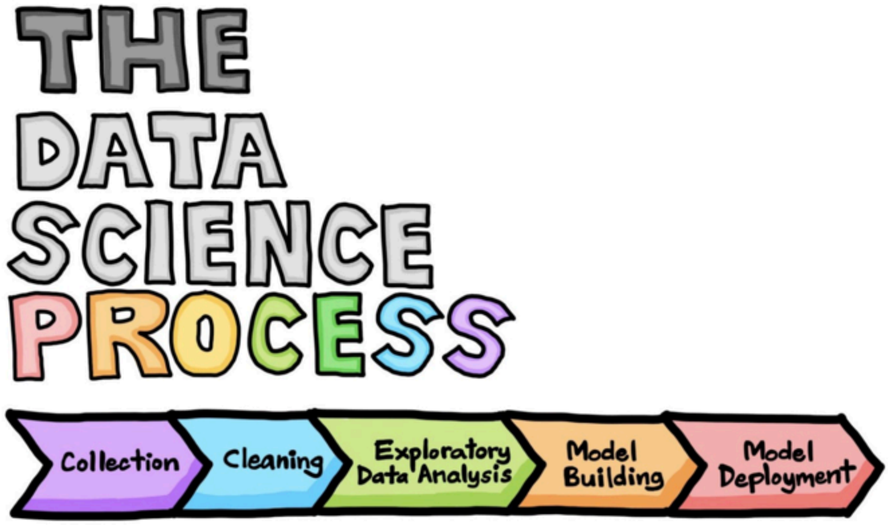
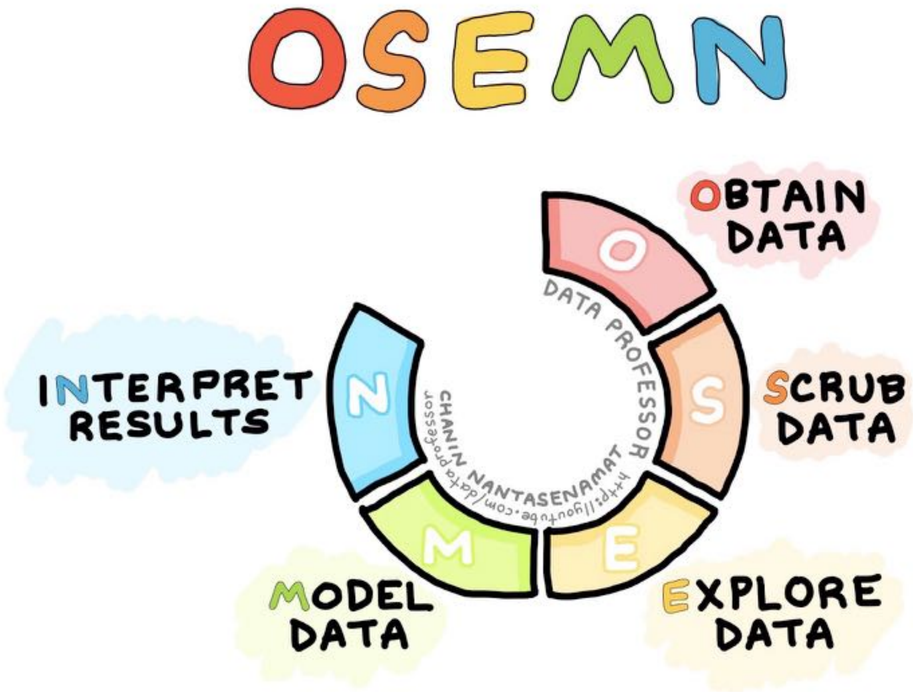
Lecture 26: The Data Science Process 2

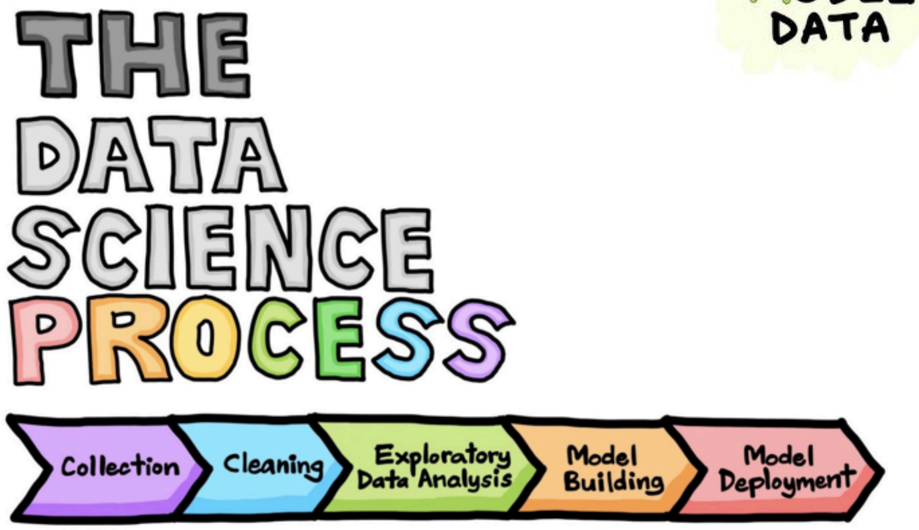
Instructor: Michael Szell
Nov 25, 2020



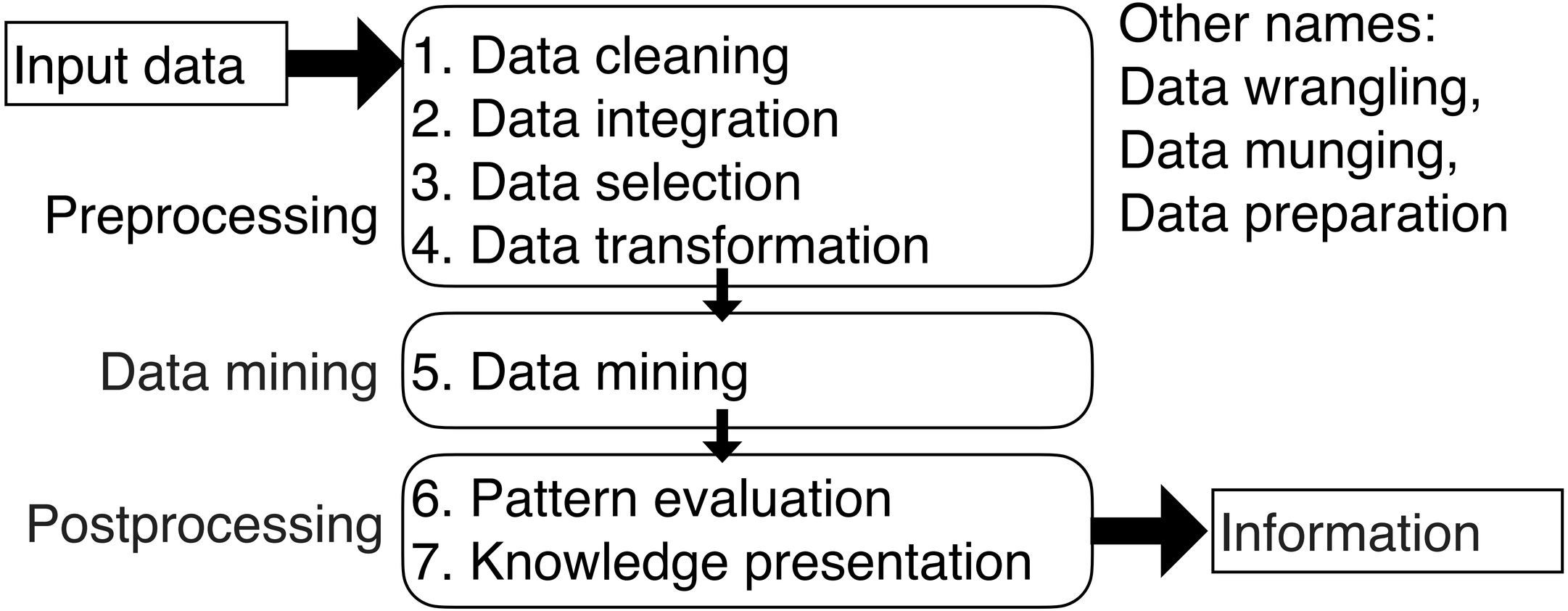
Today we will reflect on the data science process

Data quality and preprocessing



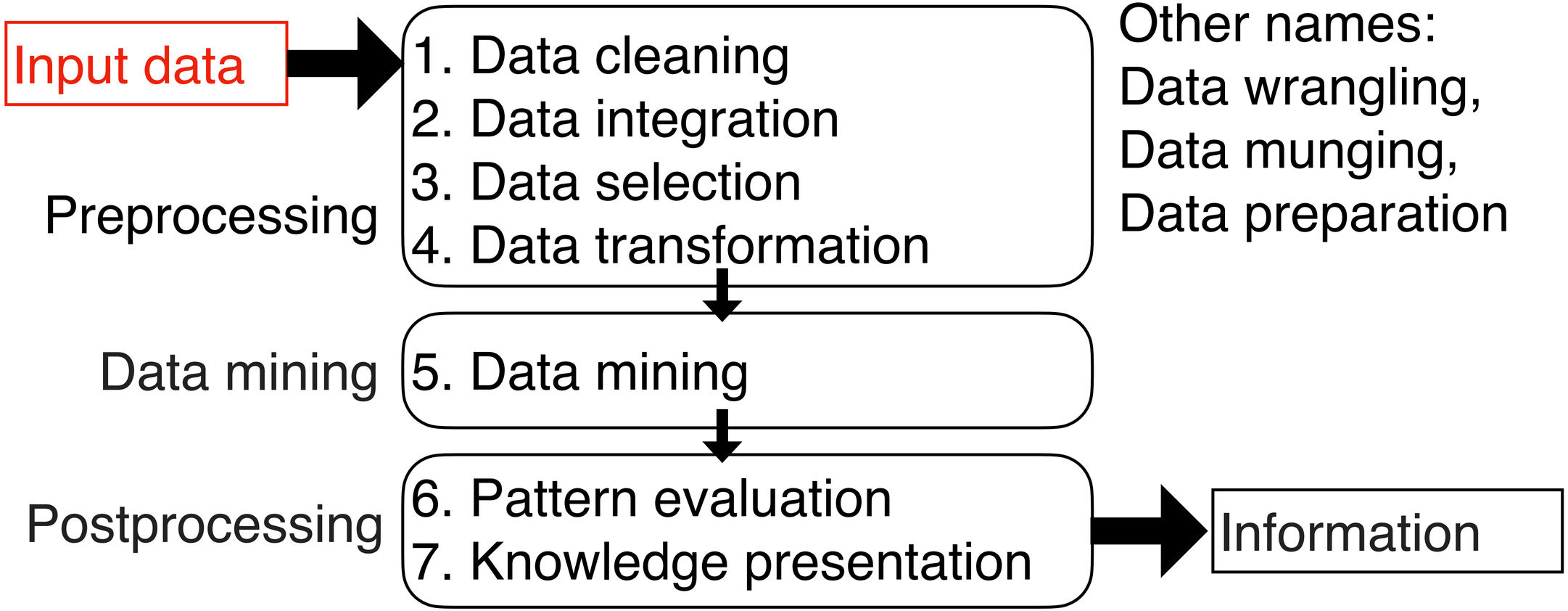


What is the most important step in data mining?



Other names:
Data wrangling,
Data munging,
Data preparation

The most important step in data mining is the first



Other names:
Data wrangling,
Data munging,
Data preparation

The most important step in data mining is the first

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

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

Position zero, Lina Faller, Marcel Mieth, Thomas Stüssi, and Susanne Weck, Degree Confluence Project, 16-Jun-2007
http://confluence.org/confluence.php?visitid=14716

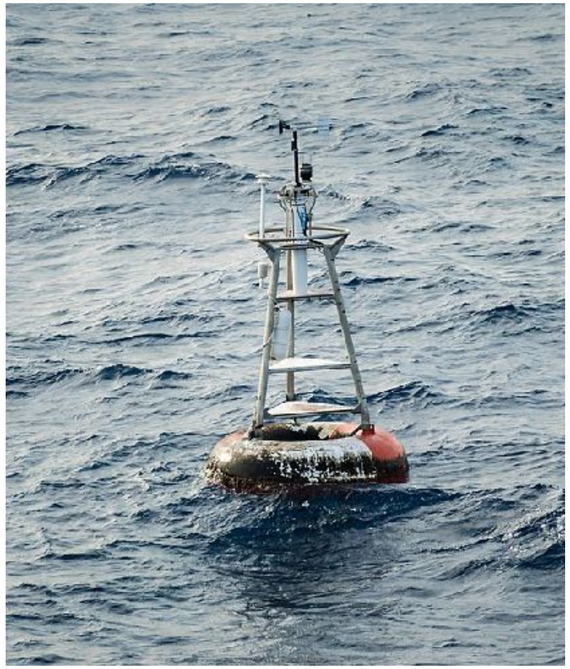
What is:

The most Geo-tagged Place on Earth

A homage to the Degree Confluence Project and the metaphysics of the GPS glitch.

?

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	PHILADELPOHIA
PHIALDELPHIA	PHILADELPPHIA
PHIDELPHIA	PHILADEPHA
PHIELADELPHIA	PHILADEPHIA
PHIILADELPHIA	PHILADEPHILA
PHILA	PHILADEPLHIA
PHILA.	PHILADERLPHIA
PHILAD	PHILADLELPHIA
PHILADALPHIA	PHILADLEPHIA
PHILADEDLPHIA	PHILADLPHIA
PHILADELAPHIA	PHILADPHIA
PHILADELHIA	PHILADRLPHIA
PHILADELHPIA	PHILAELPHIA
PHILADELLPHIA	PHILDADELPHIA
PHILADELOHIA	PHILDADLPHIA
PHILADELPH	PHILDAELPHIA
PHILADELPHA	PHILDELPHIA
PHILADELPHAI	PHILDEPPHIA
PHILADELPHI	PHILIADELPHIA
PHILADELPHIA	PHILIDELPHIA
PHILADELPHIA PA	PHILLA
PHILADELPHIA,	PHILLADELPHIA
PHILADELPHIA, PA	PHILLY
PHILADELPHIA`	PHILOADELPHIA
PHILADELPHIAP	PHLADELPHIA
PHILADELPHIAPHIA	PHOLADELPHIA
PHILADELPHILA	PHPILADELPHIA
PHILADELPHIOA	PIHLADELPHIA
PHILADELPIA	

There are many issues affecting data quality

Data collection issues:

Recording errors

Duplications

Missing values

Inconsistencies

Dirty Data

FirstName	Surname	CompanyName	Address1	Town
peter	jones	jones café	80 riverways	manchester
lisa sefton			76 the avenue	leicester
a baker		bakery baker ltd	7 main road	reading berkshire
Richard	Evans1	Richard's Treats	9 charles Street	Bracknell
Alex		The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights		Gillingham
Janine		The Janine Way	10 Fleet Place	Bracknelll
Katherine	Bolton	Bolton Foods	bond Street	
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
emma	w	The Write Way	280 Bath rd	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh
Dave	Smith	Dave's Gift	po box	Leigh Lancs

Un-Standardised

Missing or misspelled

Duplications

Clean Data

FirstName	Surname	CompanyName	Address1	Town
Peter	Jones	Jones Café	80 Riverways	Manchester
Lisa	Sefton		76 The Avenue	Leicester
A	Baker	Bakery Baker Ltd	7 Main Road	Reading
Richard	Evans	Richard's Treats	9 charles Street	Bracknell
Alex	Froy	The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights	25 Camel Lane	Gillingham
Janine	Hutton	The Janine Way	10 Fleet Place	Bracknelll
Katherine	Bolton	Bolton Foods	bond Street	London
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh

Correctly Standardised

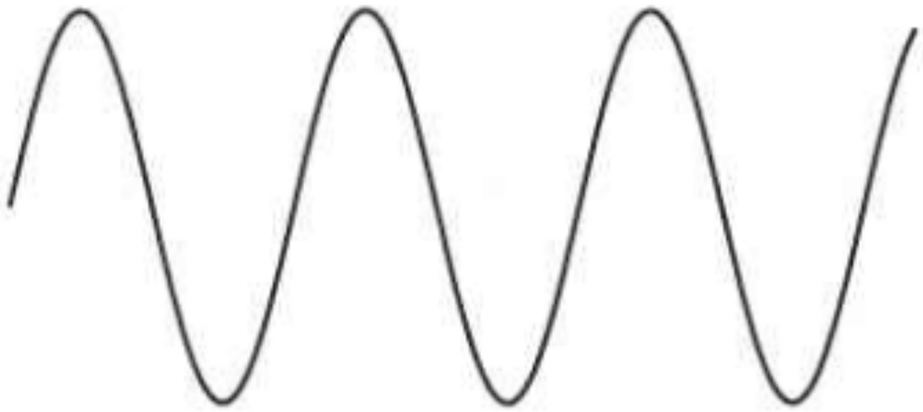
Populated and Corrected

Duplications Removed

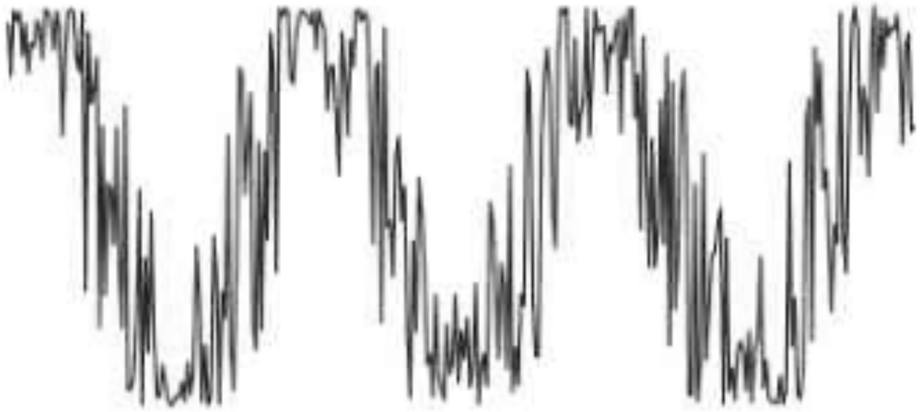
Jupyter

There are many issues affecting data quality

Measurement errors: Noise



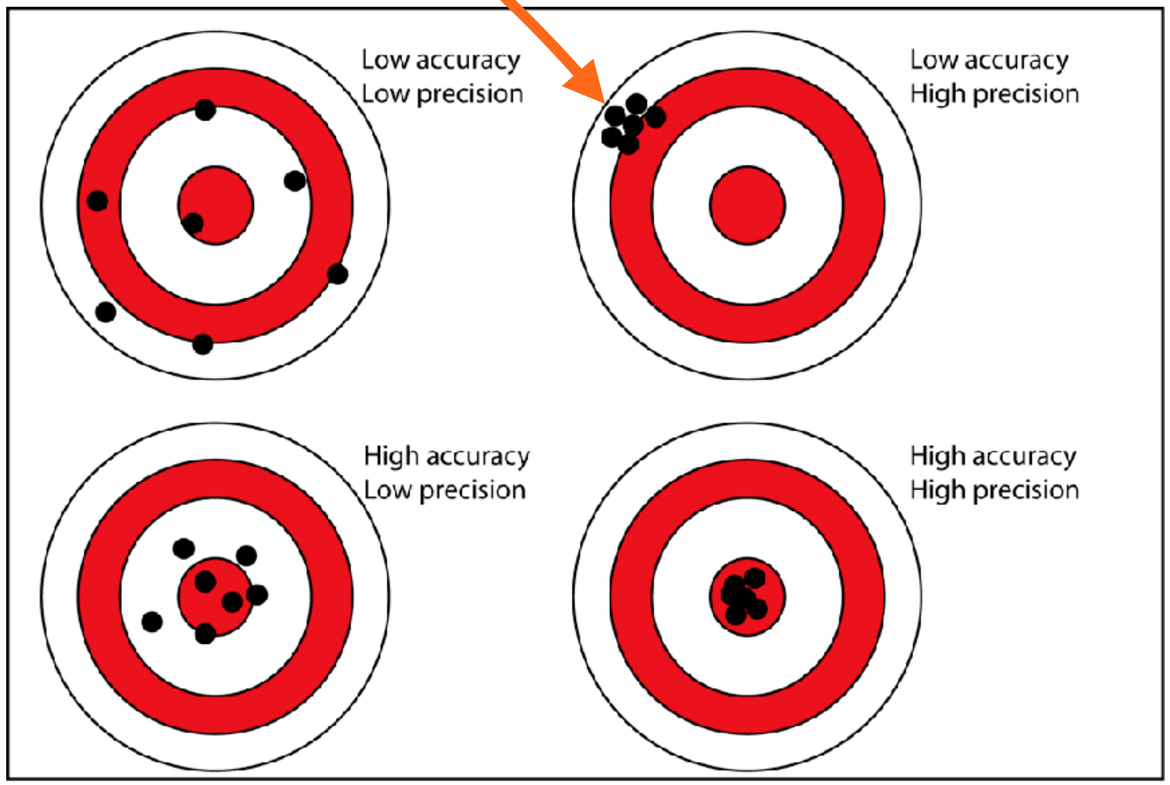
(a) Time series.



(b) Time series with noise.

There are many issues affecting data quality

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



Accuracy: The closeness of measurements to the true value

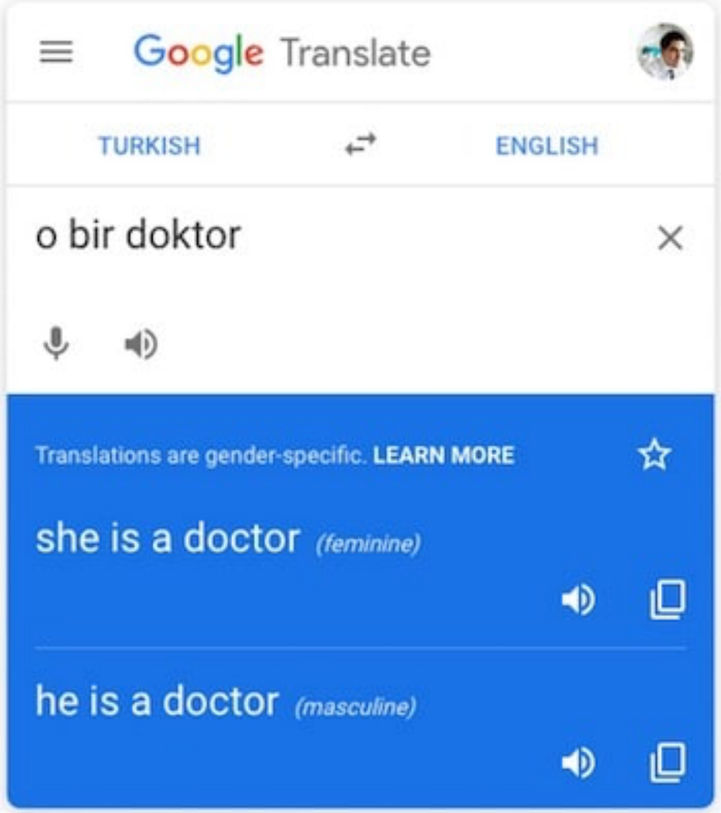
Precision: The closeness of repeated measurements to one another

There are many issues affecting data quality

Biased training data biases results

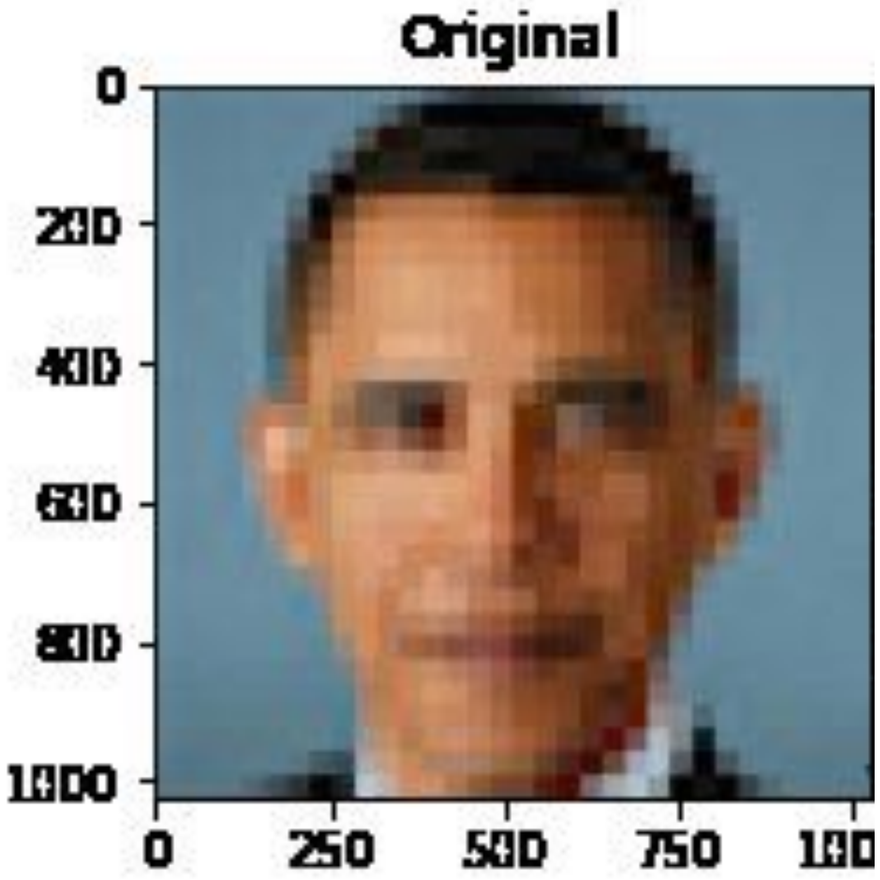
Before

After



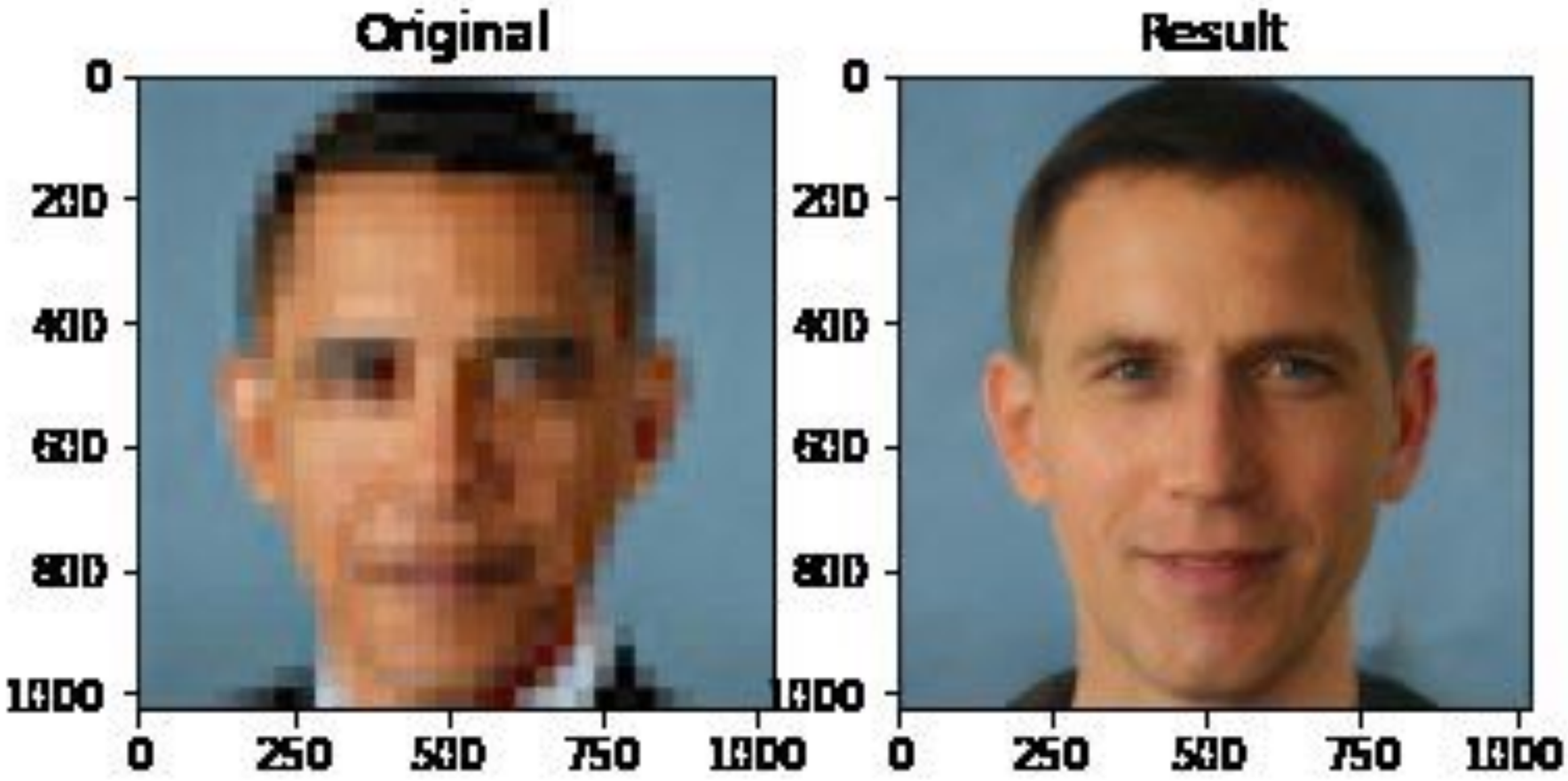
There are many issues affecting data quality

Biased training data biases results



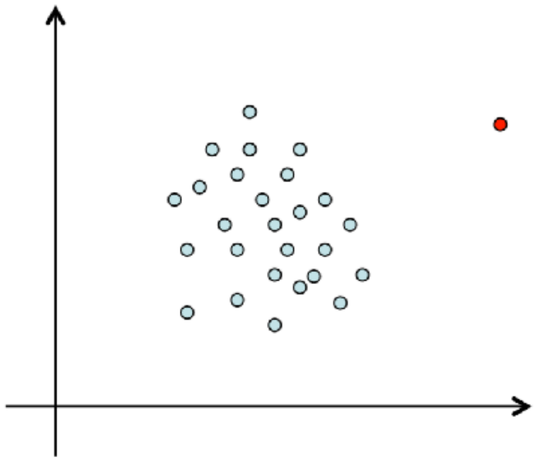
There are many issues affecting data quality

Biased training data biases results

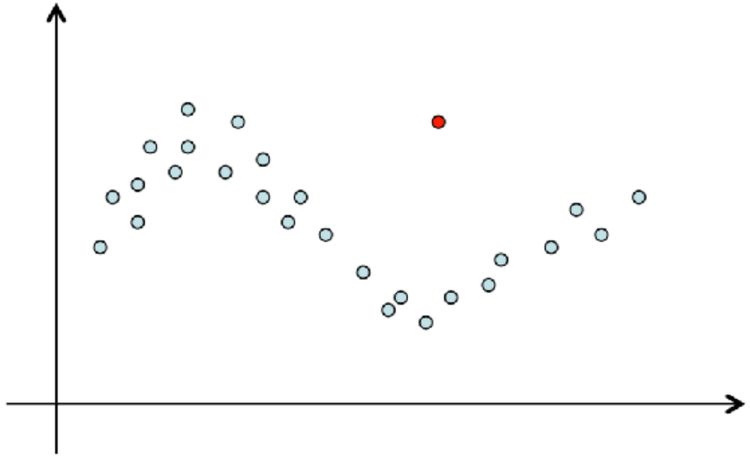


https://twitter.com/Chicken3gg/status/1274314622447820801?s=20

There are many issues affecting data quality



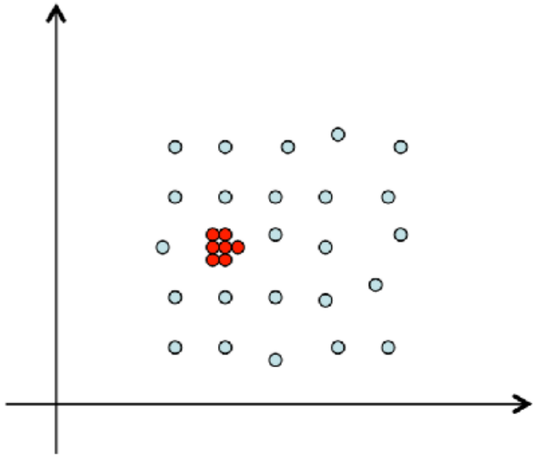
Global outliers



Contextual outliers

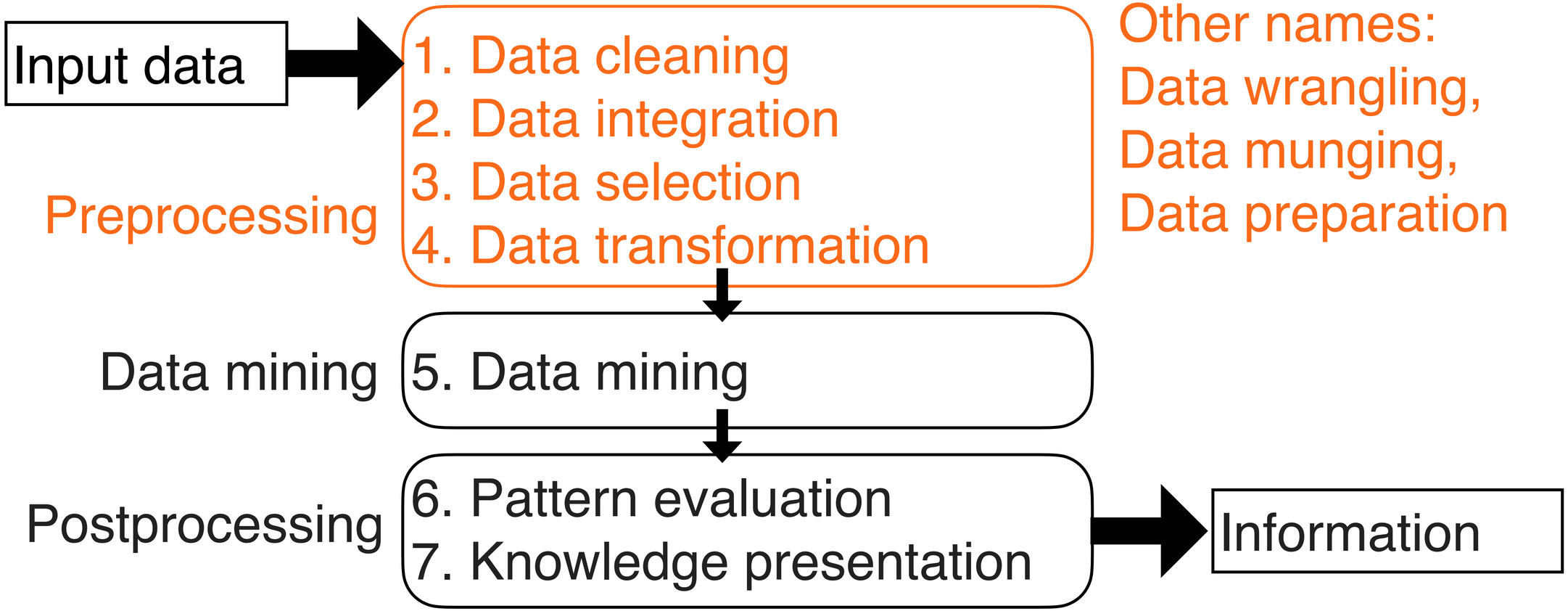
Outliers (anomalous objects or values):

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



Collective outliers

The second most important step in data analysis is the second



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

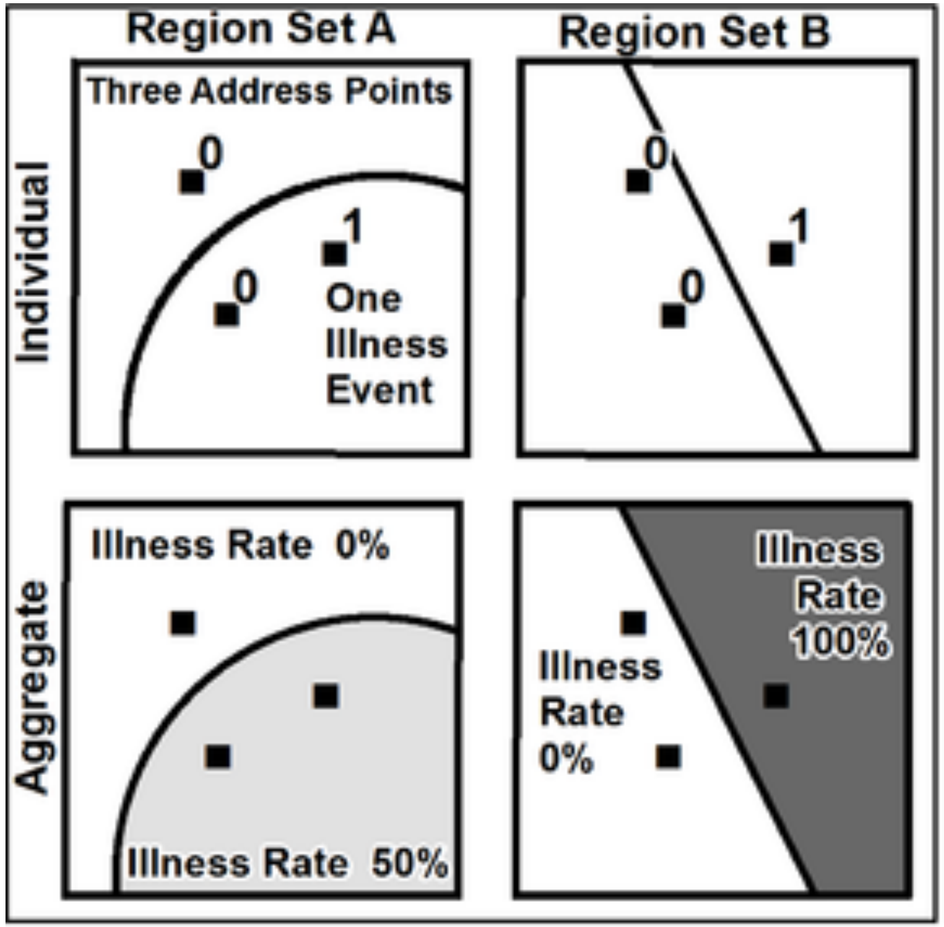
Second → Minute → Hour → Day → Week → Month → Year

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

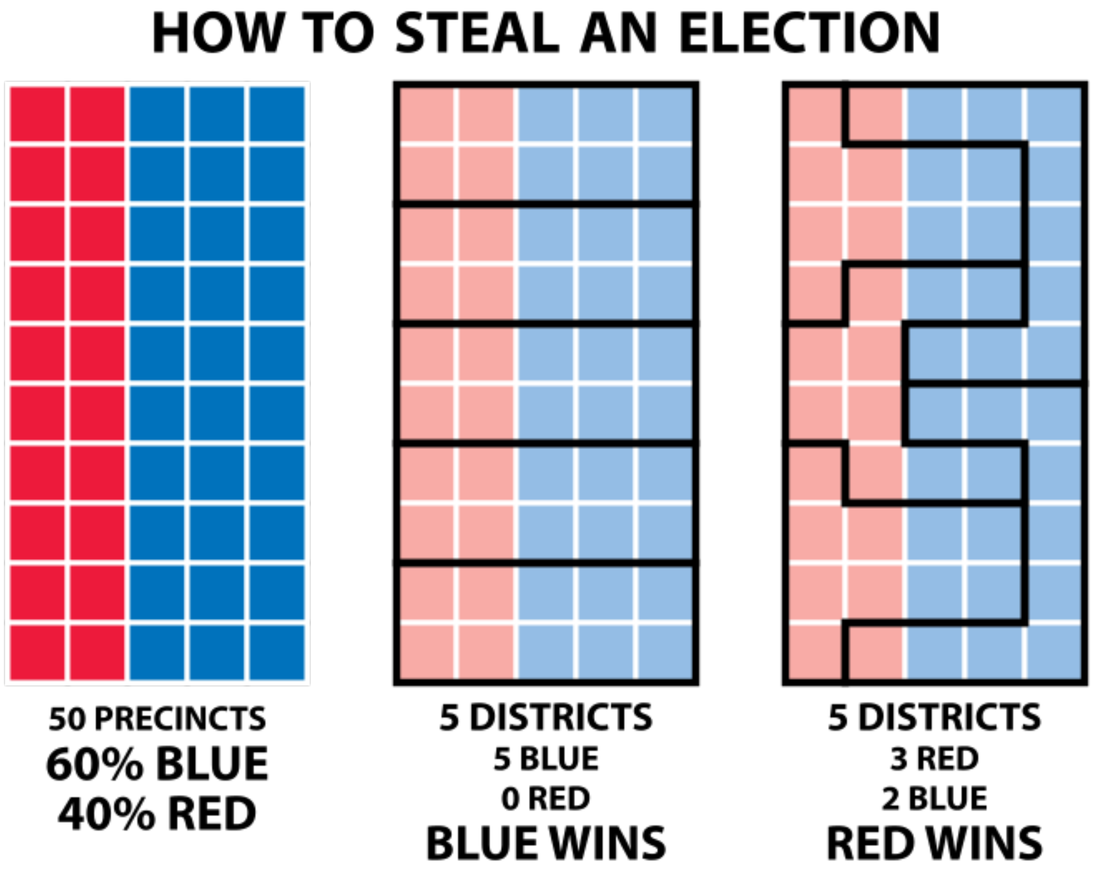
Disadvantages: Loss of details, introducing biases

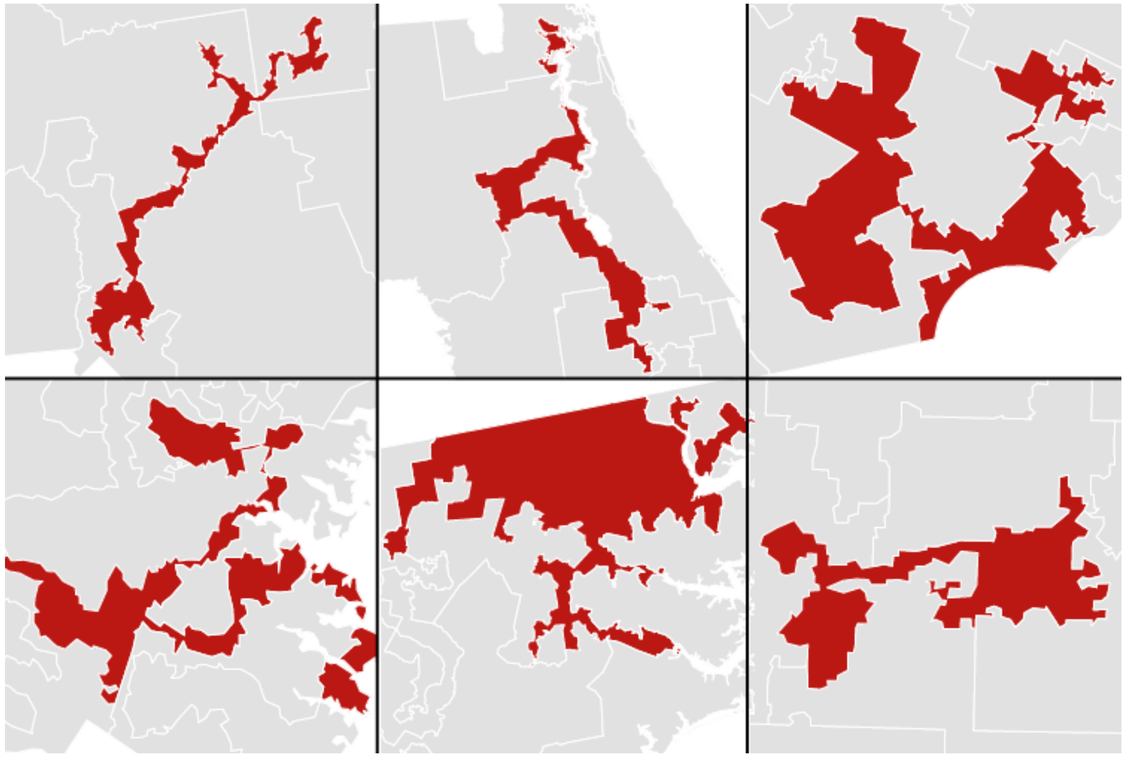
A common bias in spatial aggregation is the MAUP

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~~	~~...~~
	⋮		

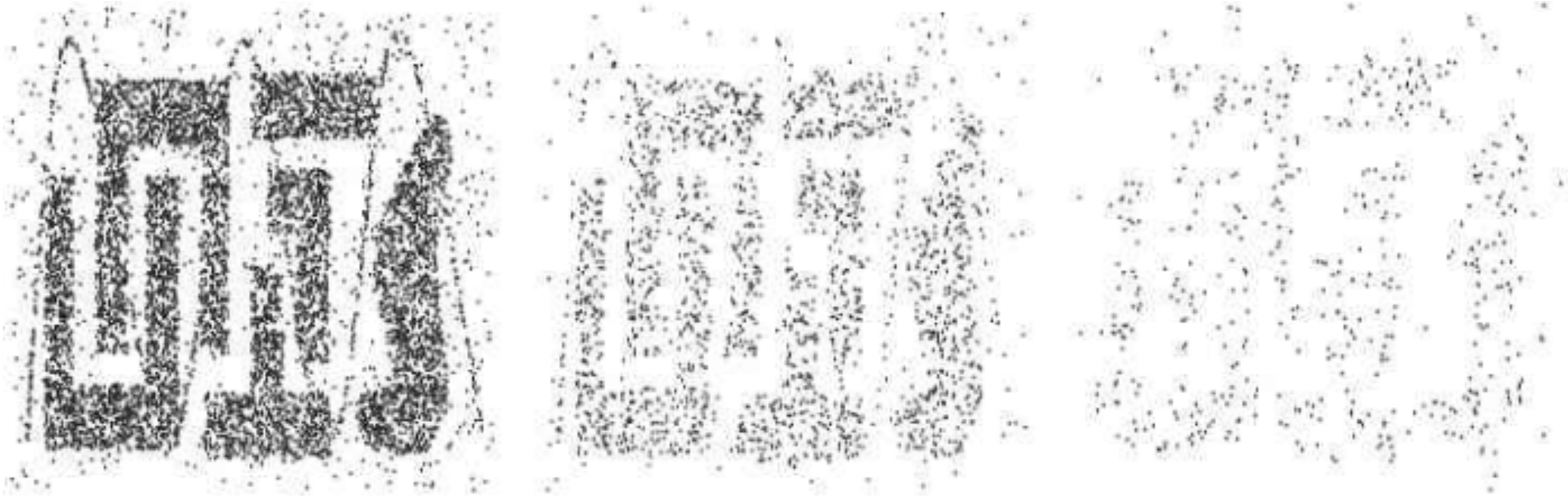
Sampling = Leaving out records

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

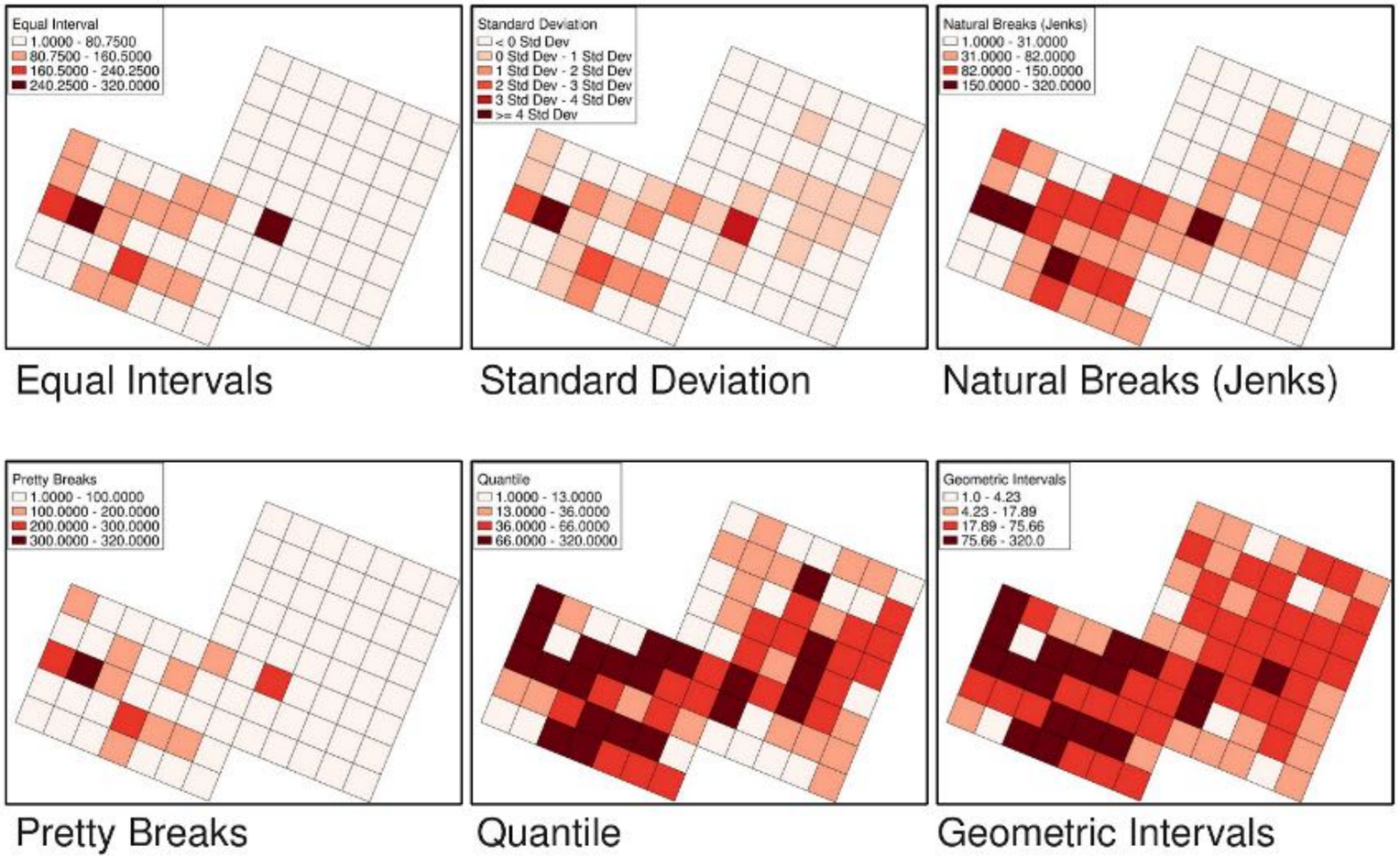
Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

reject
accept
accept

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

Data can be discretized very differently

Example: Weight of finds in an excavation grid

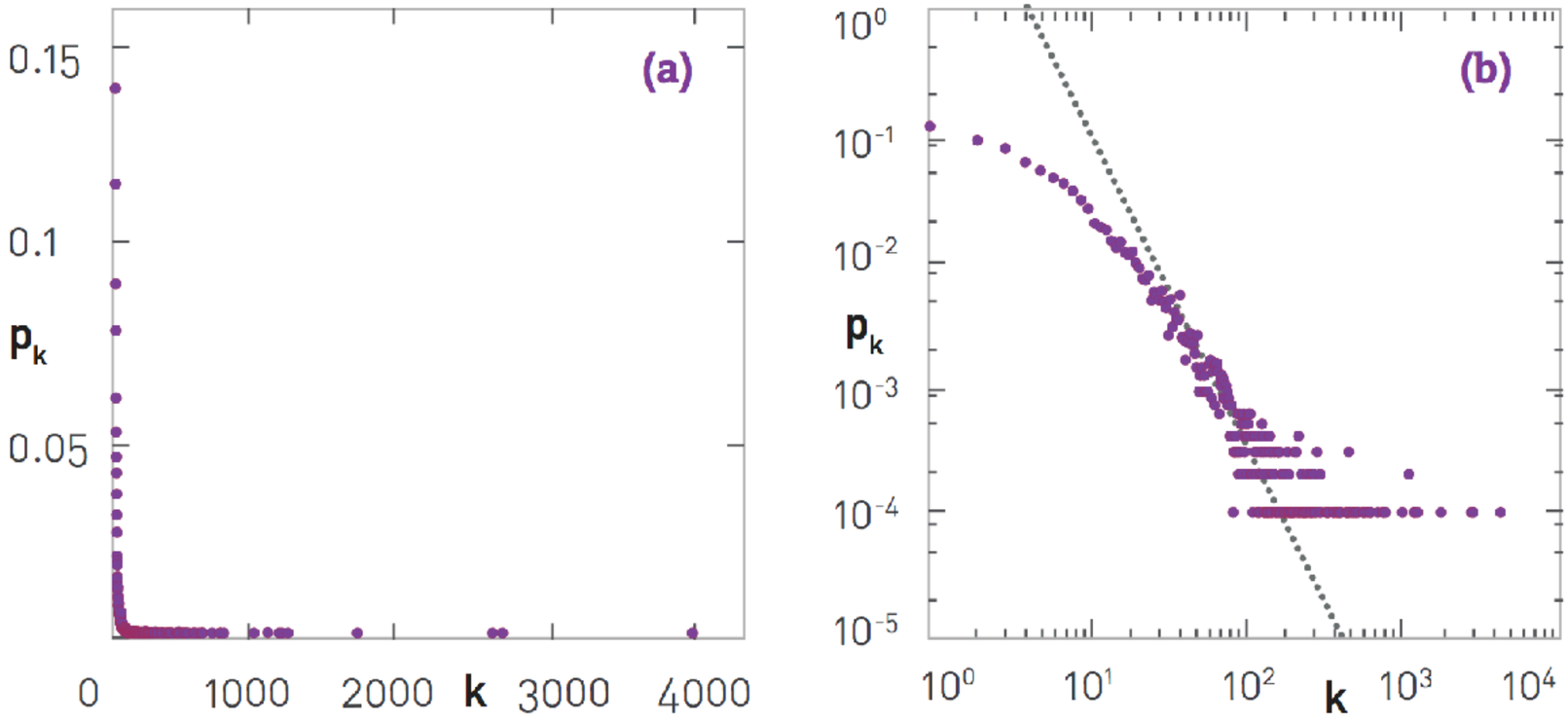


Same data, different split points

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

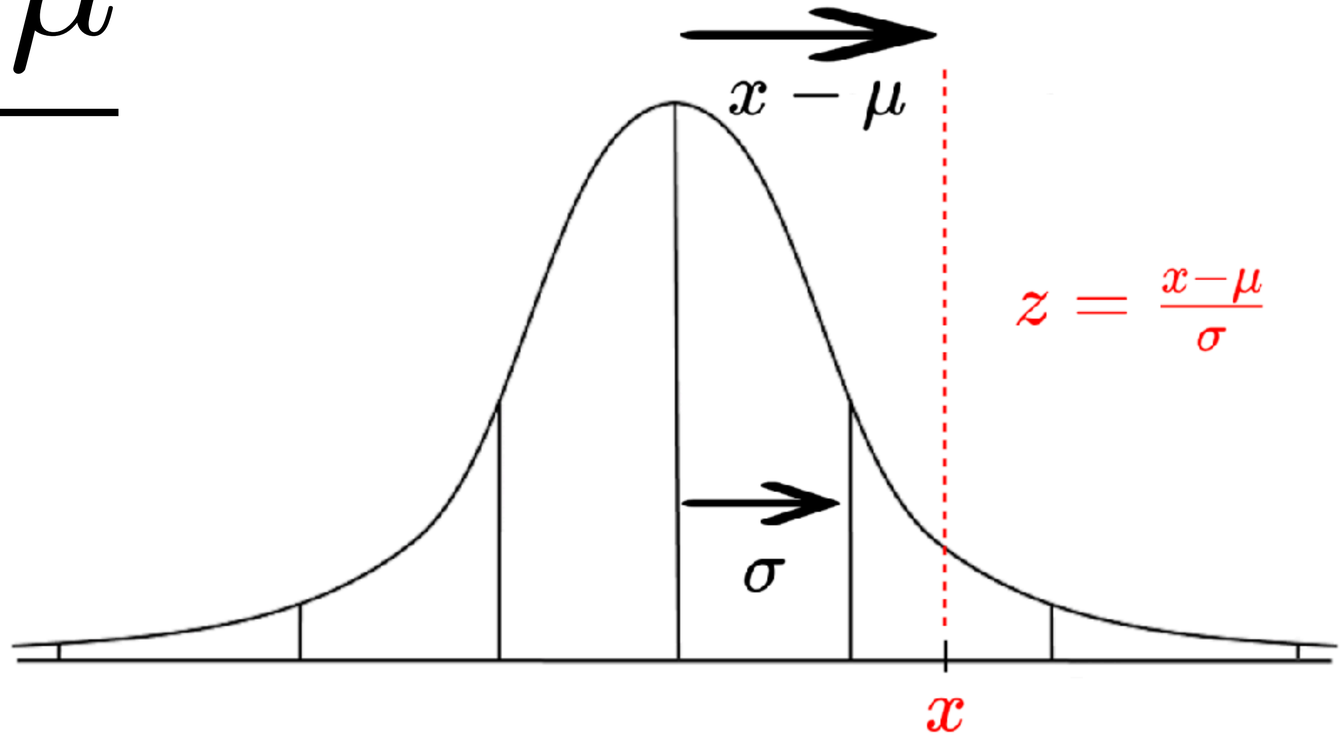


Variable transformation = Apply a function to all values

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?

.

.

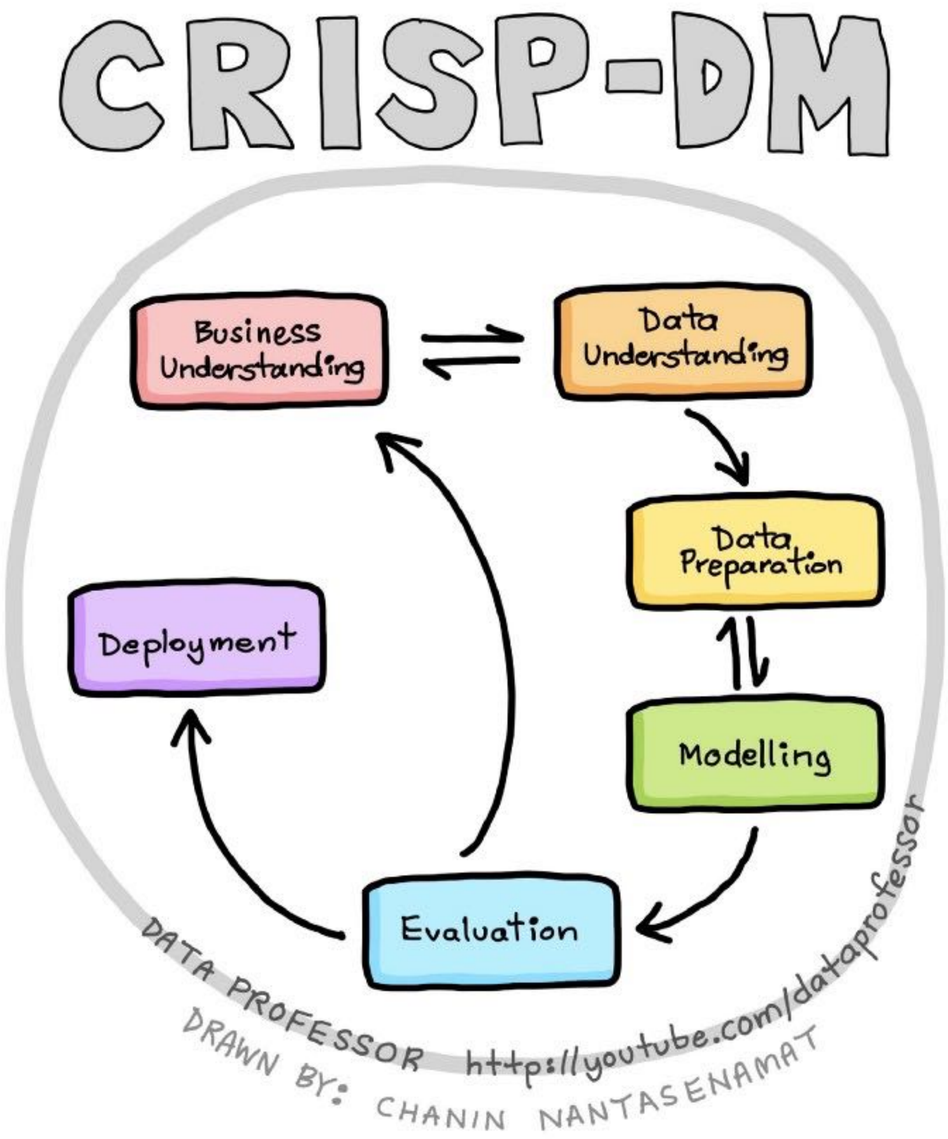
.

Therefore, data cleaning IS analysis!

You cannot separate the two.

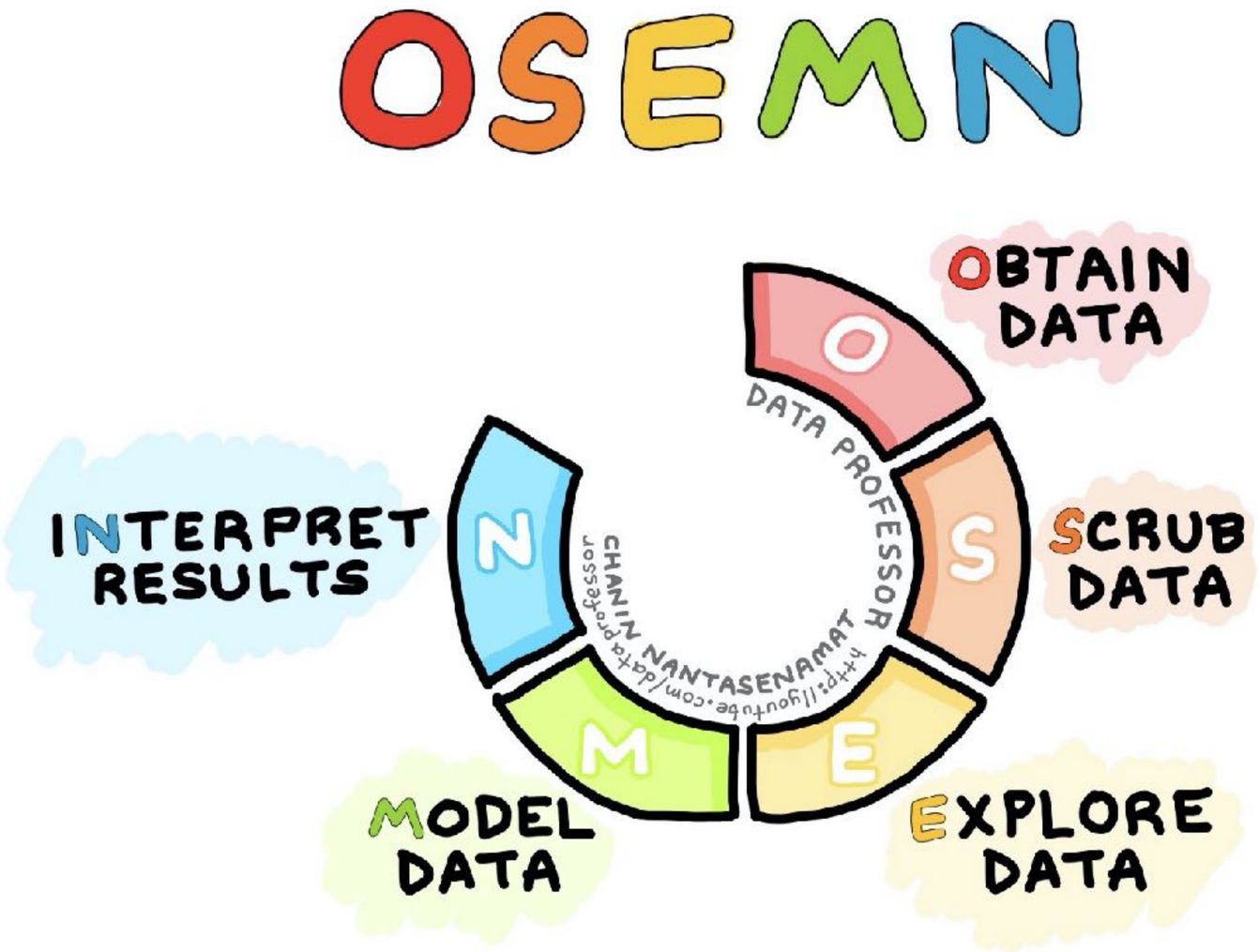
Randy Au, datacleaningisanalysis.pdf

CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (2000)



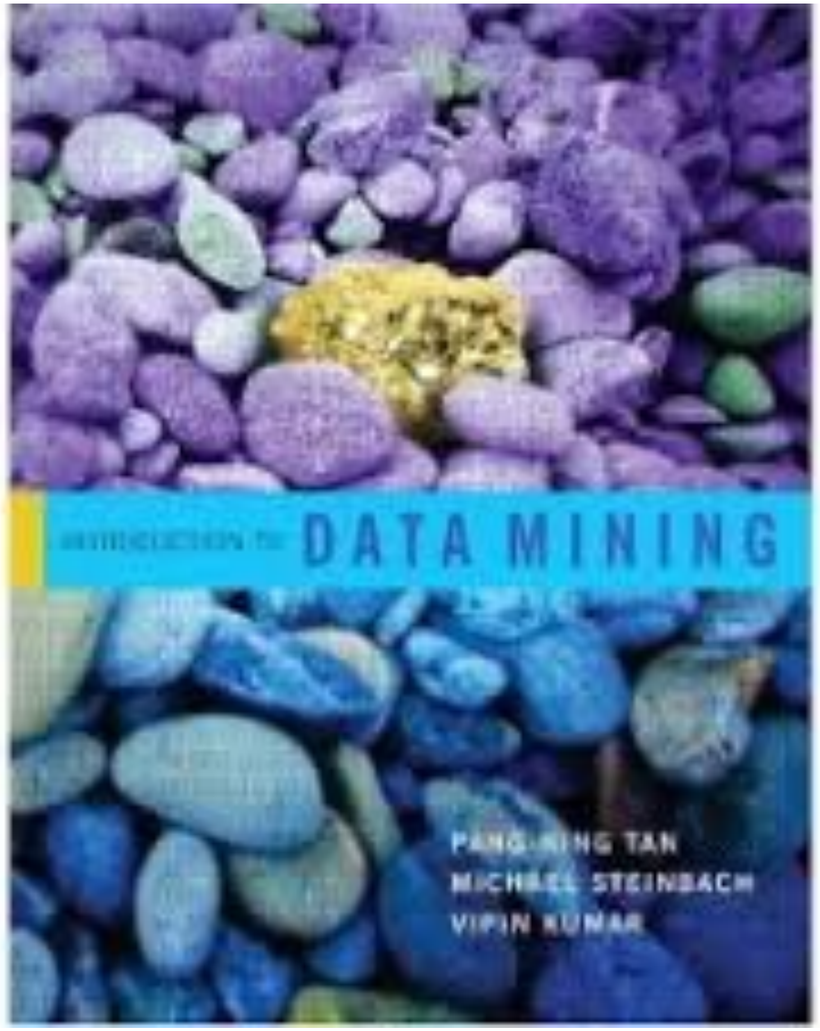
https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

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



https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

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).