2 Visual Analytics

Visual analytics is not easy to define, due to its multi-disciplinary nature involving multiple processes and the wide variety of application areas. An early definition was "the science of analytical reasoning facilitated by interactive human-machine interfaces"[125]. However, based on current practice, a more specific definition would be: "Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets".

Visual analytics combines automated analysis with interactive visualisations

So, in terms of the goal of visual analytics, we can elaborate on this definition to state that visual analytics is the creation of tools and techniques to enable people to:

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

In Section 2.2 we will look at how visual analytics strives to achieves these goals in terms of the high-level processes required to generate knowledge from data, and then in Section 2.3 in terms of the many scientific disciplines that contribute to visual analytics. But firstly, in order to give a sense of the social and economic importance of visual analytics, as well as the scale of the data being dealt with, we will look at some typical uses.

2.1 Application of Visual Analytics

Visual analytics is essential in application areas where large information spaces have to be processed and analysed. Major application fields are physics and astronomy. For example, the discipline of astrophysics offers many opportunities for visual analytics techniques: massive volumes of unstructured data, originating from different directions of space and covering the whole frequency spectrum, from continuous streams of terabytes of data that can be recorded and analysed. With common data analysis techniques, astronomers can separate relevant data from noise, analyse similarities or complex patterns, and gain useful knowledge about the universe, but the visual analytics approach can significantly support the process of identifying unexpected phenomena inside the massive and dynamic data streams that would otherwise not be found by standard algorithmic means. Monitoring climate and weather is also a domain which involves huge amounts of data collected by sensors throughout the world and from satellites, in short time intervals. A visual approach can help

Monitoring the climate involves huge amounts of data from many different sources to interpret these massive amounts of data and to gain insight into the dependencies of climate factors and climate change scenarios that would otherwise not be easily identified. Besides weather forecasts, existing applications visualise global warming, melting of the poles, the stratospheric ozone depletion, as well as hurricane and tsunami warnings.

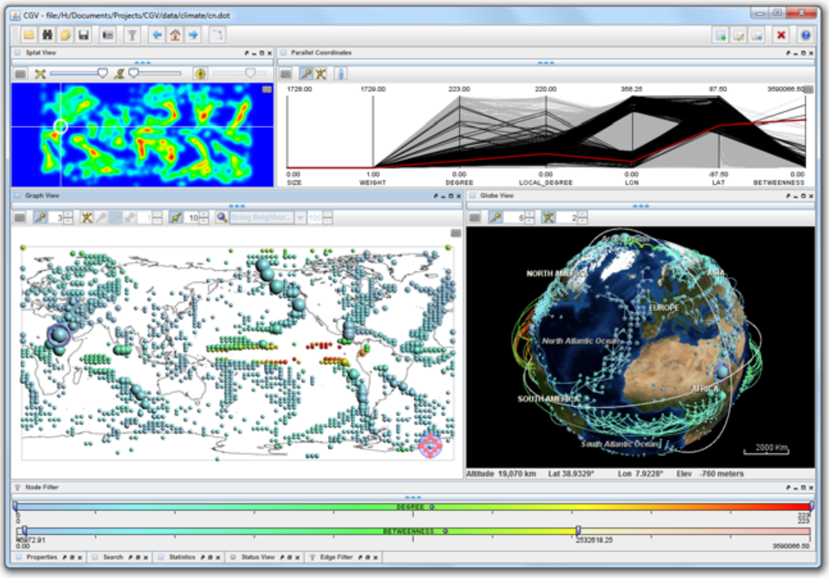

Figure 2.1: Visual analytics in action: Visual support for the simulation of climate models provided by CGV[113] (Coordinated Graph Visualization), a highly interactive graph visualisation system. To support different visualisation tasks, view ensembles can be created dynamically with the help of a flexible docking framework. CGV includes enhanced dynamic filtering, graph lenses, edge-based navigation, in addition to augmented navigation with infinite grid and radar view. Data source: Potsdam Institute for Climate Impact Research

More than 210 billion emails, 4 billion SMS and 50 million tweets per day

In the domain of emergency management, visual analytics can help determine the on-going progress of an emergency and identify the next countermeasures (e.g., construction of physical countermeasures or evacuation of the population) that must be taken to limit the damage. Such scenarios can include natural or meteorological catastrophes like flood or waves, volcanoes, storm, fire or epidemic growth of diseases (e.g. N1H1 virus), but also human-made technological catastrophes like industrial accidents, transport accidents or pollution. Visual analytics for security and geo-graphics is an important research topic. The application field in this sector is wide, ranging from terrorism informatics, border protection, path detection to network security. Visual analytics supports investigation and detection of similarities and anomalies in very large datasets. For example, on a worldwide scale, per day there are upwards of 210 billion emails, 4 billion SMS messages, 90 million tweets and the number of IP data packets exceeds 9000 billion. As an example of document processing on a European level, the Europe Media Monitor collects news documents from 2,500 news sources: media portals, government websites, and news agencies and processes 80,000-100,000 articles per day in 43 languages.

Europe Media Monitor collects and processes 100,000 news articles per day in 43 languages

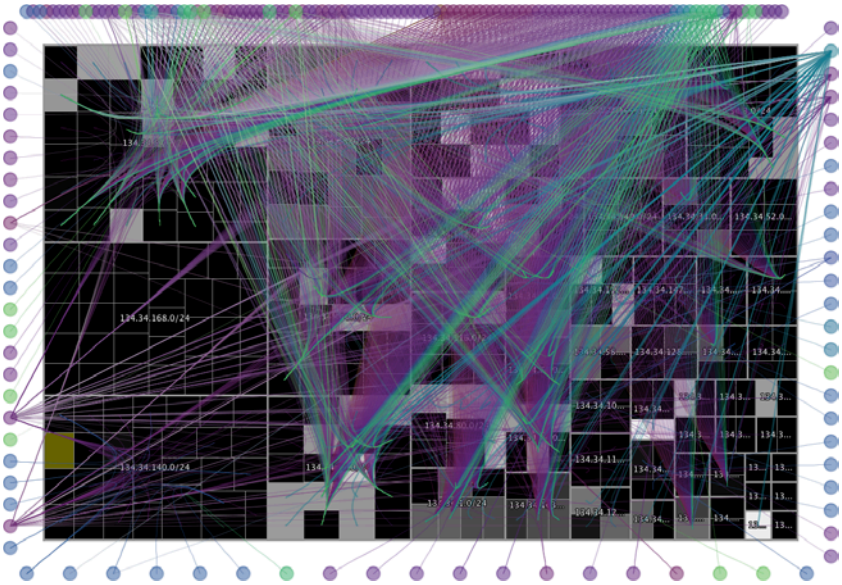

Figure 2.2: Visual analytics in action: Analysis of a distributed network attack on the SSH service of a university network using NFlowVis[76]. The TreeMap in the background represents the internal network structure with hosts as rectangles on the lowest level and external hosts as coloured dots on the outside. Hierarchical edge bundles reveal communication patterns such as the distributed attack from the hosts on the upper side

In biology and medicine, computer tomography, and ultrasound imaging for 3-dimensional digital reconstruction and visualisation produce gigabytes of medical data. The application area of bio-informatics uses visual analytics techniques to analyse large amounts of biological data. From the early beginning of sequencing, scientist in these areas face unprecedented volumes of data, like in the human genome project with three billion base pairs per human. Other new areas like proteomics (studies of the proteins in a cell), metabolomics (systematic study of unique chemical fingerprints that specific cellular processes leave behind) or combinatorial chemistry with tens of millions of compounds, add significant amounts of data every day. A brute-force computation of all possible combinations is often not possible, but interactive visual approaches can help to identify the main regions of interest and exclude unpromising areas.

Another major application domain for visual analytics is business intelligence. The financial market with its hundreds of thousands of assets generates large amounts of data on a daily basis, which results in extremely high data volumes

More than 300 million VISA credit card transaction per day over the years. For example it is estimated that there are more than 300 million VISA credit card transaction per day. The main challenge in this area is to analyse the data under multiple perspectives and assumptions to understand historical and current situations, and then monitoring the market to forecast trends or to identify recurring situations. Other key applications in this area are fraud detection, the analysis of consumer data, social data and data associated with health care services.

Further application examples of visual analytics are shown in Figures 2.5 and 2.6 at the end of this chapter.

2.2 The Visual Analytics Process

Tight coupling of automated and visual analysis through interaction

The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. Figure 2.3 shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the visual analytics process.

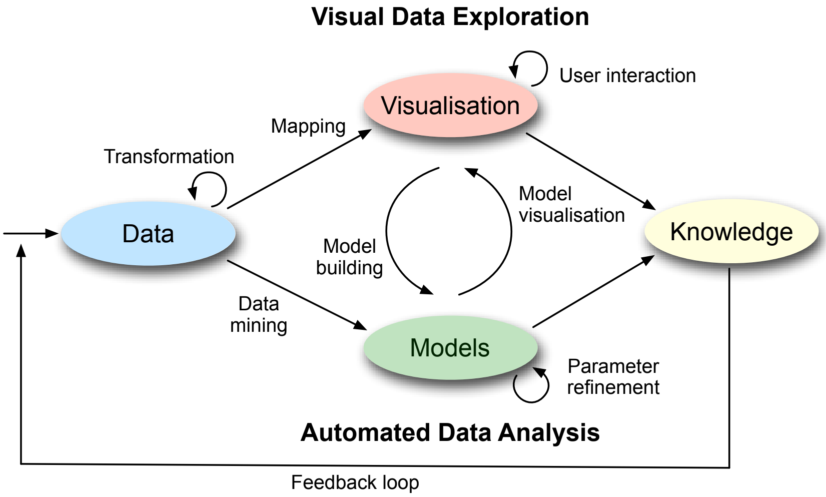

Figure 2.3: The visual analytics process is characterised through interaction between data, visualisations, models about the data, and the users in order to discover knowledge

In many application scenarios, heterogeneous data sources need to be integrated before visual or automatic analysis methods can be applied. Therefore, the first step is often to preprocess and transform the data to derive different representations for further exploration (as indicated by the *Transformation* arrow in Figure 2.3). Other typical preprocessing tasks include data cleaning, normalisation, grouping, or integration of heterogeneous data sources.

After the transformation, the analyst may choose between applying visual or automatic analysis methods. If an automated analysis is used first, data mining methods are applied to generate models of the original data. Once a model is created the analyst has to evaluate and refine the model, which can best be done by interacting with the data. Visualisations allow the analysts to interact with the automatic methods by modifying parameters or selecting other analysis algorithms. Model visualisation can then be used to evaluate the findings of the generated models. Alternating between visual and automatic methods is characteristic for the visual analytics process and leads to a continuous refinement and verification of preliminary results. Misleading results in an intermediate step can thus be discovered at an early stage, leading to better results and a higher confidence. If visual data exploration is performed first, the user has to confirm the generated hypotheses by an automated analysis. User interaction with the visualisation is needed to reveal insightful information, for instance by zooming in on different data areas or by considering different visual views on the data. Findings in the visualisations can be used to steer model building in the automatic analysis. In summary, in the visual analytics process, knowledge can be gained from visualisation, automatic analysis, as well as the preceding interactions between visualisations, models, and the human analysts.

Steer model building with visual findings

The visual analytics process aims at tightly coupling automated analysis methods and interactive visual representations. The guide to visually exploring data “Overview first, zoom/filter, details on demand”, as proposed by Shneiderman[98] in 1996 describes how data should be presented on screen. However, with massive datasets at hand, it is difficult to create an overview visualisation without losing interesting patterns, which makes zooming and filtering techniques effectively redundant as the users is given little information of what to examine further. In the context of visual analytics, the guide can usefully be extended to "Analyse first, show the important, zoom/filter, analyse further, details on demand”[65] indicating that it is not sufficient to just retrieve and display the data using a visual metaphor; rather, it is necessary to analyse the data according to its value of interest, showing the most relevant aspects of the data, and at the same time providing interaction models, which allow the user to get details of the data on demand.

Analyse first, show the important, zoom/filter, analyse further, details on demand.

2.3 Building Blocks of Visual Analytics Research

Visual analytics integrates science and technology from many disciplines, as shown in Figure 2.4. Visualisation is at the heart of the system, not only is it the means to communicate data values or the results of some analysis, but it is also increasingly being used to monitor processes in other disciplines, such as data management and data mining. We will now briefly consider the disciplines that contribute towards visual analytics.

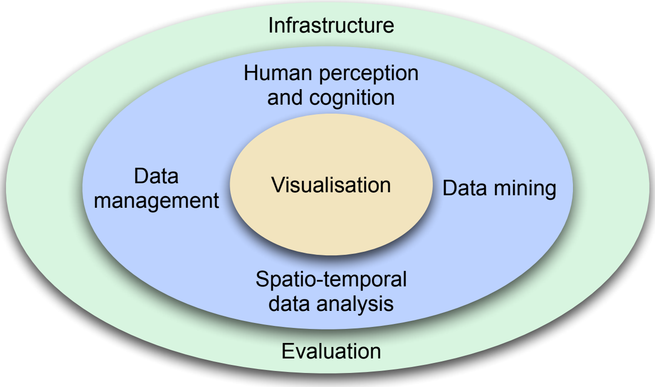

Figure 2.4: Visual analytics integrates visualisation with core adjacent disciplines and depends on the availability of appropriate infrastructure and evaluation facilities

Visualisation

Visualisation has emerged as a new research discipline during the last two decades. It can be broadly classified into scientific and information visualisation.

Scientific visualisation for 3D phenomena, such as fluid flow or molecular structures

Scientific visualisation is primarily concerned with visualising 3-dimensional (3D) data from the world of engineering, biology (whole body scans down to molecular structures), meteorology, cosmology, and so on, with the aim to represent the data, often temporal, as physical entities, such as surfaces, volumes and flows. A survey of current visualisation techniques can be found in the ’visualization handbook’[56]. Often, 3D scalar fields are visualised by isosurfaces (3D contour) or semi-transparent point clouds. Also, in recent years, significant work has focused on the visualisation of complex 3D flow data, such as in aerospace engineering[114]. While current research has concentrated mainly on improving the efficiency of the visualisation techniques in enabling interactive exploration, more and more methods have been developed to automatically derive relevant visualisation parameters. In addition, interaction techniques such as focus & context[70] have gained importance in scientific visualisation.

Information visualisation for abstract data, often with many dimensions

Information visualisation has developed methods for the visualisation of abstract data where no explicit spatial references are given[104]. Typical examples include business data, demographics data, social networks and scientific data. Not only are we having to deal with huge volumes but the data often comprises of hundred of dimensions. Also, in addition to standard numeric and textual data types, some of these dimensions may be complex data types such as graphic, video, sound, and sophisticated data types now defined for the semantic web. The data values cannot be naturally mapped to 2D or 3D display space, as with scientific visualisation, and standard charting techniques such as x-y plots, line graphs and bar-charts are ineffective with large multi-dimensional datasets. Moreover, as mentioned earlier, the capacity to interact with the data is extremely important. Novel visualisations have been developed such as parallel coordinates, treemaps, glyph and pixel-based visual data representations, to name just a few, together with a variety of techniques to reduce display clutter[41]. There are also special techniques for visualising structured data, such as graph-based approaches for networks, and for visualising spatial and temporal dimensions as found in geo-visualisation (described later in more detail). Furthermore, some visualisations make use of automatic data analysis techniques such as clustering or dimensional reduction as a preprocessing step prior to visualisation.

Data Management

The efficient management of data of various types and qualities is a key component of visual analytics, as it typically provides the input of the data, which is to be analysed. Generally, a necessary precondition to perform any kind of data analysis is an integrated and consistent database. Database research has, until the last decade, focused mainly on aspects of efficiency and scalability of exact queries on uniform, structured data. With the advent of the Internet and the easy access it provides to all kinds of diverse data sources, the focus of database research has shifted towards integration of this heterogeneous data. Finding effective representations for different data types such as numeric data, graphs, text, audio and video signals, semi-structured data, semantic representations and so on is a key problem of modern database technology. But the availability of heterogeneous data not only requires the integration of many different data types and formats but also necessitates data cleansing - such as dealing with missing and inaccurate data values. Modern applications require such intelligent data fusion to be feasible in near real-time and as automatic as possible. Also, new forms of information sources such as streaming data sources, sensor networks or automatic extraction of information from large document collections (e.g., text, HTML) result in a difficult data analysis problem; supporting this is currently the focus of database research[124]. Data management techniques increasingly make use of intelligent data analysis techniques and also on visualisation to optimise processes and inform the user.

Diverse data from the Internet imposes novel challenges to database research.

Data Mining

The discipline of data mining develops computational methods to automatically extract valuable information from raw data by means of automatic analysis algorithms[75]. There are various approaches; one is supervised learning from examples, where, based on a set of training samples, deterministic or probabilistic algorithms are used to learn models for the classification (or prediction) of previously unseen data samples. Decision trees, support vector machines and neural networks are examples of supervised learning. Another approach is unsupervised learning, such as cluster analysis[54], which aims to extract structure from data without prior knowledge being available. Solutions

Data mining: automatic extraction of valuable information from raw data in this class are employed to automatically group data instances into classes based on mutual similarity, and to identify outliers in noisy data during data preprocessing. Other approaches include association rule mining (analysis of co-occurrence of data items) and dimensionality reduction. While data analysis was initially developed for structured data, recent research aims at analysing semi-structured and complex data types such as Web documents or multimedia data. In almost all data analysis algorithms, a variety of parameters needs to be specified, a problem which is usually not trivial and often needs supervision by a human expert. Interactive visualisation can help with this, and can also be used in presenting the results of the automatic analysis – so called 'visual data mining'.

Spatio-temporal Data Analysis

Finding relations and patterns in spatial and/or temporal data requires special techniques

Spatial data, is data with references in the real world, such as geographic measurements, GPS position data, and data from remote sensing applications; essentially, data that can be represented on a map or chart. Finding spatial relationships and patterns within this data is of special interest, requiring the development of appropriate management, representation and analysis functions (for example, developing efficient data structures or defining distance and similarity functions). Temporal data, on the other hand, is a function of time, that is the value of data variables may change over time; important analysis tasks here include the identification of patterns, trends and correlations of the data items over time. Application-dependent analysis functions and similarity metrics for time-related data have been proposed for a wide range of fields, such as finance and engineering.

Scale and uncertainty impose challenges on spatio-temporal data analysis

The analysis of data with references both in space and in time, spatial-temporal data, has added complexities of scale and uncertainty. For instance, it is often necessary to scale maps to look for patterns over wide and also localised areas, and similarly for time, we may wish to look for trends that occurs during a day and others that occurs on a yearly basis. In terms of uncertainty, spatio-temporal data is often incomplete, interpolated, collected at different times or based upon different assumptions. Other issues related to spatial-temporal data include complicated topological relations between objects in space, typically very large datasets and the need for specialised data types. In addition, more and more geo-spatial data is now accessible to non-expert communities and these 'analysts' need tools to take advantage of this rich source of information.

Perception and Cognition

Design of user interfaces needs to take perception and cognition into account

Perception and cognition represent the more human side of visual analytics. Visual perception is the means by which people interpret their surroundings and for that matter, images on a computer display. Cognition is the ability to understand this visual information, making inferences largely based on prior learning. The whole system is extremely complex, and it has taken decades of research in fields such as psychology, cognitive science and neuro-science to try to understand how the visual system achieves this feat so rapidly. For many years it was thought that ’seeing’ was a generally passive activity with a detailed ’map of the world’, whereas now we recognise that it is very active, only searching for and selecting visual information, which is pertinent to the current task. Knowledge of how we ’think visually’[123] is important in the design of user interfaces and together with the practical experience from the field of human computer interaction, will help in the creation of methods and tools for design of perception-driven, multimodal interaction techniques for visualisation and exploration of large information spaces, as well as usability evaluation of such systems[36, 100].

Visual analytics relies on an efficient infrastructure to bind together many of the functions supplied by the various disciplines, in order to produce a coherent system. In addition, evaluation is critical in assessing both the effectiveness and usability of such systems. We will now consider these enabling technologies.

Infrastructure

As mentioned in the introduction to this section, infrastructure is concerned with linking together all the processes, functions and services required by visual analytic applications so they work in harmony, in order to allow the user to undertake their data exploration tasks in an efficient and effective manner. This is difficult as the software infrastructures created by the different technologies are generally incompatible at a low level and this is further complicated as one of the fundamental requirement of visual analytics applications is high interactivity. For this reason, most visual analytics applications are currently custom-built stand-alone applications, using for example, in-memory data storage rather than database management systems. The design of system and software architectures is paramount in enabling applications to successfully utilise the most appropriate technologies. In addition, the reuse of many common components will result in applications being more adaptable and built much quicker.

Appropriately designed infrastructure is vital to the success of visual analytics

Evaluation

Researchers and developers continue to create new techniques, methods, models and theories, but it is very important to assess the effectiveness, efficiency and user acceptance of these innovations in a standard way, so they can be compared and potential problems can be identified. However, as demonstrated in Chapter 8, evaluation is very difficult given the explorative nature of visual analytics, the wide range of user experience, the diversity of data sources and the actual tasks themselves. In the field of information visualisation, evaluation has only recently become more prominent[13]. It has been recognised that a general understanding of the taxonomies regarding the main data types and

Rigorous assessment of current and innovative solutions across all disciplines is imperative user tasks[4] to be supported are highly desirable for shaping visual analytics research.

The current diversification and dispersion of visual analytics research and development has focused on specific application areas. While this approach may suit the requirements of each of these applications, a more rigorous and scientific perspective based on effective and reproducible evaluation techniques, will lead to a better understanding of the field and more successful and efficient development of innovative methods and techniques.

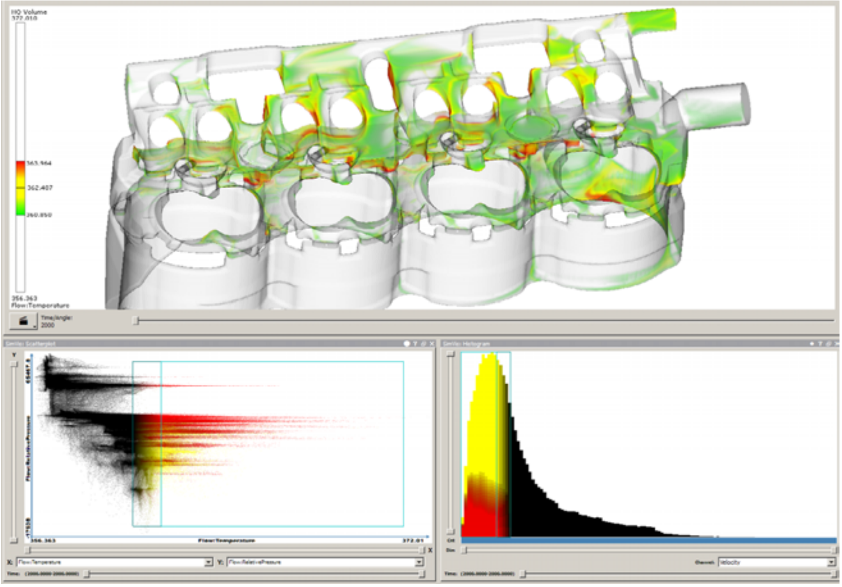

Figure 2.5: Visual analytics in action: Interactive visual analysis of a cooling jacket simulation. User has focused on critical regions of high temperatures and low flow velocities by brushing the two views (velocity histogram and temperature versus relative pressure) as they may indicate locations of insufficient cooling. Dataset is courtesy of AVL List GmbH, Graz, Austria; Interactive Visual Analysis © SimVis GmbH, 2010

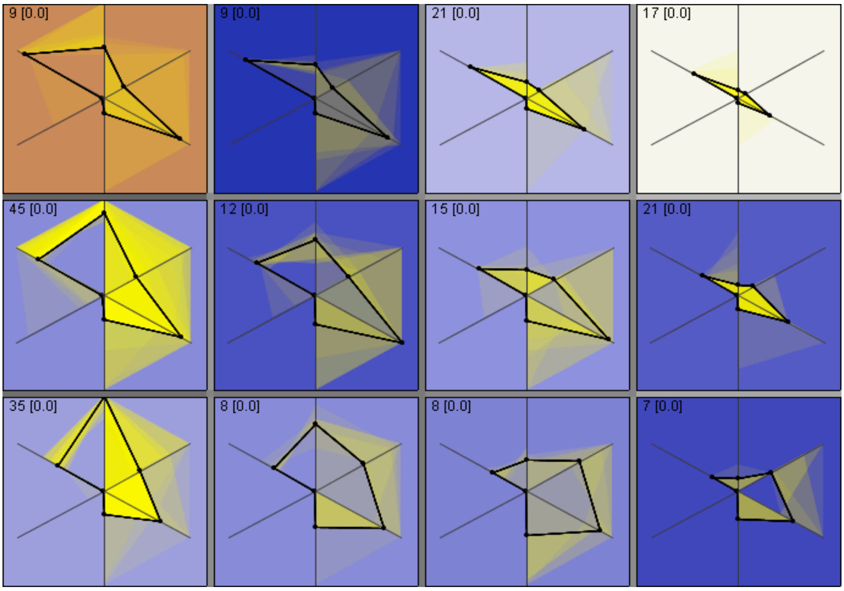

Figure 2.6: Visual analytics in action: Helping demography researchers to effectively analyse multivariate datasets. Six-dimensional demographic dataset was clustered into twelve groups, and the distribution shown by radial parallel coordinate plots. Yellow opacity bands illustrate the variance within the individual clusters and background colour coding correlates cluster with a specific target variable. Technique by Bak et al.[10]