Trends in Data Science Job Postings on Stack Overflow

Benjamin Ackerman

October 9, 2017

Introduction

- Increased demand for data scientists, hot new field, Stack Overflow's popularity

Research Aim

The purpose of this paper is to examine trends in job postings for "data scientists" on the Stack Overflow job board. This involves determining the most common computing skills that employers look for, along with their preferences of degree types and areas of study. This paper will also locate geographic regions where data science jobs are in highest demand, and if there are substantial differences in job characteristics by location. Finally, trends of these characteristics of job listings will be explored over time.

Methods

Data Collection

Data were made privately available upon request from David Robinson, a Data Scientist at Stack Overflow. The provided data consist of information from jobs posted on the Stack Overflow job board that either have "data scientist" or "data analyst" in their title between August 25, 2010 and September 25, 2017. While company names were censored from the data, the following attributes of each posting were provided in a data frame: job title, original posting date (YEAR-MM-DD), associated tags indicating relevant skills, job location (City, State, Country), salary (when included), and the full text of job descriptions and requirements.

While a fair amount of variables were already provided in a dataframe, additional information was extracted from the data and cleaned. The `geocode` function in the `ggmap` package was used to gather latitude and longitude coordinates for each job location. Preferences of academic backgrounds were extracted from the job requirements section. This included any mentions of type of degree (Bachelors, Masters, PhD) along with mentions of favorable majors and departments. To detect relevant majors, a dictionary was compiled using a comprehensive list of STEM fields provided by **Stemdegreelist.com**. Additionally, for jobs that mentioned multiple degrees (i.e, "Bachelor's degree required, Master's degree preferred"), the "highest degree preferred" for a job listing was determined. For listings that did not provide job requirement sections, the job descriptions section was used to check for these attributes.

Exploratory Data Analysis

Exploratory data analysis was conducted to summarize the most commonly listed attributes in the job postings. Skill tags, areas of study, and job locations were tabulated across all postings and ranked to determine the most common skills sought by employers, and where the most employment opportunities were geographically located. Hex maps were generated to view the distribution of the number of jobs posted by geographic location. To visualize the changes in the top ten tags, areas of study, and job locations over the last five years, code to generate a change-in-ranking plot was modified from a function described on **this Stack Overflow forum**. Number of job postings were also tabulated by month and year to determine if there were any changes in frequency of postings over time.

Statistical Analysis

In order to assess any differences in jobs by location, proportions of jobs that offer visa sponsorship, allow remote work, and assist with relocation were compared between jobs listed in the US and Europe, the two geographic regions with the highest numbers of job listings, using two-sample t-tests. The distributions of highest degree preferred were compared across regions with a Pearson's Chi-squared test.

Results

The top three computing skills listed as tags on job listings are `Python`, `R` and `SQL` (Figure 1a). Of the 995 jobs listed, 448 jobs (45%) use the `Python` tag, 281 jobs (28.2%) use the `R` tag and 249 jobs (25%) use the `SQL` tag. Employers' focus on coding abilities are also highlighted in the top three preferred areas of study for data science job candidates (Figure 1b): Computer Science (n = 469, 47.1%), Statistics (n = 436, 43.8%) and Engineering (n = 301, 30.3%). Figures 1c and 3 suggest that most data science jobs are located in big cities, led by tech hubs New York, NY and San Francisco, CA. Several European cities (London, Berlin, and Amsterdam), are represented in the top ten cities for data science jobs as well. Also, as the number of data science job listings has increased over time, so too has the proportion of jobs that are located in Europe (Figure 2).

Differences between jobs listed in the United States and jobs listed in Europe are described in Table 1. European employers are more likely to offer visa sponsorship (EU: 20.8%, USA: 5.9%, $p < .01$) and to offer assistance with relocation (EU: 35.5%, USA: 26.6%, $p < .01$) than US employers. US employers are more likely to allow employees to work remotely (USA: 9.2%, EU: 2.8%, $p < .01$) and offer more jobs for candidates with Bachelor's Degrees only ($p < .01$) than European employers.

- While Python has consistently been most tagged skill, R and machine learning have become increasingly more important to data science jobs in the last three years (as they move up the rankings)
- No major changes in top areas of study over time
- Most US jobs are in big cities like New York, San Francisco, Boston, and Chicago. While most jobs overall are located in the US, in the last two years, European cities have been becoming increasingly more popular. Biggest changes in top cities has happened between 2016 and 2017

as seen in the length of the arrows - more European cities are rising to the top while US cities are falling.

Figure 1: Most Popular Attributes of Job Listings

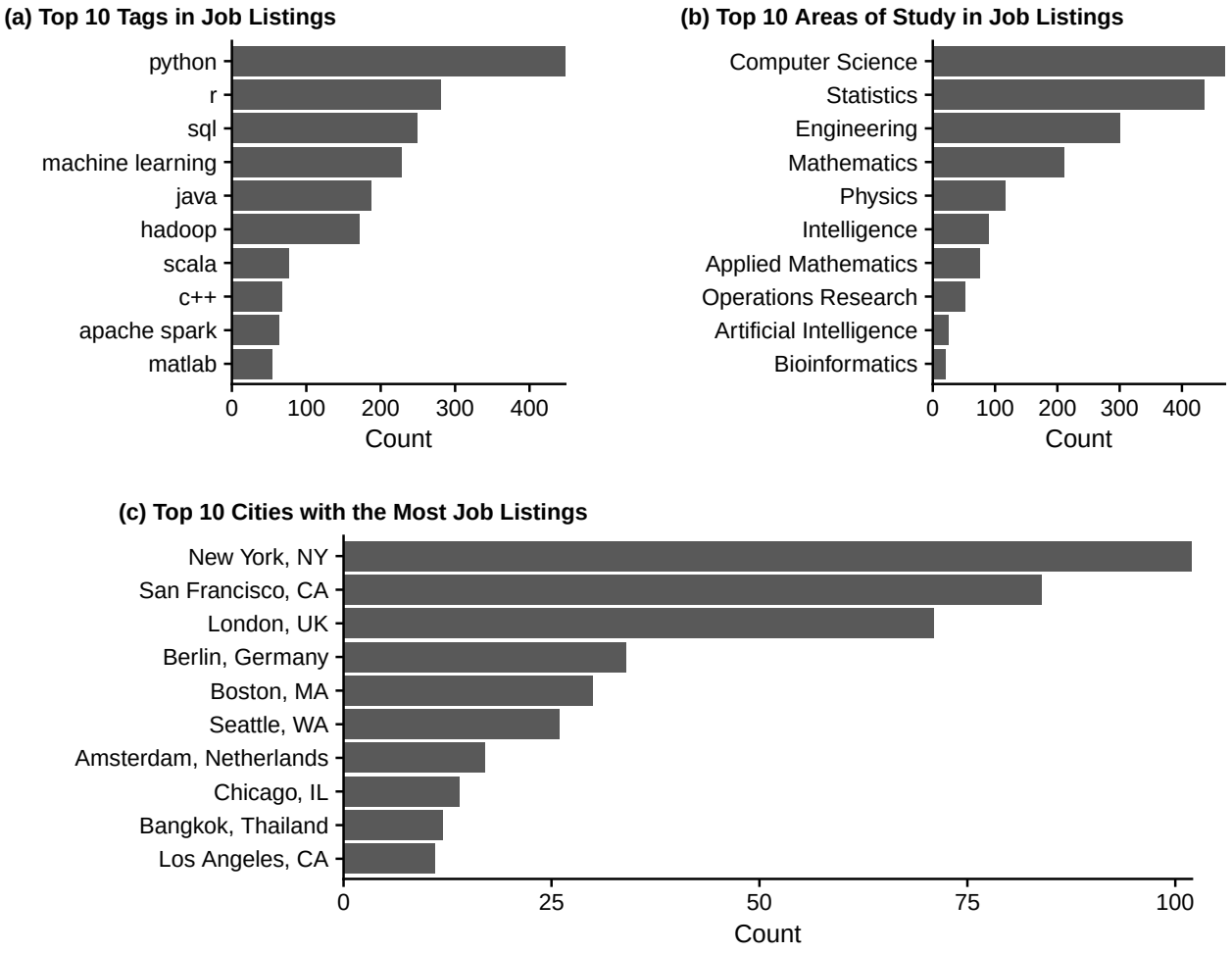


Figure 2: Number of Job Listings by Year and Geographic Region

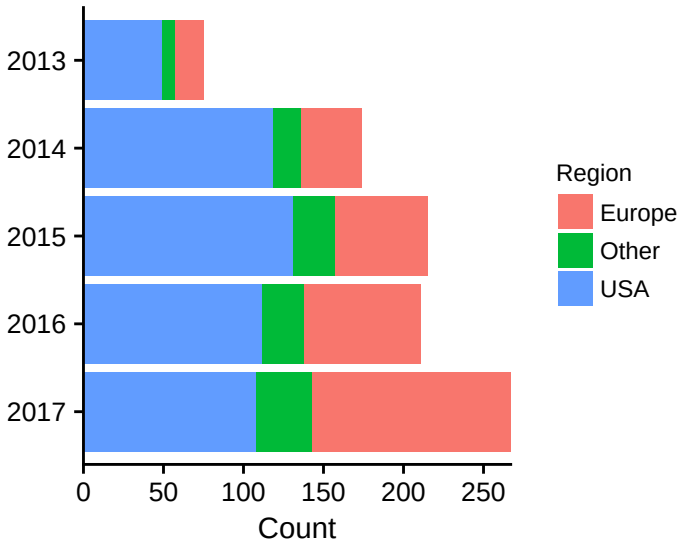


Figure 3: Geographic Distribution of Jobs in the USA vs. Europe

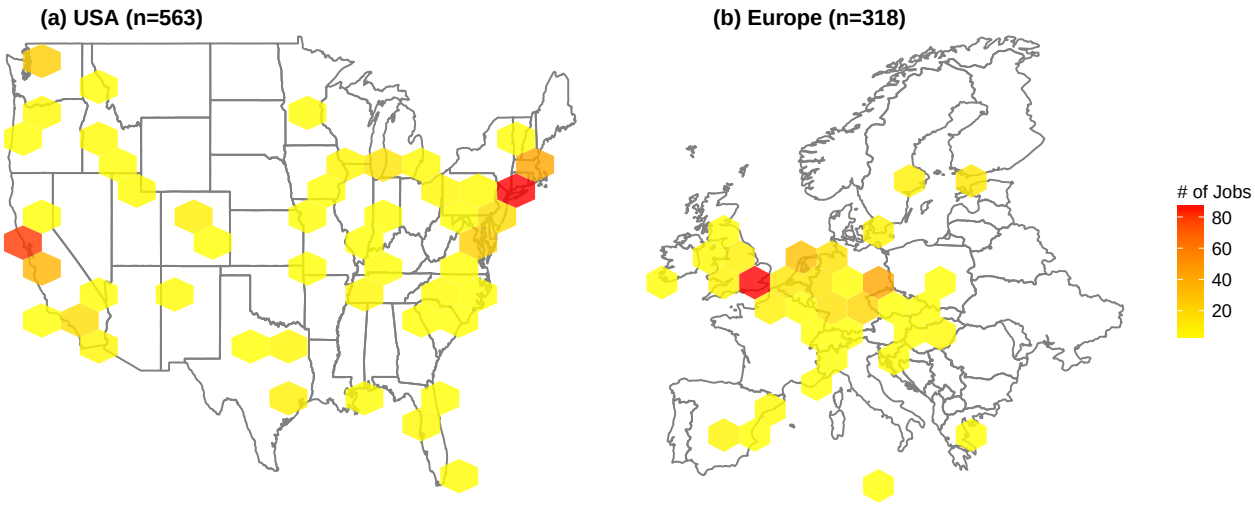


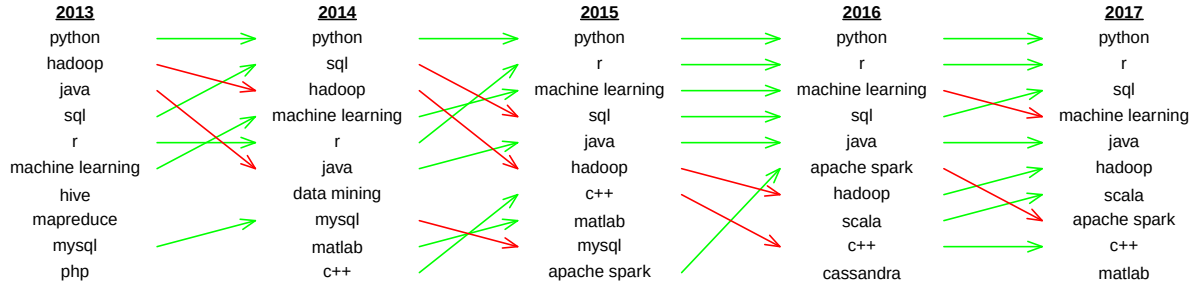
Figure 4: Changes in Job Listing Attributes over the Last Five Years

Table 1: Differences between Job Listings in the USA vs. Europe

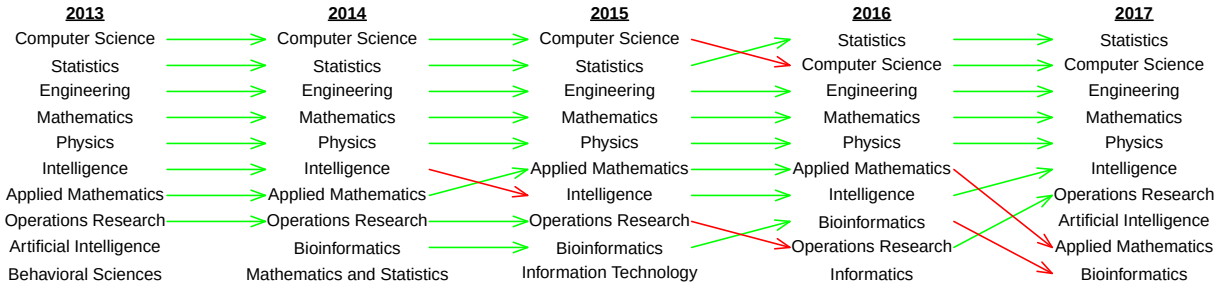
	USA	Europe	P-value
Visa Sponsorship	33 (5.9%)	66 (20.8%)	3.81e-11
Allows Remote Work	52 (9.2%)	9 (2.8%)	5.42e-04
Offers Relocation	150 (26.6%)	113 (35.5%)	7.08e-03
Highest Degree Preferred:[1]			
Bachelors	122 (35%)	26 (19.4%)	3.32e-03
Masters	79 (22.6%)	34 (25.4%)	
PhD	148 (42.4%)	74 (55.2%)	

[1] Due to missingness, percents are calculated from totals of 349 for USA and 134 for Europe.

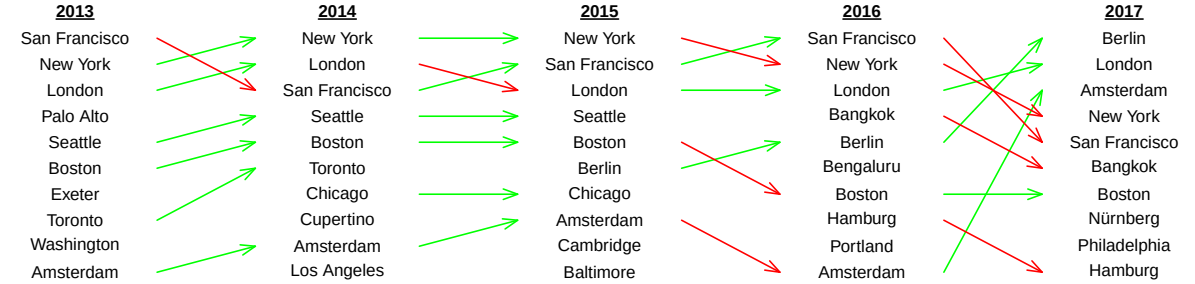
(a) Top 10 Tags by Year



(b) Top 10 Areas of Study by Year



(c) Top 10 Cities by Year



Discussion/Limitations

Discussion:

Limitations:

- results do not generalize to all data science job boards - Stack Overflow could attract jobs from particular industries, could be biased in this way
- skills are detected through tags and not through what is mentioned in the text
- estimates for 2017 jobs are only through September 25, 2017, and may change once the year is complete.

Conclusion