4 Data Mining

4.1 Motivation

In recent years, researchers of different fields have identified a phenomenon that has been coined as information tsunami or data tsunami – we live in a world where the capacity of producing and storing data is increasing daily at a very fast pace, however, our ability, as human beings, to understand such an overwhelming amount of data has not grown at the same rate. In order to deal with this problem, we undoubtedly need new technologies to unite the seemingly conflicting requirements of scalability and usability in making sense of the data.

Humans are required in the data analysis process

In the last decades, several analysis methods have been developed which were purely automatic or purely visual, but to deal with the complexity of the problem space, humans need to be included at an early stage of the data analysis process[66]. We will now consider two examples of particularly complex problems that affect us: understanding the function of genes (e.g., how can devastating diseases be cured), and understanding earth dynamics (e.g., how can natural disasters be predicted).

The 21st century has witnessed rapid development within the field of genomics. Initiatives such as the Human Genome Project and similar projects for other organisms, have begun to establish the genetic structure by identifying and locating genes in DNA sequences. Although far from perfect, these sequence-to-gene mappings will dramatically increase our understanding of genomics.

New tools and methodologies are necessary to help experts extract relevant information

At the same time, the world has been affected by some of the most catastrophic natural disasters in recent history. Some of these are of geologic origin, such as the recent L'Aquila earthquake (2009) or the Sumatra-Andaman earthquake (2004), which triggered the single worst tsunami in history; the majority are related to climatic dynamics. For example, Hurricane Katrina (2005), one of the costliest and deadliest hurricanes in American history; or El Niño (El Niño Southern Oscillation, ENSO), whose erratic periodicity cost hundreds of lives and caused billions in damage worldwide, partly through flooding in South America and partly through failed harvests in South East Asia. Natural and man-made catastrophes, coupled with increased security needs have triggered the improvement of monitoring systems (e.g., the Global Monitoring for Environment and Security, GMES[1]), capable of compiling data gathered from different sources (on the ground, from the depths of the oceans, by aircraft or balloon, or by satellites) and assembling them into usable, compatible and comparable information services.

[1] http://ec.europa.eu/gmes/

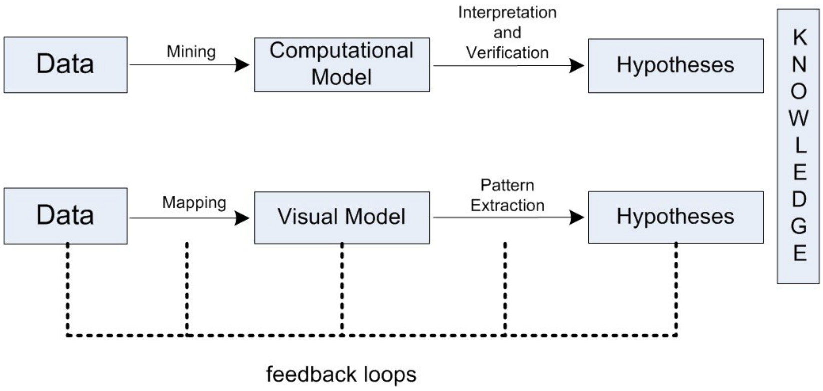


Figure 4.1: Comparing traditional data mining (top) and information visualisation (bottom) analytic processes[14]

Computers have played a key role in improving data acquisition methods thus providing us with the necessary depth of information to diagnose and prevent both diseases and natural disasters. Experts are required to assess current data sources and make predictions. Although massive amounts of data are available, it is imperative that new tools and new methodologies are developed to help these experts extract the relevant information.

KDD is useful but still limited

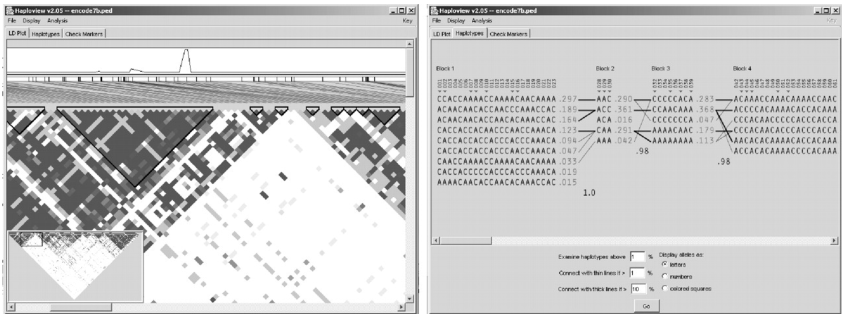
Knowledge discovery and data mining (KDD) is about semi or fully automated analysis of massive datasets and is therefore central to the problems at hand. Such automatic analysis methods are part of a discipline with a long tradition and solid, theoretical foundations. They are not focused on one application area, and the contributions of the field are more about general methodologies. KDD methods are especially suitable for analytical problems in which there exist means for assessing the quality of the proposed solutions. However, very often they become black-box methods in the hands of the end users (e.g., the prostate cancer physicians) or the algorithms provide results that do not lead to a solution to the problem, because they do not take into account relevant expert knowledge.

Limitations of visualisation methods

In contrast, visualisation methods use background knowledge, creativity and intuition to solve the problem at hand. While these approaches often give acceptable results for small datasets, they fail when the supplied data is too large to be captured by a human analyst[66]. Figure 4.1 compares the KDD and information visualisation processes.

Visual analytics approach is the third way

Nowadays, a third approach has begun to emerge, i.e., the visual analytics approach, which brings the experts' background knowledge back into the analysis process, together with the ability to interact and steer the analysis process.



```
CCACCAAAACCAAAACAACAAAA .297   AAC .290   CCCCCACA .283   ACAAACCAAACAAAACCAAC
ACAACAACACCAACCCAAACCAC .189   ACC .361   CCAACAAA .368   ACCCCACAAAACACCACAAA
ACAACAACACCAACACAAACCAC .164   ACA .016   CCCCCCCA .047   CCCACAACCCCCACCCACCA
CACCACCACAACCCAACCAAACA .123   CAA .291   AAAACAAC .179   CCCACAACACCCACCCACCA
CACCACCACCACCCACCCAAACA .094   AAA .042   AAAAAAAA .113   AACACACAAAACACCACAAA
CACCACCACCACCCAACCAAACA .047                                ACCACACAAAACCCCACAAA
CAACCAAAACCAAAACAACAAAA .033          .98             .98
CACCACCCCCACCCACCCAAACA .019
AAAACAACACCAACACAAACCAC .015
                          1.0
```

Examine haplotypes above 1 %

Connect with thin lines if > 1 %

Connect with thick lines if > 10 %

Display alleles as:
● letters
○ numbers
○ colored squares

Figure 4.2: Haploview LD display[12] with recombination rate plotted above (left) and haplotypes display (right)

4.1.1 Visual Analytics as a Combination of Automated and Visual Analysis – Success Stories

There exist a number of successful application areas in which the visual analytics approach has been used together with KDD methods. Four notable examples are discussed; bioinformatics and climate change (mentioned already in the motivation section), the pervasive problem of finding patterns in data, and spatio-temporal data mining (discussed in Chapter 5).

Bioinformatics. Bioinformatics is one of the areas where KDD methods have been used extensively in combination with visualisation methods. In fact, bioinformatics is arguably one of the great successes in the field of computational data analysis – the combination of biology and KDD has produced a whole new area of research. The multidisciplinary approach that combines biology, medicine and visualisation with advanced KDD methods have resulted to new scientific knowledge and has led to understanding and treatments for serious diseases such as cancer. The fact that KDD methods and algorithms are central in the bioinformatics field is recognised by the scientific community. The importance of the combination of such methods with visualisation can be concluded from the fact that, ten out of the fifty most-frequently cited articles in the Bioinformatics journal, currently the leading reference in the field, propose visual analysis tools or methods (see, for example, Figure 4.2, where an interactive visual interface is used for computation and analysis of linkage disequilibrium statistics and population haplotype patterns from primary genotype data). As the complexity of research increases, more and more researchers and companies are relying on visual analytics as an indispensable aid for decision making in bioinformatics. Another example of this trend is the widely use of BioConductor[2] for computational biology and bioinformatics that provides access to a large collection of KDD, machine learning and statistics methods together with advanced visualisation techniques.

Ten out of the fifty most-frequently cited articles in the Bioinformatics journal propose visual analysis tools or methods

Climate change. KDD is becoming increasingly important for measuring the impact of climate change. The massive volume of climate-related data gathered

[2] http://www.bioconductor.org/

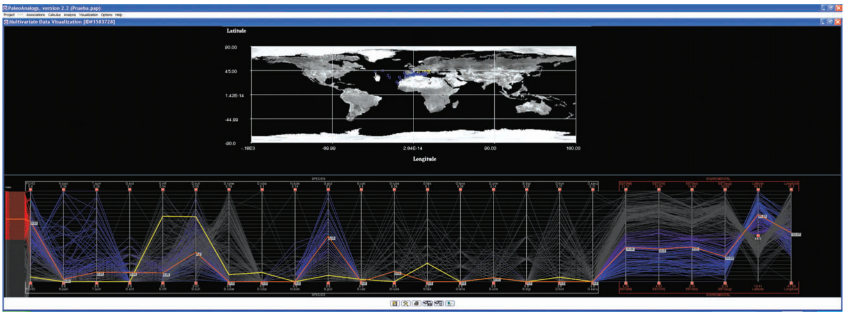

PaleoAnalogs, version 2.2 (Prueba.pap)
Project Associations Calculus Analysis Visualization Options Help
Multivariate Data Visualization [ID#1583728]
Latitude
90.00
45.00
1.42E-14
-44.99
-90.0
-.18E3
-89.99
2.84E-14
90.00
180.00
Longitude
SPECIES
ENVIROMENTAL

Figure 4.3: By means of a combination of an automatic pattern matching algorithm and an interactive visual interface, the expert is able to understand sea surface temperature changes over the past millions of years and use this to help predict future changes[109]

Visual analytics for predicting climate extremes

from remote and in-situ sensors is increasing rapidly. This vast climate database is augmented with proxy observations from the past and with data coming from simulations of global or regional climate models. In order to gain predictive insights on climate extremes and foresee events with potential impact, all these spatio-temporal data sources must be integrated, mined and presented in an understandable way. KDD methods can extract novel insights about climate extremes and regional change, while geographical information systems and multidimensional visualisation techniques can relate climate change and extremes to societal and ecological impacts. To illustrate this process, Figure 4.3 shows the distribution of micro-fossil species at different sites of the world through millions of years. These are used to reconstruct environmental features of the past by means of expert-steered k-nearest neighbour prediction; the use of linked parallel coordinate plots, maps and animations enables further analysis of the model.

Combining KDD and visualisation methods

Pattern identification. Searching for patterns is one of the main goals in KDD and it is applied to many varying domains such as medical, biological, financial and linguistic. Novel, exploratory data analysis tools and adaptive user interfaces have been developed by tailoring and combining existing KDD and visualisation methods. Variations of scatterplots, parallel coordinate plots, dendograms, heatmaps and many other visualisation techniques are used in combination with clustering, self organising maps, principal components analysis and other pattern extraction algorithms using colour linking and/or interactive brushing with excellent results. In the last five years, the success of this integration has contributed significantly to the use of visual analytics.

Spatio-temporal data mining. The availability of large repositories of spatial and spatio-temporal data has triggered the interest of the data mining community to the opportunities presented with these new data resources. However, this field presents new challenges and complexities: both the raw data (e.g., the traces of people moving in a city or flocks of animal migrating from one continent to another), and the extracted pattern (e.g., the aggregated flow from one zone of a city to another), may be too complex to be interpreted effectively by the analyst[79]. A new research field, identified by the European Project GeoPKDD[3][47], is emerging from the interaction of data mining technique with visual analytics tools for spatio-temporal data. An example of this interaction is presented in Andrienko et al.[6], where the knowledge extraction process is driven by the analyst, enabling efficient management of large datasets through stepwise refinement of the extracted model.

Combining visualisation and data mining for analysing mobility

4.1.2 Is Industry Ready for Visual Analytics?

Generally, the use of visual analytics has been well received by industry. Several companies have embraced this business model and are selling visual analytics tools and/or offering consultancy services to different industries. Arguably, the main reason to adopt this novel approach is that business users have witnessed the success stories of data mining, but they need to understand its results. Few KDD models are easy to understand and techniques need to be developed to explain or visualise existing ones. Furthermore, there is a need for techniques to translate the user's questions into the appropriate input for the data mining algorithms. Industry representatives see the need for intuitive and interactive KDD/visual analytics methods by which they can readily interact with the data and the underlying KDD models.

Techniques are required to understand the resulting KDD models

Due to its generality, KDD can be used in most visual analytics scenarios. Some good examples of its use are given below.

Marketing data. Data mining has appeared often in the media as an artificial intelligence technique capable of extracting interesting patterns out of customer activity, allowing effective marketing campaigns to launch new products and acquire new customers (see, e.g., Xtract Ltd[4]). With the rapid development of IT, exploring and analysing the vast volumes of commercial data is becoming increasingly difficult. Visual analytics can help to deal with the flood of information, since it provides a means of dealing with highly non-homogeneous and noisy data and involves the user in the data mining process (see, e.g., Visual Analytics Inc.[5]).

Process industry. The problem is that manufacturing systems are much better at collecting data than they are at helping one understand it (see, e.g., Spotfire[6]). In this context, visual analytics provides a way of making sense of the very large volume of data generated by factories related to quality parameters, process trends, maintenance events, etc. Thus, visual analysis can help solving problems, such as detecting anomalies and analysing their causes that, in turn, will lead to the development of more efficient and reliable processes.

Software industry. The complexity and size of industrial projects is currently growing rapidly, and hence there is a clear need for tools that assist during

[3] http://www.geopkdd.eu/

[4] http://www.xtract.com/

[5] http://www.visualanalytics.com/

[6] http://spotfire.tibco.com/Solutions/Manufacturing-Analytics/

the development, testing and deployment cycles. Currently, understanding the evolution of software has become a crucial aspect in the software industry. In the case of large software systems, gaining insight into the evolution of a project is challenging. Retrieving, handling and understanding the data poses problems that can only be solved by tightly coupling data mining and visualisation techniques. Thus, visual analytics can be effectively applied to support decision making in the software industry.

Pharmaceutical industry. The drug discovery process is very complex and demanding and often requires a cooperative, interdisciplinary effort. Despite the considerable methodological advances achieved through the years and the huge resources devoted to this enterprise, the results are disappointing. The recent completion of the human genome project has not only unearthed a number of new possible drug targets but has also highlighted the need for better tools and techniques for the discovery and improvement of new drug candidates. The development of these new tools will benefit from a deeper understanding of the drugs' molecular targets as well as from more friendly and efficient computational tools. With the flood of data across all aspects of the pharmaceutical industry, visual analytics is emerging as a critical component of knowledge discovery, development, and business[94].

4.2 State of the Art

The focus of the visual analytics community has been on interactive visual representation and exploration of data. But, the aim of the KDD community has focussed on developing computational methods that can be used to extract knowledge from data. There is a general awareness of the need to integrate visual analytics and KDD, but relatively few efforts have been made to address this issue. In this section, we present an overview of research and commercial systems in the following categories: statistical and mathematical tools, visually supported tools and combined methods. At the end of this section, we present several examples of KDD/visual analytics approaches from the fields of bioinformatics and graph visualisation.

As we have seen, the objective of knowledge discovery and data mining is to extract information from large datasets[55, 108]. This process is characterised by a series of operations (i.e. data pre-processing, data mining, data cleaning) that transform the data in various ways to obtain patterns and models that represent the implicit information within the data. Usually, the pre-processing steps produce a dataset in a suitable format for the data mining algorithms. The post processing steps transform the output of the mining into a form that can be understood by the analyst.

Data mining tasks are classified as predictive or descriptive

Data mining tasks can be divided into predictive tasks (e.g., classification, regression) and descriptive tasks (clustering, pattern mining, association rule discovery, etc.). In the former case, the data is analysed to build a global model, which is able to predict the value of target attributes based on the observed values of the explanatory attributes. In descriptive tasks, the objective is to summarise the data using local patterns that describe the implicit relationship and characteristics of the data itself. However, as discussed earlier, existing methods support limited user interaction and are mainly designed for homogeneous data sources. Some attempts have been made to enhance data mining with visualisation providing advanced interactive interfaces. A survey of the state of the art of current and proposed solutions that facilitate sense-making for interactive visual exploration of billion record datasets, is provided in 'Extreme visualization'[99]. Several interactive tools for information visualisation, designed for specific data types have been presented in the literature. These include graph visualisation[1], time series interactive search[20] and network visualisation[9].

We now give an overview of some research and commercial systems in the context of data mining and visualisation, categorised as follows:

- Statistical and mathematical tools
- Specific algorithmic tools
- Visual analytics libraries
- Visual data mining tools
- Web tools and packages
- Scientific visualisation tools
- Combined methods
- Computational information design

Statistical and mathematical tools. Statistical analysis has a long history of visualising the results as time series, bar charts, plots and histograms. Examples of tools providing statistical and mathematical visualisation are R[7], Matlab[8],Mathematica[9] and SAS[10] tools for statistical computing and graphics.

Specific algorithmic tools. Algorithmic tools have been developed by the research communities for a specific task or problem. Examples are Graphviz[11](see Figure 4.4), open source graph visualisation software, or Pajek[12], which is more focused on the analysis of social and complex network data by taking advantage of network/graph visualisation.

Visual analytics libraries. One example, originally aimed at providing expertise in data visualisation and visual design is BirdEye[13], a community project to advance the design and development of a comprehensive open source information visualisation and visual analytics library.

Visual data mining tools. Visual data mining creates visualisations to reveal hidden patterns from datasets. The need of new methods in data analysis has

[7] http://www.r-project.org/
[8] http://www.mathworks.com/
[9] http://www.wolfram.com/
[10] http://www.sas.com/technologies/bi/visualization/visualbi/index.html
[11] http://www.graphviz.org/
[12] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[13] http://code.google.com/p/birdeye/

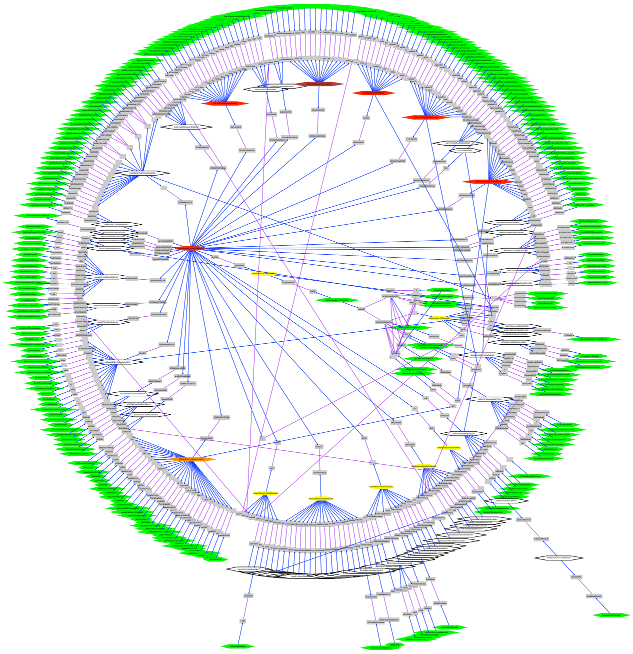


Figure 4.4: Radial layout graph visualisation made using Graphviz. A real-world network containing 300 sites over 40 countries. The diagram was made to trace network incidents and to support maintenance. Used with permission of AT&T

launched the field. Several products are on the market; often focused on 'business intelligence' such as marketing, risk analysis, sales analyses and customer relationship management. Some examples are:

KNIME[14] is a modular data exploration platform that enables the user to visually create data flows (or pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models.

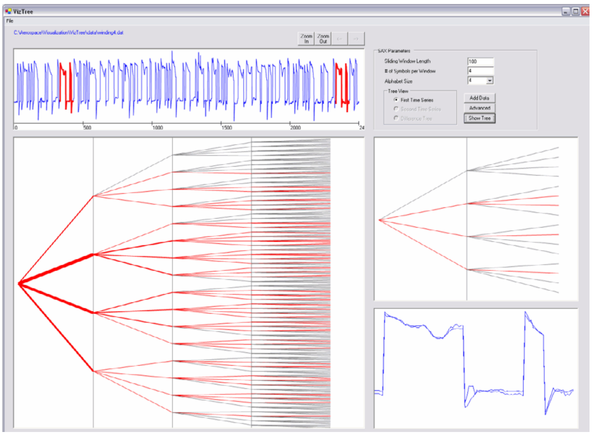
Weka[15] is a collection of machine learning algorithms for data mining tasks, which allows the user to create pipelines in order to perform data pre-processing, classification, regression, clustering, association rules, and visualisation. It is open source code, developed in Java.

Similarly to Weka, RapidMiner[16] is an environment for machine learning and data mining tasks, which allows the user to create data flows, including input and output, data pre-processing and visualisation. It also integrates learning schemes and attribute evaluators from the Weka learning environment.

[14] http://www.knime.org/

[15] http://www.cs.waikato.ac.nz/ml/weka/

[16] http://rapid-i.com/



SAX Parameters
Sliding Window Length 100
# of Symbols per Window 4
Alphabet Size 4

Tree View
● First Time Series
○ Second Time Series
○ Difference Tree

Figure 4.5: VizTree: The top panel is the input time series. The bottom left panel shows the subsequence tree for the time series. The top right window shows a zoomed in region of the tree, and the bottom window plots the actual subsequences when the user clicks on a branch

Web tools and packages. An increasing number of tools are available online, but user interaction becomes more complicated and difficult to model and optimise, when used remotely. With these tools, users can create visualisations using their own data. An example of an online social data analysis tool is ManyEyes[17], an IBM application for social data analysis.

Scientific visualisation tools. Scientific visualisation is the representation of data graphically as a means to gain understanding and insight into the data. It involves research in computer graphics, image processing, high performance computing, and many other areas. Scientific visualisation tools are often adopted for modelling complicated physical phenomenon. An example in the field of natural science is Gravity waves[18], where the Globus Toolkit has been used to harness the power of multiple supercomputers and simulate the gravitational effects of black-hole collisions. Other examples come from geography (e.g., terrain rendering) and ecology (e.g., climate visualisation).

Combined Methods. There have been some attempts to combine data mining and visualisation. For example, some concentrate on the analysis of time series by using tree visualisations and interactions (VizTree, see Figure 4.5), or propose a combination of visual data mining and time series (Parallel Bar Chart, see Figure 4.6), or combine KDD concepts and visualisations (Statigrafix[19], see Figure 4.7). However, each one lacks either effective visualisation, automatic data mining or requires a strong expertise in the application field.

[17] http://manyeyes.alphaworks.ibm.com

[18] http://www.anl.gov/Media_Center/logos20-2/globus01.htm

[19] http://statigrafix.com

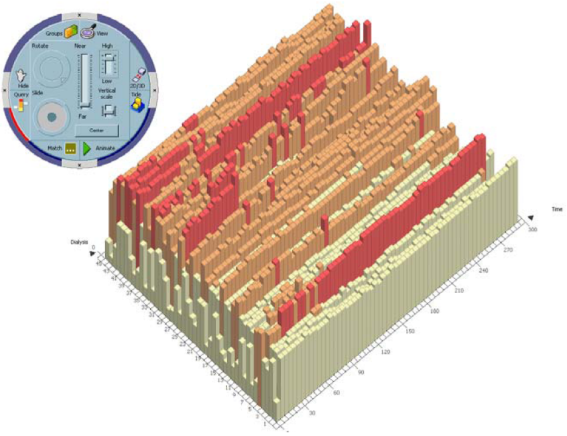


Figure 4.6: Parallel Bar Chart[30] visually represents each time-series in a bar chart format where the X axis is associated with time (the axis on the right), the Y axis with the value (height of a bar) of the series at that time and the axis on the left identifies the different time-series, ordered by date

Finally, in the bioinformatics and graph visualisation fields there are several examples of KDD/visual analytics approaches. For instance,

JUNG (Java Universal Network/Graph Framework[20]) is a software library that provides a common and extensible language for the modelling, analysis, and visualisation of data that can be represented as a graph or network. It is written in Java and it includes implementations of a number of algorithms from graph theory, data mining, and social network analysis, such as routines for clustering, decomposition, optimisation, random graph generation, statistical analysis, and calculation of network distances, flows, and importance measures.

HCE[21] (Hierarchical Clustering Explorer) for interactive exploration of multidimensional data. Genome researchers adopt cluster analysis to find meaningful groups in microarray data. Some clustering algorithms, such as k-means, require users to specify the number of clusters as an input, but users rarely know the right number beforehand. Other clustering algorithms automatically determine the number of clusters, but users may not be convinced of the result since they had little or no control over the clustering process. To avoid this dilemma, the Hierarchical Clustering Explorer (HCE, see Figure 4.8) applies the hierarchical clustering algorithm without a predetermined number of clusters, and then enables users to determine the natural grouping with interactive visual feedback (dendrogram and colour mosaic) and dynamic query controls.

[20] http://jung.sourceforge.net

[21] http://www.cs.umd.edu/hcil/hce/

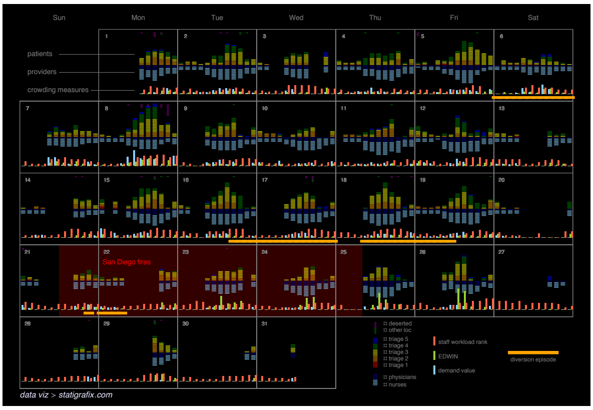


Figure 4.7: Calendar-template data visualisation of datasets captured at the visual analytics San Diego Health Service's Emergency Dept in Oct 2007 (Source: Alan Calvitti, statigrafix.com)

BicOverlapper[22] is a framework to support visual analysis of gene expression by means of biclustering. In order to improve the visualisation of biclusters, a visualisation technique (Overlapper) is proposed to simultaneously represent all biclusters from one or more biclustering algorithms, based on a force-directed layout. This visualisation technique is integrated in BicOverlapper, along with several other visualisation techniques and biclustering algorithms.

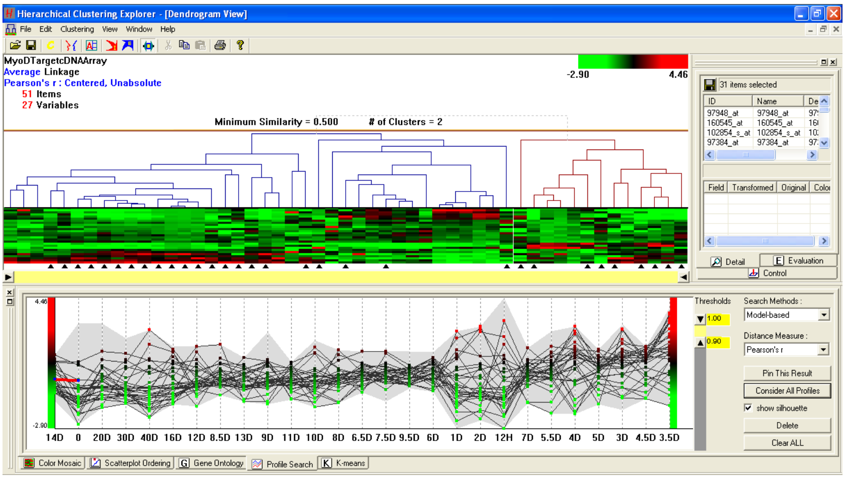
Computational Information Design. Similarly to the previous category, Computational Information Design has been suggested by Ben Fry from the Massachusetts Institute of Technology[23]. In an attempt to gain better understanding of data, fields such as information visualisation, data mining and graphic design are employed, each solving an isolated part of the specific problem, but failing in a broader sense: there are too many unsolved problems in the visualisation of complex data.

4.3 Challenges

4.3.1 Introduction

The developers of visual analytics applications face several fundamental challenges when attempting to develop integrated iterative methodologies that involve information gathering, data pre-processing, knowledge representation, interaction and decision making. One of the main purposes of this chapter is to establish the degree to which existing techniques and approaches can be integrated, and, in a wider sense, how the human-computer integration might be facilitated[14].

[22] http://vis.usal.es/bicoverlapper/
[23] http://benfry.com/phd/



Hierarchical Clustering Explorer - [Dendrogram View]

MyoDTargetcDNAArray
Average Linkage
Pearson's r : Centered, Unabsolute
51 Items
27 Variables

Minimum Similarity = 0.500
# of Clusters = 2

31 items selected

ID	Name	De
97948_at	97948_at	97
160545_at	160545_at	16
102854_s_at	102854_s_at	10
97384_at	97384_at	97

Thresholds: 1.00, 0.90
Search Methods : Model-based
Distance Measure : Pearson's r
☑ show silhouette

Figure 4.8: Hierarchical Clustering Explorer for interactive exploration of multidimensional data[97]

The five categories of Grand Challenges

According to Thomas and Cook[111], the so-called ’Grand Challenges’ faced by visual analytics can be grouped into five categories: *analytical reasoning*, *visual representations and interaction techniques*, *data representations and transformations*, *production, presentation and dissemination*, and *moving research into practice*. The first category (analytical reasoning) refers to the reasoning frameworks by which users derive insights or discover knowledge to support the decision making process. These frameworks provide the foundation for applying specific transformations, visual techniques or other operations, on the data. The second category (visual representations) covers all interactive means, methods and techniques that enable visual representation of data. The third category (data representations and transformations) refers more to the specific ways that data is represented, as well as the operations upon data (which might be noisy, incomplete, or uncertain). Representations refer to the fundamental ’structure’ of the data within an application, usually non-intuitive to users, but responsible for facilitating data transformations, calculations, etc. The fourth category (production, presentations and dissemination) refers to user activity and interaction. Finally, the fifth category (moving research into practice) refers to the practical application of methods and techniques.

In KDD, analytic reasoning, and data representations and transformations are highly relevant

In terms of KDD, the first (analytical reasoning) and third (data representations and transformations) categories are highly relevant. Next we analyse in more detail several specific technical challenges in both of these categories relating to KDD.

4.3.2 Data Issues

While data types, formats and characteristics are a key part of the motivation underlying visual analytics approaches, they also represent a key challenge to operational implementations. Here we focus specifically on data mining issues. For visual analytics to be able to fulfil its true promise, we need the capability to integrate both heterogeneous and large datasets. This includes:

- (qualitative) textual data,
- data stored in (distributed) databases,
- data received from sensors,
- spatial data such as satellite imagery,
- audio and video.

KDD approaches tend to focus on specific types of quantitative data

KDD approaches tend to focus more on specific types of (quantitative) data, however, approaches for other data types are emerging[93, 40]. There are several levels of complication:

- Some of the data could be arriving in real-time, so that ways to manage this (storage, management and interactive analysis and visualisation) are required.
- Some of the data will be of variable quality, therefore we need to know as much as possible about the data itself.
- Data may be incomplete, so we need to know what is missing, as well as having ways to handle or manage the missing data.
- Data may be of different (spatial) scales, and therefore require transformations/mappings to be compatible with other data.

The means by which data should be managed and distributed is addressed in more detail in Chapter 3, and for spatio-temporal data in Chapter 5. However, to support the data mining initiatives in visual analytics, we require methods for data cleaning, integration, data fusion etc. If we are to achieve 'real time' analytics, then the cleaning and integration methods should be automated and fast. These problems are non-trivial and significant developments are required before data mining can be integrated into a visual analytics platform.

Data standards or metadata are required

A necessary feature of these developments will be the adoption of standards across different visual analytics toolsets and software environments. These data standards[24] do not just concern data formats. More fundamentally, we require *metadata*, or documentation of the data itself: lineage/source(s), formats, method of collection, accuracy and completeness in order to support data mining approaches.

4.3.3 Visual Analytics Platforms

One of the main goals of KDD is pattern extraction. This can be applied in many application domains, as discussed in the following section. Most of the existing visual analytics related software provides some common functionality (statistical analysis, graphing tools, algorithms, visualisation), but as noted in the previous section, data needs to be represented in a format suitable for the analysis algorithms.

[24] See for example http://www.iso.org/

Most software is developed for a specific task

Functions such as linking and brushing, scatterplots and clustering are basic functions, yet are missing from many software environments. A key reason for this is that most software has developed out of the specific needs of a particular discipline, and therefore is geared towards specific types of decision making. As noted in Section 4.2, a variety of tools and environments exist which address different aspects of visual analytics. Examples include KNIME[25] and OECD explorer[26], which are developed specifically for geographic data. These are significantly different from business intelligence tools, which focus specifically on marketing and management strategies and risk analysis, and differ significantly from bioinformatics tools.

The fundamental challenge, given that we are likely to see ongoing development of these heterogeneous toolsets, is to provide the functionality so that users can easily switch between visual analytics tools and data sources. To achieve this, data sources will have to be integrated directly using applications programming interfaces (APIs). Clearly, building specific visualisation tools for every use case is not a feasible solution. Generic tools are required that can be customised with appropriate algorithms and visual tools.

Interdisciplinary initiatives are required

Many of the commonly used data mining algorithms are already well-developed and do not require expert users in order to be applied. For example, even a novice user can use a clustering algorithm, provided it has adequate documentation. This chapter has identified a wide variety of emerging software platforms both within, and closely related to visual analytics. Many of these have their own implementation of various algorithms. It has also been noted that an initial community repository for information visualisation and visual analytics algorithms is already underway (BirdEye). In order to facilitate KDD and data mining approaches, cross-disciplinary initiatives are required. Not simply to provide algorithms, but to inform the wider community (KDD, information visualisation and visual analytics) about their functionality and requirements. Cross-platform standards could also play an important role in this, in terms of defining a core set of widely used algorithms, as well as frequently used visualisation techniques.

One further issue is the provision of distributed collaboration between disciplinary experts. This has a major implication for visual analytics platform in sharing very large datasets over the Internet. Further investigation is required into the kinds of technologies that can facilitate this.

4.3.4 Towards Visually Controlled Data Mining

Advanced KDD methods require expertise

The current data mining methods support only limited user interaction. Also, existing KDD methods are not directly applicable to visual analytics scenarios. This is essentially because the more advanced KDD methods are often non-intuitive, in that a significant degree of experience is required for their successful application. As well as user expertise, many KDD methods

[25] http://www.knime.org

[26] http://www.oecd.org/gov/ regional/statisticsindicators/explorer

require substantial processing time and therefore place significant demands on computer hardware.

In complex domains, the models and patterns extracted by traditional KDD approaches may also be difficult to interpret, and relevant information may be hidden within large results sets. It is envisaged that visual analytics methods may simplify the presentation and evaluation of the models extracted. These issues should be addressed if KDD is to be able to make a significant contribution to visual analytics (and vice-versa). Work is required on identifying and implementing means by which this might occur.

Towards 'visually controlled mining'

In a review of visual analytics, information visualisation and data mining literature, Bertini and Lalanne[14] classify recent literature within these disciplines along a 'continuum' of approaches, ranging from pure data mining to pure visualisation and propose new research questions and directions. Puolamäki et al.[91] identifies a new class of data mining methods, *visually-controlled data mining*.

For a data mining method to be useful in visual analytics it should be:

1. Fast enough – sub-second response is needed for efficient interaction.
2. Parameters of the method should be representable and understandable using visualisations.
3. Parameters should be adjustable by visual controls.

Efficient interaction represents a significant hurdle in bringing KDD to visual analytics, as noted above. In terms of the second and third requirements, further investigation into what types of 'visual controls' are required to manage and adjust the algorithms is required.

Hardware and software issues are still open

There are hardware, software, and algorithmic issues involved in developing the kind of mixed-initiative approach identified above. From a hardware point of view, machine specifications should be able to handle the computations adequately. The software should be as application-independent as possible, perhaps following the plugin topology favoured by many open source research tools. These algorithms must be both efficient and robust. One could conceive of a repository for plugins to various existing and emerging platforms (similar to BirdEye as noted above), maintained for quality control and ongoing community development.

KDD and information visualisation communities should collaborate more

The research on visual analytics, using visualisation and interaction methods to analyse large datasets, and data mining have evolved separately. However, at the current time, communication and interaction between both research communities has just started in the form of workshops under the umbrella of their main international conferences (such as SIGKDD and VisWeek). The success of these events has confirmed that there are significant benefits from bringing these communities together. A challenge lies in establishing collaboration between these research communities, so that we can focus on applications. This requires that domain experts from the data mining/KDD, visual analytics and information visualisation communities, collaborate on the specific ways that the two approaches can complement one another.

4.3.5 Research and Evaluation

It is possible to identify three general categories relating to research and evaluation from the perspective of KDD and data mining. These relate to *evaluation*, *research development* and *collaboration*.

Evaluation is difficult - it is unclear what a good solution is

The evaluation of visual analytics approaches is regarded as difficult. It requires specific criteria on how to judge a visual analytics solution or application. The evaluation also requires new measures. While significant criteria exist in the separate fields which visual analytics seeks to draw together, it is difficult to envisage how these might fit together in some unified way. For example, in the discipline of visualisation, a number of techniques and criteria exist for evaluation of results such as assessment of the effectiveness of the result (through user evaluations). Similarly, it is relatively easy to judge the outcome of traditional KDD approaches through validation of the results with reference data. However, in terms of combined KDD/visual analytics solutions, it is still unclear what a 'good' solution or application should look like. We therefore expect to see ongoing development of (design and implementation) guidelines, to help identify a base upon which we can build further.

Collaboration requires workflow sharing

In terms of research collaboration, significant technical challenges exist. Several of these were identified above. The general question is "how will collaborative data mining/visual analytics approaches work?" They would require facilities for transfer of data, but also of custom algorithms or even better, entire data workflows in some way. Some collaborative approaches are currently underway, but these are by no means well developed in terms of the requirements of a mixed-initiative *visually-controlled mining* approach. More work is required to investigate the possibilities of data, software, and even full workflow-sharing approaches and their respective practical limitations.

In terms of development of the research field itself, this brings about a sociological and very practical question: how to get the referees to accept visual analytics/KDD papers? Special issues are perhaps a temporary solution, but ultimately, alongside the rapid development of software and integrated solutions, we would expect to see several dedicated academic journals to support the research discipline.

4.4 Opportunities

While the key issues identified in the previous section are significant barriers to progress, several of these also represent major opportunities. Below we discuss four general categories of these: the development of generic tools and methods, regulation and quality control, visualisation of models, and linkage of KDD and visualisation communities.

Need for a repository of generic tools and methods

Firstly, generic components are needed in order to stimulate research. This obviously includes algorithms, i.e., methods, and software libraries (preferably open source for maximal spread). It is possible to envisage some kind of 'repository' for things like plugins and software libraries with associated documentation to promote access to a range of research communities. It has already been identified that here will need to be some kind of regulation and quality control for this to develop in a controlled manner. The major opportunity in this sense is to provide the guidelines and framework for these components to develop.

In addition to the visualisation of the data we should move to *visualisation of models*. For example, why are two points clustered together? If we know some groups of people and their social interaction network, what kind of an interaction model would help to explain the data? The initial steps in achieving this are relatively simple: just bring the basic methods to visualisation of model spaces. Data mining models contain information about the phenomena. As discussed earlier in this chapter, initial approaches are already underway.

Visualisation of models could be useful

The final opportunity, already identified above, relates directly to the above issue and involves collaboration between KDD and visualisation communities. The two communities certainly share an awareness that their approaches have significant overlap. While also a cultural challenge, there are significant opportunities for cross-pollination of approaches, methods and techniques. Ways to encourage and stimulate this might be through for example expert groups or mixed-initiative 'challenges' at key research conferences. From the review in Section 4.2, as well as the VAKD '09 Workshop[91], it would appear that we are close to a breakthrough.

Collaboration between KDD and visualisation communities should be encouraged

4.5 Next Steps

Visual analytics is an emerging research field that combines the strengths of information visualisation, knowledge discovery in databases, data analysis and mining, data management and knowledge representation, human perception and user interaction. In this report we discussed the scope of visual analytics and analysed several challenges and opportunities that stem from this very promising field. Our investigation and analysis suggest that there is a clear need for integration of visual analytics and knowledge discovery and for building a community. The merging of the KDD and visual analytics communities could be achieved by two main approaches: bottom-up and top-down.

A bottom-up approach would include several dissemination activities, such as workshops, conferences and journal special issues. The VAKD '09 Workshop on Visual Analytics and Knowledge Discovery, organised by us, was a great success. The second VAKD workshop[27] will be organised in Sydney in conjunction with the 10th IEEE International Conference on Data Mining (ICDM 2010). A series of VAKD workshops will promote the development of novel visual analytics ideas and bring visual analytics research communities

[27] http://www.mpi-inf.mpg.de/conferences/VAKD10/

closer. Further, we should organise several collaborative research projects that would involve leading research groups.

Historically, challenges have been traditionally a good way to catalyse research. In VAKD ’09 workshop, the authors were encouraged to address the tasks of the IEEE VAST 2008 visual analytics challenge[50], which contain both visual analytics and KDD angles in the performance evaluation. We should organise KDD challenges in the spirit of visual analytics. For example, the evaluation of a classification algorithm should not just be the classification accuracy but should also involve several other factors, such as, user interaction, visualisation, etc. It would be essential to include both visual analytics and KDD aspect in the Grand Challenges stated in Section 4.3.1.

Knowledge discovery approach should be reconsidered and data mining processes should evolve in the direction of visual analytics processes. As part of this process, we should consider new performance evaluation measures, as it is clear that we will need more than just algorithmic measures.

One major contribution would be to develop novel visual analytics approaches that enable visualisation for both the data and the underlying model. So far, standard visual analytics only allowed visualisation of the data. For this purpose, several existing information visualisation techniques could be used and further extended and tailored, with the help of data analysis methods, to produce useful and usable data model representations.

Current data mining methods support limited user interaction. For a data mining method to be optimal in a visual analytics application, it should be fast (sub-second response is needed for efficient interaction) and the parameters of the method should be understandable and adjustable by visual controls. By using visual interaction, the visually-controlled data mining process will be more efficient than by ’blindly’ applying some data mining method, or by just interactively visualising data.

Another challenge for visual analytics is scalability of algorithms and heterogeneous data. Special emphasis should be given to methods that scale well and are applicable for indexing, accessing, analysing and visualising huge datasets. At the same time, a new trend in the area of data mining is being able to handle and combine data from large and possibly conflicting sources. Developing visual analytics algorithms that can handle this information overload and ambiguity efficiently would be another major contribution to the visual analytics community.

It is important to consider the application aspect of visual analytics. As also mentioned by Keim et al.[66], for the advance of visual analytics, several application challenges should be mastered including physics, astronomy, business, security, economics, biology and health, engineering and mechanics and GIS. Visual analytics applies to a wide range of different application fields and for our part we should encourage and enforce interdisciplinary collaboration. All the aforementioned communities should be investigated extensively and visual analytics algorithms should be developed that are tailored to their needs.