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
BIG Data
Datasets used in SoS R&D

2009


<table>
<thead>
<tr>
<th>Dataset</th>
<th># Records</th>
<th>Years covered</th>
<th>Updated</th>
<th>Restricted access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medline</td>
<td>13,149,741</td>
<td>1965–2005</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>PhysRev</td>
<td>398,005</td>
<td>1893–2006</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>PNAS</td>
<td>16,167</td>
<td>1997–2002</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>USPTO</td>
<td>3,179,930</td>
<td>1976–2004</td>
<td>Yes*</td>
<td></td>
</tr>
<tr>
<td>NSF</td>
<td>174,835</td>
<td>1985–2003</td>
<td>Yes*</td>
<td></td>
</tr>
<tr>
<td>NIH</td>
<td>1,043,804</td>
<td>1972–2002</td>
<td>Yes*</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>18,021,560</td>
<td>1893–2006</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

2019

WoS*
Scopus
Google Scholar
MS Academic Graph
UMETRICS
Social Media Data
News Data
IoT Data

* [http://iuni.iu.edu/resources/cadre](http://iuni.iu.edu/resources/cadre)
Web of Science as a Research Dataset
November 14, 2016 - November 15, 2016 | Bloomington, Indiana

Organizers:

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

https://cns.iu.edu/workshops/event/161114.html
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 means that, for the most part, the datasets are silos, with great potential to learn from each other if they were interoperable. How might this be accomplished and what potential value would it bring?

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 in 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”

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

Session Organizers
Sybille Hinze
DZHW
Berlin, Germany

Jesper Schneider
Aarhus University
Denmark

Katy Börner
Indiana University
Bloomington, IN, USA

Andrea Scharnhorst
KNAW, Amsterdam
The Netherlands

Jason Rollins
Clarivate Analytics
San Francisco, CA, USA

Theresa Velden
ZTG TU Berlin
Germany

Ludo Waltman
CWTS, University of Leiden
The Netherlands

Jesper Schneider
Aarhus University
Denmark

Workshop Agenda
1. Introduction (Sybille Hinze & Theresa Velden)
2. Reproducibility in Scientometrics: Data Enclaves, Open Code, and Open Education (Katy Börner, Jesper Schneider)
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)

https://cns.iu.edu/workshops/event/171017.html
Web of Science as a Research Dataset


Introduction

The Clarivate Analytics Web of Science (WoS) has served as a research dataset for more than 5000 scholarly articles in the past 15 years across a wide range of fields and disciplines from medicine to social sciences. WoS has become a central repository for research data and has been used extensively for research to understand the structure and dynamics of the scientific literature. The WoS includes a large number of citations to documents such as journal articles, books, proceedings, patents, and more. The data are available in multiple formats, allowing researchers to explore and analyze scientific research patterns across various fields.

The WoS data includes information on the publication year, title, authors, abstract, and keywords for each document. The dataset is updated regularly, providing a comprehensive and up-to-date view of the scientific literature. The WoS database is widely used in bibliometric studies to analyze the productivity and impact of different institutions, researchers, and countries. The dataset is also used by researchers to track the evolution of research fields and to identify emerging trends in the scientific community.

New Computational Infrastructures

Research leveraging big, scholarly datasets like WoS presents researchers with challenges related to the data scale, data heterogeneity, relational forms, and sensitive (proprietary) nature. To overcome these challenges, researchers have developed a new generation of database-supported high-performance, and infrastructure for this purpose. The WoS database is a valuable resource for researchers who wish to analyze the scientific literature and identify emerging trends in the field.

IU International Co-Affiliation Network, 2004-2013

CNS @ Indiana University 2016

References


Acknowledgments

This work was partly supported by and contributed to research for VoMI, Facebook, Jump Trading, AGH Clarivate Analytics. Collaboration between the National Science Foundation and the University of Akron, NASA LG-73247, and EO-SSD-13-01892. The authors would like to thank all the individuals and organizations that contributed to this work. The authors would like to thank the Indiana University Network Science Institute (INSI) and Indiana University Libraries for support in this project. The authors would like to thank the Indiana University Network Science Institute (INSI) and Indiana University Libraries for support in this project.

The WoS database is a valuable resource for researchers who wish to analyze the scientific literature and identify emerging trends in the field. Researchers can use the WoS database to track the evolution of research fields and to identify emerging trends in the scientific community.
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 BIG DATA GATEWAY FOR RESEARCH LIBRARIES (SBD-GATEWAY)

FEDERATED SINGLE SIGN ON
- Custom

WEB QUERY INTERFACE
- RESEARCH ASSET COMMONS
  - Saved and Shared Results, Tools and Datasets
    - Visualizations
    - Annotations
    - Data Derivatives
    - Metadata

CLOUD COMPUTE GATEWAY
- Machine Learning
  - Analytics Tool
  - Analytics Tool
- Text Analytics
  - Analytics Tool
  - Analytics Tool
- Database Analytics
  - Analytics Tool
  - Analytics Tool
- Cloud and Local Compute Resources
  - Commercial Cloud Services

AUTHENTICATION
- InCommon
- Shibboleth
- Globus

ACCESS BY PROGRAMMING AND VISUALIZATION TOOLS
- Gephi
- sigma
- tableau
- js

APPLICATION PROGRAMMING INTERFACE (API)
- REST
- JSON
- XML

CLOUD STORAGE
- Indexed Storage
  - Relational Databases
  - Graph Databases
  - Future Databases
- Raw Format Data Lake Storage
  - Web of Science
  - MS Academic Graph
  - USPTO Data

GRANULAR DATASET PERMISSIONS
- Future Databases

LOCAL AND INSTITUTIONAL STORAGE
- Future Data Sources
- Future Datasets

http://iuni.iu.edu/resources/cadre
BIG Science of Science R&D
Maps of Science & Technology
http://scimaps.org

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

We are all familiar with traditional maps that show the relationships between countries, provinces, towns, and cities. Similar relationships exist between the scientific disciplines and research fronts in science. This allows us to map the structure of science.

One of the first maps of science was developed at the Institute for Scientific Information over 30 years ago. It depicted 341 areas of science from the citation patterns in 1.8 million scientific papers. That early map was intriguing, but it didn’t convey enough of science to accurately define its structure.

Things are different today. We have enormous computing power and advanced visualization without the scale issues of the structure of science problems. This galaxy map of science (left) was generated at Sandia National Laboratories using an advanced graph-based model. (The image from the earlier paper to 2001-2004, appears published in 2004. That data in the galaxy represents one of the 95,000 research communities active in science in 2002. A research community is a group of papers that are adjacent to one another in the same research space for a five-year time period. Communities can be born, continue, split, evolve, or die.)

The map of science can be used as a road map for science policy. Here’s the idea: in which organizations and institutions bear the brunt of scientific capabilities, additional information about the scientific and research impact of each research community allows policy makers to decide which areas to expand, contract, abandon, or ignore.

We also envision the map as an educational tool. For children, the theoretical relationship between areas of science and technology is a concrete map showing how much physics, chemistry, biology, and social studies matter. For advanced students, areas of interest can be located and neighboring areas can be explored.

Nanotechnology

Most research communities in nanotechnology are concentrated in Physics, Chemistry, and Materials Science. However, many disciplines at the Life and Medical Sciences also have nanotechnology applications.

Proteomics

Research communities in proteomics are concentrated in Biochemistry. In addition, there is a heavy focus on the transgenomics area.

The balance of the proteomics communities are evenly distributed among the Life and Medical Sciences.

Pharmacogenomics

Pharmacogenomics is a relatively new field with a strong focus in Medicine. It also has many connections in both social and medical communities in the Social Sciences.
This visualization explores the activity of science, math, and technology (SMT) related articles in the English language Wikipedia (wikipedia.org/wiki/special:stats).

The central image shows 63,888 articles related to science, math, and technology. Overlaid is a 15 x 15 grid of orange half-spheres and images.

Blue, green, and yellow circles represent the 3,599 math, 6,484 science, and 3,195 technology-related articles respectively. The larger the size of a circle, the higher the likelihood that it is that type of article. The four corners show activity patterns of SMT articles.


http://scimaps.org
A Topic Map of NIH Grants 2007

The National Institutes of Health (NIH) is organized as a multitude of Institutes and Centers whose missions are primarily focused on distinct diseases. However, disease etiologies and therapies float scientific boundaries, and thus there is tremendous overlap in the kinds of research funded by each Institute. This creates a daunting landscape for decisions on research directions, funding allocations, and policy formulations. Shown here is devised an interactive topic map for navigating this landscape, online at www.scimaps.org. Institute abbreviations can be found at www.nih.gov/nci.

Cardiac Diseases Research

An area of the map focused on cardiovascular function and dysfunction. Cardiac failure primarily funded by NHLBI is typically located near to Stroke (NINDS), since these are the two major medical emergencies associated with ischemia, which results from a severe reduction in blood supply. Also localized in this area are grants focused on HIV/AIDS and HCV, suggesting a major intercellular pathway for dissemination, and grants on Myocardial, Stroke, Cell Diseases, and Myocarditis.

Neural Circuits Research

As an overview of the neural circuits, neural circuits, which shows the diversity of topics and NIH Institutes that fund research in this area, such as Cardiac primary, primarily funded by NHLBI; Visual Processing primarily funded by NINDS; Epilepsy primarily funded by NINDS. For color coding, see legend in the upper right corner.

National Cancer Institute (NCI)

TOP 10 TOPICS

- Oncology Clinical Trials
- Cancer Treatment
- Cancer Therapy
- Cancer Genomics
- Personalized Analysis
- Cancer Chemotherapy
- Vaccines
- Radiation
- Epidemiology
- Cancer Chemoprevention

National Institute of General Medical Sciences (NIGMS)

TOP 10 TOPICS

- Basic Research
- Protein Structure
- Computational Models
- Metabolism
- Metabolism
- Neuroglial Mechanisms
- Proteins Complexes
- Integrated Cell and Genetic Pathways
- Cell Cycle

National Heart, Lung, and Blood Institute (NHLBI)

TOP 10 TOPICS

- Cardiovascular
- Pulmonary
- Gastroenterological
- Cardiac and Pulmonary
- Atherosclerosis
- Hypertension
- Blood Pressure
- Asthma
- Mitochondrial Disease
- Gene Expression

National Institute of Mental Health (NIMH)

TOP 10 TOPICS

- Mood Disorders
- Schizophrenia
- Behavioral Intervention Studies
- Mental Health
- Depression
- Cognitive-Behavior Therapy
- AIDS-Related
- Genetics
- Linkage Analysis
- Antidepressants
- Antipsychotics


http://scimaps.org
Map of Scientific Collaborations from 2005-2009

Computed Using Data from Elsevier's Scopus
Check out our **Zoom Maps** online!

Visit [scimaps.org](http://scimaps.org) and check out all our maps in stunning detail!
This is the Roanoke (Raleigh) megaregion.

The News Co-occurrence Globe
An interactive visualization of how countries are mentioned together in the world's news media

2.92K
cooccurrences

UNITED KINGDOM
cooccurrences in: 2,922%
cooccurrences out: 80%

Mapping Global Society –Kalev Leetaru
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
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 Hubsands 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
Modeling and Visualizing Science and Technology Developments

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

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

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

- **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

- **Scientific prize network predicts who pushes the boundaries of science**
  Yifang Ma and Brian Uzzi

- **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
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, Shitian 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](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](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.180471115](https://doi.org/10.1073/pnas.180471115)

- **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.180714115](https://doi.org/10.1073/pnas.180714115)

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

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.
Leadership

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-%202030112.pdf

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


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 2015. We show how courses mediated 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 underemphasized 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.

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, 2018, at the Sackler Center for 21st-Century Complexity and the NSF.
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.

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


Visual Analytics Certificate

Advance your skills in one of the most in-demand careers through this six-week (6 CEUs) online course focused on understanding and creating data visualizations that translate complex data into actionable insights.

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