Data Analytics in Support of Effective Workforce Training

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 Network Science Institute (IUNI) Indiana University, Bloomington, IN, USA

InnovationDigi Conference
https://innovationdigi.com

November 5, 2020
Thursday, November 5 • 2:30pm - 3:00pm

Data Analytics in Support of Effective Workforce Training

Sign up or log in to save this to your schedule, view media, leave feedback and see who's attending!

https://sched.co/eotR

Data Analytics in support of effective Workforce Training/Katy Börner

Speakers

Katy Börner
Data Analytics, Indiana University

Katy Börner is the Victor H. Yngve Distinguished Professor of Engineering and Information Science in the Departments of Intelligent Systems Engineering and Information Science, School of Informatics, Computing, and Engineering; core faculty of the Cognitive Science Program; and founding... Read More →

Thursday November 5, 2020 2:30pm - 3:00pm EST
Virtual
Advanced Analytics

Feedback form isn't open yet.
Overview

Identify and Overcome Skill Discrepancies


Job Postings in The Substance Use Disorder Treatment Sector


Empower Yourself and Others! Take the Visual Analytics Course: (https://visanalytics.cns.iu.edu)


Map of Scientific Collaborations from 2005-2009

Computed Using Data from Elsevier’s Scopus
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 blur 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.nlm.nih.gov/]. Institute abbreviations can be found at [www.nlm.nih.gov/].

Topic modeling, a statistical technique that automatically learns semantic categories, was applied to assess projects in terms used by researchers to describe their work, without the biases of keywords or subject headings. Grant similarities were derived from their topic mixes, and grants were then clustered on a two-dimensional map using a force-directed simulated annealing algorithm. This analytic creates an interactive environment for assessing grant relevance to research categories and to NIH Institutes in which grants are localized.


Cardiac Diseases Research
An area of the map focused on cardiovascular functions and dysfunctions. Cardiac failure (primarily handled by NHLBI) is typically clustered near the left side of the map, with two other medical emergencies associated with ischemia, which result from restricted blood supply. Also located in this area are grants focused on: Mitotic (Cell Division), a unique biochemical pathway for cell division, and grants on hemodynamics, Sickle Cell Disease, and fever.

Neural Circuits Research
An area of the map focused on neural circuits, which shows the interplay of topics and NIH Institutes that fund research in this area, such as: Cardiovascular Regeneration, predominantly fundedy by NHLBI; Blood Processing, primarily handled by NCI and NIDDK, primarily handled by NHLBI. For color coding, areas in the upper-left tier.
II.8 Taxonomy Visualization of Patent Data

The US Patent Hierarchy

Impact

The United States Patent and Trademark Office data collections and analyses of patent data for gaining insights to help organizations innovate are indispensable. In this chapter, we introduce the Taxonomy of Patent Data, a method to categorize and visualize patent data. The Taxonomy provides a hierarchical structure to organize patents into categories that reflect their technological or scientific characteristics. This visualization aids in understanding the relationships between different patents and their corresponding technologies.

The Taxonomy is based on a comprehensive analysis of the patent data, which includes the classification of patents into major categories, subcategories, and further subcategories. This approach allows for a more refined understanding of the patent landscape and facilitates the identification of emerging technologies and trends.

The Taxonomy Visualization provides a graphical representation of the patent hierarchy, enabling stakeholders to identify key areas of innovation and potential research opportunities. This visualization can be used by researchers, inventors, and business strategists to make informed decisions about research directions and patent development.

Prior Art


New patents often build on older ideas from many different categories. How do they do this? The Taxonomy of Patent Data provides a way to visualize the relationships between patents and their corresponding technologies. By mapping patents to their categories, the Taxonomy helps to identify areas of innovation and track the evolution of technology over time.
Check out our Zoom Maps online!

Visit scimaps.org and check out all our maps in stunning detail!
MACROSCOPES FOR INTERACTING WITH SCIENCE

Earth
Weather on a worldwide scale

AcademyScope
Exploring the scientific landscape

Mapping Global Society
Local news from a global perspective

Charting Culture
2,600 years of human history in 5 minutes
Oct 1-Dec 3, 2020: Exhibit on display at the Dimension Mill in Bloomington, IN on https://dimensionmill.org
Identify and Overcome Skill Discrepancies


See also https://www.pnas.org/modeling
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 science and technology (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
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 rise of artificial intelligence will lead to the displacement of millions of blue-collar as well as white-collar jobs in the coming decade.

• The pandemic is speeding up automation, and 85 million jobs are on the line.
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


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

Katy Börner1,a, Olga Scrivner2, Mike Gallant3, Shutian Ma4,a, Xiaozhong Liu1, Keith Chewning2, Lingfei Wue5,a, and James A. Evans1,a,7
1School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408; 2Educational Technology & Media Centre, Dresden University of Technology, 01062 Dresden, Germany; 3Department of Information Management, Nanjing University of Science and Technology, 210094 Nanjing, China; 4Kurzweil AI Technologies, Boston, MA 02119; 5School of Journalism and Communication, Nanjing University, 210023 Nanjing, China; 6Department of Sociology, University of Chicago, Chicago, IL 60637; 7Knowledge Lab, University of Chicago, Chicago, IL 60637; Trentor Research Institute, 100080 Beijing, China; and Santa Fe Institute, Santa Fe, NM 87501

Edited by William R. Bialek, 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 advertisementst published between 2010 and 2015. We show how courses intermediate 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.

E ducation has been a critical vehicle of economic growth and social progress throughout the modern era. Higher education has been a core and specialized training ground for what jobs and offered by what schools and programs? (a) 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) Scientists: how can S&T investments improve short- and long-term prosperity? Where will advances in knowledge also yield advances in skills and technology? (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...
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.
Job Postings in The Substance Use Disorder Treatment Sector

Background
Effective treatment strategies exist for substance use disorder (SUD), however severe hurdles remain in ensuring adequacy of the SUD treatment (SUDT) workforce as well as improving SUDT affordability, access and stigma. Although evidence shows recent increases in SUD medication access from expanding Medicaid availability under the Affordable Care Act, it is yet unknown whether these policies also led to a growth in hiring in the SUDT related workforce, partly due to poor data availability. Our study uses novel data to shed light on recent trends in a fast-evolving and policy-relevant labor market, and contributes to understanding data sources to track the SUDT related workforce and the effect of recent state healthcare policies on the supply side of this sector.

Methods and data
We examine hiring attempts in the SUDT and related behavioral health sector over 2010-2018 to estimate the causal effect of the 2014-and-beyond state Medicaid expansions on these outcomes through “difference-in-difference” econometric models. We use Burning Glass Technologies (BGT) data covering virtually all U.S. job postings by employers.

Findings
Nationally, we find little growth in the sector’s hiring attempts in 2010-2018 relative to the rest of the economy or to health care as a whole. However, this masks heterogeneity in the bimodal trend in SUDT job postings, with some increases in most years but a decrease in 2014 and in 2017, as well as a shift in emphasis between different occupational categories. Medicaid expansion, however, is not associated with any statistically significant change in overall hiring attempts in the SUDT related sector during this time period, although there is moderate evidence of increases among primary care physicians.
Introduction & Motivation

• Worldwide, the direct burden of illicit drug dependence increased to 20 million disability-adjusted life years in 2010. Examples of these illicit drugs are opioids, cocaine, amphetamines, and cannabis.

• In the US, mental health and SUD together became the leading cause of disease burden in 2015, while nearly 3% of Americans aged 12 years or older reported SUDs in the same year.

• The most effective SUD treatment (SUDT) is a combination of long-acting medications (usually methadone or buprenorphine) administered as part of a cognitive behavioral approach (such as counseling, family therapy, and peer support programs).

• In 2017, there were 13,857 treatment facilities in the U.S. with over 1,356,015 clients enrolled (The National Survey of Substance Abuse Treatment Services; NSSATS).

• The SUDT workforce is deemed inadequate by almost any measure. Workforce shortages and barriers have played a prominent role in limiting treatment access among those suffering from SUDs.
Fig 1. BGT online job postings. (A) BGT Job postings for all industries (black), Healthcare industry (red) and SUDT industries (light blue). The aggregated amount for all job postings is calculated for the period from 2010 through 2018. The healthcare sector is identified by the NAICS code ‘62’. The SUDT facilities are identified by three NAICS codes ‘6222’, ‘6214’, ‘6232’ filtered at 6-digit level. The left y-axis corresponds to the logarithmic trend lines for the total of all BGT job postings (black solid line) and the total of BGT healthcare sector (red dashed line). The y-right axis represents the SUDT sector values, shown as bar graphs. (B) Break down of job postings for three SUDT sectors. Three SUDT sectors are represented by their number of annual online job postings. Average line is calculated for each SUDT sector. Data Source: Burning Glass Technologies. 2019.

https://doi.org/10.1371/journal.pone.0228394.g001
Fig 2. **Bursting top-15 SUDT occupations during 2010-2018.** Each spike in demand is shown as a horizontal bar with a start and an end date. The length of the bar corresponds to the duration of the hiring burst, the width of the bar shows the burst strength, measured as weight (e.g., in the top panel, the Mental Health Counselor occupation has the strongest and the longest burst in the years 2016–2018).

https://doi.org/10.1371/journal.pone.0228394.g002
Empower Yourself and Others!
Data Visualization Literacy

US Employers which have sent students include The Boeing Company, Eli Lilly, DOE, CDC, NSWC Crane.
References


