CNS Cyberinfrastructure for Network Science Center

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

BIG Data for BIG Science of Science Research: The Value of CADRE

Katy Börner **@katycns**

Victor H. Yngve Distinguished Professor of
Intelligent Systems Engineering & Information Science
Director, Cyberinfrastructure for Network Science Center
School of Informatics, Computing, and Engineering
Indiana University, Bloomington, IN, USA
+ 2018 Humboldt Fellow, TU Dresden, Germany

CADRE Workshop at ISSI 2019

Rome, Italy | **September 2**, 2019

CNS Cyberinfrastructure for Network Science Center

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

BIG Data

Datasets used in SoS R&D

2009

Table 1. Data sets and their properties (* future feature)

Dataset	# Records	Years covered	Updated	Restricted access
Medline	13,149,741	1965–2005	Yes	
PhysRev	398,005	1893–2006		Yes
PNAS	16,167	1997–2002		Yes
JCR	59,078	1974, 1979, 1984, 1989, 1994–2004		Yes
USPTO	3,179,930	1976–2004	Yes*	
NSF	174,835	1985–2003	Yes*	
NIH	1,043,804	1972–2002	Yes*	
Total	18,021,560	1893–2006	4	3

LaRowe, Gavin, Sumeet Adinath Ambre, John W. Burgoon, Weimao Ke, and Katy Börner. 2009. "The Scholarly Database and Its Utility for Scientometrics Research". *Scientometrics* 79 (2): 219-234.

2019

WoS*

Scopus

Google Scholar

MS Academic Graph

UMETRICS

Social Media Data

News Data

IoT Data

* http://iuni.iu.edu/resources/cadre

Web of Science as a Research Dataset

November 14, 2016 - November 15, 2016 | Bloomington, Indiana

Organizers:

https://cns.iu.edu/workshops/event/161114.html

Katy Börner

Victor H. Yngve Distinguished Professor of Information Science
Director, CI for Network Science Center Curator, Mapping Science Exhibit
School of Informatics and Computing
Indiana University, Bloomington
katy@indiana.edu

Eamon Duede

Executive Director
Knowledge Lab
University of Chicago
eduede@uchicago.edu

James Pringle

Head of Industry Development
Clarivate Analytics

Session 1: Web of Science "Outside the Box"

Facilitator: Katy Börner

The Web of Science and similar metadata datasets are housed, maintained, and enhanced in local institutional

enclaves. This

the data (nam

- Matt Hu
- Yadu Ba
  institut
- Orion P
  researc
- Nichola
  Disamb
- Vetle To
- Lee Gile

Session 3: Understanding Web of Science as Research Data

Facilitator: Jason Rollins

For over 50 years, the Web of Science evolved as a dataset in response to changing research contexts and priorities. Today, more researchers are using the Web of Science "at scale" to ask and answer powerful new questions about the shape, dynamics, and veracity of science and scholarship. The Web of Science now appears both an object of inquiry its own right and a vast sensor network for discerning large-scale trends. What is changing in this dataset to support these new uses, and what could change further? Presentations and discussion led by Clarivate Analytics team.

- Jim Pringle: **"WoS Metadata as Research Data"**
- Ted Lawless: **"Web of Science Data Integration"**
- Linge Bai: **"Data Unification and Disambiguation: Institutions and Authors"**
- Sebastien Brien: "Clarivate Analytics in the Cloud: Architecture and Analytics"

Break

Session 4: Hackathon Breakout Sessions

Facilitators: Eamon Duede, Jason Rollins, and Ted Lawless

A mix of sessions determined by 3-4 "big questions" prioritized on Day 1, grouped as:

A. *Technical Hackathon(s)*: Practical Focus on applying code across research centers in such areas as data disambiguation (names, institutions, geolocations), linking WoS data to other datasets, building models to predict gender, ethnicity, etc.

B. *Topical Hackathon(s)*: Working across research centers on Authorship & Collaboration; Gender in Science; Topic Modeling and/or other topics defined by attendee interest.

C. *Community Hackathon*: Focus on establishing an ongoing community (e.g. setting up an enclave, tools & mechanisms for sharing code, citing and acknowledging contributions, and/or what is appropriate for cross-enclave sharing).

Reproducible Scientometrics Research: Open Data, Code, and Education

Date

October 17, 2017

Meeting Place

ISSI 2017, Wuhan University
Wuhan, China

Workshop Agenda

1. Introduction (Sybille Hinze & Theresa Velden)
2. Reproducibility in Scientometrics: Data Enclaves, Open Code, and Open Education (Katy Börner,
3. Reproducibility in Scientometrics through Quality Assurance (Sybille Hinze)
4. A Vendor's View on Reproducibility — Datasets, Tools, & Partnerships (Jason Rollins)
5. Reproducibility in Scientometrics — A Journal Editor's Perspective (Ludo Waltman)
6. Reproducibility —Principles and Challenges (Jesper Schneider)
7. Reproducibility & the Performativity of Methods (Theresa Velden)

Session Organizers

Sybille Hinze
DZHW
Berlin, Germany

Jason Rollins
Clarivate Analytics
San Francisco, CA, USA

Jesper Schneider
Aarhus University
Denmark

Jesper Schneider
Aarhus University
Denmark

Theresa Velden
ZTG TU Berlin
Germany

Ludo Waltman
CWTS, University of Leiden
The Netherlands

Katy Börner
Indiana University
Bloomington, IN, USA
Slides | MP4

Andrea Scharnhorst
KNAW, Amsterdam
The Netherlands

https://cns.iu.edu/workshops/event/171017.html

Web of Science™ as a Research Dataset

Katy Börner,[1] Valentin Pentchev,[2] Matthew Hutchinson,[2] James Pringle,[3] Jason Rollins,[3] Yadu N. Babuji,[4] & Eamon Duede[5]

[1]katy@indiana.edu CNS, SOIC & Network Science Institute, Indiana University, Bloomington, IN, US
[2]maahutch@iu.edu & vpentche@iu.edu IUNI, Indiana University, Bloomington, USA
[3]jason.rollins@clarivate.com & james.pringle@clarivate.com Clarivate Analytics, USA
[4]yadunand@uchicago.edu Computation Institute, Knowledge Lab, UofChicago, USA
[5]eduede@uchicago.edu Computation Institute, Knowledge Lab, Committee on the Conceptual and Historical Studies of Science, Uof Chicago, USA

Introduction

The Clarivate Analytics Web of Science (WoS) has served as a research dataset for more than 9,000 scholarly articles in the past 15 years alone—across a wide a range of fields and disciplines from toxicology to computer science to economics. Scientists and scholars have been particularly interested in the WoS citation network, a massive graph containing billions of links that can proxy the structure and dynamics of not only scholarly communication, but knowledge diffusion, the evolution of fields, and the career lifecycles of individuals and institutions. To power these investigations, scholars are increasingly employing a number of compute-intensive methodologies, sophisticated big data infrastructures, and so called collaborative "discovery science" tools and techniques. Suddenly, in addition to deep, domain specific expertise, world-class computational knowhow appears to be a new prerequisite for analysis of scholarly data at the scale represented by WoS. While cloud-based computing and tools are more prevalent and accessible than ever before, harnessing these technologies remains both a challenge and opportunity for researchers and data providers (i.e., Clarivate Analytics and similar commercial data vendors and non-commercial aggregators). While the opportunities made possible by scholarly data at the size and scope of WoS for discovery and innovation are limited only by imagination, two general prospects come readily to mind. First, access to these data coupled with the appropriate computational and analytical capabilities opens up a wide range of funding and subsequent publishing opportunities in high impact venues. Second, data providers can pursue new business opportunities, including novel data access models, new types of analytic products, and new kinds of academic/industry partnerships. In this poster paper, we briefly explore 1) the new computational infrastructures that are being developed to enable collaborative research that leverages scholarly datasets such as WoS that are both big and proprietary; 2) some recent findings that have been made possible by these infrastructures; and, 3) new commercial offerings that have been enabled and demanded in response to increasing reliance on the WoS as a research dataset.

New Computational Infrastructures

Research leveraging big, scholarly datasets like WoS presents researchers with challenges related to the data's size, inherently relational format, and sensitive (proprietary) nature. To overcome these challenges, researchers have developed a new generation of enclave supported, high performance, and cloud-based, collaborative research environments that are both elastic enough to provide substantial computational resources when needed while remaining secure enough to protect data providers' commercial interests. Moreover, these environments have evolved to allow geographically distributed researchers to collaborate on research projects in the fast paced, iterative style that has come to dominate research in the era of Big Data—namely, "discovery science".

IUNI WoS Data Enclave

The Indiana University Network Science Institute (IUNI) acquired the complete set of Clarivate Analytics' Web of Science XML raw data (Web of Knowledge version 5). The data was parsed and stored in a well-documented Postgresql database, see entity-relationship diagram, database schema, and data dictionary on http://iuni.iu.edu/resources/web-of-science. The code used to parse the WoS XML format and to save data in the Postgresql database was made available freely on GitHub, see https://github.iu.edu/CNS/generic_parser. All data can be accessed via the IUNI WoS Data Enclave, a secure repository that uses IU's Karst high-throughput computing cluster designed to deliver large amounts of processing capacity over long periods of time. Access to the XML data and the PostgreSQL database is granted to a user's Karst account. IU faculty, staff, and qualifying sponsored affiliates can request accounts on Karst to use the data for academic research and without any sharing of data. A simple web browser based query interface to the WoS dataset was implemented to support custom queries for specific terms, journals, or authors. Datasets can be downloaded in CSV data format compatible with data mining and visualization tools such as Gephi or the Sci2 Tool (http://sci2.cns.iu.edu) (Sci2 Team, 2009). More about the IUNI WoS Data Enclave can be found at http://iuni.iu.edu/resources/web-of-science.

Cloud Kotta

One platform specifically developed with WoS in mind is Knowledge Lab's Cloud Kotta (CK). CK is a secure data enclave and analytics platform that serves the research needs of social sciences (Babuji 2016). By hosting CK in the Amazon Web Services cloud, the developers were able to take advantage of virtually limitless compute, cost-effective storage and the ability to implement a fine-grained security model ensuring the authorized collaborators could access both data and compute resources from any where in the world (Babuji 2016). Moreover, CK supports multiuser, rapid ideation and research iteration through a novel Python library that enables specific functions in an analysis code, written in a Jupyter Notebook to be seamlessly and securely submitted to the CK execution fabric (Babuji 2017). By allowing researchers to develop and share analysis code interactively over secure data like WoS, CK has removed the need for deep computational infrastructure expertise. The complete WoS XML dataset was ingested into a relational database housed in CK using a custom parser that has been made freely available on GitHub (see: https://github.com/alexander-belikov/wos_parser). The Cloud Kotta WoS database schema can be found on CK's documentation pages (see: http://docs.cloudkotta.org/dataguide/wos.html). More about Cloud Kotta can be found at http://docs.cloudkotta.org.

IU International Co-Affiliation Network, 2004-2013

CNS @ Indiana University

2016

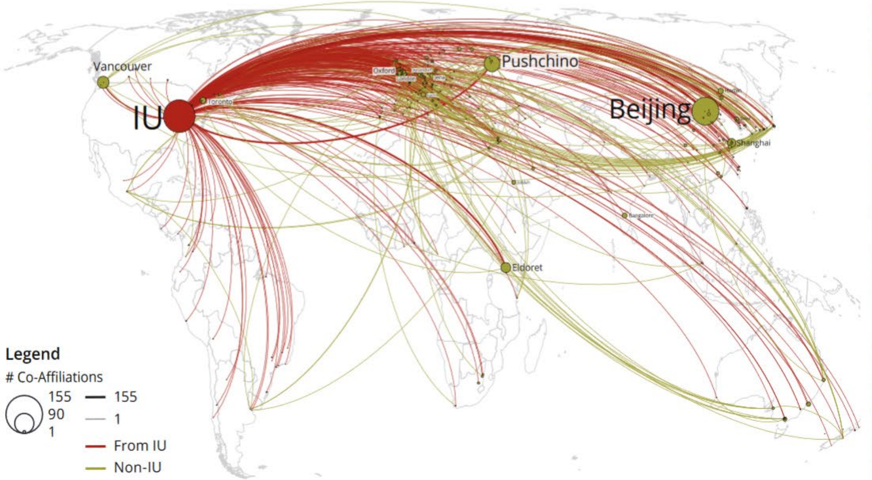


Figure 1. IU Co-Affiliation Network

New Computational Infrastructures

Fostering Global Collaboration

Among others, IU started to use the IUNI WoS data to understand existing and foster global research collaborations. The world map in Figure 1 shows the co-affiliations of authors that listed "Indiana Univ" and at least one other non-U.S. institution as affiliation on 1,590 scholarly papers published in 2004-2013. There are 344 affiliation locations (not counting IU) and 641 co-affiliation links. Nodes denote author locations and are area size coded by degree with the exception of IU, which has 1,592 co-affiliation links. Links denote co-affiliations, e.g., an author with three affiliations IU, X, Y will add three links; the two links that connect IU with X and Y are drawn in red while the link between X and Y is given in green. Links are size coded by the number of co-affiliations with the top-three being Beijing, China (155), Eldoret, Kenya (115), and Pushchino, Russia (90).

Impact vs. Disruptiveness

Researchers at the University of Chicago's Knowledge Lab and Northwestern University's NICO have used WoS data going back to 1900 to study the relationship between team size and impact and the relationship between team size and disruptiveness. This work, currently under review, finds striking differences between the scientific output of large and small teams. Looking across all fields represented in WoS, small teams are shown to disrupt science, patents, and software with new ideas and opportunities, while large teams contribute to existing ones. Figure 2 shows the relationship between impact and disruptiveness of articles (left panel) indexed by WoS, patents (middle), and software (right). In all three spaces, there is a strong, inverse relationship between citations and disruptiveness as team size increases.

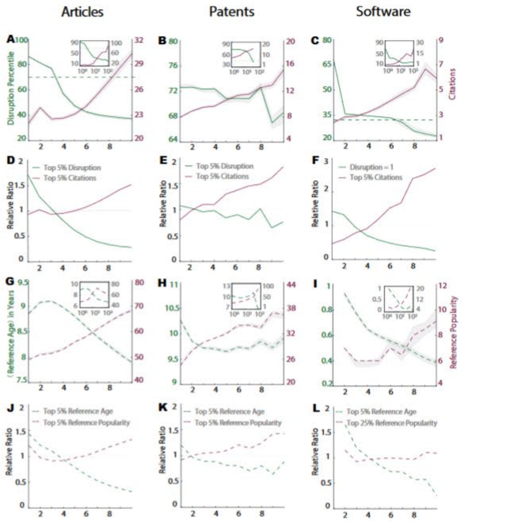


Figure 2. Tracing Inventive Teams

New Commercial Offerings

The value of the Web of Science as a search and discovery tool is well established at thousands of research institutions worldwide. But the commercial opportunities for the use of its high-quality metadata outside of the platform for big data studies are still emerging. When researchers need to study broad-scale trends in science, technology, and innovation, they very often turn to the Web of Science as the most comprehensive citation source to provide over 100 years of consistent, global publication data. Increasingly, user requests for this data take the form of custom reports, curated data sub-sets, and large-scale raw XML delivery. Clarivate Analytics is actively looking at compelling ways to meet these customer demands with new commercial products and data delivery choices. These options must balance scale and ease of use, with security and control over access to the proprietary WoS data. The lessons learned in the development of Cloud Kotta and IUNI WoS Data Enclave will very likely be instructive here, as they have proven their utility and leverage a mix of custom code built on proven commercial cloud services. Both self-service data access and secure use of analytical tools in a cloud "sandbox" seem like attractive features of these environments that could make commercial sense to meet the evolving expectation of Web of Science customers.

Acknowledgements

This work as partially supported by and contributes to research for IBM, Facebook, Jump Trading, AWS, Clarivate Analytics, the National Institutes of Health under awards P01 AG039347 and U01CA198934 and the National Science Foundation under awards NCSE-1538763, EAGER 1566393, and NCN CP Supplement 1553044. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work uses Web of Science data by Clarivate Analytics.

References

Babuji, Y. N., Chard, K., Gerow, A., & Duede, E. (2016). Cloud Kotta: Enabling Secure and Scalable Data Analytics in the Cloud. *IEEE Big Data 2016.*
Babuji, Y. N., Chard, K., Gerow, A., & Duede, E. (2016). A Secure Data Enclave and Analytics Platform for Social Scientists. *IEEE eScience 2016.*
Babuji, Y. N., Chard, K., & Duede, E. (2017). Enabling Interactive Analytics of Secure Data using Cloud Kotta. *Science Cloud Workshop: ACM International Symposium on High-Performance Parallel and Distributed Computing 2017* (Forthcoming)
Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, http://sci2.cns.iu.edu.

Börner, Katy, Valentin Pentchev, Matthew Hutchinson, James Pringle, Jason Rollins, Yadu N. Babuji, and Eamon Duede. 2017. "Web of Science™ as a Research Dataset". *16th International Conference on Scientometrics and Informetrics, Wuhan, China.*

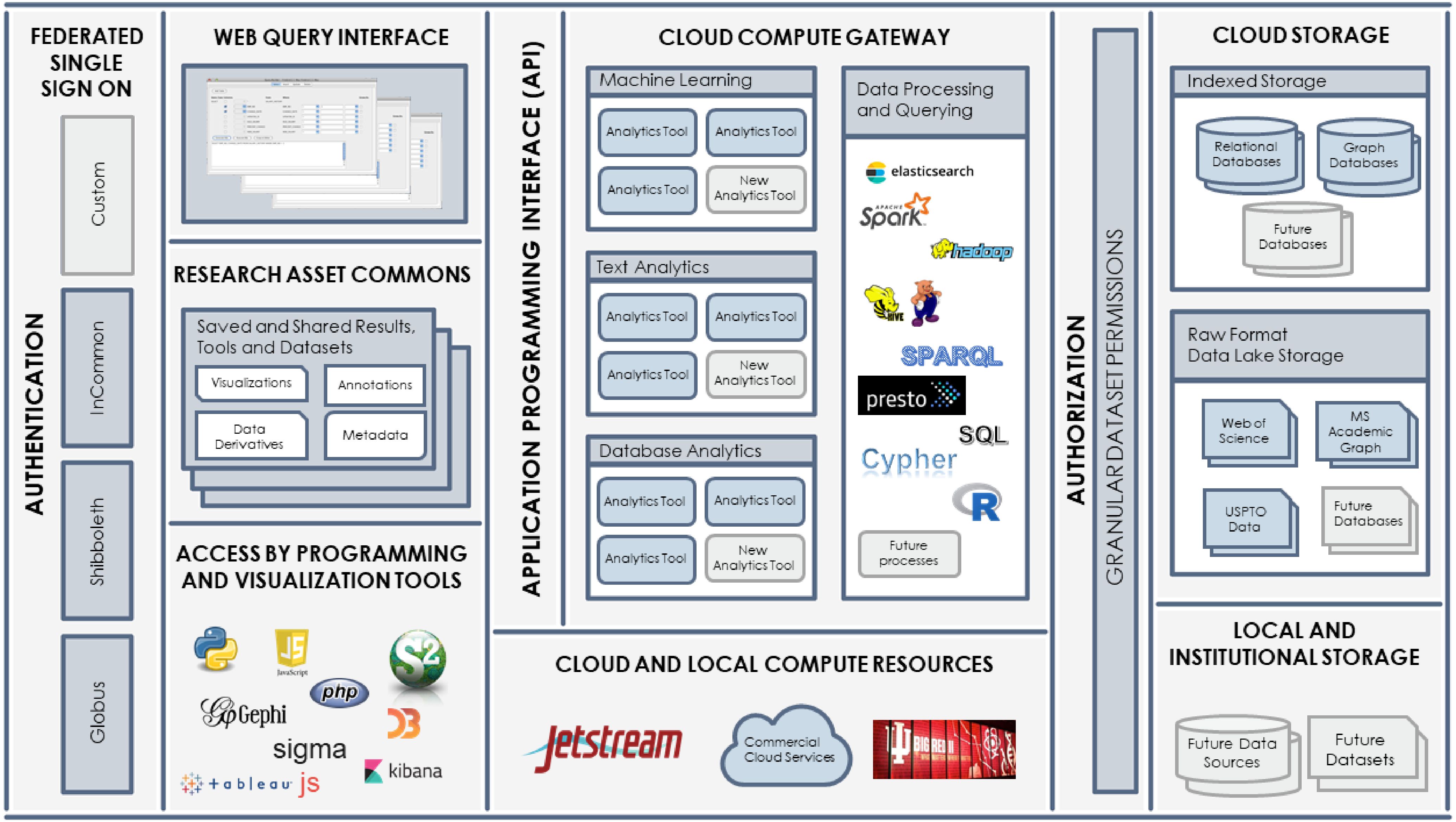
IUNI WoS Data Enclave

The Indiana University Network Science Institute (IUNI) acquired the complete set of Clarivate Analytics' Web of Science XML raw data (Web of Knowledge version 5). The data was parsed and stored in a well-documented Postgresql database, see entity-relationship diagram, database schema, and data dictionary on http://iuni.iu.edu/resources/web-of-science. The code used to parse the WoS XML format and to save data in the Postgresql database was made available freely on GitHub, see **https://github.iu.edu/CNS/generic_parser**. All data can be accessed via the IUNI WoS Data Enclave, a secure repository that uses IU's Karst high-throughput computing cluster designed to deliver large amounts of processing capacity over long periods of time. Access to the XML data and the PostgreSQL database is granted to a user's Karst account. IU faculty, staff, and qualifying sponsored affiliates can request accounts on Karst to use the data for academic research and without any sharing of data. A simple web browser based query interface to the WoS dataset was implemented to support custom queries for specific terms, journals, or authors. Datasets can be downloaded in CSV data format compatible with data mining and visualization tools such as Gephi or the Sci2 Tool (http://sci2.cns.iu.edu) (Sci2 Team, 2009). More about the IUNI WoS Data Enclave can be found at **http://iuni.iu.edu/resources/web-of-science**.

https://github.com/lightr/generic_parser

Börner, Katy, Valentin Pentchev, Matthew Hutchinson, James Pringle, Jason Rollins, Yadu N. Babuji, and Eamon Duede. 2017. "Web of Science™ as a Research Dataset". *16th International Conference on Scientometrics and Informetrics, Wuhan, China*.

SHARED BIGDATA-GATEWAY FOR RESEARCH LIBRARIES (SBD-GATEWAY)



http://iuni.iu.edu/resources/cadre

CNS Cyberinfra Network Science Center

BIG Science of Science R&D

Maps of Science & Technology

http://scimaps.org

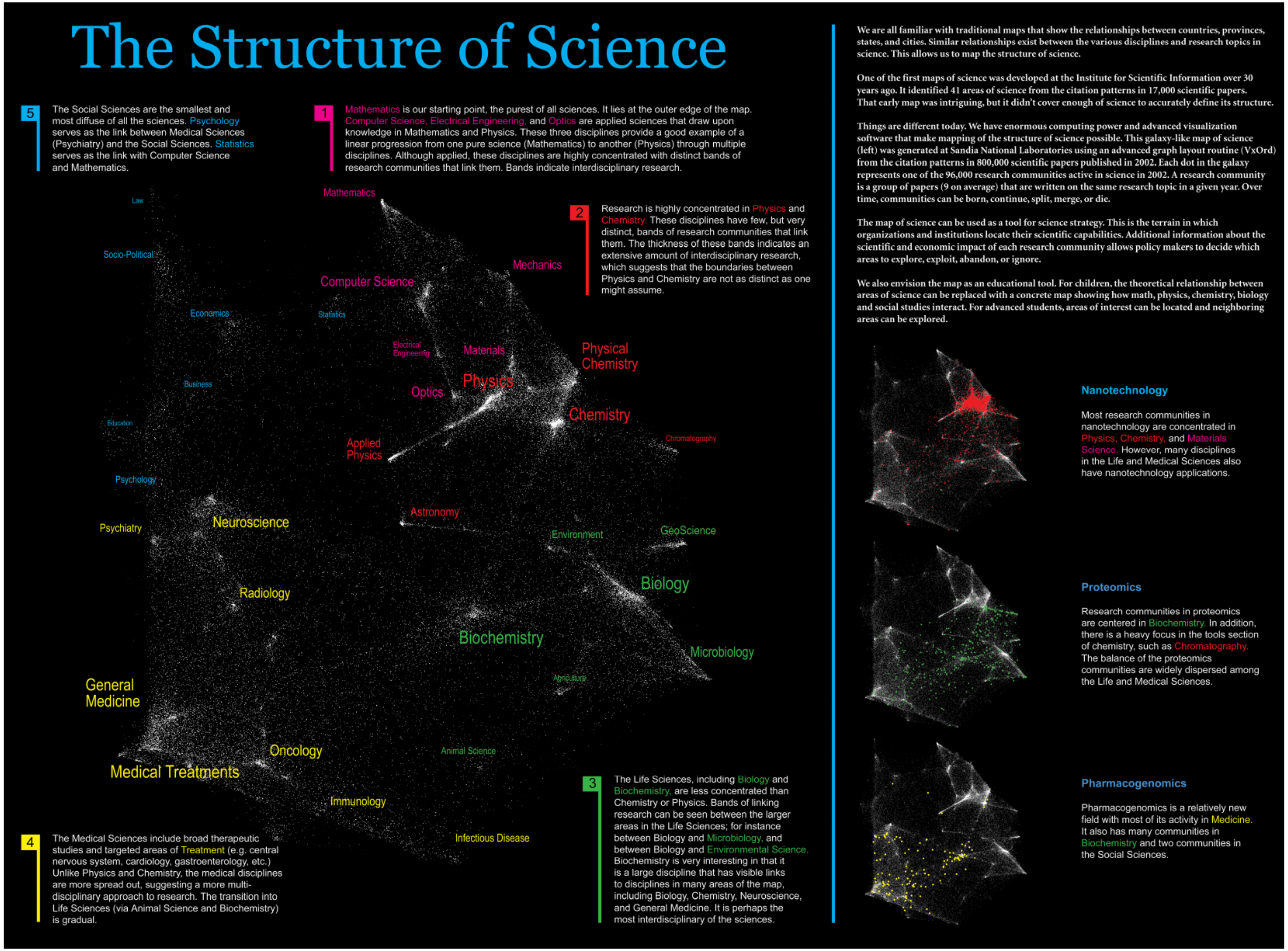
101st Annual Meeting of the Association of American Geographers, Denver, CO.
April 5th - 9th, 2005 (First showing of Places & Spaces)

University of Miami, Miami, FL.
September 4 - December 11, 2014.

Duke University, Durham, NC.
January 12 - April 10, 2015

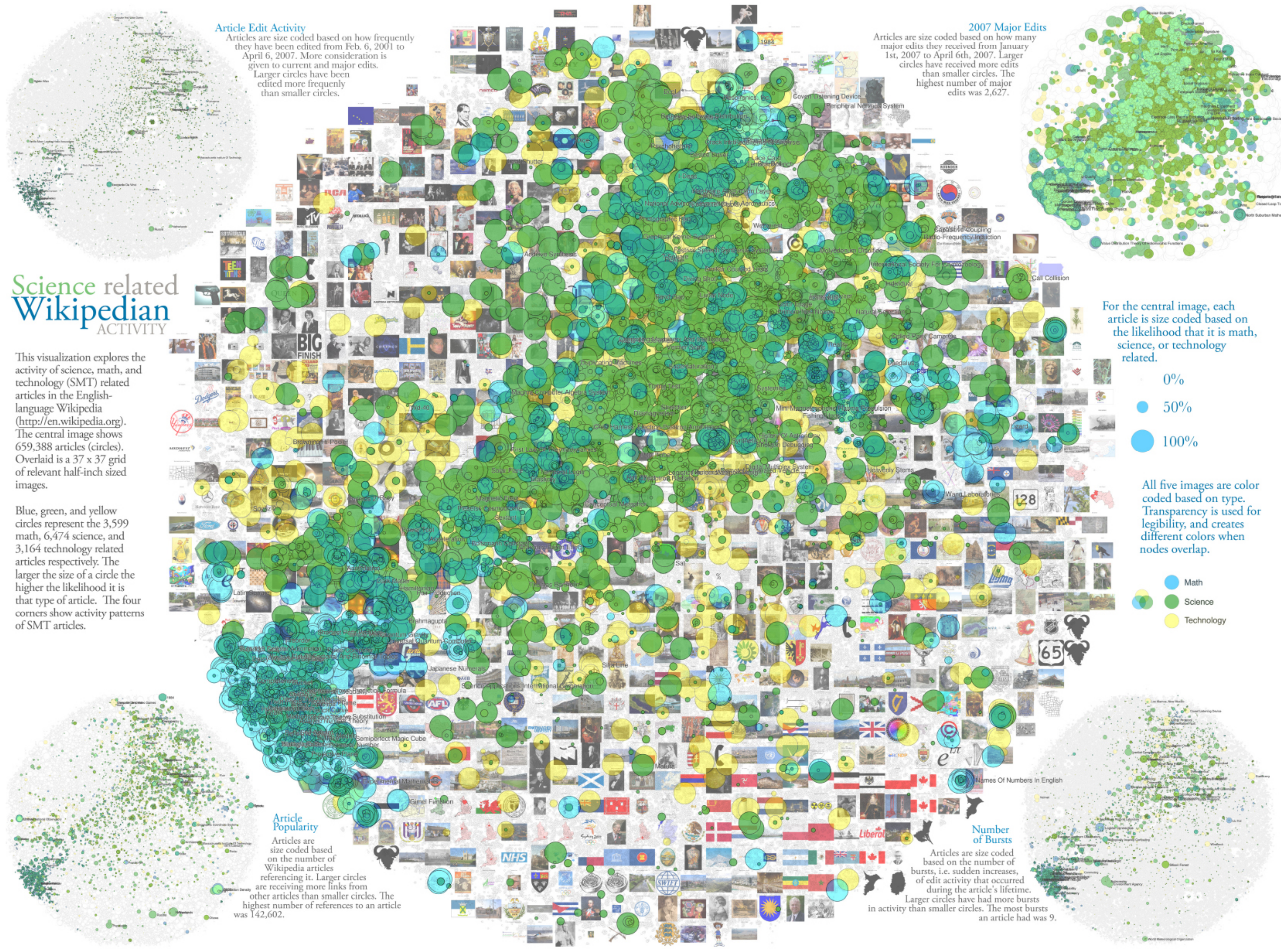
The David J. Sencer CDC Museum, Atlanta, GA.
January 25 - June 17, 2016.

100 maps and 20 macroscopes by 250+ experts on display at 350+ venues in 28 countries.



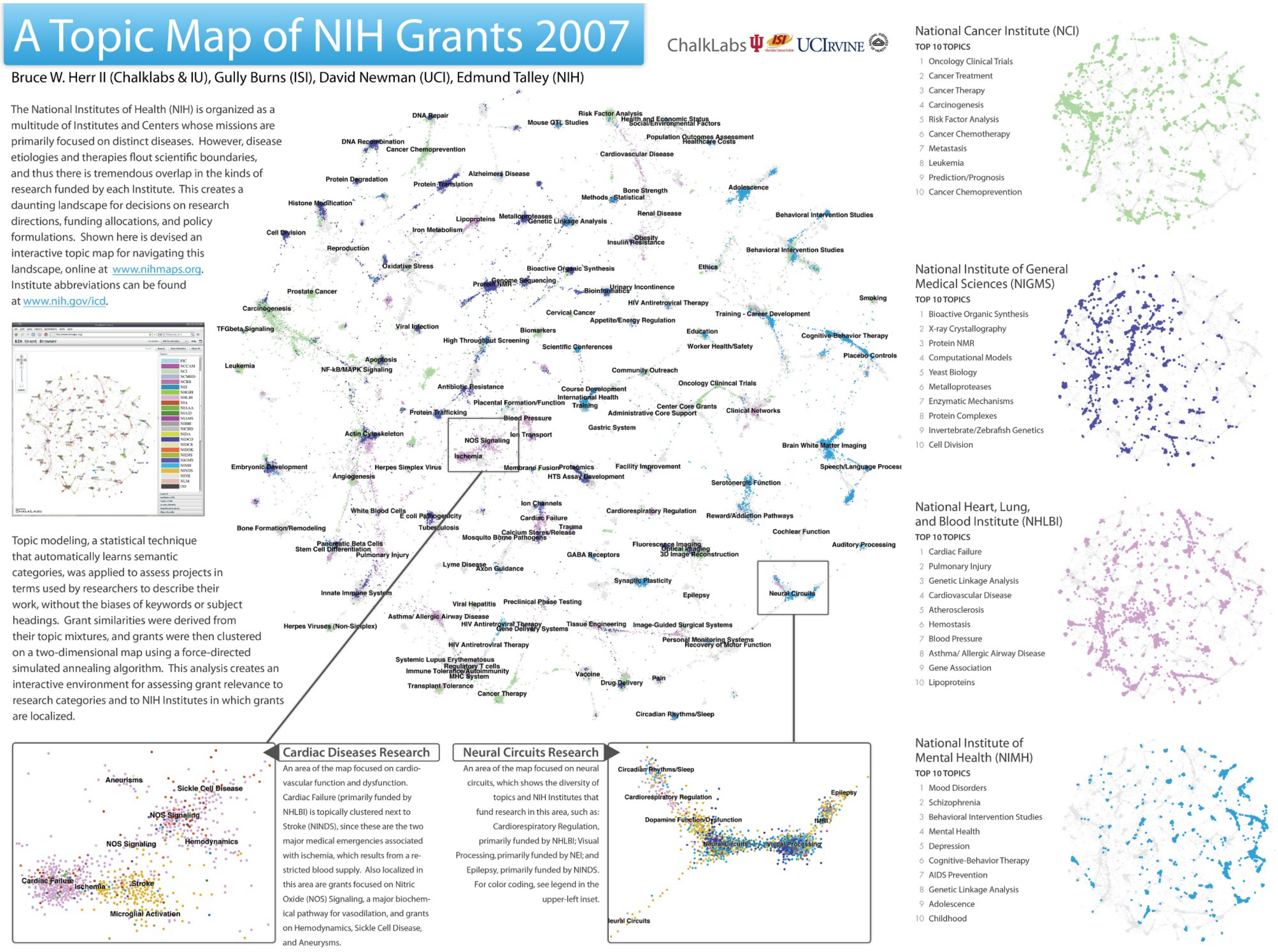
I.10 *The Structure of Science* - Kevin W. Boyack and Richard Klavans - 2005

http://scimaps.org



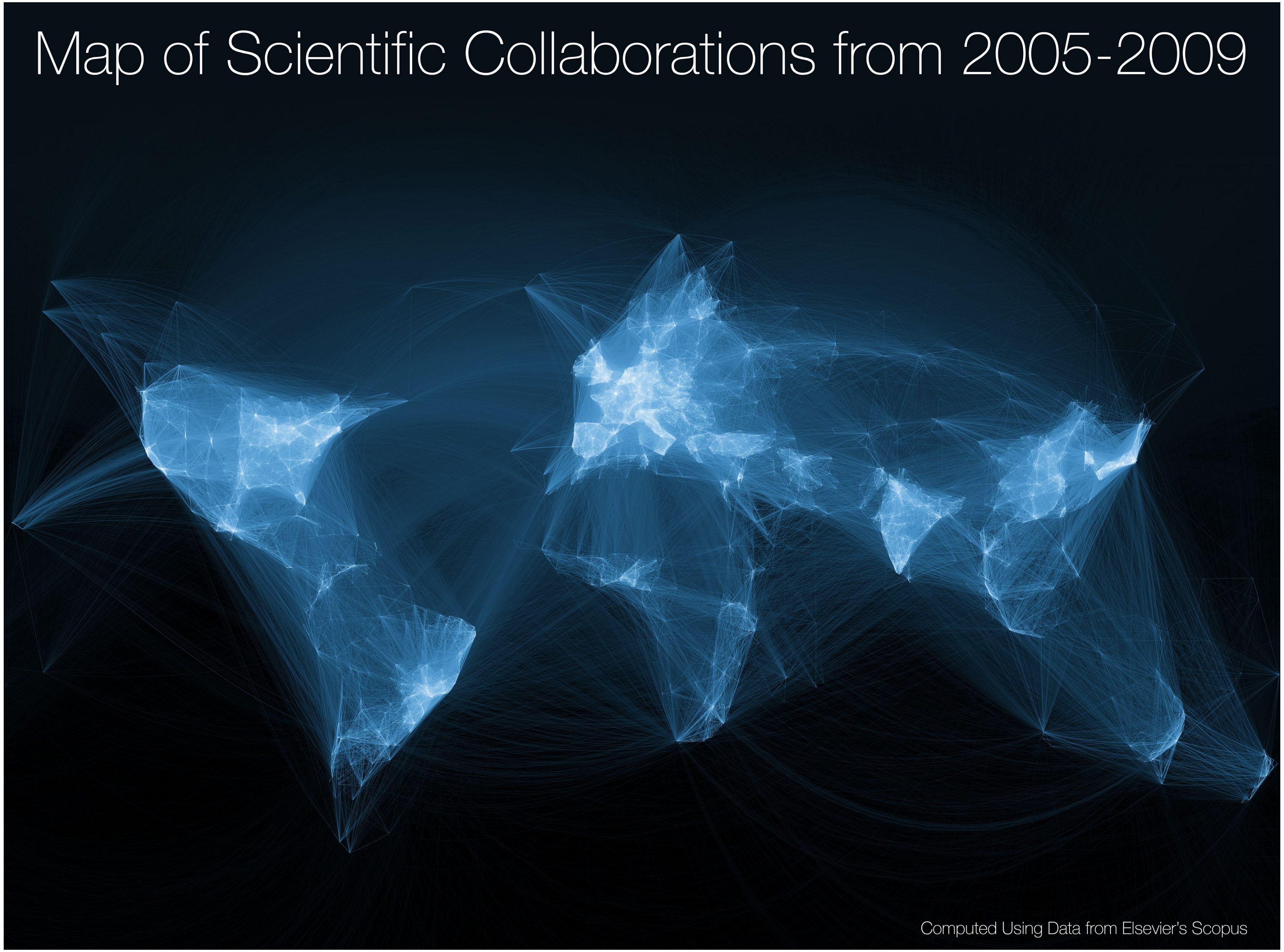
III.8 *Science-Related Wikipedian Activity* - Bruce W. Herr II, Todd M. Holloway, Elisha F. Hardy, Katy Börner, and Kevin Boyack - 2007

http://scimaps.org



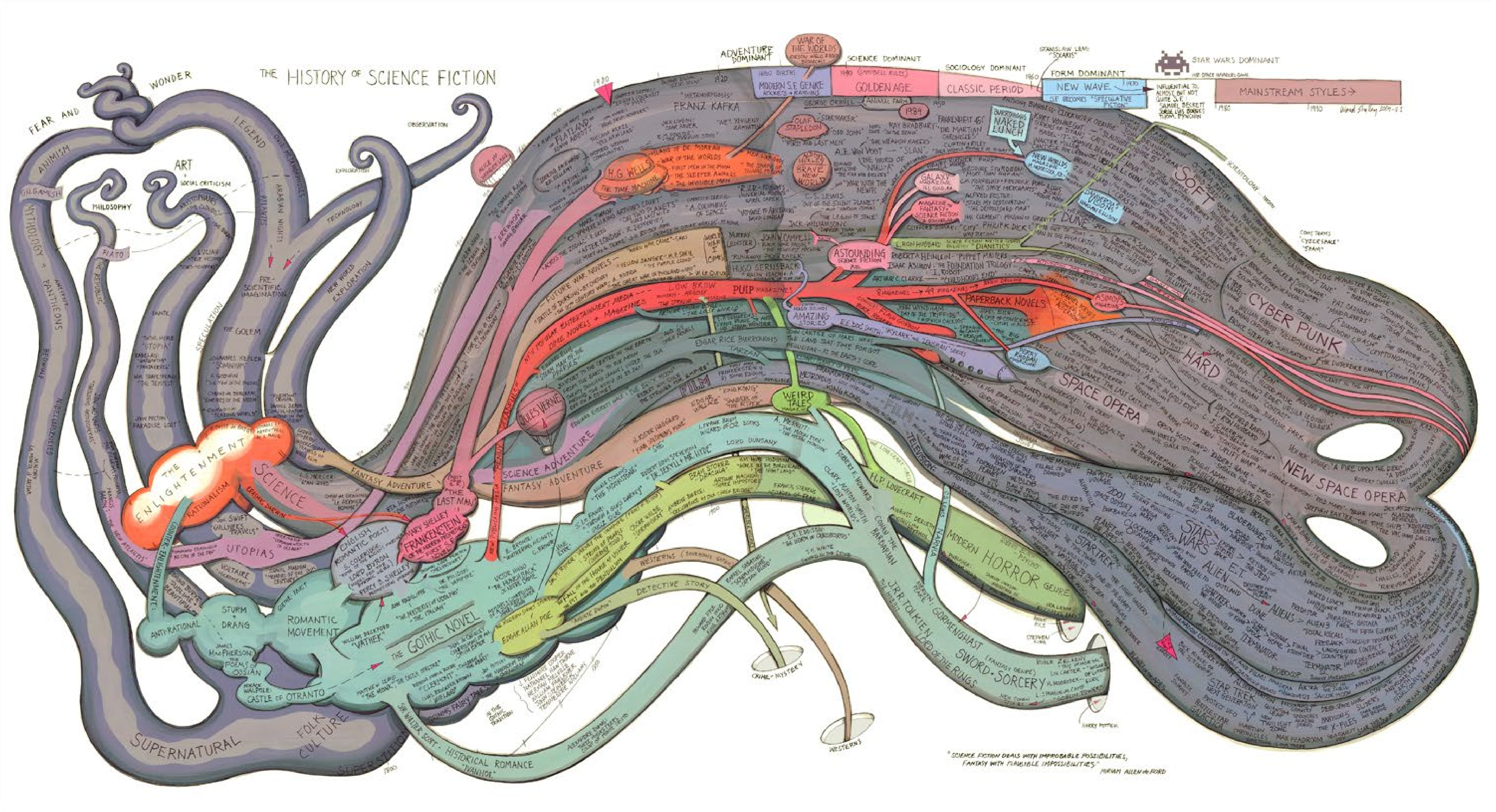
V.7 *A Topic Map of NIH Grants 2007* - Bruce W. Herr II, Gully A.P.C. Burns, David Newman, and Edmund Talley - 2009

http://scimaps.org



VII.6 Stream of Scientific Collaborations Between World Cities - Olivier H. Beauchesne - 2012

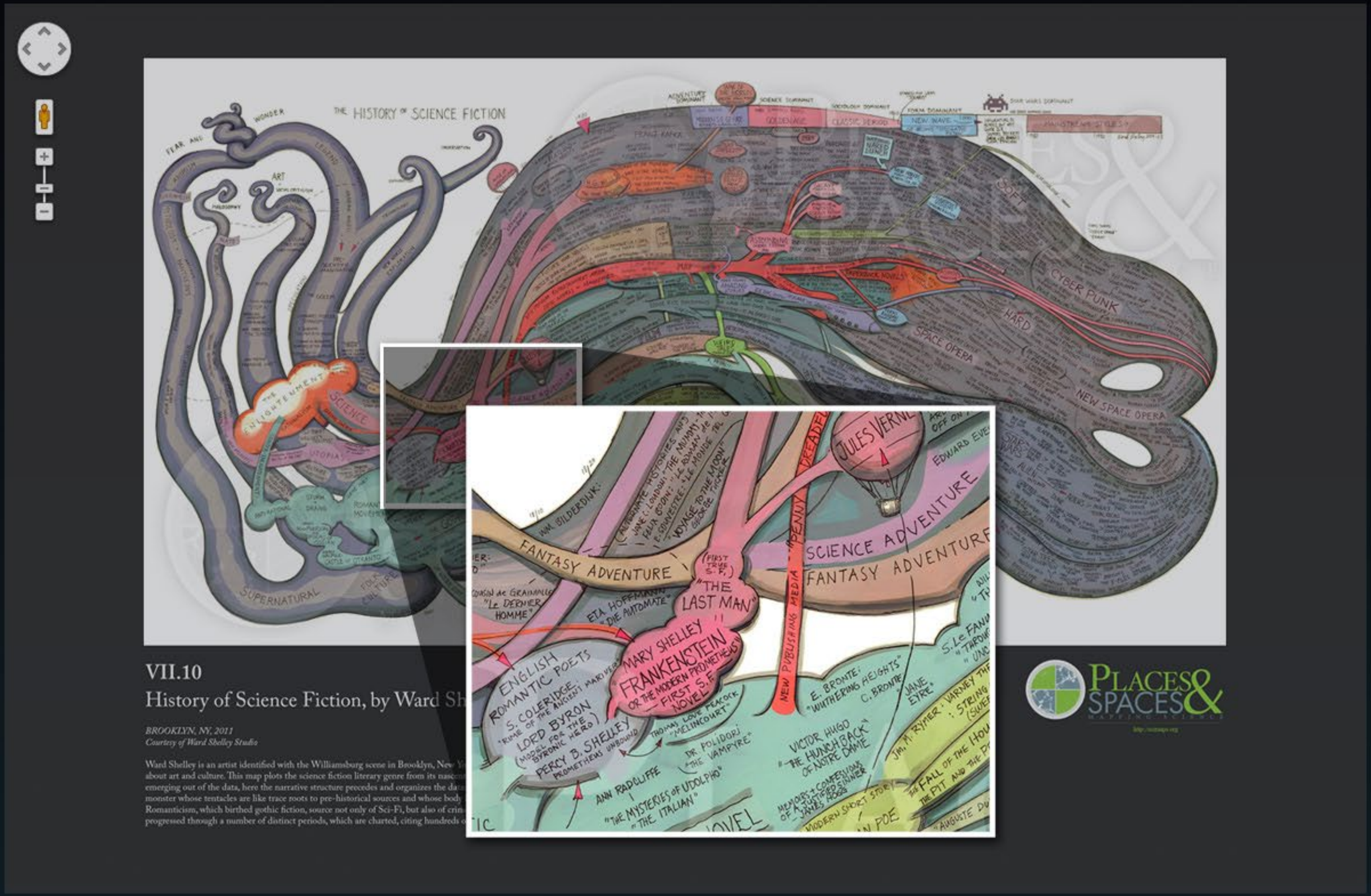
http://scimaps.org



VII.10 *History of Science Fiction* - Ward Shelley - 2011

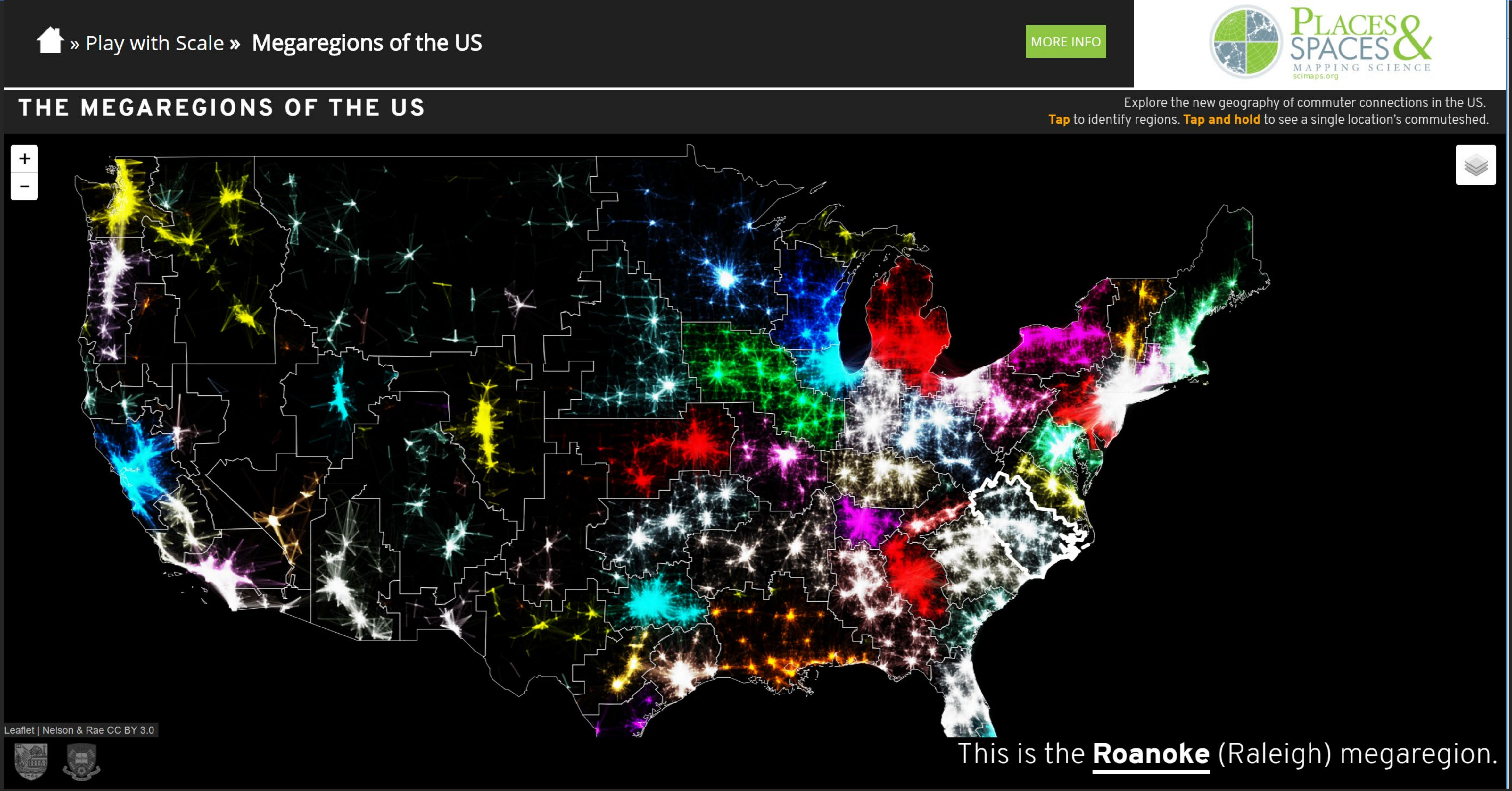
http://scimaps.org

Check out our Zoom Maps online!



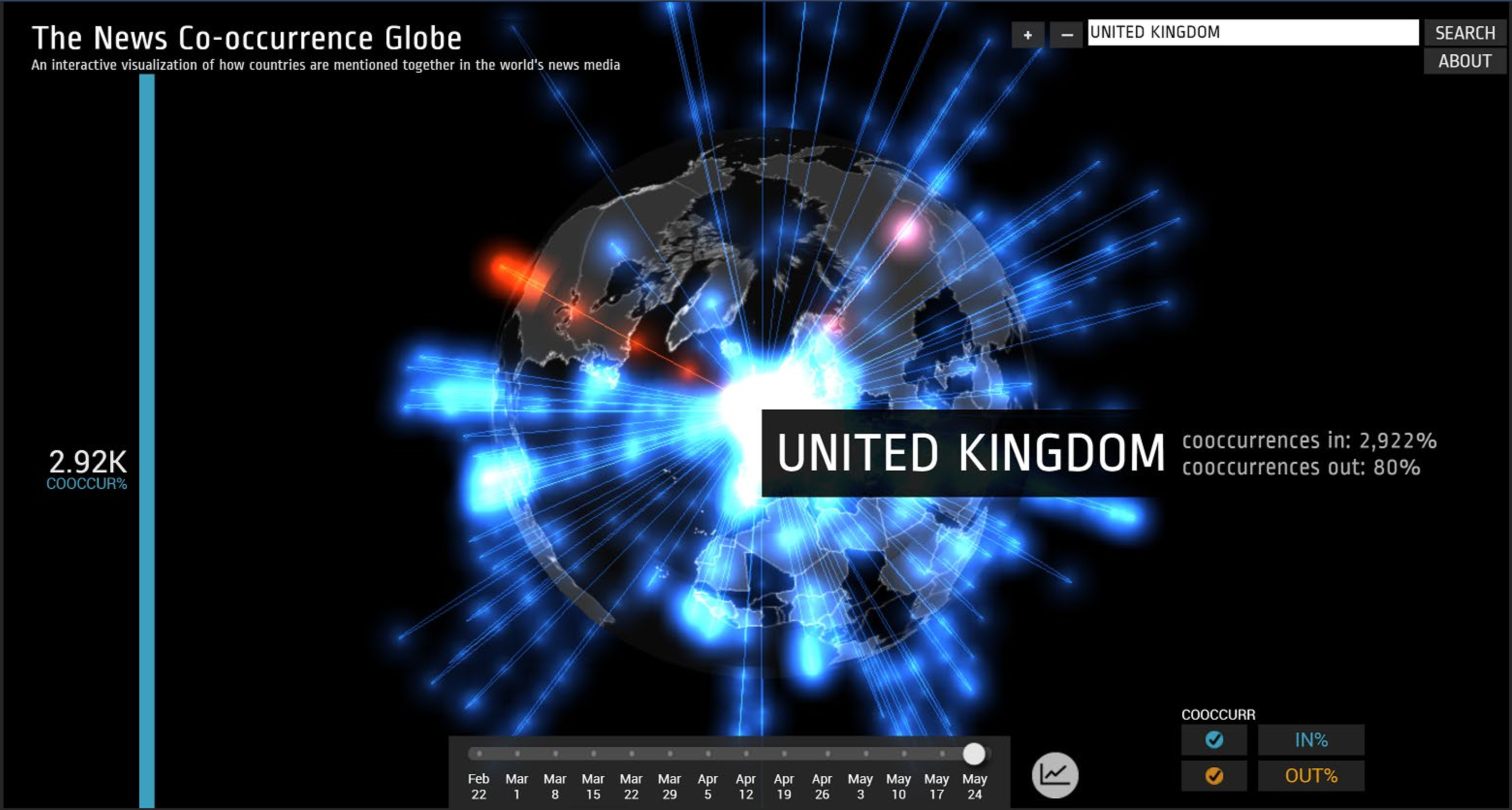
Visit scimaps.org and check out all our maps in stunning detail!

http://idemo.cns.iu.edu/macroscope-kiosk



Megaregions of the US –Garrett Dash Nelson and Alasdair Rae – 2016

Smelly Maps – Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello – 2015



Mapping Global Society –Kalev Leetaru

NSF

NETE FEDERAL IT

THOMSON REUTERS

IU NI INDIANA UNIVERSITY NETWORK SCIENCE INSTITUTE

Government, academic, and industry leaders discussed challenges and opportunities associated with using big data, visual analytics, and computational models in STI decision-making.

Conference slides, recordings, and report are available via http://modsti.cns.iu.edu/report

Arthur M. Sackler
COLLOQUIA
OF THE NATIONAL ACADEMY OF SCIENCES

Modeling and Visualizing Science and Technology Developments

National Academy of Sciences Sackler Colloquium, December 4-5, 2017, Irvine, CA

Rankings and the Efficiency of Institutions

H. Eugene Stanley | Albert-László Barabási | Lada Adamic | Marta González | Kaye Husbands Fealing | Brian Uzzi | John V. Lombardi

Higher Education and the Science & Technology Job Market

Katy Börner | Wendy L. Martinez | Michael Richey | William Rouse | Stasa Milojevic | Rob Rubin | David Krakauer

Innovation Diffusion and Technology Adoption

William Rouse | Donna Cox | Jeff Alstott | Ben Shneiderman | Rahul C. Basole | Scott Stern | Cesar Hidalgo

Modeling Needs, Infrastructures, Standards

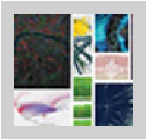
Paul Trunfio | Sallie Keller | Andrew L. Russell | Guru Madhavan | Azer Bestavros | Jason Owen-Smith

nasonline.org/Sackler-Visualizing-Science

Arthur M. Sackler COLLOQUIA OF THE NATIONAL ACADEMY OF SCIENCES | Clarivate | Microsoft

Arthur M. Sackler COLLOQUIA

Modeling and Visualizing Science and Technology Developments



December 4-5, 2017; Irvine, CA
Organized by Katy Börner, H. Eugene Stanley, William Rouse and Paul Trunfio

Overview

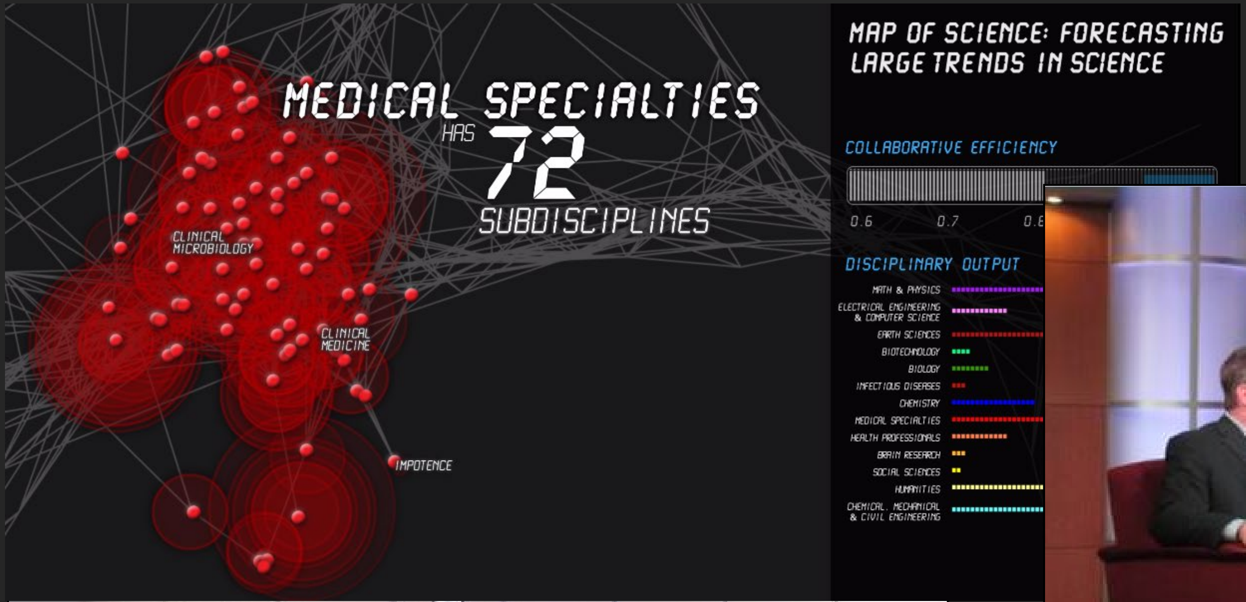
This colloquium was held in Irvine, CA on December 4-5, 2017.

This colloquium brought together researchers and practitioners from multiple disciplines to present, discuss, and advance computational models and visualizations of science and technology (S&T). Existing computational models are being applied by academia, government, and industry to explore questions such as: What jobs will exist in ten years and what career paths lead to success? Which types of institutions will likely be most innovative in the future? How will the higher education cost bubble burst affect these institutions? What funding strategies have the highest return on investment? How will changing demographics, alternative economic growth trajectories, and relationships among nations impact answers to these and other questions? Large-scale datasets (e.g., publications, patents, funding, clinical trials, stock market, social media data) can now be utilized to simulate the structure and evolution of S&T. Advances in computational power have created the possibility of implementing scalable, empirically validated computational models. However, because the databases are massive and multidimensional, both the data and the models tend to exceed human comprehension. How can advances in data visualizations be effectively employed to communicate the data, the models, and the model results to diverse stakeholder groups? Who will be the users of next generation models and visualizations and what decisions will they be addressing.

Videos of the talks are available on the **Sackler YouTube Channel**.

https://www.pnas.org/modeling

Science Forecast S1:E1



https://www.youtube.com/watch?v=lByX2_eb_QQ

PNAS Proceedings of the National Academy of Sciences of the United States of America

Arthur M. Sackler Colloquium on Modeling and Visualizing Science and Technology Developments

Twin-Win Model: A human-centered approach to research success

Ben Shneiderman

PNAS December 11, 2018 115 (50) 12590-12594; first published December 10, 2018. https://doi.org/10.1073/pnas.1802918115

Forecasting innovations in science, technology, and education

FROM THE COVER

Katy Börner, William B. Rouse, Paul Trunfio, and H. Eugene Stanley

PNAS December 11, 2018 115 (50) 12573-12581; first published December 10, 2018. https://doi.org/10.1073/pnas.1818750115

How science and technology developments impact employment and education

Wendy Martinez

PNAS December 11, 2018 115 (50) 12624-12629; first published December 10, 2018. https://doi.org/10.1073/pnas.1803216115

Scientific prize network predicts who pushes the boundaries of science

Yifang Ma and Brian Uzzi

PNAS December 11, 2018 115 (50) 12608-12615; first published December 10, 2018. https://doi.org/10.1073/pnas.1800485115

The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms

C. Jara-Figueroa, Bogang Jun, Edward L. Glaeser, and Cesar A. Hidalgo

PNAS December 11, 2018 115 (50) 12646-12653; first published December 10, 2018. https://doi.org/10.1073/pnas.1800475115

PNAS
Proceedings of the
National Academy of Sciences
of the United States of America

Keyword, Author, or DOI

Advanced Search

Arthur M. Sackler Colloquium on Modeling and Visualizing Science and Technology Developments

Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy

Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans

PNAS December 11, 2018 115 (50) 12630-12637; first published December 10, 2018. https://doi.org/10.1073/pnas.1804247115

Changing demographics of scientific careers: The rise of the temporary workforce

Staša Milojević, Filippo Radicchi, and John P. Walsh

PNAS December 11, 2018 115 (50) 12616-12623; first published December 10, 2018. https://doi.org/10.1073/pnas.1800478115

The chaperone effect in scientific publishing

Vedran Sekara, Pierre Deville, Sebastian E. Ahnert, Albert-László Barabási, Roberta Sinatra, and Sune Lehmann

PNAS December 11, 2018 115 (50) 12603-12607; first published December 10, 2018. https://doi.org/10.1073/pnas.1800471115

Modeling research universities: Predicting probable futures of public vs. private and large vs. small research universities

William B. Rouse, John V. Lombardi, and Diane D. Craig

PNAS December 11, 2018 115 (50) 12582-12589; first published December 10, 2018. https://doi.org/10.1073/pnas.1807174115

and more …

Skill Discrepancies Between Research, Education, and Jobs Reveal the Critical Need to Supply Soft Skills for the Data Economy

- Data and Crosswalks
- MaxMatch for NLP
- Causal Analyses
- Visualizations

Börner, Katy, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wue, and James A. Evans. 2018. "Skill Discrepancies Between Research, Education, and Jobs Reveal the Critical Need to Supply Soft Skills for the Data Economy." *PNAS* 115(50): 12630-12637.

Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy

Katy Börner[a,b,1], Olga Scrivner[a], Mike Gallant[a], Shutian Ma[a,c], Xiaozhong Liu[a], Keith Chewning[d], Lingfei Wu[e,f,g,h], and James A. Evans[f,g,i,1]

[a]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408; [b]Educational Technology/Media Centre, Dresden University of Technology, 01062 Dresden, Germany; [c]Department of Information Management, Nanjing University of Science and Technology, 210094 Nanjing, China; [d]Burning Glass Technologies, Boston, MA 02110; [e]School of Journalism and Communication, Nanjing University, 210008 Nanjing, China; [f]Department of Sociology, University of Chicago, Chicago, IL 60637; [g]Knowledge Lab, University of Chicago, Chicago, IL 60637; [h]Tencent Research Institute, 100080 Beijing, China; and [i]Santa Fe Institute, Santa Fe, NM 87501

Edited by William B. Rouse, Stevens Institute of Technology, Hoboken, NJ, and accepted by Editorial Board Member Pablo G. Debenedetti September 12, 2018 (received for review March 14, 2018)

Rapid research progress in science and technology (S&T) and continuously shifting workforce needs exert pressure on each other and on the educational and training systems that link them. Higher education institutions aim to equip new generations of students with skills and expertise relevant to workforce participation for decades to come, but their offerings sometimes misalign with commercial needs and new techniques forged at the frontiers of research. Here, we analyze and visualize the dynamic skill (mis-)alignment between academic push, industry pull, and educational offerings, paying special attention to the rapidly emerging areas of data science and data engineering (DS/DE). The visualizations and computational models presented here can help key decision makers understand the evolving structure of skills so that they can craft educational programs that serve workforce needs. Our study uses millions of publications, course syllabi, and job advertisements published between 2010 and 2016. We show how courses mediate between research and jobs. We also discover responsiveness in the academic, educational, and industrial system in how skill demands from industry are as likely to drive skill attention in research as the converse. Finally, we reveal the increasing importance of uniquely human skills, such as communication, negotiation, and persuasion. These skills are currently underexamined in research and undersupplied through education for the labor market. In an increasingly data-driven economy, the demand for "soft" social skills, like teamwork and communication, increase with greater demand for "hard" technical skills and tools.

science of science | job market | data mining | visualization | market gap analysis

Education has been a critical vehicle of economic growth and social progress throughout the modern era. Higher education doors. Some predictions say hundreds or even thousands of colleges and universities will close or merge in the coming years (4).

In addition, there seem to be major discrepancies and delays between leading scientific research, job market needs, and educational content. This has been particularly expressed with respect to science, technology, engineering, and mathematics jobs, where scientific and technological progress is rapid. Strategic decision making on what to teach, whom to hire, and what new research to fund benefits from a systematic analysis of the interplay between science and technology (S&T) developments, courses and degrees offered, and job market needs. Specifically, stakeholders in US higher education urgently need answers to the following questions. (*i*) Students: what jobs might exist in 5–10 years? What educational trajectories will best achieve my dream job? What core and specialized skills are required for what jobs and offered by what schools and programs? (*ii*) Teachers: what course updates are most needed? What balance of timely vs. timeless knowledge should I teach? How can I innovate in teaching and maintain job security or tenure? (*iii*) Universities: what programs should be created? What is my competition doing? How do I tailor programs to fit workforce needs? (*iv*) Science funders: how can S&T investments improve short- and long-term prosperity? Where will advances in knowledge also yield advances in skills and technology (5)? (*v*) Employers: what skills are needed next year and in 5 and 10 years? Which institutions produce the right talent? What skills are listed in job advertisements by my competition? How do I hire and train

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Modeling and Visualizing Science and Technology Developments," held December 4–5, 2017, at the Arnold and Mabel Beckman Center of the National Academies of

CNS Cyberinfrastructure for Network Science Center

Study the **(mis)match** and **temporal dynamics** of S&T progress, education and workforce development options, and job requirements.

Challenges:

- Rapid change of STEM knowledge
- Increase in tools, AI
- Social skills (project management, team leadership)
- Increasing team size

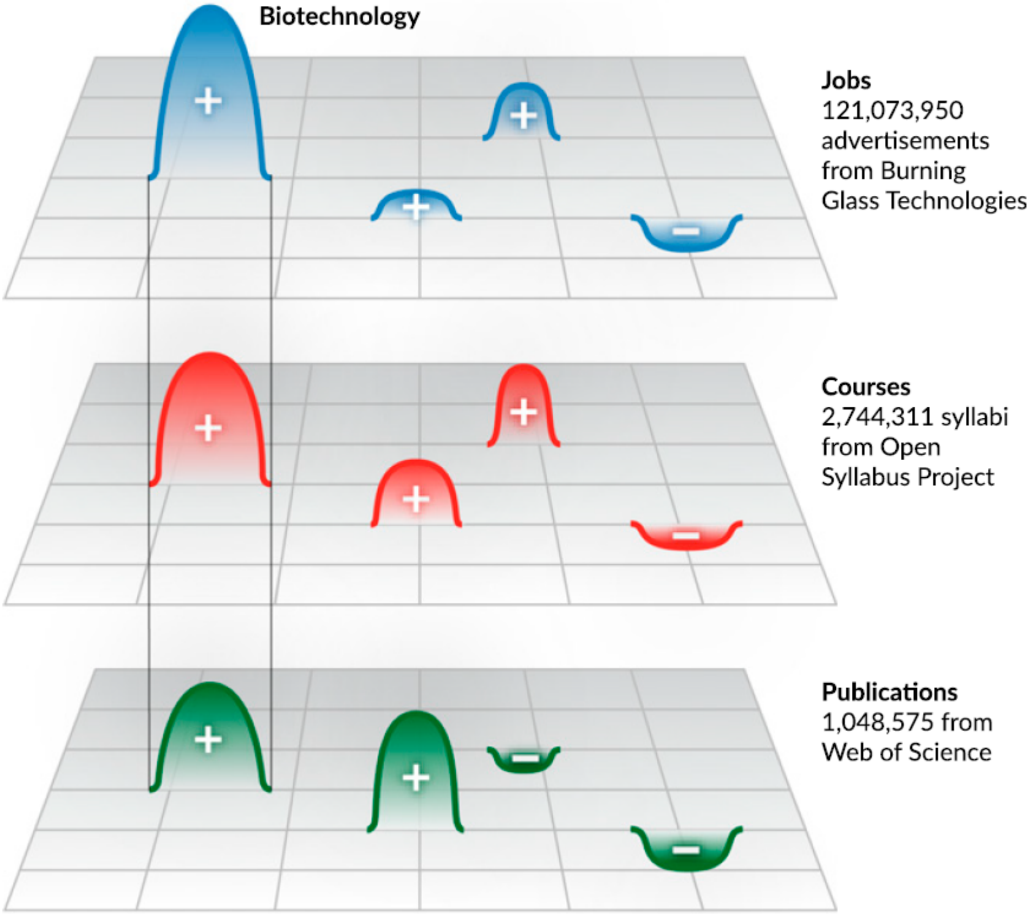
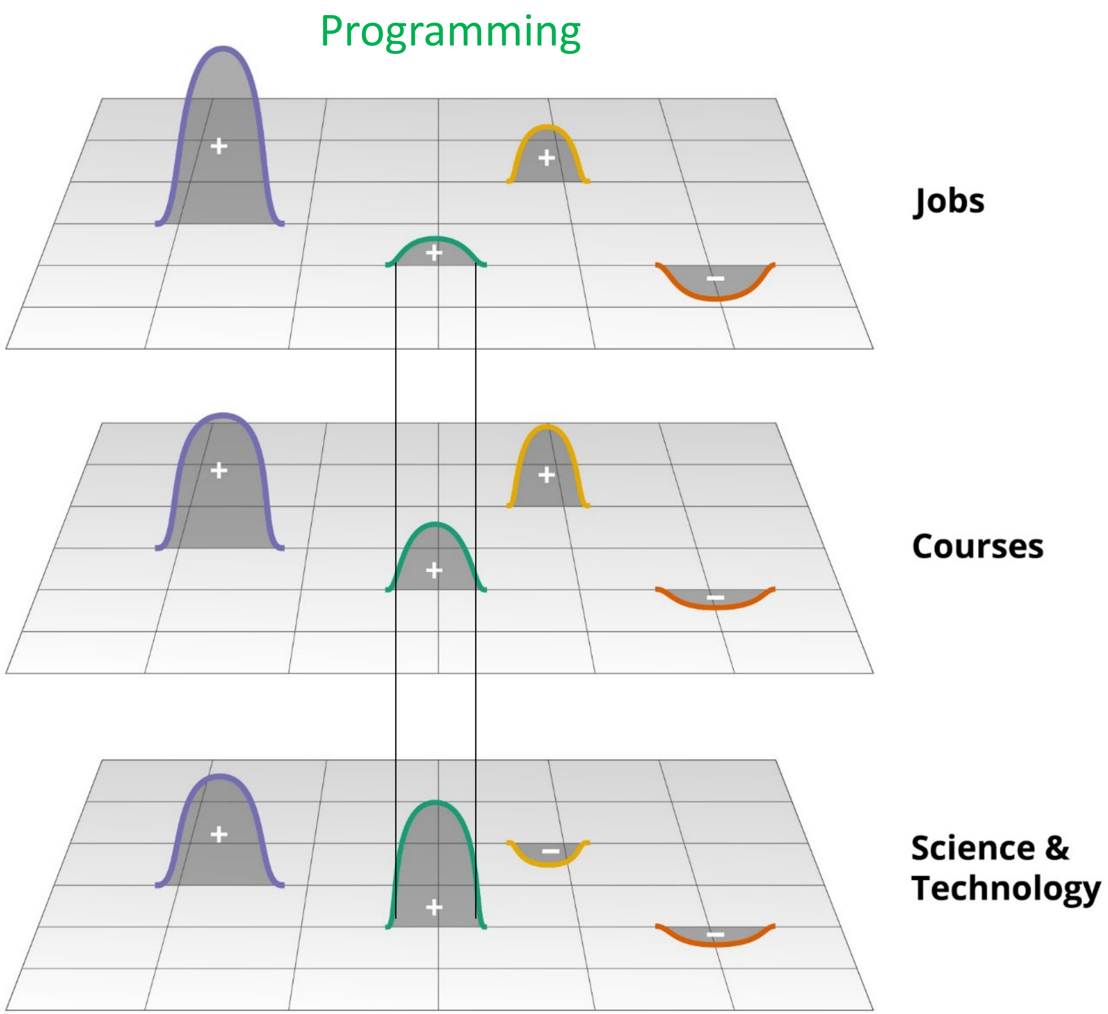
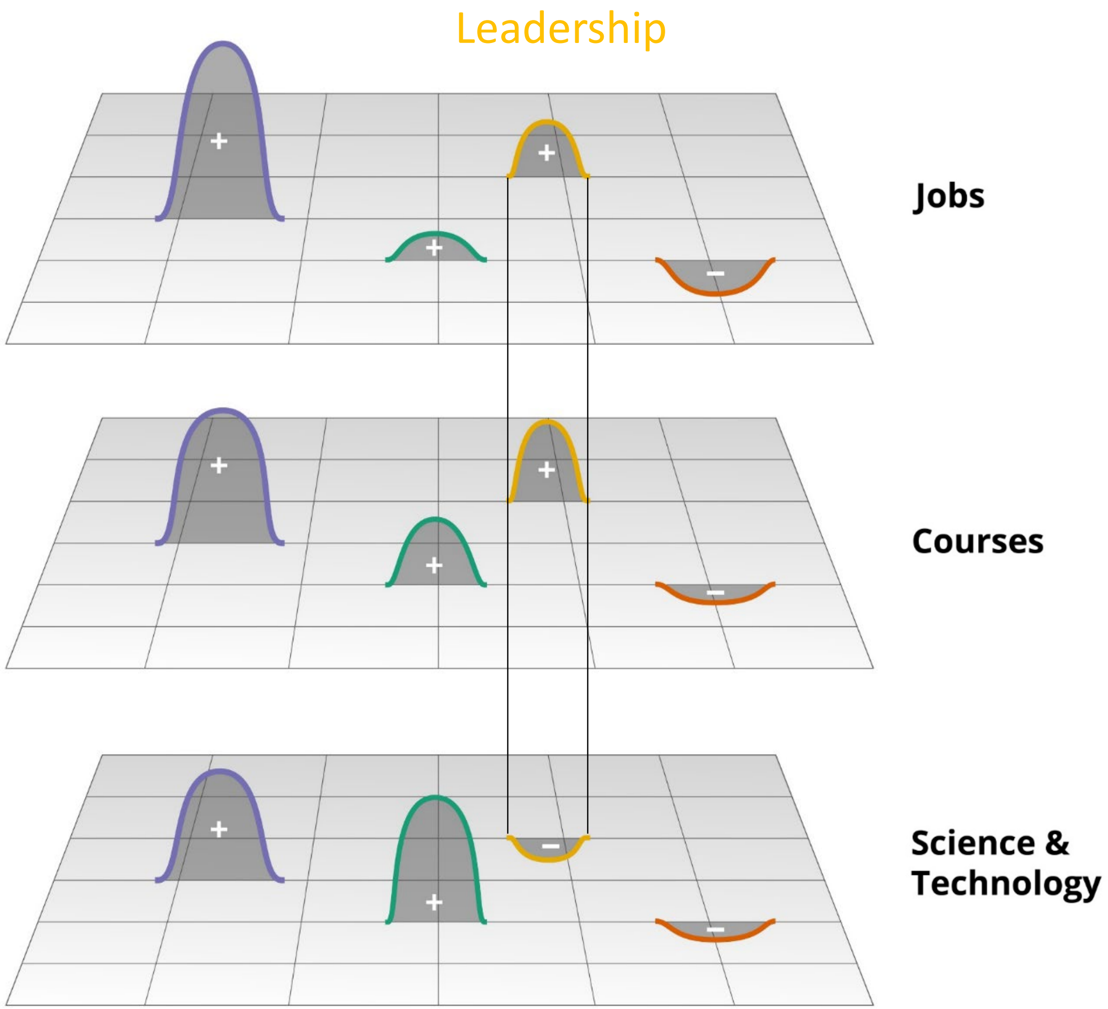
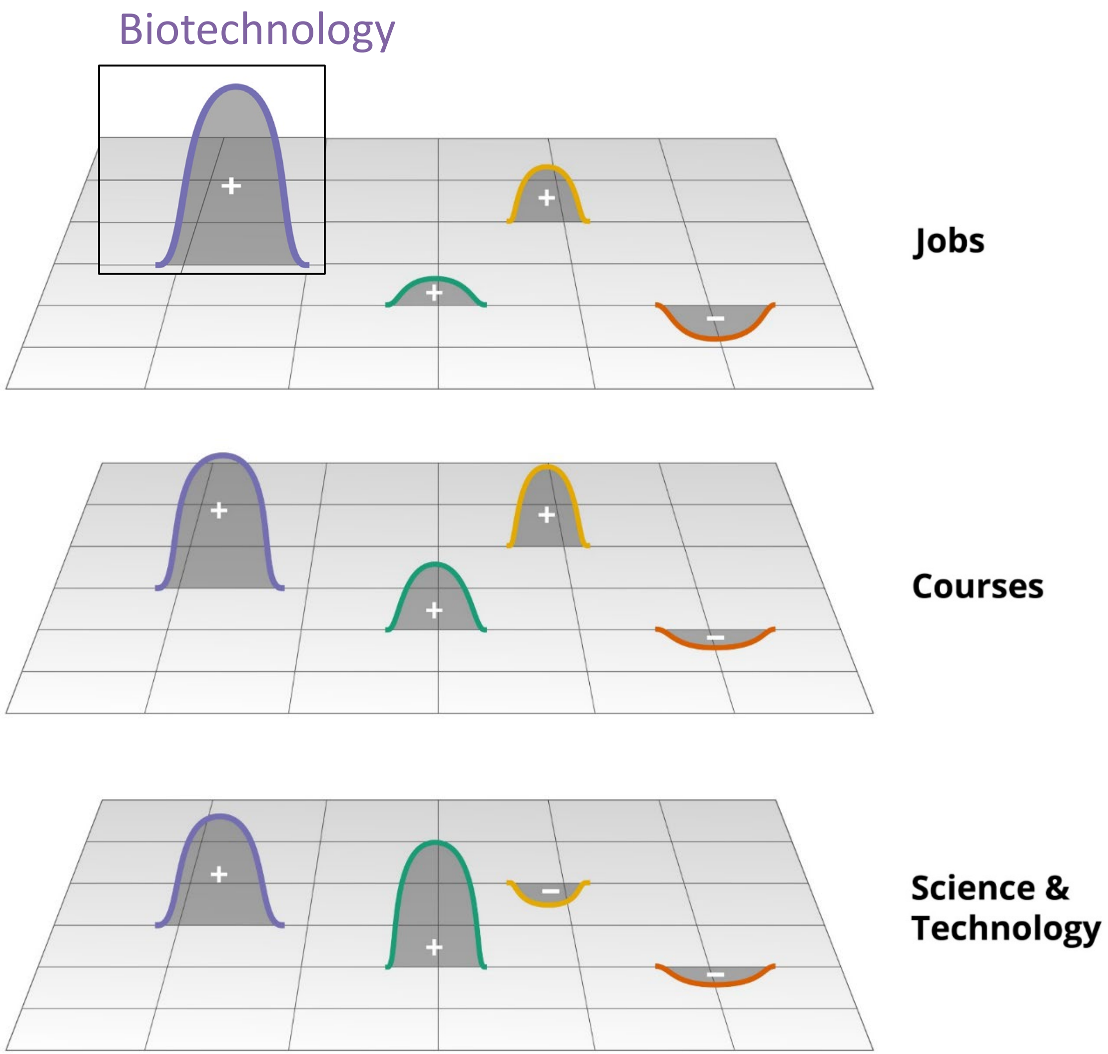
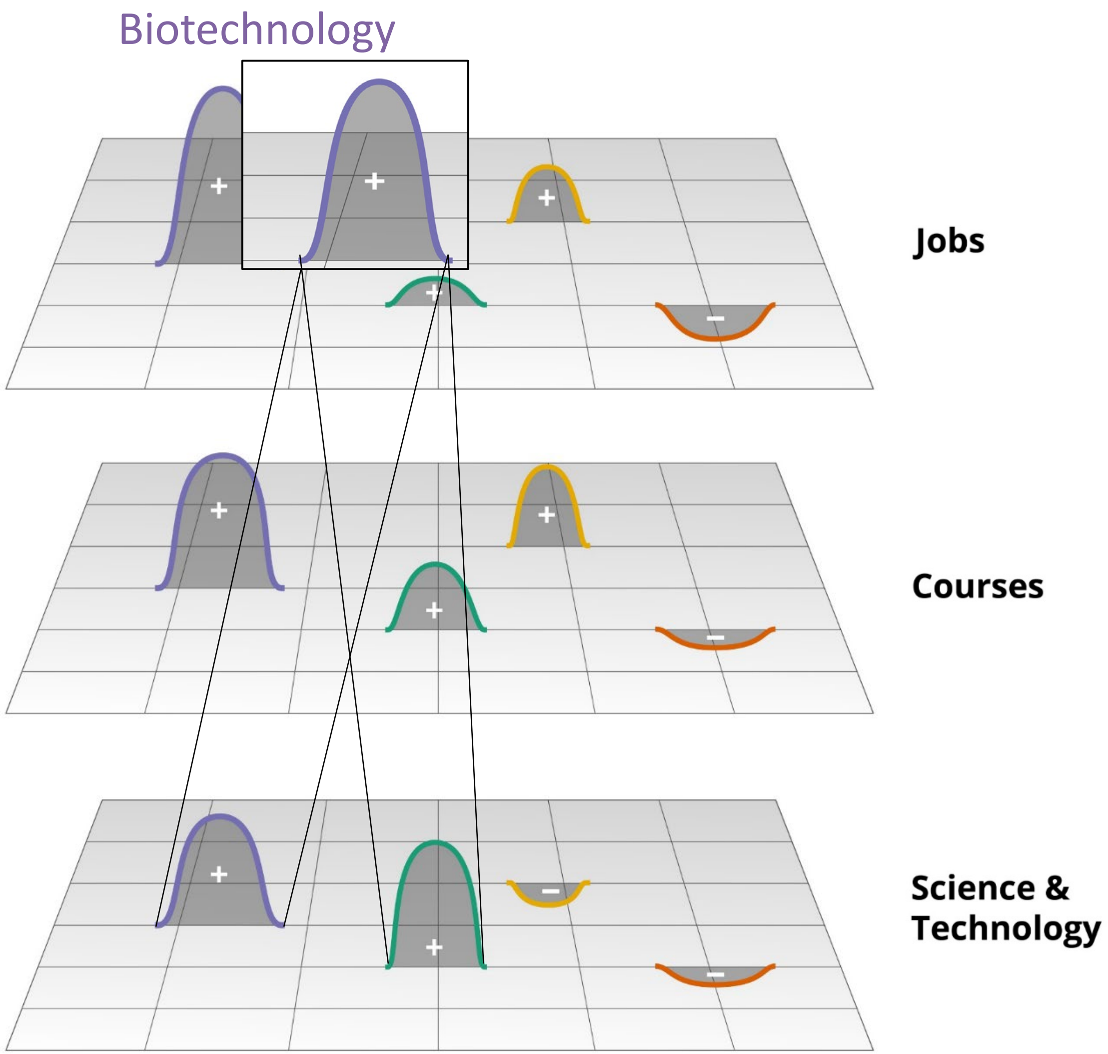


Fig. 1. The interplay of job market demands, educational course offerings, and progress in S&T as captured in publications. Color-coded mountains (+) and valleys (–) indicate different skill clusters. For example, skills related to Biotechnology might be mentioned frequently in job descriptions and taught in many courses, but they may not be as prevalent in academic publications. In other words, there are papers that mention these skills, but labor demand and commercial activity might be outstripping publication activity in this area. The numbers of jobs, courses, and publications that have skills associated and are used in this study are given on the right.


Programming
Jobs
Courses
Science &
Technology


Leadership
+
+
+
−
Jobs
+
+
+
−
Courses
+
+
−
−
Science &
Technology


Biotechnology
+
+
+
−
Jobs
+
+
+
−
Courses
+
+
−
−
Science &
Technology


Biotechnology
+
+
+
+
−
Jobs
+
+
+
+
−
Courses
+
−
+
−
Science &
Technology

Stakeholders and Insight Needs

- **Students:** What jobs will exist in 1-4 years? What program/learning trajectory is best to get/keep my dream job?
- **Teachers:** What course updates are needed? What balance of timely and timeless knowledge (to get a job vs. learn how to learn) should I teach? How to innovate in teaching and maintain job security or tenure?
- **Universities:** What programs should be created? What is my competition doing? How do I tailor programs to fit local needs?
- **Science Funders:** How can S&T investments improve short- and long-term prosperity? Where will advances in knowledge also yield advances in skills and technology?
- **Employers:** What skills are needed next year and in 5 and 10 years? Which institutions produce the right talent? What skills does my competition list in job advertisements?
- **Economic Developers:** What critical skills are needed to improve business retention, expansion, and recruitment in a region?

What is ROI of my time, money, compassion?

Urgency

- 35% of UK jobs, and 30% in London, are at high risk from automation over the coming 20 years.
  https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/uk-futures/london-futures-agiletown.pdf
- The aerospace industry and NASA have a disproportionately large percentage of workers aged 50 and older compared to the national average, and up to **half of the current workforce** will be eligible for retirement within the coming five years.
  Astronautics AIAA (2012) Recruiting, retaining, and developing a world-class aerospace workforce. https://www.aiaa.org/uploadedFiles/Issues_and_Advocacy/Education_and_Workforce/Aerospace%20Workforce-%20030112.pdf
- The rise of artificial intelligence will lead to the displacement of **millions of blue-collar as well as white-collar jobs** in the coming decade. Auerswald PE (2017) The Code Economy: A Forty-thousand-year History; Beyer D (2016) The future of machine intelligence: Perspectives from leading practitioners ; Brynjolfsson E, McAfee A (2014) The second machine age: Work, progress, and prosperity in a time of brilliant technologies; Ford M (2015) Rise of the Robots: Technology and the Threat of a Jobless Future.

Skill Discrepancies Between Research, Education, and Jobs Reveal the Critical Need to Supply Soft Skills for the Data Economy

- Data and Crosswalks
- MaxMatch for NLP
- Causal Analyses
- Visualizations

Börner, Katy, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wue, and James A. Evans. 2018. “Skill Discrepancies Between Research, Education, and Jobs Reveal the Critical Need to Supply Soft Skills for the Data Economy.” *PNAS* 115(50): 12630-12637.

PNAS PNAS PNAS PNAS

Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy

Katy Börner[a,b,1], Olga Scrivner[a], Mike Gallant[a], Shutian Ma[a,c], Xiaozhong Liu[a], Keith Chewning[d], Lingfei Wu[e,f,g,h], and James A. Evans[f,g,i,1]

[a]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408; [b]Educational Technology/Media Centre, Dresden University of Technology, 01062 Dresden, Germany; [c]Department of Information Management, Nanjing University of Science and Technology, 210094 Nanjing, China; [d]Burning Glass Technologies, Boston, MA 02110; [e]School of Journalism and Communication, Nanjing University, 210008 Nanjing, China; [f]Department of Sociology, University of Chicago, Chicago, IL 60637; [g]Knowledge Lab, University of Chicago, Chicago, IL 60637; [h]Tencent Research Institute, 100080 Beijing, China; and [i]Santa Fe Institute, Santa Fe, NM 87501

Edited by William B. Rouse, Stevens Institute of Technology, Hoboken, NJ, and accepted by Editorial Board Member Pablo G. Debenedetti September 12, 2018 (received for review March 14, 2018)

Rapid research progress in science and technology (S&T) and continuously shifting workforce needs exert pressure on each other and on the educational and training systems that link them. Higher education institutions aim to equip new generations of students with skills and expertise relevant to workforce participation for decades to come, but their offerings sometimes misalign with commercial needs and new techniques forged at the frontiers of research. Here, we analyze and visualize the dynamic skill (mis-)alignment between academic push, industry pull, and educational offerings, paying special attention to the rapidly emerging areas of data science and data engineering (DS/DE). The visualizations and computational models presented here can help key decision makers understand the evolving structure of skills so that they can craft educational programs that serve workforce needs. Our study uses millions of publications, course syllabi, and job advertisements published between 2010 and 2016. We show how courses mediate between research and jobs. We also discover responsiveness in the academic, educational, and industrial system in how skill demands from industry are as likely to drive skill attention in research as the converse. Finally, we reveal the increasing importance of uniquely human skills, such as communication, negotiation, and persuasion. These skills are currently underexamined in research and undersupplied through education for the labor market. In an increasingly data-driven economy, the demand for “soft” social skills, like teamwork and communication, increase with greater demand for “hard” technical skills and tools.

science of science | job market | data mining | visualization | market gap analysis

Education has been a critical vehicle of economic growth and social progress throughout the modern era. Higher education

doors. Some predictions say hundreds or even thousands of colleges and universities will close or merge in the coming years (4).

In addition, there seem to be major discrepancies and delays between leading scientific research, job market needs, and educational content. This has been particularly expressed with respect to science, technology, engineering, and mathematics jobs, where scientific and technological progress is rapid. Strategic decision making on what to teach, whom to hire, and what new research to fund benefits from a systematic analysis of the interplay between science and technology (S&T) developments, courses and degrees offered, and job market needs. Specifically, stakeholders in US higher education urgently need answers to the following questions. (*i*) Students: what jobs might exist in 5–10 years? What educational trajectories will best achieve my dream job? What core and specialized skills are required for what jobs and offered by what schools and programs? (*ii*) Teachers: what course updates are most needed? What balance of timely vs. timeless knowledge should I teach? How can I innovate in teaching and maintain job security or tenure? (*iii*) Universities: what programs should be created? What is my competition doing? How do I tailor programs to fit workforce needs? (*iv*) Science funders: how can S&T investments improve short- and long-term prosperity? Where will advances in knowledge also yield advances in skills and technology (5)? (*v*) Employers: what skills are needed next year and in 5 and 10 years? Which institutions produce the right talent? What skills are listed in job advertisements by my competition? How do I hire and train

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, “Modeling and Visualizing Science and Technology Developments,” held Decem-

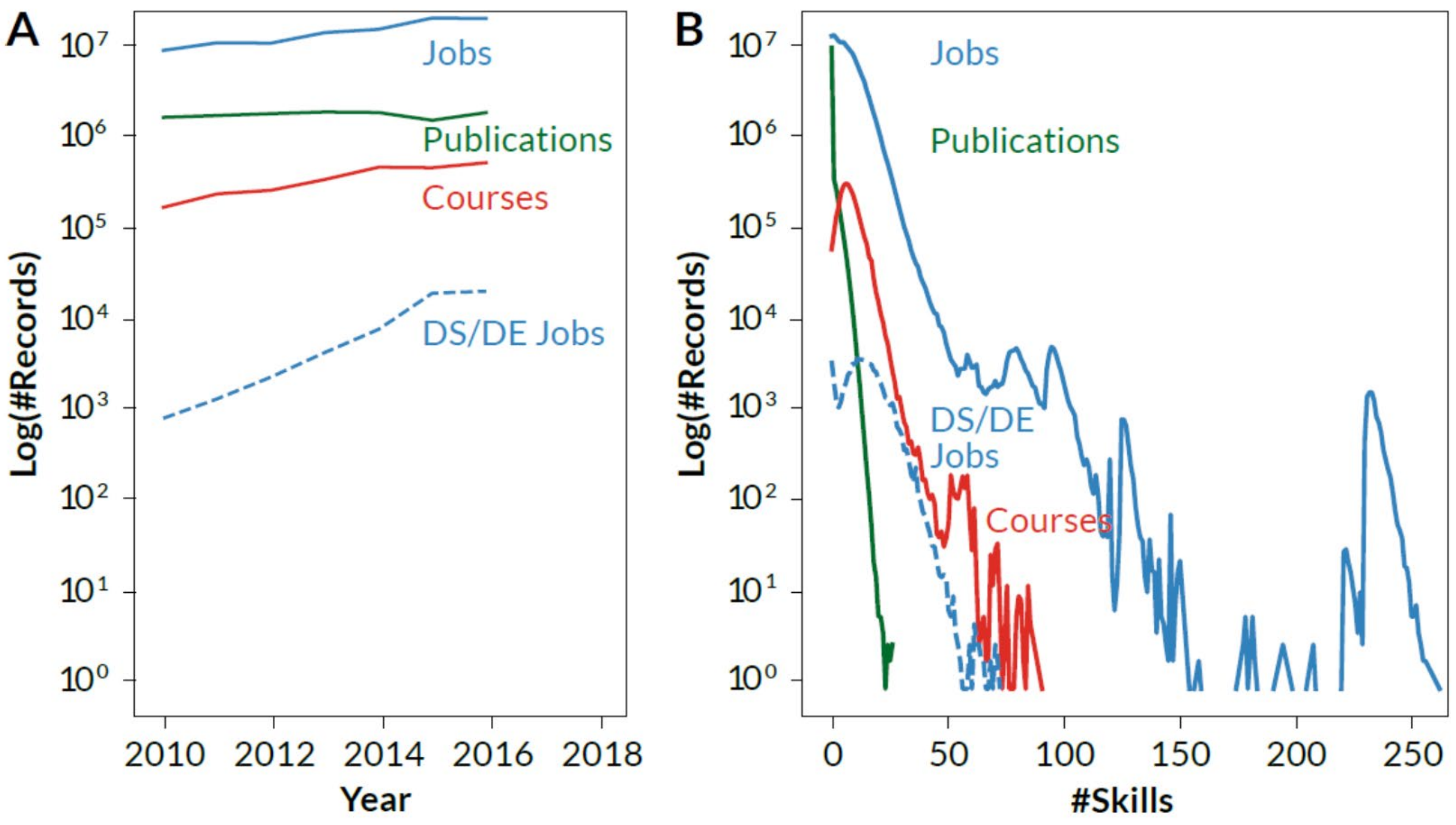
CNS Cyberinfrastructure for Network Science Center

Datasets Used

Job advertisements by Burning Glass posted between Jan 2010-Dec 2016.

Web of Science publications published Jan 2010-Dec 2016.

Course descriptions from the Open Syllabus Project acquired in June 2018 for courses offered in 2010-2016.



Data Type	#Records	#Records with skills	#Records without skills
All Courses	3,062,277	2,744,311	54,733
All Jobs	132,011,926	121,073,950	10,937,976
DSDE Jobs	69,405	65,944	3,461
All Publications	15,691,162	1,048,575	14,642,587
DSDE Publications	1,048,575	807,756	240,819

Fig. 2. Basemap of 13,218 skills. In this map, each dot is a skill, triangles identify skill clusters, and squares represent skill families from the Burning Glass (BG) taxonomy. Labels are given for all skill family nodes and for the largest skill cluster (NA) to indicate placement of relevant subtrees. Additionally, hard and soft skills are overlaid using purple and orange nodes, respectively; node area size coding indicates base 10 log of skill frequency in DS/DE jobs. Skill area computation uses Voronoi tessellation.

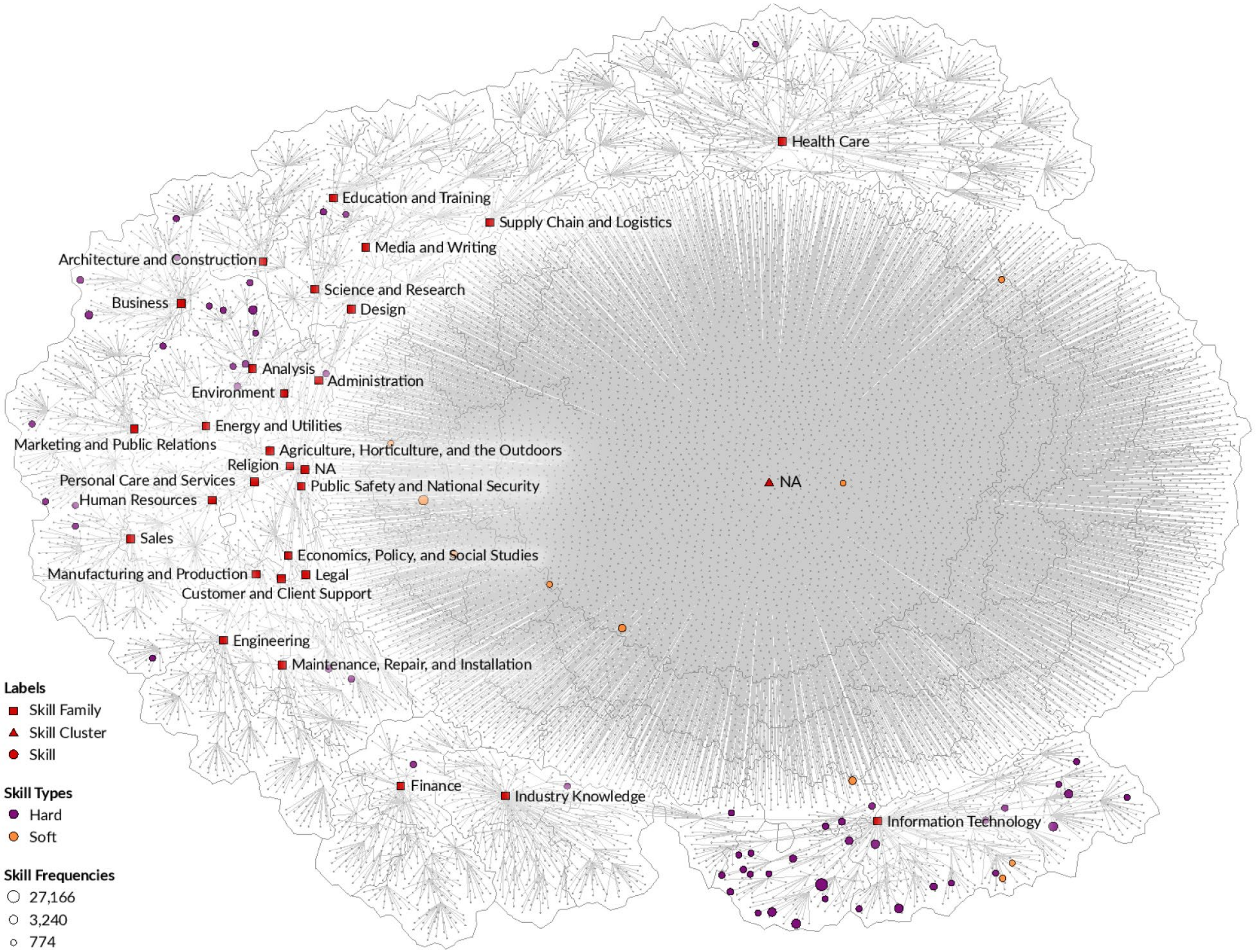
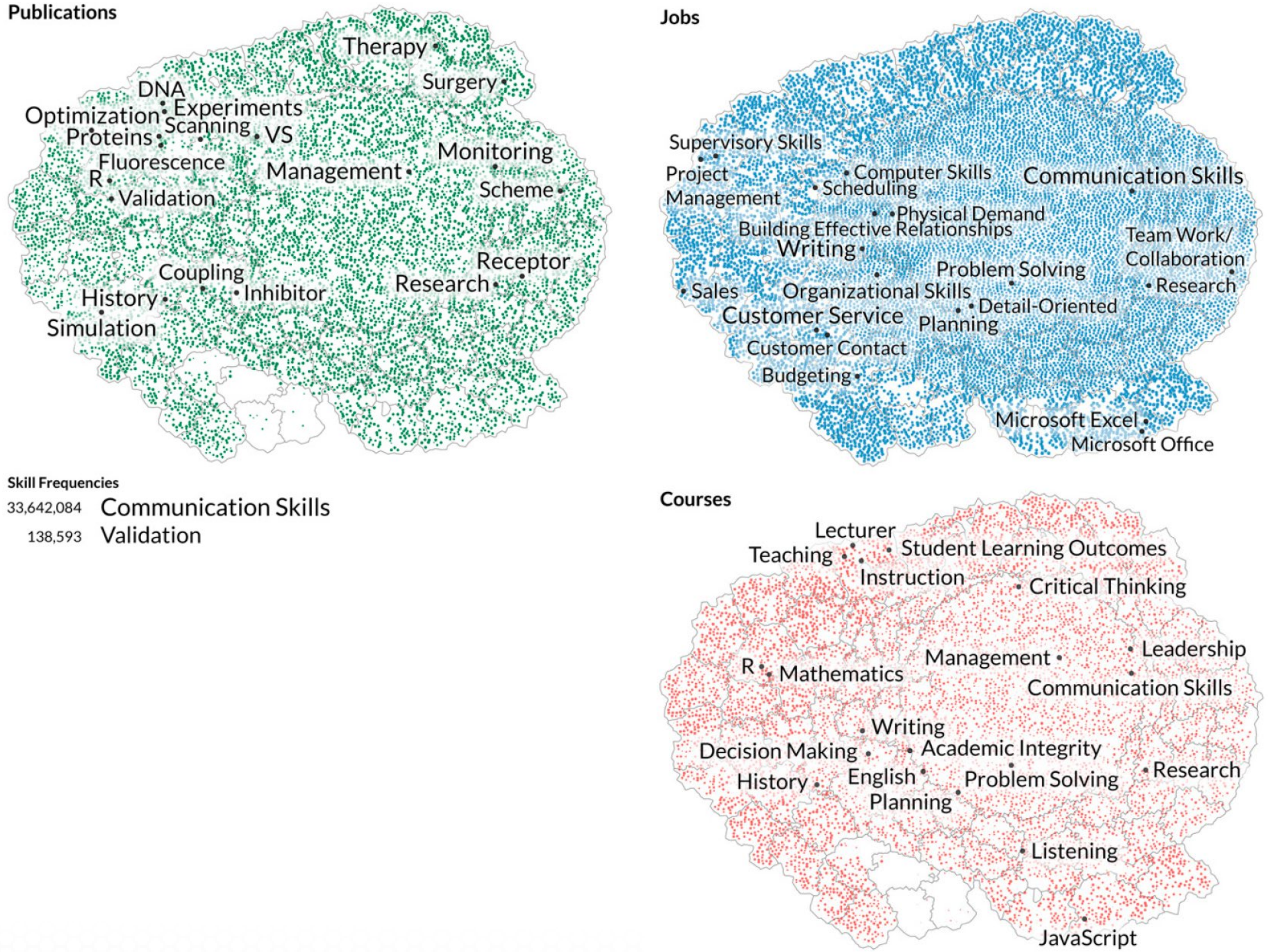


Fig. 3. Basemap of 13,218 skills with overlays of skill frequency in jobs, courses, and publications. This figure substantiates the conceptual drawing in Fig. 1 using millions of data records. Jobs skills are plotted in blue, courses are in red, and publications are in green. Node area size coding indicates base 10 log of skills frequency. The top 20 most frequent skills are labeled, and label sizes denote skill frequency.



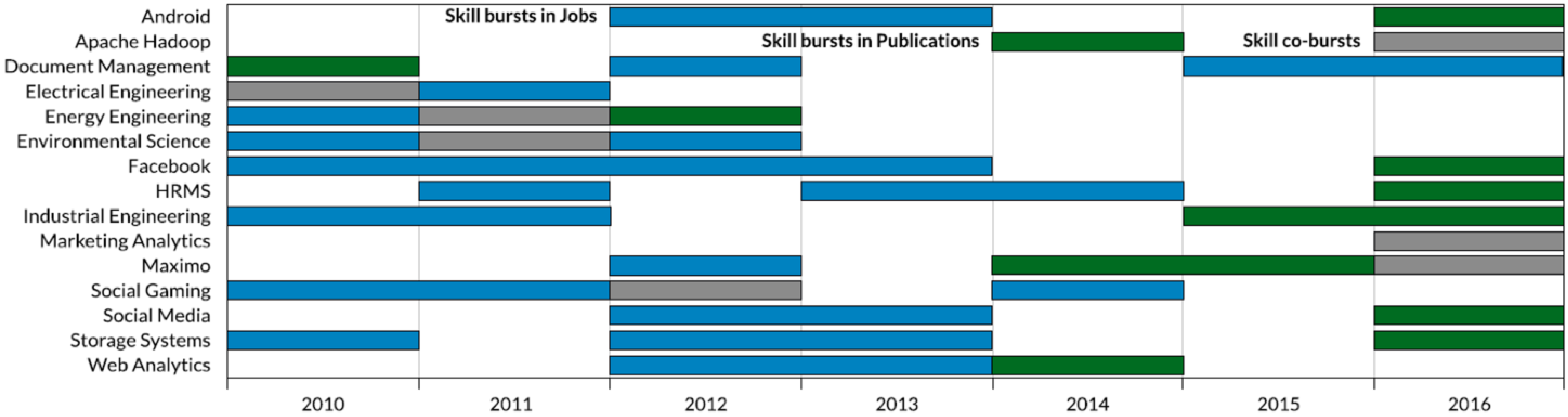


Fig. 4. Burst of activity in DS/DE skills in jobs and publications. Each burst is rendered as a horizontal bar with a start and an end date; skill term is shown on the left. Skills that burst in jobs are blue; skills bursting in publications are green. Seven skills burst in both datasets during the same years and are shown in gray. HRMS stands for human resources management system, and Maximo is an IBM system for managing physical assets.

Kullback-Leibler divergence

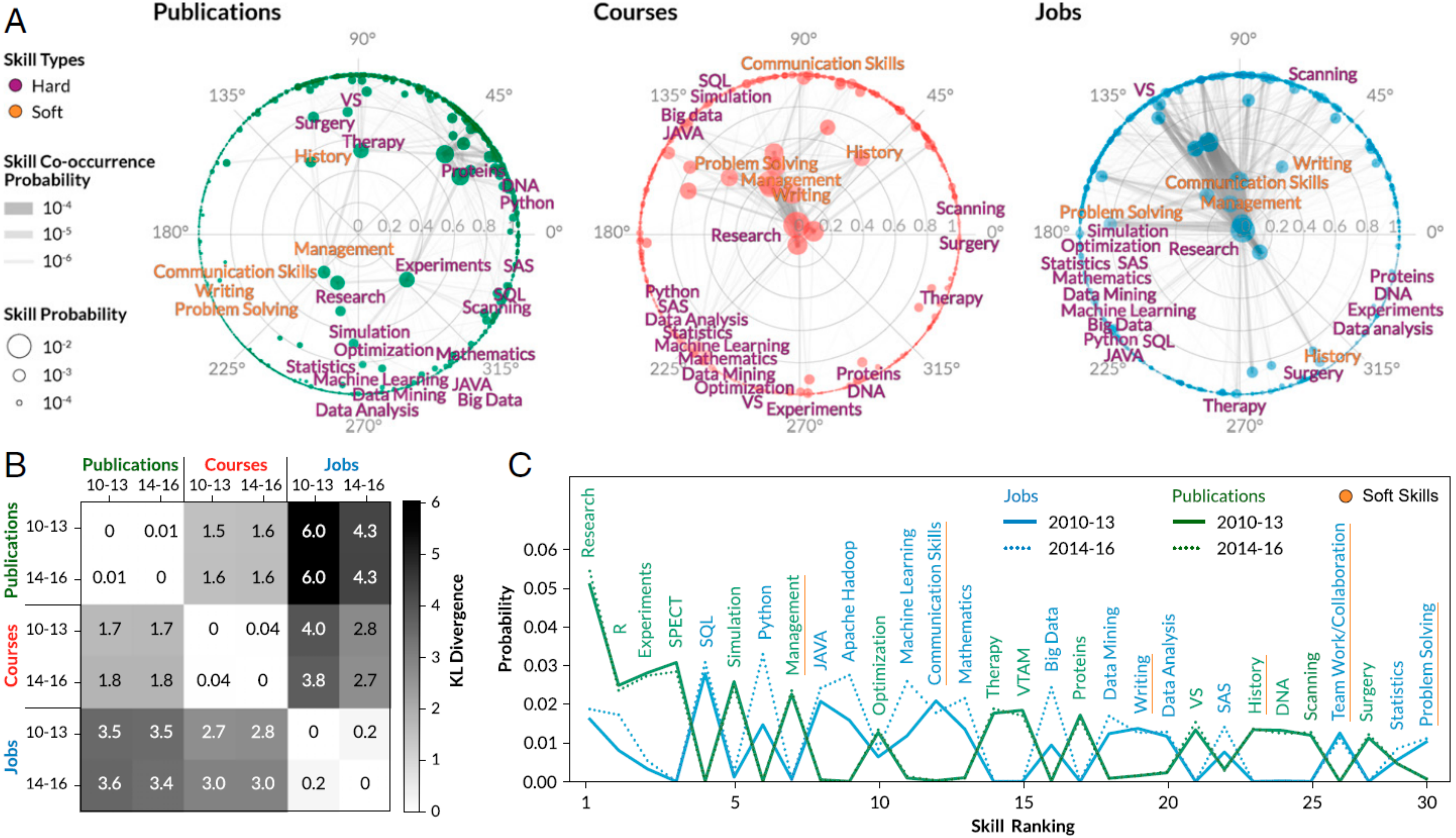


Fig. 5. Structural and dynamic differences between skill distributions in jobs, courses, and publications for 2010–2013 and 2014–2016. (*A*) Poincaré disks comparing the centrality of soft skills (orange) and hard skills (purple) across jobs, courses, and publications. (*B*) KL divergence matrix for jobs, courses, and publications in 2010–2013 and 2014–2016. (*C*) The most surprising skills in publications and jobs; *R* is a scripting language, VTAM refers to the IBM Virtual Telecommunication Access Method application, VS is the integrated development environment Visual Studio, and SAS is a data analytics software.

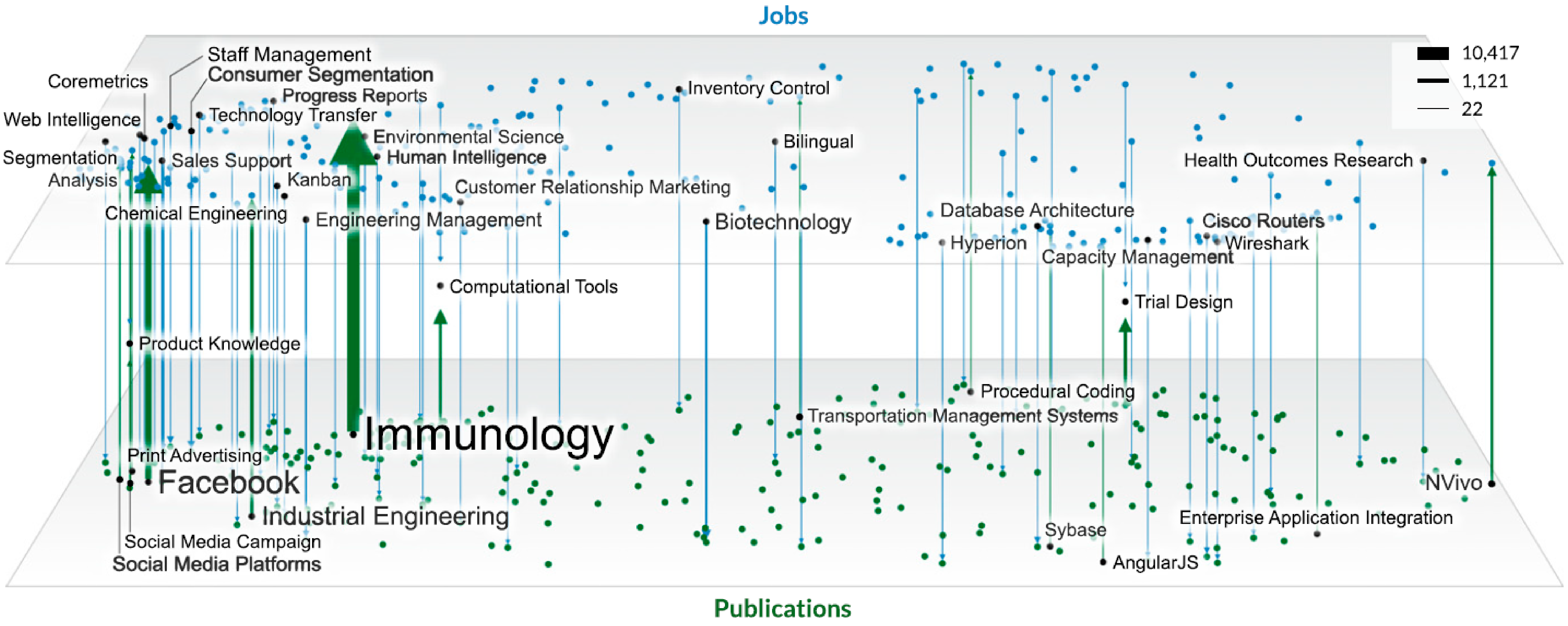


Fig. 6. Strength of influence mapping. Top 200 most frequent skills in jobs (blue) and in publications (green) plotted on the skills basemap from Fig. 2. Arrows represent skills with significant Granger causality (*P* value < 0.05). Line thickness and label size indicate skill frequency. The direction and thickness of each arrow indicate the *F*-value strength and direction.

Fig. 7. Multivariate Hawkes Process influence network of DS/DE skills within job advertisements 2010–2016. Each of the 45 nodes represents a top-frequency skill (29 soft and 16 hard skills) with a strong influence edge from/to other skill(s) in job advertisements between 2010 and 2016. Node and label size correspond to the number of times that the skill appeared in a job advertisement. Thickness of the 75 directed edges indicates influence strength.

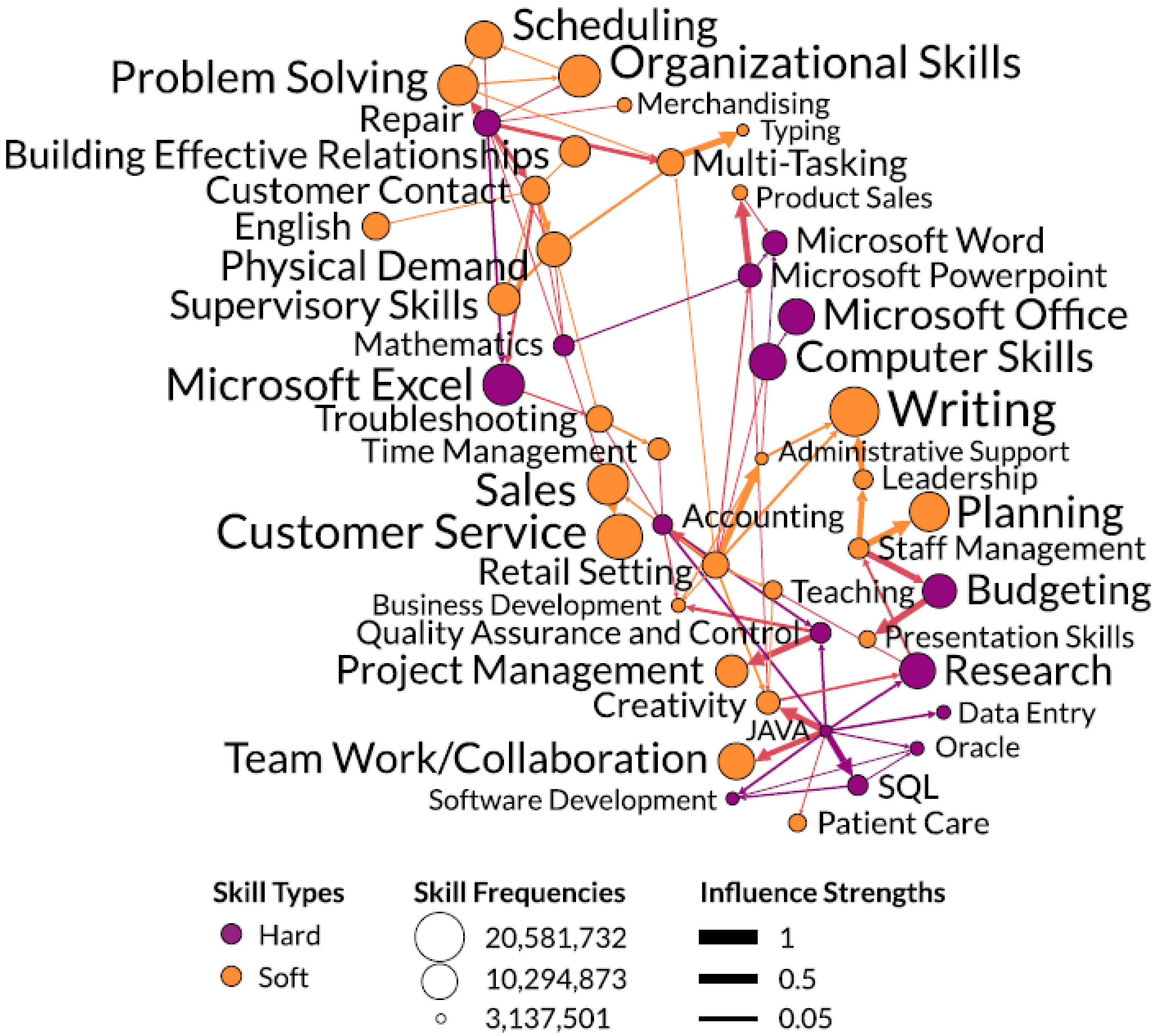
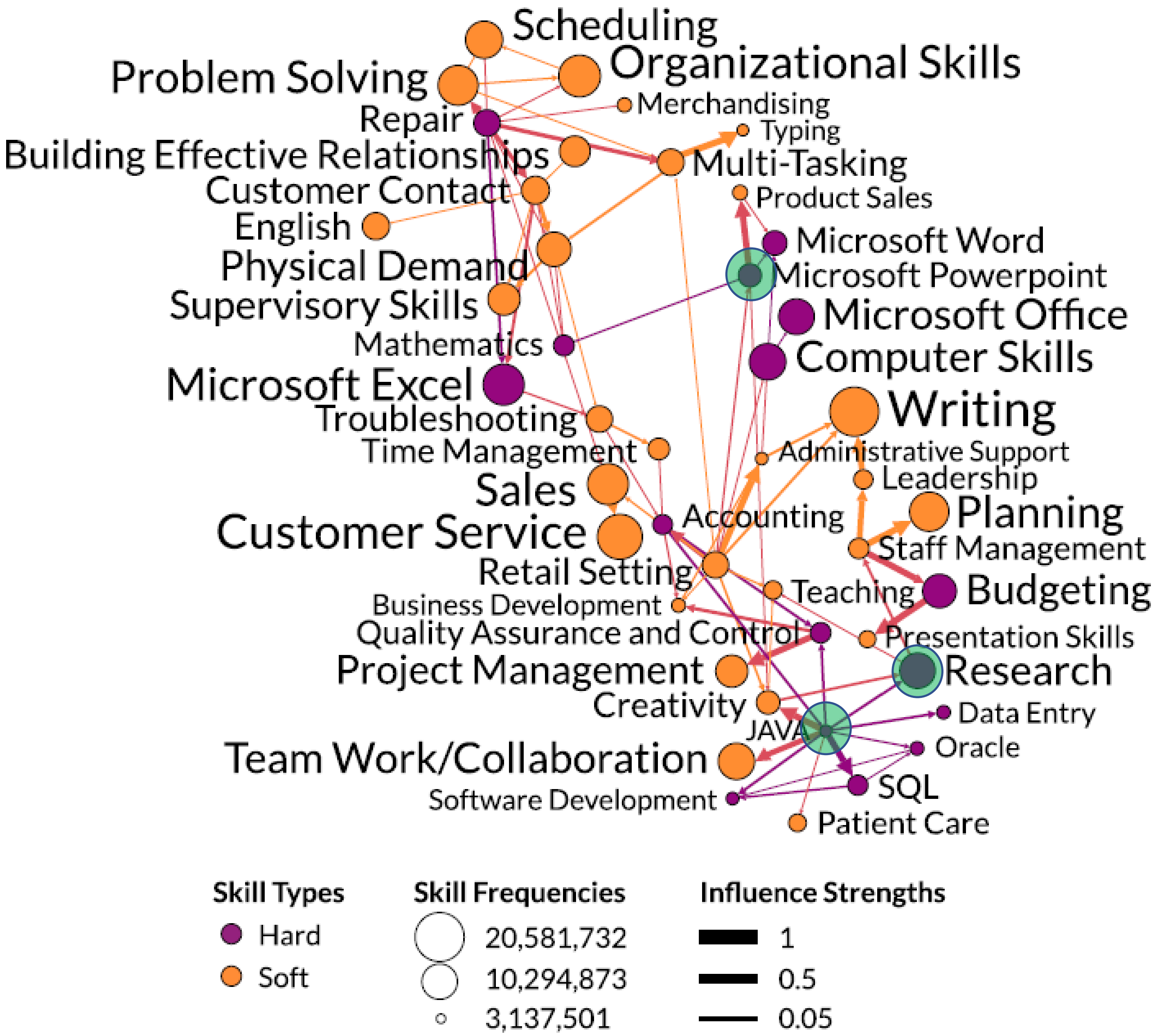


Fig. 7. Hawkes influence network of DS/DE skills within job advertisements 2010–2016. Each of the 45 nodes represents a top-frequency skill (29 soft and 16 hard skills) with a strong influence edge from/to other skill(s) in job advertisements between 2010 and 2016. Node and label size correspond to the number of times that the skill appeared in a job advertisement. Thickness of the 75 directed edges indicates influence strength.



Results

- Novel cross-walk for mapping publications, course offerings, and job via skills.
- Timing and strength of burst of activity for skills (e.g., Oracle, Customer Service) in publications, course offerings, and job advertisements.
- Uniquely human skills such as communication, negotiation, and complex service provision are currently underexamined in research and undersupplied through education for the labor market in an increasingly automated and AI economy.
- The same pattern manifests in the domain of DS/DE where teamwork and communication skills increase in value with greater demand for data analytics skills and tools.
- Skill demands from industry are as likely to drive skill attention in research as the converse.

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

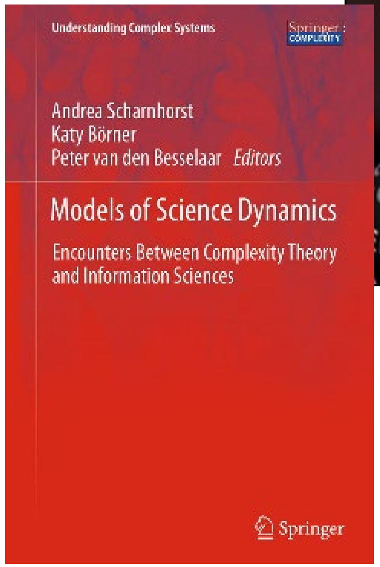
Börner, Katy (2010) **Atlas of Science: Visualizing What We Know**. The MIT Press. http://scimaps.org/atlas

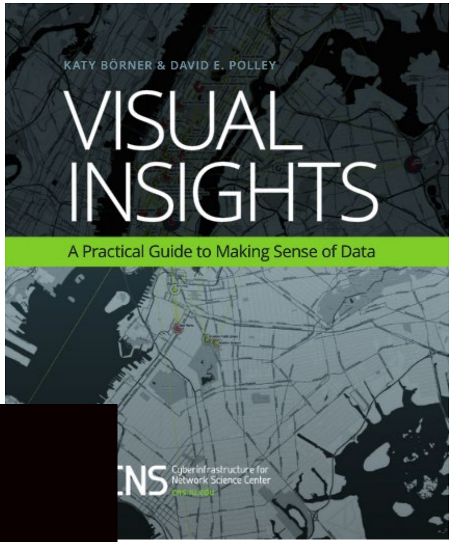
Scharnhorst, Andrea, Börner, Katy, van den Besselaar, Peter (2012) **Models of Science Dynamics**. Springer Verlag.

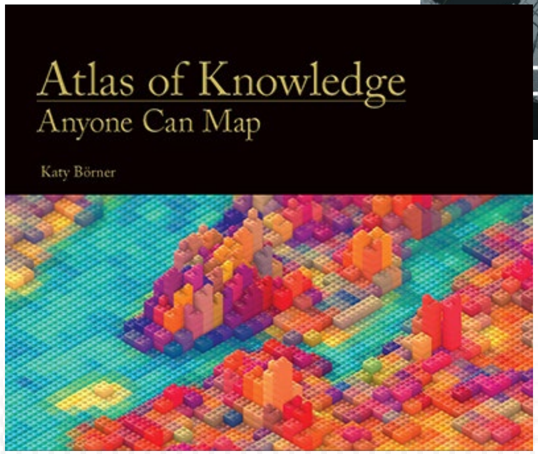
Katy Börner, Michael Conlon, Jon Corson-Rikert, Cornell, Ying Ding (2012) **VIVO: A Semantic Approach to Scholarly Networking and Discovery**. Morgan & Claypool.

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

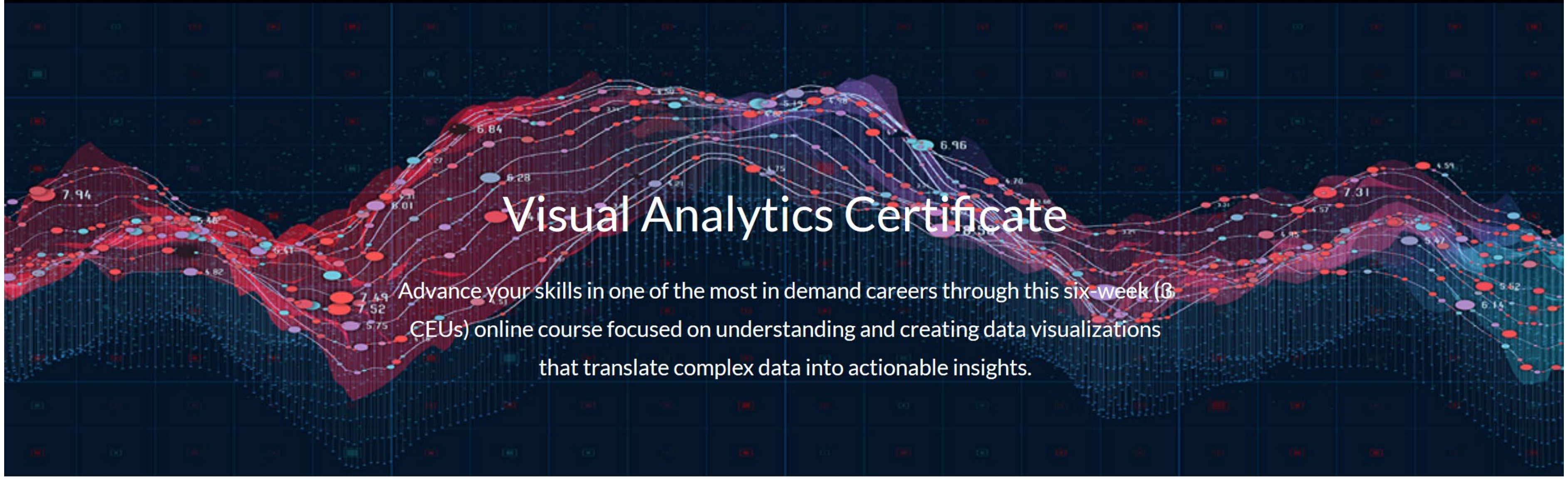






CNS Cyberinfrastructure for Network Science Center

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING



DOWNLOAD FLYER

REGISTER FOR OCT 7–NOV 17, 2019

Learn from Experts

Connect with industry professionals and leading researchers.

Evolve Yourself

Gain forever knowledge and skill-up in powerful data visualization tools.

Make a Difference

Embrace data-driven decision-making in your personal and professional life.

https://visanalytics.cns.iu.edu