

Computational Diagnostics: A Novel Approach to Viewing Medical Data

Ketan K. Mane

*Research Library, Los Alamos National Lab
Los Alamos, NM 87545
kmane@lanl.gov*

Katy Börner

*School of Library and Information Science,
Indiana University, Bloomington, IN 47405
katy@indiana.edu*

Abstract

A transition from traditional paper-based medical records to electronic health records is underway. The resulting massive amounts of electronic patient records offer tremendous potential to personalize patient diagnosis and treatment. In this paper, we discuss a computational diagnostics tool that uses digital medical records to help physicians gain better insight about a patient's medical condition. The paper details different interactive features of the tool and discusses the tool's potential in the practice of evidence-based medicine and for advancing patient diagnosis practices.

Keywords - Information Visualization, Computational Diagnosis, Electronic Health Record System

1. Introduction

Over the years, advanced medical research has led to a better understanding of different medical conditions. This knowledge is used to design improved diagnosis and treatment solutions for patients. However, traditional paper-based medical record systems make it hard or impossible to fully utilize existing knowledge and patient records for patient diagnosis [1, 2]. Hence, there is a movement towards electronic health records (EHR) that store patient medical records in digital format [2, 3]. Yet, making sense of individual or multiple EHR is challenging. Attempts have been made to replicate the original medical sheet design on the computer screen; to show the medical history of a patient in a graphical time-series format [4, 5]; and to augment EHR with 2D/3D images from imaging devices [6]. These approaches fail to take advantage of an EHR being much more than just an 'electronic view' of a paper record or an index into a collection of images from imaging devices.

In this paper, we present a 'Computational Diagnostics Tool' (CDT) that supports clinical decision making by providing context and details of EHRs together with the means to aggregate over multiple EHRs. The tool was partially inspired by a matrix visualization discussed in [7].

Specifically, the CDT provides two complementary views of medical data that are highly interactive. Physicians can load in or select a set of patients, examine their medical and clinical conditions, compare the values for one patient with all other patients, and request details. The tool helps answer questions such as: What patients show the worst values for a given set of variables? What patterns are seen in the medical variables of patients undergoing the same treatment? From a clinical dataset, what patient population has severity values for different variables? Do a selected group of patients share similar trends across different variables of interest? Do two patient groups share similar trends across different variables? How do trends and patterns for selected patients compare?

The paper is organized as follows: Section 2 describes the medical dataset used and the insight needs of physicians. Section 3 details the general architecture of the computational diagnostic tool together with the interactive visualizations it provides. Section 4 demonstrates the utility of multiple coordinated views in data analysis. The paper concludes with discussion of results and future work.

2. Details of the Medical Dataset

The sample dataset used in this study was compiled and provided by Dr. Susan Ragg, Julie Hayden, and Jada Pane, of Indiana University, Indianapolis. It consists of medical records of patients diagnosed with acute lymphoblastic leukemia (ALL). The medical dataset includes EHRs for 81 patients. Each record comprises values for 19 different medical variables that were retrieved from different data

sources: clinical and laboratory test data from the Regenstrief Medical Records System, cytogenetics data from the clinical genetics database at Indiana University, immunophenotype data from the pathology database at Indiana University, and patient demographic census data. All patient data is anonymized. For each of the 19 variables, it was important to know how doctors interpreted different variables from the diagnostic perspective. The variables were categorized into five different categories as shown below:

Category: Outcome

Patient Variables: relapse, relapse site, alive/death status, and last day known alive (LDKA).

Category: Biology

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

Category: Host

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

Category: Treatment

Patient Variables: Blast readings at day 7 (BM 7), and day 14 (BM 14).

Category: Social Factors

Patient Variables: Family Income class (MFI), patient's education level, percentage of single family members and percentage of family employment.

In sum, the sample medical dataset comprised of $81 \times 19 = 1,539$ data values that were stored in a MS-Access database.

3. Computational Diagnostic Tool

The computational diagnostic tool was developed to help analyze the acute lymphoblastic leukemia

(ALL) medical dataset introduced in section 2 as well as similar datasets. In particular, it needed to support the identification of characteristic patterns and trends and the comparison of patterns and variations in the medical conditions of patients in support of customized patient treatment.

The computational diagnostic tool uses an ODBC connection to communicate with the MS-Access database that stored the ALL medical dataset. From the patient data it generates two different visualization views to provide different perspectives on the dataset. The two views are a matrix view and a parallel coordinates view. Figure 1 shows the system architecture of the computational diagnostic tool

A Model-View-Controller (MVC) design pattern is used in the construction of the two visual views of the dataset. To generate a matrix view, a 'Database Result Parser' prunes the results from database queries, resulting in a 'Database Model'. Afterwards, the 'Matrix MVC' view is created with the help of the matrix controller. In the case of parallel coordinates, the 'Parallel Coordinates Input Generator' is used to query the database. From the query results, a parallel-coordinates data model is constructed. Later the parallel coordinates view is used to display the results. The 'Data Selection Tracker' is introduced to interlink the matrix and parallel coordinates view.

Basic visualization principles [8] are adopted to arrive at a visually pleasing and effective layout of the interface. Emphasis is on showing the data clearly to trigger viewers' thought processes about the medical dataset. Gray scales and color schemes are used to show patterns and trends. Shneiderman's information seeking mantra [9] – overview, filter, and details on demand - are adapted to show patient data values and different data characteristics.

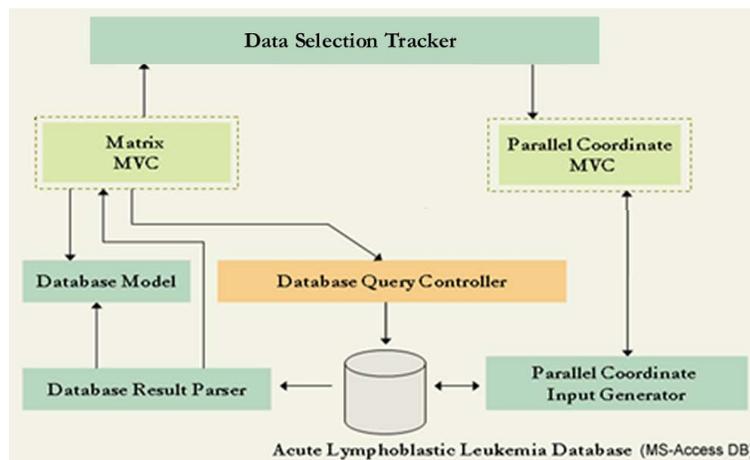


Figure 1: System architecture of the computational diagnostic tool

3.1. Matrix View

The matrix view is used to provide a global overview of the ALL medical dataset. In this view, the patient study-ids are mapped to columns and different patient variables are shown as individual rows. Based on the patient's medical condition, two different color palettes are used to show different ranges of values for independent variable data (phenotype) and dependent variable data (prognosis). Thus, the matrix cell color acts as an indicator of the patient's medical condition.

Figure 2 shows the phenotype view. In this view, a blue-to-green color gradient is used to map a range of phenotype data values. Blue represents bad medical conditions while green indicates good medical conditions.

Figure 3 shows the prognosis view of the ALL dataset. In the prognosis matrix view, the variable values are shown using a red-to-gray-to-green gradient scale. In this case, red 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 1.

Table 1: Color range and %EFS condition

<u>Color Range</u>	<u>%EFS</u>	<u>Condition</u>	<u>Risk</u>
Red	0 – 39	worse	High
Gray	40 – 89	common	Moderate
Green	90 – 100	better	Low

The matrix view also provides an option to integrate both phenotype and prognosis condition in a single view. User-interaction such as sorting is also available.

A scenario of matrix view usage for patient diagnosis is provided here. Each column represents all variable values and thus the complete medical record of a single patient. For example: column indicated by number '51' shows the medical condition of the patient with study id '51'.

In addition to a global overview of the dataset, the color codes used in the matrix view communicate visual cues on the salient characteristics of the

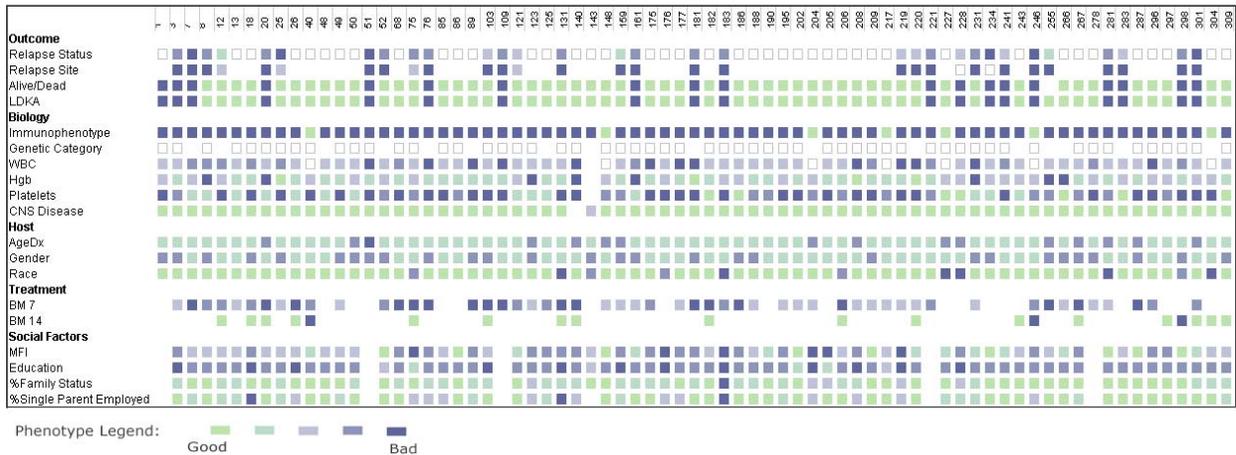


Figure 2: Phenotype view of the medical dataset

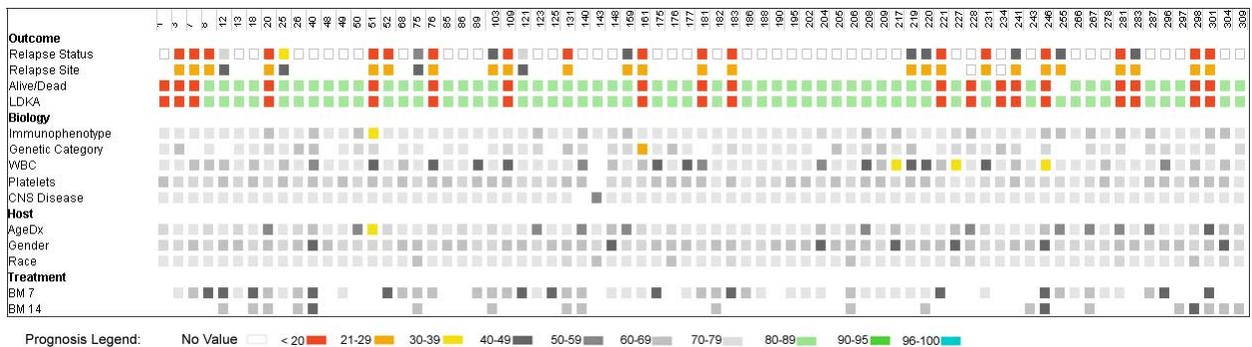


Figure 3: Prognosis view of the medical dataset

patient's condition. For example, all patients with the worst medical conditions can be easily identified: patient study id: 51, 76, 109 and 161.

Further, by placing patient data columns adjacent to each other, it is possible to perform a quick comparison and to identify the similarity/differences of the medical condition of patients. For example, the adjacency of patients with study id 51 and 52 helps to compare the medical conditions of these two patients. The color-coding supports the conclusion that the medical condition of patient with study id '51' is worse than patient with study id '52'.

3.2. Parallel coordinates view

A parallel coordinates visualization [10] was implemented by extending the open source program Parvis [11]. Normally, a parallel coordinates view is used to show quantitative data. In the CDT, the parallel coordinates view was modified to show both nominal and quantitative data values.

Figure 4 shows data in the resulting visualization. The axes represent all 19 diagnostic variables with nominal and quantitative values. 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 variable row order from the matrix view, but the order of axes can be changed as necessary.

Within the parallel coordinates view, line segments with dark color indicate that there are many patients who share the same data values. A mouse hover action over a patient's line highlights it in red to show a patient's profile. For example, Figure 4 shows the trend for patient with study id '314'. Additional interactive features of the parallel coordinates are

covered below. These features can be activated when necessary.

Tool-tip to show diagnostic values of a patient

When the cursor hovers over a line, the values for a specific patient are shown to the right of each axis. For example, the tool-tip displays values for the patient with the study id '314' in Figure 4.

Axes-labels to mark different regions/values

The axis labels can be displayed on demand for each axis. Depending upon the data values (quantitative/ nominal), respective values are displayed along the axes. The axes labels serve as landmarks to get an overview of the distribution of data being displayed in the parallel coordinates view, see Figure 5

Severity zones for different variables

For some diagnostic variables, potentially harmful value ranges can be shown by displaying zones along the axes in parallel coordinates view, see Figure 5. Triangular-shaped severity zones for quantitative variables show a proportional increase in severity with the variable value. The rectangular shaped zones show severity along axes with the nominal data.

Simultaneously display groups of patients to study their variations

Using the brush feature, two groups of patients can be selected. Different strokes and colored lines are used to clearly distinguish between groups, see Figure 6. An ability to simultaneously display multiple groups helps to compare patient group trends.

In sum, the matrix view and parallel coordinates view can be applied to identify patterns and trends in the medical condition of patients respectively.

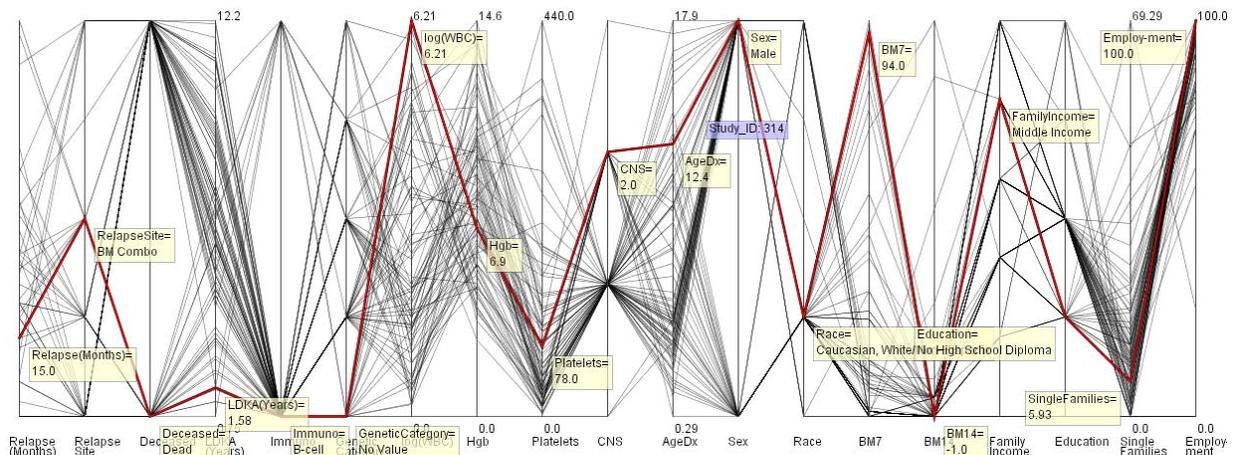


Figure 4: Parallel coordinates view of the medical dataset

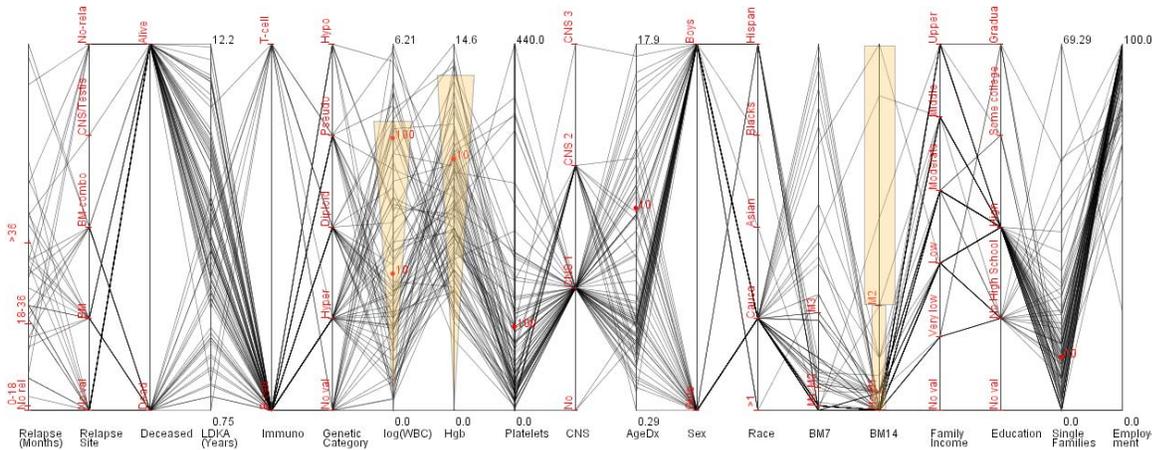


Figure 5: Parallel coordinate view with axes labels and severity zones

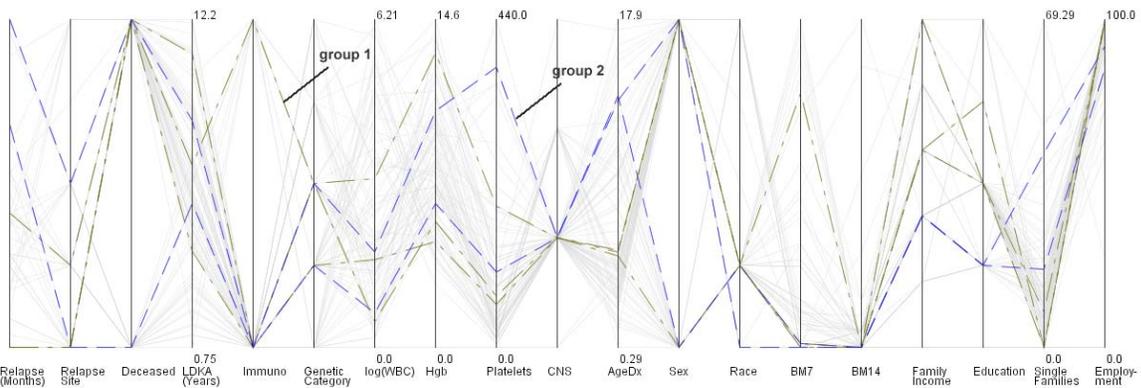


Figure 6: Simultaneous viewing of groups 1 and 2

4. Multiple Coordinated Views

The coordinated views visualization couples matrix and parallel coordinates visualizations. It helps a medical practitioner gain insight into medical data variations for selected patients. While the matrix view supports identification of similar patterns and worst case conditions, the parallel coordinates view helps to quickly identify and compare trends in patient groups. Figure 7 shows how matrix and parallel coordinates views can be used in unison to complement each other. A patient selected in one view is automatically also selected in the complementary view. To accomplish this, the user selects a patient of interest by mouse double-clicking on the patient's study-id. The user action pops up a color swatch to select a color for the patient. In the matrix view, the selected patient's study-id is highlighted using the selected color, see Figure 7-A. The same color is used to highlight the patient line in the parallel coordinates view, see Figure 7-B.

As the end-users are medical practitioners, it is

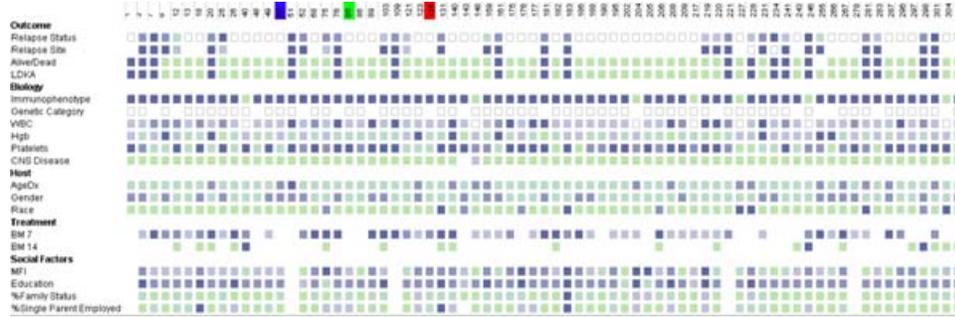
worth mentioning that the data encoding and integration of different features within the matrix and parallel coordinates view was done in consultation with Dr. Ragg and her team.

Further, the tool architecture is flexible to help visualize different medical datasets (e.g., cardiology). For different medical conditions, different value ranges are considered critical. Conditional information needs to be provided to visualize the new medical dataset in the form of visualization.

5. Discussion and Future Work

The paper presented a novel computational diagnostics tool and its application for the treatment of ALL patients. The tool was integrated in the Network Workbench, online at <http://nwb.slis.indiana.edu>. Planned next steps include a usability test to evaluate and further optimize the utility of the computational diagnostic tool over other tools used in medical practice; the application of the tool to other datasets such as diabetes, cardiology, etc.; and the extension of

A – Matrix view



B – Parallel coordinates view

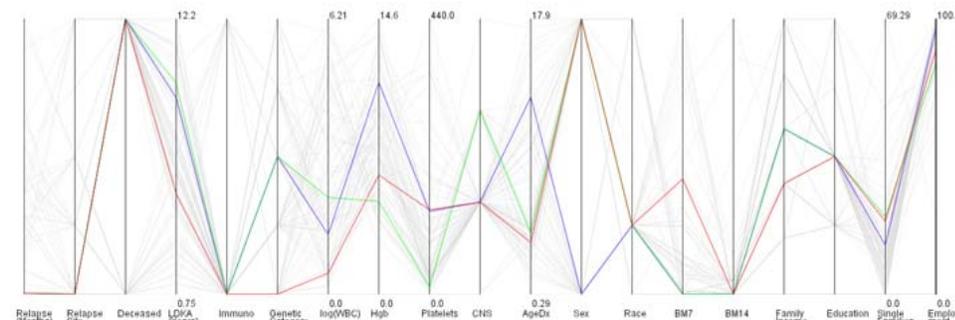


Figure 7: Multiple coordinated view

the system to a recommendation system that offers treatment plans based on a large-scale analysis of many EHRs.

Acknowledgements

All the work was done when at Indiana University, Bloomington. We would like to thank Dr. Susan Ragg for her insights on the ALL dataset. Further, we would like to thank Julie Hayden and Jada Pane for making the ALL dataset available. Also, we would like to thank Bryan Hook for proof reading the paper. This work was supported by the National Science Foundation under DUE-0333623 and a 21st Century Grant for the *Center of Excellence for Computational Diagnostics*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] Roukema, J., R. K. Los, S. E. Bleeker, A. M. Ginneken, J. Lei, and H. A. Moll, *Paper Versus Computer: Feasibility of an Electronic Medical Record in General Pediatrics*. *Pediatrics*, 2006. **117**(1): pp. 15-21.
 [2] Electronic Health Record Overview, in National Institutes of Health National Center for Research Resources. 2006, Mitre Corporation. pp. 26.

[3] Wikipedia, *EHRs*. (Accessed on March 1st, 2007) http://en.wikipedia.org/wiki/Electronic_Health_Record.
 [4] Plaisant, C., R. Mushlin, A. Snyder, J. L. D. Heller, and B. Shneiderman. *LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records In American Medical Informatic Association Annual Fall Symposium 1998*. Bethesda MD. pp. 76-80.
 [5] Powsner, S. M. and E. R. Tufte, *Graphical summary of patient status*. *Lancet*, 1994. **344**(8919): pp.386-389.
 [6] Manssour, I. H., S. Furuie, L. P. Nedel, and C. M. D. S. Freitas. *A Framework to Visualize and Interact with Multimodal Medical Images*. In *International Workshop on Volume Graphics 2001*. Stony Brook, New York: IEEE Computer Society.
 [7] Seelow, D., R. Galli, S. Mebus, H.-P. Sperling, H. Lehrach, and S. Sperling, *d-matrix – database exploration, visualization and analysis*. *BMC Bioinformatics*, 2004. **5**: pp. 168-175.
 [8] Tufte, E. R., *Envisioning Information*. 1990, Cheshire, Conn.: Graphics Press. 126 pages.
 [9] 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.
 [10] 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.
 [11] Ledermann, F. *Parvis*: (Accessed on March 1st, 2007) <http://home.subnet.at/flo/mv/parvis/>