

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.

Legal Regulations on Data Handling

HIPPA; 1996

Broadly applies to all healthcare data, especially PII

5 Steps Toward HIPAA Compliance

FOR SECURE COMMUNICATION AND COLLABORATION



Produced in collaboration with ADA Business Resources and PBHS.com

Access tips at **Success.ADA.org**



ADA Center for
Professional Success™

HIPAA FACTS

1 in 7
healthcare
organizations
have still not appointed a
HIPAA compliance officer



1 in 4 HIPAA
breaches
still not reported



50%
of healthcare
organizations
believe they would fail a
HIPAA Audit



Average cost
per record is \$363



About
600 HIPAA
Violations
referred to DoJ



80%
of healthcare
organizations
fail meaningful use audits



Unauthorized
access
to records
accounts for 20%
of HIPAA breaches



Phishing and
ransomware
the top hacker tactic



Ransomware
for medical records
accelerating rapidly



Nearly 200
million
patient records
compromized since
introduction of HIPAA



Copyright © 2018 The HIPAA Guide

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

JANUARY 1, 2020

Enforcement
begins July 1, 2020

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

- Business, service providers, third parties, and California consumers

- Personal data that is sold for monetary or other value considerations (releasing, disclosing, transferring, or even renting of the data)

- 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

WHEN DOES
THE LAW GO
INTO EFFECT?

WHAT
ORGANIZATIONS
ARE IN SCOPE?

WHO IS
AFFECTED?

WHAT DATA IS
WITHIN SCOPE?

WHAT ARE
THE FINES OF
NONCOMPLIANCE?

GDPR

MAY 25, 2018

Enforcement
in effect

ANY ORGANIZATION THAT:

- Operates inside or outside the European Union (EU) and offers goods or services to customers or businesses in the union

- EU citizens, businesses, controller, processor, and data subjects

- Personal data of any type

- 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

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

mlflow

GitHub Docs

Experiments



Search Experiments

- Default
- insurance1
- health1
- anomaly1
- kiva1
- bank1
- spx_exp
- kiva_exp
- diabetes1**

diabetes1

Experiment ID: 8

Artifact Location: file:///C:/Users/moezs/pycaret-demo-td/mlruns/8

Notes

None

Search Runs: metrics.rmse < 1 and params.model = "tree" and tags.mlflow.source.type = "LOCAL"

State:

Active

Search

Clear

Showing 32 matching runs

Compare

Delete

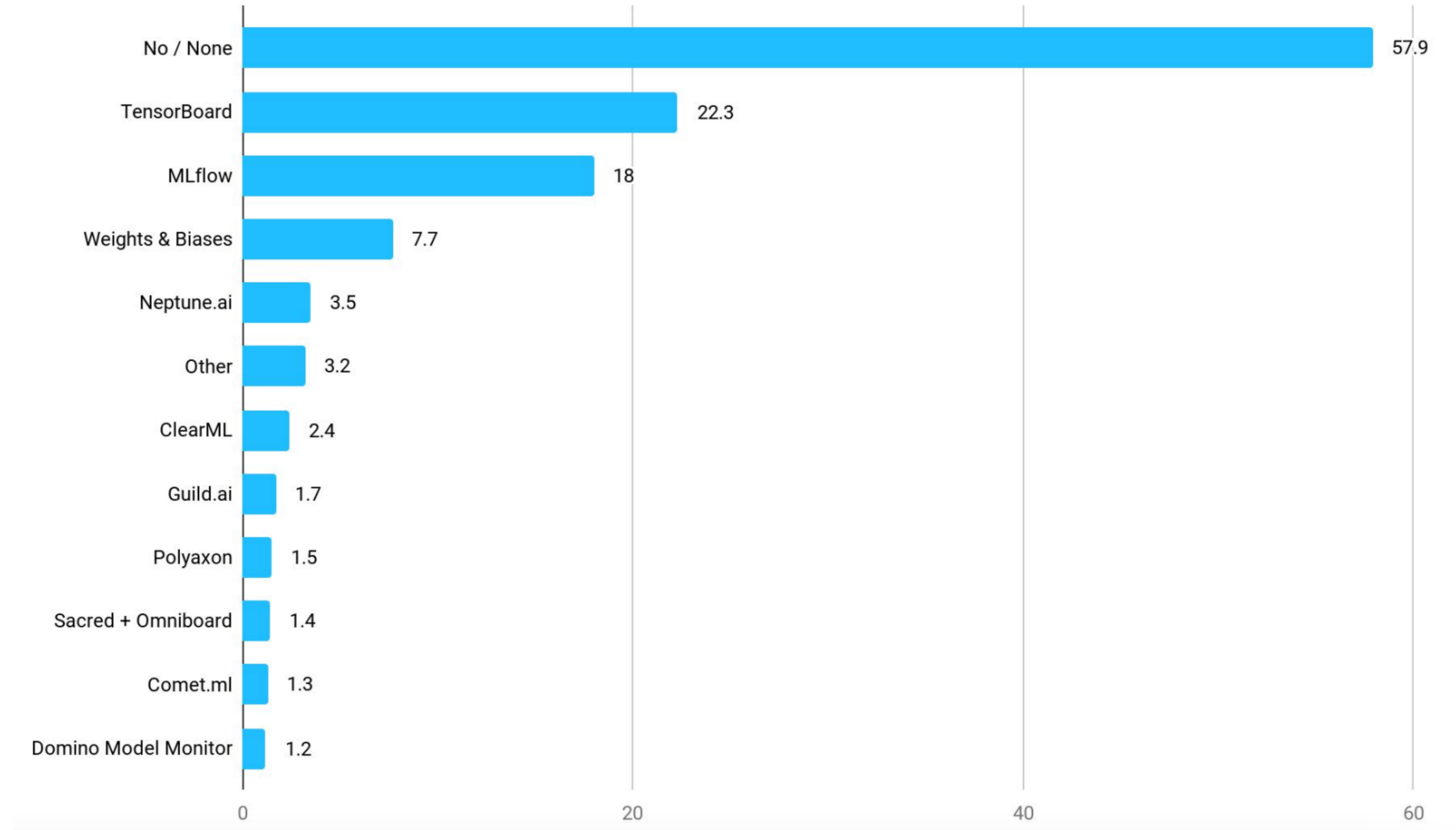
Download CSV



Columns

			Metrics <								Tags <	
<input type="checkbox"/>	Start Time	Run Name	AUC	Accuracy	F1	Kappa	MCC	Precisio	Recall	TT	Run ID	Run Time
<input type="checkbox"/>	2020-07-29 15:26:36	CatBoost Classifier	0.821	0.76	0.584	0.425	0.446	0.737	0.495	2.49	128a4c8546d6...	28.86
<input type="checkbox"/>	2020-07-29 15:26:07	Gradient Boosting Classifier	0.81	0.756	0.587	0.422	0.438	0.72	0.501	0.57	5f4bc5a973d6...	7.27
<input type="checkbox"/>	2020-07-29 15:25:59	Logistic Regression	0.778	0.75	0.595	0.42	0.428	0.682	0.534	0.22	603bf3a98e49...	3.27
<input type="checkbox"/>	2020-07-29 15:25:55	Linear Discriminant Analysis	0.804	0.773	0.63	0.471	0.483	0.738	0.555	0.02	0adbf61fd34c4...	0.94
<input type="checkbox"/>	2020-07-29 15:25:54	Ridge Classifier	0.764	0.754	0.587	0.42	0.434	0.71	0.507	0.02	861a569f0cfb4...	0.91
<input type="checkbox"/>	2020-07-29 15:25:53	CatBoost Classifier	0.817	0.763	0.598	0.438	0.453	0.725	0.516	0.38	ae9411e722cc...	34.26
<input type="checkbox"/>	2020-07-29 15:25:18	Gradient Boosting Classifier	0.794	0.752	0.594	0.421	0.431	0.694	0.523	0.25	3630b3b45d99...	4.95
<input type="checkbox"/>	2020-07-29 15:25:13	Logistic Regression	0.784	0.745	0.591	0.41	0.419	0.672	0.535	0.02	5581ef006abf4...	0.96
<input type="checkbox"/>	2020-07-29 15:25:12	Linear Discriminant Analysis	0.802	0.771	0.631	0.469	0.478	0.721	0.565	0.01	82638098e9bb...	0.53
<input type="checkbox"/>	2020-07-29 15:25:11	Ridge Classifier	0	0.752	0.591	0.42	0.432	0.699	0.518	0.01	1facf88f97f54e...	3.91

State of Tool Adoption in Practice



Outline

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

