Deep Learning for Medical Imaging

Olivier Colliot, PhD
Research Director at CNRS
ARAMIS Lab – www.aramislab.fr
PRAIRIE – Paris Artificial Intelligence Research Institute

Maria Vakalopoulou, PhD
Assistant Professor at CentraleSupelec
Center for Visual Computing

Part 3 - Validation
3.1 Introduction

A random paper in deep learning for medical imaging

(ok, a very bad paper, not really a random one)

The ZorglubFormer network for automatic classification of Alzheimer's disease from MRI

Jane Doe[1] and John Due[1]

[1]University of Random studies, Earth

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

Should we be satisfied with this?

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

What does it mean “images” with Alzheimer’s? Are these different patients?

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

Same comment for the normal images

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

How many in training set? How many in testing set?

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

What are the characteristics of the acquisition? What type of MRI sequence? Which scanner?

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

From which dataset?

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

Characteristics of the patients? Age? Sex? Disease severity?

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

Is this a relevant metric?

“We demonstrate that our new method outperformed the state of the art”

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

“We demonstrate that our new method outperformed the state of the art”

That’s a strong claim. Is it supported by evidence?

From the “experiments and results” section

Our dataset included 280 MR images with Alzheimer’s disease and 52 normal MR images.

	Accuracy
3D ResNet	87.1
TrickOfTheTradeCNN [11]	88.3
ZorglubFormer (proposed)	**89.7**

Is this due to chance?

“We demonstrate that our new method outperformed the state of the art”

That’s a strong claim. Is it supported by evidence?

Introduction

- Validation aims at **evaluating the performance of an ML model**
- **Ideally**, it should be representative of **how the model would perform in real life**
  - Difficult to achieve in practice, at least at the stage of research
- **At the very least**, it should provide an **unbiased estimate of how the model would perform on new data** that is similar to that used for training (but not the same data of course!!)
- Provide information about the **variability of the performance** and the **precision of its estimation**

Introduction

- We want a model that performs well on **new, never-before seen, data**.

- That is equivalent to saying we want our model **to generalise well**.
  - We want it to recognise only those characteristics of the data that are general enough to also apply to some unseen data
  - … while ignoring the characteristics of the training data that are overly specific to the training data

- Because of this, **we never test on training data, but use separate test data**

Introduction

- In this part, we address
  - **How to quantify the performance of the model?**
    - Performance metrics
  - **How to estimate the performance metrics?**
    - Validation strategies
  - **What kind of statistical analysis should be performed?**
  - **How to make your research reproducible?**
  - **What should you report in a paper?**

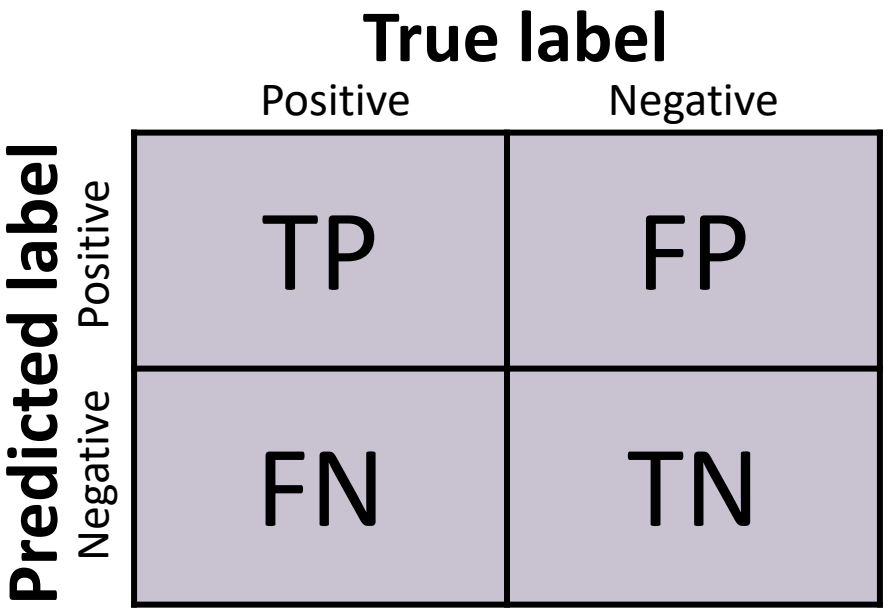
Part 3 - Validation
3.2 Performance metrics

Part 3 - Validation
3.2 Performance metrics for classification and regression

Part 3 - Validation
3.2.1 Metrics for classification

Metrics for classification

Confusion matrix



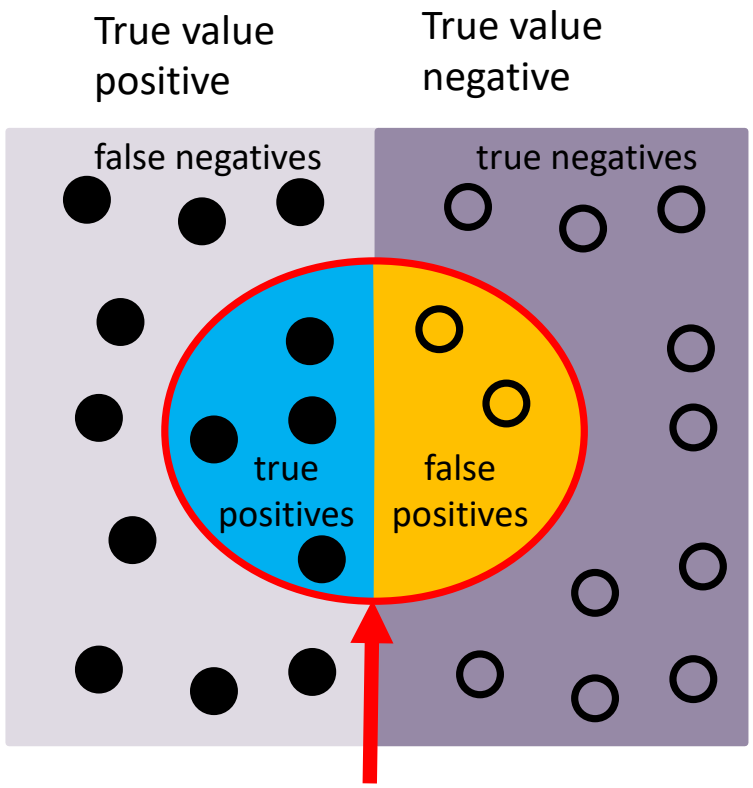
Metrics for classification

True Positives (TP): cases when the actual class of the data point was 1 and the predicted is also 1
Ex. The patient has cancer (1) and the model classifies his case as cancer(1)

True Negatives (TN): cases when the actual class of the data point was 0 and the predicted is also 0
Ex. The patient does not have cancer (0) and the model classifies his case as non-cancer (0)

False Negatives (FN): cases when the actual class of the data point was 1 and the predicted is 0
Ex. The patient has cancer (1) and the model classifies his case as non-cancer(0)

False Positives (FP): cases when the actual class of the data point was 0 and the predicted is also 0
Ex. The patient does not have cancer (0) and the model classifies his case as cancer (1)



Metrics for classification

False positives vs false negatives

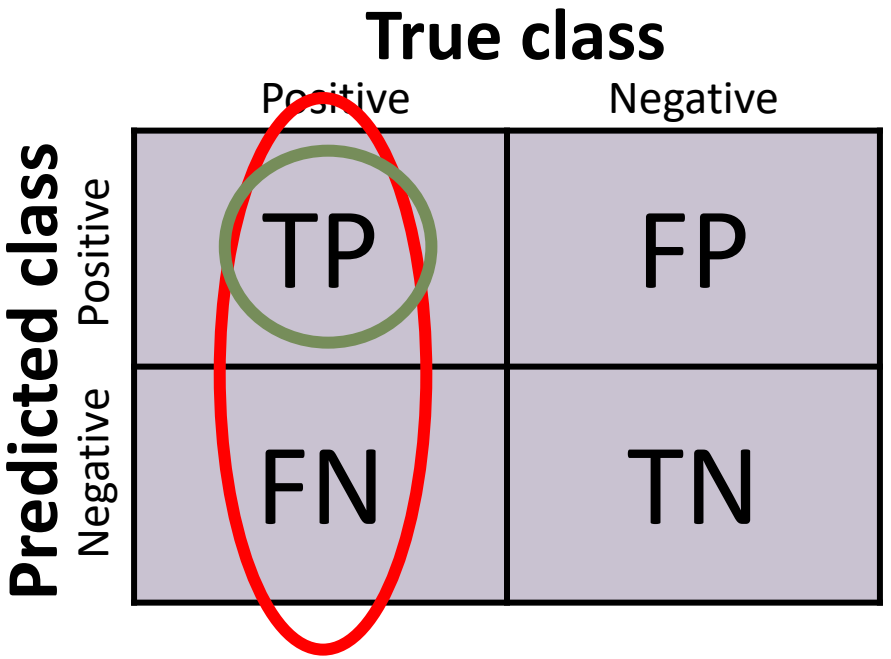
- Example 1: cancer screening
  - We should not miss any cancer cases
  - One may consider requiring have **very few false negatives even at the expense of relatively high proportion of false positives**
  - Positive cases would then be reviewed by an expert or lead to additional explorations

- Example 2: spam detection
  - We should avoid flagging legitimate emails as spam
  - One may consider requiring have **very few false postives even at the expense of relatively high proportion of false negatives**

- Example 3: segmentation
  - In many cases, one can think that false positives and negatives are equally problematic

Metrics for classification

Sensitivity (also called recall)

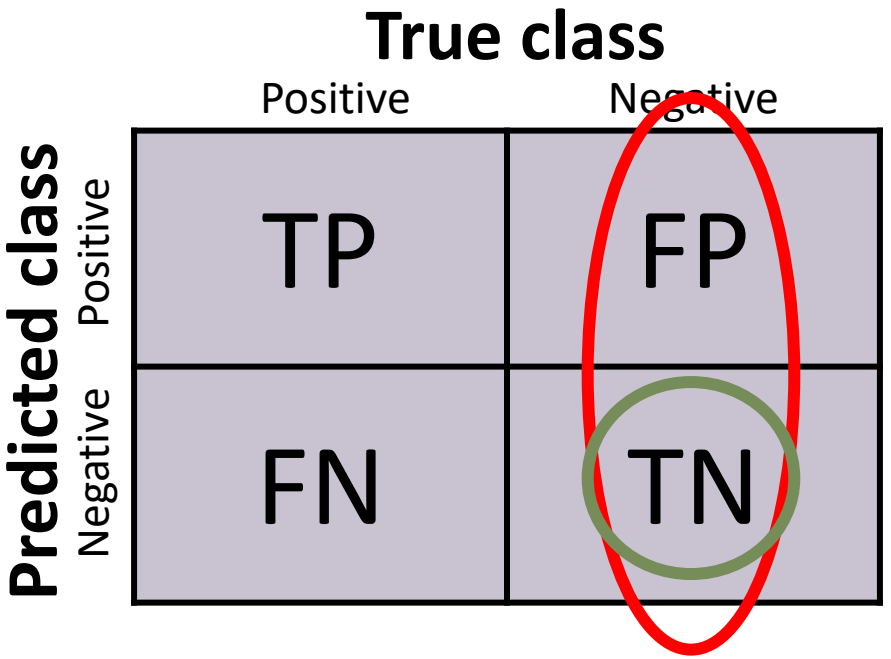


How much of the positives do we retrieve?

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

Metrics for classification

Specificity

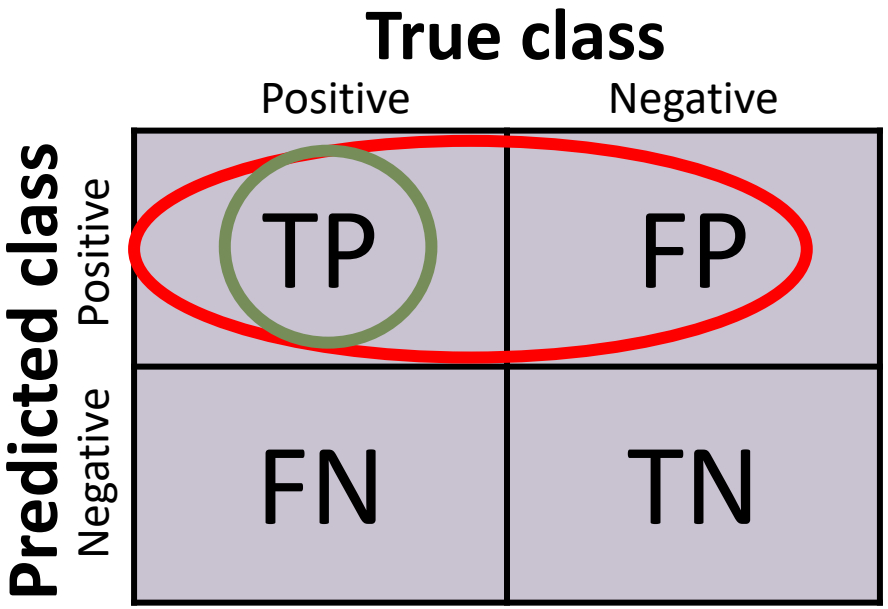


How much of the negatives do we retrieve?

$$Specificity = \frac{TN}{TN + FP}$$

Metrics for classification

Precision (also called positive predictive value - PPV)

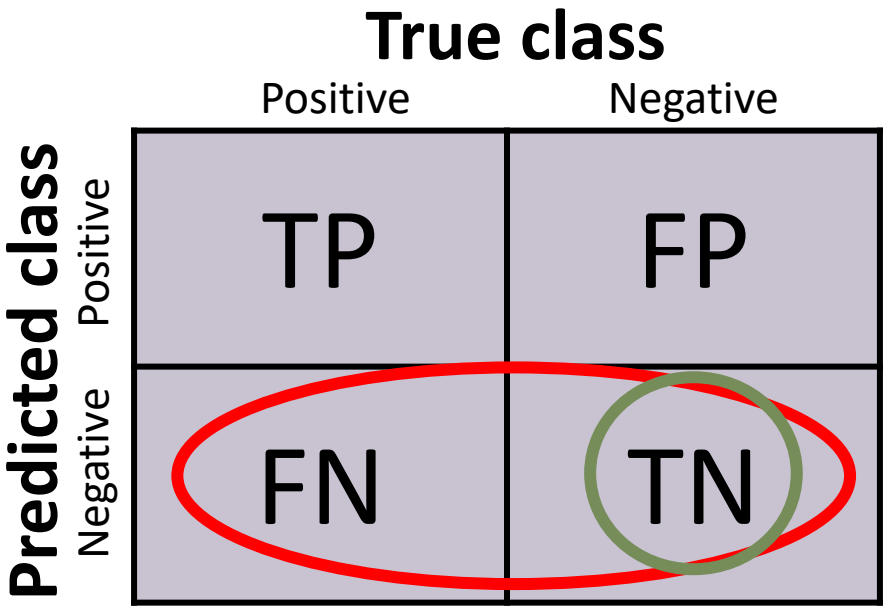


How much of those classified as positives are indeed positives?

$$PPV = Precision = \frac{TP}{TP + FP}$$

Metrics for classification

Negative predictive value - NPV



How much of those classified as negatives are indeed negatives?

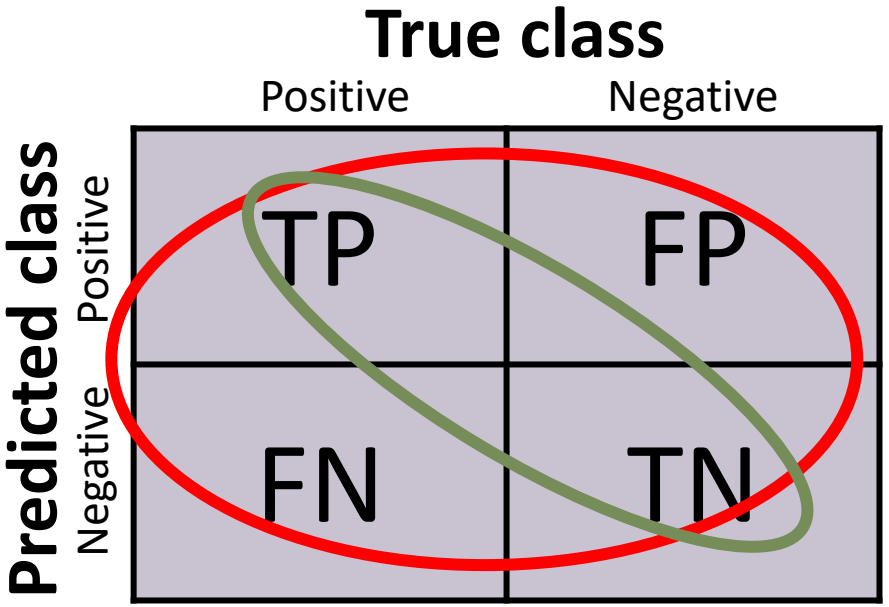
$$NPV = \frac{TN}{TN + FN}$$

Metrics for classification

The four previous metrics each describe only part of the confusion matrix

Often, one wants to have a summary in a single metric

Accuracy



Among all samples, how much are correctly classified?

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Metrics for classification

Problem with accuracy

- **Do not use when the data is imbalanced** (the number of cases in each class is not the same)
  - Example :
    - 990 non-cancer and 10 cancer
    - Trivial majority classifier: nobody has cancer
    - Accuracy: 99%

- Possible solution: **balanced accuracy (BA)**

$$BA = \frac{Sensitivity + Specificity}{2}$$

Metrics for classification

Problem with balanced accuracy

Suppose that you take a diagnostic test for a given disease

- The test turns out positive
- The **sensitivity of the test is 99%**, i.e. 99% of sick people are detected
- The **specificity is 90%**, i.e. 10% of healthy people are diagnosed as positive
- So **BA=95%** which is excellent
- What is the probability that you have the disease?

We don't have enough information.
Sensitivity=P(test positive|sick)
Specificity=P(test negative|healthy)
We are interested in P(sick|test positive)

Metrics for classification

We are interested in P(sick|test positive)

P(sick|test positive)=P(test positive|sick)*P(sick)/P(test positive)

P(sick|test positive)=**Sensitivity***P(sick)/P(test positive)

P(test positive)=P(test positive|healthy)*P(healthy)+ P(test positive|sick)*P(sick)

P(test positive)=(1-specificity)*(1-P(sick)) + sensitivity*P(sick)

Thus, we are missing P(sick) which is the **prevalence of the disease**.

Let the prevalence be 1/1000.

P(test positive)=0.10*0.999+0.99*0.001=0.0999+0.00099=10.089%

P(sick|test positive)=0.99*0.001/0.10089=0.01

So you have only 1% chance to be sick!

Metrics for classification

N= 10000 samples, prevalence= 0.001
Sensitivity: 0.99
Specificity: 0.99

PPV: 0.09
NPV: 0.9999
Accuracy: 0.99
Balanced accuracy: 0.99

is very bad: 91% of positively predicted samples are wrong

Looks good: 99% of samples are correctly classified

Balanced accuracy also looks good

Metrics for classification

Remember

- **For a diagnostic test, sensitivity and specificity are not enough**

- You need to also know the **prevalence**
  - Or the positive and negative predictive values

- Be careful at the prevalence in your sample. If you have a case-control study (for instance with equal numbers of cases and controls) the prevalence is likely wrong

- Ideally, you would need the prevalence in the situation in which the test is meant to be used (general population for a screening test)

Metrics for classification

NPV and PPV as a function of prevalence

- Sensitivity and specificity are fixed

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

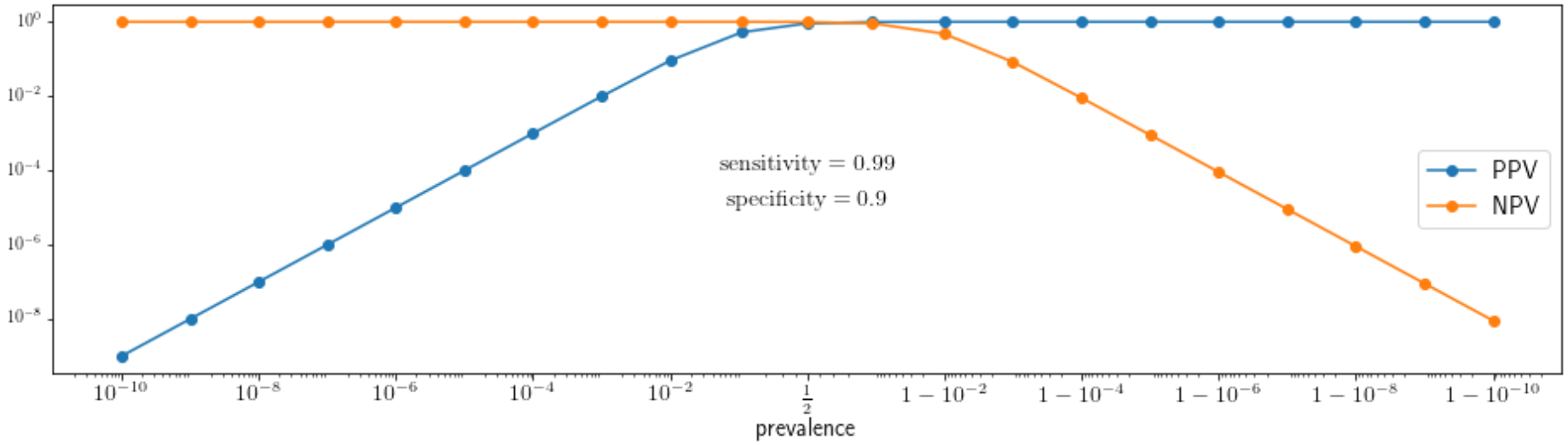
$$\text{PPV} \sim \frac{\text{sensitivity}}{1 - \text{specificity}} \times \text{prevalence} \quad \texttt{when} \quad \text{prevalence} \longrightarrow 0$$

$$\text{PPV} \longrightarrow 1 \quad \texttt{when} \quad \text{prevalence} \longrightarrow 1$$

$$\text{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{\text{specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity}) \times \text{prevalence}}$$

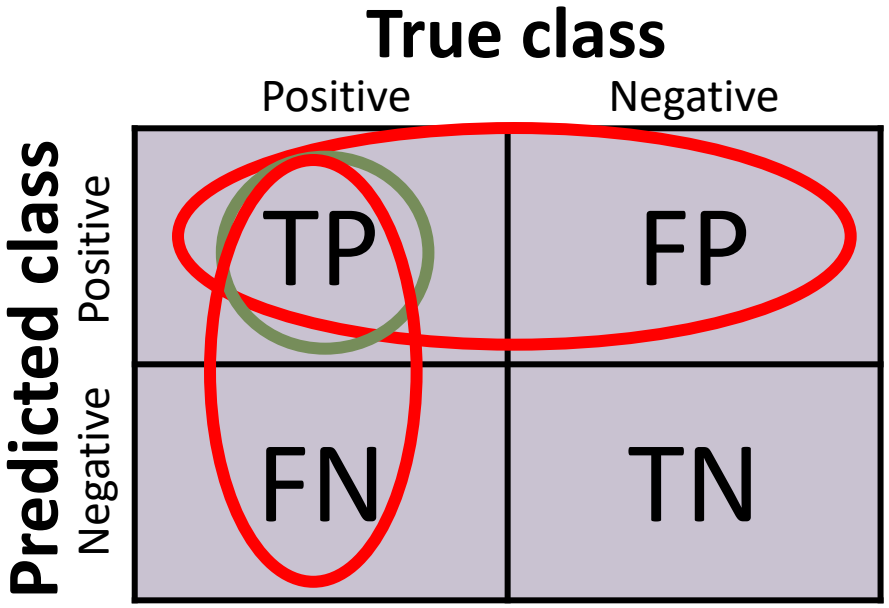
$$\text{NPV} \longrightarrow 1 \quad \texttt{when} \quad \text{prevalence} \longrightarrow 0$$

$$\text{NPV} \sim \frac{\text{specificity}}{1 - \text{sensitivity}} \times (1 - \text{prevalence}) \quad \texttt{when} \quad \text{prevalence} \longrightarrow 1$$



Metrics for classification

F1 score



Harmonic mean of precision and recall

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2\frac{Precision\ x\ Recall}{Precision + Recall}$$

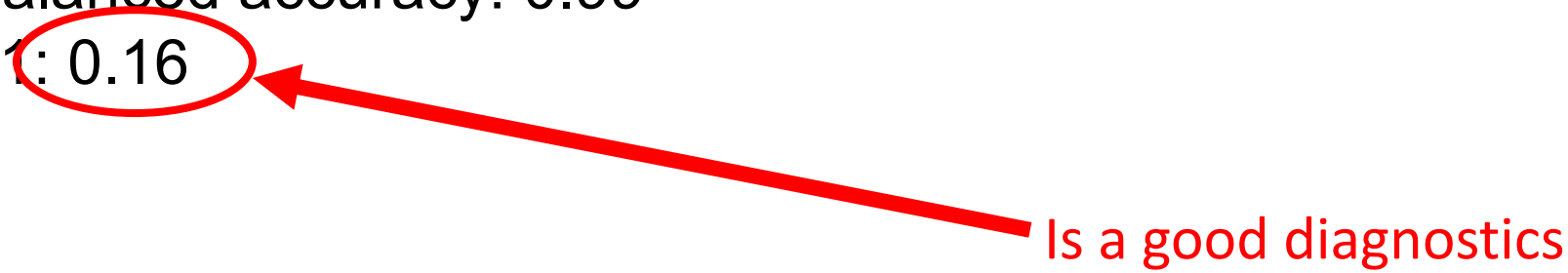
$$F1 = \frac{2TP}{2TP + FP + FN}$$

Metrics for classification

Coming back to the previous example

N= 10000 samples, prevalence= 0.001
Sensitivity: 0.99
Specificity: 0.99

PPV: 0.09
NPV: 0.9999
Accuracy: 0.99
Balanced accuracy: 0.99
F1: 0.16



Metrics for classification

N= 10000 samples, prevalence= 0.9999
Sensitivity: 0.99
Specificity: 0.99

PPV: 0.9999
NPV: 0.0098
Accuracy: 0.99
Balanced accuracy: 0.98
F1: 0.994

Is a poor diagnostics

Metrics for classification

N= 10000 samples, prevalence= 0.9999
Sensitivity: 0.99
Specificity: 0.99

PPV: 0.9999
NPV: 0.0098
Accuracy: 0.99
Balanced accuracy: 0.98
F1: 0.994

Is a poor diagnostics

Solution: switch classes
F1: 0.019

Is a good diagnostics

F1 should focus on the minority class to be informative

Metrics for classification

Matthews Correlation Coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Makes use of all the information in the confusion matrix

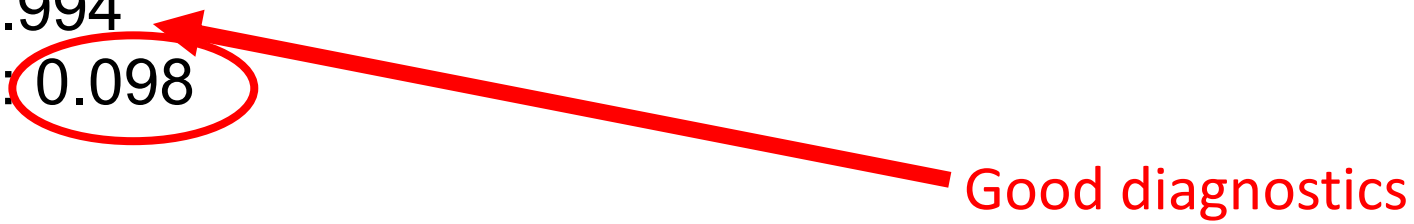
Ranges between +1 and -1
+1 is perfect prediction
0 is random prediction
-1 is perfectly wrong prediction

Metrics for classification

Coming back to the previous example

N= 10000 samples, prevalence= 0.9999
Sensitivity: 0.99
Specificity: 0.99

PPV: 0.9999
NPV: 0.0098
Accuracy: 0.99
Balanced accuracy: 0.98
F1: 0.994
MCC: 0.098



Metrics for classification

Conclusion

- Accuracy and BA **are useful because they are easy to interpret**. However, taken alone, **they are not sufficient and can be misleading**

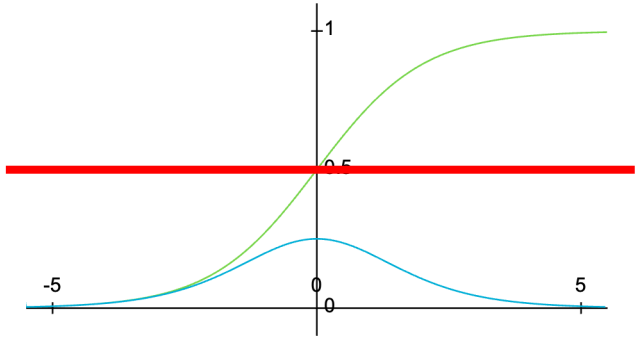
- Same thing for F1

- MCC is a good summary metric but probably less intuitive

Metrics for classification

Continuous outputs

- Many ML methods output continuous values
- This is in particular the case of neural networks
- Often one simply takes the class with highest probability
- However, there are applications where one is interested to study the performance for varying thresholds on the output

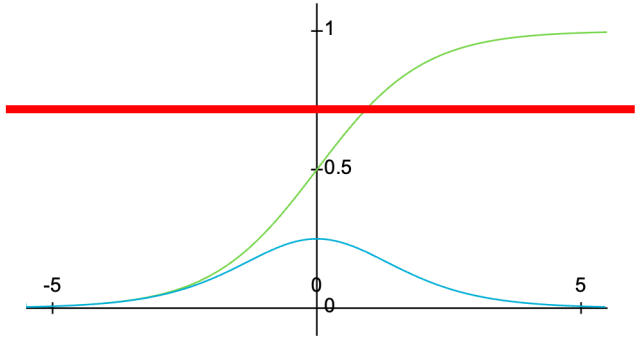


$$g(z) = \frac{1}{1 + e^{-z}}$$

Metrics for classification

Continuous outputs

- Many ML methods output continuous values
- This is in particular the case of neural networks
- Often one simply takes the class with highest probability
- However, there are applications where one is interested to study the performance for varying thresholds on the output

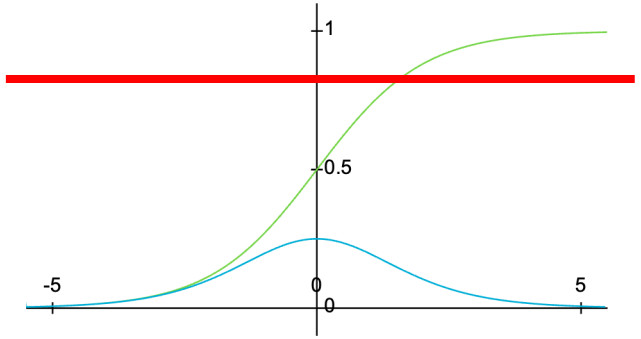


$$g(z) = \frac{1}{1 + e^{-z}}$$

Metrics for classification

Continuous outputs

- Many ML methods output continuous values
- This is in particular the case of neural networks
- Often one simply takes the class with highest probability
- However, there are applications where one is interested to study the performance for varying thresholds on the output

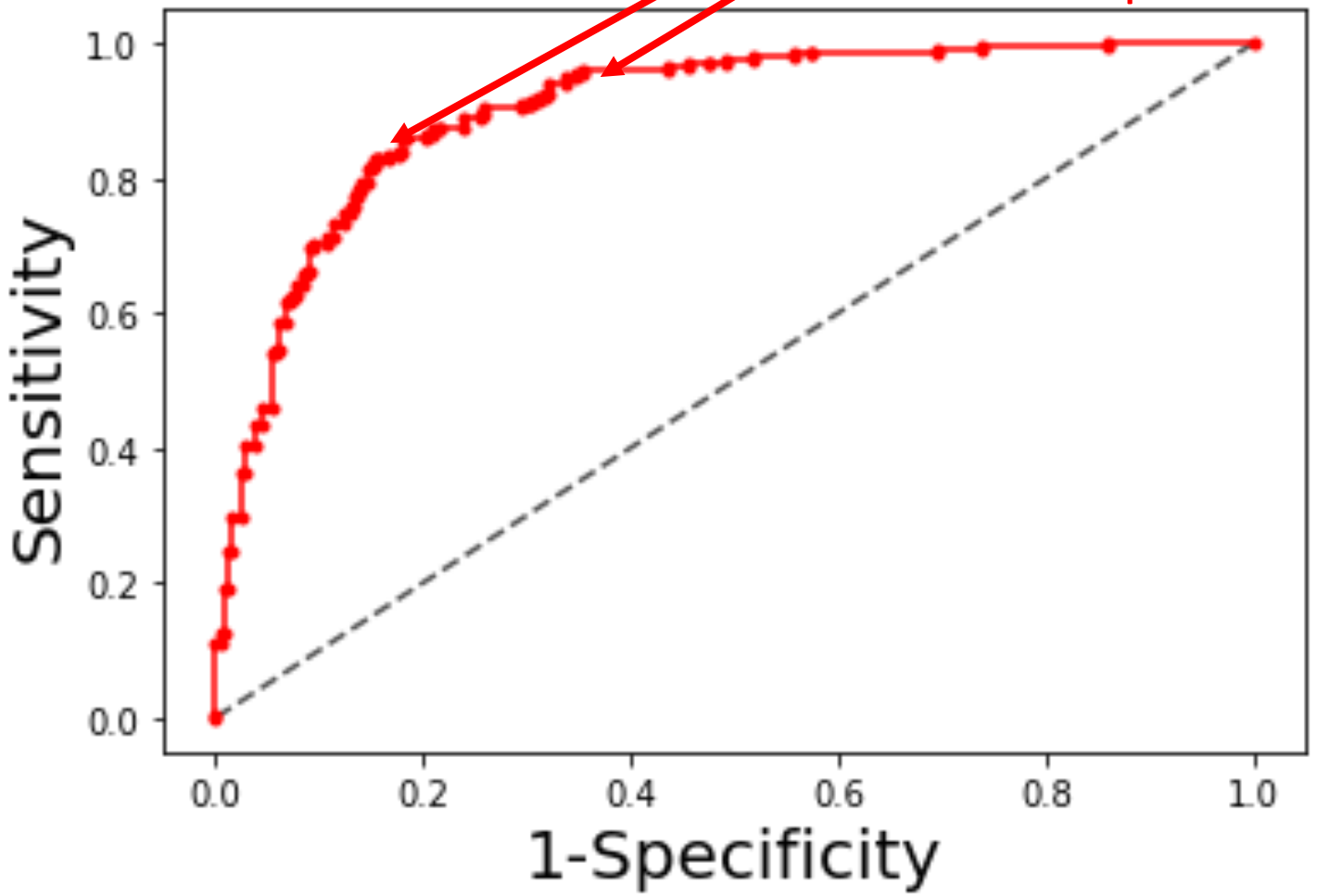


$$g(z) = \frac{1}{1 + e^{-z}}$$

Metrics for classification

ROC curve

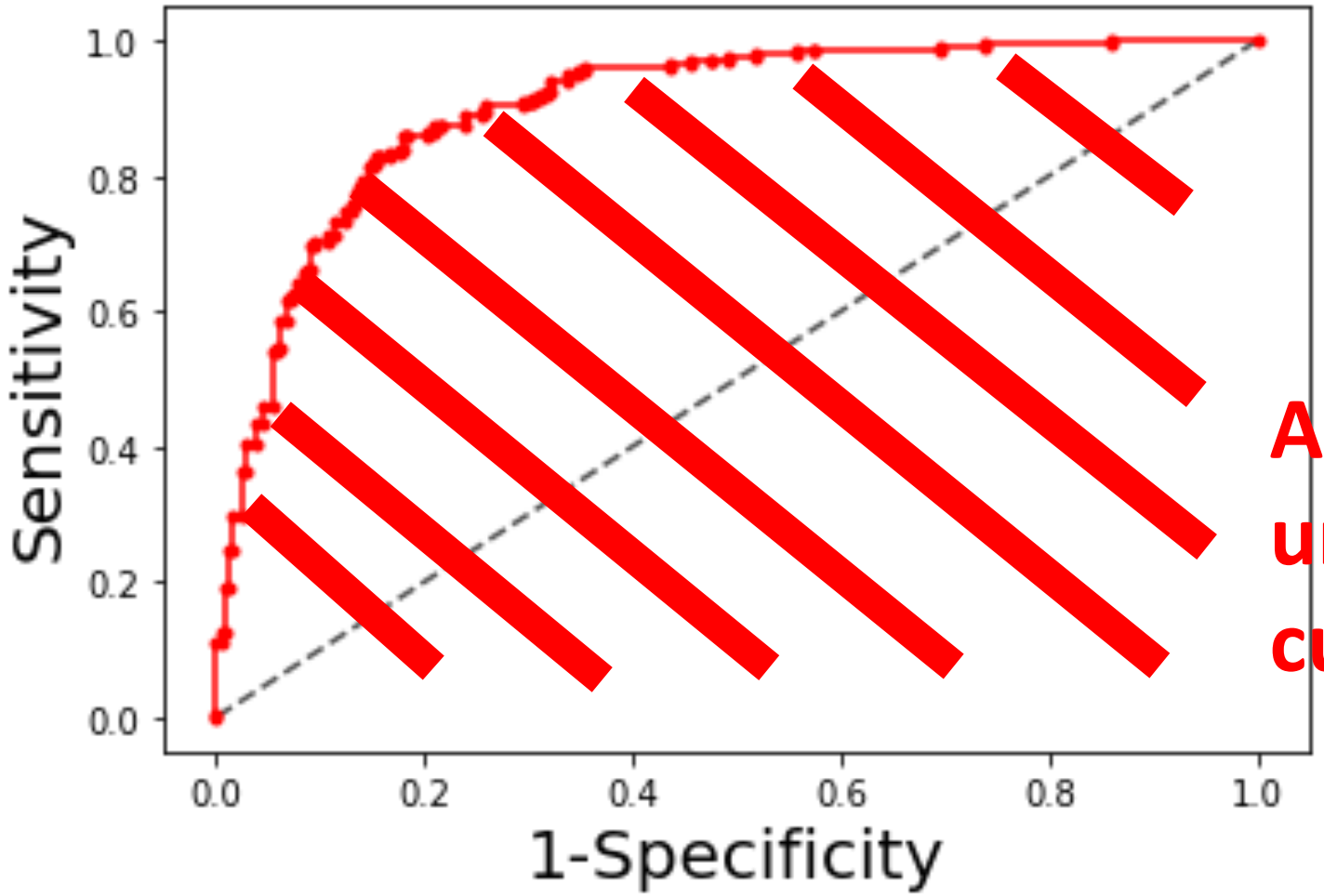
The curve corresponds to different thresholds on the output



Metrics for classification

ROC curve

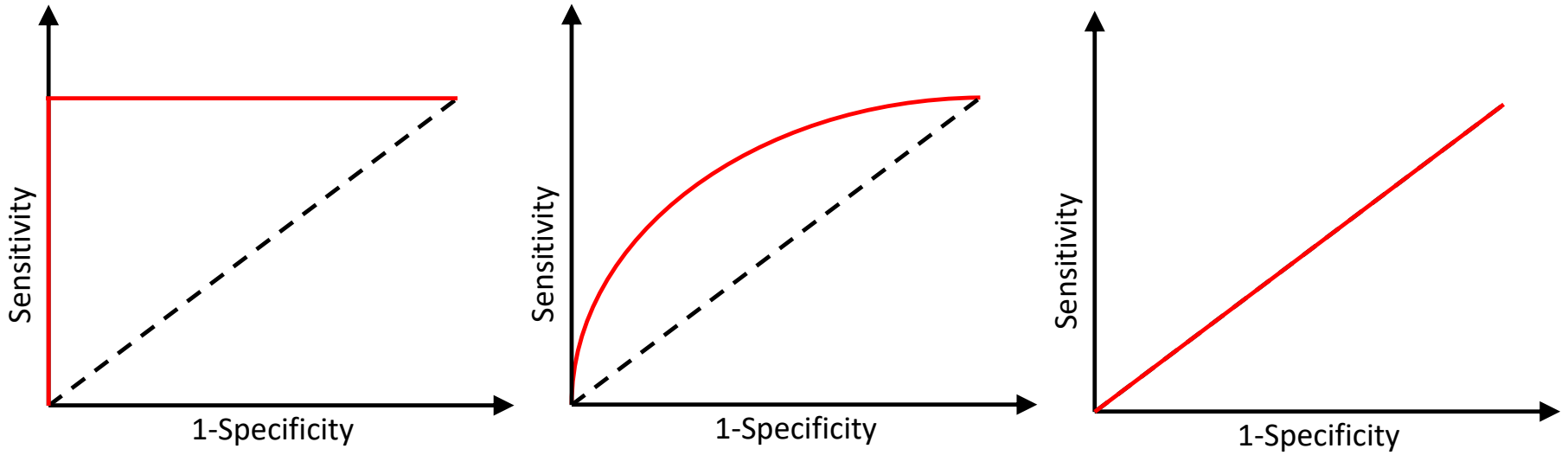
Interpretation of ROC AUC: probability that a positive sample has a higher classification score (as positive) than a negative sample



AUC (Area under curve)

Metrics for classification

ROC curve

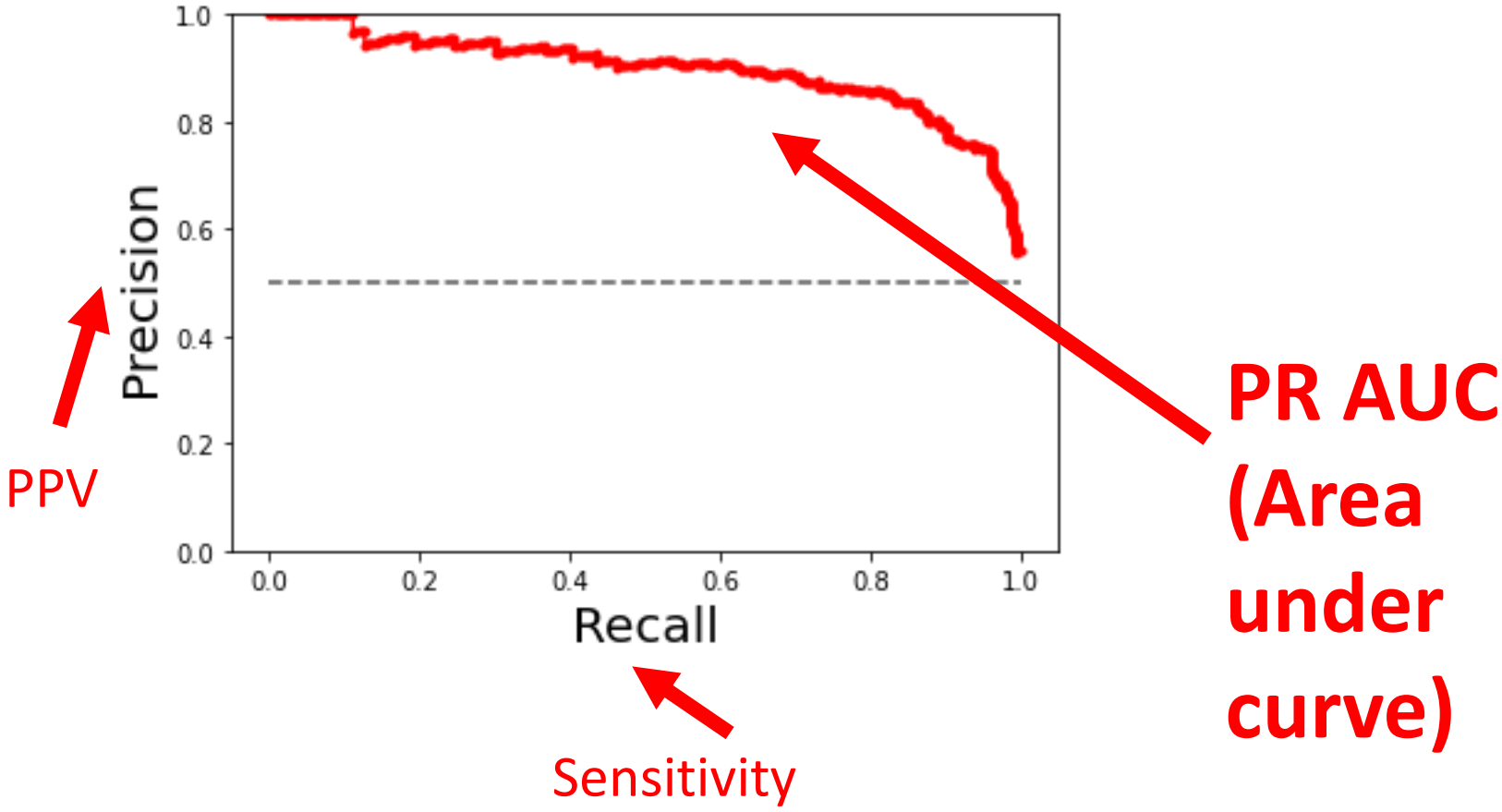


Perfect
AUC=1

Random
AUC=0.5

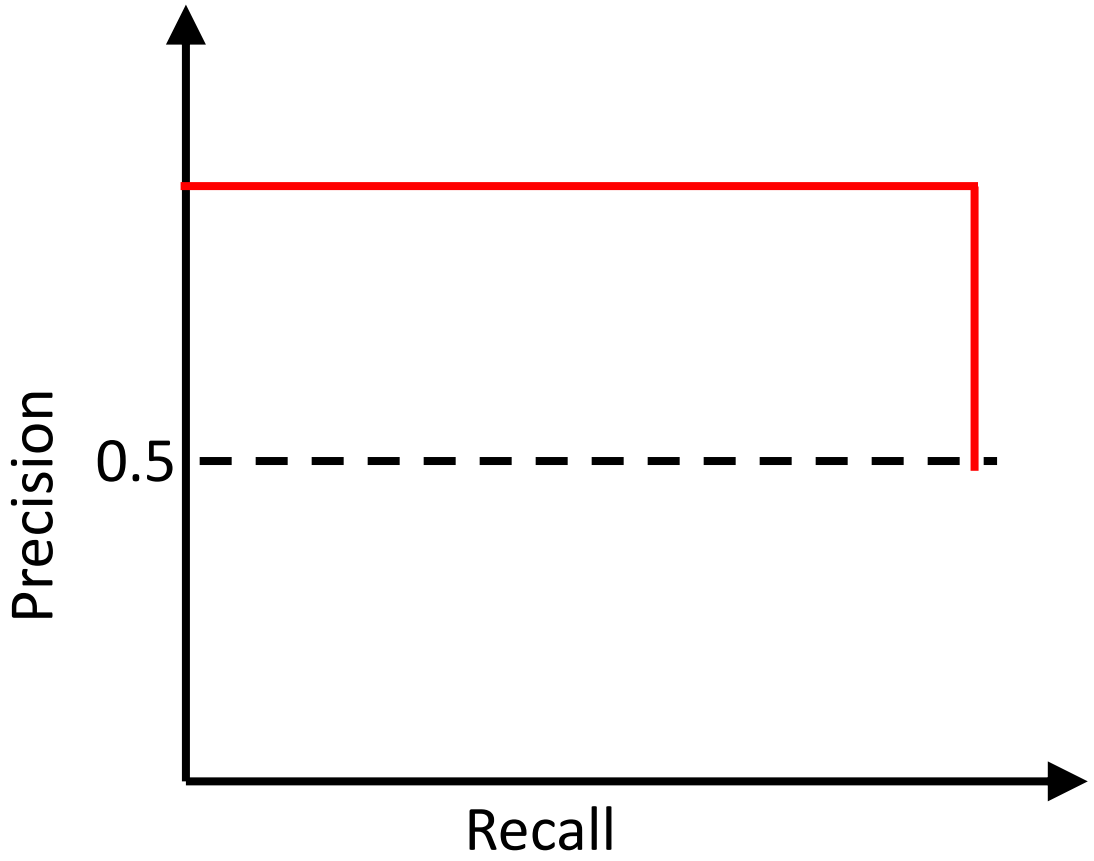
Metrics for classification

Precision-Recall (PR) curve



Metrics for classification

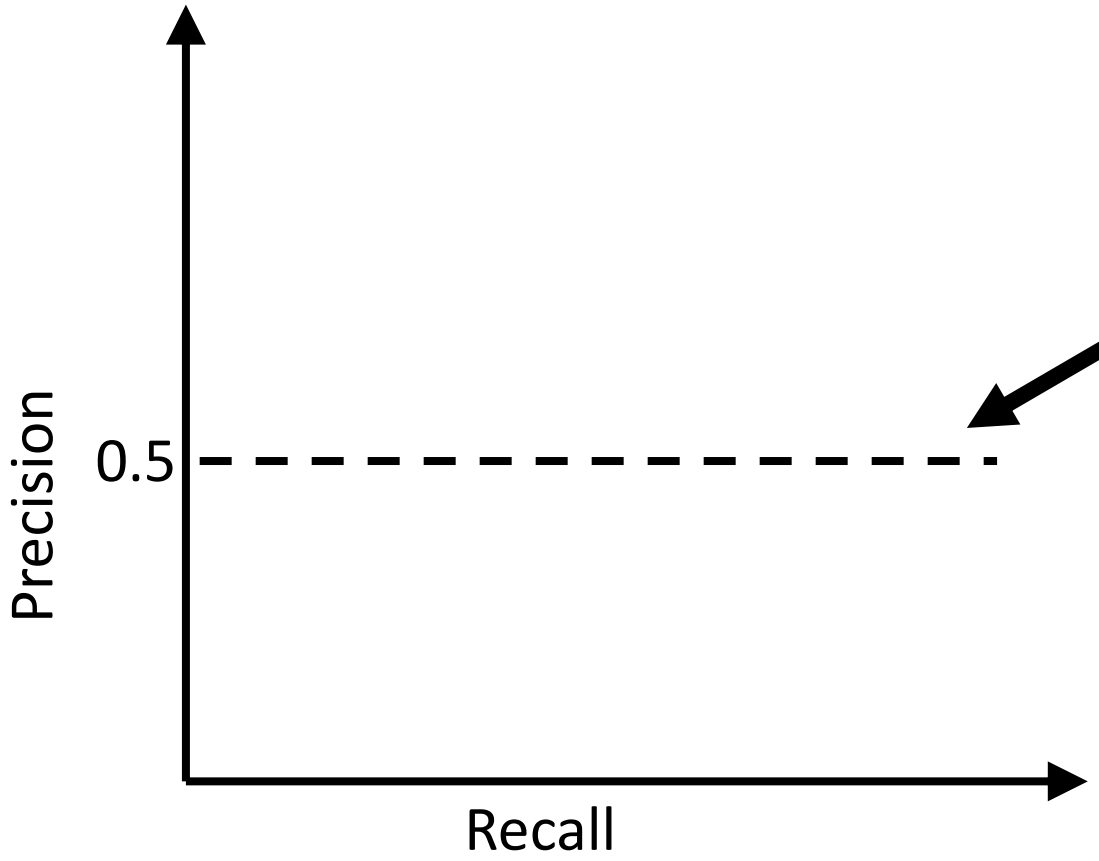
Precision-Recall (PR) curve



Perfect
PR AUC=1

Metrics for classification

Precision-Recall (PR) curve

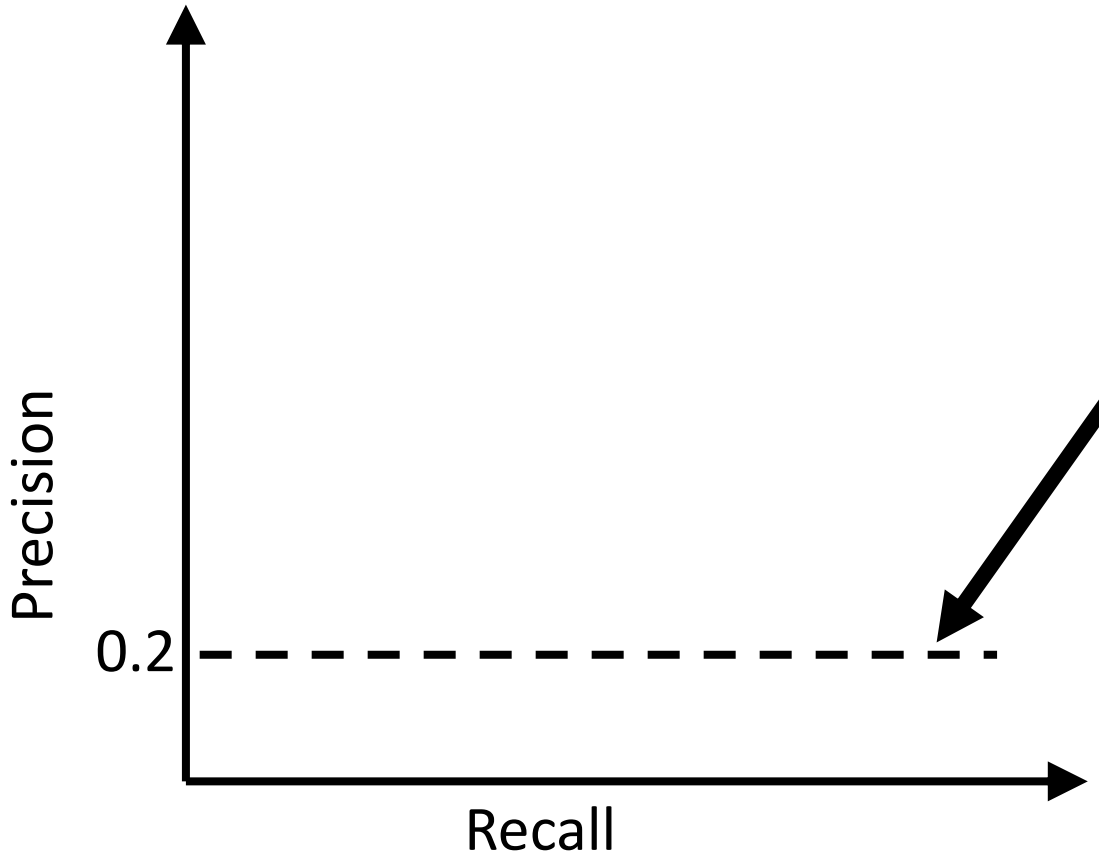


Dummy classifier performance depends on prevalence

Prevalence=0.5
PR AUC=0.5

Metrics for classification

Precision-Recall (PR) curve



Dummy classifier performance depends on prevalence

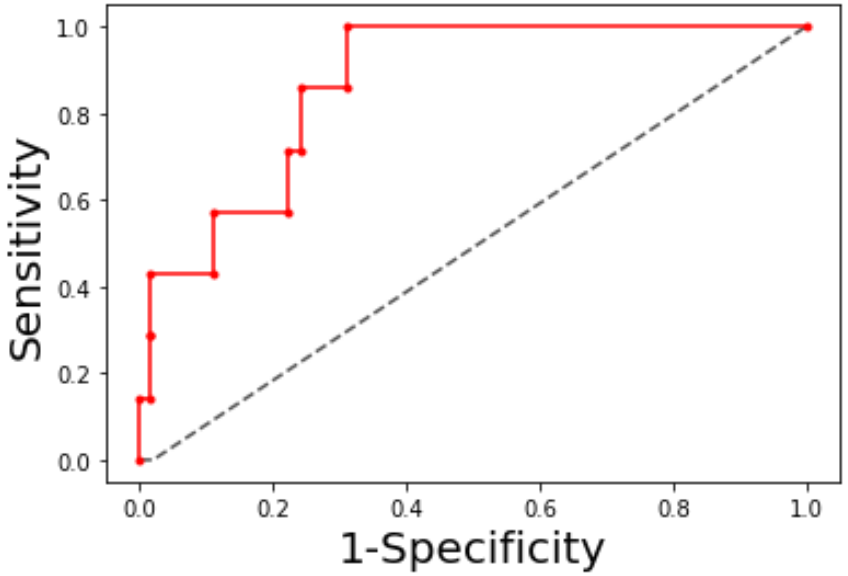
Prevalence=0.2
PR AUC=0.2

Interpretation of the PR AUC depends on the prevalence

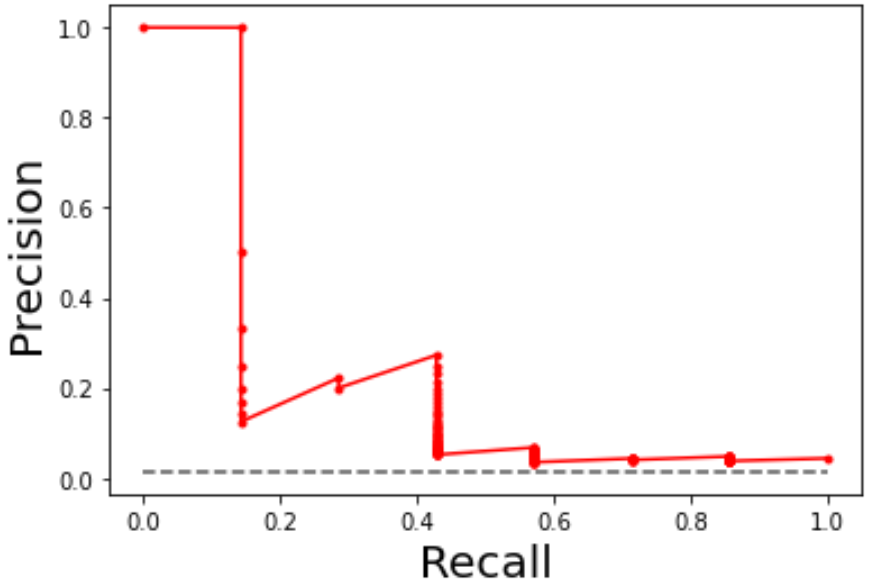
Metrics for classification

Imbalanced datasets

Example for prevalence=0.01



ROC AUC=0.87



PR AUC=0.23

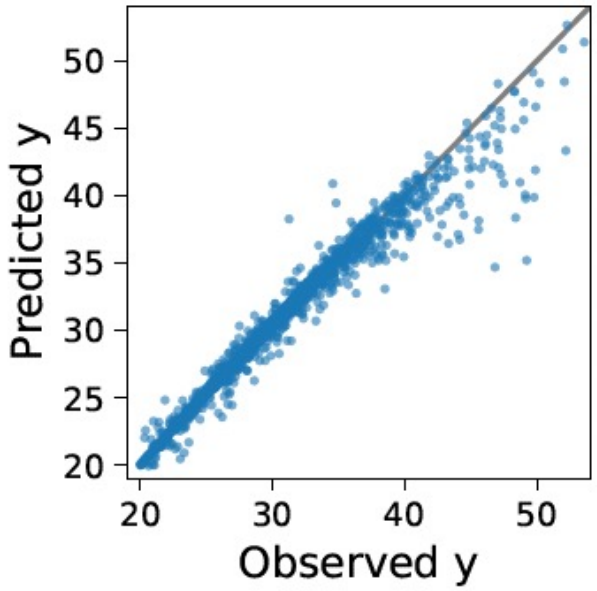
As accuracy, ROC AUC can be misleading when the dataset is imbalanced

Part 3 - Validation
3.2.2 Metrics for regression

Metrics for regression

Visualize prediction errors

Figure 6: ***Visualizing prediction errors*** *– plotting the predicted outcome as a function of the observed one enables to detect structure in the error beyond summary metric. Here the error increases for large values of y, for which there is also a systematic undershoot.*



(Varoquaux and Colliot, Preprint, 2022 - https://hal.science/hal-03682454/)

Metrics for regression

R2 - coefficient of determination

$$R2 = 1 - \frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}}$$

$$\mathrm{SS}_{res} = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

$$\mathrm{SS}_{tot} = \sum_{i=1}^{n} (y^{(i)} - \overline{y^{(i)}})^2$$

- This is computed on the test set (*out-of-sample*)
  - Can be negative (hence call it R2 and not R^2)
  - Is not the square of the correlation coefficient
- Sort of ”explained variance” (but for many authors, explained variance ignores bias)
- Do not use the correlation coefficient (between y and $\hat{y}$) because it discards errors on the mean and the scale -> **important in practice**
- Do not use to compare models because it depends on variance of y

See (Varoquaux and Colliot, Preprint, 2022 - https://hal.science/hal-03682454/)

Metrics for regression

Absolute error measures: RMSE and MAE

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{n}}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y^{(i)} - \hat{y}^{(i)}|}{n}$$

See (Varoquaux and Colliot, Preprint, 2022 - https://hal.science/hal-03682454/)

- Good to compare models
- Give an error in the scale of the outcome (e.g. outcome in years, error in years)
- MAE is easier to interpret
- RMSE will put more weight on rare large errors

$$error = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 100]$$

$$\text{MAE} = 10 \qquad \text{RMSE} \approx 30.17$$

Note that if the error was uniformly equal to the same value (10, for instance), both measures would give the same result.

Part 3 - Validation
3.2.3 Metrics for segmentation

Metrics for segmentation

Here focus on semantic (not instance) segmentation

Segmentation can be seen as classification at each pixel, however:

- Of course, don’t compute metrics across patients (need to compute one metric per patient, then average)
- TN are often meaningless since the background can be arbitrarily large
- Interest in the boundary and the shape of structures
  - Boundary-based metrics
- Interest in the volume of the structures
  - Simple volume agreement measures

Metrics for segmentation

Overlap metrics

- Dice Similarity Coefficient (a.k.a. Soerensen-Dice Coefficient)

$$\mathrm{DSC} = \frac{\mathrm{Vol}(S_r \cap S_p)}{2(\mathrm{Vol}(S_r) + \mathrm{Vol}(S_p))}$$

S_r is the reference (ground truth) segmentation
S_p is the predicted segmentation
$Vol(S)$ denotes the volume of object S

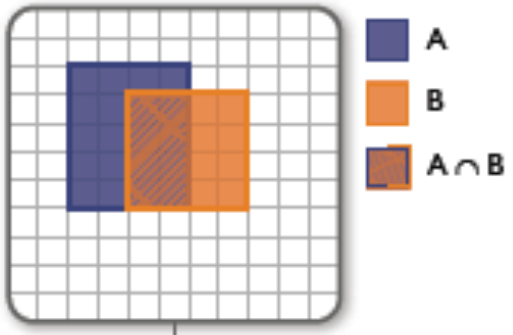


Image from: Reinke et al, Preprint, 2021
https://arxiv.org/pdf/2104.05642.pdf

- Dice is actually the same as the F_1 score

$$\mathrm{DSC} = \frac{2TP}{2TP + FN + FP}$$

Indeed:

$$\mathrm{Vol}(S_r) = TP + FN$$
$$\mathrm{Vol}(S_p) = TP + FP$$

Metrics for segmentation

Overlap metrics

- IoU (Intersection over Union, a.k.a. Jaccard Coefficient)

$$\mathrm{IoU} = \frac{\mathrm{Vol}(S_r \cap S_p)}{\mathrm{Vol}(S_r \cup S_p)}$$

- Dice and IoU provide the same information (they follow the same order) but Dice is always greater than IoU

$$\mathrm{IoU} = \frac{\mathrm{DSC}}{2 - \mathrm{DSC}}$$

$$\mathrm{DSC} = \frac{2\mathrm{IoU}}{1 + \mathrm{IoU}}$$

- No need to report both: report Dice which is more common in medical imaging

Metrics for segmentation

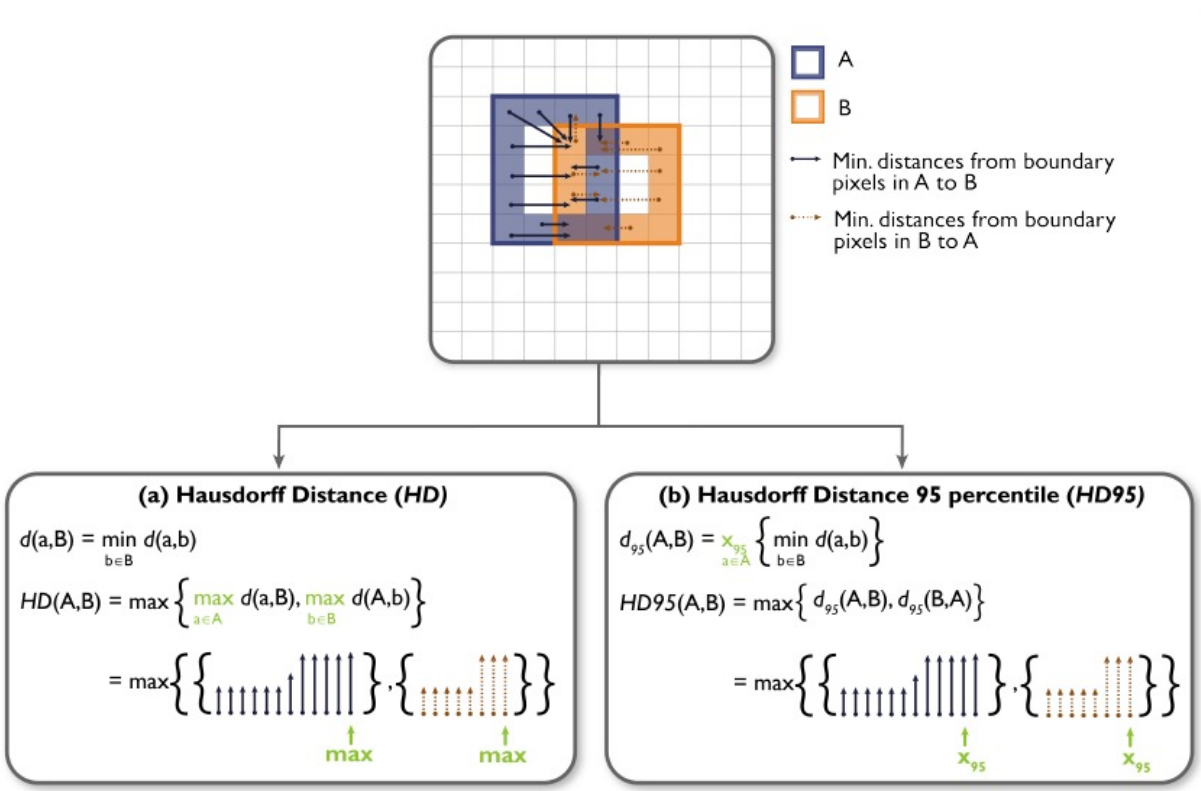
Overlap metrics

- Special cases
  - Want to put more emphasis on FP or FN?
    - Use F_β instead of Dice which is F_1
  - Dealing with tubular structures?
    - See "Center-line Dice" (cl-Dice)

Metrics for segmentation

Boundary metrics

- Hausdorff Distance (HD) and HD95



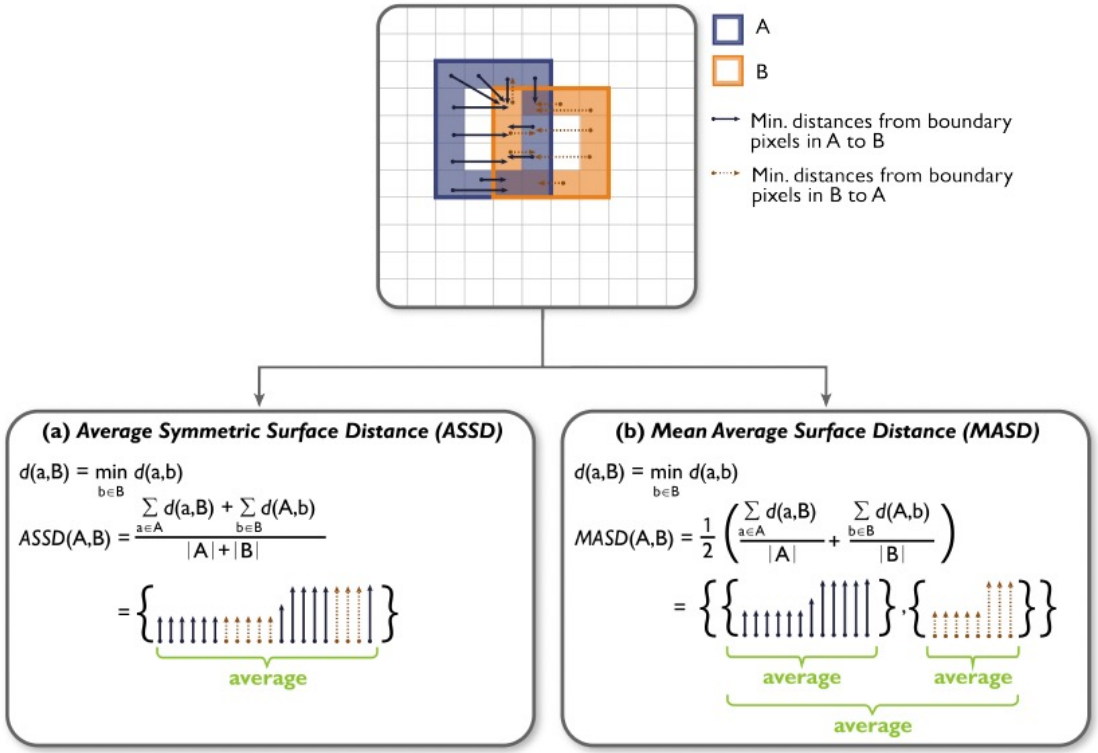
From: Reinke et al, Preprint, 2021 https://arxiv.org/pdf/2104.05642.pdf

- HD95 is less sensitive to outliers
- Allows to detect ”spikes” in the automatic segmentation
- Still the most commonly used boundary metrics
- Have drawbacks when the boundary of the ground truth is imperfect

Metrics for segmentation

Boundary metrics

- Average Surface Distances (ASSD and MASD)



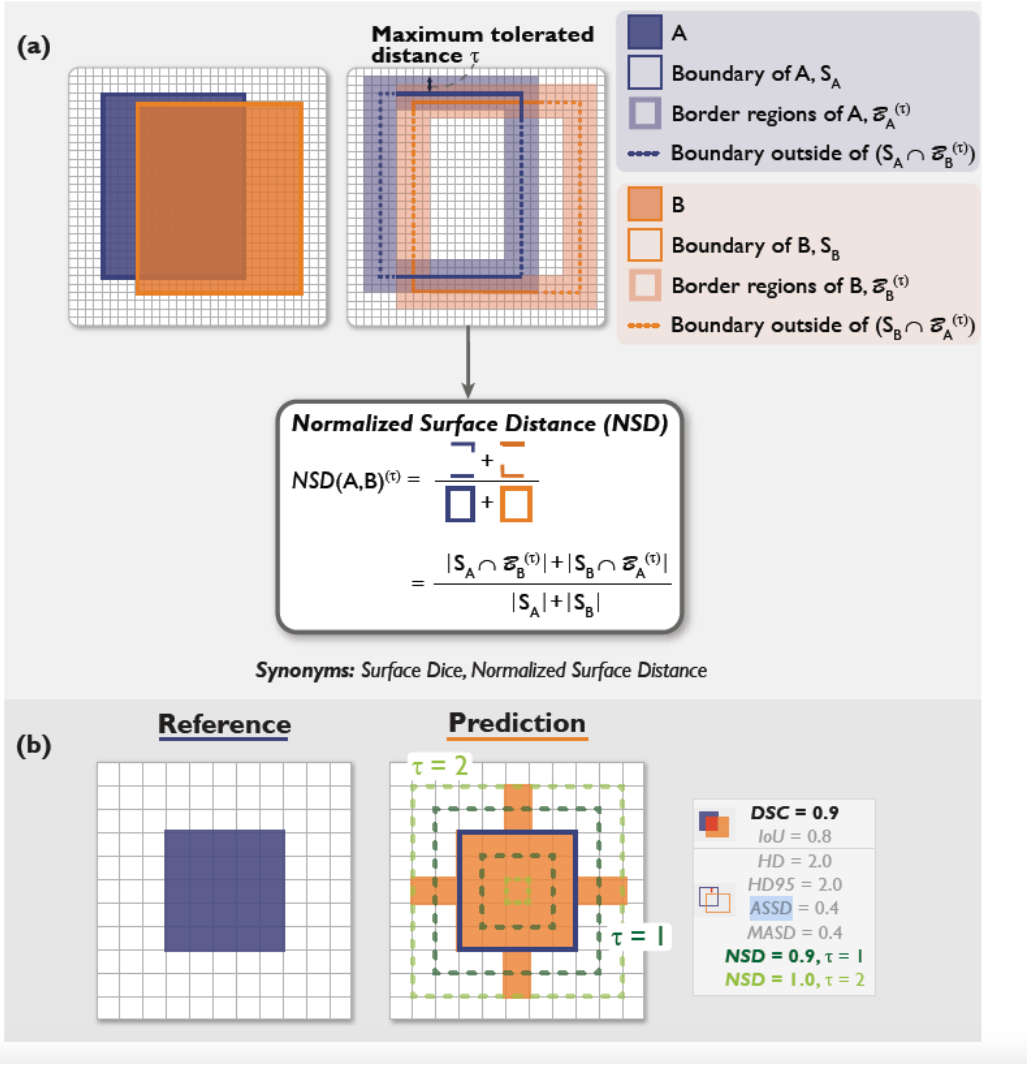
From: Reinke et al, Preprint, 2021 https://arxiv.org/pdf/2104.05642.pdf

- ASSD has the drawback that if a boundary is much larger, it will influence more the metric (solved by MASD)
- Have drawbacks when the boundary of the ground truth is imperfect

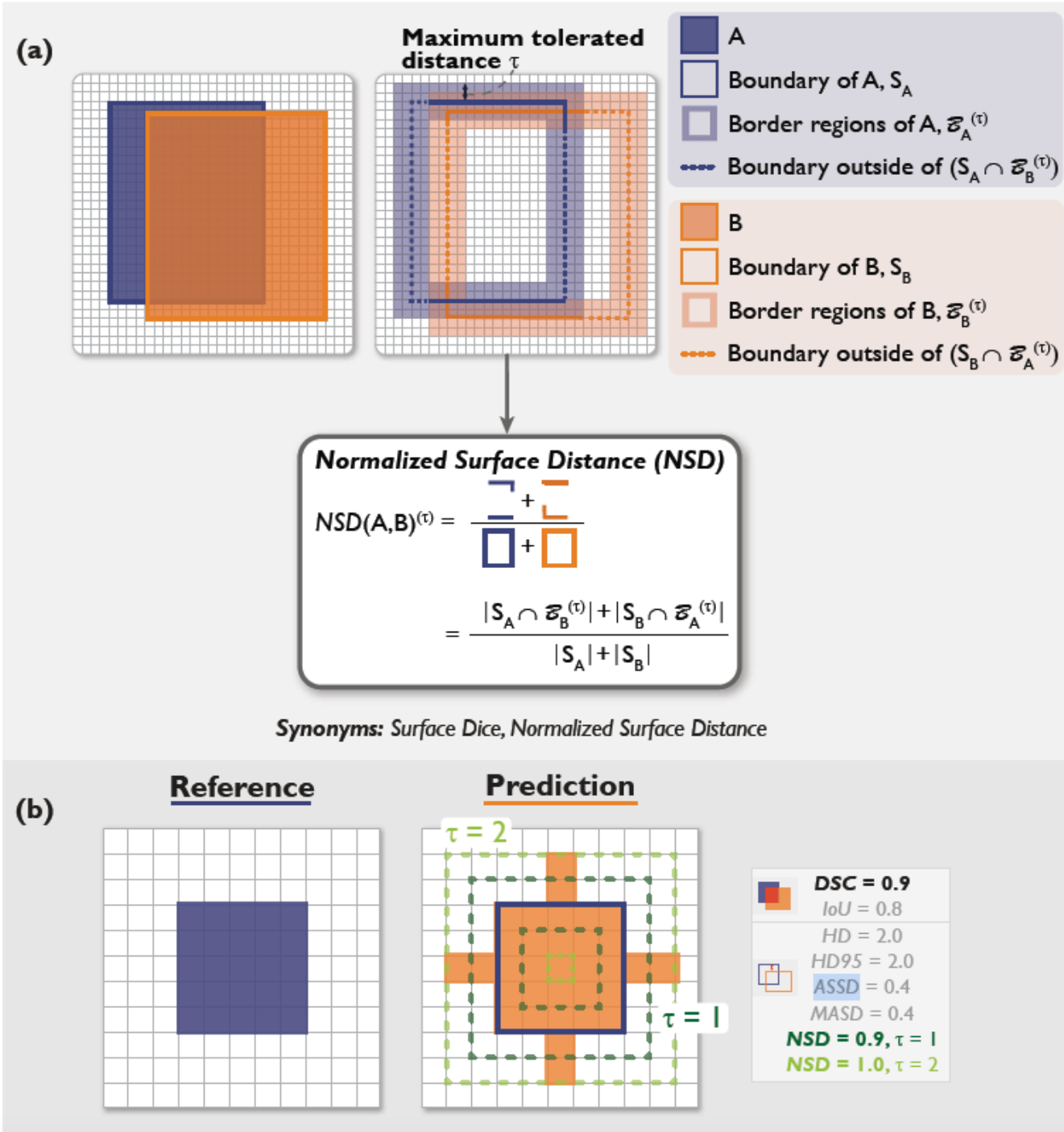
Metrics for segmentation

Boundary metrics

- Normalized Surface Distance (NSD a.k.a. Surface Dice)



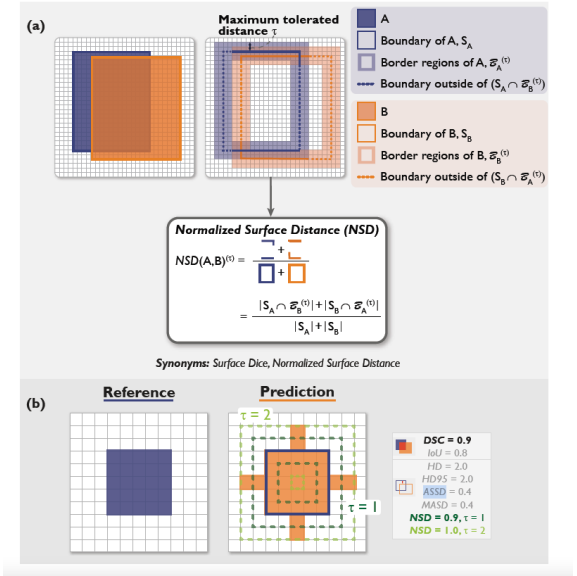
From: Reinke et al, Preprint, 2021 https://arxiv.org/pdf/2104.05642.pdf


(a)
Maximum tolerated distance τ
A
Boundary of A, S_A
Border regions of A, $\mathcal{B}_A^{(\tau)}$
Boundary outside of $(S_A \cap \mathcal{B}_B^{(\tau)})$
B
Boundary of B, S_B
Border regions of B, $\mathcal{B}_B^{(\tau)}$
Boundary outside of $(S_B \cap \mathcal{B}_A^{(\tau)})$
Normalized Surface Distance (NSD)
$NSD(A,B)^{(\tau)} = \frac{+}{+}$
$= \frac{|S_A \cap \mathcal{B}_B^{(\tau)}| + |S_B \cap \mathcal{B}_A^{(\tau)}|}{|S_A| + |S_B|}$
Synonyms: Surface Dice, Normalized Surface Distance
(b)
Reference
Prediction
τ = 2
τ = 1
DSC = 0.9
IoU = 0.8
HD = 2.0
HD95 = 2.0
ASSD = 0.4
MASD = 0.4
NSD = 0.9, τ = 1
NSD = 1.0, τ = 2

Metrics for segmentation

Boundary metrics

- Normalized Surface Distance (NSD a.k.a. Surface Dice)



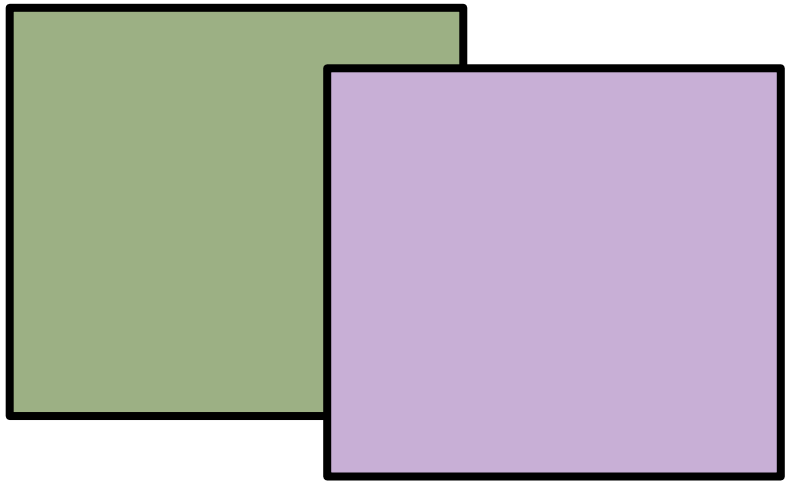
From: Reinke et al, Preprint, 2021
https://arxiv.org/pdf/2104.05642.pdf

- Advantage: defines a tolerance on the distance (in particular useful to disregard small systematic imprecisions in the ground truth)
- Drawback: need to choose maximum tolerate distance **τ**

Metrics for segmentation

Volume metrics

- They take only into account the volume of the segmented structures
  - They can be misleading



Same volume but very bad segmentation

- However, they are useful
  - Volumetry is a key medical application of segmentation
  - They are easy to interpret
- But they should not be used in isolation

Metrics for segmentation

Volume metrics

- Normalized Volume Error Rate
  - Can detect systematic under-/over-estimation (but errors cancel out when averaged across the population)

$$\mathrm{NVER} = \frac{\mathrm{Vol}(S_p) - \mathrm{Vol}(S_r)}{\mathrm{Vol}(S_r)}$$

- Absolute Normalized Volume Error Rate
  - Errors don't cancel out

$$\mathrm{ANVER} = \frac{|\mathrm{Vol}(S_p) - \mathrm{Vol}(S_r)|}{\mathrm{Vol}(S_r)}$$

- Pearson's correlation coefficient (across test set)

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(\mathrm{Vol}(S_p^{(i)}) - \overline{\mathrm{Vol}(S_p^{(i)})})(\mathrm{Vol}(S_r^{(i)}) - \overline{\mathrm{Vol}(S_r^{(i)})})}{\sqrt{\sum_{i=1}^{n}(\mathrm{Vol}(S_p^{(i)}) - \overline{\mathrm{Vol}(S_p^{(i)})})^2}\sqrt{\sum_{i=1}^{n}(\mathrm{Vol}(S_r^{(i)}) - \overline{\mathrm{Vol}(S_r^{(i)})})^2}}$$

Consensus guidelines for metric in medical imaging

Maier-Hein et al, 2022 https://arxiv.org/abs/2206.01653

https://metrics-reloaded.dkfz.de/

Metrics for other tasks cover later in the course

Part 3 - Validation
3.2.4 Aggregate vs individual metrics

Aggregate vs individual metrics

- **Aggregate metrics (at the level of a population)**
  - Classification: Accuracy, AUC, Sensitivity, Specificity…
  - Regression: R2

- **Individual metrics (at the level of a single sample, e.g. individual patient)**
  - For classification, this would typically be 0s and 1s
  - For many tasks, one will have continuous individual metrics
    - E.g.
      - Absolute error (regression) $|y^{(i)} - \hat{y}^{(i)}|$
      - Dice (segmentation)
    - Then they can be aggregated at the level of a population through averaging
      - E.g.
        - Mean absolute error` $\frac{\sum_{i=1}^{n} |y^{(i)} - \hat{y}^{(i)}|}{n}$
        - Mean Dice

Part 3 - Validation
3.3 Validation strategy


There is a a special
place in hell for
people who validate
using the training set

Validation strategies

Samples

Validation strategies

Hold out

`sklearn.model_selection.ShuffleSplit(n_splits=1)`

Larger training set:
better learning

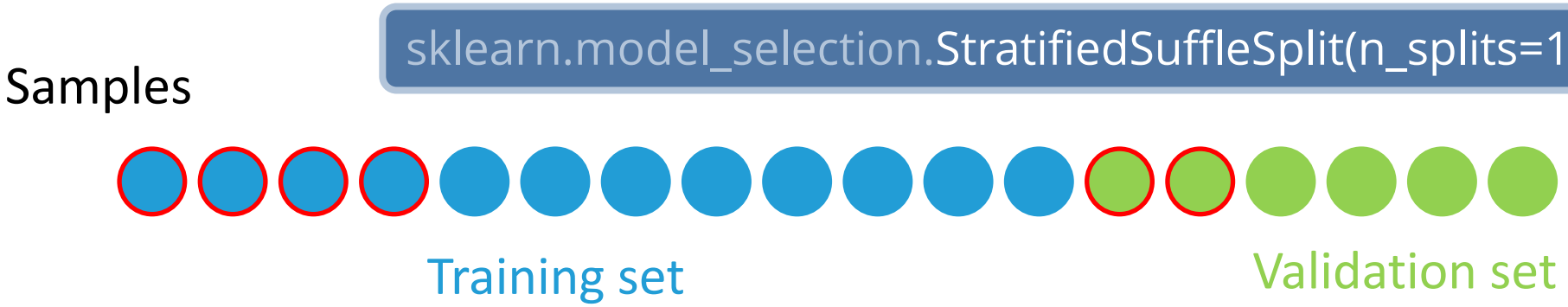
Larger validation set:
better estimation of performance

But data is not infinite → **cross validation**
Idea: repeatedly exchange training and testing data

Validation strategies

Stratification

sklearn.model_selection.StratifiedSuffleSplit(n_splits=1)



Keep the same proportion of each class in the training and validation sets

In the above example 1/3 of samples are diseased and 2/3 are healthy

Validation strategies

Stratification in a broader sense

In many cases, you want the **distribution of several variables** to be the same in the training and validation set (and not only the proportions of the different classes)

For example: age, sex…

This is **very important for medical data** (this issue may be less relevant in other areas such as computer vision)

Validation strategies

Stratification in a broader sense

Example

Table 2. Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores at baseline for ADNI.

	Subjects	Sessions	Age	Gender	MMSE	CDR
CN	330	1 830	74. 4 ± 5.8 [59.8, 89.6]	160 M / 170 F	29.1 ± 1.1 [24, 30]	0: 330
AD	336	1 106	75.0 ± 7.8 [55.1, 90.9]	185 M / 151 F	23.2 ± 2.1 [18, 27]	0.5: 160; 1: 175; 2: 1

Values are presented as mean ± SD [range]. M: male, F: female

Split into validation and test set while preserving **the most important variables**

Validation strategies

Stratification in a broader sense

It is often very difficult to acheive identical (or almost identical) distributions, in particular when controlling for many variables

In practice, one would often be happy if the mean and SD (for continuous variables) and the proportion (for categorical variables) are approximately preserved

Validation strategies

Stratification in a broader sense

Training set

	n_subjects	mean_age	std_age	min_age	max_age	sexF	sexM	mean_MMSE	std_MMSE	min_MMSE	max_MMSE
AD	236	74.995763	7.982102799	55.1	90.9	106	130	23.16949153	2.088325437	18	27
CN	230	74.42087	5.704597622	59.8	88.6	118	112	29.12173913	1.120153919	24	30

Validation set

	n_subjects	mean_age	std_age	min_age	max_age	sexF	sexM	mean_MMSE	std_MMSE	min_MMSE	max_MMSE
AD	100	74.993	7.330733319	55.9	90.3	45	55	23.25	1.986831649	19	27
CN	100	74.415	5.90662975	59.9	89.6	52	48	29.01	1.135737646	26	30

Validation strategies

Stratification in a broader sense

There is no scikit-learn function to perform this

One will often do this using ad-hoc procedures

This is usually done for a separated test set but not for a cross-validation

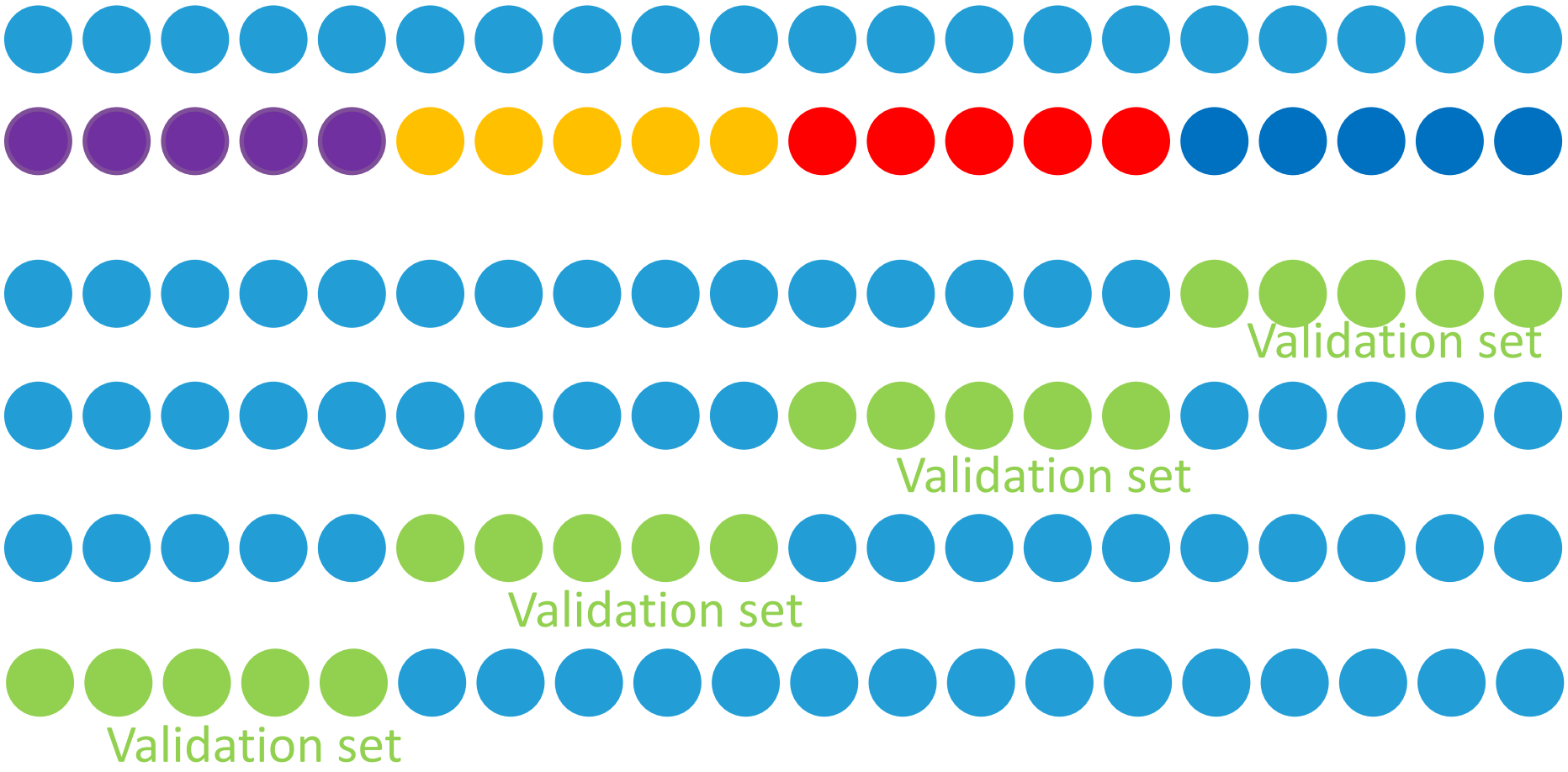
Validation strategies

k-fold cross validation

Here k=4

sklearn.model_selection.KFold

sklearn.model_selection.StratifiedKFold



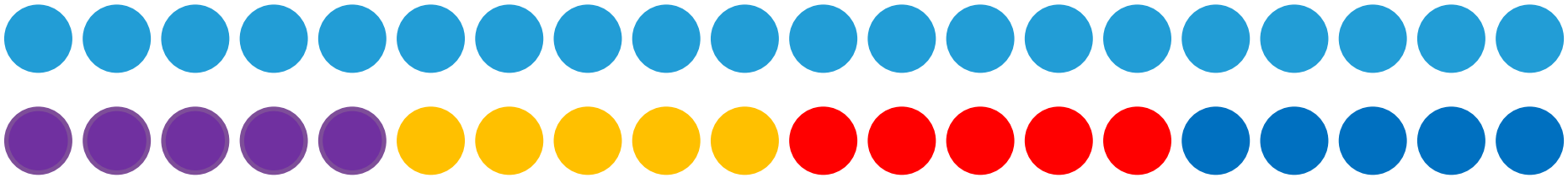
Validation strategies

k-fold cross validation

sklearn.model_selection.KFold

sklearn.model_selection.StratifiedKFold

Samples



Advantage: most efficient (efficient = less computation time) way to use all the samples for training and testing

Drawback: less comprehensive evaluation of the variability of the performance

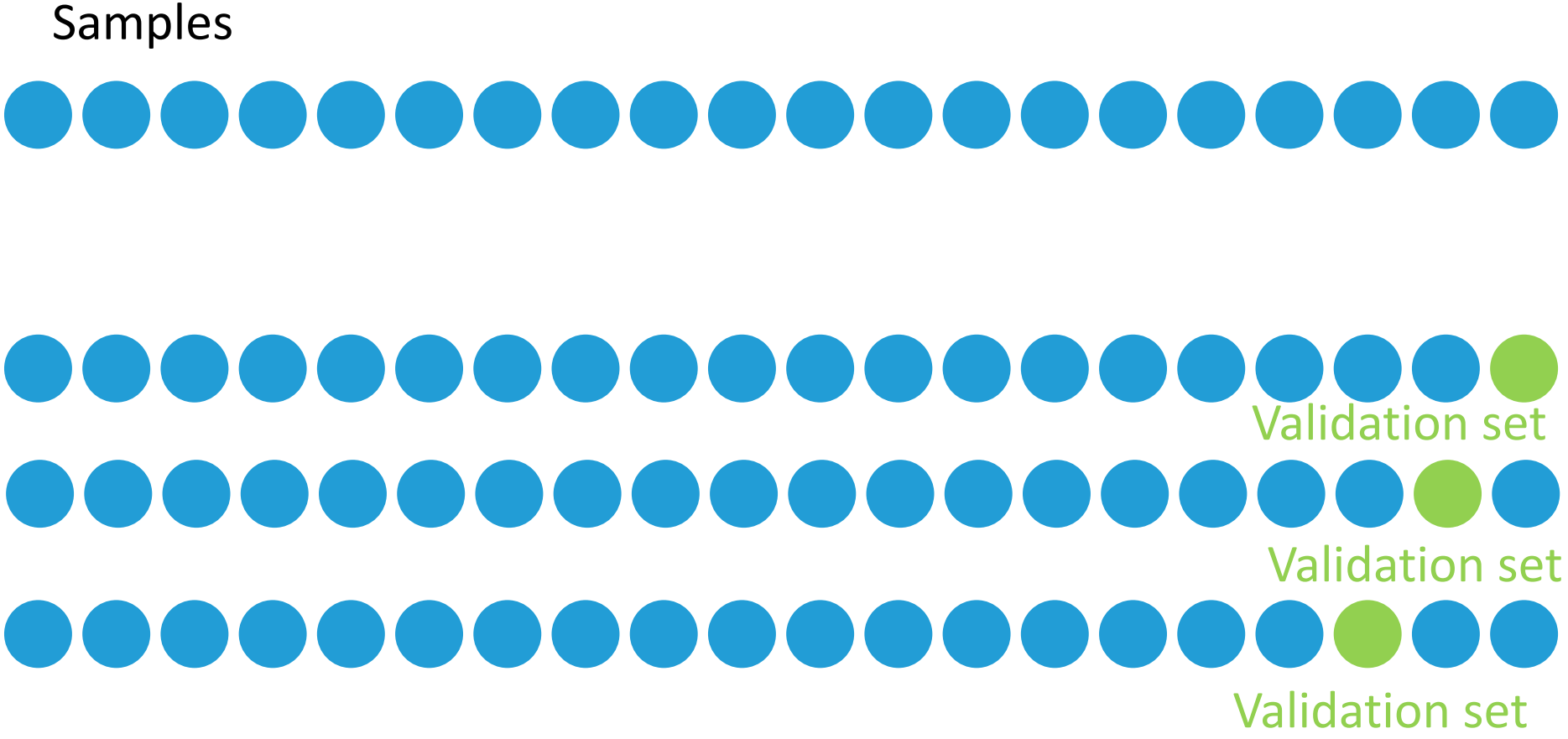
Typical values of k: 5, 10

Validation strategies

sklearn.model_selection.LeaveOneOut

Leave-one-out cross validation

Special case of k-fold with k=n



… In general, one should prefer smaller values of k unless n is really small

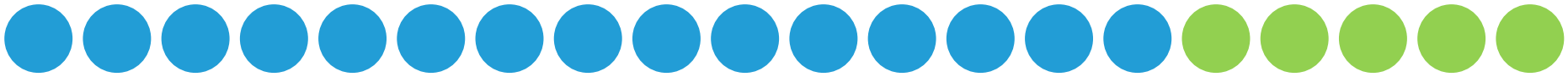
Validation strategies

Repeated hold out

sklearn.model_selection.ShuffleSplit(n_splits)

sklearn.model_selection.StratifiedSuffleSplit(n_splits)

Repeat k times (with large k, for instance 100)



Advantage: comprehensive evaluation of the variability of the performance

Drawback: computationnally expensive

Validation strategies

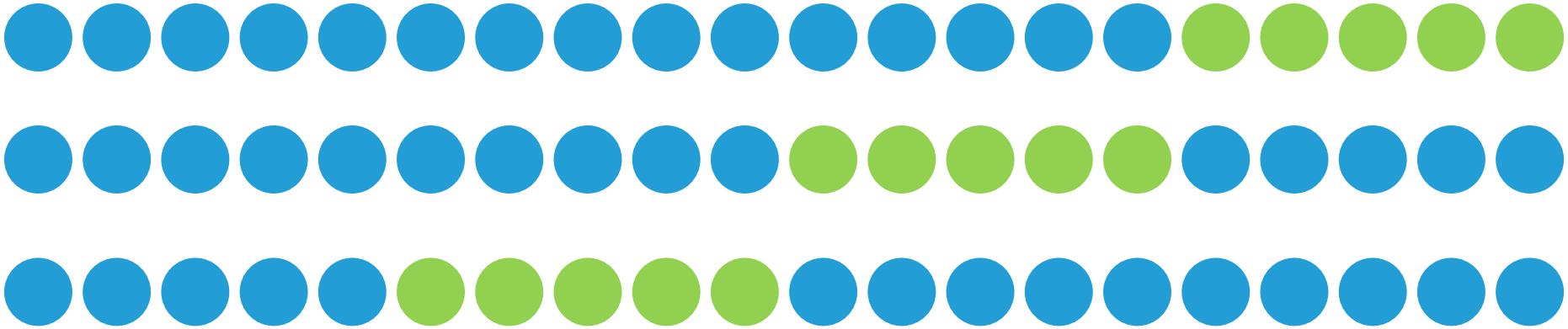
Is this enough?

- If there is no feature selection and a single model without any hyperparameter, yes
- But this is rarely the case

Bad practices

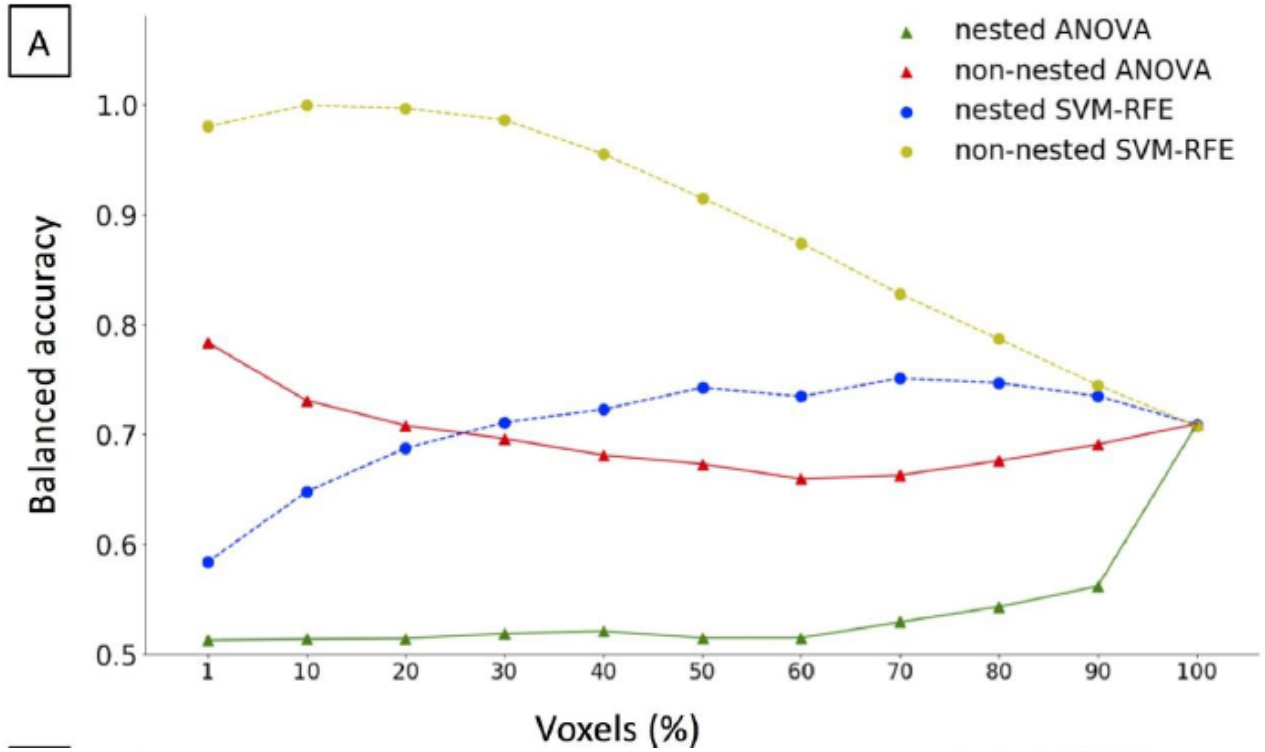
Use all samples for **feature selection**

Then cross-validate the model using the selected features as input



Bad practices

Use all samples for feature selection

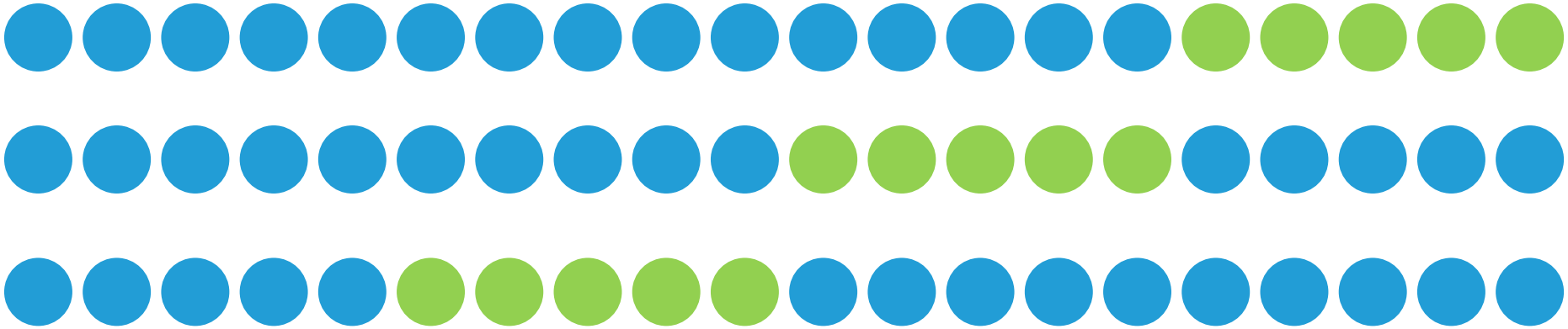


Wen et al, 2018

Bad practices

Use all samples for **dimensionality reduction (e.g PCA)**

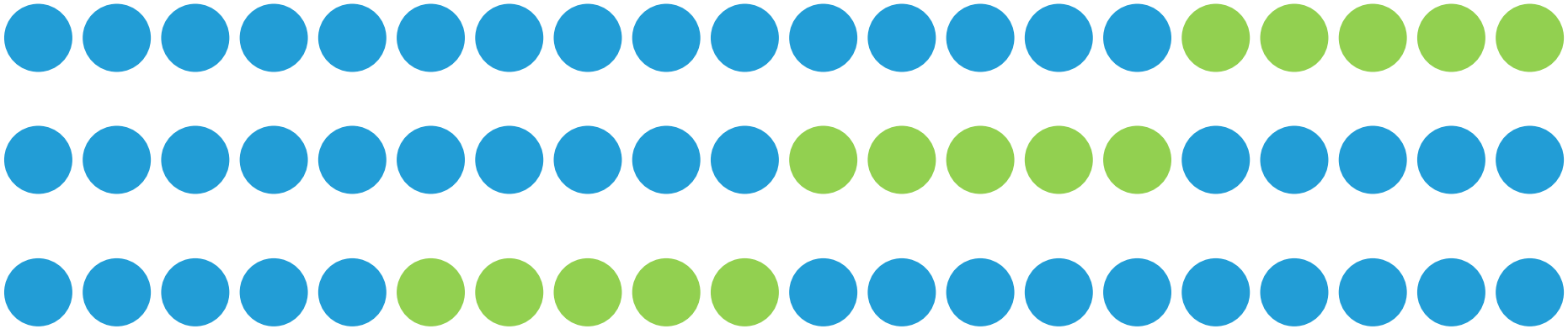
Then cross-validate the model using the reduced features as input



This should not be done but it is probably much less serious than in the case of feature selection

Bad practices

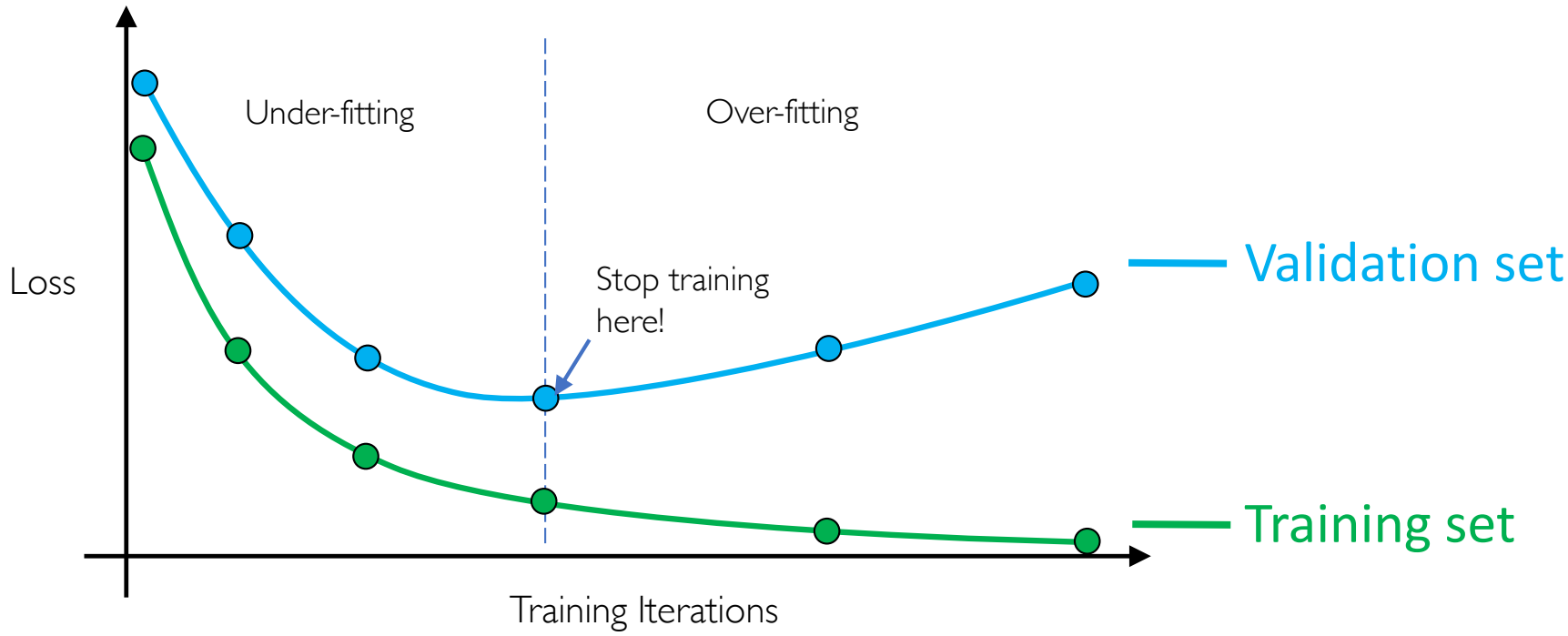
Do multiple runs of cross-validation to select the best hyperparameter



For instance the hyperparameter λ that controls the amount of regularization in l1 (LASSO) and l2 (ridge, SVM…) norm regularized approaches

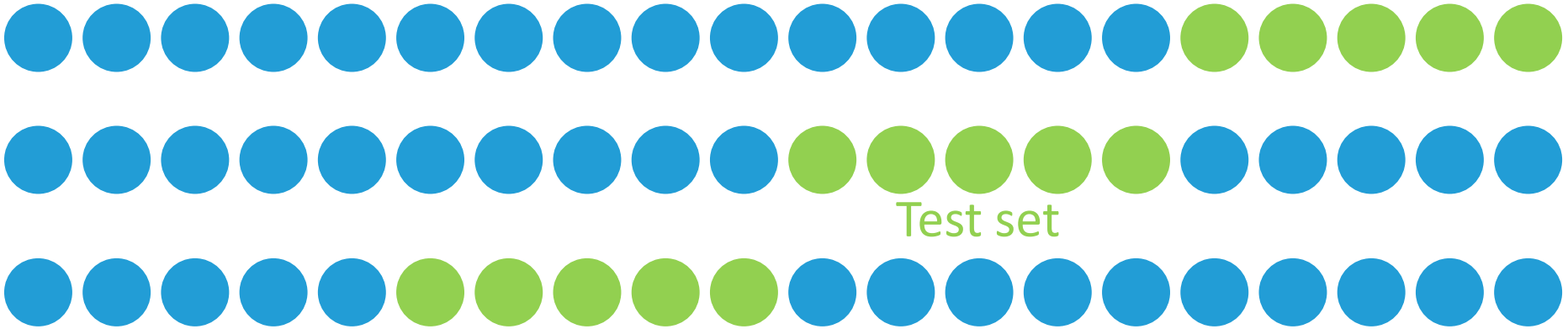
Bad practices

Report the performance obtained on the validation set that was used to decide when to stop training (in deep learning)



Bad practices

Test many possible models and architectures using multiple runs of CV and report the performance of the best performing model



Data leakage

These bad practices are called **data leakage**: some information from the validation set has leaked into the building of the model

Is data leakage prevalent?

Litterature survey of studies using CNNs for Alzheimer's classification from anatomical MRI

(A) Studies without data leakage

Study	DOI	Accuracy	Data leakage
		AD vs CN	
Aderghal et al, 2017	10.1007/978-3-319-51811-4_56	83,70%	None detected
Aderghal et al, 2018	10.1109/CBMS.2018.00067	90%	None detected
Backstrom et al, 2018 *	10.1109/ISBI.2018.8363543	90,11%	None detected
Cheng et al, 2017	10.1117/12.2281808	87,15%	None detected
Cheng and Liu, 2017	10.1109/CISP-BMEI.2017.8302281	85,47%	None detected
Islam and Zhang, 2018 **	10.1186/s40708-018-0080-3	(CN/mild/moderate/ severe: 93,18%)	None detected
Korolev et al, 2017	10.1109/ISBI.2017.7950647	80,00%	None detected
Li et al, 2018	10.1109/IST.2017.8261566	88,31%	None detected
Li et al, 2018	10.1016/j.compmedimag.2018.09.009	89,50%	None detected
Liu et al, 2018	10.1007/s12021-018-9370-4	84,97%	None detected
Liu. et al, 2018	10.1016/j.media.2017.10.005	91,09%	None detected
Liu. et al, 2018	10.1109/JBHI.2018.2791863	90,56%	None detected
Senanayake et al, 2018	10.1109/ISBI.2018.8363832	76%	None detected
Shmulev et al, 2018	10.1007/978-3-030-00689-1_9	(sMCI/pMCI: 62%)	None detected
Valliani and Soni, 2017	10.1145/3107411.3108224	81,30%	None detected

(B) Studies with potential data leakage

Study	DOI	Accuracy	Data leakage	Categories		
		AD vs CN		1	2	3
Aderghal et al, 2017	10.1145/3095713.3095749	91,41%	Unclear	X		
Hon and Khan, 2017	10.1109/BIBM.2017.8217822	96,25%	Unclear	X		X
Hosseini-Asl et al, 2018	10.2741/4606	99,30%	Unclear	X	X	
Islam and Zhang, 2017	10.1007/978-3-319-70772-3_20	(CN/mild/moderate/ severe: 73,75%)	Unclear		X	
Taqi et al, 2018	10.1109/MIPR.2018.00032	100%	Unclear		X	
Vu et al, 2017	10.1109/BIGCOMP.2017.7881683	85,24%	Unclear	X		
Wang et al, 2018	10.1007/s10916-018-0932-7	97,65%	Unclear		X	
Backstrom et al, 2018 *	10.1109/ISBI.2018.8363543	98,74%	Clear	X		
Farooq et al, 2017	10.1109/IST.2017.8261460	(AD/LMCI/EMCI/ CN: 98,88%)	Clear	X		
Gunawardena et al, 2017	10.1109/M2VIP.2017.8211486	(AD/MCI/CN: 96%)	Clear	X	X	
Vu et al, 2018	10.1007/s00500-018-3421-5	86,25%	Clear	X		X
Wang S. et al, 2017	10.1007/978-3-319-68600-4_43	(MCI/CN: 90,60%)	Clear	X		

Table 1. Summary of the studies performing classification of AD using CNNs on anatomical MRI. When d in brackets. (A) Studies without data leakage; (B) Studies with potential data leakage.
Data leakage categories: 1: Biased split; 2: No independent test set ; 3: Late split.
* (Backstrom et al., 2018) experimented two data-partitioning strategies to study the consequences of a l labels.
** Use of imbalanced accuracy on an imbalanced dataset, leading to an over-optimistic estimation of performance.

Over 40% of studies are suspect of data leakage!

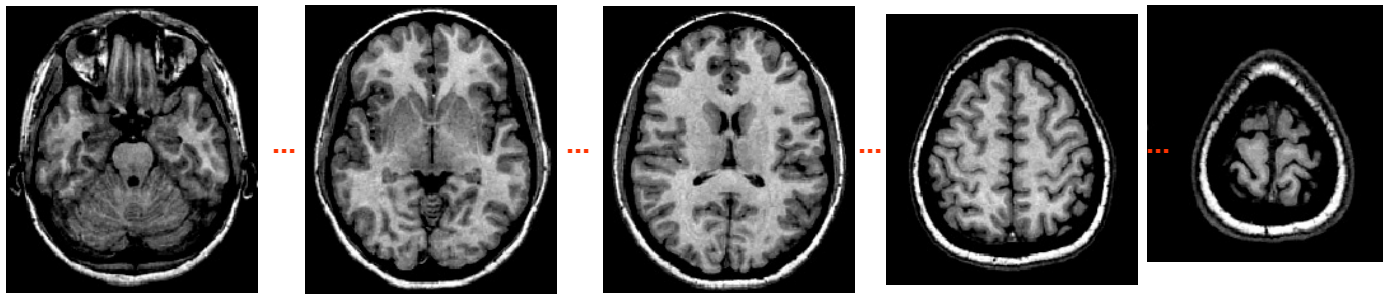
(Wen*, Thibeau—Sutre* et al, 2020)

Fifty shades of data leakage

Bad split of samples between the training, validation and test sets

3D MRI

Splitted at the slice level and not the patient level



5–fold accuracy (with data leakage) 1.00 ± 0 [1.00, 1.00, 1.00, 1.00, 1.00]

True 5 –fold accuracy 0.79 ± 0.04 [0.83, 0.83, 0.72, 0.82, 0.73]

Fifty shades of data leakage

Bad split of samples between the training, validation and test sets

Several visits per patient
Split at the visit level and not the patient level

One can use the following functions to do the split at the patient level (slices or visits will be grouped into patients)

sklearn.model_selection.LeaveOneGroupOut

sklearn.model_selection.LeavePGroupsOut

sklearn.model_selection.GroupKFold

What should we do?

Training, validation and test sets

Use to test different models, to choose when to stop training…

Often through cross-validation

Use only to report performance

What should we do?

Training, validation and test sets

Cross-validation

Standard (minimal) good practice for deep learning

Training, validation and test sets

Use cross-validation, often with k=3 to 5, to train the model, experiment with different architectures…

The test set

Where should I keep the test set?

In a safe!

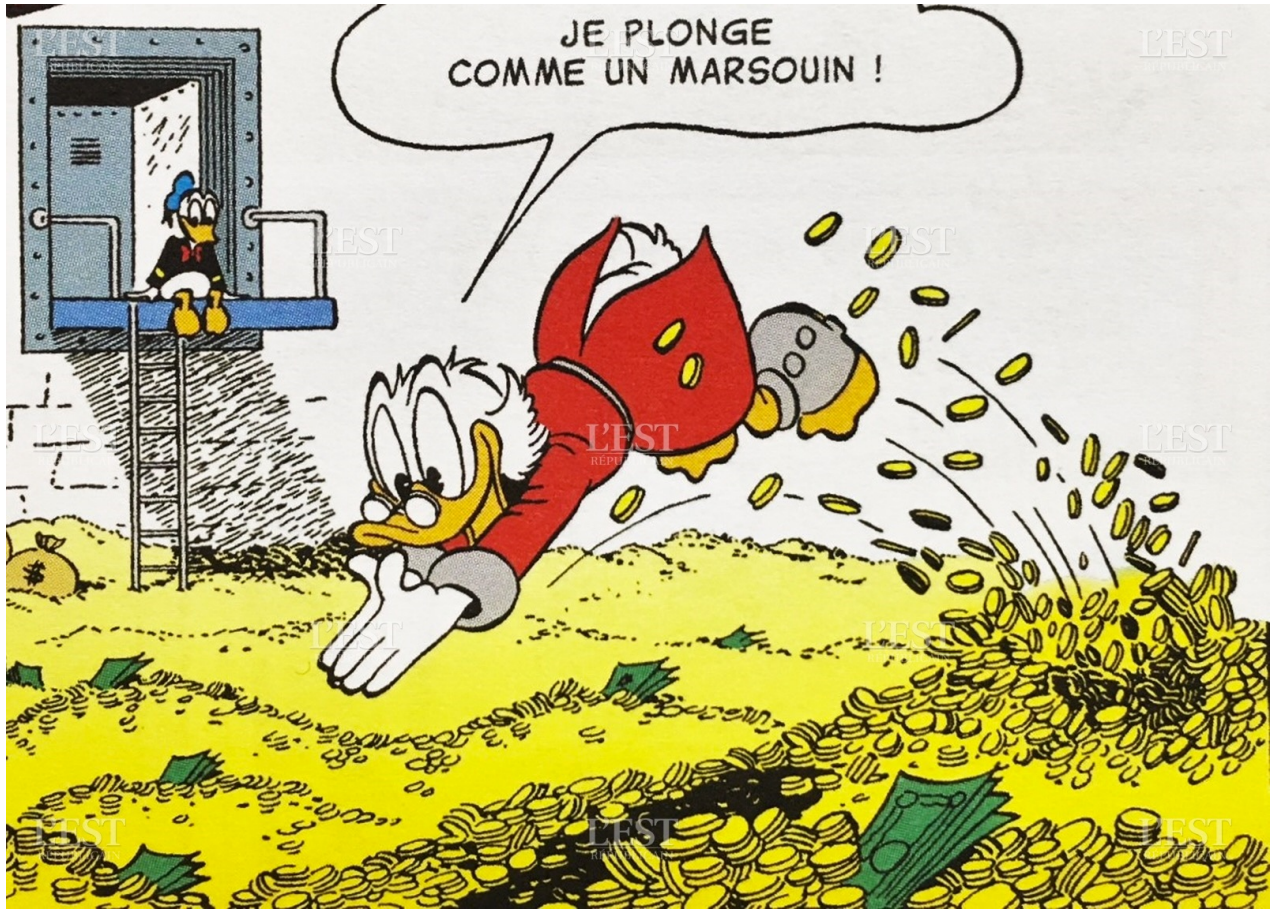
The test set

When should I separate the test set?

Before starting the work

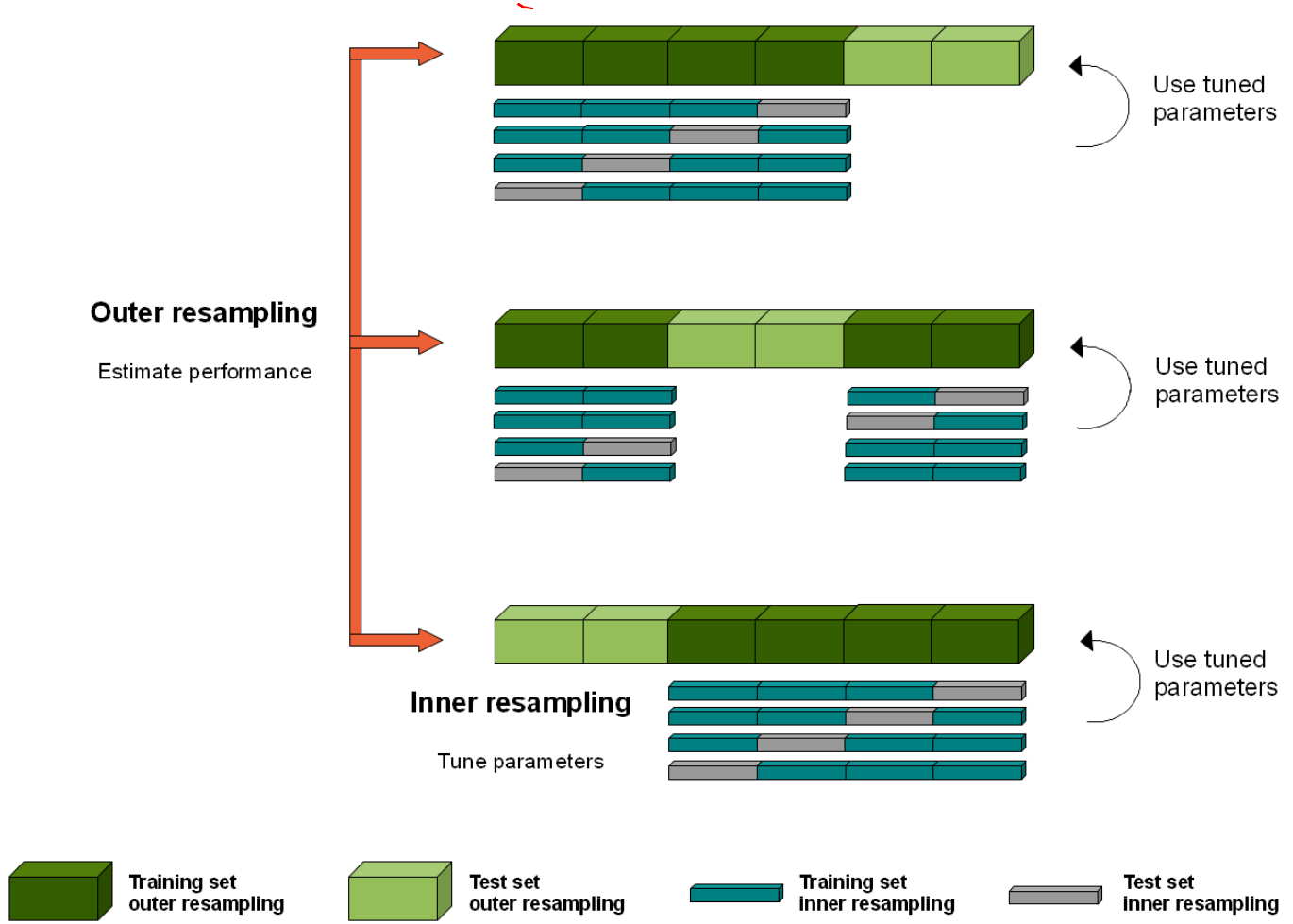
The test set

And make sure the person training the model doesn't have the key!



Other solution

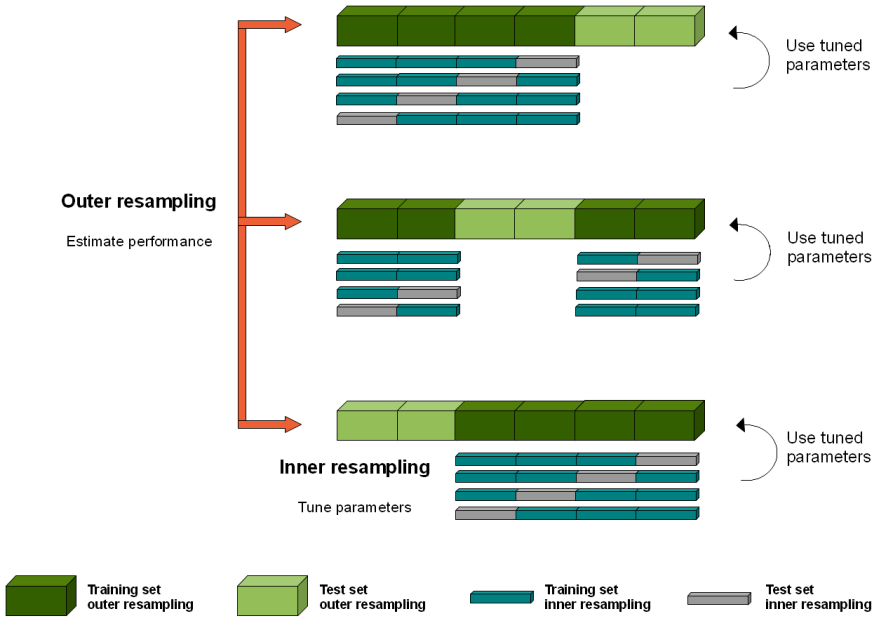
Nested cross-validation



https://i.stack.imgur.com/vh1sZ.png

Other solution

Nested cross-validation



Computationally expensive

Can be feasible with models with fast training

https://i.stack.imgur.com/vh1sZ.png

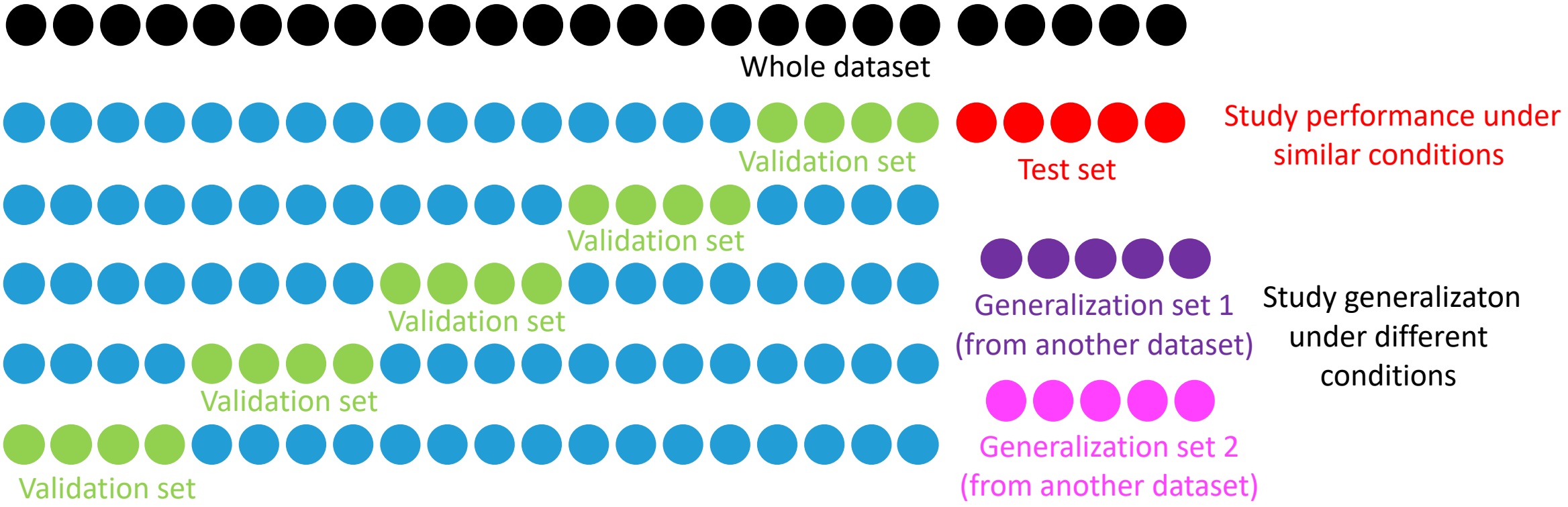
Standard (minimal) good practice for deep learning

Training, validation and test sets

Use cross-validation, often with k=3 to 5, to train the model, experiment with different architectures…

Even better

Studying generalization



Use to train the model, experiment with different architectures…

Part 3 - Validation
3.4 Statistical analysis

	Balanced accuracy
3D ResNet	72.1
TrickOfTheTradeCNN [11]	72.3
ZorglubFormer (proposed)	**72.7**

Should we be satisfied with this?

Statistical analysis

- **We are going to consider two cases**
  - **Trained models**
    - The model has been trained, the only source of variation is the test set
  - **Learning procedure**
    - You want to know how variable is the performance with respect to different sources
      - Test set
      - Training set
      - Hyperparameters
      - Initialization
      - ….

Statistical analysis

- **We are going to look at two types of statistics**
  - **Descriptive statistics**
    - How variable is your performance?
  - **Inferential statistics**
    - How precise is the estimate of your performance?
    - Can you claim that one model is better than another?

Descriptive statistics

- **Can you give some examples of ways to assess of variability the performance?**

Descriptive statistics

- **Can you give some examples of ways to assess of variability the performance?**
  - **Measures**
    - Standard-deviation
    - IQR (inter-quartile range)
    - Min, max
    - Deciles, centiles
  - **Graphs**
    - Bar plots
    - Box plots
    - Violin plots
    - Jittered points

Descriptive statistics

- **Can you give some examples of ways to assess of precision of an estimate?**

Descriptive statistics

- **Can you give some examples of ways to assess of precision of an estimate?**
  - **Measures**
    - Confidence interval
    - Standard error
  - **Graphs**
    - Bar plots
    - Box plots
    - …

Part 3 - Validation
3.4.1 Statistical analysis: variability (descriptive statistics)

Statistical analysis

- **We are going to consider two cases**
  - **Trained models**
    - The model has been trained, the only source of variation is the test set
  - **Learning procedure**
    - You want to know how variable is the performance with respect to different sources
      - Test set
      - Training set
      - Hyperparameters
      - Initialization
      - ….

Statistical analysis

- **We are going to consider two cases**
  - **Trained models**
    - The model has been trained, the only source of variation is the test set
  - **Learning procedure**
    - You want to know how variable is the performance with respect to different sources
      - Test set
      - Training set
      - Hyperparameters
      - Initialization
      - ….

Descriptive statistics: trained models

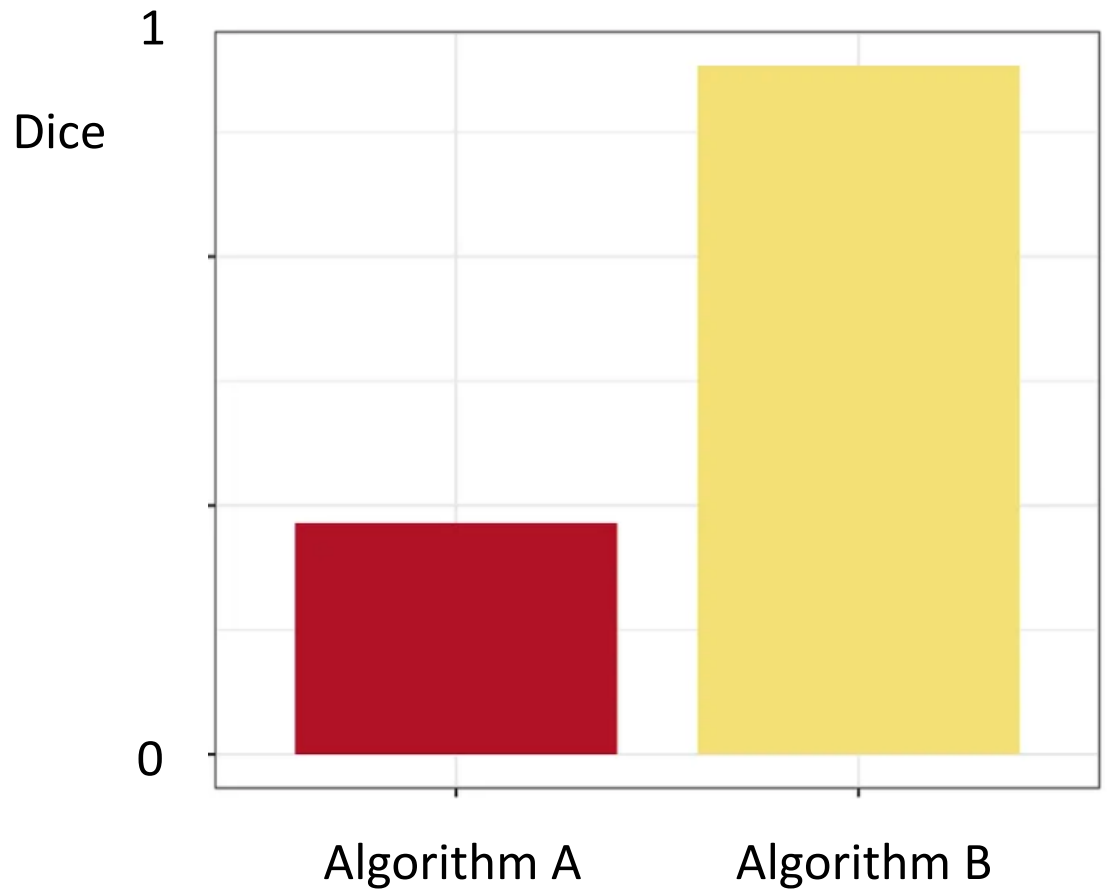
- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph

Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

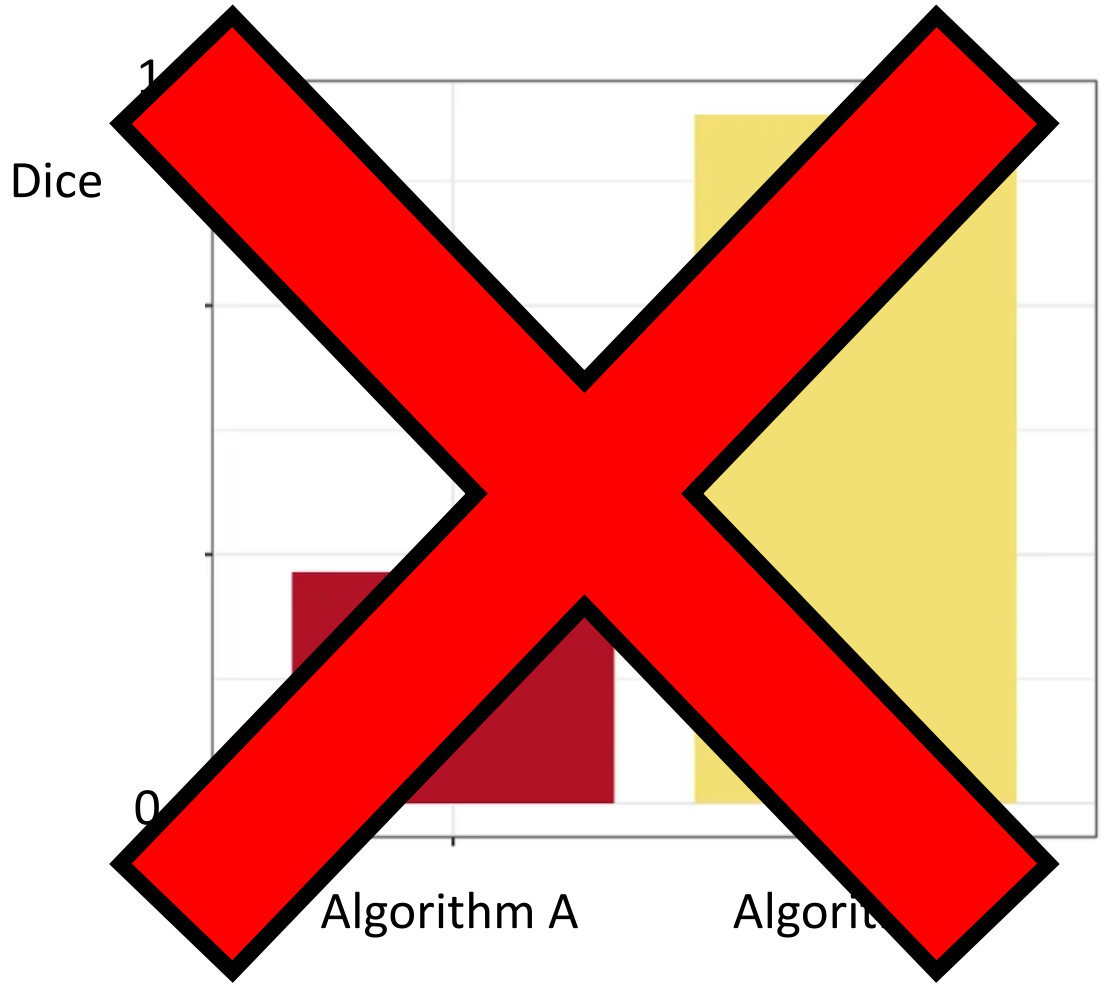
- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph



Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

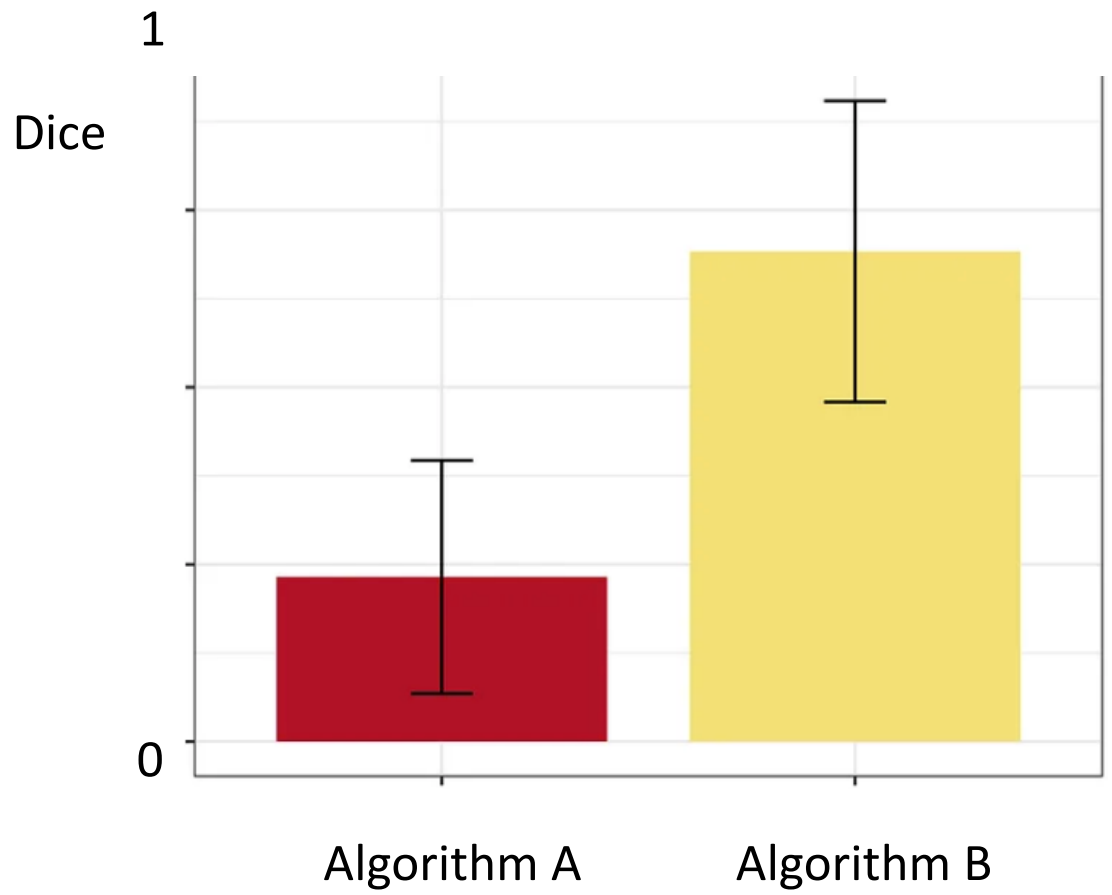
- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph



Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph

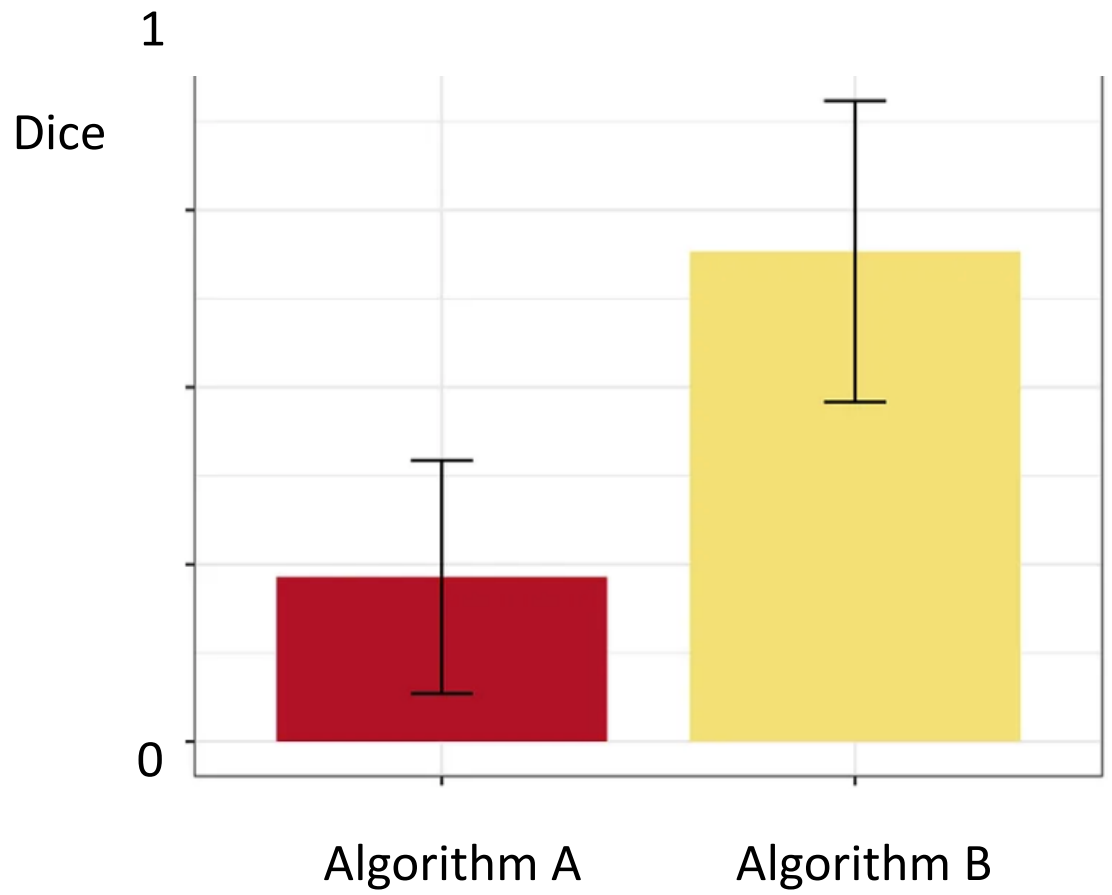


Is it enough?

Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph

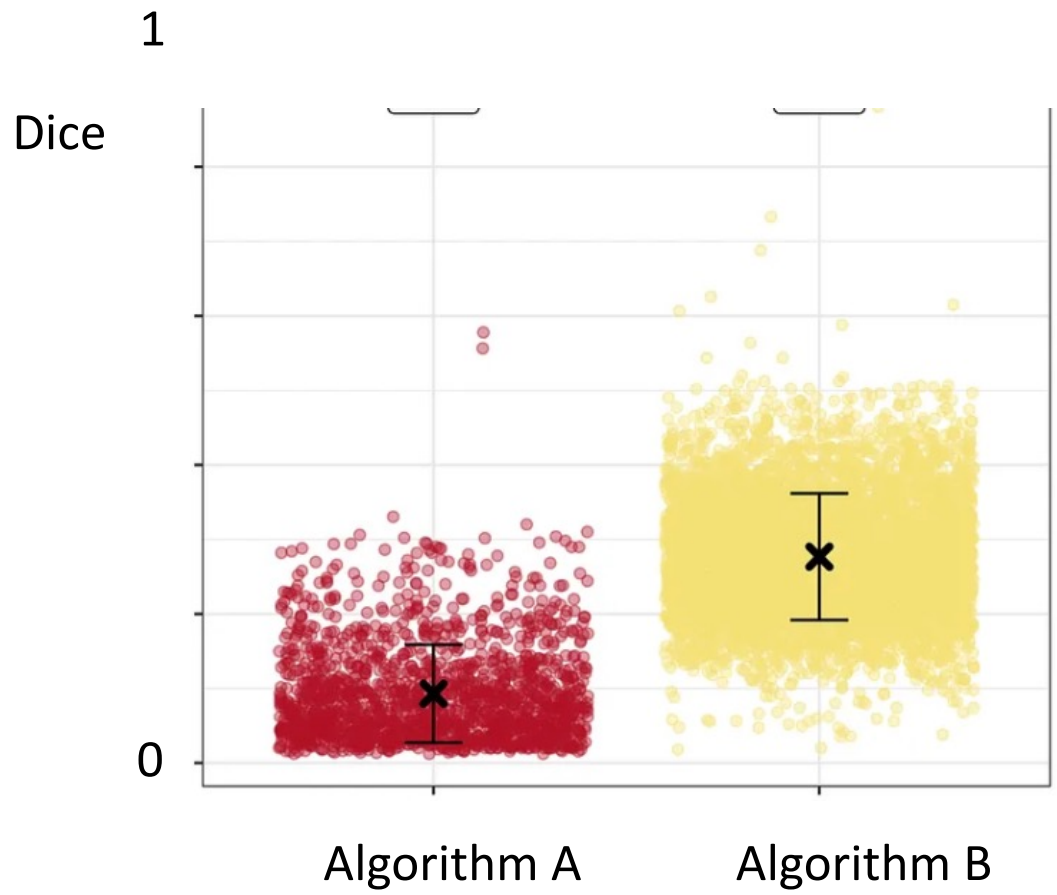


Error bars represent the standard-deviation computed on the test set

Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph

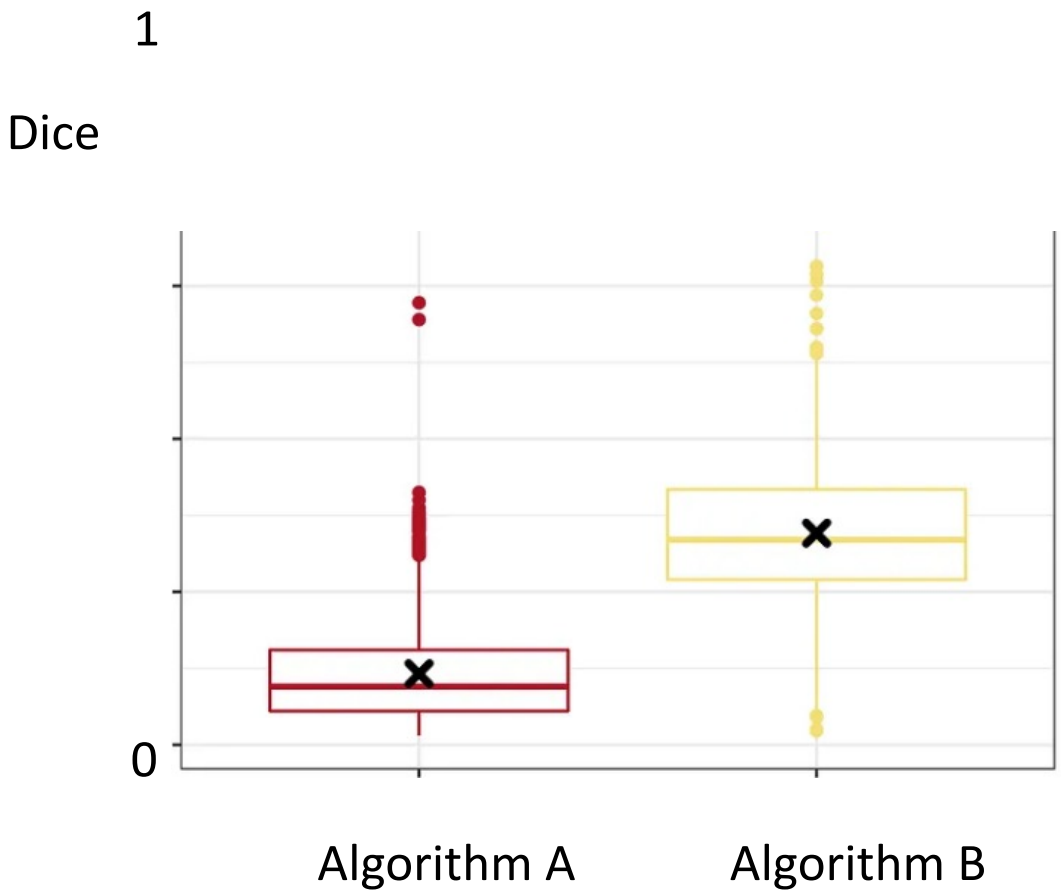


Maybe standard deviation was not enough

Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph

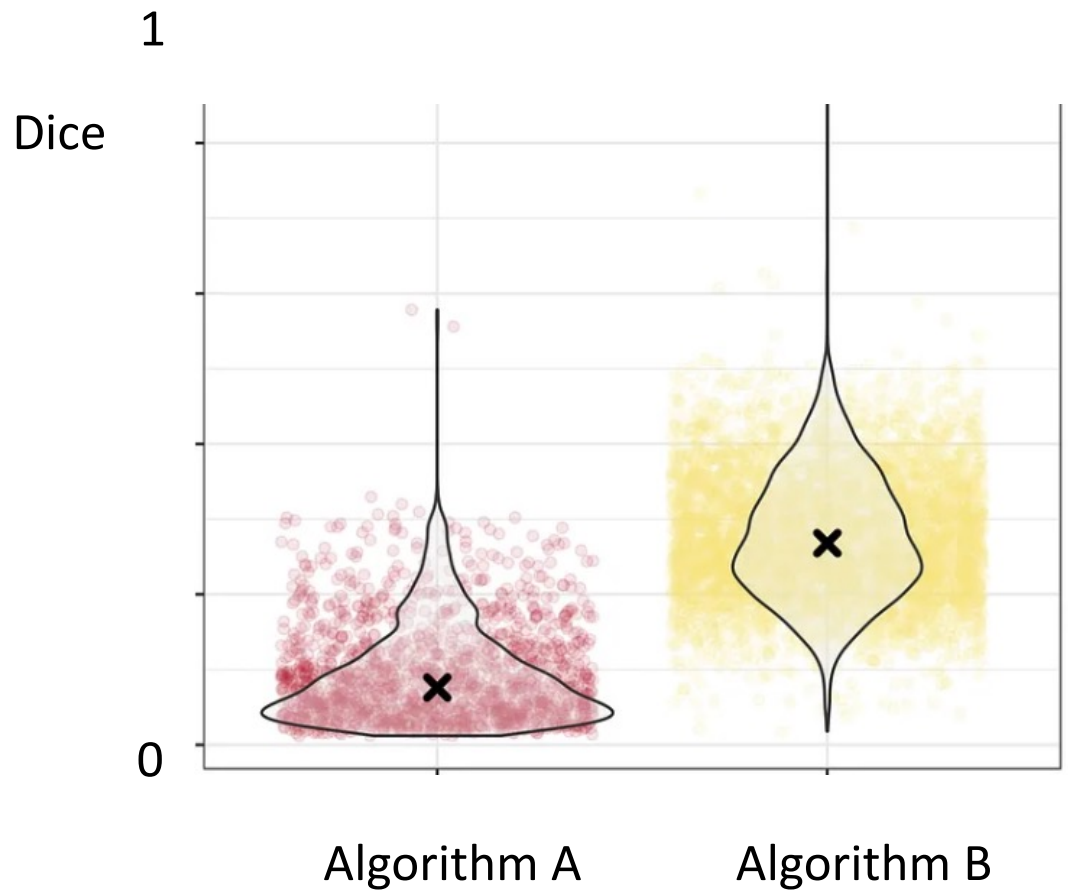


The box extends from Q1 to Q3

Descriptive statistics: trained models

- **Assess the variability of the performance**
  - Is it stable?
  - Are there extreme cases? (complete failures)

- **What should we do?**
  - Plot the distribution
    - Are mean and standard-deviation meaningful?
  - Report in tables
    - Mean and standard-deviation
      - or
    - Median and IQR
  - If enough space
    - Provide a graph



Violin plot with median and jitter points