

Sampling and overfitting

Formation IA biodiversité

Paul Tresson

May 15, 2025

UMR AMAP

Introduction

What do we want when modelling ?

- Understand things

What do we want when modelling ?

- Understand things
- **Predict things**

What do we want when modelling ?

“All models are wrong, but some are useful”

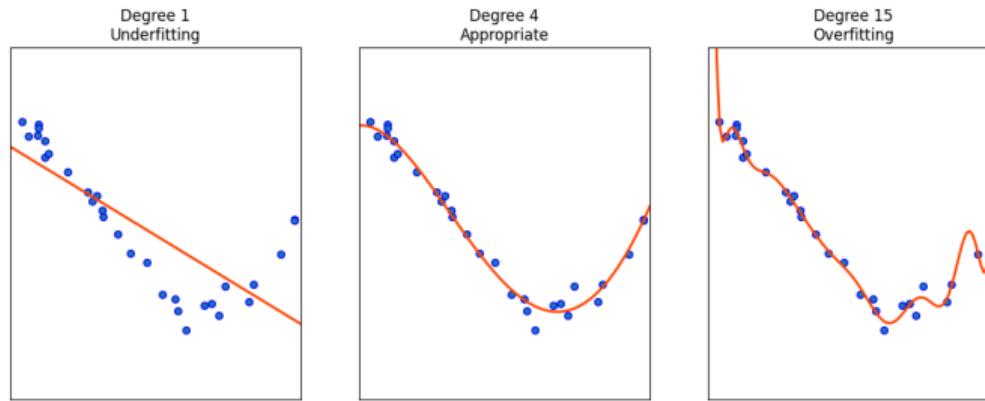
George E. P. Box

What do we want when modelling ?

- **Robustness:** Useful when mistakes
- **Generalization:** Useful applied elsewhere

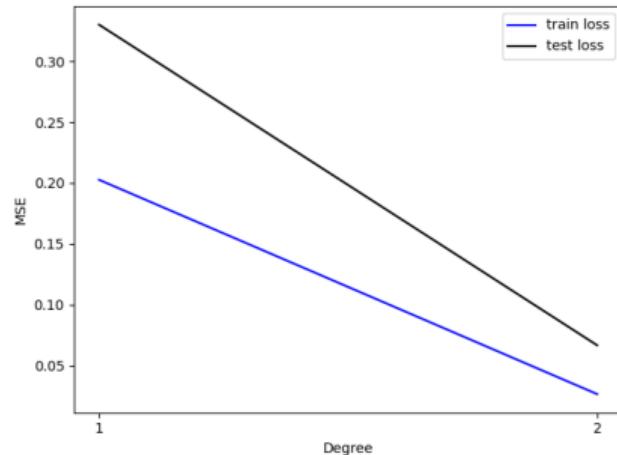
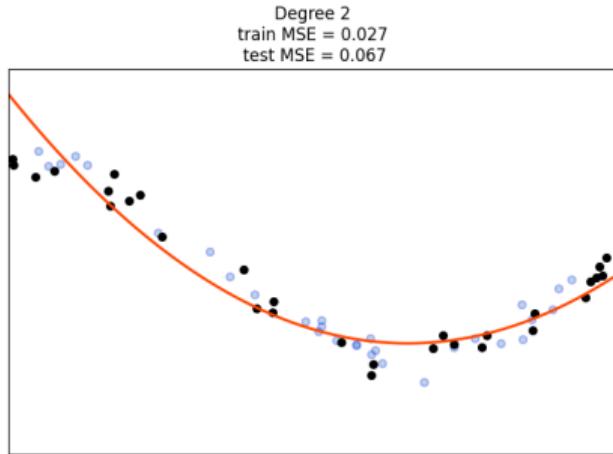
Overfitting

What is overfitting

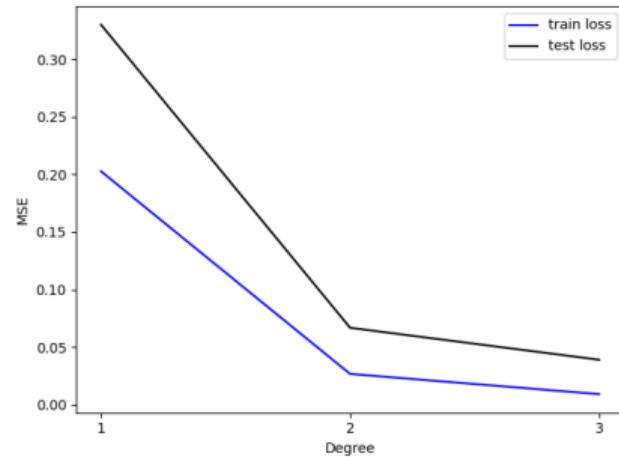
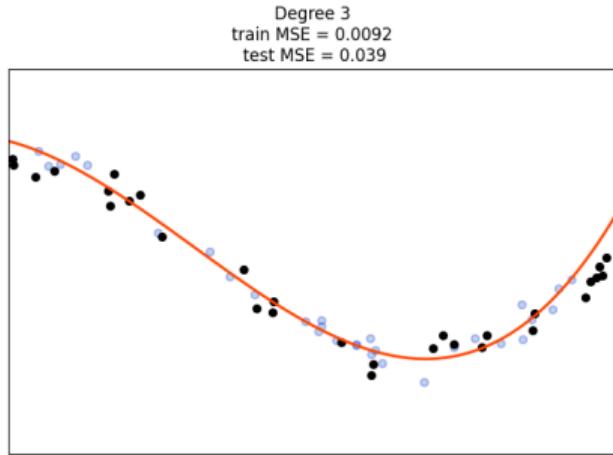


adapted from scikit-learn docs

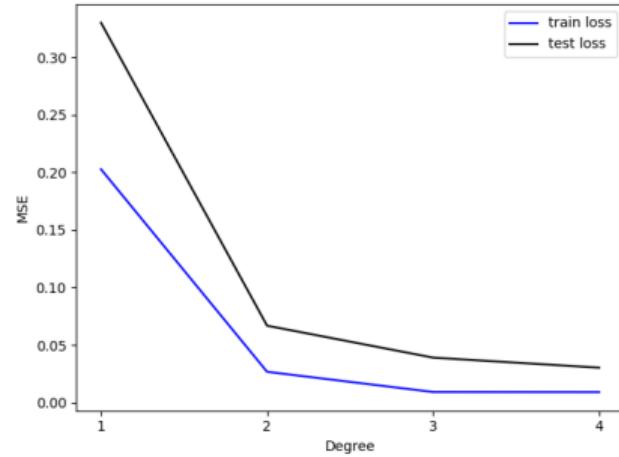
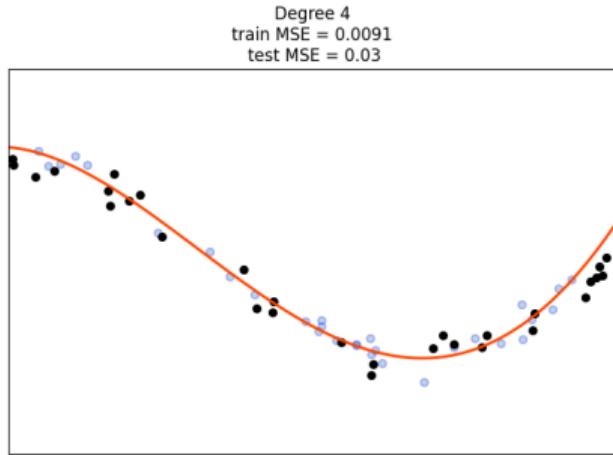
Common tools and intuitions - Train/Test loss



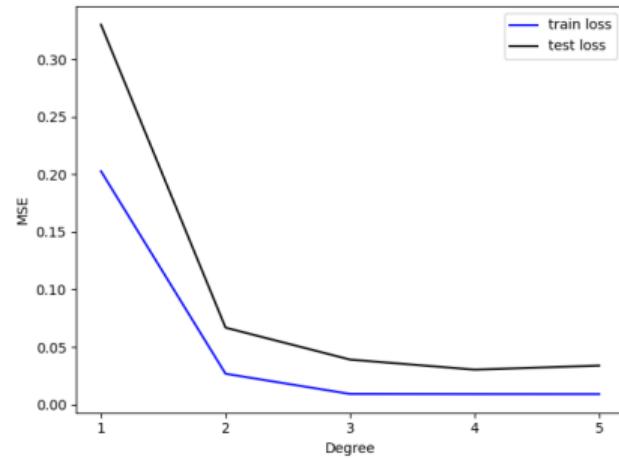
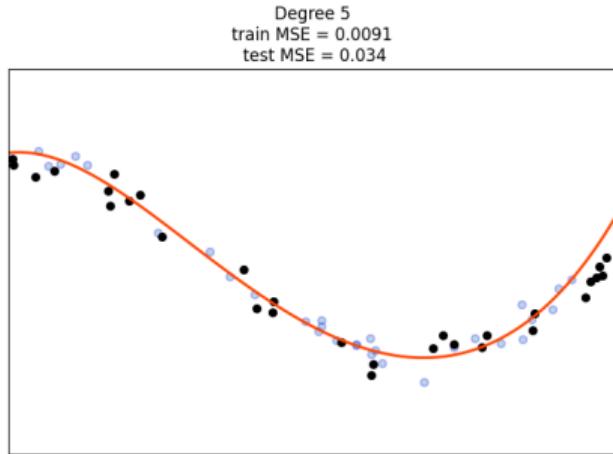
Common tools and intuitions - Train/Test loss



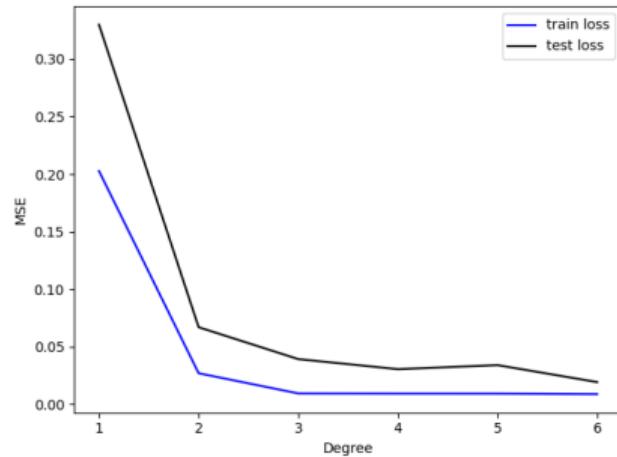
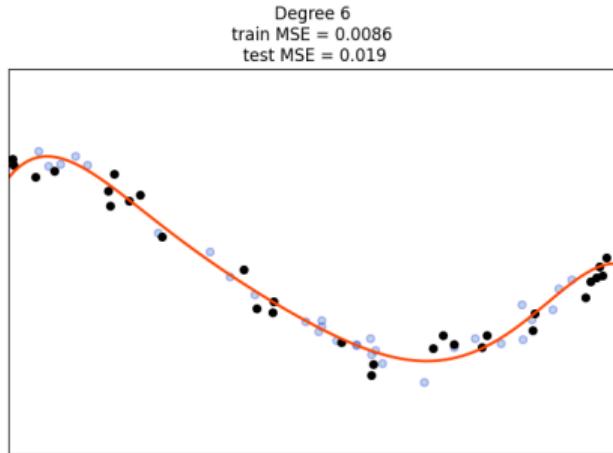
Common tools and intuitions - Train/Test loss



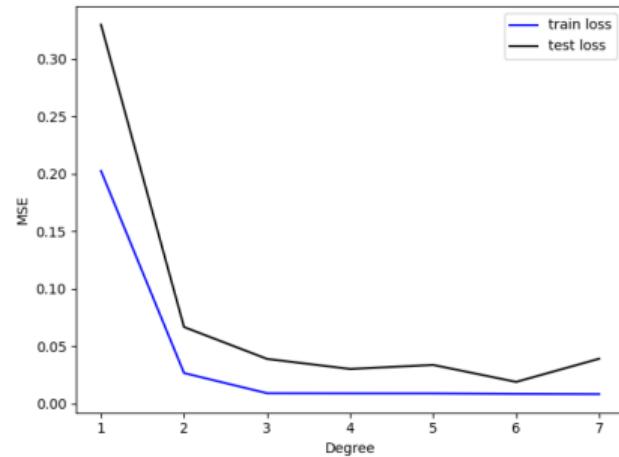
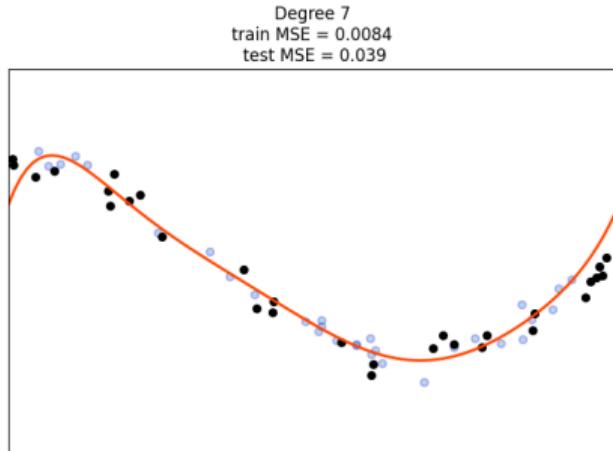
Common tools and intuitions - Train/Test loss



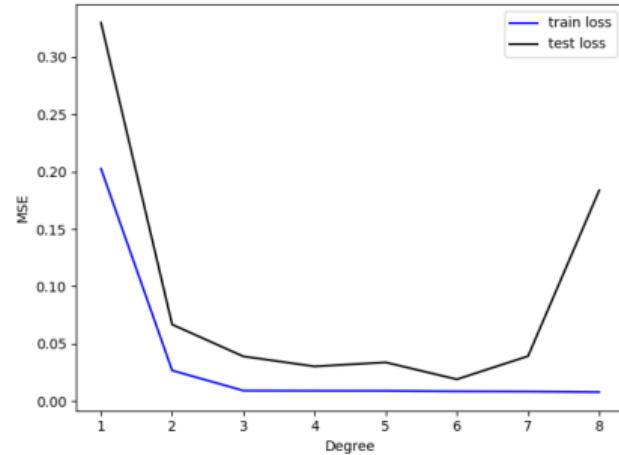
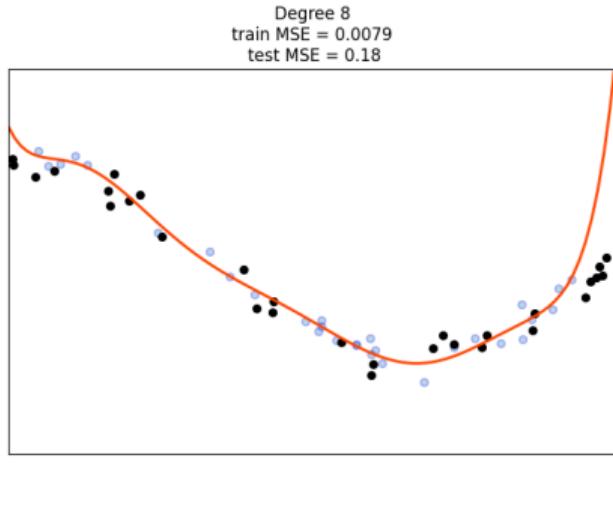
Common tools and intuitions - Train/Test loss



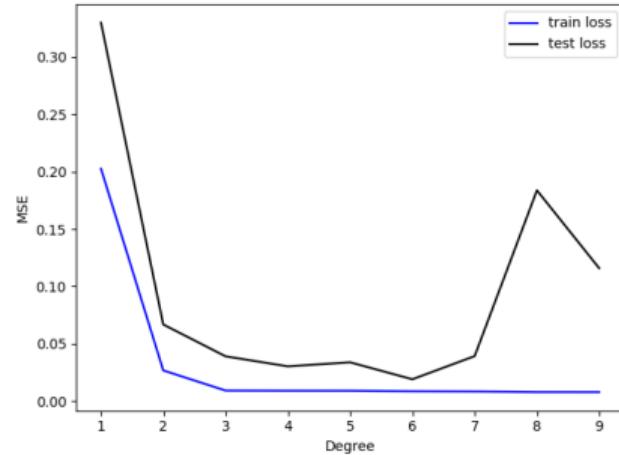
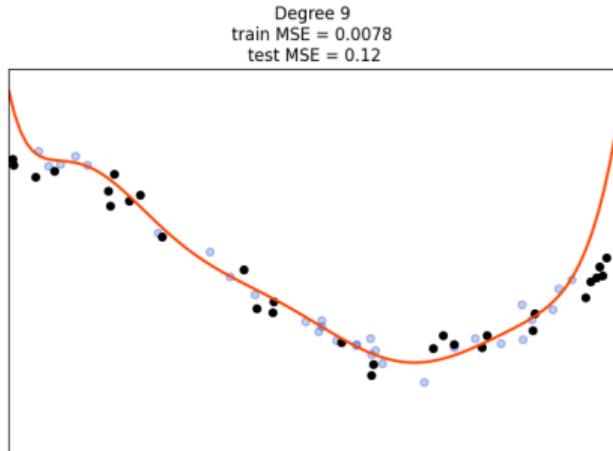
Common tools and intuitions - Train/Test loss



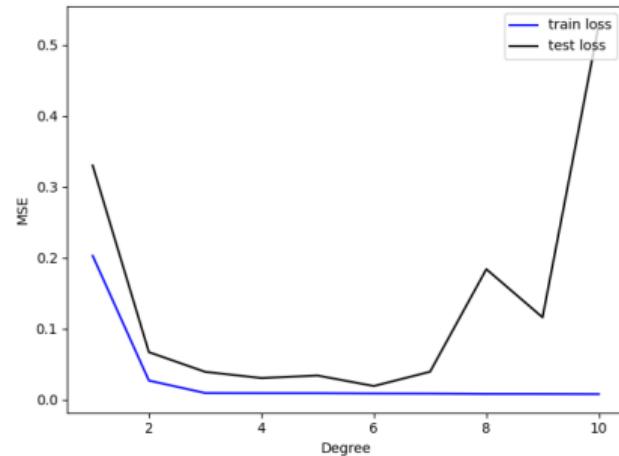
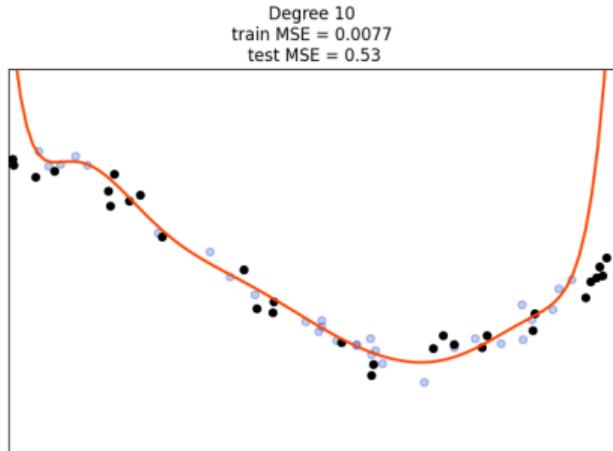
Common tools and intuitions - Train/Test loss



Common tools and intuitions - Train/Test loss



Common tools and intuitions - Train/Test loss



Common tools and intuitions - Train/Test loss

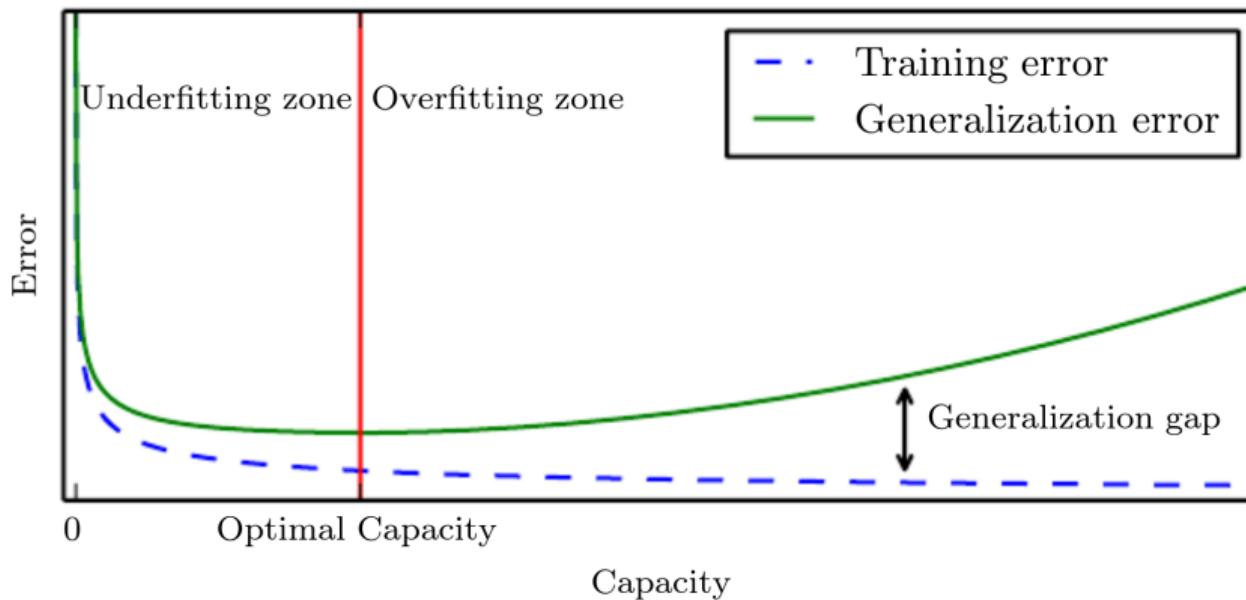


Figure from Goodfellow et al., 2016

Common tools and intuitions - AIC/BIC

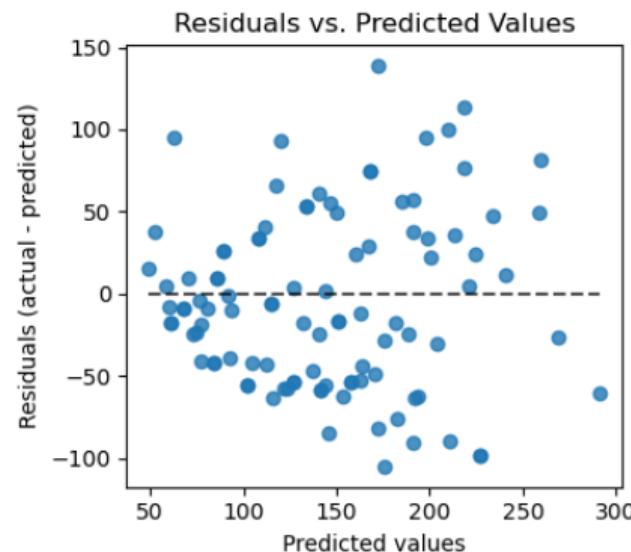
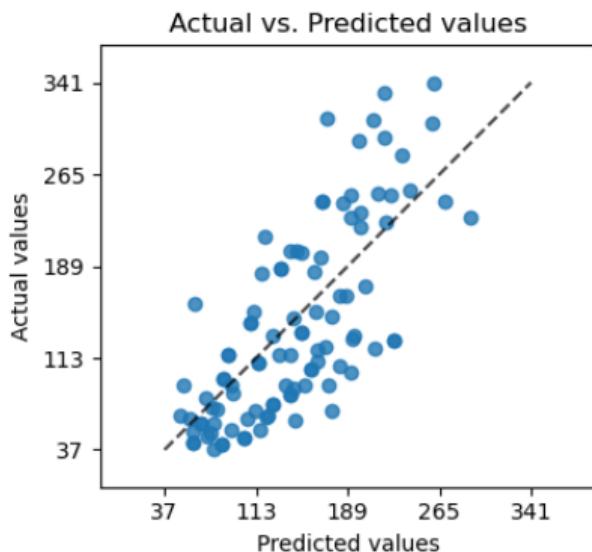
Akaike information criterion (AIC)

Bayesian information criterion (BIC)

Is the model parameter efficient ?

Common tools and intuitions - Biases

Plotting cross-validated predictions



from scikit-learn docs

And in Machine(/Deep) Learning ??

How many parameters to have

Shrek learning botany starting from random noise ?

And in Machine(/Deep) Learning ??



$\approx 2.5B$?

Root Causes

Too many parameters

Root Causes

Too many parameters

Too little training data

Root Causes

Too many parameters

Too little training data

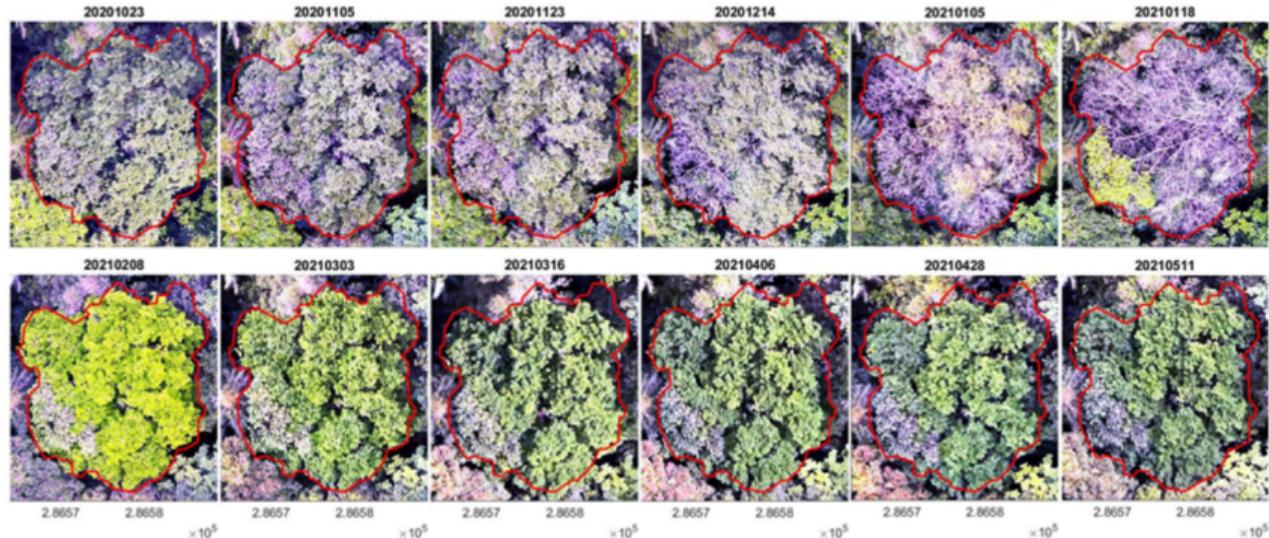
(bad) training data

Need to be very carefull on how to evaluate

Illustrated examples in Ecology

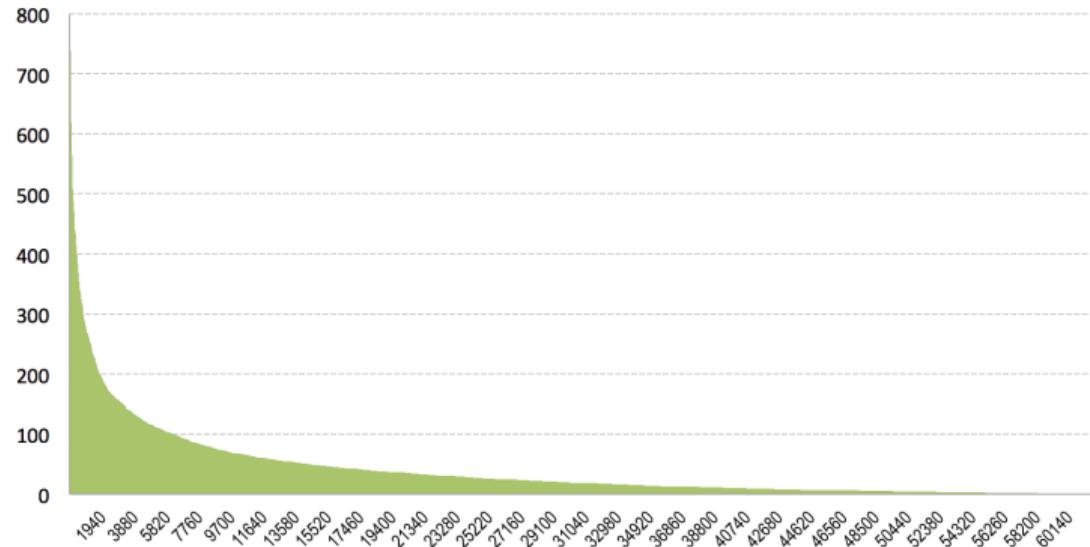
Constraints in ecology

Data from the real world is noisy,



Constraints in ecology

Data from the real world is noisy, unbalanced,



Constraints in ecology

Data from the real world is noisy, unbalanced, hard to collect,



Constraints in ecology

Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

Select all images with an Orange.

Verify

Constraints in ecology

Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

Select all images with an Orange.

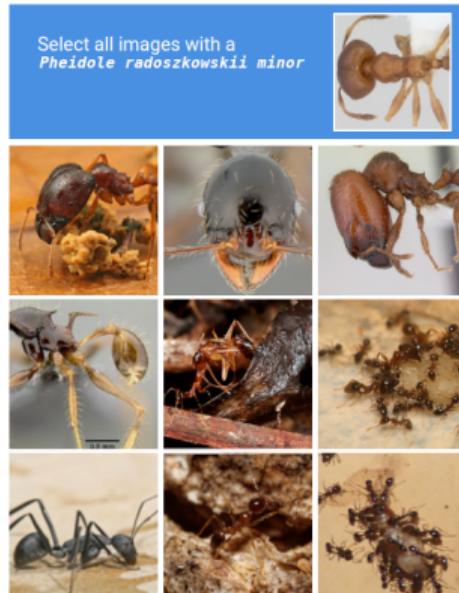
Verify

✖️ 🔍 ⓘ

Constraints in ecology

Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

Select all images with a
Pheidole radoszkowskii minor



The grid contains 12 images arranged in three rows of four. The first row has one large image on the right and three smaller images below it. The second and third rows each have four images. The images depict ants in various stages and environments, including close-ups of individual ants and wider shots of ant colonies on leaves or soil.

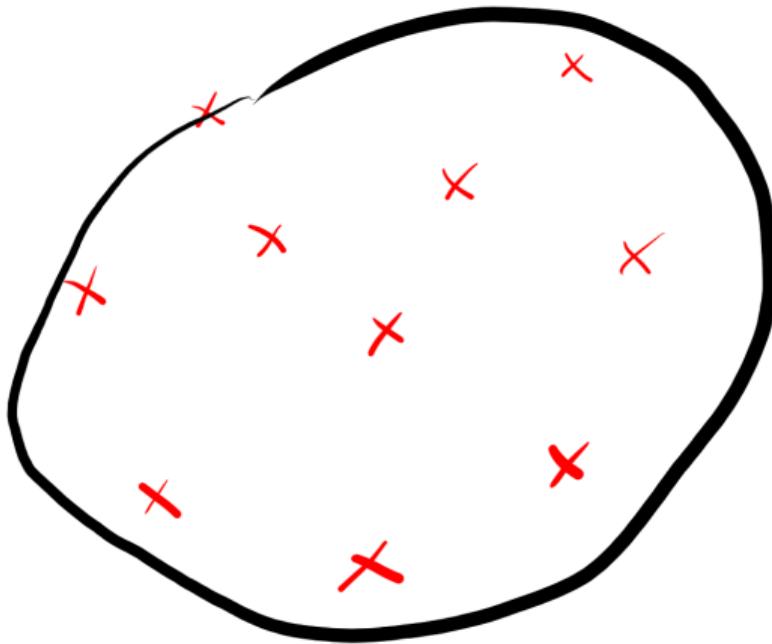
Verify



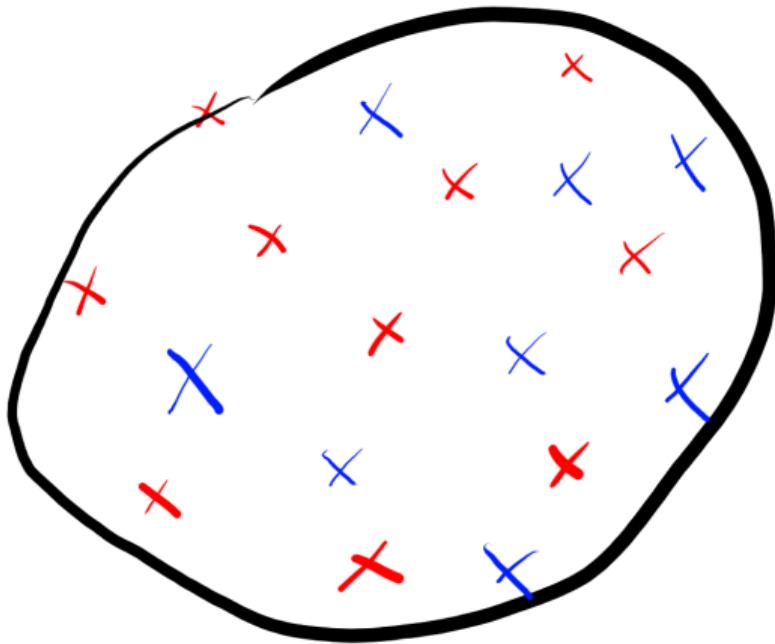
Verify



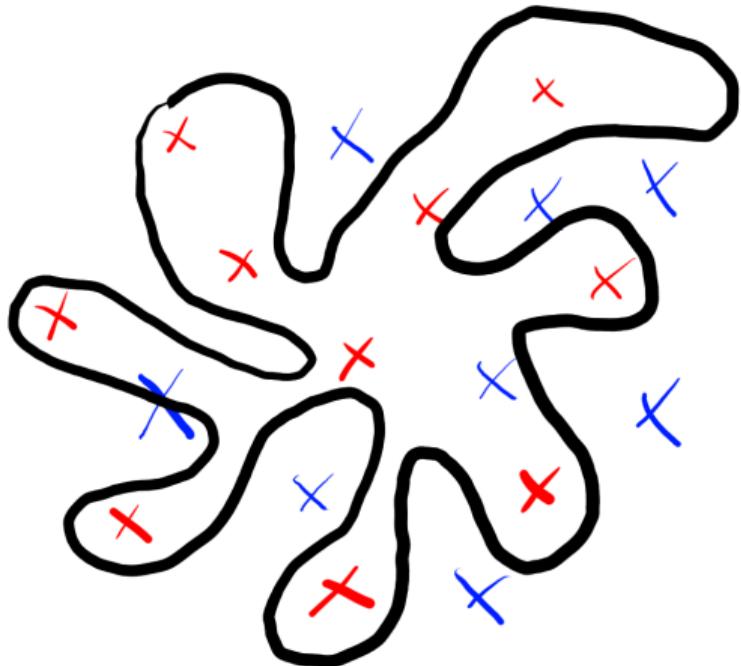
Train set



A good fitted model



Test set

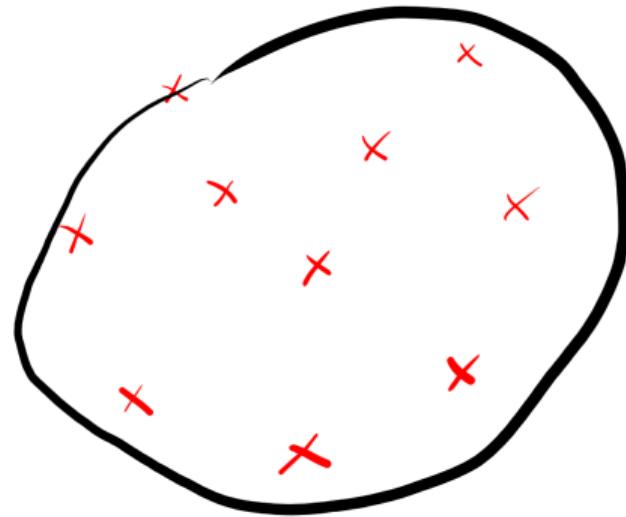


An overfitted model

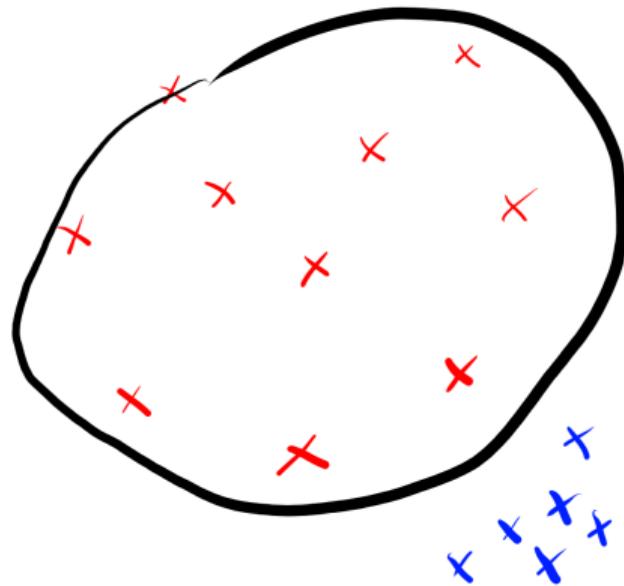
Biases in the train set



Biases in the train set



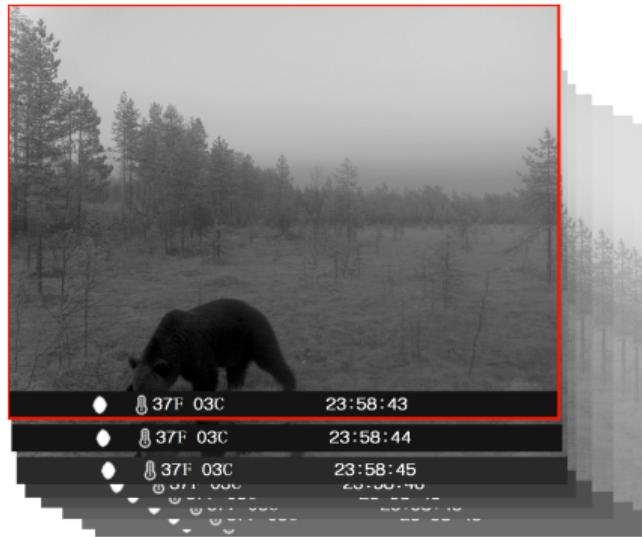
Biases in the train set



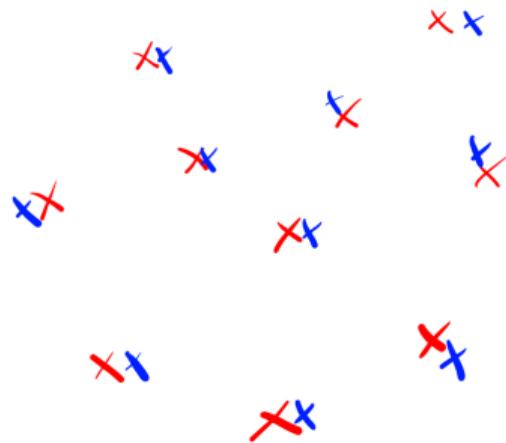
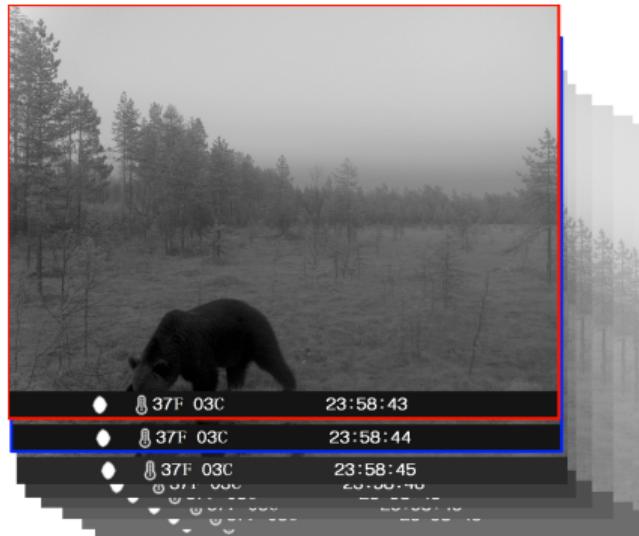
Biases in the train set - autocorrelation



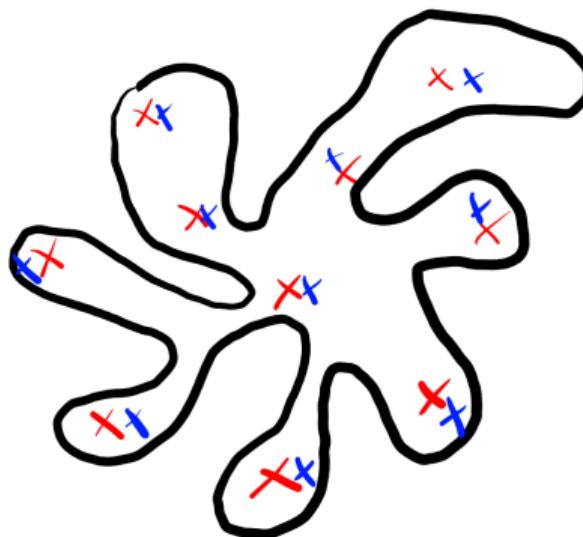
Biases in the train set - autocorrelation



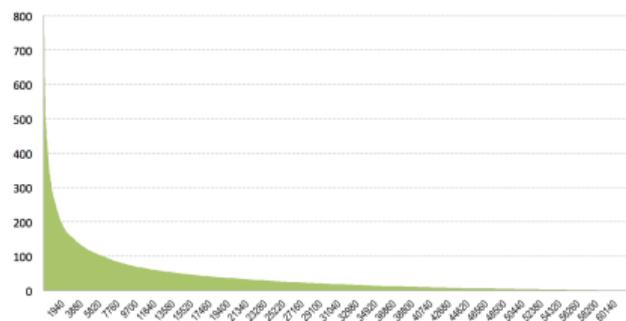
Biases in the train set - autocorrelation



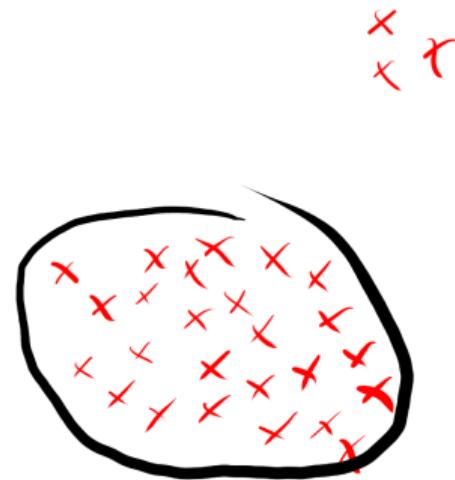
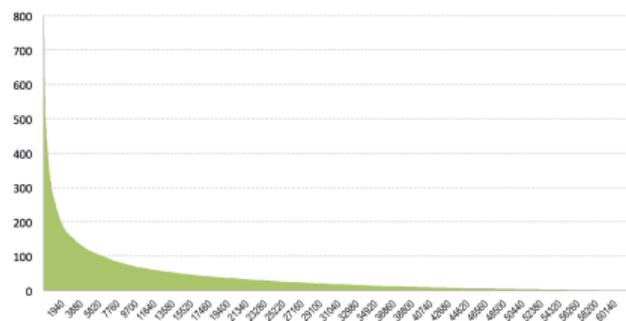
Biases in the train set - autocorrelation



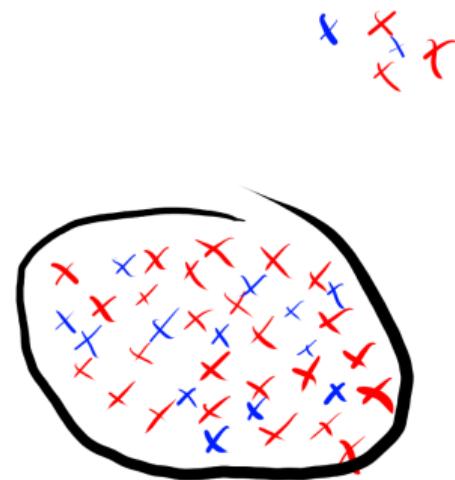
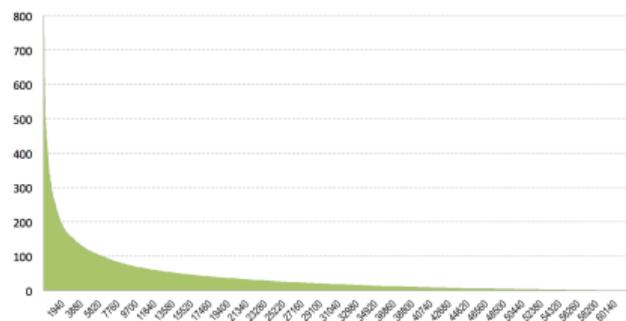
Unbalanced data



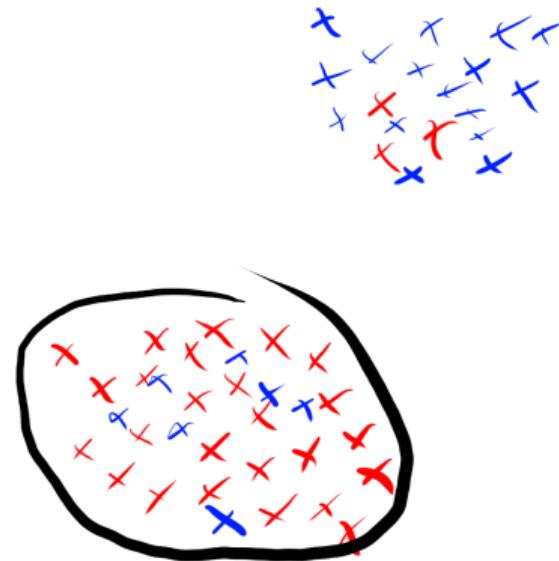
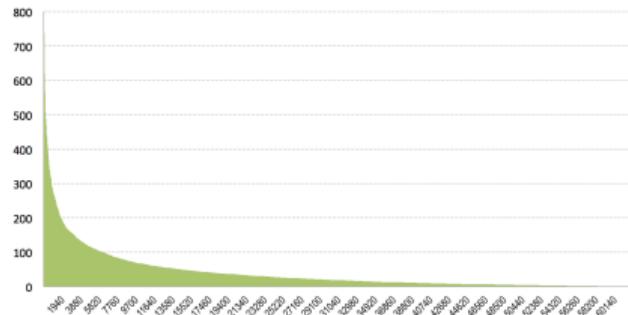
Unbalanced data



Unbalanced data



Unbalanced data



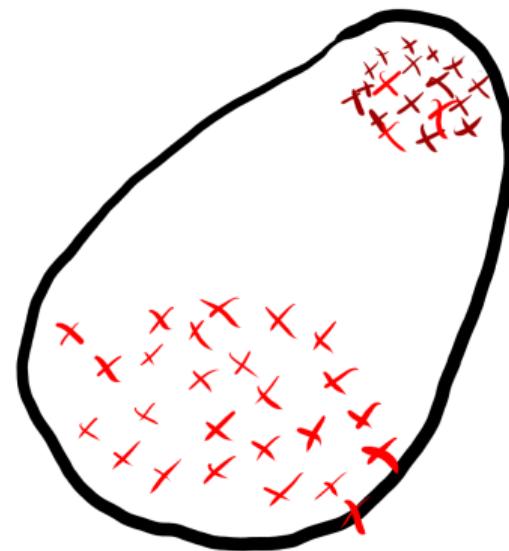
Deal with unbalanced data

- Oversample ?



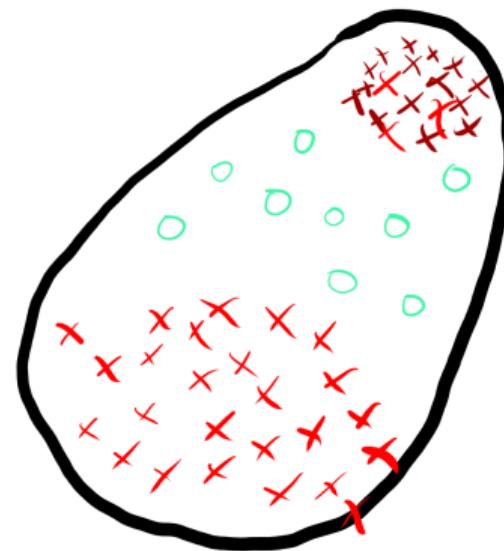
Deal with unbalanced data

- Oversample ?



Deal with unbalanced data

- Oversample ?



Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?

\times
 \times
 \times

\times \times \times \times
 \times \times \times \times
 \times \times \times \times

Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?
- Adapt loss ?



Deal with lack of data

- Data augmentation



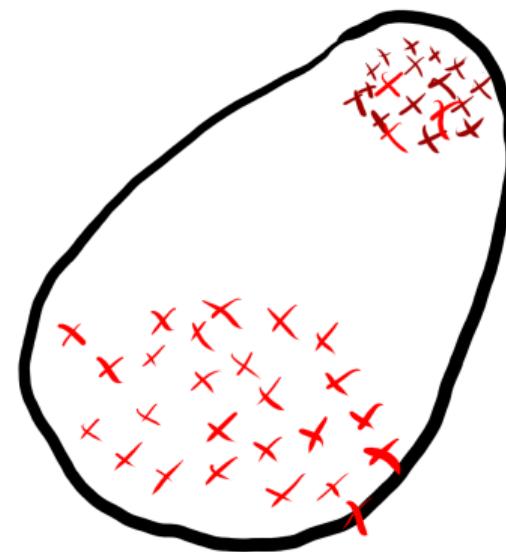
Deal with lack of data

- Data augmentation



Deal with lack of data

- Data augmentation
- Pretrained model



Deal with lack of data

- Data augmentation
- Pretrained model
- ... **collect more data**

How to sample and evaluate ?

Random split ?

Cross-validation

Case study : spatial cross-validation

Case study : Aging models ?

Usefull ressources

- scikit-learn docs !

Thanks for you attention !

Let's practice !

References i

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.