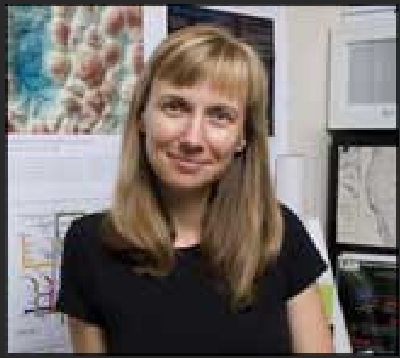
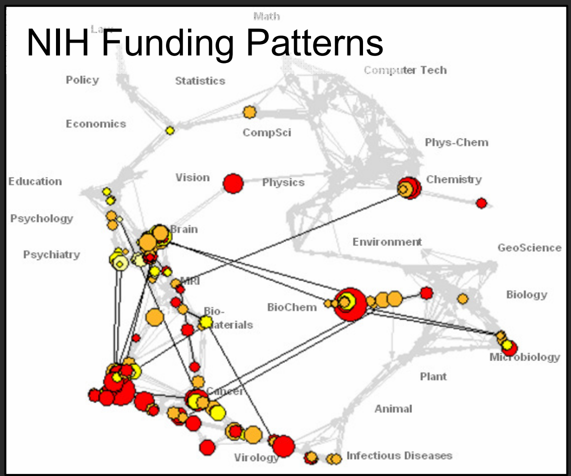
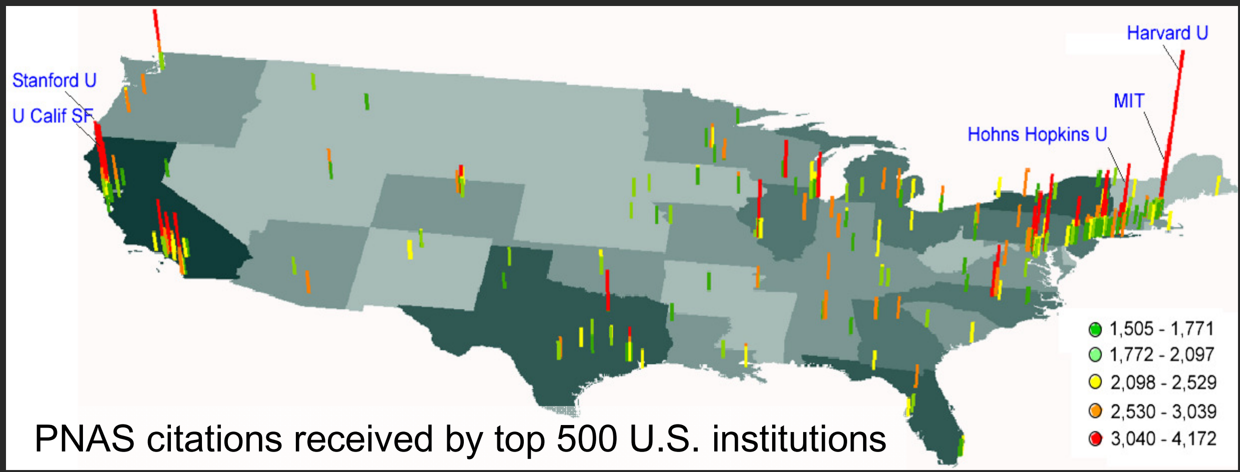
Computational Scientometrics: Mapping All of Science



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

Midnight Seminar Talk at Telcordia, NY, Aug 15, 2006.





This Talk has Three Parts:

1. Why study the structure and evolution of science?
2. What infrastructure is needed to study science?

3. Cyberinfrastructures under development: CIShell, IVC, and NWB

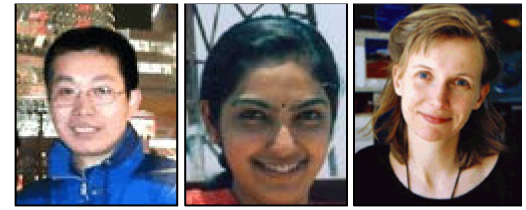
This Talk has Three Parts:

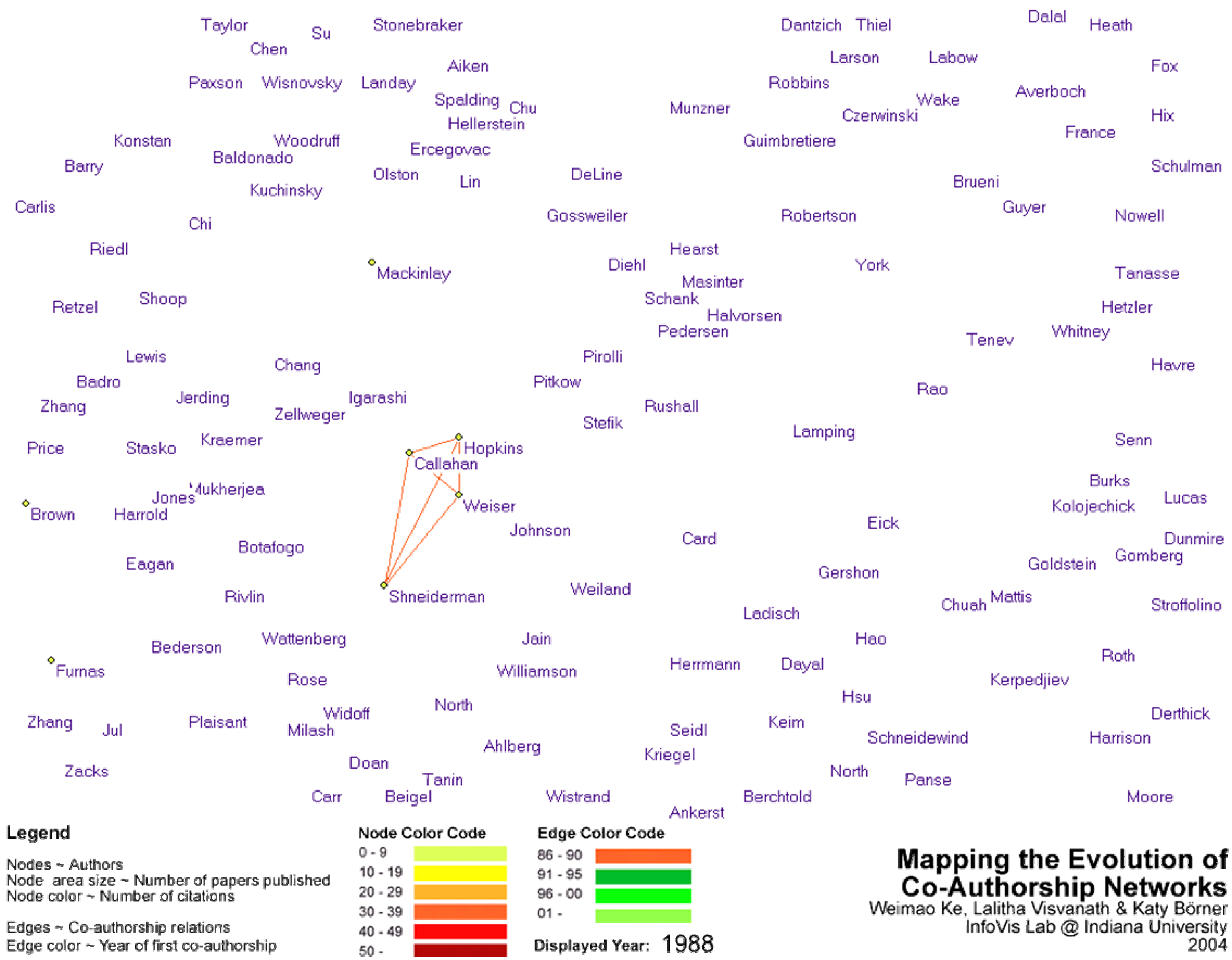
1. **Why study the structure and evolution of science?**
2. What infrastructure is needed to study science?

3. Cyberinfrastructures under development:
   CIShell, IVC, and NWB

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

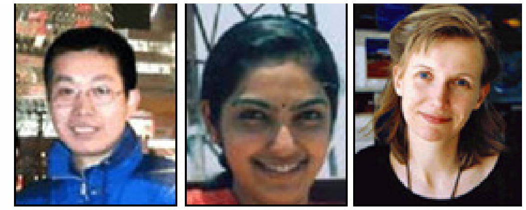
Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest.

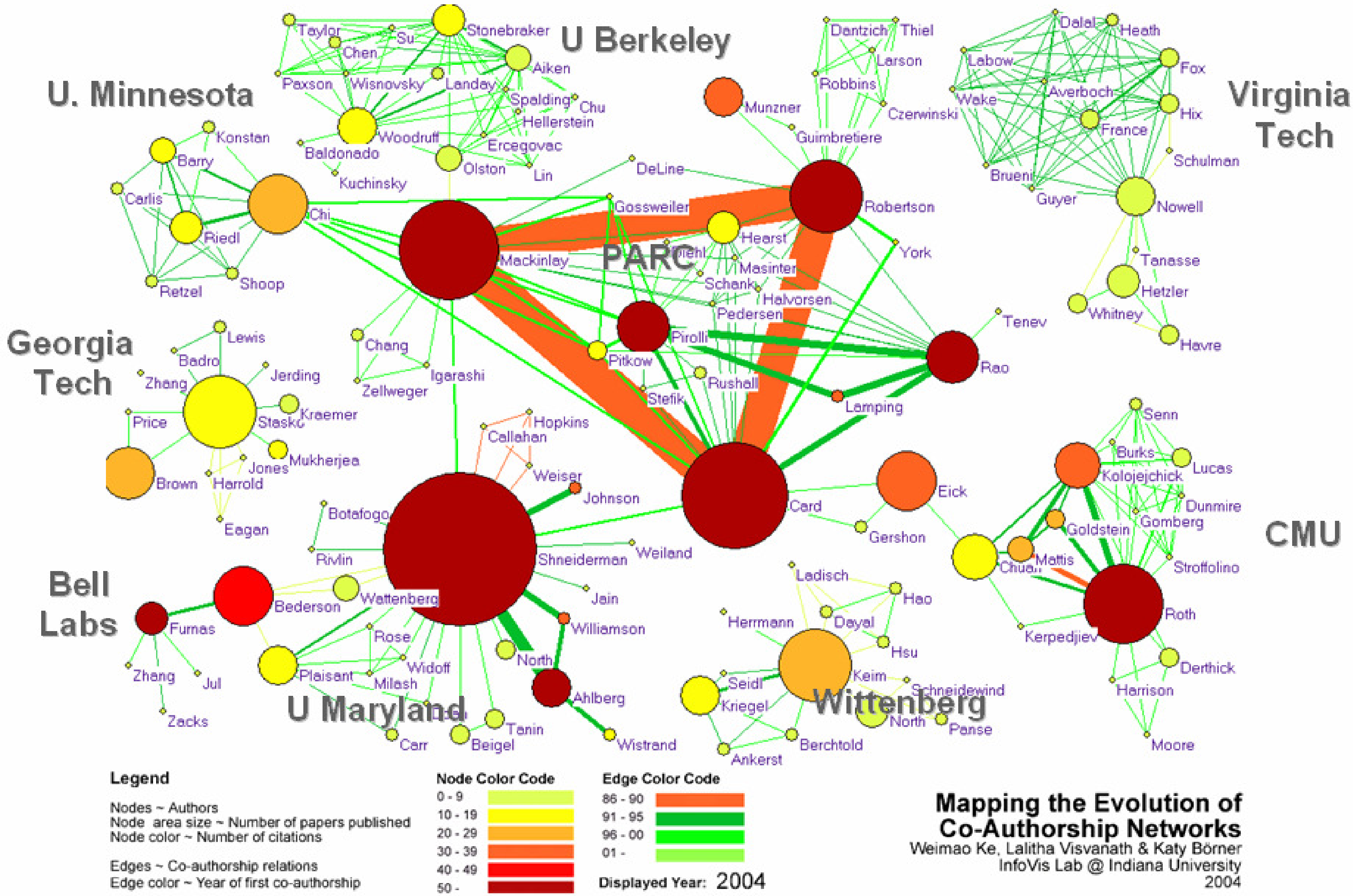


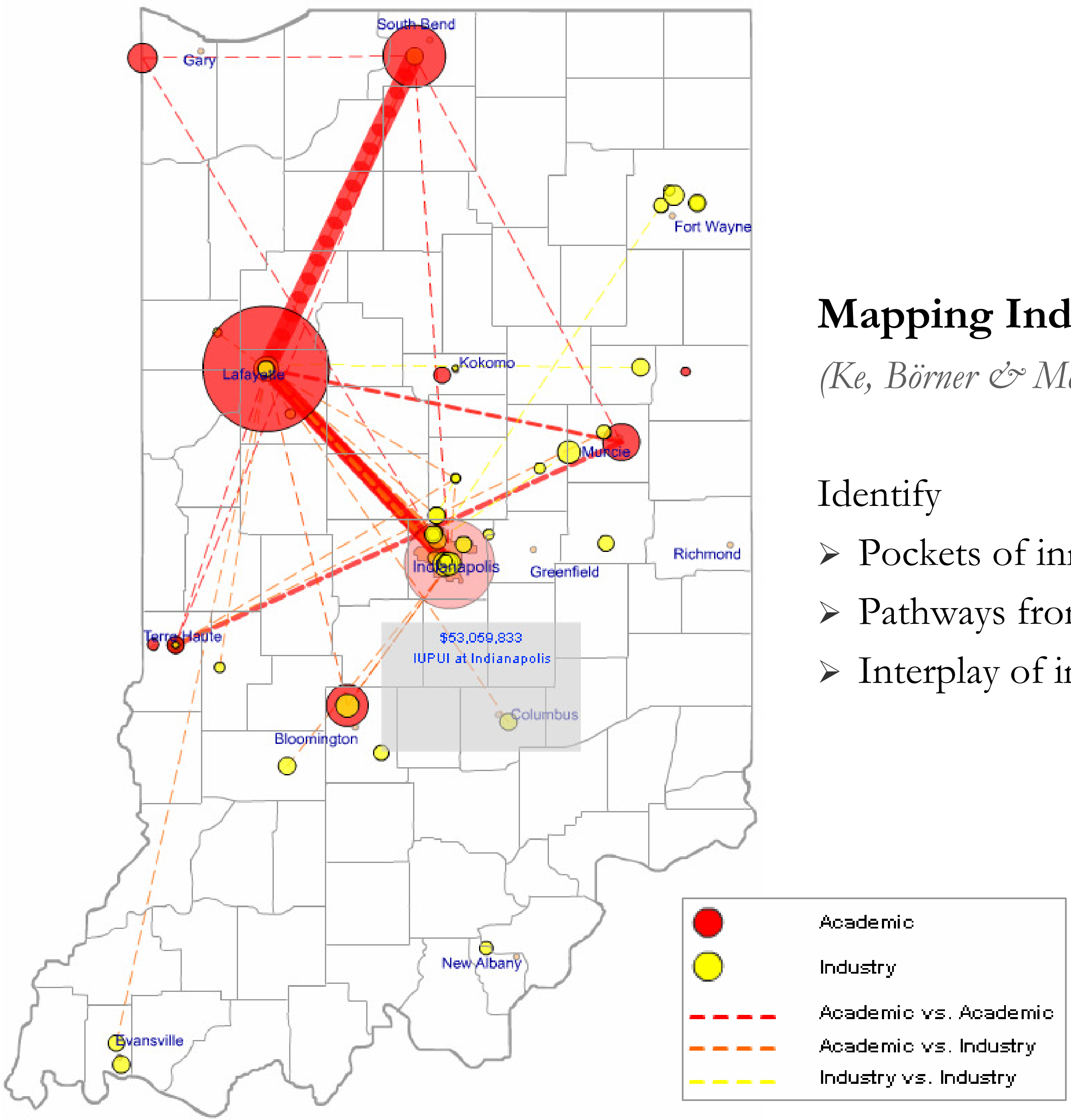


Mapping the Evolution of Co-Authorship Networks

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







Mapping Indiana's Intellectual Space

(Ke, Börner & Mei, 2005)

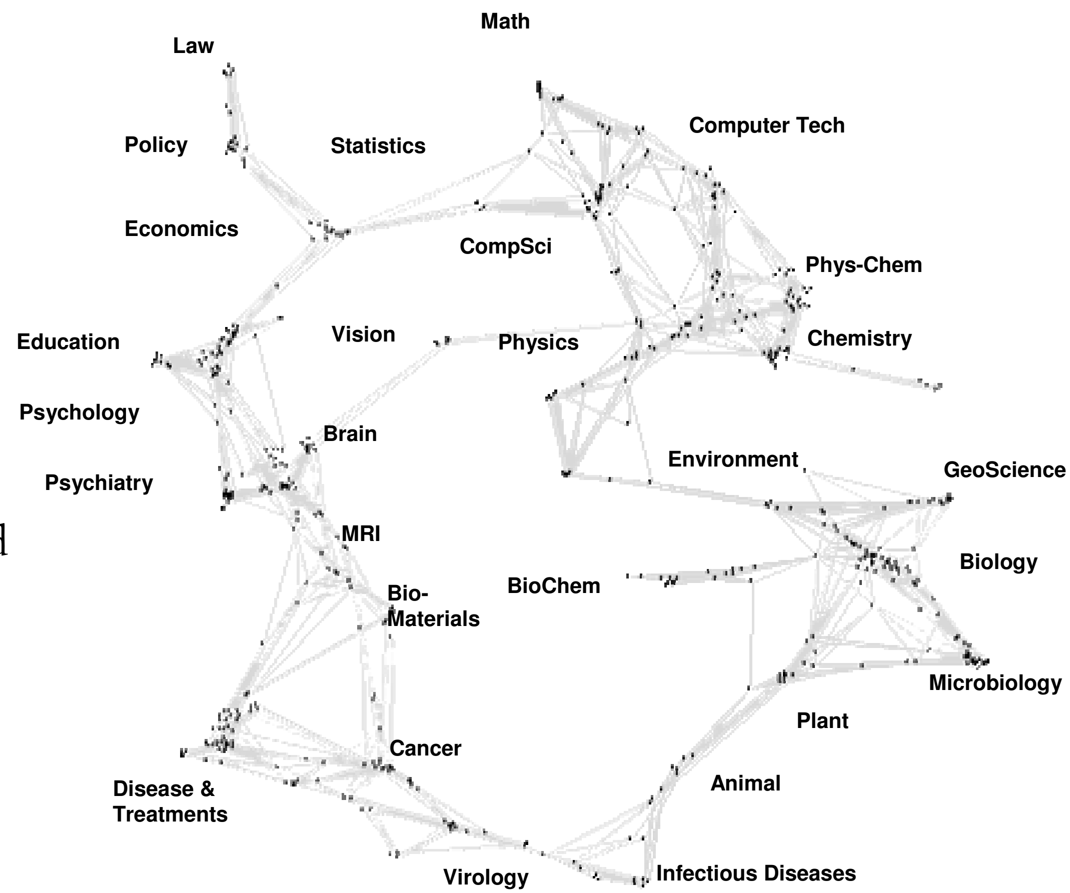
Identify

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

Latest 'Base Map' of Science

Kevin W. Boyack & Richard Klavans, unpublished work.

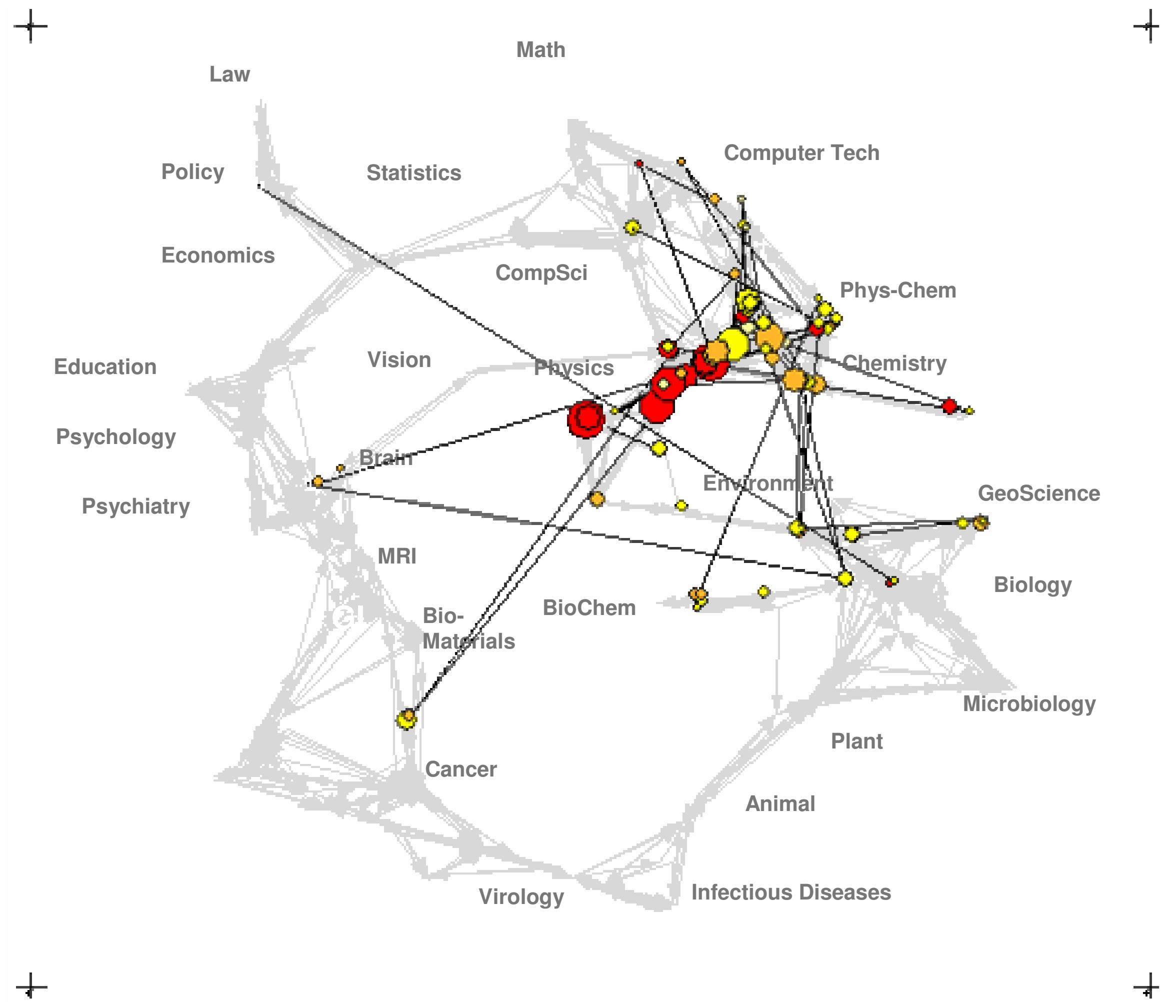
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y)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



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

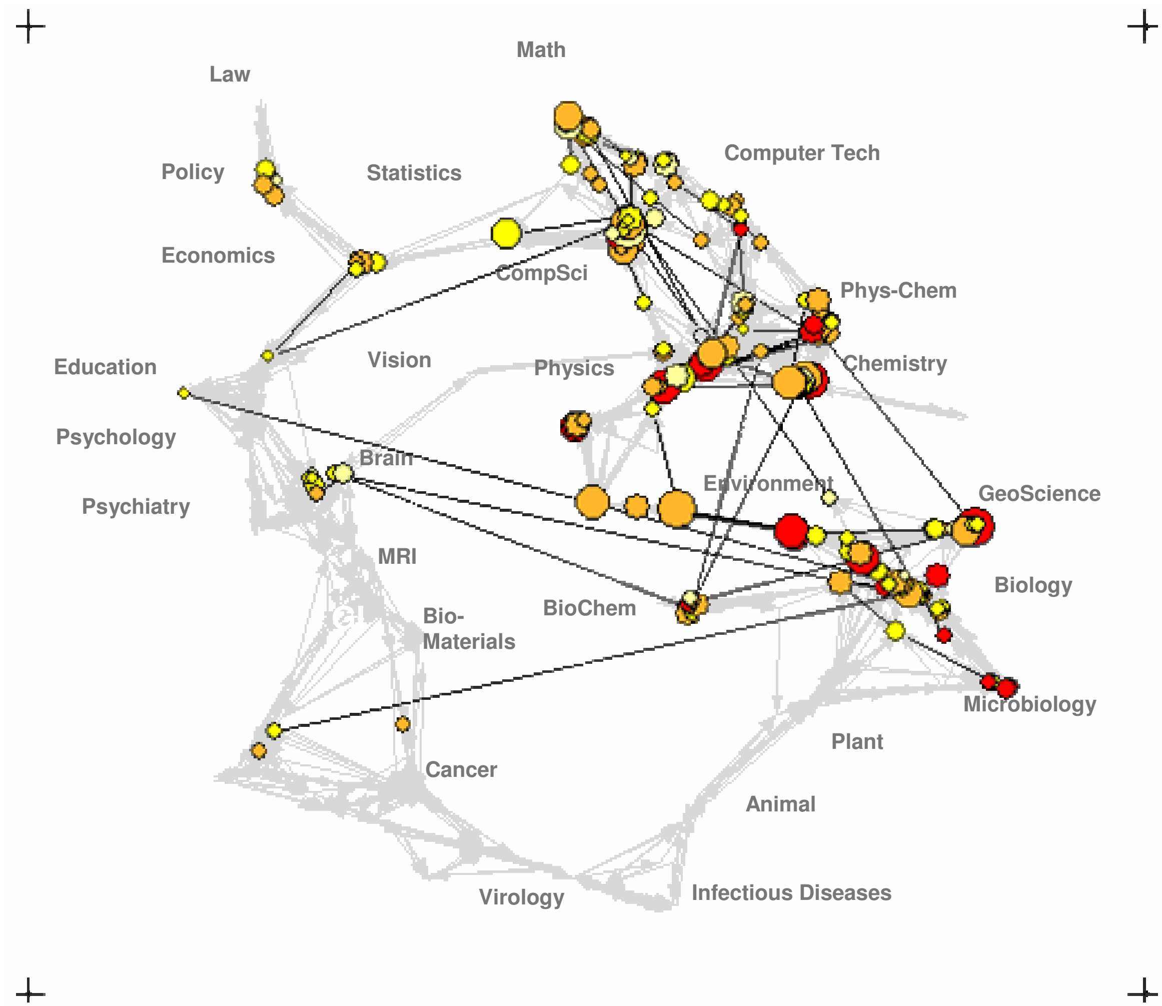
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

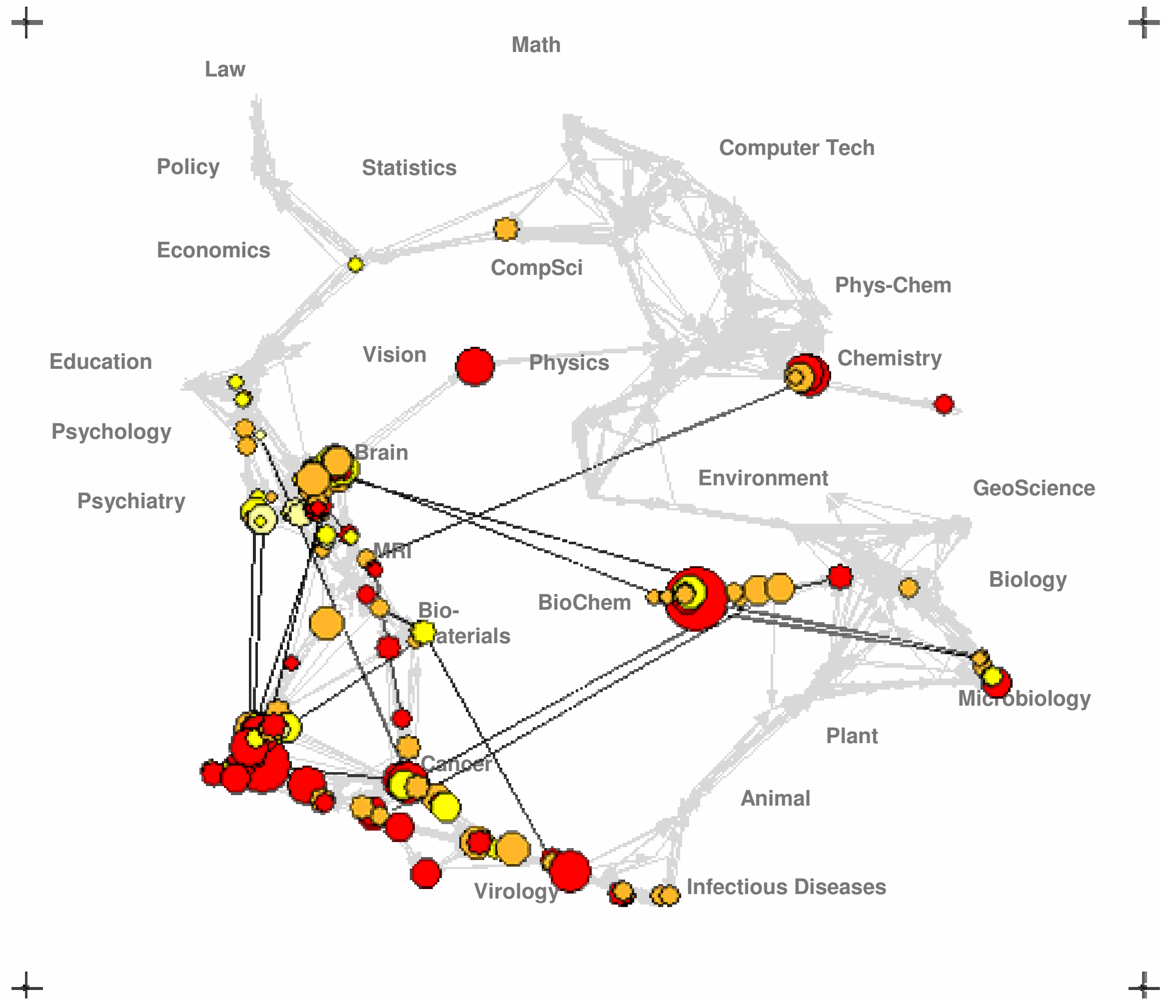
Funding Patterns of the National Science Foundation (NSF)



Science map applications: Identifying core competency

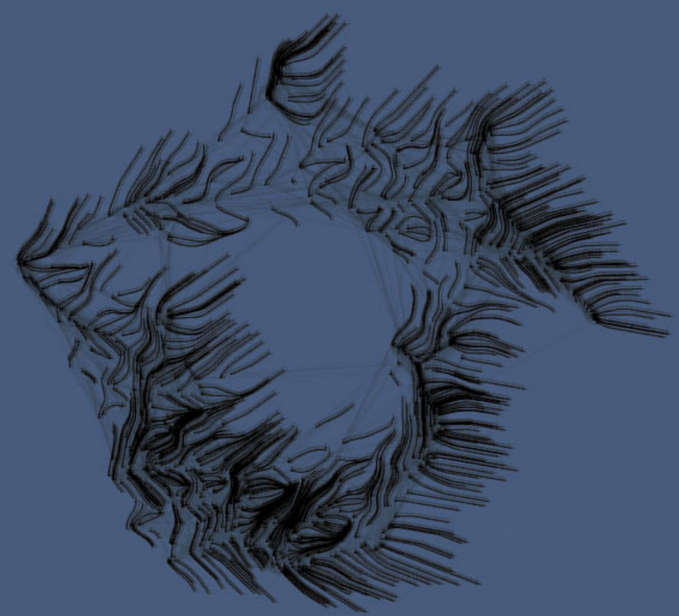
Kevin W. Boyack & Richard Klavans, unpublished work.

Funding Patterns of the National Institutes of Health (NIH)

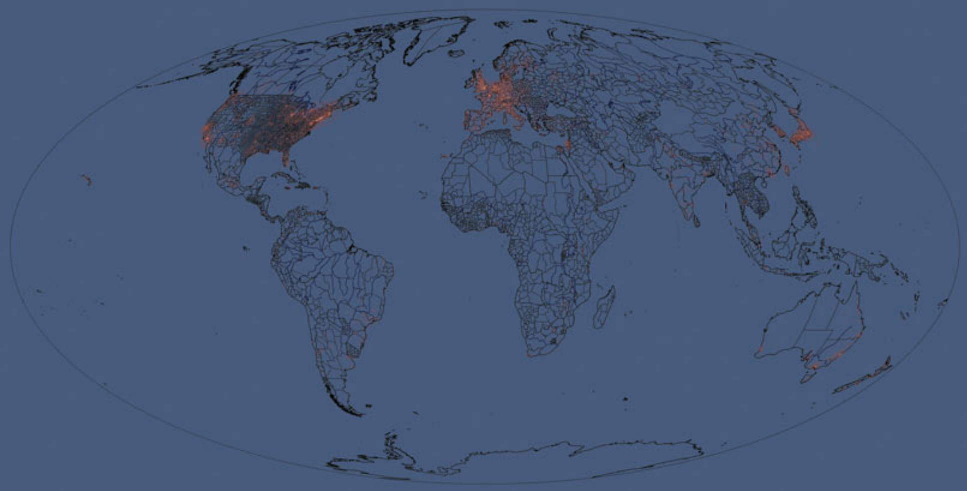


"Places & Spaces: Mapping Science"
on display at the NYPL Science, Industry, and Business Library
Madison/34th, New York City
April 3rd - August 31st, 2006.

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

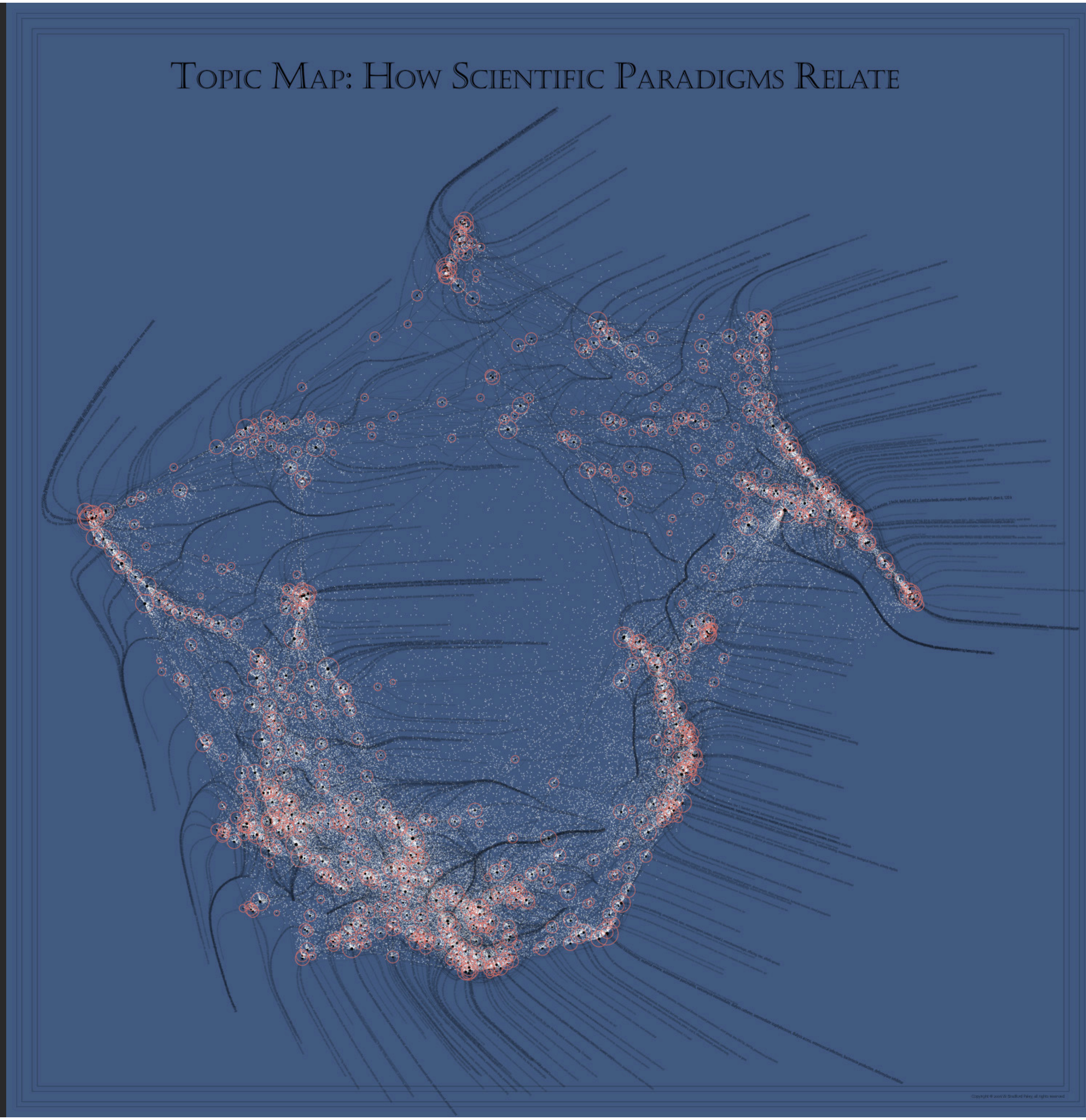
Joshua LEDERBERG
Pioneer in bacterial genetic mechanisms

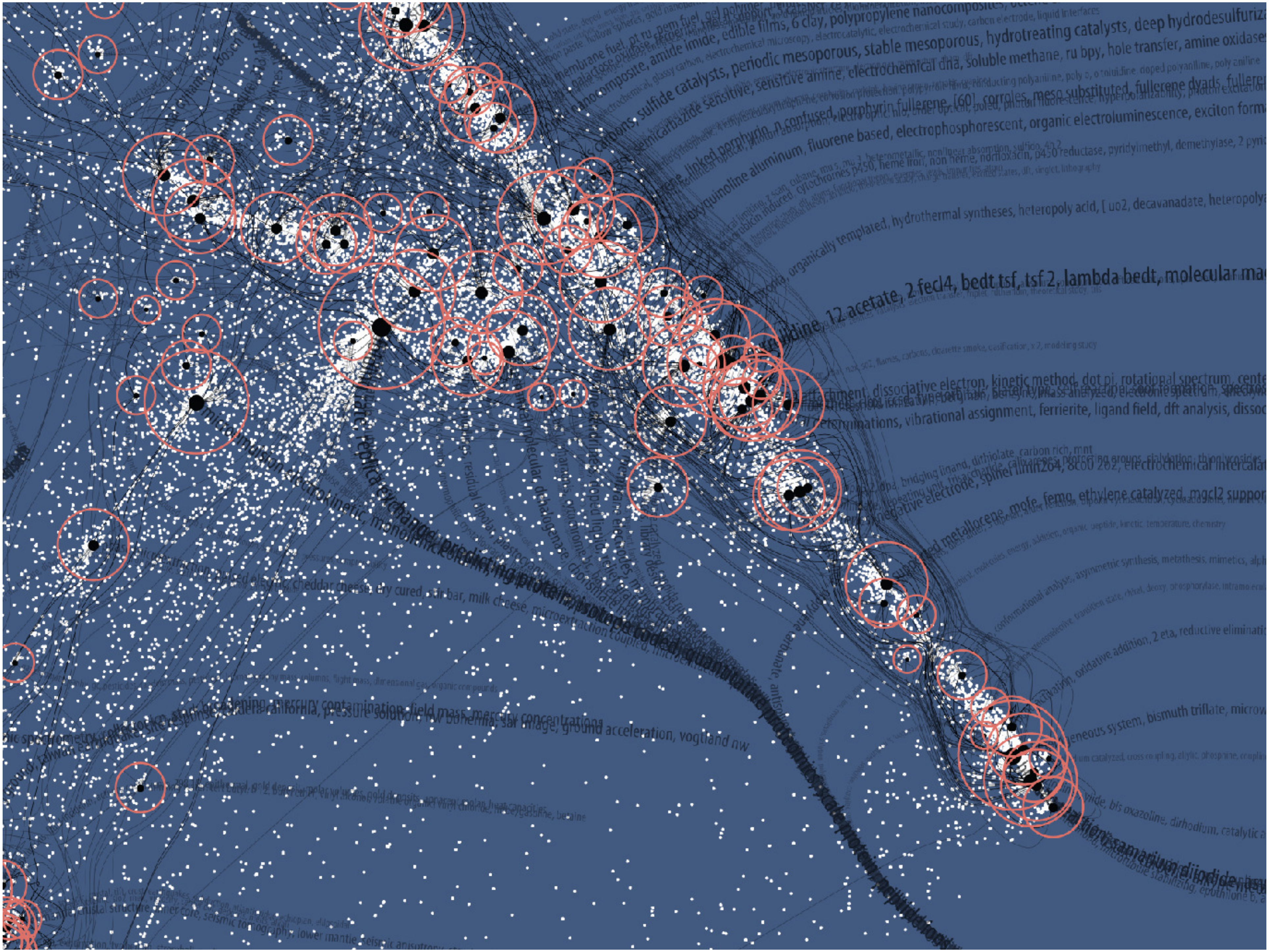
Derek J. de Solla PRICE
Known as the "Father of Scientometrics"

Richard N. ZARE
Uses laser chemistry in molecular dynamics

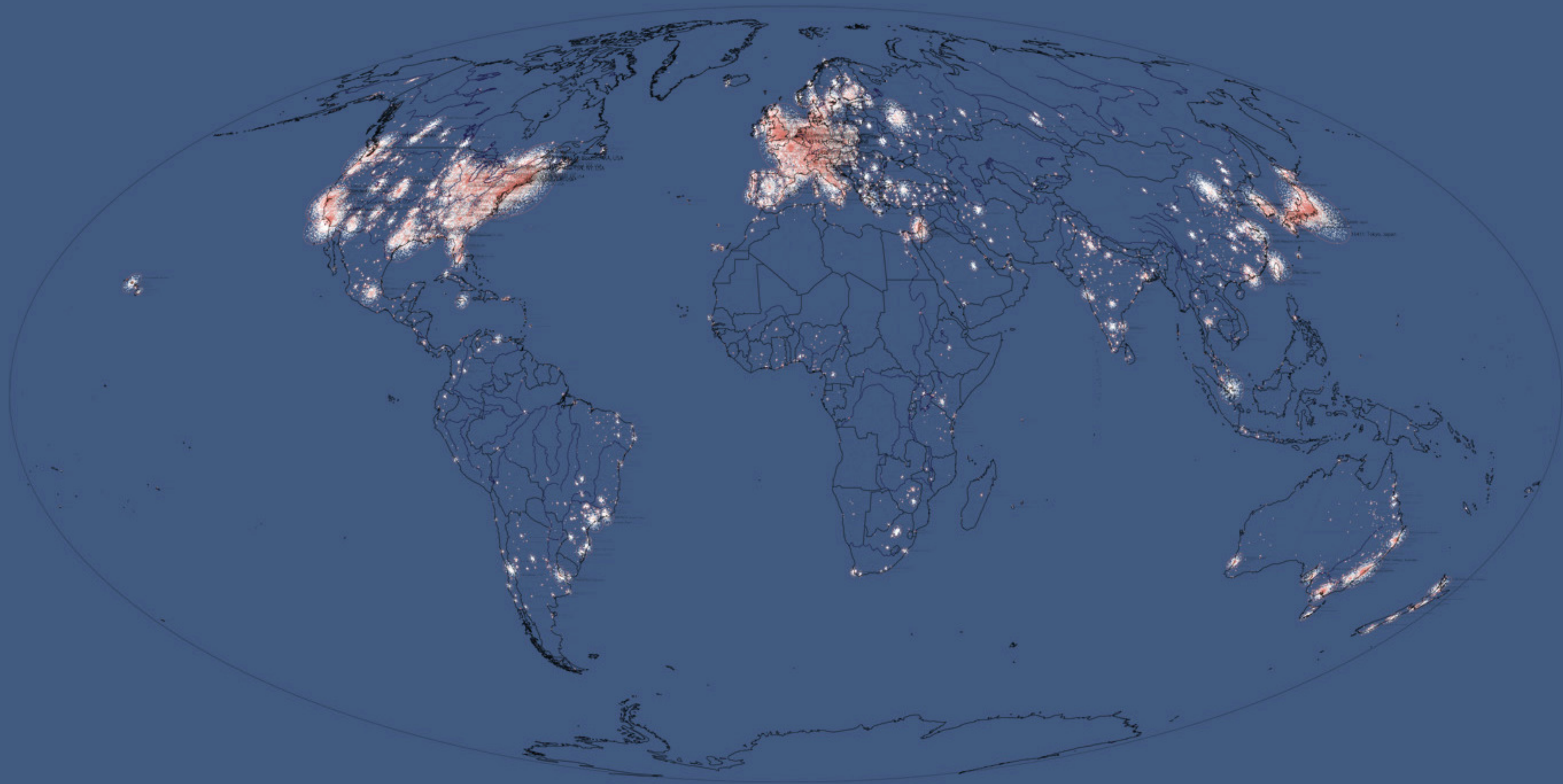
About this display
People & organizations that helped create it

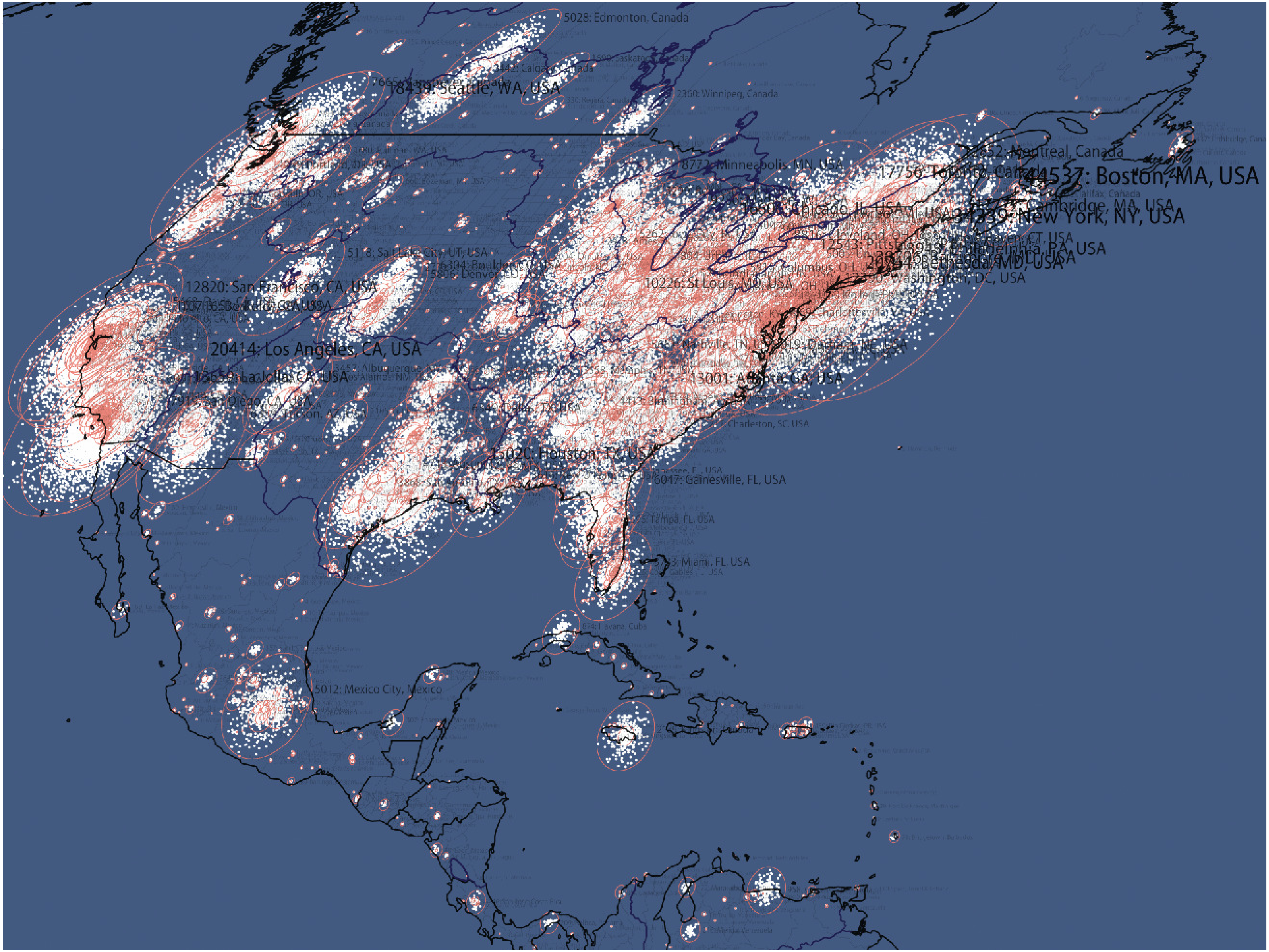
A single person's spreading influence is shown as a series of four snapshots. First, we light only topics and places relating to that person's papers—papers that are still highly cited today. The second lights everything that cites that original work. Note that this first-generation impact extends to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


Topic Map: How Scientific Paradigms Relate

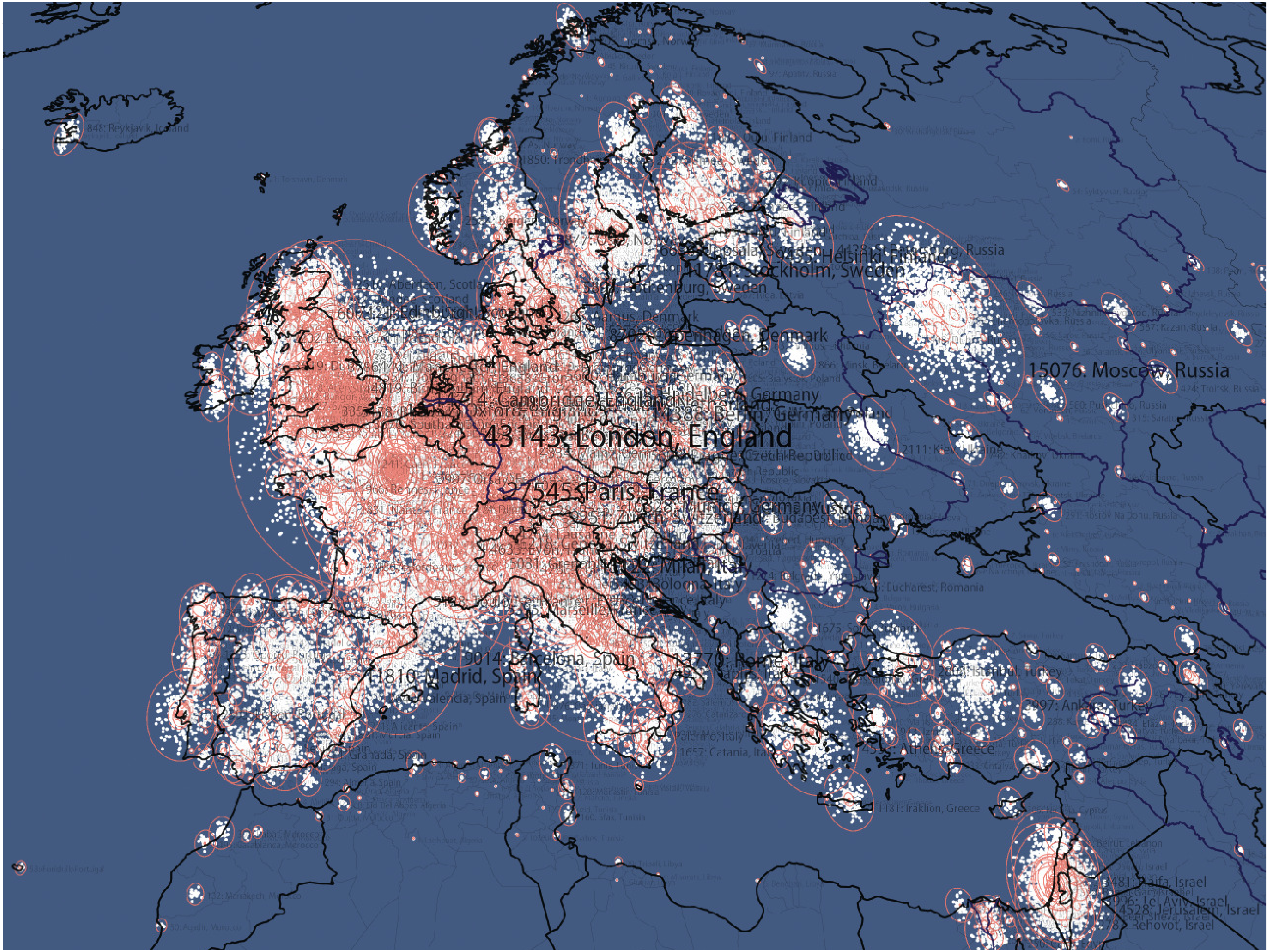

periodic mesoporous, stable mesoporous, hydrotreating catalysts, deep hydrodesulfuriz
soluble methane, ru bpy, hole transfer, amine oxidase
quinoline aluminum, fluorene based, electrophosphorescent, organic electroluminescence, exciton form
organically templated, hydrothermal syntheses, heteropoly acid, [uo2, decavanadate, heteropoly
12 acetate, 2 fecl4, bedt tsf, tsf 2, lambda bedt, molecular ma
dissociative electron, kinetic method, dot pi, rotational spectrum
vibrational assignment, ferrierite, ligand field, dft analysis
metallocene, mofe, femo, ethylene catalyzed, mgcl2 suppor
asymmetric synthesis, metathesis, mimetics
oxidative addition, 2 eta, reductive eliminati
bismuth triflate, microw
bis oxazoline, dirhodium, catalytic
cheddar cheese, dry cured
milk cheese, microextraction coupled
field mass, mercury concentrations
sar image, ground acceleration, vogtland nw
seismic tomography, lower mantle

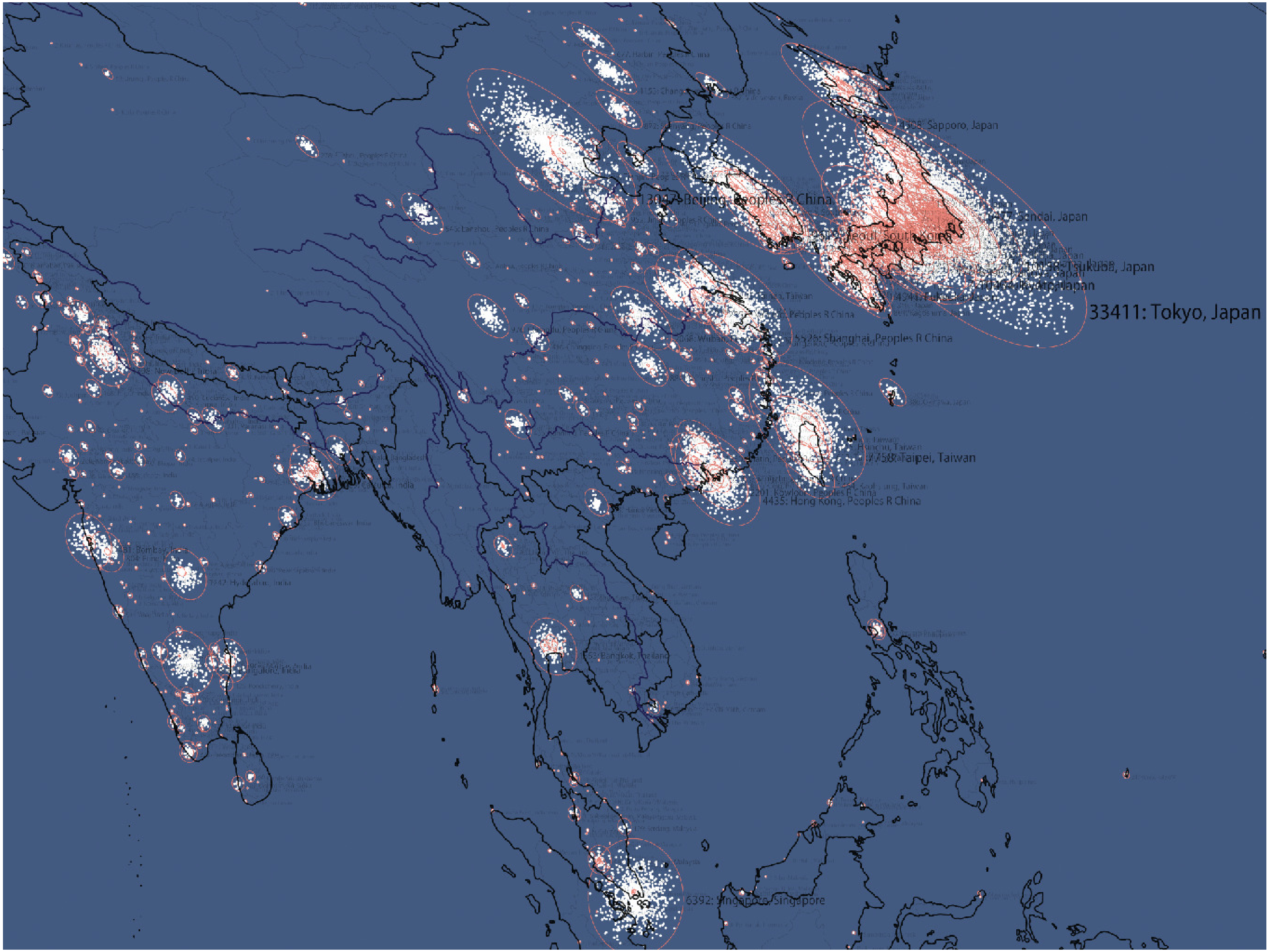
Geographic Map: Where Science Gets Done





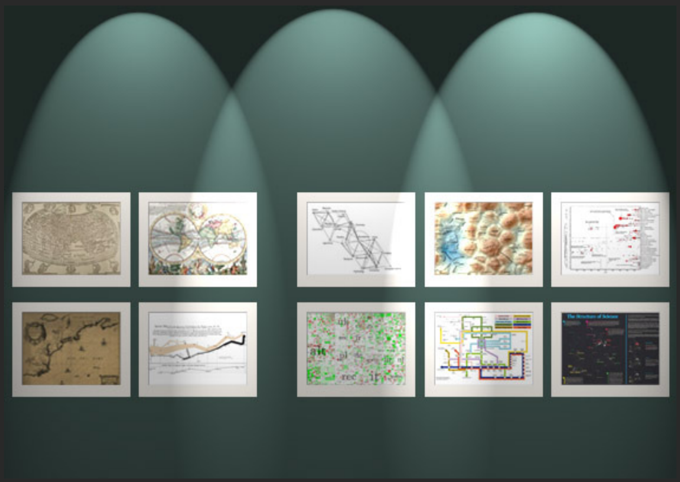
5028: Edmonton, Canada
18439: Seattle, WA, USA
2300: Winnipeg, Canada
Montreal, Canada
44537: Boston, MA, USA
New York, NY, USA
5118: Salt Lake City, UT, USA
12820: San Francisco, CA, USA
10226: St Louis, MO, USA
Washington, DC, USA
20414: Los Angeles, CA, USA
Charleston, SC, USA
15020: Houston, TX, USA
6017: Gainesville, FL, USA
5012: Mexico City, Mexico


848: Reykjavik, Iceland
43143: London, England
27545: Paris, France
15076: Moscow, Russia
11810: Madrid, Spain
9014: Barcelona, Spain
1657: Catania, Italy
1181: Iraklion, Greece
Bucharest, Romania
2097: Ankara, Turkey


33411: Tokyo, Japan
Sapporo, Japan
Sendai, Japan
Tsukuba, Japan
13047: Beijing, Peoples R China
Seoul, South Korea
Shanghai, Peoples R China
Taipei, Taiwan
4435: Hong Kong, Peoples R China
Kowloon, Peoples R China
Bangkok, Thailand
6392: Singapore, Singapore
Hyderabad, India
Bombay, India
New Delhi, India
Bangalore, India

The Power of Maps

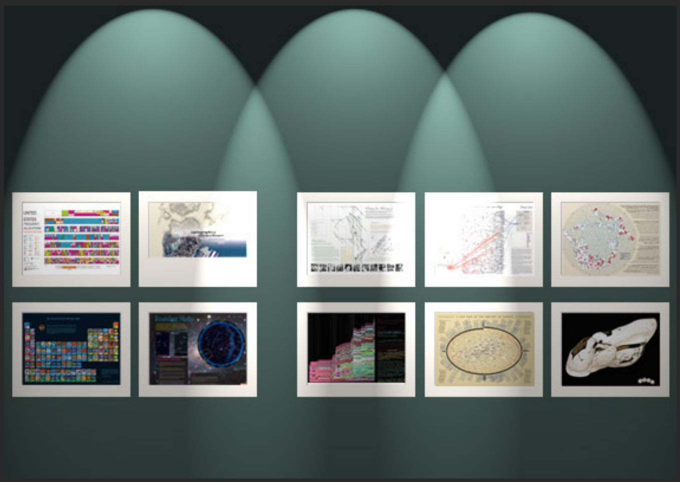
Four Early Maps of Our World
VERSUS
Six Early Maps of Science



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

The Power of Reference Systems

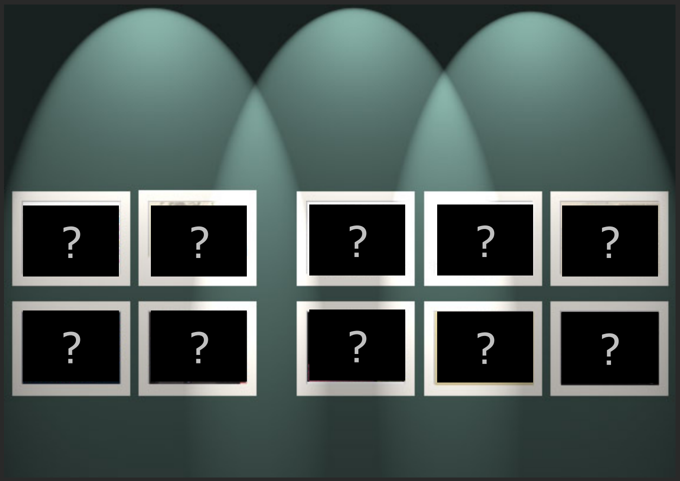
Four Existing Reference Systems
VERSUS
Six Potential Reference Systems of Science



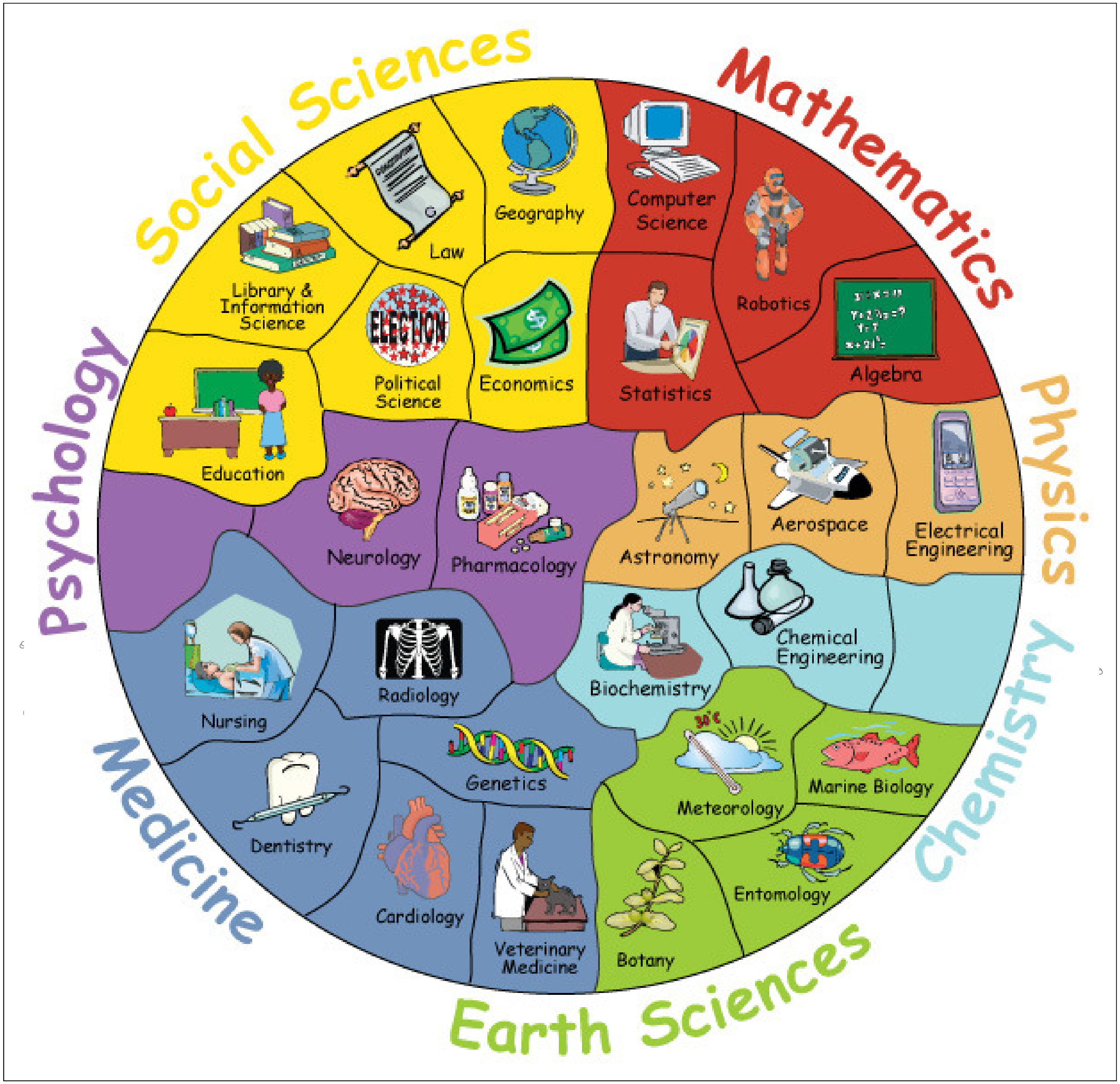
(2nd Iteration of Places & Spaces Exhibit - 2006)

The Power of Forecasts

Four Existing Forecasts
VERSUS
Six Potential Science 'Weather' Forecasts

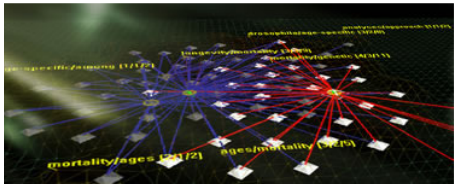


(3rd Iteration of Places & Spaces Exhibit - 2007)



Social Sciences
Mathematics
Physics
Chemistry
Earth Sciences
Medicine
Psychology
Library & Information Science
Law
Geography
Computer Science
Robotics
Political Science
Economics
Statistics
Algebra
Education
Neurology
Pharmacology
Astronomy
Aerospace
Electrical Engineering
Biochemistry
Chemical Engineering
Nursing
Radiology
Genetics
Meteorology
Marine Biology
Dentistry
Cardiology
Veterinary Medicine
Entomology
Botany

Morning Meeting
A Snowy Thursday
Trip to Art Museum
Lunch and Recess
12\8\05
54
SOCCER



Science Studies: 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.

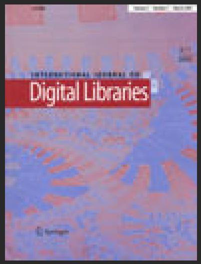
This Talk has Three Parts:

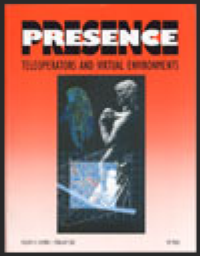
1. Why study the structure and evolution of science?
2. **What infrastructure is needed to study science?**

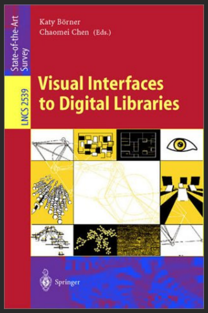
3. Cyberinfrastructures under development:
   CIShell, IVC, and NWB

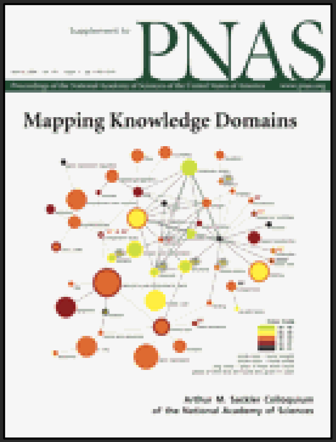
Related Work

Analyzing, Modeling, and Mapping Science

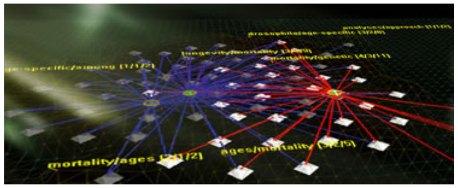








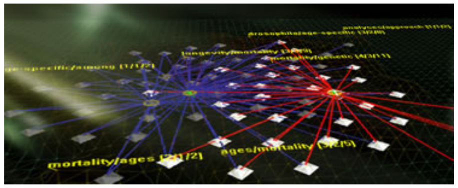
- *Shiffrin, Richard M. and Börner, Katy (Eds.) (2004).* ***Mapping Knowledge Domains.*** *Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl_1).*
- *Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003).* ***Visualizing Knowledge Domains.*** *In Blaise Cronin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



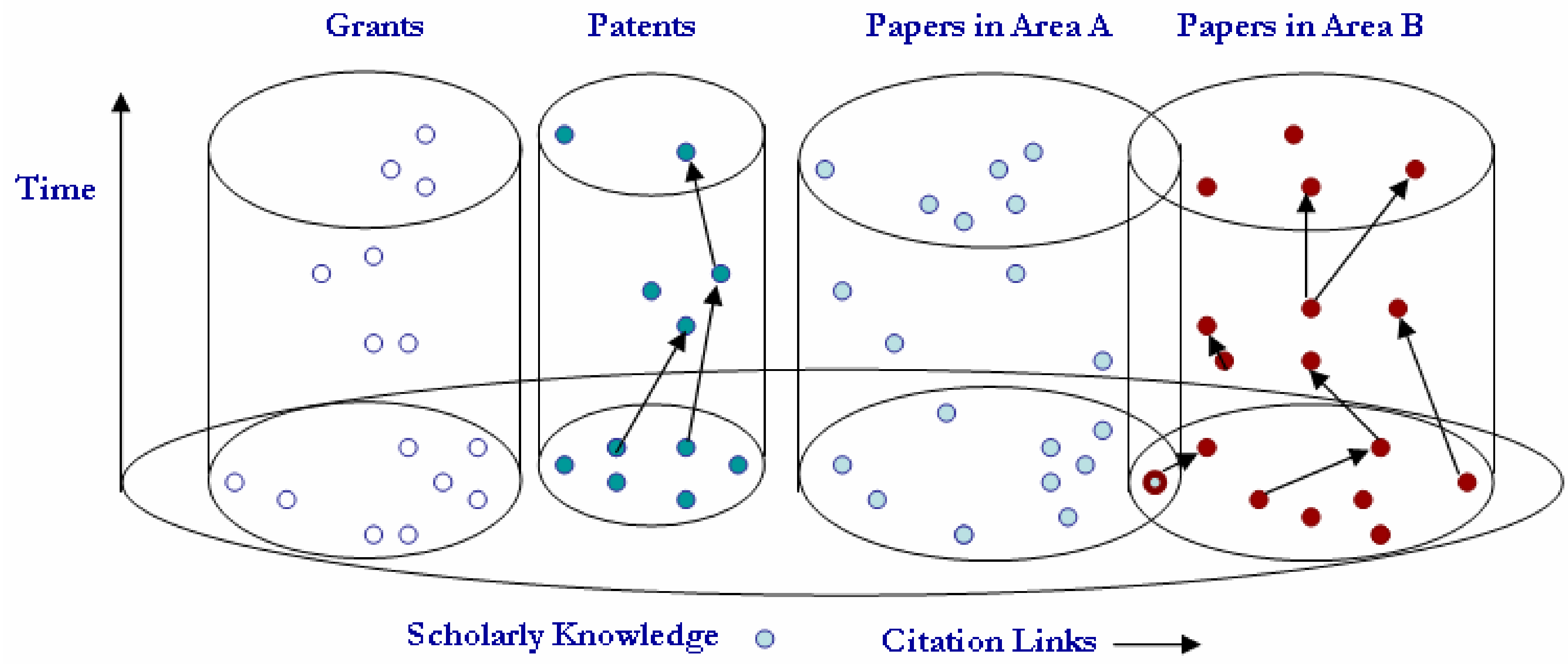
Process of Analyzing and Mapping Science

DATA EXTRACTION	UNIT OF ANALYSIS	MEASURES	LAYOUT (often one code does both similarity and ordination steps)		DISPLAY
			SIMILARITY	ORDINATION	
SEARCHES ISI INSPEC Eng Index Medline ResearchIndex Patents etc. 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

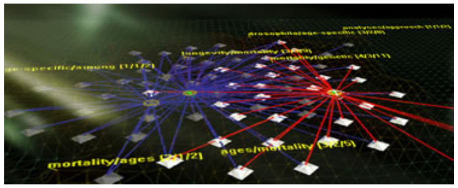
Börner, Chen & Boyack.. (2003) Visualizing Knowledge Domains. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Volume 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.



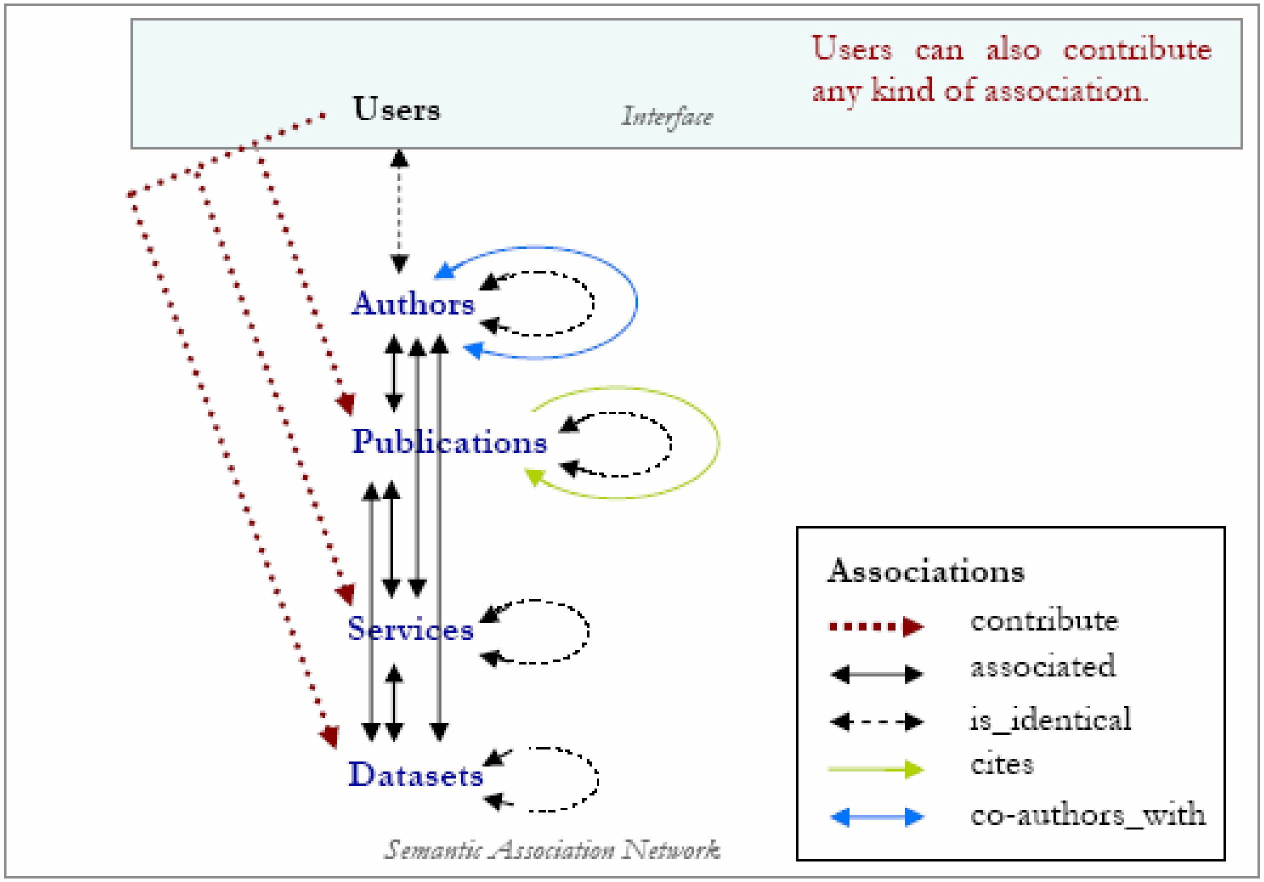
Data – Solve Interoperability Problem



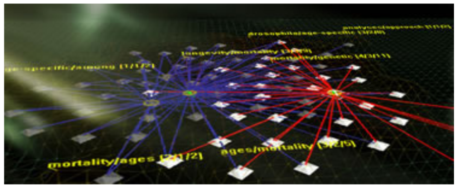
By drawing on existing efforts and by coupling automatic data integration with manual Wiki* approaches.



Data – Need Highest Quality, Coverage & Interlinkage



Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) Visualizing the Semantic Web, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.



Algorithms

Problem

There are too many different data models and data formats, different algorithms, different implementations of the same algorithm, different programming languages, different research purposes (modeling, analysis, visualization), different communities and practices.

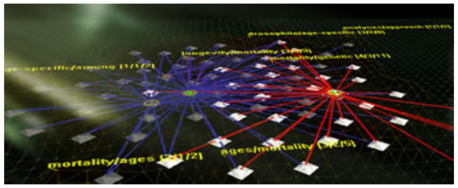
Algorithm Developers

- Are often non-computer scientists.
- Have no means to make their code available.
- Want to concentrate on developing algorithms.

Algorithms Users

- Are researchers, industry, classroom teachers, etc.
- Often have no programming or scripting skills.
- Want to concentrate on science, product development, education.

Needed is a socio-technical cyberinfrastructure that supports the
free distribution and sharing of datasets and algorithms, their descriptions and associated learning modules.



Cyberinfrastructure – Desirable Features

General

- *Extensibility:* Easily add new algorithms and data models to the framework over time.
- *Scalability:* Integrate many algorithms & process large datasets.
- *Support multiple operating systems*, e.g., Windows, Linux, Solaris, Mac OS.

Developer Specific

- *Ease of Use:* Support easy integration of many different algorithms and datasets.
- *Flexibility:* Support multiple data formats and transformation among data models. This requires a seamless exchange of data and parameters. Support multiple programming languages.

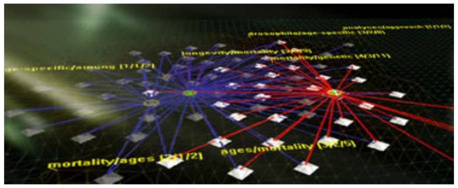
User Specific

- *Ease of Use:* Provide easy access to most popular data models & standards, the most efficient algorithm implementations, work log tracking, and scheduling. User should be able to 'fill' the CI/tool with exactly the algorithms and datasets s/he needs.
- *Adaptability*: Provide different UI solutions -- menu driven, script based, customized.

This Talk has Three Parts:

1. Why study the structure and evolution of science?
2. What infrastructure is needed to study science?

3. **Cyberinfrastructures under development: CIShell, IVC, and NWB**



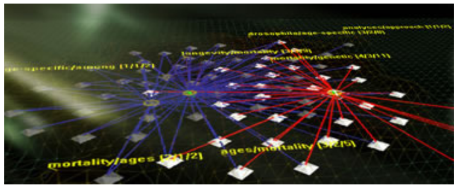
Cyberinfrastructure Shell (CIShell)

CIShell is an open source, community-driven specification for the integration and utilization of datasets, algorithms, tools, and computing resources that aims to serve the needs of three user groups:

Specification, API, and related documentation are available at http://cishell.org.

Specification and all reference implementations are open sourced under the Apache 2.0 license.

Bruce Herr, Weixia Huang, Shashikant Penumarthy, Katy Börner. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. Submitted to William S. Bainbridge (Ed.) Progress in Convergence. Annals of the New York Academy of Sciences.



CIShell – Technical Details

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

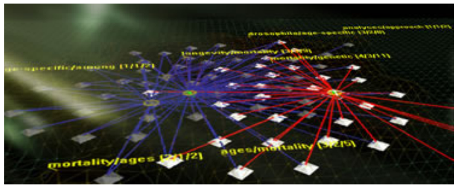
OSGi (http://www.osgi.org) is

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

Advantages of Using OSGi

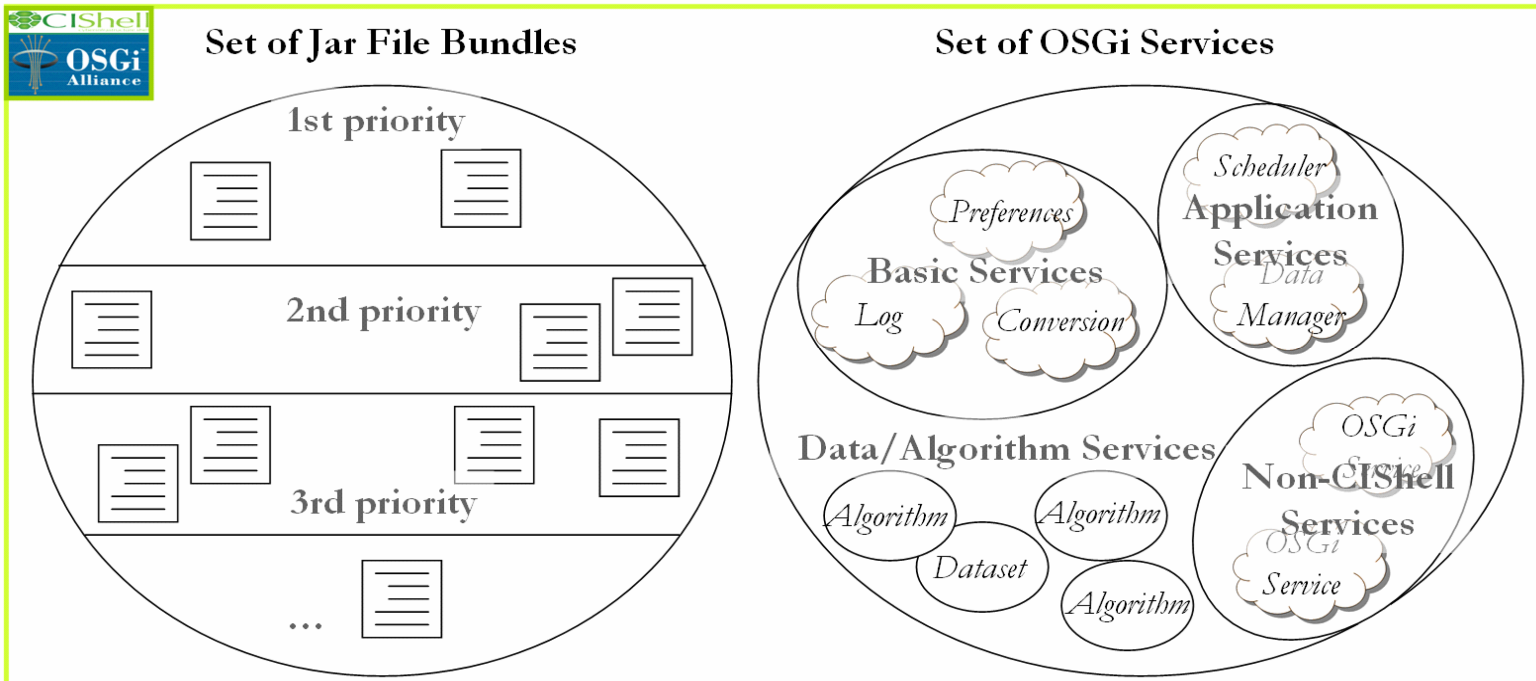
- Any CIShell algorithm is a service that can be used in any OSGi-framework based system.
- Using OSGi, running CIShells/tools can connected via RPC/RMI supporting peer-to-peer sharing of data, algorithms, and computing power.

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

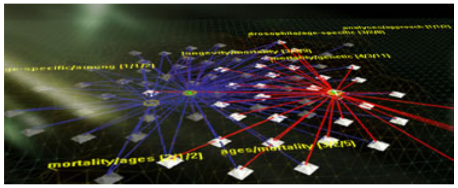


CIShell – Technical Details

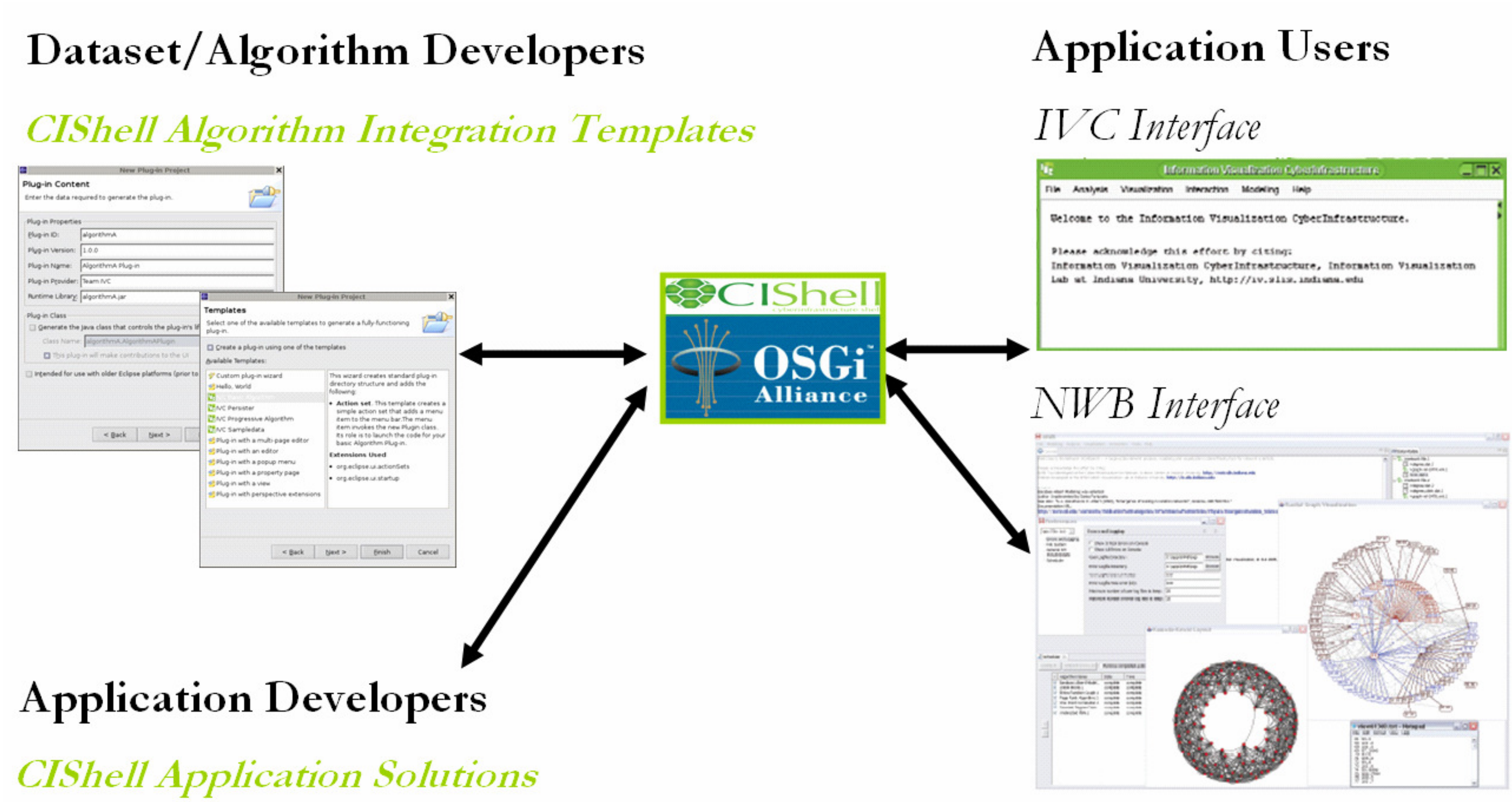
CIShell supports the design of highly modular and decentralized system architectures comprising a set of OSGi bundles (left) that upon start up instantiate a set of OSGI services (right).

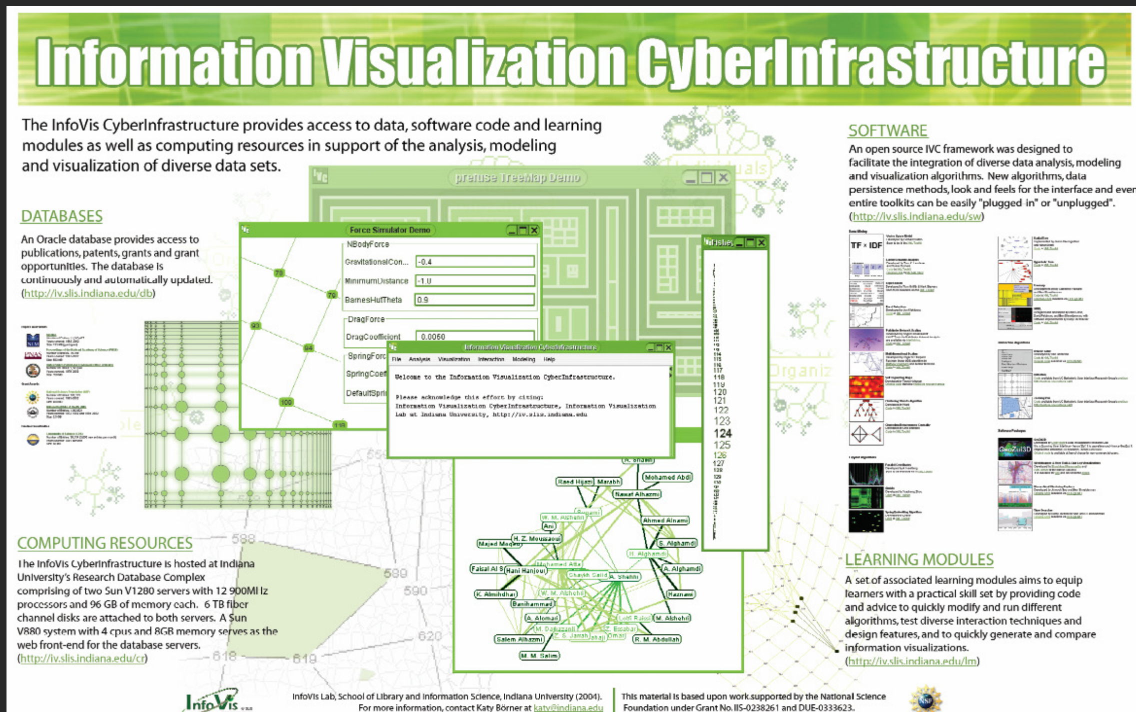


Multiple algorithm services can be registered from one bundle but each algorithm has exactly only one associated OSGi service.



CIShell – Support for Developers and Users

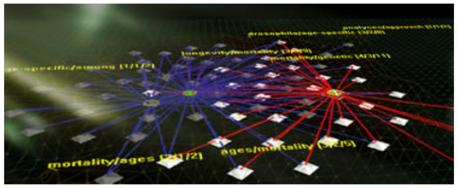




CAREER: Visualizing Knowledge Domains. NSF IIS-0238261 award (Katy Börner, $451,000) Sept. 03-Aug. 08.
http://iv.slis.indiana.edu/

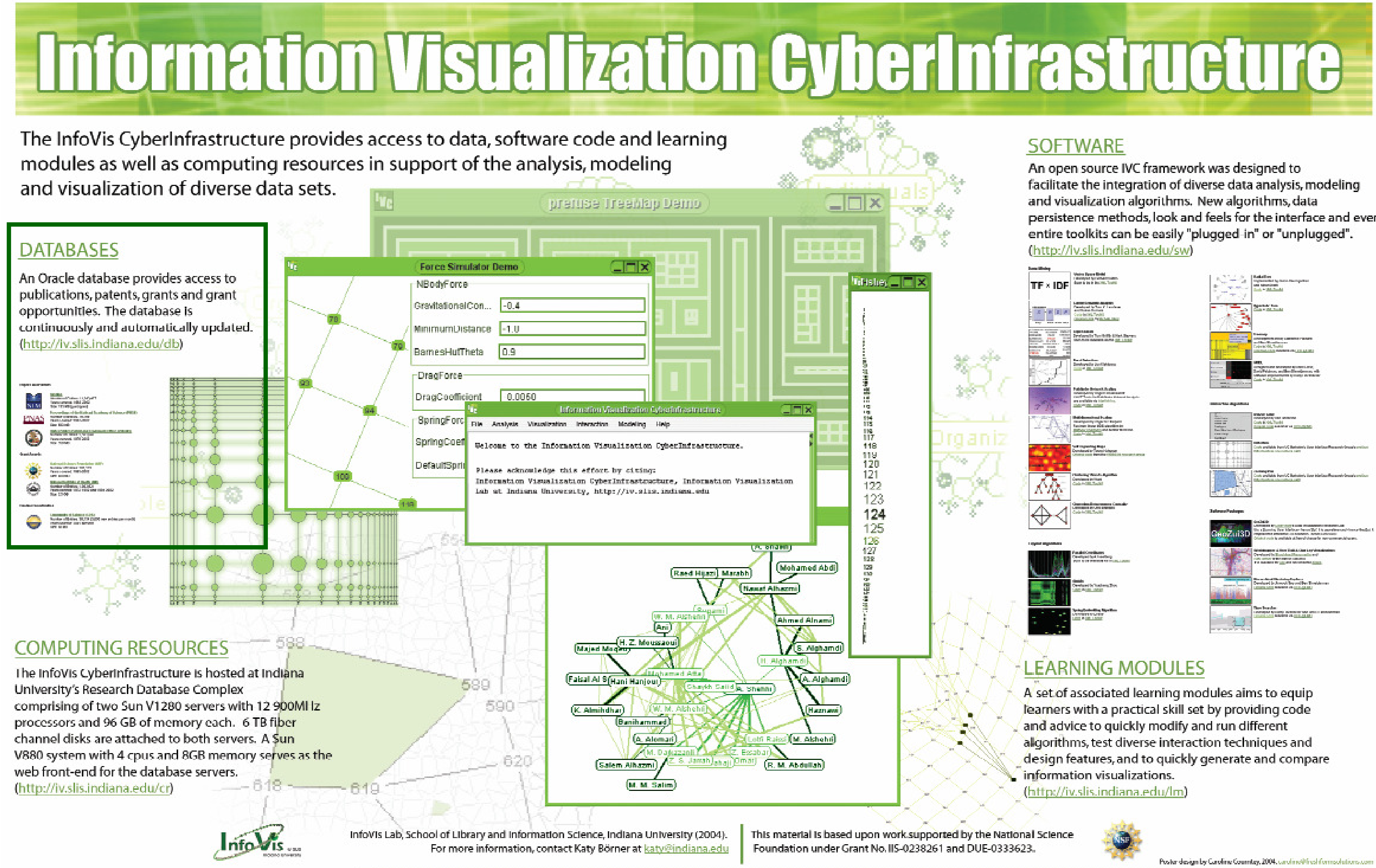
SEI: Network Workbench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. NSF IIS-0513650 award (Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert (Senior Personnel), $1,120,926)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

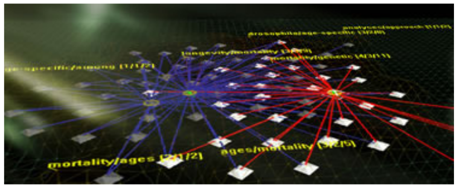
NSF SBC Sun microsystems
HPNAP High Performance Network Applications Program
INDIANA 21st Century Research & Technology Fund



InfoVis Cyberinfrastructure

http://iv.slis.indiana.edu





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

Papers and Patents

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

Proceedings of the 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

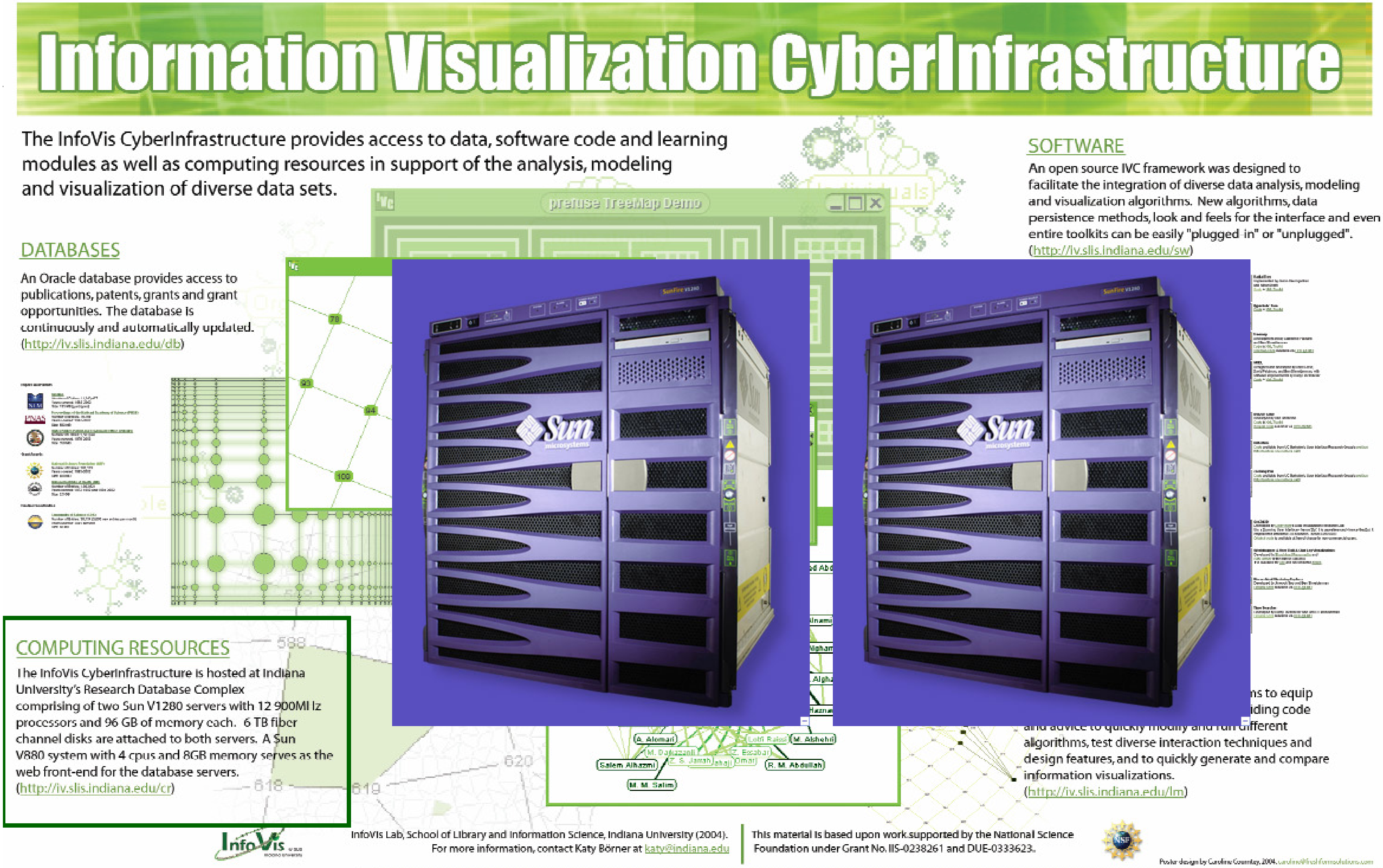
Grant Awards

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

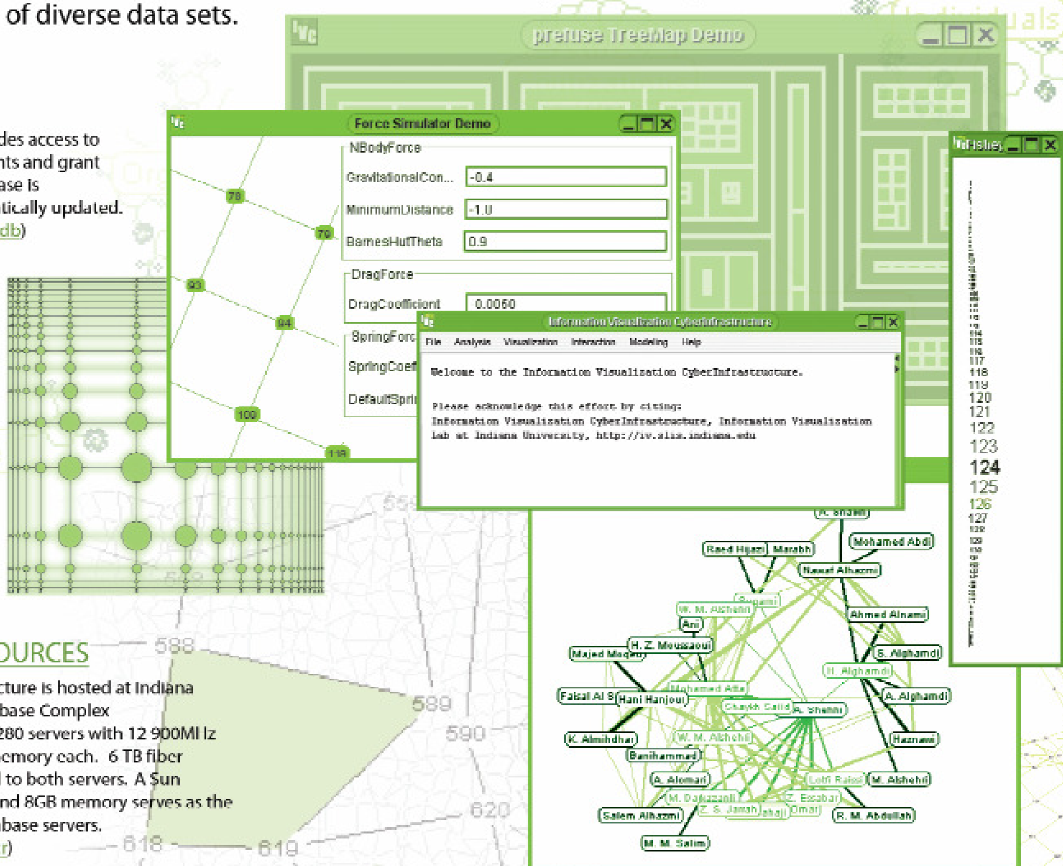

Information Visualization CyberInfrastructure
The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.
DATABASES
An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (http://iv.slis.indiana.edu/db)
SOFTWARE
An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged-in" or "unplugged". (http://iv.slis.indiana.edu/sw)
prefuse TreeMap Demo
COMPUTING RESOURCES
The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (http://iv.slis.indiana.edu/cr)
Sun microsystems
Sun microsystems
ns to equip
iding code
algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations.
(http://iv.slis.indiana.edu/lm)
InfoVis
InfoVis Lab, School of Library and Information Science, Indiana University (2004).
For more information, contact Katy Börner at katy@indiana.edu
This material is based upon work supported by the National Science Foundation under Grant No. IIS-0238261 and DUE-0333623.
NSF

Information Visualization CyberInfrastructure

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

DATABASES

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

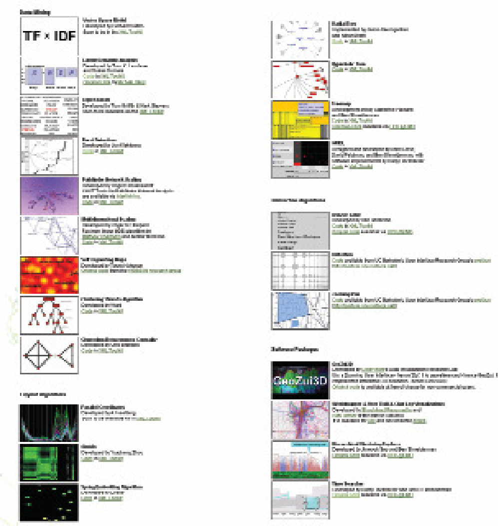


COMPUTING RESOURCES

The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (http://iv.slis.indiana.edu/cr)

SOFTWARE

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



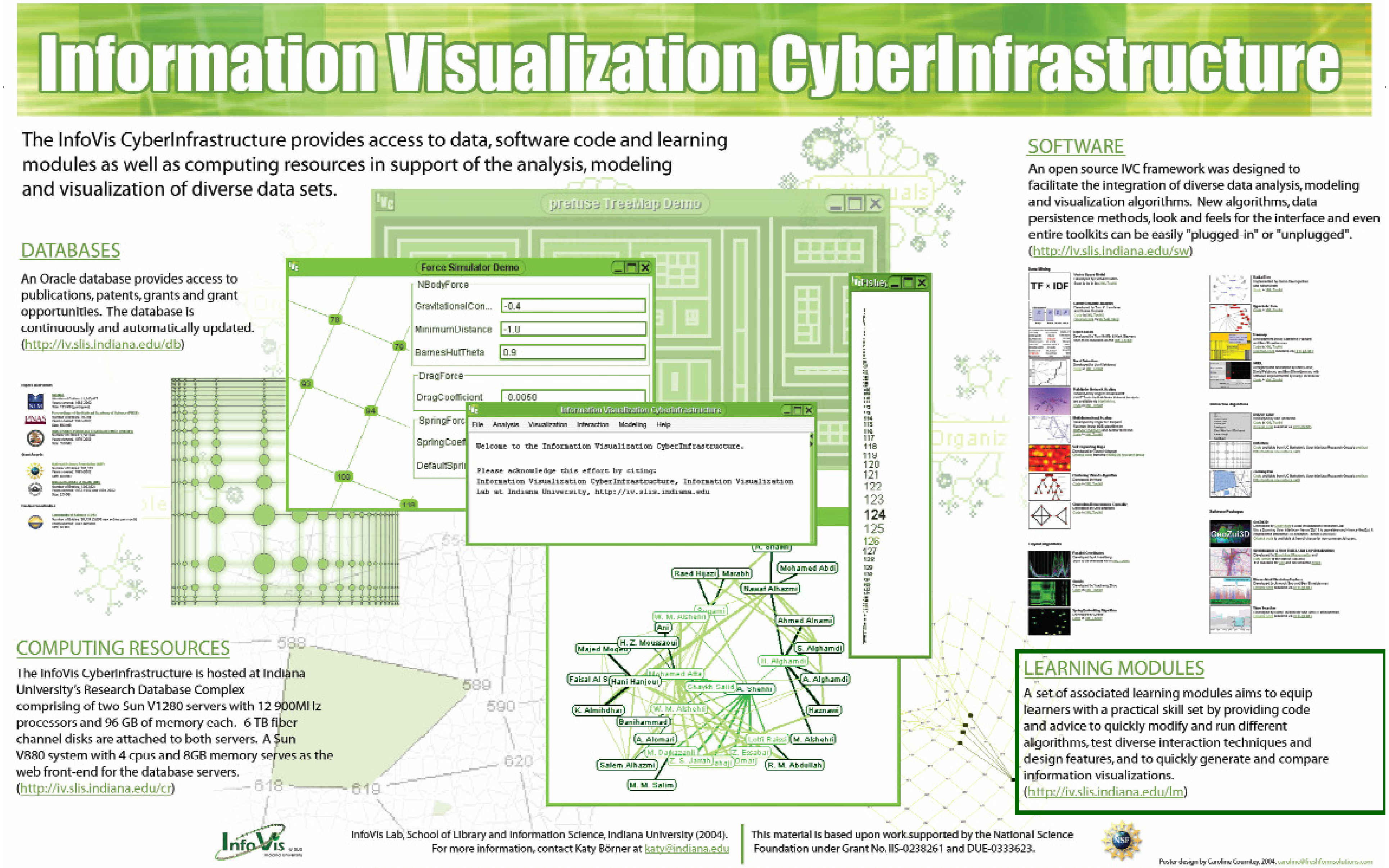
LEARNING MODULES

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

InfoVis Lab, School of Library and Information Science, Indiana University (2004). For more information, contact Katy Börner at katy@indiana.edu

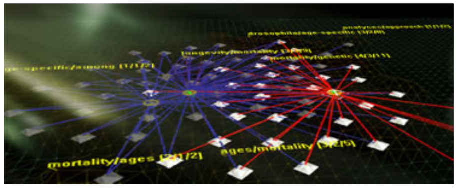
This material is based upon work supported by the National Science Foundation under Grant No. IIS-0238261 and DUE-0333623.

Poster design by Caroline Courtney, 2004.



Information Visualization CyberInfrastructure
The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.
DATABASES
An Oracle database provides access to publications, patents, grants and grant opportunities. The database is continuously and automatically updated. (http://iv.slis.indiana.edu/db)
COMPUTING RESOURCES
The InfoVis CyberInfrastructure is hosted at Indiana University's Research Database Complex comprising of two Sun V1280 servers with 12 900MHz processors and 96 GB of memory each. 6 TB fiber channel disks are attached to both servers. A Sun V880 system with 4 cpus and 8GB memory serves as the web front-end for the database servers. (http://iv.slis.indiana.edu/cr)
SOFTWARE
An open source IVC framework was designed to facilitate the integration of diverse data analysis, modeling and visualization algorithms. New algorithms, data persistence methods, look and feels for the interface and even entire toolkits can be easily "plugged-in" or "unplugged". (http://iv.slis.indiana.edu/sw)
LEARNING MODULES
A set of associated learning modules aims to equip learners with a practical skill set by providing code and advice to quickly modify and run different algorithms, test diverse interaction techniques and design features, and to quickly generate and compare information visualizations. (http://iv.slis.indiana.edu/lm)
prefuse TreeMap Demo
Force Simulator Demo
Information Visualization CyberInfrastructure
Welcome to the Information Visualization CyberInfrastructure.
Please acknowledge this effort by citing:
Information Visualization CyberInfrastructure, Information Visualization Lab at Indiana University, http://iv.slis.indiana.edu
InfoVis Lab, School of Library and Information Science, Indiana University (2004).
For more information, contact Katy Börner at katy@indiana.edu
This material is based upon work supported by the National Science Foundation under Grant No. IIS-0238261 and DUE-0333623.

IVC Learning Modules



InfoVis Cyberinfrastructure

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

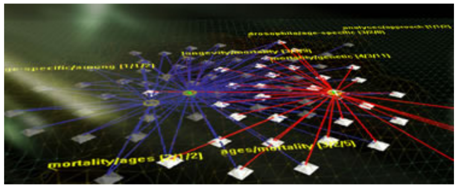
Home | Learning Modules | Software | Data Bases | Compute Resources | References

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:

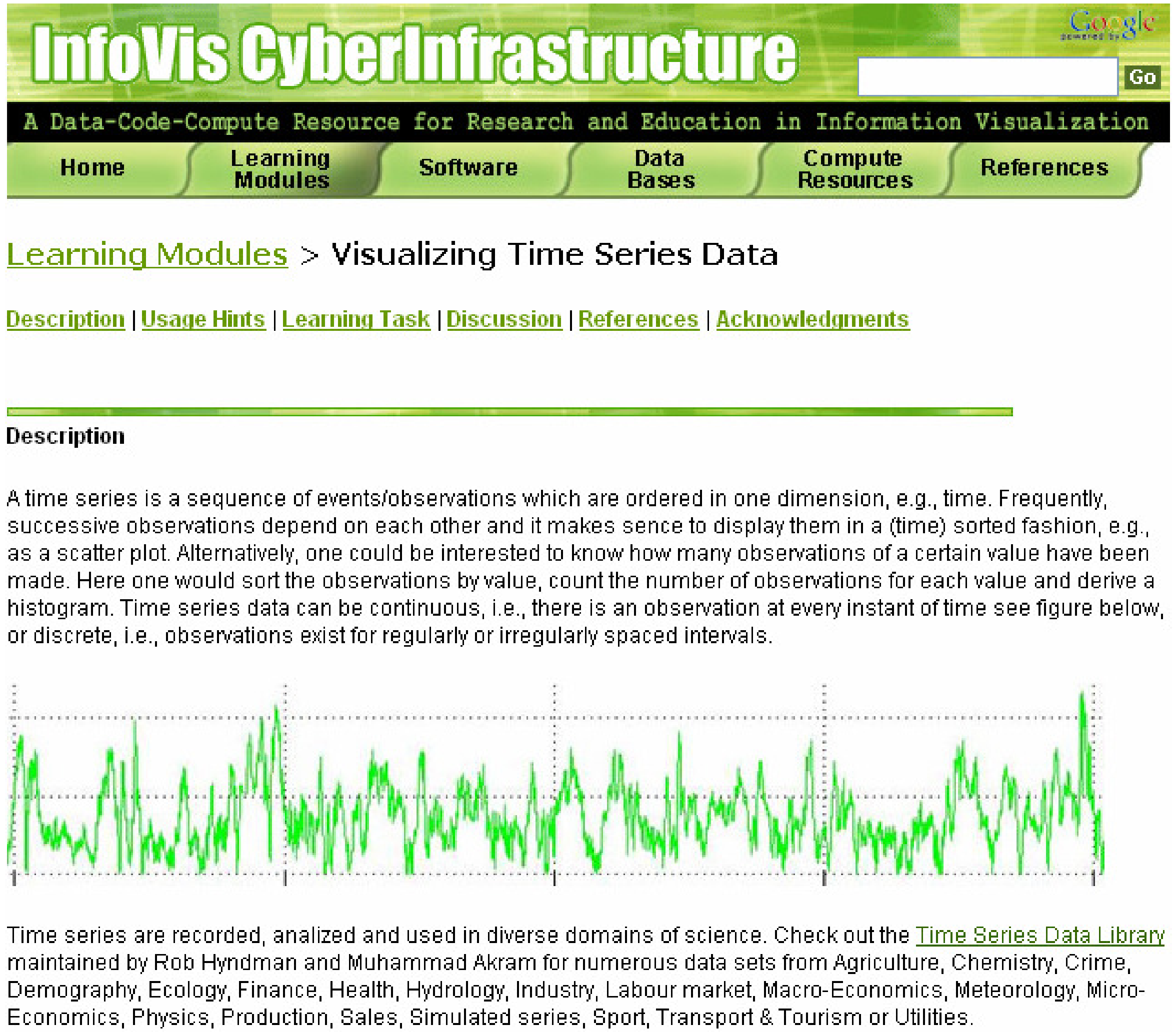
- Description of the data analysis and visualization task
- Usage hints on how to run and use a particular algorithm or tool
- Learning task - a challenging scenario to use an algorithm or to analyze and/or visualize a data set.
- Discussion of the results, and
- References to research papers, online demos, (commercial) applications.
- Acknowledgements



Time Series Analysis

Learning Module

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



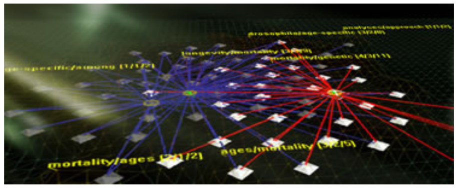
Learning Modules > Visualizing Time Series Data

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

Description

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

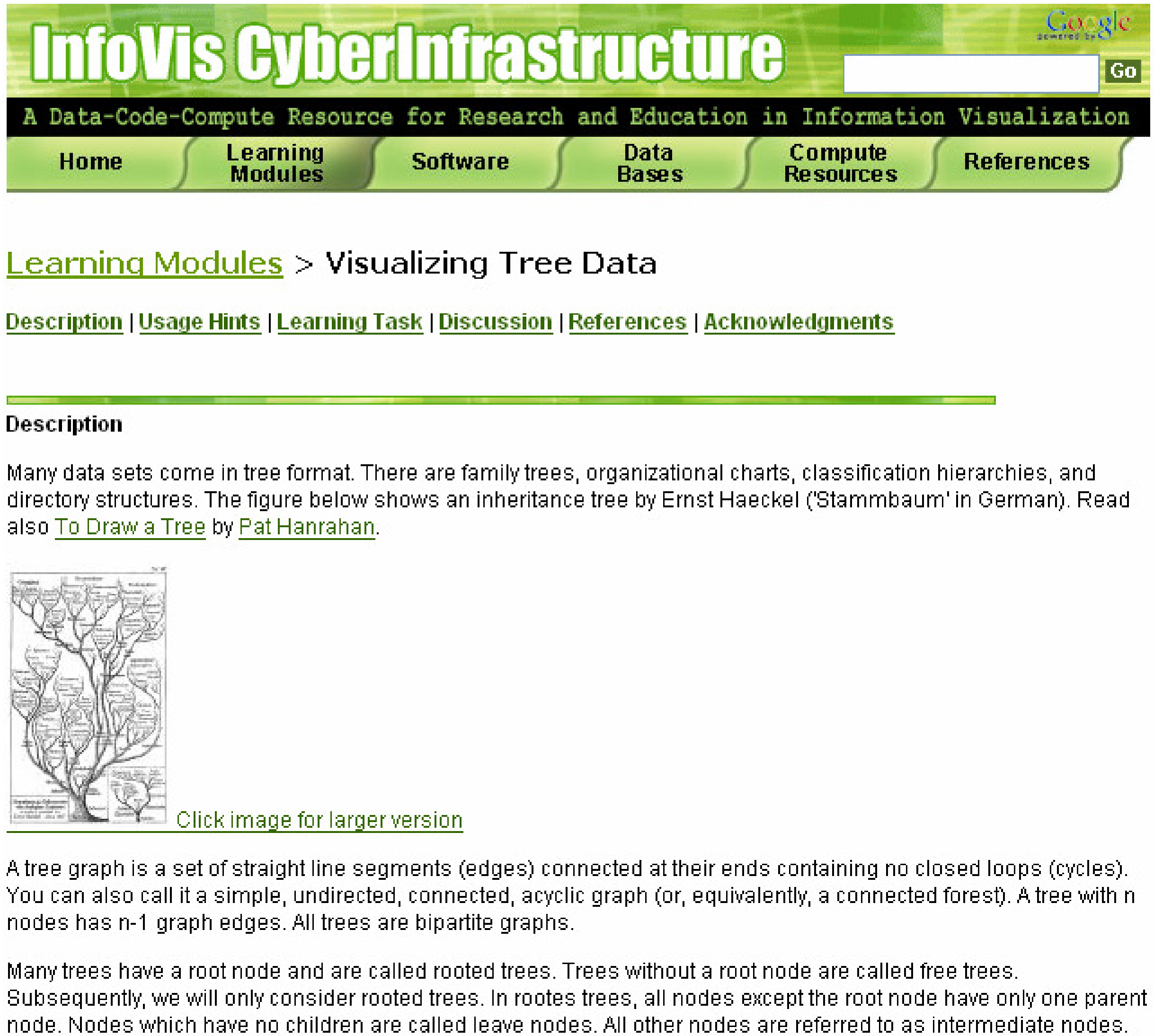
Time series are recorded, analized and used in diverse domains of science. Check out the Time Series Data Library maintained by Rob Hyndman and Muhammad Akram for numerous data sets from Agriculture, Chemistry, Crime, Demography, Ecology, Finance, Health, Hydrology, Industry, Labour market, Macro-Economics,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



Visualizing Tree Data

Learning Module

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



InfoVis CyberInfrastructure

Go

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.

Network Workbench

A Workbench for Network Scientists

MOTIVATION

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

MENU DRIVEN INTERFACE

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

WORK LOG TRACKING MODULE

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

SCHEDULER

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

ACKNOWLEDGMENTS

The NWB cyberinfrastructure is supported in part by the 21st Century Fund and the National Science Foundation under Grants No. IIS-0238261 and IIS-0513650.

PRIMARY INVESTIGATORS

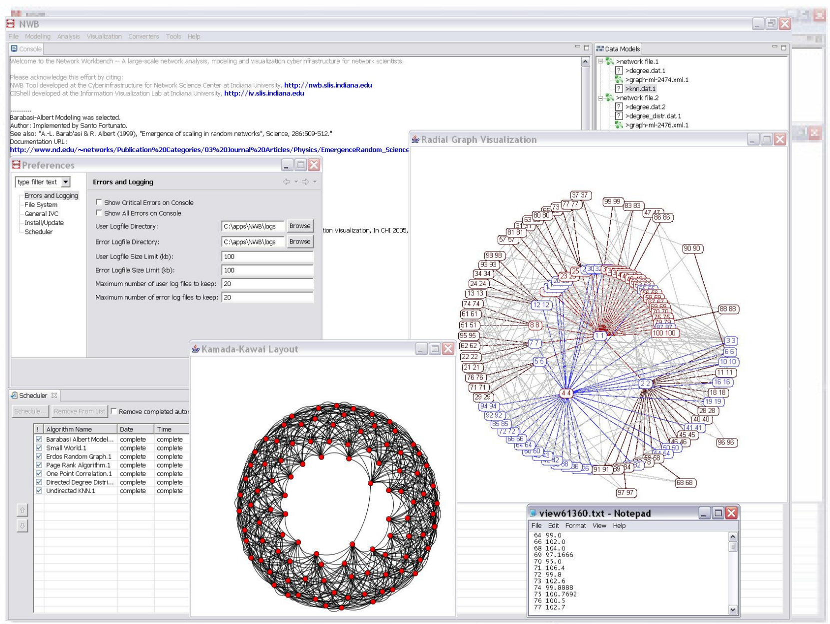
Dr. Katy Börner
Indiana University
Dr. Albert-László Barabási
University of Notre Dame
Dr. Santiago Schnell
Dr. Alessandro Vespignani
Dr. Stanley Wasserman
Dr. Eric A. Wernert
Indiana University

PROJECT MANAGER

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

DEVELOPERS

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



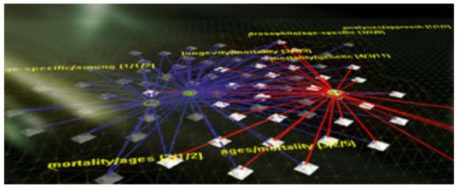
VISIT: http://nwb.slis.indiana.edu

DATA MANAGEMENT

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

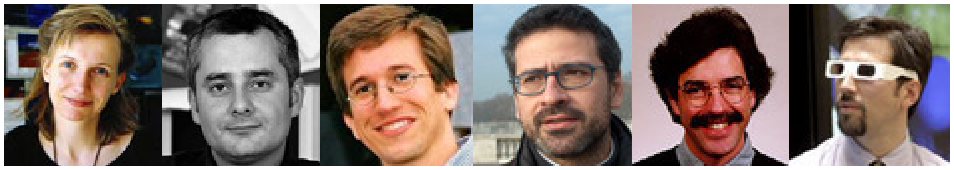
ALGORITHM INTEGRATION

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



Network Workbench

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



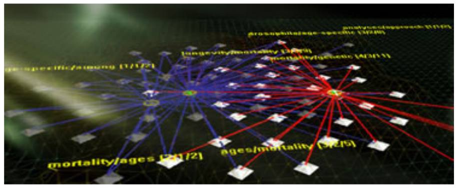
Software Team: **Team Lead: Weixia (Bonnie) Huang**
Software Developers: Bruce Herr & Ben Markines
Algorithm Developers: Santo Fortunato & Cesar Hidalgo

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

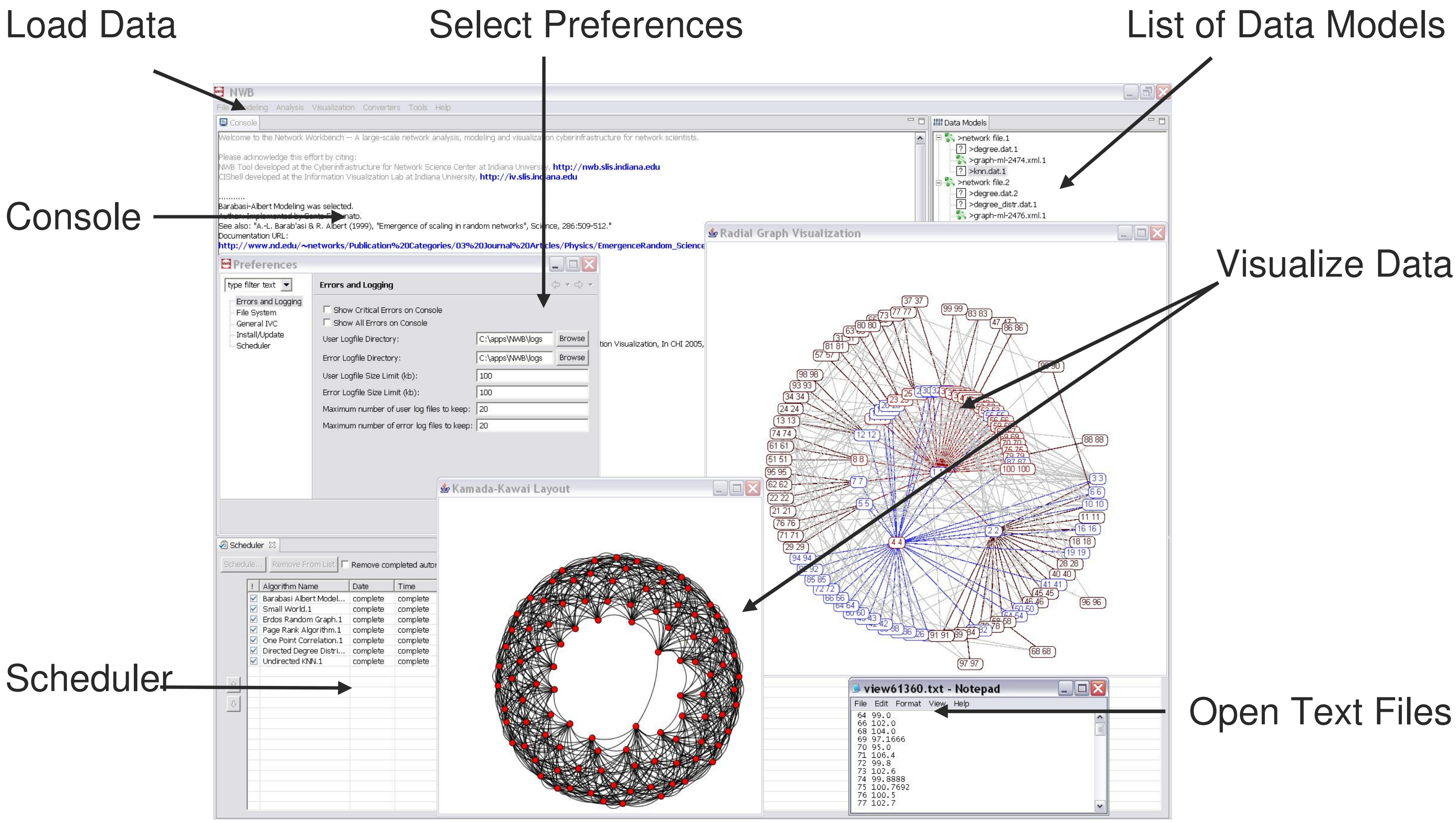
Amount: $1,120,926 NSF IIS-0513650 award.

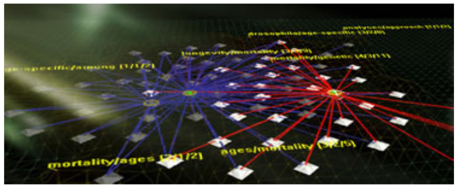
Duration: Sept. 2005 - Aug. 2008

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



NWB Tool: Interface Elements





List of Algorithms (partially implemented)

NetworkWorkbench
A Workbench for Network Scientists

Modeling
Random Network Model
Random

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

Rewiring algorithms
Rewiring based on degree distribution
Watts Strogatz Small World Model

Peer-to-Peer Models

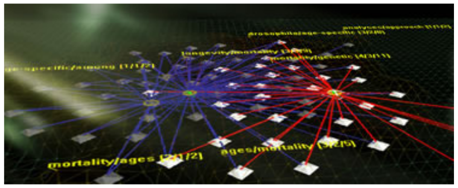
Structured
CAN Model
Chord Model

Unstructured
PRU Model
Hypergrid Model

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

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

Global
Density
Square of Adjacency Matrix
Giant Component
Strongly Connected Component
Betweenness Centrality
Diameter
Shortest Path = Geodesic Distance
Average Path Length
Motif Identification
Page Rank
Closeness centrality
Reach centrality
Eigenvector centrality
Minimum Spanning Tree



List of Algorithms (partially implemented)

Basic Processes on Networks

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

Epidemics Spreading
SIR
SIS

Graph Matching
Simple Match
Similarity Flooding
ABSURDIST

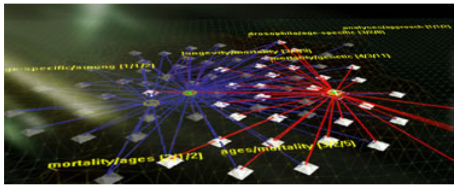
Clustering

Based on Attributes

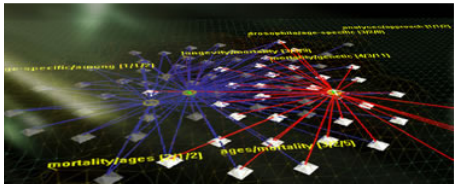
<u>Hierarchical Clustering</u>
Single Link
Complete Link
Average Link
Ward's Algorithm

<u>Based on Network Structure</u>
Newman Girvan
Clauset-Newman-Moore
Newman
Cecconi-Parisi
Simulated annealing of modularity
Caldarelli
Weak Component Clustering
vanDongen (random walk)
Cfinder (Clique percolation method)
Reichardt, Bornholdt (q-potts model)

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



Demo NWB Tool



References

- Weixia Huang, Bruce Herr & Katy Börner (May 2006) CIShell - A Plug-in Based Architecture for the Integration of Algorithms and Data Models. Network Science Workshop and Conference, Bloomington, Indiana. Available online at http://vw.indiana.edu/netsci06/conference/Huang_CIShell.pdf.
- Börner, Katy. Mapping All of Science: How to Collect, Organize and Make Sense of Mankind's Scholarly Knowledge and Expertise. Accepted for *Environment and Planning B*, Special Issue on *Mapping Humanity's Knowledge and Expertise in the Digital Domain.*
- Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao. Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. Accepted for *Scientometrics.* Dedicated issue on the *10th International Conference of the International Society for Scientometrics and Informetrics* held in Stockholm.
- Holloway, Todd, Božicevic, Miran and Börner, Katy. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. Submitted to *Complexity*, Special issue on *Understanding Complex Systems.* Also available as cs.IR/0512085.
- Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) *Visualizing the Semantic Web*, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.
- Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. *Scientometrics,* 64(3), 351-374.
- Börner, Katy, Dall'Asta, Luca, Ke, Weimao and Vespignani, Alessandro. (April 2005) Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*, special issue on *Understanding Complex Systems*, 10(4): pp. 58 - 67. Also available as cond-mat/0502147.
- Ord, Terry J., Martins, Emília P., Thakur, Sidharth, Mane, Ketan K., and Börner, Katy. (2005) Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases. *Animal Behaviour*, 69, 1399-1413. Supplementary Material.
- Mane, Ketan K. and Börner, Katy. (2004). Mapping Topics and Topic Bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5287-5290.
- Börner, Katy, Maru, Jeegar and Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl_1):5266-5273.