



# Software Analytics

## Harlan D. Mills Award Acceptance Speech

Nachi Nagappan

# About Me

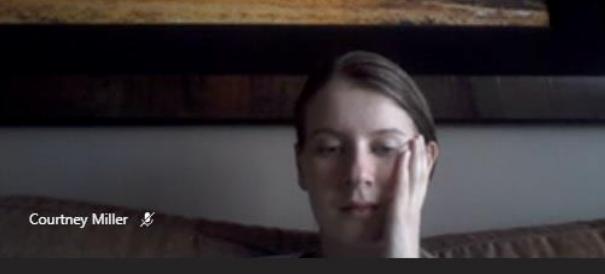
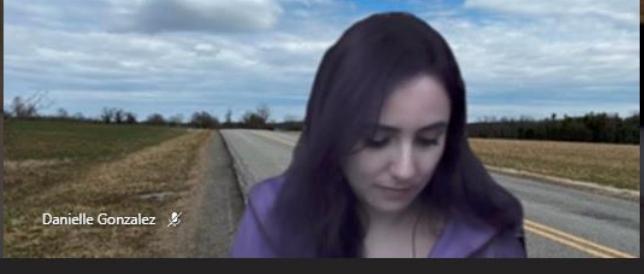


- My name is Nachiappan. I also go by Nachi.
- <https://nachinagappan.github.io/>
- Graduated with a PhD with Laurie Williams.
- I read a lot of Franco-Belgian comics (Bande dessinées)
- Attend Comic conventions
- Miniature railroad modeling (HO and G).



SAINT Social

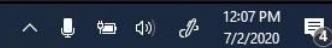
08:59



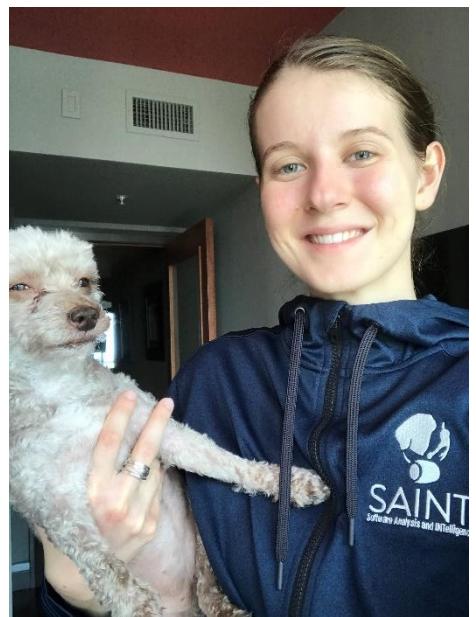
4

12:07 PM  
7/2/2020

Type here to search



Courtney Miller



Jenna Butler



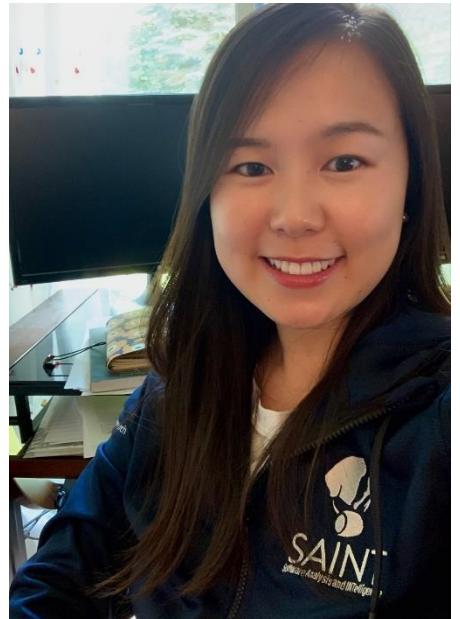
Danielle Gonzalez



Rangeet Pan



Yu Huang



Jazette Johnson



Paige Rodeghero



Rini Chattopadhyay



Courtney Miller



Jenna Butler



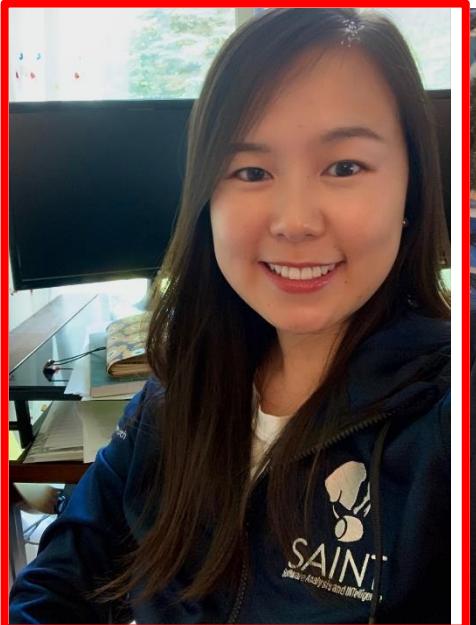
Danielle Gonzalez



Rangeet Pan



Yu Huang



Jazette Johnson



Paige Rodeghero



Rini Chattopadhyay



What metrics are the  
**best predictors of failures?**

What is the **data quality** level used in empirical studies and how much does it actually matter?

I just submitted a **bug report**.  
Will it be fixed?

How can I tell if a piece of software will have **vulnerabilities**?

Do **cross-cutting concerns** cause defects?

Does **Test Driven Development** (TDD) produce better code in shorter time?

If I increase **test coverage**, will that actually increase software quality?

Are there any **metrics that are indicators of failures** in both Open Source and Commercial domains?

Should I be writing **unit tests** in my software project?

Is strong **code ownership** good or bad for software quality?

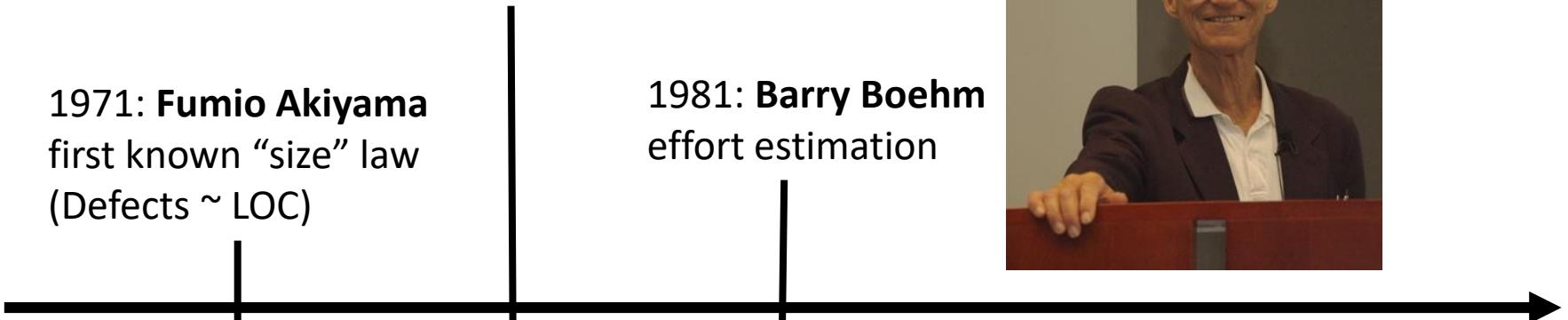
Does **Distributed/Global software development** affect quality?

# History of Software Analytics

1976: **Thomas McCabe**  
code complexity

1971: **Fumio Akiyama**  
first known “size” law  
(Defects  $\sim$  LOC)

1981: **Barry Boehm**  
effort estimation



# History of Software Analytics



**Victor Basili**



**Dieter Rombach**

Audris Mockus  
Elaine Weyuker  
Thomas Ostrand  
**and many others**

1980

1990s

2000s

Empirical Software Engineering  
Experimental Software Engineering

# History of Software Analytics



Ahmed  
Hassan



2004

Mining Software Repositories

# History of Software Analytics



Ahmed  
Hassan



2004      2005



Tim Menzies

Mining Software Repositories  
The PROMISE Repository/Conference

# History of Software Analytics



**Nachi Nagappan** founds  
Empirical Software  
Engineering at Microsoft



**Ahmed Hassan** founds  
the SAIL Group in Canada



**Dongmei Zhang** founds  
the Software Analytics  
Group at Microsoft  
Research Asia



## INFOWORLD TECH WATCH

By Paul Krill, Editor at Large, InfoWorld  
NOV 7, 2014

### About

Informed news analysis every weekday

# nature

International weekly journal of science

[Home](#) | [News & Comment](#) | [Research](#) | [Careers & Jobs](#) | [Current Issue](#) | [Archive](#) | [Audio & Video](#)

[Archive](#) > [Volume 530](#) > [Issue 7590](#) > [Research Highlights: Social Selection](#) > [Article](#)

### NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Researchers debate whether female computer coders face bias

A preliminary study suggests code-edits by female software developers are more successful — except when their gender is known.

Dalmeet Singh Chawla

[LinuxInsider](#) > [Community](#) | [Next Article in Community](#)

## MENSFITNESS

Training Nutrition Women Lifestyle Videos More ▾

# Taking short breaks during training can help you improve more quickly, video game study finds

Turns out those 36-hour Halo marathons might not actually increase think.

by Tom Briechele March 1, 2017 12:55pm

Forbes / Tech

APR 1, 2015 @ 12:32 PM 8,293

The Little Black Book of Billionaire Secrets

## Free Apps With Ads May Be Killing Your Phone's Battery And Data Plan



Alex Konrad, FORBES STAFF

Covering Silicon Alley's ad and tech scenes [FULL BIO](#) ▾

[Live Science](#) > [Tech](#)

## Older Computer Programmers<sup>13</sup> Not Out of Touch, Study Finds

By Denise Chow, Sci-Tech Editor | May 2, 2013 10:03am ET

### NEWS

Share 49 Share 176

## Code Quality Research: Functional Languages Beat Procedural/Object-Oriented

By David Ramei 10/03/2017



AGILE ARCHITECT  
CLOUD  
JAVA  
SECURITY

### NEWS

Share 49 Share 176

## Code Quality Research: Functional Languages Beat Procedural/Object-Oriented

By David Ramei 10/03/2017



ARTWORK: TAMAR COHEN, ANDREW J. BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 9.5" X 12"

DATA

# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | COMMENT | TEXT SIZE | PRINT | \$8.95 BUY COPIES

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early." Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in. He could imagine that new features capitalizing on the heuristics he was developing might

## WHAT TO READ NEXT

- [Big Data: The Management Revolution](#)
- [Making Advanced Analytics Work for You](#)
- [Google Flu Trends' Failure Shows Good Data > Big Data](#)

VIEW MORE FROM THE  
October 2012 Issue



Obsessing over our customers is everybody's job. I'm looking to the engineering teams to **build the experiences our customers love**. [...] In order to deliver the experiences our customers need for the mobile-first and cloud-first world, we will modernize our engineering processes to be **customer-obsessed, data-driven, speed-oriented and quality-focused**.



Each engineering group will have **Data and Applied Science resources** that will focus on measurable outcomes for our products and predictive analysis of market trends, which will allow us to innovate more effectively.



# Looking back...



17





## Dr. Thomas Ball

**Static analysis tools as early indicators of pre-release defect density.** [ICSE 2005](#)

**Use of relative code churn measures to predict system defect density.** [ICSE 2005](#)

**Assessing the Relationship between Software Assertions and Faults: An Empirical Investigation.** [ISSRE 2006](#)



## **Dr. Andreas Zeller**

**Mining metrics to predict component failures. ICSE 2006**

**Extrinsic influence factors in software reliability: a study of 200, 000 windows machines. ICSE 2014**



## **Dr. Prem Devanbu**

**Does distributed development affect software quality? An empirical case study of Windows Vista. ICSE 2009.**

**Putting It All Together: Using Socio-technical Networks to Predict Failures. ISSRE 2009.**

**Don't touch my code!: examining the effects of ownership on software quality. FSE 2011.**



## **Dr. Victor Basili**

**The influence of organizational structure on software quality: an empirical case study. ICSE 2008**



## Dr. Harald Gall

**Cross-project defect prediction: a large scale experiment on data vs. domain vs. process.** [ESEC/SIGSOFT FSE 2009](#)

**Does distributed development affect software quality? An empirical case study of Windows Vista.** [ICSE 2009](#)

**Software engineering for machine learning: a case study.** [ICSE \(SEIP\) 2019](#): 291-300



## Dr. Miryung Kim

A field study of refactoring challenges  
and benefits. [SIGSOFT FSE 2012](#)



## Dr. Emerson Murphy-Hill

The design of bug fixes. [ICSE 2013](#): 332-341

Cowboys, ankle sprains, and keepers of quality: how is video game development different from software development? [ICSE 2014](#)



## Dr. David Lo

**How practitioners perceive the relevance of software engineering research. [ESEC/SIGSOFT FSE 2015](#)**

**HYDRA: Massively Compositional Model for Cross-Project Defect Prediction. [IEEE Trans. Software Eng. 42\(10\): \(2016\)](#)**

**Code Coverage and Postrelease Defects: A Large-Scale Study on Open Source Projects. [IEEE Trans. Reliability 66\(4\):\(2017\)](#)**



## Dr. Audris Mockus

**Test coverage and post-verification  
defects: A multiple case  
study. [ESEM 2009](#)**

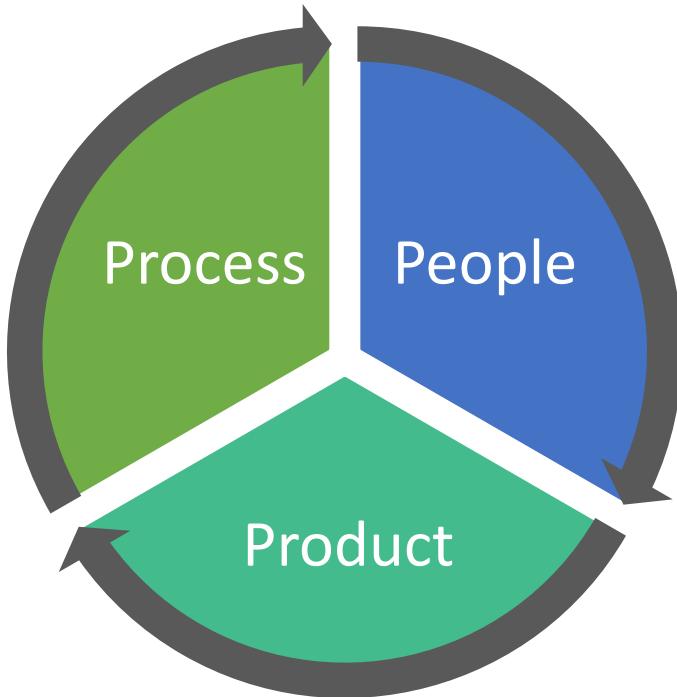




SAINT

Software Analysis and  
Intelligence

# The 3 P's of Productivity



# SAINT Focus Areas

Developer Communities	Future of Software Creation
<ul style="list-style-type: none"><li>• GitHub, Visual Studio</li><li>• Fostering empathy in developer communities</li><li>• Understanding how identity-based signals support an inclusive, open environment</li><li>• Non-traditional software engineer experience</li></ul>	<ul style="list-style-type: none"><li>• Understanding how developers and processes work now and identifying future trends</li><li>• Software 2.0</li><li>• Productivity of software teams</li><li>• Remote work</li></ul>

The common goal of our work is to understand productivity and build interventions to better support programmers.

# SAINT Focus Areas – Milestones



## Developer Communities

PR Acceptance Bias

Gender Diversity in GitHub projects



## Future of Software Creation

Effort estimation at the PR level

SE for ML

# DEVELOPER COMMUNITIES

# More specifically the goal of this study

To understand the influence of geographical location on pull request acceptance decisions in GitHub?

- Geographical location of submitters
- Same geographical location of submitters and integrators

# The Data Source

- GHTorrent data
- 1069 projects and 370,411 pull requests developed in Python (357), Java (315), Ruby (359), and Scala (38).
- Represent top 1% of the projects developed by using pull requests as the mode of collaboration.
- We use countryNameManager script by Bogdan et al and others from UC Davis.

# Observations

- Controlling for the confounding effects of
  - Project characteristics
  - Developer characteristics
  - Pull request characteristics

**Geographical location explains significant differences in pull request acceptance decisions.**

# Observations

- ✓ Compared to the United States, submitters from United Kingdom (22%), Canada (25%), Japan (40%), Netherlands (43%), and Switzerland (58%) have higher chances of getting their pull requests accepted.
- ✓ However, submitters from Germany (15%), Brazil (17%), China (24%), and Italy (19%) have lower chances of getting their pull requests accepted compared to the United States.

# Observations

- ✓ Submitters and integrators having the same nationality increases the chances of pull request acceptance decisions by 19% compared to when submitters and integrators are from different countries.

# Observations

## Submitters

SUBMITTERS

- ✓ Submitters from some countries perceive to experience bias more compared to other countries.
- ✓ Observations in agreement with quantitative analysis.

# Observations

## Integrators

INTEGRATORS

- ✓ 53% more integrators perceive that they encourage submitters from their nationality to participate.
- ✓ 8 out of 10 integrators feel that it is easy to work with submitters from the same nationality.
- ✓ Integrators do not feel that submitters from some nationalities are better at writing pull requests compared to others, except for India.

# Gender Diversity in GitHub

- Worked with a large international collaboration between MSR, SMU, DELFT, IIIT.

Region Level 1	Region Level 2	Count	Percentage		
			Man	Woman	Un- known
Africa	Northern Africa	91	91.21	5.49	3.33
Africa	Sub-Saharan Africa	273	92.67	3.66	3.66
Americas	Latin America and the Caribbean	2547	93.29	4.75	1.96
Americas	Northern America	24055	90.27	7.47	2.25
Americas	Others	5	80.00	0.00	20.00
Asia	Central Asia	34	88.24	2.94	8.82
Asia	Eastern Asia	2585	80.46	10.10	9.44
Asia	South-eastern Asia	686	87.90	6.85	5.25
Asia	Southern Asia	1463	91.46	5.47	3.08
Asia	Western Asia	529	93.19	3.40	3.40
Europe	Eastern Europe	3858	94.35	2.90	2.75
Europe	Northern Europe	7541	92.71	5.38	1.91
Europe	Southern Europe	2314	94.77	3.11	2.12
Europe	Western Europe	10637	92.94	3.88	3.18
Oceania	Australia and New Zealand	1870	92.62	5.13	2.25
Oceania	Melanesia	5	80.00	0.00	20.00
Oceania	Polynesia	5	100.00	0.00	0.00
Unknown	Unknown	12123	61.96	6.22	31.82

# Results

- There is no strong correlation between gender and geographic diversity.
- Since 2014, there has been a small and statistically significant improvement of gender diversity in North America and South-Eastern Asia, but negligible change elsewhere.

# Results

- Many of the barriers and motivations for contributing converge across geographic region.
  - Lack of resources
  - Goal alignment shift
  - Poor engineering environment
  - Poor working environment
  - Unclear onboarding
  - Inactivity on projects

# FUTURE OF SOFTWARE CREATION

# Application areas

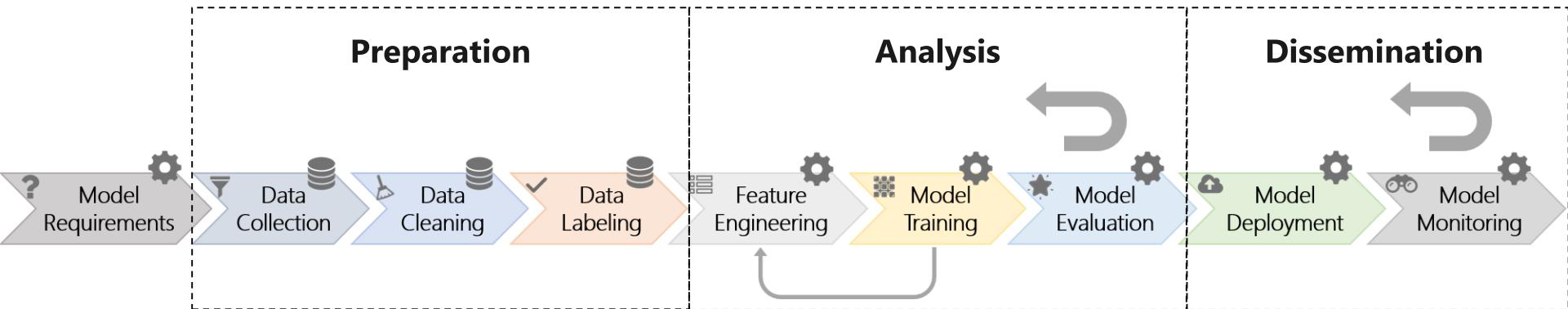
*“Please briefly describe your AI-based product, feature, or service in a few sentences.”*

Ads	Human Resources	Business
Bing	Legal	Content Moderation
Cognitive Services	Mobile	Customers
Cortana	Office	Devices
SharePoint	Research	Drawing
Skype	Security	Education
Store	Windows	Environment
Teams		Gaming
		Healthcare
		Incident Management
		Infrastructure
		Knowledge Graph
		News
		Software Engineering
		User activity / UX
		VR/AR

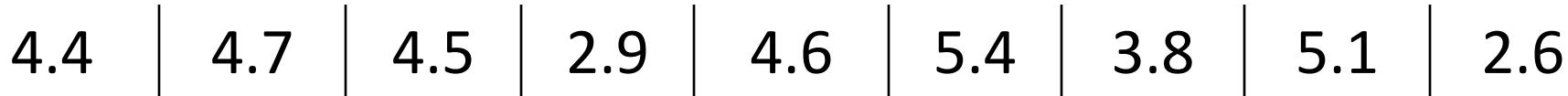
# Algorithms in use

Classification	Fraud Detection	<b>Search</b> Relevance Ranking Query Understanding	<b>Vision</b> Face Recognition Gesture Recognition Image Understanding OCR
Clustering	Navigation	<b>Recommendation</b> Profile Matching Collaborative Filtering Visual Design	<b>Speech</b> Speech-to-Text Speaker Identification
Dynamic Programming	Knowledge Graph		
Signal Processing	Root Cause Analysis		
Statistics	Social Network Analysis	<b>Prediction</b> Risk Prediction Forecasting (Sales and Marketing)	<b>NLP</b> Entity Recognition Q&A Sentiment Analysis Bots
	User Behavior Modeling		Intent Prediction Summarization Machine Translation Grammar Checking Ontology Construction Text Similarity
	Tools/Services	<b>Decision Optimization</b> Resource Optimization Planning Pricing Bidding Process Optimization	

# Machine Learning Workflow



Average reported hours/week



Percent of respondents who work on this activity



# Common challenges

End-to-end tool fragmentation

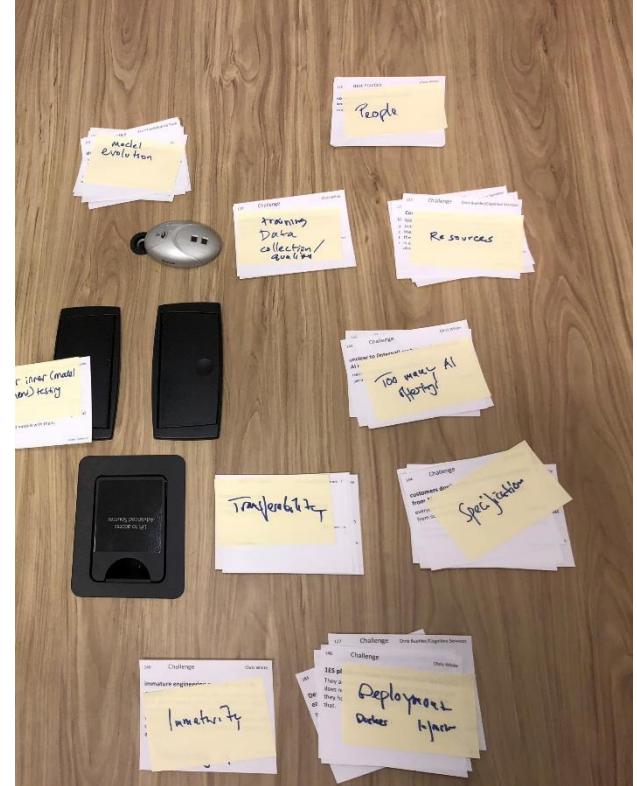
Tools can make ML too difficult

Data collection and cleaning is hard

Education and lack of expertise

Debugging is hard

Model evaluation and deployment



# Challenges differ by experience

Common to Everyone

End-to-end tool fragmentation

Data collection, cleaning, management

Low experience

Education and training

Integrating AI into larger systems

High experience

Need for specific tools

Scalability

Educating others

Model evolution, evaluation,  
and deployment

# Best practices for machine learning

ML tools need to be better stitched into the ML workflow and the workflow needs to be **automated**.

Center development around **data** (sharing, provenance, versioning).

Educating non-specialists in ML takes a lot of time but it worth the effort. Leverage **internal training and knowledge sharing**.

ML models are difficult to debug. Using **simple, explainable, and composable** models helps.

Use carefully designed **test sets, score cards for evaluating combo flights, and human-in-the-loop evaluation**.

**Do not decouple model building from the rest of the software.**

# Data, data, data

Traditional software engineering focuses mainly on code, not on data.

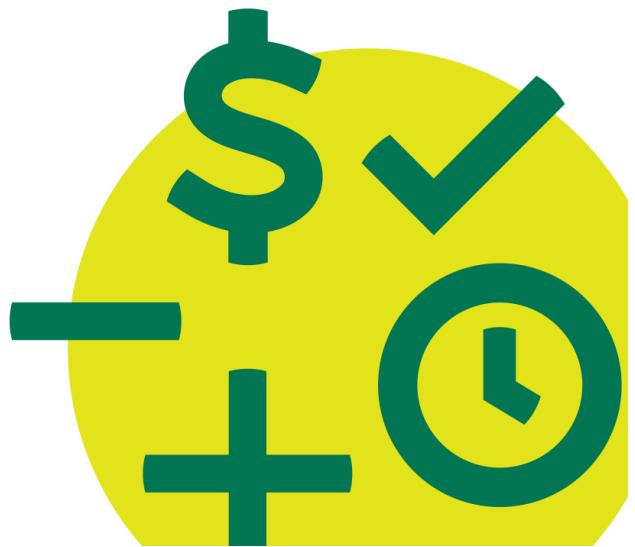
- How is data stored, versioned, and tracked in repositories?
- Data must be changed out every few months to satisfy compliance requirements.

## *Teams suggest*

- ***“Pay a lot of attention to the data.”***
- ***“Put more effort on data collection and annotation”***
- ***“Be relaxed about framework / machine learning code, but **careful & deliberate** about data & objectives.”***
- ***“Standardize on terminology and naming conventions such as the same type of user\_id”***
- ***“Reuse the modules or data as much as possible to reduce duplicate effort.”***

# Effort Estimation

“In software development, ***effort estimation*** is the process of predicting the most realistic amount of effort (expressed in terms of person-hours or money) required to develop or maintain software based on incomplete, uncertain and noisy input”



Definition from Wikipedia

# Overruns

*66% of enterprise software projects have cost overruns*

*22% of enterprise software projects go beyond estimated schedule*

*17% of IT projects go so bad that they can threaten the very existence of the company*

Based on McKinsey report

# Model

## Process

- Number of active PRs at this time, Is it a bug fix, Is it a feature,  
Number of reviewers

## Developer

- Age of the developer in current team, Age of the developer in  
current repo, Age of the developer in Microsoft

## Churn

- Class churn, Method churn, Loop churn, Class member churn,  
Loc changed

## Temporal

- Day of the week, Average age of PRs With similar paths

## Architectural

## Feature Space

- Number of paths touched, Number of distinct file types, Is  
csproj being edited, Is it a refactor, Is it a deprecate

# Feature Correlations

Positive	Negative
<ul style="list-style-type: none"><li>• Day of the week</li><li>• Average age of PRs of a developer</li><li>• Average age of PRs which touched similar paths</li></ul>	<ul style="list-style-type: none"><li>• Age of the developer in current repo</li><li>• Age of the developer in current team</li><li>• Is the PR fixing a bug?</li></ul>

# Nudge Comment

 **Sankie Service** 04/01/2019 Resolved ✓

	Analyzing historic data and trends, PRs like this tend to be completed in <i>110 hours (approx)</i> . As it is already <b>2 days</b> past the estimated time frame, you may want to consider driving this PR towards completion.
---	--

Please provide feedback/comments/questions [here](#).

This data was generated by machine learning suggestions. Please do one of the following:

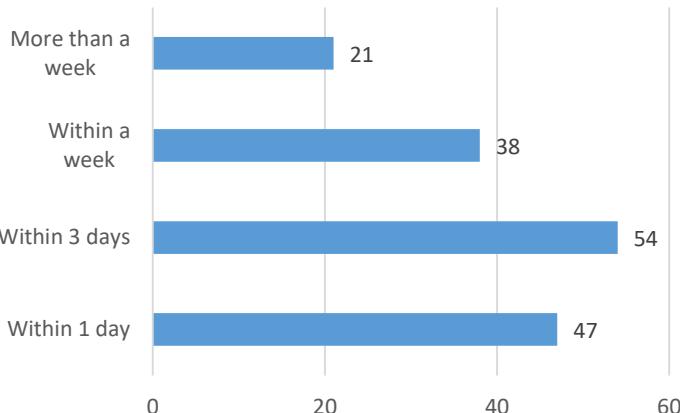
- a) Please Resolve the comment if the comment is reasonable.
- b) Please mark the comment as Won't Fix if it is not relevant.

---

 **Derek** 04/02/2019  
This was reasonable. This PR sat stale while doing work for FHL week, so it was untouched for an extended period.

 Write a reply... Reactivate

# Evaluation



No.of PRs completed after notifying

Category	Count	Time from Decoration - Completion
Non-Notified PRs	1655	103.07
Notified PRs	1069	71.27

Average time to completion



# Anecdotes

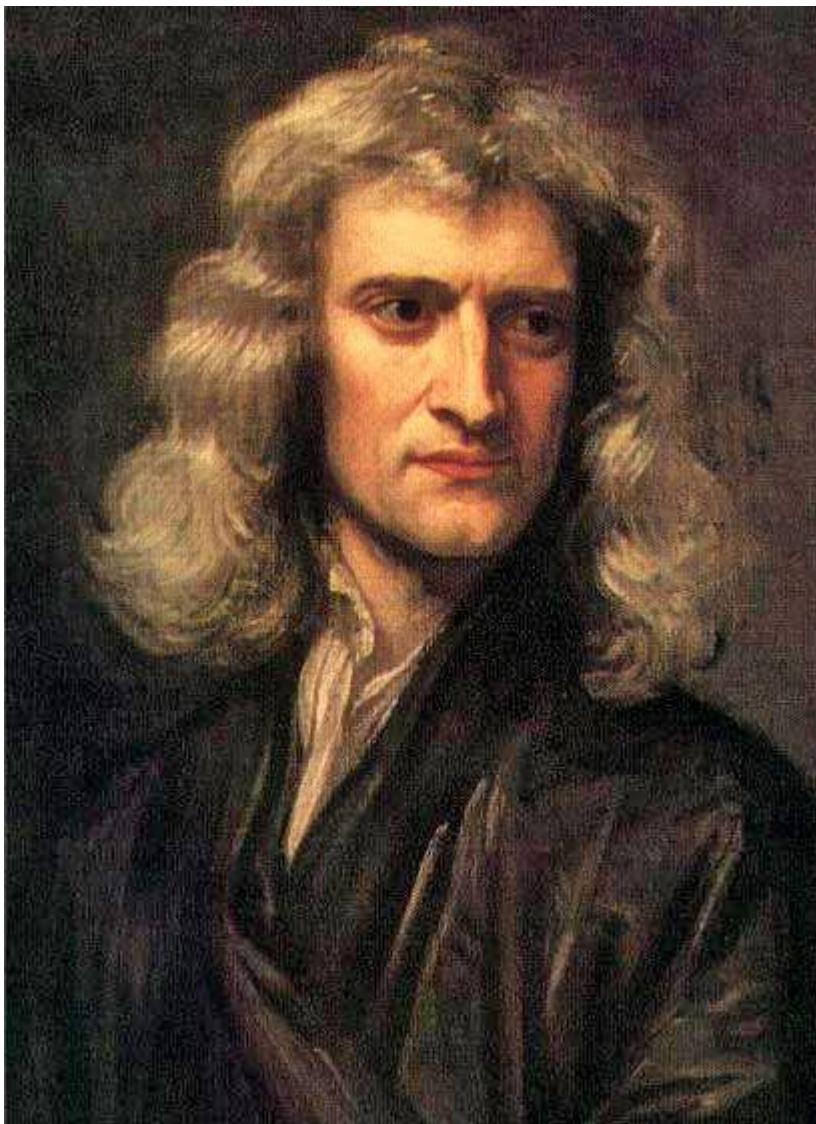
**“The pipeline is failing and blocking this check in. Followed up with an ICM incident! ”**

**“I thought the approximation was pretty good. Making few more changes and pushing this PR through! Thanks! ”**

Comment resolution percentage is

**73.3%**

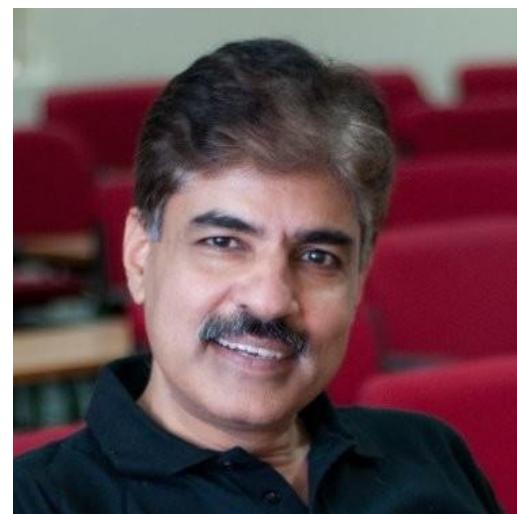
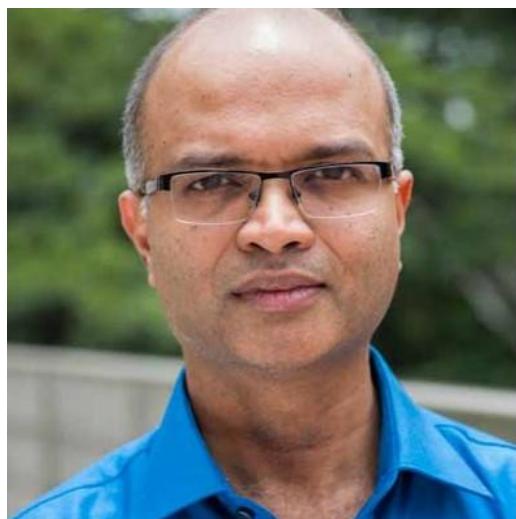
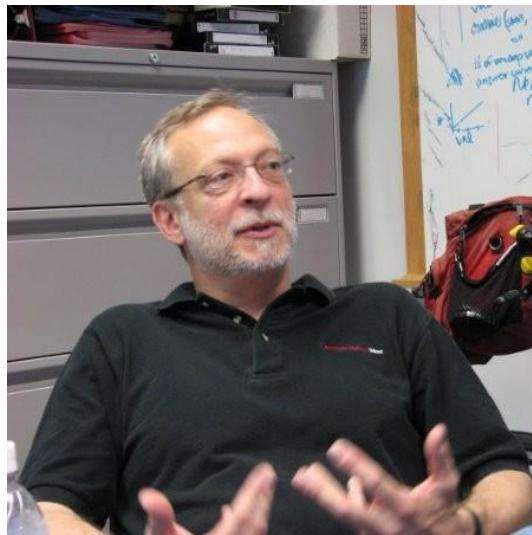
**“The approximation does sound about right. I went on-call which led to delay in check-in in this case. Normally, it would have been within about that range”**



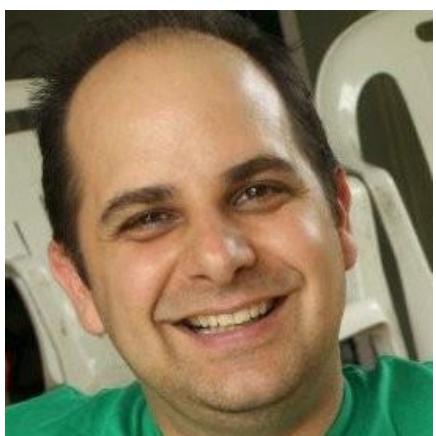
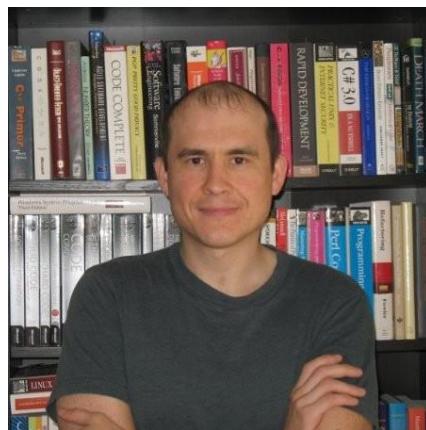
Isaac Newton in 1675: "If I have seen further it is by standing on the shoulders of Giants."

# THANKS TO...











Ayushi



Baishakhi



Erik



Foyzul



Jeff



Ken



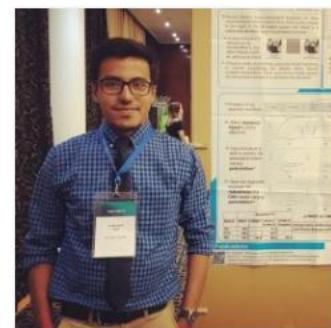
Laura



Lucas



Pavneet



Rangeet



shilpa



Song



Avushi



Baishakhi



Erik



Foyzul



Jeff



Ken



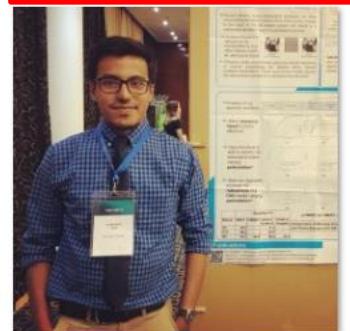
Laura



lucas



Pavneet



Rangeet



shilpa



Song

Thank you!