Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process

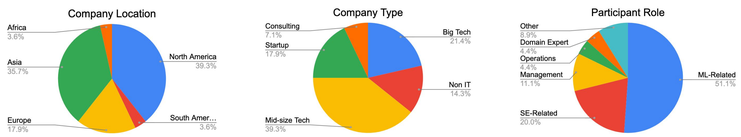
Distinguished Paper Award

Nadia Nahar*[1], Shurui Zhou[2], Grace Lewis[1], Christian Kästner[1]
[1] Carnegie Mellon University, PA, USA; [2] University of Toronto, Canada
Contact: nadian@andrew.cmu.edu

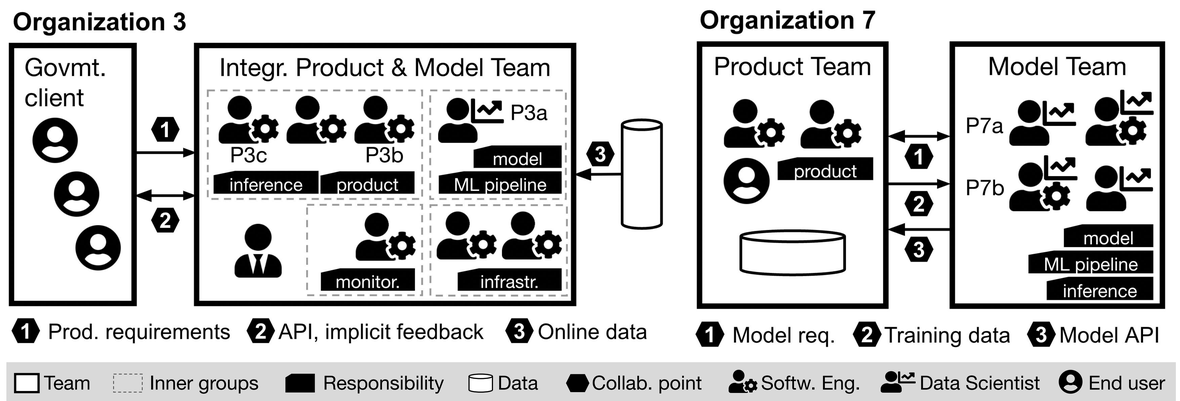
Motivation:
Machine Learning (ML) models are **components within a system.** Building such ML-enabled systems is hard and requires collaboration between different roles, including software engineers, data scientists, domain experts, and managers.

Research Question:
What are the **collaboration points** and corresponding **challenges** between data scientists and software engineers in building ML-enabled systems?

Research Approach:
We conducted **45** interviews in **28** organizations; qualitative analysis



Key Findings:
Every project is different, but problems often **at the boundary between teams**. Challenges occur in different patterns.



Inadequate Data

Data scientists have little influence over data quality and quantity

- Pay special attention to data; Budget for data collection, processing; or even dedicated data team
- Adopt a more formal contract, specifying data quantity and quality expectations

Issues of Understanding Data

Struggling with data understanding and having difficulty getting help

- Plan on accessing domain experts, establish data documentation practices

Cannot Handle Evolving Data

Undesired and unannounced changes in data sources break the data assumptions

- Offer engineering support on monitoring infrastructure to detect changes in data

Collaboration Points — *Themes*

Requirements and Planning
- Product and Model Requirements
- Project Planning

Training Data
- Negotiating Data Quality and Quantity

Product-Model Integration
- Responsibility and Cultural Clashes
- Quality Assurance for Model and Product

Communication · Documentation · Engineering · Process

Lack of AI Literacy

Lack of ML literacy leads to unrealistic requirements

- Involving data scientists early when soliciting product requirements
- ML literacy for customers and product teams: conducting training sessions

Lack of Process

Pursuing a model-first trajectory entirely without considering product requirements is problematic

- Emphasis on collaboration during requirements phase, more research on process needed

Ignored Qualities

Rarely considers any qualities other than accuracy, ignoring qualities such as latency or scalability

- Emphasis on broader system context, and plan more interaction with the product team

Look into our paper for more...

Preprint: https://arxiv.org/pdf/2110.10234.pdf
Talk: https://www.youtube.com/watch?v=EwJ1Hx4F6DY

/NadiaIIT /nadia-nahar-iit