JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 01/06 pp191−205
Volume 3, Number 3, September 2022
DOI: 10.23919/JSC.2022.0013

Misalignment Between Skills Discovered, Disseminated, and Deployed in the Knowledge Economy

Bhargav Srinivasa Desikan* and James Evans*

Abstract: The knowledge economy is a complex and dynamical system, where knowledge and skills are discovered through research, diffused via education, and deployed by industry. Dynamically aligning the supply of new knowledge with the demand for practical skills through education is critical for developing national innovation systems that maximize human flourishing. In this paper, we evaluate the complex alignment of skills across the knowledge economy by creating an integrated semantic model that neurally encodes invented, instructed, and instituted skills across three major datasets: research abstracts from the Web of Science, teaching syllabi from the Open Syllabus Project, and job advertisements from Burning Glass. Analyzing the high dimensional knowledge and skills space inscribed by these data, we draw critical insight about systemic misalignment between the diversity of skills supplied and demanded in the knowledge economy. Consistent with insights from economic geography, demand for skills from industry exhibits high entropy (diversity) at local, regional, and national levels, demonstrating dense complementarities between them at all levels of the economy. Consistent with the economics and sociology of innovation, we find low entropy in the invention of new knowledge and skills through research, as specialist researchers cluster within universities. We provide new evidence, however, for the low entropy of skills taught at local, regional, and national levels, illustrating a massive mismatch between diversity in skills supplied versus demanded. This misalignment is sustained by the spatial and institutional mismatch in the organization of education by researchers at the site of skill invention over use. Our findings suggestively trace the societal costs of tethering education to researchers with narrow knowledge rather than students with broad skill needs.

Key words: computational content analysis; knowledge economy; sociology of knowledge; natural language processing; data science and society

1 Introduction

The knowledge economy represents the "production and services based on knowledge-intensive activities that contribute to an accelerated pace of technical and scientific advance, as well as rapid obsolescence"[1]. The dynamic components of a knowledge-based economy are based on intelligent and knowledge-based capabilities and products as opposed to natural resources and physical inputs, and trace a shift in the USA and Europe from agriculture and mining to manufactured products to increasingly sophisticated knowledge-based services. Broadly, three sets of institutions play distinctive roles within the knowledge economy: private and public research centers and laboratories discover new knowledge and skills; universities, institutes, and academies disseminate established and emerging skills through instruction; and private companies deploy those skills through the

• Bhargav Srinivasa Desikan is with the Department of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland. E-mail: bhargav.srinivasadesikan@epfl.ch.
• James Evans is with the Knowledge Lab, the Department of Sociology, University of Chicago, Chicago, IL 60615, USA, and also with the Santa Fe Institute, Santa Fe, NM 87501, USA. E-mail: jevans@uchicago.edu.
* To whom correspondence should be addressed.
Manuscript received: 2022-08-19; revised: 2022-10-19; accepted: 2022-10-20

organization of industrious action.

While research, teaching, and production represent stable roles within national innovation systems[2], they do not flow linearly from one to the next just as science does not linearly mature into technology[3, 4]. Jobs forged through the capitalist process of creative destruction—entrepreneurship and industrial competition—are just as likely to generate new combinations of skills than research that invents new capacities from theory as the converse[5]. Framed in this way, one of the greatest challenges for advancing human flourishing through the knowledge economy is to improve the alignment of skills invented through research, supplied via teaching, and demanded by industry. In this paper, we align the diversity of skills across levels of the US knowledge economy to understand the potential for education to better service job needs.

The US academic and industrial institutions vary substantially in the amount and character of research, teaching, and goods or services they produce. As knowledge-intensive services and products come to represent an increasing share in major economies, we must better understand and manage the dynamics, distribution, and alignment of knowledge and skills across the knowledge economy. Here we explore the potential for knowledge and skills alignment across geographic scales, from cities to regions to the United States as a whole. The study of industrial innovation[6] and the knowledge economy has long attended to the importance of geography, largely due to the emergence of high output knowledge-based products and services in dense geographic clusters[7, 8]. With modern cities being the hubs of innovation around the world[9−11], they represent a critical interface between the supply of knowledge and skills and demand for them in the economy[12], with scaling relationships observed between population size, density, and productive outputs like the number of patents, Gross Domestic Product, and highly skilled professionals[13−16]. These scaling properties are attributed to the fundamentally social nature of knowledge-based industries, which rely on intensive human engagement to invent, transfer, deploy, and update knowledge-intensive skills. Next, we consider the region, operationalized as the US states, all of which share a policy environment and possess both educational supply and employment demand. Finally, we consider national innovation systems that account for broader educational and employment migration, which is increasingly common in knowledge-based economies[2]. Nevertheless, few studies have considered the alignment of skill discovery, dissemination, and deployment across these levels, as we do here.

We study alignment in terms of the diversity of knowledge and skills produced at each scale within the national innovation system in order to identify the potential for cities, regions, and the US as a whole to supply skills required by the workforce. We consider knowledge at the stages of discovery (research), dissemination (teaching), and deployment (industry) to be aligned with respect to its skills if there exist comparable measures of diversity for each stage. A consequence of the systematic misalignment of skill diversity at each of these stages would result in workers unprepared for the complex and diverse tasks that they are expected to perform. While scaling relationships between population size and skill diversity have been demonstrated within cities[14], they have never been explored across sectors and scales using the semantic content of the documents that natively signify it—titles and abstract for research, course syllabi for teaching, and employment advertisements for jobs. Semantic diversity[17] can be measured in a straightforward way using natural language processing tools that encode documents to allow the estimation of precise distances between them.

In this paper, we extend the geographic and scaling approach to understand the capacity for alignment across the knowledge economy through the lens of text as data and the diversity of knowledge-based content [18]. Semantic diversity has often been explored through natural language processing with a focus on inter-word and inter-document distances based on a large corpus of documents[19]. By representing words in a high dimensional space based on co-occurrence in text[17, 20], it is possible to measure distances between words and documents, and to construct well-behaved metrics that capture word ambiguity and diversity. By using large-scale datasets, we can create semantic vector representations of corpora that simultaneously capture knowledge-based skills discovery through research, skills dissemination through syllabi, and skills deployment through job ads in the knowledge economy.

By using the Web of Science for research abstracts, the Open Syllabus Project for teaching syllabi, and Burning Glass* for job demands, all embedded within a unified, self-normalizing space, we create aggregate vectors that represent the semantic content of these three major pillars of the knowledge economy. We use the popular doc2vec[21] method to create these vectors for each individual document, and aggregate them based on their organizational association (e.g., University of Chicago and Microsoft Research) and geographic location (Metropolitan Statistical Area (MSA) and State, e.g., Chicago MSA and Illinois). With these aggregate research, teaching, and job demand vectors, we can use straightforward distance metrics between entity vectors to measure the diversity and aggregation patterns across the landscape of the US knowledge economy.

Using aggregate entity vectors, we find that job ads are more diverse than research, which is more diverse than course syllabi at all levels of aggregation, reflecting the interconnection between diverse skills required for a local economy to run. Further, while jobs and teaching are equally diverse at different geographic levels, research is clustered within cities and states, such that similar semantic content reflects specialized, geographically localized clusters of skill investigation and discovery. We also examine scaling relationships between volume and semantic diversity for research, teaching, and jobs within MSA populations. This reflects the balance of supply and demand for local skilled persons. We observe a super-scaling relationship in the number of job ads, and find that while research and teaching increase in semantic diversity with city size, job demands specialise for the largest cities.

2 Related Computational Work

Textual data have been used in many contexts for social scientific analysis and to quantitatively validate and extend qualitative research. Such methods can be used to explore content across a wide variety of contexts. Text is produced as a natural byproduct of communication in social life, ranging from conversation transcripts, political and legal proceedings, and historical documents to news, cultural programming, and social media. Here we use granular text describing discrete articles, courses, and jobs, linking these artifacts by their qualities to quantitatively trace the flow of skill supply and demand across the knowledge economy. Combining relevant data sources for longitudinal comparison allows us to go beyond simply classifying or annotating text to unraveling complex social processes such as networks and hierarchies[22, 23], temporal shifts in language and meaning[24, 25], and dynamics in conversation and debate[26, 27]. The use of text as data in the realm of complex social and economic phenomena is recent[18, 24] and we aim to contribute to this growing literature.

Many recent attempts have used word embeddings and high-dimensional semantic spaces for knowledge representation and discovery. Embeddings built from corpora of common chemicals and material science documents can approximate chemical knowledge sufficient to uncover the underlying structure of the periodic table[28], capture structure-property relationships in materials[29], and predict materials that will be discovered in the future to possess a given property. By identifying the semantic dimensions of semantic spaces, Kozlowski et al.[30] demonstrated how cultural dimensions such as race, gender, and class learned from word embeddings correspond to attitudes elicited directly from contemporary cultural surveys, while tracing historical stereotypes[31], implicit cultural biases[32], and representations of human knowledge[33] with word embeddings. This research validates that relationships emergent in these learned spaces correspond in direction and magnitude with widely shared cultural meanings.

These attempts to chart semantic and knowledge structures using embedding methods have paved the way for further explorations in the social sciences, and specifically, the knowledge economy. There have been early attempts to use deep learning to predict knowledge economy indices[34], manage collective knowledge[35], and explore the relationship between diversity and performance in knowledge-producing teams[36]. Nevertheless, existing work on text data and the knowledge economy has been limited to studying gaps between education and occupation[37], and between education and innovation[38]. These recent papers attempt to use a text representation to measure the alignment and gap between different domains. Here, rather than focus on the gap between domains, we rather focus on the diversity within each, and their (mis) alignment.

*burning-glass.com (https://www.burning-glass.com)

Our attempt is the first simultaneous characterization of semantic diversity among different aspects of the knowledge economy, as well as the geographic and scaling properties of knowledge and skills in that economy.

3 Data and Computing

We use three major datasets to create semantic spaces covering time slices during 2010−2018. For the research component of the knowledge space, we use research article abstracts from Web of Science dataset; for the teaching component, we use online syllabi collected online through the Open Syllabus Project; and for the jobs component, we use advertisements curated by Burning Glass from all major online job boards in the United States. We detail each dataset below.

The motivation for using these three datasets is that they offer a way to measure multiple relationships within the knowledge economy. Research, teaching, and jobs also allow us to view the emergent role that skills play in the knowledge economy. With a view of research as inventing skills, teaching as disseminating skills, and industry as deploying skills (Fig. 1), we can measure the diversity of the skill space at different geographic levels, and their potential alignment or misalignment. By embedding these documents simultaneously in the same vector space, the distances are naturally normalized—diversity among job ads will be on the same scale as that among research abstracts and course syllabi. We note that these data represent the richest textual traces of the US research, teaching, and jobs of which we are aware.

We aggregate vectors for each entity by averaging over all the constituent document vectors associated with that entity. We use the doc2vec method of distributed representations to create our embeddings and associated document vectors.

$$\boldsymbol{V}_{entity} = \frac{1}{N_1}\sum_{n=0}^{N_1}\boldsymbol{V}_{document_n} \tag{1}$$

where N_1 is the total number of documents associated with each entity, and $\boldsymbol{V}_{entity}$ and $\boldsymbol{V}_{document_n}$ are the vector representations of the entity and document, respectively.

Fig. 1 Transformation of skills.

3.1 Web of Science

To measure the research outputs of universities, we draw on the widely used Web of Science dataset[39]. We concatenate title and abstract as article representations to build our semantic space; then we use metadata on the city, state, and institutional affiliations of the research paper to index the article. The vector centroids of article samples allow us to aggregate research papers for each city, state, and institution (e.g., university).

The dataset contains articles published across the world, predominantly in English. For the purpose of our study we extract only papers published during the years 2010−2018 in the United States of America. Because each abstract is linked to both a university and a city, we create both university and city vectors of research. In the case of multi-authored papers, the abstract is linked to cities and universities affiliated with each author. After removing data from institutions with fewer than 100 associated publications, we retain a total of 49 911 in the United States, which host a total of 2 808 749 abstracts. These institutions are not equally distributed across the US, with higher aggregation in certain states (Texas, New York, and California) and cities (Chicago, New York, and Boston).

3.2 Open Syllabus Project

We use syllabus data to represent skills instruction in universities and cities. Syllabus data have been collected and organised by the Open Syllabus Project†, undertaken by researchers at Columbia University, using a combination of web scraping and soliciting/ receiving data from departments themselves. This is an ongoing project, it has received a fair amount of attention from the press‡, and proves to be a promising way for measuring what is being taught at universities across America (and the world). To ensure that text particular to syllabi in general is not included (e.g., timetable and misconduct policies), we comprehensively cleaned the syllabi data. We use a JSON dump of the syllabi data that was gathered in 2018, using data during the years 2010−2018.

We aggregated data similar to our process for the Web of Science (WoS) dataset with syllabi linked to

†Link to website (https://blog.opensyllabus.org/)

‡Nature article (https://www.nature.com/articles/539125a), Digital Science article (https://www.digital-science.com/news/worlds-first-open-syllabus-project-expand-new-languages-awarded-global-innovation-grant-digital-science/)

both cities and universities. After removing data from institutions that have less than 100 associated syllabi, we retain a total of 1192 entities in the United States, representing a total of 860 539 syllabi.

3.3 Burning Glass job postings

Burning Glass Technologies is an analytics company that tracks job openings, advertisements, and related data to conduct analysis and serve information about the labor market. The dataset we use was provided by the company; we use all job postings collected by the company from virtually all wide-distribution digital US job boards, 2010−2018. Using a python script, we extract the contents of the job posting and pre-process them, while registering the location of the job posting and the organisation associated with it. Similar to research and teaching vectors, we create "job demand" vectors in the shared semantic space, resulting in a total of 40 857 job entities and 7 971 173 job postings in total.

We note that job demands are somewhat more diverse than research and teaching datasets. This is from diversity in the kinds of jobs being posted. Research and teaching have a higher proportion of scientific and technical content, while the jobs vary widely to include technical, but also manufacturing and service-based employments.

Moreover, while job demands represent one critical aspect of the industry space, another important addition might include patents, which link research and jobs directly through an alternative pathway of technology invention. While we have not constructed an aligned patent space for this particular analysis, it represents a natural next step.

3.4 Methods and resources used

We built pipelines to clean and organise our three diverse data sources, estimated a common doc2vec embedding model of dimension size 100, in the distributed memory training paradigm (PV-DM), using all three sources in order to construct a common, year-specific semantic space. Data were cleaned by removing: (1) words that appeared with disproportionate frequency for the corpus, (2) common stop words specific to the domain (e.g., "timetable" in syllabi), and (3) punctuation and numbers. The lemmatized form of each resulting word was added to the modeled document. We document our code, data, and results on GitHub, and link to our GitHub repository[§]. The repository contains information on data pre-processing and cleaning, embedding space creation, vector aggregation, results, and visualisations. The provided cleaning code, trained doc2vec model, and aggregate vectors allow for reproducibility.

We refer to each aggregate vector associated with an institution (e.g., university or company) or political agglomeration (e.g., city or state) with the word "entity". Examples of entities include "University of Chicago", "Houston", and "Texas". Examples of domain-specific vectors include "Columbia University Research Vector" or "Seattle Teaching Vector". Entity vectors are created by aggregating all texts associated with that aspect of the knowledge space for a given year; for example, a research vector is calculated by first creating a high dimensional representation of each document associated with that entity in the Web of Science dataset (in conjunction with all other syllabi and jobs from the same year) and averaging. The method used to create the embedding is the doc2vec[21] implementation in `gensim`. While there exist more powerful contextual neural methods such as ELMo[40] and BERT-based sentence embeddings[41], we find the noncontextual doc2vec model to be more stable and consistent in creating embeddings for paragraph-length documents normalized by each other's length and content. Moreover, our purpose is not to project documents within a pre-trained semantic space, but to project to a space entirely shaped by our temporally tagged text for precise measurement. We used noncontextual rather than contextual embeddings[40, 41] because of the time-based specificity required for the semantic distances we estimate. Contextual embeddings improve semantic distance estimates, but are notoriously difficult to fine-tune or pre-train to a specific semantic and temporal context[42]. The benefits of contextual embedding are lost because of the deviance between our own documents and the nonspecific web-based corpora on which they were pre-trained. We use cosine distance for all of our distance-based measures involving aggregate embeddings.

Our semantic-relational framework is formed from entities and their associated research, teaching, and jobs vectors. The semantic aspect of our framework is defined by the high dimensional vector associated with each entity, domain, and year; the relational aspect

[§]GitHub link (https://github.com/bhargavvader/knowledge-economy-diversity)

involves the rich landscape of distances between various social, geographical, and economic aggregates—as between medical institutions and community colleges, large and small cities, eastern and western states, etc.

Technical details of code and data

We release all of our code and the aggregated entity vectors on GitHub¶. All the code was written in python, the pipelines use `NumPy`[43] and `scipy`[44] to perform computation and linear algebra operations, `matplotlib`[45] to perform visualisations, and analysis was done using `Jupyter/IPython` notebooks[46]. The NLP libraries primarily used are `gensim`[47] and `spaCy`[48]: their usage in NLP related pipelines is described in Ref. [49]. `scikit-learn`[50] and `keras`[51] are used for machine learning and deep learning tasks.

To create the aggregate doc2vec model, we use the pre-processed, cleaned text from all three datasets, with a vector dimension size of 100. Our GitHub page organises the code so as to easily find code for embedding creation, vector creation, vector aggregation, analysis, plots, and validations.

To validate the textual representations, we calculate cosine differences within and between syllabi fields for the Open Syllabus Project corpus. We find the average cosine difference within the same field to be 64°, and for different 82°, and the Kolmogorov-Smirnov 2-sample test between the distributions of differences results in a statistic value of 0.54 with extremely high significance.

3.5 Dataset limitations

It is important to note limitations of our datasets, with the Open Syllabus Project biased towards syllabi posted online, and the Web of Science representing engineering and science disciplines at higher rates than social science and the humanities. The institutionalisation of research and syllabi formats also mean that there is less variance, while job postings are more diverse. Future work should account for bias in representation, either by sampling or weighting different disciplines. While we used only text to create entity vectors, multi-modal and more complex embedding methods can be used to incorporate a variety of data for more nuanced representation.

We also note that while our question was to better understand the knowledge economy and knowledge flows, that picture can never be fully portrayed without accounting for community-oriented knowledge[52], knowledge spread by word of mouth and passed down through families[53], street knowledge, and tacit knowledge propagated without record[54]. All of these constitute valid forms of knowledge unmeasured in our framework, which is restricted to universities and labs that publish research and syllabi available on the web. Even for institutions from which we are able to harvest data, disciplines remain unequally represented. In some disciplines, papers are published at a slower rate, contributing less to the creation of an entity's semantic representation, which might not fully capture the contribution of that discipline to the knowledge economy.

3.5.1 Open Syllabus Project

According to Open Syllabus Project (OSP) metrics, their dataset includes 5%−10% of the Anglophone curricular universe over the last 10 years. Naturally, the dataset focuses on assigned texts: syllabi that have clear assigned texts represent over 50% of the collection, although this varies by institution. Some departments have a majority of all of their syllabi available freely online, and others far less (e.g., elite business schools). Because the syllabi represented cannot provide the complete picture, we must be very tentative when drawing conclusions. We also note that all web scraping was done by OSP from websites that permitted web scraping.

3.5.2 Web of Science

Limitations of the WoS dataset include its imperfect representation of research output, due to factors ranging from publishing norms in individual disciplines (e.g., frequency of publishing, preference of books over papers in history, or conferences proceedings over articles in computer science, etc.) to the database coverage of only the most established and highly cited venues, both of which are generally biased towards greater coverage in the medical and physical sciences, as compared to the social sciences and humanities. For example, the proportion of papers in fields such as chemistry (7.3%), engineering (5.4%), neurosciences (6%), and physics (5%) is much higher than those in fields like sociology (0.66%), anthropology (0.54%), linguistics (0.54%), and archaeology (0.17%). As a result, differences in certain research areas/trends will

¶GitHub link (https://github.com/bhargavvader/knowledge-economy-diversity)

likely be more visible than others.

3.5.3 Burning Glass

Burning Glass (BG) collects information from more over 40 000 job boards and company websites. Despite representing the largest dataset about the US labor market[55], not all job ads appear online. Online recruitment represents a growing share of labor market search, even for jobs historically associated with informal recruitment and offline recruitment, but a 2013 study estimated that only 60%−70% jobs were posted online[56], growing to approach 85%[57]. To verify the representativeness of BG data on the US job market, recent work calculated occupational demand, pay level, and education requirements using BG data and found these values highly correlated with BLS statistics in 2010 and 2018[58], justifying the overall consistency and credibility of BG data during the time period of our analysis, despite coverage limitations[55].

4 Vector Exploration and Validation

Once we complete our aggregation of entity vectors associated with skill discovery (research), dissemination (teaching), and deployment (jobs), our entity vectors become the focus of analysis. To validate the consistency of our vectors, we compute the cosine similarity between a search vector and the full set of entity vectors within that domain. To explore how clusters of entities relate to each other, we perform a tSNE[59] plot on each search vector and its 7 closest vectors in doc2vec space. In Figs. 2–4, we plot the 0th and 1st dimension of the tSNE dimensionality reduction algorithm, to visualize how the entities are related.

We see that research vectors in Fig. 2 for entities with a focus on technology (e.g., MIT, IIT, and GATech) cluster together, those for health-focused entities (e.g., Johns Hopkins and University of Houston Health Center) cluster together, and high, general research-output schools such as Yale, UPenn, and University of Chicago cluster together.

For teaching entities in Fig. 3, community colleges and Texas state schools cluster, while larger research-focused state schools and private universities cluster at a distinct locus.

For jobs, each entity is also mapped to a location and in most cases, singular enterprise entities tend to post similar jobs across their varied locations, which cluster

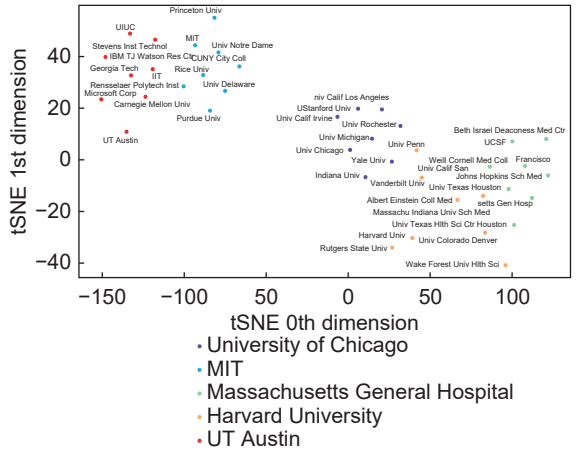

Fig. 2 Clustering of similar research entities.

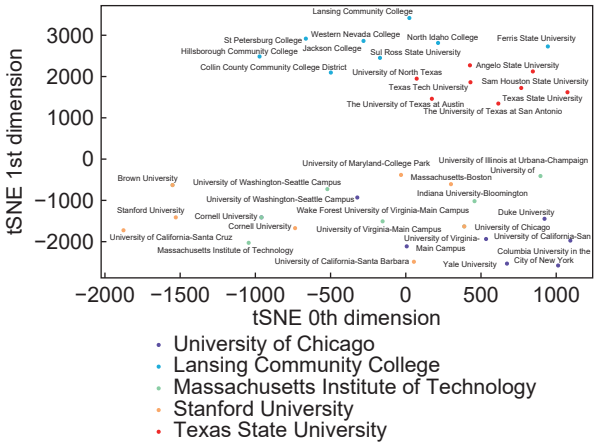

Fig. 3 Clustering of similar teaching entities.

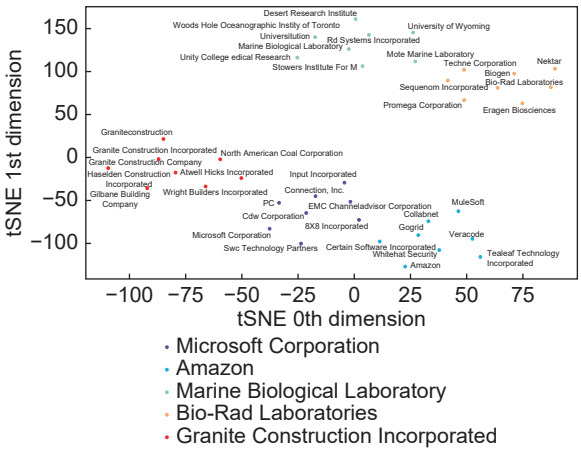

Fig. 4 Clustering of similar industrial entities.

together in semantic space. We also see in Fig. 4 that distinct industries cluster in distinct parts of the space, such as those hosting jobs defined by computational skills (e.g., Microsoft and Amazon), biomedical skills (e.g., Marine Biological Lab and Bio-Rad Labs), and construction skills (e.g., Granite Construction Incorporated).

Visualisations of entity clusters confirm our coarse-grained expectations associated with entity research, teaching, and industrial outputs. We now analyze the diversity of entities in different geographic and domain-based categories to draw conclusions about diversity and patterns of aggregation among these entities across the knowledge economy.

Validating text as a representation of skills

A key concern here may be the validity of using textual content as a proxy for skills, and whether difference in text vector representations may represent difference in skills required. Reference [5] has previously aligned research, syllabi, and job ads textual data on skill mentions, and has demonstrated coherence within areas. We perform an additional validation, where we study similarities within the fields (e.g., chemistry, history, and computer science) marked by the Open Syllabus Project. We sample differences in doc2vec vectors within a syllabus' field, and between syllabi fields, and find that syllabi within a field are substantially and statistically significantly more similar than between fields[‖].

For external validation, we use words within organizational entities (e.g., "Medical", "Technology", and "Community") as natural labels and clustered entities to demonstrate how technological, medical, and community colleges cluster such that distances within each category are substantially and statistically significantly smaller than distances between them.

5 Analyzing Patterns of Diversity Within Research, Teaching, and Industry

Our primary mode of analysis includes the calculation of normalized distances between our aggregate entity vectors. We create our measures by sampling pairs of entities from our pool of research, teaching, and jobs entity vectors. In total, we have 49 911 research entities, 1192 teaching entities, and 340 857 job entities. We constructed nonparametric confidence/credible intervals by sampling 1000 pairs with their corresponding differences. To identify whether two samples of differences come from the same distribution, we use the Kolmogorov-Smirnov test for goodness of fit.

Differences in localisation, diversity, and the relationships between research, teaching, and industry

Table 1 lists statistics for the 1000 cosine distances sampled from pairs of entities belonging to different groups. We calculated means with upper and lower intervals based on sub-sampling[60, 61].

Table 1 Upper and lower intervals of average cosine differences at different regional levels.

Entity differences in geographic aggregation	Upper interval (°)	Lower interval (°)
Research differences country	72.321	70.447
Teaching differences country	47.213	45.361
Job differences country	79.661	78.019
Research differences states	51.933	49.506
Teaching differences states	43.576	41.785
Job differences state	78.697	77.188
Research differences MSA	52.127	49.763
Teaching differences MSA	43.294	41.466
Job differences MSA	78.240	76.259
Research differences technology	72.414	70.236
Teaching differences technology	48.986	47.392
Job differences technology	76.574	75.172
Research differences medical organisation	67.789	65.703
Teaching differences medical school	56.072	54.613
Job differences medical organisation	74.681	73.338
Teaching differences community college	40.810	37.424
Job differences within organisation	52.420	49.739

We see that average cosine distance for pairs of research entities and pairs of industrial entities are nearly 25° higher than for teaching, suggesting that research and job demands are more dissimilar than teaching for a random pair of entities selected at the country level. At the state level, average values drop by 20° for research, but only 4° for teaching, and 1° for jobs. We see the highest average differences between enterprises with respect to jobs, suggesting that they have the highest degree of institutional specialization, yielding collective diversity in semantic content. Job demands are equally diverse at distinct geographic levels of aggregation, suggesting that while firms specialize, geographies generalize with similar diversity at city, state, and national levels. In short, cities approximate the skills diversity present at the level of the entire economy, while research tends to specialize more intensively within region.

These results point towards research and job demands being more diverse than teaching. Research or the invention of knowledge economy skills aggregates at MSA and state levels, with nested, clustered similarities, which both significantly diverge from

[‖]Further information can be found in the GitHub repository.

nation-wide averages. Teaching and jobs—the distribution and deployment of skills that lie behind supply and demand, manifest spatial aggregation much less pronounced than for research. In short, institutions specialize in research and jobs; but only regions specialize in research. This sheds light on the knowledge economy—regional specialization is critical for the intensive invention of new knowledge and skills, but their dissemination across persons and deployment across jobs must be distributed such that each sub-economy (e.g., each city and state) contains the full diversity of complementary skills.

We see similar patterns for technology and medical schools, although technology institutions manifest a larger research-teaching gap. For community colleges, we find that average teaching differences are the lowest among any grouping, reflecting their goal to supply the full-diversity of job-relevant skills to their local populations. We note that within a job organisation, the average distance between job demands is significantly lower. By observing the semantic diversity among the job postings within an organisation, we characterize its function within the broader economy. Examples include Microsoft and Amazon with diversities of 63° and 57°, respectively, and Burger King and Burger Lounge with 4° and 10°, respectively.

We now compare distributions of difference across domains and geographies, allowing us to pose and answer questions about how the semantic diversity is distributed at different levels of aggregation. Table 2 contains the Kolmogorov-Smirnov (KS) statistics and *p*-values for distributions of cosine distances between vectors from different domains across the country. We note almost all pairs of distributions are statistically significant ($p \leqslant 0.05$), with a range of values for the D-statistic, the absolute max distance (supremum) between the CDFs of the two samples. The closer this number is to 0 the more likely it is that the two samples were drawn from the same distribution.

We begin by comparing distributions of differences among research, teaching, and jobs across the country. We see in Table 2 that the D-statistics for research and teaching, and jobs and teaching are high, suggesting that the distribution of differences for teaching is significantly different than it is for research and jobs. We can also see this in the average statistics for teaching, where the average difference between syllabi vectors for two institutions is far less than it is for research or jobs. This suggests that while institutions and regions specialize in their research and jobs, they generalize in their teaching. We run the same KS test for research, teaching, and jobs at different geographic levels. We see the highest D-statistic value (0.582) for comparisons between community colleges and other syllabi because community colleges manifest more similar syllabi vectors, which remain highly distinct from those in medical institutions, consistent with our average statistics.

Table 2 KS test D-statistic value and *p*-value for similarity of distribution of differences at different aggregation levels.

Geographic aggregation	D-statistic	*p*-value
Research country & teaching country	0.826	0.0
Research country & jobs country	0.403	0.0
Job country & teaching country	0.937	0.0

For tests about the same domain at different geographic levels, we see that research angle differences at the country level are quite different from research at state and MSA levels. This is consistent with average statistics reported above suggesting that research at the country level is much more diverse than within a state or MSA, where regional specialization supports the genesis of new-to-the-world knowledge and skills. We also see that for teaching and jobs the distributions of differences are similar at different geographic groupings.

Our main findings are that distributions of research, teaching, and job cosine distances are significantly different, with research being more diverse than teaching, and jobs being more diverse than both. Research and teaching also specialise/localise in different ways, with research aggregating at smaller geographical units (state and MSA). Teaching does not differ much at any aggregate level. For jobs, we find that differences in diversity at the national, state, and MSA level are similar (Table 1). This is largely due to diversity in the content of job advertisements, as it includes a wider range of industries and vocations compared with research and teaching, which are constrained in their semantic regions. We see a small decrease in average differences as we move from national and state to MSA level, and we see that similar industries (i.e., medical) have more similar job advertisements, but the distribution of these angles of difference is similar. The

lowest diversity grouping for job demands remains within the entity across different geographic locations. Even within this category, there is high variation; Burger King has an average semantic diversity of nearly 5°, whereas Microsoft Corporation has a semantic diversity of 63°.

These findings provide suggestive evidence of a misalignment in the diversity of skills across research, teaching, and jobs. The misalignment here is twofold, and can be explained by the structure of knowledge and skills discovery, diffusion, and deployment across the United States. First, we see how jobs, at every geographic level are far more diverse than research and teaching, as a result of the need to combine diverse technologies, skills, and knowledge in driving the economy of a modern city, and all scales above. These findings are consistent with insights from economic geography that demonstrate demand for knowledge and skills from industry exhibits high entropy and diversity at local, regional, and national levels[62]. This reinforces how dense complementarities between skills are required at all levels of the economy. In stark contrast, however, we find low entropy in the invention of new knowledge and skills through research, as specialist researchers cluster within universities and their corresponding cities. While inter- and multi-disciplinarity enrich rare and surprising results[63], the vast majority of research production reflects deep, within-discipline specialization, which is consistent with the economics and sociology of innovation and demonstrates discrimination against interdisciplinarity in funding[64]. Our findings provide striking new evidence for the low entropy and diversity of skills taught at local, regional, and national levels. In short, skills teaching is not customized to the worksite for at least two reasons. First, the site of learning is spatially distant from the worksite. Second, the organization of education is institutionally distant from the worksite and typically staffed by researchers, with deep specialization irrelevant to the vast majority of jobs that their students will enter. These findings suggestively trace the societal cost of tethering education to researchers with narrow skill knowledge rather than to their students with broad skill needs.

We note here that these experiments and results only measure the semantic diversity and similarity between aspects of the knowledge economy, and do not provide causal association. Our results, however, are robust to sampling different pairs of entities within each category, but from concerns that misalignment could be explained by city size, we explore this relationship in detail.

6 Knowledge Diversity Scales with City Size

Here we explore how Metropolitan Statistical Areas (MSAs) scale for different metrics associated with the knowledge economy. We first observe total counts of research papers (Fig. 5), research and teaching entities (e.g., universities, schools, and institutes), and syllabi (Fig. 6) across MSA population size. While larger cities are associated with a larger number of entities, research papers, and syllabi, the presence of high research and teaching output university and college towns means that we do not see a clear superscaling property for MSAs. This is an exception to the super scaling property we would expect for knowledge-based goods and services. For research and teaching, notable outliers include the MSAs of Durham-Chapel Hill (UNC-Chapel Hill, a major state university), New Haven (Yale University), Ithaca (Cornell University), and College Station (Texas AM University), all of which have an abstract or syllabi count in the top 15 but proportionately lower populations. We note here that

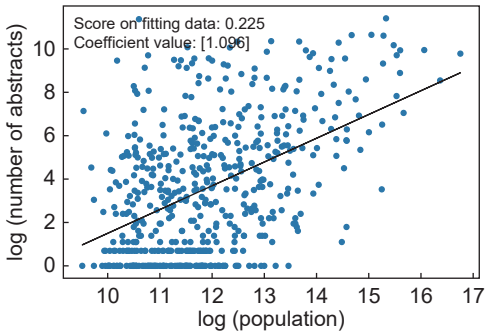

Fig. 5 log-log plot of the number of research abstracts versus the population size for MSAs.

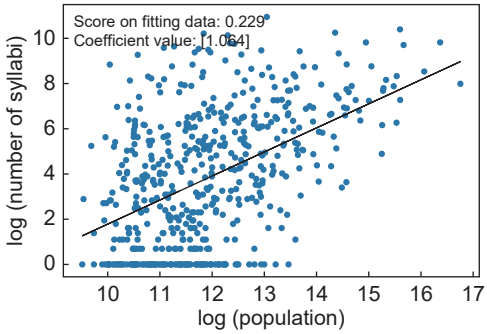

Fig. 6 log-log plot of the number of teaching syllabi versus the population size for MSAs.

these are exceptions because they are historic college towns that have crystallized to focus on research and teaching, but that job demands have not matched their research and teaching output.

We see the best fit for total number of job demands versus MSA population (Fig. 7). Job demands in a city must be diverse and increase superlinearly with population to draw persons into cities from surrounding areas, fueling their dense and interconnected economic activities.

These scaling results further demonstrate the misalignment between the diversity of skills across research, teaching, and industry settings. Jobs skill vectors superscale with city size, but research and teaching do not. This is because all skills are required present in complementarity for the knowledge economy to run. Every business environment requires contract lawyers, accountants, insurance professions, etc. By contrast, not all research and teaching environments require all skills invention or instruction based on the clustered, disciplinary nature of research and the organization of teaching. Moreover, the spatial mismatch between dense research and teaching within college towns, and sometime their under-representation in dense, industrial urban environments breaks this scaling, and highlights the limited potential for feedback from the non-local economy and the teaching that serves it.

Semantic coverage

We measure the semantic coverage of an MSA by measuring its proximity to the center of mass vector of our entire dataset, which we construct by aggregating all normalised US MSA vectors. This measure suggests how much the semantic space an MSA covers. A low cosine distance (or high coverage) suggests that an MSA vector is similar to the center of mass vector that encapsulates all semantic areas captured by the aggregate vector. We characterise the center of mass vector by averaging over all documents in a domain.

$$\boldsymbol{V}_{center\ of\ mass} = \frac{1}{N_2}\sum_{n=0}^{N_2}\boldsymbol{V}_{document_n} \tag{2}$$

where N_2 is the total number of documents within each domain of research, teaching, and jobs. We thus create a center of mass vector for each domain.

We can see in Figs. 8 and 9 that as MSAs increase in population, they grow more similar to the center of mass vector. Because vectors are normalised, a larger MSA does not contribute any more to the center of mass than a smaller MSA. MSAs with high populations have more universities, papers published, and syllabi listed, and this leads to high population MSAs covering more parts of the semantic space, leading to more diverse options for skill discovery and dissemination. However, we see outliers for all three categories that further showcase the misalignment.

For job demands which we see in Fig. 10, we see a similar overall trend with larger cities covering more of the center of mass, although we note outliers. This observation is consistent with the economies of some

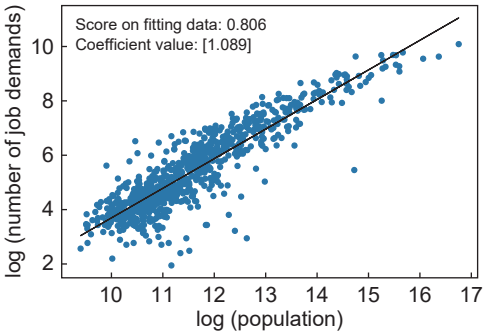

Fig. 7 log-log plot of the number of job demands versus the population size for MSAs.

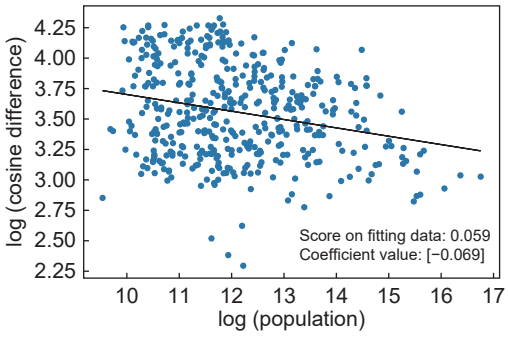

Fig. 8 log-log plot of the cosine distances between the vectors for MSAs and the center of mass for research.

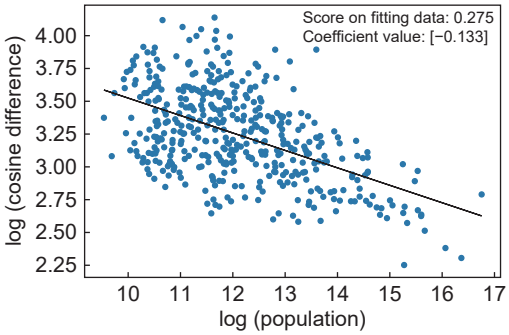

Fig. 9 log-log plot of the cosine distances between the vectors for MSAs and the center of mass for teaching.

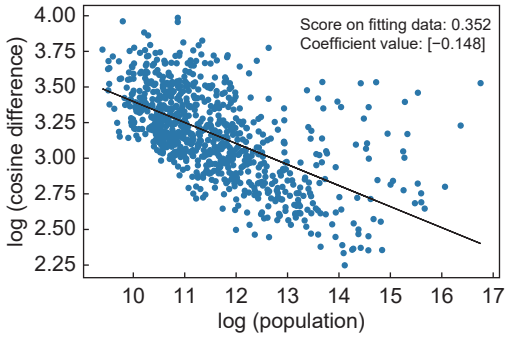

Fig. 10 log-log plot of the cosine distances between the vectors for MSAs and the center of mass for job demands.

large metropolitan areas that specialise in areas that serve the entire country, such as finance in New York City, entertainment in Los Angeles, and technology in the Bay Area, which make them deviate from the center of mass vector.

For research and teaching, we see high semantic coverage for many smaller sized MSAs, a pattern we do not see for job demands. These smaller sized, high coverage MSAs are the college towns and research centers previously mentioned.

7 Discussion

7.1 Finding

In this paper we use research abstracts, teaching syllabi, and job advertisements to create entities that chart the discovery, dissemination, and deployment of skills in research, teaching, and industry, respectively. We validate the semantic consistency of these embeddings and find clusters of similar entities in each domain. We then use two approaches to study distinct aspects of this semantic model of the knowledge economy.

In the first, we use institutional and geographic entities and the cosine differences between them to sample angles of difference within semantic categories. We then compute average statistics for distributions of difference, then calculate the Kolmogorov-Smirnov test to evaluate whether pairs of distributions are similar. We find that job demands are the most semantically diverse as they map a wide range of industrial activity (knowledge and technology based products, manufacturing, and services) widely distributed within all levels of the economy, followed by research, which clusters not only by institutions, but also by region. Both are significantly more diverse than teaching. Research is more similar at smaller geographic units, confirming how the semantic space of novel skill discovery manifests intensive geographic localisation.

In our second approach, we measure semantic coverage for different city (MSA) population sizes. To do this, we create a system-level (United States) Center of Mass vector by averaging and normalising all entities within a domain. We then measure cosine differences between each MSA and the Center of Mass, plotting this versus population size. We observe a weak super-scaling property of MSAs for coverage of teaching and industry, with larger cities being closer to the Center of Mass. The largest cities, however, specialise for job demands with industries such as finance, healthcare, and technology. For research, the presence of smaller university towns with high research output does not allow for super scaling, although we do observe larger cities covering more semantic content. We also find that total job demands manifest super-scaling with population, while teaching and research do not, largely due to the presence of smaller population research and teaching focused MSAs such as Los Alamos (for research), and College Station, TX (for teaching). These research-focused and college/university towns are exceptions to the super-scaling paradigm we may expect to observe for knowledge-based goods and services. This follows from how the creation of knowledge and skills and their embedding within persons can be specialized, and does not need to distribute geographically as knowledge and skills deployment does within the economy. In the economy, every geographical aggregation, from city to country, needs virtually all kinds of knowledge and skills to create the complex complementarities required for modern commerce.

Together, our analysis provides a novel, system-level view of the misalignment between the diversity of knowledge and skills across research, teaching, and the economy. The knowledge economy is a complex, dynamical system, but here we reveal limited feedback between the diversity of skill deployment in the workplace and skill provision through education. Specifically, our findings suggestively trace the massive social and economic costs of tethering education to researchers with narrow knowledge rather than workers with broad skill needs. The lack of diversity in education facilitates the reproduction of courses and the maintenance of programs and curricula, while allowing research specialization. In short,

standardized courses enable specialists in research to appear experts at education by narrowing the provision of education far beyond its potential for relevance to the diversity of student needs as they approach their work in the economy. This validates and extends to the 21st century early warnings by William James[65] and other education scholars regarding liabilities in collocation of teaching with research.

7.2 Future work

While it remains difficult to quantify all aspects of the knowledge economy, with textual content we now have a window into its semantic topology and geometry. Our findings are consistent with and validate existing theory on geographic aggregation of knowledge-based goods and services, while revealing novel scaling relationships. With our GitHub release, all embeddings and entity vectors can be used for replication and independent exploration. Promising avenues for further exploration include the discovery of regional clusters (greater than MSAs and cross-cutting states), the identification of optimal aggregation within research, teaching, and jobs for productivity and prosperity, introducing multi-modal representations of the knowledge economy, the incorporation of technology (e.g., patent), product, and service data in order to trace more dense interlinkages underlying the knowledge economy, and a more fine-grained analysis of temporality associated with waves of discovery and obsolescence in the knowledge economy.

Acknowledgment

We are grateful to Defense Advanced Research Projects Agency (DARPA) (No. HR00111820006) for support and Bledi Taska from Burning Glass for access to digital job advertisement data.

References

[1] W. W. Powell and K. Snellman, The knowledge economy, *Annual Review Sociology*, vol. 30, pp. 199–220, 2004.

[2] D. C. Mowery and N. Rosenberg, The US national innovation system, in *National Innovation Systems: A Comparative Analysis*, R. R. Nelson, ed. New York, NY, USA: Oxford University Press, 1993, pp. 29–75.

[3] B. Godin, The linear model of innovation: The historical construction of an analytical framework, *Science, Technology, & Human Values*, vol. 31, no. 6, pp. 639–667, 2006.

[4] T. J. Pinch and W. E. Bijker, The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other, *Social Studies of Science*, vol. 14, no. 3, pp. 399–441, 1984.

[5] K. Börner, O. Scrivner, M. Gallant, S. Ma, X. Liu, K. Chewning, L. Wu, and J. A. Evans, Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy, *Proceedings of the National Academy of Sciences*, vol. 115, no. 50, pp. 12630–12637, 2018.

[6] A. Marshall, *Industry and Trade*. New York, NY, USA: MacMillan and Co., 1919.

[7] E. Moretti, *The New Geography of Jobs*. New York, NY, USA: Houghton Mifflin Harcourt, 2012.

[8] P. -A. Balland, C. Jara-Figueroa, S. G. Petralia, M. P. A. Steijn, D. L. Rigby, and C. A. Hidalgo, Complex economic activities concentrate in large cities, *Nature Human Behaviour*, vol. 4, no. 3, pp. 248–254, 2020.

[9] R. Florida, *Cities and the Creative Class*. New York, NY, USA: Routledge, 2005.

[10] S. Y. Lee, R. Florida, and G. Gates, Innovation, human capital, and creativity, *International Review of Public Administration*, vol. 14, no. 3, pp. 13–24, 2010.

[11] R. Shearmur, Are cities the font of innovation? A critical review of the literature on cities and innovation, *Cities*, vol. 29, pp. S9–S18, 2012.

[12] G. Duranton and D. Puga, Micro-foundations of urban agglomeration economies, *Handbook of Regional and Urban Economics*, vol. 4, pp. 2063–2117, 2004.

[13] S. Arbesman, J. M. Kleinberg, and S. H. Strogatz, Superlinear scaling for innovation in cities, *Physical Review E*, vol. 79, no. 1, p. 016115, 2009.

[14] L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G. B. West, Growth, innovation, scaling, and the pace of life in cities, *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7301–7306, 2007.

[15] L. M. A. Bettencourt, J. Lobo, and D. Strumsky, Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size, *Research Policy*, vol. 36, no. 1, pp. 107–120, 2007.

[16] L. M. A. Bettencourt, J. Lobo, D. Strumsky, and G. B. West, Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities, *PloS One*, vol. 5, no. 11, p. e13541, 2010.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, in *Proc. 27th Annual Conference on Neural Information Processing Systems 2013*, Lake Tahoe, NV, USA, 2013, pp. 3111–3119.

[18] M. Gentzkow, B. Kelly, and M. Taddy, Text as data, *Journal of Economic Literature*, vol. 57, no. 3, pp. 535–574, 2019.

[19] P. Hoffman, M. A. L. Ralph, and T. T. Roger, Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words, *Behavior Research Methods*, vol. 45, no. 3, pp. 718–730, 2013.

[20] S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[21] Q. Le and T. Mikolov, Distributed representations of sentences and documents, in *Proc. 31st International*

Conference on Machine Learning, Beijing, China, 2014, pp. 1188–1196.
[22] J. A. Evans and P. Aceves, Machine translation: Mining text for social theory, *Annual Review of Sociology*, vol. 42, pp. 21–50, 2016.
[23] C. Kemp and J. B. Tenenbaum, The discovery of structural form, *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10687–10692, 2008.
[24] D. M. Blei and J. D. Lafferty, Dynamic topic models, in *Proc. of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, 2006, pp. 113–120.
[25] W. L. Hamilton, J. Leskovec, and D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, in *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 1489–1501.
[26] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo, Individuals, institutions, and innovation in the debates of the French Revolution, *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. 4607–4612, 2018.
[27] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, No country for old members: User lifecycle and linguistic change in online communities, in *Proc. 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 307–318.
[28] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S. -C. Zhang, Learning atoms for materials discovery, *Proceedings of the National Academy of Sciences*, vol. 115, no. 28, pp. E6411–E6417, 2018.
[29] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature* , vol. 571, no. 7763, pp. 95–98, 2019.
[30] A. C. Kozlowski, M. Taddy, and J. A. Evans, The geometry of culture: Analyzing the meanings of class through word embeddings, *American Sociological Review*, vol. 84, no. 5, pp. 905–949, 2019.
[31] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3635–E3644, 2018.
[32] A. Caliskan, J. J. Bryson, and A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* , vol. 356, no. 6334, pp. 183–186, 2017.
[33] G. Grand, I. A. Blank, F. Pereira, and E. Fedorenko, Semantic projection recovers rich human knowledge of multiple object features from word embeddings, *Nature Human Behaviour*, vol. 6, no. 7, pp. 975–987, 2022.
[34] A. R. Andrés, A. Otero, and V. H. Amavilah, Using deep learning neural networks to predict the knowledge economy index for developing and emerging economies, *Expert Systems with Applications*, vol. 184, p. 115514, 2021.
[35] F. Iandolo, F. Loia, I. Fulco, C. Nespoli, and F. Caputo, Combining big data and artificial intelligence for managing collective knowledge in unpredictable environment—Insights from the Chinese case in facing COVID-19, *Journal of the Knowledge Economy*, vol. 12, no. 4, pp. 1982–1996, 2021.
[36] K. Lix, A. Goldberg, S. B. Srivastava, and M. A. Valentine, Aligning differences: Discursive diversity and team performance, *Management Science*, doi: 10.31235/osf.io/8pjga.
[37] R. Yu, S. Das, S. Gurajada, K. Varshney, H. Raghavan, and C. Lastra-Anadon, A research framework for understanding education-occupation alignment with NLP techniques, in *Proc. 1st Workshop on NLP for Positive Impact*, Online, 2021, pp. 100–106.
[38] B. Biasi, D. J. Deming, and P. Moser, Education and innovation, Technical report, National Bureau of Economic Research, Cambridge, MA, USA, 2021.
[39] Thomson Reuters, Web of Science, https://www.webofknowledge.com/, 2012.
[40] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv: 1802.05365, 2018.
[41] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2018.
[42] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 8342–8360.
[43] S. V. D. Walt, S. C. Colbert, and G. Varoquaux, The NumPy array: A structure for efficient numerical computation, *Computing in Science and Engineering*, vol. 13, no. 2, pp. 22–30, 2011.
[44] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., SciPy 1.0: Fundamental algorithms for scientific computing in Python, https://scipy.org/, 2020.
[45] J. D. Hunter, Matplotlib: A 2D graphics environment, *Computing in Science and Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
[46] F. Perez and B. E. Granger, IPython: A system for interactive scientific computing, *Computing in Science and Engineering*, vol. 9, no. 3, pp. 21–29, 2007.
[47] R. Řehřek and P. Sojka, Gensim—Statistical semantics in python, https://radimrehurek.com/gensim/, 2011.
[48] M. Honnibal, SpaCy: Industrial-strength natural language processing (NLP) with Python and Cython, https://spacy.io/, 2015.
[49] B. S. Desikan, *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras*. Birmingham, UK: Packt Publishing Ltd, 2018
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[51] Keras, https://keras.io/, 2015.
[52] M. Polanyi, *Personal Knowledge*. New York, NY, USA:

Routledge, 2012.
[53] P. Bourdieu, *Habitus and Field: General Sociology*. Hoboken, NJ, USA: Wiley, 1983.
[54] H. Collins, *Tacit and Explicit Knowledge*. Chicago, IL, USA: University of Chicago Press, 2010.
[55] B. Hershbein and L. B. Kahn, Do recessions accelerate routine-biased technological change? Evidence from vacancy postings, *American Economic Review*, vol. 108, no. 7, pp. 1737–1772, 2018.
[56] A. P. Carnevale, T. Jayasundera, and D. Repnikov, Understanding online job ads data, Tech. Rep., Georgetown University, Center on Education and the Workforce, Washington, DC, USA, 2014.
[57] V. Lancaster, D. Mahoney-Nair, and N. J. Ratcliff, Technology report review of burning glass job-ad data, Technical report, Biocomplexity Institute and Initiative Social and Decision Analytics Division, University of Virginia, Charlottesville, VA, USA, 2019.
[58] D. Tong, L. Wu, and J. A. Evans, Low-skilled occupations face the highest re-skilling pressure, arXiv preprint arXiv: 2101.11505, 2021.
[59] L. V. D. Maaten and G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.
[60] D. N. Politis and J. P. Romano, A circular block-resampling procedure for stationary data, Technical report, Department of Statistics, Purdue University, West Lafayette, IN, USA, 1991.
[61] D. N. Politis and J. P. Romano, The stationary bootstrap, *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1303–1313, 1994.
[62] C. A. Hidalgo, Economic complexity theory and applications, *Nature Reviews Physics*, vol. 3, no. 2, pp. 92–113, 2021.
[63] F. Shi and J. Evans, Science and technology advance through surprise, arXiv preprint arXiv: 1910.09370, 2020.
[64] L. Bromham, R. Dinnage, and X. Hua, Interdisciplinary research has consistently lower funding success, *Nature*, vol. 534, no. 7609, pp. 684–687, 2016.
[65] W. James, *Talks to Teachers on Psychology and to Students on Some of Life's Ideals*. Cambridge, MA, USA: Harvard University Press, 1983.

James Evans received the BA degree in anthropology from Brigham Young University in 1994 and the MA and PhD degrees from Stanford University in 1999 and 2004, respectively. He is currently the Max Palevsky Professor of sociology at University of Chicago, an external faculty at the Santa Fe Institute, and the director of the Knowledge Lab and a program in Computational Social Science. His research focuses on collective systems of thinking and knowing, and employs deep learning and generative modeling for augmented intelligence. His work has been published in venues such as *Nature*, *Science*, *Proceedings of the National Academy of Sciences*, and top computer and social science outlets.

Bhargav Srinivasa Desikan received the BEng degree from BITS Pilani in 2016 and the MA degree from University of Chicago in 2020. He is currently pursuing the PhD degree in the Department of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL). His research interests include computational approaches to study knowledge, language, and culture. He has published papers in leading journals and conferences such as *ACL*, *JMLR*, and *Cognition*, and has written an introductory book on natural language processing.