

Machine learning: Basics

Bulat Ibragimov

bulat@di.ku.dk

Department of Computer Science
University of Copenhagen

UNIVERSITY OF COPENHAGEN



Learning Objectives

- Classification of machine learning methods
- Metrics for classification and segmentation evaluation
- Cross-validation and hold out for evaluation
- How to read MIA articles

Machine learning

Traditional Programming



Machine Learning



Metrics

MSE
MAE
MAPE
PSNR
SSIM
ACCURACY
PRECISION
RECALL
F1-score
AUC-ROC
.
.

Glossary

- **Sample** (x) - unit of data (observation, instance), e.g. medical image
- **Dataset** - a collection of samples
- **Ground-truth** - information provided by **direct observation** as opposed to information provided by inference, e.g. **human-expert said** that the image contains a **cat**.
- **Label** (y) - some value associated with the sample, **defining the ground-truth** annotation for that sample.
 - i. **categorical** \Leftrightarrow classification/categorization learning problem
 - ii. **continuous** \Leftrightarrow regression learning problem
 - iii. **Can be anything**: image, text, sound, vector ...
- **Prediction** - the output of an algorithm for some particular sample.

Types of learning

- **Supervised**

- a. **Labeled** dataset, each sample has a ground-truth label
- b. **The aim is to make predictions about new data**

- **Unsupervised**

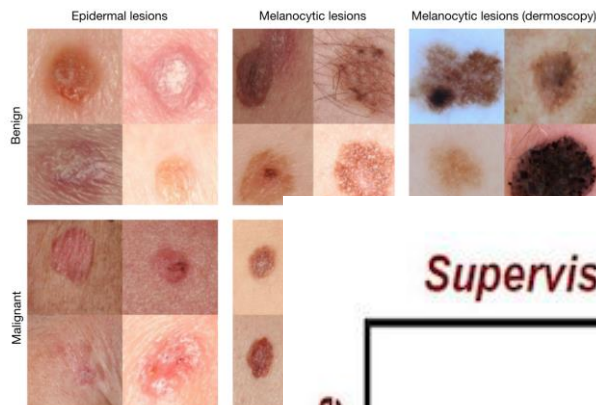
- a. **Unlabeled** dataset.
- b. Typically used to **discover the underlying structure in data**.
 - i. **Clusterisation**: dividing data into groups
 - ii. **Dimensionality reduction**

Input	Output
Chest X-ray	Pathology? 0/1
CT scan	Heart Volume in m ³
Image of tumor	Malignant/Benign ? 0/1
Radiologist report	Machine readable labels

- **Semi-supervised**

- a. The small-portion of the dataset has labels. The **majority has no labels**.

Examples for learning types



VS



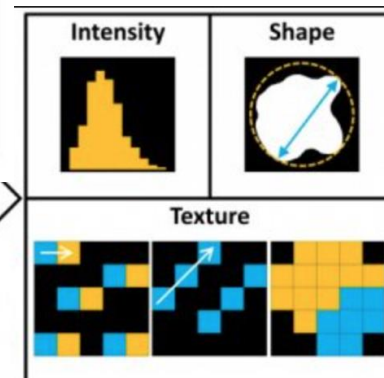
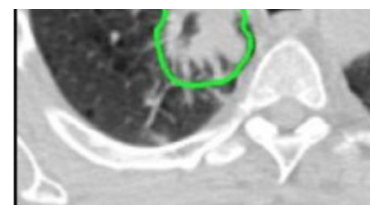
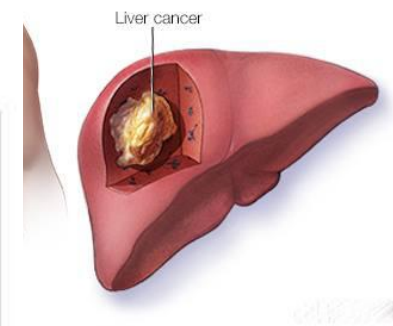
Supervised Learning

Unsupervised Learning

Discrete

Continuous

classification or categorization	clustering
regression	dimensionality reduction



Types of supervised methods

Given: the dataset $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Aim: Find function that can map $f(x_i) = \tilde{y}_i \approx y_i$

Parametric: we make an explicit assumption about the complexity of an estimated function f :

Advantages:

- Reduces the problem of estimating f to the problem of estimating the fixed set of parameters, which is generally much easier.

Disadvantages:

- Model complexity we choose usually does not match the true unknown complexity of f .

Non-parametric: no explicit assumption about the complexity of f .

Advantages:

- More flexible and not restricted by prior f from.

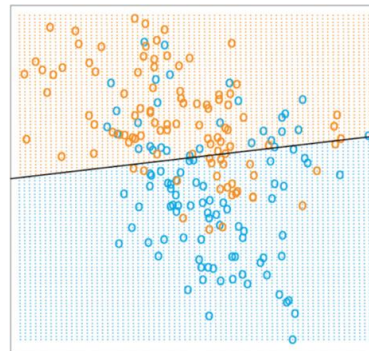
Disadvantages:

- Requires more data, not easily interpretable.

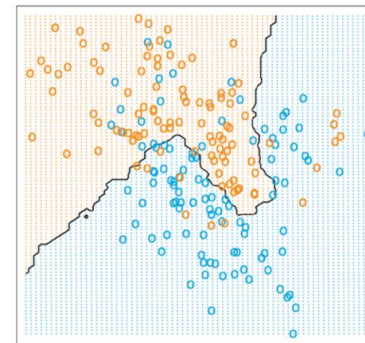
Parametric models

Parametric: we make an explicit assumption about the complexity of an estimated function f :

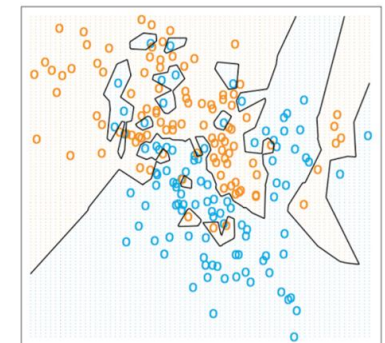
Classification



(a)



(b)



(c)

Examples of complexity assumption:

a) **Linear model:** $\hat{y} = b_0 + b_1x$

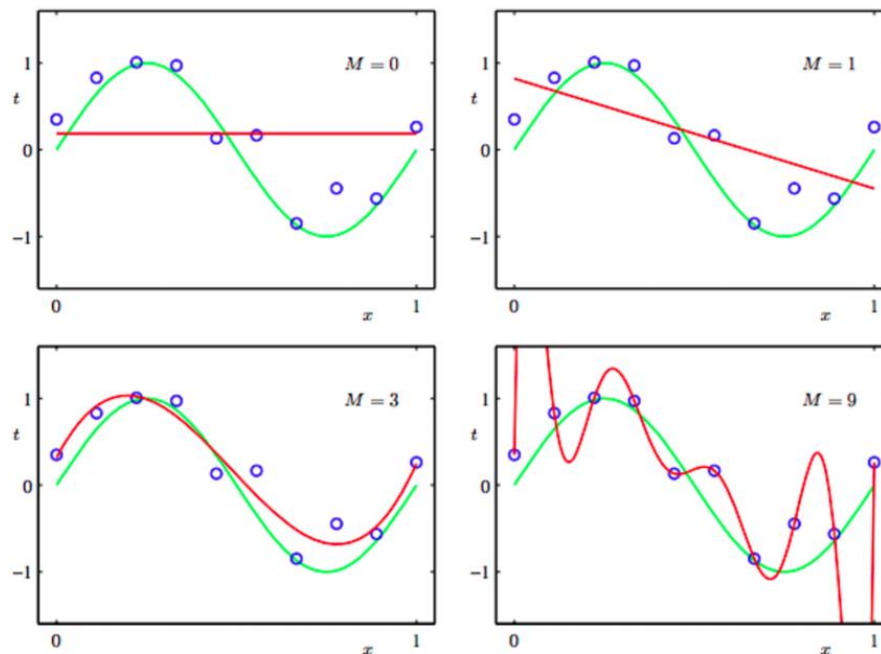
b) **Quadratic model:** $\hat{y} = b_0 + b_1x + b_2x^2$

c) **N-degree polynomial:** $\hat{y} = b_0 + b_1x + b_2x^2 + \dots + b_nx^n$

Parametric models

Parametric: we make an explicit assumption about the complexity of an estimated function f :

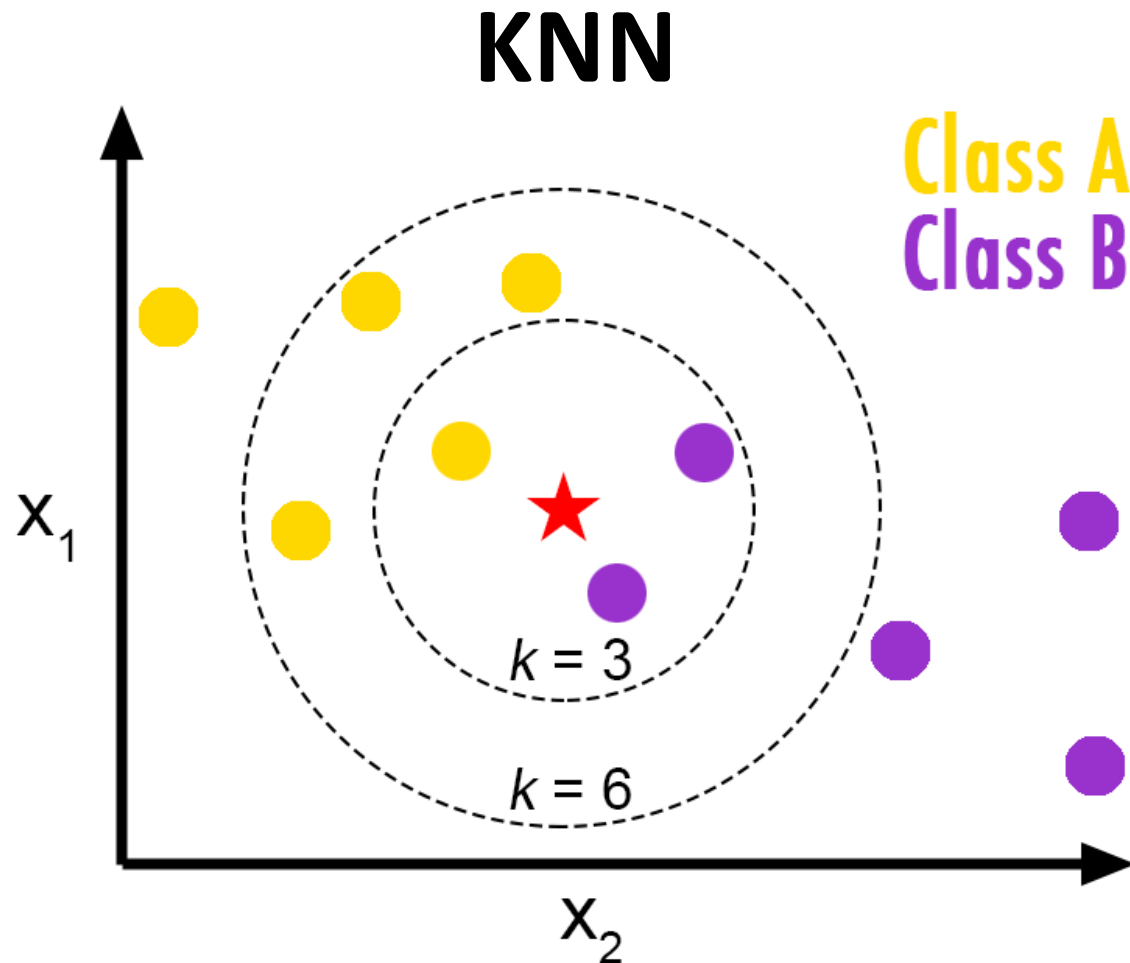
overfitting



Polynomial Curve Fitting: Polynomials having various order M (in red), fitted to the data (in blue) coming from the true underlying curve shown in green.

Non-parametric models

Non-parametric: no explicit assumption about the complexity of f :



Performance evaluation: classification

Binary classification: Labels are either 0 or 1

Binary cross-entropy

$$BCE = -y \log \hat{y} - (1 - y) \log (1 - \hat{y})$$

	Sample1
Label	0
Prediction	0.5
BCE	0.301

$$\begin{aligned} BCE &= -0 \cdot \log(0.5) - (1 - 0) \cdot \log(1 - 0.5) \\ &= 0 + 0.301 = 0.301 \end{aligned}$$

	Sample2
Label	0
Prediction	0.2
BCE	0.097

$$\begin{aligned} BCE &= -0 \cdot \log(0.2) - (1 - 0) \cdot \log(1 - 0.2) \\ &= 0 + 0.097 = 0.097 \end{aligned}$$

Performance evaluation : classification

Categorical classification: Labels are integers from 0 or N

Cross-entropy

$$CE = - \sum_i^C t_i \log(f(s)_i)$$

	Sample1		
Label	0	1	0
Prediction	0.2	0.7	0.1
BCE	0.15		

$$\begin{aligned} BCE &= -0 \cdot \log(0.2) - 1 \cdot \log(0.7) - 0 \cdot \log(0.1) \\ &= 0 \cdot 0.69 + 1 \cdot 0.15 + 0 \cdot 1 = 0.15 \end{aligned}$$

Performance evaluation: confusion matrix

Binary classification: Labels are either 0 or 1

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

1. **TP** (True positive) - number of correctly classified samples with the reference **label 1**.
2. **TN** (True negative) - number of correctly classified samples with the reference **label 0**.
3. **FP** (False positive) - number of mis-classified samples with the reference **label 0**.
4. **FN** (False negative) - number of mis-classified samples with the reference **label 1**.

Performance evaluation: confusion matrix

Binary classification: Labels are either 0 or 1

Assume: Label 1 is a patient **with disease**

Are **FP** and **FN** values equally important?

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

Performance evaluation: metrics

Metrics:

- **Accuracy** - fraction of samples, where your model **prediction was correct** against the **total** number of observations.

$$\frac{TP + TN}{N_{samples}}$$

- **Specificity** (aka. true negative rate) - a fraction of observations, where your model **prediction was positive and correct**, to the total number of **positive predictions**.

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

- **Sensitivity** (aka. true positive rate) - a fraction of observations, where your model **prediction was negative and correct**, to the total number of **negative predictions**:

$$\frac{TP}{TP + FN}$$



Example

Three algorithms for computer-aided diagnosis of lung diseases:

- **1) True negative rate = 0.97; True positive rate = 0.97**
- **2) True negative rate = 1; True positive rate = 0.75**
- **3) True negative rate = 0.75; True positive rate = 1**

$$\text{TNR} = \frac{TN}{TN+FP}; \quad \text{TPR} = \frac{TP}{TP+FN}$$

Performance evaluation: ROC-AUC

Results of prediction:

	Case1	Case2	Case3	Case4	Case5
Reference labels	0	1	1	0	1
Predictions	0.1	0.4	0.8	0.5	0.9

- Let's say we use 0.2 as a threshold to separate negative predictions:

- True negative rate $\frac{TN}{TN+FP} = \frac{1}{2} = 0.5$

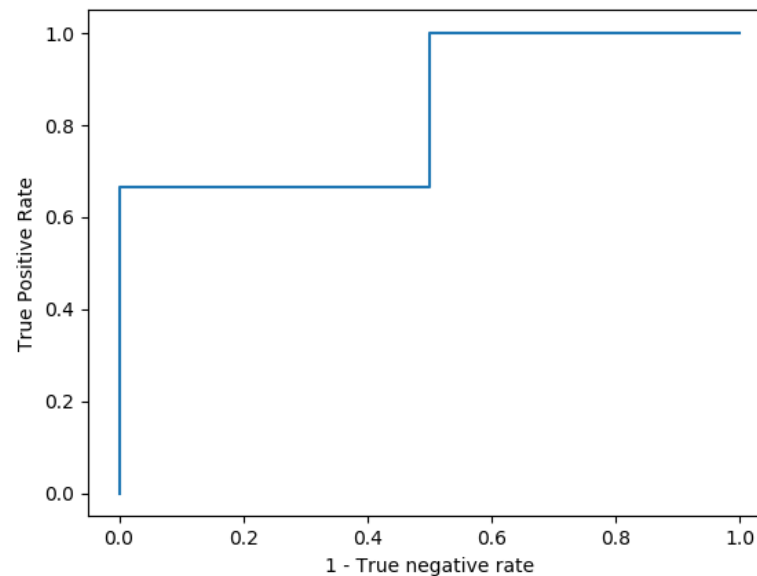
- True positive rate $\frac{TP}{TP+FN} = \frac{1}{1} = 1$

Performance evaluation: ROC-AUC

Results of prediction:

	Case1	Case2	Case3	Case4	Case5
Reference labels	0	1	1	0	1
Predictions	0.1	0.4	0.8	0.5	0.9

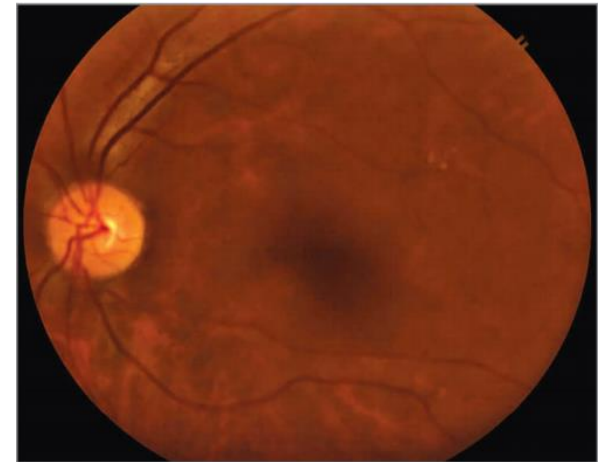
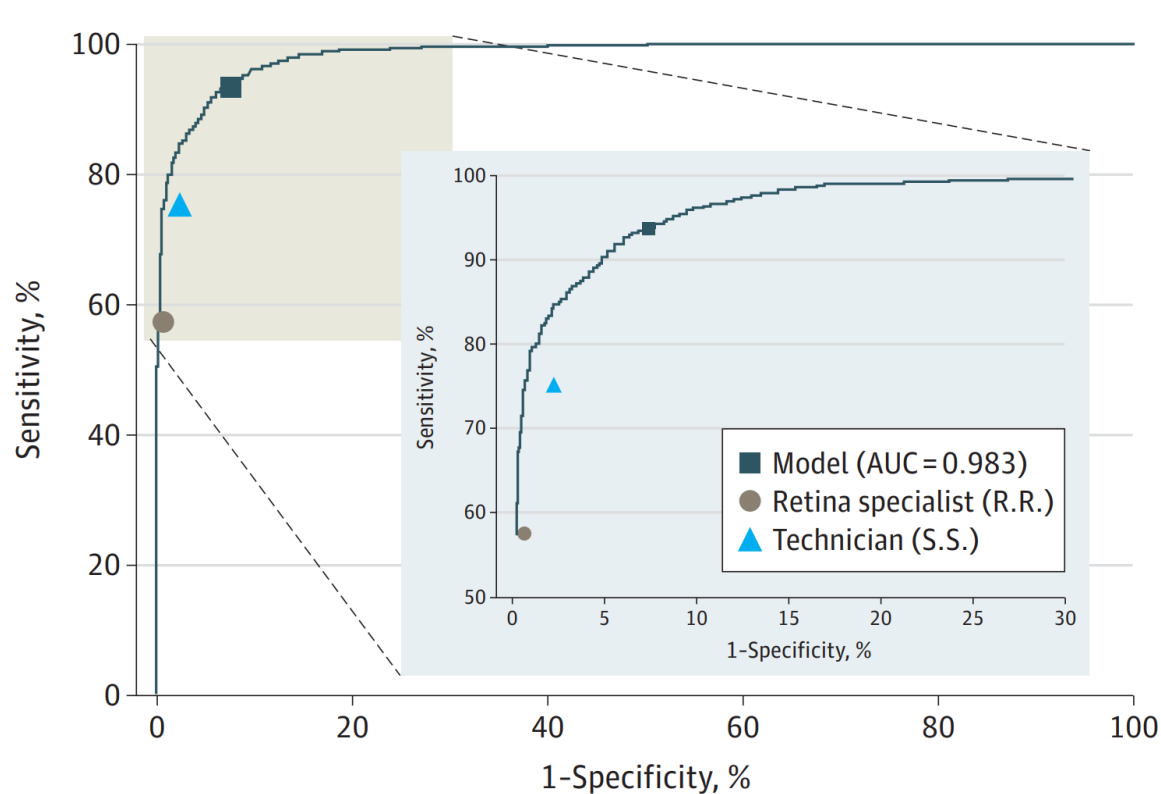
	0	0.1	0.4	0.5	0.8	0.9	1
<u>True negative rate</u>	0	0	0.5	0.5	1	1	1
<u>True positive rate</u>	1	1	1	0.66	0.66	0.33	0



Roc curve

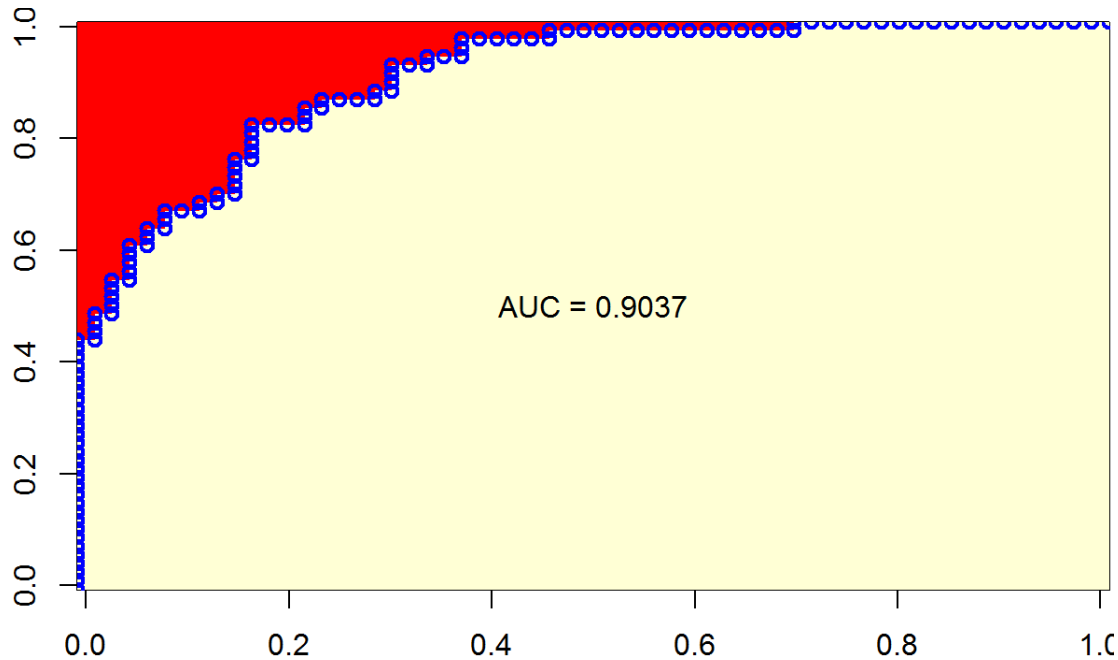
Performance evaluation: ROC-AUC

Comparison to human performance:

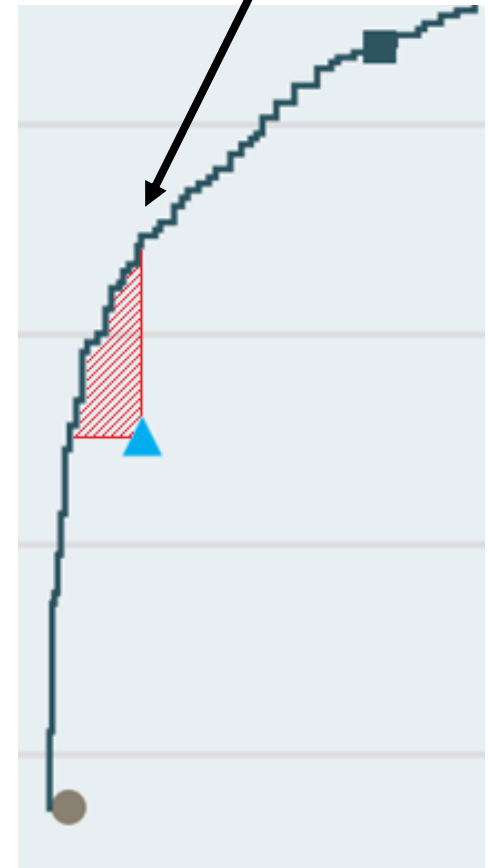
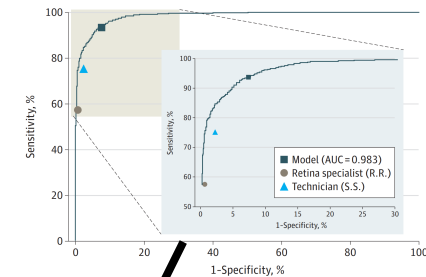


Performance evaluation: ROC-AUC

Comparison to human performance:



Closer the area under the curve to 1 the better



Performance evaluation: regression

Regression: Labels are floating/integer numbers

- Mean absolute error

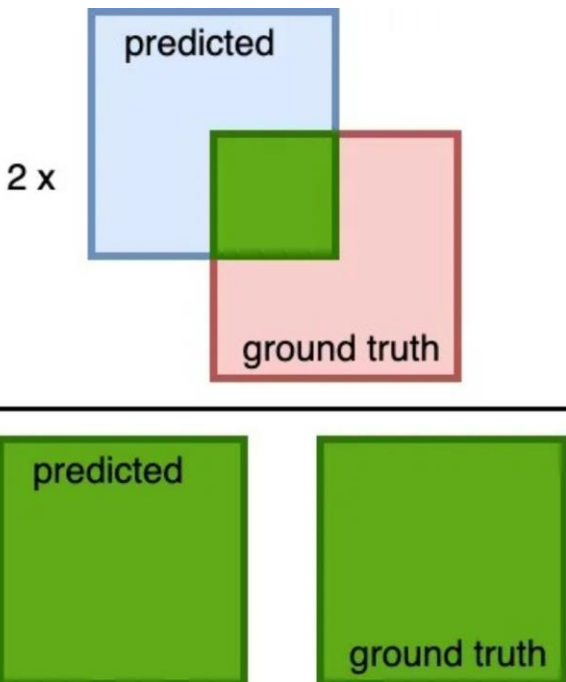
$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

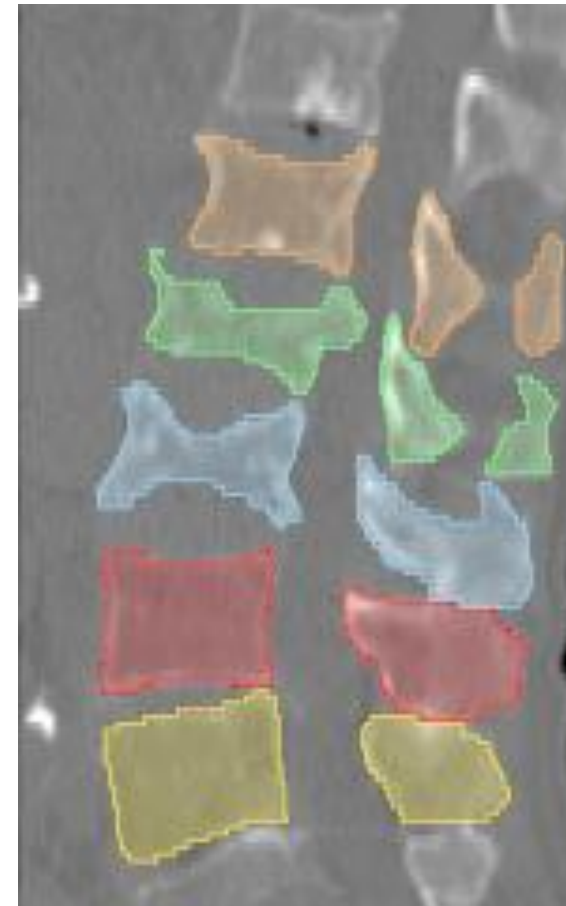
- Mean squared error

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Performance evaluation: segmentation

Dice coefficient

$$\text{Dice coefficient} = \frac{2 \times \text{area of overlapped (green)}}{\text{total area (green)}} = \frac{\text{predicted} \cap \text{ground truth}}{\text{predicted} \cup \text{ground truth}}$$




Break

Train-validation-test

Training Dataset: the part of the data used to optimize model f parameters.

Testing Dataset: the part of the data used to evaluate how good the model f work.

If the model f has too many parameters it may perfectly capture the training dataset, but works poorly on the testing dataset.
Imagine that you do not have access to the test dataset at all

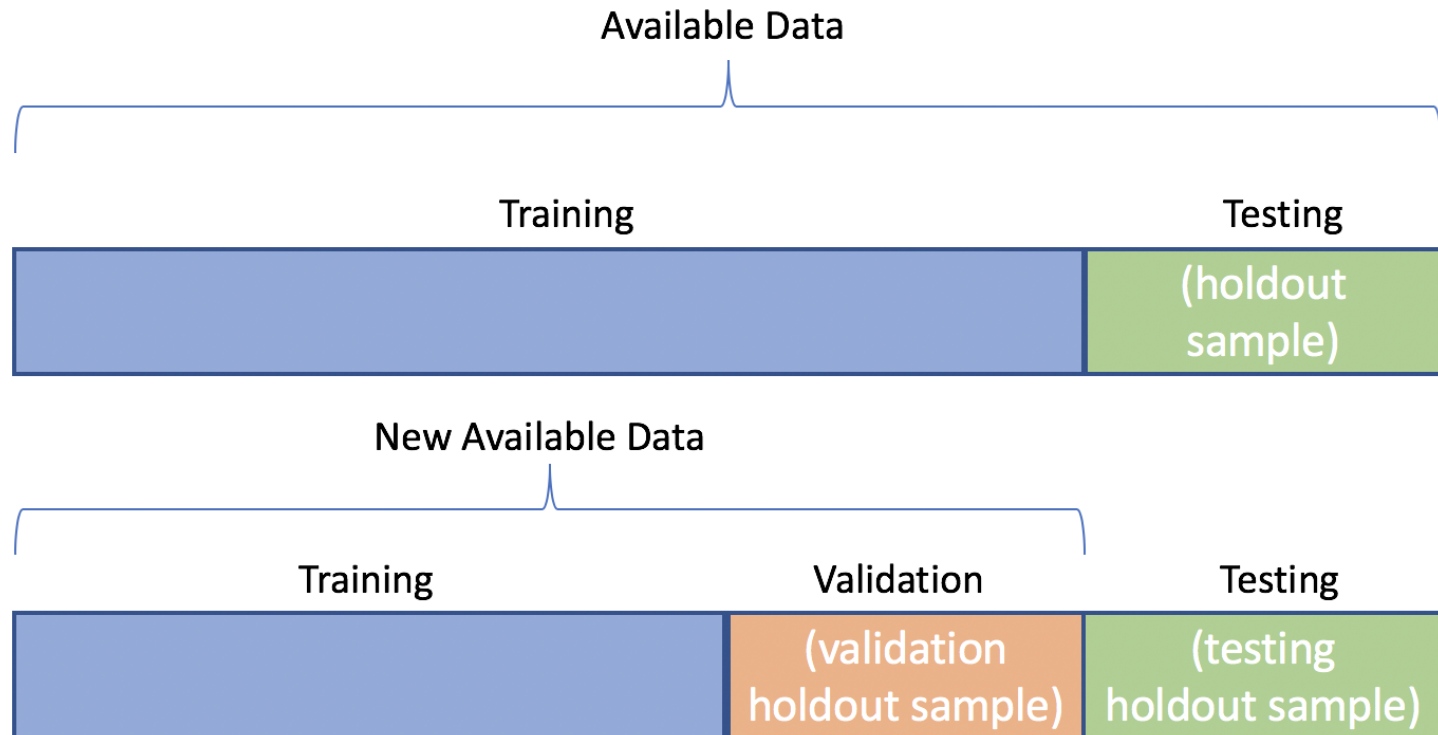


Train-validation-test: hold-out test

Training Dataset

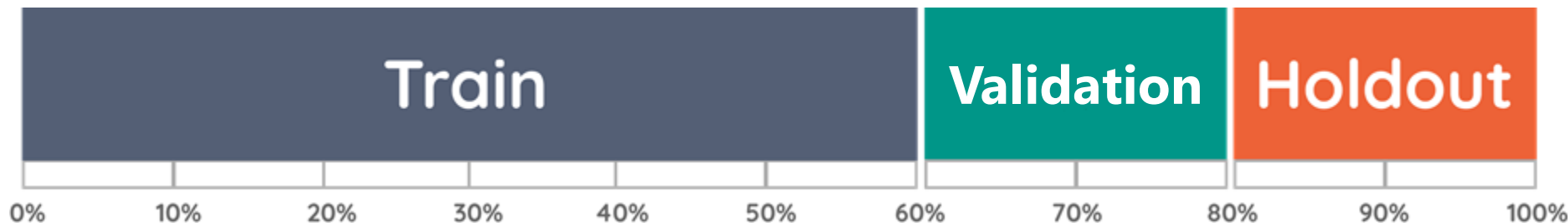
Testing Dataset

Validation Dataset: the part of the data for estimating the performance of the model before final evaluation on the test dataset.



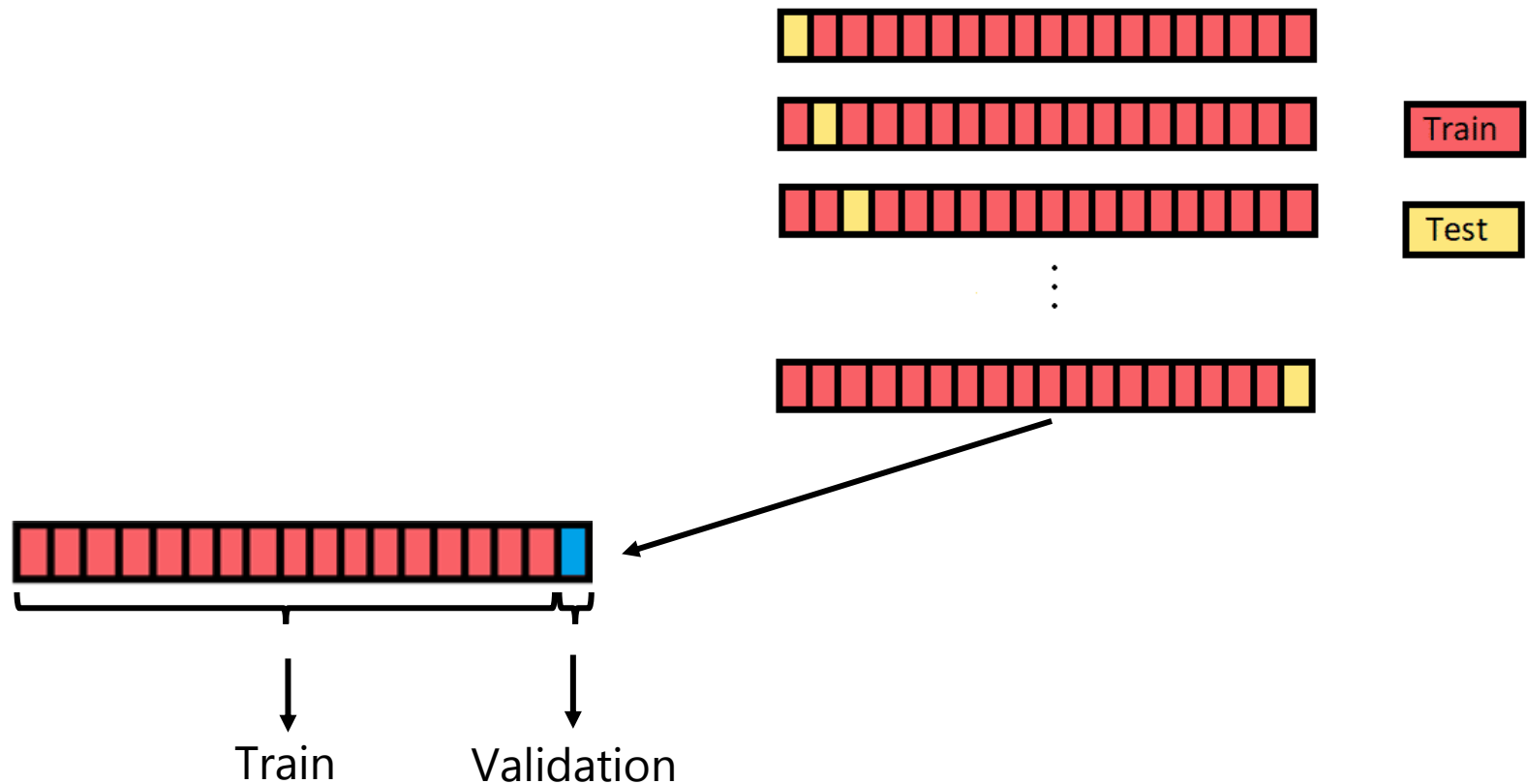
Hold-out test

The simplest strategy for model evaluation is to separate test data in advance from the complete database and only analyze it when the model development is finished



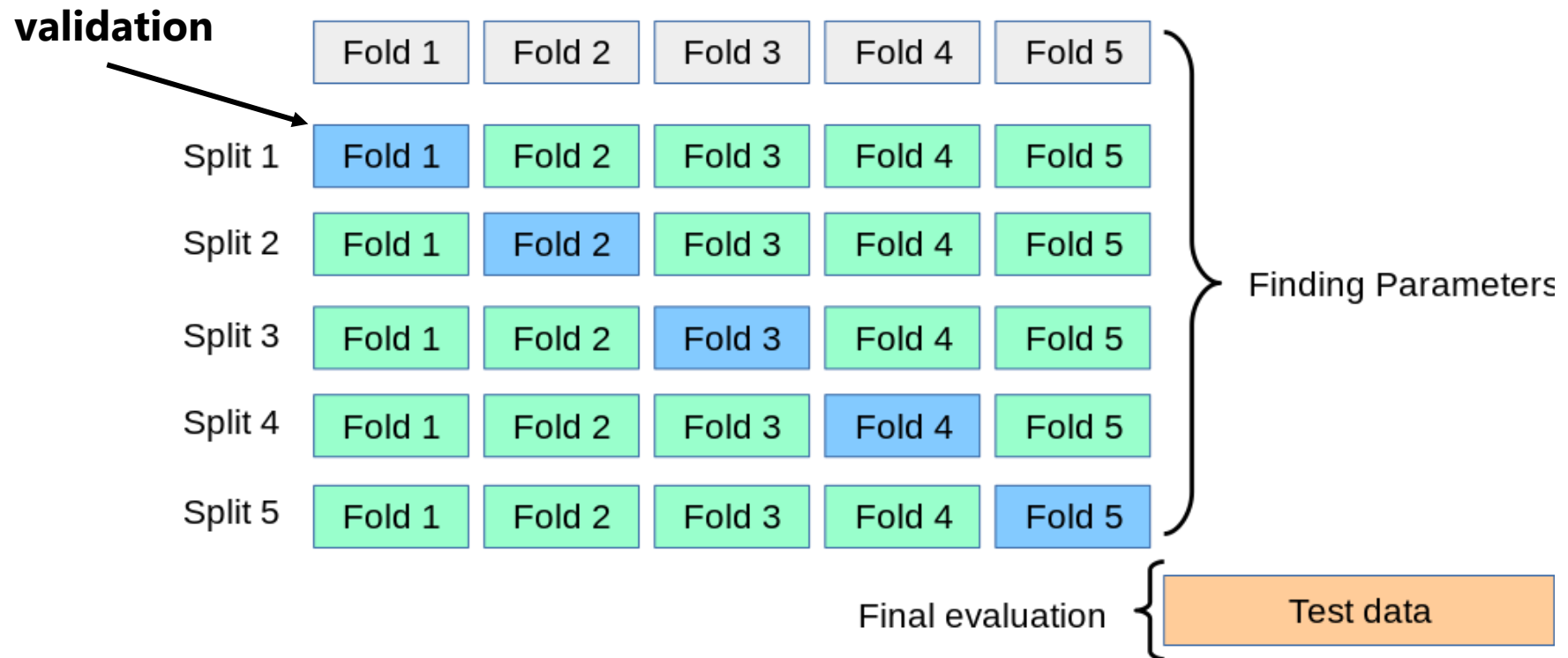
Cross-validation test

Cross-validation or 'k-fold cross-validation' is when the dataset is randomly split up into 'k' groups. One of the groups is used as the test set and the rest are used as the training set.



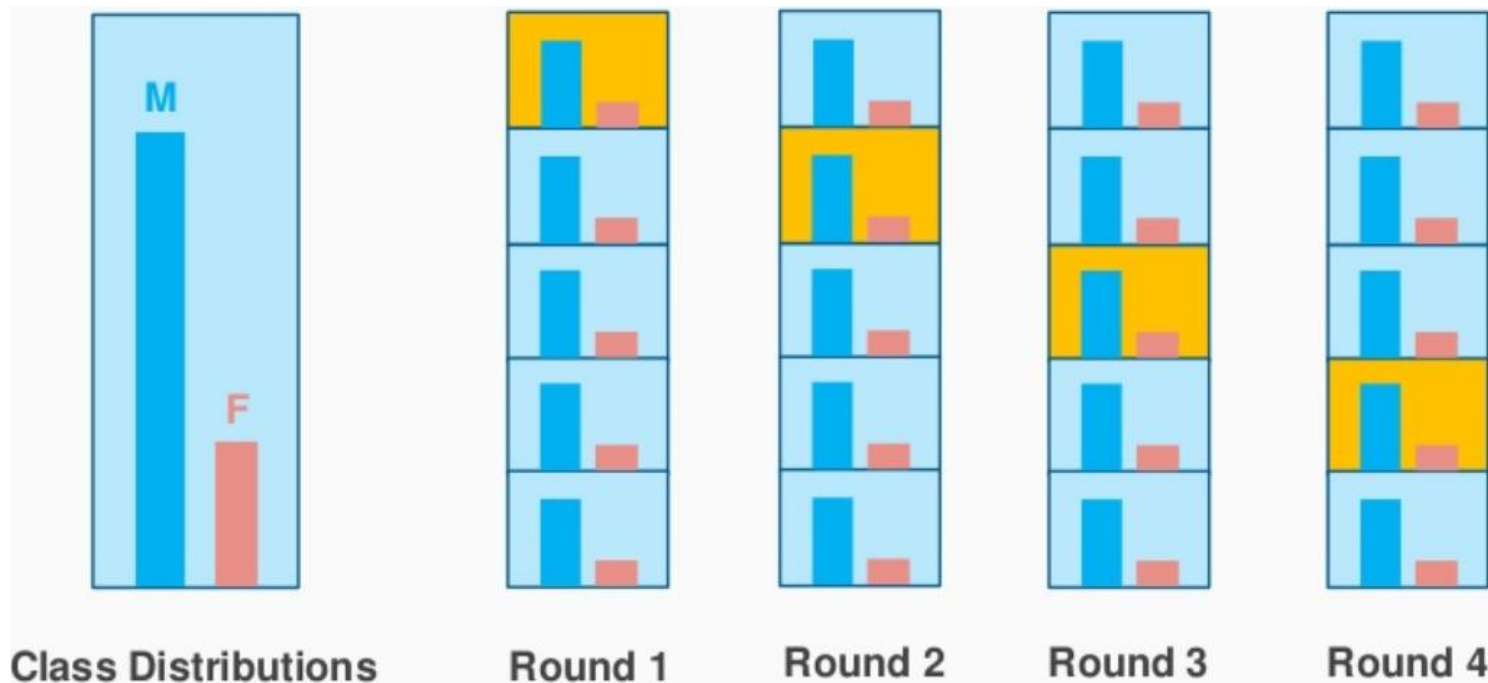
Combination of cross-validation and hold out

Separate train and test datasets. Generate k-fold cross-validation-based models training on k splits of the train dataset, and then evaluate on the test dataset.



Stratified splits

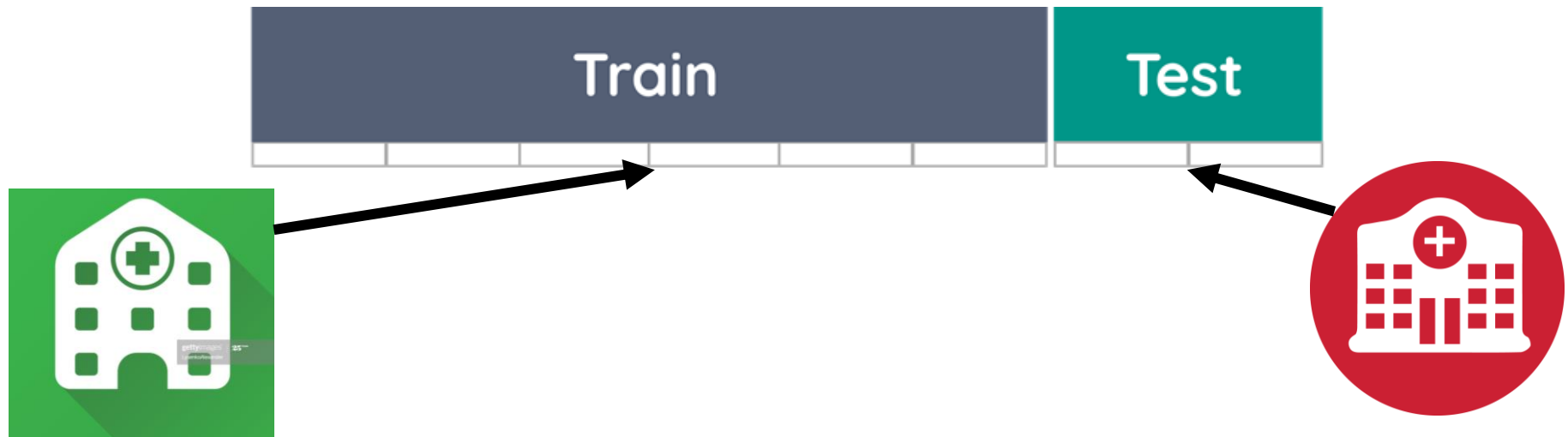
The splitting of data is governed by criteria such as ensuring that **each fold has the same proportion of observations with a certain property.**



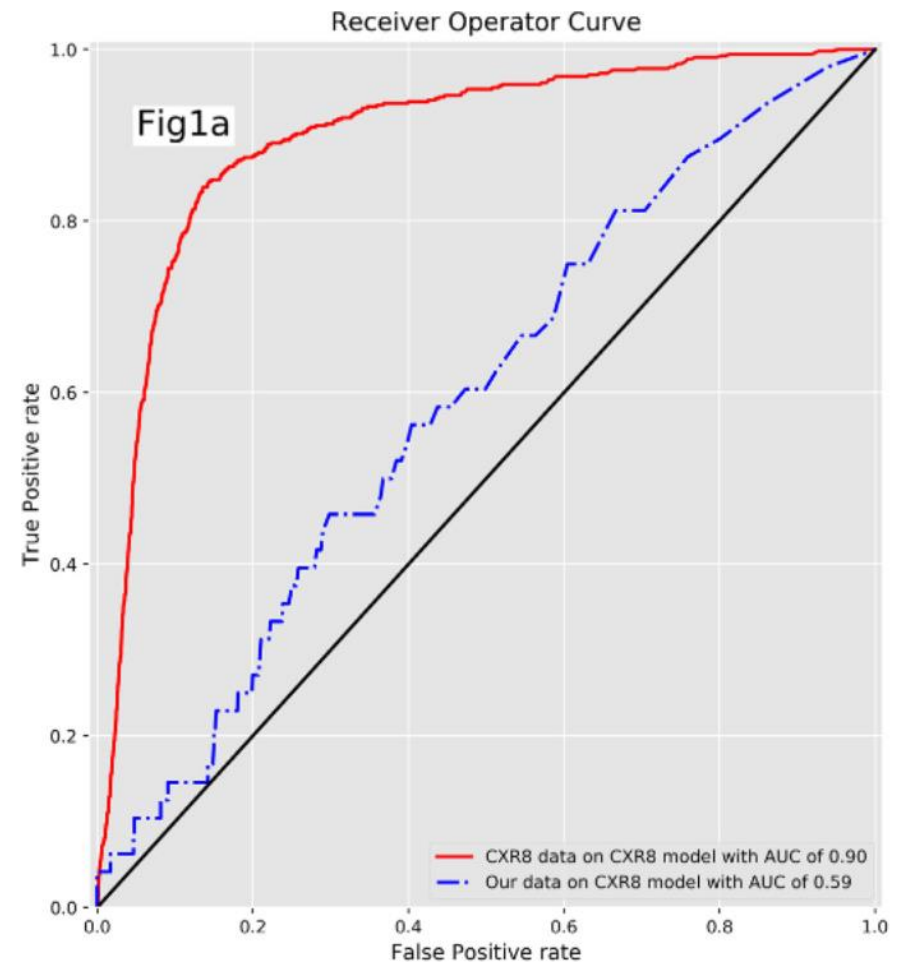
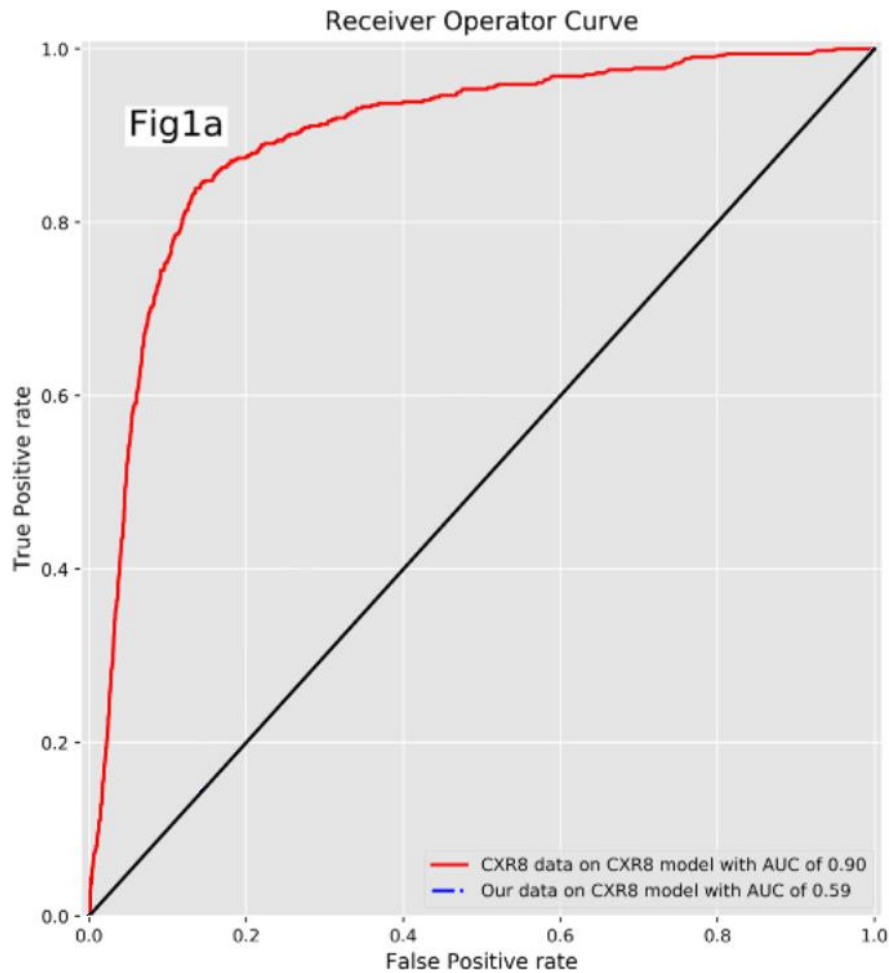
Train-test from different sources

Training Dataset comes from hospital A

Testing Dataset comes from hospital B



Train-test from different sources





Cross-validation or holdout?

What kind of evaluation approach will you use for the database:

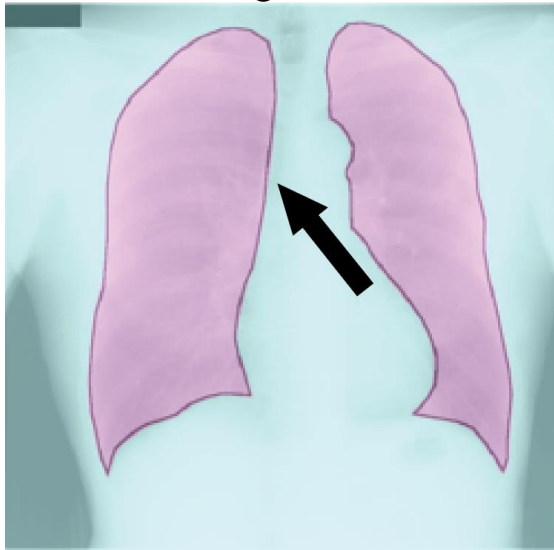
- 500 brain MRI. The aim is to automatically diagnose stroke.
- 10000 lung X-rays. The aim is to automatically diagnose pneumonia.

Example: lung field segmentation

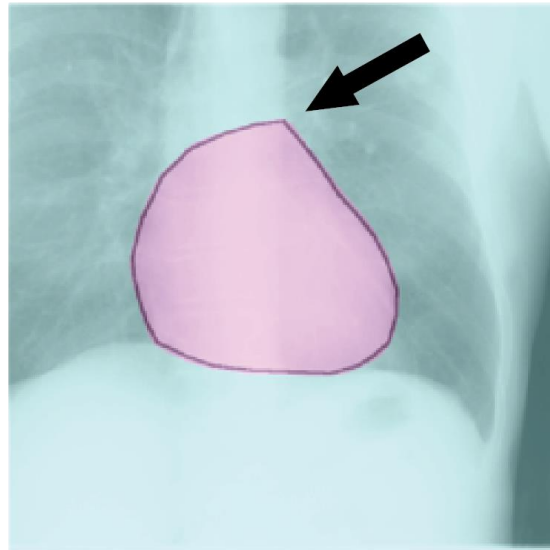
A public database of 247 chest X-rays:

- Aim: segment lung fields, heart and clavicles

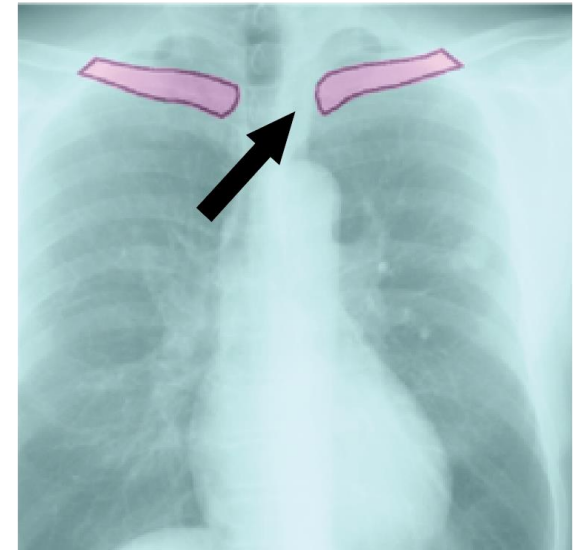
Lung fields



Heart



Clavicles





Example: lung field segmentation

Human performance

Deep learning

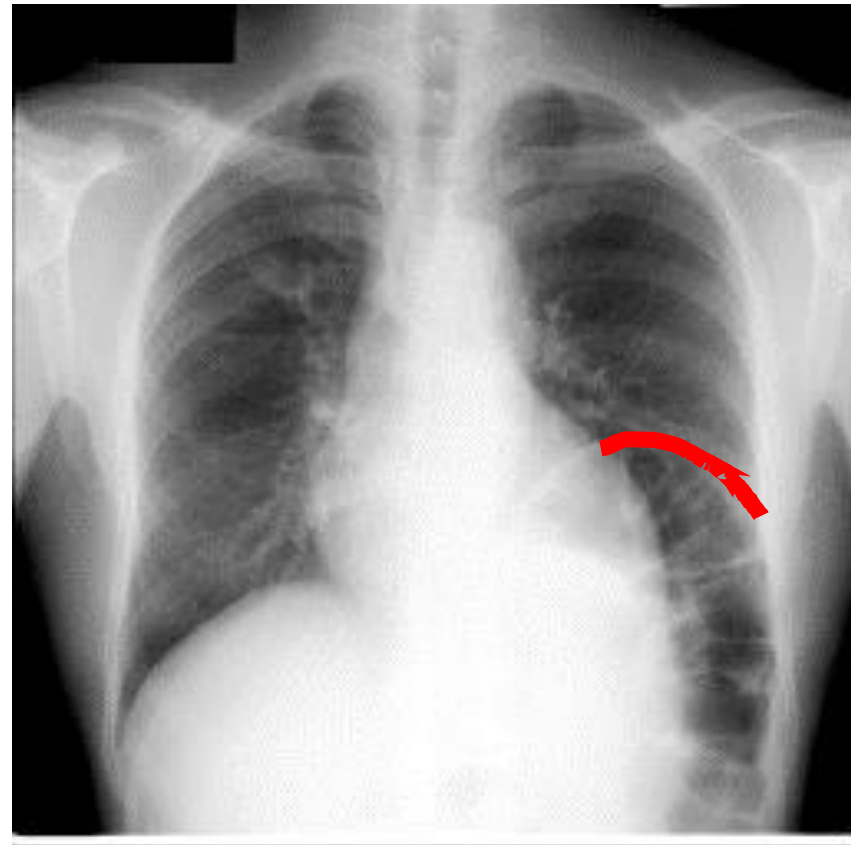
Method	Lung fields
AAM ¹ [8]	0.847
ASM ² [8]	0.927
ASM with SIFT ³ [40]	0.930
ASM & AAM [41]	0.931
ASM & AAM [42]	0.940
Atlas registration [43]	0.940
Second observer [8]	0.946
GTF ⁴ [44]	0.946
SCAN ⁵ [14]	0.947
UNet [16]	0.950
FCN-DAL ⁶ [15]	0.951
LF-SegNet [17]	0.951
MISCP ⁷ [45]	0.951
Customized ASM [46]	0.952
SED ⁸ [47]	0.952
ASLM ⁹ [48]	0.953
Atlas lung+heart [4]	0.954
LinkNet_ResNeXt50_Masks+Contours	0.955
Multi-task FCN [12]	0.959
CNN-AC ¹⁰ [13]	0.961
UNet /ImageNet [18]	0.961
Tiramisu_DenseNet56_Masks+Contours	0.961
Adaptive region growing [49]	0.963
UNet_ResNeXt50_Masks+Contours	0.971

Does this mean deep learning
can replace human here?

Example: lung field segmentation

Performance summary:

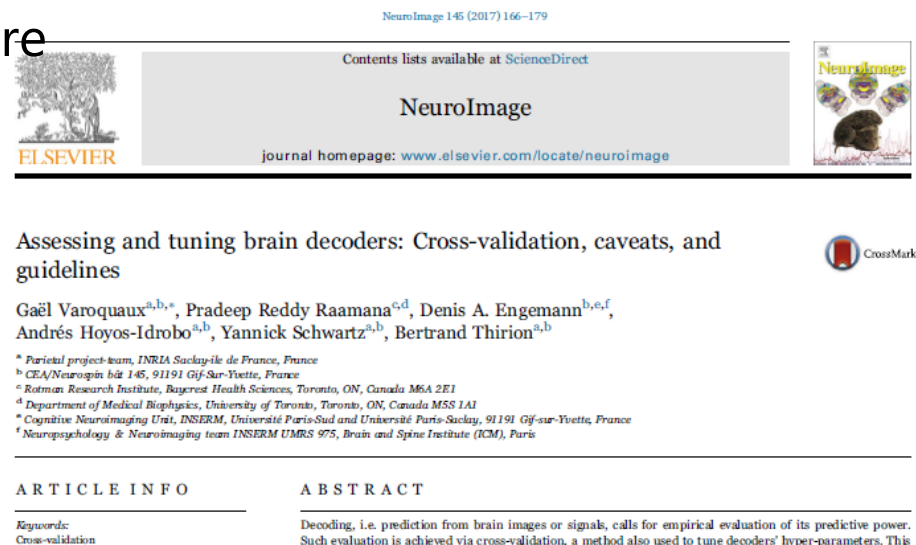
- Algorithms work very good on easy X-rays
- Algorithms learns how reference doctor creates contours and replicates him slightly better than another doctor
- Algorithms work poorly on pathological case
- Accuracy gained on many easy cases outweighs the error on few difficult cases



How to read/write articles on ML for MIA

How to search for articles:

- Google scholar
- Try "Assessing and Tuning brain decoders"
 - How many citations?
- Find related articles
 - Examine articles that cite the "core" article
 - Use the "related articles" feature



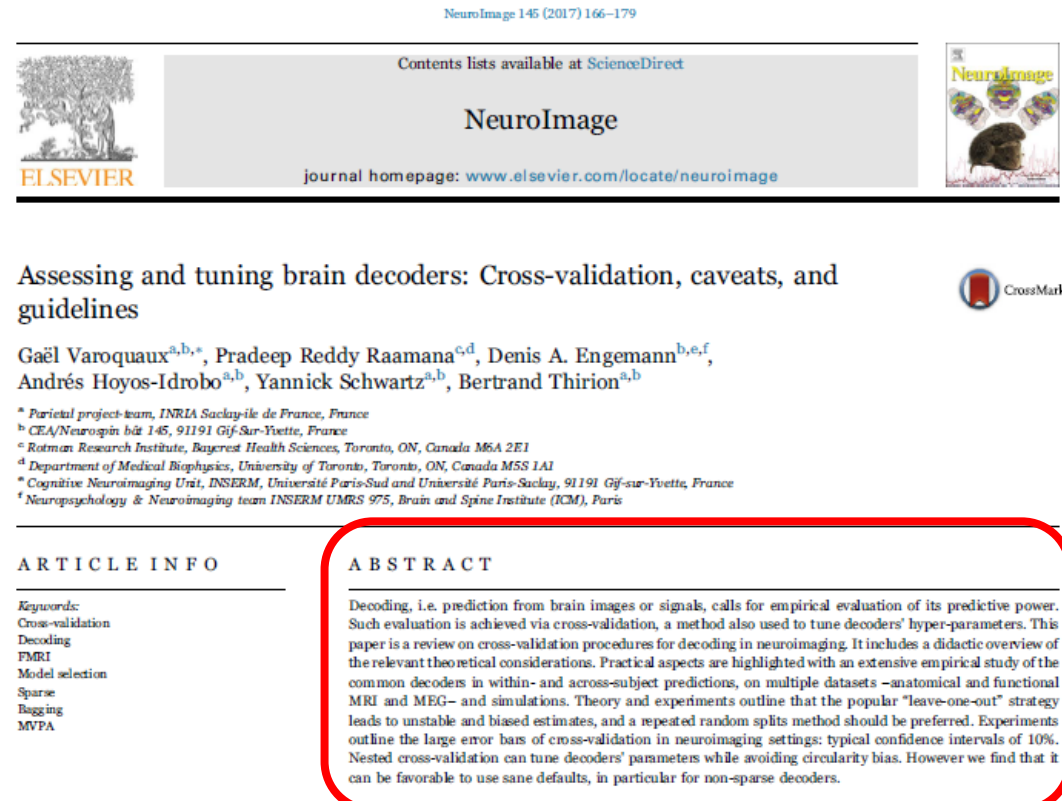
How to read/write articles on ML for MIA

Abstract sums up the paper:

- What is the problem
- Methods
 - Data
 - What did we do
- Results
- Conclusion

Many people read abstract first.

If you cannot understand the paper from the abstract – the paper is probably not good.



How to read/write articles on ML for MIA

Figures are essential in papers! Many people quickly go through images to judge the paper quality

Figures should be self-sufficient

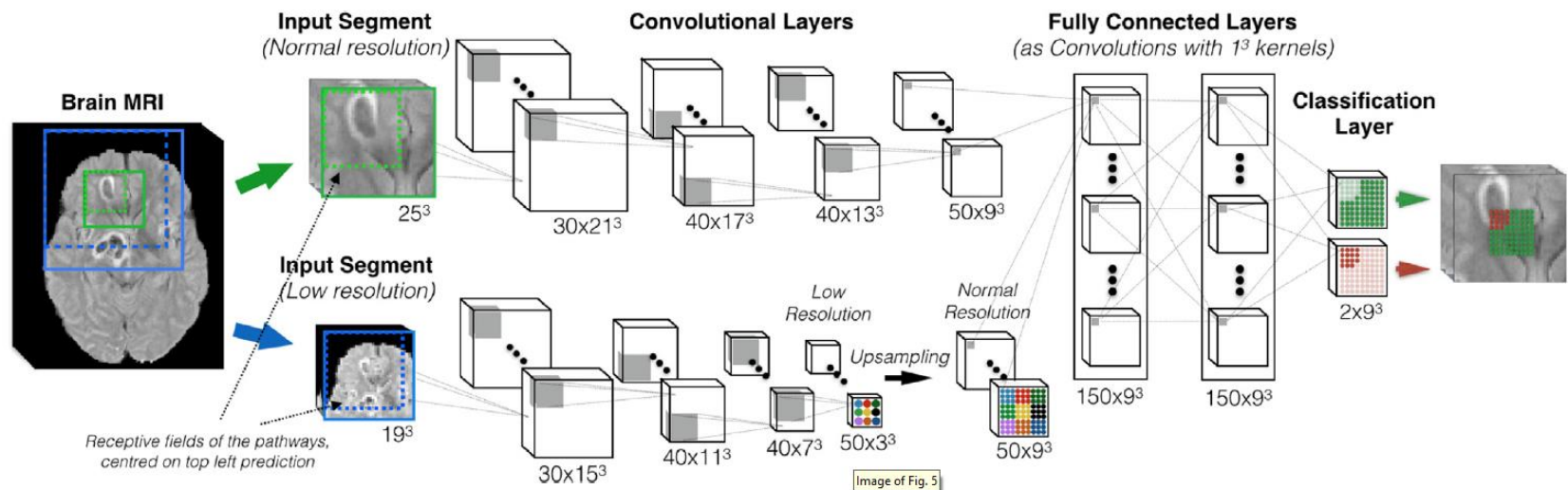


Fig. 5. Multi-scale 3D CNN with two convolutional pathways. The kernels of the two pathways are here of size 5^3 (for illustration only to reduce the number of layers in the figure). The neurons of the last layers of the two pathways thus have receptive fields of size 17^3 voxels. The inputs of the two pathways are centred at the same image location, but the second segment is extracted from a down-sampled version of the image by a factor of 3. The second pathway processes context in an actual area of size 51^3 voxels. *DeepMedic*, our proposed 11-layers architecture, results by replacing each layer of the depicted pathways with two that use 3^3 kernels (see [Section 2.3](#)). Number of FMs and their size depicted as (Number \times Size).

How to read/write articles on ML for MIA

Usually a paper consists of:

- Introduction
- Methodology
- Results
- Discussion
- Conclusion

Abstract

Purpose:

Accurate and efficient delineation of tumor target and organs-at-risks is essential for the success of radiotherapy. In reality, despite of decades of intense research efforts, auto-segmentation has not yet become clinical practice. In this study, we present, for the first time, a deep learning-based classification algorithm for autonomous segmentation in head and neck (HaN) treatment planning.

Methods:

Fifteen HN datasets of CT, MR and PET images with manual annotation of organs-at-risk (OARs) including spinal cord, brainstem, optic nerves, chiasm, eyes, mandible, tongue, parotid glands were collected and saved in a library of plans. We also have ten super-resolution MR images of the tongue area, where the genioglossus and inferior longitudinalis tongue muscles are defined as organs of interest. We applied the concepts of random forest- and deep learning-based object classification for automated image annotation with the aim of using machine learning to facilitate head and neck radiotherapy planning process. In this new paradigm of segmentation, random forests were used for landmark-assisted segmentation of super-resolution MR images. Alternatively to auto-segmentation with random forest-based landmark detection, deep convolutional neural networks were developed for voxel-wise segmentation of OARs in single and multi-modal images. The network consisted of three pairs of convolution and pooling layer, one ReLU layer and a softmax layer.

Results:

We present a comprehensive study on using machine learning concepts for auto-segmentation of OARs and tongue muscles for the HaN radiotherapy planning. An accuracy of 81.8% in terms of Dice coefficient was achieved for segmentation of genioglossus and inferior longitudinalis tongue muscles. Preliminary results of OARs segmentation also indicate that deep-learning afforded an unprecedented opportunities to improve the accuracy and robustness of radiotherapy planning.

Conclusion:

A novel machine learning framework has been developed for image annotation and structure segmentation. Our results indicate the great potential of deep learning in radiotherapy treatment planning.

How to read/write articles on ML for MIA

Introduction:

- What do we want to do
- What was done before
- How do we intent to improve what was done before



Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines

Gaël Varoquaux^{a,b,*}, Pradeep Reddy Raamana^{c,d}, Denis A. Engemann^{b,e,f},
Andrés Hoyos-Idrobo^{a,b}, Yannick Schwartz^{a,b}, Bertrand Thirion^{a,b}

^a Parietal project-team, INRIA Saclay-Île de France, France

^b CEA/Neurospin bât 145, 91191 Gif-Sur-Yvette, France

^c Rotman Research Institute, Baycrest Health Sciences, Toronto, ON, Canada M6A 2E1

^d Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada M5S 1A1

^e Cognitive Neuroimaging Unit, INSERM, Université Paris-Saclay and Université Paris-Saclay, 91191 Gif-sur-Yvette, France

^f Neuropsychology & Neuroimaging team INSERM UMR5 975, Brain and Spine Institute (ICM), Paris

ARTICLE INFO

Keywords:
Cross-validation
Decoding
fMRI
Model selection
Sparse
Roging
MYPE

ABSTRACT

Decoding, i.e. prediction from brain images or signals, calls for empirical evaluation of its predictive power. Such evaluation is achieved via cross-validation, a method also used to tune decoders' hyper-parameters. This paper is a review on cross-validation procedures for decoding in neuroimaging. It includes a didactic overview of the relevant theoretical considerations. Practical aspects are highlighted with an extensive empirical study of the common decoders in within- and across-subject predictions, on multiple datasets – anatomical and functional MRI and MEG – and simulations. Theory and experiments outline that the popular "leave-one-out" strategy leads to unstable and biased estimates, and a repeated random splits method should be preferred. Experiments outline the large error bias of cross-validation in nested settings: typical confidence intervals of 10%. Nested cross-validation can tune decoders' parameters while avoiding circularity bias. However we find that it can be favorable to use sane defaults, in particular for non-sparse decoders.

1. Introduction: decoding needs model evaluation

Decoding, i.e. predicting behavior or phenotypes from brain images or signals, has become a central tool in neuroimage data processing (Haynes and Rees, 2006; Haynes, 2015; Kamitani and Tong, 2005; Norman et al., 2006; Varoquaux and Thirion, 2014; Yarkoni and Westfall, 2016). In clinical applications, prediction opens the door to diagnosis or prognosis (Mouro-Miranda et al., 2005; Fu et al., 2008; Demirel et al., 2008). To study cognition, successful prediction is seen as evidence of a link between observed behavior and a brain region (Haxby et al., 2001) or a small fraction of the image (Kriegeskorte et al., 2006). Decoding power can test if an encoding model describes well multiple facets of stimuli (Mitchell et al., 2008; Naselaris et al., 2011). Prediction can be used to establish what specific brain functions are implied by observed activations (Schwartz et al., 2013; Poldrack et al., 2009). All these applications rely on measuring the predictive power of a decoder.

Assessing predictive power is difficult as it calls for characterizing the decoder on prospective data, rather than on the data at hand. Another challenge is that the decoder must often choose between many

different estimates that give rise to the same prediction error on the data, when there are more features (voxels) than samples (brain images, trials, or subjects). For this choice, it relies on some form of regularization, that embodies a prior on the solution (Hastie et al., 2009). The amount of regularization is a parameter of the decoder that may require tuning. Choosing a decoder, or setting appropriately its internal parameters, are important questions for brain mapping, as these choice will not only condition the prediction performance of the decoder, but also the brain features that it highlights.

Measuring prediction accuracy is central to decoding, to assess a decoder, select one in various alternatives, or tune its parameters. The topic of this paper is cross-validation, the standard tool to measure predictive power and tune parameters in decoding. The first section is a primer on cross-validation giving the theoretical underpinnings and the current practice in neuroimaging. In the second section, we perform an extensive empirical study. This study shows that cross-validation results carry a large uncertainty, that cross-validation should be performed on full blocks of correlated data, and that repeated random splits should be preferred to leave-one-out. Results also yield guidelines for decoder parameter choice in terms of prediction

* Corresponding author at: Parietal project-team, INRIA Saclay-Île de France, France.

How to read/write articles on ML for MIA

Previous work:

- The previous work summary should be specific not broad
- Previous work summary can be written as "A did this. B did that", and/or subdivided according to methodological concepts used

NeuroImage 145 (2017) 166–179

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines

Gaël Varoquaux^{a,b,*}, Pradeep Reddy Raamana^{c,d}, Denis A. Engemann^{b,e,f},
Andrés Hoyos-Idrobo^{a,b}, Yannick Schwartz^{a,b}, Bertrand Thirion^{a,b}

^a Parietal project-team, INRIA Saclay-Île de France, France
^b CEA/Neurospin bât 145, 91191 Gif-Sur-Yvette, France
^c Rotman Research Institute, Baycrest Health Sciences, Toronto, ON, Canada M6A 2E1
^d Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada M5S 1A1
^e Cognitive Neuroimaging Unit, INSERM, Université Paris-Saclay and Université Paris-Saclay, 91191 Gif-sur-Yvette, France
^f Neuropsychology & Neuroimaging team INSERM UMR5 975, Brain and Spine Institute (ICM), Paris

ARTICLE INFO

Keywords:
 Cross-validation
 Decoding
 fMRI
 Model selection
 Sparse
 Regression
 MVA

ABSTRACT

Decoding, i.e. prediction from brain images or signals, calls for empirical evaluation of its predictive power. Such evaluation is achieved via cross-validation, a method also used to tune decoders' hyper-parameters. This paper is a review on cross-validation procedures for decoding in neuroimaging. It includes a didactic overview of the relevant theoretical considerations. Practical aspects are highlighted with an extensive empirical study of the common decoders in within- and across-subject predictions, on multiple datasets – anatomical and functional MRI and MEG – and simulations. Theory and experiments outline that the popular "leave-one-out" strategy leads to unstable and biased estimates, and a repeated random splits method should be preferred. Experiments outline the large error bias of cross-validation in nested settings: typical confidence intervals of 10%. Nested cross-validation can tune decoders' parameters while avoiding circularity bias. However we find that it can be favorable to use sane defaults, in particular for non-sparse decoders.

1. Introduction: decoding needs model evaluation

Decoding, i.e. predicting behavior or phenotypes from brain images or signals, has become a central tool in neuroimaging data processing (Haynes and Rees, 2006; Haynes, 2015; Kamitani and Tong, 2005; Norman et al., 2006; Varoquaux and Thirion, 2014; Yarkoni and Westfall, 2016). In clinical applications, prediction opens the door to diagnosis or prognosis (Mouro-Miranda et al., 2005; Fu et al., 2008; Demirel et al., 2008). To study cognition, successful prediction is seen as evidence of a link between observed behavior and a brain region (Haxby et al., 2001) or a small fraction of the image (Kriegeskorte et al., 2006). Decoding power can test if an encoding model describes well multiple facets of stimuli (Mitchell et al., 2008; Naselaris et al., 2011). Prediction can be used to establish what specific brain functions are implied by observed activations (Schwartz et al., 2013; Poldrack et al., 2009). All these applications rely on measuring the predictive power of a decoder.

Assessing predictive power is difficult as it calls for characterizing the decoder on prospective data, rather than on the data at hand. Another challenge is that the decoder must often choose between many different estimates that give rise to the same prediction error on the data, when there are more features (voxels) than samples (brain images, trials, or subjects). For this choice, it relies on some form of regularization, that embodies a prior on the solution (Hastie et al., 2009). The amount of regularization is a parameter of the decoder that may require tuning. Choosing a decoder, or setting appropriately its internal parameters, are important questions for brain mapping, as these choices will not only condition the prediction performance of the decoder, but also the brain features that it highlights.

Measuring prediction accuracy is central to decoding, to assess a decoder, select one in various alternatives, or tune its parameters. The topic of this paper is cross-validation, the standard tool to measure predictive power and tune parameters in decoding. The first section is a primer on cross-validation giving the theoretical underpinnings and the current practice in neuroimaging. In the second section, we perform an extensive empirical study. This study shows that cross-validation results carry a large uncertainty, that cross-validation should be performed on full blocks of correlated data, and that repeated random splits should be preferred to leave-one-out. Results also yield guidelines for decoder parameter choice in terms of prediction

* Corresponding author at: Parietal project-team, INRIA Saclay-Île de France, France.
<http://dx.doi.org/10.1016/j.neuroimage.2016.10.038>
 Received 3 August 2016; Accepted 24 October 2016
 Available online 29 October 2016
 1053-8119/© 2016 Elsevier Inc. All rights reserved.

How to read/write articles on ML for MIA

Two strategies for previous work

- Write longer introduction with an elaborated previous work summary
- Or, write individual section named “previous work”. In brain decoders paper, the authors used has an independent section
- Write shorter introduction with previous work summary
- Just mention the main issues in the existing work that you try to address
- Write an elaborated previous work summary in the discussion by methodologically comparing your work against existing research

Articles on ML for MIA: Data

Three strategies for describing Data

Start of Methodology

- Not the best strategy, breaks the flow of the text

Start of Results

- The most common strategy. Results are also difficult to populate without Data description

Start of Methodology and Results

- Methodology will describe how the data was collected
- Results will describe what was the output of data collecting protocol

Articles on ML for MIA: Data

Data:

- How big is the database, statistical power?
- Is the data public?

3.1. Experiments on real neuroimaging data

A variety of decoding datasets. To achieve reliable empirical conclusions, it is important to consider a large number of different neuroimaging studies. We investigate cross-validation in a large number of 2-class classification problems, from 7 different fMRI datasets (an exhaustive list can be found in [Appendix E](#)). We decode visual stimuli within subject (across sessions) in the classic Haxby dataset [16], and across subjects using data from [9]. We discriminate across subjects *i*) affective content with data from [53], *ii*) visual from narrative with data from [34], *iii*) famous, familiar, and scrambled faces from a visual-presentations dataset [19], and *iv*) left and right saccades in data from [23]. We also use a non-published dataset, ds009 from openfMRI [41]. All the across-subject predictions are performed on trial-by-trial response (Z-score maps) computed

Articles on ML for MIA: Results and Discussion

Results may contain:

- Database description
- Parameter description
- Result metrics

Discussion may contain:

- Methodological comparison to previous work
- Explanation of the results

Questions?