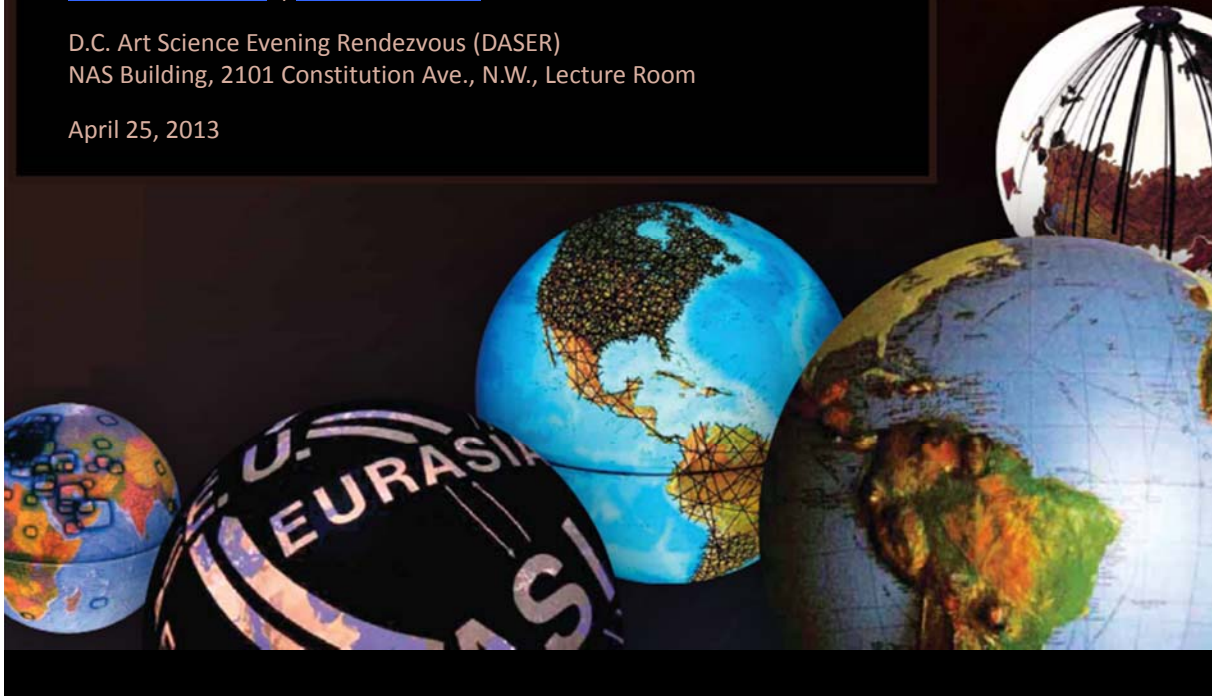


Gaining Key Insights from Big Data

Katy Börner, CNS, SLIS, Indiana University, Bloomington, Indiana
katy@indiana.edu | <http://cns.iu.edu>

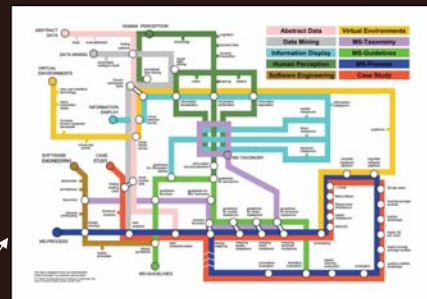
D.C. Art Science Evening Rendezvous (DASER)
NAS Building, 2101 Constitution Ave., N.W., Lecture Room

April 25, 2013



Terra bytes of data

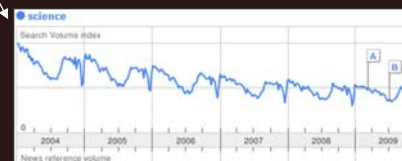
Descriptive &
Predictive
Models



Find your way



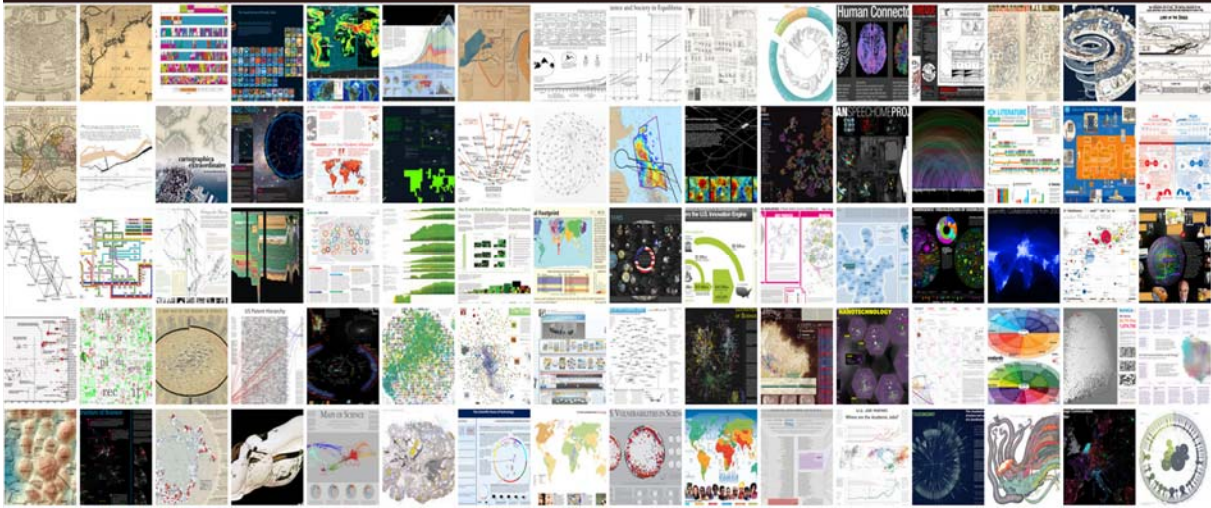
Find collaborators, friends



Identify trends

Places & Spaces: Mapping Science

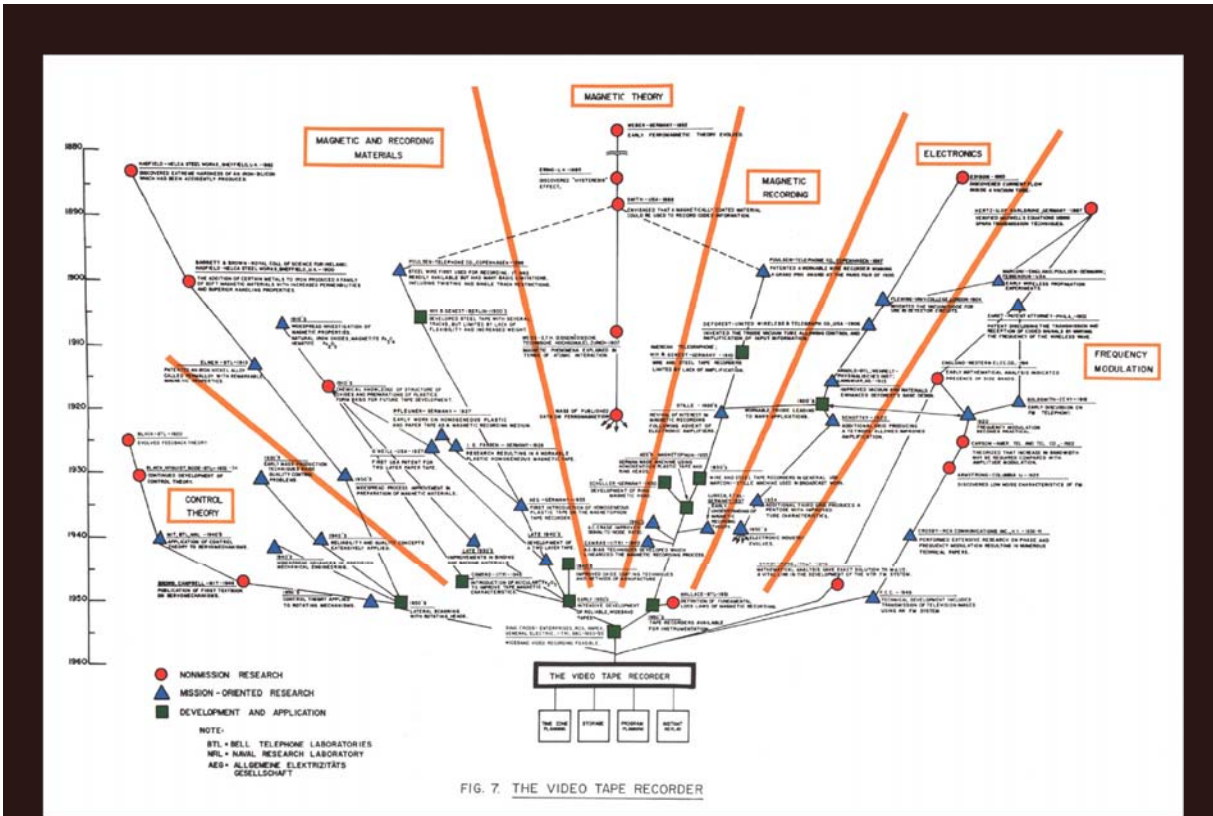
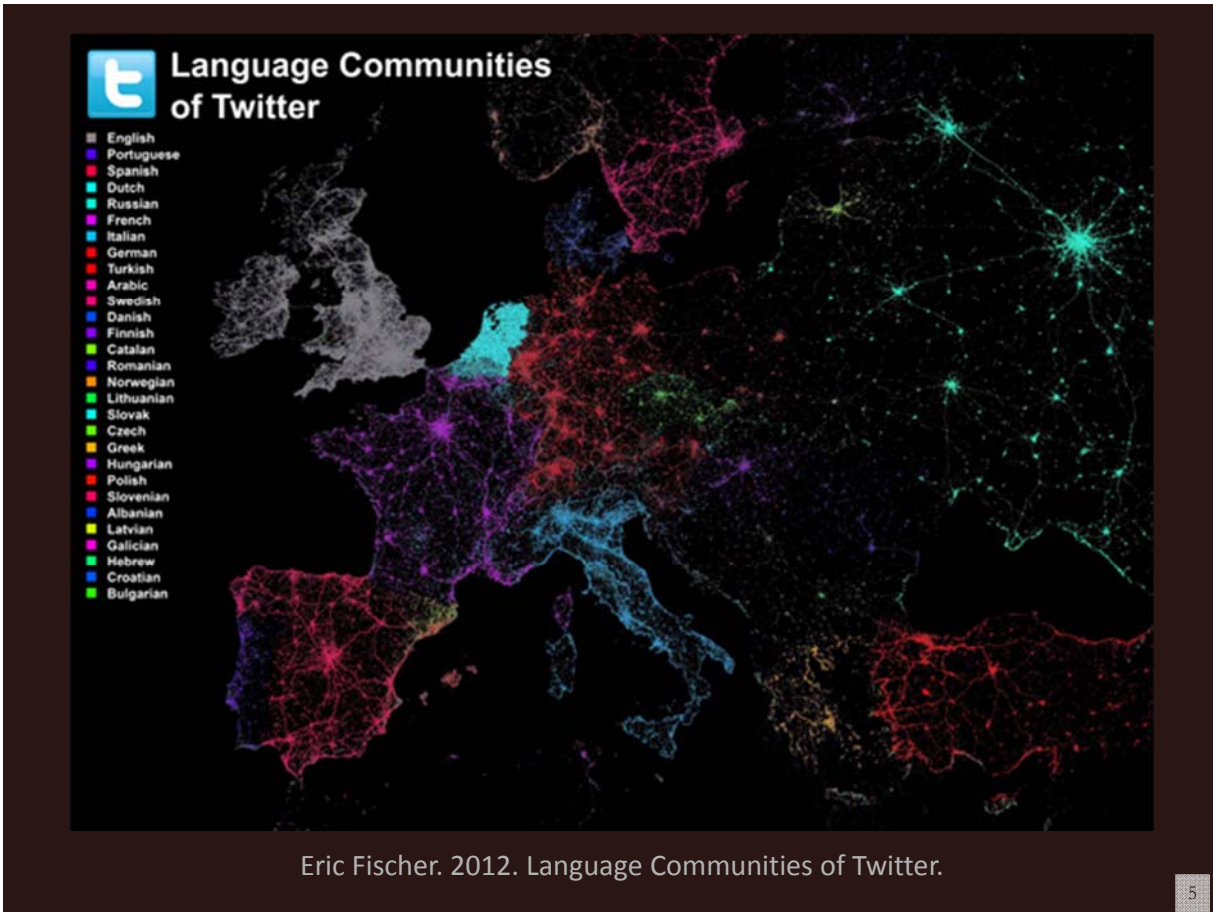
scimaps.org



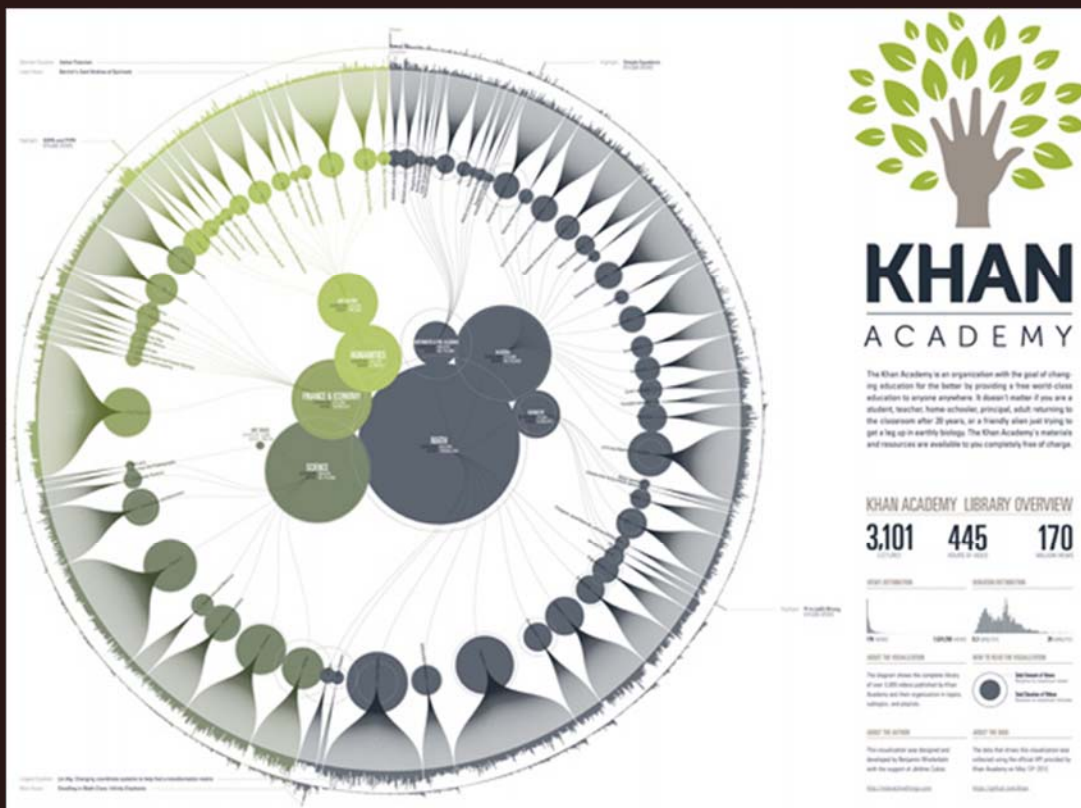
Places & Spaces: Mapping Science

scimaps.org

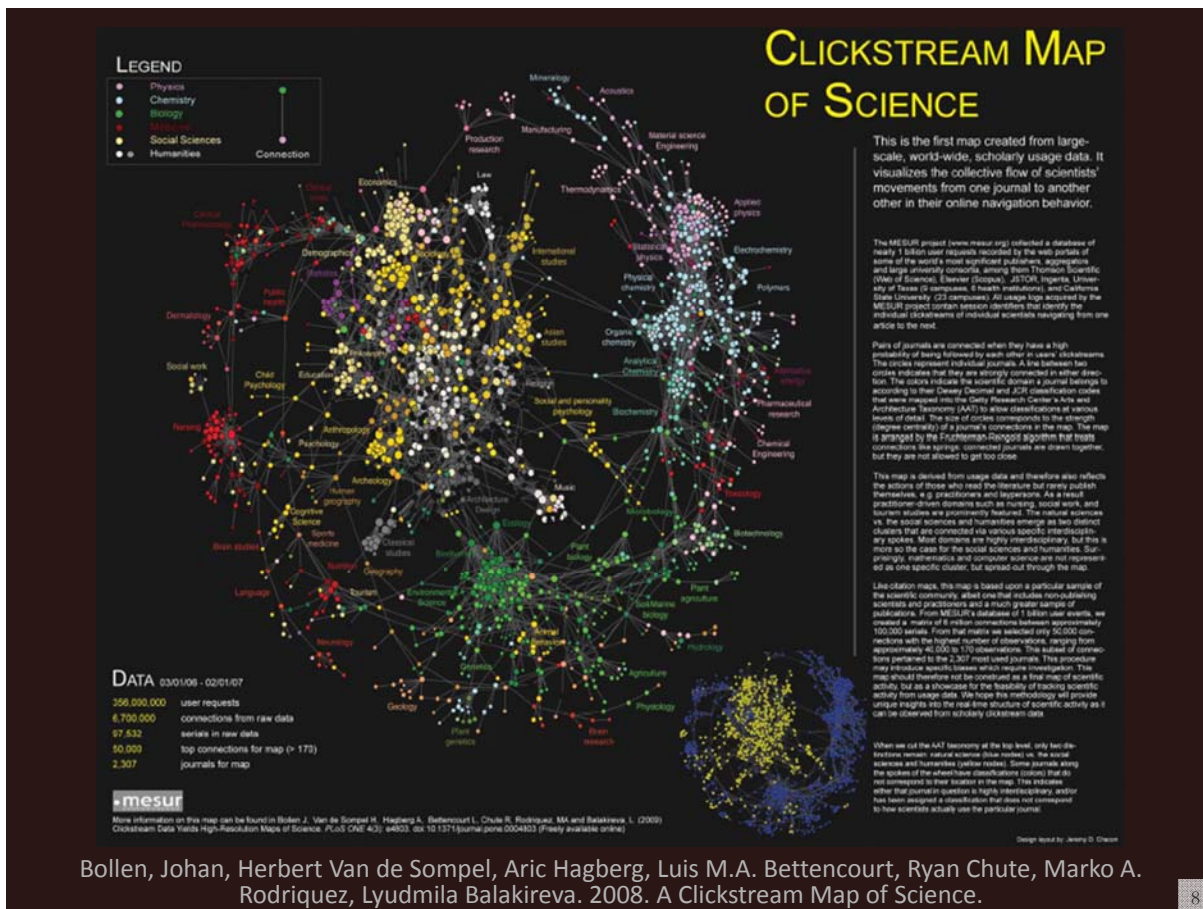




G. Benn and Francis Narin. 1968. Tracing of Key Events in the Development of the Video Tape Recorder



Benjamin Wiederkehr and Jérôme Cukier. 2012. Khan Academy Library Overview.

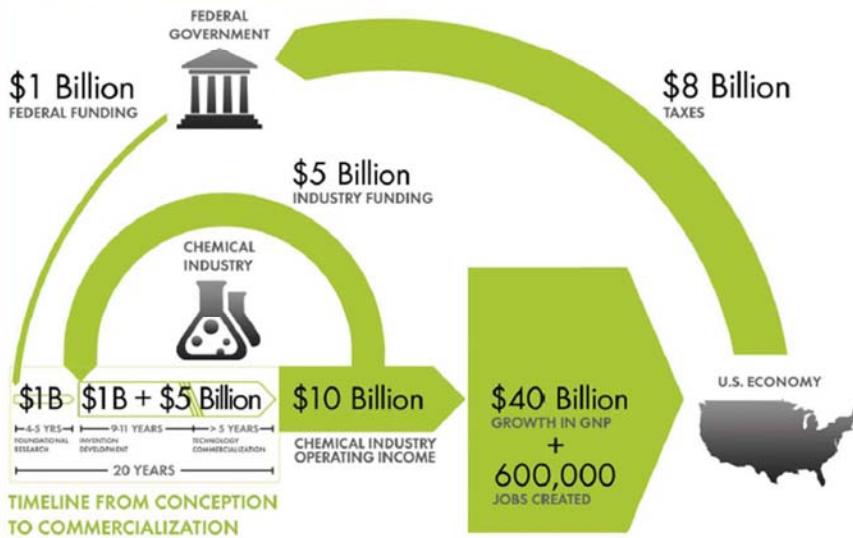


Bollen, Johan, Herbert Van de Sompel, Aric Hagberg, Luis M.A. Bettencourt, Ryan Chute, Marko A. Rodriguez, Lyudmila Balakireva. 2008. A Clickstream Map of Science.

Chemical Research & Development Powers the U.S. Innovation Engine

Macroeconomic Implications of Public and Private R&D Investments in Chemical Sciences

INVESTMENT IN CHEMICAL SCIENCE R&D



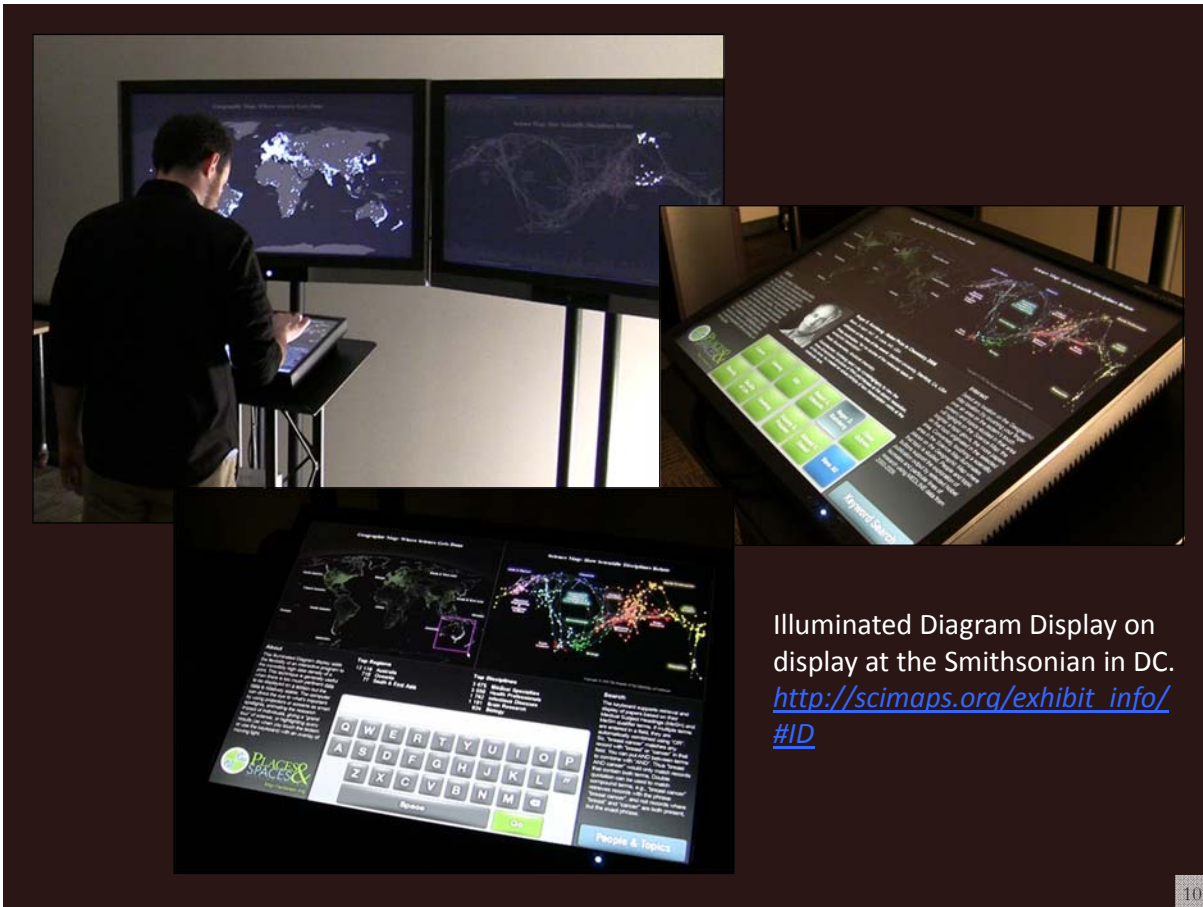
The Council for Chemical Research (CCR)

has provided the U.S. Congress and government policy makers with important results regarding the impact of Federal Research & Development (R&D) investments on U.S. innovation and global competitiveness through its commissioned 5-year two phase study. To take full advantage of typically brief access to policy makers, CCR developed the graphic below as a communication tool that distills the complex data produced by these studies in direct, concise and clear terms.



The design shows that an input of \$1B in federal investment, leveraged by \$5B industry investment, brings new technologies to market and results in \$10B of operating income for the chemical industry, \$40B growth in the Gross National Product (GNP) and further impacts the US economy by generating approximately 600,000 jobs, along with a return of \$8B in taxes. Additional details, also reported in the CCR studies, are depicted in the map to the left. This map clearly shows the two R&D investment cycles: the shorter industry investment at the innovation stage to commercialization cycle; and the longer federal investment cycle which begins in basic research and culminates in national economic and job growth along with the increase tax base that in turn is available for investment in basic research.

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



Illuminated Diagram Display on display at the Smithsonian in DC.
http://scimaps.org/exhibit_info/#ID

Geographic Map: Where Science Gets Done

Science Map: How Scientific Disciplines Relate

About

This Illuminated Diagram display adds the flexibility of an interactive program to the incredibly high data density of a print. This technique is generally useful when there is too much pertinent data to be displayed on a screen but the data is relatively stable. The computer can direct the eye to what's important by using projectors or screens as smart spotlights, animating the research impact of individuals, giving a "grand tour" of science, or highlighting query results (as when you touch the lectern or use the keyboard) with an overlay of moving light.

Top Five Continents

- North America - 4,000 records
- South & East Asia - 3,589
- Australia - 2,431
- Africa - 2,208
- South America - 1,562

Top Five Scientific Disciplines

- Math & Physics - 4,000 records
- Health Professionals - 3,589
- Social Sciences - 2,431
- Aeronautical, Chemical, Mechanical & Civil Engineering - 2,208
- Humanities - 1,562

Search

The keyboard supports retrieval and display of papers based on their Medical Subject Headings (MeSH) and MeSH qualifier terms. If multiple terms are entered in a field, they are automatically combined using "OR". So, "breast cancer" matches any record with "breast" or "cancer" in that field. You can put AND between terms to combine with "AND". Thus "breast AND cancer" would only match records that contain both terms. Double quotation can be used to match compound terms, e.g., "breast cancer" retrieves records with the phrase "breast cancer", and not records where "breast" and "cancer" are both present, but the exact phrase.

Q	W	E	R	T	Y	U	I	O	P
A	S	D	F	G	H	J	K	L	"
Z	X	C	V	B	N	M			

Space Go

<http://scinaps.org>

People & Topics

11

Geographic Map: Where Science Gets Done

Science Map: How Scientific Disciplines Relate

About

This Illuminated Diagram display adds the flexibility of an interactive program to the incredibly high data density of a print. This technique is generally useful when there is too much pertinent data to be displayed on a screen but the data is relatively stable. The computer can direct the eye to what's important by using projectors or screens as smart spotlights, animating the research impact of individuals, giving a "grand tour" of science, or highlighting query results (as when you touch the lectern or use the keyboard) with an overlay of moving light.

Elinor Ostrom - Nobel Prize in Economic Sciences 2009

Born: 7 August 1933, New York, NY, USA
Affiliation at the time of the award: Indiana University, Bloomington, IN, USA, Arizona State University, Tempe, AZ, USA
Prize motivation: "for her analysis of economic governance, especially the commons"
Field: Economic governance
Contribution: Challenged the conventional wisdom by demonstrating how local property can be successfully managed by local commons without any regulation by central authorities or privatization.

Interact

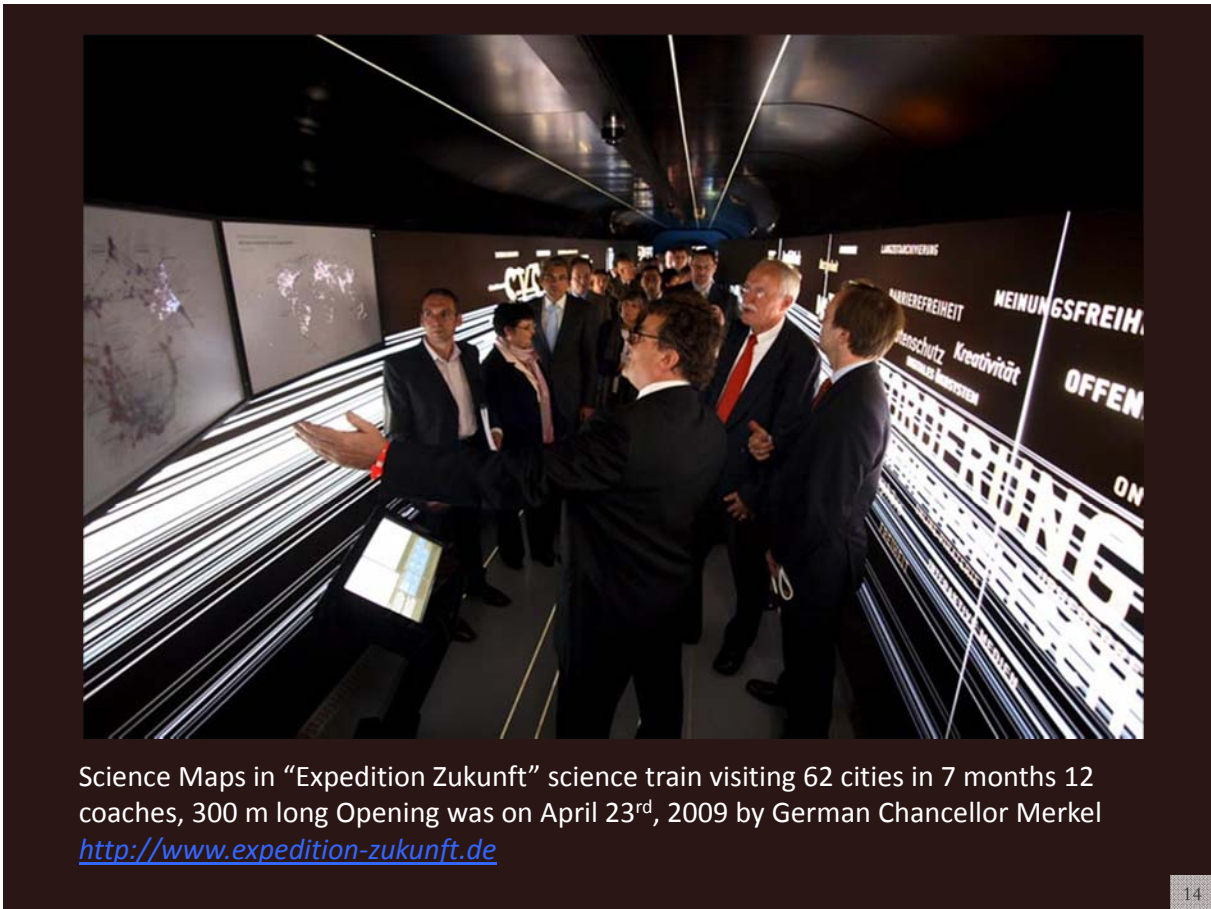
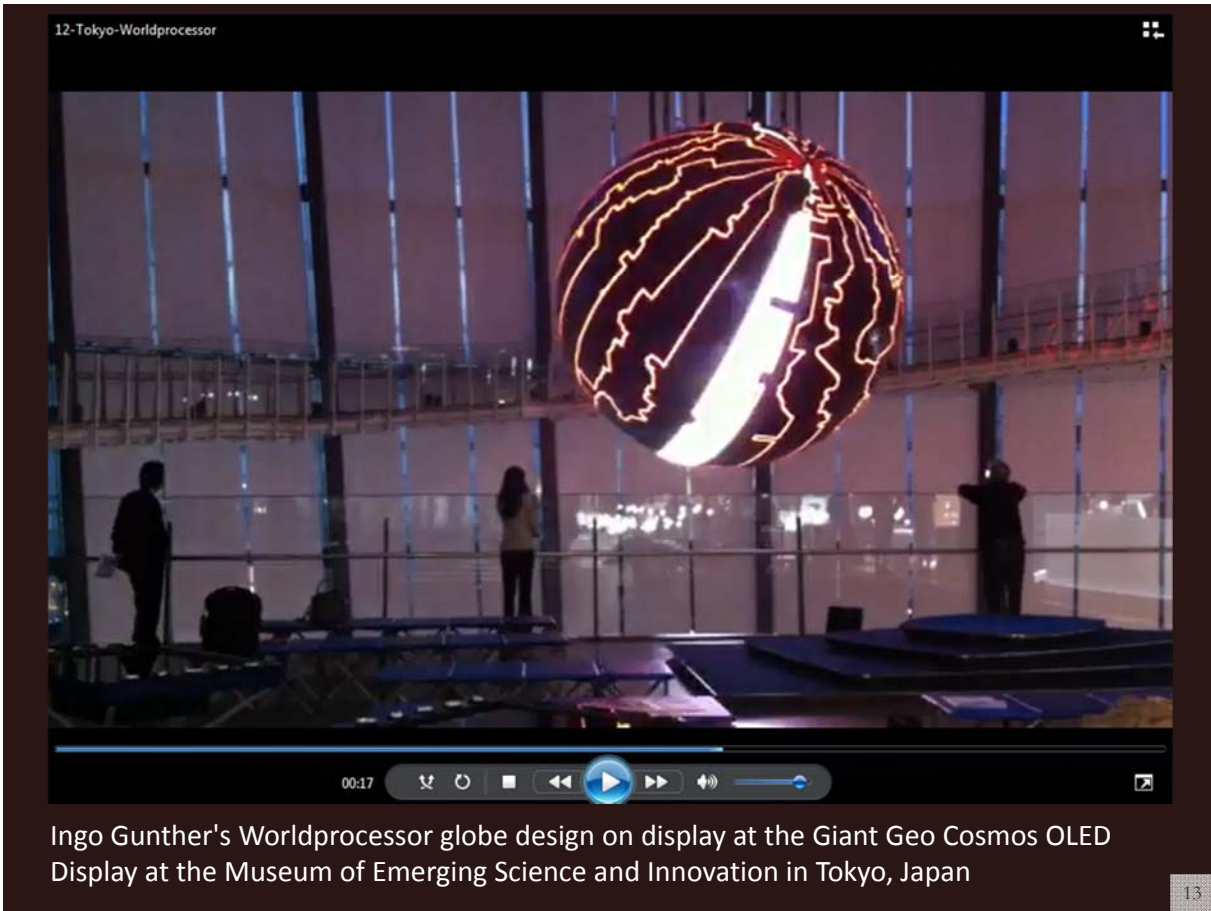
Select any location on the Geographic Map location (by brushing your finger over an area on the lectern's touch screen) and topics studied in that area will highlight on the Science Map; the brighter a topic glows, the more papers on that topic originated in the selected area. Conversely, touching a scientific area in the Science Map illuminates places on the Geographic Map where that topic is studied. People and topic buttons support the exploration of publication output by selected Noble laureates and particular lines of research using MEDLINE data from 2000-2009.

<http://scinaps.org>

Cancer	Cloning	HIV	Robert G. Edwards	Roger D. Kornberg	Elinor Ostrom
Obesity	Quality of Life	Smoking	Stanley B. Prusiner	Ahmed H. Zewail	View All

Keyword Search

12





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>



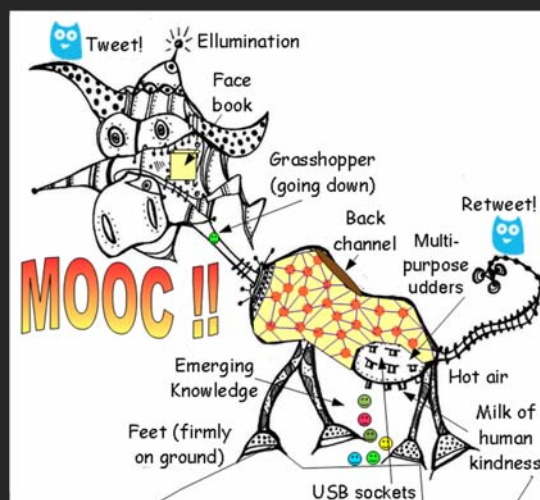
Anyone Can Cook & Anyone Can Map

17

MOOCs

In 2012, Google hosted three massive open online courses (MOOCs) collectively reaching over 400,000 registrants.

By the end of 2013 more than 250 courses will be run using the Google, Coursera, Udacity, EdX, and other platforms.



18

Overview

This course provides an overview about the state of the art in information visualization. It teaches the process of producing effective visualizations that take the needs of users into account.

Among other topics, the course covers:

- Data analysis algorithms that enable extraction of patterns and trends in data
- Major temporal, geospatial, topical, and network visualization techniques
- Discussions of systems that drive research and development.

Please watch the introduction video to get better acquainted with the course.

Everybody who registers gains free access to the Scholarly Database (26 million paper, patent, and grant records) and the Sci2 Tool (100+ algorithms and tools).

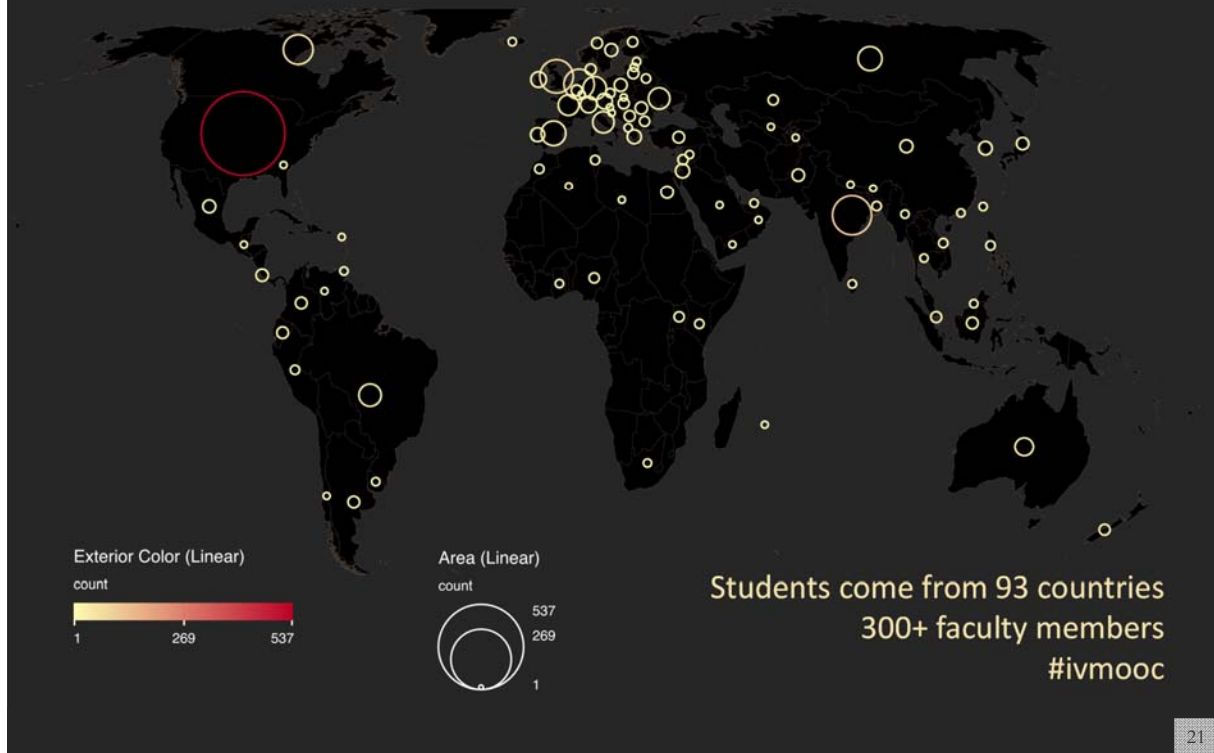
Katy Börner, Ph.D.
Indiana University



[Go To The Course](#)

ivmooc.cns.iu.edu





Instructors

Katy Börner – Theory Parts

Instructor, Professor at SLIS



David E. Polley – Hands-on Parts

CNS Staff, Research Assistant with MIS/MLS
Teaches & Tests Sci2 Tool



Scott B. Weingart – Client Work

Assistant Instructor, SLIS PhD student



Course Schedule

- **Session 1** – Workflow design and visualization framework
- **Session 2** – “When:” Temporal Data
- **Session 3** – “Where:” Geospatial Data
- **Session 4** – “What:” Topical Data

Mid-Term

Students work in teams with clients.

- **Session 5** – “With Whom:” Trees
- **Session 6** – “With Whom:” Networks
- **Session 7** – Dynamic Visualizations and Deployment

Final Exam

23

Grading

All students are asked to create a personal profile to support working in teams.



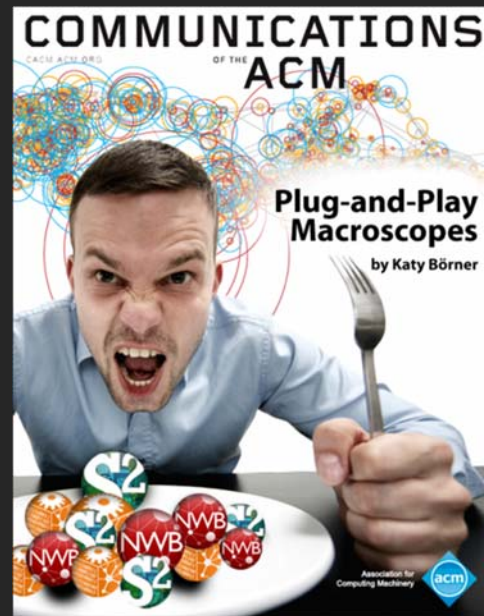
Final grade is based on Midterm (**30%**), Final (**40%**), Client Project (**30%**).

- Weekly self-assessments are not graded.
- Homework is graded automatically.
- Midterm and Final test materials from theory and hands-on sessions are graded automatically.
- Client work is peer-reviewed via online forum.

All students that receive more than **80%** of all available points get an official certificate/badge.

24

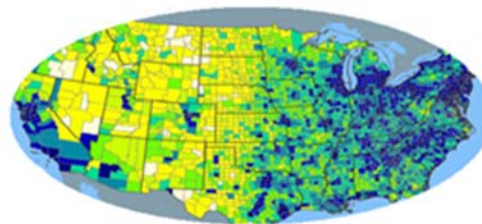
Plug-and-Play Macroscopes
cishell.org



Börner, Katy. (March 2011). Plug-and-Play Macroscopes. *Communications of the ACM*, 54(3), 60-69. <http://www.scivee.tv/node/27704>

Different Levels of Abstraction/Analysis

Macro/Global
Population Level



Meso/Local
Group Level



Micro
Individual Level

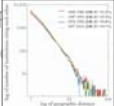

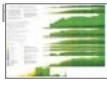

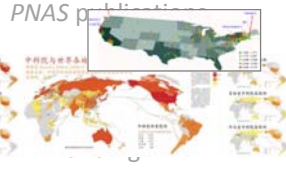
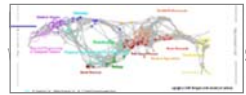



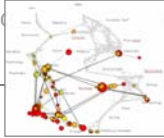


Type of Analysis vs. Level of Analysis

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

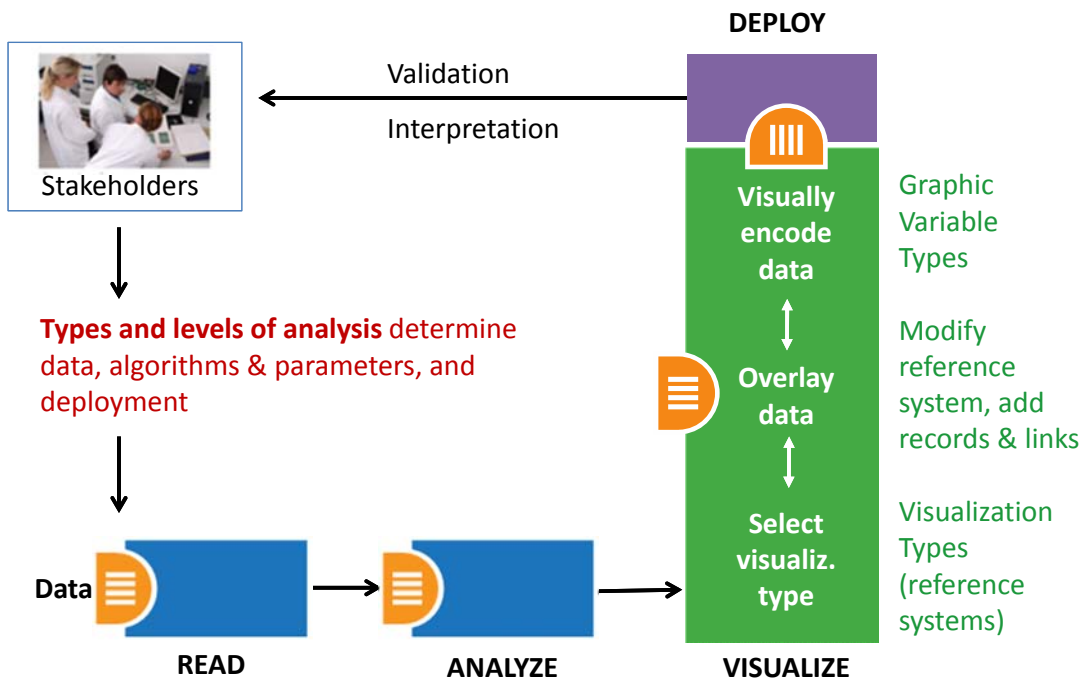
27

Type of Analysis vs. Level of Analysis

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

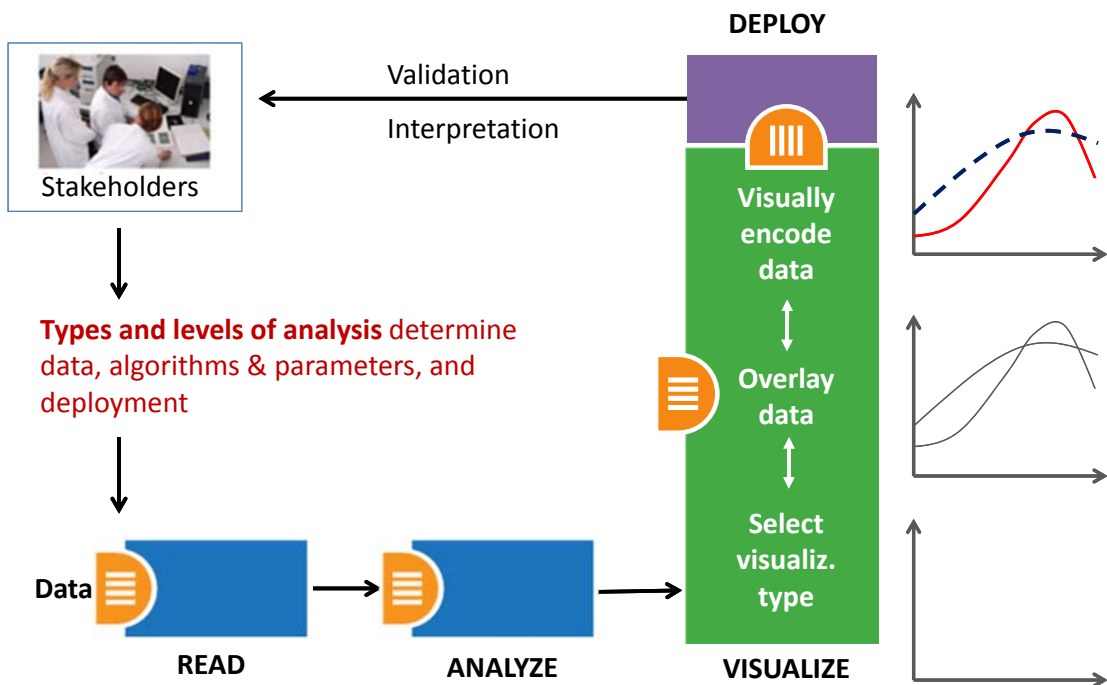
28

Needs-Driven Workflow Design



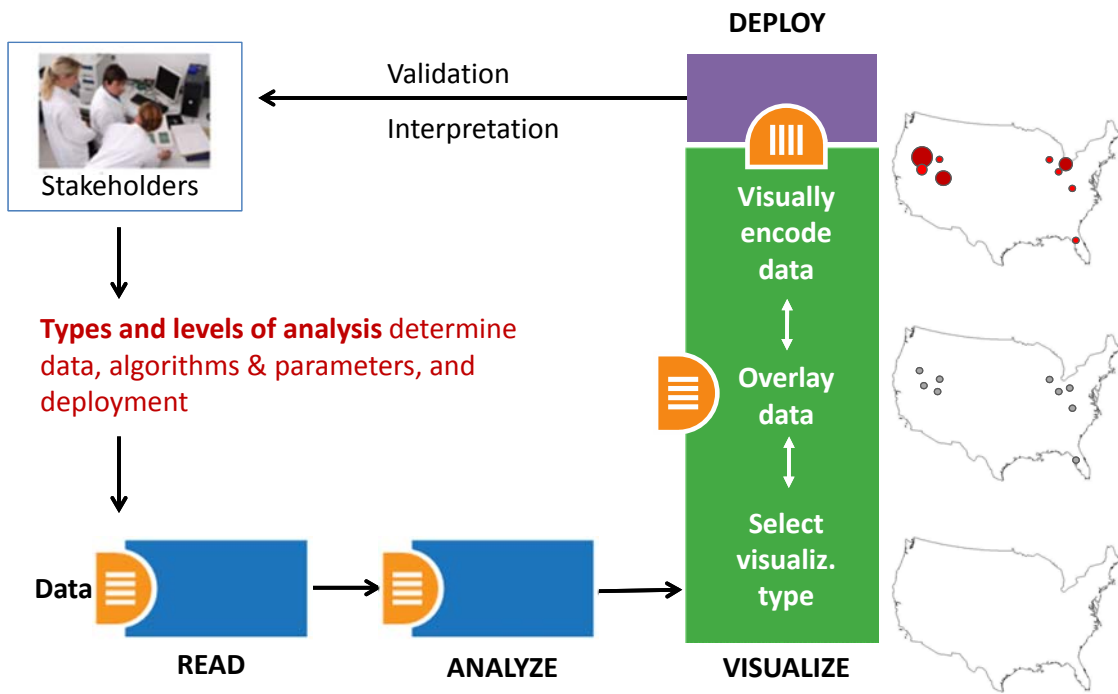
29

Needs-Driven Workflow Design






30

Needs-Driven Workflow Design



31

Visualization Types vs. Data Overlays

Visualization Type	Chart	Table	Graph	Geospatial Map	Network Graph
Modify / visually encode base map.					
Place and visually encode records/nodes.					
Place and visually encode links.					

Plus, add a title, labels, legend, explanatory text, and author info.

32

Visualization Types vs. Data Overlays

Visualization Type	Chart	Table	Graph	Geospatial Map	Network Graph
Modify / visually encode base map.					
Place and visually encode records/nodes.					
Place and visually encode links.					



Plus, add a title, labels, legend, explanatory text, and author info.

33

Clients

Information Visualization MOOC INDIANA UNIVERSITY CNS

List of Clients

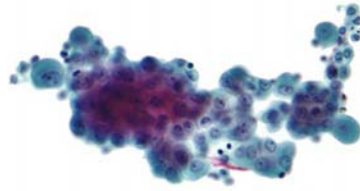
ISIS **Project Title:** Isis: 100 Years
Client Name: Jay Malone
Project goal/scientific or practical value: A visual representation Isis' contributors and locales over the past 100 years. Isis is the journal of the History of Science Society. This representation will provide a dynamic picture of how scholarship in the history of science has shifted over the past century.
Information on dataset(s) to be used: Citation information, author locale, and issue number for Isis publications.
Relevant publications, websites, etc: <http://www.press.uchicago.edu/ucp/journals/journal/isis.html>
Conditions under which students can publish results and/or add project results to their resume: Client would like to approve results.

oycib **Project Title:** e-Xploration
Client Name: Luyi
Project goal/scientific or practical value: e-Xploration is an agent-based model for the ethnographic observation and the registry, analysis, and interpretation of social practices in virtual communities for intervention in the development of collaboration and cooperation. This project will analyze the interactions between subjects and objects in a platform collaborative community called OYCIB, a project based on e-Xploration (e-crick.net).
Information on dataset(s) to be used: I can provide a data base in .graphml format for the students. The file .graphml contains the interactions between subjects and objects in a platform collaborative community called OYCIB. In the level of practice, it is not necessary that students know agent-based models for using the database. But, in another level, for example: the collaborate level for the OYCIB development, it is necessary to have basic knowledge in AMS or MAS and another competences like PHP and MySQL.
Relevant publications, websites, etc: <http://www.e-crick.net/logs>
Conditions under which students can publish results and/or add project results to their resume: If any person or institution use my dataset or another info about eXploration (e-crick.net, oycib.net), I need to approve the results and appear as co-author.

http://ivmooc.cns.iu.edu/ivmooc_clientprojects.html

34

Diogo Carmo



Mesothelioma

Main title topics in Medline papers

Mesothelioma (a rare form of cancer that develops from transformed cells originating in the mesothelium, the protective lining that covers many of the internal organs of the body) is usually caused by exposure to asbestos.

The most common anatomical site for the development of mesothelioma is the pleura (the outer lining of the lungs and internal chest walls, but it can also arise in the peritoneum (the lining of the abdominal cavity), and the pericardium (the sac that surrounds the heart) or the tunica vaginalis (a sac that surrounds the testis).

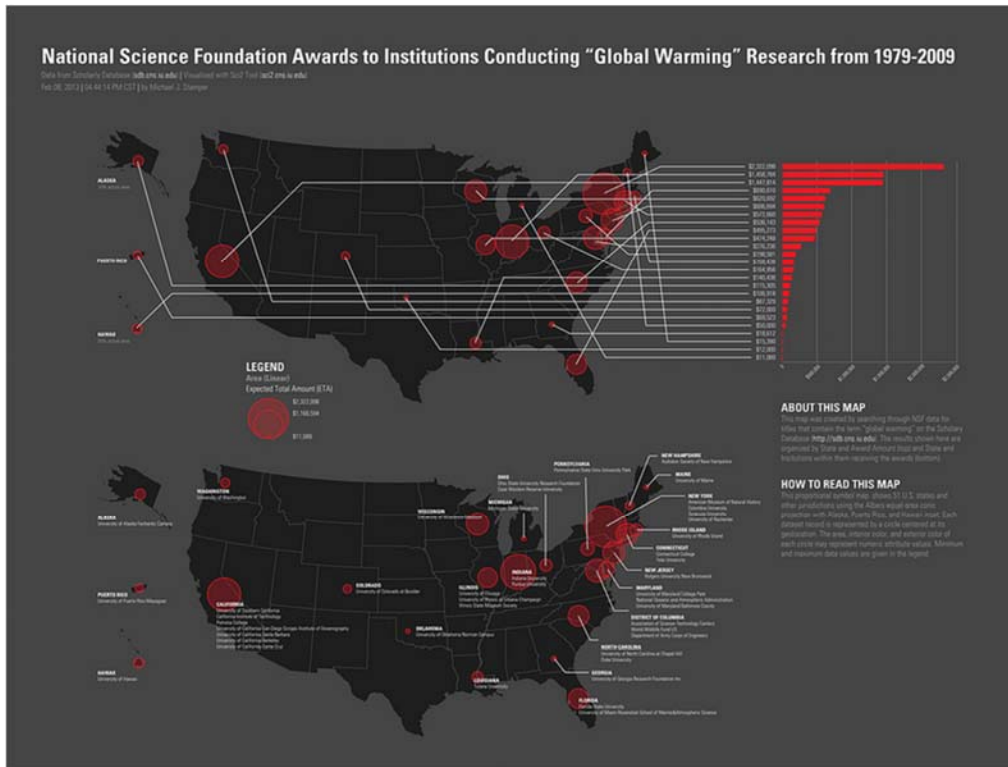
Most people who develop mesothelioma have worked in jobs where they inhaled asbestos, or were exposed to asbestos dust and fibers in other ways. It has also been suggested that working children of family members who worked with asbestos increases their risk for developing mesothelioma. Unlike lung cancer, there seems to be no association between mesothelioma and tobacco smoking, but smoking greatly increases the risk of other asbestos-induced cancers. Some people who were exposed to asbestos have inherited damage to their asbestos-related disease, including mesothelioma. Compensation via asbestos funds or class action lawsuits is an important issue in law practices regarding mesothelioma.

MALIGNANT PLEURAL CYSTIC BENIGN DIAGNOSIS



How To Read This Map
This temporal bar graph visualization represents each record as a horizontal bar with a specific start and end date and a bar label on its left side. The area of each bar encodes the number of records, and target magnitude - in the scope identified in the label (target words were summed).

Author: Diogo Carmo <http://diogo.carmo@indiana.edu> | Visualization software: Sci2 Tools, (2009) Science of Science (Sci2) Tool, Indiana University and SciTech Strategies, <http://sci2.ucis.edu/> | Dataset: Medline Papers, as available in Pubmed Database, <http://pubmed.ncbi.nlm.nih.gov/> | Text and Images: Wikipedia Mesothelioma article, available at <http://en.wikipedia.org/wiki/Mesothelioma> | Font: Sansation, by Bernd Mohr © 2011 - All Rights Reserved. This font family is freeware and is available at <http://beta1.com/berndmohr.com/>



mjstammer ivmooc

References

Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). **Visualizing Knowledge Domains**. In Blaise Cronin (Ed.), *ARIST*, Medford, NJ: Information Today, 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).

Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). **Network Science**. In Blaise Cronin (Ed.), *ARIST*, Information Today, Inc., Volume 41, Chapter 12, pp. 537-607. <http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf>

Börner, Katy (2010) **Atlas of Science**. MIT Press. <http://scimaps.org/atlas>

Scharnhorst, Andrea, Börner, Katy, van den Besselaar, Peter (2012) **Models of Science Dynamics**. Springer Verlag.



37

Acknowledgments

We would like to thank Miguel I. Lara and his colleagues at the Center for Innovative Teaching and Learning for instructional design support, Samuel Mills for designing the web pages, Robert P. Light and Thomas Smith for extending the GCB platform, and Mike Widmer and Mike T. Gallant for adding the Forum. Support comes from CNS, CITL, SLIS, SOIC, and Google.

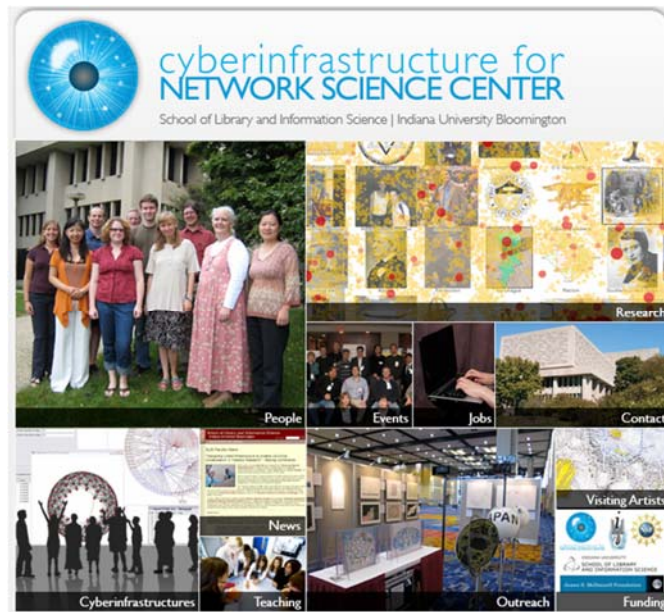
The tool development work is supported in part by the Cyberinfrastructure for Network Science Center and the School of Library and Information Science at Indiana University, the National Science Foundation under Grants No. SBE-0738111 and IIS-0513650, the US Department of Agriculture, the National Institutes of Health, and the James S. McDonnell Foundation.

Visualizations used in the course come from the Places & Spaces: Mapping Science exhibit, online at <http://scimaps.org>, and from the *Atlas of Science: Visualizing What We Know*, MIT Press (2010).



Cyberinfrastructure for Network Science Center

38



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

CNS Facebook: <http://www.facebook.com/cnscenter>

Mapping Science Exhibit Facebook: <http://www.facebook.com/mappingscience>