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**

Data Governance

- Data are "entities" with "value"—not unlike people? ☺
  - Born : created, live : used, die : deleted, protected, stewarded, etc.
  - Just as people need to be *governed*, so must data

Key aspects of governing data:

- **Privacy & Security:** Who sees what, why? No breaches!
- **Stewardship:** Who owns what, when? Access control.
- **Cataloging:** What is it, where, how to access?
- **Defining:** Data dictionaries, business knowledge.
- **Quality:** Follow conventions, reduce errors.
- **Provenance:** Track usage, changes, versions. Auditing.

Legal Regulations on Data Handling

- Just as laws exist to govern people, laws exist to govern data
  - No laws (yet) on ML “algorithms”, but yes for ML data
- Long history of laws surrounding data:

FERPA

1974

Broadly applies to all “education records” of students

FERPA is a federal law that protects the privacy of student records and applies to all schools that receive funds from the USDOE.

Types of Student Records:
- Financial information
- Disciplinary files
- Student transcripts
- Immunization & health records

To be compliant, schools can utilize a paperless system for storing student records. School’s funding is based on compliance.

https://www.recordnations.com/2019/07/ferpa-how-to-manage-student-records

Legal Regulations on Data Handling

HIPPA; 1996

Broadly applies to all healthcare data, especially PII

Legal Regulations on Data Handling

GDPR
2018

- Applies to data collected from individuals in EU and EEA
- New rights on “personal data” : right to access, right to forget/erasure, right to object, etc.
- Many Web companies scrambled; some “exited” EU area

Legal Regulations on Data Handling

GDPR
2018

- New technical challenges on making data/ML infra. GDPR-compliant: metadata handling, efficiency, etc.
- Open legal + technical questions for ML applications:
    - Are ML models under purview?
    - Any form of derived/ aggregated data?

Legal Regulations on Data Handling

	CCPA	GDPR
WHEN DOES THE LAW GO INTO EFFECT?	**JANUARY 1, 2020** Enforcement begins July 1, 2020	**MAY 25,2018** Enforcement in effect
WHAT ORGANIZATIONS ARE IN SCOPE?	**FOR-PROFIT COMPANIES THAT:** • Collect personal data on 50K+ California residents • Have annual revenues of over $25 million • Earn 50%+ of annual revenue from California residents' data	**ANY ORGANIZATION THAT:** • Operates inside or outside the European Union (EU) and offers goods or services to customers or businesses in the union
WHO IS AFFECTED?	• Business, service providers, third parties, and California consumers	• EU citizens, businesses, controller, processor, and data subjects
WHAT DATA IS WITHIN SCOPE?	• Personal data that is sold for monetary or other value considerations (releasing, disclosing, transferring, or even renting of the data)	• Personal data of any type
WHAT ARE THE FINES OF NONCOMPLIANCE?	• Up to $7,500 per violation with no ceiling on the number of violations • $100-$750 per consumer per incident for statutory damages related to breaches	• Up to 20 million euros or 4% of total global turnover from the prior fiscal year for the most severe violations • Up to 10 million euros or 2% of the worldwide annual revenue of the prior fiscal year for less severe violations

https://riskonnect.com/uk/regulatory-compliance/ccpa-and-gdpr-how-the-privacy-laws-stack-up/

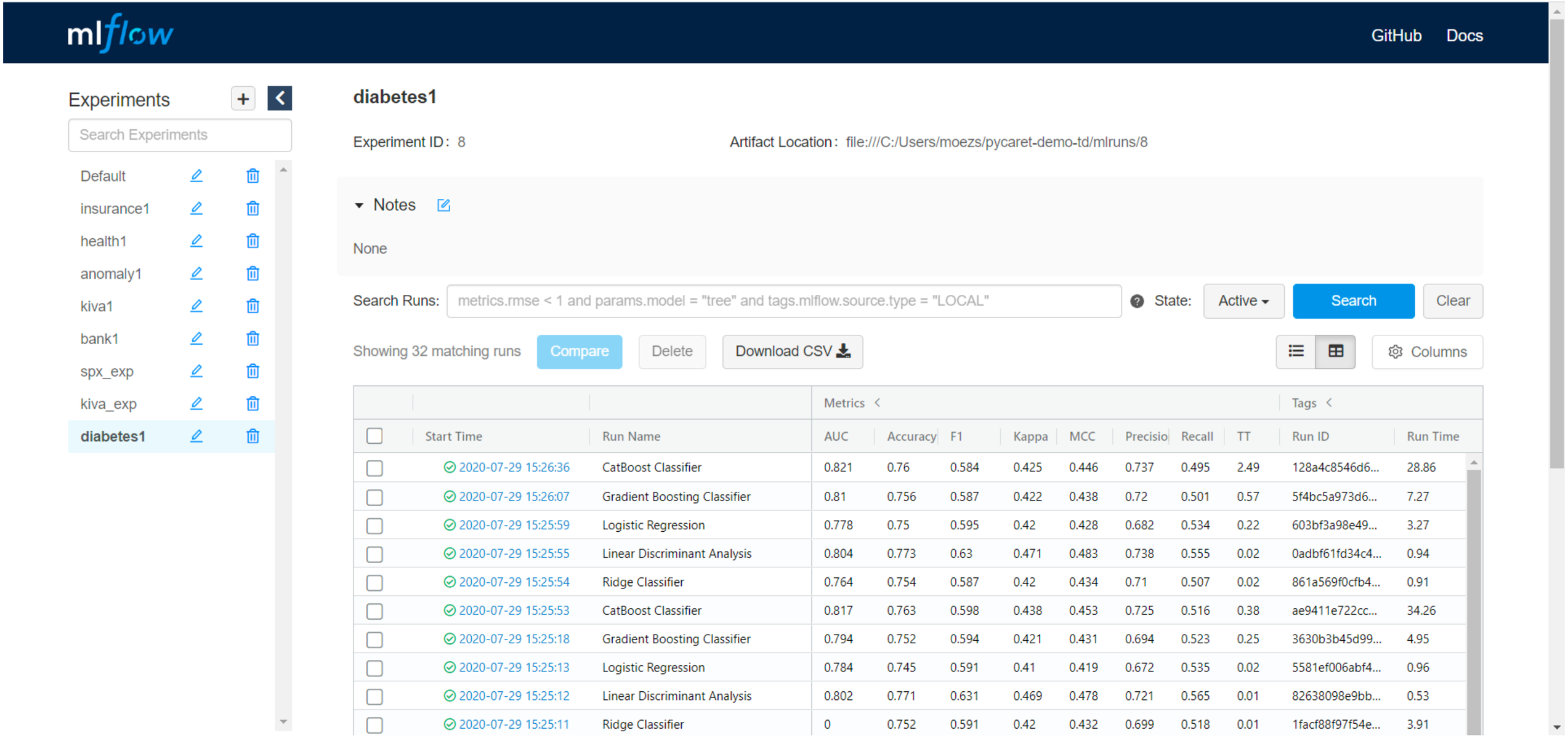
Provenance Management

- All data objects must be tracked throughout lifecycle
  - *Compliance* with data regulations; auditing
  - Makes data easier to find and consume
- **Provenance:** "Chronology of the ownership, custody or location of a historical object"
- Key aspects of provenance:
  - *Context* of data creation, deletion, access/use, etc.
  - *Evolution* of metadata
  - *Versioning* of data and all derived objects
- For ML: track derived data (e.g., feature extraction), ML artifacts (models, code/scripts, etc.), & configuration

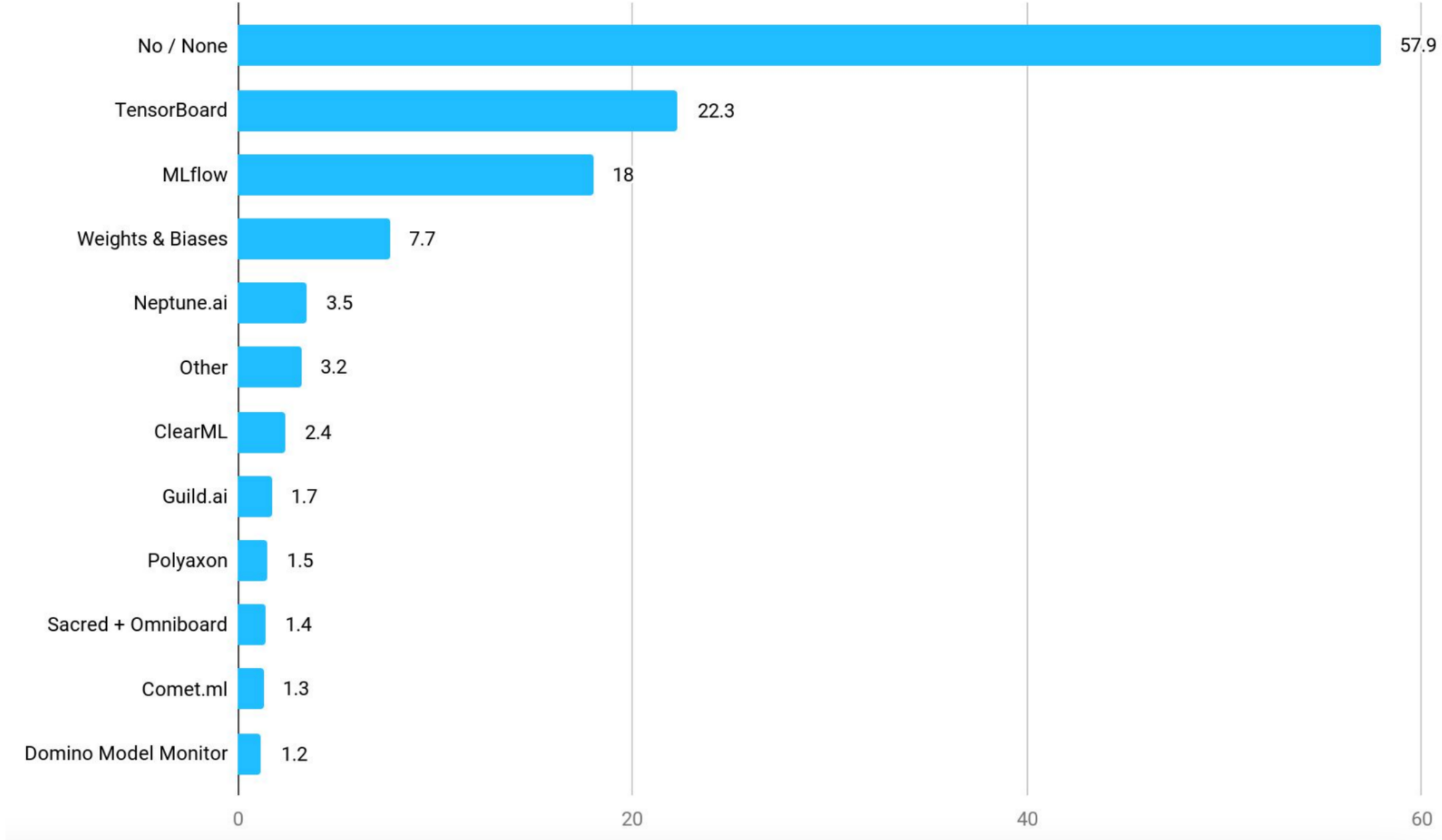
Provenance Management

- **Challenge:** Heterogeneity of data/ML platforms makes it notoriously messy/tedious
    - Metadata? Usage logs? Versioning?
- The state of the world today: ad hoc, organization-specific practices and tools
    - Need to learn org-specific practices and APIs
- Some emerging open source tools do help:
    - ML artifacts: Weights & Biases, MLFlow, TensorFlow Extended, TensorBoard, etc.
    - SQL transformation code: dbt
    - Derived data: Feature stores (e.g., Feast, Tecton)

Example: MLFlow Experiment Tracking



State of Tool Adoption in Practice



https://www.kaggle.com/kaggle-survey-2021

Outline

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