Technical Report TR-2009005, Dept of Computer Science, The Graduate Center, City University of New York, NY, USA, 2009.

TECHNICAL REPORT
Visual Analytics: A Multi-faceted Overview

Ilknur Icke[a]
iicke@gc.cuny.edu
(a) Dept of Computer Science
The Graduate Center
City University of New York
365 Fifth Avenue
New York, NY, 10016 USA

Elizabeth Sklar[a,b]
sklar@sci.brooklyn.cuny.edu
(b) Dept of Computer and Information Science
Brooklyn College
City University of New York
2900 Bedford Ave
Brooklyn, NY 11210 USA

April 16, 2009

Abstract

Visual Analytics (VA) is an emerging field that provides automated analysis of large and complex data sets via interactive visualization systems in an effort to facilitate fruitful decision making. VA is a collaborative process between the human and the machine. In this paper, we present a multi-faceted overview of this human-computer collaboration. The system facet contains everything about the data, analytical tasks, visualization types and the relationships between them. The user facet contains the number and properties of the users. The collaboration facet covers the interactions between the system and the users within the context of VA.

1 Visual Analytics and The Road to Wisdom

One of the overarching goals of science is to understand the world around us. In doing this, we rely on our senses to provide us with the observations (i.e., "data") and our brains to make sense of these observations (i.e., "analysis"). Ackoff [1] provides us with a multi-phase model of this sense making process. According to Ackoff's model, humans process data into information in order to answer questions of *who, what, where* and *when*. Further processing of information leads to knowledge that answers *how* questions, and finally an understanding of *why* helps us shape the future. We illustrate this process in Figure 1.

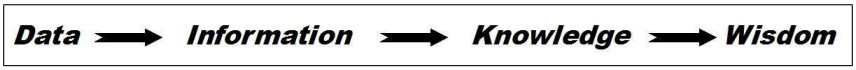

Figure 1: From Data to Wisdom

In today's world, we are becoming increasingly swamped with data. Enormous amounts of data are populating storage devices and waiting to be transformed into some sort of useful information and then knowledge that would hopefully serve a good purpose. Unfortunately, the technologies to transform data into information and knowledge lag behind the technologies that collect and store the data. Governments are storing millions of phone calls in order to catch terrorists plotting an attack, credit card companies are storing purchase and payment histories of millions of customers in order to be able to make predictions about which customers would be worthy of their credit, online retail stores are keeping records of purchases so that they can identify other products to offer to the individuals based on what they previously bought, supermarkets are recording surveillance videos to be used as evidence in case of a robbery, and so on. Many more examples can be given from our everyday lives or from specialized domains such as medicine and education.

Collecting data is only the beginning of a long journey to reach *wisdom*. Wisdom is the ultimate state of having the understanding of the *principles* of a system that is being *observed*. Observing a system (for example, stock

movements, student performance on assessments, behavior of credit card customers) starts with having the idea of what the entities and the relationships between these entities are. This is the data design stage and the most popular method in data design is the *Entity-Relationship (E-R) Model* proposed by Chen in 1976 [13]. Figure 2 shows an example entity-relationship model of customers purchasing books from an online book store.

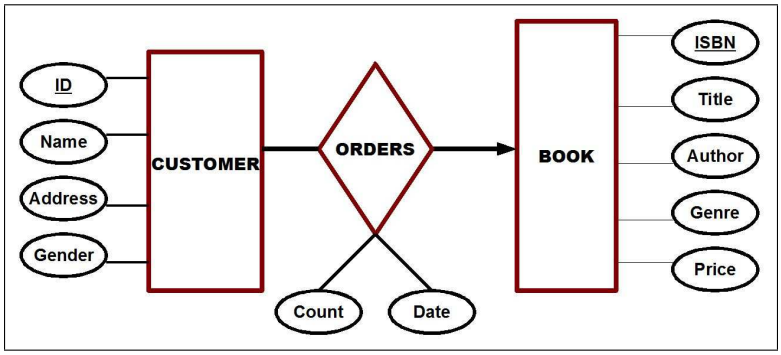

Figure 2: An example E-R diagram

Representing the data using the E-R model is possible in manufactured domains such as the business applications mentioned above. In these domains, the entities and relationships can be predefined. On the other hand, in some domains, it is hard to design a data model beforehand because the analyst does not have full understanding of the components of the system under observation. In this case, entities and relationships should also be extracted from pieces of the collected data. This is generally the case in most scientific domains. For instance, in gene expression analysis, datasets contain thousands of genes and the goal is to discover and explain the various relationships between these genes [33].

When faced with a problem to solve, it is often helpful to first create an abstraction of the problem. In most cases, abstractions contain various forms of visualization (e.g., diagrams, maps or graphs) that help us look at a problem from different angles and devise a solution. *Data Visualization* has been used and studied extensively even before computers came into our lives. Maps were the earliest visualization artifacts. The 1800's are assumed to be the beginning of modern data graphics. As mathematical and statistical methods evolved, new kinds of visualization methods emerged. Detailed information on the milestones in data visualization history can be found in [21]. Edward Tufte presents a wide variety of historical and contemporary visualizations in his well-known books *Visual Explanations* [51], *Envisioning Information* [50], *The Visual Display of Quantitative Information* [49], and *Data Analysis for Politics and Policy* [48]. The term *Exploratory Data Analysis (EDA)* was introduced by Tukey in 1977 [52]. Tukey suggested the use of statistical graphics as an aid for model design in data analysis. The field of *Information Visualization* emerged in the late 1980's and an overwhelming amount of visualization methods have been proposed since then [12, 59].

On the other side of the data analysis continuum, there have been efforts to employ purely mathematical and statistical methods with little or no emphasis on visualization of the data. *Knowledge Discovery in Databases (KDD)* became popular in the 1990's and is defined as "the process of identifying valid, novel, potentially useful, and ultimately understandable structure in data" [10]. At the heart of KDD lies a process called *Data Mining (DM)* which is defined as "a step in the KDD process concerned with the algorithmic means by which patterns or models (structures) are enumerated from the data under acceptable computational efficiency limitations" [10]. Focusing heavily on automatically created mathematical models of data comes with a number of challenges. The most important challenge is to be able to generate more intuitive and understandable models for users.

The *Visual Data Mining (VDM)* concept was introduced in the early 2000's as an interdisciplinary field that aims to facilitate human perceptual abilities in data analysis [36, 56]. The idea is that humans might catch hidden patterns in data that might have been missed by the data-mining algorithms, provided that they are given effective interactive tools to examine the datasets visually. The visualization methods employed in VDM borrow techniques from computer graphics and design theory, and they are much more complex than classical statistical graphing techniques typically used in EDA.

The field of *Visual Analytics (VA)* was initiated by the US Department of Homeland Security after the tragic events of September 11, 2001 [45]. The grand challenges were defined as preventing threats and preparing better for emergency response by analyzing the huge amounts of data that are being collected from the myriad of electronic covert means in use today. The National Visualization and Analytics Center (NVAC) published a *Research and*

Development Agenda [45] to lay the foundations of this new field. They define Visual Analytics as "the science of analytical reasoning facilitated by interactive visual interfaces". Keim *et al*. give a more elaborate definition in [35]: "Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets."

The fundamental effect of this emerging field is that it proposes a human-machine collaboration in making sense out of data. Manual exploration of large datasets is not possible but a totally automatic analysis of data is not desirable either, therefore VA promises a hybrid and more useful strategy. Figure 3 places various data analysis disciplines on a continuum. On the left lie the fields which depend on humans exploring the data via visualizations, and on the far right are the fields which depend more and more on automated analysis of data via mathematical and statistical methods and put little or no emphasis on human intervention and visualization. Visual Analytics aims to cover all aspects of what the previous fields failed to cover; it is meant to be the whole process that provides both means of visualization and analysis, starting from the data and extending to knowledge.

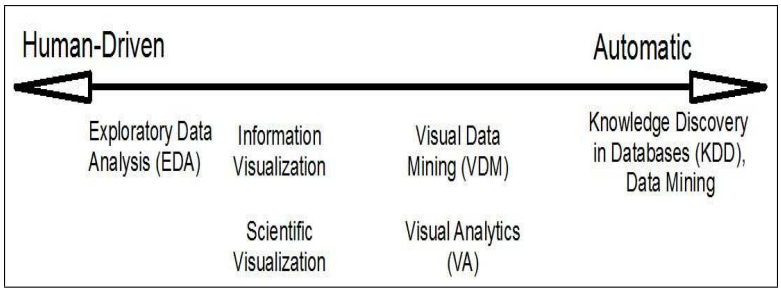

Figure 3: The data analysis continuum

Figure 4 shows the relationships between the visualization fields and how they relate to the stages of understanding on the road map from data to wisdom (figure adapted from [7]). Visual Analytics clearly is a highly interdisciplinary field, covering a wide range of fields from data management, data mining, perception and cognition, human-computer interaction and visualization. Even artists are contributing their talents [57]. The goal is to bring humans and computers together for strong collaboration in order to increase the level of understanding of the phenomena that are under observation. Interactivity is the central concept in VA and through interactivity humans are allowed to communicate with the computer to provide feedback. Wijk [55] and Keim [35] *et al*. give a conceptual view of VA, which is called *the sense-making loop*. Figure 5 illustrates our interpretation of this participatory process. Interactivity is the key ingredient: humans *interact* with visualizations which are created *automatically* from data, providing feedback so that the system can automatically generate better visualizations. Therefore, VA can be defined as a human-machine collaboration process in data-backed decision making.

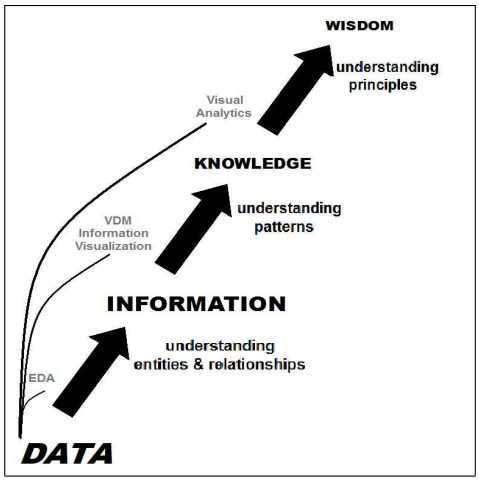

Figure 4: Visualization fields and stages of understanding (adapted from [7].

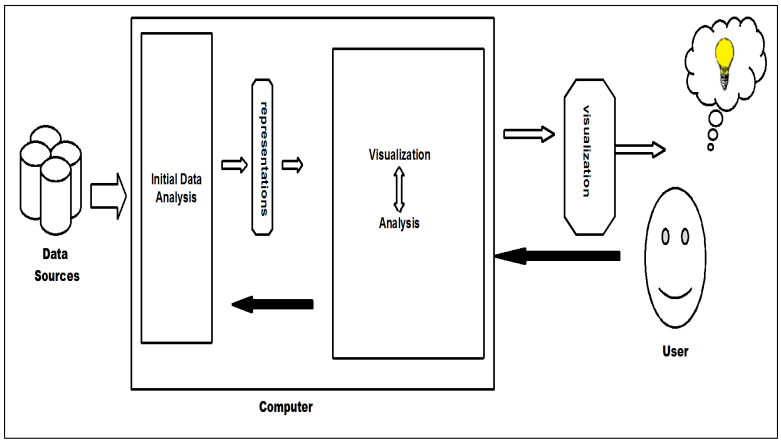

Figure 5: Sense-making loop using visual analytics

2 Aspects of Visual Analytics

A number of researchers have developed taxonomies of visualization methods. Lengler and Eppler present a *periodic table of visualization methods* [37]. In their taxonomy, they review about a hundred visualization methods and classify them based on these five aspects: *complexity of the visualization*, *main application area*, *level of detail*, *type of thinking aid* and *type of representation*. Another multi-faceted overview of visualization techniques is given by Keim [36]. In this classification, visualization techniques have been examined with respect to three aspects: *data to be visualized* (1D, 2D, higher dimensional, text/web, hierarchy or graphs, algorithms or software), *visualization technique* (standard 2D/3D display, geometrically transformed display, iconic, dense-pixel and stacked displays) and *interaction techniques* (standard, projection, filtering, zoom, distortion, link and brush).

Another way of examining visualization methods is the operator framework given by Chi *et al*. [16]. In this framework, an operator might mean any kind of system-user interaction. A value operator changes the dataset such as selecting a portion of data. A view operator changes the visualization such as zooming, rotating and scaling. In this framework, datasets turn into visualizations through a visualization pipeline. The stages in this pipeline are raw data, analytical abstraction, visualization abstraction and view. Datasets are converted into analytical abstractions via data transformations, which in turn become visualization abstractions through visualization transformations. Final visualizations are created using visual mapping transformations. A detailed analysis of 36 visualization methods with respect to this framework has been given in [15].

As an emerging field, VA has different aspects and needs than just visualizing data. In the next three sections, we dissect these aspects and develop a multi-dimensional view of VA, based on these aspects. Our overview is strongly related to the sense-making loop (Figure 5) paradigm of VA. Visual Analytics is a human-machine collaboration in decision making, therefore it has three main dimensions. The first dimension of VA is the *system*, namely the computing environment where the datasets are stored and analytical algorithms are implemented. The second dimension is the *user* who needs to make decisions based on data. The third dimension is the human-machine collaboration aspect which acts as a bridge in between the user and the system (Figure 6). In Section 3, the system aspects are examined. Section 4 presents the user aspects, and Section 5 gives a discussion on human-machine collaboration aspects.

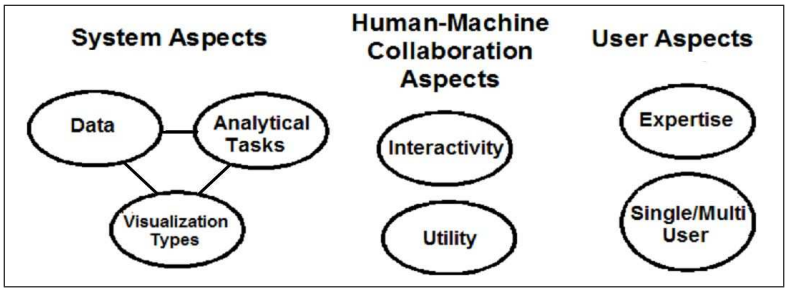

Figure 6: Aspects of VA

3 System Aspects

We categorize non-human components of the VA process as data, analytical tasks and visualizations. In this section, we will review the research aspects that are related to these components. Since these aspects relate to the computational environment, they are called the system aspects. As can be seen on the lefthand side of Figure 6, we have identified six main system aspects: properties of datasets, nature of analytical tasks, visualization types, and the relationships between each pair.

3.1 Properties of Datasets

Not all data are created equal. They come from different domains in different forms but most analytical techniques are based on representations of data as a set of multidimensional vectors. Therefore, depending on the domain, the original dataset may need to be transformed into multidimensional vector form. Here, before discussing the properties of datasets, we will give a brief overview of various domains from which datasets may come and ways of transforming data into the standard vector form.

Multimedia Domain. Humans invented different media for expressing themselves or communicating. Printed media (pictures and text) arrived first, then came multimedia where sound (speech and music) and moving pictures (video and animation) invaded our lives. Due to the advent of computer-aided design (CAD), computer games and medical imaging, virtual 3D models became another form of medium. Data in these domains are generally kept in separate files that have specific formats based on their types (such as "mp3" for audio, "mpg" for video).

Information Technology. Computer and network technologies helped merchants and governments move bookkeeping and logging operations into the digital world. As a result, databases of overwhelming amounts of data have become commonplace. Due to the design of database systems (for example, relational database systems), the data items are kept in tables which can be seen as spreadsheets consisting of rows and columns. Each row represents one individual entity (i.e., an employee, a customer, a car and so on) and each column represents an aspect that is related to the individual (such as age, gender, purchase amount, milage and so on). Generally each row (individual) is designated by a unique identifier and each column is given a specific name to specify the aspect it is describing. Sometimes, we might need to associate two individual entities–for example to answer such questions as 'who purchased what item?', 'who has called whom on the phone', 'who has whom on her social network?'. In this case, another table (a relation) is created with the unique identifiers of the individuals. When defining tables, the data designers have to assign certain data types for each column (text field, number field, date field), thus each individual row becomes just a vector of text, number and date fields. In today's commercial world, data analysis is done on terabytes of this kind of data structure and the specific discipline that is concerned with the visualization data is called information visualization.

Scientific Domain. Different scientific disciplines have different methods for conducting research. Most of them utilize sensors to capture data (for example; measuring temperature, pressure, mass, brain signals, heart rhythm, blood sugar level, recording sound, image and video). If data items are of multimedia type (sound, text, image, video, 3D model) they are stored as separate files whereas other sensor data such as temperature, pressure, mass, brain signals can be kept in database tables because they can be stored as text, numeric or date fields. Visualization of scientific data has another aspect which generally does not occur in other domains; in scientific domains, scientists frequently use animated visualizations to simulate phenomena evolving, based on time-series observations (i.e., weather forecasting, medical surgery simulations). Scientific visualization covers the methods used in such domains.

Transforming Data for Analysis. Standard data analysis tools are based on the vector model (i.e., a row of a database table) where each vector designates a data item (individual, observation) and contains different aspects (features, variables) about that item. Input data items that come from complex domains (input space), such as multimedia domains, and are transformed into a form (feature space) that can be used by the analysis tools. The aim here is to create a vector based representation for the data at hand. This process is known as *feature generation*, and it is desirable that the resulting feature vectors fairly represent the respective data items. For the most part, feature generation methods are domain specific, ad hoc and subjective. For text data, the main features are words and/or phrases. These features can tell a lot about the theme of a piece of text. The most basic feature extraction technique on text data is to compute the frequencies of individual words and/or phrases in the text. Especially after the world wide web became the largest uncharted textual territory, the problem of text data mining became a very popular topic and extensive literature exists on this problem. Also, various web-based search engines for geometric 3D models have been developed. A survey on feature generation methods from 3D geometric (CAD) models is given in [27], among others. There is also extensive literature on data transformation methods on image, sound and video domains. An example feature generation technique using mathematical morphology on range images is given in [28].

3.1.1 Variable Types

As mentioned before, due to the design of data storage mechanisms (relational databases), the data items (observations) are kept as rows (vectors) of data fields (variables). Therefore, the data types are primarily text or numeric. As far as the semantics are concerned, variables can be classified as [23]:

- Categorical (Binary, Nominal, Ordinal)
- Numerical (Interval-scaled, Ratio-scaled)

A dataset might contain data items consisting of different types of variables posing challenges in analysis.

3.1.2 Dimensionality

Dimensionality of data poses problems for visualization and for analysis. Human perception and computer screens are limited to 3 spatial dimensions therefore it would be challenging to visualize higher dimensional datasets. For analysis, high dimensionality poses theoretical and practical problems. Theoretically, the higher the dimensions get, the more data items one needs in order to perform meaningful analysis. Practically, more dimensions mean more processing time. The remedy most researchers perform is to reduce the number of dimensions. Figure 7 gives the flow of conversions starting from the input space to the data space to be used in further analysis.

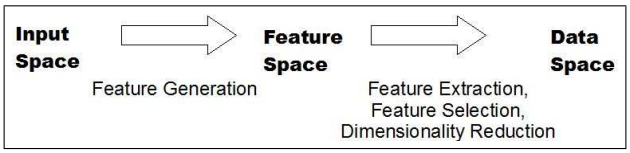

Figure 7: Input space to data space conversion

Dimensionality reduction has been treated as a preprocessing step before analyzing the data and a large number of algorithms have been proposed. Detailed review of these algorithms can be found in [20], [26] and [54]. The most commonly used method in dimensionality reduction is the *Principle Components Analysis (PCA)* [34] technique (Figure 8).

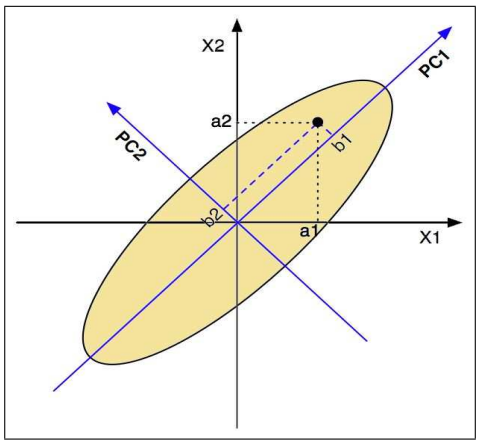

Figure 8: Principal components analysis (PCA) of 2D data

In PCA, a number of orthogonal vectors (principle components) that capture the most variations in the data are computed and the data points are projected on those vectors. The first principle component captures the most variance, and the second principle component that is orthogonal to the first captures the second most variance and so on. By projecting the original high-dimensional dataset onto a small number of top principle components, we get a lower dimensional model of the dataset (Figure 9).

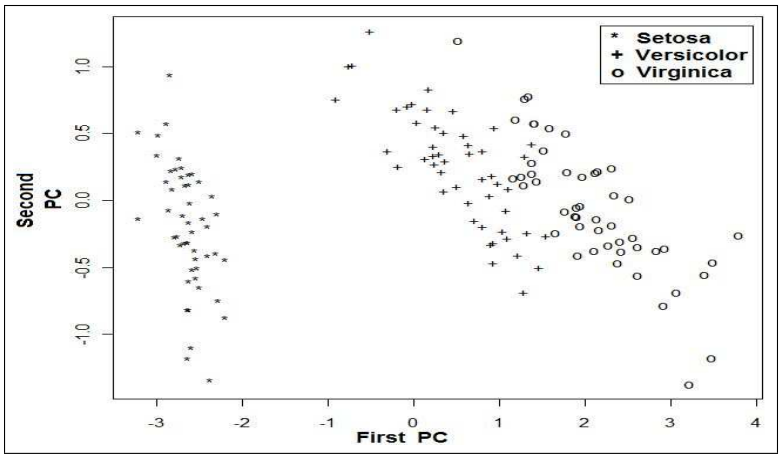

Figure 9: UCI Iris dataset [53] (four dimensional) projected onto its first two principal components (PC)

3.1.3 Time and Space Relations

Time and space have special meanings for humans and various analytical tasks aim to understand phenomena with respect to time, space or both.

- **Temporal (ordered, sequential).** Temporality can be absolute or relative. In the case of financial time series, each observation (e.g., stock price) corresponds to a specific point in time. If the exact time is not important but the order of the observations is, then the dataset is called *sequential*. The sequences of web pages visited by users are common sequential datasets used in web data mining.
- **Spatial.** If observations correspond to some underlying spatial component, this mapping can aid in data visualization. The relatively new field of *Geographical Information Systems (GIS)* focuses on analyzing and visualizing data that corresponds to geographic locations (such as a group of countries and the population of each). In the GIS case, the spatial component is a physical location, whereas in other domains, abstract spatial components can be introduced in order to facilitate visualization. For example, web-based data can be visualized by creating a network that represents a map of a web site. Another example is the use of *topological maps* in the field of *Mobile Robotics*. Topological maps are built by recognizing features that are "next to" each other. Because the robot encounters them one after the other, the map is not necessarily geographically accurate in terms of the physical distance between things. The information on the map is more useful because features that are important to the robot are emphasized.
- **Spatio-Temporal.** In this case, the observations are tied to a temporal and a spatial component. For example, Electroencephalography (EEG) provides spatio-temporal data by recording the electrical activities of various locations of the brain over a period of time.

3.1.4 Relationships

Often times the goal of data analysis is to uncover certain relationships in the data. There are two kinds of relationships one can investigate.

- **Between observations.** Observations can be displayed as a *graph*, where each observation is a vertex and edges that connect the vertices indicate relationships between observations. A single observation incorporates the values of one or more variables. For example, in the case of social networking, the existence of an edge between two people indicates that they are friends. Other kinds of semantic relationships can also be represented in this scheme (such as co-authorship of scientific publications). Edges can also have values, or *weights*, associated with them. For example, a common numeric value for an edge is to represent the *distance* between the two vertices, or observations. Observations with small distances between them are considered *similar*, whereas large distances indicate higher dissimilarity.
- **Between variables.** The study of relationships between variables is an important topic in statistics and other analytical fields. Two main types of relationships in statistics are correlational relationships and causal relationships. Two variables are correlated if they behave the same way. For example, height and weight are correlated.

Tall people tend to weigh more. But a correlational relationship does not necessarily mean a causal relationship: being tall is not necessarily the cause of weighing more. Variables that have no relationship are called independent.

3.1.5 Source and Quality

The following issues pose challenges in dealing with data:

- **Multiple Data Sources.** Datasets could be gathered at different sites, in different modalities (image, sound). Data fusion is an active research field dealing with the issue of handling multiple data sources.
- **Uncertainty.** In some domains, observations come from physical measurements. Most of the time, measurements might contain errors (noise) due to experiment conditions. In general, error is modeled as an additive Gaussian noise.
- **Missing Values.** Some datasets might have missing values for some variables of an observation. The easiest way to deal with this issue is to discard such observations. In other schemes, the missing values are estimated using various statistical techniques [2].

3.1.6 Number of Observations

There are two distinct problems as far as the number of data items (observations, individuals) are concerned:

- **Not Enough.** If the dimensionality of the data is high and the number of data items is not large, we face the issue known as the *"curse of dimensionality"*. This is a theoretical problem and it basically means that in order to be able to analyze a dataset meaningfully, more samples (observations) are needed as the data dimensionality (number of variables) increases [18].
- **Too Much.** This is a practical issue rather than a theoretical one. In general, the more data the better for analysis purposes but more data requires more processing time and in some cases, data analysis is a time critical task.

3.2 Nature of Analytics

Decision making, sense making or analytical reasoning based on observed data are very general concepts. One can not give a cookbook of recipes for these processes. But, it is possible to talk about a number of principle analytical tasks every data-backed reasoning process might contain. Figure 10 shows the most common analytical tasks that users aim to perform in order to *learn* from data.

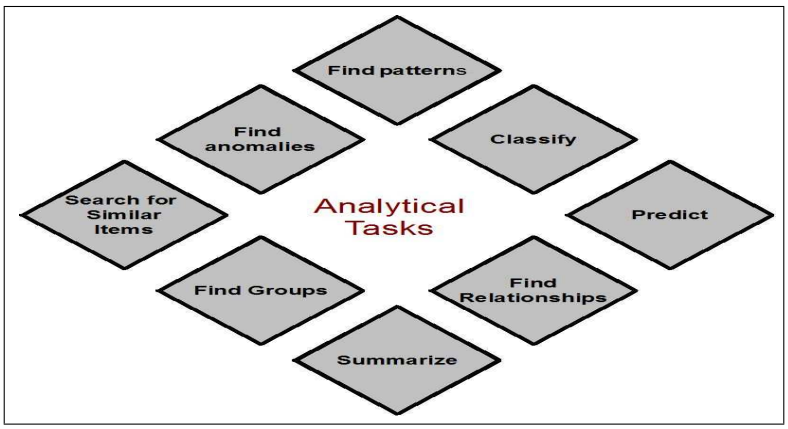

Figure 10: Analytical Tasks

In an exploratory data analysis setting, users perform these analytical tasks by inspecting and interacting with visualizations of the data. The ultimate goal is to analyze today's large and high dimensional datasets as automatically as possible, since it is not practical for humans to visually inspect such complex datasets. For each of these analytical tasks, there exists a huge collection of computational techniques from machine learning, data mining, pattern recognition and statistical learning theory. Overviews of various analytical tasks and related algorithms can be found in [3], [18] and [9].

3.3 Visualization Types

This section covers only the most common classes of data visualization techniques that are used for the purpose of exploring the original dataset. The images in this section were generated using Matlab [40], R [42] and other custom software on several publicly available data sets. The references to the individual datasets are given in the text.

3.3.1 Describing Data

Given a set of data, the first step an analyst would take is to see what story the dataset tells. This could be done by computing standard statistics on the data (i.e., mean, median) or tests to see if data follows any known distributions (for example, test for normality). Following the Exploratory Data Analysis (EDA) tradition, it is also a common place practise to utilize visualizations. In this section, we will briefly cover some of these methods.

Distributions. Bar charts, pie charts and histograms are the predominant visualizations for displaying the distribution of the values for a single variable. Examples are shown in Figure 11.

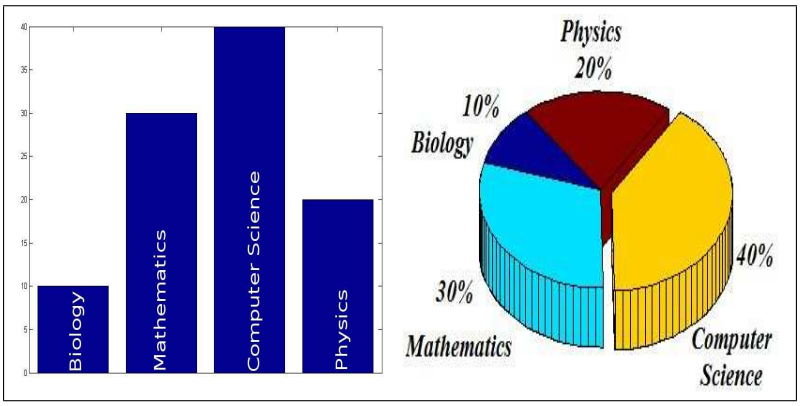

Figure 11: Bar chart and pie chart of distribution of majors

Descriptive Statistics. Instead of displaying the values of the variables, it is also possible to display various statistics of a dataset in visual form. A boxplot (figure 12) is an example of this kind of visualization.

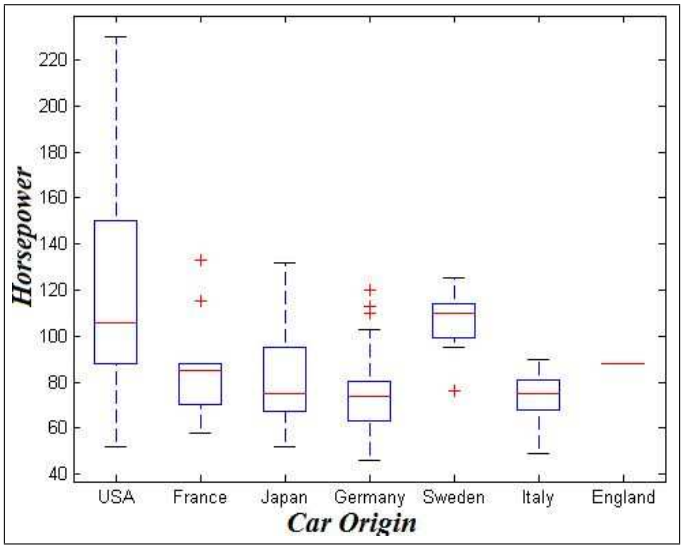

Figure 12: Boxplot visualization of UCI Car dataset [53]

On a boxplot, a five number summary (the smallest observation, lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation) of a dataset is displayed. It is also possible to view descriptive statistics for multiple datasets (populations) on the same boxplot as a means for comparisons.

3.3.2 Viewing Relationships

Between Observations. If the observations (data items) and relationships have been explicitly designed during the data modeling phase, visualization of relationships between the entities can provide a valuable tool for seeing the "big picture". A *network diagram*, such as the one shown in Figure 13, is an example of this kind of visualization.

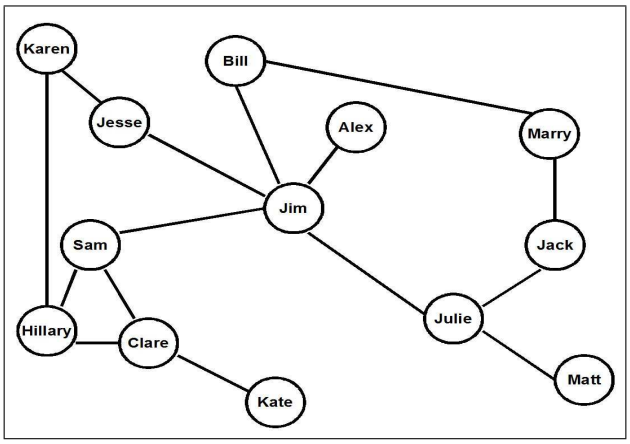

Figure 13: Social networking example

Between Variables. If data has multiple dimensions, it is more desirable to visualize them together so that any relationships between the variables would be revealed upon visual inspection. Scatterplots are used for this purpose. Since the cartesian coordinate system (orthogonal axes) is used in visualization, it is possible to display only two or three variables on a single scatterplot. For this reason, multiple scatterplots are used in order to display all of the variables (see Figure 14).

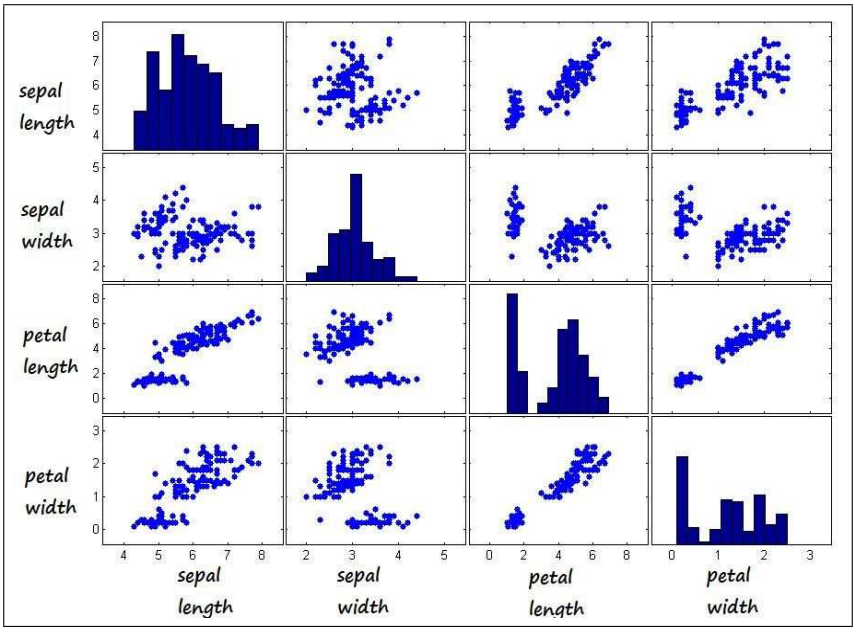

Figure 14: Scatterplots of UCI Iris dataset [53]

Inselberg designed a new technique known as *parallel coordinates* in order to overcome this limitation. An example is shown in Figure 15. In this technique, coordinate axes are drawn in parallel instead of orthogonally [32, 31]. The ordering of the dimensions might affect the usefulness of the visualization. Ankerst *et al*. [5] propose a technique that clusters the data dimensions based on their similarities to enhance visualizations.

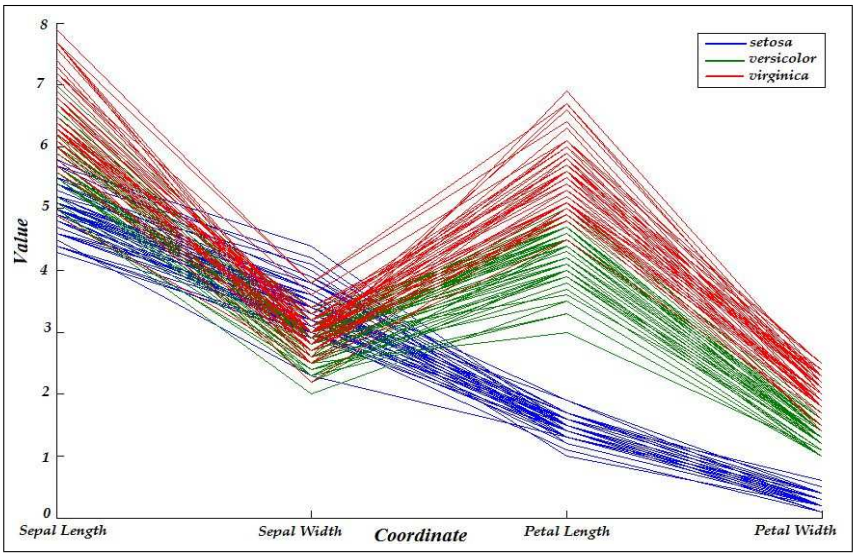

Figure 15: Parallel Coordinates visualization of UCI Iris dataset [53]

Graphical models are graphs on which each node represents a variable and (undirected) edges represent conditional independence relationships between them, directed graphical models can also indicate 'causality' relationships between the variables. Figure 16 shows a graphical model for the joint probability expression:

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_2,x_3)p(x_6|x_4)p(x_7|x_4,x_5).$$

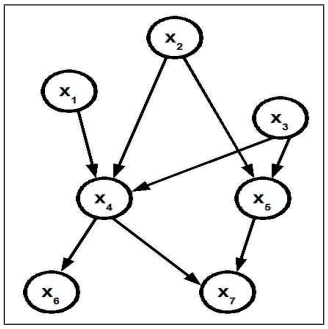

Figure 16: Graphical model visualization of a distribution

3.3.3 Picturing Data: Icons, Glyphs and Color Coding

Humans have great *preattentive processing* skills [47] for pattern recognition that have been emphasized in data visualization literature over and over again. A large number of visualization methods have been introduced aiming to utilize these skills to help make sense out of data. In these methods, the data items are mapped into easily recognizable shapes and sometimes with textures and/or colors to enhance the perceptual utility of the visualization. Mapping the most important data features onto the most salient shape features is the crucial aspect here and it is a challenging design issue. Placement of these pictoral visualizations on the screen is also an important factor on the effectiveness of the methods. Ward gives a detailed overview of placement techniques in [58]. In this section, we briefly cover various methods of picturing data.

Chernoff Faces. Introduced by Herman Chernoff in 1973, Chernoff faces [14] is by far the most famous data picturing method. It is possible to project (map) up to 18 data features onto various face features (such as size and curvature of the face, position of mouth, eyes, nose, size of features). Faces are special visual items because humans are naturally wired to recognize the faces, but the underlying mechanisms are still not well understood.

Therefore, it remains a challenge to assign the data features onto the *appropriate* face features in order to maximize the effectiveness of this visualization method.

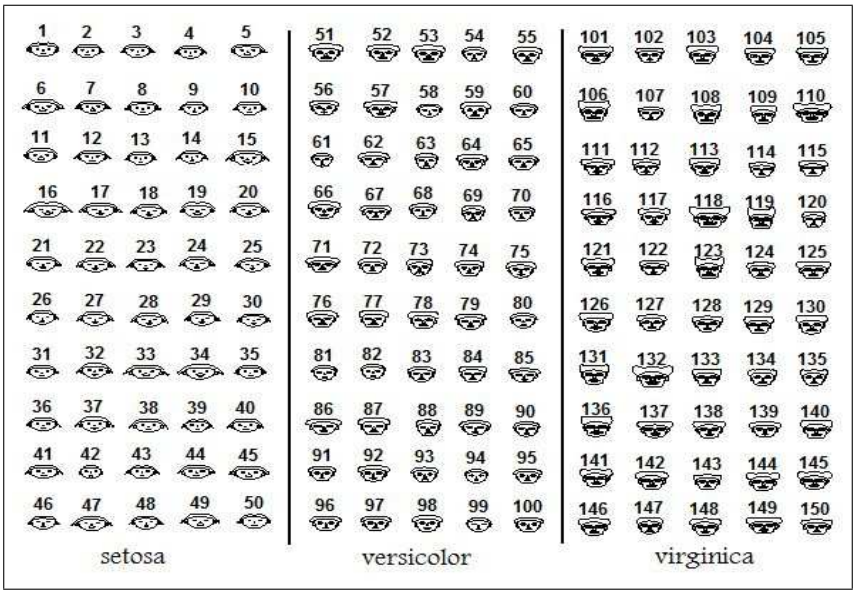

Figure 17: Chernoff Faces visualization of UCI Iris dataset [53]

Mathematical Shapes. Andrew's Plots method [4] projects each data item $X = (x_1, x_2, ..., x_N)$ from vector space into trigonometric function space. The variables (x_i) of each observation become the coefficients of the following Fourier series:

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 sin(t) + x_3 cos(t) + x_4 sin(2t) + x_5 cos(2t) + x_6 sin(3t) + x_7 cos(3t) + ...$$

where $-\pi \leq t \leq \pi$. As it can be seen from the equation, the ordering of the variables affect the shape of the curve.

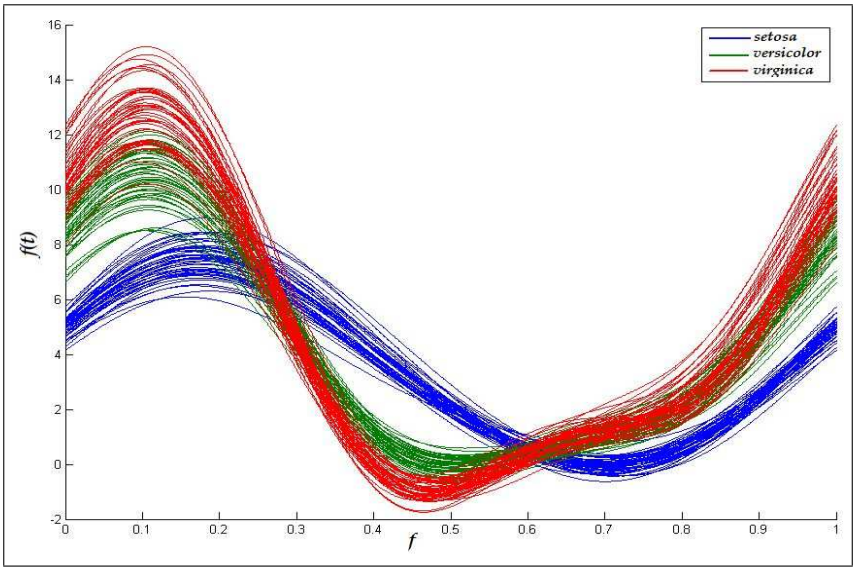

Figure 18: Andrew's plot visualization of UCI Iris dataset [53]

Another similar technique is called star glyphs which project variables onto polar coordinates in 2D and spherical coordinates in 3D. Figure 19 shows a few of the data items from the UCI Car dataset [53] visualized as star glyphs in 2D. This type of visualization can help humans easily recognize models that have similar sets of features.

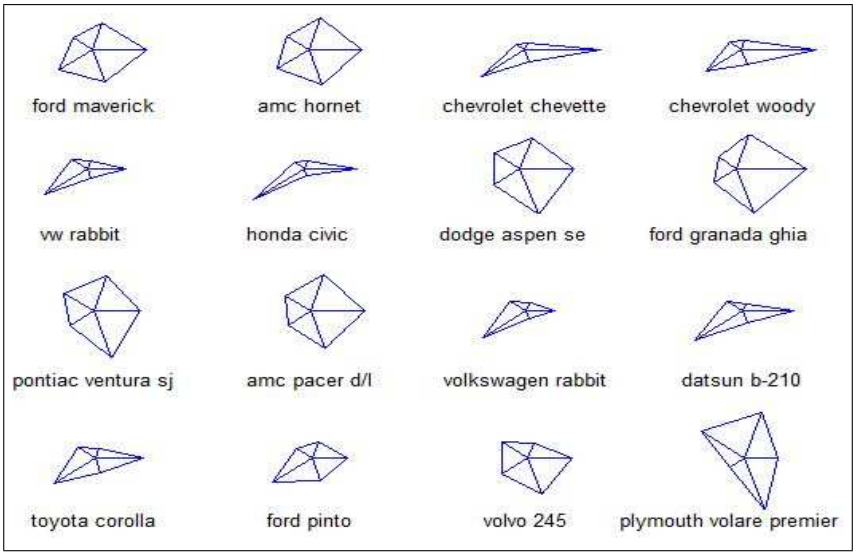

Figure 19: Star glyphs visualization of UCI Car dataset [53]

More advanced mathematical shapes have also been proposed. The parametric shape glyphs method (see Figure 20) projects variables onto the parameter space of superquadrics resulting in various 3D shapes [43].

$$S(\eta,\omega) = \begin{bmatrix} x(\eta,\omega) \\ y(\eta,\omega) \\ z(\eta,\omega) \end{bmatrix} = \begin{bmatrix} a_1 cos^{\varepsilon_1}(\eta) cos^{\varepsilon_2}(\omega) \\ a_2 cos^{\varepsilon_1}(\eta) sin^{\varepsilon_2}(\omega) \\ a_3 sin^{\varepsilon_1}(\eta) \end{bmatrix},$$

$$-\frac{\pi}{2} \leq \eta \leq \frac{\pi}{2}, -\pi \leq \omega \leq \pi$$

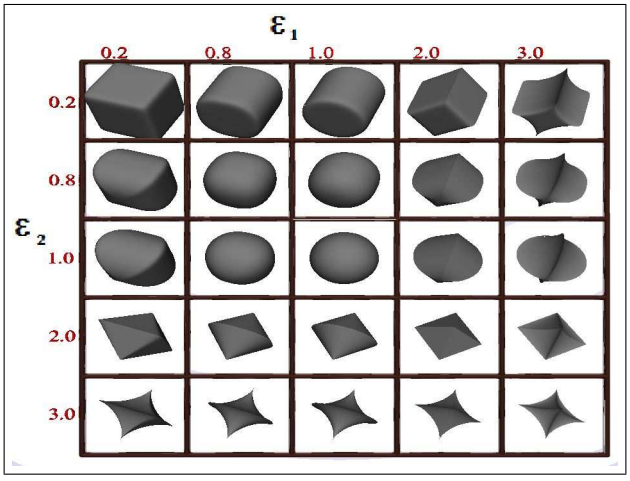

Figure 20: Superquadric shapes by varying two variables $(\varepsilon_1, \varepsilon_2)$

Daisy Maps. Icke and Sklar present a glyph based multivariate data visualization method named *daisy maps* to be used in visualizing categorical data [30]. Figure 21 shows a daisy map of color-coded scores (red:1, orange:2, blue:3, green:4, gray:no score) for one student from an educational test. Each petal of the daisy represents the score for one test topic.

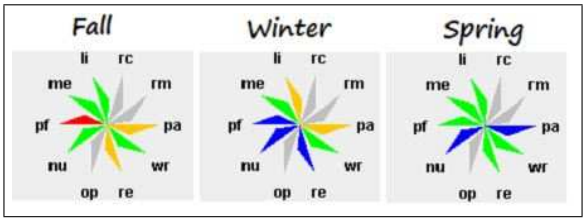

Figure 21: Example student test scores visualized as daisy maps

Heat Maps. A *heat map* is a 2D visualization of a dataset where each variable is a color coded glyph. Figure 22 shows an example gene expression dataset [38]. Each gene is represented as a color coded rectangle. Heat maps give an overall picture of the dataset where similar items can easily be pointed out.

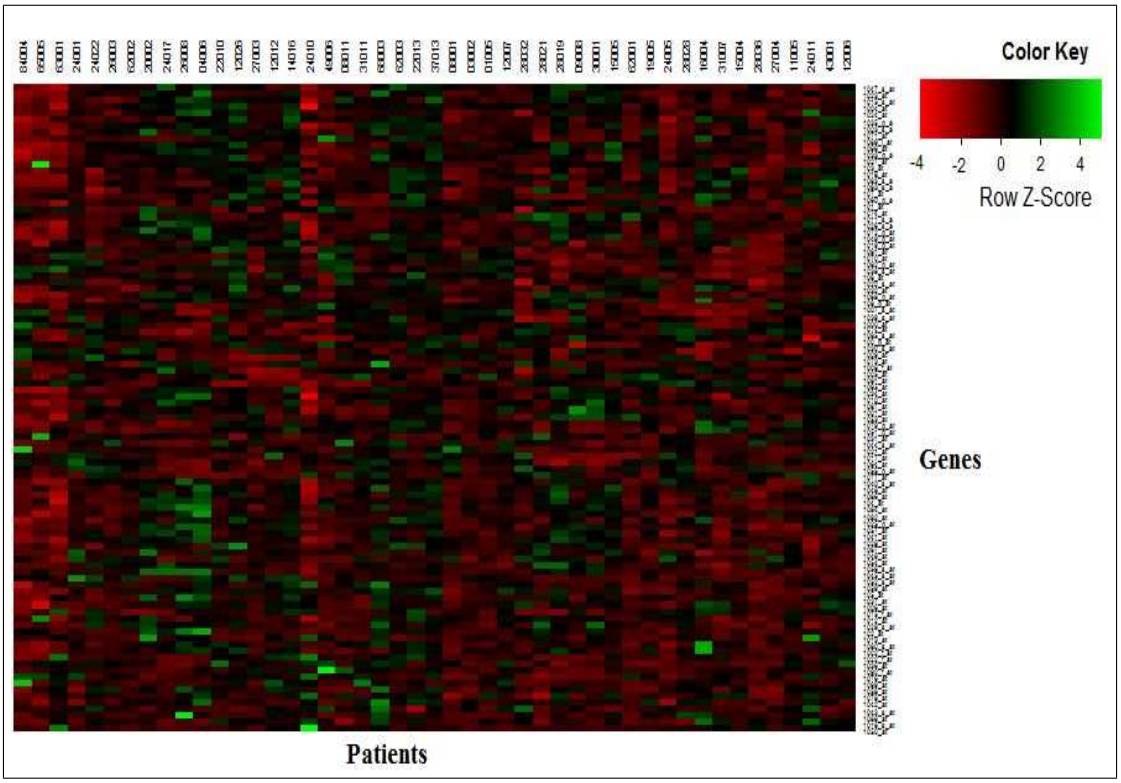

Figure 22: Gene expression dataset heat map visualization

Tag Clouds. Tag clouds are used to display the distribution of words and phrases in a given text. The icons are the graphical renderings of the words themselves and the font size of each word or phrase is proportional to the number of occurrences in the text. This visualization gives a quick summary of the topic of a given text by looking at the most common words in it. Figure 23 shows a tag cloud visualization[1] of the Universal Declaration of Human Rights [41].

[1]Generated on `http://www.wordle.net`

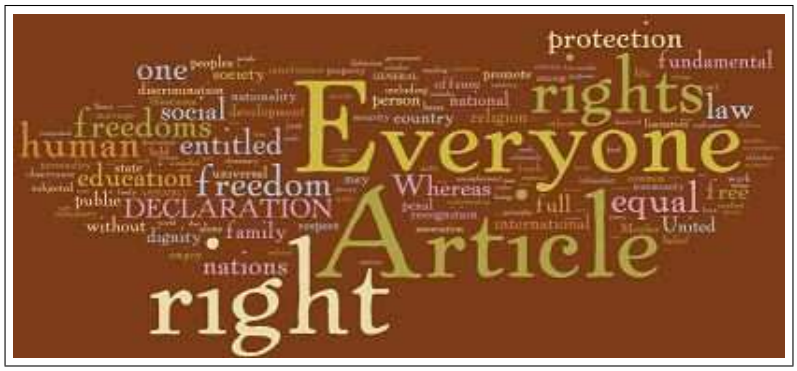

Figure 23: Universal Declaration of Human Rights as a tag cloud

3.3.4 Temporal Visualization

Visualization of temporal data poses a special problem. All variables are dependent on time and users want to visualize how each variable changes over time in order to detect patterns or anomalies in the data. Line graphs are the standard visualizations that picture the changes of one or more variables over time. 3D versions of line graphs have also been proposed [46].

Glyph representation is also used to visualize the values of the variables at one time point. Then, changes in data over time can be visualized by viewing the sequence of glyphs (animation) or by stacking the line graphs (3D time wheel) or glyphs (e.g., 3D Kiviat tube) and forming a 3D visualization of the whole dataset [46]. Figure 24 shows a multivariate (five dimensional: opening, closing, highest, lowest prices and volume per day) stock dataset [44] visualized as a series of line graphs and a snapshot of the animated star glyph.

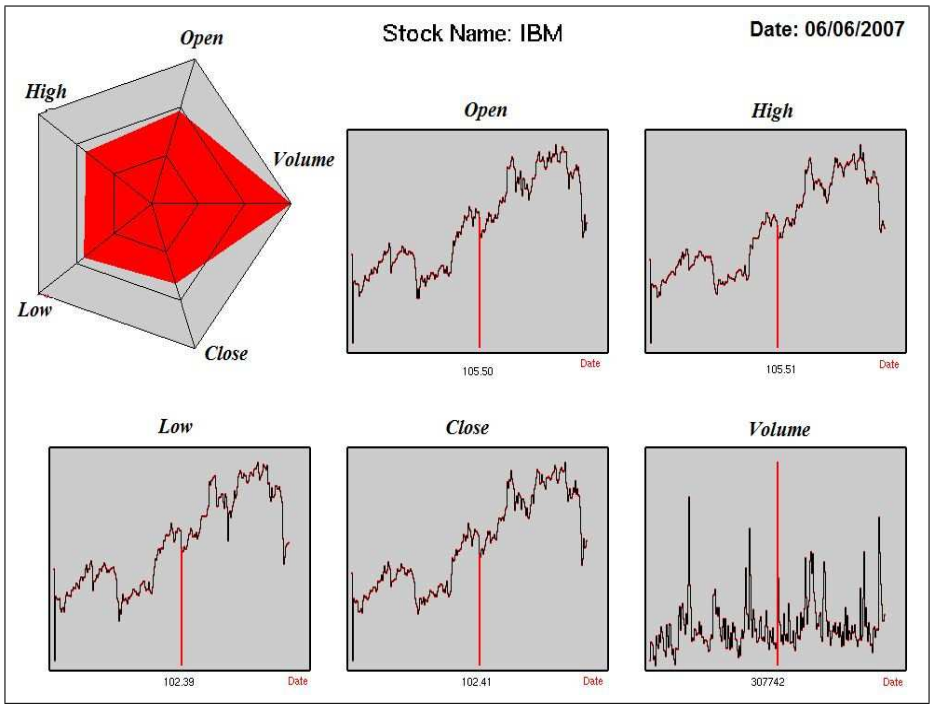

Figure 24: Line graph and animation views of multivariate stock data

3.3.5 Spatial Visualization

Spatial datasets come from various domains that relate data to a certain landscape. Use of a map (layout of the landscape) is the most natural way to visualize this kind of data.

Natural Layout: Geo-Spatial Map. In some domains the landscape corresponds to a physical locale. For example, *Geographical Information Systems (GIS)* focuses on analyzing and visualizing data that corresponds to geographic locations (such as a group of countries and the population of each). Figure 25 shows various properties of a geographical area on a map [39].

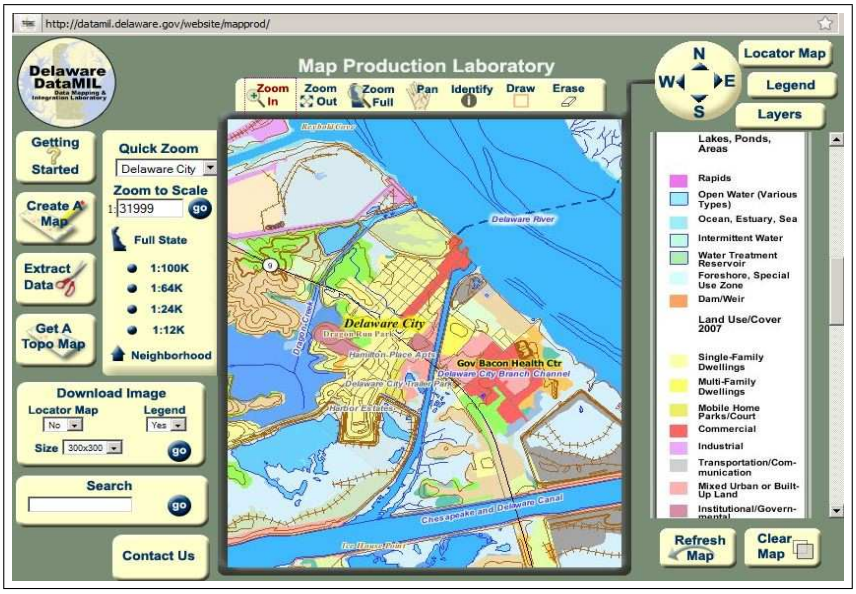

Figure 25: Geo-spatial visualization (Delaware land cover map), from [39]

Abstract Layout: Artificial Landscape Map. Some data analysis problems can be better studied by introducing an artificial problem landscape. An example is the use of *topological maps* in the field of *Mobile Robotics*. Topological maps are built by recognizing features that are "next to" each other, because the robot encounters them one after the other; the map is not necessarily geographically accurate in terms of the physical distance between things, but the information on the map is more useful because features that are important to the robot are emphasized. Also, web-based data can be visualized by creating a network that represents a map of a web site. Figure 26 shows the behavior of humans within an adaptive online environment [29]. The full network (a) shows the *landscape*, all the possible connections from one page to another in the online environment, whereas (b) illustrates the actual path taken by one human.

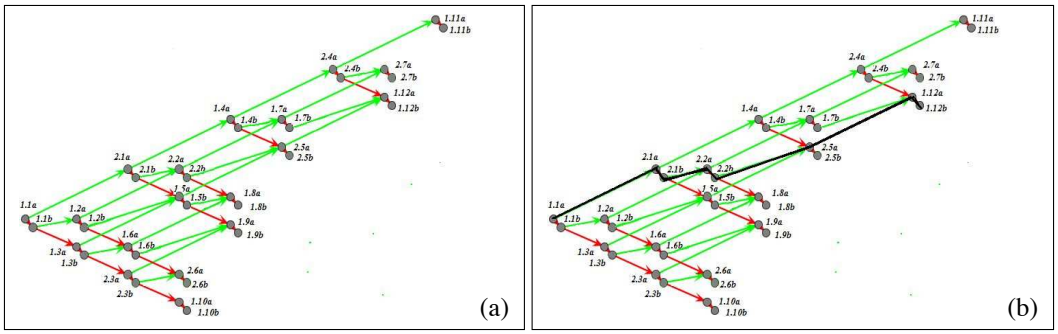

Figure 26: (a) test map, (b) an example performance path, from [29].

3.3.6 Spatio-Temporal Visualization

Spatio-temporal datasets contain both spatial and temporal aspects. The biomedical data analysis field provides an interesting example. Figure 27 shows the positions of the electrodes that are placed on human scalp in order to record EEG signals from the brain. Each electrode is marked with a specific name.

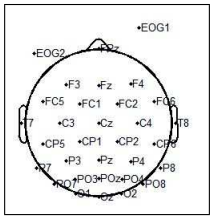

Figure 27: EEG electrode locations on human scalp

Figure 28 shows the signals recorded by each electrode for a period of time [17]. In a typical experiment, the human subject is asked to perform a simple task while the EEG readings are recorded and later analyzed in order to figure out which parts of the brain show specific activities while performing the task.

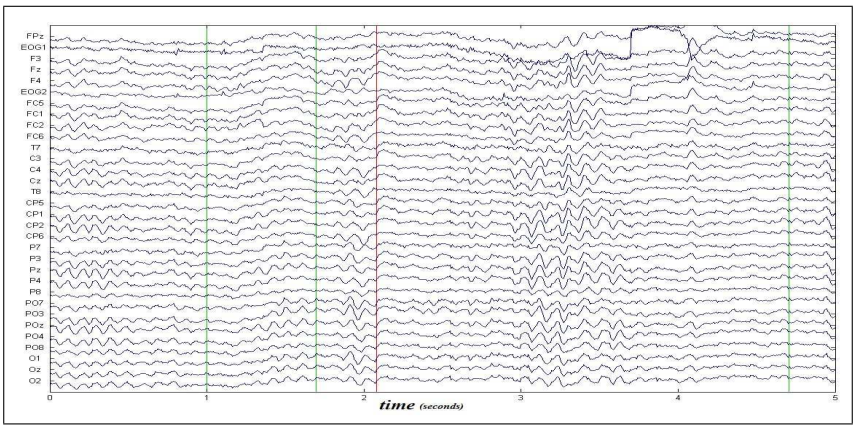

Figure 28: EEG signals

3.4 Relationship Between Data and Analytical Tasks

The relationship between the data and the analyical task is double-sided. The first side of this relationship is the choice of an appropriate algorithm for the dataset at hand. The analytical tasks outlined in section 3.2 are the high-level definitions of basic analytical paradigms. As the decision-maker, the user chooses an analytical task to perform on a given dataset. For each task, a large number of algorithms have been explored by the statistics, data mining, machine learning and related communities. The outcome of a certain analytical task on the same dataset might differ from algorithm to algorithm. Therefore, selecting the most suitable algorithm for the dataset and the analytical task at hand is an important issue. The second side of the relationship is the choice of proper data for a selected algorithm. Not everything in the dataset might be useful for the algorithm to utilize and the algorithm has to be intelligent about what bits and pieces in the dataset would increase the success of the outcome of the analysis.

3.4.1 Algorithm Selection for Dataset

Some classification (supervised learning) algorithms have been reported to perform better on some datasets but not on some other datasets. A number of researchers presented various approaches to *characterize* the datasets to see how they relate to the classification accuracy. Three methods of defining the complexity of a classification problem are proposed in [25]. These complexity measures include measures of overlap of values for each feature, measures of class separability and measures of geometric, topological or density characteristics of the dataset. Combining classifiers and various hybrid techniques have also been widely proposed in order to address the data-dependent

classifier selection problem. The field of meta-learning deals with the problem of finding the most suitable algorithm for a given dataset and an analytical task by matching meta-features of datasets to algorithms and picking the best algorithm that is known to have the best performance on datasets with similar meta-features [11].

A similar problem arises for the clustering (unsupervised learning) task as well. The goal of clustering is to find groups of *similar* items in the dataset so that the items in each group would be more similar to each other than any other item that is in a different group. The term *similarity* is a vague concept and choice of different similarity measures might affect the outcome of the clustering process. A number of algorithms have been proposed to learn a similarity (distance) metric from the given dataset in order to increase the accuracy of analytical tasks. A detailed overview of these algorithms are given in [60].

3.4.2 Data Selection for Algorithm

This is an important issue especially for high-dimensional and large datasets. An algorithm that selects the minimal number of samples (observations) from a large dataset in order to build a Support Vector Machine (SVM) Classifier is given in [22]. On the other hand, some classification algorithms such as Decision Trees perform dimensionality reduction by selecting the set of features that increases the classifier accuracy.

3.4.3 Summary

It is obvious that there is an organic relationship between the data and algorithms to analyze data because each algorithm is biased towards some characteristic of data. If the algorithm and the dataset has a good match with respect to this bias, then the performance of the algorithm will be better on that dataset.

3.5 Relationship Between Data and Visual Representations

Each visualization type has been designed to emphasize a certain aspect about a dataset. For example, line graphs aim to present a picture of *changes over time*, maps show the physical or abstract layout on which the problem is defined. A tag cloud is a specific visualization method for textual datasets. If the dataset does not exhibit those properties that a visualization method aims to picturise, then it would not be appropriate or meaningful to visualize the data using that specific method.

There is also another aspect of the data and visualization method relationship. Some visualization methods assign the variables of the dataset to certain visual components. Different assignments give different visualizations. For example, in parallel coordinates (Figure 15) different permutations of variable assignments to the parallel coordinates change the visualization. A number of methods have been proposed to find assignments so that the amount of clutter on the visualization is minimized. Chernoff faces (Figure 17) and Andrew's plots (Figure 18) methods have a similar issue with the assignment of the variables to the components of the visualizations. An adaptive method would assign the variables onto the visual components so that some criteria would be optimized whereas a static method assigns the variables in the order that they occur in the dataset.

3.6 Relationship Between Analytical Tasks and Visual Representations

The choice of visual representations also relates to the analytical task. For example, representing data items as glyphs emphasizes the similarity/dissimilarity and grouping of the items. A tag cloud is a quick way to visually summarize a textual document. Heat maps could be appropriate views of data since they might highlight abnormal patterns and outliers in the dataset. The choice of visualization methods in order to perform a certain analytical task is generally the duty of the user. Chart Tamer [19] is an attempt to provide users with tools to help them make educated choices.

3.7 Summary of System Aspects

Data is the central component of the visual analytics process. Visualization is the way data *explains* itself to the user. As we mentioned above, the choice of visualization method depends on the characteristics of the dataset and also the analytical task the user wants to perform (Figure 29). Table 1 summarizes the system aspects of VA.

Figure 29: Relationship between visualization type, data and analytical task

Properties of Datasets	Analytical Tasks	Visual Representation Types	Relationship between Data and Analytical Tasks	Relationship between Data and Visual Representations	Relationship between Analytical Tasks and Visual Representations
Variable types *Numerical Values* *Categorical Values* *Mixed* **Dimensionality** *High* *Very High* **Time/Space relations** *Temporal & sequential* *Spatial* *Spatio-Temporal* **Relationships Between Observations** *Similarity* *Semantic relationships* **Between Variables** *Independence* *Correlation* *Causality* **Source and Quality** *Multiple sources* *Uncertainty* *Missing values* **Amount of observations** *Too much* *Not Enough*	**Summarizing Simplifying** **Detecting patterns** **Detecting anomalies** **Grouping (clustering)** **Searching & Retrieval** **Discovering relationships** **Classification Prediction**	**Descriptive** *Distributions* *Statistics (mean,variance,...)* **Relationships** *between observations* *between variables* **Picturing** *Icons and Glyphs* *Color coding* *Mathematical Shapes* **Temporal visualization** *Static timeline* *Animated timeline* **Spatial visualization** *Natural layout (map)* *Artificial layout (map)* **Spatio-Temporal visualization** *Static map* *Animated map*	**Algorithm $\Leftarrow$ Data** **Data $\Leftarrow$ Algorithm**	**Static** **Adaptive**	**User selected** **Automatic**

Table 1: System aspects of Visual Analytics

4 User Aspects

The user is the reason why Visual Analytics systems exist in the first place. The user has the final say on the decision to be made based on the data. In current visual analytics realm, there are two aspects of users: the analytical skill level and number of users and collaboration between these users.

4.1 Skill Level

Users of visual analytics systems might be coming from different disciplines with different backgrounds. Some users might be domain experts while others are newcomers to the field in which the datasets come from. Moreover, some users might possess the mathematical and statistical knowledge to be able to understand the assumptions of certain data analysis algorithms and tell if the result makes sense or not while others tend to accept whatever result comes out of the *black box*.

4.2 Number of Users

Due to the developments in the computer networks, it has been possible for multiple users to interact and work on the same analytical task. Collaborative decision making has been an important research field in information management and visual analytics promises to be a research area that would provide useful tools for multi-user (collaborative) decision making as well [24, 8].

5 Human-Machine Collaboration Aspects

There are two major aspects of the human-machine collaboration in visual analytics. First one is interactivity which dictates how the collaboration takes place. The second aspect is the benefit or usefulness of the system to the user.

5.1 Interactivity

In simple data visualization systems, the burden of making sense of data is on the user. The system would provide certain kinds of visualizations and it is up to the user to select a proper way to visualize the data and then analyze. These systems are highly interactive in the sense that humans have full control. Interaction techniques help users dynamically change the visualizations by specifying certain objectives and they may also provide a number of combined/linked views to enhance the effectiveness of the exploration. A detailed overview of interaction techniques (such as filtering, projecting, zooming, distortion, linking and brushing) is given in [36].

The opposite of user-driven strategy is the automated data analysis strategy which provides visualizations of the analysis results (such as clustering results, rules generated based on the data). These systems have minimum or very little interactivity. Visual analytics systems fall into somewhere in between these two extremes. Too much interactivity puts the burden on the user and too little interactivity leaves no space for the user to control the analytical process.

5.2 Utility

The concept of utility refers to the usefulness of the visual analytics system to the user. Recently a great deal of emphasis has been put on the evaluation [55, 6] of the visualization methods from a cognitive point of view. More and more recent research on the visual analytics field contain user studies in order to prove the usefulness of their proposed visualization techniques.

6 Conclusion

In this paper we presented a multi-faceted overview of Visual Analytics (VA). Our overview is based on the sense-making loop paradigm which was given in the VA literature [55, 35]. We discussed the three main aspects of the VA process, namely the system (machine), user(s) and the machine-user interactions. We emphasize that the ultimate goal of VA is not a fully automatic analysis of data by the system but to provide the most effective medium possible for human-machine collaboration in order to help people make sense out of today's large and complex datasets. In this collaboration, both sides offer what they do best. Humans contribute their superior perceptual skills for detecting patterns in data and machines provide the computational power for *number crunching*.

7 Acknowledgements

This work was supported in part by the National Science Foundation under grants #CNS-0722177 and #IIP-0637713.

References

[1] R. L. Ackoff. From data to wisdom. *Journal of Applies Systems Analysis*, 16:3–9, 1989.

[2] P. D. Allison. *Missing Data*. Sage Publications, 2001.

[3] E. Alpaydin. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004.

[4] D. Andrews. Plots of high dimensional data. *Biometrics*, 28:125–136, 1972.

[5] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *infovis*, 00:52, 1998.

[6] *BELIV '08: Proceedings of the 2008 conference on BEyond time and errors*, New York, NY, USA, 2008. ACM. Conference Chair-Bertini,, Enrico and Conference Chair-Perer,, Adam and Conference Chair-Plaisant,, Catherine and Conference Chair-Santucci,, Giuseppe.

[7] G. Bellinger, D. Castro, and A. Mills. Data, information, knowledge, and wisdom. `http://www.systems-thinking.org/dikw/dikw.htm`.

[8] E. A. Bier, S. K. Card, and J. W. Bodnar. Entity-based collaboration tools for intelligence analysis. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 99–106, 2008.

[9] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[10] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS J. on Computing*, 11(3):217–238, 1999.

[11] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta. *Metalearning: Applications to Data Mining*. Cognitive Technologies. Springer, January 2009.

[12] C. Chen. *Information Visualization, Beyond the Horizon*. Springer, July 2004.

[13] P. P.-S. Chen. The entity-relationship model—toward a unified view of data. *ACM Trans. Database Syst.*, 1(1):9–36, 1976.

[14] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

[15] E. H. Chi. A taxonomy of visualization techniques using the data state reference model. In *INFOVIS '00: Proceedings of the IEEE Symposium on Information Vizualization 2000*, page 69, Washington, DC, USA, 2000. IEEE Computer Society.

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

[17] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134:9–21, 2004.

[18] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[19] S. Few. Chart tamer: Excel graphs done right. In *From Theory to Practice: Design, Vision and Visualization Workshop*. IEEE VisWeek 2008, `http://www.stonesc.com/Vis08_Workshop/`, 2008.

[20] I. K. Fodor. A survey of dimension reduction techniques. `https://e-reports-ext.llnl.gov/pdf/240921.pdf`, May 2002.

[21] M. Friendly. A brief history of data visualization. In C. Chen, W. Härdle, and A. Unwin, editors, *Handbook of Computational Statistics: Data Visualization*, volume III. Springer-Verlag, Heidelberg, 2006. (In press).

[22] G. Fung and O. L. Mangasarian. Data selection for support vector machine classifiers. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 64–70, New York, NY, USA, 2000. ACM.

[23] J. Han and M. Kamber. *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, September 2000.

[24] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7(1):49–62, 2007.

[25] T. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.

[26] R. Holbrey. Dimension reduction algorithms for data mining and visualization. `http://www.comp.leeds.ac.uk/richardh/astro/pdf/alg1.pdf`, February 2006.

[27] I. Icke. Content based 3d shape retrieval, a survey of state of the art. Computer Science Ph.D. program 2nd Exam Part 1, `http://web.cs.gc.cuny.edu/~iicke/academic/survey.pdf`, 2004.

[28] I. Icke, J. Hanchi, and R. Haralick. Automatic target detection using mathematical morphology. Technical report, CUNY, The Graduate Center, 2003.

[29] I. Icke and E. Sklar. Using simulation to evaluate data-driven agent-based learning partners. In *Ninth International Workshop on Multi-agent-based Simulation(MABS'08) at AAMAS 2008*, 2008.

[30] I. Icke and E. Sklar. A visualization tool for student assessments data. In *From Theory to Practice: Design, Vision and Visualization Workshop*. IEEE VisWeek 2008, `http://www.stonesc.com/Vis08_Workshop/`, 2008.

[31] A. Inselberg. Visualizing high dimensional datasets and multivariate relations (tutorial am-2). In *KDD '00: Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–94, New York, NY, USA, 2000. ACM.

[32] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90: Proceedings of the 1st conference on Visualization '90*, pages 361–378, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[33] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[34] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[35] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information Visualization*, pages 154–175. Springer, 2008.

[36] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[37] R. Lengler and M. Eppler. Towards a periodic table of visualization methods for management. In M. S. Alam, editor, *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007)*, Calgary, AB Canada, January 2007. ACTA Press.

[38] X. Li. Data of t- and b-cell acute lymphocytic leukemia from the ritz laboratory at the dfci (includes apr 2004 versions). `http://www.bioconductor.org/packages/release/data/experiment/html/ALL.html`.

[39] The map lab. `http://datamil.delaware.gov/`.

[40] MathWorks. Matlab. `http://www.mathworks.com`, 2009.

[41] G. A. of the United Nations. The universal declaration of human rights. `http://un.org/Overview/rights.html`.

[42] R project. `http://www.r-project.org/`.

[43] C. D. Shaw, J. A. Hall, C. Blahut, D. S. Ebert, and D. A. Roberts. Using shape to visualize multivariate data. In *Workshop on New Paradigms in Information Visualization and Manipulation*, pages 17–20, 1999.

[44] Historical data for s&p 500 stocks. `http://biz.swcp.com/stocks/`.

[45] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[46] C. Tominski, J. Abello, and H. Schumann. Interactive poster: 3d axes-based visualizations for time series data. In *EEE Symposium on Information Visualization (InfoVis)*, 2005.

[47] A. Treisman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156–177, August 1985.

[48] E. Tufte. *Data Analysis for Politics and Policy*. Prentice Hall, 1974.

[49] E. R. Tufte. *The Visual Display Of Quantitative Information*. Graphics Press, 1983.

[50] E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.

[51] E. R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, February 1997.

[52] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, (1977).

[53] Uci machine learning repository. `http://archive.ics.uci.edu/ml`.

[54] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: A comparative review. Submitted to Neurocomputing, `http://ticc.uvt.nl/~lvdrmaaten/Laurens_van_der_Maaten/Matlab_Toolbox_for_Dimensionality_Reduction_files/Paper.pdf`, 2008.

[55] J. van Wijk. The value of visualization. In *In: C. Silva, E. Groeller, H. Rushmeier (eds.), Proc. IEEE Visualization*, pages 79–86, 2005.

[56] *Proceedings of VDM@ECML/PKDD2001 International Workshop on Visual Data Mining*, 2001.

[57] F. B. Viégas and M. Wattenberg. Artistic data visualization: Beyond visual analytics. In D. Schuler, editor, *Artistic Data Visualization: Beyond Visual Analytics*, volume 4564 of *Lecture Notes in Computer Science*, pages 182–191. Springer Berlin / Heidelberg, 2007.

[58] M. O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3/4):194–210, 2002.

[59] C. Ware. *Information Visualization: Perception For Design*. Elsevier, 2004.

[60] L. Yang. Distance metric learning: A comprehensive survey. `http://www.cse.msu.edu/~yangliu1/frame_survey_v2.pdf`, 2006.