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 was a key factor in American leadership in economic and social spheres in the 20th century. In recent years, however, rising college costs, inconsistent student achievement, and unsatisfactory graduation rates and career outcomes have led some critics to question the value of a traditional college education and call for significant innovation in higher education (1, 2). Massively open online courses (MOOCs), microcredits, and nanocertificates aim to address the need for high-quality, timely, affordable education and workforce training. By January 2018, the MOOC universe had grown to include 9,400 courses offered by 800+ universities taken by 78 million students. The top five MOOC providers, as defined by the number of registered users, are Coursera (with 30 million users), edX (14 million users), XuetangX (9.3 million users), FutureLearn (7.1 million users), and Udacity (5 million users) (3). In 2016–2017, the number of US colleges and universities dropped by 5.6%. Most disappearing institutions were for-profit colleges, but more than 30 private nonprofits also closed their 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 Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/modeling_and_visualizing.

Author contributions: K.B., X.L., and J.A.E. designed research; K.B., O.S., S.M., L.W., and J.A.E. performed data analysis and visualization; O.S. and M.G. performed job and publication data modeling; S.M. and X.L. performed job and publication data entity extraction; K.C. performed job data modeling; L.W. and J.A.E. performed course data entity extraction; and K.B., O.S., M.G., S.M., X.L., K.C., L.W., and J.A.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. W.B.R. is a guest editor invited by the Editorial Board.

Data deposition: The data and algorithm details are provided in *SI Appendix*; the survey instruments and code are at https://github.com/cns-iu/cjobs.

[1]To whom correspondence may be addressed. Email: katy@indiana.edu or jevans@uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804247115/-/DCSupplemental.

Published online December 10, 2018.

productive teams? (*vi*) Economic developers: what critical skills are needed to improve business retention, expansion, and recruitment in a region?

Fig. 1 shows a conceptual drawing of the interplay of job market demands (Fig. 1, *Top*, blue mountains), educational course offerings (Fig. 1, *Middle*, red), and progress in S&T as captured in publications (Fig. 1, *Bottom*, green). Each level represents the same topical landscape, also called basemap, of skills (Fig. 2). Color-coded mountains (+) and valleys (−) indicate high- and low-frequency skills (listed in jobs, courses, and publications). Over time, the frequency and type of skills change. This paper uses millions of job, course, and publication records to analyze and visualize the structure and dynamics of skills requested in jobs, taught in courses, and published in scholarly publications. It shows that education not only is fueled by but also, recursively impacts scientific and technological progress and industry capabilities. Jobs in Data Science and Data Engineering (DS/DE) are a key focus, since these are evolving rapidly, requiring continuous updates of educational offerings. Resulting analyses and visualizations were presented to 20 domain experts from academia, industry, and government, and comments were used to interpret and optimize study results. Data and algorithm details are provided in *SI Appendix*; survey instruments and code are at https://github.com/cns-iu/cjobs.

Prior Work

There exists an extensive body of work on workforce analyses and projections. For example, O*NET OnLine provides detailed descriptions of different occupations and associated skills for use by job seekers, human resource professionals, students, and others (6). The US Bureau of Labor Statistics regularly publishes employment projections. The February 2018 prediction for 2016–2026 includes 158 technology skills, 83 occupations, and 178 O*NET-defined "hot technologies" (7, 8). Burning Glass Technologies (BG) harnesses more than 138 million jobs and associated skills to deliver labor market analytics to state and national governments, educational institutions, and major employers (9). The LinkedIn Economic Graph challenge (10) encourages researchers to answer core workforce questions and explore their implications. The Economic Modeling Specialist International firm (11) applies advanced techniques to analyze and map student employment trajectories using information on student majors and required job skills, with a focus on workers nearing the end of their careers and young adults just beginning them (12). The European Big Data Hackathon in 2017 used European Employment Services data (13) comprising anonymized curriculum vitae by 297,940 unique jobseekers and job vacancies published by potential employers; Opik et al. (14) created an occupation co-occurrence basemap, overlaid possible career/job retraining pathways, and studied the impact of megatrends and interventions on the labor market. Recent analyses have also attempted to trace the future of work by grafting expert opinion about automation onto labor market participation (15) or using time series to understand trends in classes of employment (e.g., the move from permanent positions to contingent employment) (16, 17). New work on skill networks suggests rich, complex measures of human capital (18) and reveals increasing polarization in physical and sensory vs. cognitive skills across US jobs (19, 20).

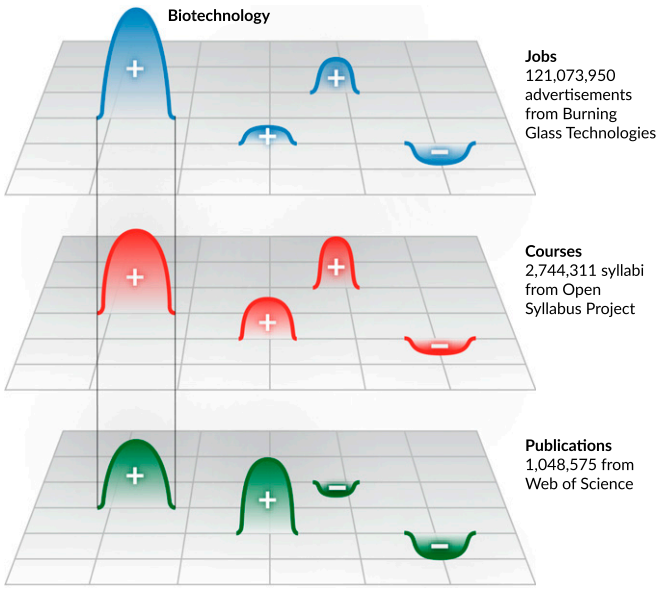

Fig. 1. The interplay of job market demands, educational course offerings, and progress in S&T as captured in publications. Color-coded mountains (+) and valleys (−) indicate different skill clusters. For example, skills related to Biotechnology might be mentioned frequently in job descriptions and taught 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.

This paper builds on and extends this prior work by aligning databases of job advertisements, course descriptions, and research publications to make two unique contributions. First, we create a basemap for job, course, and publication data and overlay skills data to compare the topical coverage of the three datasets. We show that maps can legibly convey information about the structure and dynamics of skill supply and demand, enabling the strategic cultivation of training assets. Second, we apply computational models to understand discrepancies and temporal delays between evolving job market needs, course offerings, and S&T developments. We find that university course offerings mediate the skills explored in research and exploited in jobs and that they are equidistant in "skill space" between science and industry. We also find that industry demand for job skills is just as likely to drive skill attention in research as industry is to follow research developments. Finally, we show that rising demand for "hard" technical skills often stimulates subsequent demand for "soft" social and communication skills. For example, DS/DE skills seem to condition increasing industry demand for presentation, storytelling, and sales skills. This suggests the critical importance of supplying not only hard but also, soft skills with continuing educational and research attention (21).

Cross-Walking Skills: From Jobs to Courses to Publications

The data used in this study cover January 2010 to December 2016. They comprise 132,011,926 job advertisements by BG, 3,062,277 course syllabi from the Open Syllabus Project, and 15,691,162 publication abstracts from the Web of Science (WoS) (details are in *SI Appendix*, *SI Text* and Fig. S1*A* on the number of records per year).

To model and visualize multiple data types (jobs, courses, publications), a robust mapping from each dataset to all others is required. There exist two general approaches to align datasets. The first approach cross-walks an existing classification system from one dataset to the others. Existing classification systems include the BG skills taxonomy (22), the WoS field classification system of 258 subclasses, and the University of California, San Diego map of science with 554 subdisciplines (23). The second approach uses linguistic analysis of text in jobs, courses, and publications to extract key terminology. To maximize consistency and accuracy in linking jobs, courses, and publications, we use a combination of the first and second approach here. Specifically, we anchor our analysis with the BG skills taxonomy, which organizes 13,218 unique skills into 560 skill clusters that are further aggregated into 28 skill families. We then identify the 13,218

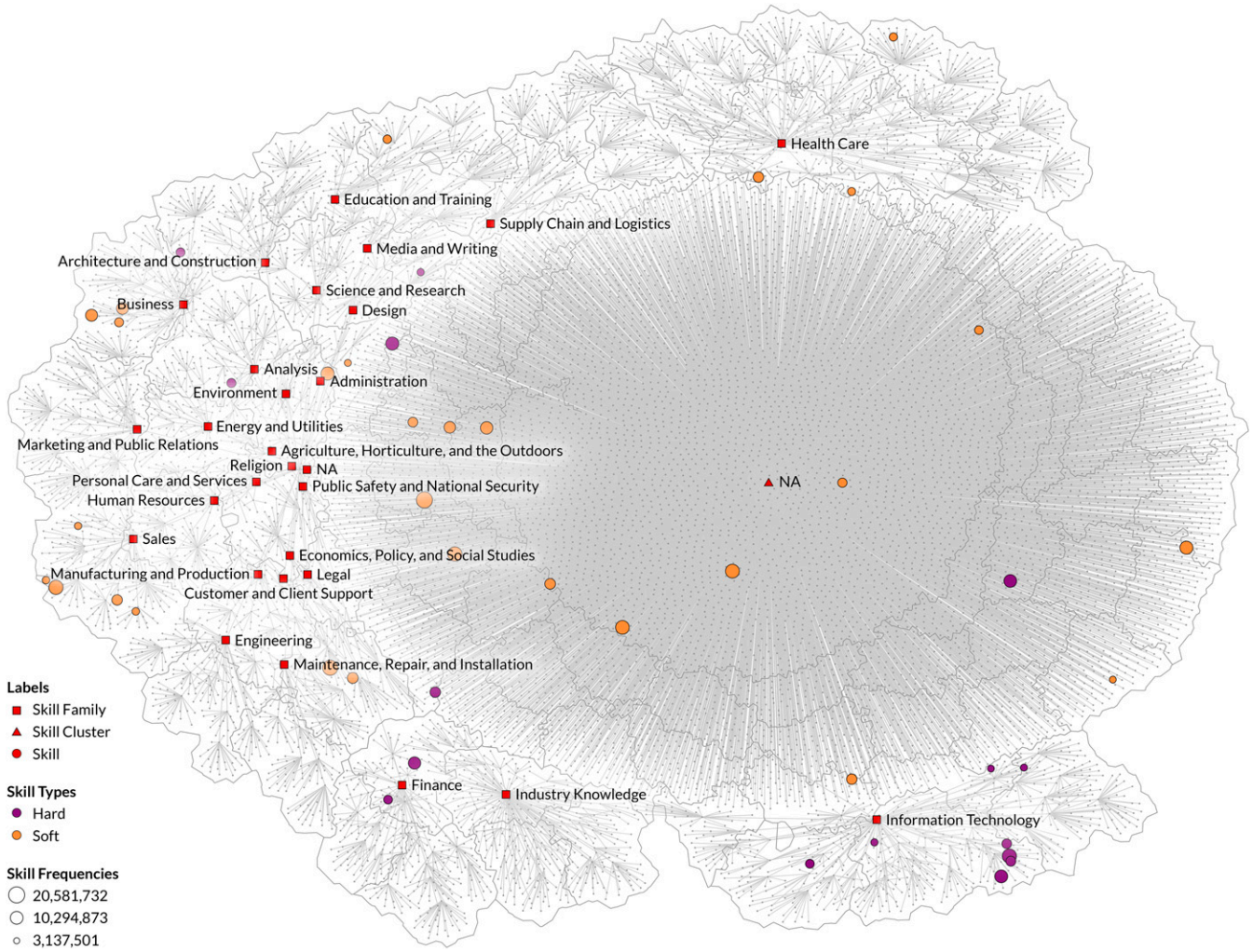

Fig. 2. Basemap of 13,218 skills. In this map, each dot is a skill, triangles identify skill clusters, and squares represent skill families from the BG taxonomy. Labels are given for all skill family nodes and for the largest skill cluster (NA) to indicate placement of relevant subtrees. Additionally, 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.

skills within job advertisements, course syllabi, and research publications to create a cross-walk.

Different supervised and unsupervised entity extraction algorithms were compared (*SI Appendix*, *SI Text*), and the forward maximum matching (MaxMatch) algorithm was selected as the algorithm with the best overall precision, recall, and F1 performance (*SI Appendix*, Table S4). The MaxMatch algorithm was then applied to job titles and descriptions, course syllabi, and publication titles and abstracts. Punctuation and stop words were removed. As a result, 121,073,950 job advertisements, 2,744,311 courses, and 1,048,575 WoS publications were associated with at least one skill in the BG skills taxonomy and are included in the analyses presented subsequently (*SI Appendix*, Fig. S1*B* and Table S1 shows the distribution of the number of skills for jobs, courses, and publications).

Building on prior work mapping S&T (24, 25) and the development of classification systems and maps of science (23), we created a basemap of the 13,218 skills from the BG skills taxonomy. The BG skills taxonomy is a proper tree with 29 family nodes, 561 cluster nodes, and 13,218 skills leaf nodes plus a total of 13,807 edges. The GMap algorithm (26) that visualizes graphs as maps that are easy to read (27, 28) was used to compute a 2D basemap layout of the BG skills taxonomy, and Gephi (29) was used to create data overlays (Fig. 2). In this basemap, each red square denotes a skill family, and circles represents skills. We also show the position of the largest skill cluster node labeled "not applicable" (NA) on the right. The disproportionate relegation of skills to the NA cluster showcases difficulties in capturing the complex structure and relationship of these unclassified skills and their resistance to organization within a strict hierarchical scheme.

The basemap in Fig. 2 assigns a 2D position to each of 13,218 skills. The MaxMatch results provide a lookup table with information on which of the 13,218 skills in the BG skills taxonomy are associated with a job, course, or publication. The BG taxonomy provides a classification of skills that distinguishes "base" skills, which are applicable to a wide range of positions (e.g., Leadership), from "technical" skills, which are often quantitative and can be job defining or specific (e.g., Oracle). Although these correspond loosely to what might be considered soft vs. hard skills by executives (30), the association is imperfect. We recoded the BG skills to more clearly distinguish quantitative and technical skills (hard) from (3) social and communication skills (soft) (31). Details on our classification of hard and soft skills are in *SI Appendix*, *SI Text*.

The combination of these three datasets makes it possible to take any subset of skills—associated with any subset of jobs, courses, and/or publications—and overlay them on the basemap.

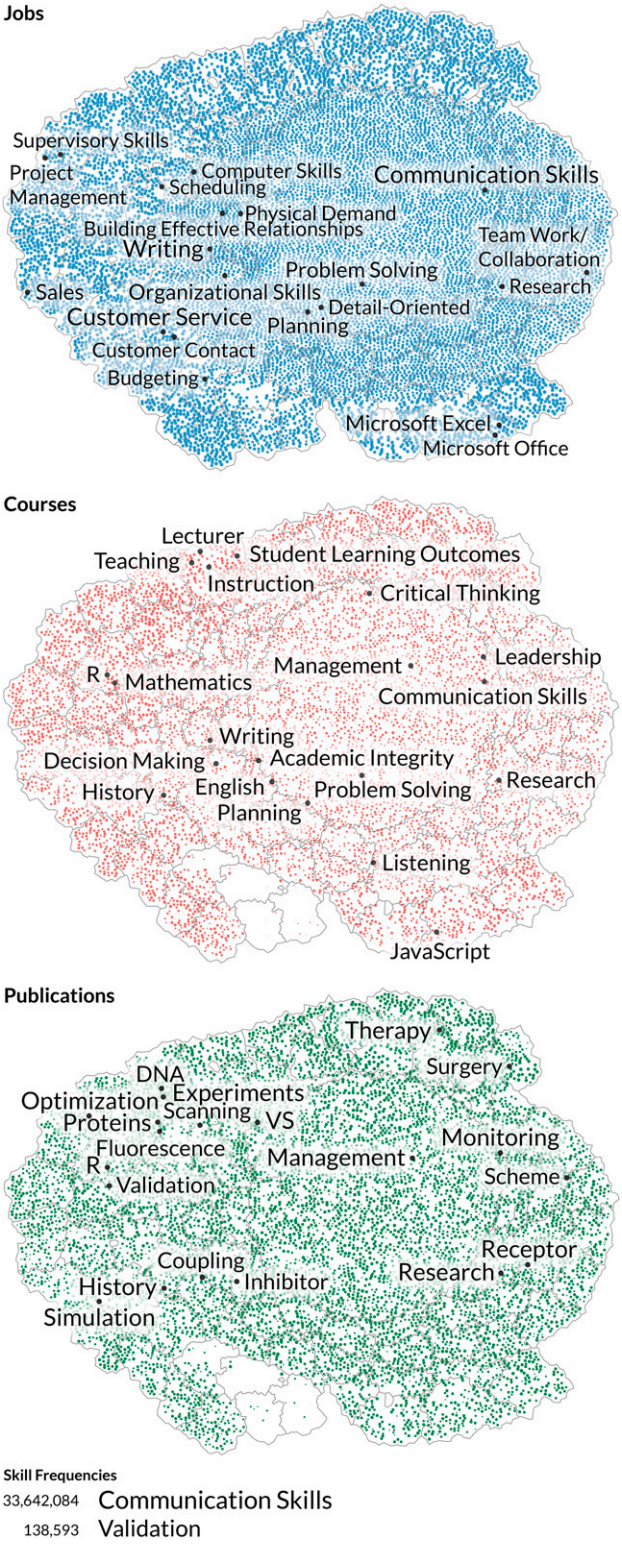

Fig. 3. Basemap of 13,218 skills with overlays of skill frequency in jobs, courses, and publications. This figure substantiates the conceptual drawing in Fig. 1 using millions of data records. Jobs skills are plotted in blue, courses are in red, and publications are in green. Node area size coding indicates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

In this paper, soft skills are rendered in orange, and hard skills are in purple. Exemplarily, Fig. 2 shows an overlay of the 45 skills discussed in *Analyzing Dynamic Skill Gaps and Flows of Influence*; each node is area size coded by skills frequency and color coded by skill type. Many soft skills are in the NA skill cluster, while several hard skills are in the Information Technology skill family area in the lower right of the basemap. Using this process, (soft and hard) skills in job, course, and publication data can be overlaid on the same basemap, and additional data variables can be used to size, color, or shape code nodes in support of comparisons.

Interested to compare the topical landscape of skills and the most frequently used skills in the three datasets, we use the basemap from Fig. 2 to rerender the conceptual drawing in Fig. 1 using all data records. The three maps in Fig. 3 show what skill terms are most frequency used in jobs (13,218 skill terms given in blue), courses (8,473 skills in red), and publications (8,856 skills in green). Each dot denotes a skill color coded by the number of times that it occurred in a job, course, or publication. The top 20 most frequent skills per dataset are labeled. For example, Communication Skills are listed in 33,642,084 jobs, and Validation is mentioned in 138,594 publications.

A comparison across maps reveals that some skills (e.g., Research) are listed frequently in all three record types; Communication Skills and Writing are often mentioned in job advertisements and course syllabi. Other skill areas show almost no commonality between the datasets (e.g., in publications, the most frequently mentioned skills are DNA, Experiments, and Proteins). Here, commonalities and differences between demandable, teachable, and researched skills can be explored using millions of data records in support of data-driven decision making.

Burst of Activity in Core Workforce Needs and Research Publications in DS/DE

In this section, we focus on skills most pertinent for the data economy. Using the BG-defined job categories, we identify 39,998 Data Science and 29,407 Data Engineering jobs posted between 2010 and 2016. The total set of 69,405 DS/DE jobs has 2,980 unique skills; 575 of these skills are assigned to the NA skill cluster (Fig. 2). Using the 2,980 DS/DE skills, we extracted 2,731,866 DS/DE courses and 803,993 DS/DE publications.

We are interested to understand what skills experience a sudden increase in usage frequency, indicating a surge in workforce needs or research and publication activity. We do this by running the burst detection algorithm, developed by Kleinberg (32) and available in the Sci2 Tool (33), on the skill terms associated with DS/DE jobs and publications. The algorithm reads a sequence of time-stamped terms, here skill terms, and identifies those terms that experienced a sudden increase in usage frequency over the 7 years. As for the skills occurring in DS/DE jobs and publications, 987 skill terms burst in jobs and 202 burst in publications between January 2010 and December 2016. The top 10 bursts in jobs and publications together with burst strength are provided in *SI Appendix*, Table S5. Several skills showed more than a single burst: 39 skills in jobs (e.g., B2B, Social Services, Decision Tree) and three in research publications (e.g., NoSQL, Apache Hadoop, MapReduce). There are 87 skill terms that burst in jobs and publications. From these, we select the top five terms with the highest burst strength from each set (jobs and publications, Android is in both sets) plus skills that burst more than once (in jobs: Social Gaming, Storage Systems, Maximo, HRMS, Document Management; in publications: Apache Hadoop) and plot them as a horizontal bar graph in Fig. 4. With time running from left to right in Fig. 4, we can examine the temporal dynamics of skills over the 7 years. Some skills burst first in jobs, then in publications, and vice versa. For example, Facebook, Industrial Engineering, and Android first appear in job postings (blue in Fig. 4) and then move to publications (green in Fig. 4). In contrast, Document Management burst in publications

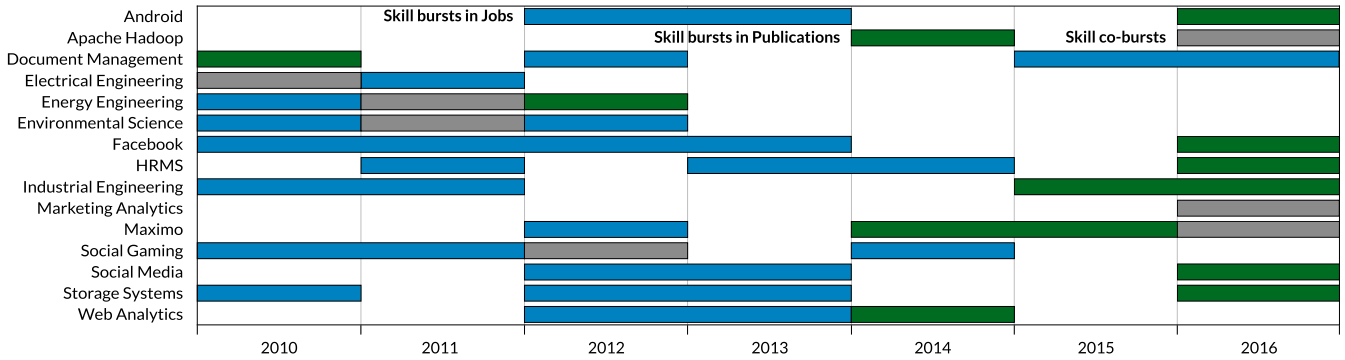

Fig. 4. Burst of activity in DS/DS skills in jobs and publications. Each burst is rendered as a horizontal bar with a start and an end date; skill term is shown on the left. Skills that burst in jobs are blue; skills bursting in publications are green. Seven skills burst in both datasets during the same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

first and then in jobs. Some skills burst several times (e.g., Storage Systems bursts first in the employment sector in 2010 and then again in 2012–2013 before it bursts in publications in 2016). Seven skills burst in both datasets during the same years and are shown in gray in Fig. 4: Apache Hadoop, Electrical Engineering, Environmental Science, Energy Engineering, Marketing Analytics, Maximo, and Social Gaming. In some cases, simultaneous bursts bridge between dataset-specific bursts. For example, Energy Engineering bursts in both publications and jobs in 2011—right between an initial burst in jobs (2010) and a subsequent in publications (2012). There are fewer bursts in 2014 and 2015, which might be due to economic cycles; lack of technological innovation or disconnect between industry needs and academic research; slowdown in innovation, funding, or both; or saturation—reasons suggested by expert survey participants that examined this visualization. A more detailed analysis of external events might be needed to explain the lower number of bursts in these 2 years. Interestingly, for most of the skills featured in Fig. 4, job skill bursts precede skill bursts in research publications. We study these dynamics more precisely in analyses presented subsequently.

Analyzing Dynamic Skill Gaps and Flows of Influence

To test the structural and dynamic alignment of skills listed in jobs, courses, and publications, we analyze skill networks and evaluate covarying time series of skills.

To examine the centrality of skills as they co-occur in job advertisements, course syllabi, and research publications, we use an embedding technique operationalized by the Facebook Artificial Intelligence (AI) group (34). The circles are Poincaré disks that use hyperbolic geometry to represent any tree-like hierarchy without distortion, locating the trunk at the center and leafs at the periphery (35, 36). The three Poincaré disks pictured in Fig. 5*A* represent the hierarchy of skill terms explored in research, taught in courses, and listed in jobs. On each disk, the position of each node I is defined by two parameters in Euclidean polar coordinates, radius r_i and angle i. The radius quantifies position in the hierarchy: skills of small radius hold a central position in the network of co-occurrences. The angle between two skills $i - j$ quantifies their proximity or structural similarity (37). As in Fig. 2, soft skills are rendered in orange, and hard skills are in purple. We see that across jobs, courses, and publications, soft skills like Management are substantially more central than hard skills, meaning that they are more likely to co-occur with hard skills like Optimization and Data Analysis than hard skills are to occur with each other. Nevertheless, soft skills are most central in jobs, somewhat less central in course syllabi, and much less central in research publications. Note especially how Communication Skills are critical to a wide variety of jobs but are only mentioned in a few courses and research publications. This highlights a substantial discrepancy between how central soft skills are to the workplace, including in technical positions, but how peripheral they are to technical courses and publications.

Interested to estimate the (mis-)match of skills in jobs, courses, and publications, we calculate Pearson correlations between the prevalence of skills and skill pairs in each dataset. The correlation between the prevalence of 13,218 individual skills in jobs and publications is very low (0.08), and the correlation of millions of skill pairs between jobs and publications is negative (−0.05). Courses strongly mediate this chasm—individual skills taught in courses correlate at 0.64 with those in jobs and 0.42 with those in publications. Correlations between skill pairs in courses and jobs/publications are somewhat lower but still large and highly statistically significant at 0.54 with jobs and 0.19 with publications.

Next, we explore skill associations between the three datasets dynamically to assess the degree to which skill emergence in publications influences industry needs in job advertisements, business priorities call forth additional research, and educational course offerings mediate the two. We compile all BG skills for papers, courses, and jobs for two time periods (2010–2013 and 2014–2016) and turn each skill frequency into a probability by normalizing by the total number of skills. Then, we assess the Kullback–Leibler (KL) divergence between the six skill probability distributions (38, 39) (*SI Appendix*, *SI Text*). The KL divergence, also called relative entropy, calculates the information gain experienced when an existing probability P_1 is confronted with a new one P_2. Specifically, when $\mathrm{KL}\,(P_1||P_2) = 0$, the distribution of two distributions is equal, which implies no information gain or surprise. The larger the KL value, the greater the divergence between them and the surprise that would be experienced by seeing one from the perspective of the other. In Fig. 5*B*, we plot the KL divergence values in a matrix, showing the divergence from skill distributions in jobs, courses, and publications. Due to the asymmetric nature of KL divergence, we need to look separately at the divergence (*i*) from publications to education and jobs, (*ii*) from education to publications and jobs, and (*iii*) from jobs to education and publication. The matrix reveals that the skill distributions from jobs, education, and publications in 2010–2013 are similar to skills in those same sectors in 2014–2016, with job skills changing the most between the two periods (0.2 vs. 0.01 or 0.04). The distribution of skills taught in the classroom is between three and four times closer to skills described in research articles than skills from job advertisements, but education still plays a critical mediating role between skills in jobs and research. Finally, gaps between skills in research and jobs and those in education and jobs appear to decrease with time, being substantially lower in 2014–2016 than in 2010–2013.

Next, we look at individual skills disproportionately associated with jobs and publications. In Fig. 5*C*, we see that industry skills

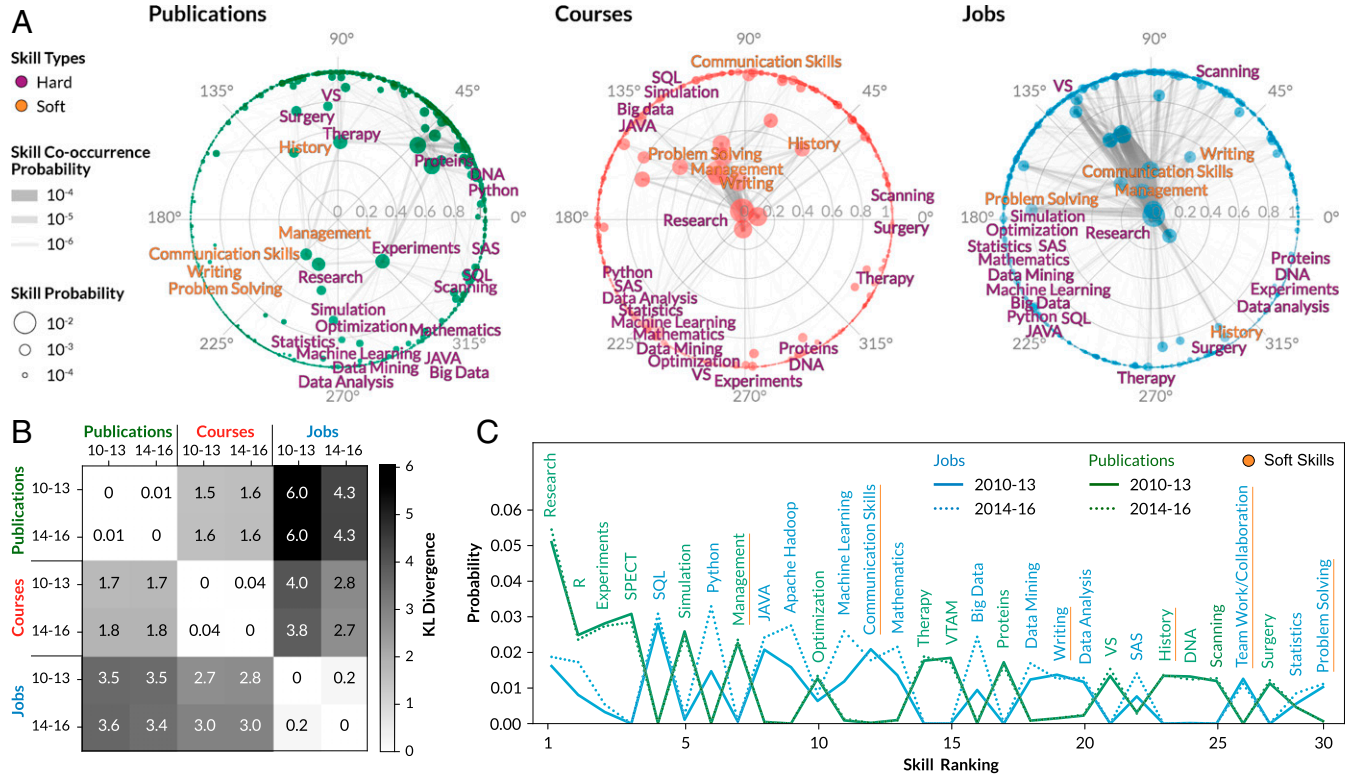

Fig. 5. Structural and dynamic differences between skill distributions in jobs, courses, and publications for 2010–2013 and 2014–2016. (*A*) Poincaré disks comparing the centrality of soft skills (orange) and hard skills (purple) across jobs, courses, and publications. (*B*) KL divergence matrix for 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.

demanded by DS/DE jobs surprisingly require social skills, such as Communication Skills, Writing, Team Work/Collaboration, and Problem Solving, while publication skills most surprising to jobs advertised by industry include the *R* scripting language, Experiments, Single Photon Emission Computed Tomography (SPECT), and Simulation. This suggests a potentially undersupplied need for soft skills alongside technical ones.

To examine shifts of attention on a skill-by-skill basis, we compute the number of times that a skill term is mentioned for each year in 2010–2016 for all skills in all jobs and publications. We then run a time series analysis using Granger causality (40) to resolve at higher resolution the degree to which skills garnering attention at the frontier of research are driving skill demands in industry or the converse. Granger causality is a statistical approach that predicts if a signal in one time series "Granger-causes" a signal in another time series or the converse. The result of analyzing all 7 years is an almost symmetrical distribution of influence: 143 publication skills posit a statistically significant effect on jobs, and 147 job skills exhibit a statistically significant effect on publications. Interestingly, only 10 skills were significant in both directions (*SI Appendix*, Table S6). Fig. 6 uses the skills basemap introduced in Fig. 2 to show the impact of jobs (blue in Fig. 6, *Upper*) on publications (green in Fig. 6, *Lower*) and the converse via arrows. Arrows are thickness coded to indicate the *F*-value strength from jobs to publications (blue arrows in Fig. 6) and from publications to jobs (green arrows in Fig. 6). Skills that have a significant Granger causality (P value < 0.05) are labeled. These skills can be used to help predict future values of the same skills in another dataset. For example, Immunology listed in publications with an F value of 10,417 is highly predictive of Immunology listed in jobs in future years. Three skills (Computational Tools, Product Knowledge, and Trial Design) have a significant Granger causality in both directions.

When aggregating the Granger causality results at the skill family level of the BG taxonomy (listing in *SI Appendix*, Table S7), we find that roughly the same number of skill families seem to be driven from industry to academy as the converse. For example, job advertisements within the Education and Training, Media and Writing, and Administration sectors tend to follow growth or decline of associated skills in the publications (publication → job). By contrast, skills in Industry Knowledge, Environment and Economics, Policy, and Social Studies job advertisements anticipate increased or decreased focus in academic research (job → publication). This suggests that shifts of skill attention within research are as likely to come from industry as the converse, reinforcing our characterization of a dynamic and symmetrical system of skill supply and demand. More data on education, research, and jobs over a longer duration will be required to identify these dynamics with greater precision.

Finally, we apply the Multivariate Hawkes Process (MHP) model (41, 42) to learn more about the relationship between soft and hard skills by exploring the predictive and potentially causal relationship between skills listed in DS/DE job advertisements. Specifically, we take the top 50 most frequent skills and apply the MHP model to compute the matrix of inferred influences between these 50 skills. We then extract the 75 directed edges with the highest influence values plus the associated 45 skill nodes (29 soft and 16 hard skills) and lay out the influence network using the ForceAtlas2 layout in Gephi (Fig. 7). In contrast to KL divergence analysis, the Hawkes analysis allows us to examine the influence between specific skills. Exploring the directed network in Fig. 7, we find that hard skills, such as JAVA programming, predict the rise in soft skills, such as Team Work/Collaboration and Creativity; that Research skills are important for Teaching; that Microsoft PowerPoint skills give rise to Product Sales; and that Budgeting is predictive of Presentation skills. The ring shape

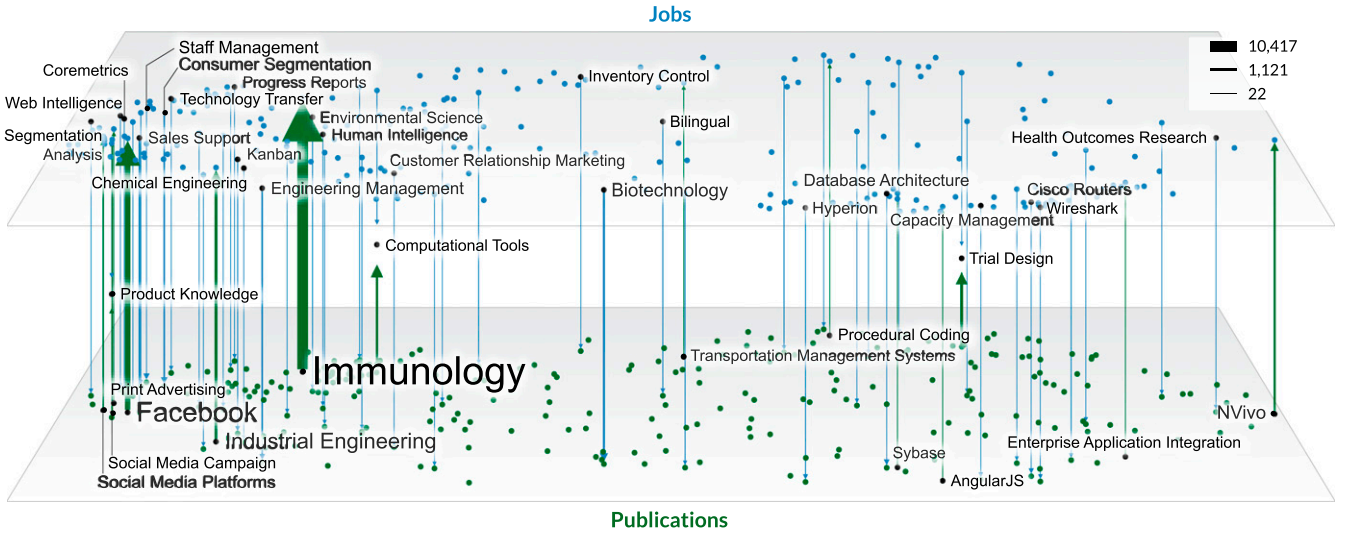

Fig. 6. Strength of influence mapping. Top 200 most frequent skills in jobs (blue) and in publications (green) plotted on the skills basemap from Fig. 2. Arrows represent skills with significant Granger causality (P value < 0.05). Line thickness and label size indicate skill frequency. The direction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

of the network is noteworthy—there is no beginning or end of influence; soft and hard skills influence each other recursively in a continuous cycle.

Expert Survey

A survey was conducted involving 20 labor market and educational domain experts from academia, industry, government, and the not-for-profit sector to examine the readability of the visualizations and the utility of results presented in this paper. The data collection and analyses were conducted per the requirements specified by the institutional review board (IRB) for human subjects research at Indiana University (protocol 1803748120). Informed consent was not required as the IRB determined that this study was exempt. Specifically, we were interested to understand what answers the visualizations provided and what new questions they inspired. Study participants completed a prequestionnaire that gathered basic demographic information and information on expertise with data visualizations (*SI Appendix*, *SI Text* has details). Next, participants were sorted randomly into two groups to give feedback on either Figs. 1–4 or Figs. 2, 5, and 6. Participants spent a median time of 25 min to complete the survey.

As expected, the needs of the diverse stakeholders differ substantially. Asked "How might you use data on job market trends, educational programs, and science and technology developments?," the 10 academic, 4 corporate, 4 government, and 1 not-for-profit experts provided the answers listed in *SI Appendix*, Table S8. Interestingly, all groups have a major need for providing policy advice to decision makers. Questions related to course development, selection, or curriculum design are mostly relevant for academia. When asked to answer questions about and give feedback on Figs. 1–6, participants were able to answer the questions posed and provided thoughtful comments that helped inform the interpretation of figures in the final version of the paper. Problems with and critiques of the legibility of figures led to redesigns.

Asked "What visualization was most interesting/useful for your decision making and why?," one participant picked Fig. 1, as it was helpful for seeing industry trends in relation to academia. Another participant picked Fig. 2, as it has the potential to show skill similarity and relationships. Two picked Fig. 3, arguing that it helps compare the presence of DS/DE skills in jobs, courses, and publications, and four selected Fig. 4, which shows bursts, as it provides information on leading and lagging trends for both industry and research.

Discussion

We presented data analyses and visualizations performed to understand discrepancies and temporal delays between evolving job market needs, course offerings, and S&T developments. As part of this work, we developed a topical basemap covering 13,218 skills that occur in job, course, and publication data; showed the evolution of skill bursts over time; conducted an analysis of skill centrality; performed a causal analysis of temporal delays between evolving job market needs and scientific developments; and examined the interplay of soft and hard skills.

Our findings suggest that educational efforts play an important role in mediating between the needs of industry and research.

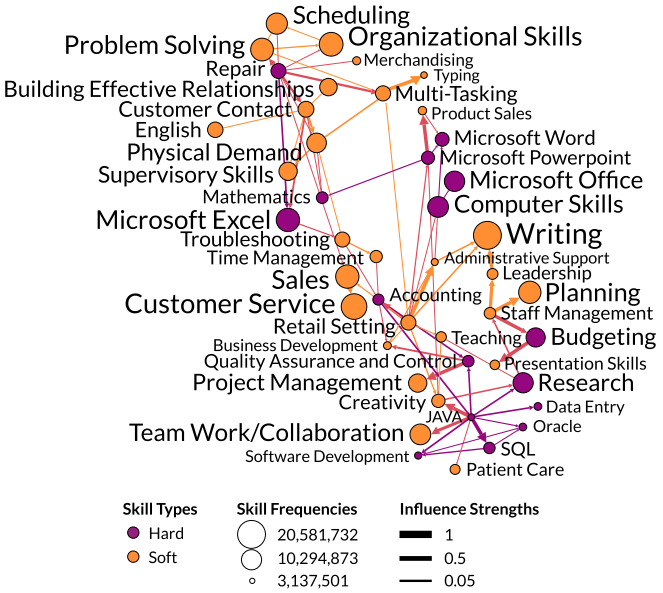

Fig. 7. Hawkes influence network of DS/DE skills within job advertisements 2010–2016. Each of the 45 nodes represents a top frequency skill (29 soft and 16 hard skills) with a strong influence edge from/to other skill(s) in job advertisements between 2010 and 2016. Node and label size correspond to the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

Moreover, we show a pattern of broad symmetry or balance between the causal flow of skill research and skill demand, suggesting the important force of industry demands on influencing research and converse. We do not find evidence of an ivory tower insularly disconnected from the needs of the enterprise. Despite gaps between skills in research, teaching, and industry, this analysis suggests that the skill system is bilaterally responsive, which bodes well for efforts to align sectors for enhanced economic growth and prosperity. Both our KL divergence and Hawkes process analyses show tight connections between technical data science and engineering skills and soft communication, presentation, and teamwork skills. This suggests that, even with the rise of data analytic, machine learning, and AI-related skills, people will continue to be needed to communicate complex ideas, negotiate, and lead. Nevertheless, our structural analysis identifies a substantial gap between the centrality of soft skills for technical jobs and their relative peripherality to technical coursework and especially, technical research publications.

Results are relevant for the stakeholder groups discussed in the Introduction and for experts in academia, industry, and government (*SI Appendix*, Table S8). In general, students can use the maps and analysis results to understand what skills are required for what jobs and what skills are taught in what courses. Teachers may examine job market and S&T developments to decide what course updates are needed, which curriculum design is best, or how to best differentiate a course in the market place of educational offerings. Universities can examine results to decide what programs should be created and how they should be marketed. Employers can use models and visualizations to identify potential skill gaps in the workforce, note which courses and institutions produce the right talent, or evaluate which skill combinations competitors list in job advertisements. Economic developers might render maps for specific regions to understand discrepancies between workforce needs and educational offerings to make more strategic decisions when it comes to workforce training as well as business retention, expansion, and recruiting in a region. Last but not least, providers of national and international statistics [e.g., the US National Science Foundation's National Center for Science Engineering Statistics or the Going Digital project by the Organisation for Economic Co-operation and Development (OECD)] might use the presented methodology in their analyses to empower policy makers and industry strategists to make data-driven decisions using high-quality data.

ACKNOWLEDGMENTS. We thank Leonard Cross and Tracey Theriault for design support for figures; Bledi Taska for arranging access to BG jobs data and providing expert comments on a previous version of the paper; Stephen Kobourov and Iqbal Hossain for expert assistance using GMap; the 20 surveyed experts for providing valuable feedback on initial figures; Elizabeth Record for copyediting, and reviewers for thoughtful comments and suggestions. This work uses WoS data by Clarivate Analytics from the Indiana University Network Science Institute (IUNI) WoS Data Enclave. K.B. and O.S. are partially supported by NIH Grants P01 AG039347 and U01CA198934 and National Science Foundation (NSF) Grants 1566393, 1839167, and 1713567. L.W. and J.A.E. are partially supported by Air Force Office of Scientific Research Grant FA9550-15-1-0162 and NSF Grants 1422902 and 1158803. 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 NSF. K.B. is Alexander von Humboldt Fellow at Dresden University of Technology, Dresden, Germany.

1. Bennett WJ, Wilezol D (2013) *Is College Worth It?* (Thomas Nelson, Nashville, TN).
2. Rouse WB, Lombardi JV, Craig DD (2018) Modeling research universities: Predicting probable futures of public vs. private and large vs. small research universities. *Proc Natl Acad Sci USA* 115:12582–12589.
3. Shah D (2017) A product at every price: A review of MOOC stats and trends in 2017. Available at https://www.class-central.com/report/moocs-stats-and-trends-2017/. Accessed February 27, 2018.
4. Rouse WB (2016) *Universities as Complex Enterprises: How Academia Works, Why It Works These Ways, and Where the University Enterprise Is Headed* (Wiley, Hoboken, NJ).
5. Stokes DE (1997) *Pasteur's Quadrant: Basic Science and Technological Innovation* (Brookings Institution Press, Washington, DC).
6. Martinez W (2018) How science and technology development impacts employment and education. *Proc Natl Acad Sci USA*, 10.1073/pnas.18201803216115.
7. O*NET Resource Center (2018) What's new? O*NET database and websites updated. Available at https://www.onetcenter.org/whatsnew.html. Accessed March 7, 2018.
8. Lewis P, Norton J (2016) Identification of "hot technologies" within the O*NET system. Available at https://www.onetcenter.org/reports/Hot_Technologies.html. Accessed July 27, 2018.
9. Markow W, Braganza S, Taska B, Miller S, Hughes D (2017) The Quant crunch: How the demand for Data Science skills is disrupting the job market. Available at https://www.burning-glass.com/research-project/quant-crunch-data-science-job-market/. Accessed December 30, 2017.
10. LinkedIn Corporation (2018) LinkedIn economic graph research. Available at https://engineering.linkedin.com/teams/data/projects/economic-graph-research. Accessed July 27, 2018.
11. EMSI: Labor Market Analytics (2018) Available at www.economicmodeling.com/. Accessed July 27, 2018.
12. CareerBuilder; EMSI (2014) In jobs recovery, boomers make up growing share of workforce while millennial employment lags. Available at https://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?ed=12%2F31%2F2014&id=pr826&sd=6%2F5%2F2014. Accessed February 27, 2018.
13. Reis F (2018) European big data Hackathon. Available at https://ec.europa.eu/eurostat/cros/EU-BD-Hackathon_en. Accessed July 27, 2018.
14. Opik R, Kirt T, Liiv I (2017) Megatrend and intervention impact analyser for jobs: A European big data hackathon entry. arXiv:1708.08262.
15. Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? *Technol Forecast Soc Change* 114:254–280.
16. Barker K, Christensen K (1998) Controversy and challenges raised by contingent work arrangements. *Contingent Work: American Employment Relations in Transition*, eds Barker K, Christensen K (Cornell Univ Press, Ithaca, NY), pp 1–20.
17. Kalleberg AL (2009) Precarious work, insecure workers: Employment relations in transition. *Am Sociol Rev* 74:1–22.
18. Anderson KA (2017) Skill networks and measures of complex human capital. *Proc Natl Acad Sci USA* 114:12720–12724.
19. Alabdulkareem A, et al. (2018) Unpacking the polarization of workplace skills. *Sci Adv* 4:eaao6030.
20. Frank MR, Sun L, Cebrian M, Youn H, Rahwan I (2018) Small cities face greater impact from automation. *J R Soc Interface* 15:20170946.
21. Deming DJ (2017) The growing importance of social skills in the labor market. *Q J Econ* 132:1593–1640.
22. Burning Glass Technologies (2018) Home page. Available at https://www.burning-glass.com. Accessed July 13, 2018.
23. Börner K, et al. (2012) Design and update of a classification system: The UCSD map of science. *PLoS One* 7:e39464.
24. Börner K (2015) *Atlas of Knowledge: Anyone Can Map* (MIT Press, Cambridge, MA).
25. Börner K (2010) *Atlas of Science: Visualizing What We Know* (MIT Press, Cambridge, MA).
26. Gansner E, Hu Y, Kobourov S (2010) GMap: Visualizing graphs and clusters as maps. *Proceedings of the 2010 IEEE Pacific Visualization Symposium (PacificVis 2010)* (Institute of Electrical and Electronics Engineers, Taipei, Taiwan), pp 201–208.
27. Saket B, Scheidegger C, Kobourov S, Börner K (2015) Map-based visualizations increase recall accuracy of data. *Comput Graph Forum* 34:441–450.
28. Saket B, Simonetto P, Kobourov S, Börner K (2014) Node, node-link, and node-link-group diagrams: An evaluation. *IEEE Trans Vis Comput Graph* 20:2231–2240.
29. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International Conference on Weblogs and Social Media*, eds Adar E, et al. (AAAI Press, Menlo Park, CA), pp 361–362.
30. Robles MM (2012) Executive perceptions of the top 10 soft skills needed in today's workplace. *Bus Commun Q* 75:453–465.
31. Andrews J, Higson H (2008) Graduate employability, 'soft skills' versus 'hard' business knowledge: A European study. *High Educ Eur* 33:411–422.
32. Kleinberg J (2002) Bursty and hierarchical structure in streams. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining–KDD '02* (ACM Press, New York), pp 91–101.
33. Sci2 Team (2009) Science of Science (Sci2) tool. Indiana University and SciTech strategies. Available at https://sci2.cns.iu.edu. Accessed July 18, 2018.
34. Nickel M, Kiela D (2017) *Poincaré Embeddings for Learning Hierarchical Representations*, Advances in Neural Information Processing Systems (NIPS), eds Guyon I, et al. (Neural Information Processing Systems Foundation, Long Beach, CA), Vol 30.
35. Papadopoulos F, Kitsak M, Serrano MÁ, Boguñá M, Krioukov D (2012) Popularity versus similarity in growing networks. *Nature* 489:537–540.
36. Krioukov D, et al. (2012) Network cosmology. *Sci Rep* 2:793.
37. Wu L, Li L, Evans J (2018) Social connection induces cultural contraction: Evidence from hyperbolic embeddings of social and semantic networks. arXiv:1807.10216v1.
38. Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley-Interscience, New York), 2nd Ed.
39. Ebrahimi N, Soofi ES, Soyer R (2010) Information measures in perspective. *Int Stat Rev* 78:383–412.
40. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37:424–438.
41. Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58:83–90.
42. Morse ST (2017) Persistent cascades and the structure of influence in a communication network. Master's thesis (Massachusetts Institute of Technology, Cambridge, MA).