

Artificial Intelligence in Ecology and Evolution : potential and limits

E2M2 webinar

Paul Tresson

26/06/25

Outline

1. Why use deep learning in ecology ?
2. What are the cases where deep learning does work ?
and other models don't
3. What are the cases where deep learning doesn't work ?
Common traps when working with living things
4. How to sample and evaluate ?
5. Perspectives

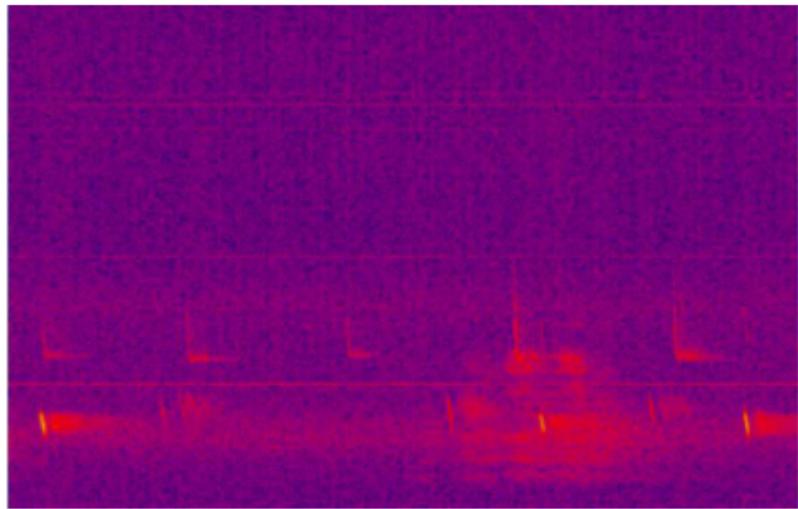
Why use deep learning in ecology ?

More and more data



- UAVs, Satellite

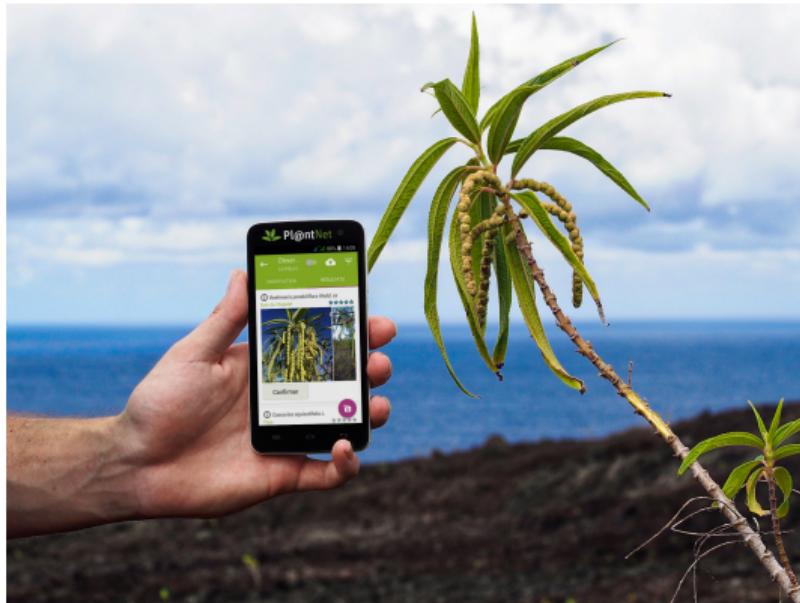
More and more data



- UAVs, Satellite
- Camera trap, acoustic

Mac Aodha *et al.* 2022

More and more data



plantnet.org

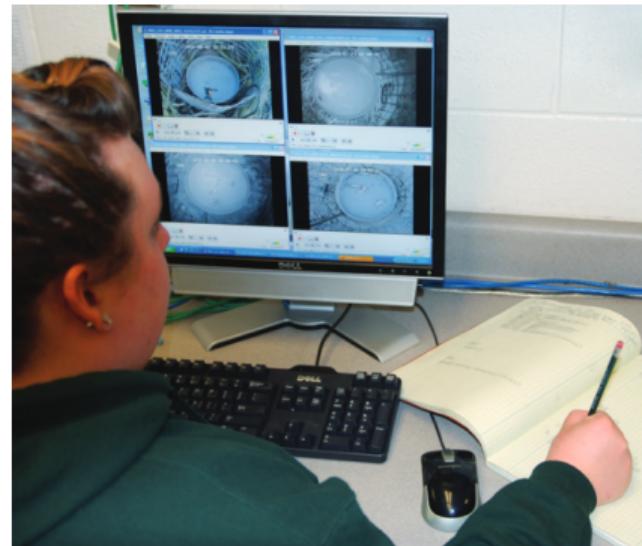
- UAVs, Satellite
- Camera trap, acoustic
- Citizen science

More and more data

- UAVs, Satellite
 - Camera trap, acoustic
 - Citizen science
- **Better coverage, better monitoring**

Data analysis and interpretation is time consuming

- A computer does not sleep
- A computer does not get tired



Grieshop et al. 2012

Data analysis and interpretation is time consuming

- A computer does not sleep
 - A computer does not get tired
- **Automation now possible**



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

xkcd 1425, 2014

**What are the cases where deep
learning does work ?**

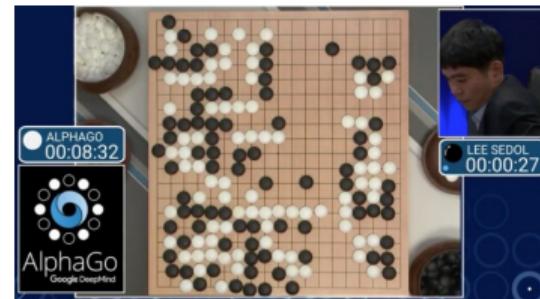
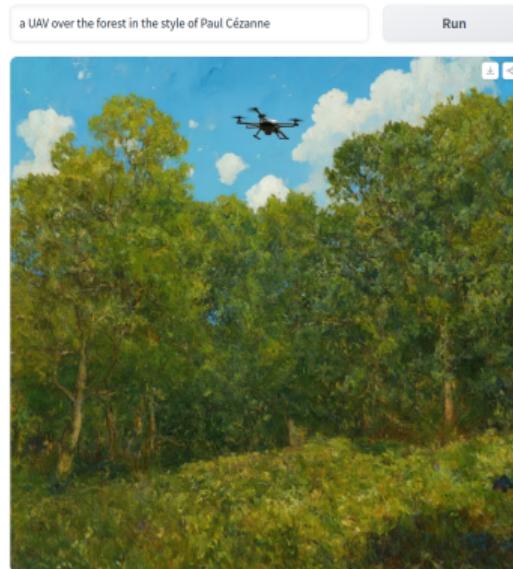
and other models don't

Model complex, non-linear and abstract relationships

Hi can you write me a haiku about forest with a subtle reference to Asterix ?

 Tall trees whispering,
Mischief stirs in Gaulish woods—
Magic roots run deep.



Interpolation and generalisation



Unonopsis stipitata Diels

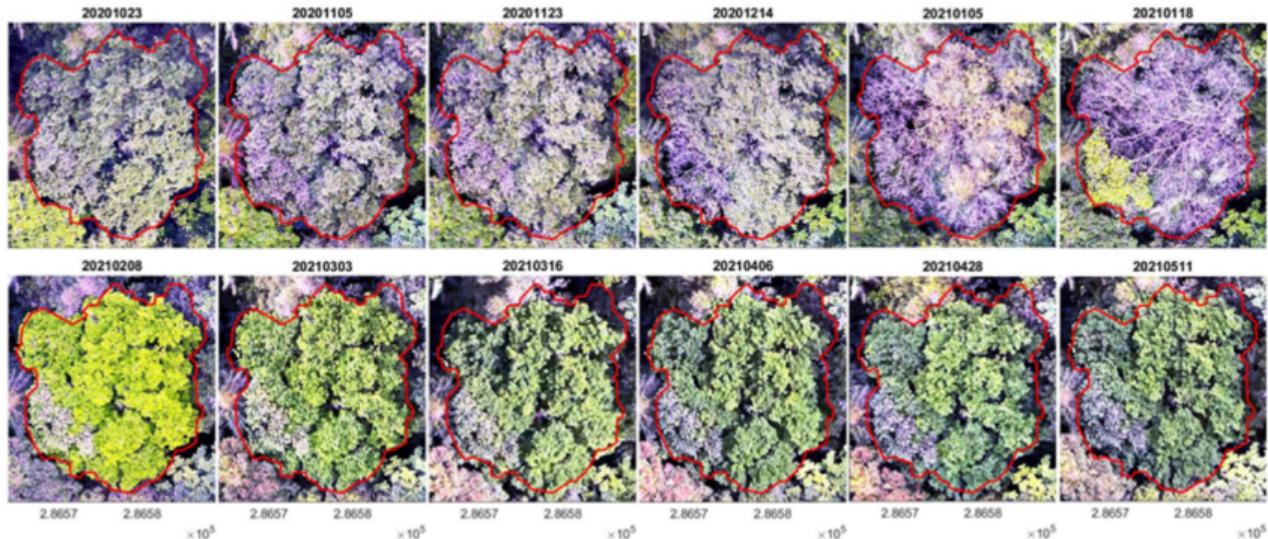
PlantClef 2020 Dataset

What are the cases where deep learning doesn't work ?

Common traps when working with living things

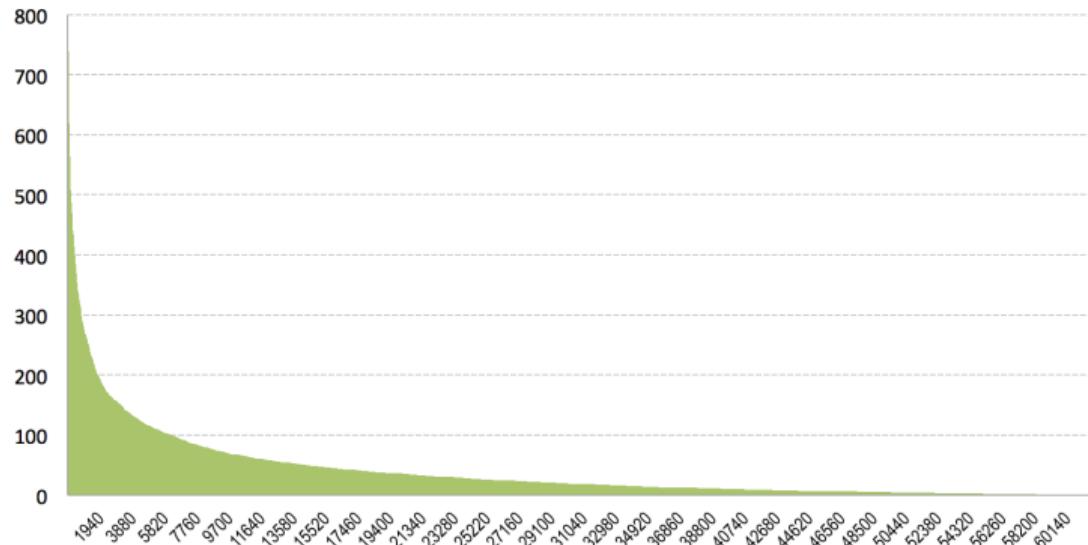
Constraints in ecology

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

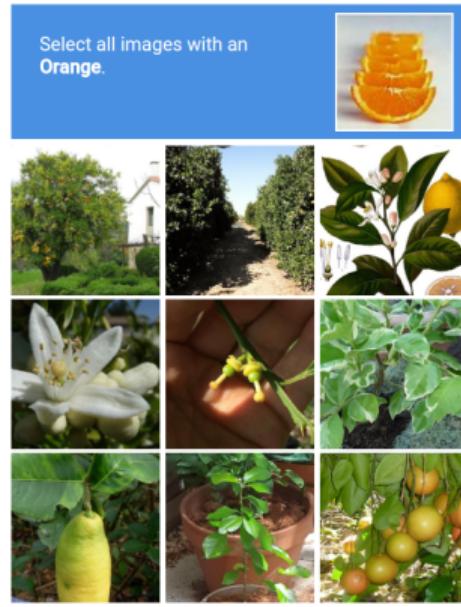
Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

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.

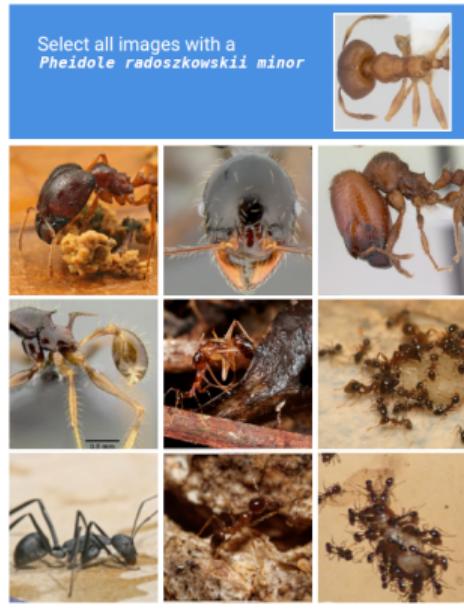


Verify

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



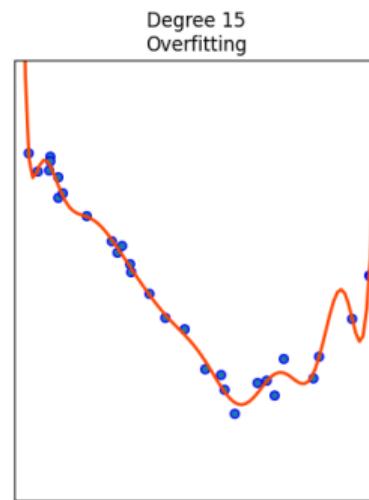
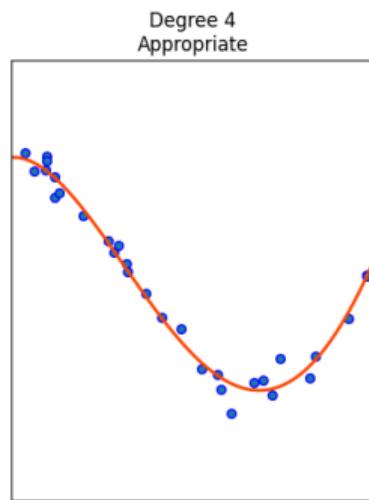
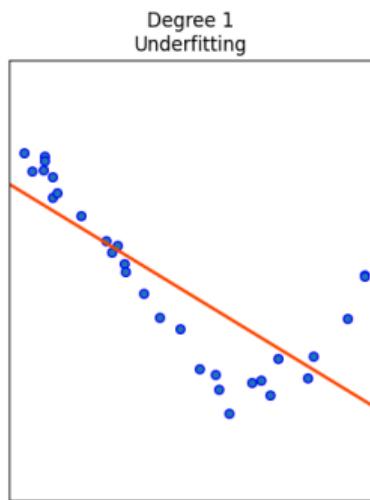
The grid contains nine images of ants, likely Pheidole species, used for a classification task. The images include various views of ants, some in groups, some in close-up, and some in their natural habitat. One image in the top row is clearly a Pheidole radoszkowskii minor ant, while the others are either different species or different morphs of the same species.



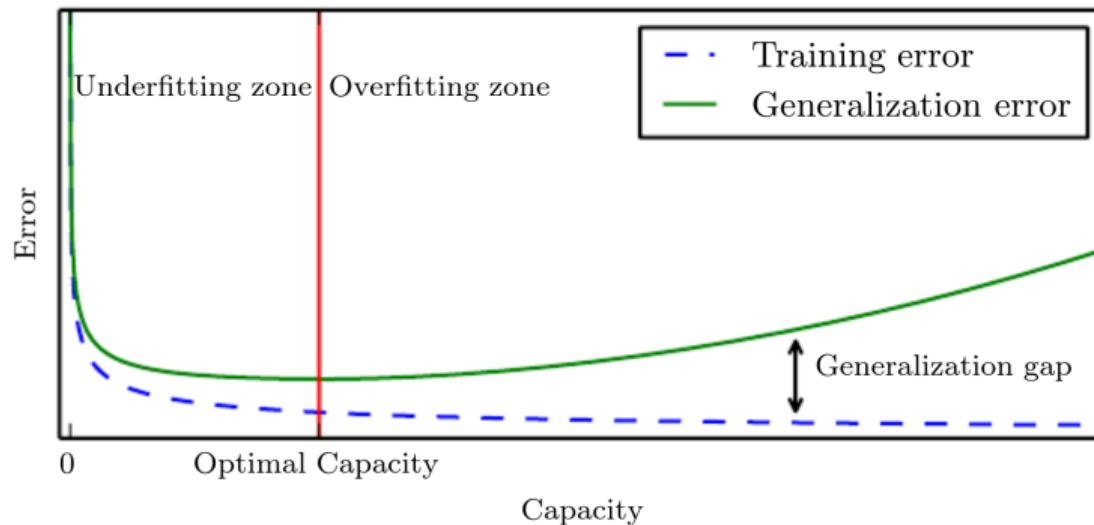
Verify

Overfitting

Overfitting



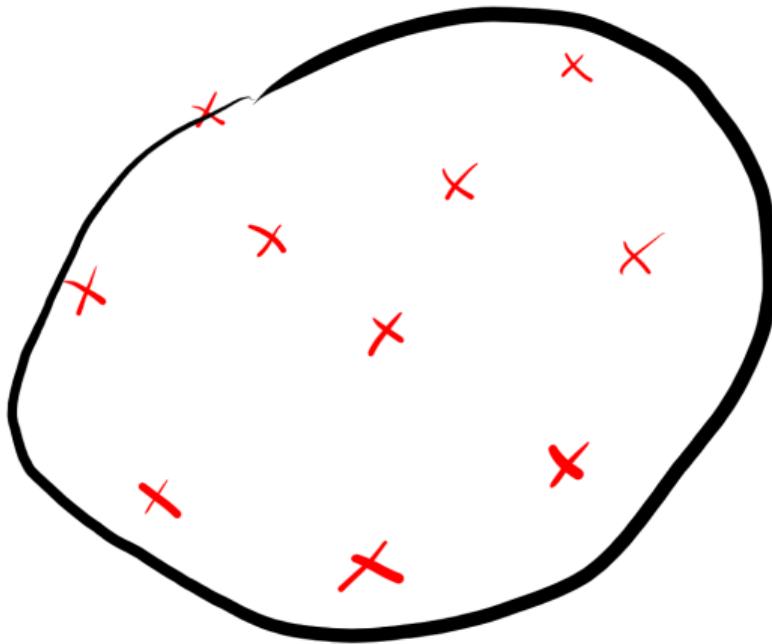
Overfitting



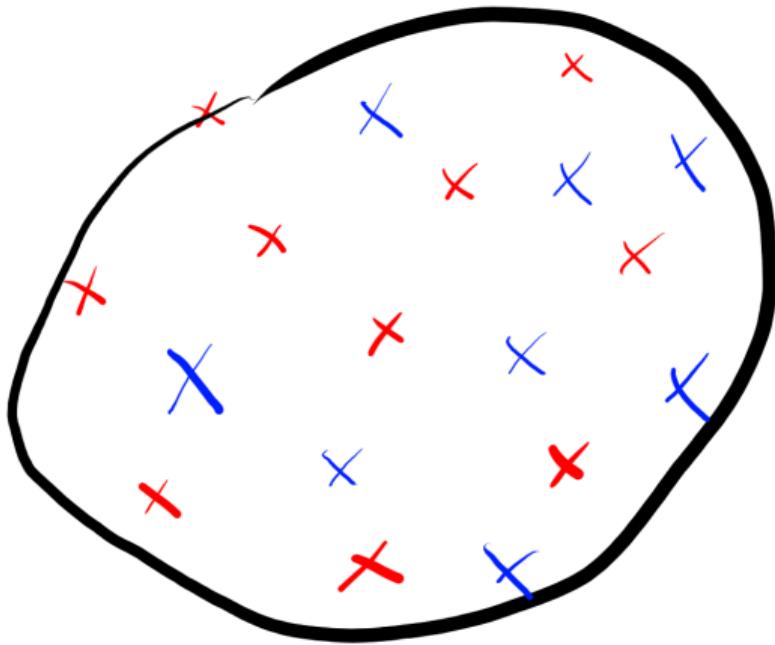
Goodfellow et al., 2016



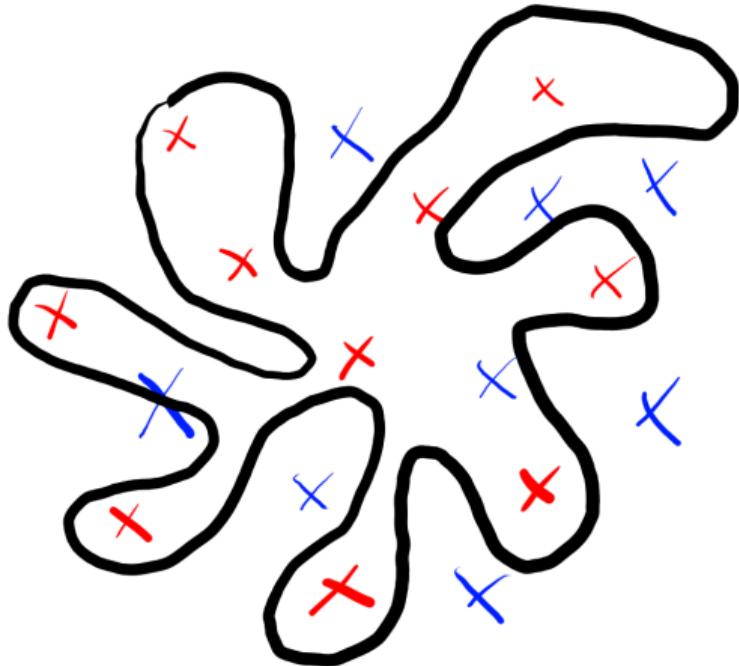
Train set



A good fitted model



Test set



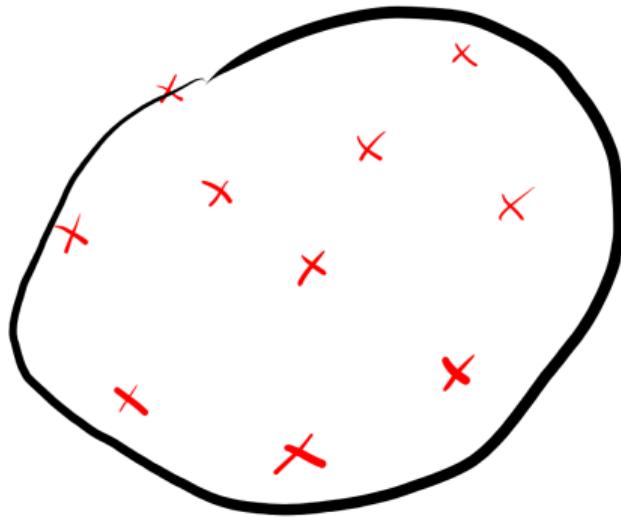
Overfitting

Biases

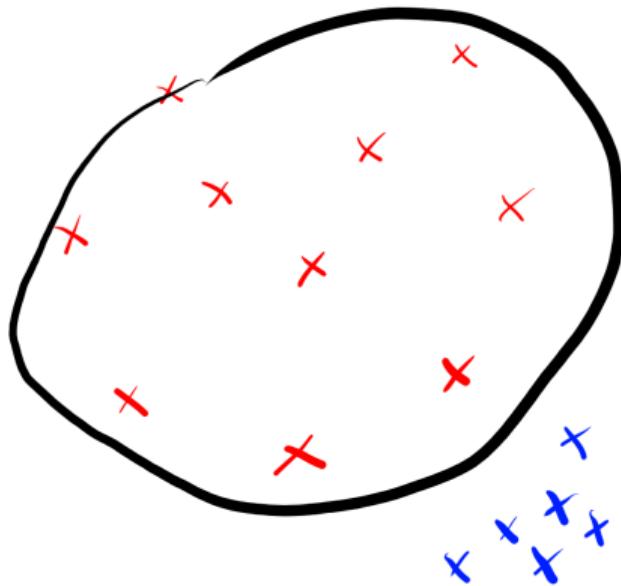
Biases in the train set



Biases in the train set



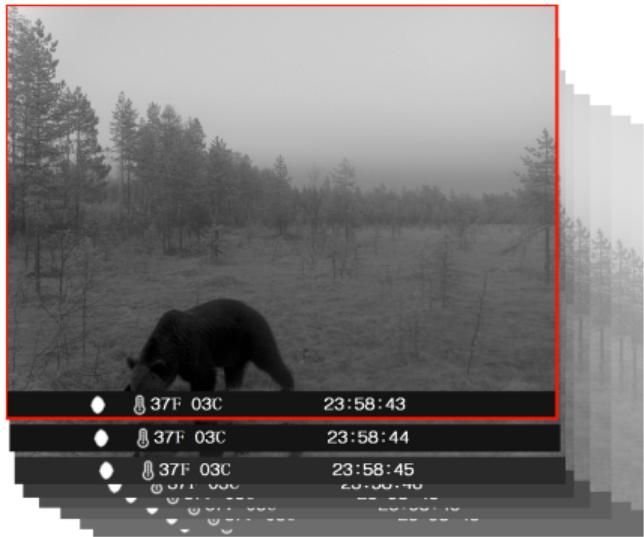
Biases in the train set



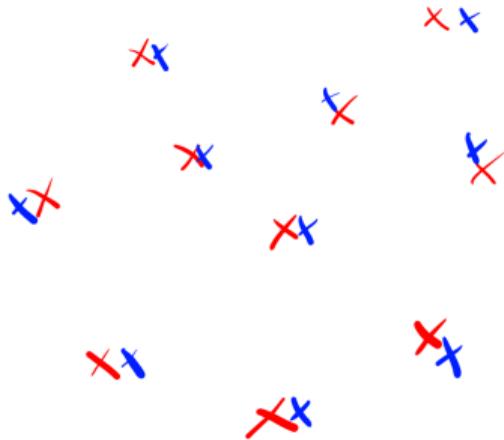
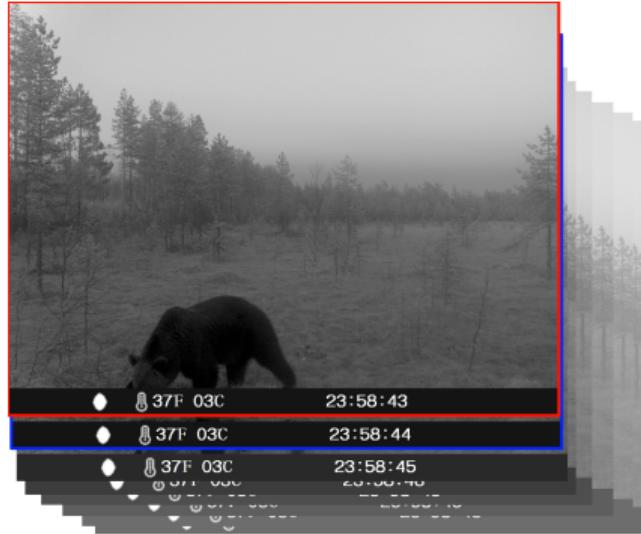
Biases in the train set - autocorrelation



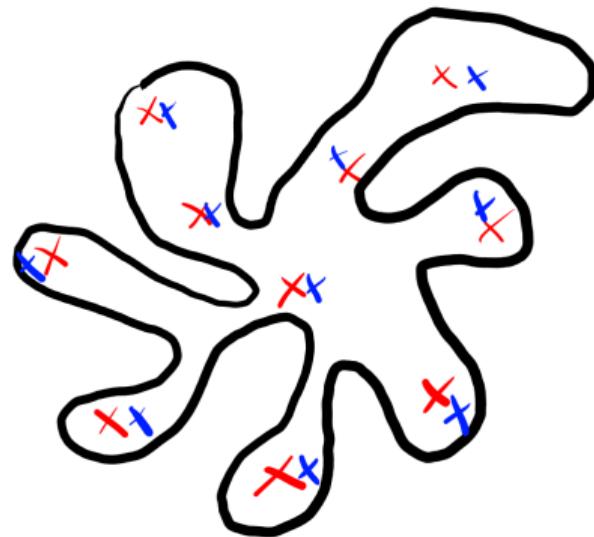
Biases in the train set - autocorrelation



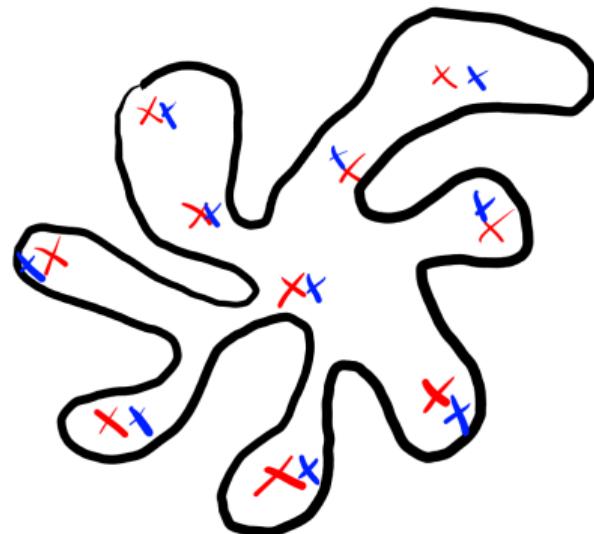
Biases in the train set - autocorrelation



Biases in the train set - autocorrelation



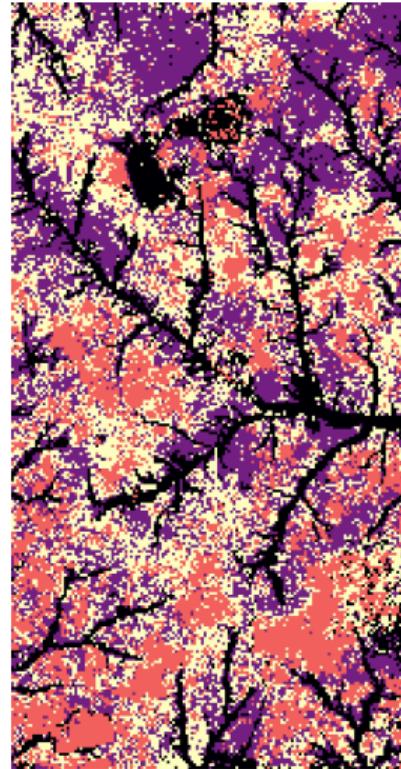
Biases in the train set - autocorrelation



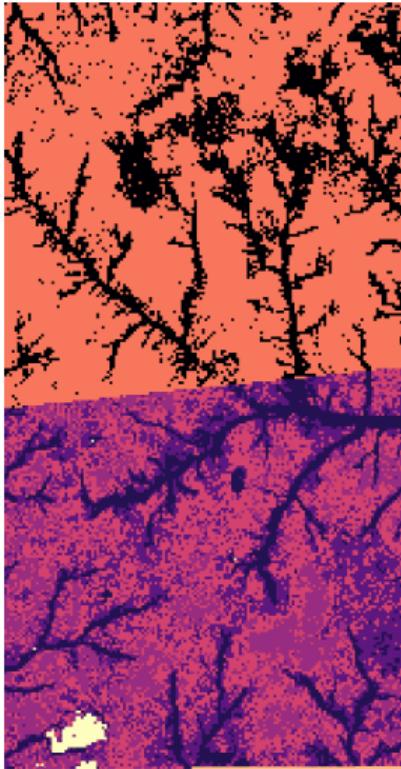
Biases in the train set - Spatial autocorrelation



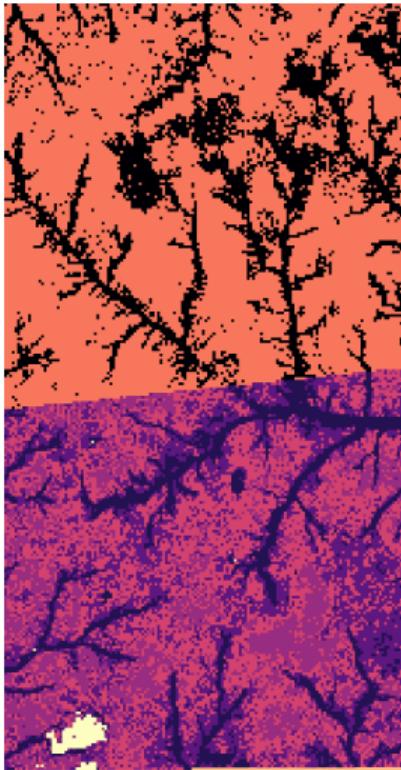
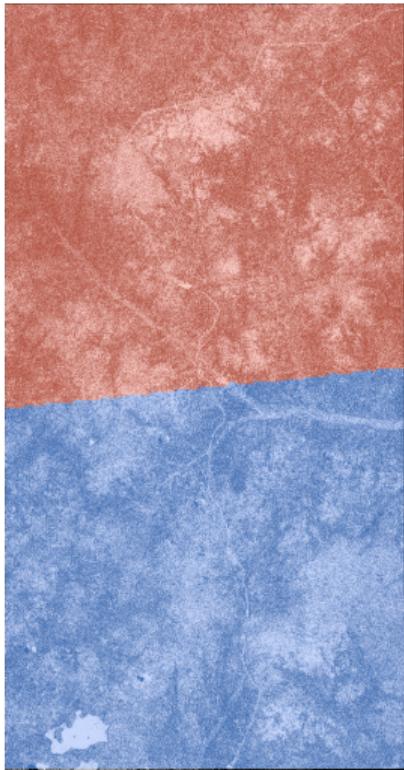
Biases in the train set - Spatial autocorrelation



Biases in the train set - Spatial autocorrelation



Biases in the train set - Spatial autocorrelation

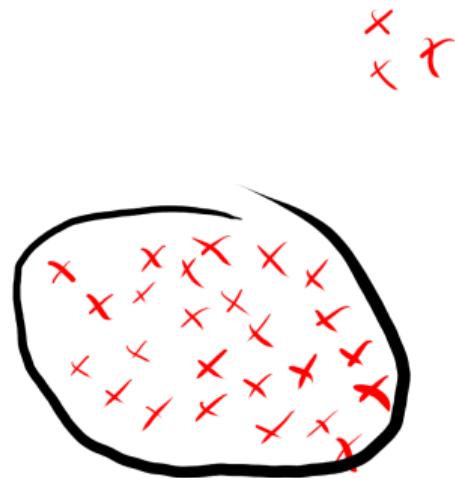


Unbalanced data

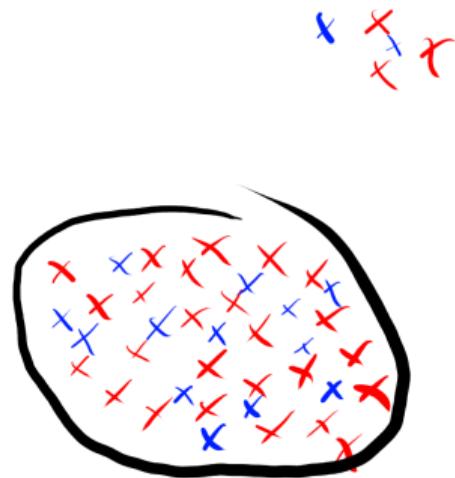
Unbalanced data



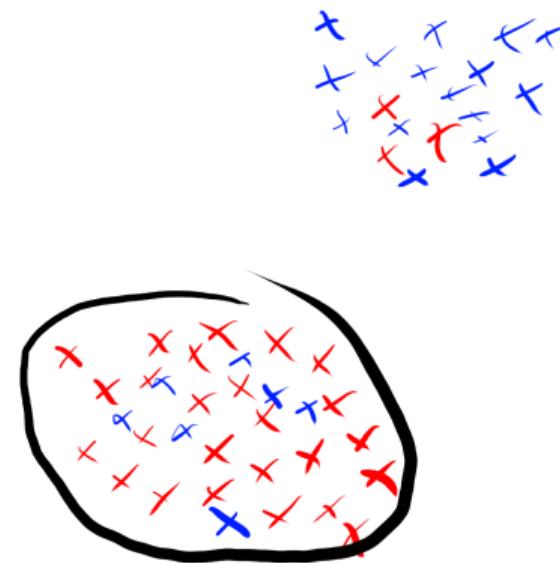
Unbalanced data



Unbalanced data



Unbalanced data



Deal with unbalanced data

- Oversample ?



Deal with unbalanced data

- Oversample ?



Deal with unbalanced data

- Oversample ?



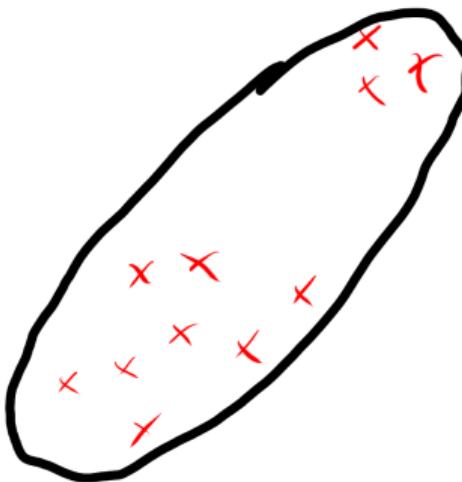
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



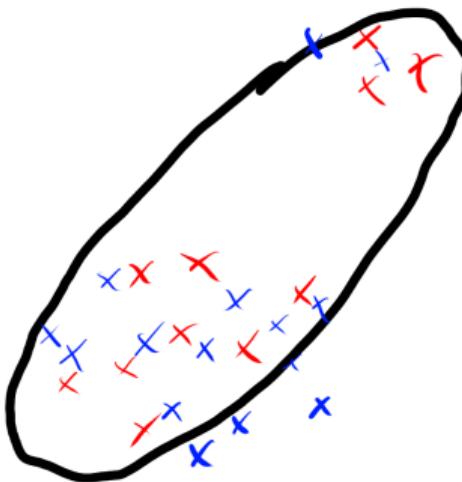
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



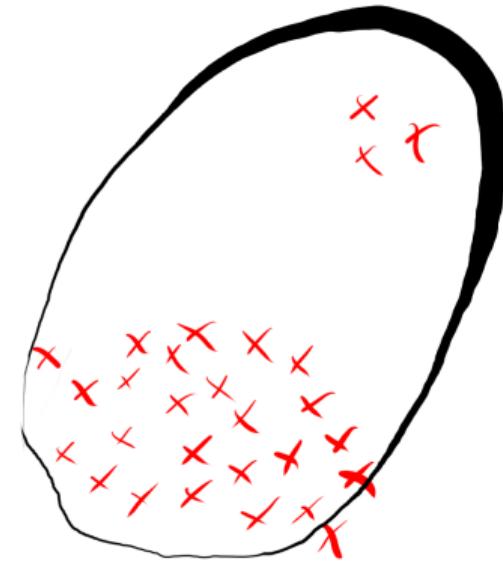
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



Deal with unbalanced data

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



Deal with lack of data

- Data augmentation



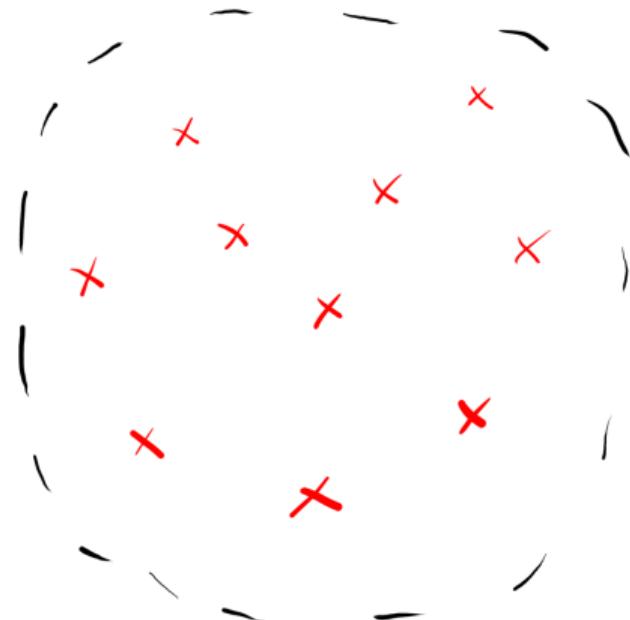
Deal with lack of data

- Data augmentation



Deal with lack of data

- Data augmentation
- Pretrained model

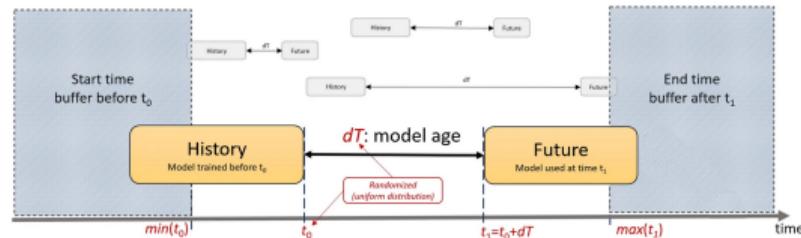


Deal with lack of data

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

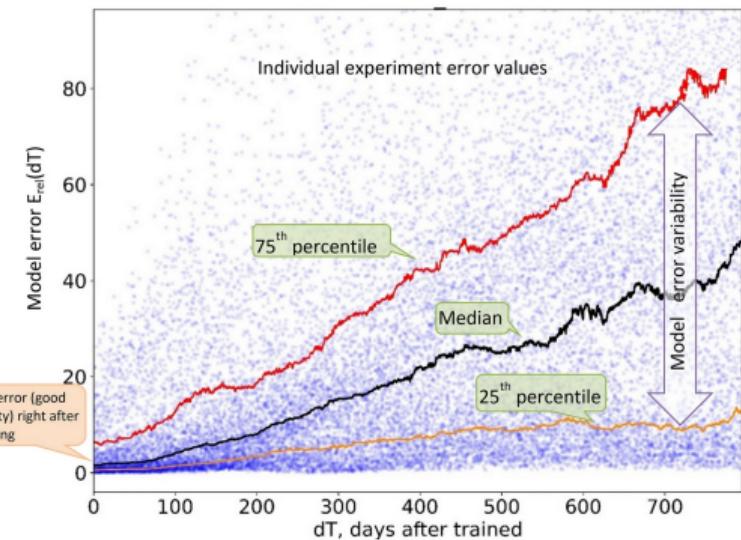
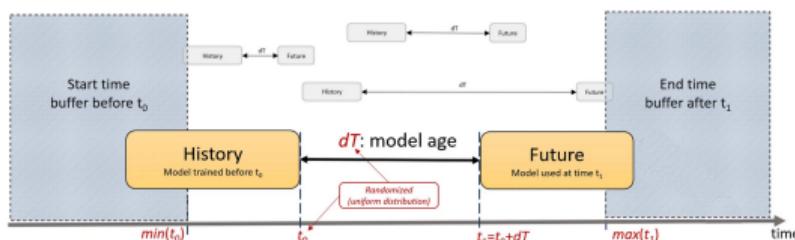
Out of distribution

Out of distribution : Evolution with time



Adapted from Vela et al. 2022

Out of distribution : Evolution with time



Adapted from Vela et al. 2022

Out of distribution : Global changes

Conditions will evolve in never seen before conditions:

- Given ecosystem in unprecedented climatic conditions

Out of distribution : Global changes

Conditions will evolve in never seen before conditions:

- Given ecosystem in unprecedented climatic conditions
- Species migrate/invoke in new territories

Out of distribution : Invasive species

New unknown species in the training test appears in a region.

- False Positive : confusion with known species

Out of distribution : Invasive species

New unknown species in the training test appears in a region.

- False Positive : confusion with known species
- False Negative : model misses the new species

Out of distribution : Invasive species

New unknown species in the training test appears in a region.

- False Positive : confusion with known species
- False Negative : model misses the new species
- Handmade check on model confidence

Need to be very careful 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 ?

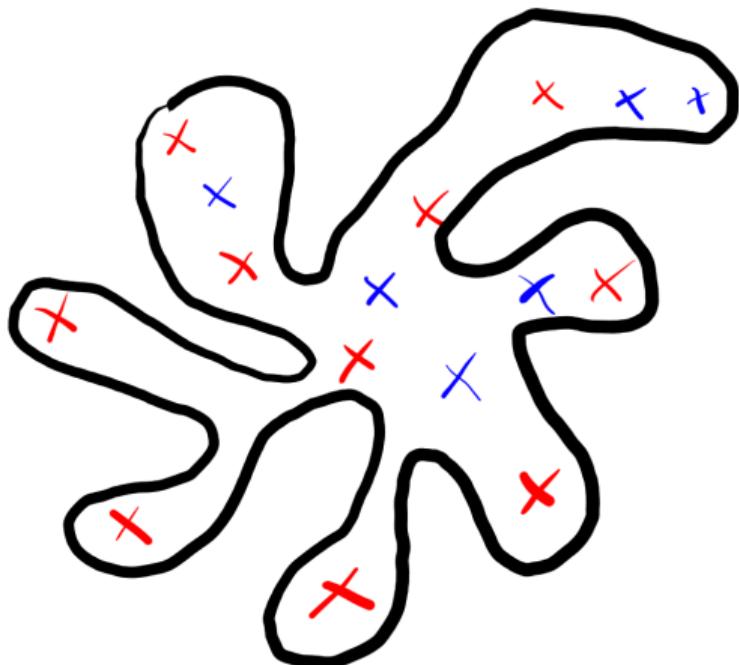
“random split training validation 80/20”

For the uncurated dataset, we randomly sample 142 million images

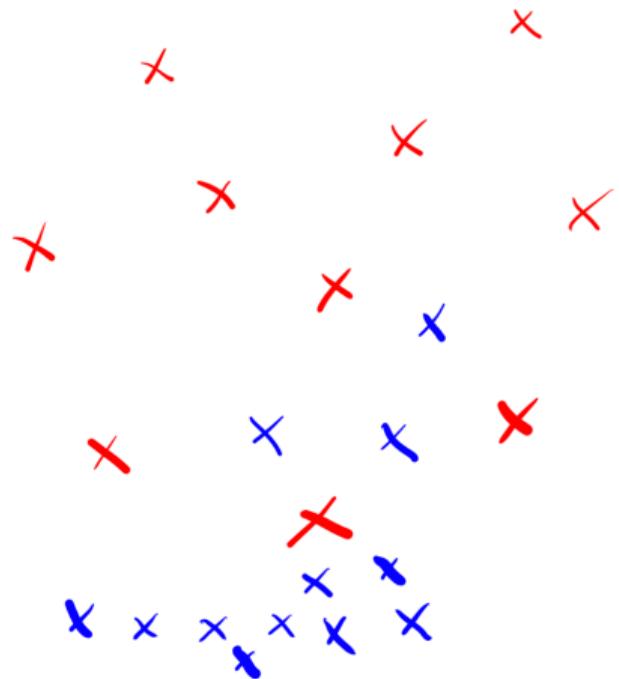
Oquab et al., 2023

Works for huge DL papers, maybe not for you

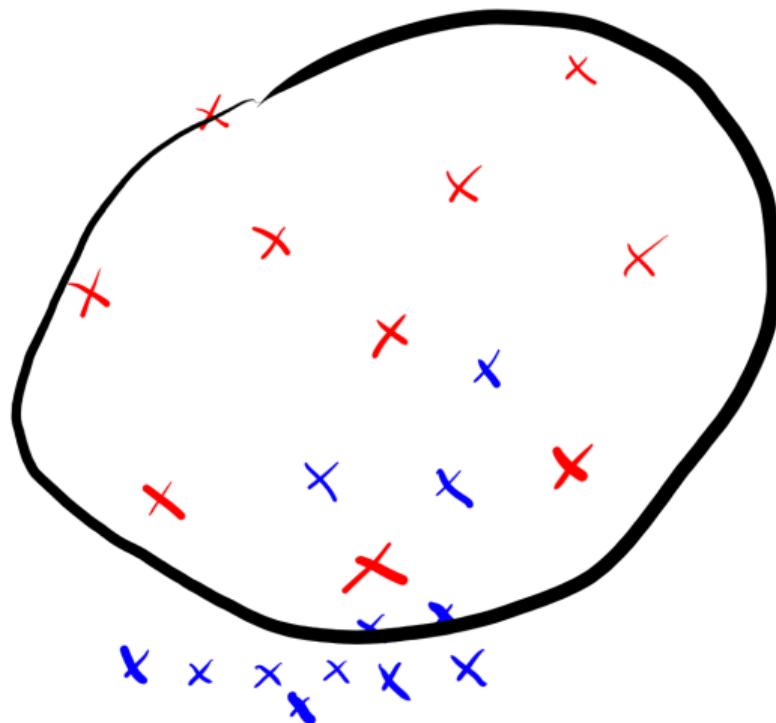
Overfitting the test set



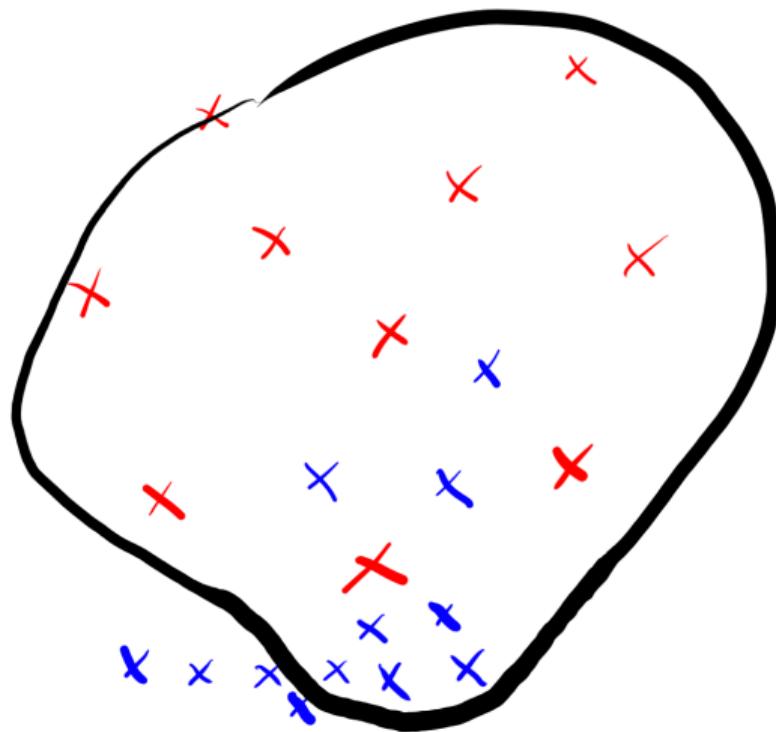
Overfitting the test set



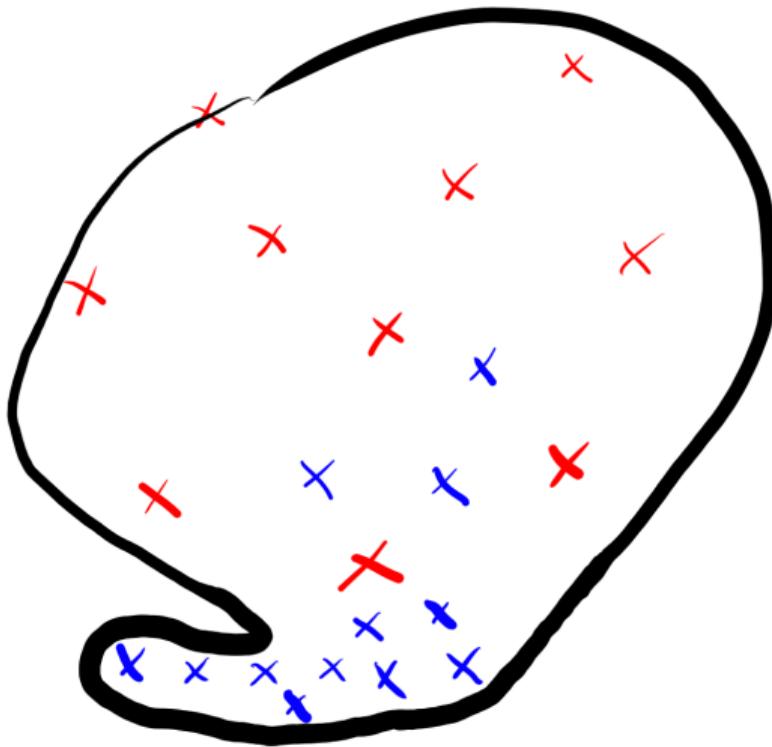
Overfitting the test set



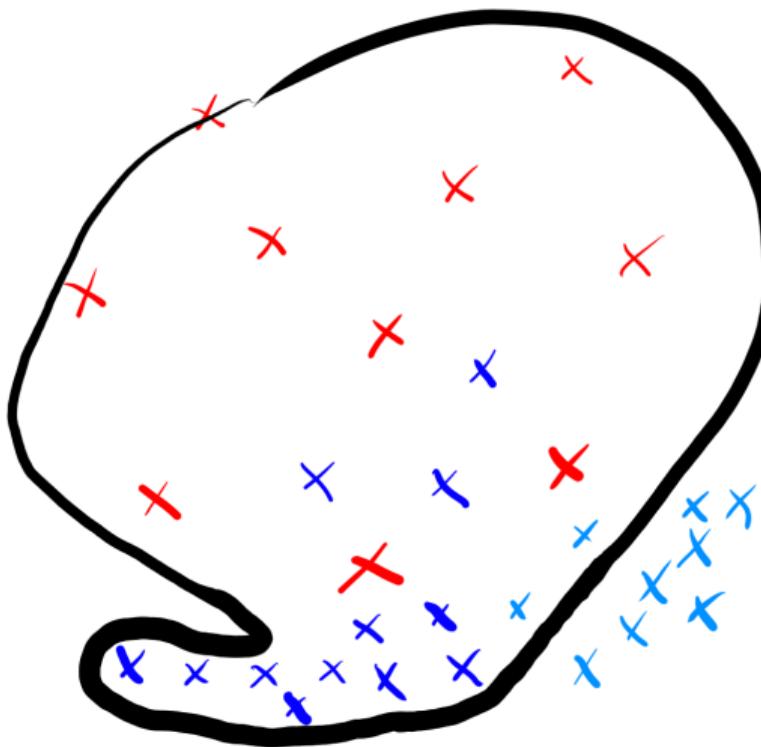
Overfitting the test set



Overfitting the test set



Overfitting the test set



Cross-validation

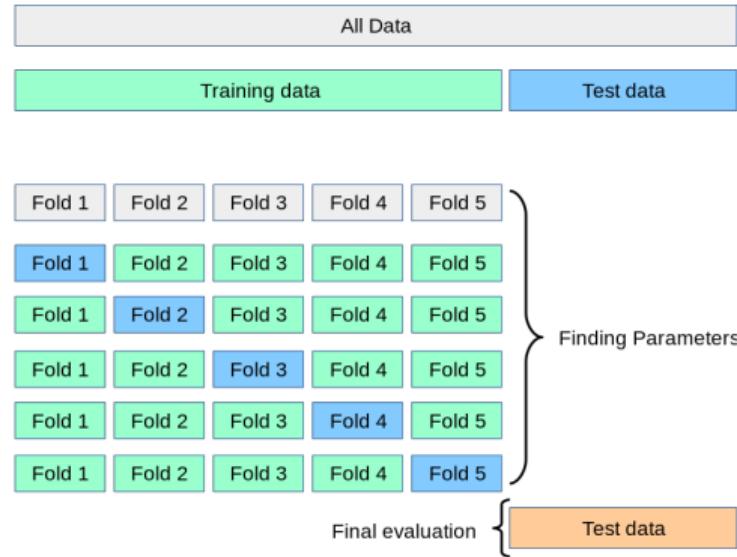


Figure from scikit-learn docs

Cross-validation

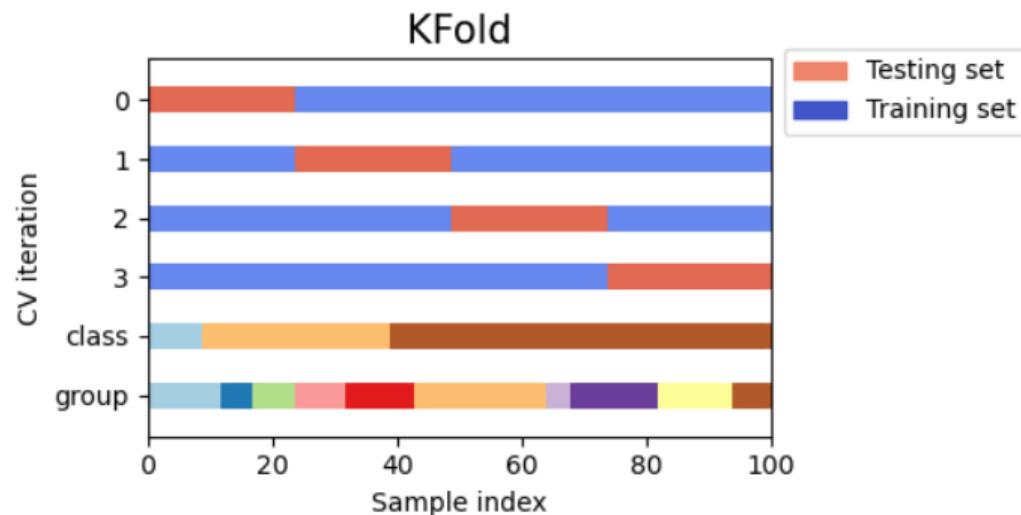


Figure from scikit-learn docs

Cross-validation

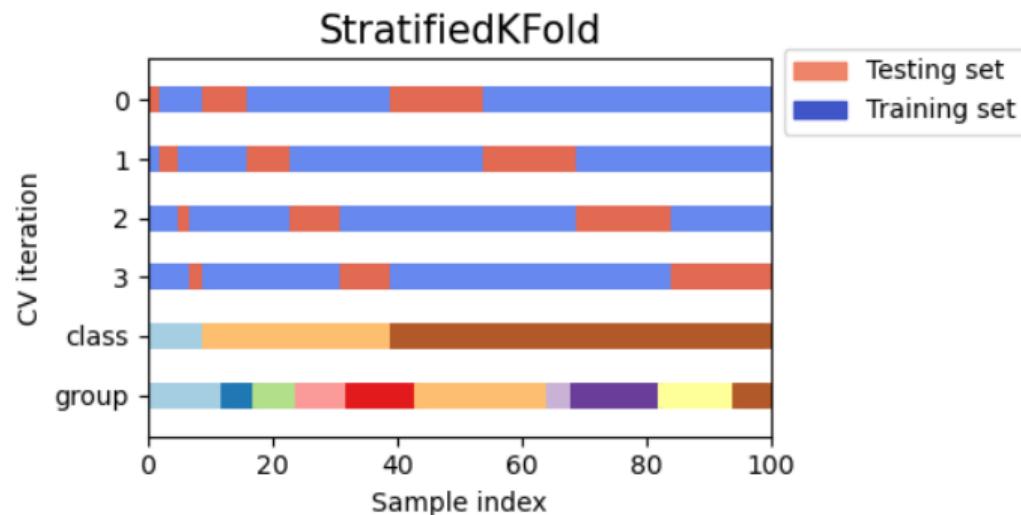
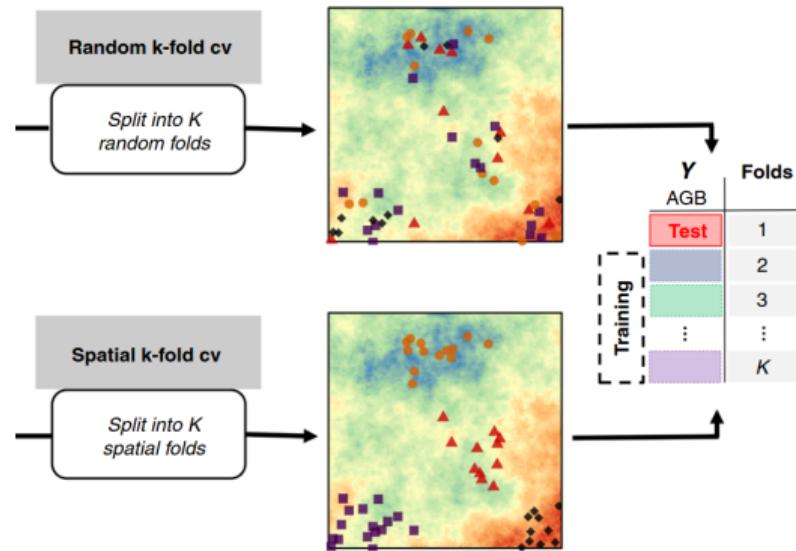


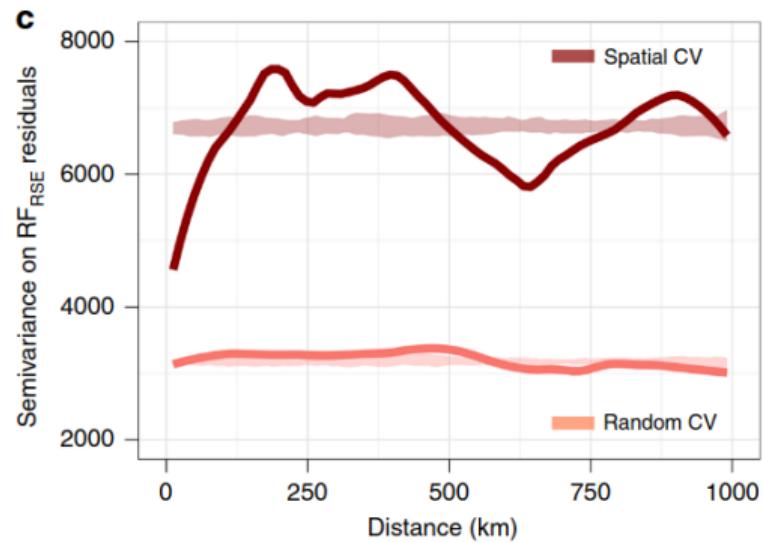
Figure from scikit-learn docs

Spatial cross-validation



See. Ploton et al., 2020

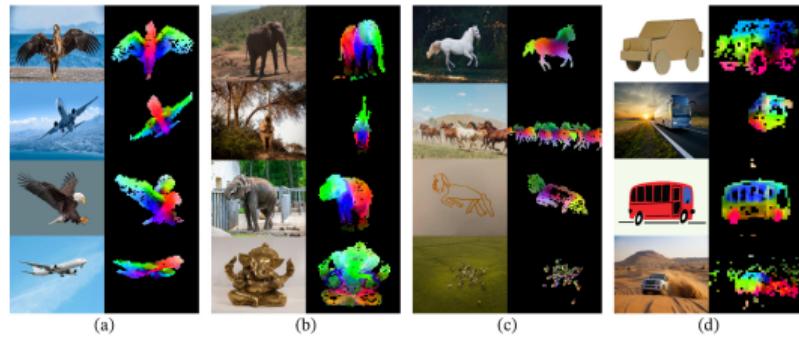
Spatial cross-validation



See. Ploton et al., 2020

Perspectives

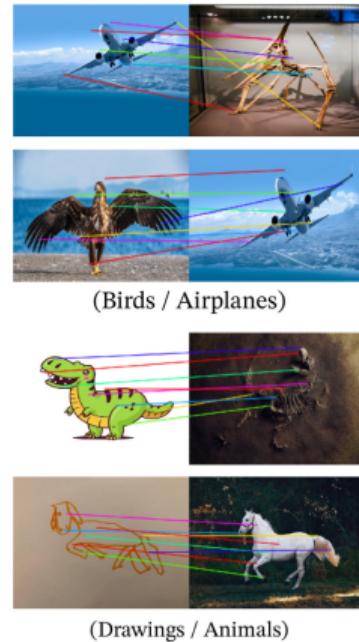
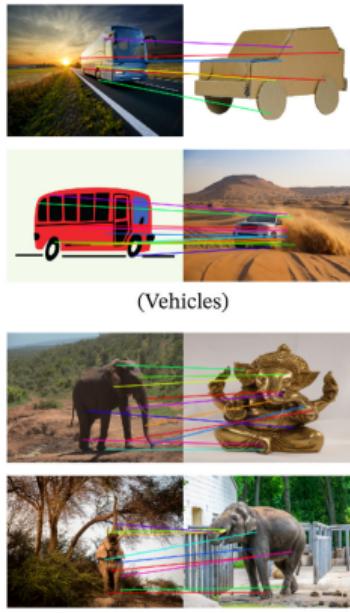
Models are more robust and generalist



- Self-supervised Learning (Pre-training)

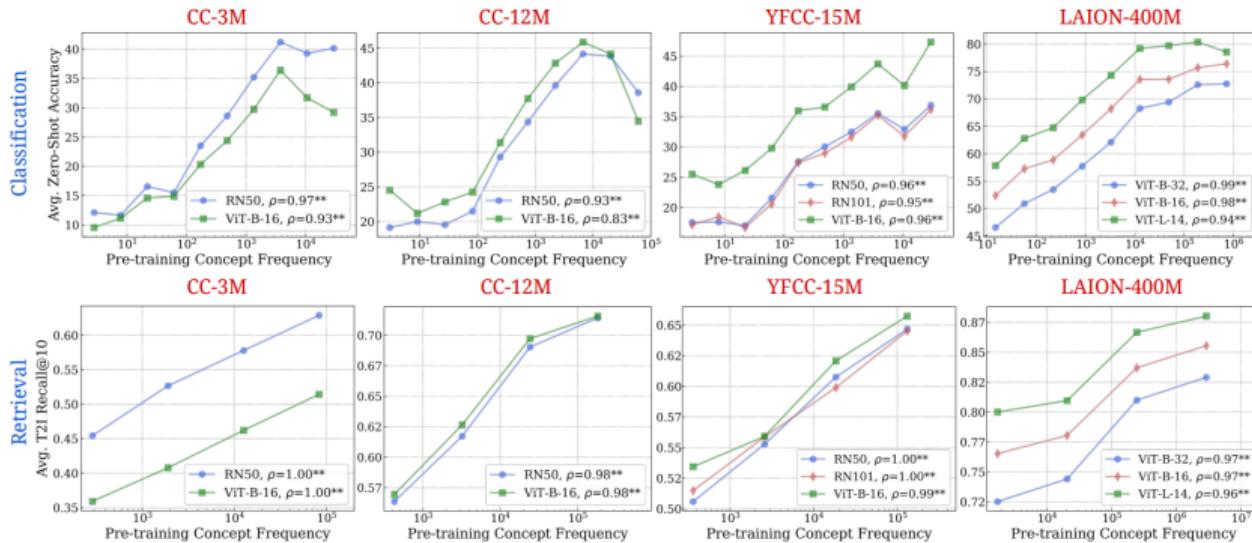
Oquab *et al.* 2024

Models are more robust and generalist



- Self-supervised Learning (Pre-training)
- Better performances and robustness

Zero-shot at a cost



Udandarao et al. 2024

Not only deep learning

Conclusion

Should I use deep learning in my research ?

- ✓ Lot of incoming data ✗ Need for explainability
- ✓ Low-level data ✗ Need for certainty
- ✓ Cumbersome but (relatively) easy to analyse ✗ Need for reliability

Thank you for your attention !

Any questions?

Useful ressources

State of the art

- Huggingface
- PapersWithCode

Getting started

- Pytorch
- Keras

Understanding papers

- Yannic Kilcher
- AI coffe break

Understanding visually

- 3blue1brown
- deepia