ENVISIONING KNOWLEDGE: TIGHTLY COUPLING KNOWLEDGE ANALYSIS & VISUALIZATION

Ketan K. Mane

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Library and Information Science of,
Indiana University
October 2006

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Katy Börner, Ph.D.
Principal Advisor

_____

Javed Mostafa, Ph.D.

_____

Luis Rocha, Ph.D.

_____

Sun Kim, Ph.D.

Oct 2nd, 2006

To my mother

Chhaya Kirtiraj Mane

Acknowledgements

Over the last four years, I had an opportunity to work with many people who made my time at Indiana University enjoyable and intellectually stimulating. I would like to take this opportunity to thank all the people who have helped me along the way and contributed towards this dissertation. Here I acknowledge people who made – "Been here (Indiana University), done that (Ph.D.)" possible for me.

I am especially grateful to have Katy Börner as my advisor. She has been a great mentor and has always been a source of inspiration for me. I appreciate her motivation, support, eye for perfection and teaching. Over the years, I enjoyed collaborating with her on several projects and many successful publications. I deeply appreciate the freedom that Katy gave me to pursue my interest in biomedical research area. Being around her, I have always admired her for her qualities and insights related to information visualization research. When I joined the Ph.D. program, I remember Katy welcoming me by saying 'Welcome to the most rewarding, exciting and stimulating years of your life'. It has indeed been true and more it was a great pleasure to work with her.

I would also like to thank Javed Mostafa, Luis Rocha and Sun Kim, who were a part of my doctoral research committee, for their time and valuable input during the course of my research. It was great to work with Javed on some of the information retrieval projects. I have always admired Luis for his research acumen, and it was always a pleasure to talk and listen to him. Sun Kim also provided good insights related to my minor in bioinformatics.

During the course of my doctoral studies, I had an opportunity to collaborate with various researchers and work on different projects. I would like to thank Kevin Boyack at Sandia National Lab for fruitful collaborations and his feedback on several occasions. I would also like to thank Dr. Susanne Ragg at the Indianapolis campus of Indiana University for her valuable insights about the medical dataset in the computational diagnostics project. It was always a great experience learning new data mining techniques with Kiduk Yang. I also appreciate the valuable insights offered by Blaise Cronin, Debra Shaw, Bradford Paley, Stephen North, Tamara Munzner and others.

For all these years, it was great to have the support of all faculty members from my department. I would also like to thank the departmental computer staff for their administrative support. I extend my thanks to Sarah Burton and Mary Kennedy for entertaining my requests related to processing my late time-sheet submissions. Further, Rhonda Spencer, Arlene Merkel, and Ericka Bodner were always great to talk to and helped me find solutions to some of my non-traditional questions related to the doctoral program.

At Indiana University, I was a member of Information Visualization Lab and also a member of Cyberinfrastructure for Network Science Center directed by Katy Börner. I enjoyed learning, working on several individual/group projects and stimulating discussions with other members of the lab. It was a pleasure to work other lab members: Shashikant Penumarthy, Weimao Ke, Gavin La Rowe, Peter Hook, Wexia (Bonnie) Huang, Bruce Herr, Soma Sanyal, Stacy Kowalczyk, John Burgoon, Sumeet Ambre and othes. I would like to thank Julie Smith for their help with proof-reading this document for grammatical errors.

I am also thankful to the research group at Los Alamos National Lab (LANL), New Mexico for hosting me as an intern in summer 2005. I really enjoyed working on different projects with Linn Marks, Rick Luce, Mark Martinez and other researchers at LANL. It was a great experience and a great place to work.

I would also like to thank friends from Bloomington for all the enjoyable and memorable times during the course of my studies. I cherish the support and memories of all the great time that I spent with my roommates: Sidharth Thakur, Vinayak Dukle and Amit Saple. It

was fun to go on hiking trips and camping with Sidharth; have eating competition with the unchallenged Vinayak Dukle, and hang out with the ever enthusiastic Amit Saple. I would also like to thank my others friends for the good times and memories while at Bloomington: Sriram Sankaran, Suchitra Mohan, Dipa Sarkar, Harshwardan Gadgil, Prajakta Vaidya, Vikas Gupta, Rahul Doshi, Avi Kevalramani, Sumit Middha, Pavan Ghaige, Deep Ved, Lalitha Viswanath, Jasleen Kaur and others. While at Indiana University, I enjoyed the support and had great time with my international friends. I enjoyed being a part of the international community and enjoyed their cuisines. For all the great memories, I would like to thank Ning Yu, Seung-min Lee, Bi-Yun Huang, Kazuhiro Seki, Yueyu Fu, Dan Kurtz, Peter Hook, Gavin La Rowe, Yi Sun and others. I would also like to extend my thanks to Kathie and Jim Lazerwitz for being a wonderful host family and making me feel at home and exposing me to American culture and traditions.

I also enjoyed the support and motivation from my other friends from my M.S.; Amol Patki, Girish Deshpande, Rakesh Patel and Rashmi Shastri. Thanks to Abhinav Chandra, Siddharth Pavithran, Milind Nemlekar, Sachin Saldana and others, my old time buddies from school for their encouragement. I would also like to thank Harmanjit Randhawa and my other colleagues at US. Army Aeromedical Research Lab in Alabama for their encouragement to pursue my doctoral studies. I appreciate the support, motivation, encouragement and enthusiasm from new friends, Supriya Naik and others. As a part of the enthusiastic undergraduate group, I always enjoyed the support and motivation from my undergraduate friends in USA and in India: Vidya Venkatsubramanam, Pulkit Desai, Deepa Narayanan, Ashutosh Khandha, Ninad Patnekar, Kaustubh Dhavse, Sarika Bhagat and others.

I would like to acknowledge the support of my family members. Although I was thousands of miles away, I always felt I was at home for which I would like to thank my cousin sister Maneesha Mulye and Madhusudhan Mulye. I would like to thank all my relatives. Especially I would like to thank my younger cousin Sneha and elder cousin Sharayu for always remembering me and making me a part of all their holiday celebrations.

Last but not the least, I acknowledge, appreciate and enjoyed the support from my mother Chhaya Mane. After my dad, she has been a great source of inspiration and motivation to me. Besides being a great mother, she has been a great mentor. She has advised me and has been a great supporter of my decisions. Her motivation, support, encouragement and confidence in me have helped me a lot. I dedicate my thesis work to my mother.

Ketan K. Mane

ENVISIONING KNOWLEDGE: TIGHTLY COUPLING KNOWLEDGE ANALYSIS & VISUALIZATION

Abstract

Analysis and visualization techniques share a symbiotic relationship when it comes to making sense of datasets. Particularly for large datasets, the coupling of data analysis and data visualization is often beneficial. While a gamut of data analysis and visualization techniques exist, it is often problematic to identify what combination of techniques is good for what task to provide maximum insight into a dataset.

This thesis introduces, exemplifies and validates a Data Analysis and Visualization Taxonomy, called 'DA-Vis taxonomy', that provides guidance to the selection of complementary analysis and visualization techniques. The DA-Vis taxonomy is validated by demonstrating its utility to develop new visualizations for real world applications. Further, the new taxonomy is applied to systematically describe and classify couplings of data analysis and visualization techniques in prior work. A user-study that evaluates the usability of the DA-Vis taxonomy was also conducted and is reported here.

The intellectual contributions of this thesis include a flexible DA-Vis layout schema that can be used to tightly couple complementary data analysis and visualization techniques. This thesis also shows a visionary computational diagnostic tool developed for data analysis and visualization of clinical data. Techniques used to generate meaningful knowledge management visualizations from a dataset are presented as a part of this thesis. This thesis concludes with a discussion of the broader impacts of the DA-Vis taxonomy, the computational diagnostic tool, and knowledge management maps.

Table of Contents

List of Figures

List of Tables

Chapter 1

Motivation

The information age has led to the large scale digitization of data. Examples of datasets include: patient medical records, genomics and proteomics data, published literature, chat interaction data, market analyses, etc. Analysis of a patient medical dataset can provide insight about new ways of diagnosis and treatment of diseases; genomic and proteomics data can be used to identify roles of genes and proteins; published literature can be analyzed for emerging research topics; market data can be used to identify trends in customer shopping, sales, etc. Results from data analysis aid the decision making process. While the amount of data is increasing at an exponential rate, our brain's capacity to adsorb and analyze data remains nearly constant. Valuable knowledge can be derived from a dataset when the raw data is processed and presented in a suitable format. Depending upon the user's tasks, different tools are needed to process and convert the raw data into a format which would help to augment user's understanding of the data. Further, owing to the scale of datasets being processed, the process of analysis needs to be automated as much as possible.

The data analysis community has developed a wide range of data analysis algorithms and techniques. The techniques help to meet different primary tasks such as: classification into pre-defined classes, clustering based on shared characteristics, identification of associations, detecting trends, exploratory data analysis (EDA), and identification of other important data characteristics. Primarily, the user's choice of a data analysis technique is dependent on the research question that needs to be answered. For example, say the user's task is to find the

semantic relation between object A and object B from a given set of documents. Within the data analysis domain, this task can be categorized under the primary task named 'identify association'. Most data analysis textbooks classify different data analysis techniques at such a primary task level. Further, textbooks lack details about the format of results obtained when a data analysis technique is used. In addition, no book covers extensive information about alternative data analysis techniques that are available to accomplish a given user task and the data input that would be supported by a given data analysis techniques. Hence for a novice user, no reference system exists that would help them identify data analysis techniques that best suit their specific task requirements (i.e., in case of the example: 'identify semantic relation').

Further, it is important to show analysis results in a format that is easy to interpret. Graphical presentations can present the data and analysis results in a meaningful format for interpretation. For centuries, graphic designers have manually drawn graphical illustrations of data to help expose embedded data characteristics [1]. Over the years, the visualization community has drawn inspiration from graphic designers to design aesthetic and meaningful visualizations by using layout algorithms. From a data analysis perspective, these layout algorithms help to satisfy different visualization goals such as: compare, distinguish, indicate, identify, relate, represent and reveal [2]. Generating meaningful data presentation is a primary focus of visual layout algorithms as it leads to superior insights. Knowledge on how to map data characteristics to visual metaphors can also be derived from cartography literature [3] and from cognitive science [4]. Further, interactivity is one of the foremost advantages offered by layout algorithms over manually drawn graphical illustrations on paper. Data interaction becomes extremely useful when large datasets need to be visualized. Data interaction engages the user in information foraging and information discovery [5]. Some visualizations offer a birds-eye view of the entire dataset. But the display of large scale datasets without any pre-processing often result in cluttered visualizations that fail to communicate meaningful information. A pre-processing data analysis step becomes essential when dealing with large scale datasets.

The above discussion shows that data analysis techniques share a symbiotic relationship with data visualization techniques in facilitating the interpretation of large datasets. However,

data analysis and visualization techniques are developed in different domains by different scholarly communities. The data analysis community has developed algorithms using various techniques from artificial intelligence, statistics and mathematics while the visualization community has utilized techniques from cartography, graphic design, scientific visualization and human computer interaction [6]. It is often a challenge to determine what techniques should be combined in what sequence. A naïve researcher from either domain who looks forward to perform a large scale data analysis is often perplexed by the number of available choices. The results from an informal questionnaire discussed in chapter 2 show that researchers are unaware of alternate techniques that are available to perform the same data analysis. Further, review of existing taxonomies in chapter 4 shows that none of the prior taxonomies cover details about establishing a bridge between data analysis techniques and data visualization techniques. No taxonomy exists to show mutually complementary data analysis and visualization techniques that can be used to make sense of large scale datasets.

Priolli and Card describe sense-making as 'a process that requires building of appropriate classification schema to promote better understanding of information' [7, 8]. It is important to address this issue not only when it comes to designing a schema but also while performing a data analysis. The task of seeing different data characteristics becomes intuitive when data characteristics are shown using appropriate design principles [9]. To facilitate easy mapping with data and its characteristics, it is important to enumerate which data mapping works with a given visualization. In addition, the new 'Visual Analytics' (details in section 2.1.3) initiative aims to develop a framework for visual data analysis by coupling techniques from data analysis and information visualization to gain a quick understanding of the data characteristics [10].

This dissertation project attempts to address this need by developing a taxonomy that couples data analysis techniques and visualization techniques to support different user tasks. The new taxonomy shows different pathways of data analysis and visualization techniques that complement each other in the analysis. It aims to support the user in selecting combinations of techniques that meet the user's task requirements. The taxonomy can also be used to organize the different techniques available to perform a given data analysis task. Further, the requirement to establish a guideline to show certain data characteristics is

addressed by the new taxonomy. Henceforth in the entire thesis, the new taxonomy is referenced as 'DA-Vis taxonomy'. While multiple definitions of the word 'taxonomy' exist, for the thesis the word 'taxonomy' is defined as 'classification of related techniques into a common category'. With the current definition, some techniques can also be classified under two different categories, if they satisfy the requirements of the category.

Figure 1 shows the scope of the proposal. As can be seen, the new coupled DA-Vis taxonomy bridges techniques from two different communities – data analysis and data visualization. The DA-Vis taxonomy covers the theoretical core for analysis and visualization of data. It is exemplarily instantiated for n-dimensional and network datasets.

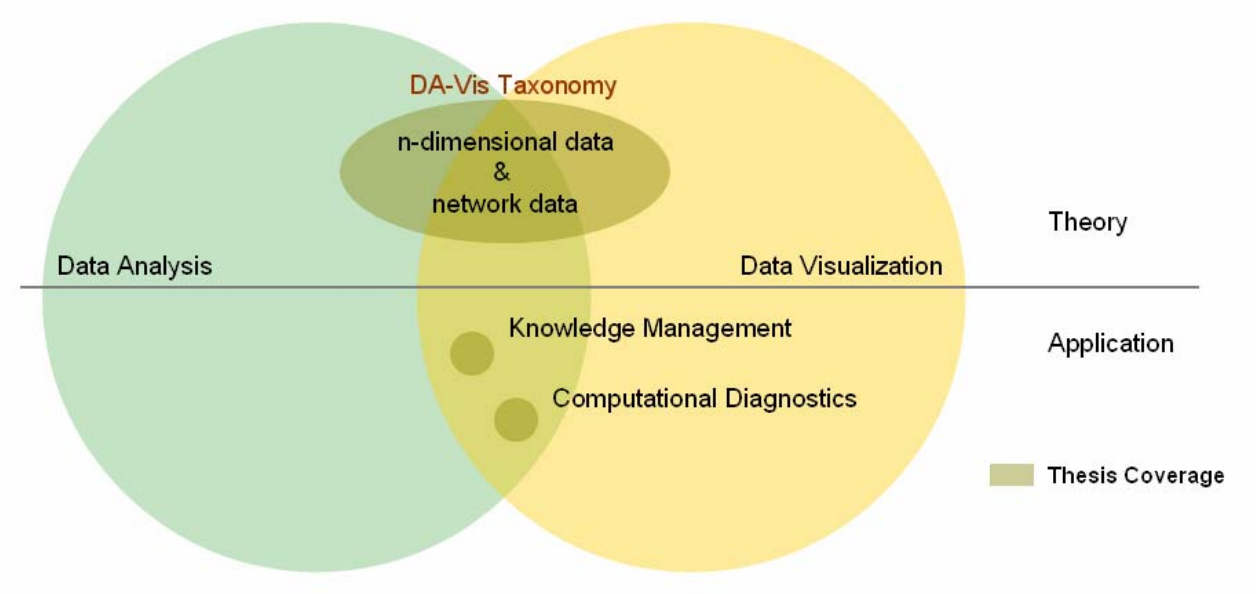


Figure 1: Scope of the thesis

The application of the DA-Vis taxonomy will be demonstrated using datasets from two different domains – Computational Diagnostics and Knowledge Management, see

Figure 1. The dataset for the former is comprised of medical records of acute lymphoblastic leukemia patients (ALL) while the latter dataset is a 20 year time slice of 'The Proceedings of National Academy of Science' (PNAS) publications. Different pathways from the newly developed DA-Vis taxonomy are used to identify the data analysis and visualization techniques that will best meet the user's task needs. The new DA-Vis taxonomy was further evaluated by demonstrating its ability to describe combinations of data analysis

and visualization techniques reported in literature. A formal user study of the DA-Vis taxonomy was also conducted to evaluate its user-friendliness and results are presented in chapter 8.

The different chapters of this thesis are organized as follows: Chapter 2 reviews existing work on sense-making, visual design principles, and visual analytics. This chapter also identifies the user tasks and the generic approach taken for data analysis. Chapter 3 details different data analysis techniques and algorithms that are used by the information visualization community. This chapter also covers details about data characteristics and data dimensions. The data analysis taxonomy is also covered in this chapter. Chapter 4 details different types of data visualization, layout algorithms and interaction techniques. This chapter reviews and compares existing taxonomies in the information visualization domain. Chapter 5 presents the theoretical core of the DA-Vis taxonomy for n-dimensional and network datasets. Chapter 6 validates the new DA-Vis taxonomy by showing its application to generate visual designs using datasets from the biomedical research domain and knowledge management. In chapter 7, a second metric for DA-Vis taxonomy validation is introduced to show its utility to describe and organize prior published research work. Chapter 8 provides a methodology to validate the usability of the DA-Vis taxonomy. The thesis concludes with a discussion of the intellectual merit and broader impact of this research in chapter 9. In the following chapters, the first occurrence of terms that occur in the glossary are underlined and annotated by '₪'. Definitions of these terms are available in Appendix I.

Chapter 2

Envisioning Knowledge: Theory & Practice

Data analysis can be carried out using different combination of data analysis techniques. Based on user-task, an identification of the sequence of steps will be helpful in the sense-making process and visual analytics process. Further, one can develop meaningful visualization by using certain known design principles. This chapter details prior work on sense-making process and design principles. It also discusses the emerging field of visual analytics. Further, it details the process model for envisioning knowledge and the user-tasks that are commonly used within the information visualization community.

2.1. Prior Work

2.1.1. Sense-Making Process

Sense-making is a process that provides a theoretical framework to the sequence of steps involved in data analysis and data cognition [8, 10]. The sense-making process is a cyclic loop that involves an ordered sequence of steps such as:

A. **Gather information** – Information foraging and the data pruning process to filter information of interest.

B. **Re-represent information** – A certain schematic data representation technique is adopted to provide a data overview or to support the data exploration process.

C. **Develop insight** – Data representations help with the cognitive interpretation of the dataset and generate hypotheses based on the observations.

D. **Produce results** – Evidence based hypothesis generation.

The above four steps are shown in Figure 2 below. The sense-making process can involve several iterations of all steps (a – d) in sequential order or several iterations between two intermediate steps. Further, depending on the situation the sense-making process presents different models. A generic model, called the 'cost structure model', is based on the evaluation of time spent in information foraging versus the information gained [8]. A time constraint based sense-making model involves quick pruning through vital data characteristics to make judicious decisions [11]. The learning sense-making model involves an effort to represent data in a format that best shows important data characteristics [7].

As analysis and visualization techniques are a part of the sense-making process, it can be inferred that coupling complementary techniques can lead to better data analysis. Recommendations from experts [10] as to how to develop this coupling include: understand the task constraints of performing different data analysis approaches on certain data types; distinguish between different task goals, and application of technique or combinations of techniques that help to meet the user goals; identify common user tasks and formulate a core conceptual schema that shows techniques that can be used to obtain data interpretations. In short, a schema or a taxonomy developed based on task, data-types and associated techniques will highly benefit the sense-making process.

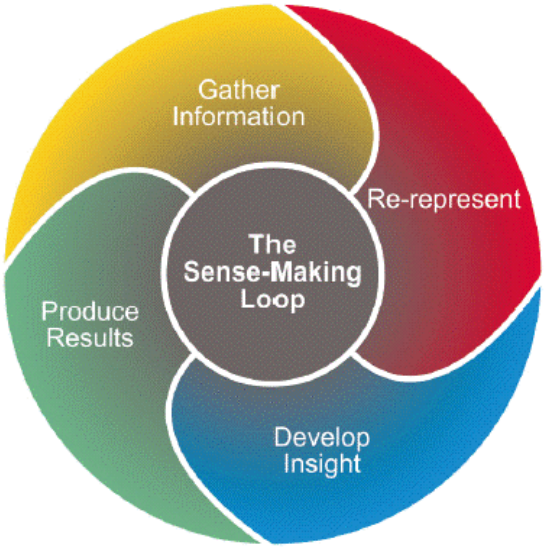


Figure 2: The sense-making loop, adopted from [10], reproduced with permission from © 2005 IEEE

2.1.2. Design Principles for Envisioning Information

A few strategies that can be used to build meaningful visualizations are highlighted by Tufte [9]. These strategies show certain design principles that help to produce data representations that can communicate information more effectively. Major strategies are listed below,

A. **Escaping the flat-land** - Provide strategies to increase the amount of data dimensions and density displayed. This strategy is normally achieved by maintaining a constant base map and overlaying information on top of it.

B. **Micro/Macro information** – A global overview of the dataset can be obtained with the display of information at macro-level. Finer dataset details are reveled in the visualization with the help of micro-level information.

C. **Layering and Separation –** This strategy is often used to reduce the clutter in the visualization. This strategy can be used to show each layer of data very distinctly. Priority can be given to show the important information in a visualization. Different colors, shapes, values, and sizes can be used to distinguish between different data entries.

Difference in color saturation can be used as a separation metric to direct one's attention to important data characteristics.

D. **Small Multiples –** Miniature images having the same background template can be used to show subtle changes in the data. By lining small multiple images in a sequence, one is able to compare the changes at a single glance.

E. **Color and Information -** Color is an important ingredient to communicate distinct pieces of information. It can be used to represent quantitative values that can be readily used for visual comparison. Colors can help to distinctly demarcate different regions in the visualization.

F. **Narratives of space and time -** Representation of more than two dimensions in 3D space always presents a problem. Animation and dynamic visualization can help to reduce the complexity involved in coding visualization in more dimensions. Even if additional space is present, it should be actively used to present different data characteristics.

By using the above strategies to envision information, one can improve the data cognition process [10, 12]. Further, using principles of Gestalt psychology, association techniques, etc., one can generate visualizations to show existing relationships within large datasets in a small display area. One can use these strategies to monitor activities in visualizations both at the global and local levels. Finally unlike traditional static visualization, an interactive visualization provides means to support the data exploration process. In short, by using design representation techniques, certain salient characteristics of an abstract data can be presented to the user for knowledge discovery.

2.1.3. Visual Analytics

Visual Analytics is a new initiative from the National Visualization and Analytics Center (NVAC) at Pacific Northwest National Lab. Visual Analytics is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [10]. The visual analytics goal

is to develop new approaches that will help to analyze and aptly gather intelligence from large amount of data. The intelligence gathered can be used to make judicious decisions.

To support tasks, it is essential to develop methods that support the selection of appropriate techniques for the data analysis task. It is important to identify techniques that can offer maximum insight about the data. In short, it becomes important to develop a framework that aims in identifying individual or combined data analysis techniques that support the user goals.

2.2. Process Model of Envisioning Knowledge

The data processing performed during different stages of data analysis and visualizations are indicated in Figure 3. A more detailed description of user tasks is covered in section 2.3 below. Chapter 3 covers details about different data analysis methods. Chapter 4 covers details about different data visualization and interaction techniques.

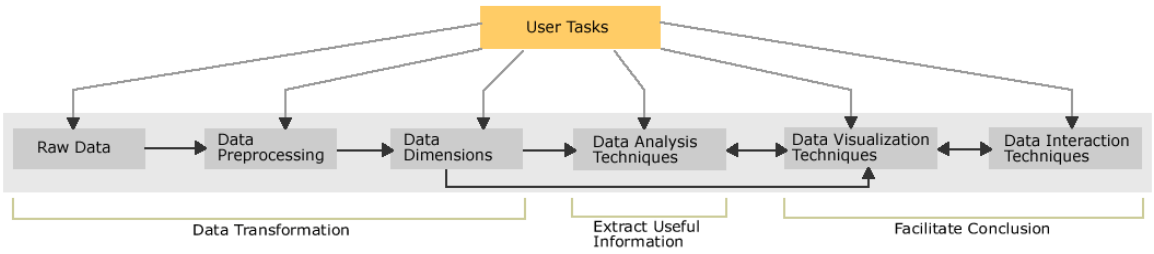


Figure 3: Different stages of data processing

Figure 3 shows three stages: 1) data transformation, 2) extract useful information and 3) facilitate conclusion. The data transformation stage involves identification of data dimensions of interest from the raw data. In the next stage, different data analysis methods are used to capture specific data characteristics that help to provide answers to a given user task. Later, the data visualization techniques are used to show data characteristics and derive conclusions about the dataset. Using visual interaction techniques helps the user to interact with and explore the dataset. An additional data analysis process is sometimes required to communicate results via interaction. The dual arrow lines are used to communicate between

data analysis, data visualization and interaction modules. Some visualization techniques (e.g., parallel coordinates; in chapter 4) can be directly used to visualize the data dimensions. Data analysis can be skipped as indicated in Figure 3.

2.3. User Tasks and Needs

An informal questionnaire was distributed among researchers to identify data analysis technique they use to perform different data analysis tasks. To start, the common data analysis algorithms were scientifically classified into sub-categories based on the user needs. In general, the questionnaire listed different data analysis techniques used by information visualization researchers to accomplish different user tasks. A copy of the distributed questionnaire is included as Appendix II. As feedback, researchers were asked to indicate which data analysis techniques they used. If the preferred data analysis technique was not given, researchers were asked to list them in the questionnaire. Feedback from eight information visualization researchers was obtained.

The feedback showed little overlap between different data analysis techniques used by researchers. This lack of much overlap could be due the fact that researchers often tend to use data analysis techniques they know. Unfamiliarity with other data analysis techniques used to perform similar tasks restricts an overlap in usage. A proper categorization of data analysis technique based on user needs would help to overcome usage unfamiliarity of different available techniques. This need is addressed by the developing DA-Vis Taxonomy; it offers a global overview of different data analysis tasks based on user needs. The details about the DA-Vis taxonomy are covered in chapter 5.

Figure 3 indicates that the user tasks play a key role in the selection and parameterization of every data processing stage. The user tasks described here are data analysis tasks that are specific to a dataset. For example, the user task can be an identification of trends in research topics within a set of documents; or with medical records the user task can be to identify common medical conditions of a selected group of patients.

Some primary tasks addressed by the information visualization community are listed below. The scope of the primary task is described using brackets. Associated specific data analysis tasks are then identified.

A. Primary Task: Identify Associations

[Scope: Identify linkages between data entries]

Specific Tasks: find correlations, compute linkage strengths, establish semantic linkages, etc.

B. Primary Task: Identify Patterns

[Scope: Apply data analysis techniques to show data characteristics]

Specific Tasks: in ordered data, based on correlation, by clustering data, etc.

C. Primary Task: Identify Trends

[Scope: Identify data characteristics by performing cumulative data operation]

Specific Tasks: sort based on data occurrence, research activity, selection criteria, etc.

D. Primary Task: Identify Clusters

[Scope: Identify set of entries with similar data characteristics]

Specific Tasks: Select based on data semantics, data partition, data hierarchy, etc.

E. Primary Task: Identify Important Data Characteristics

[Scope: Identify the data dimensions that help to describe the data]

Specific Tasks: Sort based on data dimensions, linkages, etc.

F. Primary Task: Identify Structural Patterns in Network Data

[Scope: Identify network topology]

Specific Tasks: Identify network authorities, identify network topology, identify central nodes that maintain network connectivity, etc.

The new 'DA-Vis' taxonomy discussed in chapter 5 is based on these primary data analysis tasks. The new taxonomy offers insight about complementary data analysis and visualization techniques that can be used to perform these user tasks.

The user tasks for different datasets can be abstracted to a specific data analysis (DA) task. This abstraction helps to identify corresponding primary data analysis tasks. Examples used to demonstrate this abstraction are shown below.

User Task	**Specific DA Tasks**	**Primary DA Tasks**
- Identify trends in research topics within a set of document	- Determine topical frequency	- Identify pattern
- Do the patients who are either dead/alive share common medical condition for some variables?	- Identify correlation	- Identify association

The above example shows only one specific data analysis (DA) task that can be used to accomplish the user task. But in reality, many specific tasks exist that can be used to serve the same functionality. The specific tasks that serve the same functionality can be categorized under a primary task. The unique combination of primary and specific task helps to identify a user task requirement and the data analysis method.

Figure 3 shows three main stages of the data analysis process and the data processing steps within each. These data processing steps are henceforth referenced as 'modules'. To better understand the complementary techniques shown in DA-Vis taxonomy (chapter 5), it is important to understand data analysis and data visualization techniques in more detail. Different modules of data transformation stage are covered in chapter 3. It also covers information about relevant data analysis methods which are applied to extract data characteristics from a dataset. Data analysis results can be communicated using diverse visualization techniques [5, 13, 14]. The modules which help to provide insight and facilitate conclusion about the dataset include visualization and interaction. Chapter 4 details different

visualization techniques in terms of layout options, data display, and visual encoding to show different data characteristics. Different interaction techniques are covered in chapter 4.

This chapter gives a general overview of different stages involved in the data analysis process. Based on these data analysis stages, researchers from the information visualization domain have generated various taxonomies that are covered in chapter 4.

Chapter 3

Review of Data Analysis Techniques

Data analysis is defined as "an act of transforming data with the aim of extracting useful information and facilitating conclusions" [15]. The definition shows three main stages that form the backbone of almost any data analysis process: 1) data transformation, 2) extract useful information and 3) facilitate conclusion, see also Figure 3.

3.1. Data

3.1.1. Raw Data Stage

The data analysis process starts with the raw data that comes in variety of formats, e.g., flat files, spreadsheets, or relational tables. There are three main data type categories: a) Nominal, b) Ordinal and c) Quantitative. Characteristics of each data type are discussed below [5, 16].

A. Nominal: All non-numerical data values are categorized as 'nominal'. Entries such as label names, text data, chromosomal conditions in patients, etc. are some examples of the nominal data type. Typical comparison between two nominal entries is done using string operators of equality and non-equality. For classification, numerical values need to be assigned to nominal data objects. Data objects that belong to the same category are also called 'Categorical'.

B. Ordinal: If numerical values to data objects exist in a particular order (increasing/ decreasing) then they are referred as 'ordinal'. Student ranks in a given exam or clinical values by date are examples of ordinal data. In addition to operations of equality and inequality, numerical operations of 'less than' and 'greater than' can also be applied to this category of data types. This data type is also referred to as 'rank variables'.

C. Quantitative: Data objects with no specific order to numerical values are referred to as 'quantitative' entries. Entries like blood cell count (WBC) for patients or research papers are available in different domains. Numerical values for quantitative data objects make it possible to apply all arithmetic operators of comparison on it.

3.1.2. Data Pre-processing

Usually, different data pruning techniques are applied to raw data to filter out unwanted data variables. This stage is referred to as the 'data preprocessing' stage, and is the next stage of the data analysis process. The main goals of the data processing stage are: 1) to fix problems existing in the current data and 2) to prepare the data for the next stage of analysis. Data processing can involve steps like fusing data from multiple sources, eliminating noise from the data, removing duplicates and filtering subset of data for further analysis. Some of the popular data pre-processing techniques are covered here.

A. Stemming: Within datasets, there exist many morphological variants of the same word. For example, pairs of terms such as 'computing' and 'computation' will not be recognized as equivalent without some form of natural language processing (NLP) [17]. An algorithm that helps to establish equivalence based on their stem word ('comput') is called 'Stemming Algorithm'. Usage of this algorithm helps to reduce different variants of a term to a single representative form. By doing so, the stemming algorithm eventually reduces the number of distinct terms that are essential to represent a set of documents.

B. Stop-Word Removal: List of words that do not add value to the meaning of a dataset are readily removed by 'Stop Word Removal' techniques. Typically, stop words include

grammatically used conjunctions such as ‘and’, ‘or’; articles such as ‘the’, ‘a’, ‘an’, etc. This is a simple data pruning technique.

C. Frequency Detection: Using this approach, the number of occurrences of unique data objects is computed.

D. Threshold Application: This approach is useful to obtain a subset of data of interest from the original dataset. This method imposes a condition to filter data above and/or below a chosen threshold value.

These data processing techniques help to trim unwanted data variables as well as reduce data dimensions. More information on the different data dimensions and data types are covered in the next section. Thus the contribution of the data processing stage is to prepare the data for efficient data analysis.

3.1.3. Data Types/Data Dimensions

Parameters that help to describe the characteristics of a data record are referred to as its ‘dimensions’. Different data dimension categories and their details are covered below [13, 18]. The top four options (A-D) are based on the number of data dimensions available. The other options (E-G) are based on the given information on data dimensions and their relations.

A. One-dimensional data – A dataset with the data object associated with one variable (e.g., time) are referred to as one dimensional data. For example, time-series of stock prices, time series of patient clinical data, etc.

B. Two-dimensional data – A dataset with the data objects associated to two variables are termed as two dimensional data. For example, geographical maps, floor-plans, newspaper layouts, etc.

C. Three-dimensional data – Data objects with three variables of information to describe its details are referred to as three dimensional data. For example, molecular data, flying coordinates, etc.

D. Multi-dimensional/N-dimensional data – A dataset comprised of data objects with information distributed among 'n' different variables are referred to as multi-dimensional or n-dimensional data. For example, patient medical records, article bibliography, etc.

E. Tree – A dataset that involves parent child relationship among its data objects is classified as 'Tree'. The rule applied here is that 'children' data objects have a link to a single 'parent' data object. This rule results in a hierarchical structure. For example, library classification schema, file directories, etc.

F. Network – Datasets with one-to-many or many-to-many relations between its data objects are classified as networks. For example, genes/proteins co-occurrence network, etc.

G. Text data – Datasets where each data object are comprised of a single line of text containing a string of characters are categorized under text category. For example, text data from a book, etc.

It should be noted that some researchers distinctly classify data objects with time variables under a separate category termed 'temporal' [18]. Other researchers tend to combine two data dimensions under a single category name. For example, tree and network data have been categorized under a single category called 'Hierarchies and graphs' in [13].

3.2. Data Analysis Techniques

The data analysis techniques are a part of the sense making process. This section reviews different types of data analysis techniques that are commonly used for data analysis. Only techniques commonly used by the information visualization community are covered here.

3.2.1. Similarity Measures

Similarity measures help to identify associations between different entities of the dataset. They primarily establish dataset entity-entity linkages.

A. Co-occurrence Similarity

The co-occurrence similarity measure is used to count a document collection containing two different data entities [19]. The data entities can be a term, author, paper, etc. To compute the co-occurrence similarity, the same data entity is arranged in the form of rows and columns. The matrix cell values indicate the co-occurrence strength among two data entities. For example, related topics from a dataset can be identified by computing co-occurrence similarity among terms; strength of collaboration among authors can be identified using co-authorship information to compute similarity, etc. Typically, the co-occurrence similarity space is generated for high frequency data entities.

B. Cosine Similarity

The cosine similarity is a vector-based approach used to determine the similarity among data entities. To compute the cosine similarity among terms in a data collection, co-occurrence frequency of terms in a data collection, and individual term frequency are calculated. Cosine similarity between two terms (X and Y) in a document collection (D) is computed by treating each term as a vector by using the Salton cosine formula [20], see Equation 1:

$$Sim(X,Y) = \frac{D_{X,Y}}{\sqrt{D_X \, D_Y}} \quad \textbf{(1)}$$

where:

D_X is the number of documents with term X

D_Y is the number of document with term Y

$D_{X,Y}$ is the number of documents with both terms X and Y

The cosine similarity measure ranges from '1' for the highest similarity to '0' for the lowest.

C. Jaccard Index

The Jaccard index is defined as "the size of the intersection divided by the size of the union of the sample sets" [21]. It is also known as the Jaccard distance or Jaccard similarity coefficient. It is an index that helps to compare the similarity and diversity that exists within a dataset. If A and B are two different entities then the Jaccard index is computed as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \textbf{(2)}$$

where,

$A \cap B$ is the intersection set of papers with both A and B

$A \cup B$ is the union set of papers A and B. It is the total papers with either A or B

D. Pearson Coefficient

The Pearson coefficient is a measure that can be used to describe the relation between two entities measured from the same object. The correlation coefficient is also known as Pearson Product Moment Correlation [22, 23]. It is calculated by:

$$r = \frac{\sum_{i=1}^{n}(x_i - x')(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}} \quad \textbf{(3)}$$

where,

x_i is used to represent different entity values from document x

x' is the mean of all entity values for document x

y_i is used to represent different entity values from document y

y' is the mean of all entity values for document y

The coefficient ranges from −1 to 1. A value of 1 indicates a linear relationship within the two entities. A value of −1 shows inverse relationships. A value of '0' indicates no relationship between the variables.

For example, to compute the Pearson coefficient between documents in a dataset based on the occurrence of words, we need to initially build a document-term occurrence matrix. In this matrix, rows represent documents and columns represent the terms. The matrix cells will be populated with the count of words in the document. Using values from each row, Pearson's correlation coefficient among documents can be computed. As positive value indicates that the values are related, the positive value can be treated as a strength with which the documents are related.

E. Average Relatedness Factor

The average relatedness factor is a measure of journal-journal relatedness based on journal inter-citation frequencies [24, 25]. Given journal A, the other journals it cites within its papers are termed as 'cited journal'. On the other hand, journals which cite journal A most often are referred to as 'citing journals'. The cited and citing journal information is utilized to identify the relatedness among journals. The relatedness among two journals 'A' and 'B' is calculated by using the relatedness formula shown here,

$$R_{A>B} = \frac{C_{A>B} * 10^6}{N_B * C_A} \quad \textbf{(4)}$$

where,

$R_{A>B}$ is the relatedness of journal A citing journal B

$C_{A>B}$ is the number of citation in the current year from journal A to journal B

N_B is the number of papers published in journal B in the current year

C_A is the number of references in journal A in the current year

10^6 is an arbitrary multiplier to make $R_{A>B}$ easy to interpret

Using the above formula, the relatedness of journal B citing journal A; $R_{A>B}$ is computed. The average relatedness factor between the two journals is computed using the average of the two relatedness values as shown here:

$$R_{A\&B} = \frac{R_{A>B} + R_{B>A}}{2} \quad \textbf{(5)}$$

F. Term Frequency – Inverse Document Frequency (tf-idf)

The tf-idf weight is a measure that identifies how relevant a given term is to a document. In the term-document matrix, the individual cell shows the tf-idf value for a given term and a document [26, 27]. In computing the tf-idf value, importance is given to the term based on its frequency within the document, while at the same time the term is weighted by the number of document in the collection that contain the term. Thus, the tf-idf values help to establish the similarity between documents based on their term occurrence. The tf-idf value is computed as follows,

$$W_{ik} = f_{ik} * \log_2 (N / d_k) + 1 \quad \textbf{(6)}$$

where,

W_{ik} is the tf-idf weight of the term k in document i

f_{ik} is the frequency of term k in document i

N indicates document count for a given dataset

d_k is the number of document with term k

3.2.2. Clustering Approaches

A. K-Means Clustering (partitioning clustering)

K-means clustering is a non-hierarchical approach of clustering [28]. Here K indicates the desired number of clusters. The process starts by partitioning the dataset into K number of clusters. The centroid for each cluster is computed. In successive iterations, items from different clusters are assigned to clusters with the nearest centroid (mean). The centroid of each cluster is re-computed after each round of exchange of items. These item classification iterations continue till no items are left behind to be classified.

The two main weakness of K-means clustering algorithm include: 1) the number of clusters the data needs to be split into must be pre-determined; 2) it is a non-deterministic approach as different initial conditions will yield different results.

B. Wards Clustering (hierarchical clustering)

Wards clustering algorithm is an agglomerative hierarchical clustering approach [29]. Initially a cluster is comprised of just a single node, referred to as a singleton. The euclidean sum of squares (ESS) is calculated for each cluster at every stage. The clusters (sometimes singleton) with the least increase in ESS value are combined together. The ESS value is computed as follows,

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \quad (7)$$

where, x_i is used to represent values of a cluster or a singleton

The entire process is iteratively repeated until all the clusters are combined into a single cluster. The main advantage offered by this algorithm is that different sizes of clusters can be obtained by using different threshold values. But on the other hand, incorrectly grouped objects remain fixed.

3.2.3. Latent Semantic Analysis

The approach taken by the Latent Semantic Analysis (LSA) approach is to match documents based on the concepts within the documents [30, 31]. These concepts are called as latent terms that describe the document. By using the semantic structure of the document, LSA is considered to be designed to handle the so-called vocabulary mismatch problem. LSA handles synonymy (variability in human word choice) and polysemy (one word with different meanings).

In LSA, initially a term-document matrix is constructed. Each element of the matrix indicates the number of occurrences of a term in a document. Using a mathematical

technique called singular value decomposition (SVD), the dimensionality of the original term-document matrix is approximated to a lower dimensional matrix space called SVD matrix. A computationally expensive approach is the construction of SVD matrix. The advantage offered by LSA is that the truncated matrix helps to reduce the noise in the data. Further, the reduced matrix is more accurate in establishing document similarity than semantics based solely on index terms.

3.2.4. Multi-dimensional Scaling

The multi-dimensional scaling (MDS) approach attempts to reduce high dimensional data to low dimensions based on data object distances in space to closely match it with the data similarities and dissimilarities [32, 33]. The resulting representation is based on least-squares between the objects in a low dimensional space, and helps in data interpretation. To move data objects around in the space defined by a given number of dimensions, MDS uses a minimization algorithm that evaluates different configurations with the goal of maximizing the goodness-of-fit. Using this approach, MDS checks how well the distances between objects can be reproduced by the new configuration.

The main weakness of MDS is that there is no quick and fast rule available to interpret the nature of the resulting dimensions. An analyst needs more local details and more explicit representations of structures. An MDS configuration is limited in meeting these needs.

3.2.5. Burst Detection Algorithm

The burst detection algorithm [34] from Kleinberg can be used to analyze streams of time-sorted data to find features that have high intensity over finite/limited durations of time. Rather than using plain frequencies of the occurrences of words, the algorithm employs a probabilistic automaton whose states correspond to the frequencies of individual words. State transitions correspond to points in time around which the frequency of the word changes significantly.

The algorithm is intended to extract meaningful structure from document streams that arrive continuously over time (e.g., emails or news articles). It generates a ranked list of the most significant word bursts in the document stream, together with the intervals of time in which they occurred. This can serve as a means of identifying topics or concepts (bursty words/concepts) that rose to prominence over the course of the stream, were discussed actively for a period of time, and then faded away.

3.2.6. Pathfinder Network Scaling

The Pathfinder Network Scaling (Pfnet) is a structural modeling technique that is used to identify the most meaningful linkages from a complex network data representation [35, 36]. Using this technique preserves the network backbone which helps to communicate the important characteristics of the data. Pfnet can be regarded as a more sophisticated link-reduction technique than the commonly used threshold technique. Pfnet utilizes the node and all its connection information instead of just the node linkage strength information to prune the network.

The Pfnet technique uses the triangle in-equality approach to remove redundant linkages from the network. Pfnet has two parameters 'r' and 'q' that are used to prune the complex network and preserve meaningful linkages. The r-parameter is based on the Minkowski metric approach which uses link weights to compute the distance of the path. The q-parameter is used to preserve the minimum cost path from all the 'q' links that exists between two nodes. For a network with N nodes, 'q' can take a maximum value of N-1. The resulting pruned network data obtained is termed PFNet(r, q) with specific 'r' and 'q' values that were used to get the final results. For an original network with N nodes, the PFNet(r=1, q = N-1) has the least number of linkages.

3.2.7. Betweenness Centrality

Betweenness Centrality indicates the most significant node/edge responsible for the flow of information between all other nodes/edges within the network [37-39]. The betweenness centrality for any given node in a network is calculated as follows,

$$b_i = \frac{\sum_{h\,j=1}^{N} L_{h\,i\,j}}{L_{h\,j}} \quad (8)$$

where,

$L_{h,i,j}$ indicates the total number of shortest paths between two nodes *h* and *j* passing through a particular node *i*.

L_{hj} indicates the total number of shortest paths that exist between node h and j.

The betweenness centrality acts as a measure to determine the important nodes in the network data. The removal of these nodes from the network leads to fragmentation of the network into different sub-clusters.

3.2.8. Degree Distribution

A degree is the number of edges of a node. By calculating degrees for all nodes in a network, one can easily obtain a count of the number of nodes having the same degree count. Normally, a 2D plot of number of nodes v/s node degree is used to visualize the distribution.

The degree distribution can be used as an indicator of network structure. For example, a power law curve indicates that some network nodes have a high degree. These high degree nodes are referred to as 'Hubs'. Existence and usage of hubs helps to speed up the dissemination of information. For example, in case of an epidemic, the most cost-effective solution to quickly curtail and eradicate an epidemic is to isolate or vaccinate hubs instead of random targets [40, 41].

3.2.9. Power Law Exponent

To compute the power law one needs to initially compute the values of degree distribution. A log-log plot of degree distribution results in the straight line that can be expressed in the form of a linear equation (y = mx). The slope of the line (m) is the power law exponent value. The slope m can be +/-. Typically m is negative for a real world dataset because of the small world dataset.

The value of a power law exponent γ plays an important role in explaining the behavior of a network [42]. Small values of γ indicate that hubs are important. Values of γ between 2 and 3 indicate a hierarchical network structure where hubs are in contact with a small number of nodes. Finally $\gamma > 3$, indicate hubs are not relevant.

3.2.10. Term Frequency/Frequency Distribution

This is simple count of the number of occurrence of the entity within the entire document set. The term-document matrix constructed using frequency helps to determine the distribution of the entity across the entire document set [43].

3.2.11. Thresholding

This is a simple technique that can be used to prune the data based on data values. Typically, depending upon the threshold value, the dataset is pruned to only retain values above or below the threshold value.

3.2.12. Spectral Analysis

Similar to the above term-frequency approach, spectral analysis is also a computation of the occurrence of a given entity in a sequential/time-based data.

3.3. Data Analysis Taxonomy

Based on the informal questionnaire feedback, the need for a reference system that classifies different data analysis techniques based on data analysis goals was identified, see section 2.3. The feedback was incorporated as a part of the new developed DA-Vis Taxonomy. It is often seen that data analysis techniques are commonly classified according to the task they help to achieve. Such a classification schema lacks details about how a particular technique helps to achieve the user's goals. For example, clustering techniques like 'k-means clustering' and 'Wards clustering' are both classified under a common category called 'Clustering'. It is also likely that some data analysis methods occur in different categories. So the category label is determined based on the common primary task supported by a set of data analysis methods. Using the data analysis methods covered in section 3.2, the new DA-Vis Taxonomy (for details see chapter 5) provides insight about how different data analysis techniques perform given user tasks and is covered here:

A. Identify Associations: The definition is restricted to the identification of linkages between data entries. Different approaches that can be adopted to identify association between data entries are listed here.

a. **Find Correlation**

   Data Analysis Method: Statistical technique – Pearson Coefficient

b. **Based on Linkage Strength**

   Data Analysis Methods:

   - Co-occurrence similarity
   - Cosine similarity
   - Jaccard index
   - Avg. relatedness factor (co-citation analysis)

c. **Based on Semantic Linkages**

   Data Analysis Methods:

   - Term frequency

- Latent semantic analysis

B. Identify Patterns: The definition of pattern identification is restricted to applying different data analysis techniques that help to show different data characteristics

a. **By Ordering of Non-sequential Data**

   Data Analysis Method:

   - No methods required, if visualization supports ordering
   - Simple sorting

b. **In Sequential Data**

   Data Analysis Methods: Spectral analysis

c. **Using Correlation**

   Data Analysis Methods:

   - Statistical technique – Pearson Coefficient
   - Co-occurrence correlation

d. **By Cluster Identification**

   Data Analysis Methods:

   - k-means algorithm (partition based clustering)
   - agglomerative clustering algorithm – Wards clustering (hierarchy based clustering)

C. Identify Trends: The definition to predict trend is restricted to identify data characteristics by performing some cumulative operations on the data.

a. **By Data Occurrence**

   Data Analysis Method:

   - Frequency distribution
   - Degree distribution

b. **By Data Activity**

Data Analysis Method: Burst detection algorithm

D. Identify Clusters: The clustering definition here is restricted to identification of set of entities with similar characteristics.

a. **By Data Entity Position**

Data Analysis Method:

- LSA
- Co-occurrence similarity
- Cosine similarity
- Jaccard index

b. **By Data Partition (fixed number of clusters)**

Data Analysis Method: k-means clustering algorithm

c. **By Combining Data (hierarchically)**

Data Analysis Method: Wards clustering algorithm

E. Extract Important Data Dimensions/Linkages: The definition here is restricted to the identification of data dimensions that help to describe the data or meet some conditional statements (e.g., threshold).

a. **Based on Data Dimensions**

Data Analysis Method:

- Multi-dimensional scaling
- LSA

b. **Based on Data Linkage Strength**

Data Analysis Method: Apply threshold

c. **Based on Data Connection/Max. Flow**

   Data Analysis Method: Betweenness centrality

d. **Based on Data Meaningful Pathways**

   Data Analysis Method: Pathfinder network scaling (Pfnet)

F. Detect Structural Patterns: This kind of structural pattern detection is applied more to network datasets. Only the important measures that are applied to the network dataset are covered here.

a. **Detect Network Structure**

   Data Analysis Method:

   - Degree distribution – detect network authorities
   - Power law – detect scale free topology

b. **Detect Max. Flow Nodes**

   Data Analysis Method: Betweenness centrality

Chapter 4

Review of Visualization Techniques

This chapter reviews different visualization techniques. Based on data representations, visualization techniques are classified into different types. Different data mappings that are used with different visualization techniques are also covered here [5, 13, 14, 44]. This section also discusses popular data interaction techniques [13, 18]. Researchers have developed taxonomies using some previously reviewed data analysis methods (section 3.2), data visualization techniques (section 4.1-4.2) and interaction techniques (section 4.3). Section 4.4 covers descriptions of prior taxonomies and their utilities. Section 4.5 compares different taxonomies based on their coverage of information design space.

4.1. Visualization Types

The different visualization types that are covered in research literature are described here. This description includes a brief introduction and data mapping details.

4.1.1. Scientific Visualizations

Scientific visualizations are defined as a "branch of computer graphics which is concerned with the presentation of interactive or animated digital images to scientists who interpret potentially huge quantities of laboratory or simulation data or the results from sensors out in the field" [45]. In this visualization, the spatial dimensions are mapped onto

the x-y-z plane. Examples are visualizations of the earth's ozone layer [46], Stratosphere Circulation Simulation [47], Greenhouse Gases & Sulfate Aerosols [47], etc.

4.1.2. Geographic Visualizations

This visualization type makes use of geographical coordinates to map data. Geographical longitude and latitude information is encoded on an x-y-z plane. Geographic maps are an example [3, 48]. Using geographical maps as substrate, data is represented on these maps using points, symbols, etc. Within the information visualization community, geographic based maps are used to show the contribution of various research institutions [49], US zip code distribution [44], roadmaps [50], etc.

Figure 4: US zip-code visualization [44], reproduced with permission from Ben Fry

4.1.3. Information Visualization

Information visualization is a young field which emerged approximately 15 years ago. This field primarily works to find a spatial mapping for abstract data. To find the spatial mapping, the information visualization field draws on ideas from different field like: information graphics for design principles; computer graphics for rendering of graphics; human-computer interaction to identify suitable means to interact with the data; and cognitive science and perceptual psychology to seek guidance on how to visually represent a large dataset so that the information can be better communicated visually. A brief definition of information visualization is 'the use of computer-supported, interactive, visual representations of abstract data to amplify cognition' [12].

4.2. Information Visualization Layouts and Algorithms

4.2.1. 2-D Plots

These plots are used to display and compare two sets of quantitative or numerical data by plotting two data dimensions along two orthogonal axes. Data is mapped in 2D plots using either points, bars or lines [50]. Additional data attributes can be displayed in the form of color, line width or point shapes. Examples are scatter plots, column graphs and line-graphs.

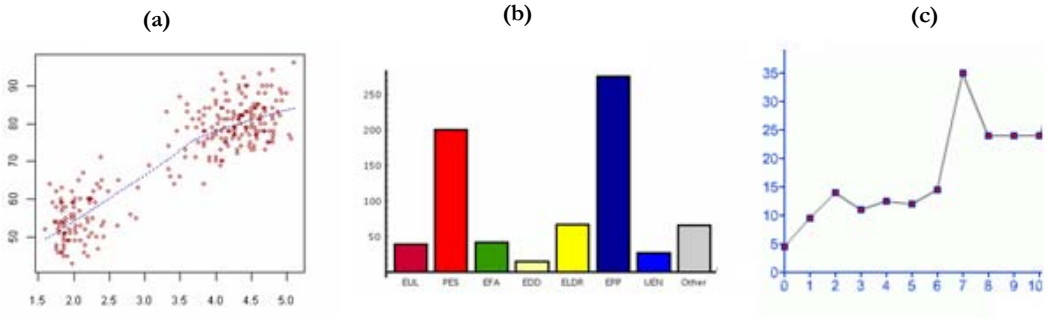


Figure 5: Examples of 2D plot. (a) scatter-plot, (b) column plot and (c) line graph

4.2.2. Iconic Displays

Visualizations map multiple data dimensions to features of an icon. Example are: chernoff faces [51], star icons, sequential icons, polygonal icons, pie icons, sun rays icon, etc. [13, 52]. For data mapping, a separate "face" icon is drawn for each data entry; relative values of the selected variables for each data entry are assigned to shapes and sizes of individual facial features.

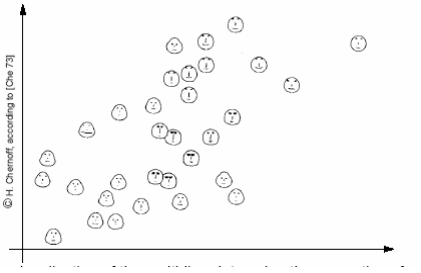


Figure 6: Iconic display of faces showing different characteristics, adopted from [52].

4.2.3. Multi-Dimensional Scatter-Plot Visualizations

This visualization type supports the mapping of multiple non-spatially related variables onto the same X-Y plane. Different filtering mechanism can be applied to parse different sets of information. Color and text formats can be used to encode different data characteristics. E.g.: FilmFinder [53].

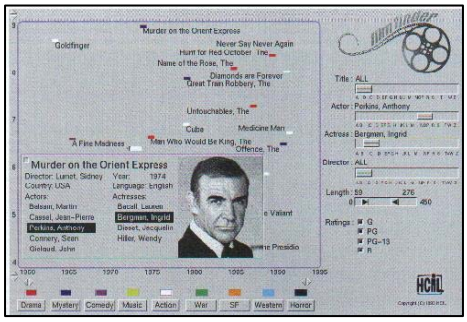


Figure 7: Multi-dimensional scatter plot usage for FilmFinder [53], reproduced with permission from Univ. of Maryland, Human-Computer Interaction

4.2.4. Geometrically Transformed Displays

Visualizations that involve the projection of all data dimensions on a 2D plane are included under this category [13]. In parallel coordinates visualization [54], each data dimension is represented by an axis. The data is shown by the lines that intersect different axes based on its variable value along the axis. This representation is unique as it replaces the orthogonal coordinate system to n-dimensions.

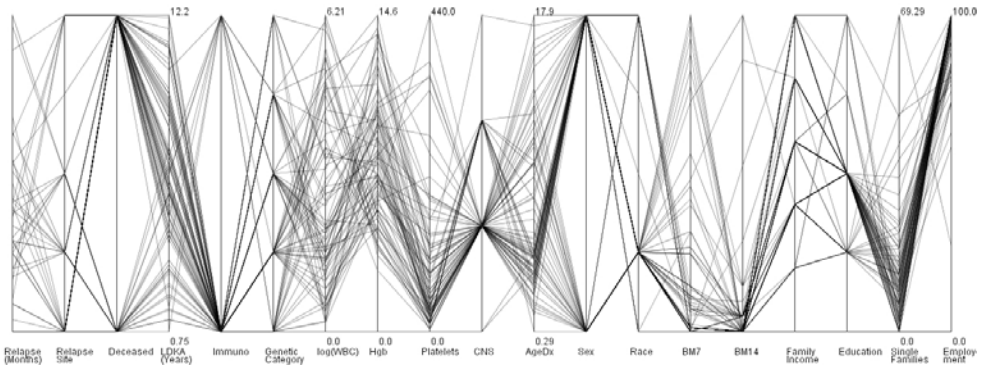


Figure 8: Parallel coordinates images

4.2.5. Special Data Transformation: Text Visualizations

Visualization that support the direct representation of textual data fall under this type. These kinds of visualizations map text directly into the XY plane. Data highlighting and text size variation are used to show the important characteristics of the data. For example, TextArc [55], SeeSoft [56], etc.

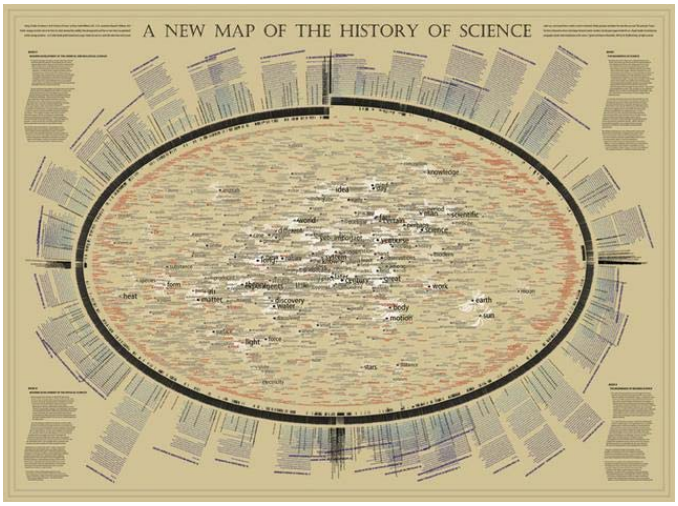


Figure 9: Text visualization example [55], reproduced with permission from Bradford Paley

4.2.6. Dense Pixel Display Visualizations

Visualizations map each data dimension to a color pixel and provide a means to rearrange data points based on data dimensions. This view can also be used to show time-dependent data. In Figure 10, concentric circles show different time periods and hence the evolution of different attributes over time [57]. Some example are: Recursive patterns [13], Circle segments techniques [13], Circleview [57], etc.

Figure 10: Circle segment example to show evolution of attributes overtime [57] © 2004 ACM Inc.

4.2.7. Multi-Dimensional Tables/Matrix Visualizations

Data presented in tabular format or matrix format (rows and columns) fall under this type. Matrix display simplifies the analysis of multidimensional data by enabling the user to study multiple data characteristics at a single glance. Each single matrix cell is used to show different data. Color can be added as a visual cue to show matrix cell data. This kind of visual encoding helps to show patterns in the data. TableLens [58] is an example.

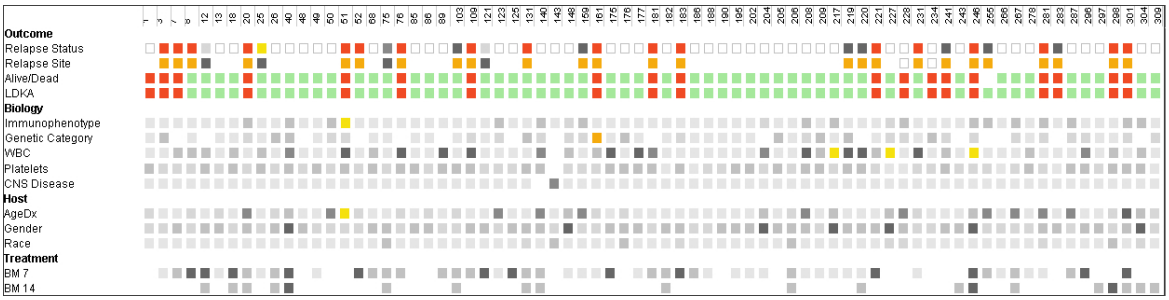
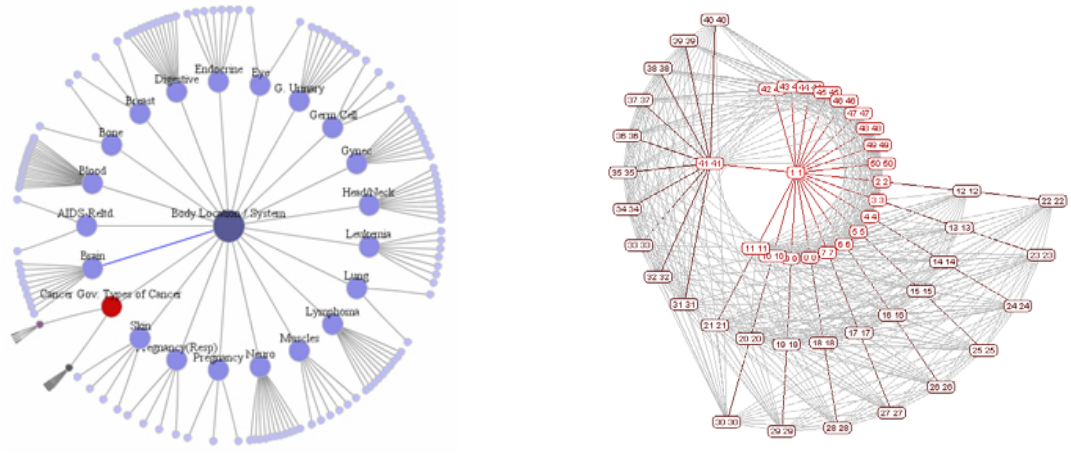


Figure 11: Matrix example with cell color showing data characteristics

4.2.8. Node and Link Diagrams/Networks

Representation of data objects as nodes and relations between data objects by edges is done in node and link based visualizations. These visualizations are ideal to show the linkage information between the data objects/entities. Visual data mapping for nodes is in the form of color, size, and shape. Line data mapping is done in the form of thickness and texture. Hierarchical and non-hierarchical data display is supported using node and link based representations. Examples are hyperbolic tree [59] and radial tree [60, 61] for hierarchical data, and network diagrams for non-hierarchical data [42, 62-66].



(a) (b)

Figure 12: Examples of node and link diagram (a) shows hierarchical data as radial tree and (b) shows a non-hierarchial data as network

4.2.9. Dendrogram Layout

A dendrogram layout (see Figure 13) is an alternate representation of the node and link diagram. This arrangement is useful to show clusters identified by a clustering algorithm. Within a dendrogram layout, data objects are connected using U-shaped lines to form a hierarchical tree. The height of each U line is representative of the distance between the two objects. Dendrogram offers the advantage of representing nested clusters. Additionally, there is no need to pre-define the number of clusters [67] .

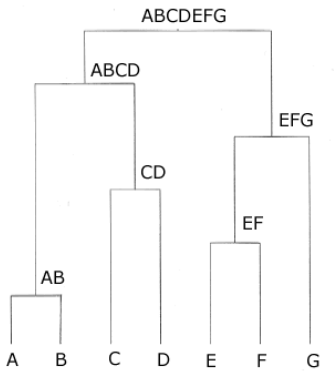


Figure 13: Dendrogram layout to show hierarchical clustering of data objects

4.2.10. Information Landscapes

With this type of visualization, data is typically laid out on a surface (XY plane) using certain data properties like similarity between data objects, etc. This type of visualization uses force-directed algorithms to create visualization instead of geographical longitude and latitude information which is used in geographic visualization (see 4.1.B). Typically color is used to show the different data characteristics. E.g.: VxInsight [68, 69].

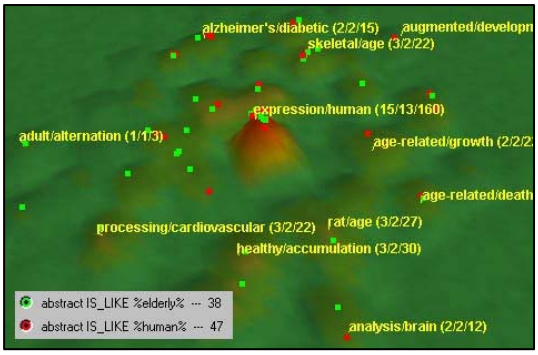


Figure 14: Information landscape layout showing the structure of aging research, adopted from[70]

4.2.11. Stacked Display Visualizations

This category of visualizations can also be used to present hierarchical data as partitioned data. Different data categories are represented as nested rectangles. Data characteristics can be shown using colors. Treemaps [71] are an example.

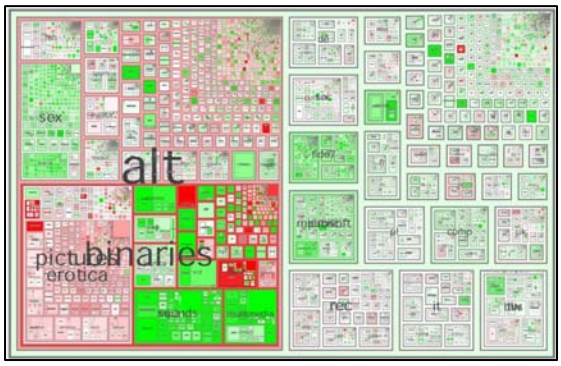


Figure 15: Treemap showing social cyberspace in Netscan Project [72], reproduced with permission from Marc Smith

4.2.12. Visualization spreadsheets

Visualizations of this kind help to build multiple visual representations of different dimensions of one or more datasets. As multiple views are available, this type of visualization is ideal for exploratory purposes. Ed Chi's visualization spreadsheet [14, 73] is an example.

Multi-dimensional datasets can be visualized using some of the above methods. For network visualizations, the major layout algorithm used are covered below.

4.2.13. Force-Directed Spring Embedded Algorithms

Visualizations should be aesthetically pleasing and help in meaningful data interpretation. Several algorithms have been developed to generate visualizations. To generate optimum layout, these algorithms follow some basic aesthetic guidelines like: reduction in edges cross-over; uniform distribution of nodes and edges; uniform edge lengths; minimizing bend edges and symmetrical display of a graph [74, 75]. Based on the layout methodology, algorithms are classified into different categories. Major tree layouts, linear, force-directed, and planar algorithms are discussed in my Ph.D. proposal [1] and Ph.D. qualifying exam document [2]. The force directed algorithm is discussed below as it is a commonly used network layout algorithm, and used in the DA-Vis taxonomy (see chapter 5).

Eades proposed the original idea to consider a network as physical system of mass particles (nodes) connected by spring edges [76]. The algorithm's goal is to reposition nodes in order to reduce the overall spring force, Sp (distance separation of nodes) for the entire system. An algorithm developed using this approach is referred to as a force directed

[1] Ph.D. proposal can be accessed online at:
http://ella.slis.indiana.edu/~kmane/proposal/proposal_ketan-mane.pdf

[2] Ph.D. Qualifying exam document can be accessed online at:
http://ella.slis.indiana.edu/~kmane/quals/quals-ketan-mane.pdf

placement algorithm. Force variations include gravity force (GF), electric (EF) and magnetic force (MF) [74]. Formulae for the computation of different forces are:

$Sp = k(1-l)$, where k is the stiffness of the spring and l is the natural length of the spring.

$GF = \frac{g}{r^2}$, where g is associated with the mass of the particle, usually equals 1 and r is the distance between the two objects.

$EF = eE$, where E is the electric field strength and e is a constant.

$MF = qB$, where B is the magnetic field strength and q is a constant.

2D and 3D network layouts can be generated using this algorithm, see Figure 16 [77].

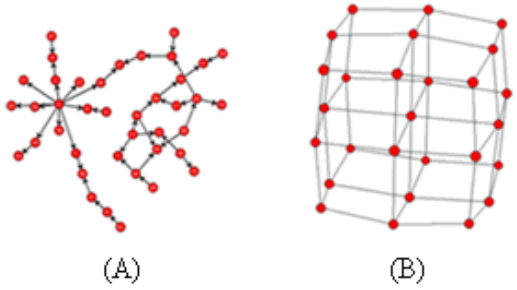


Figure 16: Sample layouts for spring embedded algorithm, adopted from [78]

However, Eades version of the algorithm ignores Hooke's law which states that "the power of any springy body is directly proportion to its extension". This law is used to calculate spring energy. Further it stricts the energy computation to its neighbors instead of the entire graph. Fruchterman-Rheingold [79] and Kamada Kawaii [80] are improved versions of this algorithm.

Kamada Kawaii uses the least square method under a set of conflicting constraints while Fruchterman-Rheingold considers the system of interacting particles with repelling forces

[81]. Best layout is obtained for undirected sparse, tree-like graphs [82] and symmetric graphs [83]. These algorithms take multiple iterations to converge.

4.3. Visual Mapping

Visual mapping deals with the perception principles. To cover these details is beyond the scope of this thesis. For detailed explanation on perception principles, one could refer to either of the following references: Colin Ware's book titled 'Information Visualization' [84], and Stephen Palmer's book titled 'Vision Science' [85].

4.4. Interaction Techniques

Data interaction techniques are applied to the visualization to show different data characteristics. One of the most popular interaction techniques is the 'information seeking mantra' from Ben Shneiderman which includes – overview, zoom and details-on-demand. Different techniques of interaction are shown below [18].

4.4.1. Zooming

Using this technique, a geometric scaling of the visualization is achieved. The zoom operation performed by the user helps to gather more information about an area of interest. For large datasets, the zoom feature helps to take a closer look at regions with dense information overlay. This kind of zooming is referred to as 'geometric zooming'. Zooming where additional details related to the region are progressively revealed as one zooms in is called 'semantic zooming'.

4.4.2. Details-On-Demand

This feature helps to obtain additional information about an item of interest from the dataset.

4.4.3. Interactive Filtering

Using filtering mechanisms, all the interesting items are selected from the dataset. This technique generates dynamic queries that can be applied to items within a visualization to filter out uninteresting parts. A technique called 'direct manipulation' is useful in rapidly filtering out and maintaining the data of interest in the view [86].

4.3.4. Brushing and Linking

This interaction technique supports data selection in one view and highlighting of related data in the other view [87]. Multiple coordinated views are supported using this technique. This technique shows the relationships among the items [53].

4.3.5. Distortion

This technique supports focus-and-context within visualizations. Global as well as local views of the data are maintained at the same time [58]. Examples of distortion techniques that supports local zooming while preserving the global context include: perspective wall [88], bifocal display [89], graphical fish-eye [60]. etc.

4.5. Existing Information Visualization Taxonomies

This section reviews earlier attempts to map the structure of the information visualization design space by means of taxonomy. Taxonomy helps to organize the information, so that it can be better understood. Once again, it is worth mentioning that multiple definitions of taxonomy exist, however the word 'taxonomy' used in this paper refers to the classification of related techniques into a common category. Same techniques can be classified under two different categories, if they satisfy the requirements of the category.

Prior taxonomies categorized information visualization techniques at different levels such as: 1) data type representation level, 2) user task and data type level 3) visualization technique level, 4) processing step level to obtain visualization and 5) the level of visual data analysis.

Using the terminology introduced in section 2, details of the taxonomies are discussed below. Techniques are categorized based on the demands of the tasks they best serve and the data type they operate on.

4.5.1. Taxonomy of Data-Type Representation by Bertin

This early work by Bertin is based on data value and structure [90, 91]. It identifies two forms of data: a) data value – which is data driven and specifically addresses a single issue at the local level, and b) data structure – which is more conceptually driven and helps to address the issue at the global level. Representation details for each form of data are covered here.

A. Data value representation: In this representation, the data attributes are mapped onto the available axes. For example, in a scatter-plot, two different data attributes define the axes. The scatter-plot shows a local data relation at two attribute levels. Hence data value representation is also referred to as the 'low level view' of the dataset. Data value representation is used for representations of quantitative data type information.

B. Data structure representation: The representation of existing relations between different entities of the dataset reveals the structure of the dataset.

Figure 17 shows five types of data representation which provide a global overview of the dataset. In all these views, nodes represent data entities and lines show the relations between data entities.

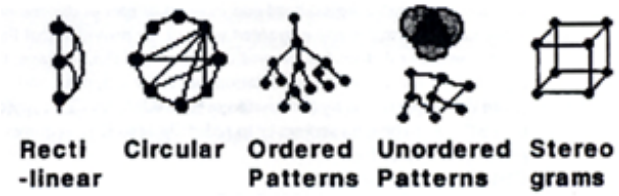


Figure 17: Five types of structural representation [91] © 1997 AMC, Inc.

Distinct advantages are offered by different forms of data structure representation. 'Rectilinear' view shows the order as well as relations within data entities. In 'Circular' view, the circular representation offers a compact view of the data at the cost of losing the data order information. The 'Ordered Patterns' view shows the hierarchical structure of the dataset. This view also shows the parent-child relation between the data. The 'Unordered Patterns' view is a variant of the circular view. In this view, data entries are not arranged on the circumference of the circle but are dispersed in the available space for display. The advantage offered by this view is the ability to cluster similar data entries together. All the above views are available to generate 2D data representations. But the last view, 'Stereographs' can be used to draw 3D data representations. This taxonomy was adopted as a framework by other researchers to classify different types of data visualization, see section 4.1.

4.5.2. Taxonomy of Task by Data-Type by Ben Shneiderman

The Task by Data-Type taxonomy by Ben Shneiderman tries to establish a mapping between different data types and user tasks [18, 92]. The taxonomy approach is guided by visual design principles of overview first, followed by zoom and filter, and finally details on demand. This visual design principle is commonly known as the 'Information Seeking Mantra' [18]. Shneiderman's taxonomy is particularly useful when designing user-interface functionality to prune data based on data-type information.

Information seeking is a process of selecting items that satisfy values within a given set of attributes. These attributes strongly reflect data type information. There are seven attribute data type categories: 1-, 2-, 3-dimensional data, temporal data, n-dimensional data (n-D), trees and network data. Detailed information for each data type was covered earlier in section 3.1.

User tasks depend upon the data type information available within the dataset. User tasks at a higher level of abstraction are divided into seven different categories: 1) Overview, 2)

Zoom, 3) Filter, 4) Details-on-Demand (D-o-D), 5) Relate, 6) History and 7) Extract. These user tasks are a subset of data interaction techniques covered in section 4.3.

A data overview helps to maintain a global overview of the dataset. It also helps to identify the contextual relations between different entities. A detail-on-demand in global view requires a local data zooming feature. Distortion techniques (section 4) can be used for this purpose. An data overview helps to accomplish data driven user tasks such as: item counts (1-dimensional data), number of adjacent items (2-dimensional data), identification of data positions (3-dimensional data), identification of correlation (n-dimensional data), identification of gaps (n-dimensional data), identification of outliers (n-dimensional data), traversing paths (network data) and before and after events (temporal data) [3]. The zooming technique (section 4) is particularly useful where the density of information displayed is high. This data driven user task helps to identify the node-link relations (network data) and eliminate occlusions (3-dimensional data). Filtering tasks applicable to different data types are: identification of items with certain attributes (1-dimensional data), identification of patterns (n-dimensional data) and clusters (n-dimensional data). The details-on-demand technique becomes applicable at the local as well as the global level. At the local level, details-on-demand provides all the information about a single entity. But at global level, it helps in the identification of items with all their attributes (1-dimensional data and n-dimensional data), identification of structural properties like: structure level information (trees) and number of child nodes for a given parent node (1-dimensional data). The relate technique helps to fulfill user tasks such as cluster identification (n-dimensional data), containment of one item within another (2-dimensional data), node and link relations (tree and network).

This taxonomy also includes some non-data interaction techniques as user tasks. The user action tracking called 'History' can be used for undo, replay, and for progressive refinement of data. This is a general purpose task that can be used to retrace the path taken. Logical operators in combination with recorded history steps can be used as data filters. Another

[3] Temporal variables are treated different from 1-dimensional data as they have a start time and an end time.

user task of data extraction provides a feature to save data. This task can be used to extract a subset of data when large datasets are involved. Recursive use of this strategy on the data of interest can help to reduce the data to a manageable size.

4.5.3. Taxonomy of Information Visualization Techniques by Stuart Card and Jock Mackinlay

This work by Stuart Card and Jock Mackinlay extend Bertin's earlier taxonomy of data representation (section 4.5.1). This information visualization taxonomy [5] helps to group different visualization techniques (section 4.1) based on the similarity of the underlying data. Further, it highlights different data mappings used in different visualization techniques to make sense of the data. The utility of this taxonomy is seen when it comes to understanding the data type support offered by each visualization technique.

Visualizations are generated by mapping data in the form of marks (such as points, lines, areas, surfaces or volumes or positions in space and time). These marks are utilized for encoding data characteristics by varying their size, position, color gradient, connections, and proximity. Data encodings in the form of marks are automatically processed by our visual channels. The other form of data encoding uses text. These textual representations undergo controlled processing by our visual channels.

Based on the data type information, marks and text are used to represent data. This taxonomy compares different techniques that are covered in section 4.1. From the taxonomy, it can be seen that all visualization techniques require some pre-processing stage to format the data that has to be visualized. The taxonomy shows that most of the visualization techniques differ, it can be clearly seen in the usage of marks and encoding formats. For example, color is used to show different types of information in multi-dimensional scatter-plots. On the other hand, color is used to show association information in the node-link diagram.

4.5.4. Taxonomy of Visualization Processing Steps by Ed Chi

Ed Chi's taxonomy is based on processing operations required by different visualization techniques [93]. The basis of this approach is the 'Data State Model' that describes the process required for generating different visualizations [14, 94]. This taxonomy helps to identify the common processing steps involved in generating different types of visualizations. It is a good reference to identify common data pre-processing steps required for multiple visualizations.

Within the 'Data State Model', the process of generating any visualization technique is split into four data stages (value, analytical abstraction, visualization abstraction, and view), three data transformation stages (data transformation, visualization transformation, and visual mapping transformation) and four types of "within" stage operators (within value, within analytical abstraction, within visualization abstraction and within view). Figure 18 shows the order of the processing steps. The data state model provides information about data transformation that takes place from raw data to visualization. But it does not cover any information about the data analysis techniques that are involved during any of the data transformation process.

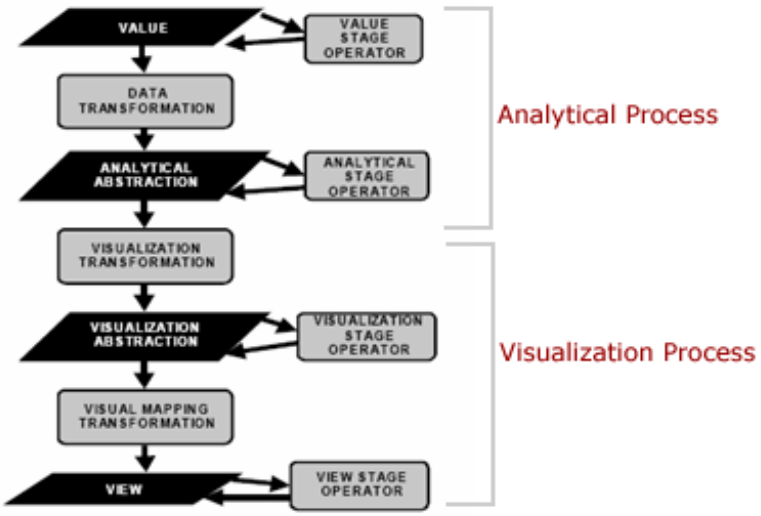


Figure 18: Data state model in information visualization [14], reproduced with permission from Ed Chi

The work in section 4.5.3 is used as a foundation for this taxonomy. For each visualization category, the process starts with raw data as input to the system. Based on the data state model for each visualization category, data processing at each stage in the pipeline is constructed. To facilitate understanding, different stages of the data state model are combined into two main processing stages: 'analytical process' and 'visualization process' as indicated in Figure 18.

For example, to generate a node-diagram (network) visualization, the analytical process is comprised of extracting node information, linkage information from the original dataset and transforming the data into graph or network format. At the visualization technique level, this taxonomy covers details about representations adopted by different systems. The visualization process information is very specific to a visualization application that is used to represent the data. For example, the network display, GraphViz [95] uses different layout algorithms to position data entries in the visualization. On the other hand, the SeeNet [96] uses a matrix data representation format. Explanation at analytical and visualization process level is available for different visualization techniques (section 4.1).

4.5.5. Taxonomy of Visual Data Analysis Techniques by Daniel Keim

The most comprehensive taxonomy was introduced by Daniel Keim. It covers details about different data types, visualizations, and interaction techniques [13, 97]. Figure 19 shows different options available under each category along three orthogonal axes. This taxonomy provides a good overview of information visualization techniques.

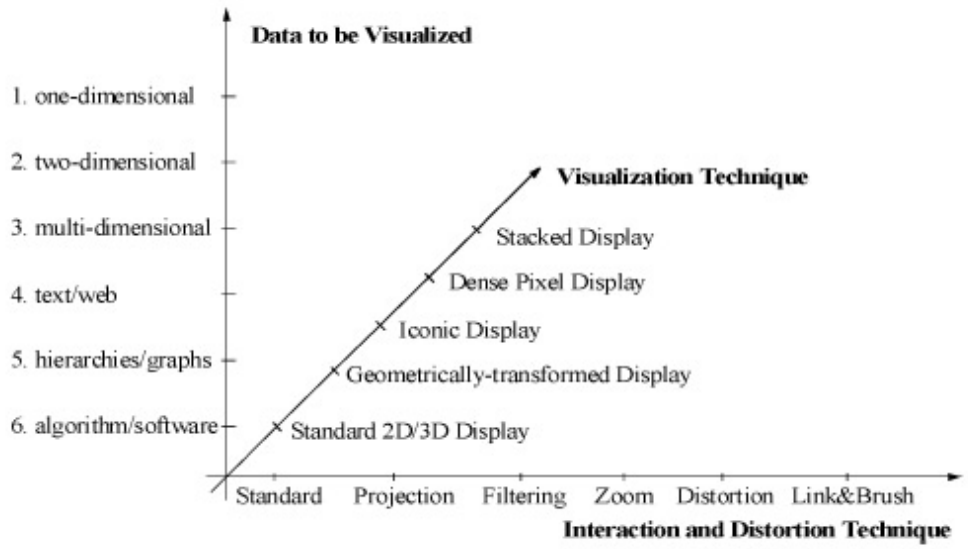


Figure 19: Classification of information visualization techniques, adopted from [97] © 2002 IEEE

Available data types in visualization include: 1-dimensional, 2-dimensional, multi-dimensional, text and hypertext. Classification of temporal data as 1-dimensional data is the only difference introduced by this taxonomy as compared to earlier ones covered in this section [18].

A different set of visualization techniques is available through this taxonomy. Visualization category name and different techniques classified under it are as follow: standard 2D/3D, geometrically transformed displays, iconic displays, dense pixel displays and stacked displays. Details of the different visualization techniques mentioned above were covered earlier in section 4.1.

The data interaction and dynamically changing visualizations are based on user tasks. Different interaction techniques are: dynamic projections, interactive filtering, zooming, distortion, brushing and linking. Based on exploration support, categories of techniques are discussed below.

A. Navigation techniques: This technique deals with the change of data projection on a display screen. The ability to view data from different perspectives makes this technique

ideal for multi-dimensional datasets. Dynamic projection, a popular navigation technique, has been adopted in systems like XGobi [98].

B. View enhancement techniques: Exposure of data details and emphasis of data subsets can be done using view enhancement techniques. Zooming is the most popular view enhancement technique, e.g.: TableLens [58]. Distortion technique is an alternate view enhancement technique e.g., Fisheye View [99]. View enhancement leads to the exposure of relevant information. Hence this technique can also be used in interactive data filtering operations.

C. Selection techniques: Highlighting, filtering and quantitative analysis are options available through selection techniques. Interactive filtering process is a part of this technique. Brushing and linking techniques are also a part of this technique.

4.6. Discussion

This section compares existing taxonomies. It includes a discussion about their major contributions and omissions. Figure 20 shows that information space can be split into 5 distinct regions: data dimensions (section 3.1.C), user tasks (section 2.3), data analysis techniques (section 3.2), data visualization techniques (section 4.1), and data interaction techniques (section 4.4). The rectangular colored boxes show areas covered by different taxonomies and the section number in which they are briefly discussed.

Figure 20: Coverage of previously established taxonomy

The early taxonomy from Bertin (section 4.5.1) targeted the data visualization region at a very fundamental level. It covered details about different forms of data representation based on the data type information. The taxonomy also offers a basic framework to classify different kinds of data representation. It further highlights distinct advantages offered by different forms of visualization. But it lacks detailed classification for different kinds of visualization techniques.

The taxonomy from Stuart Card and Jock Mackinlay (section 4.5.2) extends the work of Bertin by categorizing different visualization techniques into sub-categories based on the data type they represent best. This categorization helps to easily identify different forms of representation that can be used to show data characteristics. Although it shows a wide range of visualization, it does not communicate any information related to what data transformation is required to get a particular type of visualization.

The more recent taxonomy by Ed Chi (section 4.5.4) extends the taxonomy developed by Stuart Card and Jock Mackinlay to include details about processing steps needed to generate different kinds of visualizations. The taxonomy offers an advantage in identifying common steps required to generate visualizations. This taxonomy comes in handy when identifying what visualizations can be generated using common data processing step.

The taxonomy from Ben Shneiderman (section 4.5.3) targets the user task area of information space. But these tasks are more at a data interaction level. The interactions can be mapped to data analysis goals to show what different user tasks can be accomplished. But it fails to account for different visualization layout and data analysis techniques that can be used for data representation.

The most current taxonomy for visual data analysis from Daniel Keim (section 4.5.5) is the most comprehensive taxonomy. It covers more details about the latest visualizations and interaction techniques. It extends the taxonomy by Stuart Card and Jock Mackinlay (section 4.5.2) to include new types of visualization techniques that are not covered in the previous taxonomies.

Figure 20 shows the coverage areas of prior taxonomies. It can be seen that none of the prior taxonomies covered information about data analysis techniques. No existing taxonomy links data analysis techniques to data visualization techniques. Further, bridging analysis and visualization techniques is only useful if it matches the user requirements. The DA-Vis taxonomy proposed by current research fills in this void. It shows a taxonomy of complementary data analysis and data visualization techniques that can be used together to meet the user task requirements. Details of the DA-Vis taxonomy are covered in chapter 5.

Chapter 5

Coupled Data Analysis and Visualization Taxonomy (DA-Vis)

This chapter presents the centerpiece of this thesis: DA-Vis Taxonomy interlinking analysis and visualization techniques. Given the scope of this work, see Figure 1, the DA-Vis taxonomy will be restricted to commonly used techniques of analysis and visualization in n-dimensional and network datasets.

The visual layout schema adopted for the new coupled taxonomy is shown in Figure 21. Within the schema, the top level (level 1) includes information about the primary goals. Level 2 includes information on specific data analysis tasks to be undertaken. Based on the available data type information, level 3 indicates different techniques for data analysis and visualization. Within each specific task (level 2), data analysis techniques are arranged based on the complexity involved in performing a given data analysis technique. The techniques are shown under the 'Data Analysis Methods' column. Also given at level 3 under the 'Visualization Layout' column are different visualization techniques which can be used to represent results from data analysis. Further, based on the selected visualization technique, different visual mappings become appropriate. These visual mappings are shown under the 'Visual Mapping' column.

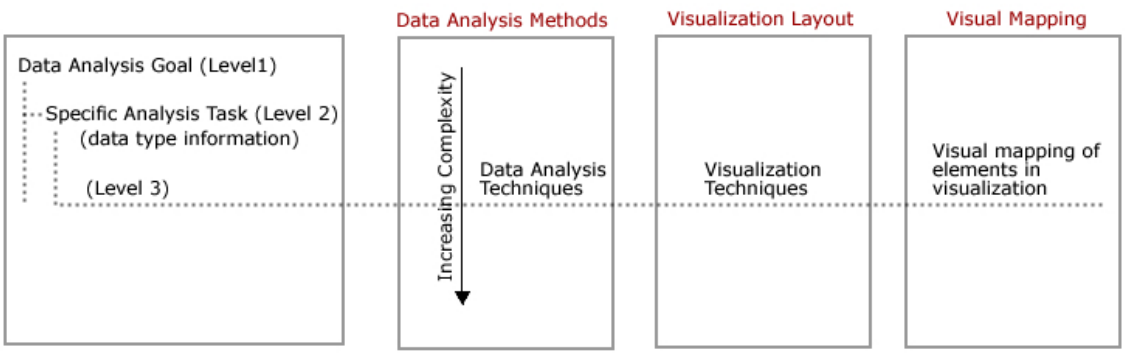


Figure 21: New coupled DA-Vis taxonomy layout schema

Within the DA-Vis taxonomy, a pathway is defined as 'a choice of data analysis goals (level 1), specific data analysis tasks (level 2), and identified data analysis methods, visualization layouts, and visualization mappings that can be used to accomplish the user task'.

The visual representation language in Figure 21 is used to generate a taxonomy for different data analysis goals such as: identifying associations (Figure 22), identifying patterns (Figure 23), identifying trends (Figure 24), identifying clusters (Figure 25), extract important data dimensions/linkages (Figure 26), and detect structural patterns (Figure 27). All these data analysis goals have been covered in section 3.3.

A user's task is specific. For example, the task could be to find the correlations between two data objects A and B. This task falls under a broader category called 'identifying associations' which is one of the primary tasks of data analysis (section 3). The new coupled taxonomy embeds information at both primary task level and specific task level. Availability of this information helps the user to quickly identify the primary task as well as to acquire a list of data analysis techniques that can be used to obtain results. Based on the data analysis goal, the DA-Vis coupled taxonomy classifies different data analysis techniques that can be applied to both n-dimensional and network data types.

5.1. Identifying Associations

Figure 22 shows the taxonomy to 'identify associations'. The diagram gives one a quick overview of the available data analysis techniques. The diagram also includes data type level information that would be supported by a given analysis technique. For example, Figure 22 shows that associations in n-dimensional data can be determined using any of the three specific tasks: 1) finding data correlations, 2) computing linkages and 3) establishing semantic linkages. In turn, each of the specific tasks show available data analysis techniques. Data analysis techniques that help to determine co-citations are categorized under 'compute linkage strength' because their metric is based on the number of paper citations (linkages) for a given paper. However, for network data, linkage information is available. Hence no special data analysis technique is required to compute linkage information.

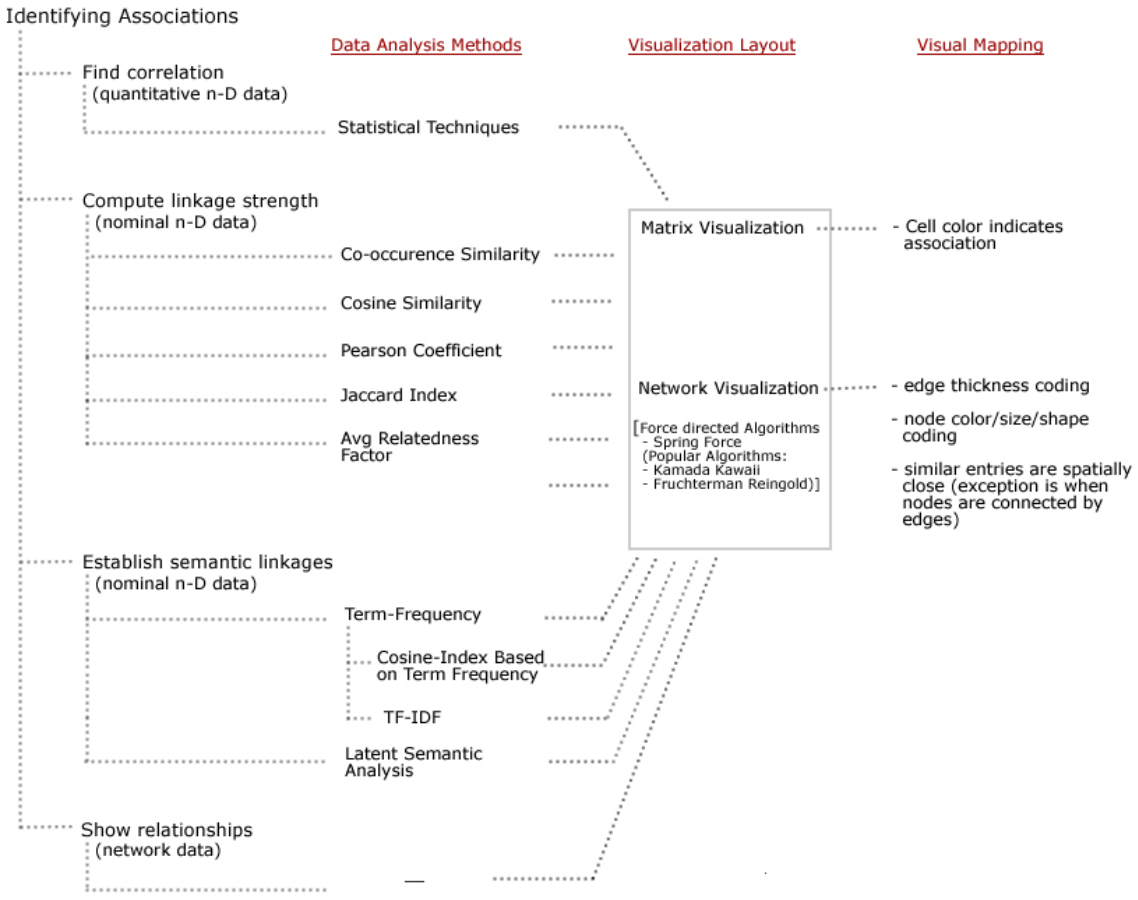


Figure 22: Data analysis and visualization coupled taxonomy to identify data association

Further, Figure 22 also shows a tight coupling between the data analysis techniques and data visualization techniques. Only those visualization techniques that can be used to generate meaningful visualizations from the suggested data analysis techniques are covered. For n-dimensional data, many-to-many relationships exist among data entries. Among the visualization techniques, matrix and network based visualization techniques can be used to show associations among n-dimensional data. Depending upon the choice of the visualization technique, different visual mapping properties are deployed to communicate association information. For example, in matrix visualizations, associations among data entities are often indicated by cell color. In network visualizations, the proximity of points in the layout and the connections among data entries communicate basic associations. Information related to the strength of association can be encoded using other visual mapping as shown in Figure 22.

5.2. Identifying Patterns

Figure 23 shows the sub-taxonomy to 'Identifying Patterns'. Different user tasks that support the identification of patterns include: data ordering, pruning time-series (sequential) data, using data correlation, and clustering. Corresponding data analysis tasks along with complementary visualization techniques are shown under respective columns.

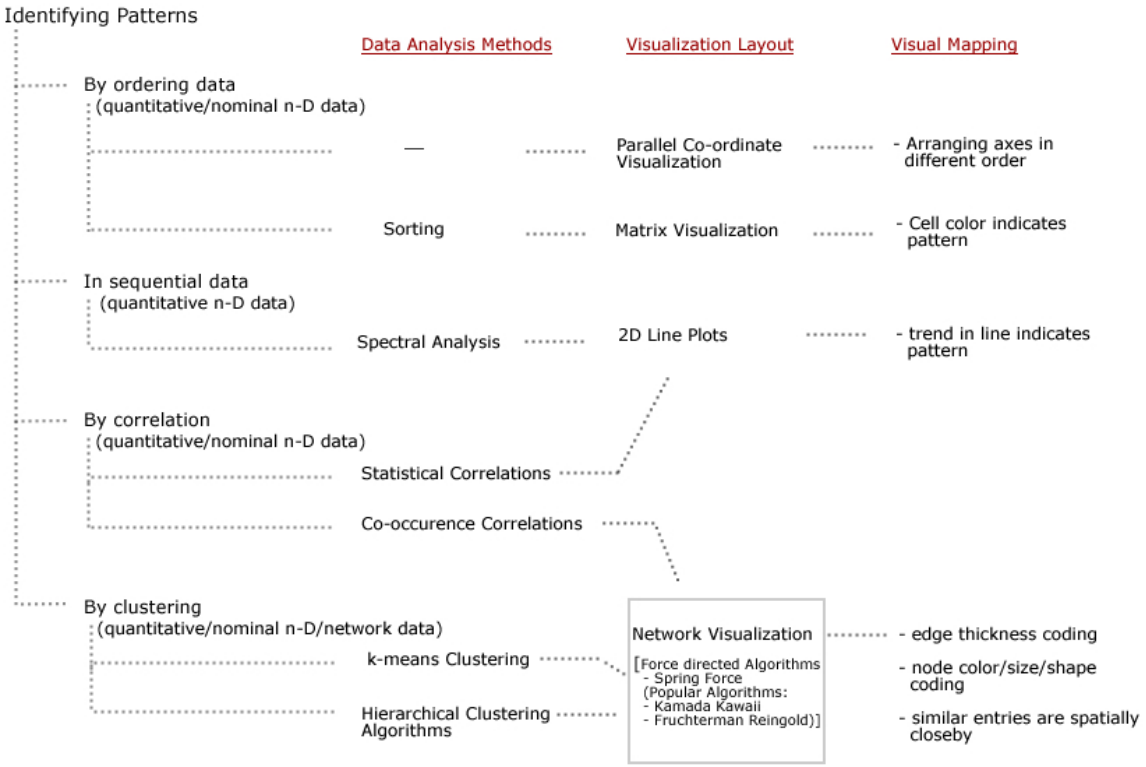


Figure 23: Data analysis and visualization coupled taxonomy to identify patterns

5.3. Identifying Trends

Figure 24 shows the taxonomy to 'Identifying Trends'. Depending upon the data type, different specific user tasks are shown in the taxonomy. The data trends are identified based on their occurrence, data selection, and data activity. Depending upon the user task, different data analysis methods can be used. Based on the results of the data analysis, complementary visualization techniques can be used. For example, to visualize the results of the burst detection algorithm, one can use a 2D histogram. But, if the burst time period has to be visualized then, 2D floating bars become an ideal visualization option to show patterns of different burst activity for an identified burst item.

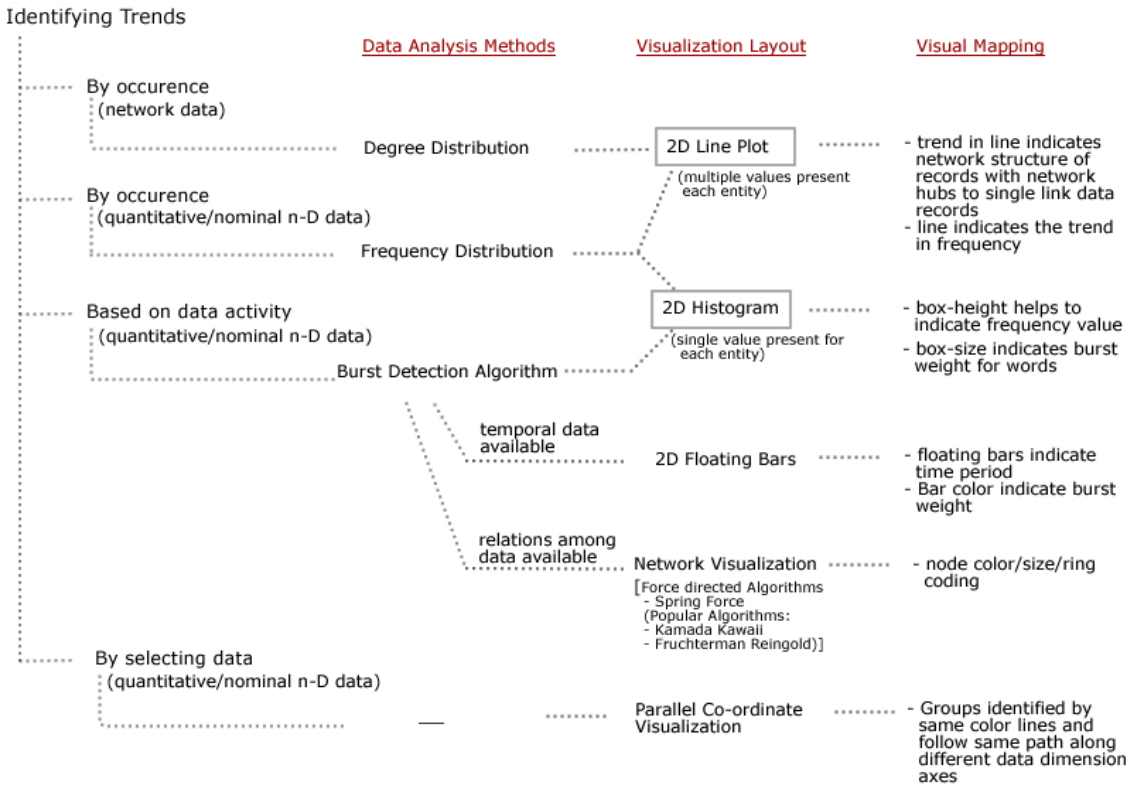


Figure 24: Data analysis and visualization coupled taxonomy to identify trends

5.4. Identify Clusters

Figure 25 shows the taxonomy to 'Identify Clusters'. Clusters in the dataset can be identified based on the data entity position, by data partition, and data hierarchy. All identified user tasks make use of nominal n-dimensional (n-D) data. To identify clusters based on the data entity position, a similarity measure needs to be computed among data entities of a dataset. The data analysis method column indicates different data analysis methods available to compute similarity strength. Further, a force directed layout algorithm can be used to position similar data entities in close vicinity of each other to be identified as a cluster.

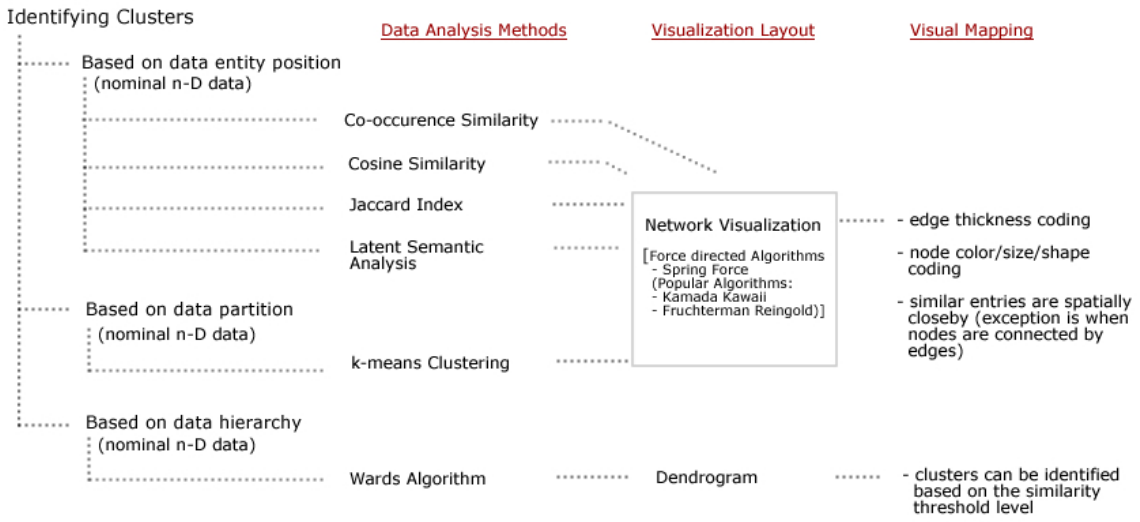


Figure 25: Data analysis and visualization coupled taxonomy to identify clusters

5.5. Extract Important Data Dimensions/Linkages

Figure 26 identifies dimensionality reduction techniques to capture important data dimensions during data analysis operations on nominal n-dimensional data. For network data, threshold operation can be performed only when linkage strength information is available. If linkage strength information is not available then, techniques like betweenness centrality, and pathfinder network scaling can be used to identify important data linkages in a network.

5.6. Detect Structural Patterns

Figure 27 shows the taxonomy to 'detect structural patterns' in network data. Common specific user tasks are identified. Based on the user task, an appropriate data analysis technique can be identified in the data analysis method column. Complementary visualization techniques are given in the visualization layout column.

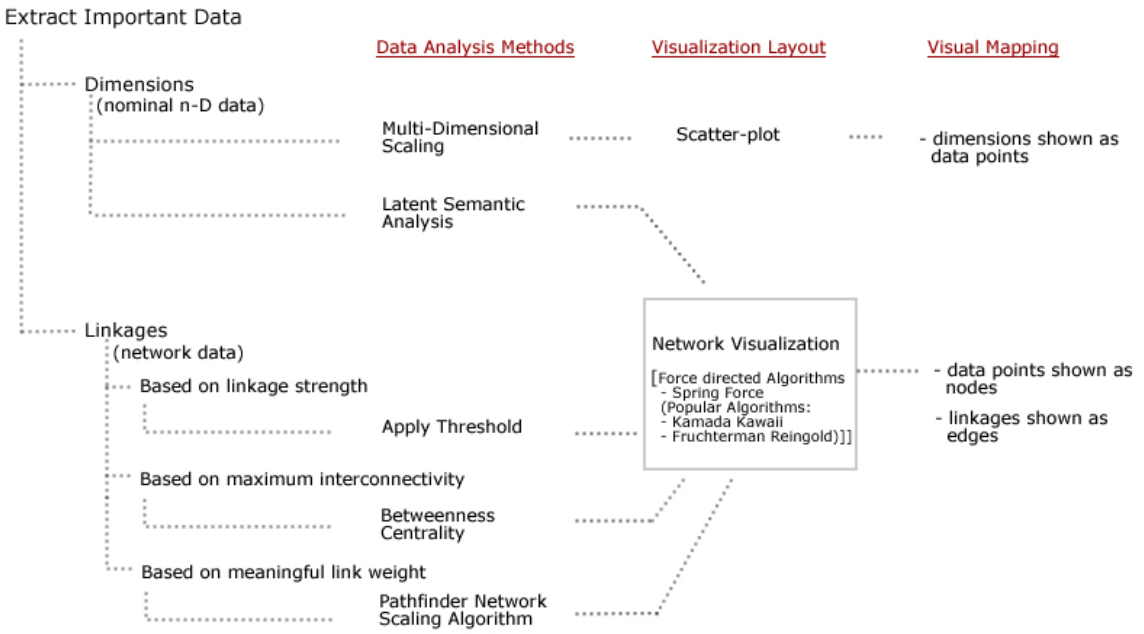


Figure 26: Data analysis and visualization coupled taxonomy to extract important data dimensions/linkages

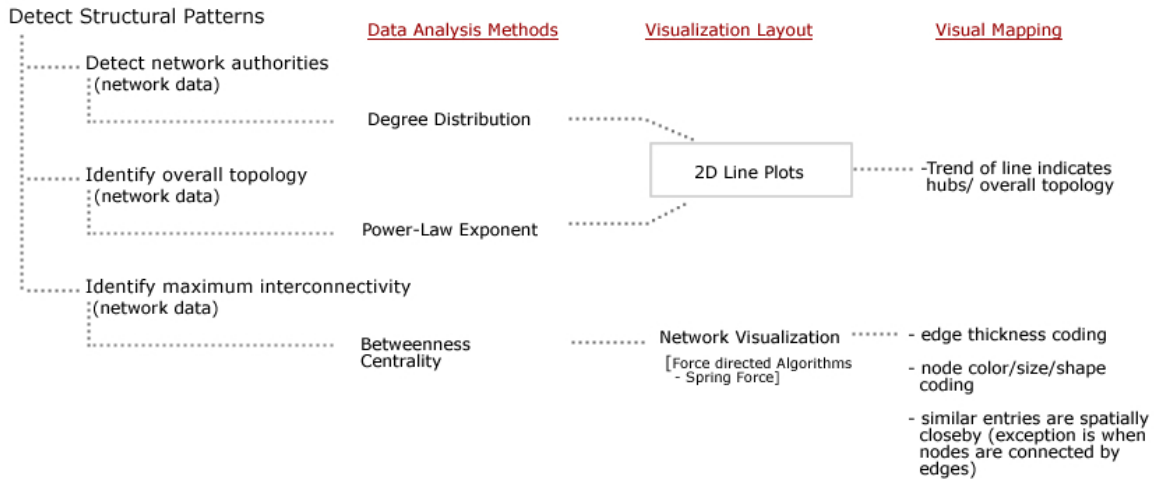


Figure 27: Data analysis and visualization coupled taxonomy to detect structural patterns

Chapter 6

Validation: Using DA-Vis Taxonomy to Develop New Visualizations

This chapter applies the DA-Vis taxonomy to two application domains: 'Computational Diagnostics' and 'Knowledge Management'. For each application domain, the data analysis goals and user task abstractions are identified. Next, the DA-Vis taxonomy is used to select valid pathways.

6.1. Computational Diagnostics

Clinical trials are common in medical domain to identify potential new treatment for different diseases. During clinical trials, new drugs are tested to study their effects. From the selected patient group, variations in patient responses to new drug treatment are captured to evaluate the effects of the drug under different medical conditions. A comparison of the patient responses helps to identify the patient medical conditions under which the drug performs better. For the comparison, the user needs to have a global overview of the patient medical conditions, and a means to quickly identify the patients with positive or negative side-effects.

Computational diagnostics focuses on the development of novel computational techniques to look at medical data. The goal is to help identify characteristic patterns in

patient medical data based on clinical trials or treatment information. Other data such as census data can be added as well. Ability to compare patterns and variation in medical condition of patients helps to identify customized treatment for patients with almost identical medical conditions.

Subsequently, the computational diagnostics approach is being applied to medical records of acute lymphoblastic leukemia patients (ALL). The goal of the computational diagnostics approach and other details related to ALL dataset are covered below.

6.1.1. Data Analysis Goal

The main analysis goal is to identify medical conditions that cause a relapse in acute lymphoblastic leukemia (ALL) patients. Sometimes patients share almost same medical condition, but when undergoing the same medical treatment one patient survives while the other one dies. As their medical conditions look identical, it is difficult to quickly identify data characteristics or combination of data characteristics that cause relapse in a subset of the patient group which is subjected to similar medical treatment.

To help identify data characteristics that causes relapse, the envisioned computational diagnostics tool must provide a global overview; ability to align patients adjacent to each other for comparison; provide the ability to quickly identify patterns among data variables; facilitate the observation of a selected group of patients as groups to compare trends, etc. By adopting an approach to develop interactive visualization tools, all of the above requirements can be met. Hence, a data visualization based computational diagnostic tool is envisioned.

6.1.2. User Task Abstraction

The data analysis goal to identify factors that causes relapse in patients is very general. To accomplish this goal, several related questions serve as user tasks. The questions are covered in the first column of Table 1. The second column from Table 1 shows the mapping between the user tasks and different primary data analysis goals, see section 3.3. The third

column from Table 1 shows specific data analysis goals that will help to meet demands of the user task.

From Table 1, it can be seen that task 7 and task 9 can be accomplished by user interaction. In performing the different user tasks, the motive is also to keep a global overview of the dataset.

Table 1: Mapping of user tasks to data analysis task for medical dataset

User Tasks	Primary Data Analysis Goal	Specific Data Analysis Goal
1. Can we get an overview of all patient medical condition that are a part of the dataset?	-	Global overview
2. Is it possible to identify patients with the worst values at a variable level?	Identify Patterns	Find Order
3. Is it possible to identify patients with the overall worst medical conditions?	Identify Patterns	Find Order
4. Do the patients who are either dead/alive share common medical conditions for some variables?	Identify Associations	Find Correlation
5. Can we have a demographic distribution of medical conditions for all patients in the dataset?	Identify Trends	Find Occurrence
6. Is there a correlation between different variables available within a medical dataset?	Identify Associations	Find Correlation
7. Can we get a phenotype/prognostic/combo view of the patient values?	-	Toggle Between Views
8. Can we identify groups of patients who share similar trends across multiple variables?	Identify Trends	Show Trend
9. Can we compare the trends seen in two patient groups?	Identify Trends	Show Selected Groups

6.1.3. Dataset Details

The medical records include patients diagnosed with acute lymphoblastic leukemia (ALL) at Indiana University in the time-period 1992-2005. The medical data contains different category variables of patients such as:

A. Clinical data and laboratory data [Source: Regenstrief Medical Records System]
Patient variables: haemoglobin (Hgb), white blood cells (WBC), platelets (PLT), CNS diseases (CNS), relapse condition, relapse site, last day known alive (LDKA), diagnostic age (ageDx), blast % readings for day 7, and day 14 in response to treatment.

B. Cytogenetics data [Source: clinical genetic database at Indiana University]
Patient variables: genetic category (chromosomal state), chromosomal structural defect information.

C. Immunophenotype data [Source: pathology database at Indiana University]
Patient variables: B-cell/T-cell lineage condition.

D. Patient demographics [Source: Census data]
Patient variables: gender, race or ethnic background, family income, educational level, %single family and %employment.

Some of the variables identified above (A-D) act as primary sources of information and are of interest to doctors. These variables along with their categories and information sources are listed below. Detailed descriptions of medical variables are available in Appendix III.

Further, to facilitate easy communication of variable information to doctors, it becomes important to know how they perceive the information. Through communication with a

doctor, it was known that doctors try to gain insight about patient condition by analyzing variable information using different categories and associated variables such as:

A. Outcome

Patient Variables: relapse, relapse site, alive/death status, and LDKA.

B. Biology

Patient Variables: immunophenotype, genetic condition, WBC, Hgb, platelets, and CNS.

C. Host

Patient Variables: diagnostic age (ageDx), gender, and race.

D. Treatment

Patient Variables: BM 7, and BM 14.

E. Social Factors

Patient Variables: MFI-class, patient's education level, percentage of single family members and percentage of family employment.

By communication with the doctor, it was also known that viewing the data at two levels (phenotype and prognosis) would help the doctors gain better insight into the patient condition during diagnosis. In phenotype view, the variables shown are independent of other variables, so henceforth these variables are referred to as 'independent variables'. On the other hand, in prognosis view the variable value is dependent on other variables and hence these variables are henceforth referred to as 'dependent variables'.

In the phenotype view, the display shows most data in its raw format. In this view, the patient medical condition is indicated by 'hazard ratio', which is based on the patient age. But the hazard values are only available for some variables (WBC, Hgb, platelets). Higher values of hazard ratio indicate a worse condition for a given patient. More information related to hazard values for different variables is available in Appendix IV.

The prognostic view gives information about the event free survival percentage (%EFS) of the patient. The %EFS is another indicator of the patient medical condition. Lower values for %EFS indicate critical patient conditions while higher values indicate that the patient is in a healthy condition and less likely to relapse. The %EFS conditions differ for different variables. Details about the %EFS values applicable to known variables are given in Appendix IV.

6.1.4. Task Based Application of DA-Vis Taxonomy

Based on the data analysis tasks, pathways from the coupled DA-Vis taxonomy are identified. The initial step of identifying a taxonomy pathway is based on the user task that needs to be supported. A generalization of the user tasks (see Table 1) to a primary data mining task helps to identify the sub-taxonomy that can be used. Identified pathways that meet the user requirements are shown below:

Task 1: Ordering to see patterns

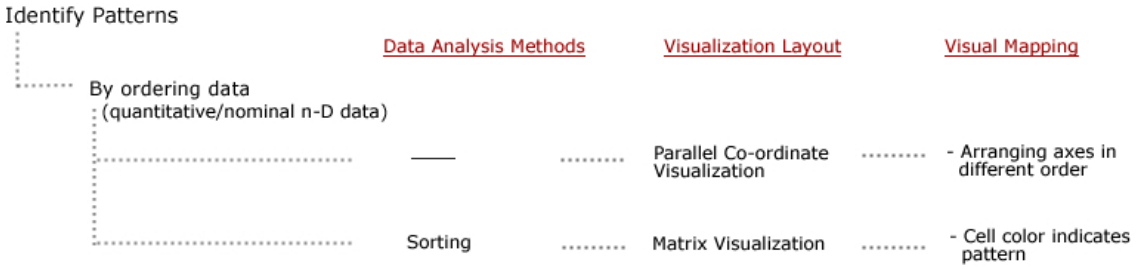


Figure 28: Pathway from DA-Vis taxonomy to order data

Task 2: Find correlations

Figure 29: Pathway form DA-Vis taxonomy to find correlation in data

Task 3: See trend based on selection

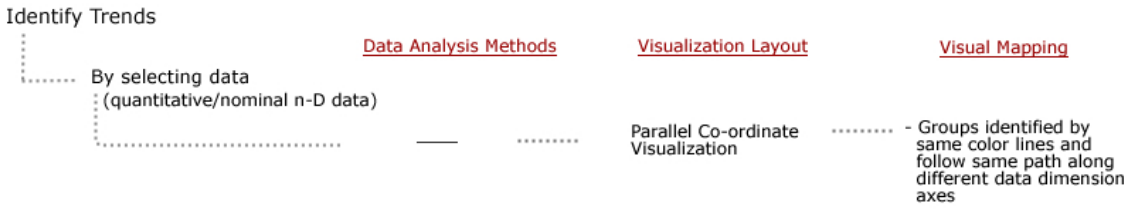


Figure 30: Pathway from DA-Vis taxonomy to identify trends by data selection

Task 4: See distribution based on trends within the dataset

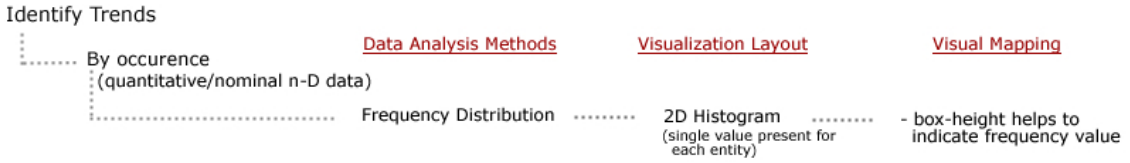


Figure 31: Pathway from DA-Vis taxonomy to identify trends based on data occurrence

By mapping user goals to pathways in the DA-Vis taxonomy, appropriate techniques are identified as shown. The identified pathways show that both matrix and parallel coordinate visualizations are needed to satisfy the user task requirements. The matrix view provides a good overview and serves as an easy way to communicate associations, and patterns. On the other hand, the parallel coordinate view shows data in the form of lines. In the real world, we are trained to see the trend in the form of a line (e.g.: business profit charts). The parallel coordinate makes use of lines to represent different values of a data entity to see the trend in the data.

6.1.5. Visual Design

Using the DA-Vis taxonomy, matrix and parallel coordinate visualization techniques were identified. In this section, we cover details about mapping the data characteristics onto these two visualizations. The system architecture that is used to support coordinated viewing of the medical data is also presented.

6.1.5.1. Matrix Visualization

The matrix view is used to visualize patient data. In this view, the patient study-ids are mapped to columns and different patient variables are shown as individual rows. The independent variables (phenotype) and dependent variables (prognosis) data are represented in the form of matrix cell color. Two different color palates are used within the view to distinguish between phenotype and prognosis variables. Patient severity condition for individual variable types can be inferred using colors from respective palates.

Figure 32 shows the phenotype view, Figure 33 shows the prognosis view of the dataset, and Figure 34 shows integrated view representing variables with either phenotype or prognosis conditions.

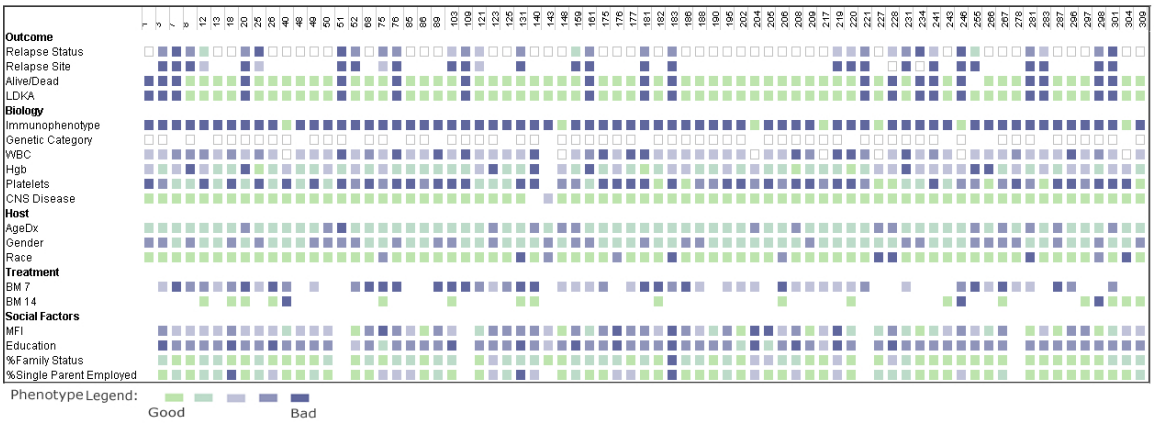


Figure 32: Phenotype view of the patient medical dataset

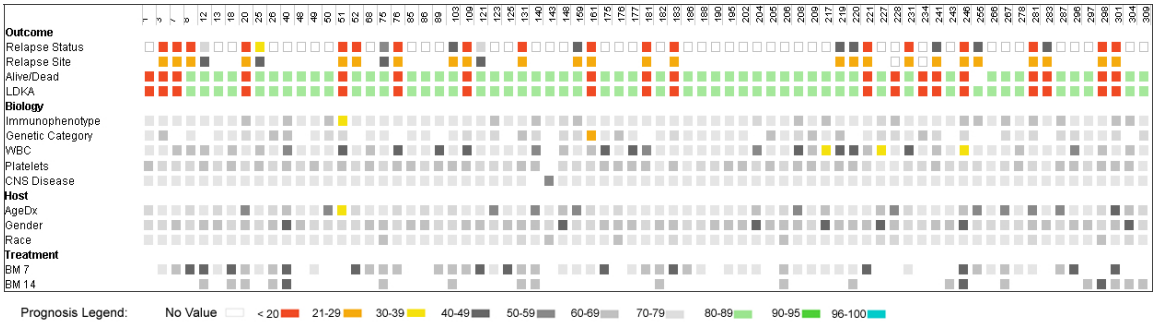


Figure 33: Prognosis view of the patient medical dataset

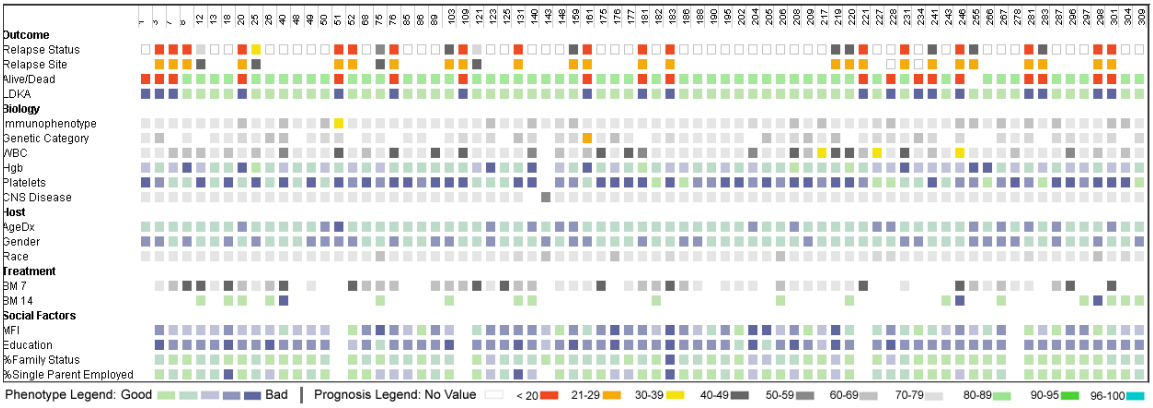


Figure 34: Combined phenotype and prognosis view of the patient medical dataset

A. Data Mapping

A blue to green color gradient is used to map phenotype data values. Blue is used to represent bad medical conditions while green is used to indicate good medical conditions.

In the prognosis matrix view, the derived variables are shown using red-to-gray-to-green color gradient scale. In this case, red color indicates the worst medical condition. The medical condition severity decreases as the saturation of red decreases. Color gradient range and associated event free survival percent (%EFS) severity are shown in Table 2.

Table 2: Color range and associated %EFS condition

Color Range	%EFS	Patient Condition	Risk of Relapse
Red gradient	0 – 39	worse	High
Gray gradient	40 – 89	commonly seen	Moderate
Green gradient	90 – 100	better	Low

B. Usage for Data Interpretation

The color coding supports the easy identification of patients with similar medical conditions. The distribution of data values is an indicator of the general trends. For the medical dataset, unique occurrences of individual values are computed. Depending upon the view (phenotype or prognosis), the values are binned into different value ranges and a color

is assigned to them. The matrix view here is also used to support user tasks such as sorting and detecting patterns.

The sorting functionality is used to reorder patient data in the matrix view. Depending on the data values, matrix cells are rearranged in increasing or decreasing order. This functionality helps to determine different groups of patients who share similar medical conditions.

6.1.5.2. Parallel Coordinate Visualization

A parallel coordinate visualization was generated using an open source program called 'Parvis' [100]. Modifications needed at axes level to optimize the parallel coordinate view for diagnosis purpose are covered below. The modified parallel coordinate view is shown in Figure 35.

A. Data Mapping

Primarily, the parallel coordinate view is used to show quantitative data. But the view shown in Figure 35 is modified to handle both nominal and quantitative data values. For example, variable 'Relapse Site' includes data values such as: BM, BM combined, CNS/Testis, and no relapse. The nominal values are shown as separate regions on the respective axes. Similarly, other data variables with nominal data values include: relapse, deceased, immunophenotype, genetic category, CNS, race, BM 7, BM 14, family income, and education.

On the other hand, quantitative data variable values are arranged with increasing values from bottom to the top of the axis. Quantitative variables include: LDKA, WBC, Hgb, platelets, ageDx, single family percentile, and employment percentile. Value markings are made on axes with quantitative values to help visually separate patients with higher and lower variable values.

Additional information on severity values for different variables is overlaid in the form of gray boxes. For nominal variables, rectangular gray boxes are used. Triangular gray boxes are used to indicate severity levels for quantitative variables. Special treatment is used for WBC

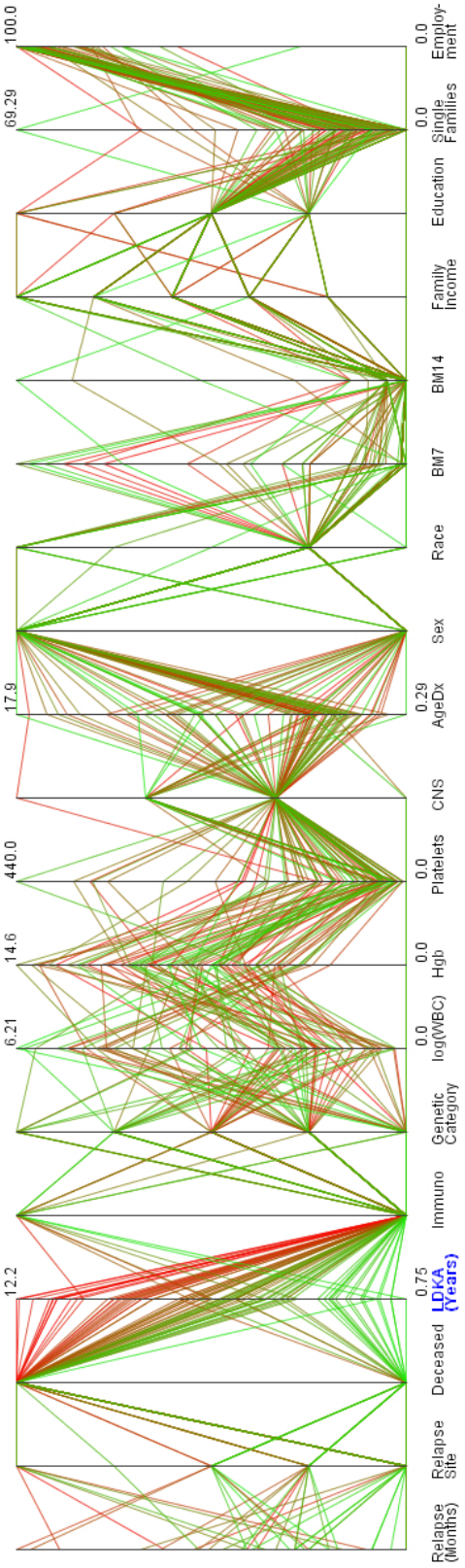


Figure 35: Parallel coordinate visualization for acute lymphoblastic leukemia dataset

variables because patients' value can range from 1000 to 50,000. To show the entire spectrum of values shown on the WBC axis, a log scale is used. All numerical values are mapped with lower numerical value variable at the bottom and the higher numerical value at the top of the axes.

B. User Data Interaction & Usage in Data Interpretation

Figure 39 shows data in the parallel coordinate view. The axes represent different diagnostic variables (both nominal and quantitative). Within the view, each patient is represented by a line. The patient value for a variable is indicated by the point of intersection of the patient line along each variable axis. The initial arrangement of axes is identical to the order available in the matrix view. But the order of axes in parallel coordinate view can be changed. Different order lead to different patterns in the parallel coordinates view. For example, Figure 36 shows a unique line pattern that is created when axes: (A) relapse (months), (B) LDKA, and (C) genetic category are in the order ABC. For axes arrangement as BAC; see Figure 37, a different line pattern is created by the same dataset. An axes arrangement of CAB brings forth another line pattern; see Figure 38. Thus, similar arrangement of other axes in the parallel coordinate view would reveal different data patterns. The user can place different axes closer together to closely study the correlations between values of different data variables.

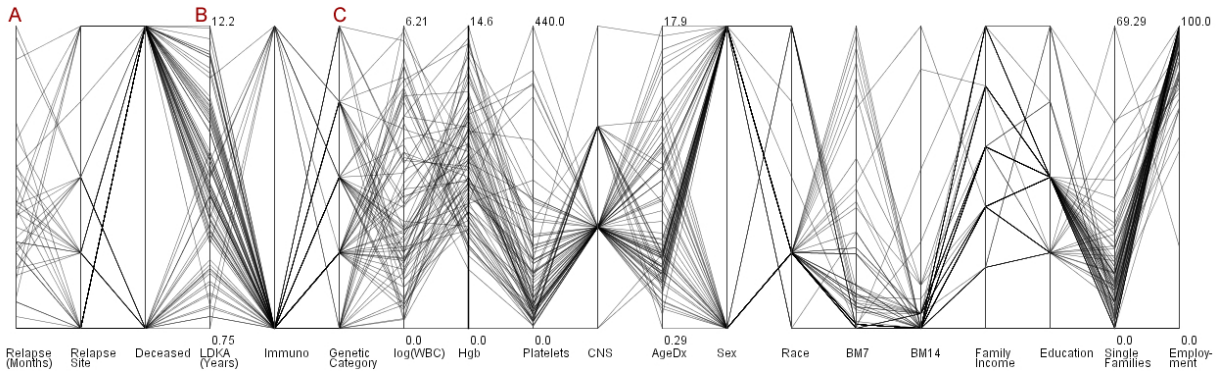


Figure 36: Parallel coordinate view to show data line pattern with selected axes order as A, B and C

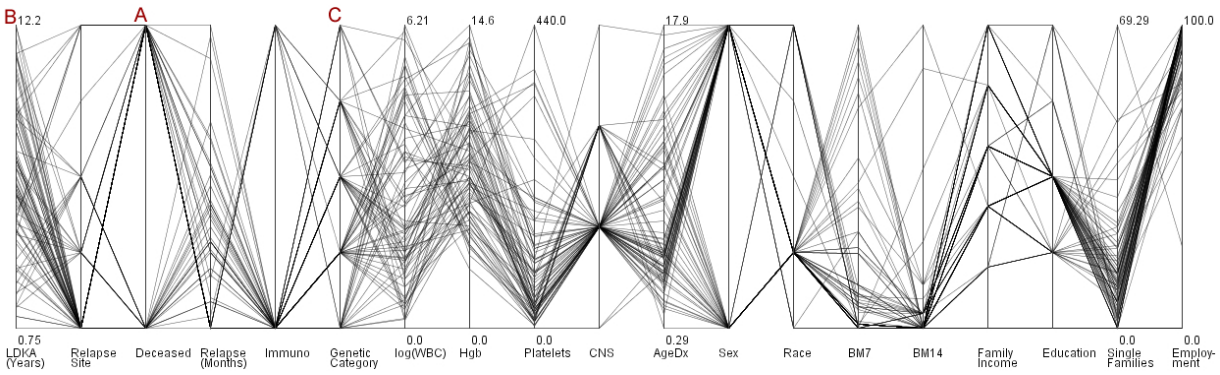


Figure 37: Parallel coordinate view to show data line pattern with selected axes order as B, A, and C

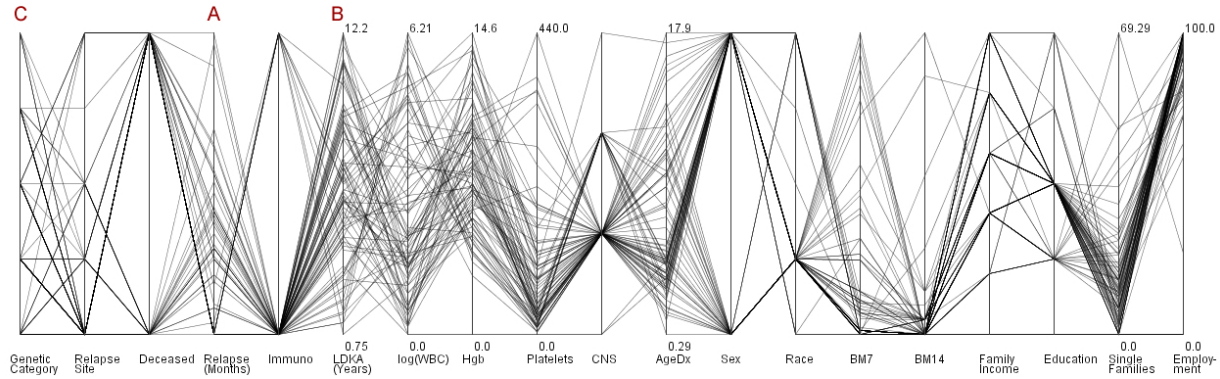


Figure 38: Parallel coordinate view to show data line pattern with selected axes order as C, A, and B

Figure 39 shows 81 patients (lines) with 19 different variables (axes) each. Line segments with dark color indicate that there are many patients who share the same trend. To show the line for a single patient, a mouse-hover event was implemented. When the cursor hovers over a line, the specific patient line is highlighted in red color. For example, in Figure 39 the line highlighted in red shows the trend for patient with a study id of 314.

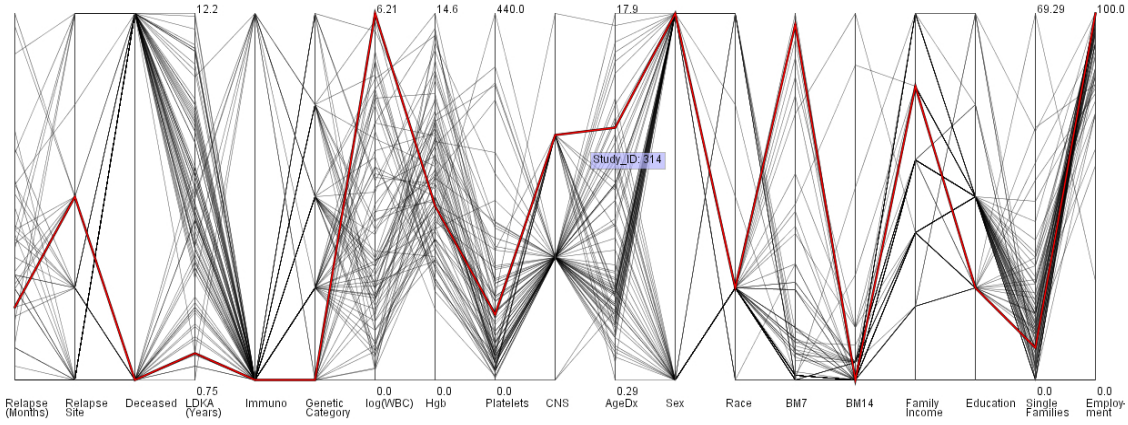


Figure 39: Parallel coordinate view showing data for 81 patients with acute lymphoblastic leukemia

a. Tool-tip display to show diagnostic values of the selected patient

The tool-tip display can be activated by check-box selection in the interface. When the cursor hovers over a line and the tool-tip display check-box is selected, then values for the specific patient are shown to the right of each axis. For example, in Figure 40, the tool-tip displays values for the patient with the study id of 314.

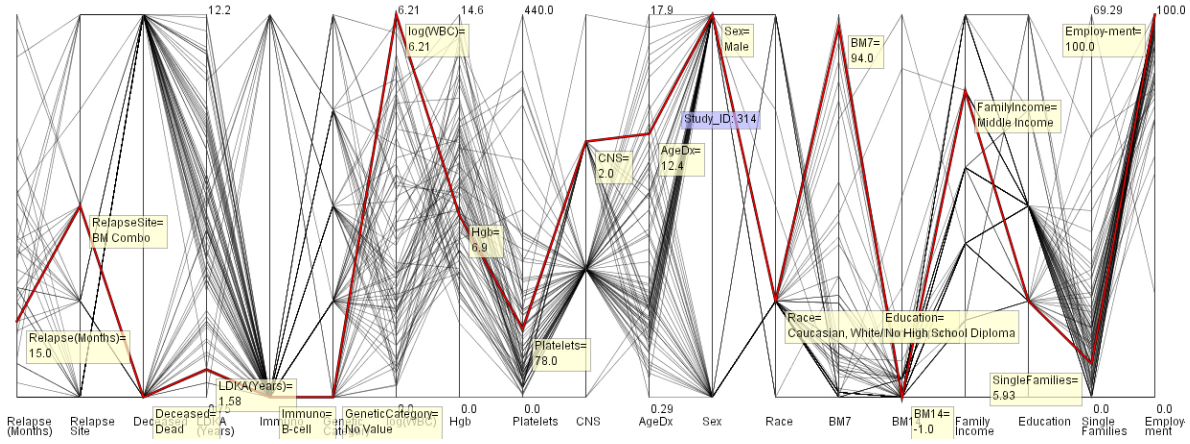


Figure 40: Parallel coordinate with tool-tips showing data for a single patient

b. Display axes-labels to mark different regions/values along axes

The axis labels can be displayed on demand for each displayed axis. For axes representing quantitative data, numerical values are shown. On the other hand for axes showing nominal data, data-labels are displayed. The axis labels act as landmarks to obtain an overview of the data distribution for the data being displayed in the parallel coordinate view, see Figure 41.

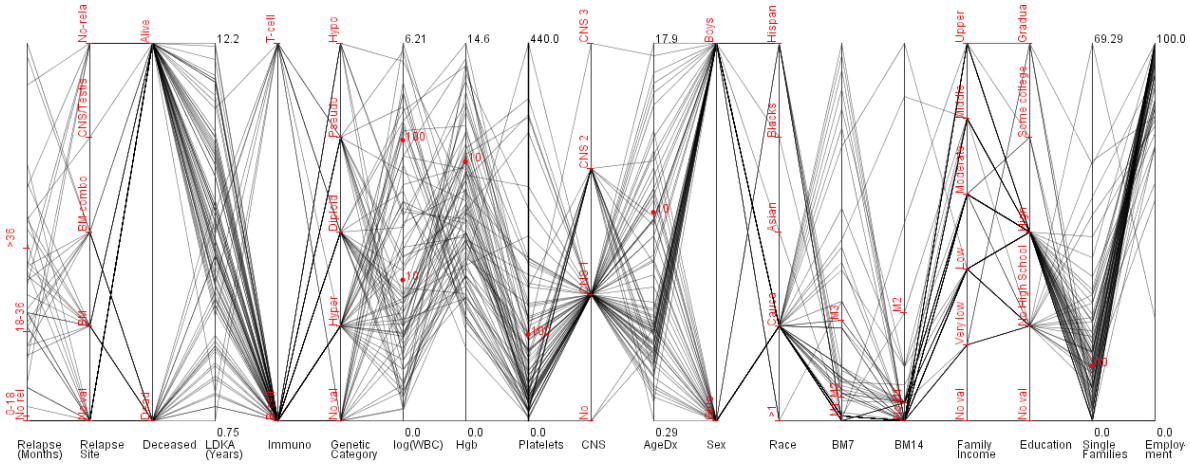


Figure 41: Parallel coordinate view with axis labels

c. Display zones to show severity values for different variables

For some diagnostic variables, potentially harmful value ranges have been identified. These ranges can be shown by displaying zones along the axes shown in the parallel coordinate view, see Figure 42. User-interaction is provided to turn on/off the severity zones from the display. Triangular shaped zones are used to show severity along axes showing quantitative variables (e.g., Hgb in Figure 42). The increasing width of the triangle is used to show that increase in variable value caused increase in severity. The rectangular shaped zones are used to show the constant severity along axes showing nominal data (e.g., BM14 in Figure 42).

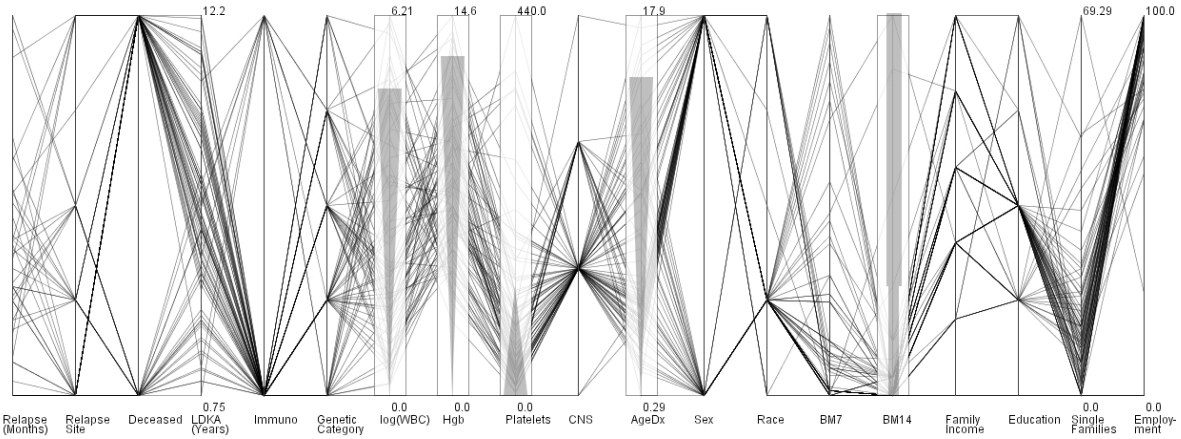


Figure 42: Parallel coordinate view showing regions with severity value

d. Axis selection to study global variations in patient values

To show the trend in patient lines based on a single variable selection, a single axis selection feature is included. A mouse double-click action is used as a user interaction to select one axis at a time. The selected axis variable name is highlighted in blue to clearly distinguish it from the other axes. Based on the values along the selected axis, patient lines are color coded on a red-to-green gradient. Higher values are indicated in red, a gradient is used for intermediate value range, and green color is used to indicate the lower range values. For example, in Figure 43 patient lines are color-coded based on the 'LDKA' values. The color established based on the selected axis are maintained across different axes in the entire view. This feature offers the ability to study the global variation in patient data based on the selected axis values.

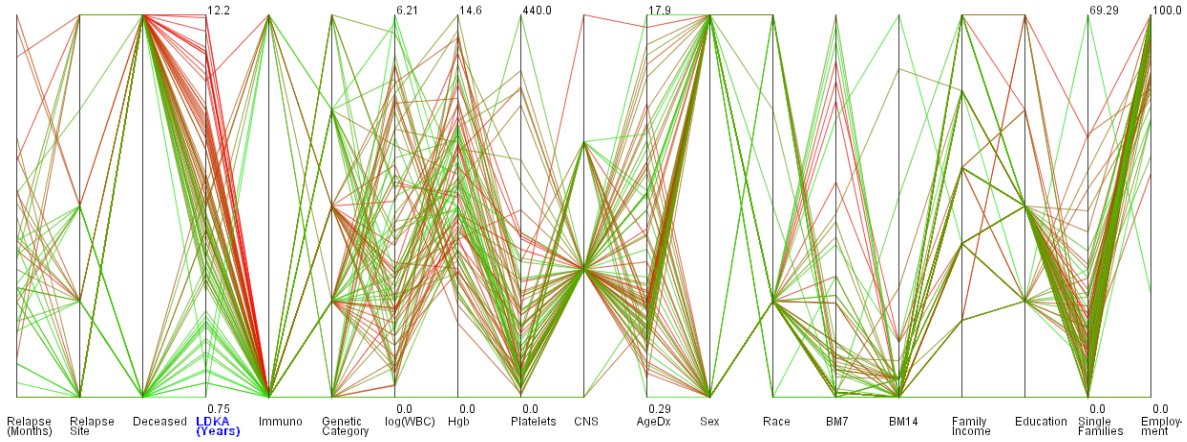


Figure 43: Patient lines highlighted based on selected axis (LDKA) in parallel coordinate view (81 patients shown)

e. Select and categorize patients as a group

To identify the trend in values for a group of patients, the brushing functionality can be used to select a group of patients. Figure 44 shows a group of patients selected by brushing along the 'AgeDx' axis.

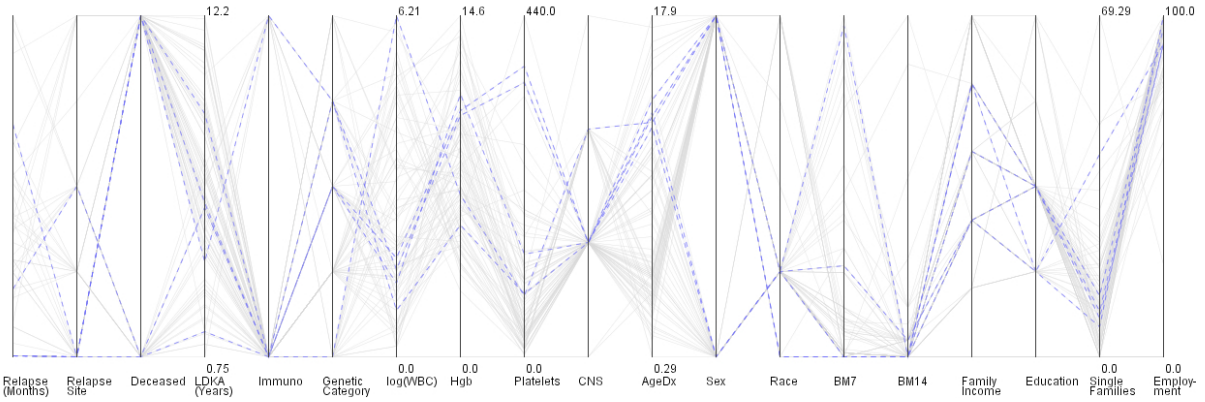


Figure 44: Parallel coordinate view shows the selected patients as a group

f. Simultaneously display groups of patients to study their variation

Using the brushing functionality, two groups of patients can be selected. Differently stroked and colored lines are used to clearly distinguish the two groups. The two groups can be saved as 'selected groups' using the GUI interface. Further, different combinations of groups from the saved selection can be combined together by keeping the 'Shift' key pressed and selecting the groups from the selected group panel. Figure 45 and Figure 46 show the

selection of patient lines as group 1 and group 2 respectively. Figure 47 shows a view where two group of patients are displayed simultaneously. This feature of the ability to display multiple prior selected groups is useful when comparing trends across different groups.

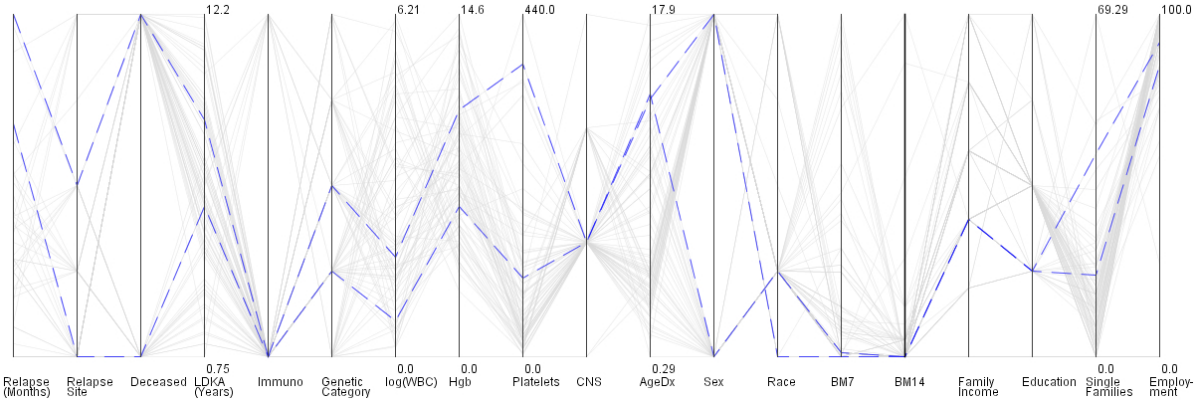


Figure 45: Patient lines identified as group 1

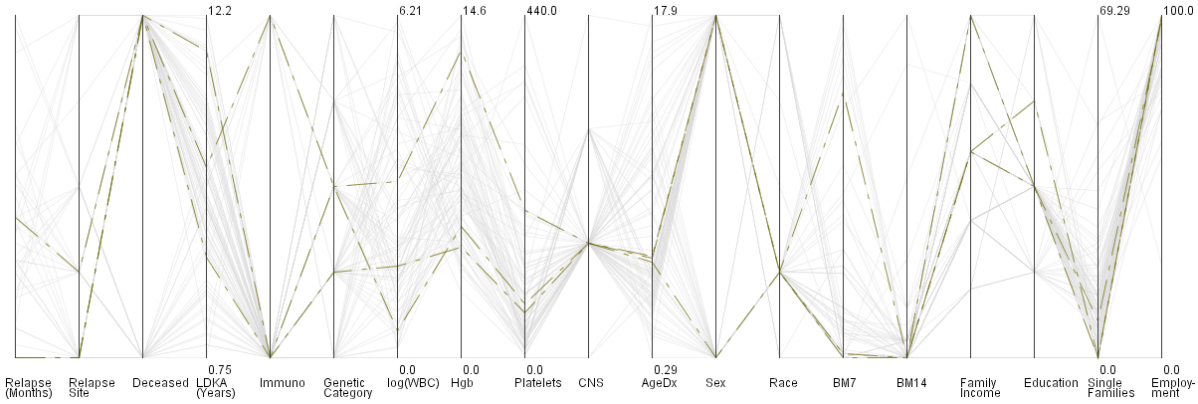


Figure 46: Patient lines identified as group 2

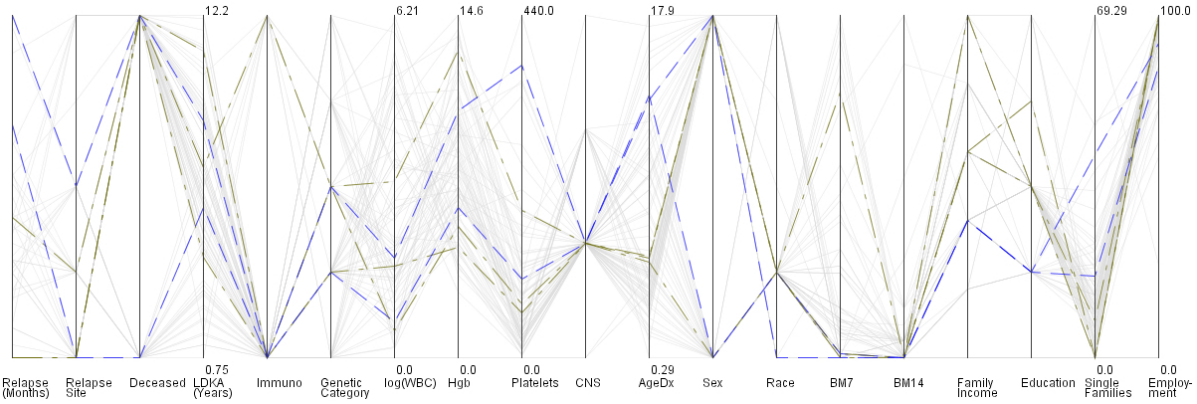


Figure 47: Simultaneous views of group 1 and group 2

6.1.6. System Architecture for Coordinated Viewing of Medical Dataset

The patient data that is shown in matrix view and parallel coordinated view is stored in a MS-Access database. The system architecture that is used to query the database, to acquire and parse the results, to apply conditional transformations, and to generate visualizations (matrix visualization and parallel coordinate visualization) is described below. Further, it is important to interlink the matrix and parallel coordinate visualization as each view offers a distinct advantage. The matrix visualization is helpful when quickly determining patterns in the dataset. The parallel coordinate visualization is useful when checking trends in patients or patient groups. The system architecture also describes details related to the coupling of matrix and the parallel coordinated view. The schematic representation of the system architecture is shown in Figure 48.

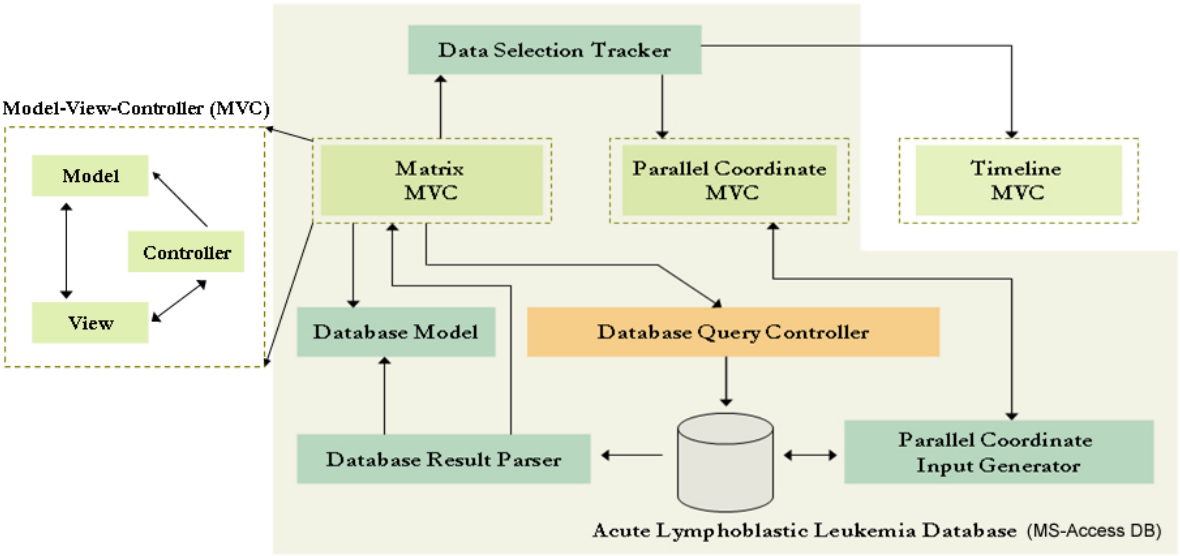


Figure 48: System architecture of the designed computational diagnostic application

The application being developed is based on the Model-View-Controller (MVC) design pattern. The above diagram shows three views that were implemented – matrix view, parallel coordinate view, and timeline view. The matrix and parallel view are covered in the sub-section above. The timeline view is used to show the progression of certain variables over time. The timeline view is included to show the complete system architecture diagram but it will not be implemented as a part of the thesis.

The MVC design pattern helps to cleanly separate objects into three categories - the model is used for maintaining the data, the view is used for data display, and the controller is used to handle events that affect the model and the view.

To generate the matrix visualization, user interactions to display different kinds of data (phenotype and/or prognosis) at variable level are captured by the controller for the matrix view. As different variable information is available in different database tables, the database query must be specific. Depending upon the variable data to be accessed, the database query controller issues specific SQL queries to query the database and get results. Each variable data results are in turn passed across to database result parsers. The result parsers include variable conditions level information to prune individual data and store it as a database model. When all variable data is available, then the matrix controller uses the matrix view to create the desired view (phenotype/prognosis/combined view). Depending upon the user requirements to visualize phenotype or prognosis data at variable level, the individual matrix cells are color coded with different color codes depending upon their data values.

To generate a parallel coordinate view, it is important to identify different variable data type information (numerical or categorical). This identification is important as parallel coordinate supports viewing of only numerical data. To support the viewing of quantitative information in parallel coordinate view, specific conditions are written in MS-access database to transform categorical data into numerical data. The transformed numerical data for all patient variables are obtained and a parallel coordinate data model is constructed. The parallel coordinate view is used to display the data. Different user interactions are captured by the controller to make relevant changes in the dataset.

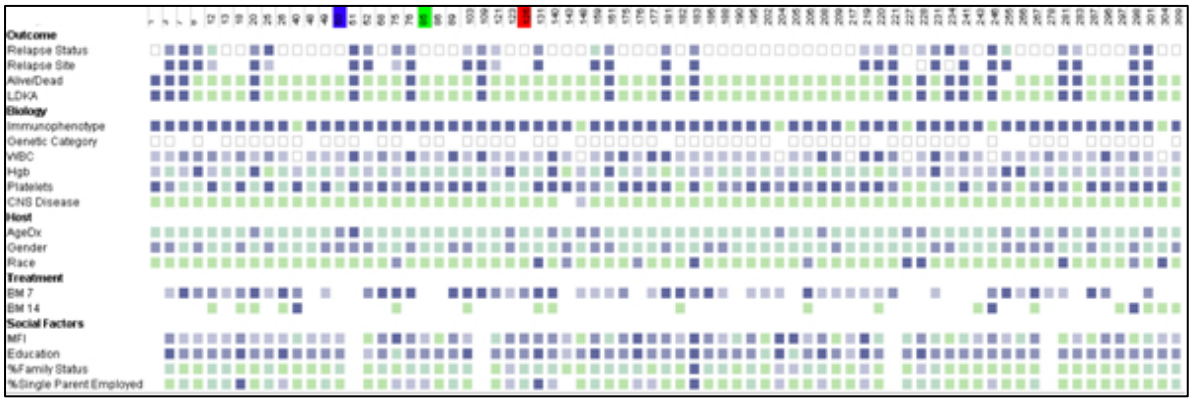
To interlink the matrix and parallel coordinate view, a data selection tracker is introduced in matrix view. The tracker keeps track of the selected patients in the matrix view. The selected patient information is passed across to the parallel coordinate view where the selected patient line is highlighted. The patient selection in the matrix view and corresponding highlighting of patient line in parallel coordinate view is shown in Figure 49.

6.1.7. User Case Scenario for Multiple Coordinated Views

The user case scenario describes the use of the matrix view and the parallel coordinate view to work in unison to provide different perspective for the same acute lymphoblastic leukemia dataset. Individually both views display data in their own format.

The goal here is to simultaneously view selected patient data in both views. To accomplish this, the user selects a patient of interest by mouse double-click action on the patient's study id. The action pops up a color swatch to select a color for the patient. The selected color is used as a background color for the study id of the patent in matrix view to visually distinguish the selected patient study id from the other selection, see Figure 49-A. The same color code is used to highlight the patient line in parallel coordinate view, see Figure 49-B.

(A)



(B)

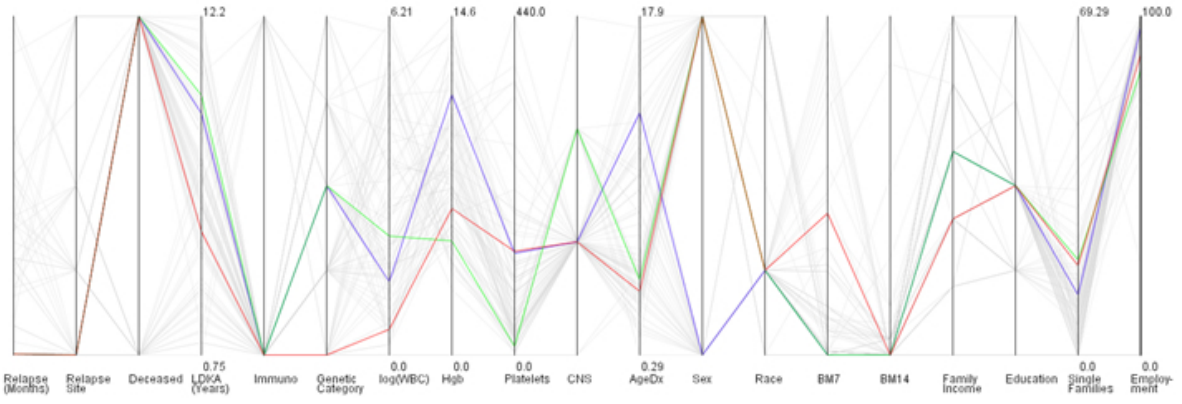


Figure 49: Multiple coordinated view showing selected patients and respective color codes in (A) matrix view and (B) parallel coordinated view

6.1.8. Insight Offered by the Computational Diagnostic Tool

The computational diagnostic tool can be used in the study of patient clinical trials. The tool supports 'Evidence Based Medicine' (EBM) – a new paradigm in medical practice. Wikipedia defines EBM as 'the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients' [101].

By using the designed tool, the medical practitioner can load the clinical trial data. The matrix view provides a global overview and also provides a quick way of looking at patterns in the data. With color codes, patients with the worst values can easily be identified from the rest of the population in the dataset. The parallel coordinate offers a quick overview of the patient data distribution. In-built interactivity helps in data exploration of a single patient or group of patients.

The multiple coordinated views provide a medical practitioner with different perspectives of the same data. From a diagnostics perspective, the medical practitioner gains insight about data variations among selected patients. The matrix view helps to quickly identify similar patterns and worst case conditions. On the other hand, the parallel coordinate view helps to quickly identify and compare trends shared by groups of patients. So the matrix view and parallel coordinate view complement each other to help the medical practitioners gain quick insight about the data.

The existing tool template can be expanded to incorporate medical details that would help in the study of other clinical trials. Examples of clinical trials are: cardiology patient datasets; behavioral science data studies of the correlation between smoking and diabetics, etc.

6.2. Knowledge Management

The information age has led to the explosive growth of new data. Traditional methods are still being used to prune and sift through the data. But the use of traditional methods makes

it difficult to keep the pace with the information growth. In the scientific domain, it becomes increasingly difficult to get a global overview of the entire dataset.

The knowledge domain visualization (KDVs) is an emerging field of study that helps to generate maps to show the structure and evolution of a field. KDVs help to obtain a global overview of the dataset which shows the structure and evolution of a scientific field. KDVs use sophisticated data analysis and visualization techniques to identify major research areas, subject domain experts, institutions, grants, publications, etc.

As KDVs provides a global overview of the dataset. They help to quickly identify the area of interest. Rather than sifting through the entire dataset for the desired information, the KDVs knowledge maps exploit the power of human vision and spatial cognition to help humans mentally organize, electronically access, and manage large complex information spaces. The visualizations offer a unique advantage in organizing and managing information. Using these kinds of knowledge management techniques, one can use existing knowledge to study the diffusion of knowledge among institutions, potential collaborators for a given subject domain, effect of institution proximity for knowledge transfer, etc. The use of KDVs to study a dataset comprising 20 years of published papers in PNAS journal is covered below.

6.2.1. Data Analysis Goal

The goal behind the analysis of the PNAS dataset is to identify the major topics and trends in biomedical research from the 20 years time-slice between1982 - 2002 of published literature.

6.2.2. User Task Abstraction

The list of questions that are targeted towards answering the data analysis goals are covered in the first column of Table 3. The second column shows the mapping between the user tasks and primary data analysis goals. The third column shows specific data analysis goals that will help to meet the demands of the user task.

Table 3: Mapping of user tasks to data analysis task for PNAS dataset

User Tasks	Primary Data Analysis Goals	Specific Data Analysis Goals
a. To identify trends in research of hot topics in the given time frame.	Identify Trends	Determine occurrence
b. To identify the burst of research topics in the given time frame.	Identify Trends	Identify data activity
c. To identify relations between hot topics and burst research topics.	Identify Association	Identify linkages
d. To view the change of focus in certain areas (dynamics) of the bioscience domain during the 20 year time period.	Identify Trends	Identify data activity

6.2.3. Dataset Details

The Proceedings of National Academy of Science (PNAS) is one of the most distinguished journals. It publishes new research in the biomedical domain. The PNAS dataset used for analysis includes a 20 year time slice (1982-2001) dataset of published biomedical literature [102]. The dataset includes information about different fields like: title, keywords, journal number, issue number, etc. From all these fields, the title and keyword variables are a primary focus during our analysis.

The original dataset consists of 47,073 published articles. Assuming that highly cited work will include information about major topics and trends, a data subset containing top 10% of the most highly cited publications were considered for data analysis. After excluding papers without titles, the final dataset considered for analysis consisted of 4,699 papers.

6.2.4. Task Based Application of DA-Vis Taxonomy

In this section, the identified data analysis tasks are mapped onto the new DA-Vis taxonomy to identify techniques needed to accomplish the user task. The identified pathways are shown below.

Task 1: Identify topic frequency to see the trend over the years (multiple values are available as the word frequency is computed for each year).

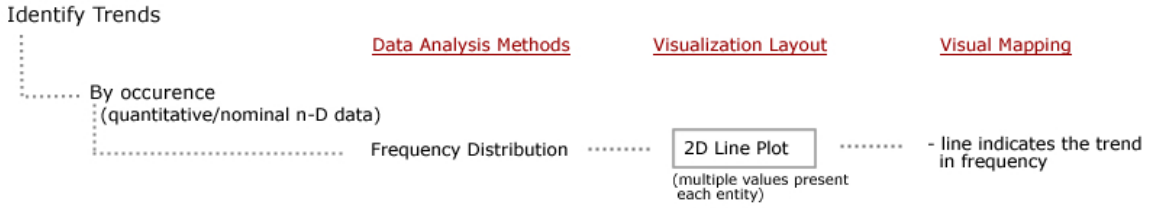


Figure 50: Pathway from DA-Vis taxonomy to identify trend by occurrence of data

Task 2: Identify data activity

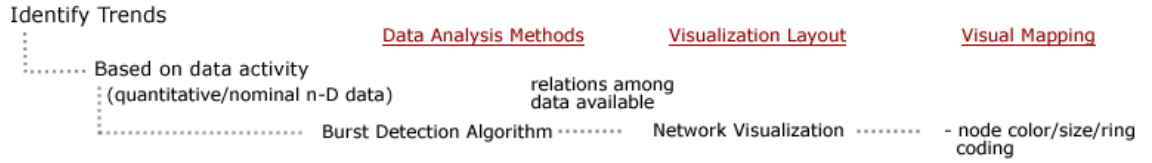


Figure 51: Pathway for DA-Vis taxonomy to identify trend by data activity

Task 3: Identify linkages is to determine association among topics

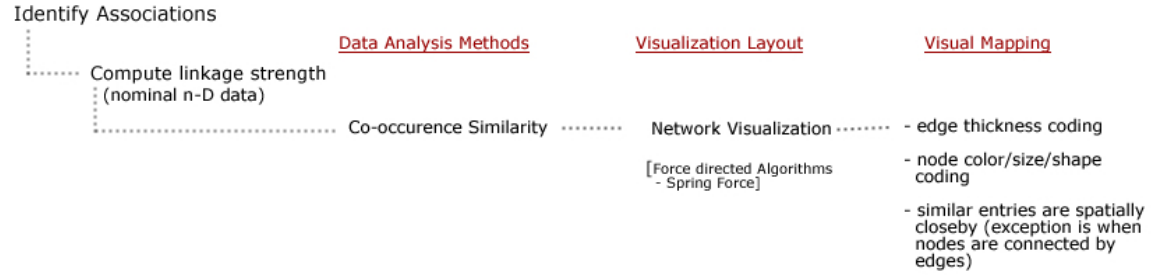


Figure 52: Pathway from DA-Vis taxonomy to identify association based on linkage strength

Based on the taxonomy, the frequency distribution of research topics over time can be visualized using a line plot. The data analysis method for detecting bursts is the 'Burst Detection Algorithm'. Further, co-occurrence similarity can be used to identify linkages.

The original network visualization generated based on co-occurrence similarity results shows 1,082 non-zero entries characterizing the complex co-occurrence relationships among the top 50 bursty research topics. The resulting visualization was very cluttered. Therefore an alternative goal was adopted to generate a visualization where only the meaningful relations

in co-word space were preserved. The DA-Vis taxonomy shows the pathfinder network scaling (Pfnet) algorithm as a data analysis technique that can be applied to accomplish the data analysis goal. The selected pathway from the new coupled DA-Vis taxonomy is shown below.

Task 4: Identify meaningful associations

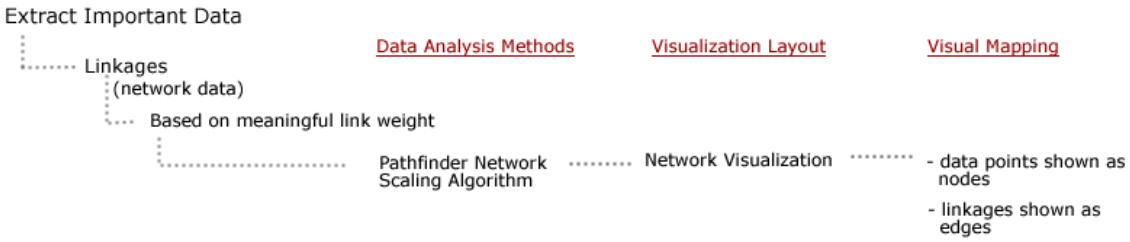


Figure 53: Pathway from DA-Vis taxonomy to identify important data linkages based on meaningful pathways

6.2.5. Visual Design

Using the taxonomy, it was identified that a 2D Line plot can be used to show trends in popular topics which were determined based on their frequency count. Further, a network visualization can be used to display topical burst information and semantic linkages between high frequency topical keywords and bursty keywords. The trend in top 10 high frequency topical research keywords in the time frame 1982-2001 is shown as 2D plot in Figure 54. In the graph, each line is used to represent the value of the number of occurrence of high frequency keyword for each year. The network visualization technique was adopted to simultaneously display the results of the PFnet algorithm and the Burst detection algorithm. The network visualization is generated using a force-directed 2D graph layout algorithm called Fruchterman–Reingold, see section 4.

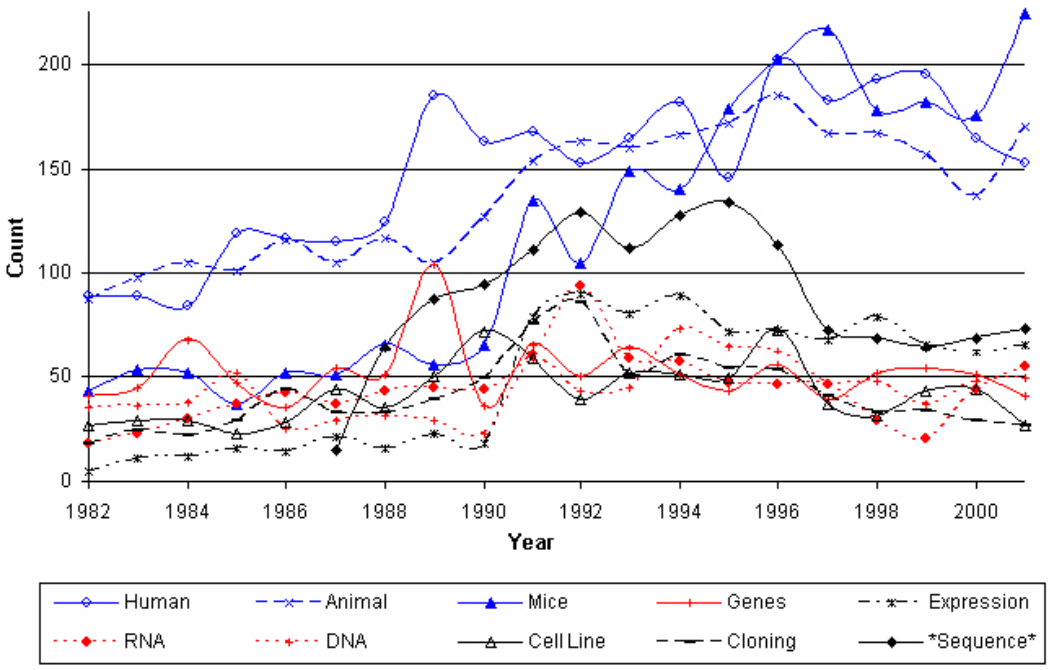


Figure 54: Frequency count for the most frequently used words in the top 10% of most highly cited PNAS publications from 1982 to 2001, adopted from [102]

Nodes in the network visualization show 50 high frequency and bursting research words. Within the visualization, the size of the node circle corresponds to the maximum burst level this word achieved from the burst analysis [34]. Color coding is used to denote the years in which the word was used most often as well as the year of the maximum burst. Five time durations and respective colors were used: 1982–1985, green; 1986–1989, yellow; 1990–1993, orange; 1994–1997, red; and 1998–2001, black. The year of the maximum frequency and the starting year of the first burst of this word were decoded by circle border colors and inner circle area colors, respectively. For example, the word molecular sequence data, represented by a large circle with an orange inner area and a red ring, showed the highest, large burst between 1990 and 1993 and had a high frequency of usage in the later years 1994–1997. Edge thickness is proportional to the number of word co-occurrences. The resulting network visualization is shown in Figure 55.

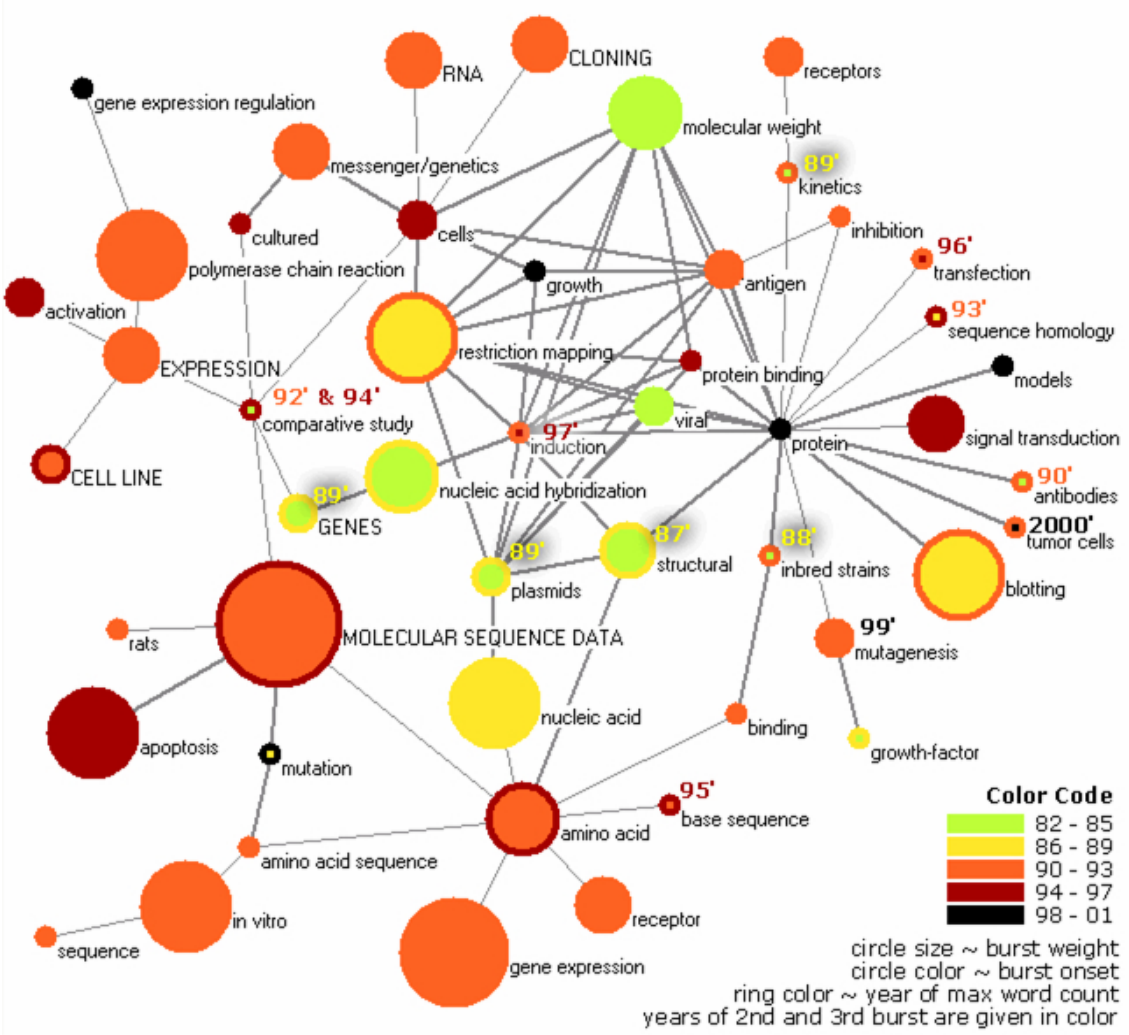


Figure 55: Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982–2001, adopted from [102]

6.2.6. Insight Offered by the Knowledge Management Viz.

The knowledge management visualization (see Figure 55) offers a global view of the dataset. Such a global view provides insight about key research areas in the domain. Further, it offers insight into how research in a particular area benefits other research domains or helps in the emergence of new research areas. By using time as one of the variable, one is able to identify the peak activity year of different research topics. Similar to Figure 55, knowledge management visualizations can be generated for different datasets to represent information in an effective and simple format.

Chapter 7

Validation: Using DA-Vis Taxonomy to Categorize and Describe Prior Work

This chapter demonstrates the utility of the DA-Vis taxonomy in categorizing and describing scholarly published work. For datasets from different subject domains, the taxonomy helps to describe data analysis and visualization techniques that meet user insight needs.

7.1. Identify Related Research Areas in Animal Behavior Domain

Goal: To identify semantic relationships among prominent research areas in animal behavior research across a 40 year time slice [103].

Data Analysis Technique Used: Latent Semantic Analysis (LSA)

Visualization Technique Used: Network Visualization.

Data Analysis Details: Titles and keyword information of animal behavior publications were used to identify the semantic associations between different trends in animal behavior

research. LSA was used to obtain semantic linkages among title words and other keywords. The results were visualized using Kamada-Kawaii network layout algorithm to communicate relation between keywords.

User Task Abstraction: The user task to identify relations among research areas can be abstracted to the identification of associations. In this particular case, the areas of research have to be semantically related to obtain a meaningful interpretation. Hence, it can be inferred that one needs to find semantic linkages between research areas.

Identified Pathway from DA-Vis Taxonomy

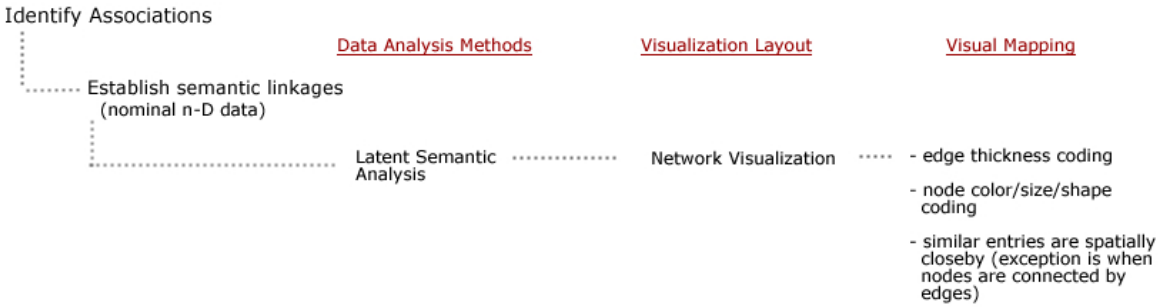


Figure 56: Pathway from DA-Vis taxonomy to identify associations based on semantic linkages

Visual Design Application: In the network visualization, topics are shown by nodes. Edges indicate semantic relations lies between topics. Using the Kamada-Kawaii algorithm ensures that nodes with higher semantic similarity are placed closer to each other. Figure 57 shows the resulting visualization of major topics in animal behavior research during the time-period 1991-1992.

Visualization:

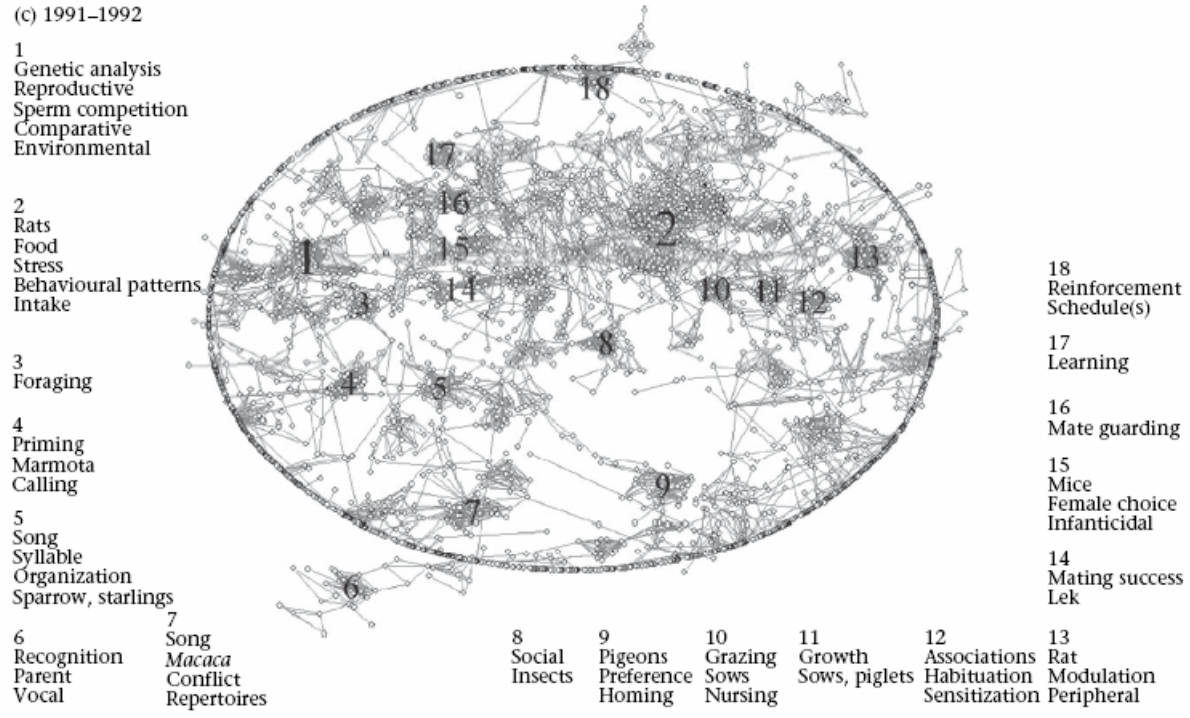


Figure 57: Association between topics in animal behavior research for time period 1991-1992, adopted from [103], reproduced with permission from Elsevier[4]

7.2. Mapping Melanoma Research

Goal: Based on published research on melanoma, identify the relationship between papers, genes, and proteins. Additionally, to identify a surge of interest in certain genes and proteins related to melanoma research [104].

Data Analysis Technique Used: Cosine similarity (to identify relationship between papers,

[4] Reprinted from Animal Behavior Journal, Vol 69, Ord, T.J., E.P. Martins, S. Thakur, K.K. Mane, and K. Börner, Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases. Animal Behaviour, pp. 1399 -1413, Copyright 2005, with permission from Elsevier.

genes, and proteins) and burst detection algorithm (to identify bursty researched genes and proteins).

Visualization Technique Used: Network visualization using force-directed algorithm (to show relations) and floating bars (to show genes/proteins and their activity period)

Data Analysis Details: For data analysis, the single term 'melanoma' was used to collect melanoma related publication data from Medline, gene data from Entrez-Gene[5], and protein data from the UniProt[6] database. The initial query resulted in 54,016 papers from Medline (1960 – Feb 2004), 304 genes from Entrez-Gene database, and 566 proteins from UniProt.

The cosine similarity measure was used to calculate similarity between paper-paper, gene-gene, protein-protein, and gene-protein. Using the similarity, an integrated map of papers, genes, and proteins was generated using a force-directed algorithm in VxOrd. The generated map shows the relationship between melanoma related papers, genes, and proteins. To detect the highly researched genes and proteins, titles and MeSH terms were used in the analysis. The burst detection algorithm from Kleinberg was used to identify the bursty words.

User Task Abstraction: The identification of linkages between papers, genes, and proteins would identify relationships between them. Further, a surge in the interest of genes and proteins can be identified by identifying their bursts.

[5] Entrez-Gene can be accessed online at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[6] Uniprot database can be accessed online at: http://www.pir.uniprot.org/

Identified Pathway from DA-Vis Taxonomy:

A) To identify linkages

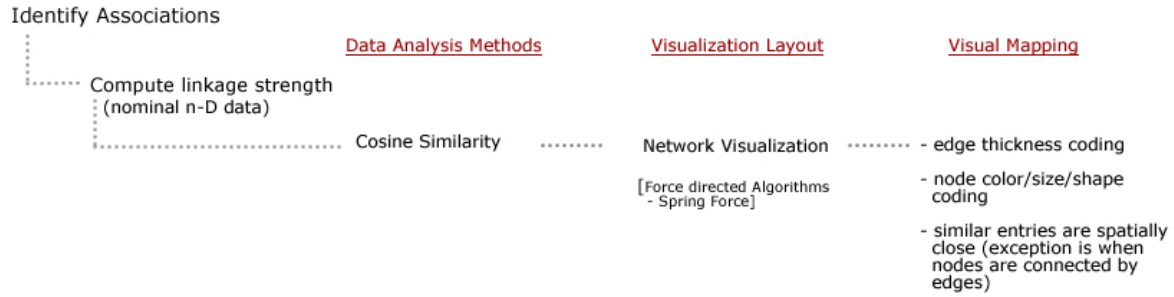


Figure 58: Pathway from DA-Vis taxonomy to identify association based on linkage strength

B) To identify surge of interest

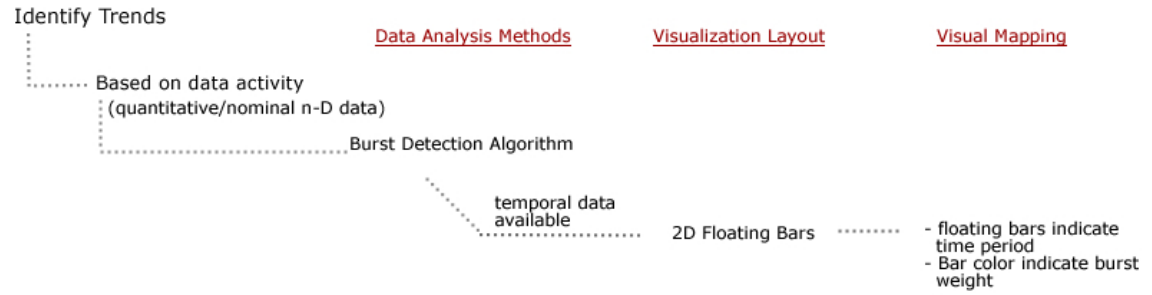


Figure 59: Pathway from DA-Vis taxonomy to identify trends based on data activity

Visual Design Application: The network visualization shows three different node-types – papers (in gray), genes (in blue) and proteins (in red), see Figure 60. It shows the relationship between papers, genes, and proteins. Owing to the layout algorithm property, nodes with higher similarity are placed closer to each other. The edges are not displayed. The floating bar layout is used to indicate the time-period. Figure 61 shows highly researched melanoma related genes along with their burst activity period. Figure 62 shows highly researched melanoma related proteins along with their burst activity period.

Visualizations:

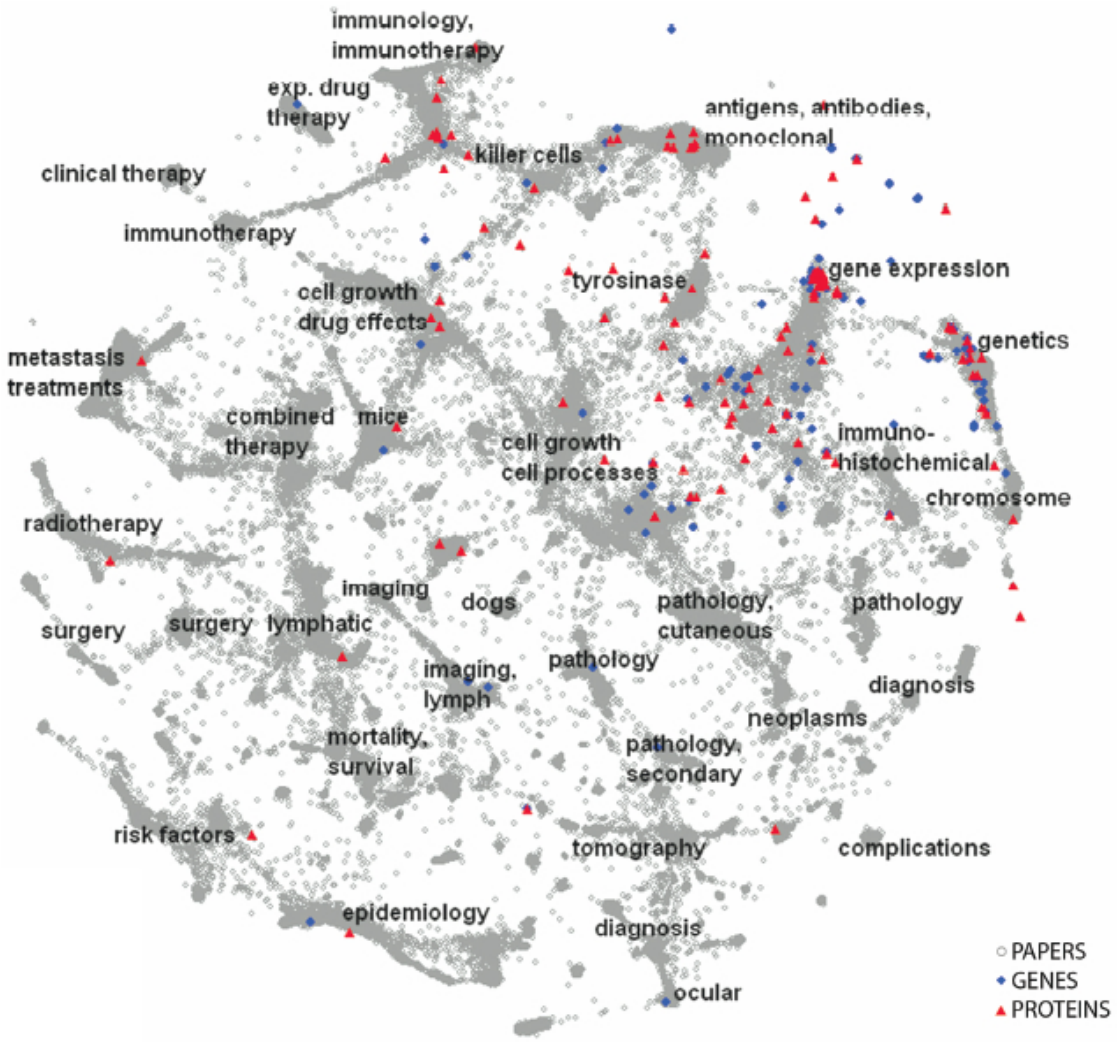


Figure 60: Melanoma paper-gene-protein map, adopted from [104], reproduced with permission from © 2004 IEEE

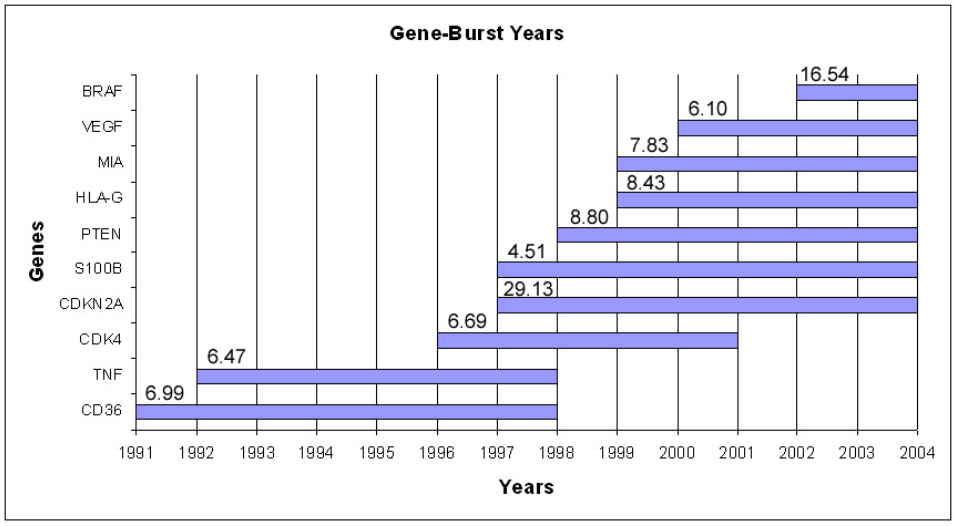


Figure 61: Highly researched melanoma related genes and their activity period, adopted from [104], reproduced with permission from © 2004 IEEE

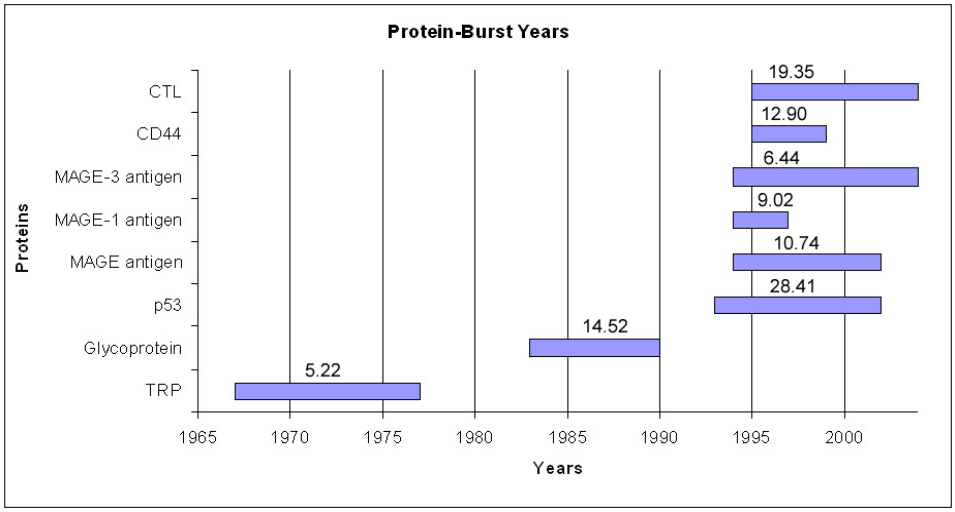


Figure 62: Highly researched melanoma related proteins and their activity period, adopted from [104], reproduced with permission from © 2004 IEEE

7.3. Identify Co-citation Network in Theoretical Physics

Goal: From the co-citation network, clearly identify two main articles that contributed to 'Superstring' theory in theoretical physics. Further, preserve the meaningful co-citation network where the intellectual contribution of these articles exists [105].

Data Analysis Technique Used: Pathfinder network scaling algorithm [35].

Visualization Technique Used: Network Visualization

Data Analysis Details: A co-citation network of 624 articles on superstring theory (1985-2003) was generated. There were two main articles that triggered the development of superstring theory. But from the original network, only one article was clearly seen whereas the other remained hidden. Further, as per the goal, meaningful linkages resulting from these main articles were preserved using a pathfinder network scaling algorithm. The resulting network was visualized using network visualization in CITESPACE tool.

User Task Abstraction: The goal is to preserve important co-citation linkages among authors can be abstracted to preserving important data dimensions.

Identified Pathway from DA-Vis Taxonomy:

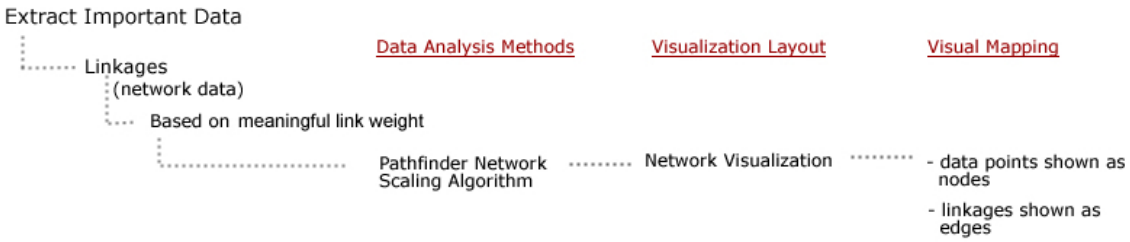
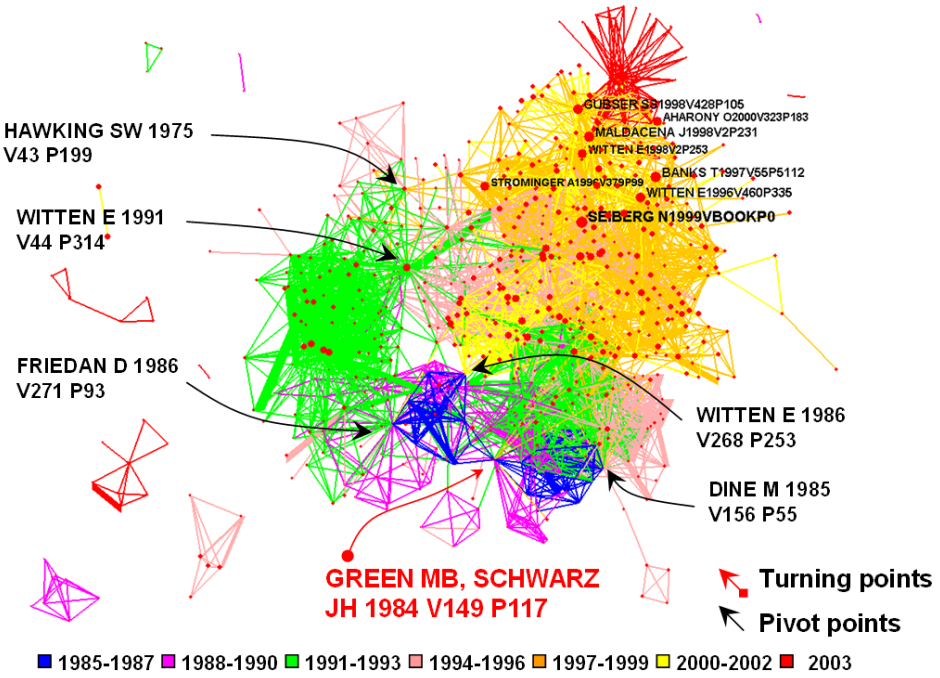


Figure 63: Pathway from DA-Vis taxonomy to identify important data linkages based on meaningful pathways

Visual Design Application: Nodes are used to show papers. The node size and node label size is proportional to the citation count for a given paper. The links are used to show the co-citation coefficient. The color of the link shows the early citation time of an article. Different node label colors are used to indicate turning points, pivot points, and hubs, see Figure 64.

Visualizations:

(A)



(B)

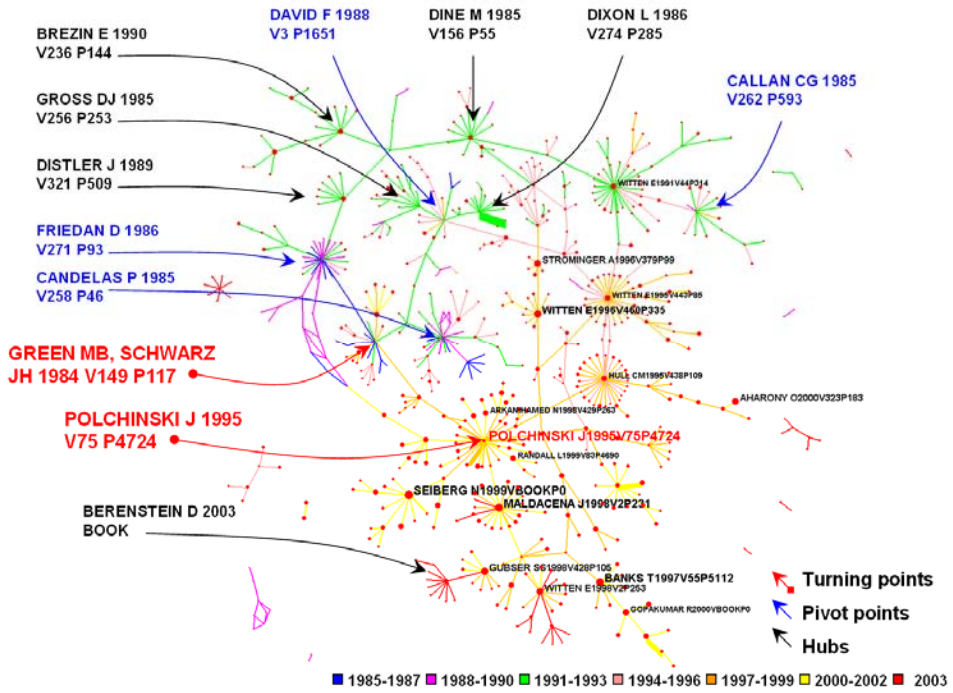


Figure 64: A co-citation network comprising of 624 nodes. (A) original network and (B) network pruned after using pathfinder network scaling, adopted from [105], reproduced with permission from Chaomei Chen

7.4. Protein-protein Interaction in Mouse Genome

Goal: From the protein-protein network in mouse genome, identify the important proteins that are responsible for maintaining interconnectivity in the network [106] .

Data Analysis Technique Used: Betweenness Centrality

Visualization Technique Used: Network Visualization

Data Analysis Details: The protein-protein interaction data for Mus musculus (mouse) was obtained from Database of Interacting Proteins (DIP) [107]. The data was imported into the application called CentiBiN [108] which helps in the calculation and visualization of centralities for biological networks. The central-flow betweenness algorithm [109] was applied to identify the important proteins in the network. The result was visualized using network visualization. Within the network visualization, important proteins are highlighted.

User Task Abstraction: The user task is to identify important proteins that help to maintain connectivity in the network. Connectivity can be maintained if the network flow is maintained. So the task can be abstracted to the identification of proteins that maintain flow in the network structure.

Identified Pathway from DA-Vis Taxonomy:

Figure 65: Pathway from DA-Vis taxonomy to detect structural patterns based on identification od maximum flow nodes

Visual Design Application: In the network visualization, the proteins are shown as nodes. Edges show connections between proteins. The proteins responsible for maintaining the connections in the network are identified using the betweenness centrality algorithm, and are highlighted in red color, see Figure 66.

Visualizations:

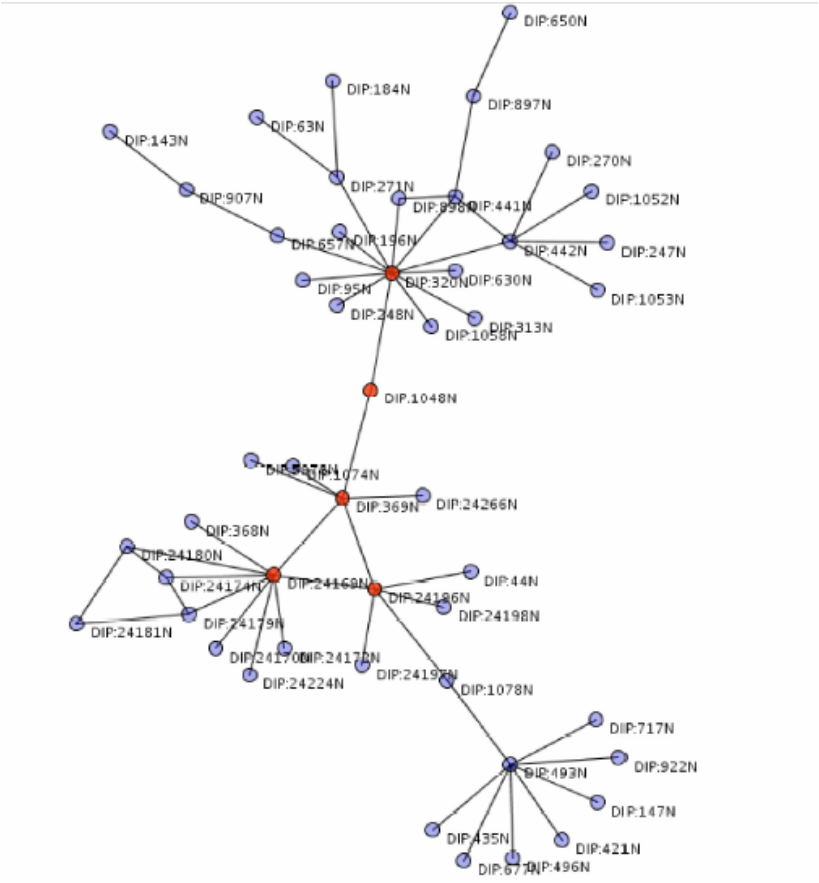


Figure 66: Protein-protein interaction network in Mus musculus (mouse) showing central proteins in red [106], reproduced with permission from Dirk Koschutzki

7.5. Major Author/Scholarly Communities in Complex Network Research

Goal: Based on research topics, identify different author collaboration groups for complex network domain [110].

Data Analysis Technique Used: Co-occurrence Similarity (to identify co-authorship) and Wards clustering (to identify author community clusters).

Visualization Technique Used: Dendrogram

Data Analysis Details: Using seminal papers as query terms, a dataset comprised of papers that cite these seminal paper was put together using ISI's Web of Science. Using co-authorship count as a similarity metric, a wards clustering approach was used to identify author collaboration groups. The relationship among authors was visualized using a modified form of matrix visualization identified by Steven Morris as 'crossmaps'. The author groups were visualized using dendrogram.

User Task Abstraction: The user task is to identify author collaboration groups. The task can be abstracted to identifying associations among authors. Further this information can be used to identify groups of authors that work together on similar research topics.

Identified Pathway from DA-Vis Taxonomy:

A) To identify linkages

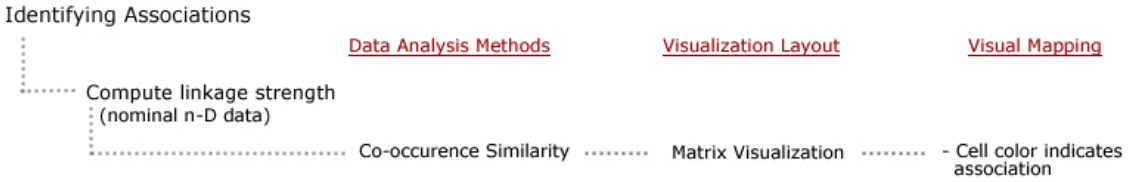


Figure 67: Pathway from DA-Vis taxonomy to identify linkages based in linkage strength

B) To identify clusters

Figure 68: Pathway from DA-Vis taxonomy to identify cluster from data hierarchy

Visual Design Application: Crossmaps, a modified form of matrix visualization is used to show the relationship among authors. The node size is used to indicate the publication count for each author. Using dendrograms, the most closely related authors are connected together to form a cluster, see Figure 69.

Visualizations:

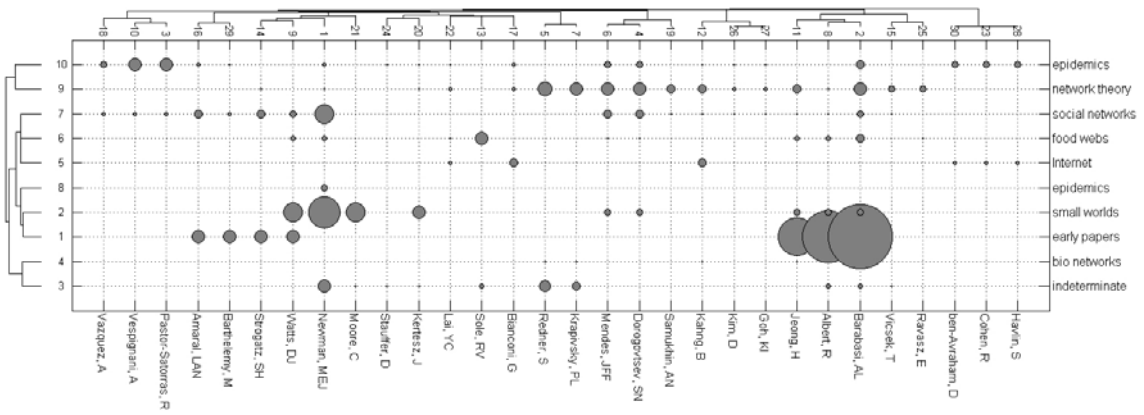


Figure 69: Author collaboration groups shown using dendrogram, used with permission from Steve Morris [110], reproduced with permission from Steven Morris

7.6. Discussion

The above examples of scholarly research demonstrate that the data analysis approach can be mapped onto different pathways in the DA-Vis Taxonomy. Match between the data analysis approach taken and the complementary techniques within pathways helps to validate the DA-Vis Taxonomy. Subsequently, different pathways within DA-Vis taxonomy offer experts and novice an insight into complimentary data analysis and visualization techniques. Using the taxonomy, an expert can quickly identify other techniques that can be used to carry out a similar data analysis. For a novice, the DA-Vis taxonomy offers a global overview of different data analysis and visualization techniques. Only a few features to use pathways from DA-Vis taxonomy are highlighted here, the pathways can also be used for comparison and other purposes.

Chapter 8

Validation: Usability of DA-Vis Taxonomy

A formal usability test was conducted to evaluate the user-friendliness of the DA-Vis taxonomy to naïve and expert users. As the current taxonomy includes techniques from the information visualization domain, only participants who are conversant with data analysis and visualization techniques from information visualization domain were used. To test the DA-Vis taxonomy, 14 participants were recruited. The test subjects received $10 each for participating in the study.

8.1. Methodology

At the start of the test, each participant was asked to fill out a pre-test questionnaire to gather demographic information and data related to profession & expertise. As it is less likely to find subjects who are experts in both the data analysis and data visualization domains, subjects were queried for their level of expertise in data analysis and data visualization. Based on their feedback as 'naïve', 'moderate', or as an 'expert', the total participants were split into two groups. For the purpose of the test, the participants who reported 'naïve' and 'moderate' were considered as novice users. In addition, participants were asked about their familiarity with any kind of taxonomy.

Further to familiarize each user with the different features of the DA-Vis Taxonomy, each user was briefed with information that was encoded at three different levels in the

taxonomy using a simple example. The participants were also informed that the pathway included pieces of information at all of the three levels. Subsequently, each participant was asked to find answers to the task-list questions.

The task-list included eight questions that were based on the data analysis goals that needed to be accomplished. Special emphasis was placed on different sections of the DA-Vis taxonomy. Further, all questions can be classified into two categories: 1) Topical – where direct mapping can be done to user tasks available in the DA-Vis taxonomy (Q#1-4) and 2) Non-topical – where ambiguity will be maintained in specifying the user task goals (Q#5-8). For non-topical questions, the participants would need to take an extra step of identifying the task and then map it to a user task covered in the DA-Vis taxonomy. The purpose behind framing non-topical questions was to determine the user-friendliness of the DA-Vis taxonomy when different words were used to describe user goals. This part tests the amount of additional information that needs to be made available with the task-list questions to use the taxonomy.

The amount of time it takes to answer a single task-list question was used as a measure to evaluate the efficiency. As the taxonomy is concise and easy to scan, a time period of three minutes was considered sufficient to go through the DA-Vis taxonomy and find answers to any task list question. If more than three minutes of time were used to find an answer to a question with/without any hint then it would be considered to negate the user-friendliness of the DA-Vis taxonomy.

The questions given to the participants are listed below:

Topical Questions

1. From the DA-Vis taxonomy, identify one of the pathways that help to establish associations based on semantic linkages. Write it down.

2. From the DA-Vis taxonomy, identify a data analysis method that helps to identify trends in network data. Write it down.

3. Name the visualization technique that can be used to visualize a cluster hierarchy.

4. Name the visualization technique that can be used to visualize a network structure.

Non-topical Questions

5. What data characteristics can be visualized using parallel-coordinates visualization? Write down the pathway in the DA-Vis taxonomy that you used.

6. How can data bursts activity be represented? Write it down.

7. What techniques exist to re-order data?

8. What user goals can 'Latent Semantic Analysis' (LSA) support?

After the task list was completed, a post-test questionnaire was handed to each test subject. The post-test questions were designed to gauge participant's response to the tasks. Questions evaluated: quick identification of utility, ease of use, ability to quickly identify combination of techniques and the ability to find information to complete task. Further they were asked to comment on the features that they liked or dis-liked about the taxonomy. The list of post-test questions is given below:

1. I quickly understood the utility of the DA-Vis taxonomy.

2. The DA-Vis taxonomy is easy to use.

3. The DA-Vis taxonomy helps to quickly identify available techniques based on a user's task.

4. The DA-Vis taxonomy helps combine different data analysis and visualization techniques available for a given user task.

5. It was easy to find the information that I needed to complete the task.

6. What did you like most about the DA-Vis taxonomy?

7. What did you like least about the DA-Vis taxonomy?

8. Do you have any recommendations to improve the DA-Vis taxonomy?

Participants were asked to respond to questions using a Likert scale of 1 (strongly disagree) – 5 (strongly agree).

8.2. Results

The results can be summarized as follows: 10 men and 4 women participated in the usability test. There were eight participants in the '21-30' age group, four participants in the '31-40' age groups, and two belonged to the '41 and above' age group. Except for one individual who had an undergraduate degree, all other participants were working on or had advanced degrees, either MS or Ph.D. To perform the usability test, some familiarity with data analysis and/or data visualization techniques was expected. The analysis revealed that a mixed user-group involved in the study comprised of: four subjects who were naïve to both data analysis and data visualization techniques; the other four reported moderate experience with data analysis and data visualization; two reported mixed experience of naïve and moderate experience with the techniques; and the remaining four reported themselves as experts. So based on our grouping strategy, there were 10 participants who were categorized into the 'naïve' user-group and 4 participants who were categorized into the 'expert' user group. The question 'Have you used any kind of data taxonomy before?' was answered with 'Yes' by 5 participants and 'No' by the remaining 9 participants. The known data taxonomies listed were Rainbow classifier, Legal West topic and key numbers, Lexis legal taxonomy, Library of Congress classification system, United States Patent and Trademark Office (USPTO), Medical Subject Heading (MeSH), and Gene ontology.

Figure 70 shows the average time taken by different participants in answering topical and non-topical questions. Except for Q7, there is no significant difference in the average time that all participants took to answer the task list questions. A big difference is seen in the time taken to answer Q7 as it was confusing for most of the participants to map 're-order word' to 'ordering' at level 2 in the DA-Vis taxonomy. Only five participants out of the total 14 participants asked for a hint to understand Q7 correctly. The hint included pointers to which column in the Da-Vis Taxonomy, they should look to find answer to the task question. Also, from the total participant group only one participant was confused about answering Q8 but managed to find the correct answers after receiving hints about mapping tasks with data goals in the DA-Vis taxonomy.

The post-test questionnaire was analyzed to evaluate the user-friendliness of the DA-Vis taxonomy. The results are presented in Figure 71. It can be seen that on an average, participant's score was above four out of the total score of five for all questions.

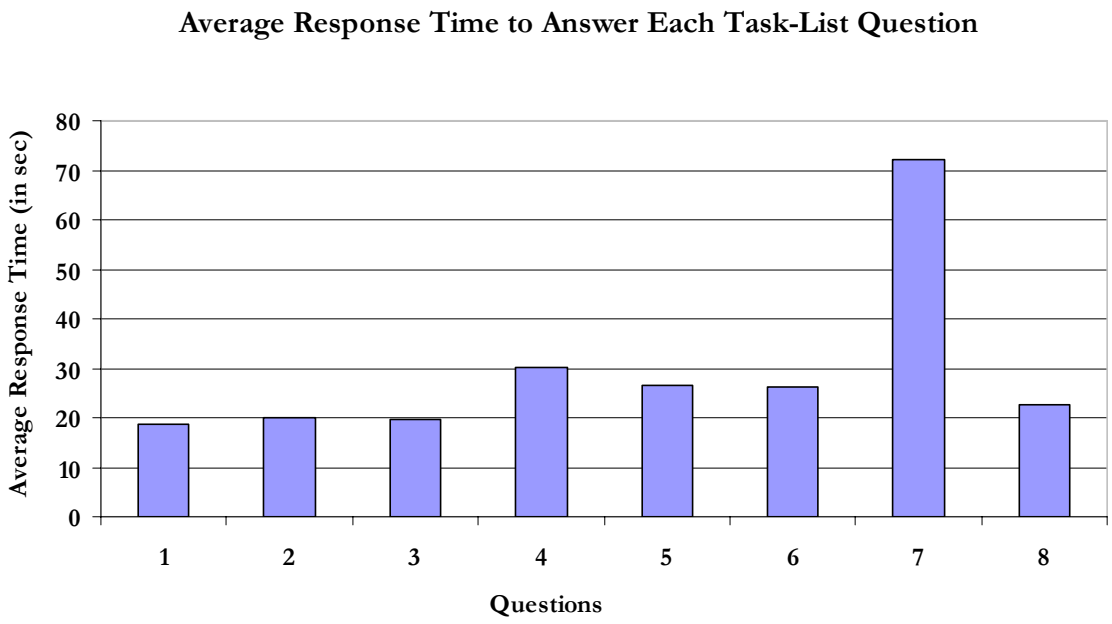


Figure 70: Average response time in sec taken to answer each task list question

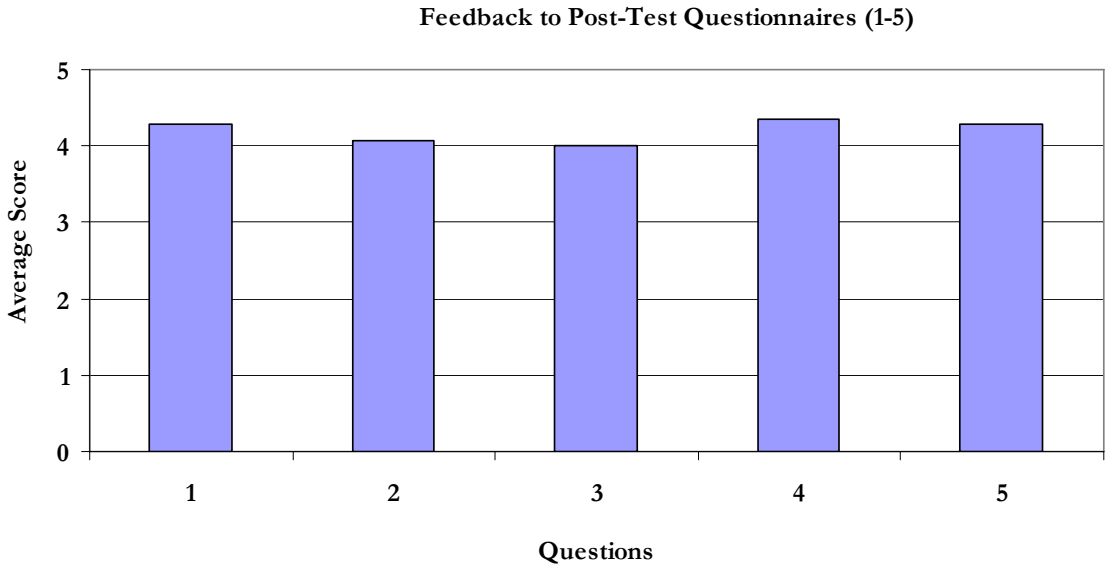


Figure 71: Average score of feedback for Q1-5 for 14 participants

When asked about what they liked about the DA-Vis taxonomy, most participants reported that it offered a good organizational structure to couple analysis and visualization techniques, it offers an easy way to find proper techniques, the clear layout helps to remember different options, the simple structure shows a range of options available for analysis at each level, and it was a good coupling strategy to layout information in hierarchical structure based on user tasks. When asked what they liked least, the participants mentioned the need for visuals for different visualization layout suggested within the taxonomy. Further, because of the paper based version, it was difficult to know if the same technique was available somewhere else within the taxonomy. Recommendations for the DA-Vis taxonomy include: online and interactive version of the taxonomy, color coding of different columns of information, highlighting popular techniques, pointer to different reference resources. In the interest of maintaining the structural simplicity and avoid clutter in the paper-based version of the DA-Vis taxonomy, some recommendations from the participant were not included in the current paper-based version of the taxonomy. An online version of the DA-Vis taxonomy with in-built mouse event would be more suited to include some of the proposed recommendations.

Chapter 9

Intellectual Merits and Broader Impact

This thesis presented a new coupled DA-Vis taxonomy that can be used to ease the design of effective visualizations. The DA-Vis taxonomy advances theory of data analysis and visualization by providing a unified framework that identifies valid algorithm (techniques) combinations. In this thesis, the DA-Vis layout schema has been used to classify different data analysis and visualization techniques that can be applied to n-dimensional data and network data. Similarly, different data types and related algorithms can be plugged into the DA-Vis schema. The schema also shows the tight coupling that exists between different data analysis techniques, and visualization techniques. The DA-Vis taxonomy can be adapted to classify data analysis techniques from different subject domains. Further, it can be used to establish a tight coupling between different data analysis and visualization techniques.

This thesis also demonstrates the utility of the DA-Vis taxonomy in creating meaningful visualizations for biomedical data and scholarly data (chapter 6). Using real world clinical data of acute lymphoblastic leukemia, the DA-Vis taxonomy was used to identify different data analysis and visualizations (matrix view and parallel coordinate view) that help to fulfill the data analysis goals. Further, this thesis also presents a novel data analysis and visualization approach (section 6.1) to correlate clinical data of acute lymphoblastic leukemia patients.

Further, this thesis also applies the DA-Vis taxonomy to support knowledge management (chapter 6). Using a scholarly dataset, the DA-Vis taxonomy was used to identify data analysis and visualization techniques (network visualizations) that best meet the user requirements (section 6.2). The resulting visualization (Figure 55) helps to quickly identify major research topics in the 20-year time slice of PNAS journal data that was considered in the analysis. The generated visualization provides a good overview of major research topics, other domain research areas that are connected, time-based emergence of associated research areas, etc. To our knowledge, this was the first attempt to draw a knowledge map based on co-word space of highly frequent and highly researched bursty topical words. Such knowledge maps also help identify research frontiers, or change of focus in certain areas. Identification of these frontiers forms a critical component for resource allocation decisions in research laboratories, governmental institutions, and corporations. Similar knowledge management maps created for different subject domains can be used for educational purposes. For example, it can be used to study trends in melanoma research.

Further, the thesis validates the proposed DA-Vis taxonomy by showing its usefulness in categorizing and describing prior work (chapter 7). User goals from prior work have been mapped to different pathways from the DA-Vis taxonomy to validate the identified coupling between different data analysis and visualization techniques.

This thesis presents the results of a user study (chapter 8) performed to identify the user-friendliness of the DA-Vis taxonomy. In general, participants commented that the structure of DA-Vis was easy to understand. Further, the simplicity of the DA-Vis taxonomy structure seemed to help users remember different options and helped to quickly identify techniques that would meet their data analysis goals.

The DA-Vis taxonomy can be utilized as an 'external aid' to identify analysis and visualization techniques which can be used to accomplish different user goals. The schema provides a simple and flexible framework where techniques from different subject domains can be added. This flexibility helps to customize the DA-Vis taxonomy to meet the requirements of individual subject domain data analysis requirement.

A naïve user can use the DA-Vis taxonomy to develop a 'learning module'. Based on the task goal and data type, a naïve user can quickly acquire knowledge about different data analysis techniques that can be used to perform data analysis and visualization. Further, with the DA-Vis taxonomy schema, additional information used to map data characteristics onto final visualizations is readily available to the user. Creation of a taxonomy based in the DA-Vis schema can help improve the teaching, training, and learning experience of complementary data analysis and visualization techniques.

The DA-Vis taxonomy schema can be used to guide the selection of algorithms in cyber-infrastructure tools such as NetworkBench, Information Visualization Cyberinfrastructure, etc. Further, an advanced computing environment which allows users to establish a work-flow to perform different data analyses can be created.

The novel data analysis approach shown for a clinical dataset can help in the practice of evidence based medicine. It offers the ability to look at different data variables at a glance, visually identify data patterns, and data trends help to compare different patients. Using these features will benefit the medical practitioners to perform better diagnosis of the patient's medical condition.

A demonstration of the computational diagnostics tool was presented at the National Institute of Cancer (NCI) series on 'Informatics in Action 2006' on 'Finding Patterns in a Sea of Data: How Information Visualization can Support NCI's Fight Against Cancer'. The goal of the series was to improve the outcomes in implementing health, medical, and bioinformatics technologies through the science of user-centered informatics research. Further, a talk and demonstration of the developed computational diagnostic tool was presented at the National Institute of Health (NIH) Interdisciplinary Methodology and Technology Summit (M&T Summit). The researchers were intrigued by the new methodology of looking at clinical data with the developed computational diagnostic tool. Researchers gave positive feedback on the utility of the developed tool. One of the most positive comment about the computational diagnostic tool developed as a part of the thesis was 'the work has developed the seeds of what can probably become a successful way of transforming data visualization into a decision support tool for clinical decisions'.

References

1. Tufte, E., *The Visual Display of Quantitative Information*. 1997, Cheshire, CT: Graphics Press. 156 pages.

2. Keller, P.R. and M.M. Keller, *Visual Cues: Practical Data Visualization*. 1993, Piscataway, NJ: IEEE Press. 229 pages.

3. MacEachren, A.M., *How Maps Work : Representation, Visualization, and Design*. 2004: The Guilford Press. 513 pages.

4. Anderson, J.R., *Architecture of Cognition*. 1983: Harvard University Press. 345 pages.

5. Card, S.K. and J.D. Mackinlay. *The Structure of the Information Visualization Design Space*. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis'97)*. 1997. Pheonix, AZ. pp. 92-99.

6. Fayyad, U., G. Grinstein, and A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*. 2001, San Francisco, CA: Morgan Kaufmann. 407 pages.

7. Russell, D.M., M.J. Stefik, P. Pirolli, and S.K. Card. *The Cost Structure of Sense-making*. In *Proceedings of the SIGCHI conference on Human factors in Computing Systems*. 1993. Amsterdam, The Netherlands. pp. 269-276.

8. Pirolli, P. and S. Card. *Sense-making Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis*. In *Proceeding of the 2005 International Conference on Intelligence Analysis*. 2005. McLean, Virginia. pp. 6.

9. Tufte, E.R., *Envisioning Information*. 1990, Cheshire, Conn.: Graphics Press. 126 pages.

10. Thomas, J.J. and K.A. Cook, eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. 2005, IEEE Press. 200 pages.

11. Patterson, E.S., E.M. Roth, and D.D. Woods, *Predicting Vulnerabilities in Computer-Supported Inferential Analysis Under Data Overload*. Cognition, Technology and Work, 2001. **3**: pp. 224-237.

12. Card, S.K., J.D. Mackinlay, and B. Shneiderman, *Reading in Information Visualization: Using Vision to Think*. 1999, San Francisco: Morgan Kauffman. 712 pages.

13. Keim, D. and M. Ward, *Intelligent Data Analysis: Chapter 11: Visualization*. 2nd ed, ed. M. Berthold and D. Hand. 2003: Springer. 403-427 pages.

14. Chi, E.H. *A Taxonomy of Visualization Techniques Using the Data State Reference Model* In *IEEE Symposium on Information Vizualization* 2000. Salt Lake City, Utah: IEEE Computer Society. pp. 69-76.

15. Data-Analysis, *Wikipedia Reference Page: http://en.wikipedia.org/wiki/Data_Analysis*.

16. Wikipedia, *Reference page on 'Level of measurement': http://en.wikipedia.org/wiki/*.

17. Wikipedia, Reference Page on Natural Language Processing: http://en.wikipedia.org/wiki/Natural_language_processing.

18. Shneiderman, B. *The eyes have it: A task by data type taxonomy for information visualization*. In *Proc. IEEE Symposium on Visual Languages '96*. 1996. Los Alamos, CA: IEEE. pp. 336-343.

19. Callon, M., J. Law, and A. Rip, *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* 1986, London: Macmillan. 260 pages.

20. Salton, G., *Introduction to Modern Information Retrieval*. 1983, Auckland: McGraw-Hill. 448 pages.

21. Wikipedia, *Reference Page on: http://en.wikipedia.org/wiki/Jaccard_index*.

22. Wikipedia, *Reference Page on Pearson Correlation.* http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

23. *Pearson Correlation Reference Page.* http://www.vias.org/tmdatanaleng/cc_corr_coeff.html.

24. Pudovkin, A.I. and E. Garfield, *Algorithmic procedure for finding semantically related journals*. Journal of the American Society for Information Science and Technology, 2002. **53**(13): pp. 1113 -1119

25. Boyack, K., R. Klavans, and K. Börner, *Mapping the backbone of science*. Scientometrics, 2005. **64**: pp. 351-374.

26. Wikipedia, *Reference Page on Term Frequency - Inverse Document Frequency (TF-IDF)`.* http://en.wikipedia.org/wiki/Tf-idf.

27. Korfhage, R.R., *Information Storage and Retrieval*. 1997, Canada: John Wiley and Sons, Inc. 349 pages.

28. McQueen, J.B. *Some methods of classification and analysis of multivariate observations*. In *Proceedings of 5th Berkeley Symp. on Mathematical Statistics and Probability*. 1967 pp. 281-297.

29. Ward, J.H., *Hierarchical Grouping to optimize an objective function*. Journal of American Statistical Association, 1963. **58**(301): pp. 236-244.

30. Berry, M., S.T. Dumais, and G.W. O'Brien, *Using linear algebra for intelligent information retrieval*. SIAM: Review, 1995. **37**(4): pp. 573-595.

31. Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, *Indexing By Latent Semantic Analysis*. Journal of the American Society For Information Science, 1990. **41**: pp. 391-407.

32. *InfoVis Cyberinfrastructure, Multi-Dimensional Scaling Reference Page: .* http://iv.slis.indiana.edu/sw/mds.html.

33. Wikipedia, *Reference Page on Multi-Dimensional Scaling.* http://en.wikipedia.org/wiki/Multidimensional_scaling.

34. Kleinberg, J.M. *Bursty and hierarchical structure in streams*. In *In 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*. 2002. Edmonton, Alberta, Canada: ACM Press. pp. 91-101

35. Schvaneveldt, R., *Pathfinder Associative Networks: Studies in Knowledge Organization*. 1990, Norwood, NJ: Ablex Publishers. 240 pages.

36. Chen, C., *Information Visualization: Beyond the horizon*. 2004: Springer. 304 pages.

37. Newman, M.E.J. and M. Girvan, *Finding and evaluating community structure in networks*. Phys. Rev. E., 2004. **69**(026113): pp. 15.

38. Newman, M.E.J., *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality*. Physical Review E, 2001. **64**(016132).

39. Girvan, M. and M.E.J. Newman, *Community structure in social and biological networks*. PNAS. USA, 2002. **99**: pp. 7821-7826, also available arXiv.cond-mat/0112110v1.

40. Pastor-Satorras, R. and A. Vespignani, *Epidemics and Immunization in Scale Free Networks in Handbook of Graphs and Networks: From the Genome to the Internet*, ed. S.B.a.H.G. Schuster. 2003: Wiley-VCH. 111-130 pages.

41. Dezso, Z. and A. Barabasi, *Halting viruses in scale-free networks*. Phys Rev E Stat Nonlin Soft Matter Phys, 2002. **65**(055103): pp. 4.

42. Barabási, A.-L. and Z.N. Oltvai, *Network Biology: Understanding the cell's functional orgainization*. Nature Reviews - Genetics, 2004. **5**: pp. 101-114.

43. Lewis, D.D., *Text representation for intelligent text retrieval: A classification-oriented view.* Text-based intelligent systems: Current research and practice in information extraction and retrieval, ed. P.S. Jacobs. 1992, Hillsdale, NJ: Erlbaum. 179-197 pages.

44. Fry, B., *Computational Information Design*, in *MIT Media Lab.* 2004, MIT: Boston.170 pages.

45. Wikipedia, *Reference Page on Scientific Visualization:* *http://en.wikipedia.org/wiki/Scientific_Visualization*.

46. Treinish, L., *Ozone Animation: IBM.* 1994.

47. *National Center for Atmospheric Research (NCAR) - Scientific Visualizations:* *http://www.vets.ucar.edu/vg/index.shtml*.

48. Skupin, A., *The World of Geography: Visualizing a Knowledge Domain with Cartographic Means.* Proceedings of the National Academy of Sciences, 2004. **101**(Suppl. 1): pp. 5274-5278.

49. Börner, K. and S. Penumarthy. *Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions.* In *Proceedings of ISSI 2005*. 2005. Stockholm: Karolinska University Press. pp. 635-641.

50. Harris, R.L., *Information Graphics: A Comprehensive Illustrated Reference: Visual Tools for Analyzing, Managing, and Communicating.* 1999, New York: Oxford University Press. 448 pages.

51. Chernoff, H., *The Use of Faces to Represent Points in k-Dimensional Space Graphically.* Journal American Statistical Association, 1973. **68**: pp. 361-368.

52. *Selected Topics in Graphical Analytic Techniques - Brief overviews of types of graph:* *http://www.statsoft.com/textbook/stgraph.html#icon%20plots*.

53. Ahlberg, C. and B. Shneiderman. *Visual Information Seeking Using the FilmFinder.* In *Conference Companion on Human factors in Computing Systems (CHI'94).* 1994. Boston, Massachusetts: ACM Press, NY. pp. 433-434.

54. Inselberg, A. and B. Dimsdale. *Parallel coordinates: A tool for visualizing multi-dimensional geometry* In *Proceedings of visualization '90*. 1990. San Francisco: IEEE Press. pp. 361-370.

55. Paley, B., *TextArc Website:* *http://www.textarc.org/*pages.

56. Eick, S.G., J.L. Steffen, and E.E. Sumner, *Seesoft-A tool for visualizing software.* IEEE Transactions on Software Engineeering, 1992. **18**: pp. 957-968.

57. Keim, D.A., J. Schneidewind, and M. Sips. *CircleView: a new approach for visualizing time-related multidimensional data sets.* In *Proceedings of the working conference on Advanced visual interfaces.* 2004. Gallipoli, Italy. pp. 179-182

58. Rao, R. and S.K. Card. *The Table Lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information*. In *Proceedings CCHI'94 Conference on Human Factors in Computing Systems*. 1994. New York: ACM. pp. 318-322.

59. Munzner, T., *Interactive visualization of large graphs and networks*, in *Computer Science*. 2000, Stanford University (Thesis).167 pages.

60. Herman, I., G. Melançon, and M.S. Marshall, *Graph Visualization and Navigation in Information Visualization: a Survey*. IEEE Transactions on visualization and computer graphics, 2000. **6**: pp. 1-21.

61. Yee, K.P., D. Fisher, R. Dhamija, and M. Hearst. *Animated exploration of dynamic graphs with radial layout*. In *IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*. 2001. San Diego, CA IEEE Press. pp. 43-50.

62. Albert, R. and A.-L. Barabási, *Statistical mechanics of complex networks*. Review of Modern Physics, 2002. **74**: pp. 47-97.

63. Redner, S., *How popular is your paper? An empirical study of the citation distribution*. European Phys. J., 1998. **B4**: pp. 131-134.

64. Watts, D.J. and S.H. Strogartz, *Collective dynamics of small-world network*. Nature, 1998. **393**: pp. 440-442.

65. Faloutsos, M., P. Faloutsos, and C. Faloutsos, *On Power-Law Relationships of the Internet Topology*. Comp. Comm. Rev., 1999. **29**: pp. 251-262.

66. Wuchty, S., E. Ravasz, and A.-L. Barabasi, *The Architecture of Biological Networks*. Complex Systems in Biomedicine, ed. T.S. Deisboeck, J.Y. Kresh, and T.B. Kepler. 2003, New York: Kluwer Academic Publishing. - pages.

67. Wikipedia, *Reference Page on Dendrogram*. http://en.wikipedia.org/wiki/Dendrogram.

68. Boyack, K.W., B.N. Wylie, and G.S. Davidson, *Domain visualization using VxInsight for science and technology management*. Journal of the American Society for Information Science and Technology, 2002. **53**: pp. 764-774.

69. Boyack, K.W., B.N. Wylie, G.S. Davidson, and D.K. Johnson. *Analysis of patent databases using VxInsight*. In *New Paradigms in Information Visualization and Manipulation*. 2000. McLean, VA. pp. 7.

70. Boyack, K.W. and K. Börner, *Indicator-Assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers*. Journal of the American Society of Information Science and Technology, Special Topic Issue on Visualizing Scientific Paradigm, 2003. **54**(5): pp. 447-461.

71. Shneiderman, B., *Treemaps for space-constrained visualization of hierarchies, www.cs.umd.edu/hcil/treemaps*.

72. Fiore, A. and M. Smith, *Tree Map Visualizations of Newsgroups*. 2001.

73. Chi, E.H., *A Framework for Information Visualization Spreadsheets*. 1999, University of Minnesota: Minnesota.160 pages.

74. Battista, G., P. Eades, R. Tamassia, and I.G. Tollis, *Algorithms for drawing graphs: An annotated bibliography*. Computational Geometry : Theory and Applications, 1994. **4**(5): pp. 235-282.

75. Battista, G., P. Eades, R. Tamassia, and I.G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*. 1999: Prentice-Hall Inc. 397 pages.

76. Eades, P., *A heuristic for graph drawing*. Congressus Numerantium, 1984. **42**: pp. 149-160.

77. Kumar, A. and R.H. Fowler, *A Spring Modeling Algorithm to Position Nodes of an undirected Graph in Three Dimensions*. Technical Report Department of Computer Science, University of Texas - Pan American. http://bahia.cs.panam.edu/info_vis/spr_tr.html, 1996.

78. Mehlhorn, K. and S. Näher, *LEDA: a platform for combinatorial and geometric computing*. Communications of the ACM, 1995 **38**(1): pp. 96-102.

79. Fruchterman, T.M.J. and E.M. Reingold, *Graph Drawing by Force-Directed Placement*. Software-Practice & Experience, 1991. **21**(11): pp. 1129-1164.

80. Kamada, T. and S. Kawai, *An algorithm for drawing general undirected graphs*. Information Processing Letters, 1989. **31**(1): pp. 7-15.

81. Leydesdorff, L., *Clusters and Maps of Science Journals Based on Bi-connected Graphs in the Journal Citation Reports*. Journal of Documentation, 2004. **60**(4): pp. 371-427.

82. User Manual, *Algorithms for Graph Drawing*: http://www.ads.tuwien.ac.at/AGD/pages.

83. Alberts, D., C. Gutwenger, P. Mutzel, and S. Naher. *AGD-library: a library of algorithms for graph drawing*. In *Proceedings of 1st Int. Workshop on Algorithm Engineering (WAE 97)*. 1997. pp. pages 112-123.

84. Ware, C., *Information Visualization, Second Edition: Perception for Design*. 2004: Morgan Kaufmann. 486 pages.

85. Palmer, S.E., *Vision Science: Photons to Phenomenology* 1999: MIT Press. 760 pages.

86. Ahlberg, C., C. Williamsons, and B. Shneiderman. *Dynamic queries for information exploration: An implementation and evaluation*. In *Proceeding of ACM CHI'92: Human Factors in Computing Systems*. 1992. New York: ACM. pp. 619-626.

87. North, C. and B. Shneiderman. *Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata.* In *Proc. AVI 2000*. 2000. Palermo, Italy. pp. 1-9.

88. Mackinlay, J.D., G.G. Robertson, and S.K. Card. *The Perspective Wall: Detail and context smoothly integrated.* In *In Proceedings of CHI' 91*. 1991. New York: ACM. pp. 173-179.

89. SPENCE, R. and M.D. APPERLEY, *Database navigation: An office environment for the professional.* Behau. Inf Tech, 1982. **1**(1): pp. 43-54.

90. Bertin, J., *Graphics and Graphic Information Processing.* Walter de Gruyter, 1977/1981.

91. Tweetie, L. *Characterizating Interactive Externalization.* In *Proceedings of the ACM CHI 97 Human Factors in Computing Systems Conference.* 1997. Atlanta, GA: ACM. pp. 375-382.

92. *OLIVE: On-Line Library of Information Visualization Environments.* http://otal.umd.edu/Olive, 1999.

93. Chi, E.H. and J.T. Riedl. *An Operator Interaction framework for visualization systems.* In *Symposium on Information Visualization (InfoVis' 98)*. 1998. Research Triangle Park, North Carolina. pp. 63-70.

94. Chi, E.H., *A framework for information visualization spreadsheets.* March 1999, University of Minnesota: Minnesotapages.

95. Ellson, J., E.R. Gansner, E. Koutsofios, S.C. North, and G. Woodhull, *Graphviz and Dynagraph - Static and Dynamic Graph Drawing Tools*, in *Graph Drawing Software*, M. Junger and P. Mutzel, Editors. 2003, Springer-Verlag. pp. 127-148.

96. Becker, R.A., S.G. Eick, and A.R. Wilks, *Visualizing Network Data.* IEEE Transactions on Visualizationand Computer Graphics, 1995. **1**(1): pp. 16-21.

97. Keim, D., *Information Visualization and Visual Data Mining.* IEEE Transactions on Visualization and Computer Graphics, 2002. **7**(1): pp. 100-107.

98. Swayne, D.F., D. Cook, and A. Buja, *Xgobi: interactive dynamic data visualization in the X window system.* Journal of Computational and Graphical Statistics, 1998. **7**: pp. 113-130.

99. Sarkar, M. and M.H. Brown, *Graphical fisheye views.* Communcations of the ACM, 1994. **37**(12): pp. 73-84.

100. Ledermann, F., *URL: http://home.subnet.at/flo/mv/parvis/index.html.*

101. Wikipedia, *Reference Page on Evidence Based Medicine (EBM).* http://en.wikipedia.org/wiki/Evidence_based_medicine.

102. Mane, K.K. and K. Börner, *Mapping Topics and Topic Bursts in PNAS.* Proceedings of National Academy of Science, 2004. **101**(Suppl. 1): pp. 5287-5290.

103. Ord, T.J., E.P. Martins, S. Thakur, K.K. Mane, and K. Börner, *Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases.* Animal Behaviour, 2005. **69**: pp. 1399-1413.

104. Boyack, K.W., K. Mane, and K. Börner. *Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research.* In *Eighth International Conference on Information Visualisation (IV'04).* 2004. London, UK. pp. 965-971.

105. Chen, C., *Searching for intellectual turning points: Progressive knowledge domain visualization.* Proceedings of National Academy of Science, 2004. **101**(Suppl.1): pp. 5303-5310.

106. Junker, B.H., D. Koschuetzki, and F. Schreiber, *Exploration of biological network centralities with CentiBiN.* BMC Bioinformatics, 2006. **7**(219): pp. 1-14.

107. Salwinski, L., C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg, *The Database of Interacting Proteins.* Nucleic Acids Res, 2004. **32**: pp. 449-451.

108. *Centrality in Biological Networks (CentBiN), availabe for download at: http://centibin.ipk-gatersleben.de/*.

109. Brandes, U. and D. Fleischer. *Centrality Measures Based on Current Flow.* In *In Proc. 22nd Symp. Theoretical Aspects of Computer Science (STACS 05), In Lecture Notes in Computer Science (LNCS).* 2005: Springer-Verlag pp. 533-544.

110. Morris, S.A. and G.G. Yen, *Crossmaps: Visualization of overlapping relationships in collections of journal papers.* Proceedings of National Academy of Science, 2004. **101**(Suppl. 1): pp. 5291-5296.

111. Donadieu, J., M.-F. Auclerc, A. Baruchel, Y. Perel, P. Bordigoni, J. Landman-Parker, T. Leblanc, G. Cornu, D. Sommelet, G. Leverger, G. Schaison, and C. Hill, *Prognostic study of continuous variables (white bloodcell count, peripheral blast cell count, haemoglobinlevel, platelet count and age) in childhood acutelymphoblastic leukaemia. Analysis of a population of1545 children treated by the French AcuteLymphoblastic Leukaemia Group (FRALLE)* British Journal of Cancer, 2000. **83**(12): pp. 1617-1622.

Appendix I: Glossary

Tree Layouts	A tree is a connected graph without any cycles
Graph	A set of items connected by edges. Each item is called a vertex or node. Formally, a graph is a set of vertices and a binary relation between vertices
Planar	Graph with no overlapping lines
Automatic Processing	It is based on the visual properties such as position and color. It is highly parallel, but limited in power [5]
Controlled Processing	It works on example text. It has a powerful operations, but is limited in capacity [5]
Relapse	A falling back into a former state, especially after apparent improvement
Isomorphic graphs	Two graphs with one-to-one mapping from the nodes of one graph to the nodes of the other that preserves the adjacency of nodes

Appendix II: Questionnaire About Commonly Used Data Analysis & Visualization Algorithms

The questionnaire that was used to acquire information about various data analysis techniques used by researchers to accomplish data analysis goals is included below.

Informal Questionnaire

The goal of the questionnaire is to identify different algorithms used by researchers to perform various data analysis tasks. Your feedback provides invaluable input to my Ph.D. research. It would help to identify trends of algorithm usage by researchers to accomplish a given task.

Listed below are some common data analysis tasks (a-g). Below each task is a list of algorithms that can be used to gather different task relevant information. From the list, please check those algorithms that you find relevant. Please add to the list, any other algorithms that are missing.

Thank you for your time
Ketan Mane

a) Identify Associations

1. ***Based on linkage strength*** – information can be obtained through data analysis approaches like:
   - ☐ Co-occurrence matrix
   - ☐ Cosine-similarity

- ☐ Pearson coefficient
- ☐ Jaccard index

- o *Any other methods:*

2. ***Based on semantic association*** – information can be obtained through data analysis approaches like:
   - ☐ Term Frequency matrix
   - ☐ Latent Semantic analysis and Similarity Matrix

   - o *Any other methods:*

b) Identify Patterns - information can be obtained through data analysis approaches like:

- ☐ Sorting
- ☐ Spectral analysis
- ☐ Statistical correlation techniques
- o *Any other methods:*

c) Predict Trends

- ☐ Frequency distribution
- ☐ Burst detection algorithm

- o *Any other methods:*

d) Extract Important Data Dimensions

- ☐ Latent semantic analysis (LSA)

- o *Any other methods:*

e) Extract Important Linkages

- ☐ PathFinder network
- ☐ Betweenness centrality

- ☐ Applying thresholds

- o *Any other methods:*

e) Identify Clusters

- ☐ Latent semantic analysis (LSA)
- ☐ k-means algorithms
- ☐ Wards clustering algorithm
- ☐ Network analysis measure – Structural holes
- ☐ Network analysis measure – Structural equivalence

- o *Any other methods:*

f) Classify Data

- ☐ Simple data grouping

- o *Any other methods:*

g) Detect Structural Patterns – for network data types where linkage information is known.

1. **Detect topology**
   - ☐ Detect hubs – degree distribution
   - ☐ Detect scale free topology – power law exponent
   - ☐ Central nodes – betweenness centrality
   - o *Any other methods:*

Appendix III: Acute Lymphoblastic Leukemia – Dataset Variable Information

Variable information from acute lymphoblastic leukemia medical dataset are covered here,

- ***Relapse -*** gives information about the time to relapse from the diagnostic date.
- ***Relapse Site -*** in what body region did the relapse occur – bone marrow (BM), CNS, Testis or any combination of above sites.
- ***Death -*** information whether the person is alive or dead.
- ***LDKA -*** the date the person was last known to be alive.
- ***Immunopheno -*** immunophenotype conditions (B/T-cell) responsible for causing ALL.
- ***Genetic Category -*** includes information on the chromosomal conditions. Different known conditions under genetic category include: hypodiploid (<45), diploid (46), hyperdiploid (>46) and pseudodiploid.
- ***WBC -*** counts number of white blood cells in $1000/mm^3$ of blood.
- ***HGB -*** counts haemoglobin content in g/dl of blood.
- ***PLT -*** counts number of red blood cell in $1000/mm^3$ of blood.
- ***CNS diseases -*** information on different stages of CNS disease is available. The CNS condition can be CNS1a, CNS1b, CNS1c, CNS2a, CNS2b, CNS3.
- ***BM7 -*** blast % at day 7 is recorded.
- ***BM14 -*** blast % at day 14 is recorded.
- ***AgeDx -*** this value is referred to as the diagnostic age of the patient.
- ***Gender –*** Male/Female values are recorded.
- ***Race -*** patient race is an important element in the diagnostic factors as some races have less probability to get cancer than the other.
- ***Median Family Income (MFI) -*** based on the income, there exist four different MFI classes - 1: Very low income (<50% MFI), 2: Low income (50-80% MFI), 3: Moderate income (80 – 120% MFI) and 4: Upper income (> 120% MFI).

- ***Education*** **-** there exist four different categories for this variable – 1: No high school diploma, 2: High school diploma, 3: Some college and 4: College graduate.
- ***% Single Family*** **-** % record of single members living with patients.
- ***% Employment*** **-** % record of employed parents in the patient's family.

Appendix IV: Acute Lymphoblastic Leukemia – Hazard Ratio Conditions for Phenotype Display

Researchers in the biomedical domain have studied the vulnerability relationship between certain variables of the patients [111]. For blood count information, values of haemoglobin level, WBC count, peripheral blast cell count and platelets count were taken into account.

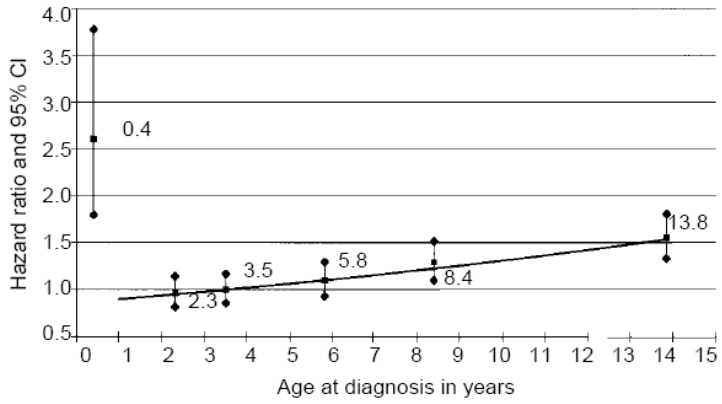


Figure 72: Potential hazard ratio % identified based on the age of diagnosis (years) of patients, adopted from [111]

Figure 72 shows the percentage risk for patients based on their age of diagnosis when the ALL condition is 'B-cell'. From Figure 72, it is clear that the children below one year of age are at a risk that is 2.6 times higher than that of the patients between 3 and 4.3 years of age of diagnosis [111]. As the age increases, the hazard ratio % also increases almost linearly. There is an increase of just 0.5% in hazard ratio for the diagnostic age range of 2-14 years. The risk in the highest quintile is 1.6 times higher than in the lowest quintile. There is no difference in the age of diagnosis outcome for 'T-cell' ALL.

For WBC's, the hazard ratio is computed based on the count of WBC's in one cubic-cm ($1000/mm^3$) of blood. The hazard ratio % or risk of relapse increases exponentially with

increase in the count of WBC's. Figure 73 shows the % hazard ratio increase that has been identified based on the number of WBC's (on logarithmic scale). From the figure, it can be deduced that the risk of relapse in highest quintile is 1.9 times higher than in lowest quintile. The NCI cut-point shows the threshold in the count set by the National Cancer Institute.

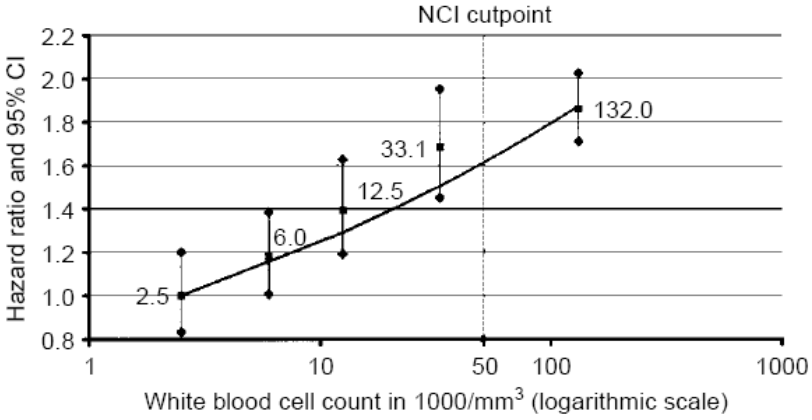


Figure 73: Potential hazard ratio % identified based on the WBC count, adopted from [111]

The peripheral blast cells numbers strongly correlate with the count of WBC. This is because WBC's represent a large portion of peripheral blast cells. Hence, the prognostic curve information for peripheral blast cells is similar to that of the WBC (same curve as shown in Figure 73). The hazard ratio % or risk of relapse for HGB is 1.3 times higher than the lowest quintile, shown in Figure 74. However the risk of relapse decreases when there is an increase in the platelets count. Figure 75 shows the risk in lowest quintile is l.2 times than that in the highest quintile.

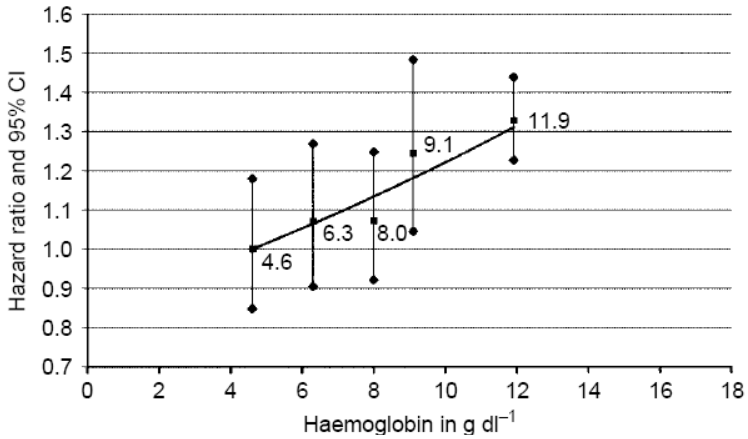


Figure 74: Potential hazard ratio % identified based on the HGB count, adopted from [111]

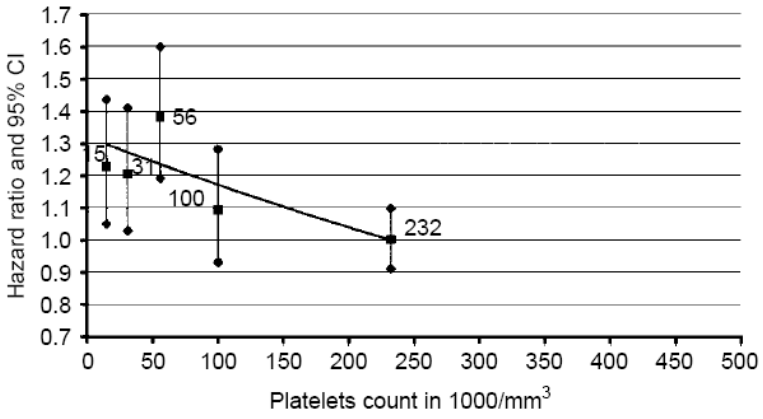


Figure 75: Potential hazard ratio % identified based on the Platelet count, adopted from [111]

Appendix V: Acute Lymphoblastic Leukemia – %EFS Value Conditions for Prognosis Display

Another prognostic factor that can influence the occurrence of relapse is the measurement of event free survival (EFS). It is computed as a percentage value which gives us an idea about the condition of the patient being free from encountering a relapse condition. A higher percentage value indicates good condition of the patient i.e., it is less likely for the patient to go into a state of relapse.

The percentile values for EFS are determined based on immunophenotype condition (B-cell/T-cell). EFS values for some variables that are important for diagnosis are available and hence can be used in the prognosis. The EFS values that are available for different variables are discussed below,

- ***Relapse*** **-** The % EFS values for relapse condition is dependent upon the site of relapse. The EFS values for different sites of relapse are tabulated in Table 4. From the table, it is clear that the relapse condition is more likely to occur in patient where the relapse site is 'Bone Marrow (BM)'. A combination of sites of occurrence of relapse is the second worse condition. Patients with relapse site CNS and Testis have relatively lower chances of encountering a relapse as compared to patients with relapse conditions BM and combination. From Table 4, it is clear that patients with relapse site 'Testis' have a better chance of recovery without undergoing a relapse.

Table 4: %EFS values based on relapse site condition

Time Period	Site of Relapse			
	BM	**CNS**	**Testis**	**Combination**
0 - 17 months:	6	33	52	9
18 - 36months	11	59	57	11
> 36 months	43	72	81	49

- ***Relapse Site-*** If the relapse site condition is considered explicitly apart from the relapse (months) information, then patients with Testis as relapse site have better chances of recovery than the patients with relapse site as bone marrow. The %EFS values are: BM (20%), CNS (48%) and Testis (70%).

- ***Immunophenotype -*** Based on T-lineage or B-lineage immunophenotype condition, the %EFS value for B-cell lineage (80%) is greater than the T-cell lineage (67%). However, for an infant (< 1 year), the B-cell lineage has the worst EFS value of 38%.

- ***Genetic Category -*** EFS percent values are available for patients with diagnostic age profile value that is greater than 10, WBC count > 50,000 and treated by protocol #1882 followed by protocol#1961. Based on the B-cell/T-cell lineage, the %EFS values for different genetic conditions are tabulated in Table 5 below.

Table 5: %EFS values based on genetic category condition

	Immunophenotype	
	B-Cells	**T-cells**
Genetic Categories		
Hypodiploid:	40	40
Diploid:	75	80
Pseudodiploid:	70	73
Hyperdiploid:	72	70
Duplication:	75	-

- ***Genetic Structural Changes*** **–** includes information about duplications in the chromosome. The structural condition is considered worse, if there is a B-cell condition and a chromosomal change at loci '20'. The last row is Table 5 shows that %EFS value for patients with B-cell and structural change in chromosome at position 20 is equal to 75%.

- ***#WBC*** **–** For B-cell condition, higher values for %EFS are an indicators that the patient is likely to relapse. The %EFS values for B-cell conditions are shown in Table 6 below. But in case of a T-cell condition, the reverse condition holds true. The %EFS values for different range of #WBC (in thousands) are indicated here: 10 – 50 (EFS = 60%), > 50 (EFS = 50%) and < 10 (EFS = 38%).

Table 6: %EFS values based on WBC count

B-cells		T-cells	
(Count) x 1000	**%EFS**	**Count(in K)**	**%EFS**
0 -10	89%	10-50	60
10 - 2	69	> 50	50
25 - 50	60	< 10	38
50 -100	53		
>100	47		

- ***Platelets*** **-** Patients profiles which have a higher count of platelets are less prone to undergo a relapse condition. Different ranges of platelet count are indicated with the %EFS values in the adjoining brackets: <50 (EFS = 65 %), 50-100 (EFS = 72%), 100-200 (EFS = 76%), and > 200 (EFS = 80%).

- ***CNS diseases*** **-** For the CNS condition, irrespective of the subtype condition (a/b/c) the %EFS is high for CNS 1 (EFS = 80%) and CNS 2 (EFS = 80%) conditions. Patients with CNS 3 (EFS = 50%) condition are more likely to undergo a relapse condition.

- ***BM7*** **-** For %blast condition at day 7, the M3 condition has the worst %EFS value. The different values for different categories are: M1 (EFS= 80%), M2 (EFS = 61%), and M3 (EFS = 44%).
- ***BM14*** **-** For %blast condition at day 7, the M3 condition has a lower %EFS value than for the BM7 condition. The different values for different categories are: M1 (EFS= 60%), M2 (EFS = 44%) and M3 (EFS = 29%).

- ***Protocol*** **-** Patients being treated with selective protocol# have a better chance of avoiding a relapse condition. Different protocol # adopted in the treatment along with the %EFS values are indicated here: 1953 (EFS = 35%), 1891 (EFS = 83%), 1881

(EFS = 79%), 1882 (EFS = 78%), 1922 (EFS = 83%), 1952 (EFS = 83%), 1991 (EFS = 83 %), 1961 (EFS = 80%).

- ***AgeDx -*** Similar trend to the one seen in Figure 72 (ageDx v/s %hazard ratio plot) is seen with the %EFS values. The age range and their corresponding %EFS values for B-cell conditions are indicated in Table 7 below. The infants below one year have the worst %EFS values. For T-cell condition irrespective of the age group the EFS value equals 60%.

Table 7: %EFS values based on AgeDx values

AgeDx Range	%EFS value
0-1	36
1-5	80
5-10	73
10-15	53
15-20	44

- ***Gender -*** The %EFS values are indicated in Table 8 for both genders. For both B-cell/T-cell conditions, it is seen that the boys are more susceptible to relapse than girls.

Table 8: %EFS values based on patient's gender

Gender	B-cell	T-cell
Girls	76	63
Boys	66	47

- ***Race -*** The race of the patient is also one of the factors that determine the %EFS values. Based on the general observation, Caucasian race (EFS = 82 %) is less susceptible to relapse than any other race. The values for other races are indicated

here: African American (EFS = 72%), Hispanic (EFS = 69 % [high risk]), and for any other race (EFS = 73%).

All the variables for which the EFS values were know are discussed above. The impact of other social factor variables like: Median Family Income (MFI), Education, % Single Family, %Employment is unknown.

Curriculum Vitae

Ketan Mane

800 N. Union Street, Apt # 402, Bloomington, IN 47408
(812) 391-2744 | kmane@indiana.edu | http://ella.slis.indiana.edu/~kmane

---

OBJECTIVE

Seeking a position, where I can utilize my analytical abilities, problem-solving capabilities to design information systems.

EDUCATION

- **Ph.D. in Information Science** May 2002 – Oct 2006
  Indiana University, Bloomington, Indiana, USA
  Thesis: Envisioning Knowledge: Tightly Coupling Knowledge Analysis and Vis.
  Advisor: Dr. Katy Börner, Indiana University

- **M.S. in Human Factors Engineering** Dec 1999 – Mar 2001
  Wright State University, Dayton, Ohio, USA

- **Bachelor of Science in Biomedical Engineering** Jun 1995 – May 1999
  Mumbai University, Mumbai, India

TECHNICAL SKILLS

- Perl-CGI
- Java
- R
- Matlab
- Windows
- Macintosh
- Unix
- HTML, XML
- Dreamweaver
- Photoshop
- Fireworks
- JavaScript
- PostgreSQL
- MySQL
- MS-Access
- Perl-DBI

RESEARCH INTERESTS

Information Visualization, Data Analysis, Network Analysis, Knowledge Management, Bioinformatics.

PUBLICATIONS

- Mane, Ketan, Börner, Katy. (2006). SRS Browser: A visual interface to Sequence Retrieval System, Visualization and Data Analysis Conference, San Jose, CA, SPIE-IS&T, Jan 15-19, 2006.
- Collins, Linn Marks, Mane, Ketan, K., Martinez, Mark, L. B., Hussell, Jeremy, A. T., Luce, Rick, E., (2005) ScienceSifter: Facilitating activity awareness in collaborative research groups through focused information feeds, 1st IEEE international conference on e-Science and grid computing (e-Science 2005), Melbourne, Australia. pp. 40-47.
- Susanne Ragg, Marc B. Rosenman, Eve M. Doucette, Zhong Yan, Julie C. Haydon, Jada H. Paine, Nadine D. Lee, Terry Vik, Ketan Mane, and Katy Börner. (2005) Data Visualization of Multiparameter Information in Acute Lymphoblastic Leukemia Expands the Ability To Explore Prognostic Factors. Blood, 106: pp. 862.
- Mane, Ketan, K. and Mostafa, Javed (2005) Exploring the impact of information visualization on medical information seeking using the web. HCI International 2005, Las Vegas, Nevada.
- Ord, Terry J., Martins, Emília P., Thakur, Sidharth, Mane, Ketan K., and Börner, Katy. (2005) Trends in animal behavior research (1968-2002): Ethoinformatics and mining library databases, Animal Behavior, 69, pp. 1399-1413.
- Mane, Ketan and Börner, Katy. (2004) Mapping Topics and Topic Bursts in PNAS. PNAS, 101(Suppl. 1), pp. 5287-5290. Also available as cond-mat/0402380.
- Boyack, Kevin W., Mane, Ketan K. and Börner, Katy. (2004) Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. IV2004 Conference, London, UK, pp. 965-971.
- Thakur, Sidharth, Mane, Ketan, Börner, Katy, Martins, Emilia and Ord, Terry. (2004) Content Coverage of Animal Behavior Data. In Visualization and Data Analysis, San Jose, CA, SPIE-IS&T, 5295, pp. 305-311.
- Mane Ketan and Thakur Sidharth. (2003) Oncosifter - A Customized Approach to Cancer Information. Worshop on Information Visualization Interfaces for Retrieval and Analysis (IVIRA), JCDL, 2003.
- Sheth, Nihar, Börner, Katy, Baumgartner, Jason, Mane, Ketan, Wernert, Eric. (2003) Treemap, Radial Tree and 3D Tree Visualizations. Poster Compendium, IEEE Information Visualization Conference, pp. 128-129, 2003. This entry won 2nd place in the InfoVis 2003 Contest.

TECHNICAL REPORT

- Mane, Ketan and Börner, Katy (2004) SRS Browser: A Visual Interface to Sequence Retrieval System. Technical Report SLISWP-04-04, SLIS, Indiana University.
- Penumarthy, Shashikant, Mane, Ketan and Börner, Katy. (2004) A Toolkit for Large Scale Network Analysis. Technical Report SLISWP-04-02, SLIS, Indiana University.

- Mane, Ketan and Börner, Katy. (2003) Content Coverage of PNAS in 1982-2001. Technical Report SLISWP-03-02, SLIS, Indiana University. Presentation at the SLIS Student Conference at Indiana University won the 1st prize.

POSTER PRESENTATIONS

- Jan 15-19, 2006: SRS Browser: A Visual Interface to Sequence Retrieval System, Poster Presentation at the Visualization and Data Analysis Conference 2006, San Jose, California.
- Sept 21-22, 2005: SRS Browser: A Visual Interface to Sequence Retrieval System, Poster Presentation at the I-Light Symposium 2005, IUPUI, Indianapolis, Indiana.
- Aug 3-4, 2005: Data Visualizations of Co-Authorship Networks, Taxonomies and Information Feeds, Poster presentation at the Symposium 2005: Highlighting Student and Postdoctoral Research, Los Alamos, New Mexico.
- May 9-11, 2005: SRS Browser: A Visual Interface to Sequence Retrieval System, Poster Presentation at the BioComplexity 7 Conference, Indiana University, Bloomington, Indiana.
- Nov 14-17, 2004: 'Mapping Medline Papers, Genes and Proteins related to Melanoma Research'. Poster presented at the NSDL Conference, Chicago, Illinois.
- Sept 10, 2004: Visual Interface to SRS, Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. Presentation at the IV lab Open House, School of Informatics, Indiana University, Indiana.
- May 27, 2004: Mapping Medline Papers, Genes and Proteins related to Melanoma Research. Presentation at the 1st Annual Indiana Bioinformatics Conference, Indianapolis, Indiana.
- Dec 8, 2003: Mapping Topics and Topic Bursts of PNAS Publications in 1982-2001, Correlating Publication, Gene, and Protein Data, Reviewing 50 Years of Animal Behavior Research, Visualizing the Space of Faculty and Courses in SLIS, CS, and Informatics at IUB, Visualization and Pairwise Comparison of Trees. Presentation at IV lab Open House, SLIS, Indiana University, Indiana.
- Aug 14-17, 2003: Content Coverage of PNAS 1982-2001. Presentation at the Biocomplexity Vs Multiscale Modeling in Biology, Notre Dame, Indiana.
- May 9-11, 2003: Content Coverage of PNAS 1982-2001. Presentation at the Mapping Knowledge Domains, Arthur M. Sackler Colloquium, NAS' Beckman Center, Irvine, California.

HONORS/ACHIEVEMENTS

- **2006**
  - Invited to give a talk at the National Institute of Health (NIH) Roadmap Interdisciplinary Methodology and Technology Summit (M & T Summit) in Rockville, Maryland.
  - Demonstrated the visionary data analysis approach to visualize clinical trial information at the 'Informatics in Action' talk series on 'How Information Visualization Can Support NCI's Fight Against Cancer', National Cancer Institute in Rockville, Maryland.

- **2005**
  - Making IT Happen Award for developing SRSBrowser, an interactive visualization tool to query Sequence Retrieval System (SRS) that is among the biology community for their research.
  - Making IT Happen Award for developing InfoVis Cyberinfrastructure Database that provides provide researchers with easy access to scholarly records from digital libraries of publications, patents and grants.

- **2004**
  - Won 1st prize at the Annual SLIS student conference for poster presentation on 'SRS Browser: A Visual Interface to Sequence Retrieval System'.
  - Making IT Happen Award for developing a ' Toolkit for Large Scale Network Analysis'.

- **2003**
  - Won the first prize at the 1st Annual SLIS student conference for paper presentation on 'Content Coverage of PNAS in 1982-2001'.
  - Won the 2nd prize at the InfoVis Contest 2003 for paper on 'Treemap, Radial Tree and 3D Tree Visualizations'.

- **2002**
  - Recipient of SLIS Merit Scholarship 2002-2003, Indiana University.

WORK EXPERIENCE

- **Intern** May 2005 – Aug 2005
  Los Alamos National Lab (LANL), Los Alamos, New Mexico
  - Developed Sciencesifter, an RSS based collaborative tool to promote collective intelligence based on knowledge sharing among research scientist at LANL.
  - Analyzed a dataset and generated visualization that gives a global overview of collaboration trend between LANL and strategic UC campuses over the past 10 years, existing and potential collaborators for a specific subject, variations in collaboration strength and years of successful publication.

- **Faculty Research Assistant** Apr 2002 – present
  SLIS, Indiana University, Bloomington, Indiana
  - Working on Indiana 21st Century Research and Technology funded project on Center of Excellence for Computational Diagnostics that aims to develop novel visualizations that would help in diagnosis and treatment of leukemia patients.
  - Worked on NSF funded project Enable which focuses on developing association based learning modules for the bioinformatics domain.
  - Generated a 'Map of Melanoma Research' that provides a global overview of activity in different research domains, emergence of new topics, association between genes, proteins and papers.
  - Lead a team of five to develop InfoVis Cyberinfrastructure Database - a database of public, patents and grants data which will serve researchers and practitioners interested in large scale analysis, modeling and visualization of scholarly data.
  - Developing a system called OncoSifter that would provide easy access to diagnostic information on cancer.

- **Graduate Assistant** July 2002 – Dec 2002
  Life Science Library, Indiana University, Bloomington, Indiana
  - Assisted students with the database and research material search.
  - Developed a website for students in the school of nursing for efficient access to instructional materials.

- **Engineer** June 2001 – Apr 2002
  UES, Inc., U.S. Army Aeromedical Research Laboratory (USAARL), Fort Rucker, Alabama
  - Supported U.S. Army initiative to improve ground and aviation vehicle restraints.
  - Analyzed historical data and identified human tolerance limits and biomechanics in crashes.
  - Developed new techniques to simulate the static and dynamic loading conditions.
  - Performed simulation data analysis and made recommendation for a crashworthiness design.

- **Intern** Sept 2000 – Mar 2001

Delphi Energy & Chassis Systems, Dayton, Ohio
  - Analyzed and redesigned workstation as per ergonomic standards to reduce stress.
  - Gained knowledge about applying ergonomic principles in manufacturing process.
  - Maintained database of human resources department using Microsoft Access.

PROJECTS

- **Computational Diagnostic Project** (Faculty: Dr. Katy Borner, SLIS, IU & Susanne Ragg, IUPUI)
  - Developed a visual diagnostic tool to help detect patterns and trends in clinical data
  - Incorporated interactivity features to support the diagnostic decision making process.

- **ENABLE Project:** (Faculty: Dr. Javed Mostafa & Dr. Katy Börner, SLIS, IU)
  - Developed SRS Browser, an interactive visualization tool developed to visualize query results from Sequence Retrieval System (SRS) that is widely used by the biology research community.
  - Involved in developing an associative based learning module for bioinformatics.
  - Generated the map of melanoma research involving association between genes, proteins and Medline literature.

- **Taxonomy Validator Project:** (Faculty: Dr. Katy Börner, SLIS, IU)
  - Led a team of 4 members in developing an alternate visualization for hierarchical dataset.
  - Involved in designing the system architecture, work-plan, database schema of the project.

- **Cyberinfrastructure & Knowledge Domain Visualization:** (Faculty: Dr. Katy Börner, SLIS, IU)
  - Contributed towards the development of 'InfoVis Cyberinfrastructure', a data-software-computing repository for research and teaching in information analysis and visualization.
  - Involved in analyzing and visualizing the space of faculty and courses among different departments at Indiana University, Bloomington.

- **Information Retrieval System – OncoSifter:** (Faculty: Dr. Javed Mostafa, SLIS, IU)
  - Developed a system using Perl-CGI, to retrieve latest cancer diagnosis & treatment information.
  - Integrated different modules for query submission, data retrieval and filters for displaying results.
  - Introduced personal profile creation feature to customize the latest news to user preferences.

- **Network Analysis Code Repository:** (Faculty: Dr. Katy Börner, SLIS, IU)
  - Brainstorming for various properties involved in network analysis.
  - Developed a repository that supports serialization of the process in analyzing networks using different algorithms.

- **Cancer Data Visualization:** Hierarchical data visualization (Faculty: Dr. Katy Börner, SLIS, IU)
  - Implemented the hyperbolic tree algorithm to represent hierarchical data of different cancer types.
  - Data retrieval and information filters were implemented using Java-Perl combination to parse relevant information.

- **Patient Medical Record System:** Enhancing the visualization (Researcher: Dr. Susan Ragg, IUPUI)
  - The existing system was studied for the visual interfaces offered for interaction with data.
  - User-perspective study was done to understand the additional requirements.

- Additional features of zooming and filtering of the data was applied to enrich the user experience.

- **Voice Google:** Using Natural Language Processing (NLP) (Faculty: Dr. Javed Mostafa, SLIS, IU)
  - Explored the potential of data retrieval from Google's directory structure through speech.
  - Using Java-Swing, mapped the developed GUI interface with search engine for query submission.
  - Applied information filters to parse data prior to displaying to the user.

- **Checkmate:** Web page information validator (Faculty: Dr. Kiduk Yang, SLIS, IU)
  - Implemented Perl-CGI based errors check program designed to validate the links on the web page.
  - Incorporated report generating feature for user feedback on the functional and non-functional links.
  - Usage of agents was done to provide detailed system function information.

- **Library Journal Database Management:** Life-science library, IUB, (Faculty: Dr. Kiduk Yang, SLIS, IU)
  - Brainstorming operation was carried out to understand the workflow of library.
  - A prototype MS-Access database was developed to improve the efficiency of information management and retrieval.

- **Usability Study:** Web pages for Office of Academic Support & Diversity, Indiana University
  - Performed user's perspective to generate a task-oriented questionnaire for use in usability test.
  - Recorded the test usability test session for in-depth analysis activities of the participants.
  - Recommended web-page interfaces and information layout to enrich the user's experience.

- **Client Website Design:** BeebleBrox Electronic Press Kit
  - Conducted a needs assessment to identify desired web-content required by the client.
  - Brainstorming of ideas was applied to build website based on user-task analysis.
  - Usability testing was conducted for feedback and optimization.

TEACHING EXPERIENCE

- **Assistant Instructor**
  - **Course: L542 - Introduction to Human Computer Interaction – Sum 2002**
    (Instructor: Katy Börner)
    Responsibilities include: answering student's questions related to the course assignment, maintaining
    the course web-page, validating the timely submission of assignment by students.

- **Course: L597 – Structural Data Mining and Modeling – Fall 2003 and Fall 2004**
  (Instructor: Katy Börner)
  Responsibilities include: helping students during the lab sessions, assisting student's in writing scripts for data parsing and data analysis, script writing to convert different data formats to Pajek input format, a network analysis tool used in for data visualization in the course.

- **Design of Education Material**
  I helped in the design of education material that was used by students enrolled in 'Structural Data Mining and Modeling' course. Different activities carried out in this role include: creating tutorial pages, For example, Pajek tutorial page that briefly explains all steps required from data upload to visualization. The role also included writing scripts for data parsing, data conversion and data analysis, customizing data for demo of different visualization application, creating web-pages explaining different graph matching algorithms (ABSURDIST, Similarity Flooding and Simple Matching) along with its pros and cons and its application in graph analysis.

SERVICES

- Reviewing paper for the Visualization and Data Analysis Conference, San Jose, CA, 2007.
- Chaired the open laptop software demo session at Network Science (NetSci) 2006 conference, Bloomington, IN, 2006.
- Chaired sessions for the Visualization and Data Analysis Conference, San Jose, CA, 2006.
- Reviewed paper for major science journal – Scientometrics, 2005
- Presented the Places and Spaces Exhibit at Los Alamos National Lab, Aug 2nd, 2005.
- Presentation of the Melanoma Map at the SLIS, Brown Bag Conference, Mar 4th, 2005.

REFERENCES

Katy Börner, Ph.D.
School of Library and Information Science (SLIS)
Indiana University, Bloomington, IN
(812) 855-3256 | katy@indiana.edu

Javed Mostafa, Ph.D.
School of Library and Information Science (SLIS)
Indiana University, Bloomington, IN
(812) 856-4182 | jm@indiana.edu

Kevin Boyack, Ph.D.
Sandia National Lab
New Mexico, NM
(505) 844-7556 | kboyack@sandia.gov