The Emerging Role of Data Scientists on Software Development Teams

MIRYUNG KIM

UNIVERSITY OF CALIFORNIA, LOS ANGELES

TOM ZIMMERMANN, ROBERT DELINE, ANDREW BEGEL

MICROSOFT RESEARCH

Take Away Messages

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

There is also **demand for experimenting with real users**.

Data scientists are **new emerging roles** within SW teams and shaping how software is developed and tested.

We identified **five working styles** of data scientists in SW teams: *Insight Provider, Modeling Specialists, Platform Builder, Polymath, Team Leader*

Research Questions

Q1: Why are data scientists needed on SW teams?

Q2: What are the educational and training backgrounds of data scientists in SW teams?

Q3: What kinds of problems and activities do data scientists work on?

Q4: What are the working styles of data scientists in SW teams?

Methodology

Interviews with 16 participants

- 5 women and 11 men from eight different organizations at Microsoft
- Ads, Azure, Bing, Exchange, Office, R&D, Skype, Windows, and Xbox

Snowball sampling

- data-driven engineering meet-ups and technical community meetings
- word of mouth

Coding with Atlas.TI

Clustering of participants using affinity diagram and card sorting

Q1. Why are Data Scientists Needed on SW Teams?

Software companies want to **experiment with real users**, e.g., A/B testing, flighting, games and rewards.

People demand results with **statistical rigor,** e.g, confidence interval and normalization**.**

Q1. Why Are Data Scientists Needed on SW Teams?

Software companies want to **experiment with real users**, e.g., A/B testing, flighting, games and rewards.

People demand results with **statistical rigor,** e.g, confidence interval and normalization**.**

Quality assurance is moving towards statistical approaches rather than traditional testing and debugging.

“Instead of having an army of testers to go off and generate a bunch of data, that data's already here. It's more authentic because it's real customers on real machines, real networks. You no longer have to simulate and anticipate what the customer's gonna do.” [P10]

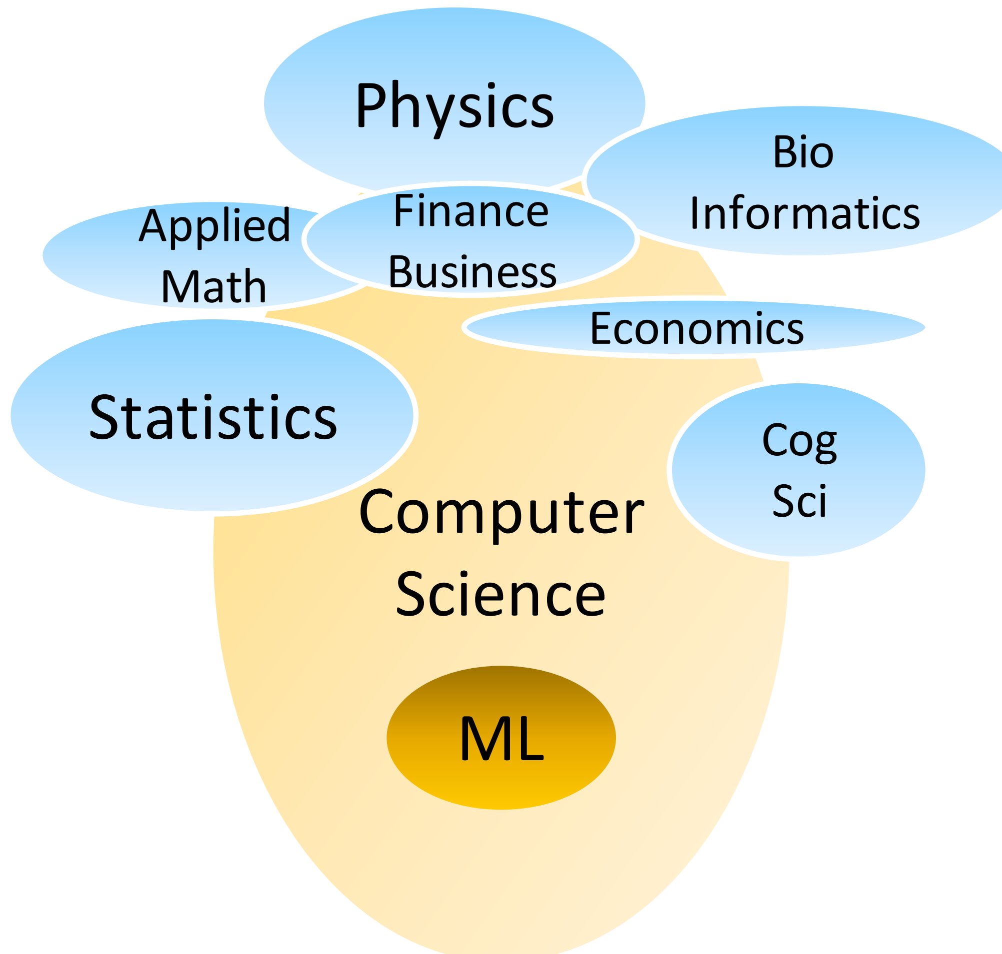
Q2: What Are the Educational and Training Backgrounds?

Most CS, many interdisciplinary backgrounds

11 in CS but many with joint degrees

Many have higher education PhD or MS degrees

PhD training contributes to working style



Q3. What Do Data Scientists Work On?

Performance Regression

Are we getting better in terms of crashes or worse? [P3]

Requirements Identification

If you see the repetitive pattern where people don't recognize, the feature is there. [P3]

Root Cause Analysis

What areas of the product are failing and why? [P3]

Bug Prioritization

Oh, cool. Now we know which bugs we should fix first. Then how can we reproduce this error? [P5]

Server Anomaly Detection

Is this application log abnormal w.r.t. the rest of the data? [P12]

Failure Rate Estimation

Is the beta ready to ship? [P8]

Customer Understanding

How long do our users use the app? [P1]

What are the most popular features? [P4]

Cost Benefit Analysis

How many customer service calls can we prevent if we detect this type of anomaly? [P9]

Activities

Collecting	Building Data Collection Platform
	Telemetry Injection
	Building Experimentation Platform
Analyzing	Data Merging Cleaning
	Sampling
	Shaping, Feature Selection
	Define Sensible Metrics
	Build Predictive Models
	Define Ground Truth
	Hypothesis Testing
Using Disseminating	Operationalize Predictive Models
	Define Actions and Triggers
	Translate Predictive Models to Domain Specific Insights

Q4: What Are Working Styles of Data Scientists?

Insight Provider

Specialists

Platform Builder

Polymath

Team Leader


Insight Providers
THINKER
ThinkingCap

Insight Providers

Coordinate between managers and engineers within a product group

Generate insights and to guide managers in decision making

Strong communication and coordination skills are key

Example: P2 worked on a product line to inform managers needed to know whether an upgrade was of sufficient quality to push to all products in the family.

Insight Providers Success Strategies

Get data from engineers but need to understand the rationale behind instrumentation

> I basically tried to eliminate from the vocabulary the notion of "You can just throw the data over the wall ... She'll figure it out." There's no such thing. I'm like, "Why did you collect this data? why did you measure this many samples, not this many?" [P2]

Engage with the stakeholders who plan to consume results, e.g. weekly data meet-up


Modelling Specialists

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

Example: P7 is an expert in time series analysis and works with a team on automatically detecting anomalies in their telemetry data.

Modelling Specialists Success Strategies

Operationalize predictive models—build features based on predictive models

Translate findings into business values such as dollars saved, customer calls prevented.

> In terms of convincing, if you just present all these numbers like precision and recall factors, that is import from the knowledge sharing perspective. But if you are out there to sell your model or ideas, this will not work. [P12]


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

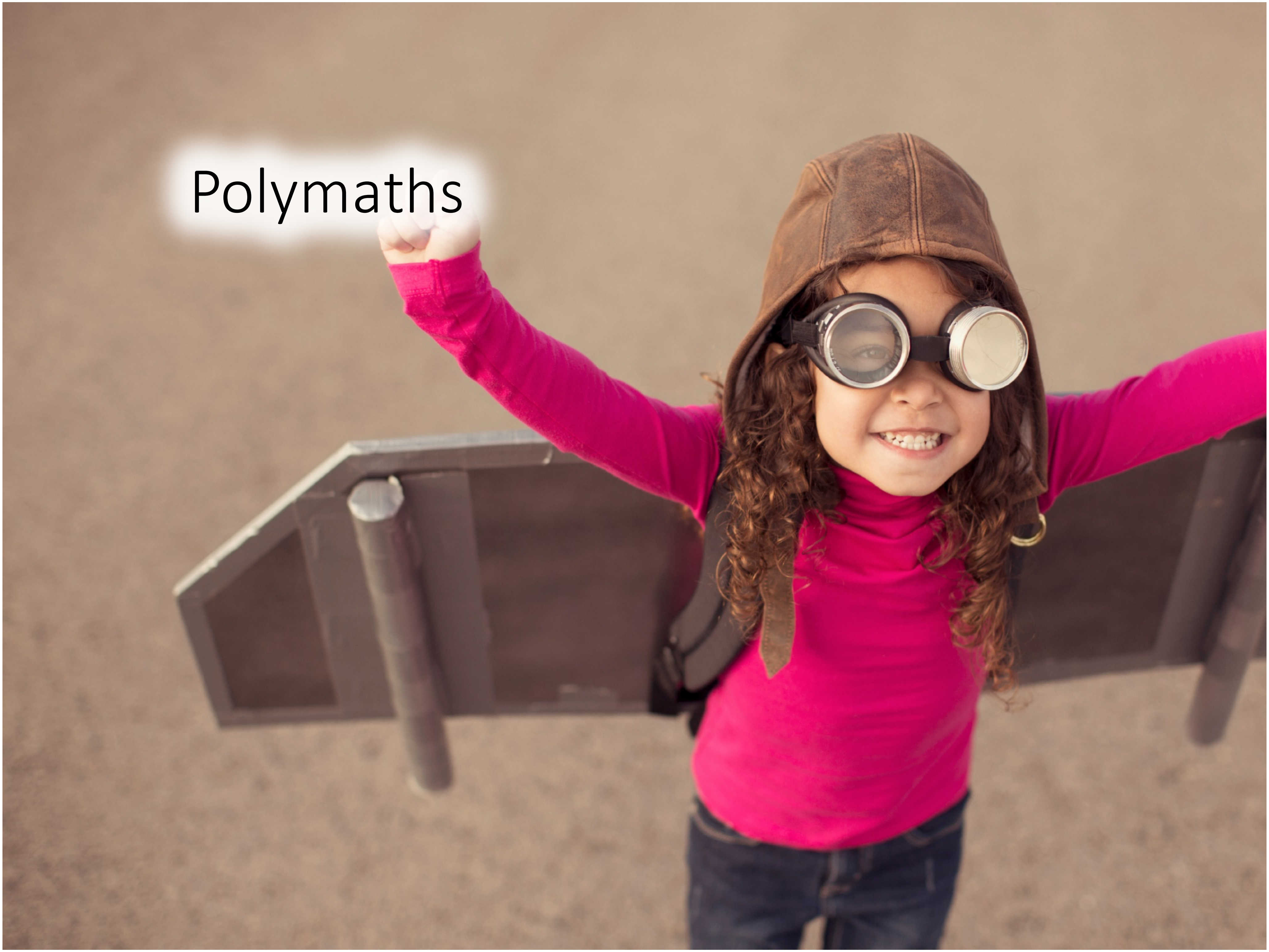
Example. P4 worked on platform to collect crash data.

Platform Builders Success Strategies

Triangulate multiple data sources to increase their confidence.

Validate quantitative data through qualitative channels

> If you could survey everybody every ten minutes, you don't need telemetry. The most accurate is to ask everybody all the time. So what we typically is 10% are surveyed and we get telemetry. And then we calibrate and infer what the other 90% have said. [P4]


Polymaths

Polymaths

Data scientists who “do it all”:

Form a business goal

Instrument a system to collect data

Do necessary analyses or experiments

Communicate the results to managers

Example. P13 works on a product that serves ads and explores her own ideas for new advertisement data models.

Polymaths Success Strategies

Polymaths set up regular channels such as "brown bag lunches" to deliver their project outcomes to their team.


Team Leaders
5

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

Example. P10 and his team of data scientists estimated the number of bugs that would remain open when a product was scheduled to ship.

Team Leaders Success Strategies

Choose the right questions for the right team

> (a) Is it a priority for the organization (b) is it actionable, if I get an answer to this, is this something someone can do something with? and, (c), are you as the feature team — if you're coming to me or if I'm going to you, telling you this is a good opportunity — are you committing resources to deliver a change? If those things are not true, then it's not worth us talking anymore.[P5]

Explain findings in simple terms.

Summary

Quality assurance is moving towards statistical approaches rather than traditional testing and debugging.

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

We have identified five distinct working styles of data scientists and strategies for improving the impact and actionability of their work.

Insight Provider

Specialists

Platform Builder

Polymath

Team Leader

Summary

Come to "BigDebug" Talk
Thu 4pm @ Colorado Room

Quality assurance is moving towards statistical approaches rather than traditional testing and debugging.

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

We have identified five distinct working styles of data scientists and strategies for improving the impact and actionability of their work.

Insight Provider

Specialists

Platform Builder

Polymath

Team Leader