Sampling and overfitting

AI for ecologists

Paul Tresson

21/05/25

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”

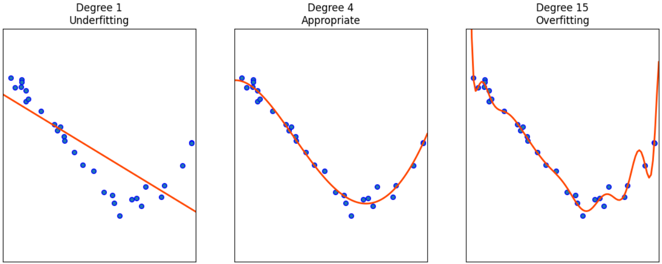
George E. P. Box

What do we want when modelling ?

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

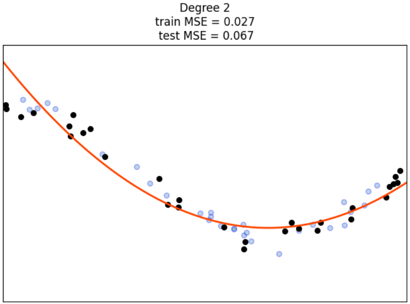
Overfitting

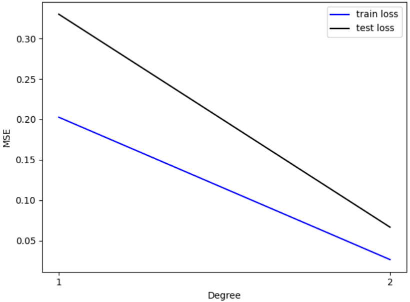
What is overfitting



adapted from scikit-learn docs

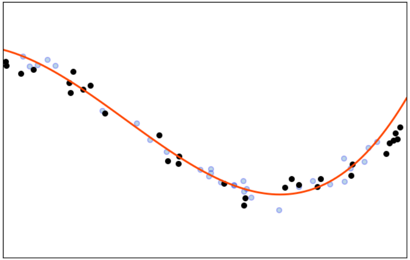
Common tools and intuitions - Train/Test loss

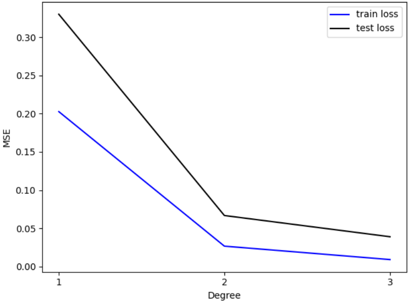




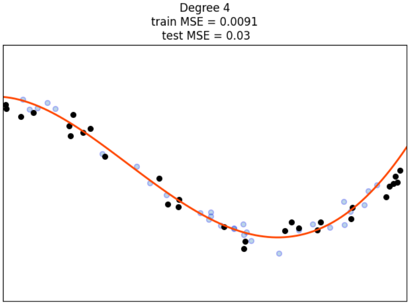
Common tools and intuitions - Train/Test loss

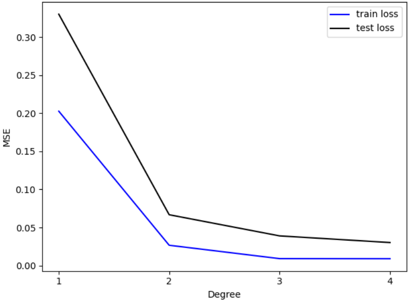
Degree 3
train MSE = 0.0092
test MSE = 0.039



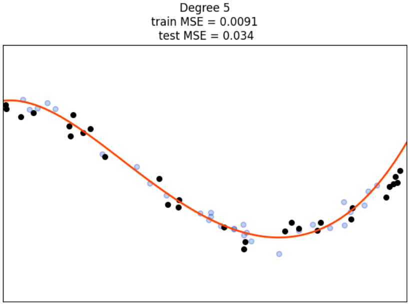


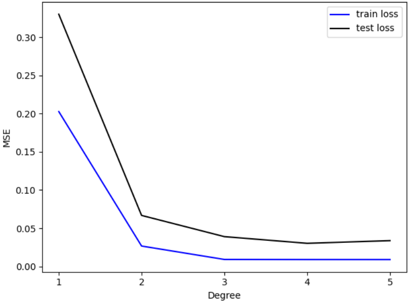
Common tools and intuitions - Train/Test loss



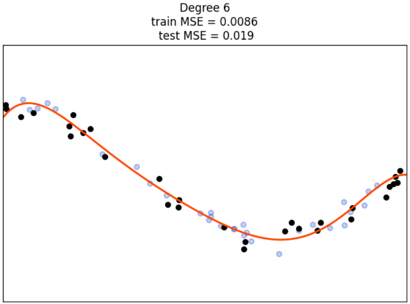


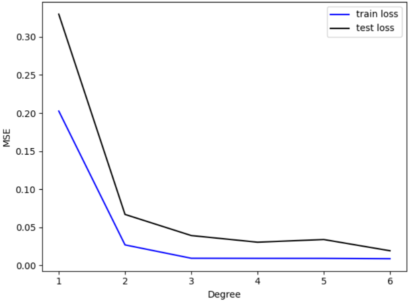
Common tools and intuitions - Train/Test loss



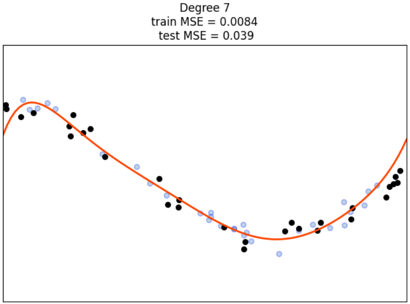


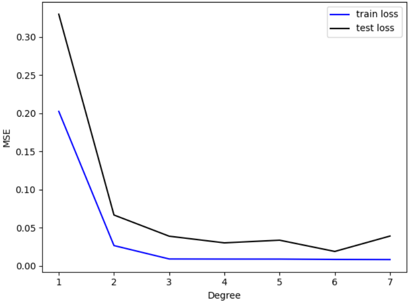
Common tools and intuitions - Train/Test loss





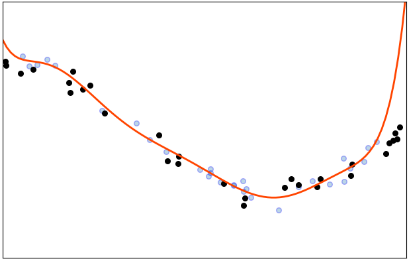
Common tools and intuitions - Train/Test loss

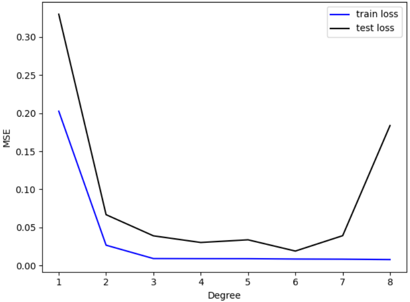




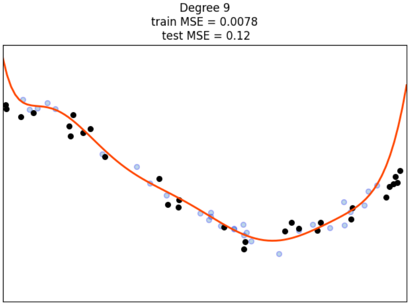
Common tools and intuitions - Train/Test loss

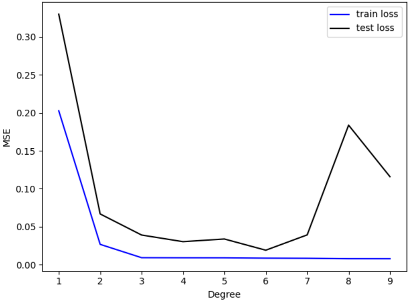
Degree 8
train MSE = 0.0079
test MSE = 0.18



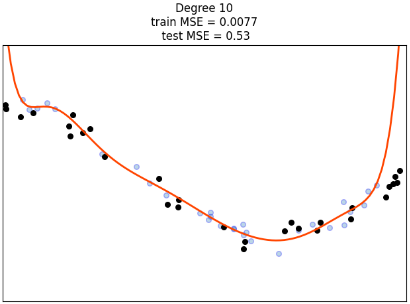


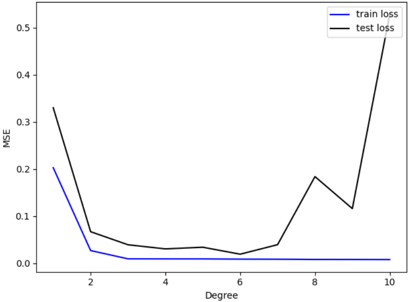
Common tools and intuitions - Train/Test loss





Common tools and intuitions - Train/Test loss





Common tools and intuitions - Train/Test loss

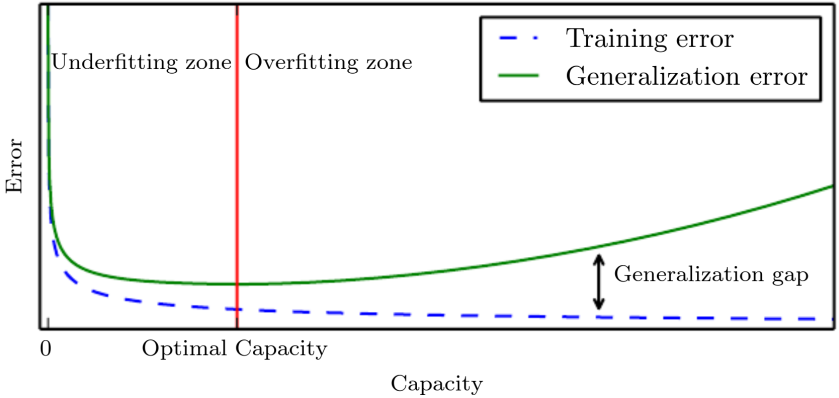


Figure from Goodfellow et al., 2016

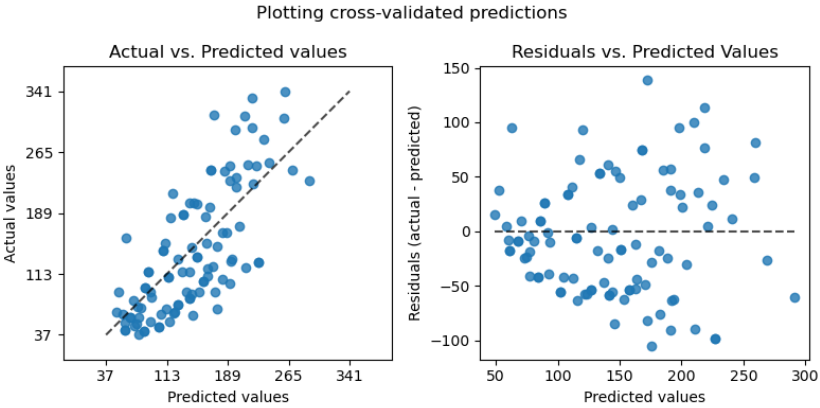
Common tools and intuitions - AIC/BIC

Akaike information criterion (AIC)

Bayesian information criterion (BIC)

Is the model parameter efficient ?

Common tools and intuitions - Biases



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

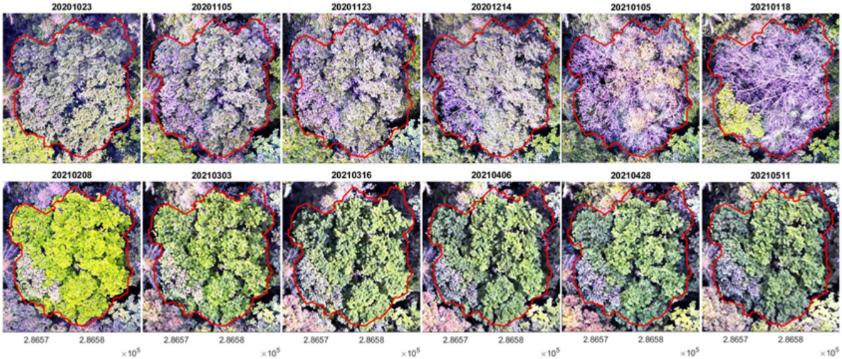
Too little training data

(bad) training data

Illustrated examples in Ecology

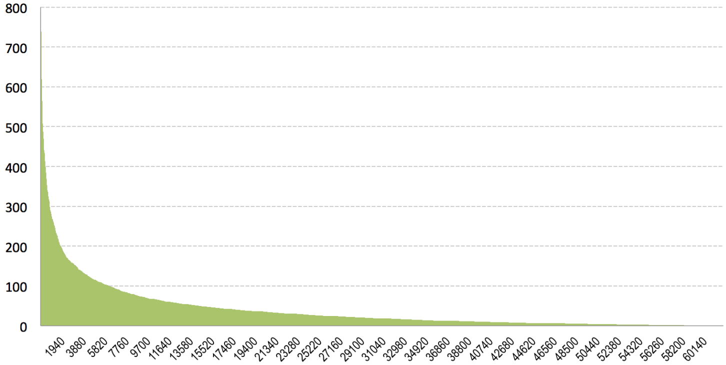
Contraints in ecology

Data from the real world is noisy,



Contraints in ecology

Data from the real world is noisy, unbalanced,

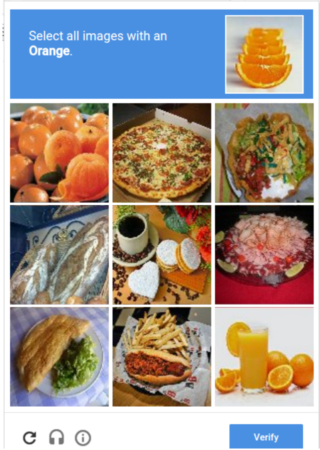


Contraints in ecology

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

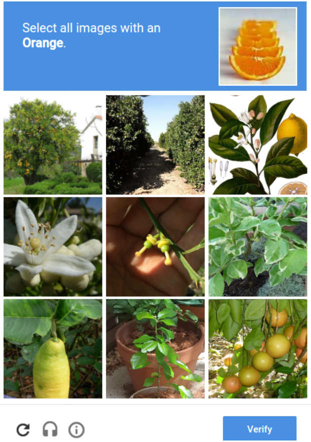
Contraints in ecology

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



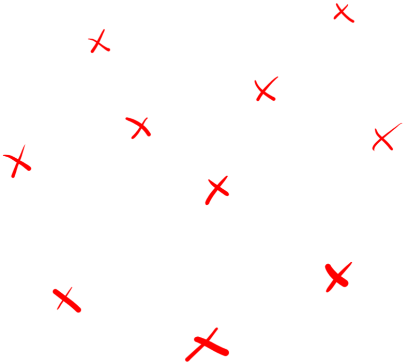
Contraints in ecology

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

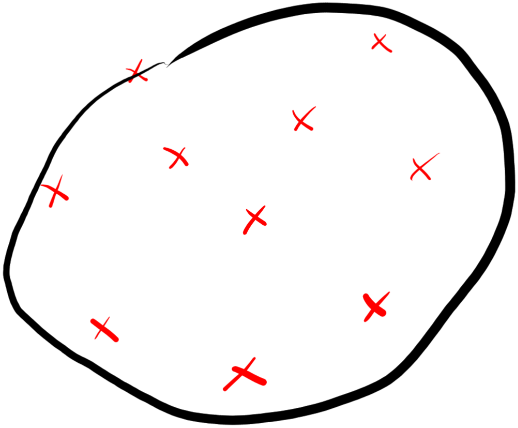


Contraints in ecology

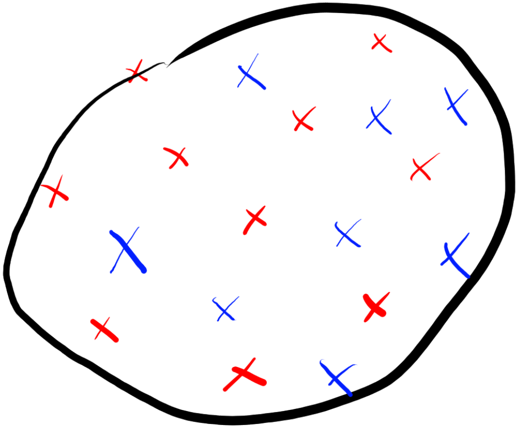
Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.



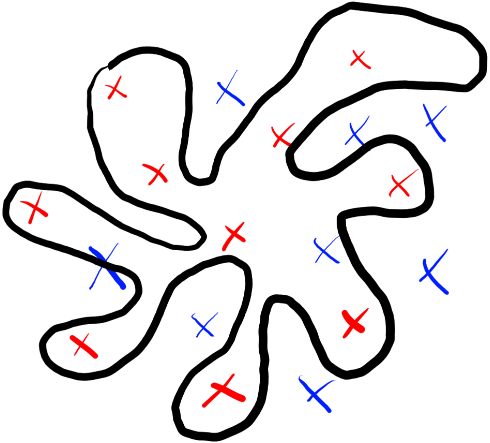
Train set



A good fitted model

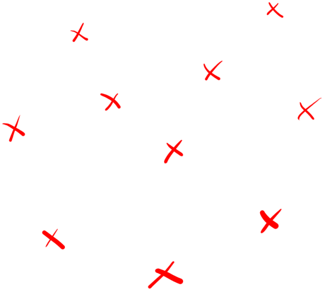


Test set

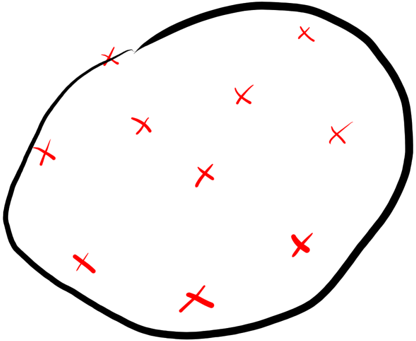


An overfitted model

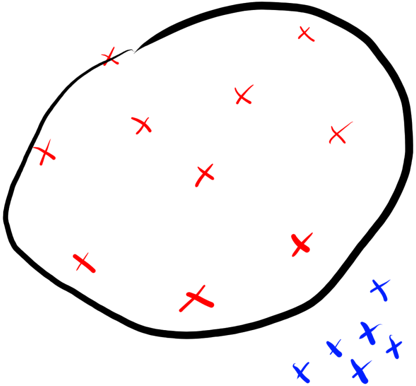
Biases in the train set



Biases in the train set

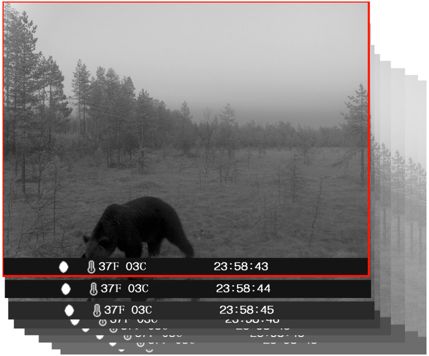


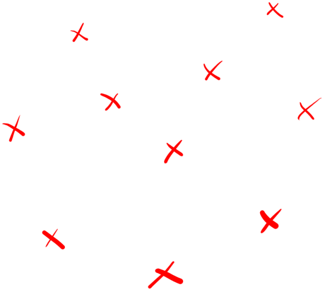
Biases in the train set



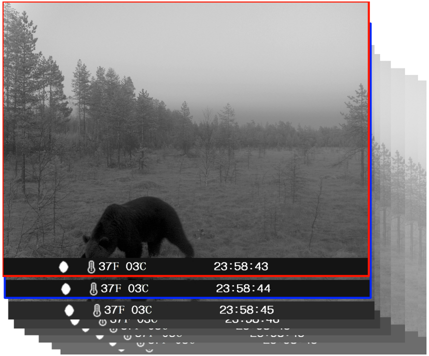
Biases in the train set - autocorrelation

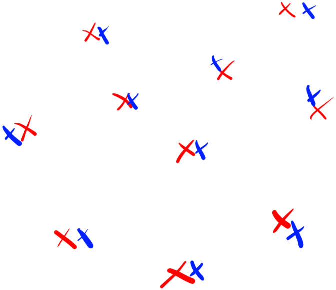
Biases in the train set - autocorrelation



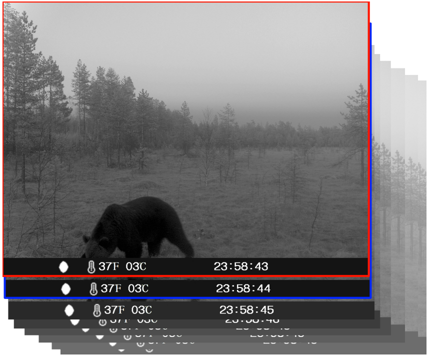


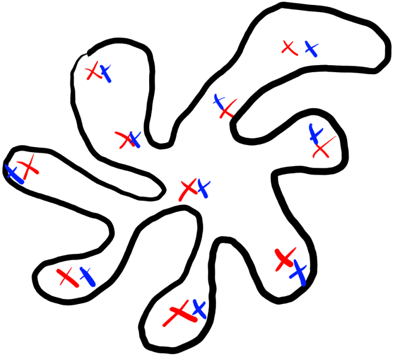
Biases in the train set - autocorrelation



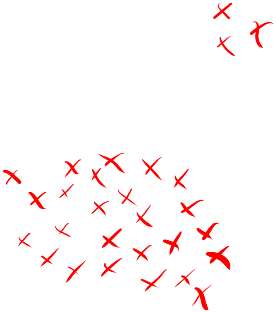


Biases in the train set - autocorrelation

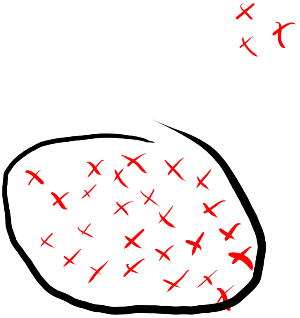




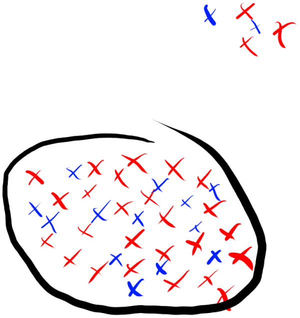
Unbalanced data



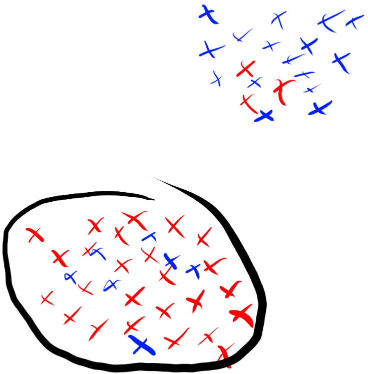
Unbalanced data



Unbalanced data



Unbalanced data

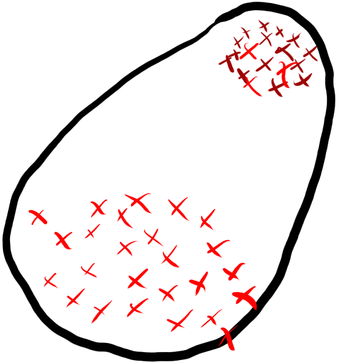


Deal with unbalanced data

- Oversample ?

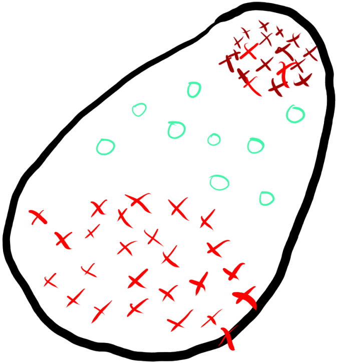
Deal with unbalanced data

- Oversample ?



Deal with unbalanced data

- Oversample ?

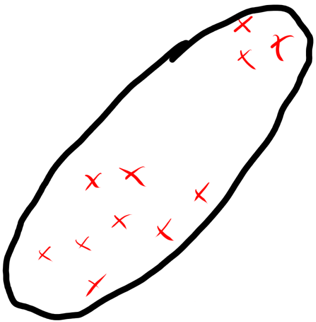


Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?

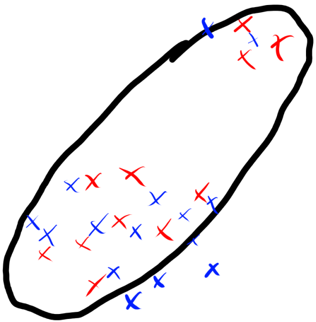
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



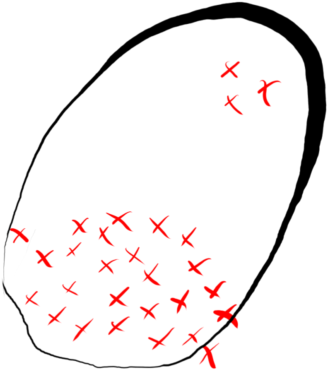
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



Deal with unbalanced data

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

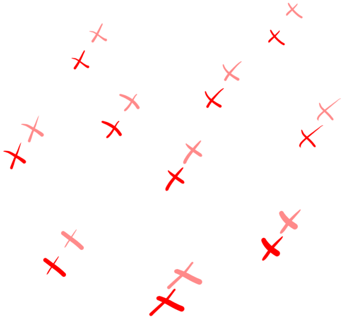


Deal with lack of data

- Data augmentation

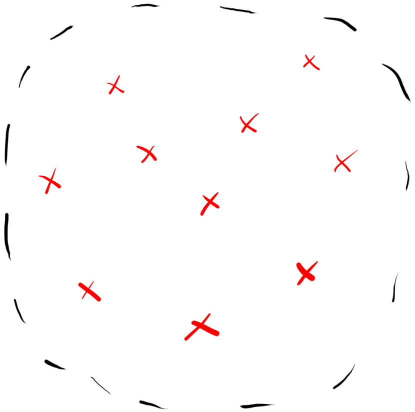
Deal with lack of data

- Data augmentation



Deal with lack of data

- Data augmentation
- Pretrained model



Deal with lack of data

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

Play with your model

- Dropout
- Pruning
- Ablation studies
- Ensembles

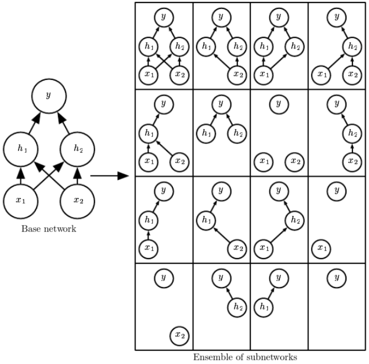


Figure from Goodfellow et al., 2016

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

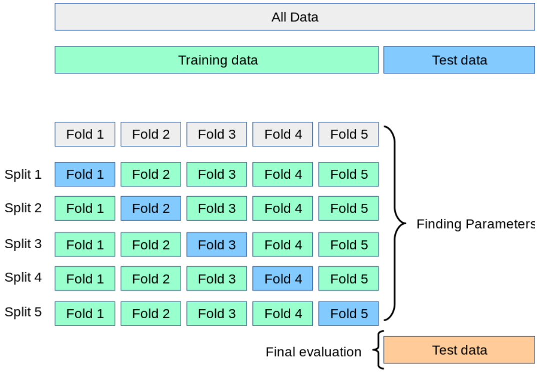


Figure from scikit-learn docs

Cross-validation

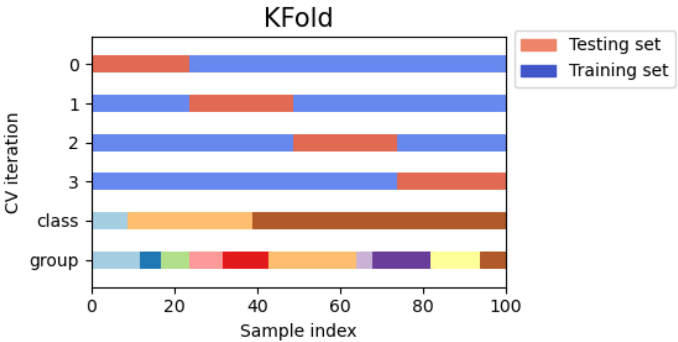


Figure from scikit-learn docs

Cross-validation

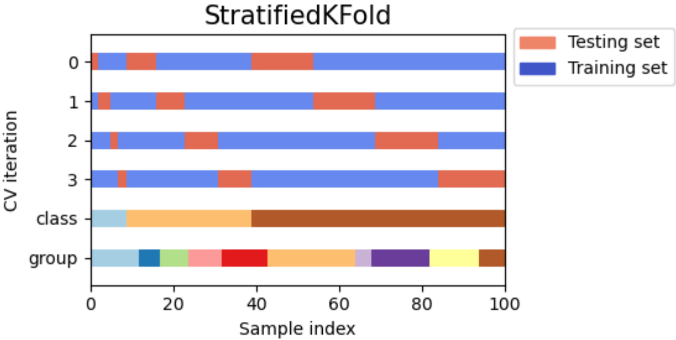
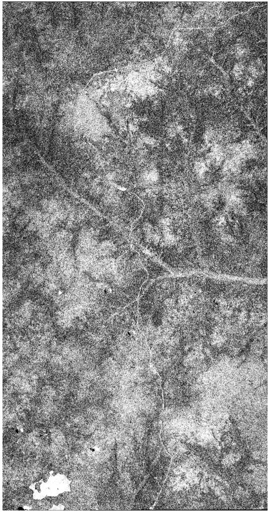
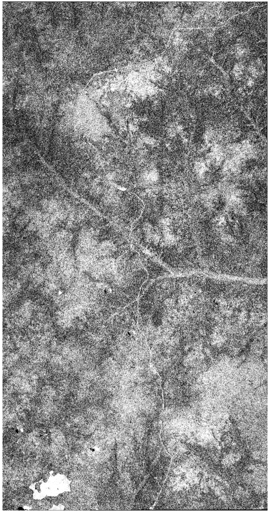


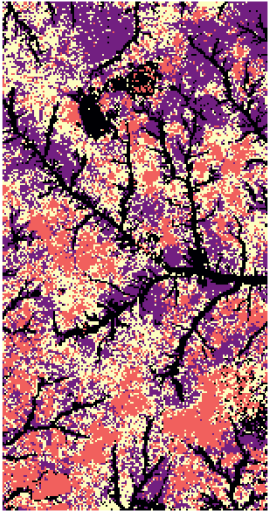
Figure from scikit-learn docs

Case study : Spatial cross-validation

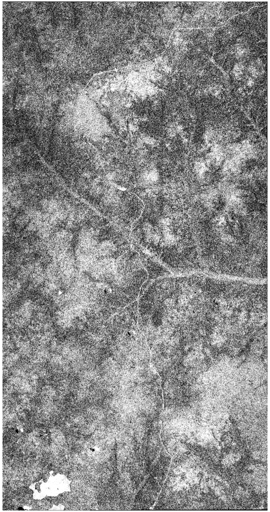


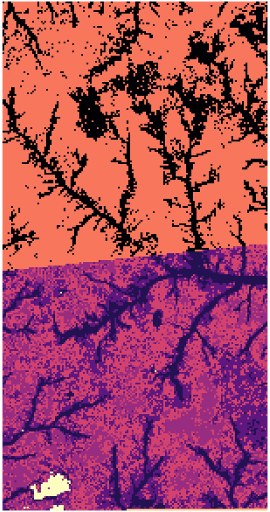
Case study : Spatial cross-validation



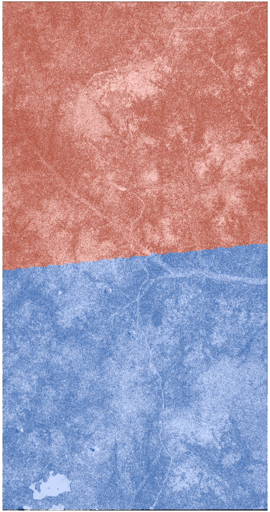


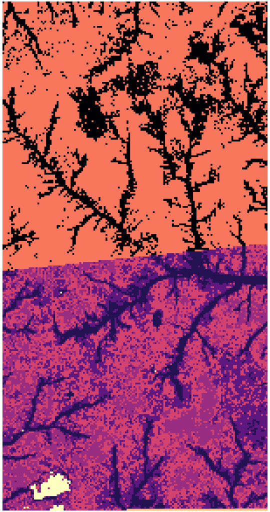
Case study : Spatial cross-validation



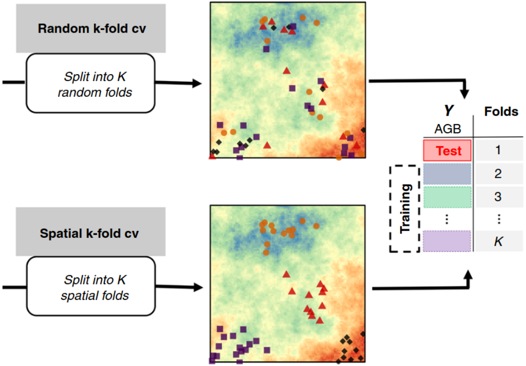


Case study : Spatial cross-validation



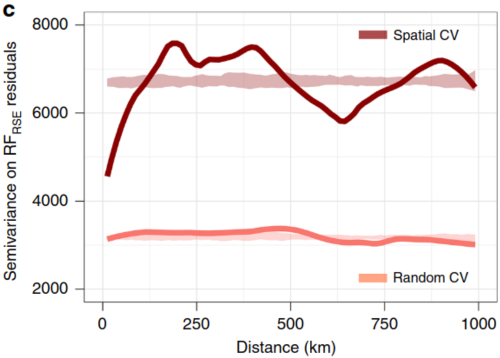


Case study : Spatial cross-validation



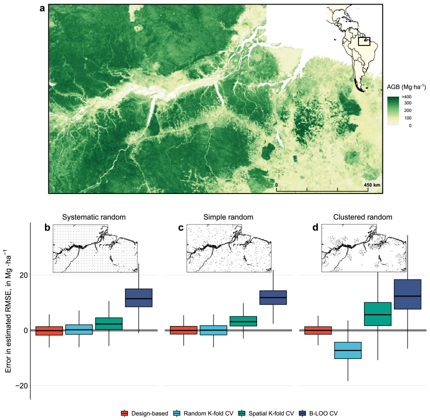
See. Ploton et al., 2020

Case study : Spatial cross-validation



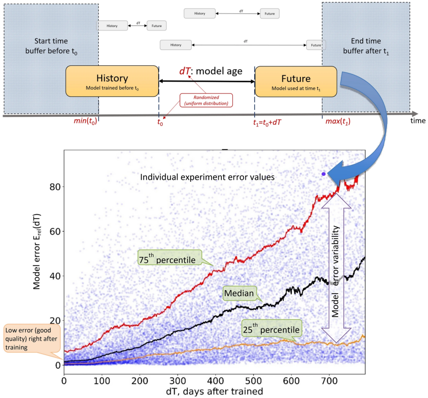
See. Ploton et al., 2020

Case study : Spatial cross-validation



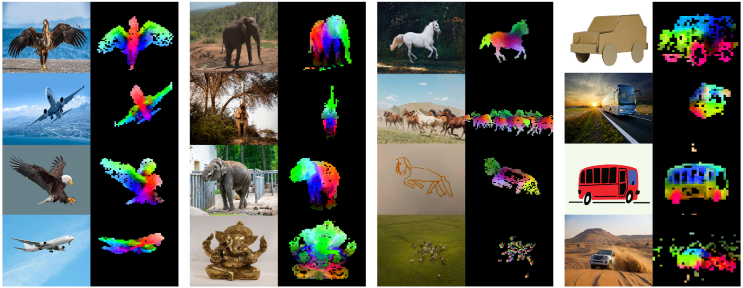
See. Wadoux et al., 2021

Case study : Aging models ?



See. Vela et al., 2022

Perspective : Foundation models ?



(a) (b) (c) (d)

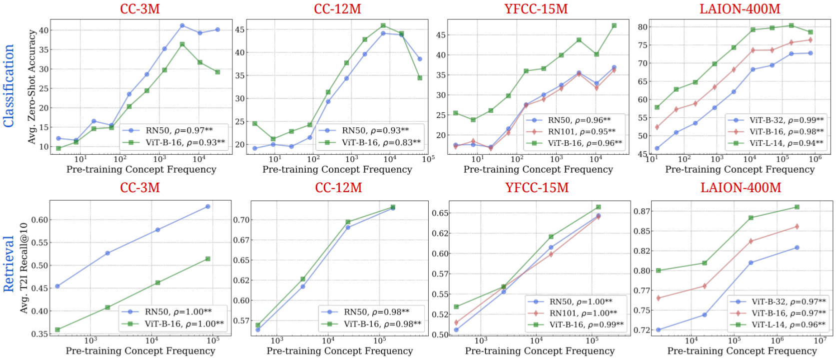
See. Oquab et al., 2023

Perspective : Foundation models ?

A photograph of Anne Graham Lotz included in the training set of Stable Diffusion, a text-to-image model

An image generated by Stable Diffusion using the prompt "Anne Graham Lotz"

Perspective : Foundation models ?



See. Udandarao et al., 2024

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.

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.

Udandarao, Vishaal et al. (2024). **“No zero-shot without exponential data: Pretraining concept frequency determines multimodal model performance”.** In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

References ii

Vela, Daniel et al. (2022). **“Temporal quality degradation in AI models”.** In: *Scientific reports* 12.1, p. 11654.

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.