Towards an Privacy-Aware Reproducible Machine Learning Pipeline for Open Data

Igor Jakovljevic, CERN & Graz University of Technology
Christian Güetl, Graz University of Technology
Andreas Wagner, CERN

Introduction

- Machine Learning - Development and Growth
- Difficult to reproduce many key results, raising doubts about research methods and publication protocols
- Machine Learning Pipelines - What are they?
- Private and Sensitive Data
- Open Data and Open Science

Why ML and ML Pipelines?

- **Increase Service Adoption** - It is necessary to create innovative features in order to make the service more appealing to users. One of these features is the ability to recommend interesting items to the users and to increase interaction with the system.

- **Research and Innovation** - Fostering innovaton and research is an important aspect of every project, it is necessary to try out and evaluate SOTA solutions quickly and efficiently.

- **User Productivity Boost** - To reduce the time the user needs to find relevant and interesting information it is necessary for the system to automate this task.

Focus of the Paper

- **RQ1:** How can Open Data facilitate the creation of privacy-respecting ML Pipelines?
- **RQ2:** How can Open Data-based ML pipelines be used for the implementation of ML algorithms while ensuring reproducibility using search as the application domain for ML algorithms?

Reproducibility and Open Science

Reproducibility can be defined as the ability to replicate a model that produces the same result as the original model given the same input data

A movement to conduct science transparently by making code, data, scientific communications, and any other research artifact publicly available and easily accessible over the long-term is called Open Science

Machine Learning and Privacy

> *Need to provide **personalized** and evolving **artificial intelligence (AI) services***

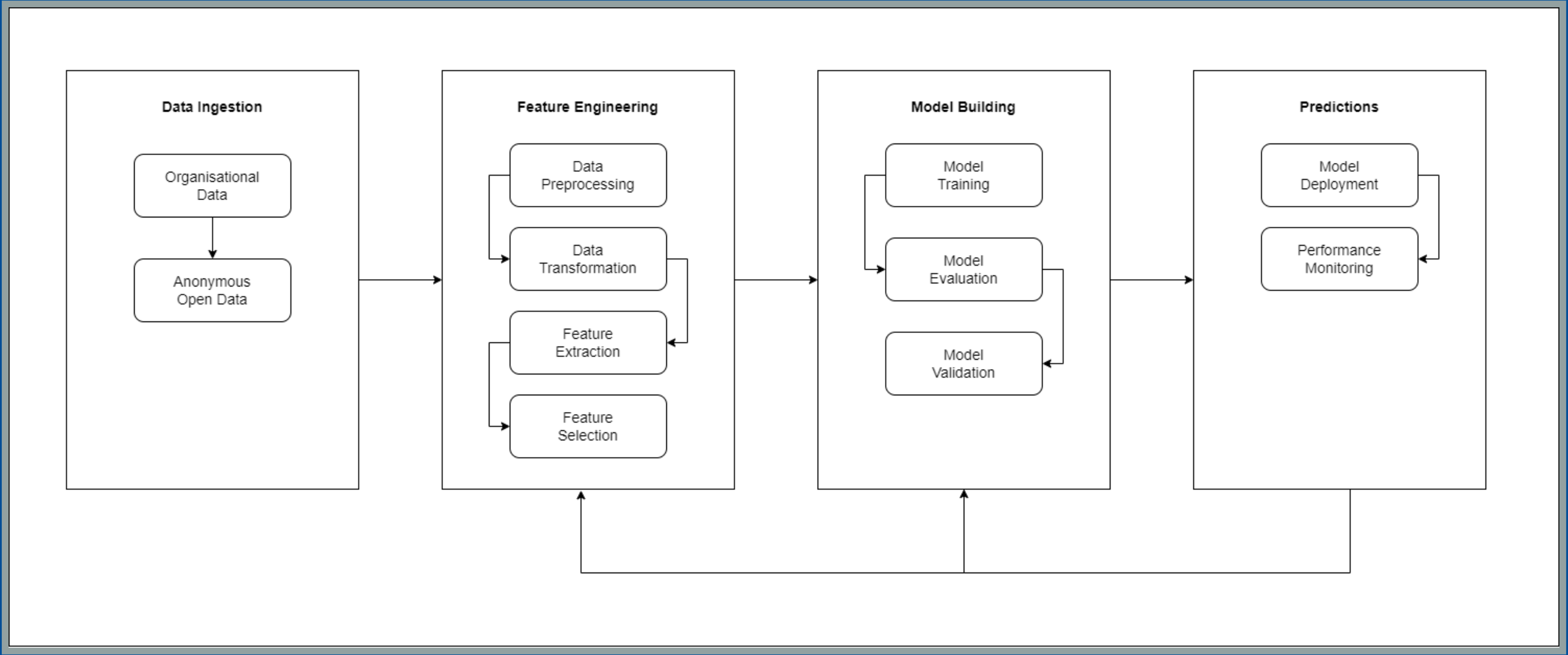
- Diverse Scandals: Netflix Prize, Facebook Data Leak, AOL Query Logs
- Local Differential Privacy and Federated Machine Learning

Machine Learning Pipelines

Main Steps:

- **Data Ingestion**
- **Feature Engineering**
- **Model Building**
- **Predictions**

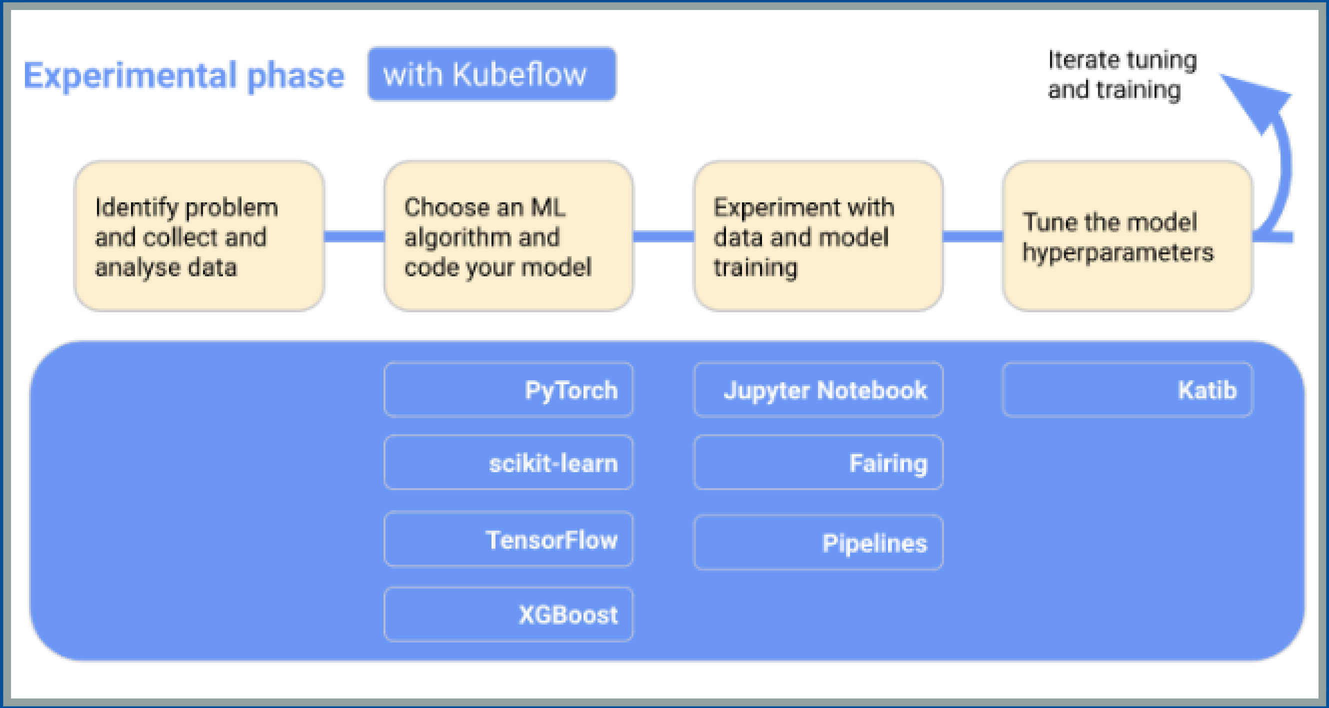
Machine Learning Pipelines



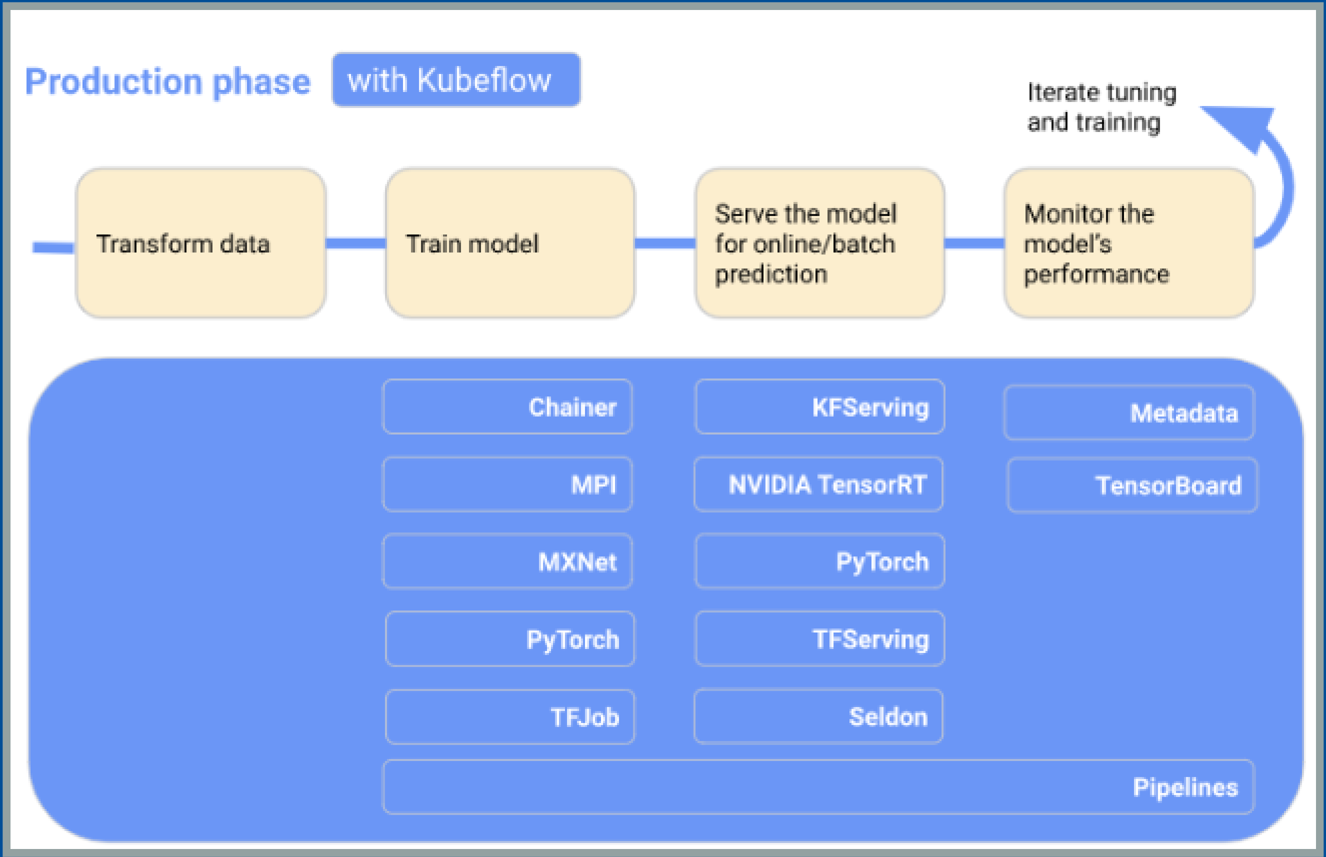
Open Source Software Solutions for ML Pipeline Creation

	Kubeflow	MLFlow	Flyte	ML Run
Open Source	Yes	Yes	Yes	Yes
Language	Python	Python/R/Java	Python	Python
Documentation	Very Good	Good	Poor	Good
Tracking and Versioning	Yes	Yes	Yes	Yes
Pipeline Orchestration	Yes	No	No	Yes
Model Deployment	Yes	No	No	Yes
Scheduler	Yes	No	Yes	No
Dashboard	Yes	Yes	Yes	Limited Functionality

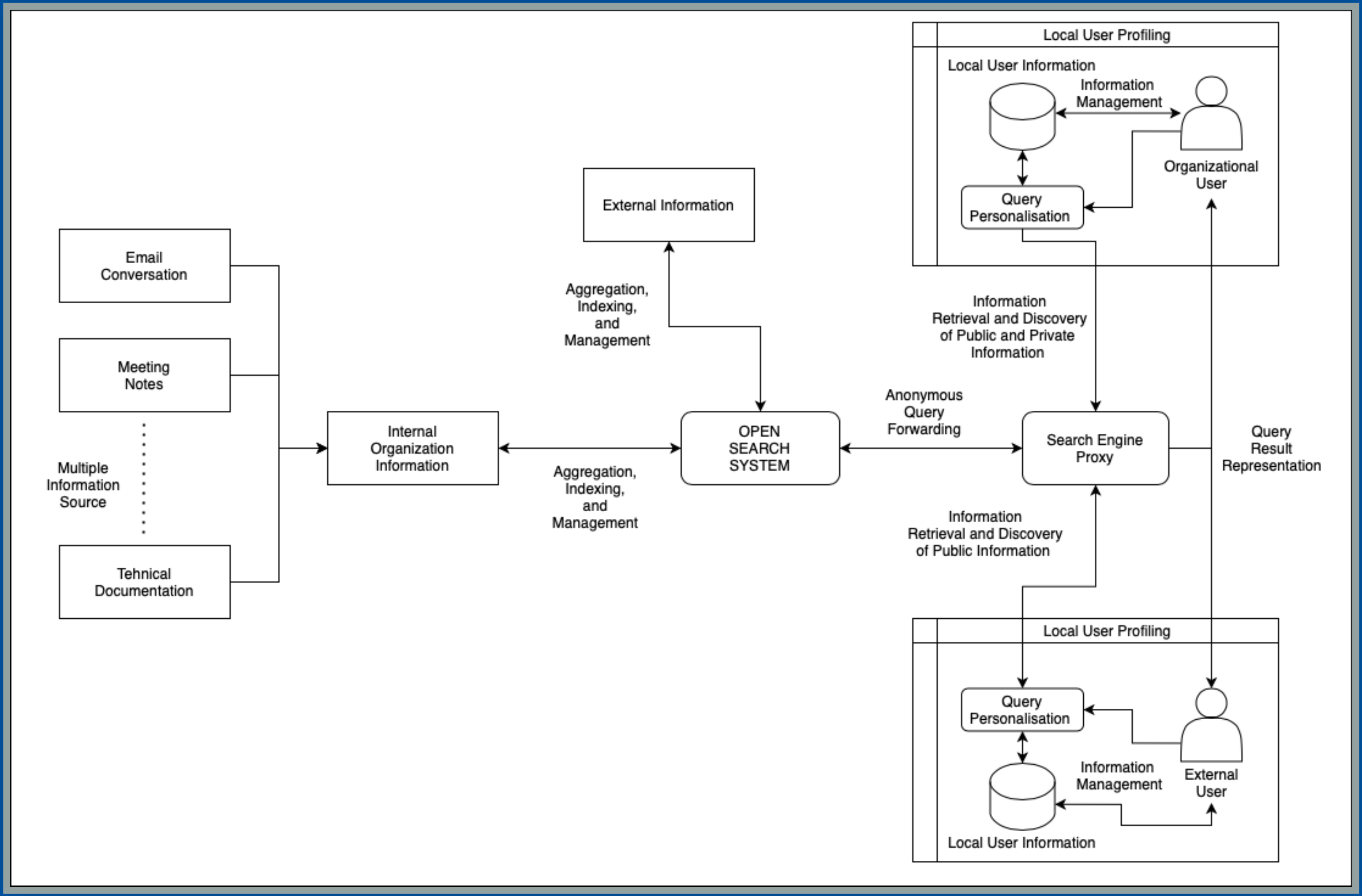
Machine Learning Pipelines - Kubeflow Pipeline



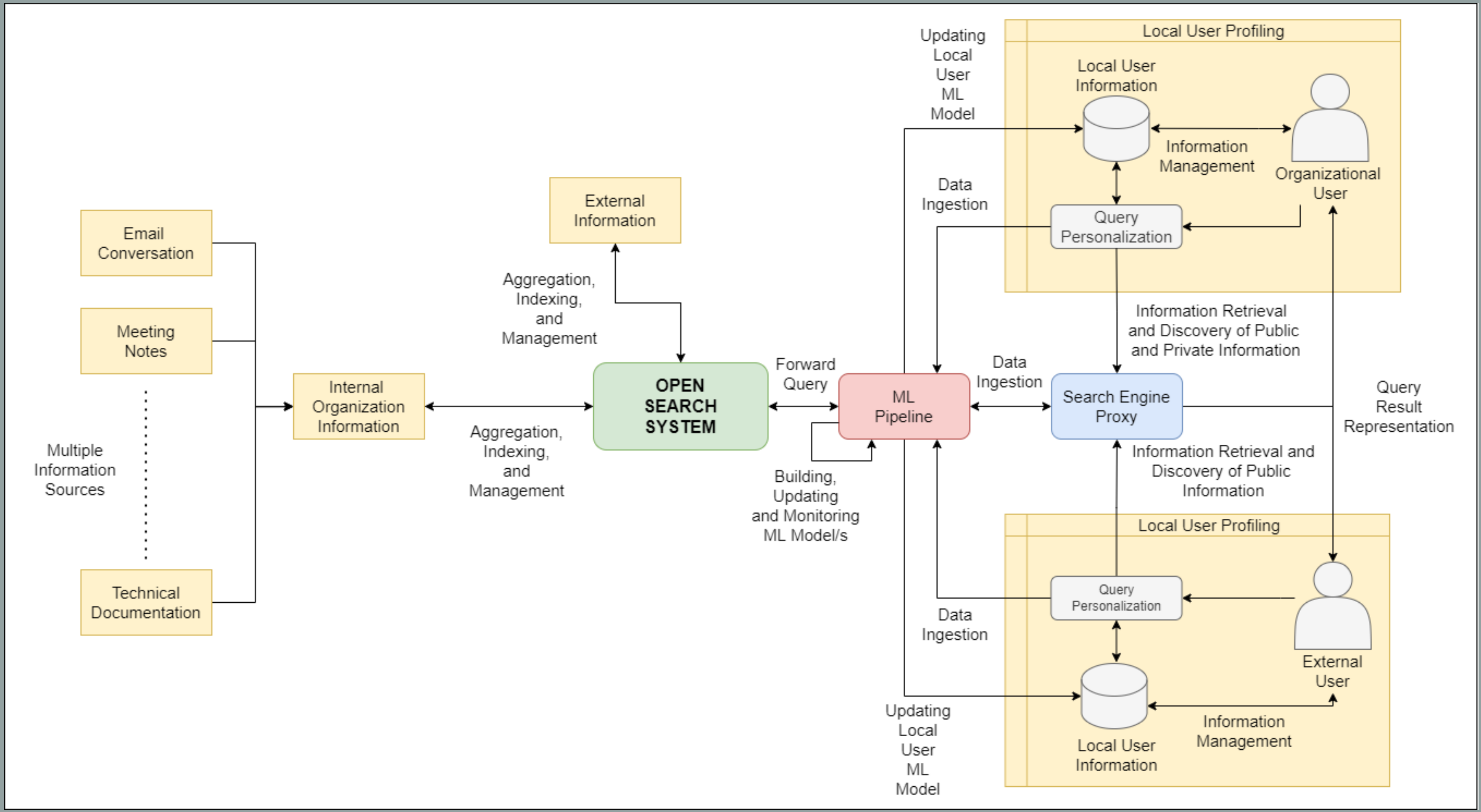
Machine Learning Pipelines - Kubeflow Pipeline



Conceptual Integration Diagram of the Open Search System



Integration of ML Pipelines into an Open Search System



Drawbacks and Benefits

- Efficient storing, generating, maintaining, and sharing of anonymous user information
- Generating Reproducible Results

- Community Validation and Better Understanding of ML/ML Algorithms
- Accountability for Data and Algorithms

Thank you for your Attention

Contacts

- Twitter: @IgoJJ
- Email:
  - igor.jakovljevic@outlook.com
  - igor.jakovljevic@cern.ch