

Data Scientists in Software Teams: State of the Art and Challenges

Miryung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel

Abstract— The demand for analyzing large scale telemetry, machine, and quality data is rapidly increasing in software industry. Data scientists are becoming popular within software teams, e.g., Facebook, LinkedIn and Microsoft are creating a new career path for data scientists. In this paper, we present a large-scale survey with 793 professional data scientists at Microsoft to understand their educational background, problem topics that they work on, tool usages, and activities. We cluster these data scientists based on the time spent for various activities and identify 9 distinct clusters of data scientists, and their corresponding characteristics. We also discuss the challenges that they face and the best practices they share with other data scientists. Our study finds several trends about data scientists in the software engineering context at Microsoft, and should inform managers on how to leverage data science capability effectively within their teams.

Index Terms—data science, development roles, software engineering, industry



1 INTRODUCTION

Software teams increasingly analyze data to inform their engineering and business decisions and to build data solutions that deploy data in software products. The people behind the data gathering and analysis are called *data scientists*, a term coined by DJ Patil and Jeff Hammerbacher in 2008 to define their jobs at LinkedIn and Facebook [1]. The mission of a data scientist is to transform data into insights that guide the team's actions [2].

Recent studies document this emerging role of data scientists. Fisher et al. interviewed sixteen data analysts at Microsoft working with large datasets, with the goal of identifying pain points from a tooling perspective [3]. Kandel et al. conducted interviews with 35 enterprise analysts in healthcare, retail, marketing, and finance [4]. The study focuses on recurring pain points, challenges, and barriers for adopting visual analytics tools. In our own prior work, through interviews with sixteen data scientists at Microsoft, we identified five distinct working styles of data scientists and cataloged strategies for increasing the impact and actionability of their work [5]. However, all these studies are based on a relatively small number of data scientists, and therefore do not provide a broader perspective on data science work and how different types of data scientists differ in terms of educational background, tool usage, topics that they work on, and types of data that they work with.

This paper reports the findings of a comprehensive survey with 793 professional data scientists at Microsoft. The survey covers their skills, tool usage, challenges, and best

practices. The respondents include both people who work as a data scientist (38%), as well as those who do data science while working as software engineers (24%), program managers (18%), and other job roles (20%). Our research questions cover the following in the context of Microsoft data scientists:

- RQ1. What is the demographic and educational background of data scientists at Microsoft? (Section 3)
- RQ2. How do data scientists work? What tasks do they work on, how do they spend their time, and what tools do they use? (Sections 4 and 5)
- RQ3. What challenges do data scientists face and what are the best practices and advice to overcome those challenges? (Sections 6 and 7)
- RQ4. How do data scientists increase confidence about the correctness of their work? (Section 8)

Our study finds several trends about data science in the software development context. There is heavy emphasis on understanding customer and user behavior through automated telemetry instrumentation and live monitoring. Data science is also used as an introspective tool for assessing developer productivity and software quality.

In comparison to prior studies on data scientists, our study reveals a new category of data scientists, called *moonlighters* who are initially hired into non-data-science roles and but have incorporated data analysis as a part of their engineering work. Due to this transitional nature of adopting new responsibilities, many respondents emphasize the need of formal training and express the desire to have shared repositories for mentoring.

Data scientists in our survey spend a significant amount

M. Kim is with University of California, Los Angeles, 420 Westwood Plaza, 4532-H Boelter Hall, Los Angeles, 90095. E-mail: miryung@cs.ucla.edu.

T. Zimmermann, R. DeLine and A. Begel are with Microsoft Research, One Microsoft Way, Redmond, WA 98052. E-mails: tzimmer@microsoft.com, rdeline@microsoft.com, and abegel@microsoft.com.

of time querying data; building platforms for instrumentation; cleaning, merging, and shaping data; and analyzing data with statistics and machine learning. During these activities, poor data quality, missing or delayed data, and the mundane work of shaping data to fit the diverse suite of analytics tools become barriers. To overcome these challenges, data scientists recommend consolidating analytics tools and constructing data standards for instrumentation.

Despite the rising importance of data-based insights, our respondents find it difficult to increase confidence in the quality of their analysis work. To ensure the correctness of their work, more structured processes and tool support are needed for validating data science work, like peer logic review, dogfood simulation (using or creating their own data to test the software), live monitoring and debugging, and checking implicit constraints and data schema.

This paper provides a comprehensive description of the data scientist types as the roles emerge at a large company and a survey instrument that others can use to study data scientists.

2 METHODOLOGY

The findings in this paper are based on survey responses. The design of the questionnaire (Section 2.1) was informed by our previous interview study of sixteen data scientists [5] and existing studies on data scientists [6]. The survey was distributed to 2397 employees (Section 2.2). To analyze the data, we used a combination of statistical analysis and card sorting (Section 2.3).

2.1 Survey Design

Our questionnaire included questions about the following topics (the complete questionnaire is available as supplemental material and a technical report [7]):

- **Demographics:** We asked questions about organization, gender, having direct reports, job discipline, geographic location, experience in years (overall professional experiences, years at Microsoft, and years in analyzing data), and educational background.
- **Skills and self-perception:** We replicated questions from Harris et al. [6], who asked about (1) skills (“Please rank your skills” using a predefined list of skills) and (2) self-perceptions about professional categories (“I think of myself as a/an...” like scientist, engineer, business person, artist, etc.). Harris et al. used the answers to classify data scientists into four groups: *Data Businesspeople*, *Data Creatives*, *Data Engineers*, and *Data Researchers*.
- **Working styles:** Motivated by our previous study [5], we included a checkbox question to map respondents to one or more *working styles*. We also asked about the *tools* that respondents use and the *types of data* they analyze as part of their work. In an open-ended question, we asked for a concrete example of a data science task that respondents worked on in the past six months.
- **Time spent:** We asked respondents how much time they spend on certain activities. The list of activities was derived from existing work on data science workflows [8] [9] [10]. In addition to the activities, we asked how much time people spend in meetings. We also ask how much time they spend on activities *not* related to data science.
- **Challenges:** We asked an open-ended question about the challenges that respondents face: “What challenges do you frequently face when doing data science?”
- **Best practices:** To distill a set of best practices for data science, we asked two open-ended questions: “What advice related to data science would you give a friend who is looking to get started with data science?” and “What new features, tools, processes or best practices could improve how we do data science?”
- **Correctness:** To learn how data scientists ensure the quality of their work, we asked the two open-ended questions: “How do you ensure that the input data to your analysis is correct?” and “How do you ensure that you have high confidence about your analysis results?”

We followed a pilot protocol [11] to design the survey, i.e., an earlier version of the survey was sent to a small subset of the population (about 20) and their feedback was used to improve the survey and increase clarity of the questions. For example, an improvement made during the pilot was to ask participants about the “hours per week” instead of the “percentage of time” with respect to the data science activities. The responses of the pilot population were not included in the data analysis of the actual study.

2.2 Participant Selection

We sent the final survey to 2,397 Microsoft employees who we identified from two target populations:

- **Full-time data scientist employees** (population “*data science employees*”). By using the organizational database, we identified 599 Microsoft employees working in the new “Applied Science & Data” discipline or with “Data Sci” in the job title (Data Scientist, Data Science Lead, etc.).
- **Employees with interest in data science** (population: “*data science enthusiasts*”). We identified 1798 Microsoft employees who were subscribed to one or more large mailing lists related to data science or machine learning and did not belong to the “data science employees” population.

We included the population “data science enthusiasts” because in our previous interview study [5], we found there are hidden populations of data science practitioners. As one of the interview respondents pointed out, employees often transition to a full-time data science role from an engineering role (“a lot of people kind of moonlighted as data scientists besides their normal day job”, P15 in Kim et al. [5]). A significant portion (30%) of the survey respondents who

self-identified as part of the official data science discipline were from the “data science enthusiasts” population.

We invited respondents through a personalized email to participate in a survey on “Data Culture at Microsoft.” As an incentive to participate, respondents could enter a raffle of one of four \$125 Visa Gift Cards after completion of the survey. We received a total of 793 responses (response rate 33%), of which 216 responses were from the population of 599 “data science employees” (response rate 36%) and 577 from the population of 1798 “data science enthusiasts” (response rate 32%). This rate is comparable to the 14% to 20% [12] and 6% to 36% [13] that were reported for other surveys in software engineering.

In terms of demographics, 24% of survey respondents are female and 74% are male. Respondents vary in terms of geographic locations: North America (82%), Asia (9%), and Europe (7%). Of the respondents, 23% have direct reports (i.e., managers), while 75% do not. This indicates that a high number of our survey respondents are individual contributors without direct management responsibility.

2.3 Data Analysis

To analyze the responses, we used a combination of descriptive statistics, cluster analysis, statistical testing, and card sorting.

For the 532 responses to the **time spent** question, we first normalized the hours spent on each activity by computing the relative time spent on an activity per week (in percent). Next, we ran the *Partitioning Around Medoids* (PAM) clustering algorithm [14] using the *pamk* implementation from the *fpc* package in R and varying the number of clusters (k) from one to twenty. For each k , the algorithm computes a clustering and then returns the clustering with the optimum average silhouette width [15]. In our case, the optimum number of clusters was $k=9$. We discuss the results of the clustering in Section 5.

To further describe the clusters, we performed a series of **statistical tests** to identify how respondents in each cluster differ from the respondents outside the cluster in terms of demographics, skills, and tool usage.

For example, one of the nine clusters corresponds to the data scientists who “do it all,” i.e., spend significant time in most of the data science activities; we call it the *Polymath* cluster. We compared whether the presence of a PhD degree is significantly different between respondents in the *Polymath* cluster vs. respondents not in the *Polymath* cluster. We do similar comparisons for other demographics (experience, education, and role), skills (from the Harris et al. [6] survey, e.g., Algorithms, Machine Learning), and tool usage (R, Python, Scope, SQL, etc.). For binary variables (e.g., presence of a PhD degree, presence of a certain skill), we used *Fisher Exact Value* tests [16]. For numerical variables (e.g., years of professional experience), we used *Wilcoxon Mann Whitney* tests [16]. In this paper, we report differences that are significant at the level of $p < 0.05$.

This analysis was of exploratory nature. We wanted to identify any potentially interesting differences in the data. The statistical tests served as a filtering technique rather than a hypothesis testing technique. There is the possibility of making false discoveries due to multiple statistical tests. According to McDonald [17], there is “*no universally accepted approach for dealing with the problem of multiple comparisons.*” Any correction brings trade-offs between false and missed discoveries. Due to the exploratory nature of the analysis, we are more liberal with including discoveries. Any discovery that we report in this paper with respect to the cluster difference will need further validation on an independent context.

The questionnaire asked a few **open-ended questions**, covering (1) problem topics, (2) challenges, (3) best practices, (4) advice, (5) how to ensure input correctness, and (6) how to check for output correctness. To analyze the responses to these items, we used *card sorting* [18] [19], a technique that is widely used to create mental models and derive taxonomies from data. In our case, card sorting also helps us to deduce a higher level of abstraction and identify common themes. Our card sort was *open*, meaning we had no predefined groups; instead, we let the groups emerge and evolve during the sorting process. Each card sort was initially performed by one or two authors. When the card sort was performed by only one author, it was subsequently validated by another author. The resulting groups were then mapped back to the survey responses to check if the nine clusters responded differently to some of the open-ended questions. Results from the card sort analysis are discussed in Sections 4 and 6-8.

3 WHO ARE DATA SCIENTISTS AT MICROSOFT?

In this section, we characterize data scientists at Microsoft with respect to professional experience, skills, education, working style and time spent on work activities.

Professional experience. The respondents have 13.6 years of professional experience on average (median 12.6 years; respondents were allowed to report experience as a decimal number). They have worked at Microsoft 7.4 years on average (median 6 years). They have 9.8 years of experience in analyzing data (median 7.8 years).

Job title. Of the respondents, 38% identify as part of the data science discipline, 24% identify as software engineers, SDE, or engineering managers, 18% identify as program managers or PMs, and 20% identify with other disciplines.

Education. 34% have bachelor’s degrees, 41% have master’s degrees, and 22% have PhDs. Harris et al. [6] note that 70% of enterprise analysts in their study have at least a master’s degree and that post-graduate education provided hands-on experience working with real data to evaluate a theory or argue a position. Kim et al. [5] also note that PhD training contributes to the working style where

data scientists identify important questions on their own and iteratively refine questions and approaches.

Skill. To understand the skill sets that data scientists bring to their work, we asked the respondents to rank their skills in order. The list of 22 skills came from the Harris et al.'s survey [6]. Respondents dragged and dropped skills into an ordered list, with their introspectively identified top skill on top.

The list of skills ranges from Business, ML/Big Data, Math and Operations Research, Programming, and Statistics. To help the survey respondents to understand each skill category, several tools and keywords are mentioned with each skill set: for example, "Machine Learning (ex: decision trees, neural nets, SVM, clustering)" and "Unstructured Data (ex: noSQL, text mining)."

The results indicate that the respondents come with strong skill sets in Product Development, Business, and Back End Programming. Querying and manipulating structured data, data manipulation, and Big Data and distributed systems are most frequently reported skills. On the other hand, spatial statistics (geographic covariates, GIS), surveys and marketing (multinomial modeling), simulation (discrete, agent-based, and continuous simulations), Bayesian and Monte-Carlo statistics (MCMC) are less frequently reported skills.

Working styles. Each respondent checked statements that apply to their working styles. The given statements characterize five representative working styles that we identified in our prior work based on interviews with data scientists [5]. Participants could select multiple statements and the working styles are not mutually exclusive.

- 81% report that they analyze product and customer data;
- 76% communicate results and insights to business leaders (**Insight Provider**);
- 60% use big data cloud computing platforms to analyze large data;
- 51% build predictive models from the data (**Modeling Specialist**);
- 36% build data engineering platforms to collect and process a large quantity of data and use big data cloud computing platforms (**Platform Builders**);
- 31% add logging code or other forms of instrumentation to collect the data required for analysis (**Polymaths**);
- 12% manage a team of data scientists (**Team Leaders**).

The percentages add up to more than 100% because the respondents could check more than one statement. 48% of the respondents selected three or more statements.

Time spent. Figure 1 shows a boxplot of the time spent for each activity category. The thick vertical line in each box shows median hours per week that the respondents spend for each activity. We discuss more details on the clustering of respondents based on relative time spent in Section 5.

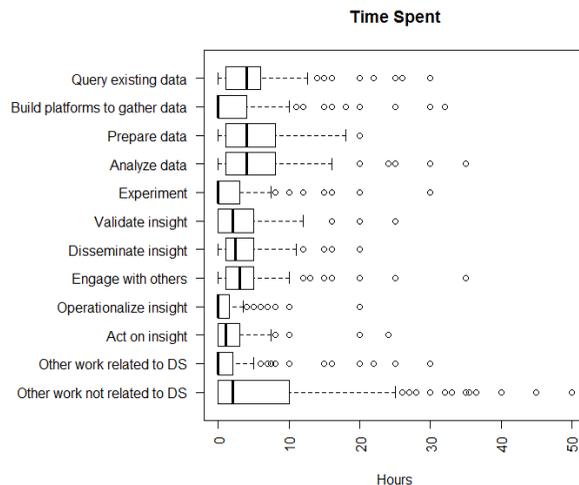


Figure 1. The time that respondents spent in each activity.

4 HOW DO DATA SCIENTISTS WORK?

We asked Microsoft data scientists about problem topics that they work on (Section 4.1) and tools and data used (Section 4.2).

4.1 What Problem Topics do Data Scientists Work on?

We categorized the problem topics that Microsoft data scientists work on into four topics using an open card sort: (1) user engagement; (2) software productivity and quality; (3) domain-specific problems; and (4) business intelligence. Below we describe the subcategories of each high-level category and illustrate them with quotes from respondents.

User Engagement. With usage data from software, teams pay attention to how customers use their software like which features are most often used. Based on telemetry logs, data scientists quantify user satisfaction, analyze complaints, and assess customer adoption and retention rates.

☛ *I work on understanding user activity and engagement impacts based on usage for new and existing users separately, using [...] user data collected from the [...] clients for each user. I built user journey from these data sources using modelling techniques. [P756]*

☛ *I used customer survey data to analyze for correlations / relationships with order size and frequency, [...] transaction cycle time, concession/discount. Using this data, I identified the relationship between customer satisfaction and concession/discount level to help decision makers determine the optimal discount level based on customer profile. [P698]*

Software Productivity and Quality. The respondents work on software teams; therefore, their own work leaves digital traces that can be analyzed. They assess engineering productivity and software quality through analysis of software artefacts. The types of tasks are often the topics stud-

ied by the Mining Software Repositories research community, like performance modelling, requirements elicitation, bug finding and repair, release planning, and system health and service quality monitoring.

☛ *Bug prioritization. Anomalous data activity indicating user failure. Reduce time to detection. Choose which build goes out to next set of customers [P36]*

☛ *I then used these visualizations to help product teams identify and prioritize major performance bugs in [our product]. [P60]*

☛ *I worked on prioritizing next release features based on customer feedback data. [P304]*

☛ *Can we determine the componentization of a system based on build data, which targets get built together and is it possible to determine componentization health using build data, check-ins, and branching tree data? [P577]*

☛ *I worked on calculating availability of our service. I use telemetry data collected from our service [P620]*

Domain-Specific Problems. The respondents work on product-specific data analysis topics or act as expert consultants on their clients' or customers' problems. These domain-specific problems include: assessing a speech-based natural language processing platform; addressing bottlenecks in image collection; creating predictive models for stock pricing; modeling advertiser value for ads; identifying malware; predicting power consumption; analyzing game user population; analyzing search behavior, and analyzing device churns across different hardware devices.

☛ *I worked on analyzing textual feedback and trying to figure out the relationship between the feedback and the ratings. [P17]*

Business Intelligence. With the background in finance modeling, many respondents work on predicting investment, demand, revenue, adoption, and growth of sales. These topics are what traditional business enterprise analysts work on.

☛ *I have a constant need to know data related to partner compensation, the number of customers that partners deal with, which customer has been compensated for, and how much revenue the partner is being attributed. This comes from multiple systems, and we typically have to use BI from several systems, exports and/or SQL queries from [product] systems and run comparison data from the [product] platform. [P237]*

Discussion. Based on the four categories, we conclude that data science work in software development teams is more than just business intelligence topics catalogued from Kandel's study on enterprise analysts [4]. The software productivity and quality topics are unique in that sense that data science is being used as an introspective tool to assess their own productivity and quality of software teams. This is an important trend to note, since the topics

studied by the Mining Software Repository research community are now being addressed by practitioners. However, we also note that topics are much more focused on user and customer behavior, e.g., user engagement, adoption, retention, migration, customer sentiment, and satisfaction. While academics do not generally have access to such large-scale user behavior data, data is available and analysis is more relevant in industry settings.

4.2 What Tools and Data Do Data Scientists Use?

Tool Usage. We asked the respondents to specify tools that they use for data science tasks. At Microsoft, SQL and Excel are popular (48% and 59%). Many data scientists use statistics tools and packages, like R (29%), MATLAB (5%), Minitab (4%), SPSS (3%), and JMP (2%). Respondents also rely on scripting and data manipulation tools, like Python (17%).

While we note that data scientists use small-scale, data analysis tools, like Excel (59%) and Office BI (25%), many are also using big data analytics platform, like Microsoft's map-reduce platform, called SCOPE (34%) and large scale machine learning libraries like Azure ML (15%) and TLC (9%). Since many respondents come from engineering roles, they are proficient with mainstream programming languages, like C, C++, and C# (33%).

Types of Data. In terms of data that they work with, 47% of the respondents analyze customer usage, (such as telemetry data, feature usage, game play data); 36% analyze business data, like purchases and transactions; 26% analyze execution behavior of the product (e.g., crashes, performance data, load balancing); 17% investigates engineering data of the product, like check-ins, work items, code reviews; and 17% use customer survey data.

5 THE TYPES OF DATA SCIENTISTS

Using the clustering method described in Section 2.3, we find nine distinct clusters for 532 responses of Microsoft data scientists. Table 1 shows each cluster with the percentage of their work hours and the average hours for each activity. Each row corresponds to a cluster and each column to an activity. The top number in each cell indicates the average relative time spent on an activity (in percent, used for clustering) and the bottom number indicates the absolute average number (in hours). For example, the 33 respondents in Cluster 4, **Data Shaper**, spend on average 10.9 hours in analyzing data on average, which is 25.7% of their work hours.

Throughout the discussion, we contrast the characteristics of respondents in each cluster against the rest of the respondents in terms of demographics, skills, and tool usage. Table 3 in the appendix reports only the characteristics that are significant with $p < 0.05$. For example, for the Cluster 1, **Polymath**, the statement

↑ PhD degree 31% vs. 19%

TABLE 1. THE NINE CLUSTERS OF DATA SCIENTISTS, BASED ON NORMALIZED TIME SPENT ON SURVEYED ACTIVITIES.

Each row corresponds to a cluster and each column to an activity. The top number in each cell corresponds to the percentage of time a person spends on an activity. The bottom number show to how many hours this time corresponds. For example, the 156 people in Cluster 1, “Polymath”, spend on average 10.4% of their time on querying existing data; this corresponds to 4.4 hours.

Entire population 532 people	12.0% 4.7h	7.2% 2.9h	11.7% 4.9h	12.5% 5.2h	4.8% 2.1h	6.9% 3.0h	8.5% 3.5h	9.2% 3.8h	2.4% 1.1h	5.5% 2.1h	4.1% 1.9h	15.1% 6.7h
Cluster 1 Polymath 156 people	10.4% 4.4h	8.5% 3.6h	11.5% 5.1h	15.1% 6.7h	9.1% 4.0h	7.7% 3.6h	7.4% 3.5h	7.9% 3.6h	3.2% 1.5h	5.2% 2.3h	4.0% 2.0h	10.1% 4.5h
Cluster 2 Data Evangelist 71 people	6.8% 2.2h	2.1% 1.0h	6.7% 2.5h	7.7% 2.9h	2.4% 1.2h	7.0% 2.6h	12.0% 4.5h	23.0% 8.6h	3.7% 1.3h	9.5% 3.3h	13.4% 6.0h	5.7% 2.6h
Cluster 3 Data Preparer 122 people	24.5% 9.4h	4.9% 1.9h	19.6% 7.8h	10.0% 4.0h	3.0% 1.3h	9.0% 4.1h	11.6% 4.5h	8.8% 3.5h	1.5% 0.7h	3.9% 1.3h	1.5% 0.7h	1.8% 0.8h
Cluster 4 Data Shaper 33 people	5.6% 2.5h	1.8% 0.7h	27.0% 11.5h	25.7% 10.9h	6.0% 2.6h	8.9% 3.8h	7.6% 3.3h	7.5% 3.2h	2.1% 1.0h	3.3% 1.4h	2.5% 1.1h	1.9% 0.9h
Cluster 5 Data Analyzer 24 people	9.9% 3.7h	0.9% 0.3h	5.8% 2.4h	49.1% 18.4h	4.6% 2.2h	6.6% 2.7h	5.2% 2.2h	5.8% 2.4h	1.8% 0.9h	4.2% 1.6h	2.8% 1.3h	3.2% 1.3h
Cluster 6 Platform Builder 27 people	12.5% 4.4h	48.5% 18.4h	6.1% 2.6h	4.3% 1.9h	3.8% 1.1h	2.7% 1.2h	4.4% 2.0h	4.1% 1.9h	2.1% 0.9h	3.0% 1.1h	1.4% 0.6h	6.9% 3.1h
Cluster 7 Moonlighter 50% 63 people	7.3% 3.1h	5.0% 2.2h	5.0% 2.1h	5.5% 2.4h	2.8% 1.2h	4.2% 2.0h	7.8% 3.3h	5.9% 2.4h	1.8% 0.8h	5.7% 2.3h	2.5% 1.1h	46.5% 20.0h
Cluster 8 Moonlighter 20% 32 people	2.9% 1.2h	1.4% 0.6h	1.9% 0.9h	1.6% 0.7h	0.4% 0.2h	1.5% 0.7h	1.7% 0.8h	2.3% 1.0h	0.6% 0.3h	2.1% 1.0h	2.9% 1.3h	80.9% 36.1h
Cluster 9 Insight Actor 4 people	0.9% 0.1h	2.1% 1.0h	1.8% 0.2h		0.9% 0.1h	5.7% 1.5h	18.5% 4.8h	10.1% 1.6h	3.0% 1.1h	57.1% 11.8h		
	Query existing data	Build platforms to gather data	Prepare data	Analyze data	Experiment	Validate insight	Disseminate insight	Engage with others	Operationalize insight	Act on insight	Other work related to DS	Other work not related to DS

means that 31% of the respondents who belonged to the Polymath cluster have a PhD degree, whereas only 19% of all other respondents have a PhD degree. As another example, the statement

↓ Years at Microsoft 6.3 yr vs. 7.5 yr

indicates that this cluster’s average years of experience at Microsoft is 6.3 years, whereas the rest of the respondents’ average years at Microsoft is 7.5 years.

Cluster 1: Polymath. This cluster is characterized by engaging in all kinds of activities, ranging from analyzing data, preparing data, querying data, and validating insights, etc. We name this cluster *polymaths* for consistency with our prior study [5], where we observed a working style of data scientists who “do it all,” e.g., forming a business goal, instrumenting a system to collect the required data, doing necessary analyses or experiments, and communicating the results to business leaders. 156 respond-

ents belong to this cluster of polymaths. 46% of them belong to the data science discipline, while 16% and 20% are program managers and software engineers, respectively. Polymaths have a relative high representation of people with PhDs (31% compared to 19%).

In terms of skill sets, this group of data scientists shows true versatility. They are more likely to have knowledge of machine learning (62% of polymaths vs. 47% of the rest), graphical models, spatial statistics and Bayesian statistics, and are also familiar with such scripting tools as Python and SCOPE (59% of polymaths vs. 44% of the rest). In terms of analysis topics, polymaths more frequently work on domain-specific problems than the rest (40% vs. 29%).

Cluster 2: Data Evangelist. This cluster is characterized by spending a significant amount of time with others about data and analysis (8.6 hours, 23.0% of their time), disseminating insights from data (4.5 hours, 12%), and acting on insight (3.3 hours, 9.5%). We name this cluster as “Data Evangelists,” because in our prior study, we observed data

scientists who push for the adoption of data-driven decision making within the organization or the company as “evangelists.” 71 respondents belong to this cluster. 38% belong to the data science discipline, 24% are program managers, and 8% are software engineers. Data evangelists are more likely to have longer experience at Microsoft (average 8.6 years) and the overall data analysis experience (average 11.9 years). They are more likely to be individual contributors (37% of data evangelists vs. 22% of the rest).

In terms of skill sets, data evangelists are more likely to have business and product development skills (65% of data evangelists vs. 38% of the rest), while less likely to have skills related to structured data (55% vs. 71%). Similarly, data evangelists are familiar with tools for business intelligence like Office BI (49% of data evangelists vs. 33% of the rest). In terms of analysis topics, data evangelists work more frequently on domain-specific problems given by clients and customers (24% vs. 10%).

Cluster 3: Data Preparer. This cluster is characterized by spending a significant amount of time in querying for existing data (9.4 hours, 24.5% of their time) and preparing data (7.8 hours, 19.6% of their time). 122 respondents belong to this cluster. 46% belong to the data science discipline, 14% and 19% are program managers and software engineers, respectively. Data preparers are less likely to be individual contributors (14% of data preparers vs. 26% of the rest). In terms of skill sets, data preparers have proficiency working with structured data (86% of data preparers vs. 63% of the rest), while slightly less likely to have expertise in algorithms (38% vs. 50%). Most data preparers are familiar with SQL (85% vs. 65). Data preparers more often mention the challenge of stitching together different streams of data (15% vs. 7%).

Cluster 4. Data Shaper. This cluster is characterized by spending a significant amount of time in analyzing data (10.9 hours, 25.7% of their time) and preparing data (11.5 hours, 27.0% of their time). 33 respondents belong to this cluster. Data shapers predominantly belong to the discipline of data science (79%), while only 3% and 9% are program managers and software engineers, respectively. They more likely have post-graduate degrees (54% with PhD and 88% with Master’s degree vs. 21% and 61% respectively for the rest) and skills in algorithms, machine learning, and optimization algorithms. They are less likely to have skills related to business, front end programming, and product development. In terms of tools, they indicate familiarity with MATLAB (30% vs. 5%), Python (48% vs. 22%), and the machine learning library TLC (35% vs. 11%). These skills are crucial to extracting and modeling relevant features from data. In terms of analysis topics, they work on search top query ranking problems (17% vs. 4%) and speech analysis (8% vs. 0%) more frequently than the rest.

Cluster 5. Data Analyzer. This cluster is characterized by spending about a half of their time in analyzing data (18.4 hour, 49.1% of their time). 24 respondents belong to

this cluster. 54% of data analyzers belong to the data science discipline, while 17% and 25% are program managers and software engineers, respectively. They are more likely to hold Master’s degrees than the rest (82% vs. 61%). Data analyzers are more likely to be familiar with statistics (76% vs. 47%), math (66% vs. 47%), Bayesian Monte Carlo statistics (42% vs. 18%), and data manipulation (82% vs. 54%) than the rest. They are less likely to have skills related to front-end programming and product development. Many data analyzers are R users (64% vs. 38%). They often mention handling and transforming data as a challenge.

Cluster 6. Platform Builder. This cluster is characterized by spending a significant amount of time in building platforms to instrument code to collect data (18.4 hours, 48.5% of their time). 27 respondents belong to this cluster. Platform builders are less likely to belong to the data science discipline (only 4%). 70% are software engineers, and 19% are program managers. They are more likely to have a background in big data and distributed systems (81% vs. 50%), and back-end (70% vs. 36%) and front-end programming (65% vs. 31%). In terms of tools, 89% use SQL and 70% are proficient with main stream languages like C, C++, and C#. In this group of data scientists, very few hold a job title as a data scientist (only 3.7%). They identify as software engineers who contribute to the development of a data engineering platform and pipeline. They frequently mention the challenge of data cleaning (15% vs. 2%).

Cluster 7. Fifty-percent Moonlighter. This cluster is characterized by spending about a half of their time in activities not related to data science (20 hours, 46.5% of their time). 63 respondents belong to this cluster. Only 10% of them belong to the data science discipline. 29% are program managers, and 35% are software engineers. In the open questions, they often mentioned that data science is not their day job. They have longer professional experience (16 years) and job experience at Microsoft (8.6 years). They are less likely to have a PhD degree (10% vs. 24%). While maintaining different engineering roles, they adopt data science work as a part of their responsibilities. In terms of skills and tools, they are less likely to be familiar with Bayesian Monte Carlo statistics, (8% vs. 21%), unstructured data (17% vs. 36%), Python (11% vs. 25%), the machine learning library TLC (3% vs. 13%), and Scope (33% vs. 50%).

Cluster 8. Twenty-percent Moonlighter. This cluster is characterized by engaging in mostly non-data-science work (36.1 hours, 80.9% of their time). 32 respondents belong to this cluster. Only 3% belong to the discipline of data science, 41% are program managers, and 43% are software engineers. Like the other moonlighter cluster, few have an job title as a data scientist, and only 6% have a PhD degree (vs. 23% for the rest). They have longer professional experience (17 years vs. 13.7 years) and job experience at Microsoft (10.8 years vs. 6.9 years). They are more likely to be skilled in product development (66% vs. 44%), while less familiar with R-like tools (15% vs. 42%). More people in

this group emphasize the importance of formal training (29% vs. 10%) and getting a mentor (23% vs. 1%) than the rest. In terms of problem topics, this group of moonlighters is more likely to work on the analysis of user and customer behavior (37% vs. 14%) and the analysis of bugs, crashes, and failures (36% vs. 9%). Assessing developer productivity and software quality is done more frequently by this cluster (37% vs. 14%).

Cluster 9. Insight Actors. This cluster is characterized by spending more than half of their time in acting based on the insights drawn from the data (57.1%) and disseminating insights from the data (18.5%). This cluster consists of only 4 respondents. Because of the small size, we excluded the cluster from any of the statistical analysis.

6 WHAT CHALLENGES DO DATA SCIENTISTS FACE?

When we asked data scientists at Microsoft about the challenges that respondents face, their answers fall into three categories: *data*, *analysis*, and *people*. The following subsections detail each category.

6.1 Challenges Related to Data

Data Quality. Respondents pointed to data quality as a challenge. They identified several reasons for having poor quality, due to bugs in the data collection and sampling procedure. Some respondents mentioned that there is an expectation that it is a data scientist's job to correct data quality issues, even though they are the main consumers of data. The data quality issue also makes it difficult for data scientists to have high confidence about the correctness of their work (see also Section 8 for how data scientists ensure the quality of their work).

- *Poor data quality. This combines with the expectation that as an analysis, this is your job to fix (or even your fault if it exists), and not that you are one of the main consumers of this poor-quality data. [P754]*

Data Availability. Respondents described a lack of data, missing data values, and delayed data as challenges. They mention either not having necessary data or having too much meaningless data to sift through. Data may not be available because of lack of instrumentation in legacy systems or an absence of data curation in the past.

- *Not enough data available from legacy systems. Adding instrumentation to legacy systems is often considered very expensive. [P304]*

Even when there is enough data, dealing with missing or incomplete data can be a challenge, e.g., missing samples. Another challenge is a very long time taken to receive live data. In particular, this challenge is mentioned by the cluster of Data Shapers more frequently than the rest (17% of Data Shapers vs. 4% of the rest.) Locating the right data

sources for analysis can be also a challenge. Some also mentioned the issue of data permission and access, since relevant data could be held by other teams.

- *Data has long delay, so it is hard to review live information. [P576]*

- *Finding the right data sources to use. Most streams aren't well documented and it's difficult to know if you should use them. [P233]*

Data Preparation. Another challenge is the integration of data from multiple sources and shaping of the data. Often data lives in different systems (i.e., "data islands") and must be combined to enable data analysis. One respondent called data integration "making sense of the spaghetti data stream." This concern of merging from spaghetti data is more frequently mentioned by Data Preparers than the rest (15% vs. 7%).

- *We have a lot of data from a lot of sources, it is very time consuming to be able to stitch them all together and figure out insights [P365]*

- *Data is created in silos, so our job is to find and make keys that connect disparate sources into patterns that help us learn and improve customer experience. [P367]*

During the data preparation phase, data scientists must understand what the data mean. Factors that complicate data understanding include lack of documentation, inconsistent schemas, and multiple possible interpretations of data labels. Figuring out the meaning of data requires talking to the people who collected the data. Several respondents emphasized that data are never clean and that they must account for bias.

- *Little documentation can be found, and usually not up to date, making it hard to ramp up with new dataset. [P235]*

Many respondents mentioned the challenge of data shaping and wrangling, i.e., shaping data into a right format to be consumed by various tools. The same challenge was noted in Kandel et al.'s study [4].

- *Getting data I need from whatever source and dealing with parsing, format manipulating and clean up. [P42]*

- *Massaging data into the right format to fit various tools. [P646]*

6.2 Challenges Related to Analysis

Scale. Because of the huge data size, batch processing jobs like Hadoop can take a while. Thus, iterative work flow can be expensive and quick visualization of large data sets is painful.

☛ *It takes too long to collect and analyze data due to long running time and sometimes long queue on Cosmos [P651]*

Despite a large suite of diverse tools, it is difficult for data scientists to access the right tools, because generic tools do not work for specific problems that they have.

☛ *Though we have lot of data science tools, there is no one tool that helps to solve most of the problems. [P182]*

Machine Learning. When it comes to building predictive models, respondents discussed the difficulty of knowing key tricks related to feature engineering and evaluating different models for optimal performance. For them, it is also difficult to infer the right signal from the data and what confidence level should be appropriate for the analysis. Respondents from the population “data science enthusiasts”, who do occasional data science work, mention that due to limited resources and restricted data access, it is difficult to define the scope of needed analysis.

☛ *There is no clear description of a problem, customers want to see magic coming out of their data. We work a lot on setting up the expectations in terms of prediction accuracy and in terms of time to develop end-to-end solutions. [P220]*

6.3 Challenges Related to People

Respondents mentioned the challenge of convincing their team of the value of data science and getting buy-in from the engineering team to collect high quality data. Respondents from the population “data science enthusiasts” also found it difficult to stay current with evolving tools, as they have other responsibilities and occasionally engage in data science work. Respondents mentioned that it is hard to convey the resulting insights to leaders and stakeholders in an effective manner.

☛ *Convincing teams that data science actually is helpful. Running behind people to get data. Helping to demystify data science. [P29]*

☛ *It's something that I don't do on a daily basis so my skills get rusty and need few hours to feel productive. [P161]*

☛ *Communicating to the team and getting all stakeholders on the same page. [P372]*

7 WHAT BEST PRACTICES CAN IMPROVE DATA SCIENCE?

We asked data scientists at Microsoft to share (1) *best practices* to overcome the challenge of data science work and (2) *advice* that they would give to novice data scientists. We combine the discussion of both questions because the responses were related to each other.

The responses falls into four categories: (1) formal training; (2) standardization; (3) clarifying the goals; and (4) understanding the caveats of data. The following subsections

describe the details for each category.

7.1 Learning and Practicing Data Science

The most frequently requested practice was the desire for training, through formal coursework, knowledge repositories, and mentoring. One popular piece of advice was to learn statistics—modelling and distributions. On the surface, this seems fundamentally obvious, but some respondents were very specific, suggesting to learn as many statistical and probabilistic models and distributions as one can for mastering the art of reasoning with statistics and dissecting one another's arguments. This phenomenon is interesting because the top suggestion for becoming a programmer is not likely to be “learn to program.”

Many respondents were not initially hired as data scientists and they have transitioned from other engineering roles. Therefore, there was strong emphasis on recognizing the need for formal training. Similarly, one popular suggestion was to take coursework. Respondents offered specific course suggestions, like MOOCs, courses from Udacity, Coursera, Code Academy, and PluralSight. Respondents suggested only online coursework, which made sense for adult learners in a full-time position, with little time for university degree programs.

Understanding machine learning is another popular suggestion, specifically to learn regression algorithms, classification algorithms, feature extraction, and feature generation, and to understand their assumptions and caveats.

Several respondents explicitly mentioned that data science is often only part of an employee's job and cuts across different job roles. They suggested that there should be a structured program of learning and certification that everyone should go through, since picking up a new discipline on the job is tough. Because of this emerging nature of the data science discipline in software teams, some respondents expressed the desire to avoid amateurism by professionalizing the practice.

☛ *Statistical rigor and peer review. [P459]*

☛ *Force creation of a null hypothesis into experimentation. [P633]*

☛ *More emphasis on data validation/testing models, establishing peer review of models as a routine practice. [P440]*

Respondents frequently mentioned the desire for hands-on training and practical case studies. Similarly, some respondents called for internal data science competitions to provide practical experience.

Respondents expressed the goal of fostering a *community of practice* [20] across the company and centralizing the learning resources in knowledge repositories.

☛ *1. There is no one uber wiki for data science, explaining various data science tools, processes and best practices. 2. Data science teams can host workshops or office hours weekly to increase awareness about various tools. [P259]*

☛ *Encourage development of practice communities around various data mining and analysis tool chains to make people more productive through reuse of tools, data, and code. [P311]*

7.2 Heterogeneous Tools and Data Standards

Respondents recommended learning the tools of the trade: R, LINQ, Excel, SQL, PowerBI, Python, Matlab, Azure ML, etc. The R tool was the most commonly suggested tool. Some suggested that data scientists need a balance between programming tools and data tools.

However, when it comes to this diversity of tools, respondents also expressed frustration about having too many tools and incompatible tools. Many responses called for integrating specific tools, drawn from the set of R, Python, F#, Excel, Cosmos (a map/reduce and distributed storage framework), Azure, SQL, AzureML, PowerBI, and the .NET framework. This reflects their frustrating workflows that are spread across many tools. Along this line, they complained about a proliferation of tools across the company. This heterogeneity makes it harder to reuse work across teams. Also, the company's engineering effort is spread across multiple competing tool efforts, rather than a single centralized effort. Further, the heterogeneity also creates analysis difficulties, particularly integrating temporal data from multiple systems.

☛ *There are already TOO MANY tools for doing data science. Invest in making a few easier to learn and use. R, Python, Unix-type commands, SCOPE, TLC/AzureML are good enough, adding more just takes too much time figuring out. [P461]*

☛ *Integrating / unifying separate tools/platforms [P780]*

This problem of heterogeneity does not just apply to tools but also extends to the data itself.

☛ *Manifest of available data sources across the org. Streamlined access/permission process for different data sources. Standardized nomenclature/types of similar data [P380]*

☛ *Centralize data and their definition. [P510]*

7.3 Clarifying the Goal of Data Science in Projects

Several respondents talked about the need to integrate data science early in the product life cycle, namely determining a project's success metrics in the early planning phases and then designing the remaining steps around these goals.

☛ *Do more to clearly define the business decision that has to be informed by data. Analysis, reporting, and modeling are not end-goals – they are tools to improve decisions that have monetary benefit. Once the problem is crisply defined in business decision terms, it focusses effort. [P710]*

☛ *Telemetry requirements should be part of specs. Success metrics should be defined before designing a product. [P304]*

This emphasis on specifying goals is also reflected this popular advice: choose the right business goals, needs, and

problems to have the most impact on your organization, and closely with the team with identifying good goals. Some respondents advised newcomers to become familiar with business objectives so that their work can be applied and explained to business decision makers. However, business decision makers do not always know what they are asking for, so it is important for data scientists to come up with their own conclusions and provide actual value to the recipients of the analyses.

Related to this point, respondents gave the advice that it is important to start from a specific problem and figure out what questions data scientists want answered before trying to analyze any of their data. Focusing first on tools and techniques can help you understand what can be done with your data, but there may be too many options to choose from due to the diversity of available and applicable techniques. Worse, if data scientists learn techniques before solving concrete problems, they will be tempted to use those techniques everywhere, even if the techniques are not going to help them solve the right problem.

7.4 Understand the Nuances of Data

“Just because the math is right, doesn't mean the answer is right,” said one of the survey respondents [P307]. *“When it comes to data, trust nothing,”* responded another [P59]. The importance of questioning the data itself, looking for unknowns, nulls, and blanks where the real data hides is another popular advice on the list.

Respondents spoke of getting their hands dirty and learning how to recognize biased or sketchy data. They worried that newcomers would presume the instrumentation code that gathered the data was correct, or that the data pipeline was lossless or eliminated noise already. They emphasized the importance of challenging one's own assumptions. Playing with small or toy data sets was reported to be helpful for coming up with hypotheses and gaining insights. Some suggested that one could develop an analysis on a sample data set and then analogically apply the same process to one's real data. Most respondents used words like ‘play’ and ‘fun,’ demonstrating their intrinsic interest in explorative analysis activities. At the same time, many respondents stressed that newcomers must focus on *real* data sets and *meaningful* problems.

☛ *It's hard to get a feel for data science if you are not working on something where you actually care about the results. [P47]*

Respondents said it is important to understand the data and how it was collected:

☛ *Interpreting [data] without knowing why it looks like it does will most likely lead you into a wrong direction. [P577]*

They warned against overreliance on aggregate metrics – to gain real insight, you must go one level deeper. More practically, if you can understand the context in which the data was collected and the rationale by which

the pipeline was designed, you can more easily detect the nuances crucial to proper understanding of the data.

8 HOW DO DATA SCIENTISTS ENSURE THE QUALITY OF THEIR WORK?

Despite the rising importance of data science work in the software industry and at Microsoft, respondents perceived that validation is a major challenge and the current validation methods can be improved.

☛ *“There is no empirical formula but we take a look at the input and review in a group to identify any discrepancies.” [P147]*

This problem of validation is frequently mentioned by Data Evangelists than the rest (21% of Data Evangelists vs. 8% of the rest). They emphasize that there is no perfect method for validation currently (24% of Data Evangelists vs. 11% of the rest). In the rest of this section, we discuss the validation technique that are currently used by data scientists.

Cross Validation is Multi-Dimensional. Developers use multiple data sources to triangulate their results and perform a held-out comparison with respect to different sources of data, other competition data sets, or benchmarked or curated data in the warehouse. Data scientists compare their analysis results against historical data or previously published data to see whether their new results fit within the boundary of historical norm and baseline.

☛ *“Cross reference between multiple independent sources and drill down on discrepancies, Drill down on interesting observations and patterns.” [P193]*

Another dimension of cross validation is having other team members or peers to validate their analysis, similar to peer code review. However, because data scientists often work as an external consultant for other organization’s problems, it is extremely important to have subject matter experts and stakeholders to be directly involved in the validation.

☛ *Code reviews and logic reviews with other team members. Presentations and requesting feedback on results. [P18]*

Data scientists also compare analysis results against human-labelled ground truth, and this need of having human labelled ground truth is mentioned more frequently by Data Shapers than the rest (16% vs. 4%).

☛ *We acquire and use evaluation data labeled by humans. I also compare against open data sets. [P758]*

Check Data Distribution. To build intuition about data, data scientists explore and understand the underlying distributions by computing descriptive statistics (e.g., histo-

grams), measuring confidence intervals, and measuring inter-rater agreement. Data scientists also build intuition by plotting data using visualization tools or spot checking. Polymaths more frequently discuss the importance of performing statistical tests and measuring inter-rater agreement than the rest (16% vs. 6%).

Dogfood Simulation. One unique aspect of software-oriented data science work is that, because input data is often collected from instrumented software, data scientists can create new data through simulation. This idea of scenario-based testing involves logging your own behavior, creating corresponding ground truth, and validating the results simulated through the same data collection and analysis pipeline.

☛ *“I will reproduce the cases or add some logs by myself and check if the result is correct after the demo.” [P384]*

Data scientists collect data through live, on-line monitoring and apply their analysis on these data. By leveraging this feedback loop of intentional data creation and analysis of collected data, data scientists test and cross-check their analysis. This idea of is more frequently mentioned by the Data Preparers than the rest (10% vs. 3%).

Type and Schema Checking. To ensure data quality and integrity, data scientists often check their format, type and schema to see whether individual fields are well-defined. Some even write scripts to verify a metadata and whether table columns are well defined. This type checking can help the data scientists to ensure that data are clean and not corrupted with malformed inputs.

Repeatability. To increase confidence in the correctness of results when processing and ingesting data, data scientists repeat the same procedure multiple times to replicate the same results. In other words, their analysis is often iterative by nature, and they rerun the same analysis on multiple data points.

Check Implicit Constraints. Data scientists often check implicit constraints, like assertions in software testing. Such constraints are not about single data points, but rather how the subgroups of data relate to other subgroups.

☛ *“If 20% of customers download from a particular source, but 80% of our license keys are activated from that channel, either we have a data glitch, or user behavior that we don’t understand and need to dig deeper to explain.” [P695]*

9 LIMITATIONS

Drawing general conclusions from empirical studies in software engineering is difficult because any process depends on a potentially large number of relevant context variables [21]. Since our survey was conducted within a single company, we cannot assume that the results will generalize outside of the company.

However, the survey respondents came from eight different organizations, each working on different kinds of products ranging from operating systems, databases, cloud software, software tools, to productivity software. There is nothing specific or different in the study that prevents replication at other companies or in the open source domain. Replicating our study in different organizational contexts will help generalize its results and build an empirical body of knowledge. To facilitate replication, our survey is available as a technical report [7].

We believe that the nature of data science work in this context is meaningful for others, given the scale of the company in terms of company size, project size, and range of products. Some of the challenges, best practices, and advice that we discussed in this paper might be less applicable to small companies [22], which deal with data on a smaller scale. For example, we expect that employees in small companies (or projects) will need to communicate with fewer people to understand the meaning of a piece of data. In many cases, the person who analyzes the data will be the same person who collected the data – similar to the *Polymath* role that we identified. We expect that small companies have more people with broad knowledge of data science, while large companies will benefit from having experts in a specific field of data science such as data platforms, data analysis, or prediction models.

The survey operated on a self-selection principle, which means that participation in the survey was voluntary. Results might be biased towards people who are more likely to answer the survey, such as employees with extra spare time. Avoiding the self-selection principle is impossible. For example, a sponsorship or an encouragement from the senior company leaders might increase participation, it would not have eliminated any potential bias. As pointed out by Singer and Vinson, the decision of responders to participate “could be unduly influenced by the perception of possible benefits or reprisals ensuing from the decision” [23]. We do not expect any systematic difference in the responses. Non-respondents will likely mention working styles, tasks, challenges, advice, and strategies for quality control that are similar to the data scientists that responded to our survey.

10 RELATED WORK

Data Science has become popular over the past few years as companies have recognized the value of data, either in data products, to optimize operations or to support decision making. Not only did Davenport and Patil [24] proclaim that data scientists would be “the sexiest job of the 21st century,” many authors have published data science books based on their own experiences (see books by O’Neill and Schutt [25], Foreman [26], or May [27]). Patil summarized strategies to hire and build effective data science teams based on his experience in building the data science team at LinkedIn [1].

10.1 Empirical Studies of Data Scientists

We found a small number of studies that focused on how data scientists work inside a company. Fisher et al. interviewed sixteen data analysts at Microsoft working with large datasets, with the goal of identifying pain points from a tooling perspective [3]. They uncovered tooling challenges like data integration, cost estimation problems for cloud computing, difficulties with shaping data to the computing platform, and the need for fast iteration on the analysis. The results on the challenges that data analytics developers face in ensuring correctness (Section 6) are complementary to their investigation on tooling pain points. However, their study is limited to only sixteen data scientists and does not provide large-scale, quantitative perspectives on tool usage.

Kandel et al. conducted interviews with 35 enterprise analysts in healthcare, retail, marketing, and finance [4]. Their study focuses on recurring pain points, challenges, and barriers for adopting visual analytics tools. They study general business intelligence analysts, as opposed to data scientists in software teams. Our study is done at a much larger scale, with 700+ data scientists, as opposed to 35 enterprise analysts. Regarding the challenges that data scientists face, both studies mention data quality issues, data availability issues, and data comprehension issues.

10.2 The Roles of Data Scientists

In a survey, Harris et al. asked 250+ data science practitioners how they viewed their skills, careers, and experiences with prospective employers [6]. Based on the respondents’ self-ranked skills and the extent that they self-identify with a variety of professional categories, the authors clustered the survey respondents into four roles: *Data Businesspeople*, *Data Creatives*, *Data Developers*, and *Data Researchers*. They also observed evidence for so-called “T-shaped” data scientists, who have a wide breadth of skills with depth in a single skill area.

While both our work and Harris et al. use a survey-based research method, the focus is different—our work focuses on data scientists in software development, while Harris et al. focus on general business intelligence analysts recruited from meet up groups. Harris et al.’s survey is also limited to two dimensions only: (1) skill sets and (2) the extent they agreed with various professional categories, e.g., “*I think of myself as an X.*” Our survey supplements Harris et al.’s survey with tool usage, challenges, best practices, and time spent on different activities. We also cluster data scientist based on time spent for each category of activities, as opposed to skill sets used in Harris et al. Therefore, our new clustering (Section 5) provides another complementary angle in classifying the population of data scientists. We discover a new category of data scientists, called “moonlighters” who have begun integrating data analysis as a part of their other engineering roles; this new category of moonlighters would not have been found unless the time spent for non-data science related activities is considered for clustering.

In our own prior work [5], we interviewed sixteen data

TABLE 2. COMPARISON OF THE DIFFERENT TYPES OF DATA SCIENTISTS.

	THIS PAPER	KIM ET AL. 2016 [5]	HARRIS ET AL. 2013 [6]
Generalists	Polymath	Polymath, “describes data scientists who ‘do it all’ ”	Data Creatives, “data scientists [who] can often tackle the entire soup-to-nuts analytics process on their own”
Specialists	Data Preparer		
	Data Shaper		
	Data Analyzer	Insight Provider, “main task is to generate insights and to support and guide their managers in decision making”	
	Platform Builder	Platform Builder, “build shared data platforms used across several product teams”	Data Developer, “people focused on the technical problem of managing data”
		Modelling Specialist, “data scientists who act as expert consultants and build predictive models”	Data Researcher, people with “deep academic training in the use of data to understand complex processes,”
Manager	Data Evangelist	Team Leader, “senior data scientists who run their own data science teams [...] act as data science ‘evangelists’ ”	Data Businesspeople, people who “are most focused on the organization and how data projects yield profit”
	Insight Actor		
Moonlighter	50% Moonlighter		
	20% Moonlighter		

scientists across several product groups at Microsoft and identified five working styles: *Insight Provider*, *Modeling Specialists*, *Platform Builder*, *Polymath*, and *Team Leader* as well as the corresponding strategies for increasing impact and actionability. This identification of five working styles was done qualitatively. In this paper, we cluster data scientists based on self-reported time spent for various activities, and we identify five additional groups: *Data Preparer*, *Data Shaper*, *Fifty-percent Moonlighter*, and *Twenty-percent Moonlighter* as well as the *Insight Actor*. In addition to characterizing data scientists in terms of time spent for various activities, we also contrast different clusters of data scientist in terms of problem topics, tool usage, challenges, etc. The challenges identified in our study could guide the development of new analytics tools that data scientists need.

Table 2 maps between the four types of data scientists identified by Harris et al. [6], the five working styles identified in Kim et al. [5] and the nine clusters of data scientists found from our large-scale survey. We can further group into four categories of data scientists: *generalists* with broad knowledge of data science, *specialists* who are experts in a specific field of data science (data preparation, data shaping, data analysis, prediction models, and data platforms), *managers*, who run data science teams and evangelize data science, and *moonlighters*, who have adopted data analysis work as a part of other job roles.

10.3 Software Analytics

Begel and Zimmermann conduct surveys on the questions

that software engineers would like data scientists to investigate about software and rate the resulting 145 questions in terms of importance [28]. The top 10 questions identified by Begel and Zimmermann like “*how do users typically use my application?*” and “*what parts of a software product are most used and loved by customers?*” are indeed the problem topics being worked on by data scientists in practice. Our results on problem topics (Section 4.1) indicate that customer behavior and user engagement analysis is one of the top five categories of problems that data scientists work on.

Software Analytics is a subfield of *analytics* with the focus on *software data*. Software data can take many forms like source code, changes, bug reports, code reviews, execution data, user feedback, and telemetry information. Software analytics has been the dedicated topic of tutorials and panels at the International Conference on Software Engineering [29, 30], as well as special issues of IEEE Software (July 2013 and September 2013). Zhang et al. [31] emphasized the trinity of software analytics in the form of three research topics (development process, system, and users) as well as three technology pillars (information visualization, analysis algorithms, and large-scale computing). Buse and Zimmermann argued for a dedicated data science role in software projects [32] and presented an empirical survey with software professionals on guidelines for analytics in software development [33]. None of this work has focused on the characterization of data scientists on software teams, which is one of the contributions of this paper.

11 CONCLUSION

For this paper, we conducted a survey with over seven hundred professional data scientists at Microsoft. Our survey had a comprehensive look at the educational background of data scientists, activities and time spent, tool usage, challenges that they face, and the best practices to overcome the challenges.

Our study finds that data scientist is a new emerging role in software teams – only 38% respondents are part of the data science discipline, and the rest were initially hired as other engineering roles and have taken the new responsibility of analyzing data as a part of their work. We named this new category of data scientist as *moonlighters*. Due to this transitional nature of their responsibility, many respondents stressed the importance of formal training, including coursework, shared knowledge repositories, and mentoring.

What makes data science unique in software development is that there is heavy emphasis on understanding customer and user behavior through automated instrumentation and monitoring. Another trend to note is that data science is being used as an introspective tool for assessing the organization's own productivity and software quality. We also note that many data scientists come with strong proficiency in mainstream programming languages like C, C++, and C# as well as big data analytics platforms like SCOPE, since the scale of data is so huge that the analytics work cannot be done using Excel or R-like tools alone. This emphasis on engineering scalability differs from traditional business enterprise analysts who rely on desktop analysis tools like Excel or Office BI.

Respondents spent a significant portion of their time on querying databases, building instrumentation platforms, manipulating data, and analyzing data with statistics and machine learning. During these activities, they face the challenges of poor data quality, missing or delayed data, or needing to shape the data to fit the diverse set of tools that they have to work with. To overcome these challenges, data scientists suggest consolidating heterogeneous tool suites and creating data standards for instrumentation.

Our study also finds that *validation* is a major challenge in data science and currently there are no good methods for ensuring correctness. For data scientists to increase confidence about the correctness of their work, there must be more structured tool support for peer review, cross validation, automated dogfood simulation, and checking implicit constraints and schema.

There are several research opportunities to further support data scientists.

- We observed diverse set of characteristics of data scientists with respect to activities, skill sets, and tool usage. We believe that the different types of data scientists on software teams have their own set of requirements for tool support. For example, a Moonlighter data scientist will need different tools than a Polymath.
- The heterogeneity of diverse tools and data standards makes it hard to reuse work across teams. It is necessary

to centralize data and to have standardized nomenclature and to develop *software processes* that account for the new role of data science in software projects.

- Tools to support *reuse* of data science work in software teams are an important research direction as well since many data scientists are transitioning from traditional engineering roles and they need formal training, shared knowledge repositories and mentoring.
- Validation is a major challenge in data science work and automated tool support is needed for cross-validation, debugging, and dogfood simulation. *Debugging* data-driven software is very challenging because it often involves tracking data across multiple steps in the data pipeline and talking to many people. Data scientists cannot assume that existing instrumentation code is correct or already collected data is clean. They are often the consumers of poor quality data. Tools supporting data scientists with identifying nuances or “sketchy” data can be very helpful to data scientists.

We hope that this paper will inspire research in these directions. With the rising number of data scientists, more research is needed to support the work of data scientists in software teams. To facilitate replication of this work, we provide the text of the survey as a technical report [7].

REFERENCES

- [1] D. Patil, *Building Data Science Teams*, O'Reilly, 2011.
- [2] T. H. Davenport, J. G. Harris and R. Morison, *Analytics at Work: Smarter Decisions, Better Results*, Harvard Business Review Press, 2010.
- [3] D. Fisher, R. DeLine, M. Czerwinski and S. M. Drucker, "Interactions with big data analytics," *Interactions*, vol. 19, no. 3, pp. 50-59, 2012.
- [4] S. Kandel, A. Paepcke, J. Hellerstein and J. Heer, "Enterprise Data Analysis and Visualization: An Interview Study," in *IEEE Visual Analytics Science & Technology (VAST)*, 2012.
- [5] M. Kim, T. Zimmermann, R. DeLine and A. Begel, "The Emerging Role of Data Scientists on Software Development Teams," in *ICSE'16: Proceedings of 38th IEEE/ACM International Conference on Software Engineering*, Austin, TX, USA, 2016.
- [6] H. D. Harris, S. P. Murphy and M. Vaisman, *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*, O'Reilly, 2013.
- [7] M. Kim, T. Zimmermann, R. DeLine and A. Begel, "Appendix to Everything You Wanted to Know About Data Scientists in Software Teams," Microsoft Research. Technical Report. MSR-TR-2016-1127. <https://www.microsoft.com/en-us/research/publication/appendix-everything-wanted-know-data-scientists/>, 2016.
- [8] P. Guo, "Data Science Workflow: Overview and Challenges," Blog@CACM, <http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext>, 2013.
- [9] B. A. Hamilton, "The Field Guide to Data Science, Second Edition," <http://www.boozallen.com/insights/2015/12/data-science-field-guide-second-edition>, 2015.
- [10] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>, 2011.
- [11] I. Brace, *Questionnaire design: How to plan, structure and write survey material for effective market research*, Kogan Page Publishers, 2013.
- [12] T. Punter, M. Ciolkowski, B. G. Freimut and I. John, "Conducting online surveys in software engineering," in *ISESE '03: Proceedings of*

- the International Symposium on Empirical Software Engineering*, 2003.
- [13] E. Smith, R. Loftin, E. Murphy-Hill, C. Bird and T. Zimmermann, "Improving Developer Participation Rates in Surveys," in *Cooperative and Human Aspects of Software Engineering*, 2013.
- [14] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, 1990.
- [15] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Computational and Applied Mathematics*, vol. 20, p. 53–65, 1987.
- [16] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer, 2010.
- [17] J. H. McDonald, *Handbook of biological statistics*, Baltimore, MD: Sparky House Publishing, 2009.
- [18] W. Hudson, "Card Sorting," in *Encyclopedia of Human-Computer Interaction*, 2nd Ed., M. Soegaard and R. F. Dam, Eds., The Interaction Design Foundation, 2013.
- [19] T. Zimmermann, "Card-sorting: From text to themes," in *Perspectives on Data Science for Software Engineering*, Morgan Kaufmann, 2016.
- [20] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, 1991.
- [21] V. Basili, F. Shull and F. Lanubile, "Building knowledge through families of experiments," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 456-473, July/August 1995.
- [22] R. Robbes, R. Vidal and M. C. Bastarrica, "Are Software Analytics Efforts Worthwhile for Small Companies? The Case of Amisoft," *IEEE Software*, vol. 30, no. 5, pp. 46-53, 2013.
- [23] J. Singer and N. G. Vinson, "Ethical Issues in Empirical Studies of Software Engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 12, pp. 1171-1180, 2002.
- [24] T. H. Davenport and D. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, pp. 70-76, October 2012.
- [25] C. O'Neil and R. Schutt, *Doing Data Science: Straight Talk from the Frontline*, O'Reilly Media, 2013.
- [26] J. W. Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Wiley, 2013.
- [27] T. May, *The New Know: Innovation Powered by Analytics*, Wiley, 2009.
- [28] A. Begel and T. Zimmermann, "Analyze This! 145 Questions for Data Scientists in Software Engineering," in *ICSE'14: Proceedings of the 36th International Conference on Software Engineering*, Hyderabad, India, 2014.
- [29] D. Zhang and T. Xie, "Software analytics: achievements and challenges," in *ICSE '13: Proceedings of the 2013 International Conference on Software Engineering*, 2013.
- [30] D. Zhang and T. Xie, "Software Analytics in Practice," in *ICSE '12: Proceedings of the International Conference on Software Engineering*, 2012.
- [31] D. Zhang, S. Han, Y. Dang, J.-G. Lou, H. Zhang and T. Xie, "Software Analytics in Practice," *IEEE Software*, vol. 30, no. 5, pp. 30-37, September 2013.
- [32] R. P. Buse and T. Zimmermann, "Analytics for software development," in *FOSER '10: Proceedings of the Workshop on Future of Software Engineering Research*, 2010.
- [33] R. P. Buse and T. Zimmermann, "Information needs for software development analytics," in *ICSE '12: Proceedings of 34th International Conference on Software Engineering*, 2012.



Miryung Kim is an associate professor in the Department of Computer Science at the University of California, Los Angeles. Her research focuses on software engineering, specifically on software evolution. She develops software analysis algorithms and development

tools to improve programmer productivity and her recent research focuses on software engineering support for big data systems and understanding data scientists in software development organizations. She received her B.S. in Computer Science from Korea Advanced Institute of Science and Technology in 2001 and her M.S. and Ph.D. in Computer Science and Engineering from the University of Washington under the supervision of Dr. David Notkin in 2003 and 2008 respectively. She received various awards including an NSF CAREER award, Google Faculty Research Award, and Okawa Foundation Research Award. Between January 2009 and August 2014, she was an assistant professor at the University of Texas at Austin before joining UCLA as an Associate Professor with tenure.



Thomas Zimmermann is a Senior Researcher in the Research in Software Engineering group at Microsoft Research, Redmond, USA. His research interests include software productivity, mining software repositories, software analytics, recommender systems, and

games research. He is best known for his research on systematic mining of software repositories to conduct empirical studies and to build tools to support developers and managers. His work received several awards, including Ten Year Most Influential Paper awards at ICSE'14 and MSR'14, '15, and '17, five ACM SIGSOFT Distinguished Paper Awards, and a CHI Honorable Mention. He is Co-Editor in Chief of the Empirical Software Engineering journal and serves on the editorial boards of several journals, including the IEEE Transactions on Software Engineering. He received his PhD in 2008 from Saarland University in Germany. His Twitter handle is @tomzimmermann and his homepage is <http://thomas-zimmermann.com>.



Robert DeLine, a Principal Researcher at Microsoft Research, has spent the last twenty-five years designing programming environments for a variety of audiences: end users making 3D environments (Alice); software architects composing systems (Unicon); professional programmers exploring unfamiliar code (Code Thumbnails, Code Canvas,

Debugger Canvas); and, most recently, data scientists analyzing streaming data (Tempe). He is a strong advocate of user-centered design and founded a research group applying that approach to software development tools. This approach aims for a virtuous cycle: conducting empirical studies to understand software development practices; inventing technologies that aim to improve those practices; and then deploying these technologies to test whether they actually do.



Andrew Begel is a Senior Researcher in the VIBE group at Microsoft Research. Andrew studies software engineers to understand how communication, collaboration and coordination behaviors impact their effectiveness in collocated and distributed development. He then

builds software tools that incentivize problem-mitigating behaviors. Andrew's recent work focuses on the use of biometrics to better understand how software developers do their work, and on autistic software engineers. He received his Ph.D. in Computer Science from the University of California, Berkeley in 2005. He currently serves as an Associate Editor for the Transactions on Software Engineering Journal.

TABLE 3. OVERVIEW OF THE DIFFERENCES BETWEEN THE NINE CLUSTERS.

DEMOGRAPHIC	SKILLS	TOOLS
<u>Polymath</u>		
↑ Population “Data Science Employees”: 37% vs. 24%	↑ Bayesian Monte Carlo Statistics: 26% vs. 17%	↑ Python: 33% vs. 20%
↑ PhD degree: 31% vs. 19%	↑ Big Distributed Data: 60% vs. 48%	↑ Scope: 59% vs. 44%
↓ Years at Microsoft: 6.3yr vs. 7.5yr	↓ Business: 35% vs. 45%	
	↑ Graphical Models: 24% vs. 15%	
	↑ Machine Learning: 62% vs. 47%	
	↑ Science: 46% vs. 35%	
	↑ Spatial Statistics: 13% vs. 8%	
<u>Data Evangelist</u>		
↑ Individual contributor: 37% vs. 22%	↑ Business: 65% vs. 38%	↓ SQL: 57% vs. 71%
↑ Year at Microsoft: 8.6yr vs. 6.9yr	↑ Product Development: 61% vs. 43%	↑ Office BI: 49% vs. 33%
↑ Year of data analysis: 11.9yr vs. 9.6yr	↓ Structured Data: 55% vs. 71%	
<u>Data Preparer</u>		
↓ Individual Contributor: 14% vs. 26%	↓ Algorithms: 38% vs. 50%	↑ SQL: 85% vs. 65%
↓ Bachelor’s Degree: 93% vs. 97%	↑ Structured Data: 86% vs. 63%	↑ Office BI: 45% vs. 33%
<u>Data Shaper</u>		
↑ PhD degree: 54% vs. 21%	↑ Algorithms: 71% vs. 46%	↑ MATLAB: 30% vs. 5%
↑ Master’s degree: 88% vs. 61%	↓ Business: 13% vs. 43%	↑ Python: 48% vs. 22%
↓ Years at Microsoft: 4.1yr vs. 7.3yr	↓ Front End Programming: 13% vs. 34%	↓ Excel: 57% vs. 84%
	↑ Machine Learning: 92% vs. 49%	↓ Office.BI: 9% vs. 37%
	↑ Optimization: 42% vs. 19%	↑ TLC: 35% vs. 11%
	↓ Product Development: 13% vs. 47%	
	↓ Structured Data: 46% vs. 69%	
<u>Data Analyzer</u>		
↑ Population “Data Science Employees”: 64% vs. 26%	↑ Bayesian Monte Carlo Statistics: 42% vs. 18%	↑ R: 64% vs. 38%
↑ Master's degree: 82% vs. 61%	↑ Classical Statistics: 76% vs. 47%	↓ Office.BI: 15% vs. 37%
↓ Professional experience: 8.4yr vs. 14.3yr	↑ Data Manipulation: 82% vs. 54%	
↓ Years at Microsoft: 3.7yr vs. 7.4yr	↓ Front End Programming: 12% vs. 34%	
	↑ Math: 66% vs. 47%	
	↓ Product Development: 27% vs. 46%	
<u>Platform Builder</u>		
↓ Population “Data Science Employees”: 4% vs. 29%	↑ Back End Programming: 70% vs. 36%	↑ SQL: 89% vs. 68%
↓ PhD degree: 0% vs. 23%	↑ Big and Distributed Data: 81% vs. 50%	↑ C/C++/C#: 70% vs. 45%
↓ Years of data analysis: 5.4yr vs. 10.2yr	↓ Classical Statistics: 30% vs. 50%	
	↑ Front End Programming: 63% vs. 31%	
<u>Fifty-percent Moonlighter</u>		
↓ Population “Data Science Employees”: 3% vs. 31%	↓ Bayesian Monte Carlo Statistics: 8% vs. 21%	↓ Python: 11% vs. 25%
↓ PhD degree: 10% vs. 24%	↓ Unstructured Data: 17% vs. 36%	↓ Scope: 33% vs. 50%
↑ Professional experience: 16yr vs. 13.6yr		↓ TLC: 3% vs. 13%
↑ Years at Microsoft: 8.6yr vs. 7.0yr		
<u>Twenty-percent Moonlighter</u>		
↓ Population “Data Science Employees”: 3% vs. 30%	↓ Data Manipulation: 34% vs. 57%	↓ R: 16% vs. 42%
↓ PhD degree: 6% vs. 23%	↑ Product Development: 66% vs. 44%	
↑ Professional experience: 17yr vs. 13.7yr	↓ Temporal Statistics: 16% vs. 35%	
↑ Years at Microsoft: 10.8yr vs. 6.9yr		