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

Original paper by:Miryung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel
Summary by: Liam Day

Introduction

Data scientists analyze data to make informed decisions regarding business and engineering.

793 professional data scientists at Microsoft were surveyed

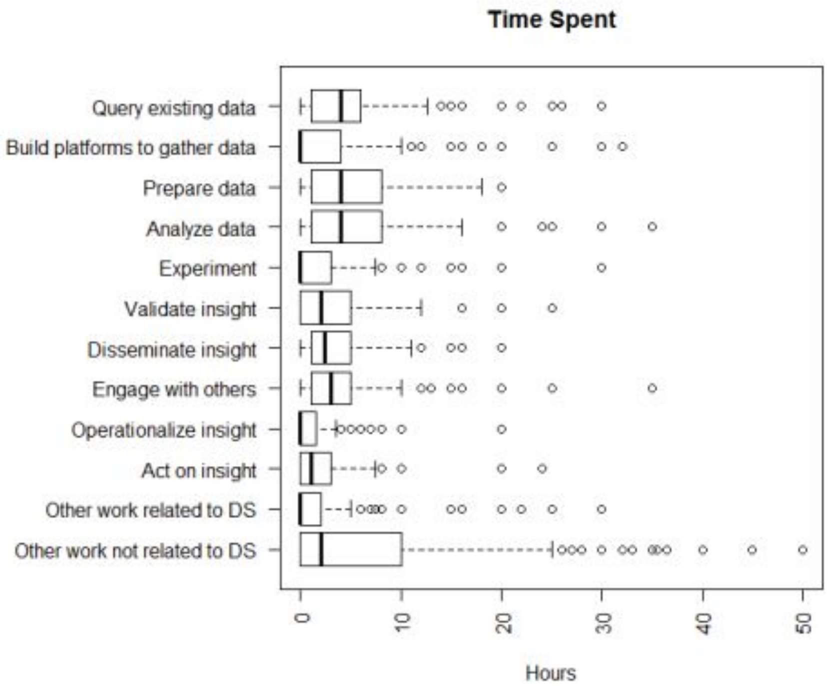
- Some of the problems Data Scientists work on
  - User Engagement
  - Software Productivity and Quality
  - Domain Specific
  - Business Intelligence
  - Discussion
- Actual Discipline of Data Scientists
  - 38% Data Scientists
  - 24% Software Engineers
  - 18% Program Managers
  - 20% Other Disciplines

What is the demographic and educational background of data scientists at Microsoft?

- Professional Experience
  - 13.6 Years on average
  - 7.4 Years at Microsoft
  - 9.8 Years analyzing data
- Educational Background
  - 34% Bachelors
  - 41% Masters
  - 22% PhD's
- Skills
  - Skill sets strong for things like Product Development, Business, and Backend Programming
  - Structured Data, Data Manipulation, and Big Data/Distributed Systems were most frequently reported
  - Spatial Statistics, Surveys/Marketing, Simulation, Bayesian/Monte-Carlo statistics were less frequently reported

Working Styles

- 81%Analyze product and customer data
- 76% Communicate results and insights
- 60% Big data cloud computing platforms
- 51% Build predictive models from the data
- 36% build data engineering platforms to collect and process data
- 31% add instrumentation to collect data
- 12% manage a team of data scientists



How do data scientists work?, and what tools do they use?

- Popular tools include
  - SQL, Excel, R, MATLAB, Minitab, SPSS, JMP, Python, Office BI, SCOPE, Azure ML, TLC
- One of the big insights of the paper was that their are too many tools.
- There are issues moving between different platforms so work has limited reusability

- Polymath
  - Jack of all trades
  - Above average PHD
- Evangelist
  - Explain data
- Preparer
  - Querying and manipulating data
  - Deal with many data streams
- Shaper
  - Analyzing and preparing
  - Most are dedicated data scientists
- Analyzer
  - Around half time analyzing
- Platform Builder
  - Building platforms and instrumentation
  - Mostly software engineers
- Moonlighter 50/20%
  - Less data focus

	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 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		

What challenges do data scientists face?

- Data Quality
  - Limits confidence
- Data Availability
  - Missing, Incomplete, and Access
- Data Preparation
  - Format and Documentation
- Scale
  - Time
  - Generic Tools lack feature
- Machine Learning
  - Problems are undefined
- Buy-In
  - Dedicated work is limited

Data scientists advice to overcome those challenges?

- Better Learning
  - Things like MOOCS are currently used
- Professionalizing the practice
  - Some basis training
- Community of Practice
  - Hands on learning was a recurring request

- R was the most popular tool
- Too many tools with limited interoperability
  - Work reuse limited
- Defining and translating between goals and data
- Having a gut instinct about the data

How do data scientists increase confidence about the correctness of their work?

- Group review
  - With peers
- Cross referencing
  - Other sources
- Human labeled ground truth
  - Need human knowledge to decide what is true
- Simulation
  - To build up a cross reference
- Repeatability

Questions & Discussion

How would you solve a too many tools issue?

Reference

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

By: Miryung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel

Accessed from: http://web.cs.ucla.edu/~miryung/Publications/tse2017-datascientists.pdf

Accessed on: 2018/07/17