Introduction to Open & Reproducible Science (IORDS)

Michael Dayan, Data Scientist Manager

Methods & Data facility
Human Neuroscience Platform
Foundation Campus Biotech Geneva

Virtual machine info

To get an IP, please fill the form at:
https://tinyurl.com/IORDS2021-IP-ML2

START VS CODE AND JUPYTER ON YOUR BROWSER:

- *Start an internet browser on your own machine*
- *Jupyter:* `<your_IP>:8888`

PASSWORD: braincode!

Connecting your:	WIFI SSID	WIFI Password
Laptop (no phones)	NIDS_course	reproduciblescience
Phone	CAMPUS_VISITORS	welcomecampus

On site support (including coding):

Maël

Nathan

Remote support (including coding):

Serafeim

ANY PROBLEM? Please raise your hand or ask questions on Slack: channel **#machine-learning**

HNP METHODS & DATA

LECTURE OBJECTIVES

Introduction to machine learning lectures objectives (you should be able to…):

- Understand the typical form of the data characterizing a machine learning problem (N samples x p features, with p labels for supervised machine learning)
- Understand what fitting a model means
- Understand how to score predictions and why an independent test set is essential
- Know how to implement cross-validation, and know why this is useful

ML Part 1

- Understand the goal of classification and how performance is measured
- Understand the concept of overfitting and underfitting
- Understand the principles of regularization and it could be implemented
- Understand the main idea behind dimensionality reduction
- Understand the goal of clustering, how the performance can be measured, and how it is implemented with k-means

ML Part 2

MAKING PREDICTIONS

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

.EAST BAD" MODEL?

Fitted model

$r = \hat{\beta}_{\Psi}\ \Psi + \hat{\beta}_{0}$

X → y_{pred}

Test score (e.g. MSE)

ORIGINAL TRAINING DATASET

X: Features — y: Labels

Observations (data points / samples)

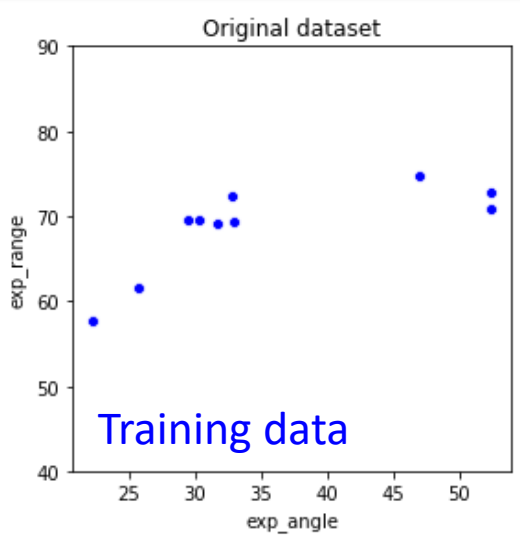
	Angle Ψ (psi)	Range r
Shot 1	60	32.4
Shot 2	45	36.2
Shot 3	45	37.7
…	…	…

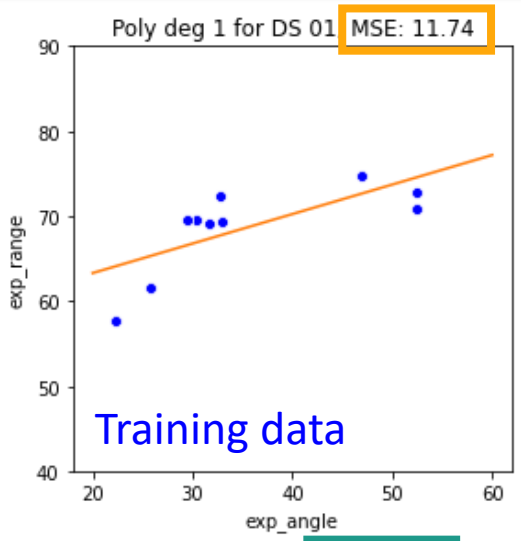
NEW TESTING DATASET

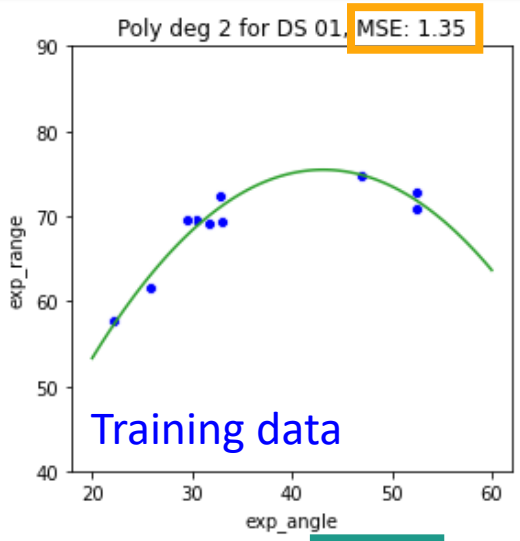
Features — Labels — y

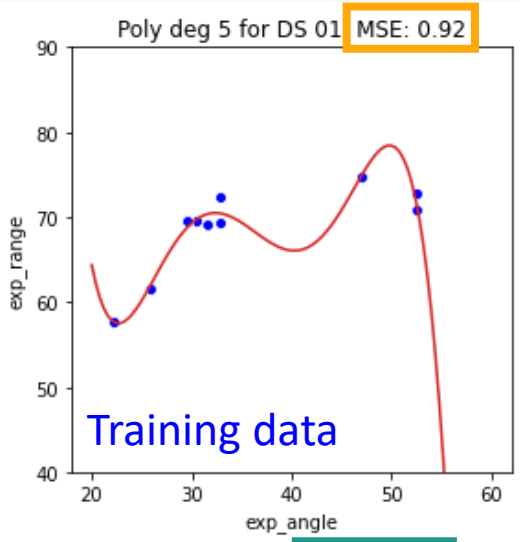
	Angle Ψ (psi)	Range r
Shot 1	37	33.1
Shot 2	24	29.6
Shot 3	54	41.4
…	…	…

MAKING PREDICTIONS

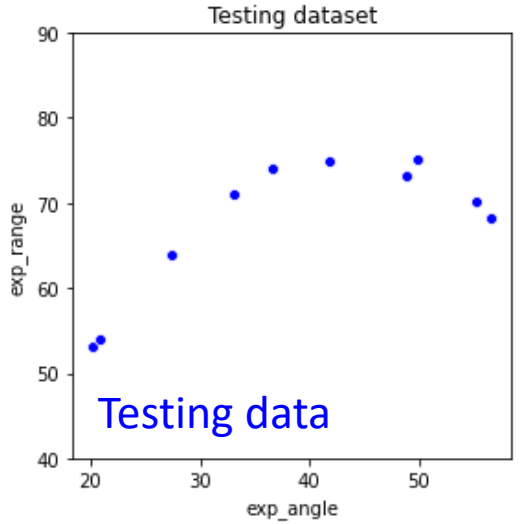


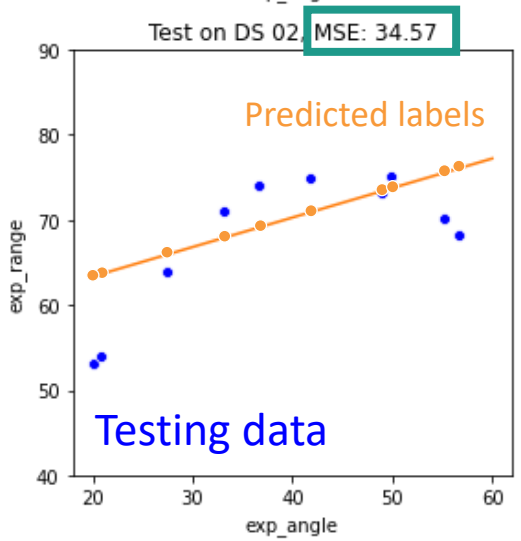


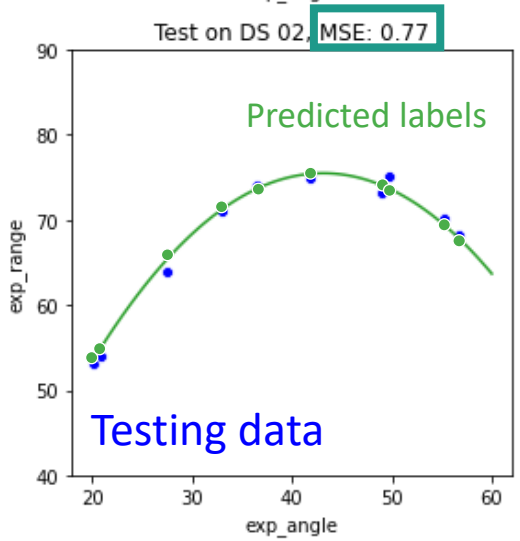


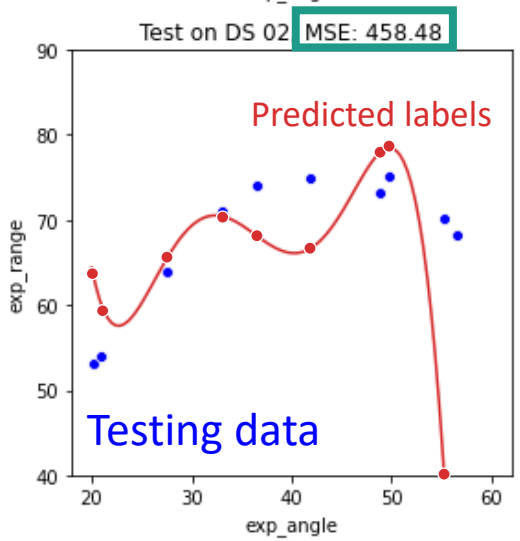


Model fitted on the training data







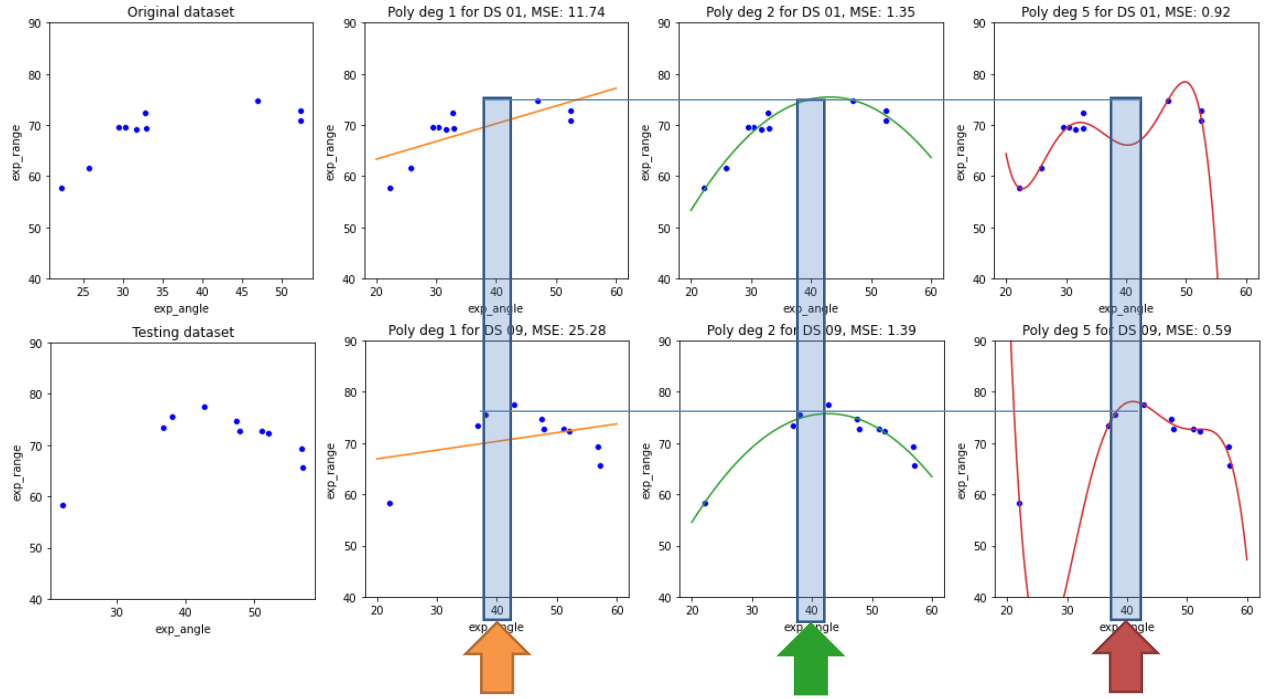


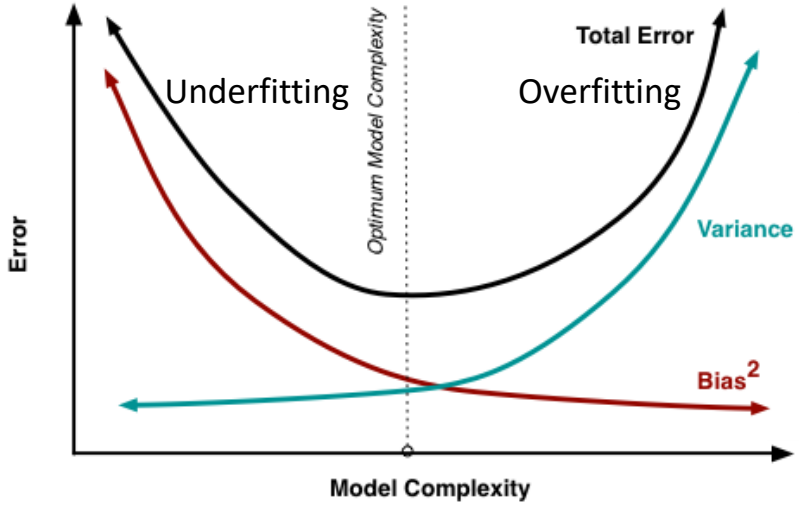
Fitted model used to predict unseen data

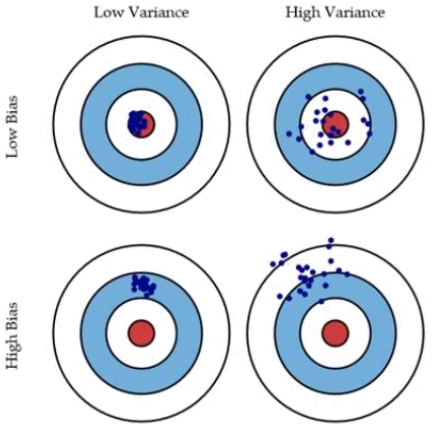
→ comparison between predictions and true labels provides perf. score

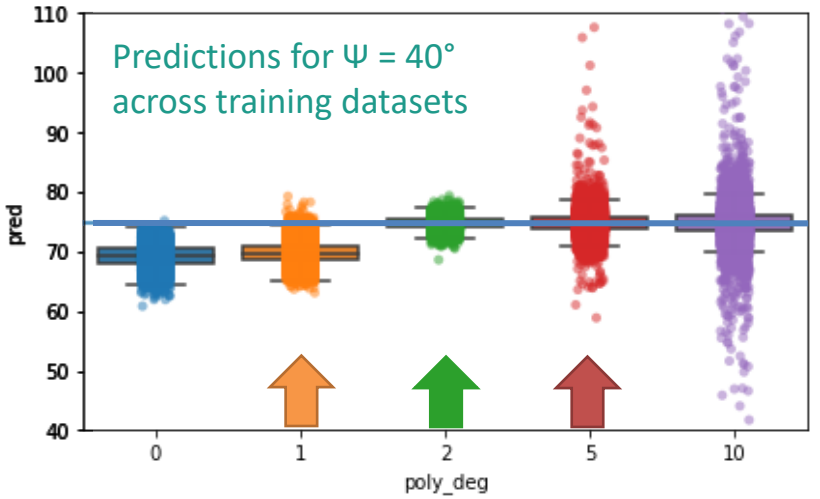
MAKING PREDICTIONS

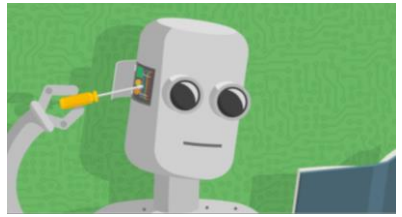
- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions







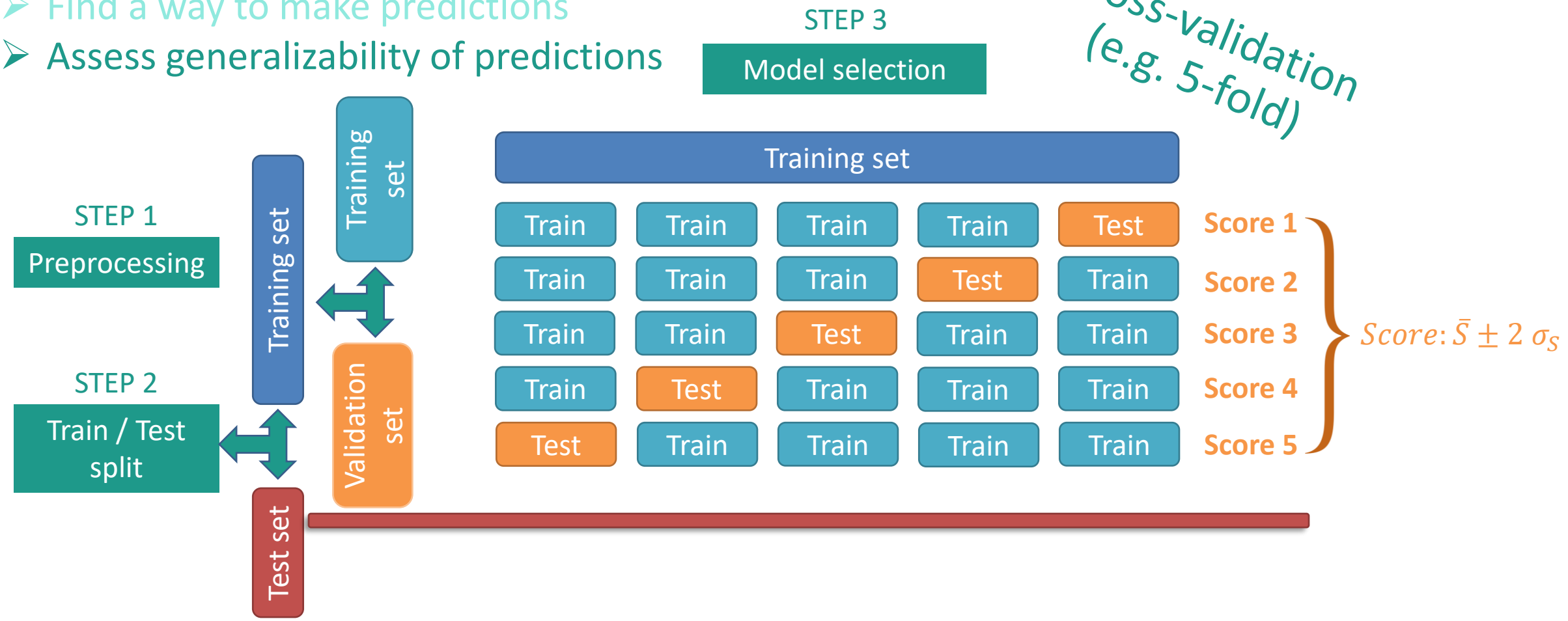




Ground Truth

GENERALIZABILITY OF ML MODELS (& MODEL SELECTION)

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions



SUPERVISED LEARNING: REGRESSION & CLASSIFICATION

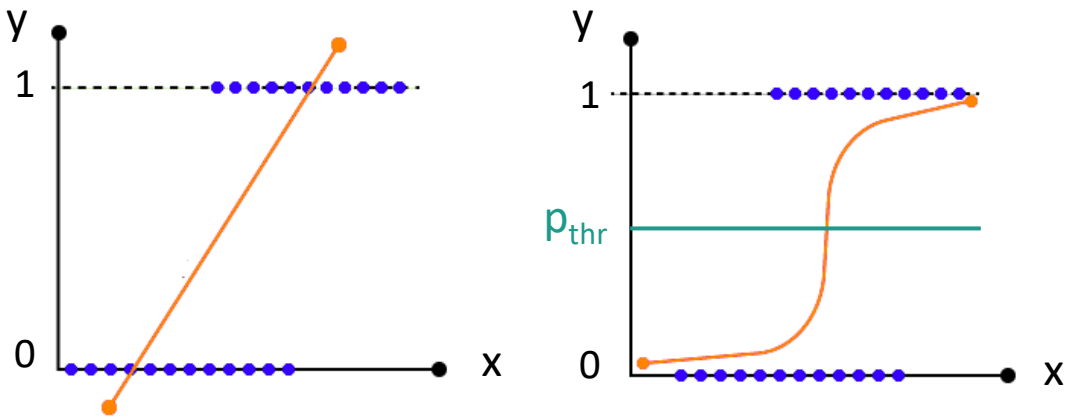
➢ Supervised learning

When the model is associating features to a label, the task is called supervised learning (the labels provided by the user is the "supervision")

Depending if the label is continuous (e.g. house price) or categorical (e.g. diagnosis), the supervised learning task is called differently:

- **regression** when the label is continuous
- **classification** when the label is categorical

➢ Example of classification algorithm: logistic regression



Many classification algorithm output a probability value p between 0 and 1

➔ have to choose threshold p_{thr} to define labels, e.g. $p_{thr} = 0.5$:

- if y <= 0.5 then label is 0
- If y > 0.5 then label is 1

There can be reasons to choose $p_{thr} \neq 0.5$ (e.g when focus on avoiding false positives or false negatives)

CLASSIFICATION EVALUATION

➢ Classification scores

		Actual	
		Positive	Negative
Predicted	Positive	**True positive**	**False positive**
	Negative	**False negative**	**True negative**

Accuracy = (TP + TN) / (TP + FP + FN + TN)

Sensitivity (or TPR) = TP / (TP+FN)

False Positive Rate (FPR) = FP / (FP+TN)

Specificity = TN / (FP+TN)

Precision (PPV) = TP / (TP+FP)

→ **Depends on which probability threshold you choose to define positive cases**
p=0.5? p=0.95?

The choice of outcome(s) (and features) is fully part of the research design

DATASET

Feature matrix X | Label array y

Observations (data points / samples)

	Age	Sex	ROI 1	…	Has disease
Subj 1	60	F	42.0	…	No
Subj 2	45	M	29.1	…	Yes
Subj 3	45	F	31.7	…	No
Subj 4	35	F	25.4	…	Yes
…	…	…	…	…	…

CLASSIFICATION EVALUATION

- Classification scores

		Actual	
		Positive	Negative
Predicted	Positive	True positive	False positive
	Negative	False negative	True negative

Accuracy = (TP + TN) / (TP + FP + FN + TN)

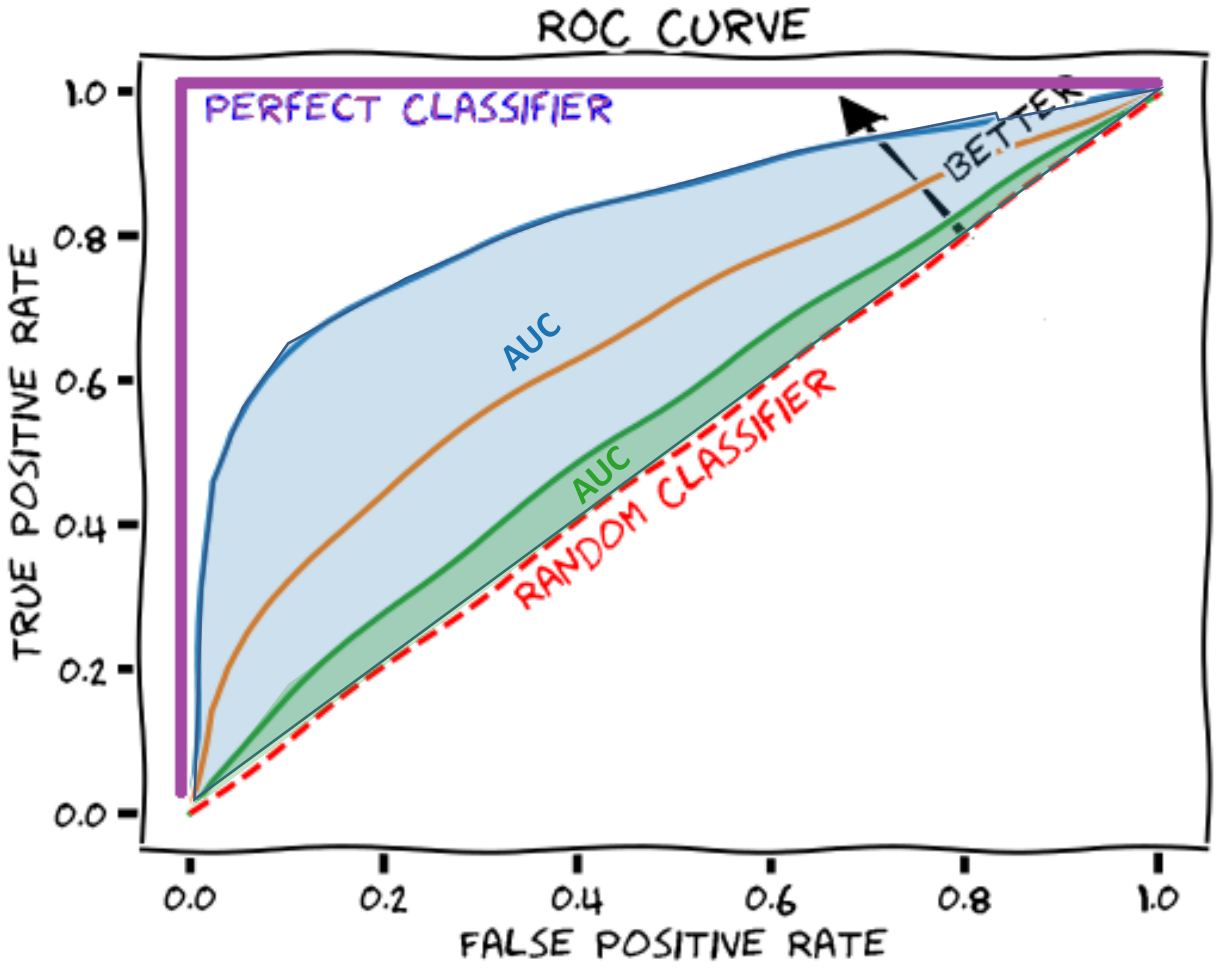
Sensitivity (or TPR) = TP / (TP+FN)

False Positive Rate (FPR) = FP / (FP+TN)

Specificity = TN / (FP+TN)

Precision (PPV) = TP / (TP+FP)

→ **Depends on which probability threshold you choose to define positive cases**
p=0.5? p=0.95?

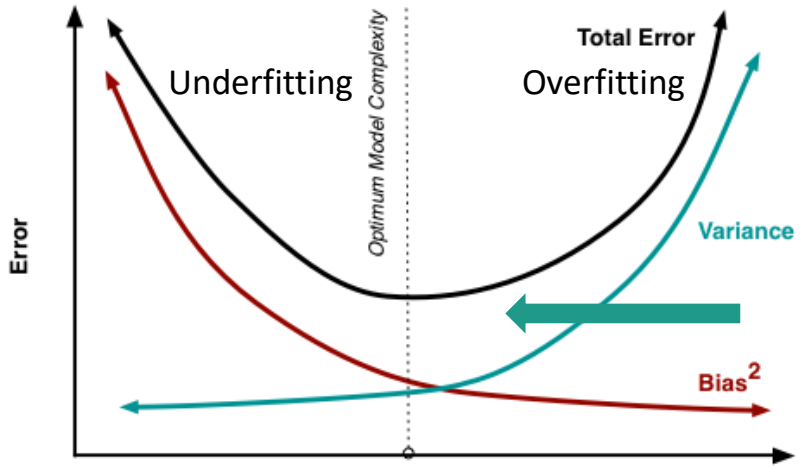


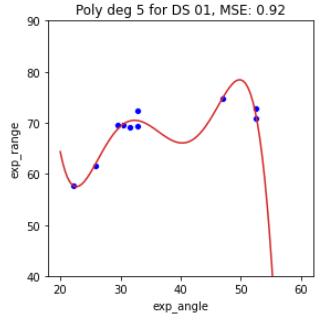
REGULARIZATION

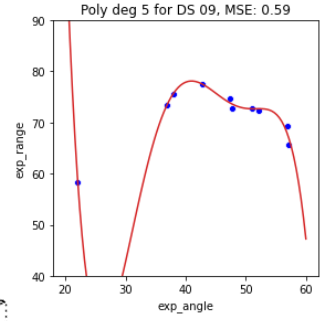
➢ How to prevent overfitting?

- Feature selection
- Regularization (in linear regression), forcing model parameters (β coefficients) to be small or zero

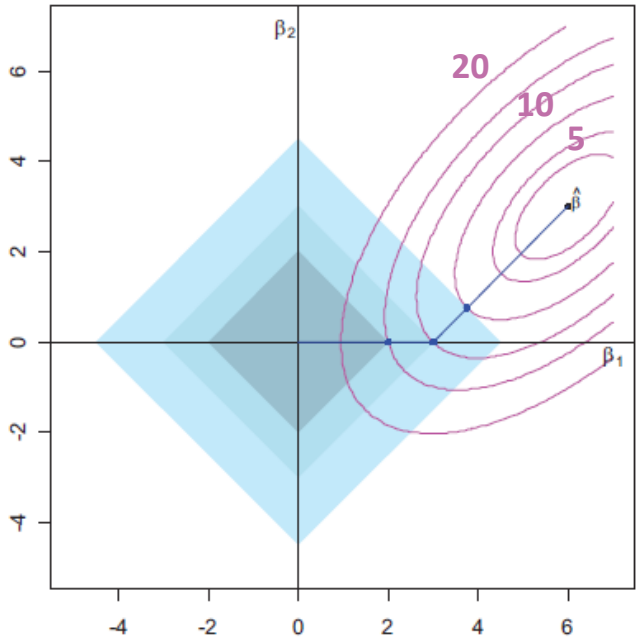
λ is an hyper-parameter (not fitted)



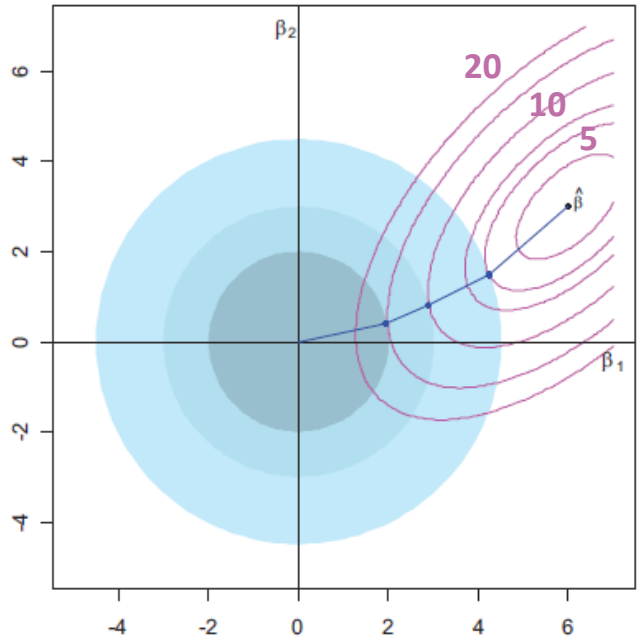


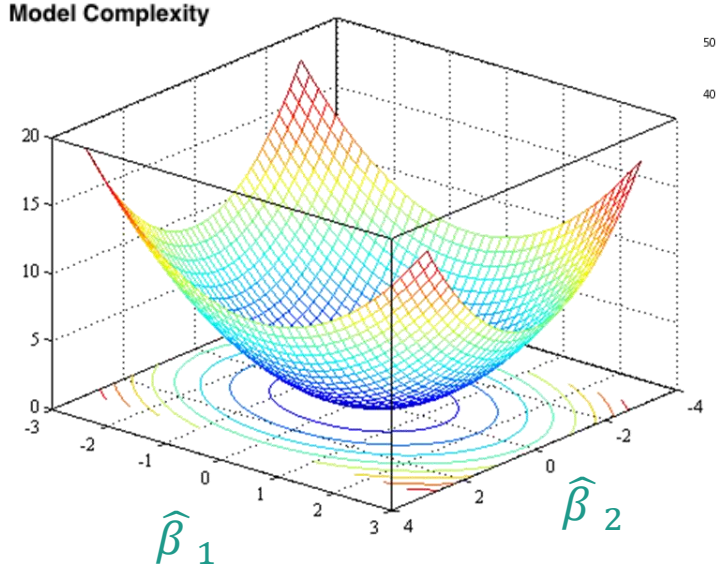


$$L_{lasso} = \sum_{i=0}^{n}(y_i - \hat{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{i=0}^{p}|\hat{\beta}_j|$$



$$L_{ridge} = \sum_{i=0}^{n}(y_i - \hat{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{i=0}^{p}\hat{\beta}_j{}^2$$





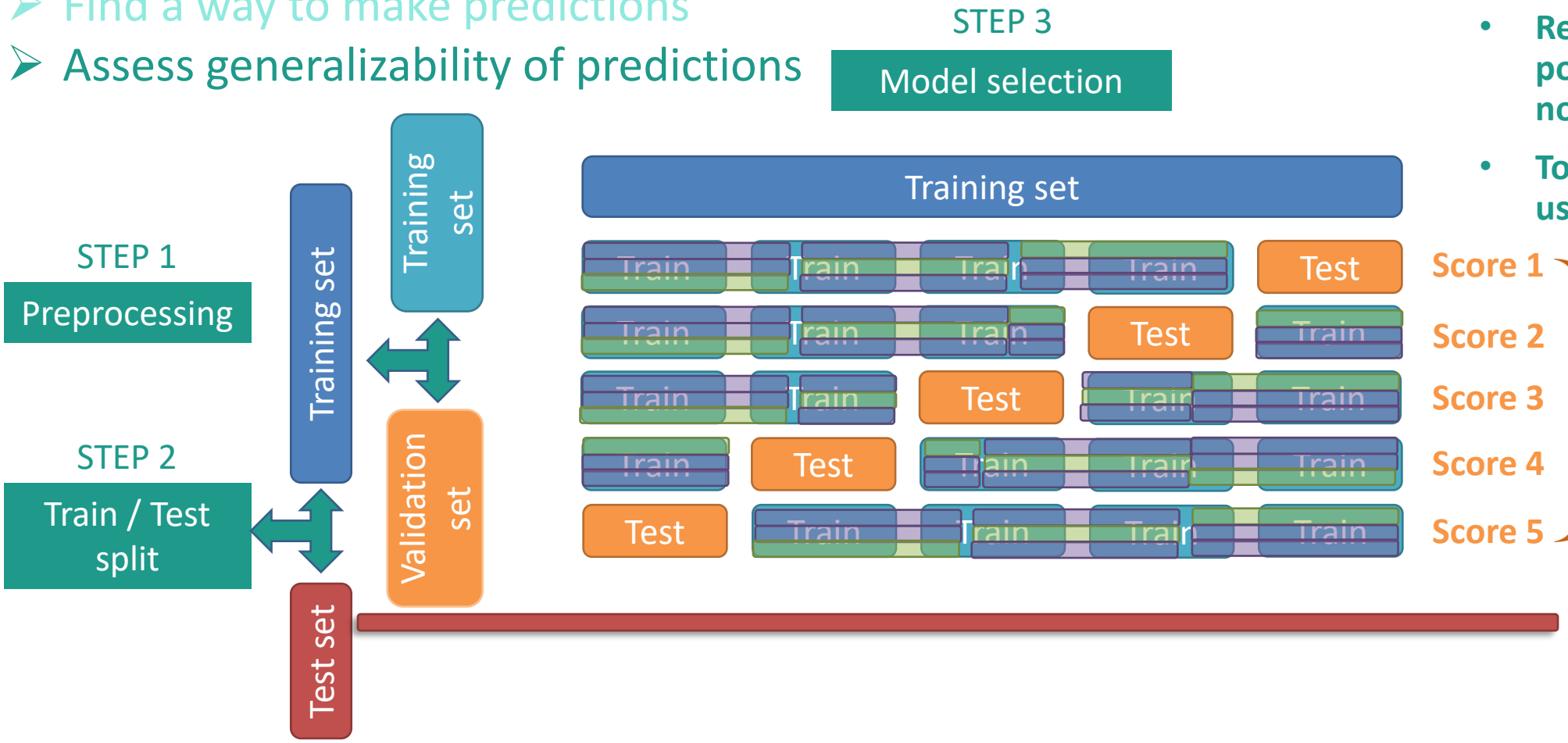
→ **Need to find best hyper-parameter**

CHOOSING HYPERPARAMETERS WITH NESTED CROSS-VALIDATION

- Take measurements
- Find a way to make predictions
- Assess generalizability of predictions

→ **How to find best hyper-parameter λ (e.g. among λ_1, λ_2 and λ_3)?**

- **Repeating CV for each possible λ possible but not ideal**
- **To assess generalizability, use nested CV**



One can investigate combinations of hyper-parameter values in a "grid-search"

OTHER EXAMPLE OF CLASSIFICATION MODEL: SVM

$$\text{margin} = \frac{1}{\|w\|}$$

$$\underset{\mathbf{w},b}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to:} \quad y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1 \quad i = 1, \ldots, n\,.$$

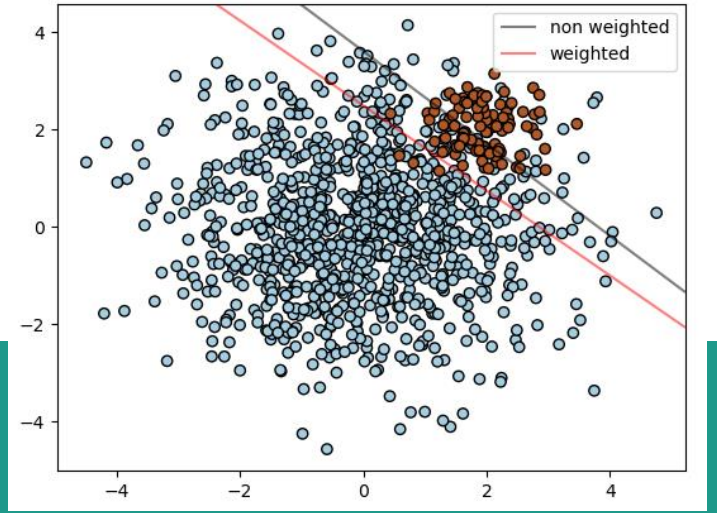
$$\underset{\mathbf{w},b}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

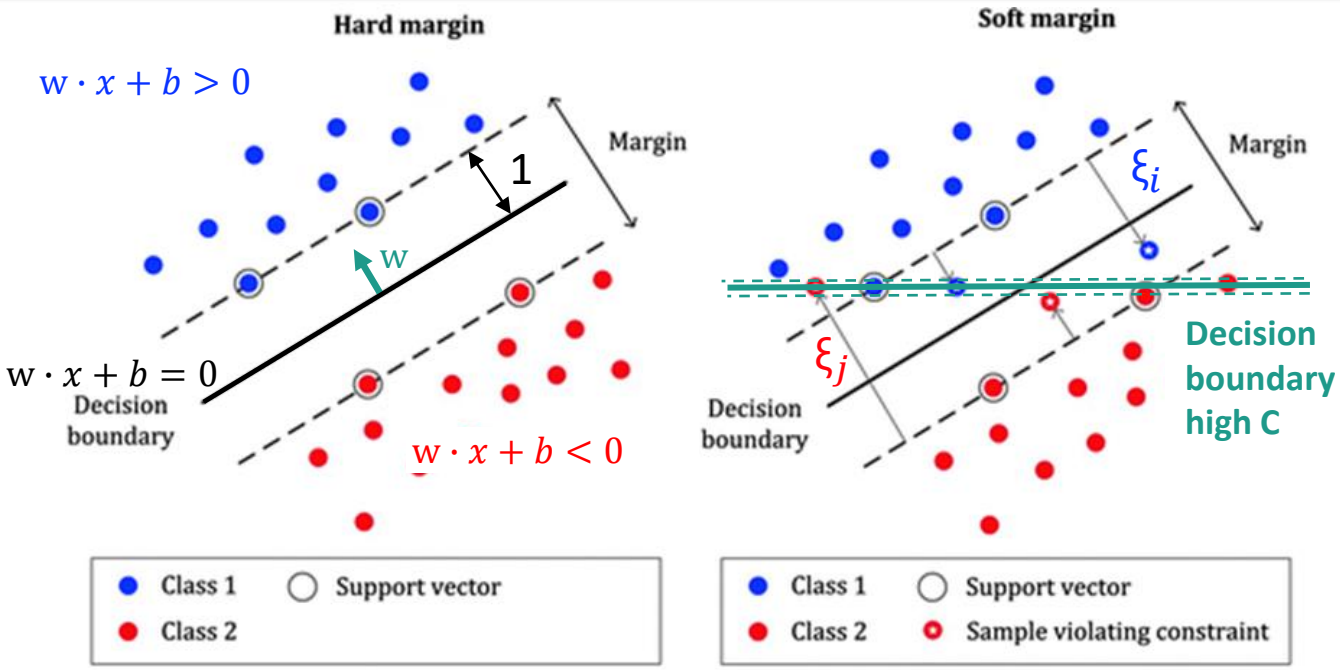
$$\text{subject to:} \quad y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

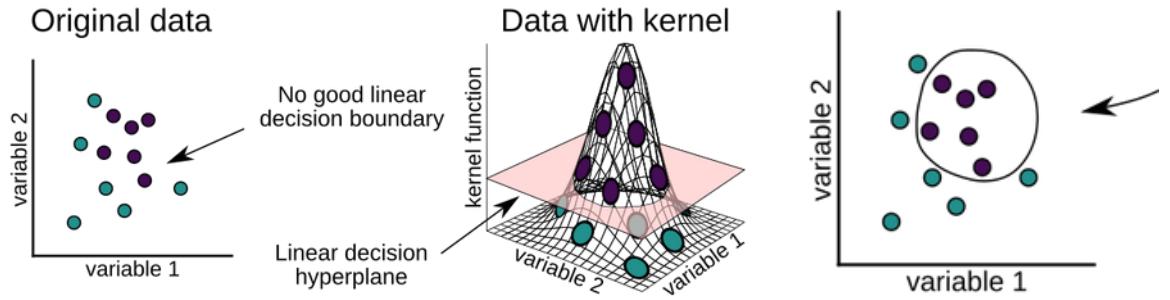
$$C\sum_{i=1}^{n}\xi_i \longrightarrow C_+\sum_{i\in I_+}\xi_i + C_-\sum_{i\in I_-}\xi_i$$

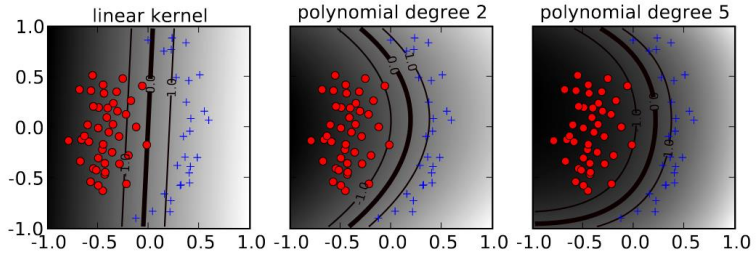
$$C_+ n_+ = C_- n_-$$

$$\frac{C_+}{C_-} = \frac{n_-}{n_+}.$$

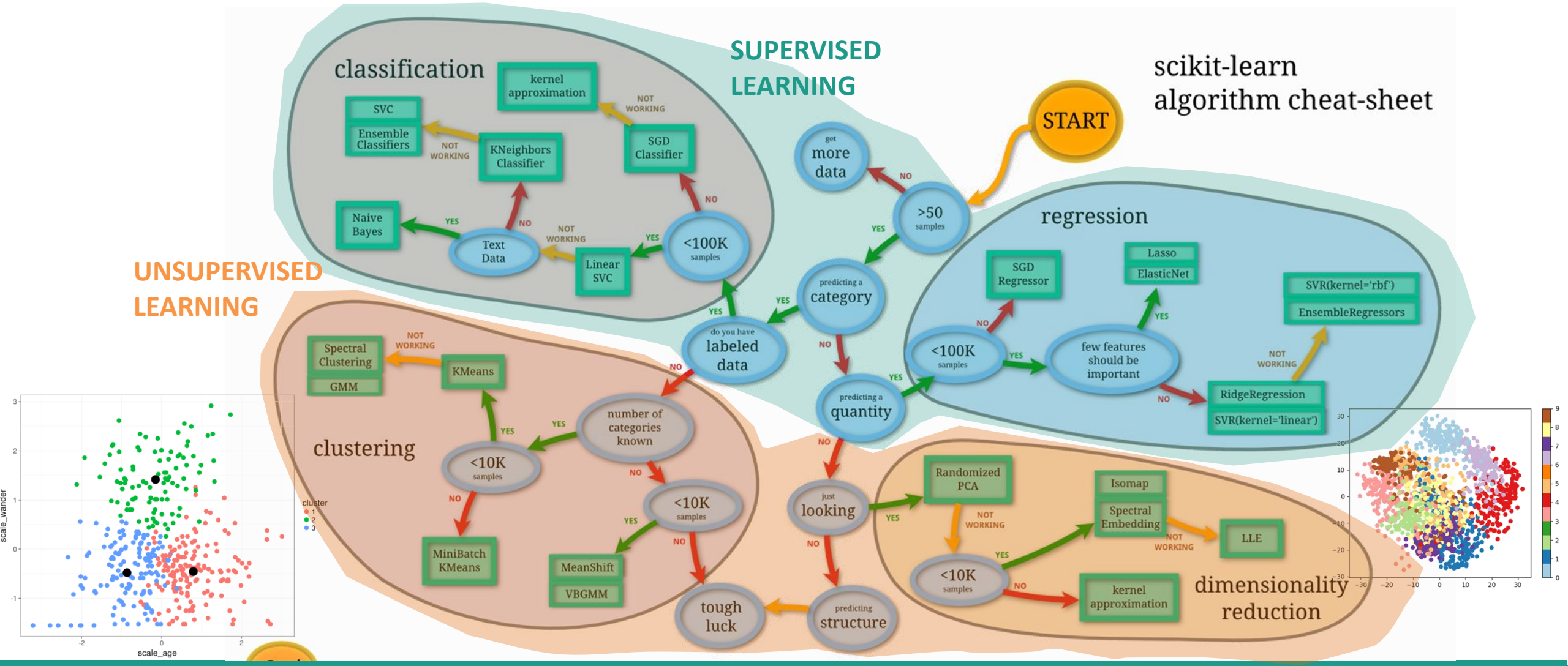








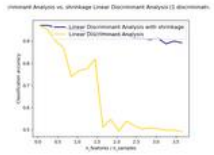
MACHINE LEARNING ESTIMATORS



THANK YOU FOR YOUR ATTENTION!

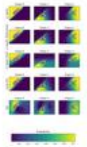
Classification

General examples about classification algorithms.



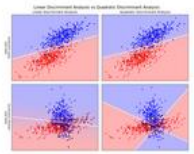
Normal and Shrinkage Linear Discriminant Analysis for classification

Recognizing hand-written digits



Plot classification probability

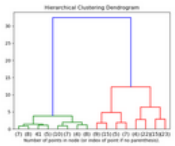
Classifier comparison



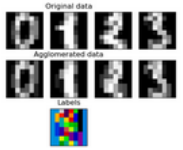
Linear and Quadratic Discriminant Analysis with covariance ellipsoid

Clustering

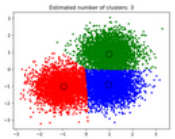
Examples concerning the `sklearn.cluster` module.



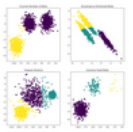
Plot Hierarchical Clustering Dendrogram



Feature agglomeration



A demo of the mean-shift clustering algorithm



Demonstration of k-means assumptions

COURSE SUPPORT

SLACK (iords2021.slack.com)

- Course main channel: `#general`
- Topic channels: `#linux`, `#linux-capstone`, `#git`, `#git-capstone`, `#python`, `#full-example`, `#machine-learning`

→ Check regularly for course info (esp. pinned items)

→ Do not hesitate to ask questions (please reply "in thread")

1-to-1 OFFICE HOURS for course questions:

- 20-min slots every Friday morning between 9AM and 11AM

→ Book a time slot here: https://tinyurl.com/IORDS-office-hours

→ Do not hesitate to ask any kind of question, this is a beginner course !

EMAIL: methods@fcbg.ch ← Please whitelist!


Thank You!
Michael Dayan: methods@fcbg.ch
H N P
METHODS & DATA
campus biotech
Fondation Campus Biotech Geneva