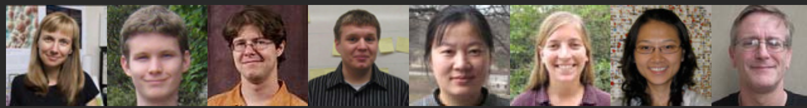


Open Data and Open Code for S&T Assessment

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

With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Angela M. Zoss, Hanning Guo, Mark A. Price, Scott Weingart

Northwestern Institute on Complex Systems (NICO) Annual Conference
Northwestern University, IL
September 3, 2009

Overview

Science of Science Studies

Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu):

- **Scholarly Database (SDB)** (http://sdb.slis.indiana.edu) that provides free access to 23 million scholarly records
- **Sci^2 Tool** which reads SDB data and supports the identification of activity bursts, the extraction and display of co-author/inventor/investigator networks, and topical analysis, among others.

Mapping Science Exhibit

Overview

Science of Science Studies

Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu):

- **Scholarly Database (SDB)** (http://sdb.slis.indiana.edu) that provides free access to 23 million scholarly records
- **Sci² Tool** which reads SDB data and supports the identification of activity bursts, the extraction and display of co-author/inventor/investigator networks, and topical analysis, among others.

Mapping Science Exhibit

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

- *Börner, Katy, Ma, Nianli, Duhon, Russell Jackson & Zoss, Angela. (2009).* ***Science & Technology Assessment Using Open Data and Open Code.*** *IEEE Intelligent Systems. Vol. 24(4), 78-81, IEEE Computer Systems..*

- ***Places & Spaces: Mapping Science*** *exhibit, see also* http://scimaps.org.

Computational Scientometrics Opportunities

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.

Process of Computational Scientometrics

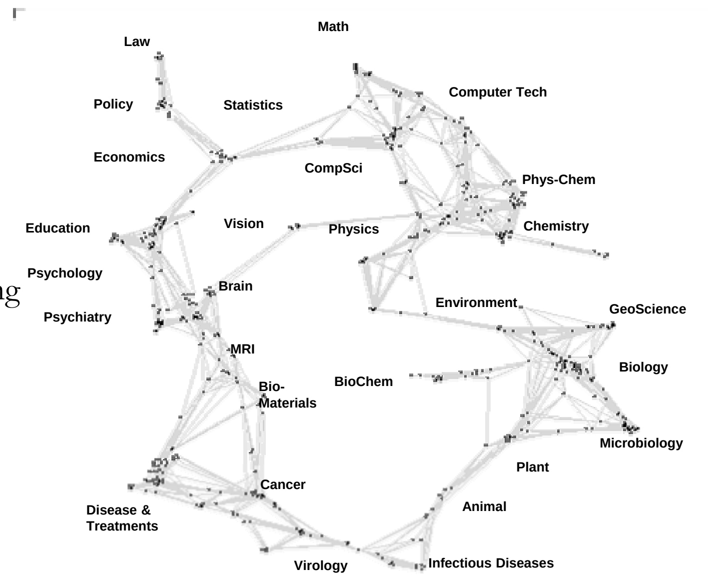
DATA EXTRACTION	UNIT OF ANALYSIS	MEASURES	LAYOUT (often one code does both similarity and ordination steps)		DISPLAY
			SIMILARITY	ORDINATION	
SEARCHES	COMMON	COUNTS/FREQUENCIES	SCALAR (unit by unit matrix)	DIMENSIONALITY REDUCTION	INTERACTION
ISI	CHOICES	Attributes (e.g. terms)	Direct citation	Eigenvector/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

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.

Latest 'Base Map' of Science

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007). Mapping the Structure and Evolution of Chemistry Research. 11th International Conference on Scientometrics and Informetrics. pp. 112-123.

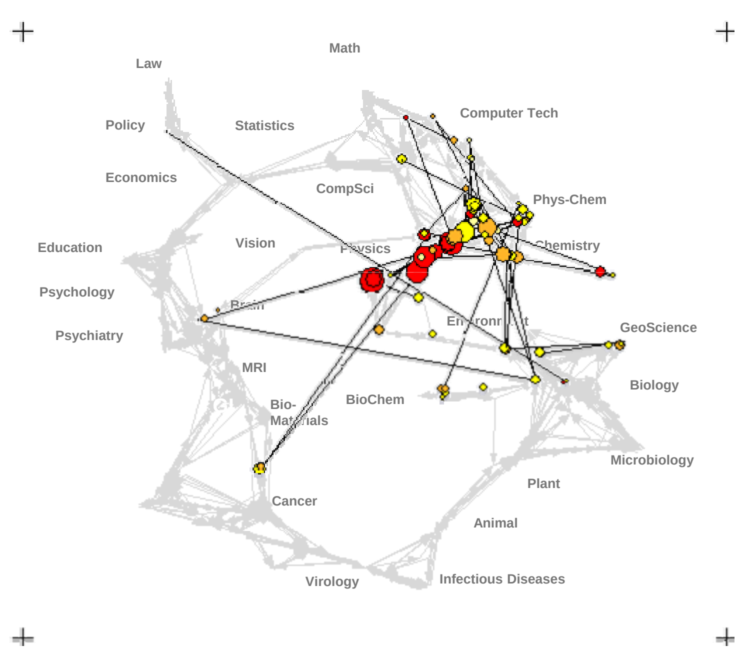
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y) positions for each journal cluster
  - by association, (x,y) positions for each journal



Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

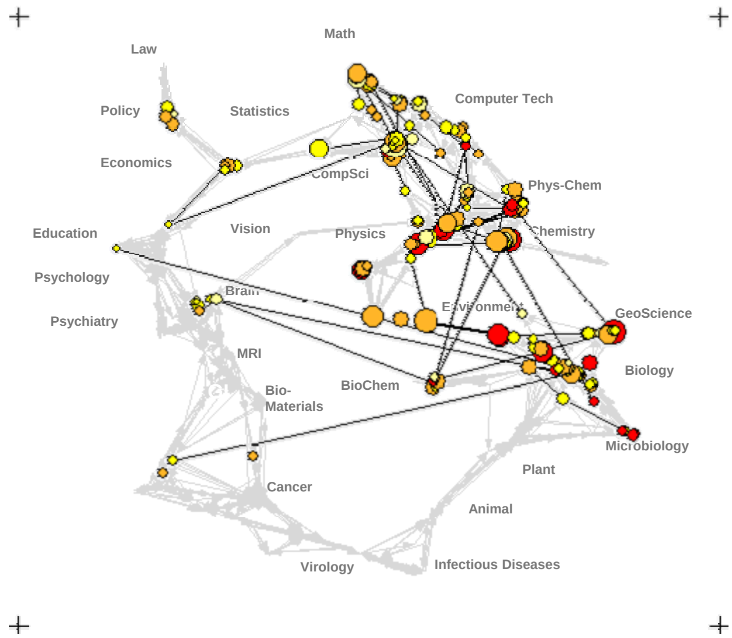
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

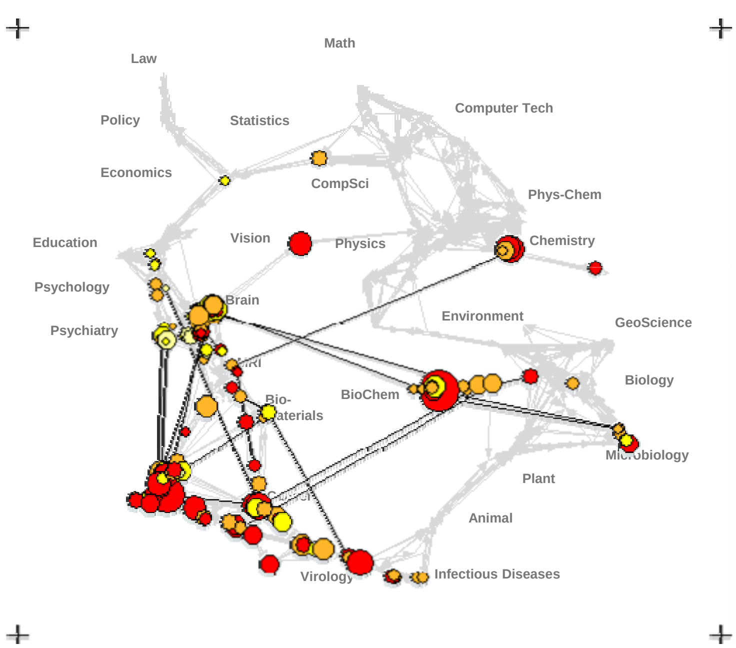
Funding Patterns of the National Science Foundation (NSF)



Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

Funding Patterns of the National Institutes of Health (NIH)



Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

Funding Patterns of the National Institutes of Health (NIH)

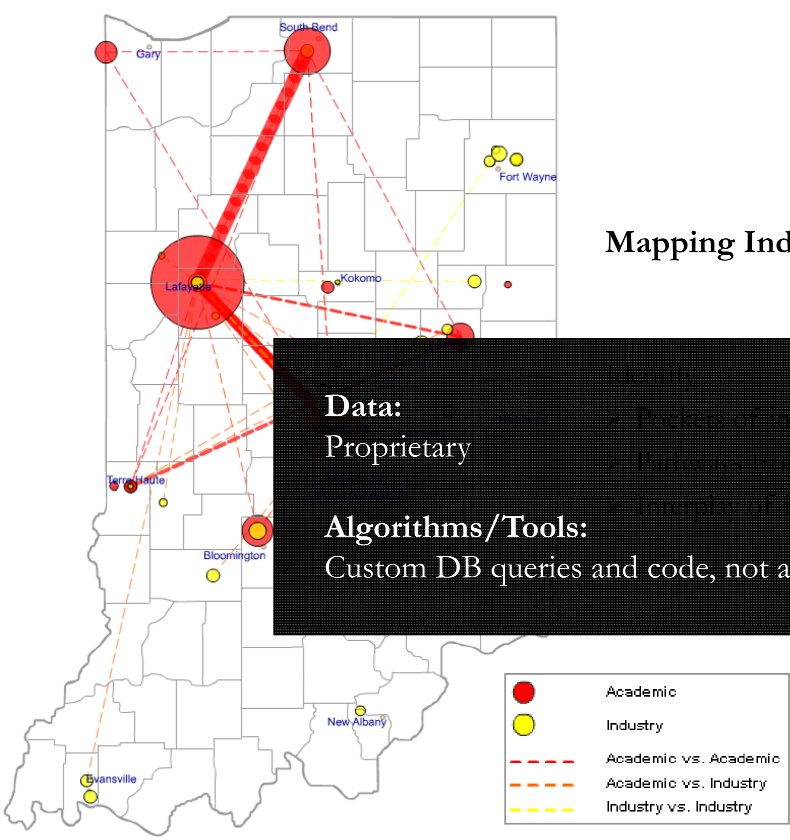
Data:

SCI/SSCI 2002:	proprietary
DOE:	FOIR
NIH:	http://projectreporter.nih.gov
NSF:	http://www.nsf.gov/awardsearch

SciMap to DOE/NIH/NSF linkage data not available.

Algorithms/Tools:

DrL available



Mapping Indiana's Intellectual Space

Identify

- Pockets of innovation
- Pathways from ideas to products
- Interplay of industry and academia

Data:

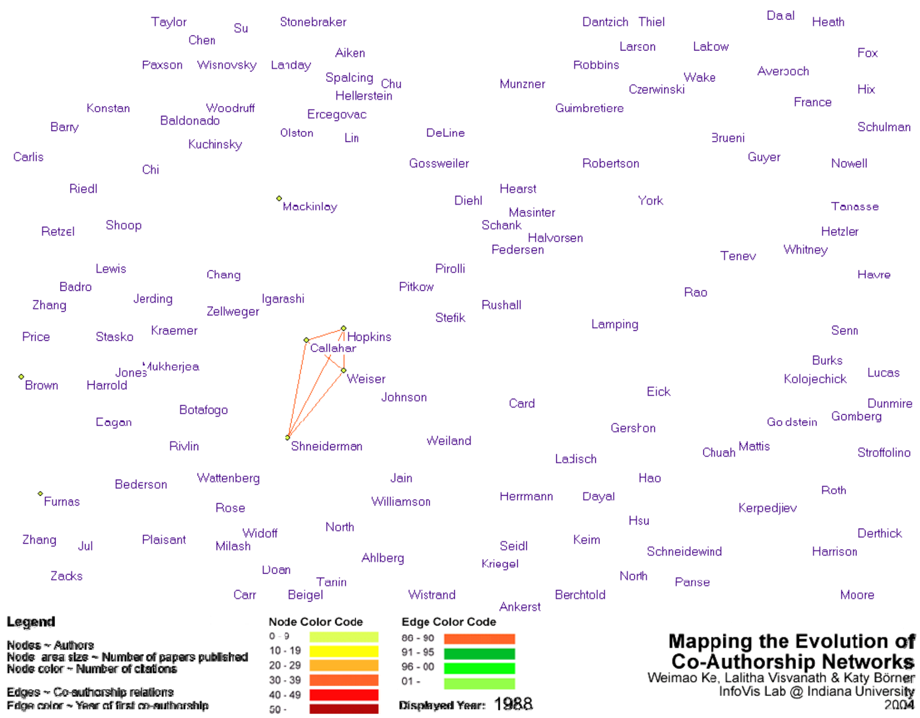
Proprietary

Algorithms/Tools:

Custom DB queries and code, not available.

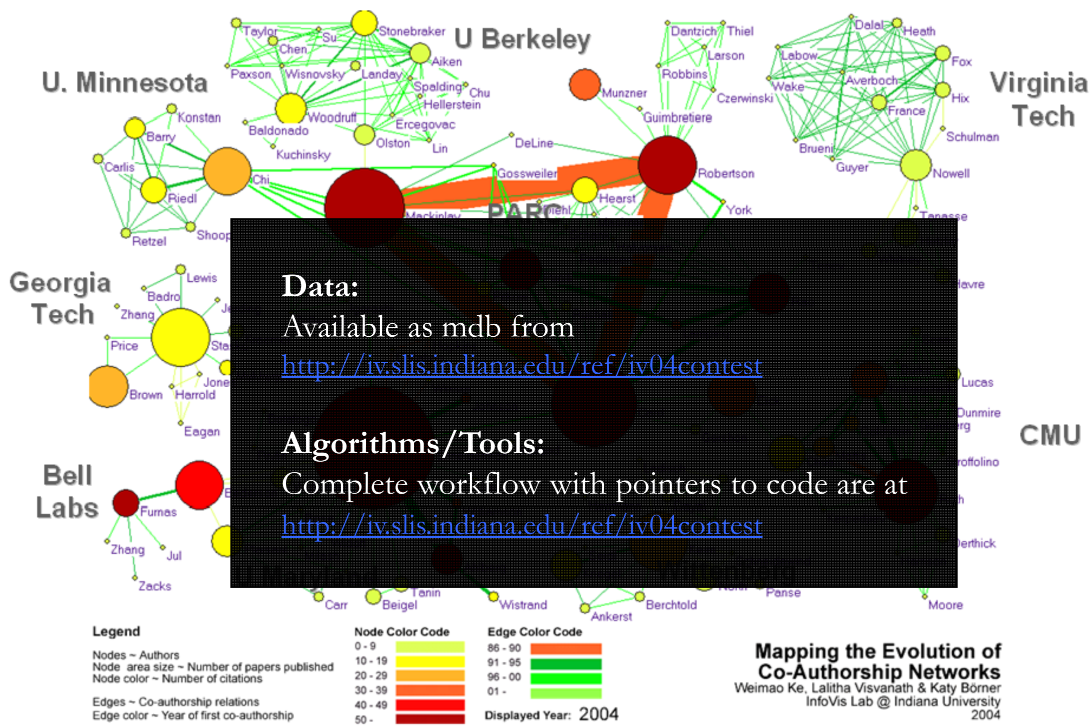
Mapping the Evolution of Co-Authorship Networks

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



Mapping the Evolution of Co-Authorship Networks

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



Data:
Available as mdb from
http://iv.slis.indiana.edu/ref/iv04contest

Algorithms/Tools:
Complete workflow with pointers to code are at
http://iv.slis.indiana.edu/ref/iv04contest

Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams

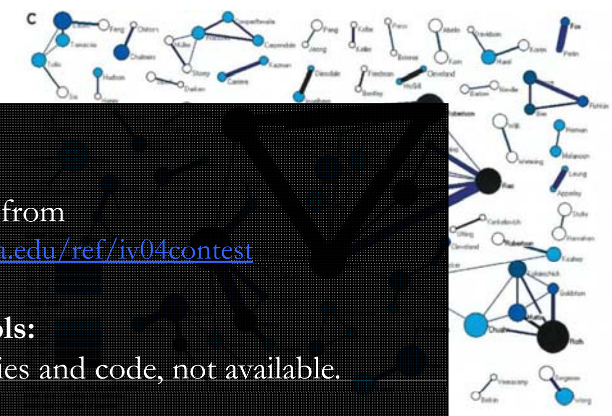
Börner, Dall'Asta, Ke & Vespignani (2005) Complexity, 10(4):58-67.

Research question:

- Is science driven by prolific single experts or by high-impact co-authorship teams?

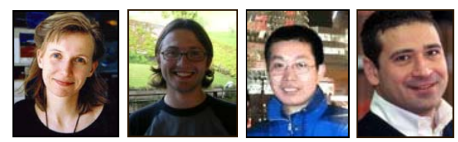
Contributions:

- New approach to allo... credit.
- Novel weighted gr...
- Visualization of th... co-author network...
- Centrality measure... impact.
- Global statistical analysis of paper production and citations in correlation with co-authorship team size over time.
- Local, author-centered entropy measure.



Data:
Available as mdb from
http://iv.slis.indiana.edu/ref/iv04contest

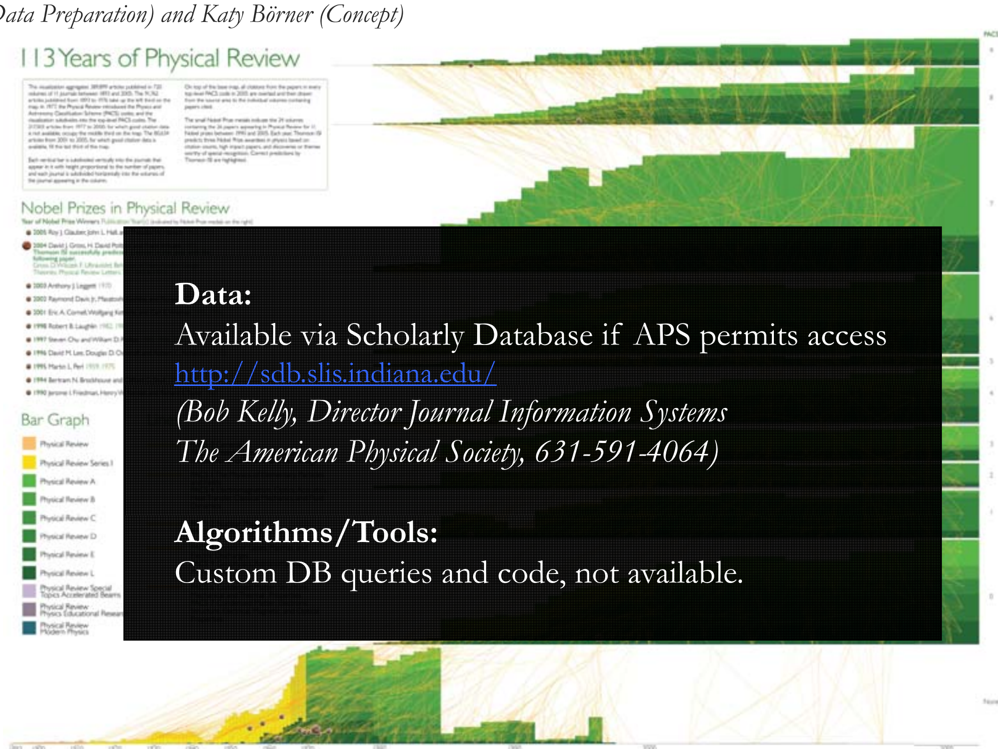
Algorithms/Tools:
Custom DB queries and code, not available.



113 Years of Physical Review

http://scimaps.org/dev/map_detail.php?map_id=171

Bruce W. Herr II and Russell Duhon (Data Mining & Visualization), Elisha F. Hardy (Graphic Design), Shashikant Penumarthy (Data Preparation) and Katy Börner (Concept)



Data:
Available via Scholarly Database if APS permits access
http://sdb.slis.indiana.edu/
(Bob Kelly, Director Journal Information Systems
The American Physical Society, 631-591-4064)

Algorithms/Tools:
Custom DB queries and code, not available.

Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions

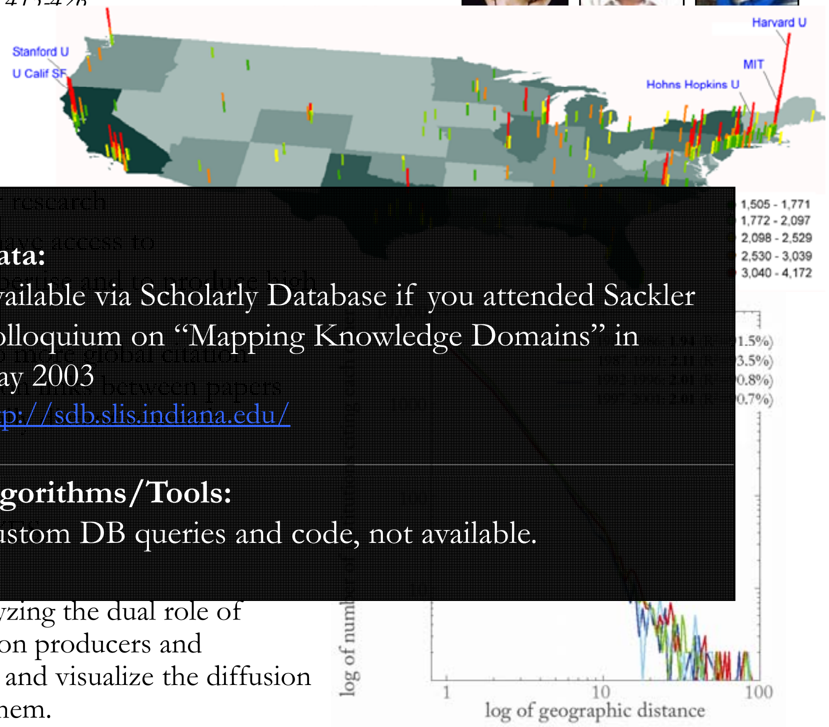
Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao. (2006) Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. Scientometrics. 68(3), pp. 415-426.

Research questions:

1. Does space still matter in the Internet age?
2. Does one still have to study and work at major research institutions in order to have access to high quality data and expertise and to produce high quality research?
3. Does the Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?

Contributions:

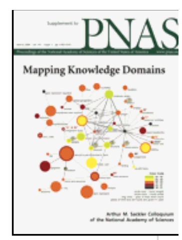
- Answer to Qs 1 + 2 is YES.
- Answer to Qs 3 is NO.
- Novel approach to analyzing the dual role of institutions as information producers and consumers and to study and visualize the diffusion of information among them.



Data:
Available via Scholarly Database if you attended Sackler Colloquium on "Mapping Knowledge Domains" in May 2003
http://sdb.slis.indiana.edu/

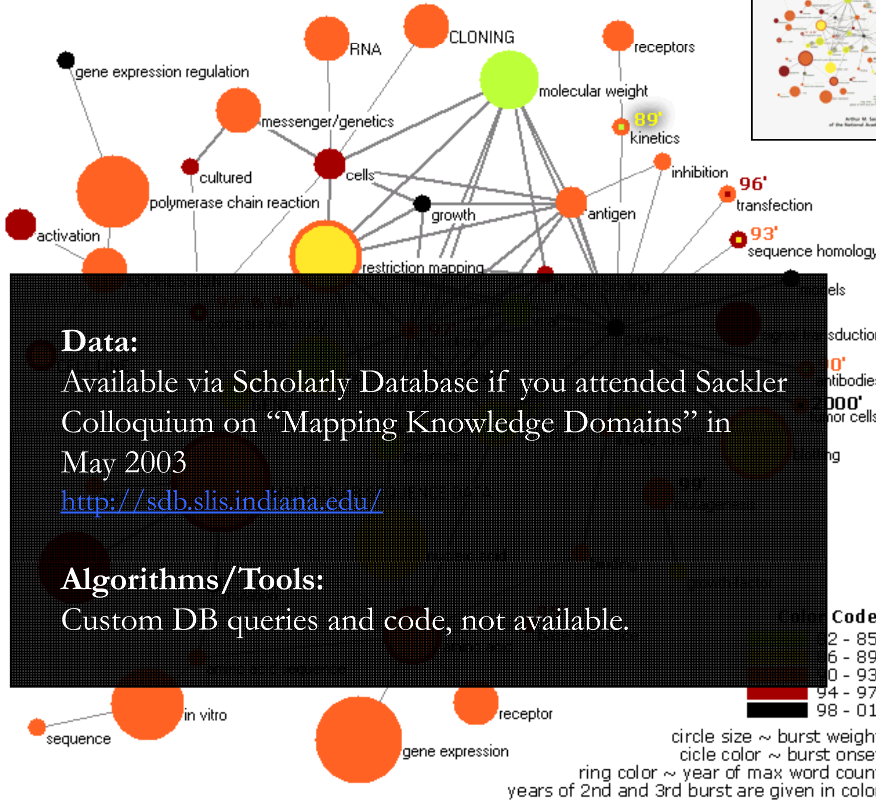
Algorithms/Tools:
Custom DB queries and code, not available.

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.



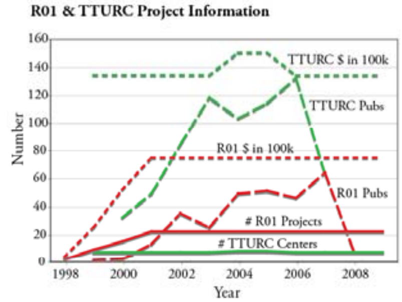
Data:
Available via Scholarly Database if you attended Sackler Colloquium on "Mapping Knowledge Domains" in May 2003
http://sdb.slis.indiana.edu/

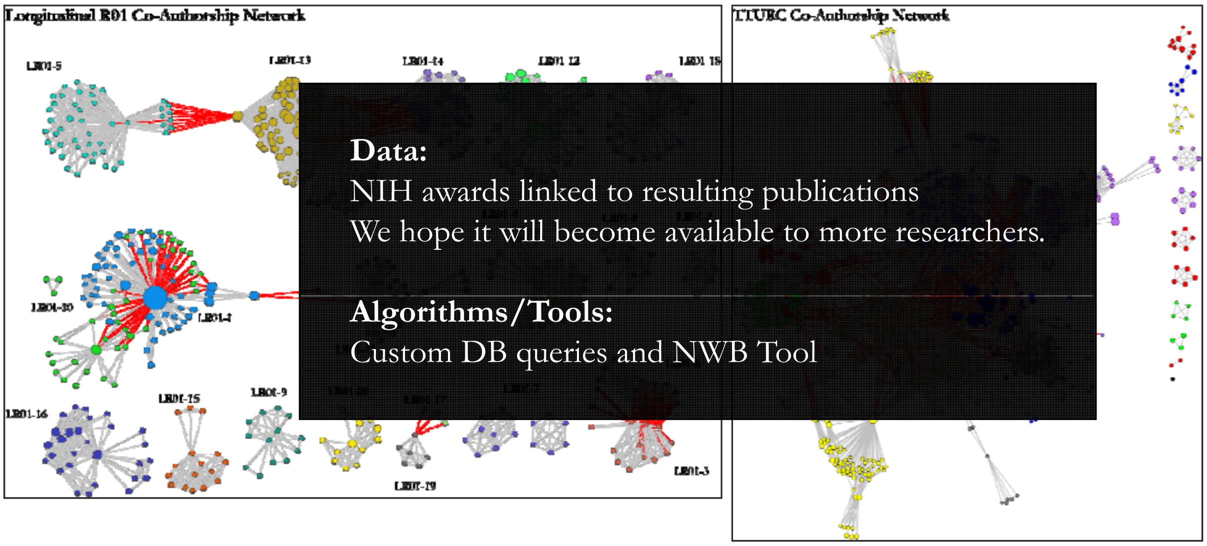
Algorithms/Tools:
Custom DB queries and code, not available.

Mapping Transdisciplinary Tobacco Use Research Centers Publications

Compare R01 investigator based funding with TTURC Center awards in terms of number of publications and evolving co-author networks.

Zoss & Börner, forthcoming.



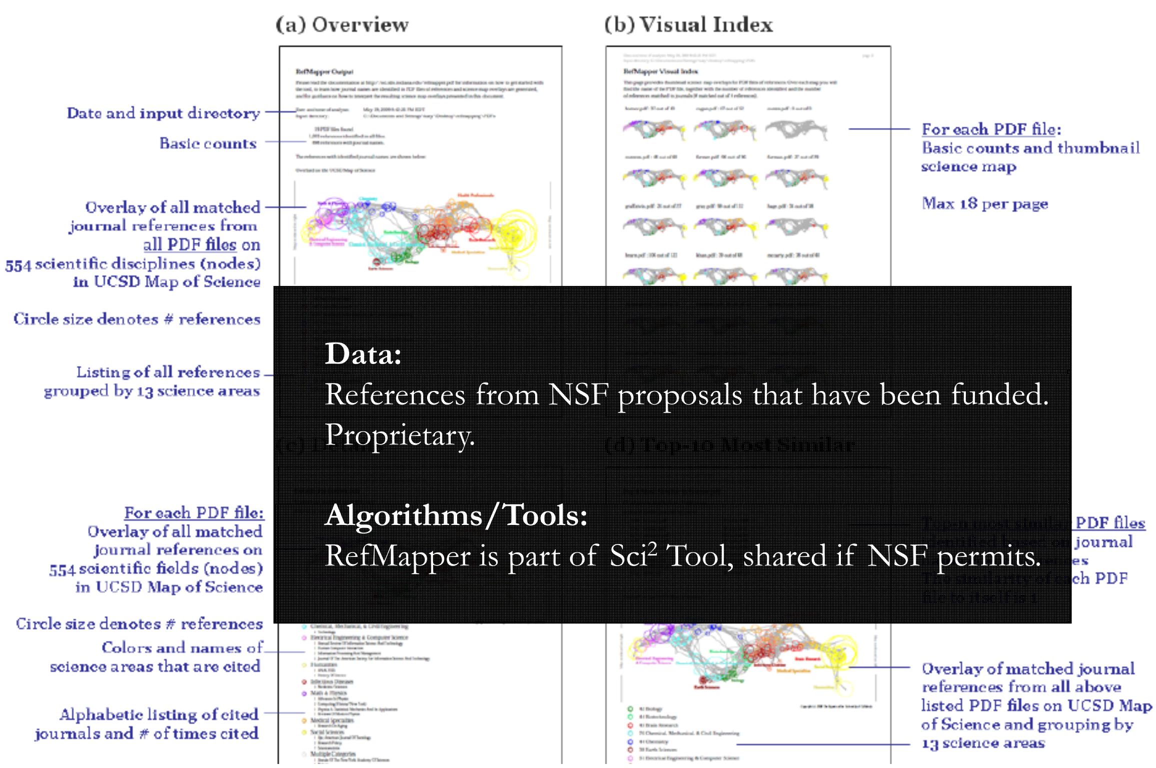


Data:
NIH awards linked to resulting publications
We hope it will become available to more researchers.

Algorithms/Tools:
Custom DB queries and NWB Tool

Reference Mapper

Duhon & Börner, forthcoming.



Data:
References from NSF proposals that have been funded.
Proprietary.

Algorithms/Tools:
RefMapper is part of Sci^2 Tool, shared if NSF permits.

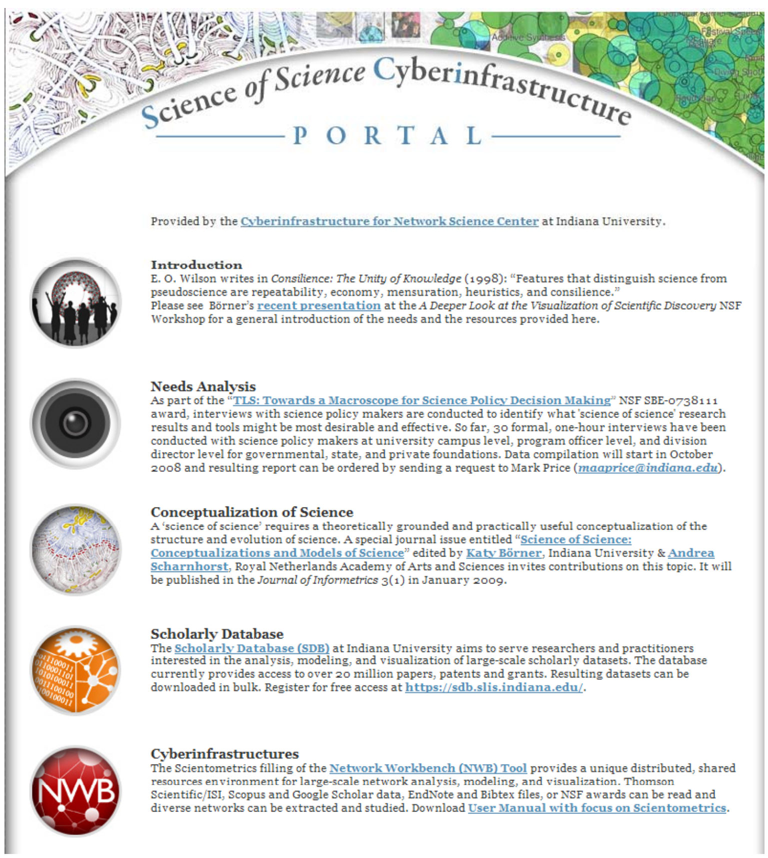
Overview

Science of Science Studies

Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu):

- **Scholarly Database (SDB)** (http://sdb.slis.indiana.edu) that provides free access to 23 million scholarly records
- **Sci² Tool** which reads SDB data and supports the identification of activity bursts, the extraction and display of co-author/inventor/investigator networks, and topical analysis, among others.

Mapping Science Exhibit



http://sci.slis.indiana.edu

Scholarly Database

http://sdb.slis.indiana.edu

Nianli Ma

"From Data Silos to Wind Chimes"

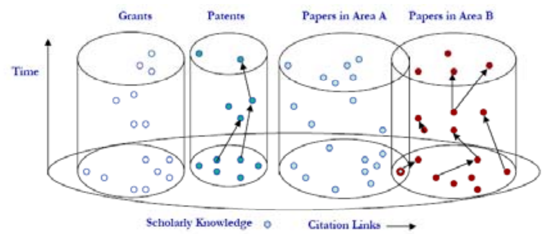
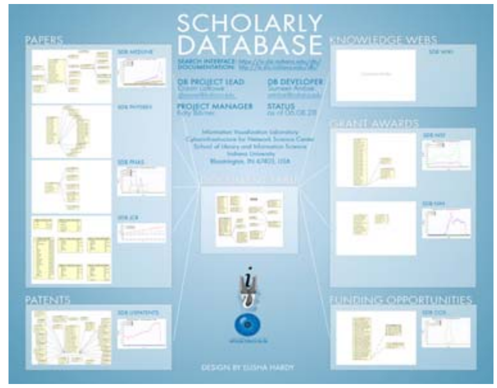


Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type are 'horizontally' interlinked.



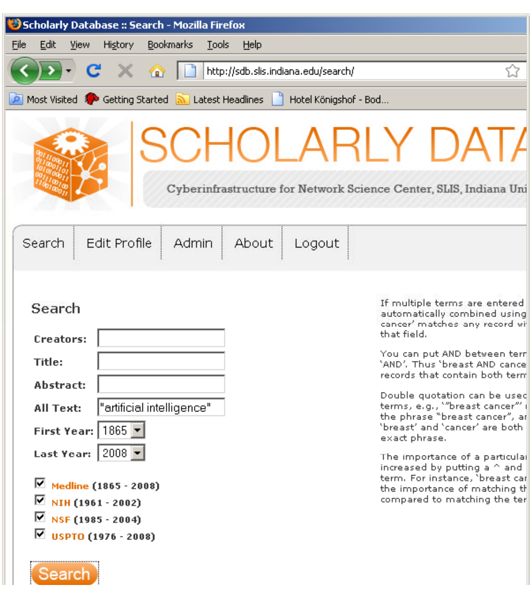
- Create public databases that any scholar can use. Share the burden of data cleaning and federation.
- Interlink creators, data, software/tools, publications, patents, funding, etc.

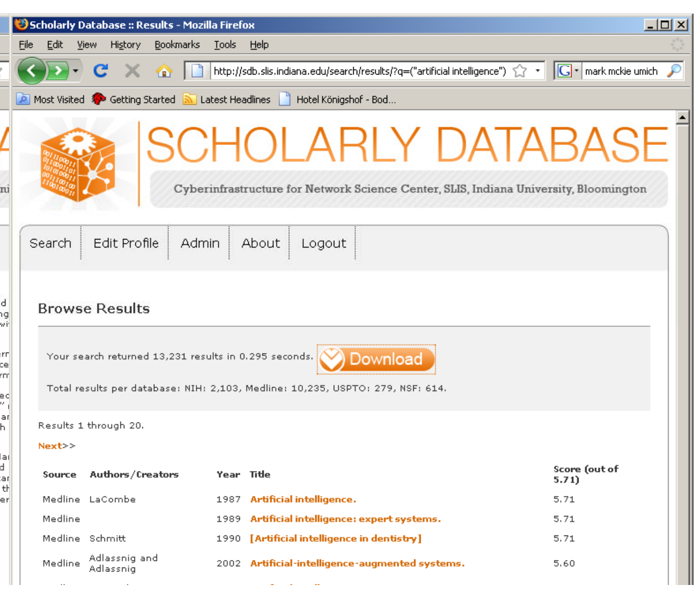
La Rowe, Gavin, Ambre, Sumeet, Burgoon, John, Ke, Weimao and Börner, Katy. (2007) The Scholarly Database and Its Utility for Scientometrics Research. In Proceedings of the 11th International Conference on Scientometrics and Informetrics, Madrid, Spain, June 25-27, 2007, pp. 457-462. http://ella.slis.indiana.edu/~katy/paper/07-issi-sdb.pdf

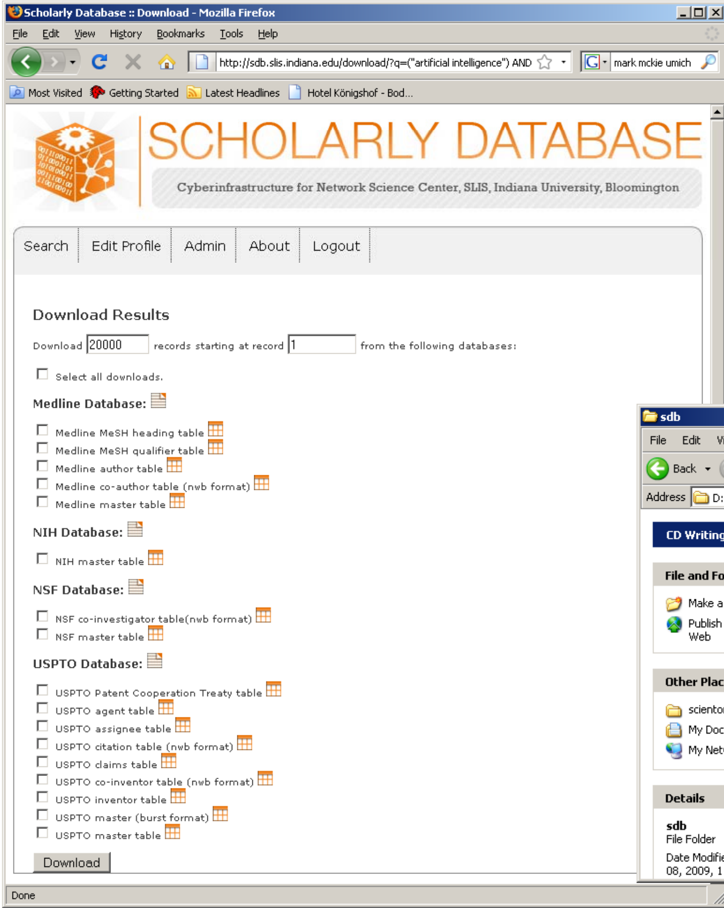
Scholarly Database: Web Interface

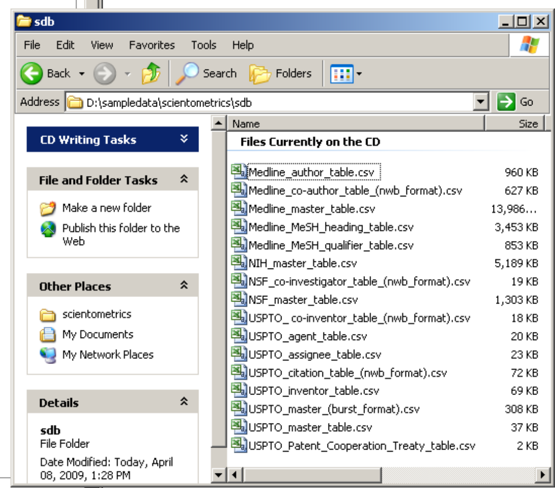
Anybody can register for free to search the about 23 million records and download results as data dumps.

Currently the system has over 130 registered users from academia, industry, and government from over 60 institutions and four continents.









Since March 2009:
Users can download networks:
- Co-author
- Co-investigator
- Co-inventor
- Patent citation

and tables for
burst analysis in NWB.

SDB Demo

http://sdb.slis.indiana.edu

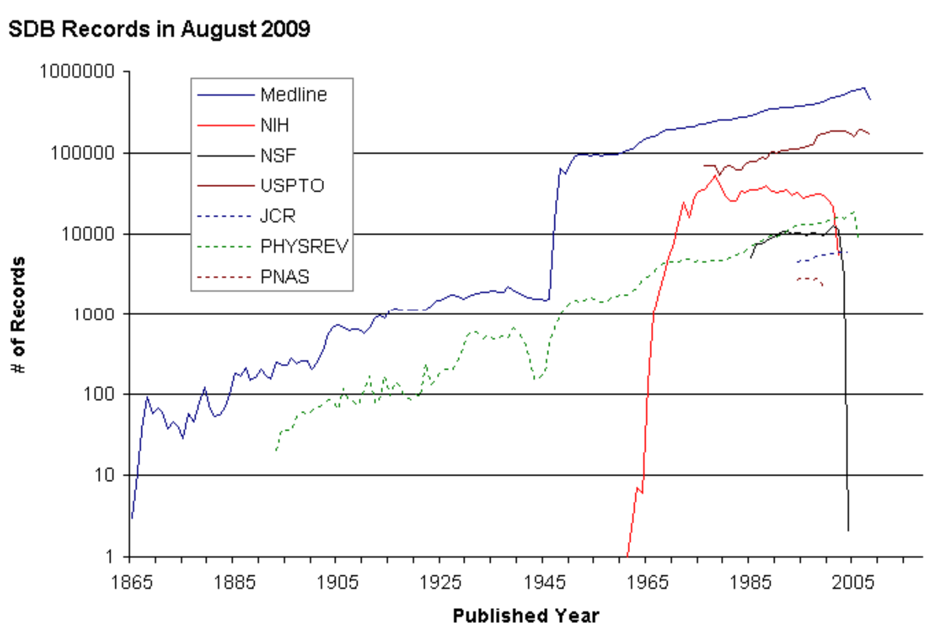
Scholarly Database: # Records, Years Covered

Datasets available via the Scholarly Database (* internally)

Dataset	# Records	Years Covered	Updated	Restricted Access
Medline	17,764,826	1898-2008	Yes	
PhysRev	398,005	1893-2006		Yes
PNAS	16,167	1997-2002		Yes
JCR	59,078	1974, 1979, 1984, 1989 1994-2004		Yes
USPTO	3, 875,694	1976-2008	Yes*	
NSF	174,835	1985-2002	Yes*	
NIH	1,043,804	1961-2002	Yes*	
Total	**23,167,642**	**1893-2006**	**4**	**3**

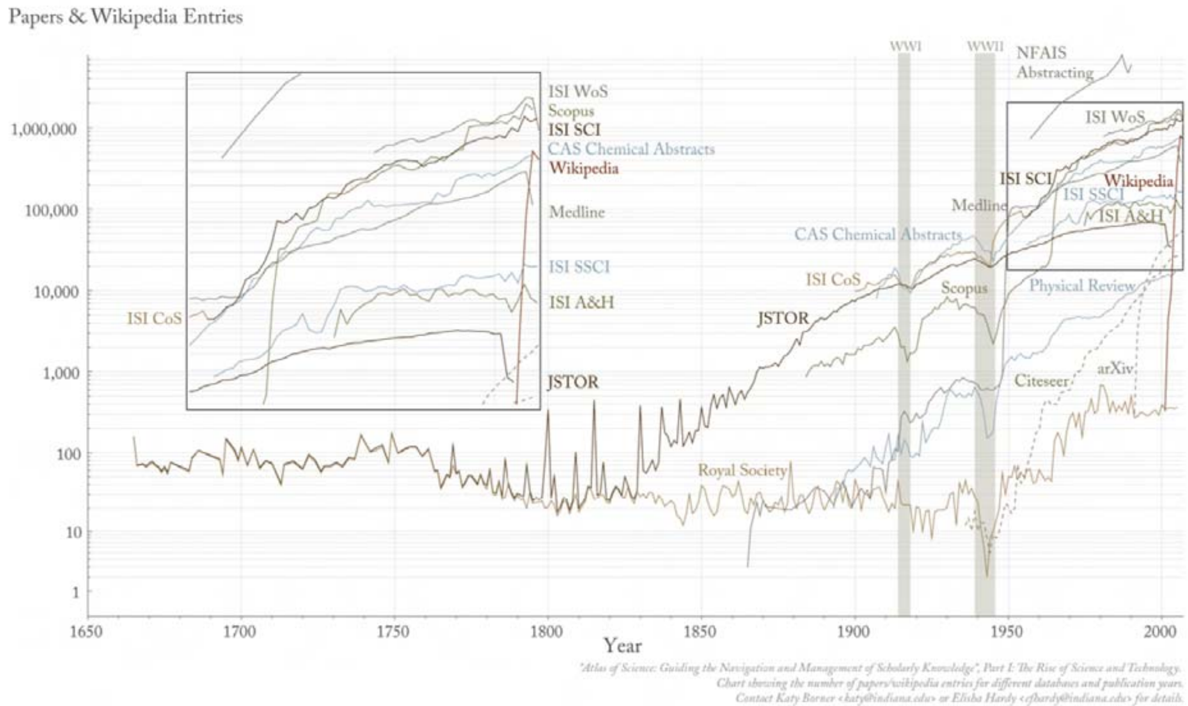
Aim for comprehensive time, geospatial, and topic coverage.

Temporal and Geospatial Coverage

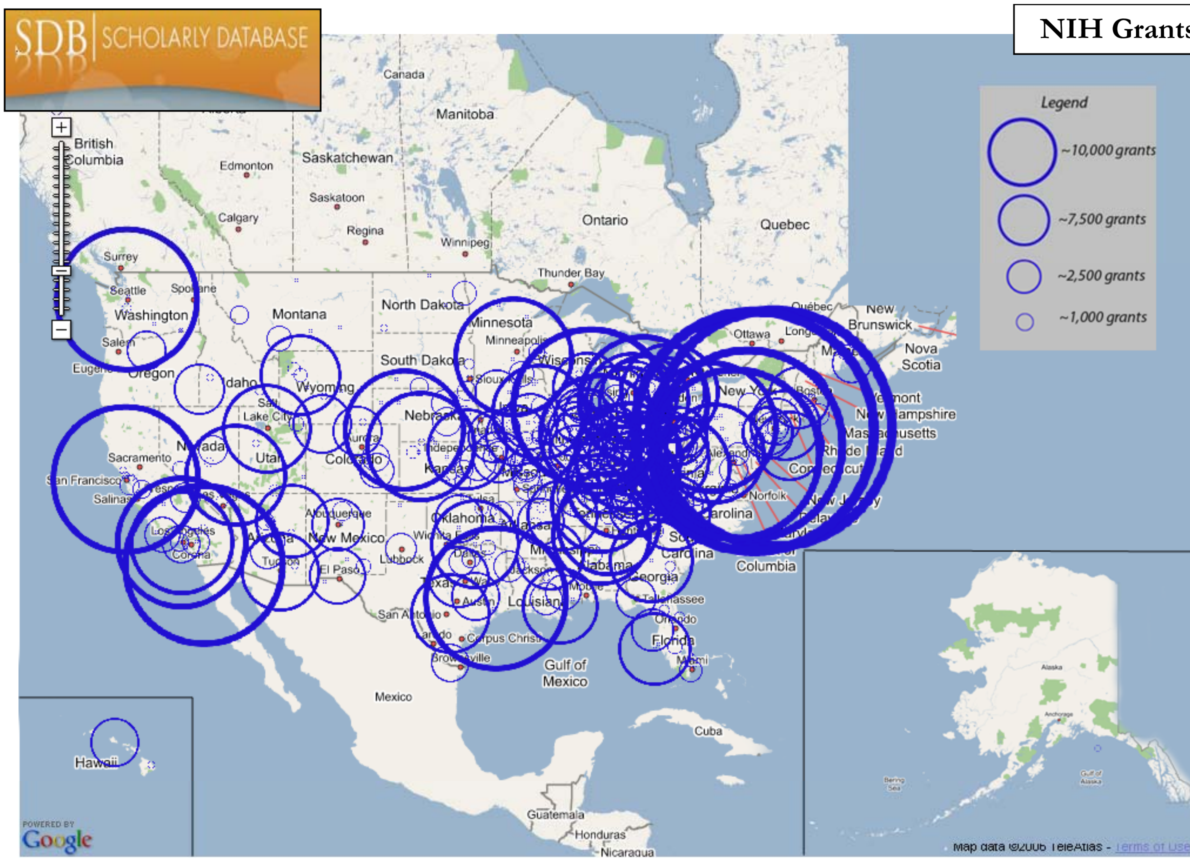


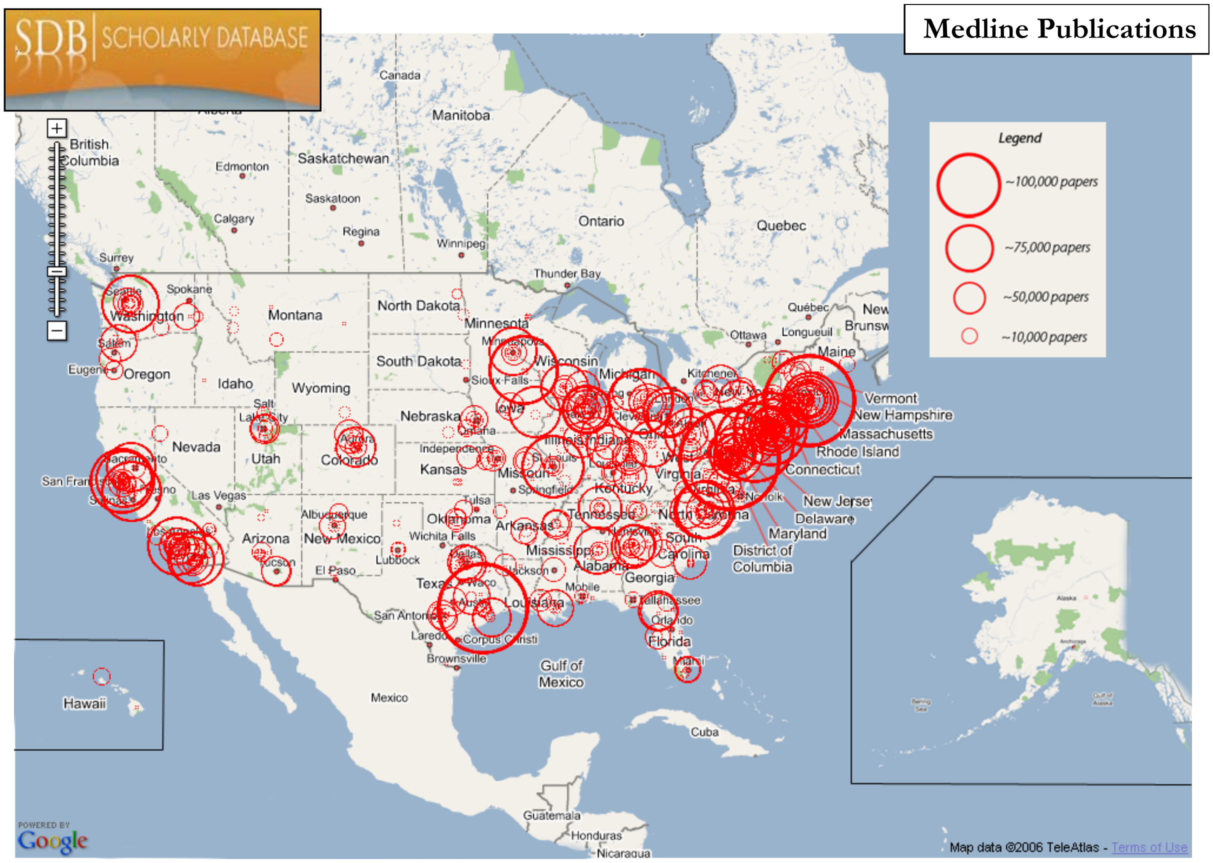
Comparison with Major Publication Data

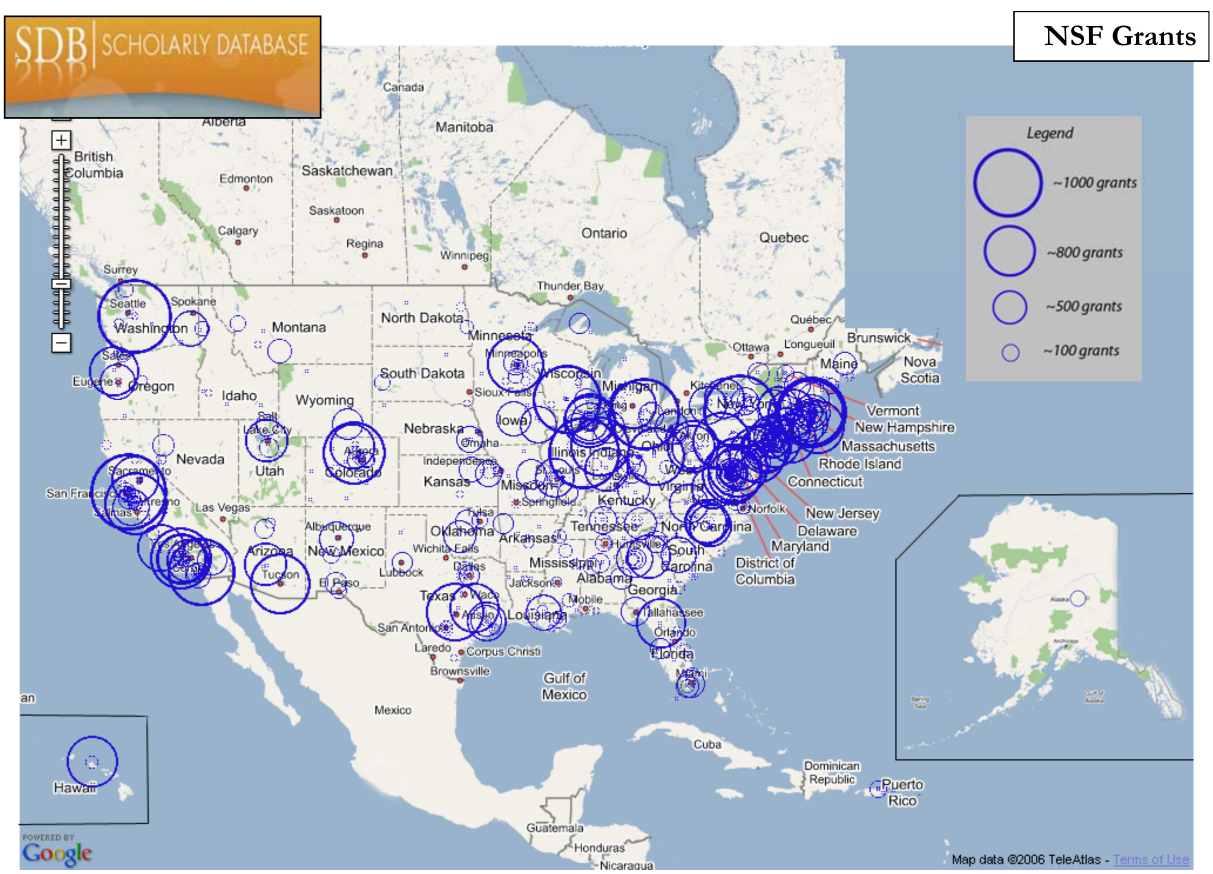
commonly used in scientometric studies



NIH Grants

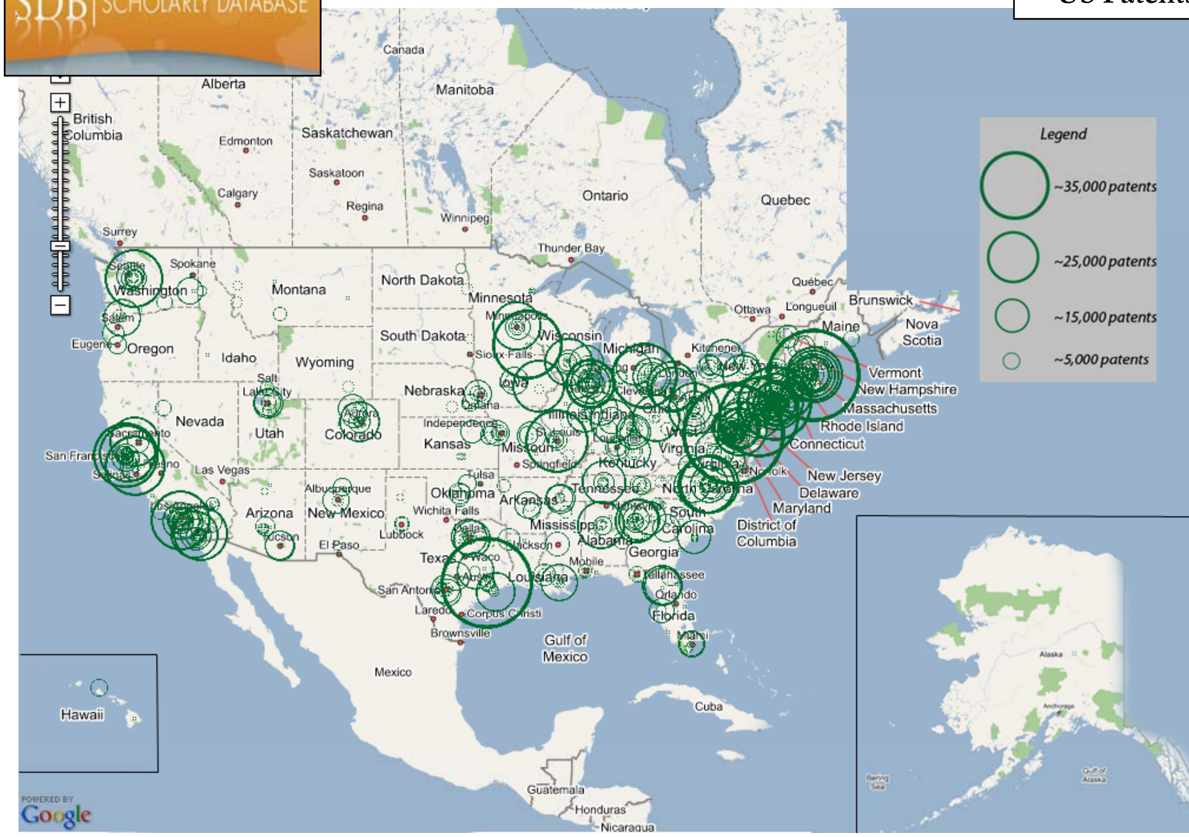



SDB | SCHOLARLY DATABASE
Medline Publications
Legend
~100,000 papers
~75,000 papers
~50,000 papers
~10,000 papers
Map data ©2006 TeleAtlas - Terms of Use


SDB | SCHOLARLY DATABASE
NSF Grants
Legend
~1000 grants
~800 grants
~500 grants
~100 grants
Map data ©2006 TeleAtlas - Terms of Use

SDB | SCHOLARLY DATABASE

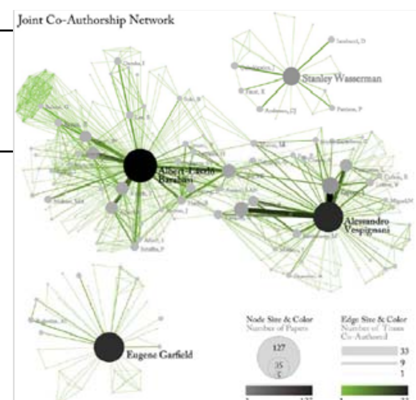
US Patents



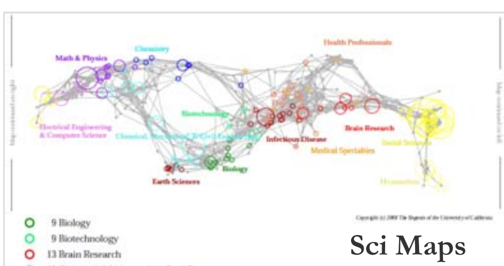
Sci² Tool

[*http://sci.slis.indiana.edu*](http://sci.slis.indiana.edu)

"Open Code for S&T Assessment"
Branded OSGi/CIShell based tool with NWB plugins and many new plugins.



GUESS Network Vis



Sci Maps

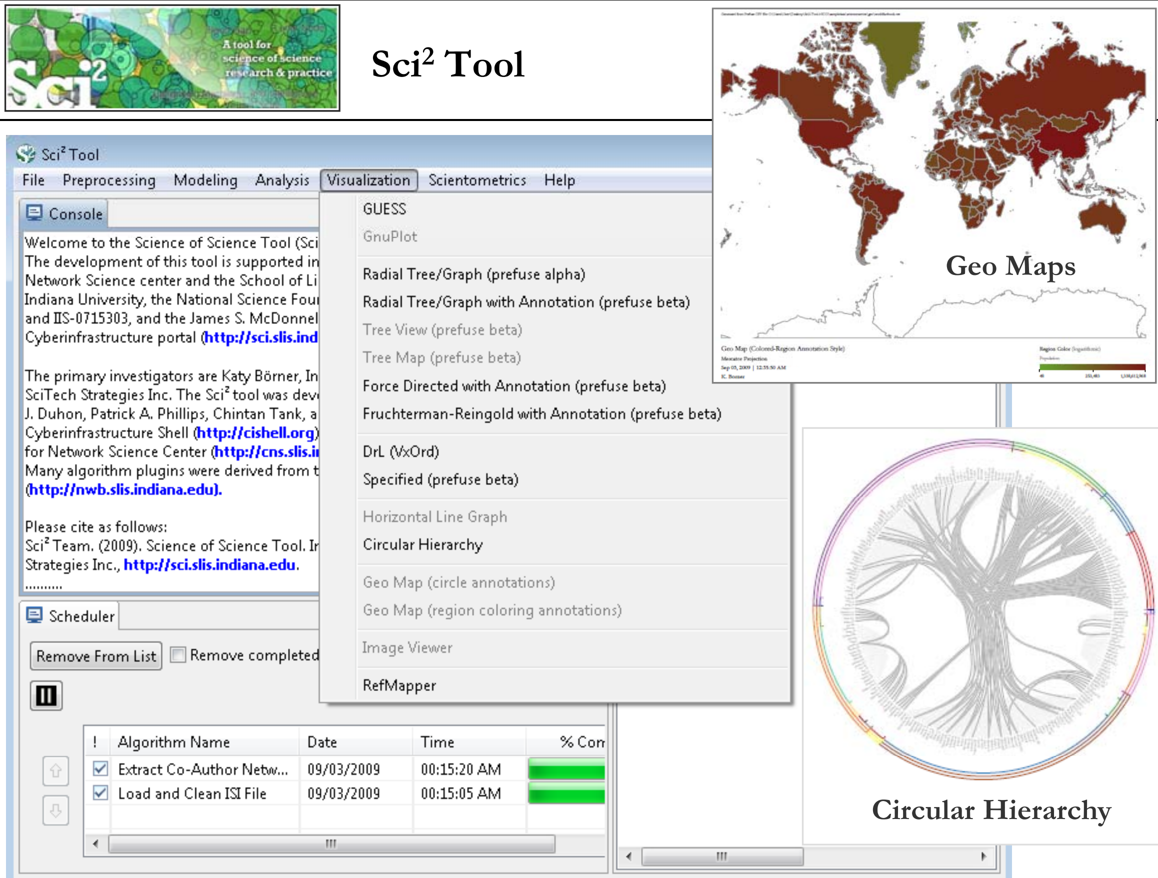
Geo Maps

Hierarchical Circular Visualization

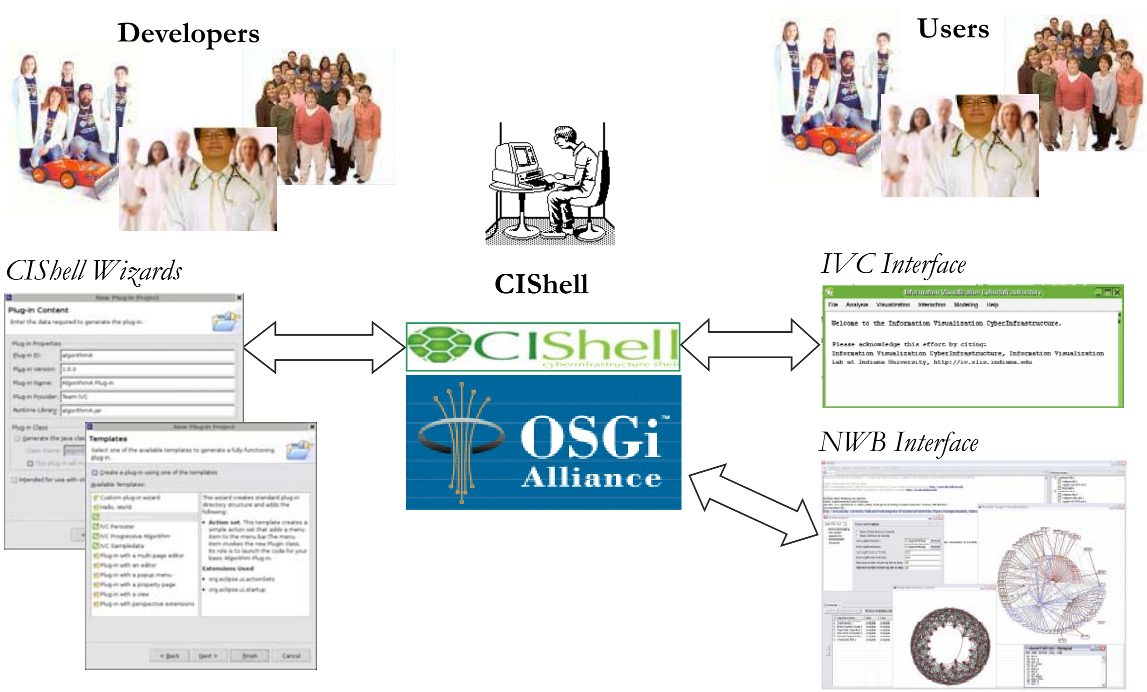
Horizontal Time Graphs

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.

Sci² Tool



Serving Non-CS Algorithm Developers & Users



Sci² Tool: Supported Data Formats

Personal Bibliographies
- Bibtex (.bib)
- Endnote Export Format (.enw)

Data Providers
- Web of Science by Thomson Scientific/Reuters (.isi)
- Scopus by Elsevier (.scopus)
- Google Scholar (access via *Publish or Perish* save as CSV, Bibtex, EndNote)
- Awards Search by National Science Foundation (.nsf)

Scholarly Database (all text files are saved as .csv)
- Medline publications by National Library of Medicine
- NIH funding awards by the National Institutes of Health (NIH)
- NSF funding awards by the National Science Foundation (NSF)
- U.S. patents by the United States Patent and Trademark Office (USPTO)
- Medline papers – NIH Funding

Network Formats
- NWB (.nwb)
- Pajek (.net)
- GraphML (.xml or .graphml)
- XGMML (.xml)

Burst Analysis Format
- Burst (.burst)

Other Formats
- CSV (.csv)
- Edgelist (.edge)
- Pajek (.mat)
- TreeML (.xml)

NWB=Sci² Tool: Algorithms (July 1st, 2008)

See https://nwb.slis.indiana.edu/community and handout

Preprocessing Edit

- **Remove Nodes**
  - Extract Top Nodes
  - Extract Nodes Above or Below Val
  - Delete High Degree Nodes
  - Delete Random Nodes
  - Delete Isolates
- **Remove Edges**
  - Extract Top Edges
  - Extract Edges Above or Below Valı
  - Remove Self Loops
  - Trim By Degree?
  - Pathfinder Network Scaling
- **Sampling**
  - Snowball Sampling (n nodes)
  - Node Sampling
  - Edge Sampling
- **Transformations**
  - Symmetrize
  - Dichotomize
  - Multipartite Joining

Modeling Edit

- **General**
  - Random Graph
  - Watts-Strogatz Small World
  - Barabási-Albert Scale-Free
- **Structured**
  - CAN
  - Chord
- **Unstructured**
  - Hypergrid
  - PRU
- **Other**
  - TARL
  - Discrete Network Dynamics

Analysis Edit

- **General Purpose**
  - Network Analysis Toolkit?
- **Unweighted & Undirected**
  - Based on degree/
    - Node Degree
    - Node Distribution
  - Based on clustering
    - k-Nearest Neighbor
    - Watts Strogatz Clustering Coefficient
    - Watts Strogatz Clustering Coefficient Over k
  - Based on path
    - Diameter
    - Average Shortest Path
    - Shortest Path Distribution
    - Node Betweenness Centrality
  - Based on components
    - Connected Components
    - Weak Component Clustering
  - K-Core
    - Extract K-Core?
    - Annotate K-Coreness?
- **Unweighted & Directed**
  - Based on degree
    - Node Indegree
    - Node Outdegree
    - Indegree Distribution
    - Outdegree Distribution
  - Based on local graph structure
    - k-Nearest Neighbor
    - Single Node In-Out Degree Correlations?
  - Unnamed Category?
    - Page Rank
  - Based on local graph structure #2
    - Dyad Reciprocity?
    - Arc Reciprocity?

tion Edit

- **Tools**
  - GUESS
  - GnuPlot?
- **Predefined Positions Layout**
  - DrL (VxOrd)
  - Pre-defined Positions (prefuse beta)?
- **Move**
  - Circular
- **Tree Layouts**
  - Radial Tree (prefuse alpha)
  - Radial Tree with Annotations (prefuse beta)?
  - Tree Map
  - Tree View
  - Balloon Graph (prefuse alpha)?
- **Network Layouts**
  - Force Directed with Annotation (prefuse beta)
  - Kamada-Kawai (JUNG)
  - Fruchterman-Reingold (JUNG)
  - Fruchterman-Reingold with Annotation (prefuse beta)
  - Spring (JUNG)
  - Small World (prefuse alpha)
- **Other Layouts**
  - Parallel Coordinates (demo)?
  - LaNet (k-Core Decomposition)

etrics Edit

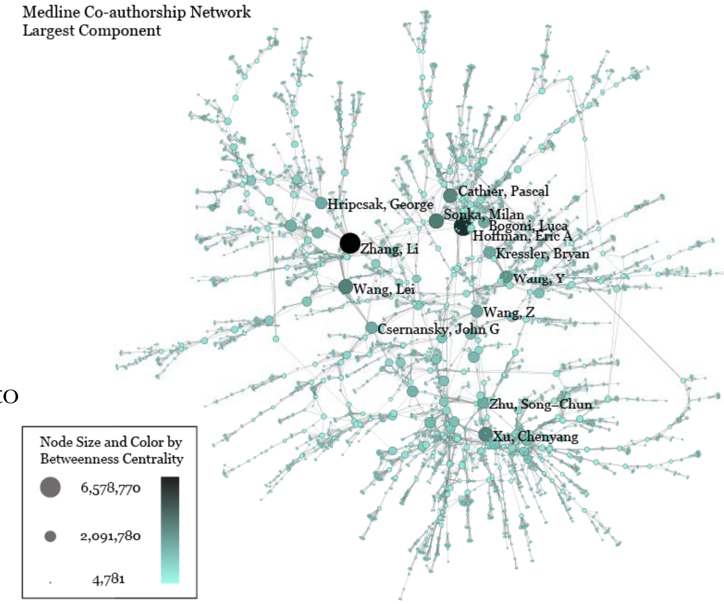
- **Extract Network From Table**
  - Extract Co-Authorship Network
  - Extract Co-Occurrence Network From Table?
  - Extract Directed Network From Table?
- **Extract Network From Another Network**
  - Extract Bibliographic Coupling Similarity Network
  - Extract Co-Citation Similarity Network?
- **Cleaning**
  - Remove ISI Duplicate Records

NWB=Sci^2 Tool: Output Formats

- NWB tool can be used for data conversion. Supported output formats comprise:
- CSV (.csv)
- NWB (.nwb)
- Pajek (.net)
- Pajek (.mat)
- GraphML (.xml or .graphml)
- XGMML (.xml)

- GUESS
  Supports export of images into common image file formats.

- Horizontal Bar Graphs
- saves out raster and ps files.



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.

B. Loading NSF [redacted] chers at Indiana U

Institution Level

C. Indiana U, C [redacted] Co-PI networks.

Scientific Field L [redacted]

D. Extracting co [redacted] ng bursts in SD [redacted]

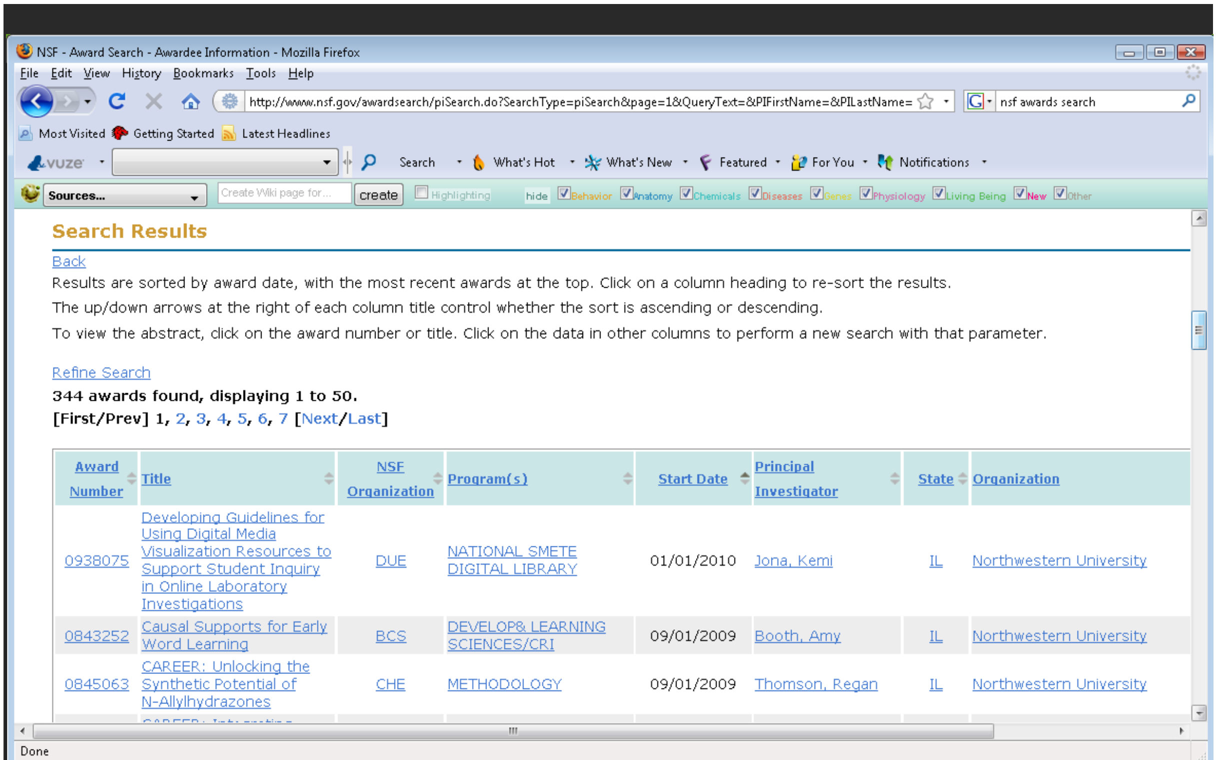
Will be presented in hands-on Workshop on Thursday Sept 3, 2009, 1-5pm

Together with guidance on how to design workflows using 100+ algorithms and how to dissect and design effective visualizations.

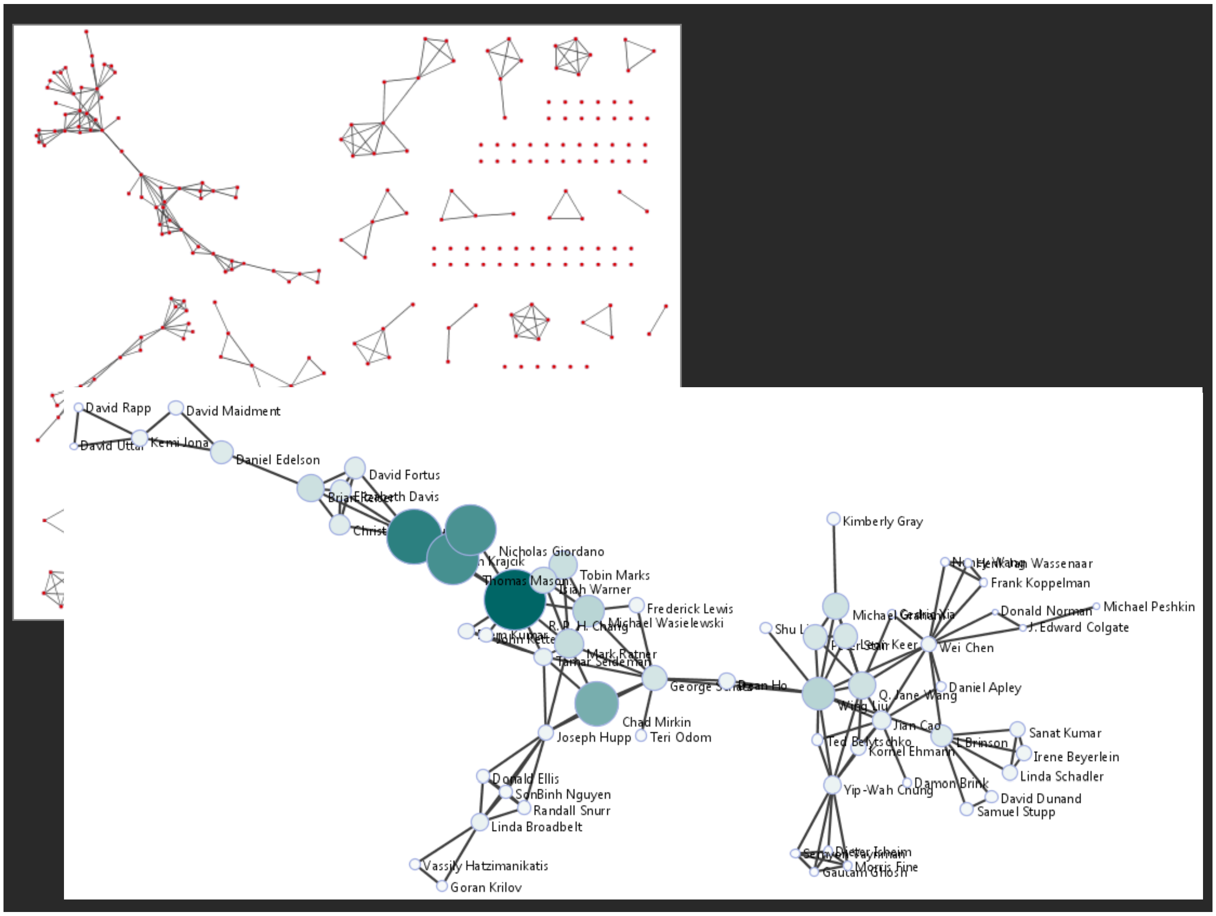
Bonus: Create your custom tool.

Sci² Tool Demo

http://sci.slis.indiana.edu



http://www.nsf.gov/awardsearch



Outlook

CIShell/OSGi is at the core of different CIs and a total of 169 unique plugins are used in the

- ***Information Visualization*** (http://iv.slis.indiana.edu),
- ***Network Science (NWB Tool)*** (http://nwb.slis.indiana.edu),
- ***Scientometrics and Science Policy (Sci² Tool)*** (http://sci.slis.indiana.edu), and
- ***Epidemics*** (http://epic.slis.indiana.edu) research communities.

Most interestingly, a number of other projects recently adopted OSGi and one adopted CIShell:

Cytoscape (http://www.cytoscape.org) lead by Trey Ideker, UCSD is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data (Shannon et al., 2002).

Taverna Workbench (http://taverna.sourceforge.net) lead by Carol Goble, University of Manchester, UK is a free software tool for designing and executing workflows (Hull et al., 2006). Taverna allows users to integrate many different software tools, including over 30,000 web services.

MAEviz (https://wiki.ncsa.uiuc.edu/display/MAE/Home) managed by Shawn Hampton, NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research.

TEXTrend (http://www.textrend.org) lead by George Kampis, Eötvös University, Hungary develops a framework for the easy and flexible integration, configuration, and extension of plugin-based components in support of natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component.

As the functionality of OSGi-based software frameworks improves and the number and diversity of dataset and algorithm plugins increases, the capabilities of custom tools or macroscopes will expand.

Overview

Science of Science Studies

Science of Science Cyberinfrastructure (http://sci.slis.indiana.edu):

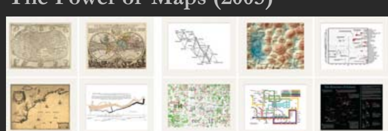
- **Scholarly Database (SDB)** (http://sdb.slis.indiana.edu) that provides free access to 23 million scholarly records
- **Sci² Tool** which reads SDB data and supports the identification of activity bursts, the extraction and display of co-author/inventor/investigator networks, and topical analysis, among others.

Mapping Science Exhibit

Mapping Science Exhibit – 10 Iterations in 10 years

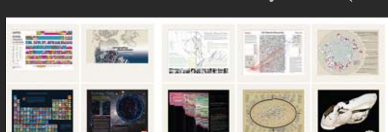
http://scimaps.org

The Power of Maps (2005)



Science Maps for Economic Decision Makers (2008)

The Power of Reference Systems (2006)



Science Maps for Science Policy Makers (2009)
Science Maps for Scholars (2010)
Science Maps as Visual Interfaces to Digital Libraries (2011)
Science Maps for Kids (2012)
Science Forecasts (2013)

The Power of Forecasts (2007)

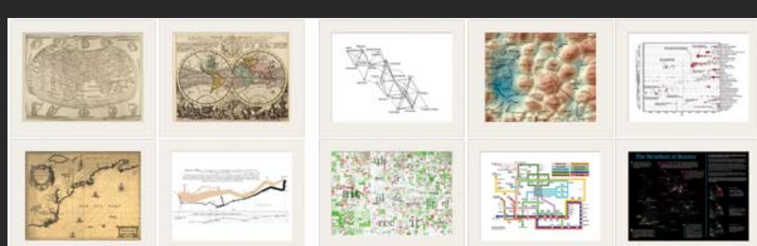
How to Lie with Science Maps (2014)

Exhibit has been shown in 72 venues on four continents. Currently at
- NSF, 10th Floor, 4201 Wilson Boulevard, Arlington, VA
- Wallenberg Hall, Stanford University, CA
- Center of Advanced European Studies and Research, Bonn, Germany
- Science Train, Germany.

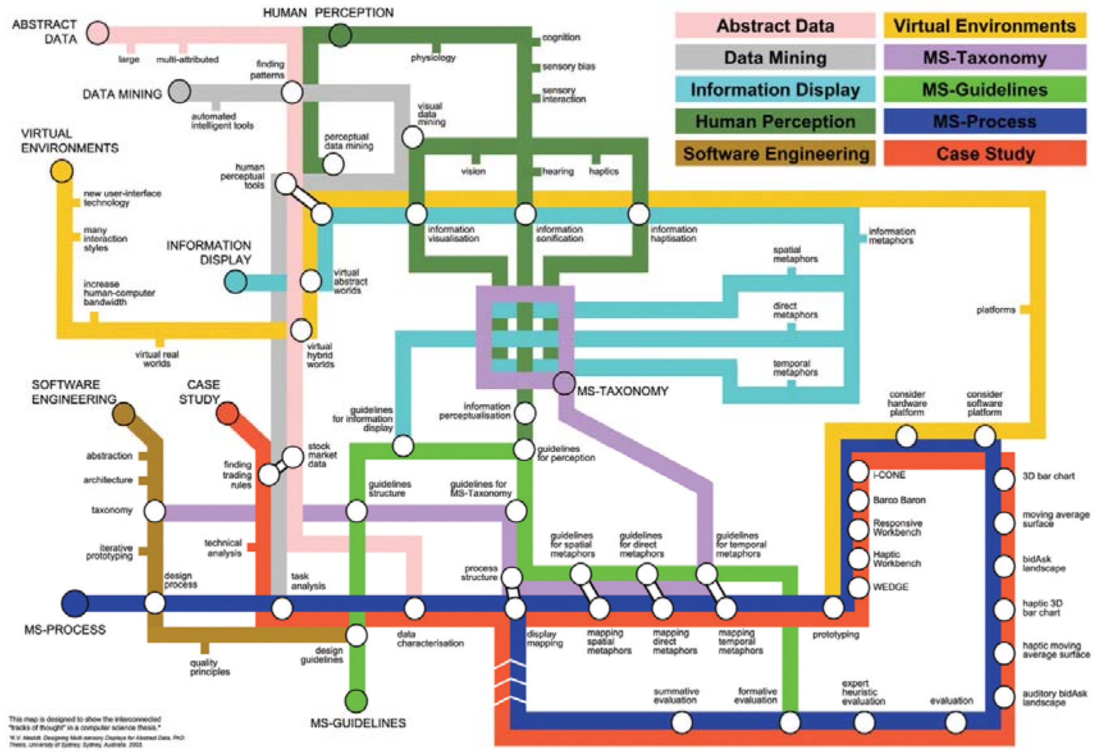
Debut of 5th Iteration of Mapping Science Exhibit at MEDIA X was on May 18, 2009 at Wallenberg Hall, Stanford University, *http://mediax.stanford.edu*, *http://scaleindependentthought.typepad.com/photos/scimaps*

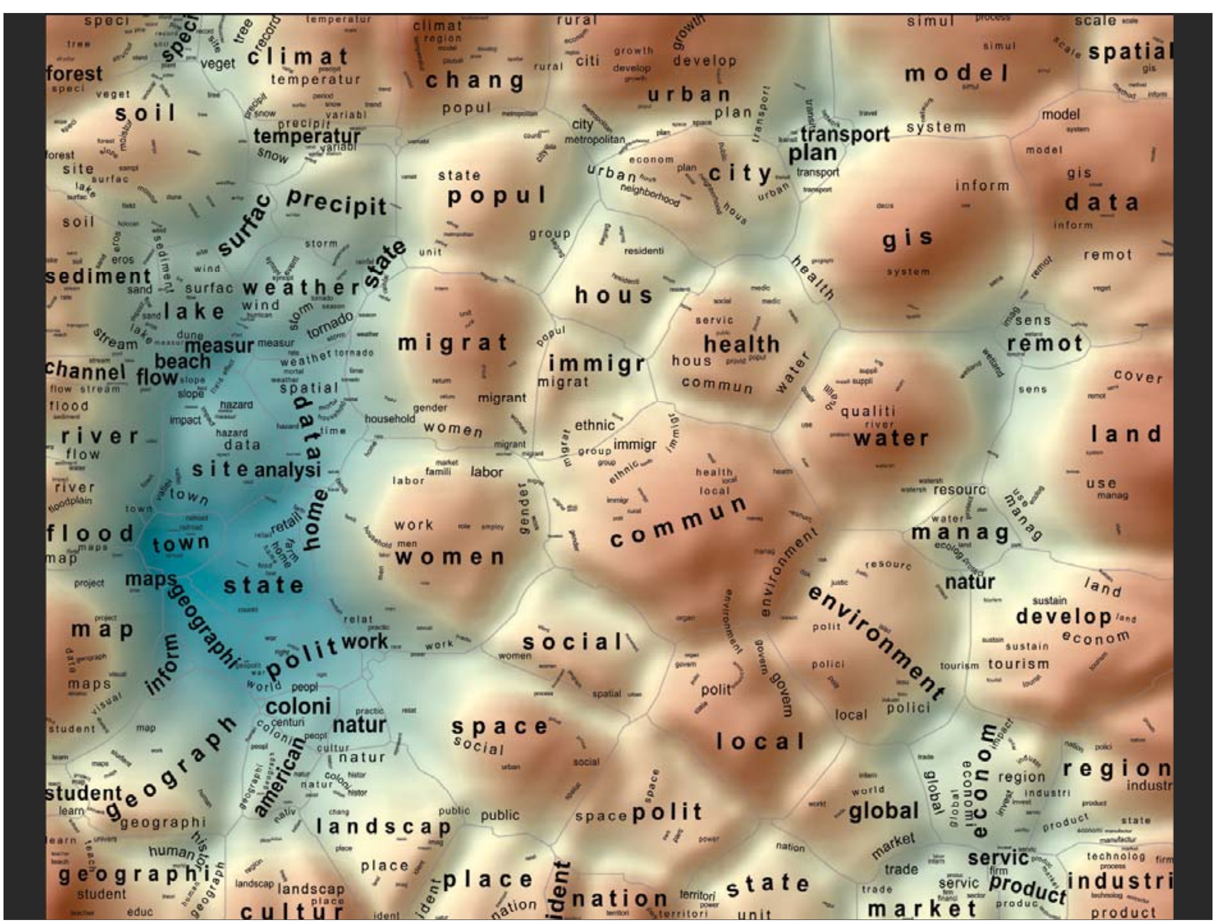
The Power of Maps

Four Early Maps of Our World
VERSUS
Six Early Maps of Science



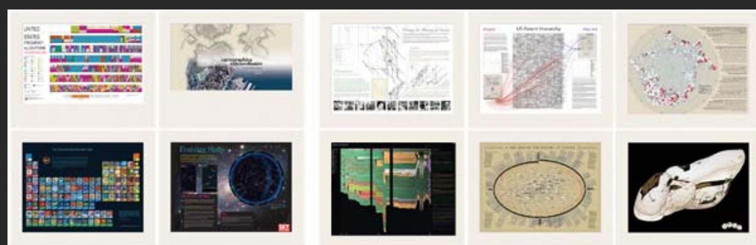
(1st Iteration of Places & Spaces Exhibit - 2005)


ABSTRACT DATA
HUMAN PERCEPTION
DATA MINING
VIRTUAL ENVIRONMENTS
INFORMATION DISPLAY
SOFTWARE ENGINEERING
CASE STUDY
MS-TAXONOMY
MS-PROCESS
MS-GUIDELINES
Abstract Data
Virtual Environments
Data Mining
MS-Taxonomy
Information Display
MS-Guidelines
Human Perception
MS-Process
Software Engineering
Case Study
large
multi-attributed
finding patterns
physiology
cognition
sensory bias
sensory interaction
automated intelligent tools
visual data mining
perceptual data mining
human perceptual tools
vision
hearing
haptics
new user-interface technology
many interaction styles
increase human-computer bandwidth
information visualisation
information sonification
information haptisation
information metaphors
spatial metaphors
direct metaphors
temporal metaphors
platforms
virtual abstract worlds
virtual real worlds
virtual hybrid worlds
information perceptualisation
guidelines for information display
guidelines for perception
consider hardware platform
consider software platform
abstraction
architecture
taxonomy
iterative prototyping
stock market data
finding trading rules
technical analysis
guidelines structure
guidelines for MS-Taxonomy
I-CONE
Barco Baron
Responsive Workbench
Haptic Workbench
WEDGE
3D bar chart
moving average surface
bidAsk landscape
haptic 3D bar chart
haptic moving average surface
auditory bidAsk landscape
guidelines for spatial metaphors
guidelines for direct metaphors
guidelines for temporal metaphors
process structure
design process
task analysis
data characterisation
display mapping
mapping spatial metaphors
mapping direct metaphors
mapping temporal metaphors
prototyping
quality principles
design guidelines
summative evaluation
formative evaluation
expert heuristic evaluation
evaluation
This map is designed to show the interconnected "tracks of thought" in a computer science thesis.*

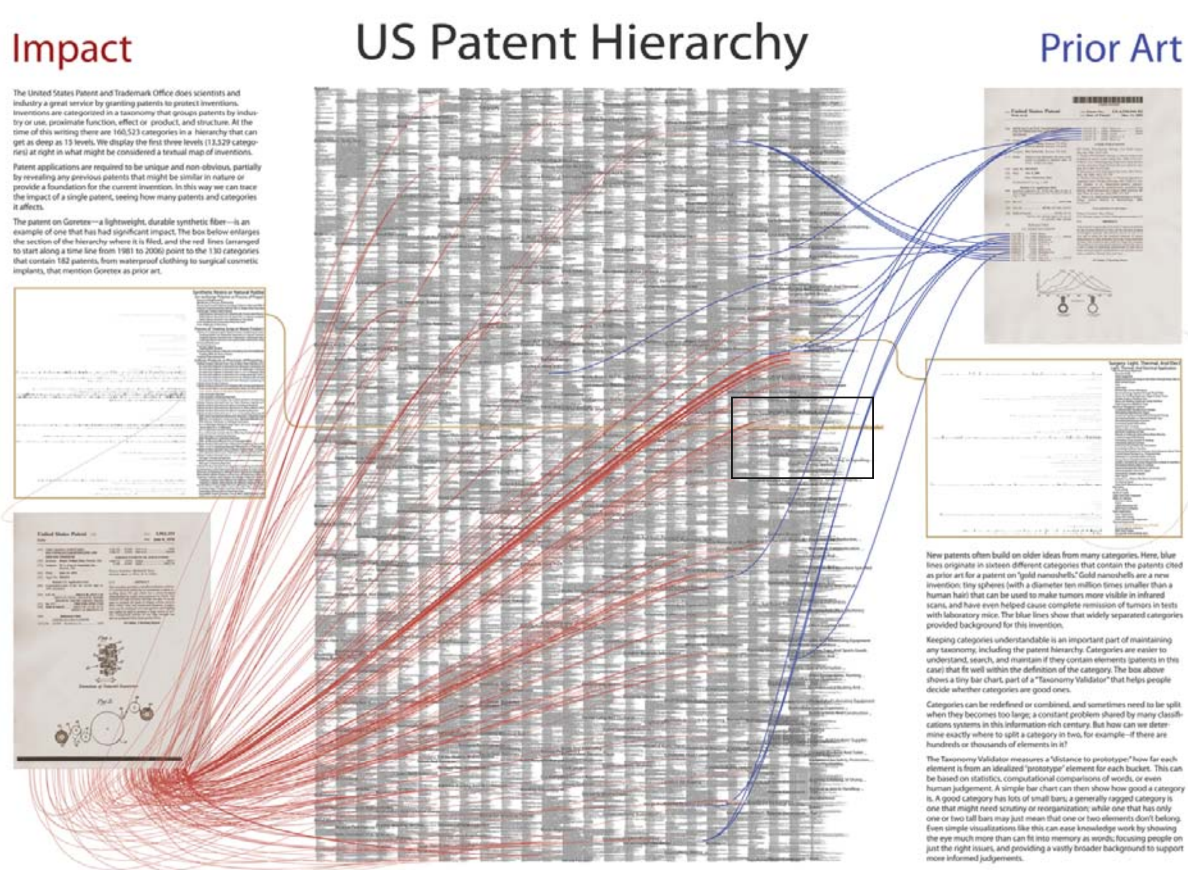

climat
chang
urban
model
spatial
forest
soil
temperatur
precipit
popul
city
transport
plan
data
gis
surfac
sediment
weather
state
hous
lake
measur
beach
channel
flow
migrat
immigr
health
remot
river
data
site analysi
water
land
flood
town
commun
manag
state
women
environment
natur
develop
map
polit
work
social
maps
inform
geographi
coloni
natur
tourism
geograph
student
american
space
local
econom
region
landscap
polit
global
geographi
place
ident
nation
state
servic
product
industri
cultur
market

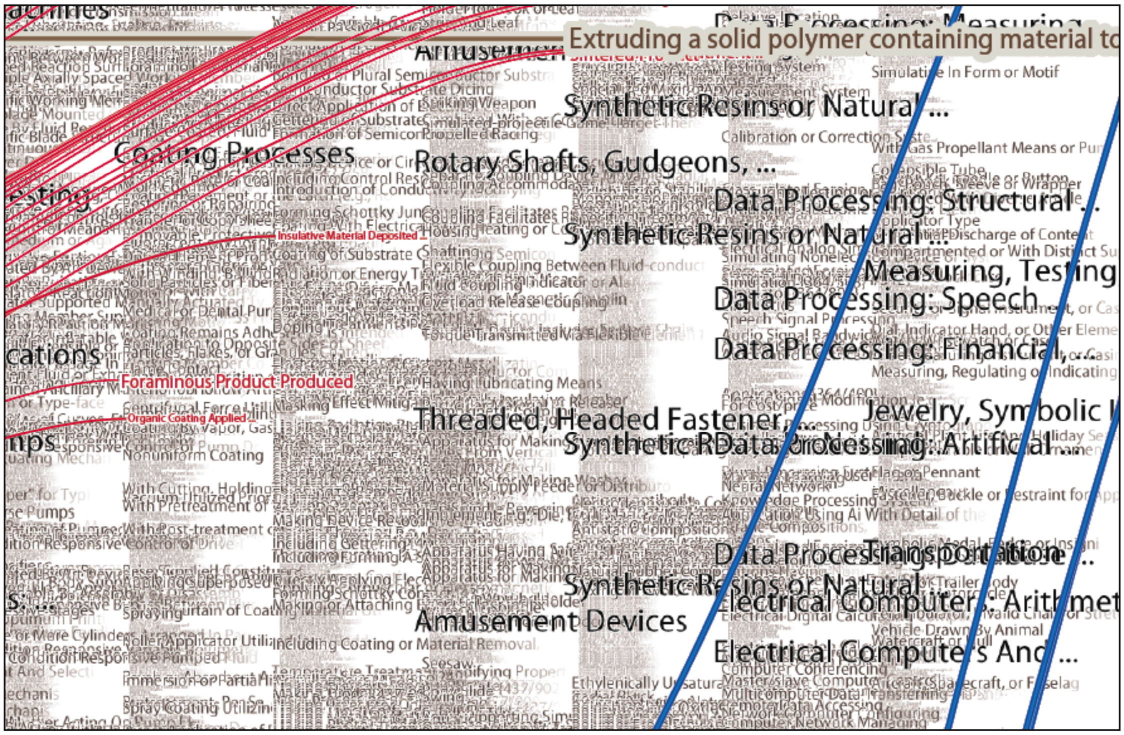
The Power of Reference Systems

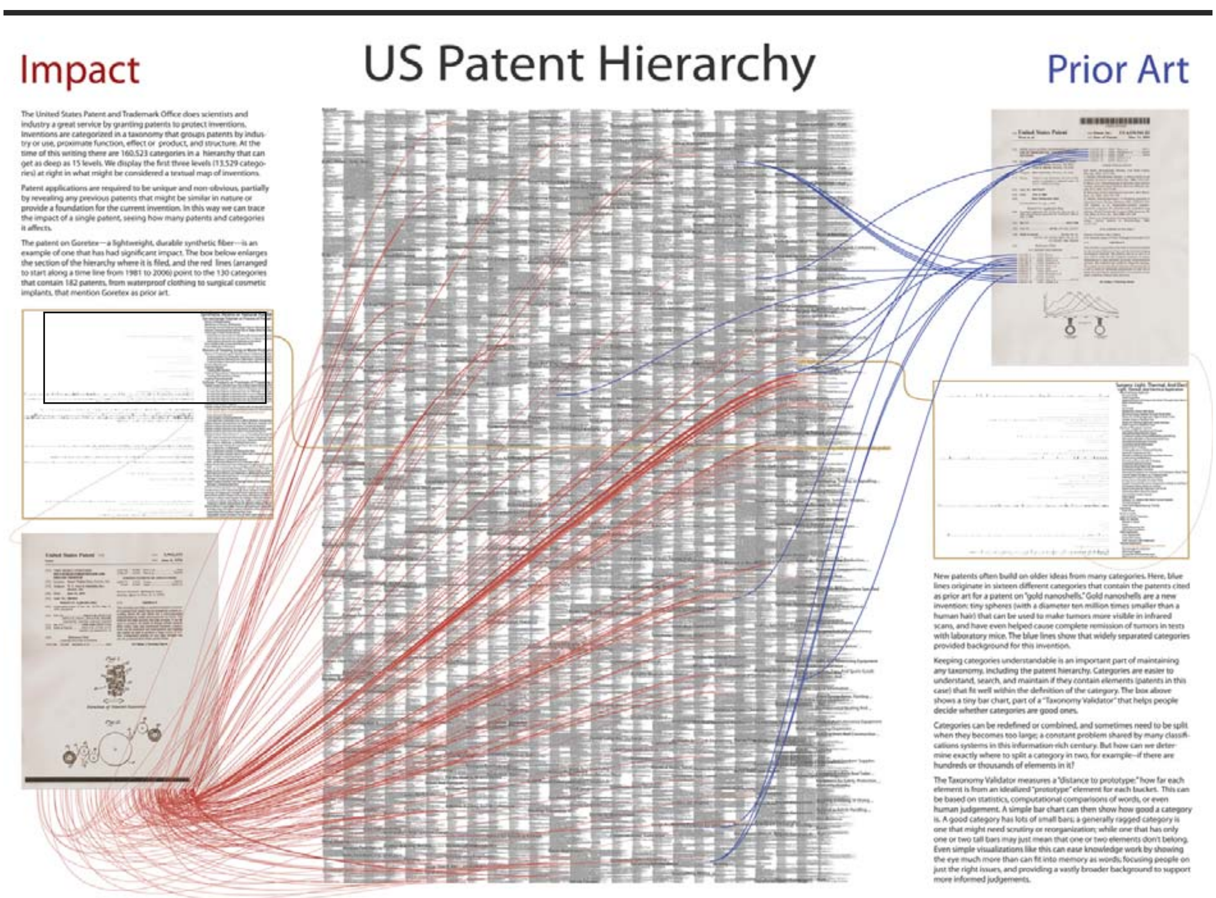
Four Existing Reference Systems
VERSUS
Six Potential Reference Systems of Science

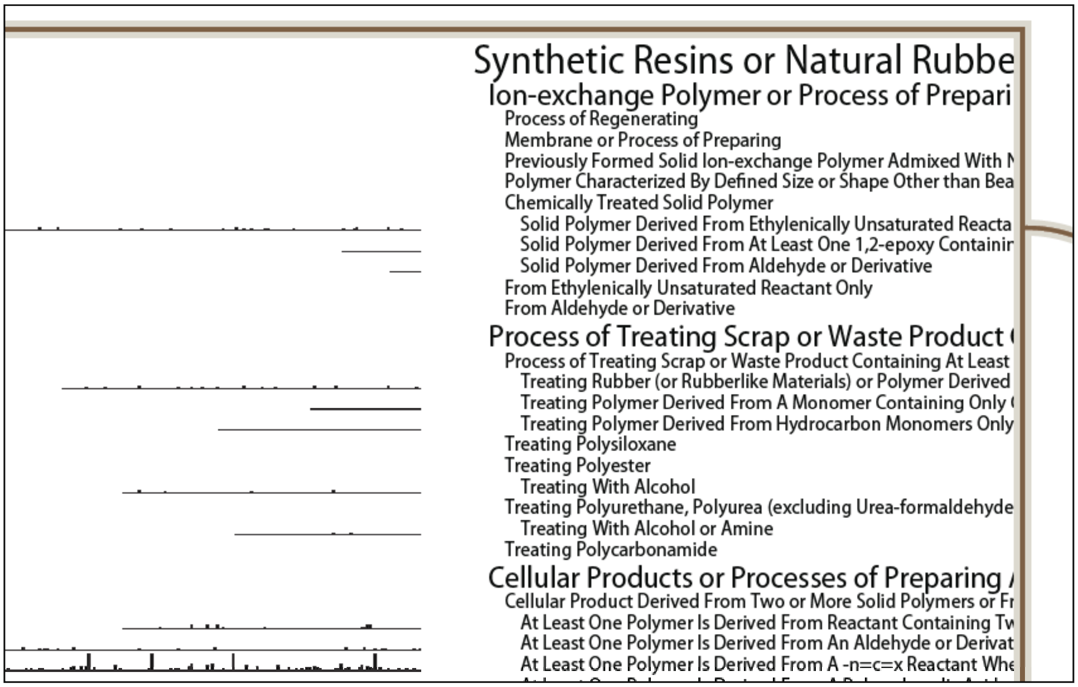


(2nd Iteration of Places & Spaces Exhibit - 2006)

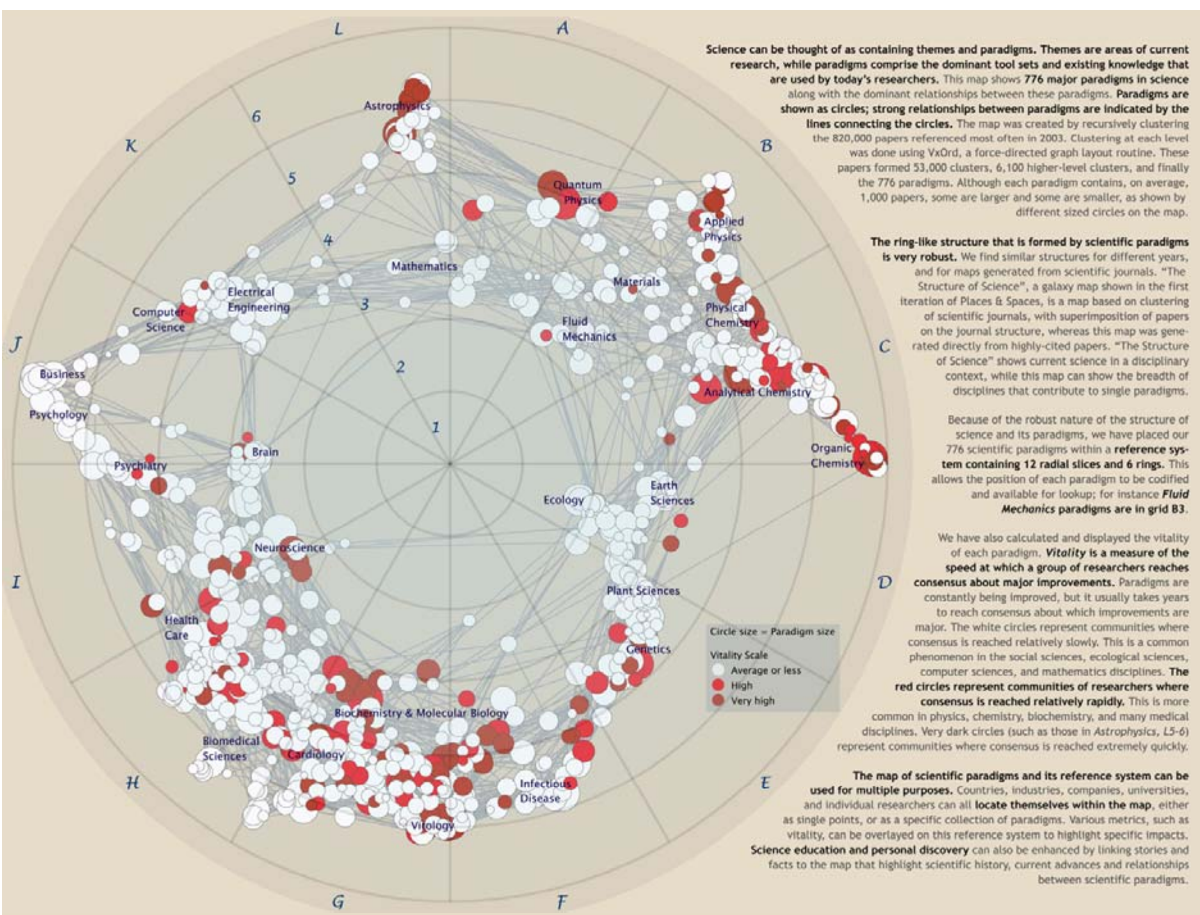



Extruding a solid polymer containing material to
Coating Processes
Rotary Shafts, Gudgeons, ...
Synthetic Resins or Natural ...
Data Processing: Structural ...
Measuring, Testing
Data Processing: Speech
Data Processing: Financial, ...
Foraminous Product Produced
Insulative Material Deposited
Organic Coating Applied
Threaded, Headed Fastener, ...
Jewelry, Symbolic
Amusement Devices
Electrical Computers: Arithmet
Electrical Computers And ...
Nonuniform Coating
Simulative In Form or Motif
Calibration or Correction System
Computer Conferencing
Vehicle Drawn By Animal


US Patent Hierarchy
Impact
The United States Patent and Trademark Office does scientists and industry a great service by granting patents to protect inventions. Inventions are categorized in a taxonomy that groups patents by industry or use, proximate function, effect or product, and structure. At the time of this writing there are 160,523 categories in a hierarchy that can get as deep as 15 levels. We display the first three levels (13,529 categories) at right in what might be considered a textual map of inventions.
Patent applications are required to be unique and non-obvious, partially by revealing any previous patents that might be similar in nature or provide a foundation for the current invention. In this way we can trace the impact of a single patent, seeing how many patents and categories it affects.
The patent on Goretex—a lightweight, durable synthetic fiber—is an example of one that has had significant impact. The box below enlarges the section of the hierarchy where it is filed, and the red lines (arranged to start along a time line from 1981 to 2006) point to the 130 categories that contain 182 patents, from waterproof clothing to surgical cosmetic implants, that mention Goretex as prior art.
Prior Art
New patents often build on older ideas from many categories. Here, blue lines originate in sixteen different categories that contain the patents cited as prior art for a patent on "gold nanoshells." Gold nanoshells are a new invention: tiny spheres (with a diameter ten million times smaller than a human hair) that can be used to make tumors more visible in infrared scans, and have even helped cause complete remission of tumors in tests with laboratory mice. The blue lines show that widely separated categories provided background for this invention.
Keeping categories understandable is an important part of maintaining any taxonomy, including the patent hierarchy. Categories are easier to understand, search, and maintain if they contain elements (patents in this case) that fit well within the definition of the category. The box above shows a tiny bar chart, part of a "Taxonomy Validator" that helps people decide whether categories are good ones.
Categories can be redefined or combined, and sometimes need to be split when they becomes too large; a constant problem shared by many classifications systems in this information-rich century. But how can we determine exactly where to split a category in two, for example—if there are hundreds or thousands of elements in it?
The Taxonomy Validator measures a "distance to prototype:" how far each element is from an idealized "prototype" element for each bucket. This can be based on statistics, computational comparisons of words, or even human judgement. A simple bar chart can then show how good a category is. A good category has lots of small bars; a generally ragged category is one that might need scrutiny or reorganization; while one that has only one or two tall bars may just mean that one or two elements don't belong. Even simple visualizations like this can ease knowledge work by showing the eye much more than can fit into memory as words; focusing people on just the right issues, and providing a vastly broader background to support more informed judgements.



Synthetic Resins or Natural Rubbe
Ion-exchange Polymer or Process of Prepari
Process of Regenerating
Membrane or Process of Preparing
Previously Formed Solid Ion-exchange Polymer Admixed With N
Polymer Characterized By Defined Size or Shape Other than Bea
Chemically Treated Solid Polymer
Solid Polymer Derived From Ethylenically Unsaturated Reacta
Solid Polymer Derived From At Least One 1,2-epoxy Containir
Solid Polymer Derived From Aldehyde or Derivative
From Ethylenically Unsaturated Reactant Only
From Aldehyde or Derivative
Process of Treating Scrap or Waste Product (
Process of Treating Scrap or Waste Product Containing At Least
Treating Rubber (or Rubberlike Materials) or Polymer Derived
Treating Polymer Derived From A Monomer Containing Only (
Treating Polymer Derived From Hydrocarbon Monomers Only
Treating Polysiloxane
Treating Polyester
Treating With Alcohol
Treating Polyurethane, Polyurea (excluding Urea-formaldehyde
Treating With Alcohol or Amine
Treating Polycarbonamide
Cellular Products or Processes of Preparing /
Cellular Product Derived From Two or More Solid Polymers or Fr
At Least One Polymer Is Derived From Reactant Containing Tw
At Least One Polymer Is Derived From An Aldehyde or Derivat
At Least One Polymer Is Derived From A -n=c=x Reactant Whe

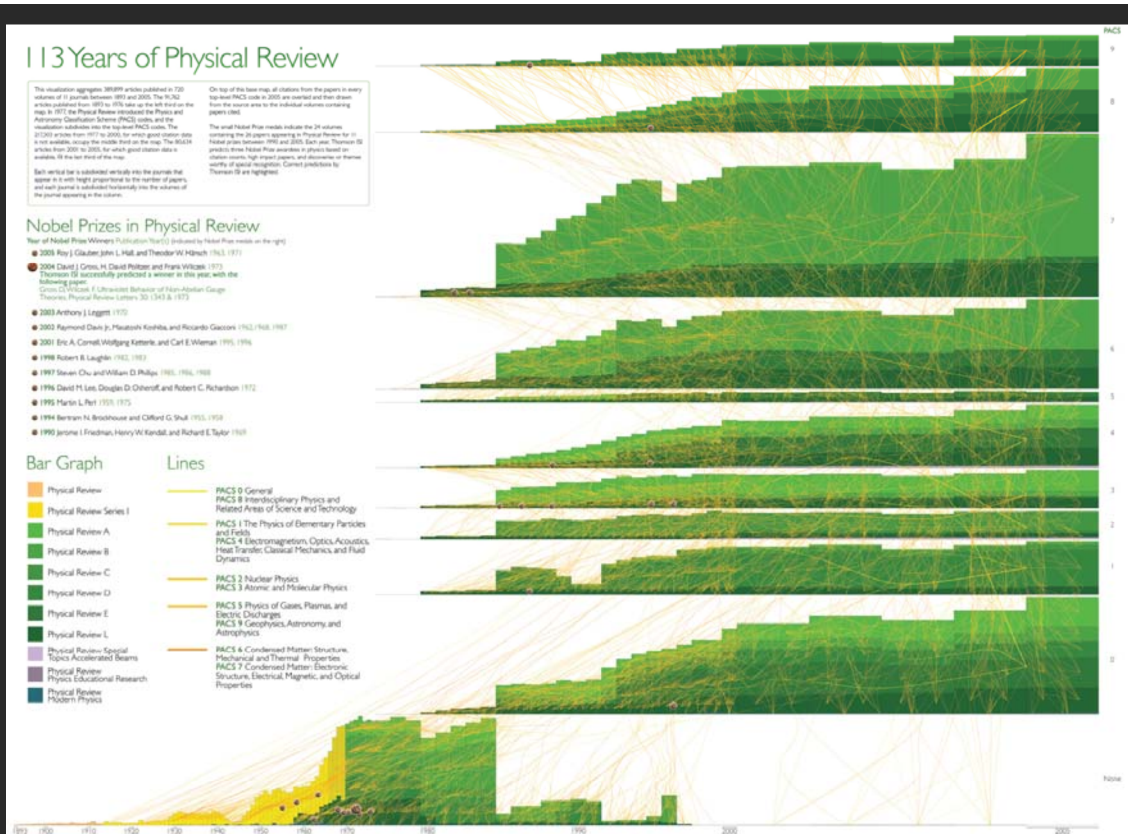


L
A
B
C
D
E
F
G
H
I
J
K
1
2
3
4
5
6
Astrophysics
Quantum Physics
Applied Physics
Mathematics
Materials
Electrical Engineering
Computer Science
Physical Chemistry
Fluid Mechanics
Analytical Chemistry
Business
Psychology
Brain
Organic Chemistry
Psychiatry
Ecology
Earth Sciences
Neuroscience
Plant Sciences
Health Care
Genetics
Biochemistry & Molecular Biology
Biomedical Sciences
Cardiology
Infectious Disease
Virology
Circle size = Paradigm size
Vitality Scale
Average or less
High
Very high
Science can be thought of as containing themes and paradigms. Themes are areas of current research, while paradigms comprise the dominant tool sets and existing knowledge that are used by today's researchers. This map shows **776 major paradigms in science** along with the dominant relationships between these paradigms. **Paradigms are shown as circles; strong relationships between paradigms are indicated by the lines connecting the circles.** The map was created by recursively clustering the 820,000 papers referenced most often in 2003. Clustering at each level was done using VxOrd, a force-directed graph layout routine. These papers formed 53,000 clusters, 6,100 higher-level clusters, and finally the 776 paradigms. Although each paradigm contains, on average, 1,000 papers, some are larger and some are smaller, as shown by different sized circles on the map.
The ring-like structure that is formed by scientific paradigms is very robust. We find similar structures for different years, and for maps generated from scientific journals. "The Structure of Science", a galaxy map shown in the first iteration of Places & Spaces, is a map based on clustering of scientific journals, with superimposition of papers on the journal structure, whereas this map was generated directly from highly-cited papers. "The Structure of Science" shows current science in a disciplinary context, while this map can show the breadth of disciplines that contribute to single paradigms.
Because of the robust nature of the structure of science and its paradigms, we have placed our 776 scientific paradigms within a **reference system containing 12 radial slices and 6 rings.** This allows the position of each paradigm to be codified and available for lookup; for instance ***Fluid Mechanics* paradigms are in grid B3**.
We have also calculated and displayed the vitality of each paradigm. ***Vitality* is a measure of the speed at which a group of researchers reaches consensus about major improvements.** Paradigms are constantly being improved, but it usually takes years to reach consensus about which improvements are major. The white circles represent communities where consensus is reached relatively slowly. This is a common phenomenon in the social sciences, ecological sciences, computer sciences, and mathematics disciplines. **The red circles represent communities of researchers where consensus is reached relatively rapidly.** This is more common in physics, chemistry, biochemistry, and many medical disciplines. Very dark circles (such as those in *Astrophysics, L5-6*) represent communities where consensus is reached extremely quickly.
The map of scientific paradigms and its reference system can be used for multiple purposes. Countries, industries, companies, universities, and individual researchers can all **locate themselves within the map**, either as single points, or as a specific collection of paradigms. Various metrics, such as vitality, can be overlayed on this reference system to highlight specific impacts. **Science education and personal discovery** can also be enhanced by linking stories and facts to the map that highlight scientific history, current advances and relationships between scientific paradigms.

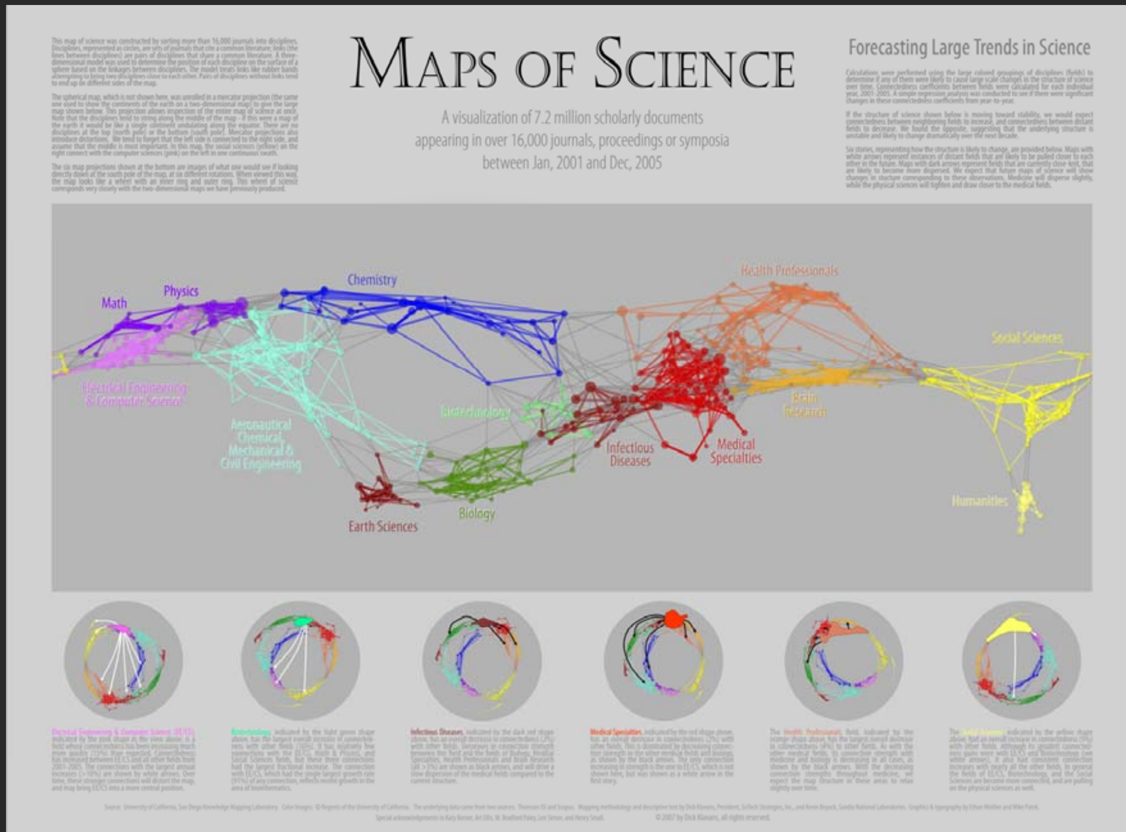
The Power of Forecasts

Four Existing Forecasts
VERSUS
Six Potential Science 'Weather' Forecasts

(3rd Iteration of Places & Spaces Exhibit - 2007)



114 Years of Physical Review - Bruce W. Herr II, Russell Duhon, Katy Borner, Elisha Hardy, Shashikant Penumarthy - 2007



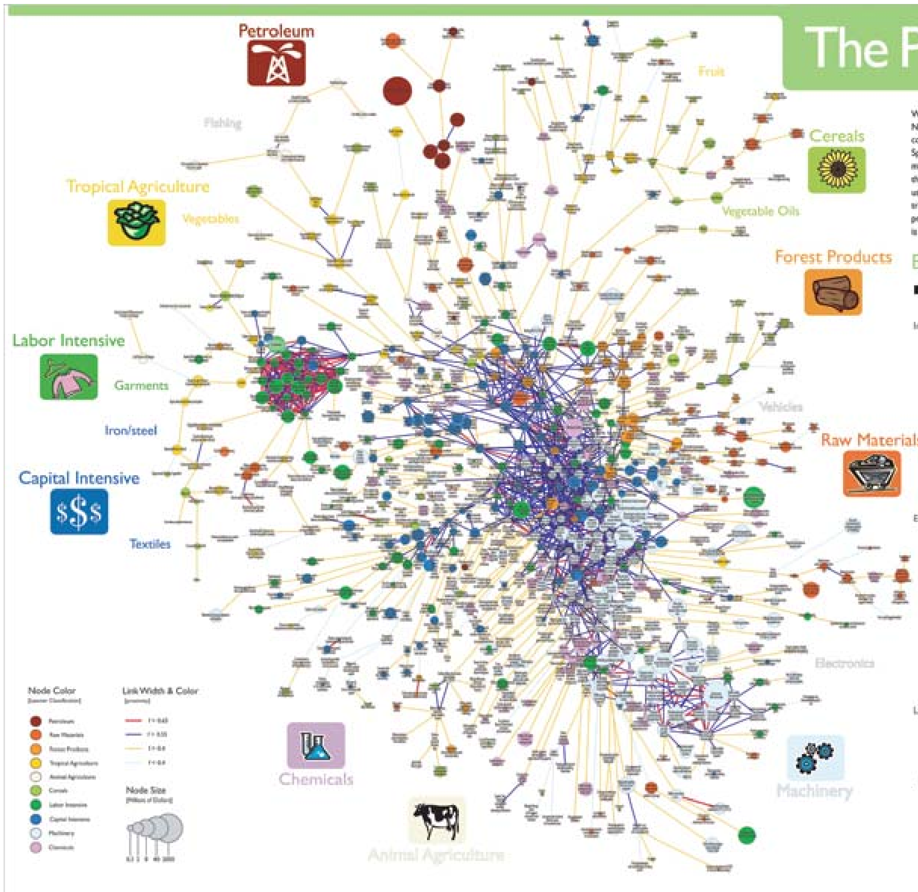
Maps of Science: Forecasting Large Trends in Science - Richard Klavans, Kevin Boyack - 2007

Science Maps for Economic Decision Making

Four Existing Maps
VERSUS
Six Science Maps

(4th Iteration of Places & Spaces Exhibit - 2008)

The Product Space



World trade flow data compiled by Feenstra et al. and available at the National Bureau of Economic Research were used to identify the complete co-export matrix of 775 industrial products for 1998-2000. A Maximum Spanning Tree (MST) algorithm was used to reduce the complete co-export matrix to less than 1% of the links. The resulting network, which combines the MST plus all links with a co-export frequency of at least 0.55, was laid out using a force-directed layout algorithm. Node sizes represent the value of traded products in millions of U.S. dollars. Their color corresponds to ten product groups identified using the Leamer classification. Each product class is labeled by an icon. Link color and width indicate the frequency of joint exports.

Economic Footprint

■ Indicate Relevant Exports

Industrialized Countries

East Asia Pacific

Latin America & the Caribbean

The network has a core-periphery structure with higher value product classes, e.g., machinery, and chemicals, in the core and lower quality classes, e.g., fishing and garments, in the periphery. Products at the core of the network are highly interconnected while products in the periphery are sparsely interlinked.

Each country has a certain product export footprint. Relevant exports by 'Industrialized Countries', 'East Asia Pacific' and 'Latin America & the Caribbean' are given on the right.

Traditional growth theory assumes that there is always a more sophisticated product within reach. However, given the core-periphery structure of the product space, the distances between products differ considerably.

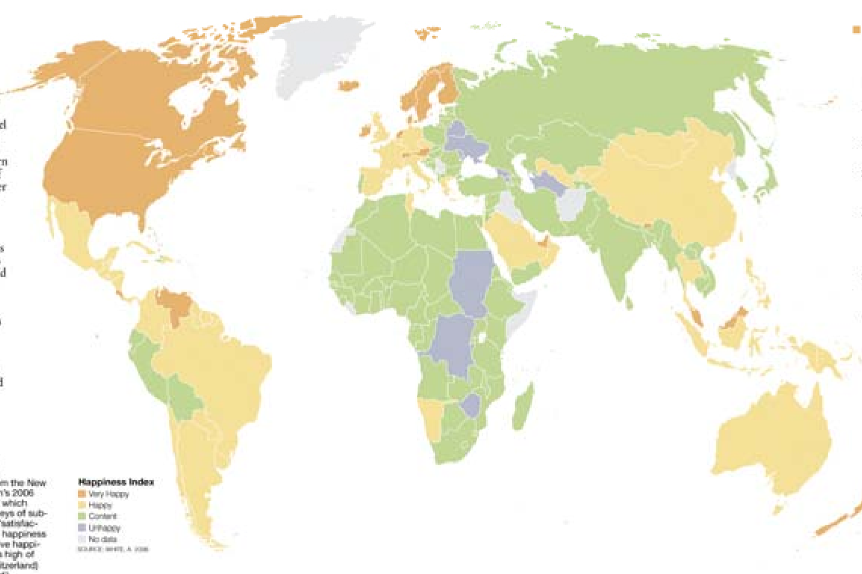
Countries that operate at the core have capabilities to develop and manufacture a wide range of products. Yet, countries that mostly operate in the periphery of the product space have much fewer opportunities for diversification. A country's current footprint and the structure of the product space have a major impact on a country's future development.

Happiness Depends on Various Factors

"It's time we admitted there's **more to life** than money."
- David Cameron, U.K. leader of the opposition, 2006

Social scientists are starting to include relative happiness with hard data on economic status, health, and other factors as they assess quality of life. They rely on surveys of "subjective well-being"—how good people feel about their lives. A world map of one "happiness index" shows many, but not all, wealthy northern countries faring well. Residents of sub-Saharan Africa and the former Soviet Union, meanwhile, report particularly low levels of contentment.

Any attempt to measure happiness will fall short—each life is a series of joys, struggles, and sorrows, and satisfaction can depend as much on outlook as on circumstances. Averages obscure the happy moments in struggling nations, as well as people who suffer from poor health, poverty, or discrimination in countries that rank high. Still, happiness indices can help researchers move beyond simple economics as they track progress—or backsliding—over time.



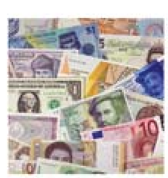
MEASURING THE INTANGIBLE
The map is derived from the New Economics Foundation's 2006 "Happy Planet Index," which drew on over 100 surveys of subjective well-being. Its "satisfaction with life scale"—a happiness index—ranks the relative happiness of nations, from a high of 273 (Denmark and Switzerland) to a low of 100 (Burundi).

RANKING THE WORLD'S HAPPIEST PLACES
Northern Europe, North America, and several wealthy countries make the list, but so do many less prosperous island nations.

1. DENMARK
   SWITZERLAND
2. AUSTRIA
   ICELAND
3. BAHAMAS
   FINLAND
   SWEDEN
4. BHUTAN
   BRUNEI
   CANADA
   IRELAND
   LUXEMBOURG
5. COSTA RICA
   MALTA
   NETHERLANDS
6. ANTIGUA AND BARBUDA
   MALAYSIA
   NEW ZEALAND
   NORWAY
   SEYCHELLES
   ST. KITTS AND NEVIS
   UNITED ARAB EMIRATES
   UNITED STATES
   VANUATU
   VENEZUELA

DEFINING WELL-BEING
By comparing the happiness index to data from the UN, the CIA, and other sources, a U.K. psychologist determined that good health and health care, enough money for fundamental needs, and access to basic education are the most important factors for subjective well-being. European countries top all three measures.

HEALTH
Japan boasts the world's longest life expectancy—one measure of overall health. Swaziland, at the other end of the scale, is plagued by poverty, disease, and violence. Disparities in access to health care divide many countries into haves and have-nots.



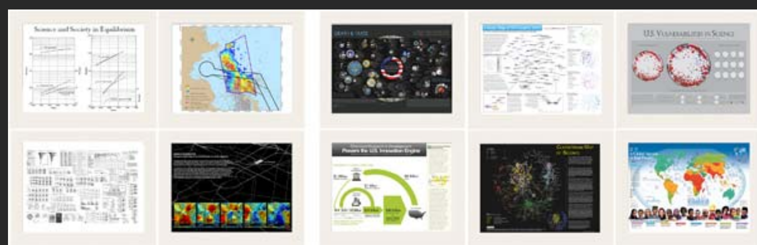
WEALTH
Money still can't buy love, or happiness, and wealthier people aren't always more content. Still, tiny Luxembourg, which takes top rank in per capita Gross Domestic Product (GDP), also rates a 253 on the happiness index. Real poverty means real misery, a fate shared by billions.

EDUCATION
Residents of Australia can expect to spend more time in school—an average of almost 21 years—than citizens of any other country. But only a basic education is needed to see a significant jump in overall happiness. Around the world, hundreds of millions lack even that.

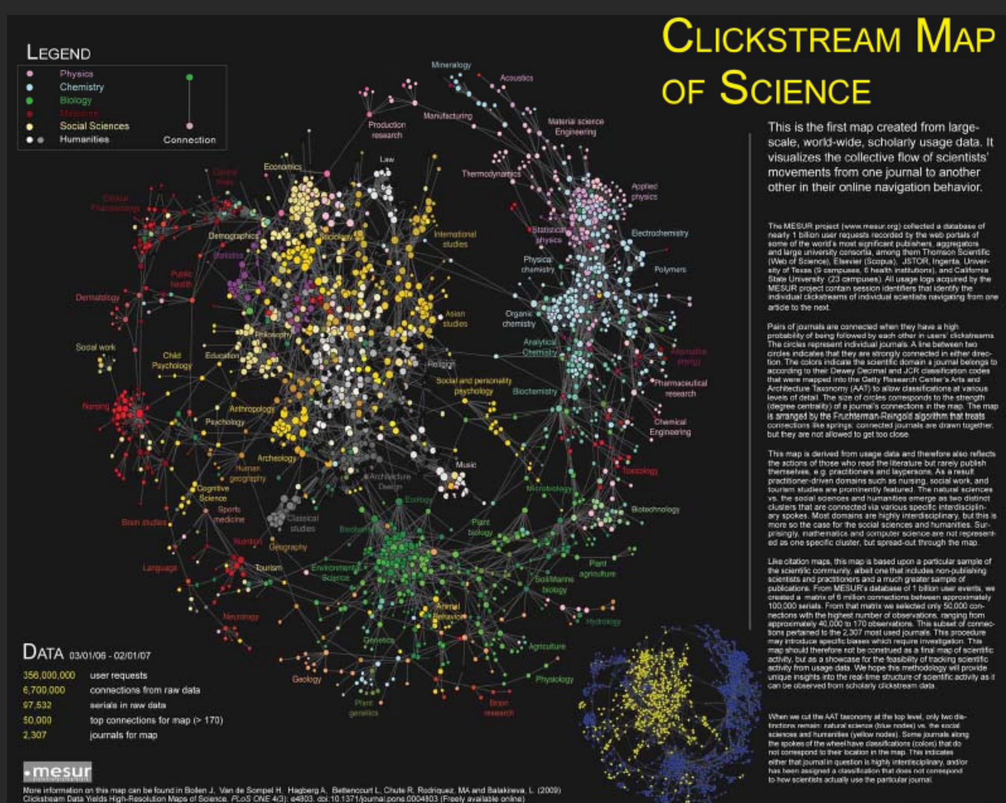
*Photos different than those used in EarthPulse publication

Science Maps for Science Policy Making

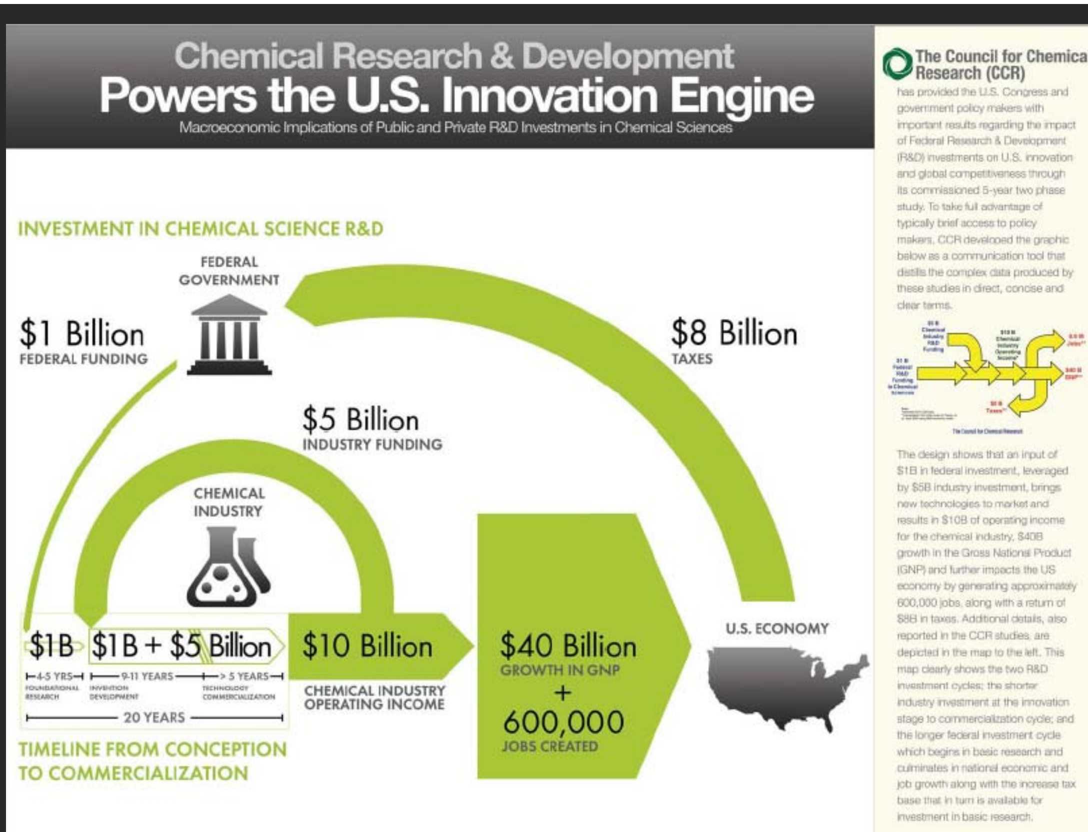
Four Existing Maps
VERSUS
Six Science Maps



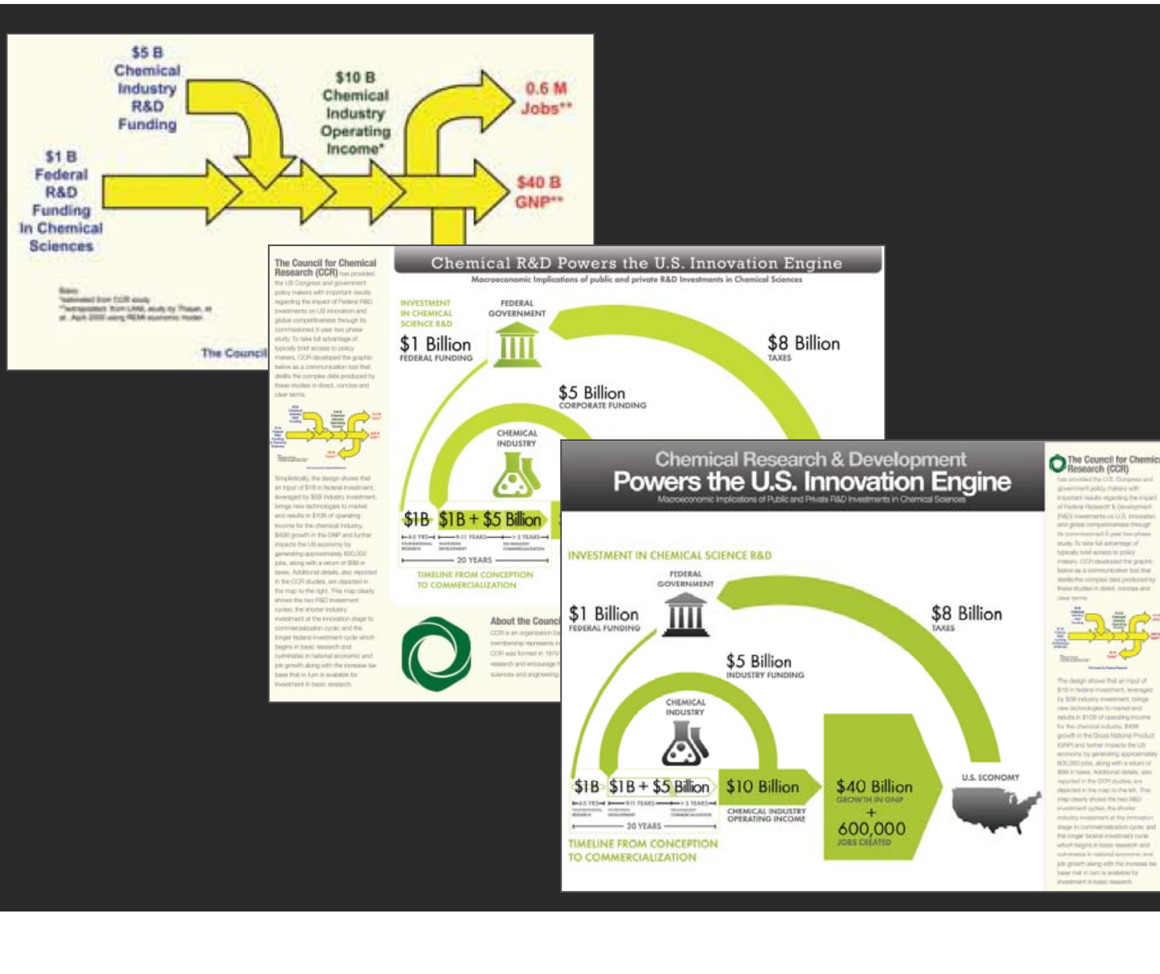
(5th Iteration of Places & Spaces Exhibit - 2009)



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



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



Additional Elements of the Exhibit

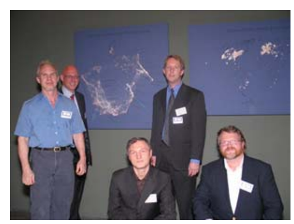
Illuminated Diagram Display

Hands-on Science Maps for Kids

Worldprocessor Globes

Illuminated Diagram Display

W. Bradford Paley, Kevin W. Boyack, Richard Kalvans, and Katy Börner (2007) Mapping, Illuminating, and Interacting with Science. SIGGRAPH 2007.

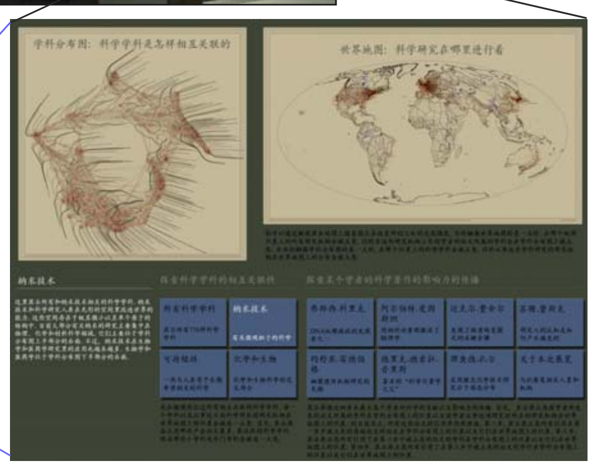


Large-scale, high resolution prints illuminated via projector or screen.

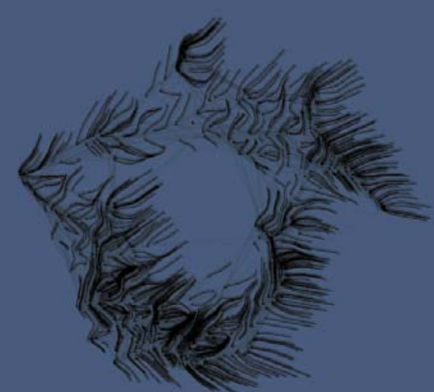
Questions:
- Who is doing research on what topic and where?
- What is the 'footprint' of interdisciplinary research fields?
- What impact have scientists?

Interactive touch panel.

Contributions:
- Interactive, high resolution interface to access and make sense of data about scholarly activity.



Topic Map: How Scientific Paradigms Relate



Geographic Map: Where Science Gets Done

You may run your finger over each of these maps to control the lighting on the other: touching a place on the world map will light up topics studied in that place; touching a paradigm on the topic map will light up the places that study that topic.

Nanotechnology

This overlay shows the distribution of nanotechnology within the paradigms of science. The majority of current work in nanotechnology takes places in physics, chemistry, and materials science, at the upper right portion of the map. However, an increasing amount of nanotechnology is being applied in the biological and medical sciences, at the lower right.

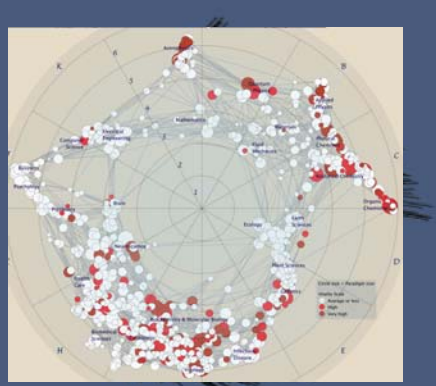
- **All Topics** — *Sweep through all 776 scientific paradigms*
- **Nanotechnology** — *Science on the tiny scale of molecules*
- **Sustainability** — *The science behind our long-term hopes*
- **Biology & Chemistry** — *The interface between these two vital fields*

We sweep slowly through adjoining related topics, lighting up the places in the world that study each topic. You may select a subset of the topics that deal with these three interesting subjects by touching it.

- **Francis H. C. CRICK** — *Co-discovered DNA's double helix*
- **Albert EINSTEIN** — *Revitalized physics with Relativity theories*
- **Michael E. FISHER** — *Models critical phase transitions of matter*
- **Susan T. FISKE** — *Connects perception and stereotypes*
- **Joshua LEDERBERG** — *Pioneer in bacterial genetic mechanisms*
- **Derek J. de Solla PRICE** — *Known as the "Father of Scientometrics"*
- **Richard N. ZARE** — *Uses laser chemistry in molecular dynamics*
- **About this display** — *People & organizations that helped create it*

A single person's spreading influence is shown as a series of four snapshots. First, we light only topics and places relating to that person's papers—papers that are still highly cited today. The second lights everything that cites that original work. Note that this first-generation impact extends to far more topics than did the original work. The third snapshot lights science that cites the second; and the fourth lights science that cites the third.

Topic Map: How Scientific Paradigms Relate



Geographic Map: Where Science Gets Done

You may run your finger over each of these maps to control the lighting on the other: touching a place on the world map will light up topics studied in that place; touching a paradigm on the topic map will light up the places that study that topic.

Nanotechnology

This overlay shows the distribution of nanotechnology within the paradigms of science. The majority of current work in nanotechnology takes places in physics, chemistry, and materials science, at the upper right portion of the map. However, an increasing amount of nanotechnology is being applied in the biological and medical sciences, at the lower right.

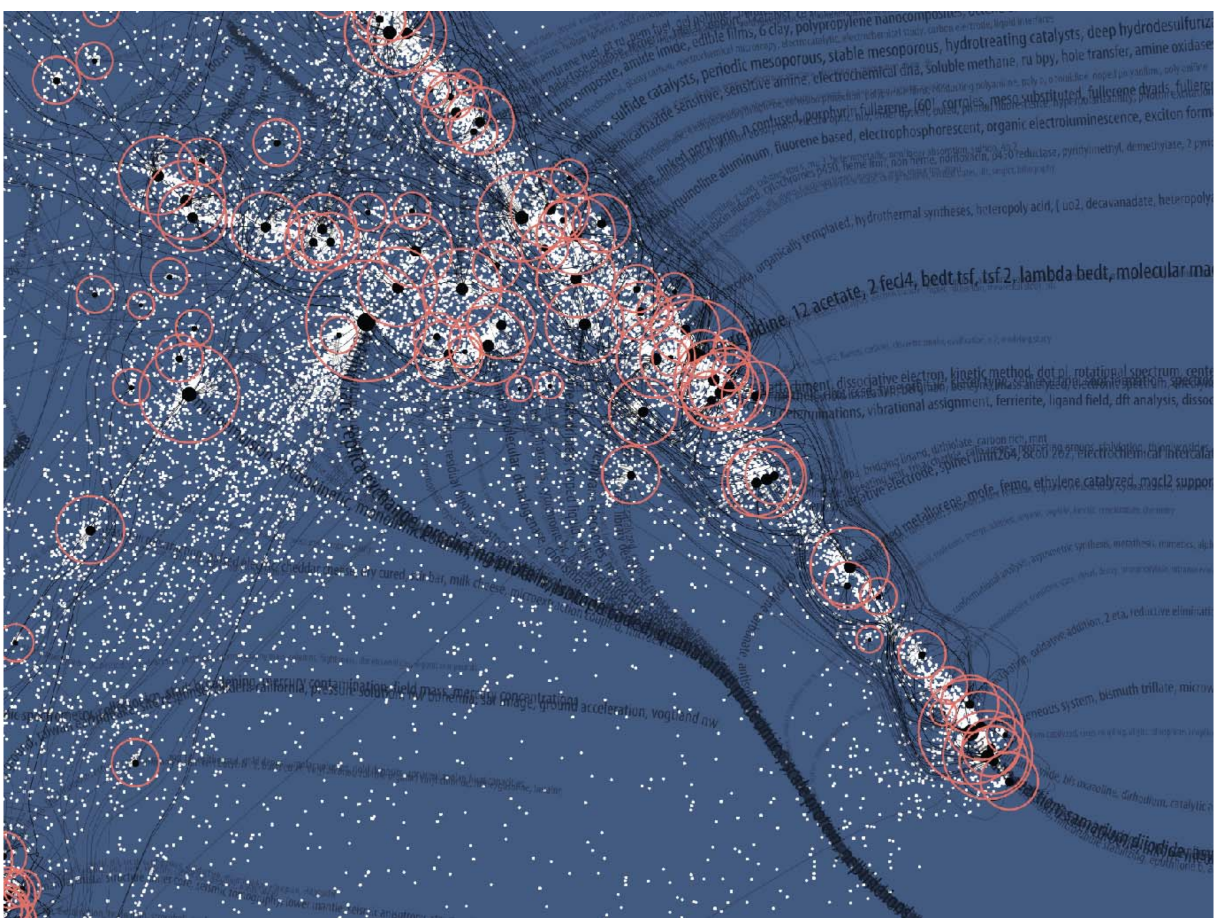
- **All Topics** — *Sweep through all 776 scientific paradigms*
- **Nanotechnology** — *Science on the tiny scale of molecules*
- **Sustainability** — *The science behind our long-term hopes*
- **Biology & Chemistry** — *The interface between these two vital fields*

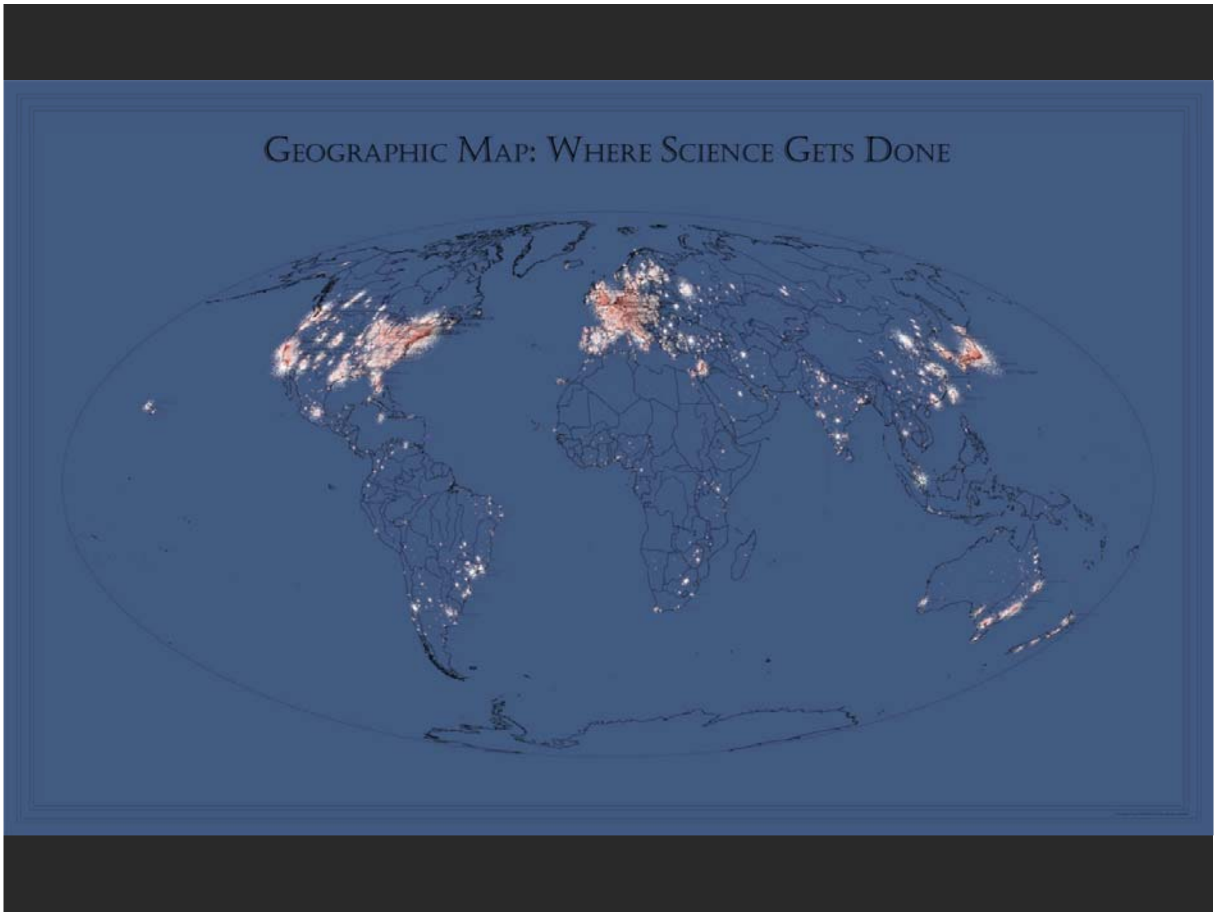
We sweep slowly through adjoining related topics, lighting up the places in the world that study each topic. You may select a subset of the topics that deal with these three interesting subjects by touching it.

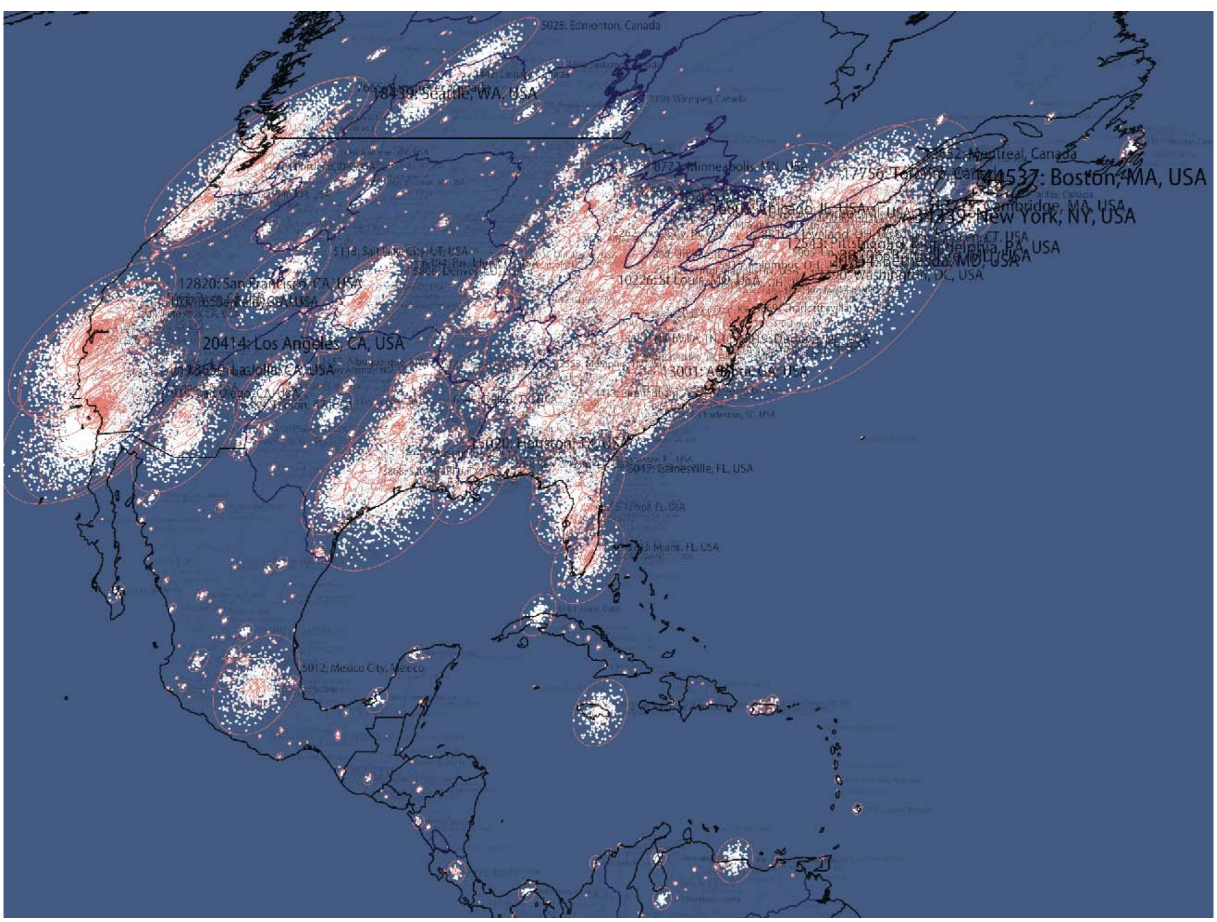
- **Francis H. C. CRICK** — *Co-discovered DNA's double helix*
- **Albert EINSTEIN** — *Revitalized physics with Relativity theories*
- **Michael E. FISHER** — *Models critical phase transitions of matter*
- **Susan T. FISKE** — *Connects perception and stereotypes*
- **Joshua LEDERBERG** — *Pioneer in bacterial genetic mechanisms*
- **Derek J. de Solla PRICE** — *Known as the "Father of Scientometrics"*
- **Richard N. ZARE** — *Uses laser chemistry in molecular dynamics*
- **About this display** — *People & organizations that helped create it*

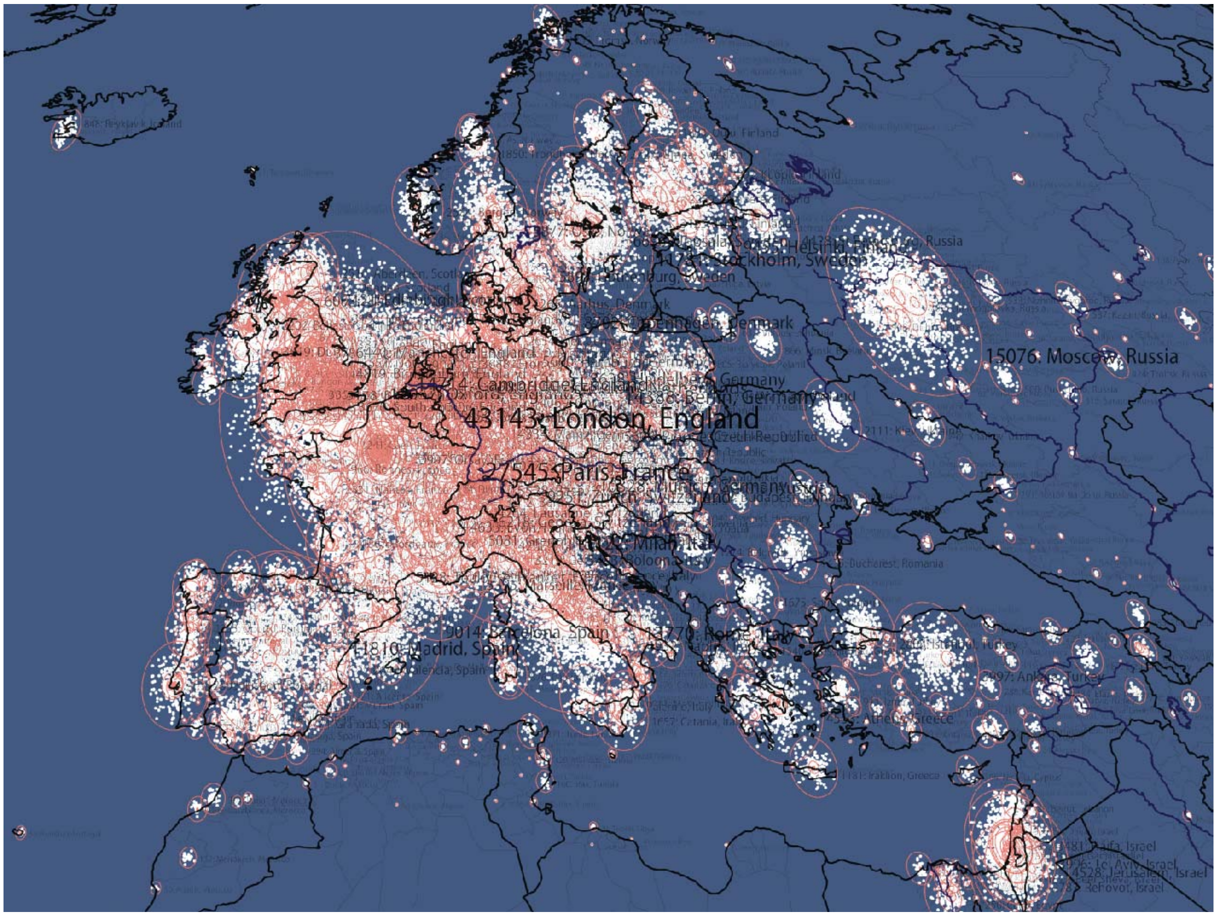
A single person's spreading influence is shown as a series of four snapshots. First, we light only topics and places relating to that person's papers—papers that are still highly cited today. The second lights everything that cites that original work. Note that this first-generation impact extends to far more topics than did the original work. The third snapshot lights science that cites the second; and the fourth lights science that cites the third.

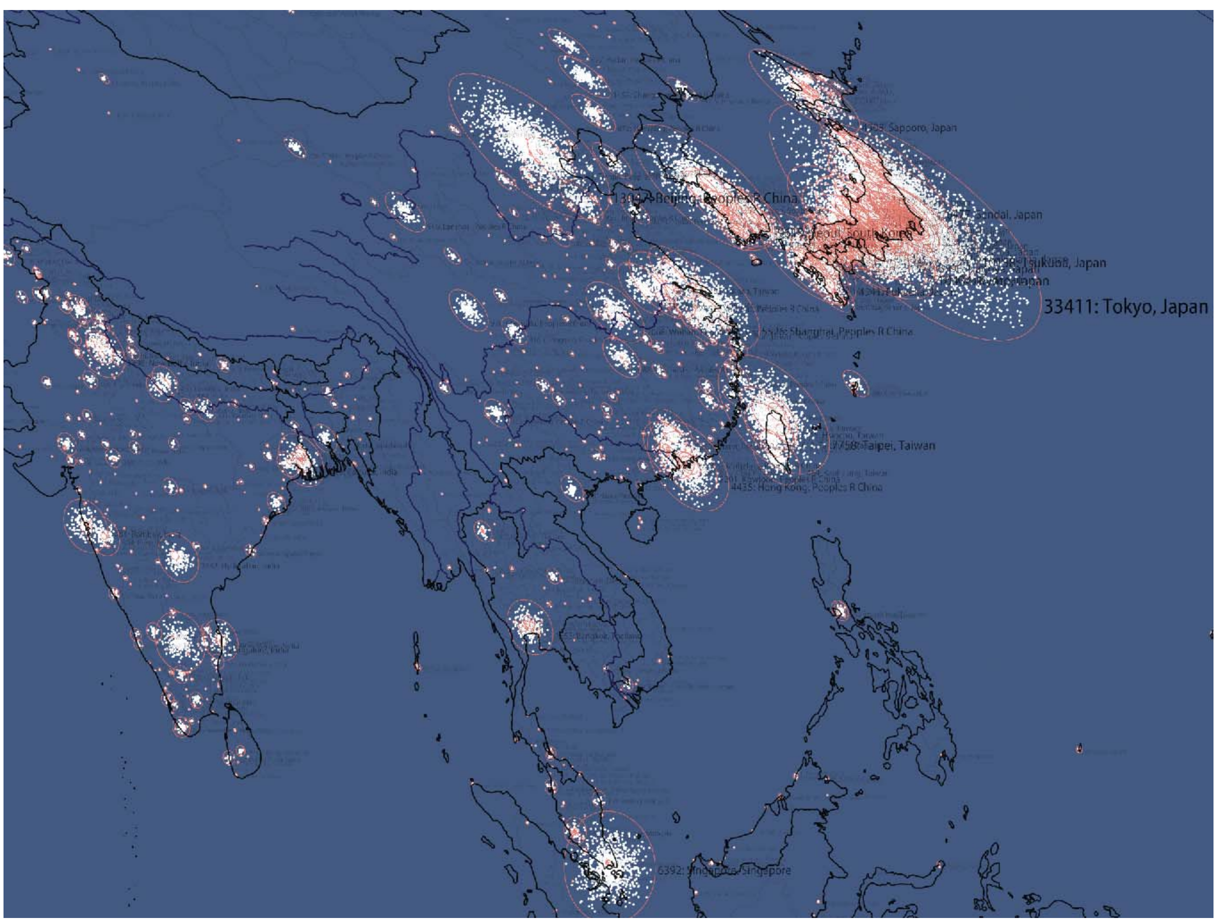

TOPIC MAP: HOW SCIENTIFIC PARADIGMS RELATE

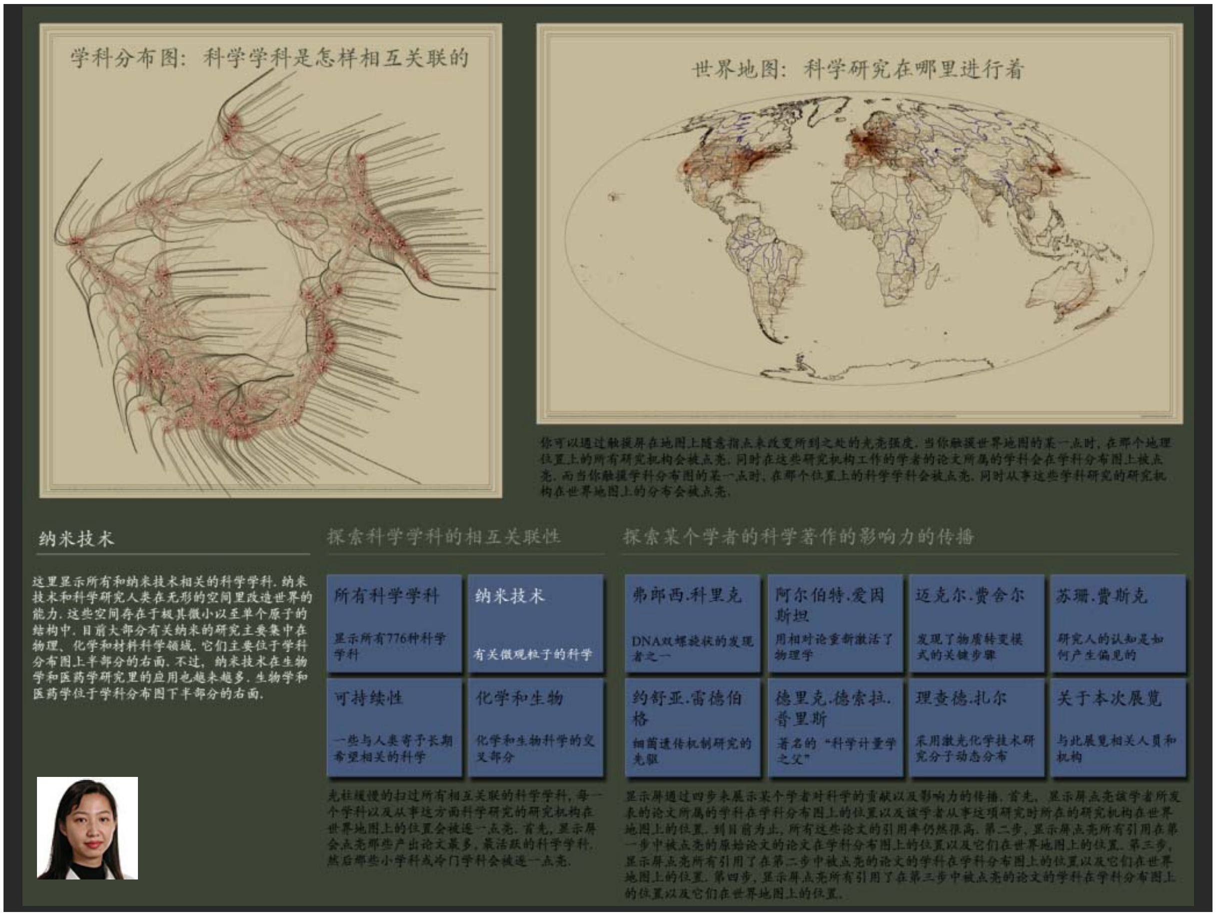



GEOGRAPHIC MAP: WHERE SCIENCE GETS DONE


5028: Edmonton, Canada
Seattle, WA, USA
Montreal, Canada
Boston, MA, USA
New York, NY, USA
20414: Los Angeles, CA, USA
5012: Mexico City, Mexico
5017: Gainesville, FL, USA


43143: London, England
15076: Moscow, Russia


33411: Tokyo, Japan



Re-implementation of Illuminated Diagram Software
by Advanced Visualization Lab, Indiana University

Drives unlimited number of ID screens.

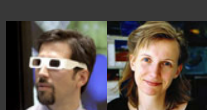
Science World

Medicine Asia

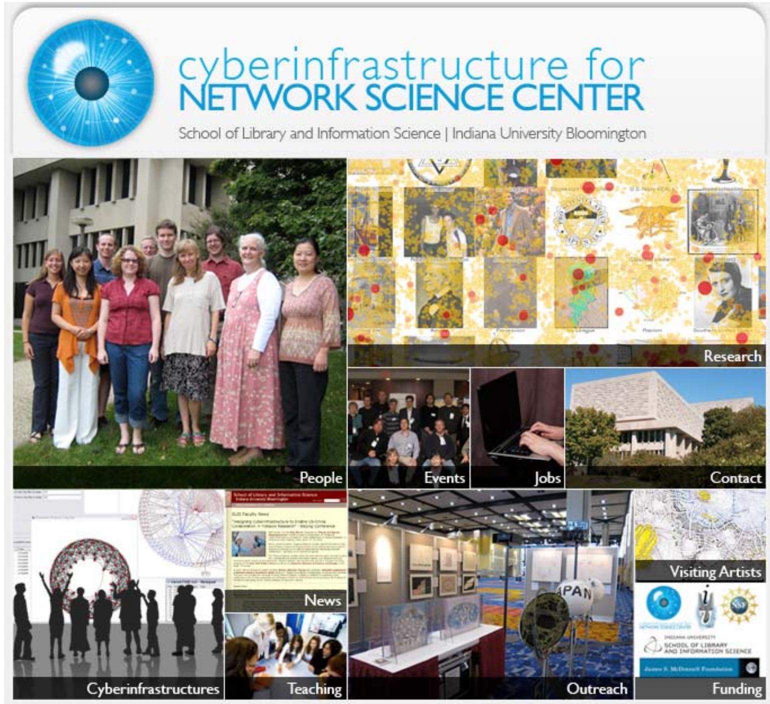
Cancer Japan

... ...

Touch screen for direct interaction.
Keyword and name search.
Selection of canned queries for
- interdisciplinary research areas
- famous people
- activity patterns, e.g., bursts, trends, etc.



Science Maps in "Expedition Zukunft" science train visiting 62 cities in 7 months, 12 coaches, 300 m long. Opening was on April 23rd, 2009 by German Chancellor Merkel, [*http://www.expedition-zukunft.de*](http://www.expedition-zukunft.de)



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