.frost.com (3 October 2012)
[9] http://academic.reportlinker.com (3 October 2012)
[10] http://www.gartner.com (3 October 2012)

3 Framework for visual analytics

Visual analytics is a mixture of technologies, theories and methodologies. A framework for visual analytics was constructed as part of this work in order to structure the elements of the multi-discipline research area, to review the state of the art of them, and to identify the building blocks of visual analytics tools. Another motivation for this quite an extensive framework was that there is no text book or other comprehensive material introducing this new research area. The present literature consists mainly of visions and challenges or detailed application descriptions. The only wider introductions are in the Visual Analytics Agenda (Thomas and Cook, 2005) and in the report of the VisMaster CA project (Keim et al, 2010), which are the main sources of this framework.

The framework divides the visual analytics elements to *cornerstones* and *building blocks*. The cornerstones are *problem*, *data* and *insight*. The building blocks represent the knowledge, technologies and methods used in the construction of visual analytics applications. They include *human capabilities, communication, enabling technologies,* and *infrastructure*. Figure 5 shows the framework.

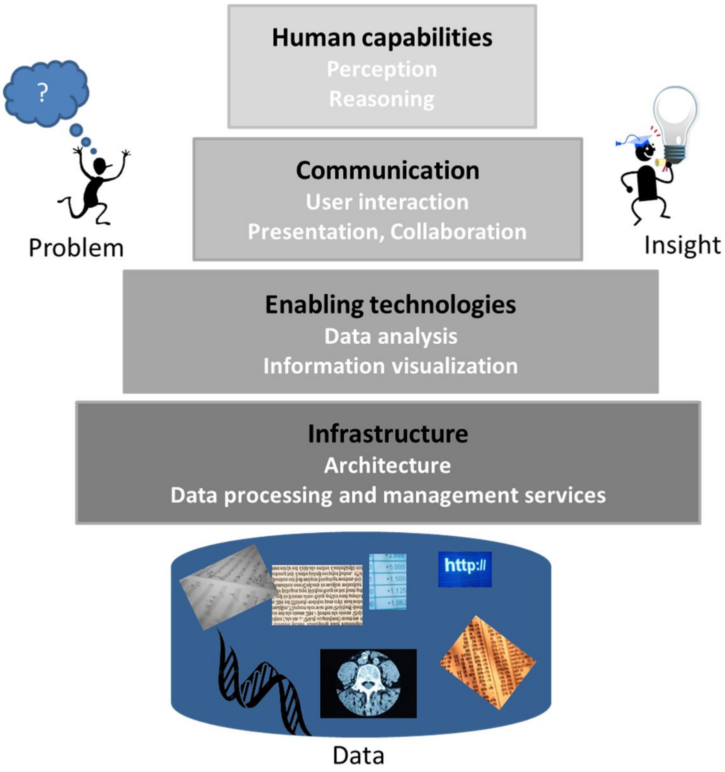


Figure 5. Visual analytics framework**.**

The cornerstones *problem, data* and *insight* are stated clearly in the definition of visual analytics by Keim et al (2006): “The ultimate goal is to gain insight into the problem at hand which is described by vast amounts of scientific, forensic or business data from heterogeneous sources”. The need for visual analytics arises from a problem or issue that needs to be solved with the help of data, which is expected to contain hidden knowledge that that can be used to support reasoning and decision making. The ultimate goal is to gain insight.

Visual analytics solutions rely on human capabilities. The most essential of these is the human perception and reasoning process. Communication refers to methods for communication with the outside world, users and the audience. It includes user interaction, annotation, collaborative problem solving, and production of analytic outputs and presentations. Enabling technologies provide methods for insight-searching, including data analysis and information visualization. All of these components are tied together by the infrastructure, which comprises an architectural model and services for various data processing and management tasks.

Each of the elements has its specific, more or less mature theories and methods. When these are tied together in a visual analytics tool new challenges emerge, both within each element and in the co-operation between elements.

This chapter introduces the framework elements and cornerstones and discusses the special questions arising in relation to them. Human capabilities, including perception and reasoning, are discussed, and user interaction, data analysis and information visualization are reviewed. The final section considers the role of infrastructure. Presentation and collaboration issue are omitted from the introduction having less importance in this study.

3.1 Visual analytics problems

Problems arise where there is a lack of relevant knowledge needed to produce an immediate solution (Eysenck and Keane, 2005). Problems can be divided into well-defined and ill-defined problems. Well-defined problems are clearly specified in all aspects. They include the initial state or situation, the range of possible moves or strategies, and the goal, such that it is clear when the goal has been reached. A maze is an example of a well-defined problem. Ill-defined problems are unspecified, and these account for many of the everyday problems commonly encountered. Problems can also be knowledge-rich or knowledge-lean. Knowledge-rich problems can only be solved by individuals possessing a considerable amount of specific knowledge. Knowledge-lean problems require no special knowledge. A significant problem to one group of people may be of no importance to another, and where a person with less expertise may see a problem, someone with the relevant expertise may see none (Eysenck and Keane, 2005).

The Visual Analytics Agenda (Thomas an Cook, 2006) classifies visual analytics problems as (1) making judgement based on evidence about an issue, (2) discovering new understanding, and (3) supporting decision making. These all fall into the category of ill-defined knowledge-rich problems.

Keim et al (2006) define the characteristics of visual analytics problems: there needs to be a large information space, possibly integrated from disparate data sources, and human expertise in finding solutions. A well-defined problem where an optimal estimation can be calculated automatically should not be classed as a visual analytics problem. In addition, by Keim et al visualization problems not involving methods of automatic data analysis do not fall into the problem category of visual analytics problems.

Visual analytics problems are typically highly application domain and require expertise. For example, bioinformatics has different problems to marketing.

3.2 Data

The data used in visual analytics comes in a huge variety of forms. It can be different in type: numeric or textual, image, video or audio; or in structure: unstructured, structured or semi-structured, in addition to which structured data can also have a hierarchical or network structure. Data can also have different dimensions; it can be 1-dimensional, 2-dimensional or multidimensional. It can be temporal, spatial or spatio-temporal; it can also be static or streaming. It can originate from various sources: collections of documents, spreadsheets, e-mails, Web pages, structured databases or measurements by devices. There can be huge amounts of data. A dataset is considered large when it contains more than 100,000 data items, and "massive" when there are more than hundreds of millions items (Ferreira de Oliveira and Levkowitz, 2003). Big data is defined by the McKinsey Global Institute[11] as a "Dataset whose size is beyond the ability of typical database software tools to capture, store, manage and analyse."

In addition to the actual data, there can be metadata – information about the data, describing its properties or semantics. For example, relational databases and ontology models, in particular, store data about relationships between data elements.

Data types

The Visual Analytics Agenda (Thomas and Cook, 2005) characterizes data according to characteristics that have an impact on the data representation. These are: (1) data type, (2) level of structure, (3) geospatial characteristics and (4) temporal characteristics. Data can be numeric, non-numeric or both. Sensor data is often numeric and language data non-numeric. The level of structure varies from completely structured, such as categorical data, to completely unstructured, such as a narrative description on a web page. Unstructured does not always mean that there is no structure, but rather that the structure is only humanly interpretable. Geospatial characteristics are associated with a particular location or region. Any data can have a geospatial association. Furthermore, all data types can have a temporal association, which may be either discrete or continuous.

Each data type has its own specific methods of analysis and visualization. Shneiderman (1996) introduces a data type taxonomy for information visualization consisting of seven data types: 1-dimensional, 2-dimensional, 3-dimensional, temporal, multidimensional, tree, and network. Examples of 1-dimensional data are written documents or lists. Maps, floor plans or newspaper layouts represent 2-dimensional data. 3-dimendional data represents real-world objects, such as molecules, the human body or buildings. Temporal data represents timelines. Multi-dimensional data consists of data items with several data attributes. Trees are collections of data structured as hierarchies, and networks represent more complex relationships among data items. All of these data types can occur in combination.

In statistics, data is described as nominal, ordinal, interval-scaled, ratio-scaled, discrete or continuous. In nominal data, values are assigned to code labels. Data items

[11] http://www.mckinsey.com/insights/mgi (accessed 1 October 2012)

can be counted but not ordered. ID numbers, eye colour, and zip codes are examples of nominal data. If data is ordinal, values can be ranked and put in order, but not measured. Ordinal data examples include rankings (e.g., taste of chocolate on a scale from 1-5), grades, height categories: tall, medium or short. Interval-scaled data has a scale, and the distance between any two values can be measured, but the zero point is arbitrary. Scores on an interval scale can be added and subtracted but not be meaningfully multiplied or divided. Interval data examples are calendar dates or temperatures in Celsius or Fahrenheit. Ratio-scaled data has a definite zero point. Ratio examples are temperature in Kelvin, length, time, and frequency counts. A set of data is said to be continuous if the values may take on any value within a finite or infinite interval. Continuous data can be counted, ordered and measured. Examples are height, weight and temperature. A set of data is said to be discrete if the values are distinct and separate, i.e. they can be counted. (andrews.edu, 2005)

This works mainly concerns multivariate data consisting of attributes categorized as nominal, ordinal, or quantitative. Quantitative data can be both interval- and ratio-scaled data. Attributes can be single-valued or time series data that has a time stamp attached to each value.

3.3 Insight

There is almost a community-wide consensus that insight is the ultimate goal of visual analytics (Chen 2010). But what does it mean? In cognition science, insight refers to “an experience of suddenly realizing how to solve a problem”, known commonly as an “aha” or “eureka” moment (Eysenck and Keane, 2005). Insight occurs suddenly, it “pops out”. It does not necessarily require gradual accumulation of information, but past experience is beneficial.

In visual analytics there is no precise or agreed definition of insight, although a number of suggestions have been proposed. A precise definition would be especially useful in insight-based evaluations in order to know how to measure insights. Chen (2010) characterizes insights as “unexpected discoveries”, “a deepened understanding”, or “a new way of thinking”. Chang et al (2009) define two distinct types of insight. One is spontaneous insight, a “moment of enlightenment”; the other is knowledge-building insight, “an advance in knowledge or a piece of information“ a form of learning. Smuc et al (2009) define insight for evaluation purposes as “understanding gained by an individual using a visualization tool (or parts thereof) for the purpose of data analysis, which is a gradual process towards discovering new knowledge.”

North et al (2006) take a different point of view. In “Towards measuring visualization insight” they consider that any formal definition would be either too restrictive to capture the essence of insight or too general and vague to be useful. Instead, characteristics of insight are listed. Insight is “complex, involving large amounts of information, deep – building over time, qualitative – uncertain and subjective, often unexpected, and relevant to the data domain”. They also introduce different kinds of insight: when a pattern is selected in visualization, when a cognitive script is identified for data analysis, or when a mental model of data is completed.

In this work insights are considered vaguely as “aha experiences” that users have when making expected and unexpected findings.

3.4 Human cognition and perception

Visual analytics combines analytical reasoning with interactive visualization, both of which are subject to the strengths and limitations of human perceptual and cognitive abilities. Building effective visual tools requires deep understanding of the capabilities and limits of the human information and visual system. Visualizations designed without knowledge about human perceptual and cognitive principles can lead to poor or misleading ad hoc solutions (Johnson et al, 2006). If information is presented in an inappropriate way it can, in the worst case lead, to incorrect decisions. A classic example is the O-ring damage of the space shuttle Challenger (Tufte, 1983). A scatterplot made by E. Tufte would have shown the fatal damage that the original visualization did not reveal. In contrast, good visualizations can improve the efficiency, effectiveness and capabilities of decision makers and analysts (Ware, 2004). A good visualization can tell a story, help make discoveries, and show the "big picture" or different views of phenomena (Chen, 2010).

Cognition science studies human cognition, the purpose of which is to make sense of the environment and decide what action(s) might be appropriate (Eysenck and Keane, 2005). It includes aspects such as attention, perception, learning, memory, problem solving, reasoning and thinking. Human perception research studies the understanding of sensory information. Perception, as defined by Secular and Bake (in Eysenck and Keane, 2005) means "The acquisition and processing of sensory information in order to see, hear, taste, or feel objects in the works; also guides an organism's action with respect to those objects".

This section introduces the areas of human cognition and perception that are the most important from the point of view of visual analytics, the visual perception and memory systems (Thomas and Cook, 2005). Knowing the principles of human visual perception helps to favour the visual patterns, symbols, colours and other effects that are interpreted easily and avoid such visual elements that are misleading or difficult to perceive. The limitations and strengths of the human memory system are important to know in creating fluent reasoning processes that do not overload the human memory system and distract attention. The topics covered here are: the human visual process; perception of colour, depth and size; recognition of objects; motion and action detection and perception; attention and searching; and the architecture of memory and memory processes. The last topic is about data graphics that introduces design principles for visual interfaces in line with the findings of cognition research. Human reasoning is introduced in Section 3.5.

3.4.1 Visual perception

What happens when a visual stimulus reaches the eye? The perception process proceeds in stages, each at different rates. The process begins with rapid parallel processing, including the extraction of features, orientation, colour, texture and movement (reception), continues with pattern perception (transduction), and ends with slow sequential goal-driven processing. During this process some visual symbols are understood quickly. These are sensory symbols. Others, arbitrary symbols, are interpreted slowly. These are learned and easily forgotten. Letters and numbers are examples of arbitrary symbols, whereas a line connecting two areas is a sensory symbol. (Ware, 2004)

Acuity is the ability to see detail. Examples of acuities are the ability to differentiate distinct points or letters or to distinguish a pattern of bright and dark bars. Humans have certain superacuities where the human eye is especially adept. These include perception of depth, co-linearity of line segments, and stereoscopic depth. However, up to 20% of the population are stereoblind, and often unaware of it.

Contrast sensitivity is the visual ability to see objects that may not be outlined clearly or that do not stand out from their background. The ability varies between individuals and decreases with age. Striped patterns covering large areas flicker and cause visual stress or, in the worst case, can induce epileptic seizures.

The human eye is able to work in varying light levels. Physical luminance and perceived brightness do not necessarily correspond to each other. Deep contrast differences can cause strange effects: ghost effects near dark and bright bands (Hermann grid illusion). Another effect is that the same grey patches look different on different backgrounds.

Colour perception (Ware, 2004)

Colour vision is helpful in distinguishing objects. Humans recognize three qualities of colour: hue, brightness and saturation. Interpretation of colours is affected by lighting conditions and the colour environment (colour constancy). Colour blind individuals cannot extinguish all colours. The most common forms of colour blindness are red–green deficiency and blue–yellow deficiency. 12% of men and 1% of women are colour blind. The ability to distinguish colours decreases with very small objects (small-field colour blindness). Colour contrast effects can also distort the reading of colour-coded values.

The human eye does not fully correct against chromatic aberration, which leads to certain special colour perception effects. Blue text on a black background is unreadable when in close proximity to white or red. People also see red and blue on a black background differently, and red seems to “pop up” on a blue background.

Human perception of shape or motion with colours is limited, and for that reason colours are recommended only for labelling. In the selection of colours for labelling distinctness, hue, contrast with background, size, number (5-10 colours are rapidly distinguished) and convention (red associated with hot, blue with cold, etc.) should be considered. The use of basic colours: red, green, yellow, blue, black and white is recommended. If more colours are needed, these should be pink, purple, orange or aqua. Pure colours are easily remembered and can be named by 75% of individuals. Connecting meanings to colours depends on familiarity, expectations and cultural background.

With respect to colour scales there is no general agreement. The “rainbow” scale is the default choice, although this is not recommended where shape is important. With greyscales it is difficult to distinguish shades, and only 2 shades and black and white should therefore be used. A blue-yellow scale is recommended for the red-green colour blind. Some guidelines on the use of colour scales with different data types have been proposed: for nominal data distinctive colours should be used; for ordinal data a fixed colour scale can be used; and with interval data colour changes should reflect differences in the data. Ratio is impossible to represent using colour scales.

Depth and size (Ware, 2004*;* Eysenck and Keane, 2005)

The human visual system transforms the two-dimensional retinal image into a three-dimensional world. The physical world has many sources of depth information. The transformation occurs with the help of various distance and depth cues. These principles can be used to create the impression of distance, depth effects, or three-dimensional scenes.

The perspective cue, occurring when lines converge at a single point, gives a powerful impression of depth. Image blur is another depth cue. When an object boundary is sharp, the object seems to be nearer. Distant objects lose contrast and appear hazy. Changes in object texture, shading and contour are also cues for distance. An object that occludes another appears to be closer to the viewer. Objects placed higher are seen more distant. Object familiarity – for example objects that are known to be very big or small – affects the interpretation of size.

Shading is a cue for 3-dimensional depth. As two-dimensional surfaces do not cast shadows, shading can be used to give a three-dimensional impression. Cast shadows determine the height of objects. Shadow effects are especially powerful with objects that are in motion. The movement of objects within a 3D space gives depth cues. Near objects pass by rapidly, distant ones slowly.

When several cues are used the effects can be additive, selective or multiplicative. Additivity means that all information from the different cues is added together. In selectivity, information from a single cue is used and the others ignored. Multiplication interacts information in a multiplicative fashion. There is not yet clear understanding of how the cues interact. Experiments have shown that depth perception degrades when cues conflict. There is evidence that occlusion is the strongest depth cue.

Object recognition (Eysenck and Keane, 2005)

Object recognition is a complex process. Humans are surrounded by a numerous array of objects and must identify where one ends and the next starts. In addition to dealing with overlapping objects, objects have to be recognized accurately over a wide range of viewing distances and orientations and ultimately categorized.

There are numerous theories regarding the *perception of visual objects*. The most influential are Marr’s sketch-based theory and recognition by composition by Irwin Biedermann (Eysenck and Keane, 2005). Marr’s theory is founded on efficient image recognition based on three main kinds of representation sketches. The first phase identifies the outline, the second the depth and orientation, and the final creates a 3D model representation. Composition theory suggests that images are recognised as structures built from three-dimensional primitives called geons.

Face recognition is thought to differ from other forms of object recognition. It is important for humans to distinguish between familiar and unfamiliar faces, name persons and remember their doings. A specifically associated brain area dedicated to face processing has been recognized. Numerous kinds of information can be extracted from the face. The ability to distinguish between familiar and unfamiliar faces is learned by humans earlier than the ability to recognise many other objects. Humans are also experts at recognizing facial expressions. Cross-cultural studies have identified six universal expressions: anger, disgust, fear, happiness, sadness and surprise (Ware, 2004). These are expressed in popular "smileys".

Some visual objects are processed pre-attentively, before conscious attention (Ware, 2004). Such objects seem to "pop out" at the observer. Bolded text is an example of a pre-attentive feature that pops out. Other pre-attentive features include line orientation, length, width, collinearity, size, curvature, spatial grouping, added marks, numerosity (up to four), colour (hue, intensity), blur, motion (flicker, direction), and spatial position (2D position, stereo depth, convex/concave shape of shading). Combining pre-attentive features does not necessarily lead to pre-attentive processing; for example, combining width and height or conjunction searches such as "find red and circular" are not pre-attentive. Grouping size and colour or spatial features, motion and stereo can, however, lead to pre-attentive features; for example, "find red moving things" is pre-attentive.

Perception of patterns (Ware, 2004)

Pattern perception is summed up in Gestalt laws[12], which can be translated directly into design principles. These laws include proximity, similarity, connectedness, continuity, symmetry, closure and relative size. Proximity, or nearness, means that things that are near to each other appear to be grouped together. Similarity groups similar things together, for instance similar colours and shapes. Connectedness means connecting graphical objects with lines. Continuity means that humans are more likely to construct visual entities out of visual elements that are smooth and continuous rather than ones that contain abrupt changes in direction. In symmetry, symmetrically arranged pairs are strongly perceived to be together. Closure refers to the tendency to close contours that have gaps in them. Relative size refers to the tendency for smaller components of a pattern to be perceived as objects. Visual grammars such as Mind Maps or UML diagrams are applications of Gestalt laws. UML, for instance, utilizes similarity and continuation.

Humans are not consciously aware of all objects that they see. There is evidence of perception without awareness, including detection and localization of light, orientation, shape, direction of movement and flicker (Eysenck and Keane, 2005).

Motion and action (Eysenck and Keane, 2005)

Humans have an acute ability to perceive movement, and there is evidence that our brains have a special area dedicated to motion detection. Biological motion, that of humans and animals, is perceived especially accurately and effectively. Our ability to detect motion is based on structural and dynamic cues. Causality is a form of motion detection involving the ability to perceive motion that causes motion in other objects. Motion, animations and causality can be used effectively to reveal patterns in data. Moving objects can be used to reveal structures in data.

Change blindness is a phenomenon in which observers do not notice an unexpected object appearing in a visual display, or fail to detect that an object has moved or disappeared. The main factor of change blindness is whether the changed object was

[12] Gestalt School of Psychology (Germany, 1912 onward)

attended before the change. Other factors include the observers' intentions, sensitivity, similarity of unexpected objects and attended objects, and the extent of change.

Attention (Eysenck and Keane, 2005)

Visual attention refers to the process of seeking out visual stimuli and focusing on them. Attention is selective and its capacity is limited. There are two modes of attention, active and passive. Active attention has goals and expectations and is a top-down process, whereas passive attention is based on bottom-up processes. The attention capacity varies depending on the perceptual load, and is determined by the number of units in the field of vision and the nature of processing required for each unit. The attention capacity is greater when the task difficulty is high and increases in conditions of high effort or motivation.

Visual attention resembles a spotlight. Individuals focus their attention initially on a general area and then on a specific object or objects. Everything in this fairly small space is seen clearly, but not the area surrounding it.

When an individual is doing several things at a time, information needs to be coordinated from two or more senses. This requires interpretation of visual, auditory and tactile modalities. Tasks involving the same sense modality seem to interfere with each other more than those involving different sense modalities. Practice improves dual-task performance. Performance also depends on the task difficulty.

3.4.2 Memory system

Architecture of human memory (Eysenck and Keane, 2005)

Human memory is used for numerous purposes in everyday life. We use memory when talking, reading, writing, and in making sense of the world around us. Memory use has personal qualities and is influenced by situational demands. Human memory research has produced many theories, but there is no consensus concerning the nature of the memory system. The most influential approaches propose the division of the human memory into sensory, working and long-term memory. Sensory memory is a brief memory store. It is a modality-specific store, comprising an iconic image store for vision and an echoic store for hearing, in which information is held very briefly to be processed further.

Working memory is for coping in everyday life. It is a short-term storage of very limited capacity and can contain only about seven digits at a time. It is in a constant state of activation and can be accessed immediately and effortlessly. It is also fragile, and any distraction easily causes forgetting. Working memory is suggested to have four components: (1) a central unit that controls what visual information is held and stored, (2) a phonological loop, which holds information in speech-based form, (3) an episodic buffer, which holds and integrates diverse information, and (4) a visio-spatial sketchpad, also called the visual working memory, which is specialized for spatial and visual coding. The latter's capacity is limited to a small number (3–5) of simple visual objects and patterns. The visual working memory stores information on position, shape, colour and texture. An example of the implications the limitations of the visual working memory is data glyph design. A glyph is a visual object that represents one or more data variables. A data glyph should be held in visual memory and encoded according to the visual memory capacity.

Long-term memory is a life-long information store and has unlimited capacity. It is argued that there might be several long-term memories. Four types are suggested: (1) episodic memory involves recalling personal events from the past, (2) semantic memory involves knowledge of the world, (3) procedural memory is used in skill learning, and (4) the perceptual representation system is used for improving performance.

Memory processes

How do things end up in the long-term memory? One theory is the levels-of-processing framework proposed by Craik and Lockhart (in Eysenck and Keane, 2005), which assumes that attentional and perceptual processes of learning determine what information is stored in long-term memory. There are various levels of processing, ranging from shallow or physical analysis of a stimulus (e.g. detecting specific letters in words) to deep or semantic analysis that extracts the meaningfulness from the stimulus. The level of depth of processing of a stimulus has a large effect on its memorability. Deeper levels of analysis produce more elaborate, longer, and stronger memory traces than shallow levels of analysis. As regards memory rehearsal, a distinction is drawn between maintenance and elaborative rehearsal. Maintenance involves repeating previous analyses, whereas elaborative rehearsal involves deeper or more semantic analysis of the learning material. There is no agreement as to which method better enhances long-term memory. The existing theories explain incidental and intentional learning. Some people have exceptional memory skills that may be due to either deliberate strategies or natural outstanding properties.

3.4.3 Data graphics

Constructing good visualizations is challenging. They should invite users to explore the data, they should be based on cognitive, perceptual and graphic design principles, and they should improve the efficiency, effectiveness and capabilities of analysts. Data graphics is the theory underpinning effective visualization. Many of the principles have been developed by Edward Tufte (1983), who has offered valuable guidance on presenting static information. The principles were originally intended for graphics design, but are also valid for computer-based visualizations. They are not an exact theory, but more a collection of rules of thumb. They provide guidelines for building visualizations that are intuitive and convey the intended meaning clearly to the observer.

The main guideline is that representations should match the task performed by the user. Attention should be on the data; the main purpose of visualization is to show the data. Superfluous information can be distracting and make the task more difficult. “Chart junk”– the decoration of graphics that does not tell the viewer anything new should be avoided. The purpose of chart junk may be to make the graphics appear more scientific or lively (or to give the designer an opportunity to exercise their artistic skills). Gridlines, decorations, vibrations and redundancy are all chart junk. In the worst cases, the design overwhelms the data. In addition, the proportions of the visual representation should match the information being represented. If the graphics are not proportional to the numerical quantities, the visualization will be misleading. Multifunction elements can effectively display complex multivariate data, but they must be used with caution. A map that uses shading and colour to represent coordinate data and other properties is an example of multifunction graphics. Complex multifunction elements can easily turn data graphics into graphical puzzles.

The Visual Analytics Agenda (Thomas and Cook, 2005) introduces three principles of data graphics, adapted from Norman and Dunaeff (1994): appropriateness, naturalness and matching. The appropriateness principle states that visual representations should provide neither more nor less information than that needed for the task at hand. Naturalness calls for visual representations that most closely match the information being presented; new visual metaphors are only useful for representing information when they match the user's cognitive model of the information. The matching principle states that representations are most effective when they match the task to be performed by the user.

3.5 Analytical reasoning

Fluent analytical reasoning is essential to visual analytics. Psychology and cognition science have a long history of research into human reasoning, yet no unifying theory has emerged (Green, Ribarsky and Fisher, 2008). Instead, there exists a number of competing theories, heuristics and principles for sub-processes. The visual analytics community has defined processes, frameworks, precepts and elements that support reasoning, but it, too, lacks universal consensus, although a couple of dominating approaches do exist. A need for an agreed definition and a cognitive model of the visual analytics process has been recognized as essential to be able to construct fluent reasoning environments (Liu and Stasko, 2010).

Cognition science has three major theories of sense-making (Eysenck and Keane, 2005). First, there are *abstract rule theories*, according to which people use mental logic when confronted by a reasoning task. The second theory is the *mental model approach*, where people form mental models or representations and use these mental models – rather than rules – to draw conclusions. The third is the *probabilistic approach*, which is based on the notion that humans use cognitive processes developed to cope with the uncertainties of everyday life and these habitual ways of thinking continue in deductive reasoning tasks. In visual analytics the mental model approach is the most influential, and many visual analytics reasoning models and tools are founded on it. These include the works by Green et al (2009), Liu and Stasko (2010), and Wright et al (2006).

This section introduces approaches to sense-making in visual analytics, discusses the human mental model, and precepts for reasoning.

3.5.1 Sense-making in visual analytics

Many visual analytics tools have adopted the sense-making model by Pirolli and Card (2005), shown in Figure 6. The model is based on the mental model approach, having an emphasis on the role of schematic knowledge structures – a set of schemas used to organize information. It consists of four phases: (1) information gathering, (2) representation of the information in a schema that aids analysis, (3) the development of insight through the manipulation of schema representation, and (4) the creation of some knowledge product of direct action based on the insight. The representations may be informally in the analyst's mind, or aided by paper and pencil or a computer-based system. The process includes several back loops. The *Foraging loop* is the cycle of activities around finding information. It starts from external data sources and proceeds with producing a subset or "shoebox" of information. Evidence files are snippets of data from the shoebox. The loop involves processes that aim at seeking information, searching and filtering it, and reading and extracting information into schema. The *Sense-making loop* involves making sense of information has plenty of interaction between other loops. The process ends with a presentation of the work product. The model has been criticised for leaving the concept too open and for revealing little about what goes on within each step (Pottenger, Fisher and Ribarsky, 2009).

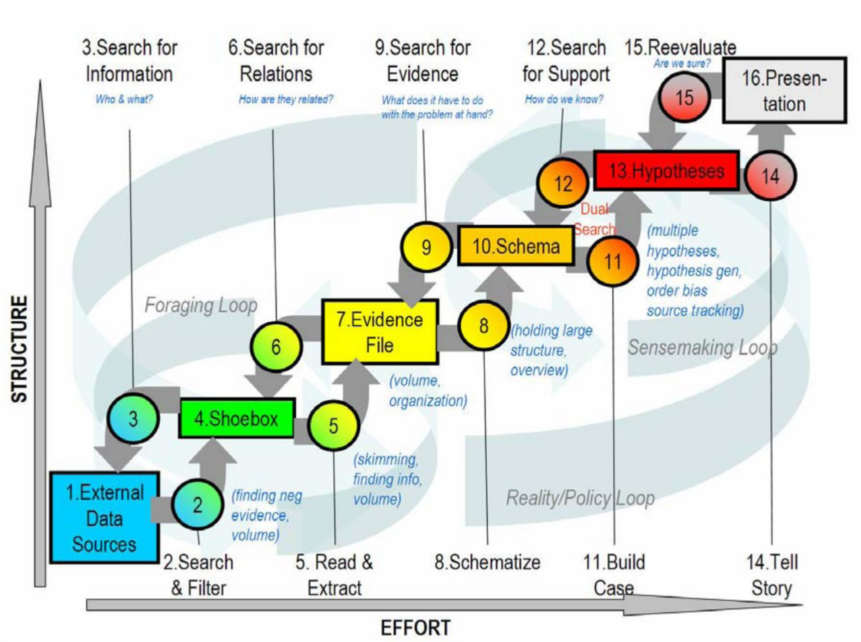


Figure 6. Model of sense-making for intelligence analysis (Pirolli and Card, 2005).

Daniel Keim et al (2008) have drafted a visual analytics process model (Figure 7) based on the visualization processes proposed by van Wijk (2005) and Johnson et al (2006). The model consists of stages and their transition. Complex and heterogeneous data from different sources and various types and levels of quality need to be filtered, noise removed and transformed in order to be processed jointly. The data is processed and abstracted using mathematical, statistical and data mining algorithms and models. Visualizations highlight the important features, including commonalities and anomalies, making it easy for users to perceive new aspects of the data. The principle of a visual analytics tool has been summed up in the visual analysis mantra by Keim et al (2006):

“Analyse First – Show the important – Zoom, Filter and Analyse Further – Details on demand”.

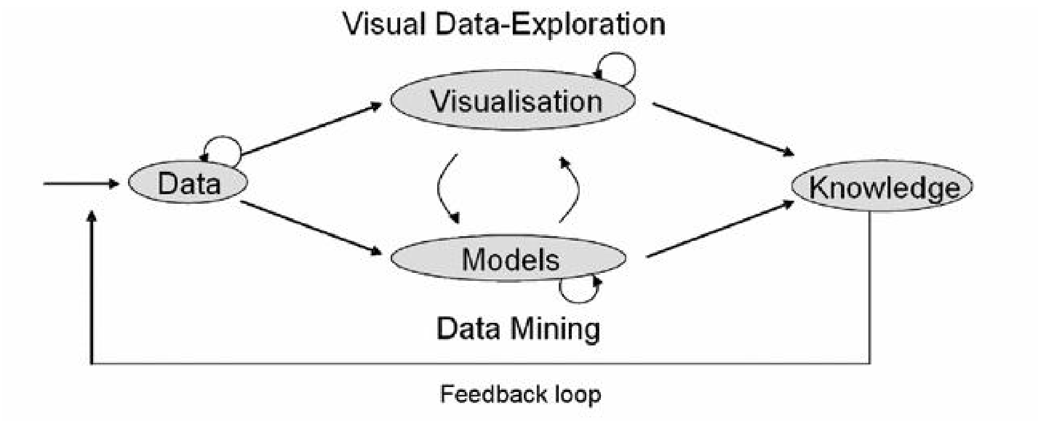


Figure 7. Visual analytics process (Keim et al, 2008).

Kang and Stasko (2011) studied the intelligent reasoning process of visual analytics with empirical testing. Their study involved a group of students doing intelligent analysis studies. Based on the findings, Kang and Stasko introduce an overall analysis process with four iterative phases: (1) constructing a conceptual model, (2) collection of information, (3) analysis, and (4) production. In their guidelines for planning visual analytics tools they question the Pirolli and Card sense-making process. The process is not sequential, instead the analyst works on all four tasks in parallel. According to Kang and Stasko, the key to intelligent analysis is not the analysis of a specific data set, but the construction of a frame. Tools should support all four reasoning phases. The authors emphasize that the results are from the intelligence analysis domain, and can differ from statistical analysis. They also provide guidelines on collaboration, stating that whereas collaboration is part of the process, little collaboration actually takes place during the analysis stage when working on the data and content. The authors recommend asynchronous collaboration.

Shirinivasan and Wijk (2008) claim that in the Pirolli model support of reasoning is limited to the information foraging loop. They suggest a three-view model that also takes into account the sense-making loop: (1) the Data view is a container of interactive visualization tools for exploring data, (2) the Navigation view provides an overview of the exploration process by capturing the visualization states automatically – a history tracking mechanism, (3) the Knowledge view is a diagrammatic representation to record the finding. The three views support the three phases of the analytical reasoning process: model construction, revisiting and falsification. A prototype of the framework (Aruvi) is implemented and evaluated with four analysts having their own data and representing different domains. The analysts agreed that the knowledge view helped the analysis process by building a bridge between visualizations and the knowledge gained.

3.5.2 Human mental model

The mental model or human cognitive model, as the visual analytics community calls it, is defined as a representation of a possible state of affairs in the world (Eysenck and Keane, 2005). There has been discussion that an agreed concept of a human cognitive model for visual analytics is required to be able to construct fluent reasoning environments. No model has yet emerged, but precepts and requirements have been listed.

Liu and Stasko (2010) suggest a mental model that is a functional analogue representation of an external interactive visualization system. Having a mental model enables the layout of visualizations to be kept consistent with users' mental maps. They suggest a model where text, images and coarse spatial relations are overlaid. Green et al (2008) suggest a framework for a human cognition model and a set of guidelines. The central point of the model is discovery. In their approach, a computer presents the information in an ontological-like structure within a relevant, possibly human-defined context, and humans directly interact with the visualized information. New knowledge creation produces relationships and annotations. The model requires multiple views of the same information, direct interaction, insulation of reasoning flow, searching by pattern, and creation and analysis of hypotheses. Green et al evaluate five visual analytics systems against their principles. They state that multiple practical problems are encountered among these systems, but do not specify them.

Instead, they highlight the following areas for further study: what number of multiple views can be used; how to suggest information to users without interrupting reasoning; how to maintain a fluent interactive process; and how to manage annotations.

3.5.3 Precepts and artefacts for reasoning

Several precepts and requirements for visual analytics have been proposed to support analytical reasoning. The earliest precepts are from Amar, Eagan and Stasko (2005), who define primitives for “low-level” analysis tasks. Here, low-level tasks refers to the acquisition of knowledge needed to gain insight into higher phenomena, such as understanding trends and predicting future behaviour. Such tasks include the kinds of specific questions that a person might ask when working with data sets. Amar et al extracted the tasks by interviewing students and analysing their responses. The resulting list of analytical tasks is shown in Table 1. Other tasks, such as comparison, can be constructed using these primitive tasks. The authors list as “high-level tasks” complex decision making under uncertainty, learning a domain, identification of trends, and predicting the future.

Table 1. Analytical tasks.

Retrieve value	Find attributes of specific cases.
Filter	Find data cases satisfying defined conditions.
Computer-derived value	Average, median, count, more complex values.
Find extremum	Find data cases having the highest and lowest value of a defined attribute.
Sort	Rank data cases according to some ordinal metric of a selected attribute.
Determine range	Find a span of values of an attribute of data cases.
Characterize distributions	Create distribution of a set of data cases with a quantitative attribute, e.g. to understand “normality”.
Find anomalies	Find unexpected values from the attributes of the data case. These are usually a fruitful source for further exploration.
Cluster	Find clusters of similar attributes.
Correlate	Find correlations between two attributes of a data set.

The Visual Analytics Agenda (Thomas and Cook, 2005) lists pieces of information, artefacts, that the analyst identifies and creates during the reasoning process. These are classified as: (1) *elemental artefacts* that are derived from isolated pieces of information, (2) *pattern artefacts* that are derived from collections of information, (3) *hierarchical knowledge constructs*, and (4) *complex reasoning constructs*.

3.6 User interaction

The objective of user interaction in visual analytics is to support the analytical reasoning process. User interaction is a broad topic that has been researched intensively since the appearance of graphical user interfaces in 1980s. It studies the relationship between people and technology and looks for methods for utilizing knowledge in the design and development of solutions (Sears and Jacko, 2009), and combines computing, psychology, education, graphic design and industrial engineering. It has been argued that interaction is a neglected component in information visualization, and that the major focus has been on the presentation of information (Yi et al, 2007). Recently, however, the emphasis in visualization has shifted from static visualizations that show all data in one glance to interactive visualizations.

This section focuses on the special questions of user interaction in visual analytics. It introduces the requirements for user interaction in visual analytics, interaction tasks, interacting techniques and special questions concerning big amounts of data, and spatial and temporal data. Specific interaction problems arise when there are huge amounts of data and are introduced in section 3.8.6.

3.6.1 Requirements for user interaction

According to the Visual Analytics Agenda (Thomas and Cook, 2005), user interaction should provide high level dialogue between analyst and information and visual representations is the interface or view into data. The analyst observes the data representations, interprets and makes sense of them, considers the next question to ask, and formulates a strategy for how to proceed. The task of user interaction is to provide ways of gaining new perspectives on the data, filter out results, and request new visual representations. The user should be protected against cognitive overload, overcome biases, and be encouraged to continue the analysis until the analysis produces the ultimate outcome. There should be a fluent dialog between analyst and information, where the user can engage fully in the analytical reasoning process.

The Visual analytics agenda categorises the interaction needs into four groups. The first is interactions *for modifying data transformation (filtering)*. These include techniques such as direct manipulation, dynamic queries, brushing, and details-on-demand. The next group is interactions *for modifying visual mappings*. These include techniques that allow users to interactively change mappings between the data and their visual representations. The third group is interactions for *modifying view transformations (navigation)*. These include selecting and highlighting objects of interest, panning and zooming and balancing between overview and detail. The final group is *interaction for human information discourse*. These are interactions to support creating abstractions, comparing and categorizing data, extracting, recombining, creating and testing hypothesis and annotating data.

Scholtz (2006) proposes requirements from the usability viewpoint, distinguishing five areas where support for reducing cognitive workload is needed. These are: (1) *Situation awareness*: awareness of where the user is in the process, and what is the best step forward; (2) *Collaboration:* where multiple people are working together, synchronously or asynchronously, in order to keep track of doings and findings, (3) *Interaction*: capability to view occluded information, move the level of abstraction of the view up and down, high-level un-dos, data sharing (4) *Support of creativity,* and (5) *Utility*, bringing improvements to the process and product.

By Ware (2004) interaction with visualizations refers to the dialog between the user and the system as the user explores the data set to uncover insights. It is a process made up of a number of interlocking feedback loops divided into three phases. At the lowest level is the *data manipulation* loop, through which objects are selected and moved using eye-hand coordination. At the intermediate level is an *exploration and navigation loop* through which the analyst finds their way in the large visual space. At the highest level is the *problem-solving loop* through which the analyst forms hypotheses about the data and refines them by repeating the process. Each step has a time scale for human action reflecting what the user is cognitively and perceptually capable of doing.

3.6.2 Interaction tasks

According to the Visual analytics agenda, visual analytics needs two kinds of interaction technique: *high-level operations*, and *basic interactions* that support these higher processes. However, there is no clear understanding of what the high and low-level operations are. The agenda calls for taxonomies of user interaction techniques that support analytical reasoning.

There are a variety of taxonomies that list and classify the interaction tasks or techniques to be used among visualizations. Yi et al (2007) have listed eleven taxonomies and categorized them as (1) taxonomies of low-level interaction techniques, (2) taxonomical dimensions of interaction techniques, (3) taxonomies of interaction operations, and (4) taxonomies of user tasks.

Yi et al have also created their own taxonomy by reviewing the literature. They propose seven general categories of interaction: (1) Select: “mark something interesting”; (2) Explore: “show me something else”; (3) Reconfigure: “show me a different arrangement” – where different perspectives of the data are shown; (4) Encode: “show me a different representation” – where colour, size, fonts, shapes and orientation are used to visually encode data points with attributes of interest; (5) Abstract/Elaborate: “show me more or less detail” – means abstracting data into higher-level component and inversely, to details; (6) Filter: “show me something conditionally” – focusing on important data, decreasing complexity, e.g., leaving out statistically unimportant data points. All kinds of filters, sliders, and dynamic queries are useful here; and (7) Connect: “show me related items” – where associations and relationships between data items are highlighted (e.g., brushing).

Zhou and Gotz (2009) have surveyed the actions of several different visual analysis environments and identified twenty distinct actions. These are (listed in alphabet order): annotate, bookmark, brush, create, metaphor, delete, edit, filter, inspect, merge, modify, pan, query, redo, undo, remove, restore, revisit, split, zoom. They further categorized these actions as *exploration actions*, *insight actions* and *meta actions*. The exploration actions (query, filter, pan and zoom) are performed as users access and explore data in search of new insight. Insight actions are performed as they discover or manipulate the insights obtained over the course of analysis, such as annotating and bookmarking. Meta actions, such as undo and redo, are related to the users' action history, not to the visual presentation or the data set.

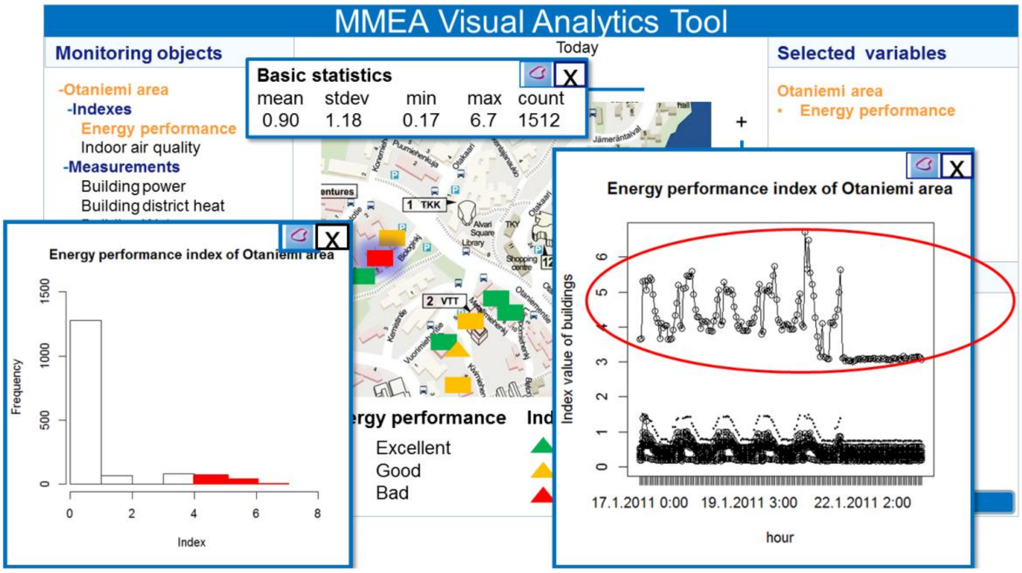
The interaction tasks vary depending on the visualization representations. Different operations are required for spatio/temporal visualizations or hierarchical and network structures. Chi (2002) has categorized the operations of different kinds of visualizations. They include scientific, geographical-based, 2D, multi-dimensional plots, information landscapes and spaces, trees, text, web visualizations, and visualization spread sheets. Chi also introduces 50 visualizations and their location in the taxonomy. The taxonomy introduces each visualization, gives example data, analytical needs, visualization transformations, abstraction and visual mapping, and operations within view.

Shneiderman (1996) introduces seven tasks for information seeking when interacting with large data sets: overview, zoom, filter, details-on-demand, relate, history, and extract.

3.6.3 Interaction techniques

In visual interaction, there are two basic interaction techniques: *Direct manipulation*, which allows the user to filter or select elements of visualizations, and *dynamic queries* where the user interacts with sliders, menus and buttons. Direct manipulation techniques are recommendable because they do not distract attention from the analysis process. The menus, buttons and sliders are often scattered around the user interface, and using them requires extra effort (Roberts, 2007).

A popular technique in visual analytics is using *coordinated multiple views* (Roberts, 2007), which is a specific exploratory visualization technique. Data is represented in multiple windows and operations in the views are coordinated. This means that data elements which are selected and highlighted in one view are highlighted concurrently in all other views that include the same data element. This operation is called brushing. The user can change the style of brush, the bounding region and the brushing effects. The method is effective for discovering outliers. An example of brushing is shown in Figure 8.



Basic statistics

mean	stdev	min	max	count
0.90	1.18	0.17	6.7	1512

Figure 8. Brushing. The rounded area is highlighted in the histogram and on the map.

3.6.4 Spatial and temporal data

Temporal and spatial data have operation lists of their own. Aigner et al (2007) give a list of requirements for temporal data. Browsing the time axis is viewed as important as well as switching between different levels of temporal aggregation: days, weeks and months. The choice of temporal scale is also considered important, as relationship patterns may not be detectable when examined using certain scales. Scaling is also needed to integrate data measured at different scales.

Spatial data operations are listed by Andrienko et al (2010). The operations rotate, zoom in and out, fly-over, annotate points of interest, and hyperlink, are viewed as basic requirements. Scaling is also considered important and may significantly affect the results, as with temporal data.

3.7 Data analysis

Data analysis has a key role in visual analytics. The goal of data analysis is to examine data with the purpose of drawing conclusions about the information. Several techniques are employed, each using similar methods but having a slightly different focus. These include data mining, knowledge discovery, machine learning, and statistics, among others (Figure 9). Data mining is defined as "a science of extracting useful information from large data sets or databases (Tan et al, 2006). Machine learning is "programming computers to optimize a performance criterion using example data or past experience" (Alpaydin, 2004). Sometimes the division between machine learning and data mining is done based on data sets. Data mining is focused on analysing large databases, whereas in machine learning the focus is on learning patterns from data. The roots of data analysis are in statistics. The development of computers and ability to store and manage large amounts of data has made possible large-scale statistical computation, and has launched the development of new methods that would be tedious to perform manually.

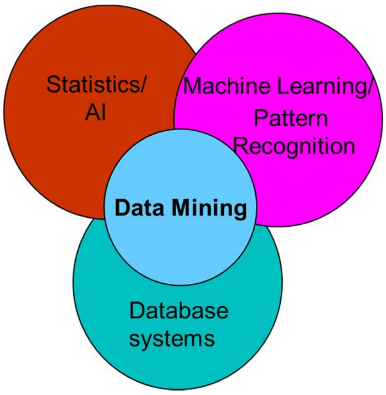


Figure 9. Data analysis techniques (Tan, Steinbach and Kumar 2006).

Data analysis has been studied intensively and numerous algorithms exist. It has applications in different business, science, and social science domains. A wide range of tools and commercial applications are available, some of which are highly competitive in markets, such as Customer Relationship Management (CRM). There are also several statistics programs and packages available, both for casual users and specialists (Excel[13], SAS[14], SPSS[15], R[16]).

A recent area of data analysis is visual data mining. Information visualization, data mining and user interaction have evolved as separate fields in the past, but since the turn of the 2000s have become increasingly integrated as visual data mining. The idea of visual data mining first emerged in 1999 when Wong (1999) argued that rather than using visual data exploration and analytical mining algorithms as separate tools, a stronger data mining strategy would be to couple the visualizations and analytical

[13] http://www.microsoft.com (accessed 18 October 2012)
[14] http://www.sas.com/technologies/analytics/statistics/ (accessed 18 October 2012)
[15] http://www-01.ibm.com/software/analytics/spss/ (accessed 18 October 2012)
[16] http://www.r-project.org/ (accessed 18 October 2012)

processes into one data mining tool. Many data mining techniques involve mathematical steps that require user intervention, and visualization could support these processes. Visual data mining is not just about using visualization to exploiting data, it is an analytical mining process in which visualizations play a major role (Ferreira de Oliveira and Levkowitz, 2003).

The newest branch (Wong, 1999) of data mining is isually-controlled data controlled mining, a new class of data mining methods characterised by Puolamäki, Papapetrou and Lijffijt (2010), who define the requirements for a useful data mining method in visual analytics as follows: it should be fast enough – sub-second response is needed for efficient interaction, the parameters of the method should be representable and understandable using visualisations, and parameters should be adjustable by visual controls.

This section represents the data analysis process and gives an overview of data analysis methods. Numerous methods have been developed for data analysis. The methods introduced here represent the most common and those that are considered useful in the point of view of this work. They cover (1) methods for data exploration, (2) descriptive methods, (3) predictive methods and (4) methods for anomaly detection. This introduction omits a large number of methods, including methods for text analysis and other unstructured data. Examples of basic visual data mining techniques are introduced in the next section on visualization.

3.7.1 Data analysis process

Data analysis is an iterative process starting with selecting the target data from the raw material and pre-processing and transforming it into a suitable form (Figure 10).

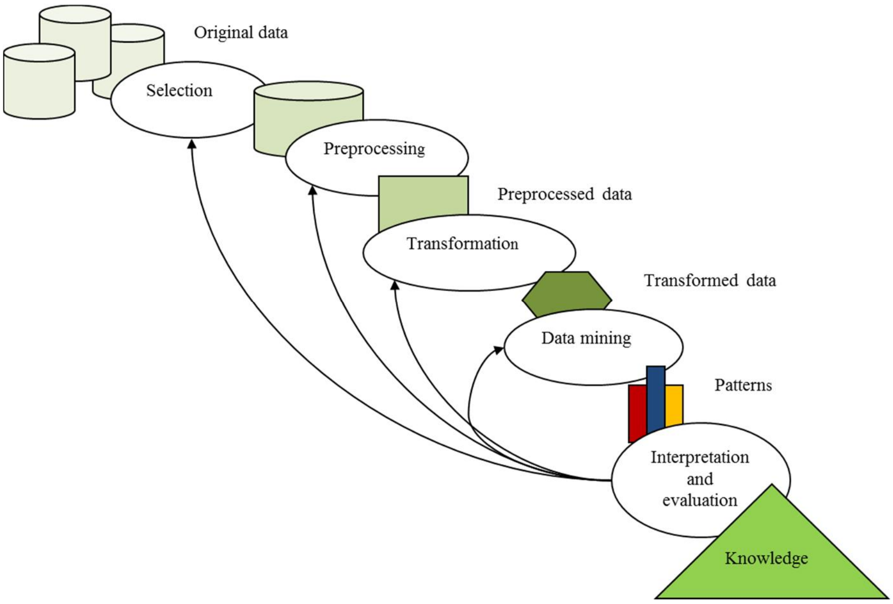


Figure 10. The data mining process (Figure adapted from Tan, Steinbach and Kumar (2006).

As in visual analytics, data analysis uses several data types: database records, matrix data, documents, graphs, links, transaction data, transaction sequences, sequence data, genomes, spatio-temporal data. Similarly, the quality of data causes problems. The data can contain noise, there may be missing values and duplicate data, and a thus data cleaning phase is required before using the data. Other kinds of pre-processing may also be required, such as data aggregation, sampling, dimensionality reduction, subset selection, feature creation, and attribute transformation (Tan et al 2006).

Next, the data is run through a data mining algorithm that creates patterns from the data. The user interprets and evaluates the results and starts a new iteration with possible modifications to the raw data, algorithm and algorithm parameters.

3.7.2 Methods for data exploration

The purpose of data exploration is to gain a better understanding of the characteristics of data (Tan et al, 2006). The central methods are summary statistics and visualizations. Summary statistics are numbers that summarize properties of the data. Amar et al (Amar, Eagan and Stasko, 2005) have classified the statistical methods as (1) computer-derived values; average, median, count, more complex values, (2) finding extremum; finding data cases having the highest and lowest value of a defined attribute, (3) determining range: finding a span of values of an attribute of data cases, and (4) characterizing distributions: creating a distribution of a set of data cases with a quantitative attribute, e.g. to understand “normality”. The visual methods utilize humans’ ability to recognize patterns. Single variables are expressed in visual form, for instance as histograms and line charts.

Correlation is a basic statistical method of studying two variables. The prevailing method is calculation of the Pearson correlation coefficient (r), where the correlation between two variables, x and y, is calculated with the formula:

$$r = \sum_{i=0}^{n} \frac{(x_i - x)(y_i - y)}{\mathrm{n} S_x S_y}$$

where n is the number of observation pairs, S_x and S_y are the standard deviations, and x and y the means of the variables x and y. The correlation produces positive or negative values within the range -1 to 1. If the result is zero, there is no correlation between the variables. Values -1 and 1 indicate complete linear dependence between the variables, either negative or positive. Often the square of the correlation coefficient R^2 is calculated. This value ranges from 0 to 1, and indicates how much one variable explains the variance of the other, and is often expressed as a percentage. For instance, if R^2 is 0.32, 32% of the variance of a variable is explained by the other.

Correlations are visualized in the form of scatterplots. Exploration methods for higher dimensions use projections of data on a two-dimensional plane. These are called dimension reduction methods. They include principal component analysis (PCA) and multidimensional scaling. The result of PCA can be visualized as a 2-dimensional plot. Visual methods are introduced in more detail in Section 3.8.

3.7.3 Descriptive methods

The goal of descriptive methods is to discover patterns and rules in data. The methods focus on finding clusters, patterns and associations from data. (Tan et al, 2006)

Clustering looks for groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The similarity of objects is defined based on similarity (or distance) measures. Euclidean distance can be used if attributes are continuous; otherwise problem-specific measures are needed. Clustering has been an active research topic and lots of algorithms are available. Algorithms include K-means clustering and its variants, hierarchical clustering, agglomerative clustering and density-based clustering. Market segmentation is an application of clustering.

Pattern detection involves finding combinations of items that occur frequently in databases. Sequential pattern discovery finds rules that predict strong sequential dependencies among different events.

Association rule mining involves the prediction of occurrences of an item based on occurrences of other items. It produces dependency rules such as "buyers of milk and diapers are likely to buy beer".

3.7.4 Predictive methods

The purpose of predictive modelling is to build models that predict the value of one variable from the known values of other variables (Hand, Mannila and Smyth, 2001). The predicted objects are predefined. Regression and classification are two much used predictive methods.

Regression predicts a value of a continuous variable based on other variables using linear or nonlinear models (Tan et al, 2006). Linear regression is easy to visualize, often shown as a line on a scatterplot diagram. The area is studied extensively and has its origins in statistics. It has various uses, both in commerce and science. Application examples include predicting sales based on advertising expenditure, stock markets, or wind as a function of temperature or humidity.

Classification creates a model for a class attribute as a function of the values of other attributes (training set). Unseen records are then assigned to the class. The accuracy of the models is evaluated with a test set. Several techniques have been developed including decision trees, Bayesian methods, rule-based classifiers and neural networks. Classification is a much used method and commercial applications are also available. Examples include classification of credit card transactions as legitimate or fraudulent, classification of e-mails as spam, or classification of news stories as finance, weather, entertainment or sports (Tan et al, 2006).

3.7.5 Methods for anomaly detection

Anomalies are observations whose characteristics differ significantly from the normal profile. Methods of anomaly detection look for sets of data points that are considerably different from the remainder of the data. The methods build a profile of "normal" behaviour and detect significant deviations from it. The profile can be patterns or summary statistics for the overall population. Types of anomaly detection schemes can be graphical-based, statistical-based, distance-based or model-based. Credit card fraud detection, telecommunication fraud detection, network intrusion detection and fault detection are examples of application areas (Tan et al, 2006).

3.8 Information visualization

Visualizations are an efficient way to find insight from data. Visual representations translate data into a visible form that highlights important features and make it easy for users to perceive salient aspects quickly. Visualization takes advantage of the capabilities of the human eye to see, explore and understand large amounts of information at once, providing external aids to cognition (Thomas and Cook, 2005).

Visualizations are entrusted to amplify human cognitive capabilities. Card et al (Card, Mackinlay and Shneiderman, 1999) list six basic ways: (1) creating cognitive resources by using a visual resource to expand human working memory; (2) reducing the search, e.g., by representing a large amount of data in a small space; (3) enhancing recognition of patters, e.g., by organizing information according to its time relationships; (4) supporting the easy perceptual inference of relationships; (5) supporting perceptual monitoring of a large number of potential events; (6) providing a manipulation medium that enables exploration of a space of parameter values.

Visual presentations have a long history, originating from the early maps of ancient China. The Konya Town Map dating from 6200 BC is the first known visualization. Computer-aided visualization has been studied actively since 1980 when the development of computing power made image processing possible. Improved rendering, real-time interactivity and lowering costs have accelerated this development. In the early years of information visualization the focus was on viewing the entirety of a data set at a glance, to discover interesting and hidden patterns and connections. Visualizations showing the entirety of the data set have been replaced by tools that support the process of seeking insight (Chen, 2010).

Collaborative and social visualization is a new phenomenon that has developed along with social media. Users publish easy-to-use tools and visualizations on the web, and communicate findings in easy-access visual form with other people (e.g., IBM's Many Eyes). The inclusion of visualization features in office tools, such as Excel, and free visualization tools on the Web have brought the creation of visualizations within reach of all.

Nowadays, a vast amount of visualizations are readily available. Lengler and Martin (2007) have identified 160 visualization methods in their study classifying management visualizations. Of these, they selected 100 methods and classed them into six categories, forming a “periodic table of visualization methods”. These include: (1) data visualizations including pie charts, area charts and line graphs; (2) information visualizations including semantic networks or treemaps; (3) concept visualizations such as concept maps or Gantt charts; (4) metaphor visualizations such as metro maps; (5) strategic visualizations such as the strategy canvas; and (6) compound visualizations consisting of several of the aforementioned formats. Chen (2010) classifies visualizations into scientific visualizations (based on physical data) and information visualizations (abstract data), or functional and aesthetic visualizations. Visualization is further classified as abstract or representative.

The sheer variety of visualization makes it difficult for users to identify the most relevant visual presentation for the task at hand. Selecting the right visualization depends on the properties of the data and the purpose of use and on the dimensionality of the data, the data type and structure, and the size of data set. The visualizations for single-variable ordinal data are different from multivariable quantitative data, and some visualizations work well only with small datasets.

This section introduces a sample of the variety of visualizations. Many of them are familiar from statistics and belong to the basic exploration and descriptive methods of data analysis. They are general-purpose and easy to interpret, and have proven their value during their long mileage of use. The introduction is structured based on the data types: univariate data, bivariate data, multivariate data, and structured data, spatial and temporal data, and the approaches for visualization of large amounts of data. Most of the examples are made with the help of the R statistical package.

3.8.1 Univariate data

Bar charts (Figure 11) and pie charts (Figure 12) are the basic methods used to visualize univariate ordinal data. Data is represented in classes based on levels or factors. In bar charts, the value of the bar represents the sum of the data points belonging to the class represented by the bar. The bar height is proportional to the number of data points. The height can be representative of frequency or proportion.

In *pie charts* values are represented as proportional sectors. Pie charts are not very effective at showing differences between sectors, especially if the differences are small (Venables and Smith, 2012).

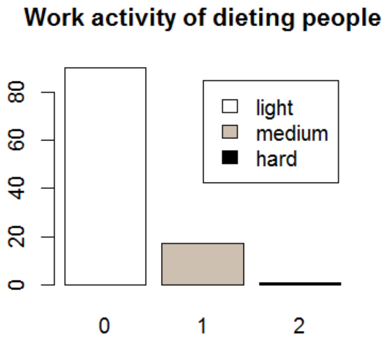


Figure 11. Bar chart.

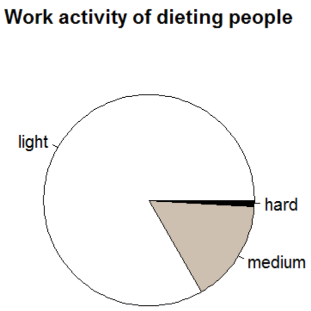


Figure 12. Pie chart

Histograms (Figure 13) are similar to bar charts, but are used to represent quantitative data. The histogram defines a sequence of breaks and then counts the number of observation in the bins formed by the breaks. It plots these with a bar similar to the bar chart. In addition, histograms can add a point at the top of the rectangle and then connect these points with straight lines. This is called the *frequency polygon* (Venables and Smith, 2012).

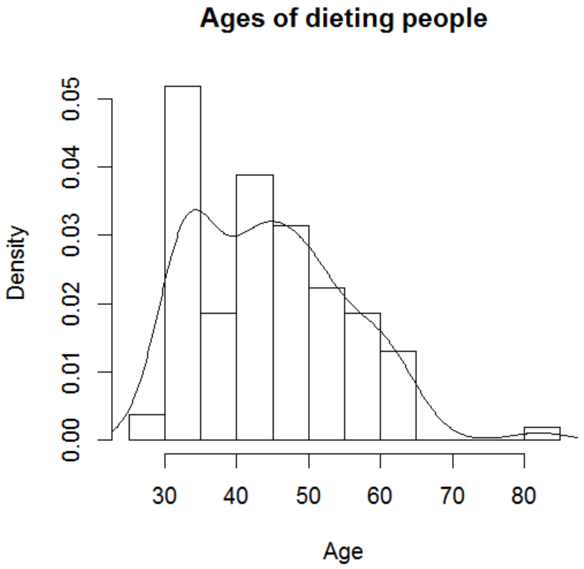


Figure 13. Histogram.

Line graphs are used for displaying quantitative data as a continuous function of a single variable. Common uses are showing frequency distributions (Figure 14) and time series. Time series show the evolution of data over time by placing the data values on a time axis (Verzani, 2002).

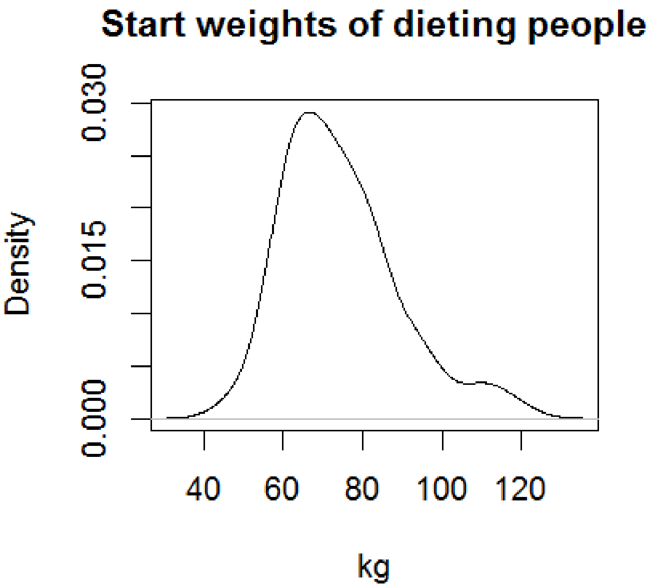


Figure 14. Line chart.

3.8.2 Bivariate data

The basic visual method for analysing bivariate data is the *scatterplot* (Figure 15) (Hand et al, 2001). Scatterplots are a good means of finding correlations, clusters and outliers between two attributes. A third dimension can also be added by using a visual effect such as colour and size of plot (bubble charts), or animations (animated bubble charts). A linear regression line can be added to the plot to indicate correlations between variables. Scatterplots are inadequate for higher dimensions and if there is a very high density of data points, in which case it becomes difficult to distinguish interesting features (Hand et al, 2001).

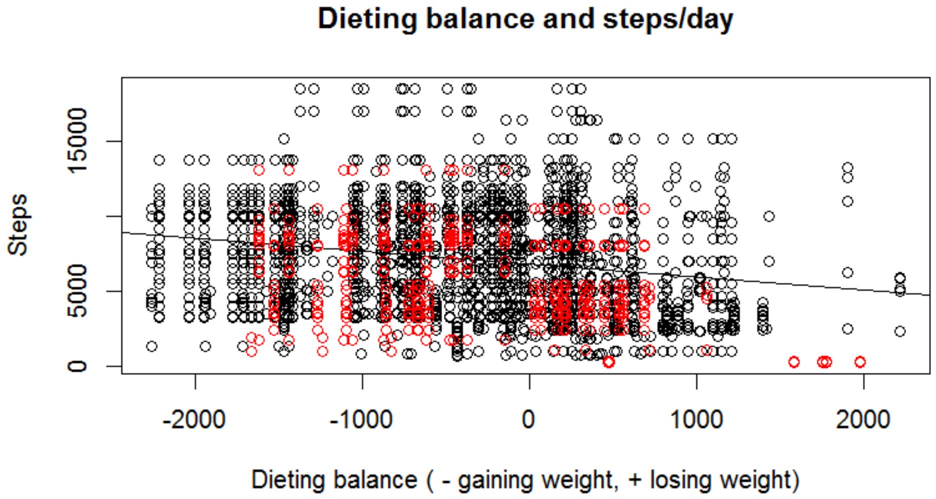


Figure 15. The scatterplot shows the number of daily steps taken by the dieting people and the dieting balance. The regression line does not indicate much correlation between number of steps taken and dieting balance. The red spots represent individuals who are feeling low.

3.8.3 Multivariate data

Often, real-world data is multidimensional, consisting of many data items or without a clear hierarchy. Dimension reduction methods aims at projecting data into a low dimensional space (1D-3D) while maintaining the correct relations between the nodes. There are several methods with different optimization goals and complexities. One of the best known is *Principal Component Analysis (PCA)*. It tries to find a linear subspace that has maximal variance. It is based on matrix algebra (Alpaydin, 2004).

An example of a PCA plot of the daily nutrition of dieters is shown in Figure 16.

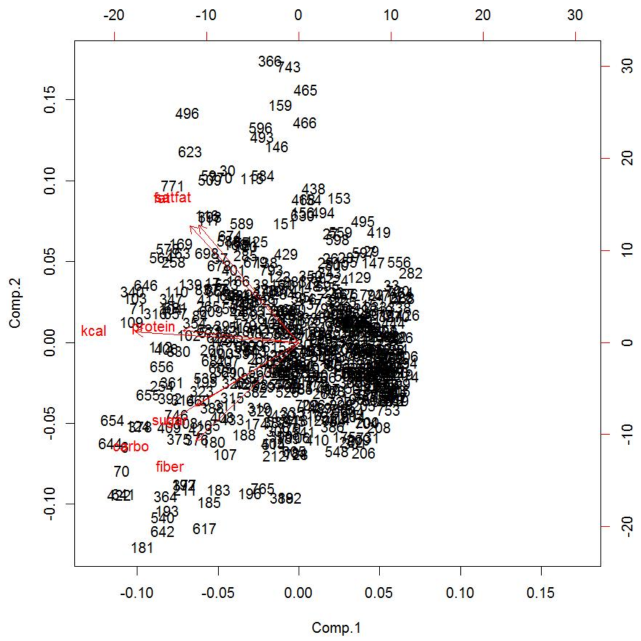


Figure 16. A PCA of food nutrients, indicating a group of sugar-carb-fibre eaters, a group of protein eaters, and a group of fat eaters among the dieters.

In *parallel coordinates* visualization each of the dimensions corresponds to a vertical axis and each data element is displayed as a series of connected points along the dimensions/axes. To show different combinations the order of the variables is changed (Hand et al, 2001). An example is shown in Figure 17. Subsets can be highlighted with colours.

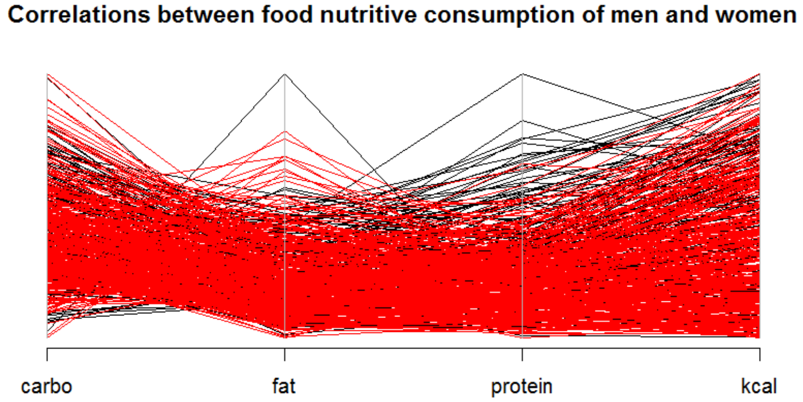


Figure 17. Parallel coordinates visualization of nutrient consumption. In the figure, black represents men, red women.

If there are more than two variables in the dataset *scatterplot matrixes, correlation matrixes* or *correlation networks* can be used to show pairwise correlations for all variable combinations (Hand et al, 2001). Figure 18 shows a scatterplot matrix, Figure 19 a correlation matrix and Figure 20 a correlation network of the same data. Each brings slightly different aspects to light. The data represents the daily nutrient consumption of dieting people.

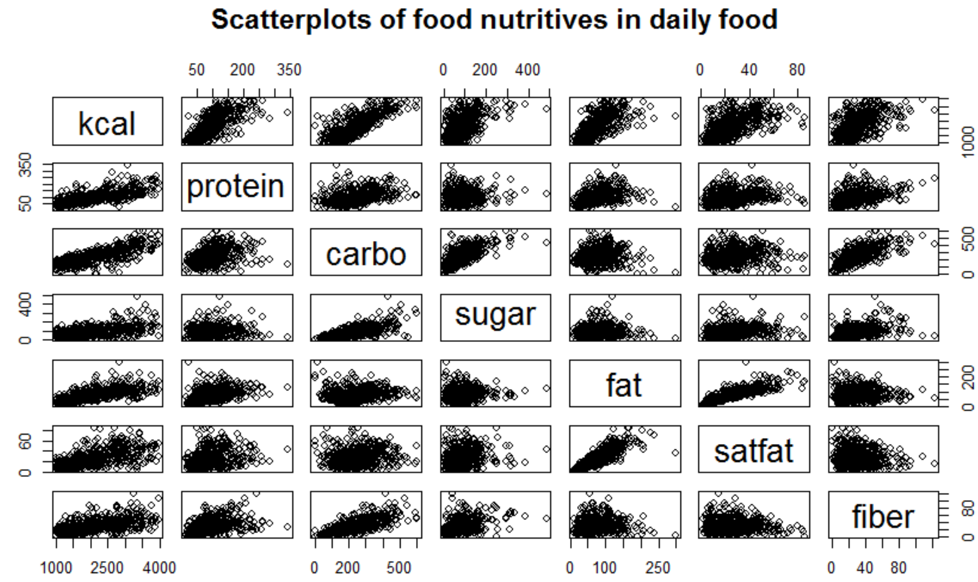


Figure 18. Scatterplot matrix, enabling shapes and forms to be examined. The plot between kcal and carbohydrates has a clear linear upward shape, indicating positive correlation. The shape between fibre and fat is descending and indicates a slight negative correlation. Shapes without a clear form indicate low correlation between the variables.

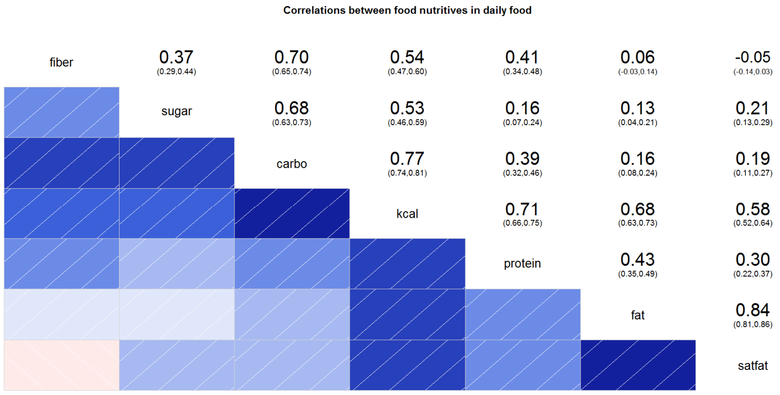


Figure 19. In the correlation matrix, colours indicate the strength of correlations. Blue is positive and red negative. The matrix shows a strong correlation between kcals and carbohydrates, suggesting most daily energy derives from carbohydrates. Another interesting feature is the negative correlation between fibre and saturated fat, suggesting that people who eat saturated fats eat less fibre.

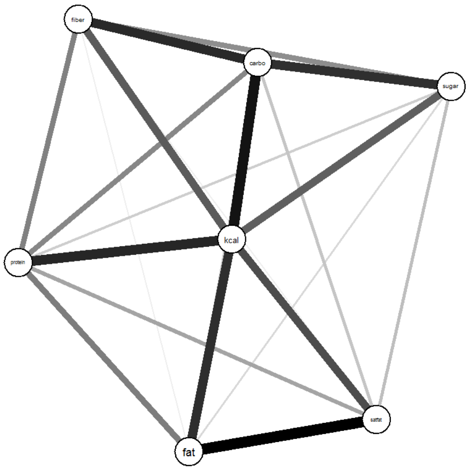


Figure 20. In the correlation network the nodes express the variables and the link thickness indicates the correlation value. Here again, the correlations between kcal, fat and carbohydrates are strong. The strong correlation between fat and saturated fat is expected.

A *cluster dendrogram* (Figure 21) is a tree-like graphical representation of hierarchical clustering. The entire sequence of merging clusters is shown (Hand et al, 2001).

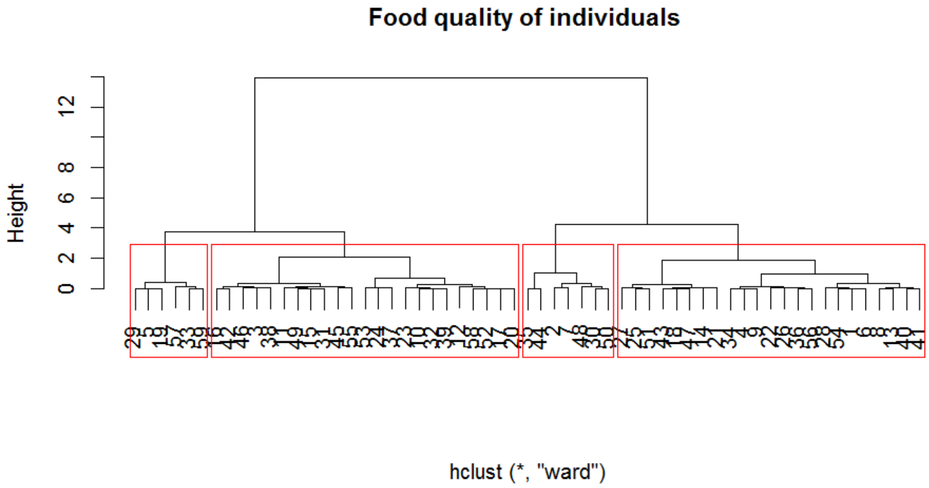


Figure 21. Individuals categorized into four clusters based on the food quality index.

Glyphs are symbols that are used to describe multivariate discrete data. A single glyph corresponds to one sample in a data set. Data values are mapped to the visual properties of the glyph. Glyphs can be constructed from pre-attentive visual features that “pop out”, such as: spatial position, colour, shape, orientation, texture, motion, and blinking. Glyphs can represent 4–8 dimensions, but most effectively four dimensions. Glyphs can be constructed so that they are perceived pre-attentively. An example of glyphs is the Chernoff face (Figure 22) where data is mapped to facial expressions. Other popular glyphs are star glyphs, where dimensions are represented as angular spokes radiating from the centre, and stick figure icons where data dimensions are mapped to the rotation angles of the limbs (Ware, 2004).

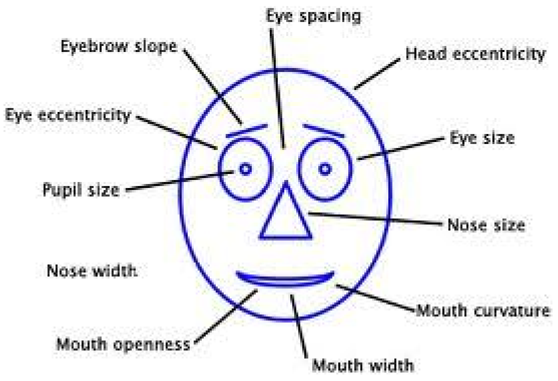


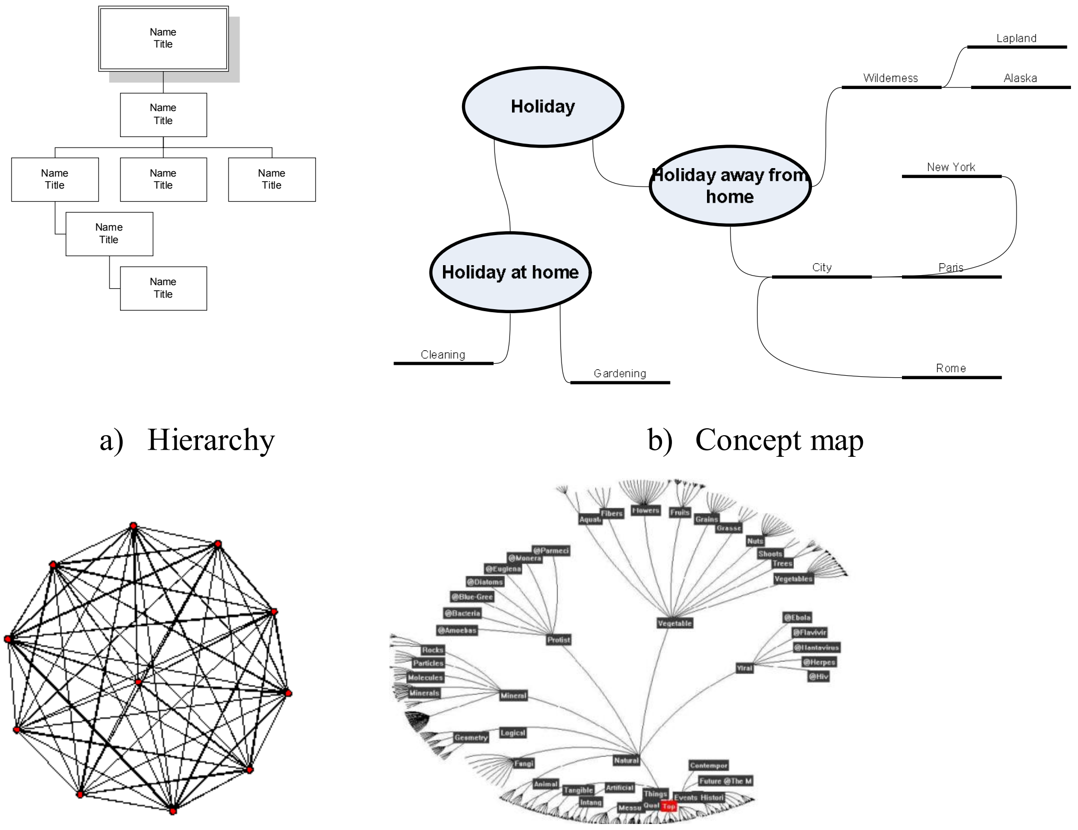
Figure 22. Chernoff faces [17]

3.8.4 Hierarchical and network structures

If the data has a hierarchical or network structure it can be represented as trees and graphs. There are two major approaches: node link diagrams and enclosures, such as treemaps (Card, Mackinlay and Shneiderman, 1999). Node-link diagrams have several variants, including traditional vertical or horizontal hierarchies, networks, hyperbolic (star) trees and concept maps (Figure 23). The UML model used in this work is one application of a node-link diagram. There are numerous drawing algorithms for optimizing the graph layout of node-link diagrams (Card et al, 1999).

Node-link diagrams are useful in topology-based visual analytics tasks, such as finding clusters and connected components, detecting patterns and outliers, and determining the shortest path between nodes. But there are also limitations. They do not scale up well. They produce cluttered overviews with few readable labels (Kang, Getoor and Singh, 2007).

[17] Chernoff faces**:** http://kspark.kaist.ac.kr/Human%20Engineering.files/Chernoff/Chernoff%20Faces.htm
(accessed 18 October 2012)



a) Hierarchy

b) Concept map

c) Network

d) Hyperbolic tree (Pirolli, Card and Van Der Wege, 2003)

Figure 23. Different kind of node-link diagrams.

Treemapping, developed by Johnson and Shneiderman (in Card, Mackinlay and Shneiderman, 1999), is a method for visualizing hierarchically structured information. A treemap shows data in square cells, the size which is proportional to the cell frequency. It allows very large hierarchies to be displayed in their entirety. Figure 24 shows an example.

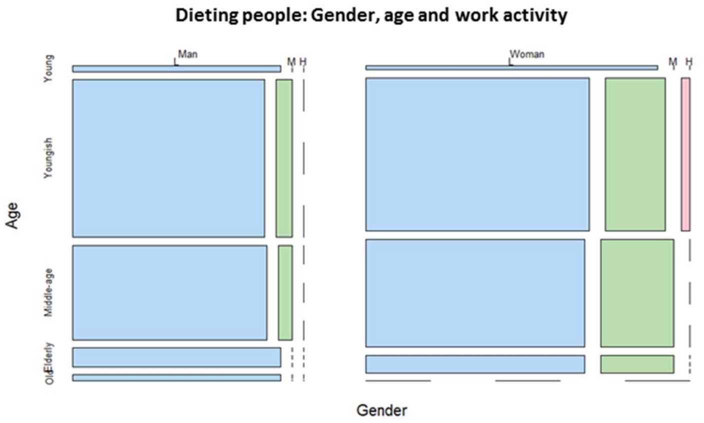


Figure 24. Treemap showing the age, gender and work activity of dieting people. The biggest group is youngish women who do light work.

3.8.5 Temporal and spatial data

Temporal data

Temporal data embodies the change in properties over time. Time has a hierarchical system of granularities including seconds, minutes, hours, days, weeks, months, years, decades and centuries. These are organized in different calendar systems. Time is also cyclic, consisting of natural cycles and re-occurrences. Some are regular and predictable, such as seasons, or less regular, such as social or economic cycles. Time can also be branching. Branching time represents alternative time scenarios. Linear time corresponds to the natural perception of time, and most visualizations have a linear time axis (Aigner et al, 2007)

Time series diagram (Figure 25) is the basic visualization of linear time. It shows the evolution of the data over time by placing data values on a linear time axis, thus supporting the detection of temporal trends and patterns. In addition to time series, time points can be interpreted as an ordinary numeric variable and used in any kinds of visualizations that apply to quantitative data.

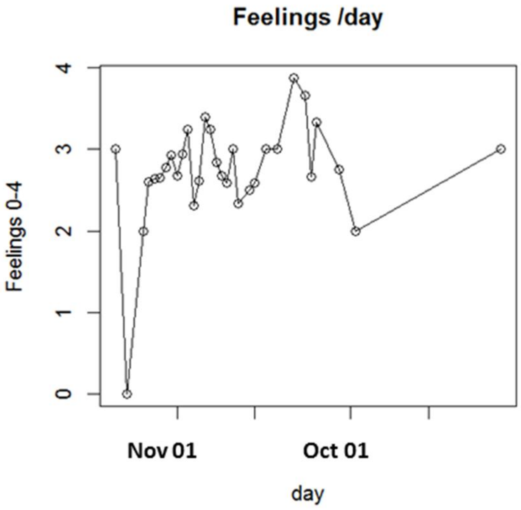


Figure 25. Time series. The figure shows the variation in mood of the dieting people during the dieting period.

Correlations between time series can be revealed by *cross-correlation*, a standard method of estimating the degree to which two series are correlated (Bourke, 1996). An example of cross-correlation is shown in Figure 26.

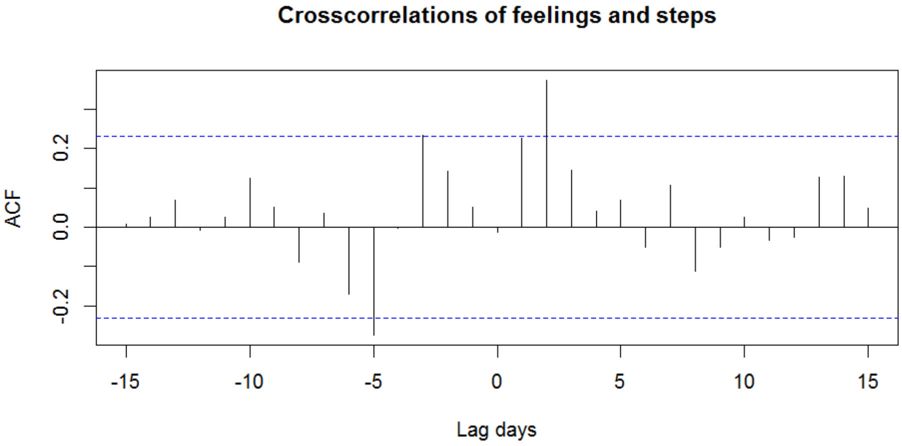


Figure 26. Cross-correlation of feelings and exercise. The biggest correlation has a lag of three days. This may be interpreted as indicating that experiencing good feelings on day one results in a high exercise level on day three, or the opposite. The lowest correlation has a lag of 5 days, possibly indicating that an active day results in feeling low five days later. A more accurate interpretation would require examination of the data values.

Auto-correlation is the cross-correlation of a variable with itself. Figure 27 shows an example of auto-correlation.

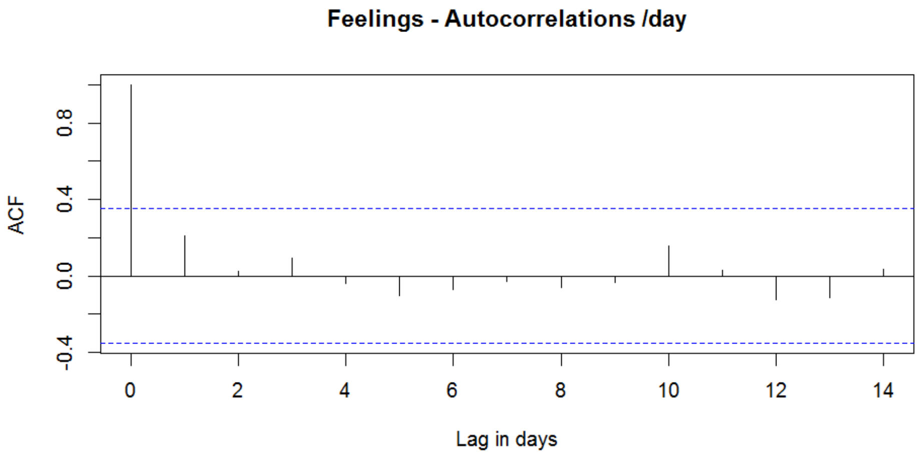


Figure 27. Auto-correlation. The figure indicates no strong correlation. The feelings of today do not predict the feelings of tomorrow.

Cyclic time is composed of a set of recurring temporal elements, such as seasons of the year or days of the week. Rolling data points on a cyclic time axis at different speeds is an efficient way to find cyclic patters of time. Figure 28 (Aigner et al, 2008) is an example of a cyclic presentation. Branching time can be represented with charts such as a Gantt chart. A comprehensive overview of time-oriented visualization can be found in the study by Aigner et al (2011).

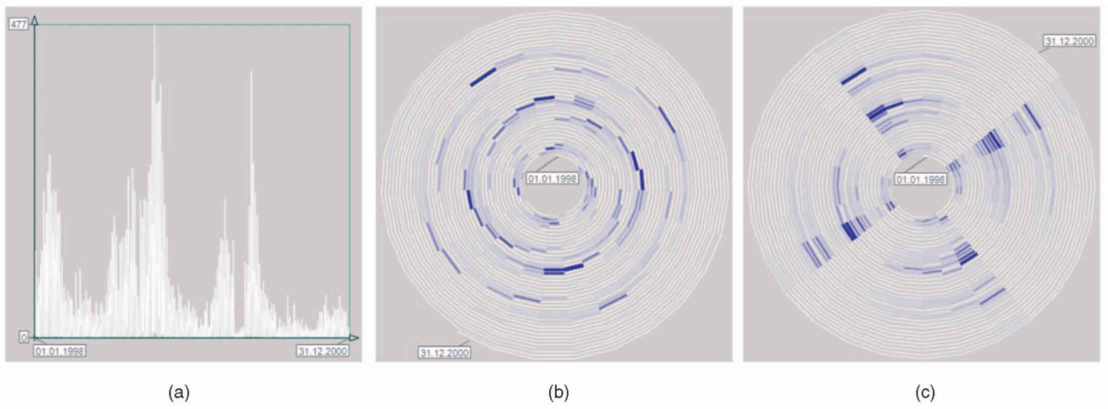
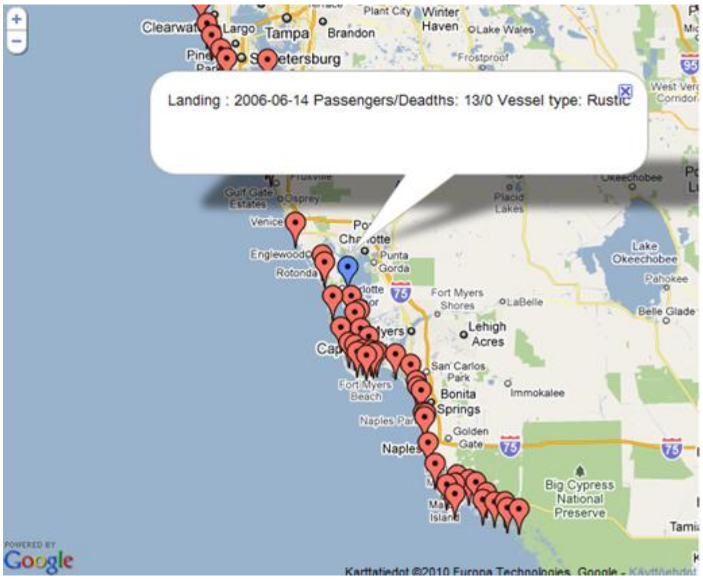


Figure 28. Cyclic time (Aigner et al, 2008).

Spatial data

Spatial data has a geographic or spatial component and can be mapped to a location. Spatial data has a number of properties that distinguish it from other types of data. Spatial data is mapped to a location, and properties such as distance, proximity, direction and scale variance are meaningful. Andrienko et al (2007) list the structure and properties of geographic spaces. Physical spaces are heterogeneous, consisting of islands, mountains, oceans, state boundaries or cities. There are also specific metric properties and topological relations, e.g. distances in actual geographic spaces differ from their metric properties, and places near to each other are more related than distant spaces. Phenomena and events can occur in a physical space at different points in time. Heterogeneity is therefore a complicating factor in fully automatic processing.

Spatial data has been traditionally presented by means of maps. Geovisualization has a long tradition and conventions of showing data on maps. Advances in computing have brought interactive maps, dynamic maps, three-dimensional maps and maps combined with other graphics. Through web applications with open API (e.g., Google maps) users can add information of their own to maps. An example is shown in Figure 29.



Landing : 2006-06-14 Passengers/Deadths: 13/0 Vessel type: Rustic

Figure 29. Google Map Visualization[18]. Landing places of boat refugees presented on Google Map.

Spatio-temporal data

Spatio-temporal data combines the properties of temporal and spatial data. Andrienko et al (2010) describe methods for presenting spatio-temporal data. These include (1) map animation; (2) the space-time cube, in which two dimensions represent space and the third dimension represents time; and (3) coordinated multiple views, which can show spatial, temporal and thematic aspects of data simultaneously.

[18] Personal material

3.8.6 Large collections of information

The problem of representing, navigating and finding details in large collections of data is known as the *focus + context problem*. The basic principle of exploring information collections is given by Ben Shneiderman (1996):

> "Overview first, zoom and filter, details on demand".

The principle is to keep a view of the whole data available (context), while pursuing detailed analysis of a part of it (focus). Information needed in the overview may be different to that needed in detailed analysis. These two types of information should be combined within a display.

Card et al (Card, Mackinlay and Shneiderman, 1999) classify the methods for information selection and reduction that are used among the techniques. These include filtering, selective aggregation, micro-macro-readings, highlighting and distortion. *Filtering* is the selection of cases according to whether variables are within specified ranges. *Selective aggregation* creates new cases that are aggregated from other cases. *Micro-macro readings,* defined by Tufte (1983), are graphics in which "detail cumulates into larger coherent structures". The overall set of items provides a macro environment against which the micro reading of individual highlighted items can be interpreted. In *highlighting* individual items are made visually distinctive. *Distortion* means relative changes in the number of pixels in the space, such as change in size of objects, change in size due to perspective, or change in size of the space. Distortion techniques magnify regions of interest and shrink irrelevant regions (Card, Mackinlay and Shneiderman, 1999).

Figure 30 shows a Table Lens visualization, which is an example of distortion with tabular information (Rao and Card, 1994).

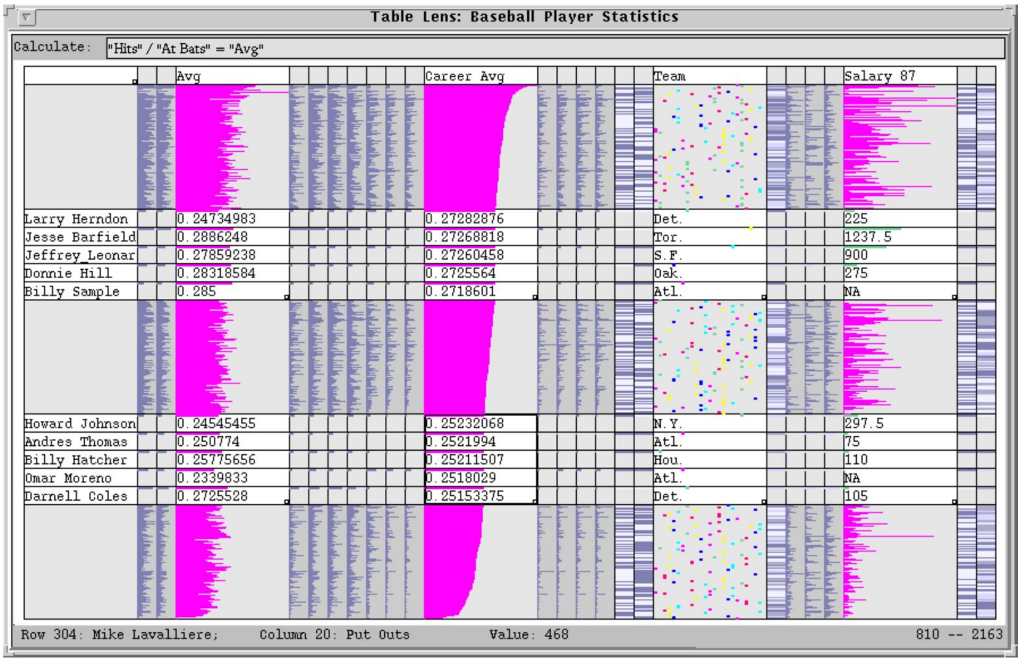


Table Lens: Baseball Player Statistics

Calculate: "Hits" / "At Bats" = "Avg"

	Avg	Career Avg	Team	Salary 87
Larry Herndon	0.24734983	0.27282876	Det.	225
Jesse Barfield	0.2886248	0.27268818	Tor.	1237.5
Jeffrey_Leonar	0.27859238	0.27260458	S.F.	900
Donnie Hill	0.28318584	0.2725564	Oak.	275
Billy Sample	0.285	0.2718601	Atl.	NA
Howard Johnson	0.24545455	0.25232068	N.Y.	297.5
Andres Thomas	0.250774	0.2521994	Atl.	75
Billy Hatcher	0.25775656	0.25211507	Hou.	110
Omar Moreno	0.2339833	0.2518029	Atl.	NA
Darnell Coles	0.2725528	0.25153375	Det.	105

Row 304: Mike Lavalliere; Column 20: Put Outs Value: 468 810 -- 2163

Figure 30. Table lens (Rao and Card, 1994)

Examples of visualization where parts of the structure are hidden until needed include Furnas' fisheye view and Cone Tree (Figure 31). Rapid zooming techniques enable users to zoom in and out of regions of interest (Pad++), while multiple windows enable the user to view an overview as well as other content in different windows.

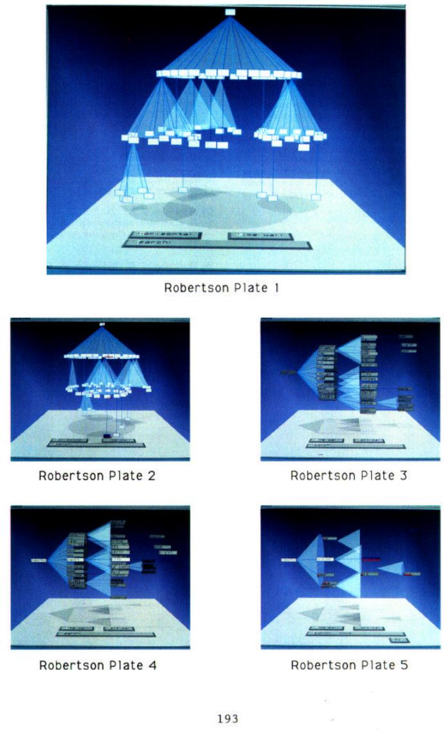


Figure 31. Cone Tree (Robertson, Mackinlay and Card, 1991)

Dense layouts and *pixel-oriented visualization* techniques are other approaches to visualizing large data sets (Keim, 1996). In these, each attribute value is represented by one pixel. The techniques can be divided into query-independent techniques that directly visualize the data or a certain portion of it, and query-dependent techniques that visualize the data in the context of a specific query. Figure 32 represents the principle of pixel techniques.

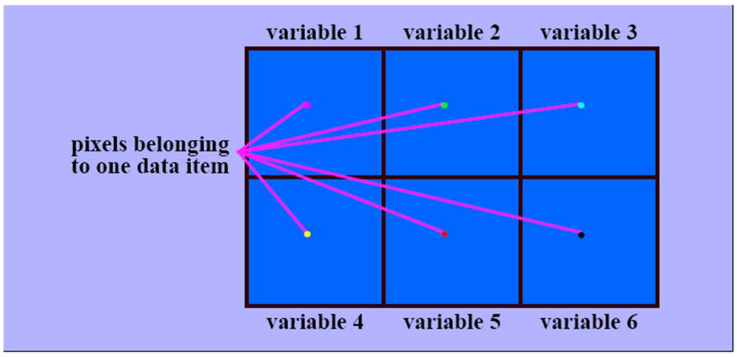


Figure 32. Pixel-oriented visualization (Keim, 1996)

3.9 Infrastructure

Infrastructure glues the visual analytics elements together as a working software environment. There is no standard definition of software infrastructure. One definition is given by CIO.gov[19] (in Galorath, 2008)

"IT infrastructure consists of the equipment, systems, software, and services used in common across an organization, regardless of mission/program/project. IT Infrastructure also serves as the foundation upon which mission/program/project-specific systems and capabilities are built."

In software engineering, infrastructures include system architecture, services, component libraries and interfaces. Infrastructure is expressed with the help of an architectural model that defines the decomposition of components and interfaces for different kind of services.

The need for infrastructure is recognized and discussed by the visual analytics community. Currently, each visual analytics system build their own infrastructure. Services that a visual analytics infrastructure could provide include launching and blending different kinds of analysis and visualizations, rendering visualization in different environments (2D, 3D), data pre-processing, data transformation, data management, scalability, interpreting semantics, support for collaboration, storing, and dissemination and communication of results (Keim et al, 2010).

Many of these tasks are similar in each application. With the help of a common infrastructure more or less of the tasks could be implemented as components and be used through open application interfaces (APIs). Nevertheless, there will be application-specific features to implement.

Construction of an infrastructure requires distinguishing the concepts and processes that are required in the visual analytics environment. Listings and taxonomies of reasoning tasks and operations required in visualization are part of this work. Developing approaches to distinguish process phases and data processing requirements during reasoning are also steps towards building infrastructure.

Suggestions for infrastructure or infrastructure elements have been made both in information visualization and visual analytics. Chi et al (2002) have developed a model for visualization infrastructure that includes an information visualization pipeline and an operator model for visualization. It divides the processing into three phases: (1) *data transformation*, (2) *visualization transformation* and (3) *visual mapping transformation*. Data transformation turns raw data into mathematical form; visualization transformation establishes a visual-spatial model of the data. Visual mapping transformation determines the appearance of the visual-spatial model to the user. The authors list the elementary operations required throughout the model. These are classified as: Data transformation, Visualization transformation, and Visual Mapping transformation. Visualizations sharing similar operations can be standardized, taking advantage of their similarities, and can be re-used in many problem domains, in constructing modular systems, and to provide ready-made components for the information visualization community.

[19] the US Chief Information Officers Council (CIO.gov) http://www.cio.gov/ (accessed 18 October 2012)

Bull (2008) has used methods of model-driven software engineering for generating information visualizations. Model-driven engineering uses a series of models to specify, design, implement and deploy software. Bull has provided a collection of formal view models for common information visualization techniques, a method for designing and customising information visualization, and a code generation technique. These are installed as an open source visualization toolkit. Bull has also provided a collection of platform-independent UML models for visualization, including a bar chart, node link diagram and treemap.

Another set of software patterns for visualization is by Heer and Agrawala (2006). They present twelve design patterns for information visualization software as well as a reference model and UML models for elements of information visualization software. The models include Data column, Cascaded Table, Relational graph, Proxy tuple, Expression, Scheduler, Operator, Renderer, Production rule, Camera and Dynamic query binding. The patterns form a network of interactions between patterns.

A visual analytics infrastructure is also introduced by Garg et al (2008). Their approach is based on logic programming, and they encode the relations as rules and facts on which the computing is based. Using this system, the analyst is able to construct models of arbitrary relationships in the data, explore the data for scenarios that fit the model, refine the model and query the model. The framework components are knowledge base, ILP (Inductive Logic Programming) system, and visualization system.

Brennan et al (2006) describe an architecture for a visual analytics framework for collaboration among multiple analysts. The framework supports sharing, translating between and fusing representations while keeping track of the source information. The perspectives support a variety of visualization techniques and representational styles. The framework consists of: (1) a common knowledge base capturing the factual aspects of the application; (2) a knowledge base local to an analyst containing cached facts from the common knowledge-base and additional or modified facts generated during analysis; and (3) inference engines for deriving conclusions based on the factual information present in the local knowledge base.

4 Evaluation of visual analytics

The purpose of evaluation in visual analytics is to confirm whether a tool has a positive impact, worth and significance (Plaisant, Grinstein and Scholtz, 2009). At present, there are no agreed methods and guidelines for how visual analytics should be evaluated. The prevailing view, however, is that traditional methods and metrics are not sufficient and novel approaches are required.

This chapter briefly reviews and introduces the traditional methods and approaches of visual analytics.

4.1 Traditional techniques

Human computer interaction (HCI) studies have developed several methods, techniques and variants of these for usability evaluation. The central methods are heuristic analysis, cognitive walkthrough and user observation. In heuristic analysis, an expert assesses how well the design is in line with user interface guidelines, principles and rules of thumb. Cognitive walkthrough is a task-specific usability inspection method in which users accomplish predefined tasks using the thinking aloud approach. The tasks are performed by a test user representing the intended target group. The method aims to find answers to questions such as: "Will the user know what to do", "Will the user try to achieve the right effect" or "Will the user notice that the correct action is available". In user observation, an evaluator observes how users actually use the product in their natural environment (Benyon, Turner and Turner, 2005).

Evaluation methods measure specific features of the target system. Several measures available, both quantitative and qualitative. The choice of metrics depends on the testing objectives, such as overall usability or learnability. The common usability metrics according to the ISO/IEC 25062 (2006) usability standard are effectiveness, efficiency and satisfaction. Effectiveness can be measured, for example, according to percentage of task completed successfully. Efficiency is measured according to the time taken to complete a task. Qualitative metrics are used to measure behaviour. Physical and physiological measures can also be used. Eye-movement tracking can map the user's focus of attention on screen. Emotions, anxiety and pleasure can be gauged using physiological measures, such as changes in heart rate or rate of perspiration (Benyon, Turner and Turner, 2005)

.

4.2 Evaluation in visual analytics

The visual analytics community has argued that the present methods do not fit well to visual analytics evaluation. The call for new methods is based on several grounds. In information visualization, evaluation methods have traditionally consisted of predefined benchmarking tests under controlled conditions. These are typically low-level tasks such as: "Find minimum and maximum value from a data set with the help of visualizations". Predefined tests are viewed as problematic because visual analytics is by nature exploratory, and the set of tasks that users want to perform may not be known beforehand (North, 2006). Predefined tests also leave little room for the unexpected. The duration of analytic studies, which can last for days, has also caused concern. Furthermore, controlled studies may not effectively represent real situations and are difficult to arrange. Predefined, premature completion times also leave little room for insight. One key problem is that traditional metrics are not well suited to visual analytics (Plaisant, Grinstein and Scholtz, 2009). Measures such as performance and accuracy do not necessarily measure the goal of visual analytics – insight. The purpose of evaluation should be to determine to what degree visualizations achieve this goal (North, 2006).

The suggested approaches can be categorized in two groups. The first is "insight-based" methods, which are based on the notion that if the purpose of visual analytics is insight then the purpose of evaluation should be to determine to what degree the tools achieve this purpose (North, 2006). The metrics would thus be user-reported insights. The other approach incorporates evaluation into the design process, giving a set of appropriate evaluation methods for each process phase. These latter methods include both traditional evaluation methods and insight-based methods.

Insight-based methods

The use of insight-based methods raises several questions, such as (North, 2006): What is the agreed definition of insight? What kinds of methods and metrics adequately represent insight? Are controlled experiments useful? Should predefined tasks be used? What kind of participants should be used? North introduces an insight-based approach and suggests using controlled experiments both with and without predefined tasks. If task are used, they should be more complex than traditional benchmarking tasks, such as estimation tasks, or task that involve uncertainty. Users should be directed to interpret the visualizations in articulate written answers so their insights can be captured. Another alternative is to eliminate tasks entirely and study what insights the users gain on their own, letting the user explore the data in the way that they choose. In this approach users are initially oriented with the help of starting questions, but then left to go on to freely explore the data and report their insights as and when they find them. The users verbalize their findings in a think-aloud protocol, enabling the evaluator to capture their insights. Findings are marked as insight occurrences which can be quantified based on different metrics: complexity, time to generate, errors, insight depth category and so on. During the evaluation, other kinds of usability data can be also collected, such as which features helped to gain insight, and what caused problems for users. According to North, both types of methods are needed.

Smuc et al (2009) refines the insight-based method suggested by North by introducing a three-level method: (1) Counting insights (as suggested by North). The user analyses data in a controlled think-aloud test setting. During the process, a moderator measures certain variables, such as number of insights; time taken for insight generation, and correctness. (2) Predefined insight categories are defined, related to the tool's intended purpose. If an insight type is missing from the insights that users find, redesign is required. (3) Relational insight organizer – a graph used to study and compare the insight generation of participants. It visualizes the prior knowledge of users and their insight generation process. The tool can be used to easily compare the insight generation of users with a priori knowledge. The authors conclude that no superior method exists, and that the method must be chosen according to the evaluation goals, time available, and the research question. In addition to domain experts, they recommend using "semi experts" in the evaluations if real experts are difficult to obtain.

Design process methods

T. Munzner (2009) introduces a “nested model” in which evaluation is integrated into the design process. The design process is divided into nested phases, and distinct evaluation methods are suggested for each level. The levels are: (1) domain problem and data characterization; (2) operations and data type abstraction; (3) visual encodings and interaction design; and (4) algorithm design. Munzner defines both the threats and validation approaches for each level. The validation methods are: (1) observe and interview target users; (2) test on target users; (3) field study of use to justify designs, qualitative/quantitative results analysis; (4) lab study, measure system time/memory usage.

Perer and Shneiderman (2009) combine the design process and insight-based evaluation. Their methodology consists of five phases, supported by development of the tool: (1) interview, in order to understand the intentions of the domain experts; (2) training; (3) early use (2-4 weeks); (4) mature use (2-4 weeks); (5) outcome. The system is installed for use during the development process, and observers then visit and interview the user about their recall of insights. Updates of the software are then carried out based on the results. Later, domain experts explain the impact of the tool to their work. The authors have performed four case studies. They claim that the method led to new insights and discoveries that might not have been found using traditional usability experiments. The drawback is that the methods require a lengthy period.

Levels of evaluation

According to Scholtz (2006), evaluation should not focus only on the visualizations or the usability of the user interface, but should address a range of aspects. He suggests aspects such as situation awareness, collaboration, interaction, creativity, and utility need. Other suggested aspects by Scholtz are derived from the standard ISO 9241:10 “Dialogue principles” (ISO 9241-110, 2006): suitability for the task, self-descriptiveness of the action, controllability, conformity with user expectations of consistency, error tolerance, suitability for individualization of customization, and suitability to learning.

The Visual Analytics Agenda suggests that no particular technique will be suitable for all problems and evaluation must therefore be targeted at different levels and different methods and metrics applied to each level. The agenda suggests three levels of evaluation: component, system and work environment. Components that do not involve interaction can be evaluated using methods that measure performance, accuracy, or identification of limitations. If user interaction is included, empirical user evaluation can be used, with standard metrics of effectiveness, efficiency and user satisfaction. The system level evaluates the target uses. It can be performed in lab conditions, and its objectives are learnability and user satisfaction. The work environment level measures technology adoption, adoption rate, trust and productivity. Case studies and ethnographic studies are recommended at this level.

The recommendations given in the report by Keim et al (2010) call for more research. New evaluation methods and techniques, standards, guidelines, repositories containing datasets for benchmarking and showcases are required.

5 Data modelling

A data model is an abstract model that defines the concepts and their relationships of a subset of real-world information. Data modelling is a widely used technique in software engineering. Data models have been used since the 1990s when the use of relational databases became widespread. The models have several uses among information systems. This chapter introduces data modelling and how it has been used in visual analytics and information visualization.

The key concepts of a data model are *business objects*, which represent the concepts of interest, *relations between objects*, and *attributes*, which are properties of the model objects. In addition, there are *classification objects* for categorising the business objects, such as code values or taxonomies. The relations have cardinalities. There can be one-to-one, one-to-many and many-to-many relationships. If a relationship contains additional information, it can also be a model object. *Model instances* represent real data items that are structured according to the model.

The model structure is defined in a *model schema*. A data model schema is often expressed in visual form as an entity relation diagram, which is a form of node-link diagram. There are several styles of model schema, with differing notations of nodes, links and relation cardinalities. Figure 33 shows an example data model expressed in UML[20] (Unified Modelling Language™). It defines a cookery book, including categorized recipes and wine recommendations.

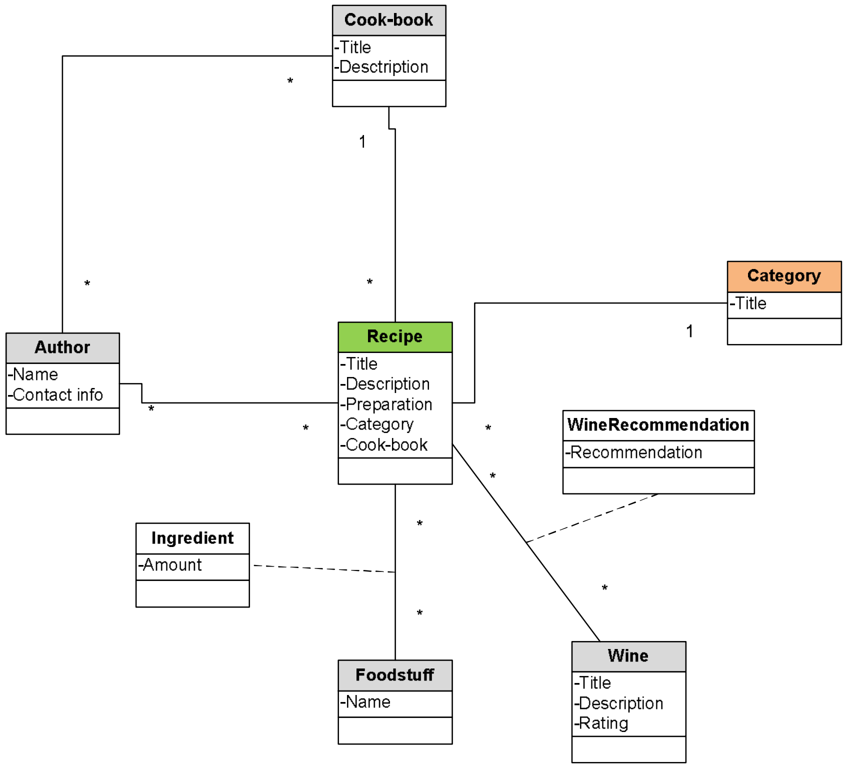


Figure 33. UML model.

[20] http://www.uml.org/ (accessed 18 October 2012)

Recipe is the key business object (green rectangle). Other business objects are *Cook-book, Author, Foodstuff* and *Wine (*grey rectangles*)*. *Category* is a classification object that categorises the recipes as starters, main dishes and desserts (orange rectangle). The relationships between objects are expressed by links between the objects, and the cardinalities by numbers and asterisks (* = many) at the end of the links. The white rectangles are objects that contain information on relationships. The relationships are:

- *Cook-book–Recipe* (one-to-many): collects the individual recipes of the cook-book.
- *Recipe–Foodstuff* (many-to-many): lists the foodstuff used in a recipe. The relationship is many-to-many, because the same foodstuff can be used in several recipes. The relationship has an object (*Ingredient*) that tells the amount of the foodstuff needed in the recipe.
- *Recipe–Wine* (many-to-many): tells which wines are recommended with the food. The relationship is many-to-many because the same wine can be recommended with many recipes. The relationship object WineRecommendation contains special information about serving a wine with a specific food.
- *Recipe–Author* (many-to-many): tells the authors of the recipe. Many-to-many relationship means that one recipe can have several authors and one author can have written several recipes.
- *Recipe–Category* (one-to-many): collects the recipes in different categories. In this case, one recipe can belong only to one category.
- *Cook-Book–Author* (many-to-many): names the editor of the whole cook-book. The relation is many-to-many because an editor can have edited several cook-books and one cook-book can have several editors.

Example instances from the model are shown below:

Cook-book

Title	Description	Authors (from the relationship)
100 Cakes	100 delicious cakes for everybody and everyday	Paula

Author

Name	Contact info	Cook-Books	Recipes(from the relationship)
Paula	Meet only at home	100 Cakes	Sugar cake
Sanni	tel…	-	Chocolate cake

Foodstuff

Name
Egg
Sugar
Flour
Chocolate
Almond

Wine

Title	Description	Rating
Cake wine 1	Smooth, dark…	3
Cake wine 2	Yellow, sparkling…	4

Category

Title
Chocolate cakes
Other cakes

Recipe

Title	Description	Preparation	Category	Cook-book	Authors
Chocolate cake	A delicious cake	Mix all and bake	Chocolate cakes	100 cakes	Sanni
Sugar cake	A quick …	Mix eggs and sugar…	Other cakes	100 cakes	Paula

Ingredient

Recipe	Foodstuff	Amount
Chocolate cake	Almond	100 g
	Egg	3
	Chocolate	50 g
Sugar cake	Sugar	150 g
	Egg	2
	Flour	100 g

Wine recommendation

Recipe	Wine	Recommendation
Chocolate cake	Cake wine 1	Serve ice cold
Sugar cake	Cake wine 2	Serve room temp

Data models are usually constructed with a specific application in mind and define only the data needed in that application. For example, the previous example omits data that could relate to the domain: what other books the authors have written, the environmental and nutritional details of the foodstuff, or the grapes the wine is made of.

Why useful

The data model serves as a template for the whole application. Often several levels of models are constructed varying in purposes of use and detail: one for defining the software structure (object model), another for constructing the database (database model) and one for data exchange. Data models are also useful in communication and documentation. The visual model provides a quick overview of the data. It can be used in discussions between IT developers and business people. People with different backgrounds and levels of experience can easily understand the conceptual structure from the model (Simsion and Witt, 2005).

Conceptual data modelling is applicable only to structured data, not for unstructured data such as images, video or textual content.

History and development

Data modelling originated in the 1970s when P.P.S. Chen published his articles on entity relationships modelling (Chen 1975, 1977). He introduced the basic concepts: entities, attributes and relations, and a graphical presentation "Entity-relationship diagram". The model was intended for construction of databases.

Modelling came into use along with the development of the relational model and databases, which began in the beginning of 1970 after Ted Codd published his article "A relational model for shared data bank" (Codd, 1970). Entity relationship modelling lent itself well to planning relational databases. As the popularity of relational databases grew towards the end of the 1980s, several other styles of modelling also emerged, along with data modelling languages and software tools supporting them. CASE tools (Computer Aided Software Engineering) generated schemas for relational databases and application code directly from the entity–relationship model. Examples of commercial tools were AD/Cycle by IBM, based on structured analysis and design methodologies, SA/SD by Yourdon de Marco (DeMarco, 1979), and tools supporting "information engineering" methods by James Martin (Martin and Finkelstein, 1981), including IEF (Information Engineering Facility) by Texas instruments, and IEW (Information engineering Workbench).

The emergence of object-oriented modelling in the early 1990s expanded the use of data models towards software construction. Traditionally, data and procedures had been separate, but object orientation combined data objects and their processing methods. Object modelling defines the data objects, their relationships, object attributes and the processing methods for each data object. Object modelling is nowadays a widely applied technique. The most popular method is UML, Unified Data Modelling (Booch, Rumbaugh and Jacobson, 2005). In addition to data models, it provides many other techniques for modelling business and information systems, including functional models. Tools supporting these methodologies also exist, such as

Microsoft Visio[21] and Rational Rose by IBM[22]. In this work, the UML data models are constructed using Microsoft Visio.

Along with the Semantic Web in 2000, semantic models and ontologies have also become popular. Ontologies are often mentioned in the context of data models. They represent another kind of data modelling. Ontologies provide a generic view of data, referring to a shared and common understanding about a specific domain (Ding and Foo, 2002). Ontologies can define both the concepts and the instances of the domain. The difference between ontologies and data models is defined by Spyns et al (Spyns, Meersman and Jarrar, 2002) as follows:

"Unlike data models, the fundamental asset of ontologies is their relative independence of particular applications, i.e. an ontology consists of relatively generic knowledge that can be reused by different kinds of applications/tasks."

Ontologies are often specified in ontology specification languages such as OWL (Web Ontology Language)[23] and RDF (Resource Description Framework)[24] defined by W3C, although ER modelling techniques can be applied as well. The practice of using the term ontology is not established. Often conceptual data models defined with ontology definition languages are called ontologies, while graphs showing instance data ontologies are called semantic graphs.

Generic and standard models

Data modelling is time-demanding work. It requires co-operation between domain experts and model designers to extract and define the concepts and their relationships. The notion of reusing models has also gained ground. Generic and standard models for a variety of domains are published both by standards bodies and business or industrial consortiums. Examples of standard models are ISO 14926[25] (Integration of life-cycle data for process plants including oil and gas production facilities) for data exchange, and the ISO Standard for the Exchange of Product Model Data STEP (ISO 10303)[26], using EXPRESS modelling language. MIMOSA[27] is an example of a standard model by an industry trade association, providing open information standards for operations and maintenance (Mathew et al, 2006).

In addition, the research community suggests standard models. One example is the data model for monitoring data by D. Inaudi (2002), which is proposed as an open standard aiming at the archival of long-term monitoring data. The model is also intended for data acquisition, data analysis and representation. Another example is the model by Manolescu et al (2009), which is proposed as a model for continuously updated data for data analysis.

[21] http://www.microsoft.com (accessed 18 October 2012)

[22] http://www-01.ibm.com/software/awdtools/developer/rose/ (accessed 18 October 2012)

[23] http://www.w3.org/TR/owl-features/ (accessed 18 October 2012)

[24] http://www.w3.org/TR/#tr_RDF (accessed 18 October 2012)

[25] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29556 (accessed 18 October 2012)

[26] http://www.steptools.com/library/standard/ (accessed 18 October 2012)

[27] http://www.mimosa.org/ (accessed 18 October 2012)

5.1 Data models in information visualization

Data models have been used in information visualization for mapping data and visualizations together. Three approaches are used. The first approach focuses on how to generate visualizations from the relational model. The second group of methods generate visualizations by means of manual mapping. The third group encompasses automatic methods for mapping.

Early approaches

The earliest uses of data models with visualizations presented the contents of relational databases. Catarci et al (1997) thoroughly explore and classify the early attempts from the beginning of the 1980s to the mid-1990s. A total of 78 systems were evaluated. The objectives of these pioneering systems were the same as in visual analytics today: visual representations that are effective at expressing different kinds of knowledge and that can be used to understand database content, to focus on meaningful items, and to find query patterns and reason from the query results. The visual representations used in them are classified by Catarci et al as form-based (tables), diagram-based, icon-based and hybrid. The data model is used as a starting point for browsing data, either for top-down refinements, selective or hierarchical zooms or removing irrelevant objects. Another use is to form visual queries with the help of a data model, thus building visual query systems for databases. These early systems produce only static visualizations, without any intention for deeper analysis.

Manual mappings

The "Snap" system by North and Shneiderman (2000) uses a relational model in the creation of coordinated visualizations in which basic relational concepts are modelled and mapped to visualization concepts. A model object corresponds to a whole visualization, and an instance (tuple) to an item of the visualization. Each visualization displays a single model object. The relations between the objects are used to create the coordinated visualization. One-to-one relationship is used in brushing-and-linking, showing overview, details-on-demand and synchronized scrolling. One-to-many relationship is used for drilling down to details. Snap is intended for an open application interface that can be added as a visualization tool for researchers and developers.

A related approach is Rivet (Bosch et al, 2000), a visualization system for computer system visualizations, although the scope is wider than in Snap. It supports the development of the visualizations themselves. Here, the basic data element is a model object and instance (tuple). Visual primitives, metaphors, are used to define the drawing attributes of a tuple and for mapping the fields of a tuple to a location, colour, fill pattern and size. Complex visualizations can be composed with the help of these primitives. There are also mechanisms to support coordination of multiple views and brushing.

Another similar use of data models is presented by Falconer et al (2009) for creating visualizations of ontology instances. The user can define ontology visualizations with the help of the tool, which is used to map ontology concepts to visualization concepts, for example to show a concept as a node and another as a link in a node link diagram.

Both the ontology concepts and visualization concepts are defined as data models. Example models exist for node-link diagram, list, line chart, pie chart, map, heat map, table, parallel coordinates and tree visualizations. The work by Falconer et al is based on the PhD thesis by Bull (2008) on generating and customizing visualizations with the help of Model Driven Engineering, which proposes a model-based approach for creating visualizations and viewers for visualizations.

Automatic mapping

Automatic methods are a popular choice for web visualizations. Unstructured data is the starting point, and visualizations are generated with the help of a model generated from the data.

Cammarano et al (2007) use an RDF-like data model in the creation of visualizations automatically from heterogeneous web data. The user can make a query and select a desired visualization, such as a map, timeline of scatterplot. The visualizations have specific requirements and the system uses schema matching to couple the underlying heterogeneous data and the specification of visualization.

A related approach derives from Gilson et al (2008). They use three ontology models to create visualizations automatically from web data. External visualization toolkits can be used to render the visualizations. The starting point is tabular data that the user has extracted from the web, and the intended visualization selected by the user. Source data ontology is generated from the user data with a schema mapping process using similarity measures for creation of the ontology concepts. The next step is mapping from the source data ontology to the visual representation ontology. This is done with the help of a semantic bridging ontology that stores information about visualization tools, styles and parameters. Finally, the visualization toolkit is invoked to show the visualization.

5.2 Data models in visual analytics

The use of data models in visual analytics is not widespread. Ten visual analytics approaches were found that claim to use data models. The approaches vary in terms of the kinds of model used and the ways that they are utilized. The majority use data models to extract concepts from original data and create an ontology of them. The ontology instances can then be analysed with the help of different kinds of visualizations, typically a node-link diagram. A couple of approaches use a data model as part of the system architecture and generate applications using of a set of models. These models are either fixed application models, general-purpose models or models generated from the concepts existing in the data. A summary of the studied approaches is represented in Table 2, which shows the main purpose, main use and type of model. The approaches are categorized as: (1) ontology networks; (2) fixed application models; (3) general-purpose domain models; and (4) general-purpose models. The approaches are briefly introduced thereafter.

Table 2. Data models in visual analytics.

Approach	**Purpose**	**Use of data model**	**Data model type**
1) Ontology networks			
(Liu, Navathe and Stasko 2011)	Framework for network presentation	Network generation and presentation	Model generated from data tables
(Alani et al 2005)	Community analysis	Network presentation	Existing ontology model
(Shen, Ma and Eliassi-Rad, 2006)	Social networks	Network presentation, guiding analysis	Either existing ontology or derived from network data
"Harvest" (Gotz, Zhou and Aggarwal, 2006)	Visual investigation	Adding new concepts to ontology, analysis and presentation	Existing ontology
2) Fixed application models			
"ER-explorer" (Dai et al, 2008)	Text analysis	Discovering entities from text, presentation	Fixed application ER-model
"Jigsaw" (Görg, Spence and Stasko, 2008)	Investigative analysis	Discovering Entities form text	Fixed application model ER-model
3) General-purpose domain models			
(Manolescu et al, 2009)	Framework for visual analytics for dynamic data, example from dataflow of publications and projects	Database accumulation and queries, presentation with an external toolkit	ER-model for dynamic data
Inaudi (2002)	Monitoring structures, especially bridge monitoring	Storing data, presentation and analysis with an external toolkit	ER-model of monitoring with sensors
4) General-purpose models			
NetLens (Kang et al, 2007)	Digital libraries analysis	Application generation, analysis and presentation	Two levels of ER-models: General model and application model
Streit et al, (2012)	General-purpose, example from biomedical data	Tool setup, analysis process guidance, data retrieval and presentation.	Three levels of ER-models: setup model, domain model, analysis session model

Ontology networks

In “Network-based visual analytics” (Liu, Navathe and Stasko, 2011), a network representation is created from relational tabular data and the network can be manipulated in numerous ways: assigning weights to edges, doing projections, grouping nodes, dividing networks into sub-networks. Individual nodes have interactions for selecting, filtering, moving, hiding, showing and expanding. Several network layout algorithms and analytical measures are available. The user can modify the network by dragging and dropping model attributes from a schema view and making connections between them. The purpose is to provide a general approach for multi-dimensional and multi-level visual analytics.

A related approach is introduced by Alani et al (2005) for community analysis. Here, the starting point is an ontology model and knowledge base. The conceptual model – ontology model – is used for searching connections between instances of the model objects. The user can select relationships of interest from the model and the system finds corresponding instances and shows them as lists of instances.

Gotz et al (Gotz, Zhou and Aggarwal, 2006) introduce Harvest, a tool with which users can add new concepts to ontologies and relate concepts together. Tools for visual exploration of the original and new synthesized concepts are available. The concepts can represent instances of data and model concepts. The data can be analysed with the help of graph-like structures and timelines.

Shen et al (Shen, Ma and Eliassi-Rad, 2006) describe a tool for examining social networks, expressed as semantic graphs representing terrorist connections. It uses an ontology graph to guide the analysis. The ontology graph is a variant of the entity-relation model, describing the relationships of actors in the semantic graph. The user can select entity attributes of the diagram and the whole network is visualized in light of the selected entity attribute, showing the connections to other instances of another entity. The authors term the method semantic abstraction. Subsets of entity values can be selected to show connections between selected groups. The graph shows the number of frequencies of the links as numbers of edges. The node sizes represent the disparity of connected entity types for each node type.

Fixed application models

ER-explorer by Dai et al (2008) supports the discovery of text data with the help of an entity-relationship model. It uses a predefined model defining 24 entity types, including people, geo-political entities, date and others. Their algorithm extracts the data and shows the entity instances in text cube view and a network view.

Görg et al (Görg, Spence and Stasko, 2008) have a related fixed-model approach. They introduce Jigsaw, a tool for investigative analysis. It derives instances of entity types from textual data and makes connections between them. Four predefined entity types are used: person, place, data and organization. The results can be viewed in different ways: as a tabular connection view, a semantic graph view, a scatterplot view and a text view showing the original data. The views are interactive and users can examine the relationships between the entities in multiple ways.

General-purpose domain models

ReaViz (Manolescu et al, 2009) is a visual analytics application for dynamic, continuously updated data. It is built on a data model that comprises three kinds of entities: Application-dependent entities that model the data used by the specific visualization/analysis applications, workflow-related entities that capture the definition and instances of workflows, and visualization-related entities that capture the information items required by the data visualization modules. An application prototype demonstrates the flow of publications and projects.

Inaudi (2002) suggests using a relational model for handling data flow and introduces a data model for monitoring structures with sensors. The analysis and visualization is performed with the help of a package that allows exploration of the database. It is designed especially for analysis of curvature measurements of bridge monitoring projects using special bridge analysis tools.

General-purpose models

NetLens (Kang et al, 2007) is an approach to analyse textual data from digital libraries with the help of an entity–relation model. Their motivation is to avoid network overviews of whole data, as such networks easily grow too large and become cluttered. They use an abstract entity–relation model, called a "content–action" data model. The model has two entity types and includes relations between the entities and within an entity. An application schema is constructed based on the content–actor model and the data is stored according to the application model. A general-purpose user interface is constructed on the abstract model, which can be applied to any application data set built using the abstract model. The models have a viewer and different tools to explore the data, including histograms and filters. An example model is applied to express papers and people, and their attributes.

The work of Streit et al (2012) is a general-purpose approach with the objective of supporting user orientation in data and guiding the analysis with the help of data models. It uses three kinds of models: a setup model, model of the domain, and a model of the analysis session. The set up model defines the data sources and ways to interact with it, both for data retrieval and visualization. The analysis session model defines the tasks and analysis steps for different analyses. The domain model maps the analysis tasks and the set up model elements together. Building the system requires a tedious authoring phase in which the data sources and their interfaces and analysis task are modelled and mapped together, and resembles a normal software construction process of defining user tasks and use cases. Analysis support is implemented with predefined analysis processes, although the user is not obliged to use them.

6 Monitoring data

Nearly all branches of human life generate monitoring data. Machinery, industrial processes and structures such as buildings and bridges are equipped with sensors that gather data on the state of the monitored object. Natural and environmental phenomena are also monitored with sensors. Finance, business and public authorities also collect vast amounts of data. Telecommunication software and other computer software have built-in data collection methods. People are also monitored in numerous ways, such as through medical records and the tracking of behaviour and consuming habits.

Collecting monitoring data is not a new phenomenon, but advances in sensor and data processing technology have extensively broadened its reach. Sensors are readily available and easy to install and use. The majority of sensor data is collected for useful purposes, but also merely as it is possible. The collected data is generally expected to contain important knowledge. However, the speed at which data can be accumulated far outpaces our ability to understand the data we collect.

Essentially, monitoring data consists of value–timestamp pairs that form a time series. The measurement values depend on the monitored object and can represent anything from temperature, vibration or position to opinions. The measurement frequencies also vary. Measurements can be of ongoing real-time data or history data providing a snapshot of the past. The focus in this work is on history data. Stream data manipulation adds complexity to data processing. How the principles presented in this work could be extended to stream data is a subject for further research.

This chapter is written from the point of view of industrial monitoring, but also discusses other areas. Industrial monitoring has a long tradition and its concepts and procedures are well defined. The chapter introduces the area, indicators, monitoring process, and analysis and presentation methods. A survey on the use of visual analytics in monitoring was also conducted as part of this work.

What is monitoring?

Industrial monitoring focuses on process performance, on ensuring that the process remains stable, within its predefined limits, and operates at its full potential, also after process modifications and improvements. Monitoring is related to quality control and looks for areas that may conceal sources of variation in quality. Monitoring tasks include following the monitoring objects, recognizing new and unknown phenomena, and reacting based on the observations. Methods of statistical process control (SPC) are applied to monitoring and controlling a process (Fassò and Pezzetti 2007).

Several standards exist in this area. MIMOSA (2012) provides open specifications for Enterprise Application Integration (EAI) and Condition-Based Maintenance (CBM). ISO 13374 is a standard for condition monitoring and diagnostics of machines (ISO 13374, 2007). IEEE 1451.1 defines a standard for Smart Transducer Interface for Sensors and Actuators (IEEE 1451.1, 2004). SEMI E10-0304 provides a specification for the Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM) (SEMI E10-0304).

Business performance is often monitored with the help of data warehouses. Data warehouses gather knowledge from various data sources from organizations in order to form a consolidated view of the business. They provide data views in the form of dashboards, “data marts”, and enterprise reporting systems. Data warehouses provide online analytical processing (OLAP), that is, tools for interactive analysis of multidimensional data of varied granularities. Data warehouses do not usually have links to the source systems and it is not possible to trace back to the original data. Data warehouses are heavy systems, tedious to build up and require specialized technical expertise (Azarm, Nargesian and Peyton, 2011a).

Key Performance Indicators

The performance of monitored objects is followed with the help of key performance indicators or indexes that express the health of the monitoring object. Examples of industrial indicators are energy performance indexes (Heilala et al, 2010a, 2010b). The finance and business world uses key performance indicators that reflect the critical success factors of an organization. Human and wellbeing studies calculate a variety of indexes such as the Body Mass Index or Burnout Index. A quick Google search identified a variety of indicators for all branches of human activity, from football to hotel websites. Performance indicators are usually long-term considerations. The definition of what they are and how they are measured are domain-specific and do not change often.

Monitoring process

The monitoring process is represented in both the MIMOSA and ISO 13374 standards by means of a level structure (Figure 34).

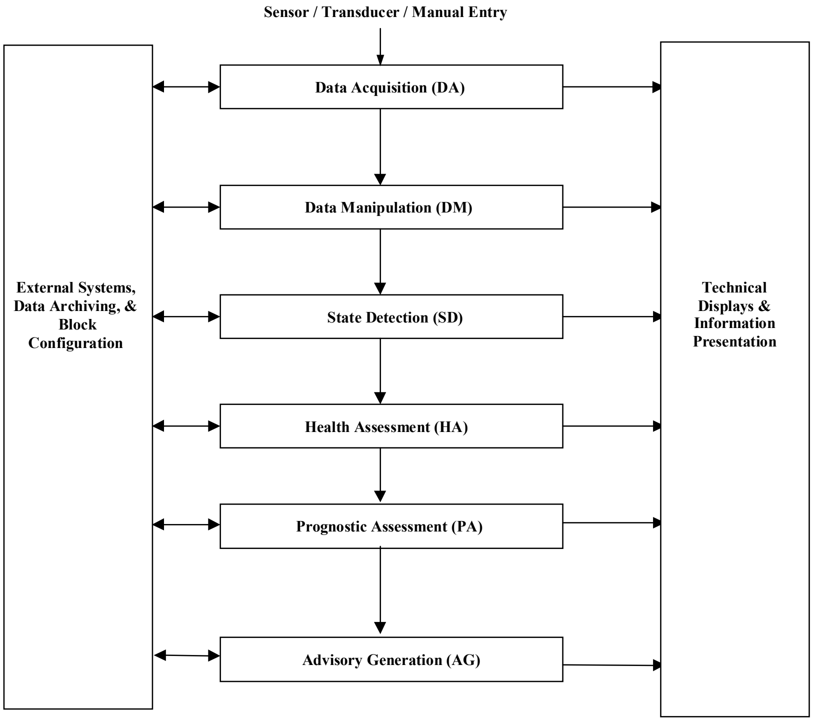

Sensor / Transducer / Manual Entry
Data Acquisition (DA)
Data Manipulation (DM)
State Detection (SD)
Health Assessment (HA)
Prognostic Assessment (PA)
Advisory Generation (AG)
External Systems, Data Archiving, & Block Configuration
Technical Displays & Information Presentation

Figure 34. Level structure (ISO TC 108/SC 5, 2005).

Functions of monitoring comprises the following:

a) Data Acquisition (DA) block: converts an output from the transducer to a digital parameter representing a physical quantity and related information (such as time, calibration, data quality, data collector utilized, and sensor configuration).

b) Data Manipulation (DM) block: performs signal analysis, computes meaningful descriptors, and derives virtual sensor readings from the raw measurements.

c) State Detection (SD) block: facilitates the creation and maintenance of normal baseline "profiles", searches for abnormalities whenever new data is acquired, and determines in which abnormality zone, if any, the data belongs (e.g., "alert" or "alarm").

The final three blocks normally attempt to combine monitoring technologies in order to assess the current health of the machine, predict future failures, and provide recommended action steps for operations and maintenance personnel. These three blocks and the functions they should support are as follows:

d) Health Assessment (HA) block: diagnoses any faults and rates the current health of the equipment or process, considering all state information.

e) Prognostic Assessment (PA) block: determines future health states and failure modes based on the current health assessment and projected usage loads on the equipment and/or process, as well as remaining useful life predictions.

f) Advisory Generation (AG) block: provides actionable information regarding maintenance or operational changes required to optimize the life of the process and/or equipment.

MMOSA presents a cyclic version of the model (Figure 35):

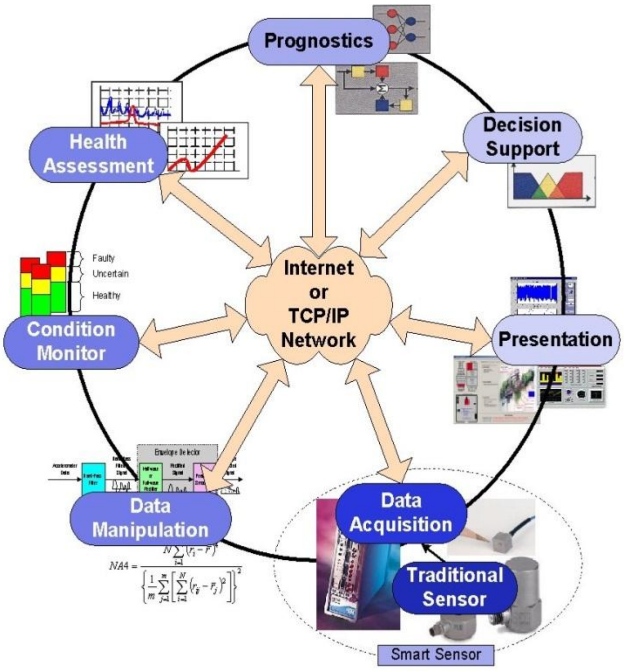


Figure 35. Cyclic monitoring model (ISO TC 108/SC 5, 2005).

Analysis methods

Monitoring data is analyzed with the help of statistical methods. Time series analysis methods are feasible, and detecting trends and patterns over time is central to the analysis. Monitoring data uses control charts, which are a form of line graph with limits for variation (Gravois, 2007). CUSUM (Cumulative Sums) is one key method (Chang, 2012). As measurements are taken, the difference between each measurement and the benchmark value is calculated, and this is cumulatively summed up. If the processes are under control, the measurements do not deviate significantly from the benchmark, so measurements greater than the benchmark and those less than the benchmark average each other out, and the CUSUM value should vary narrowly around the benchmark level. If the processes are out of control, measurements are more likely to be on one side of the benchmark, so the CUSUM value will progressively depart from the benchmark. Data mining methods are also used.

Presentation

In industry and business, process control data is traditionally presented by means of dashboards. Dashboards show the status and trends of the processes in numbers and using static visualization. Industrial dashboards are often called SCADA (supervisory control and data acquisition) systems. They are attached to the process control system, such as process automation or building automation system. In business, measurements are stored and analysed with general-purpose business intelligence tools. There is huge variety of tools on the market, including tools that enable users to create tailored dashboards. Application-specific tools are also available, as well as tools connected to enterprise resource planning (ERP) systems. General-purpose tools such as Excel and SAS are also used. Market research uses statistical packages such as SPSS, which provide basic statistics, graphs and plots, and tailored tools.

Visual analytics and monitoring data

The visual analytics community associated with monitoring data. Only Inaudi (2002) discusses visual analytics in connection with monitoring structures with sensors. Inaudi suggests using a relational model for handling data flow and introduces a data model for this purpose. The model is proposed as an open and free standard, as discussed earlier in Section 5.2, and the author also introduces a database structure (Figure 36). Analysis and visualization is performed with the help of a package that allows exploration of the database. The package is intended for analysis of curvature measurements in bridge monitoring using special bridge analysis tools.

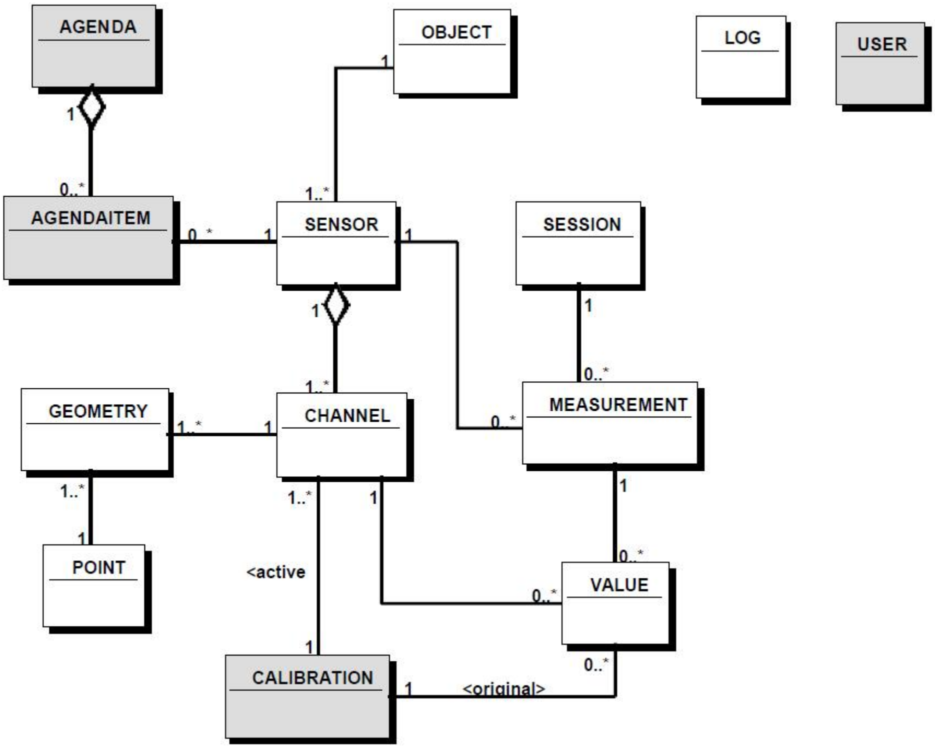


Figure 36. Data model for monitoring structures with sensors (Inaudi, 2002).

7 A data model based approach for visual analytics

This chapter introduces the suggested data model based approach to visual analytics. It presents the idea of a domain model and its uses in visual analytics. Domain means here a field or scope of knowledge or activity, such as monitoring of objects, marketing or bioinformatics, comprising a variety of applications. A domain data model is a general purpose data model for the whole domain. A domain model for monitoring data is presented and the approach compared to other data models and uses of data models in visual analytics. A tool concept based on the domain model is presented in the next chapter.

7.1 Domain data model

In the measurement data domain, data consists of similar concepts: *objects of interest*, *measurements*, *indexes* and *background information,* regardless of how the data is collected or stored. For example, in monitoring the energy efficiency of a preheating furnace for steel bars, the object of interest is the furnace, while the measurements include temperatures in different areas of the furnace, and gas consumption (Heilala et al, 2010a, 2010b), and the background information includes the surface area of the furnace. The indexes are energy use parameters (EUPs) – functions calculated from the data to identify the state of energy use. In indoor air quality monitoring the object of interest is a building space, and the measurements include temperature, CO2 and the number of occupants, while the background information comprises the volume and purpose of use of the space. An indoor air quality index is calculated to define weather the air quality is within accepted limits [28]. The same elements can also be distinguished in body weight management, a completely unrelated application area (Järvinen, 2007). The object of interest here is a person, the measurements are the food eaten, energy expenditure in exercise and daily weight, and the background information includes start weight, age, gender and height. The indexes are body mass index, food quality index and energy balance.

This gives an idea to construct a *domain specific data model* for visual analytics of monitoring data. The model would include the key concepts of monitoring: object of interest, measurements, background properties, and indexes. On the other hand, visualizations and analysis methods require the data to be in specific forms, regardless of the data domain. If the model takes into account the specific needs of visualizations and analysis methods, the model could be used as a template for visual analytics applications in the domain.

Three uses of the model are suggested: (1) for construction of visual analytics applications, (2) for supporting reasoning, and (3) as a user interface element.

Construction of visual analytics applications

The model could serve as an architectural model for a platform of visual analytics applications. The platform would contain a database constructed based on the model. The application data would be stored there in a "domain standard" form. Predefined visualization and analysis methods applicable to the data would be stored in a library

[28] https://www.rakennustieto.fi/kortistot/rt/kortit/10946

that could be accessed via standard interfaces. An application-independent user interface could then be constructed on top of the model and would include features that are useful in the monitoring domain. When a new monitoring application requires analysis, the data from the original data sources would be loaded to the database. The predefined analysis and visualization methods would be automatically available through the library interface. The user would be able to use the application with the help of the application-independent user interface. This is possible because the underlying domain model is similar regardless of the specific application. In addition, a tailored user interface could be built on top of the general-purpose interface. Figure 37 outlines the architecture.

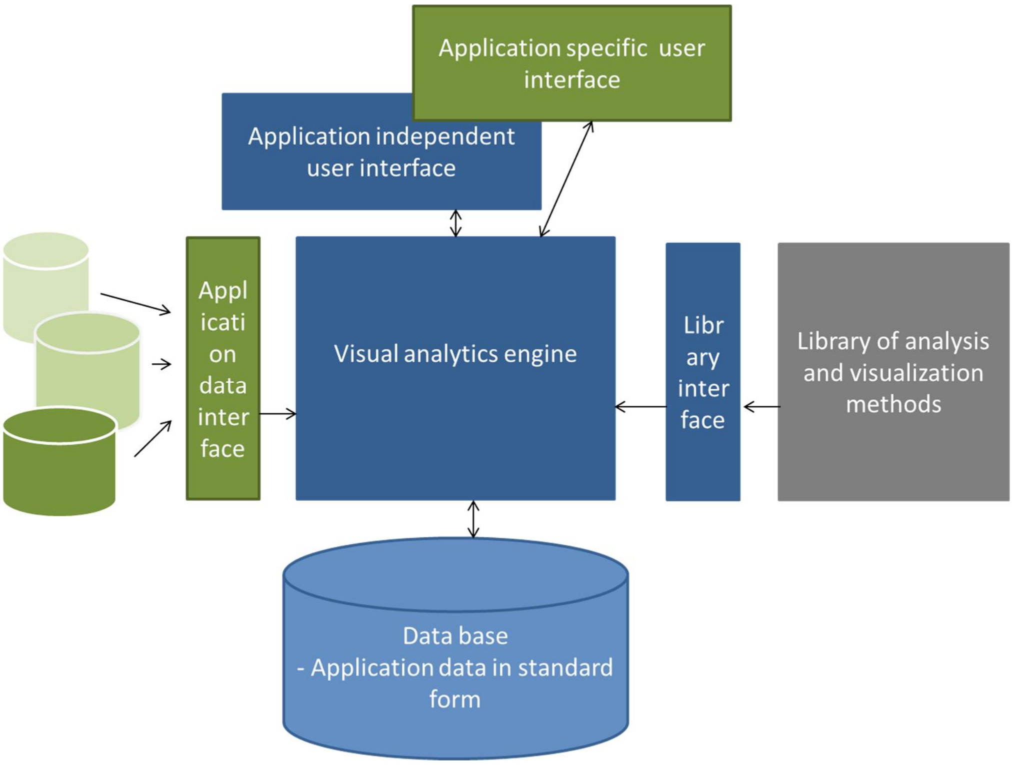

Application specific user interface
Application independent user interface
Appl icati on data inter face
Visual analytics engine
Libr ary inter face
Library of analysis and visualization methods
Data base
- Application data in standard form

Figure 37. System architecture. The domain database stores the application data, which is loaded from the external application sources. The analysis and visualization methods, chosen for the domain, are stored in a library. The data and the methods are available to the user via an application-independent user interface. The core of the system is the visual analytics engine, which delivers the data between the different components.

In the above approach, users are not required to have any detailed knowledge about the analysis or visualization methods. They simply obtain suitable pre-selected methods from the library via the user interface. This approach could be beneficial to users who know the application domain but are not familiar with the analysis methods. Using a domain specific approach also provides flexibility, which is lacking from many of the present visual analytics applications. It also avoids the need to build case-by-case solutions. However, it does not solve all problems, as the data still has to be extracted, cleaned and harmonized as before, and scalability requires additional attention.

Figure 38 outlines the suggested solution represented as a level structure, as in industrial models (MIMOSA, 2012; ISO 13374, 2007). At the *Metering level*, measurement data is collected from disparate sources. The *Measurements level* stores the measurement and background data into diverse data storages: repositories, databases and files. At the *Data model level*, the user defines and enters the application concepts to the system database according to the domain model. The set of analysis and visualization methods can be checked and new methods can be installed if required. At the *Mapping level* the data coming from the disparate sources is mapped to the database objects. This includes cleaning, harmonization and transformation. At the *Visual Analytics level* all predefined visualizations and analyses are automatically available through the domain specific user interface. If required, a more elaborate application-specific user interface can be constructed on top of the general-purpose user interface. This work covers the three top levels.

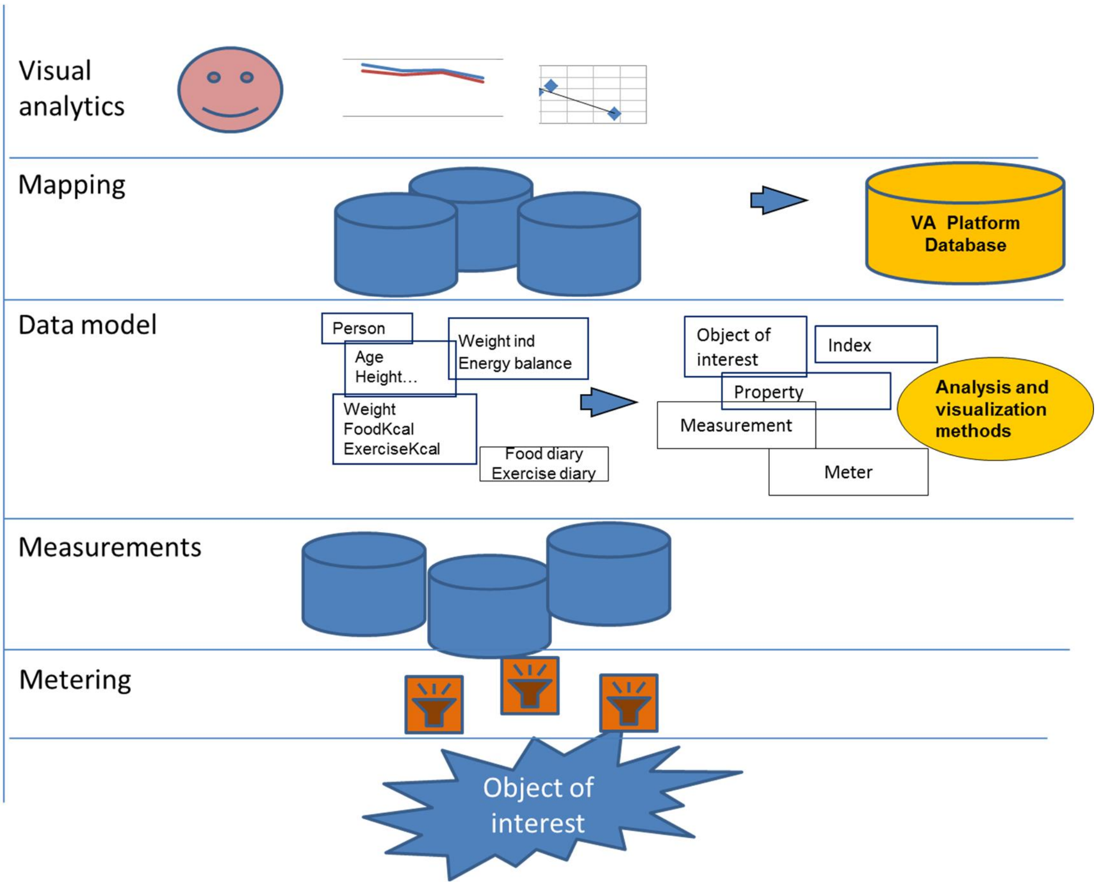


Figure 38. Level architecture of measurement data analysis.

Support of reasoning

Visual analytics systems should avoid cognitive overload, such as inconvenient user tasks that interrupt the reasoning process. Measurement data often includes a very wide range of variables. This presents a huge number of visualization and analysis options, which can be confusing for the user. Users would therefore benefit from guidance on appropriate methods, thus limiting the number of choices open to them. Ferreira de Oliveira and Levkowitz (2003) claim that at present the selection of the right method is a largely intuitive and ad hoc process.

In this work, reasoning is supported with the help of metadata that is added to the domain model. Metadata is “data about data”, such as information about data types, units, value ranges, etc. Metadata is used in two ways: (1) for selecting the suitable visualization and analysis methods in different situations, and (2) for supporting the interpretation of the analysis results.

From the technical point of view, the applicability of visualization and analysis methods often depends, at least with respect to the methods used in this work, on the number of variables, the data types of the variables, the number of observations, and the time scale. As introduced in Section 3.8, there are specific methods for single variables, for two or more variables, for time series, and for ordinal and quantitative data. Combining the metadata from the data model and user selections, the appropriate analysis methods can be suggested for users. For example, if the user has selected two numerical variables for analysis and the amount of data items is moderate, the system can suggest calculation of correlations and showing scatterplots.

Value ranges and other kinds of application metadata can be shown in the visualizations to help interpret the visualization result. In the monitoring domain, it is important to compare the monitored values to the limits and reference values. Alarms are activated if the value is not within the defined limits. This kind of metadata, such as limits, goal values and reference values of variables, can be easily included in the data model and shown in visualizations.

Data model as a user interface element

Monitoring data can accumulate in large amounts. This requires a method for visualizing large collections of data. In this work the data model is used as an approach to the *focus–context* problem of information visualization. Data models are known to be efficient means of communication. A data model shows clear concepts and their relationships to the user. Here, the overview of the data is given with the help of a data model (*context*). The user can select interesting concepts from the model (*focus*) and apply visualization and analysis methods to them. This kind of user interface can be generated from the domain model.

7.2 Domain model for monitoring data

There are many ways to create a data model. A domain-specific model preserves semantics effectively and is comprehensible, but is applicable only in the application domain for which it is built. A generic model loses semantics and can make applying and understanding the model complicated. A good model can make the application flexible, easy to build and update. A poor model can make the application complicated and difficult to maintain. Here, a semi-generic model is suggested as an approach that is both comprehensible and applicable. It defines the key concepts, that is, those that remain similar from one application to another in the domain and are understandable and meaningful to the user. The model takes into account the analysis and visualization needs and includes metadata that can be used to support reasoning.

Suggested data model

The key concepts of the measurement data domain are:

- *objects of interest*, *measurements* (value, timestamp pairs),
- *indexes* (key performance indicators), and
- *background knowledge* of the object of interest.

A simple data model showing these key concepts is given below:

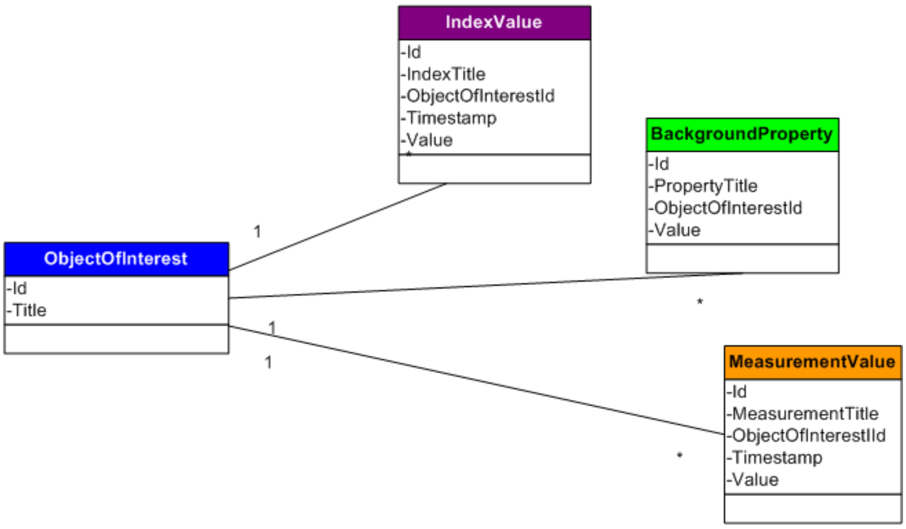


Figure 39. Key concepts of measurement data.

An object of interest is a phenomenon that is monitored. It can be equipment, a building, or an individual person. In one application there can be several objects of interest with their specific background data, measured properties and indexes. Objects of interest can form hierarchies. Measurements are measured values of a property of the object of interest, such as temperature of a room, energy usage of a machine, exercise length or food eaten by a person. Background knowledge consists of static properties of the object of interest. Examples of these are the age and type of machine, the structure and material of a building, or the date of birth of a person. Indexes are values calculated from the measurements and background knowledge that users want to follow. Examples of indexes are energy performance of a building, body mass index of a person, or air quality index of a room.

This simple model does not take into account the features that could be helpful to support reasoning. Useful additions to the model could be:

- *The data types of the measurements, background properties and indexes.* These can be nominal, ordinal or quantitative. Measurements can be absolute or cumulative.
- *Properties of properties*: Quantitative properties can have an unit, value limits, goal and reference values. Nominal and ordinal properties have code values. Measured properties have a measurement frequency.
- *Hierarchies of objects of interest.* For example, an area contains buildings, buildings consist of spaces, and equipment consists of parts.
- *Grouping of properties.* Measurements can sometimes relate to each other or serve as joint measures for a given aspect. These can be grouped together to assist the user. An example grouping is food nutrients: fat, protein, carbohydrates, sugar and fibre.
- *Variables used in index calculation.* Knowledge of these can be used for exploring details and finding explanations behind the indexes.

The model, complemented with these features is shown in Figure 40. The grey objects above the line represent metadata and the rest the actual data.

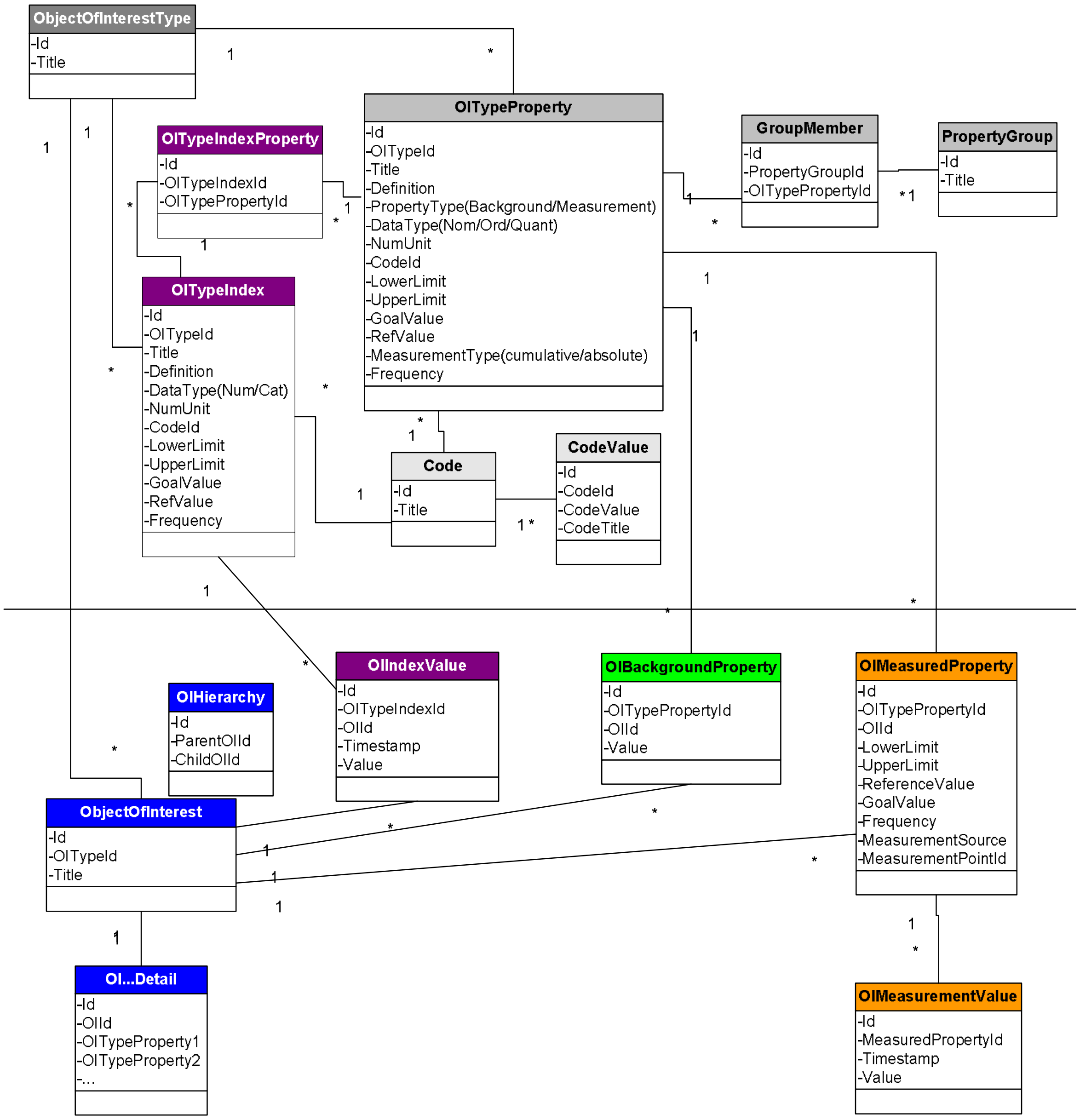


Figure 40. Domain model for monitoring data.

The concepts of the final model are:

- *ObjectOfInterestType* (acronym OIType) lists the phenomena that are monitored. Example object of interest types are building, building space, area, or person.
- *OITypeProperty* describes the kinds of properties objects of interest types can have. Properties can be background properties or measured properties. Their data types can be nominal, ordinal or quantitative. Categorical properties are coded and the coding and the code values are defined in *Code* and *CodeValue* objects. If the property is numerical it has an unit, lower and upper value limits, and goal and reference values. Measured properties also have a measurement frequency and type, type indicating whether absolute or

cumulative values are measured. Dependent properties can be grouped together in *PropertyGroup* and *GroupMember*.

- An *OITypeIndex* defines the kind of index related to each object of interest type. *OITypeIndexProperty* links indexes to the properties used in index calculation.
- *ObjectOfInterest (*acronym OI*)* lists the real objects that are monitored.
- *OIHierarchy* defines the hierarchical structures of ObjectsOfInterest.
- *OI...Detail* is an object for descriptive information of ObjectsOfInterest. A building can have a location on a map, or a 3D model, for example. There can be OIDetail objects for different kind of ObjectOfInterestTypes, named accordingly (e.g. OIBuidlingDetail, OIPersonDetail).
- *OIMeasuredProperty* stores limits, references and goal values of an individual object of interest. They can differ from the type level values of the Property Type object.
- *OIMeasurementValue* stores the actual time-stamped measurement values.
- *OIBackgroundProperty* stores the background properties of the real objects.
- *OIIndexValue*, stores index values of the real monitored objects. These can be time stamped.

7.3 Discussion

Why a new model?

Instead of creating a new model, another alternative could have been to use the models presented by the official standards and the research community and enlarge these with new features. There are, however, many reasons for developing a new data model. The focus of the industrial standards, such as MIMOSA (2012) and ISO 13374 (ISO 13374, 2007), is on sensors and data collection, which is not the aim of the present work. Furthermore, the standard models are huge in scope, seeking to take into account every possible alternative and detail. The inclusion of concepts not relevant to the goals of this work was considered unnecessary. While the model suggested by Inaudi (2002) shares some of the same areas of interest as the present work – analysis and visualization – and is of reasonable size, its scope is more limited. The Inaudi model does not store background data on the monitored objects or metadata on measurements, which can be useful in data analysis. The model is focused more on modelling sensors and collecting units. Expansion of the earlier models with the concepts suggested in this work is, however, considered a possible approach, and offers one direction for further research.

Comparison to related work

In comparison to the other approaches that use data models, introduced in Section 5.2, this approach falls between the classes "General-purpose domain model" and "General model". It has a domain model, but it can be applied to all applications inside the domain, being more of a "General-purpose domain model". On the other hand, it is not applicable outside the domain. As it is restricted to a domain that shares similar concepts, constructing applications is more straightforward than with the general models. There is no need to create an application model as in the approaches by Kang et al (2007) and Streit et al (2012), instead the application data is instance data of the domain model. In addition, the user interface can be built to serve the needs of the domain. This is in line with the claim by the Visual Analytics Agenda that visual analytics systems that do not adequately take into account the context of the data and their use will likely fail.

The author's search for visual analytics approaches that use metadata to support reasoning produced little results. Kang et al (2007) mention using metadata and that "the quality of the interface depends on the richness and value of metadata", but provide no further explanation. The only other example found comes from healthcare (Azarm, Nargesian and Peyton, 2011b), and has a data architecture that combines report delivery with business analysis applications to produce a "healthcare analytics dashboard". It has a metadata repository represented as a conceptual model. The metadata comprises details of captured data, minimum and maximum values, data type, technical metadata, business metadata, and information on data availability and sources. The repository is connected to business intelligence tools that store and report the data. The metadata is used to keep track of the lineage between the results and the original report data, not actually to support reasoning.

Furthermore, none of the studied visual analytics approaches support the dynamic context-sensitive selection of suitable visualization and analysis methods. They used either one kind of visualization, typically network, or different kinds of visualizations that were fixed in the user interface. Some flexibility was found in the works of Streit et al (2012), who propose reconfigured analysis steps that can accommodate different visualization methods, but the reconfiguration is performed prior to the analysis.

Data models or elements of them have been used as user interface elements, but not in the way specifically suggested in the present work. From the studied works presented in Chapter 5, only Shen et al (Shen, Ma and Eliassi-Rad, 2006) enabled the selection of concepts from the ontology graph in order to form a new kind of graph. Others used concept lists to form new networks or other pre-selected views.

8 Tool concept

In this chapter a concept for a visual analytics tool based on the defined domain model in introduced. The concept consists of a reasoning process, a user interface, and a selection of analysis and visualization methods. A step-by-step process for applying the concept in the construction of new applications is defined. Finally, this work is reviewed against the requirements for visual analytics tools, and compared to other ways to implement analytics solutions.

8.1 Reasoning process

The tool is constructed around a reasoning process. The reasoning process is based on the model by Pirolli and Card (2005). The flow of the process is represented in Figure 41. First, the user is shown the available data objects (1) and their relationships in the form of a data model or some other visualization. The user can browse the objects and view the basic visualizations and statistics. Then user then selects model objects for analysis (2). The system recommends to the user suitable analysis and visualization methods to apply based on the data types of the selected objects, the number of objects, and the time scale. Based on the user's selection, an analysis and visualization is performed and the results are shown to the user (3). The user interprets the result and looks for patterns (4). The user continues the analysis by making a new analysis and visualization selection (5). The user can select another analysis and visualization with the same data, get new data objects, remove uninteresting or irrelevant objects, and continue the analysis with this new data set. Alternatively, the user can select a subset of data directly from the visualization and include this subset in the data to be analysed. The process continues with the user exploring the data. User can go deeper into details until raw data – or until insight is gained.

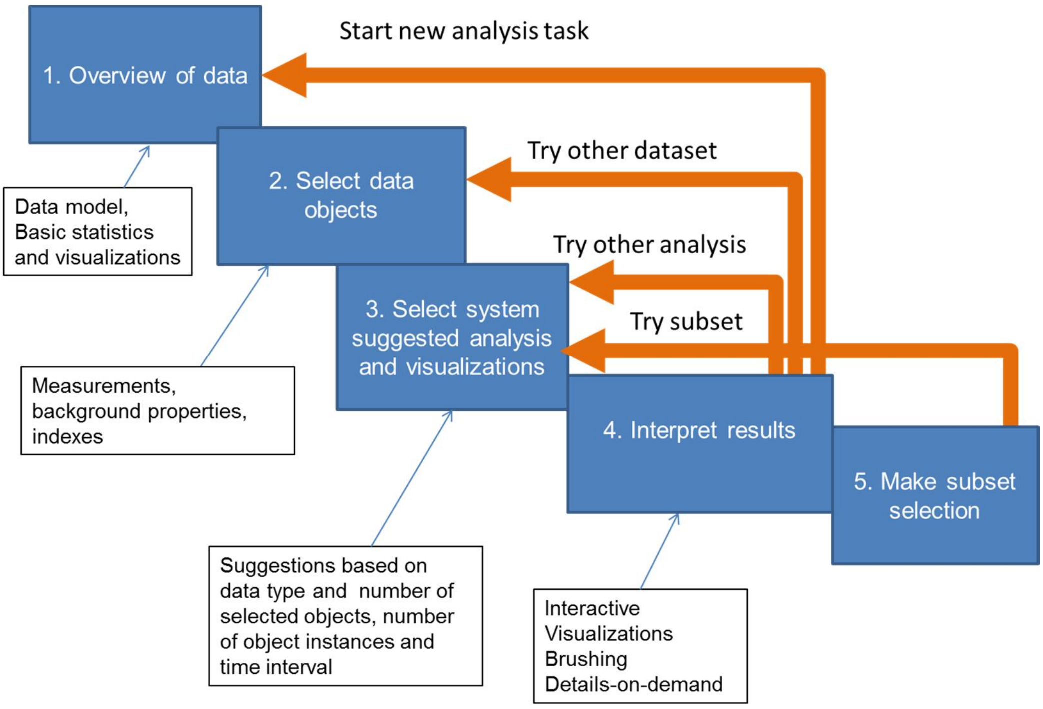

1. Overview of data
Start new analysis task
2. Select data objects
Try other dataset
Data model, Basic statistics and visualizations
Try other analysis
3. Select system suggested analysis and visualizations
Try subset
Measurements, background properties, indexes
4. Interpret results
5. Make subset selection
Suggestions based on data type and number of selected objects, number of object instances and time interval
Interactive Visualizations Brushing Details-on-demand

Figure 41. Reasoning process.

Steps 1 and 2 correspond to Pirolli's information gathering, step 3 to Pirolli's representation of the information, and step 4 to the development of insight. The model includes Pirolli's *Foraging loop* – a cycle of activities around finding information (steps 1 and 2), and *Sense-making loop* for making sense of the information (steps 3 and 4). In step 5 the user can create new concepts for analysis as the work of Gotz et al (Gotz, Zhou and Aggarwal, 2006) suggests.

8.2 User interface

A general-purpose user interface is designed to support the reasoning process. It supports the reasoning process, which sets the requirements for the user interface functionality. The tasks that the user should be able to do include:

- get an overview of the data,
- filter a dataset for analysis,
- select an analysis and visualization method for the dataset,
- see the result in visual form,
- and pick subsets of the analysis results and form a new dataset for further analysis.

The techniques used to support the analysis are direct manipulation, allowing the user to filter or select elements from the visualizations; coordinated multiple views; and brushing. Use of dynamic queries, menus and buttons is minimized to avoid interrupting the user's attention.

This section introduces the user interface layout, how the reasoning tasks are performed and how direct manipulation, coordinated multiple views and brushing work. The examples are from the HyperFit case introduced in detail in Chapter 9. A slightly different user interface is represented in the MMEA case in Chapter 10.

Layout

Figure 42 shows the user interface layout. The surface is divided into a data area (left), user work area (right), and tool panel (top). The data area shows the data contents as a node link diagram constructed from the data model. Colours are used to distinguish the background properties (green), measurements (orange) and indexes (violet). The thickness of the links corresponds to the volumes of the dataset. The work area is used for selecting data sets, launching analyses and showing visualizations. The tool panel has tools for picking variables from the model , selecting subsets (icon image can be modified) and adjusting the time scale and period .

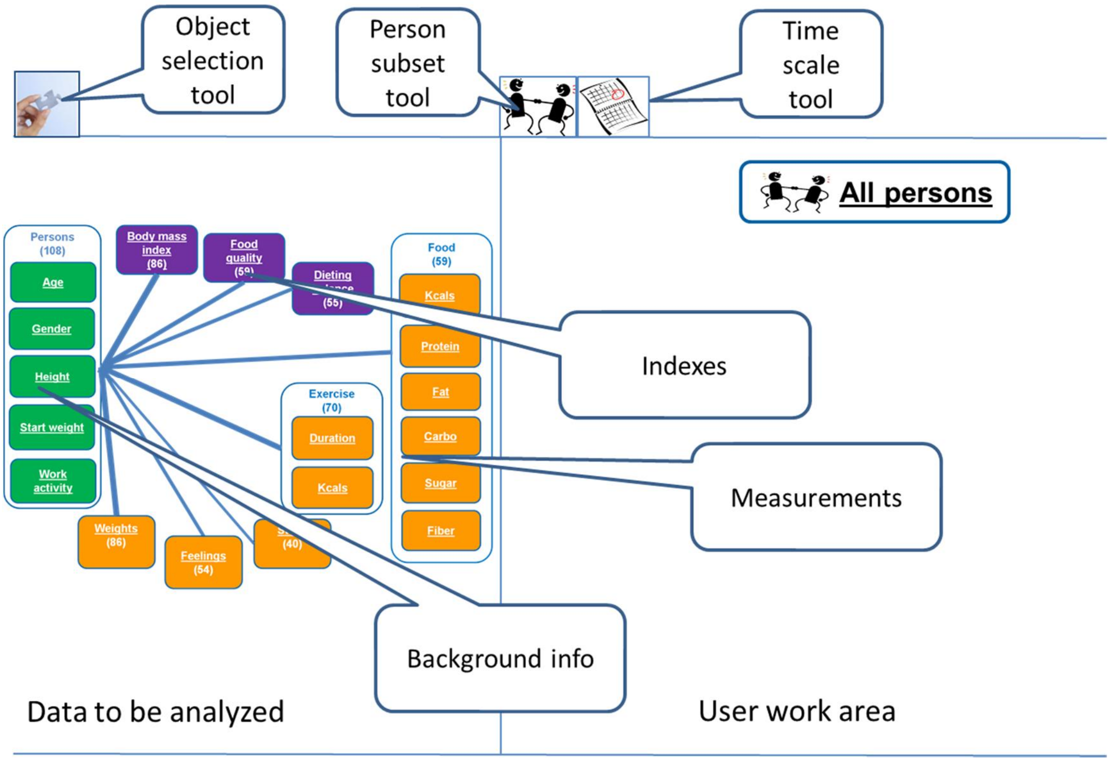
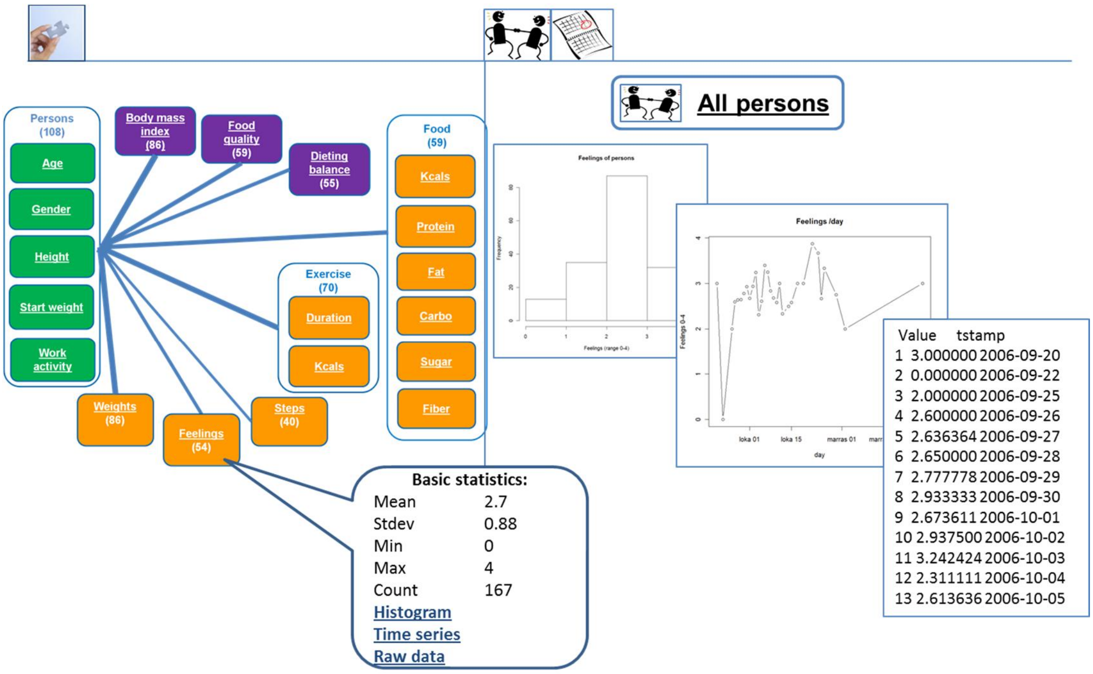


Figure 42. User interface layout.

Overview of the data

The data model provides an overview of the available data. In the example given there is only one kind of object of interest, and the diagram shows only the properties of dieting people. Clicking the model objects gives the user quick access to basic statistics, histograms, time series and raw data (Figure 43).



Basic statistics:

Mean	2.7
Stdev	0.88
Min	0
Max	4
Count	167

Histogram
Time series
Raw data

	Value	tstamp
1	3.000000	2006-09-20
2	0.000000	2006-09-22
3	2.000000	2006-09-25
4	2.600000	2006-09-26
5	2.636364	2006-09-27
6	2.650000	2006-09-28
7	2.777778	2006-09-29
8	2.933333	2006-09-30
9	2.673611	2006-10-01
10	2.937500	2006-10-02
11	3.242424	2006-10-03
12	2.311111	2006-10-04
13	2.613636	2006-10-05

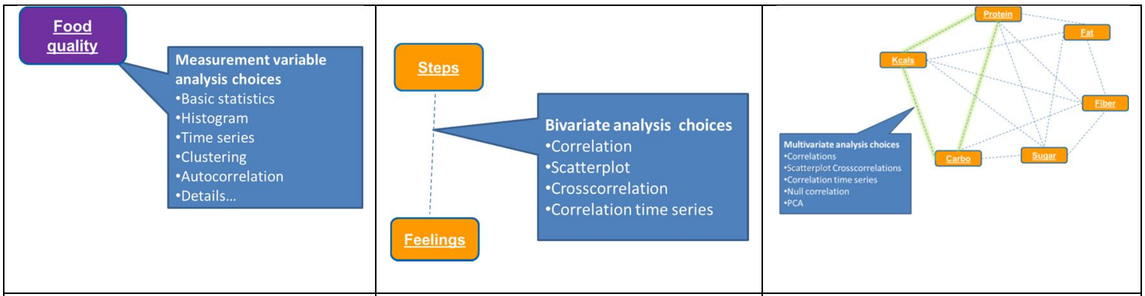
Figure 43. Overview of the data.

Filter datasets for analysis

The user forms the dataset for analysis by picking properties, measurements and indexes of interest from the data model and placing them in the work area using the data picker tool. By default, all objects of interest and the whole time period is used in the analysis, but the user can define a subset using the subset tool . The time scale tools allow the user to adjust the time period and time granularity (day/week/month). These tools are not defined here further as they are basic user interface elements and not essential here.

Selecting an analysis and visualization method for the dataset

After the dataset is selected the system suggests a suitable analysis for the user. The suggestions are based on the data types of the selected data variables (measurement/background data, ordinal, quantitative, time series etc.), the number of selected variables, the number of objects of interest the user has selected, and the length of the time period. The user receives the suggestions in a menu that appears in the work area. If there is only one object, the user receives the suggestions by clicking the object (Figure 44 a). If there is more than one object, they are linked together with a dashed line. Clicking the line gives the suggestion (Figure 44 b). When there are more variables, the user selects the objects included into the analysis and receives the suggestions for the selected group (Figure 44 c).



a) Single variable

b) Two variables

c) Multiple variables

Figure 44. Selecting analysis and visualization methods

Showing the analysis results in visual form

The user receives the analysis results in visual form in the work area. Correlations between variables are expressed as the thickness and colour of the connecting line. Some visualizations, such as scatterplots, time series and histograms, are shown as independent windows and the user can either keep them open in the work area or close them (Figure 45).

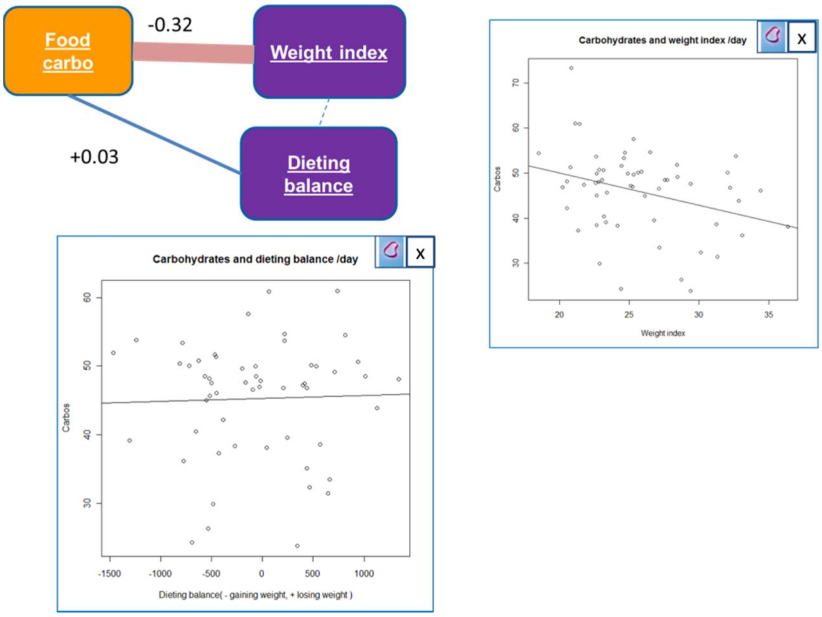

Food carbo
-0.32
Weight index
+0.03
Dieting balance
Carbohydrates and weight index /day
X
Carbos
Weight index
20
25
30
35
30
40
50
60
70
Carbohydrates and dieting balance /day
X
Carbos
30
40
50
60
-1500
-1000
-500
0
500
1000
Dieting balance(- gaining weight, + losing weight)

Figure 45. Showing the analysis results**.**

Picking subsets of analysis results for further analysis

Each visualization has a visual subset selection tool, the “Magic rubber band” , in the top right corner of the visualization window. The user can select a data subset from the visualization with the tool, create a new object from the subset, and continue the analysis with it (Figure 46).

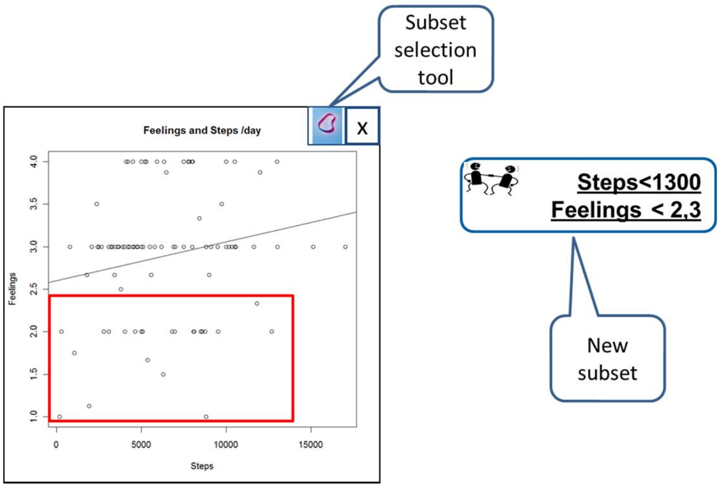


Figure 46. Creating subsets.

Brushing

The user can simultaneously keep several visualization windows open and highlight the same data objects in each visualization. The user can select objects in the visualizations using the subset selection tool (Figure 47).

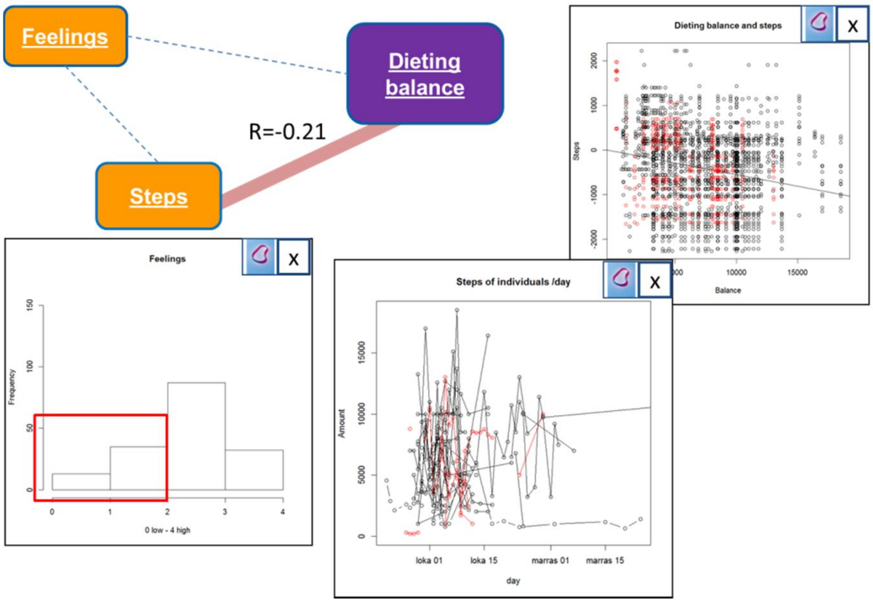
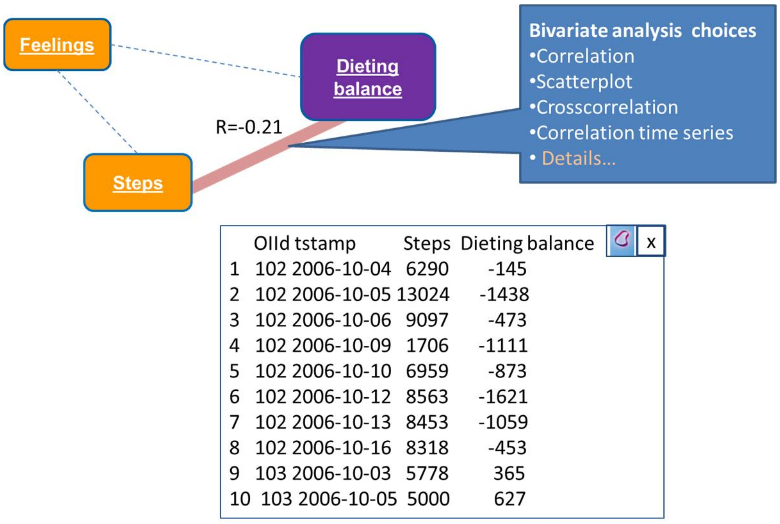


Figure 47. Brushing.

Details-on-demand

The user can click an object or a relationship and get the underlying raw data (Figure 48).



Bivariate analysis choices
- Correlation
- Scatterplot
- Crosscorrelation
- Correlation time series
- Details...

	OIId	tstamp	Steps	Dieting balance
1	102	2006-10-04	6290	-145
2	102	2006-10-05	13024	-1438
3	102	2006-10-06	9097	-473
4	102	2006-10-09	1706	-1111
5	102	2006-10-10	6959	-873
6	102	2006-10-12	8563	-1621
7	102	2006-10-13	8453	-1059
8	102	2006-10-16	8318	-453
9	103	2006-10-03	5778	365
10	103	2006-10-05	5000	627

Figure 48. Viewing details.

8.3 Analysis and visualization

The suggested approach does not require a fixed set of analysis and visualization methods. The only requirement is that the methods are compatible with the data types of the domain model. The analysis and visualization methods used in this work were chosen with the following requirements in mind: the methods have to be applicable to monitoring data that consists of time series and static variables, nominal, ordinal and quantitative variables, and single and groups of variables. The methods should be applicable to simple table-like data structures that the data model supports. They should be comprehensible to non-expert users. The applications users can be domain professionals with no advanced data or statistical analysis skills.

In this work a set of data exploration and descriptive methods data have been chosen. They include methods the majority of the methods introduced in Section 3.8. In addition to existing methods, a new visual method for analysing correlations between multiple variables, the *residual networks* method, is introduced later in this section.

The used methods are listed in Table 3. They include basic statistics, pie charts and histograms, time series, cross and auto-correlations of time series, scatterplots with linear regression, correlation matrixes and networks, PCA, parallel coordinates and hierarchical clustering. They are categorized by data type (nominal, ordinal, and quantitative), single values or time series, and the number of variables (univariate, bivariate, multivariate). Single values can be background values or single measurements, or means or sums of time series values.

Table 3. Analysis and visualization methods.

	No time point/single time point Background variables Means or sums of time series measurements of each object of interest			**Time series** Measurements/Indexes Hourly, daily, weekly Sums/means of several objects of interest		
	Nominal	**Ordinal**	**Quantitative**	**Nominal**	**Ordinal**	**Quantitative**
Univariate	Pie chart Bar chart	Basic statistics: Mean, var, min, max, Pie chart Bar chart	Basic statistics: Mean, var, min, max, count Histograms Clustering		Time series line graph Auto-correlations	Time series line graph Auto-correlations
Bivariate		Scatterplot with regression line Correlation r and coefficient R^2 (with caution)	Scatterplot with regression line Correlation r and coefficient R^2		Cross-correlations (with caution)	Cross-correlations
Multivariate			Multiple scatterplots Correlation matrix or network Residual network Parallel coordinates PCA			

Residual network

Residual network is a new visual method for analysing correlations between multiple variables. The starting point is the correlation network. The thickness of the link and the link colour indicates the value of the correlation. Positive correlation is blue and negative reddish. Figure 49 shows the correlations between kcal intake and energy percentages of nutrients in the daily diet of dieting people.

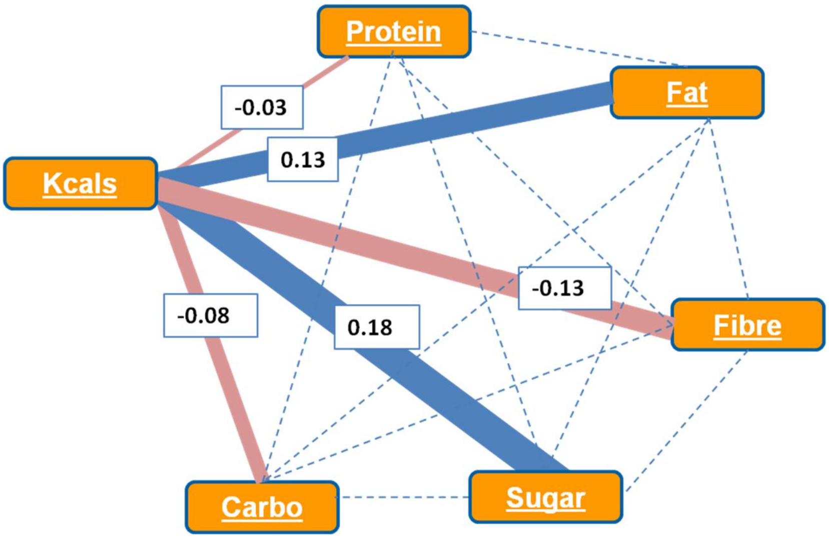


Figure 49. Starting point showing all correlations.

If the user considers a correlation to be not of interest or self-evident, the correction can be “discarded” by replacing it with the corresponding residual. A new network is created using the residual. In the previous figure the correlation between kcals and sugar was the strongest and self-evident. In Figure 50 it is discarded and the correlations between kcals and fat and kcals and carbohydrates become the most meaningful. The inference could be that people who do not get energy from sugar get it from fat.

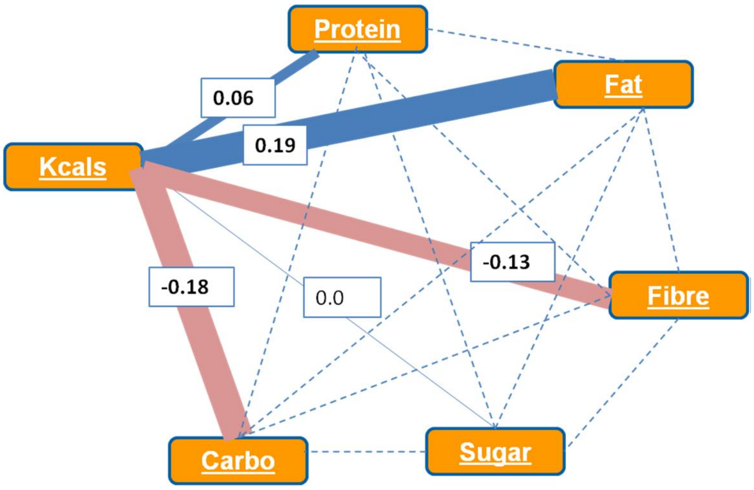


Figure 50. Correlation between kcals and sugar is discarded**.**

In Figure 51, the correlation between kcals and fat is discarded. The inference here might be that those who do not eat fibre, carbohydrates or fat obtain kcals from protein (pure low-carb diet).

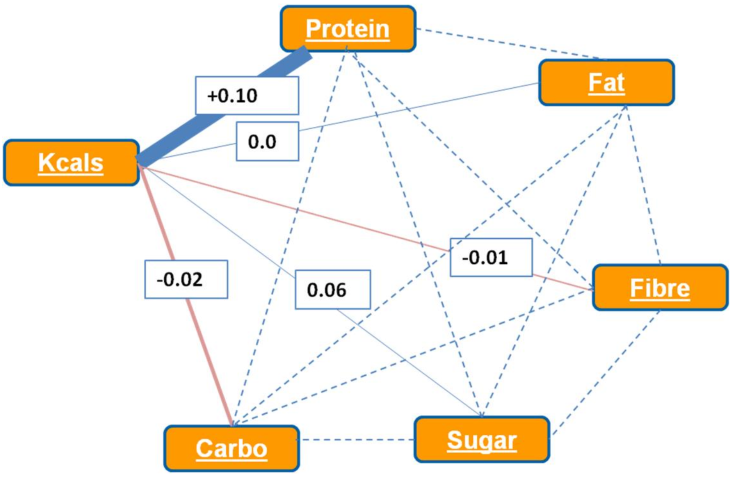
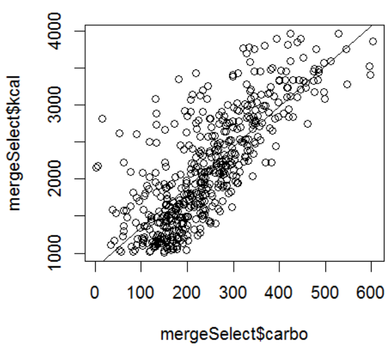
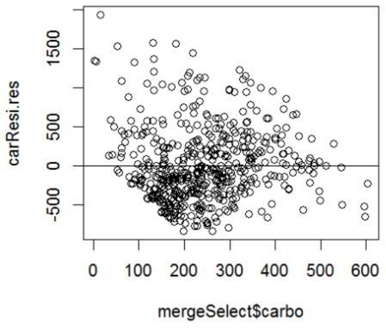


Figure 51. Correlation between kcals and fat is discarded.

The principle behind the method is based on linear regression. In regression analysis one variable (dependent value) is explained with the help of one or more other variables (independent values). The residual is the difference between the observed value of the dependent variable (kcals) and the predicted value by the model (the line value). The residual represents the variance of the dependent variable that is not explained by the independent value. So, if the dependent variable is replaced with the residual in the correlation network, the network shows the correlation as if the effects of the independent variable had been discarded.

In Figure 52 on the left is a regression line that explains correlation between kcals and carbohydrates. On the right-hand side is a residual plot showing the residuals on the vertical axis and the independent variable (carbohydrates) on the horizontal axis. In the example, the residual represents the variance in kcals that carbohydrates do not explain. Now, if the kcals are replaced with the residual, the network shows the correlations to other variables as if the effects of carbohydrates had been discarded.

<table>
<tr>
<td>


</td>
<td>


</td>
</tr>
<tr>
<td>r= 0.773161</td>
<td>r=1.934176e-17</td>
</tr>
</table>

Figure 52. Scatterplot with regression line.

8.4 Construction of a new application

The construction of a new application process consists of six steps, as shown in Figure 53.. The prerequisites are that the database is constructed according to the data model, a set of analysis and visualization methods are mapped to the model, and that a general-purpose user interface exists.

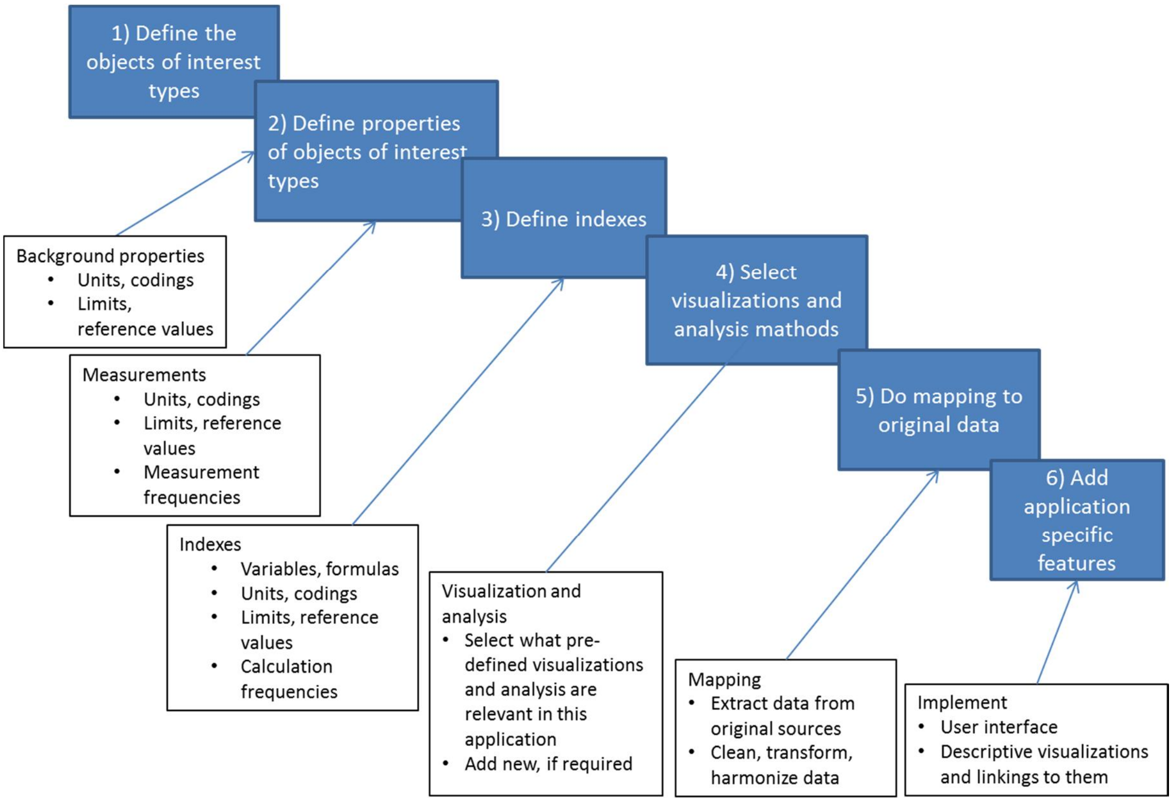

1) Define the objects of interest types
2) Define properties of objects of interest types
3) Define indexes
4) Select visualizations and analysis mathods
5) Do mapping to original data
6) Add application specific features
Background properties
• Units, codings
• Limits, reference values
Measurements
• Units, codings
• Limits, reference values
• Measurement frequencies
Indexes
• Variables, formulas
• Units, codings
• Limits, reference values
• Calculation frequencies
Visualization and analysis
• Select what pre-defined visualizations and analysis are relevant in this application
• Add new, if required
Mapping
• Extract data from original sources
• Clean, transform, harmonize data
Implement
• User interface
• Descriptive visualizations and linkings to them

Figure 53. Construction of a new application**.**

The steps include the following tasks:

Step 1: Define object of interest types

> Object of interest types define the phenomenon being monitored, for example a building, person or machinery. The OIType object in the model is updated.

Step 2: Define property types

> All possible property types for the different object of interest types are defined. For example, for buildings age, volume and energy consumption are defined.
>
> The properties are classified as background / measured properties, nominal/ordinal/quantitative, cumulative / absolute. Units, coding, value

ranges, goal and reference values are defined. If properties are dependent on each other, they can form a property group. The model objects Property, PropertyGroup, GroupMember, Code and CodeValue are updated.

Step 3: Define indexes and calculated values

The indexes and calculated values and their formulas are defined. The property types that are used in the formulas are attached to the index. Index calculation formulas are defined and implemented. Indexes require implementation because they are unique to each application. The model objects Index and IndexProperty are updated.

Step 4: Define visualizations and analysis

The platform provides a predefined set of visualisations and analysis. In this step the set is reviewed. New methods can be added and unnecessary ones discarded.

Step 5: Do the mapping

Mappings from the original data sources are performed. There are three types of mapping:

1) Objects of interest: the real objects of interest are defined. These are the actual objects being monitored, e.g. buildings or persons. They can form hierarchies, e.g. a building belongs to an area, a person lives in a building. The model objects ObjectOfInterest and OIHierarchy are updated.

2) Measured properties: the actual measurements are listed and mapped to the property types. E.g., water consumption is measured for the Otaniemi Digitalo building. The measurement frequency is added. The measurement details, e.g. value ranges, reference and goal values, can be redefined if the values differ from the values given in the property definition. The model objects OIMeasuredProperty and OIMeasurementValue are updated.

3) Background properties: the background variables area added. The OIBackgroundProperty object is updated.

The database can then be loaded with the data from the original sources. This requires cleaning, transformation and harmonization. This may involve coding and defining SQL queries, depending on the system implementation. Index value calculation can be launched as soon as the required variable values are available. OIIndexValue is updated.

Step 6: Add application-specific features

Details of objects of interest are added. These vary depending on the objects (OIDetail). The user interface can have application-specific descriptive visualizations that are mapped to the user interface.

8.5 Discussion

Comparison to requirements

The design fulfils many of the requirements for a visual analytics tool given by the Visual Analytics Agenda, presented in Section 3.6. It includes filtering using techniques such as direct manipulation and brushing and details-on-demand. Navigation is supported in many ways, including browsing within data by selecting and highlighting objects of interest. Human information discourse is supported, providing ways to extract, recombine, create abstractions and compare. Cognitive load is reduced by direct manipulation and auto-suggestion of analysis methods. The user interfaces also fulfils special requirements such as browsing the time axis and switching between deferent levels of temporal aggregation: daily, weekly, monthly.

However, there are some requirements that the user interface does not carry out. These include categorizing data, creating and testing hypotheses, annotating data, and allowing the user to interactively change the mapping between the data and their visual representations. There are also several other interesting techniques that could be used, such as using different kinds of time axes and animations.

A comparison to other kind of visual analytics solutions

There are several ways to analyse data with the help of visual analytics. One way is to invite or employ a data analysis expert to solve the problem at hand or to construct an application with the help of an analysis package and its visualization features, such as Matlab or R. Another possibility is to buy a commercial package, either general-purpose or application-specific, and train personnel to use it. This work suggests a third approach, a domain specific platform that can be used to quickly construct visual analytics applications in the domain.

A comparison with alternative ways of analysing data using visual analytics was conducted to identify the advantages and disadvantages of this approach compared to the others. The comparison was made with the help of the software quality model by ISO/IEC 9126 standard (ISO/IEC 9126, 2001). The model includes a set of characteristics and sub-characteristics that can be evaluated for each product: *Functionality, Reliability, Usability, Efficiency, Maintainability* and *Portability*. The evaluation was performed as a self-evaluation and expresses only the subjective views of the author, based on the author's 30 years' experience in the software industry. The full evaluation criteria and the results of each characteristics of the different approach are presented in Appendix E. The results of the evaluation are summarised below.

Tailored solutions by a data analysis expert can produce highly accurate and efficient solutions to customer problems. However, it relies heavily on the skills of the implementer. Usability, reliability, maintainability and portability may receive less attention. Interactive visualizations and their interoperability may be limited. Such solutions are also not cost-efficient; development and modifications require expensive expert work and the acquired knowledge is transferred – or lost – though individual experts. This alternative might be recommendable for distinct and complex special cases, situations that have special performance requirements, or when no other solutions are feasible.

The solutions provided by commercial packages can be reliable, usable, and support security and integration well. However, they may not include all required analysis methods, and possibilities for modification can be limited. They may also require a tedious configuration and learning phase and may need a trained person to use the system. General-purpose solutions are not necessarily very efficient and require considerable computer resources. These kinds of solutions might be recommendable for enterprises that have varying data analysis needs and do not need the most accurate analysis methods.

An application-specific solution can provide an accurate and fine-tuned solution only if a suitable solution exists. Such a solution incorporates expert knowledge of the application area. Usability, security, maintainability and portability depend on the maturity of the product.

Domain approach can provide an accurate solution within the domain by capturing the special expertise of the domain. The standard data model makes this approach flexible. Functionality can be adjusted, and security and interoperability features can be provided as standard services. The solutions become stable when they mature. These kinds of solutions could be recommended for companies analysing several kinds of data from a given domain, or for consultants that want to quickly construct analysis applications for customers.

9 Case HyperFit

HyperFit is an Internet service for personal management of nutrition and exercise (Jarvinen et al 2008). It provides tools for promoting a healthy diet and physical activity. The service includes food and exercise diaries and tools for self-evaluation and personal analysis. It hosts a database of product-specific or average nutritional data on approximately 2,500 foods. The product-specific data comes from Finnish food producers (Fazer Bakeries, Raisio, Valio, Lännen tehtaat) and the averages data from the Fineli® Finnish Food Composition Database[29]. The application was built during a two-year project in 2005-2007 as part of the FENIX Technology Programme (FENIX – Interactive Computing) run by the Finnish National Technology Agency (TEKES).

The application was used during the project in several field tests by individual users, weight management groups and professional nutritionists. Many users continued using the system after the testing periods. The accumulated database currently contains information on the eating and exercise habits of 393 users, including 1,578 exercise diary entries and 14,688 food diary entries, providing an interesting source of information for nutrition research.

This chapter continues below with an introduction of the HyperFit data and the HyperFit visual analytics tool, followed by a step-by-step presentation of the tool's generation process. Examples of reasoning using the tool are then given, followed by evaluation of the approach and discussion on the experiment. The objective in the evaluation is to examine how users are able to reason and gain insight using this kind of tool.

9.1 HyperFit data

The original HyperFit database (data model in Figure 54) contains the background information of persons (birth date, gender, height, start weight, waistline, hipline and work activity), personal goals (ideal weight and weight target, targets for exercise and daily steps), food diaries recording what people eat, when and how much, exercise diaries recording how people exercise, what sports, when and how long for, including the number of daily steps, weight measurements and feelings. In addition to personal data the database contains information on food products including identification and classification information, nutritional content, nutritional composition (e.g. alcohol and salt content), vitamins, minerals and ingredients. Products are classified according to a food taxonomy. Exercise also has a taxonomy based on name, type and physical load factor.

[29] www.fineli.fi

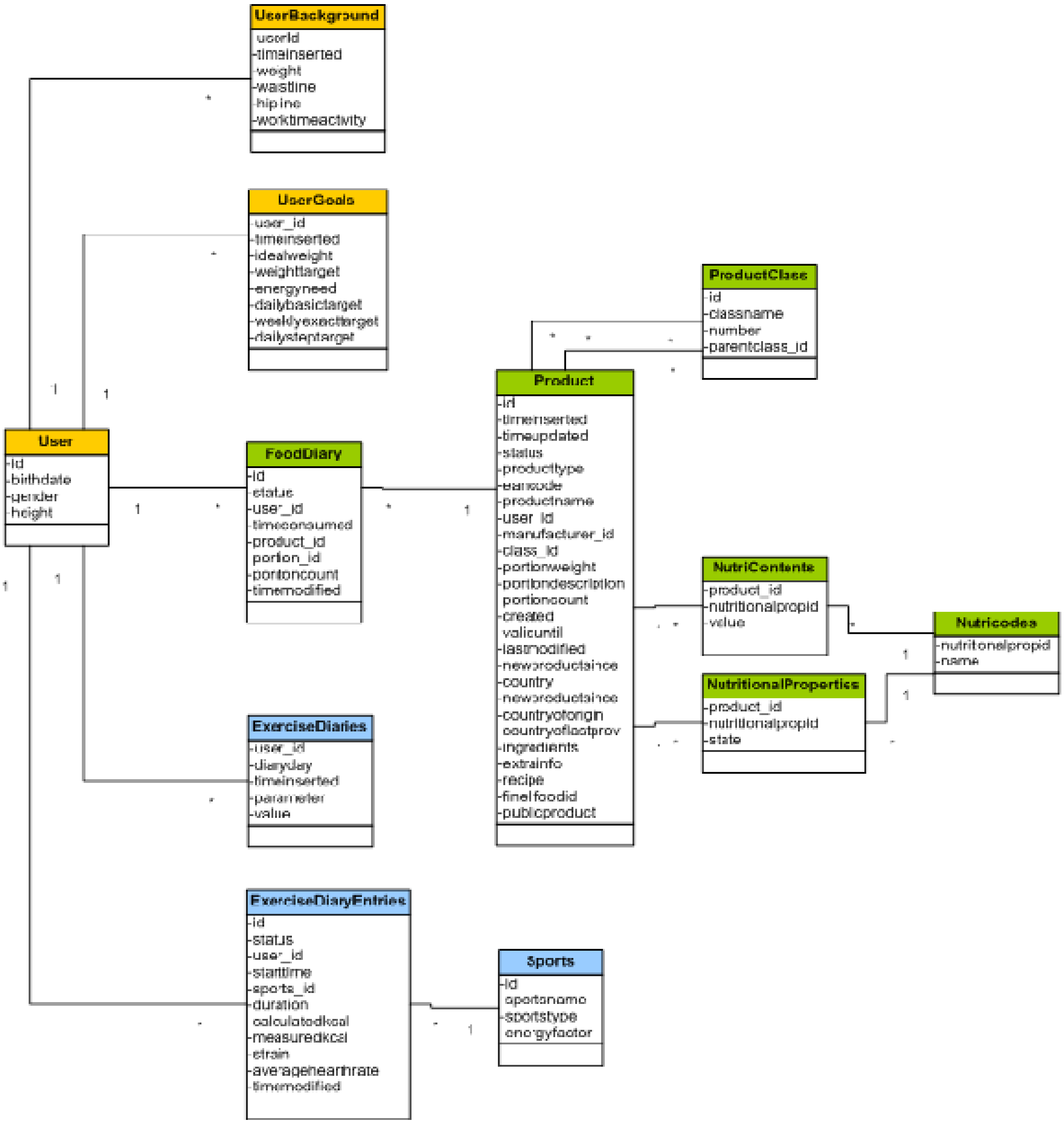


Figure 54. Original HyperFit data (Järvinen, 2007).

9.2 HyperFit visual analytics tool

The HyperFit visual analytics tool is a software application constructed as part of the present work according to the principles described in Chapter 7. The tool enables users to explore the contents of the HyperFit database using a variety of analysis and visualization means. The tool could be used by professional nutritionists wishing to study the dieting and exercise habits of people, for example to discover new connections between mental health, sports and dieting. The application could also serve food retailers and producers and actors in the sports and sports equipment sector. Retail businesses, for example, could use the tool to determine the kinds of food products consumers eat and how they exercise.

A subset of HyperFit data contents was selected for the visual analytics application (defined in detail in Section 9.2). The data completeness and the tool requirements guided the selection. Some data attributes were not useable due to insufficient user data entries (e.g. hipline and waistline). As the main purpose in this work was to test the feasibility of the concept, the amount of similar types of data attributes was not significant. To keep the implementation within reasonable limits, the product information was limited to the most important nutrients (kcal, protein, carbohydrates, sugar, fat, fibres). After cleaning, the databases consisted of data from 108 persons during the time period from 1 December 2003 to 15 September 2006.

For analysis, a set of basic methods for single, bivariate and multivariate variables was selected. As the intended users were not necessarily familiar with data analysis, the objective was to use methods that are easy to understand and interpret. The user interface of the tool is the general-purpose user interface presented in Section 8.2. The tool was implemented as a mixed working and paper prototype. The data management and analysis were implemented with the R Stats Package[30], while the user interface was simulated with a paper prototype.

9.3 Creating the application

Construction of the visual analytics application followed the process defined in Section 8.4. An overview of the process steps is provided below. The detailed process is documented in Appendix B.

Step 1: Define object of interest types

In this case only one object of interest is used: dieting people. The focus of interest is on what they eat, how they exercise, and how they feel.

Other possibilities for objects of interest could be food products and exercise types.

Step 2: Define properties

The background properties are gender, age, height, work activity (1 - light, 2 - medium, 3 -normal) and start weight. The measured properties are daily weight (kg),

[30] A programming language for statistical computing under GNU General Public License

daily feeling (scores from 1 to 5), daily steps, daily exercise duration and kcals, daily kcal from food, daily kcals from protein, fat, carbohydrates, sugar and fibre. A detailed definition with types, units, coding, limits and frequencies is the Appendix B.

Step 3: Define indexes

Three indexes are used: body mass index, dieting balance and food quality index.

Body mass index is defined as the individual's body weight divided by the square of his or her height. The index is a widely used measure for body fat, defining the “thickness” or “thinness” of a person. Index values less than 18.5 mean underweight, values from 18.5 to 25 normal weight, from 25 to 30 overweight, from 30 to 35 significant overweight, from 35 to 40 difficult overweight and values over 40 unhealthy overweight.

Dieting balance is an index developed in the HyperFit project. It compares a person’s daily energy (= energy from food – energy expended through exercise) with the daily energy target. Zero means dieting is in balance, a negative value indicates that the user has not eaten enough compared to their energy expenditure, and a positive value the opposite.

The food quality index is a simplified version of the food quality analysis developed in the HyperFit project and is a combination of dieting balance and food quality. The amounts of fat, hard fat, protein, carbohydrates, fibre and sugar are calculated and compared to the recommendations of the Finnish National Nutrition Council31. The user receives a score from 0 to 6 depending on how closely these values match the official recommendations. The food quality index was developed exclusively for this work and is not an official index.

Step 4: Define visualization and analysis methods

Analysis and visualizations for single variables include: basic statistics (mean, standard deviation, min, max), histogram, pie chart, clustering, time series (all measurements, means/sums of a time point), and auto-correlations. Bivariate analyses include calculating correlation coefficients, scatterplot, linear regression and cross-correlations. Multivariate methods included correlation networks and residual networks. A detailed table of the analysis and visualization methods is presented in Appendix B.

Step 5: Map to original data

In this step the values from the original HyperFit database are transformed to the model objects and properties. The mappings from the original HyperFit database are presented in Appendix B.

Step 6: Add application-specific features

Indexes and calculated value formulas are implemented. The user interface is constructed from the data model. In this case no descriptive visualizations are used.

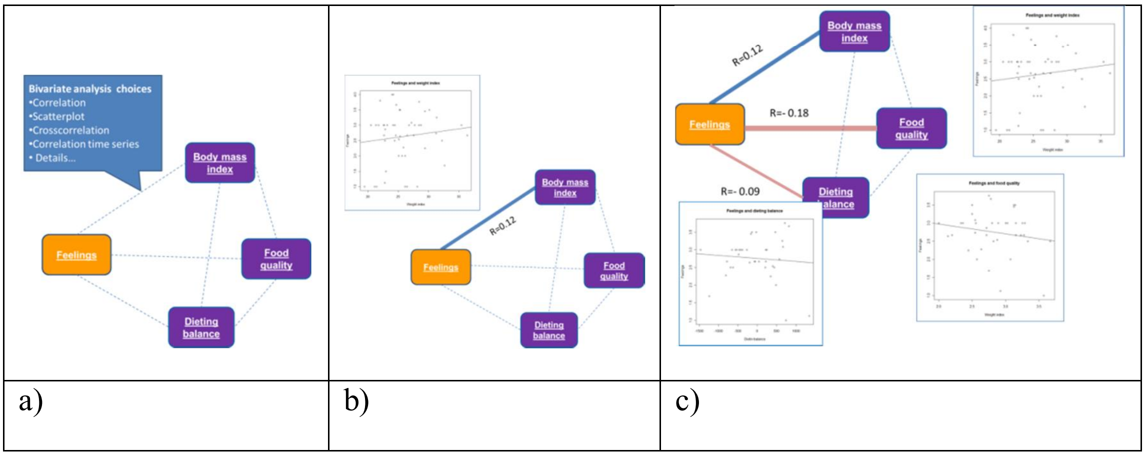
[31] http://www.ravitsemusneuvottelukunta.fi/portal/en/national_nutrition_council/ (accessed 08.11.2012)

9.4 Reasoning examples

Two examples of how reasoning is conducted with the tool are presented in this section. The examples cover two kinds of analysis task: answering questions and finding explanations. The tasks are “How feelings relate to dieting” and “Find explanations for overweight”. The cases were used as part of the user evaluation.

How feelings relate to dieting

The user selects the Feelings object and the indexes that measure dieting – Body mass index, Dieting balance and Food quality – from the data model to the work area. These form a network (Figure 55a). Firstly, the user wishes to study the correlation between Feelings and Dieting balance. The user selects the edge between these objects and receives suggestions for analysis (Figure 55a). The user selects correlation and scatterplot and gets the resulting visualization (Figure 55b), which shows that there is only a small correlation (R=0.12) between feelings and dieting balance. Similarly, the user receives the other correlations and the final results with scatterplots as shown in Figure 55c. There is a slight negative correlation between food quality and feelings, indicating that as food quality improves, feelings deteriorate and vice versa.



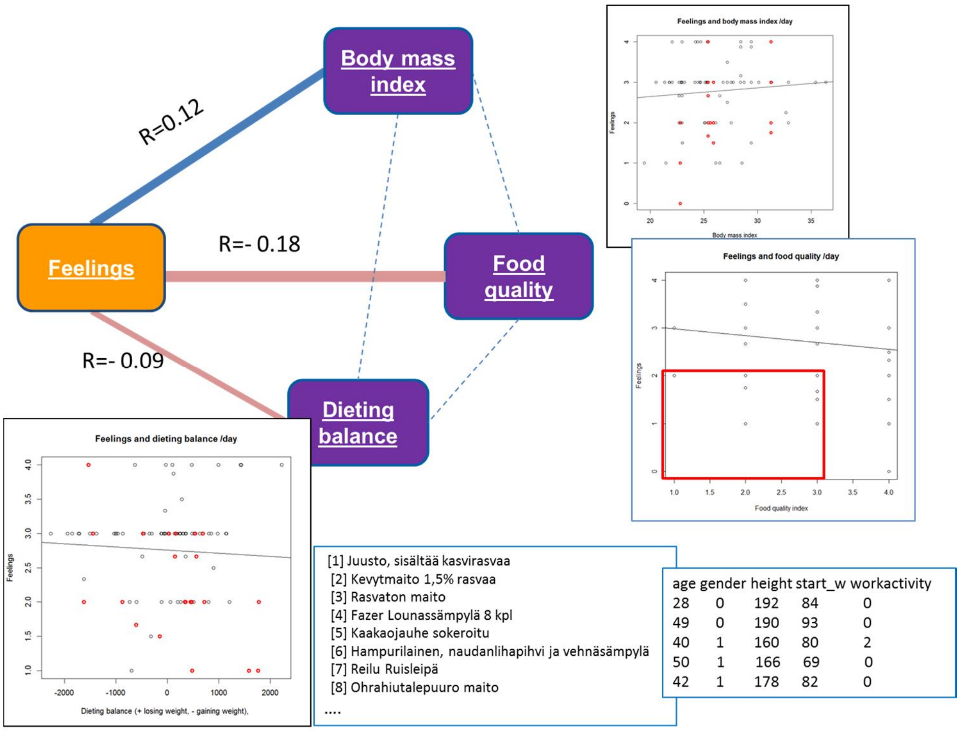
a)

b)

c)

Figure 55. Feelings and dieting

From here the user can continue the analysis, for instance by selecting and brushing subsets and continuing to examine questions such as “How do feelings and food quality correlate with people who exercise a lot”, or by searching for details on a specific group “Which people have low feelings and poor food quality, and what do they eat”. Figure 56 shows the results of the latter case.


Body mass index
R=0.12
Feelings
R=- 0.18
Food quality
R=- 0.09
Dieting balance
Feelings and body mass index /day
Feelings
Body mass index
Feelings and food quality /day
Food quality index
Feelings and dieting balance /day
Dieting balance (+ losing weight, - gaining weight),

```
[1] Juusto, sisältää kasvirasvaa
 [2] Kevytmaito 1,5% rasvaa
 [3] Rasvaton maito
 [4] Fazer Lounassämpylä 8 kpl
 [5] Kaakaojauhe sokeroitu
 [6] Hampurilainen, naudanlihapihvi ja vehnäsämpylä
 [7] Reilu Ruisleipä
 [8] Ohrahiutalepuuro maito
....
```

```
age gender height start_w workactivity
28     0     192     84        0
49     0     190     93        0
40     1     160     80        2
50     1     166     69        0
42     1     178     82        0
```

Figure 56. Continuing the analysis

Finding explanations for overweight

In the next example the user examines what causes some people to have normal and others high body mass indexes, and the differences in eating or exercising behaviour between these groups. First, the user selects the Body Mass Index object from the data model and plots a basic histogram. From there, the user select two subsets: normal-weight (BMI under 25) and overweight people (BMI 25 or over). The system then generates two data subset objects in the work area (Figure 57).

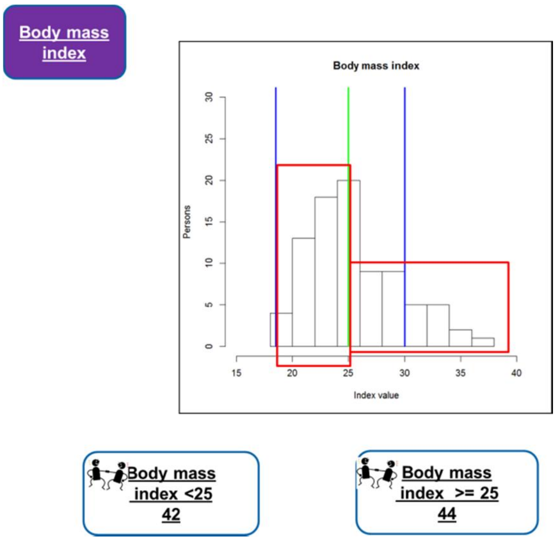
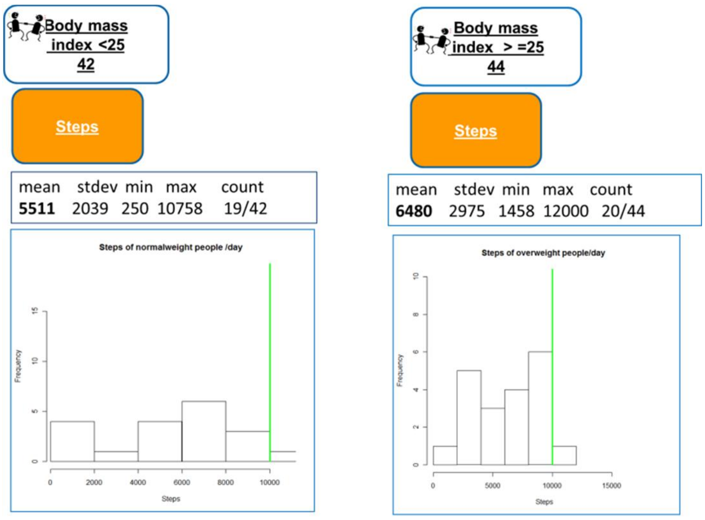


Figure 57. Subsets of body mass index.

Next, the user wants to study the differences in number of daily steps, exercise duration and consumed kcals. The user picks the corresponding variables onto the work area and plots the basic statistics and histograms. An example is shown in Figure 58. Surprisingly, overweight people take more steps than normal-weight people.



mean	stdev	min	max	count
5511	2039	250	10758	19/42

mean	stdev	min	max	count
6480	2975	1458	12000	20/44

Figure 58. Daily steps of normal and overweight people

The exercise results are summarized in Table 4. Exercise of normal and overweight people. No other dramatic differences were found from this data set. Normal-weight people seem to exercise slightly longer and expend slightly more kcals during exercise than overweight people.

Table 4. Exercise of normal and overweight people

Exercise measurement	Unit	Mean, normal-weight	Mean, overweight
Steps	daily steps	5511	6480
Exercise duration	min	49	42
Exercise kcals	kcal	333	326

Users can similarly study differences in food constituent consumption, as shown in Table 5. Food consumption of normal and overweight people. Both groups seem to have quite similar eating habits. A surprise finding was the level of sugar consumption, with normal-weight people found to consume more sugar than overweight people.

Table 5. Food consumption of normal and overweight people.

Food constituent	Unit	Mean consumption, normal-weight	Mean consumption, overweight	Recommendation /day
Fat	E%	32	34	25-35 E%
Sugar	E%	17.5	17.3	max 10 E%
Fibre	g	27.1	28.0	25-35 g
Protein	E%	17.7	18.5	10-20 E%
Carbohydrates	E%	47.7	44.0	55-60 E%

9.5 User evaluation

The objective of the user evaluation was to determine whether users were able to find interesting observations and insights, both expected and unexpected, from the data, and whether the constructed application was usable and useful. The specific evaluation questions were:

- Is the data model a good starting point for gaining an overview of the data?
- Is the guidance obtained from the data model understandable and useful?
- Is the set of visualizations that the system provides sufficient and easy to interpret?
- Is the user able to continue the analysis using the visualizations by selecting subsets and adding new data objects to the analysis? and
- Do the users find useful information, expected and unexpected, and get new insight with the help of the application?

9.5.1 Test settings

Evaluation method

The selected evaluation method was a combination of a controlled insight-based method and traditional usability testing. The users performed predefined test tasks in a controlled environment, thinking aloud during the testing. The tests tasks were planned based on the principles defined by North (2006), according to which the user is allowed to explore the data in a way that they choose, starting with the help of initial questions. Insights are captured throughout the session, along with traditional usability metrics.

Three test sessions were arranged: (1) traditional “thinking aloud” usability test, (2) pair testing where feedback was gathered from the discussion between the test persons, and (3) a pilot test to verify the test settings. The test results of the pilot test were used only to develop the actual tests and they are not included in the results.

Users

The test users represented both domain experts and semi-experts. One was a professional nutritionist; two were researches of user experience and one a casual user interested in nutrition. A summary of the test persons is in Table 6.

Table 6. Test users.

Test	Gender	Age	Role in testing	Experience
Test 1	Female	62	Professional nutritionist, specialised in counselling	Familiar with nutrition science and the original HyperFit system
Test 2	Female	40	User experience (UIX) researcher	Familiar with UX research
	Male	37	User experience researcher (UIX) and multivariate data analyst	Familiar with UX research and multivariate data analysis
Pilot	Female	39	Casual user	Interested in nutrition

Test environment

A paper prototype was constructed. The data model and the work area were represented by laminated paper sheets. The data objects consisted of laminated pieces of paper that the user could pick from the model and place in the work area. The suggested analysis alternatives also consisted of laminated pieces of paper that the test moderator added to the work area as a computer would have done. Based on the user's choice, the moderator added the pre-calculated visualizations and analysis results to the work area. A rubber band represented the subsetting tool. The test sessions were recorded. Figure 59 shows the test setting.

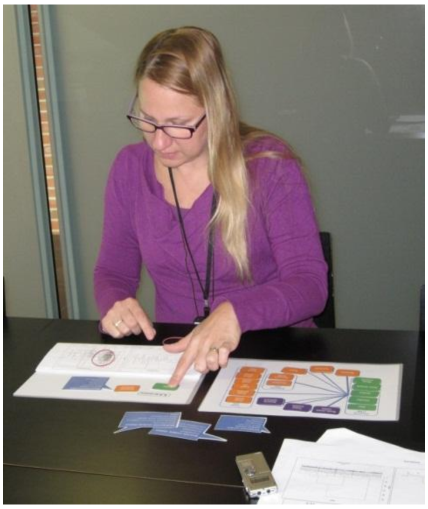


Figure 59. Test setting.

Flow of the test

The test session started with an introduction to the test process. The user was asked to fill in a questionnaire on the user's background. It was confirmed to the user that the purpose was to test the concept, not the user. Permission to record the test was requested.

The user interface was introduced by means of a brief example. The testing scenario was introduced and the test tasks given. The user was encouraged to think aloud when performing the tests.

The tests were not repeated exactly in both test rounds. After the test with the professional nutritionist a couple of corrections were made to the data. The absolute grams of the food constituents were converted to energy percentages (E%). This had a slight effect on the analysis results and on the comments and conclusions of the users in the second test. Also the class limits for overweight and normal-weight people were redefined.

After performing the test tasks, the users were asked a couple of interview questions (Appendix C) to obtain qualitative information. Finally, the user was presented with a small gift for their participation. Each test took approximately 1.5 hours.

Test scenario and tasks

The test scenario was

> "Imagine that you are a nutrition researcher and you have been given this interesting database and a new tool to study it. You have all kinds of questions, beliefs and suspicions that you want to get answers to."

The users were given five tasks to complete. These included free browsing, finding answers to questions, looking for explanations, and confirming beliefs and suspicions. These required studying basic statistics, time series, histograms, and correlations between two and more variables, scatterplots, selecting subsets from visualizations and continuing the analysis with subsets. The tasks were:

1. Browse and familiarize yourself with the data.

   Examine the background variables and indexes to see what kinds of people are dieting. Did you discover anything of interest?

2. What is the relationship between feelings and dieting?

   Study the correlations between feelings and the indexes: body mass index, food quality and dieting balance. Did you notice anything of interest? What would you like to examine further?

3. Find explanations for overweight.

Have a look at body mass indexes. Form subgroups of overweight and normal-weight people (index value over/under 25). Do they exercise differently? Compare steps, exercise duration and kcals. Do they eat differently? Compare the fat, sugar and fibre intake of both groups. What do you find?

4. Is eating less carbohydrates a good way to diet?

   Examine the correlations between carbohydrates and the weight index and dieting balance index. What do you notice?

5. Examine the correlation between food constituents.

   What is the biggest source of energy? Try omitting this from the analysis and see what happens.

Metrics

The testing focused on measuring user satisfaction, drawn from the observations and interviews and the insights gained. The individual metrics that were used were the number of comments indicating overall satisfaction, usefulness and ease of use, the number of assists by the moderator, indicating ease of use. Insights, "aha experiences", were calculated from the amount of findings, expected and unexpected. In addition, the number of deductions made based on the findings and new ideas of how to continue examining the data were also calculated.

The typical quantitative metrics for efficiency and effectiveness, such as completion rate and fail rate, were not used here as the main emphasis was not on the efficient use of the user interface – especially when there was not an implemented user interface in use – but in finding interesting information using the tool.

Analysis of results

For analysis of the results, the session recordings were transcripted. The findings related to the predefined metrics were then identified from the transcriptions and interviews and documented as Excel sheets. One sheet documented user satisfaction and contained categorized expressions of overall satisfaction, usefulness, ease of use and assists. Another sheet documented each task, categorizing expected findings (user expected), unexpected findings (surprise for user), conclusions (this means that…) and ideas for continued action (next I would like to…).

9.5.2 Results

The main objectives of the user testing were to determine whether the constructed application is usable, useful, understandable, and sufficient for performing analyses, and whether it helps the user to find interesting things from data. The test results revealed a working user interface and satisfied users. After brief initial uncertainty, users quickly learned to use the user interface independently to make findings and draw conclusions.

The test results are presented in detail below, classified as overall satisfaction, usefulness, ease of use and insights.

Overall satisfaction

The comments related to overall satisfaction were positive: “It was fun”, “Positive experience”, “Really interesting”, “Clearly positive”, “A really interesting tool”. Specific comments were also given related to the data: “This data is interesting” “This food is interesting”. The appearance of the user interface was also liked: “Easy, visually successful” “Nice graphical layout, variables visible”. No negative comments were given.

Usefulness

All users found uses for the tool. According to the professional nutritionist, its primary use would be in nutrition research. She did not consider the tool to be very useful in counselling individuals, but possibly for counselling groups. The UX researchers considered the tool to be suitable for multivariate analysis and for showing and studying the results. The data analyst researcher considered possible uses for the tool for data analysis in user experience research.

Ease of use

The tool was considered to be simple and straightforward to use. Initially, the users experienced a degree of uncertainty, but quickly learned their way around the user interface and the analysis process: “Easy tool; a bit confusing at first – but it soon became clear during the first two tasks”

Selecting objects for analysis, selecting analysis and visualizations methods, and selecting subsets with the “rubber band” were considered easy and fun: “Variables visible and easy to select to the work area; same applies to the set of analyses”

Users quickly developed their own ideas of how to progress further, select new subsets, and study new correlations with other variables. Unfortunately, the paper prototype, with its limited number of pre-processed analyses, restricted the users’ curiosity.

Basic statistics with histograms, networks of correlations and scatterplots were the most used visualizations. Histograms were the most familiar and easy to interpret. The users studied the shapes of the histograms, the tails, and individual instances. Interpreting correlations proved less straightforward, although the use of scatterplots helped in this respect. The users studied patterns and outliers in the scatterplots with interest.

Interpreting the visualizations was easy for the nutritionist, for whom the data domain was more familiar. The other test persons felt that more nutritional background knowledge would be useful to interpret the results. The non-professionals had difficulty interpreting negative correlations, while the nutritionist experienced no problems in this respect.

Assistance was needed to explain concepts (e.g. dieting balance) and the use of certain features, and when viewing a visualization for the first time. In the first test assistance was needed ten times, in the second eight times.

The paper prototype does not give a full picture of the ease of use, but rather introduces the idea. Preserving ease of use in actual implementation requires special attention. Two improvement areas were highlighted by the users: firstly, the addition of an overview picture of the whole dataset and, secondly, a help service to explain the meaning of different analyses and nutritional concepts.

Insights

In both tests the users made numerous findings, both expected and unexpected, and were able to draw conclusions based on them. Table 7. shows the numbers of insights, unexpected findings, conclusions and ideas for further analysis.

Table 7. Insights.

Test	Insights	Unexpected findings	Conclusions	Ideas for further analysis
Test1	15	5	8	5
Test2	20	1	11	5

Examples of the main findings and conclusions from the task are presented below.

Who is dieting?

> "This is funny, why are the normal-weight people dieting?"
> "Surprisingly many men, quite unexpected!"
> "A surprising number over thirty and middle-aged people"
> "One anorexic, 46 kg!"
> "Mainly light work"

How do they eat?

> "People seem to eat fiber quite well, unbelievable good quantity!"
> "Looks like people eat too much fat."
> "But people eat terribly much sugar!”
> "No extremes in food quality, goes up and down, getting worse towards Christmas”
> "Strongest sources of energy are sugar and fat, biggest negative is fibre"
> "So switching to a sugarless diet means most energy comes from fat?"
> "Fibre seems to be an independent variable"

How do they exercise?

> "Unbelievable, they don't exercise at all, just walk for half an hour"
> "Overweight burn more energy during exercise, normal-weight exercise longer. Maybe the overweight try harder while normal-weight people exercise more steadily"
> "The overweight take more steps than the others, they try to lose weight by walking"
> "This one's a sports addict"

How do they feel?

> "Between two and three, a bit on the positive side".
> "One seems to be quite low."
> "Negative correlation with dieting balance, but very small, nobody seems to eat to sorrow"
> "Negative correlation with food quality, when food quality improves feelings deteriorate – no wonder, it means stress"

How's the diet?

> "People seem to gain more weight than lose"
> "Looks like people burn too little energy and eat too much"
> "No real zero-carb people here, i.e. getting less than 30% from carbs; can't say based on this data whether cutting out carbs is beneficial"

Ideas for continuing the analysis

In both tests users came up with new ideas about what to examine. These included other interesting subsets, searching details, using different time granularities, finding effects of changes: "It would be interesting to know what foods people are getting most fibre from, bread or vegetables?", "As food quality changes, feelings change – what happens here? What do people eat more or less of".

9.5.3 Answers to the evaluation questions

Based on the results, the following answers to the evaluation questions can be suggested:

Is the data model a good starting point for gaining an overview of the data?

The model does not provide an explicit overview of the data, rather it provides a good starting point for analysis and selecting variables. It shows the variables that are available and enables their easy selection. Another kind of overview could be considered.

Is the guidance obtained from the data model understandable and useful?

Users were content with the analysis method suggestions. For a user familiar with data analysis the current model would be sufficient, but for others some additional help with the method descriptions might be useful.

Is the set of visualizations that the system provides sufficient and easy to interpret?

The visualizations used, namely basic statistics, histograms, time series, scatterplots and correlation networks, were relatively easy to interpret. The users could study the visualizations and focus on interesting patterns and details. For the person familiar with the domain the visualizations revealed more than for the others, and she was able to make the interpretations more quickly. The only problematic issue was interpretation of negative correlation.

Is the user able to continue the analysis using the visualizations by selecting subsets and adding new data variables?

Yes, these were easily learned and much liked features.

Do the users find useful information, expected and unexpected, and get new insight with the help of the application?

Yes, the analysis confirmed users' assumptions, brought to light numerous findings, both expected and unexpected, and helped to gain new insight.

9.6 Discussion

In the application construction, extracting data, cleaning and harmonizing took most effort. Data extraction proved out to be complicated. For example, getting the daily kcal intake of each person from the HyperFit database food diaries of dieting people included several steps. The database does not store the required attribute. Instead, there is a table showing what food products each person has eaten, at what time of day, and how many portions. In order to determine the daily kcal intake, a chain of processing steps is required. Firstly, for each product eaten the product category and portion size in grams must be determined. Next, the product information must be examined to determine the number of kcals per 100g of the product. Even then, the kcal information is not directly accessible. First, one has to find the code that represents kcals and search for the value corresponding to this code from a separate nutritive table. The kcal intake from this product can then be calculated. After summing all products eaten by each person during the day, the information is available for visualization and analysis.

Another time-demanding step was deciding what variables the user would be interested in, and in what form the variables should be shown. A domain expert would be useful in this step. Furthermore, a domain expert would be useful in defining the indexes. In this case, only the body mass index is an official index; the other two were constructed specifically for user testing in the present study. As the user interface was not implemented in practice, one can only speculate as to what its actual implementation might entail.

The clear benefit of this approach was that once the data and indexes were entered into the database, using the analysis and visualizations was straightforward.

These results of the user evaluations suggest that the constructed system fulfils the expectations well. They also suggest that the system would be most beneficial to domain experts who are familiar with the domain but not with data analysis. This kind of tool could free such experts to focus on the data instead of pondering what analysis methods to use or seeking the assistance of a data analyst.

10 Case MMEA

The focus of the MMEA case was to determine how well the concept fits with other kinds of data, and on building an application-specific user interface. A visual analytics tool was constructed to analyse the energy efficiency and indoor conditions of buildings. The user interface included descriptive visualizations of the objects of interest. In this case, no user evaluation was performed, only self-evaluation of experiences of constructing the application.

MMEA Indoor is an on-going pilot project carried out as part of the MMEA (Measurement, Monitoring and Environmental Assessment) research programme under the Finnish Cluster for Energy and Environment (CLEEN)[32]. The MMEA develops applications for environmental and measurement data. The data is accessible from the MMEA Platform, which serves as a “market place” for environmental and measurement data. Data collected from sensors and databases is delivered to applications in a unified format by means of a message service provided by the platform. The focus of the MMEA Indoor pilot is energy efficient indoor environments. The project develops tools for building users and operators to enable better control of their indoor environments, aiming at energy efficient, comfortable and productive conditions.

This chapter begins with an introduction to the MMEA visual analytics tool. Next, the tool generation is presented step-by-step. The user interface is then introduced and examples are given of how reasoning is carried out with the tool. The chapter ends with a self-evaluation and discussion on the experiment.

10.1 MMEA visual analytics tool

The MMEA tool is a visual analytics application constructed based on the principles described in chapter 7. The tool is intended for use by specialists for problem identification and diagnostics. Users can explore and navigate the data, launch analyses and visualizations, request details and examine interesting subsets. The intended user is an expert or researcher in building indoor conditions and energy usage. Casual web users could also use the tool to check indoor quality and energy efficiency information. The knowledge obtained could be used to produce alerts and warnings and to control building systems.

The monitored objects are office buildings in the research campus area in Otaniemi, Finland. The data consists of measurements of power, reactive power, district heat and water consumed by each building, building gross volume, gross floor area and building age, indoor conditions of the rooms including (from one building only) room temperature, CO2 and space occupancy, and environmental measurements including outdoor temperature and relative humidity of the campus area.

The measurements are stream data with an accumulation rate of one measurement per hour. In the case study a snapshot of one week (January 17 – January 23, 2011) was

[32] CLEEN = Cluster for Energy and Environment, a company managing the Strategic Centre for Science, Technology and Innovation for energy and environment in Finland

used. The used data consisted of information on 24 buildings, including 2,856 power measurements, 2,109 reactive power, 678 water and 2,016 district heat measurements, the indoor conditions of one building, including room temperature of 217 rooms (36,456 entries), occupancy of people from 255 rooms (42,840 entries) and CO2 measurements of 7 meeting rooms (1,176 entries), and temperature and humidity measurements for the Otaniemi campus area (186 entries).

The system calculates indexes for the energy performance of buildings and indoor conditions of building spaces. The set of visualizations and analysis methods includes basic statistics (min, max, mean, standard deviation), histograms, time series with different time granularities (hour/day/week), clustering, correlations, scatterplots, linear regression, auto-correlations and cross-correlations.

In this case the user interface was constructed using descriptive visualizations of the objects of interest. A map of the Otaniemi campus area and an IFC[33] building model of one of the buildings were used.

The tool was implemented as a mixed working and paper prototype, similarly to the HyperFit. Data management and analysis was implemented with the R Stats package, while the user interface was simulated with a paper prototype.

10.2 Creating the application

The application construction followed the process steps defined in Section 8.4. Further details are provided in Appendix D.

Step 1: Define object of interest types

Three object of interest types are used: geographical area, area buildings and building spaces. These form a hierarchy: building spaces belong to buildings, which belongs to a geographical area.

Step 2. Define properties

The background properties of each object of interest type are (1) Buildings: Gross volume (m3), gross floor area (m2) and age (years); (2) Building spaces: purpose of use (meeting room / office room).

Measured area properties are temperature (°C) and humidity (%). Building measurements are power kWh, react power (kVARh), water consumption (m3) and district heat (kWh). Building space properties are indoor temperature (°C), CO2 (ppm) and space occupancy (occupied/not occupied). A detailed definition including types, units, codings, limits and frequencies is given in Appendix D.

Step 3. Define indexes

Two kinds of indexes are used: (1) Building energy performance index: formed by dividing the building energy consumption by the building volume; and (2) Indoor air quality index: calculated using the classification of indoor climate by the Finnish

[33] Industry Foundation Classes (IFC),an open, neutral and standardized specification for Building Information Models (BIM), (buildingsmart.com/standards/ifc) (accessed 18 December 2012)

Building Information Foundation (RTS)[34]. There are three classes indicating the quality of indoor conditions: S1 individualized, S2 good, S3 satisfactory. The class is determined by the room temperature, room humidity and outdoor temperature. Both indexes are calculated per hour. A detailed definition of indexes with types, units, codings, limits and frequencies is given in Appendix D.

Step 4. Select visualization and analysis methods

Analyses and visualizations for single variables include basic statistics (mean, standard deviation, min, max), histograms, pie charts, clustering, time series (all measurements, means/sums of a time point), and auto-correlations. Methods for bivariates include correlation coefficient, scatterplot, linear regression, cross-correlations; and for multivariate data correlation matrix, correlation network, residual plots and PCA. A detailed table of analysis and visualization methods used is given in Appendix D.

Step 5. Map to original data

In this case there is no need for mapping as the measurement data comes from the MMEA Platform. Pre-processing is performed already on the platform.

Step 6. Add application-specific features

Application related data is added, including details of the buildings and building spaces: the names and addresses of the buildings, building coordinates, purpose of use of a building space, and the location of the space in the building model.

10.3 The user interface

The user interface is more complicated than in the HyperFit case, as there are now several objects of interest forming a hierarchy and descriptive visualizations. A map of the area and a building IFC model is used in the user interface to locate the buildings. Interacting with the map and the building model requires application-specific implementation.

The area map or building model was used as the starting points for navigation. Menus were used as an alternative way to navigate data, especially to support navigation in hierarchies. The data model was not used in this case, although the model could be added as an alternative means of browsing the data.

This section introduces the user interface layout, how the reasoning tasks are performed and how direct manipulation, coordinated multiple views and brushing work.

[34] https://www.rakennustieto.fi/index/english.html (accessed 12 March 2012)

Layout

Figure 60 represents the user interface layout. The surface is divided into three sections. From the left area the user can select objects of interest for analysis, including the whole area, buildings and building spaces, and their properties. The right side shows the user selections and the analysis suggestions. The middle area is for the descriptive visualizations and the analysis results. The descriptive visualizations and analysis results are interlinked so that highlighting an object or area is shown in both kinds of visualizations, if possible. The user can select subsets of data for further analysis from both the descriptive and abstract visualizations.

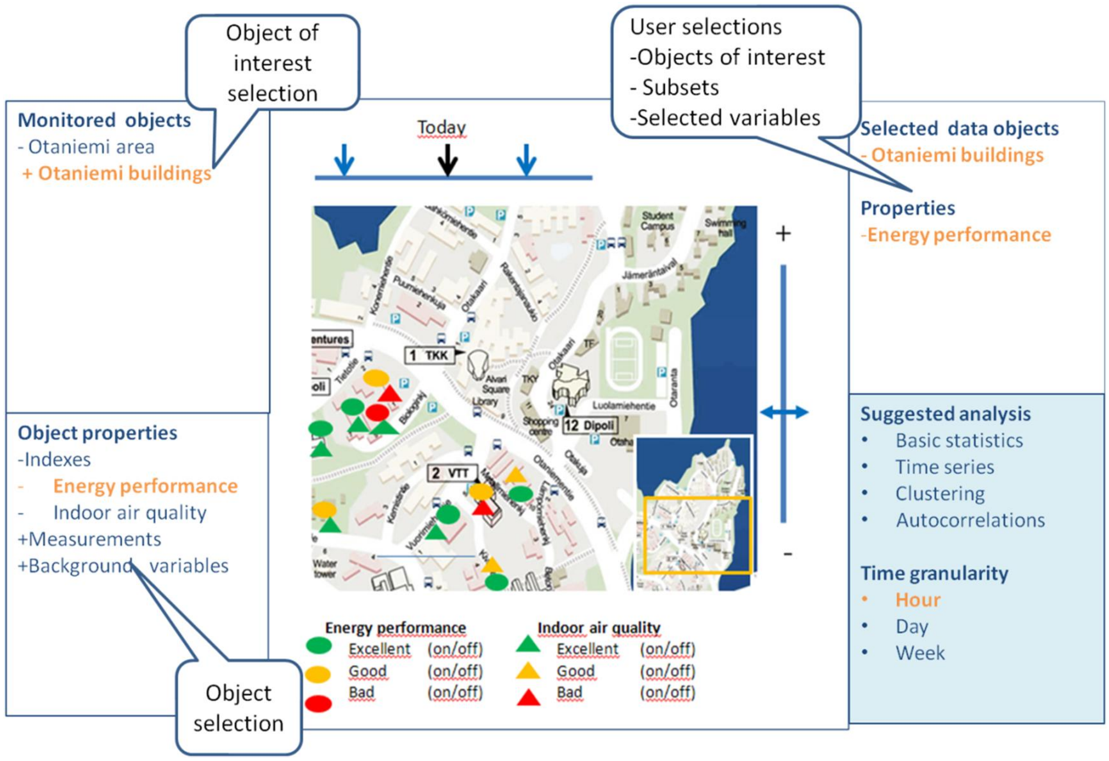


Figure 60. User interface layout.

Giving overview

The overview shows on the descriptive visualization the status of the indexes for the object of interest that the user selected from the hierarchy. Figure 61 shows an overview of all Otaniemi buildings, and Figure 62 one Otaniemi building. The building indexes are visualised using different symbols and colours. By clicking the index symbols the user can quickly get the basic statistics, histograms, time series and raw data.

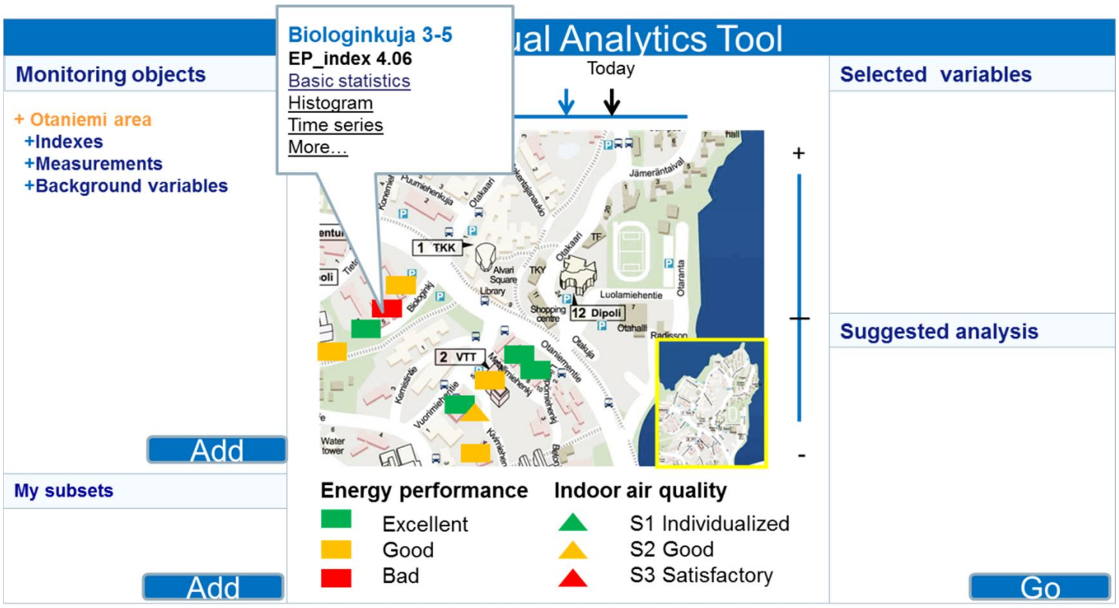


Figure 61. MMEA Otaniemi overview.

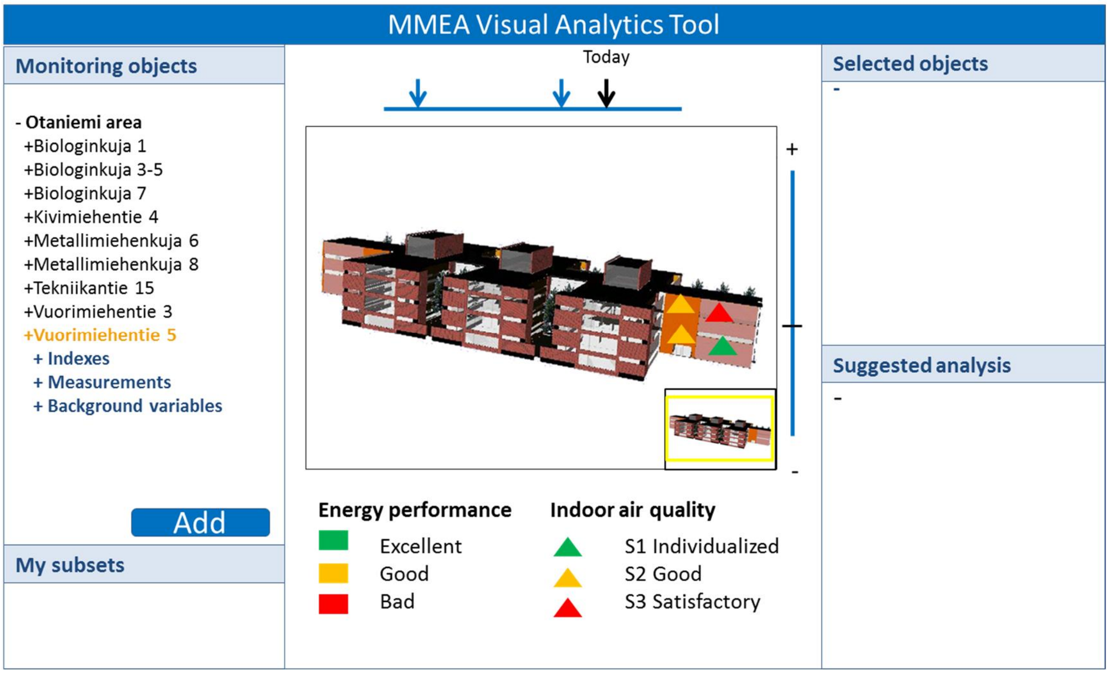


Figure 62. Digitalo overview.

Forming datasets for analysis

The user forms the dataset for analysis by selecting the objects of interest and their properties from the menus on the left. The resulting selections are shown on the right under Selected variables. For example, in Figure 63 the user has selected one building and two variables for analysis. The user can select several variables from one or more objects of interest. The time scale tool allows the user to adjust the time period. The selection of variables is not described here in detail as it is a basic user interface element and is not essential here.

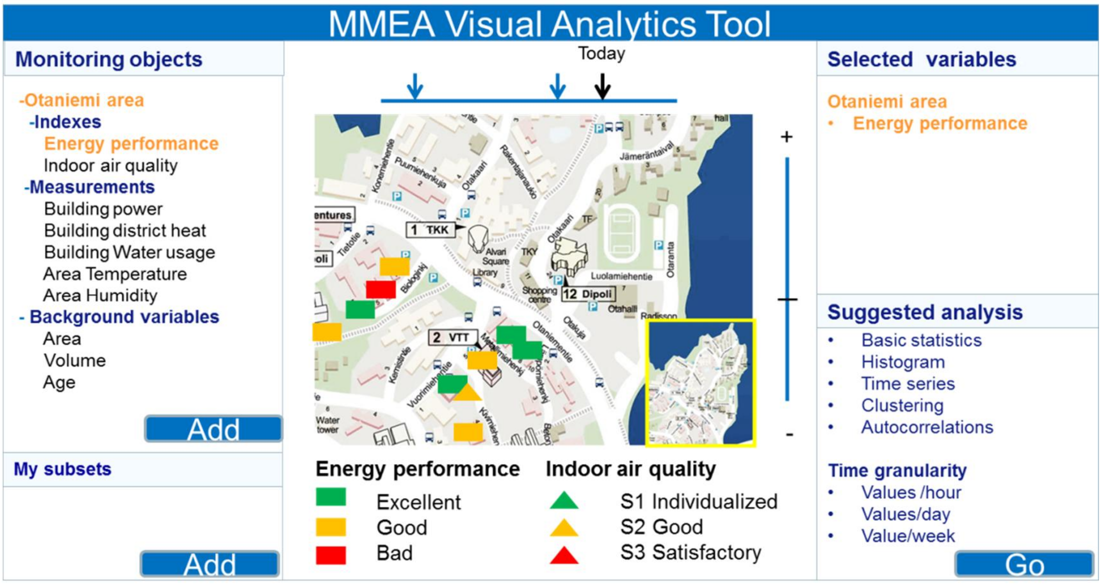


Figure 63. Selecting variables.

Selection of the analysis and visualization method

After the dataset is selected, the system suggests suitable analyses for the user. The user selects the variables by clicking the variables and receives the suggestions. The user can also adjust the time granularity of the analysis (hour/day/week). In the example in Figure 64 the user has selected three variables and received suggestions for multivariate analysis.

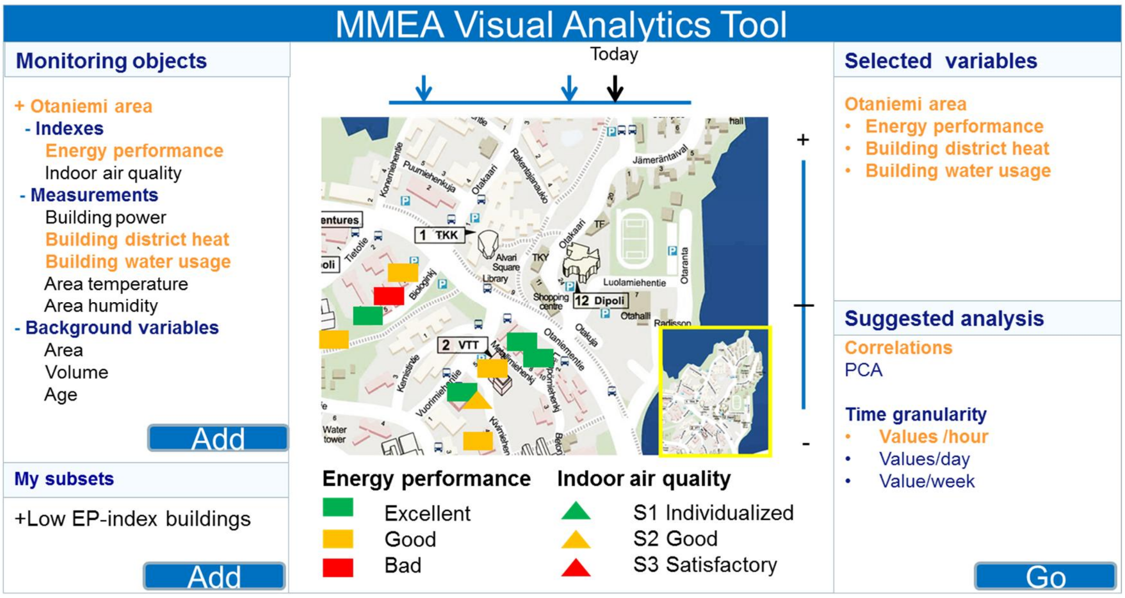


Figure 64. Analysis suggestions.

Showing the analysis results in visual form

The results are shown in the middle area in independent windows similar to the HyperFit case (Figure 65). The descriptive visualization remains in the background.

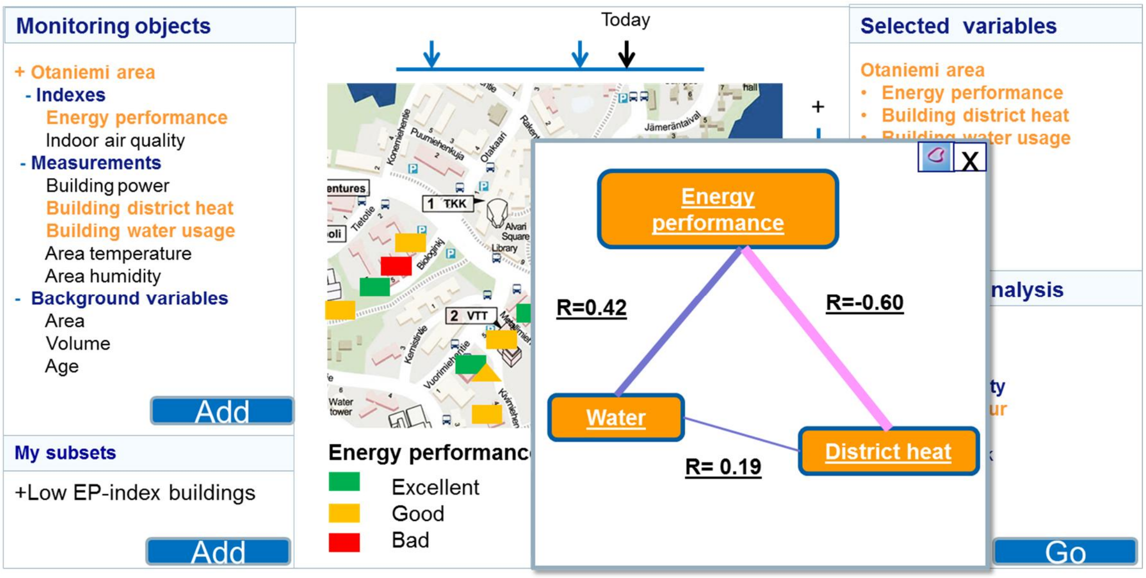


Figure 65. Analysis results.

Picking subsets

The user can pick subsets in a similar way to the HyperFit tool. The selected objects are also highlighted on the descriptive visualization (Figure 66).

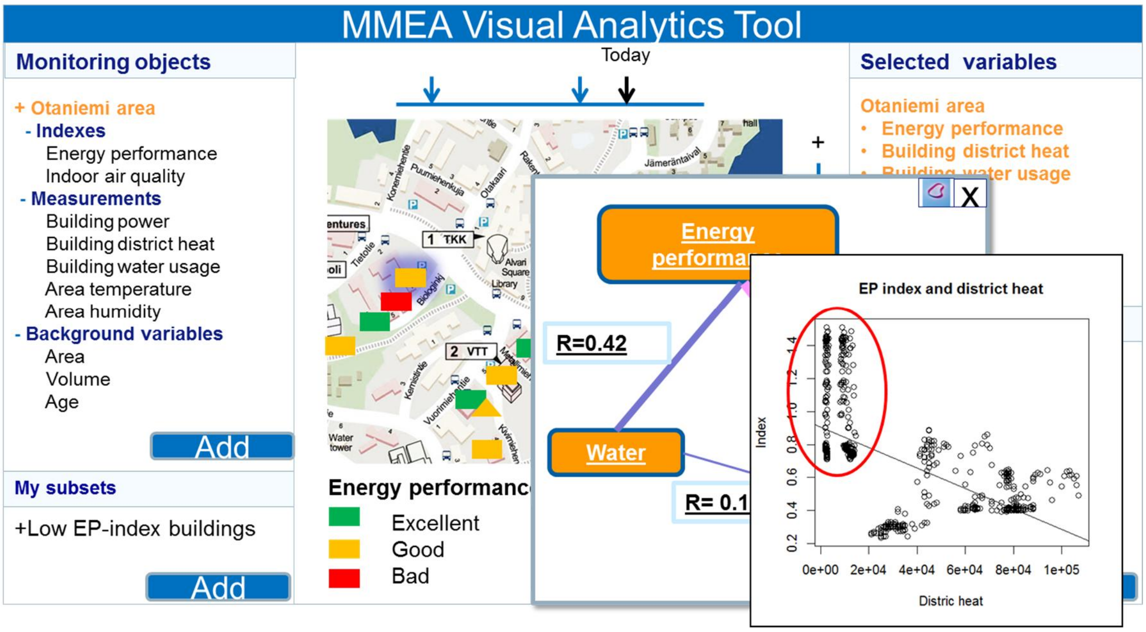


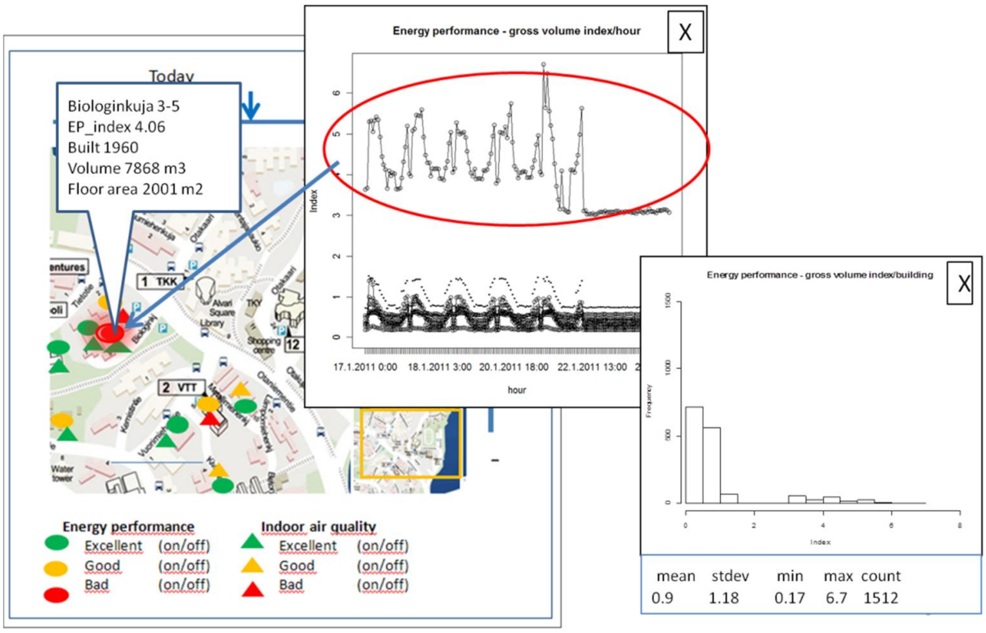
Figure 66. Selecting subsets.

Brushing and details on demand

Brushing is performed in the same way as with the HyperFit tool, but the selected objects are also highlighted on the map visualization. The user can get details from both descriptive and abstract visualizations.

10.4 Reasoning example

Let us assume that the user is interested in the energy efficiency of the campus area buildings. First, the user views an overview of the energy efficiency of the buildings on the map. The different colours indicate the index values of each building, showing a rough estimate of the energy performance of each building. To get a better overview, the user wants to see the basic statistics and time series (Figure 67). These show two interesting things: most buildings have a low energy performance index (i.e. low energy usage), but some have very high values. Naturally, energy usage changes considerably at weekends and during day and night time. To link the analysis results to the actual buildings, the user selects the area of the visualization in question to highlight the corresponding buildings on the map.



Biologinkuja 3-5
EP_index 4.06
Built 1960
Volume 7868 m3
Floor area 2001 m2

mean	stdev	min	max	count
0.9	1.18	0.17	6.7	1512

Figure 67. Basic statistics and time series.

Next, the user wants to determine possible reasons for the high variance in energy performance index values. The first idea is to compare building age with the index values. A scatterplot showing correlation coefficient and regression line gives an indication that there is some correlation between age and energy performance index (Figure 68).

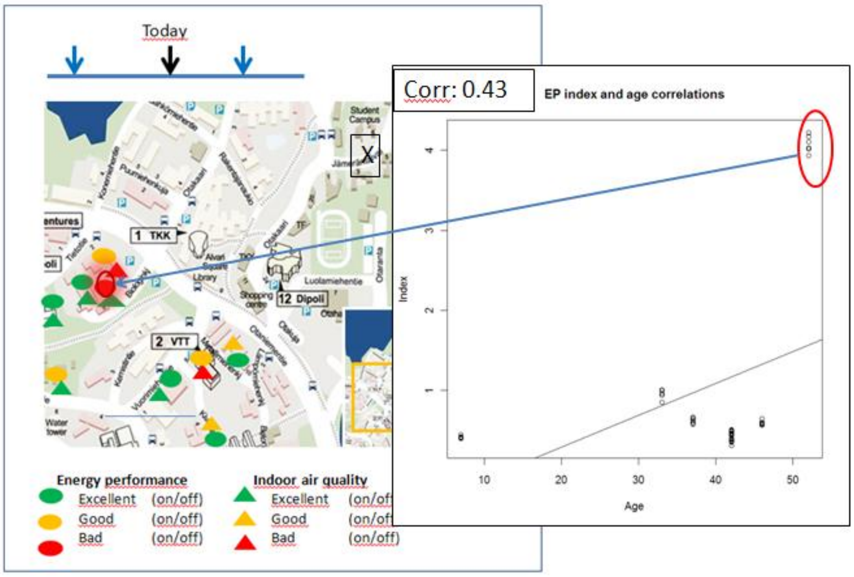
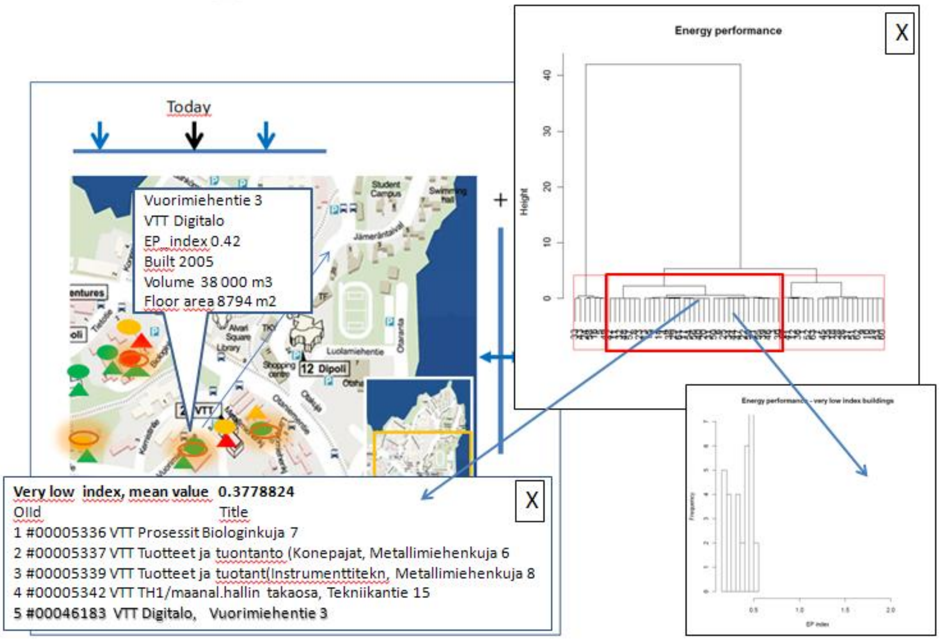


Figure 68. Correlation with age.

The user can select an object from the scatterplot and view the corresponding buildings on the map. In the example, the user is interested in plots that indicate a high index value and high building age. These are shown to come from a building that is highlighted on the map. The user can also select data subsets of visualizations and perform further analysis with these subsets. In the example in Figure 69 three clusters are formed based on the indexes, and the statistics and details of each cluster are shown.



Vuorimiehentie 3
VTT Digitalo
EP_index 0.42
Built 2005
Volume 38 000 m3
Floor area 8794 m2

Very low index, mean value 0.3778824
OIId Title
1 #00005336 VTT Prosessit Biologinkuja 7
2 #00005337 VTT Tuotteet ja tuontanto (Konepajat, Metallimiehenkuja 6
3 #00005339 VTT Tuotteet ja tuotant(Instrumenttitekn, Metallimiehenkuja 8
4 #00005342 VTT TH1/maanal.hallin takaosa, Tekniikantie 15
5 #00046183 VTT Digitalo, Vuorimiehentie 3

Figure 69. Analysis of subsets.

The second example is of a multivariate analysis aimed at studying correlations between energy performance index, district heat and water consumption (Figure 70). The example shows a strong correlation between district heat and energy performance. The user wants to examine this correlation further by producing a scatterplot. The scatterplot shows an interesting pattern, which the user also wishes to investigate further. The pattern proves to represent a building, which is shown on the map.

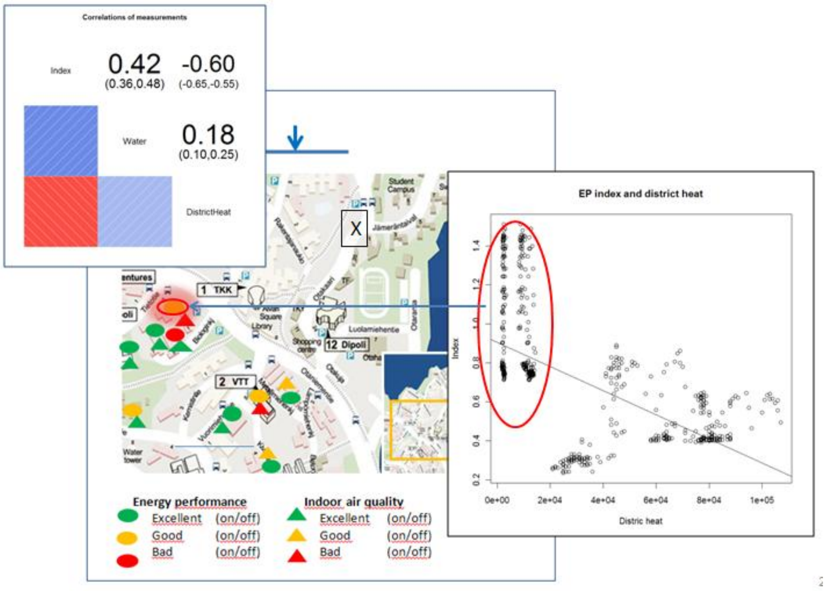


Figure 70. Multivariate analysis example**.**

10.5 Discussion

Two conclusions can be drawn from the experiment. The first relates to the construction of the application, and the second to the user interface including application-specific descriptive visualizations.

Constructing the application

Applying the step-by-step generation of the application was straightforward and did not require changes to the concept. However, identifying the objects of interest, in other words the objects to be monitored, might be problematic for the user. In the example case, the MMEA project had done the preparatory work and identified these beforehand. The same problem applies to the properties and indexes. Choosing which variables and indexes to form and monitor requires prior examination and decision making. In the example case the variables and indexes were specified by domain experts. This suggests that, as with any other software project, a requirements analysis phase conducted by domain experts is necessary before implementation.

The set of visualization and analysis methods was feasible and did not require modifications. Mapping to the original data required minimal effort as the data was already in the right format, coming from the MMEA platform. Only implementation of the indexes required coding.

In this case, too, the user interface was not fully implemented in practice, and thus one can only speculate as to what its actual implementation might entail. The hierarchy of objects of interest and the use of descriptive visualizations both bring added complexity to the user interface compared to the HyperFit tool.

The experiment suggests that applications can be constructed in this way inside the domain.

User interface

The main difference compared to the HyperFit solution was the use of descriptive visualizations. This was possible because the objects of interest were the kinds of objects that could be positioned on maps and construction models. Informal demonstrations suggested that using the descriptive visualizations was helpful by providing further context to the findings. This suggests that combining different kinds of visualizations can result in more powerful reasoning tools than using only abstract visualizations. This is possible only in cases where natural descriptive visualizations exist.

11 Summary and conclusions

This work examined the possibilities and advantages of utilizing data models in visual analytics. Data modelling is an established method in software engineering, but not common practice in information visualization and data analysis. Visual analytics is a recent approach for obtaining knowledge from data masses. The approach has its roots in information visualization and data analysis and combines the strengths of automatic data processing and the visual perception and analysis capabilities of the human user.

This work introduced the concept of visual analytics, the state of the art of visual analytics research, and commercial markets. The building blocks of this multi-discipline research area were represented in the form of a framework constructed as a part of this work. In addition, the methods for evaluation of visual analytics were reviewed. The two other key elements of this work, data modelling and monitoring data, were also introduced.

The backbone of this work is the domain data model. The model incorporates the main concepts of a given domain, which remain similar regardless of the application, but which can be tuned for visualization and analysis purposes. Three uses of the domain model were studied. The first was using the domain model in the construction of visual analytics applications. The second was to support reasoning by giving suggestions of suitable visualizations and analysis methods with the help of metadata added to the model. The third was using the data model as a user interface element, offering a new approach for navigation in large collections of data. To keep the work within reasonable limits the study focused on the monitoring domain.

A concept for a visual analytics tool for analysing monitoring data was drafted to study the applicability of the domain model approach. The concept includes a reasoning process, a user interface, and a set of visualization and analysis methods for monitoring data. In addition, the process of creating a new application within the domain was defined. The concept was evaluated using two cases from different application areas within the monitoring domain. The first was HyperFit, a visual analytics tool intended for professional nutritionists to support research on dieting and exercise habits. The other was the MMEA indoor pilot for studying energy efficient indoor environments. The purpose of the evaluations was to determine whether users can reason and gain insight using this kind of tool and determine the applicability of the concept to other kinds of monitoring data.

Discussion

The suggested approach utilized several the results and methods introduced in visual analytics framework and also brought new contribution to the area.

Figure 71 illustrates these aspects. The area to which this work brings the greatest new contribution is infrastructure. The work proposes a domain specific infrastructure, based on the domain data model. The model forms a standard template for storing data and constructing applications in the domain. User interaction is another area of contribution. Using metadata to support reasoning, and the data model based approach to navigating data are new concepts in visual analytics. The reasoning process was constructed based on the existing approaches. The most influential of these was the process by Pirolli and Card (2005). The variety of visualization and analysis methods presented a range of potential methods for this work. The methods chosen for the experiments were data exploration and descriptive methods. The emphasis was on familiarity and being self-explanatory. The user interface techniques used were direct manipulation, coordinated multiple views, and brushing. The human perception results guided the interface design.

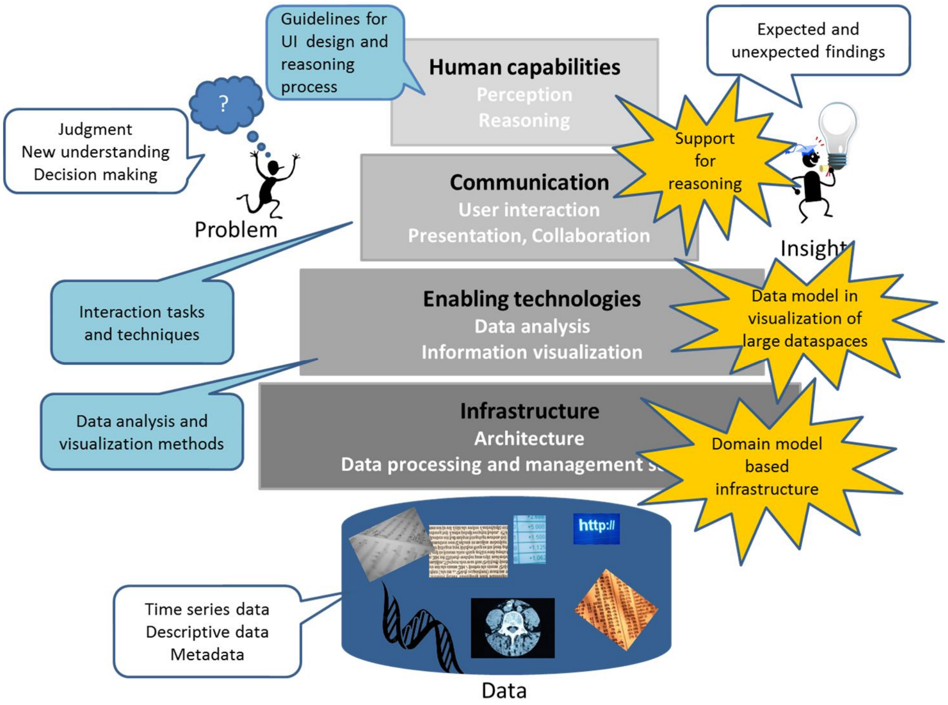

Guidelines for UI design and reasoning process
Human capabilities
Perception
Reasoning
Expected and unexpected findings
?
Judgment
New understanding
Decision making
Support for reasoning
Problem
Communication
User interaction
Presentation, Collaboration
Insight
Interaction tasks and techniques
Enabling technologies
Data analysis
Information visualization
Data model in visualization of large dataspaces
Data analysis and visualization methods
Infrastructure
Architecture
Data processing and management s
Domain model based infrastructure
http://
Time series data
Descriptive data
Metadata
Data

Figure 71. Relation of this work to the visual analytics framework. Blue bubbles represent utilized areas and orange stars contributed areas. The cornerstones, problems, data and insight, express the boundaries of this work. The problems were limited to the monitoring data domain which falls into the categories "making judgement on an issue based on evidence", "discovering new understanding" or "supporting decision making". This work considered static history data, added with metadata and descriptive data. Insight was understood as interesting findings obtained from the data, expected or unexpected.

A set of research questions was presented in the beginning of the study. The answers to these emerging from the study are discussed here in the form of a self-interview.

Are data models useful in visual analytics?

The results of this work suggest that they are. All three suggested uses proved to be useful. The data model based architecture provided a flexible mechanism for constructing new applications within a single domain. Support for reasoning, based on metadata, freed the user to focus on the domain instead of finding suitable methods for analysis. As regards the user interface element, the data model showed clear and understandable concepts to the user.

What kind of data model is useful?
A semi-generic domain model was suggested, being a general purpose data model for the whole domain. It includes the key concepts of the domain, those that remain similar from one application to another in the domain. The model is in a form compatible with visualization and analysis methods, enforced with metadata to be used to give guidance to users. The domain model preserved some semantics of the domain while providing a general-purpose approach.

What kind of visual analytic tool can be constructed based on the suggested model?
This work suggests one alternative with two different user interface solutions, a general purpose user interface and a user interface including descriptive visualizations of the objects of interest. Many other alternatives could, however, be implemented as the model sets no limits in this respect. The suggested approach comprised a reasoning process and provided a set of basic analysis and visualization methods that can be easily extended. The only requirement is that the methods are compatible with the data types of the domain model.

Does this approach help users find useful knowledge from the data?
The user evaluation with HyperFit indicates that this is possible. The user guidance proved helpful. Users did not need to have expertise in analysis methods, all that was required was the ability to interpret the visual results. Users familiar with the domain were more able to gain insight compared to non-experts. Guidance was given for the interpretation of the results by showing the limits and reference values for measurements. In the MMEA case, descriptive visualizations were shown to help the user understand the context of the results and could add further value to the data analysis. Adding more user guidance based on semantics or user action history offer one direction for future research.

How straightforward it is to apply the approach to a new case in the domain?
In the simplest cases this requires mapping the original data sources to the tool's data storage and implementing the calculations for application-specific indexes. In principle, nothing else is required. If an application-specific user interface is needed, extra implementation is required. Additionally, if new analysis and visualization methods are needed, these will need to be implemented and mapped to the data model. This does not, nevertheless, eliminate arduous data preprocessing and formatting. Domain expertise is important in the definition of the application content.

Can the results be generalized to other domains than monitoring data?
This question cannot be answered by the present study alone. Applying the results to some other domain that has a set of key concepts that remain similar from one application to another is a good candidate of for further research.

What are the advantages and disadvantages of this approach compared to other kinds of visual analytics solutions?
Four kinds of solutions were distinguished: Tailored solutions by a data analysis expert, commercial packages, application-specific solutions and the "domain approach" of this work. Compared to others the domain approach can provide an accurate and flexible solutions within the domain by capturing the special expertise of the domain. They can be reliable, stable, secure and interoperable, as those features can be provided as standard service. Tailored solutions may surpass the domain

approach in functionality, but usability, reliability, maintainability and portability may receive less attention. *Commercial packages* can be reliable, usable, and support security and integration well, but the functionality, efficiently and maintainability can be limited. *An application-specific solution* can provide an accurate and fine-tuned solution but only for this specific application.

Comparison to related work

The use of data models in visual analytics is not widespread. Only ten studies were found, the majority of which used data models to extract concepts from data and to create ontology models, while the rest used data models as part of system architecture in the form of either fixed application models, domain models or general-purpose models. The general-purpose model approaches generated independent applications with their own database schemas and concepts and required a tedious application definition and generation phase. Two domain model approaches were identified. One was used for handling work flows in publishing, and the other for monitoring structures (e.g. bridges) with sensors. The latter (Inaudi, 2002) was the nearest approach to the present study, although its focus was more on gathering sensor data than analysis. Combining the data models of the present work and Inaudi's approach could produce a powerful domain model for monitoring applications.

Only one example of metadata use was found, although in that case the metadata was used to keep track of the lineage between results and original data as opposed to guiding the end user. A data model or elements of such a model were used as user interface elements, although not in the exact way suggested in the present study.

Conclusions

Data modelling can be considered a useful method in visual analytics. The domain approach might facilitate the construction of new visual analytics applications in the domain. The suggestion of suitable analysis methods to the user can be helpful, especially for users that are unfamiliar with analysis methods but know the application domain well. Use of a data model as a user interface element shows the concepts clearly to the user and makes communication with the data easy. Combining abstract and descriptive visualizations can produce powerful tools for analytical reasoning. Such solutions could bring new alternatives to the data analysis market. These could be recommended for companies that need to analyse different kinds of data within a given domain, or for consultants that want to quickly construct analysis applications for customers.

The uses for data models suggested in this work hardly represent the only possibilities. Many important areas of visual analytics are not addressed in the study. It applies only a fraction of the overall potential of visualization and analysis methods. The study does not take into account complex data structures. In addition, special questions of scalability, streaming and big data sets are not addressed, and annotation, collaboration and presentation issues are not included. These limitations leave the path open for the future uses for data models.

References

Aigner, W., Miksch, S., Muller, W., Schumann, H. and Tominski, C. (2008) "Visual methods for analyzing time-oriented data", *IEEE Transactions on Visualization and Computer Graphics,* vol. 14, no. 1, pp. 47-60.

Aigner, W., Miksch, S., Müller, W., Schumann, H. and Tominski, C. (2007) "Visualizing time-oriented data--A systematic view", *Computers and Graphics,* vol. 31, no. 3, pp. 401-409.

Aigner, W., Miksch, S., Schumann, H. and Tominski, C. (2011) *Visualization of time-oriented data,* 1st edn., Springer-Verlag, NY, USA.

Alani, H., Dasmahapatra, S., O'Hara, K. and Shadbolt, N. (2005) "Identifying communities of practice through ontology network analysis", *Intelligent Systems, IEEE,* vol. 18, no. 2, pp. 18-25.

Alpaydin, E. (2004) *Introduction to machine learning,* 1st edn., MIT press, Cambridge, MA, USA.

Amar, R., Eagan, J. and Stasko, J. (2005) "Low-level components of analytic activity in information visualization", *IEEE Symposium of Information Visualization (INFOVIS) 2005*, eds. J.T. Stasko and M.O. Ward, IEEE Computer Society, Minneapolis, MN, USA, 23-25 Oct., pp. 111.

andrews.edu. (2005), *An Introduction to Statistics*, [Online]. Available:http://www.andrews.edu/~calkins/math/edrm611/edrm01.htm [2012, October 12].

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S.I., Jern, M., Kraak, M.J., Schumann, H. and Tominski, C. (2010) "Space, time and visual analytics", *International Journal of Geographical Information Science,* vol. 24, no. 10, pp. 1577-1600.

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M.J., MacEachren, A. and Wrobel, S. (2007) "Geovisual analytics for spatial decision support: Setting the research agenda", *International Journal of Geographical Information Science,* vol. 21, no. 8, pp. 839-857.

Azarm, M., Nargesian, F. and Peyton, L. (2011a), "Managing and mapping data lineage for business intelligence and analytics applications in health care", *2011 International Conference on Information Society (i-Society),* IEEE, London, UK, 27-29 June, pp. 120.

Azarm, M., Nargesian, F. and Peyton, L. (2011b) "Tool Support and Data Management for Business Analytics Applications in Healthcare", *International Journal for Infonomics (IJI),* vol. 4, no. 3/4, pp. 484-493.

Benyon, D., Turner, P. and Turner, S. (2005) *Designing interactive systems: People, activities, contexts, technologies,* Pearson Education, Essex, UK.

Booch, G., Rumbaugh, J. and Jacobson, I. (2005) *Unified Modeling Language User Guide,* 2nd edn., Addison-Wesley.

Bosch, R., Stolte, C., Tang, D., Gerth, J., Rosenblum, M. and Hanrahan, P. (2000) "Rivet: a flexible environment for computer systems visualization", *ACM SIGGRAPH Computer Graphics,* vol. 34, no. 1, pp. 73-82.

Bourke, P. (1996) *Cross Correlation* [Homepage of Paul Burke], [Online]. Available: http://paulbourke.net/miscellaneous/correlate/ [2012, October 12].

Brennan, S.E., Mueller, K., Zelinsky, G., Ramakrishnan, I., Warren, D.S. and Kaufman, A. (2006) "Toward a multi-analyst, collaborative framework for visual analytics", *2006 IEEE Symposium On Visual Analytics Science And Technology* IEEE Computer Society, Baltimore, Maryland, USA, Nov. 2, pp. 129.

Bull, R.I. (2008) *Model driven visualization: towards a model driven engineering approach for information visualization*, University of Waterloo, Department of Computer Science.

Cammarano, M., Dong, X.L., Chan, B., Klingner, J., Talbot, J., Halevey, A. and Hanrahan, P. (2007) "Visualization of heterogeneous data", *IEEE Transactions on Visualization and Computer Graphics,* vol. 13, no. 6, pp. 1200-1207.

Card, S.K., Mackinlay, J.D. and Shneiderman, B. (1999) *Readings in information visualization: using vision to think,* 1st edn., Morgan Kaufmann.

Catarci, T., Costabile, M.F., Levialdi, S. and Batini, C. (1997) "Visual query systems for databases: A survey", *Journal of visual languages and computing,* vol. 8, no. 2, pp. 215-260.

Chang, A. (2012) Oct-last update*, CUSUM explained* [Homepage of statTools.net], [Online]. Available: http://www.stattools.net/CUSUM_Exp.php [2012, Oct 18].

Chang, R., Ziemkiewicz, C., Green, T.M. and Ribarsky, W. (2009) "Defining insight for visual analytics", *IEEE Computer Graphics and Applications,* vol. 29, no. 2, pp. 14-17.

Chen, C. (2010) "Information visualization", *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 2, no. 4, pp. 387-403.

Chen, P.P.S. (1977) "The entity-relationship model: a basis for the enterprise view of data", *Proceedings of the 1977 ACM national computer conference,* ACM, New York, NY, USA, June 13-16, pp. 77.

Chen, P.P.S. (1975) "The entity-relationship model", *ACM SIGIR Forum,* vol. 10, no. 3, pp. 9-9.

Chi, E.H. (2002) "A taxonomy of visualization techniques using the data state reference model", *Information Visualization 2000. InfoVis 2000. IEEE Symposium on Information Visualization 2000,*IEEE Computer Society, 9-10 Oct., pp. 69.

Codd, E.F. (1970) "A relational model of data for large shared data banks", *Communications of the ACM,* vol. 13, no. 6, pp. 377-387.

Cook, K., Earnshaw, R. and Stasko, J. (2007) "Guest editors' introduction: discovering the unexpected", *Computer Graphics and Applications, IEEE,* vol. 27, no. 5, pp. 15-19.

Dai, H., Lim, E., Hady Wirawan, L. and Hweehwa, P. (2008) "Visual Analytics for Supporting Entity Relationship Discovery on Text Data", *Intelligence and Security Informatics, Lecture Notes in Computer Science Volume 5075*, eds. C.C. Yang, H. Chen, M. Chau, et al, Springer, Taipei, Taiwan, June 17, pp. 183.

DeMarco, T. (1979) *Structured analysis and system specification,* 1st edn., Prentice Hall, Englewood Cliffs, NJ, USA.

Ding, Y. and Foo, S. (2002) "Ontology research and development. Part 1-a review of ontology generation", *Journal of Information Science,* vol. 28, no. 2, pp. 123-136.

Eysenck, M.W. and Keane, M.T. (200), *Cognitive psychology: A student's handbook,* 5th edition edn., Taylor and Francis, Psychology Pr, NY, USA.

Falconer, S.M., Bull, R.I., Grammel, L. and Storey, M.A. (2009) "Creating visualizations through ontology mapping", *International Conference on Complex, Intelligent and Software Intensive Systems. CISIS'09.*IEEE Computer Society, Burgos, Spain, 16-19 March, pp. 688.

Fassò, A. and Pezzetti, G. (2007) "Statistical Methods for Monitoring Data Analysis", *7th International Symposium on Field Measurements in Geomechanics (FMGM),* American Society of Civil Engineers, Reston, VA, USA, Sept 24-27, pp. 1.

Ferreira de Oliveira, M.C. and Levkowitz, H. (2003) "From visual data exploration to visual data mining: A survey", *Visualization and Computer Graphics, IEEE Transactions on,* vol. 9, no. 3, pp. 378-394.

Galorath, D. (2008) "Towards a Standard Definition of IT Infrastructure" [Online] Available: http://www.galorah.com [2012, Oct 18].

Garg, S., Nam, J.E., Ramakrishnan, I. and Mueller, K. (2008) "Model-driven visual analytics", *IEEE Symposium on Visual Analytics Science and Technology VAST'08*IEEE Computer Society, Columbus, Ohio, USA, 19-24 Oct., pp. 19.

Gilson, O., Silva, N., Grant, P.W. and Chen, M. (2008) "From web data to visualization via ontology mapping", *Computer Graphics Forum,* vol. 27, no. 3, pp. 959-966.

Görg, C., Spence, R. and Stasko, J. (2008) "Jigsaw: supporting investigative analysis through interactive visualization", *Information Visualization,* vol. 7, no. 2, pp. 118-132.

Gotz, D., Zhou, M.X. and Aggarwal, V. (2006) "Interactive visual synthesis of analytic knowledge", *IEEE Symposium On Visual Analytics Science And Technology,* IEEE, Baltimore, Maryland, USA, Oct. 31 - Nov 2, pp. 51.

Gravois, M. (2007) *Data Monitoring and Analysis Program Manual*, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Green, T.M., Ribarsky, W. and Fisher, B. (2009) "Building and applying a human cognition model for visual analytics", *Information Visualization,* vol. 8, no. 1, pp. 1-13.

Green, T.M., Ribarsky, W. and Fisher, B. (2008) "Visual analytics for complex concepts using a human cognition model", *IEEE Symposium on Visual Analytics Science and Technology VAST'08*IEEE Computer Society, Columbus, Ohio, USA, 19-24 Oct, pp. 91.

Hand, D.J., Mannila, H. and Smyth, P. (2001) *Principles of data mining,* First edition, MIT press.

Heer, J. and Agrawala, M. (2006) "Software design patterns for information visualization", *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, no. 5, pp. 853-860.

Heilala, J., Klobut, K., Salonen, T., Järvinen, P. and Shemeikka, J. (2010a) "Energy Use Parameters for Energy Efficiency Enhancement in Discrete Manufacturing Process", *7th CIRP International Conference on Intelligent Computation in Manufacturing Engineering,*Dept. of Materials and Production Engineering, University of Naples "Federico II", Capri, Italy, June 23 - 25, pp. 1.

Heilala, J., Klobut, K., Salonen, T., Järvinen, P., Siltanen, P. and Shemeikka, J. (2010b) "Energy Efficiency Enhancement in Discrete Manufacturing Process with Energy Use Parameters", *International Conference on Advances in Production Management Systems, Competitive and Sustainable Manufacturing Products and Services - APMS 2010*, eds. M. Garetti, M. Taisch , T. Cavalieri S. and M. Tucci, Politecnico di Milano, Rhodes island, Greece, Oct 11-13, pp. 1.

Icke, I. (2009) *Visual Analytics: A Multifaceted Overview*, Citeseer, CUNY, The Graduate Center.

IEEE 1451.1 (2004) *Standard: Smart Transducer Interface for Sensors and Actuators*, 1st edn., IEEE 1451.4 Standard Working Group.

Inaudi, D. (2002) "Development of reusable software components for monitoring data management, visualization and analysis", *SPIE, International Symposium on Smart Structures and Materials*, eds. S.-. Liu and D.J. Pines, Smart Structures and Materials, San Diego, CA, USA, March 17-21.3, pp. 10.

ISO 13374 (2007) *Standard: Condition monitoring and diagnostics of machines*, 1st edn., ISO International Organization for Standardization, Geneva, Switzerland.

ISO 9241-110 (2006) *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 10: Dialogue principles*, 2nd edn., ISO, Geneva, Switzerland.

ISO TC 108/SC 5 (2005) *Standard draft: Condition monitoring and diagnostics of machines — Data processing, communication, and presentation — Part 2: Data processing*, 1st edn., ISO TC 108/SC 5/WG 6. Secretariat, Geneva, Switzerland.

ISO/IEC 25062 (2006) *Usability standard*, 1st edn., ISO International Organization for Standardization, Geneva, Switzerland.

ISO/IEC 9126 (2001), *Standard ISO/IEC 9126-4 Software engineering — Product quality*, 2nd edn., ISO International Organization for Standardization, Geneva, Switzerland.

Järvinen, P. (ed.) (2007), *Hybrid media in personal management of nutrition and exercise*, VTT publications 656, Espoo, Finland.

Jarvinen, P., Jarvinen, T., Lahteenmaki, L. and Sodergard, C. (2008) "HyperFit: hybrid media in personal nutrition and exercise management", *Second International Conference on Pervasive Computing Technologies for Healthcare,*IEEE, January 30th - February 1st, pp. 222.

Järvinen, P., Puolamäki, K., Siltanen, P. and Ylikerälä, M. (2009) *Visual analytics. Final report.*, VTT Technical Research Centre of Finland.

Johnson, C., Moorhead, R., Munzner, T., Pfister, H., Rheingans, P. and Yoo, T.S. (2006) *NIH/NSF visualization research challenges report*, IEEE Computing Society, Los Alamitos, CA, USA.

Kang, H., Getoor, L. and Singh, L. (2007) "Visual analysis of dynamic group membership in temporal social networks", *ACM SIGKDD Explorations Newsletter,* vol. 9, no. 2, pp. 13-21.

Kang, H., Plaisant, C., Lee, B. and Bederson, B.B. (2007) "NetLens: iterative exploration of content-actor network data", *Information Visualization,* vol. 6, no. 1, pp. 18-31.

Kang, Y. and Stasko, J. (2011) "Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study", *IEEE Conference on Visual Analytics Science and Technology (VAST)*, eds. S. Miksch and M. Ward, IEEE Computer Society, Providence, RI, USA, 23-28 Oct., pp. 21.

Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J. and Melançon, G. (2008) "Visual analytics: Definition, process, and challenges" in *Information visualization*, eds. A. Kerren, J. Stasko, J. Fekete and C. North, First edition, Springer, Berlin Heidelberg, pp. 154-175.

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

Keim, D., Mansmann, F., Oelke, D. and Ziegler, H. (2008a) "Visual analytics: Combining automated discovery with interactive visualizations", *Discovery Science,* Springer, pp. 2.

Keim, D., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. (2008b) "Visual analytics: Scope and challenges" in *Visual Data Mining*, eds. S.J. Simoff, M.H. Böhlen and A. Mazeika, First edition edn., Springer, Heidelberg, pp. 76-90.

Keim, D.A. (1996) "Pixel-oriented visualization techniques for exploring very large data bases", *Journal of Computational and Graphical Statistics,* vol. 5, no. 1, pp. 58-77.

Keim, D.A., Mansmann, F., Schneidewind, J. and Ziegler, H. (2006) "Challenges in visual data analysis", *Tenth International Conference on Information Visualization* IEEE, London, UK, July 5-7, pp. 9.

Lengler, R. and Eppler, M.J. (2007) "Towards a periodic table of visualization methods for management", *Proceedings of the IASTED Conference on Graphics and Visualization in Engineering (GVE 2007),*ACTA Press Anaheim, CA, USA, Anaheim, CA, USA, January 3 – 5, pp. 83-88.

Liu, Z., Navathe, S.B. and Stasko, J.T. (2011) "Network-based visual analysis of tabular data", *IEEE Conference on Visual Analytics Science and Technology (VAST),*IEEE, Providence, RI, USA, Oct. 23-28, pp. 41.

Liu, Z. and Stasko, J.T. (2010) "Mental models, visual reasoning and interaction in information visualization: A top-down perspective", *IEEE Transactions on Visualization and Computer Graphics,* vol. 16, no. 6, pp. 999-1008.

Manolescu, I., Khemiri, W., Benzaken, V. and Fekete, J.D. (2009) "ReaViz::Reactive workflows for visual analytics", *Data Management and Visual Analytics Workshop,* Berlin, June 9.

Martin, J. and Finkelstein, C. (1981) *Information engineering,* First edition edn., Prentice Hall, USA.

Mathew, A., Zhang, L., Zhang, S. and Ma, L. (2006) "A review of the MIMOSA OSA-EAI database for condition monitoring systems", *World Congress on Engineering Asset Management* Springer, Gold Coast, Australia, July 11-14, pp. 837.

MIMOSA (2012) *Open specifications for Enterprise Application Integration (EAI) and Condition-based Maintenance (CBM).*, Alliance of Operations and Maintenance (O&M), http://www.mimosa.org.

Munzner, T. (2009) "A nested process model for visualization design and validation", *Visualization and Computer Graphics, IEEE Transactions on,* vol. 15, no. 6, pp. 921-928.

Norman, D. and Dunaeff, T. (1994) *Things that make us smart: Defending human attributes in the age of the machine,* Basic Books, USA.

North, C. (2006) "Toward measuring visualization insight", *Computer Graphics and Applications, IEEE,* vol. 26, no. 3, pp. 6-9.

North, C. and Shneiderman, B. (2000) "Snap-together visualization: a user interface for coordinating visualizations via relational schemata", *Proceedings of the working conference on Advanced visual interfaces,* ACM Press, Napoli, Italy, May 28-30, pp. 128.

Perer, A. and Shneiderman, B. (2009) "Integrating statistics and visualization for exploratory power: From long-term case studies to design guidelines", *Proceedings of International Conference on Intelligence Analysis* IEEE, 8 May, pp. 39.

Pirolli, P. and Card, S. (2005) "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis", *Proceedings of International Conference on Intelligence Analysis*, 8 May, pp. 2.

Pirolli, P., Card, S.K. and Van Der Wege, M.M. (2003) "The effects of information scent on visual search in the hyperbolic tree browser", *ACM Transactions on Computer-Human Interaction (TOCHI),* vol. 10, no. 1, pp. 20-53.

Plaisant, C., Grinstein, G. and Scholtz, J. (2009) "Visual-analytics evaluation", *IEEE Computer Graphics and Applications,* vol. 29, no. 3, pp. 16-17.

Pottenger, W., Fisher, B. and Ribarsky, W. (2009) "Science of analytical reasoning", *Information Visualization,* vol. 8, no. 4, pp. 254-262.

Puolamaki, K., Papapetrou, P. and Lijffijt, J. (2010) "Visually Controllable Data Mining Methods", *IEEE International Conference on Data Mining Workshops (ICDMW),* IEEE, Sydney, Australia, 13-13 Dec, pp. 409.

Rao, R. and Card, S.K. (1994) "The table lens: merging graphical and symbolic representations in an interactive focus context visualization for tabular information", *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence,* ACM, Boston, MA, USA, pp. 318.

Roberts, J.C. (2007) "State of the art: Coordinated and multiple views in exploratory visualization", *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV'07.*IEEE, 2-2 July, pp. 61.

Robertson, G.G., Mackinlay, J.D. and Card, S.K. (1991) "Cone trees: animated 3D visualizations of hierarchical information", *Proceedings of the SIGCHI* *conference on Human factors in computing systems: Reaching through technology,* ACM, April 27 - May 02, pp. 189.

Scholtz, J. (2006) "Beyond usability: Evaluation aspects of visual analytic environments", *IEEE Symposium On Visual Analytics Science And Technology,* IEEE, Baltimore, Maryland, USA, October 31 - November 2, pp. 145.

Sears, A. and Jacko, J.A. (2009) *Human-computer interaction: Development process,* First edition edn., CRC Press, Boca Raton, FL.

SEMI E10-0304, S. . Available: www.semi.org/en/standards/ctr_031244.

Shen, Z., Ma, K.L. and Eliassi-Rad, T. (2006) "Visual analysis of large heterogeneous social networks by semantic and structural abstraction", *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, no. 6, pp. 1427-1439.

Shneiderman, B. (1996) "The eyes have it: A task by data type taxonomy for information visualizations", *Proceedings, IEEE Symposium on Visual Languages,* IEEE, September 3-6, pp. 336.

Shrinivasan, Y.B. and van Wijk, J.J. (2008) "Supporting the analytical reasoning process in information visualization", *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems,* ACM, April 5-10, pp. 1237.

Simsion, G. and Witt, G. (2005) *Data modeling essentials,* Third edition, Morgan Kaufmann, San Francisco, CA.

Smuc, M., Mayr, E., Lammarsch, T., Aigner, W., Miksch, S. and Gartner, J. (2009) "To score or not to score? Tripling insights for participatory design", *Computer Graphics and Applications, IEEE,* vol. 29, no. 3, pp. 29-38.

Spyns, P., Meersman, R. and Jarrar, M. (2002) "Data modelling versus ontology engineering", *ACM SIGMOD Record,* vol. 31, no. 4, pp. 12-17.

Streit, M., Schulz, H., Lex, A., Schmalstieg, D. and Schumann, H. (2012) "Model-Driven Design for the Visual Analysis of Heterogeneous Data", *IEEE Transactions on Visualization and Computer Graphics,* vol. 18, no. 6, pp. 998-1010.

Tan, P.N., Steinbach, M. and Kumar, V. (2006) *Introduction to data mining,* First edition edn., Addison Wesley.

Thomas, J.J. and Cook, K.A. (2006) "A visual analytics agenda", *Computer Graphics and Applications, IEEE,* vol. 26, no. 1, pp. 10-13.

Thomas, J.J. and Cook, K.A. (2005) *Illuminating the path: The research and development agenda for visual analytics,* 1st edn., IEEE Computer Society, Los Alamitos, CA.

Tufte, E.R. (1983) *The visual display of quantitative information,* Graphics press, Cheshire, CT.

Van Wijk, J.J. (2005) "The value of visualization", *IEEE Visualization 2005*IEEE, Baltimore, Maryland, 23-28 October, pp. 79.

Venables, W.N. and Smith, D.M. (2012*) An introduction to R*. Available: http://cran.r-project.org/doc/manuals/R-intro.pdf [2012, October 12].

Ware, C. (2004) "Information visualization: perception for design"*,* 1st edn., Morgan Kaufmann.

Verzani, J. (2002) "simpleR - Using R for Introductory Statistics". Available: http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf [2012, October 12].

Wong, P.C. (1999) "Guest editor's introduction: Visual data mining", *IEEE Computer Graphics and Applications,* vol. 19, no. 5, pp. 20-21.

Wong, P.C. and Thomas, J. (2004) "Guest Editors' Introduction--Visual Analytics", *IEEE Computer Graphics and Applications, 24 (5): 20-21,* vol. 24, no. 5.

Wong, P.C. (2007) "Visual analytics science and technology", *Information Visualization,* vol. 6, no. 1, pp. 1-2.

Wright, W., Schroh, D., Proulx, P., Skaburskis, A. and Cort, B. (2006) "The Sandbox for analysis: concepts and methods", *Proceedings of the SIGCHI conference on Human Factors in computing systems,* ACM, Montreal, April 24-27, pp. 801.

Yi, J.S., ah Kang, Y., Stasko, J.T. and Jacko, J.A. (2007) "Toward a deeper understanding of the role of interaction in information visualization", *Visualization and Computer Graphics, IEEE Transactions on,* vol. 13, no. 6, pp. 1224-1231.

Zhou, M.X. and Gotz, D. (2009) "Characterizing users' visual analytic activity for insight provenance", *Information Visualization,* vol. 8, no. 1, pp. 42-55.

Appendix A: Visual analytics tool survey

Product and web-link (3 October 2012)	Company	Cat	Application areas	Description
D2K Data to Knowledge http://d2k.tamu.edu/	D2K Project	App	Learning and teaching	Data to Knowledge Project is a partnership of teachers, administrators, and researchers dedicated to creating useful solutions to face the complex responsibilities of today's teachers analytics and pattern discovery.
ILOG Visual CP http://pauillac.inria.fr/~contraintes/OADymPPaC/Public/oadDiscovery/ilog_discovery.html	ILOG comm, Download free preview version.	App	interactive visualization software designed to help analyze and debug constraints-based programs.	Interactive visualization software designed to help analyze and debug constraints-based programs.
iThink (isee) http://www.iseesystems.com/	isee systems, inc, commercial	App	Thinking for Business	Predictive analytical tools. Built-in functions facilitate mathematical, statistical, and logical operations, simulation.
Leadscope model applier http://www.leadscope.com/	Leadscope Commercial	App	biology, chemistry	Predictive data miner, provides specialized data mining and visualization software for the pharmaceutical industry.
Partek http://www.partek.com/	Partek comm. free trial	App	sequencing, microarray, integrated genomics	Visualization and analysis environment for Genomics analysis
DataDesk http://www.datadesk.com/	Data Description Inc Commercial	Env	General-purpose	Dynamic data exploration and visualization tools with interactive statistics procedures.
DataDetective http://www.sentient.nl/	sentient information systems Commercial	Env	Business intelligence, crime analysis	Users are able to analyze very large data sets to find all hidden patterns and complex relations by using artificial intelligence, statistical techniques and visualization methods. Relations, trends, clusters.
DecisionQ http://www.decisionq.com/	DecisionQ corporation Commercial	Env	Financial sector, care industries, diagnostics, insurance, goods, logistics, national defence and other industries	Predictive statistical products.
Eudaptis Viscovery http://www.eudaptics.com/	Viscovery,commercial	Env	Telecom, insurance, banking, retail, e-business, media, finance, fund raising, voting analysis, industry, science, document classification, and much more	Interactive data visualization software and solutions. Critical insight from all types of data. Explorative data miningPrediction & scoringCluster analysis & segmentation, SOM

Product and web-link (3 October 2012)	Company	Cat	Application areas	Description
GeneXproTools http://www.gepsoft.com/	Commercial	Env	risk management, response models, financial services, marketing, and in most scientific and engineering research	Modelling and data mining software for data analysis, designed for Function Finding, Classification, Time Series Prediction, and Logic Synthesis
groupvisual io http://groupvisual.io/	groupvisual io, consulting	Env	General-purpose	Dashboarding, reporting, visualization tools, and technologies, find patterns, distributions, correlations and/or anomalies across multiple data types using visual cues to explore and understand the information.
InforSense http://www.inforsense.com/	inforsense.com Commercial consultant	Env	Business, Healthcare, Life sciences	Interactive web-based analytical applications linking multiple tables and charts together. Classification and Predictive Modelling, Clustering and Segmentation Analysis, Association Analysis, Multivariate and Regression Analysis, Advanced data preparation methodologies
KNIME Konstanz Information Miner http://www.knime.org/	Open source, Univ. Of Konstanz	Env	General-purpose	Modern data analytics platform, statistics, data mining, analyze trends and predict potential results. Its visual workbench combines data access, data transformation, initial investigation, powerful predictive analytics and visualization.
Knowledge Seeker, Knowledge studio www.angoss.com	Agnoss, commercial consultant	Env	Business intelligence	Focused industry services expertise in the areas of deployment, integration and use of predictive analytics within enterprise environments. Interactive Decision Trees, Predictive Modelling and Scoring.
KXEN http://www.kxen.com/	KXEN Commercial consultant	Env	Communications, business, retail, finances	Classification, Regression, and Attribute Importance.
NovoSpark http://www.novospark.com/	NovoSparkcommercial Tool, 30 days free trial	Env	Oceanology Insurance and Finance Medical Research Aeronautics	Visualizing multidimensional datasets Human eye to perceive visual patterns. Presents each multidimensional observation as a single observation curve.
Nuggets http://www.data-mine.com/	Data Mining Technologies, commercial	Env	Direct Marketing, Healthcare, Pharmaceuticals, Genetics, CRM, Telecommunications, Utilities, Financial Services	Builds models that uncover hidden facts and relationships. These models can predict for new data, allow you to generalize and reveal which indicators (i.e. variables) most impact your decisions"
OpenVIz http://www.openviz.com/	Advanced Visual	Env	mi Business Intelligence, Risk & Financial Customer Analytics , Drug Discovery, Life Sciences, Medicine, Oil & Gas, Earth Sciences, Engineering, Scientific Research, Education	A selection of built-in standard analytics procedures or incorporate your own for modelling, trending and grouping analysis. Highly interactive features such as filtering, sampling and thresholding...

Product and web-link (3 October 2012)	**Company**	**Cat**	**Application areas**	**Description**
Panopticon http://www.panopticon.com/	Panopticon	Env	General-purpose	Data visualization software supports fast, efficient visual analysis of real-time and historical time series data as well as very large data sets stored in standard relational databases.
RapidMiner (YALE) http://rapid-i.com/	Rapid-I, Commercial , GNU General Public License	Env	business intelligence solutions	RapidMiner provides an environment for data mining, machine learning and business intelligence in Java. Integrates learning schemes and attribute evaluators of the Weka environment.
SALFORD SYSTEMS (CART, MARS, Treenet, RandomForest) http://www.salfordsystems.com	Salford Systems Commercial	Env	business intelligence: telecommunications, retail, catalog, healthcare and financial services markets	"... data mining and web mining tools", iCART Decision Tree Software MARS Predictive Modeling Software, TreeNet® Stochastic Gradient Boosting Software, LOGIT Software, Random Forests
Spotfire 5http://spotfire.tibco.com	Tibco softwarecommercial	Env	Business decisions, Enterprise analysis, Decision analysis, Big data	Interactive, visual data analysis. Data visualization helps users interpret critical relationships in multidimensional data. Spotfire information visualization allows users to easily query and comprehend complex data.
Statistica http://www.statsoft.com/	Statsoft, commercial	Env	Banking , Finance, Chemical / Petrochemical, Consumer Product Goods, Food Manufacturing, Healthcare, Insurance, Marketing, Pharmaceuticals, Power Industry, semiconductors	Suite of analytics software products. Data analysis, data management, data visualization, and data mining procedures. Predictive modelling, clustering, classification, and exploratory techniques. Vector Machines, EM and k-Means Clustering, Classification & Regression Trees, Generalized Additive Models, Independent Component Analysis, Stochastic Gradient Boosted Trees, Ensembles of Neural Networks, Automatic Feature Selection, MARSplines, CHAID Trees, Nearest Neighbour Methods, Association Rules, Random Forests
Tableau http://www.tableausoftware.com/	Tableausoftware Commercial free trial	Env	Business intelligence, marketing, finance, sales departments, supply-chains, national security, engineering	“A visual spreadsheet for databases that allows you to visually explore, analyze and create reports. Fast analytics + visualization for everyone Tableau combines data exploration and visualization in easy-to-use applications you can quickly master. "
Think EDM (former K.wiz) http://www.thinkanalytics.com/	thinkAnalytics commercial	Env	Telecommunications, Media, Banking & Retail industries, CRM and Business Intelligence	"... knowledge discovery and data mining techniques designed to provide business users with intelligent analysis capabilities...", "provides a range of Java visualization components including heatmaps, decision trees, 3D scatterplots and association rules"
Visokio Omniscope http://www.visokio.com/	Visokio product and consulting	Env	home use	"Filter, analyse and edit information in interactive point-and-click graphs, charts and maps, import web content, and create slider-driven models"

Product and web-link (3 October 2012)	Company	Cat	Application areas	Description
VisuMap http://www.visumap.net	VisuMap Technologies product and consulting	Env	Pharmaceutics, Bioinformatics, Financial analysis, Market analysis, Telecommunication industry, Internet traffic analysis, Quality control	"...information visualization tools for high dimensional non-linear data. VisuMap provides a window into the patterns, relationships and correlations hidden in your data and enables you to interact with them in real time"
XLMiner http://www.resample.com/xlminer/	Resampling Stats, Inc., Commercial	Env	XLMiner	Data mining add-in for Excel, with neural nets, classification and regression trees, logistic regression, linear regression, Bayes classifier, K-nearest neighbours, discriminant analysis, association rules, clustering, principal components, more …
Birdeye http://code.google.com/p/birdeye/	Open Source community	Misc	General-purpose	BirdEye is a community project to advance the design and development of a comprehensive open source information visualization and visual analytics library for Adobe Flex.
C4.5/C5.0/See5	Open source data mining tool	Misc	General-purpose	
Weka http://www.cs.waikato.ac.nz/ml/weka/	The University of Waikato, GNU General Public License	Misc	a collection of algorithms for any data mining tasks	"... an open source software that supports data mining tasks, classification, data preprocessing, clustering, regression and visualization. It's a Java standalone application with a very nice GUI, issued under GNU General Public License. Open source machine learning software in Java, Weka is a collection of algorithms for any data mining tasks"
BusinessObjects XI http://www.sap.com/solutions/analytics/business-intelligence/index.epx	SAP company	SWH	Business intelligence	”BusinessObjects XI is full spectrum software tool that can do reporting, query and analysis, dashboards and visualization, intuitive discovery and advanced predictive analytics capabilities”
i2 Analysts Notebook http://www.i2group.com/uk	IBM i2 Analyst's Notebook comm	SWH	General-purpose	"A rich visual analysis environment underpinned by a local repository"
IBM Spatiotemporal Visual Analytics Workbench http://researcher.watson.ibm.com	Commercial	SWH	General-purpose	"IBM Spatiotemporal Visual Analytics Workbench converts raw data into interactive visualizations that are comprehendible by domain experts. The key capabilities include: Support for multiple data layers, Interactive visualization, including filtering, highlighting, and selection, In-context analytical algorithms, Capturing and communicating insight, Converting insight from historical data into live event monitoring rules."
Inxight http://www.inxight.com/	Inxight, SAP company commercial	SWH	text analysis	"Get valuable insight into big data based on the location of an event or transaction. SAP Data Services can help you leverage the power of location-based intelligence, geocoding, and geographic data services"

Product and web-link (3 October 2012)	Company	Cat	Application areas	Description
Oracle http://www.oracle.com	Oracle	SWH	General-purpose	"New Oracle engineered systems deliver big data and high-speed visual analytics."
SAP visual intelligence		SWH		" SAP Visual Intelligence point and click interface and engaging visualizations allow you to quickly analyze data for rapid time to insight and business agility – no scripting required"
SAS Visual analytics http://www.sas.com	SAS, Commercial	SWH		"Visually explore big data using high-performance, in-memory capabilities to understand your data, discover new patterns and publish reports to the web and mobile devices"
SPSS Clementine (VTT), http://www.spss.com/clementine/	SPSS Inc, Commercial,	SWH	Business, government agencies, academic institutions	Cluster analysis (Kohonen, K-Means and Two-Step), Missing value imputation based on both rule- and algorithm-based methods, Enhanced Logistic Regression node, Chi-squared, t-test, and ANOVA, Time-Series Modelling node estimates exponential smoothing and ARIMA models to produce forecasts, CHAID, Exhaustive CHAID, and QUEST.
SQL Server	Microsoft	SWH	data analysts, business decision makers, and information workers	" Power View is a browser-based application that enables users to present and share insights with others in their organization through interactive presentations. Many capabilities for analysis, visualization, and sharing available with Power View."
WebFocus http://www.informationbuilders.com	Information Builders, commercial	SWH	General-purpose	
Bayesia http://www.bayesia.com/	Bayesia commercial	Tech	Marketing, Industry, Health, Risk management	Bayesian networks and graphics
DataEngine http://www.dataengine.de	MIT GmbH commercial, Germany	Tech	Process control, quality control	"DataEngine is the software tool for intelligent data analysis which unites statistical methods with neural networks and fuzzy technologies."
Exelis http://www.exelisvis.com/	ITT, comm	Tech	Defence and intelligence, monitoring natural resources, federal and local governments, earth observation, mining, oil, and gas exploration, and biotechnology	"…access, analyze, and share all types of data and imagery. Interactive Data Language, enables in-depth data analysis through industry-leading visualization" , built-in library of math, statistics, image processing and signal processing routines
inflow Network Mapping Software http://www.orgnet.com	orgnet.com commercial consultant	Tech	Social Network Analysis software & services for organizations, communities, and their consultants	"Quickly and easily navigate large data sets to spot trends, detect outliers, and uncover data quality problems", Network Centrality, Small-World Networks, Cluster Analysis, Network Density... "

Product and web-link (3 October 2012)	Company	Cat	Application areas	Description
Miner 3Dhttp://www.miner3d.com/	commercialfree trial for 15 days	Tech	Life Sciences, Clinical Trials, Bank and investment analysts, researchers in pharmaceutical, biotech or material research, process engineers, sales managers, geologists	"Create engaging data visualizations and live data-driven graphics! Miner3D delivers new data insights by allowing you to visually spot trends, clusters, patterns or outliers." Kohonen's Self-Organizing Maps,Trellis Charts, Principal Component Analysis, K-Means Clustering, Visual Clustering, Data analysis and visualization in 3D
NETMAP http://www.netmap.com.au/	NetMap analytics commercial	Tech	Insurance, retail, corporate & government, crime investigation and marketing industries	"innovative combination of link analysis and data visualization, with applications to fraud detection and claims analysis."
Oculus GeoTime™ http://www.oculusinfo.com/	oculus commercial consultant	Tech	General-purpose	State-of-the-art class libraries and rendering engines, visualization of geospatial and temporal data with complex data over space and time within a single, highly interactive 3D view.
Purple Insight MineSet http://www.algorithmic-solutions.com/ leda/projects/mineset.htm	commercial	Tech	General-purpose	"...easy-to-use and scalable program for data mining and real-time 3D visualization. Featuring powerful interactive tools for data access as well as analytical and visual data mining"
Sentinel Visualizer http://www.fmsasg.com/ products/SentinelVisualizer/	FMS commercial	Tech	Law enforcement, investigators, researchers, and information workers,	Advanced social network analysis and visualization, link analysis, social network analysis, geospatial, and timelines.
POLYANALYST http://www.megaputer.com/	Megaputer intelligence Amer. Commercial	Tech	Data mining, text mining	"… derives actionable knowledge from large volumes of text and structured data, delivers this knowledge to decision makers in the form of predictive models" , Categorization, Clustering, Prediction, Link Analysis, Keyword and entity extraction, Pattern discovery, Anomaly detection"
ScienceGL http://www.sciencegl.com/	commercial	Tech	Health, Industry, Forensics, Business	3D and 4D data visualization software for science, healthcare, technology and business applications.
Data clarity suite http://www.visualanalytics.com	VAI Visual Analytics inc	VA	General-purpose	Pattern discovery and link analysis
Nuix Visual Analytics http://www.nuix.com / visual-analytics	General-purpose	VA	General-purpose	Nuix Visual Analytics extends Nuix's powerful investigative review and analysis capabilities with a fully interactive data visualization and workflow framework.
PV-WAVE http://www.roguewave.com/ products/pv-wave-family.aspx	commercial	VA		"Development environment for desktop visual data analysis applications for rapid importing, manipulation and analysis of data"

Appendix B: HyperFit tool construction steps

Step 1. Define object of interest types

ObjectOfInterestType

Id	Title
	Person

Step 2: Define properties

OITypeProperty

Id	**OITypeId**	**Title**	**Definition**	**Property type**	**Data type**	**NumUnit**	**Code Id**	**Lower limit**	**Upper limit**	**Goal value**	**Ref value**	**Meas.Type**	**Freq**
	<Person>	Age		B	Q	year	-	18	100	-	-	-	-
	<Person>	Height		B	Q	cm	-	80	230	-	-	-	-
	<Person>	Start Weight		B	Q	kg	-	40	150	-	-	-	-
	<Person>	Gender		B	N	-	<Gender code>	-	-	-	-	-	-
	<Person>	Work Activity		B	N	-	<WA code>	-	-	-	-	-	-
	<Person>	Weight		M	Q	kg	-	40	150	-		A	1/day
	<Person>	Feeling		M	O	score	-	0	4			A	1/day
	<Person>	Steps		M	Q	step	-	0	30 000	10 000[1]		A	1/day
	<Person>	Exercise duration		M	Q	min	-	0	3600	45		A	1/day
	<Person>	Exercise kcal		M	Q	kcal	-	0				A	1/day
	<Person>	Food kcal		M	Q	kcal	-	800	4000			A	1/day
	<Person>	Food protein		M	Q	E%	-	0	100 E%	15 E%.[1]		A	1/day
	<Person>	Food carbo-hydrates		M	Q	E%	-	0	100 E%	55% E%.[1]		A	1/day
	<Person>	Food Fat		M	Q	E%	-	0	100 E%	30 E%.[1]		A	1/day
	<Person>	Food Sugar		M	Q	E%	-		100 E%	10 E%.[1]		A	1/day
	<Person>	Food Fibre		M	Q	g	-	0	-	25 g		A	1/day

Property group

Id	Title
	Food
	Exercise

GroupMember

Id	PropertyId	GroupId
	<Food kcal>	<Food>
	<Food protein>	<Food>
	<Food carbohydrates>	<Food>
	<FoodFat>	<Food>
	<FoodSugar>	<Food>
	<FoodFiber>	<Food>
	<Exercise kcal>	<Exercise>
	<Exercise duration>	<Exercise>

Code

Id	Title
	Gender
	WorkActivity

CodeValue

Id	CodeId	CodeValue	CodeTitle
	<genderCode>	1	Man
	<genderCode>	2	Woman
	<WAcode>	1	Light
	<WAcode>	2	Medium
	<WAcode>	3	Hard

Step 3. Define indexes

OITypeIndex

Id	OITypeId	Title	Definition	DataType	NumUnit	CodeId	Lower limit	Upper limit	Ref value	Goal value	Frequency
	<Person>	Weight index	Height/ weight ^{2}	Q	-	-	10	50		18,5 – 25	1/day
	<Person>	Energy balance	(basic consumption + exercise kcal) – food kcal	Q	-	-	-5000	+5000		0	1/day
	<Person>	Food quality	1 - fat g * 9/ * 100 < = 30 1 - fibre over 25 g 1 -	O	-	<QualityCode>	0	6		3-6	1/day

OITypeIndexProperty

Id	OITypeIndexId	OITypePropertyId
	<Weight index>	<Height>
	<Weight index>	<Weight>
	<Energy balance>	<Exercise kcal>
	<Energy balance>	<Food kcal>
	<Food quality>	<Food protein>
	<Food quality>	<carbohydrates>
	<Food quality>	<fat>
	<Food quality>	<sugar>
	<Food quality>	<fibre>

Step 4. Select visualization and analysis methods

	No time point/single time point -Background variables -Means of measurements			**Time series** Measurements/Indexes -hourly, daily, weekly -sums/means		
	Nominal	**Ordinal**	**Quantitative**	**Nominal**	**Ordinal**	**Quantitative**
Variables	Gender Work Activity	-	Age Height Start weight Means of measurement	-	Feelings Food quality	Measurements: Daily weight Steps Exercise duration Exercise kcals Food kcals, protein, carbohydrates, fat, sugar, fibre Indexes: Body mass index, Energy balance
Univariate	Pie chart		Mean, var, min, max, count Histogram Clustering		Time series Auto-correlations	Time series Auto-correlations
Bivariate			Scatterplot with regression line Correlation r and coefficient R^2		Cross-correlations	Cross-correlations Time series diagram of values calculated from two variables
Multivariate			Correlation matrix or network, Residual network PCA Parallel coordinates			

Step 5. Map to original data

ObjectOfInterest	HyperFit	
OIId	User-id	
Title	-	
Background properties		
Age	User-Birthdate	
Gender	User-Gender	
Height	User-Height	
StartWeight	UserBackground-Weight	First given weight
WorkActivity	UserBackground-Worktimeactivity	
Weight		
Value, TimeStamp	User_Background_Weight	daily mean
GoalValue	UserGoals-WeigthTarget	
RefValue	UserGoals-IdealWeight	
Feeling		
Value, TimeStamp	ExerciseDiaries (parameter=2) -value	daily mean
Steps		
Value,TimeStamp	ExerciseDiaries (parater=1) – value	daily sum
GoalValue	UserGoals-DailyStepTarget	
Exercise duration		
GoalValue	UserGoals- DailyBasicTarget+WeeklyExactTarget/7	
Value, TimeStamp	ExerciseDiariesEntries-duration	daily sum
ExerciseKcal		
Value, TimeStamp	ExerciseDiariesEntries-duration*Sports-energyfactor*UserBackground-Weight	daily sum
FoodKcal		
RefValue	UserGoals-EnergyNeed	
Value, TimeStamp	FoodDiary-ProductId, NutriContents-Value (Nutritionalpproid=11)	
FoodProtein		
RefValue	15*UserGoals-EnergyNeed/100	
Value, TimeStamp	FoodDiary-ProductId, NutriContents-Value (Nutritionalpproid=12)	400 * Daily sum/FoodKcal
FoodCarbo		
RefValue	55*UserGoals-EnergyNeed/100	
Value, TimeStamp	FoodDiary-ProductId, NutriContents-Value (Nutritionalpproid=13)	400 * Daily sum/FoodKcal
FoodFat		
RefValue	30*UserGoals-energyneed/100	
Value, TimeStamp	FoodDiary-ProductId, NutriContents-Value (Nutritionalpproid=16)	900 * Daily sum/FoodKcal
FoodSugar		
RefValue	10*UserGoals-energyneed/100	
Value, TimeStamp	FoodDiary-ProductId, NutriContents-Value (Nutritionalpproid=14)	400 * Daily sum/FoodKcal
FoodFibre		
Value, TimeStamp	FoodDiary-ProductId, NutriContents-Value (Nutritionalpproid=18)	Daily sum

Step 6. Add application specific features

No application-specific features were added

Appendix C: HyperFit tool evaluation

Background query

Date ______________

Test user

Name ____

Age __________

Gender Man ____ Woman ____

Profession ____

Education ____

Work description:

____

_______________ Have you used other similar products?

____

_______________Expectations of the test?

After session interview

CASE Hyperfit
After session interview
Date ________________
Test users ____

1. What was good about the tool?

2. Was anything confusing or difficult about the tool?

3. What improvements would you suggest?

4. Would you find this kind of tool useful? How?

5. For whom would you recommend this tool and for what purpose?

6. Other comments?

Appendix D: MMEA tool construction steps

Step 1. Define object of interest types

ObjectOfInterestType

Id	Title
	Area
	Building
	Space

Step 2. Define properties

OITypeProperty

Id	OI TypeId	Title	Definition	Property type	Datatype	Num Unit	Code Id	Lower limit	Upper limit	Goal value	Ref value	Meas. Type A	Freq
	<Building>	Power	Power usage of a building	M	Q	kWh	-	0				Abs	1/hour
	<Building>	React power	React power of a building	M	Q	kVARh	-	0				Abs	1/hour
	<Building>	Water	Water consumption of a building	M	Q	m^3	-	0				Abs	1/hour
	<Building>	District heat	District heat usage of a building	M	Q	kWh	-	0				Abs	1/hour
	<Building>	Gross volume	Volume of a building	B	Q	m^3	-	0	-	-	-	Abs	-
	<Building>	Floor area	Floor area of a building	B	Q	m^2	-	0	-	-	-	-	-
	<Building>	Age	Age of a building	B	Q	years	-	0	-	-	-	-	-
	<Space>	CT	Indoor temperature of a room	M	Q	deg.C	-	0				Abs	1/hour
	<Space>	CCO	Indoor CO2 of a room	M	Q	ppm	-					Abs	1/hour
	<Space>	Occupancy	Occupancy of a room	M	N	-	<Presence >	-				-	1/hour
	<Area>	WOT	Weather operative temperature	M	Q	deg.C	-	-50				Abs	1/hour
	<Area>	WRH	Humidity	M	Q							Abs	1/hour

Code

Id	Title
	Presence
	Indoor Index

CodeValue

Id	CodeId	CodeValue	Title
	<Presence>	0	No presence
	<Presence >	1	Presence
	<Indoor Index>	S1	Individualized
	<IndoorIndex>	S2	Good
	<IndoorIndex>	S3	Satisfactory

Step 3. Define indexes

OITypeIndex

Id	OITypeId	Title	Definition	DataType (nominal/ ordinal/ quantitative)	CodeId	Lower limit	Upper limit	Ref value	Goal value	Frequency
	Building	Energy efficiency	Power/ gross volume	Quantitative	-					1/hour
	Space	Indoor air quality	Defined by [35]	Ordinal	<Indoor Index>				S2	1/hour

OITypeIndexProperty

Id	OITypeIndexId	OITypePropertyId
	BuildingEnergyEfficiency	Power,
	BuildingEnergyEfficiency	gross volume
	SpaceIndoorAirQuality	CT
	SpaceIndoorAirQuality	CCO
	SpaceIndoorAirQuality	WOT

[35] https://www.rakennustieto.fi/index/english.html

Step 4. Select visualization and analysis methods

	No time point/single time point -Background variables -Means of measurements			**Time series** Measurements/Indexes -hourly, daily, weekly -sums/means		
	Nominal	**Ordinal**	**Quantitative**	**Nominal**	**Ordinal**	**Quantitative**
Variables			Building - Gross volume - Floor Area - Age Means of measurements	Room -presence -indoor index		Building: -Power -Reactive power -Water -District heat Room: -CT (Indoor temperature) -CCO (CO_2) Area: -WOT (Area temperature) -WRH (Area humidity) Energy performance index
Univariate	Pie chart		Mean, var, min, max, count Histogram Clustering		Time series Auto-correlations	Time series Auto-correlations
Bivariate			Scatterplot with regression line Correlation r and coefficient R^2		Cross-correlations	Cross-correlations
Multivariate			Correlation matrix or network, Residual network PCA Parallel coordinates			

Step 5. Mapping to original data

No mapping required; data comes from the platform.

Step 6. Application-specific features

Coordinates and model details are required.

Object of interest	Title	Co-ordinates	IFC model
VTT digitalo		8794	<>

Appendix E: Comparison of visual analytics approaches

Full criteria

- *Functionality* - A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs. The sub-characteristics are *suitability, accuracy, interoperability, security, and functionality compliance.*

- *Reliability* - A set of attributes that bear on the capability of software to maintain its level of performance under stated conditions for a stated period of time. The sub-characteristics are *maturity, fault tolerance, recoverability, and reliability compliance*.

- *Usability* - A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users. The sub-characteristics are *understandability, learnability, operability, attractiveness, and usability compliance.*

- *Efficiency* - A set of attributes that bear on the relationship between the level of performance of the software and the amount of resources used, under stated conditions. The sub-characteristics are *time behaviour, resource utilisation, and efficiency compliance*.

- *Maintainability* - A set of attributes that bear on the effort needed to make specified modifications, including *analyzability, changeability, stability, testability, and maintainability compliance*.

- *Portability* - A set of attributes that bear on the ability of software to be transferred from one environment to another. These are *adaptability, installability, co-existence, replaceability, and portability compliance*.

Analysis approaches				
ISO/IEC 9126 software quality evaluation criteria	Traditional data mining package	General-purpose visual analytics solutions	Specific visual analytics application	Domain approach
Functionality				
Suitability	Good, tailored specifically to the problems at hand. Visualization capabilities can be limited.	Average, does not necessarily include the best possible analysis methods.	Good, but only if a suitable application exists.	Good inside the domain.
Accuracy	Good, tailored to the problem.	Average	Good if matches with the application needs	Good, can be fitted to the problem at hand by adding analysis and visualization methods.
Interoperability	Limited, distinct file-based solutions.	May provide interoperability features.	Limited, distinct applications, interoperability may not be supported.	Standard database supports interoperability.
Security	Requires extra skills and implementation	May be included.	Typically do not cover security, not main focus.	Standard solution supports implementation.
Functionality Compliance	Changes require human	Data contents and methods can be	Changes may require human	Data contents and methods can be

Analysis approaches				
ISO/IEC 9126 software quality evaluation criteria	Traditional data mining package	General-purpose visual analytics solutions	Specific visual analytics application	Domain approach
	implementation	modified easily, but within the limits of the solution.	implementation.	modified easily inside the domain.
Reliability				
Maturity	Single-shot solutions, maturity with time.	Matures through multiple users and versions.	Solutions, mature through multiple use.	Standard solution, maturity with time.
Fault Tolerance	Depends heavily on the implementation and testing.	Supposedly well tested before launch.	Supposedly well tested before launch.	Standard solution, good fault tolerance can be achieved.
Recoverability	Not good in file-based systems	If built on databases, include standard database recoverability.	If built on databases, include standard database recoverability.	Built on database, includes standard features of databases.
Reliability Compliance	No	May be included.	May be included.	Can be implemented.
Usability				
Understandability	Depends on implementation	Can be confusing initially due to many features. May require considerable learning to apply specific application data.	Focus on the specific application, may be clear.	Usability testing suggests good understandability.
Learnability	Depends on the implementation and personal guidance.	Typically includes good tutorial, examples and guides to support learnability.	Depends on the implementation and heavily on the solution.	Usability testing suggests good learnability. Built-in user guidance lessens need for learning.
Operability	Depends on the implementation.	Important for selling the product.	Important for selling the product.	Standard solution and libraries improve operability.
Attractiveness	Not the focus.	Important, does not sell otherwise.	Not necessarily the focus.	Depends on the final installation of user interface.
Usability Compliance	Not the focus.	May provide restricted possibilities to comply to needs.	Not the focus.	User interface can be tailored on standard interfaces to comply to needs.

Analysis approaches				
ISO/IEC 9126 software quality evaluation criteria	Traditional data mining package	General-purpose visual analytics solutions	Specific visual analytics application	Domain approach
Efficiency				
Time Behaviour	Can be tuned (depends on the skills of the implementer).	Overall solutions are not known for efficiency.	Depends heavily on the implementation.	Unknown, depends on the database and its tuning capabilities.
Resource Utilisation	Can be tuned (depends on the skills of the implementer)	Overall solutions tend to demand resources.	Depends heavily on the implementation.	A standard solution, can be tuned.
Efficiency Compliance	Good	Tuning not necessarily possible.	Tuning not necessarily possible.	Standard ways can be constructed.
Maintainability				
Analysability	Depends on the implementation, not necessarily very large solutions.	Not necessary possible, code not available or very large (open source).	Not necessarily possible, code not available.	Depends on the implementation, standard solution may support analysability.
Changeability	Possible, requires rebuilding the scripts.	New version updates.	New version updates.	Standard interfaces can be implemented to support changing data, analysis and visualization methods, and alternative user interfaces.
Stability	Depends on the implementation.	Commercial products need to be stable.	Depends on the implementation.	Depends on the implementation, standard solutions more stable.
Testability	Normal	May include automatic testing features.	Normal	Normal
Maintainability Compliance	If implementer is available at reasonable price.	Requires dealings with product maintenance.	Requires dealings with product developer.	Modifications outside data, analysis and visual methods, and UI Requires dealings with product developer.
Portability				
Adaptability	Not good, built for specific cases.	Good	Good inside the application.	Good inside the domain.
Installability	Ad-hoc installation procedures.	Typically includes automatic installation. May	Typically includes a good installation	An installation procedure can be built. Requires a configuration

Analysis approaches				
ISO/IEC 9126 software quality evaluation criteria	Traditional data mining package	General-purpose visual analytics solutions	Specific visual analytics application	Domain approach
		require a configuration phase.	procedure.	phase.
Co-Existence	no hindrances	no hindrances	no hindrances	no hindrances
Replaceability	Can be replaced with other kinds of solutions with some loss of functionality.	Other alternatives exist, can be replaced.	Can be replaced with a general-purpose solution, with some loss of functionality.	Can be replaced with a general-purpose solution, with some loss of functionality and usability.
Portability Compliance	Same scripts can be reused, packages are portable.	Unknown	Unknown	Can be implemented.