Responsible Data Science

The data science lifecycle

February 24 2025

Prof. Emily Black

Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

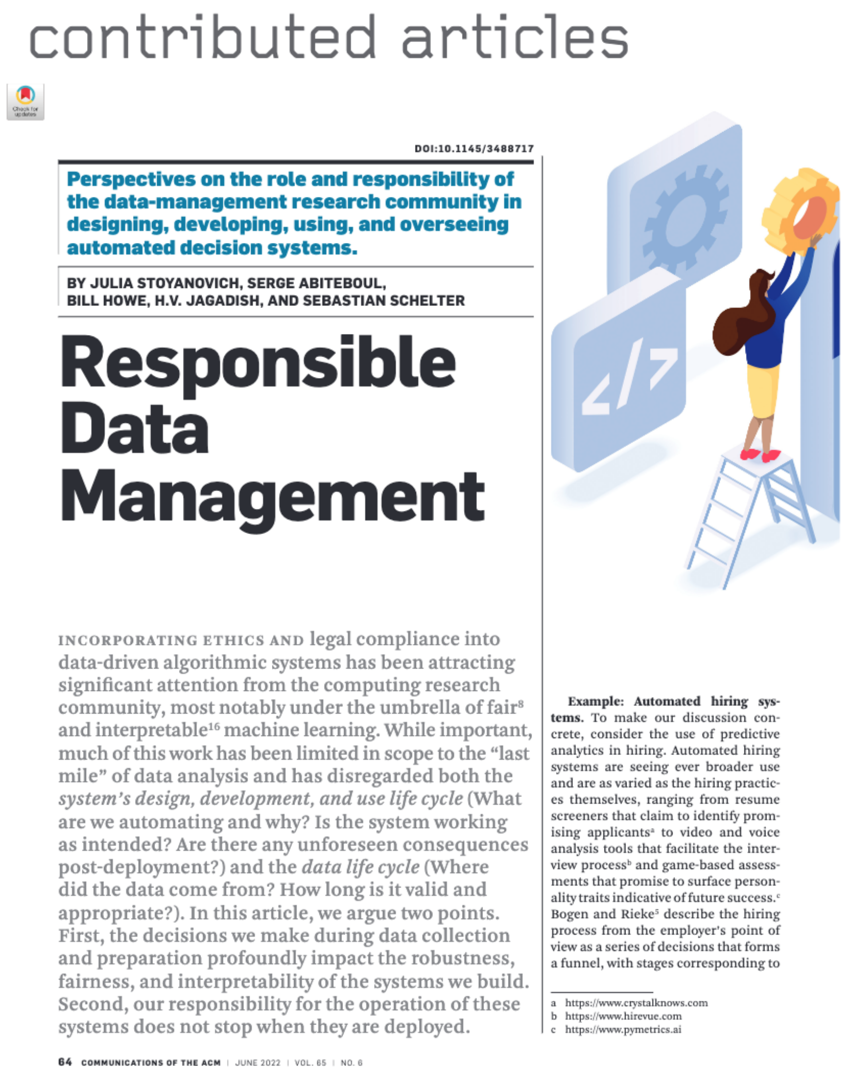
r/ai center for responsible ai

- https://docs.google.com/forms/d/e/1FAIpQLSe4f3DYVpuyz78SIqBY271EYx6RIl7G0-79w_laVCH_oCHhlQ/viewform

Outline

- Problem formulation: IRS
- Data collection: gender shades
- Data cleaning: shades of null
- Data bias: word embeddings, predict and serve
- Statistical modeling: bias amplification
- Testing and validation: d-hacking/GenAI stuff
- Model integration & deployment: something on the huma—AI paper (I’ll have to make that later)

This week's reading



contributed articles

DOI:10.1145/3488717

Perspectives on the role and responsibility of the data-management research community in designing, developing, using, and overseeing automated decision systems.

BY JULIA STOYANOVICH, SERGE ABITEBOUL, BILL HOWE, H.V. JAGADISH, AND SEBASTIAN SCHELTER

Responsible Data Management

INCORPORATING ETHICS AND legal compliance into data-driven algorithmic systems has been attracting significant attention from the computing research community, most notably under the umbrella of fair[8] and interpretable[16] machine learning. While important, much of this work has been limited in scope to the "last mile" of data analysis and has disregarded both the *system's design, development, and use life cycle* (What are we automating and why? Is the system working as intended? Are there any unforeseen consequences post-deployment?) and the *data life cycle* (Where did the data come from? How long is it valid and appropriate?). In this article, we argue two points. First, the decisions we make during data collection and preparation profoundly impact the robustness, fairness, and interpretability of the systems we build. Second, our responsibility for the operation of these systems does not stop when they are deployed.

Example: Automated hiring systems. To make our discussion concrete, consider the use of predictive analytics in hiring. Automated hiring systems are seeing ever broader use and are as varied as the hiring practices themselves, ranging from resume screeners that claim to identify promising applicants[a] to video and voice analysis tools that facilitate the interview process[b] and game-based assessments that promise to surface personality traits indicative of future success.[c] Bogen and Rieke[5] describe the hiring process from the employer's point of view as a series of decisions that forms a funnel, with stages corresponding to

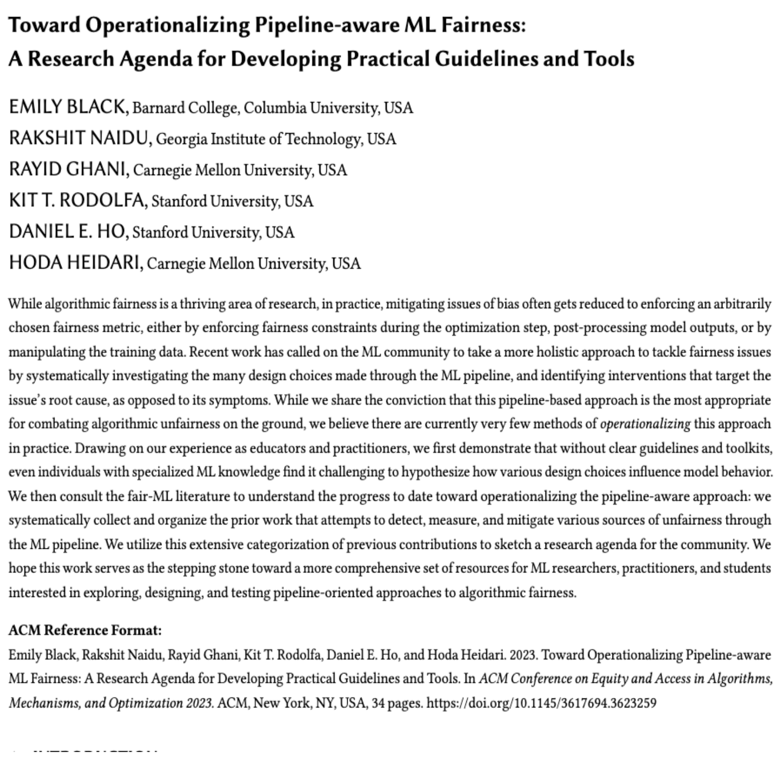
[a] https://www.crystalknows.com
[b] https://www.hirevue.com
[c] https://www.pymetrics.ai

64 COMMUNICATIONS OF THE ACM | JUNE 2022 | VOL. 65 | NO. 6

IN DETAIL

To predict and serve?

Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data? **Kristian Lum** and **William Isaac** consider the evidence – and the social consequences



Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools

EMILY BLACK, Barnard College, Columbia University, USA
RAKSHIT NAIDU, Georgia Institute of Technology, USA
RAYID GHANI, Carnegie Mellon University, USA
KIT T. RODOLFA, Stanford University, USA
DANIEL E. HO, Stanford University, USA
HODA HEIDARI, Carnegie Mellon University, USA

While algorithmic fairness is a thriving area of research, in practice, mitigating issues of bias often gets reduced to enforcing an arbitrarily chosen fairness metric, either by enforcing fairness constraints during the optimization step, post-processing model outputs, or by manipulating the training data. Recent work has called on the ML community to take a more holistic approach to tackle fairness issues by systematically investigating the many design choices made through the ML pipeline, and identifying interventions that target the issue's root cause, as opposed to its symptoms. While we share the conviction that this pipeline-based approach is the most appropriate for combating algorithmic unfairness on the ground, we believe there are currently very few methods of *operationalizing* this approach in practice. Drawing on our experience as educators and practitioners, we first demonstrate that without clear guidelines and toolkits, even individuals with specialized ML knowledge find it challenging to hypothesize how various design choices influence model behavior. We then consult the fair-ML literature to understand the progress to date toward operationalizing the pipeline-aware approach: we systematically collect and organize the prior work that attempts to detect, measure, and mitigate various sources of unfairness through the ML pipeline. We utilize this extensive categorization of previous contributions to sketch a research agenda for the community. We hope this work serves as the stepping stone toward a more comprehensive set of resources for ML researchers, practitioners, and students interested in exploring, designing, and testing pipeline-oriented approaches to algorithmic fairness.

ACM Reference Format:
Emily Black, Rakshit Naidu, Rayid Ghani, Kit T. Rodolfa, Daniel E. Ho, and Hoda Heidari. 2023. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization 2023.* ACM, New York, NY, USA, 34 pages. https://doi.org/10.1145/3617694.3623259

Recall: Bias in computer systems

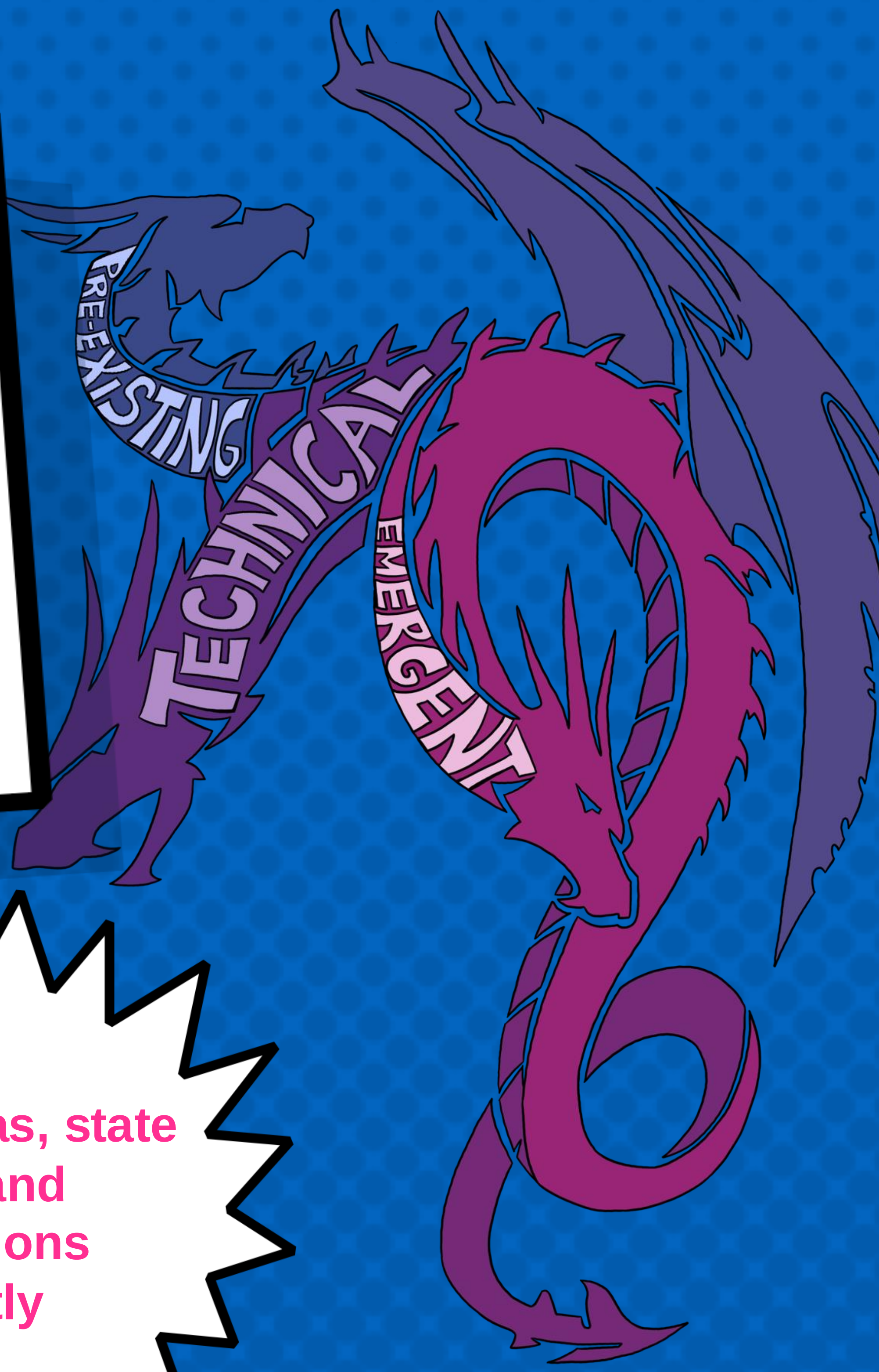
Pre-existing is independent of an algorithm and has origins in society

Technical is introduced or exacerbated by the technical properties of an ADS

Emergent arises due to context of use

[Friedman & Nissenbaum (1996)]

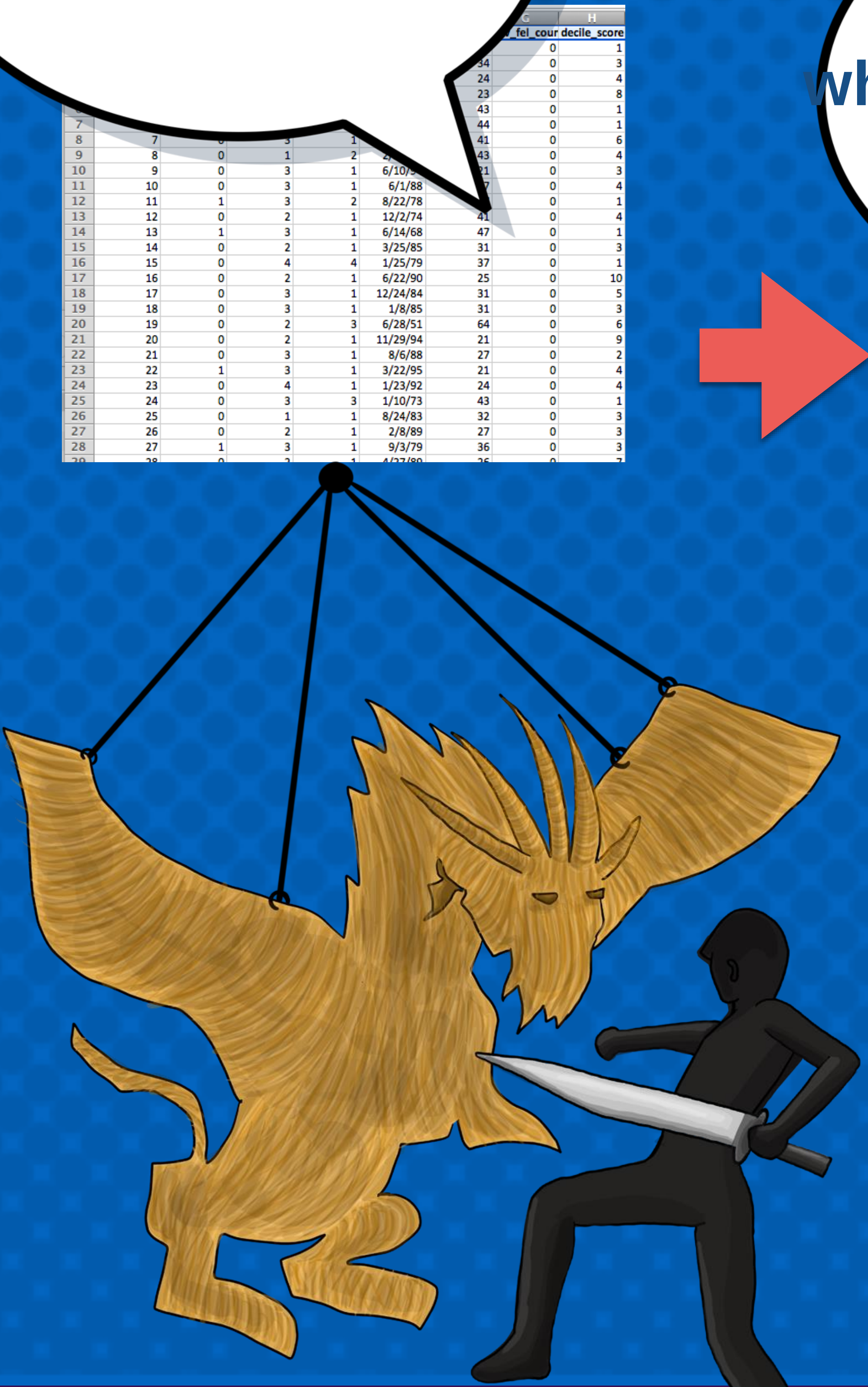
to fight bias, state beliefs and assumptions explicitly



r/ai center for responsible ai

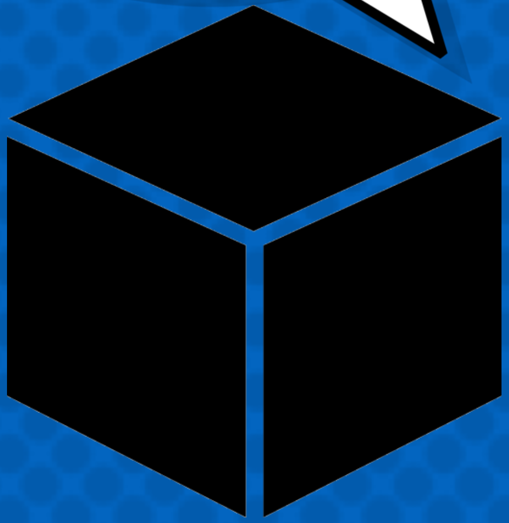
The “last-mile” view of responsible AI

where did the data come from?

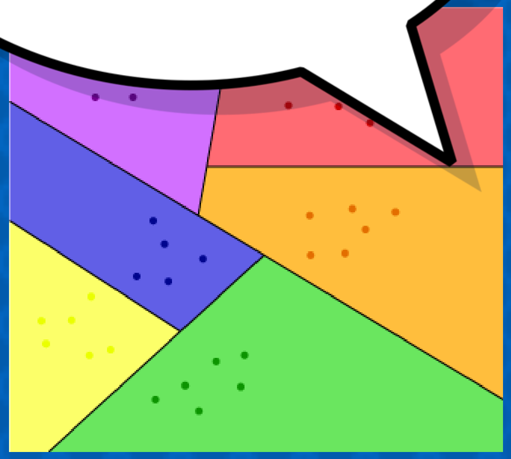


							G	H
							_fel_cour	decile_score
							0	1
						34	0	3
						24	0	4
						23	0	8
						43	0	1
7						44	0	1
8	7		3	1		41	0	6
9	8	0	1	2		43	0	4
10	9	0	3	1	6/10/	21	0	3
11	10	0	3	1	6/1/88		0	4
12	11	1	3	2	8/22/78		0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3

what happens inside the box?

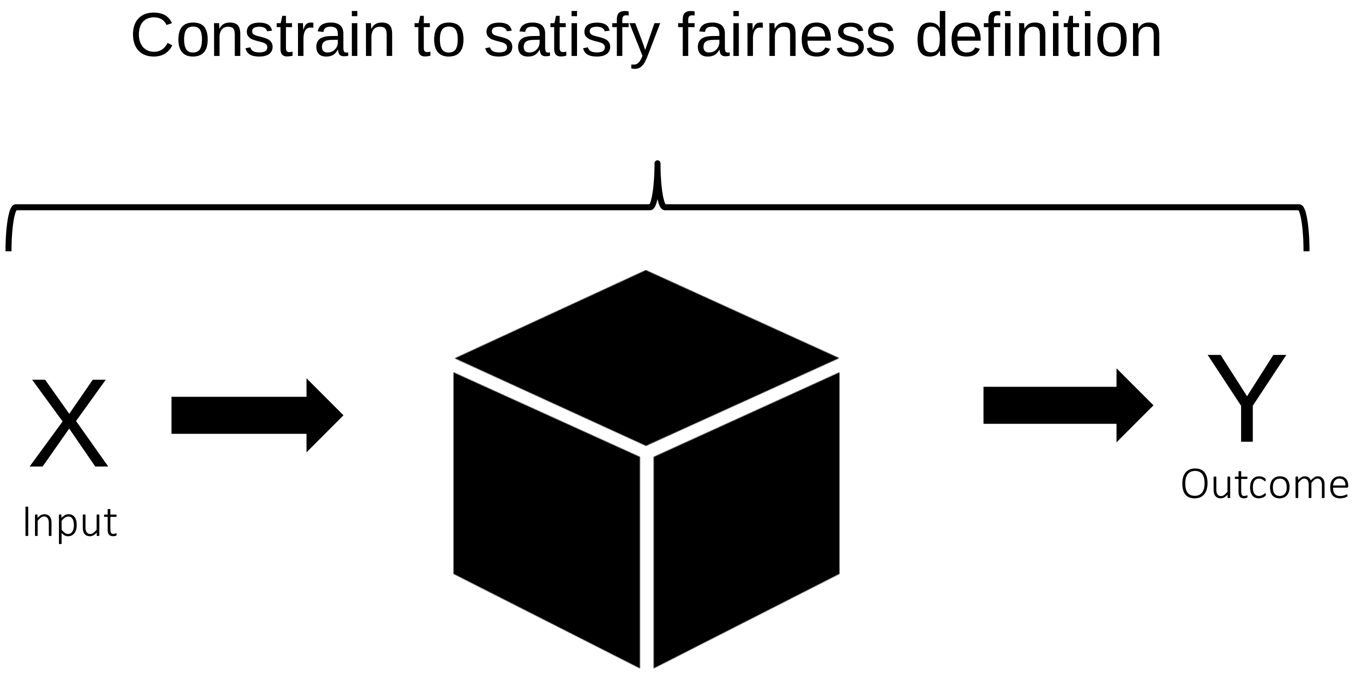


how are results used?



r/ai center for responsible ai

Common approaches to ML fairness



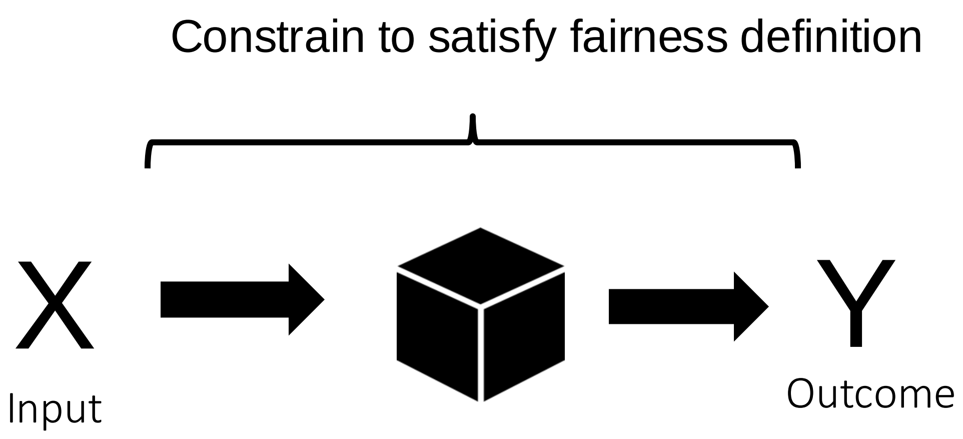
Fairness Definitions

$\hat{Y}$: model prediction
Y : ground truth
Z: protected class (e.g. race)

- Demographic Parity → $P(\hat{Y} = 1 \mid Z = z) = P(\hat{Y} = 1 \mid Z = z')$
- Equal False Positive Rates → $P(\hat{Y} = 1 \mid Y = 0, Z = z) = P(\hat{Y} = 1 \mid Y = 0, Z = z')$
- Equal Accuracy → $P(\hat{Y} = Y \mid Z = z) = P(\hat{Y} = Y \mid Z = z')$

e.g. Dwork et al. 2012, Hardt et al. 2016, Chouldechova 2016, Berk et al 2017

A common approach to AI fairness



Constrain to satisfy fairness definition

X

Input

Y

Outcome

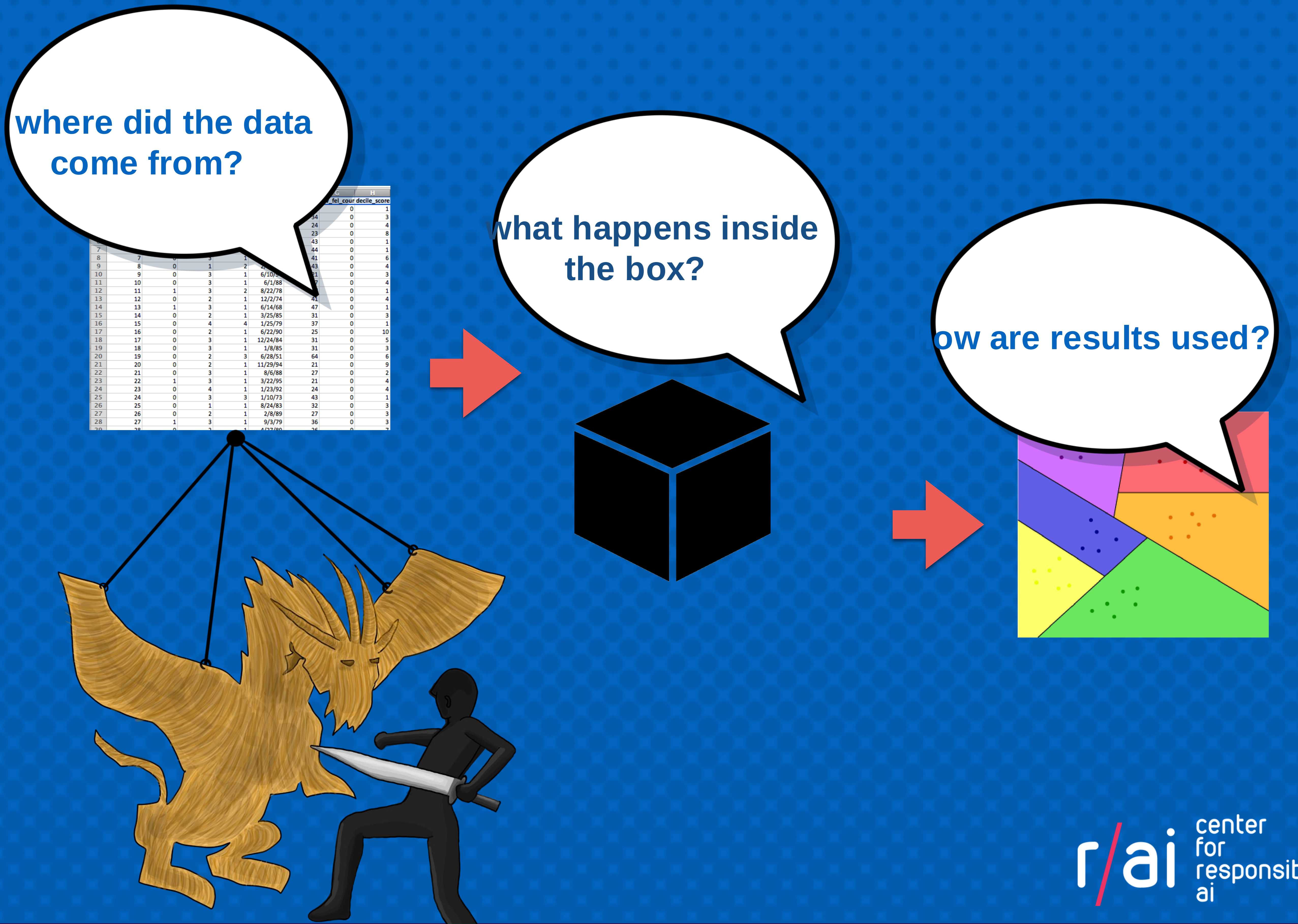
Pros:

- Generalizable across contexts
- Because of this, often easily deployable

Cons:

- May not address certain unfair behaviors
- Often does not target *core* issue
- **Often not legally viable in regulated contexts**

The “last-mile” view of responsible AI



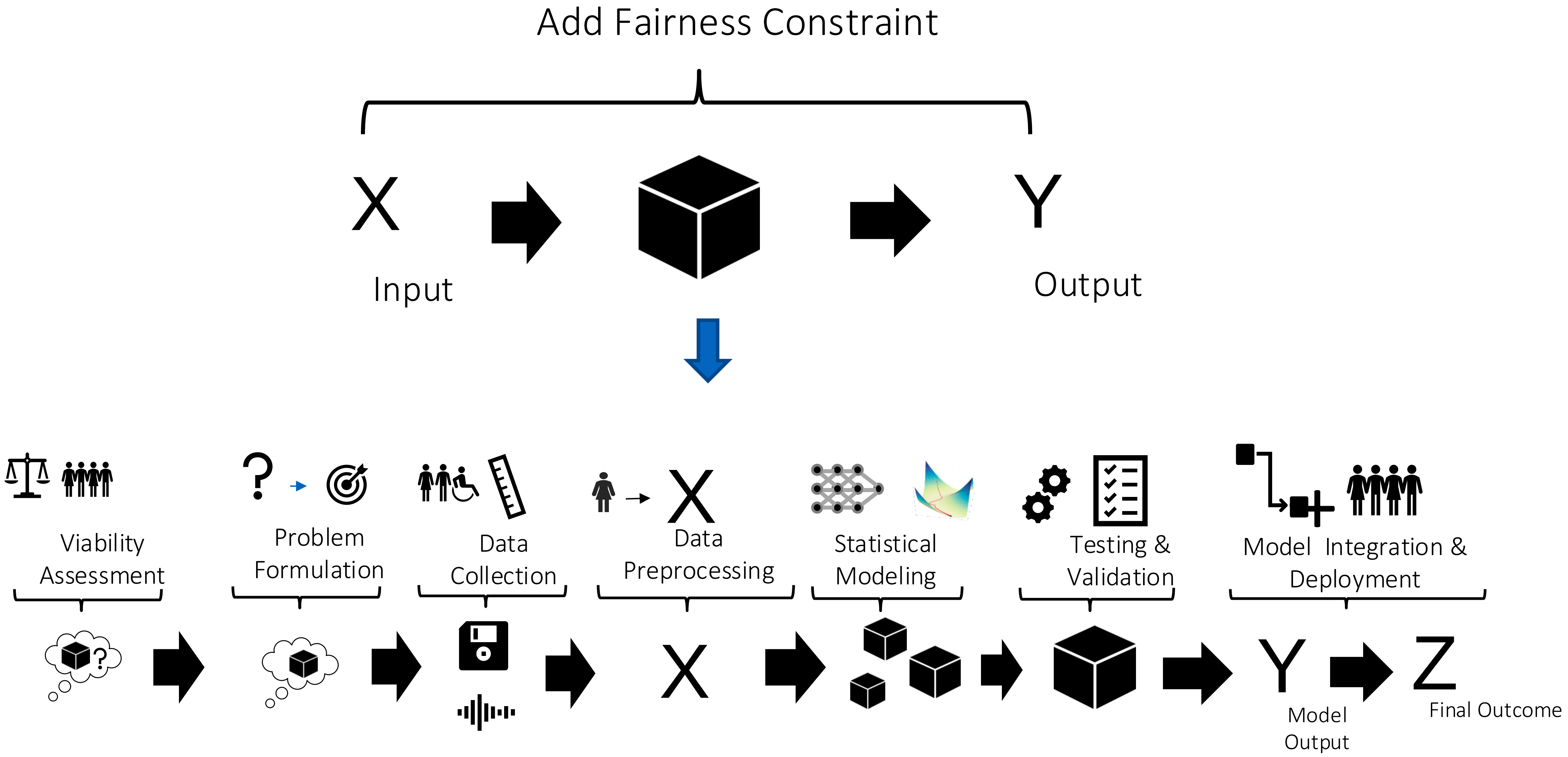
where did the data come from?

what happens inside the box?

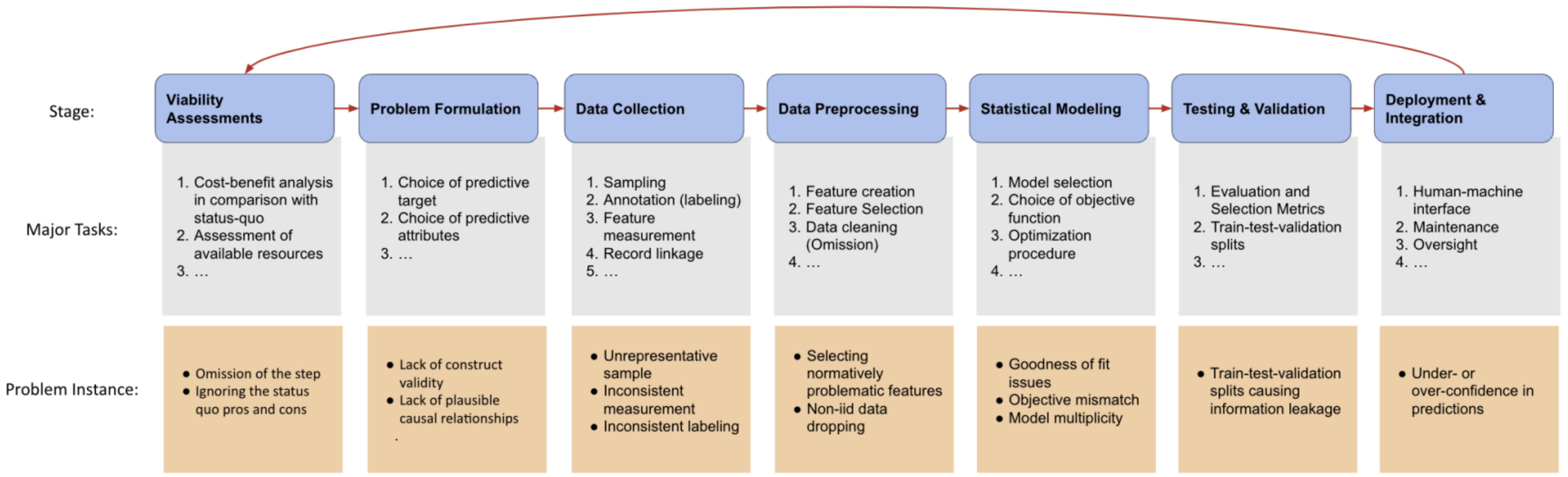
ow are results used?

center for responsible ai

There's a lot more going on under the hood



Examples of Pipeline Steps



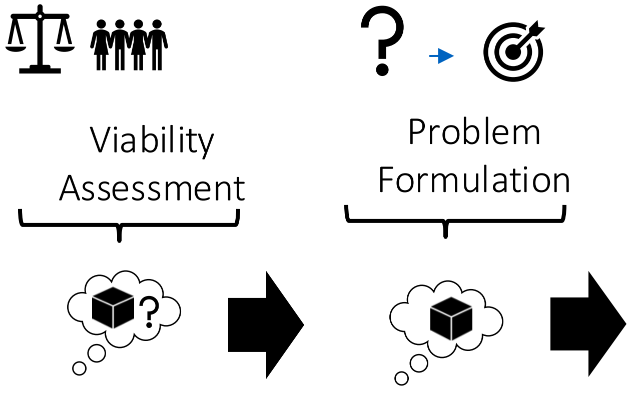
Stage:	**Viability Assessments**	**Problem Formulation**	**Data Collection**	**Data Preprocessing**	**Statistical Modeling**	**Testing & Validation**	**Deployment & Integration**
Major Tasks:	1. Cost-benefit analysis in comparison with status-quo 2. Assessment of available resources 3. …	1. Choice of predictive target 2. Choice of predictive attributes 3. …	1. Sampling 2. Annotation (labeling) 3. Feature measurement 4. Record linkage 5. …	1. Feature creation 2. Feature Selection 3. Data cleaning (Omission) 4. …	1. Model selection 2. Choice of objective function 3. Optimization procedure 4. …	1. Evaluation and Selection Metrics 2. Train-test-validation splits 3. …	1. Human-machine interface 2. Maintenance 3. Oversight 4. …
Problem Instance:	• Omission of the step • Ignoring the status quo pros and cons	• Lack of construct validity • Lack of plausible causal relationships	• Unrepresentative sample • Inconsistent measurement • Inconsistent labeling	• Selecting normatively problematic features • Non-iid data dropping	• Goodness of fit issues • Objective mismatch • Model multiplicity	• Train-test-validation splits causing information leakage	• Under- or over-confidence in predictions

Fig. 1. A simplified view of the ML pipeline, its key stages, and instances of design choices made at each stage.

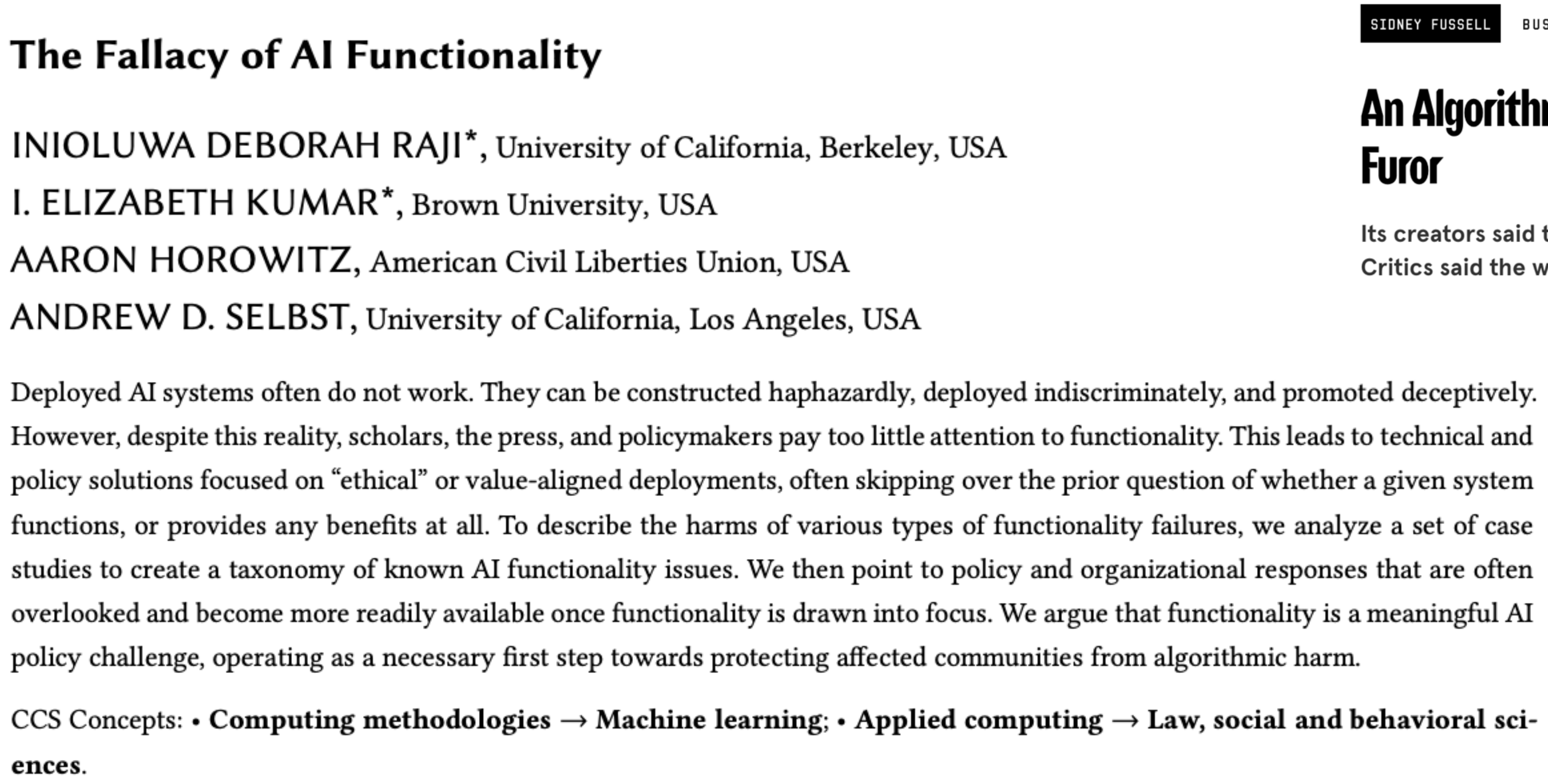
Viability Assessment, Problem Formulation

- Viability assessment: *is making an AI decision-making system a good idea? Do we have the resources to do so?*

- Problem formulation: *How can we use the data we have available to predict what we want to predict-- like "creditworthiness" or "employability"?*



Viability Assessment, Problem Formulation

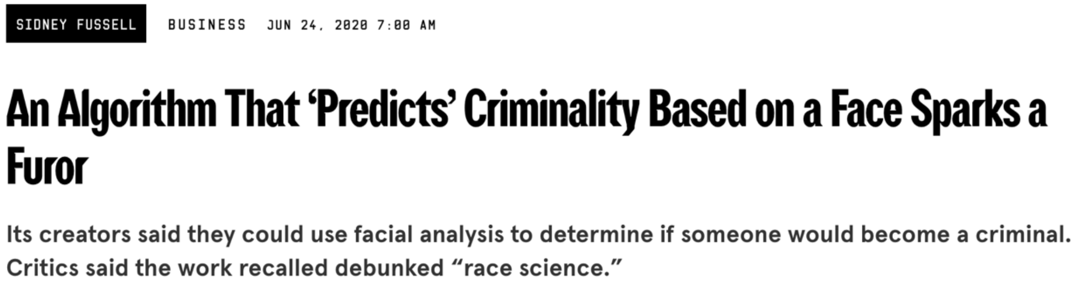


The Fallacy of AI Functionality

INIOLUWA DEBORAH RAJI*, University of California, Berkeley, USA
I. ELIZABETH KUMAR*, Brown University, USA
AARON HOROWITZ, American Civil Liberties Union, USA
ANDREW D. SELBST, University of California, Los Angeles, USA

Deployed AI systems often do not work. They can be constructed haphazardly, deployed indiscriminately, and promoted deceptively. However, despite this reality, scholars, the press, and policymakers pay too little attention to functionality. This leads to technical and policy solutions focused on "ethical" or value-aligned deployments, often skipping over the prior question of whether a given system functions, or provides any benefits at all. To describe the harms of various types of functionality failures, we analyze a set of case studies to create a taxonomy of known AI functionality issues. We then point to policy and organizational responses that are often overlooked and become more readily available once functionality is drawn into focus. We argue that functionality is a meaningful AI policy challenge, operating as a necessary first step towards protecting affected communities from algorithmic harm.

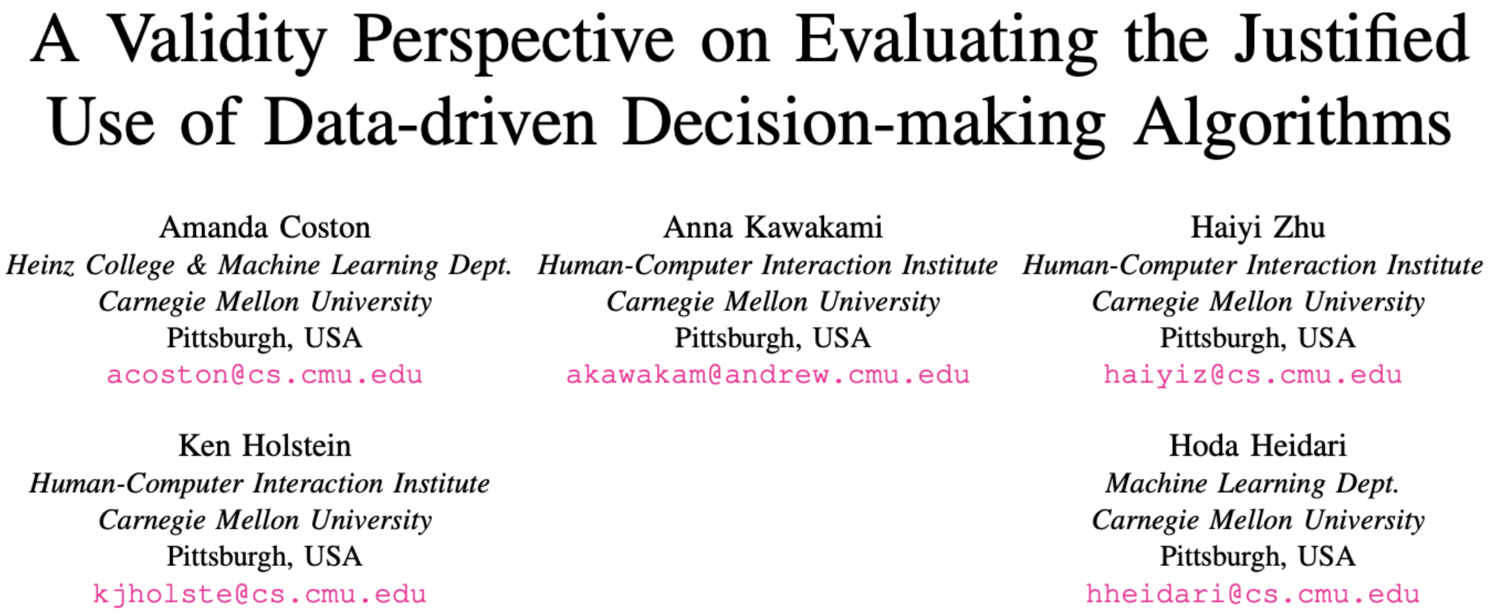
CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Applied computing** → **Law, social and behavioral sciences**.



SIDNEY FUSSELL BUSINESS JUN 24, 2020 7:00 AM

An Algorithm That 'Predicts' Criminality Based on a Face Sparks a Furor

Its creators said they could use facial analysis to determine if someone would become a criminal. Critics said the work recalled debunked "race science."



A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms

Amanda Coston
Heinz College & Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
acoston@cs.cmu.edu

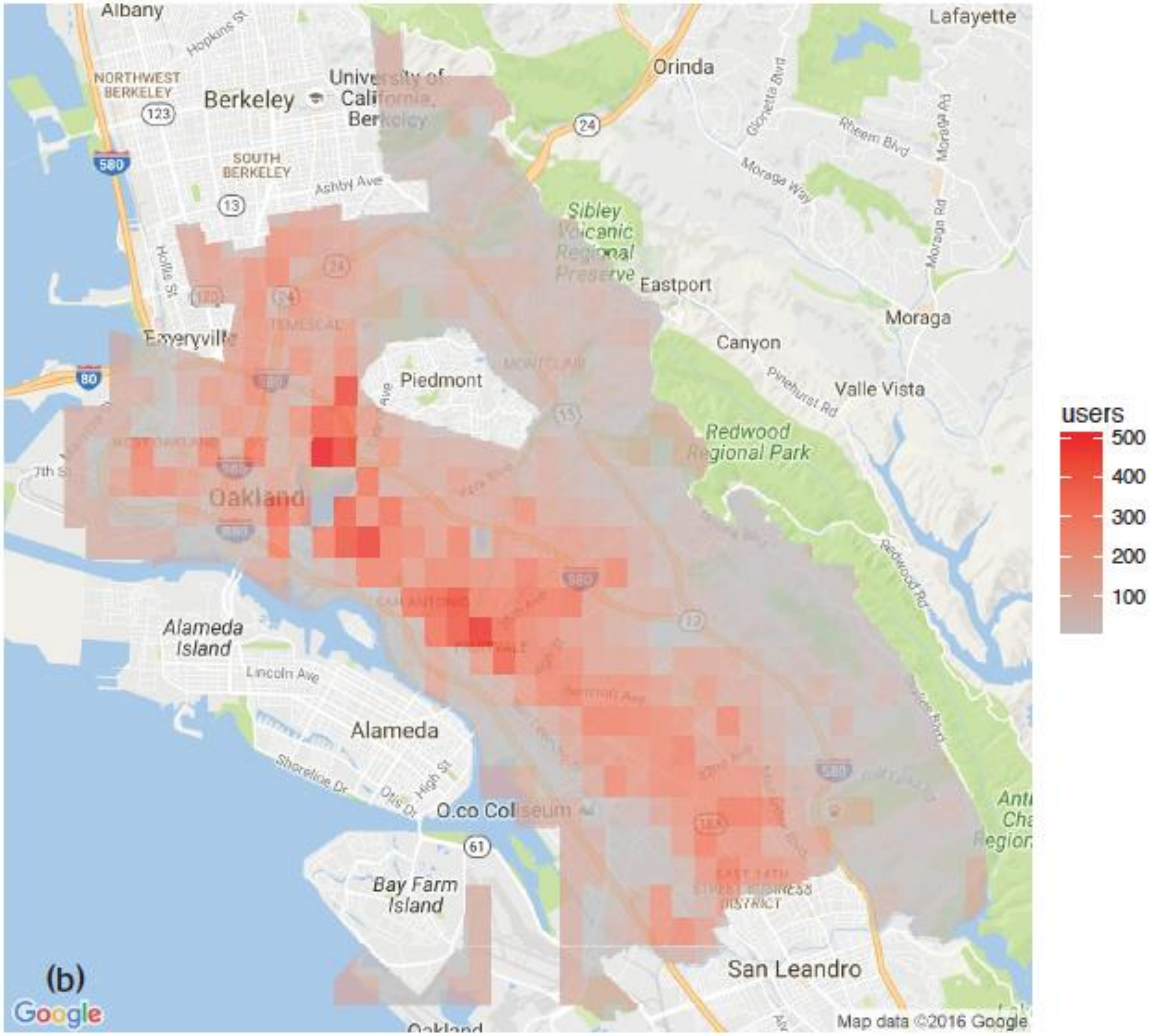
Anna Kawakami
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, USA
akawakam@andrew.cmu.edu

Haiyi Zhu
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, USA
haiyiz@cs.cmu.edu

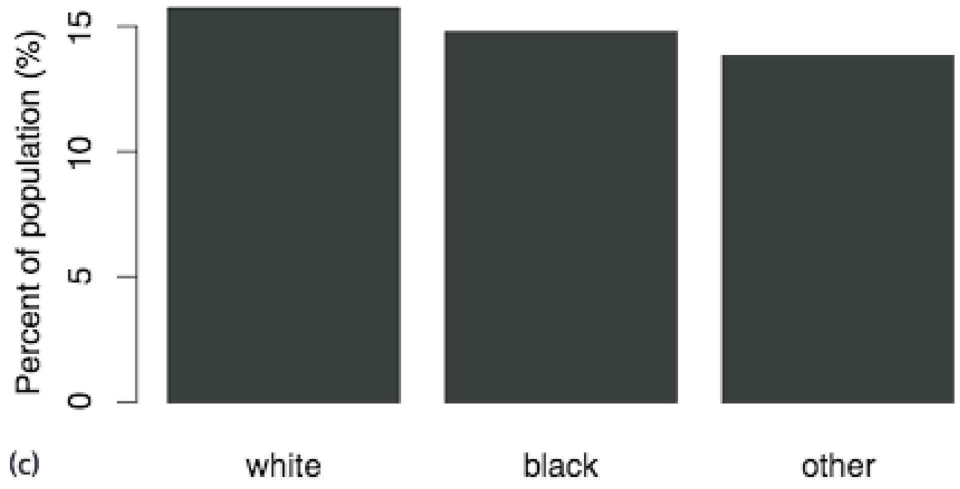
Ken Holstein
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, USA
kjholste@cs.cmu.edu

Hoda Heidari
Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
hheidari@cs.cmu.edu

Is my data biased? (histograms + geo)



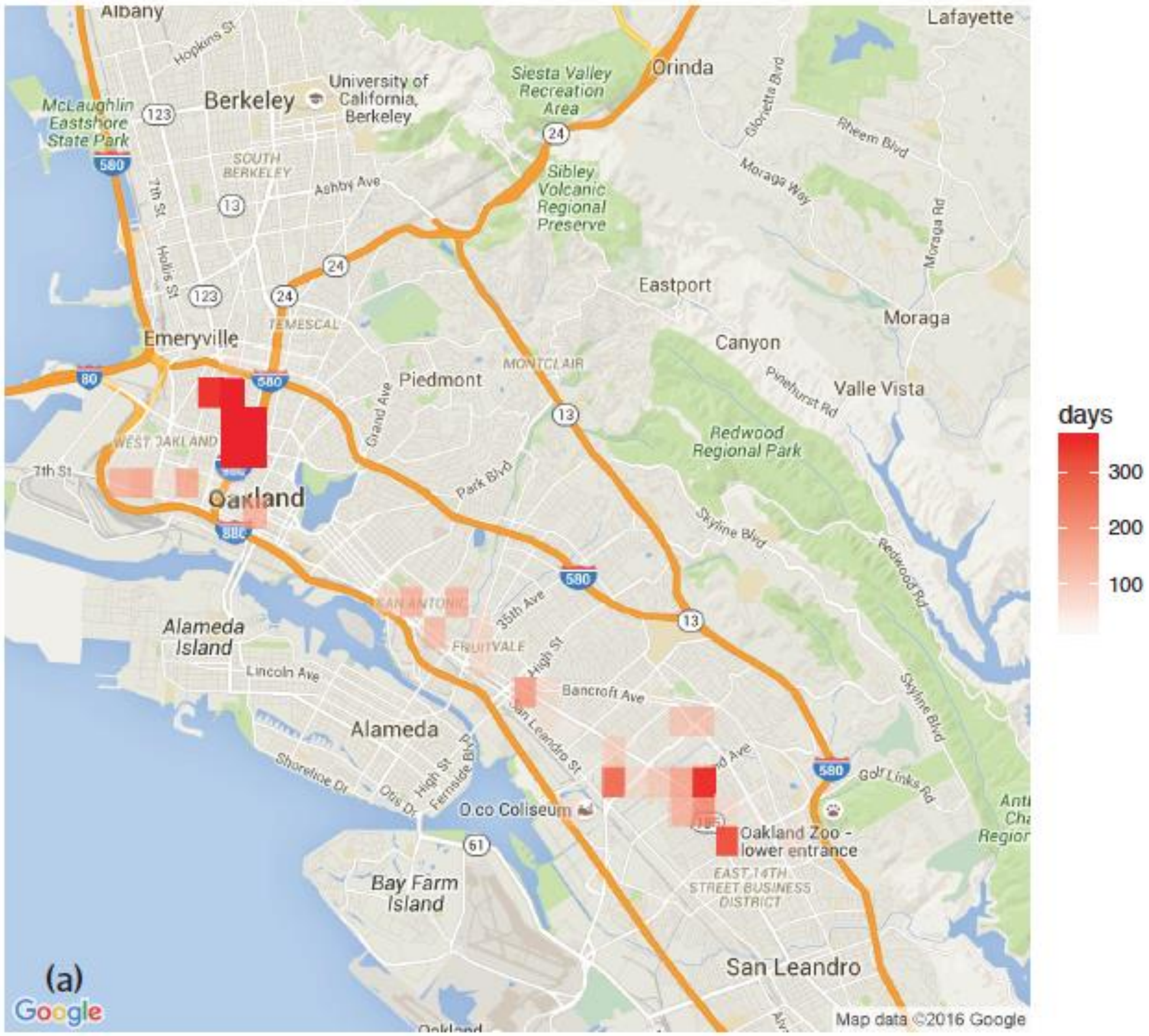
Estimated number of drug users, based on 2011 National Survey on Drug Use and Health, in Oakland, CA



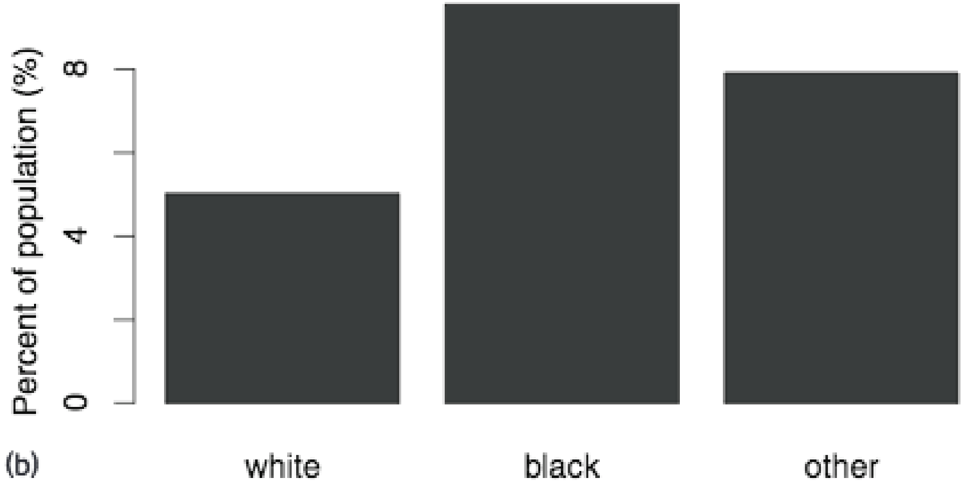
Estimated drug use by race

[Lum & Isaac (2016)]

Is my data biased? (histograms + geo)



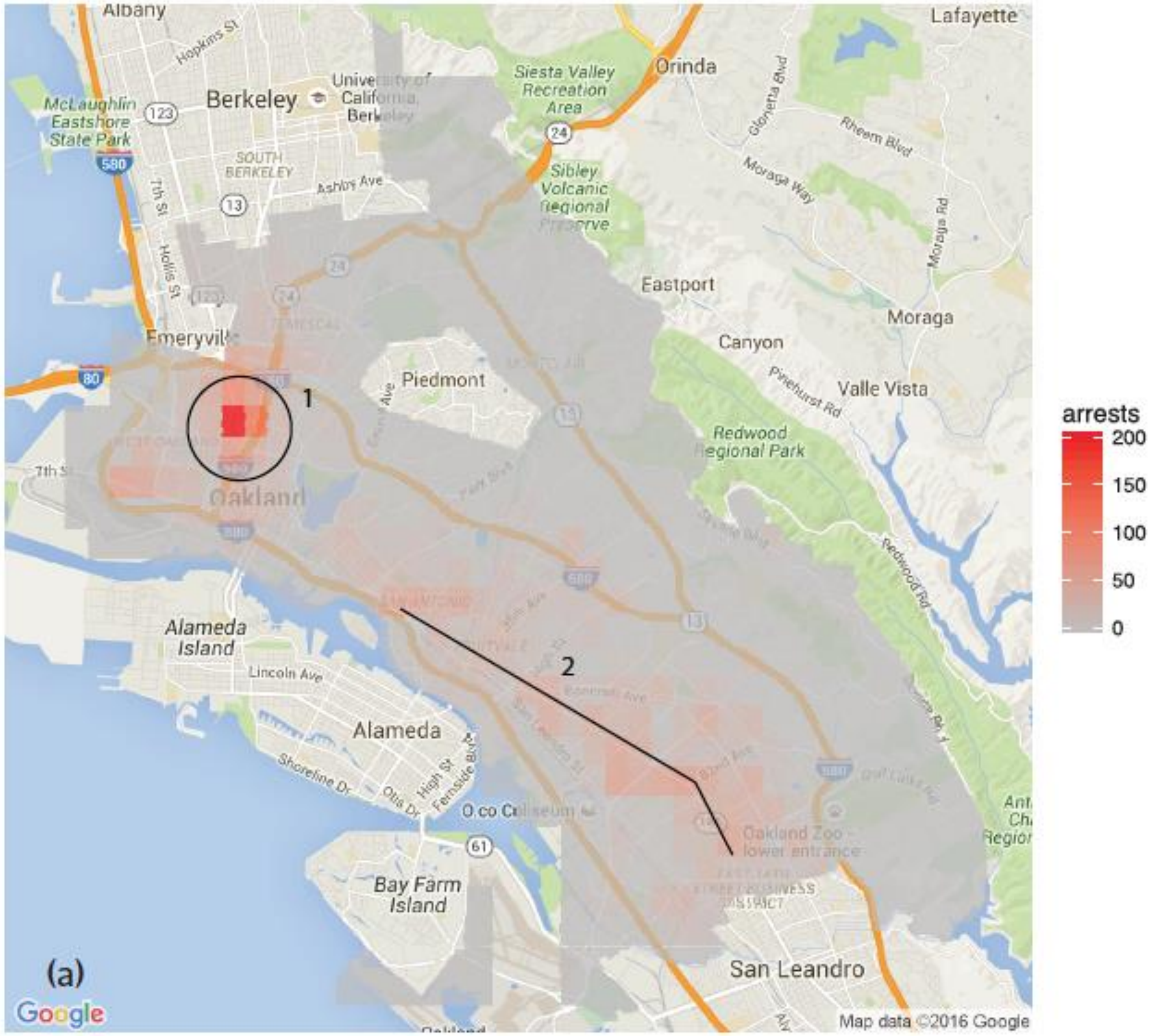
Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland, CA, police data for 2011



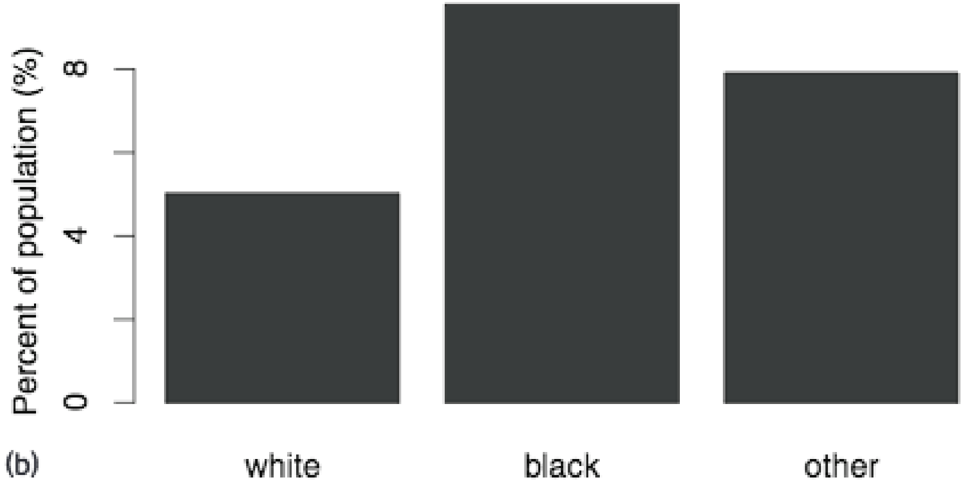
Targeted policing for drug crimes by race

[Lum & Isaac (2016)]

Is my data biased? (histograms + geo)



Number of drug arrests made by the Oakland, CA, police department in 2010



Targeted policing for drug crimes by race

[Lum & Isaac (2016)]

Data Collection

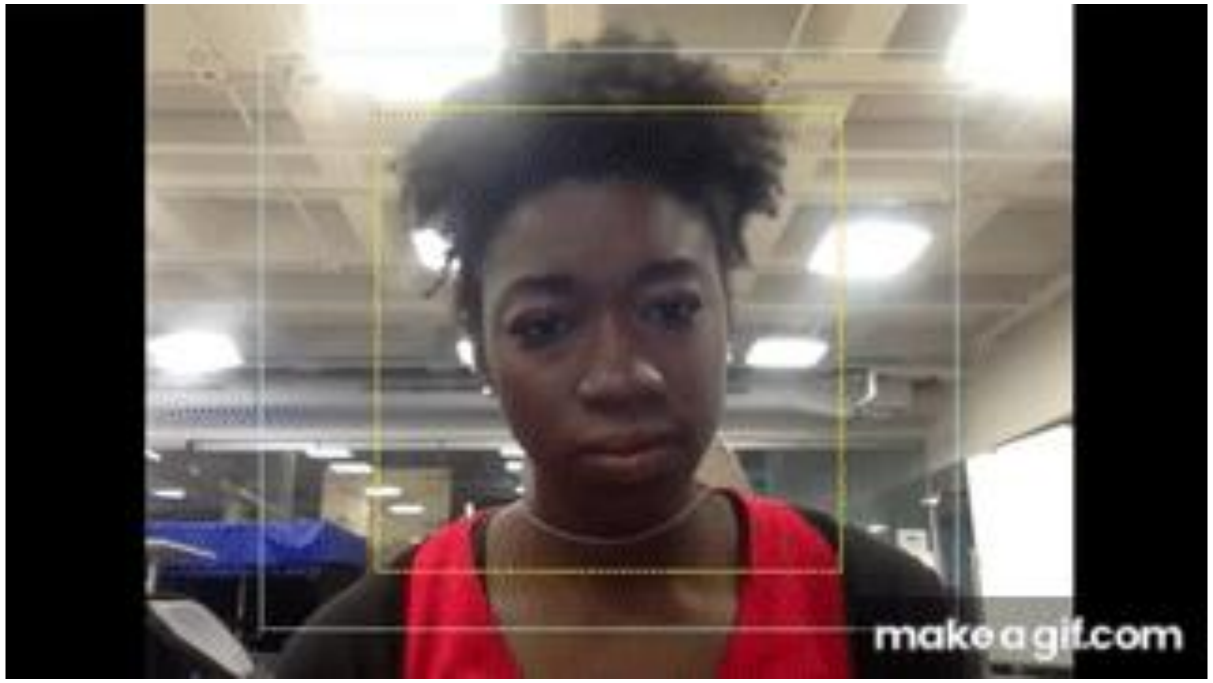
- Data collection: collecting or compiling data to train the model

Data
Collection

- *What population will we sample to build our model?*
- *How will we collect this data?*

Unfairness in Data Collection

- Facial Recognition technologies are notoriously bad at recognizing darker faces



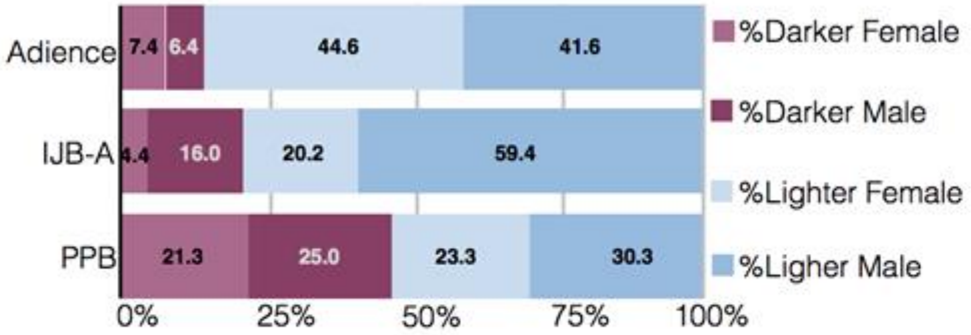
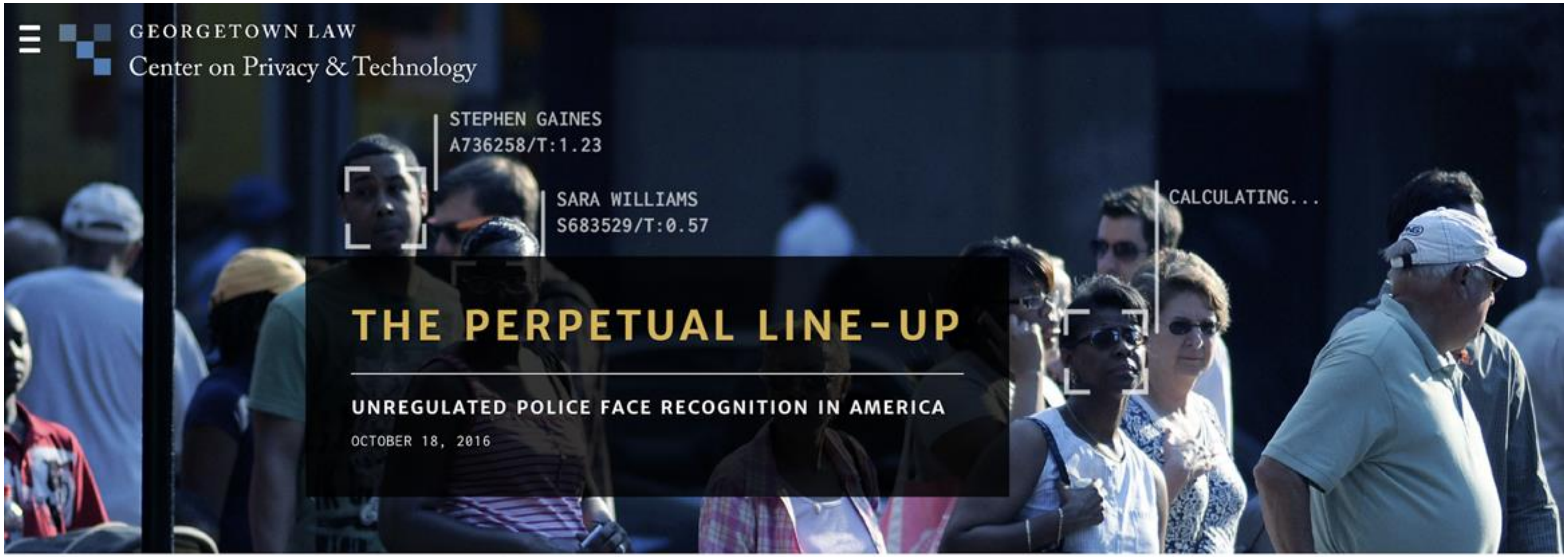


Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

Complications: Should we fix that?

- What are the implications of accurate facial recognition technology?

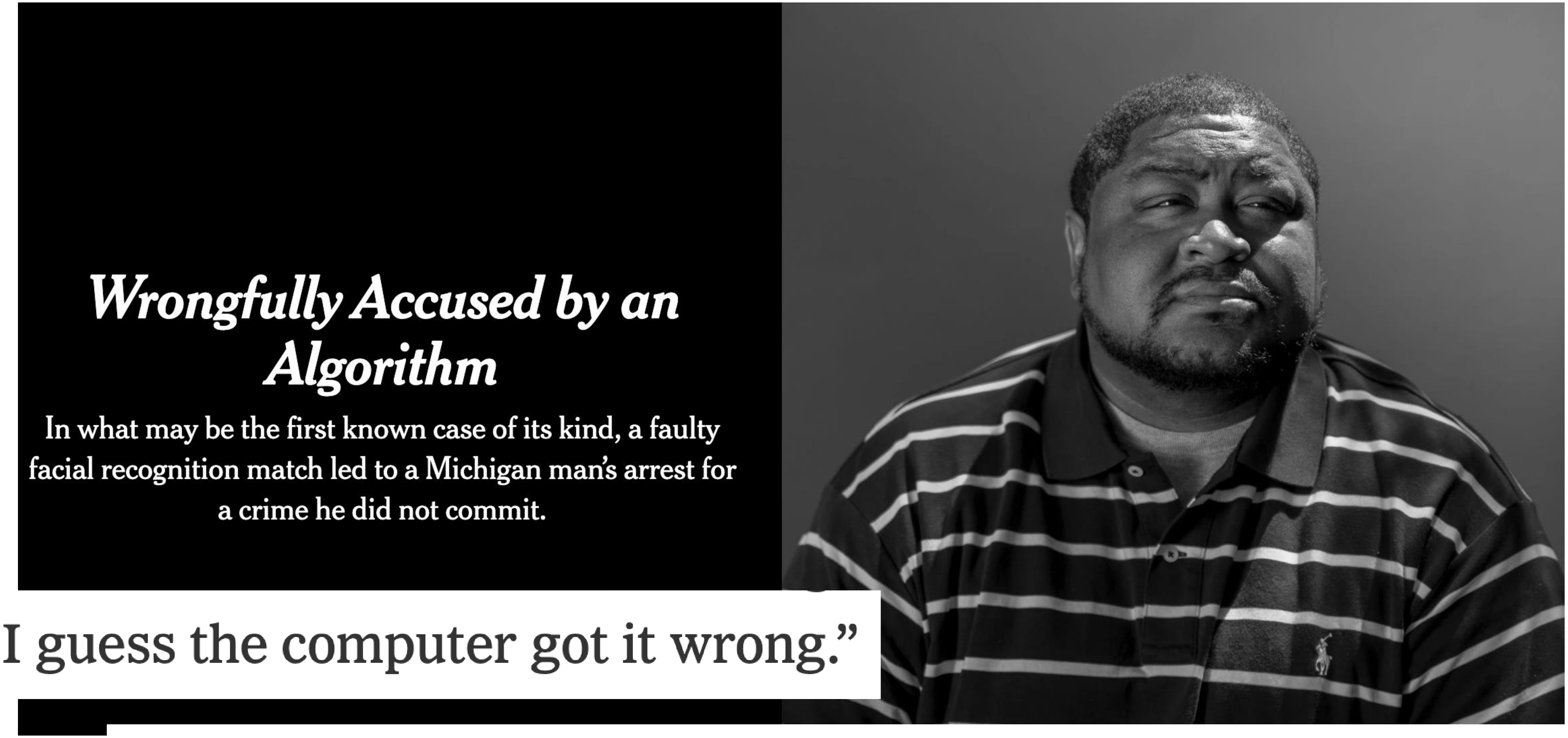


GEORGETOWN LAW
Center on Privacy & Technology

THE PERPETUAL LINE-UP

UNREGULATED POLICE FACE RECOGNITION IN AMERICA

OCTOBER 18, 2016



Wrongfully Accused by an Algorithm

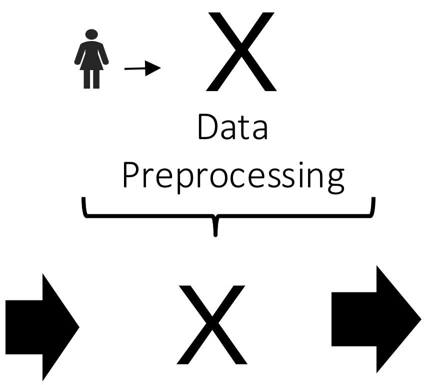
In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

"I guess the computer got it wrong."

Mr. Williams asked if he was free to go. "Unfortunately not," one detective said.

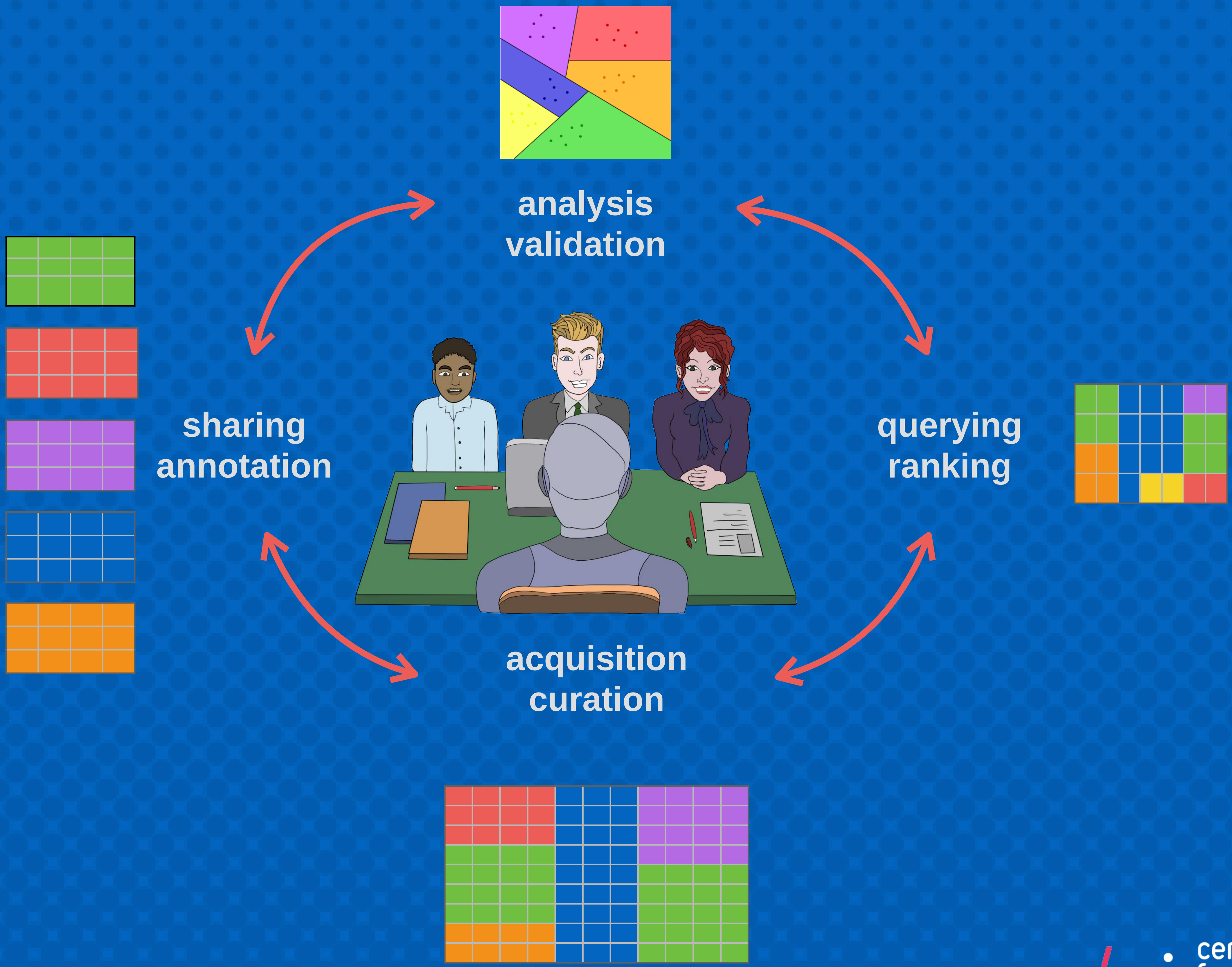
Data Pre-processing

- Data preprocessing: steps taken to make data usable by the AI model



- *How do we construct features from our data?*
- *What do we do with missing or incomplete data?*

Zooming out to the lifecycle view



analysis
validation

sharing
annotation

querying
ranking

acquisition
curation

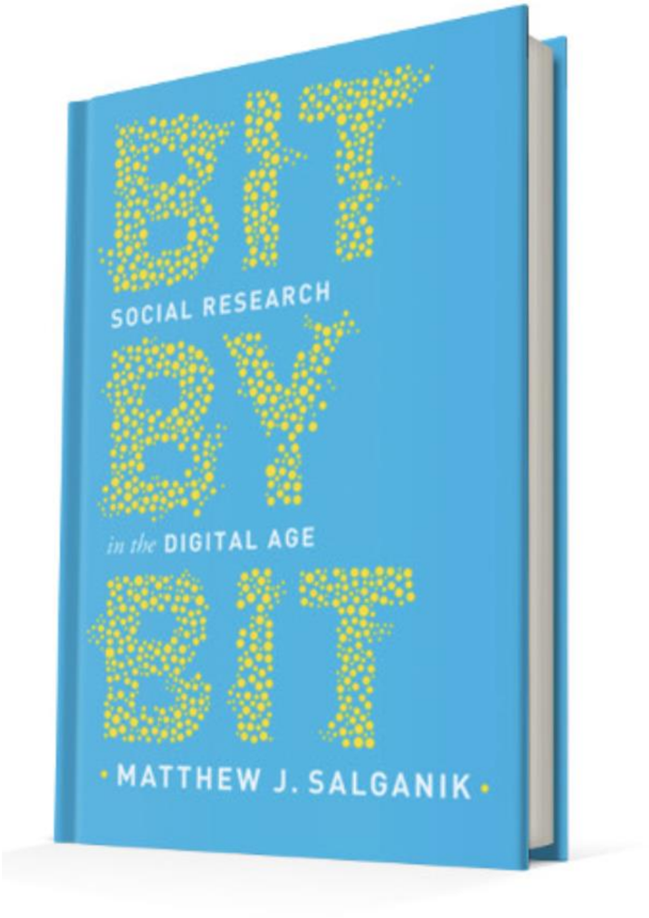
r/ai center for responsible ai

Understand your data!

"Given the heterogeneity of the flood of data, it is **not enough merely to record it and throw it into a repository**. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is **unlikely anyone will ever be able to find, let alone reuse**, any of this data. With adequate **metadata**, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure."

https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

Understand your data!

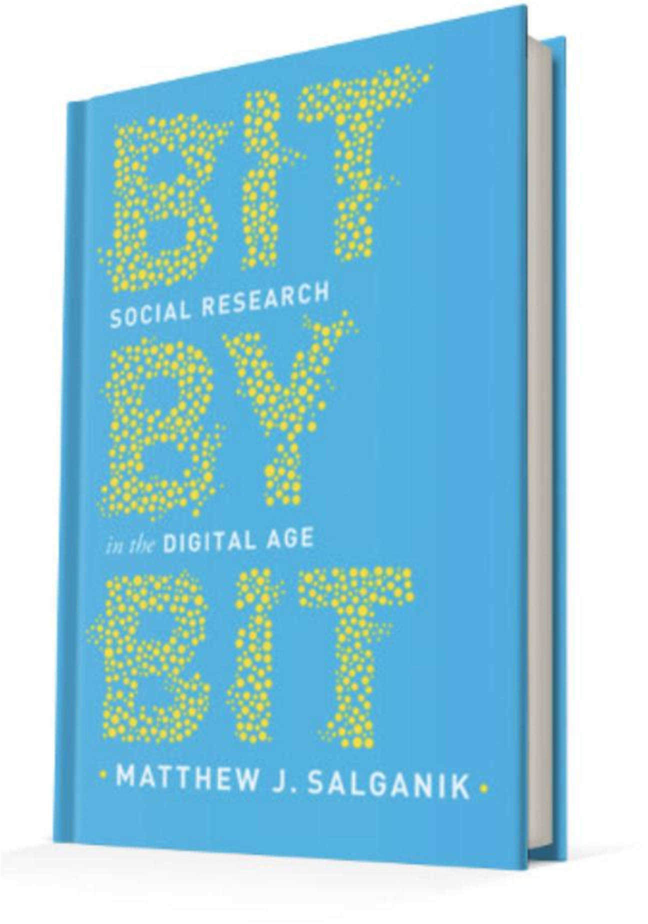


2.2 Big data

In the analog age, most of the data that were used for social research was created for the purpose of doing research. In the digital age, however, a huge amount of **data is being created by companies and governments for purposes other than research**, such as providing services, generating profit, and administering laws. Creative people, however, have realized that you can **repurpose** this corporate and government data for research.

https://www.bitbybitbook.com/en/1st-ed/observing-behavior/data/

Understand your data!



2.2 Big data

… from the perspective of researchers, big data sources are “found,” they don’t just fall from the sky. Instead, data sources that are “found” by researchers are **designed by someone for some purpose**. Because “found” data are designed by someone, I always recommend that you **try to understand as much as possible about the people and processes that created your data**.

https://www.bitbybitbook.com/en/1st-ed/observing-behavior/data/

r/ai center for responsible ai

Understand your data!

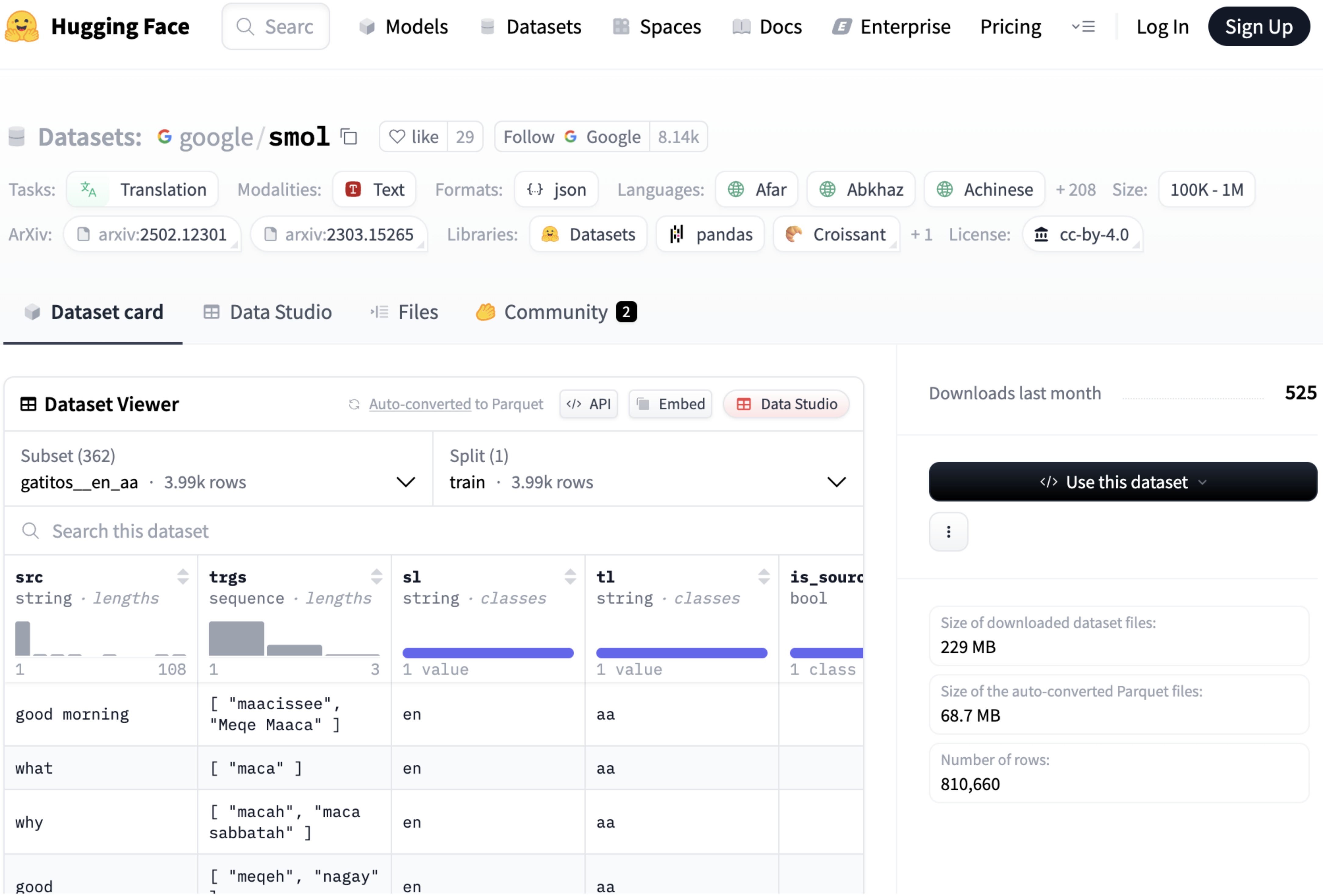
Need **metadata** to:

- enable data **re-use** (have to be able to find it!)
- determine **fitness for use** of a dataset in a task
- help establish **trust** in the data analysis process and its outcomes

Data is considered to be of high quality if it's "**fit for intended uses** in operations, decision making and planning"

[Thomas C. Redman, "Data Driven: Profiting from Your Most Important Business Asset." 2013]

Common metadata example



Datasets: google/smol like 29 Follow Google 8.14k

Tasks: Translation Modalities: Text Formats: json Languages: Afar Abkhaz Achinese + 208 Size: 100K - 1M

ArXiv: arxiv:2502.12301 arxiv:2303.15265 Libraries: Datasets pandas Croissant + 1 License: cc-by-4.0

Dataset Viewer Auto-converted to Parquet

Subset (362)
gatitos__en_aa · 3.99k rows

Split (1)
train · 3.99k rows

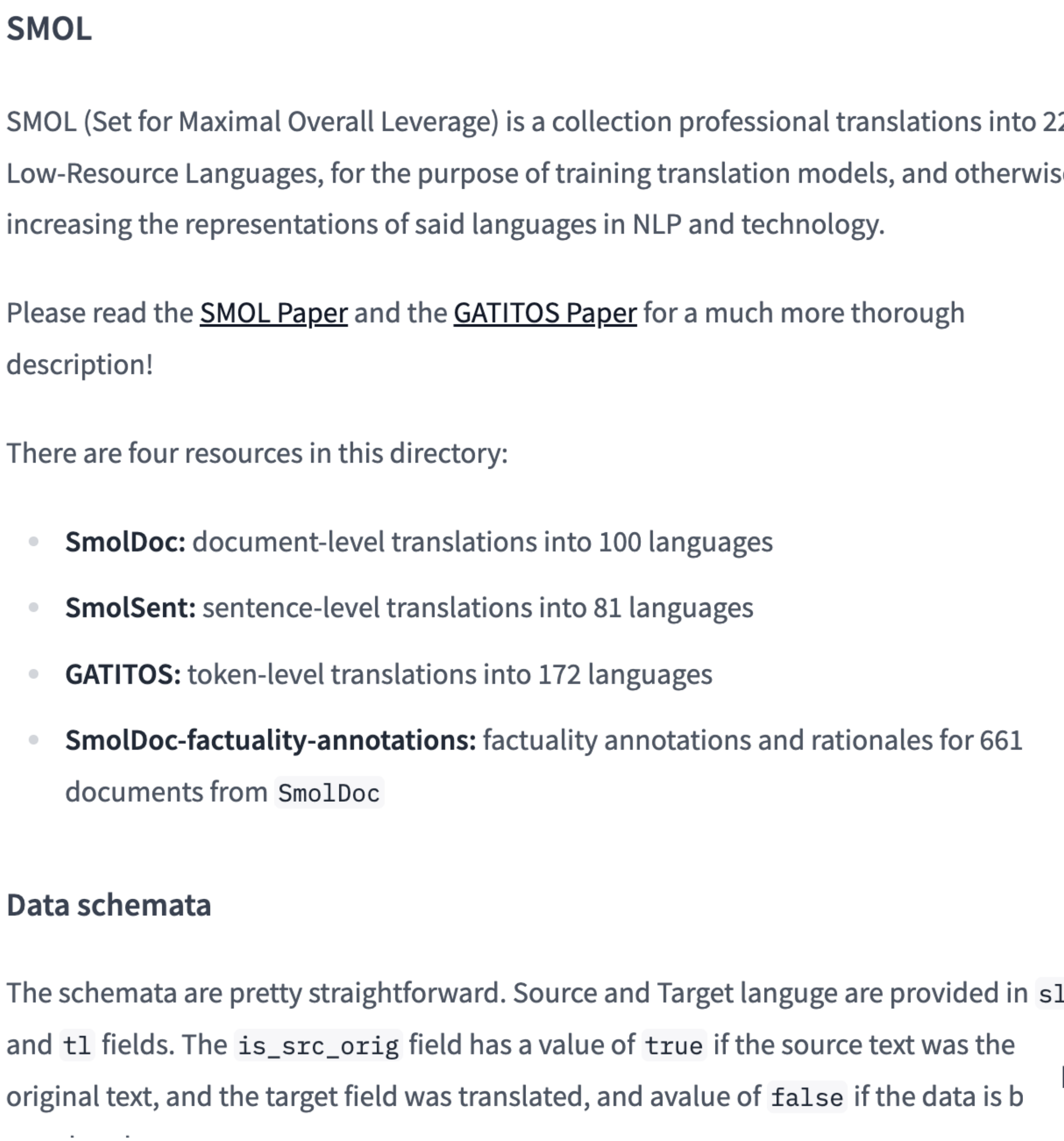
src string · *lengths* 1 108	trgs sequence · *lengths* 1 3	sl string · *classes* 1 value	tl string · *classes* 1 value	is_sourc bool 1 class
good morning	["maacissee", "Meqe Maaca"]	en	aa	
what	["maca"]	en	aa	
why	["macah", "maca sabbatah"]	en	aa	
good	["meqeh", "nagay"]	en	aa	

Downloads last month 525

Size of downloaded dataset files:
229 MB

Size of the auto-converted Parquet files:
68.7 MB

Number of rows:
810,660



SMOL

SMOL (Set for Maximal Overall Leverage) is a collection professional translations into 221 Low-Resource Languages, for the purpose of training translation models, and otherwise increasing the representations of said languages in NLP and technology.

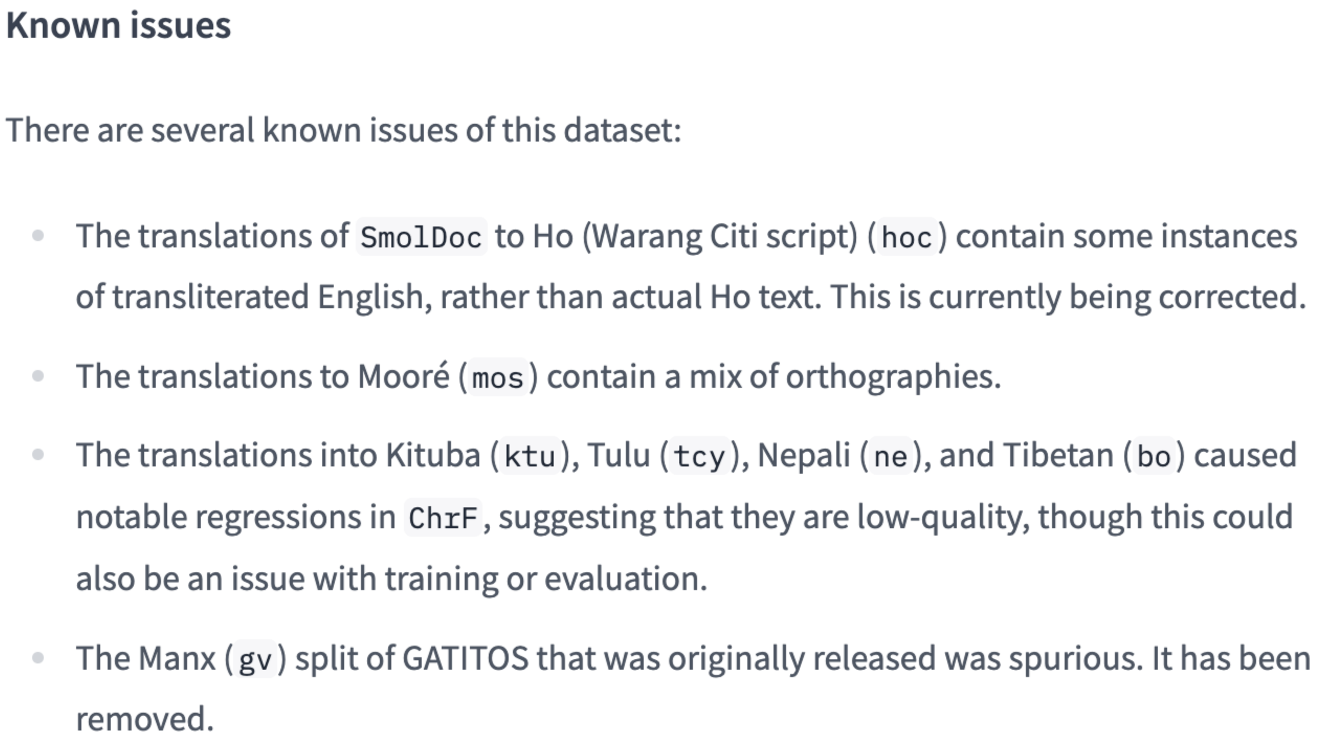
Please read the SMOL Paper and the GATITOS Paper for a much more thorough description!

There are four resources in this directory:

- **SmolDoc:** document-level translations into 100 languages
- **SmolSent:** sentence-level translations into 81 languages
- **GATITOS:** token-level translations into 172 languages
- **SmolDoc-factuality-annotations:** factuality annotations and rationales for 661 documents from `SmolDoc`

Data schemata

The schemata are pretty straightforward. Source and Target languge are provided in `sl` and `tl` fields. The `is_src_orig` field has a value of `true` if the source text was the original text, and the target field was translated, and avalue of `false` if the data is b

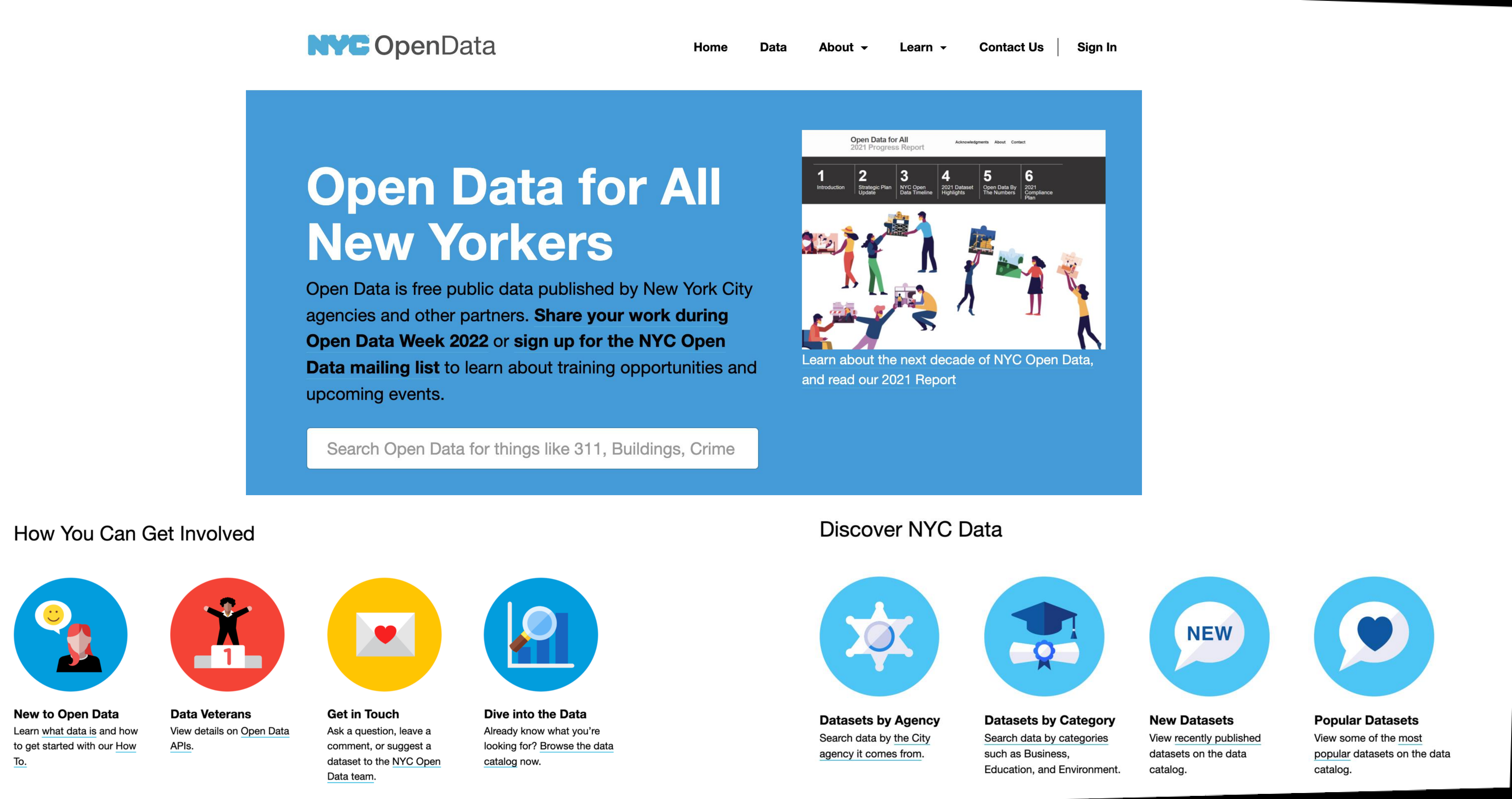


Known issues

There are several known issues of this dataset:

- The translations of `SmolDoc` to Ho (Warang Citi script) (`hoc`) contain some instances of transliterated English, rather than actual Ho text. This is currently being corrected.
- The translations to Mooré (`mos`) contain a mix of orthographies.
- The translations into Kituba (`ktu`), Tulu (`tcy`), Nepali (`ne`), and Tibetan (`bo`) caused notable regressions in `ChrF`, suggesting that they are low-quality, though this could also be an issue with training or evaluation.
- The Manx (`gv`) split of GATITOS that was originally released was spurious. It has been removed.

NYC Open Data



Open Data for All New Yorkers

Open Data is free public data published by New York City agencies and other partners. **Share your work during Open Data Week 2022** or **sign up for the NYC Open Data mailing list** to learn about training opportunities and upcoming events.

Search Open Data for things like 311, Buildings, Crime

Learn about the next decade of NYC Open Data, and read our 2021 Report

How You Can Get Involved

New to Open Data
Learn what data is and how to get started with our How To.

Data Veterans
View details on Open Data APIs.

Get in Touch
Ask a question, leave a comment, or suggest a dataset to the NYC Open Data team.

Dive into the Data
Already know what you're looking for? Browse the data catalog now.

Discover NYC Data

Datasets by Agency
Search data by the City agency it comes from.

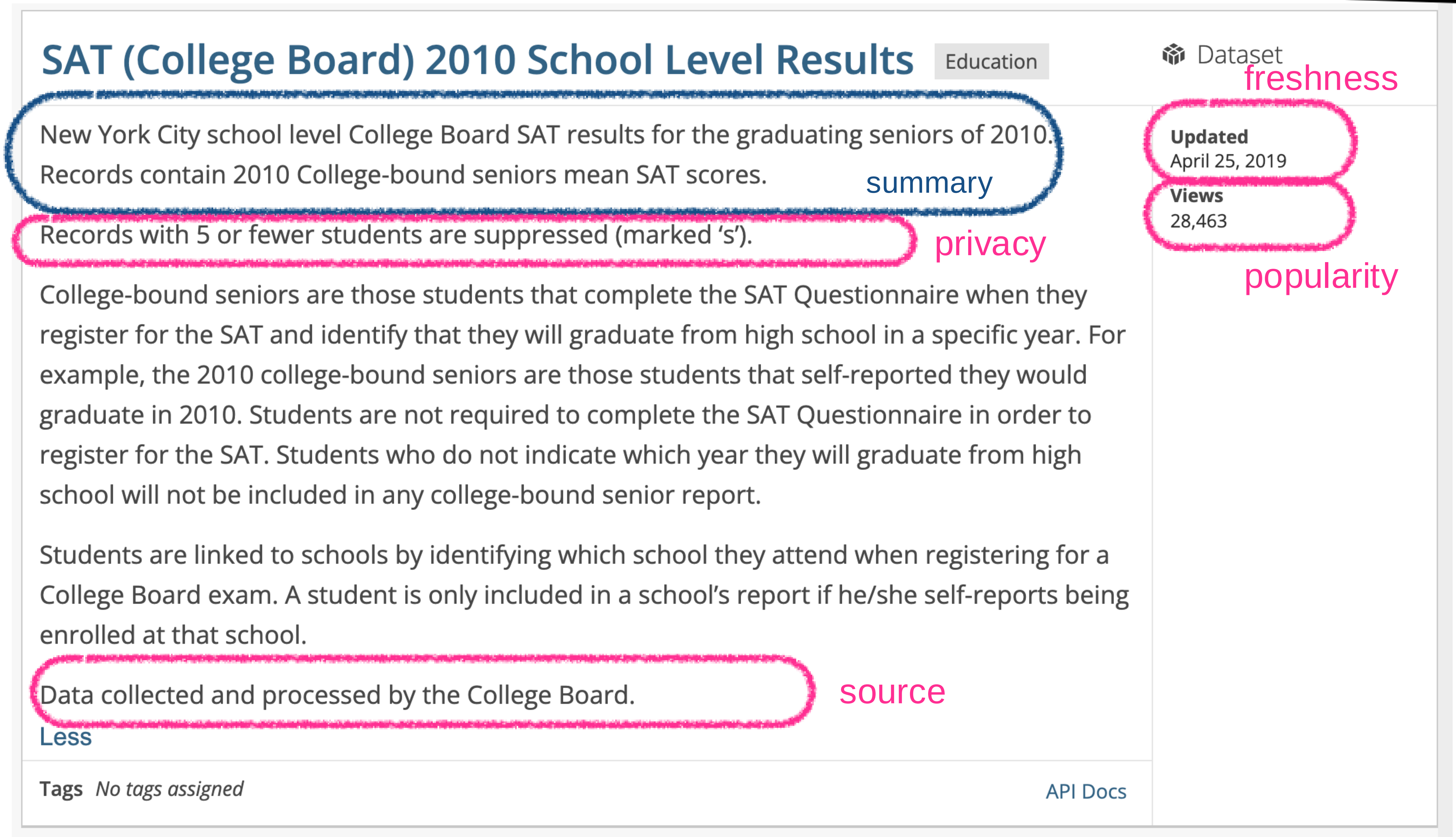
Datasets by Category
Search data by categories such as Business, Education, and Environment.

New Datasets
View recently published datasets on the data catalog.

Popular Datasets
View some of the most popular datasets on the data catalog.

https://opendata.cityofnewyork.us/

NYC Open Data



SAT (College Board) 2010 School Level Results

Education

Dataset

New York City school level College Board SAT results for the graduating seniors of 2010. Records contain 2010 College-bound seniors mean SAT scores.

summary

Records with 5 or fewer students are suppressed (marked ‘s’).

privacy

College-bound seniors are those students that complete the SAT Questionnaire when they register for the SAT and identify that they will graduate from high school in a specific year. For example, the 2010 college-bound seniors are those students that self-reported they would graduate in 2010. Students are not required to complete the SAT Questionnaire in order to register for the SAT. Students who do not indicate which year they will graduate from high school will not be included in any college-bound senior report.

Students are linked to schools by identifying which school they attend when registering for a College Board exam. A student is only included in a school’s report if he/she self-reports being enrolled at that school.

Data collected and processed by the College Board.

source

Less

Tags *No tags assigned*

API Docs

freshness

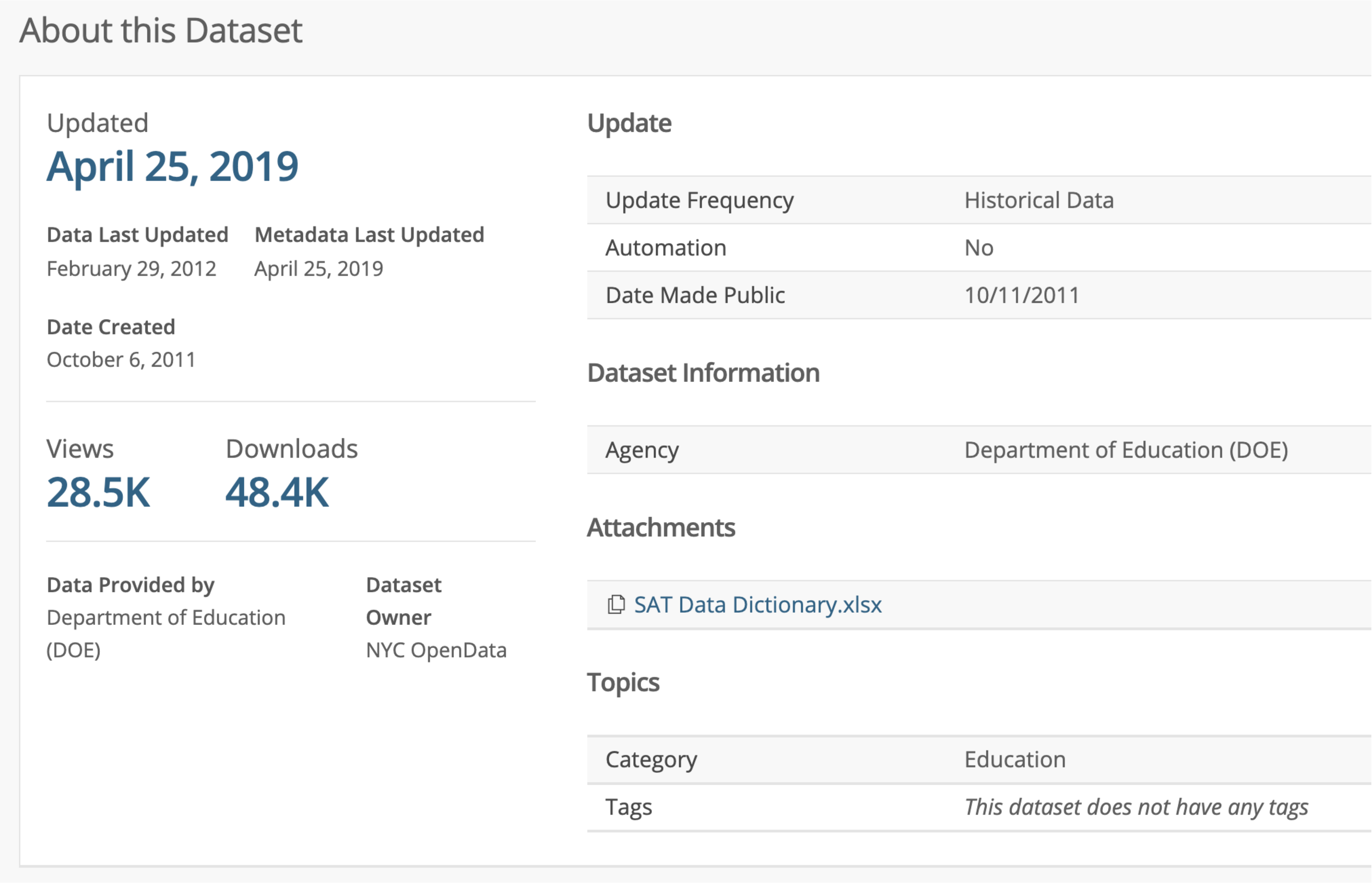
Updated
April 25, 2019

Views
28,463

popularity

https://opendata.cityofnewyork.us/

NYC Open Data



About this Dataset

Updated

April 25, 2019

Data Last Updated	Metadata Last Updated
February 29, 2012	April 25, 2019

Date Created

October 6, 2011

Views	Downloads
28.5K	48.4K

Data Provided by	Dataset Owner
Department of Education (DOE)	NYC OpenData

Update

Update Frequency	Historical Data
Automation	No
Date Made Public	10/11/2011

Dataset Information

Agency	Department of Education (DOE)

Attachments

SAT Data Dictionary.xlsx

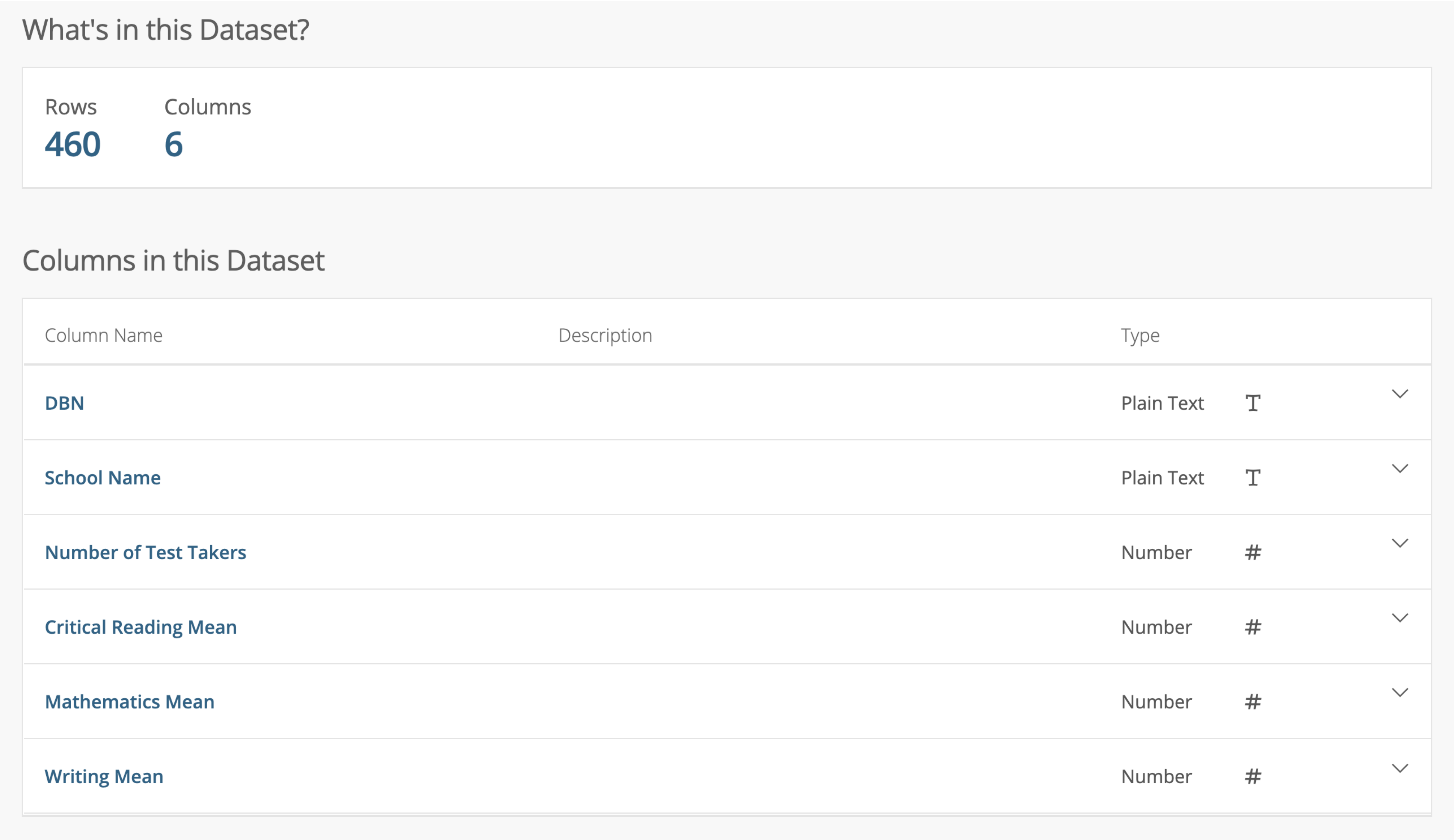
Topics

Category	Education
Tags	*This dataset does not have any tags*

https://opendata.cityofnewyork.us/

r/ai center for responsible ai

NYC Open Data



What's in this Dataset?

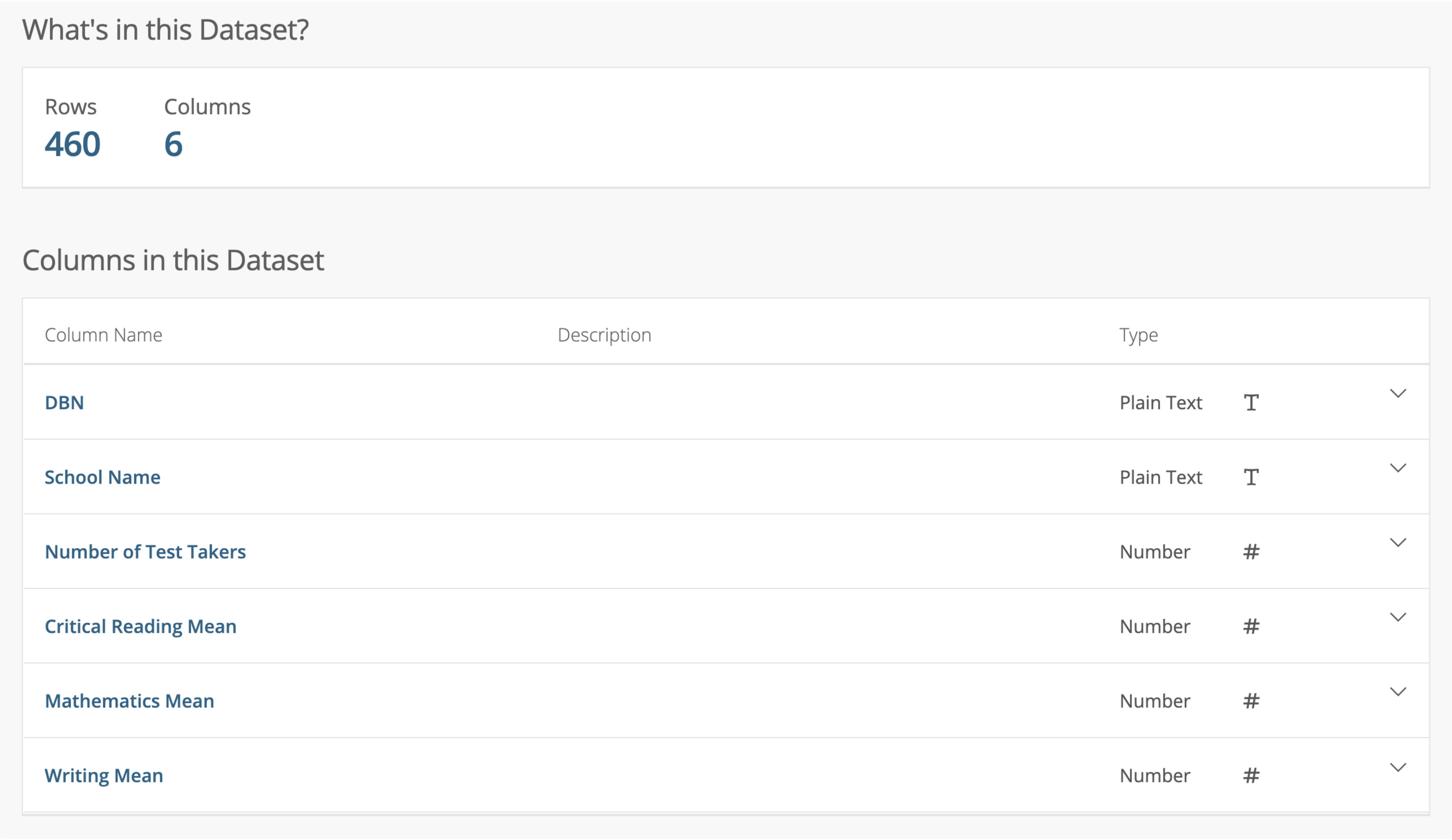
Rows	Columns
460	6

Columns in this Dataset

Column Name	Description	Type
DBN		Plain Text T
School Name		Plain Text T
Number of Test Takers		Number #
Critical Reading Mean		Number #
Mathematics Mean		Number #
Writing Mean		Number #

https://opendata.cityofnewyork.us/

NYC Open Data



What's in this Dataset?

Rows	Columns
460	6

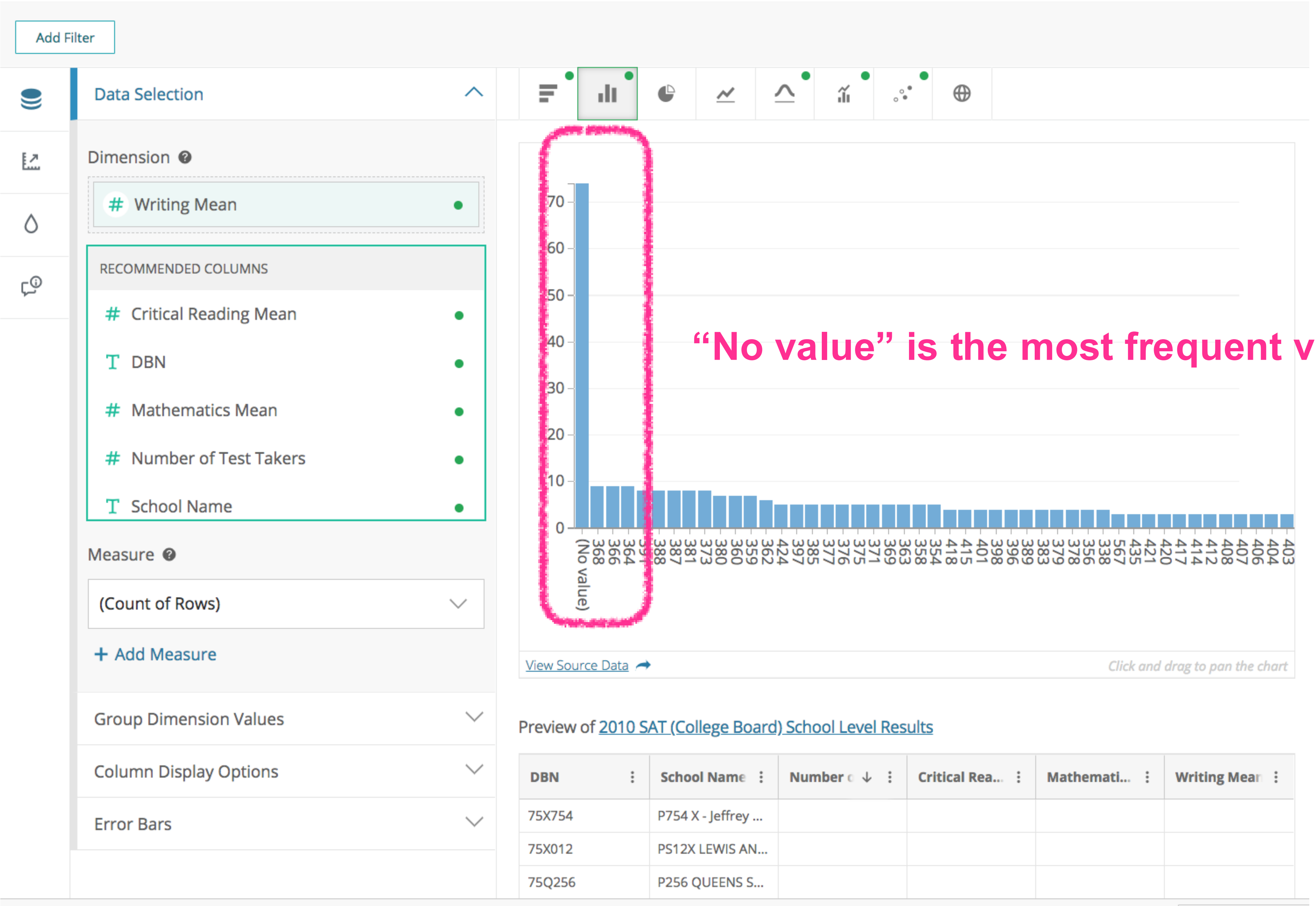
Columns in this Dataset

Column Name	Description	Type
DBN		Plain Text T
School Name		Plain Text T
Number of Test Takers		Number #
Critical Reading Mean		Number #
Mathematics Mean		Number #
Writing Mean		Number #

https://opendata.cityofnewyork.us/

r/ai center for responsible ai

NYC Open Data



Add Filter

Data Selection

Dimension

# Writing Mean

RECOMMENDED COLUMNS

- # Critical Reading Mean
- T DBN
- # Mathematics Mean
- # Number of Test Takers
- T School Name

Measure

(Count of Rows)

+ Add Measure

Group Dimension Values

Column Display Options

Error Bars

View Source Data

Click and drag to pan the chart

Preview of 2010 SAT (College Board) School Level Results

DBN	School Name	Number o	Critical Rea...	Mathemati...	Writing Mean
75X754	P754 X - Jeffrey ...				
75X012	PS12X LEWIS AN...				
75Q256	P256 QUEENS S...				

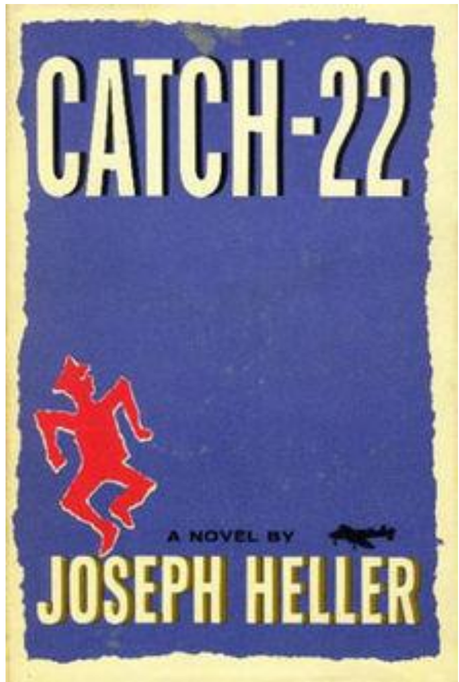
“No value” is the most frequent value

https://opendata.cityofnewyork.us/

r/ai center for responsible ai

Data profiling

- **Data profiling** refers to the activity of creating **small** but **informative** summaries of a database
- What is informative depends on the task, or set of tasks, we have in mind



should profiling be task-agnostic or task-specific?

A related activity is **data cleaning**

Data cleaning

Data cleansing or **data cleaning** is the process of detecting and repairing corrupt or inaccurate records from a data set in order to improve the **quality of data**.

Erhard Rahm, Hong Hai Do: Data Cleaning: Problems and Current Approaches, IEEE Data Engineering Bulletin, 2000.

… **data** is generally considered high **quality** if it is "**fit for [its] intended uses** in operations, decision making and planning“

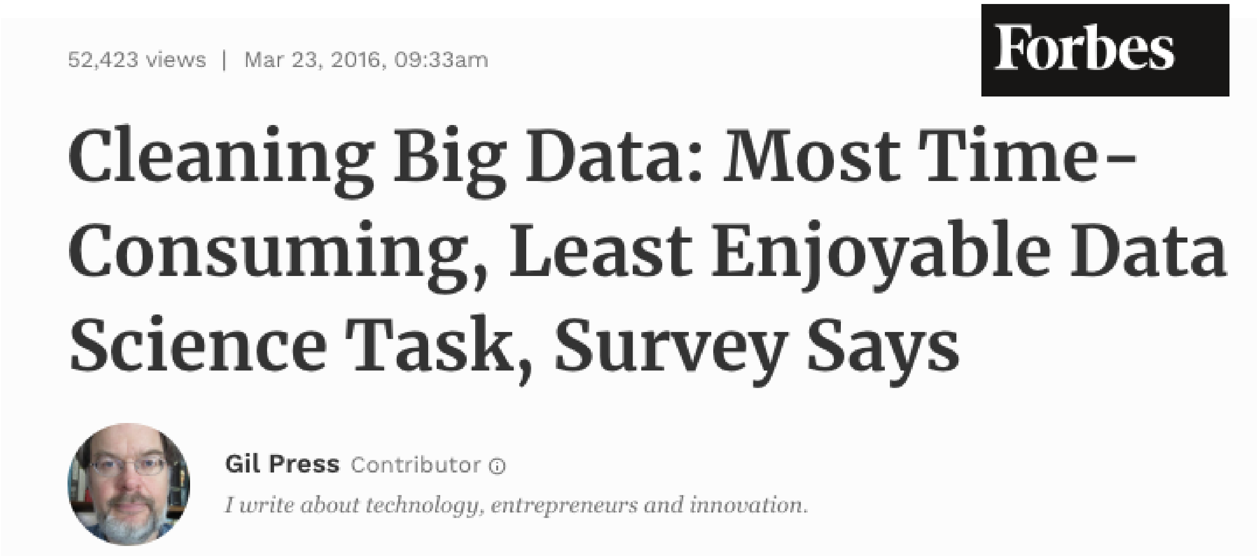
Thomas C. Redman, Data Driven: Profiting from Your Most Important Business Asset. 2013

Even though quality cannot be defined, you know what it is.

Robert M. Prisig, Zen and the Art of Motorcycle Maintenance, 1975

slide by Heiko Mueller

Data cleaning



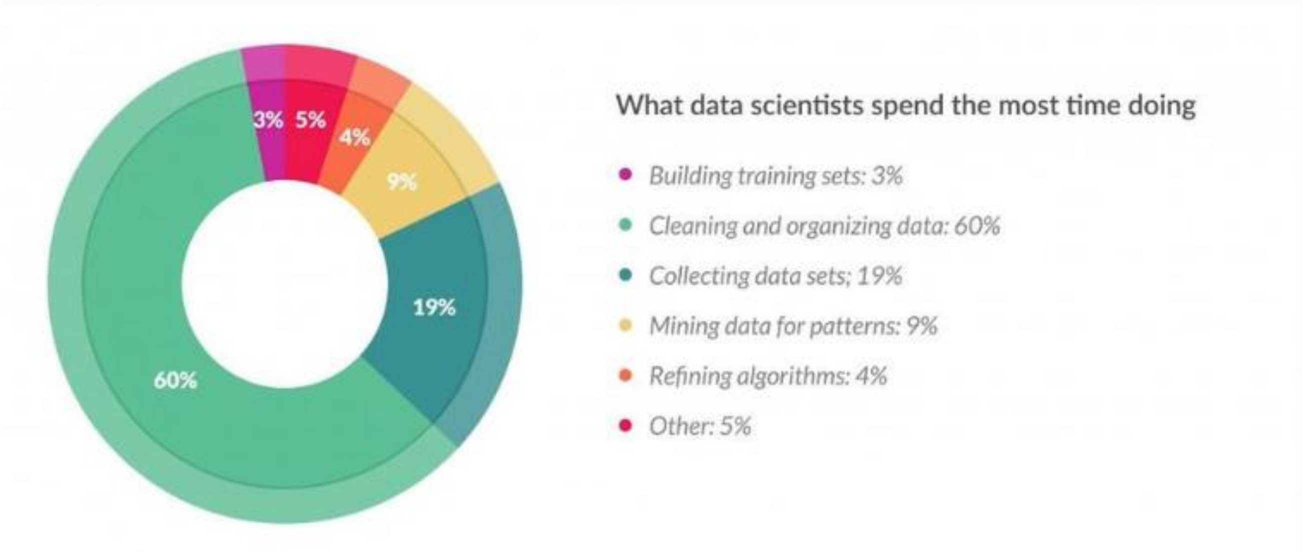
52,423 views | Mar 23, 2016, 09:33am

Forbes

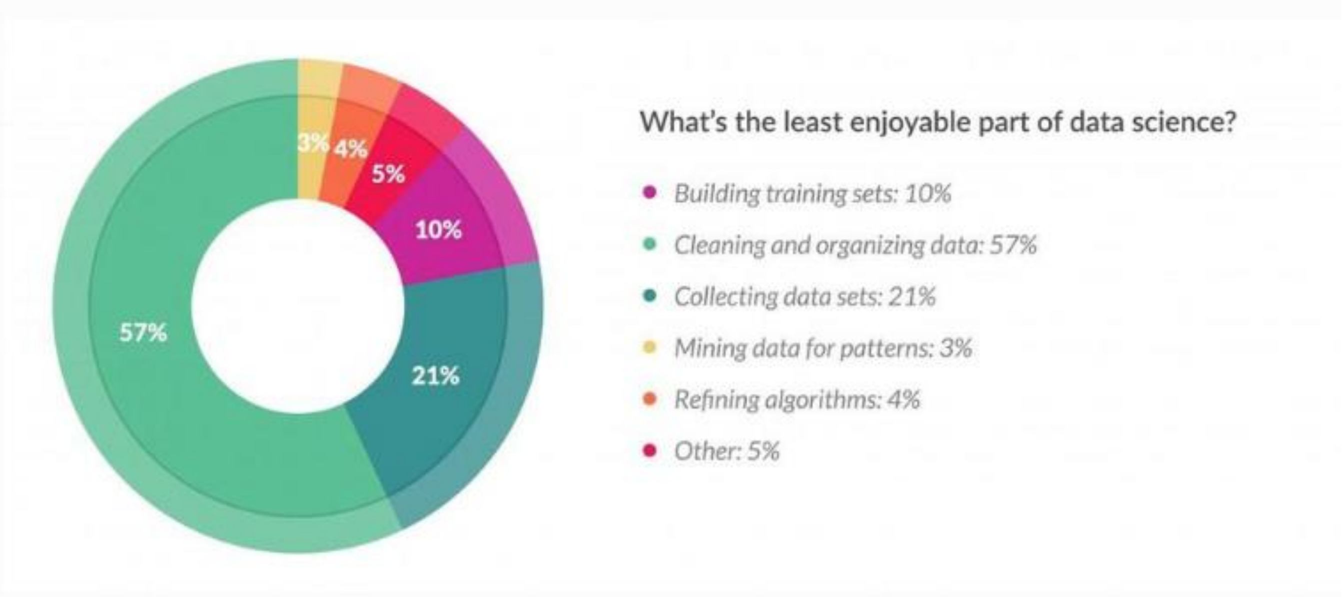
Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press Contributor

I write about technology, entrepreneurs and innovation.



What data scientists spend the most time doing



What's the least enjoyable part of data science?

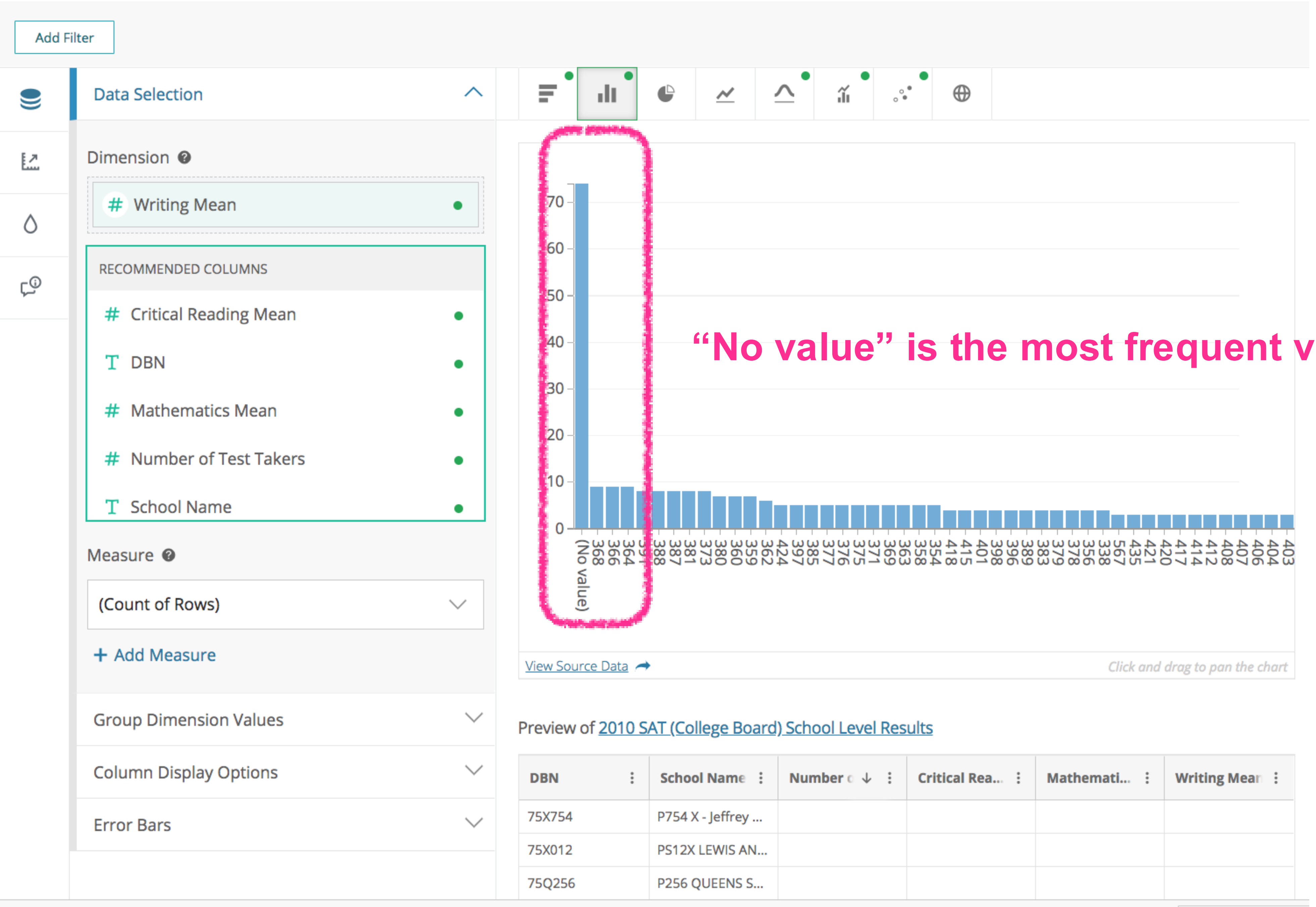
Spend most time doing
Collecting data (19%)
Cleaning and organizing data (60%)

Find least enjoyable
Collecting data (21%)
Cleaning and organizing data (57%)

slide by Heiko Mueller

data profiling

NYC Open Data



Dimension: # Writing Mean ●

- # Critical Reading Mean ●
- T DBN ●
- # Mathematics Mean ●
- # Number of Test Takers ●
- T School Name ●

Measure: (Count of Rows)

+ Add Measure

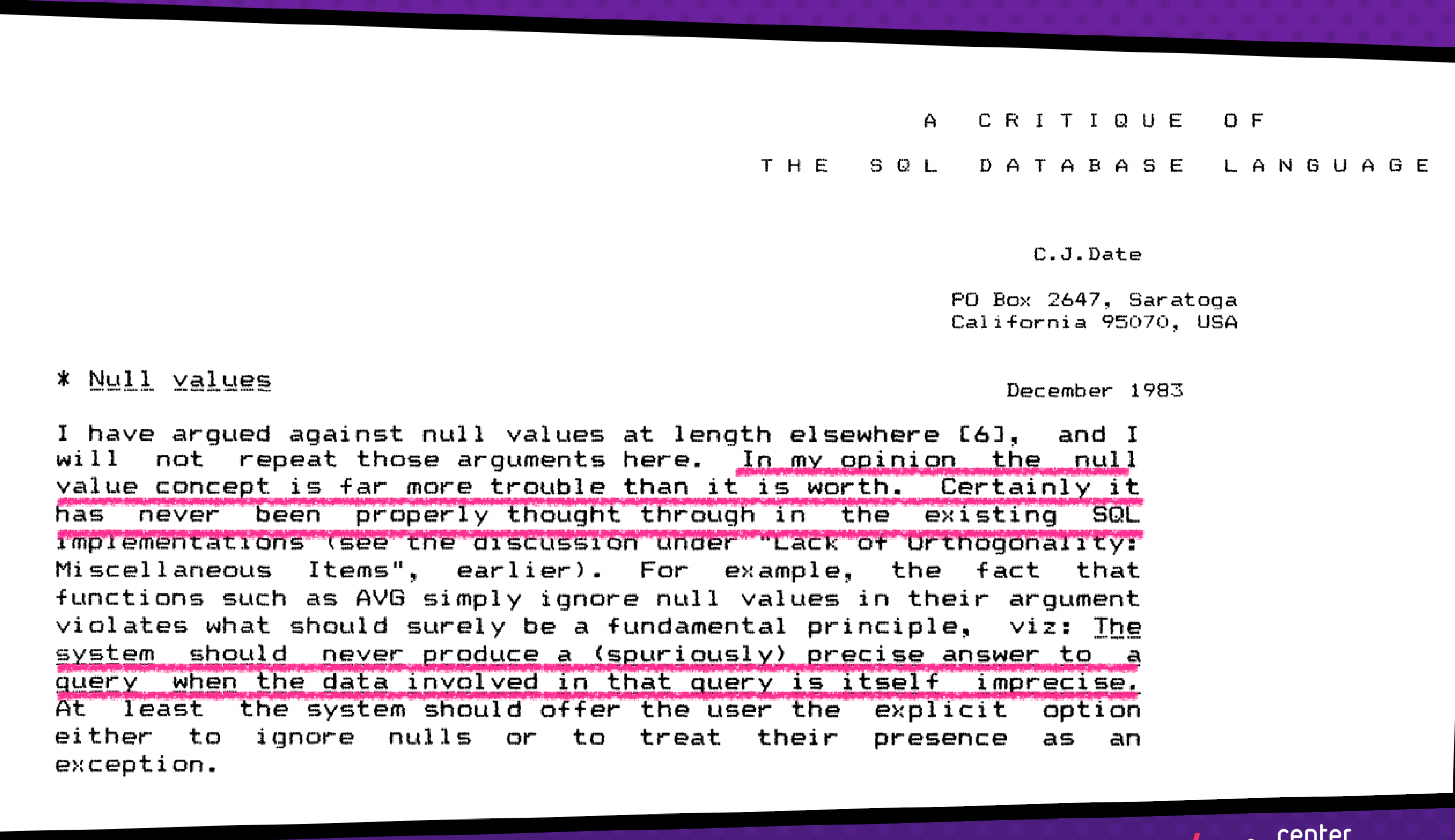
“No value” is the most frequent value

Preview of 2010 SAT (College Board) School Level Results

DBN	School Name	Number o ↓	Critical Rea...	Mathemati...	Writing Mear
75X754	P754 X - Jeffrey ...				
75X012	PS12X LEWIS AN...				
75Q256	P256 QUEENS S...				

https://opendata.cityofnewyork.us/

The trouble with *null* values



A CRITIQUE OF

THE SQL DATABASE LANGUAGE

C.J.Date

PO Box 2647, Saratoga
California 95070, USA

December 1983

* Null values

I have argued against null values at length elsewhere [6], and I will not repeat those arguments here. In my opinion the null value concept is far more trouble than it is worth. Certainly it has never been properly thought through in the existing SQL implementations (see the discussion under "Lack of Orthogonality: Miscellaneous Items", earlier). For example, the fact that functions such as AVG simply ignore null values in their argument violates what should surely be a fundamental principle, viz: The system should never produce a (spuriously) precise answer to a query when the data involved in that query is itself imprecise. At least the system should offer the user the explicit option either to ignore nulls or to treat their presence as an exception.

50 shades of *null*

- **Unknown** - some value definitely belongs here, but I don’t know what it is (e.g., unknown birthdate)
- **Inapplicable** - no value makes sense here (e.g., if marital status = single then spouse name should not have a value)
- **Unintentionally omitted** - values is left unspecified unintentionally, by mistake
- **Optional** - a value may legitimately be left unspecified (e.g., middle name)
- **Intentionally withheld** (e.g., an unlisted phone number)
- …..

(this selection is mine, see reference below for a slightly different list)

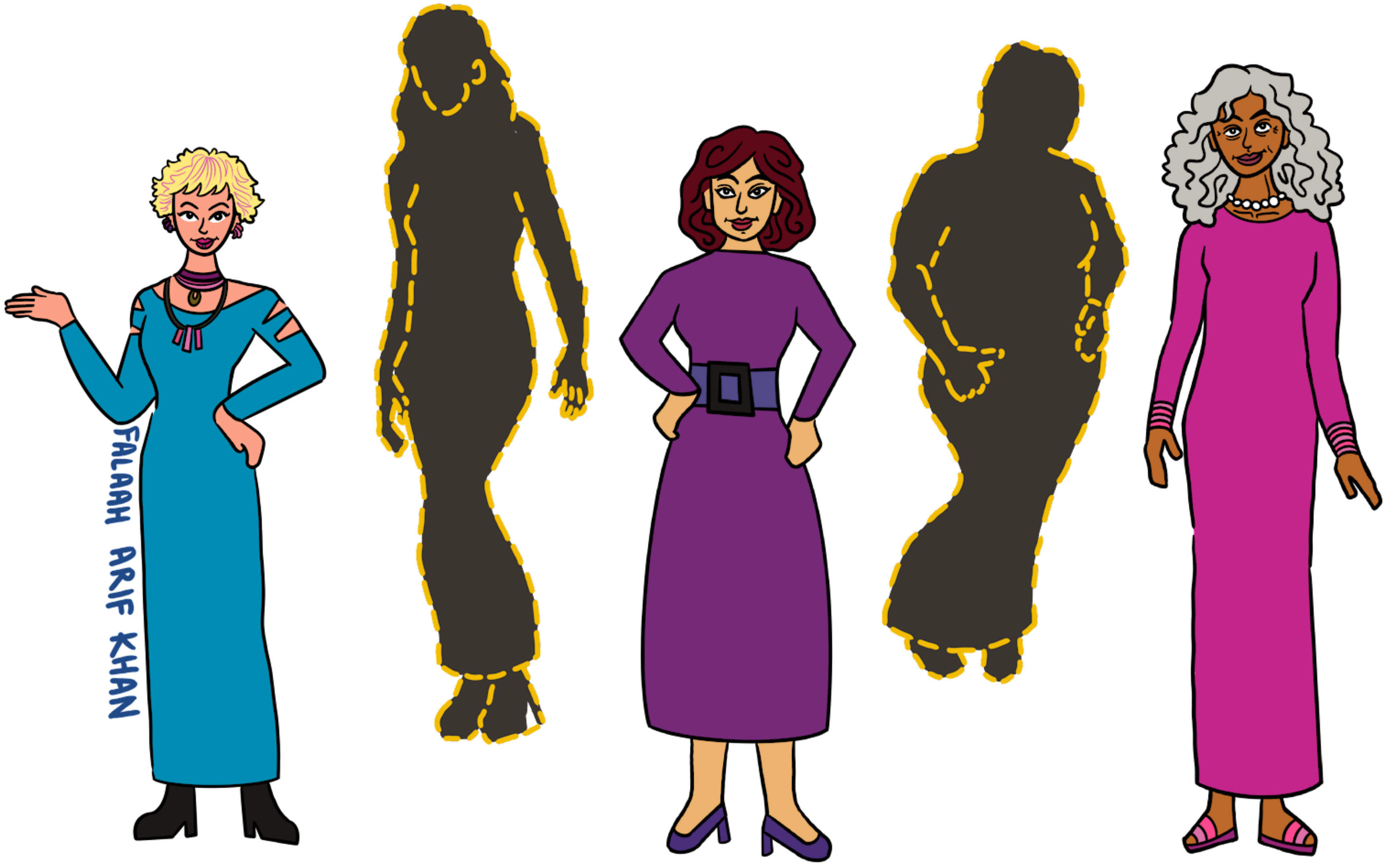
https://www.vertabelo.com/blog/technical-articles/50-shades-of-null-or-how-a-billion-dollar-mistake-has-been-stalking-a-whole-industry-for-decades

50 shades of *null*… and it gets worse

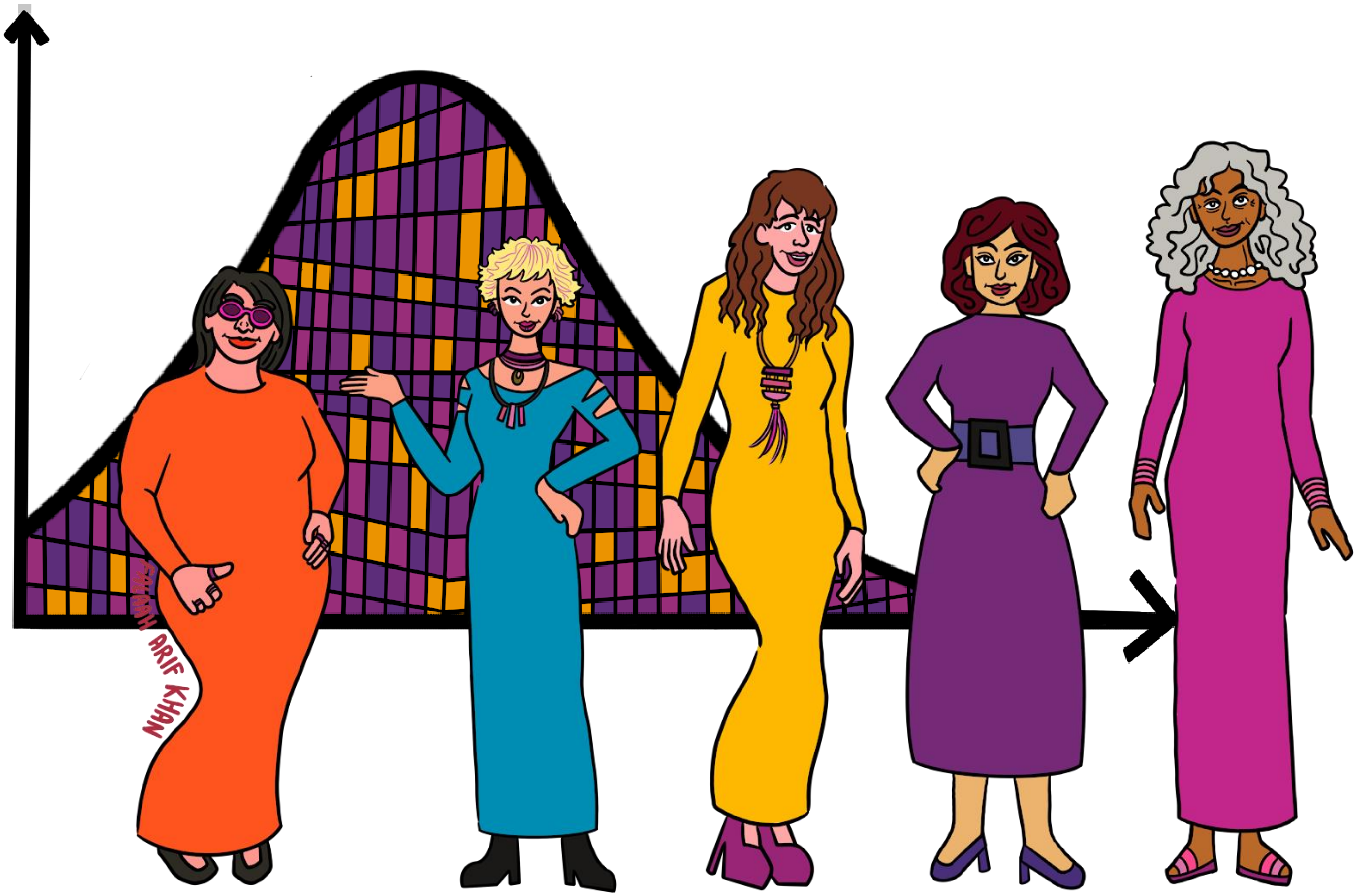
- **Hidden missing values** -
  - 99999 for zip code, Alabama for state
  - need data cleaning….
- lots of houses in Philadelphia, PA were built in 1934 (or 1936?) - not really!

how do we detect hidden missing values?

Missing values: Observed data



Missing values: Imputed distribution



Missing values: True distribution

FALAAH ARIF KHAN

r/ai center for responsible ai

Missing value imputation

are values **missing at random** (e.g., gender, age, disability on job applications)?

are we ever interpolating **rare categories** (e.g., Native American)

are **all categories** represented (e.g., non-binary gender)?

Data filtering

“filtering” operations (like selection and join), **can arbitrarily change demographic roup proportions**

select by zip code, country, years of C++ experience, others?

age_group	county
60	CountyA
60	CountyA
20	CountyA
60	CountyB
20	CountyB
20	CountyB

50% vs 50%

→

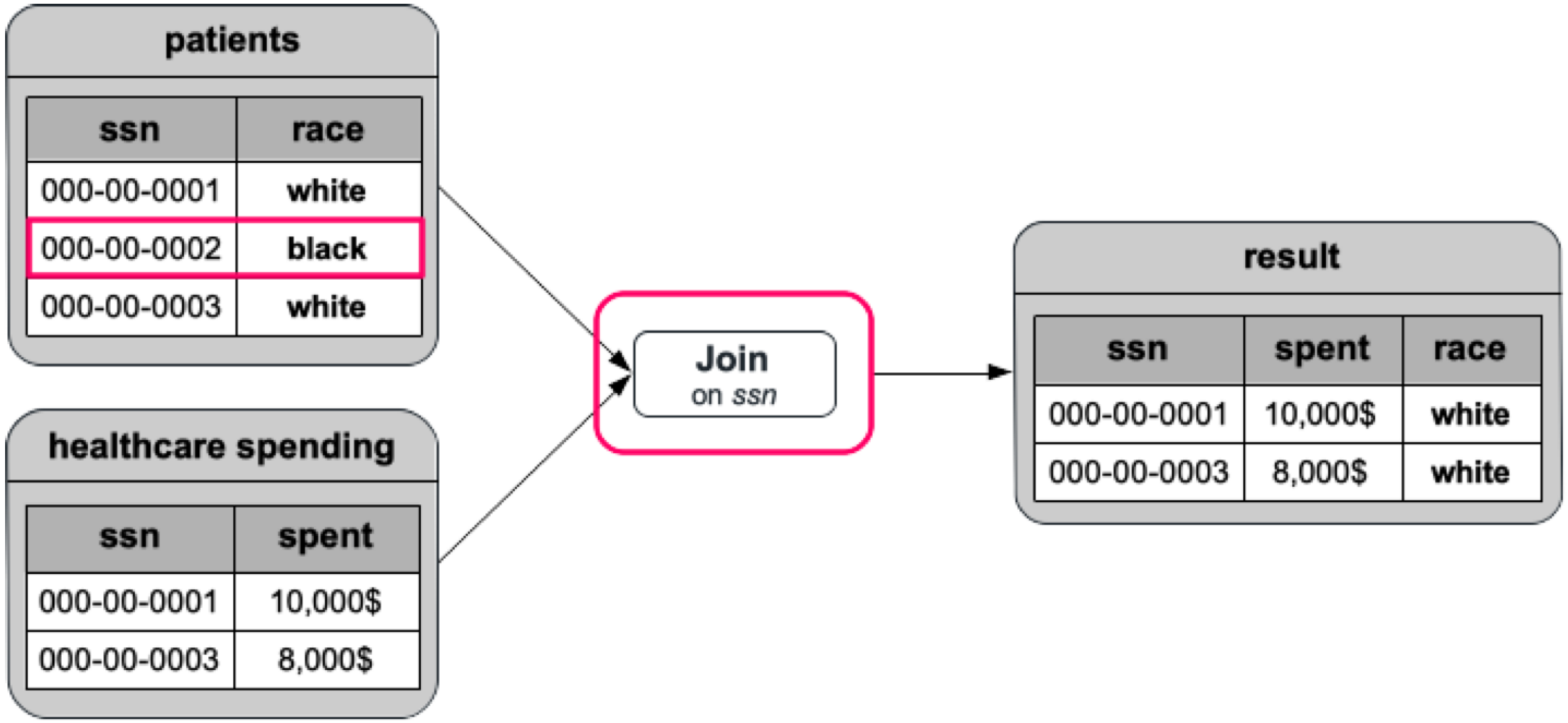
age_group	county
60	CountyA
60	CountyA
20	CountyA

66% vs 33%

Data filtering

“filtering” operations (like selection and join), **can arbitrarily change demographic**
roup proportions

select by zip code, country, years of C++ experience, others?



patients

ssn	race
000-00-0001	white
000-00-0002	black
000-00-0003	white

healthcare spending

ssn	spent
000-00-0001	10,000$
000-00-0003	8,000$

result

ssn	spent	race
000-00-0001	10,000$	white
000-00-0003	8,000$	white

Data distribution debugging: mlinspect

Potential issues in preprocessing pipeline:

1 Join might change proportions of groups in data

2 Column ‘age_group’ projected out, but required for fairness

3 Selection might change proportions of groups in data

4 Imputation might change proportions of groups in data

5 ‘race’ as a feature might be illegal!

6 Embedding vectors may not be available for rare names!

Python script for preprocessing, written exclusively with native pandas and sklearn constructs

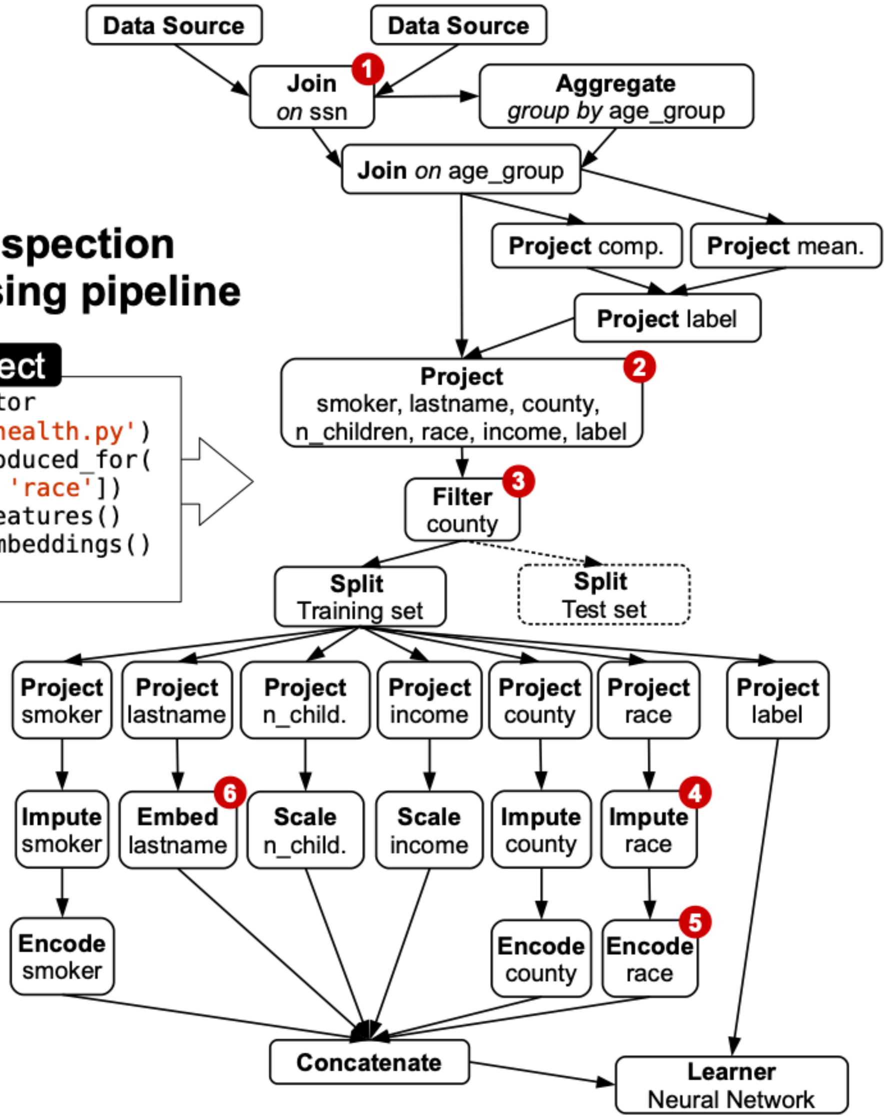
```
# load input data sources, join to single table
patients = pandas.read_csv(…)
histories = pandas.read_csv(…)
data = pandas.merge([patients, histories], on=['ssn'])

# compute mean complications per age group, append as column
complications = data.groupby('age_group')
 .agg(mean_complications=('complications','mean'))
data = data.merge(complications, on=['age_group'])

# Target variable: people with frequent complications
data['label'] = data['complications'] >
  1.2 * data['mean_complications']

# Project data to subset of attributes, filter by counties
data = data[['smoker', 'last_name', 'county',
             'num_children', 'race', 'income', 'label']]
data = data[data['county'].isin(counties_of_interest)]

# Define a nested feature encoding pipeline for the data
impute_and_encode = sklearn.Pipeline([
  (sklearn.SimpleImputer(strategy='most_frequent')),
  (sklearn.OneHotEncoder())])
featurisation = sklearn.ColumnTransformer(transformers=[
  (impute_and_encode, ['smoker', 'county', 'race']),
  (Word2VecTransformer(), 'last_name')
  (sklearn.StandardScaler(), ['num_children', 'income']])

# Define the training pipeline for the model
neural_net = sklearn.KerasClassifier(build_fn=create_model())
pipeline = sklearn.Pipeline([
  ('features', featurisation),
  ('learning_algorithm', neural_net)])

# Train-test split, model training and evaluation
train_data, test_data = train_test_split(data)
model = pipeline.fit(train_data, train_data.label)
print(model.score(test_data, test_data.label))
```

Declarative inspection of preprocessing pipeline

mlinspect

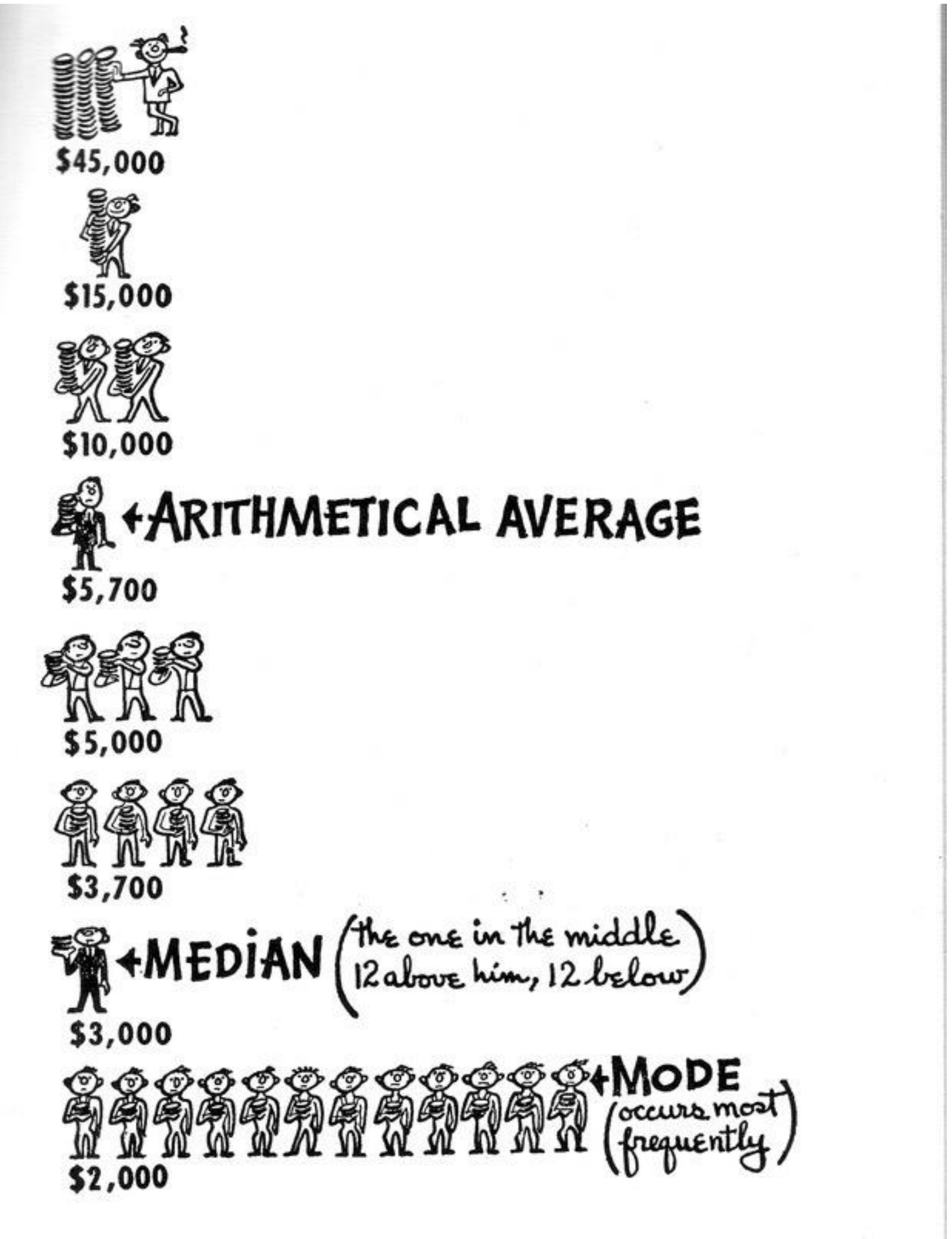
```
PipelineInspector
.on_pipeline('health.py')
 .no_bias_introduced_for(
 ['age_group', 'race'])
 .no_illegal_features()
 .no_missing_embeddings()
.verify()
```

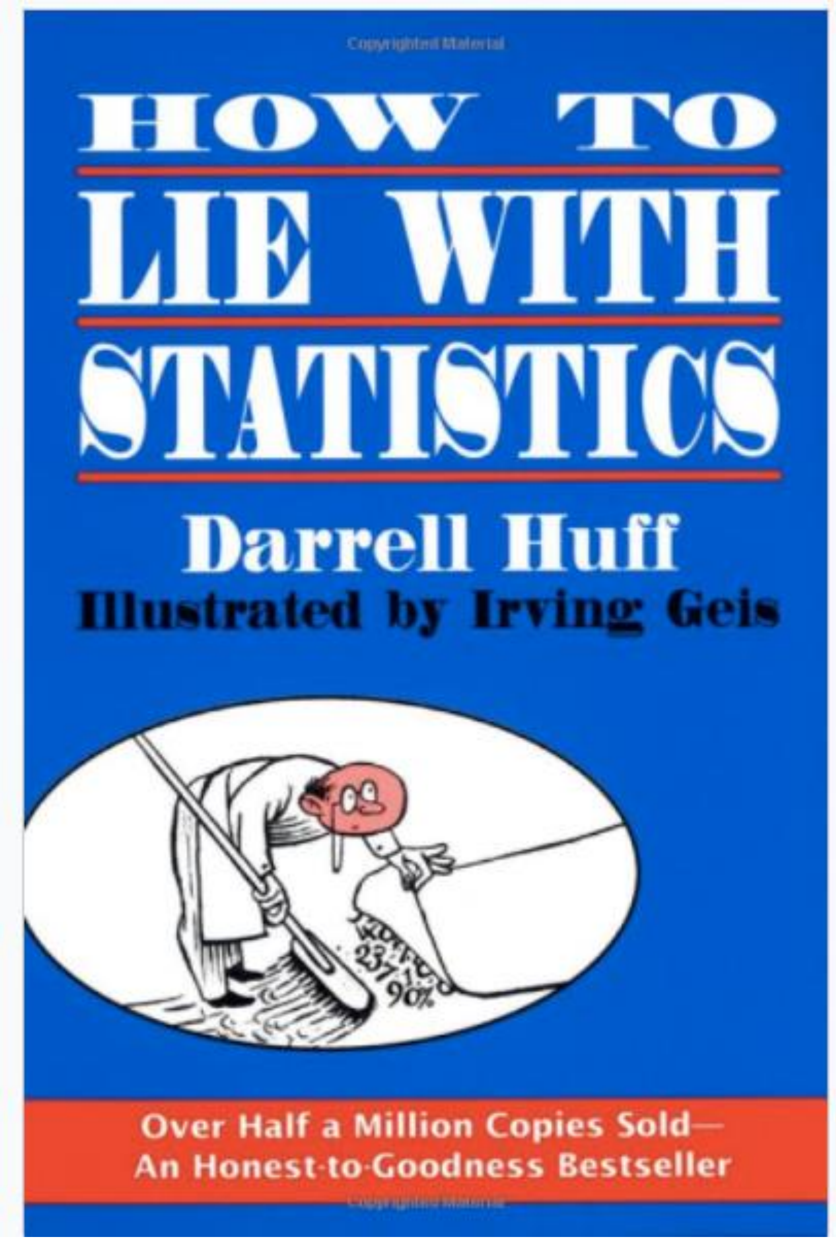
Corresponding dataflow DAG for instrumentation, extracted by *mlinspect*



[Grafberger, Stoyanovich, Schelter (2022)]

The well-chosen average

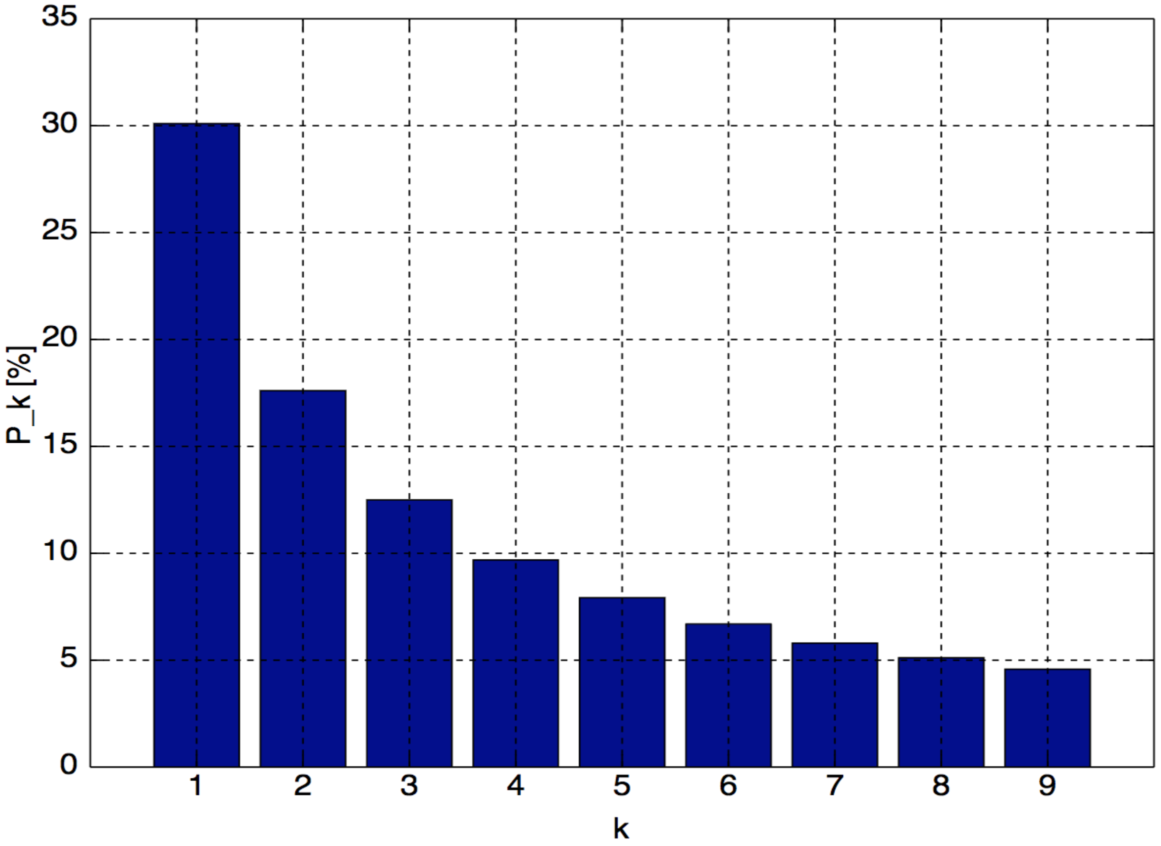

$45,000
$15,000
$10,000
←ARITHMETICAL AVERAGE
$5,700
$5,000
$3,700
←MEDIAN (the one in the middle 12 above him, 12 below)
$3,000
←MODE (occurs most frequently)
$2,000


HOW TO LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

Benford Law

The distribution of **the first digit d** of a number, in many naturally occurring domains, approximately follows

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$



1 is the most frequent leading digit, followed by 2, etc.

https://en.wikipedia.org/wiki/Benford%27s_law

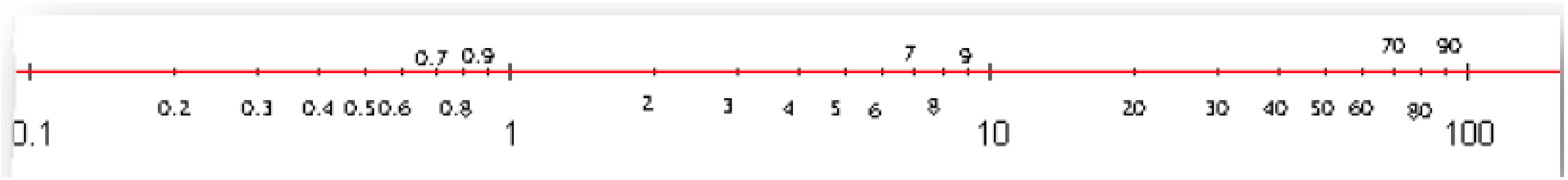
[Benford: “The law of anomalous numbers” *Proc. Am. Philos. Soc.*, 1938]

Benford Law

The distribution of **the first digit d** of a number, in many naturally occurring domains, approximately follows

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

Holds if **log(x) is uniformly distributed**. **Most accurate** when values are distributed across multiple orders of magnitude, especially **if the process generating the numbers is described by a power law** (common in nature)



A logarithmic scale bar. Picking a random x position uniformly on this number line, roughly 30% of the time the first digit of the number will be 1.

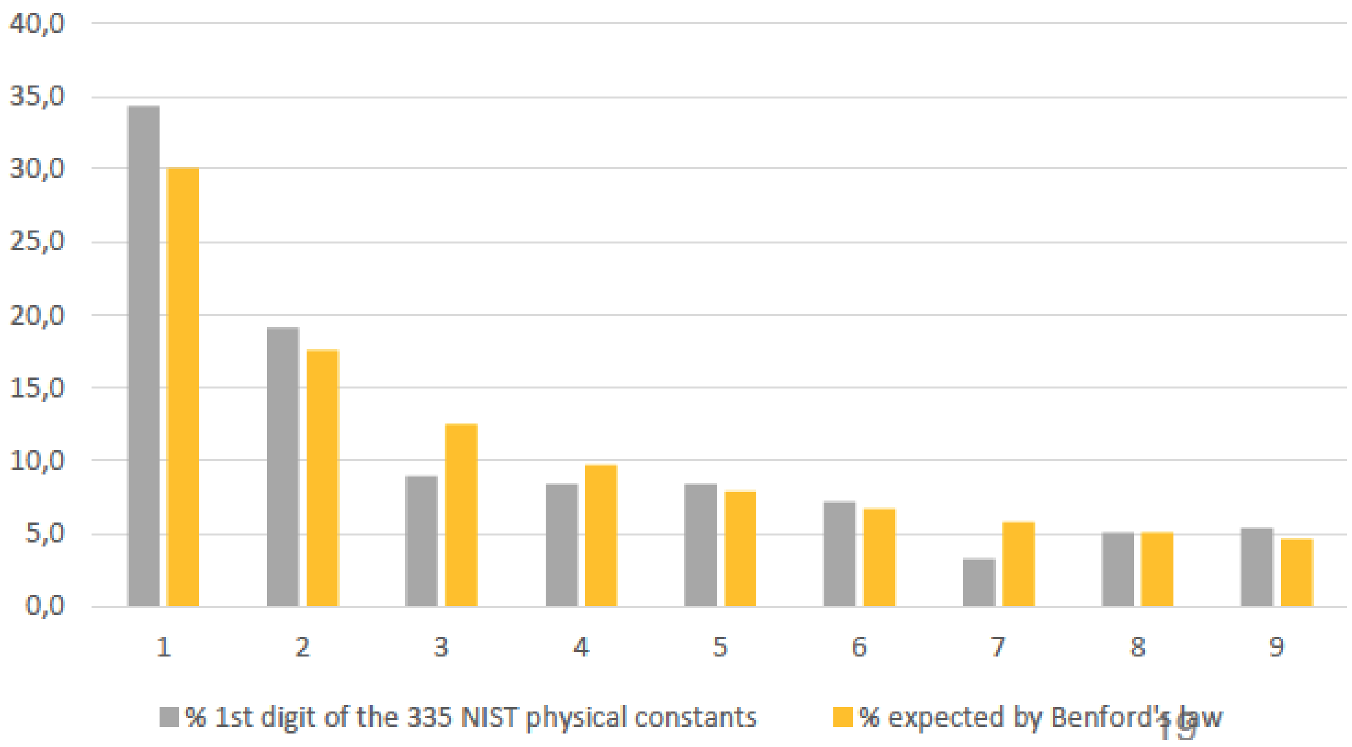
https://en.wikipedia.org/wiki/Benford%27s_law

[Benford: “The law of anomalous numbers” *Proc. Am. Philos. Soc*., 1938]

Examples of Benford Law

- surface area of 355 rivers
- sizes of 3,259 US populations
- 104 physical constants
- 1,800 molecular weights
- 308 numbers contained in an issue of Reader’s Digest
- Street addresses of the first 342 persons listed in American Men of Science
- ….

used in fraud detection!

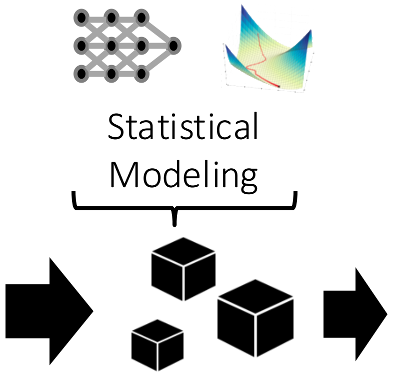


physical constants

[Abedjan, Golab & Naumann (2015)]

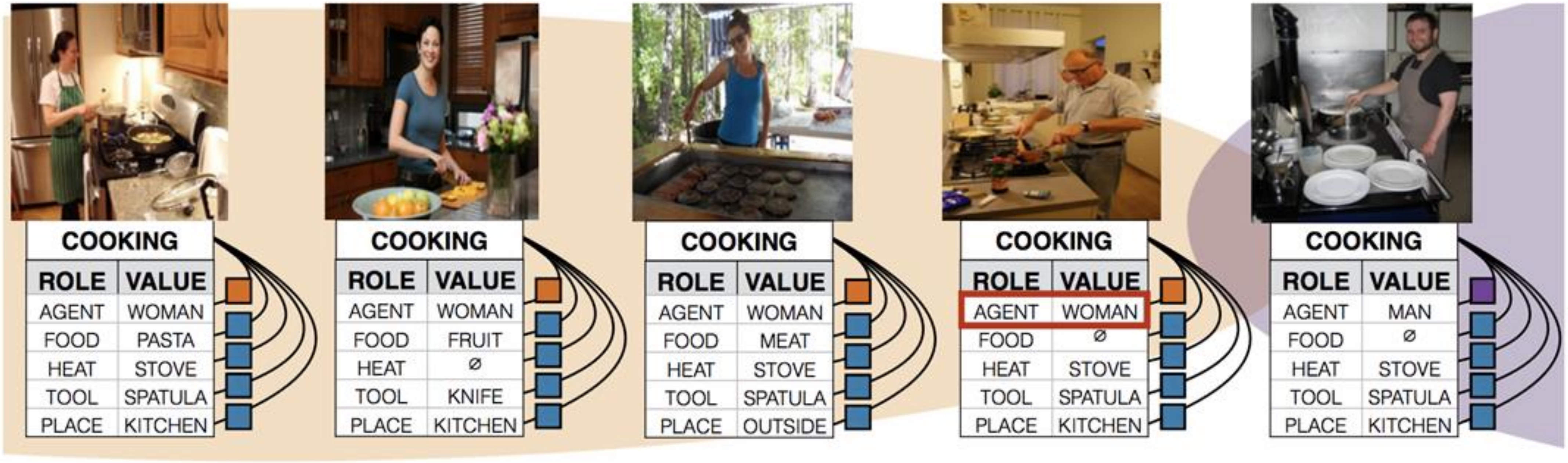
Statistical Modeling

- Statistical modeling: deciding how to crate a model for data and how it will be trained



- *What model do we use? What training algorithm?*

Unfairness in Model: Bias Amplification



COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

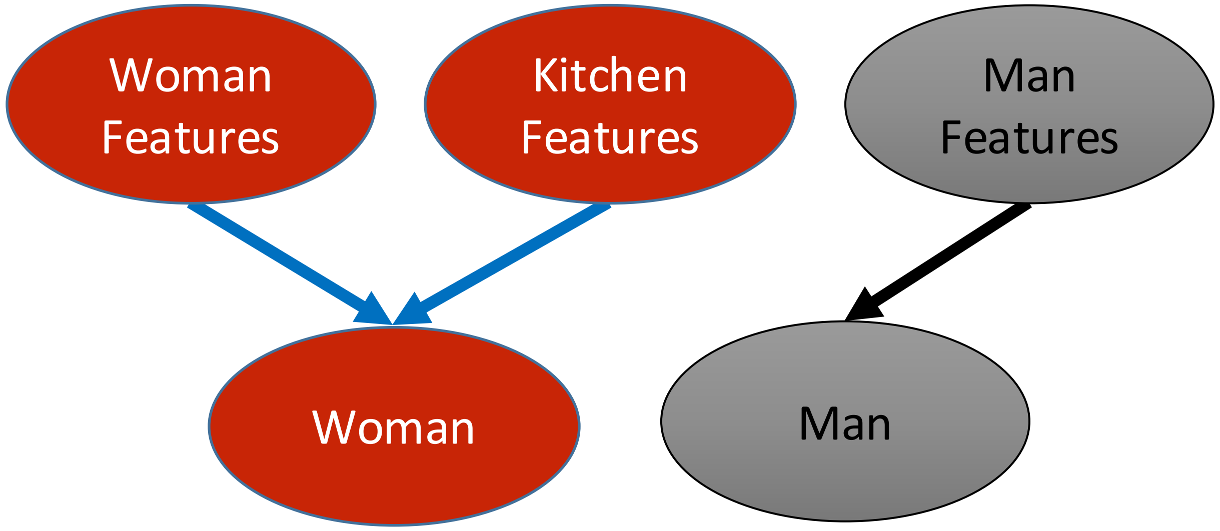
COOKING

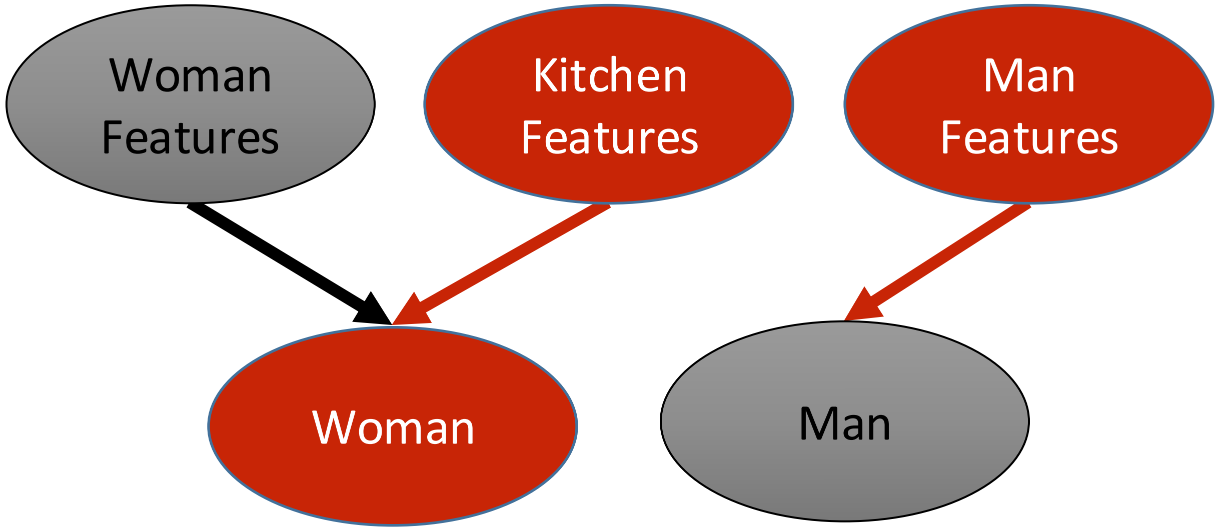
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Class	Man	Woman
Data prior	33%	67%
Pred. prior	16%	84%

Unfairness in Model: Bias Amplification

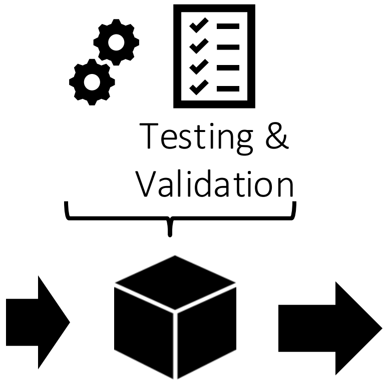
We say a model exhibits *bias amplification* if the prior distribution of the model's predictions does not match that of the data: in particular, we don't want the model to *create* or *exaggerate disparities* in the training data.





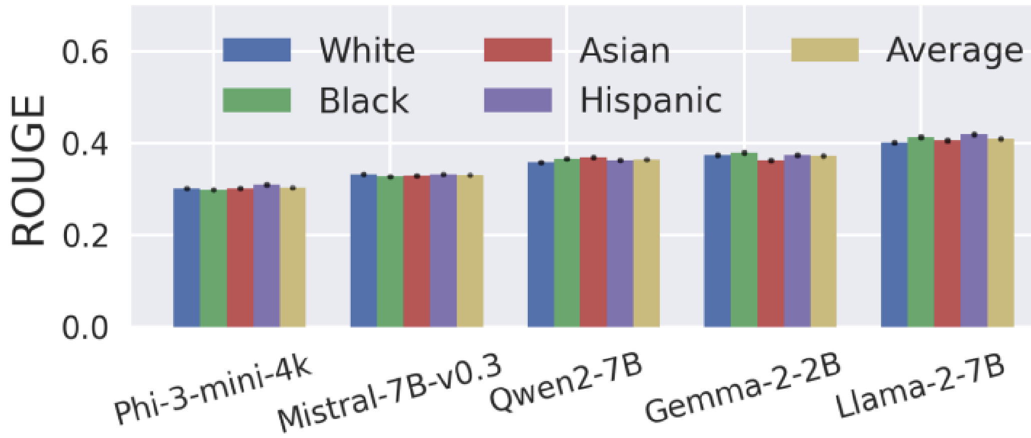
Testing and Validation

- Testing and validation: the processes by which a model is determined to be performing well, both in relation to other models in the training set, but also on unseen data.

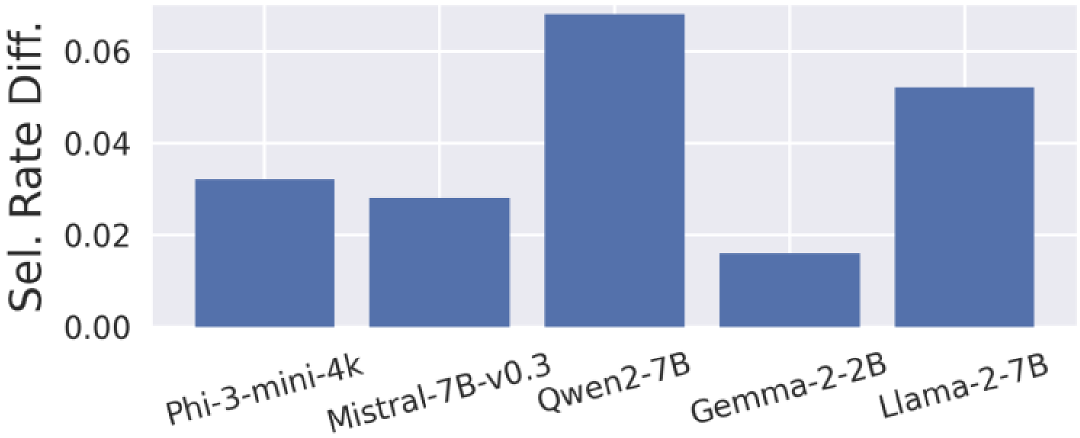


- *What data will we test our model with? What metrics will we use?*

Equalizing "Generation Quality" Metrics May not Capture Discriminatory Impacts



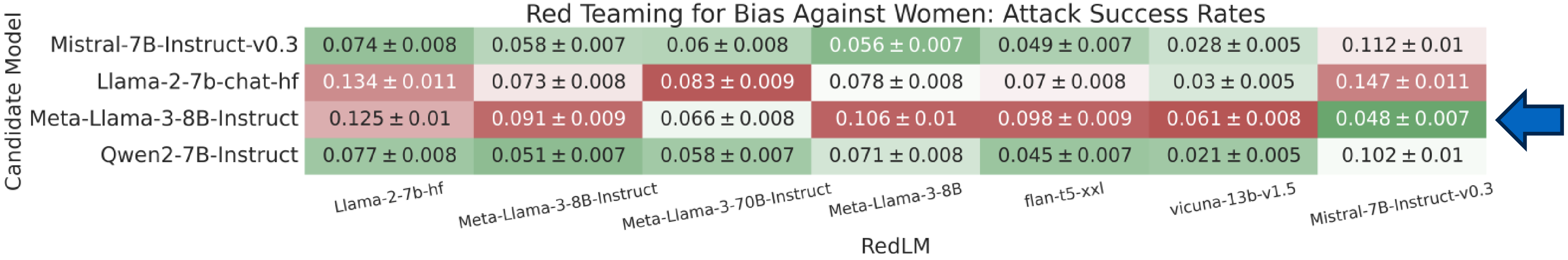
Rouge score for resume summaries across different demographic groups



Selection rate (hiring) differences based on those summaries

Towards Effective Discrimination Testing for Generative AI. Zollo, Rajaneesh, Zemel, Gillis, Black

Evaluation schema that do exist are incredibly unstable



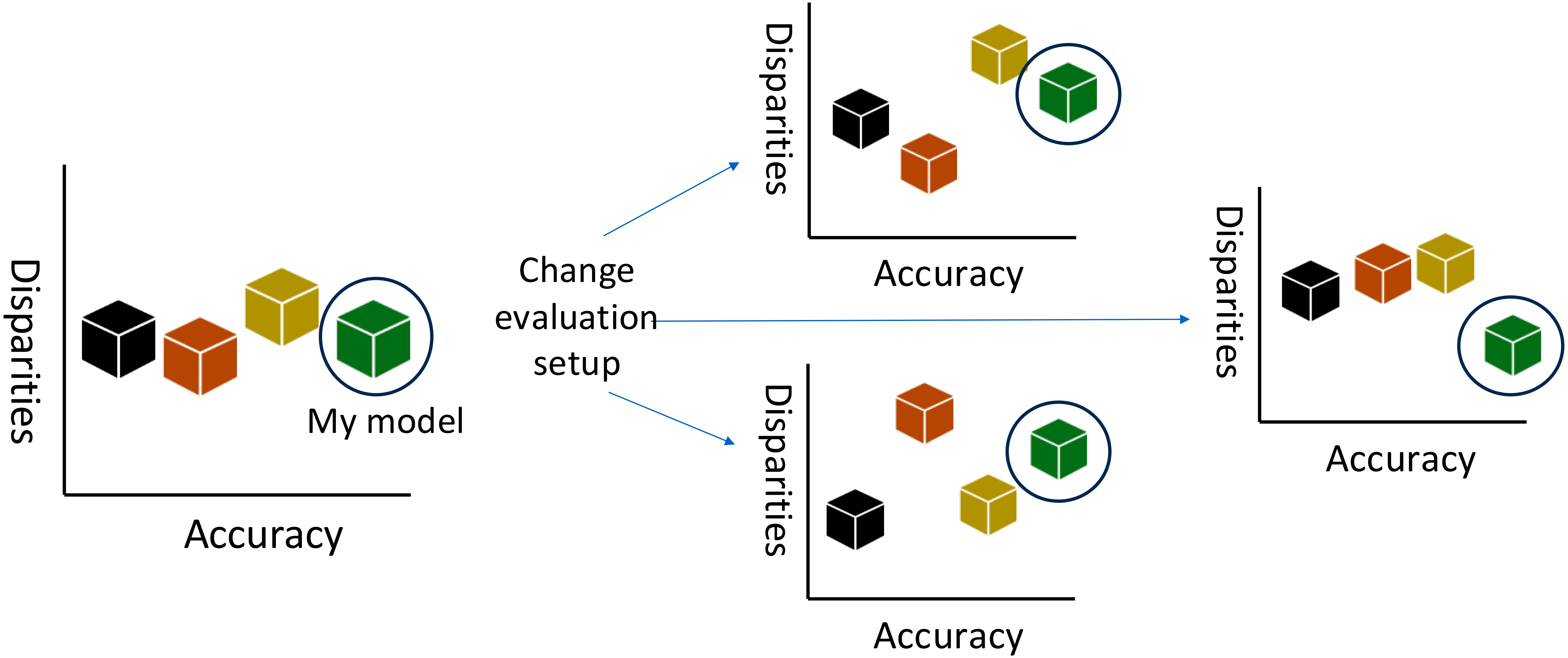
Red Teaming for Bias Against Women: Attack Success Rates

Candidate Model \ RedLM	Llama-2-7b-hf	Meta-Llama-3-8B-Instruct	Meta-Llama-3-70B-Instruct	Meta-Llama-3-8B	flan-t5-xxl	vicuna-13b-v1.5	Mistral-7B-Instruct-v0.3
Mistral-7B-Instruct-v0.3	0.074 ± 0.008	0.058 ± 0.007	0.06 ± 0.008	0.056 ± 0.007	0.049 ± 0.007	0.028 ± 0.005	0.112 ± 0.01
Llama-2-7b-chat-hf	0.134 ± 0.011	0.073 ± 0.008	0.083 ± 0.009	0.078 ± 0.008	0.07 ± 0.008	0.03 ± 0.005	0.147 ± 0.011
Meta-Llama-3-8B-Instruct	0.125 ± 0.01	0.091 ± 0.009	0.066 ± 0.008	0.106 ± 0.01	0.098 ± 0.009	0.061 ± 0.008	0.048 ± 0.007
Qwen2-7B-Instruct	0.077 ± 0.008	0.051 ± 0.007	0.058 ± 0.007	0.071 ± 0.008	0.045 ± 0.007	0.021 ± 0.005	0.102 ± 0.01

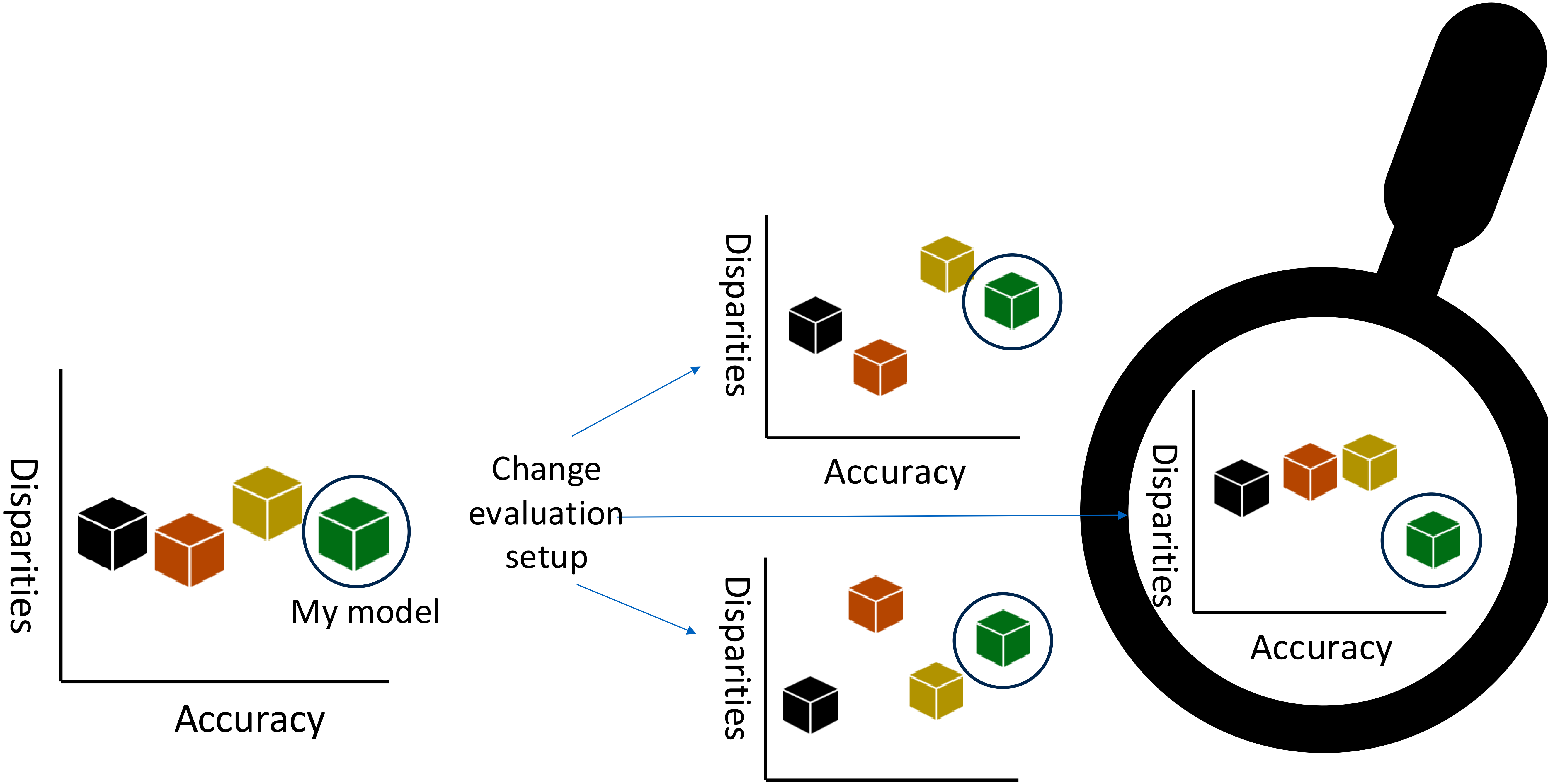
Instability of toxicity scores based on red-teaming with various base RedLM models

Towards Effective Discrimination Testing for Generative AI.
Zollo, Rajaneesh, Zemel, Gillis, Black

Discrimination hacking (d-hacking)



Discrimination hacking (d-hacking)

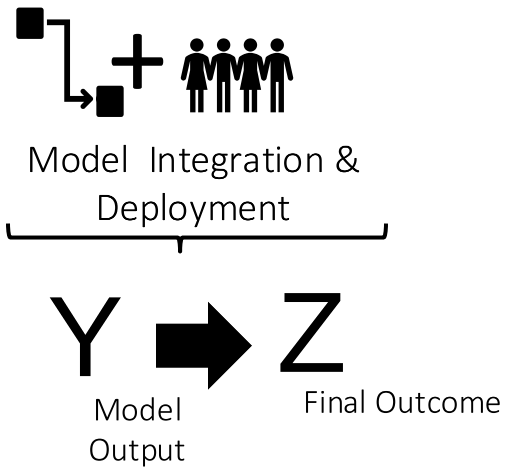


Change evaluation setup

Look, lower disparity!!!

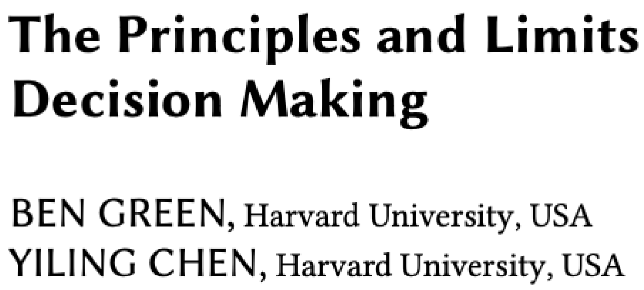
Model Integration and Deployment

- Model Integration/ Deployment: collecting or compiling data to train the model



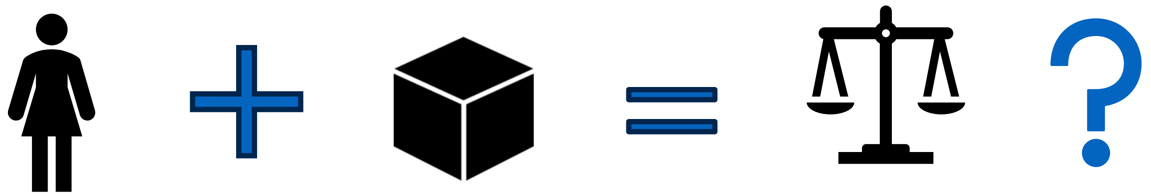
- *How do we communicate model output to a human decision-maker?*
- *How do we ensure good performance over time?*

Model Integration and Deployment



The Principles and Limits of Algorithm-in-the-Loop Decision Making

BEN GREEN, Harvard University, USA
YILING CHEN, Harvard University, USA



- How do humans deviate from AI predictions? Do they do so in a biased way?
- How consistent are human + AI decisions versus just human/just AI?

- How well do humans understand the information coming from the model? Overreliance, under-reliance, etc…

What do we do about all of this?

- Be a data detective!
- Design your evaluations carefully!
- Data science is also an art: But be explicit about your choices an assumptions
- Document your code, your data–so that the next person understands your choices and assumptions
- We’re still working on it---join the fun!

Responsible Data Science

The data science lifecycle

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai center for responsible ai

association rule mining

The early days of data mining

- Problem formulation due to Agrawal, Imielinski, Swami, SIGMOD 1993
- Solution: the **Apriori** algorithm by Agrawal & Srikant, VLDB 1994
- Initially for **market-basket data** analysis, has many other applications, we'll see one today
- We wish to answer two related questions:
  - **Frequent itemsets:** Which items are often purchased together, e.g., milk and cookies are often bought together
  - **Association rules:** Which items will likely be purchased, based on other purchased items, e.g., if diapers are bought in a transaction, beer is also likely bought in the same transaction

Market-basket data

- $\boldsymbol{I = \{i_1, i_2, \ldots, i_m\}}$ is the set of available items, e.g., a product catalog of a store
- $\boldsymbol{X \subseteq I}$ is an **itemset**, e.g., {milk, bread, cereal}
- **Transaction** $\boldsymbol{t}$ is a set of items purchased together, $\boldsymbol{t \subseteq I}$, has a transaction id (TID)

  $\boldsymbol{t_1}$: {bread, cheese, milk}

  $\boldsymbol{t_2}$: {apple, eggs, salt, yogurt}

  $\boldsymbol{t_3}$: {biscuit, cheese, eggs, milk}

- Database $\boldsymbol{T}$ is a set of transactions $\boldsymbol{\{t_1, t_2, \ldots, t_n\}}$
- A transaction $\boldsymbol{t}$ **supports** an itemset $\boldsymbol{X}$ if $\boldsymbol{X \subseteq t}$
- Itemsets supported by at least ***minSupp*** transactions are called **frequent itemsets**

minSupp, which can be a number or a percentage, is specified by the user

Itemsets

TID	Items
1	A
2	A C
3	A B D
4	A C
5	A B C
6	A B C

minSupp = 2 transactions

How many possible itemsets are there (excluding the empty itemset)?

$2^4 - 1 = 15$

itemset	support
A	6
B	3
C	4
D	1
A B	3
A C	4
A D	1
B C	2
B D	1
C D	0
A B C	2
A B D	1
B C D	0
A C D	0
A B C D	0

Association rules

An **association rule** is an implication $X \rightarrow Y$, where $X, Y \subset I$, *and* $X \cap Y = \varnothing$

example: {milk, bread} → {cereal}

"A customer who purchased X is also likely to have purchased Y in the same transaction"

we are interested in rules with a **single item** in Y

can we represent {milk, bread} → {cereal, cheese}?

Rule $X \rightarrow Y$ holds with **support** *supp* in T if *supp* of transactions contain $X \cup Y$

Rule $X \rightarrow Y$ holds with confidence *conf* in T if *conf* % of transactions that contain X also contain Y

$conf \approx \Pr(Y \mid X)$

$conf\ (X \rightarrow Y) = supp\ (X \cup Y)\ /\ supp\ (X)$

Association rules

minSupp = 2 transactions
minConf = 0.75

supp = 3

$A \rightarrow B$ conf = 3 / 6 = 0.5
$B \rightarrow A$ conf = 3 / 3 = 1.0 ★

supp = 2

$B \rightarrow C$ conf = 2 / 3 = 0.67
$C \rightarrow B$ conf = 2 / 4 = 0.5

supp = 4

$A \rightarrow C$ conf = 4 / 6 = 0.67
$C \rightarrow A$ conf = 4 / 4 = 1.0 ★

supp = 2

$AB \rightarrow$ X conf = 2 / 3 = 0.67
$AC \rightarrow B$ conf = 2 / 4 = 0.5
$BC \rightarrow A$ conf = 2 / 2 = 1.0 ★

	itemset	support
★	A	6
★★	B	3
	C	4
★	D	1
★	A B	3
★	A C	4
	A D	1
	B C	2
★	B D	1
	C D	0
	A B C	2
	A B D	1
	B C D	0
	A C D	0
	A B C D	0

***conf* (*X* → *Y*) = *supp* (X U Y) / *supp* (X)**

Association rule mining

- Goal: find all association rules that satisfy the user-specified minimum support and minimum confidence

- Algorithm outline
  - Step 1: find all frequent itemsets
  - Step 2: find association rules

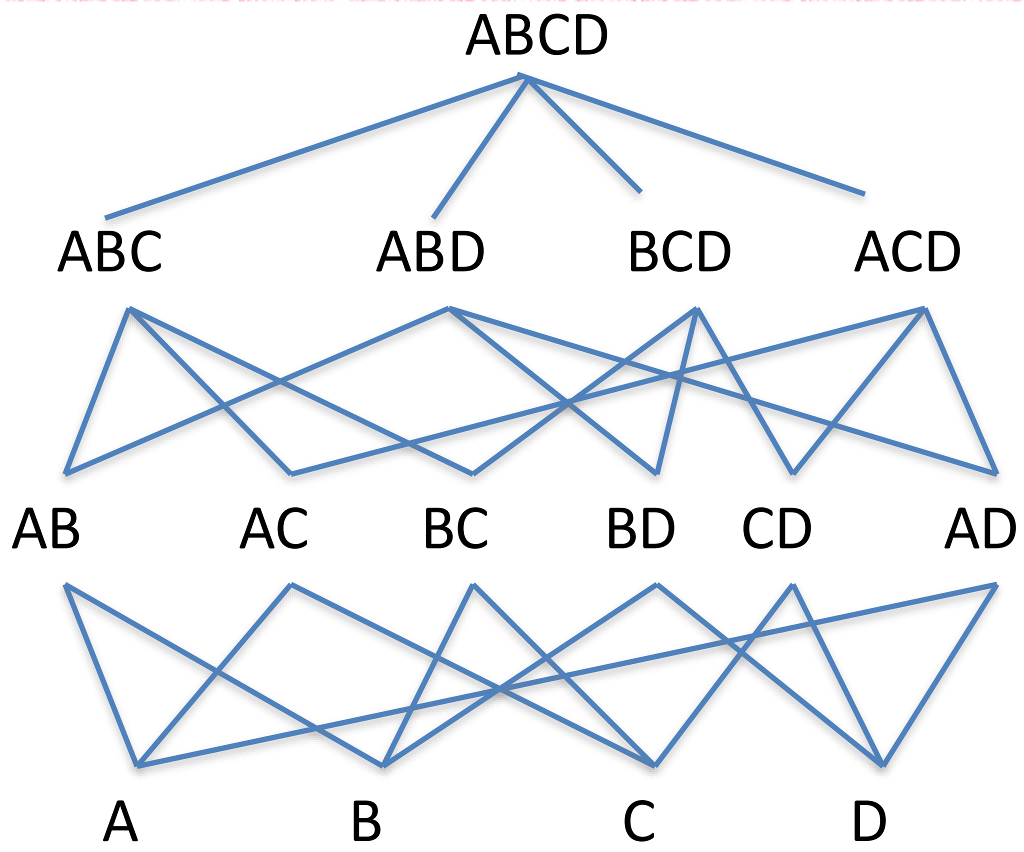
- Take 1: naïve algorithm for frequent itemset mining
  - Enumerate all subsets of ***I***, check their support in ***T***
  - **What is the complexity?**

Key idea: downward closure

	itemset	support
★	A	6
★★	B	3
	C	4
★	D	1
★	A B	3
★	A C	4
	A D	1
	B C	2
★	B D	1
	C D	0
	A B C	2
	A B D	1
	B C D	0
	A C D	0
	A B C D	0

All subsets of a frequent itemset ***X*** are themselves frequent

So, if some subset of X is infrequent, then X cannot be frequent, we know this **apriori**



The converse is not true! If all subsets of ***X*** are frequer

X is not guaranteed to be frequent

The *Apriori* algorithm

```
Algorithm Apriori($T$, $minSupp$)
    $F_1$ = {frequent 1-itemsets};
    for ($k = 2$; $F_{k-1} \neq \emptyset$; $k$++) do
        $C_k \leftarrow$ candidate-gen($F_{k-1}$);
        for each transaction $t \in T$ do
            for each candidate $c \in C_k$ do
                if $c$ is contained in $t$ then
                    $c.count$++;
            end
        end
        $F_k \leftarrow \{c \in C_k \mid c.count \geq minSupp\}$
    end
return $F \leftarrow \bigcup_k F_k$;
```

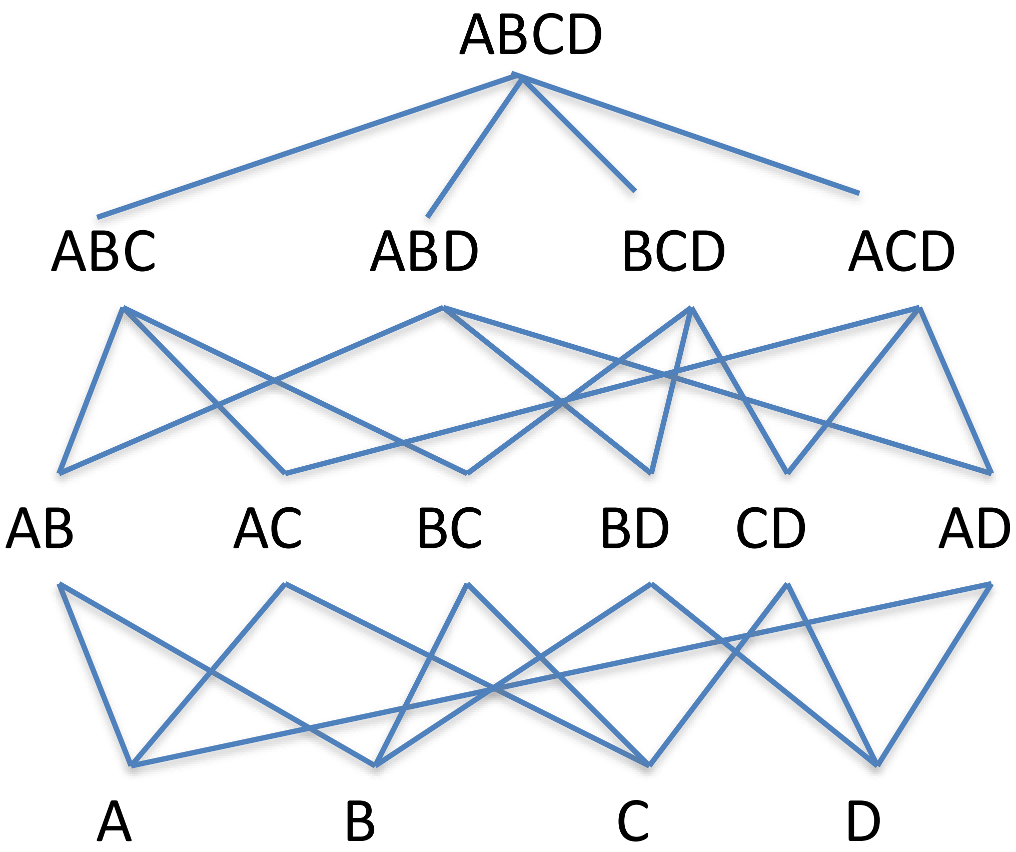
itemset	support
A	6
B	3
C	4
D	1
A B	3
A C	4
A D	1
B C	2
B D	1
C D	0
A B C	2
A B D	1
B C D	0
A C D	0
A B C D	0

Candidate generation

The **candidate-gen** function takes F_{k-1} and returns a superset (called the candidates) of the set of all frequent k-itemsets. It has two steps:

Join: generate all possible candidate itemsets C_k of length k

Prune: optionally remove those candidates in C_k that have infrequent subsets



Candidate generation: join

```
Insert into $C_k$ (
 select   p.$item_1$, p.$item_2$, …, p.$item_{k-1}$, q.$item_{k-1}$
 from     $F_{k-1}$ p, $F_{k-1}$ q
 where    p.$item_1$ = q.$item_1$
  and       p.$item_2$ = q.$item_2$
     and  …
     and  p.$item_{k-1}$ < q.$item_{k-1}$ )
```

F_1 as p

A
B
C

F_1 as q

A
B
C

C_2

A	**B**
A	**C**
B	**C**

	itemset	support
★	A	6
★★	B	3
	C	4
★	D	1
★	A B	3
★	A C	4
	A D	1
	~~B C~~	~~2~~
★	B D	1
	C D	0
	A B C	2
	A B D	1
	B C D	0
	A C D	0
	A B C D	0

Candidate generation: join

```
Insert into $C_k$ (
 select   $p.item_1, p.item_2, …, p.item_{k-1}, q.item_{k-1}$
 from     $F_{k-1}$ p, $F_{k-1}$ q
 where    $p.item_1 = q.item_1$
  and        $p.item_2 = q.item_2$
    and   …
    and   $p.item_{k-1} < q.item_{k-1}$ )
```

F_2 as p

A	**B**
A	C
B	C

F_2 as q

A	B
A	**C**
B	C

C_3

A	**B**	**C**

	itemset	support
★★	A	6
★	B	3
	C	4
★	D	1
★	A B	3
★	A C	4
	A D	1
	~~B C~~	~~2~~
★	B D	1
	C D	0
	A B C	2
	A B D	1
	B C D	0
	A C D	0
	A B C D	0

Candidate generation

Assume a lexicographic ordering of the items

Join

```
Insert into $C_k$ (
    select  p.$item_1$, p.$item_2$, …, p.$item_{k-1}$, q.$item_{k-1}$
    from    $F_{k-1}$ p, $F_{k-1}$ q
    where       p.$item_1$ = q.$item_1$
  and     p.$item_2$ = q.$item_2$
    and       …
    and       p.$item_{k-1}$ < q.$item_{k-1}$ )
```

why not p.$item_{k-1} \neq$ q.$item_{k-1}$?

Prune

```
for each c in $C_k$ do
    for each (k-1) subset s of c do
        if (s not in $F_{k-1}$) then
            delete c from $C_k$
```

Generating association rules

```
Rules = Ø
for each frequent k-itemset X do
        for each 1-itemset $A \subset X$ do
          compute conf $(X / A \rightarrow A)$ = supp(X) / sup $(X / A)$
          if conf $(X / A \rightarrow A) \geq$ minConf then
            Rules $\leftarrow$ □X / A → A"
        end
    end
end
return Rules
```

Performance of *Apriori*

- The possible number of frequent itemsets is exponential, O($\boldsymbol{2^m}$), where ***m*** is the number of items
- Apriori exploits sparseness and locality of data
  - Still, it may produce a large number of rules: thousands, tens of thousands, ….
  - So, thresholds should be set carefully. What are some good heuristics?

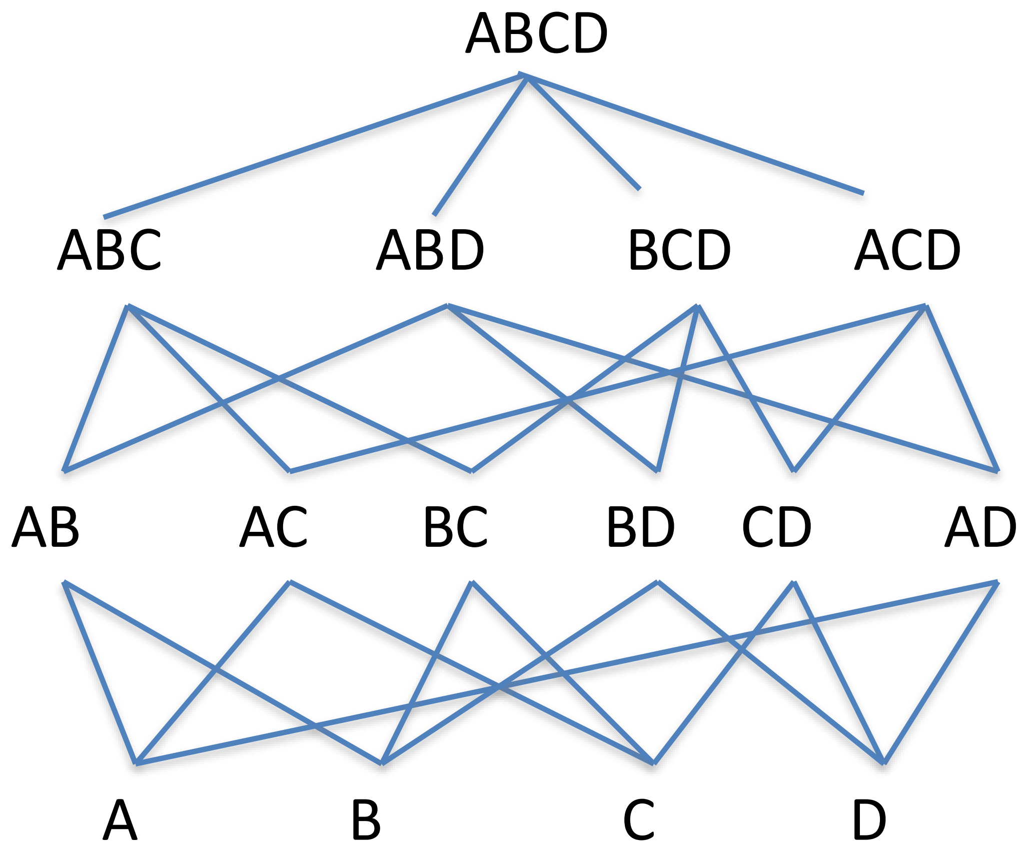
Discovering uniques

Given a relation schema ***R** (A, B, C, D)* and a relation instance ***r***, a **unique column combination** (or a **"unique"** for short) is a set of attributes ***X*** whose **projection** contains no duplicates in ***r***

Given a relation schema ***R** (A, B, C, D)* and a relation instance ***r***, a set of attributes ***Y*** is **non-unique** if its projection contains duplicates in ***r***

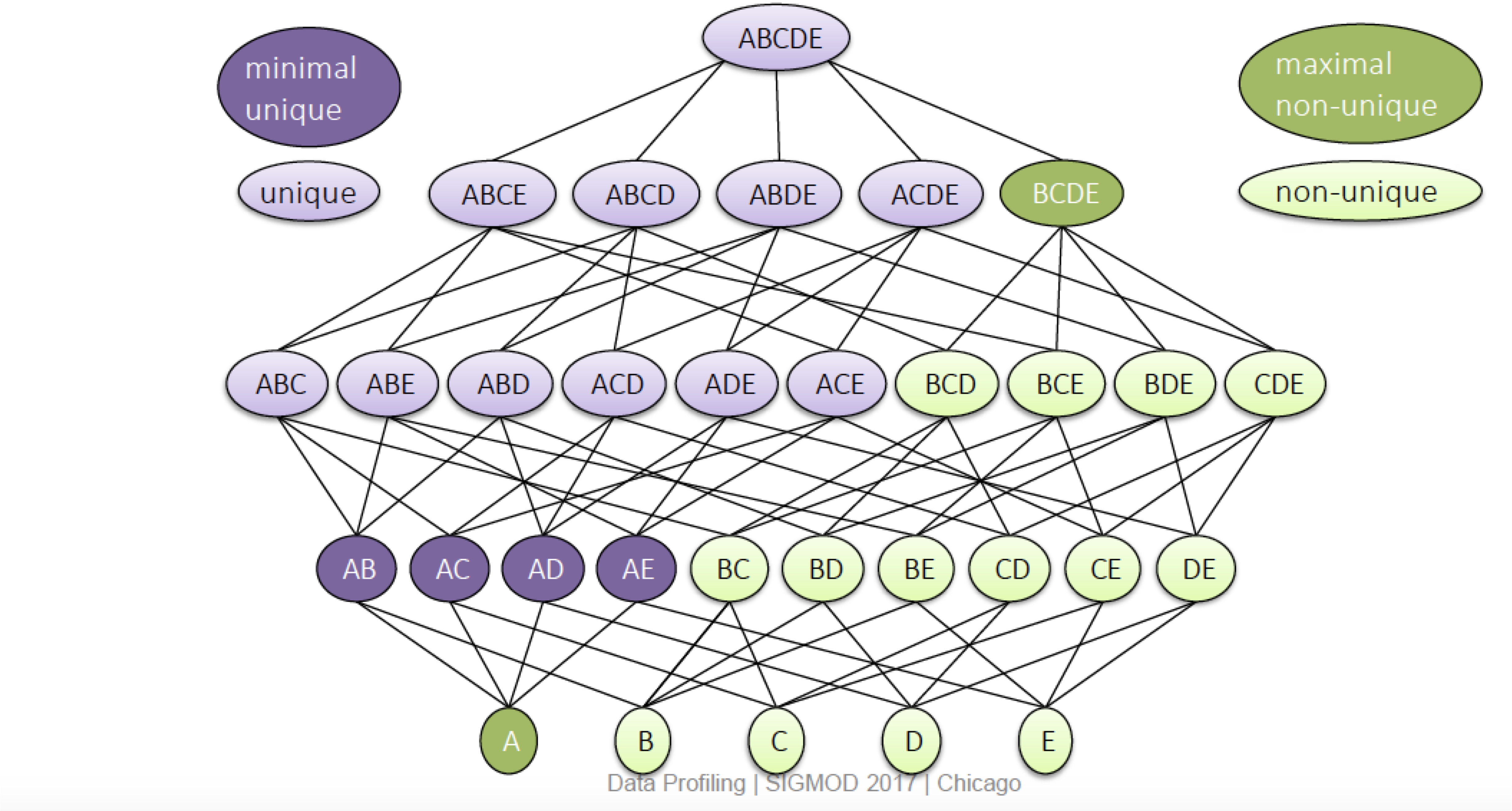
X is **minimal unique** if every subset ***Y*** of ***X*** is non-unique

Y is maximal non-unique if every superset ***X*** of ***Y*** is unique



[Abedjan, Golab, Naumann; *SIGMOD 2017*]

Output



From uniques to candidate keys

Given a relation schema ***R (A, B, C, D)*** and a relation instance ***r***, a **unique column combination** is a set of attributes ***X*** whose **projection** contains no duplicates in ***r***

$\mathit{Episodes}(\mathit{season}, \mathit{num}, \mathit{title}, \mathit{viewers})$

season	num	title	viewers
1	1	Winter is Coming	2.2 M
1	2	The Kingsroad	2.2 M
2	1	The North Remembers	3.9 M

A set of attributes is a **candidate key** for a relation if:
(1) no two distinct tuples can have the value values for all key attributes (candidate key **uniquely identifies** a tuple), *and*
(2) this is not true for any subset of the key attributes (candidate key **is minimal**)

A minimal unique of a relation instance is a (possible) candidate key of the relation schema. To find such possible candidate keys, find all minimal uniques in a given relation instance.

Apriori-style uniques discovery

[Abedjan, Golab, Naumann; *SIGMOD 2017*]

A minimal unique of a relation instance is a **(possible) candidate key** of the relation schema.

```
Algorithm Uniques  // sketch, similar to HCA
    $U_1$ = {1-uniques}    $N_1$ = {1-non-uniques}
    for ($k = 2; N_{k-1} \neq \emptyset; k{+}{+}$) do
        $C_k \leftarrow$ candidate-gen($N_{k-1}$)
        $U_k \leftarrow$ prune-then-check ($C_k$)
            // prune candidates with unique sub-sets, and with value distributions that cannot be unique
            // check each candidate in pruned set for uniqueness
        $N_k \leftarrow C_k \setminus U_k$
    end
return $U \leftarrow \bigcup_k U_k$;
```

breadth-first bottom-up strategy for attribute lattice traversal