IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 17, 2018, accepted August 27, 2018, date of publication October 1, 2018, date of current version November 8, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2870535*

# How to Find a Perfect Data Scientist: A Distance-Metric Learning Approach

**HAN HU[1], YONG LUO[2], YONGGANG WEN[2], (Senior Member, IEEE), YEW-SOON ONG[2], AND XINWEN ZHANG[3]**

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2]Data Science and Artificial Intelligence Research Centre, School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798
[3]Hiretual, Mountain View, CA 94043, USA

Corresponding author: Yong Luo (yluo@ntu.edu.sg)

**ABSTRACT** The title of data scientist has been described as one of the sexiest jobs of the 21st century. Numerous efforts have been made to define the job of a data scientist in a qualitative manner by, for example, listing the job functions and required skill sets of data scientists. However, to the best of our knowledge, no attempt has been made to define the term data scientist in a scientific manner. In this paper, we address this issue by using a data-driven approach to answer three questions: 1) What is a proper definition of the term data scientist from a market-demand perspective? 2) Do self-described data scientists meet the market demand? and 3) Finally, how can companies efficiently recruit data scientists that match their openings? To answer these questions, we crawl two data sets for the supply and demand sides. For the former, we collect a set of data scientist user profiles from LinkedIn; for the latter, we collect a set of data scientist job descriptions from Monster. We first parse the set of data scientist job descriptions via natural language processing techniques and derive a scientific definition of the job of a data scientist via a clustering algorithm. Second, we use the same approach to determine that, under the aforementioned definition, self-claimed data scientists on the market would meet the market demand with a high probability. Finally, we introduce a distance-metric learning approach that can be used by companies to find data scientist candidates that match their openings. We achieve an average precision of 12.31%; i.e., one in ten candidates with matching qualifications would accept a given offer. The application of this quantitative approach could significantly reduce the human-resource costs incurred by companies in recruiting matching data scientists.

**INDEX TERMS** Data scientist, natural language processing, distance metric learning.

## I. INTRODUCTION

The demand for data scientists, who are said to have the sexiest job of the 21st century [1], has emerged and dominates the job market. The major driving force underlying the rise of data scientists comes from the advent of the big data era. Given the vast amounts of data that are now available (i.e., big data), companies in the information technology industry have begun to hire more data scientists and to focus on exploiting data and extracting insights to obtain competitive advantages. Evidence of the rapid growth in data scientist positions is also provided by a worldwide employment-related search engine; i.e., Indeed.com. Searching for ''data scientist'' on Indeed.com shows that this title accounted for approximately 0.1% of all job postings by the end of 2016. According to our prediction results based on the historical data (please refer to Appendix A), by the end of 2025, the data scientist job would account for at least 0.5% of the demand for all jobs. Furthermore, the market size of data scientists will surpass that of other relevant data analysis positions, such as business analysts, by August 2021. Moreover, Hiretual [2], the leading platform for AI sourcing, states that 2.7% of automated recruiting tasks during 2017 involved searches for data scientists.

This explosive demand for data scientists has led to some confusion between the demand and supply sides (i.e., companies and job seekers). First, no definition of a data scientist exists that is mutually agreed upon by the supply and demand sides. For the former, companies commonly define data scientists based on their own diverse needs, such as challenges, data products, responsibility, and applications. For the latter, some job seekers claim that they are data scientists based on personal experience, such as skills, college major and


2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

projects. Second, many people have suddenly begun to call themselves data scientists [3] because it is the latest fad, although they lack the typical skill set of data scientists. Third, it is laborious for companies to identify matching candidates for their job openings. Due to the first and second challenges, the matching of job openings defined by the supply side and user profiles generated by the demand side becomes difficult. To hire top data scientists, considerable amounts of manpower are consumed by manual screening.

Many efforts have been made to study the aforementioned challenges in defining and recruiting data scientists. Anjul Bhambhri [4], the vice president of big data products at IBM, provided a high-level description: "*A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who truly wants to learn and bring change to an organization.*" Patil and Davenport [1] described a data scientist as a hybrid of a data hacker, an analyst, a communicator and a trusted adviser and stated that a data scientist should be able to extract insights from the current data tsunami. Other efforts have also defined data scientists in a more explicit way. Dhar [5] specified an integrated skill set for data scientists that includes mathematics, statistics, machine learning and databases. These definitions tend to be based on the experiences of individuals and rely heavily on industry contexts. Regarding job recruitment, current HR practice normally uses keywords to find potential candidates and then adopts rule-based strategies to screen candidates. However, to identify more candidates, HR needs to define more keywords. This practice leads to an explosion in keywords without addressing the real need for job recruitment. In addition, rule-based screening is either too coarse, thus identifying many qualified candidates, or too harsh, missing some qualified candidates.

In response to the aforementioned challenges, we adopt a data-driven approach to answer three essential questions for data scientists, namely:

- *What is a proper definition of the term "data scientist" from a market-demand perspective?*
- *Do self-described data scientists meet the market demand?*
- *Finally, how can companies efficiently recruit data scientists that match their openings?*

We aim to demystify the job of a data scientist using data science methods. Our approach builds upon two datasets, which are collected from the demand and supply sides. For the demand side, we collect a list of data-scientist job descriptions (∼5000) from Monster; for the supply side, we collect a list of user profiles with the job title of data scientist (∼6000) from LinkedIn. The details of our quantitative analysis can be summarized as follows:

- We present two types of definitions, including a descriptive empirical definition and an analytical quantitative definition, from the demand perspective. The empirical descriptive definition summarizes the work of researchers and experts and focuses on responsibilities and skills. The common responsibilities are obtaining, cleaning, exploring, modeling and interpreting data to extract insights from data and developing data products. The typical skill set includes mathematics, machine learning, artificial intelligence, statistics, databases, and optimization. The quantitative analytical definition is based on natural language processing methods. We observe that most job postings can be categorized into a dominant group (i.e., reflecting a common understanding); a typical data scientist is expected to hold a bachelors degree (with a probability of 73%) in Computer Science or Data Science (with a probability of 69%), have 2-5 years of work experience (with a probability of 42%) and be familiar with "analysis", "development", "Python", "R", "machine learning", and "databases" (with a probability of 33%).
- We compare the statistical characteristics of self-claimed data scientists with the definition drawn from the demand-side dataset to show the understanding bias. Using the same NLP method applied to the demand-side dataset, we extract a feature representation of each social media user profile. We observe that most user profiles can be categorized into a dominant group and that approximately 55% of user profiles fall into the range of a typical data scientist drawn from the demand side. This result indicates that both the demand and supply sides share a similar understanding of the term "data scientist". In addition, the distribution of the distances of typical job postings or job seekers related to "data scientist" to the common understanding can be modeled by Rician distributions.
- To reduce the inefficiency in accuracy and labor consumption of traditional rule-based recruitment strategies, we develop a distance metric learning-based recruitment algorithm for use by employers. As some social media users included in the supply-side dataset work at companies that also release job openings that are included in the demand-side dataset, we filter out similar pairs of ⟨social media user profile, job post⟩ (i.e., belonging to the same company) when constructing the associative dataset. These similar pairs are utilized to learn a suitable distance metric that matches candidates efficiently for a given job post. The precision of our algorithm varies from 10.47% to 12.31%, which indicates that at least one candidate will join the company for every 10 recommended job seekers.

It is believed that our quantitative findings can provide deep insights into data scientist employment from both the supply and demand perspectives. In particular, job seekers can build their skill sets according to our definitions, especially the skills that appear with in the definitions with high frequency, to meet the market demand and further accumulate project experience related to the responsibilities of data scientists. Employers can write data scientist openings accurately according to our definitions and judiciously invite matching candidates for job interviews using our proposed distance

metric learning-based algorithm. This practice would significantly reduce the labor costs of recruitment.

The remainder of this paper is structured as follows. Section II introduces the data collection process to crawl data scientist profiles from LinkedIn and data scientist job descriptions from Monster. Section III presents two alternative definitions for data scientists: a qualitative definition summarized from the work of industry experts and a quantitative definition derived from the data scientist job description dataset. Section IV analyzes the understanding bias for the term "data scientist" between the supply and demand sides. Section V develops a distance metric learning approach that can be used by companies to find qualified data scientist candidates with high probability. Finally, Section VI concludes the entire paper.

## II. DEMAND AND SUPPLY DATASETS

In this section, we introduce the process of preparing the two datasets for the supply and demand sides. For the demand side, we crawl a set of job postings with the title of data scientist from Monster [6]. For the supply side, we crawl a set of profiles from LinkedIn [7] in which the users describe themselves as data scientists in their latest job titles.

### A. DEMAND SIDE DATASET - JOB POSTS FROM MONSTER

#### 1) DATA SOURCE

We choose Monster as the data source and crawl job postings with the job title of "data scientist" from Monster to obtain the demand-side dataset. Monster is one of the largest employment websites on which companies release job postings. Compared to the other two popular job listing websites (i.e., LinkedIn and Indeed), Monster stands out as a good source of demand-side data for the following reasons. First, job posting information cannot be easily retrieved from LinkedIn, due to its restrictive data privacy policy. Second, it is generally believed that the quality of job postings on Monster is better than that of the job postings on Indeed because the former charges more for recruitment advertisements and employs a more rigorous review process [8].

#### 2) DATA ENTRY

For each job posting, we crawl the whole webpage and filter out some critical items related to job recruitment to construct the demand-side dataset. A typical job posting consists of the position description, responsibilities, requirements, and company background. Of all the items, we extract the position description, responsibilities and requirements, as shown in Figure 1(a),[1] which are commonly used to screen job seekers, to support our analysis.

#### 3) CRAWLING STRATEGY

We adopt a keyword-based query in which "data scientist" is used as the only keyword, and we employ webpage crawling to gather the job postings related to data scientist positions.

[1] Some items (e.g., company background) may be missed.

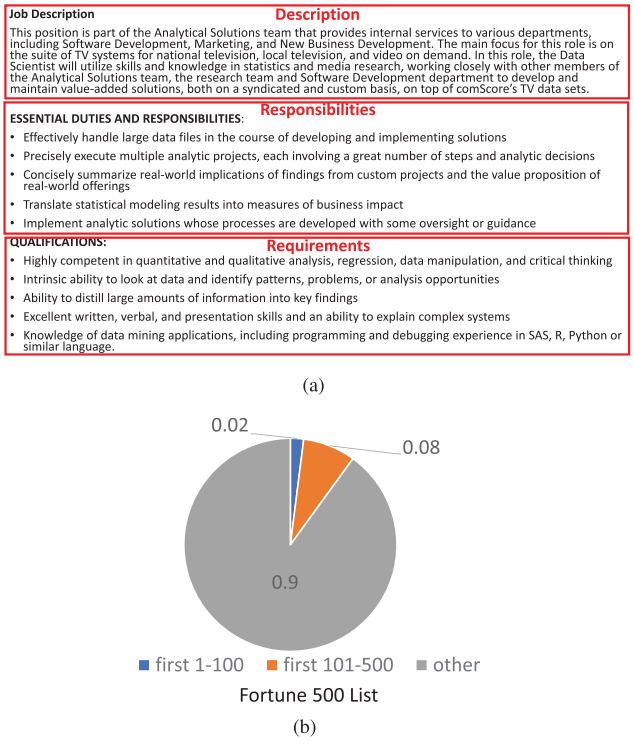



**FIGURE 1. A snapshot of a typical job post with the title "data scientist" from Monster and the summary of companies that released the job openings in our dataset: a) a job post on data scientist, including job description, responsibilities and requirements; b) company scale distribution according to the latest Fortune 500 list.**

The key challenge is that Monster displays no more than 1000 results for a given query. However, the total number of data scientist postings exceeds 10,000 when we crawled the data. To address this challenge, we split the original query into multiple subqueries by combining the keyword "data scientist" and the location qualifier (e.g., Singapore or California, US). In this way, each subquery produces the data scientist results for a location of interest, and the number of job postings returned for each subquery is less than 1000. The list of subqueries is fed into our distributed crawling engine [9].

#### 4) STATISTICS

By preprocessing the job postings extracted from the downloaded webpages, we construct the demand-side dataset. The preprocessing stage utilizes two rules to filter out some incomplete or irrelevant job postings:

- *Removing incomplete job postings*: We remove the job postings from which some critical items are missing, such as the job description, requirement, or responsibilities.
- *Removing irrelevant job postings*: We remove the job postings for which the job titles are not exactly "data scientist".

In this way, we acquire 4,915 job postings. These postings are derived from different types of companies that range

from small startup companies to large firms to guarantee the representativeness of the data. As shown in Figure 1(b), according to the latest Fortune 500 list, 2% of the companies in our dataset are in the first 1-100, and 8% are in the first 101-500.

### *B. SUPPLY SIDE DATASET - USER PROFILES FROM LINKEDIN*

#### 1) DATA SOURCES

We choose LinkedIn as the data source and crawl social media user profiles with the current job title of data scientist from LinkedIn to produce the supply-side dataset. LinkedIn is one of the largest employment-oriented social networking services and provides a platform on which job seekers release personal details. This information is commonly considered to be credible [10].

#### 2) DATA ENTRY

For each user profile, we crawl the entire webpage and filter out some critical items related to job recruitment to construct the supply-side dataset. A typical social media user profile consists of different types of information, such as current affiliation, work experience, and social connections. To support our analysis, we extract only certain critical features, including work experience, educational background and skills, which are commonly used for a job recommendation [11]. A snapshot of such information from a social media user who self-describes as data scientist is shown in Figure 2(a).

#### 3) CRAWLING STRATEGY

We adopt a data crawling approach that is similar to that described in the previous subsection to gather user profiles from LinkedIn. However, LinkedIn sets rigorous limitations on data crawling, including:

- *Access speed*: Regular members can issue only a limited number of requests (e.g., searching and viewing profiles) within a given period of time. Once the limitation is reached, the requests are disabled until the next calendar month.
- *Display of the query results*: Given a query request, no more than 1,000 profiles are shown for each specific search.

Given these two limitations, it would take more than 1 month to download 16,000 user profiles when we conducted the data crawling. To address these challenges and accelerate the crawling speed, we upgrade our accounts to the Recruiter Lite level to acquire more information within a specific time. Furthermore, we utilize a combination of the keyword "data scientist" and different locations to generate the raw queries and then employ a hierarchical division strategy (please refer to Appendix B for more details) to generate the subqueries, thus guaranteeing the completeness of the data.

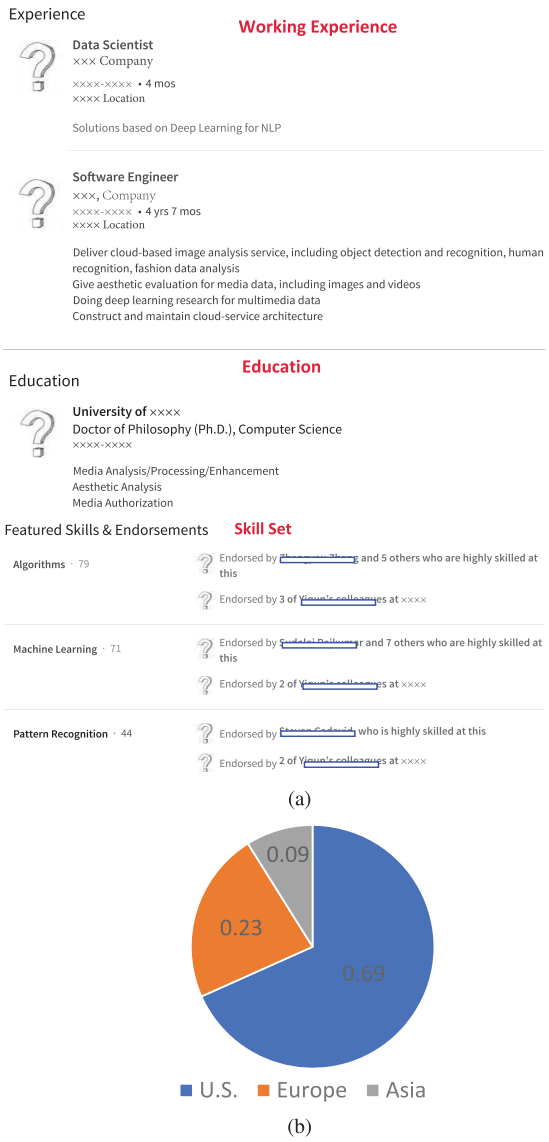



**FIGURE 2. A snapshot of a typical social user profile (self-claimed as data scientist) from LinkedIn and a summary of social users in our supply side dataset: a) a social user profile on data scientist, including working experience, educational background, and skills; b) the geographical distribution of these social users.**

#### 4) STATISTICS

By preprocessing the social media user profiles extracted from the downloaded webpages, we construct the supply-side dataset. We employ three rules in filtering the data:

- *Removing empty social media user profiles*: Social media user profiles that are not available to (or seen by) the public due to privacy issues are removed.
- *Removing incomplete social media user profiles*: Social media user profiles from which some critical information (e.g., work experience) related to job recruitment is missing are removed;
- *Removing irrelevant social media user profiles*: Social media user profiles with job titles that are not exactly "data scientist" are removed.

In this way, we acquire 5,876 social media user profiles; this number corresponds to 36% of the raw social media user profiles. These users are from different regions; thus, the representativeness of the data can be guaranteed. As shown in Figure 2(b), 69% of the users are from the U.S., followed by users from Europe (23%) and Asia (9%).

## III. DEFINING "DATA SCIENTIST" ANALYTICALLY

In this section, we introduce a natural language processing (NLP) approach and derive an analytical definition of the term "data scientist". Specifically, we map the set of job postings from Monster into a feature set. This set includes the highest education, major area of study, work experience and skill set. We then adopt an unsupervised learning approach to identify the main cluster within the group of all job postings. The center of the learned cluster is finally interpreted to produce an analytical definition of the term "data scientist". To our knowledge, this is the first analytically produced definition of the term "data scientist".

### A. TF-IDF METHODOLOGY

This subsection introduces the term frequency-inverse document frequency (TF-IDF) methodology, which is from the field of natural language processing [12]. We utilize this method from to analyze the textual information contained in the Monster dataset.

The TF-IDF method determines how important a given word is in a document contained in a corpus. Words that are common in a single document or a small group of documents tend to have higher TF-IDF values. Given a corpus $D$, the TF-IDF value of word $t$ in document $d$, tf-idf$(t, d, D)$, can be calculated as:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D), \tag{1}$$

tf-idf$(t, d, D)$ is a product of two metrics, including the term frequency tf$(t, d)$ and the inverse document frequency idf$(t, D)$. The term frequency tf$(t, d)$ measures the number of times that word $t$ occurs in document $d$. The inverse document frequency idf$(t, D)$ is a measure of whether the word is common or rare across all documents. For instance, the common words "the" and "and" are not significant. Obviously, a large TF-IDF value is achieved by a high term frequency in a given document and a low document frequency in the corpus.

Using the TF-IDF method, we can construct the feature representation of each job posting via two steps:

- *Feature words selection:* We can jointly use the domain knowledge and TF-IDF methods to select feature words, which characterize documents. For instance, based on our prior knowledge, the word "rain" describes the weather. We can also calculate the TF-IDF value of each word in entire documents and identify the words with higher TF-IDF values as feature words without prior knowledge.
- *TF-IDF value calculation:* For each word in the feature word set, we calculate the corresponding TF-IDF value in a document.

In this way, a document can be simply represented by the feature word set with the associated TF-IDF values in this document.

### B. FEATURE EXTRACTION FOR MONSTER DATASET

This subsection details how we extract the feature representation of each job posting in the Monster dataset, including feature word selection and TF-IDF value calculation.

We use domain knowledge and the TF-IDF method jointly to construct the feature word set. According to [11], educational background, major area of study, work experience, and skill set are commonly utilized in producing job recommendations. We extract these four distinctive types of textual information to construct the set of features, which can be described as follows:

- *Highest education*: This feature refers to the degree qualification that candidates must have. We consider three degrees, "bachelors", "masters", and "Ph.D".
- *Major of study*: This feature is the academic discipline to which candidates formally belong. We categorize different discipline descriptions into 23 majors according to the QS subject list [13].[2]
- *Working experience*: This feature refers to the duration of prior employment measured in terms of the number of years worked. To facilitate calculation, we divide this continuous value into 7 intervals, including "0", "(0, 2]", "(2, 5]", "(5, 10]", "(10, 15]", "(20, 30]", and "(30, $\infty$)".
- *Skill set*: This feature refers to the abilities that candidates should grasp to become qualified data scientists. Some typical skills related to "data scientist" include "analysis", "Python", and "machine learning".

The feature words related to "highest education", "major area of study" and "work experience" are manually selected as mentioned. In contrast, there is no explicit skill set defined for "data scientist". In this work, we utilize the TF-IDF method to select skill set-related words/phrases. 1) We first download the skill set defined by LinkedIn, which consists of more than 3000 words/phrases that describe skills [14]. 2) Next, we calculate the TF-IDF values for these skills in each job posting. 3) Finally, we select the top 100 words/phrases with the largest average TF-IDF values as the feature words. These words and their corresponding term frequencies are shown in the form of a word cloud in Figure 3. As a result, the final feature set consists of $3 + 23 + 7 + 100 = 133$ words/phrases.

Using the feature word set, we calculate the feature representation of each job posting. Given a job posting, we calculate the TF-IDF value for each word/phrase in the feature

[2]We consider 23 majors, including Computer Science, Statistics, Physics, Mathematics, Economics, Data Science, Electrical and Electronics Engineering, Operations Research, Biological Engineering, Industrial Engineering, Business, Chemical Engineering, Information Science, Management Science and Engineering, Astrophysics, Sociology, Cognitive Science, Environmental Engineering, Linguistics, Geography, Transportation Engineering, Psychology, and Engineering.

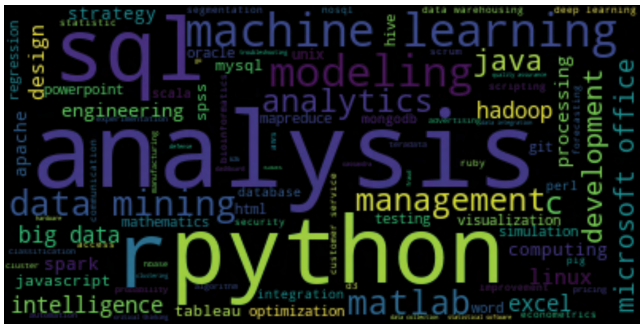



**FIGURE 3.** Word cloud of skills related to "Data Scientist". Some typical skills are "analysis", "Python", "machine learning", "SQL", etc.

word set using Eq. (1). All 133 TF-IDF values make up a vector that represents a job posting.

### C. UNSUPERVISED LEARNING TO DEFINE DATA SCIENTIST DEFINITION ANALYTICALLY

Based on the feature representation of each job posting, we utilize the hierarchical clustering algorithm [15] to explore the latent commonality of these job postings. Hierarchical clustering seeks to build hierarchies of clusters, which are more informative than the unstructured sets of clusters produced by flat clustering. This method requires no prior knowledge of the number of clusters required. The clustering procedure works as follows:

- *t-SNE based dimensionality reduction:* Because each feature vector is 133-dimensional, we utilize t-distributed stochastic neighbor embedding (t-SNE) [16] to perform dimensionality reduction and intuitively represent some characteristics of the job postings. t-SNE is a nonlinear dimensionality reduction technique that embeds high-dimensional data into a low-dimensional space (e.g., two or three dimensions). Similar objects are modeled by nearby points, and dissimilar objects are modeled by distant points. We utilize the t-SNE method to project each feature vector into a 2-dimensional space and show them in Figure 4(a). We note that most of the points are concentrated in space, and some points are located farther apart.
- *Initial hierarchical clustering:* After performing dimensionality reduction, we apply the hierarchical clustering algorithm to these job postings. In particular, we utilize the Euclidean distance to measure the distance between each pair of job postings and group the closer pairs into clusters. The value of the cophenetic correlation coefficient [17], which measures the performance of clustering, is 0.7244.
- *Optimal cluster number determination:* As the number of clusters is unknown, we use the Davies-Bouldin index [18] to explicitly find the optimal value. The Davies-Bouldin index measures both the separation between clusters and the cohesion within clusters; thus, it mathematically guarantees good clustering results. A smaller Davies-Bouldin index reflects better clustering performance. We vary the number of clusters from 1 to 10 and show the corresponding value of the Davies-Bouldin index in Figure 4(b). We note that the smallest value of the Davies-Bouldin index occurs when the number of clusters is 3. This result suggests that the optimal number of clusters is 3.

We categorize the job postings into 3 clusters based on the cluster tree and find that almost all of the job postings (approximately 98%) fall into the same class (the green points in Figure 4(a)), which excludes 2% of the job postings (the blue and yellow points). Thus, there exists a dominant cluster within all of the job postings. In other words, these companies share a common understanding of the term "data scientist".

Based on the observations shown in Figure 4(a), we can use the center (the black diamond shown in Figure 4(a)) of all of the job postings to represent a typical data scientist and further present a quantitative definition of the term "data scientist". We calculate the Euclidean distance between each point (i.e., job posting) and the centroid and we show the cumulative distribution function of these distances in Figure 4(c). We note that the distance to 80% of the points from the centroid is less than 53.5801. We can conclude that **given a recruitment description, if the Euclidean distance between the corresponding feature vector and the centroid in Figure 4(a) is less than 53.5801, the description is related to a data scientist.**

Furthermore, we study the probability density of these distances using some well-known distributions. We utilize 14 distributions to fit the cumulative distribution function of the distance to the centroid shown in Figure 4(d) and calculate the corresponding log-likelihood value as shown in Table 1. A larger log-likelihood value indicates a better fit. We note that the Nakagami distribution leads to the best fitting result. The Nakagami distribution has the following density function:

$$f(x;\mu,\omega) = 2\left(\frac{\mu}{\omega}\right)^{\mu}\frac{x^{(2\mu-1)}e^{-\frac{\mu}{\omega}x^2}}{\Gamma(\mu)},$$

where $\mu$ is the shape parameter and $\omega$ is the scale parameter. The Nakagami distribution is related to the Gamma distribution. In particular, if $x$ has a Nakagami distribution, then $x^2$ has a Gamma distribution. In Figure 4(d), the parameters for the Nakagami distribution are $\mu = 0.9181 \pm 0.0163$ and $\omega = 1706 \pm 25.5625$.

We also employ the Rician distribution, which achieves the best fitting performance when applied to the LinkedIn dataset discussed in the following section, to fit the probability density curve shown in Figure 4(d). The Rician distribution has the following density function:

$$f(x;s,\sigma) = I_0\left(\frac{xs}{\sigma^2}\right)\frac{x}{\sigma^2}e^{-\frac{x^2+s^2}{2\sigma^2}},$$

where $s \geq 0$ is the noncentrality parameter and $\sigma$ is the scale parameter. $I_0(\cdot)$ is the zeroth-order modified Bessel function of the first kind. In Figure 4(d), the parameters

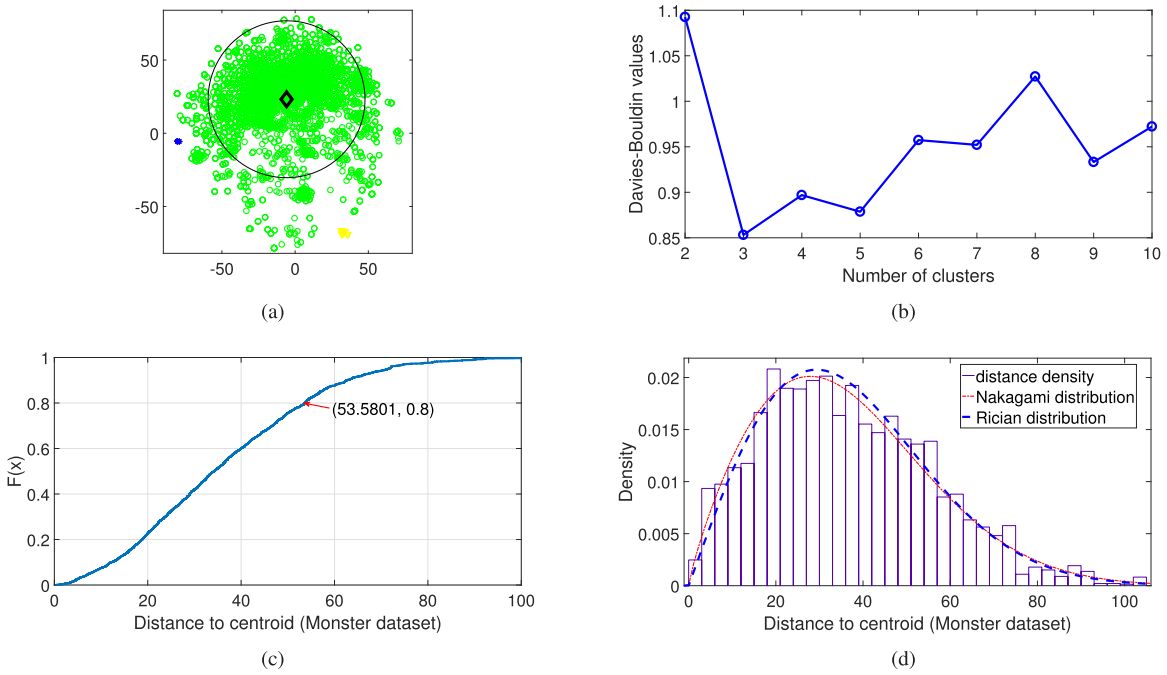



**FIGURE 4.** Hierarchical clustering on the demand side dataset: a) projection on 2-D space using the t-SNE method on the raw feature vectors of all the job posts; b) the Davies-Bouldin index under different cluster number; c) CDF of distance between points and the centroid; d) distribution fitting for the distances between job posts and the centroid. The results show that: 1) all the job posts share a common understanding on "Data Scientist"; 2) the distance between job posts and the centroid of the dominant cluster can be modeled by the Nakagami distribution or Rician distribution; and 3) we can use a circle, defined by the centroid and a radius corresponding the distance upper bound of 80% of the job posts to the centroid, to represent the analytical definition for "Data Scientist".

**TABLE 1.** Log Likelihood value using different distribution functions.

| | BS[a] | Exponential | EV[b] | Gamma | GEV[c] | IG[d] | LL[e] | Logistic | Lognormal | Nakagami | Normal | Rayleith | Rician | Weibull |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Demand | -22026 | -22286 | -22004 | -21205 | -21150 | -22203 | -21489 | -21378 | -21676 | **-21067** | -21304 | -21079 | -21079 | -21073 |
| Supply | -23639 | -24826 | -23317 | -22553 | -22271 | -23770 | -22721 | -22250 | -23195 | -22306 | -22263 | -22492 | **-22216** | -22267 |
| Joint | -26207 | -27513 | -25576 | -25436 | -25152 | -26315 | -25732 | -25359 | -25939 | -25204 | -25209 | -25319 | **-25124** | -25152 |

[a] Birnbaum-Saunders [b] Extreme Value [c] General Extreme Value [d] Inverse Gaussian [e] Log Logistic

**TABLE 2.** Mean and variance of 2 different distributions.

| | Nakagami | | Rician | |
|---|---|---|---|---|
| | mean | variance | mean | variance |
| Demand | 36.2391 | 392.863 | 36.6324 | 364.532 |
| Supply | 30.6442 | 178.008 | 30.8787 | 163.582 |
| Joint | 49.5027 | 504.982 | 49.8322 | 472.341 |

for the Rician distribution are $s = 12.8554 \pm 6.8969$ and $\sigma = 27.7597 \pm 1.6068$. In fact, the Nakagami distribution and the Rician distribution are widely used in communication theory to model the scattered signals that reach a receiver by multiple paths. The Nakagami distribution is used for dense scatters, while the Rician distribution models fading with a stronger line-of-sight. For the density curve shown in Figure 4(d), the log-likelihood values are quite close to one another ($-21067$ and $-21079$ respectively) when we employ these two distributions. The corresponding mean and variance are shown in Table 2, and these values are also quite similar.

In addition, we investigate the common characteristics of the job postings by calculating the probability distribution of the top feature words, such as "bachelor" and "computer", and we find that these job postings are highly consistent in terms of certain features. In particular, we observe that 73% of the job postings require a "bachelors" degree or above, 69% of the job postings limit the major area of study to "computer science" or "data science", 42% of the job postings expect the candidates to have 2-5 years of work experience, and 33% of the job postings require the skill set to include "analysis", "development", "Python", "R", "machine learning", and "databases".

Therefore, we can conclude that a typical data scientist is expected to hold a bachelors degree in computer science or data science, be familiar with "analysis", "development", "Python", "R", "machine learning", and "databases" and have 2-5 years of prior work experience.

## IV. UNDERSTANDING THE BIAS BETWEEN THE DEMAND AND SUPPLY SIDES

In this section, we aim to investigate the second question, i.e., "do self-described data scientist meet the market demand", in an analytical manner. Specifically, we adopt the

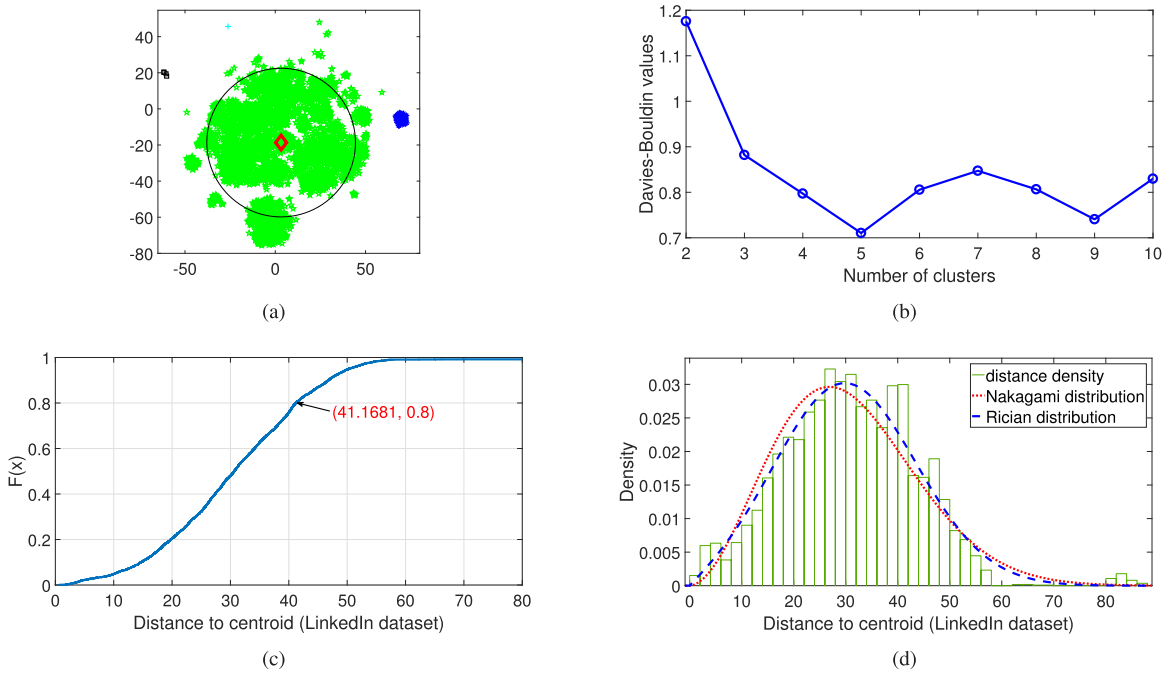



**FIGURE 5.** **Hierarchical clustering on the supply side dataset: a) projection on 2-D space using the t-SNE method on the raw feature of all the user profiles; b) the Davies-Bouldin index under different cluster number; c) CDF of distance between points and the centroid; d) distribution fitting for the distances between user profiles and the centroid. The results show that: 1) all the user profiles share a common understanding on "Data Scientist"; 2) the distance between user profiles and the centroid of the dominant cluster can be modeled by the Rician distribution or Nakagami distribution; and 3) we can use a circle to represent the analytical definition for "Data Scientist" from the supply side.**

same NLP approach to analyze the two datasets from Monster and LinkedIn and derive some quantitative measurements of the understanding bias between the demand and supply sides.

### A. DATA SCIENTIST UNDERSTANDING FROM SUPPLY SIDE

This subsection details how we utilize the feature extraction and unsupervised learning method described in Section III to investigate the understanding of the term "data scientist" using the LinkedIn dataset.

To perform feature extraction using the LinkedIn dataset, we utilize the same feature word set described in Section III-B and the TF-IDF method (Eq. (1))[3] to build a feature representation of each user profile. In this way, each user profile is represented by as a 3-dimension vector.

We then utilize the hierarchical clustering algorithm to explore the latent commonality of these user profiles. The clustering procedure works as follows:

- *t-SNE-based dimensionality reduction*: Similarly, we utilize the t-SNE method to perform dimensionality reduction[4] and obtain some intuitive characteristics of the user profiles. Each user profile is projected into a 2-dimensional space, as shown in Figure 5(a). We note that most of the points are concentrated in this space, and several points are located farther apart.
- *Initial hierarchical clustering:* After dimensionality reduction, we apply the hierarchical clustering algorithm to the user profiles. Similarly, the Euclidean distance metric is selected to measure the distance between pairs of user profiles. The cophenetic correlation coefficient is 0.6284.
- *Determination of the optimal number of clusters:* We categorize the user profiles into different numbers of clusters (varying from 1 to 10) and calculate the corresponding Davies-Bouldin index to find the optimal number of clusters. The results are shown in Figure 5(b). The optimal number of clusters is clearly 5; this value leads to the smallest Davies-Bouldin index.

Therefore, we cluster the user profiles into 5 groups based on the cluster tree. We observe that almost all of the user profiles (approximately 96%) fall into the same group (the green points shown in Figure 5(a)), and this group excludes less than 4% of the user profiles. Thus, there exists a dominant cluster for all of the user profiles. This result indicates that these job seekers share a common understanding of the term "data scientist".

We can utilize the dominant cluster to extract a quantitative understanding of the term "data scientist" from the supply side. In particular, the centroid (the red diamond shown in Figure 5(a)) can be considered as a typical job seeker for data scientist positions. We then calculate the Euclidean distance between each point (i.e., user profile) and the centroid, and we show the cumulative distribution in Figure 5(c).

[3]The corpus is the union of LinkedIn dataset and Monster dataset.

[4]We apply t-SNE to all of the feature vectors for the job postings and user profiles simultaneously.

We note that the distance of 80% of the points is less than 41.1681. Using a radius corresponding to this value and the centroid, we draw the circle shown in Figure 5(c) to represent a quantitative understanding of the term "data scientist".

Furthermore, we study the probability density of these distances using 14 well-known distributions. From Table 1, we note that the Rician distribution leads to the best fitting performance. We utilize the Rician distribution and the Nakagami distribution to fit the density curve shown in Figure 5(d). The parameters for the Rician distribution are $s = 26.7798 \pm 0.2441$ and $\sigma = 14.1407 \pm 0.1742$. The parameters for the Nakagami distribution are $\mu = 1.4136 \pm 0.0242$ and $\omega = 1117.07 \pm 12.5495$. By comparing the log-likelihood values, means and variances of these two distributions shown in Table 1 and Table 2, we note that they yield similar performance.

### B. UNDERSTANDING BIAS

This subsection uncovers the bias and commonality between the demand and supply sides by jointly analyzing the demand- and supply-side datasets.

We project the feature vectors of the job postings and user profiles into the 2-D space shown in Figure 6(a) and obtain some intuitive characteristics. We obtain the following observations:

- *Observation 1:* The two clouds of points intersect one another. The points related to the user profiles (i.e., the red points) can be considered to be a subset of the set defined by the points related to the job postings (i.e., the green points). This result indicates that the understanding of the term "data scientist" derived from any user profile is coincident with that from certain companies.
- *Observation 2:* Many points related to the user profiles fall into the circle derived from the Monster dataset (i.e., the typical understanding from the demand side). In particular, 55% of the red points fall into the circle derived from the Monster dataset, and 52% of the red points fall into the intersection between the two circles from the Monster dataset and the LinkedIn dataset. This result reveals that the understanding of the term "data scientist" derived from more than half of the job seekers is consistent with the typical definition of the term "data scientist" derived from the demand-side dataset.
- *Observation 3:* The centroid of the LinkedIn dataset falls into the circle defined by the Monster dataset. This result indicates that the typical job posting for data scientists and a typical job seeker share a similar understanding.

We can conclude that **both the demand and supply sides share a similar understanding of the term "data scientist"**.

Similarly, we study the probability density of the distances between the user profiles and the centroid defined by the Monster dataset. From Table 1, we note that the Rician distribution leads to the best fitting performance. We utilize the Rician distribution and the Nakagami distribution to fit the density curve shown in Figure 6(b). The parameters for the Rician distribution are $s = 41.8654 \pm 0.4619$ and $\sigma = 24.5234 \pm 0.3227$; on the other hand, the parameters for the Nakagami distribution are $\mu = 1.3059 \pm 0.0222$ and $\omega = 2955.5 \pm 34.5456$. From Figure 4(d), Figure 5(d), and Figure 6, we find that the **Rician distribution can be utilized to model the distribution of the distances from typical job postings or job seekers related to the term "data scientist" to the common understanding of this term**.

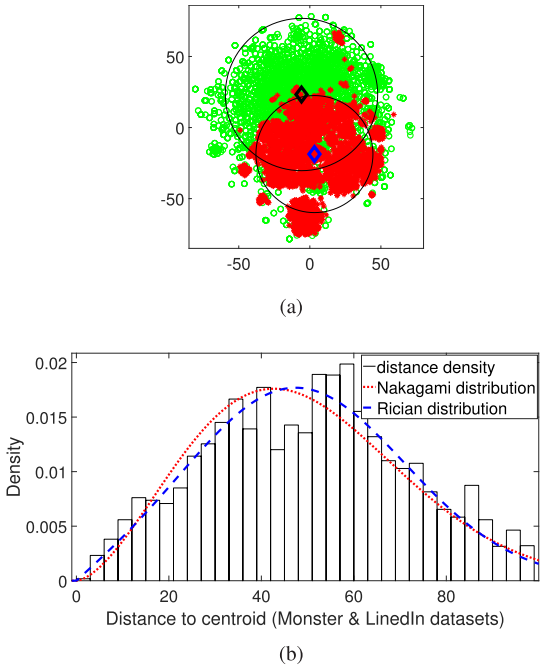



**FIGURE 6.** Joint analysis on the demand and supply side datasets: a) projection on 2-D space using the t-SNE method on the raw feature. Green points represent the job posts and red points are the user profiles; b) distribution fitting for distances between typical user profiles related to data scientists and the centroid of Monster dataset. The results show that: 1) both the demand side and supply side share similar understanding on "Data Scientist"; and 2) Rician distributions can be utilized to model the distance distribution of typical job posts or job seekers related to "Data Scientist" to the common understanding.

## V. RECRUITMENT ALGORITHM FOR DATA SCIENTIST

In this section, we proceed to investigate the third question; i.e., "How can companies efficiently recruit data scientists that match their openings", via a distance-metric learning approach. We start by adopting a rule-based query approach that companies can use to look for potential matching candidates in profession-based social networks (e.g., LinkedIn). This approach may guide employers to better design their job requirements. This approach also motivates a distance-metric learning framework that a company can use to interview qualified candidates with greater chances of joining the company if an offer is extended. Finally, we verify our proposed algorithm using the two datasets in this paper, and our numerical analysis shows that our algorithm can provide a one-in-ten chance for a candidate to accept the offer, significantly reducing the recruitment costs incurred by employers.

### A. RULE-BASED CANDIDATE QUERY

The conclusions presented in the last section indicate that it is possible for companies to find potential candidates from professional social networks. We employ a straightforward rule-based strategy, which is supported by LinkedIn, to investigate some preliminary recruitment results. For example, how many candidates comply with the requirements? The rule-based strategy employs two rules for candidate screening:

- A qualified candidate must satisfy the requirements related to the educational degree, major area of study, and work experience;
- A qualified candidate is expected to be familiar with a certain percentage of the entire skill set.

For instance, if a job posting requires applicants with a Masters degree in computer science or statistics, 2 years of work experience, and 5 different skills, a job seeker with a Masters degree in computer science and 3 years of work experience meets the basic requirements, and he/she should have at least 2 of the skills of interest to be a qualified candidate if the threshold is 40%. We consider 6 different algorithms. The first algorithm, termed *base*, only requires the potential candidates to meet the requirements related to degree, major area of study, and work experience. The other five algorithms require the candidates to satisfy the additional requirement related to a certain percentage of the desired skill set (from 10% to 50%), termed *base+10% (or 20%, 30%, 40%, 50%) skills*.

Figure 7 shows a comparison of the 6 different algorithms. For the *base* algorithm, we note that approximately 1-500 job seekers satisfy the basic requirement for 31% of the companies, and 69% of the companies can find more than 500 potential candidates. For the other 5 algorithms, clearly fewer job seekers satisfy the more stringent skill requirements. In particular, 21% of the companies cannot find any qualified candidates using the *base+10%skills* algorithm, and this value is 93% for the *base+50%skills* algorithm. 22% of the companies can find 1-500 qualified candidates using the *base+10%skills* algorithm, and this value is 2% for the *base+50%skills* algorithm. These results reveal two phenomena:

- **Case 1: A considerable number of companies (21%-93%) cannot find a suitable candidate.** For these companies, we investigate the distance between the corresponding job postings and the centroids derived from the LinkedIn and Monster datasets shown in Figure 7(b) and Figure 7(c).[5] The distance to the two centroids is clearly larger than that of companies that have greater numbers of candidates. This result indicates that understanding bias between the demand and supply sides may result in recruitment failure. The result suggests that *the recruitment information should be consistent with the typical understanding from both the demand and supply sides.*

[5]These results were obtained using the *base+30%skills* algorithm. The other algorithms lead to similar conclusions.

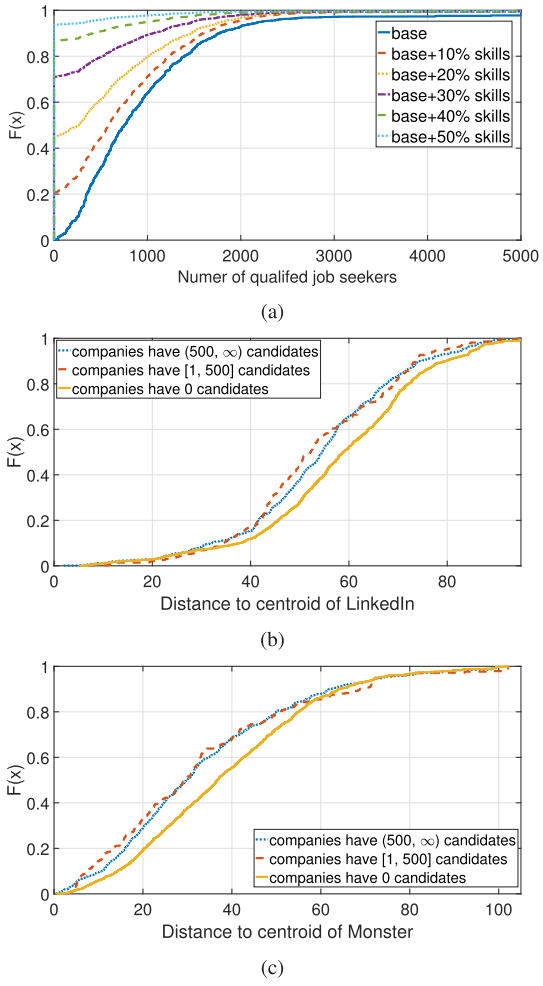



**FIGURE 7. Recruitment results using 6 rule-based algorithms: a) CDF of number of candidates who are qualified to job posts; b) CDF of distance between job posts to the centroid of supply side dataset; c) CDF of distance between job posts to the centroid of demand side dataset. The results show that: 1) A considerable number of companies (21%-93%) cannot find a suitable candidate, and some companies can find a large number of candidates; and 2) The distance of job posts to the two centroid is larger than that of companies which have more candidates.**

- **Case 2: Some companies can find a large number of candidates.** Using the aforementioned algorithms, approximately 2%-22% of the companies can find 1-500 candidates, and 5%-57% of the companies can find more than 500 candidates. As shown in Figure 7(b) and Figure 7(c), the distances of these two types of companies to the two centroids is difficult to distinguish but smaller than that of companies with 0 candidates. *To identify qualified candidates within this large pool is labor-intensive for job recruitment. This case motivates us to design an efficient recruitment algorithm to alleviate labor consumption in candidate screening.*

### B. DISTANCE-METRIC LEARNING FRAMEWORK FOR CANDIDATE RECOMMENDATION

We introduce a distance metric learning method to recommend candidates for companies according to job postings

and social media user profiles; our objective is to improve recruitment efficiency and reduce labor costs. Distance metric learning algorithms learn a transformation optimized to yield small distances between similar pairs of points and large distances between dissimilar pairs of points [19]. Let $q$ and $x$ denote a job posting created by company $A$ and the profile of social media user $U$, respectively. If user $U$ holds a position in company $A$, $q$ and $x$ can be viewed as a similar pair; otherwise, they are a dissimilar pair. The candidate recommendation problem can be formally stated as follows:

$$\min_{W\succeq 0,\mathrm{rank}(W)=m} \sum_{q\in\mathcal{Q}} \sum_{x^+\in\mathcal{X}_q^+} \mathcal{L}(r_q(x^+)) + \lambda\Omega(W), \tag{2}$$

where

$$2f_q(x) = -\|q-x\|_W^2 = (q-x)^T W(q-x), \tag{3}$$

$$r_q(x^+) = \sum_{x^-\in\mathcal{X}_q^-} \mathbf{I}[f_q(x^-) - f_q(x^+)], \tag{4}$$

Here, $W \in \mathbb{R}^{d\times d}$ is the distance metric we wish to learn, $d$ is the dimension of $x$ or $q$ (i.e., 133), and $m < d$ is a factor that is used to control the scalability and computational complexity of the computations for large-scale datasets; $\mathcal{Q}$ is the set of job postings; $\mathcal{X}_q^+$ and $\mathcal{X}_q^-$ are the training sets of the social media users who are relevant and irrelevant to job posting $q$, respectively, and $\langle q, x^+\rangle$ and $\langle q, x-\rangle$ are the similar and dissimilar pairs, respectively; $f_q(x)$ is the distance between $q$ and $x$, $\mathbf{I}(x) = 1$ if $x > 0$ and 0 otherwise; $r_q(x^+)$ is the rank of $x^+$, which we define as the number of the points in $\mathcal{X}_q^-$ that are closer to $q$ than $x^+$; $\Omega(W)$ is a regularizer on $W$, and $\lambda$ is a weight factor; and $\mathcal{L} : \mathbb{Z}^+ \to \mathbb{R}^+$ is a mapping that transforms $r_q(x^+)$ into a loss. The objective function in Eq. (1) consists of two terms, including a loss function and a regularizer. Roughly, given $q$, we aim to minimize the loss function and $r_q(x^+)$, which indicates that few dissimilar points $x^-$ have smaller distances to $q$ than the similar points $x^+$. The regularizer is used to prevent overfitting.

Based on the algorithm described in [20], we learn the optimal distance[6] metric $W$ and recommend job seekers for companies. In particular, given a job posting $q$ and a collection of user profiles $\mathcal{X}$, we can calculate the distance between any profile $x \in \mathcal{X}$ and the job posting $q$ using $W$ and then rank these user profiles according to the distances. In general, a user profile with a smaller distance to $q$ will have a higher probability of joining the company.

### C. NUMERICAL RESULTS

As employees commonly meet the needs of working companies, we jointly utilize the LinkedIn and Monster datasets to construct the required dataset for training and testing based on the following assumption:

[6]In Section III-C, we utilize the Euclidean distance for clustering. Here, the learned distance aims to measure the similar pairs from the two datasets (LinkedIn and Monster) and cannot be utilized to perform clustering in one dataset.

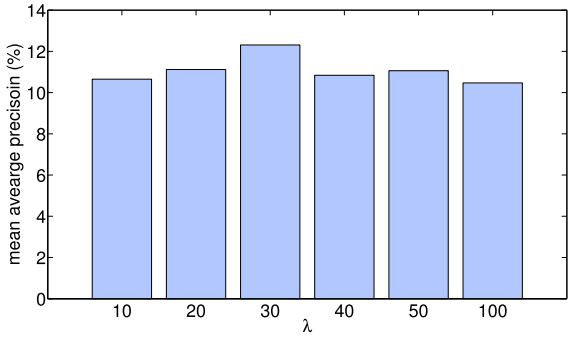



**FIGURE 8.** The MAP of our distance metric learning based algorithm for recruitment under different $M$. The MAP can be up to 12.31%, i.e., one in ten candidates with matching qualifications would have accepted the offer.

*Assumption 1:* For an employee U who works at a company A, his/her profile matches the requirements (at least approximately) of the corresponding job posting at that company.

Real cases include considerable noise because employees may not update their profile for extended periods. This effect can be reduced by filtering the candidates who have recently joined a company. The construction procedure is as follows: 1) For the job postings from Monster, we can extract the names of companies who release the recruitment messages; 2) For the social media user profiles from LinkedIn, we can also extract the names of companies where the users work. 3) If a job posting and a social media user are related to the same company, we can construct a query $q$ (i.e., a job posting), a point $x$ (i.e., a social media user profile), and the resulting similar pair $\langle q, x\rangle$. In this way, we extract 230 job postings and 3802 social media users. Given a job posting or query $q$, the relevant set $\mathcal{X}_q^+$ is defined as all of the social media users from the same company, and the irrelevant set $\mathcal{X}_q^-$ is the set of all social media users from other companies.

We run our algorithms with different settings to study the recommendation performance. In the experiments, for a given $m$, $W$ is initialized as the product $LL^T$, where the entries in $L$ are generated from the standard normal distribution. The weight factor $\lambda$ is fixed at 0.1. We vary $m$ from 10 to 100 and utilize 10-fold cross validation; i.e., 90% of the job postings are used for training, and 10% are used for testing. We report the mean average precision (MAP) of the predicted rankings in Figure 8. For the rule-based algorithms, we conduct experiments to find the most appropriate parameter settings and utilize the Euclidean distance to perform the ranking. As the MAP of these algorithms is quite small (approximately 0.1%), we do not show the results in Figure 8. Given different values of $m$, the MAP of the distance metric learning algorithms varies from 10.47% to 12.31%, which indicates that one candidate is expected to join the company if we recommend 10 job seekers.

## VI. CONCLUSION

Based on collected data relating to data scientists from Monster and LinkedIn, we investigated the properties of

data scientist employment from the perspectives of both demand and supply. We introduced a data crawling procedure to gather sufficient and unbiased data to support our analysis. Subsequently, we presented two types of definitions, including a descriptive empirical definition and an analytical quantitative definition. The descriptive empirical definition was summarized from the work of researchers, experts or tech bloggers, and the analytical quantitative definition was derived from the demand-side dataset using natural language processing approaches. Finally, we found that both the demand and supply datasets shared a similar definition of a data scientist. Motivated by the observation that companies still encounter challenges in finding qualified data scientists, we introduced a distance metric learning-based algorithm for candidate recommendation, which is shown to provide improved precision for the data scientist market. In our future work, we will use the same datasets to mine other hidden patterns pertaining to data scientist employment, such as how to improve the skills of data scientists and what the key feature of a good data scientist is.

## APPENDIX

### A. MARKET TRENDS OF DATA SCIENTISTS

The job demand for data scientists has experienced tremendous growth in the past several years, as shown in Figure 9. As of the beginning of 2012, there were hardly any job postings for data scientists. Since then, the demand for data scientists has grown dramatically. By the end of 2016, data scientists accounted for approximately 0.1% of all job postings. One major driving force for such phenomenal growth is the data explosion. In each minute in 2016, over 2 million searches on Google, 3 million shares on Facebook, and 150 million email exchanges were generated; moreover, 2 million videos were watched on YouTube and 200 thousand minutes of audio chat took place through WeChat [21]. Currently, the rapid generation of data occurs almost everywhere. As the new oil in the 21st century, data act as the source of advanced analytics [22]. To extract actual value from data, data scientists use data in creative ways, and the market demand for data scientists is growing continuously. To better model the increase in demand for this job, we derive two types of prediction functions. One is more pessimistic, while the other is more optimistic. The red line shown in Fig. 9 corresponds to a pessimistic prediction that is derived using a polynomial function, $(0.07\,x^3 + 1.46\,x^2 + 4.29\,x + 3.06) \times 10^{-3}$, where $x$ is the number of months since the beginning of 2012. This function approximates the actual values and ensures that at least 95% of the actual values are above the line. Similarly, the green line refers to the optimistic prediction expressed as $(0.7x^3 - 3.9\,x^2 + 19\,x + 2.3) \times 10^{-3}$, and 95% of the actual values are below this curve. According to the prediction results, by the end of 2025, data scientists will account for at least 0.5% of the whole job demand, and this fraction could be as high as 1.3%, according to the optimistic prediction.

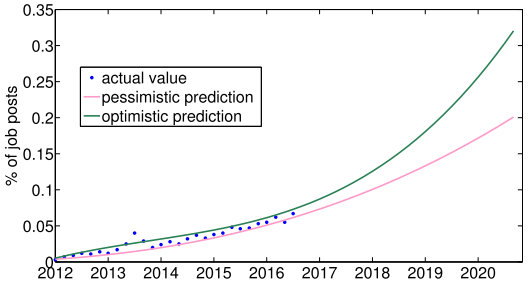



**FIGURE 9.** The percentage of real job postings for data scientist from 2012 to 2016. Higher percentage value means higher market demand for data scientist. The percentage grows out of almost nothing to around 0.1% of the whole job postings in 2016. Both the pessimistic and optimistic predicting functions are approximate the real values and ensure 95% real values are above/below the line accordingly. [Source: www.indeed.com].

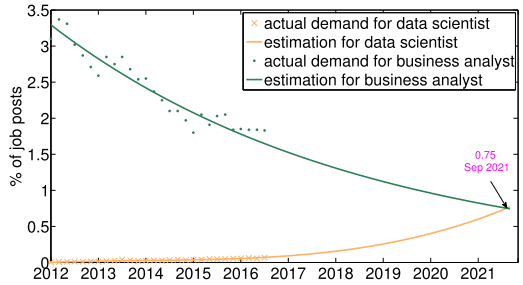



**FIGURE 10.** The percentage of job postings for data scientist and one of its traditional counterparts, business analyst, from 2012 to 2016. Higher percentage value means higher market demand for data scientist. Although the demand gap between data scientist and its traditional counterpart narrows in time, data scientist is still in its initial development stage. Depending on the prediction result, these two job positions can achieve the same proportion, 0.75%, of the whole job demand by August 2021. [Source: www.indeed.com].

In addition, we show the changes in job demand for business analysts and data scientists in Figure 10. On the one hand, in contrast to data scientists, the traditional counterpart exhibits the opposite pattern of shrinking demand in recent years. In many scenarios, the data scientist is an expert who is proficient with big data analysis skills and has outstanding business understanding. Thus, the increase in the prevalence of data scientists has diminished the market share of its traditional counterparts. On the other hand, although the demand gap between data scientists and their counterparts has narrowed over time, this gap is still significant. By the end of 2016, there were still 20 times more job postings for business analysts than for data scientists. This result again suggests the huge potential future growth in data scientist positions. To better model these two curves, we derive a regression function for each job position, $(1.5\,x^3 - 9.6\,x^2 + 28\,x - 2.9) \times 10^{-3}$ for data scientists and $3.4\,e^{-0.15x}$ for business analysts. As we can see from Fig. 10, data scientists do not account for the same proportion of overall job demand as business analysts until September 2021.

### B. CHALLENGES AND METHODS FOR DATA CRAWLING

When crawling information from LinkedIn and Monster, we take advantage of their APIs by using the ''field selectors''

syntax, such as keywords. To guarantee strong consistency, we only use "data scientist" as a keyword for data crawling and ignore similar terms, such as "data analyst" and "data science engineer". However, two limitations are set by LinkedIn and Monster. 1) *Speed limitation*: We can send only a limited number of requests to gather information within a given period of time. For example, LinkedIn limits the profile viewing times for normal members. Once the given search limit is reached, the profile viewing function is disabled until the next calendar month. In addition, the frequency of profile viewing is also restricted within a certain limit, i.e. no additional viewing is permitted within three seconds since the beginning of the current viewing. 2) *Amount limitation*: Moreover, given a query request, we can view only a limited number of pages, regardless of the total number of pages that match. For instance, LinkedIn shows no more than one thousand profiles for each specific search.

To address these challenges and guarantee the completeness of the data, we design two corresponding strategies: 1) *Parallel crawler*: To accelerate the crawling speed, we upgrade our accounts to Recruiter Lite to acquire more information within a specific period of time. In addition, based on our previous work, we devise a parallel crawler system and utilize multiple servers to execute data crawling tasks cooperatively. 2) *Layered division*: Moreover, when the number of requested pages exceeds the limitation (e.g., 1000), we split the original query into multiple queries via layered division.

Taking data crawling for "data scientist" in the U.S. as an example (as shown in Figure 11), 14,481[7] profiles were located on LinkedIn by searching for "data scientist" in October 2016; however, fewer than 6% of these profiles are shown in the profile list for the search. Except for the keyword "data scientist", the first layer specifies location, i.e., the greater New York City area, the San Francisco Bay area, the Greater Chicago area, and the Washington, D.C. metropolitan area. Inevitably, some locations return more than one thousand profiles, i.e., 3566 data scientist profiles were found in the San Francisco Bay area. Thus, we further divide the search by the second layer, years of employment, including less than one year, 1-2 years, 3-5 years, 6-10 years and more than 10 years. Again, taking the San Francisco Bay area as an example, while there are 284 data scientists with no more than two years of employment, there are 1461 data scientists with 6-10 years of employment. Thus, we move on to the third layer; i.e., seniority, which is a LinkedIn-defined term with the labels entry, senior, manager, director and so on. For the San Francisco Bay area and 6-10 years of employment, there are 969 entry-level data scientists and 492 data scientists of other seniority levels. At this point, each specific search corresponds to fewer than one thousand data scientists, and there is no need to further divide the search. In this way, a coarse-grained query is split into multiple fine-grained queries with more syntax specifications. The division satisfies two requirements: 1) each fine-grained query corresponds to fewer than one thousand data scientists; and b) the profiles returned by different queries do not overlap. In our example, the three layers are location, years of employment and seniority.

[7]This number increases over time. Unless stated otherwise, the following numbers are based on data collected in October 2016.

### C. DESCRIPTIVE DEFINITION

Researchers, experts and even tech bloggers provide various definitions of the term "data scientist". Most of the definitions are summarized by a specific person or group at an abstract level. These definitions provide people with an initial understanding of the work of a data scientist and attract increasing amounts of attention from various areas. In fact, the majority of the definitions are derived from industrial sources because the position of data scientist is usually considered to be popular in industry, although a few of these definitions are provided by academic researchers. Here, we classify the existing definitions into three categories in terms of their focuses, including composite metaphors, responsibilities, and skill sets, and describe them one by one.

- *Composite Metaphor:* To enable rapid understanding of the work of a data scientist, several definitions use well-known positions or knowledge to produce composite metaphors. For instance, Anjul Bhambhri [4], the vice president of big data products at IBM, uses a high-level metaphor in describing a data scientist as a Renaissance individual: "*A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who truly wants to learn and bring change to an organization.*" He believes "*A data scientist is part digital trendspotter and part storyteller stitching various pieces of information together.*" Patil and Davenport [1] describe a data scientist as a hybrid of a data hacker, an analyst, a communicator and a trusted adviser and stated that a data scientist should be able to extract insights from the current data tsunami. Likewise, Monica Rogati [23], a senior data scientist at LinkedIn, says that a data scientist is "*half hacker, half analyst, [and] they use data to build products and find insights.*" She says that a data scientist is "*Columbus meet Columbo – starry eyed explorers and skeptical detectives.*"
- *Responsibility*: The most direct way to define the work of a data scientist is to specify their responsibilities. One typical definition comes from Baiju NT [24], who is the editor in chief at CMO and the driving force of *Big Data Made Simple*. He says that a "*data scientist is a person who has the knowledge and skills to conduct sophisticated and systematic analyses of data. A data scientist extracts insights from datasets for product development and evaluates and identifies strategic opportunities.*" In addition, Daniel Tunkelang, a principal data scientist at LinkedIn, describes a data scientist as "*someone who can obtain, scrub, explore, model and interpret data,*

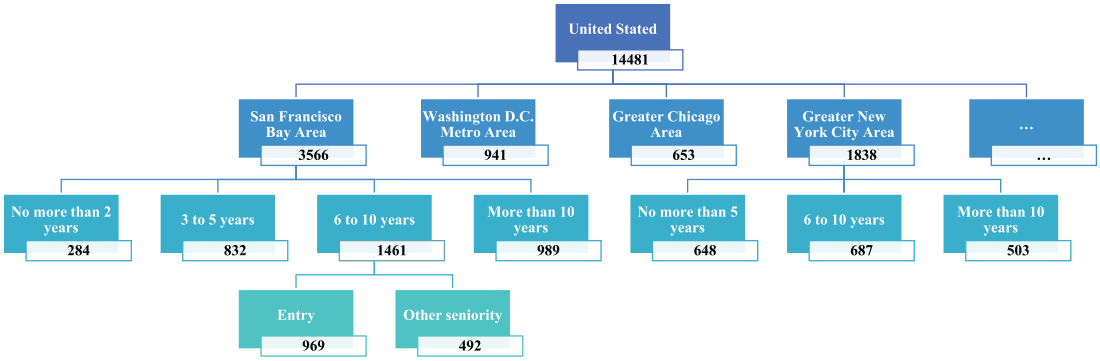



**FIGURE 11.** Three-layer scheme for U.S., where the layers are marked with different colors. The number attached to each block refers to the total number of matched data scientist profiles of the corresponding search. The first layer specifies the location of the top 100 metropolitan areas in U.S. Locations with over one thousand matched profiles, i.e. San Francisco Bay Area, fans out the second layer on years of employment. Again, the second layer searches with over one thousand matched profiles introduces the third layer, which specifies the seniority.

*blending hacking, statistics and machine learning. Data scientists not only are adept at working with data but also appreciate data itself as a first-class product.*'' [24]. In [22], Van der Aalst also describes data scientists as people who both ''*should be creative and able to realize solutions using IT*'' techniques and are ''*able to convey the message well*'' based on domain knowledge. Other similar definitions can also be found in [25]–[29].

- *Skill*: In addition to their responsibilities, many researchers or experts prefer to define data scientists as having certain specific skills. For instance, Dhar [5] believes that ''*A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.*'' Similarly, Hilary Mason [30], the founder and CEO of Fast Forward Labs, prefers to define a data scientist as someone ''*who blends, math, algorithms, and an understanding of human behavior with the ability to hack systems together to get answers to interesting human questions from data.*'' Additionally, Steve Hillion [30], the vice president of analytics at EMC Greenplum, believes data scientists should be ''analytically minded, statistically and mathematically sophisticated data engineers'.

IBM researchers [31] provide a comprehensive definition: ''*A data scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems, they will pick the right problems that have the most value to the organization.*''

These descriptions present a macro-scale definition of the term ''data scientist'' from three distinct perspectives. Their common element is that a data scientist is someone who analyzes data. These descriptions present an intuitive understanding for consumption by the general public or managers of what a data scientist is. However, a more elaborate and quantitative definition is clearly lacking. In particular, we may not know what the differences between data scientists and data analysts are or how to write a specific recruitment advertisement for data scientist positions.

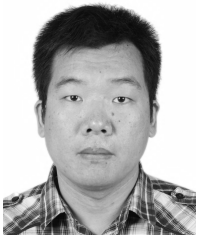
## ACKNOWLEDGMENT

H. Hu was a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

## REFERENCES

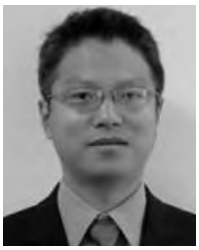
[1] T. H. Davenport and D. J. Patil, ''Data scientist: The sexiest job of the 21st century,'' *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 70–76, Oct. 2012.
[2] *Hiretual*. Accessed: Oct. 20, 2017. [Online]. Available: https://hiretual.com
[3] *Why So Many Fake Data Scientists?* Accessed: Oct. 20, 2017. [Online]. Available: https://www.linkedin.com/pulse/why-so-many-fake-data-scientist-bernard-marr
[4] *Being a Data Scientis*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.linkedin.com/pulse/being-data-scientist-carla-gentry
[5] V. Dhar, ''Data science and prediction,'' *Commun. ACM*, vol. 56, no. 12, pp. 64–73, 2013.
[6] *Monster*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.monster.com
[7] *Linkedin*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.linkedin.com/feed
[8] *Which Job Boards are Most Useful for Jobseekers?* Accessed: Oct. 12, 2018. [Online]. Available: https://theundercoverrecruiter.com/job-boards-useful-jobseekers/
[9] H. Hu, Y. Wen, Y. Gao, T.-S. Chua, and X. Li, ''Toward an SDN-enabled big data platform for social TV analytics,'' *IEEE Netw.*, vol. 29, no. 5, pp. 43–49, Sep./Oct. 2015.
[10] A. Archambault and J. Grudin, ''A longitudinal study of Facebook, LinkedIn, & Twitter use,'' in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 2741–2750.
[11] I. Paparrizos, B. B. Cambazoglu, and A. Gionis, ''Machine learned job recommendation,'' in *Proc. 5th ACM Conf. Recommender Syst.*, Oct. 2011, pp. 325–328.
[12] C. Manning, *Foundations of Statistical Natural Language Processing*, vol. 999. Cambridge, MA, USA: MIT Press, 1999.
[13] *Qs World University Rankings by Subject*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.topuniversities.com/subject-rankings/2017
[14] *Linkedin Skill Set List*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.linkedin.com/directory/topics-a

[15] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Hoboken, NJ, USA: Wiley, 1997.
[16] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[17] J. S. Farris, "On the cophenetic correlation coefficient," *Systematic Zoology*, vol. 18, no. 3, pp. 279–285, 1969.
[18] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.
[19] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, Tech. Rep., 2006.
[20] D. Lim and G. Lanckriet, "Efficient learning of Mahalanobis metrics for ranking," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1980–1988.
[21] *Linkedin Skill Set List*. Accessed: Oct. 20, 2017. [Online]. Available: http://www.go-globe.com/blog/60-seconds
[22] W. M. P. Van der Aalst, "Data scientist: The engineer of the future," in *Enterprise Interoperability VI*. Cham, Switzerland: Springer, 2014, pp. 13–26.
[23] *LinkedIn's Monica Rogati On 'What Is A Data Scientist?'*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.forbes.com/sites/danwoods/2011/11/27/linkedinsmonicarogationwhatisadata-scientist/#5976d4603e15.
[24] *What is a Data Scientist? 14 Definitions of a Data Scientist!* Accessed: Oct. 20, 2017. [Online]. Available: http://bigdatamadesimple.com/whatis-adatascientist14definitionsofadatascientist
[25] *Amazon's John Rauser on 'What Is a Data Scientist?'*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.forbes.com/sites/danwoods/2011/10/07/amazons-john-rauser-on-what-is-a-data-scientist/#26f56e4b108e.
[26] *Tableau Software's Pat Hanrahan on 'What Is a Data Scientist?'*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.forbes.com/sites/danwoods/2011/11/30/tableausoftwarespathanrahanonwhatisadata-scientist/#7097b6bb5eb1.
[27] *Building Data Science Teams*. Accessed: Oct. 20, 2017. [Online]. Available: http://radar.oreilly.com/2011/09/buildingdatascienceteams.html
[28] *Data Scientists: The Definition of Sexy*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/#140b7e8e5f96
[29] *What is a Data Scientist?* Accessed: Oct. 20, 2017. [Online]. Available: https://www.theguardian.com/news/datablog/2012/mar/02/data-scientist
[30] *Who is a Data Scientist? 14 Statements to Understand*. Accessed: Oct. 20, 2017. [Online]. Available: http://www.dailytenminutes.com/2017/05/who-is-data-scientist-14-statements-to.html
[31] *Data Scientists*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.ibm.com/analytics/us/en/technology/clouddataservices/data-scientist



**HAN HU** received the B.S. and Ph.D. degrees from the University of Science and Technology of China in 2007 and 2012, respectively. He joined the National University of Singapore and Nanyang Technological University, Singapore, as a Research Fellow, in 2012 and 2014, respectively. He is currently an Associate Professor with the School of Information and Electronics, Beijing Institute of Technology, China. He has published over 40 papers in top journals (e.g., IEEE JSAC, IEEE TCSVT, and IEEE TMM) and prestigious conferences (e.g., INFOCOM and ACM MM). His research interests include multimedia networking and data analytics. He received three best paper awards, including the 2013 IEEE Globecom Best Paper Award, the 2015 IEEE Multimedia Best Paper Award, and the 2015 Chinacom Best Paper Award. His work on multimedia networking and data center networking have been awarded the 2013 ASEAN ICT Awards (Gold Medal) and the 2015 Datacentre Dynamics Awards.

**YONG LUO** received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2009, and the D.Sc. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2014. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University. He has authored several scientific articles at top venues including IEEE T-PAMI, T-NNLS, IEEE T-IP, IEEE T-KDE, IEEE T-MM, IJCAI, and AAAI. His research interests are primarily on machine learning and data mining with applications to visual information understanding and analysis. He received the IEEE GLOBECOM 2016 Best Paper Award, and was nominated as the IJCAI 2017 Distinguished Best Paper Award.



**YONGGANG WEN** (S'99–M'08–SM'14) received the Ph.D. degree in electrical engineering and computer science (minor in western literature) from the Massachusetts Institute of Technology, Cambridge, USA, in 2007. He was with Cisco to lead product development in content delivery network, which had a revenue impact of three Billion U.S. dollars globally. He has worked extensively in learning-based system prototyping and performance optimization for large-scale networked computer systems. He is currently an Associate Professor and the Director of the Innovation Lab, School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore. He also serves as the Acting Director of the Nanyang Technopreneurship Centre, NTU. His research interests include cloud computing, green data center, distributed machine learning, big data analytics, multimedia network, and mobile computing. His work in multi-screen cloud social TV has been featured by global media (over 1600 news articles from over 29 countries) and received 2013 ASEAN ICT Awards (Gold Medal). His work on Cloud3DView, as the only academia entry, has received the 2016 ASEAN ICT Awards (Gold Medal) and the 2015 Datacentre Dynamics Awards 2015 C APAC (Oscar award of data centre industry). He was a co-recipient of the 2015 IEEE Multimedia Best Paper Award, and a co-recipient of Best Paper Awards at 2012 IEEE EUC, 2013 IEEE GLOBECOM, 2014 IEEE WCSP, 2015 EAI/ICST Chinacom, 2016 IEEE GLOBECOM, and 2016 IEEE Infocom MuSIC Workshop. He received the 2016 IEEE ComSoc MMTC Distinguished Leadership Award. He was elected as the Chair for the IEEE ComSoc Multimedia Communication Technical Committee (2014–2016). He serves on editorial boards for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *IEEE Wireless Communication Magazine*, the IEEE COMMUNICATIONS SURVEY AND TUTORIALS, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, the IEEE ACCESS and Elsevier *Ad Hoc Networks*.

**YEW-SOON ONG** received the B.S. and M.Eng. degrees in electrical and electronics engineering from Nanyang Technological University (NTU), Singapore, in 1998 and 1999, respectively, and the Ph.D. degree in artificial intelligence in complex design from the Computational Engineering and Design Center, University of Southampton, Southampton, U.K., in 2003.

He is currently an Associate Professor and the Director of the A*Star SIMTECH-NTU Joint Laboratory on Complex Systems and Programme, School of Computer Engineering, NTU. He is also a Principal Investigator of the Rolls-Royce@NTU Corporate Laboratory on Large Scale Data Analytics, Singapore. His research interests include computational intelligence spans across memetic computation, evolutionary design, machine learning, and big data. His research work on memetic algorithm was featured in the Emerging Research Fronts of the Essential Science Indicators in 2007.

Dr. Ong received the 2012 IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION Outstanding Paper Award and the 2015 *IEEE Computational Intelligence Magazine* Outstanding Paper Award for his published works on memetic computation. He is the Founding Technical Editor-in-Chief of *Memetic Computing* Journal; the Founding Chief Editor of *Studies in Adaptation, Learning, and Optimization* (Springer); and an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, the IEEE TRANSACTIONS ON NEURAL NETWORK AND LEARNING SYSTEMS, the *IEEE Computational Intelligence Magazine*, the IEEE TRANSACTIONS ON CYBERNETICS, and the IEEE TRANSACTIONS ON BIG DATA.

**XINWEN ZHANG** received the Ph.D. degree in information security from George Mason University, Fairfax, VA, USA. He was a Principal Engineer and the Senior Director of the Mobile Communication Lab, Secure Enterprise Group, Samsung Research America, Mountain View, CA, USA. He is currently the CTO of Hiretual, Mountain View, CA, USA. His research interests include security policies, models, architectures, and mechanisms in general computing and networking systems. He currently focuses on security enhanced Android platform and MDM systems for enterprise.