Deep learning for medical imaging

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

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

Master 2 - MVA

Course website: http://www.aramislab.fr/teaching/DLMI-2019-2020/
Piazza (for registered students): https://piazza.com/centralesupelec/spring2020/mvadlmi/

Part 6 - Validation
6.1 Introduction

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!!)

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**

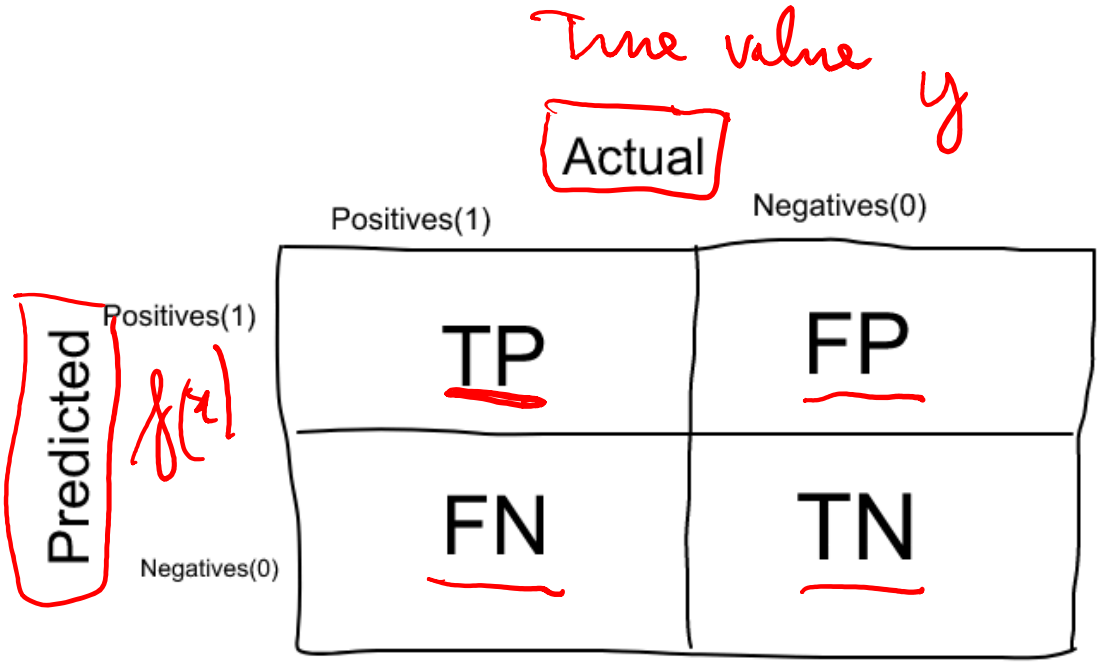
Introduction

- In this part, we address
  - **How to quantify the performance of the model?**
    - Performance metrics
  - **How to estimate the performance metrics?**
    - Validation strategies
  - **How to make your research reproducible?**
  - **How to make interpret a neural network?**

Part 6 - Validation
6.2 Performance metrics

Metrics for classification

Confusion matrix



Source of images: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
https://en.wikipedia.org/wiki/Precision_and_recall

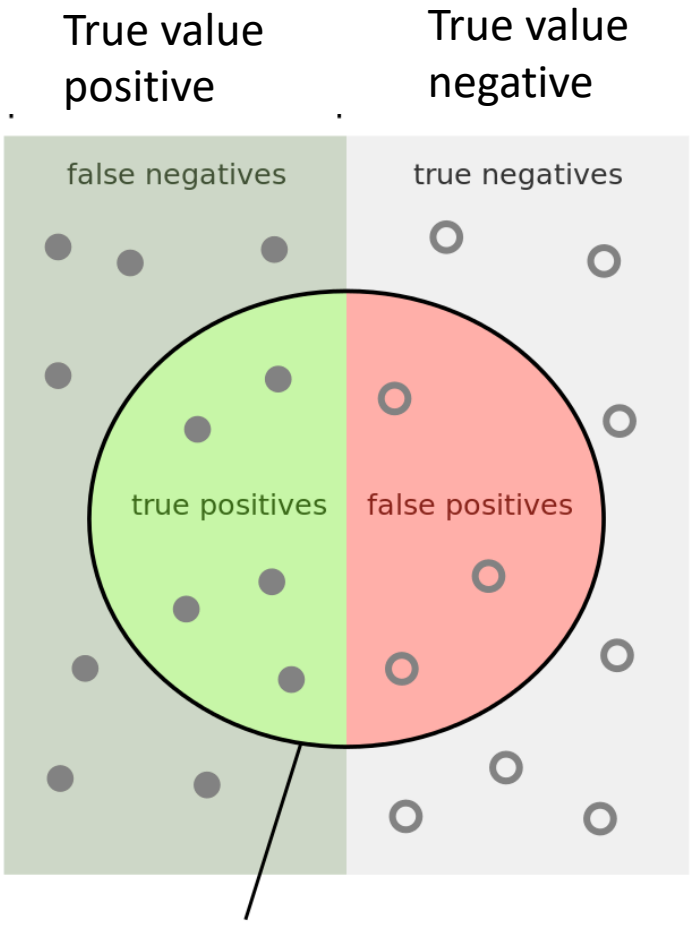
Metrics for classification

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

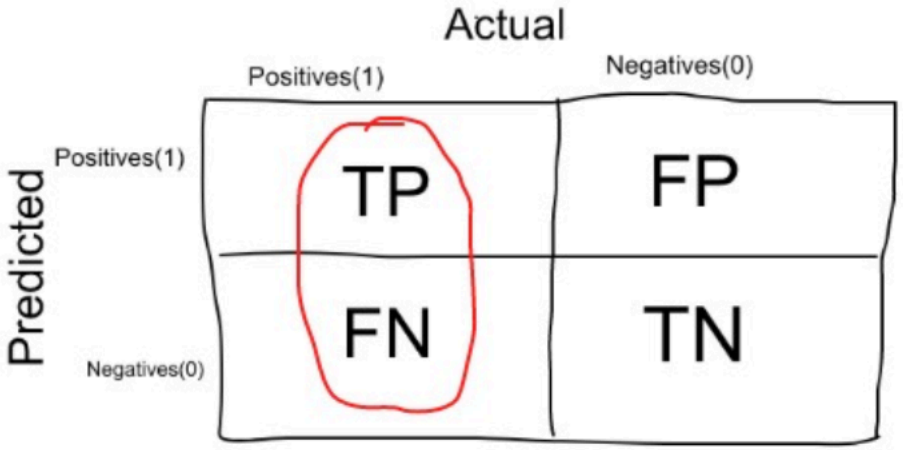
- Example 1: cancer screening
  - We should not miss any cancer cases
  - One may consider requiring have **very few false negatives even at the expense of relatively high proportion of false positives**
  - Positive cases would then be reviewed by an expert or lead to additional explorations

- 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)

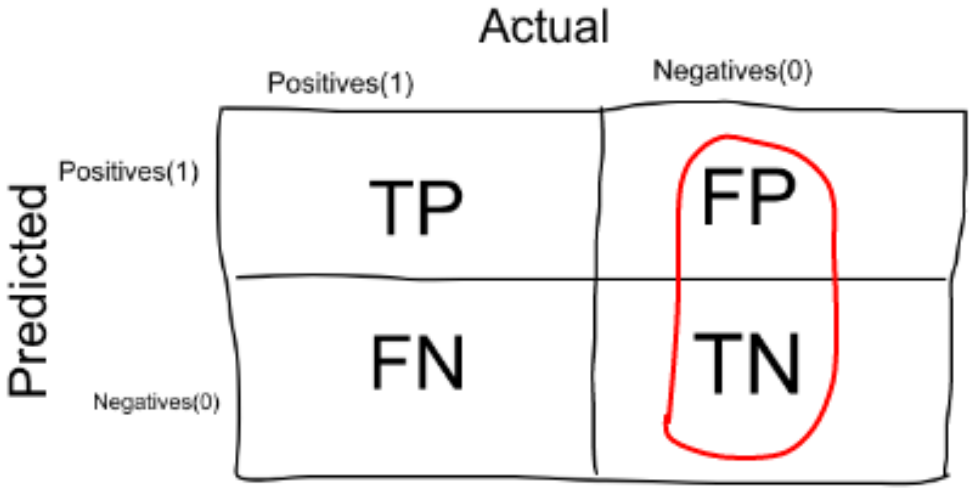


$$\text{Recall} = \frac{TP}{TP + FN}$$

Source of images: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Metrics for classification

Specificity

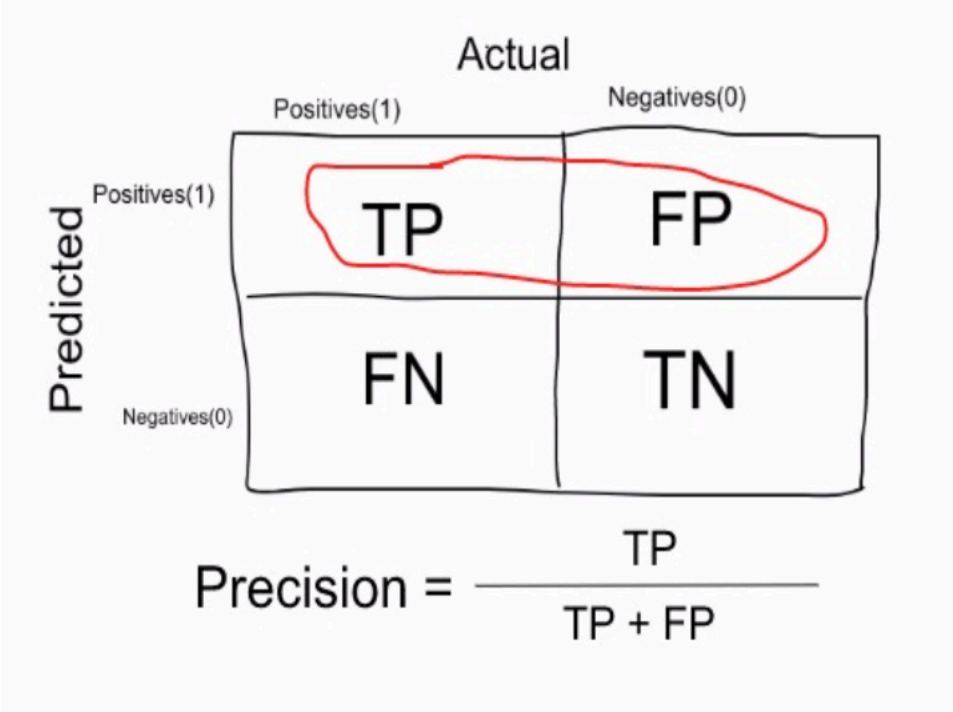


$$\text{Specificity} = \frac{TN}{TN + FP}$$

Source of images: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Metrics for classification

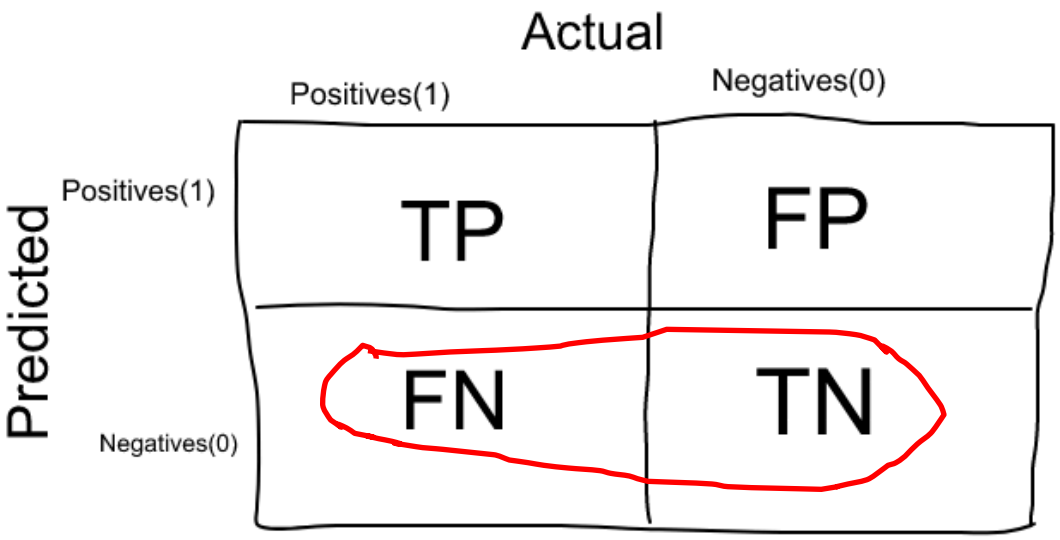
Precision (also called positive predictive value)



Source of images: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Metrics for classification

Negative predictive value

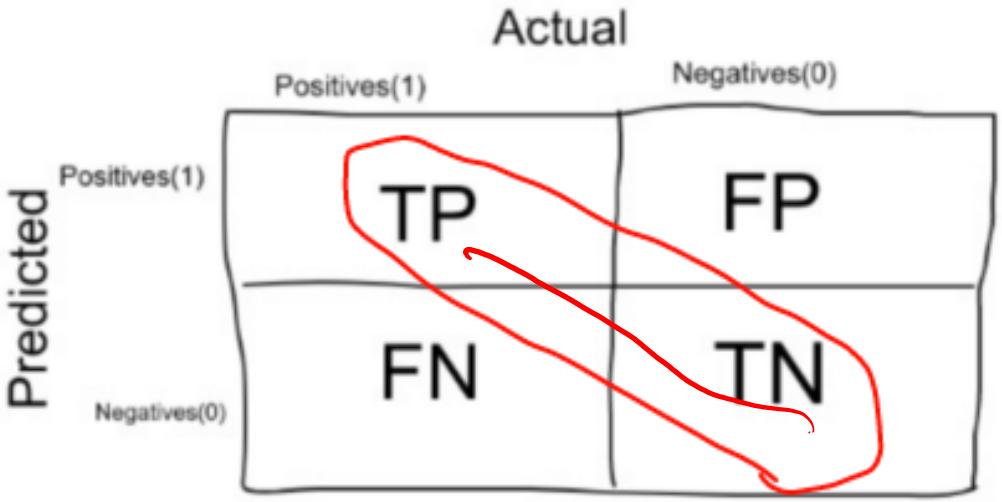


$$\frac{TN}{FN+TN}$$

Source of images: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
https://en.wikipedia.org/wiki/Precision_and_recall

Metrics for classification

Accuracy



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

Source of images: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Metrics for classification

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
- 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

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

Problem with accuracy

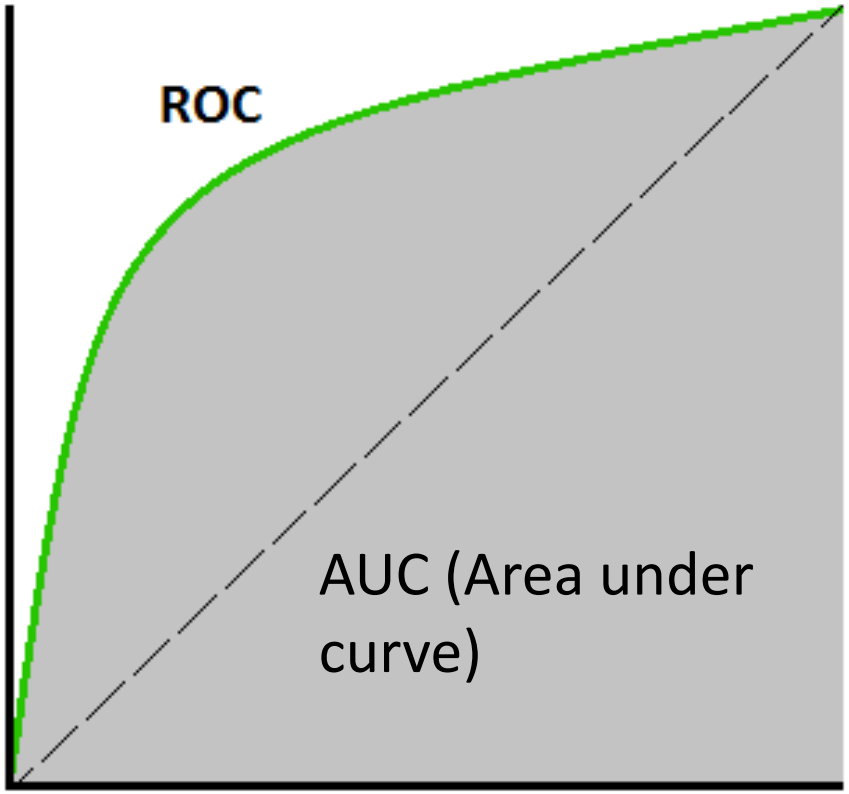
- **Do not use when the data is unbalanced** (the number of cases in each class is not the same)
  - Example :
    - 990 non-cancer and 10 cancer
    - Dummy classifier: nobody has cancer
    - Accuracy: 99%

- Solution: **balanced accuracy**

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

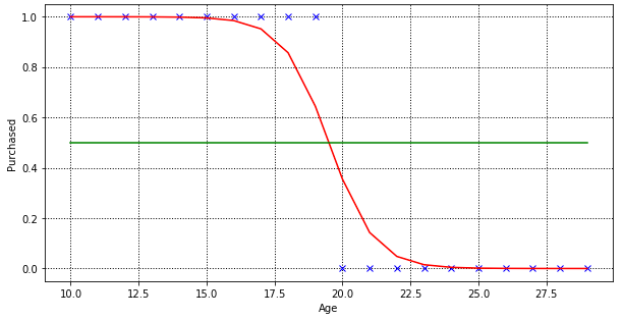
Metrics for classification

ROC curve



The curve is plotted by varying the threshold on the activation function

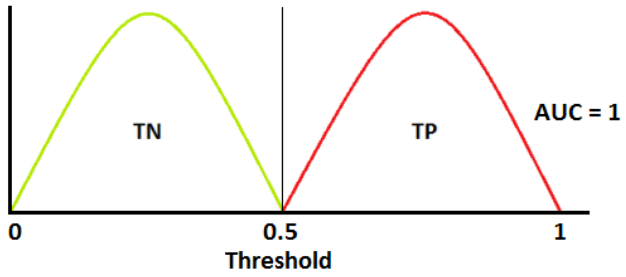
$$g(z) = \frac{1}{1+e^{-z}}$$

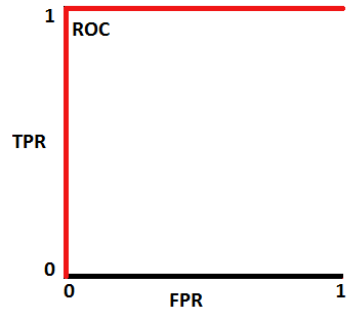


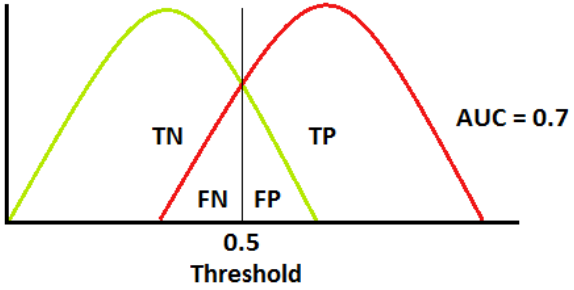
Source of images: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

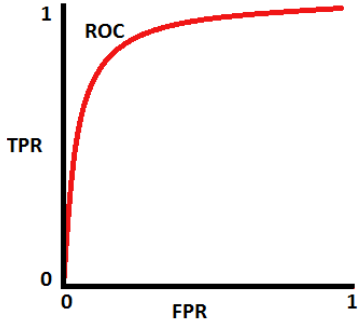
Metrics for classification

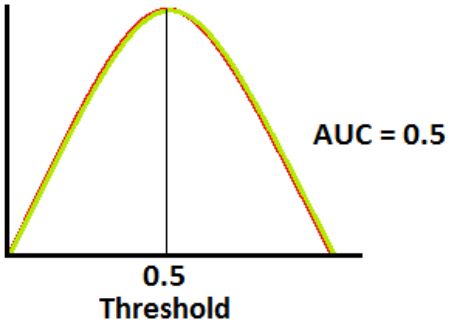
ROC curve

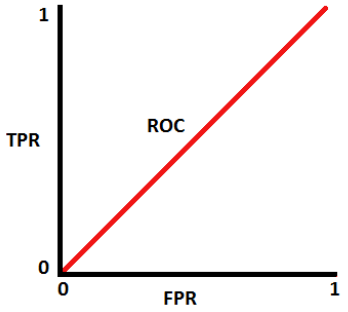












https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Metrics for classification

Conclusion

- Different metrics and tools exist
- They provide complementary information
- Emphasis can be put on specific metrics depending on the problem at hand

Other metrics

There are many other metrics depending on the task

- Image segmentation (seen in Course 4 – Segmentation)
  - Dice similarity coefficient (DSC)
  - Jaccard coefficient
  - Volume similarity
  - Surface distance measures

- Image synthesis (will be seen in Course 7 – Generative models)
  - Root Mean Square Error
  - Peak Signal to Noise Ratio
  - Structural Similarity

Part 6 - Validation
6.2 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)`

Samples

Training set

Validation set

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)

Samples

Training set

Validation set

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

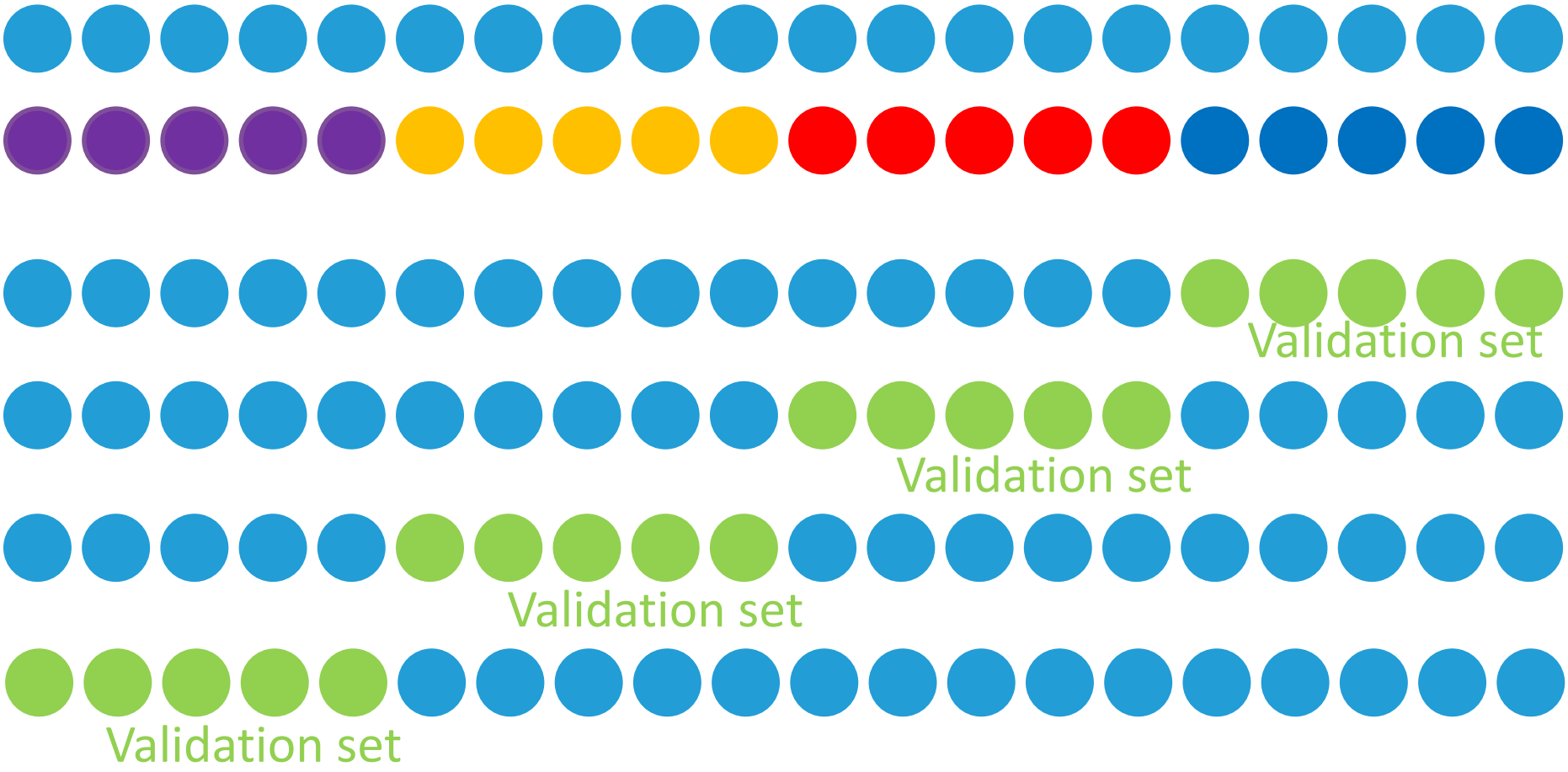
Validation strategies

k-fold cross validation

Here k=4

sklearn.model_selection.KFold

sklearn.model_selection.StratifiedKFold



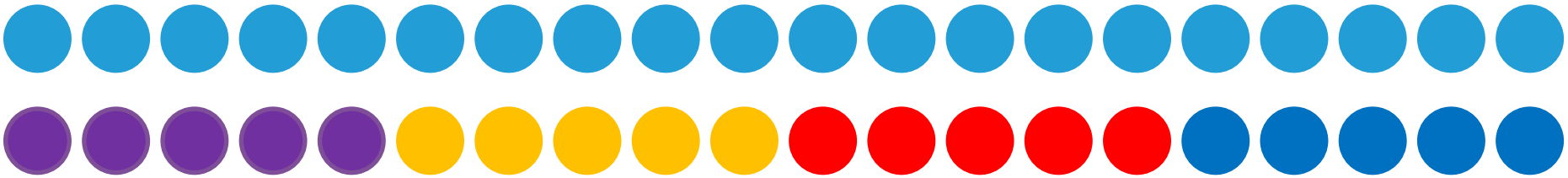
Validation strategies

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

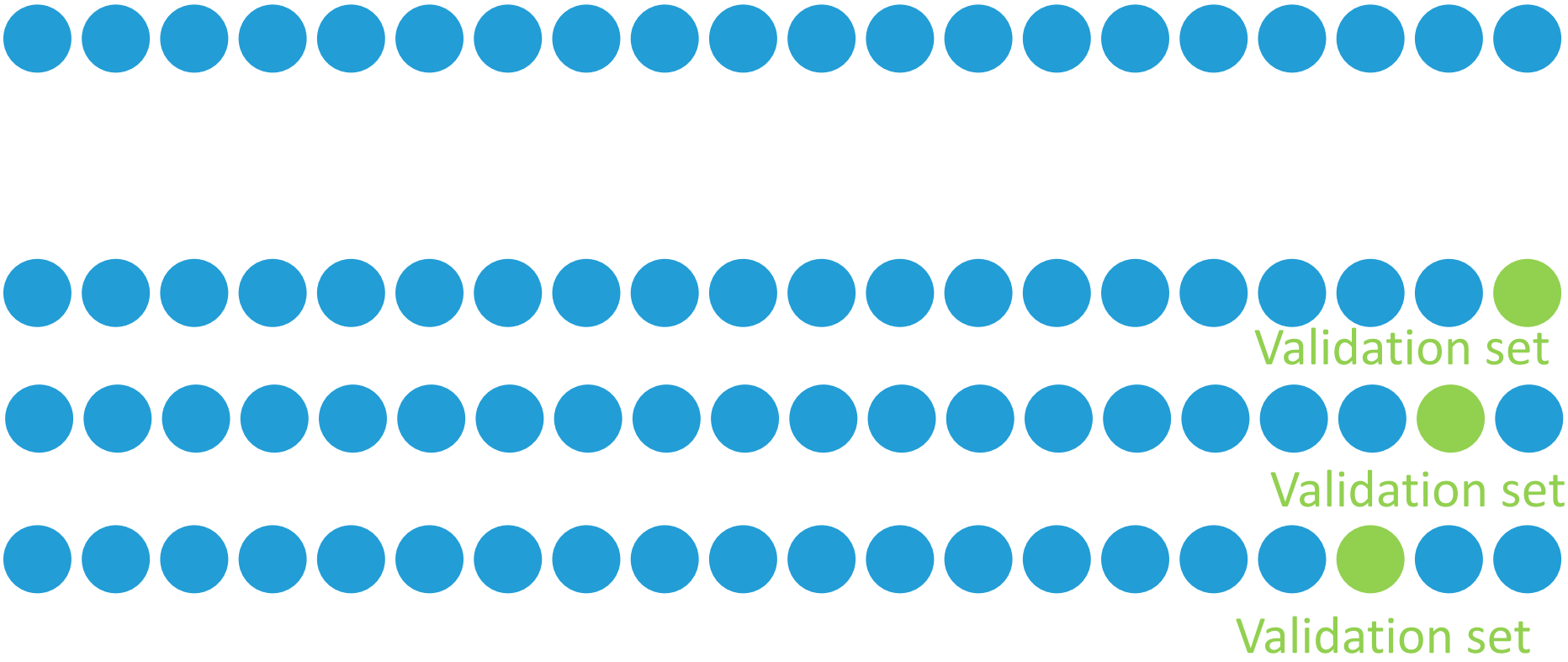
Validation strategies

sklearn.model_selection.LeaveOneOut

Leave-one-out cross validation

Special case of k-fold with k=n

Samples



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

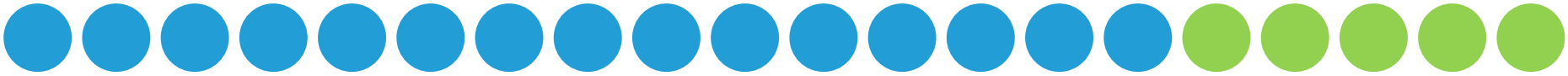
Validation strategies

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

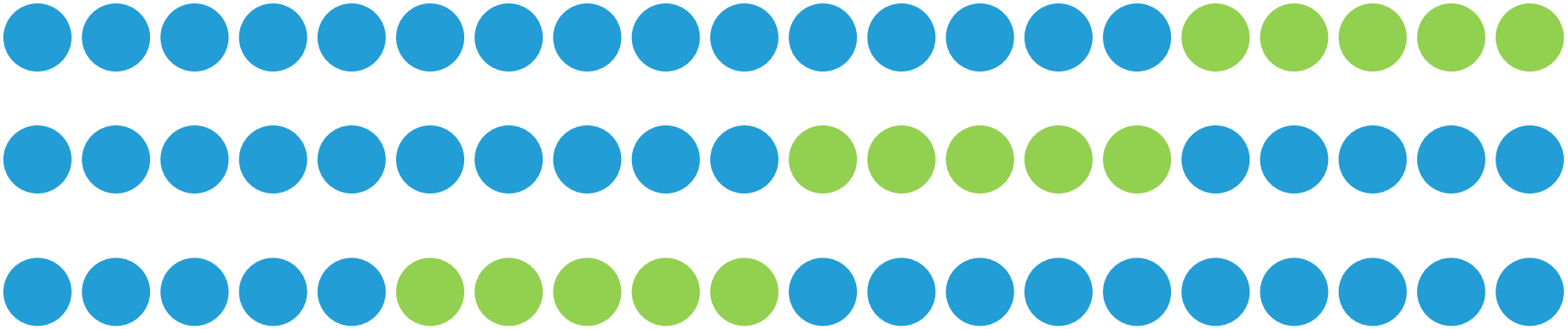
Is this enough?

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

Bad practices

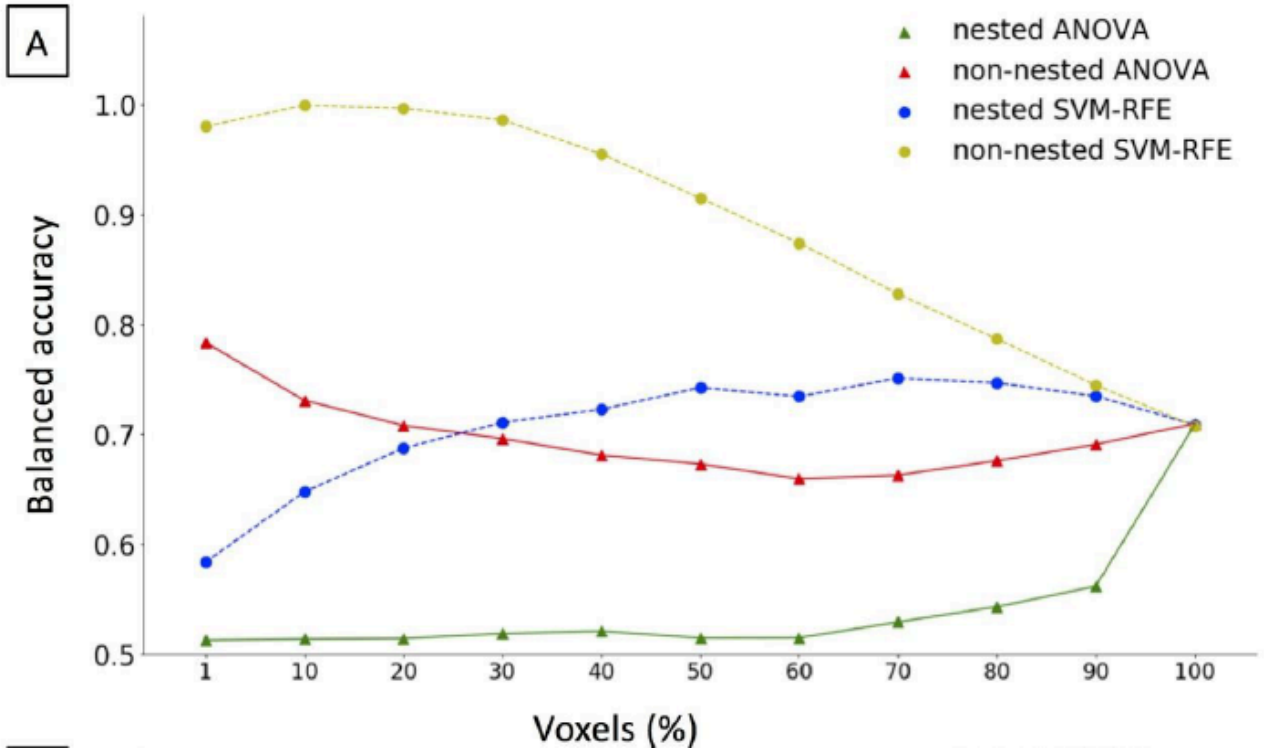
Use all samples for **feature selection**

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



Bad practices

Use all samples for feature selection

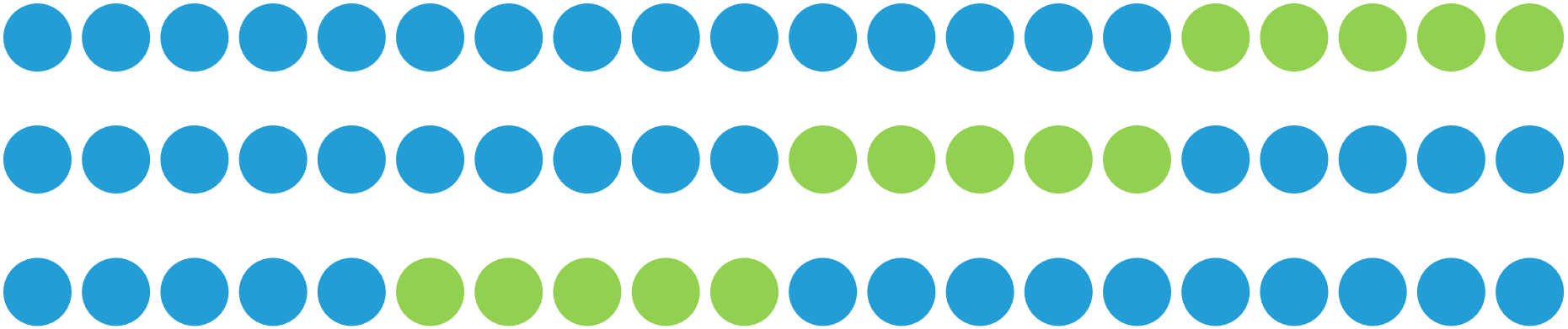


Wen et al, 2018

Bad practices

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

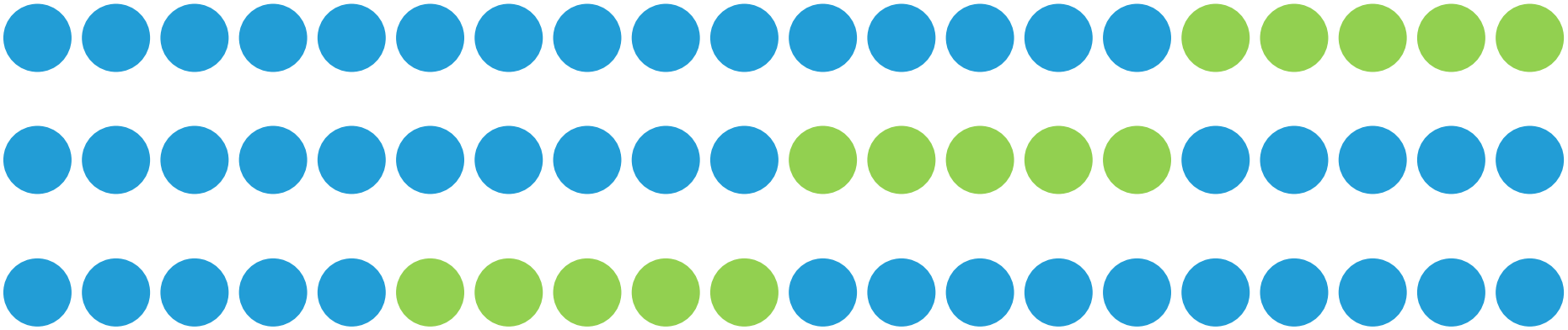
Then cross-validate the model using the reduced features as input



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

Bad practices

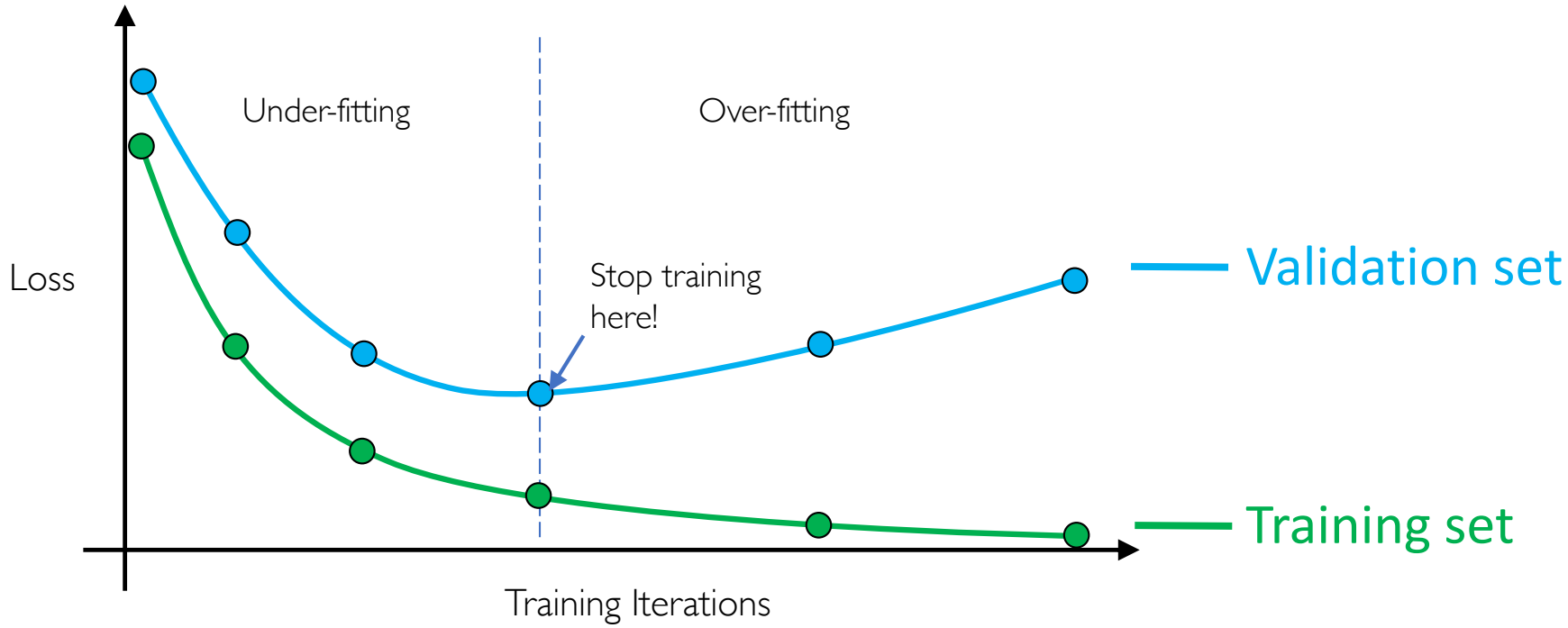
Do multiple runs of cross-validation to select the best hyperparameter



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

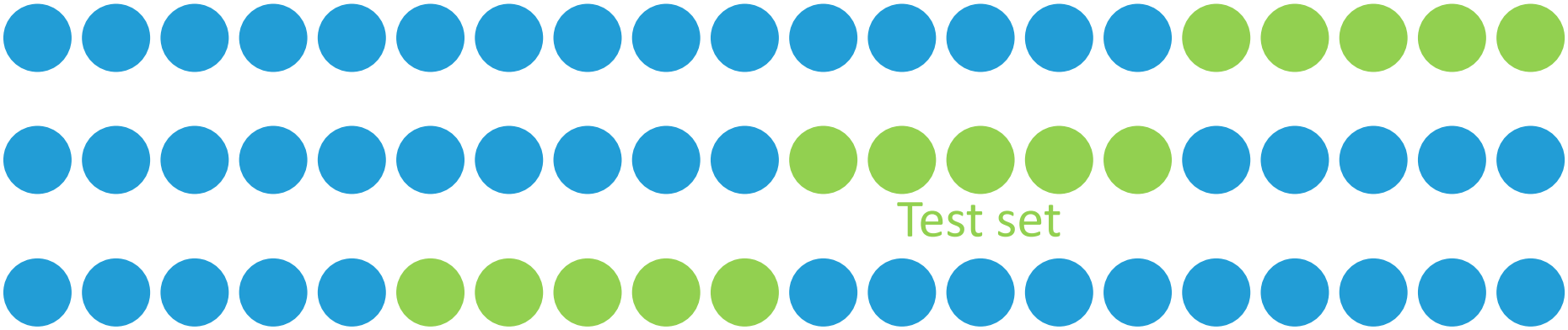
Bad practices

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, 2019)

What should we do?

Training, validation and test sets

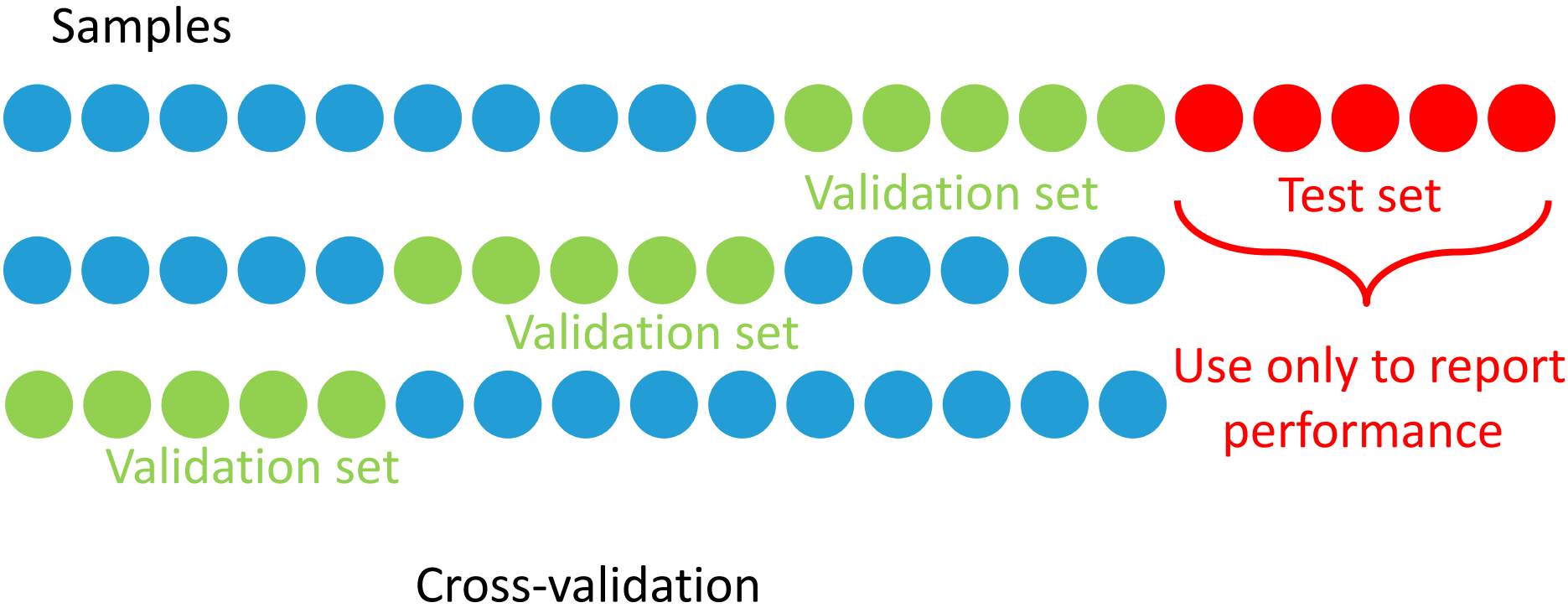
Use to test different models, to choose when to stop training…

Often through cross-validation

Use only to report performance

What should we do?

Training, validation and test sets



Cross-validation

Standard 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!

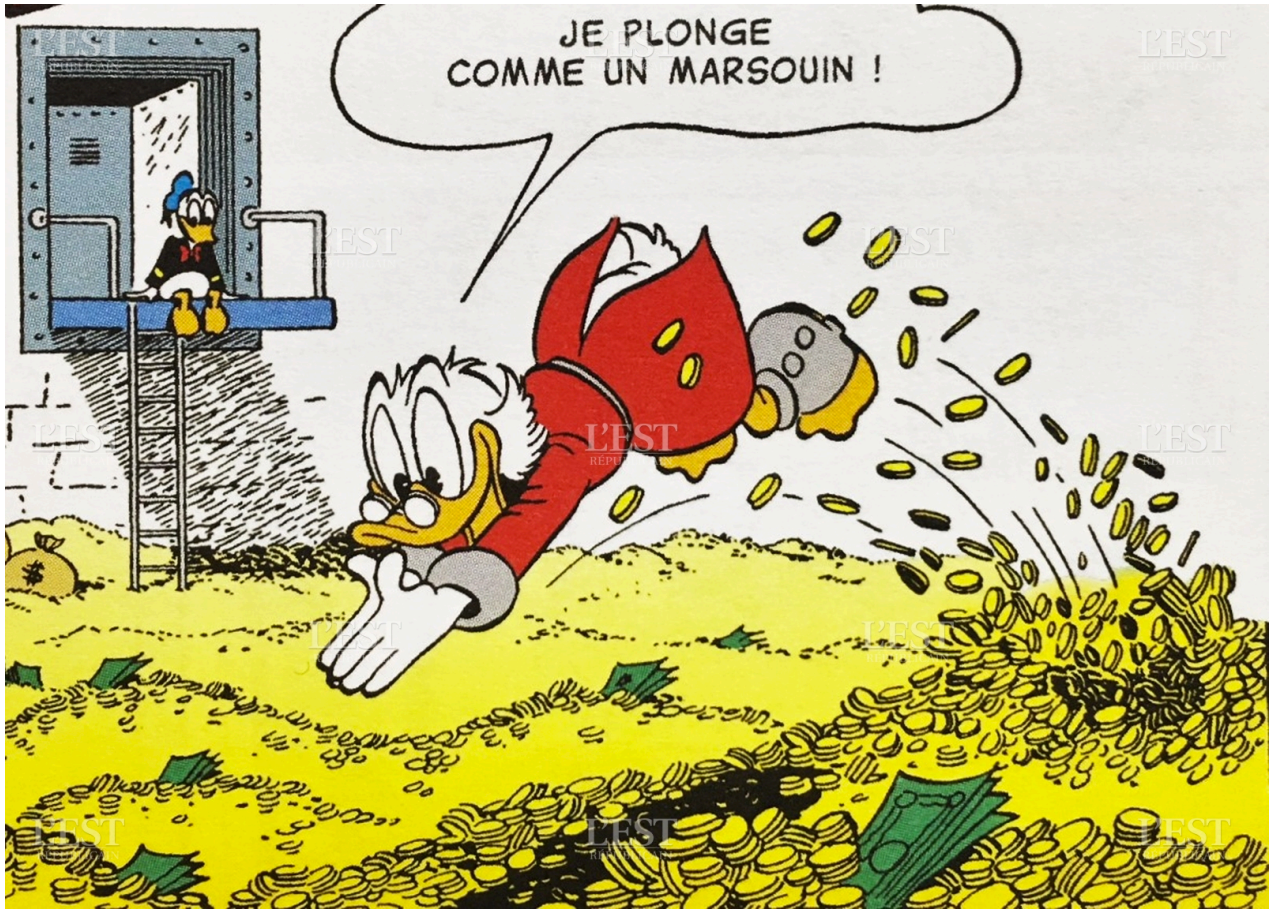
The test set

When should I separate the test set?

Before starting the work

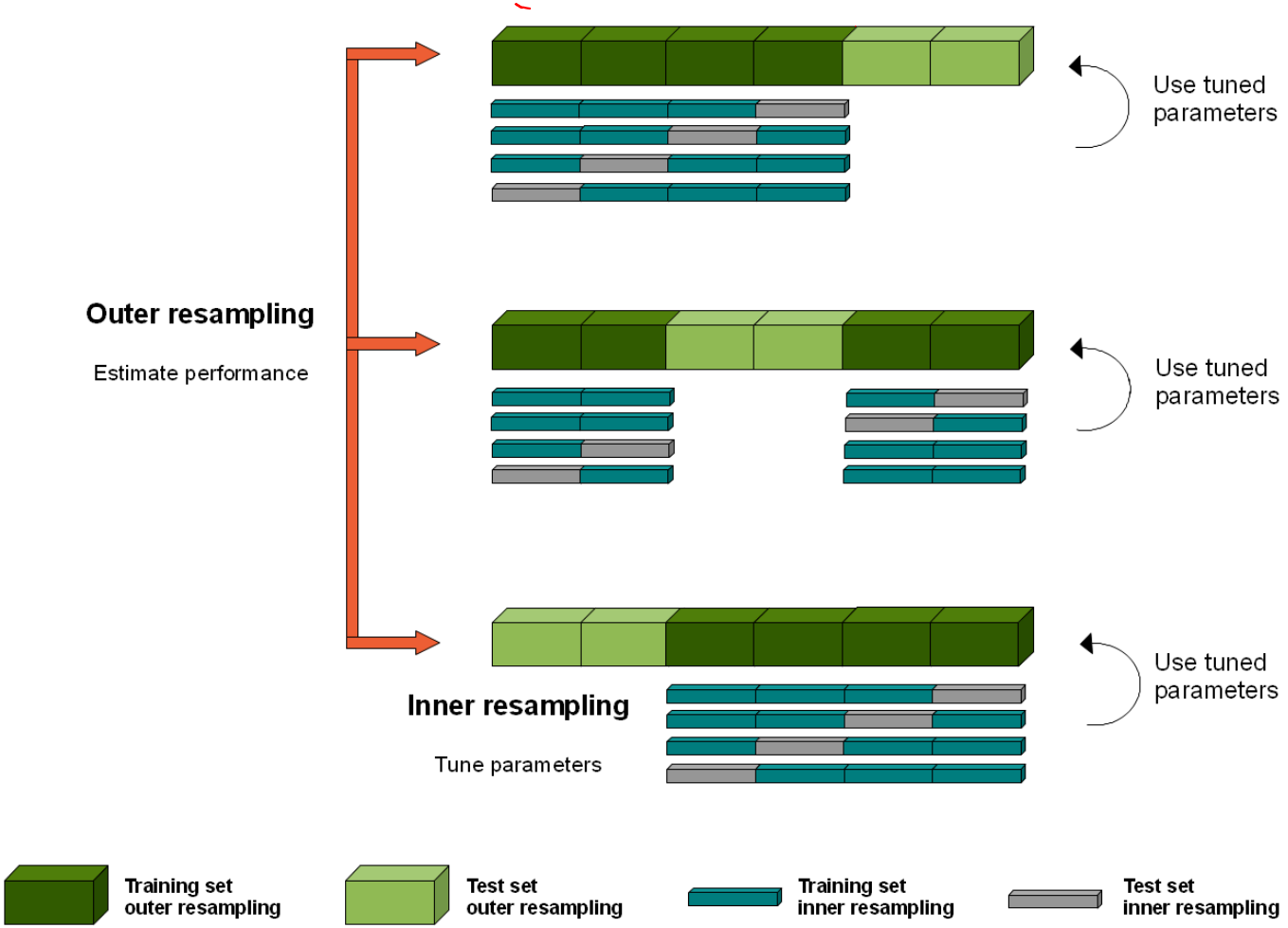
The test set

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



Other solution

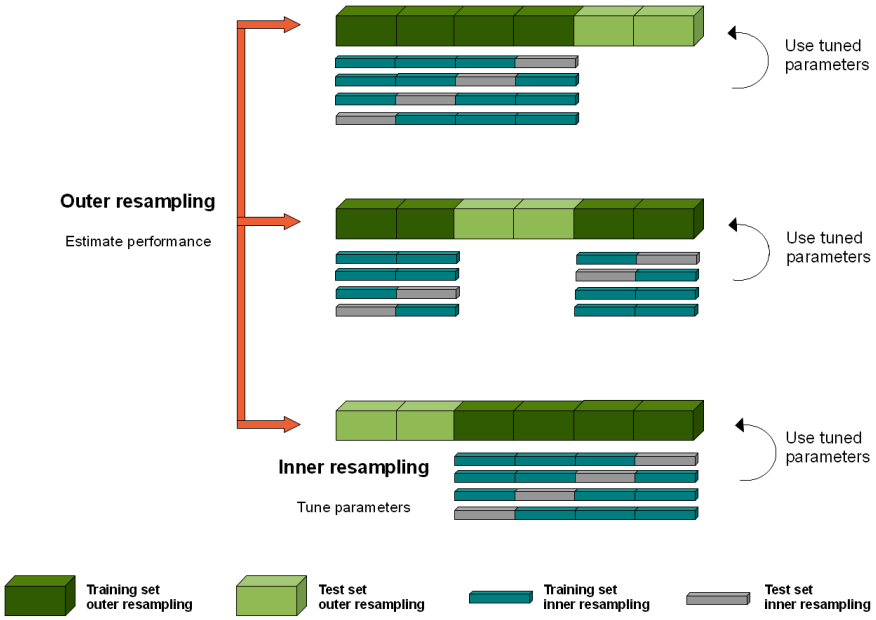
Nested cross-validation



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

Other solution

Nested cross-validation



Computationally expensive

Can be feasible with models with fast training

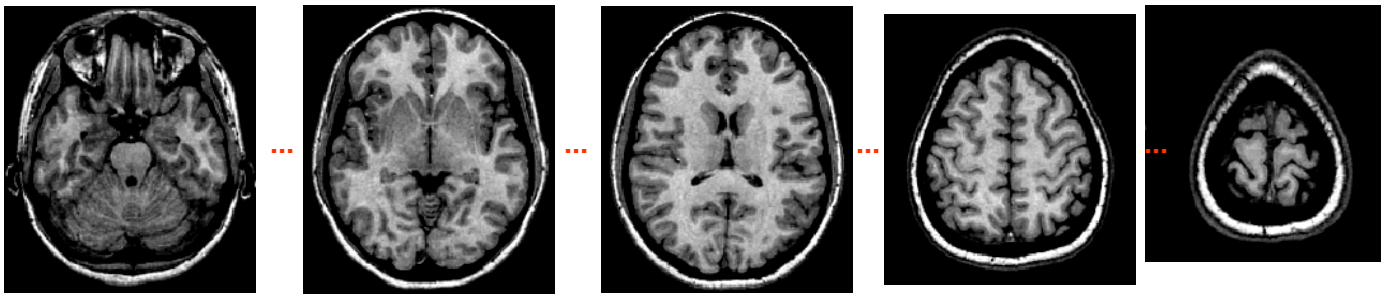
https://i.stack.imgur.com/vh1sZ.png

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

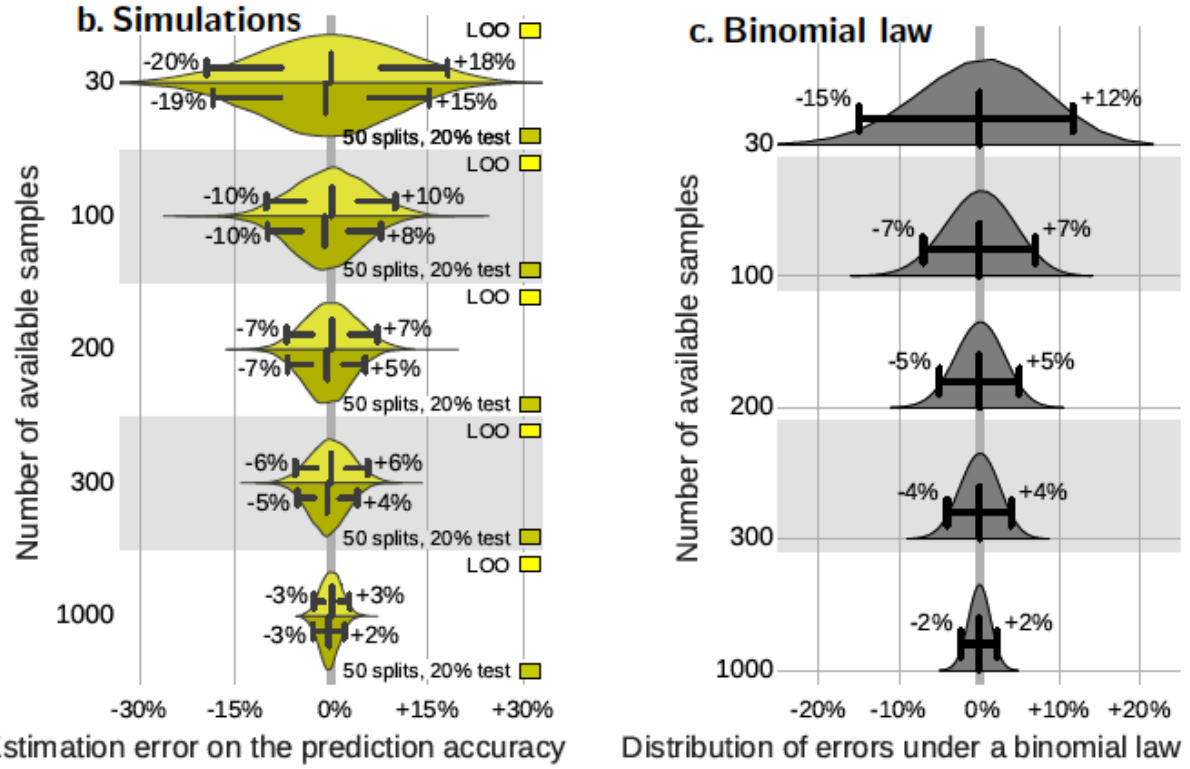
One can use the following functions to do the split at the patient level (slices or visits will be grouped into patients)

sklearn.model_selection.LeaveOneGroupOut

sklearn.model_selection.LeavePGroupsOut

sklearn.model_selection.GroupKFold

How about sample size?



b. Distribution of errors between the prediction accuracy as assessed via cross-validation and as measured on 10 000 new data points (for **data simulated** such that the classifier acheives 75% accuracy)

c. Distribution of errors as given by a **binomial law**: difference between the observed prediction error and the population value of the error (p = 75%)

From (Varoquaux, NeuroImage, 2017)

How about significance testing?

One often wants to compare different approaches

	Alzheimer vs controls	Stable vs progressive MCI
3D Whole-brain CNN	84%	73%
3D Hippocampus CNN	86%	73%
2D ResNet	76%	N/A
Linear SVM	86%	72%

Here the 2D ResNet is likely less good than the three others.
But how about the 2 other approaches?

How about significance testing?

One would like to performance hypothesis testing

	Alzheimer vs controls	Stable vs progressive MCI
3D Whole-brain CNN	84%	72%
3D Hippocampus CNN	86%	73%
2D ResNet	76%	N/A
Linear SVM	86%	72%

P<0.05?

How about significance testing?

If you are using cross-validation, this is difficult to do….

Journal of Machine Learning Research 5 (2004) 1089–1105 Submitted 05/03; Revised 9/03; Published 9/04

No Unbiased Estimator of the Variance of K-Fold Cross-Validation

Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc, H3C 3J7, Canada

Yves Grandvalet YVES.GRANDVALET@UTC.FR
Heudiasyc, UMR CNRS 6599
Université de Technologie de Compiègne, France

You can report the variance over folds (it is better than nothing) but always add a note that the estimator is biased

Don’t trust a t-test applied to cross-validation results

How about significance testing?

If you are using cross-validation, this is difficult to do….

Machine Learning, 52, 239–281, 2003

Inference for the Generalization Error

CLAUDE NADEAU jcnadeau@altavista.net
Health Canada, AL0900B1, Ottawa ON, Canada K1A 0L2

YOSHUA BENGIO Yoshua.Bengio@umontreal.ca
CIRANO and Dept. IRO, Université de Montréal, C.P. 6128 Succ. Centre-Ville, Montréal, Quebec, Canada H3C 3J7

Editor: Lisa Hellerstein

There are some corrections that can be made to the variance and to the statistical tests. But their validity is not universal

How about significance testing?

If you are using separate test set

Some valid tests exist such as the McNemar test for comparing classifiers

Conclusion

- **Metrics**
  - **Use the appropriate metrics**
    - E.g, never use accuracy with unbalanced datasets
  - Several metrics are necessary to get a comprehensive view of the performance
    - Do not only report accuracy!

- **Population:**
  - Always report a socio-demographic table

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

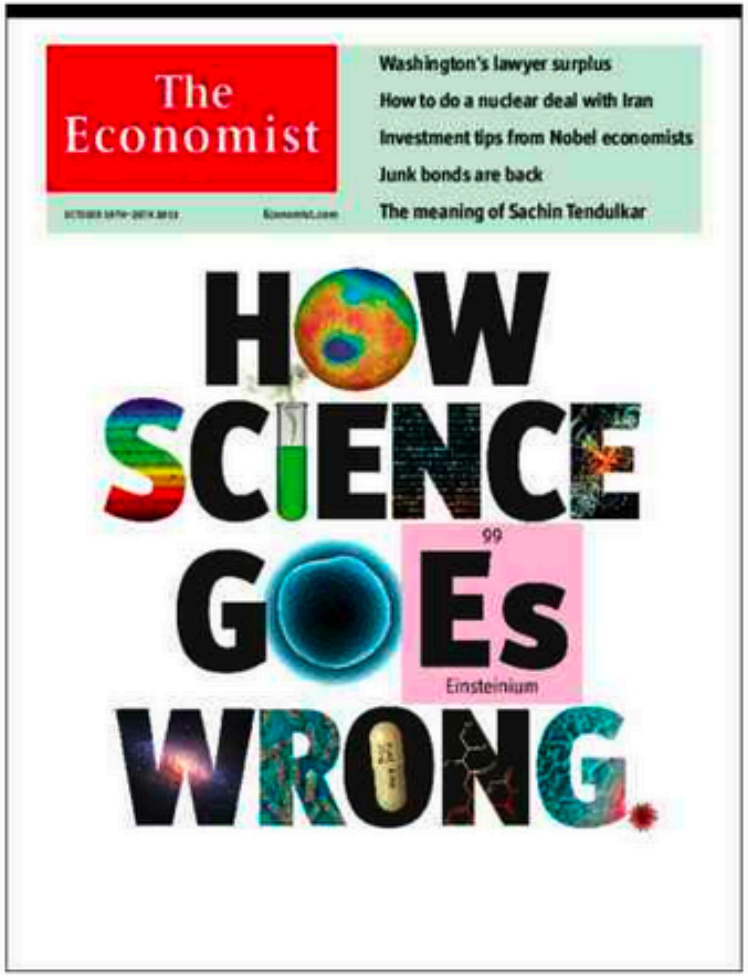
Conclusion

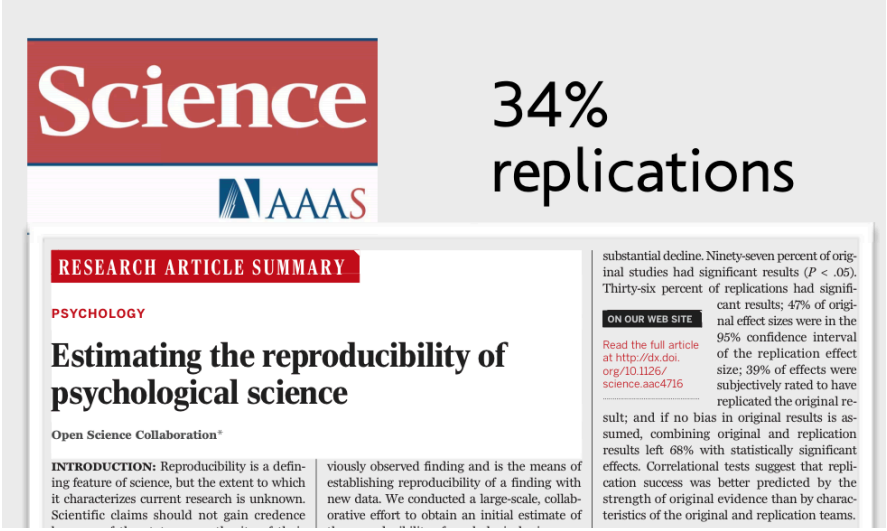
- **Validation**
  - Beware of data leakage
  - If you do deep learning, always use a separate test set, in addition to training and validation sets
    - Leave the test set untouched until the end

- **Be aware of the limitations of variance estimation for CV and of statistical testing**

Part 6 - Validation
6.3 Reproducibility

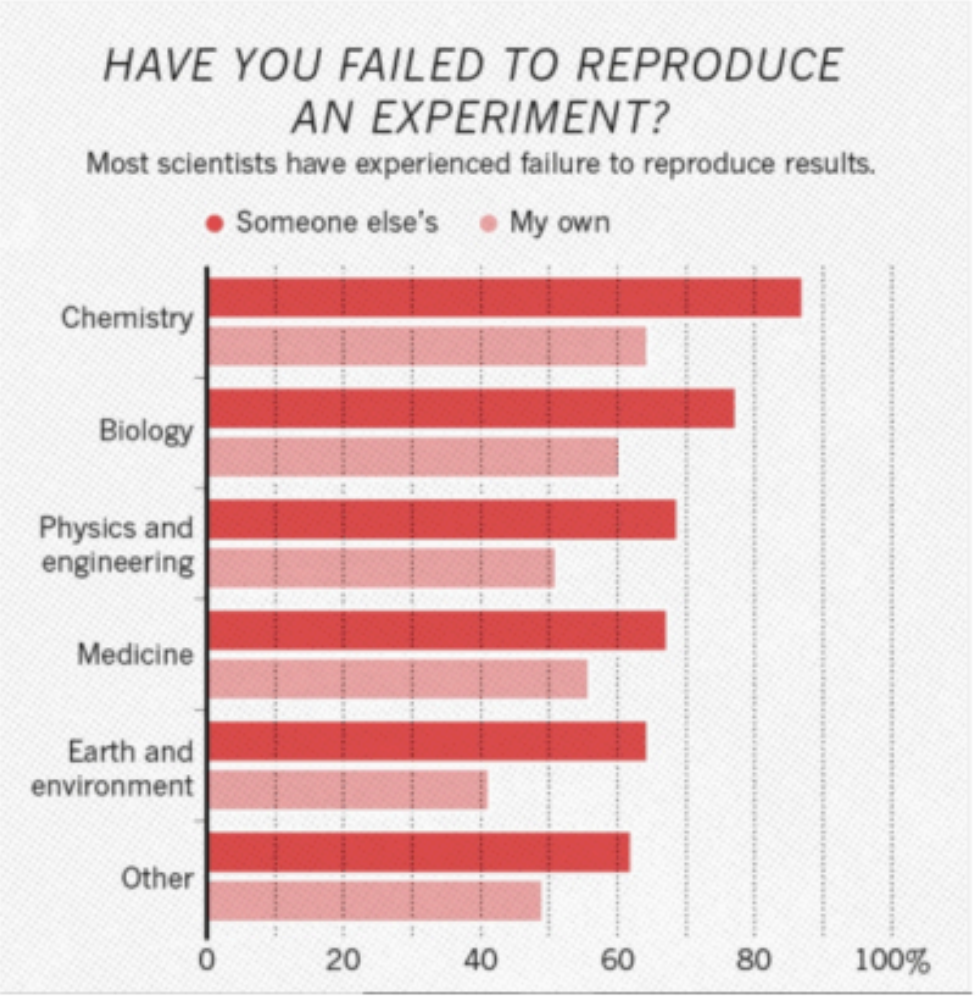
Replication crisis





AI in medicine is facing it too!

Reproducibility



Reproducibility

- What is reproducibility?
  - **The ability to reproduce scientific results**

- More specifically
  - **Technical Replicability:** Can results be replicated under **technically identical conditions**?
  - **Statistical Replicability:** Can results be replicated under **statistically identical conditions**?
  - **Conceptual Replicability:** Can results be replicated under **conceptually identical conditions**?

Based on: McDermott et al, Reproducibility in Machine Learning for Health, RML@ICLR Workshop, 2019

Reproducibility

- **Technical Replicability:**
  - Obtain exactly the same results as reported in the paper
  - Requires
    - Dataset release
      - Can raise some difficulties in healthcare
      - When using a public dataset, specify which subpart of the dataset was used
    - Code release
      - Including possible random number generators
      - Including any possible preprocessing pipelines
    - Release of trained models

Based on: McDermott et al, Reproducibility in Machine Learning for Health, RML@ICLR Workshop, 2019

Reproducibility

- **Statistical Replicability:**
  - Obtain same results under statistically equivalent conditions
    - For instance on random subsets of the data used
    - Or on other datasets with the same characteristics
  - Requires:
    - Reporting variance
      - Can be biased in some settings
    - Large testing datasets
    - No cherry picking of the samples

Based on: McDermott et al, Reproducibility in Machine Learning for Health, RML@ICLR Workshop, 2019

Reproducibility

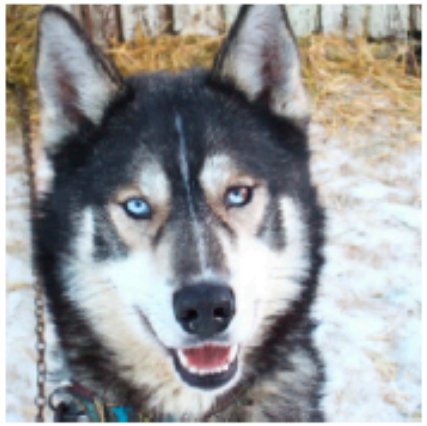
- **Conceptual Replicability:**
  - Obtain same results under conceptually equivalent conditions
    - For instance another set of patients with the same disease but from a different hospital
  - Gets closer to real-life performance
  - Requires:
    - Testing on multiple datasets acquired in different conditions

Based on: McDermott et al, Reproducibility in Machine Learning for Health, RML@ICLR Workshop, 2019

Part 6 - Validation
6.4 Interpretation

Why intrepretation?

- **Build confidence in neural networks systems**
- **Analyse failure modes**



(a) Husky classified as wolf

(b) Explanation

Training
- Wolves with snow
- Huskies without snow

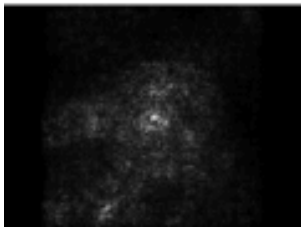
Testing
Husky with snowy background classified as wolf

Source: Elina Thibeau-Sutre

Interpretation

Two main classes of methods

original image

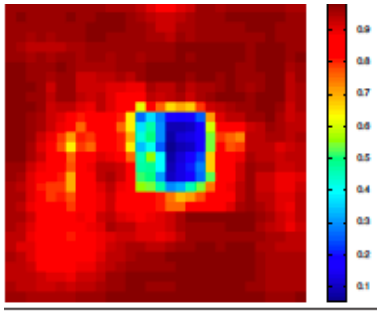


saliency map

Gradients visualization or "saliency maps"

[Simonyan et al, 2013]

occluded image



classifier output probability depending on occlusion position

Occlusion heatmaps

[Zeiler and Fergus, 2016]

→ Difficult interpretation…

→Zones are studied independently

Saliency Maps

Concept:

Backpropagation of gradients in the network

Pros:

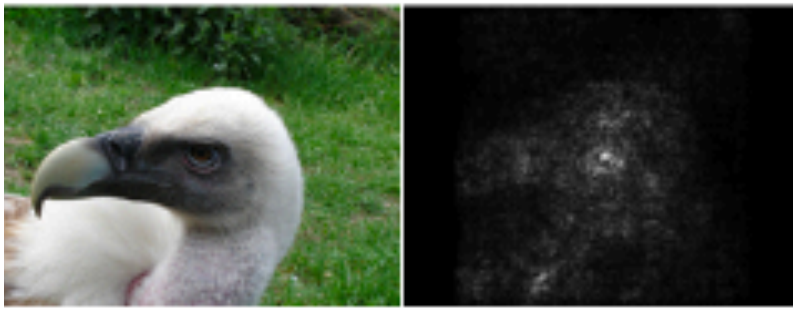
- Simple

Cons:

- Noisy maps, difficult to interpret
- Cannot interpret the different parts of the network

Original image

Saliency Map



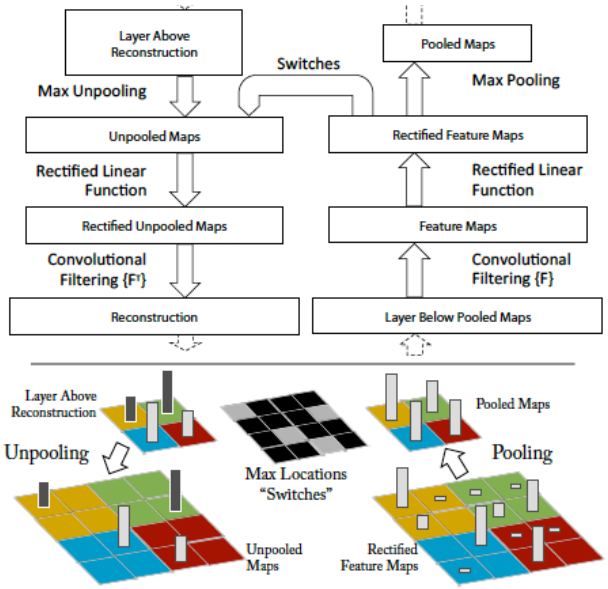
[Simonyan et al, 2013]

Source: Elina Thibeau-Sutre

Deconvolution

Concept

Inverting the transformations of the CNN layers



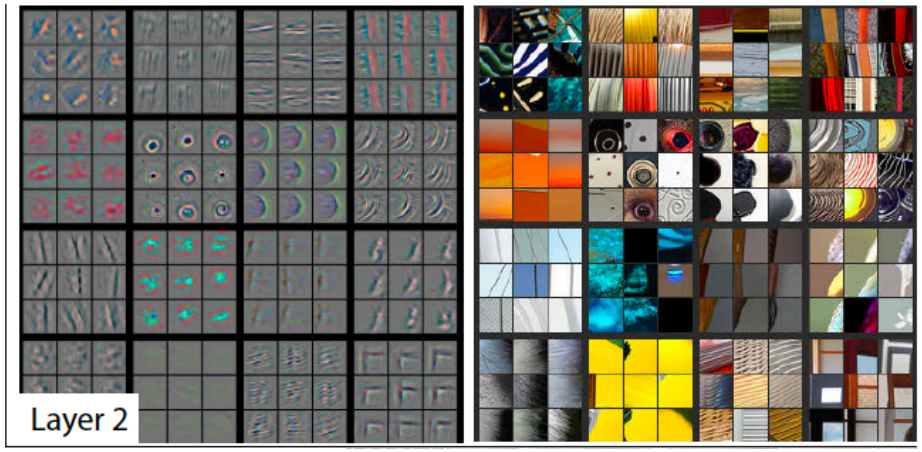
Source: Elina Thibeau-Sutre

Pros:

- Can be used to interpret different parts of the network

Cons:

- Does not always select the relevant object



[Zeiler and Fergus, 2014]

Guided Back-propagation

Concept:

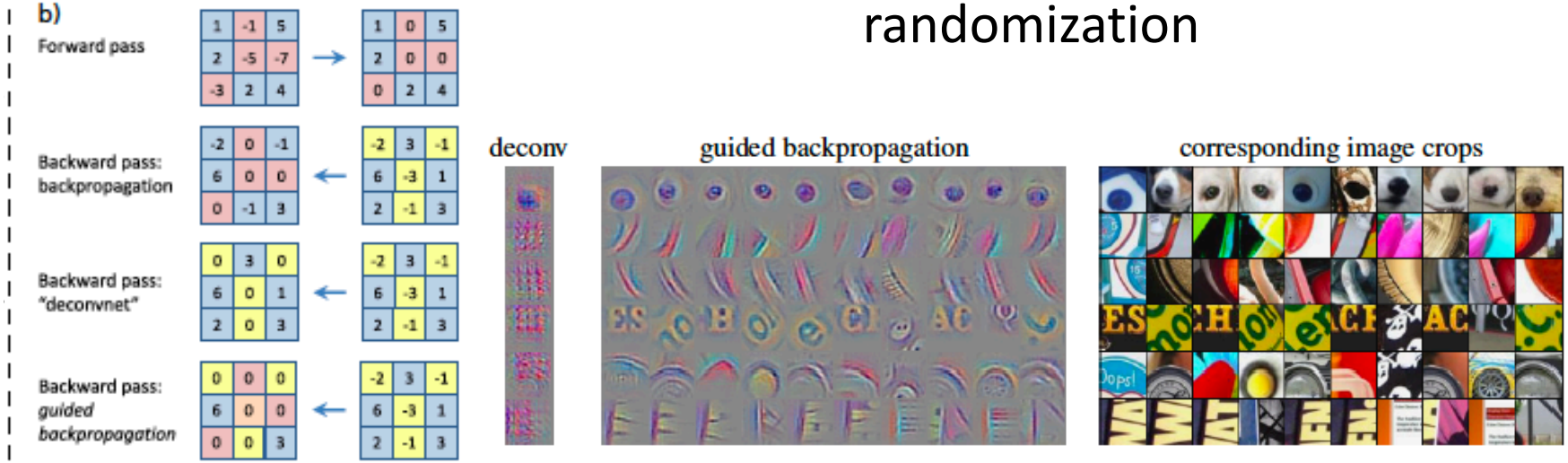
Mix of deconvolution and saliency maps

Pros:

- Can be used to interpret different parts of the network
- Selects the relevant object

Cons:

- Does not change with randomization



Source: Elina Thibeau-Sutre

[Springenberg et al, 2015]

Grad-CAM

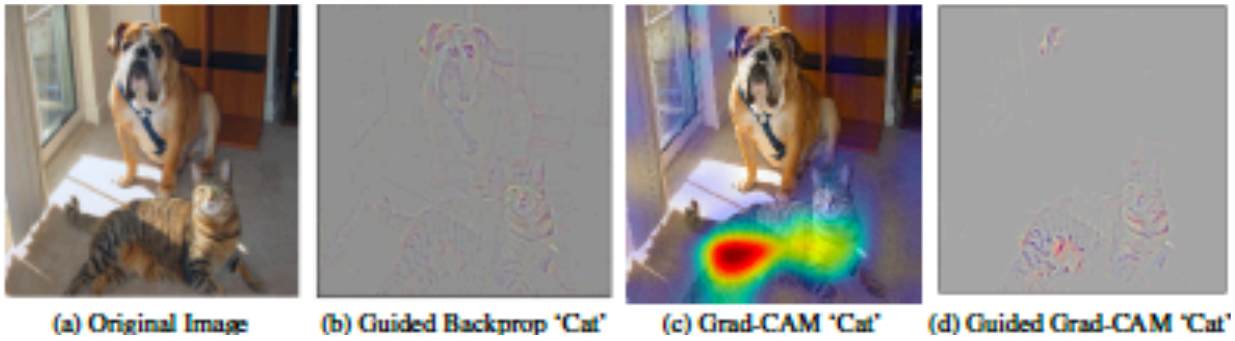
Concept:

Linear combination of feature maps weighted by the sum of the gradients

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

We want to preserve only the positive influence



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat'

[Selvaraju et al, 2017]

Source: Elina Thibeau-Sutre

Occlusion

Concept:

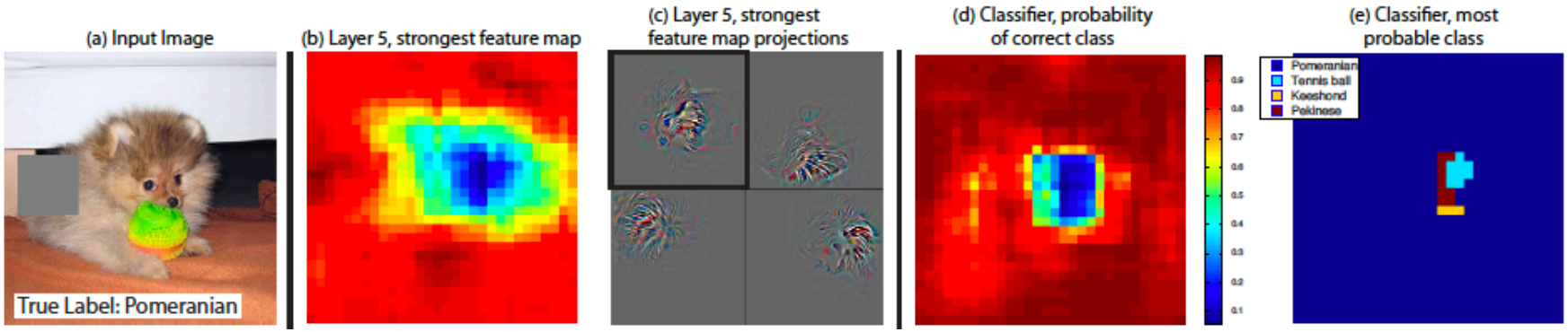
Occluding part of the image with a grey square

Pros:

- Simple
- Different concept

Cons:

- Poor resolution (depends on size of occlusion)
- What about grey objects?



[Zeiler and Fergus, 2016]

Source: Elina Thibeau-Sutre

Greedy Occlusion

Concept:

Occluding part of the image with an adaptive mask

Different types of occlusion:

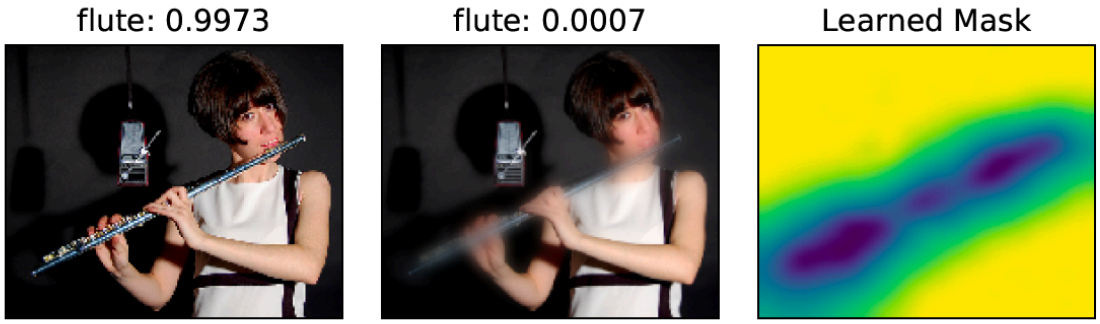
- Constant value
- Noise
- Blurring

Pros:

- Adaptive mask size

Cons:

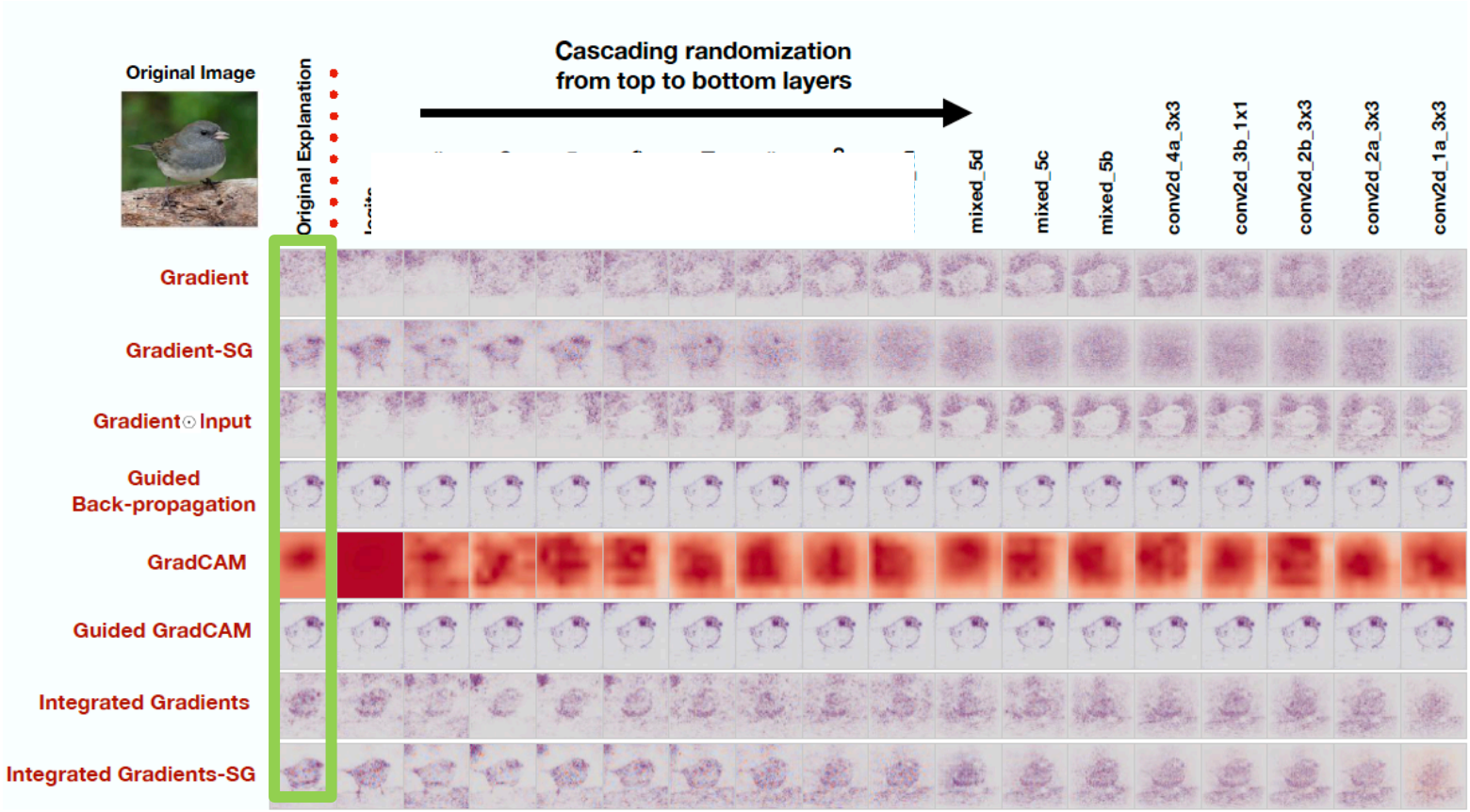
- Hyperparameters
- May create artifacts



[Fong and Vedaldi, 2017]

Source: Elina Thibeau-Sutre

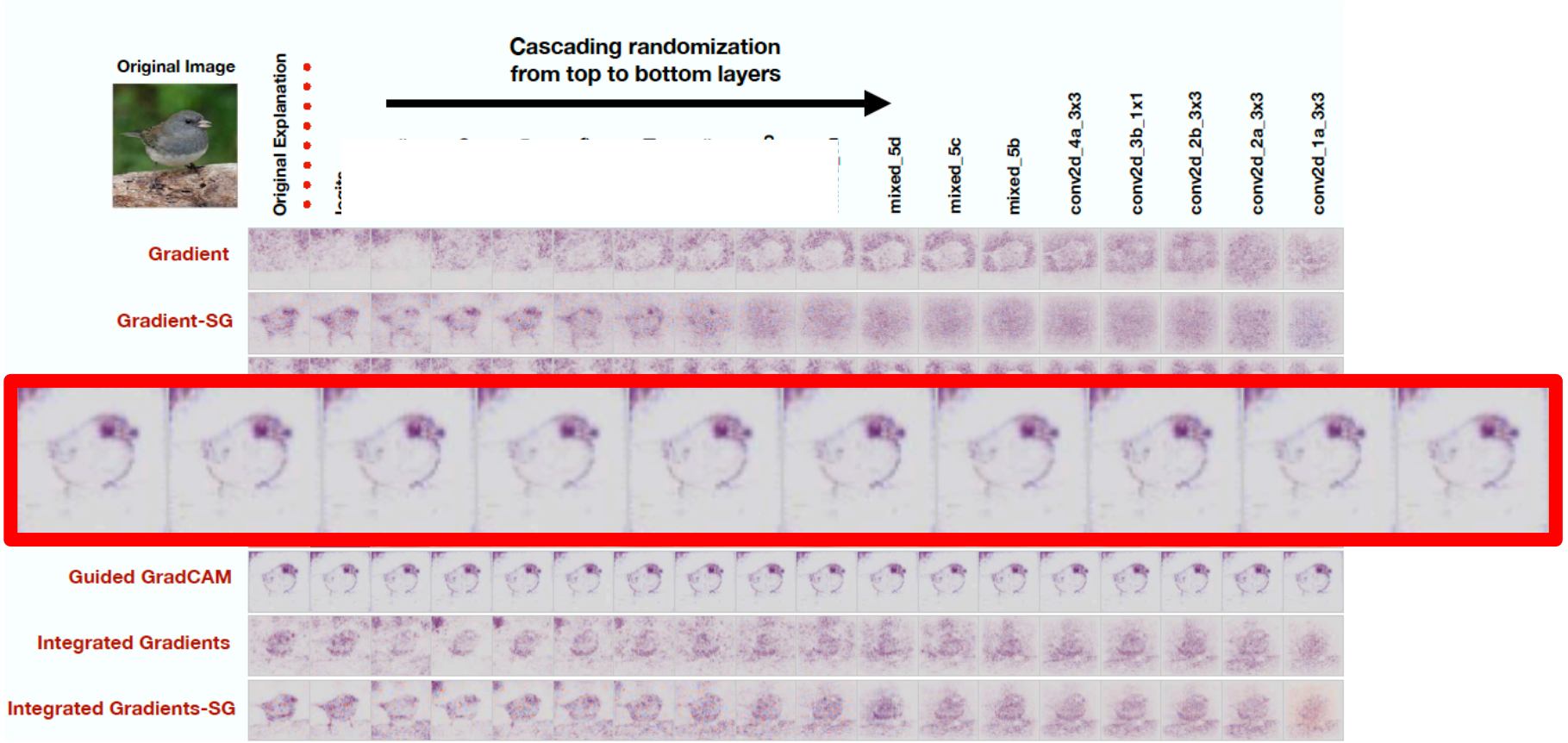
Robustness of saliency maps



Source: Elina Thibeau-Sutre

[Adebayo et al, 2018]

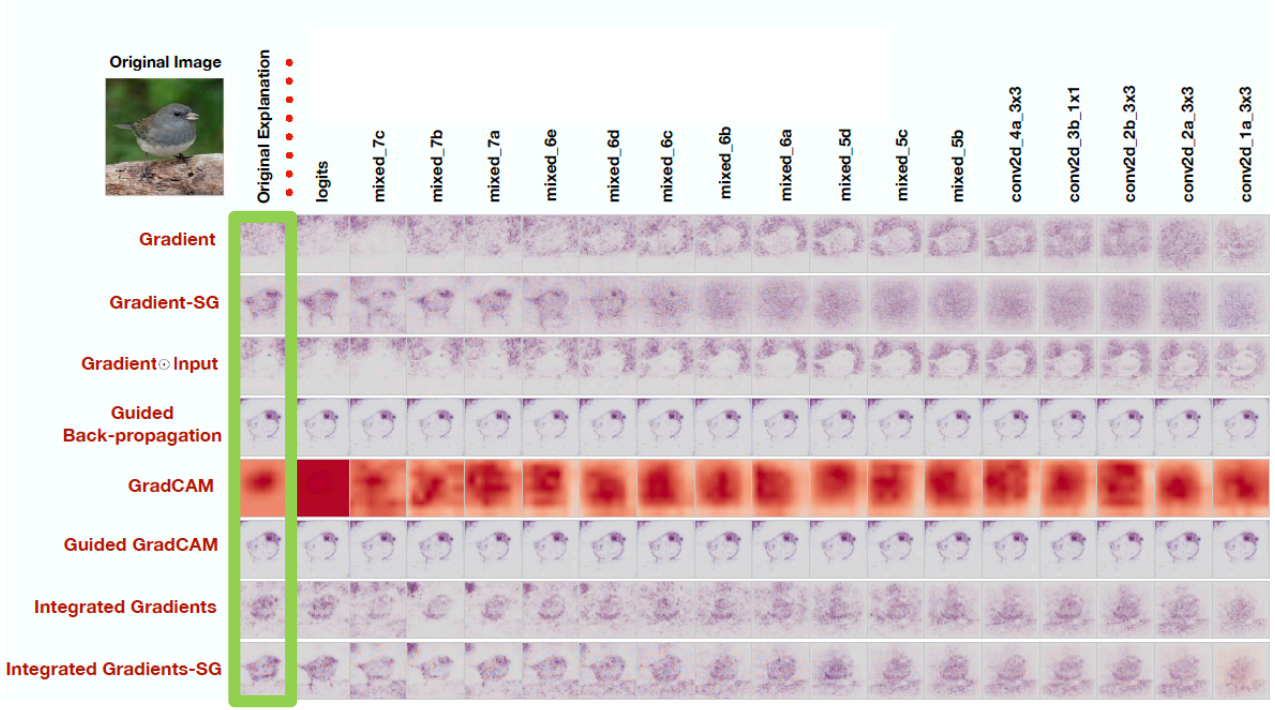
Robustness of saliency maps



Source: Elina Thibeau-Sutre

[Adebayo et al, 2018]

Robustness of saliency maps

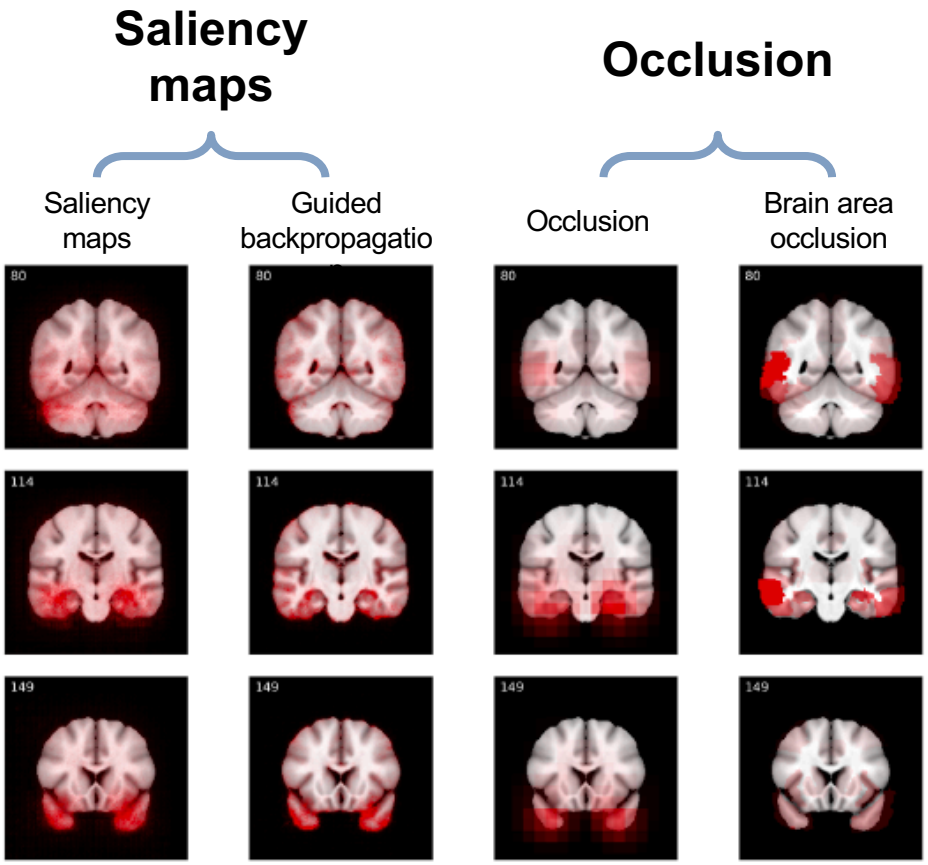


Independent from learnt parameters

[Adebayo et al, 2018]

Source: Elina Thibeau-Sutre

Application to MRI classification in AD



Binary classification
- Alzheimer's disease patients
- Cognitively normal participants

[Rieke et al, 2018]

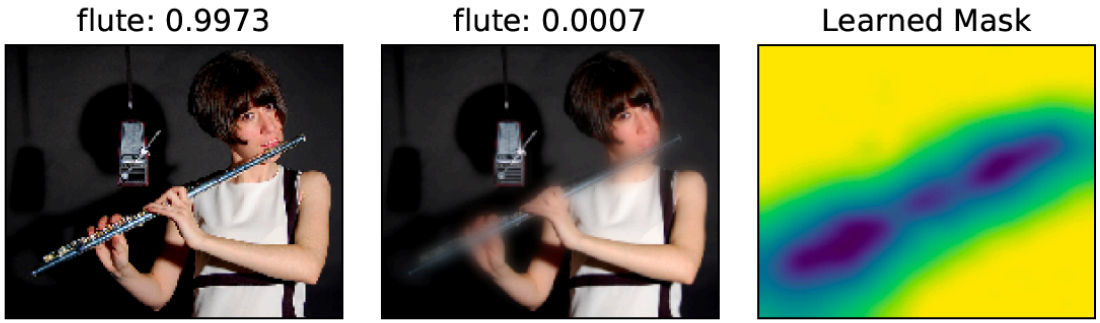
Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Interpretability method**

The image X'_m masked by m at voxel u is defined as:

$$X'_m(\mathrm{u}) = \mathrm{m(u)X(u)} + \big(1 - \mathrm{m(u)}\big)\mathrm{X_{perturbed}(u)}$$



Meaningful perturbation

[Fong and Vedaldi, 2017]

Goal: mask m minimizing the score of the CNN f on a set of occluded images X'_m by covering a minimal amount of pixels in connected parts.

$$m^* = \underset{m}{\operatorname{argmin}}\, f(X'_m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \|\nabla m\|_{\beta_2}^{\beta_2}$$

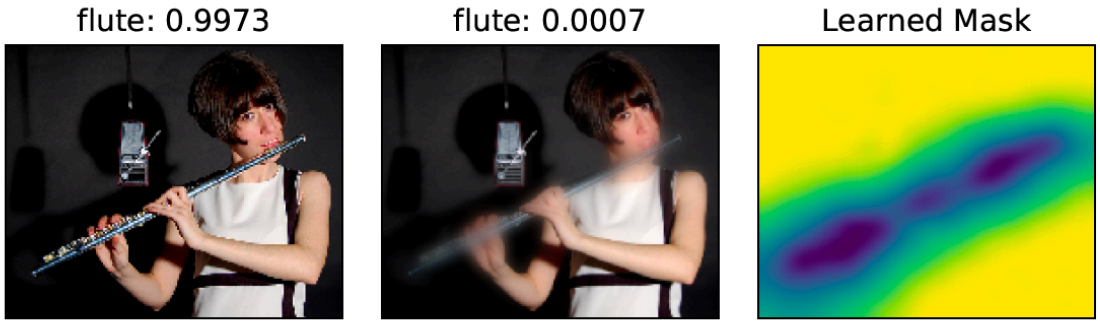
Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Interpretability method**

The image X'_m masked by m at voxel u is defined as:

$$X'_m(\mathrm{u}) = \mathrm{m(u)X(u)} + \big(1 - \mathrm{m(u)}\big)\mathrm{X_{perturbed}(u)}$$



Meaningful perturbation

[Fong and Vedaldi, 2017]

Goal: mask m minimizing the score of the CNN f on a set of occluded images X'_m by covering a minimal amount of pixels in connected parts.

$$m^* = \underset{m}{\operatorname{argmin}}\, f(X'_m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \|\nabla m\|_{\beta_2}^{\beta_2}$$

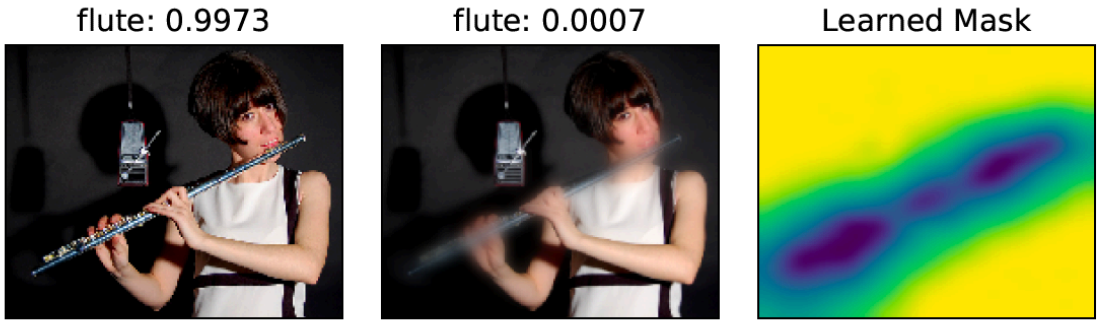
Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Interpretability method**

The image X'_m masked by m at voxel u is defined as:

$$X'_m(\mathrm{u}) = \mathrm{m(u)X(u)} + \big(1 - \mathrm{m(u)}\big)\mathrm{X_{perturbed}(u)}$$



Meaningful perturbation

[Fong and Vedaldi, 2017]

Goal: mask m minimizing the score of the CNN f on a set of occluded images X'_m by covering a minimal amount of pixels in connected parts.

$$m^* = \underset{m}{\operatorname{argmin}}\, f(X'_m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \|\nabla m\|_{\beta_2}^{\beta_2}$$

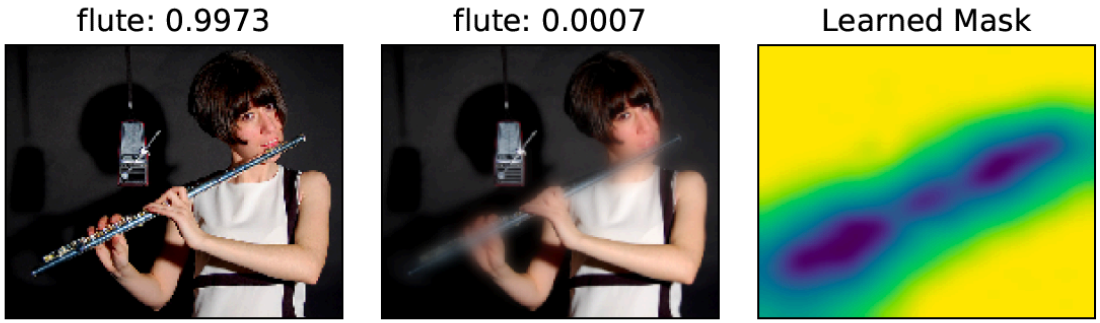
Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Interpretability method**

The image X'_m masked by m at voxel u is defined as:

$$X'_m(\mathrm{u}) = \mathrm{m(u)X(u)} + \big(1 - \mathrm{m(u)}\big)\mathrm{X_{perturbed}(u)}$$



Meaningful perturbation

[Fong and Vedaldi, 2017]

Goal: mask m minimizing the score of the CNN f on a set of occluded images X'_m by covering a minimal amount of pixels in connected parts.

$$m^* = \underset{m}{\operatorname{argmin}}\, f(X'_m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \|\nabla m\|_{\beta_2}^{\beta_2}$$

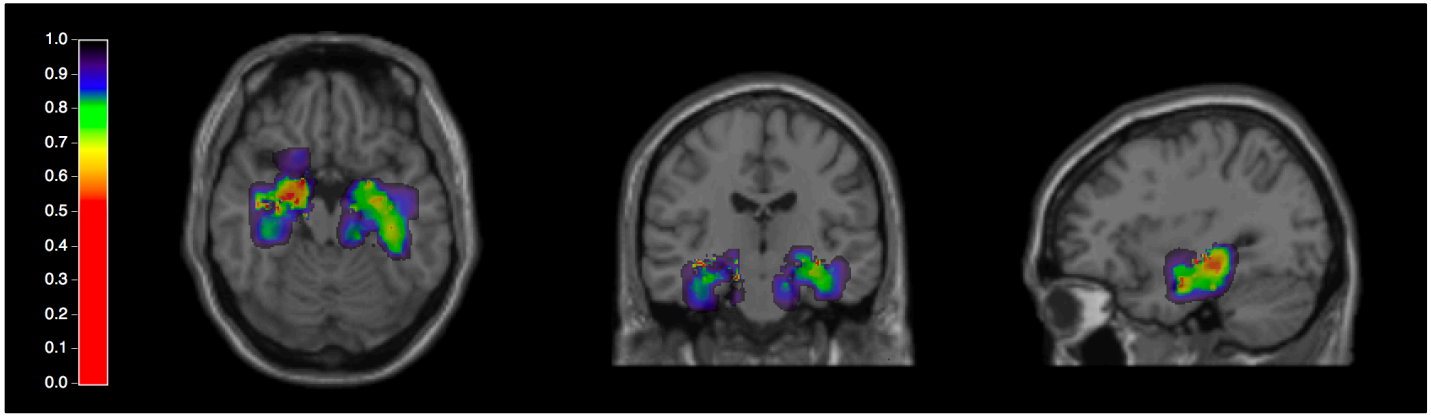
Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Results on AD vs CN classification**

Group level masking

- Mask computed on a set of images
- Possible because of image registration



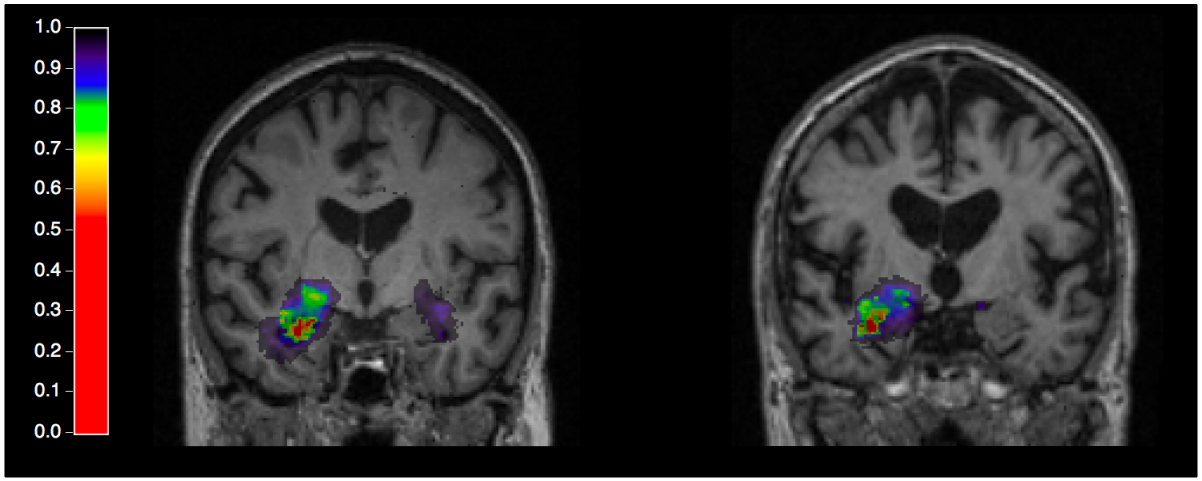
Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Results on AD vs CN classification**

Individual level masking

- **Mask computed on one single image**



Patient A Patient B

Source: Elina Thibeau-Sutre

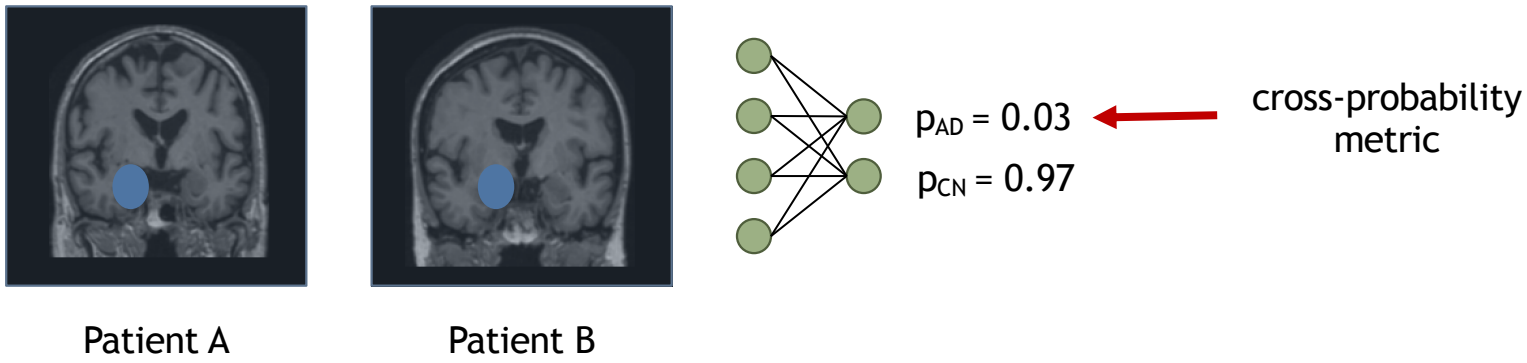
Application to MRI classification in AD

- **Metrics**

  - **Cross-probability ($prob_{CNN}$)**

    apply a mask tuned on context 1 to context 2

    - 0 → similar
    - 1 → dissimilar



Source: Elina Thibeau-Sutre

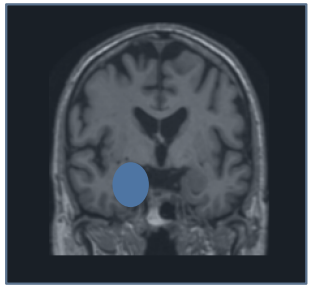
Application to MRI classification in AD

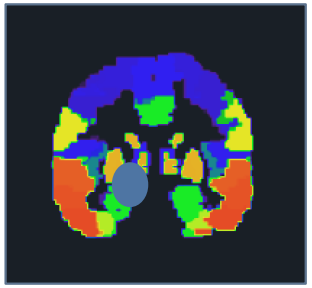
- **Metrics**

  - **Region-based similarity**

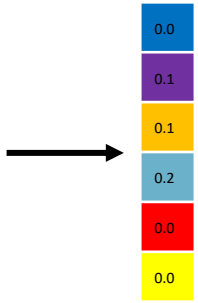
    compare the densities of two masks on all ROIs of AAL2

    - 0 → dissimilar
    - 1 → similar





Atlas AAL2



ROI-vector

$$sim = \frac{v_1 . v_2}{\|v_1\| * \|v_2\|}$$

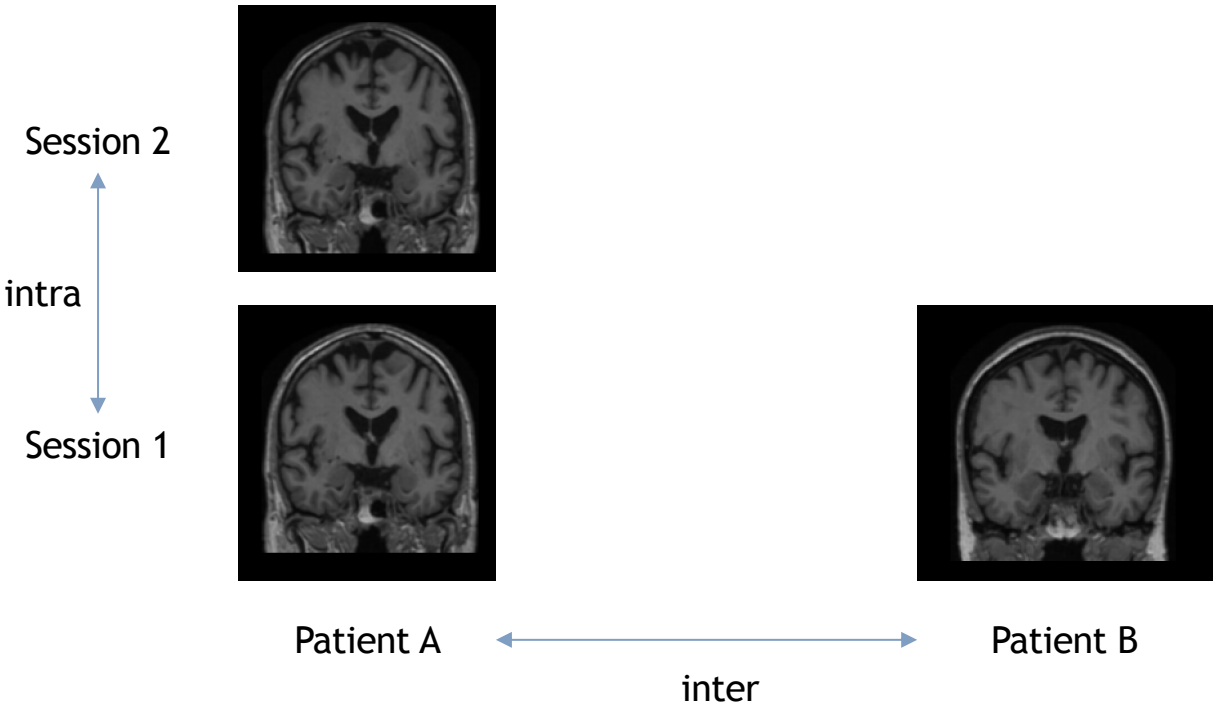
cosine similarity

Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Robustness of the masking method**

➢ **Comparison of intra- and inter-subject similarity**



Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Robustness of the masking method**

  - **Comparison of intra- and inter-subject similarity**

Analysis	Intra-subject	Inter-subject
Cross-probability (dissimilarity)	0.11	0.58
regional-based (similarity)	0.94	0.80

Individual masks are more similar within subjects than between subjects

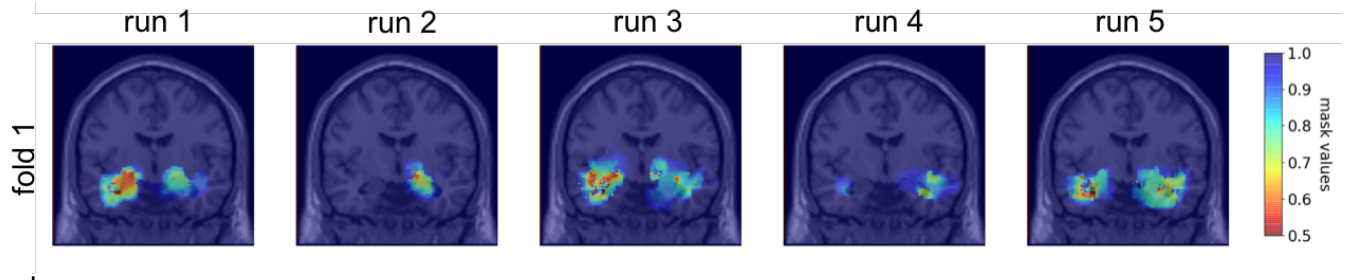
→ Coherence of the masking method

Source: Elina Thibeau-Sutre

Application to MRI classification in AD

- **Robustness of CNN training**

➢ **Interpretation of CNN trainings on different runs**



Test (ADNI)	0.89	0.89	0.83	0.88	0.90
Test (AIBL)	0.88	0.85	0.85	0.87	0.90

Analysis	Inter-subject	Inter-runs
Cross-probability (dissimilarity)	0.58	0.82
ROI-based (similarity)	0.80	0.69

→ CNN training does not robustly identify the relevant regions

Different CNN trainings lead to different interpretations

Conclusion

- It is often a good idea to use interpretability methods to ensure the decision of the network is not based on artifacts
- Many methods exists, some of them having serious drawbacks