ONLINE MASTERS IN DATA SCIENCE

DSC 208R - Data Management for Analytics

Data Collection and Governance

Arun Kumar

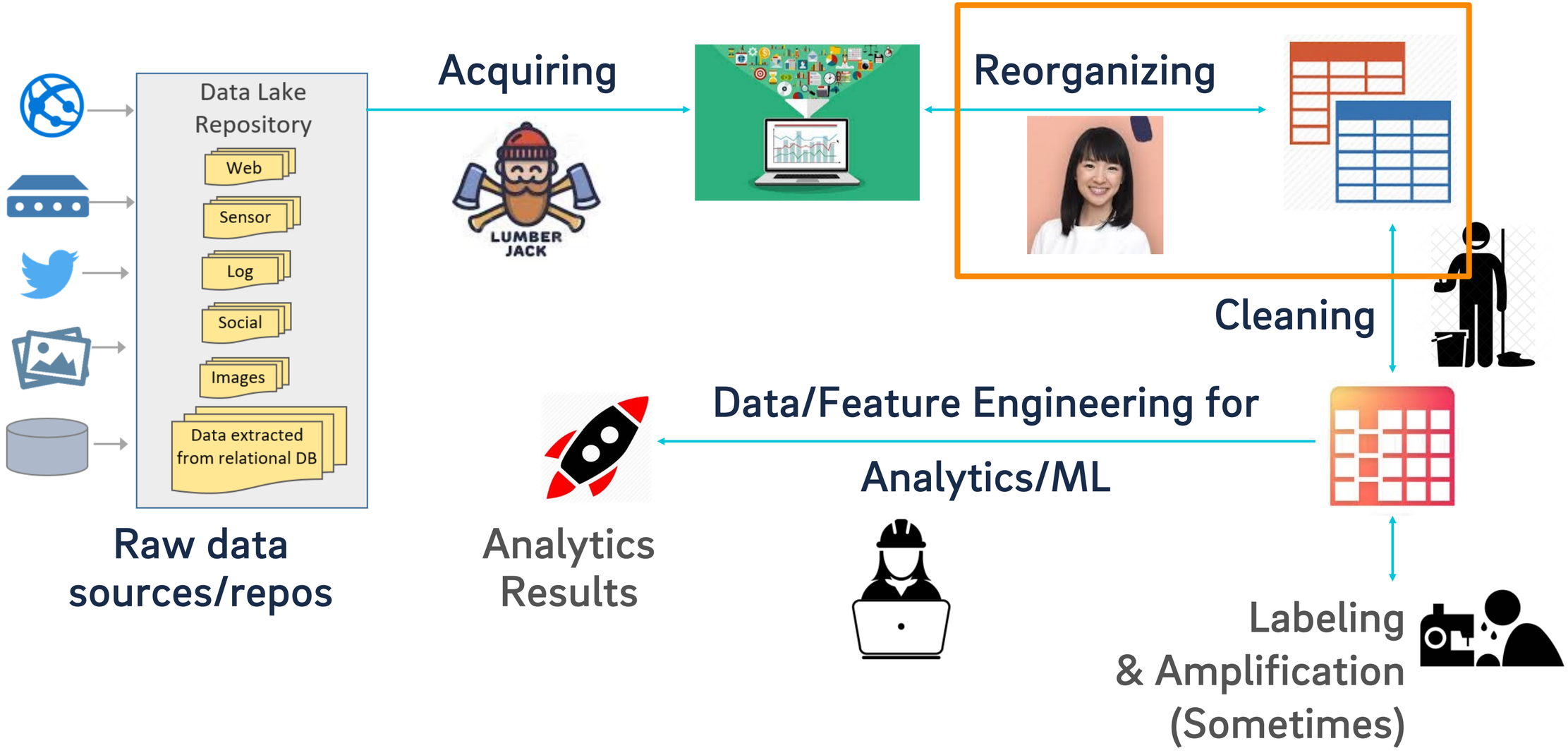
UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

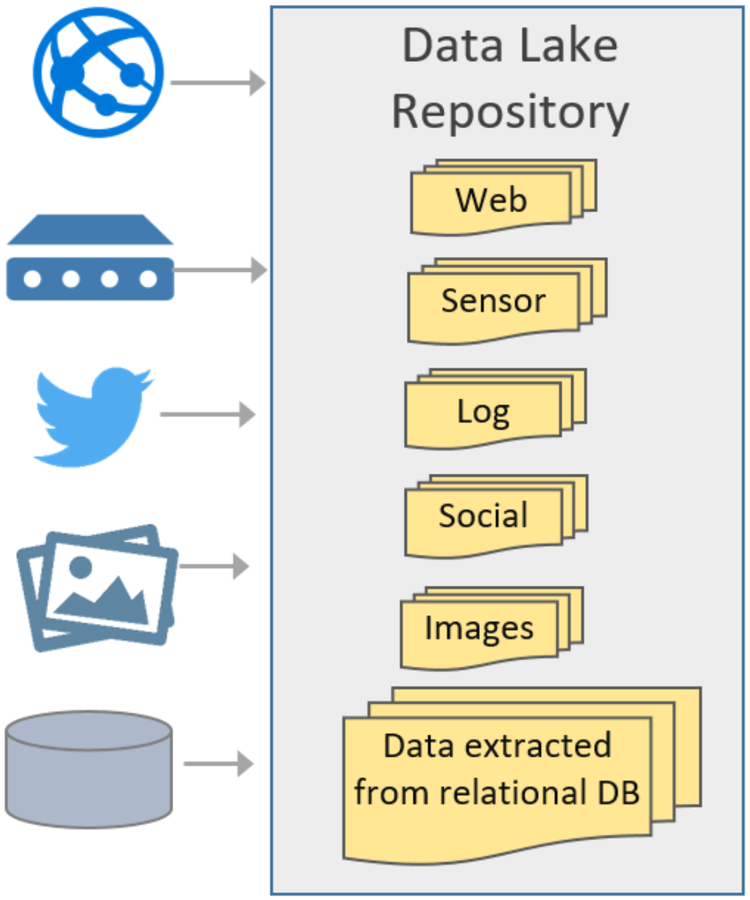
Outline

- Overview
- Data Organization and File Formats
- Data Acquisition
- **Data Reorganization and Preparation**
- Data Labeling and Amplification
- Data Governance and Privacy

Reorganizing and Preparing Data



Reorg/Prep Data for Analytics/ML



Raw data sources/repos

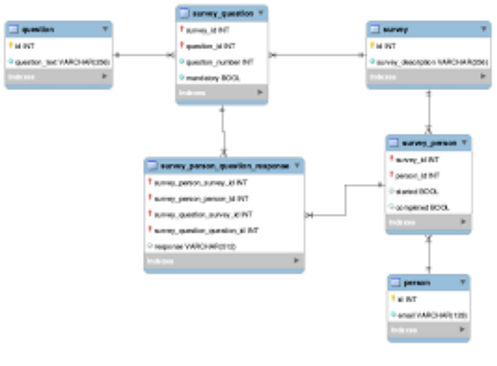
- Raw datasets sit in source systems in their own formats
- Need to unify and *reorganize* them for analytics/ML
- How to reorganize depends on data types and analytics/ML task at hand
- Use SQL, MapReduce, file I/O APIs, etc.

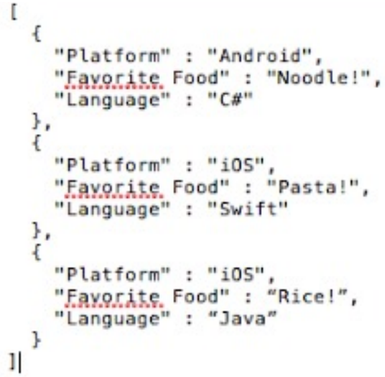
Common Steps:

- Change file formats (e.g., export table -> CSV -> TFRecords)
- Decompression (e.g., multimedia)
- Key-key joins for multimodal data
- Key-FK joins for relational data

Reorg/Prep Data: Examples

Prediction App: Fraud detection in banking





```
[
  {
    "Platform" : "Android",
    "Favorite Food" : "Noodle!",
    "Language" : "C#"
  },
  {
    "Platform" : "iOS",
    "Favorite Food" : "Pasta!",
    "Language" : "Swift"
  },
  {
    "Platform" : "iOS",
    "Favorite Food" : "Rice!",
    "Language" : "Java"
  }
]
```

Joins to denormalize

Flatten JSON records

Large single-table CSV file on HDFS

Prediction App: Image captioning on social media

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Fuse JSON records

Extract image tensors

Large binary file with 1 image tensor and 1 string per line

Reorg/Prep Data for Analytics/ML

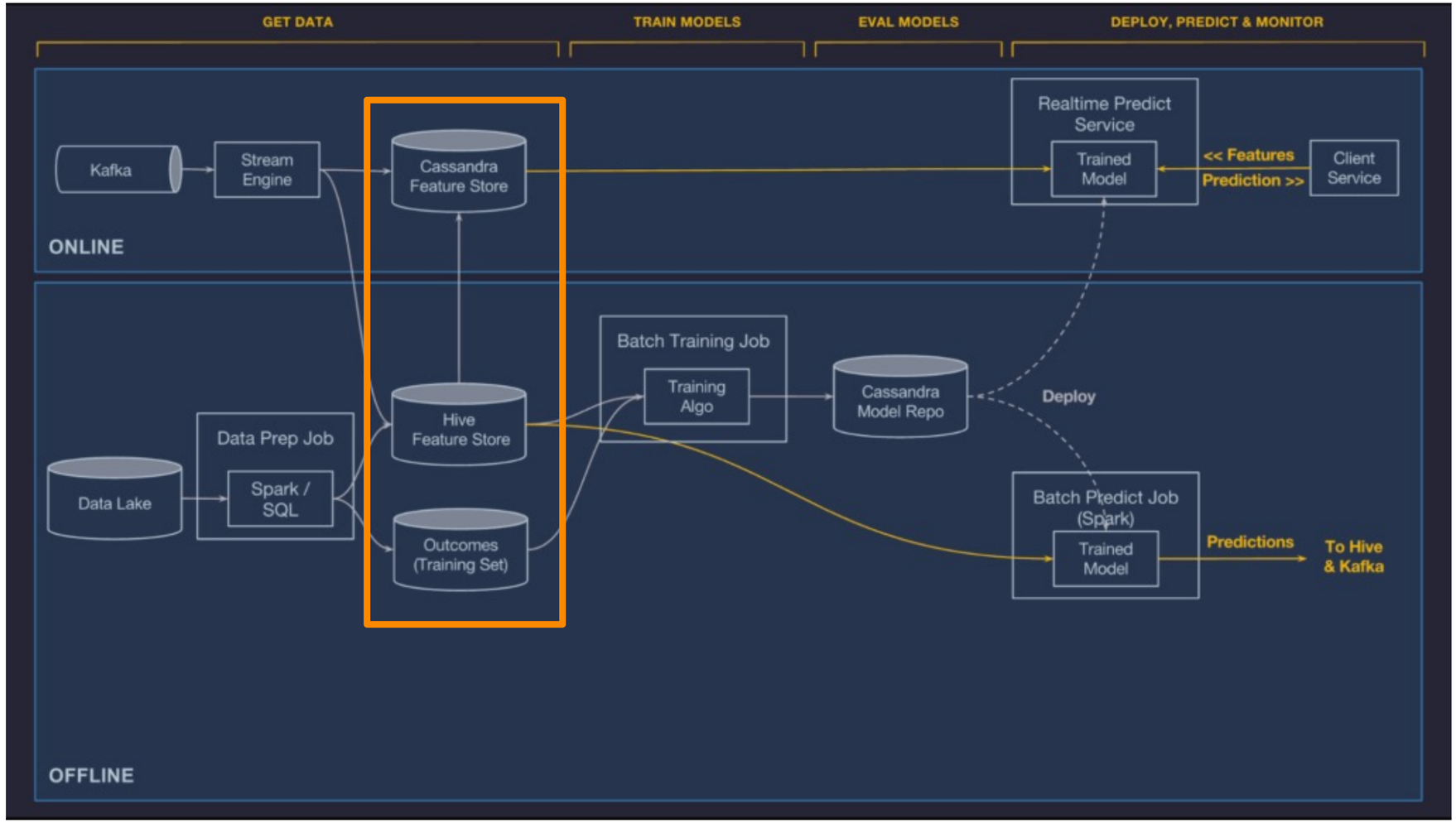
- Typically need both code (SQL, Python) and scripts (bash)

Some best practices:

- **Automation:** Use scripts for reorg workflows; Airflow
- **Documentation:** Maintain notes/READMEs for code
- **Provenance:** Manage metadata on source/rationale for each data source and feature
- **Versioning:** Reorg. is never one-and-done! Maintain logs of what version has what and when

Reorg/Prep Data for Analytics/ML

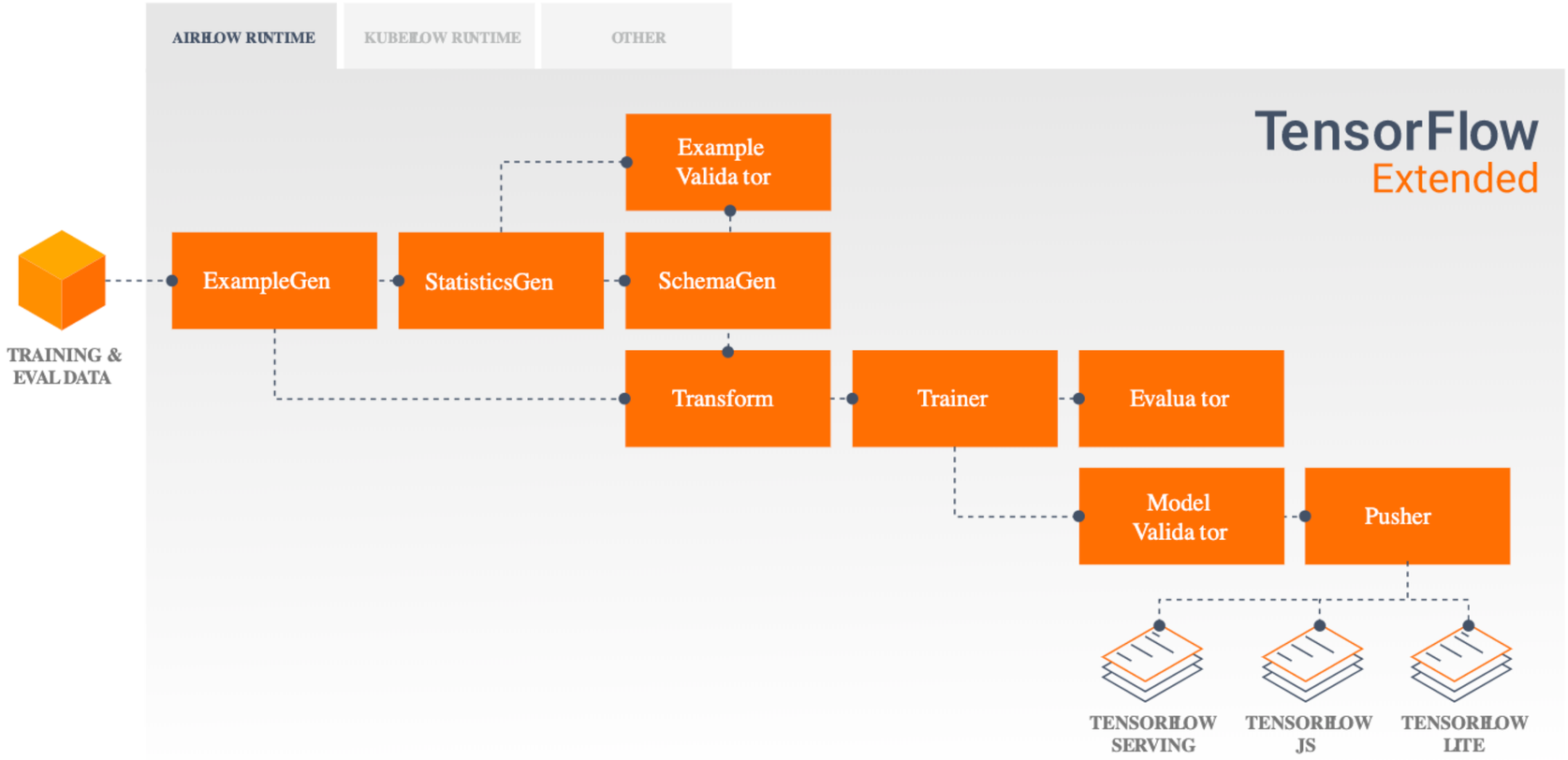
- "Feature stores" in industry help catalogue ML data



https://eng.uber.com/michelangelo/

Reorg/Prep Data for Analytics/ML

- "ML platforms" help streamline reorg/prep workflows
  - Lightweight and flexible schemas now common
  - Makes it easier to automate data *validation*




```
...
feature {
  name: "age"
  value_count {
    min: 1
    max: 1
  }
  type: FLOAT
  presence {
    min_fraction: 1
    min_count: 1
  }
}
feature {
  name: "capital-gain"
  value_count {
    min: 1
    max: 1
  }
  type: FLOAT
  presence {
    min_fraction: 1
    min_count: 1
  }
}
...
```

https://www.tensorflow.org/tfx/guide