

Trends in Data Science Job Postings on Stack Overflow

Benjamin Ackerman

October 11, 2017

Introduction

- Overall importance of statistics and data-driven jobs - “if there are data, there’s need for a statistician”
- Increased demand for data scientists as more data become available (“big data”)
- “hot new field”
- Stack Overflow’s popularity
- So much change in best coding languages to know and where to look for jobs
- *Find some papers or blog posts that talk about data science as a field and its growth*

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 Stemdegreeelist.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 year and geographic region to determine if there were any changes in frequency of postings by region 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 number of yearly data science job listings posted on Stack Overflow has increased over time, most notably more than doubling between 2013 (n=75) and 2014 (n=174) (Figure 2). Figure 1 displays the overall top ten skill tags, areas of study, and cities, while Figure 4 contains the top ten lists by year, and depicts how the rankings have changed over time. Differences between jobs listed in the United States and jobs listed in Europe are described in Table 1, and the geographic distributions of jobs in both regions are portrayed in Figure 3.

Skill Tags

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. While Python has consistently been the most tagged skill in data science job postings, R has become increasingly more important to employers hiring data scientists over the last three years, as it has jumped from the 5th-most frequent tag to the 2nd-most frequent tag from 2013 to 2017 (Figure 4a). Similarly, knowledge of machine learning algorithms has gained more popularity among employers hiring data scientists in a similar timeframe.

Areas of Study

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%). Few to no changes in preferred areas of study of job candidates have occurred in the past five years, indicating that employers hiring data scientists have consistently sought candidates with quantitative and programming-based backgrounds (Figure 4b).

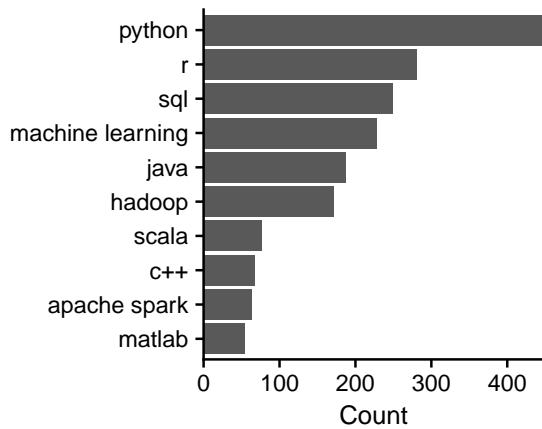
Location

In both the United States and Europe, the highest number of job listings appear to cluster regions where there are major cities. Darker reds appear over big cities like New York and San Francisco (Figure 3a), and London and Berlin (Figure 3b), indicating higher numbers of job listings, while less dense and less industrial areas have lighter shades of yellow, indicating fewer data science job opportunities. Interestingly, while New York and San Francisco are the two cities with the most overall job listings (Figure 1c), and have consistently ranked in the top two cities between 2013-2016, they have dropped to the fourth and fifth spots in 2017, as European cities like Berlin and London have risen to the top (Figure 4c). This geographic trend is also noticable in figure 2, where it is apparent that the proportion of jobs posted in 2017 that are located in Europe is much larger than that of earlier years.

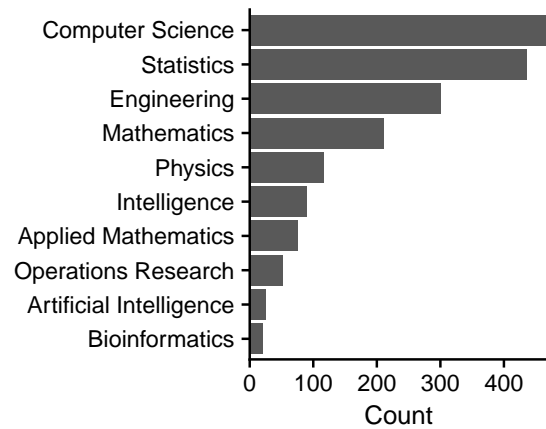
While there are similarities in the types of cities with the most jobs in the US and Europe, there are several differences between the job benefits and characteristics by region. European employers are more likely to offer visa sponsorship (EU: 20.8%, USA: 5.9%, $p < .01$) and to offer assistance with 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 (Table 1).

Figure 1: Most Popular Attributes of Job Listings

(a) Top 10 Tags in Job Listings



(b) Top 10 Areas of Study in Job Listings



(c) Top 10 Cities with the Most Job Listings

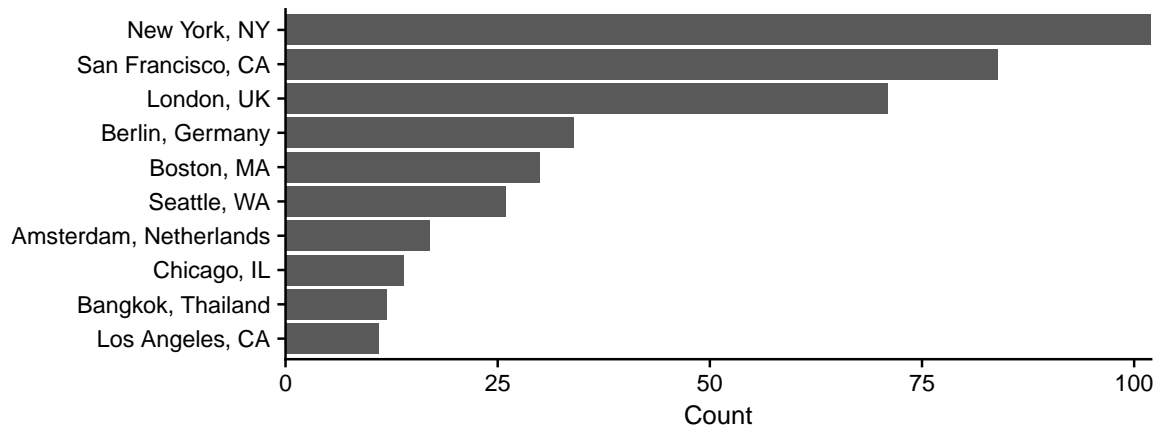


Figure 2: Number of Job Listings by Year and Geographic Region

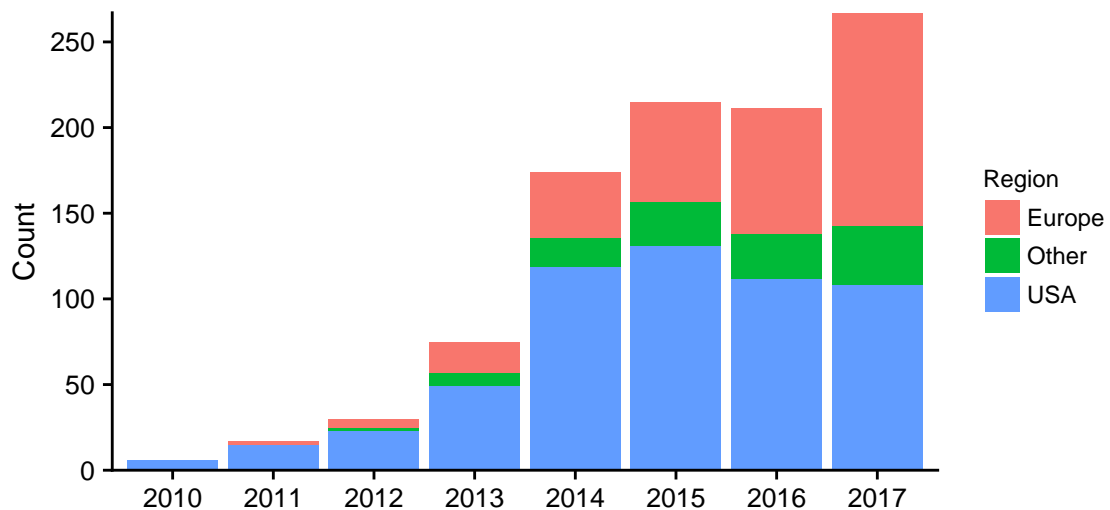


Figure 3: Geographic Distribution of Jobs in the USA vs. Europe

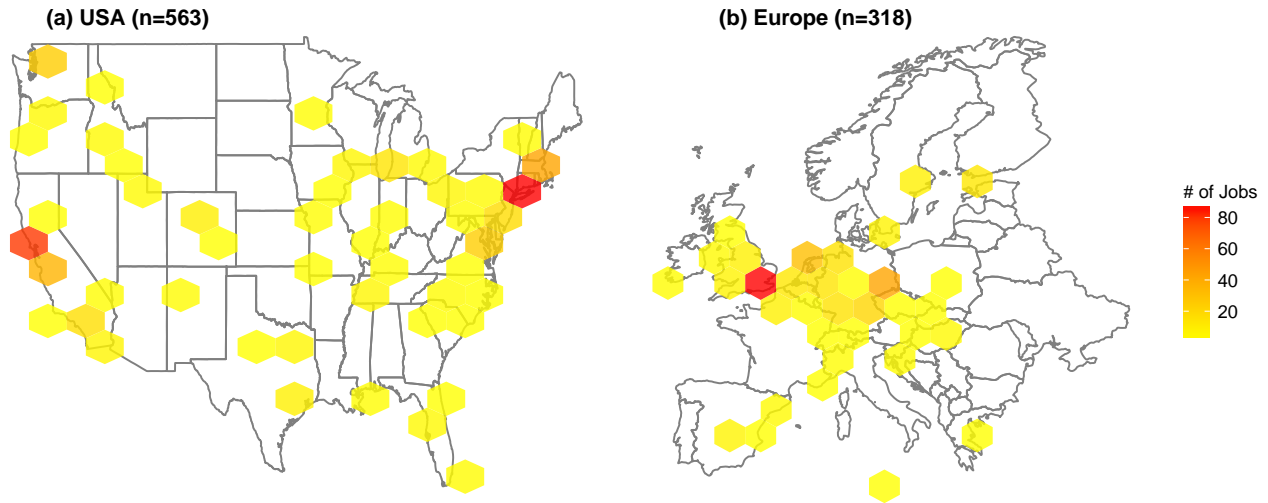


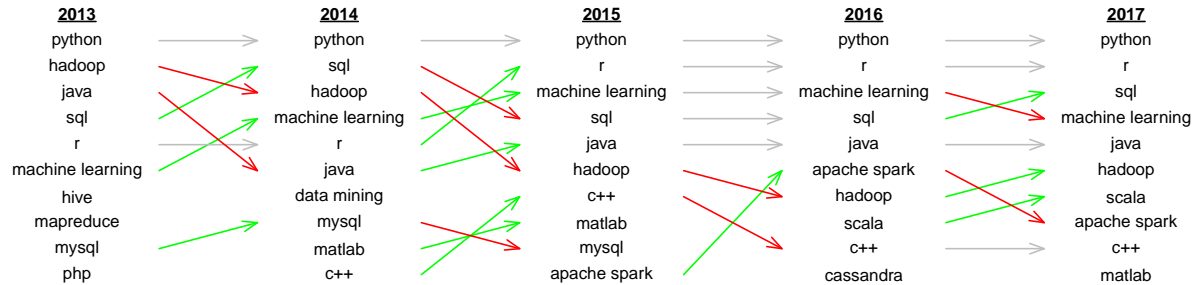
Table 1: Differences between Job Listings in the USA vs. Europe

	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%)	

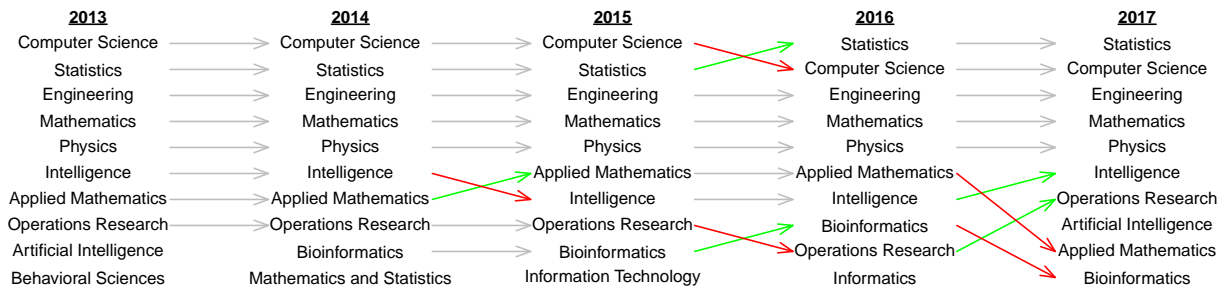
¹ Due to missingness, percents are calculated from totals of 349 for USA and 134 for Europe.

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

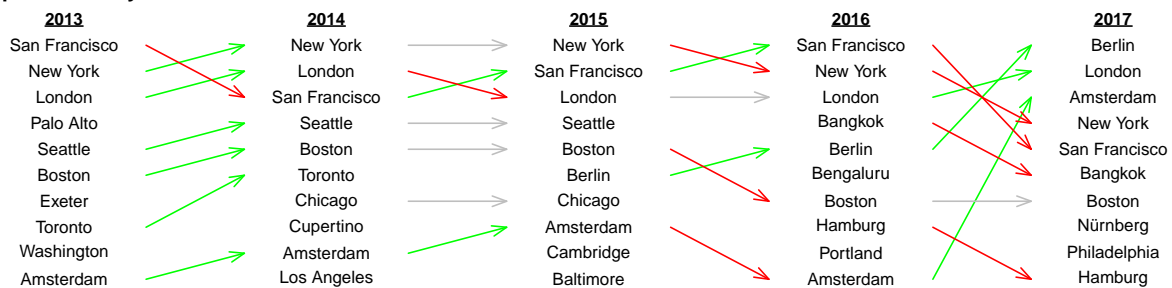
(a) Top 10 Tags by Year



(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, or in particular cities, and could be biased in this way
 - no way to assess mechanism by which SO obtains job listings
- skills are detected through tags and not through what is mentioned in the text
- variability in how job descriptions and requirements are written up, and when they were missing, could result in underestimates of true skills and attributes required
- estimates for 2017 jobs are only through September 25, 2017, and may change once the year is complete.