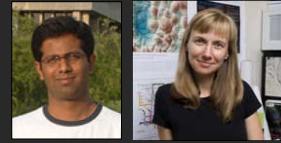
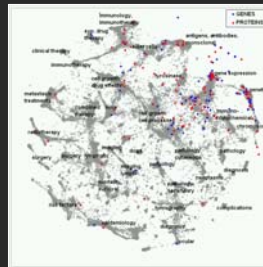


Cyberinfrastructures in Service of Health

Dr. Katy Börner & Ketan Mane
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu

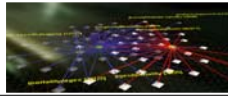


Cancer Institute's (NCI) speaker series Informatics in Action, Bethesda, Maryland, July 20, 2006.



This Talk has Three Parts:

1. Why do we need Cyberinfrastructures (CI)?
2. CI applied to map 'melanoma' related literature, genes, and proteins.
3. CI applied to support computational diagnostics of Acute Lymphoblastic leukemia patients



Why Do we Need Cyberinfrastructures?

Problem

- There are too many and too complex datasets that need to be correlated and understood to arrive at the best possible decisions.
- There are too many different data formats, different algorithms, different implementations of the same algorithm, different programming languages, different research purposes (modeling, analysis, visualization), different communities and practices.
- The analysis, modeling, and visualization of large datasets requires powerful computing infrastructures.
- Managing 1000+ of different data sets and 100+ of different algorithms requires a means to quickly select the best dataset(s)/algorithm(s).

Needed is a socio-technical cyberinfrastructure that supports

Easy access to datasets and algorithms, computer resources, their descriptions, and associated learning modules and access to expertise.

Katy Börner, Cyberinfrastructures in Service of Health, NCI Speaker Service, July 20, 2006.

3

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES
An Oracle database provides access to publications, papers, grants and grant opportunities. The database is continuously and automatically updated.

COMPUTING RESOURCES
The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex, consisting of two Sun E9000 servers with 12 GB RAM, 100 GB hard disk and 100 GB network cache. A Sun X86 server acts as a proxy and file server between users and the database server.

SOFTWARE
An open source IIC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. The algorithms, data analysis methods, tools and the framework code are available for users to be easily "plugged in" or "unplugged" from the framework.

LEARNING MODULES
A set of associated learning modules aims to equip learners with a practical skill set for processing, analyzing and visualizing network data and use different algorithms and diverse visualization techniques and visualization capabilities for the generation and comparison of network visualizations.

CAREER: Visualizing Knowledge Domains. NSF IIS-0238261 award

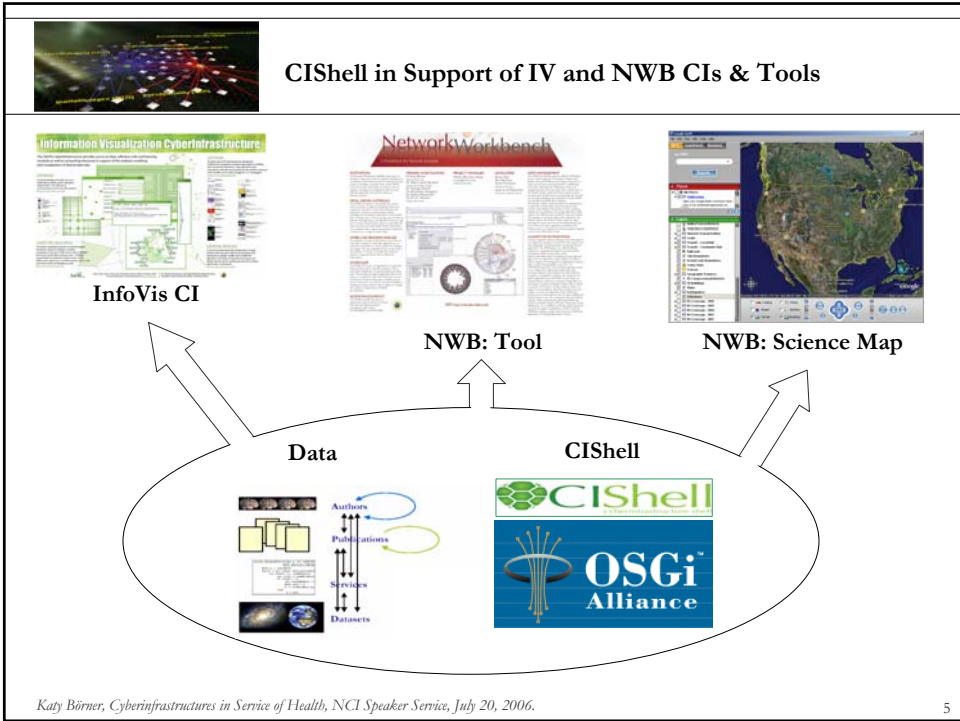
(Katy Börner, \$451,000) Sept. 03-Aug. 08.

<http://in.slis.indiana.edu/>

NetworkWorkbench
A Workbench for Network Scientists



SEI: Network Workbench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. NSF IIS-0513650 award (Katy Börner, Albert-László Barabási, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Werner) (Senior Personnel), \$1,120,926) Sept. 05 - Aug. 08.
<http://nwb.slis.indiana.edu>



CIShell in Support of IV and NWB CIs & Tools

InfoVis CI NWB: Tool NWB: Science Map

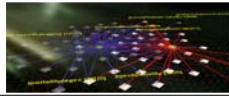
CIShell Framework: <http://sourceforge.net/projects/cishell>

InfoVis Cyberinfrastructure: <http://iv.slis.indiana.edu>

Network Workbench: <http://nwb.slis.indiana.edu>

SciMaps.org: <http://www.SciMaps.org>

Katy Börner, Cyberinfrastructures in Service of Health, NCI Speaker Service, July 20, 2006.



Cyberinfrastructure Shell (CIShell)



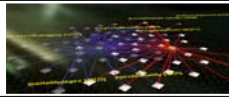
CIShell is an 'empty shell' that supports

- Easy integration of new datasets and algorithms by algorithm developers and
- Easy usage of algorithms by algorithm users.

Its plug-and-play architecture supports the integration and utilization of diverse

- Datasets, e.g., stored in files, databases, streaming data.
- Algorithms, e.g., data processing, analysis, modeling, visualization.
- Interfaces, e.g., remote services, scripting engines, peer-to-peer clients.
- Services, e.g., workflow support, scheduler.

Hence, it can be used for custom UI/Toolkit development.



CIShell – Technical Details



CIShell is built upon the Open Services Gateway Initiative (OSGi) Framework.

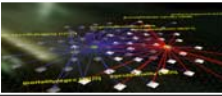
OSGi (<http://www.osgi.org>) is

- A standardized, component oriented, computing environment for networked services.
- Successfully used in the industry from high-end servers to embedded mobile devices since 7 years.
- Alliance members include IBM (Eclipse), Sun, Intel, Oracle, Motorola, NEC and many others.
- Widely adopted in open source realm, especially since Eclipse 3.0 that uses OSGi R4 for its plugin model.


Advantages of Using OSGi

- Any CIShell algorithm is a service that can be used in any OSGi-framework based system.
- Using OSGi, running CIShells/tools can be connected via RPC/RMI supporting peer-to-peer sharing of data, algorithms, and computing power.

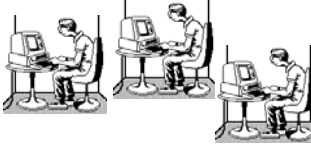
Ideally, CIShell becomes a standard for creating OSGi Services for algorithms. Developed Tools/CI, e.g., IVC & NWB, provide a reference GUI for underlying services.




Serve Algorithms Developers & Users



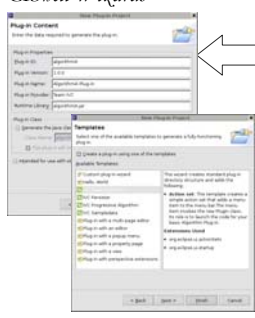
Developers




Users




CIShell Wizards



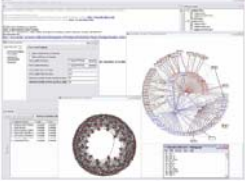
CIShell



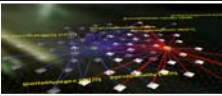
IVC Interface



NWB Interface



Katy Börner, Cyberinfrastructures in Service of Health, NCI Speaker Service, July 20, 2006.



InfoVis Cyberinfrastructure

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES

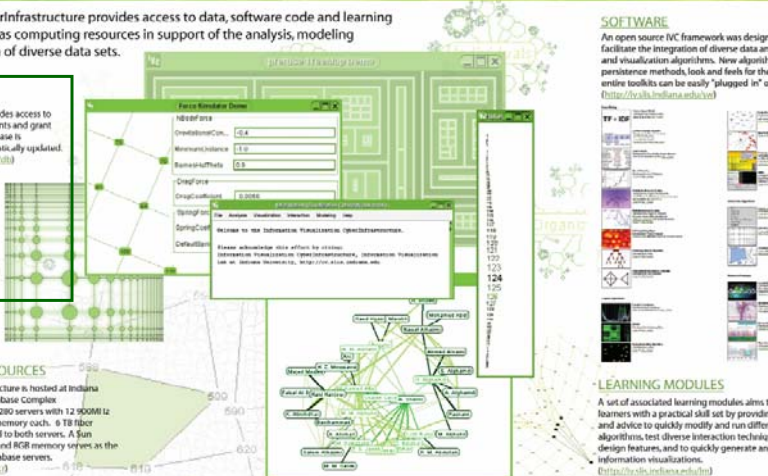
An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (<http://ivis.indiana.edu/db/>)

SOFTWARE

An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged in" or "unplugged". (<http://ivis.indiana.edu/ivc/>)

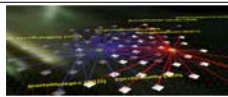
COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun X2200 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpu and 8GB memory serves as the web front-end for the database servers. (<http://ivis.indiana.edu/>)



InfoVis Lab, School of Library and Information Science, Indiana University (2004).
 For more information, contact Katy Börner at kborner@ivylib.org

This material is based upon work supported by the National Science Foundation under Grant Nos. 0328261 and 0308330.



IVC Database (<http://iv.slis.indiana.edu/db>)

Papers and Patents



Medline
Number of Entries: 11,693,477
Years covered: 1963-2002
Size: 135 MB (gunzipped)



Proceedings of the National Academy of Science (PNAS)
Number of Entries: 16,169
Years covered: 1997-2002
Size: 583 MB



United States Patent and Trademark Office (Patents)
Number of Entries: 2,582,847
Years covered: 1976-2003
Size: 350 MB

Grant Awards



National Science Foundation (NSF)
Number of Entries: 181,132
Years covered: 1965-2002
Size: 400 MB



National Institute of Health (NIH)
Number of Entries: 1,003,521
Years covered: 1972-1992 and 1994-2002
Size: 2.3 GB

Funding Opportunities



Community of Science (COS)
Number of Entries: 38,154 (5,000 new entries per month)
Years covered: 2001-present
Size: 60 MB

Katy Börner, *Cyberinfrastructures in Service of Health*, NCI Speaker Service, July 20, 2006.

11

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

SOFTWARE

An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged in" or "unplugged". (<http://iv.slis.indiana.edu/ivc>)

DATABASES

An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (<http://iv.slis.indiana.edu/db>)

COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex, comprising of two Sun V2300 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 CPUs and 4GB memory serves as the web front-end for the database servers. (<http://iv.slis.indiana.edu/ivc>)



For service to quickly modify and test different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (<http://iv.slis.indiana.edu/ivc>)



InfoVis Lab, School of Library and Information Science, Indiana University (2006). For more information, contact Katy Börner at kborner@indiana.edu

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0228261 and DUE-0339623.



Photo: Angkor Capital Group, Pte. Ltd.

Katy Börner, *Cyberinfrastructures in Service of Health*, NCI Speaker Service, July 20, 2006.

12

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES

An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (<http://ivis.indiana.edu/db/>)



COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex consisting of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V980 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (<http://ivis.indiana.edu/>)



InfoVis Lab, School of Library and Information Science, Indiana University (2006). For more information, contact Katy Börner at kborner@indiana.edu

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-0238261 and DUE-0339623.

SOFTWARE

An open source IV framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and entire toolkits can be easily "plugged in" or "unplugged". (<http://ivis.indiana.edu/iv/>)



LEARNING MODULES

A set of associated learning modules aims to equip learners with a practical skill set by providing code and advice to quickly modify and run different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (<http://ivis.indiana.edu/iv/>)

Photo: Angelo Corbelli/Getty Images, 2004

Information Visualization CyberInfrastructure

The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.

DATABASES

An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (<http://ivis.indiana.edu/db/>)



COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex consisting of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V980 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (<http://ivis.indiana.edu/>)



InfoVis Lab, School of Library and Information Science, Indiana University (2006). For more information, contact Katy Börner at kborner@indiana.edu

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-0238261 and DUE-0339623.

SOFTWARE

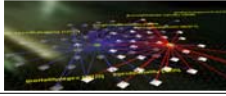
An open source IV framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and entire toolkits can be easily "plugged in" or "unplugged". (<http://ivis.indiana.edu/iv/>)



LEARNING MODULES

A set of associated learning modules aims to equip learners with a practical skill set by providing code and advice to quickly modify and run different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (<http://ivis.indiana.edu/iv/>)

Photo: Angelo Corbelli/Getty Images, 2004



Time Series Analysis

Learning Module

<http://iv.slis.indiana.edu/lm/lm-time-series.html>

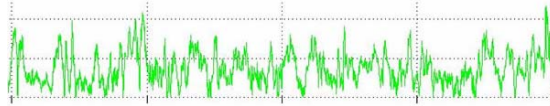


[Learning Modules](#) > Visualizing Time Series Data

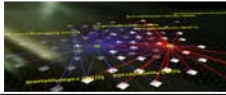
[Description](#) | [Usage Hints](#) | [Learning Task](#) | [Discussion](#) | [References](#) | [Acknowledgments](#)

Description

A time series is a sequence of events/observations which are ordered in one dimension, e.g., time. Frequently, successive observations depend on each other and it makes sense to display them in a (time) sorted fashion, e.g., as a scatter plot. Alternatively, one could be interested to know how many observations of a certain value have been made. Here one would sort the observations by value, count the number of observations for each value and derive a histogram. Time series data can be continuous, i.e., there is an observation at every instant of time see figure below, or discrete, i.e., observations exist for regularly or irregularly spaced intervals.



Time series are recorded, analyzed and used in diverse domains of science. Check out the [Time Series Data Library](#) maintained by Rob Hyndman and Muhammad Akram for numerous data sets from Agriculture, Chemistry, Crime, Demography, Ecology, Finance, Health, Hydrology, Industry, Labour market, Macro-Economics, Meteorology, Micro-Economics, Physics, Production, Sales, Simulated series, Sport, Transport & Tourism or Utilities.



Visualizing Tree Data

Learning Module

<http://iv.slis.indiana.edu/lm/lm-trees.html>



[Learning Modules](#) > Visualizing Tree Data

[Description](#) | [Usage Hints](#) | [Learning Task](#) | [Discussion](#) | [References](#) | [Acknowledgments](#)

Description

Many data sets come in tree format. There are family trees, organizational charts, classification hierarchies, and directory structures. The figure below shows an inheritance tree by Ernst Haeckel ('Stammbaum' in German). Read also [To Draw a Tree](#) by Pat Hanrahan.



[Click image for larger version](#)

A tree graph is a set of straight line segments (edges) connected at their ends containing no closed loops (cycles). You can also call it a simple, undirected, connected, acyclic graph (or, equivalently, a connected forest). A tree with n nodes has $n-1$ graph edges. All trees are bipartite graphs.

Many trees have a root node and are called rooted trees. Trees without a root node are called free trees. Subsequently, we will only consider rooted trees. In rooted trees, all nodes except the root node have only one parent node. Nodes which have no children are called leaf nodes. All other nodes are referred to as intermediate nodes.

NetworkWorkbench

A Workbench for Network Scientists

MOTIVATION

The Network Workbench (NWB) project aims to develop a large-scale network analysis, modeling, and visualization cyberinfrastructure for biomedical, social science, and physics research. Users of the NWB tools and portals will be able to perform network analysis, modeling and visualization with the most effective algorithms and the best reference datasets available.

MENU DRIVEN INTERFACE

The NWB tool shown in the middle has a menu-driven interface. It supports file/dataset load, view, conversion, and save as well as the selection and application of diverse preprocessing, analysis, modeling, and visualization algorithms on the loaded data. To guide users' choices among many and diverse datasets and algorithms, only algorithms that can read the currently activated data model are selectable. All data entry forms provide default values, information on acceptable value ranges, instantaneous feedback if a value is out of range, as well as help.

WORK LOG TRACKING MODULE

The sequence of steps performed by a user such as what file is loaded or saved, what algorithm is run with what parameters, as well as preference changes are logged. The log is displayed in the console and is also saved as a record in a log file. Error logs are saved in a separate file and can be refilled as big reports.

SCHEDULER

A scheduler lets users run algorithms at a particular date and time and in a specified sequence. This is particularly suitable for non-remotely accessible jobs. The number and type of algorithms that run in series or in parallel is only restricted by the amount of memory and processing power available. At any point in time, users can see all currently scheduled or running processes, monitor their progress, or change the sequence of algorithms scheduled for execution.

ACKNOWLEDGMENTS

The NWB cyberinfrastructure is supported in part by the 31st Century Fund and the National Science Foundation under Grants No. IIS-0238281 and IIS-0513650.

PRIMARY INVESTIGATORS

Dr. Katy Börner
Indiana University
Dr. Albert-László Barabási
University of Notre Dame
Dr. Santiago Schnell
Dr. Alessandro Vespignani
Dr. Stanley Wasserman
Dr. Eric A. Wernert
Indiana University

PROJECT MANAGER

Weixia (Bonnie) Huang
(huang@indiana.edu)
Indiana University

DEVELOPERS

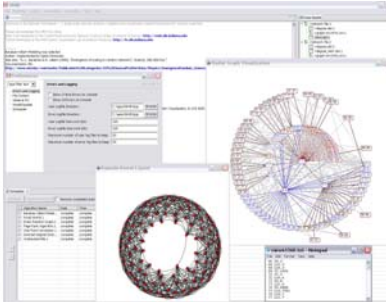
Bruce Herr
Ben Markines
Santo Fortunato
Indiana University
Cesar A. H. Ramaciotti
University of Notre Dame

DATA MANAGEMENT

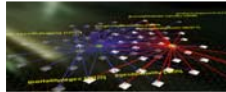
The NWB tool defines a generic, efficient NWB data format which supports the storage of million node graphs. Using the NWB parser plugin, the tool can load, view, and save a network from a NWB data format file. Although the NWB data model is the fundamental data structure, other data models, such as the Preface Graph model and Matrix model, and the parsers that handle those corresponding data formats can be easily developed and integrated into the NWB tool by following NWB data templates. Several data model converters have been developed to conduct the transformation between diverse data models. This facilitates the pipeline of data modeling, analysis, and visualization even though algorithms might require very different data models for input and output. For example, a converter plugin that transforms the NWB model to the Preface Graph model has been developed so that users can use the Radial Graph and Force Directed Layout algorithms provided by the Preface library to visualize the network dataset originally saved in the NWB data format.

ALGORITHM INTEGRATION

A major computer science challenge is the development of an algorithm integration framework that supports the easy integration and dissemination of existing and new algorithms. The NWB utilizes the C2DWeb software architecture originally developed in the Information Visualization Cyberinfrastructure (IVC) (<http://indiana.edu/ivc>) to facilitate the easy plug and play of diverse algorithms. While C2DWeb is written in JAVA it supports the integration of algorithms written in other programming languages, viz. in C++ or FORTRAN. In practice, a pre-compiled algorithm needs to be wrapped as a plugin that implements basic interfaces defined in the C2DWeb Local APIs. Different templates are available to facilitate the integration of diverse algorithms into the NWB. In most cases, no programming is required to integrate an algorithm as a new plugin. A plugin developer simply needs to fill out a sequence of forms for creating a plugin, reports the plugin to the workspace directory, and then users are ready to use the new algorithm via the NWB tool graphical menu. Drawing from the IVC, other JUNG and Preface libraries have been integrated into the NWB as plugins. After converting the generated NWB data model into JUNG Graph and Preface Graph data model, NWB users can run JUNG and Preface graph layouts to interactively explore visualizations of their networks. NWB also supplies a plugin that invokes the XMGraice application for plotting data analysis results.



VISIT: <http://nwb.slis.indiana.edu>



Network Workbench



Investigators: Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert



Software Team: Team Lead: Weixia (Bonnie) Huang
Software Developers: Bruce Herr & Ben Markines
Algorithm Developers: Santo Fortunato & Cesar Hidalgo



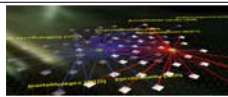
Goal: Develop a large-scale network analysis, modeling and visualization toolkit for biomedical, social science and physics research.

Amount: \$1,120,926 NSF IIS-0513650 award.

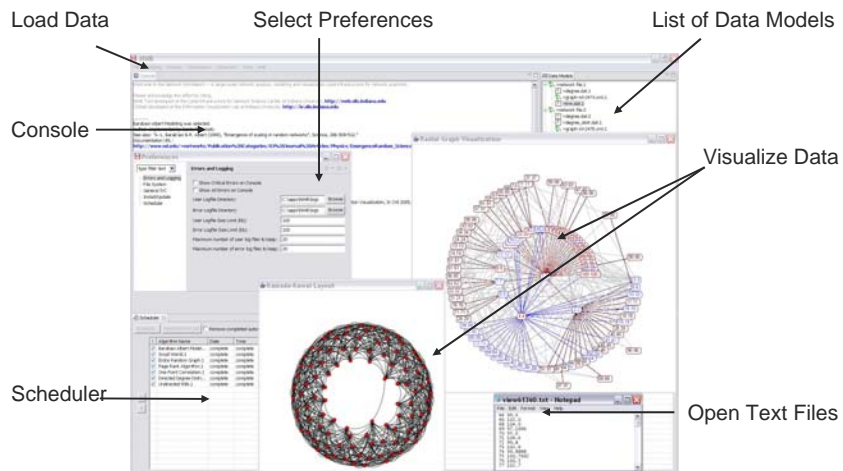
Duration: Sept. 2005 - Aug. 2008

Website: <http://nwb.slis.indiana.edu>



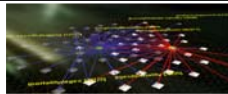


NWB Tool: Interface Elements



Katy Börner, Cyberinfrastructures in Service of Health, NCI Speaker Service, July 20, 2006.

19



List of Algorithms (partially implemented)



Modeling

Random Network Model
Random

Preferential Attachment Algorithms

Barabasi-Albert Model
Dorogovtsev-Mendes-Samukhin
Fitness
Vertices/edges deletion
Copying strategy
Finite vertex capacity
TARL

Rewiring algorithms

Rewiring based on degree distribution
Watts Strogatz Small World Model

Peer-to-Peer Models

Structured
CAN Model
Chord Model

Unstructured

PRU Model
Hypergrid Model

Measurement

Edge/Node level
node degree
BC value of nodes/edges
Max flow edge
Hub/Authority value for nodes
Distribution of node distances (Hop plot)

Local (directed and weighted versions)

Clustering Coefficient (Watts Strogatz)
Clustering Coefficient (Newman)
k-Core Count
Distributions (Plot and gamma, and R^2)
Degree Distributions (in, out, total) (Directed/Total Degree Distribution)
Degree Correlations (in-out, out-out, out-in, in-in, total-total)
Clustering Coefficient over k
Coherence for weighted graphs
Distribution of weights
Probability of degree distribution

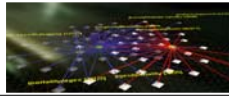
Global

Density
Square of Adjacency Matrix
Giant Component
Strongly Connected Component
Betweenness Centrality
Diameter
Shortest Path = Geodesic Distance
Average Path Length

Motif Identification
Page Rank
Closeness centrality
Reach centrality
Eigenvector centrality
Minimum Spanning Tree

Katy Börner, Cyberinfrastructures in Service of Health, NCI Speaker Service, July 20, 2006.

20



List of Algorithms (partially implemented)



Basic Processes on Networks

Search

k Random-Walk Search
Depth First Search
p-rand Breadth-First Search
P2P
CAN Search
Chord Search

Epidemics Spreading

SIR
SIS

Graph Matching

Simple Match
Similarity Flooding
ABSURDIST

Clustering

Based on Attributes

Hierarchical Clustering
Single Link
Complete Link
Average Link
Ward's Algorithm

Based on Network Structure

Newman Girvan
Clauaset-Newman-Moore
Newman
Cecconi-Parisi
Simulated annealing of modularity
Caldarelli
Weak Component Clustering
vanDongen (random walk)
Cfinder (Clique percolation method)
Reichardt, Bornholdt (q-potts model)

Visualization

Distribution
Scatterplot
Histogram
Geospatial
Circle layout
Grid-based
Dendrogram
Treemap
Hyperbolic tree
Radial Tree
Sparse Matrix Visualization
Kamada-Kawaii
Fruchterman-Rheingold
Orthogonal Layout
k-core visualization

Katy Börner, Cyberinfrastructures in Service of Health, NCI Speaker Service, July 20, 2006.

21

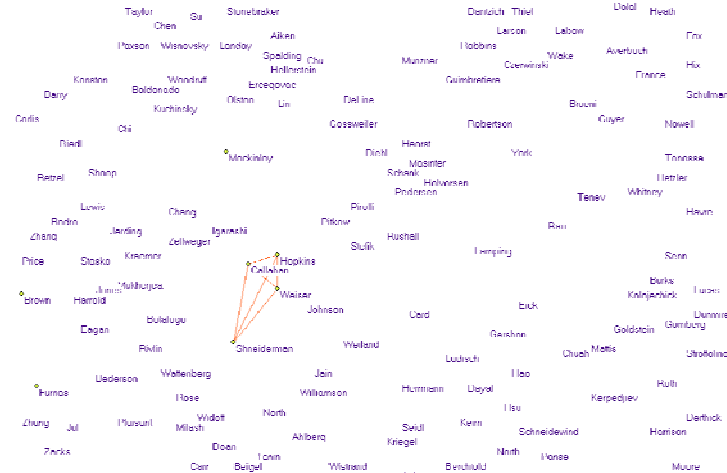
This Talk has Three Parts:

1. Why do we need Cyberinfrastructures (CI)?
2. CI applied to map 'melanoma' related literature, genes, and proteins.
3. CI applied to support computational diagnostics of Acute Lymphoblastic leukemia patients



Mapping the Evolution of Co-Authorship Networks in Information Visualization, 1988 - 2004

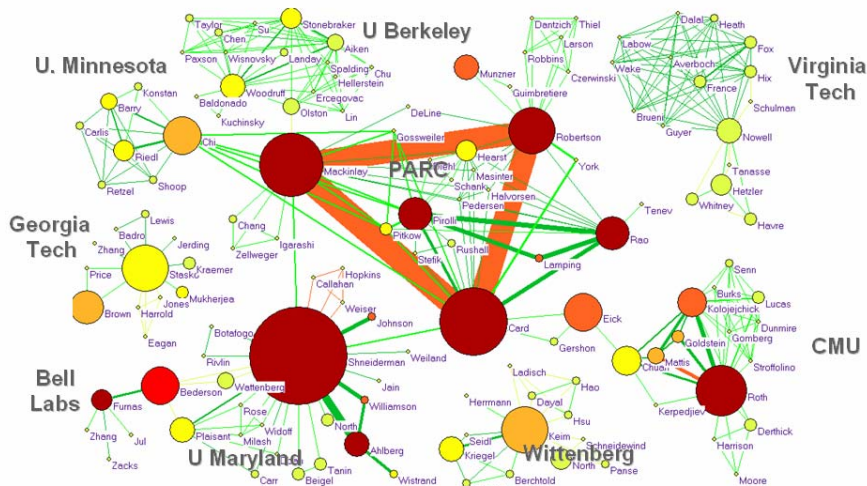
Ke, Viswanath & Börner, (2004) Won 1st prize at the IEEE InfoVis Contest.



Mapping the Evolution of Co-Authorship Networks
Weimao Ke, Lalitha Viswanath & Katy Börner
InfoVis Lab @ Indiana University
2008

Mapping the Evolution of Co-Authorship Networks

Ke, Viswanath & Börner, (2004) Won 1st prize at the IEEE InfoVis Contest.

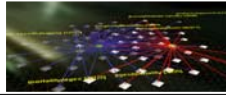


Mapping the Evolution of Co-Authorship Networks
Weimao Ke, Lalitha Viswanath & Katy Börner
InfoVis Lab @ Indiana University
2004

Analyzing, Modeling, and Mapping Science



- Shiffrin, Richard M. and Börner, Katy (Eds.) (2004). **Mapping Knowledge Domains**. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Suppl_1).
- Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). **Visualizing Knowledge Domains**. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Volume 37, Medford, NJ: Information Today, Inc./ American Society for Information Science and Technology, chapter 5, pp. 179-255.
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (in press). **Network Science**. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Information Today, Inc./ American Society for Information Science and Technology, Medford, NJ.



Process of Analyzing and Mapping Science

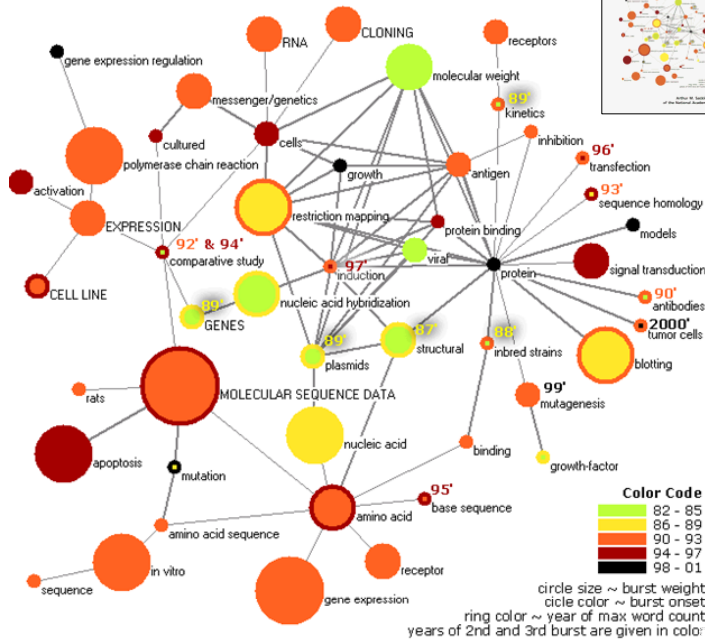
| DATA EXTRACTION | UNIT OF ANALYSIS | MEASURES | LAYOUT (often one code does both similarity and ordination steps) | | DISPLAY |
|---|---|---|--|--|---|
| | | | SIMILARITY | ORDINATION | |
| SEARCHES ISI INSPEC Eng Index Medline ResearchIndex Patents etc. | COMMON CHOICES Journal Document Author Term | COUNTS/FREQUENCIES Attributes (e.g. terms) Author citations Co-citations By year THRESHOLDS By counts | SCALAR (unit by unit matrix) Direct citation Co-citation Combined linkage Co-word / co-term Co-classification VECTOR (unit by attribute matrix) Vector space model (words/terms) Latent Semantic Analysis (words/terms) incl. Singular Value Decomposition (SVD) CORRELATION (if desired) Pearson's R on any of above | DIMENSIONALITY REDUCTION Eigenvector/ Eigenvalue solutions Factor Analysis (FA) and Principal Components Analysis (PCA) Multi-dimensional scaling (MDS) LSA, Topics Pathfinder networks (PFNet) Self-organizing maps (SOM) includes SOM, ET-maps, etc. CLUSTER ANALYSIS SCALAR Triangulation Force-directed placement (FDP) | INTERACTION Browse Pan Zoom Filter Query Detail on demand ANALYSIS |
| BROADENING By citation By terms | | | | | |

Börner, Chen & Boyack.. (2003) *Visualizing Knowledge Domains*. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Volume 37, Medford, NJ: Information Today, Inc./ American Society for Information Science and Technology, chapter 5, pp. 179-255.

Mapping Topic Bursts

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.

Mane & Börner. (2004) PNAS, 101(Suppl. 1): 5287-5290.



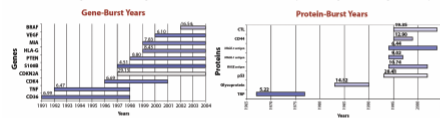
Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research

GOAL

To provide researchers, practitioners and students with a global map of a research domain, to help them answer questions such as: What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in a certain area of research? Which areas are most insular? What are the main connections for each area? What is the relative speed of areas? What new research areas are evolving? How are the objects of study (e.g., genes, proteins) interconnected via papers?

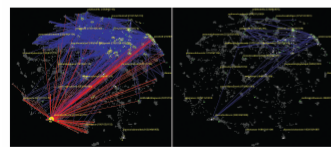
TOP-RESEARCHED GENES & PROTEINS

Identification of sudden interests in research/published papers on certain genes and proteins using Kleinberg's burst detection algorithm. The diagrams show the amount and the time spans of major burst for genes and proteins.



ASSOCIATION MAPS

A gene-gene, gene-paper, gene-protein, protein-paper and protein-protein map was generated. The figure shows the gene-paper (left) and gene-gene (right) network. Highlighted in red is a single gene (CIMA) and all its connections within the given network.



DATASETS

- 53804 Medline publication (1960 - Feb 2004)
- 299 Genes downloaded from Entrez-Gene
- 367 Proteins downloaded from Uniprot

Kevin W. Boyack, Ketan K. Mane, Katy Börner (in press) Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. Accepted for the Information Visualization Conference 2004.

For more information, contact Katy Börner at katy@indiana.edu

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0238261 and DUE-0333623.



PAPER-GENE-PROTEIN MAP

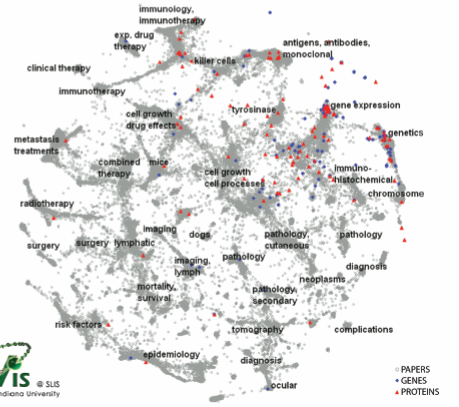
Shown here is the melanoma research area over the last 40 years. Gray dots represent papers, red dots denote proteins, blue dots indicate genes. Experts classified the shown research areas into two main categories:

- Applied Medical Sciences (left side) where research work occurs at organism level.
- Basic Molecular Sciences (right side) with studies related to genes and proteins.

TIME SERIES ANALYSIS

The structure & dynamics of melanoma research was examined:

- 1964-1973: Diagnostic and immunity based approaches dominate.
- Chemotherapy emerges as a new area for cancer treatment.
- 1974-1983: Chemotherapy gains popularity as viable treatment. Monoclonal studies involving tagging cancerous cells using antigens start.
- 1984-1993: Research on metastasis behavior of cancer dominates.
- 1994-2003: Gene-expression and mutation related studies gain popularity.



Boyack, Kevin W., Mane, Ketan and Börner, Katy. (2004). Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. IV2004 Conference, London, UK, pp. 965-971.

SRS Browser: A Visual Interface to The Sequence Retrieval System

Ketan K. Mane & Katy Börner
School of Library and Information Science, Indiana University, Bloomington, IN 47405
Email: {kmane, katy}@indiana.edu



Abstract

The sequential retrieval system (SRS) facilitates retrieval of information from a large number of biological, biochemical and biomedical databases. Currently not supported by SRS is the navigation of associations to answer questions such as: What relation exists between genes that are known to cause melanoma? Which proteins are associated with a particular disease? What is the most well researched gene in cancer research? What papers refer to many genes and/or proteins related to a melanoma?

The 'SRS Browser' system helps researchers to find answers to these questions via association identification in SRS query results. The associations among genes, proteins and papers are presented to the user as highly interactive, navigable visualizations.

System Architecture

The SRS Browser consists of four components:

> SRS System

The UBio-SRS system at Indiana University is queried through the web service program.

> SRS Visual Interface

Client-side interface to query the SRS system and to visualize data associations.

> Web Service

An interface used to access SRS system via SOAP.

> Data-Processing Module

Query results are obtained in xml format.

Association among genes, proteins and literature are identified. Underlying assumption

is that genes/proteins co-occurring in the same paper are related. In the case of literature, if two papers are related if they refer to the same gene or protein.

Features of the SRS Browser

> Extends UBio-SRS system.

> Discovers associations among results from SRS system for genes, proteins and literature.

> Supports interactive visualizations of association for genes, proteins and literature.

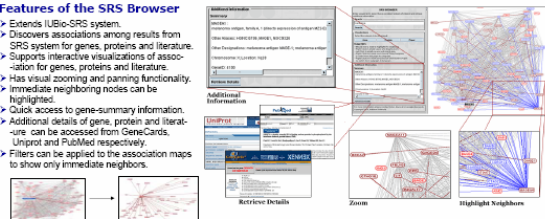
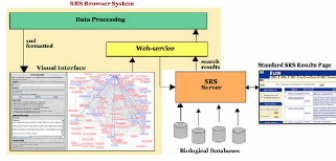
> Has visual zooming and panning functionality.

> Immediate neighboring nodes can be highlighted.

> Quick access to gene-summary information.

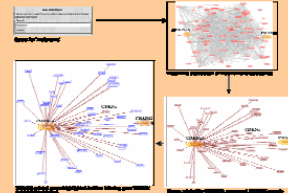
> Additional details of gene, protein and literature can be accessed from GeneCards, Uniprot and PubMed respectively.

> Filters can be applied to the association maps to show only immediate neighbors.



Prototypical Application Scenarios

1. Examining the co-occurrence of genes in papers



2. Finding relevant and related papers



3. Finding highly interlinked proteins



Acknowledgments: We would like to thank Shaikant Penumarty, Josh Goodman, Don Gilbert, Santiago Schnell and Kranthi Varala for their valuable input throughout the project. This work is supported by National Science Foundation grant DUE-0333523.



Mane, Ketan & Börner, Katy. (2006). SRS Browser: A visual interface to Sequence Retrieval System Visualization and Data Analysis, San Jose, CA, SPIE-IS&T, Jan 15-19, 2006.

This Talk has Three Parts:

1. Why do we need Cyberinfrastructures (CI)?
2. CI applied to map 'melanoma' related literature, genes, and proteins.
3. CI applied to support computational diagnostics of Acute Lymphoblastic Leukemia patients

Center of Excellence for Computational Diagnostics

21st Century Grant, Sept. 04 - Aug. 06, \$1,994,951.

Investigators: Susanne Ragg (PI), David Clemmer, Sven Rahmann, and Ilka Ott, Terry Vik, R Clement McDonald, Nunroe Pecock, Zina Ben Miled & Katy Börner.

Visualization Subproject:

Aims to develop visualizations that help identify factors which cause relapse in Acute Lymphoblastic Leukemia (ALL) patients.

Current visualizations provide a

- Global overview of medical condition and diagnostic variable(s) of all patients in the dataset.
- Means to identify patients with best or worst phenotype/prognosis data values.
- Ability to compare patients or patient groups that share similar medical/diagnostic variables.

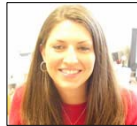
Visualization Team:



Dr. Susanne Ragg



Julie Haydon



Jada Pane



Ketan Mane



Dr. Katy Börner

Computational Diagnostics – Interactive Visualizations

Coupling of different interactive visualizations such as

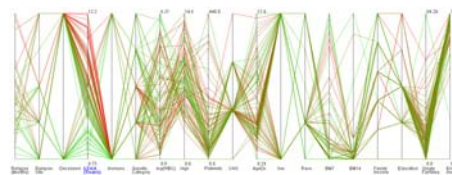
Matrix Visualization

phenotype, prognosis, combined view.

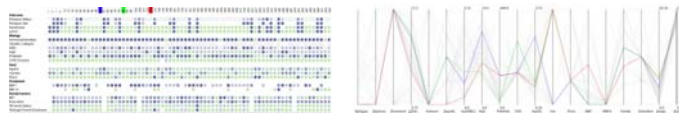


Parallel Coordinate Visualization

phenotype view.

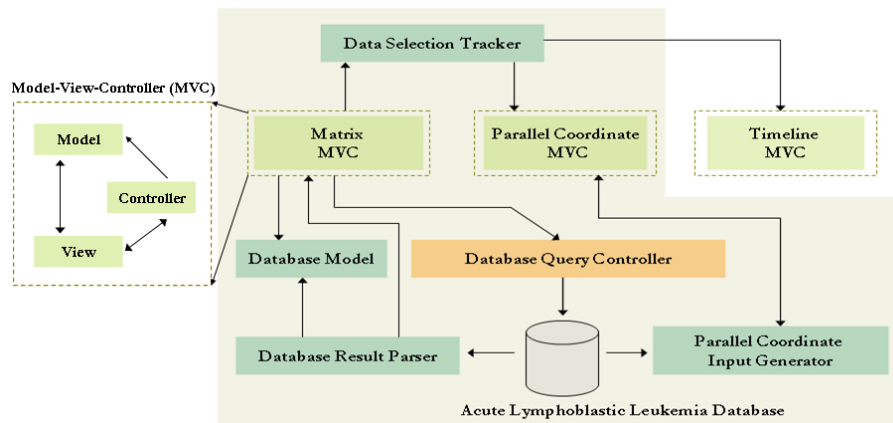


Coupled Windows



Computational Diagnostics – Interactive Visualizations

System Architecture



Computational Diagnostics - Dataset Details

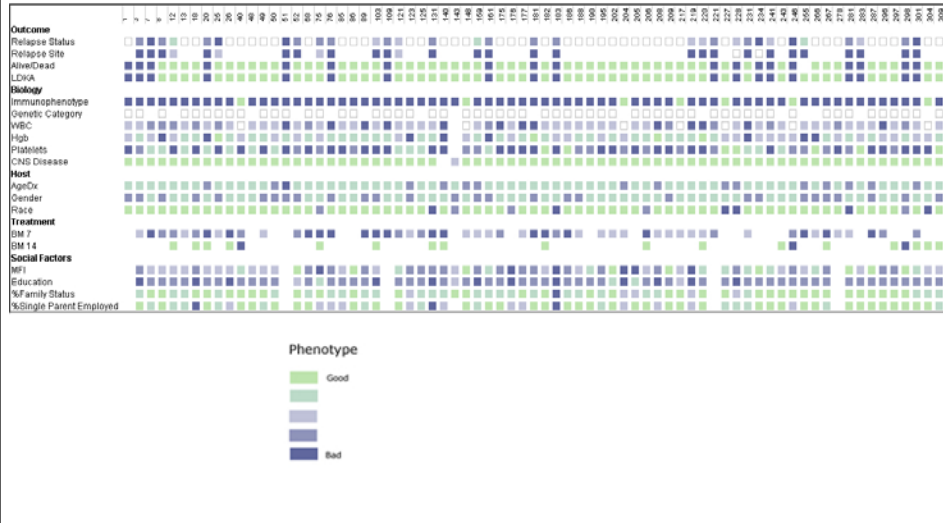
Diagnostic data variables from medical records for Acute Lymphoblastic Leukemia (ALL) patients are grouped into

- 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, education level, %single family members, and % family employment.

All data was provided by Dr. Susanne Raggs, Julie Haydon and Jada Pane.

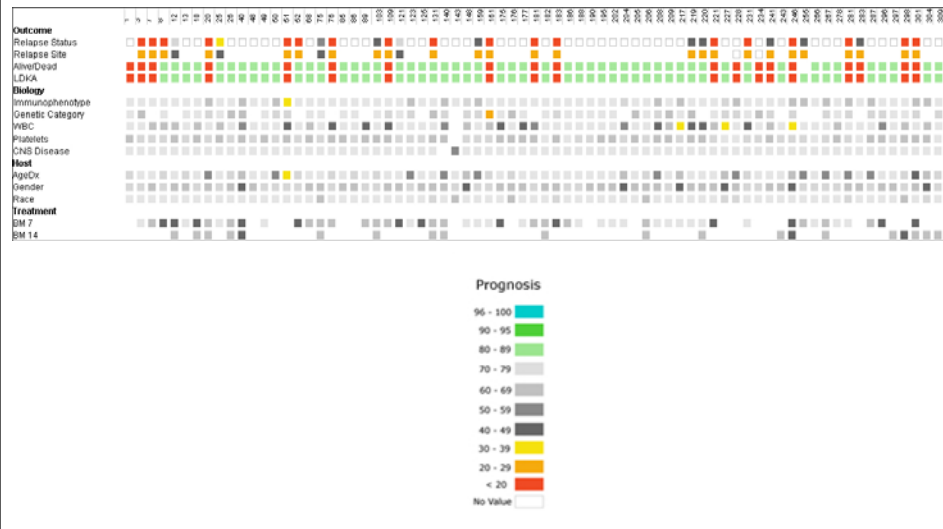
Matrix Visualization – Phenotype View

- Data (vertical axis) - patient (horizontal axis) matrix.
- Data values are color coded according to their severity.



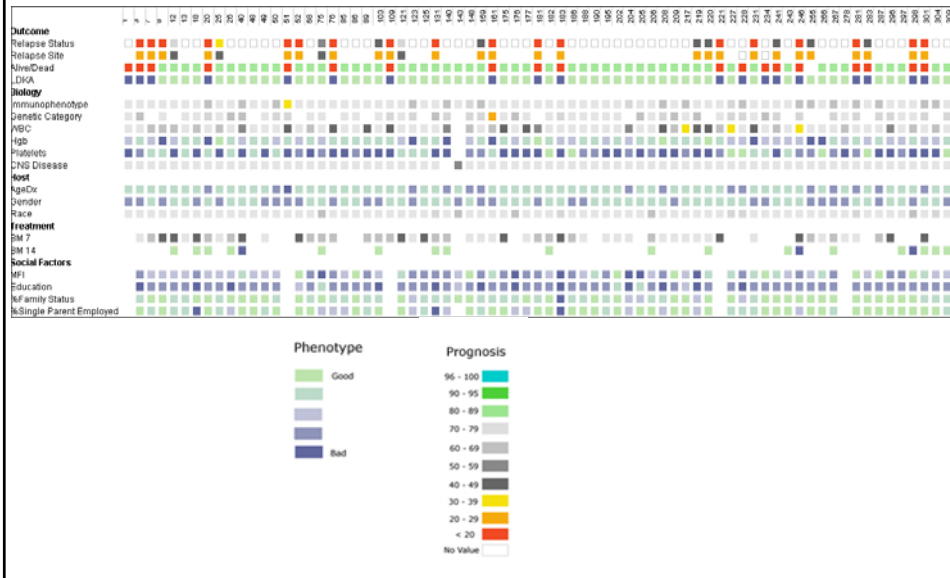
Matrix Visualization – Prognosis View

- Data (vertical axis) - patient (horizontal axis) matrix.
- Color codes indicate event free survival in percent (%EFS).



Matrix Visualization – Combined View

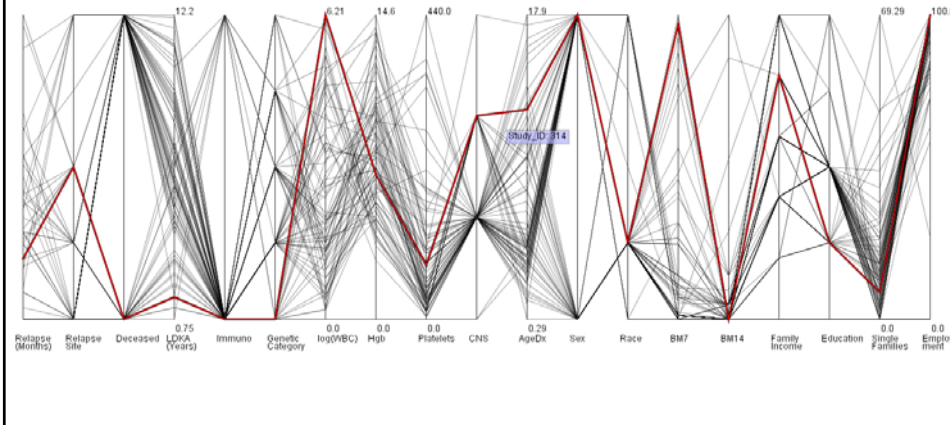
- Facilitates selection of phenotype/prognosis view for individual diagnostic variables.



Parallel Coordinates Visualization

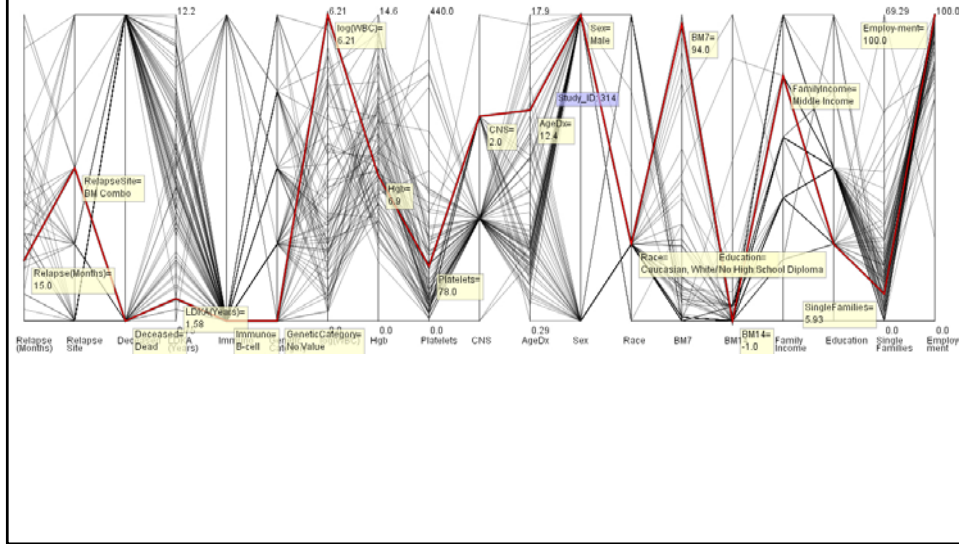
- Uses one axis for each data variable.
- For each patient, all data values on the different parallel axes are connected.
- All patient graphs are shown here.

Single or multiple patients can be selected and studied in detail.



Parallel Coordinates Visualization

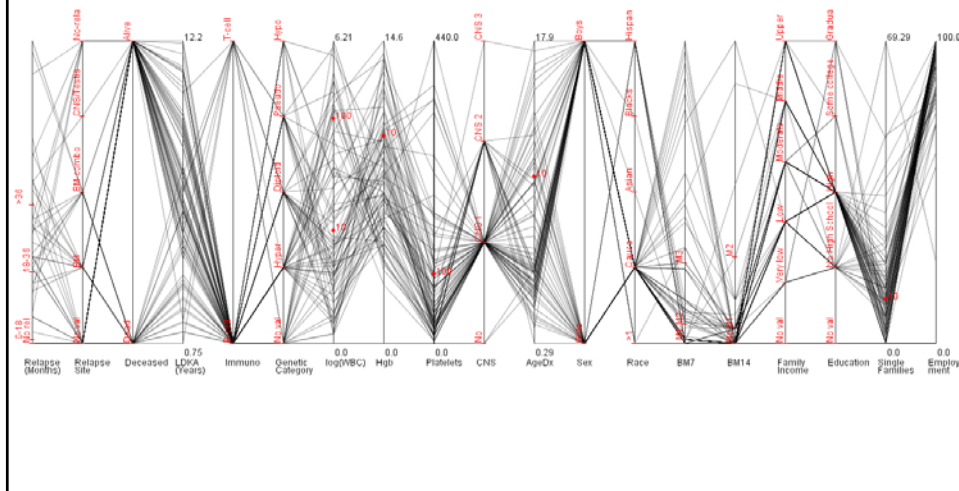
Tool-tip display to show diagnostic values of selected patient.



Parallel Coordinates Visualization – User Interactions

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

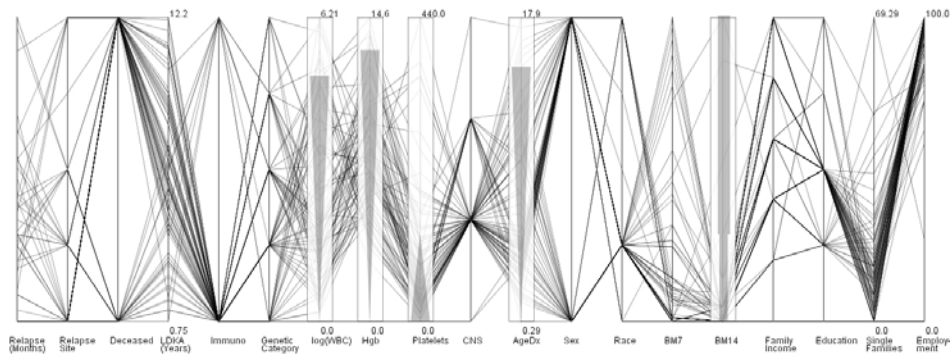
- Numerical landmarks along axes showing values for quantitative variables.
- Category labels along axes show values for nominal variables.



Parallel Coordinates Visualization – User Interactions

Display zones to show severity values for different variables

- Triangular zones indicate variables with quantitative values.
- Rectangular zones are used for variables with nominal values.

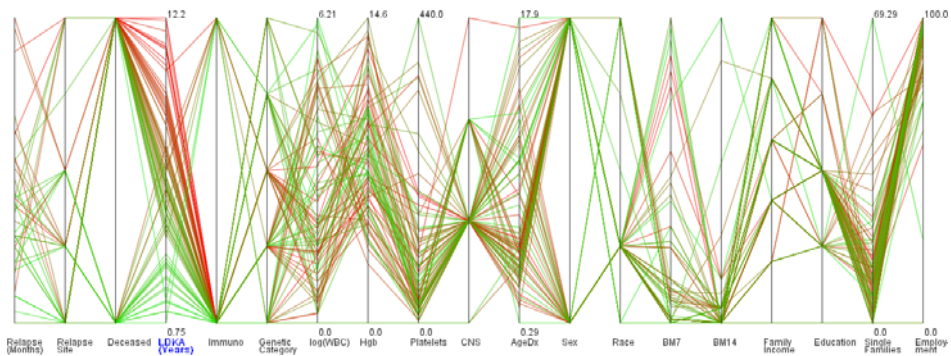


Parallel Coordinates Visualization – User Interactions

Axis selection to study global variations in patient values

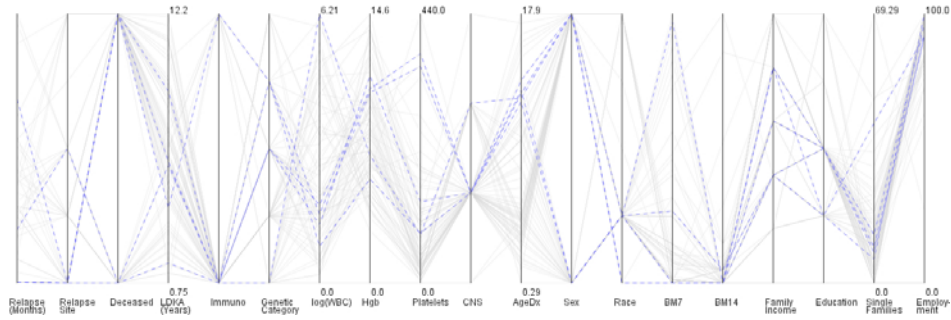
- Single axis can be selected to study the trend in patient values.
- Red-to-green gradient used to indicate values along the selected axis.

[Red = High value, Green = Low value]



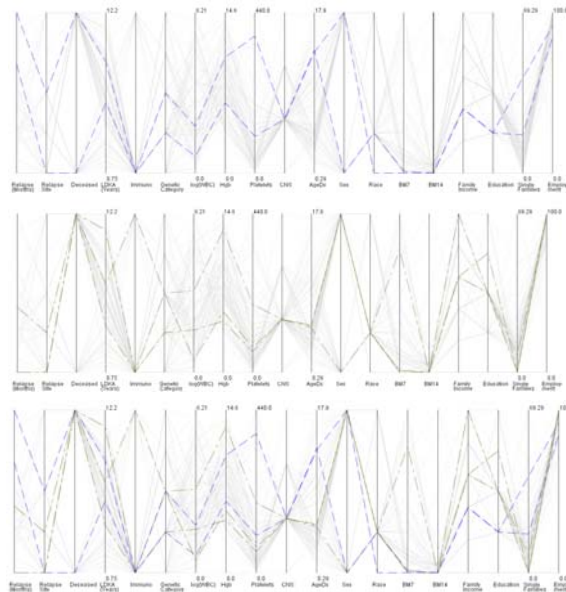
Parallel Coordinates Visualization

A subset of patents can be selected and examined as a group.



Parallel Coordinates Visualization

Simultaneous display of patient groups to study differences.



Patient Group 1

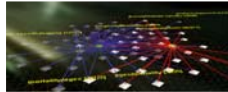
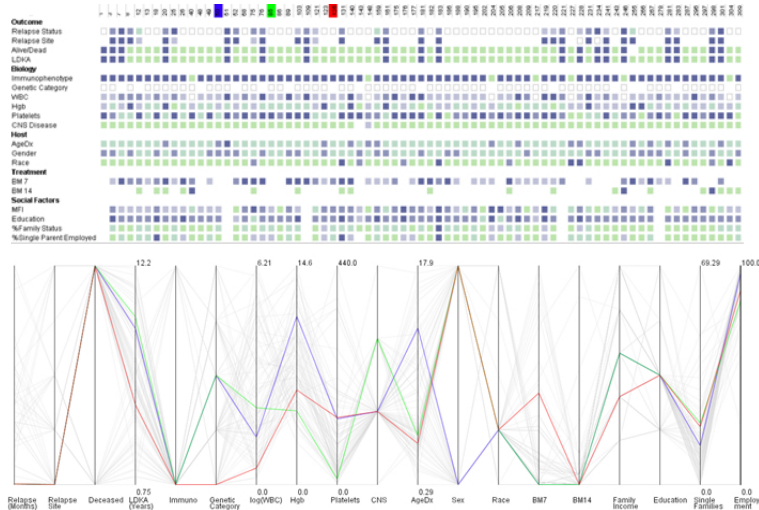
Patient Group 2

Patient Group 1 & 2

Parallel Coordinates Visualization

Multiple Coordinated Views

- Patient can be selected and color coded in matrix view.
- Corresponding patient lines are highlighted in parallel coordinate view.



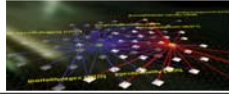
References

Mapping Science

- Mane, Ketan & Börner, Katy. (2006). SRS Browser: A visual interface to Sequence Retrieval System Visualization and Data Analysis, San Jose, CA, SPIE-IS&T, Jan 15-19, 2006.
- Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. *Scientometrics*, 64(3), 351-374.
- Börner, Katy, Dall'Asta, Luca, Ke, Weimao and Vespignani, Alessandro. (April 2005) Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*, special issue on *Understanding Complex Systems*, 10(4): pp. 58 - 67. Also available as [cond-mat/0502147](#).
- Ord, Terry J., Martins, Emilia P., Thakur, Sidharth, Mane, Ketan K., and Börner, Katy. (2005) Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases. *Animal Behaviour*, 69, 1399-1413. [Supplementary Material](#).
- Mane, Ketan K. and Börner, Katy. (2004). [Mapping Topics and Topic Bursts in PNAS](#). *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5287-5290. Also available as [cond-mat/0402380](#).
- Börner, Katy, Maru, Jeegar and Goldstone, Robert. (2004). [The Simultaneous Evolution of Author and Paper Networks](#). *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl.1):5266-5273. Also available as [cond-mat/0311459](#).
- Boyack, Kevin W., Mane, Ketan and Börner, Katy. (2004). Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. IV2004 Conference, London, UK, pp. 965-971.

Computational Diagnostics

- Mahoui, Malika, Kulkarni, Harshad, Li, Nianhua, Ben-Miled, Zina and Börner, Katy. Semantic Correspondence in Federated Life Science Data Integration Systems. Accepted for the 2nd International Workshop on Data Integration in the Life Sciences (DILS'05).
- S. Ragg, T. Vik, D. N. Lee, N. Li, Z. Ben-Miled, M. Mahoui, K. Mane, K. Borner. Combination of Database Integration and Data Visualization for Biomarker Detection in Cancer. Abstract accepted for 37th Congress of the International Society of Pediatric Oncology, Vancouver, September 21-25, 2005.



Thank you.



Please feel free to attend demonstrations of the diverse tools by Ketan Mane.