Responsible Data Science

The data science lifecycle

February 22 & March 1, 2022

Prof. Julia Stoyanovich

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

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

This week's reading

Responsible Data Management

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

H.V. Jagadish
University of Michigan
Ann Arbor, MI, USA
jag@umich.edu

ABSTRACT

The need for responsible data management intensifies with the growing impact of data on society. One central locus of the societal impact of data are Automated Decision Systems (ADS), socio-legal-technical systems that are used broadly in industry, non-profits, and government. ADS process data about people, help make decisions that are consequential to people's lives, are designed with the stated goals of improving efficiency and promoting equitable access to opportunity, involve a combination of human and automated decision making, and are subject to auditing for legal compliance and to public disclosure. They may or may not use AI, and may or may not operate with a high degree of autonomy, but they rely heavily on data.

In this article, we argue that the data management community is uniquely positioned to lead the responsible design, development, use, and oversight of ADS. We outline a technical research agenda that requires that we step outside our comfort zone of engineering for efficiency and accuracy, to also incorporate reasoning about values and beliefs. This seems high-risk, but one of the upsides is being able to explain to our children what we do and why it matters.

PVLDB Reference Format:
Julia Stoyanovich, Bill Howe, H.V. Jagadish. Responsible Data Management. *PVLDB*, 13(12): 3474 - 3488, 2020.
DOI: https://doi.org/10.14778/3415478.3415570

1. INTRODUCTION

We are in the midst of a global trend to regulate algorithms, artificial intelligence, and automated decision systems. This flurry of activity hardly comes as a surprise. As reported by the recent One Hundred Year Study on Artificial Intelligence [58]: "AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human aware and trustworthy." In the European Union, the General Data Protection Regulation (GDPR) [66] offers protections to individuals regarding the collection, processing, and movement of their personal data, and applies broadly to the use of such data by governments and private-sector entities. Regulatory activity in several countries outside of the EU, notably Japan [46] and Brazil [42], is in close alignment with the GDPR.

In the US, many major cities, a handful of states, and even the Federal government are establishing task forces and issuing guidelines about responsible development and use of technology, often starting with its use in government itself—rather than in the private sector—where there is, at least in theory, less friction between organizational goals and societal values. Case in point: New York City rightfully prides itself on being a trendsetter—in architecture, fashion, the performing arts and, as of late, in its very publicly made commitment to opening the black box of the government's use of technology. In May 2018, an Automated Decision Systems (ADS) Task Force was convened, the first such in the nation, and charged with providing recommendations to New York City's agencies about becoming transparent and accountable in their use of ADS. The Task Force issued its report in November 2019, making a commitment to using ADS where they are beneficial, reducing potential harm across their lifespan, and promoting fairness, equity, accountability, and transparency in their use [5].

Can the principles of the responsible use of ADS — of *socio-legal-technical systems* that may or may not use AI, and may or may not operate with a high degree of autonomy, but that rely heavily on data — be operationalized as a matter of policy [2]? Can this be done in the face of a crisis of trust in government, which extends to the lack of trust in the government's ability to manage modern technology in the interest of the public [73]? What will it take to instill responsible ADS practices beyond government?

In this article, we hope to convince you that the data management community should play a central role in the responsible design, development, use, and oversight of ADS. By engaging in this work, we have a critical opportunity to help make society more equitable, inclusive, and just; make government operations more transparent and accountable; and encourage public participation in ADS design and oversight. To make progress, we may need to step outside our engineering comfort zone and start reasoning in terms of values and beliefs, in addition to checking results against known ground truths and optimizing for efficiency objectives. This seems high risk, but one of the upsides is being able to explain to our children what we do and why it matters.

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3415478.3415570

This week's reading

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-

Quantitative Data Cleaning for Large Databases

Joseph M. Hellerstein*
EECS Computer Science Division
UC Berkeley
http://db.cs.berkeley.edu/jmh

February 27, 2008

1 Introduction

Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission. Data analysis typically drives decision-making processes and efficiency optimizations, and in an increasing number of settings is the *raison d'etre* of entire agencies or firms.

Despite the importance of data collection and analysis, data *quality* remains a pervasive and thorny problem in almost every large organization. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decades on various aspects of *data cleaning*: computational procedures to automatically or semi-automatically identify – and, when possible, correct – errors in large data sets.

In this report, we survey data cleaning methods that focus on errors in *quantitative* attributes of large databases, though we also provide references to data cleaning methods for other types of attributes. The discussion is targeted at computer practitioners who manage large databases of quantitative information, and designers developing data entry and auditing tools for end users. Because of our focus on quantitative data, we take a statistical view of data quality, with an emphasis on intuitive outlier detection and exploratory data analysis methods based in *robust statistics* [Rousseeuw and Leroy, 1987, Hampel et al., 1986, Huber, 1981]. In addition, we stress algorithms and implementations that can be easily and efficiently implemented in very large databases, and which are easy to understand and visualize graphically. The discussion mixes statistical intuitions and methods, algorithmic building blocks, efficient relational database implementation strategies, and user interface considerations. Throughout the discussion, references are provided for deeper reading on all of these issues.

1.1 Sources of Error in Data

Before a data item ends up in a database, it typically passes through a number of steps involving both human interaction and computation. Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful both in designing data collection and curation techniques that mitigate

*This survey was written under contract to the United Nations Economic Commission for Europe (UNECE), which holds the copyright on this version.

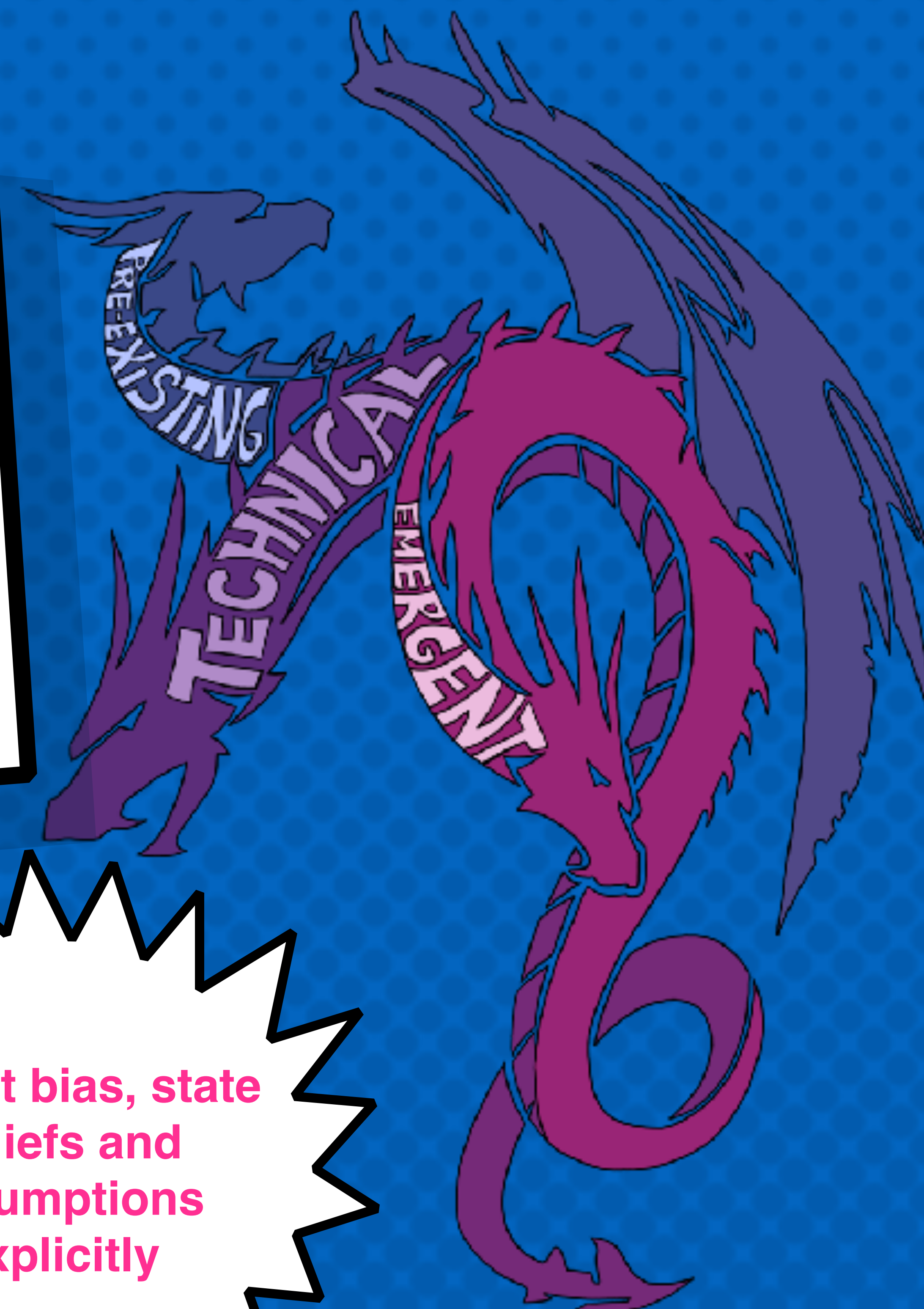
r/ai

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

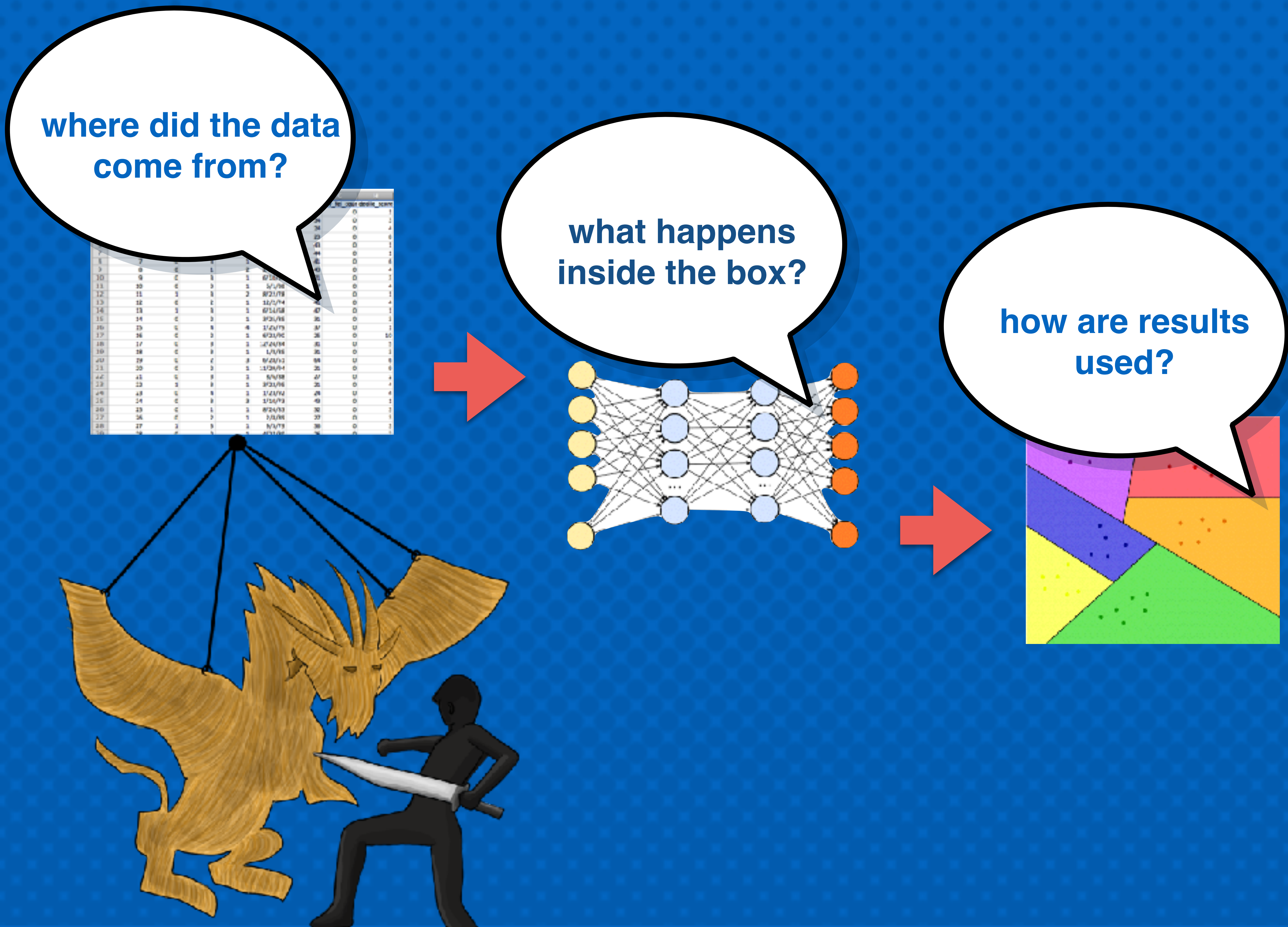
Emergent arises due to context of use



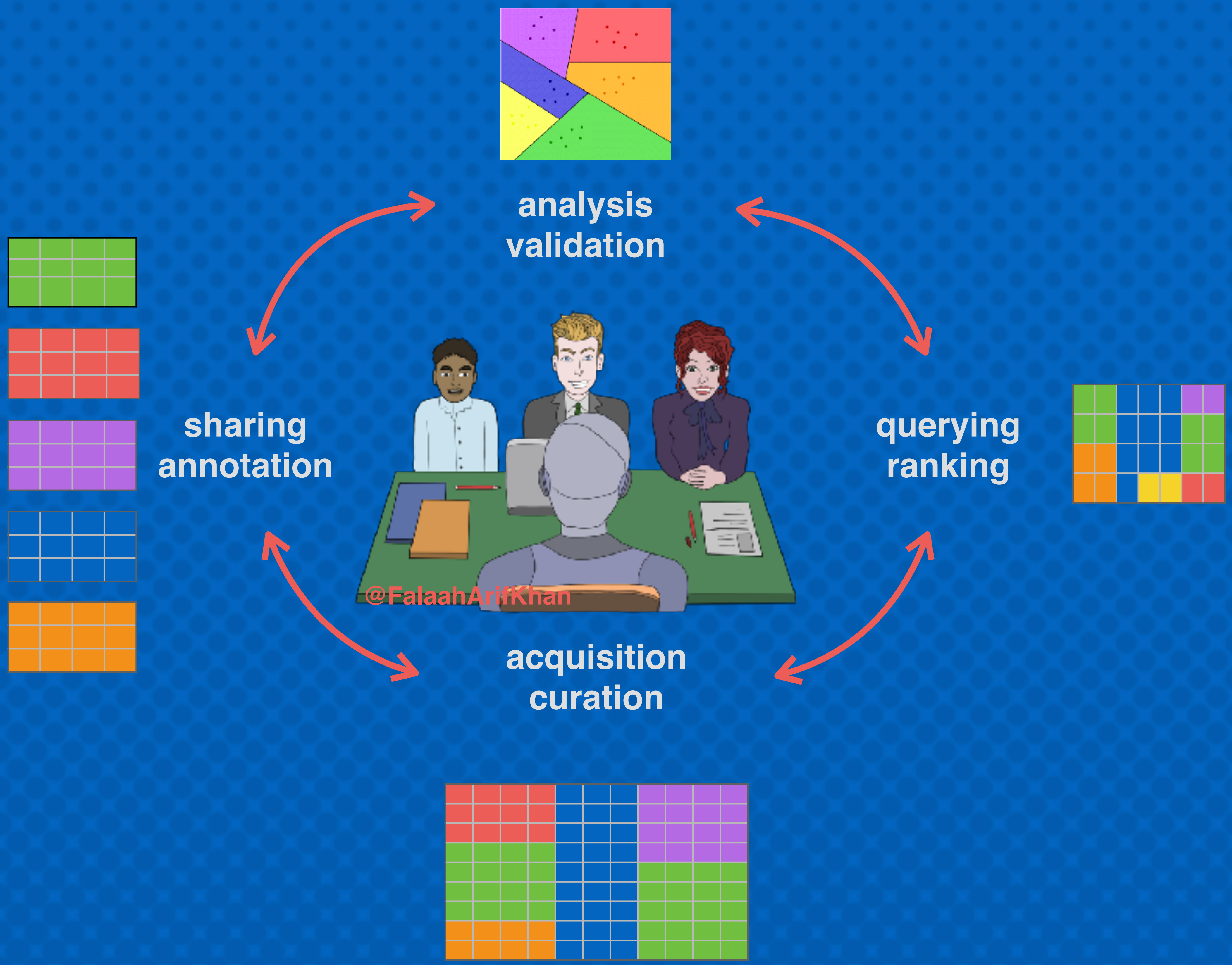
to fight bias, state beliefs and assumptions explicitly

[Friedman & Nissenbaum (1996)]

The “last-mile” view of responsible AI



Data lifecycle of an ADS



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/

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

Understand your data!

Need **metadata** to:

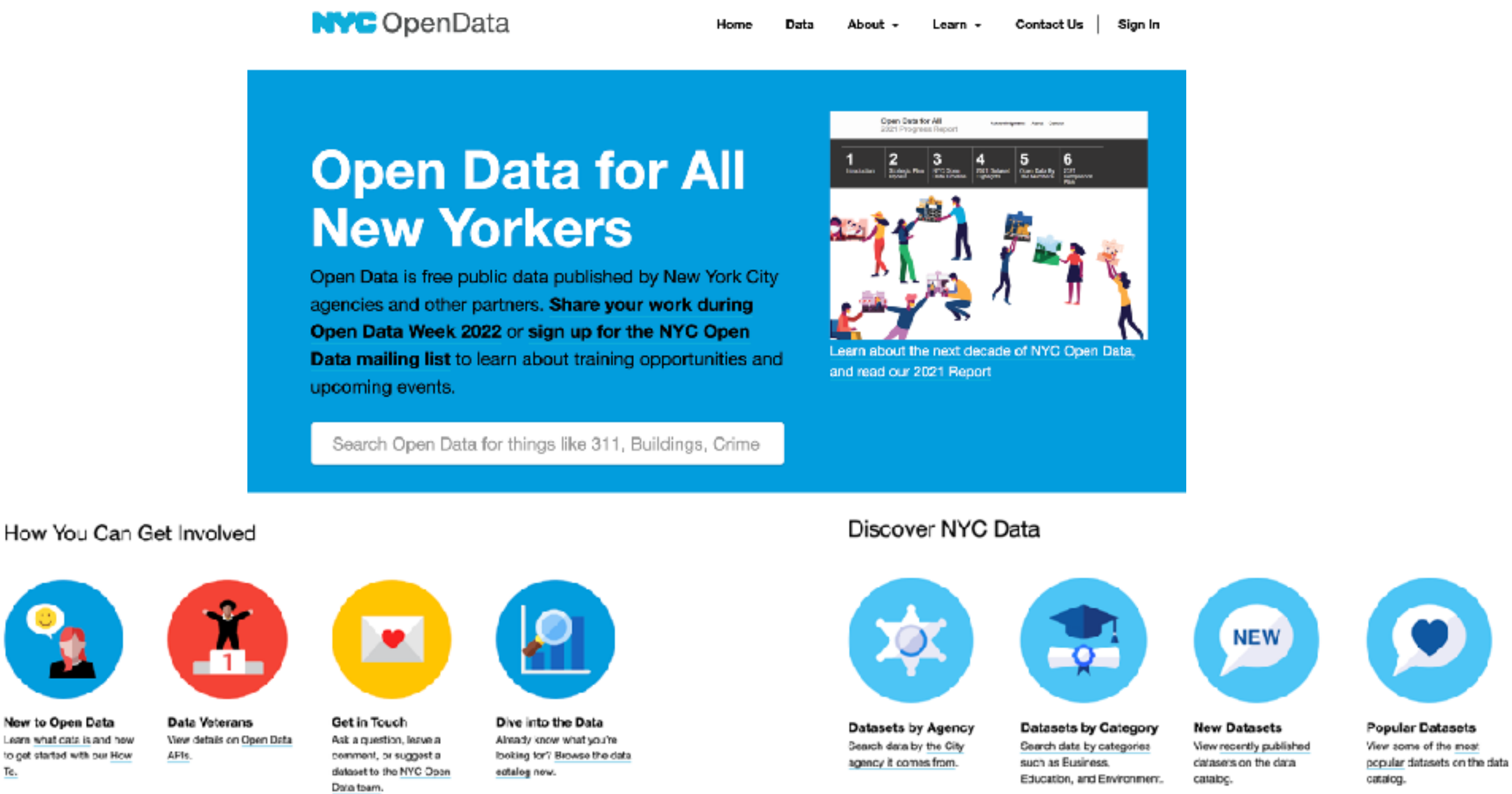
- enable data **re-use** (have to be able to find it!)
- determine **fitness for use** of a dataset in a task
- help establish **trust** in the data analysis process and its outcomes

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]

r/ai

NYC Open Data



https://opendata.cityofnewyork.us/

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

Updated
April 25, 2019

Views
28,463

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

privacy

popularity

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

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/

r/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 #

NYC Open Data

What's in this Dataset?

Rows	Columns
460	6

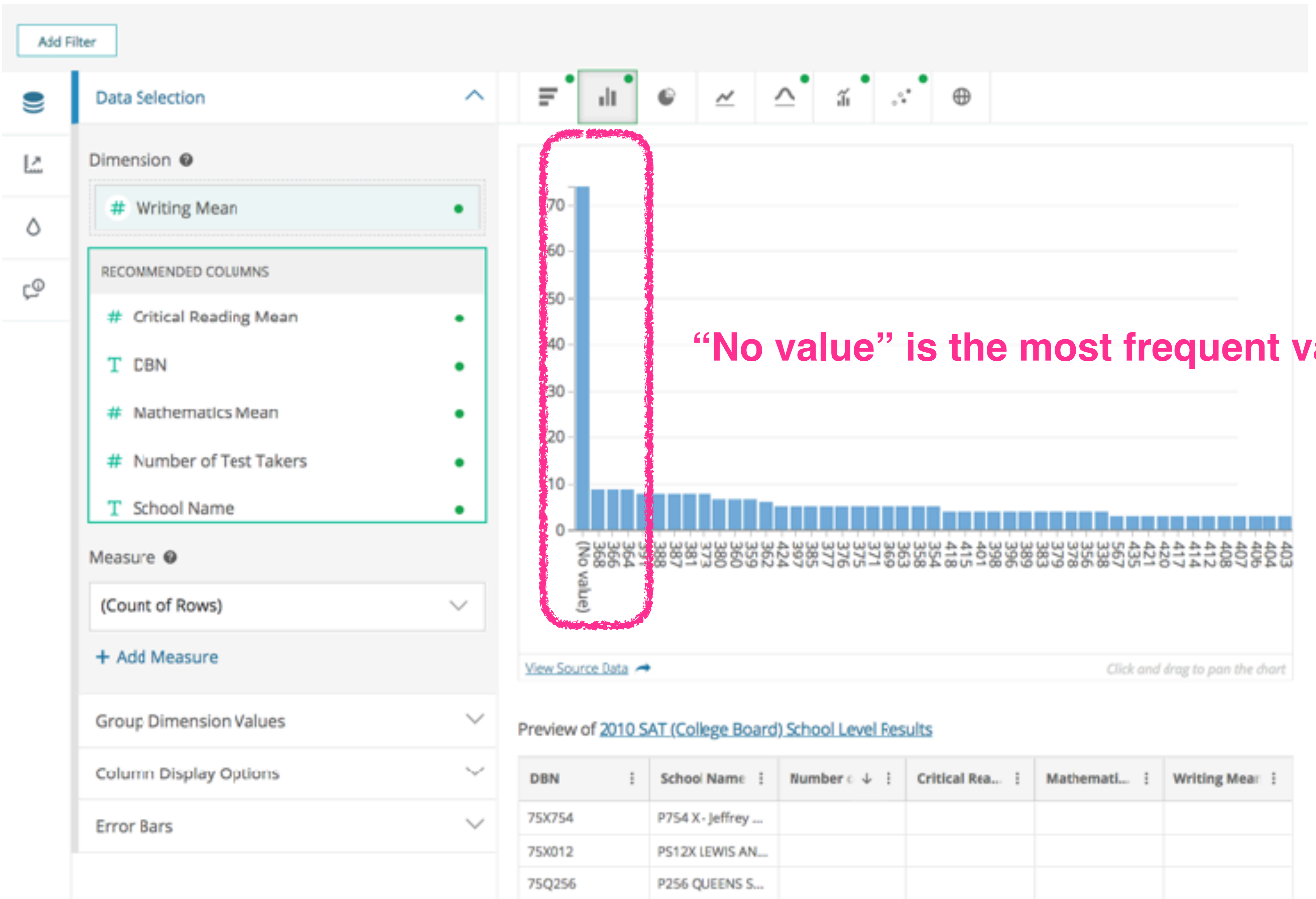
Columns in this Dataset

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

NYC Open Data



“No value” is the most frequent value

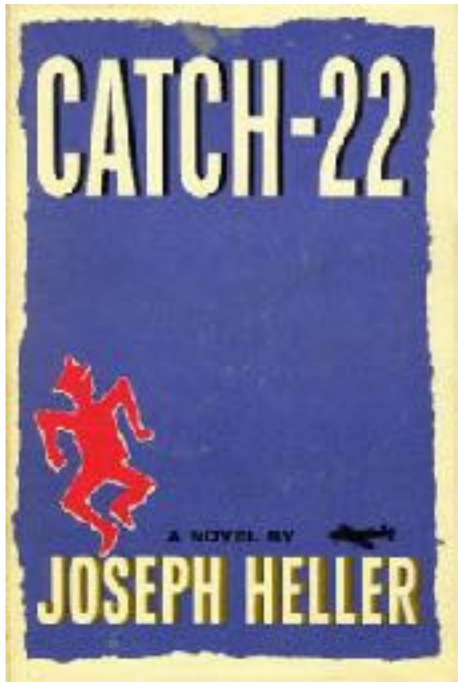
r/ai

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

r/ai

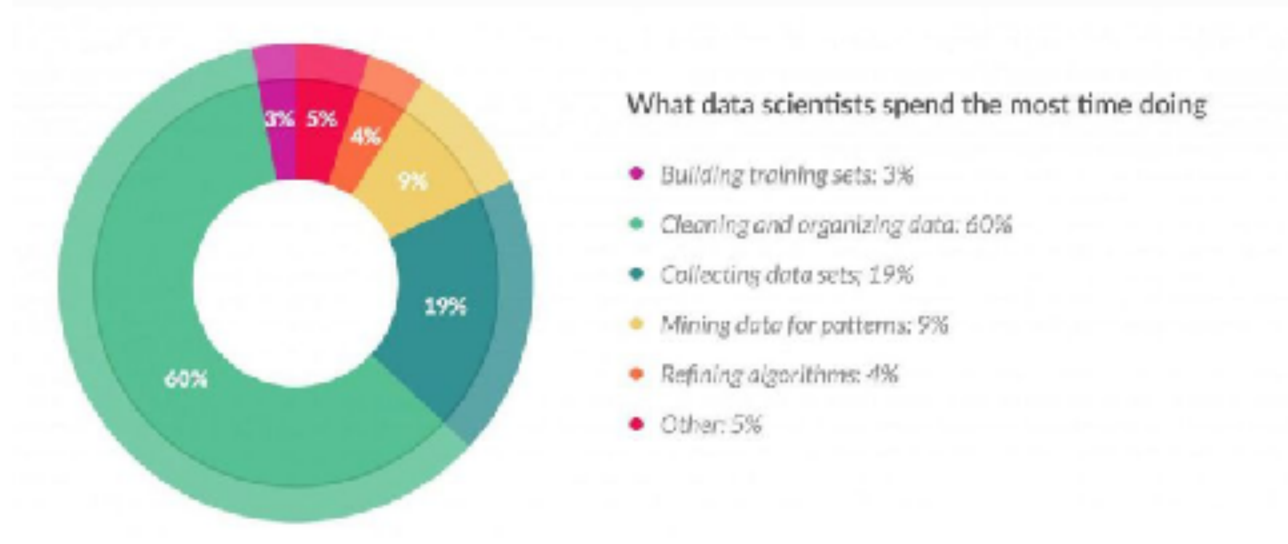
Data cleaning

Forbes

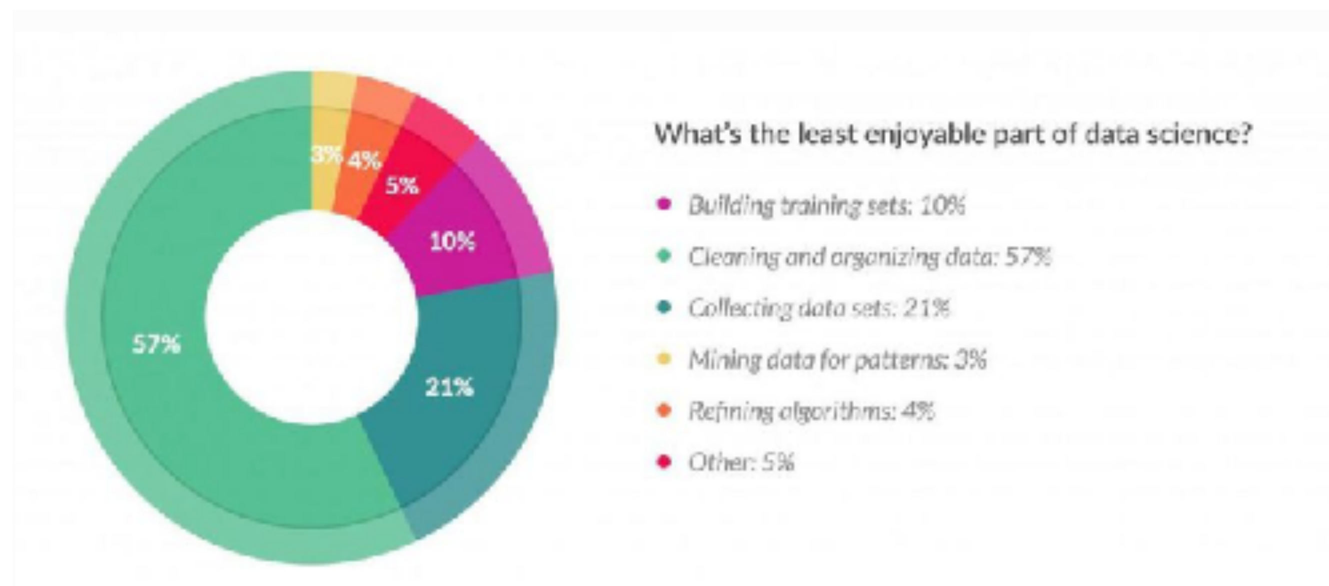
Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press Contributor

I write about technology, entrepreneurs and innovation.



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

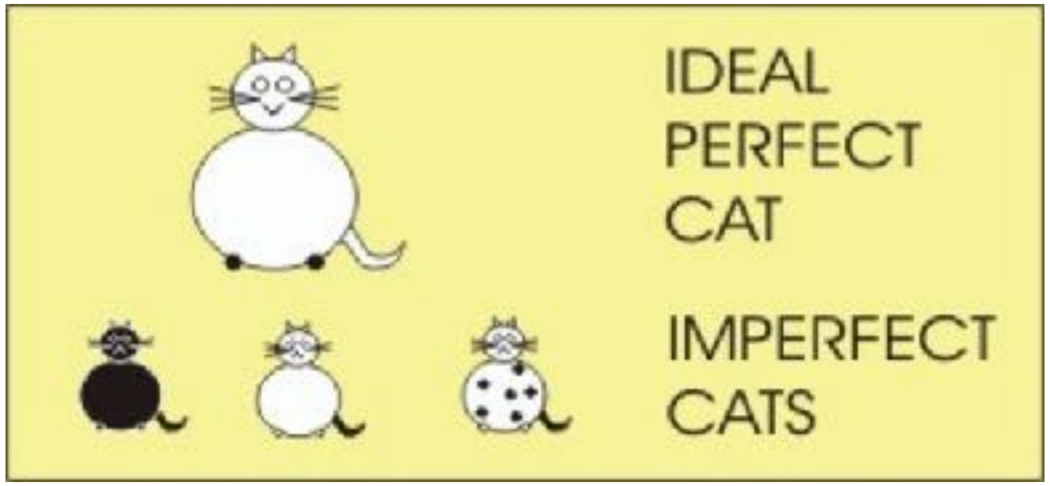


Find least enjoyable
Collecting data (21%)
Cleaning and organizing data (57%)

slide by Heiko Mueller


data profiling

DB (databases) vs DS (data science)

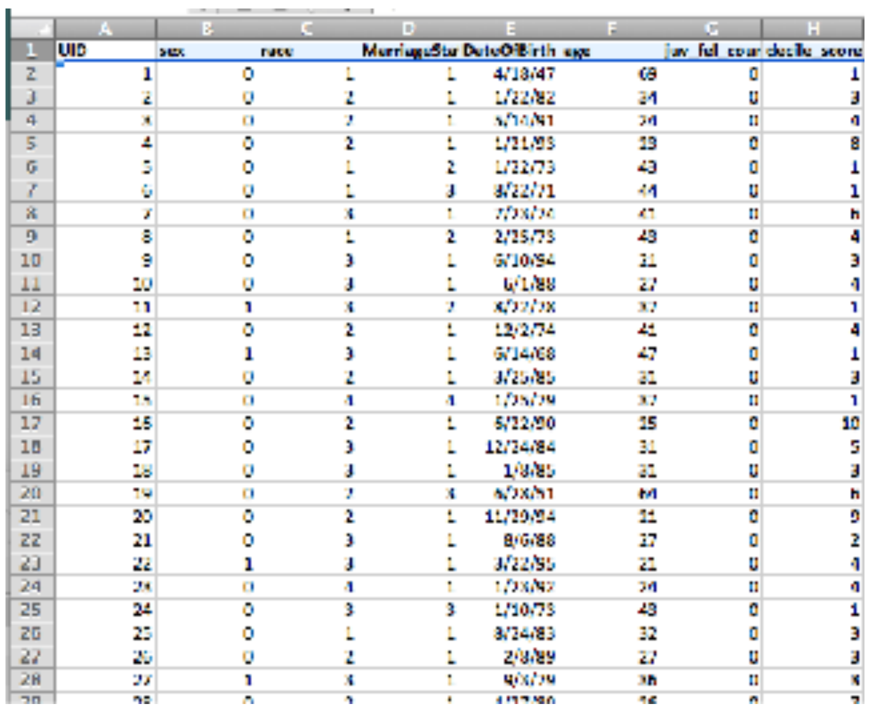


https://midnightmediamusings.wordpress.com/2014/07/01/plato-and-the-theory-of-forms/

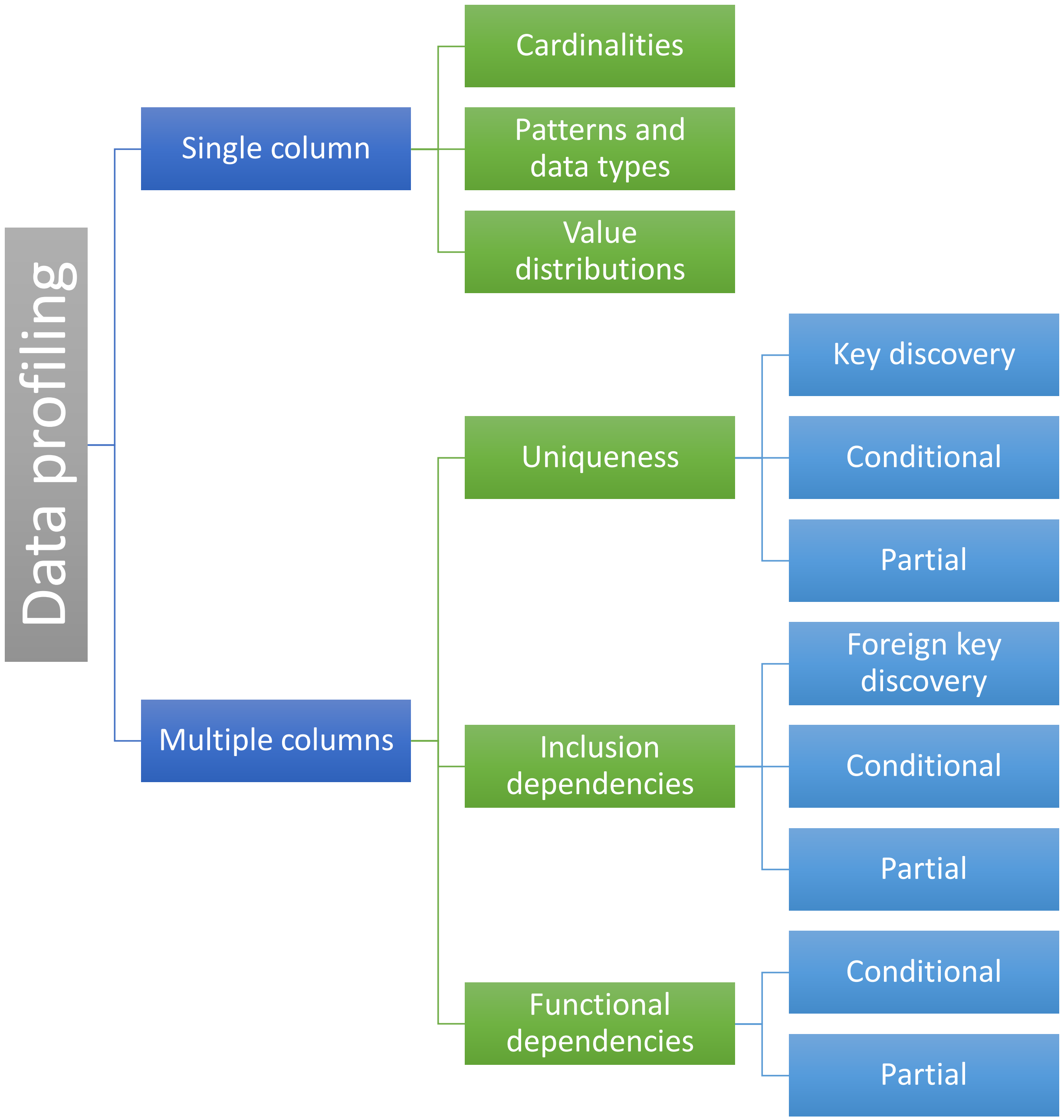
- **DB**: start with the schema, admit only data that fits; iterative refinement is possible, and common, but we are still schema-first
- **DS**: start with the data, figure out what schema it fits, or almost fits - reasons of usability, repurposing, low start-up cost

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

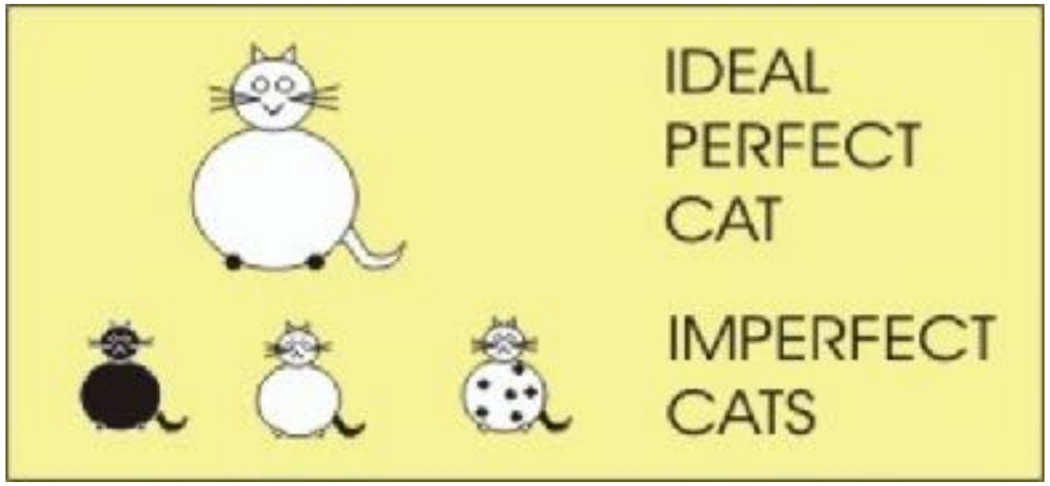
Data profiling



relational data (here: just one table)



An alternative classification

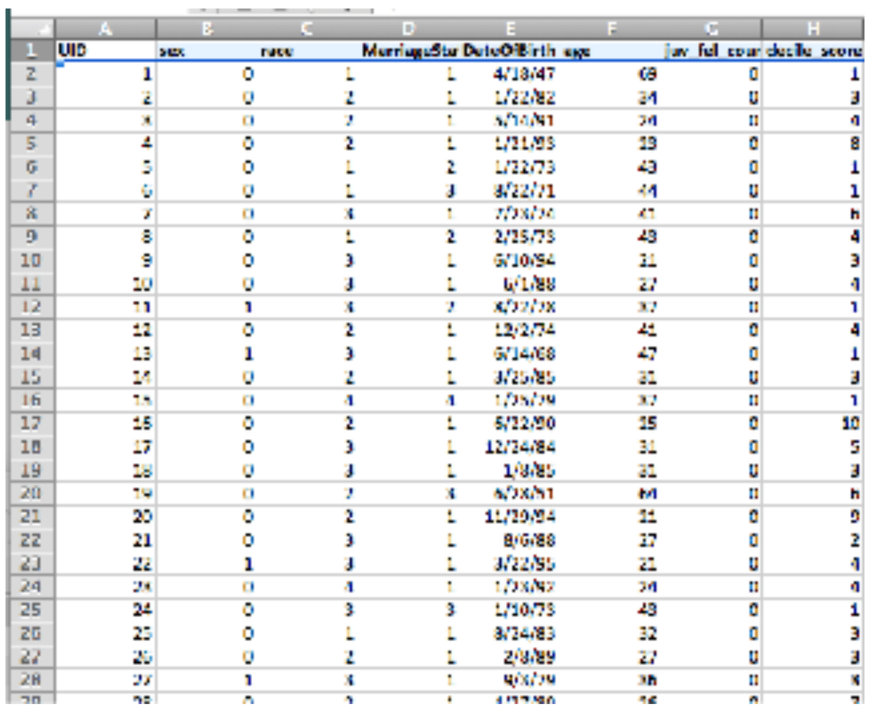


- To help understand the **statistics**, we look at value ranges, data types, value distributions per column or across columns, etc

- 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

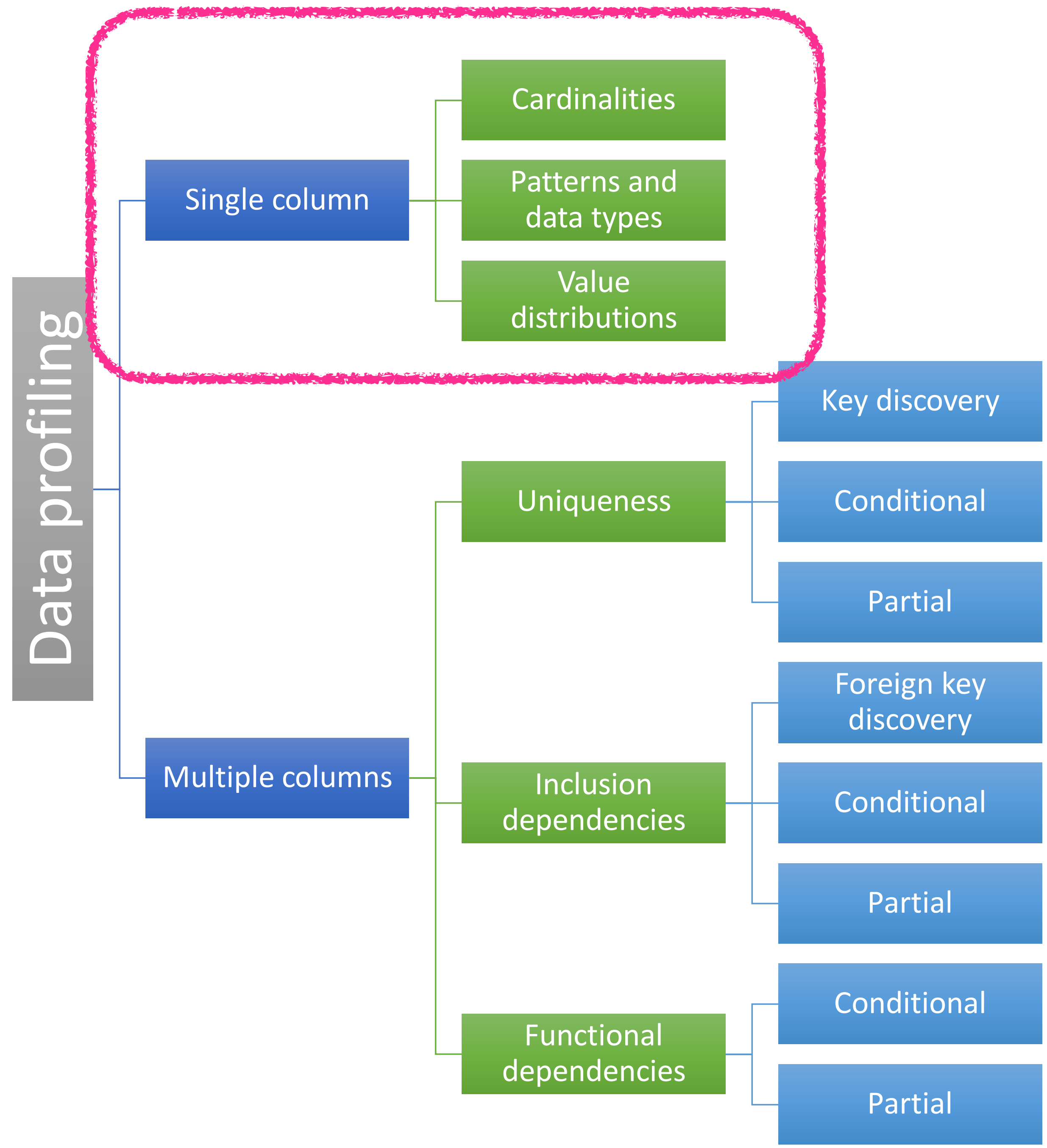
- We need both statistics and structure, they are mutually-reinforcing, and help us understand the **semantics** of the data - it's meaning

Data profiling



	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_coun	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	[illegible]	1	[illegible]	34	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	[illegible]	1	[illegible]	47	0	[illegible]
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	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	0	4	4	[illegible]	[illegible]	0	1
17	16	0	2	1	5/22/90	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	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]
21	20	0	2	1	11/20/94	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	[illegible]	34	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	[illegible]	1	[illegible]	[illegible]	0	[illegible]

relational data (here: just one table)



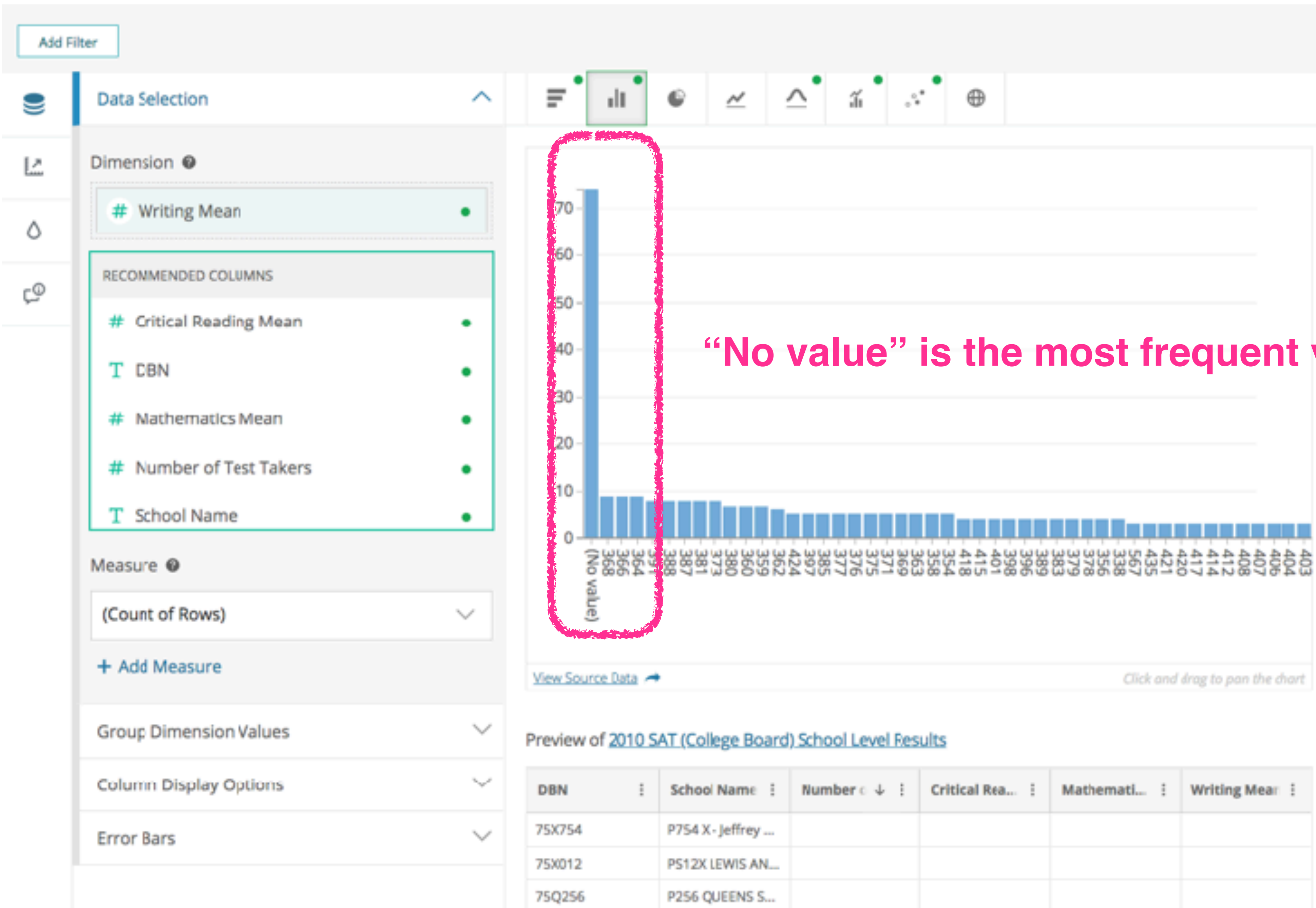
Single column: cardinalities, data types

- 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)]

r/ai

NYC Open Data



"No value" is the most frequent value

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?

r/ai

Single column: cardinalities, data types

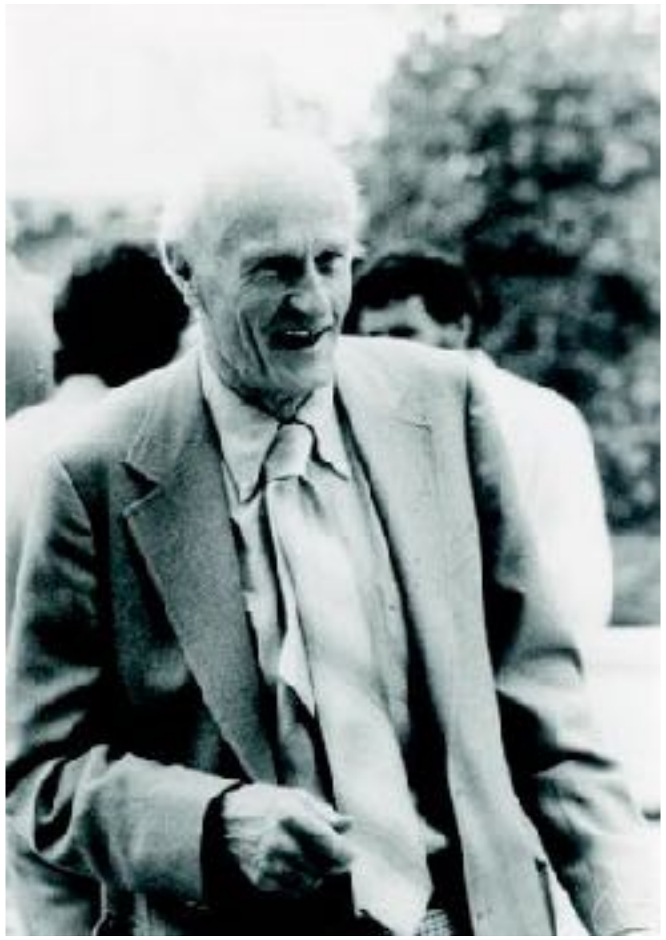
- cardinality of relation **R** - number of rows
- domain cardinality of a column **R.a** - number of **distinct** values
- attribute value **length**: min, max, average, median
- **basic data type**: string, numeric, date, time, ….
- number of percentage of **null** values of a given attribute
- **regular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

[Abedjan, Golab & Naumann (2017)]

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!**

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.



Stephen Kleene

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

r/ai

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
(718) 206-0545
(814) 681-6200
(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

r/ai

Oakham’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
(718) 206-0545
(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

(	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
(718) 868-2300
(718) 206-0545
(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

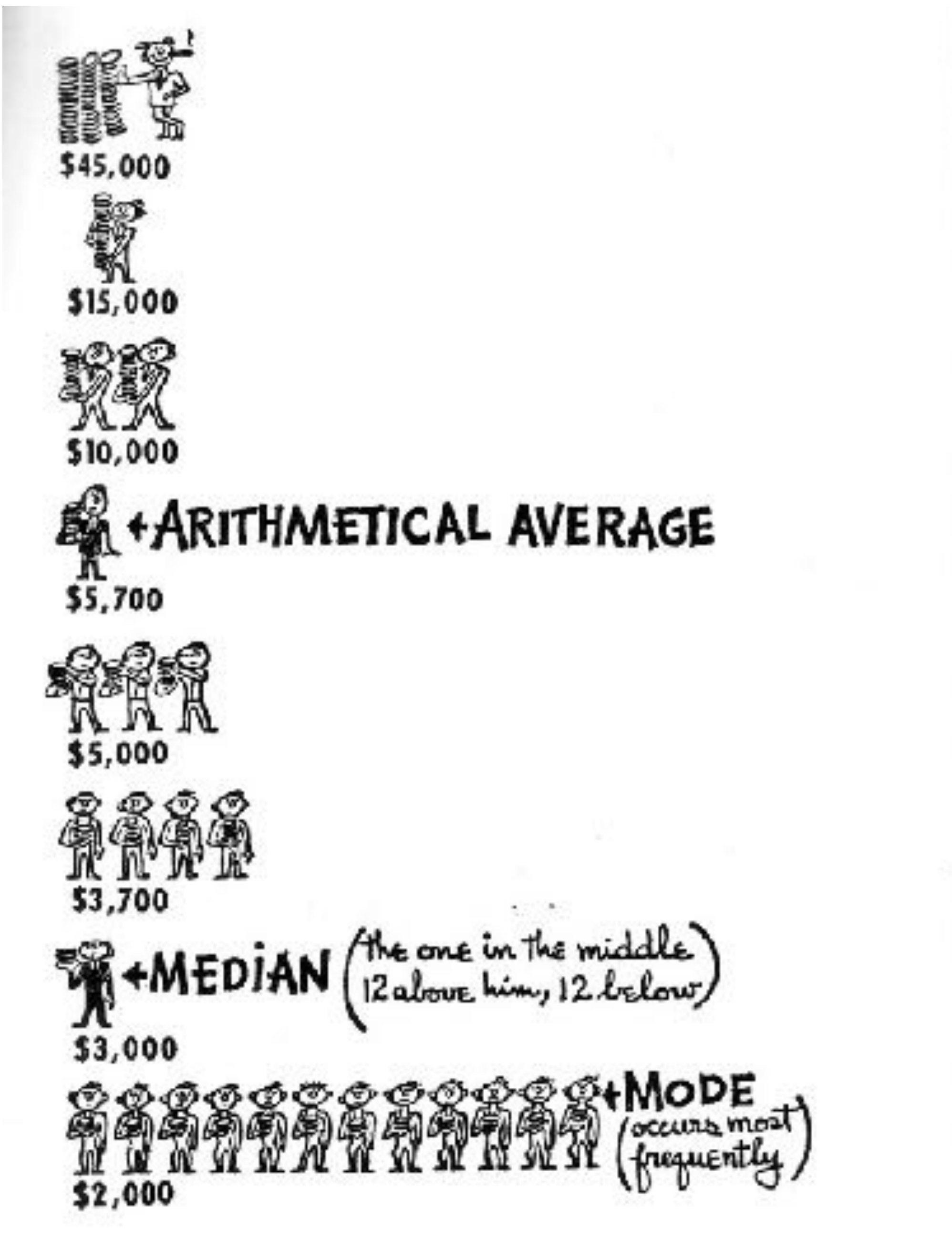
```
(\d{3}) \d\w{2}.{2}\w{3}
```

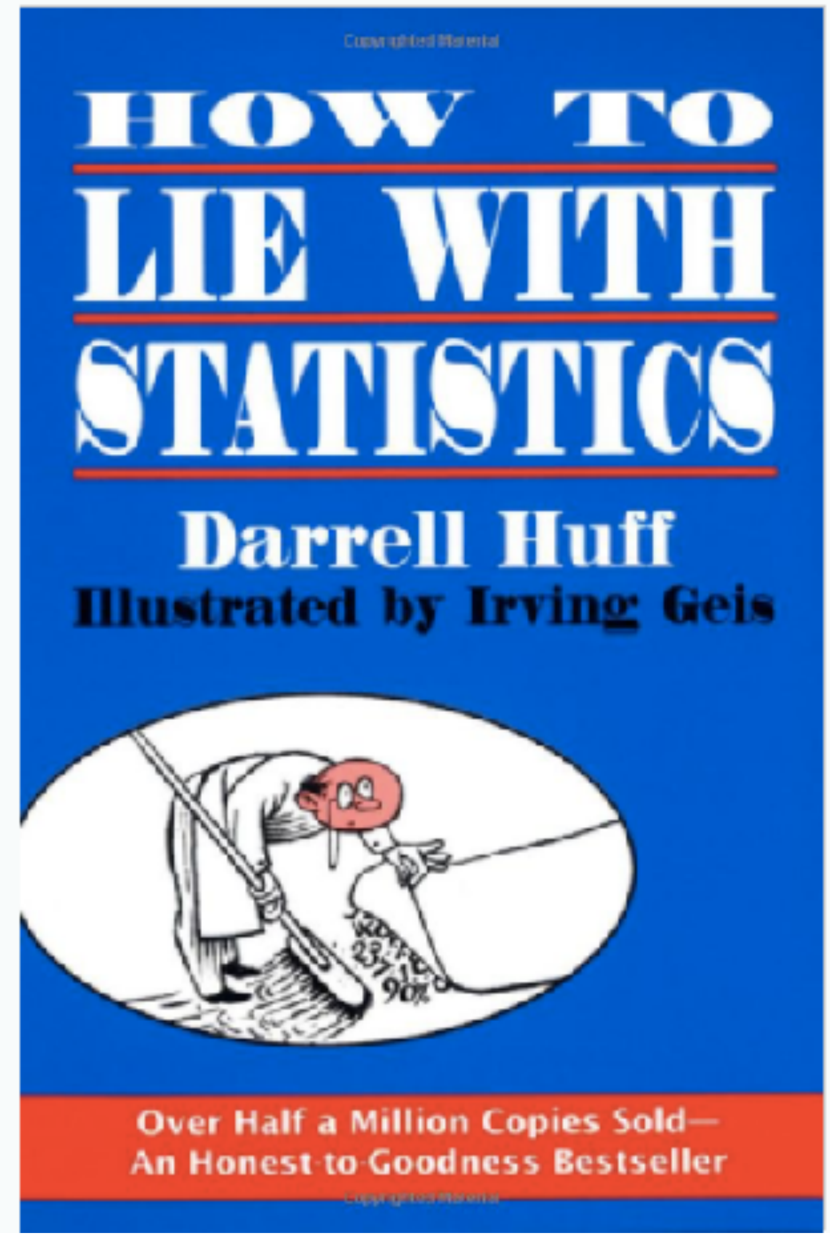

r/ai

Single column: basic stats, distributions

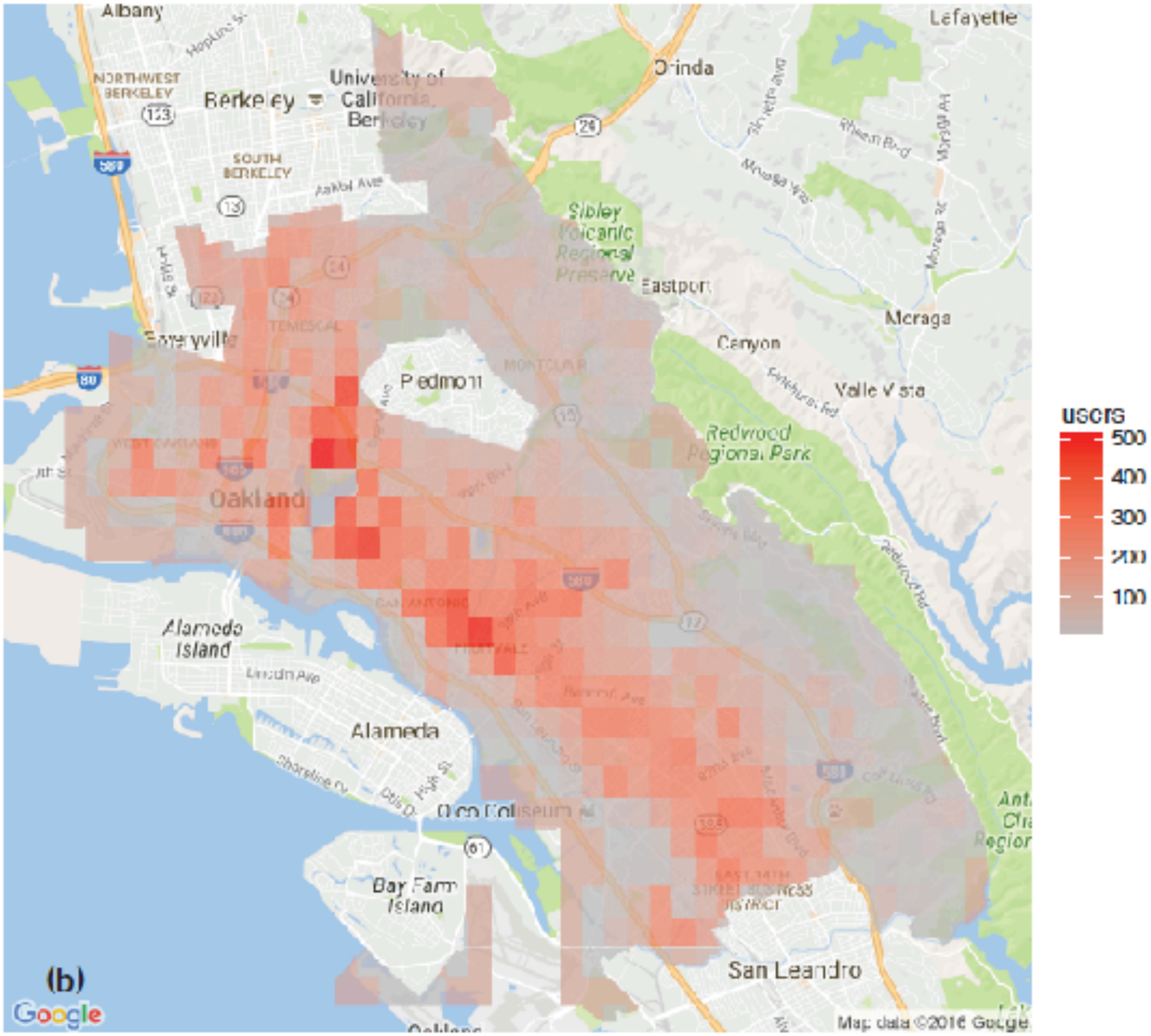
- min, max, **average**, median value of **R.a**
- **histogram**
  - equi-width - (approximately) the same number of distinct values in each bucket (e.g., age broken down into 5-year windows)
  - equi-depth (approximately) the same number of tuples in each bucket
  - biased histograms use different granularities 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
- …

The well-chosen average

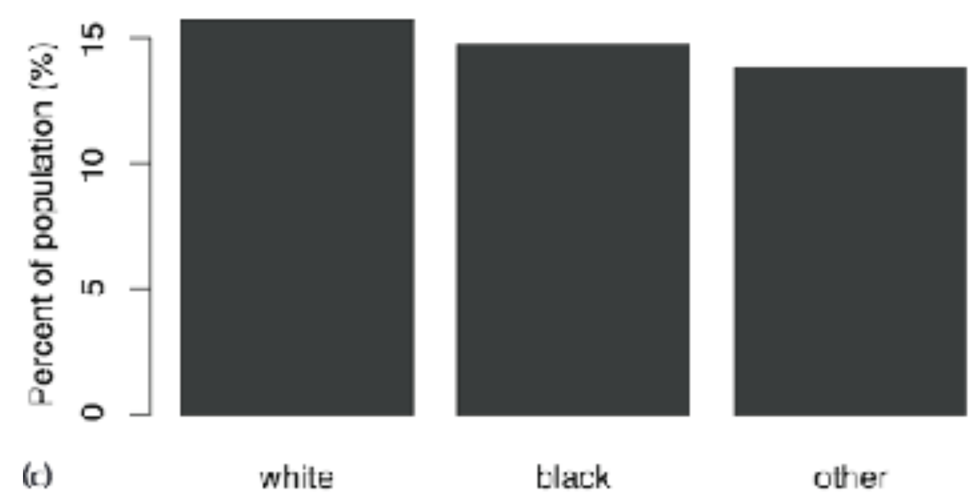




Is my data biased? (histograms + geo)



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

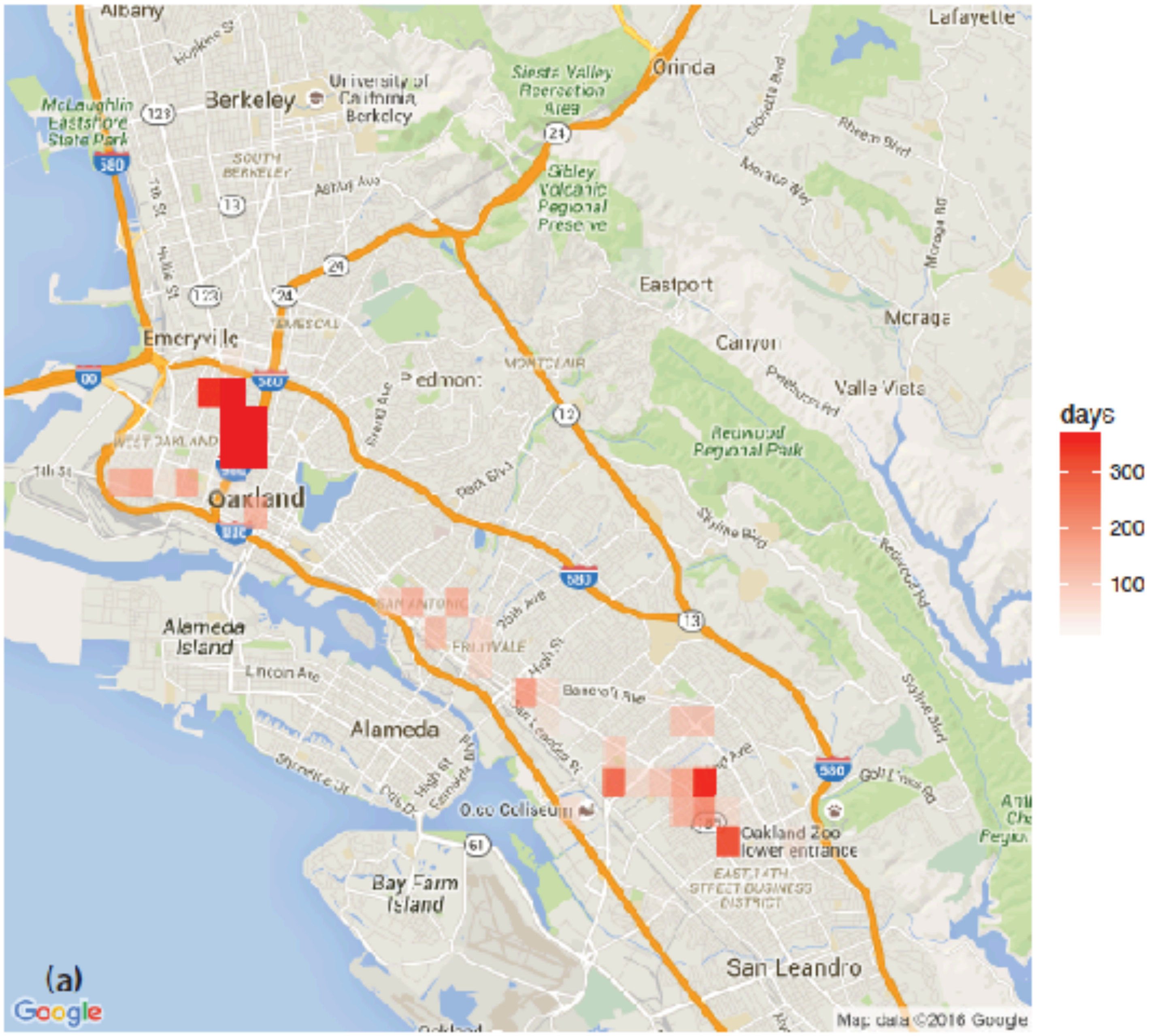


Estimated drug use by race

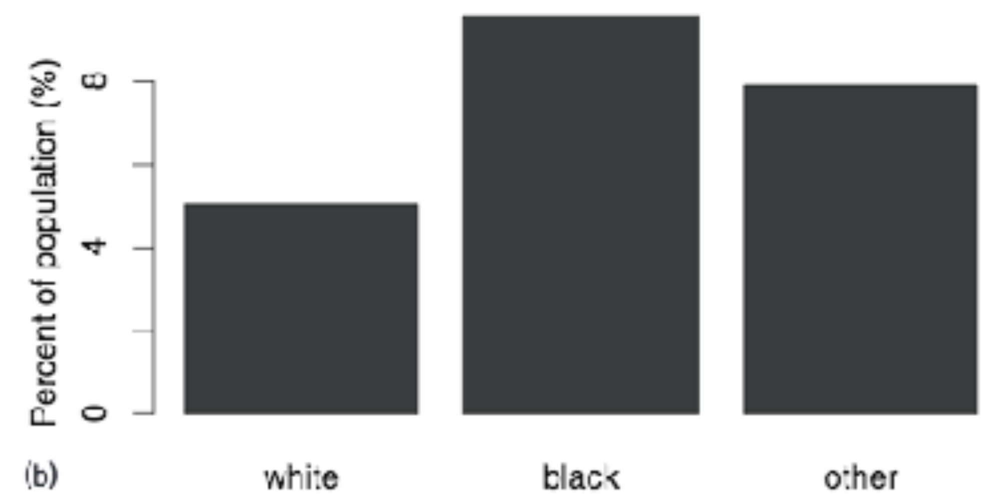
[Lum & Isaac (2016)]

r/ai

Is my data biased? (histograms + geo)

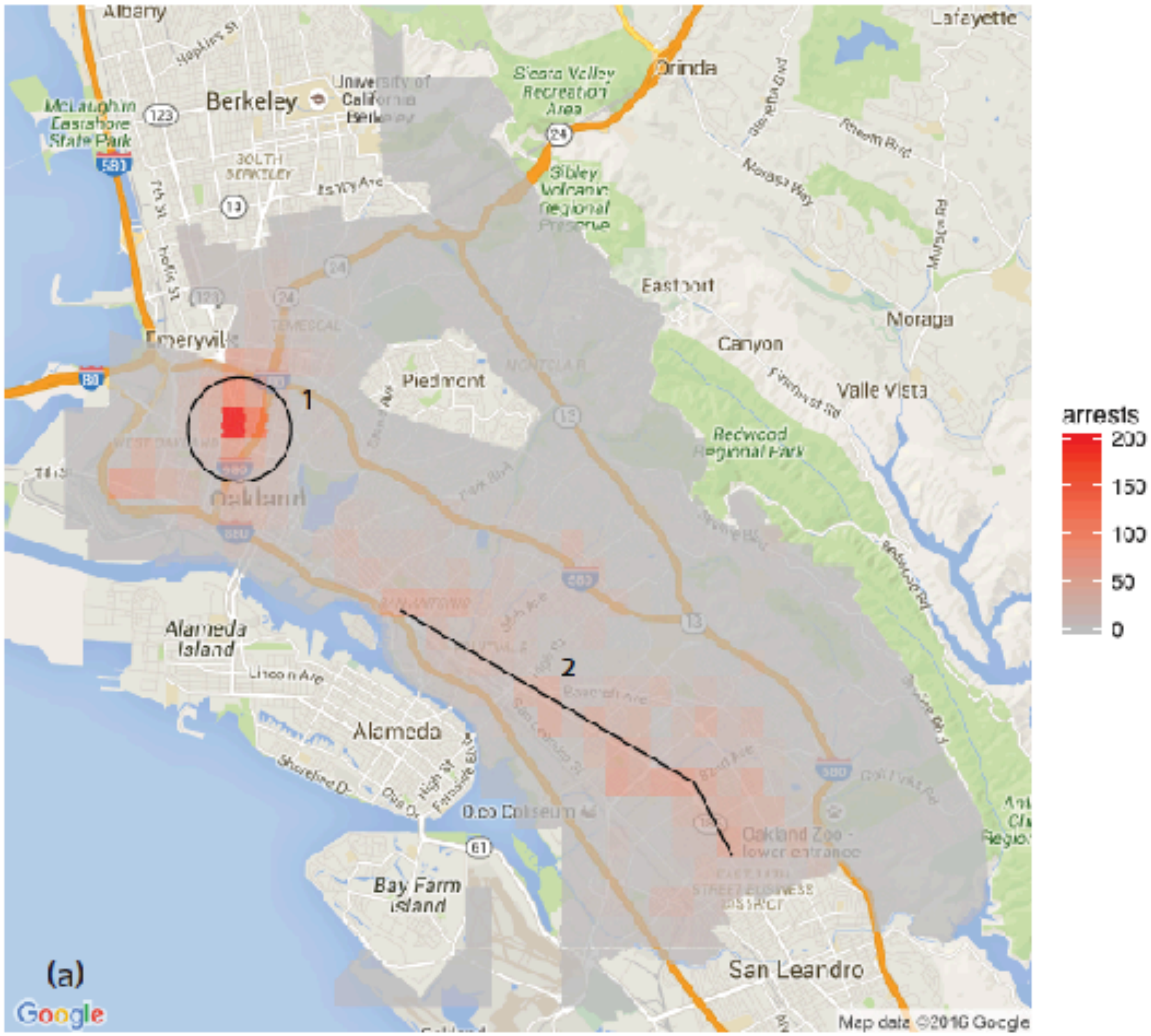


Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland, CA, police data for 2011

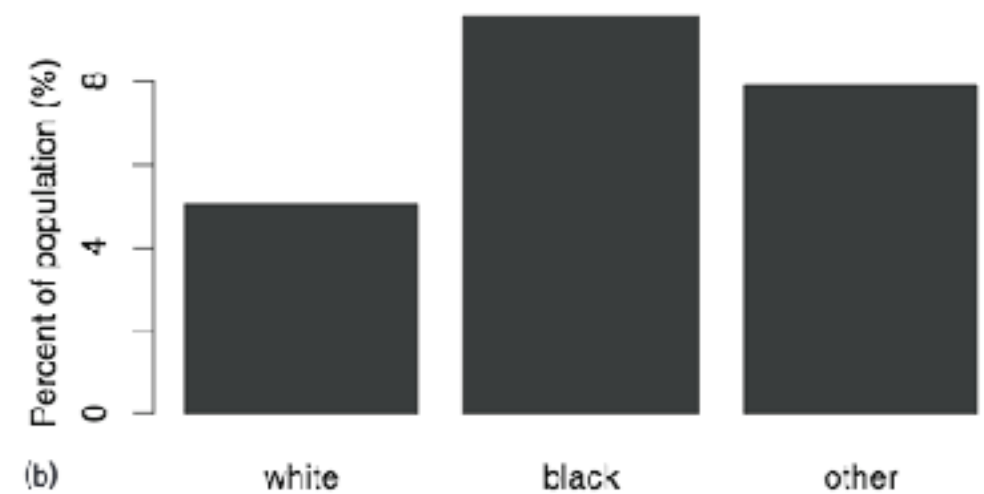


Targeted policing for drug crimes by race

Is my data biased? (histograms + geo)



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



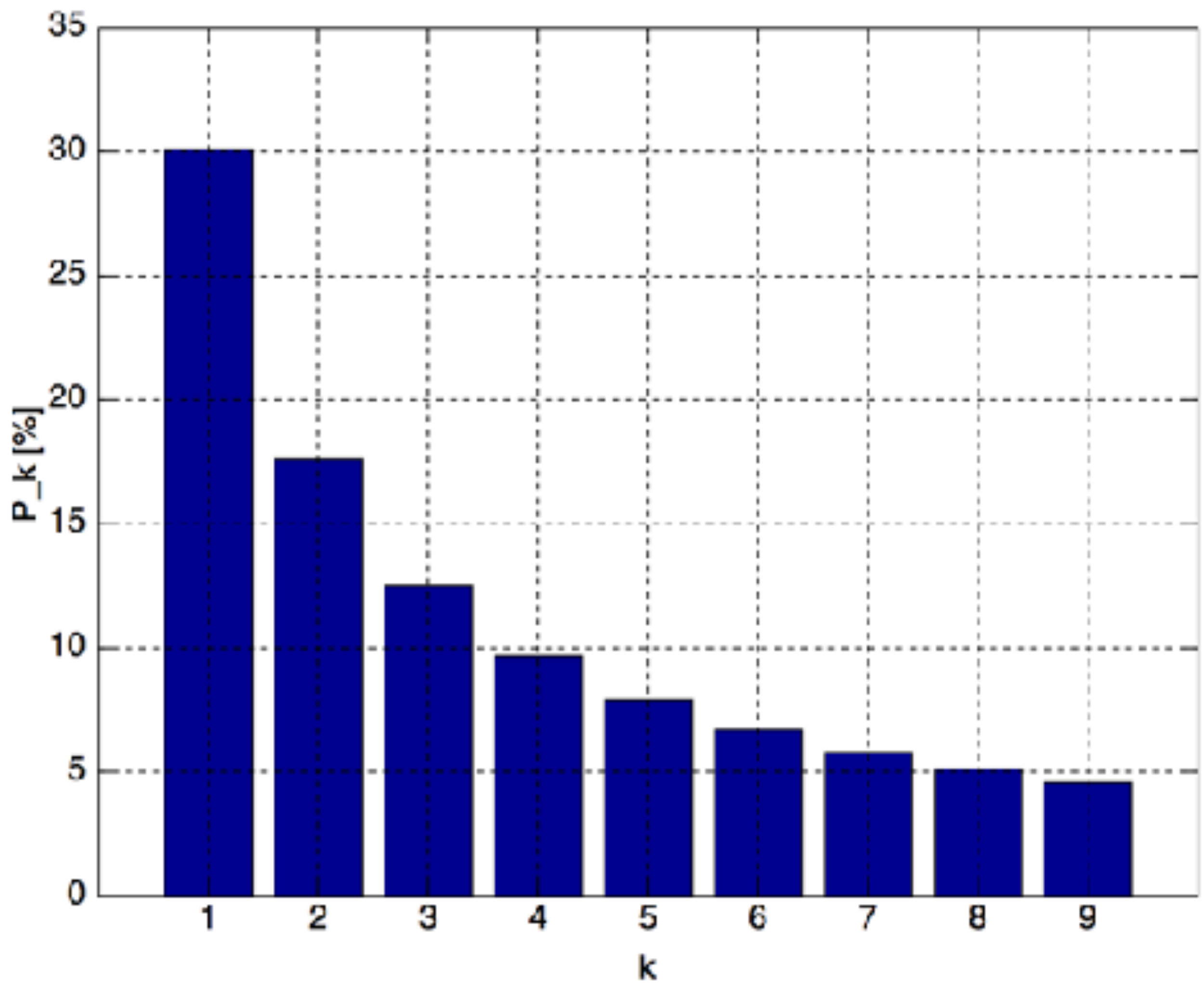
Targeted policing for drug crimes by race

[Lum & Isaac (2016)]

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.

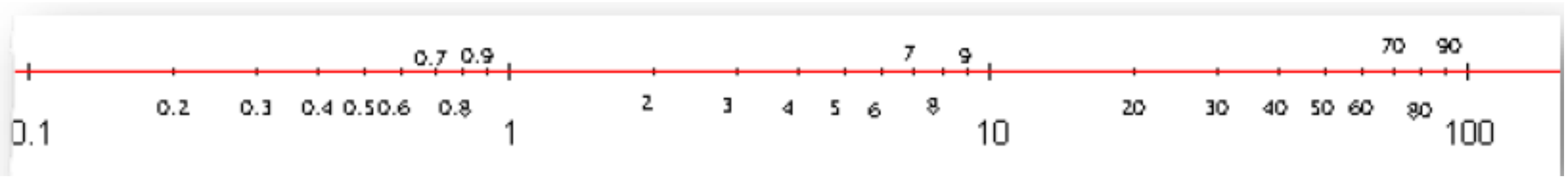
[Benford: "The law of anomalous numbers" *Proc. Am. Philos. Soc.*, 1938]

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.

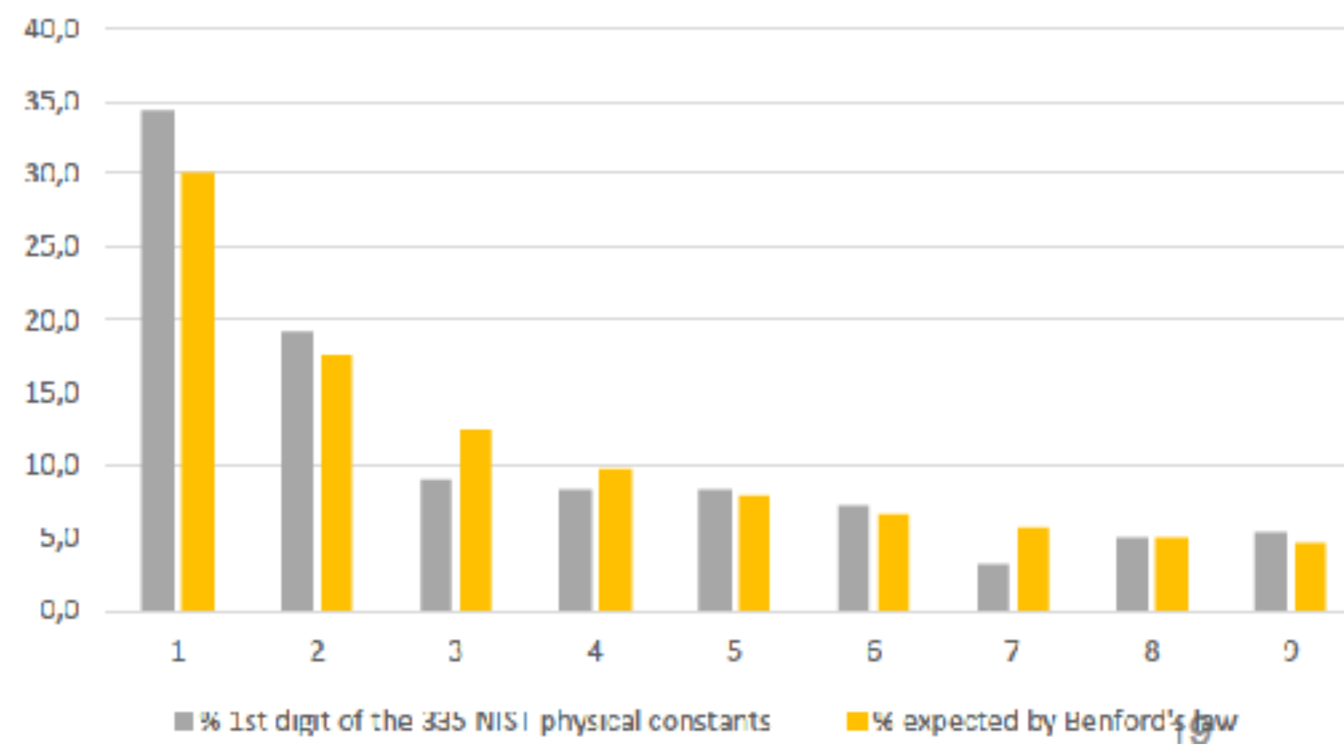
https://en.wikipedia.org/wiki/Benford%27s_law

[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
- ….

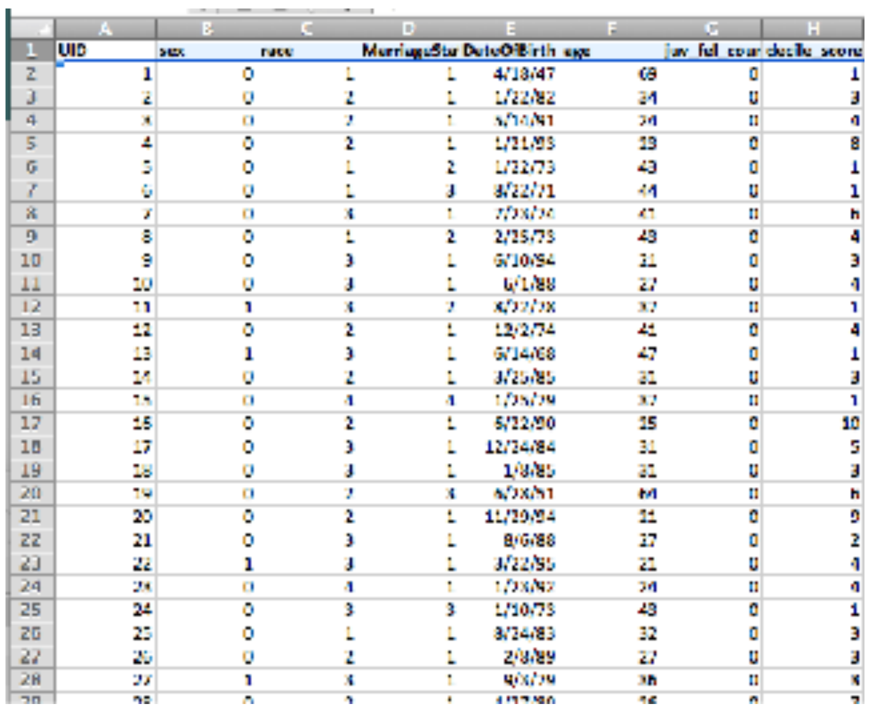
used in fraud detection!



height of tallest structures

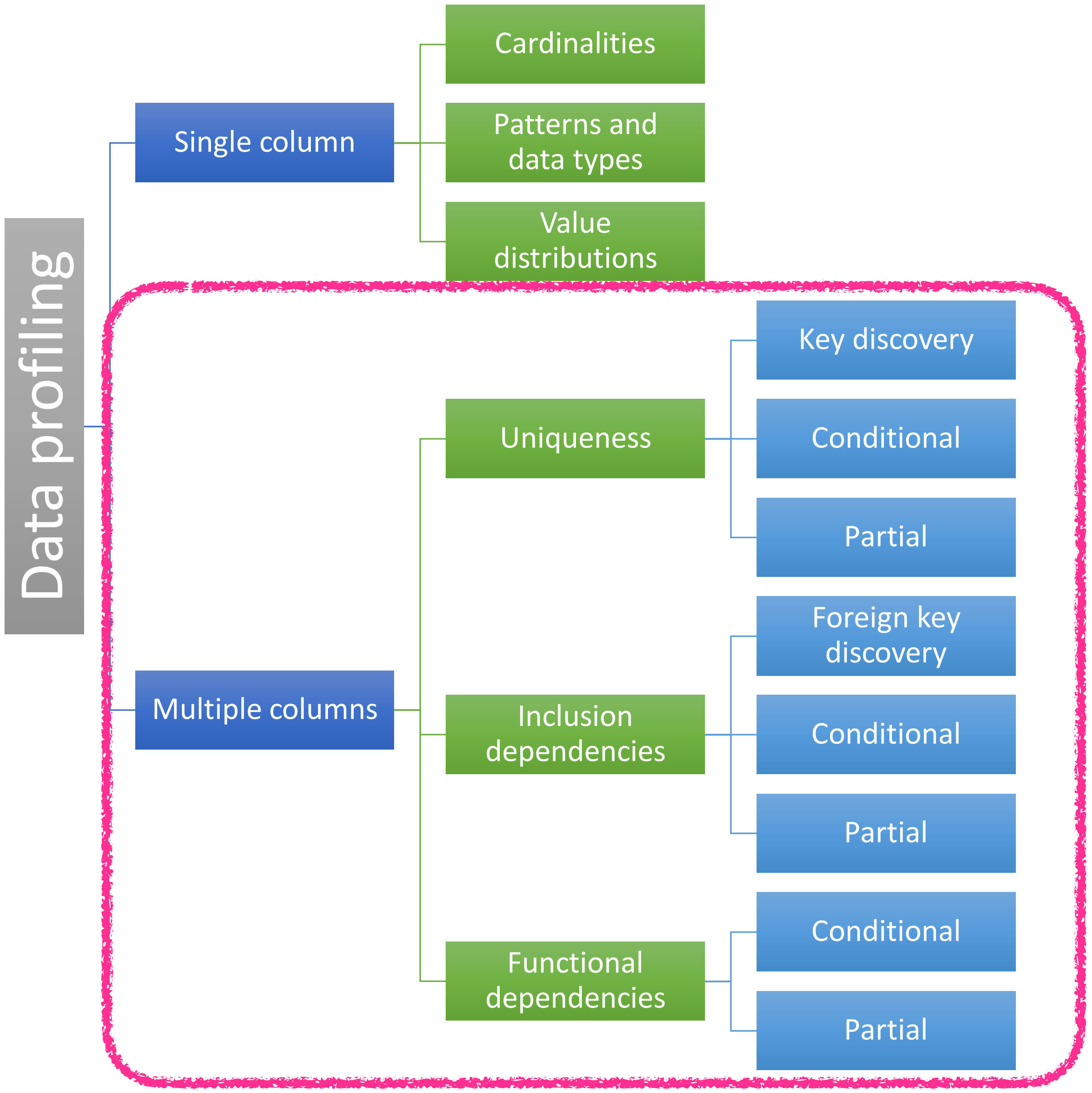
[Abedjan, Golab & Naumann (2017)]

Data profiling



	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageStatus	DateOfBirth	age	juv_fel_count	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	[illegible]	1	[illegible]	34	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	[illegible]	1	[illegible]	[illegible]	0	[illegible]
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	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	0	4	4	[illegible]	[illegible]	0	1
17	[illegible]	0	2	1	5/22/90	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	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]
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	[illegible]	0	4	1	[illegible]	[illegible]	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	[illegible]	1	[illegible]	[illegible]	0	[illegible]

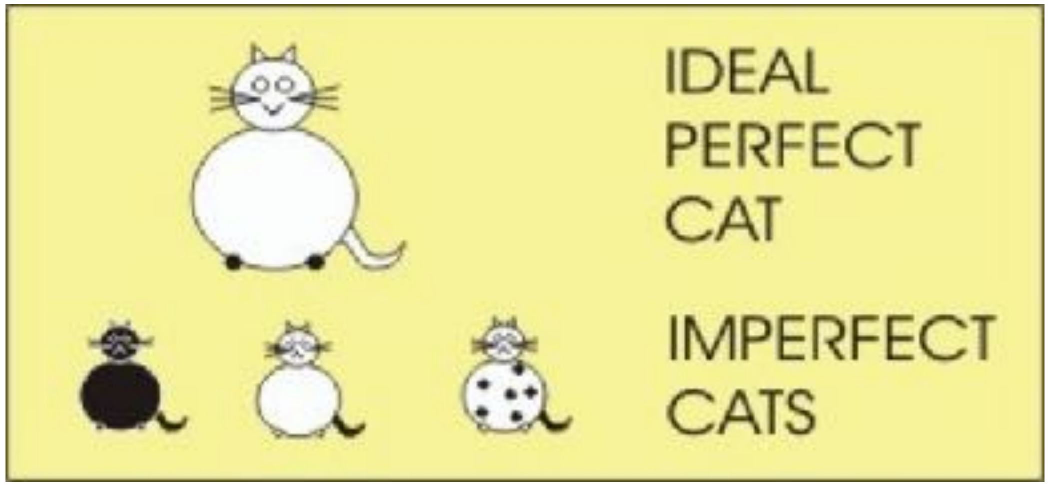
relational data (here: just one table)



[Abedjan, Golab & Naumann (2017)]

r/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

r/ai

taming technical bias

r/ai

Reading for this part

Taming Technical Bias in Machine Learning Pipelines *

Sebastian Schelter
University of Amsterdam & Ahold Delhaize
Amsterdam, The Netherlands
s.schelter@uva.nl

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Abstract

Machine Learning (ML) is commonly used to automate decisions in domains as varied as credit and lending, medical diagnosis, and hiring. These decisions are consequential, imploring us to carefully balance the benefits of efficiency with the potential risks. Much of the conversation about the risks centers around bias — a term that is used by the technical community ever more frequently but that is still poorly understood. In this paper we focus on technical bias — a type of bias that has so far received limited attention and that the data engineering community is well-equipped to address. We discuss dimensions of technical bias that can arise through the ML lifecycle, particularly when it's due to preprocessing decisions or post-deployment issues. We present results of our recent work, and discuss future research directions. Our over-all goal is to support the development of systems that expose the knobs of responsibility to data scientists, allowing them to detect instances of technical bias and to mitigate it when possible.

1 Introduction

Machine Learning (ML) is increasingly used to automate decisions that impact people's lives, in domains as varied as credit and lending, medical diagnosis, and hiring. The risks and opportunities arising from the wide-spread use of predictive analytics are garnering much attention from policy makers, scientists, and the media. Much of this conversation centers around *bias* — a term that is used by the technical community ever more frequently but that is still poorly understood.

In their seminal 1996 paper, Friedman and Nissenbaum identified three types of bias that can arise in computer systems: pre-existing, technical, and emergent [9]. We briefly discuss these in turn, see Stoyanovich et al. [33] for a more comprehensive overview.

- *Pre-existing bias* has its origins in society. In ML applications, this type of bias often exhibits itself in the input data; detecting and mitigating it is the subject of much research under the heading of algorithmic fairness [5]. Importantly, the presence or absence of pre-existing bias cannot be scientifically verified, but rather is postulated based on a belief system [8, 12]. Consequently, the effectiveness — or even the validity — of a technical attempt to mitigate pre-existing bias is predicated on that belief system.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work was supported in part by NSF Grants No. 1926250, 1934464, and 1922658, and by Ahold Delhaize. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

The VLDB Journal
https://doi.org/10.1007/s00778-021-00726-w

SPECIAL ISSUE PAPER

Data distribution debugging in machine learning pipelines

Stefan Grafberger[1] · Paul Groth[1] · Julia Stoyanovich[2] · Sebastian Schelter[1]

Received: 27 February 2021 / Revised: 9 September 2021 / Accepted: 3 December 2021

Abstract
Machine learning (ML) is increasingly used to automate impactful decisions, and the risks arising from this widespread use are garnering attention from policy makers, scientists, and the media. ML applications are often brittle with respect to their input data, which leads to concerns about their correctness, reliability, and fairness. In this paper, we describe mlinspect, a library that helps diagnose and mitigate technical bias that may arise during preprocessing steps in an ML pipeline. We refer to these problems collectively as *data distribution bugs*. The key idea is to extract a directed acyclic graph representation of the dataflow from a preprocessing pipeline and to use this representation to automatically instrument the code with predefined *inspections*. These inspections are based on a lightweight annotation propagation approach to propagate metadata such as lineage information from operator to operator. In contrast to existing work, mlinspect operates on declarative abstractions of popular data science libraries like estimator/transformer pipelines and does not require manual code instrumentation. We discuss the design and implementation of the mlinspect library and give a comprehensive end-to-end example that illustrates its functionality.

Keywords Data debugging · Machine learning pipelines · Data preparation for machine learning

1 Introduction

Sebastian Schelter
s.schelter@uva.nl

Stefan Grafberger
s.grafberger@uva.nl

Paul Groth
p.t.groth@uva.nl

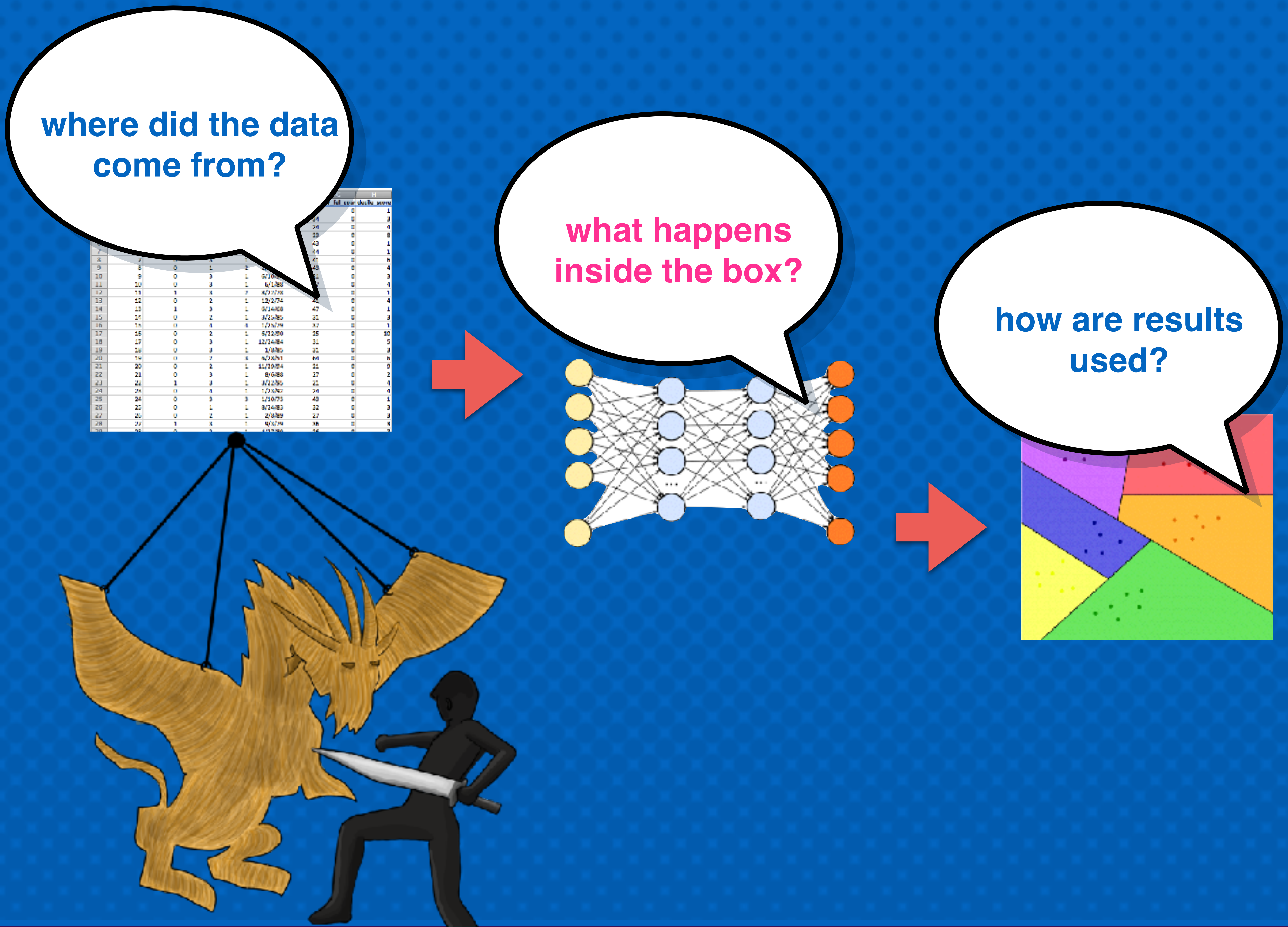
Julia Stoyanovich
stoyanovich@nyu.edu

1 University of Amsterdam, Amsterdam, Netherlands
2 New York University, New York, USA

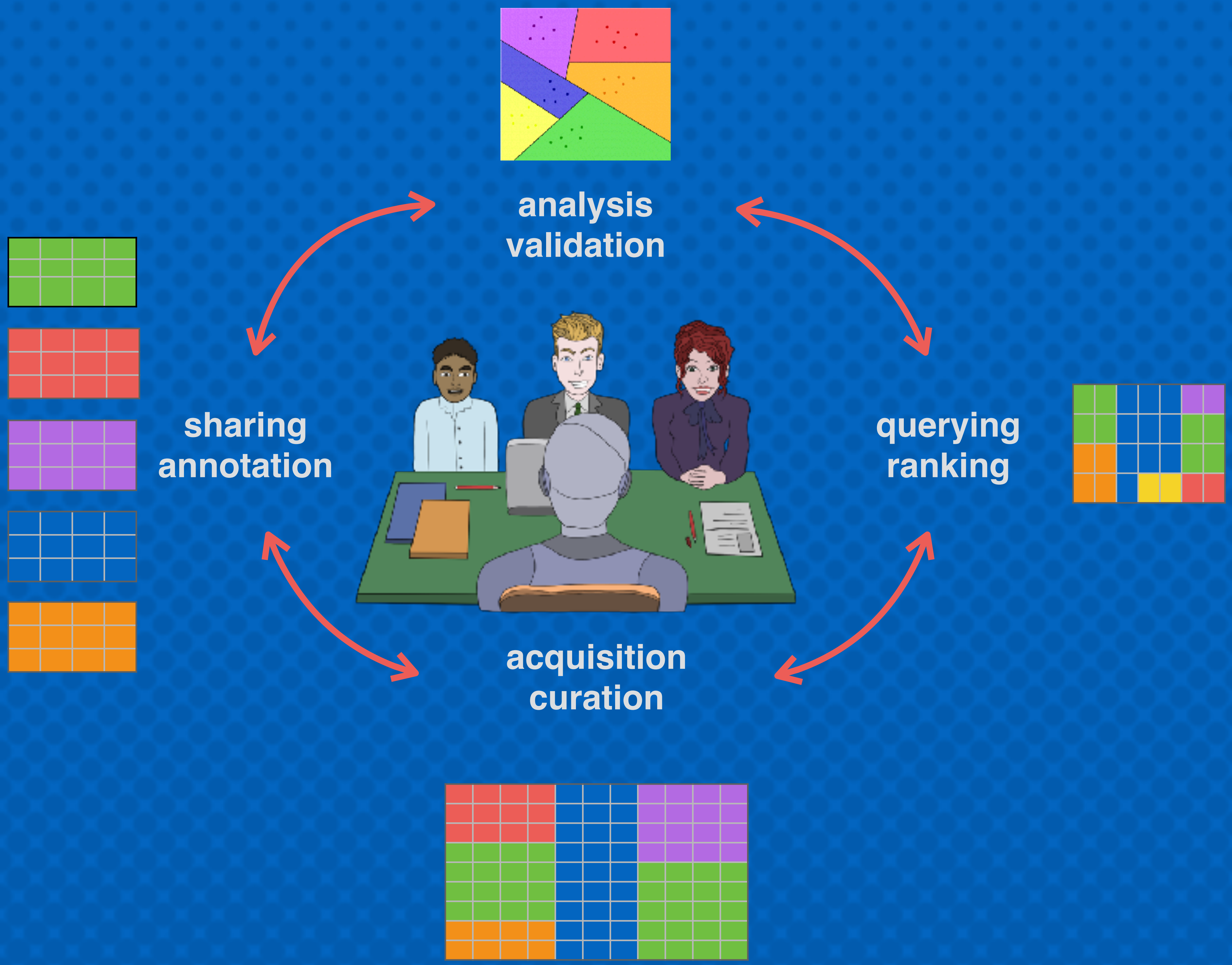
Published online: 21 January 2022

r/ai

The "last-mile" view of responsible AI



Zooming out to the lifecycle view



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)]

r/ai

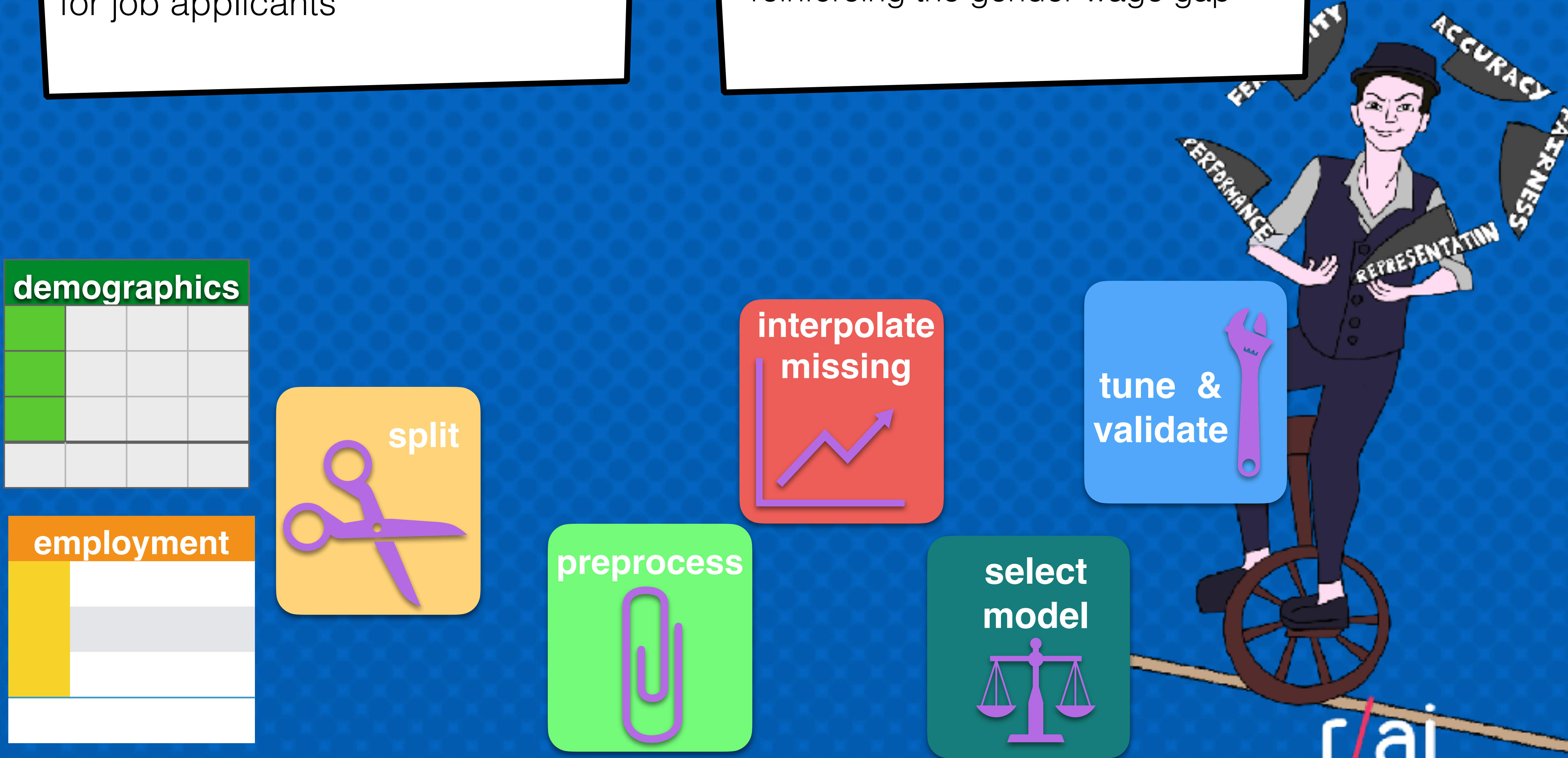
Model development lifecycle

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



[Schelter, He, Khilnani, Stoyanovich (2020)]

dimensions of technical bias

r/ai

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)
- …..

should we be filling these in? if so, how?

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

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)?

r/ai

Data filtering

"filtering" operations (like selection and join), **can arbitrarily change demographic group 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%

→

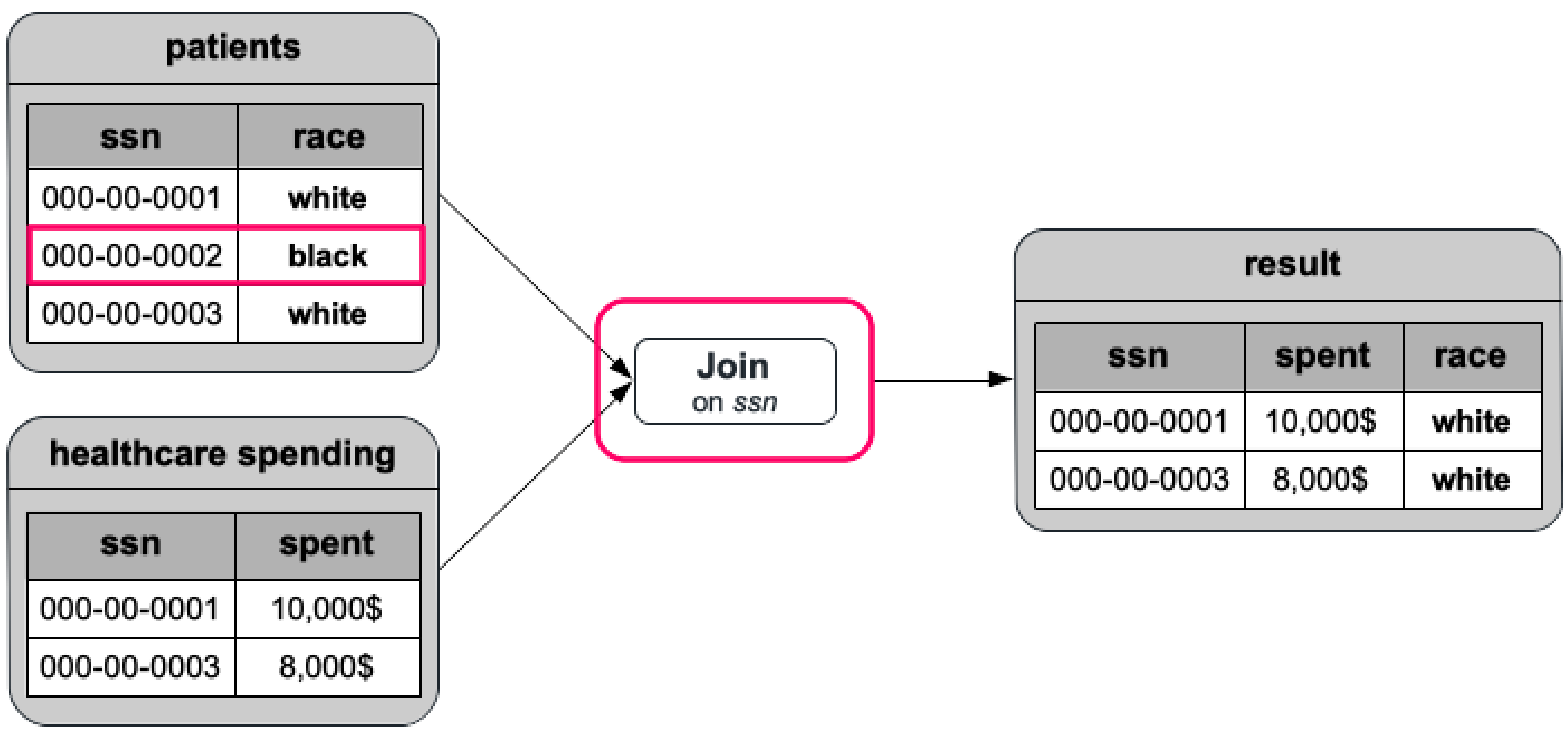
age_group	county
60	CountyA
60	CountyA
20	CountyA

66% vs 33%

Data filtering

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

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



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

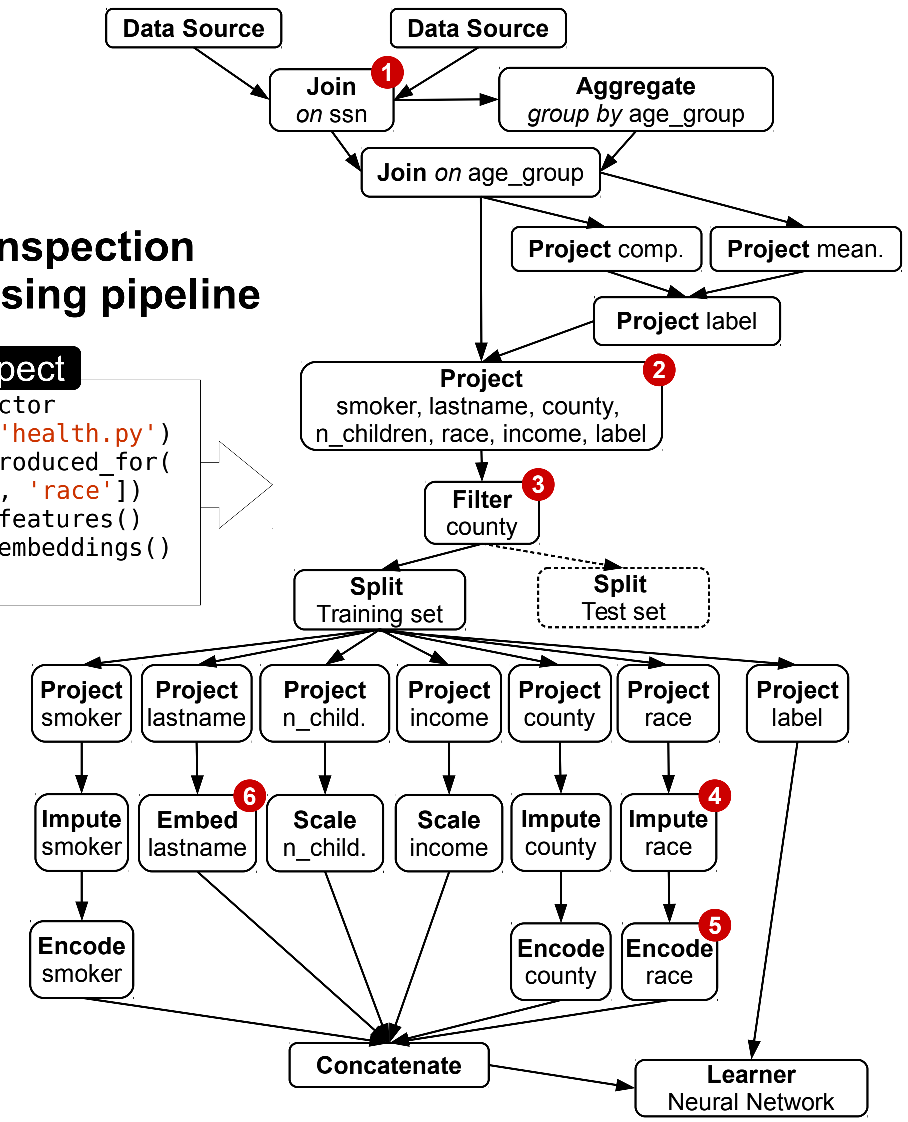
```
# load input data sources, join to single table
patients = pandas.read_csv(…)
histories = pandas.read_csv(…)
data = pandas.merge([patients, histories], on=['ssn'])

# 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 (2021)]

r/ai

Data debugging: mlinspect

- similar to code inspection in modern IDEs, but specifically for data
- works on existing pipeline code using libraries like pandas and scikit-learn
- negligible performance overhead

ACM SIGMOD 2021 demo (4 min)

https://surfdrive.surf.nl/files/index.php/s/ybriyzsdc6vcd2w

CIDR 2021 talk (10 min)

https://www.youtube.com/watch?v=Ic0aD6lv5h0

https://github.com/stefan-grafberger/mlinspect

Sound experimentation

> "A theory or idea shouldn't be scientific unless it could, in principle, be proven false."
>
> *Karl Popper*

- software-engineering and data science best-practices
- data isolation: training / validation / test
- accounting for **variability** when observing trends
- tuning hyper-parameters: **for what objective**?