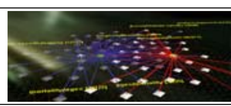
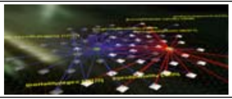
L701, Guest Lecture
Data Mining and Information Visualization

Katy Börner
School of Library and Information Science
INDIANA UNIVERSITY
BLOOMINGTON
katy@indiana.edu



Research Projects



Information Visualization

"Information Visualization is a process of transforming data and information that are not inherently spatial, into a visual form allowing the user to observe and understand the information."

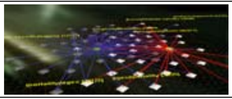
(Source: Gershon and Eick, First Symposium on Information Visualization)

- Rooted in geograp
- Not even 15 years
- Far reaching (IR,
- Tremendous pote

Humans can detect a single dark pixel in a 500 x 500 array of white pixels in less than a second. This screen can be replaced every second by another, enabling a search of 15 million pixels in a minute (Ware, 2000).

Also, people have a truly remarkable ability to recall pictorial images. In one study, Standing, Conezio, & Haber (1970) showed S's 2560 pictures, each for 10 seconds over 7 hours, in a 4-day period. Afterwards, S's were asked to classify pictures presented at a rate of 16 pictures/min and they achieved better than 90% accuracy.

http://ella.slis.indiana.edu/~katy/gallery/04-openhouse/

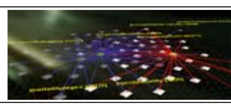


Knowledge Domain Analysis and Visualization

To answer questions such as:

- What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in xx research?
- Which areas are most insular?
- What are the main connections for each area?
- What is the relative speed of areas?
- Which areas are the most dynamic/static?
- What new research areas are evolving?
- Impact of xx research on other fields?
- How does funding influence the number and quality of publications?

Answers are needed by funding agencies, companies, and researchers.



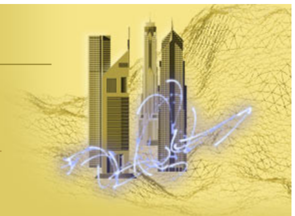
Process of Mapping Knowledge Domains

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. BROADENING By citation By terms	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 Decomp (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

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.

http://vw.indiana.edu/aag05/

Mapping Humanity's Knowledge and Expertise in the Digital Domain



At the 101st Annual Meeting of the Association of American Geographers Denver, CO: April 5-9, 2005.

Session Organizers

Katy Börner, Indiana University
André Skupin, University of New Orleans

Sponsors

Cartography and GIS specialty groups

Description

This session will bring together leading researchers and practitioners that aim to develop techniques, tools, and infrastructures to map humanity's knowledge and expertise for the improvement of science and education.

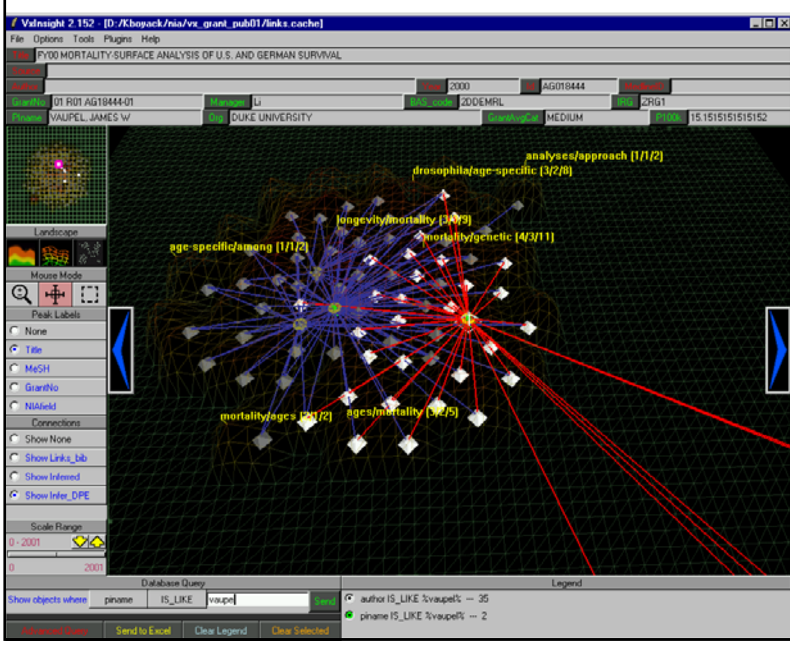
Knowledge and expertise is typically extracted from digitally available literature, news, computer mediated communication data as well as from information about the producers and consumers of those data sets. Advanced data analysis techniques in combination with spatial metaphors, geographic principles, and cartographic methods are applied to organize, visualize, and communicate the semantic relationships inherent in the data.

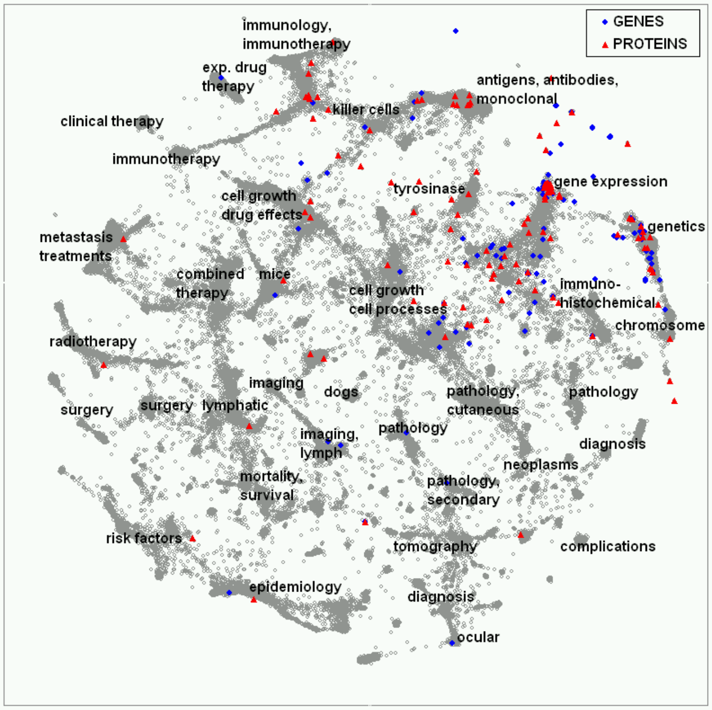
The ultimate goal of this work might be an interactive cartographic map of all of science, with continents representing the major research areas such as, e.g., biology or physics, dots denoting major authors, PIs, papers or news, dynamically evolving research frontiers, blinking 'hot' papers and topics, etc. This map could be used to teach and understand the evolving structure of all of science, to identify major experts, to find and read the most relevant papers and news, to see the effects of resource allocation decisions, to study social networks, etc. Last but not least, it would provide a unique bird's eye view of major experts in specific areas and mankind's knowledge in general.

Some of the leading-edge research on this topic is found where geography intersects with information/library science, computer science, and cognitive science. We invite papers on the broad foundations, computational methods, software systems, and evaluation of such data analyses and visualizations, as they have emerged in this interdisciplinary endeavor.

Indicator-Assisted Evaluation and Funding of Research

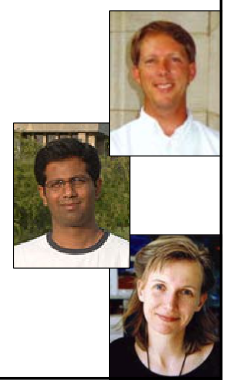
Visualizing the influence of grants on the number and citation counts of research papers (Boyack & Börner, 2003)





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

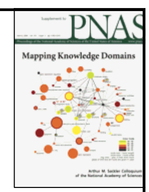
(Boyack, Mane & Börner, 2004)

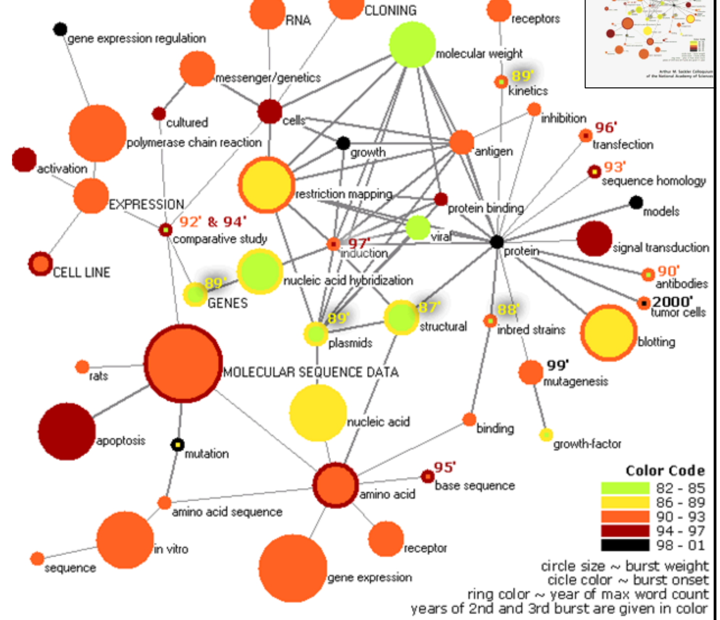


Mapping Topic Bursts

(Mane & Börner, 2004)

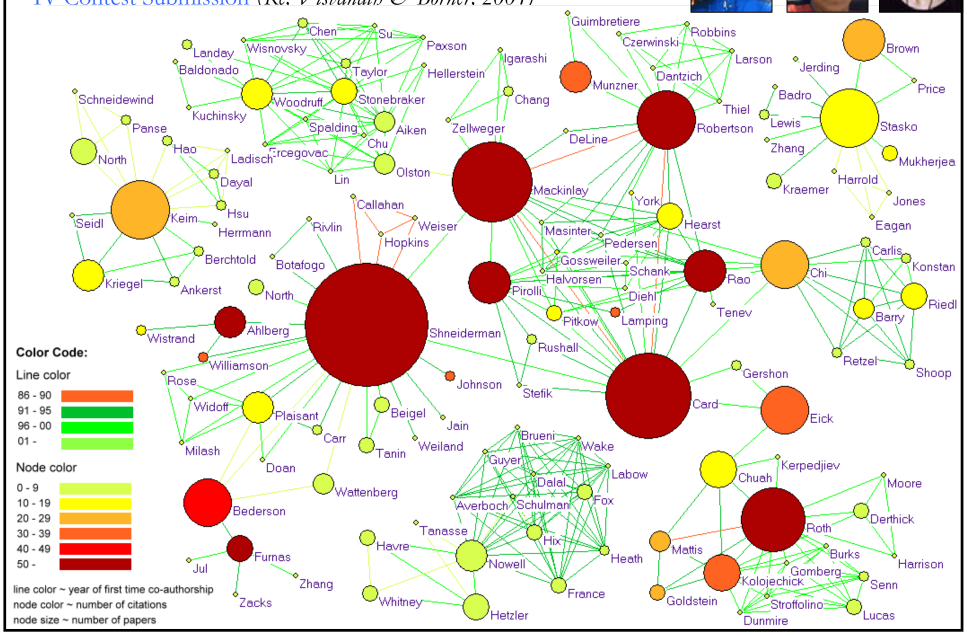
Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.





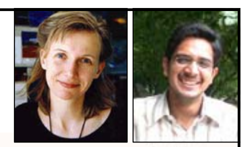
Mapping InfoVis Co-Authorships (Interactive Map)

IV Contest Submission *(Ke, Visvanath & Börner, 2004)*



Information Diffusion Patterns

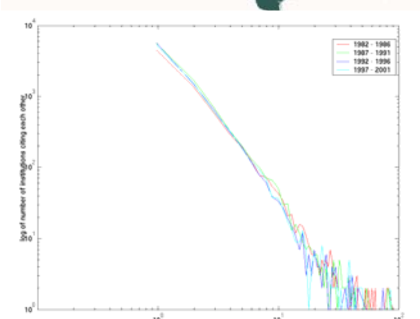
(Börner & Penumarthy, 2004)



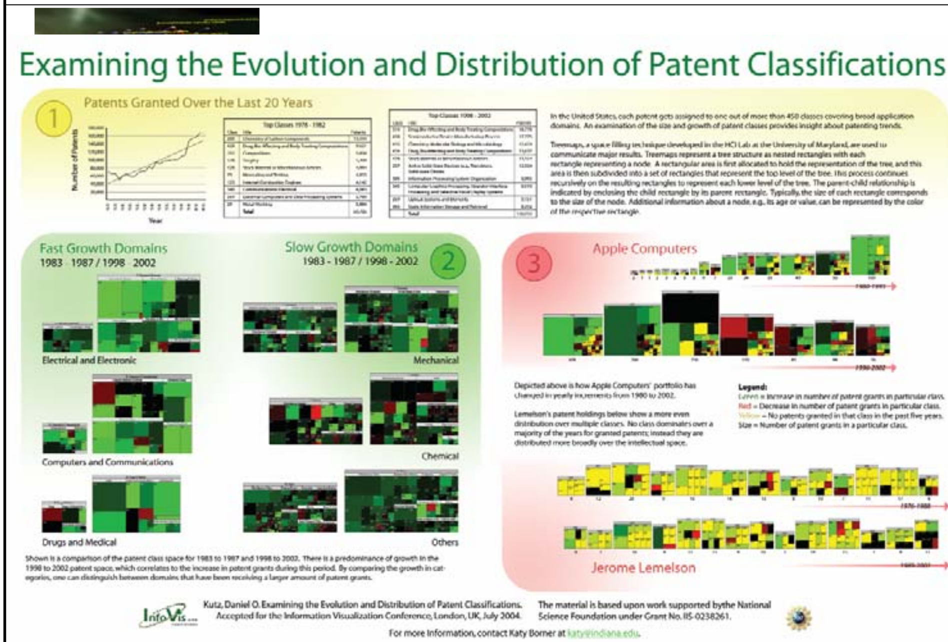
Top 500 most highly cited U.S. institutions.

Each institution is assumed to produce and consume information.

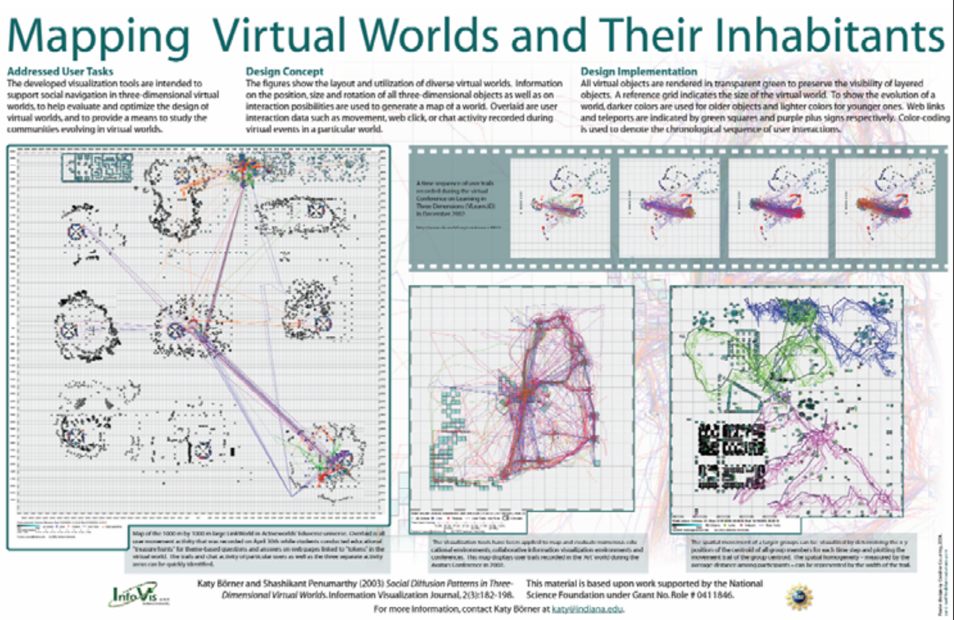
Does Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?



Katy Börner: SLIS Orientation Session, Fall 2004.



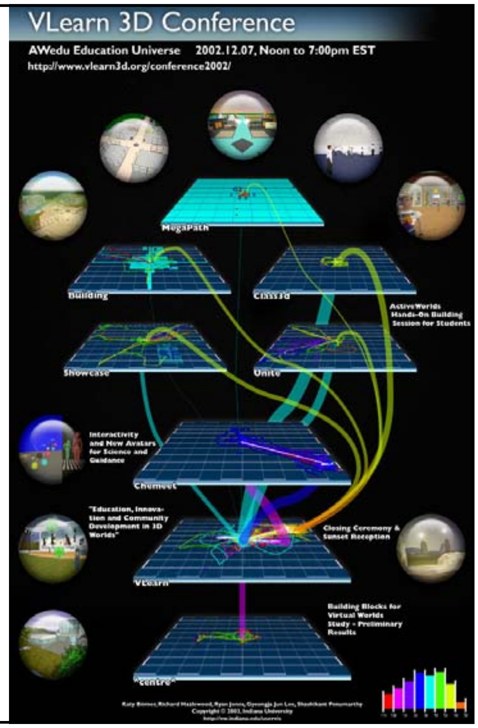
Katy Börner: SLIS Orientation Session, Fall 2004.

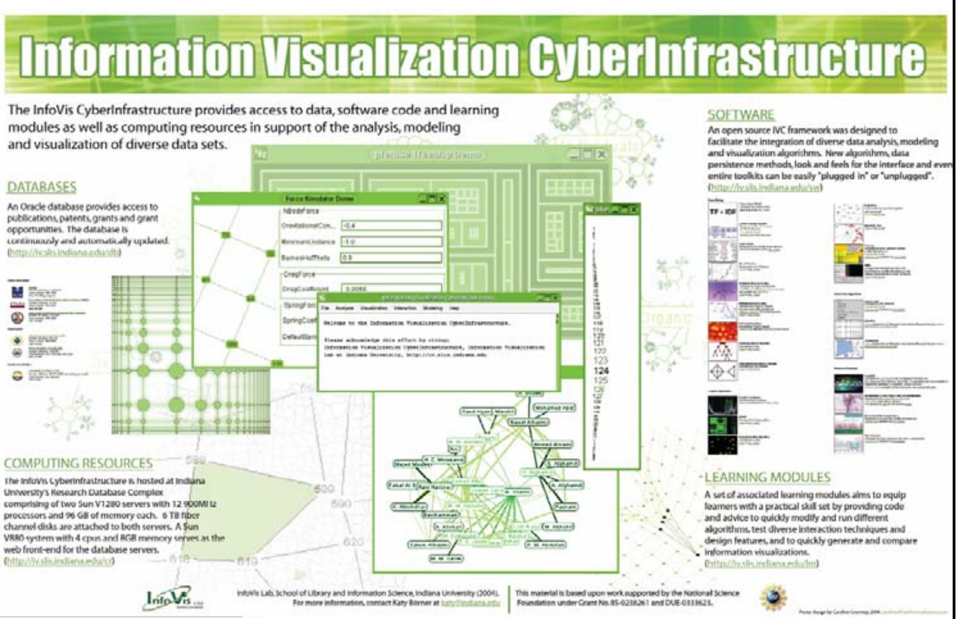


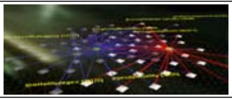
VLearn 3D Vis
(Börner, Hazlewod, Jones, Lee & Penumarthy, 2003)

Temporal-spatial distribution of Conference attendees

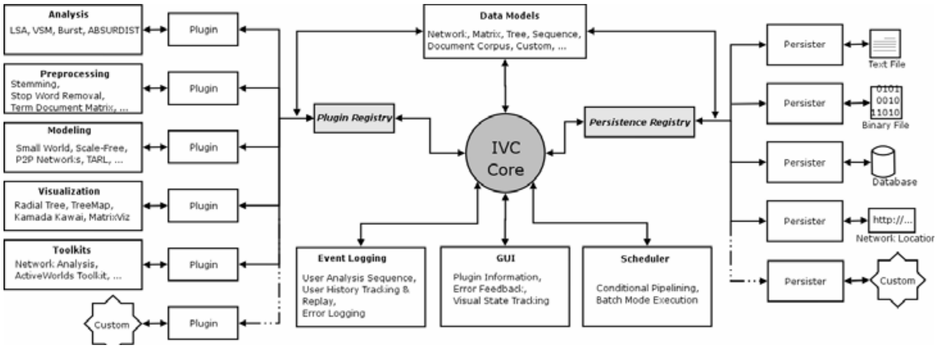
- Conference worlds are represented by square, perspective maps, each labeled by its name.
- Worlds accessed at the beginning of the conference are placed at the bottom, worlds accessed later toward the top.
- Next to each world is a circular snapshot of the virtual venue. Short descriptions of the main sessions are added as text.
- Major jumps between worlds are visualized by transparent lines. The thickness of each line corresponds to the number of traveling users. Color coding was used to denote the chronological paths of the conference sessions.





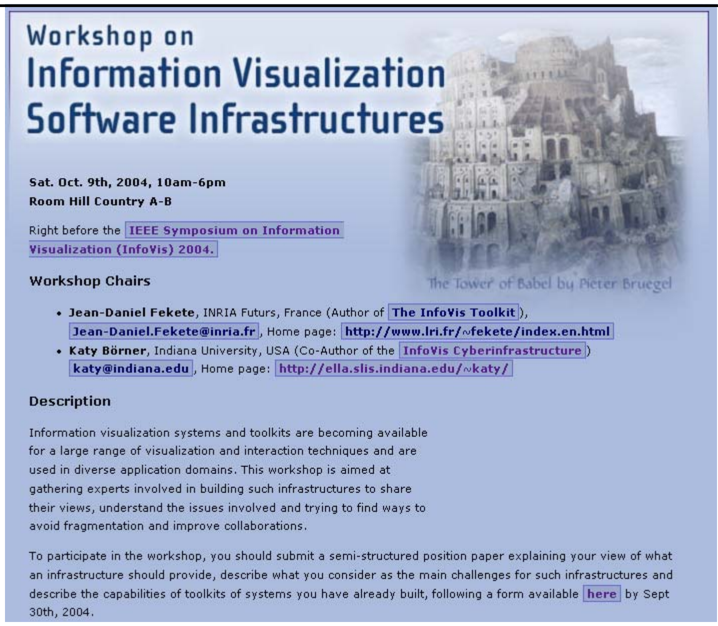


IVC Software Framework



See SLIS Colloquium talk on the *InfoVis Cyberinfrastructure* on Nov 19, 2004.

Katy Börner: SLIS Orientation Session, Fall 2004.



Workshop on
Information Visualization Software Infrastructures

The Tower of Babel by Pieter Bruegel

Sat. Oct. 9th, 2004, 10am-6pm
Room Hill Country A-B

Right before the IEEE Symposium on Information Visualization (InfoVis) 2004.

Workshop Chairs

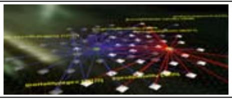
- **Jean-Daniel Fekete**, INRIA Futurs, France (Author of The InfoVis Toolkit), Jean-Daniel.Fekete@inria.fr, Home page: http://www.lri.fr/~fekete/index.en.html
- **Katy Börner**, Indiana University, USA (Co-Author of the InfoVis Cyberinfrastructure) katy@indiana.edu, Home page: http://ella.slis.indiana.edu/~katy/

Description

Information visualization systems and toolkits are becoming available for a large range of visualization and interaction techniques and are used in diverse application domains. This workshop is aimed at gathering experts involved in building such infrastructures to share their views, understand the issues involved and trying to find ways to avoid fragmentation and improve collaborations.

To participate in the workshop, you should submit a semi-structured position paper explaining your view of what an infrastructure should provide, describe what you consider as the main challenges for such infrastructures and describe the capabilities of toolkits of systems you have already built, following a form available here by Sept 30th, 2004.

http://vw.indiana.edu/ivsi2004/



InfoVis Learning Modules

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

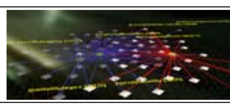
Home | Learning Modules | Software | Data Bases | Compute Resources | References

Learning Modules

Most information visualizations are highly interactive. While a number of excellent textbooks exist, the two-dimensional printouts on paper often cannot convey their true visual appearance and interactive performance. Several textbooks come with accompanying web sites that contain snapshots of user interfaces as well as animations and movies. However, none of them facilitates the exploration, application, evaluation, and comparison of algorithms.

This web page will provide access to a number of learning modules. Each learning module comes with an:

- Introduction
- Discussion of Existing Algorithms
- Learning Task
  - A challenging scenario to use the code in the XML Toolkit to visualize a data set to support a specfic user group.
- Programming Exercise
  - An programming exercise plus an explanation of possible solutions.
- Opportunities & Challenges, and
- References to research papers, online demos, (commercial) applications)



Visualizing Tree Data

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

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

Home | Learning Modules | Software | Data Bases | Compute Resources | References

Learning Modules > Visualizing Tree Data

Description | Usage Hints | Learning Task | Discussion | References | Acknowledgments

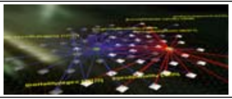
Description

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 rootes trees, all nodes except the root node have only one parent node. Nodes which have no children are called leave nodes. All other nodes are referred to as intermediate nodes.



Student's Project Results

User & Task Analysis for Visualizing Tree Data

- Visualizing the structure of IU's Decision Support System
- Visualizing the co-occurences of keywords in DLib Magazine articles.
- Visualization of the Java API
- Visualizing the the Library of Congress Classification System to retrieve legal materials in a library.

See Handin pages at
http://ella.slis.indiana.edu/~katy/handin/L579-S04/cgi/handinlogin.cgi

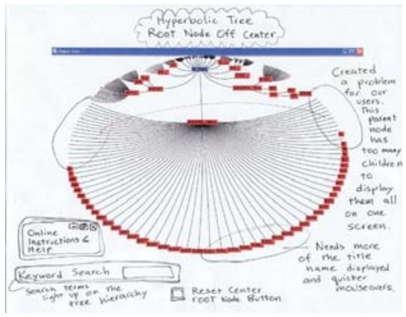
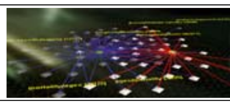
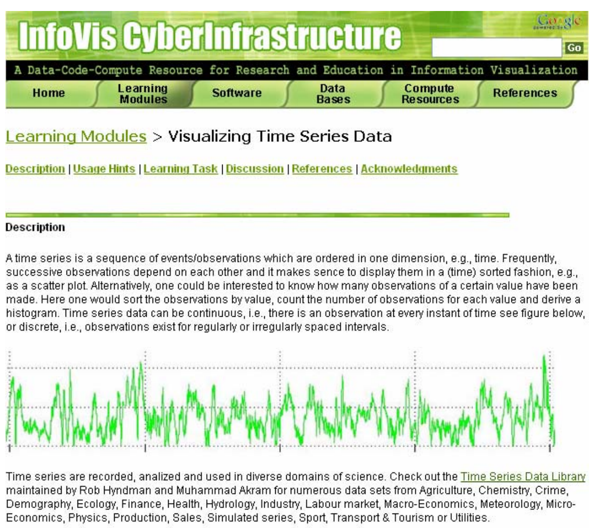


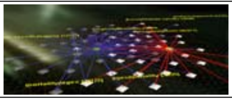
Image by Peter Hook and Rongke Gao



Time Series Analysis & Visualization

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





Student's Project Results

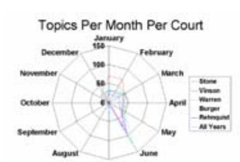
Time Series Analysis & Visualization

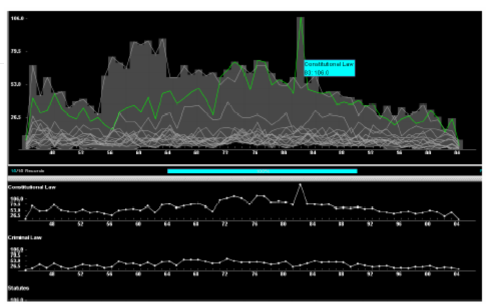
- Using Timesearcher and the Burst Detection Algorithm to Analyze the Stock Market from 1925 to 1945
- Applying Burst and TimeSearcher to Chat Data
- Lab Access Trends
- Quest Atlantis Chat Log Data

See Handin pages at
http://ella.slis.indiana.edu/~katy/handin/L579-S04/cgi/handinlogin.cgi

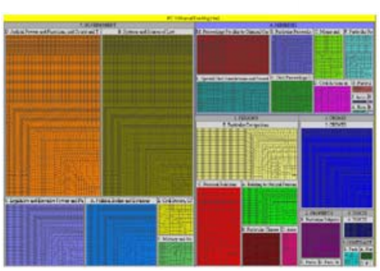
Visualizing the Work of the United States Supreme Court Based on Time Data and Top Level West Topics

by Peter A. Hook & Rongke Gao

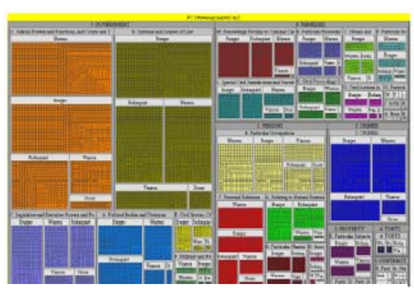




Top fifteen most occurring topics from 1944 to 2004 in Timesearcher



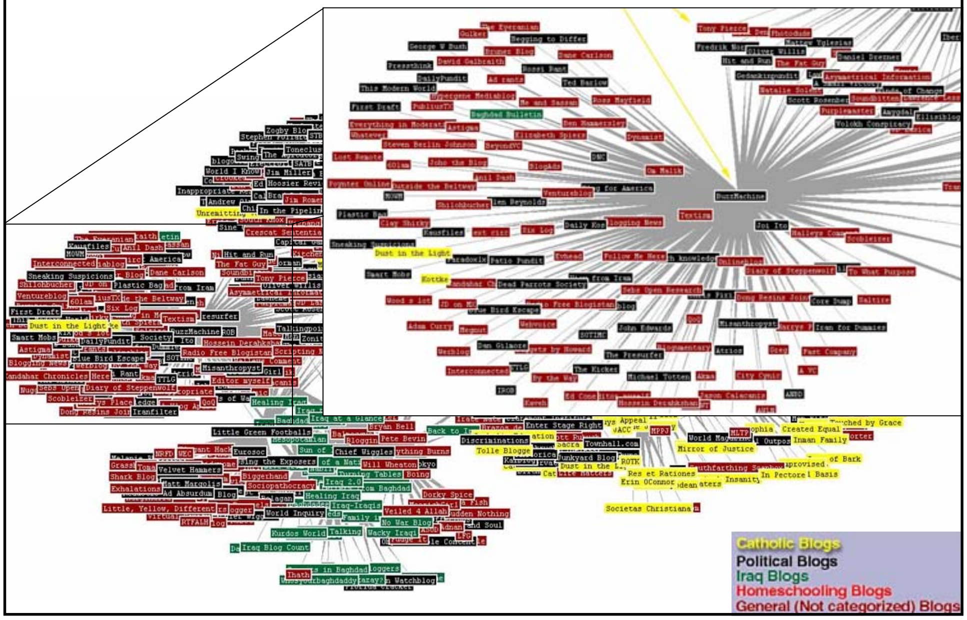
All topics grouped by West Category and Sub-Category grouped over the entire lengths of the data set

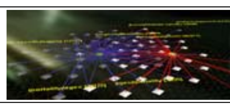


All topics by West Category and Sub-Category grouped corresponding to the five chief justices

Visualizing Niches of the Blog Universe

BY Mike Tyworth and Elijah Wright





IVC DB Data Sets *(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 Natioanl 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: 2,582,647
Years covered: 1976-2003
Size: 350 MB

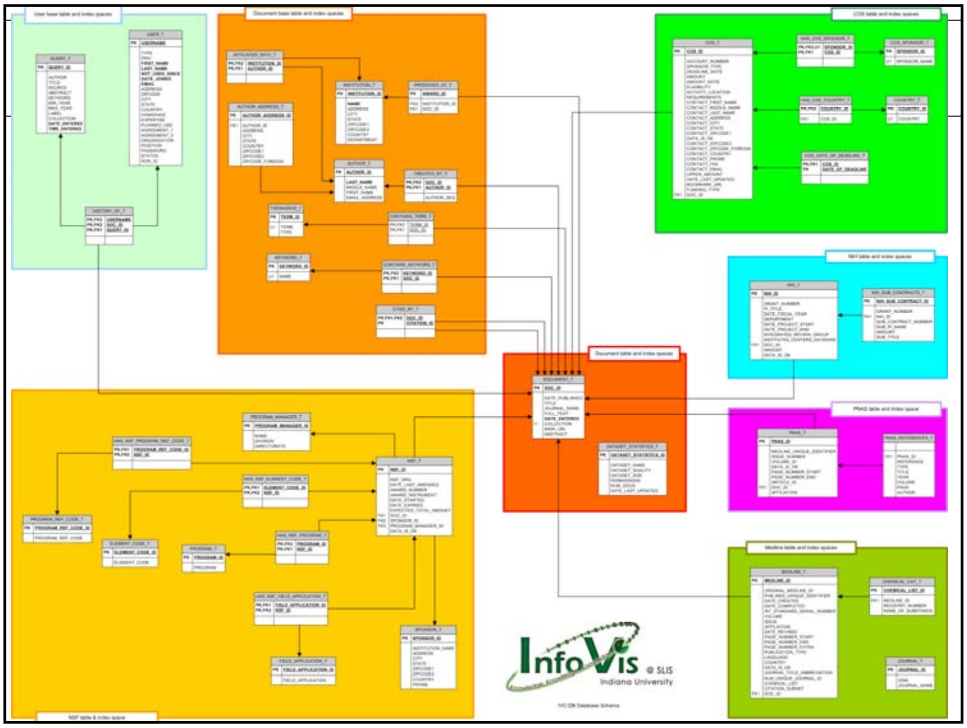
Grant Awards

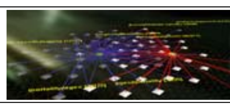
National Science Foundation (NSF)
Number of Entries: 181,132
Years covered: 1985-2002
Size: 400 MB

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

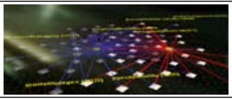
Funding Opportunities

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





Questions?

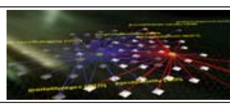


Courses

- **L597** Structural Data Mining and Modeling
  Fall 2004 (http://ella.slis.indiana.edu/~katy/L597)

- **L579** Information Visualization (formerly L697)
  Spring 2005 (http://ella.slis.indiana.edu/~katy/L579)

Both courses have final projects in which students work with outside collaborators.

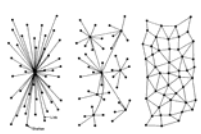


L597 Structural Data Mining and Modeling

This course

- Introduces students to major methods, theories, and applications of structural data mining and modeling.
- Covers elementary graph theory and matrix algebra, data collection, structural data mining, data modeling, and applications.

Upon taking this course students will be able to analyze and describe real networks (power grids, WWW, social networks, etc.) as well as relevant phenomena such as disease propagation, search, organizational performance, social power, and the diffusion of innovations.

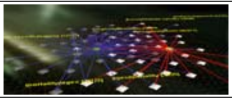


Format: Lectures and 4-5 labs.

Class Webpage: http://ella.slis.indiana.edu/~katy/L597

Fall 2004 Talk Series on Networks and Complex Systen

http://vw.indiana.edu/talks-fall04/



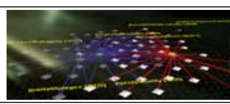
L579 Information Visualization

This course covers

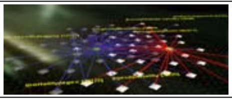
- Perceptual basis of information visualization.
- Data mining algorithms that enable extraction of relationships in data.
- Visualization and interaction techniques.
- Discussions of systems that drive research and development, and
- Future trends and remaining fundamental problems in the field.

Students do weekly readings, provide a presentation on specific readings, do Java projects, and participate in class & online discussion.

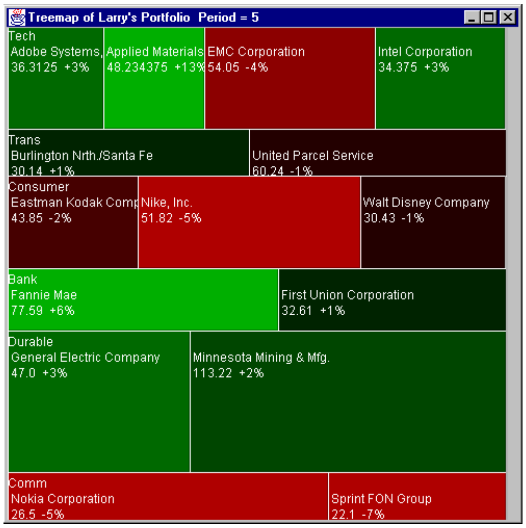
Class Webpage: http://ella.slis.indiana.edu/~katy/L579

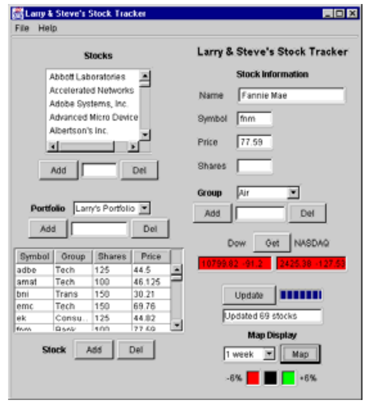


Student Projects

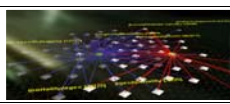


Stock Tracker
Larry Mongin & Steve Rice

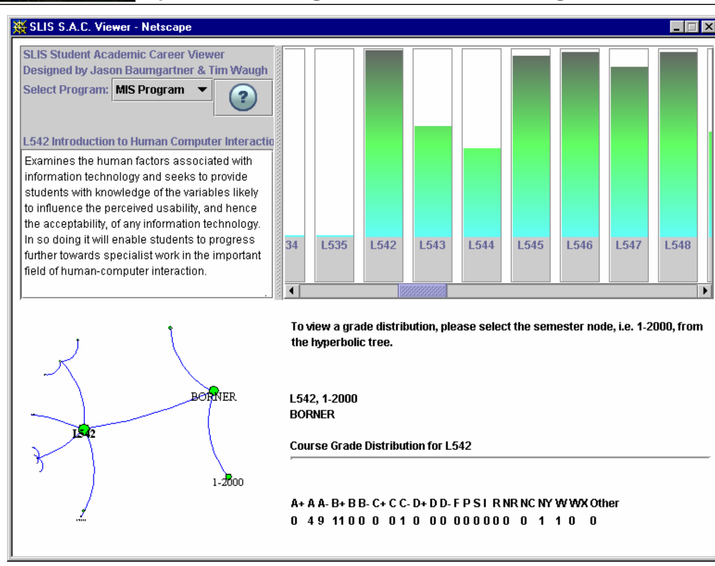




Katy Börner: SLIS Orientation Session, Fall 2004.



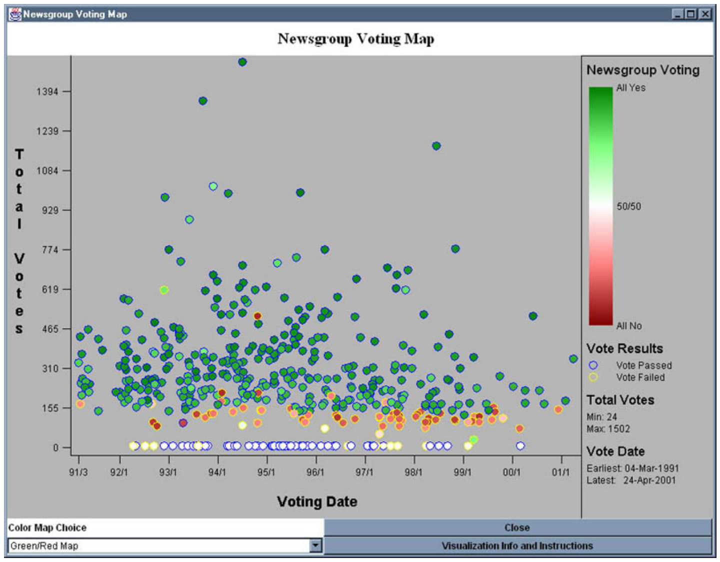
SLIS Student Academic Career Viewer
Jason Baumgartner & Tim Waugh



Katy Börner: SLIS Orientation Session, Fall 2004.

Newsgroup Votes Visualization

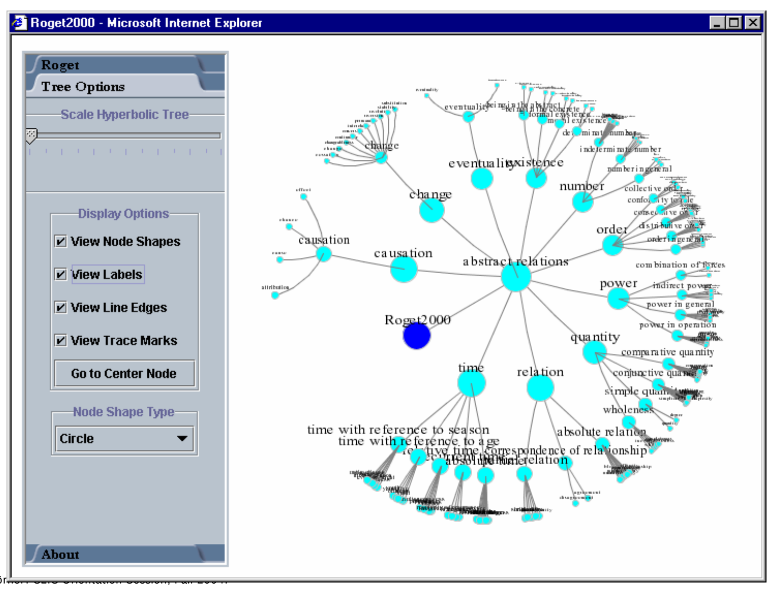
David Heald (Collaborator John C. Paolillo)



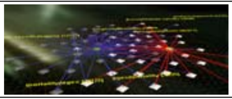
Katy Börner: SLIS Orientation Session, Fall 2004.

Hyperbolic Tree Visualization of Roget's Thesaurus

Jason Baumgartner & Tim Waugh (Collaborator John Old)



Katy Börner: SLIS Orientation Session, Fall 2004.

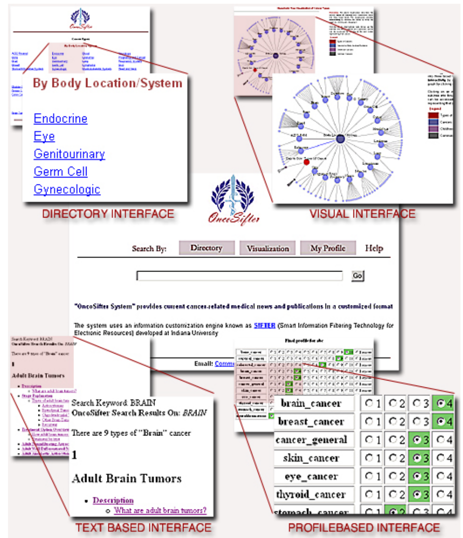


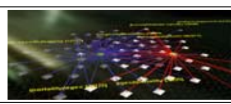
Oncosifter – Personalized Cancer Information

- Filters cancer related news and medical information from Medlineplus and Cancer.gov.
- Provides hierarchical search & browsing interface.

Faculty Client: Javed Mostafa, SLIS, IUB

To be presented at the Information Visualization Interfaces for Retrieval and Analysis Workshop @ JCDL, Houston, TX, May 31, 2003.

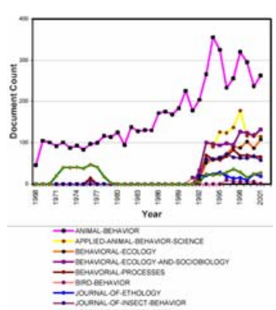


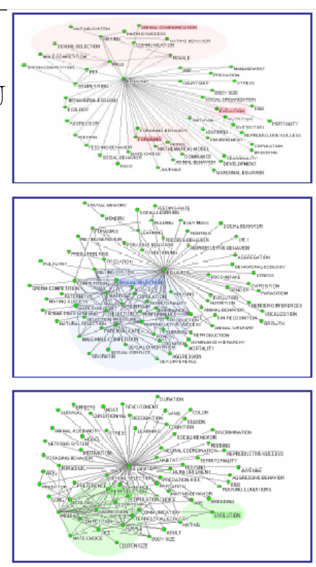


Visualizing the Animal Behavior Domain

Faculty Client:

Emilia Martins & Terry Ord, Biology, IU





Sidharth Thakur, Ketan Mane, Katy Börner, Emilia Martins & Terry Ord. Content Coverage of Animal Behavior Data. *Submitted to IEEE InfoVis*, 2003.