Analysis and Visualization of Science

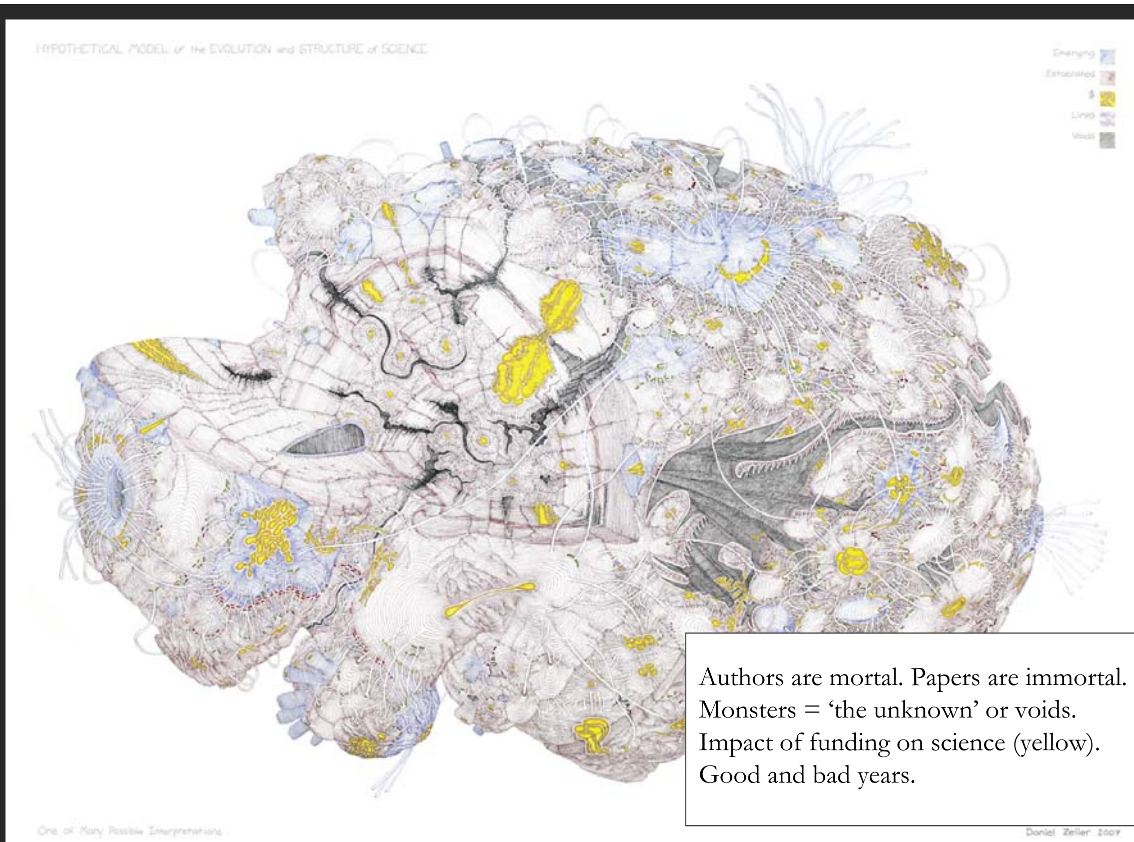
Angela Zoss, M.S.
Research Assistant, Cyberinfrastructurefor Network Science Center
Doctoral Student, School of Library and Information Science
Indiana University, Bloomington, IN
amzoss@indiana.edu

February 8, 2011 – Workshop on Scholarly Communication and Informetrics
iConference 2011, Seattle, Washington

Analysis and Visualization of Science

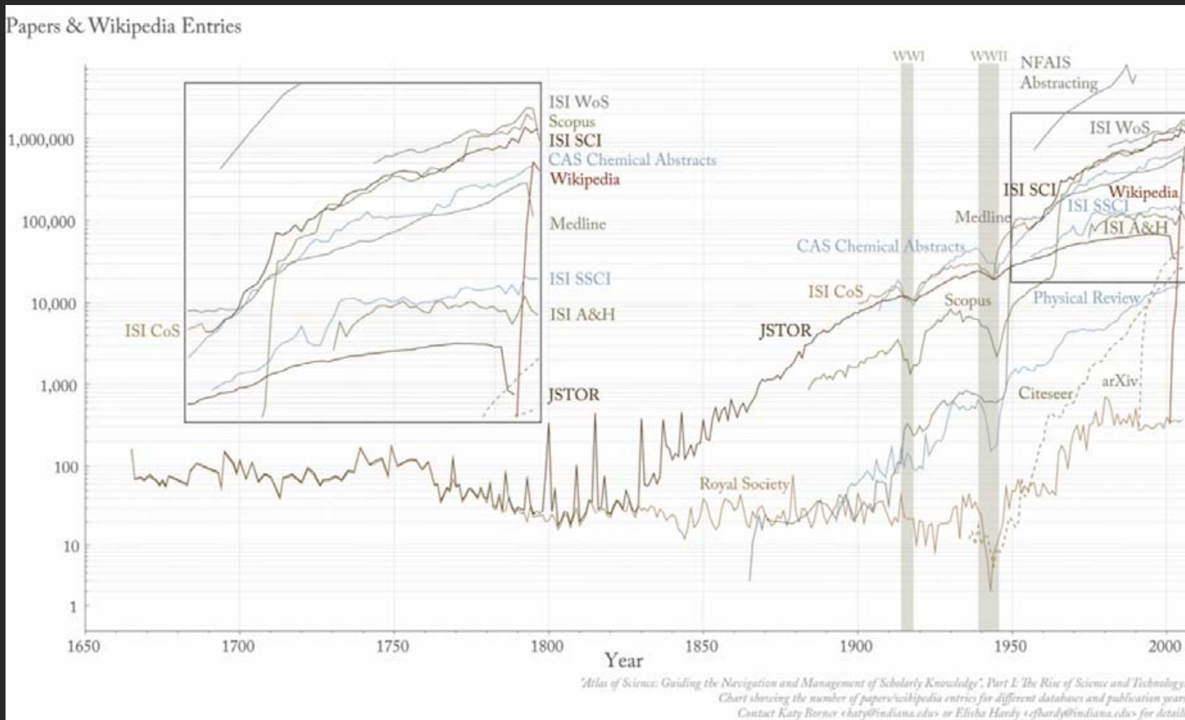
- What is science?
- Why do we analyze and visualize science?
- How do we analyze and visualize science?

Conceptualizing Science

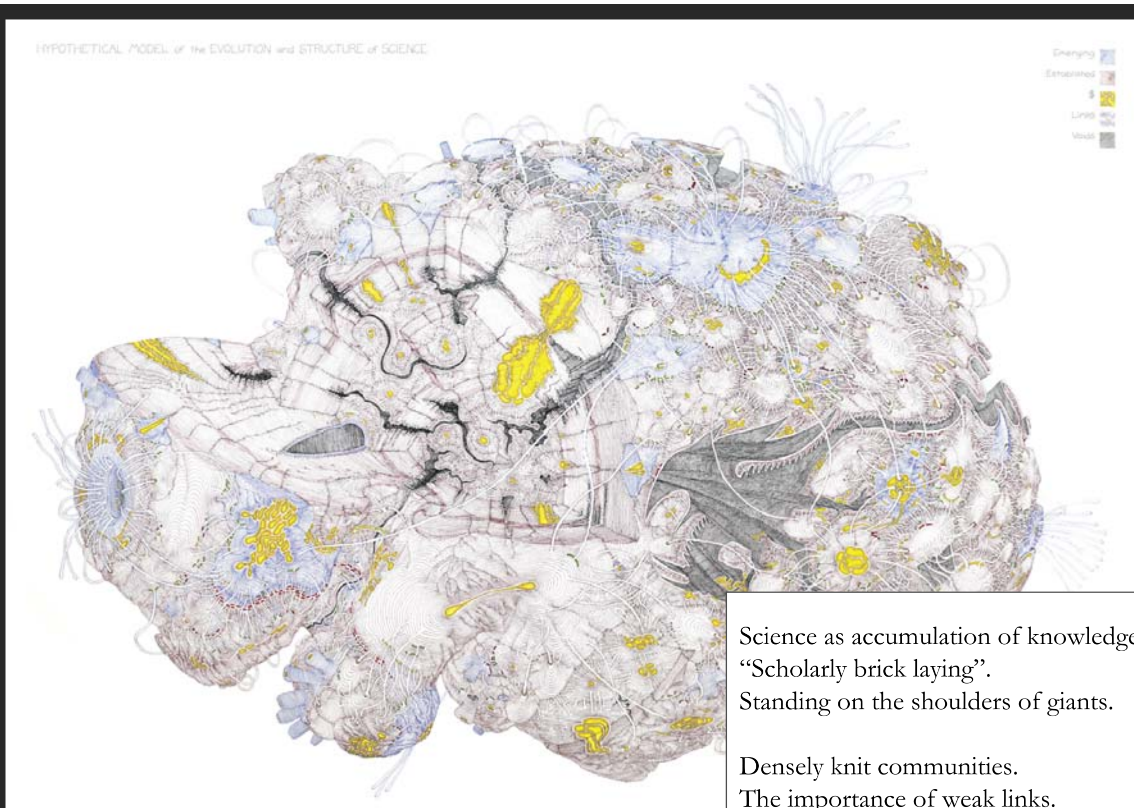


Authors are mortal. Papers are immortal.
Monsters = 'the unknown' or voids.
Impact of funding on science (yellow).
Good and bad years.

Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



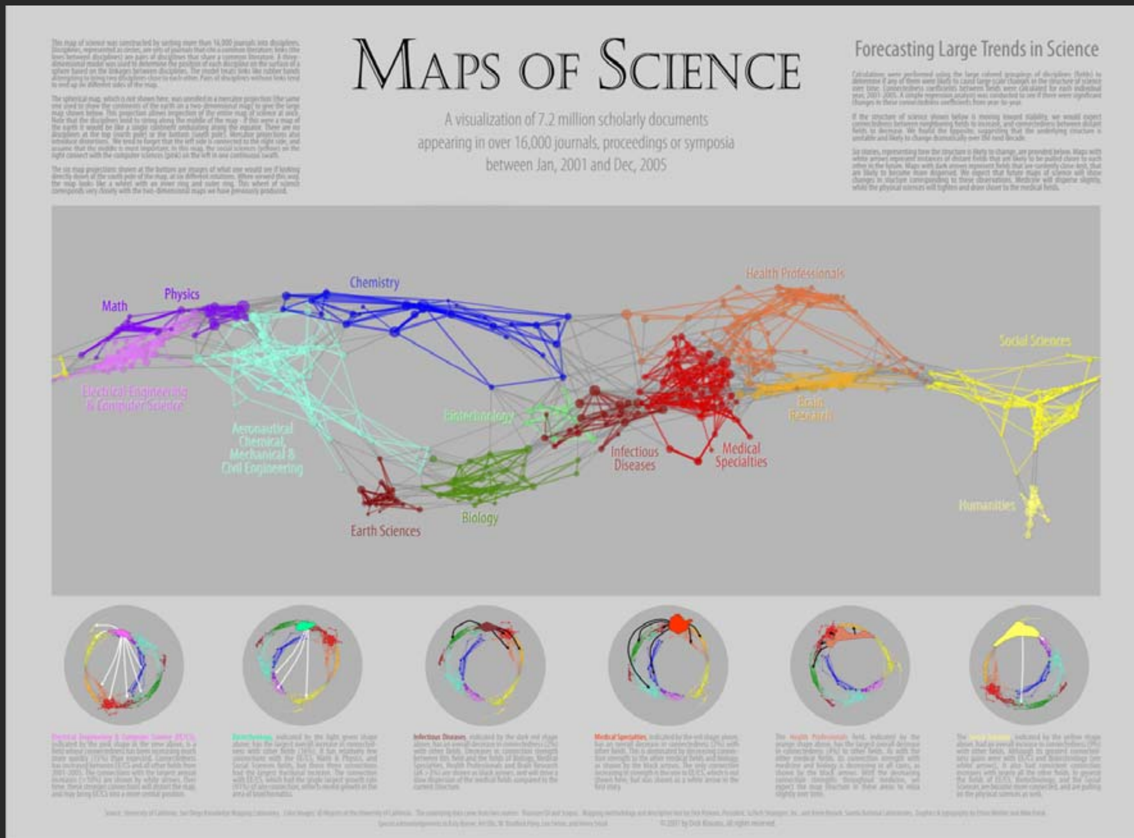
Atlas of Science - Katy Borner - 2010



Science as accumulation of knowledge.
"Scholarly brick laying".
Standing on the shoulders of giants.

Densely knit communities.
The importance of weak links.

Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007

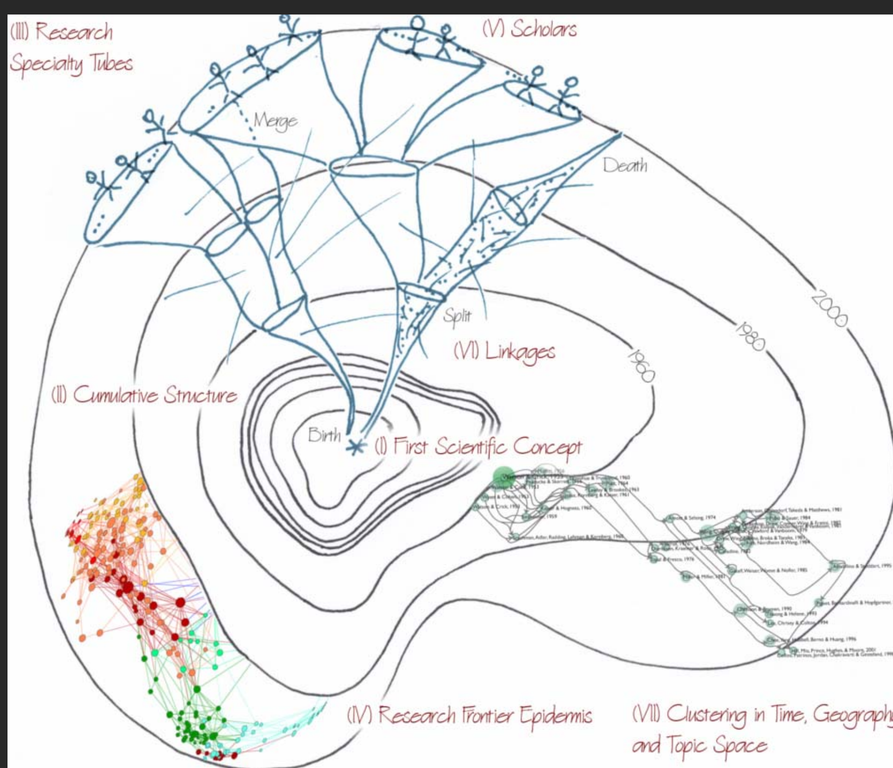


Maps of Science: Forecasting Large Trends in Science - Richard Klavans, Kevin Boyack - 2007

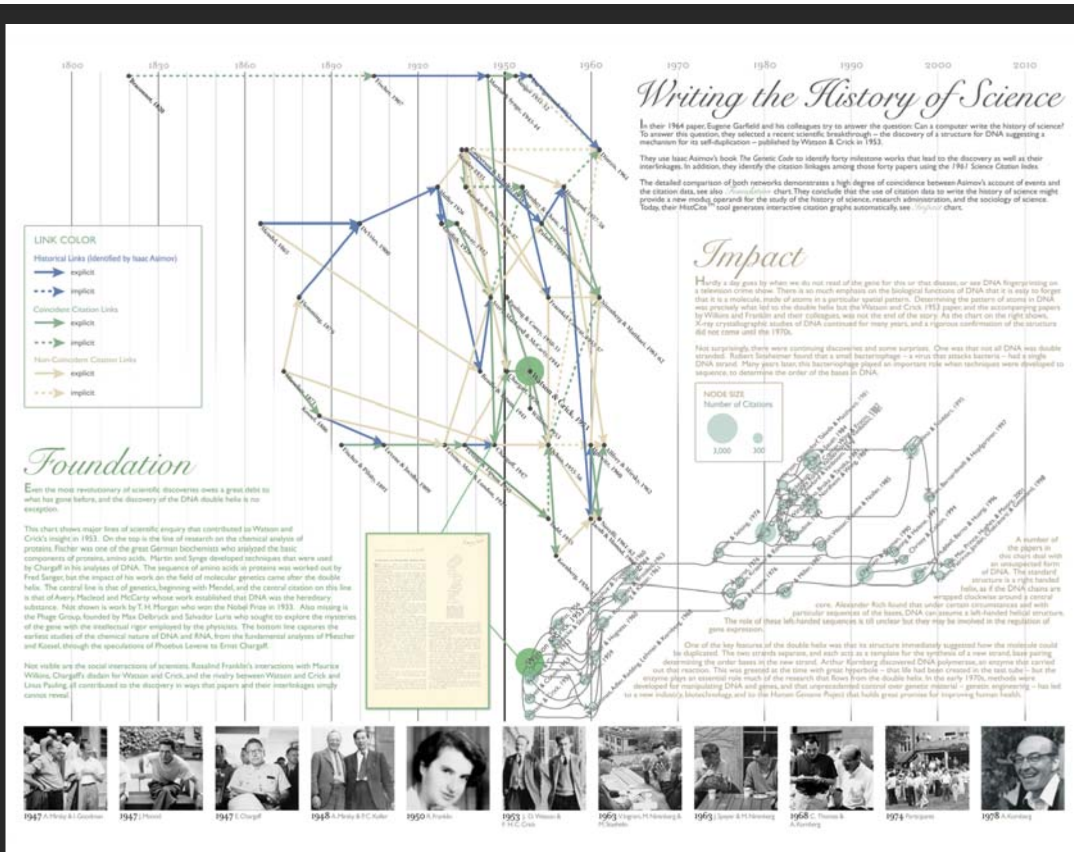
In Terms of Geography - Andre Skupin - 2005

Areas of science are tube shaped.

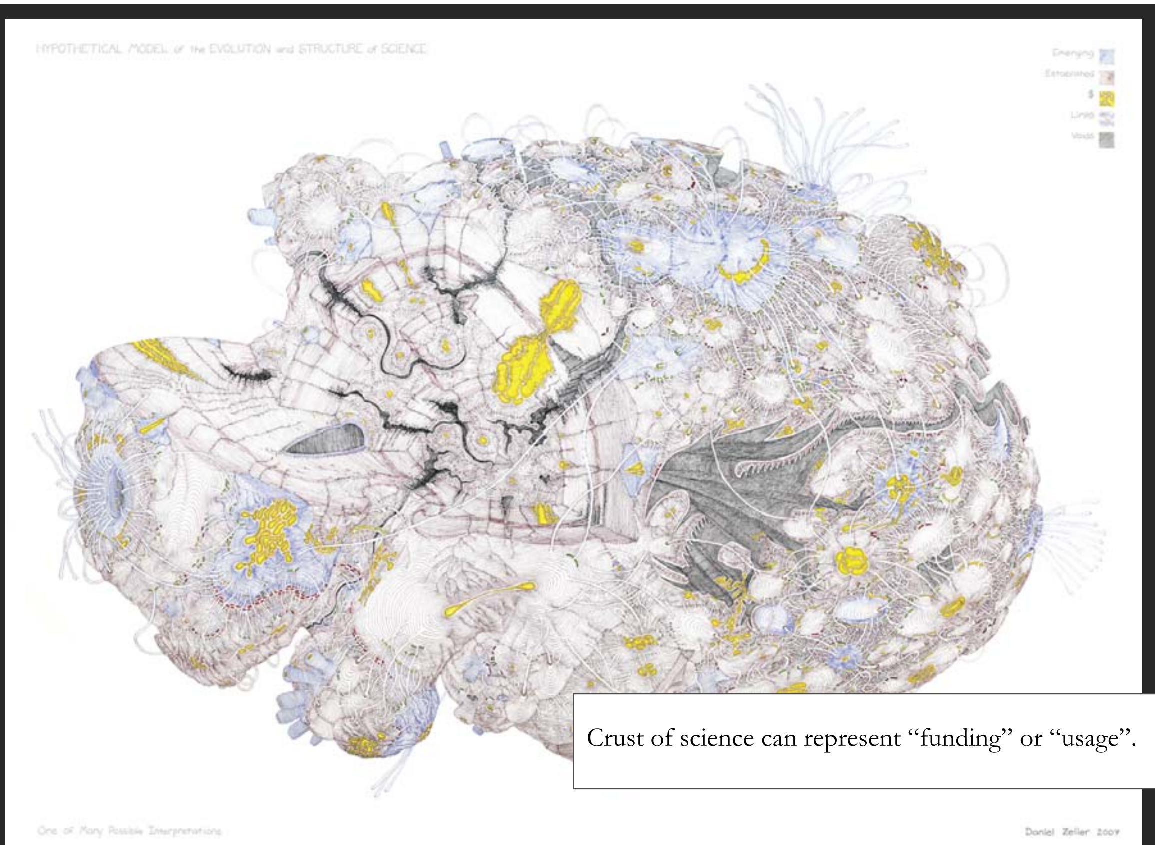
Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



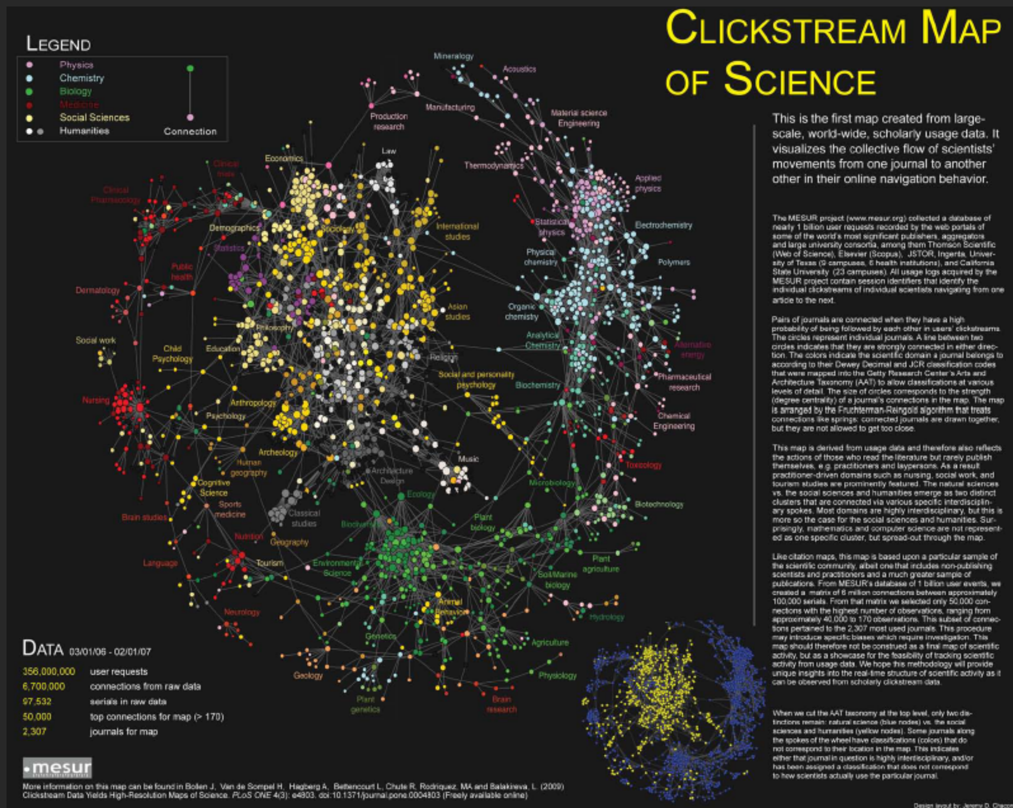
Atlas of Science - Katy Borner - 2010



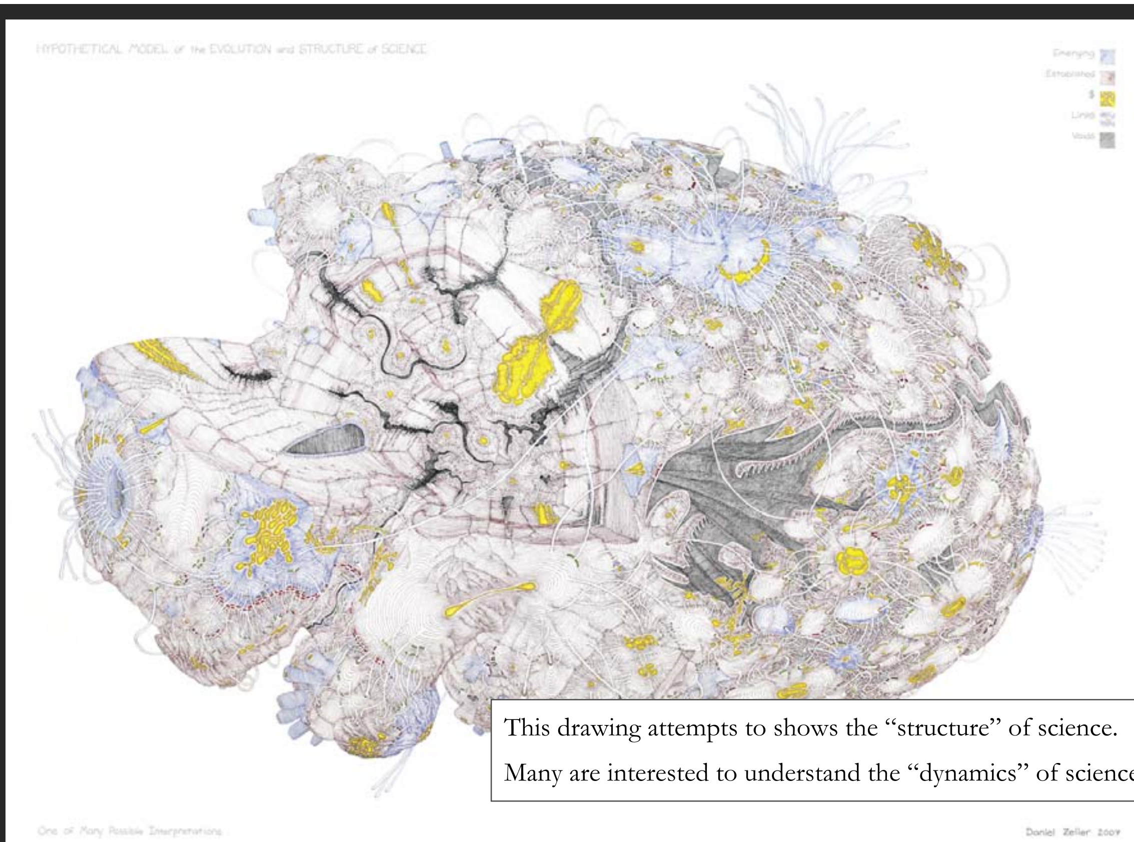
HistCite™ Visualization of DNA Development - Eugene Garfield, Elisha Hardy, Katy Borner, Ludmila Pollock, Jan Witkowski- 2006



Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



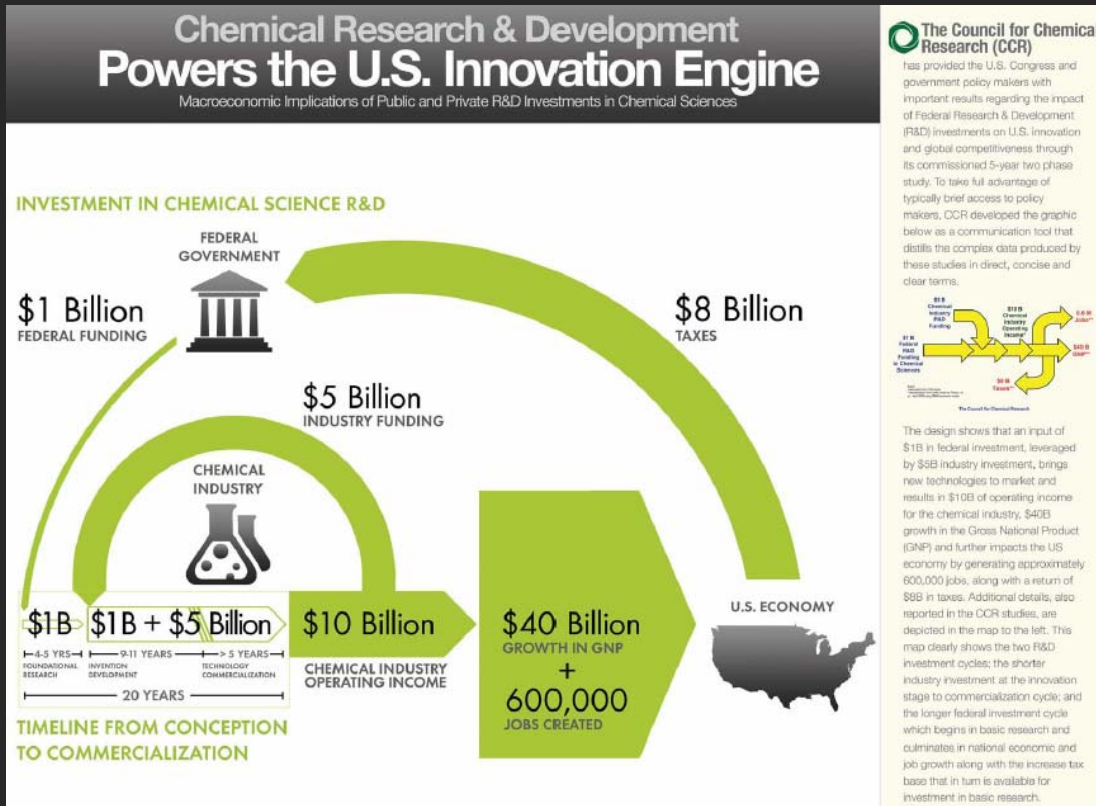
A Clickstream Map of Science – Bollen, Johan, Herbert Van de Sompel, AricHagberg, Luis M.A. Bettencourt, Ryan Chute, Marko A. Rodriquez, Lyudmila Balakireva - 2008



This drawing attempts to shows the "structure" of science.

Many are interested to understand the "dynamics" of science.

Hypothetical Model of the Evolution of Science - Daniel Zeller - 2007



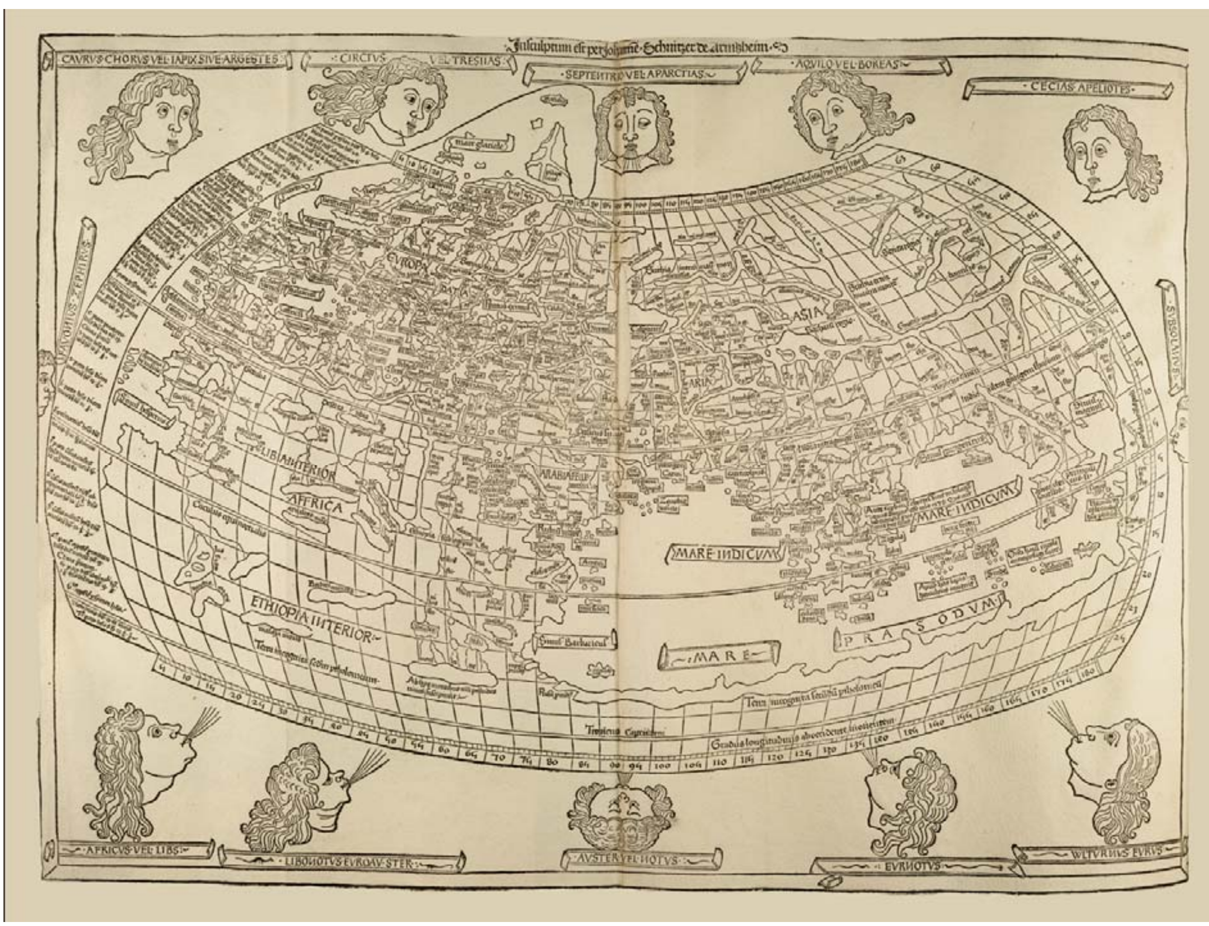
Council for Chemical Research - Chemical R&D Powers the U.S. Innovation Engine. Washington, DC. Courtesy of the Council for Chemical Research - 2009

Why Map Science?

Cartographic maps of physical places have guided mankind's explorations for centuries.

They enabled the discovery of new worlds while also marking territories inhabited by the unknown.

Without maps, we would be lost.



Domain maps of abstract semantic spaces aim to serve today's explorers navigating the world of science.

These maps are generated through a scientific analysis of large-scale scholarly datasets in an effort to connect and make sense of the bits and pieces of knowledge they contain.

They can be used to identify objectively major research areas, experts, institutions, collections, grants, papers, journals, and ideas in a domain of interest. Science maps can provide overviews of "all-of-science" or of a specific area.

They can show homogeneity vs. heterogeneity, cause and effect, and relative speed. They allow us to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research.

Information Needs for Science Map User Groups

Advantages for Funding Agencies

- Supports monitoring of (long-term) money flow and research developments, evaluation of funding strategies for different programs, decisions on project durations, funding patterns.
- Staff resources can be used for scientific program development, to identify areas for future development, and the stimulation of new research areas.

Advantages for Researchers

- Easy access to research results, relevant funding programs and their success rates, potential collaborators, competitors, related projects/publications **(research push).**
- More time for research and teaching.

Advantages for Industry

- Fast and easy access to major results, experts, etc.
- Can influence the direction of research by entering information on needed technologies **(industry-pull)**.

Advantages for Publishers

- Unique interface to their data.
- Publicly funded development of databases and their interlinkage.

For Society

- Dramatically improved access to scientific knowledge and expertise.

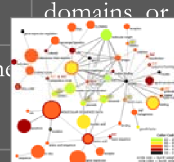
Analysis and Visualization of Science

Type of Analysis vs. Scale of Level of Analysis

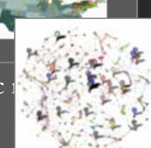
	Micro/Individual (1-100 records)	Meso/Local (101–10,000 records)	Macro/Global (10,000 < records)
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science.
Temporal Analysis (When)	Funding portfolio of one individual	Mapping topic bursts in 20-years of PNAS	113 Years of physics Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a states intellectual landscape	PNAS publications
Topical Analysis (What)	Base knowledge from which one grant draws.	Knowledge flows in Chemistry research	VxOrd/Topic maps of NIH funding
Network Analysis (With Whom?)	NSF Co-PI network of one individual	Co-author network	NSF's core competency

Type of Analysis vs. Scale of Level of Analysis

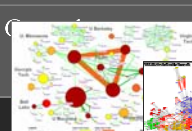
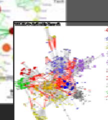
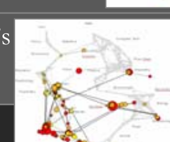
	Micro/Individual (1-100 records)	Meso/Local (101–10,000 records)	Macro/Global (10,000 < records)
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS SA, all of scie
Temporal Analysis (When)	Funding portfolio of one individual	ic bursts f PNAS	113 Years of pl Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a st intellectual la	PNAS p
Topical Analysis (What)	.	flows in esearch	VxOrd/Topic NIH funding
Network Analysis (With Whom?)	NSF work of one	C k	NSF's cy







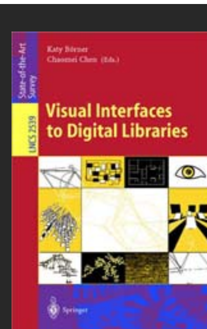




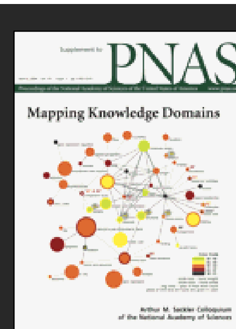
Process of Computational Scientometrics

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 •Atributes (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 •Pathfinder networks (PFNet) •Self-organizing maps (SOM) incl. 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.

Computational Scientometrics: Studying Science by Scientific Means





- Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, Volume 37, Chapter 5, pp. 179-255. http://ivl.slis.indiana.edu/km/pub/2003-borner-arist.pdf
- 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). http://www.pnas.org/content/vol101/suppl_1/
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). Network Science. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, Volume 41, Chapter 12, pp. 537-607. http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf
- Places & Spaces: Mapping Science exhibit, see also http://scimaps.org.
- Börner. Katy. (2010). Atlas of Science: Visualizing What We Know. MIT Press. http://scimaps.org/atlas
- Börner, Katy. (March 2011). Plug-and-Play Macroscopes.*Communications of the ACM*.

Science of Science Cyberinfrastructure

Overview

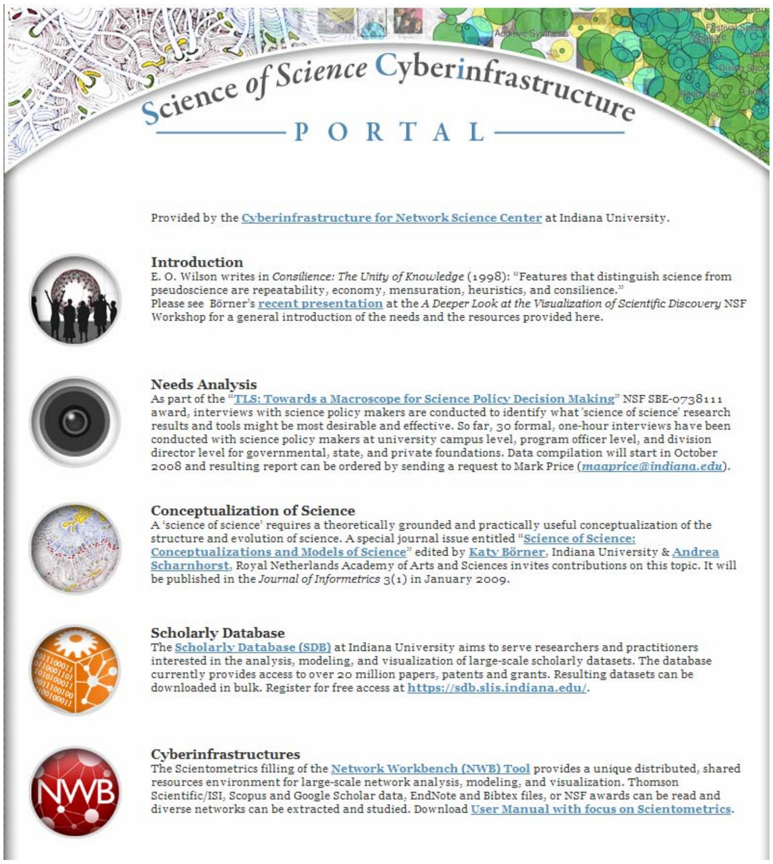
What cyberinfrastructure will be required to measure, model, analyze, and communicate scholarly data and, ultimately, scientific progress?

This talk presents our efforts to create a science of science cyberinfrastructure that supports:

- Data access and federation via the **Scholarly Database**, http://sdb.slis.indiana.edu,
- Data preprocessing, modeling, analysis, and visualization using plug-and-play cyberinfrastructures such as the **Sci^2 Tool**, http://sci2.cns.iu.edu, and
- Communication of science to a general audience via the **Mapping Science Exhibit** at http://scimaps.org.

The following demos should be particularly interesting for those interested to

- Map their very own domain of research,
- Test and compare data federation, mining, visualization algorithms on large scale datasets,
- Use advanced network science algorithms in their own research.



Science of Science Cyberinfrastructure

PORTAL

Provided by the Cyberinfrastructure for Network Science Center at Indiana University.

Introduction
E. O. Wilson writes in *Consilience: The Unity of Knowledge* (1998): "Features that distinguish science from pseudoscience are repeatability, economy, mensuration, heuristics, and consilience."
Please see Börner's recent presentation at the *A Deeper Look at the Visualization of Scientific Discovery* NSF Workshop for a general introduction of the needs and the resources provided here.

Needs Analysis
As part of the "TLS: Towards a Macroscope for Science Policy Decision Making" NSF SBE-0738111 award, interviews with science policy makers are conducted to identify what 'science of science' research results and tools might be most desirable and effective. So far, 30 formal, one-hour interviews have been conducted with science policy makers at university campus level, program officer level, and division director level for governmental, state, and private foundations. Data compilation will start in October 2008 and resulting report can be ordered by sending a request to Mark Price (maaprice@indiana.edu).

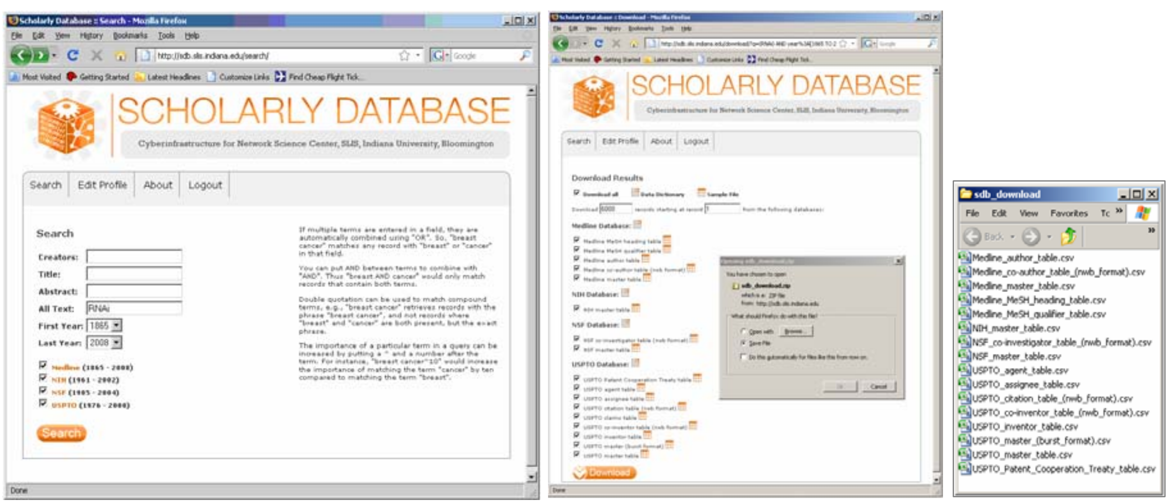
Conceptualization of Science
A 'science of science' requires a theoretically grounded and practically useful conceptualization of the structure and evolution of science. A special journal issue entitled "Science of Science: Conceptualizations and Models of Science" edited by Katy Börner, Indiana University & Andrea Scharnhorst, Royal Netherlands Academy of Arts and Sciences invites contributions on this topic. It will be published in the *Journal of Informetrics* 3(1) in January 2009.

Scholarly Database
The Scholarly Database (SDB) at Indiana University aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale scholarly datasets. The database currently provides access to over 20 million papers, patents and grants. Resulting datasets can be downloaded in bulk. Register for free access at https://sdb.slis.indiana.edu/.

Cyberinfrastructures
The Scientometrics filling of the Network Workbench (NWB) Tool provides a unique distributed, shared resources environment for large-scale network analysis, modeling, and visualization. Thomson Scientific/ISI, Scopus and Google Scholar data, EndNote and Bibtex files, or NSF awards can be read and diverse networks can be extracted and studied. Download User Manual with focus on Scientometrics.

http://sci.slis.indiana.edu

Scholarly Database (*http://sdb.cns.iu.edu*)



The Scholarly Database at Indiana University provides free access to 25,000,000 papers, patents, and grants. Since March 2009, users can also download networks, e .g., co-author, co-investigator, co-inventor, patent citation, and tables for burst analysis.

Sci² Tool for Science of Science (*http://sci2.cns.iu.edu*)

- Explicitly designed for SoS research and practice, well documented, easy to use.
- Empowers many to run common studies while making it easy for exports to perform novel research.
- Advanced algorithms, effective visualizations, and many (standard) workflows.
- Supports micro-level documentation and replication of studies.
- Is open source—anybody can review and extend the code, or use it for commercial purposes.


nature Vol 464|25 March 2010

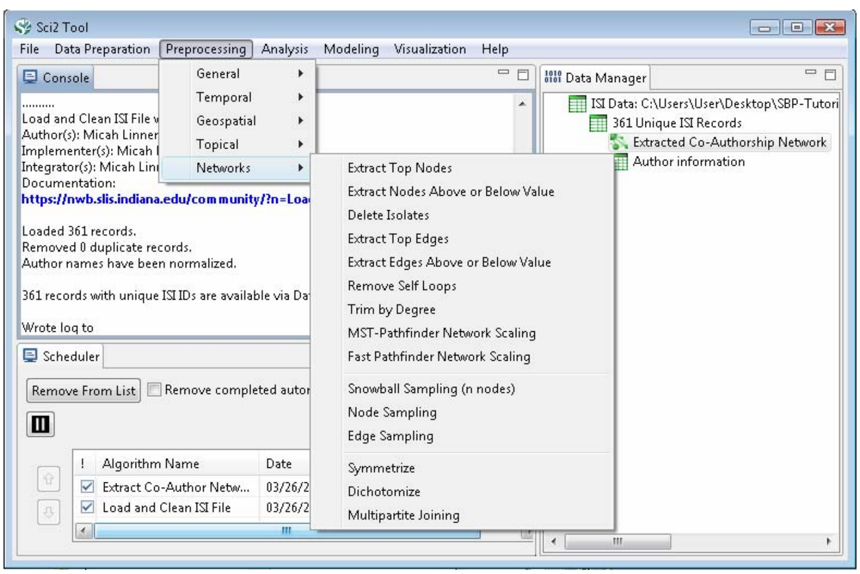
OPINION

Let's make science metrics more scientific

To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity, says **Julia Lane**.

SUMMARY
- Existing metrics have known flaws
- A reliable, open, joined-up data infrastructure is needed
- Data should be collected on the full range of scientists' work
- Social scientists and economists should be involved

Sci² Tool for Science of Science Research and Practice



Acknowledgments
This work is supported in part by the Cyberinfrastructure for Network Science center and the School of Library and Information Science at Indiana University, the National Science Foundation under Grant No. SBE-0738111 and IIS-0513650, and the James S. McDonnell Foundation.

INDIANA UNIVERSITY SCHOOL OF LIBRARY AND INFORMATION SCIENCE
James S. McDonnell Foundation

Sci² Tool for Science of Science Research and Practice

Supported Input file formats:

- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net) &Pajek .Matrix (*.mat)
- NWB (*.nwb)
- TreeML (*.xml)
- Edge list (*.edge)
- **CSV (*.csv)**
- **ISI (*.isi)**
- **Scopus (*.scopus)**
- **NSF (*.nsf)**
- **Bibtex (*.bib)**
- **Endnote (*.enw)**

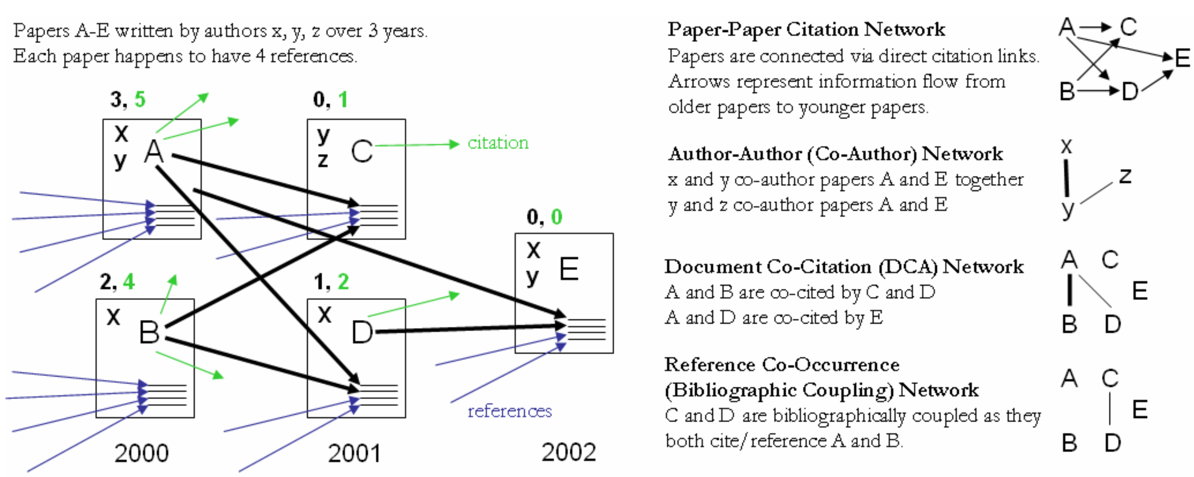
Output file formats:

GraphML (*.xml or *.graphml)
Pajek .MAT (*.mat)
Pajek .NET (*.net)
NWB (*.nwb)
XGMML (*.xml)
CSV (*.csv)

http://sci2.wiki.cns.iu.edu/2.3+Data+Formats

Network Extraction

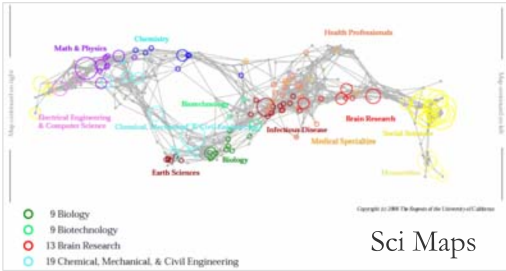
Sample paper network (left) and four different network types derived from it (right).
From ISI files, about 30 different networks can be extracted.

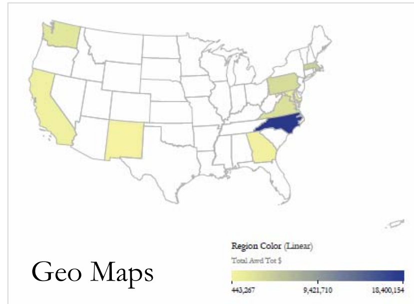


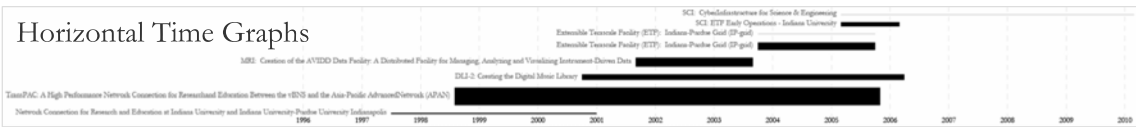
Local citation counts (within this dataset) are given in **black** and global citation counts (ISI times cited) are given in green above each paper.

Sci^2 Tool

Plugins that render into Postscript files:







Börner, Katy, Huang, Weixia (Bonnie), Linnemeier, Micah, Duhon, Russell Jackson, Phillips, Patrick, Ma, Nianli, Zoss, Angela, Guo, Hanning & Price, Mark. (2009). Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Scholarly Database and the Network Workbench Tool. Proceedings of ISSI 2009: 12th International Conference on Scientometrics and Informetrics, Rio de Janeiro, Brazil, July 14-17 . Vol. 2, pp. 619-630.

Exemplary Analyses and Visualizations

Individual Level

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks (p. 54-65)

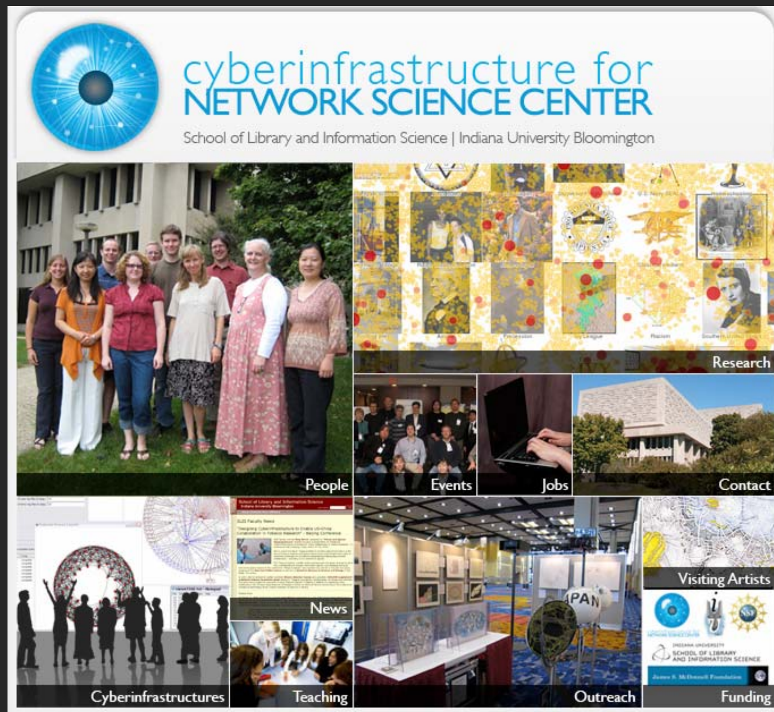
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U (p. 49-53)

Institution Level

C. Indiana U, Cornell U, and Michigan U, extracting, and comparing Co-PI networks (p. 65-69)

Scientific Field Level

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data (p. 77-85)



All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.iu.edu