

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

Nadia Nahar

nadian@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Shurui Zhou

University of Toronto
Toronto, Ontario, Canada

Grace Lewis

Carnegie Mellon Software Engineering Institute
Pittsburgh, PA, USA

Christian Kästner

Carnegie Mellon University
Pittsburgh, PA, USA

44TH International Conference On Software Engineering (ICSE 2022)

発表者：井町 瑠貴

背景

機械学習(ML)は、研究や実務において大きな注目と資金提供を受けている

プロトタイプのMLモデルから本番システムに移行することは非常に困難である

- 2019年：MLプロジェクトの87%が失敗すると報告^[107]
- 2020年：53%がプロトタイプから生産に至らない^[69]

[107] Staff, V. B. 2019. Why do 87% of data science projects never make it into production?

URL: <https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/>.

[69] Panetta, K. 2020. Gartner Identifies the Top Strategic Technology Trends for 2021.

URL: <https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021>.

背景

| 人的要因に課題があると考えられる

- データサイエンスとソフトウェアエンジニアリングの作業の分担と調整
- インターフェースと責任の調整と文書化
- システムの運用と進化の計画 など

本研究では、実務者へのインタビューを通じて、MLシステムを構築し導入する際にチームが直面するコラボレーション課題を明らかにする

アプローチ

文献調査を基にインタビューを作成

- キーワード検索と独自の知識によって15の論文を選択
- 9つのコラボレーション領域に凝縮し，インタビューガイドを作成

インタビューの実施

- 28の組織から参加した45名に対して，30～60 分の半構造化インタビューを実施

アプローチ

結果を文献的知見と三角測量

- インタビューで浮上したトピックのキーワード検索や、後方・前方への雪だるま式アプローチ
- 300以上の論文を読み、61の論文を関連性があると判断

インタビュー対象者との妥当性の確認

- 以下の2点を尋ねた
 - 正しさや同意・不同意があるか（＝適合性）
 - 得られた知見があるか（＝適用性）

組織構造

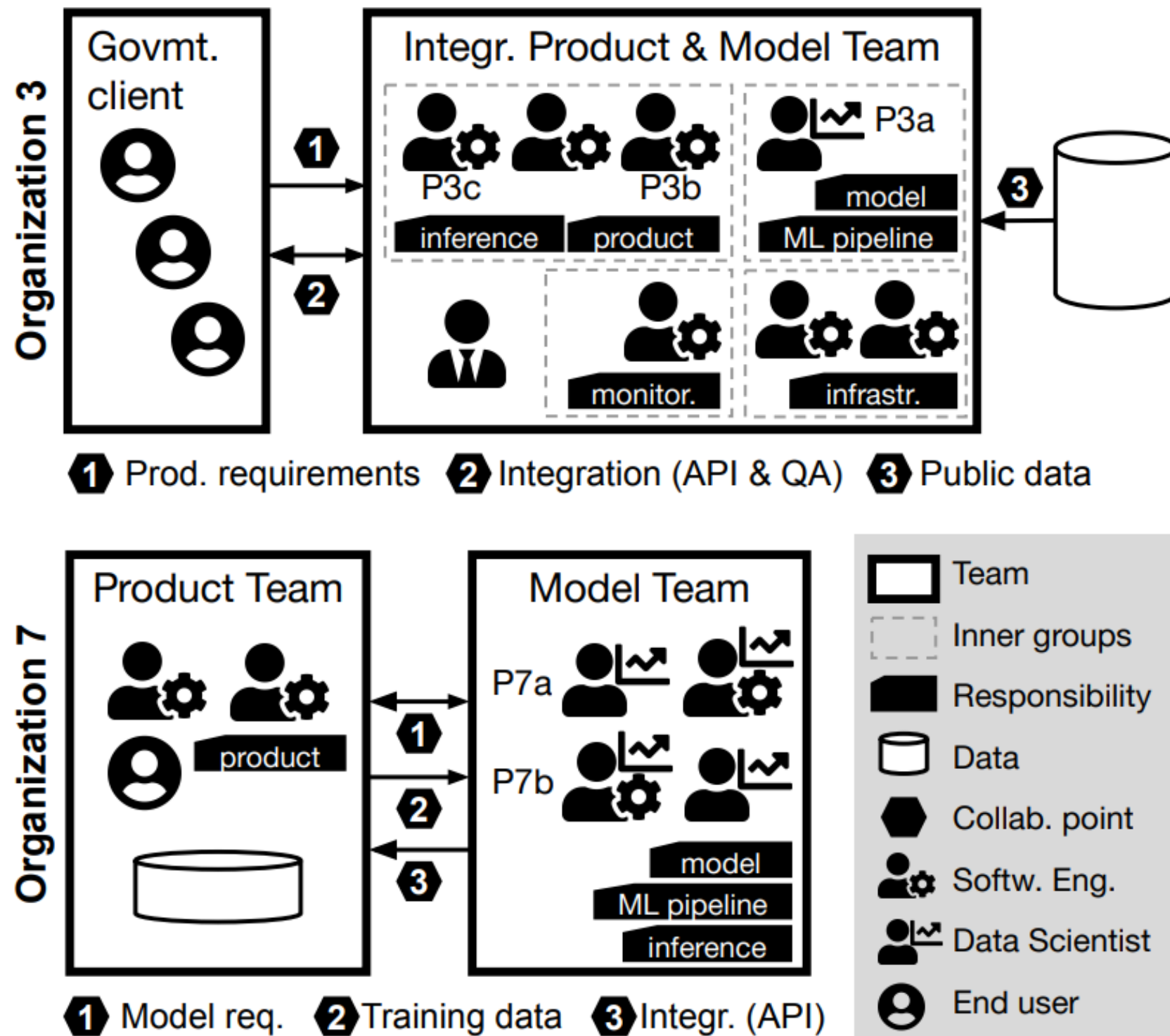


Figure 1: Structure of two interviewed organizations

インタビュー結果

インタビューの結果、以下の3つの側面にパターンが見られた

- 要件とプロジェクト計画
- トレーニングデータ
- プロダクトとモデルの統合

要件とプロジェクト計画

プロダクトの要件はモデルチームからの
インプットを必要とする

- MLの機能を十分に理解せずにプロダクトの要件を引き出すことは困難
- データサイエンティストはモデル能力に関する非現実的な期待に直面することが多い

モデル要件が不明確なままモデル開発を
行うことはよくある

- MLモデルに対して明確に要求をすることは難しい

要件とプロジェクト計画

提供されるモデルの要件が、精度やデータの安全性を超えることはほとんどない

- 与えられた要件は、主に精度の概念に関連
- データセキュリティとプライバシーの要件は、データ所有者や法的要件によって課される
- 公正性と説明可能性の要求はほとんどない

MLの不確実性は今後の見通しを難しくする

- 労力、コスト、精度に関する期待値や契約内容を設定することが難しい

トレーニングデータ

提供されたデータ，公開されたデータは不十分なことが多い

- 受け取ったデータは低品質であることが多く，データクリーニングに多大な投資が必要
- 提供者は，データ品質の重要性を理解していない

データの理解と専門家へのアクセスはボトルネックである

- 既存のデータ文書は，データを理解するのに十分でない場合がほとんどである

トレーニングデータ

進化するデータを扱う必要性

- データは常に提供されなければならないことが理解されていないことがある
- データソースが突然変更される場合があり、長期的な一貫性の確保に苦勞する

社内の優先順位やセキュリティ上の懸念により、データアクセスが阻害されることが多い

- モデルチームはデータを要求するための交渉力をほとんど持たない

プロダクトとモデルの統合

| 負う責任が能力や好みと一致しない

- データサイエンティストは、自分ですべてを行うよりもエンジニアリングのサポートを好む

| 技術的な専門用語がコミュニケーションの障害になる

- 例：パフォーマンス
 - データサイエンティスト：予測精度
 - ソフトウェアエンジニア：応答時間

プロダクトとモデルの統合

モデルの目標を設定するのが難しい

- 精度以上の品質要件が提示されることがない
- そのため、待ち時間、メモリ消費量、公平性のテストに対して責任を感じない

システムテストの責任の所在が不明確

- モデルチーム
 - プロダクトの品質に関する責任は負わず、責任はモデルを提供することにある
- プロダクトチーム
 - システム全体のテストを計画しておらず、アドホックなシステムテストを実施しているにすぎない

結論

| インタビューを通じて、3つの中心的な
コラボレーションポイントを特定した

- 要件とプロジェクト計画
- トレーニングデータ
- プロダクトとモデルの統合

| エンジニアリングの実践と研究の両面で
もっと注目されるべき4つの広範なテーマが
あると考える

結論

コミュニケーション

- ソフトウェアエンジニアやマネージャーのためのMLリテラシーを含む教育が重要
- データサイエンティストを訓練することも必要

ドキュメンテーション

- チーム間で期待されることを明確に文書化することが重要
- モデル要件, データ期待値, および保証されたモデルの品質を文書化するための手法は十分に確立されていない

結論

エンジニアリング

- 組織は、モデルを製品として運用・維持するために必要なエンジニアリングの負担を過小評価している
- プロジェクトのML部分と非ML部分に十分なエンジニアリング能力を持たせ、プロダクトと運用の考え方を育成する必要がある

プロセス

- ML対応システムのための統合プロセスライフサイクルに関するより多くの研究が必要である

所感

- 優秀論文賞だったので読んでみた
- 理想的に考えると、モデルとプロダクトの両方の知識を持つことが重要（必須？）であるが、簡単に解決できる問題ではないと感じた
- 自分が仕事をする際にも、このような状況に遭遇するかもしれないと考えると、今のうちに広い知識を身につけたいと感じた

補足資料

補足 | 参加者の分布

Company Location	Count
North America	11
South America	1
Europe	5
Asia	10
Africa	1
= 28	

Current Position	Count
ML-Related	23
SE-Related	9
Management	5
Operations	2
Domain Expert	2
Other	4
= 45	

Company Type	Count
Big Tech	6
Non IT	4
Mid-size Tech	11
Startup	5
Consulting	2
= 28	

