

# Converting Data Into Actionable Insights: Open Data, Open Code, and Open Education

**Katy Börner**

Victor H. Yngve Professor of Information Science  
Director, Cyberinfrastructure for Network Science Center  
School of Informatics and Computing and Indiana University Network Science Institute  
Indiana University, USA

Commerce Data Advisory Council (CDAC) Meeting  
National Oceanic and Atmospheric Administration (NOAA)  
David Skaggs Research Center, Boulder, CO

October 29-30, 2015

*Olivier H. Beauchesne, 2011. Map of Scientific Collaborations from 2005-2009.*

Computed Using Data from Elsevier's Scopus

## Map of Scientific Collaborations from 2005-2009



*Olivier H. Beauchesne, 2011. Map of Scientific Collaborations from 2005-2009.*

Computed Using Data from Elsevier's Scopus

# Open Data

**Data** - economic, scientific, environmental, patent, geospatial

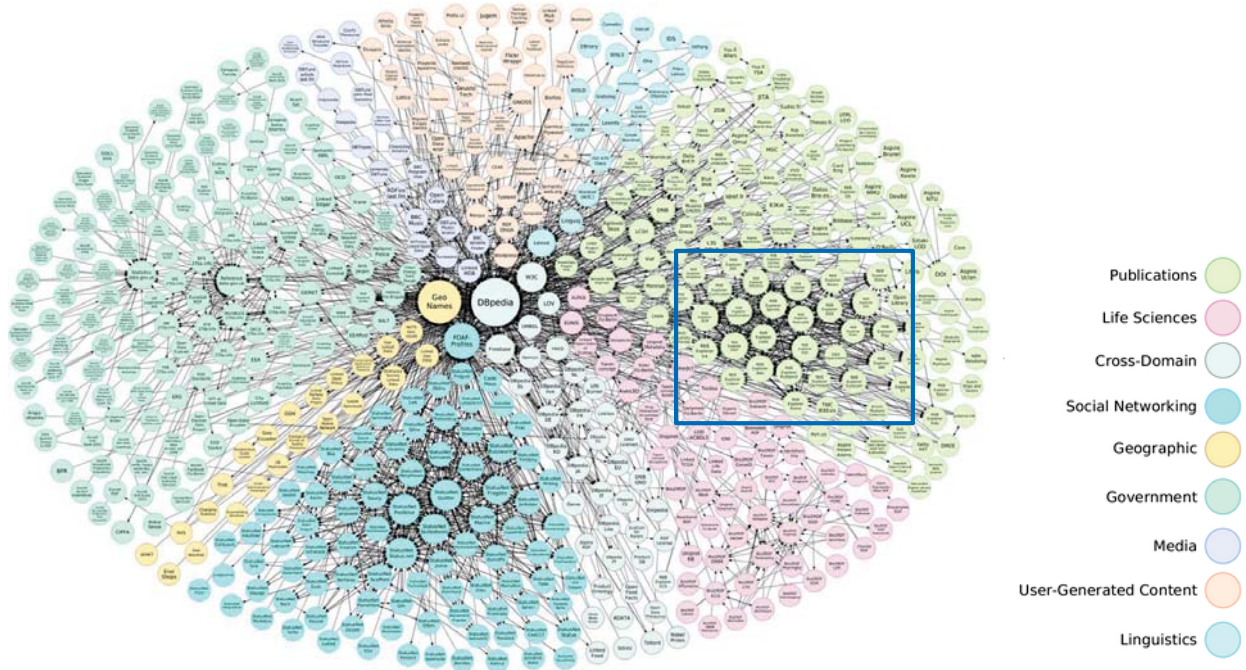
Core CDAC Mission:

- data management practices
- common, open data standards
- policy issues related to privacy, latency, and consistency
- effective models for public-private partnership
- external uses of Commerce data
- methods to build new feedback loops between the Department and data users.

<http://www.esa.doc.gov/content/cdac-mission-goals>

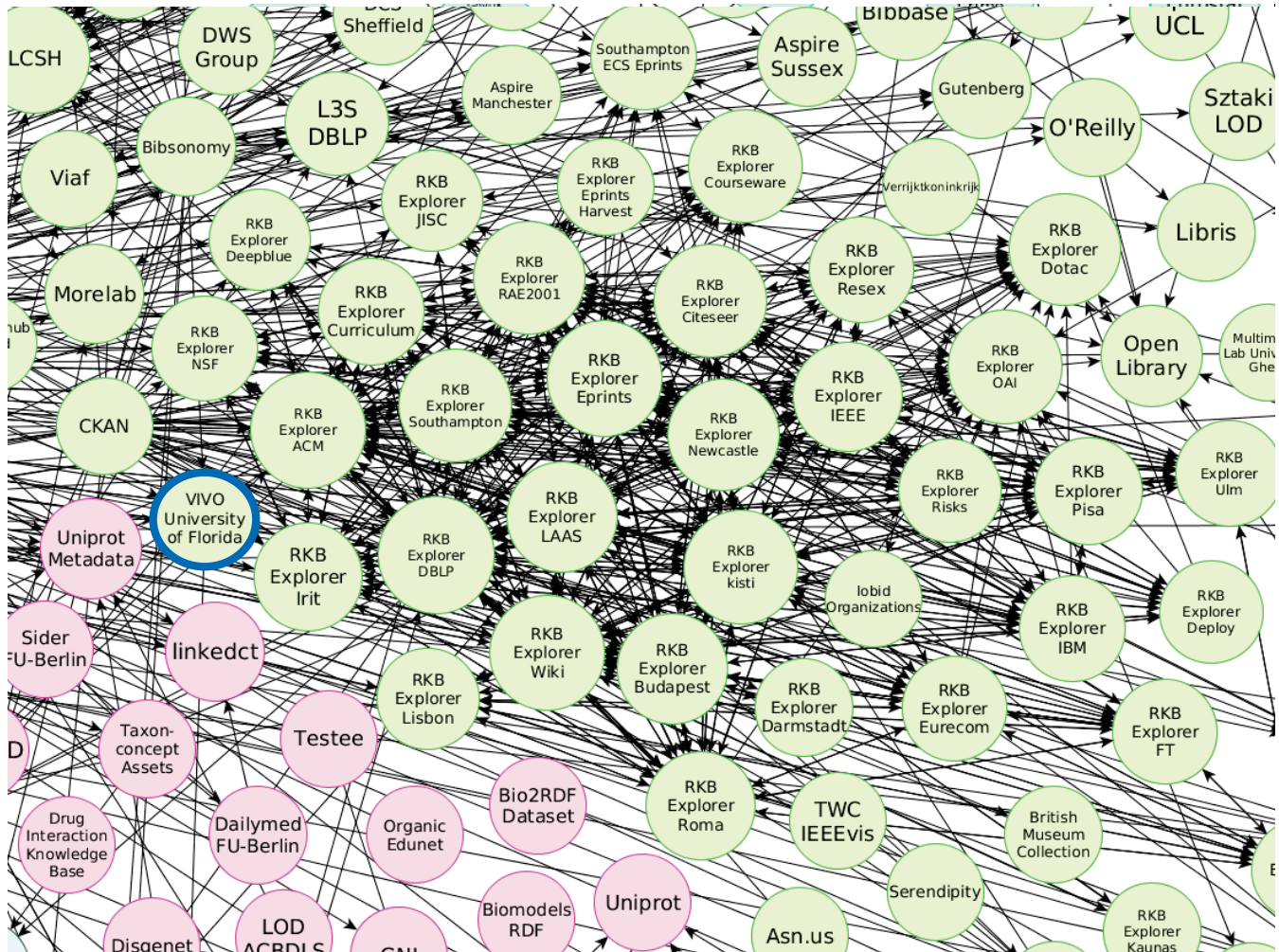


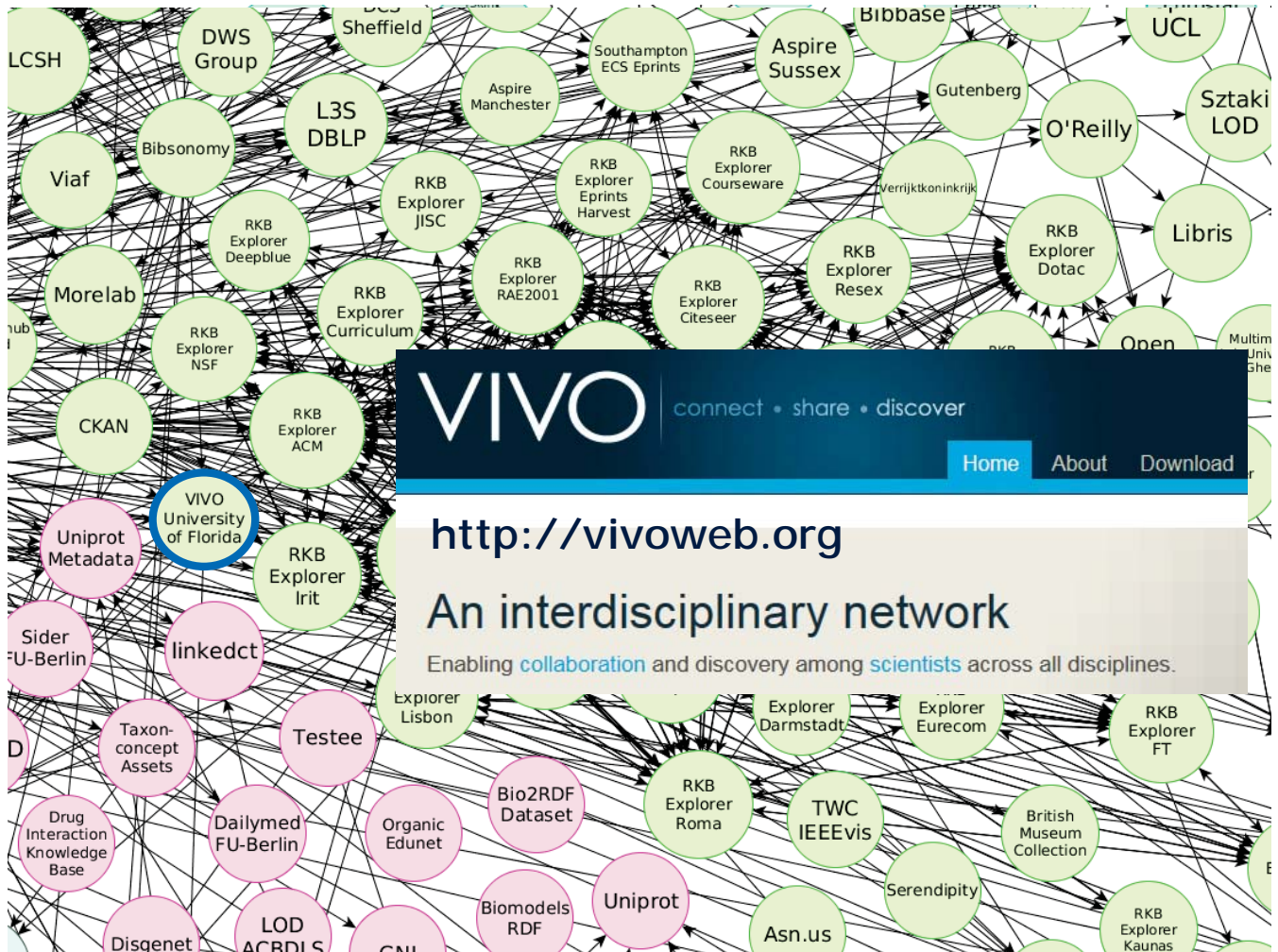
# Data – Expertise



Linked Open Data Cloud Diagram 2014

<http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/>





<http://nrn.cns.iu.edu>

# Open Data Portals in Europe

Open Data Portals as part of the Europe 2020 Initiative

<http://ec.europa.eu/digital-agenda/en/open-data-portals>

European Union Open Data Portal

<http://open-data.europa.eu/en/data/>

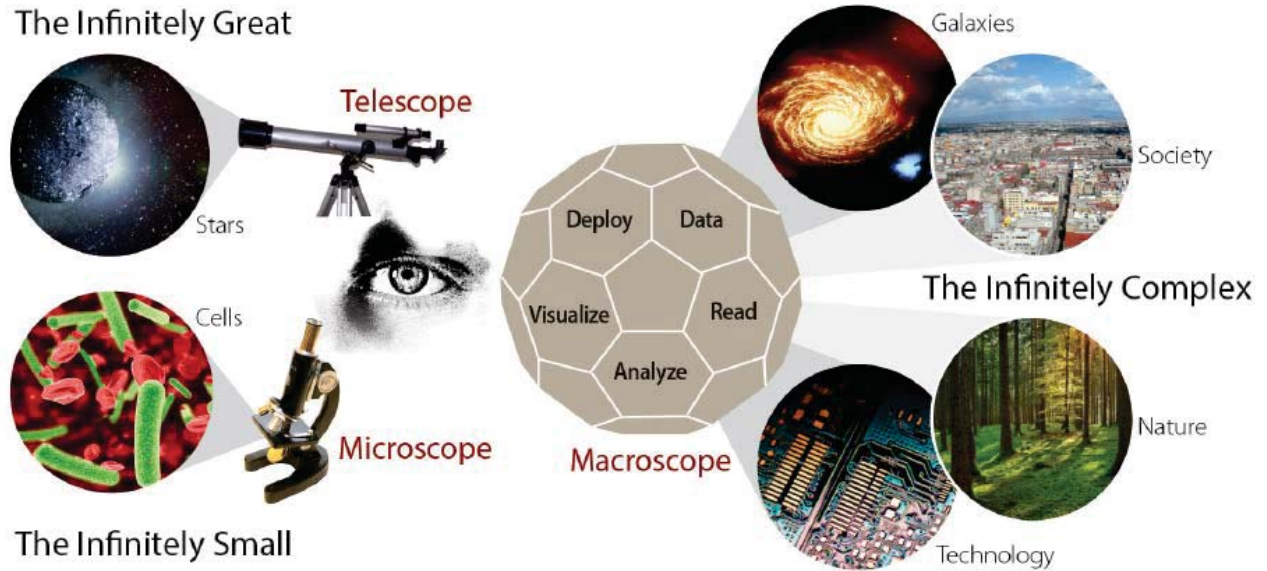
European cloud of public administration data and services

<http://open-dai.eu>

**Open Code**

# Code – analysis, simulation, visualization

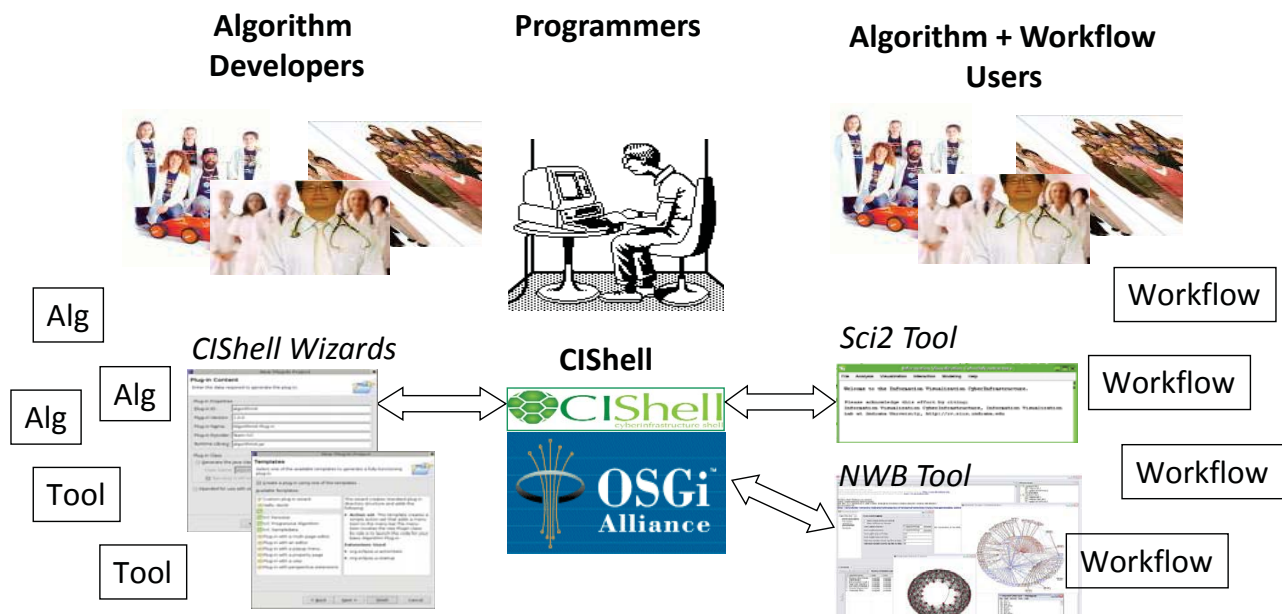
## Microscopes, Telescopes, Macrosopes Plug-and-Play Macrosopes



13

# Code – analysis, simulation, visualization

## Plug-and-Play Macrosopes

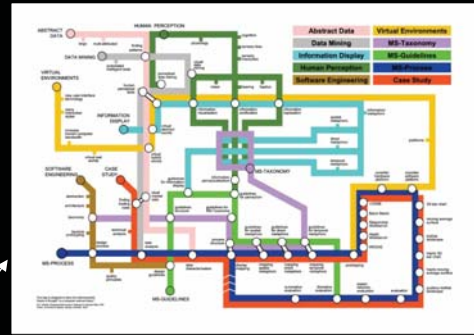


14



Terra bytes of data

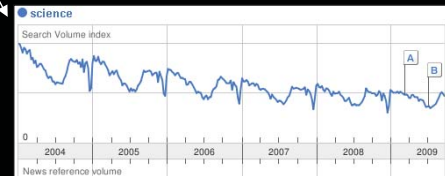
Macrosopes



Find your way

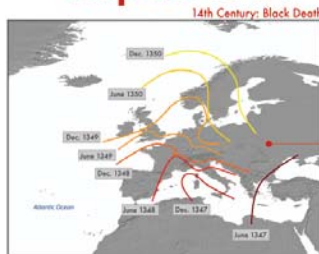


Find collaborators, friends



Identify trends

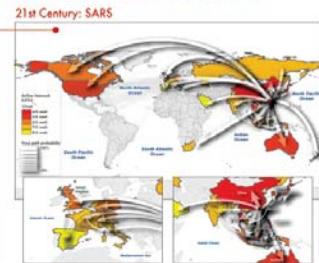
## Impact of Air Travel on Global Spread of Infectious Diseases



**Epidemic spreading pattern changed dramatically after the development of modern transportation systems.**

In pre-industrial times disease spread was mainly a spatial diffusion phenomenon. During the spread of Black Death in the 14th century Europe, only few traveling means were available and typical trips were limited to relatively short distances on the time scale of one day. Historical studies confirm that the disease diffused smoothly generating an epidemic front traveling as a continuous wave through the continent at an approximate velocity of 200-400 miles per year.

The SARS outbreak on the other hand was characterized by a patched and heterogeneous spatio-temporal pattern mainly due to the air transportation network identified as the major channel of epidemic diffusion and ability to connect far apart regions in a short time period. The SARS maps are obtained with a data-driven stochastic computational model aimed at the study of the SARS epidemic pattern and analysis of the accuracy of the model's predictions. Simulation results describe a spatio-temporal evolution of the disease (color coded countries) in agreement with the historical data. Analysis on the robustness of the model's forecasts leads to the emergence and identification of epidemic pathways as the most probable routes of propagation of the disease. Only few preferential channels are selected (arrows; width indicates the probability of propagation along that path) out of the huge number of possible paths the infection could take by following the complex nature of airline connections (light grey; source: IATA).



## Forecasts of the Next Pandemic Influenza

### Seasonal



Forecasts are obtained with a stochastic computational model which explicitly incorporates data on worldwide air travel and detailed census data to simulate the global spread of an influenza pandemic.

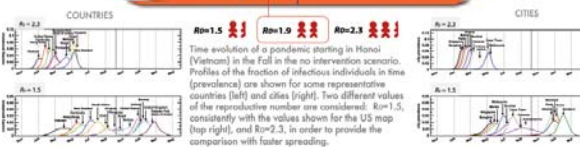
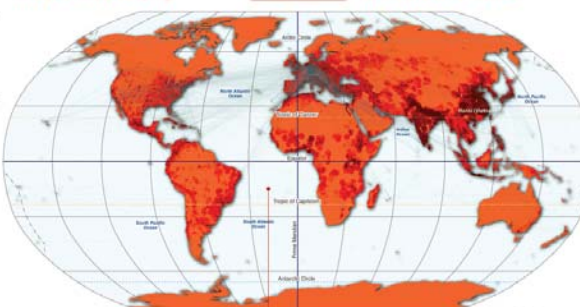
The modeling approach considers infection dynamics (i.e., virus transmission, onset of symptoms, infectiousness, recovery, etc.) among individuals living in urban areas around the world, and assumes that individuals are allowed to travel from one city to another by means of the airline transportation network.

### Geographical



Numerical simulations provide results for the temporal and geographic evolution of the pandemic influenza in 3,100 urban areas located in 220 different countries. The model allows to study different spreading scenarios, characterized by different initial outbreak conditions, both geographical and seasonal.

The central map represents the cumulative number of cases in the world after the first year from the start of a pandemic influenza with  $R_0=1.9$  originating in Hanoi (Vietnam) in the Spring.



The US maps focus on the situation in the US after one year, and show the effect of changes in the original scenario analyzed. Different color coding is used for the sake of visualization.

### Reproductive Number ( $R_0$ )



The model includes the worldwide air transportation network (source: IATA) composed of 3,100 airports in 220 countries and  $E=17,182$  direct connections, each of them associated to the corresponding passenger flow. This dataset accounts for 99% of the worldwide traffic and is complemented by the census data of each large metropolitan area served by the corresponding airport.

Additional spreading scenarios can be obtained by modeling different levels of infectiousness of the virus, as expressed in terms of the reproductive number  $R_0$ , representing the average number of infections generated by a sick person in a fully susceptible population.

Intervention strategies modeling the use of antiviral drugs can be considered. Two scenarios are compared: an uncooperative strategy in which countries only use their own stockpiles, and a cooperative intervention which envisions a limited worldwide sharing of the resources.

### Intervention

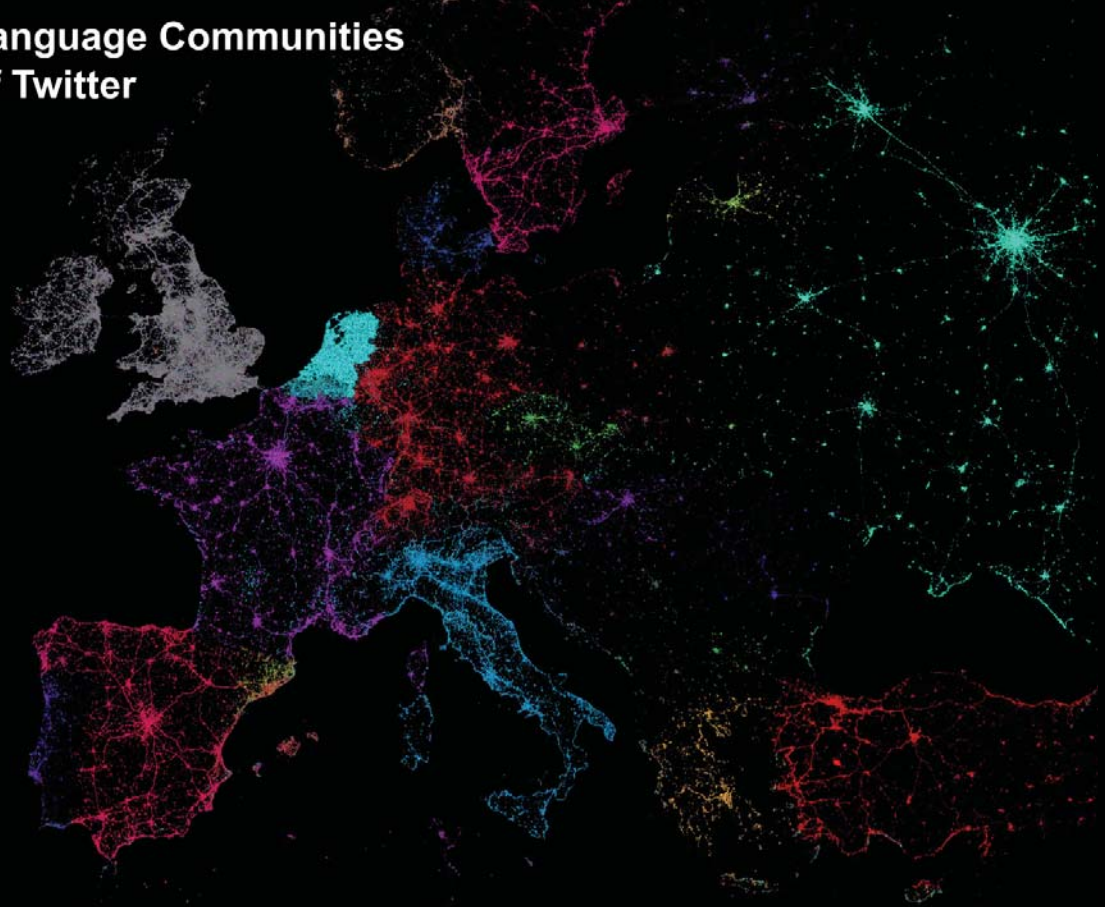






# Language Communities of Twitter

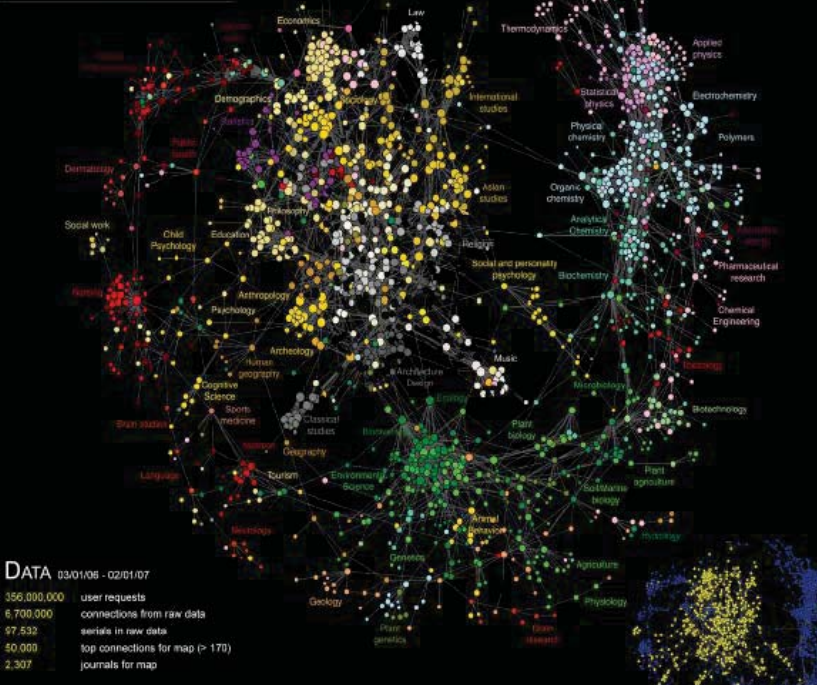
- English
- Portuguese
- Spanish
- Dutch
- Russian
- French
- Italian
- German
- Turkish
- Arabic
- Swedish
- Danish
- Finnish
- Catalan
- Romanian
- Norwegian
- Lithuanian
- Slovak
- Czech
- Greek
- Hungarian
- Polish
- Slovenian
- Albanian
- Latvian
- Galician
- Hebrew
- Croatian
- Bulgarian



Language Communities of Twitter - Eric Fischer - 2012

## LEGEND

- Physics
- Chemistry
- Biology
- Social Sciences
- Humanities



DATA 03/01/06 - 02/01/07

356,000,000	user requests
6,700,000	connections from raw data
97,532	serials in raw data
50,000	top connections for map (> 170)
2,307	journals for map



More information on this map can be found in Bollen J., Van de Sompel H., Hagberg A., Bettencourt L., Chute R., Rodriguez, MA and Balakireva, L. (2009) Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4(3): e4903. doi:10.1371/journal.pone.0049003 (Freely available online)

# CLICKSTREAM MAP OF SCIENCE

This is the first map created from large-scale, world-wide, scholarly usage data. It visualizes the collective flow of scientists' movements from one journal to another in their online navigation behavior.

The MESUR project ([www.mesur.org](http://www.mesur.org)) collected a database of nearly 1 billion user requests recorded by the web portals of some of the world's most significant publishers, aggregators and large university consortia, among them Thomson Scientific (Web of Science), Elsevier (Scopus), JSTOR, Ingenta, University of Texas (iCampus), iHealth institutions, and California State University (23 campuses). All usage logs acquired by the MESUR project contain session identifiers that identify the individual clickstreams of individual scientists navigating from one article to the next.

Pairs of journals are connected when they have a high probability of being followed by each other in users' clickstreams. The circles represent individual journals. A line between two circles indicates that they were properly connected in either direction. The colors indicate the scientific domain a journal belongs to according to their Dewey, Decennial and JCR classification codes that were mapped into the Getty Research Center's Arts and Architecture Taxonomy (AAT) to allow classifications at various levels of detail. The size of circles corresponds to the strength (degree centrality) of a journal's connections in the map. The map is arranged by the Fruchterman-Reingold algorithm that treats connections like springs; connected journals are drawn together, but they are not allowed to get too close.

This map is derived from usage data and therefore also reflects the actions of those who read the literature but rarely publish themselves, e.g. practitioners and laypersons. As a result practitioner-driven domains such as nursing, social work, and tourism studies are prominently featured. The natural sciences vs. the social sciences and humanities emerge as two distinct clusters that are connected via various topics. Interdisciplinary spokes. Most domains are highly interdisciplinary, but this is more so the case for the social sciences and humanities. Surprisingly, mathematics and computer science are not represented as one specific cluster, but spread-out through the map.

Like citation maps, this map is based upon a particular sample of the scientific community, albeit one that includes non-publishing scientists and practitioners and a much greater sample of publications. From MESUR's database of 1 billion user events, we extracted a matrix of 6 million connections between approximately 100,000 serials. From that matrix we selected only 50,000 connections with the highest number of observations, ranging from approximately 40,000 to 370 observations. This subset of connections pertained to the 2,307 most used journals. This procedure may introduce specific biases which require investigation. This map should therefore not be considered as a final map of scientific activity, but as a showcase for the feasibility of tracking scientific activity from usage data. We hope this methodology will provide unique insights into the real-time structure of scientific activity as it can be observed from scholarly clickstream data.

When we cut the AAT taxonomy at the top level, only two dimensions remain: natural science (blue nodes) vs. the social sciences and humanities (yellow nodes). Some journals along the spokes of the wheel have classifications (colors) that do not correspond to their location in the map. This indicates either the journal in question is highly interdisciplinary, and/or has been assigned a classification that does not correspond to how scientists actually use the particular journal.

Design layout by: Jeremy D. Chace

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.

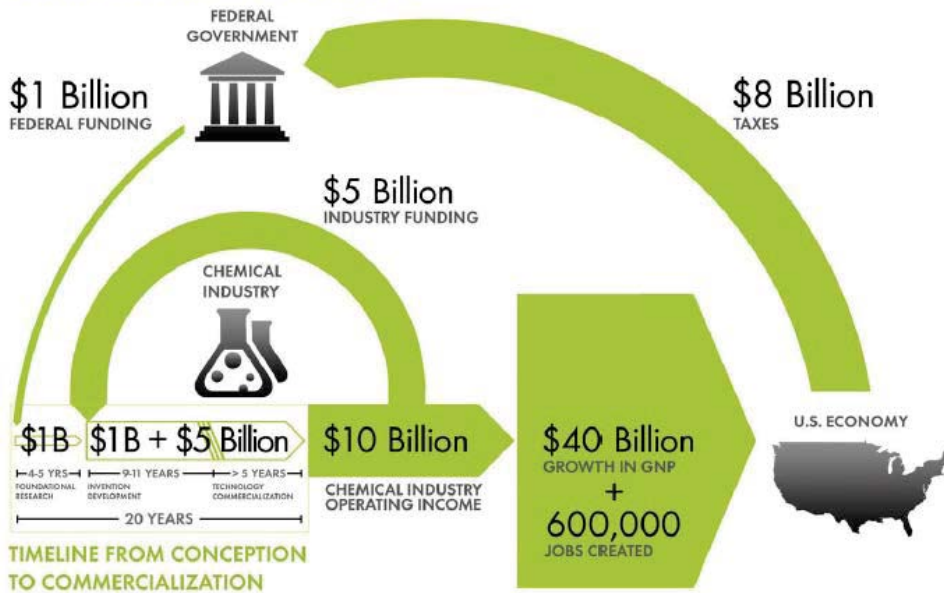
# Chemical Research & Development Powers the U.S. Innovation Engine

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



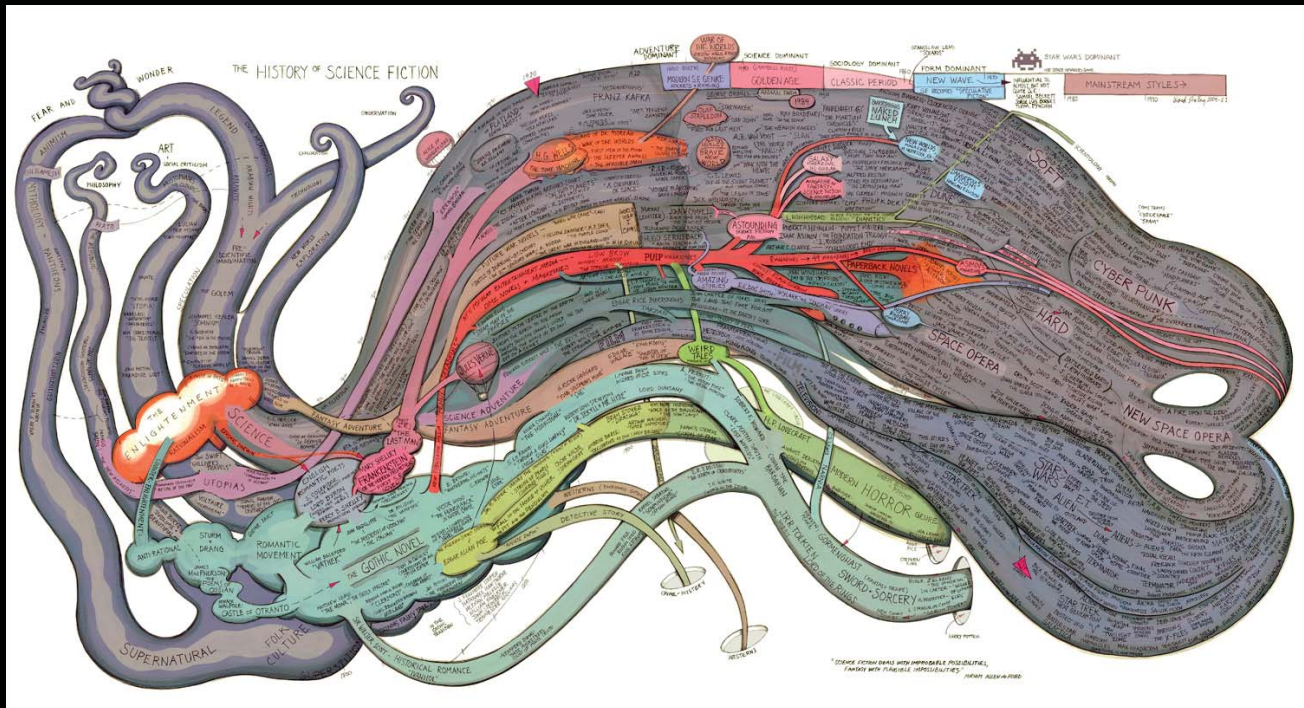
has provided the U.S. Congress and government policy makers with important results regarding the impact of Federal Research & Development (R&D) investments on U.S. innovation and global competitiveness through its commissioned 5-year two phase study. To take full advantage of typically brief access to 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.

## INVESTMENT IN CHEMICAL SCIENCE R&D



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 studies, 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.



Ward Shelley . 2011. History of Science Fiction.



Illuminated Diagram Display on display at the Smithsonian in DC. [http://scimaps.org/exhibit\\_info/#ID](http://scimaps.org/exhibit_info/#ID)

### Geographic Map: Where Science Gets Done

**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.

### Science Map: How Scientific Disciplines Relate

**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

Input your search query here.

Space      Go

**Search**

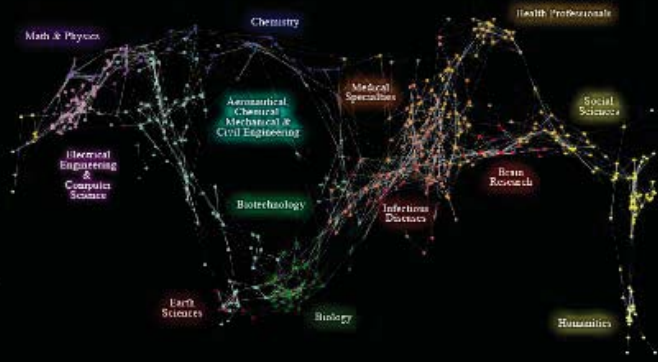
The keyboard supports retrieval and display of papers based on their Medical Subject Headings (MeSH) and MeSH qualifier terms. If multiple terms are entered in a field, they are automatically combined using "OR". So, "breast cancer" matches any record with "breast" or "cancer" in that field. You can 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.

**People & Topics**

## Geographic Map: Where Science Gets Done



## Science Map: How Scientific Disciplines Relate



Copyright © 2009 The Regents of the University of California

### 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.



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

# Open Education

# Education – K-12, college, grad., informal, professional training



“Expedition Zukunft” German science train visiting 62 cities in 7 months. 12 coaches, 300 m long.



North Carolina State’s Immersion Theater



Hidalgo, et al., 2007.  
See also The Product Space map in <http://scimaps.org>

25



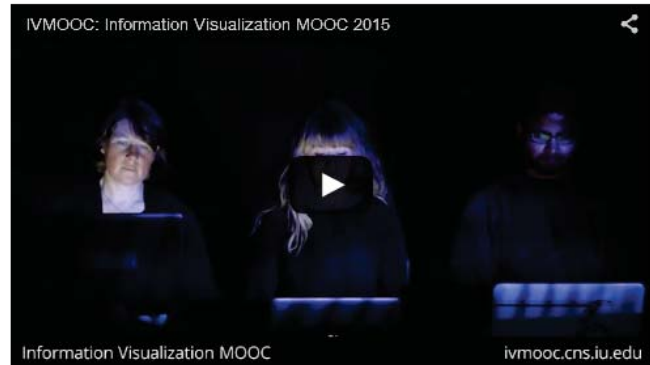
Places & Spaces Exhibit at the David J. Sencer CDC Museum, Atlanta, GA  
January 25-June 17, 2016

26

## 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.

The course can be taken for three Indiana University credits as part of the **Online Data Science Program**, as part of the **Information and Library Science M.S. program**, and as part of the online **Data Science M.S. Program** offered by the School of Informatics and Computing. Students seeking enrollment information should contact Rhonda Spencer at 812-855-2018, [ilsmain@indiana.edu](mailto:ilsmain@indiana.edu) or [datasci@indiana.edu](mailto:datasci@indiana.edu).



[Register for Course](#)

Already registered? [Click here to go to the course.](#)  
 Forgot your password? [Click here to reset it.](#)

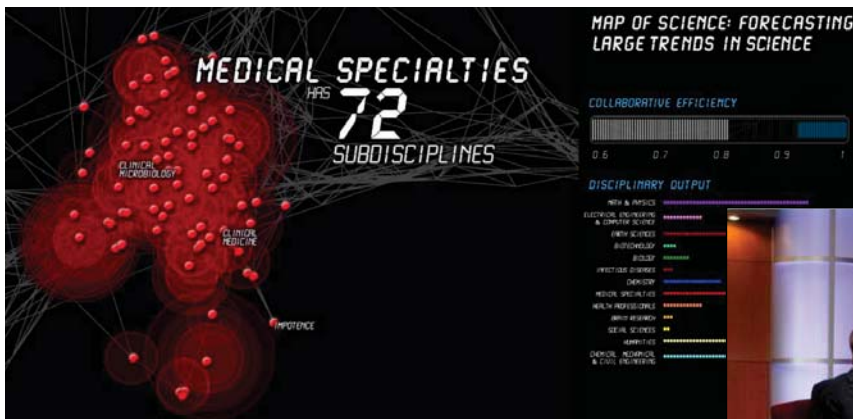
Among other topics, the course covers:

- 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.

Register for free at <http://ivmooc.cns.iu.edu>. Class restarts January 12, 2016.

50 courses available		Add	Course Name	Start Date	Rating
By start date		Ad	Introduction to Programming for the Visual Arts with p5.js <small>University of California, Los Angeles via <b>Kadenze</b></small>	4th Nov, 2015	★★★★★
Recently started or starting soon (12)		+	Data Visualization <small>University of Illinois at Urbana-Champaign via <b>Coursera</b></small>	20th Jul, 2015	★★★★☆ <span>11</span>
Just Announced (2)		+	Information Visualization <small>Indiana University via <b>Independent</b></small>	13th Jan, 2015	★★★★☆ <span>0</span>
Courses in Progress (4)		+	Data Management and Visualization <small>Wesleyan University via <b>Coursera</b></small>	26th Oct, 2015	★★★★★ <span>1</span>
Future courses (14)		+	Data Visualization and D3.js <small>via <b>Udacity</b></small>	Self paced	★★★★☆ <span>1</span>
Self Paced (9)		+	Communicating Results: Visualization, Ethics, Reproducibility <small>University of Washington via <b>Coursera</b></small>	1st Nov, 2015	★★★★☆ <span>0</span>
Finished courses (20)		+	Data Visualization and Communication with Tableau <small>Duke University via <b>Coursera</b></small>	1st Nov, 2015	★★★★☆ <span>0</span>
By subject					
Computer Science (23)					
Health & Medicine (5)					
Mathematics (4)					
Business & Management (6)					

<https://www.class-central.com/search?q=visualization>



## 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). [http://www.pnas.org/content/vol101/suppl\\_1](http://www.pnas.org/content/vol101/suppl_1)

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

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

Katy Börner, Michael Conlon, Jon Corson-Rikert, Cornell, Ying Ding (2012) **VIVO: A Semantic Approach to Scholarly Networking and Discovery**. Morgan & Claypool.

Katy Börner and David E Polley (2014) **Visual Insights: A Practical Guide to Making Sense of Data**. The MIT Press.

Börner, Katy (2015) **Atlas of Knowledge: Anyone Can Map**. The MIT Press. <http://scimaps.org/atlas2>





**CNS** Cyberinfrastructure for Network Science Center

search  Search

About Us Research Development Teaching Outreach Videos News & Events Connect With Us

We work closely with clients to provide custom-made data, visualization, and software solutions

**Research**  
Open Data and Open Code for Big Science of Science Studies

**Latest News**  
Put your money where your citations are: a proposal for a new funding system (website accessed 9/05/13)

**Upcoming Events**  
OCT 1 Katy Börner attends PIUG 2013 Northeast Conference  
10.13 Katy Börner presents Mapping Science Exhibit at WSSF  
10.15 Ted Polley & Google Team present IVMOOC at EDUCAUSE  
10.22 Katy Börner presents at the SciELO 15 Years Conference

**Development**  
Behind the scenes of the design and development of *AcademyScope*

**Outreach**  
See some of the most fascinating data visualizations in the world.

**Videos**  
Watch Katy Börner's full presentation from TEDxBloomington

**Teaching**  
Successful IVMOOC will be offered again in January of 2014

**Our Products**  
We work closely with clients to provide custom-made data, visualization, and software solutions

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

These slides are at <http://cns.iu.edu/docs/presentations>

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

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