Performance Evaluation
(Epistemological challenges in distant seeing)

A. Nicolaou

Automating the humanities

Machine learning is like magic!

Does this makes us wizards?

Automating the humanities

Machine learning is like magic!

Does this makes us wizards?

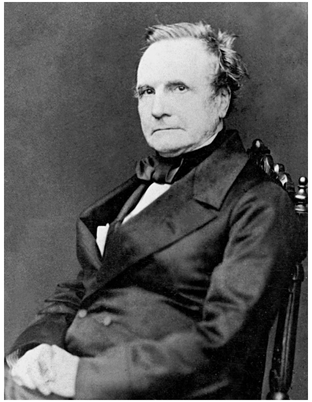
Only apprentices!

Sometimes confusion is obvious

-Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?

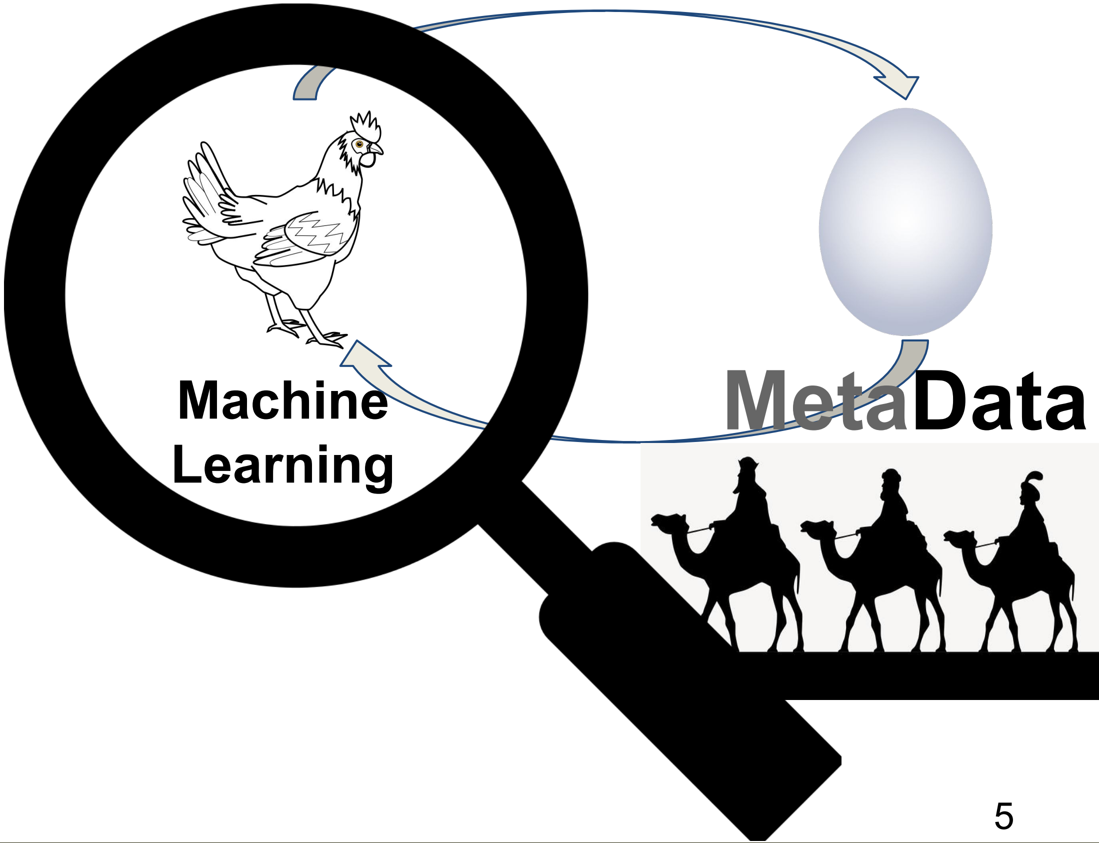
-I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question

C. Babbage 1791- 1871



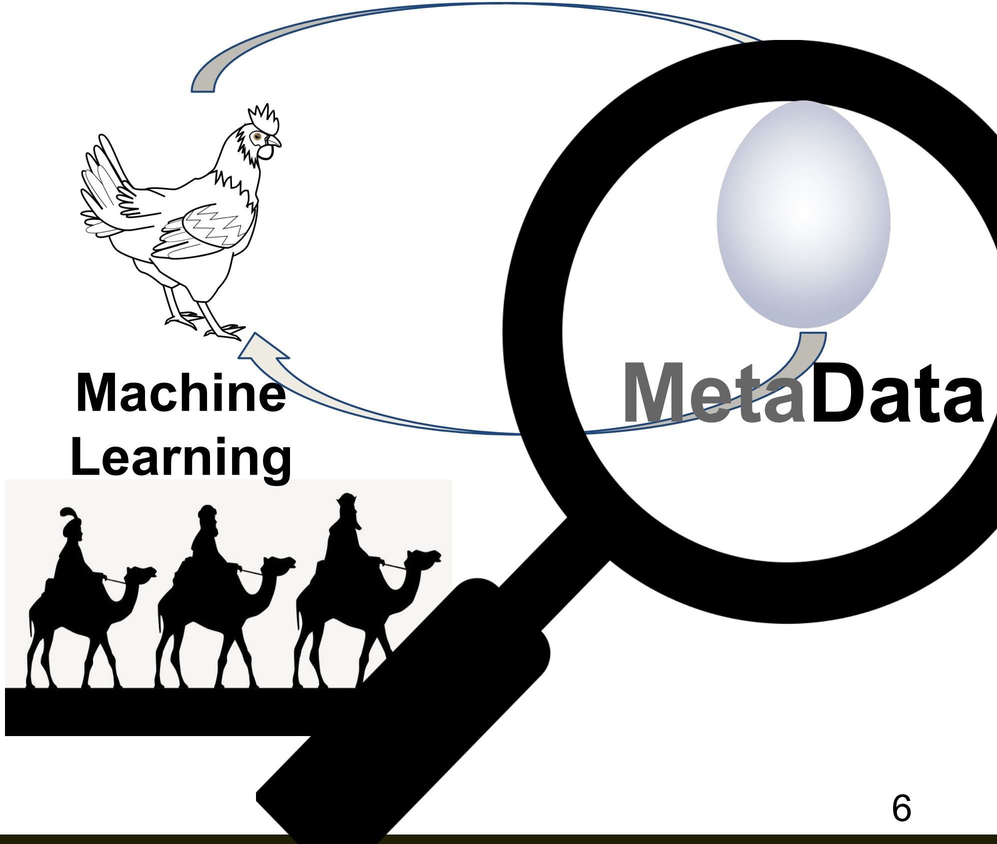
An engineers perspective

- **I scrutinise the machine: learning method**
  - **Does it work?**
  - **Does it overfit?**
  - **Is it reliable?**
- **I trust the data:**
  - **It is uncontested**
  - **It is meaningful**
  - **It was scientifically sampled**
  - **It is objective**



A humanicist's perspective

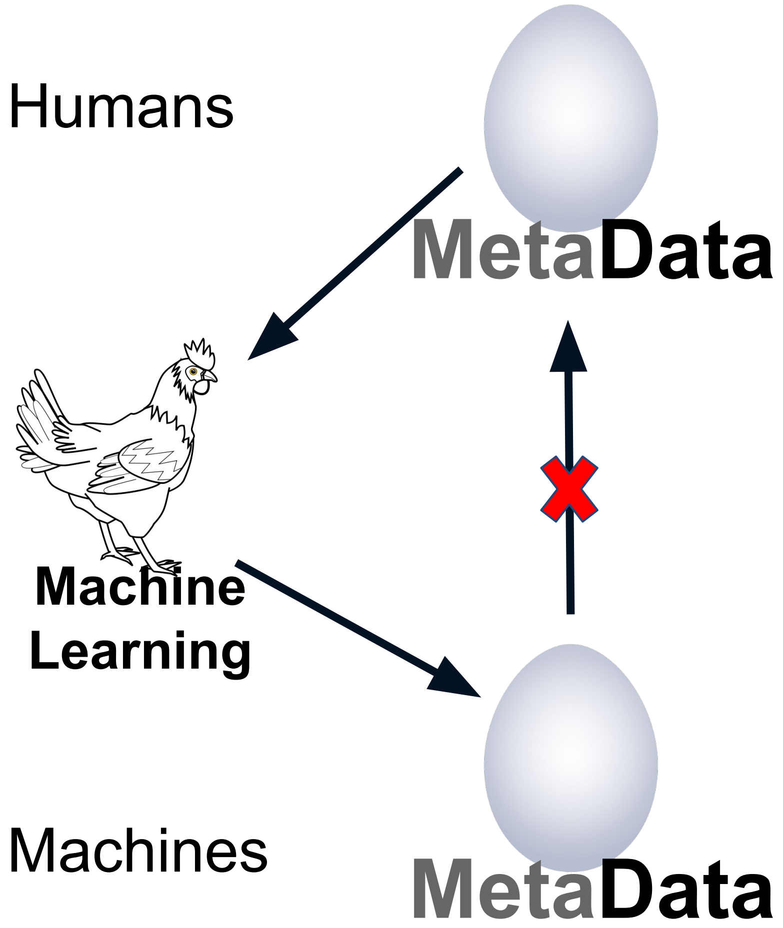
- I trust the machine learning method
  - It kind of works
  - Once it's trained we can use it
- I work on the data
  - I understand it
  - I assign nuanced
  - It was scientifically compiled



Methodology

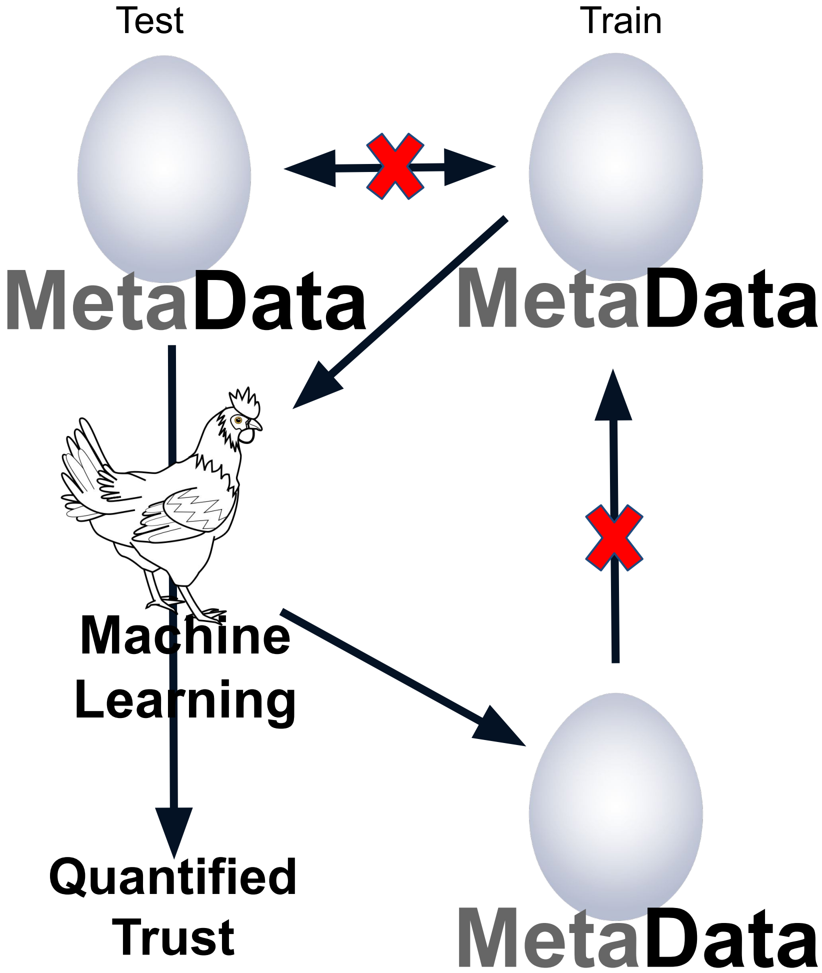
- ML / CV Research:
    - You **trust** the data and analyse the method
- Domain specific research / Diplomatics:
    - You **trust** the method and analyse the data
- **Trust**:
    - externalising the source of doubt
    - requires consensus among experts
- **"If you torture the data long enough, it will confess to anything" R. Coase 1910-2013**
- Interdisciplinary research:
    - We can no longer "externalize" responsibility

- "If you torture the data long enough, it will confess to anything" R. Coase 1910-2013
- Interdisciplinary research:
    - To unconstrained
- Solutions:
    - **Never mix human generated data with machine generated data.**
    - Always keep separate data for performance evaluation
    - Commit to any experiment measurement before the results are out.



Methodology

- "If you torture the data long enough, it will confess to anything" R. Coase 1910-2013
- Interdisciplinary research:
  - To unconstrained
- Solutions:
  - Never mix human generated data with machine generated data.
  - **Always keep separate data for performance evaluation**
  - Commit to any experiment measurement before the results are out.



Probability

- A precise estimate on the chances of something occuring that can be theoretically founded
- Can be interpreted as probability:
  - Any set of variables between 0 and 1 that sum up to one
  - Any single variable that is bound between 0 and 1
  - The math will work but the answers might be wrong

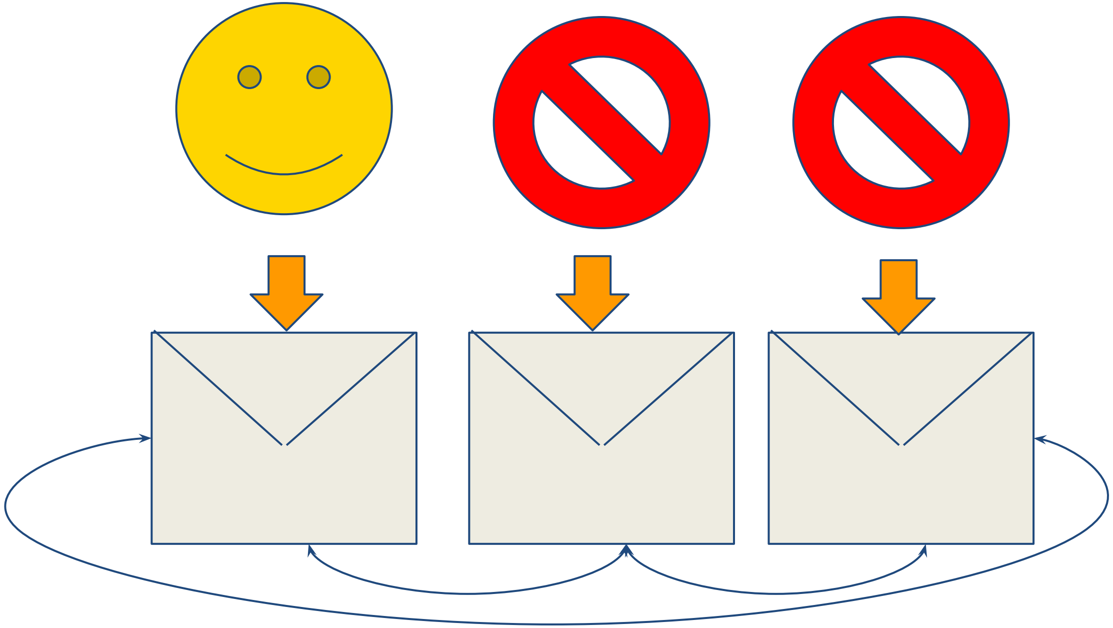
Probability

- Every time we assume a mathematical model can be used for something, we make an axiom out of a hypothesis

Probabilities are tricky!

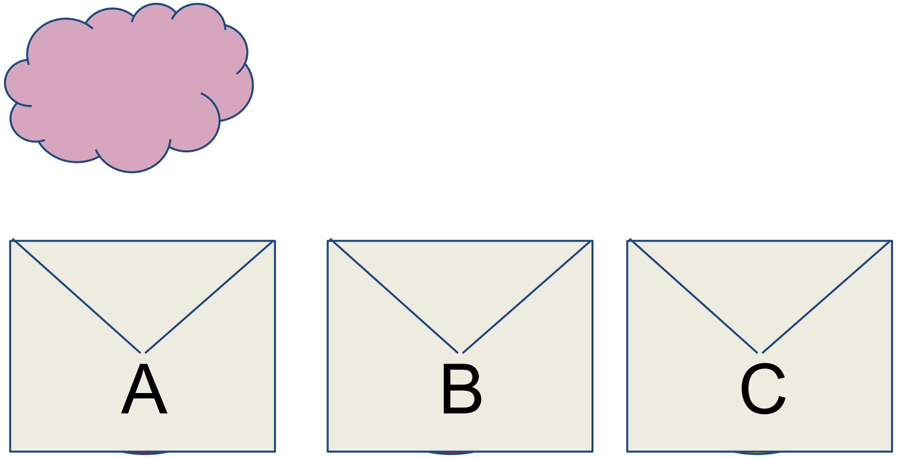
- One winning ticket, Two losing tickets

Probabilities are tricky!



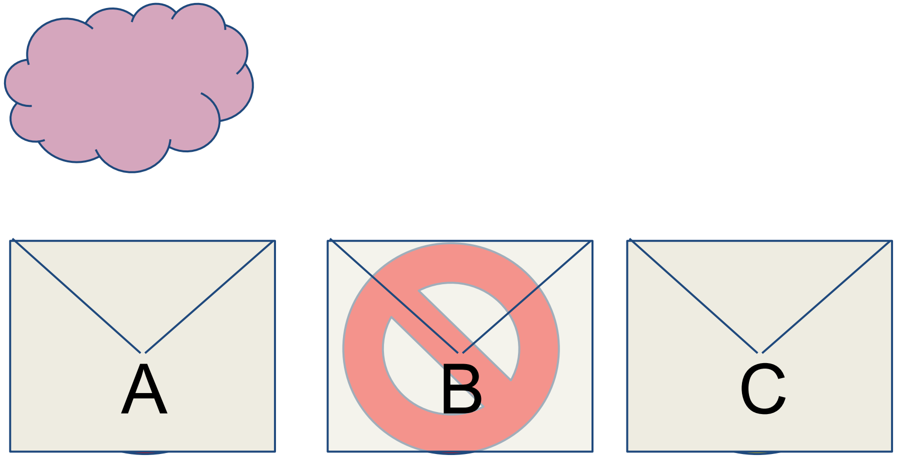
- Are put into three envelopes which are then shuffled

Probabilities are tricky!



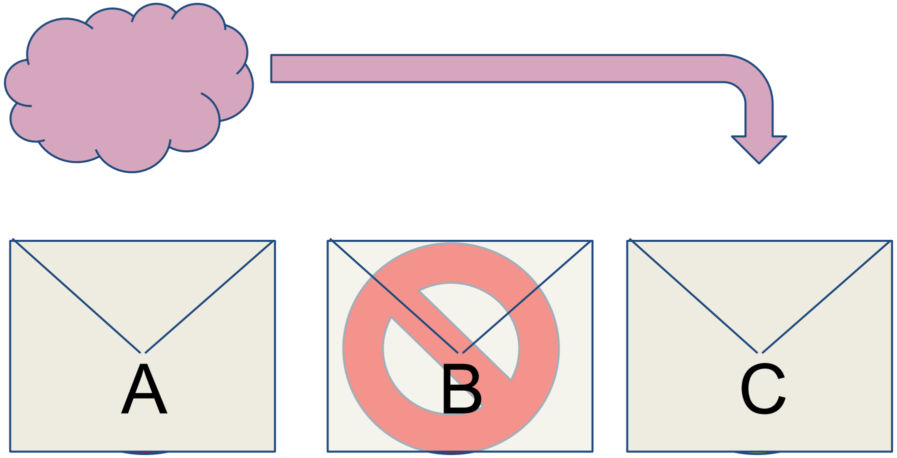
- Let's assume you choose envelope A

Probabilities are tricky!



- Let's assume you choose envelope A
- And then you learn B is a losing ticket

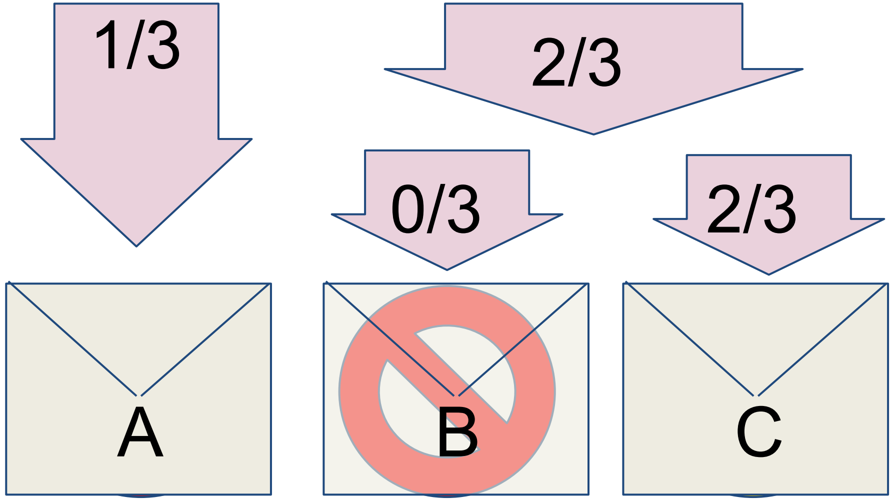
Probabilities are tricky!



- Let's assume you choose envelope A
- And then you learn B is a losing ticket
- Should you switch your choice to C if given the option?

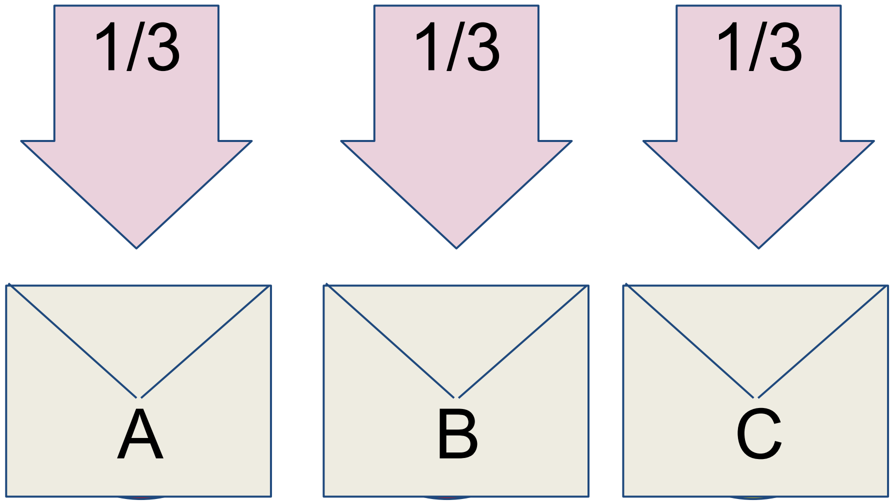
UNI GRAZ

Probabilities are tricky!



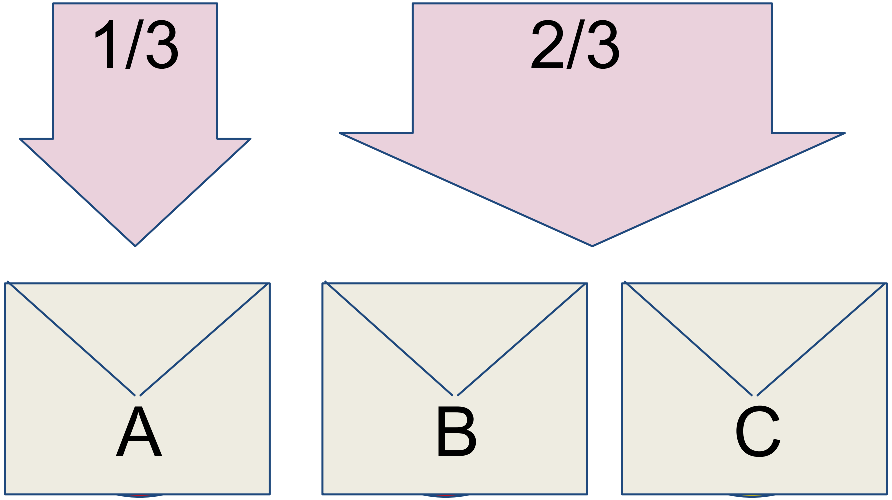
- Let's assume you choose envelope A
- And then you learn B is a losing ticket
- Should you switch your choice to C if given the option

Probabilities are tricky!



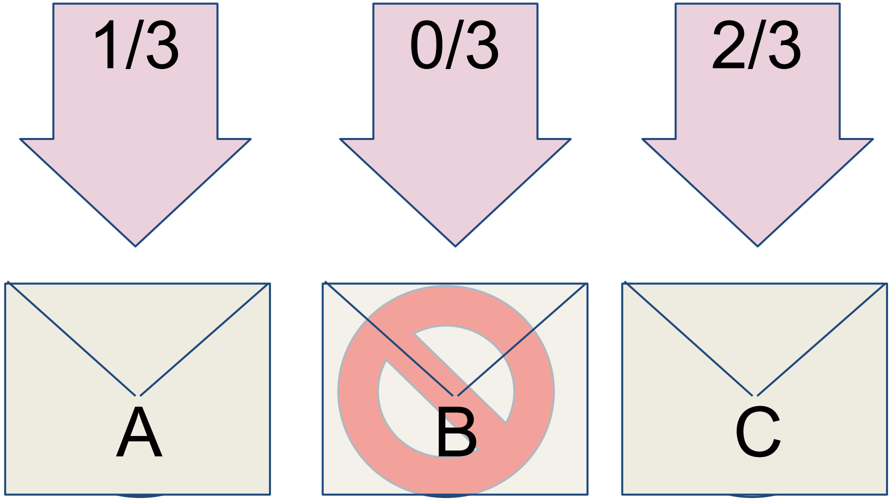
- Before we choose

Probabilities are tricky!



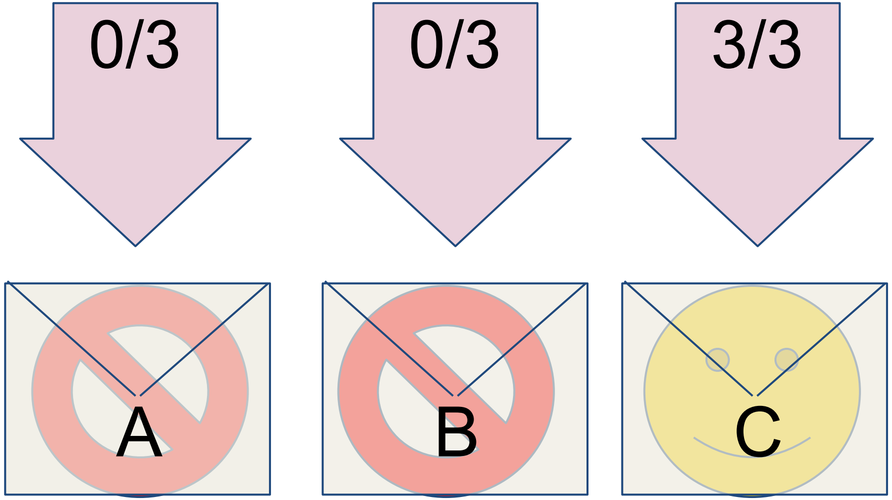
- After we choose

Probabilities are tricky!



- After we learn that B loses

Probabilities are tricky!



- After we learn that C wins

Probabilities are tricky!

- There is always a sampling space Ω
- The sampling space has always P(Ω) = 100%
- In ML performance evaluation, everything has already happened.
  - Who knew what?
  - When did they know it?

Binary Classification

- Classes:
  - Positives
  - Negatives
- Predictor:
  - Classifier
  - Groundtruth

		Predictions	
		Positive	Negative
Groundtruth	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

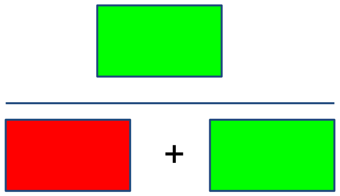
Ω

UNI GRAZ

Binary Classification: Accuracy

- Probability of a correct prediction
- Always remember the best blind classifier

		Predictions	
		Positive	Negative
Groundtruth	Positive	True Positive	False Negative
	Negative	False Positive	True Negative



Binary Classification: Accuracy

- Is 90% Accuracy good or bad?

Binary Classification: Accuracy

- Example problems
- Recto-Verso:
  - we see an image of a 1-sheet document is it front or back?
  - Balanced dataset: 500 recto 500 verso
- Forgery detection:
  - We see an image of a document
  - Is it forged or not?
  - Unbalanced dataset: 36 forgeries in 1000 documents

Binary Classification: Accuracy

- Trivial Baselines
- Random predictor
  - Performance when predicting at random
  - Recto-verso performance: 50%
  - Forgery detection performance: 50%
- Best blind predictor
  - Performance of the best possible prediction that ignores input
  - Recto-verso performance: 50%
  - Forgery detection performance: 96.4%

Binary Classification: Accuracy

- Is 90% accuracy good or bad?
- Recto-verso
  - Can we live with a system that is wrong 90% of the time?
  - We are 5 times better than random and best blind classifiers (10% error vs 50% error)
- Forgery detection
  - We are 5 times better than random classifier
  - We are ~2.5 times worse than best blind predictor

Binary Classification: Accuracy

- Is 90% accuracy good or bad?
- YES

Binary Classification: Recall

- Not all mistakes are equal

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	True Positive	False Negative
	Authentic	False Positive	True Negative

UNI GRAZ

Binary Classification: Recall

System A:

- Accuracy 94.0%

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	25	11
	Authentic	49	915

System B:

- Accuracy: 95.2%

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	22	14
	Authentic	34	930

Binary Classification: Recall

- True positives
- Relevant (Forgeries)
- Recall:

$$\frac{\boxed{TP}}{\boxed{Relevant}}$$

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	True Positive	False Negative
	Authentic	False Positive	True Negative

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	Relevant	
	Authentic	Irrelevant	

Binary Classification: Recall

System A:

- Accuracy 94.0%
- Recall: **69.4%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	25	11
	Authentic	49	915

System B:

- Accuracy: 95.2%
- Recall: **61.1%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	22	14
	Authentic	34	930

Binary Classification: Recall

- AKA Detection rate
- When we must not miss things
- When FN are worse than FP
- When outcome will be verified by humans

Binary Classification: Recall

- When we must not miss things
- When FN are worse than FP
- When outcome will be verified by humans

Binary Classification: Precision

- What if we go to criminal court
- How does the presumption of innocence play into it?

UNI GRAZ

Binary Classification: Precision

- True positives
- Relevant (Forgeries)
- Recision:

$$\frac{\text{TP}}{\text{Detected}}$$

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	True Positive	False Negative
	Authentic	False Positive	True Negative

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	Detected	Undetected
	Authentic		

Binary Classification: Precision

System A:

- Accuracy 94.0%
- Recall: 69.4%
- Precision: **33.8%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	25	11
	Authentic	49	915

System B:

- Accuracy: 95.2%
- Recall: 61.1%
- Precision: **39.3%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	22	14
	Authentic	34	930

UNI GRAZ

Binary Classification: Precision

- When FP are worse than FN
- When we want to filter high confidence outputs
- When we don't want to be wrong on any detection

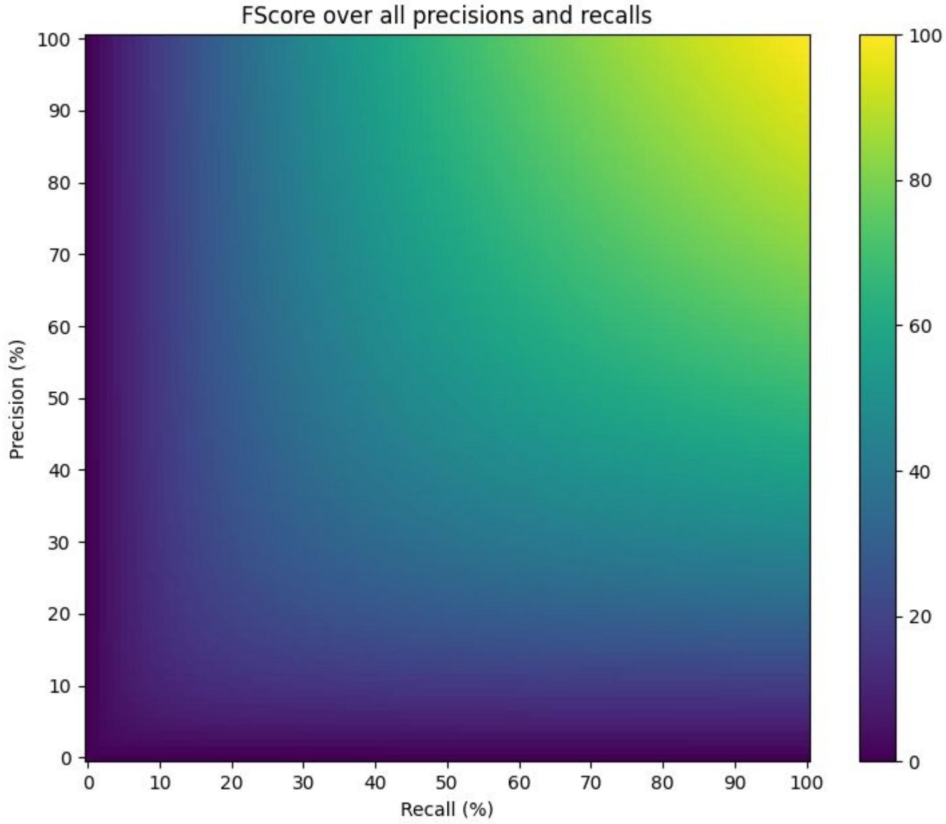
Binary Classification: FScore

- So which system is better overall
- If we can only choose one system, which should it be?
- Both for paleographers who search for forgeries
- And for advising criminal court

Binary Classification: FScore

$$\text{FScore} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- AKA:Harmonic mean of precision and recall
- AKA: F-Measure



UNI GRAZ

Binary Classification: FScore

System A:

- Accuracy 94.0%
- Recall: 69.4%
- Precision: 33.8%
- FScore: **45.5%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	25	11
	Authentic	49	915

System B:

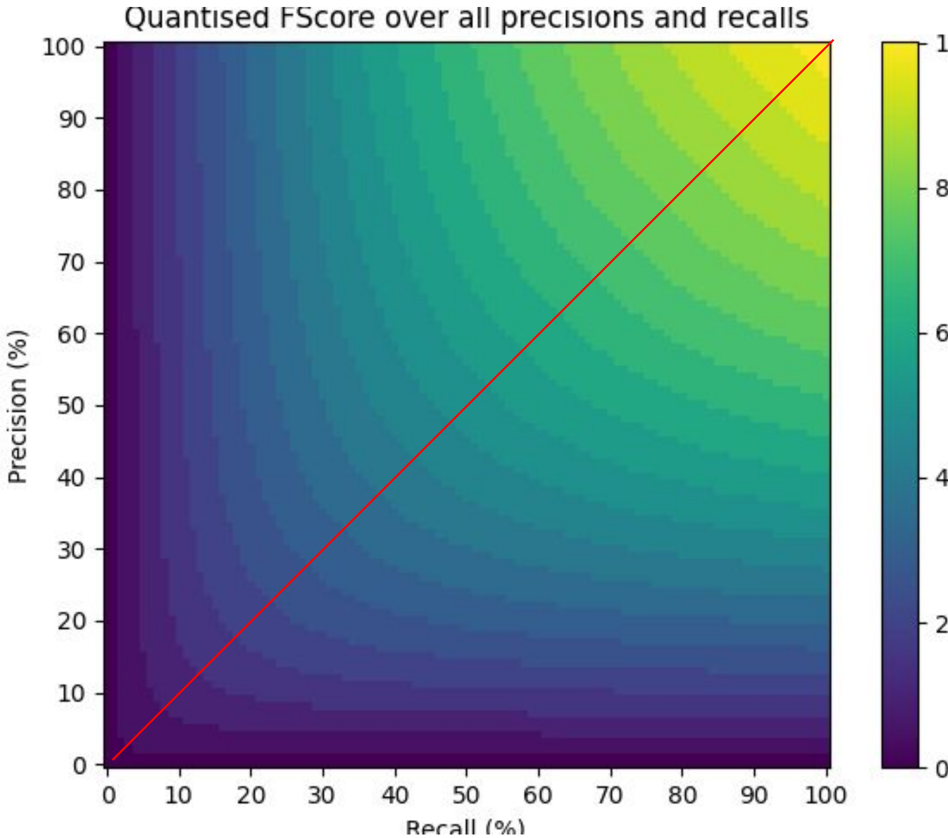
- Accuracy: 95.2%
- Recall: 61.1%
- Precision: 39.3%
- FScore: **47.8%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	22	14
	Authentic	34	930

Binary Classification: FScore

$$\text{FScore} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- AKA:Harmonic mean of precision and recall
- AKA: F-Measure
- You can't be bad at either
- Punishes imbalance
- The better you perform the less imbalance is punished
- Fscore <= Accuracy



Binary Classification: Naive Baselines

Random Classifier:

- Accuracy 50.0%
- Recall: 50.0%
- Precision: 3.6%
- FScore: **6.7%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	18	18
	Authentic	482	482

Best Blind Predictor:

- Accuracy: 96.4%
- Recall: 0.0%
- Precision: 0/0%
- FScore: **0.0%**

		Predictions	
		Forgery	Authentic
Groundtruth	Forgery	0	36
	Authentic	0	964

UNI GRAZ

Multiclass Classification

- One vs Rest:
  - Can turn into a set of binary classifiers
  - How do we weight the metrics?
  - Metrics are not linear
  - FScore not linear
- Pure Multiclass
  - Confusion matrix
  - Qualitative

Multiclass Classification

- Trivial multiclass dataset:
  - Script detection
  - 130 Documents in English
  - 122 Documents in German
  - 89 Document in Italian
  - 117 Document in Russian
  - 458 Documents in total

Ground Truth	E	130
	G	122
	I	89
	R	117

Multiclass Classification

- System output

	P(E)	P(G)	P(I)	P(R)	GT
Sample 1	0.9	0.05	0.03	0.02	E
Sample 2	0.4	0.15	0.15	0.15	I
...	...	...	...	...	...
Sample 458	0.41	0.1	0.04	0.45	R

Multiclass Classification

- System output
- Winner takes all

	P(E)	P(G)	P(I)	P(R)	GT	Ac.
Sample 1	1.0	0.0	0.0	0.0	E	1.0
Sample 2	1.0	0.0	0.0	0.0	I	0.0
...	...	...	...	...	...	
Sample 458	0.0	0.0	0.0	1.0	R	1.0

Multiclass Classification: Confusion Matrix

- Rows sum up to detected
- Columns sum to relevant
- Diagonal / sum is accuracy

Ground Truth

Predictions

	E	G	I	R	
E	111	6	9	2	128
G	7	112	5	2	126
I	10	3	74	3	90
R	2	1	1	110	114
	130	122	89	117	458

Multiclass Classification: Confusion Matrix

- Recall:
  - Column wise (per class)

Ground Truth

Predictions

	E	G	I	R	
E	111	6	9	2	128
G	7	112	5	2	126
I	10	3	74	3	90
R	2	1	1	110	114
	130	122	89	117	458

Multiclass Classification: Confusion Matrix

- Precision:
  - Row wise (per class)

Ground Truth

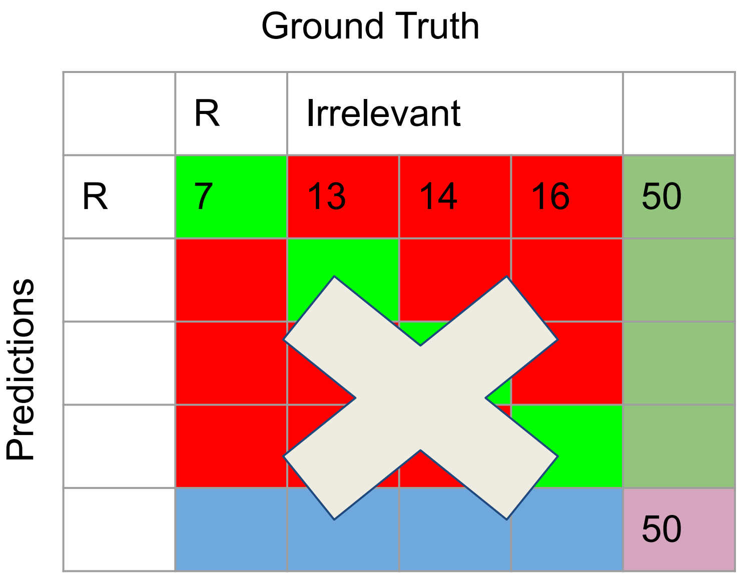
Predictions		E	G	I	R	
	E	111	6	9	2	128
	G	7	112	5	2	126
	I	10	3	74	3	90
	R	2	1	1	110	114
		130	122	89	117	458

Retrieval

- What if we cast our classification problem as a ranking one
- Nearest neighbor classifier

Retrieval

- Needles in haystack
- Lets focus on a single class **R**(elevant)
- The class is quite rare
- Our model ranks the database



Retrieval

- We look at the top ***n*** samples of the sorted DB
  - And measure Recall on it
  - And measure Precision on it
- What if we see what happens for all possible ***n***
  - ***Recall(n)***
  - ***Precision(n)***

UNI GRAZ

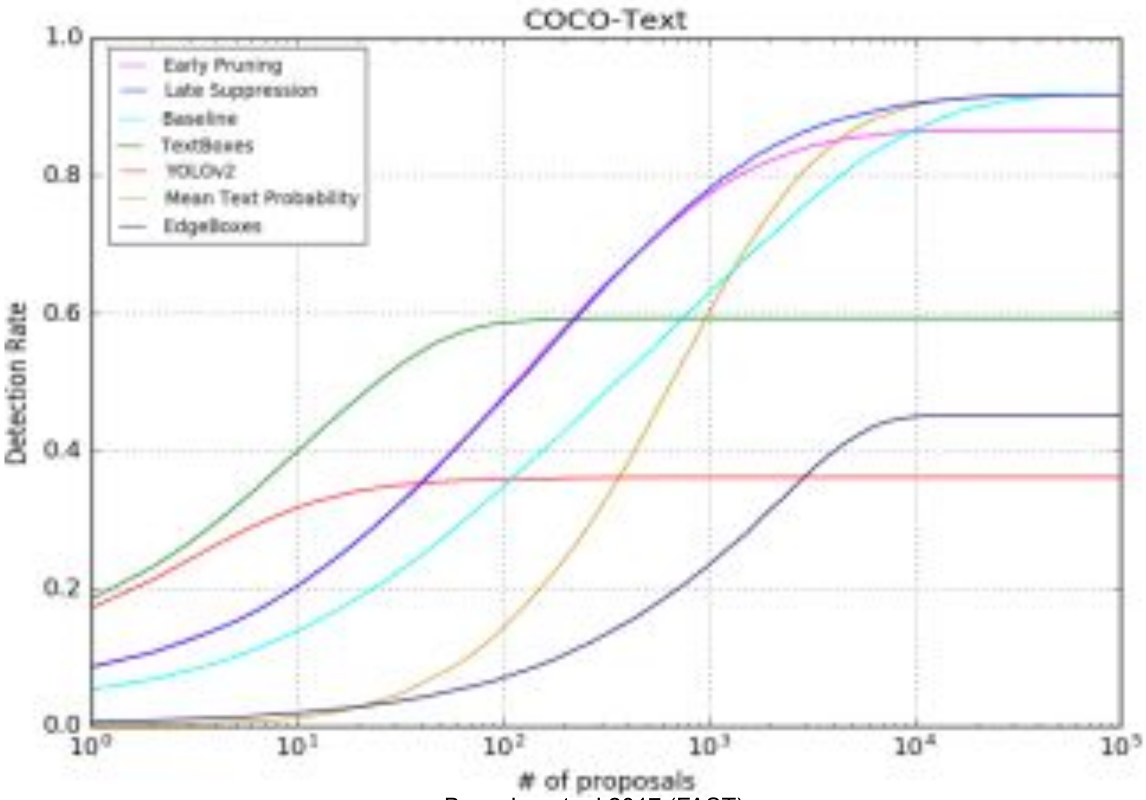
Retrieval

- How are we doing when we select the top word hypotheses?

Bazazian et. al 2017 (FAST)

Retrieval

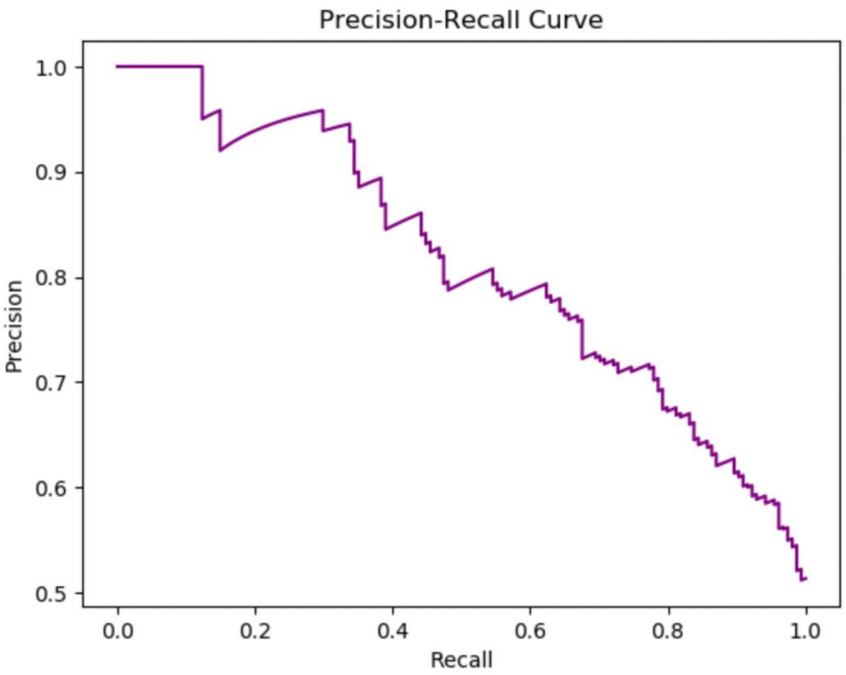
- Recall @
- How many hypotheses should we entertain before



Bazazian et. al 2017 (FAST)

Retrieval

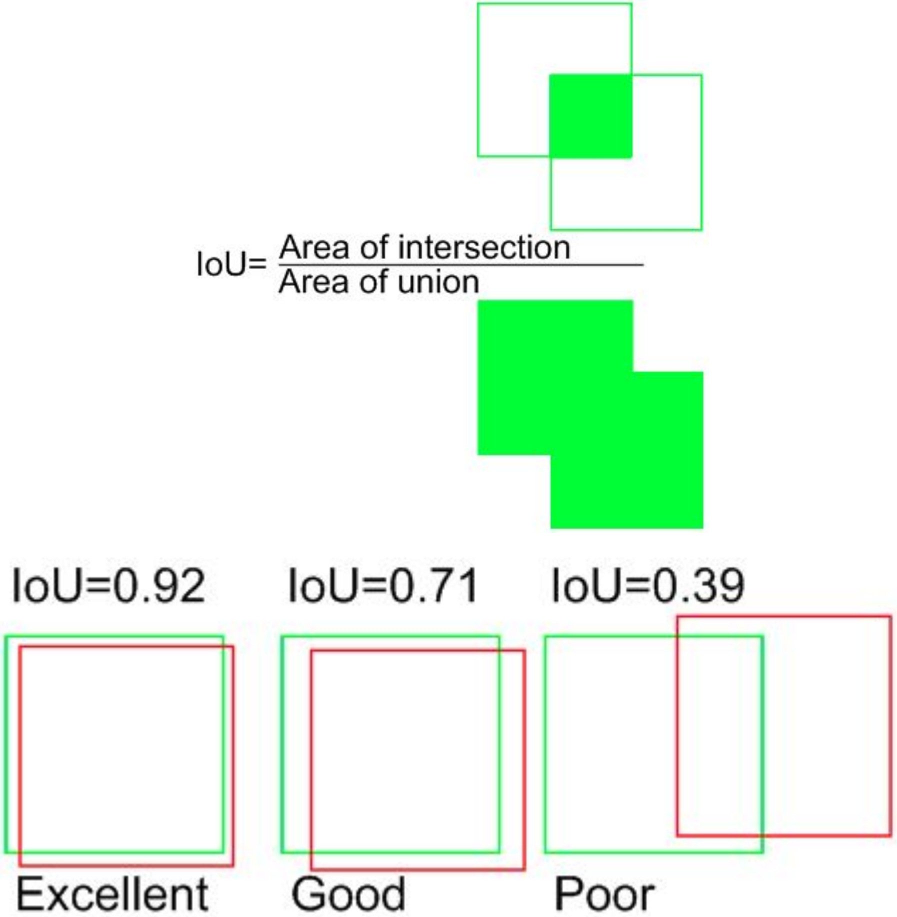
- Precision-Recall curve
- How many mistakes can we tolerate?
- How many things must be there?
- The average of the values sampled at every upward edge is called mAP



https://www.statology.org/precision-recall-curve-python/

Retrieval

- When we have localised predictions
- Are we predicting the same thing that is there?
- Or maybe something else?



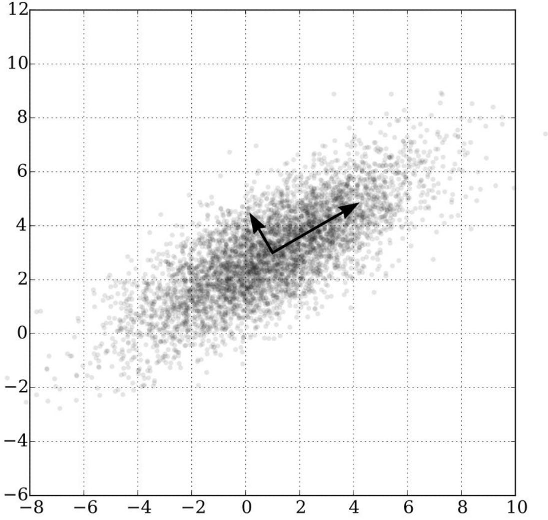
http://ronny.rest/tutorials/module/localization_001/iou/

The black arts:
Non quantifiable visualisations

- It is very hard control our perceptual biases
- Point confusions are interpreted by our eye like color blending

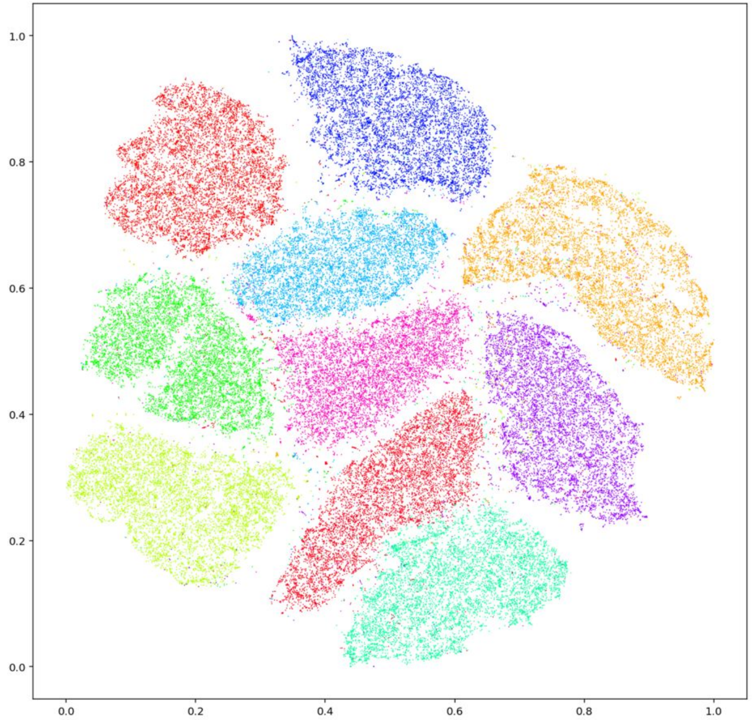
The black arts: PCA

- Invented in 1901 by Karl Pearson
- Rotate your representation in order to maximise the explained variance
- Output dimensions are sorted by importance of information
- Deterministic on all but the signs
- First 2D can put data on the plane



The black arts: T-SNE

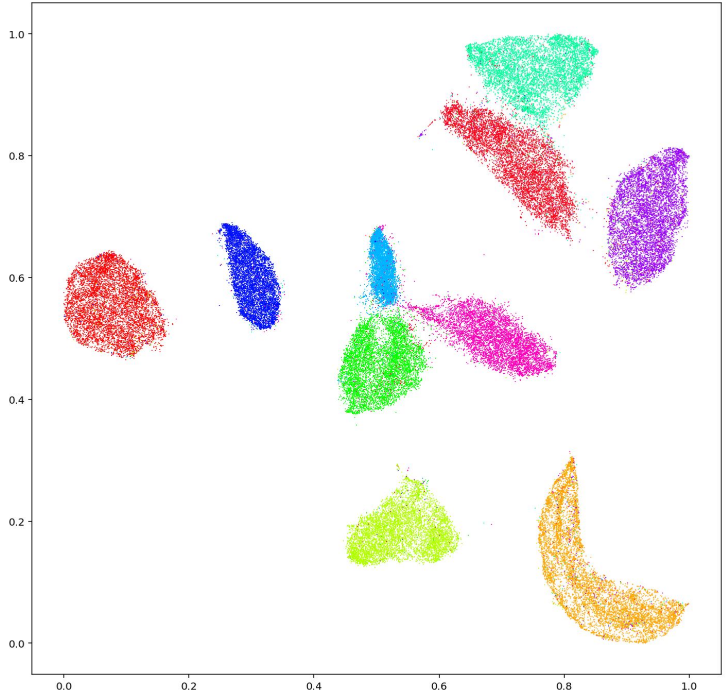
- Manipulate a high dimensional representation so that locality is preserved
- How far things are doesn't matter
- Close things does



https://www.flickr.com/photos/kylemcdonald/albums/72157662596196708

The black arts: UMap

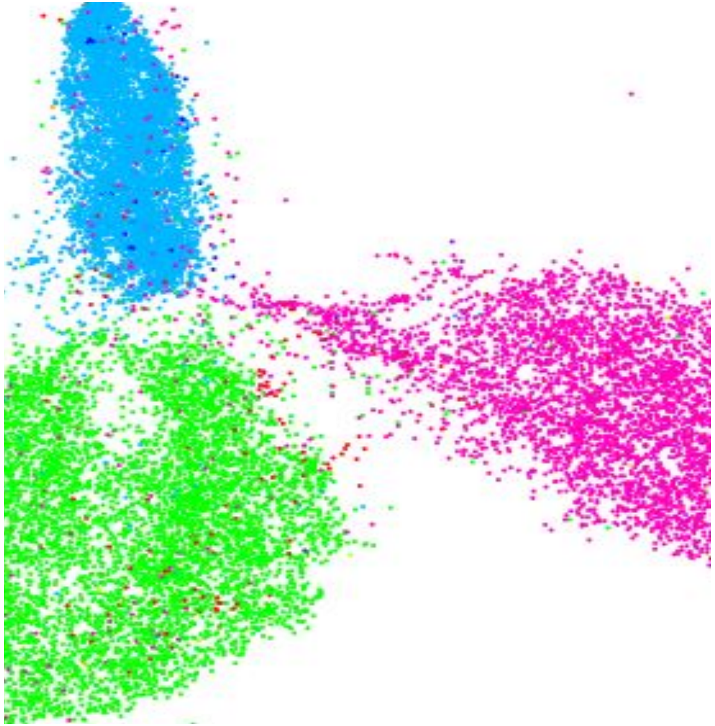
- Same as t-sne but preserving global structure if possible
- Faster



https://www.flickr.com/photos/kylemcdonald/albums/72157662596196708

The black arts: UMap

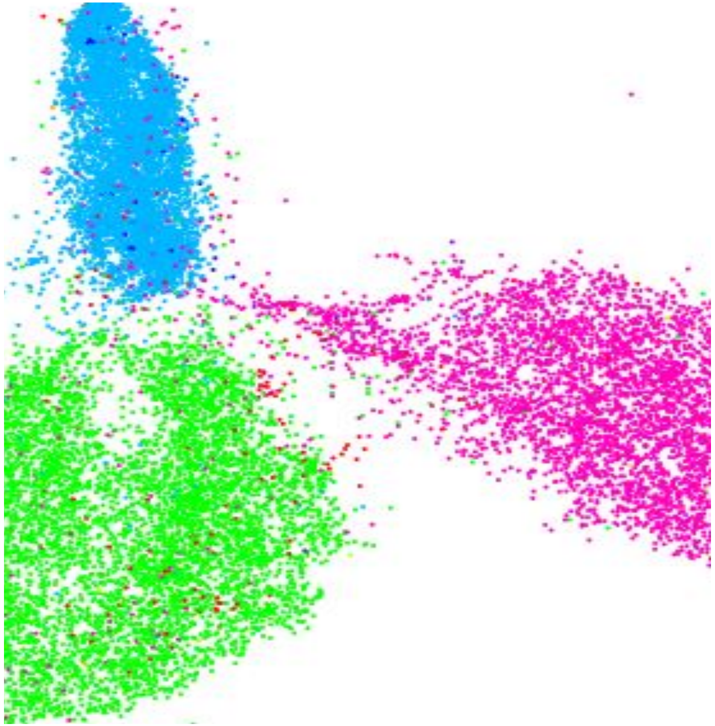
- Same as t-sne but preserving global structure if possible
- Faster



https://www.flickr.com/photos/kylemcdonald/albums/72157662596196708

The black arts: UMap

- Same as t-sne but preserving global structure if possible
- Faster

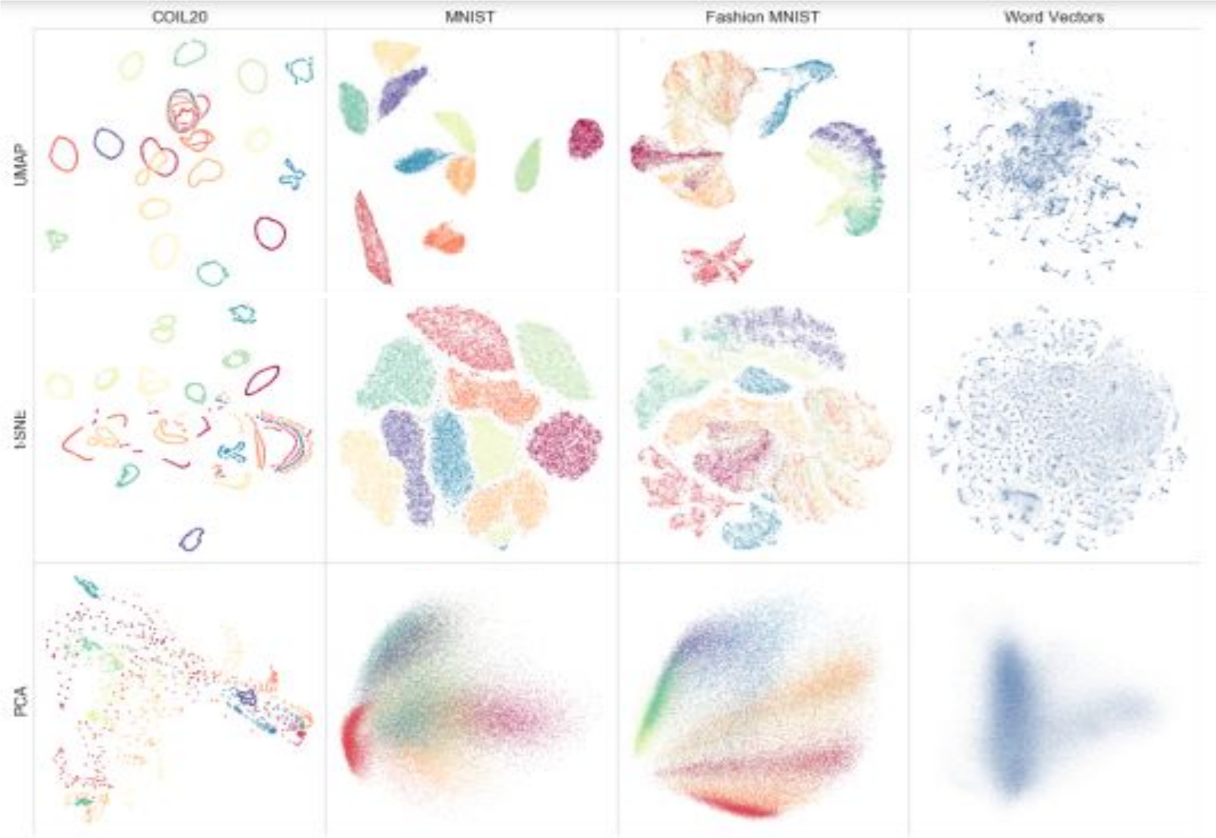


https://www.flickr.com/photos/kylemcdonald/albums/72157662596196708

The black arts: Comparisson

There is order in raw data

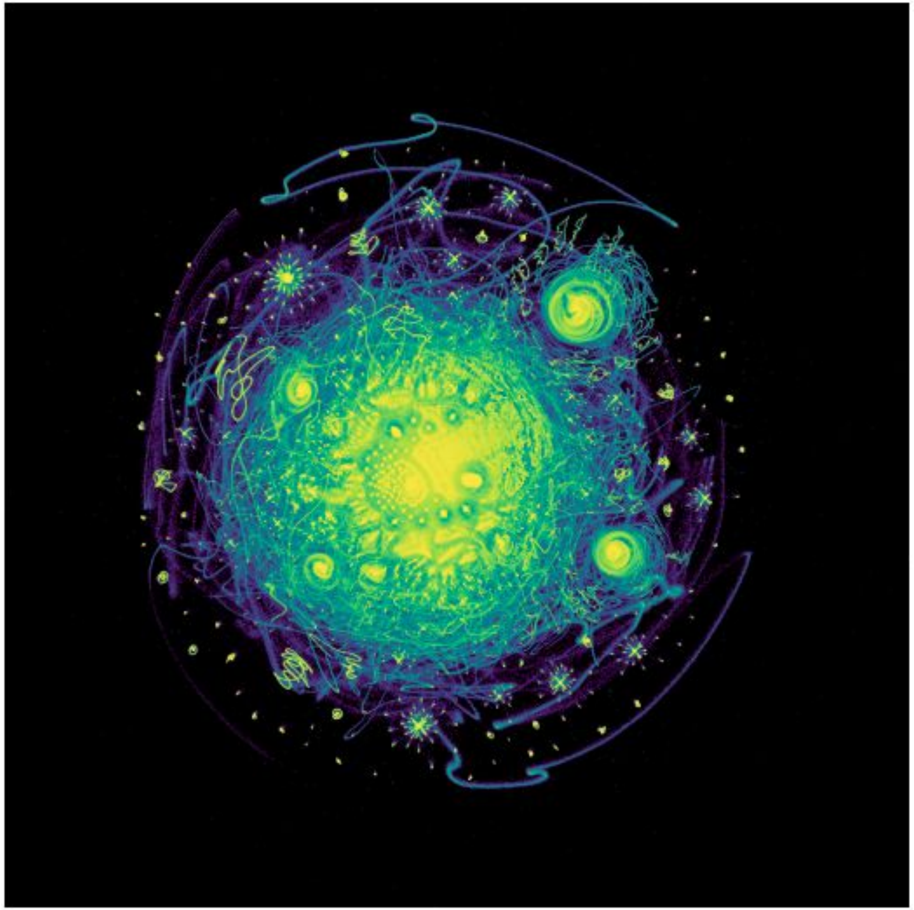
But what does it mean?



J. Healy et al. 2020 (UMAP)

The black arts:

- What are we looking at after all?
- 30,000,000 integers as represented by binary vectors of prime divisibility



J. Healy et al. 2020 (UMAP)

UNI
GRAZ