QA: Different strokes for different systems.

Michael Hilton

Learning goals

Interview: **Tammy Butow**

Retrospective: Recitation

- Blameless retrospective.
- What went wrong?
- Replacement Engineer
- What was the purpose of this recitation?
- What can you learn from this?
- How could you tell this story in an interview?

QA for ML

What does it mean to do QA for a ML System?

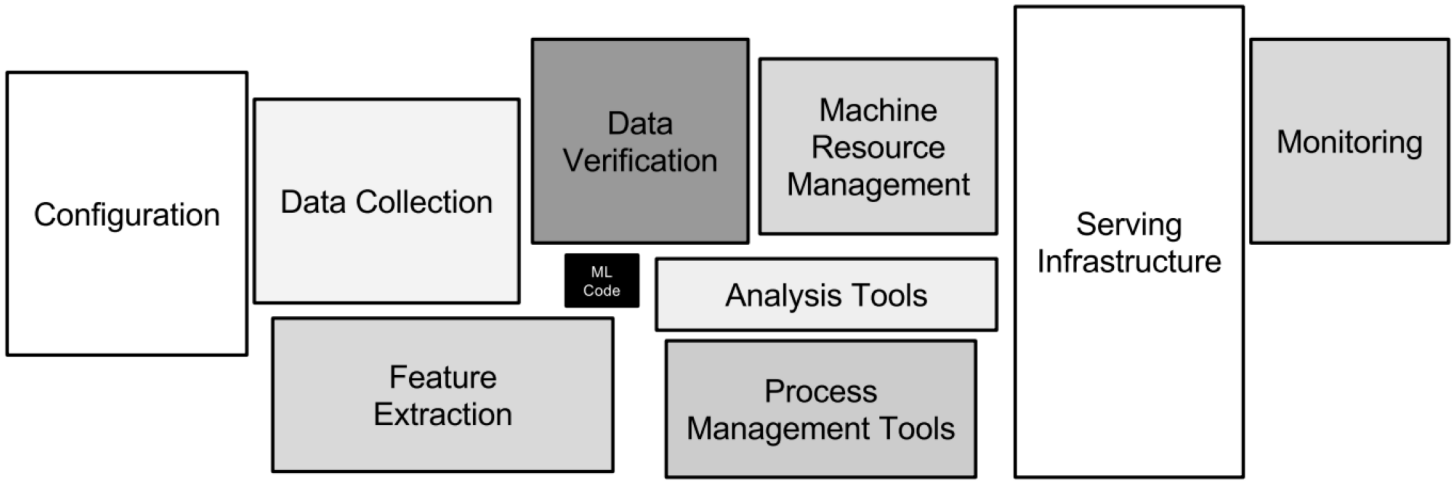


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Broad considerations when testing ML

- Data debugging, validation, and testing
- Model debugging, validation, and testing
- Service debugging, validation, and testing

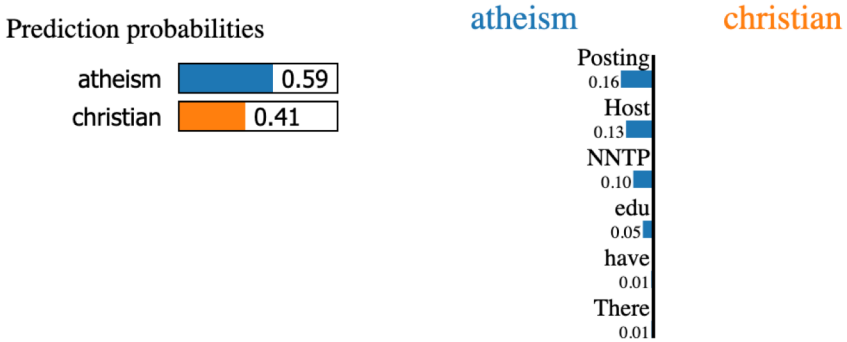
Data Debugging

- Validate Input Data Using a Data Schema
  - For your feature data, understand the range and distribution. For categorical features, understand the set of possible values.
  - Encode your understanding into rules defined in the schema.
  - Test your data against the data schema.
- Test Engineered Data: For example:
  - All numeric features are scaled, for example, between 0 and 1.
  - One-hot encoded vectors only contain a single 1 and N-1 zeroes.
  - Missing data is replaced by mean or default values.
  - Data distributions after transformation conform to expectations.
  - Outliers are handled, such as by scaling or clipping.

Model Debugging

- Check that the data can predict the labels.
  - Use 10 examples from your dataset that the model can easily learn from. Alternatively, use synthetic data.
- Establish a baseline
  - Use a linear model trained solely on most predictive feature
  - In classification, always predict the most common label
  - In regression, always predict the mean value

Beyond Correctness….



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

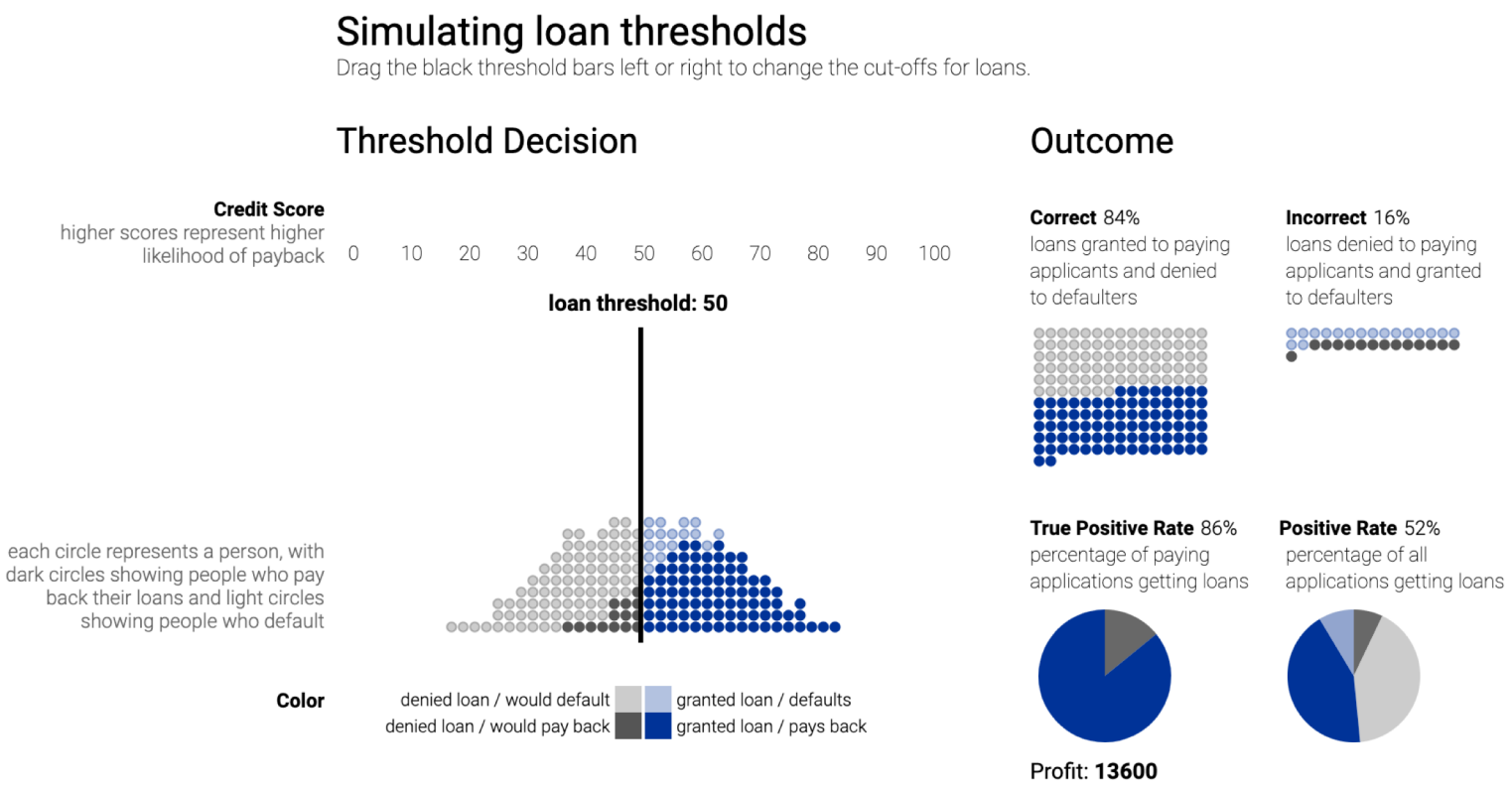
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Thanks,

john chadwick

Explainability



https://research.google.com/bigpicture/attacking-discrimination-in-ml/

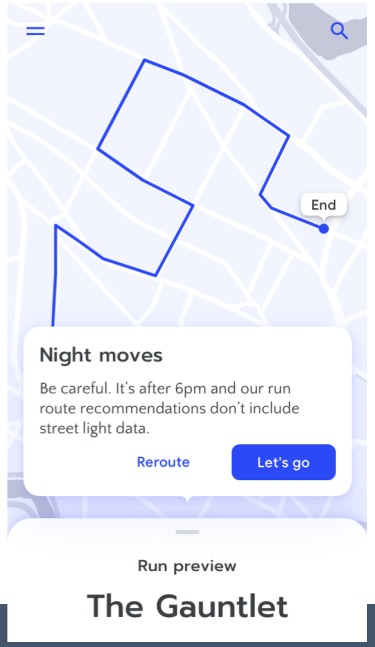
Varieties of fairness

- Group unaware
- Group thresholds
- Demographic parity
- Equal opportunity
- Equal accuracy

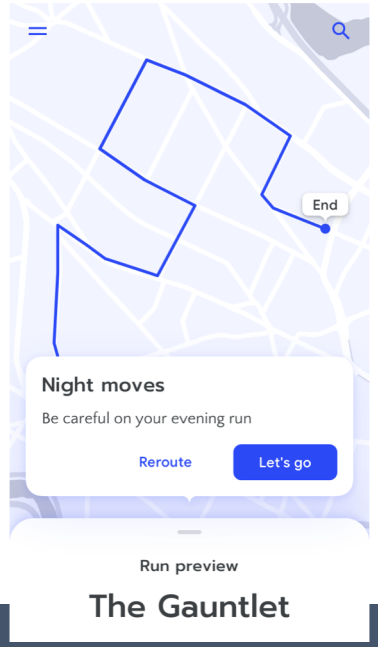
Which of these do we want?

Trust Calibration https://pair.withgoogle.com/

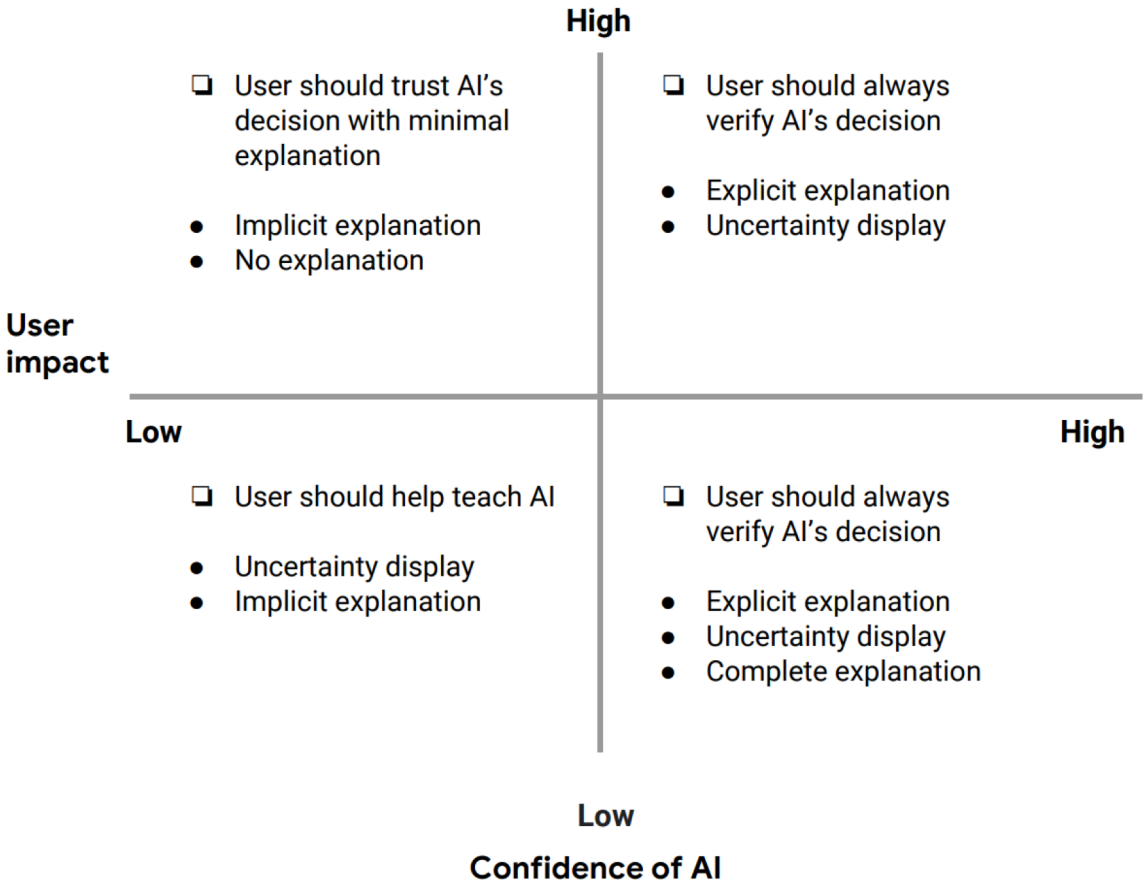
- As a team, brainstorm what kinds of experiences and interactions would decrease, maintain, or inflate trust in your feature's AI. Identify the underlying data sources, system data and user knowledge, that could impact the calibration.



Vs



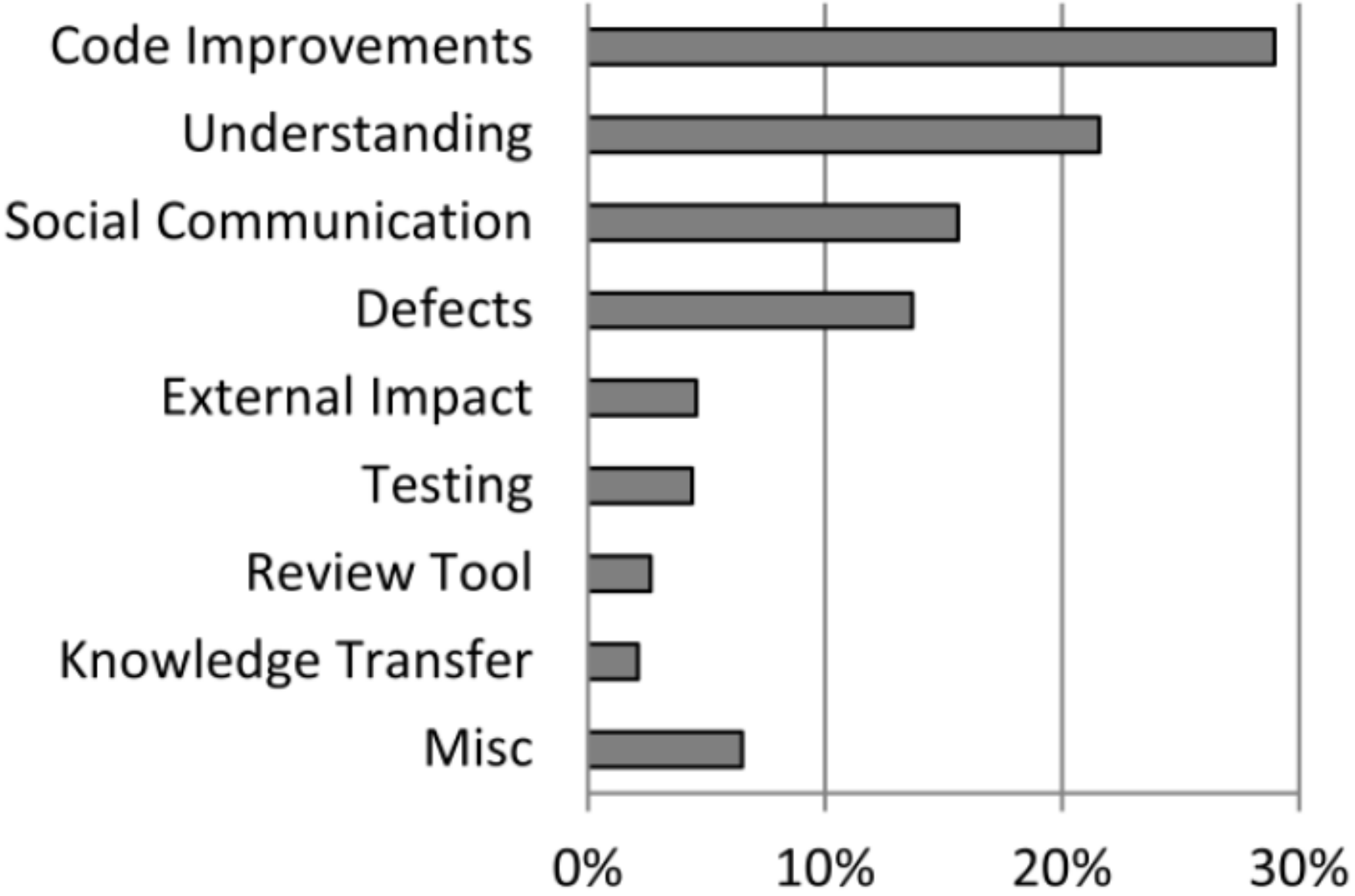
Explanation Strategy



Code Reviews

What are Code Reviews?

Outcomes (Analyzing Reviews)



Bacchelli, Alberto, and Christian Bird. "Expectations, outcomes, and challenges of modern code review." *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013.

Mismatch of Expectations and Outcomes

- Low quality of code reviews
  - Reviewers look for easy errors, as formatting issues
  - Miss serious errors
- Understanding is the main challenge
  - Understanding the reason for a change
  - Understanding the code and its context
  - Feedback channels to ask questions often needed
- No quality assurance on the outcome

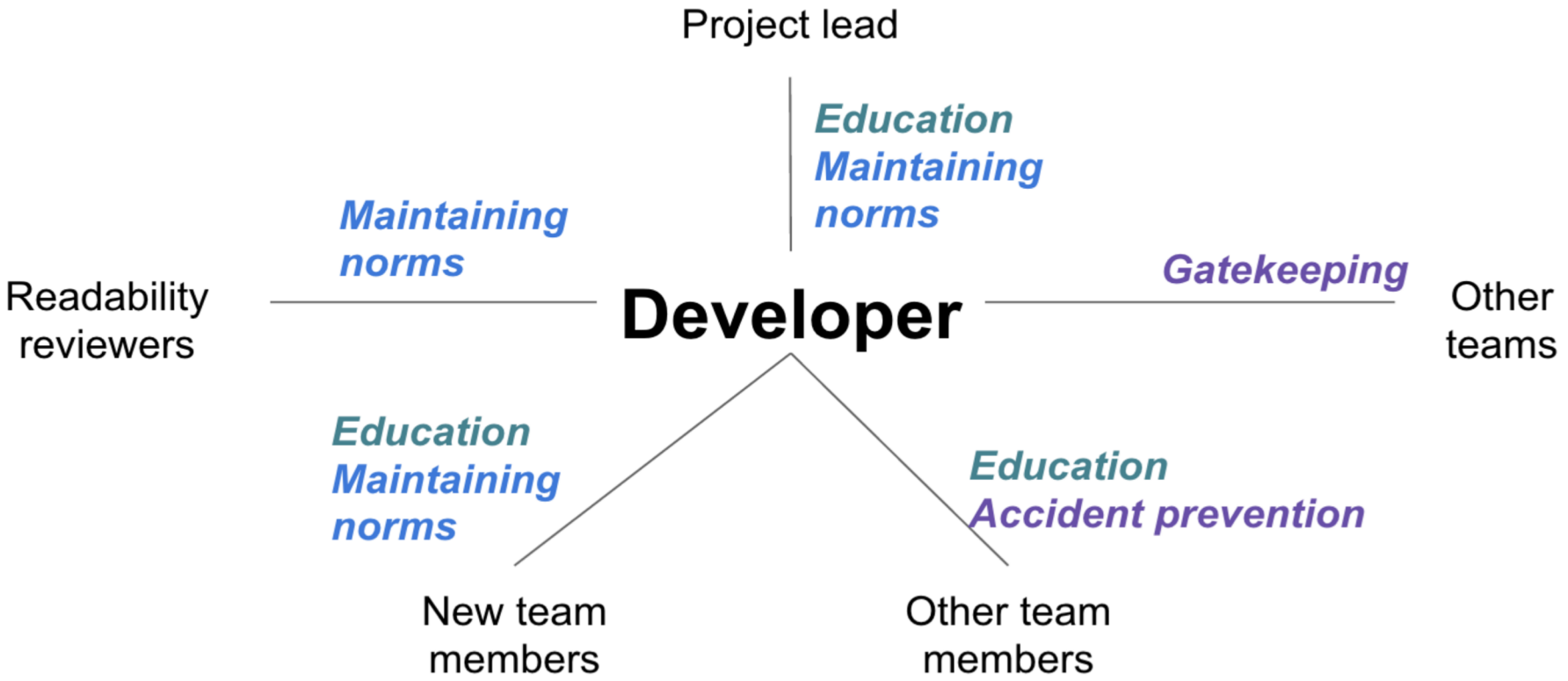
Bacchelli, Alberto, and Christian Bird. "Expectations, outcomes, and challenges of modern code review." *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013.

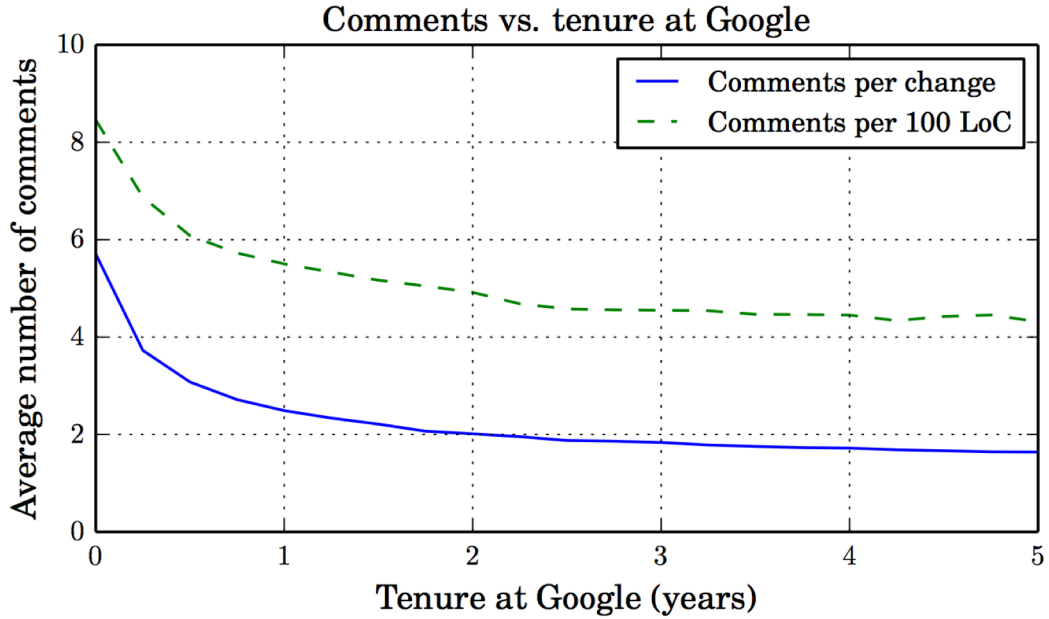
Code Review at Google

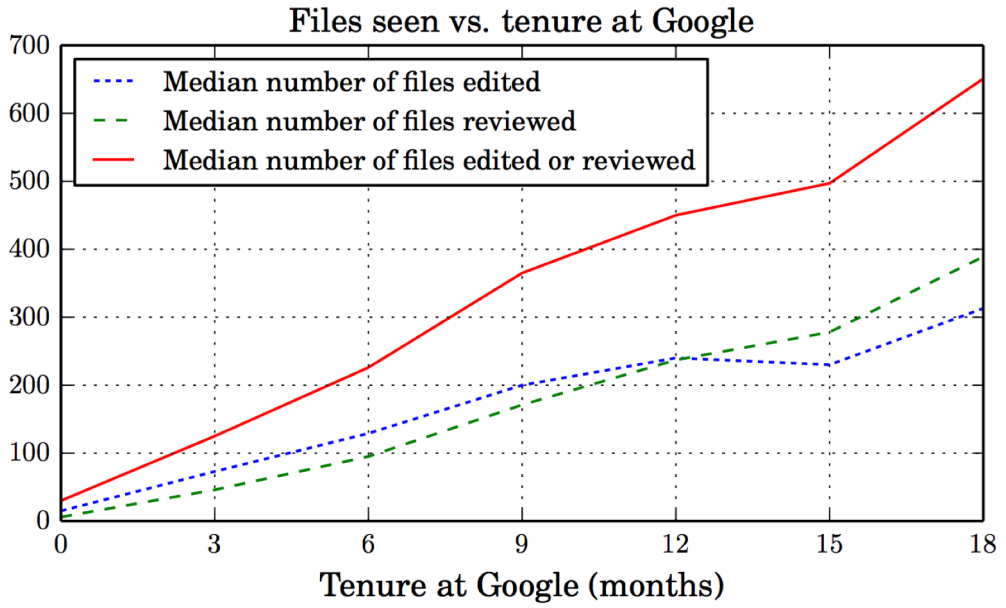
- Introduced to "*force developers to write code that other developers could understand*"
- 3 Found benefits:
  - checking the consistency of style and design
  - ensuring adequate tests
  - improving security by making sure no single developer can commit arbitrary code without oversight

Caitlin Sadowski, Emma Söderberg, Luke Church, Michal Sipko and Alberto Bacchelli. 2018. Modern Code Review: A Case Study at Google. International Conference on Software Engineering

Reviewing relationships




Comments vs. tenure at Google
Comments per change
Comments per 100 LoC
Average number of comments
Tenure at Google (years)
0
2
4
6
8
10
0
1
2
3
4
5


Files seen vs. tenure at Google
Median number of files edited
Median number of files reviewed
Median number of files edited or reviewed
Tenure at Google (months)
0
100
200
300
400
500
600
700
0
3
6
9
12
15
18