Carnegie Mellon University Qatar

Quality Assurance for ML/AI

17-313, Foundations of Software Engineering, Fall 2023

(based on the slides of 17-445 by Christian Kästner)

S3D
Software and Societal Systems Department

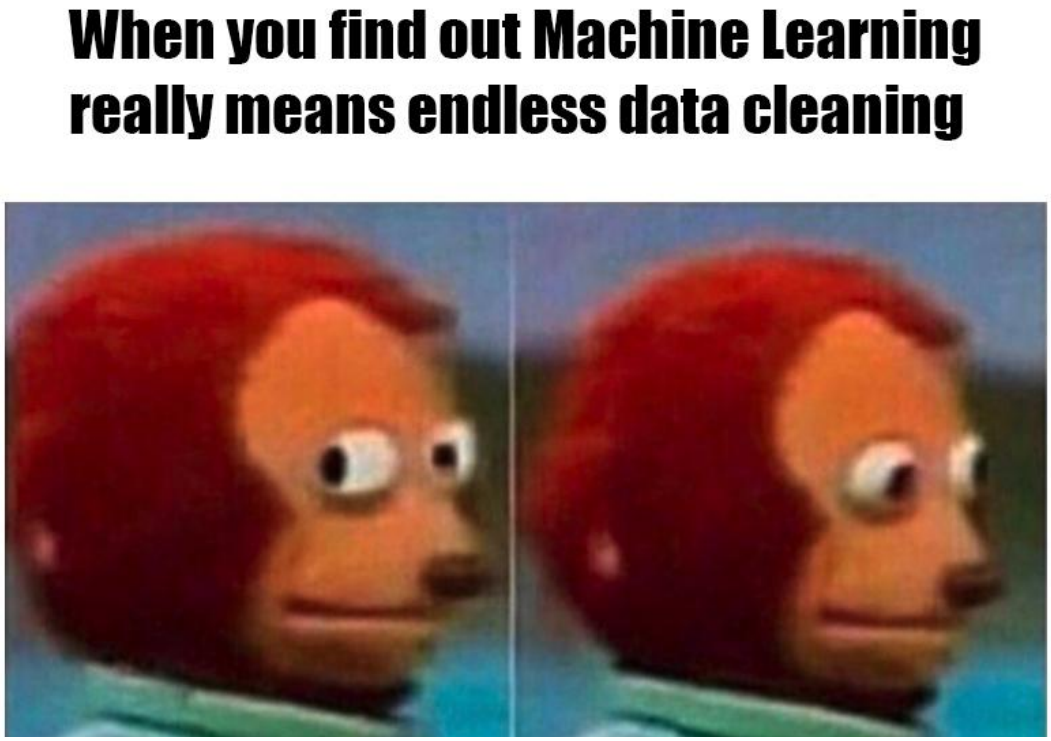
Learning goals

- Understand challenges for QA of ML systems
- Be able to test assumptions about the data
- List quality attributes to consider in building ML systems
- Understand different definitions of fairness
- Discuss methods for measuring fairness
- Outline interventions to improve fairness at the model level

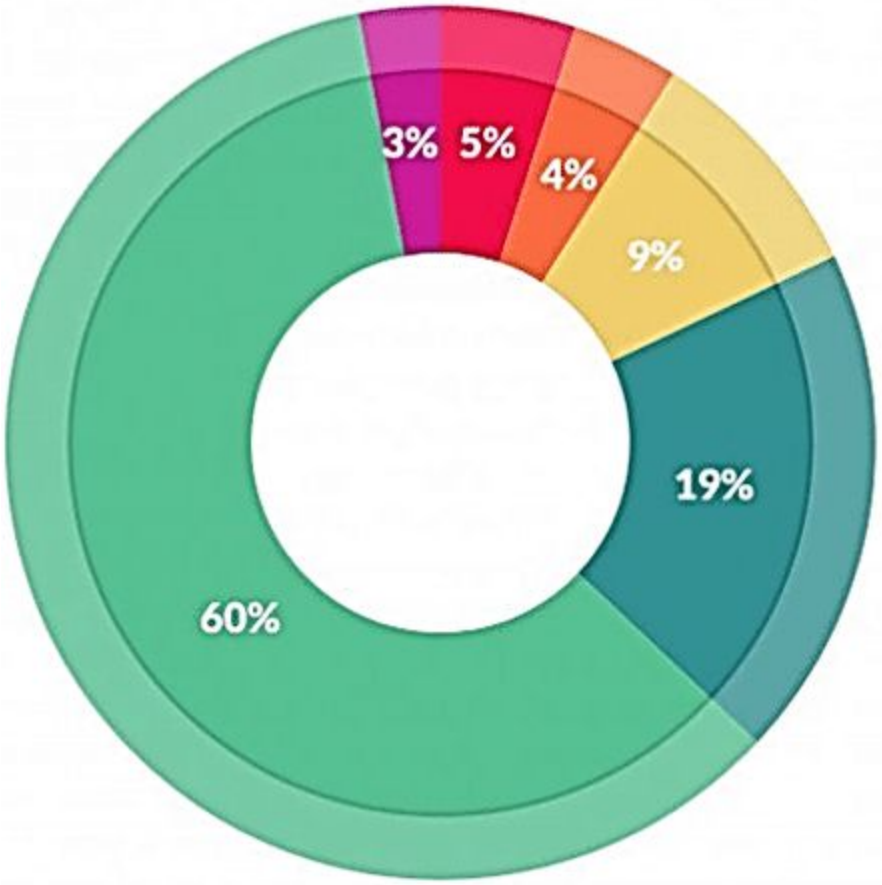
Outline

- Data Qualities
- Model Qualities
- Fairness

“***Data***” science



"Data cleaning and repairing account for about 60% of the work of data scientists."



- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

What makes good quality data?

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - The entries are recorded once.
- Consistency
  - Format, units, data agrees with itself
- Timeliness
  - The data is kept up to date.

Data is noisy

- Multiple sources
- Unreliable sensors or data entry
- Wrong results and computations, crashes
- Duplicate data, near-duplicate data
- Out of order data
- Data format invalid

Data quality and ML

- More data -> better models (up to a point)
- Noisy data (imprecise) -> less confident models
  - Some ML techniques are more or less robust to noise
- Inaccurate data: misleading models, biased models

- Need the "right" data
- Invest in data quality, not just quantity

Dirty data: Example

TABLE: CUSTOMER

ID	Name	Birthday	Age	Sex	Phone	ZIP
3456	Ford, Harrison	18.2.76	43	M	9999999999	15232
3456	Mark Hamil	33.8.81	43	M	6173128718	17121
3457	Kim Kardashian	11.10.56	63	M	4159102371	94016

TABLE: ADDRESS

ZIP	City	State
15232	Pittsburgh	PA
94016	Sam Francisco	CA
73301	Austin	Texas

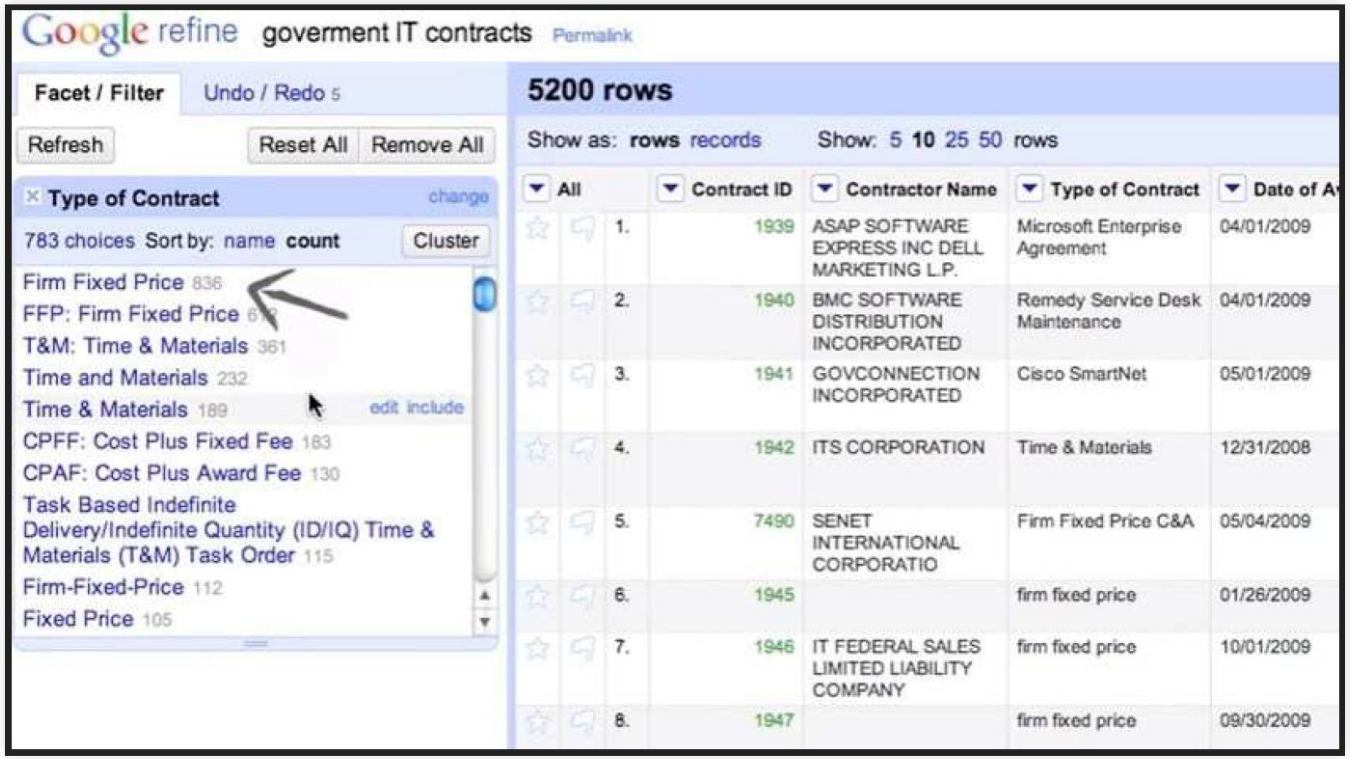
Q. Can we (automatically) detect errors? Which errors are problem-dependent?

How do you avoid bad data?

Common strategies

- Enforce schema constraints
  - e.g., delete rows with missing data or use defaults
- Explore sources of errors (Data exploration)
  - e.g., debugging missing values, outliers
- Remove outliers
  - e.g., Testing for normal distribution, remove > 2σ
- Normalization
  - e.g., range [0, 1], power transform
- Fill in missing values

Data cleaning tools

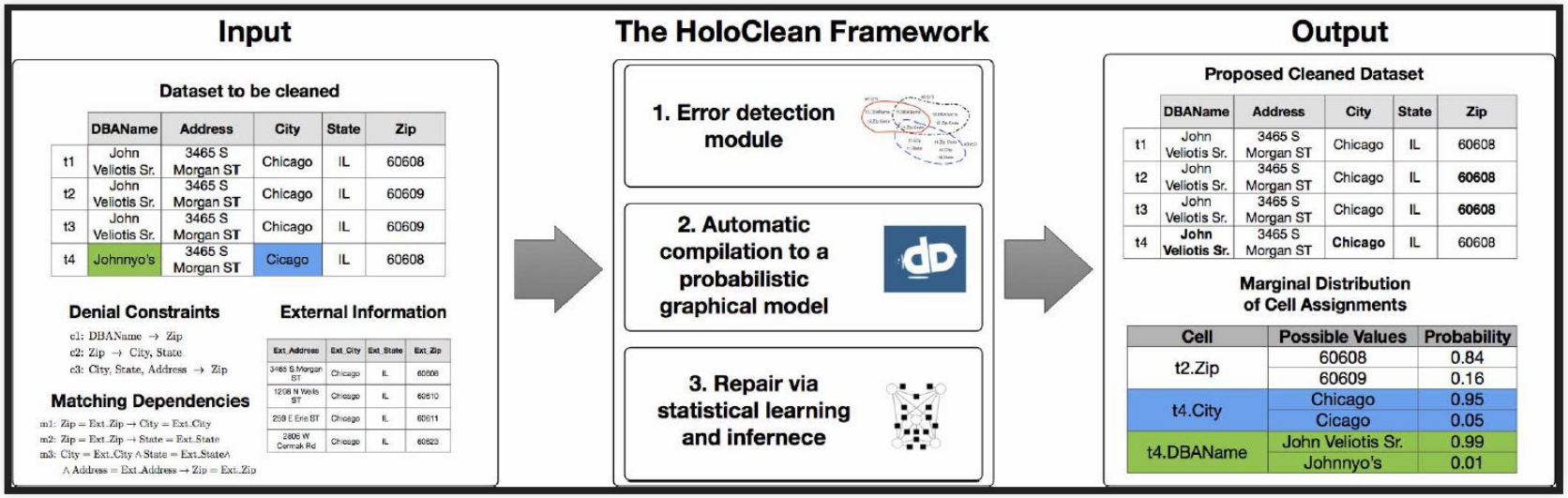


OpenRefine (formerly Google Refine), Trifacta Wrangler, Drake, etc.,

Probabilistic repair

- Use rules to identify inconsistencies and the more likely fix
- If confidence high enough, apply automatically
- Show suggestions to end users (like spell checkers) or data scientists
- Many tools in this area

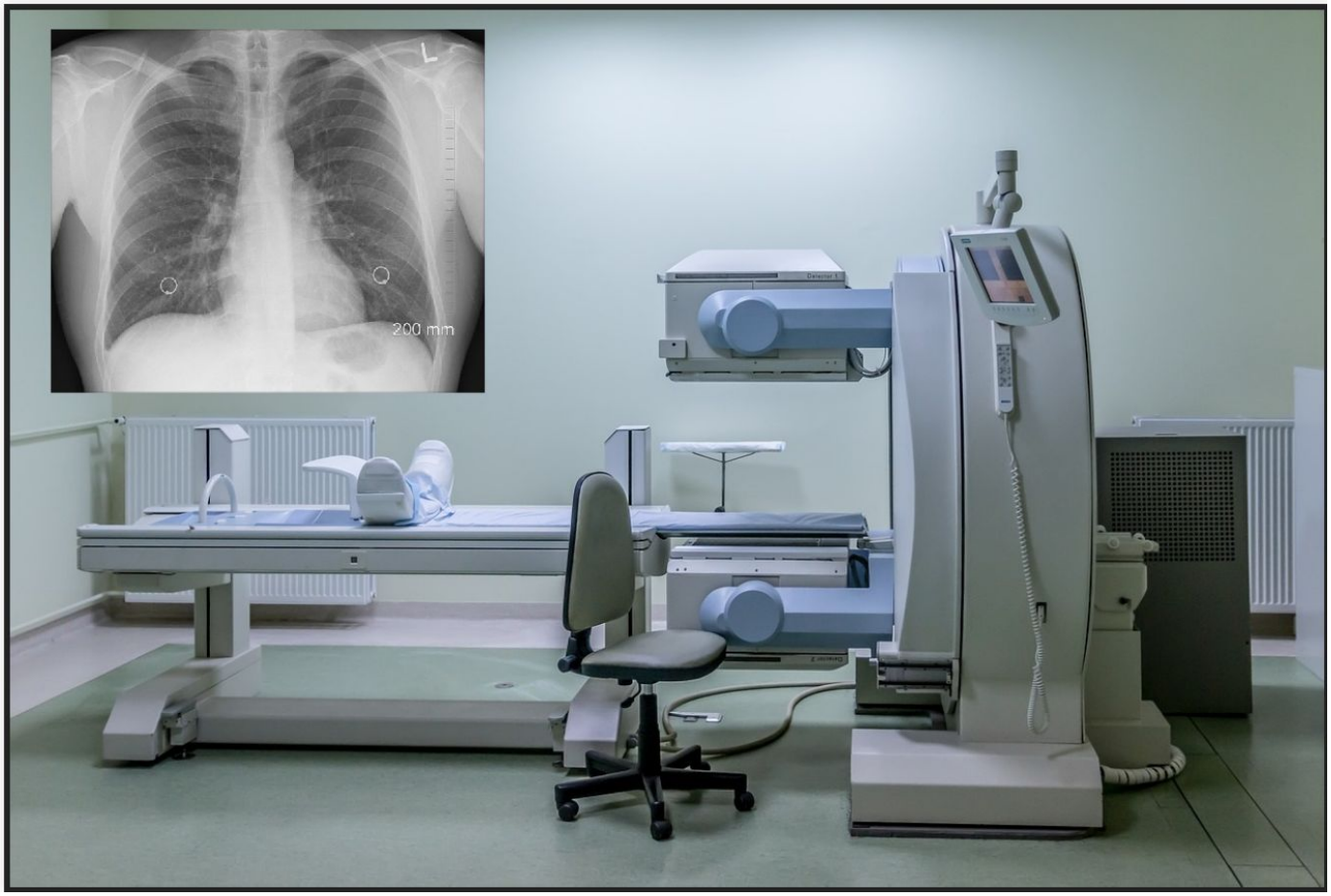
HOLOCLEAN: AI to create more AI



Data drift

- Structural drift
  - Data schema changes, sometimes by infrastructure changes
    - e.g., 4124784115 -> 412-478-4115
- Semantic drift
  - Meaning of data changes, same schema
  - e.g., Netflix switches from 5-star to +/- rating, but still uses 1 and 5
- Distribution changes
  - e.g., credit card fraud differs to evade detection
  - e.g., marketing affects sales of certain items

Case study: Cancer detection



Model qualities

- Prediction accuracy of a model is important
- But many other quality matters when building a system:
  - Model size
  - Inference time
  - User interaction model
  - Kinds of mistakes made
  - How the system deals with mistakes
  - Ability to incrementally learn
  - Safety, security, fairness, privacy
  - Explainability
- *Today: Narrow focus on prediction accuracy of the model*

A note on terminology

In machine learning, "performance" typically refers to accuracy

"this model performs better" = it produces more accurate results

Be aware of ambiguity across communities.

- When speaking of "time", be explicit: "learning time", "inference time", "latency",

Classification Accuracy

- Binary classification: Positive / Negative
- Possible classification outcomes:
  - TN: True Negatives
  - TP: True Positives
  - FN: False Negatives
  - FP: False Positives

Actual Class	Predicted Class	
	Negative	Positive
Negative	**TN**	**FP**
Positive	**FN**	**TP**

Key metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Example: Cancer prediction

Number of cases: **100,000**

Actual State	Predicted Negative	Predicted Positive
Negative	TN **97750**	FP **150**
Positive	FN **330**	TP **1770**

Accuracy

- Proportion of **correct** classifications (true positives and negatives) from **overall** number of cases

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{97,750 + 1,770}{100,000} = 0.9952$$

Recall

- Proportion of **correct positive** classifications (true positives) from cases that are **actually positives**
  - aka true positive rate, hit rate, sensitivity; *higher is better*
  - False negative rate = 1 - recall
    - aka miss rate; *lower is better*

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{1770}{1770 + 330} = 0.8428$$

Precision

- Proportion of **correct positive** classifications from cases that are **predicted as positive**

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{1770}{1770 + 150} = 0.9219$$

Is 99% accuracy good?

- depends on problem; can be excellent, good, mediocre, terrible
- 10% accuracy can be good on some tasks (information retrieval)

Always compare to a base rate!

$$\text{Reduction in error} = \frac{(1 - accuracy_{\text{baseline}}) - (1 - accuracy_f)}{1 - accuracy_{\text{baseline}}}$$

- from 99.9% to 99.99% accuracy = 90% reduction in error
- from 50% to 75% accuracy = 50% reduction in error

Baselines?

- Suitable baselines for cancer prediction? For recidivism?

Consider the baseline probability

Predicting unlikely events -- 1 in 2000 has cancer (stats)

Random predictor

	Cancer	No c.
Cancer pred.	3	4998
No cancer pred.	2	4997

.5 accuracy, .6 recall, 0.001 precision

Never cancer predictor

	Cancer	No c.
Cancer pred.	0	0
No cancer pred.	5	9995

.999 accuracy, 0 recall, .999 precision

See also Bayesian statistics

Case study: Mitosis detection

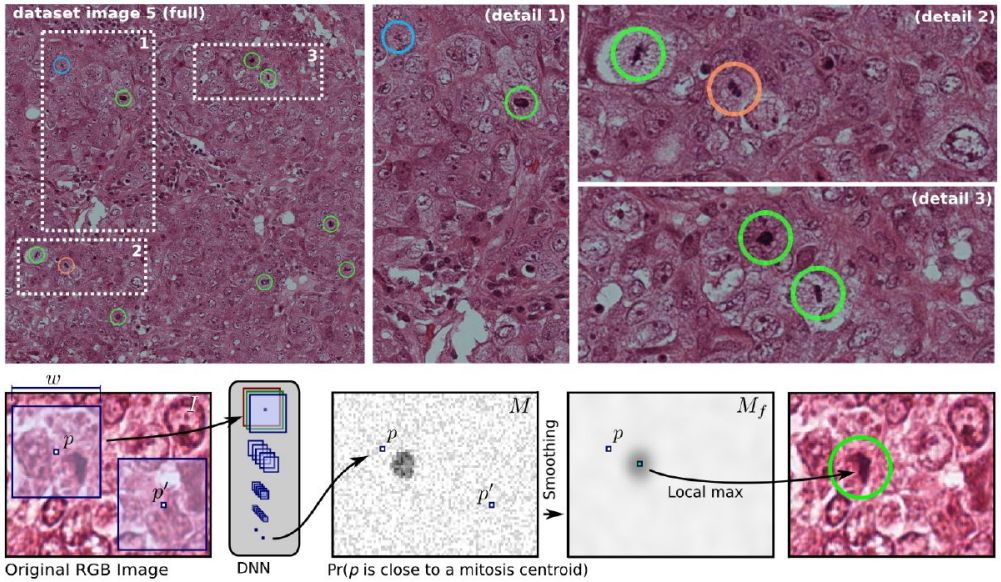
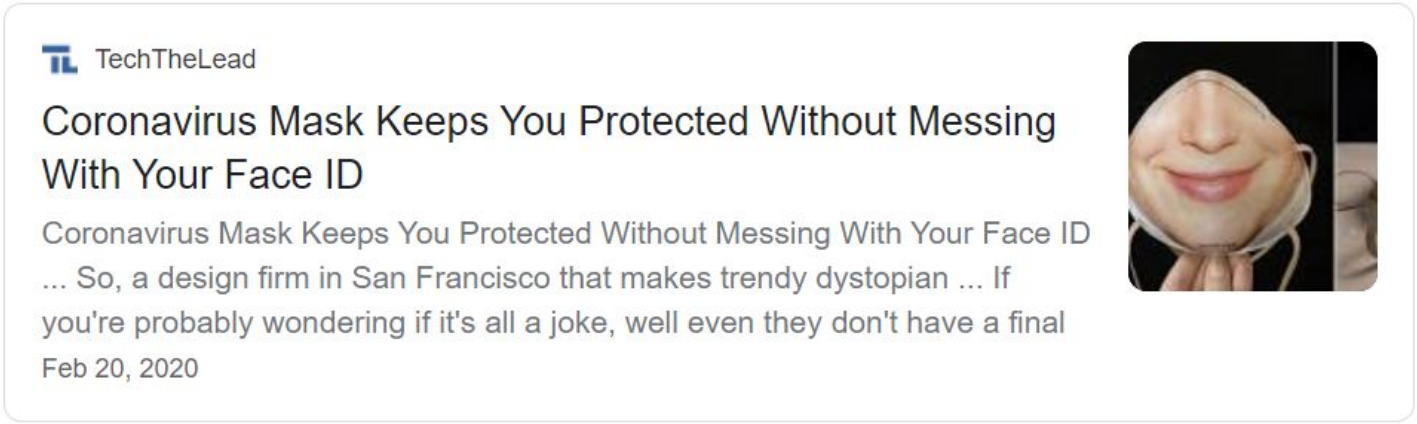


Fig. 1. *Top left:* one image (4 MPixels) corresponding to one of the 50 high power fields represented in the dataset. Our detected mitosis are circled green (true positives) and red (false positives); cyan denotes mitosis not detected by our approach. *Top right:* details of three areas (full-size results on the whole dataset in supplementary material). Note the challenging appearance of mitotic nuclei and other very similar non-mitotic structures. *Bottom:* overview of our detection approach.

Validating the model

- Validation data should reflect usage data
- Be aware of data/concept drift? (face recognition during pandemic, new patterns in credit card fraud detection)



Independence of data

Kaggle competition on detecting distracted drivers

Fairness: What is fair?

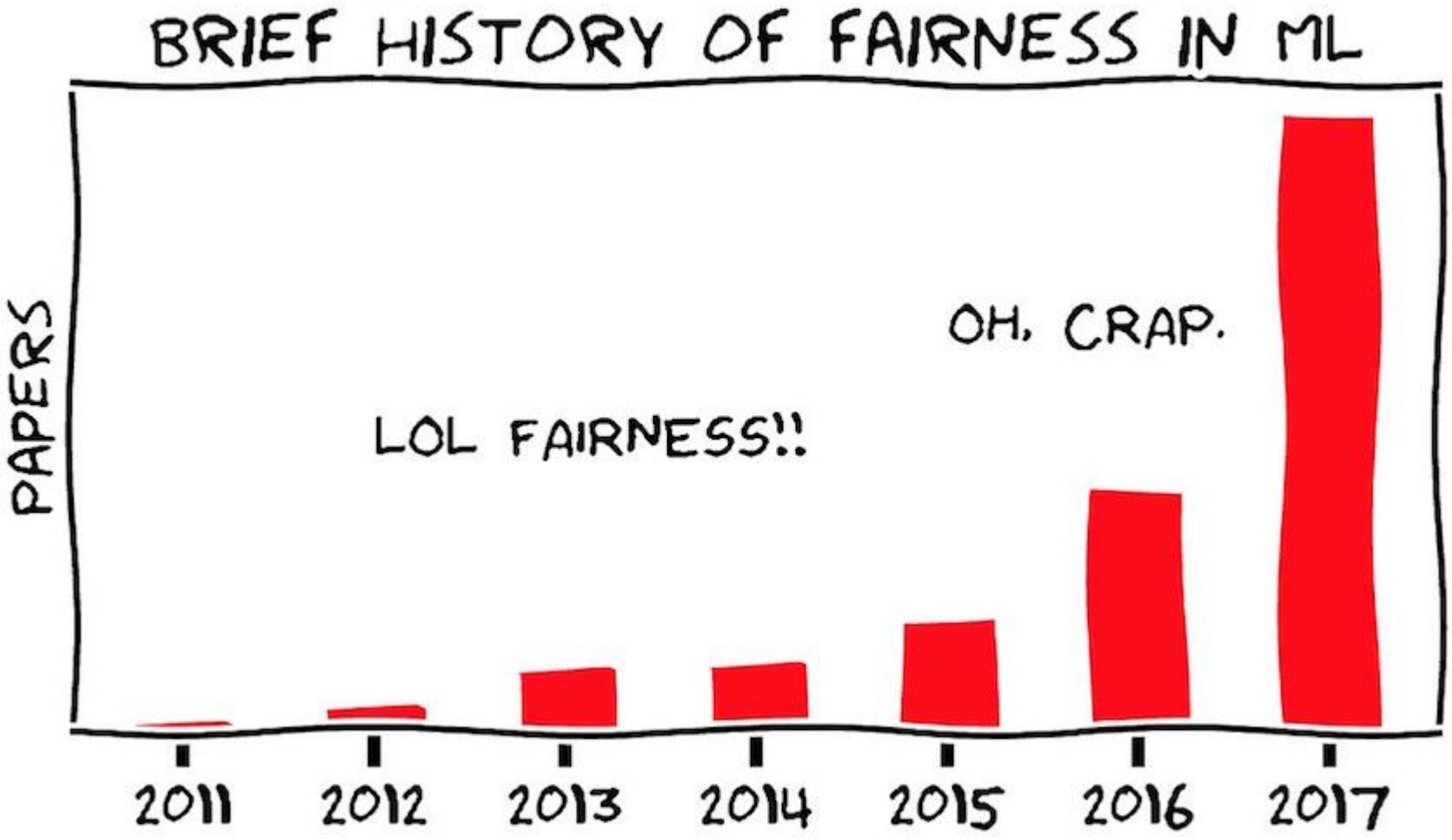
Fairness discourse asks questions about how to treat people and whether treating different groups of people differently is ethical. If two groups of people are systematically treated differently, this is often considered unfair.

Philosophy: "what is fair is also what is morally right."

Law: "protect individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories"

S3D

Fairness is still an actively studied & disputed concept!



Fairness:
Running Example - Mortgage Applications

- Running Example: Mortgage Applications
- Home ownership is key path to build generational wealth
- Past decisions often discriminatory (redlining)
- Replace biased human decisions by objective and more accurate ML model
  - income, other debt, home value
  - past debt and payment behavior (credit score)
- Reduce operational costs and turn times within the mortgage process.

FORBES > MONEY

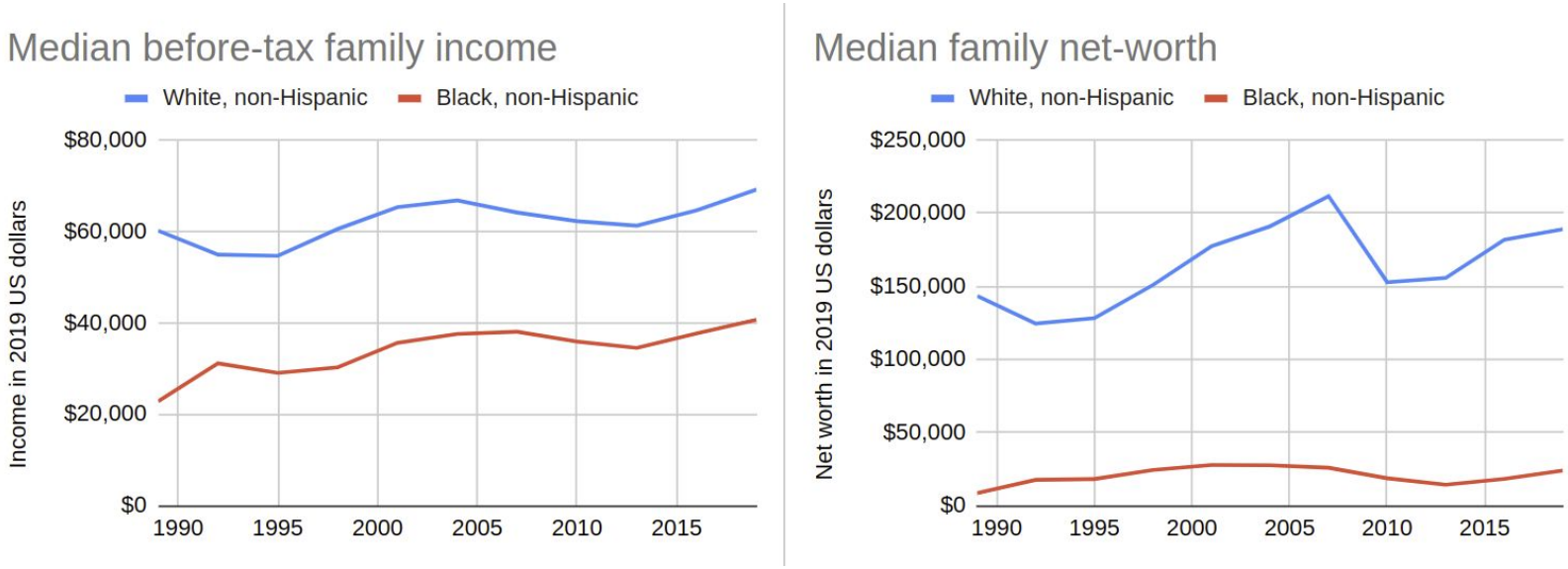
The Future Of Mortgage Lending: How AI And Humans Can Coexist

Alec Hanson Forbes Councils Member
Forbes Finance Council COUNCIL POST | Membership (Fee-Based)

Mar 9, 2023, 07:30am EST

S3D

Past bias, different starting positions



Source: Federal Reserve's Survey of Consumer Finances

S3D

Q. What is fair in mortgage applications?

Varieties of fairness

- Group unaware
- Demographic parity
- Equalized Odds

Varieties of fairness

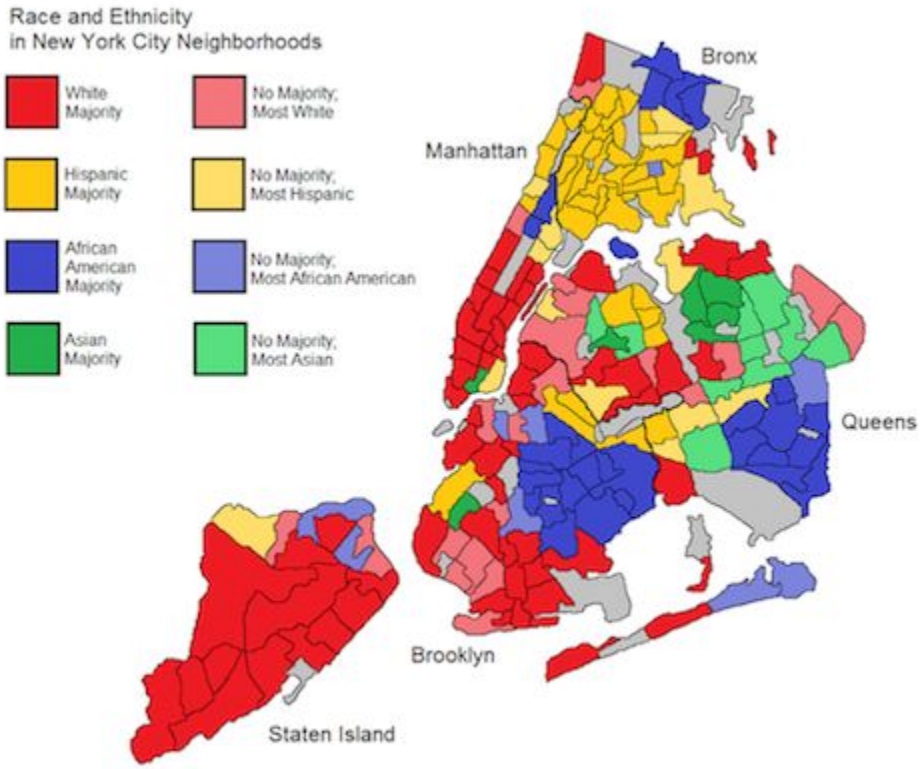
- **Group unaware (blindness)**
- Demographic parity
- Equalized Odds

Group Unaware

- Also called fairness through blindness or fairness through unawareness
- Ignore certain sensitive attributes when making a decision
- Example: Remove gender and race from mortgage model
- Easy to implement, but any limitations?

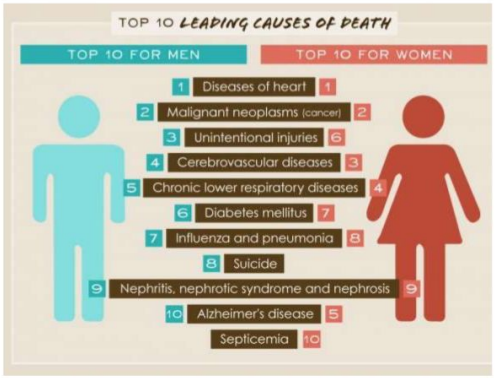
Group Unaware: Issues

- Proxies: Features correlate with protected attributes



Group Unaware:
Is all discrimination is harmful?

Group Unaware:
Not all discrimination is harmful



- Loan lending: Gender and racial discrimination is illegal.
- Medical diagnosis: Gender/race-specific diagnosis may be desirable.
- Discrimination is a domain-specific concept!

S3D

Ensuring Group Unawareness

- How to train models that are fair wrt. group unawareness?
  - Simply remove features for protected attributes from training and inference data
  - If you can't edit the model
    - Null/randomize protected attribute during inference
- How to test if models are fair wrt. group unawareness?
  - $\forall x.f(x[p \leftarrow 0]) = f(x[p \leftarrow 1])$
  - Test with any test data, e.g., purely random data or existing test data
  - Any single inconsistency shows that the protected attribute was used. Can also report percentage of inconsistencies.

Varieties of fairness

- Group unaware (blindness)
- **Demographic parity (independence)**
- Equalized odds

Demographic parity

Key idea: Compare outcomes across two groups

- Similar rates of accepted loans across racial/gender groups?
- Similar chance of being hired/promoted between gender groups?
- Similar rates of (predicted) recidivism across racial groups?
- Outcomes matter, not accuracy!

Varieties of fairness

- Group unaware (blindness)
- Demographic parity (independence)
- **Equalized odds (separation)**

Equalized Odds

Key idea: Focus on accuracy (not outcomes) across two groups

- Similar default rates on accepted loans across racial/gender groups?
- Similar rate of "bad hires" and "missed stars" between gender groups?
- Similar accuracy of predicted recidivism vs actual recidivism across racial groups?
- Accuracy matters, not outcomes!

Usually implemented by training different models