ONLINE MASTERS IN DATA SCIENCE

DSC 208R - Data Management for Analytics

Data Collection and Governance

Arun Kumar

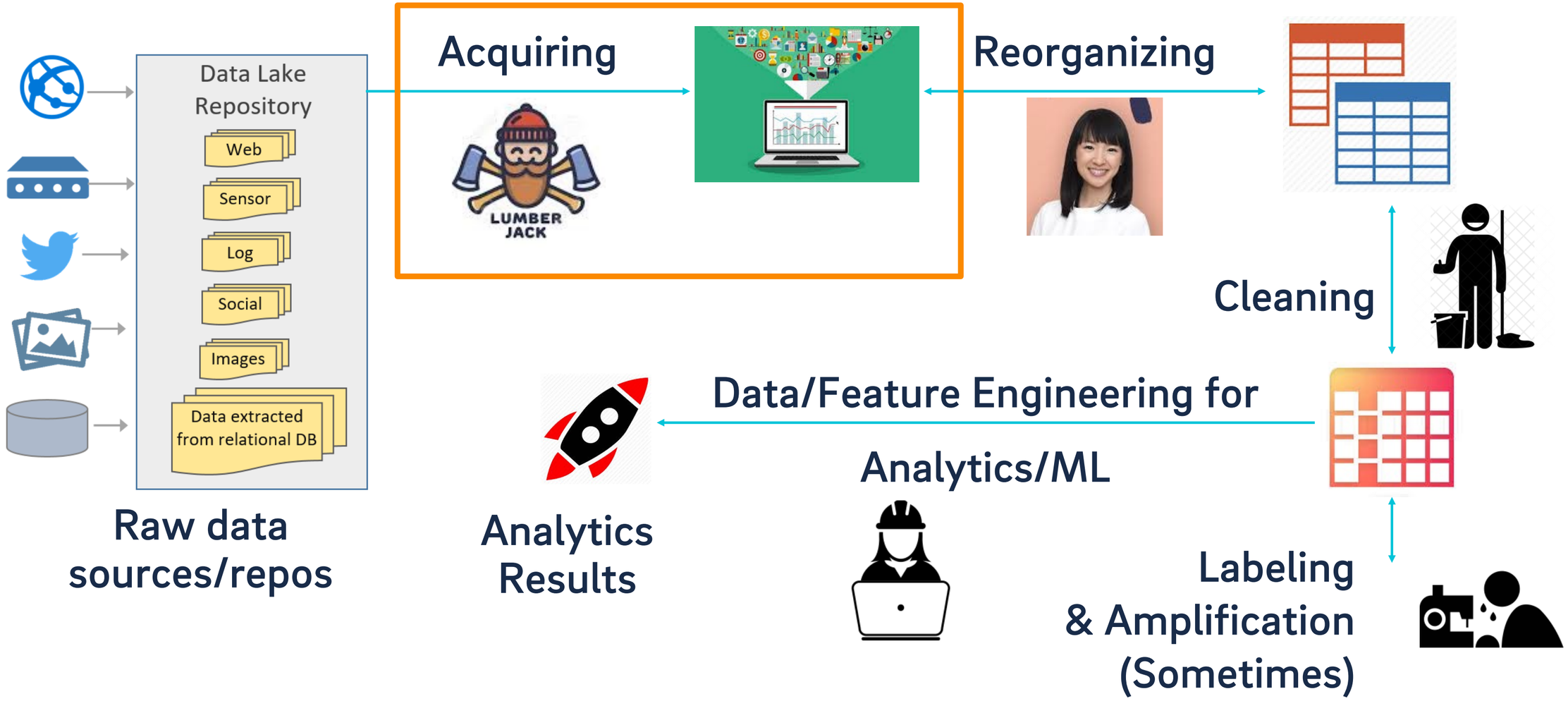
UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

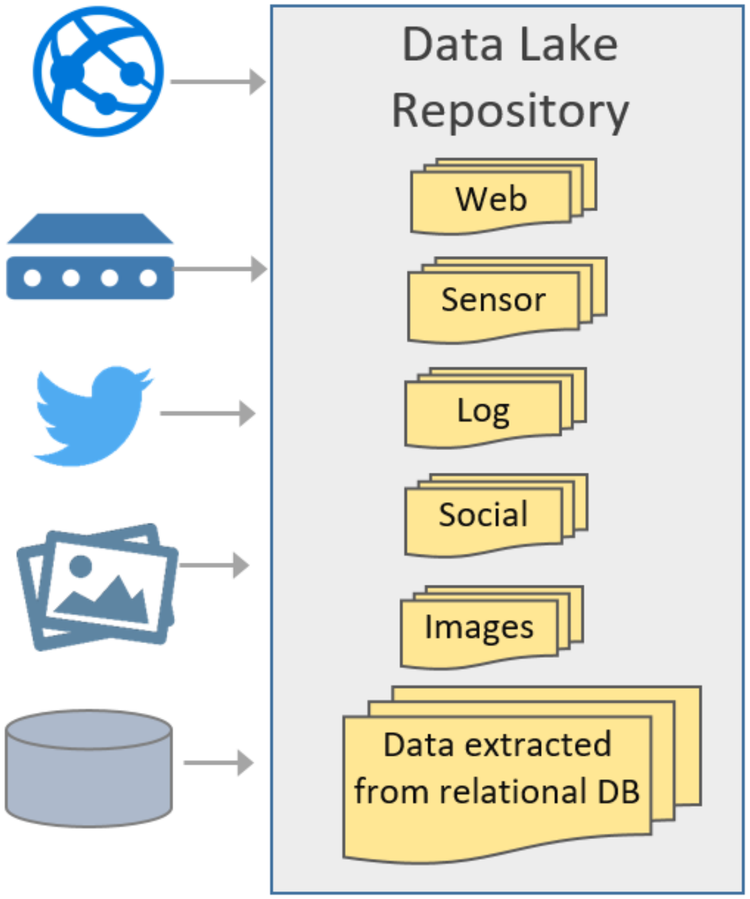
Outline

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

Acquiring Data



Acquiring Data: Data Sources



Raw data sources/repos

- Different sources have different "query languages" and/or APIs to acquire data
- **Structured data:** Typically managed by RDBMSs; queried using SQL
- **Semistructured data:** Exported from key-value stores (e.g., MongoDB)
- **Graph data:** Typically managed by graph DBMSs such as Neo4j
- JSON logs, text files, multimedia, etc.: typically just files on S3, HDFS, etc.

Acquiring Data: Examples

Example: Recommendation System (e.g., Netflix)

Prediction App: Identify top movies to display for user

Data Sources:

User data and past click logs

Movie data

Movie images

Example: Social media analytics for social science

Prediction App: Predicts which tweets will go viral

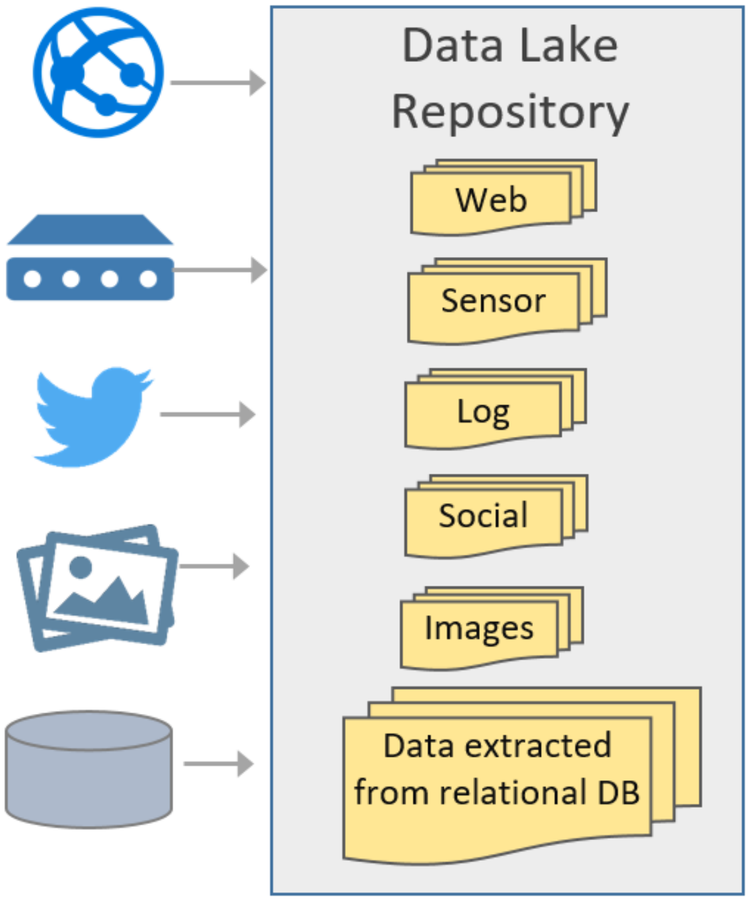
Data Sources:

Tweets as JSON
Structured metadata

Entity Dictionaries

Graph data

Acquiring Data: Challenges



Raw data sources/repos

- Different sources have different "query languages" and/or APIs to acquire data

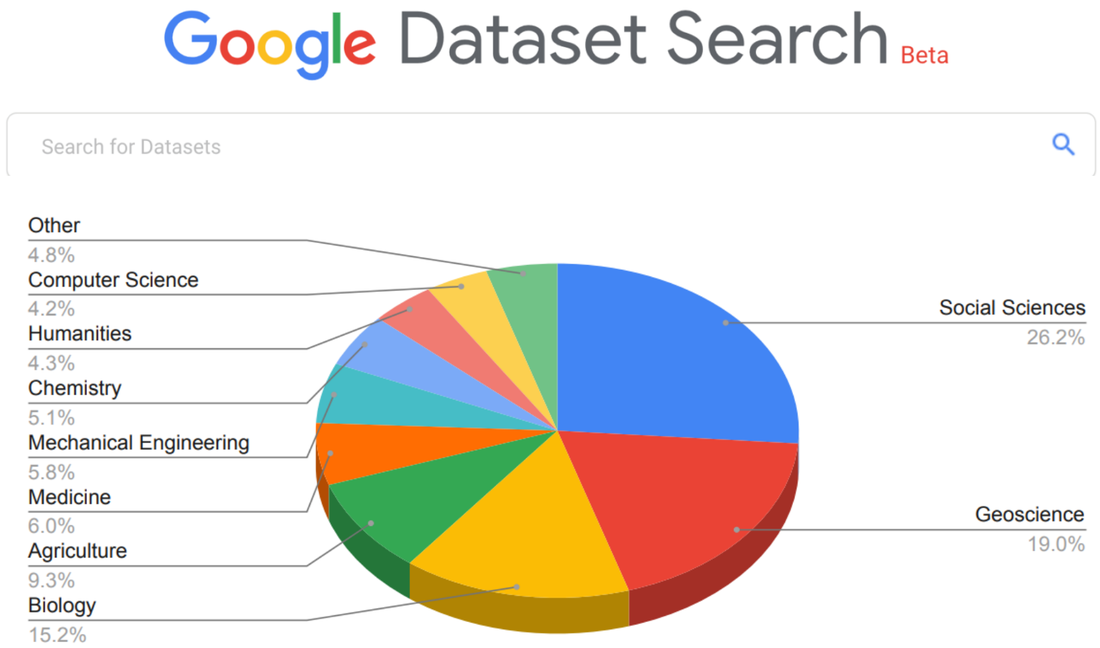
Potential challenges and mitigation:

- Heterogeneity: Do you really need *all* data sources/modalities?
- Access control: Learn organization's data security and authentication policies
- Volume: Do you really need *all* data?
- Scale: Avoid copying files one by one
- Manual errors: Use automated workflow tools such as AirFlow

Acquiring Data: Dataset Discovery

- Some orgs have built “data discovery” tools to help ML users
- **Goal:** Make it easier to find relevant datasets
- **Approach:** Relevance ranking over schemas/metadata

Example:



- Metadata: schema.org/Dataset

Category	Number of datasets	% of total	Sample formats
Tables	7,822K	37%	CSV, XLS
Structured	6,312K	30%	JSON, XML, OWL, RDF
Documents	2,277K	11%	PDF, DOC, HTML
Images	1,027K	5%	JPEG, PNG, TIFF
Archives	659K	3%	ZIP, TAR, RAR
Text	623K	3%	TXT, ASCII
Geospatial	376K	2%	SHP, GEOJSON, KML
Computational biology	110K	<1%	SBML, BIOPAX2, SBGN
Audio	27K	<1%	WAV, MP3, OGG
Video	9K	<1%	AVI, MPG
Presentations	7K	<1%	PPTX
Medical imaging	4K	<1%	NII, DCM
Other categories	2,245K	11%	

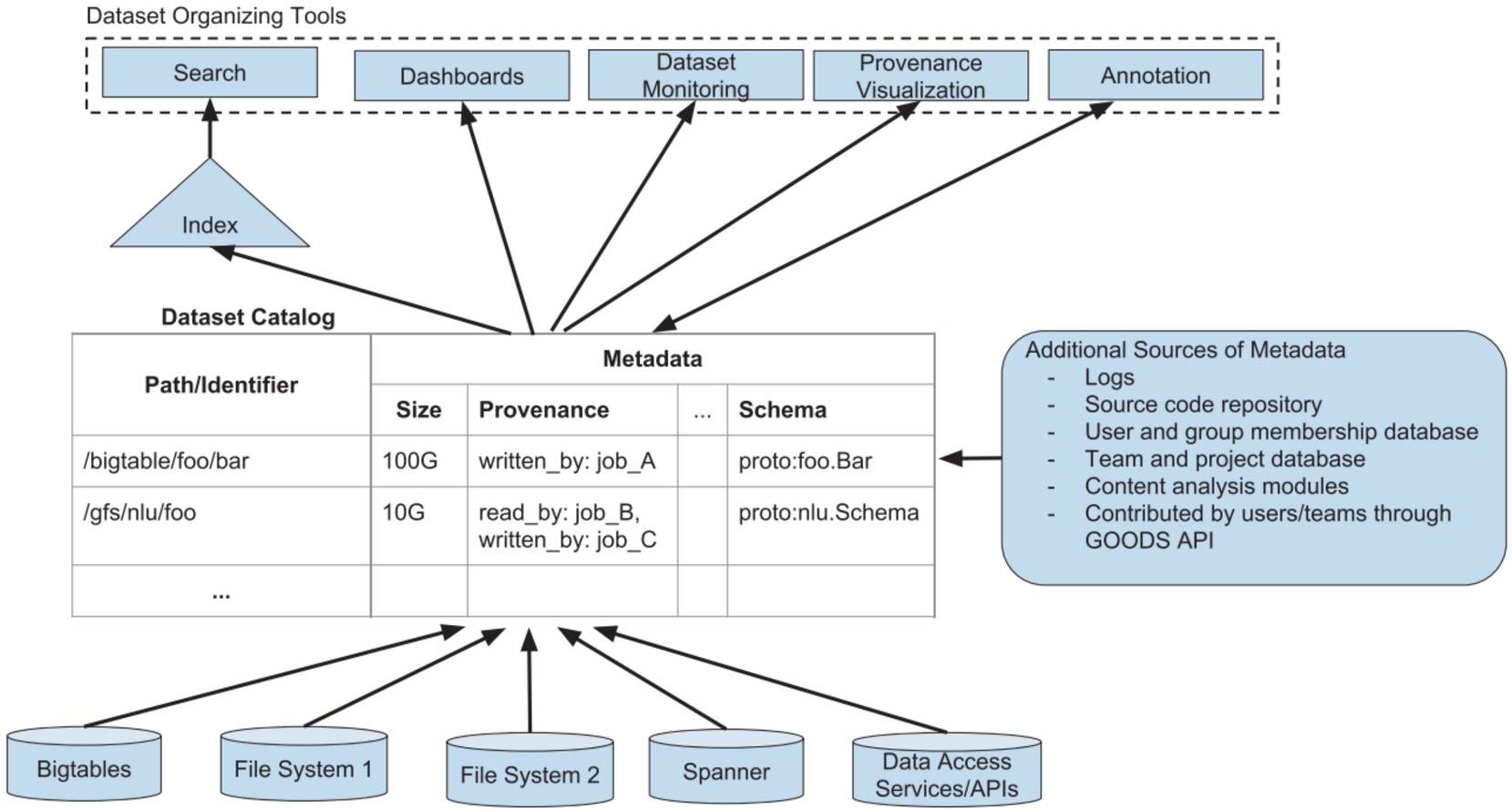
https://storage.googleapis.com/pub-tools-public-publication-data/pdf/afd0602172f297bccdb4ee720bc3832e90e62042.pdf

Acquiring Data: Dataset Discovery

- Tabular datasets especially amenable for augmentation
  - Foreign keys (FK) implicitly suggest possible *joins*

Example:

- GOODS catalogs billions of tables within Google
- Extracts schema from file
- Assigns versions, owners
- Search and dashboards



https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45390.pdf
https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45a9dcf23dbdfa24dbced358f825636c58518afa.pdf