

# **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

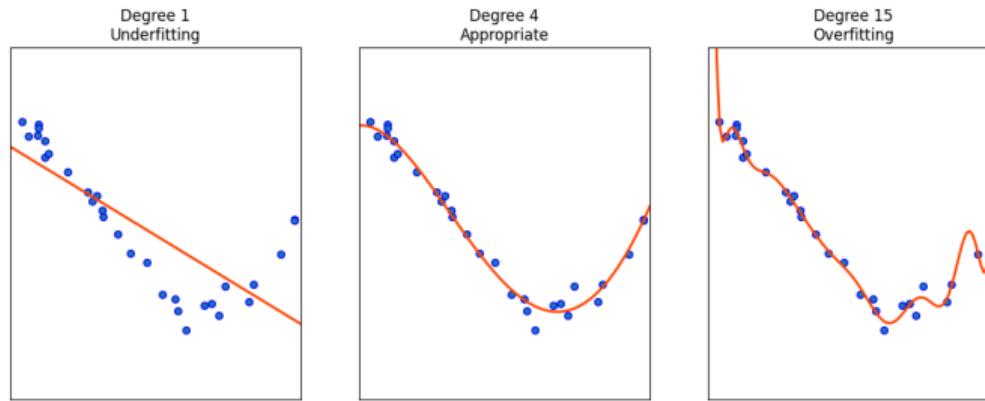
## What do we want when modelling ?

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

# Overfitting

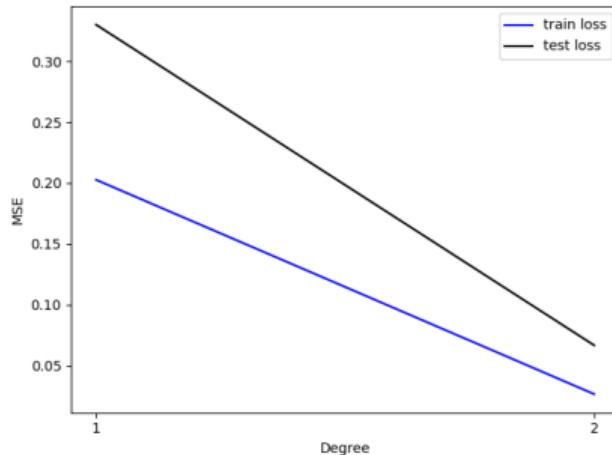
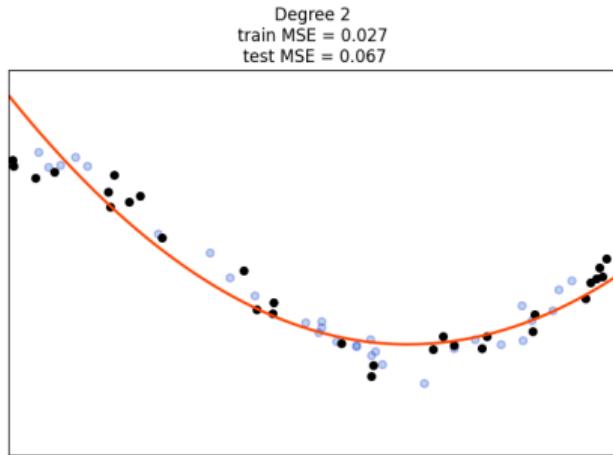
---

# What is overfitting

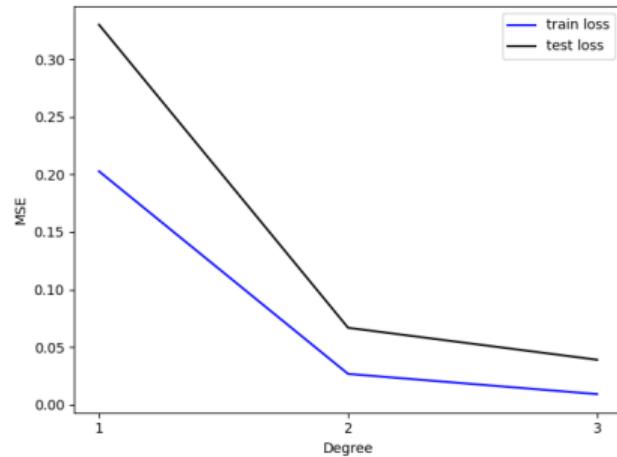
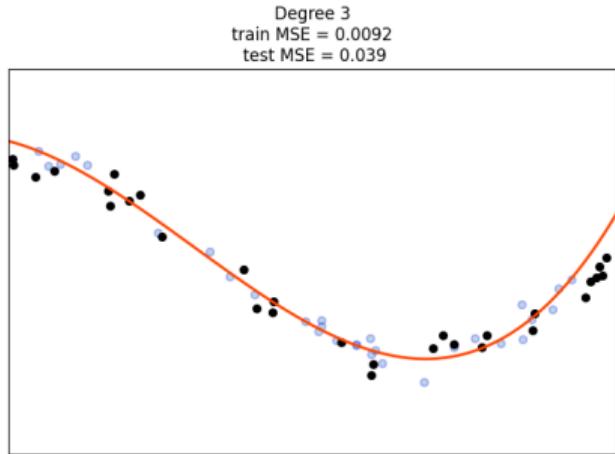


adapted from scikit-learn docs

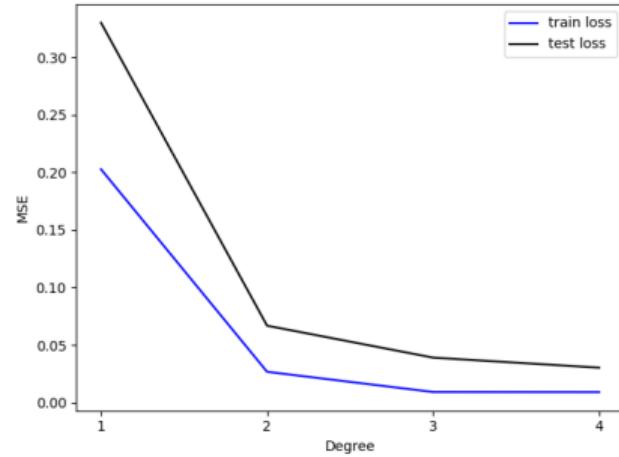
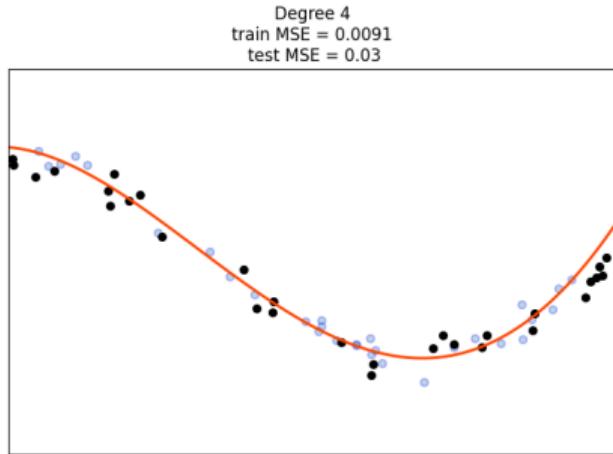
## Common tools and intuitions - Train/Test loss



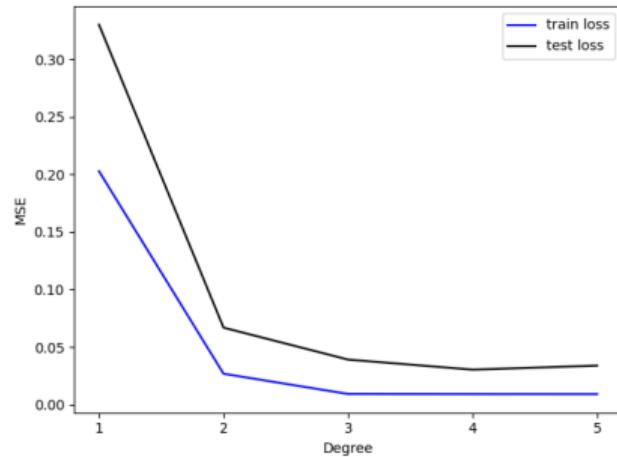
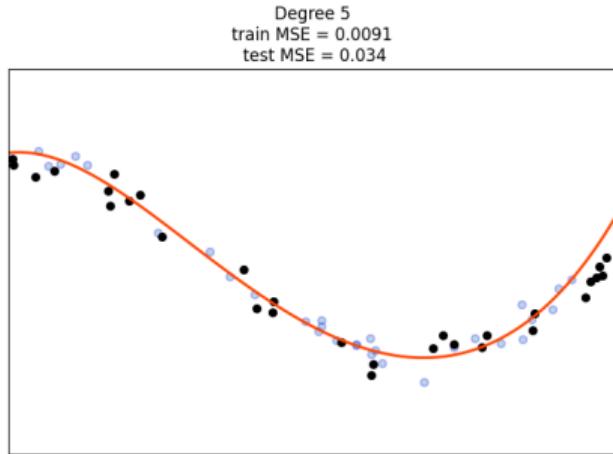
## Common tools and intuitions - Train/Test loss



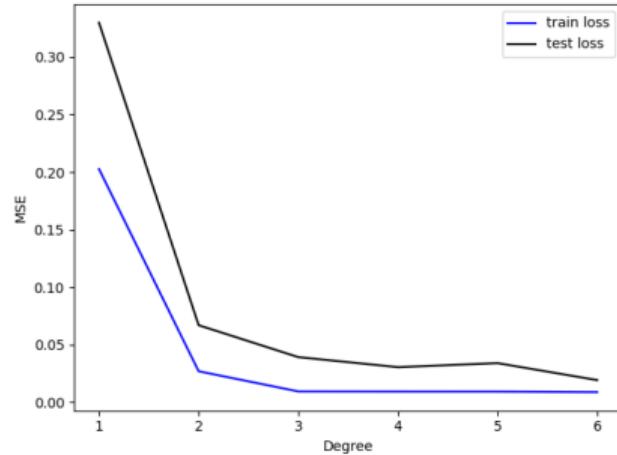
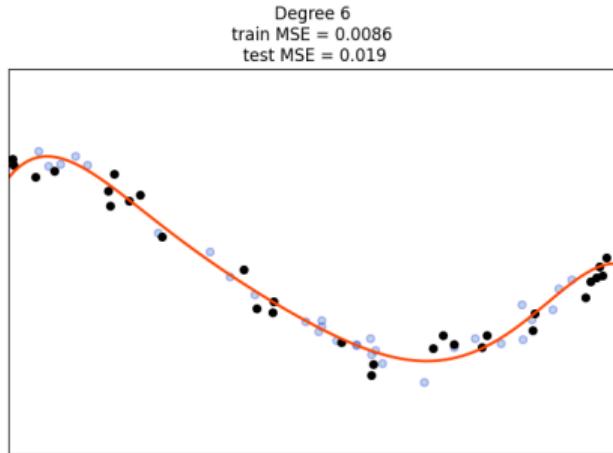
## Common tools and intuitions - Train/Test loss



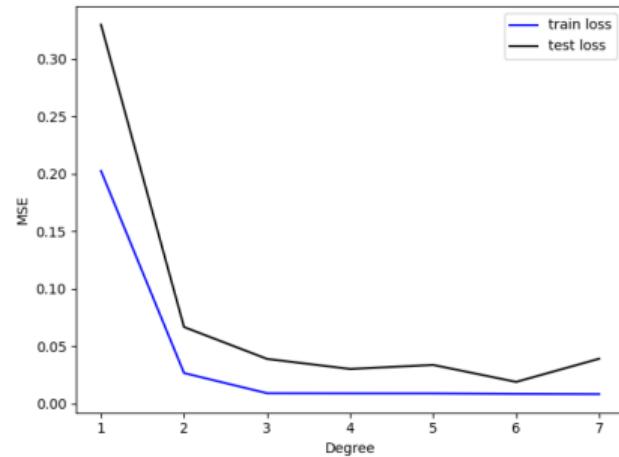
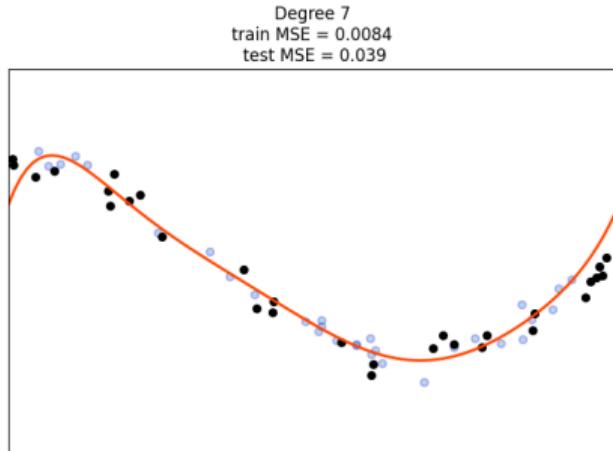
## Common tools and intuitions - Train/Test loss



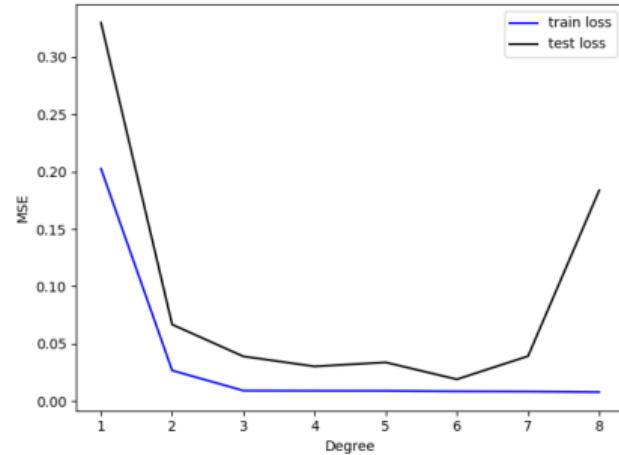
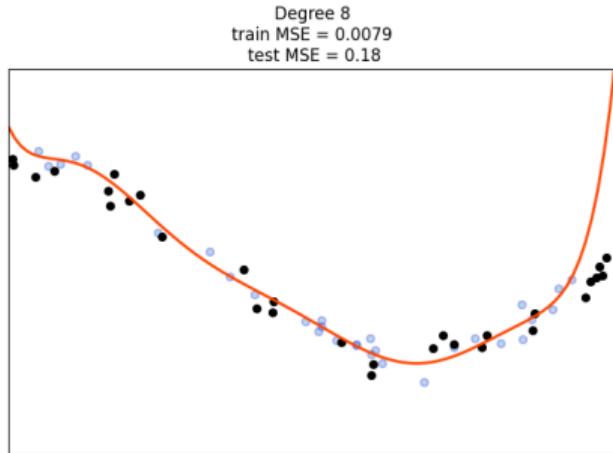
## Common tools and intuitions - Train/Test loss



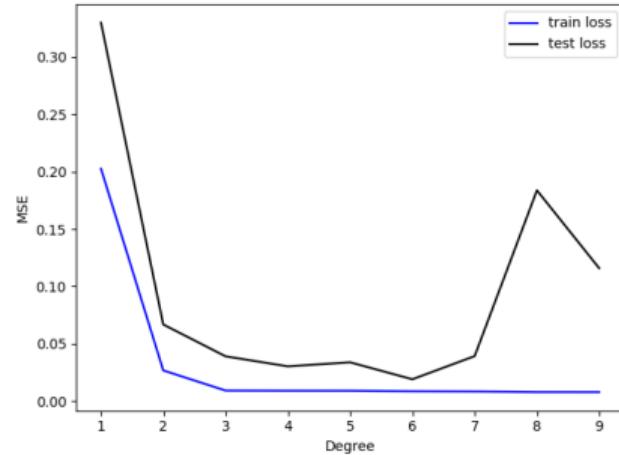
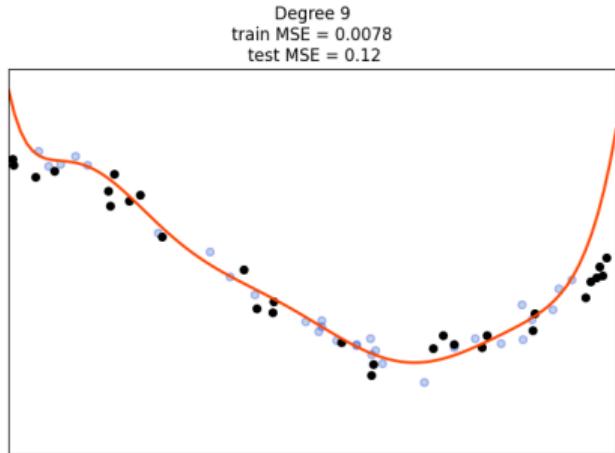
## Common tools and intuitions - Train/Test loss



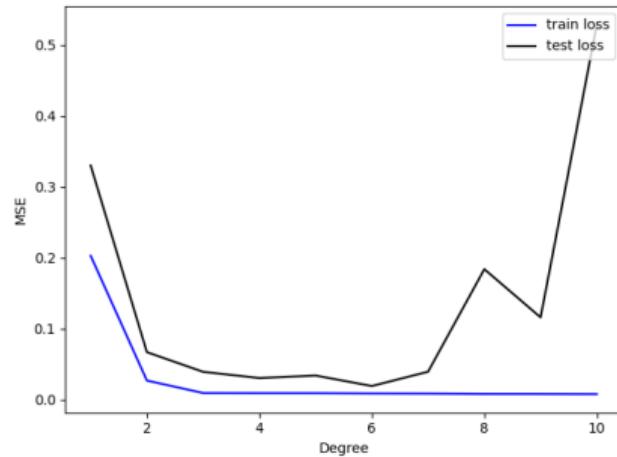
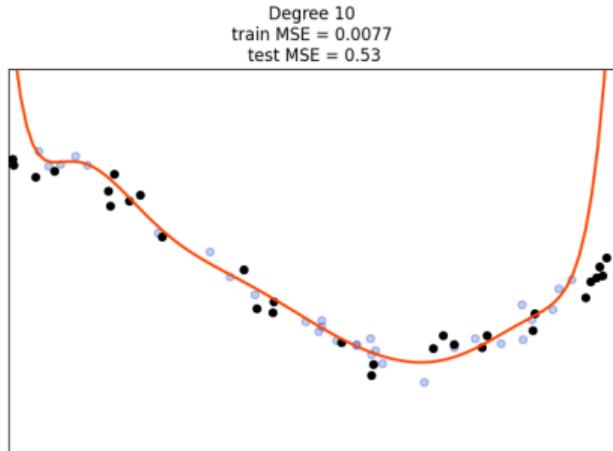
## Common tools and intuitions - Train/Test loss



## Common tools and intuitions - Train/Test loss



## Common tools and intuitions - Train/Test loss



## Common tools and intuitions - Train/Test loss

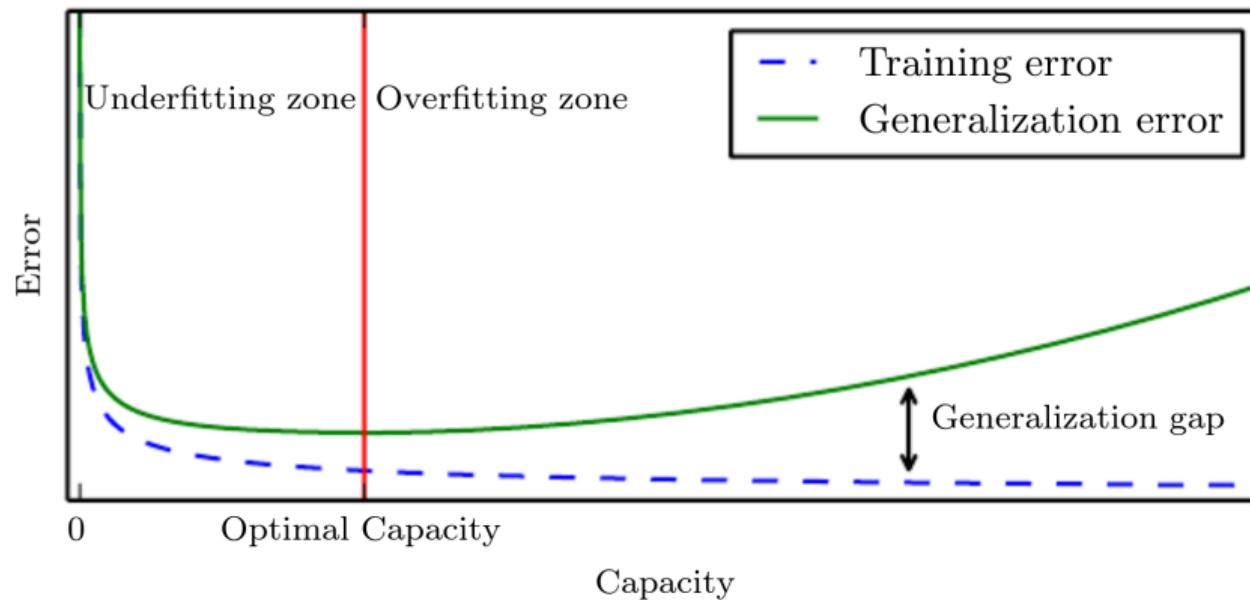


Figure from Goodfellow et al., 2016

## Common tools and intuitions - AIC/BIC

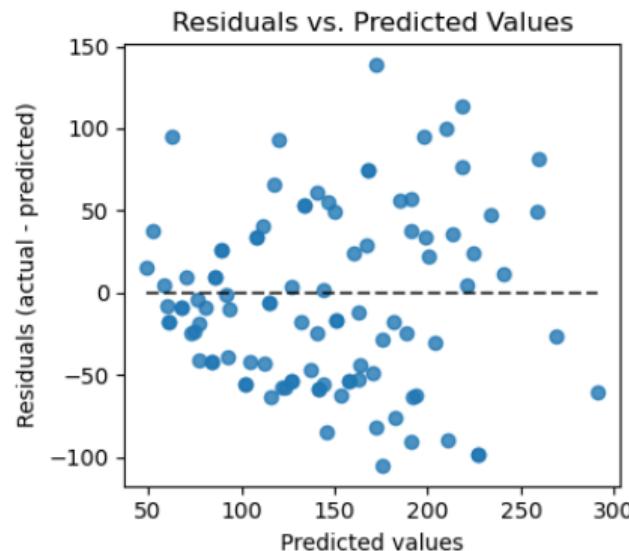
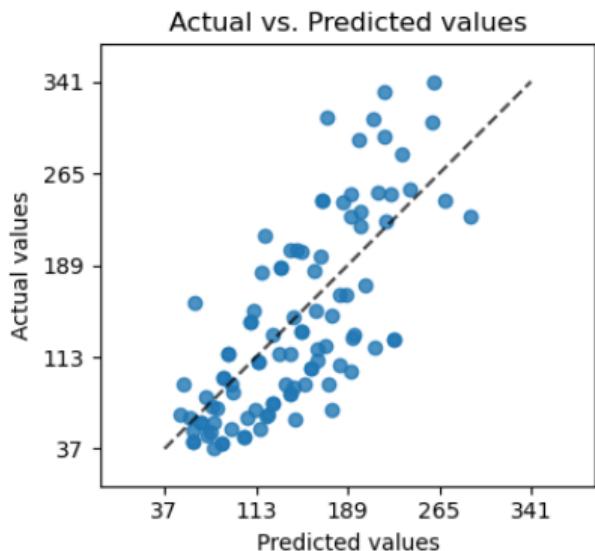
**Akaike information criterion (AIC)**

**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

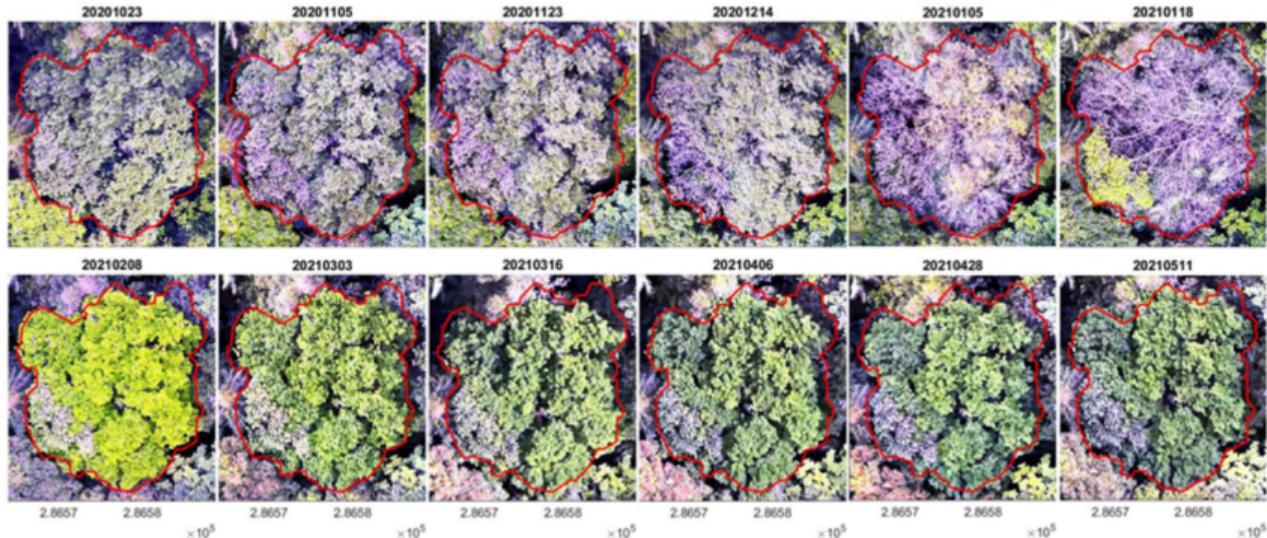
(bad) training data

## Illustrated examples in Ecology

---

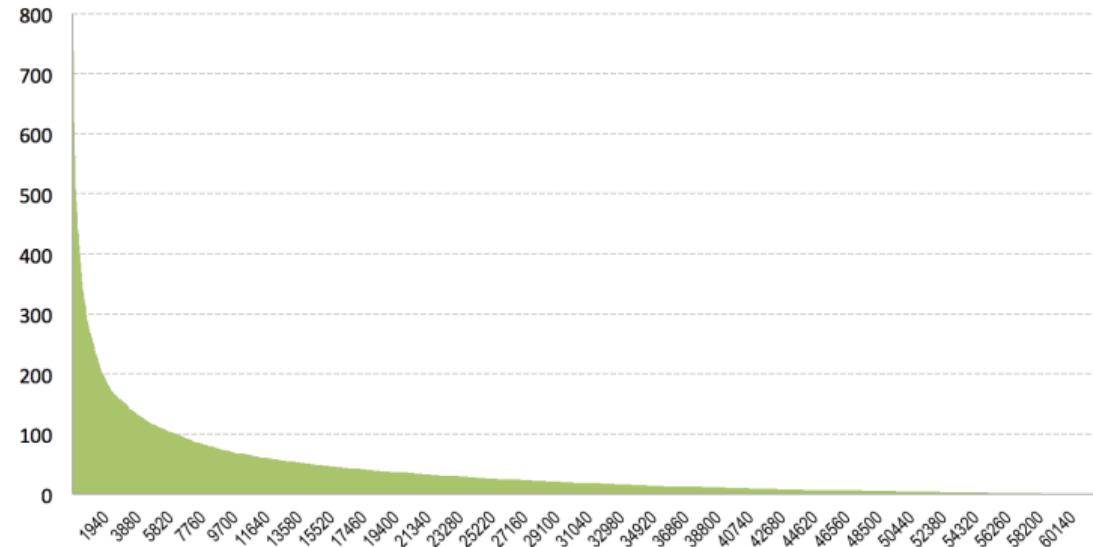
# Constraints in ecology

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.

C    Verify

# Constraints in ecology

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

Select all images with an Orange.

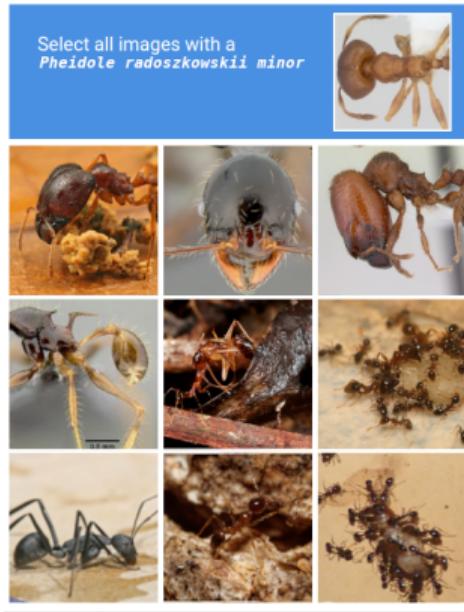
The interface shows a blue header with the text "Select all images with an Orange." Below the header is a 4x3 grid of 12 smaller images. The images depict various stages of orange plants and flowers, including a close-up of orange slices, a tree with fruit, a blossoming flower, a hand holding a small plant, a potted seedling, a large yellow fruit on a branch, a small green sprout in a pot, and a cluster of oranges on a tree. At the bottom of the grid are three small icons: a circular arrow, a speaker, and an information symbol. To the right of the grid is a blue "Verify" button.

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 nine images of ants, likely Pheidole species, used for a classification task. The images include various views of ants, some carrying food, and some in groups. One image in the top row is a clear match for the target species, while the others are different ant species or stages.

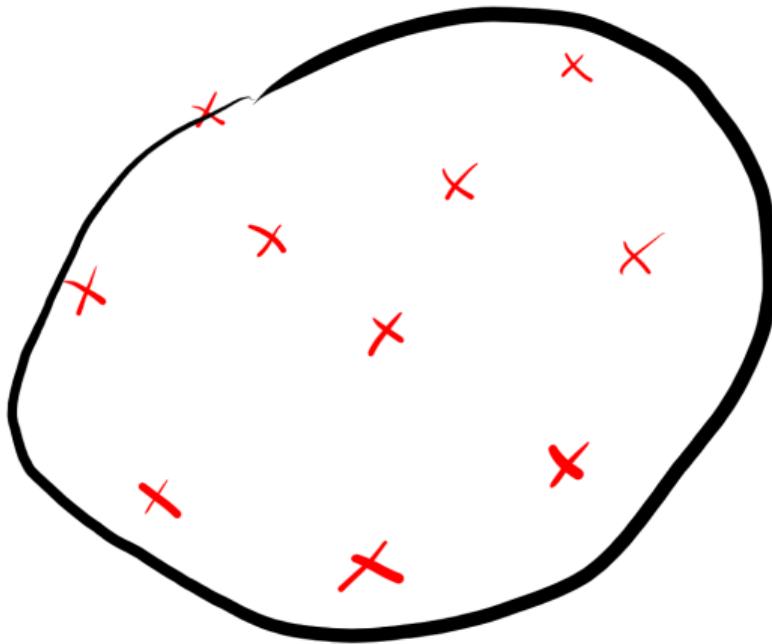
		
		
		



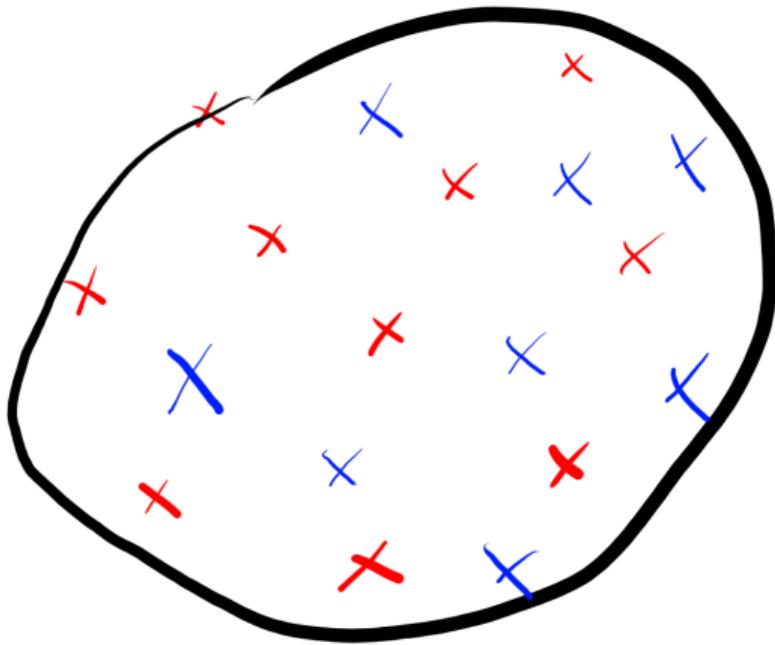
Verify



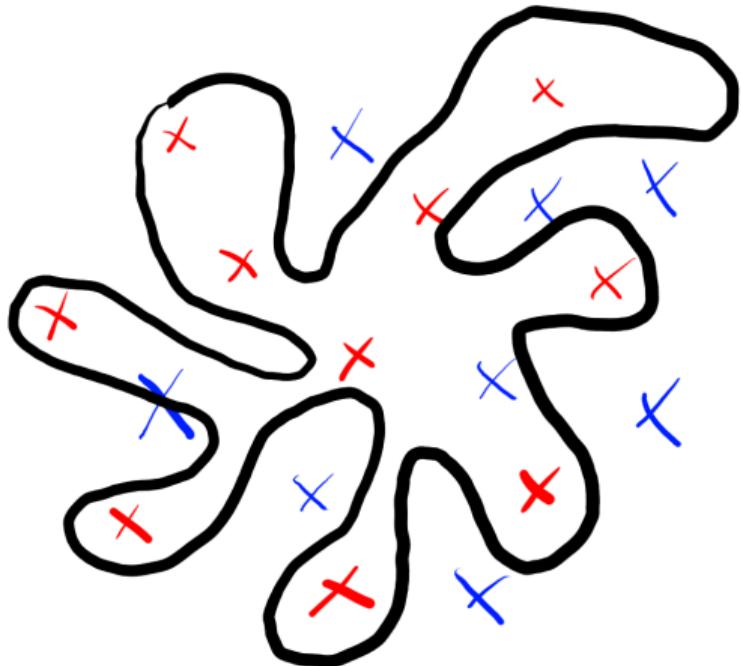
Train set



A good fitted model



Test set

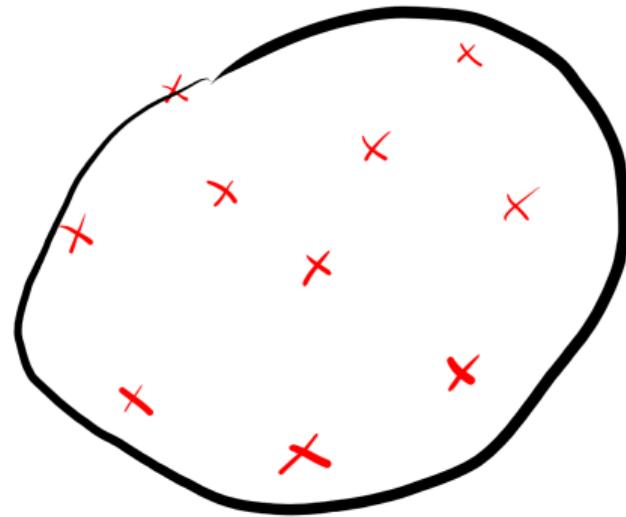


An overfitted model

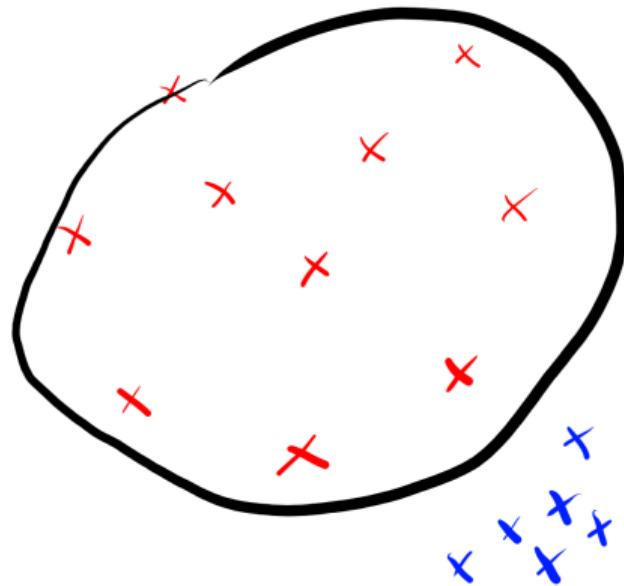
# Biases in the train set



## Biases in the train set



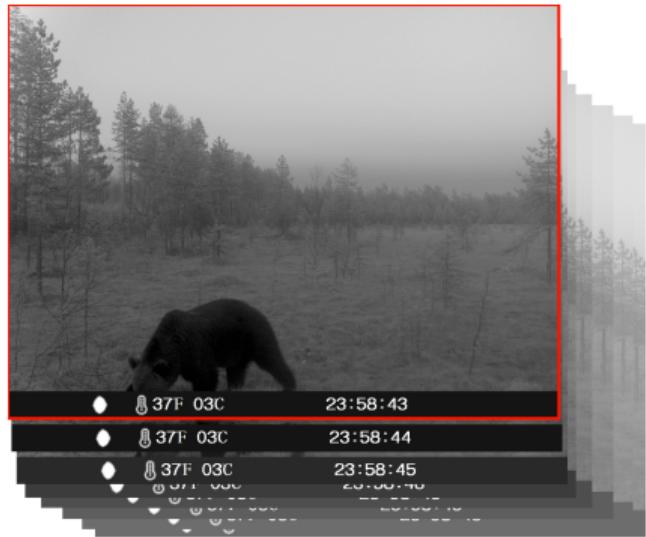
## Biases in the train set



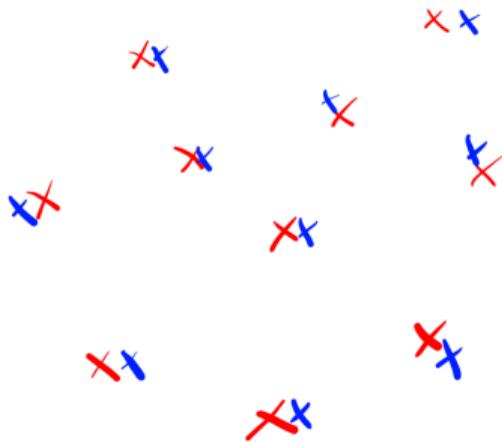
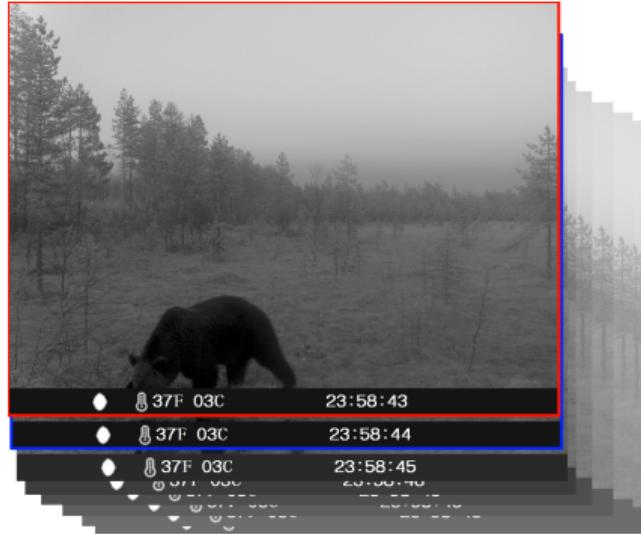
## Biases in the train set - autocorrelation



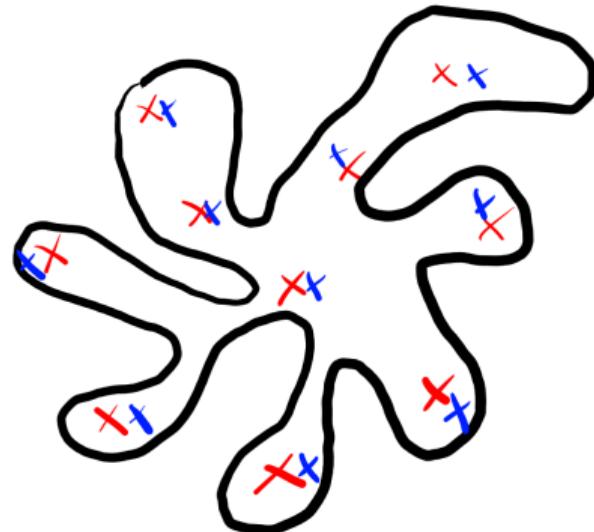
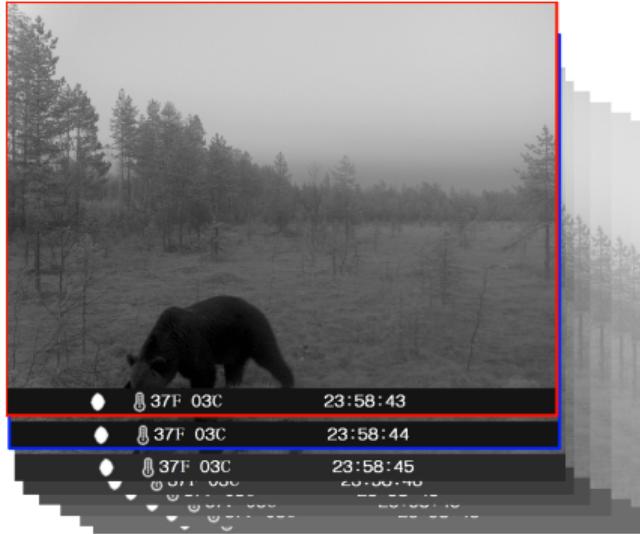
## Biases in the train set - autocorrelation



## Biases in the train set - autocorrelation



## Biases in the train set - autocorrelation



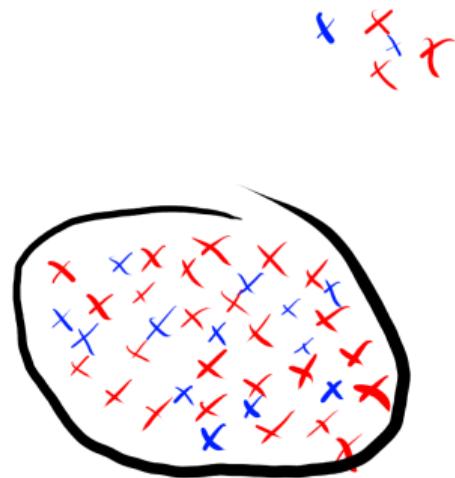
## Unbalanced data



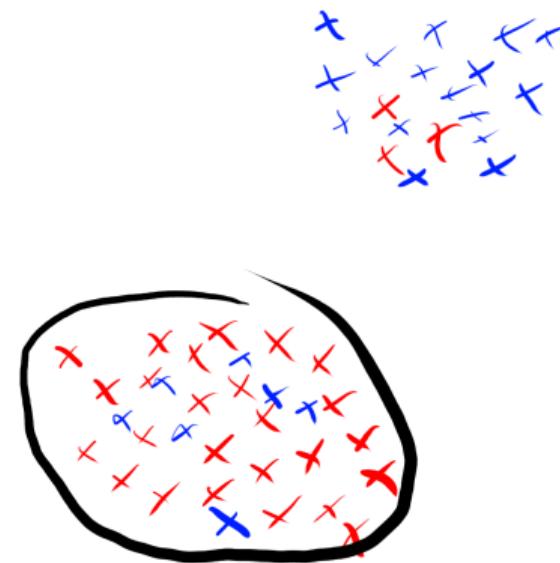
## Unbalanced data



## Unbalanced data



## Unbalanced data



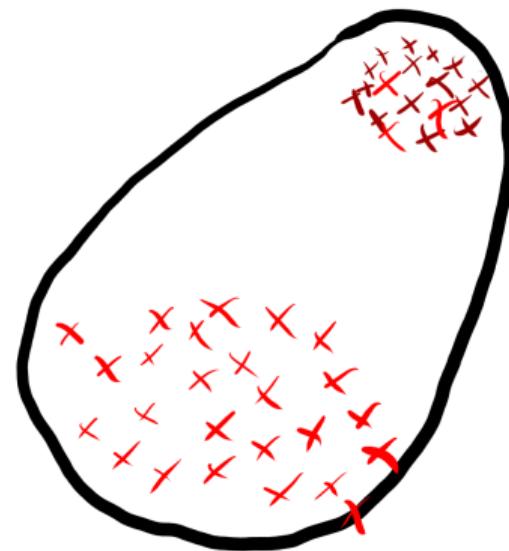
## Deal with unbalanced data

- Oversample ?



## Deal with unbalanced data

- Oversample ?



## Deal with unbalanced data

- Oversample ?



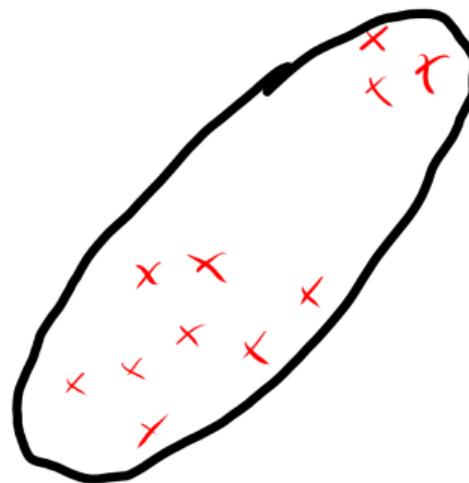
## Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



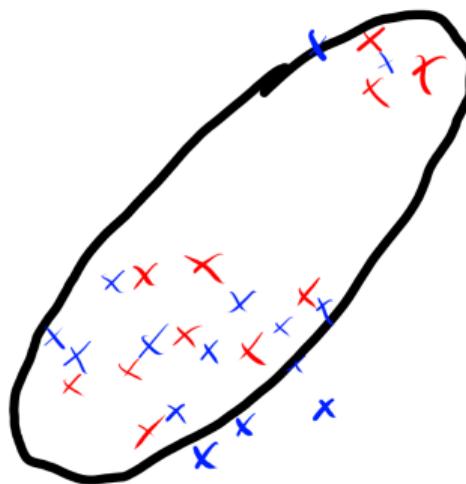
## Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



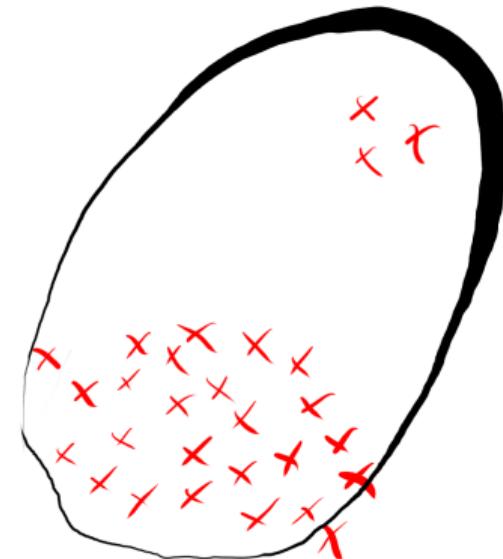
## Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



## Deal with unbalanced data

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



## Deal with lack of data

- Data augmentation



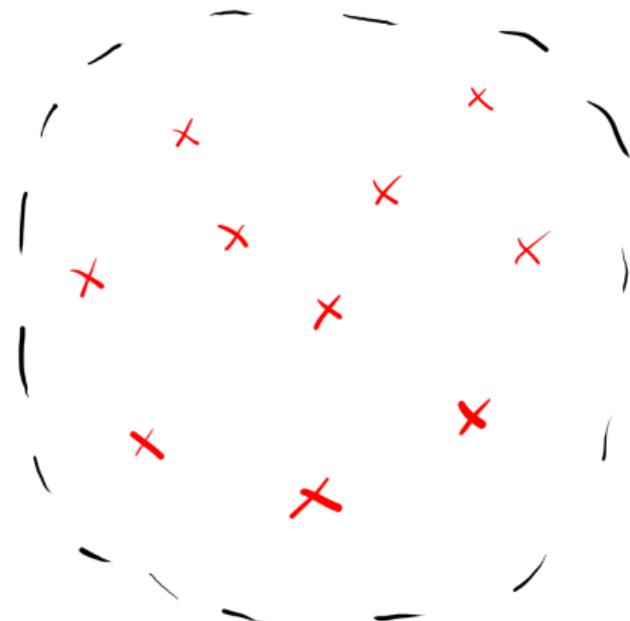
## Deal with lack of data

- Data augmentation



## Deal with lack of data

- Data augmentation
- Pretrained model



## Deal with lack of data

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

**Need to be very carefull on how to evaluate**

## How to sample and evaluate ?

---

## Random split ?

“random split training validation 80/20”

## Random split ?

“random split training validation 80/20”

For the uncurated dataset, we randomly sample 142 million images

Oquab et al., 2023

## Random split ?

“random split training validation 80/20”

For the uncurated dataset, we randomly sample 142 million images

Oquab et al., 2023

Works for huge DL papers, maybe not for you

# Cross-validation

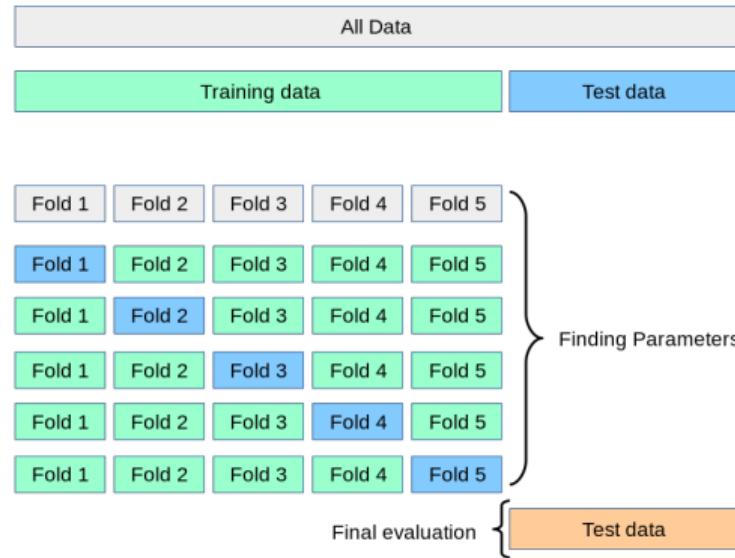


Figure from scikit-learn docs

# Cross-validation

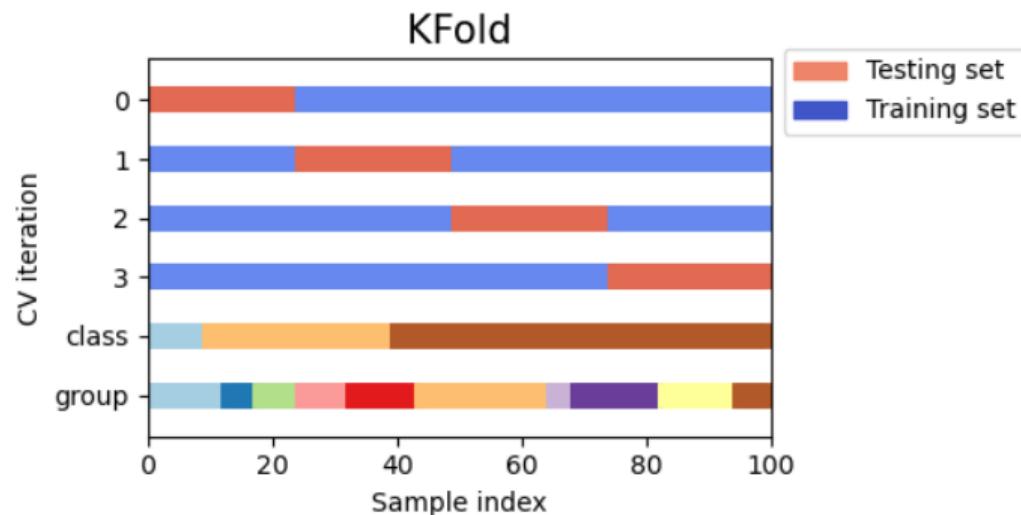


Figure from scikit-learn docs

# Cross-validation

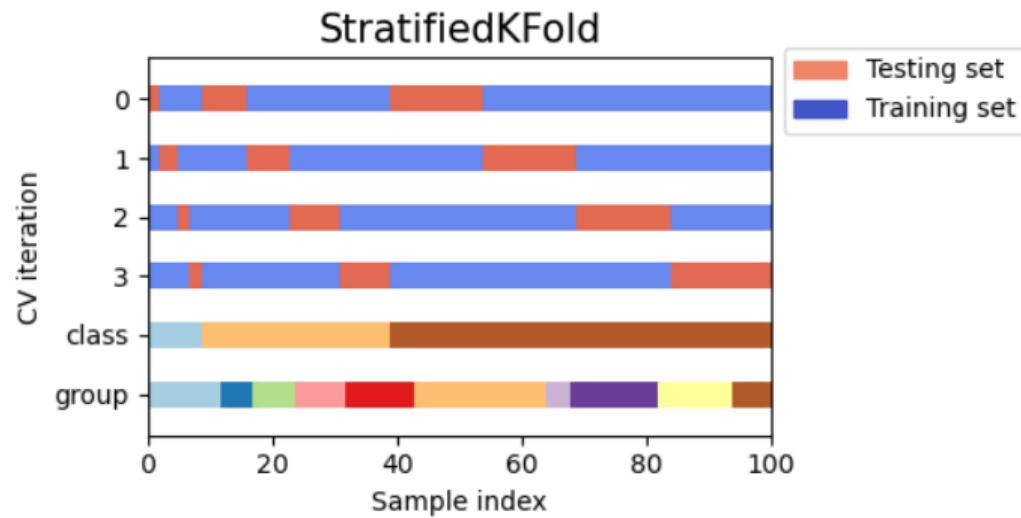
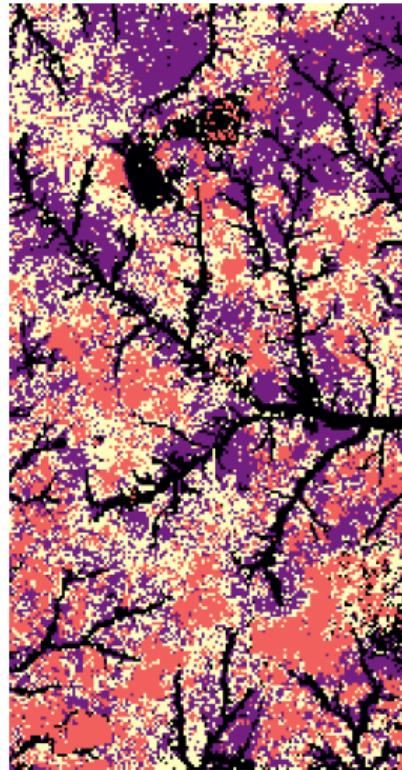


Figure from scikit-learn docs

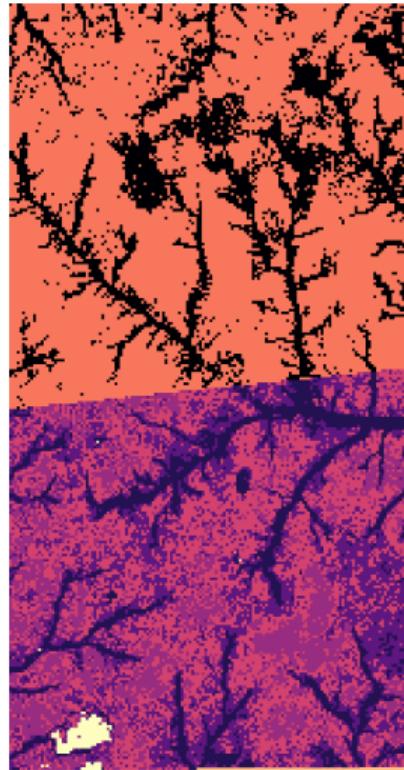
## Case study : Spatial cross-validation



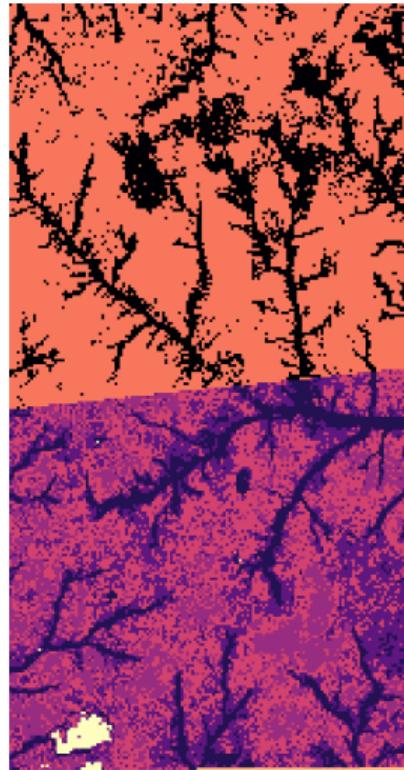
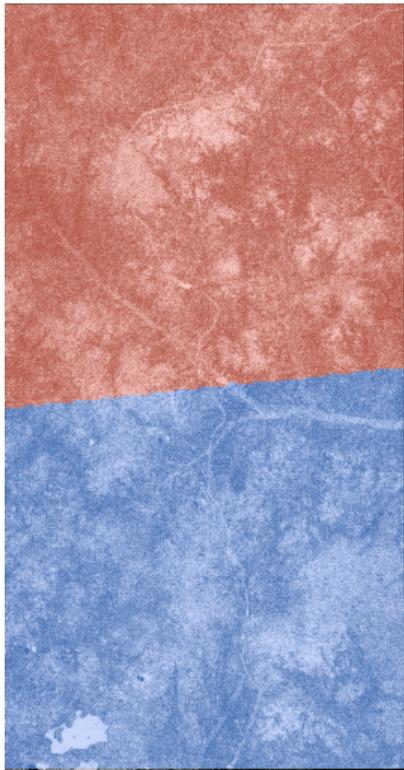
## Case study : Spatial cross-validation



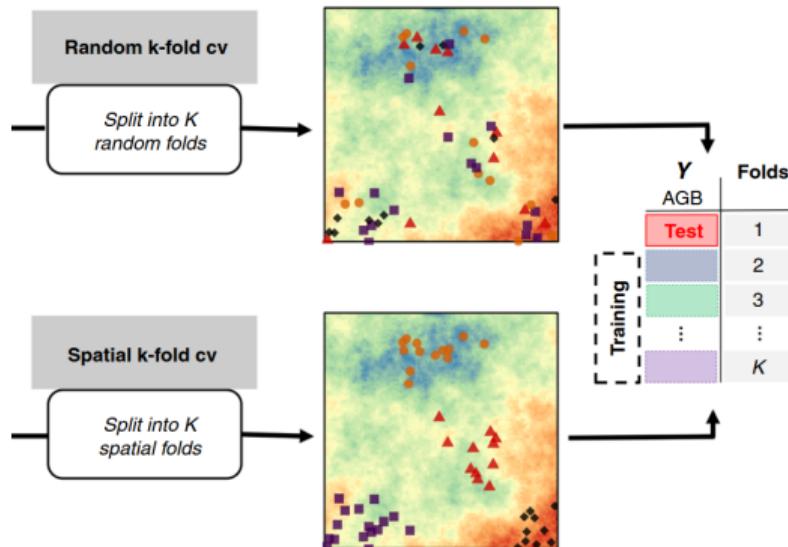
## Case study : Spatial cross-validation



## Case study : Spatial cross-validation

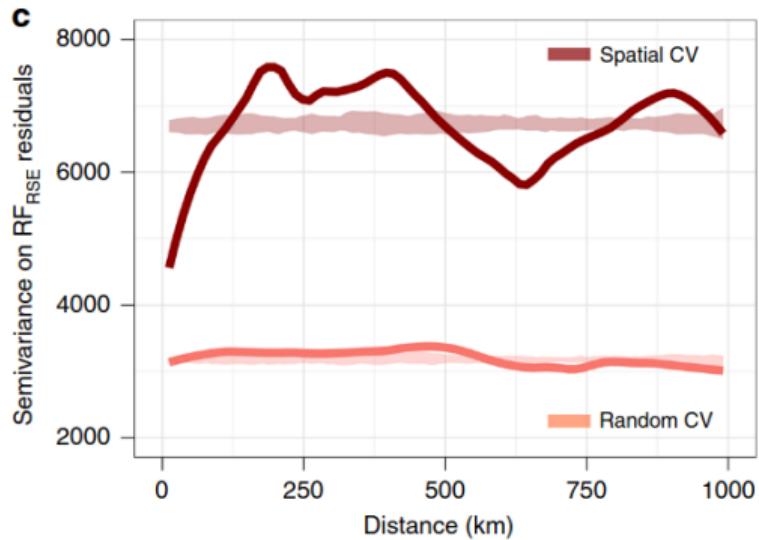


# Case study : Spatial cross-validation



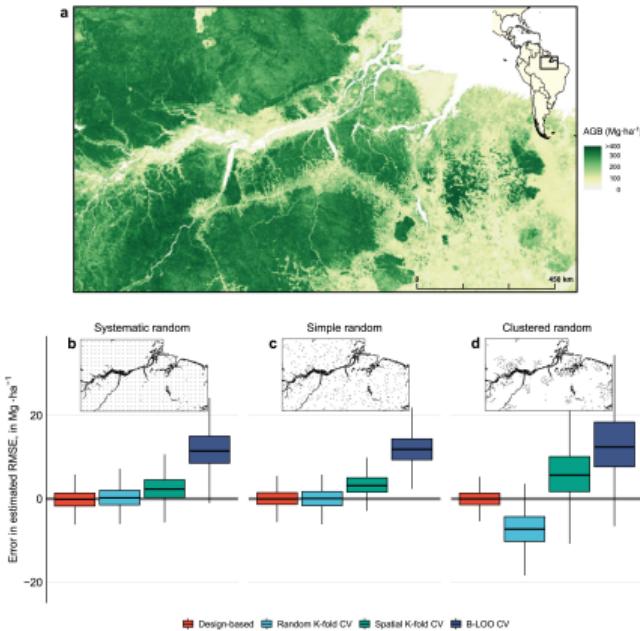
See. Ploton et al., 2020

## Case study : Spatial cross-validation



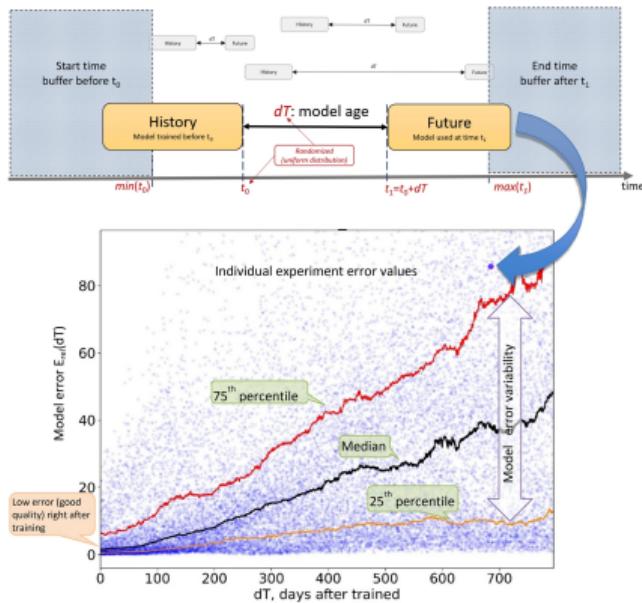
See. Ploton et al., 2020

# Case study : Spatial cross-validation



See. Wadoux et al., 2021

# Case study : Aging models ?



See. Vela et al., 2022

**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.
- Oquab, Maxime et al. (2023). “**Dinov2: Learning robust visual features without supervision**”. In: *arXiv preprint arXiv:2304.07193*.
- Ploton, Pierre et al. (2020). “**Spatial validation reveals poor predictive performance of large-scale ecological mapping models**”. In: *Nature communications* 11.1, p. 4540.
- Vela, Daniel et al. (2022). “**Temporal quality degradation in AI models**”. In: *Scientific reports* 12.1, p. 11654.

## References ii

Wadoux, Alexandre MJ-C, Gerard BM Heuvelink, Sytze De Bruin, and Dick J Brus (2021). “**Spatial cross-validation is not the right way to evaluate map accuracy**”. In: *Ecological Modelling* 457, p. 109692.