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

Miryung Kim
UCLA
Los Angeles, CA, USA
miryung@cs.ucla.edu

Thomas Zimmermann
Microsoft Research
Redmond, WA, USA
tzimmer@microsoft.com

Robert DeLine
Microsoft Research
Redmond, WA, USA
rdeline@microsoft.com

Andrew Begel
Microsoft Research
Redmond, WA, USA
abegel@microsoft.com

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. For example, 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.

CCS CONCEPTS

• **Software and its engineering → Software creation**;

KEYWORDS

Software productivity, data science

ACM Reference Format:
Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. In *ICSE '18: 40th International Conference on Software Engineering , May 27-June 3, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/3180155.3182515

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. The mission of a data scientist is to transform data into insights that guide the teams actions. Previous studies on data scientists in software teams are based on a relatively small number of people, and therefore do not provide a broader perspective on data science work.

ICSE '18, May 27-June 3, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5638-1/18/05.
https://doi.org/10.1145/3180155.3182515

In the talk, we report 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 full-time 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 back-ground of data scientists at Microsoft?
- **RQ2.** How do data scientists work? What tasks do they work on, how do they spend the time, and what tools do they use?
- **RQ3.** What challenges do data scientists face? What are best practices and advice to overcome the challenges?
- **RQ4.** How do data scientists increase confidence about the quality of their work?

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 used as an introspective tool for assessing developer productivity and software quality. 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 work. Due to the transitional nature of adopting new responsibilities, many respondents emphasize the need of formal training and shared repositories for mentoring.

Data scientists spend a significant amount 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. To ensure the correctness of their work, more structured processes and tool support are needed for validating data science work.

The full paper [1] 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 in other companies.

REFERENCES

[1] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions in Software Engineering*. To appear. DOI: https://doi.org/10.1109/TSE.2017.2754374