SYSTEMS ENGINEERING RESEARCH CENTER

RT-206: Data science approaches to prevent failures in systems engineering

Sponsor: DASD(SE)

By

Prof. Karen Marais (AAE) and Prof. Bruno Ribeiro (CS)

11th Annual SERC Sponsor Research Review

November 19, 2019

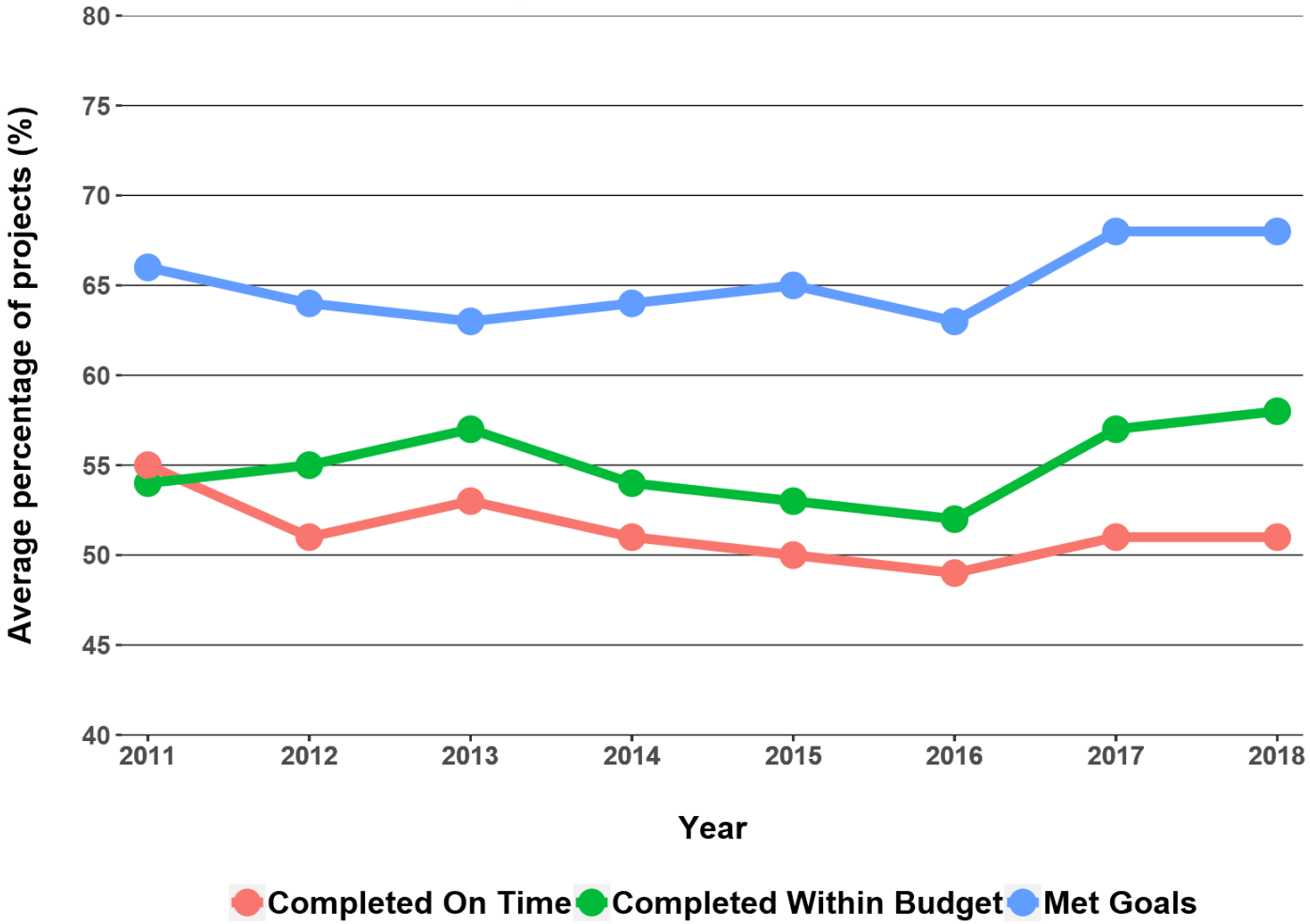
FHI 360 CONFERENCE CENTER

1825 Connecticut Avenue NW, 8th Floor

Washington, DC 20009

www.sercuarc.org

Project failures occur despite systems engineering best practice



Project delays, cost overruns, quality concerns…

Data obtained from Project Management Institute (PMI): Pulse of the Profession 2018

Humans are an integral part of project failures

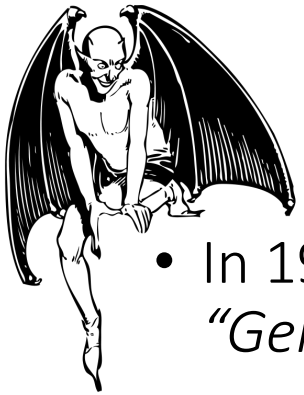
- People usually do not willfully cause failures but make mistakes that lead to failures
  - E.g. Get distracted while conducting equipment testing, communicate poorly during design phase

- Activities, behaviors, and personality can lead to poor practices and individual performance
  - E.g. Poor team coordination decreases productivity[1]

- **Our core ideas:**
  - Risk assessment based on these human actions that lead to failures
  - Predict failures and get actionable insight as to how to avoid them
  - Have an adaptable process that is tailored to the organization

[1]Eccles, D.W. and Groth, P.T., 2006. Agent coordination and communication in sociotechnological systems: Design and measurement issues. *Interacting with Computers*, *18*(6), pp.1170-1185.

Year 1: A Quest to Find Lurking Human Failure Root Causes in Projects…



- In 1904 Spearman published his *"General Intelligence" Objectivity Determined and Measured*
  - Observed human ability as a manifestation of **latent** general intelligence **factors**
  - Spearman's paper was pinnacle of 20th century statistical tools: **factor analysis**
- Influential because **latent factors** can explain why seemly unrelated observations are correlated in practice
  - *Spoiler alert:* They share hidden common "root causes"

Presentation Roadmap (Today)

- *Year 1* in 25 minutes:

  Capturing Latent Failure "Root Cause" Factors from Crowd Signals

- *Q1: How to get failure signals*?
  A: Ask the crowd

  1. *Don't ask if someone's team is properly assessing risk*, ask
     - "Did you spend time discussing trivial matters with your team last week?"
     - "Did you disagree with an idea because you did not understand all the potential implications?"
     - "How often did you notice a "silent room"?"
     - …
  2. Ask the all team members (crowd) rather than trusting a single individual
     - Rarely one person has a complete picture of a project

- Q2: How to extract latent *failure "root cause" factors from crowd signals*?

  A: Latent factors + state-of-the-art machine learning tools

  1. Spearman's latent factor cannot be applied to new teams

  2. More predictive factors with less data: From 20th century Spearman to state-of-the-art **machine learning**

Presentation Roadmap (Future)

Year 2 Ideas:

1. *Causation from data alone*
   - From latent "root cause" factors → corrective action
   - Knowing latent "root causes" ≠ knowing how to intervene
   - Correlation is not causation
     - But causation may be inferred from correlation through novel machine learning tools

2. *Industrial Partnerships:* Student projects → industry projects
   - Deployment on "real" projects
     - Industrial partnerships
       - Year 1 student projects → Year 2 "real" industrial projects
     - Can we ask fewer questions without loss of power?

Why Crowd Signals?

A Crowd-Based Risk Assessment Tool for Organizations

Easy to collect

Collect data in a way that is not cumbersome, via an online crowd-signal app

Frequent & expanded data collection

We ask a large number of employees about the projects frequently, collecting risk information continuously

Adaptable

Use data science and machine learning to make the approach adaptable to the organization

Based on the common reasons of failure

Previous research has identified 21 “real reasons” that projects fail (e.g., poor documentation)

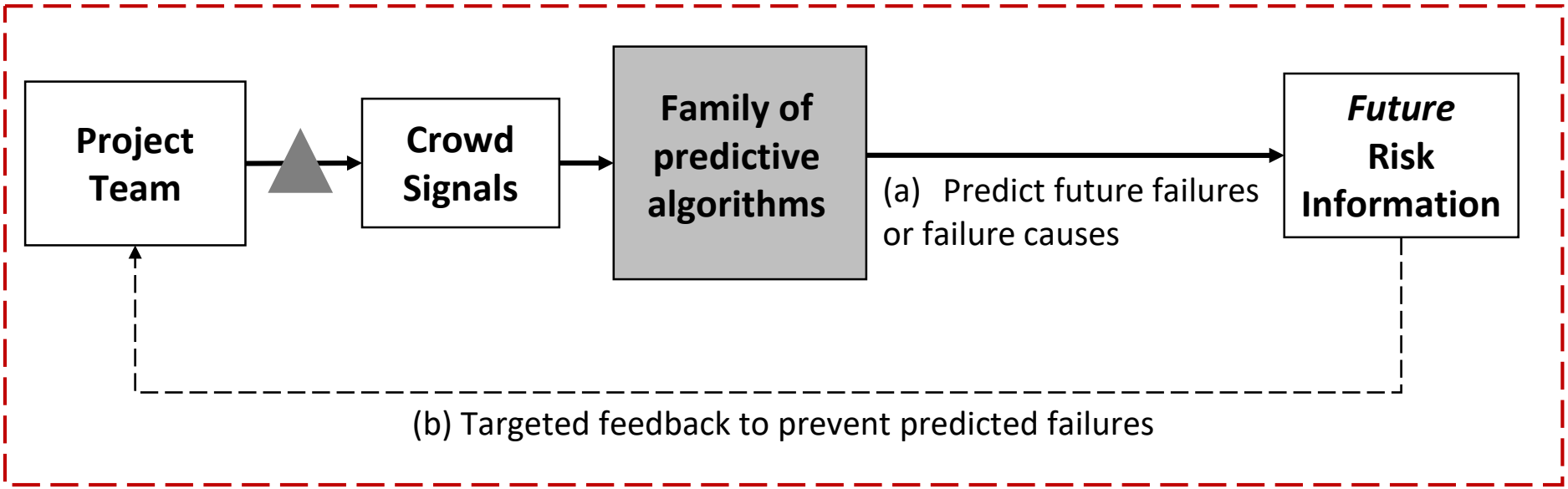
“Hard-to-game”

Questions that do not have obvious answers, but capture human behavior

Human behavior at the core

Include a variety of signals from factors that impact project and team performance

Capture “the pulse of the organization” frequently, continuously, and as efficiently as possible



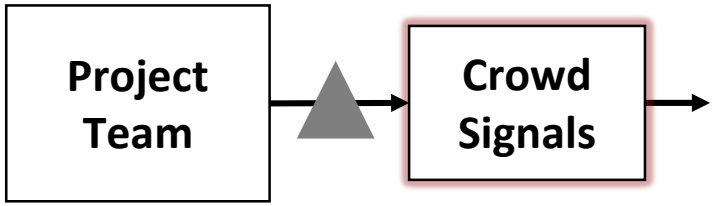
- Crowd signals
  - Data collected directly from members of a project team
  - Questions that people answer about their behaviors and activities related to their projects
  - Start with a large set of questions, then narrow down

Student teams as testbed for industry teams

Project Team

- Smaller-scale engineering projects allow students to design, manufacture, test, or operate equipment
- Students work in teams and make weekly progress with meetings and interaction with the instructor
- We have ready-access to such teams
- The teams are doing “real engineering”
- The students come from diverse backgrounds, reflecting what industry looks like

Literature as a guide to develop the crowd signals



- Factors that affect individual, team, and project performance
- Cognitive biases and safety archetypes
- Indirect metrics we suspect relate to failure
  - E.g. Personal habits, number of ordered parts, number of unscheduled team meetings
- 49 questions in total
- The students respond to these 49 questions every week

When asking people questions, there is a problem….

- Bandwagon effect: Tendency to do or believe what others do or believe

Do you suffer from bandwagon effect?

- Focusing effect: Tendency to place too much importance on one aspect

Do you suffer from focusing effect?

No, of course not!

- When possible, we phrase questions as hard to game and in context of a student project

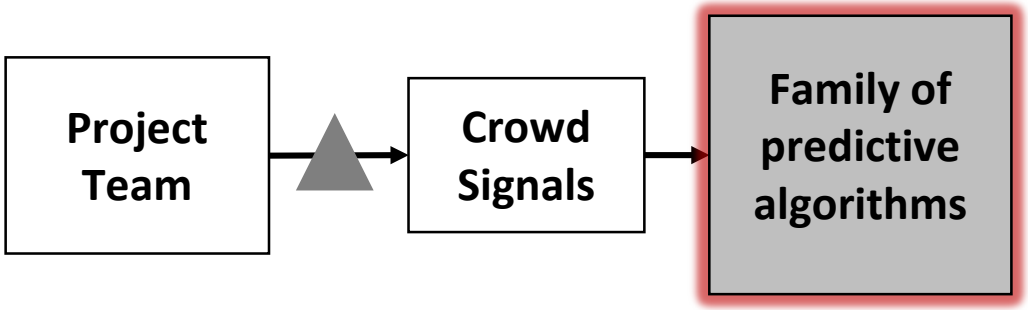
Bandwagon effect

Focusing effect

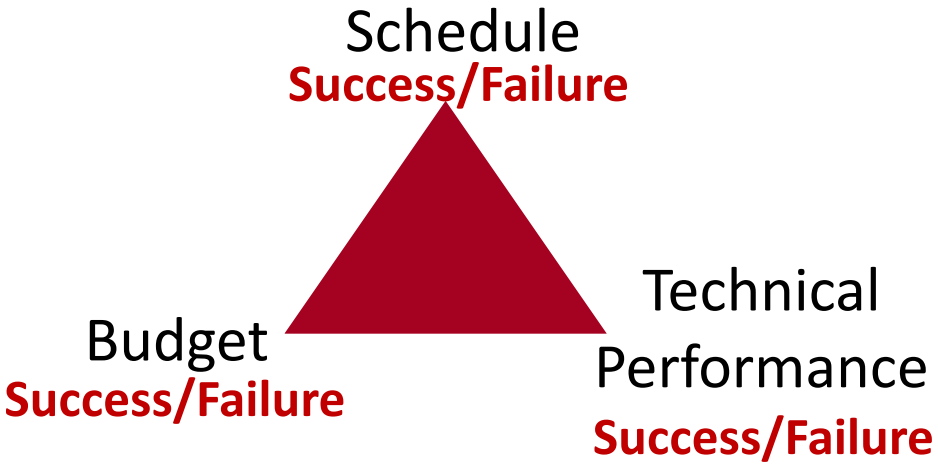
Fixing symptoms rather than root causes:

Using symptomatic solutions that become less effective over time resulting in problem resurfacing

Instructors provide the failure events we need to train the predictive models



- Three qualitative metrics of project success
  - One question for each
  - Dependent variables



We asked the instructors three questions, one for each project metric

What is currently true about the project spending, compared to what you initially planned?

	Project spending
Project 1	On budget
Project 2	On budget
Project 3	On budget
Project 4	On budget
Project 5	On budget

Cli ns...

Under budget
On budget
Over budget

How to extract latent failure “root cause” factors from crowd signals?

Challenge: Spearman's latent factors cannot be applied to new teams

Q1: Are the Crowd questions useful to predict failures?

Simple logistic regression models to predict failures

- Problem: To predict whether a student team will have a failure in the next week, based on their responses this week
  - Occurrence of failure is binary, so we use classification
  - Mixed effects model to account for correlation between responses from the same student
  - 3 models in total: budget, schedule, technical requirements failure
  - Logistic regression includes lots of assumptions, but hints to which questions correlate with which failure

Probability of failure of team of student *i* next week (t+1)

Student answers this week

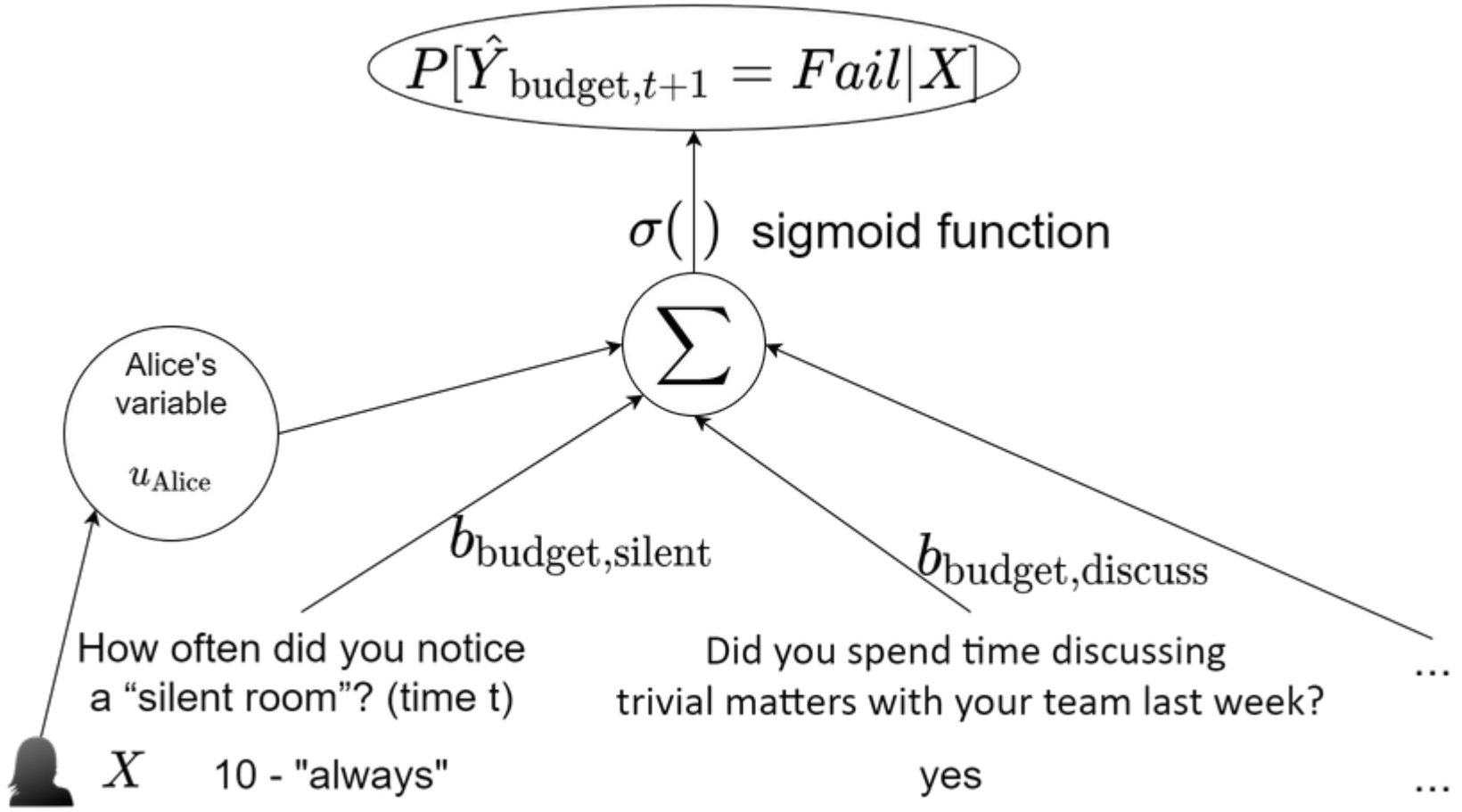
Random effects to account for student-specific effects

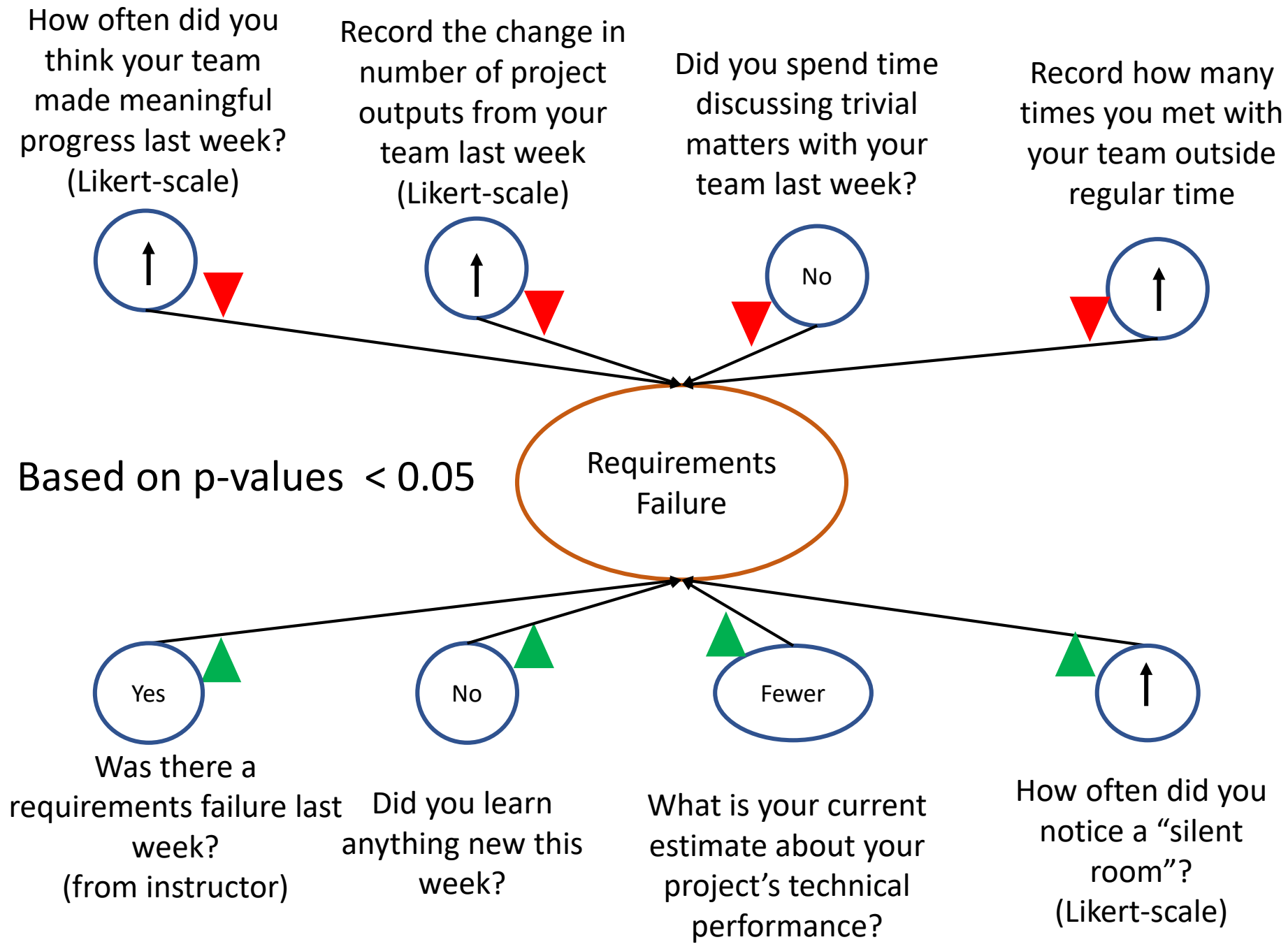
$$\log P[\hat{Y}_{i,t+1} = Fail|X_{it}] \propto a + bX_{it}{}^{T} + u_i + \varepsilon_{it}$$

Error of model compared to reality

Current Logistic Regression Model

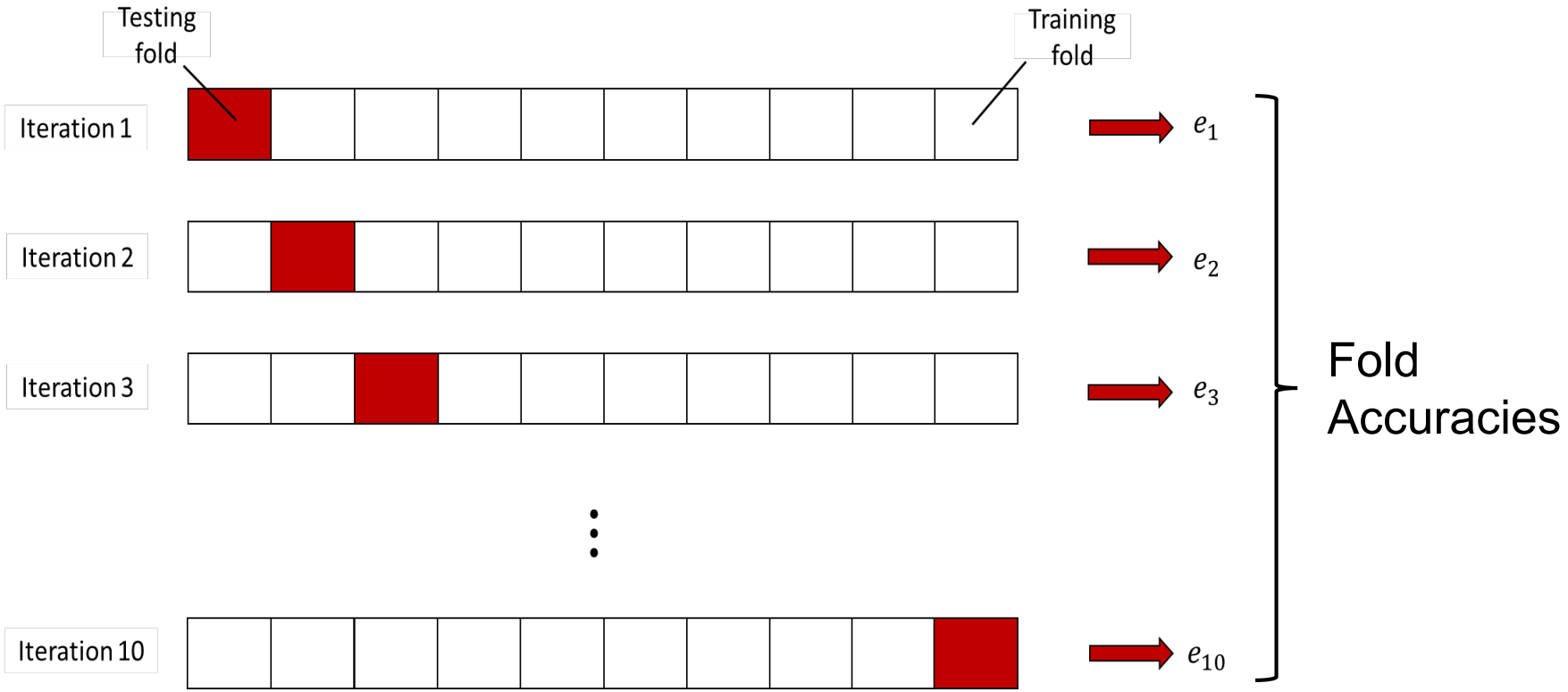
Prediction of project outcome:



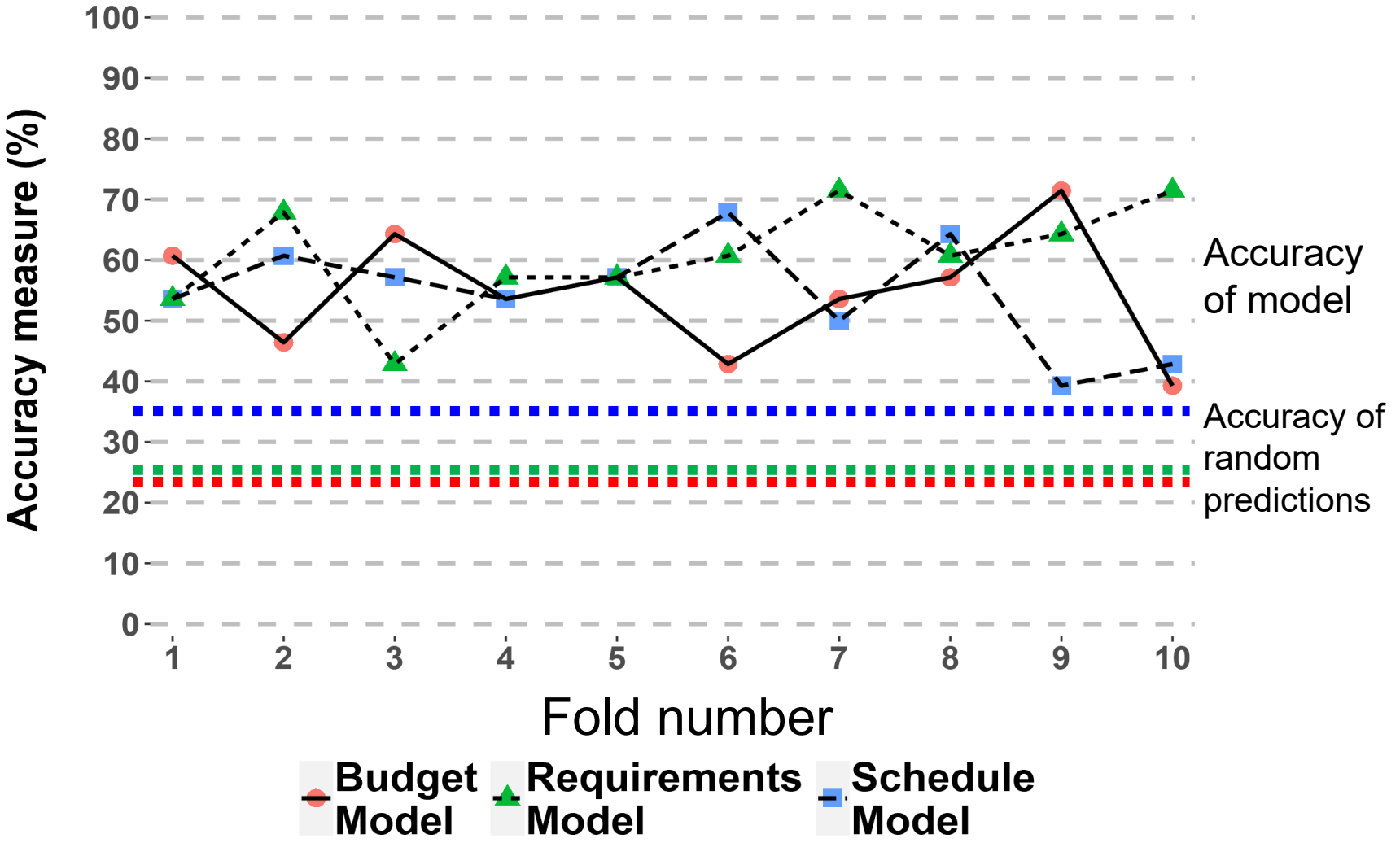

How often did you think your team made meaningful progress last week? (Likert-scale)
Record the change in number of project outputs from your team last week (Likert-scale)
Did you spend time discussing trivial matters with your team last week?
Record how many times you met with your team outside regular time
No
Based on p-values < 0.05
Requirements Failure
Yes
No
Fewer
Was there a requirements failure last week? (from instructor)
Did you learn anything new this week?
What is your current estimate about your project's technical performance?
How often did you notice a "silent room"? (Likert-scale)

k-fold cross-validation to gauge the ability of the models to generalize

- *K*=10 folds
  - Find accuracy in binary predictions (failure/no failure) from unknown data



Budget and schedule model are on average 54% accurate, requirements model is 60% accurate



Q2: Can we find latent factors for new teams / new students?

Small Datasets: Information is precious, simple models are not data-efficient

- Improving accuracy & obtaining latent factors for new teams requires going from simple to complex:
  - *Simple models (e.g., logistic regression, latent factor models)*
    - *Data-wasteful*
    - *Latent factors cannot be used in new teams / new students*
  - *Complex models (e.g., neural network factor models, boosted decision trees)*
    - *Data-efficient*
    - *Latent factors on new teams / new students*

Simple models

e.g., logistic regression, (Spearman's) latent factor models

Cons (simple):

- *Data-wasteful*: Cannot capture all information from data
- Either no latent factors or cannot apply method to new teams

Pros (simple):

- Will not overfit the training data
- Parameters can be "interpreted"
  - Not quite if model is wrong (misspecified)

Complex models

e.g., neural factor models

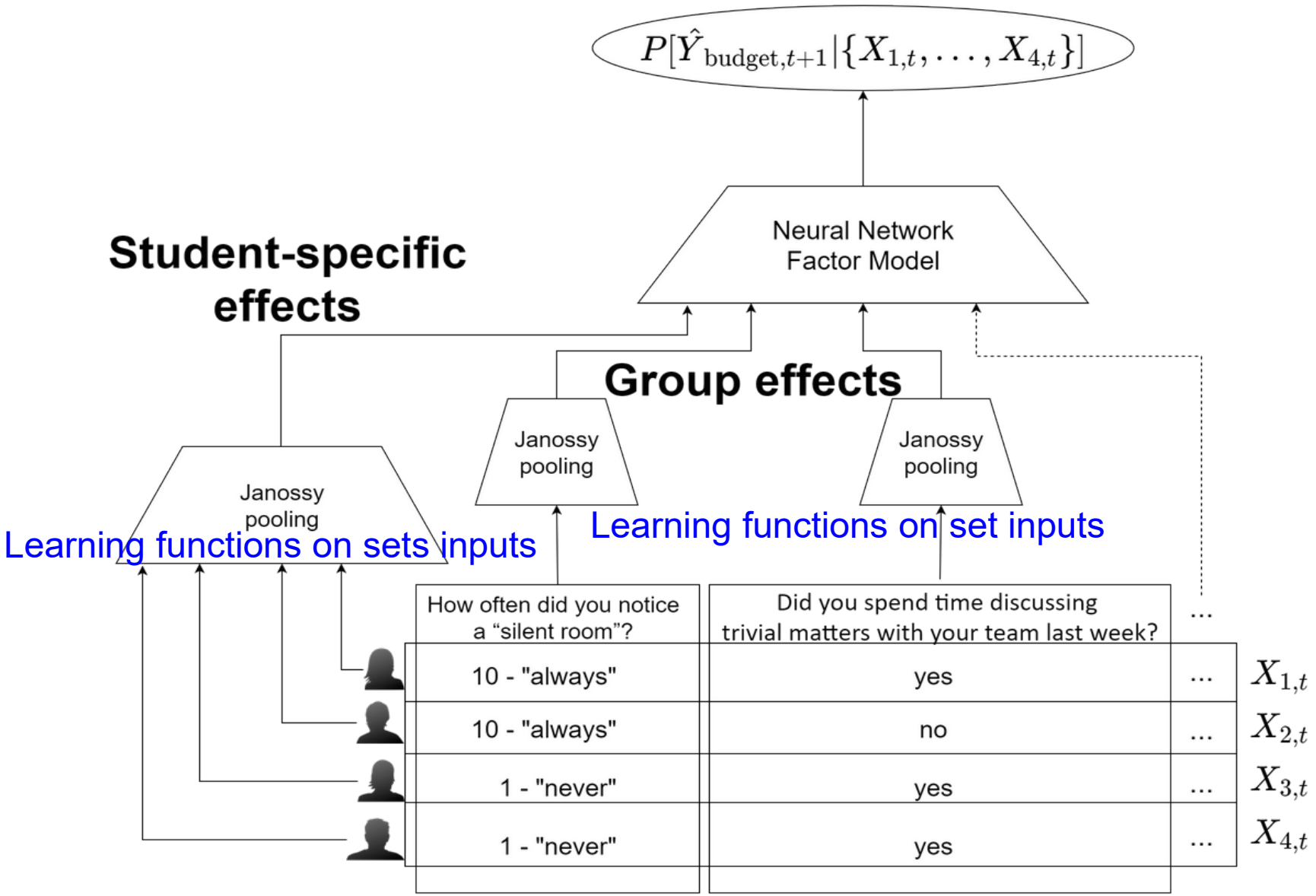
Cons (complex):

- *Will overfit the data*
  - Less of an issue with modern optimization methods
- Interpretation can be more challenging

Pros (complex):

- *Data-efficient*: Will extract most information from data
- Can obtain latent factors directly applied to new students, new classes

Advanced Model



Murphy, Srinivasan, Rao, Ribeiro, 2019, "***Janossy pooling**: Learning deep permutation-invariant functions for variable-size inputs*," International Conference on Learning Representations (ICLR), May 2019.

Logistic vs Advanced Model

Advanced Model is More Accurate than Logistic Regression

- Single task = just predict either budget, requirements, **or** schedule
- Multi-task = jointly predict budget, requirements, **and** schedule

Advanced model generalizes better to test data
(results from 5-fold cross-validation)

- Multi-task learning better than predicting single task

- Obtains latent "root cause" factors for new students

Project Failure	Logistic Regression (Single Task)	Our Model (Single Task)	Our Model (Multi-task)
Budget	0.642 ± 0.080	0.689 ± 0.09	$\mathbf{0.729 \pm 0.068}$
Schedule	0.523 ± 0.072	$\mathbf{0.586 \pm 0.068}$	0.580 ± 0.041
Technical Requirements	0.580 ± 0.062	0.643 ± 0.035	$\mathbf{0.688 \pm 0.088}$

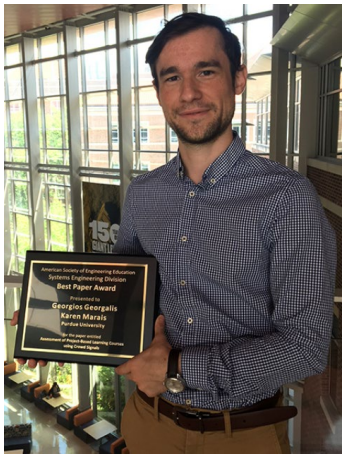
$\pm$ standard deviation

Year 1 resulted in 5 publications overall from the research team and one best conference paper award

1. Georgalis, G and Marais, K 2019, “Can we use Wisdom-of-the-Crowd to Assess Risk of Systems Engineering Failures?” ***INCOSE 2019 International Symposium***, Orlando, FL, July 2019.
2. Georgalis, G and Marais, K 2019, “Assessment of Project-Based Learning Courses using Crowd Signals.” ***ASEE 2019 Annual Conference & Exposition***, Tampa, FL, June 2019.
   Selected as the ASEE Best Overall PIC Paper
3. Murphy, R, Srinivasan, B, Rao, V and Ribeiro, B 2019, “Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs,” **International Conference on Learning Representations (ICLR)**, 2019, New Orleans, LA, May 2019.
4. Murphy, R, Srinivasan, B, Rao, V and Ribeiro, B 2019, “Relational Pooling for Graph Representations,” ***International Conference on Machine Learning (ICML)***, Long Beach, CA, June 2019.
5. Meng, C, Yang, J, Ribeiro, B and Neville, J 2019, “HATS: A Hierarchical Sequence-Attention Framework for Inductive Set-of-Sets Embeddings” ***ACM SIGKDD Conference on Knowledge Discovery and Data Mining***, Anchorage, AK, August 2019.

ASEE Annual Conference and Exposition 2019

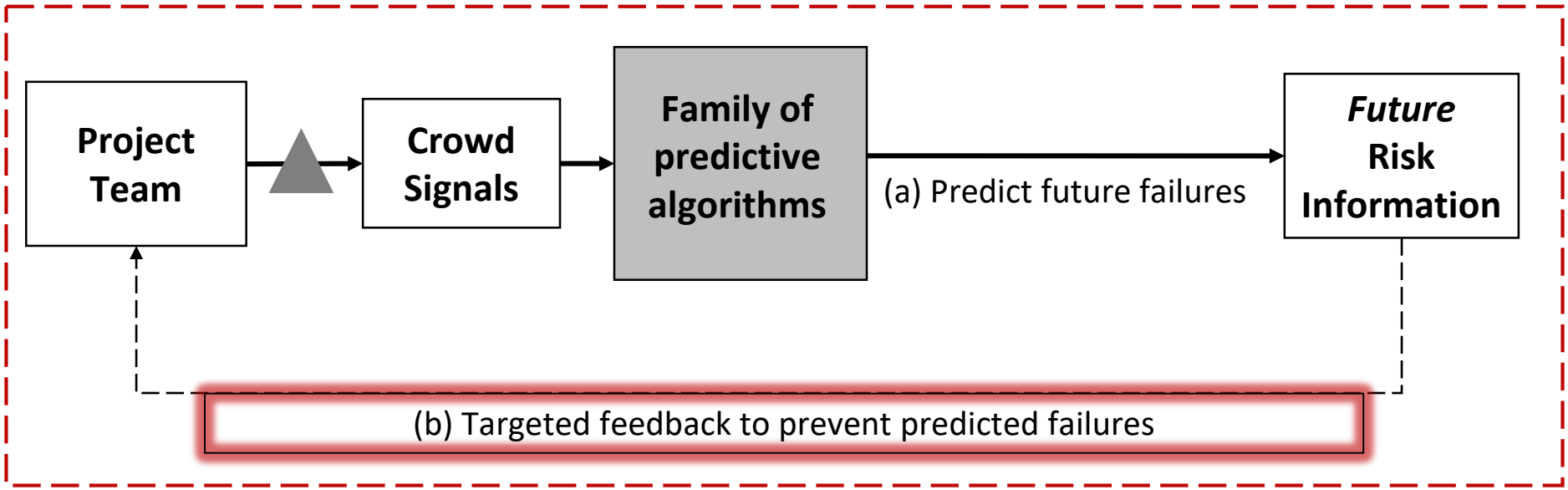
In this paper we compared the frequency of certain types of systems engineering failure causes occurring in industry vs. in student design projects.



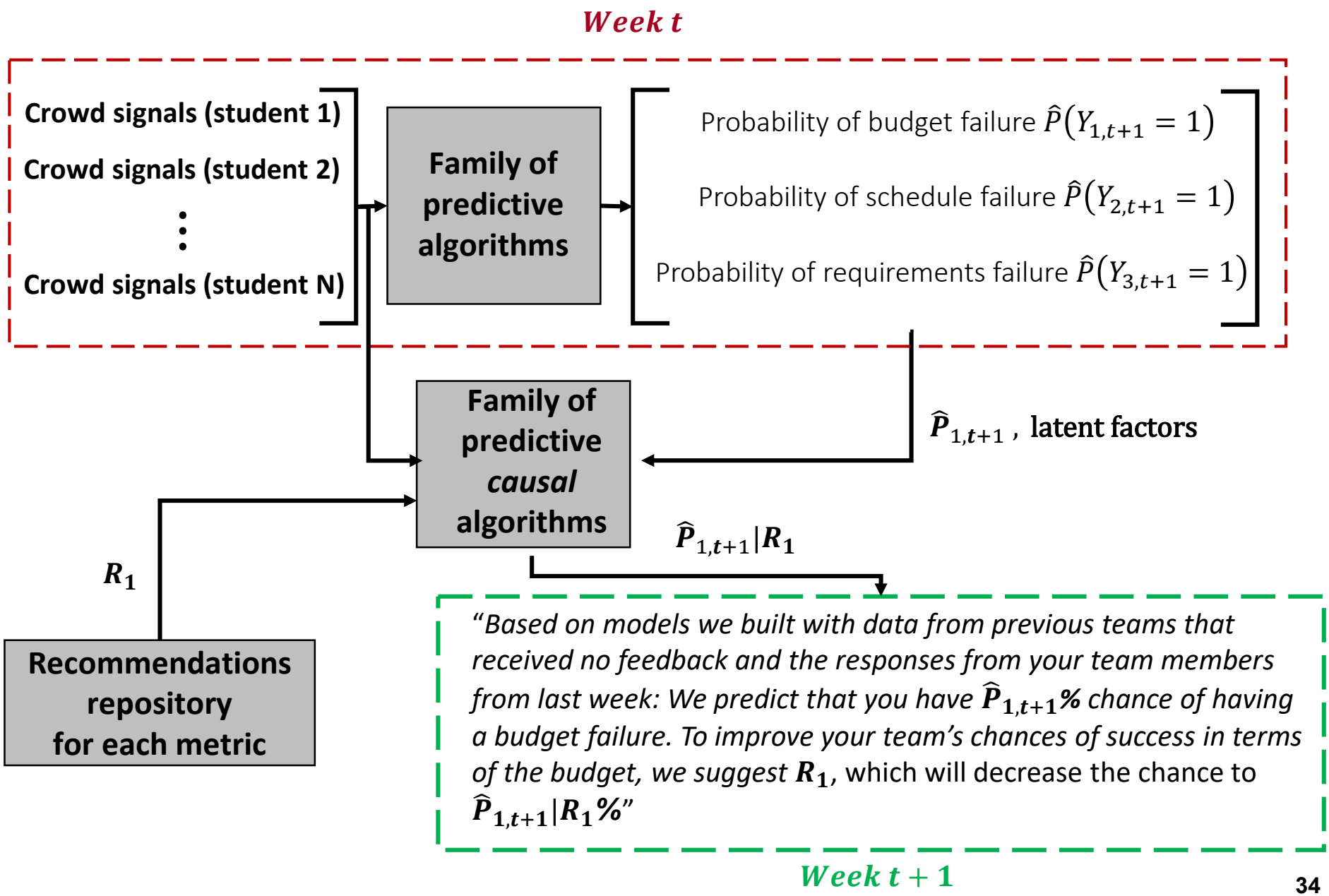
Georgios Georgalis is part of the research team under Dr. Karen Marais

Year 2 Ideas

We are currently working on the feedback process



- Feedback has two parts
  - A prediction of budget, schedule, and requirements failure
  - Recommendations to prevent them (from a repository)
- But
  - Logistic regression/neural latent factor model captures correlation only
    - Recommendations are causal
  - Under multiple causes, causal model could be learned from our data
    - "The blessing of multiple causes" [Wang & Blei, JASA 2019]


Week t
Crowd signals (student 1)
Crowd signals (student 2)
Crowd signals (student N)
Family of predictive algorithms
Probability of budget failure $\hat{P}(Y_{1,t+1} = 1)$
Probability of schedule failure $\hat{P}(Y_{2,t+1} = 1)$
Probability of requirements failure $\hat{P}(Y_{3,t+1} = 1)$
$\hat{P}_{1,t+1}$, latent factors
Family of predictive *causal* algorithms
$\hat{P}_{1,t+1}|R_1$
R_1
Recommendations repository for each metric
"Based on models we built with data from previous teams that received no feedback and the responses from your team members from last week: We predict that you have $\hat{P}_{1,t+1}$% *chance of having a budget failure. To improve your team's chances of success in terms of the budget, we suggest* R_1, which will decrease the chance to $\hat{P}_{1,t+1}|R_1$%"
Week t + 1

Reducing the prediction model input size

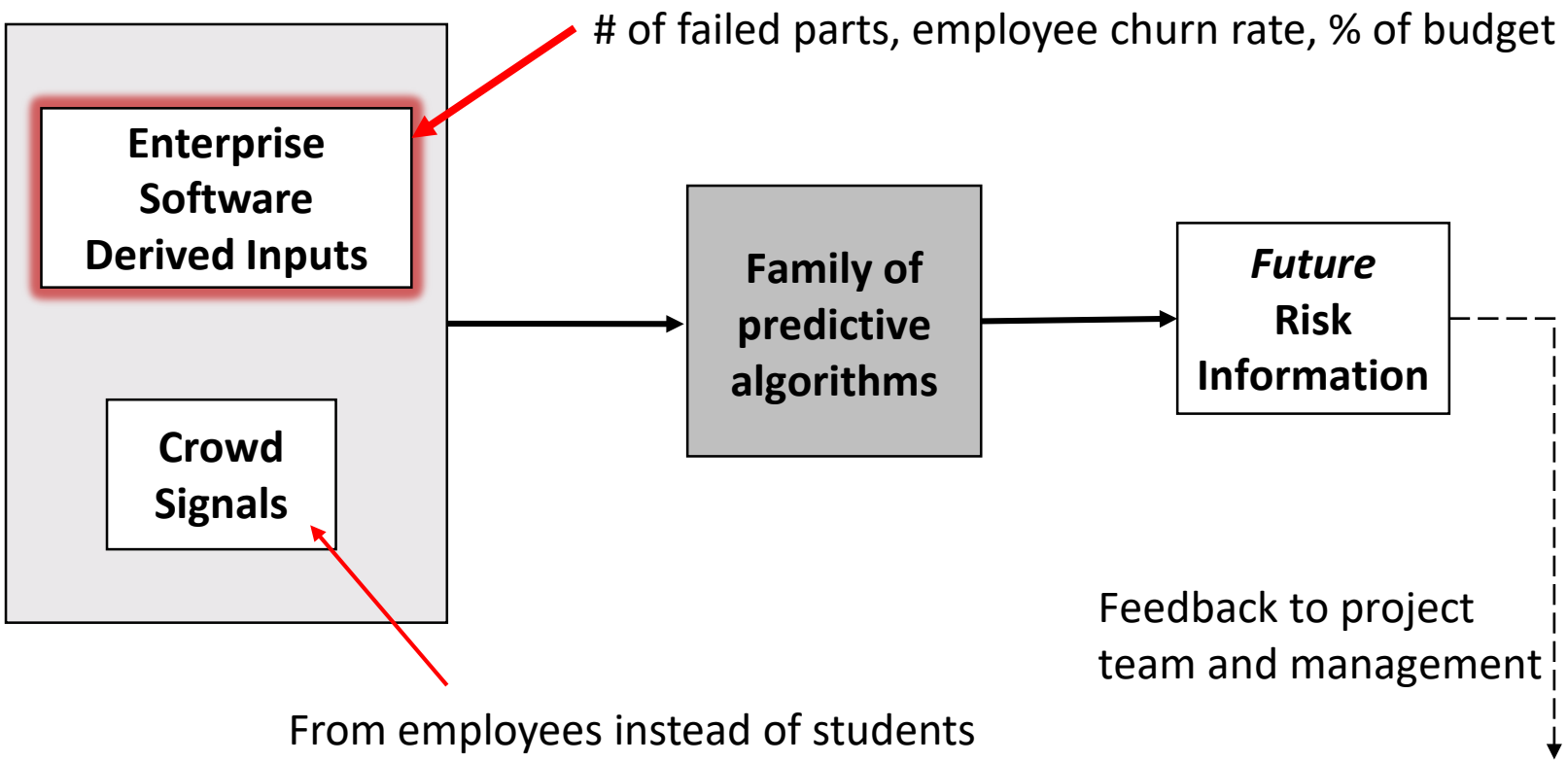
- Simple **logistic regression** needs 49 questions as the input to predict the failures with reasonable accuracy

Would like to reduce the number of questions we ask:

- Reduce the burden on the respondents
- Make it more “realistic” as an application
- We only care about the questions that matter (correlate with failure)
- It will require advanced models
- It will require causal models

- To accomplish this we are currently doing the following analyses:
  - Feature selection: get feedback from model on which questions are really relevant
  - Study advanced models that can still work well with fewer question
  - Study feature selection in causal models (new machine learning tools)

For Year 2 we propose to partner with an organization to expand our inputs and test in “real” projects



- We will deliver value to our partners by leveraging the predictive capability of our prototype. Our process gives us the opportunity to provide feedback to decision makers, alerting them of upcoming failures, and suggesting corrective actions