The Emerging Role of Data Scientists on Software Development Teams

-Shruthi Nagaraj

Carleton University

Who is a Data Scientist ?

"The people who do collection and analysis are called *data scientists!!",*

-*DJ* Patil and Jeff Hammerbacher

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

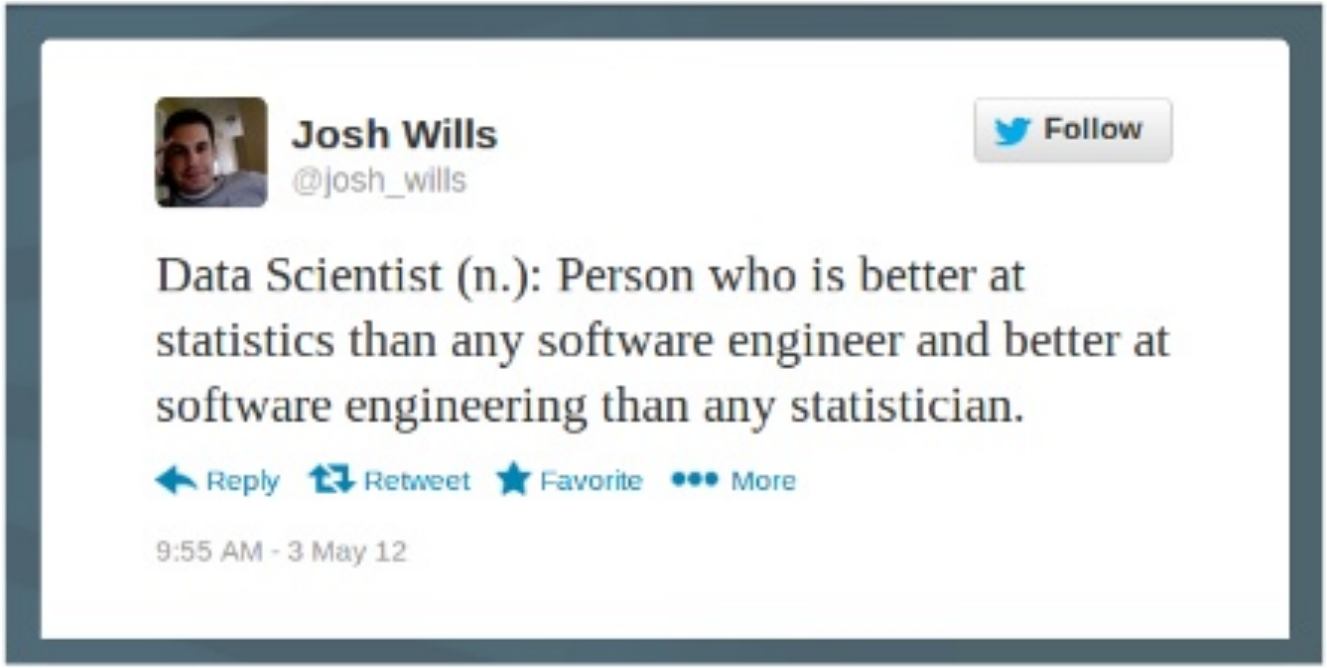
DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Enter the Data Scientist



Methodology

- Interviews with 16 participants { P1 to P16}
  - 5 women and 11 men from eight different organizations at Microsoft
- Snowball sampling
  - data-driven engineering meet-ups and technical community meetings
  - word of mouth
- Clustering of participants

DATA SCIENTISTS IN SOFTWARE DEVELOPMENT TEAMS

- Data science is not a new field, but the prevalence of interest in it has grown rapidly.

- Observed an evolution of data science in , both in Microsoft terms of technology and people

Why are Data Scientists Needed in Software Development Teams?

- **Demand for Experimentation**
  - need for designing experiments with real user data
- **Demand for Statistical Rigor**
  - conduct formal hypothesis testing, report confidence intervals, and determine baselines through normalization.
- **Demand for Data Collection Rigor**
  - data scientists discuss how much data quality matters and how many data cleaning issues they have to manage.

Background of Data Scientists

- Most CS, many interdisciplinary backgrounds
- Many have higher education degrees
- Strong passion for data
- PhD training contributes to working style

Activities of Data Scientists

- **Collection**
- *Data engineering platform, Experimentation platform*
- **Analysis**
- *Data merging and cleaning, Data shaping including selecting and creating features*
- **Use and Dissemination**
- *Defining actions and triggers, Translating insights and models to business values*

Problems that Data Scientists Work on

- Performance Regression
- Requirements Identification
- Fault Localization and Root Cause Analysis
- Bug Prioritization
- Customer Understanding
- …….etc

Organization of Data Science Teams

- *The "Triangle" model*
- *The "Hub and Spoke" model*
- *The "Consulting" model*
- *The "Individual Contributor"*
- *The "Virtual Team" model.*

Working Styles of Data Scientists

Insight Provider

Modelling Specialists

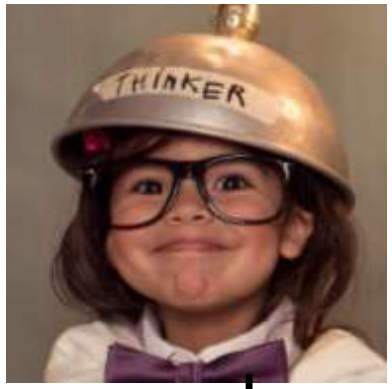
Platform Builder

Polymath

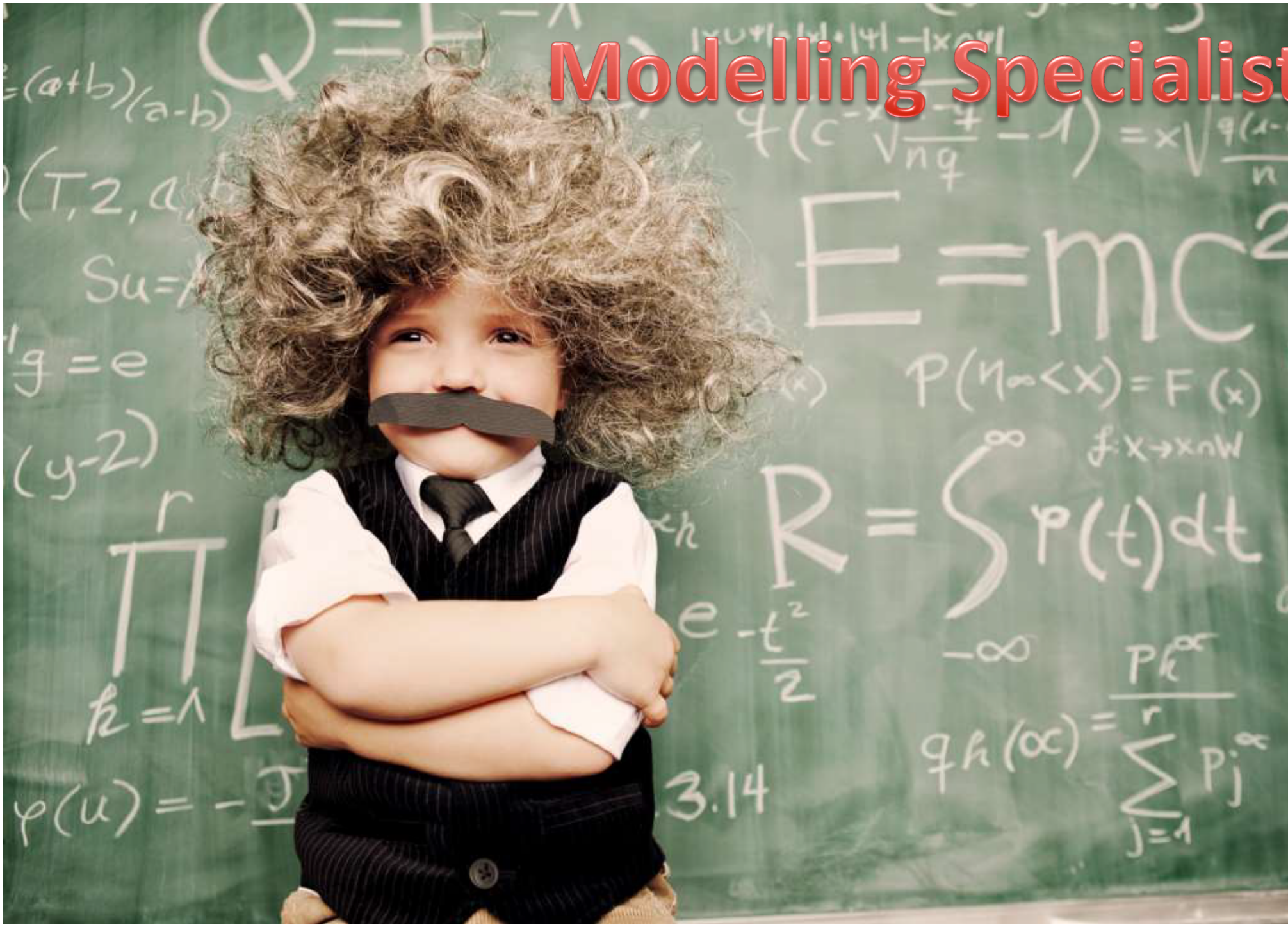
Team Leader

Insight Providers

Insight Providers



- Play an interstitial role between managers and engineers within a product group
- Generate insights and to support and guide their managers in decision making
- Analyze product and customer data collected by the teams' engineers
- Strong background in statistics
- Communication and coordination skills are key


Modelling Specialist

Modelling Specialists

- Act as expert consultants
- Build predictive models that can be instantiated as new software features and support other team's data-driven decision making
- Strong background in machine learning
- Other forms of expertise such as survey design or statistics would fit as well

Modelling Specialists

- Modeling Specialists sometimes partner with Insight Providers to define ground truths to assess the quality of their predictive models

- They believe - building new software features based on the predictive models is extremely important for demonstrating the value of their work

Platform Builders

Platform Builders

- Build data engineering platforms that are reusable in many contexts
- Strong background in big data systems
- Make trade-offs between engineering and scientific concerns

Platform Builders

- They think data collection software must be **reliable, performant, low-impact, and widely deployable**.
- On the other hand, the software should provide data that are sufficiently **precise, accurate, well-sampled, and meaningful** enough to support statistical analysis.
- Their expertise in both software engineering and data analysis enables them to make tradeoffs between these concerns.

Polymaths

Polymaths

- Data scientists who “do it all”:
  - Forming a business goal
  - Instrumenting a system to collect data
  - Doing necessary analyses or experiments
  - Communicating the results to managers

Team Leaders

Team Leaders

- Senior data scientists who typically run their own data science teams
- Act as data science “evangelists”, pushing for the adoption of data-driven decision making
- Work with senior company leaders to inform broad business decisions

IMPLICATIONS

- **Research**
- for researchers this new team composition changes the context in which problems are pursued.
- **Practice**
- how to improve the impact and actionability of data science work from the strategies shared by other data scientists.
- **Education**
- combine a deep understanding of software engineering problems,

Conclusion

- Demand for designing experiments with real user data and reporting results with statistical rigor.
- Shared activities, several success stories, and five distinct styles of data scientists.
- Reported strategies that data scientists use to ensure that their results are relevant to the company

Discussions

- Why are data scientists needed in software development teams ?
- What kinds of problems and activities do data scientists need to work on in software development teams?
- Should big companies start using this idea?


THANK
YOU!