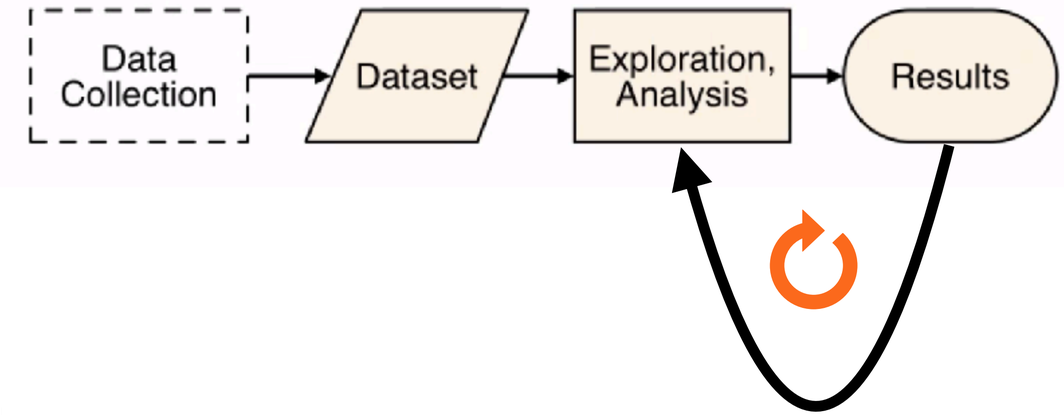
Lecture 25: Data cleaning, bias, and pitfalls

Instructor: Michael Szell
Nov 29, 2023

Today we will reflect on the data science process

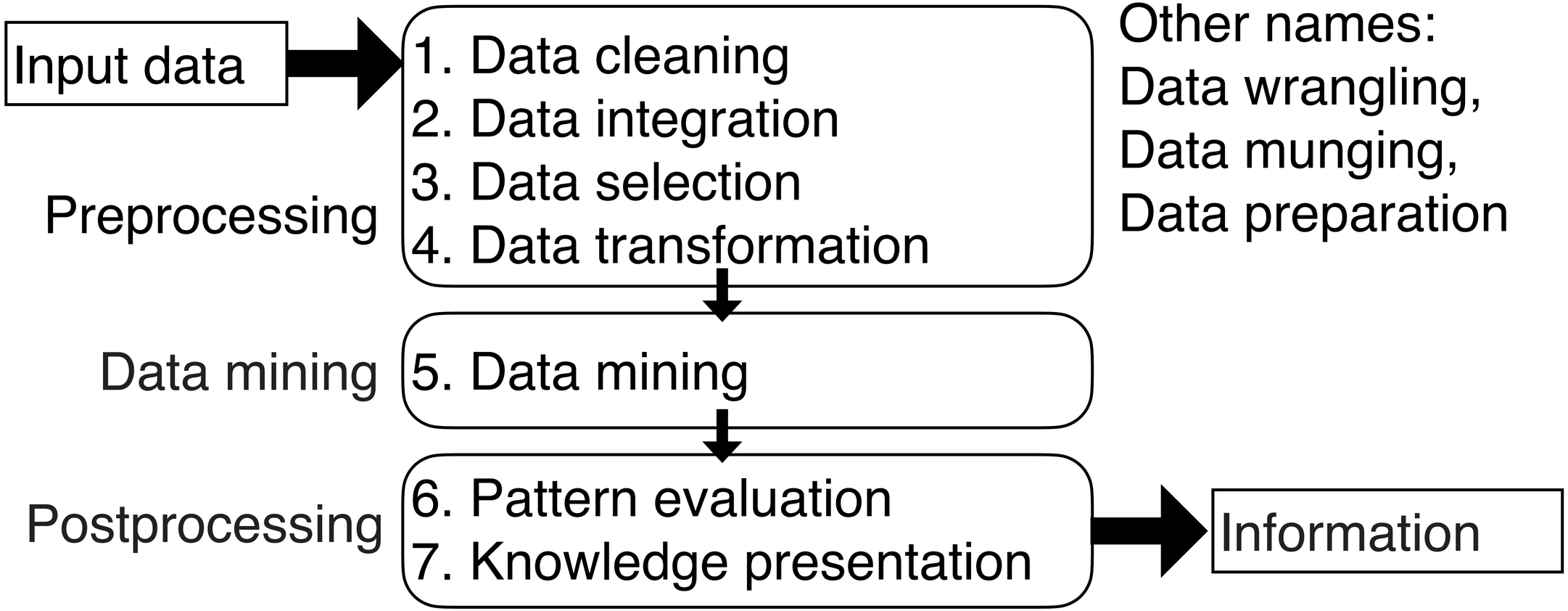
Data quality and preprocessing

Pitfalls in Data Science Applications and Research

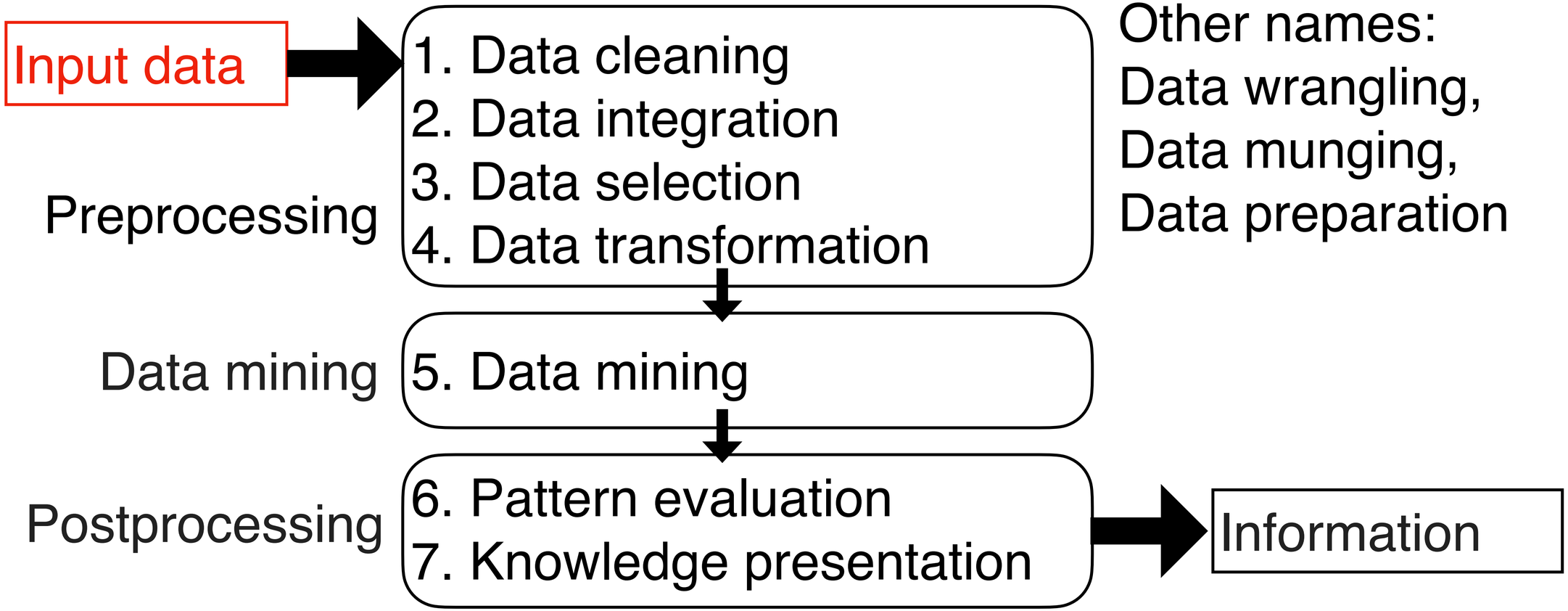


Bias

What is the most important step in data mining?



The most important step in data mining is the first



Problem 0: McNamara


U.S. ARMY

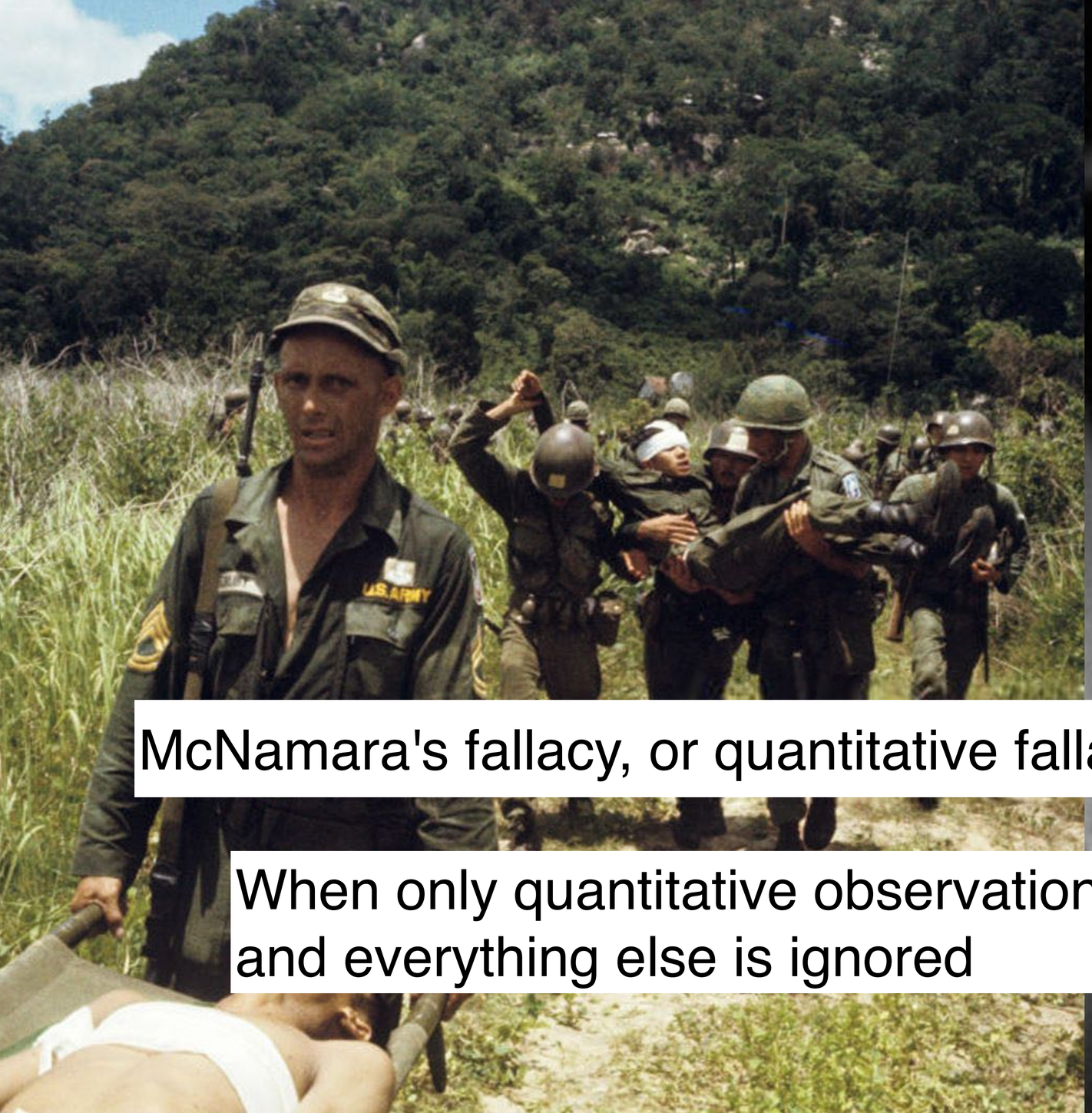
McNamara, 1960s: Dead enemies = success

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

McNamara, 1960s: Dead enemies = success

1) *Measure whatever can be easily measured.*
2) *Disregard that which can't easily be measured or given a quantitative value.*
3) *Presume that what can't be measured easily really isn't important.*
4) *Say what can't be easily measured really doesn't exist.*
   This is suicide.

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



McNamara's fallacy, or quantitative fallacy:

When only quantitative observations are considered, and everything else is ignored

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

Problem 0 in data science: McNamara fallacy

What has/has not been measured?

McNamara's fallacy, or quantitative fallacy:

> When only quantitative observations are considered,
> and everything else is ignored

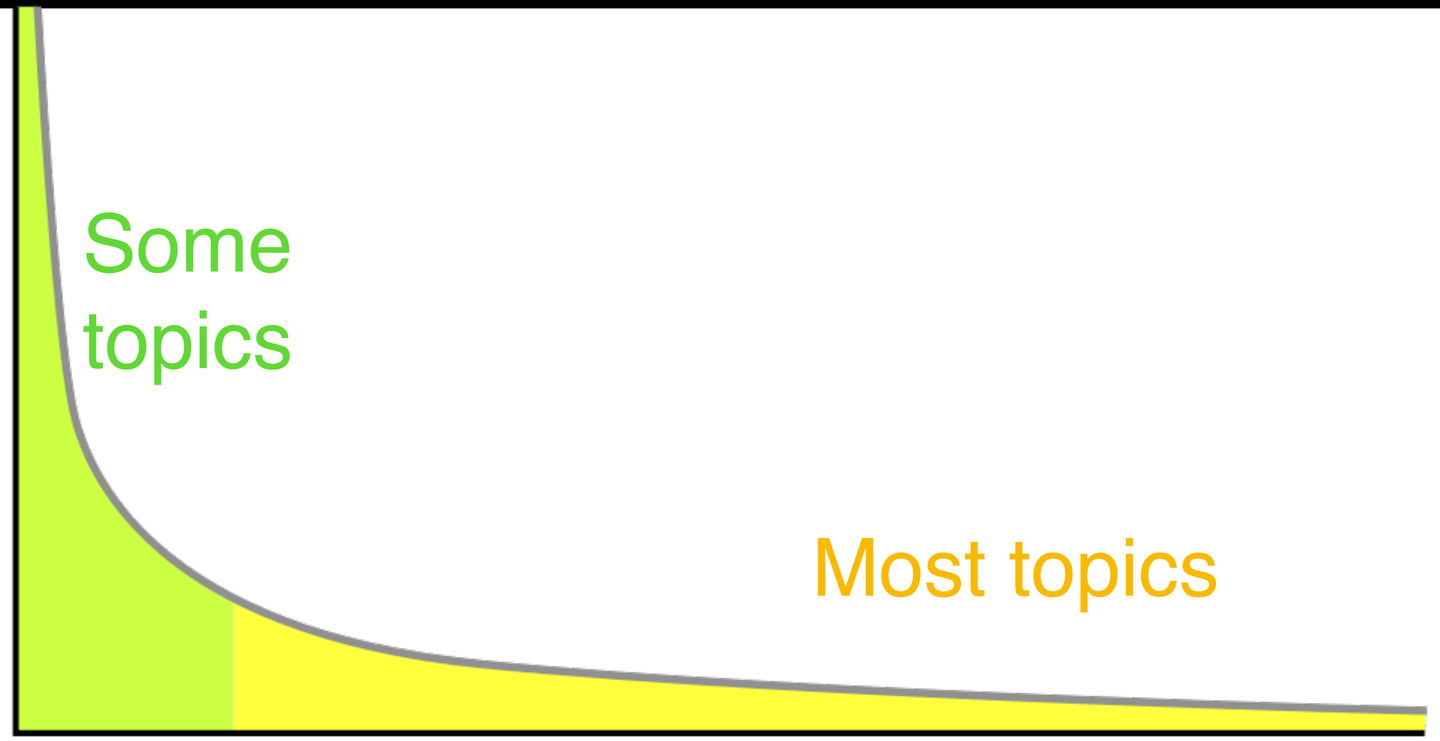
https://en.wikipedia.org/wiki/McNamara_fallacy

Problem 0 in data science: McNamara fallacy

War
Homelessness
Bicycle traffic/crashes
Domestic violence
Workplace harassment
Child abuse
....

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

Problem 0 in data science: McNamara fallacy



Facebook: Clicks
Google: GPS positions
Tesla: Driving behavior
...

Often siloed data

War
Homelessness
Bicycle traffic/crashes
Domestic violence
Workplace harassment
Child abuse
....

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


May stop quickly

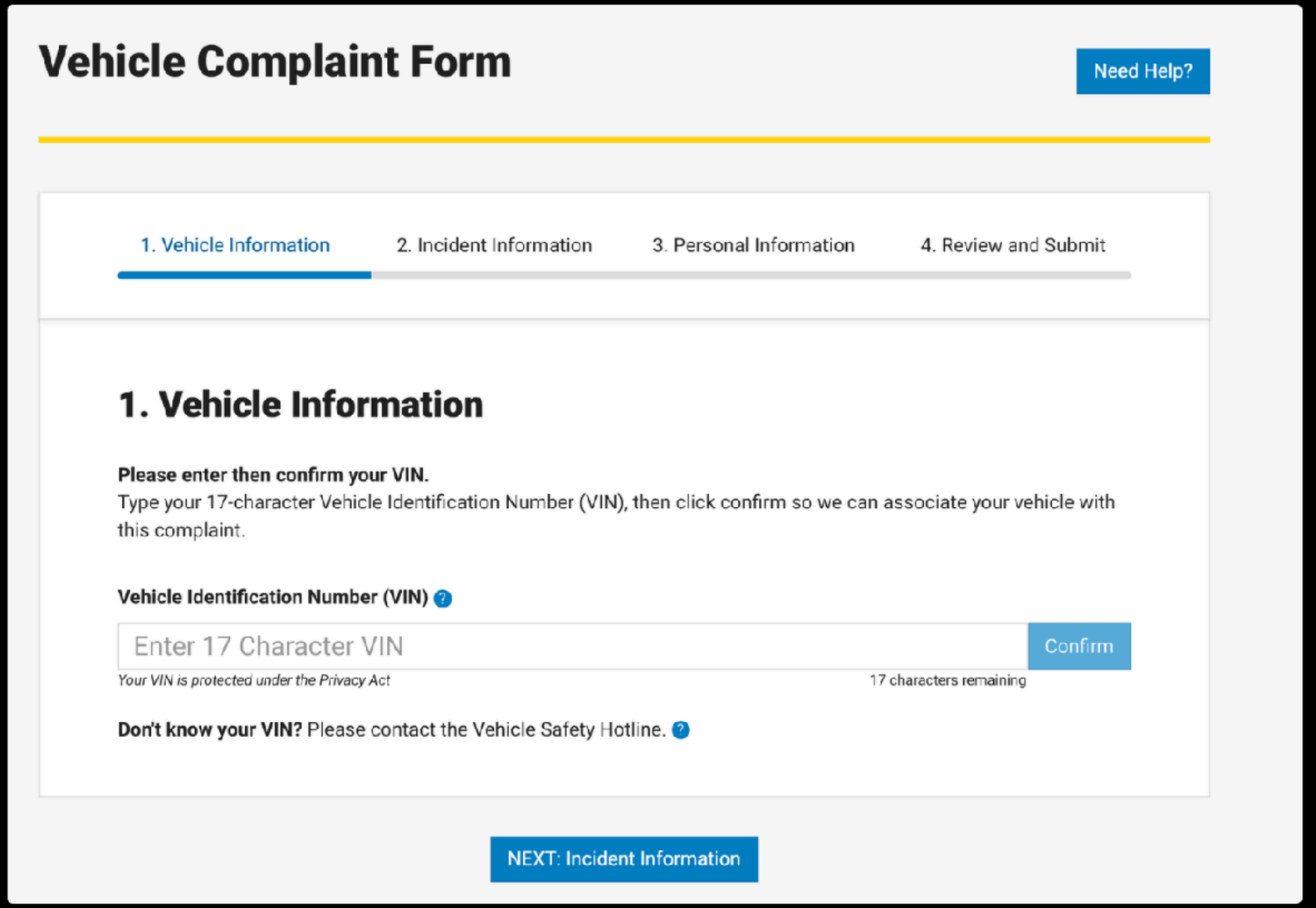
FOIA

Feds Have No Idea How Many Times Cruise Driverless Cars Hit Pedestrians

JASON KOEBLER · NOV 15, 2023 AT 1:41 PM

'Not ideal:' NHTSA is learning about driverless car incidents from Reddit videos because its safety incident reporting form is deeply flawed, internal emails show.

https://www.404media.co/feds-have-no-idea-how-many-times-cruise-driverless-cars-hit-pedestrians/



THE CURRENT "VEHICLE COMPLAINT FORM."

https://www.404media.co/feds-have-no-idea-how-many-times-cruise-driverless-cars-hit-pedestrians/

Problem 0 in data science: McNamara fallacy

When we analyze data, we analyze only
1) What has been measured
2) What is measurable

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

Problem 1: GIGO

Problem 1: GIGO: Garbage In - Garbage Out

Is the quality of my data set good? (for my problem)

Problem 1: GIGO: Garbage In - Garbage Out

GIGO is the most common reason why data science solutions fail

Your perfect algorithm

Data sets
in tutorials

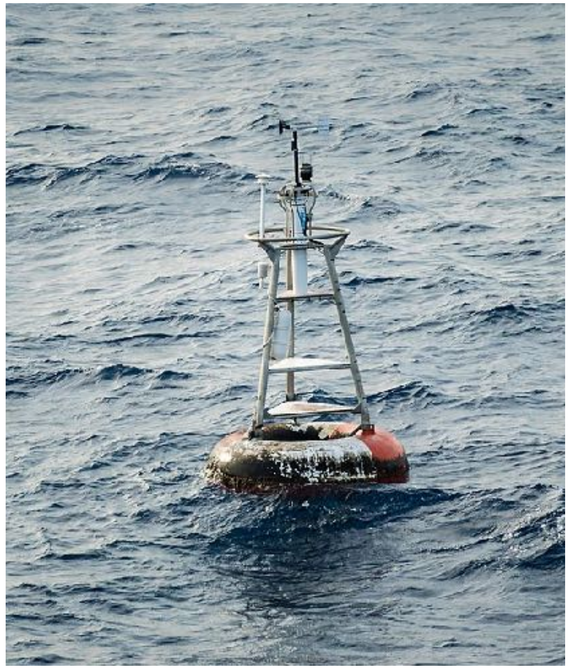
Data sets
in tutorials

Data sets in
the wild

What is: **The most Geo-tagged Place on Earth** ?

https://www.confluence.org/photo.php?visitid=14716

The most geo-tagged place on earth is Null Island



A troubleshooting country has been added with an Indeterminate sovereignty class called ***Null Island (1, 2)****. It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode failures which are routed to 0,0 by most mapping services. Aside: "Null Islands" exist for all local coordinate reference systems besides WGS84 like State Plane (and global if not using modern Greenwich prime meridian). Null Island in Natural Earth is scaleRank 100, indicating it should never be shown in mapping. Side note: Rank 30 (zoom 29 in Google speak)*

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

There are many issues affecting data quality

Data collection issues:

Recording errors

57 ways of spelling Philadelphia in one loan data set

PHIADELPHIA
PHIALDELPHIA
PHIDELPHIA
PHIELADELPHIA
PHIILADELPHIA
PHILA
PHILA.
PHILAD
PHILADALPHIA
PHILADEDLPHIA
PHILADELAPHIA
PHILADELHIA
PHILADELHPIA
PHILADELLPHIA
PHILADELOHIA
PHILADELPH
PHILADELPHA
PHILADELPHAI
PHILADELPHI
PHILADELPHIA
PHILADELPHIA PA
PHILADELPHIA,
PHILADELPHIA, PA
PHILADELPHIA`
PHILADELPHIAP
PHILADELPHIAPHIA
PHILADELPHILA
PHILADELPHIOA
PHILADELPIA

PHILADELPOHIA
PHILADELPPHIA
PHILADEPHA
PHILADEPHIA
PHILADEPHILA
PHILADEPLHIA
PHILADERLPHIA
PHILADLELPHIA
PHILADLEPHIA
PHILADLPHIA
PHILADPHIA
PHILADRLPHIA
PHILAELPHIA
PHILDADELPHIA
PHILDADLPHIA
PHILDAELPHIA
PHILDELPHIA
PHILDEPPHIA
PHILIADELPHIA
PHILIDELPHIA
PHILLA
PHILLADELPHIA
PHILLY
PHILOADELPHIA
PHLADELPHIA
PHOLADELPHIA
PHPILADELPHIA
PIHLADELPHIA

There are many issues affecting data quality

Data collection issues:

Recording errors

Duplications

Missing values

Inconsistencies

Why we generally do not use Excel in Data Science

Lack of:

- Reproducibility
- Version control
- Testing
- Maintainability
- Accuracy

https://datascience.stackexchange.com/questions/5443/do-data-scientists-use-excel

Why we generally do not use Excel in Data Science

Comment | Open Access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

127k Accesses | **45** Citations | **2567** Altmetric | Metrics

Abstract

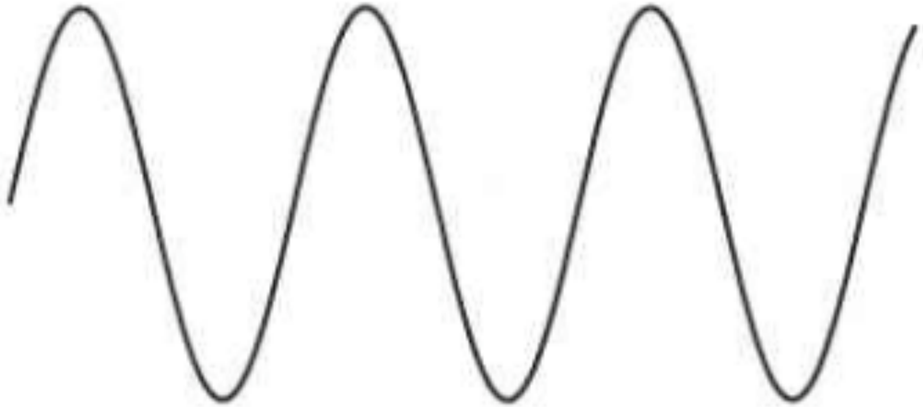
The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. '*SEPT2*' converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are frequently reused. Our aim here is to raise awareness of the problem.

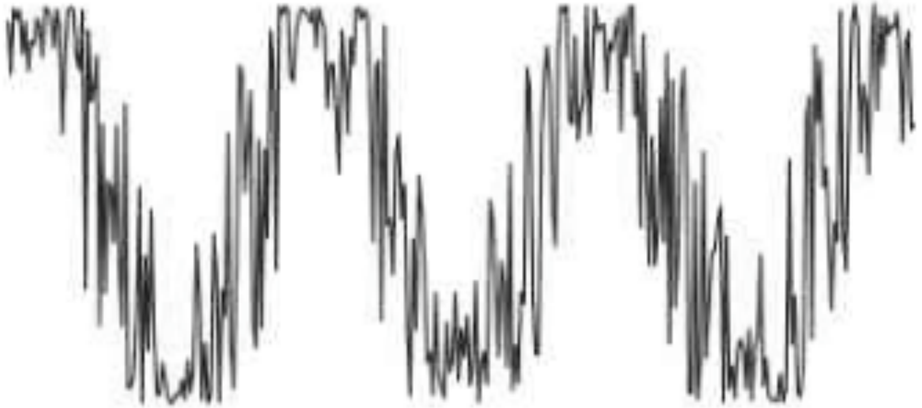
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7

There are many issues affecting data quality

Measurement errors: Noise

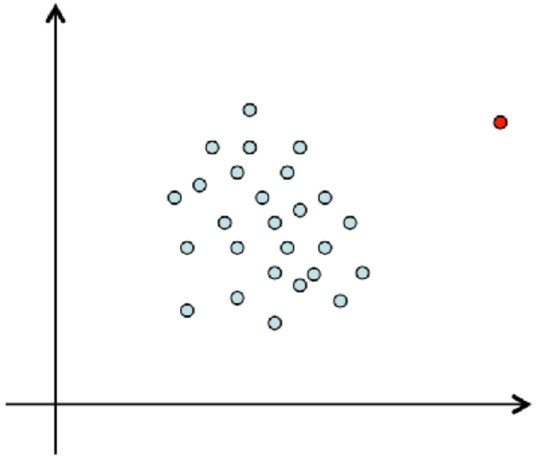


(a) Time series.

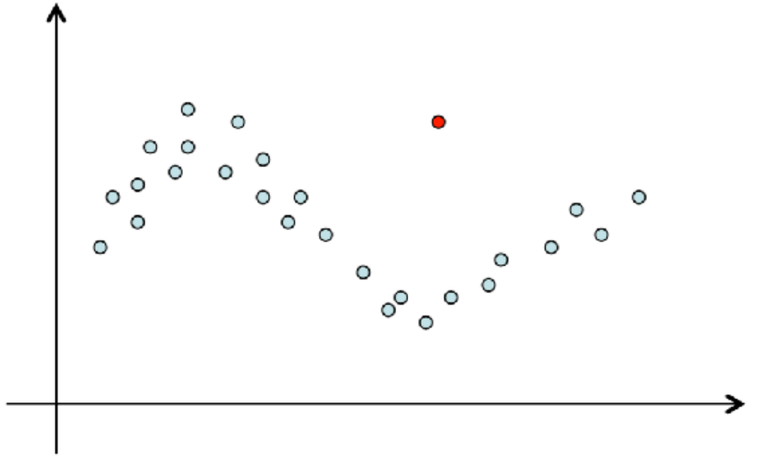


(b) Time series with noise.

There are many issues affecting data quality



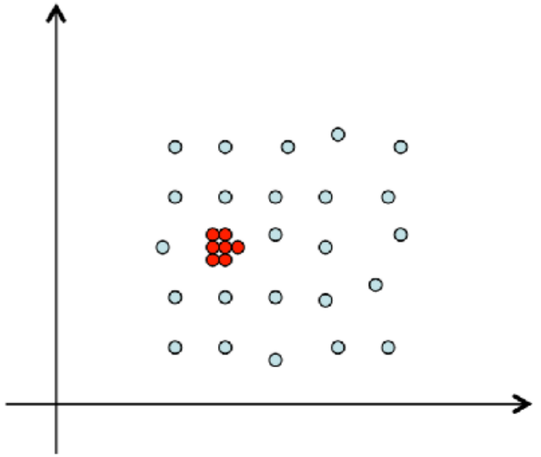
Global outliers



Contextual outliers

Outliers (anomalous objects or values):

1) Data objects that have characteristics different from most others, or
2) Values of an attribute that are unusual

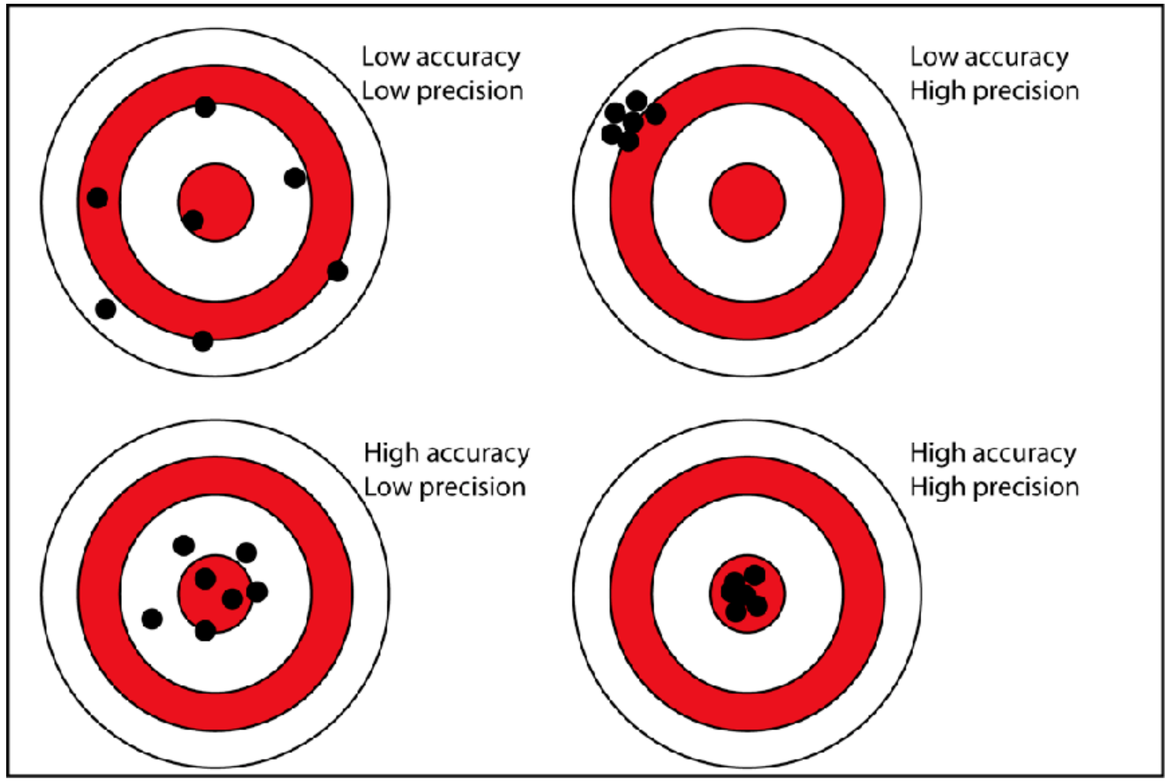


Collective outliers

Bias

There are many issues affecting data quality

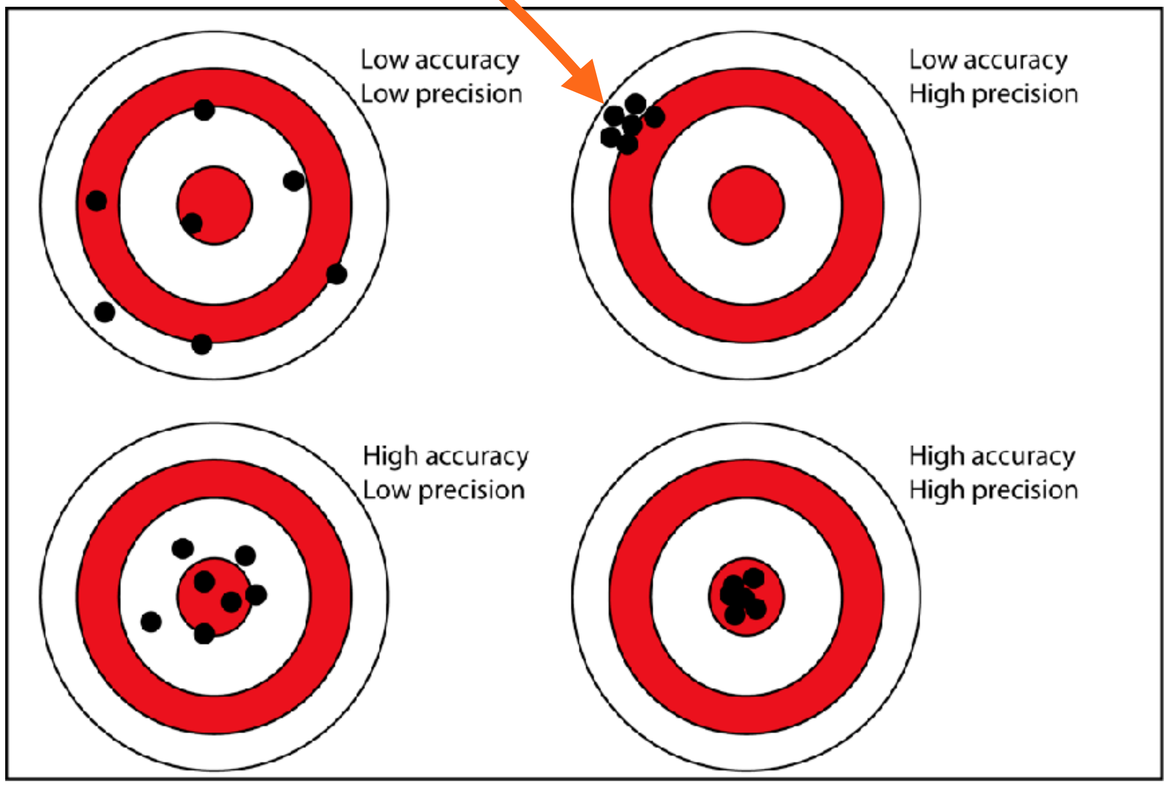
Accuracy: The closeness of measurements to the true value



Precision: The closeness of repeated measurements to one another

There are many issues affecting data quality

Bias: A systematic variation of measurements from the quality being measured



Accuracy: The closeness of measurements to the true value

Precision: The closeness of repeated measurements to one another

There are many issues affecting data quality

Biased training data biases results, Example: AI upscaling

Low-res original

There are many issues affecting data quality

Biased training data biases results, Example: AI upscaling

Low-res original

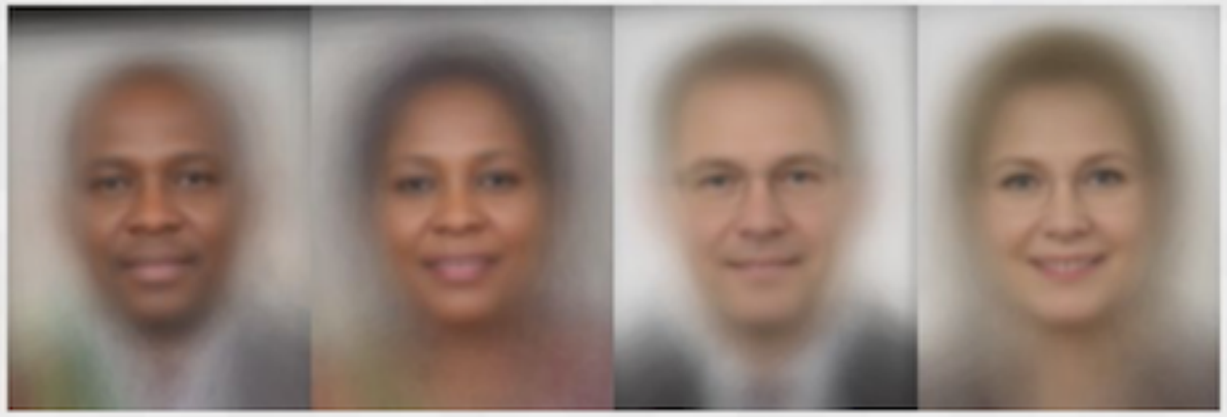
Hi-res result

https://twitter.com/Chicken3gg/status/1274314622447820801?s=20

There are many issues affecting data quality

Biased training data biases results, Example: Gender classifier

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



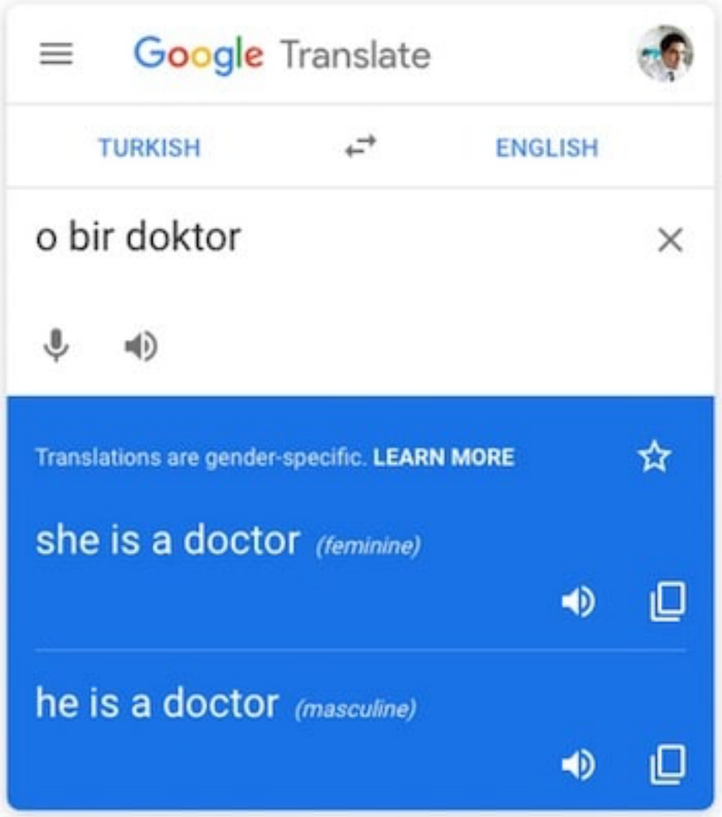
https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

There are many issues affecting data quality

Biased training data biases results, Example: Translation

Before

After



Data bias

The available data is not representative of the population or phenomenon of study

Representation

Data bias

The available data is not representative of the population or phenomenon of study	Representation
Data lacks variables that properly capture the phenomenon	Omission
Data includes content produced by humans that contains bias against groups of people	Human biases

Data bias

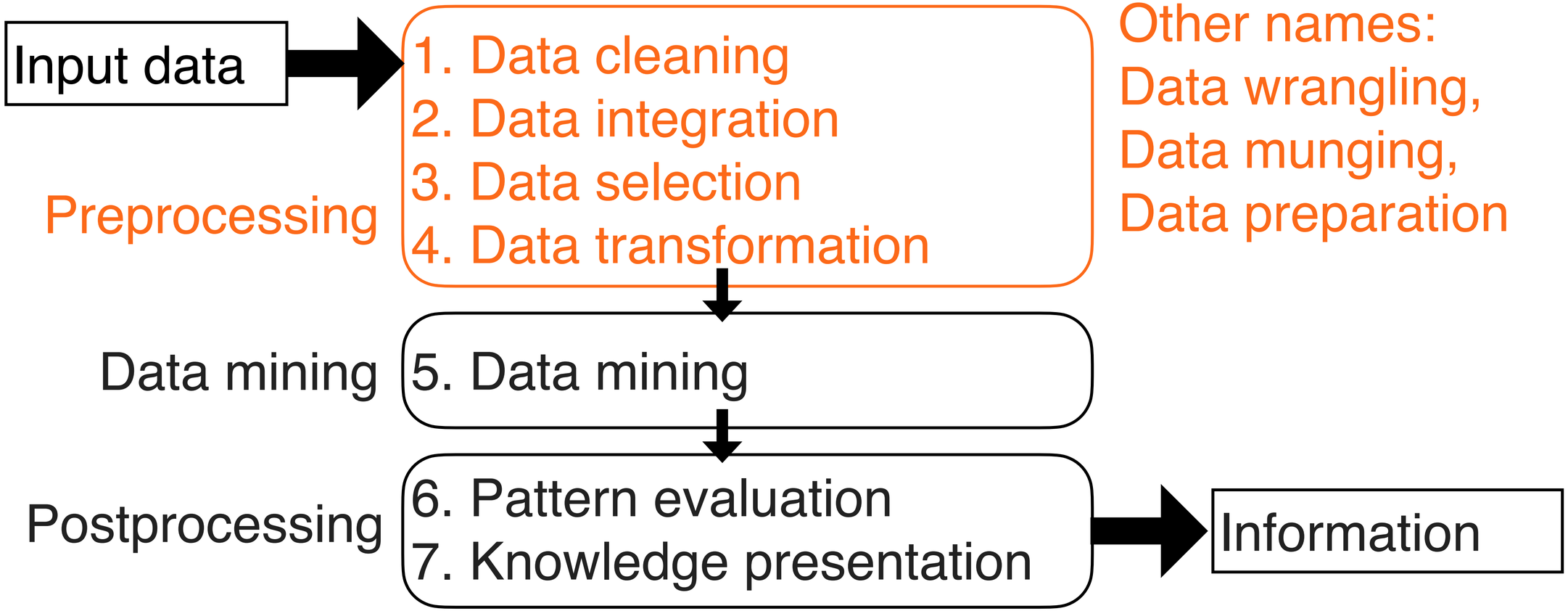
All data sets are biased

Data bias

All data sets are "political"

There is always somebody who decides to collect (or to NOT collect) some data

The second most important step in data analysis is preprocessing



The most common steps in Data Preprocessing are:

Aggregation

Sampling

Dimensionality reduction

Discretization

Variable transformation

Aggregation = Combining objects into a single one

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

NULL	Non-Freshman	3.375

Aggregation = Combining objects into a single one

Examples:

GPS coordinate → Zip Code → City → Country

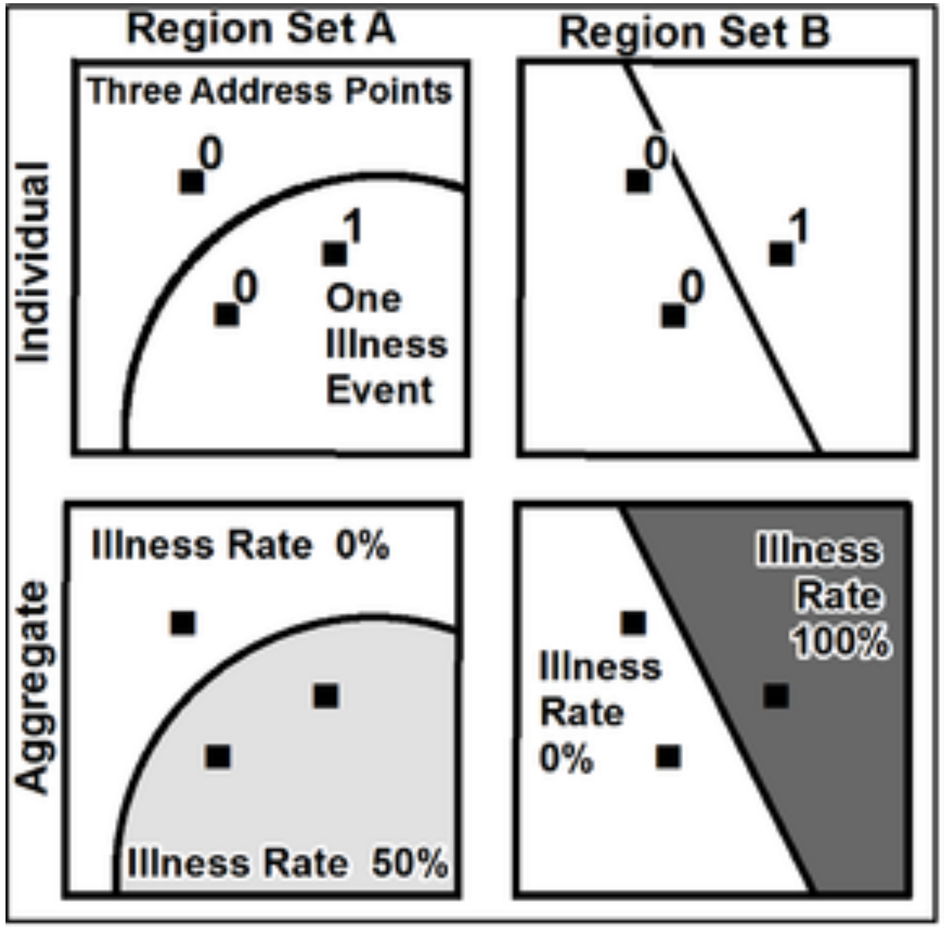
Second → Minute → Hour → Day → Week → Month → Year

Advantages: Data reduction, easier to process, high-level view, smaller statistical fluctuations

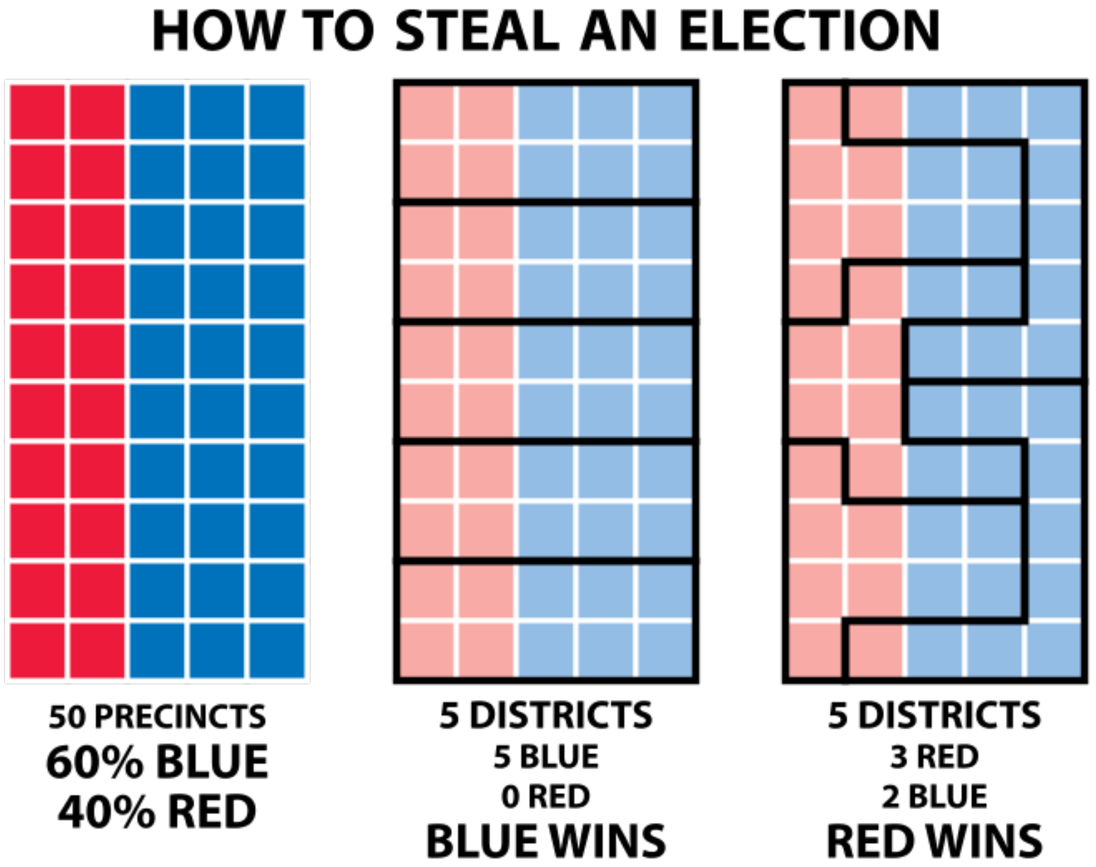
Disadvantages: Loss of details, introducing biases

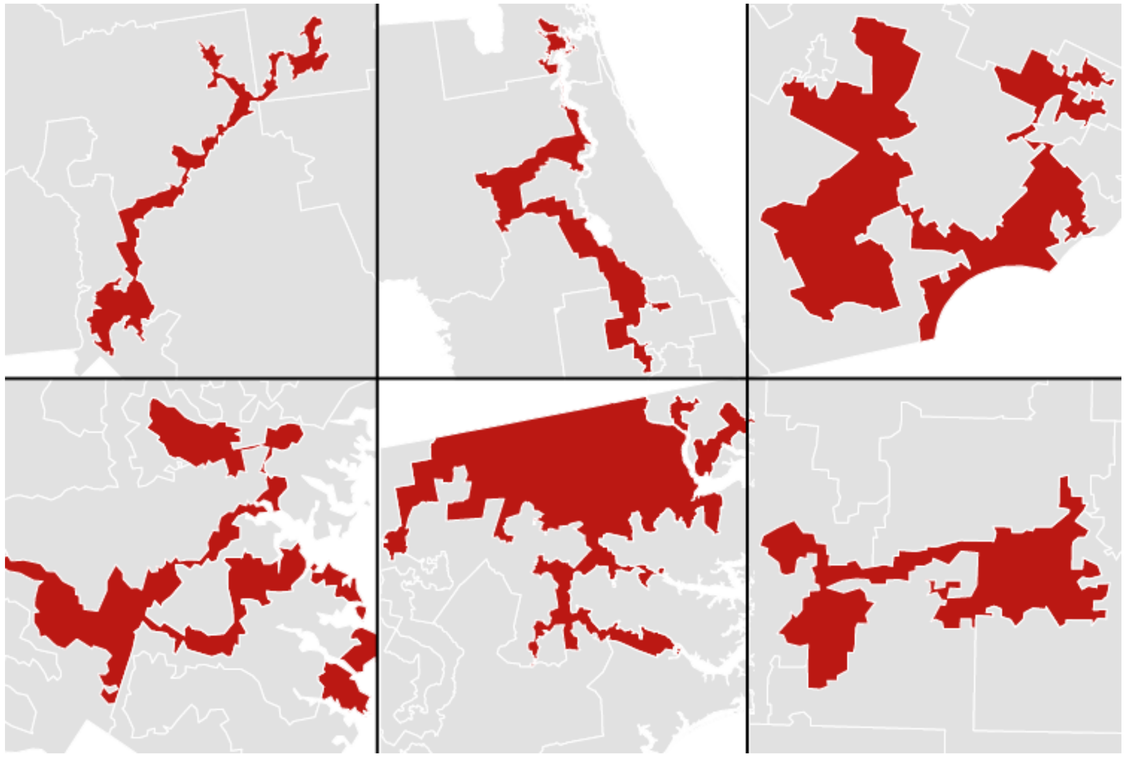
A common bias in spatial aggregation is the MAUP

Modifiable Areal Unit Problem (MAUP)



The MAUP is abused for Gerrymandering





Sampling = Leaving out records

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
~~1034262~~	~~Senior~~	~~3.24~~	~~...~~
1052663	Sophomore	3.51	...
~~1082246~~	~~Freshman~~	~~3.62~~	~~...~~
	⋮		

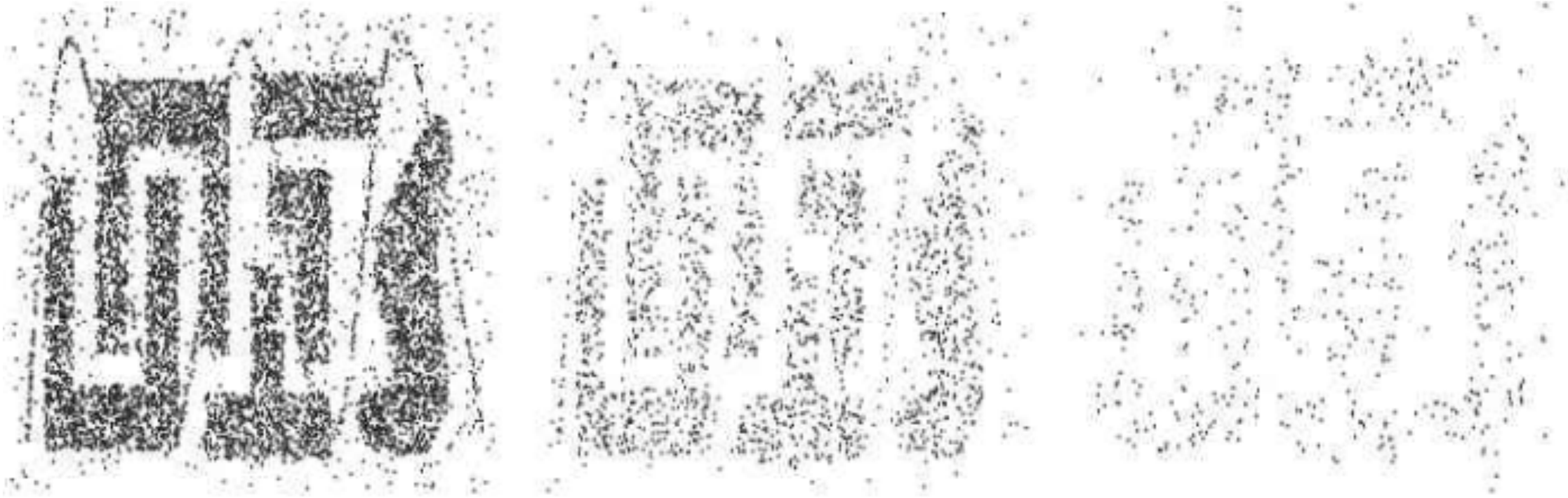
Sampling = Leaving out records

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
~~1034262~~	~~Senior~~	~~3.24~~	~~...~~
1052663	Sophomore	3.51	...
~~1082246~~	~~Freshman~~	~~3.62~~	~~...~~
	⋮		

Done if too expensive or time consuming to process all the data.

Different from statistics, where sampling is done because obtaining the entire data set is not feasible.

The sample must be representative



(a) 8000 points (b) 2000 points (c) 500 points

The sample must preserve the same properties of interest as the original data set

Dimensionality reduction = reducing the attributes

Student ID	Year	Grade Point Average (GPA)	..
	⋮		
1034262	Senior	3.24	..
1052663	Sophomore	3.51	..
1082246	Freshman	3.62	..
	⋮		

Feature subset selection = Selecting a subset of attributes

Q: If you have n attributes, how many possible subsets are there?

Dimensionality reduction = reducing the attributes

Student ID	Year	Grade Point Average (GPA)	..
	⋮		
1034262	Senior	3.24	..
1052663	Sophomore	3.51	..
1082246	Freshman	3.62	..
	⋮		

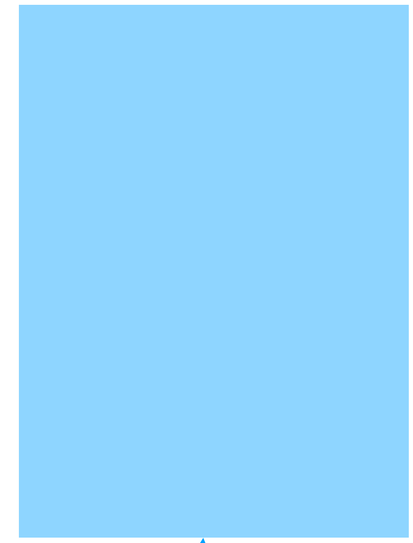
Feature subset selection = Selecting a subset of attributes

Q: If you have n attributes, how many possible subsets are there?

A: 2^n

Dimensionality reduction = reducing the attributes

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		



Principle Components Analysis (PCA) = Make a new attribute from a linear combination of old ones

Discretization = Transforming continuous into categorical

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

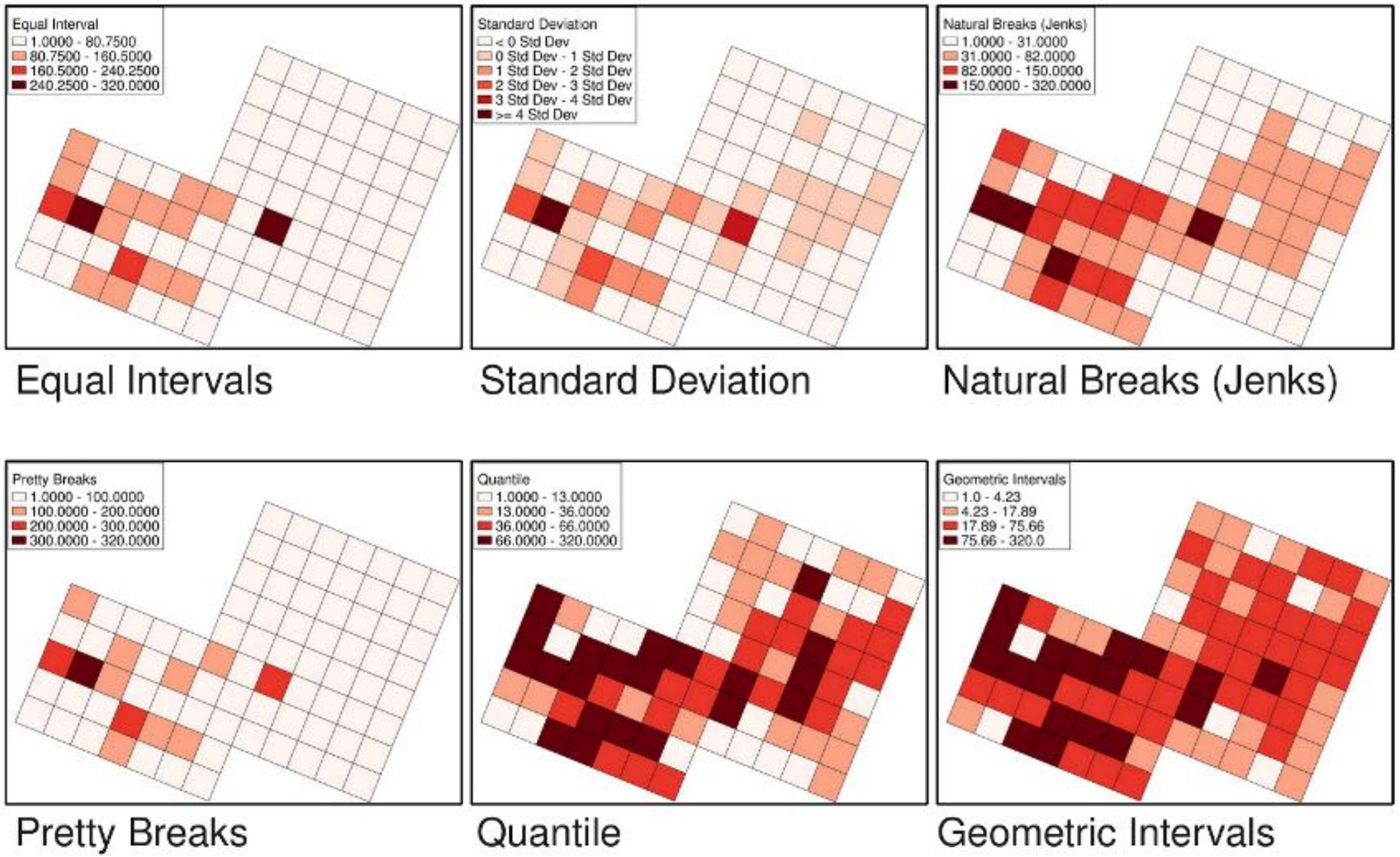
reject
accept
accept

How many categories should there be? If 2: Binarization
How should the values be mapped?

https://www.nature.com/articles/518288a

Data can be discretized very differently

Example: Weight of finds in an excavation grid

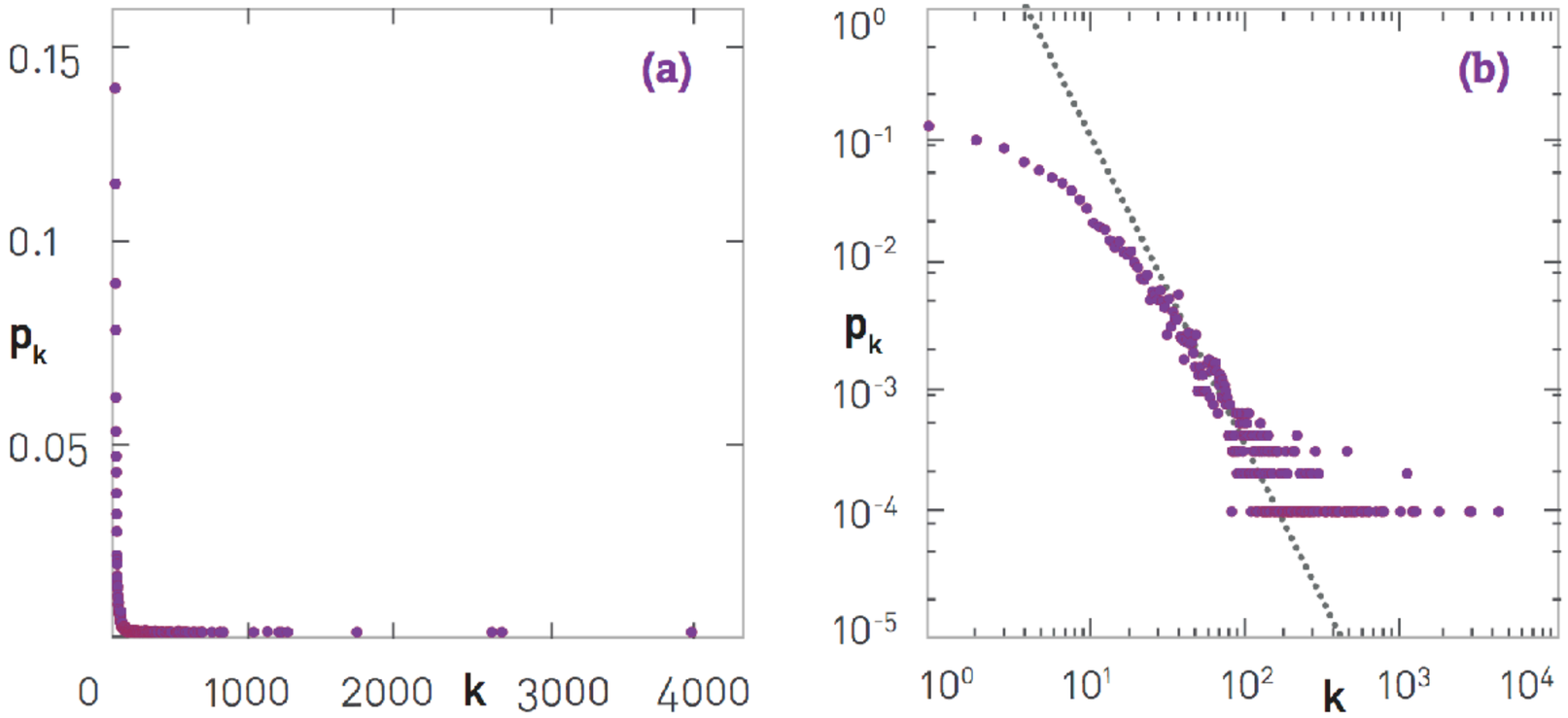


Same data, different split points

http://184.106.205.13/uselessarchaeology/CMSimple_UA/?Programming___QGIS_Geometric_Class

Variable transformation = Apply a function to all values

Common in skewed data: Logarithm

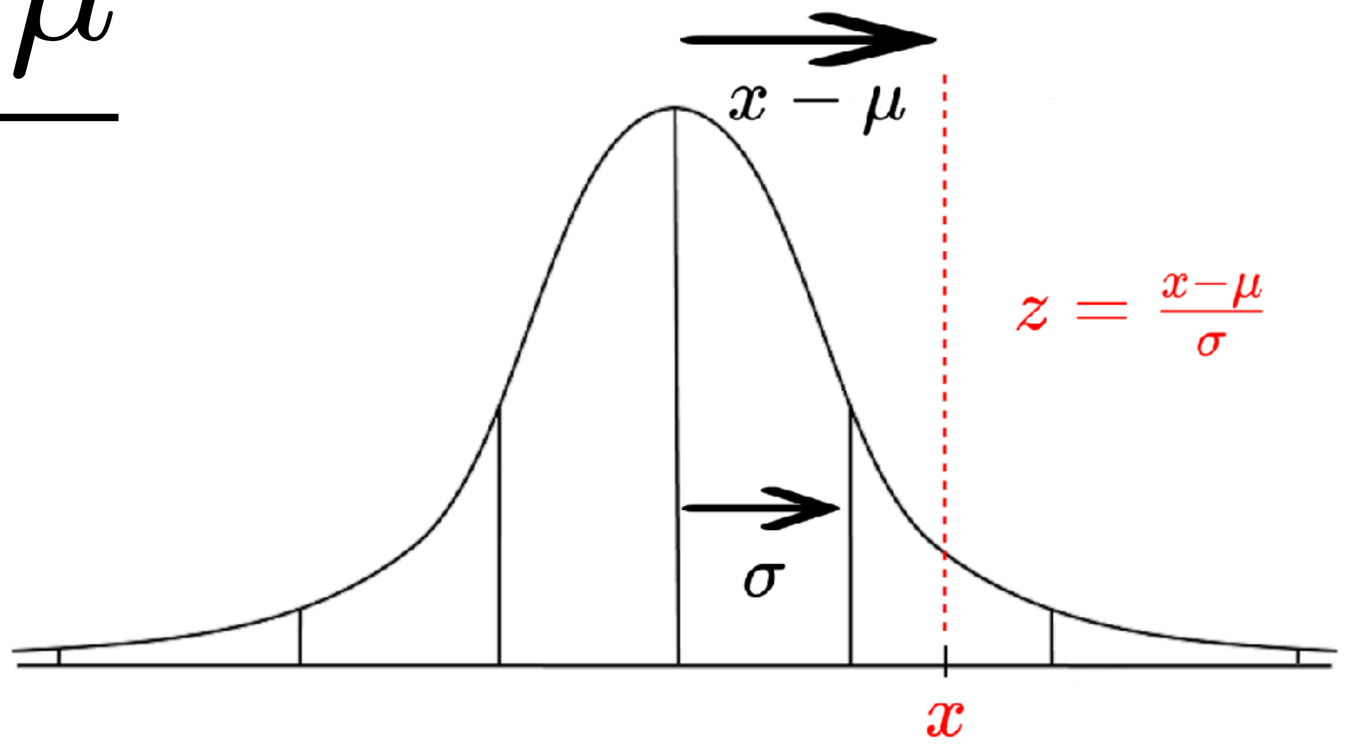


Variable transformation = Apply a function to all values

Common in normally distributed data: Standardization

Rescaled to have a mean of 0 and a standard deviation of 1.

$$z = \frac{x - \mu}{\sigma}$$



There are many questions to ask in data preparation

1) What is the problem I want to solve?

2) Is data for this available or do I need to collect it?

3) Is the quality and quantity of my data set good enough?

4) What parts of the data set are relevant?

5) How do I need to reshape my data to solve the problem?

6) How do I need to reshape my data to solve the problem efficiently?

.

.

.

There are many questions to ask in data preparation

1) What is the problem I want to solve?

2) Is data for this available or do I need to collect it?

3) Is the quality and quantity of my data set good enough?

4) What parts of the data set are relevant?

5) How do I need to reshape my data to solve the problem?

6) How do I need to reshape my data to solve the problem efficiently?

. Therefore, data cleaning IS analysis!

. You cannot separate the two.

Randy Au, datacleaningisanalysis.pdf

Data Science Pitfalls in Real Applications

Bias is not just in the data, but also in algorithms

Machine Learning

AI Detection Tools Falsely Accuse International Students of Cheating

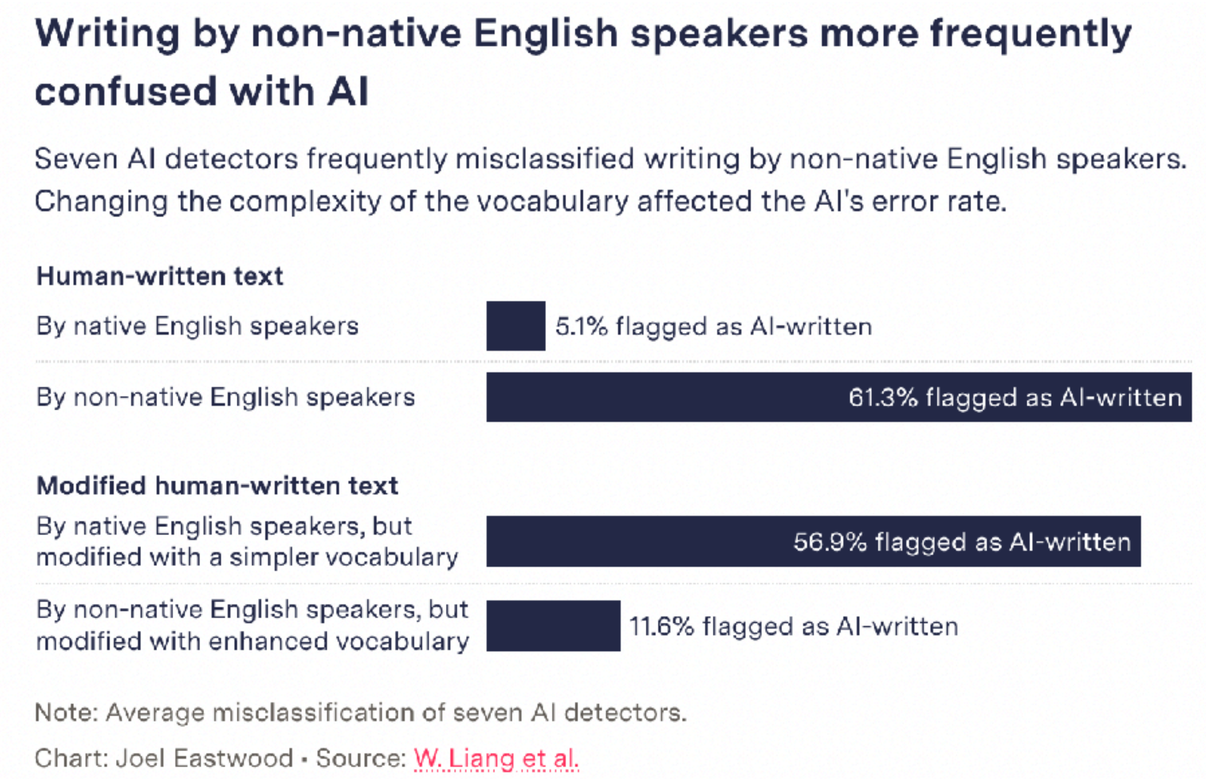
Stanford study found AI detectors are biased against non-native English speakers

By Tara García Mathewson

August 14, 2023 08:00 ET

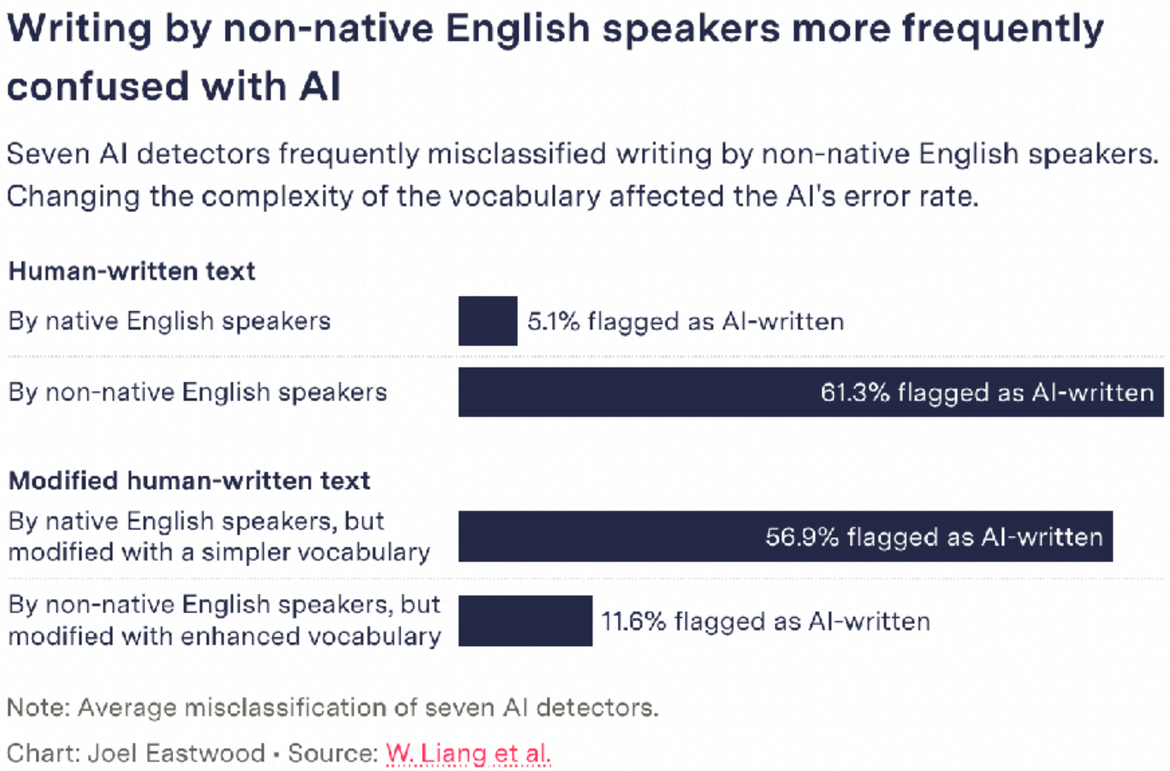
https://themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating

Bias is not just in the data, but also in algorithms



https://themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating

Bias is not just in the data, but also in algorithms



AI detectors tend to be programmed to flag writing as AI-generated when the word choice is predictable and the sentences are more simple. As it turns out, writing by non-native English speakers often fits this pattern, and therein lies the problem.

https://themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating

Bias is not just in the data, but also in algorithms

https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773

MACHINE BIAS

What Algorithmic Injustice Looks Like in Real Life

A computer program rated defendants' risk of committing a future crime. These are the results.

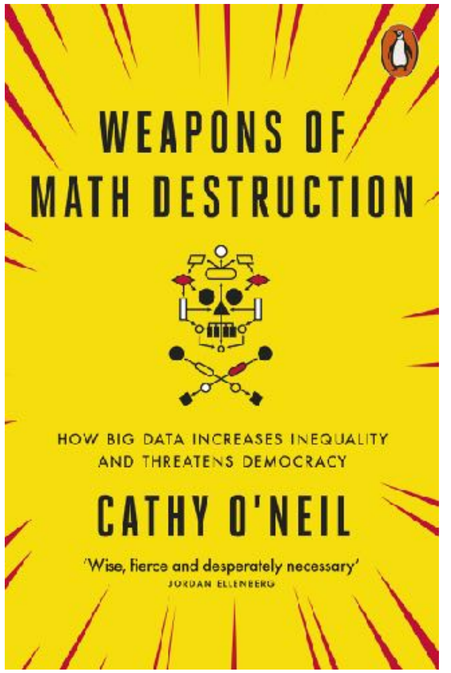
https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

MACHINE BIAS

What Algorithmic Injustice Looks Like in Real Life

A computer program rated defendants' risk of committing a future crime. These are the results.

Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It



https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/

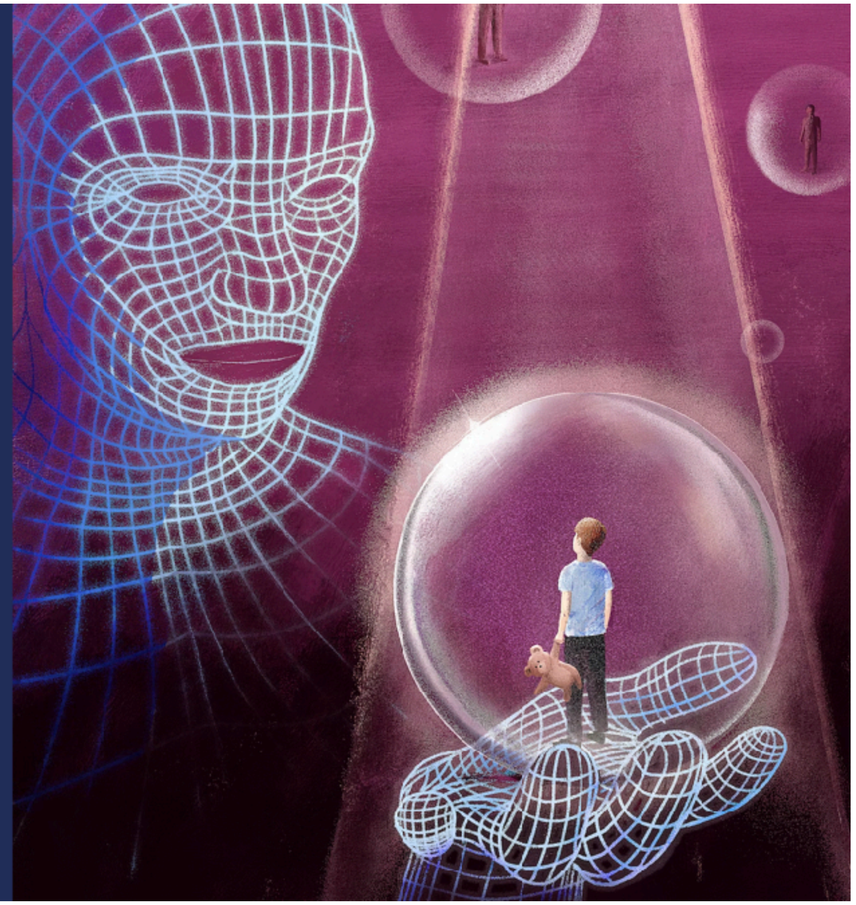
https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

Problematic algorithms are also applied in Denmark

Therese Moreau

Kan algoritmer se ind i et barns fremtid? I Hjørring og Silkeborg eksperimenterede man på udsatte børn

I et stort eksperiment testede to kommuner, om en kunstig intelligens kan assistere socialrådgivere i at træffe en af de sværeste beslutninger i velfærdssamfundet: at anbringe et barn uden for hjemmet.



https://www.zetland.dk/historie/s8YxAamr-m8aLzmJK-375f5

https://marycenter.ku.dk/netvaerk/cph-tech-policy-committee/cph-tech-policy-brief-6/CPH_TECH_POLICY_BRIEF_no._6.pdf

Problematic algorithms are also applied in Denmark

Overall, the model suggests that older children are at substantially higher risk of maltreatment, everything else being equal.

As there is no welfare crisis among teenagers this does not make sense.

https://www.zetland.dk/historie/s8YxAamr-m8aLzmJK-375f5

https://marycenter.ku.dk/netvaerk/cph-tech-policy-committee/cph-tech-policy-brief-6/CPH_TECH_POLICY_BRIEF_no._6.pdf

Problematic algorithms are also applied in Denmark

McNamara Fallacy

Overall, the model suggests that older children are at substantially higher risk of maltreatment, everything else being equal.

The data sample is biased as it only contains information for children which the social services have received at least one notification about.

the algorithm is not getting the full picture, which can lead it to make incorrect judgments

https://www.zetland.dk/historie/s8YxAamr-m8aLzmJK-375f5

https://marycenter.ku.dk/netvaerk/cph-tech-policy-committee/cph-tech-policy-brief-6/CPH_TECH_POLICY_BRIEF_no._6.pdf

The solution to many bias problems in tech:

Better testing, on a representative sample of the population

The solution to many bias problems in tech:

Better testing, on a representative sample of the population

Legal regulations, audits +

The solution to many bias problems in tech:

Better testing, on a representative sample of the population

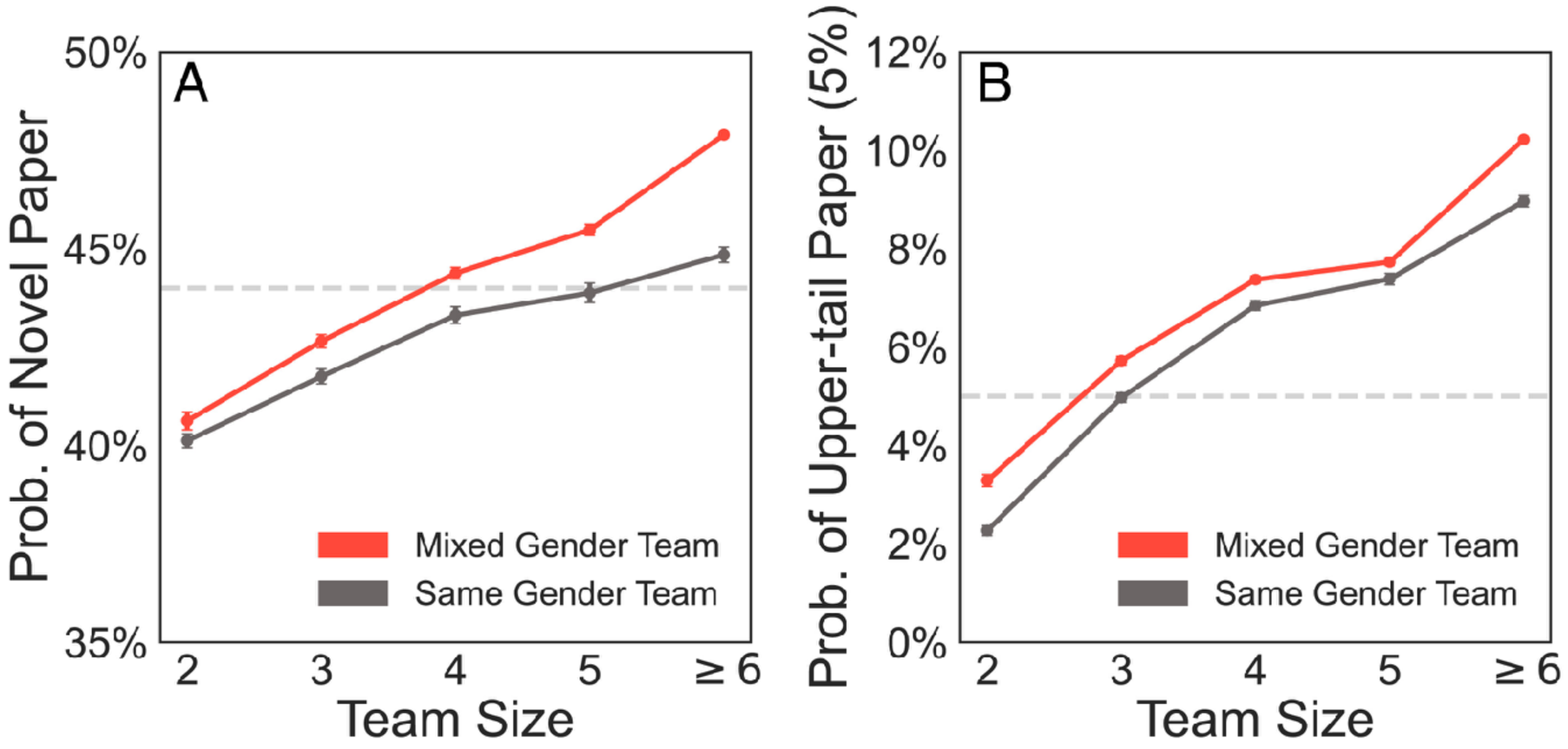
Legal regulations, audits +

If the software developers and their bosses are representative they will think about more cases to test

"Diversity" is not just a buzz word

RESEARCH ARTICLE | SOCIAL SCIENCES |

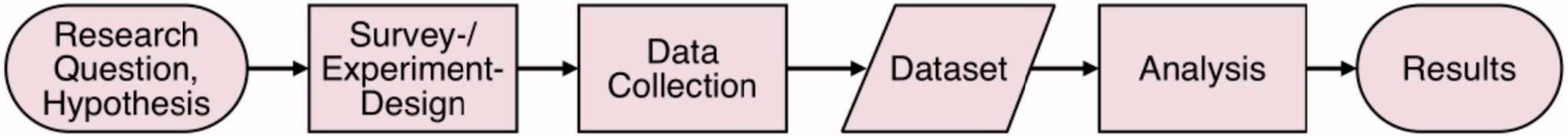
Gender-diverse teams produce more novel and higher-impact scientific ideas



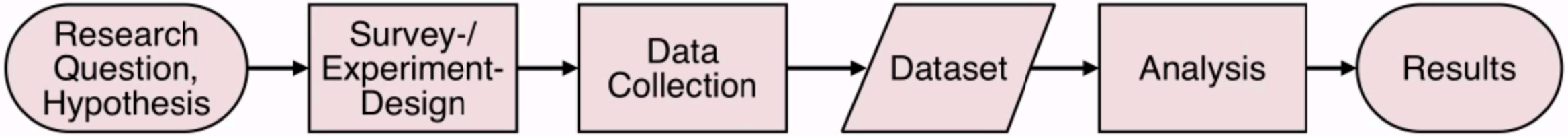
https://www.pnas.org/doi/full/10.1073/pnas.2200841119

Data Science Pitfalls in Research

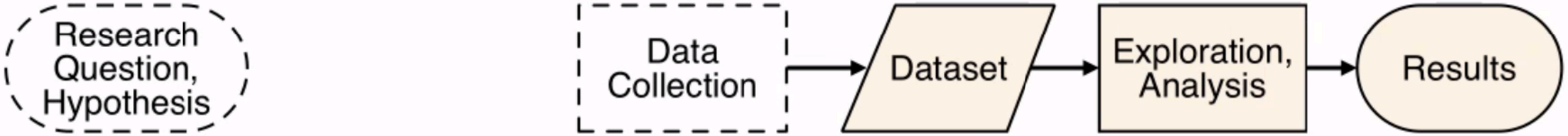
"Traditional" research workflow



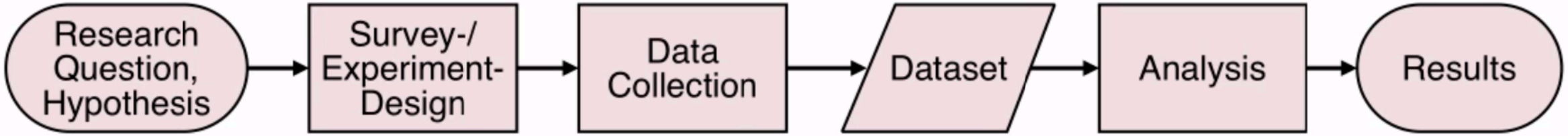
"Traditional" research workflow



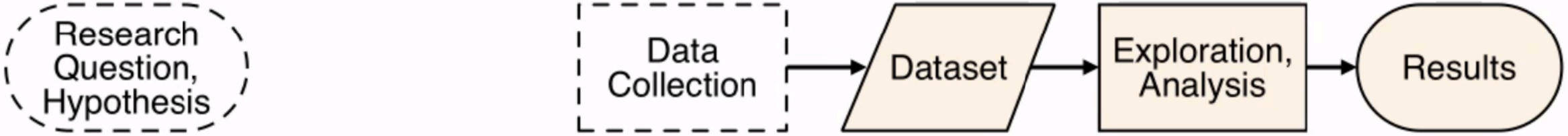
"Big Data" research workflow



"Traditional" research workflow



"Big Data" research workflow



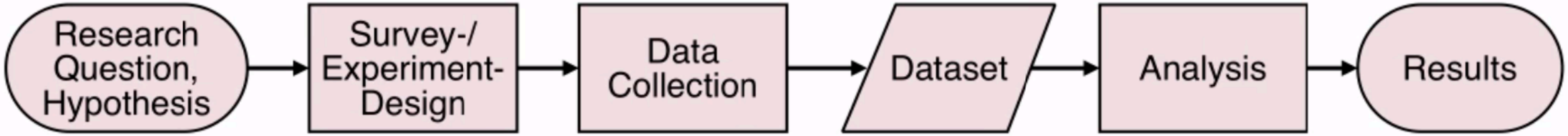
How was the sample selected?

What pre-processing was performed?

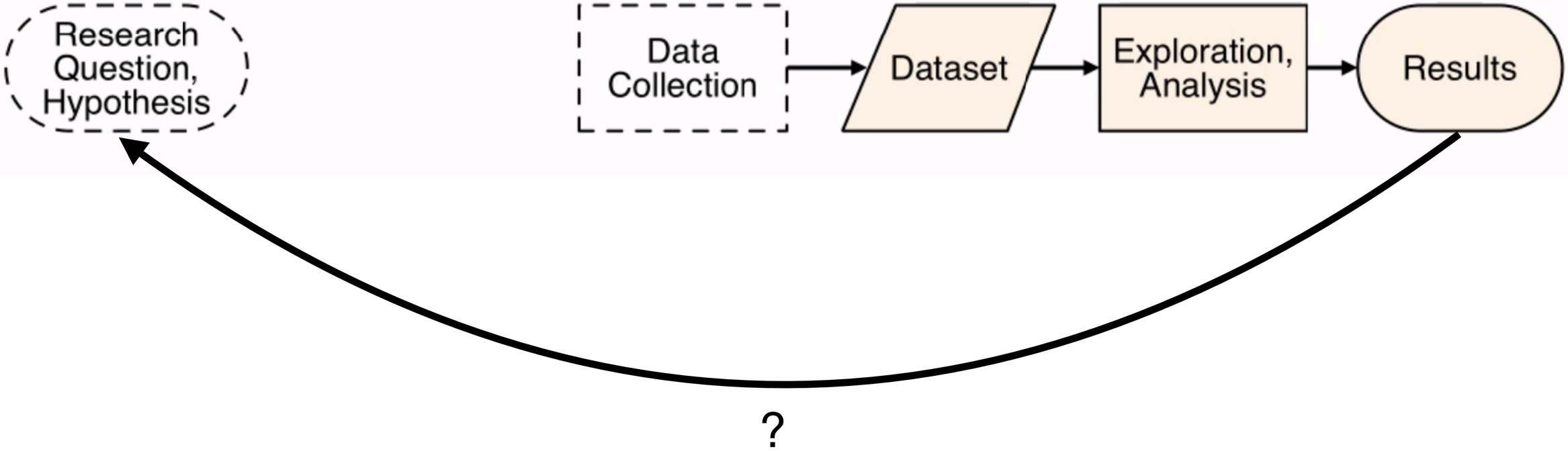
Was the measurement process uniform?

Who is represented in the data?

"Traditional" research workflow



"Big Data" research workflow



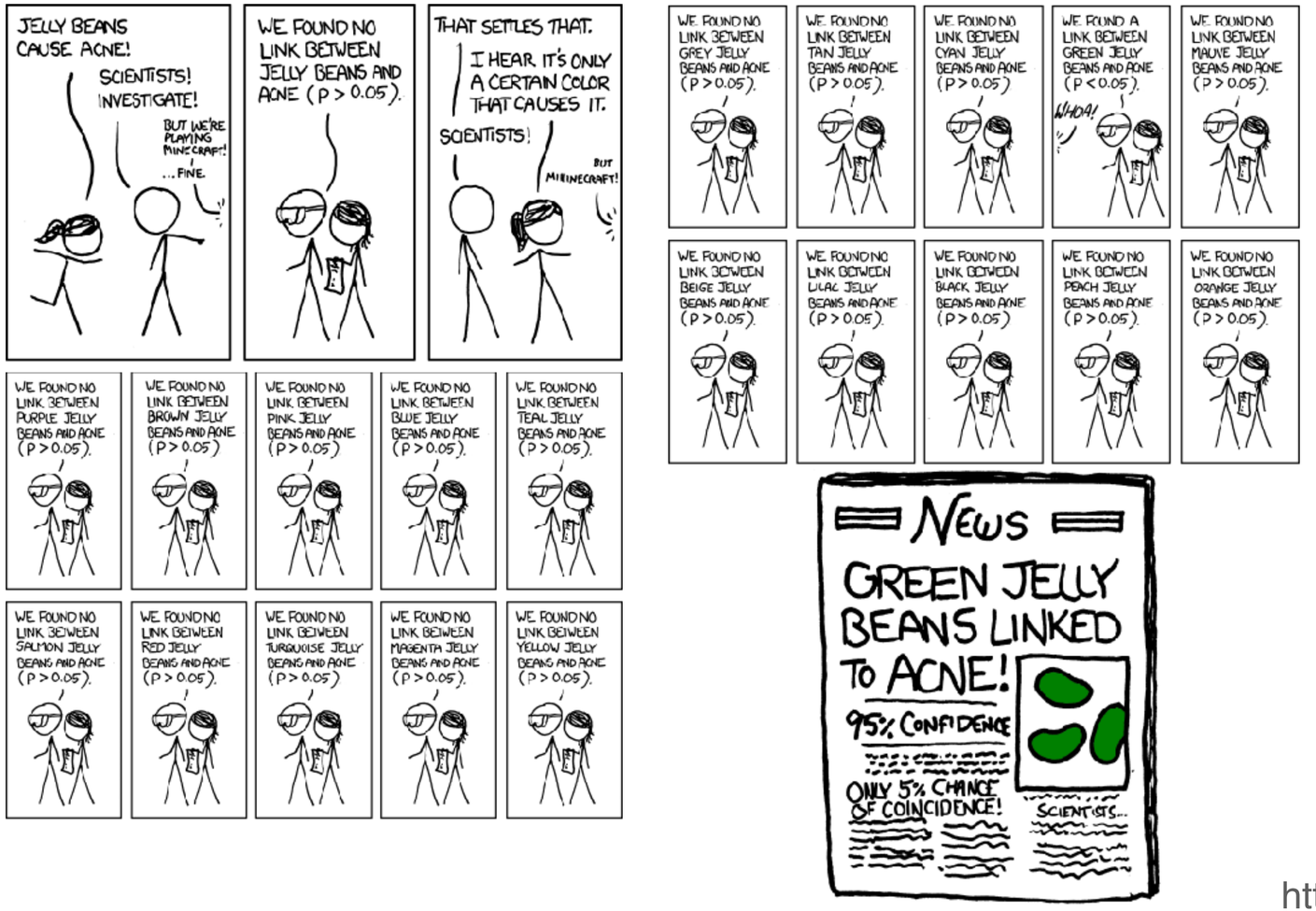
Beware of data dredging

Data dredging is a misuse of data analysis: To run many statistical tests and to only report those with a significant result.

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

Beware of data dredging

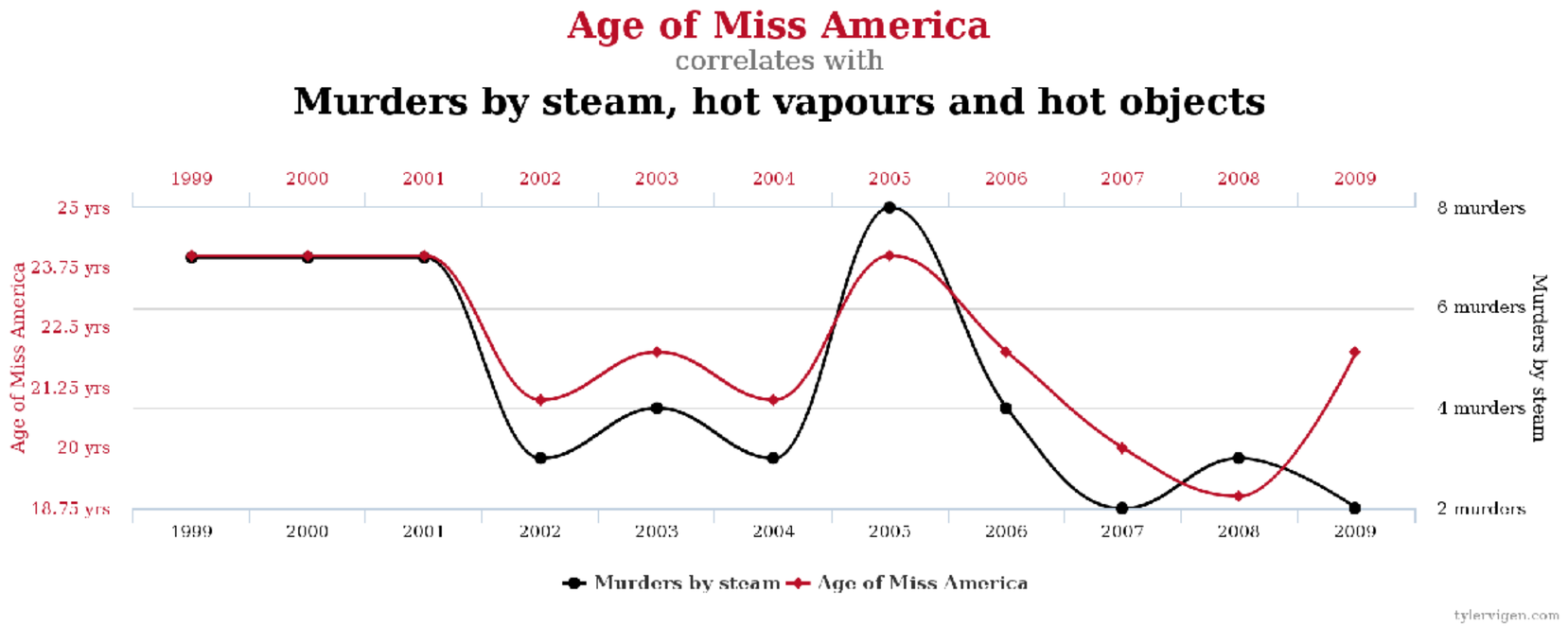
Data dredging is a misuse of data analysis: To run many statistical tests and to only report those with a significant result.



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

Beware of data dredging

Data dredging is a misuse of data analysis: To run many statistical tests and to only report those with a significant result.

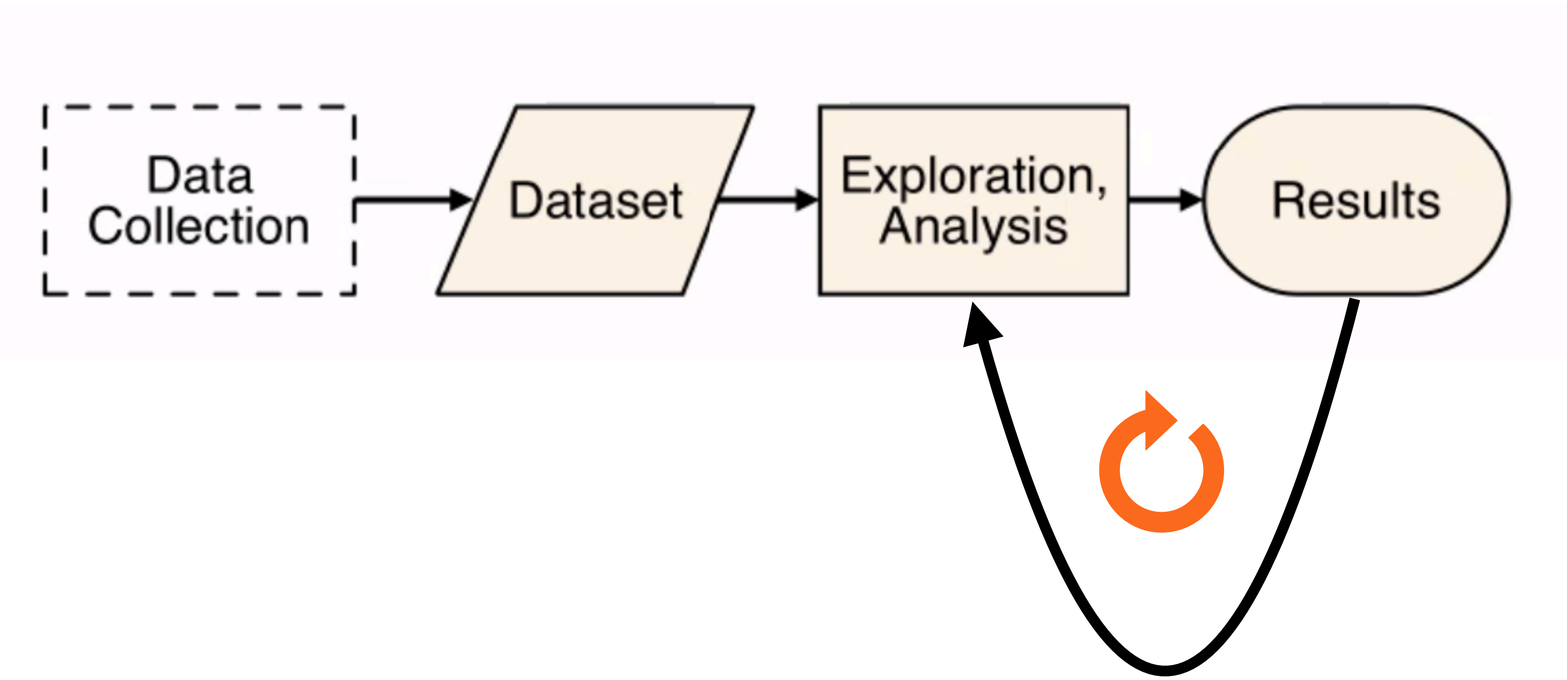


Eventually you will just find a spurious correlation

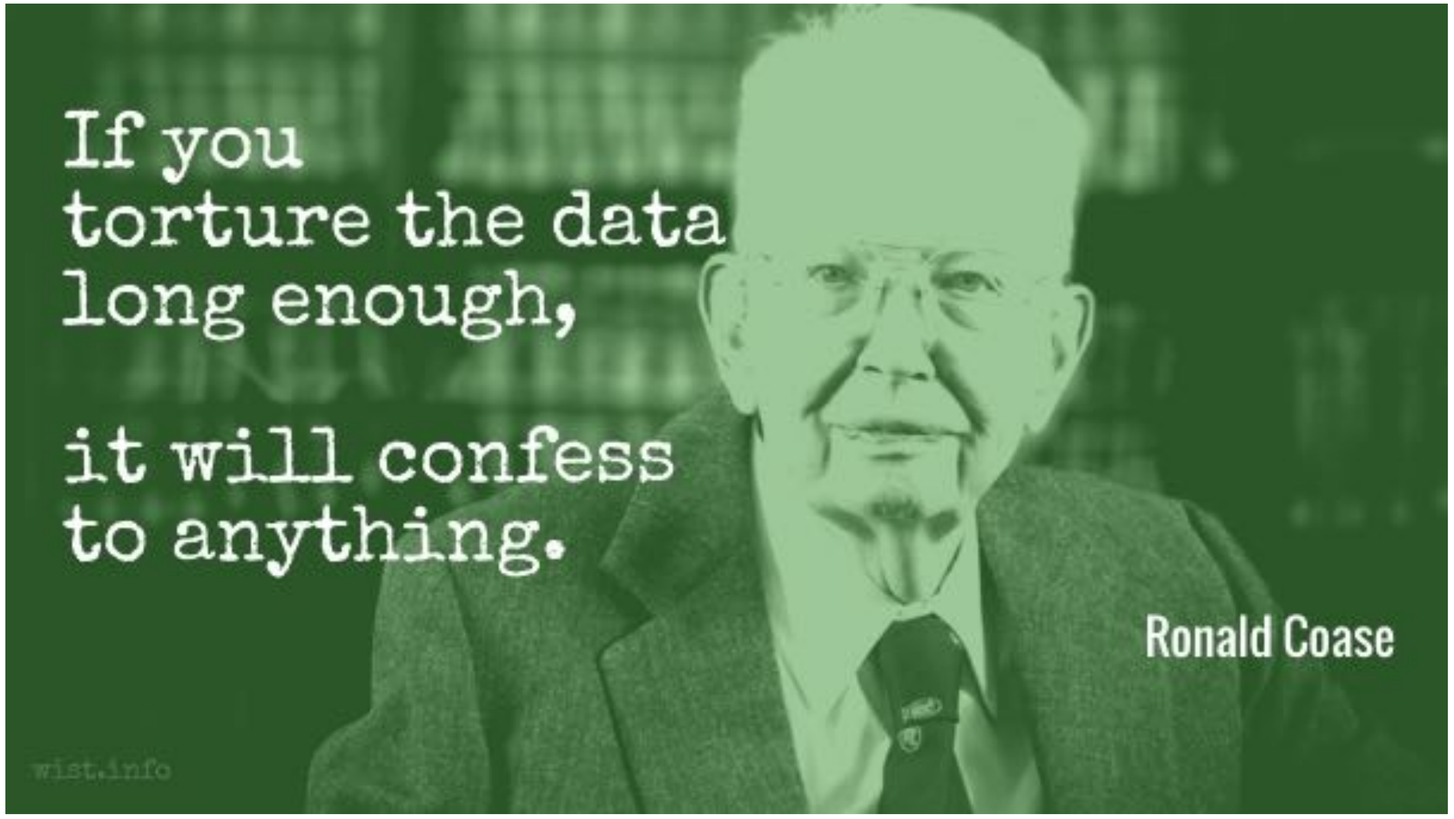
https://en.wikipedia.org/wiki/Data_dredging

Beware of data dredging

Data dredging is a misuse of data analysis: To run many statistical tests and to only report those with a significant result.



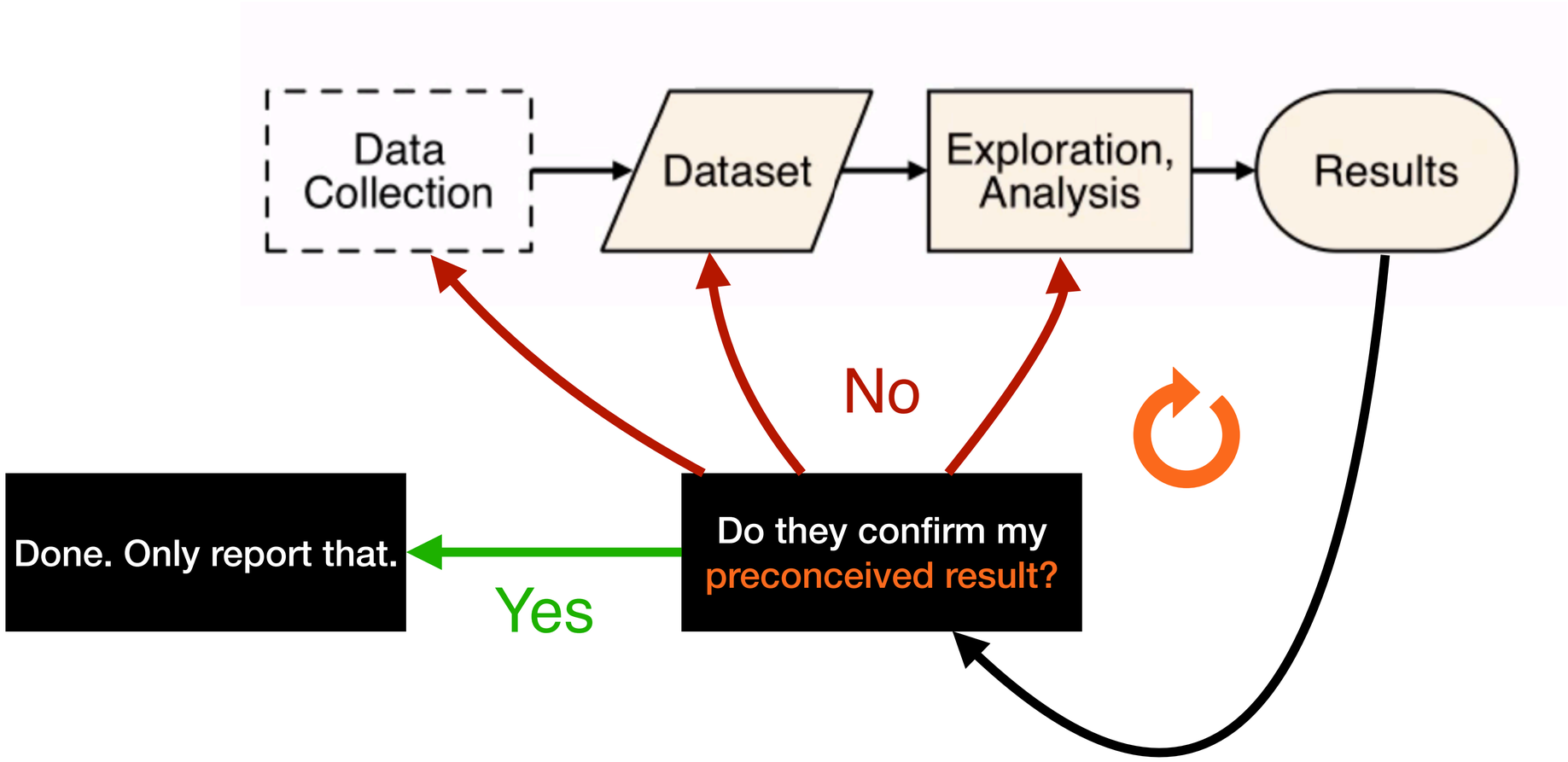
https://en.wikipedia.org/wiki/Data_dredging


If you
torture the data
long enough,
it will confess
to anything.
Ronald Coase
wist.info

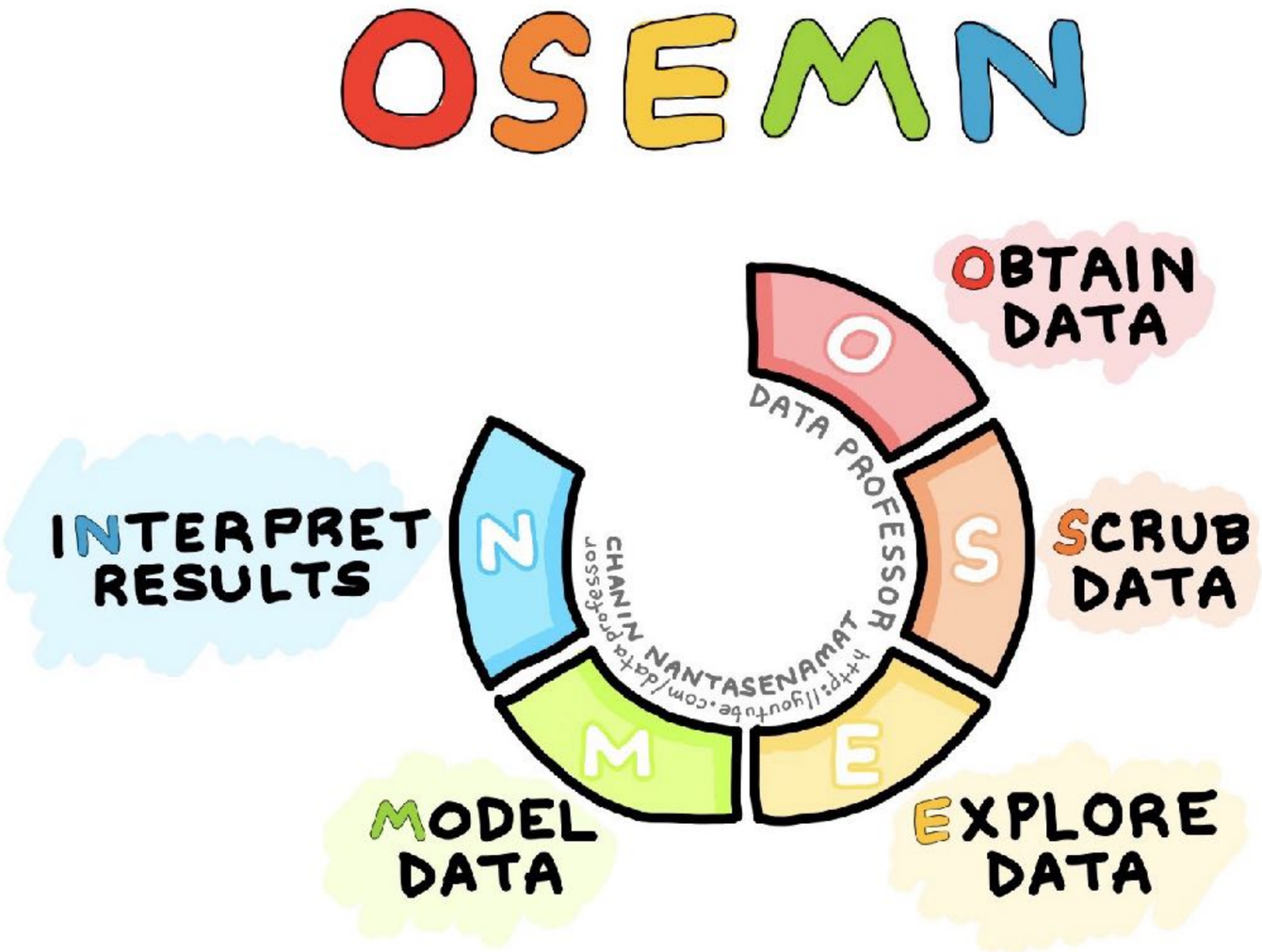

DATA
DATA SCIENTIST
imgflip.com

Beware of data torture for a preconceived result

"Data torture" is a misuse of data analysis: To run many statistical tests, or to use different data, to find support for a preconceived result.

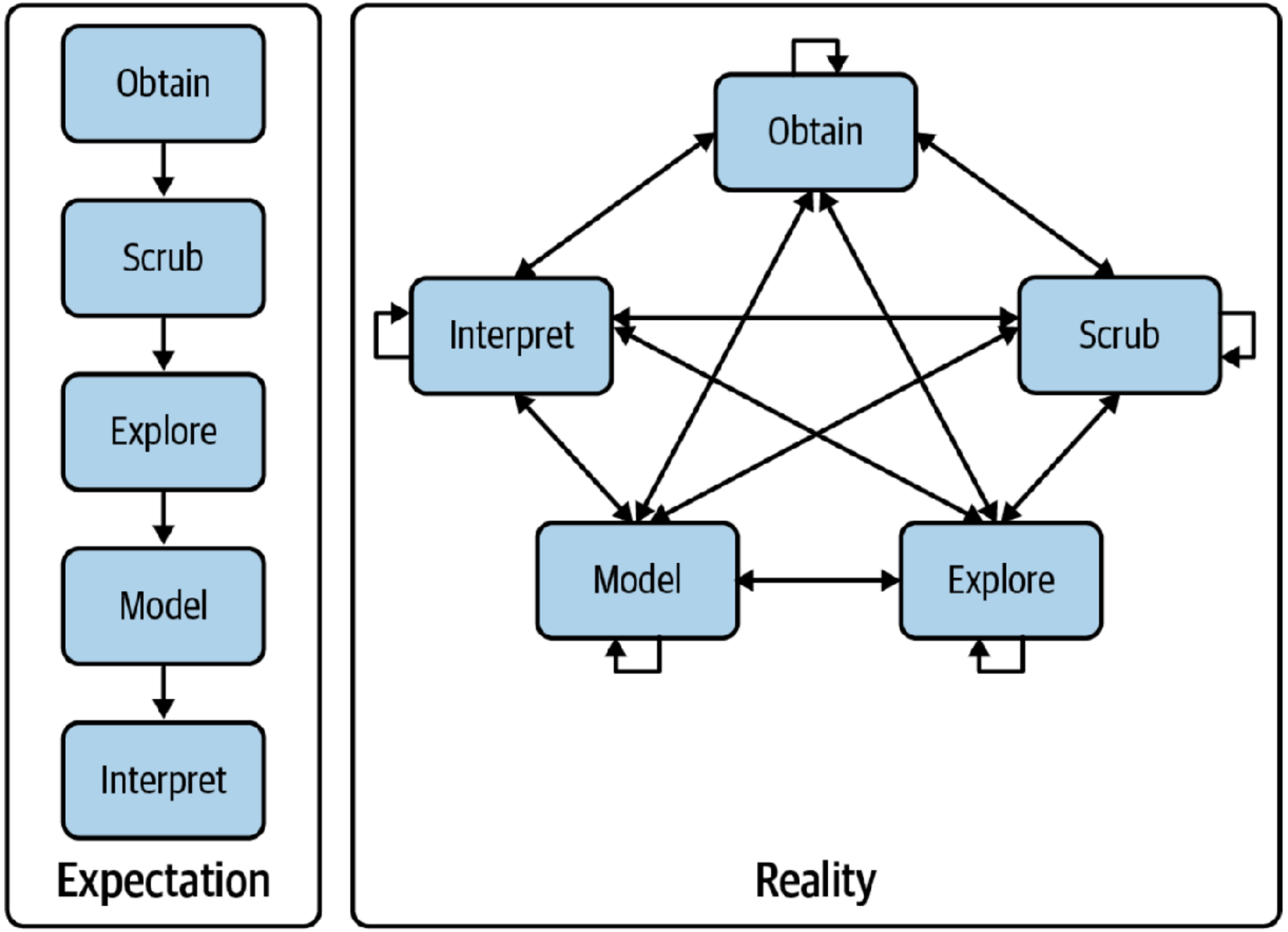


Obtain - **S**crub - **E**xplore - **M**odel - I**N**terpret (2010)



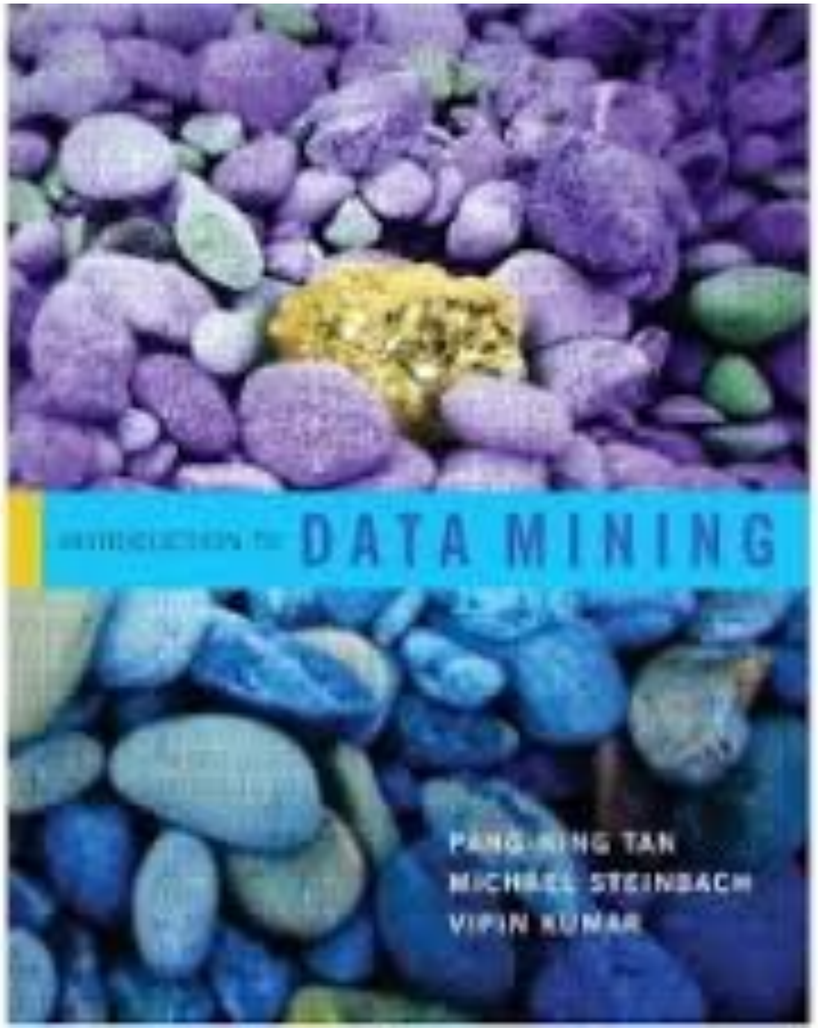
https://www.datascience-pm.com/osemn/

Obtain - **S**crub - **E**xplore - **M**odel - I**N**terpret (2010)



Data science at the command line, Fig 1.1

Sources and further materials for today's class



Most important insights from today

Data science is a non-linear process.
You iterate: Make mistakes, learn, go back, reformulate, over and over..

Asking about data quality is most important!
Garbage in - Garbage out

Data cleaning and analysis cannot be separated.
Document ALL the steps (in Jupyter notebooks).