Data Scientists in Software Teams: State of Art and Challenges
[IEEE Transactions on Software Engineering, ICSE 2018 Journal First]

Miryung Kim
University of California, Los Angeles
Thomas Zimmermann, Rob DeLine, and Andrew Begel
Microsoft Research
Motivation: *The Emerging Roles of Data Scientists on Software Teams*

We are at a **tipping point** where there are large scale telemetry, machine, process and quality data.

Data scientists are emerging roles in SW teams due to an increasing demand for **experimenting with real users** and reporting results with statistical rigor.

We reported **the first in-depth interview study** with 16 data scientists in software teams [Kim et al. ICSE 2016].
Synopsis: *Data Scientists in Software Teams—State of Art and Challenges*

We conducted a comprehensive study of 793 professional data scientists at Microsoft.

We identified 9 distinct clusters and quantified their characteristics in terms of background, skill sets, activities, tool usage, challenges, and best practices.
Participant Demographic

Sent to 2397 employees
- 599 data science employees
  full time data scientists or the applied science & data discipline
- 1798 data enthusiasts
  subscribed to one or more lists on data science

793 responses (response rate 33%)

Job title. 38% data scientists, 24% software engineers, 18% program managers, and 20% others

Experience. 13.6 years on average (7.4 years at Microsoft)

Education. 34% bachelor’s degrees, 41% have master’s degree, and 22% have PhDs

Gender. 24% female, 74% male
Survey Design and Example Questions

Demographics

Skills and self-perception: “Please rank your skills.” “I think of myself as an ...”

Working style, Tools, Types of data, etc.

Problem topics: “Please give an example of a program related to data science that you worked in the last six months.”

Time spent: “Please enter roughly how many hours per week you typically spend on each of the activities.”

Challenges: “What challenges do you frequently face when doing data science?”

Best Practices: “What advice related to data science would you give to a colleague?”

Correctness: “How do you ensure that your analysis is correct?”
## Data Analysis Method

<table>
<thead>
<tr>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Card sorting for open-ended questions</strong></td>
<td><strong>Clustering (K-means)</strong> based on time spent on activities</td>
</tr>
<tr>
<td>Problem topics</td>
<td><strong>Statistical tests</strong> to identify how respondents in each cluster differ from the rest</td>
</tr>
<tr>
<td>Challenges</td>
<td></td>
</tr>
<tr>
<td>Best practices</td>
<td></td>
</tr>
<tr>
<td>Advice</td>
<td></td>
</tr>
<tr>
<td>How to ensure input correctness / output correctness</td>
<td></td>
</tr>
</tbody>
</table>
Time Spent on Activities

Hours spent on certain activities (self reported, survey, N=532)
Time Spent on Activities

Cluster analysis on relative time spent (k-means)

532 data scientists at Microsoft
9 Distinct Categories of Data Scientists based on Work Activities

Data Scientists in Software Teams: State of the Art and Challenges, Kim et al.
IEEE Transactions on Software Engineering
Category 1: Data Shaper

↑ PhD Degree: 54% vs. 21%
↑ Master’s Degree: 88% vs. 61%
↑ Algorithms: 71% vs. 46%
↑ Machine Learning: 92% vs. 49%
↑ Optimization: 42% vs. 19%

↓ Structured Data: 46% vs. 69%
↓ Front End Programming: 13% vs. 34%
↑ MATLAB: 30% vs. 5%
↑ Python: 48% vs. 22%
↑ TLC: 35% vs. 11%
↓ Excel: 57% vs. 84%
Category 2: Platform Builder

↑ Back End Programming: 70% vs. 36%
↑ Big and Distributed Data: 81% vs. 50%
↑ Front End Programming: 63% vs. 31%
↑ SQL: 89% vs. 68%

↑ C/C++/C#: 70% vs. 45%
↓ Classic Statistics: 30% vs. 50%
## Category 3: Data Evangelist

<table>
<thead>
<tr>
<th></th>
<th>Entire Population</th>
<th>Cluster 2 Data Evangelist (71 people)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Contributors</td>
<td>37% vs. 22%</td>
<td>6.8% vs. 2.1%</td>
</tr>
<tr>
<td>Years of Data Analysis</td>
<td>11.9 yr vs. 9.6 yr</td>
<td>6.7% vs. 2.5%</td>
</tr>
<tr>
<td>Product Development</td>
<td>61% vs. 43%</td>
<td>23.0% vs. 8.6%</td>
</tr>
<tr>
<td>Business</td>
<td>65% vs. 38%</td>
<td>9.5% vs. 3.3%</td>
</tr>
<tr>
<td>Structured Data</td>
<td>45% vs. 71%</td>
<td>45% vs. 71%</td>
</tr>
<tr>
<td>SQL</td>
<td>57% vs. 71%</td>
<td>57% vs. 71%</td>
</tr>
<tr>
<td>Office BI</td>
<td>49% vs. 33%</td>
<td>49% vs. 33%</td>
</tr>
</tbody>
</table>
## Category 4: Polymath

<table>
<thead>
<tr>
<th>Category</th>
<th>Entire Population 532 people</th>
<th>Cluster 1 Polymath 156 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>Build platforms to gather data</td>
<td>12.0% (4.7h)</td>
<td>10.4% (4.4h)</td>
</tr>
<tr>
<td>Prepare data</td>
<td>7.2% (2.9h)</td>
<td>8.5% (3.6h)</td>
</tr>
<tr>
<td>Query existing data</td>
<td>11.7% (4.9h)</td>
<td>11.5% (5.1h)</td>
</tr>
<tr>
<td>Experiment</td>
<td>12.5% (5.2h)</td>
<td>15.1% (6.7h)</td>
</tr>
<tr>
<td>Validate insight</td>
<td>4.8% (2.1h)</td>
<td>9.1% (4.0h)</td>
</tr>
<tr>
<td>Disseminate insight</td>
<td>6.9% (3.0h)</td>
<td>7.7% (3.6h)</td>
</tr>
<tr>
<td>Engage with others</td>
<td>8.5% (3.5h)</td>
<td>7.4% (3.5h)</td>
</tr>
<tr>
<td>Operationalize insight</td>
<td>9.2% (3.8h)</td>
<td>7.9% (3.6h)</td>
</tr>
<tr>
<td>Act on insight</td>
<td>2.4% (1.1h)</td>
<td>3.2% (1.5h)</td>
</tr>
<tr>
<td>Other work related to DS</td>
<td>5.5% (2.1h)</td>
<td>5.2% (2.3h)</td>
</tr>
<tr>
<td>Other work, not related to DS</td>
<td>4.1% (1.9h)</td>
<td>4.0% (2.0h)</td>
</tr>
<tr>
<td>Scope</td>
<td>15.1% (6.7h)</td>
<td>10.1% (4.5h)</td>
</tr>
</tbody>
</table>

↑ PhD Degree: 31% vs. 19%
↑ Big and Distributed Data: 60% vs. 48%
↓ Business: 35% vs. 45%
↑ Graphical Models: 24% vs. 15%

↑ Machine Learning: 62% vs. 47%
↑ Spatial Statistics: 13% vs. 8%
↑ Python: 33% vs. 20%
↑ Scope: 59% vs. 44%
Category 5: Moonlighter

- **Population: “Data Science Employees”**: 3% vs. 30%
- **Professional Experience**: 17yr vs. 13.75 yr
- **PhD degree**: 6% vs. 23%
- **Data Manipulation**: 34% vs. 57%
- **Product Development**: 66% vs. 44%
- **Temporal Statistics**: 16% vs. 35%
- **R**: 16% vs. 42%
Challenges that Data Scientists Face

Data

- Analysis
- People

DATA

Chart, magnifying glass, and silhouette figures.
Challenges Related to Data

Expected to Fix Incorrect Data

“Poor data quality. This combines with the expectation that as an analyst, this is your job to fix (or even your fault if it exists), not that you are the main consumer of this poor quality data.” [P754]

Lack of Data, Missing Values, and Delayed Data

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

Making Sense of the Spaghetti Data Stream

“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]
Challenges Related to Analysis

**Scale**

“Because of the huge data size, batch processing jobs like Hadoop make iterative work expensive and quick visualization of large data painful.” [P193]

**Difficulty of Knowing Key Tricks of Feature Engineering for ML**

“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.” [P220]
Challenges Related to People

**Convincing the Value of Data Science**
“Convincing teams that data science actually is helpful. Helping to demystify data science.” [P29]

**Buy-In from the Engineering Team to Collect High Quality Data**
“It is a lot of work to get engineering teams to collect high quality usage data (they depend heavily on system generated telemetry, rather than explicit usage logging).” [P594]
Ensuring Correctness
Challenges in Ensuring “Correctness”

Validation is a major challenge.

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

“Not possible most of the time... Intuition suffices most of the time.” [P27]
Success Strategies for Ensuring Correctness

Cross Validation and Peer Reviews

“Cross reference between multiple independent sources and drill down on discrepancies” [P193]

Dogfood Simulation

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

Check Implicit Constraint

“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]
Big Data Debugging in the Dark

Develop locally → Hope it works → Run in cloud → Bug!

Google Map Reduce

Hadoop

Spark

Hive

Debugging for Big Data Analytics in Spark
- Interactive Debugger [ICSE ’16]
- Automated Debugging [SoCC ‘17]
- Data Provenance [VLDB ‘16]

ACM Student Research Competition Poster: Muhammad Gulzar
Summary

Data scientist is a new emerging role in software teams.

In order to provide scientific, empirical understanding of data scientists, we clustered data scientists into sub-categories and quantified their characteristics.

Despite the rising importance of data-based insights, validation is a major challenge, motivating a new line of research on SE tools for increasing confidence in data science work.