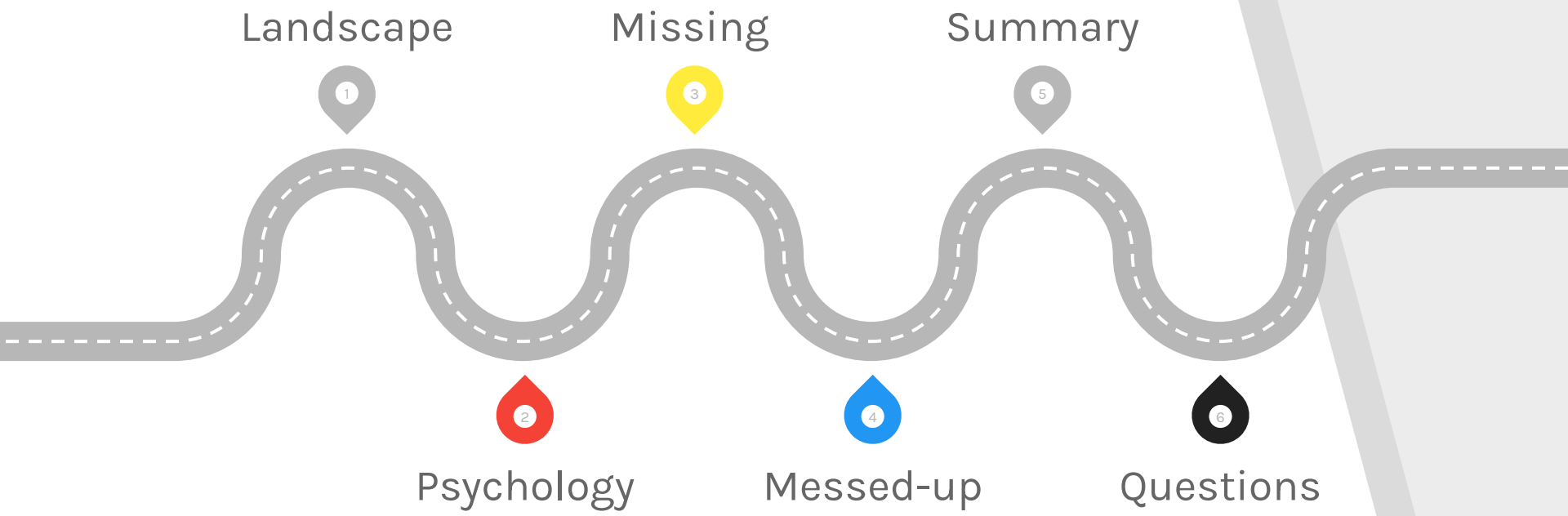
The Colour of Cleaning

ODSC Europe 2021

Marta Markiewicz
InPost, Wroclaw University of Economics and Business

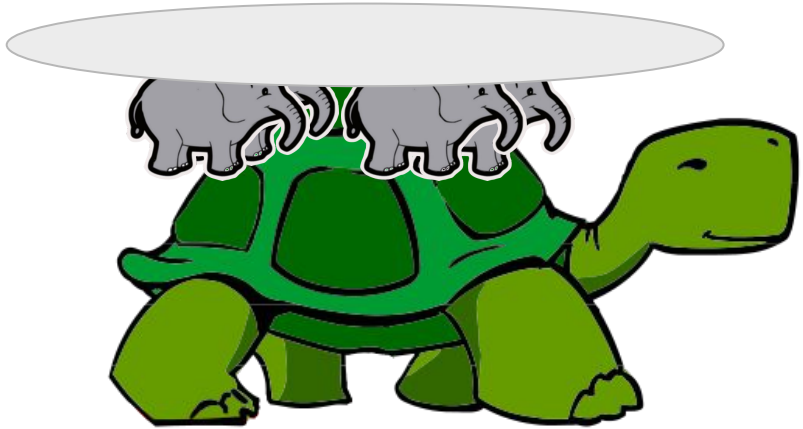
ROADMAP


Landscape
1
Psychology
2
Missing
3
Messed-up
4
Summary
5
Questions
6

This Thing Called Cleaning

Why care?

Data



https://clipartix.com
https://imgur.com/gallery/dQcny7b

Business
Cleaning
Modelling
Deployment

Manually introduced data

No proper system for data input

Incorrect data types

Wrong formats

Out-of-domain observations

All possible dates formats

Typos

Single value columns

Lack of uniqueness

Systems integrations

Duplicated not bringing value

Missing data

Improperly gathered data

Improper cross field dependencies

Lack of data input rules

1.

Psychological barriers

RoomEscapeArtist.com

Why the dislike?

Pressure

Satisfaction

Never ending

Personal development

ual & boring

Fame

No fame

Hack your psychology

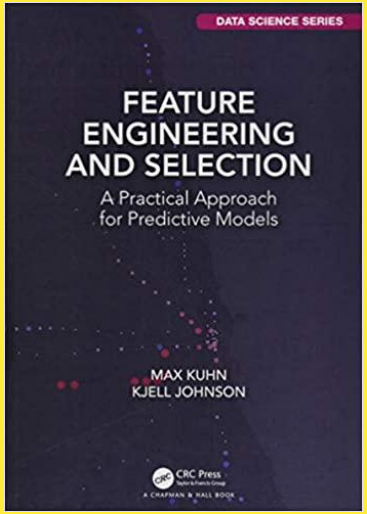
High iterativeness

Evangelisation

ML for cleaning

2.

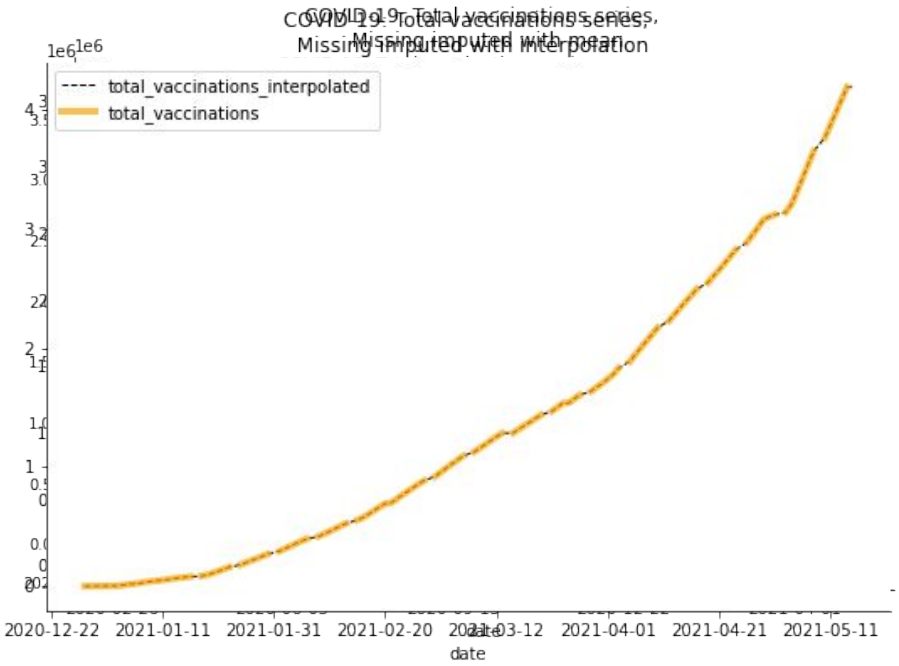
Missing data imputation



Simple approaches

73% empty

1 out of 60



Dropping column

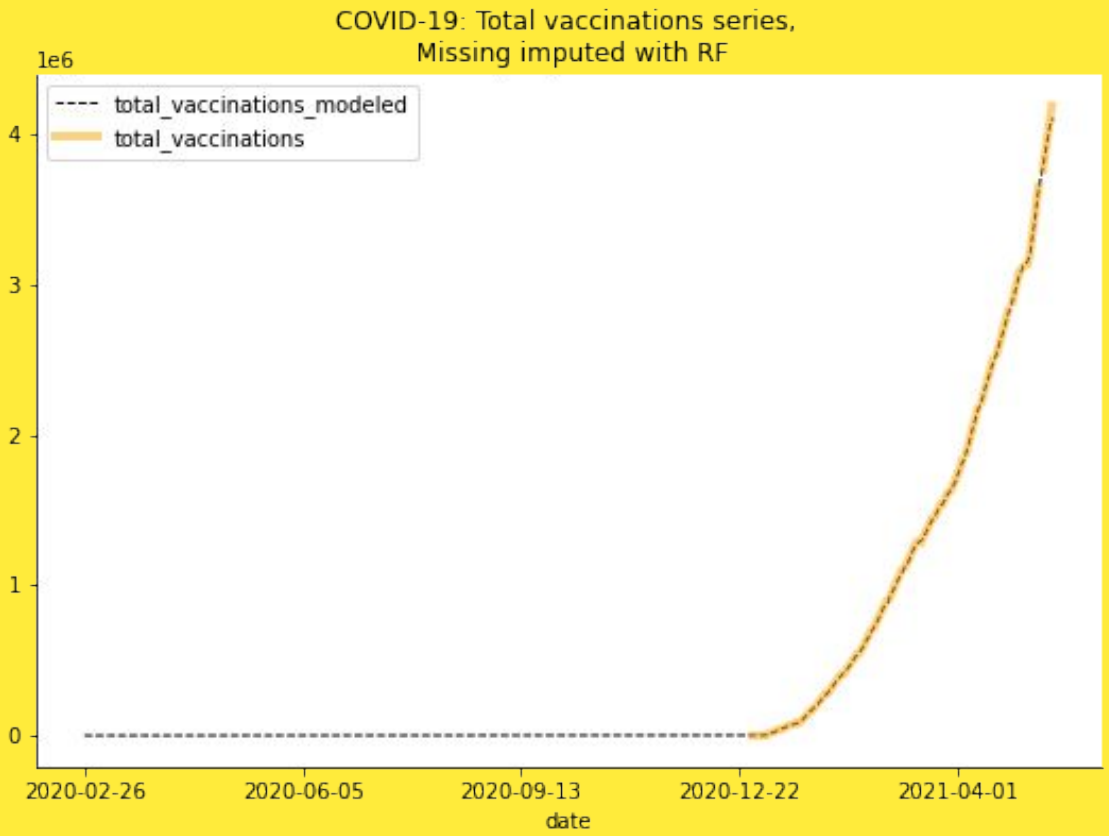
Dropping rows

Mean / median / other statistic

Last observation carried forward
Next observation carried backward

Interpolation

Building a model



missForest

Pclass	Gender	Fare
1	female	71.2833
3	female	7.925
1		263
1	male	27.7208
3	male	7.8958
1	female	
2		26

Pclass	Gender	Fare
1	female	71.2833
3	female	7.925
1	male	263
1	male	27.7208
3	male	7.8958
1	female	32.17
2	male	26

Pclass	Gender	Fare
1	female	71.2833
3	female	7.925
1	male	263
1	male	27.7208
3	male	7.8958
1	female	107.08
2	male	26

Pclass	Gender	Fare
1	female	71.2833
3	female	7.925
1	female	263
1	male	27.7208
3	male	7.8958
1	female	107.08
2	male	26

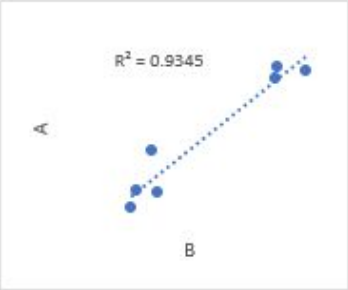
Gender	Fare
0	0
0	0
1	0
0	0
0	0
0	74.91
0	0

Multiple iterations!

MICE - Multiple Imputation by Chained Equations

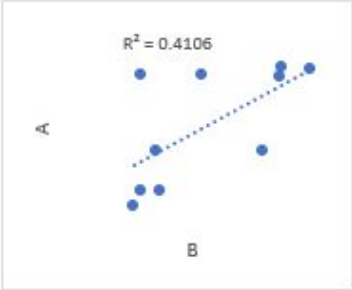
Missing data is in red. There is a strong correlation between A and B, so let's try to impute A using B and C.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



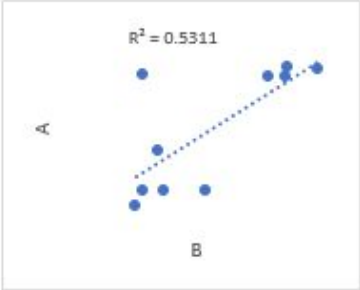
Missing data is filled in randomly. This dillutes the correlations, but allows us to impute using all available data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45



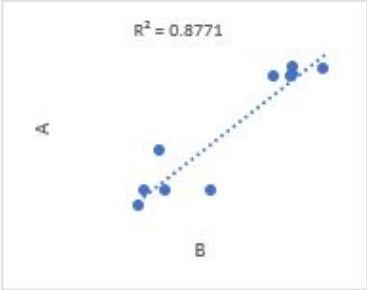
A random forest is used to predict A with B and C. Notice the correlation between A and B improved.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45



After Imputing B using A and C, we have achieved a correlation between A and B much closer to the original data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45



PMM - Predictive Mean Matching

The predicted value of A (E[A|B,C]) is shown to the left. We are interested in imputing the bold missing value below

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

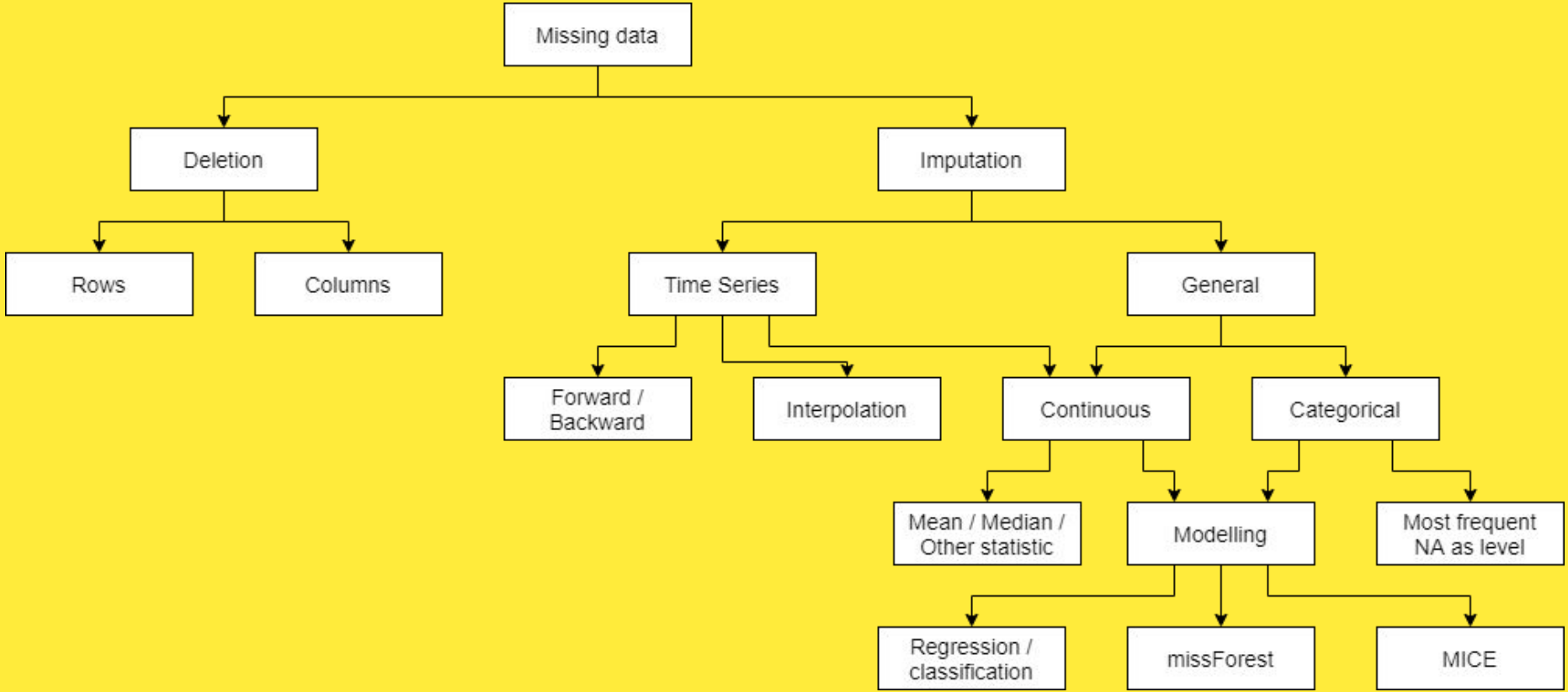
Our predicted value for the first missing sample is 0.60. The closest predicted value is 0.62. We find the closest values for all of our missing samples.

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

We then impute the value corresponding to the original data.

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60	0.24	0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20	0.89	1.14	1.28
1.24	0.89	1.23	1.45

Summary


Missing data
Deletion
Imputation
Rows
Columns
Time Series
General
Forward / Backward
Interpolation
Continuous
Categorical
Mean / Median / Other statistic
Modelling
Most frequent NA as level
Regression / classification
missForest
MICE

3.

Messed up data

LAST time

Text

Typos

Inconsistent categories

Swaps

Dates in Time Series

Geo for Map

Incorrect target labels

Active quality learning

Labels

One dimensional

Multi dimensional

Anomalies

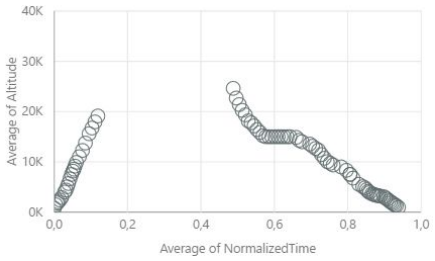
Incorrect target labels

	VOC 2012	COCO 2017	Udacity – Self Driving Car
Training Set Images	17 177	94 439	11 992
Training Set Labels	20	80	9
Training Set Objects	49 834	686 385	78 230
Manual Correction	**21.1%**	**23.6%**	**25%**

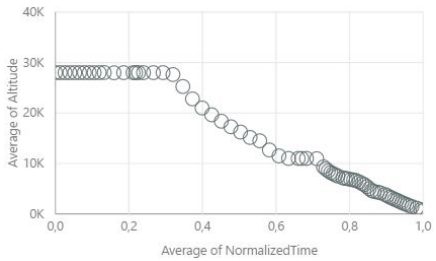
https://deepomatic.com/en/how-we-improved-computer-vision-metrics-by-more-than-5-percent-only-by-cleaning-labelling-errors

Breaking quality issues

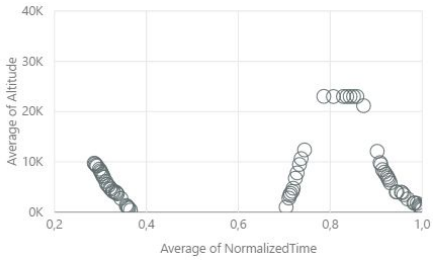
Lack of middle phase



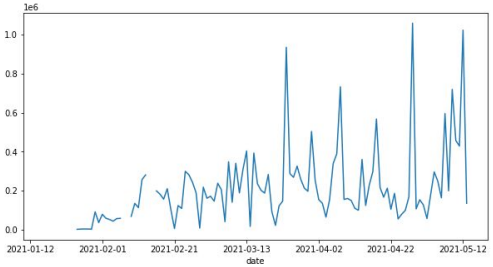
Lacks half of a flight



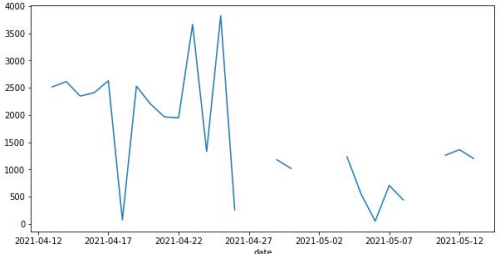
Data gathering error



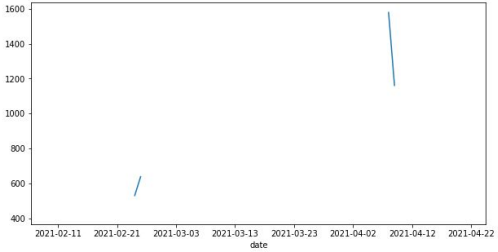
Complete vaccinations series



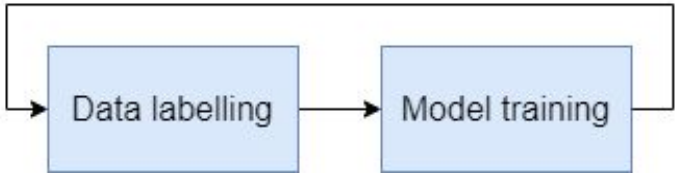
Incomplete vaccinations series

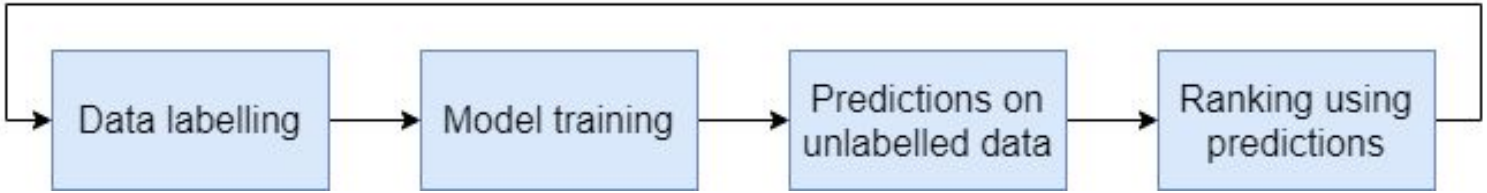


(Almost) absent vaccinations series

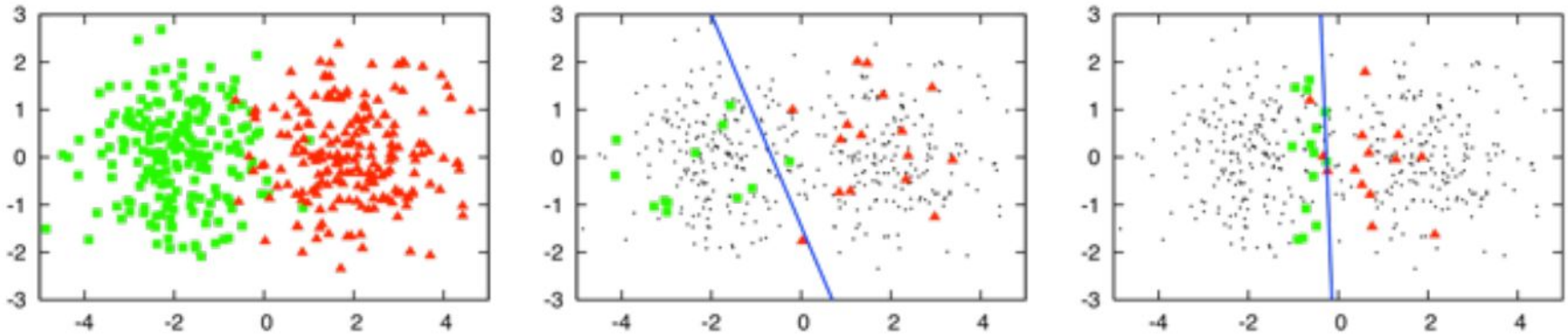


Active learning



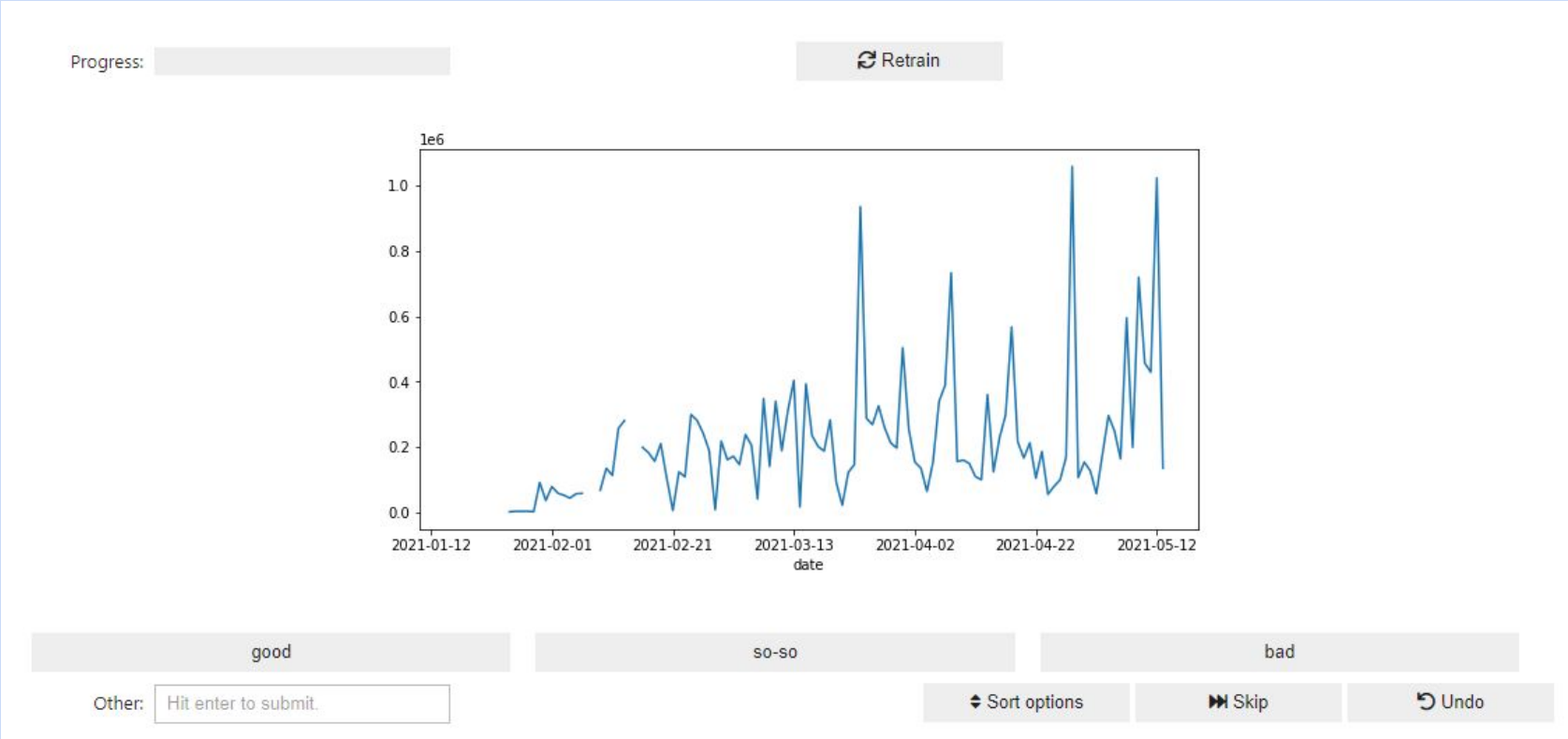


Active learning - intuition

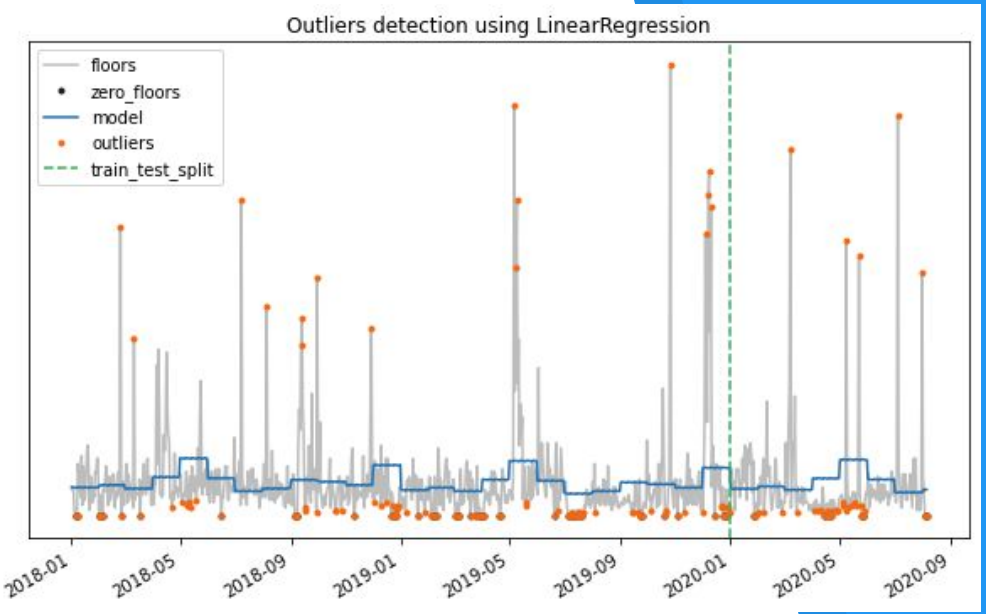


https://towardsdatascience.com/introduction-to-active-learning-117e0740d7cc

Active quality learning



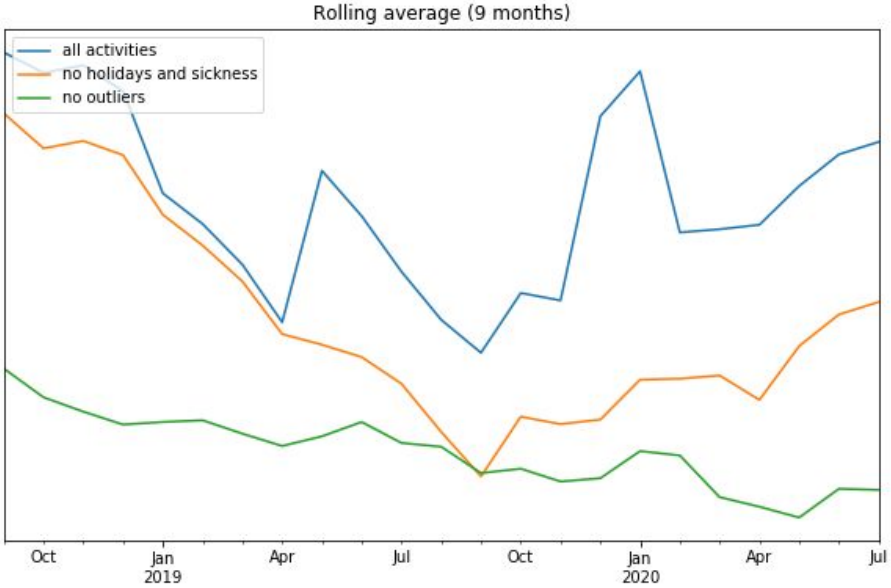
Anomalies - examples

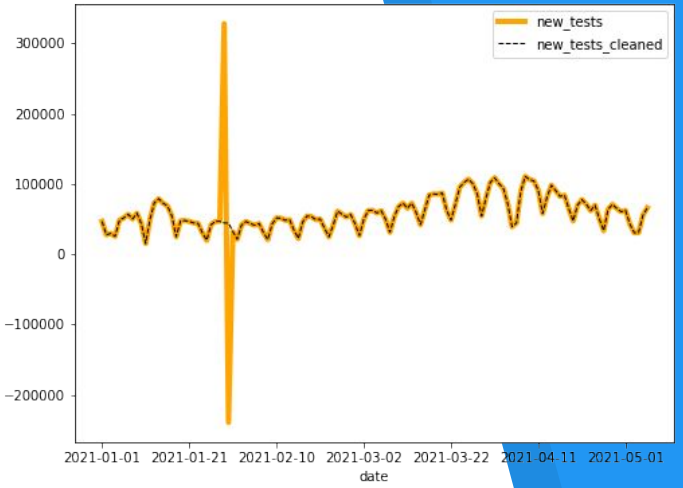


Anomalies

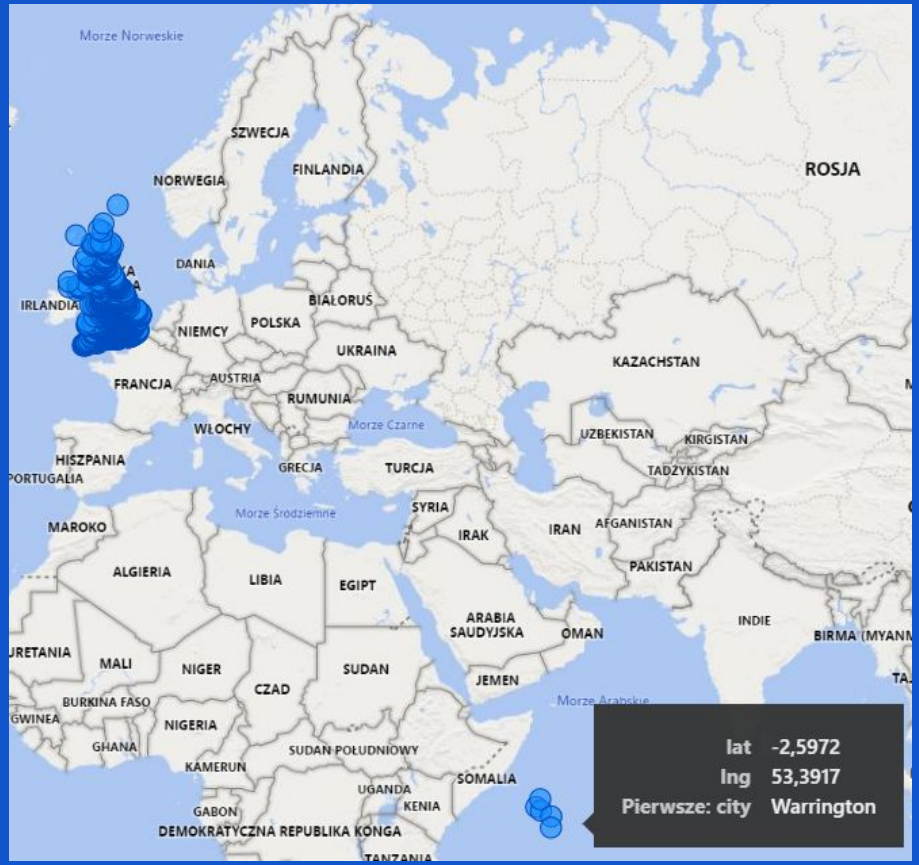
	Intended supervised	A side effect supervised	Unsupervised	Mixed
Manual	A lot	No	No	A little
Immunity to pattern changes	No, unless repeated periodically	Yes	Yes	Partial

Anomalies - what to do with them?



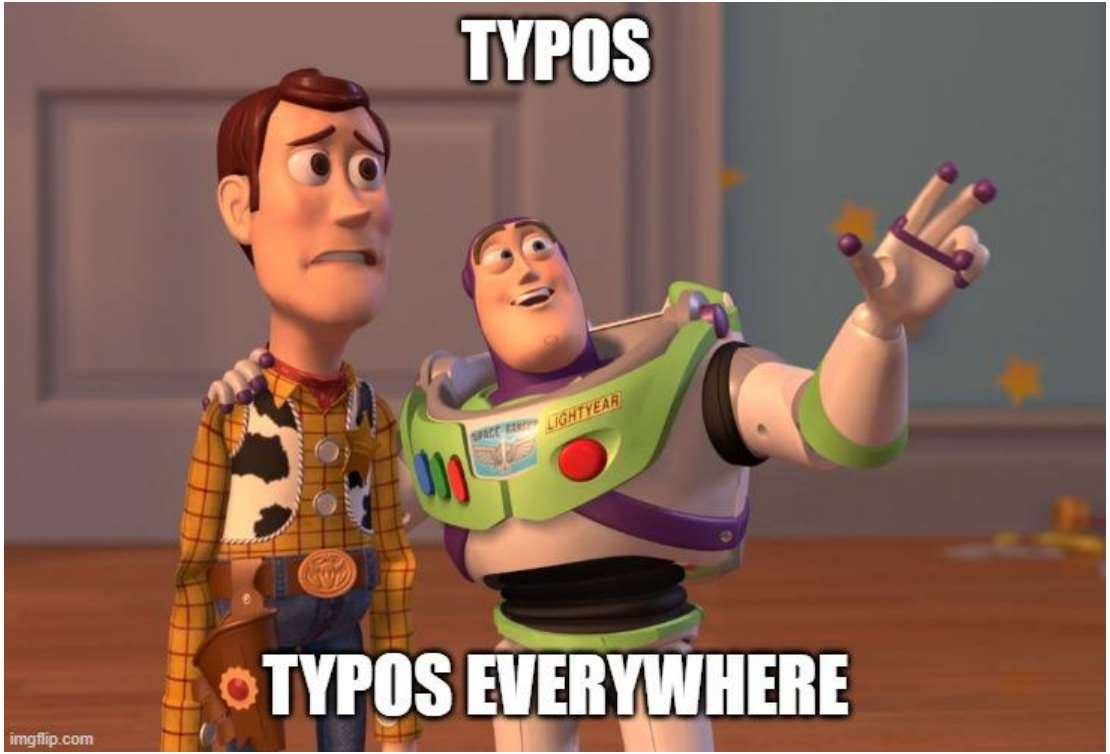


Geo



1. Anomalies detection
2. Swapping
3. Anomalies detection

Typos world



male	375
female	200
Mr.	68
facet	63
monsieur	51
femme	27
kobiet	25
m ale	20
mujer	19
kobieta	18
dziewczyna	16
famale	9

NLP evolution

Bag-of-Words

TF-IDF

oVe

FastText

Transformer

ELMO

GPT

BERT

RoBERTa

ERNIE 2.0

GPT-2

XLNet

Sentence-Bert

GPT-3

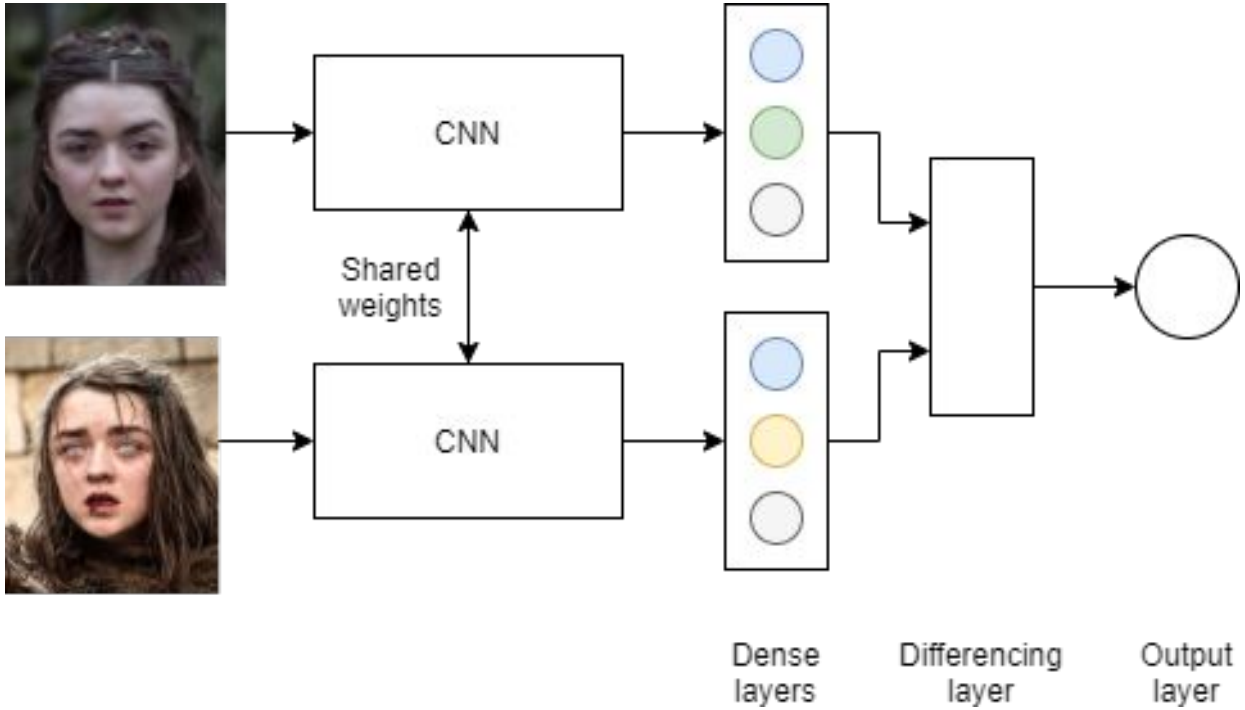
Big Bird

1954

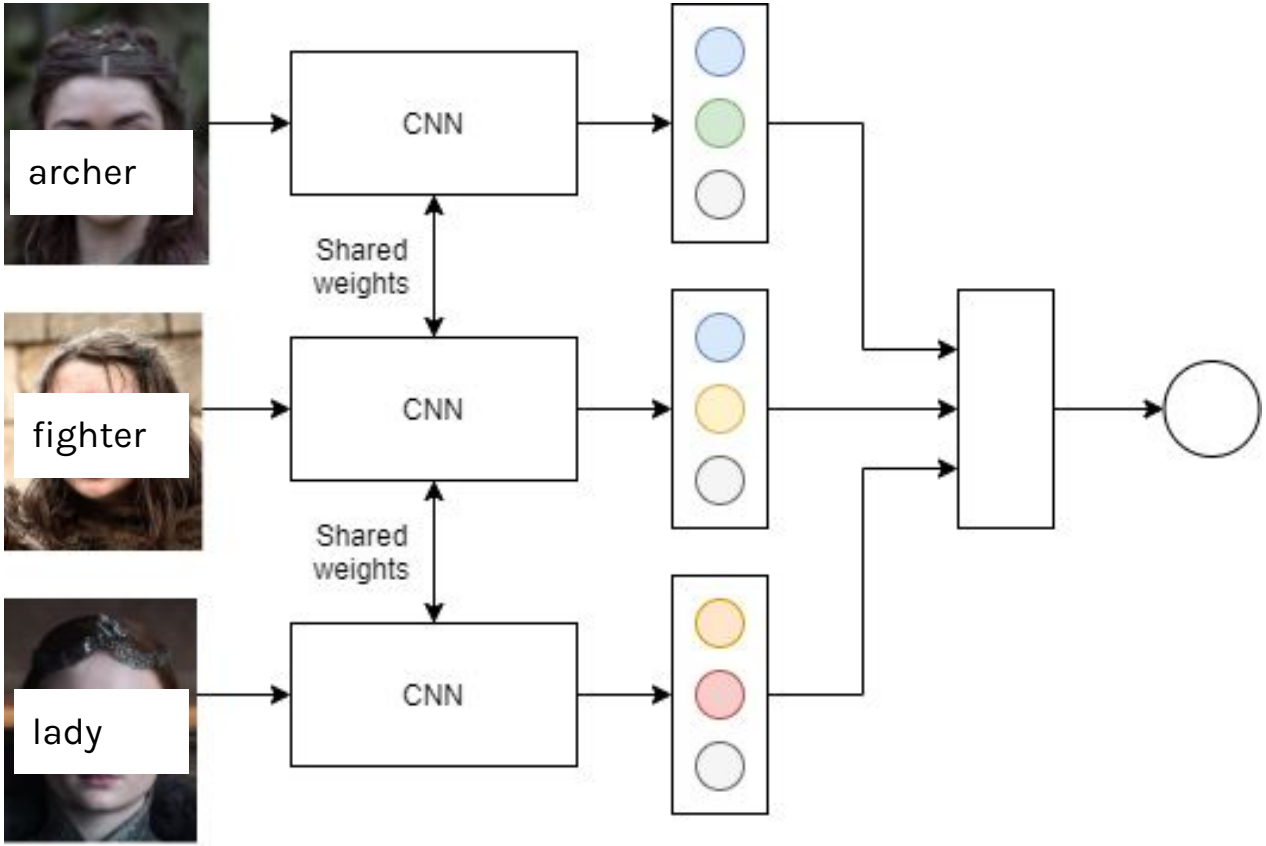
2013

2017

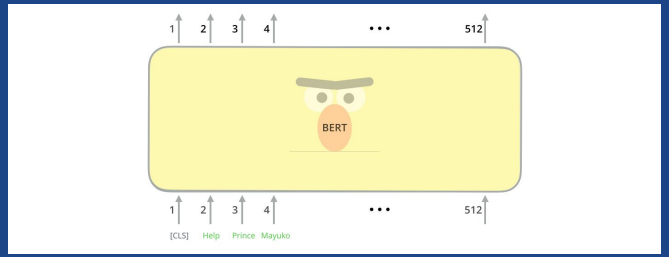
Siamese / Triplets Networks

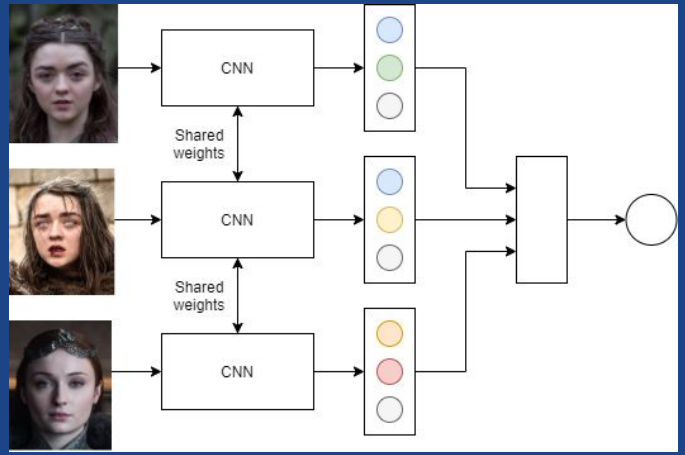


Siamese / Triplets Networks

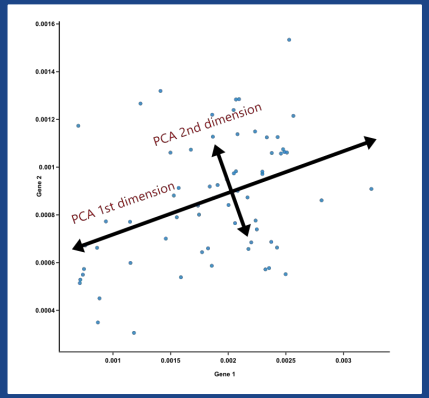


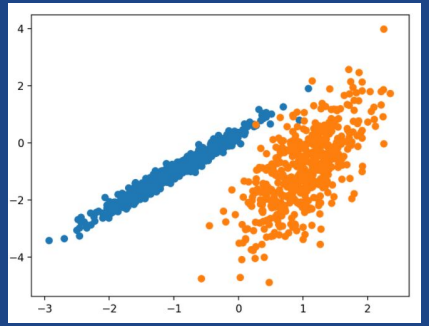
Embeddings + dimension reduction / clustering





https://arxiv.org/pdf/1908.10084.pdf

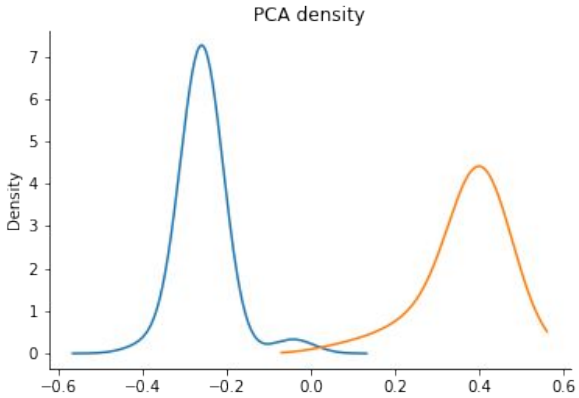


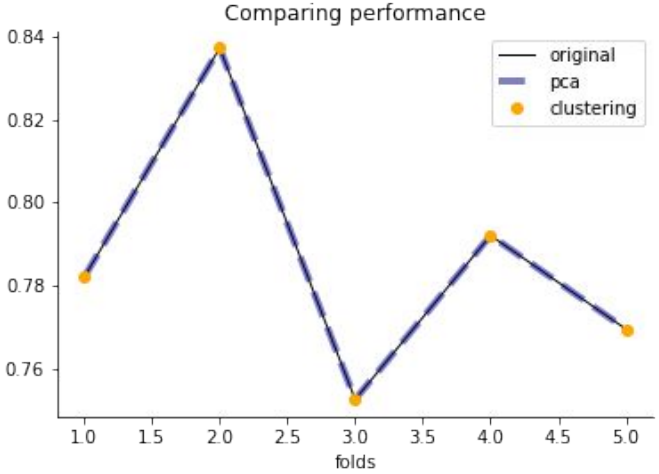


https://machinelearningmastery.com/clustering-algorithms-with-python/

Cleaning categories

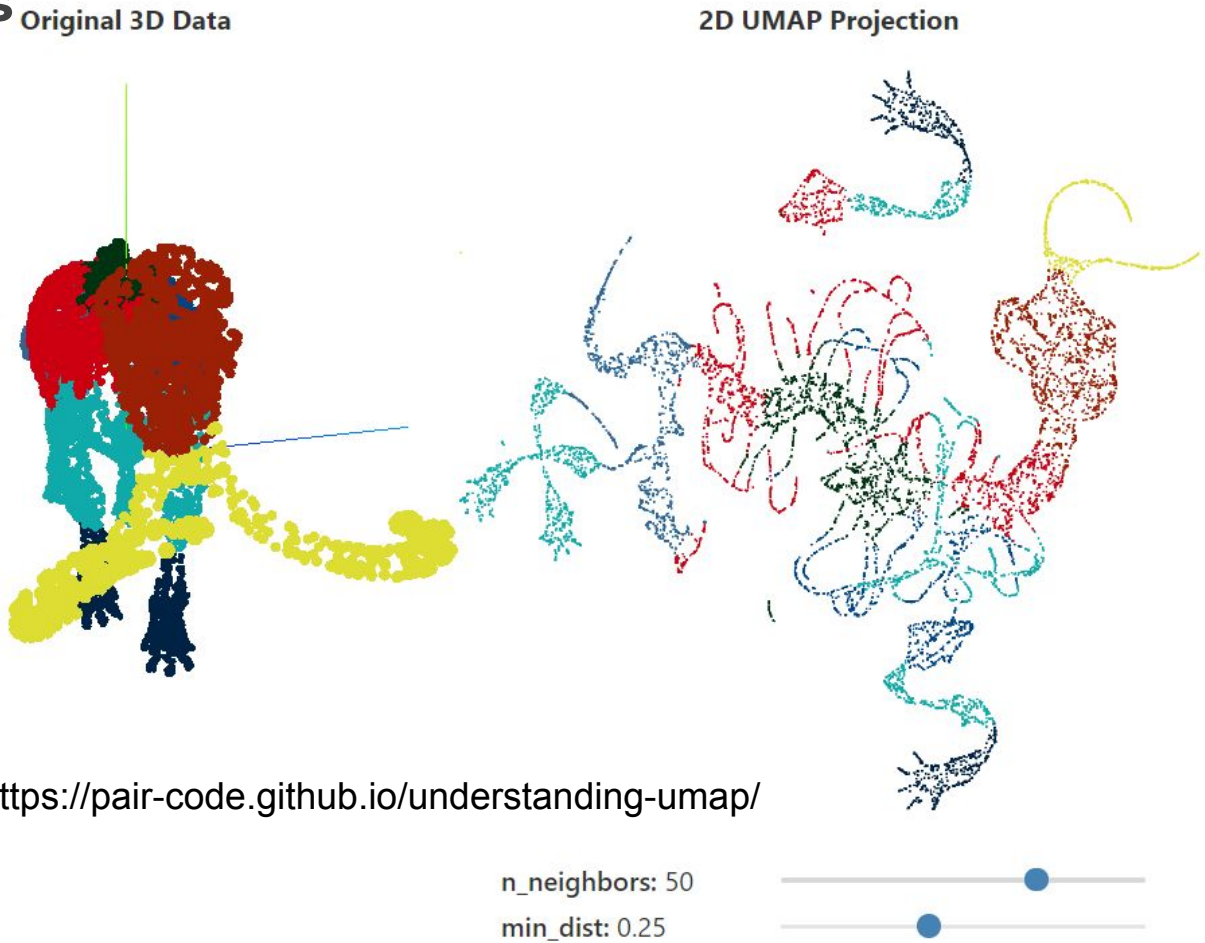
male	462
female	251
mle	24
man	24
monsieur	19
m ale	18
facet	16
mujer	14
Mr.	14
kobiet	11
kobieta	11
femme	10
femalle	10
dziewczyna	7





Cleaning categories

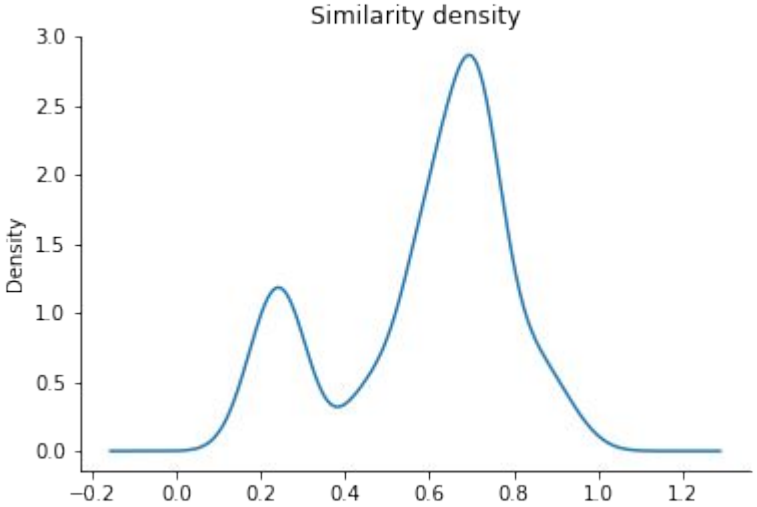
male	462
female	251
mle	24
man	24
monsieur	19
m ale	18
facet	16
mujer	14
Mr.	14
kobiet	11
kobieta	11
femme	10
femalle	10
dziewczyna	7



https://pair-code.github.io/understanding-umap/

Identifying improper categories

category	subcategory
house	bungalow
	bus
	cottage
	log-house
	sem-detached house
	semi-detached house
	short-sleeved
vehicle	calf-length
	car
	caravan
	lorry
	lory
	minivan
	motorcoach
	motorcycle
	truck
	two-storey hous



category	subcategory	similarities
vehicle	two-storey hous	0.204250
house	short-sleeved	0.260327
vehicle	calf-length	0.264653
house	bus	0.450055
vehicle	caravan	0.553825

Key takeaways

- Iterate
- Evangelate
- Leverage ML:
  - Regression / classification
  - Multiple imputation
  - Quality active learning
  - Anomalies detection
  - Embeddings
  - Dimension reduction
  - Clustering
  - …

Thank you!

Any Questions?

Marta Markiewicz

m.markiewicz.pl@gmail.com

https://github.com/lady-pandas/cleaning-is-coming