Study Information Block

Discrepancies between Scientific Discovery, Education, and Market Demands

You are invited to participate in a research study on the effectiveness of visualizations that compare skills extracted from labor market, education, and scientific research data. We ask that you read this form and ask any questions you may have before agreeing to be in the study. The study is being conducted by Dr. Katy Börner (katy@indiana.edu) and Dr. Olga Scrivner (obscrivn@indiana.edu) from the School of Informatics, Computing, and Engineering at Indiana University.

STUDY PURPOSE

The purpose of this study is to evaluate the effectiveness of different data visualizations and to understand what answers the visualizations provide and what (new) questions they inspire.

PROCEDURES FOR THE STUDY

If you agree to be in the study, you will do the following things:
You will complete a prequestionnaire that will gather basic demographic information as well as some information about your expertise with data visualizations. After this, you will see a set of data visualizations and answer a few questions about each of them. The study will take approximately 30 minutes of your time.

RISKS AND BENEFITS

The risks of participating in this research are discomfort answering questions about unfamiliar visualizations. Please be aware that you can terminate your participation in the study at any time.

CONFIDENTIALITY

We do not collect information regarding your personal identity. Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the study investigator and his/her research associates, the Indiana University Institutional Review Board or its designees, and (as allowed by law) state or federal agencies, specifically the Office for Human Research Protections (OHRP), who may need to access your research records.

CONTACTS FOR QUESTIONS OR PROBLEMS

For questions about the study, please contact researcher Olga Scrivner at obscrivn@indiana.edu. For questions about your rights as a research participant or to discuss problems, complaints or concerns about a research study, or to obtain information, or offer input, contact the IU Human Subjects Office at 8128564242.

VOLUNTARY NATURE OF STUDY

Taking part in this study is voluntary. You may choose not to take part or may leave the study at any time. Leaving the study will not result in any penalty or loss of benefits to which you are entitled. Your decision whether or not to participate in this study will not affect your current or future relations with the School of Informatics, Computing, and Engineering at IU.

Pre-Q

Please answer the following questions below to help us understand your background and experience.

Please indicate your age:

- ○ <20
- ○ 21-30
- ○ 31-40
- ○ 41-50
- ○ 51-60
- ○ >60

Please indicate your gender:

- ○ Male
- ○ Female
- ○ Other
- ○ I prefer not to answer

What is your native language?

- [ ] English
- [ ] Other(s), please specify:

Where do you work/study?

- ○ Academia
- ○ Company
- ○ Government
- ○ Not for profit
- ○ Other, please specify:

What is your general job title?

What department, specialization, or industry do you work in? For example, do you work in engineering, the social sciences, manufacturing, etc.

How might you use data on job market trends, educational programs, and science and technology developments? Check all that apply.

- [ ] Understanding what course(s) I should take next.
- [ ] Planning my career or choosing my next job.
- [ ] Deciding on needed course revisions.
- [ ] Developing a new curriculum or degree program.
- [ ] Understanding delays between job market needs and educational offerings.
- [ ] Providing policy advice to decision makers.
- [ ] Hiring the best employees with the right skills
- [ ] Reviewing what jobs my competitors are trying to fill.

- [ ] Writing future job advertisements.
- [ ] Other, please specify: _____

How comfortable are you reading graphs and maps?

	Very much	Somewhat	Undecided	Not really	Not at all
I am...	○	○	○	○	○

Have you ever taken a visualization course, training, or seminar?

- ○ Yes
- ○ No

Please specify which course, training, or seminar you took.

_____

Group 1

The following section includes a number of figures (disregard figure numbers). Please examine the figures and then answer the questions posed. Note that some questions go beyond what can be seen in the current figures to help us understand how people interpret visualizations.

Figure 1

Figure 1 shows a conceptual drawing of the interplay of job market demands, course offerings, and science and technology progress as written up in scientific publications. Each of the three levels uses the same topical landscape. Mountains (+) indicate surplus skills and valleys (-) indicate deficiencies. For example, the Biotechnology mountain is in the same position at all three levels. Many more Biotechnology skills are listed in job advertisements than in courses or publications.

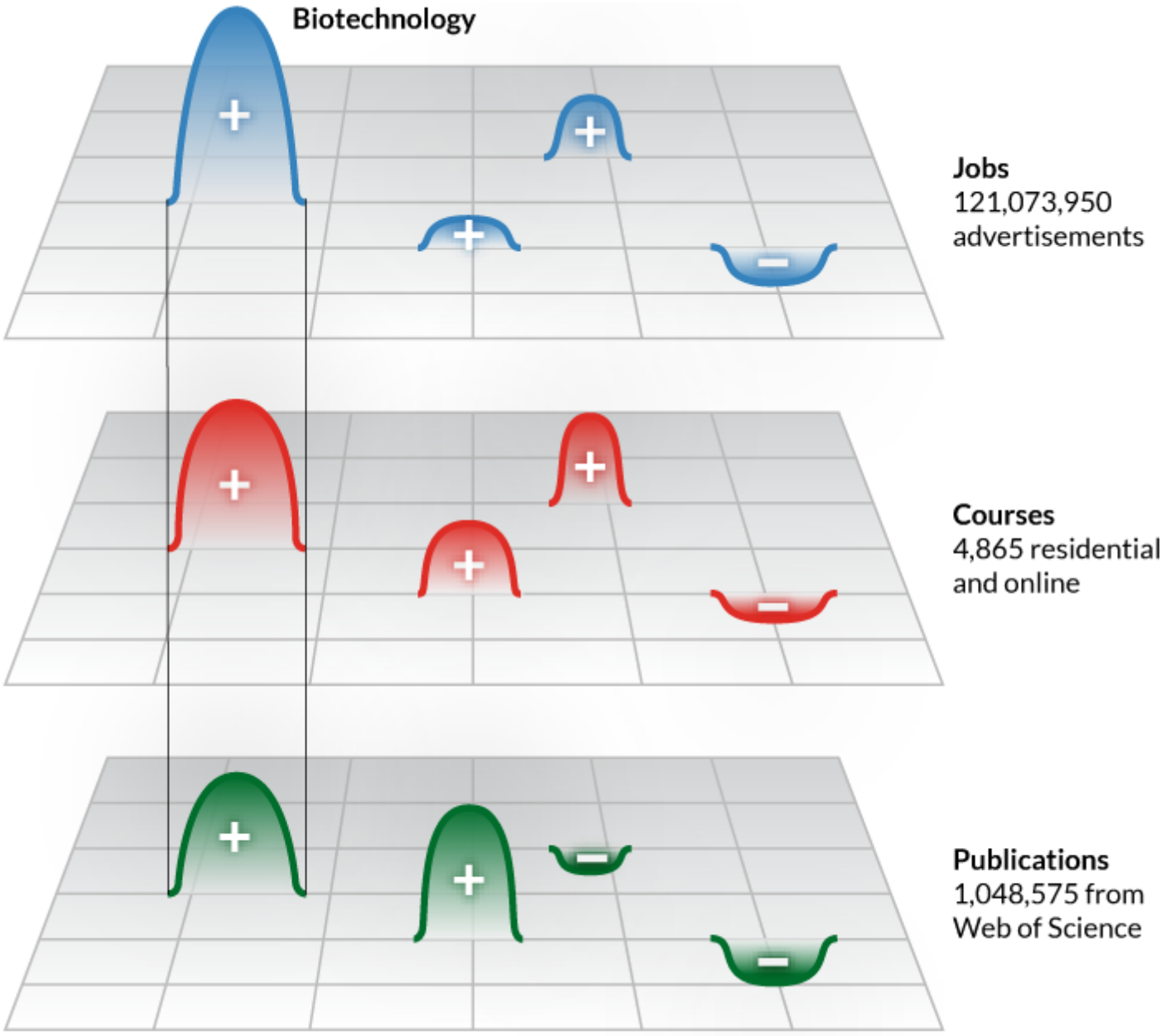


Figure 1: Conceptual drawing of the interplay of job market demands, course offerings, and publications.

Why might Biotechnology skills be more frequently listed in job advertisements than in courses or publications?

What skills or areas might be more frequently listed in publications than in course descriptions and in job advertisement?

What skills or areas might be more frequently listed in courses than in publications and in job advertisements?

Figure 3

Figure 3 below shows a topical space of skills that appear in job advertisements. In this topical space, jobs that contain similar skills are closer together. Each tiny gray dot is a skill that is connected via a gray line to a skill cluster; skill clusters are connected to skill families. The resulting tree layout shows the topical space of a widely-used skills taxonomy. Labels are given for all skills family nodes (red circles) and for the largest skill cluster, which is shown as a red triangle labeled NA. Topical areas are defined by boundary lines.

Hard (programming, debugging) and soft skills (communication, management) are identified with purple and orange circles respectively. For hard and soft skills, the size of the circle indicates the number of times the skills are listed in Data Science and Data Engineering job announcements.

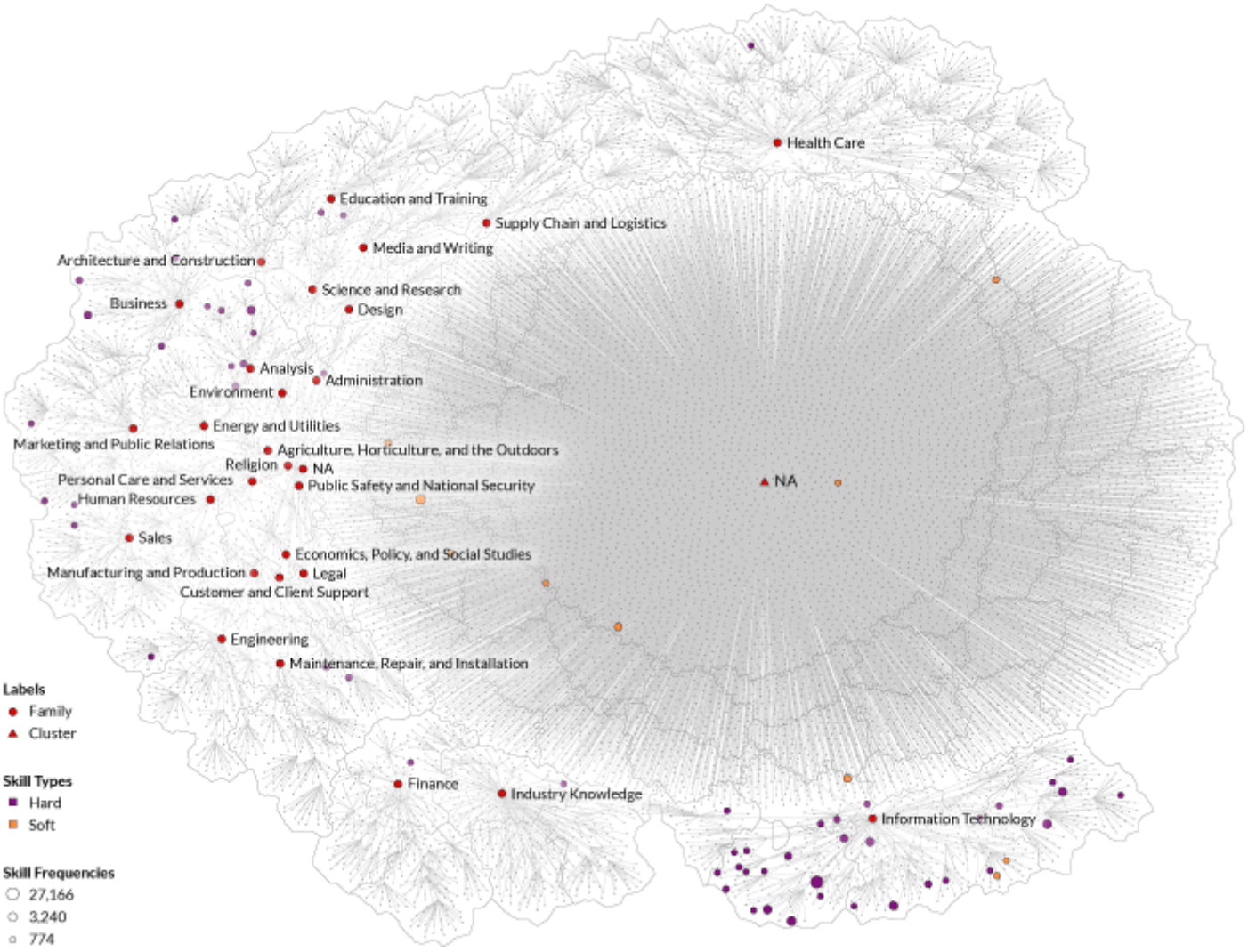


Figure 3. Topical map of 13,218 skills.

Find the Information Technology skills area in the lower right corner. Are there more hard skills or soft skills in that area? Check one:

- ○ Hard
- ○ Soft

What skills area has the second largest number of hard skills?

What is the name of the largest skill family?

What is the name of the largest skill cluster that contains the most soft skills?

Figure 4

The map shown in Figure 3 can be used to map the number of Data Science and Data Engineering skills in job advertisements (panel A), course descriptions (B), and publications (C). Each dot is a skill; dots are area size coded by the number of times a skill occurs. Label sizes indicate skill frequency.

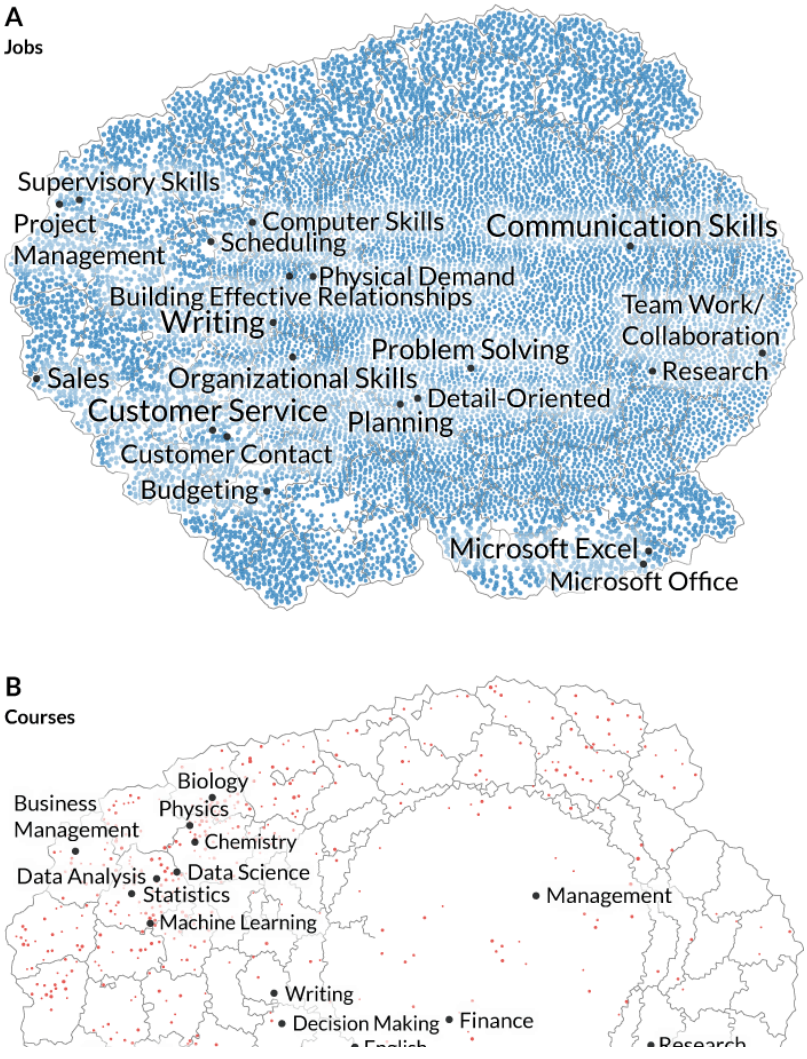


Figure 4. Map of skill terms found in jobs, publications, and courses.

List 3 skills that are labeled and that only occur in Jobs:

List 3 skills that only occur in Courses:

List 3 skills that only occur in Publications:

What does it mean if a skill only occurs in one level/dataset?

Using skills area labels from the map shown in Figure 3, in what area are Microsoft Excel/Office skills?

Figure 5

Burst analysis can be run to determine sudden increases in the frequency of skills. In the below figure, each burst is rendered as a vertical bar with a start and an end date; the burst term is rendered on the left. Skills that burst in jobs are given in blue; skills bursting in publications are shown in green. Seven skills burst in both datasets during the same years and are shown in gray: Apache Hadoop, Electrical Engineering, Energy Engineering, Marketing Analytics, Maximo, Social Gaming.

Figure 5. Burst of skills in jobs and publications. HRMS stands for human resources management system and Maximo is an IBM system for managing physical assets.

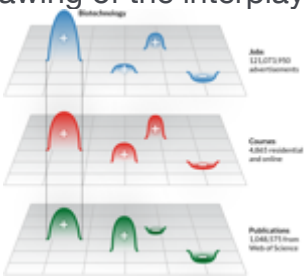
What might co-bursting of terms indicate?

What might be the reason that there are so few bursts in 2014 and 2015?

Name skills that burst 3 times but never in both datasets together.

Based on your answer to the previous question - What might this indicate?

What visualization was most interesting/useful for your decision making?

- ○ Conceptual drawing of the interplay of job market demands, course offerings, and publications. 
- ○ Topical map of 13,218 skills. 
- ○ Map of skill terms found in jobs, publications, and courses. 
- ○ Burst of skills in jobs and publications. 

Why did you find this visualization interesting/useful?

Thank you very much for your input!

Group 2

The following section includes a number of figures (disregard figure numbers). Please examine the figures and then answer the questions posed. Note that some questions go beyond what can be seen in the current figures to help us understand how people interpret visualizations.

Figure 3

Figure 3 below shows a topical space of skills that appear in job advertisements. In this topical space, jobs that contain similar skills are closer together. Each tiny gray dot is a skill that is connected via a gray line to a skill cluster; skill clusters are connected to skill families. The resulting tree layout shows the topical space of a widely-used skills taxonomy. Labels are given for all skills family nodes (red circles) and for the largest skill cluster, which is shown as a red triangle labeled NA. Topical areas are defined by boundary lines.

Hard (programming, debugging) and soft skills (communication, management) are identified with purple and orange circles respectively. For hard and soft skills, the size of the circle indicates the number of times the skills are listed in Data Science and Data Engineering job announcements.

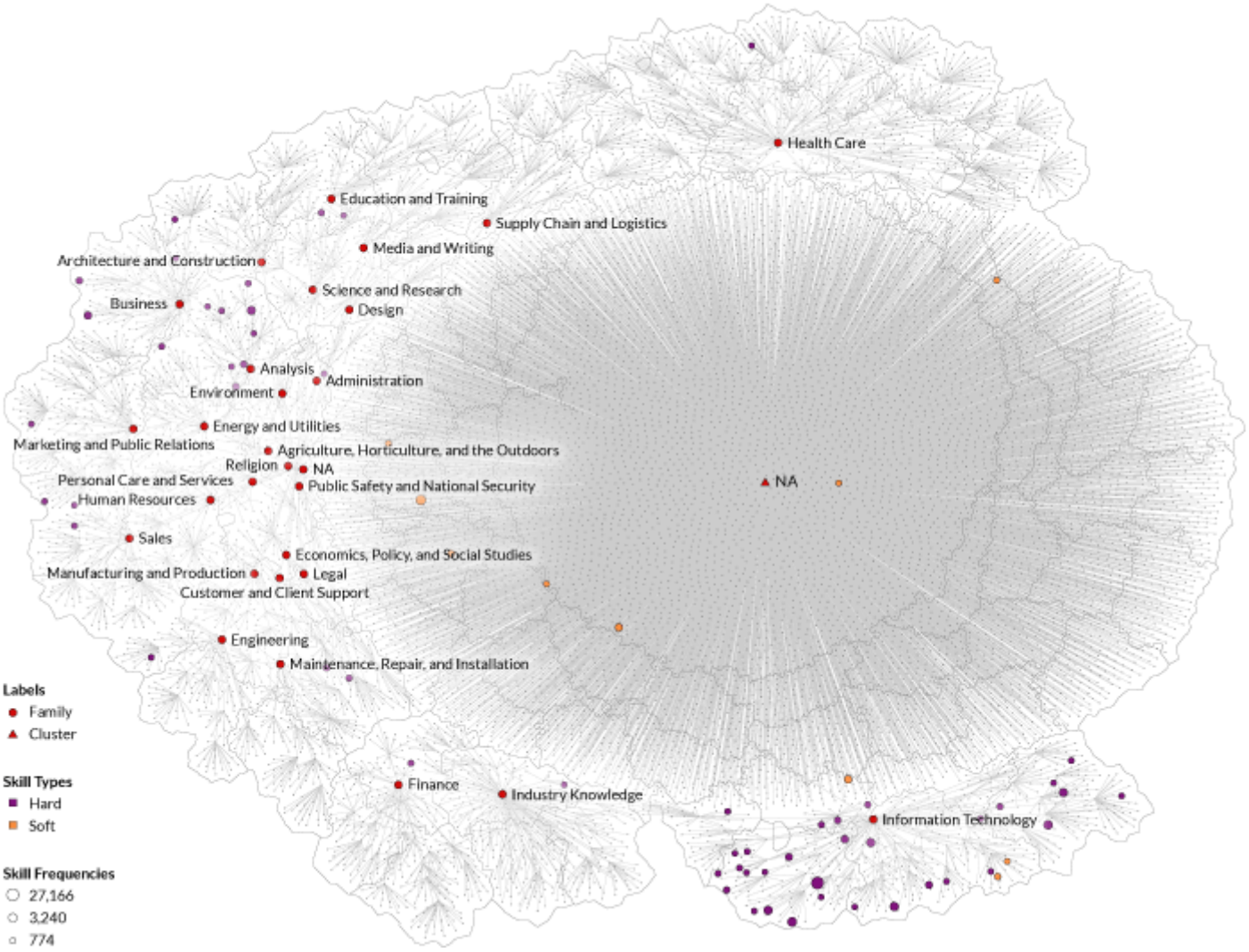


Figure 3. Topical map of 13,218 skills.

Find the Information Technology skills area in the lower right corner. Are there more hard skills or soft skills in that area? Check one:

- ○ Hard
- ○ Soft

What skills area has the second largest number of hard skills?

What is the name of the largest skill family?

What is the name of the largest skill cluster that contains the most soft skills?

Figure 6

The visualization shown in Figure 6 compares skills distributions in jobs, courses, and publications. Kullback-Leibler (KL) Divergence, a measure of relative entropy, was run to calculate the information gained or lost when confronted with a new probability distribution given an existing one.

In panel A on the left, the KL probability values are plotted as a matrix, showing the divergence of skill distributions found in publications, courses, and jobs. When both distributions are equal, there is no information gain or surprise—see zero values in the diagonal. The larger the difference between two distributions, the more divergence or surprise exists.

Publications and jobs are segmented into two timeframes, namely 2010-2013 and 2014-2016. Note that the matrix is asymmetric, i.e., the divergence values from publications in 2010-2013 to courses (value 5.1 in top row) is different from the divergence from courses to those publications (value 2.9 in left column). The highest surprise/divergence exists for publications to jobs 2010-2013 with the value of 6.1, followed by a decrease in 2014-2016 to value 4.8. Note that there is little information gain when comparing for different years of publications and jobs—the values are between 0.01 and 0.4.

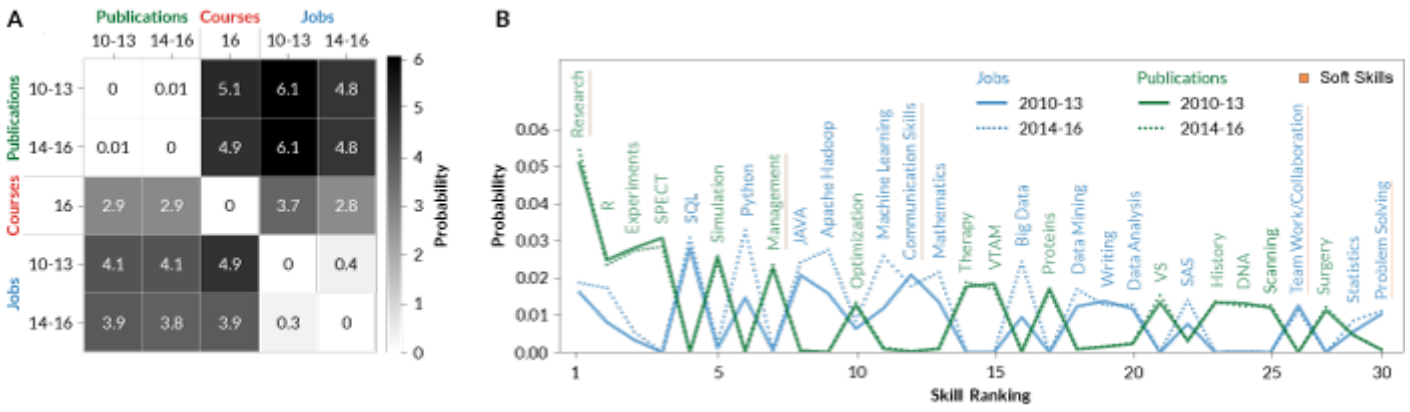


Figure 6. Differences between skill distributions in academic abstracts 2010-13 and 2014-16, course descriptions, and job advertisements in 2010-13 and 2014-16 (A) and skills ranking (B). R is a scripting language, VTAM refers to the IBM Virtual Telecommunication Access Method application, VS is the integrated development environment Visual Studio, and SAS is a data analytics software.

Panel A: Do skill distributions for publications and jobs become more similar over time?

- ○ Yes
- ○ No

Panel A: In what years are skills in publications and jobs most surprising?

- ○ 10/13
- ○ 14/16

Panel B: Name the 2 most popular skills:

Panel B: Name the 2 most popular skills in jobs:

Figure 7

The below figure shows the skills topical landscape for Jobs (blue, on top) and Publications (green, below). Skills that have high burst values and that have a significant Granger causality (p-value < 0.05) are labelled. These skills can be used to predict future values of the same skills in another dataset. For example, Immunology listed in Publication (green) is highly influencial on the usage of Immunology listed in Jobs (blue)

in future years. Arrows coded by thickness indicate the F-value strength from jobs -> publications (blue) and from publications -> jobs (green). Three skills (Computational Tools, Product Knowledge, and Trial Design) have a significant Granger causality in both directions.

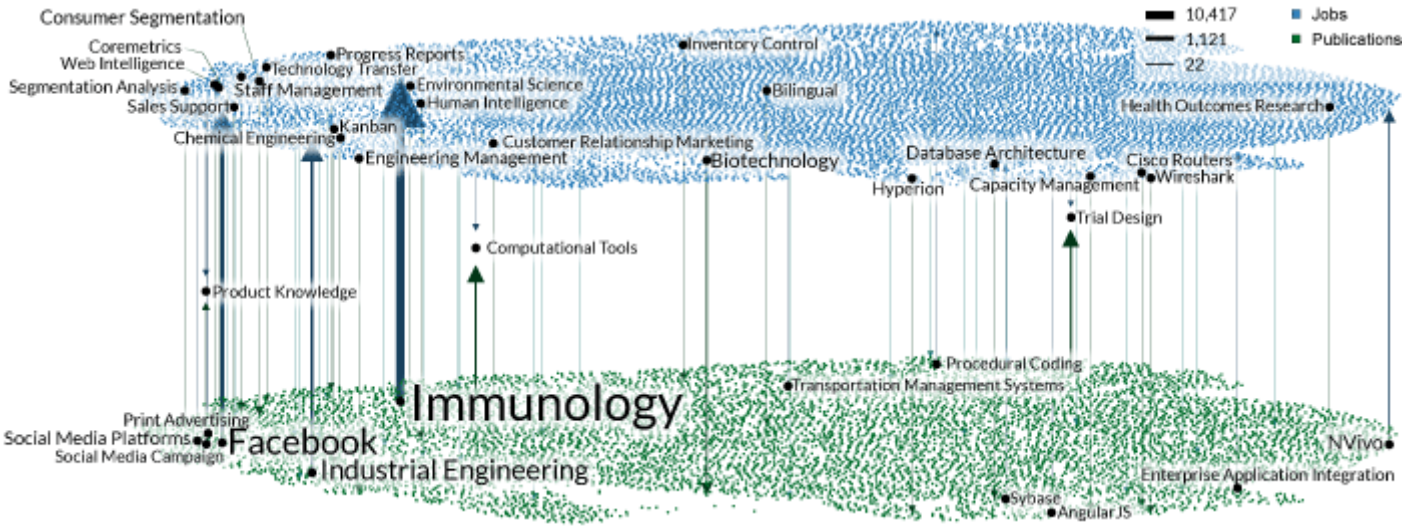


Figure 7: Strength of influence between jobs (blue) and publications (green).

What publication skill has the most influence on jobs?

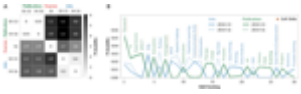
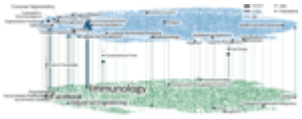
Based on your previous answer, what does this mean?

The skill Computational Tools has a significant Granger causality in both directions. What does this mean?

The job skill Biotechnology has high influence. What does this mean?

What else did you find interesting?

What visualization was most interesting/useful for your decision making?

- Topical map of 13,218 skills. 
- Differences between skill distributions between publications and jobs 
- Strength of influence between jobs and publications. 

Why did you find this visualization interesting/useful?

Thank you very much for your input!