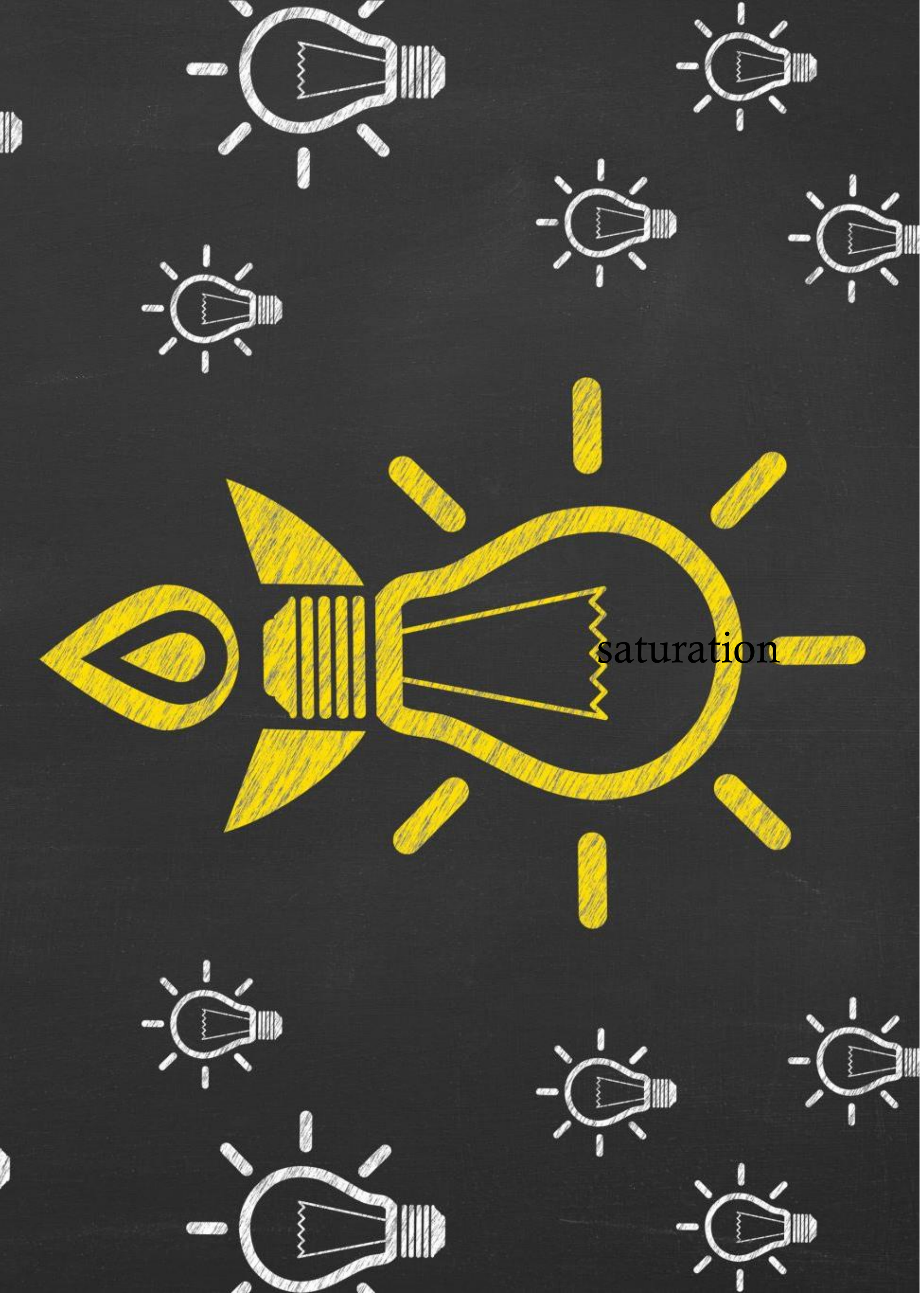


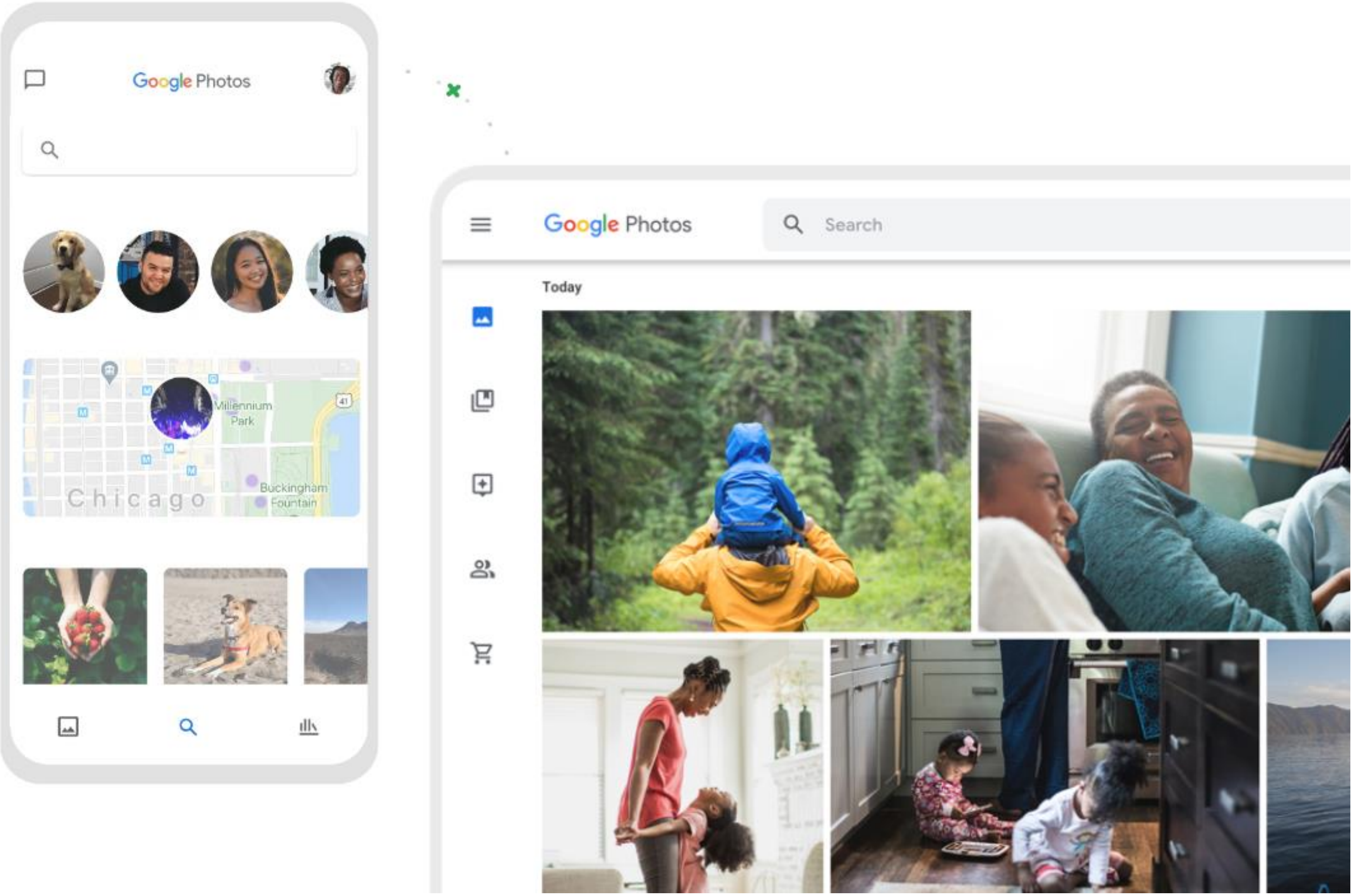
COLLABORATION CHALLENGES IN BUILDING PRODUCTION MACHINE LEARNING SYSTEMS

Research Project Proposal
17803 – Empirical Methods

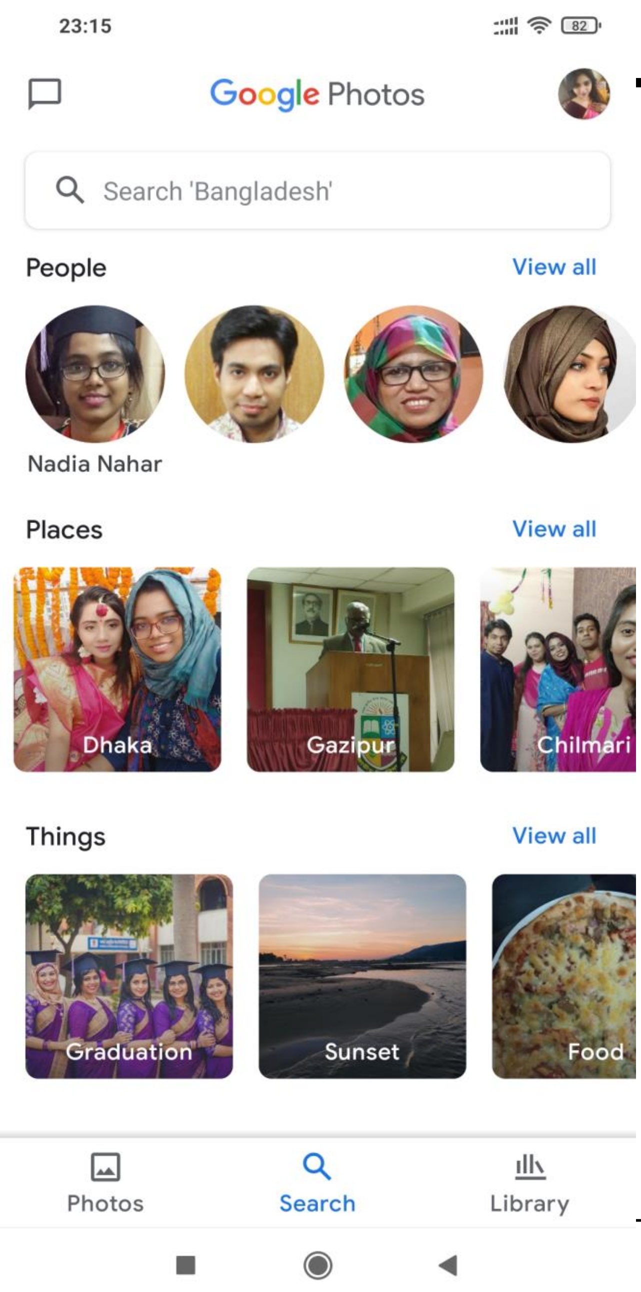
Presenter: Nadia Nahar

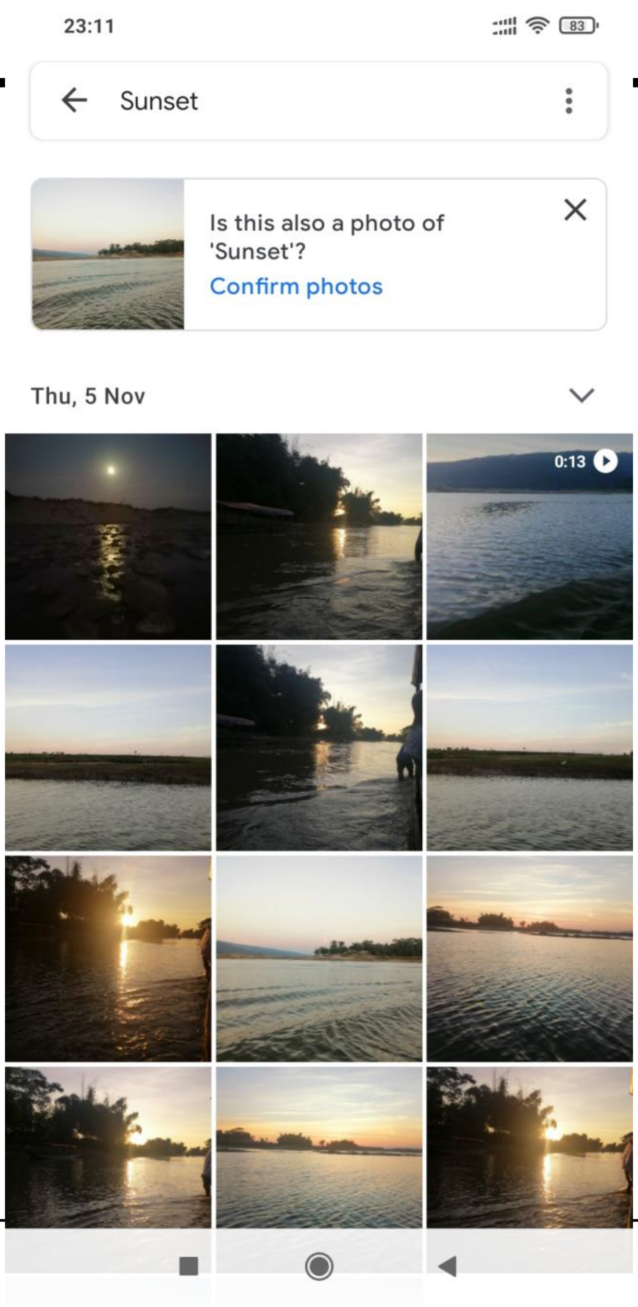
WHAT DO WE MEAN BY ML SYSTEMS?

Google Photos

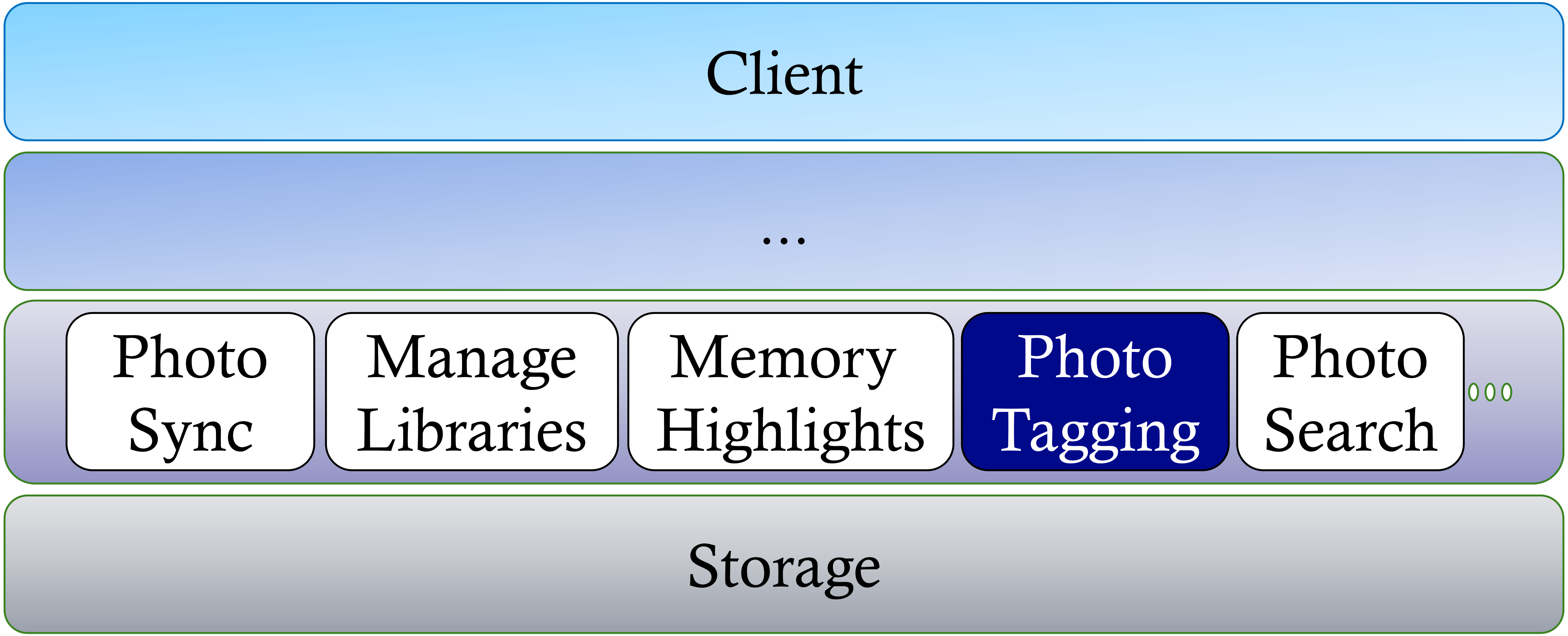


https://www.google.com/photos/about/





Is this also a photo of 'Sunset'?



Client

…

Photo Sync

Manage Libraries

Memory Highlights

Photo Tagging

Photo Search

Storage

MOTIVATION

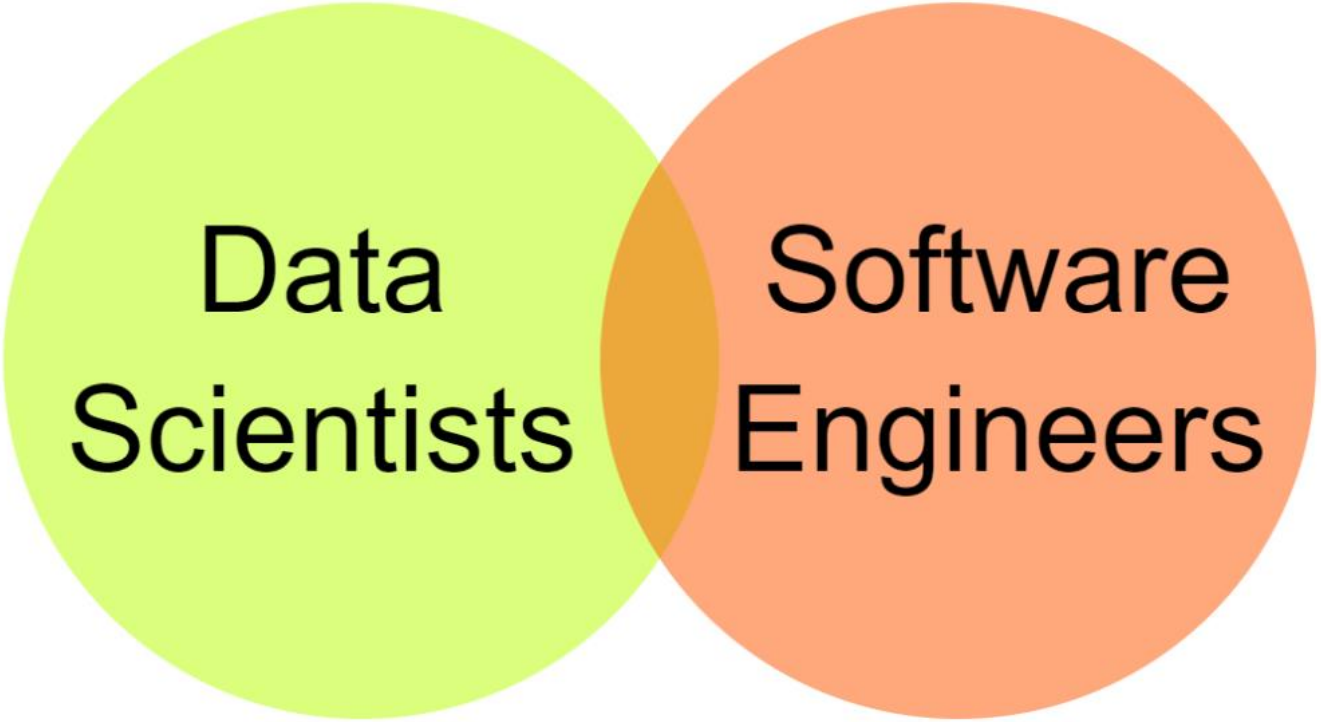
JOURNEY

- Have been exploring papers in the domain of SE4ML
- Started with notebook papers
- Went into the papers that talks about the **software view** rather than the **model view**

JOURNEY

- Found scattered mentions about the **challenges** here and there

  - Establishes that ML-project are **different** from traditional SE project
  - Talks about challenges of uncertainty, code integration, difference in priority, problem of communication due to different language jargons, etc.



and Domain specialists + Operators + Business team + Project managers + Designers, UI Experts + Safety, security specialists + Lawyers + Social scientists + ...

https://github.com/ckaestne/seai/tree/F2020/lectures

WHY IS THIS HARD?

HIGH-LEVEL THEORY

“Projects Containing Machine Learning Parts Are Different From Traditional SE Projects, And Raises Additional Challenges in Collaboration Between Different Roles.”

Gap: We don’t have enough understanding of the challenges like why, how, who, etc.

Hook: All the stakeholders related to the software having machine learning components.

RESEARCH QUESTION

- How do data scientists and software engineers collaborate when building production-level machine learning systems?

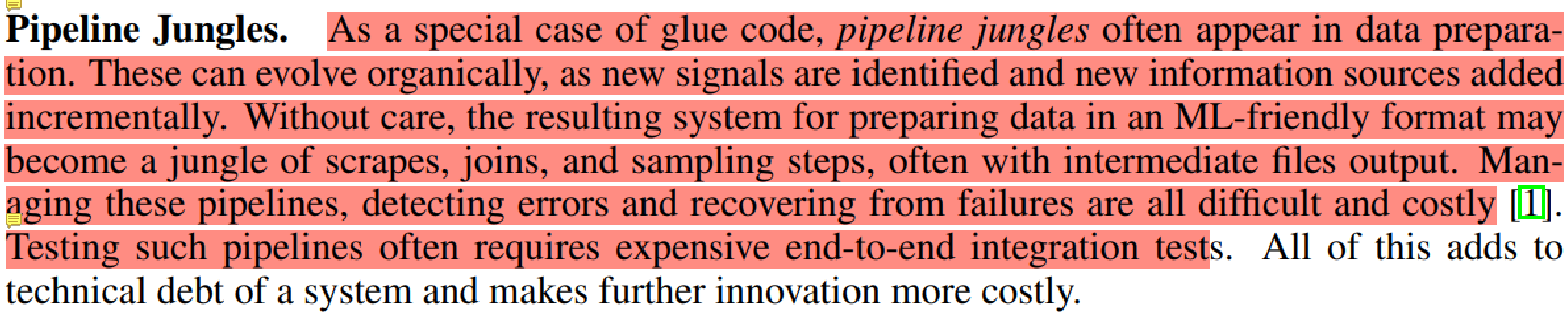
  - What do they collaborate on?
  - What other stakeholders/roles do they collaborate with?
  - What are the collaboration points?
  - What are the challenges in interdisciplinary collaboration?

STUDY DESIGN

- Literature Survey and Coding Challenges in Papers
- Defining the Codebook and Defining Questions to Ask
- Conducting Interview (Qualitative Study)
- Coding Interview Scripts
- Analysis and Discovering Patterns

CODING CHALLENGES IN PAPERS

3 decomposition + 3.4 maintainability + pipeline component -- pipeline jungle makes it difficult manage, detect error, recover from failure



Pipeline Jungles. As a special case of glue code, *pipeline jungles* often appear in data preparation. These can evolve organically, as new signals are identified and new information sources added incrementally. Without care, the resulting system for preparing data in an ML-friendly format may become a jungle of scrapes, joins, and sampling steps, often with intermediate files output. Managing these pipelines, detecting errors and recovering from failures are all difficult and costly [1]. Testing such pipelines often requires expensive end-to-end integration tests. All of this adds to technical debt of a system and makes further innovation more costly.

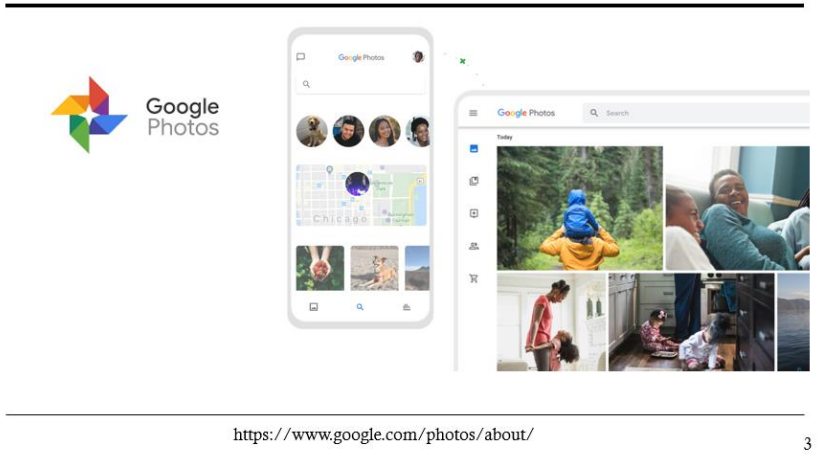
DEFINING THE CODEBOOK AND INTERVIEW GUIDE

- Codebook – https://docs.google.com/document/d/1mk3BW9OaP0cMjM4H03lTVBk0B2XhuTaP6nrb35jF7bg/edit?usp=sharing
- Interview Guide – https://docs.google.com/document/d/1pCmZ0jOcTwobwPx8vu9tO-Ko9t4wn88EPV-byAIuUfs/edit?usp=sharing

INTERVIEW DESIGN

- Maximum Variation Sampling
  - Different Roles
  - Different Regions
  - Different Company Setups
- Quantity?
  - The magic word – “saturation”

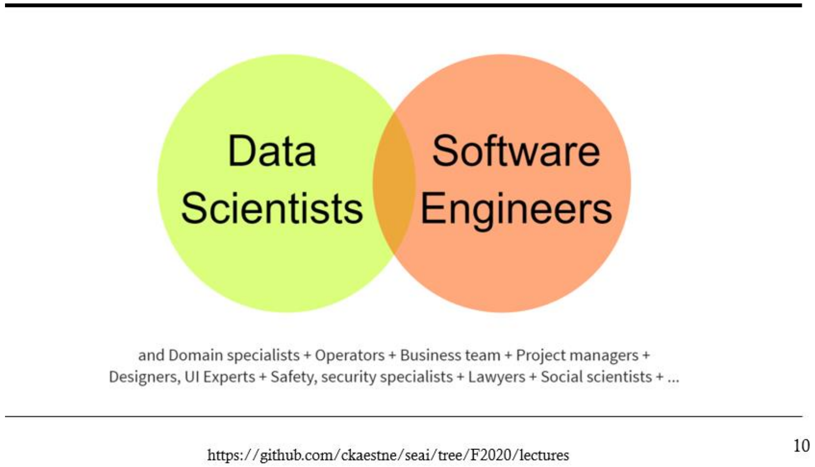
SUMMARY



https://www.google.com/photos/about/

3

Production Machine Learning Systems

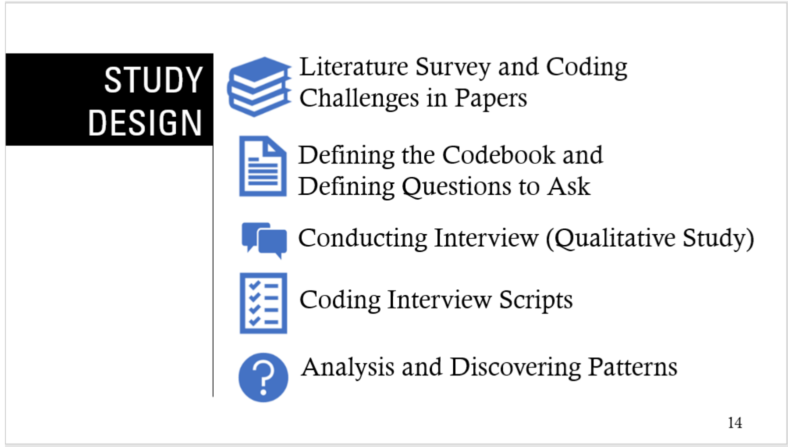


and Domain specialists + Operators + Business team + Project managers + Designers, UI Experts + Safety, security specialists + Lawyers + Social scientists + ...

https://github.com/ckaestne/seai/tree/F2020/lectures

10

Inter-disciplinary Collaboration is Challenging.



STUDY DESIGN

- Literature Survey and Coding Challenges in Papers
- Defining the Codebook and Defining Questions to Ask
- Conducting Interview (Qualitative Study)
- Coding Interview Scripts
- Analysis and Discovering Patterns

14

Research Goal: Understand Collaboration Challenges