Responsible Data Science

Understanding our data: Data profiling

Prof. Julia Stoyanovich

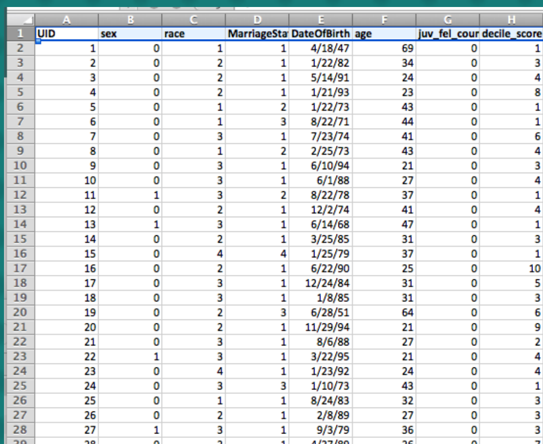
Center for Data Science &
Computer Science and Engineering
New York University

dataRESPONSIBLY

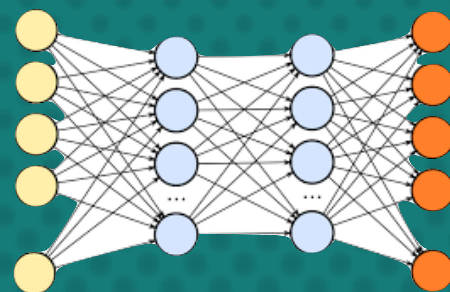
NYU

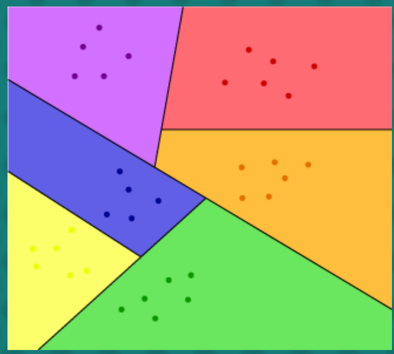
module 1: algorithmic fairness

“Bias” in predictive analytics



	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





Statistical

model does not summarize the data correctly

Societal

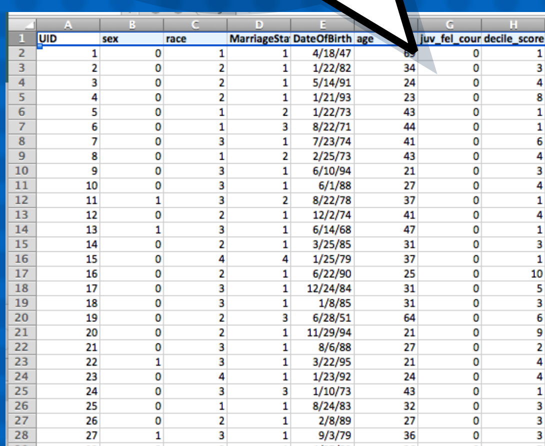
data does not represent the world correctly

module 2:

the data science lifecycle

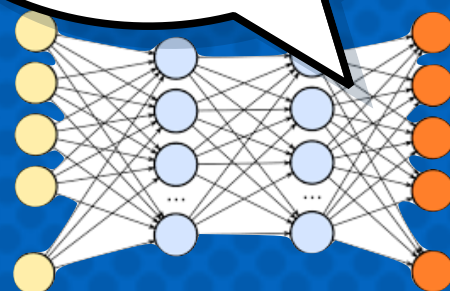
Frog's eye view

where did the data come from?

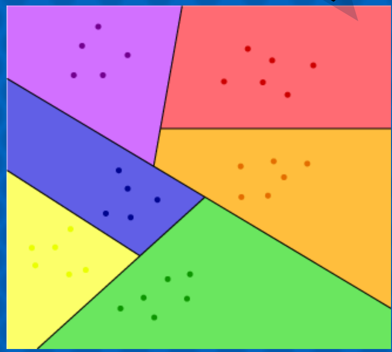


	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	[illegible]	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
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

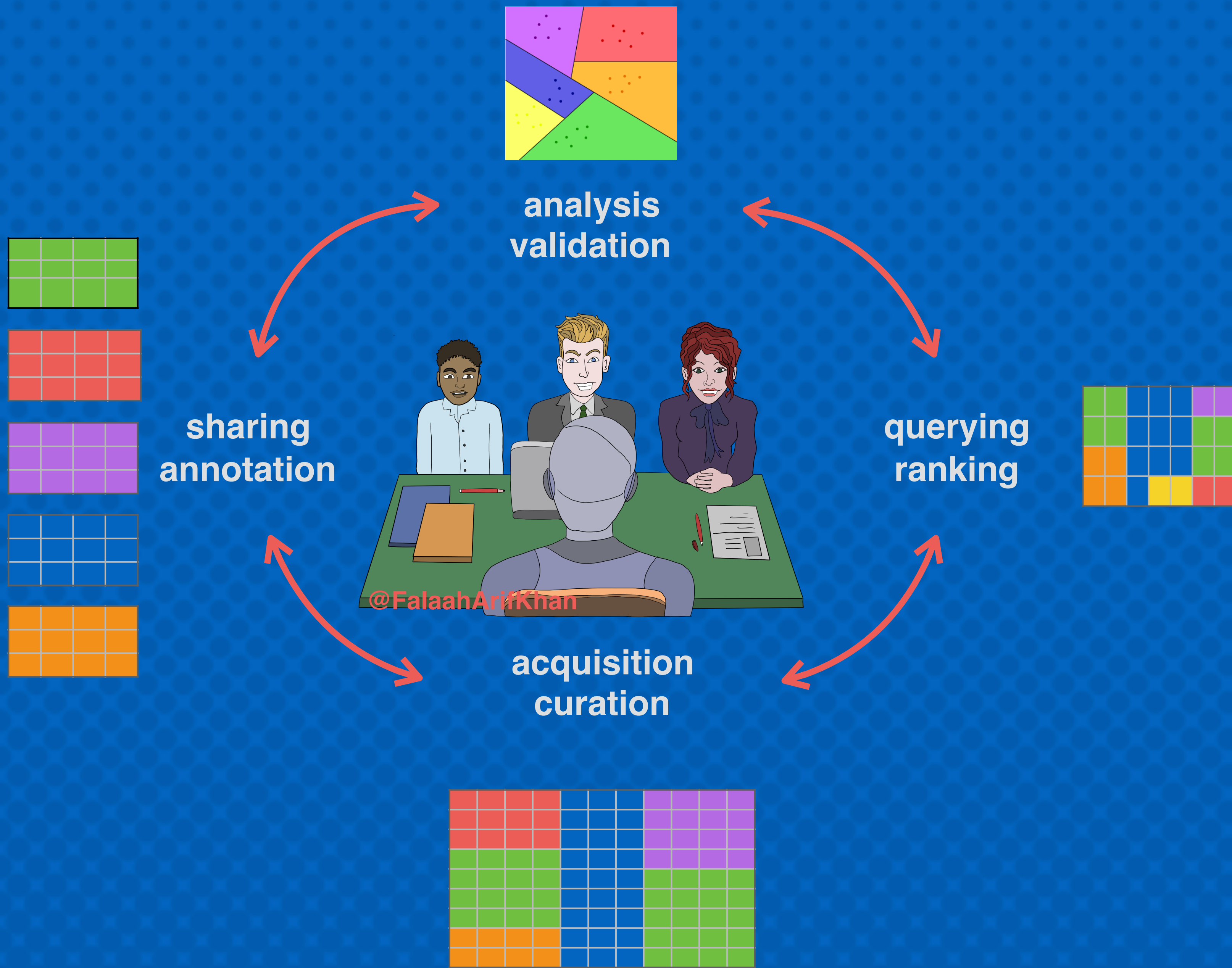
what happens inside the box?



how are results used?



Data lifecycle of an ADS



Understand your data!

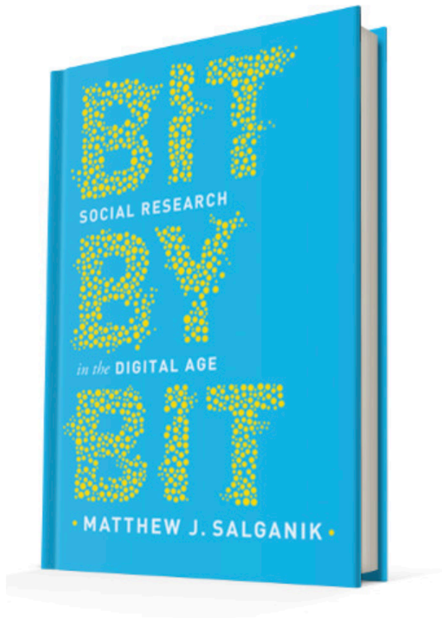
CRA
Computing Research Association

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/

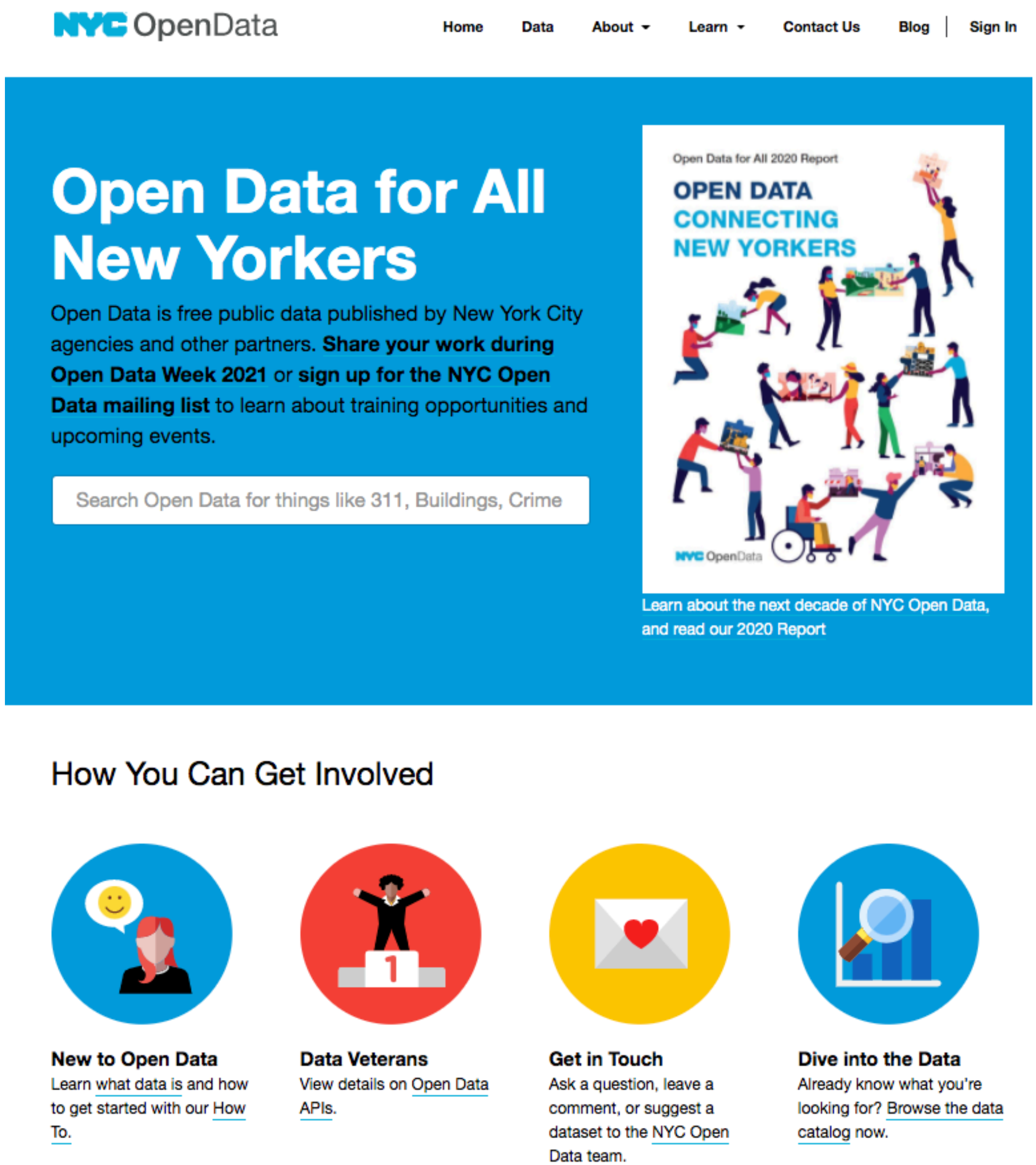
Understand your data!

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 OpenData
Home
Data
About
Learn
Contact Us
Blog
Sign In
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 2021 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
Open Data for All 2020 Report
OPEN DATA CONNECTING NEW YORKERS
NYC OpenData
Learn about the next decade of NYC Open Data, and read our 2020 Report
How You Can Get Involved
1
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.

NYC OpenData

SAT (College Board) 2010 School Level Results

Education

Dataset

freshness

Updated
April 25, 2019

Views
27,142

popularity

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.

description

Data collected and processed by the College Board.

source

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
27.1K	43.1K

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*

What's in this Dataset?

NYC OpenData

Rows: 460

Columns: 6

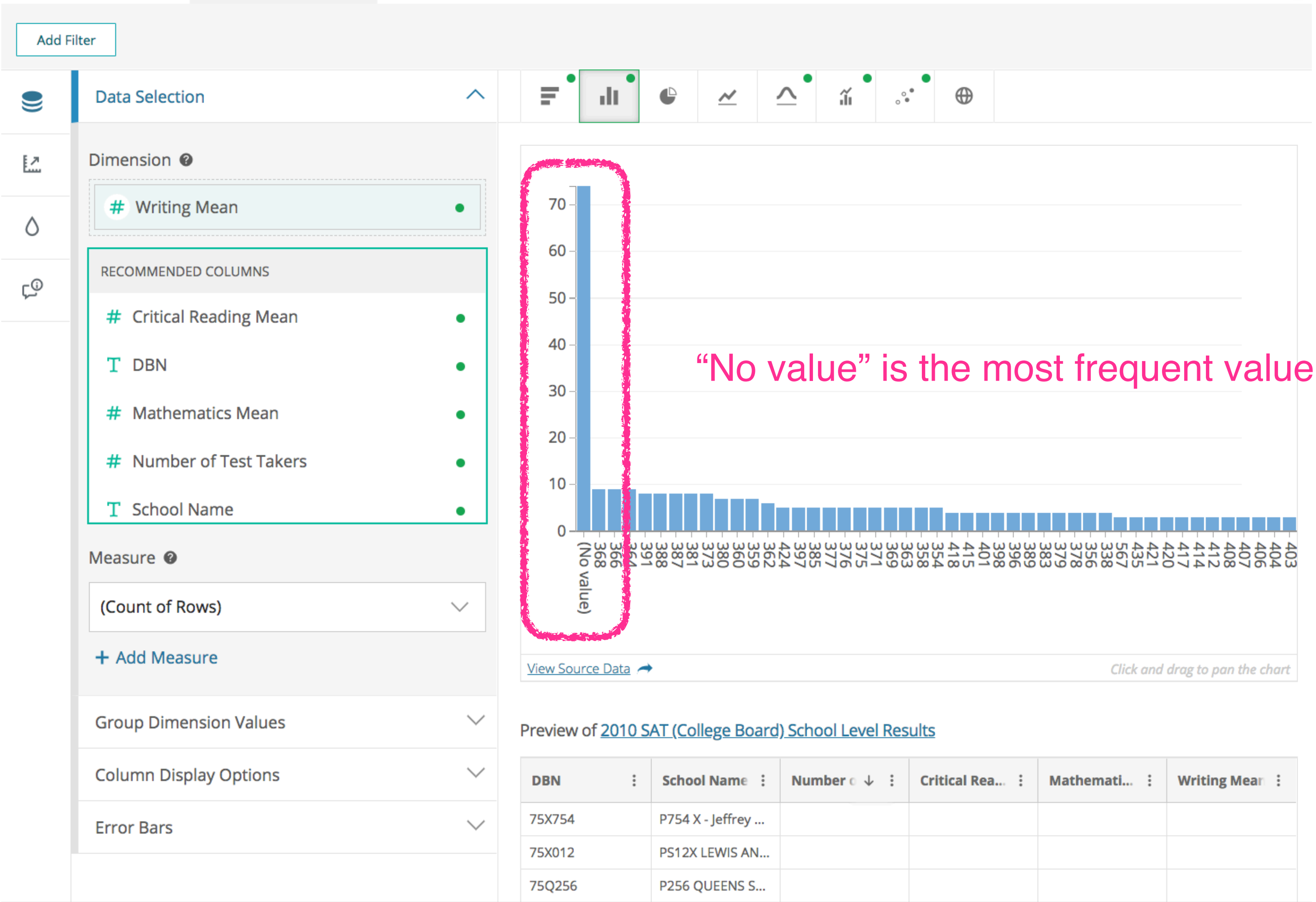
Columns in this Dataset

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 #

2010 SAT (College Board) School Level Results

Education

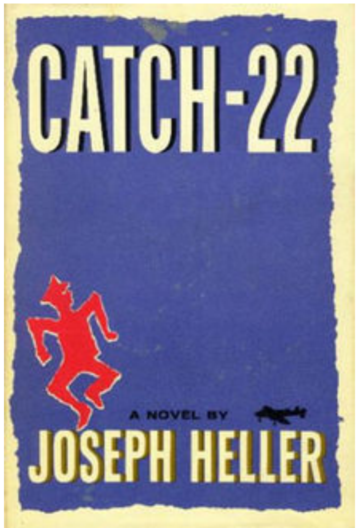
NYC OpenData



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.

https://en.wikipedia.org/wiki/Data_cleansing & 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

Data of poor quality is lacking rich metadata.

Divesh Srivastava, AT&T Research

slide by Heiko Mueller

Data cleaning

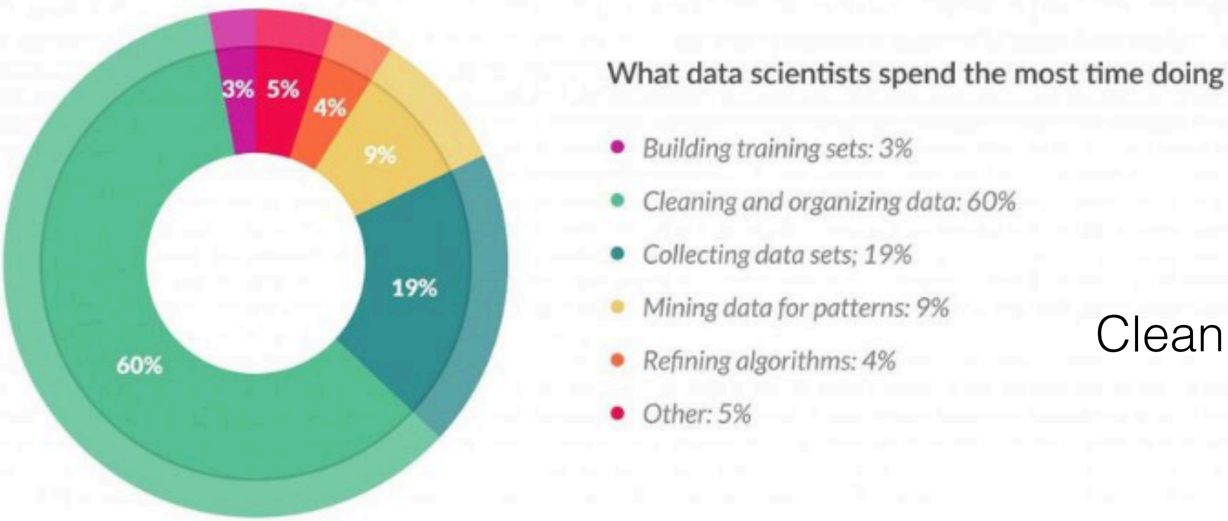
Forbes

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

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%)

slide by Heiko Mueller

Data cleaning

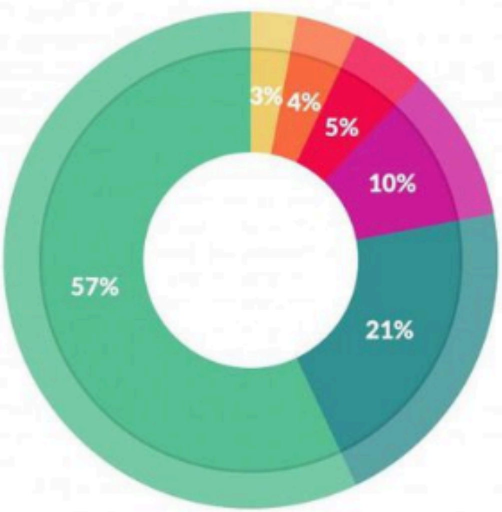
52,423 views | Mar 23, 2016, 09:33am

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's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

Find least enjoyable

Collecting data (21%)

Cleaning and organizing data (57%)

slide by Heiko Mueller

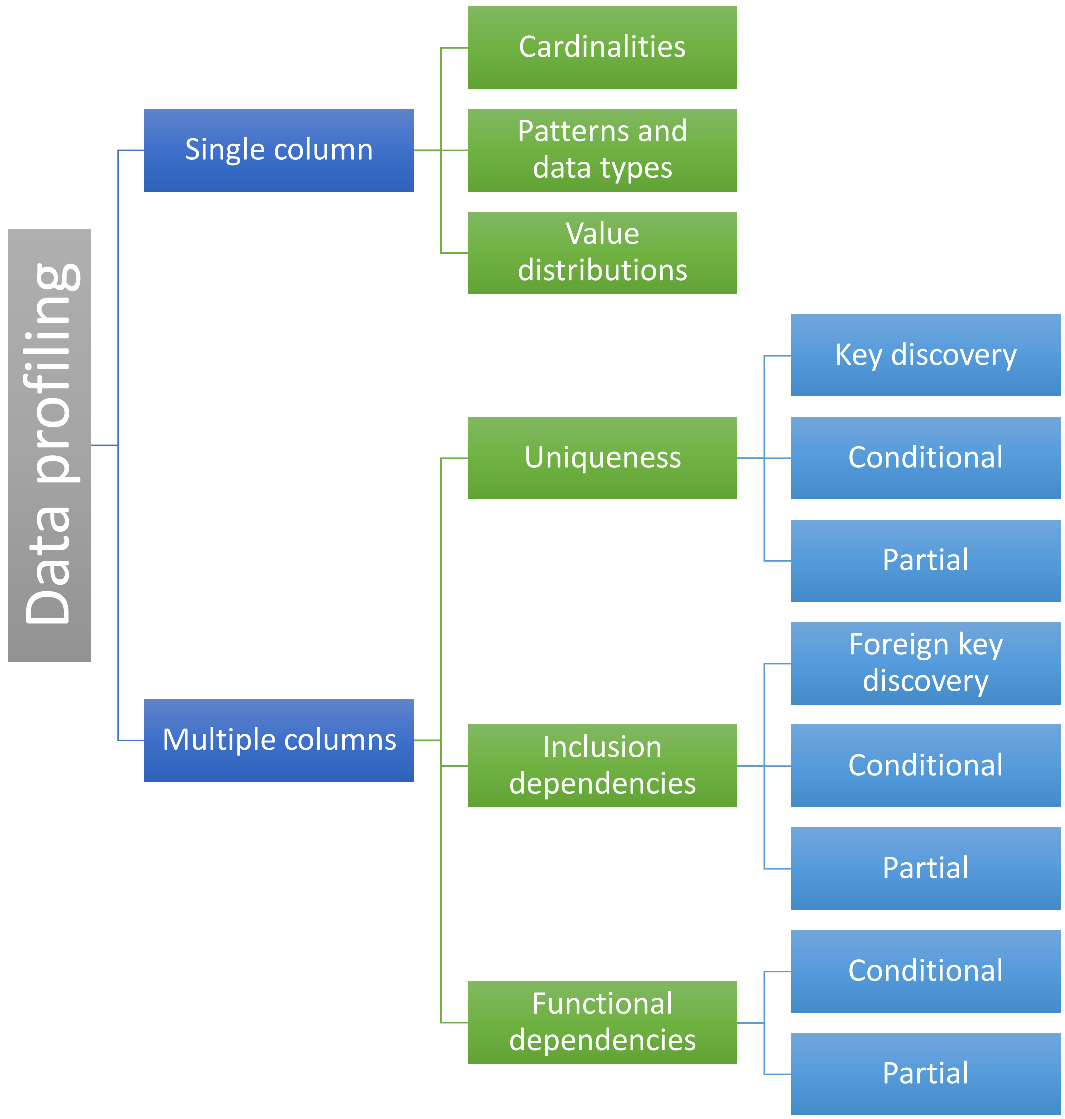

data profiling

A classification of data profiling tasks

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

	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

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

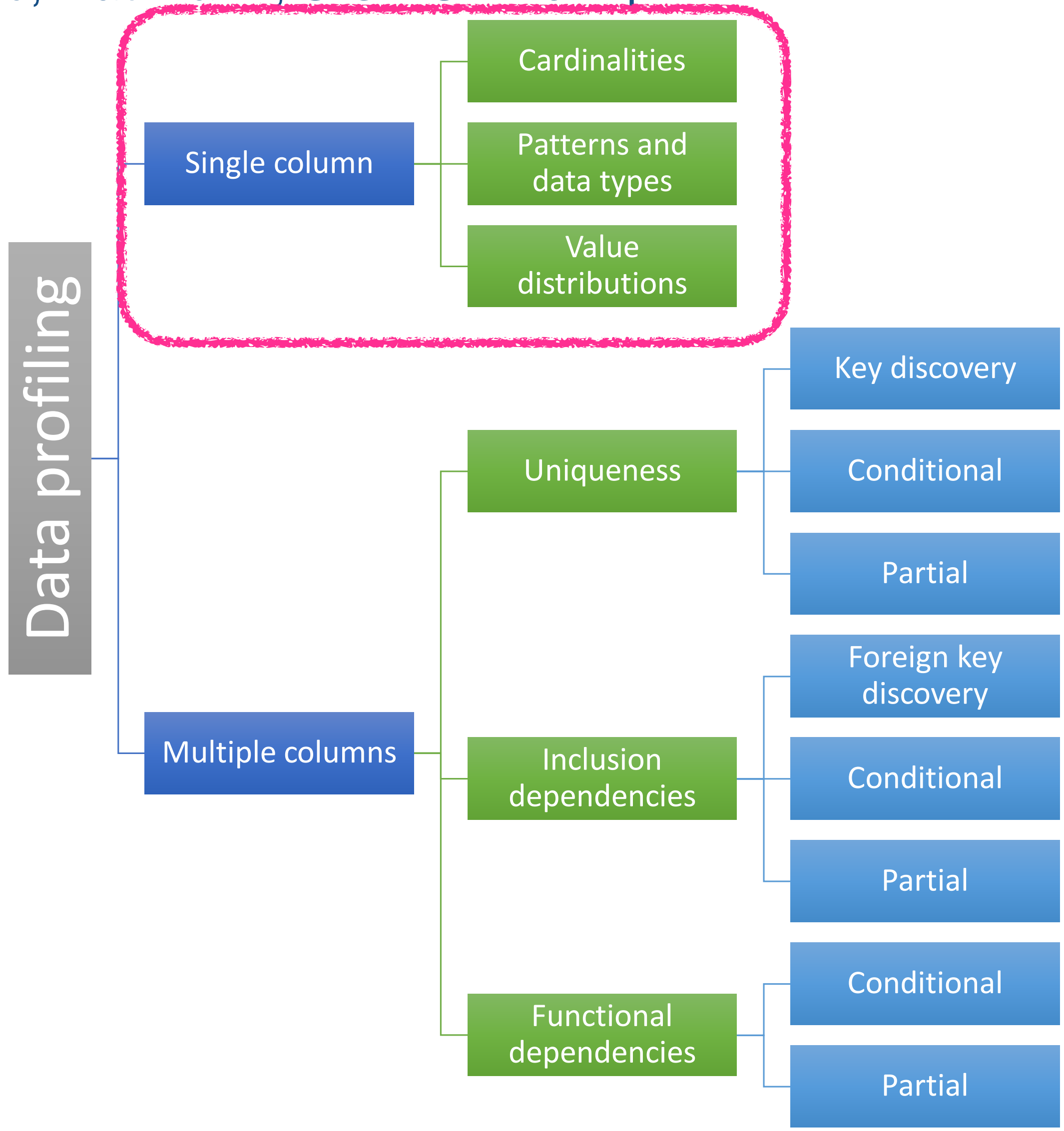
- We need both statistics and structure, they are mutually-reinforcing, and help us understand the **semantics** of the data - it's meaning

A classification of data profiling tasks

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

	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

relational data (here: just one table)



Single column: cardinalities, data types

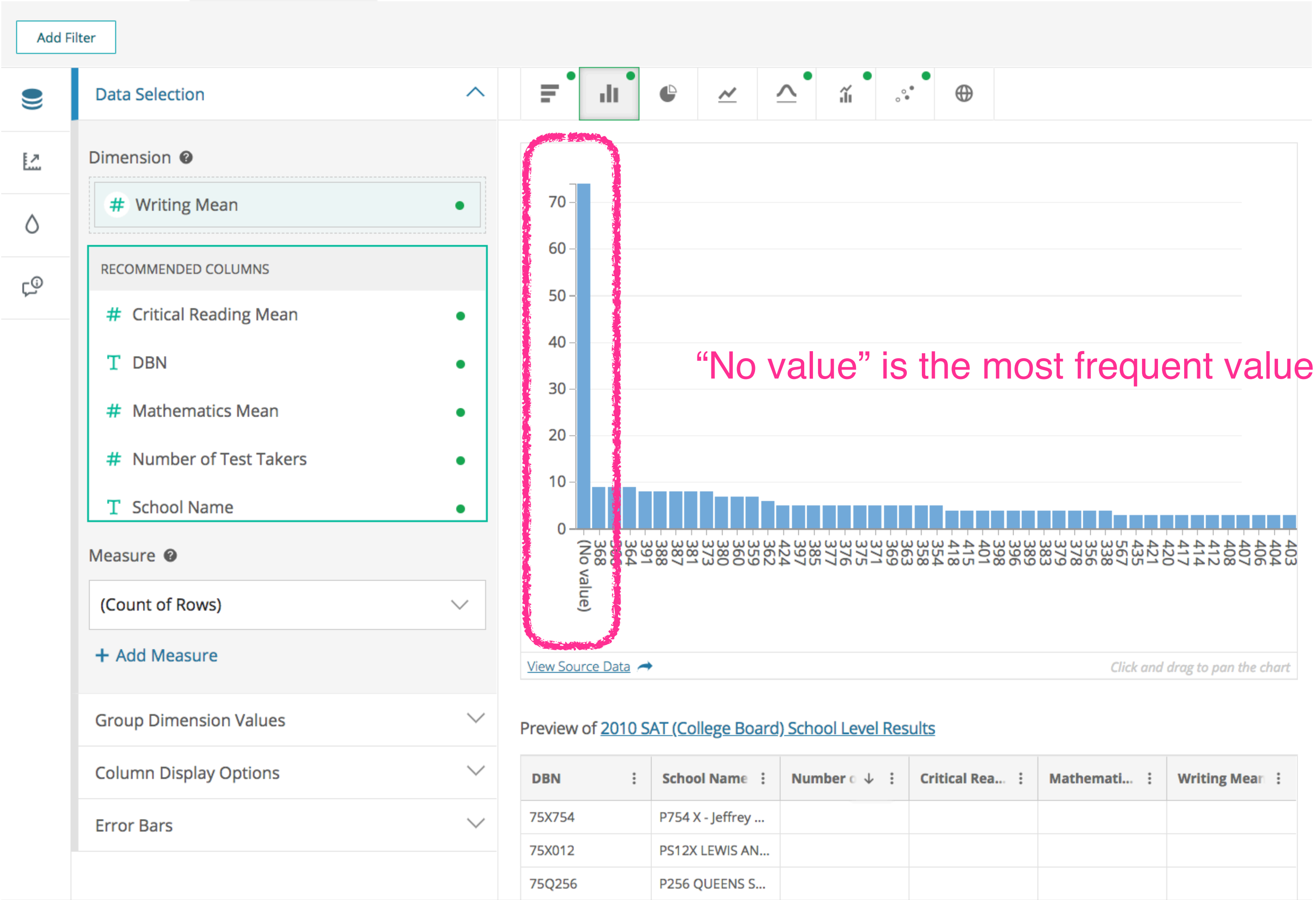
[Abedjan, Golab, Naumann; *SIGMOD 2017*]

- 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
- ….

2010 SAT (College Board) School Level Results

Education

NYC OpenData



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?

Single column: cardinalities, data types

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

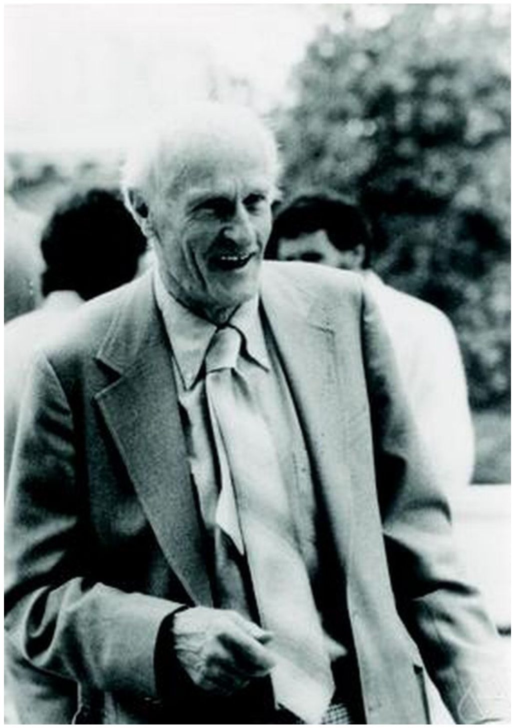
- cardinality of relation **R** - number of rows
- domain cardinality of a column **R.a** - number of **distinct** values
- attribute value **length**: min, max, average, median
- **basic data type**: string, numeric, date, time, ….
- number of percentage of **null** values of a given attribute
- **regular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

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.

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



Stephen Kleene

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

based on a slide by Heiko Mueller

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

based on a slide by Heiko Mueller

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}
```

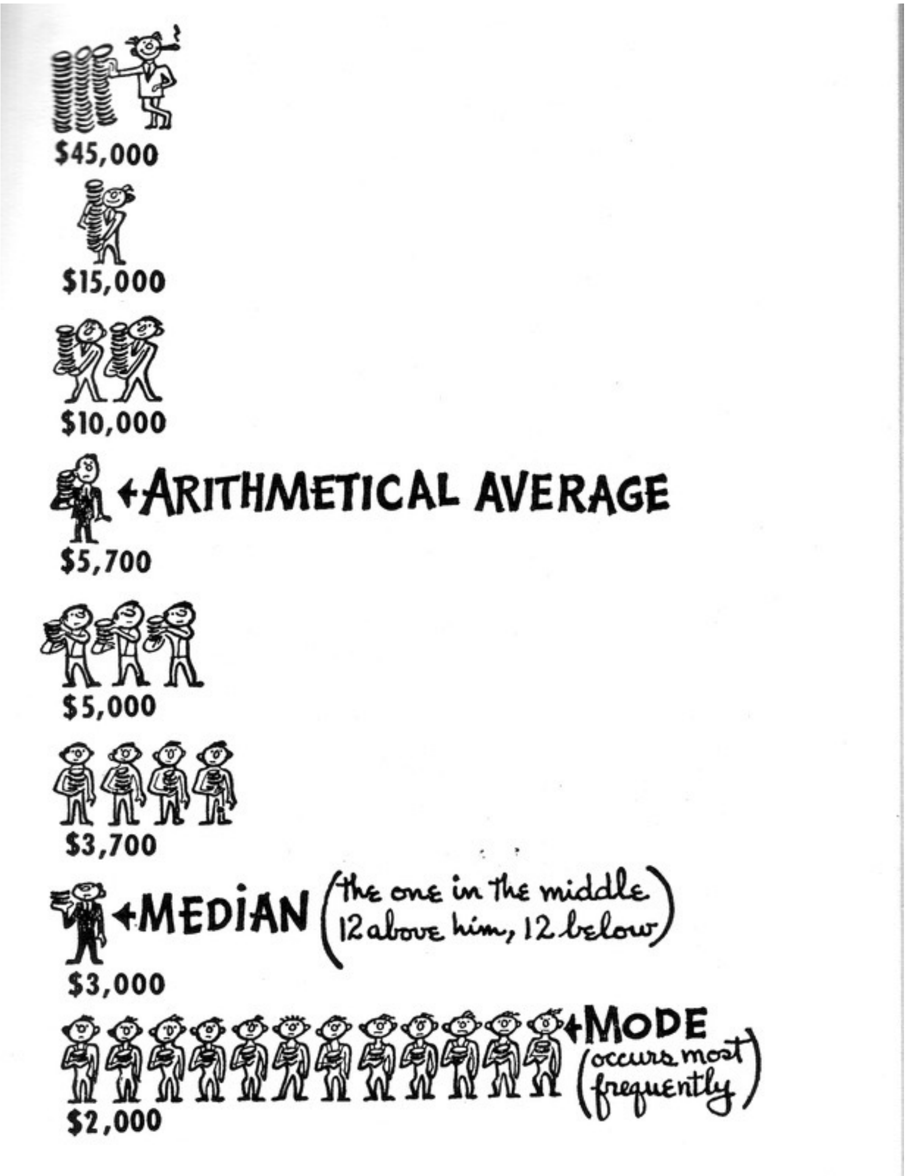
based on a slide by Heiko Mueller

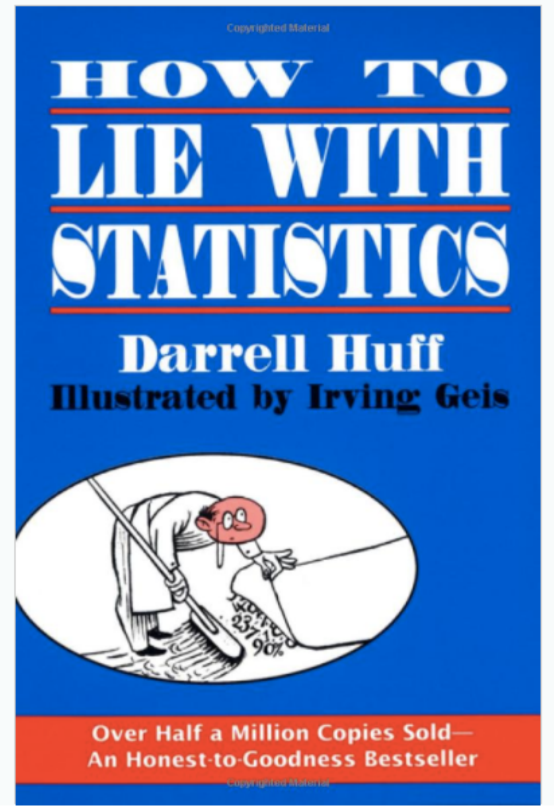
Single column: basic stats, distributions

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

- 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
- …

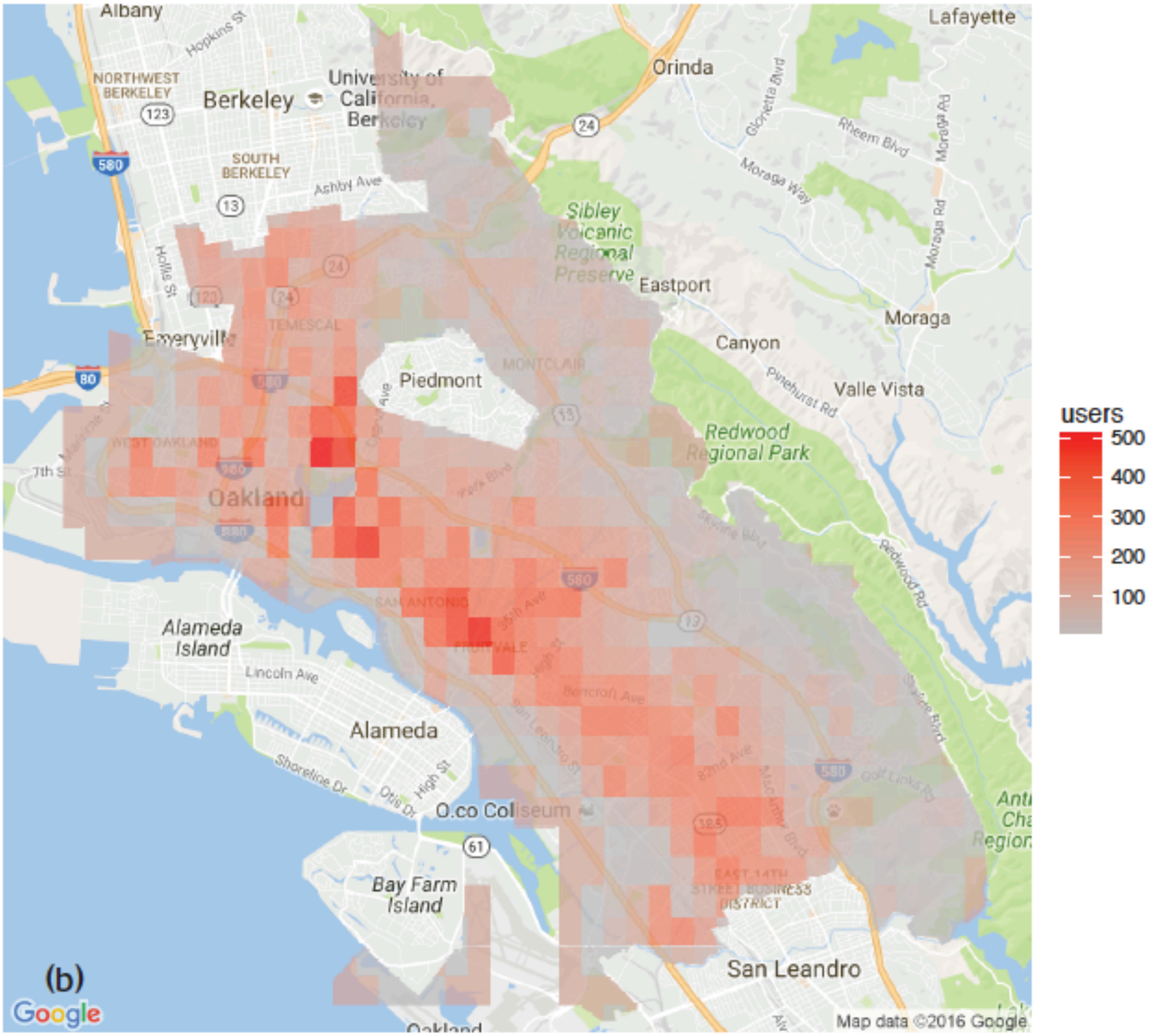
The well-chosen average



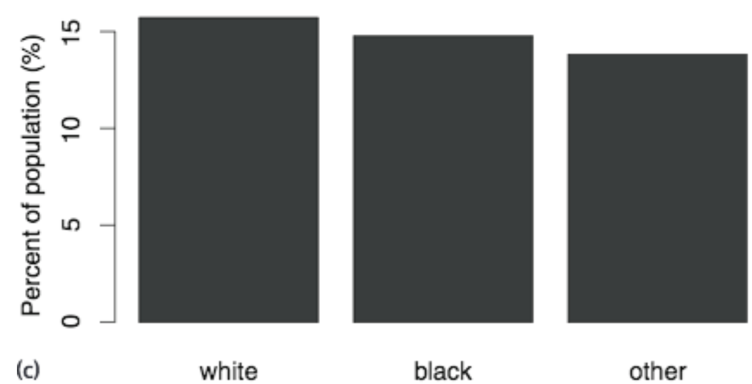


Is my data biased? (histograms + geo)

[Lum, Isaac; *Significance, 2016*]



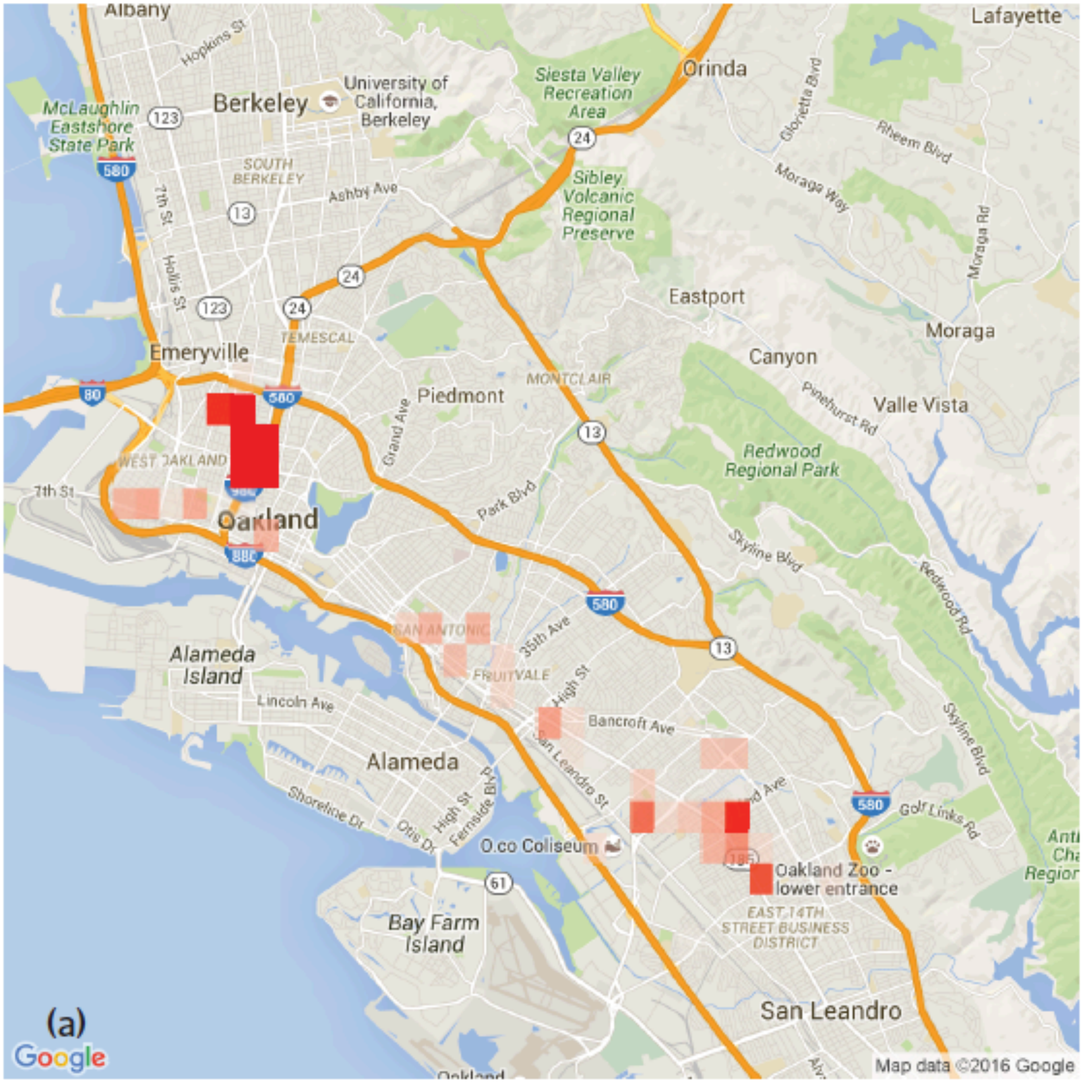
Estimated number of drug users, based on 2011 National Survey on Drug Use and Health, in Oakland, CA



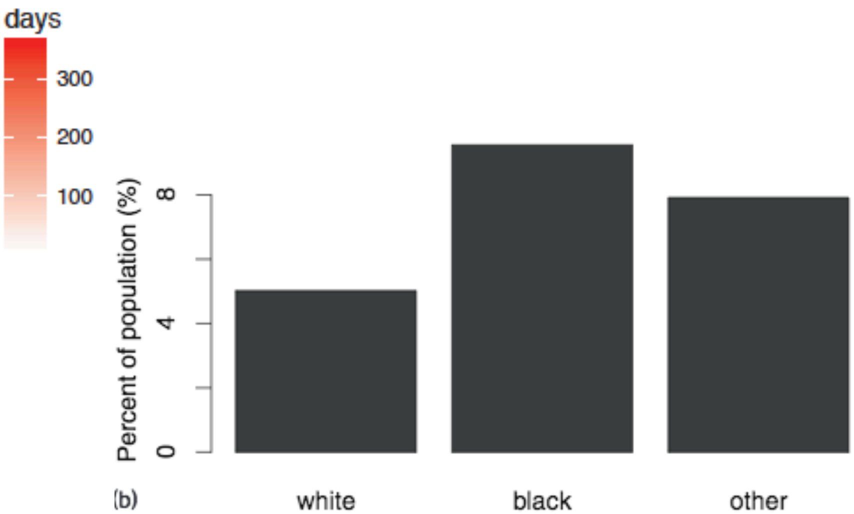
Estimated drug use by race

Is my data biased? (histograms + geo)

[Lum, Isaac; *Significance, 2016*]



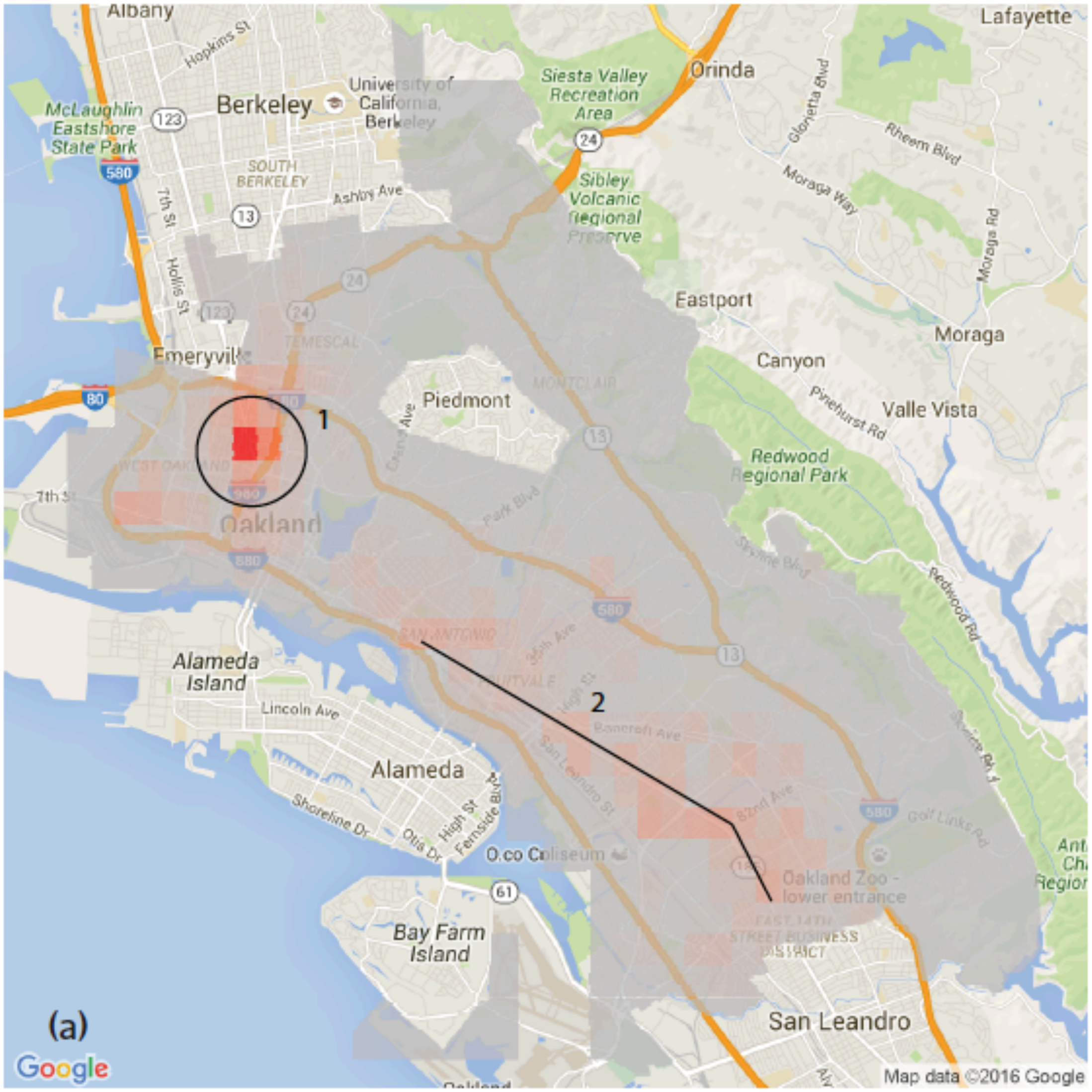
Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland, CA, police data for 2011



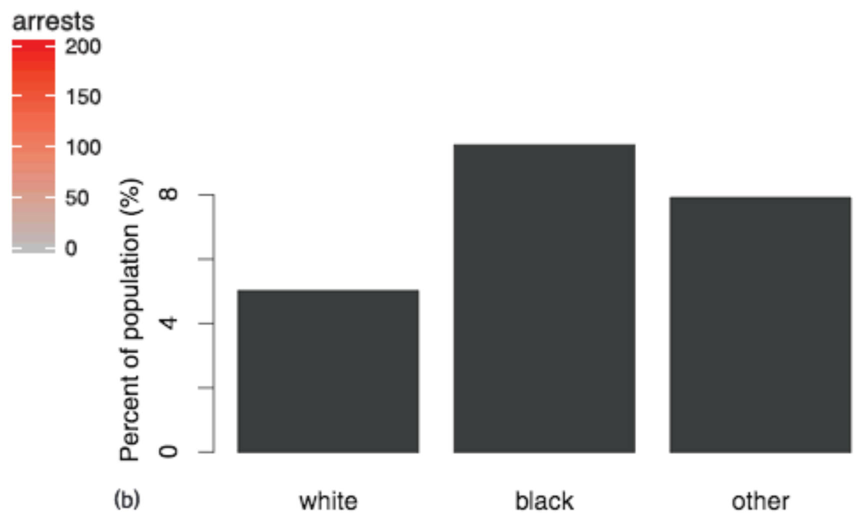
Targeted policing for drug crimes by race

Is my data biased? (histograms + geo)

[Lum, Isaac: *Significance, 2016*]



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



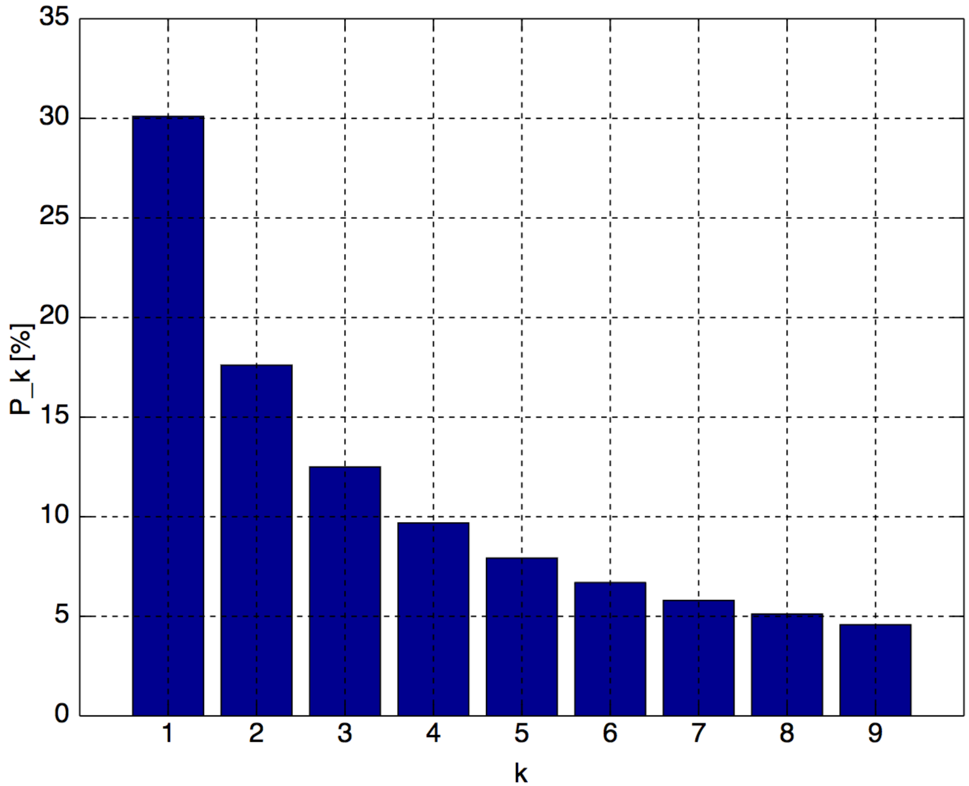
Targeted policing for drug crimes by race

Benford Law (first digit law)

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

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.

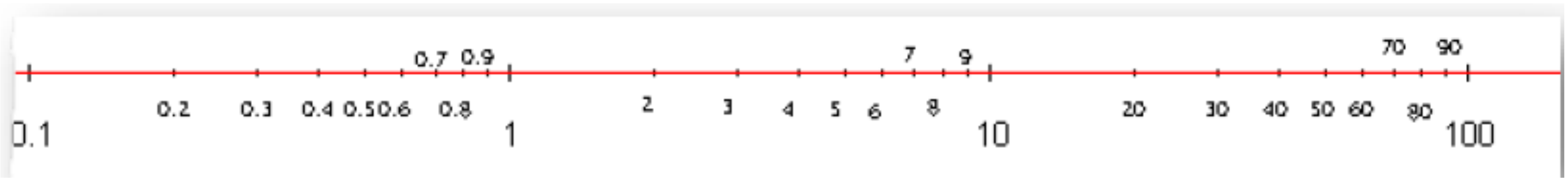
Benford Law (first digit law)

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

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)



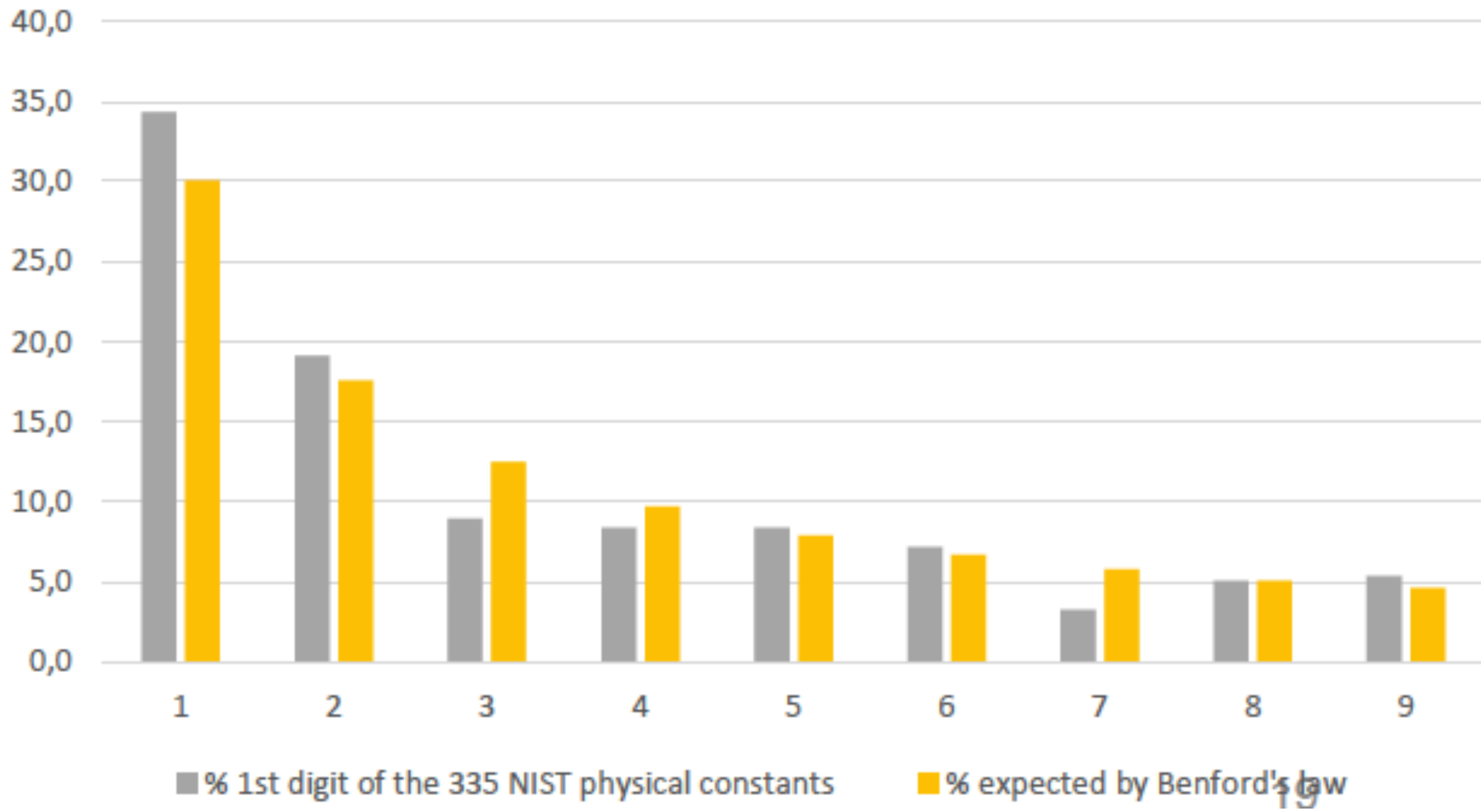
A logarithmic scale bar. Picking a random x position uniformly on this number line, roughly 30% of the time the first digit of the number will be 1.

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

Benford Law: an example

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

height of tallest structures



Benford Law: other examples

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

- 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!

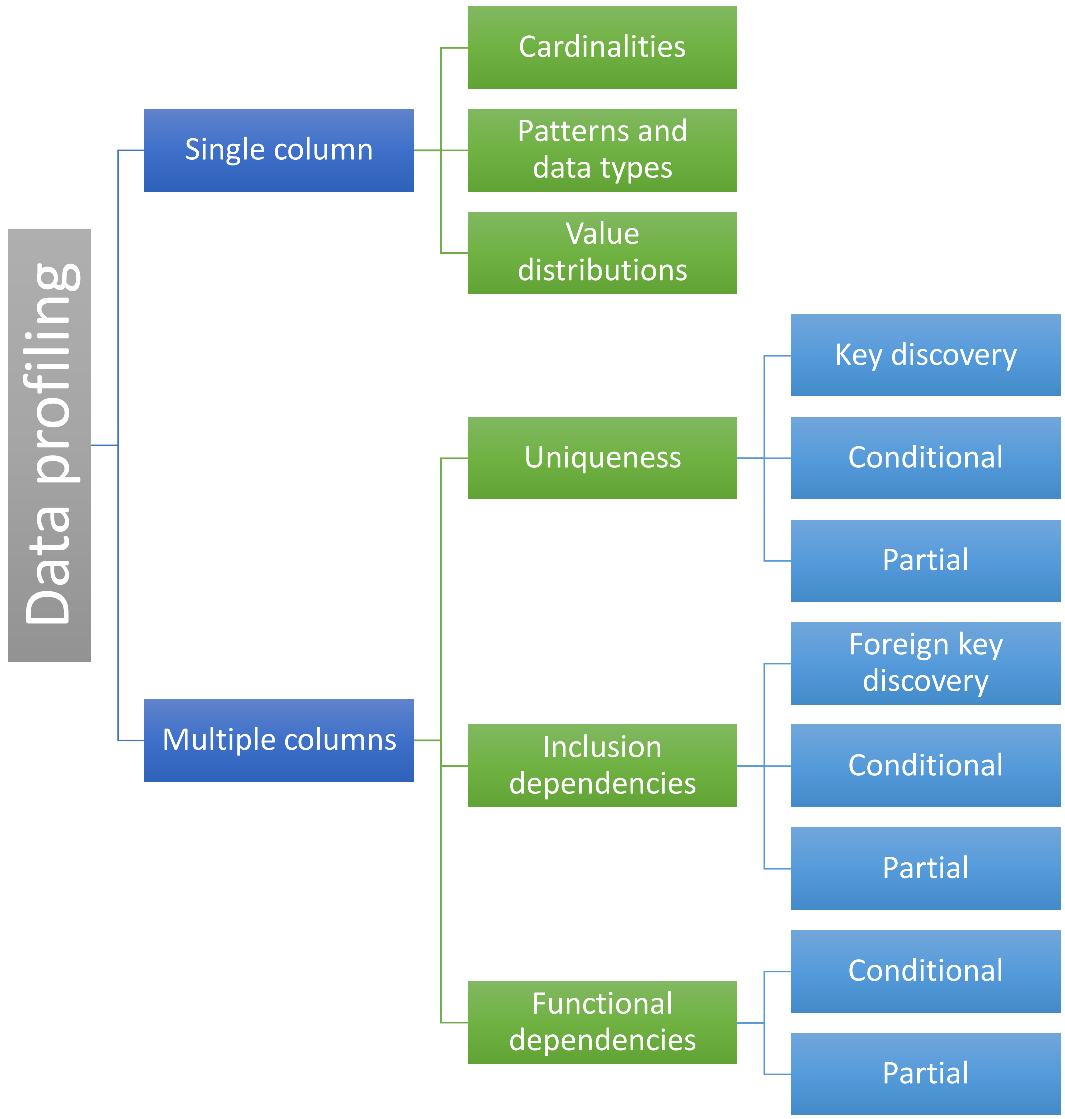
[Benford: "The law of anomalous numbers" *Proc. Am. Philos. Soc.*, *1938*]

Classification of data profiling tasks

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

	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

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 or natural) 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

next up: relational model basics