Responsible Data Science

The data science lifecycle

February 20, 2025

Prof. Jonathan Colner

Center for Data Science
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai center for responsible ai

Team Project

DS-UA 202, Responsible Data Science, Spring 2024
Course Project: Technical Audit of an Automated Decision System
assigned on February 20, 2025; see description for due dates

Objectives

In this project, you will work in **teams of two** to conduct a technical audit of an automated decision system (ADS) of your choice. We suggest that you audit one of the systems developed in response to a Kaggle competition of your choice, but you should feel free to use other systems that are of interest to you. **Do not focus on Northpointe's COMPAS** in this assignment, since this tool was already covered extensively during class. Be sure to prominently cite your sources of code and data!

Both team members should work together on all parts of the project. You should not discuss your project submission or components of your solution with any students other than your project partner. If you have questions about this assignment, please send a private question to all instructors over email.

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

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

The VLDB Journal (2015) 24:557–581
DOI 10.1007/s00778-015-0389-y

REGULAR PAPER

Profiling relational data: a survey

Ziawasch Abedjan[1] · Lukasz Golab[2] · Felix Naumann[3]

Received: 1 August 2014 / Revised: 5 May 2015 / Accepted: 13 May 2015 / Published online: 2 June 2015

Abstract Profiling data to determine metadata about a given dataset is an important and frequent activity of any IT professional and researcher and is necessary for various use-cases. It encompasses a vast array of methods to examine datasets and produce metadata. Among the simpler results are statistics, such as the number of null values and distinct values in a column, its data type, or the most frequent patterns of its data values. Metadata that are more difficult to compute involve multiple columns, namely correlations, unique column combinations, functional dependencies, and inclusion dependencies. Further techniques detect conditional properties of the dataset at hand. This survey provides a classification of data profiling tasks and comprehensively reviews the state of the art for each class. In addition, we review data profiling tools and systems from research and industry. We conclude with an outlook on the future of data profiling beyond traditional profiling tasks and beyond relational databases.

✉ Felix Naumann
felix.naumann@hpi.de

Ziawasch Abedjan
abedjan@csail.mit.edu

Lukasz Golab
lgolab@uwaterloo.ca

1 MIT CSAIL, Cambridge, MA, USA
2 University of Waterloo, Waterloo, Canada
3 Hasso Plattner Institute, Potsdam, Germany

1 Data profiling: finding metadata

Data profiling is the set of activities and processes to determine the metadata about a given dataset. Profiling data is an important and frequent activity of any IT professional and researcher. We can safely assume that any reader of this article has engaged in the activity of data profiling, at least by eye-balling spreadsheets, database tables, XML files, etc. Possibly, more advanced techniques were used, such as keyword searching in datasets, writing structured queries, or even using dedicated data profiling tools.

Johnson gives the following definition: "Data profiling refers to the activity of creating small but informative summaries of a database" [79]. Data profiling encompasses a vast array of methods to examine datasets and produce metadata. Among the simpler results are statistics, such as the number of null values and distinct values in a column, its data type, or the most frequent patterns of its data values. Metadata that are more difficult to compute involve multiple columns, such as inclusion dependencies or functional dependencies. Also of practical interest are approximate versions of these dependencies, in particular because they are typically more efficient to compute. In this survey we preclude these and concentrate on exact methods.

Like many data management tasks, data profiling faces three challenges: *(i)* managing the input, *(ii)* performing the computation, and *(iii)* managing the output. Apart from typical data formatting issues, the first challenge addresses the problem of specifying the expected outcome, i.e., determining which profiling tasks to execute on which parts of the data. In fact, many tools require a precise specification of what to inspect. Other approaches are more open and perform a wider range of tasks, discovering all metadata automatically.

The second challenge is the main focus of this survey and that of most research in the area of data profiling: The com-

Springer

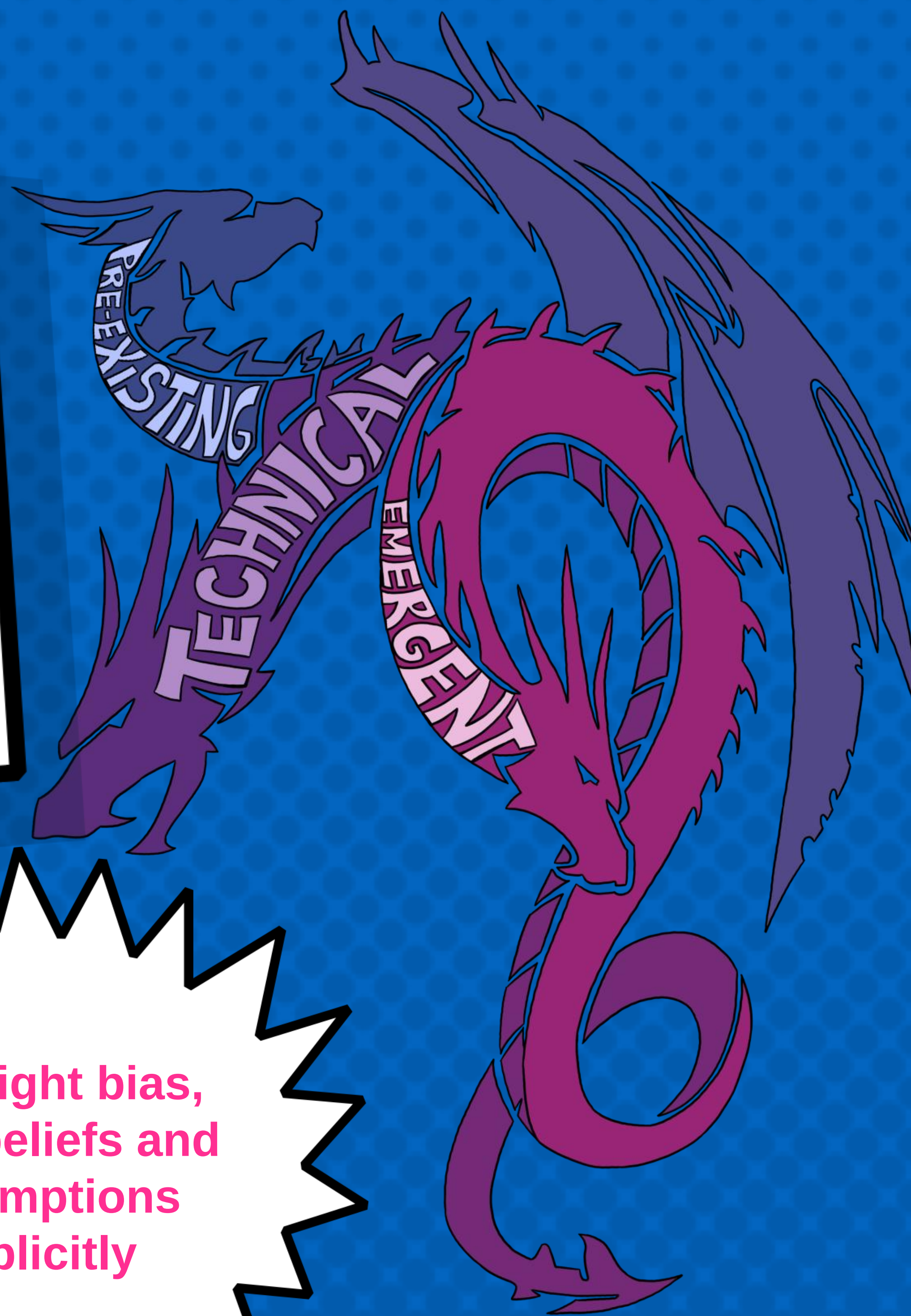
center for responsible ai

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



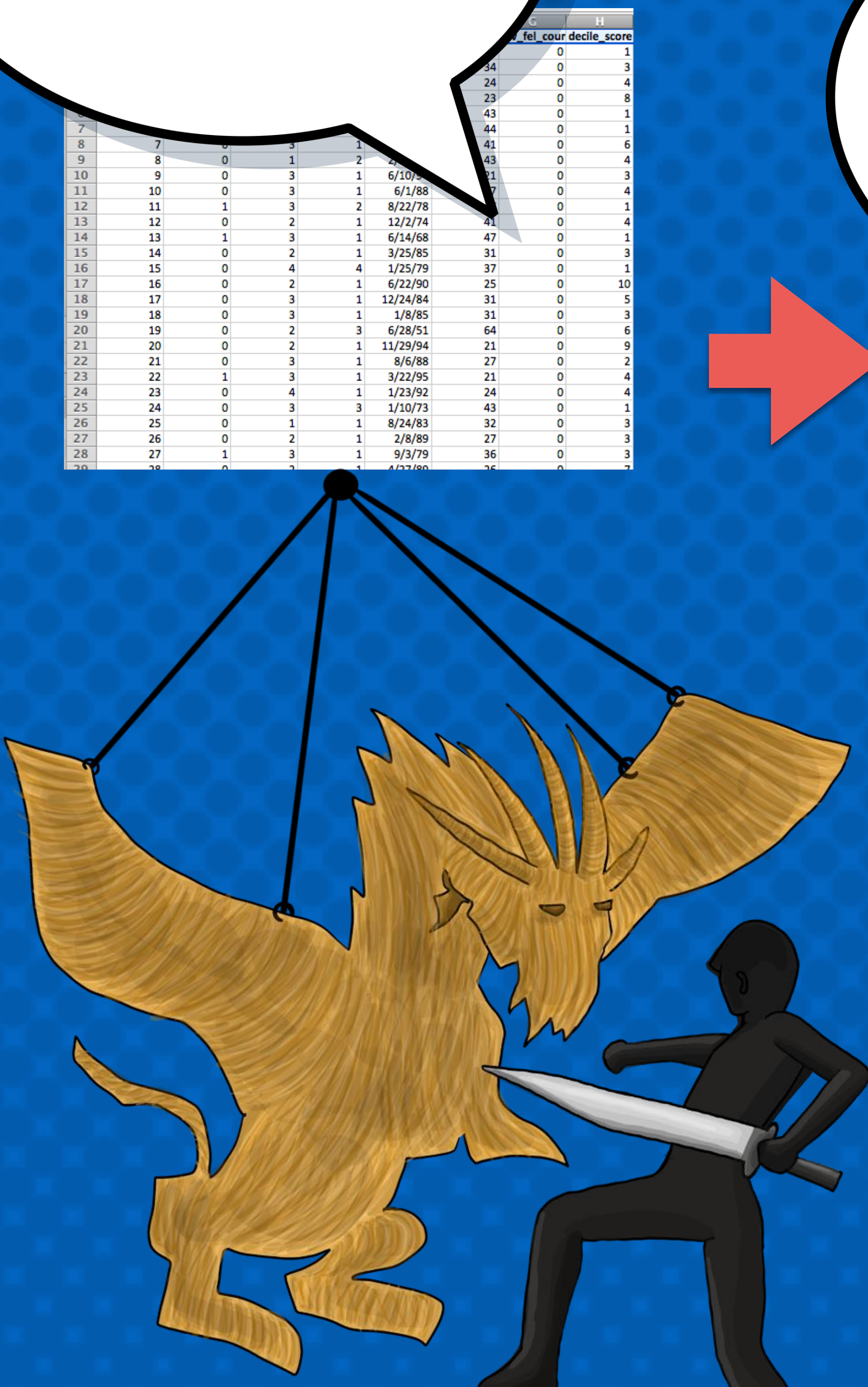
to fight bias, state beliefs and assumptions explicitly

[Friedman & Nissenbaum (1996)]

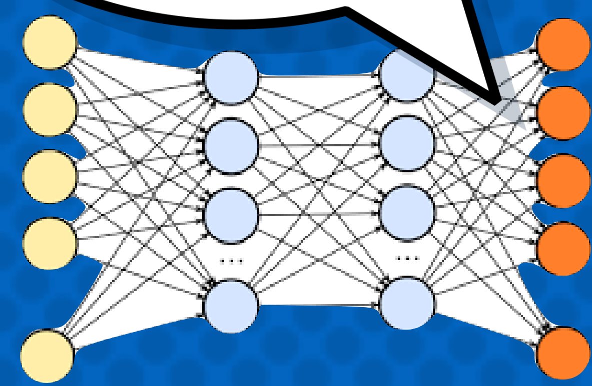
r/ai center for responsible ai

The “last-mile” view of responsible AI

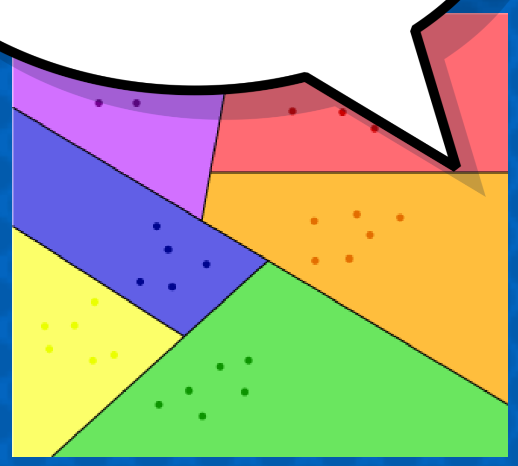
where did the data come from?



what happens inside the box?

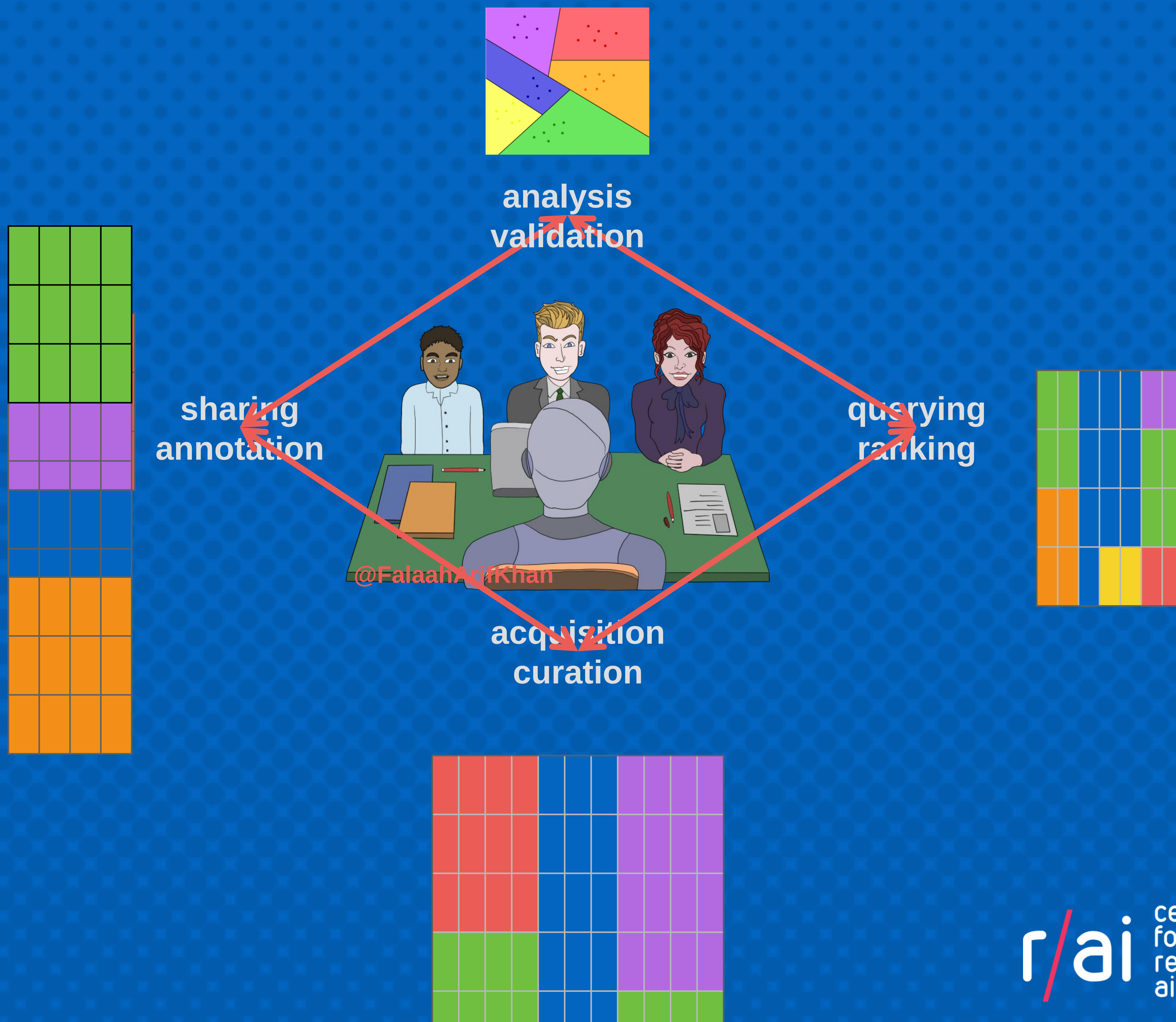


how are results used?



r/ai center for responsible ai

Data lifecycle of an ADS



r/ai center for responsible ai

Understand your data!

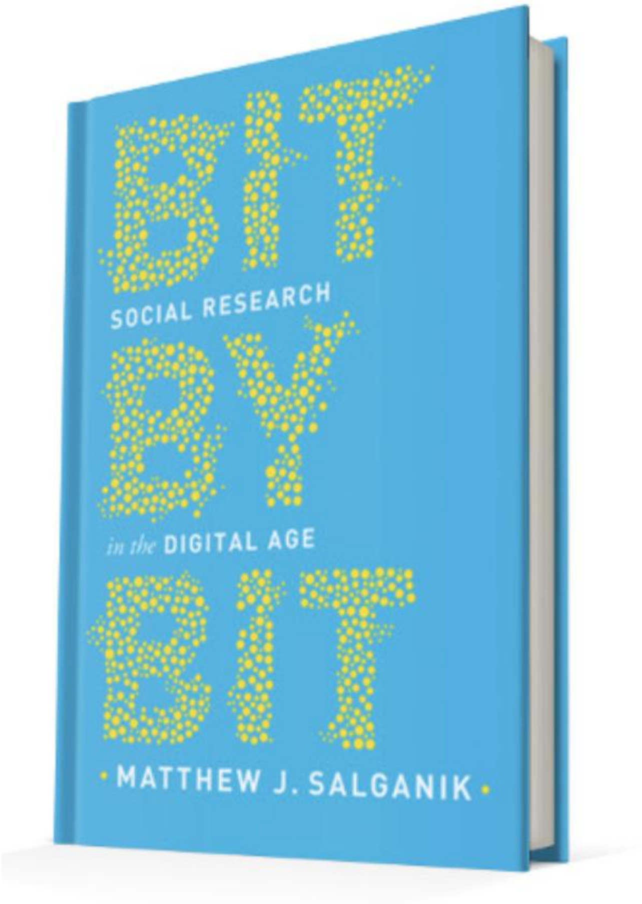
CRA
Computing Research Association

COMPUTING COMMUNITY CCC CONSORTIUM

"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!



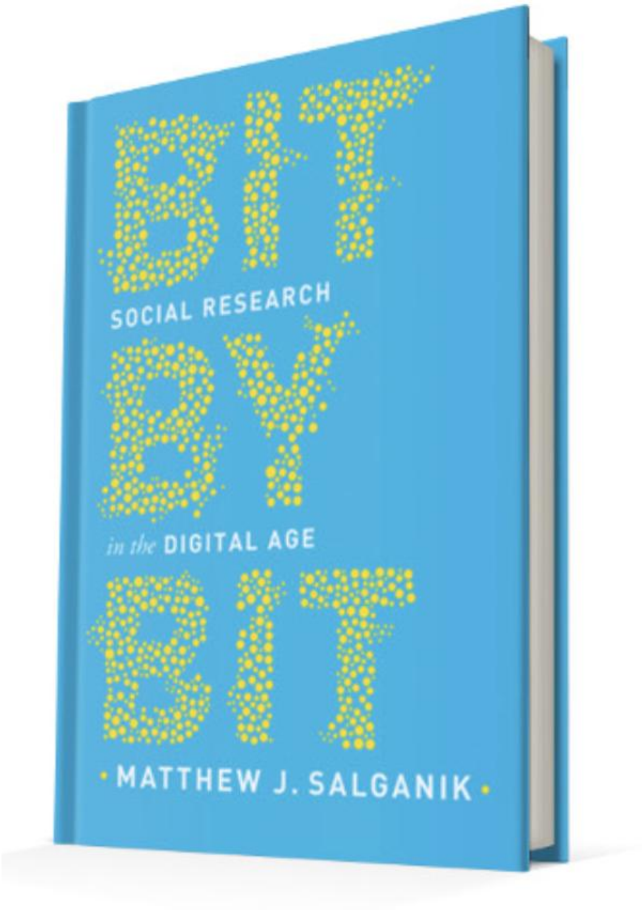
2.2 Big data

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/

center for responsible ai

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/

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]

NYC Open Data

https://opendata.cityofnewyork.us/

r/ai center for responsible ai

NYC Open Data

SAT (College Board) 2010 School Level Results

Education

Dataset

freshness

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

Updated
April 25, 2019

Views
28,463

popularity

Tags *No tags assigned*

API Docs

https://opendata.cityofnewyork.us/

NYC Open Data

About this Dataset

Updated
April 25, 2019

Data Last Updated
February 29, 2012

Metadata Last Updated
April 25, 2019

Date Created
October 6, 2011

Views
28.5K

Downloads
48.4K

Data Provided by
Department of Education (DOE)

Dataset Owner
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/

NYC Open Data

What's in this Dataset?

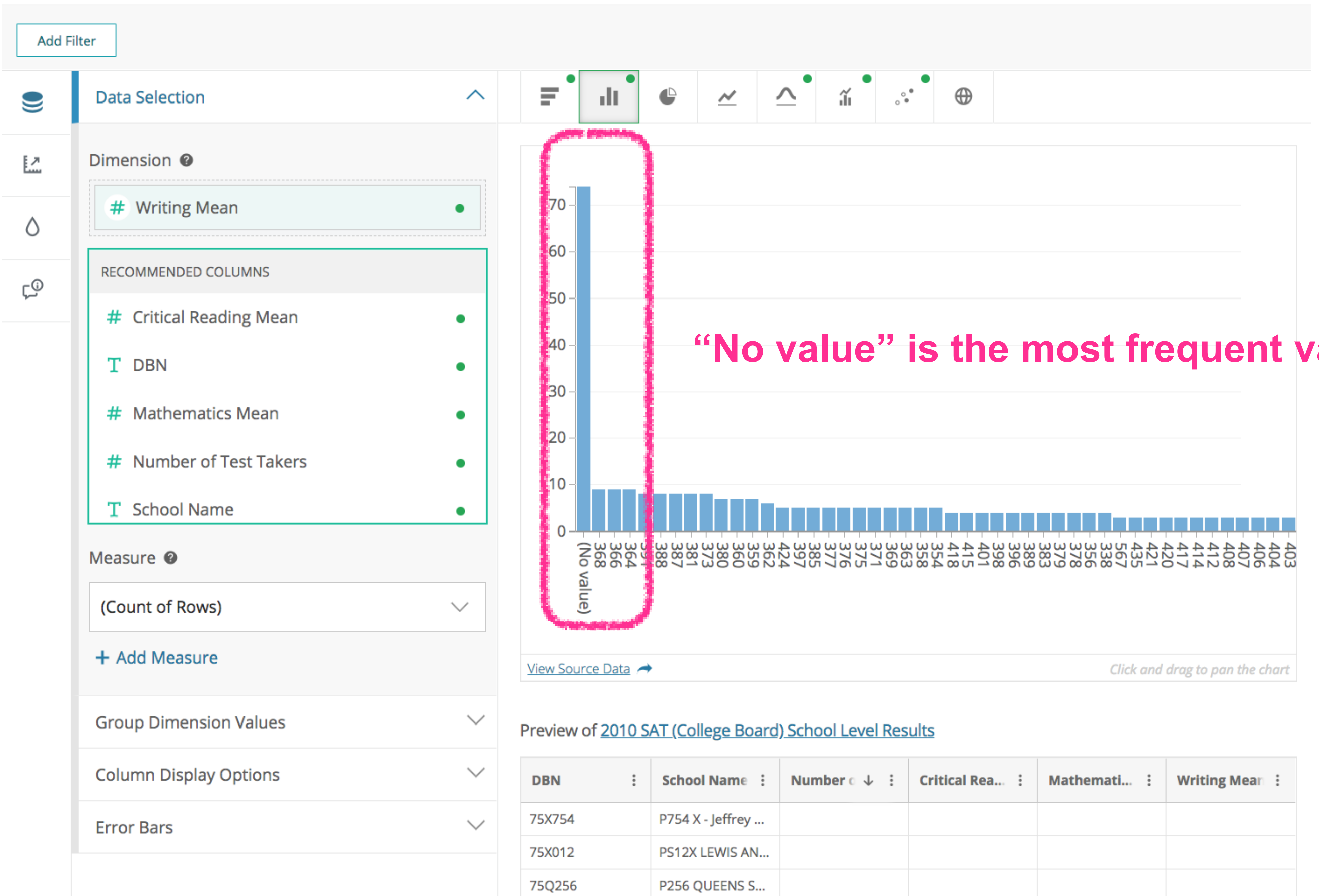
Rows	Columns
460	6

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



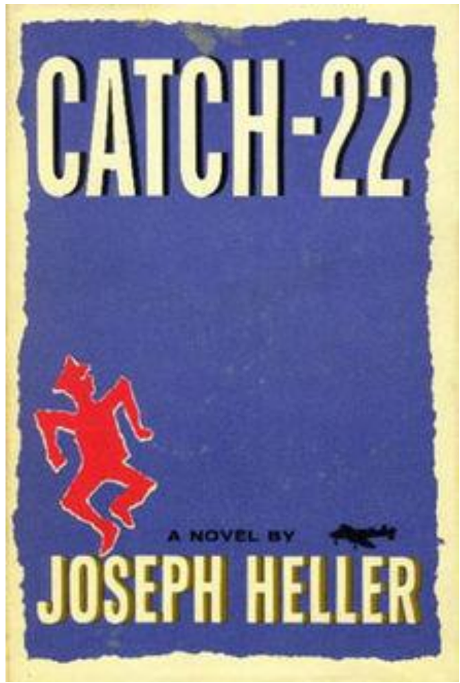
“No value” is the most frequent value

https://opendata.cityofnewyork.us/

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

center for responsible ai

Data cleaning

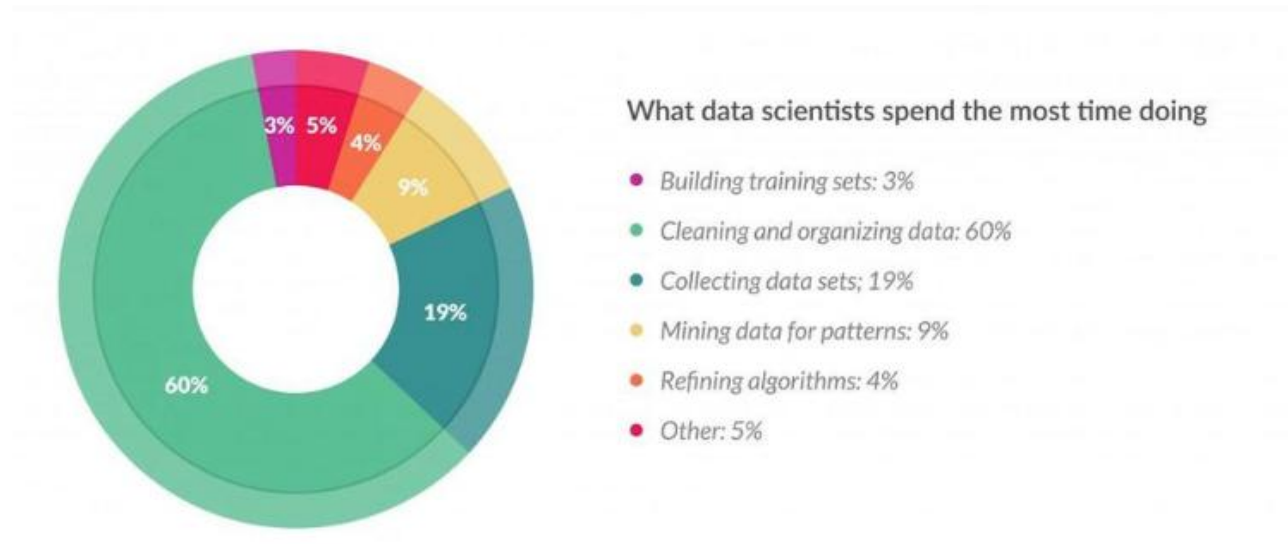

52,423 views | Mar 23, 2016, 09:33am

Forbes

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

Gil Press Contributor

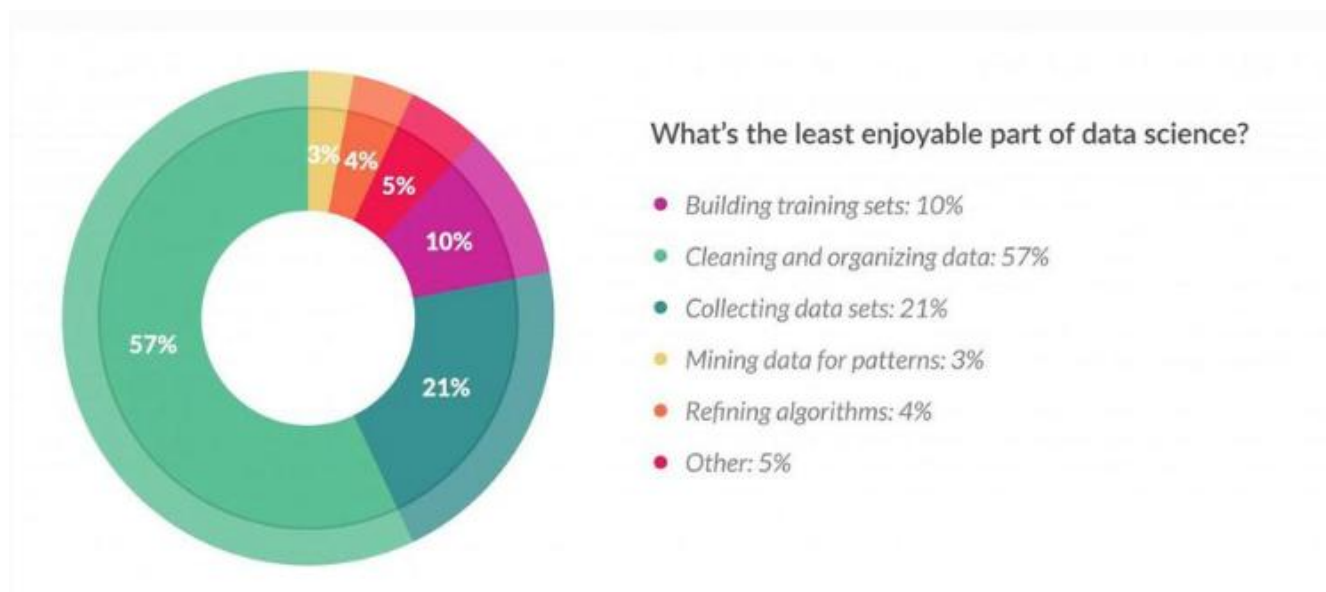
I write about technology, entrepreneurs and innovation.



Spend most time doing

Collecting data (19%)

Cleaning and organizing data (60%)



Find least enjoyable

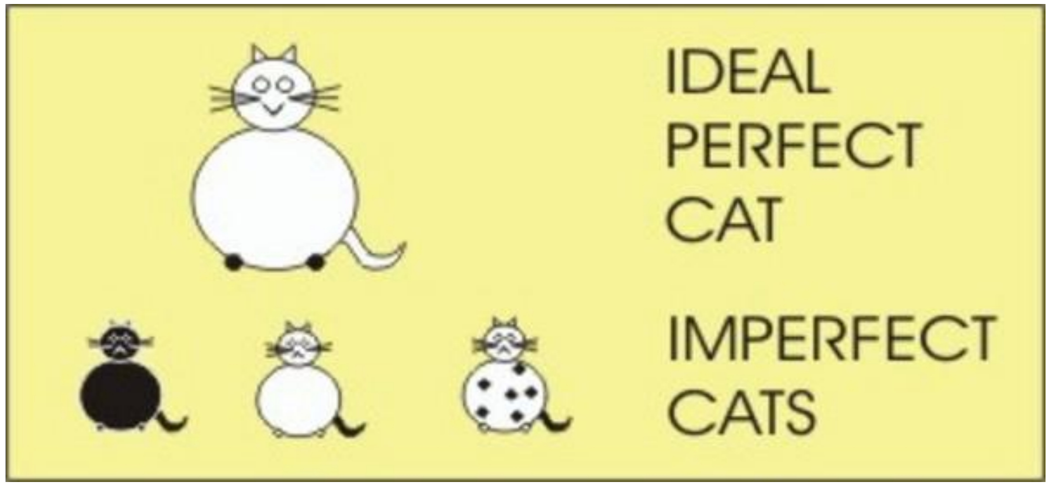
Collecting data (21%)

Cleaning and organizing data (57%)

data profiling

r/ai center for responsible ai

DB (databases) vs DS (data science)



https://midnightmediamusings.wordpress.com/2014/07/01/plato-and-the-theory-of-forms/

- **DB**: start with the schema, admit only data that fits; iterative refinement is possible, and common, but we are still schema-first
- **DS**: start with the data, figure out what schema it fits, or almost fits - reasons of usability, repurposing, low start-up cost

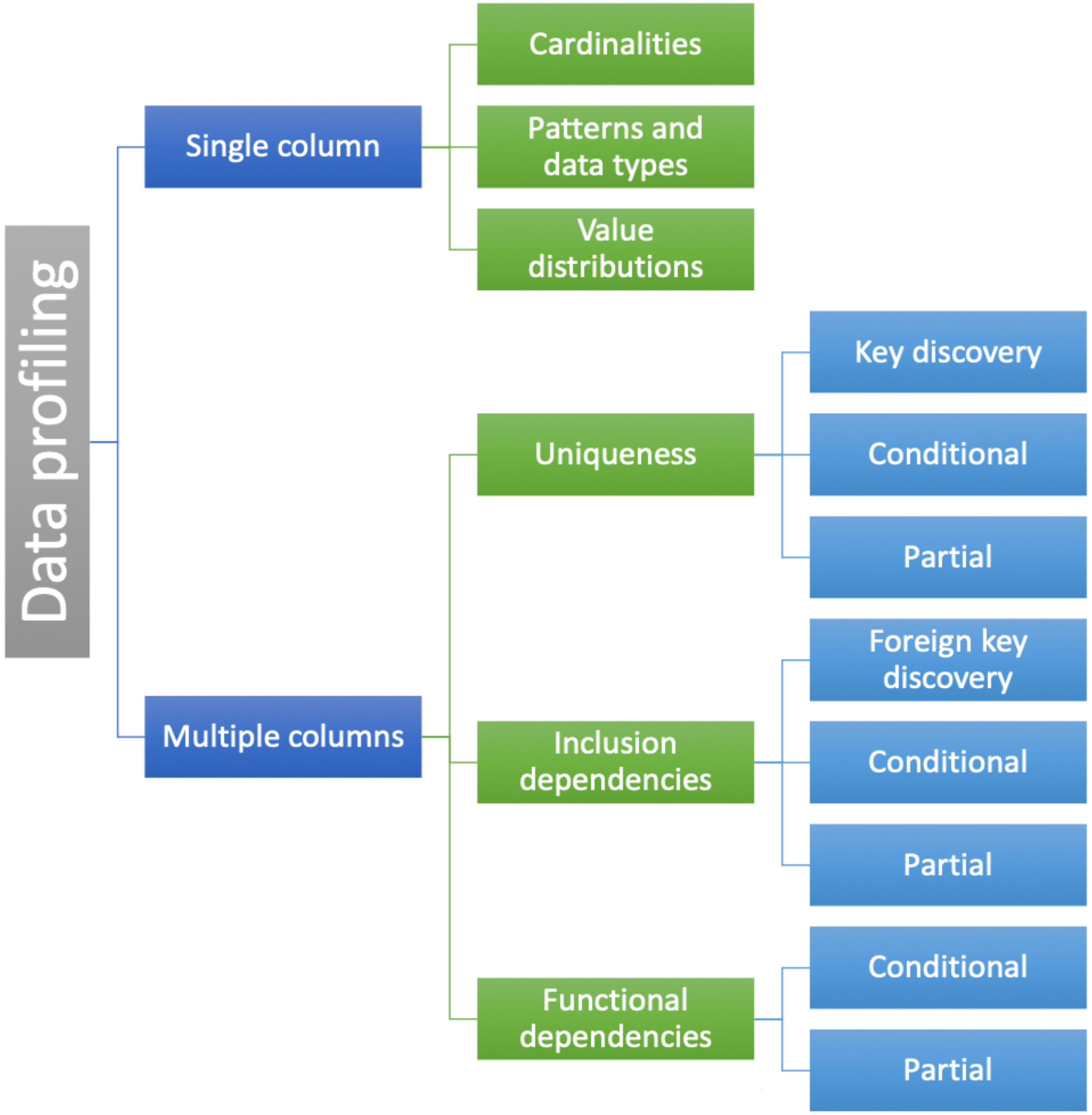
the "right" approach is somewhere between these two, **data profiling aims to bridge** between the two world views / methodologies

[Abedjan, Golab & Naumann (2017)]

Data profiling

	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	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

ational data (here: just one table)



[Abedjan, Golab & Naumann (2017)]

center for responsible ai

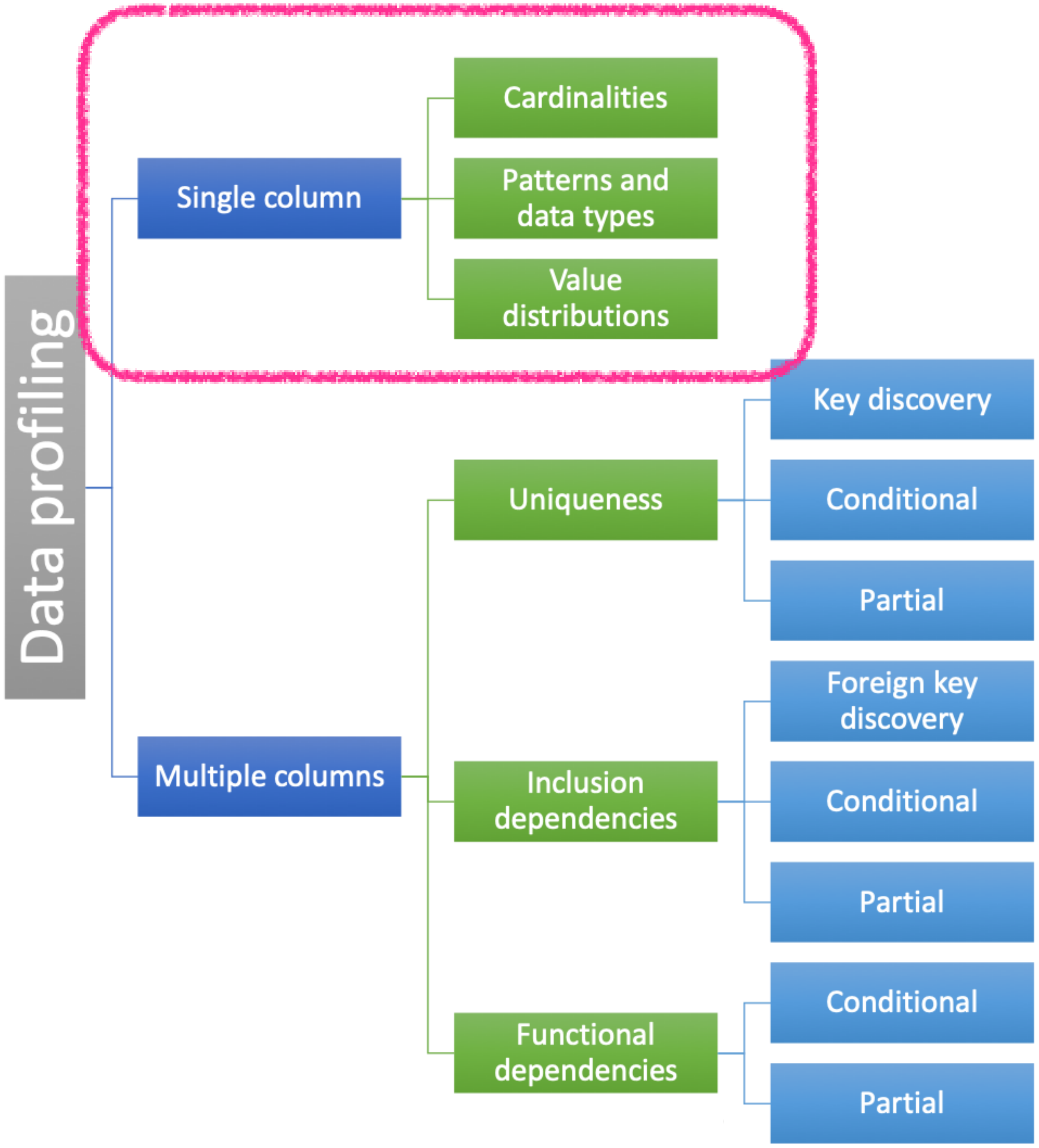
An alternative classification

- To help understand the **statistics**, we look at value ranges, data types, value distributions per column or across columns, etc
- To help understand the **structure** - the (business) rules that generated the data - we look at unique columns / column combinations, dependencies between columns, etc - **reverse-engineer the relational schema** of the data we have
- We need both statistics and structure, they are mutually-reinforcing, and help us understand the **semantics** of the data - it's meaning

Data profiling

	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	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

ational data (here: just one table)



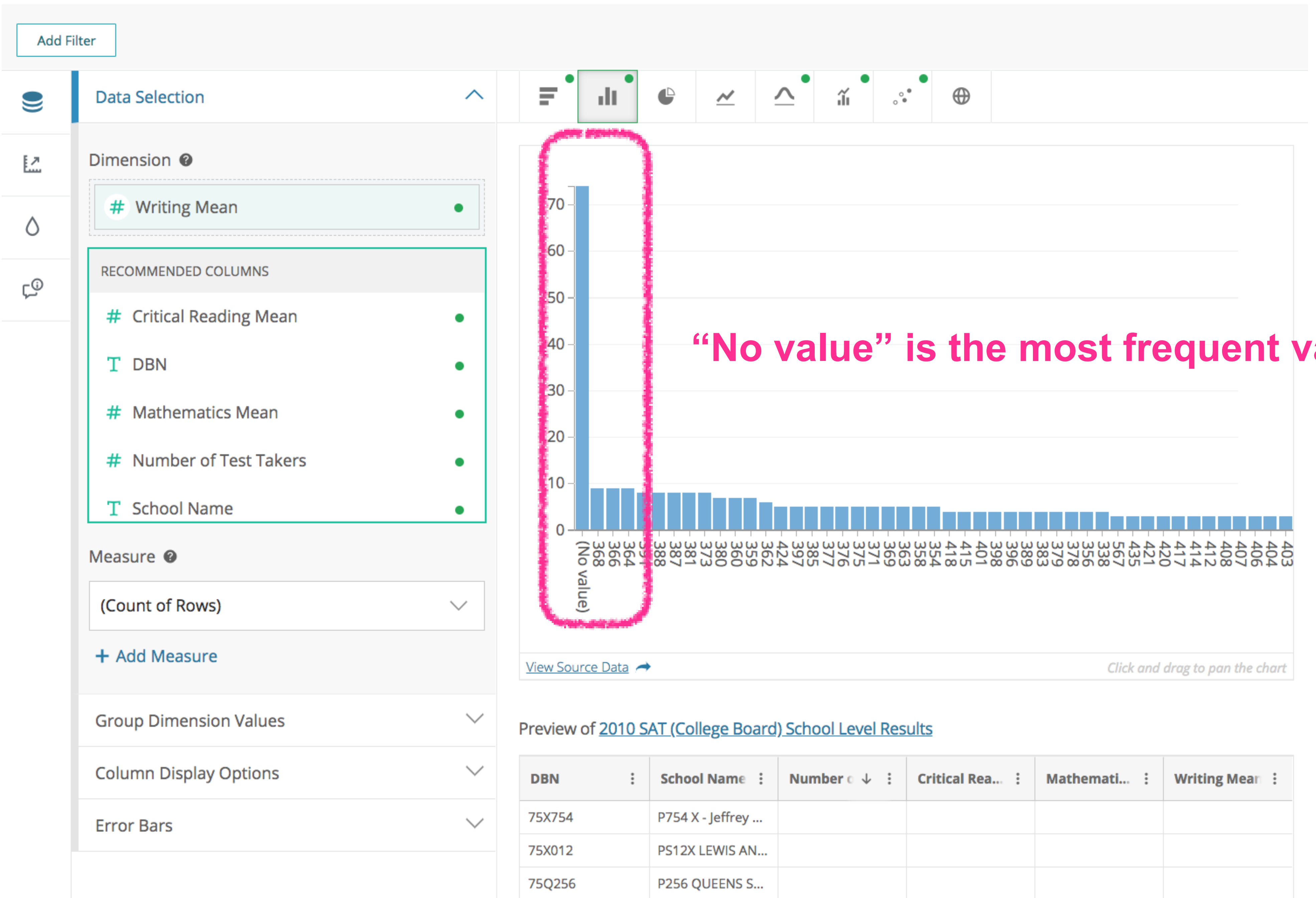
[Abedjan, Golab & Naumann (2017)]

Single column: cardinalities, data types

- cardinality of relation **R** - number of rows
- domain cardinality of a column **R.a** - number of **distinct** values
- attribute value **length**: min, max, average, median
- **basic data type**: string, numeric, date, time, ….
- number of percentage of **null** values of a given attribute
- regular expressions
- semantic domain: SSN, phone number
- ….

[Abedjan, Golab & Naumann (2017)]

NYC Open Data



"No value" is the most frequent value

https://opendata.cityofnewyork.us/

center for responsible ai

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

r/ai center for responsible ai

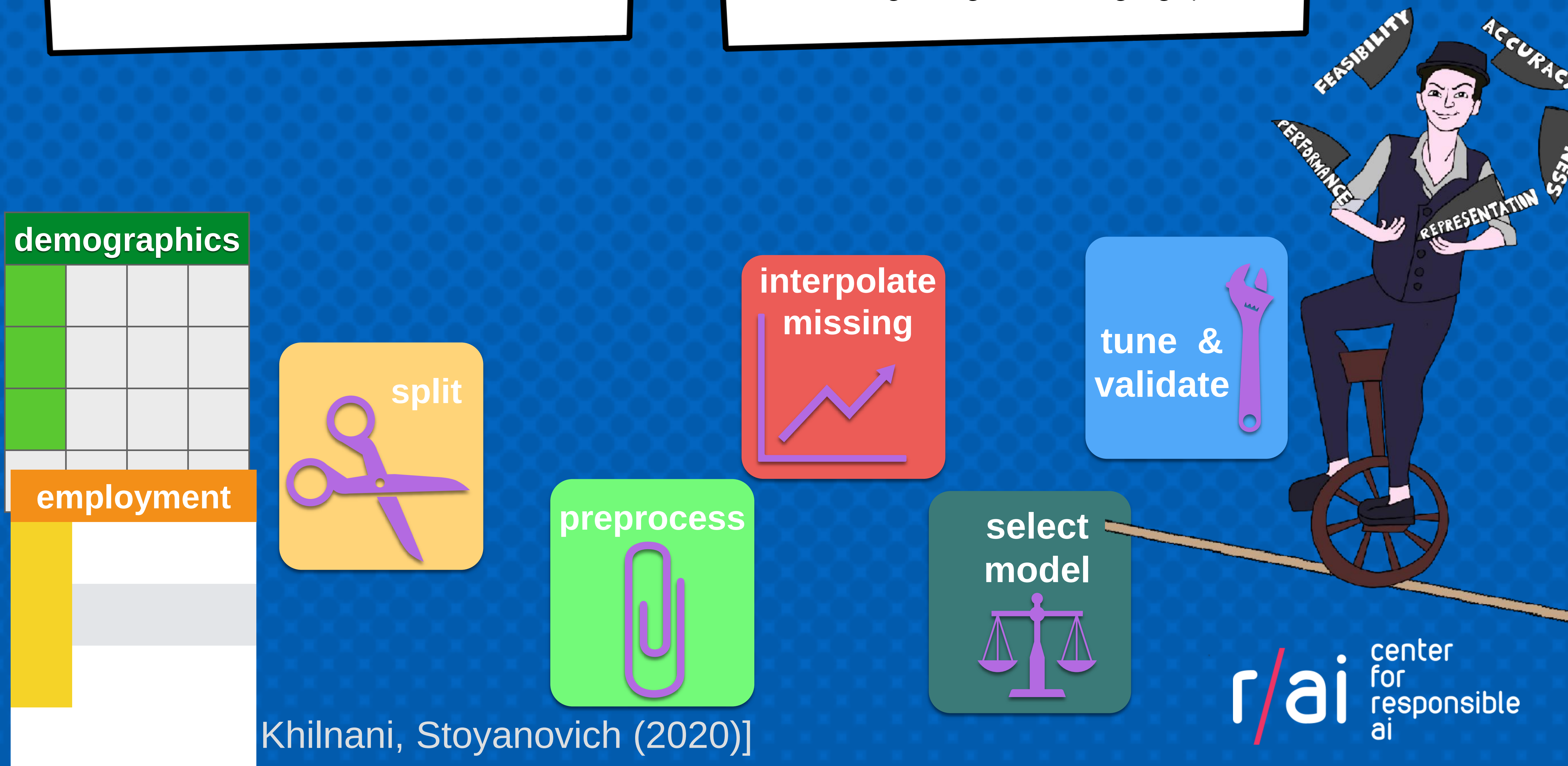
Model development lifecycle

Goal

design a model to predict an appropriate level of compensation for job applicants

Problem

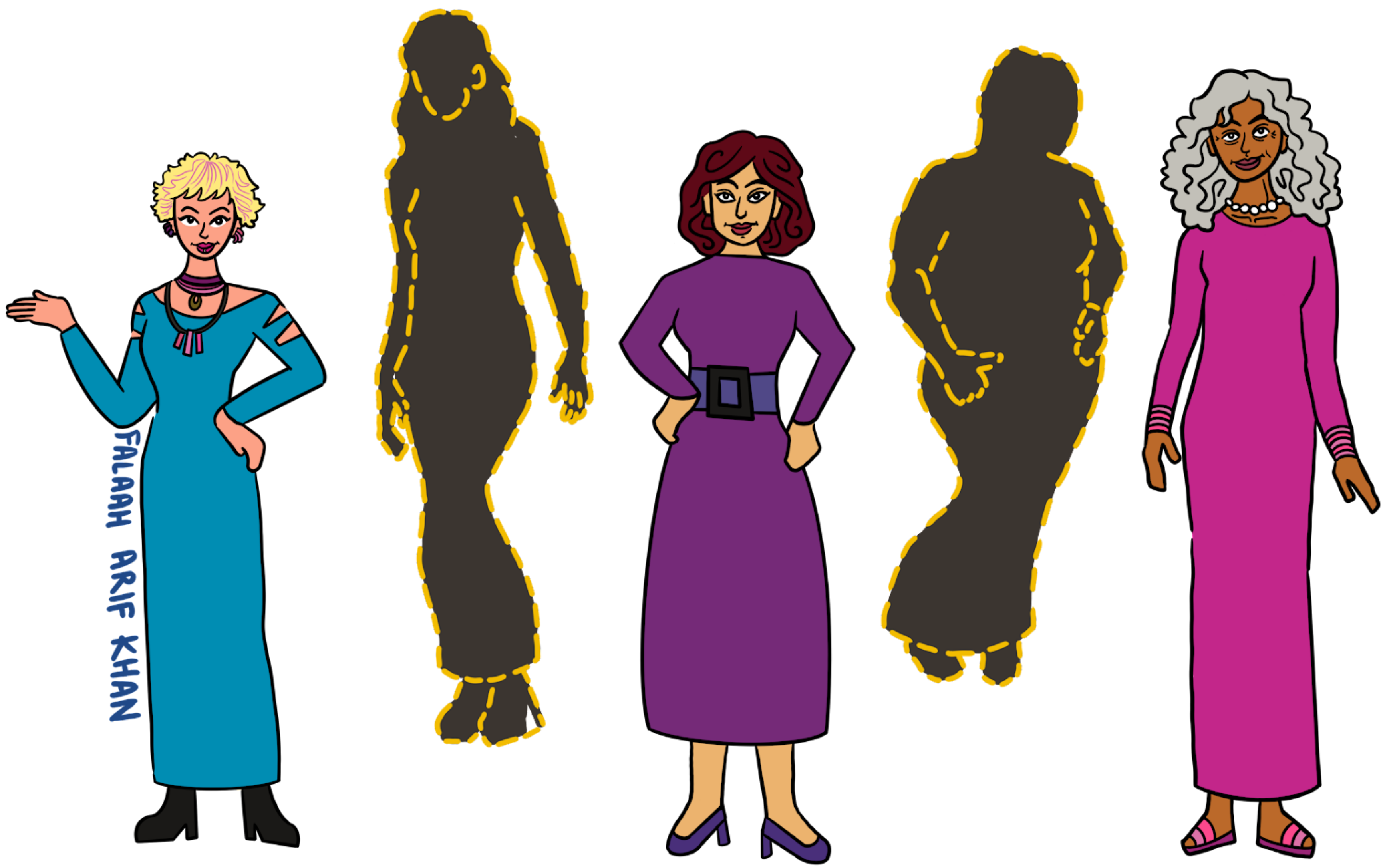
women are offered a lower salary than they would expect, potentially reinforcing the gender wage gap



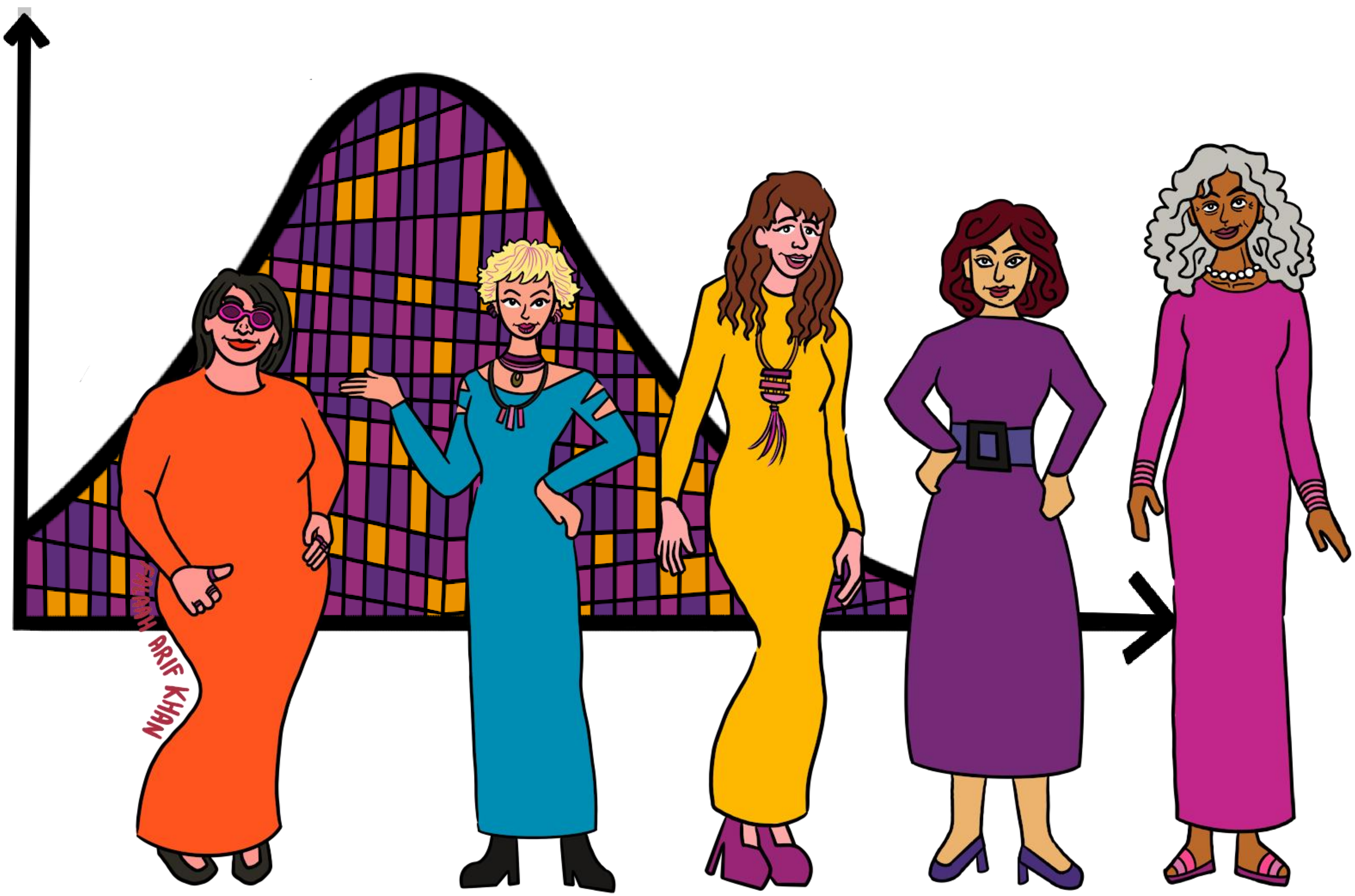
Khilnani, Stoyanovich (2020)]

center for responsible ai

Missing values: Observed data



Missing values: Imputed distribution



r/ai center for responsible ai

Missing values: True distribution

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)?

center for responsible ai

50 shades of *null*… and it gets worse

- **Hidden missing values** -
  - 99999 for zip code, Alabama for state
  - need data cleaning….
- Potential explanation for the "150 year-olds" receiving Social Security?
  - Social security uses and old version of COBOL that bases dates counting from

how do we detect hidden missing values?

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%

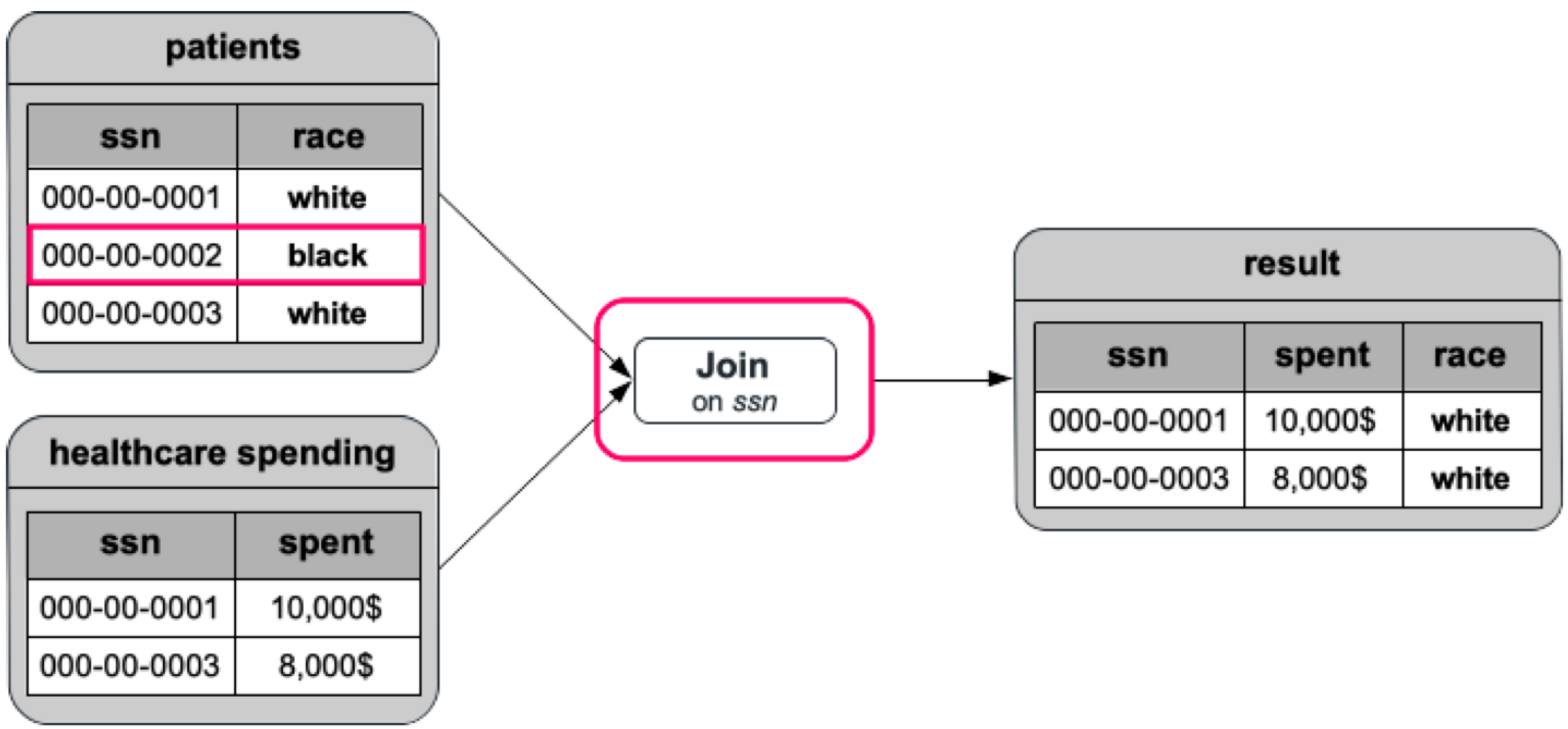
age_group	county
60	CountyA
60	CountyA
20	CountyA

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?



center for responsible ai

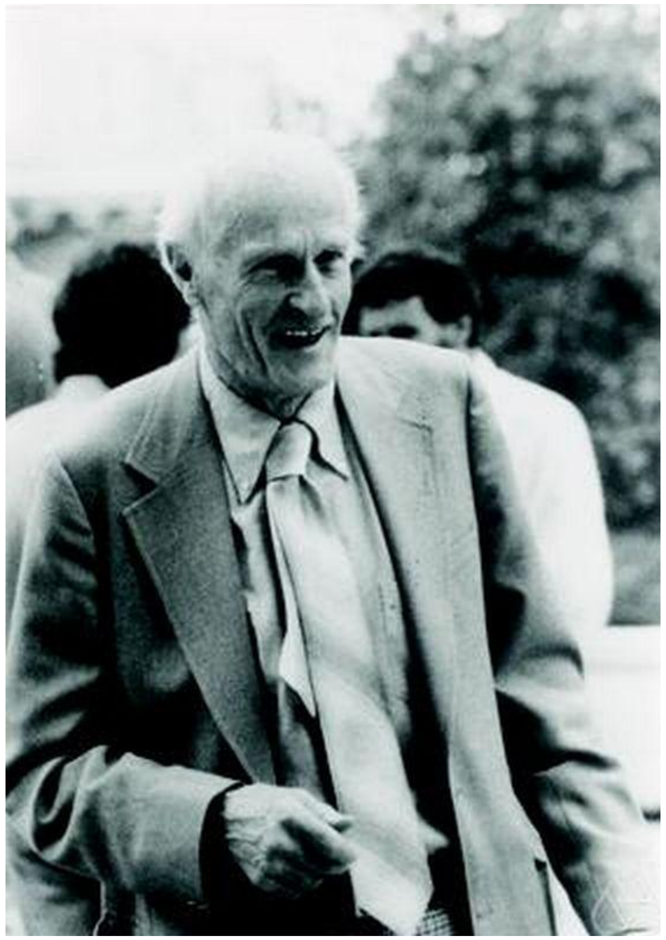
Single column: cardinalities, data types

- cardinality of relation **R** - number of rows
- domain cardinality of a column **R.a** - number of **distinct** values
- attribute value **length**: min, max, average, median
- **basic data type**: string, numeric, date, time, ….
- number of percentage of **null** values of a given attribute
- **regular expressions**
- semantic domain: SSN, phone number
- ….

[Abedjan, Golab & Naumann (2015)]

Regular expressions

- some attributes will have values that follow a regular format, e.g, telephone numbers: 212-864-0355 or (212) 864-0355 or 1.212.864-0355
- we may want to identify a small set of **regular expressions** that match all (or most) values in a column
- challenging - **very many possibilities!**



Stephen Kleene

A **regular expression**, **regex** or **regexp** … is a sequence of characters that define a search pattern. Usually this pattern is used by string searching algorithms for “find” or “find and replace” operations on strings, or for input validation. It is a technique that developed in theoretical computer science and formal language theory.

https://en.wikipedia.org/wiki/Regular_expression

Inferring regular expressions

- we may want to identify a small set of **regular expressions** that match all (or most) values in a column
- challenging - **very many possibilities!**

telephone
(201) 368-1000
(201)373-9599
(718) 206-1088
(718) 206-1121
(718) 206-1420
(718) 206-4420
(718) 206-4481
(718) 262-9072
(718) 868-2300
(888) 8NYC-TRS
800-624-4143

Example Regular Expression Language

. Matches any character
abc Sequence of characters
[abc] Matches any of the characters inside **[ ]**
***** Previous character matched zero or more times
? Previous character matched zero or one time
{m} Exactly **m** repetitions of previous character
^ Matches beginning of a line
$ Matches end of a line
\d Matches any decimal digit
\s Matches any whitespace character
\w Matches any alphanumeric character

Ockham's razor

Lex parsimoniae

If multiple hypotheses explain an observation, the simplest one should be preferred.

Ockham's motivation: can one prove the existence of God?

Used as a heuristic to help identify a promising hypothesis to test

Many applications today: biology, probability theory, ethics - also good for inferring regular expressions :)

William of Ockham (1285-1347)

Inferring regular expressions

telephone
800-624-4143
(201)373-9599
(201) 368-1000
(718) 206-1088
(718) 206-1121
(718) 206-1420
(718) 206-4420
(718) 206-4481
(718) 262-9072
(718) 868-2300
(888) 8NYC-TRS

Simple Algorithm

(1) Group values by length

(2) Find pattern for each group

- Ignore small groups
- Find **most specific character** at each position

(	2	0	1	)		3	6	8	-	1	0	0	0
(	2	0	1	)		2	0	6	-	1	0	8	8
(	7	1	8	)		2	0	6	-	1	1	2	1
(	7	1	8	)		2	0	6	-	1	4	2	0
(	7	1	8	)		2	0	6	-	4	4	2	0
(	7	1	8	)		2	0	6	-	4	4	8	1
(	7	1	8	)		2	6	2	-	9	0	7	2
(	7	1	8	)		8	6	8	-	2	3	0	0
(	7	1	8	)		2	0	6	-	0	5	4	5
(	8	1	4	)		6	8	1	-	6	2	0	0
(	8	8	8	)		8	N	Y	C	-	T	R	S

(	\d	\d	\d	)		\d	\w	\w	.	.	\w	\w	\w

Inferring regular expressions

telephone
800-624-4143
(201)373-9599
(201) 368-1000
(718) 206-1088
(718) 206-1121
(718) 206-1420
(718) 206-4420
(718) 206-4481
(718) 262-9072
(814) 681-6200
(888) 8NYC-TRS

Simple Algorithm

(1) Group values by length

(2) Find pattern for each group

- Ignore small groups
- Find **most specific character** at each position

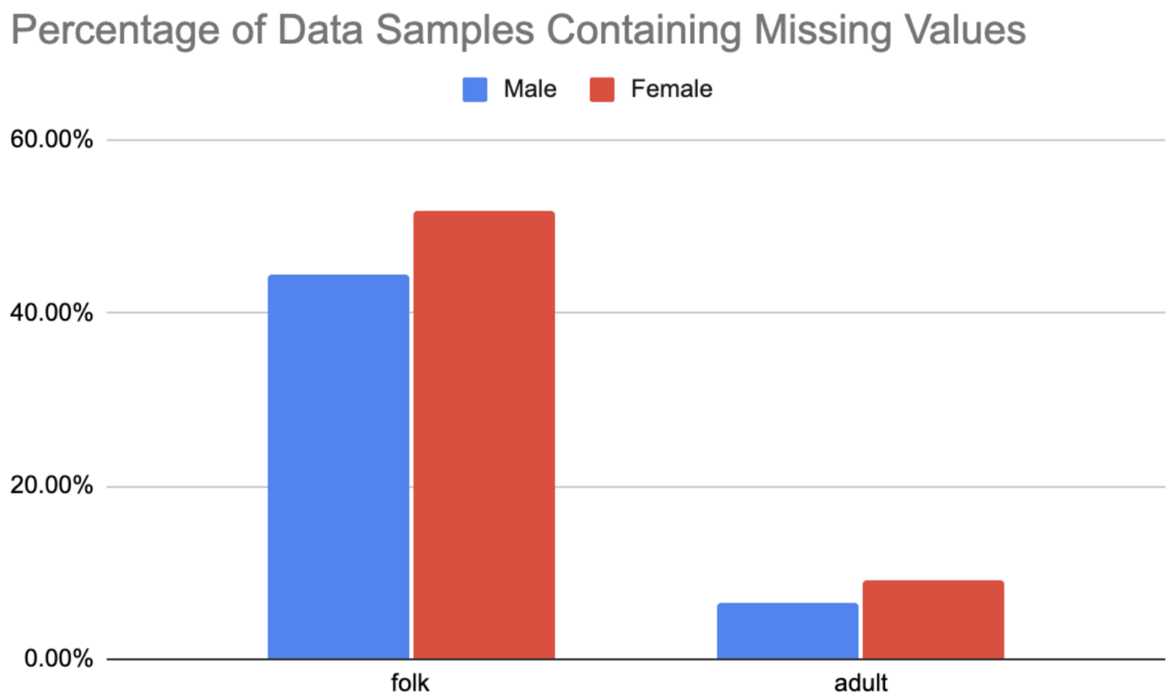
ignoring small groups: alternatives?

(	\d	\d	\d	)		\d	\w	\w	.	.	\w	\w	\w

```
(\d{3}) \d\w{2}.{2}\w{3}
```

Data quality and fairness

- poor-quality data can hurt ML model accuracy
- data from historically disadvantages groups may suffer from poorer quality
- systematic differences in data quality may hurt performance of predictors - a fairness concern
- **RQ1**: Does the incidence of data errors track demographic group membership in ML fairness datasets?



[Guha, Arif Khan, Stoyanovich, Schelter (2023)]

Single column: basic stats, distributions

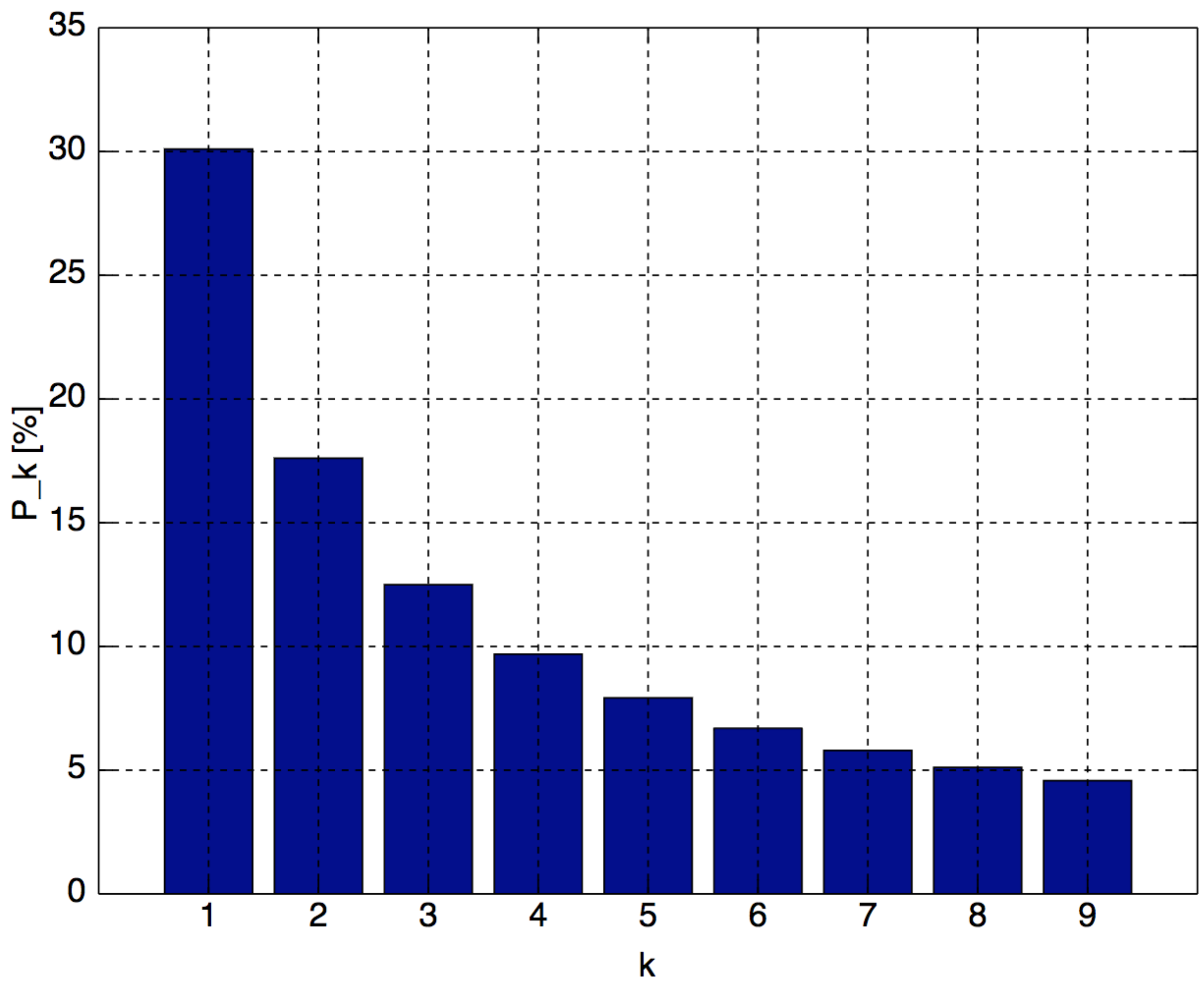
- min, max, **average**, median value of **R.a**
- **histogram**
  - equi-width - (approximately) the same number of distinct values in each bucket (e.g., age broken down into 5-year windows)
  - equi-depth (approximately) the same number of tuples in each bucket
  - biased histograms use different granularities for different parts of the value range to provide better accuracy
- quartiles - three points that divide the numeric values into four equal groups - a kind of an equi-depth histogram
- **first digit** - distribution of first digit in numeric values, to check Benford law
- …

[Abedjan, Golab & Naumann (2015)]

Benford Law

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

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



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

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

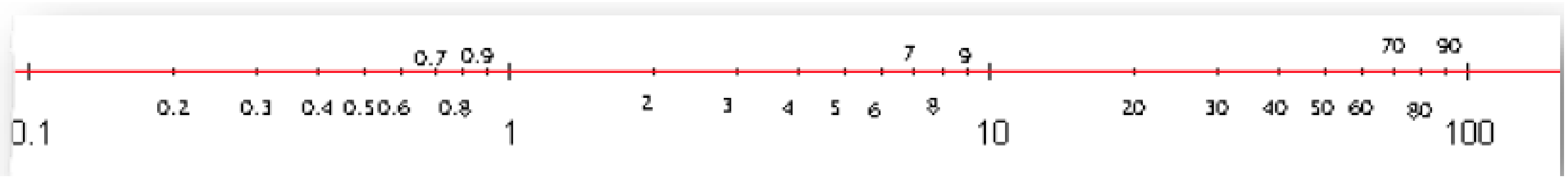
[Benford: "The law of anomalous numbers" Proc. Am. Philos. Soc., 1938]

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.

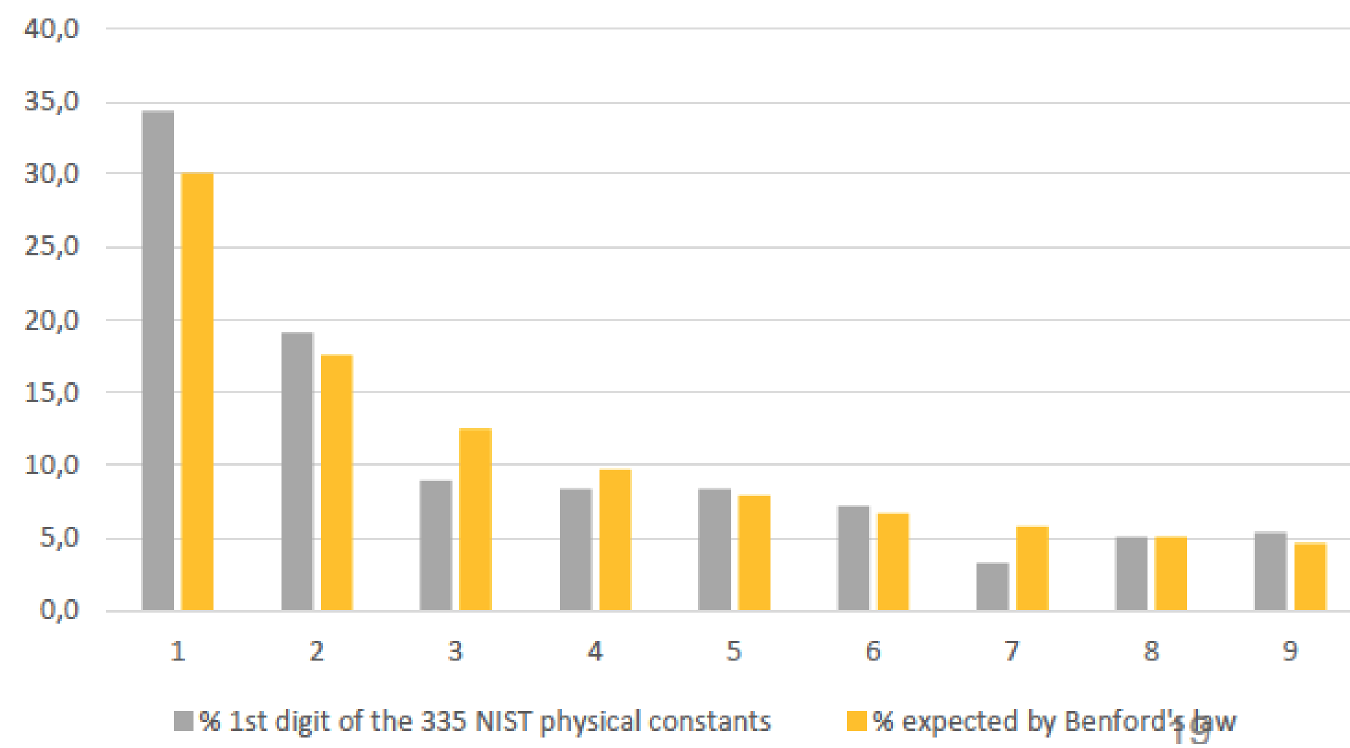
https://en.wikipedia.org/wiki/Benford%27s_law

[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!

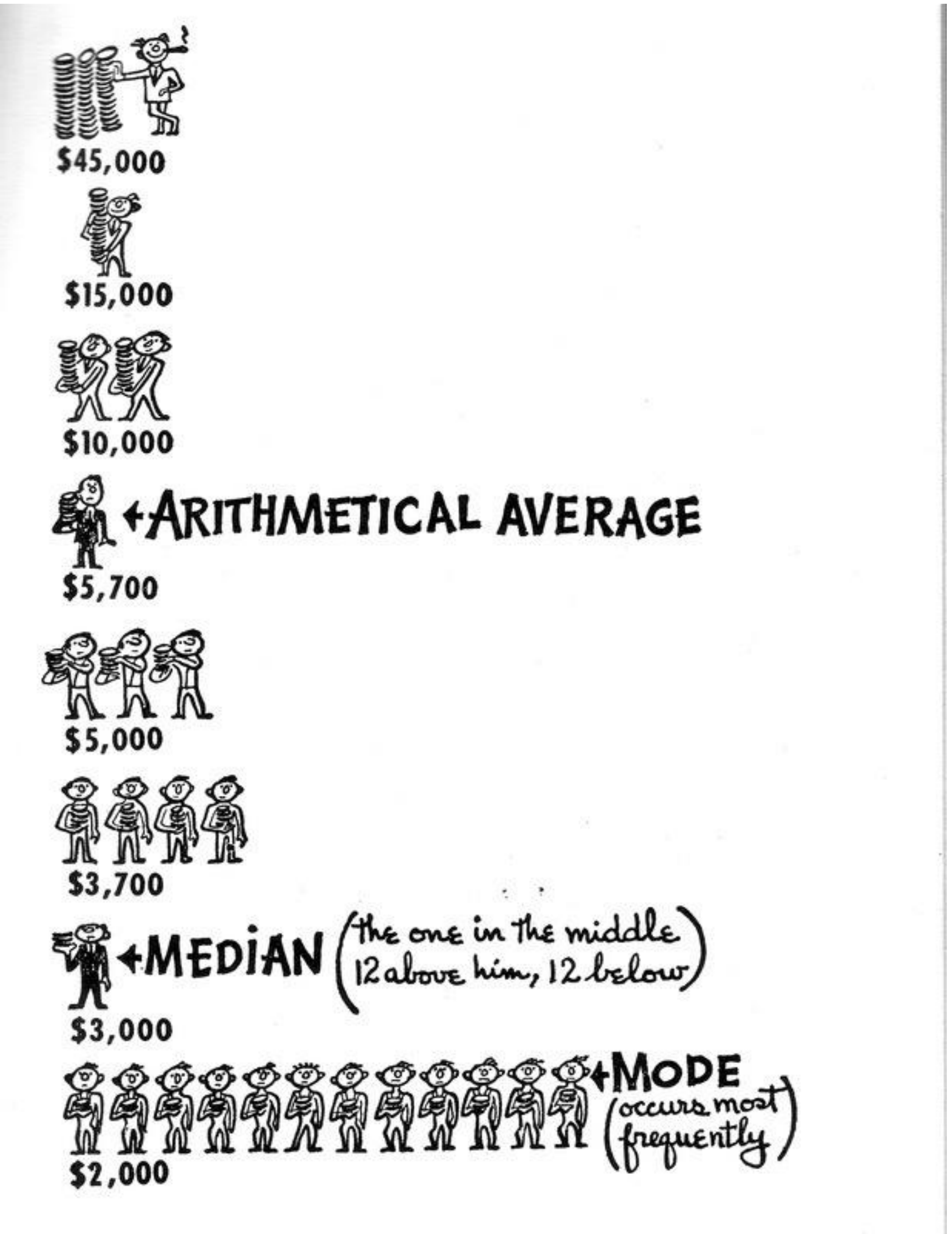


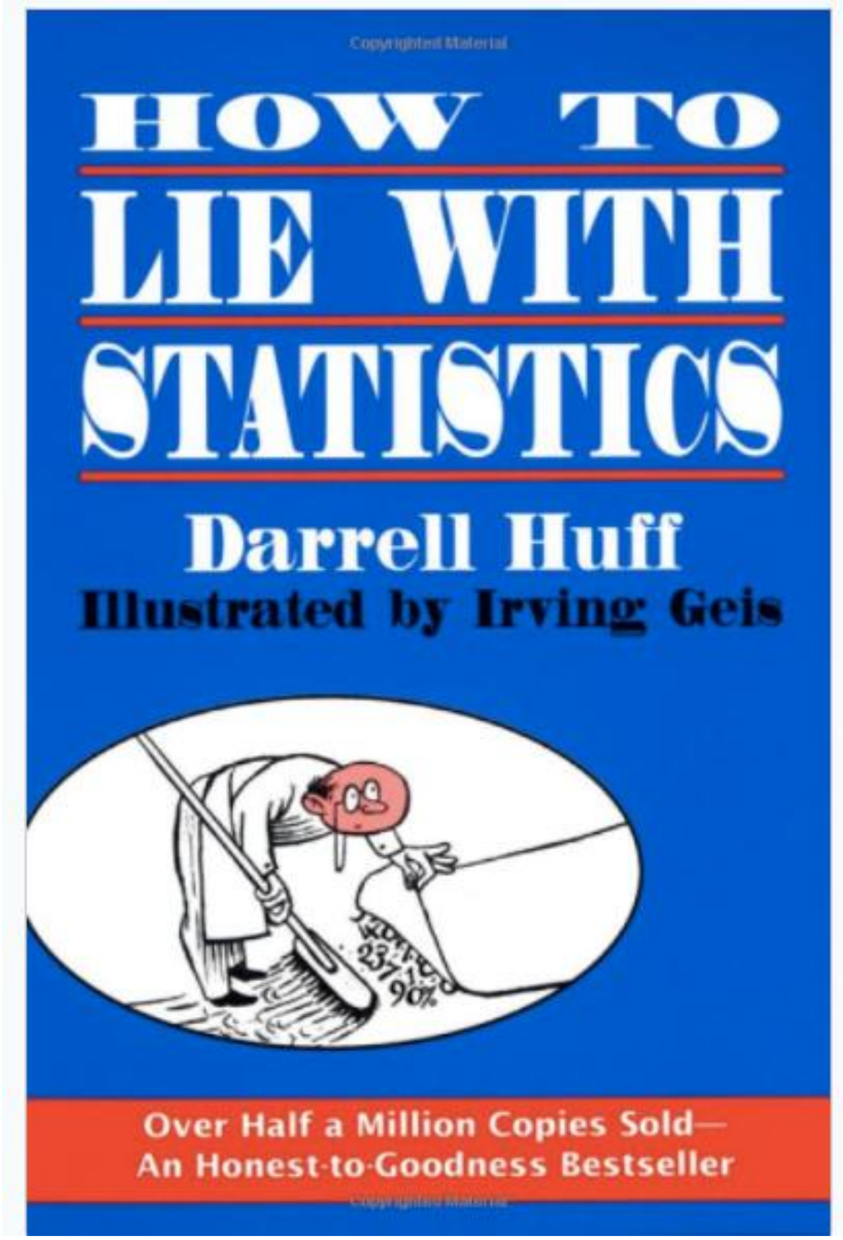
physical constants

[Abedjan, Golab & Naumann (2015)]

center for responsible ai

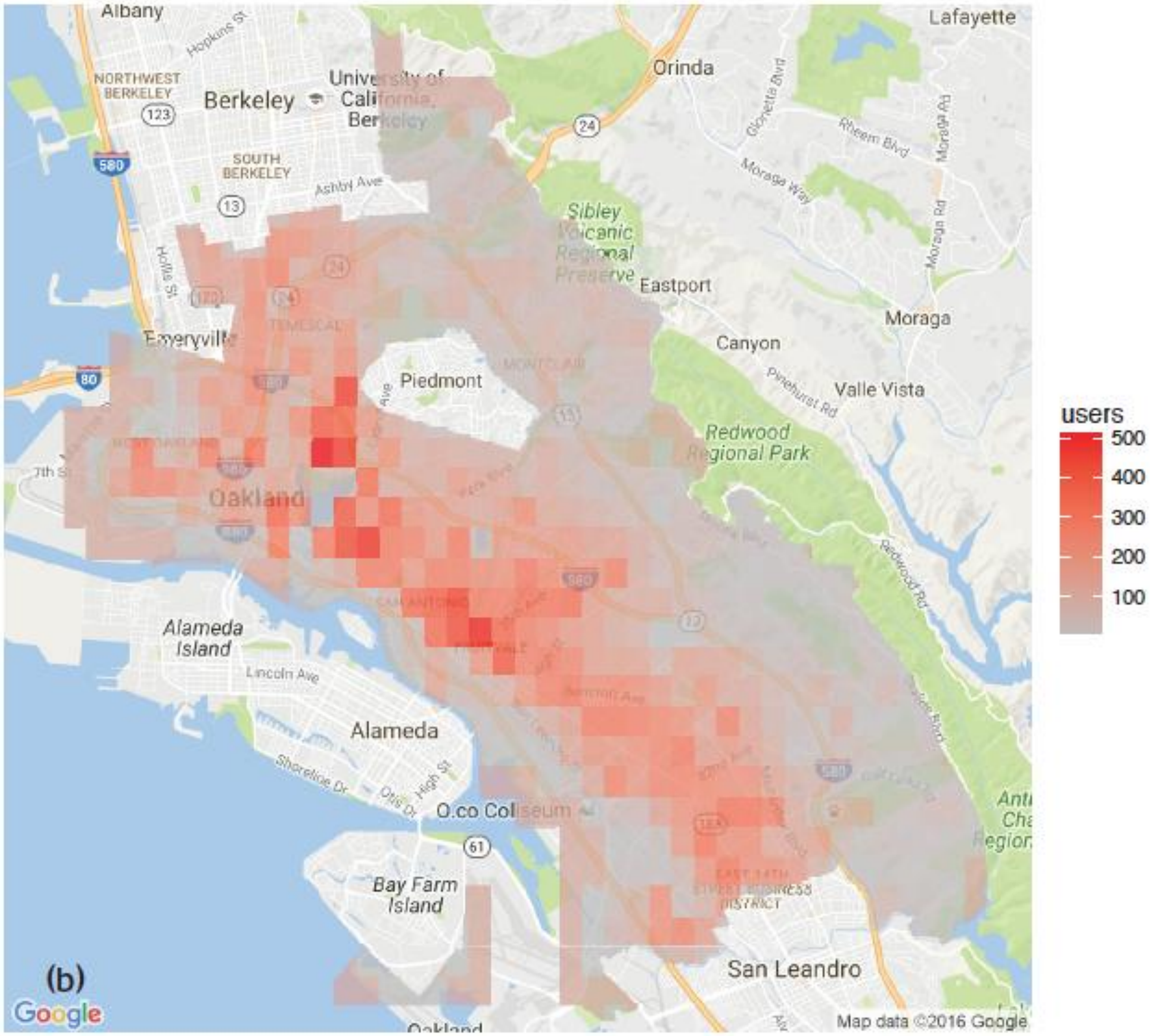
The well-chosen average



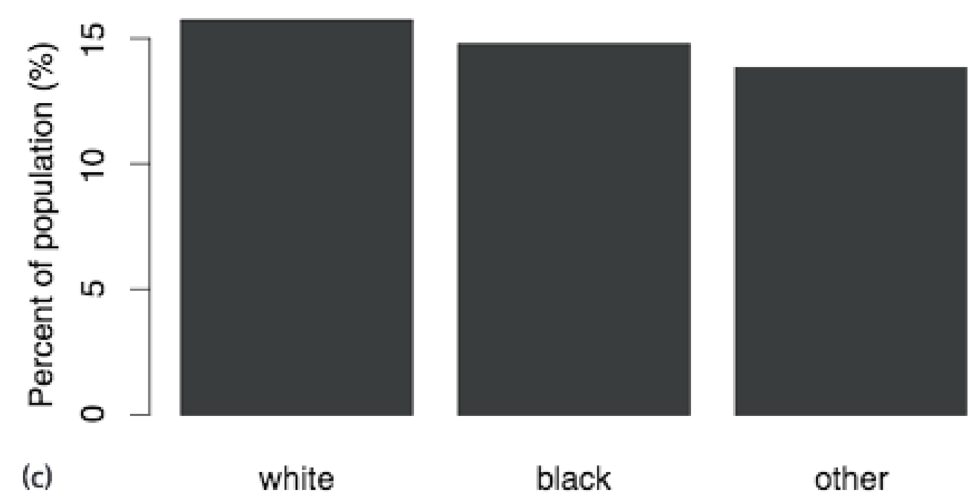


r/ai center for responsible ai

Is my data biased? (histograms + geo)

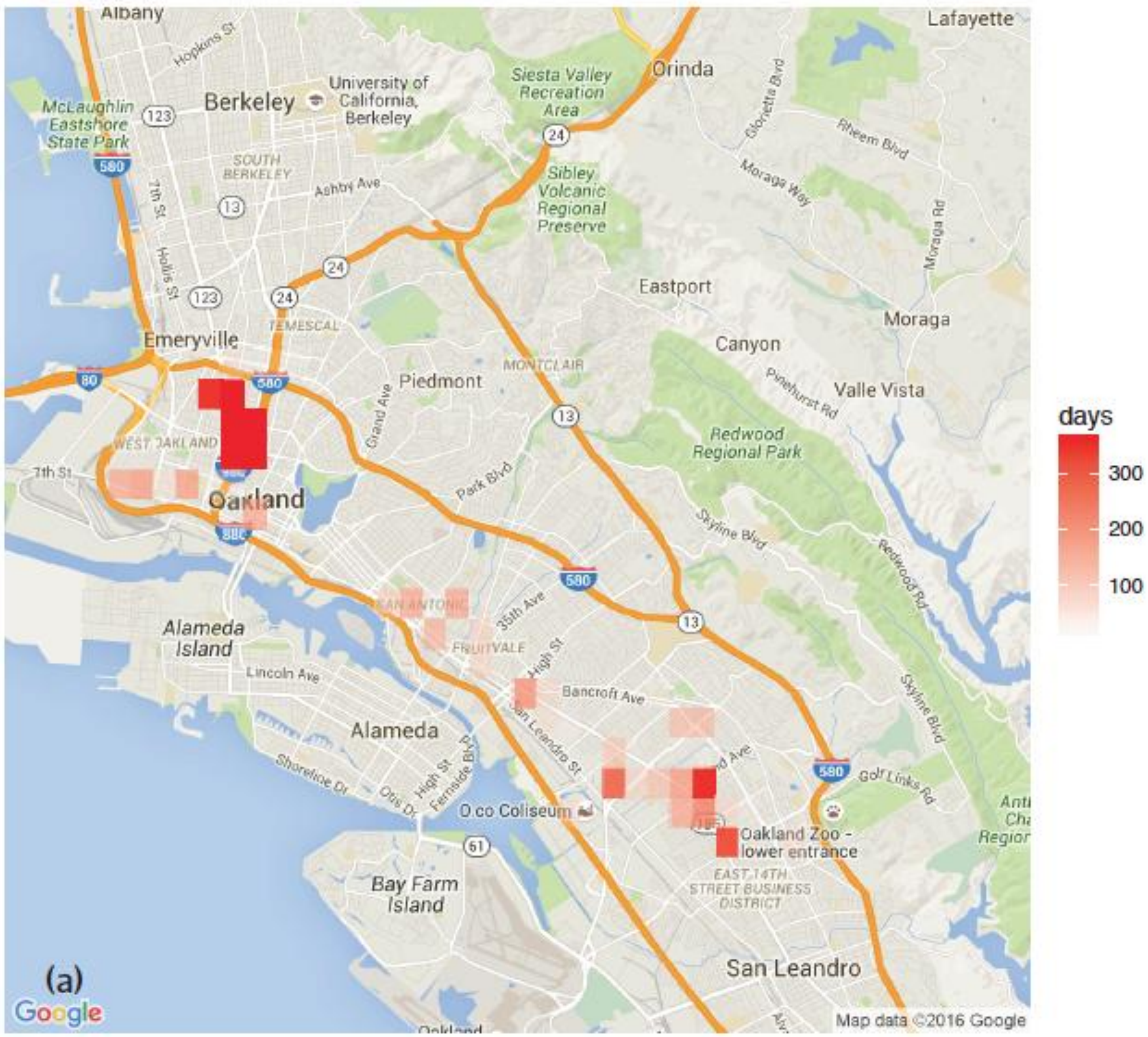


Estimated number of drug users, based on 2011 National Survey on Drug Use and Health, in Oakland, CA

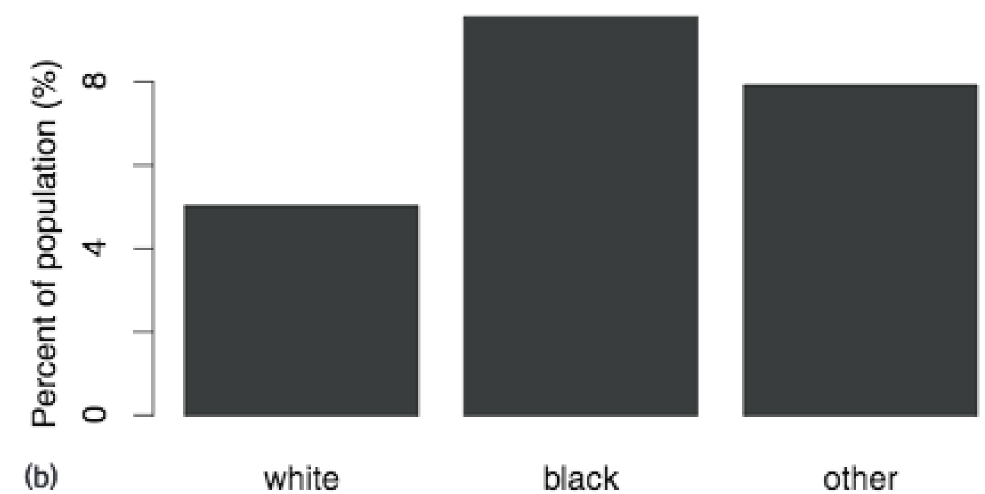


Estimated drug use by race

Is my data biased? (histograms + geo)

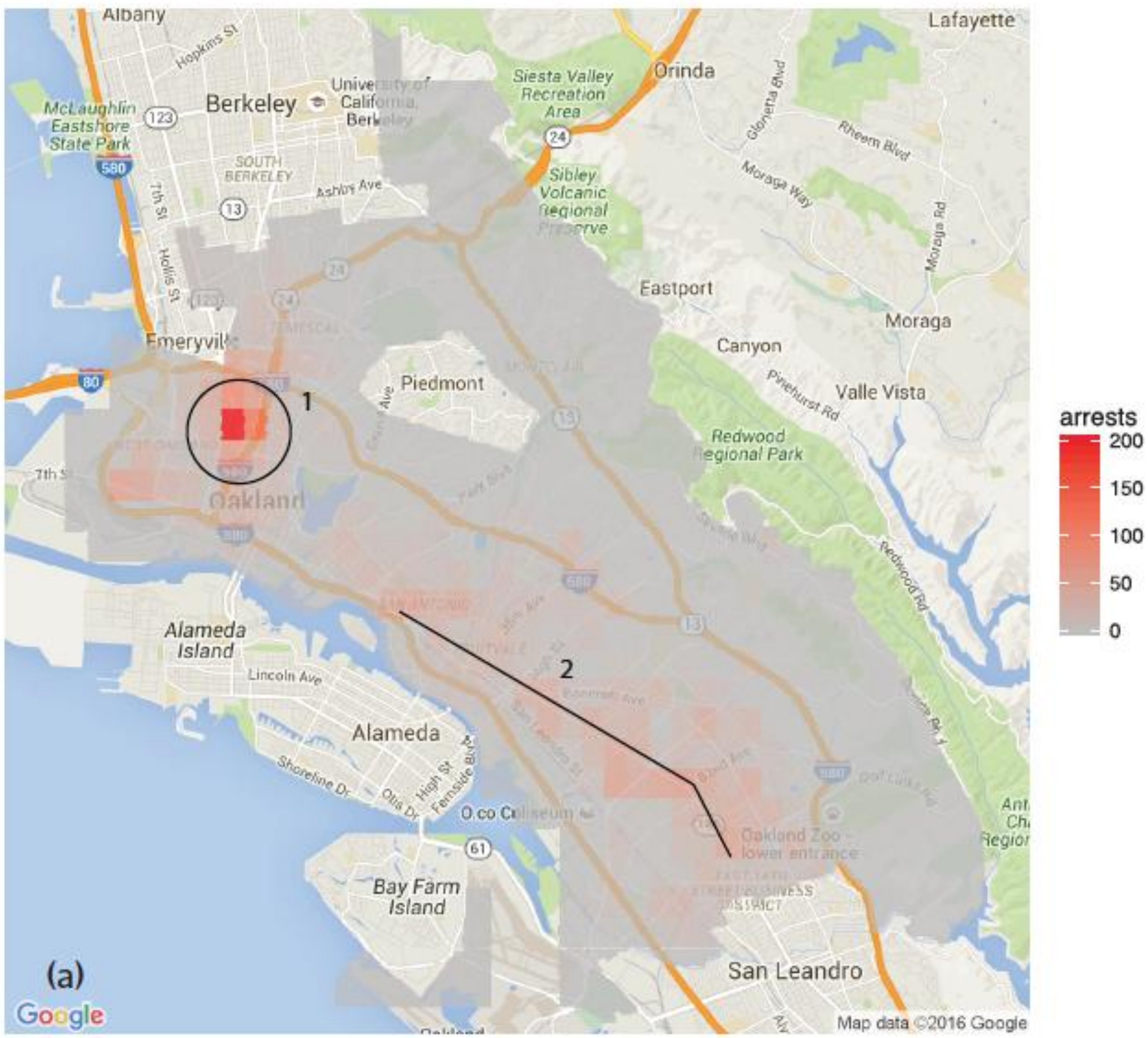


Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland, CA, police data for 2011

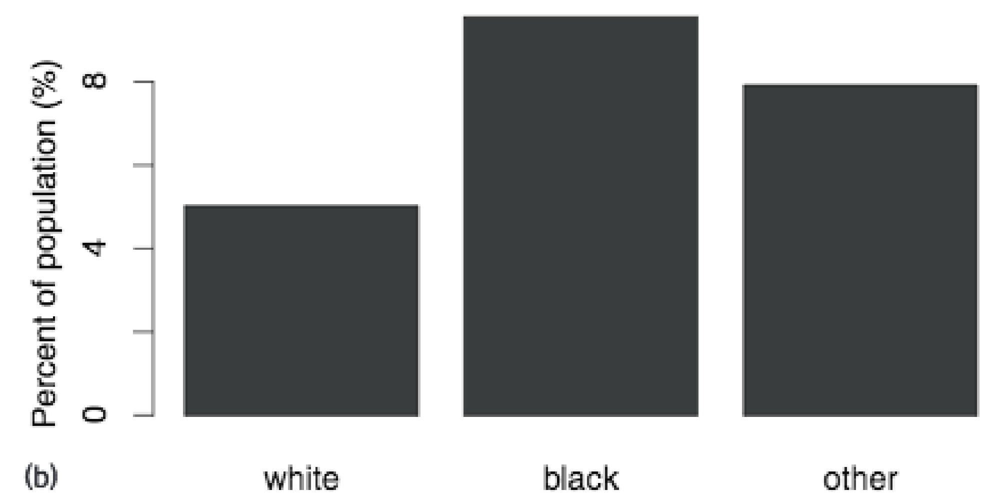


Targeted policing for drug crimes by race

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

Responsible Data Science

The data science lifecycle

Thank you!