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.

Data Shaper

Platform Builder

Polymath

Data Evangelist

Moonlighter

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

Qualitative

Card sorting for open-ended questions

Problem topics
Challenges
Best practices
Advice
How to ensure input correctness / output correctness

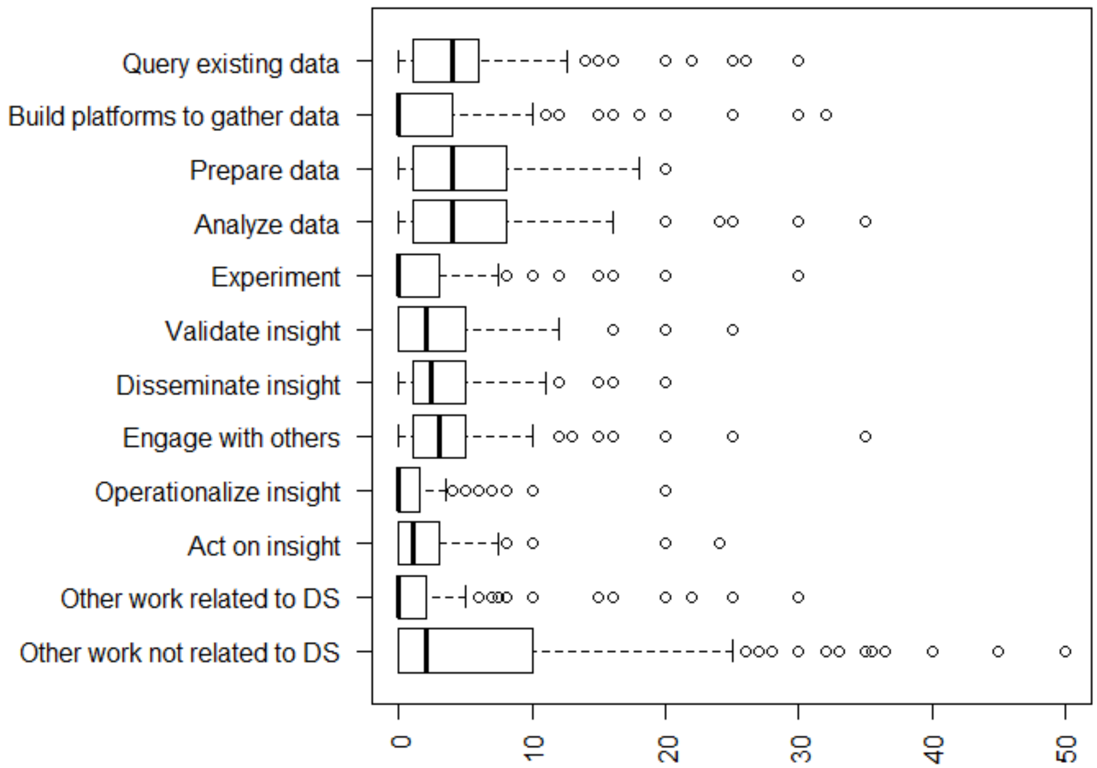
Quantitative

Clustering (K-means) based on time spent on activities

Statistical tests to identify how respondents in each cluster differ from the rest

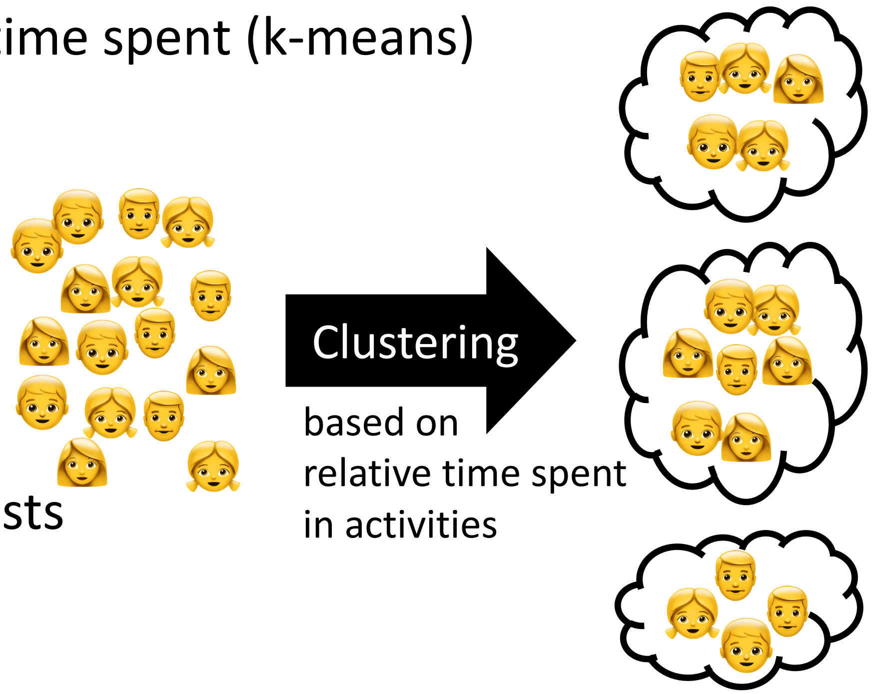
Time Spent on Activities

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

←Clusters

	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
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 10% 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 Act on Insight 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		

Activities →

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

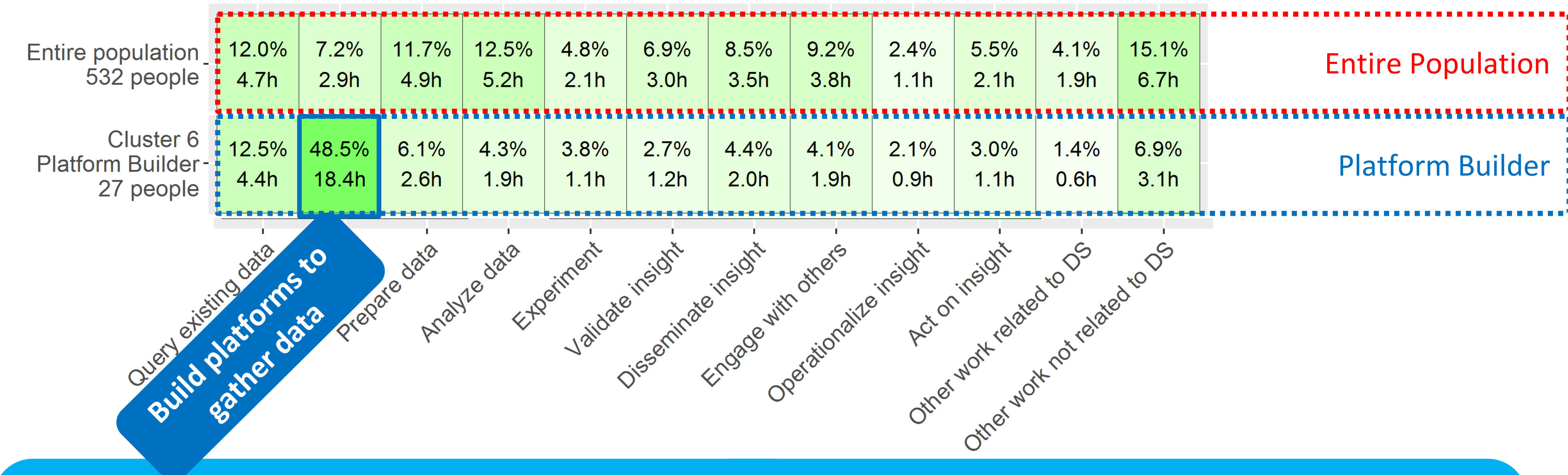
Category 1: Data Shaper

	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	
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	Entire Population
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	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



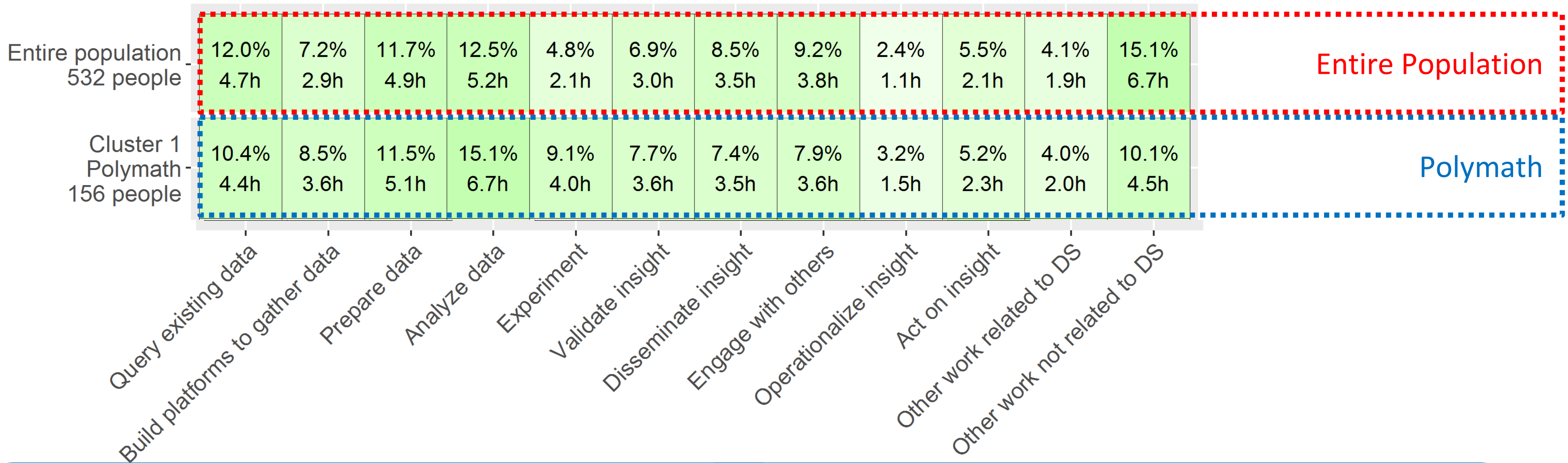
↑Back End Programming: 70% vs. 36%
↑Big and Distributed Data: 81% vs. 50%
↑Front End Programming: 63% vs. 31%
↑SQL: 89% vs. 68%
↑C/C++/C#: 70% vs. 45%
↓Classic Statistics: 30% vs. 50%

Category 3: Data Evangelist

	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	
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	Entire Population
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	Data Evangelist

- ↑Individual Contributors: 37% vs. 22%
- ↑Years of Data Analysis: 11.9 yr vs. 9.6 yr
- ↑Product Development: 61% vs. 43%
- ↑Business: 65% vs. 38%
- ↓Structured Data: 45% vs. 71%
- ↓SQL: 57% vs. 71%
- ↑Office BI: 49% vs. 33%

Category 4: Polymath



↑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

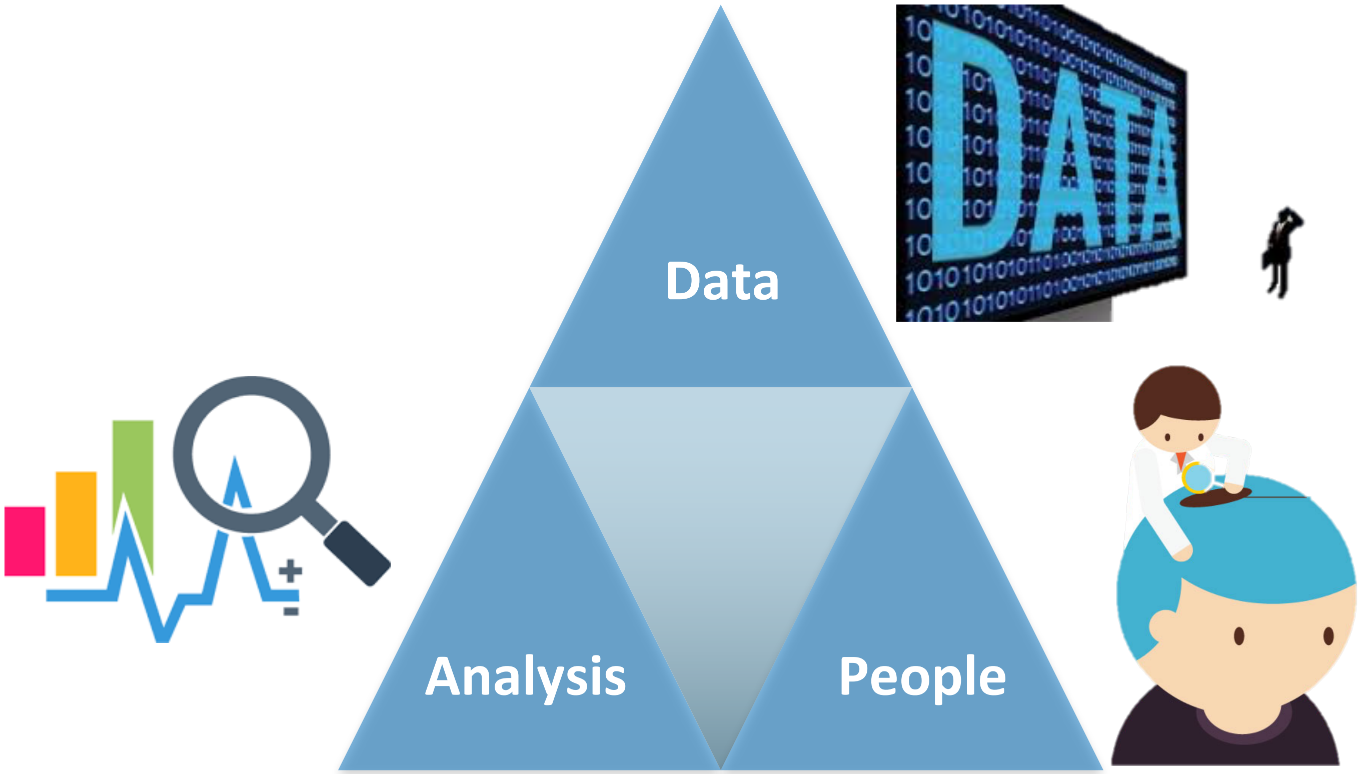
	Query existing data	Build platforms to gather data	Prepare data	Analyze data	Experiment	Validate insight	Disseminate insight	Engage with others	Operationalize insight	Act on insight	Other work related to DS	Other work not related to DS	
Entire population 532 people	12.0% 4.7h	7.2% 2.9h	11.7% 4.9h	12.5% 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	Entire Population
Cluster 8 Moonlighter 10% 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	Moonlighter

Other work not related to DS

↓ 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



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

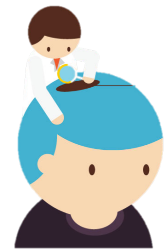
Scale

“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]

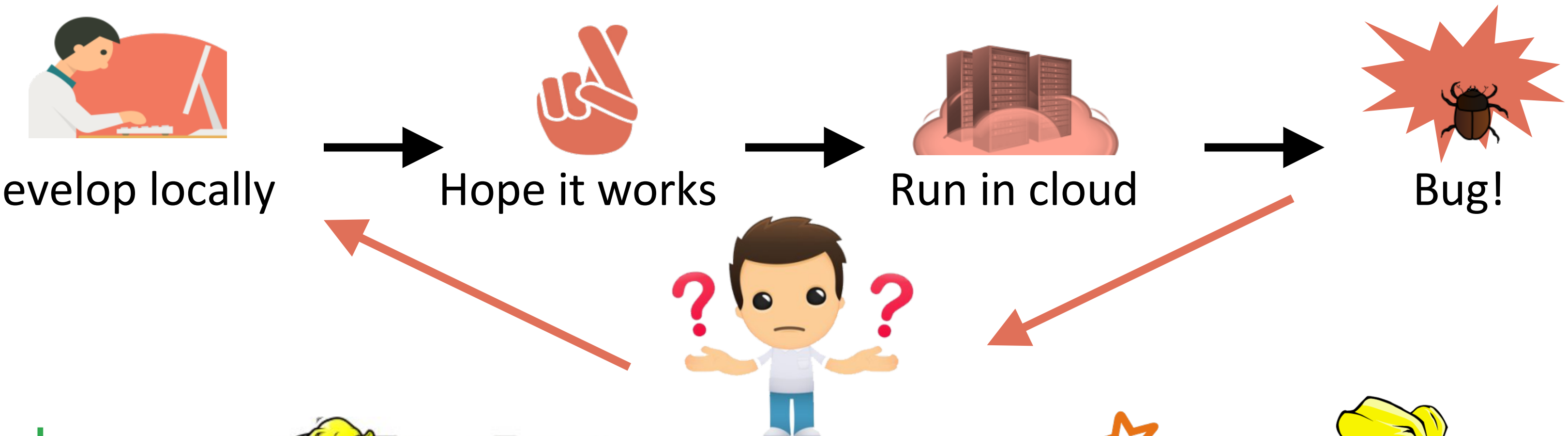
Dogfood Simulation

“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



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.**