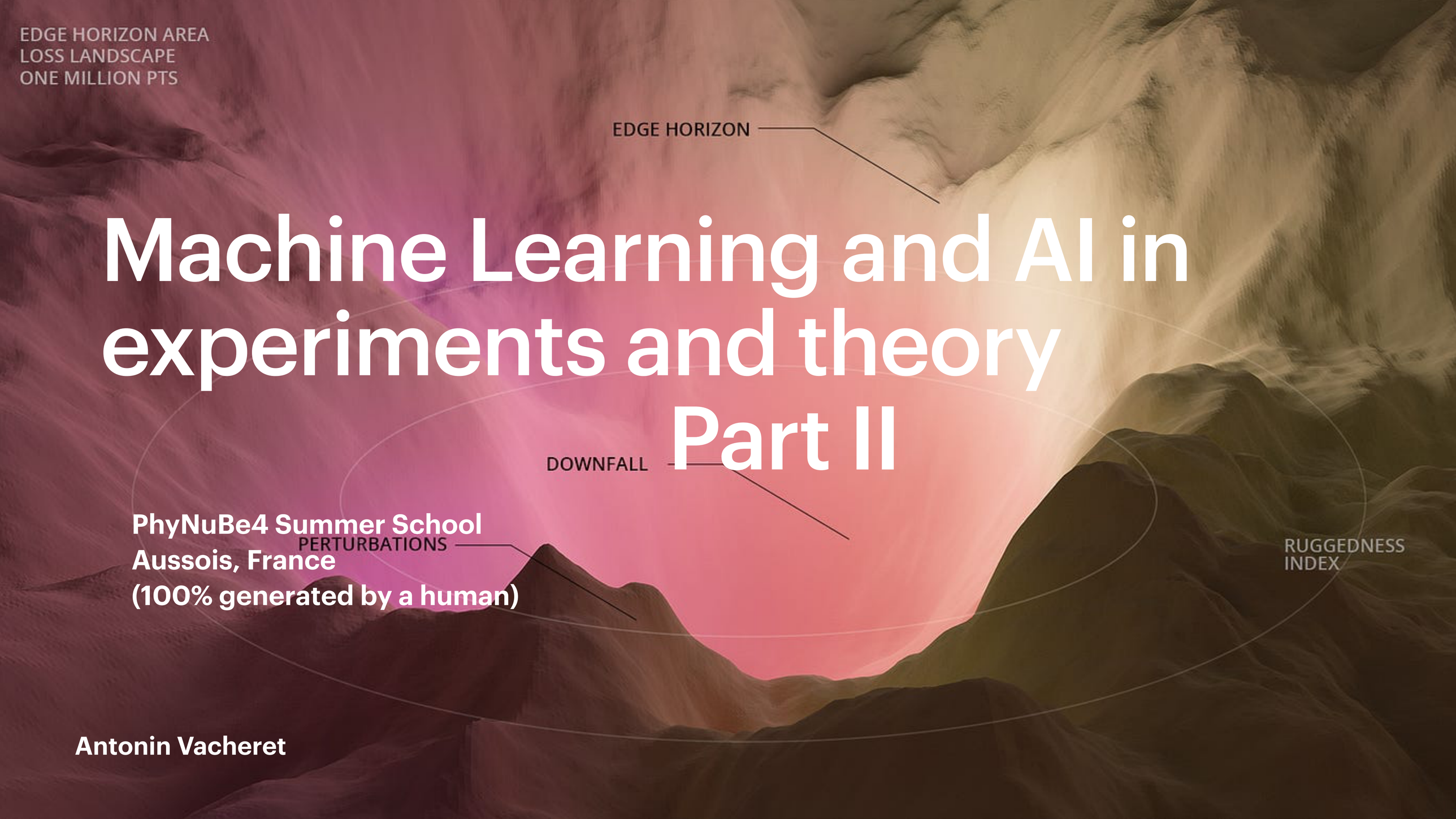


Machine Learning and AI in experiments and theory Part II

PhyNuBe4 Summer School
Aussois, France
(100% generated by a human)

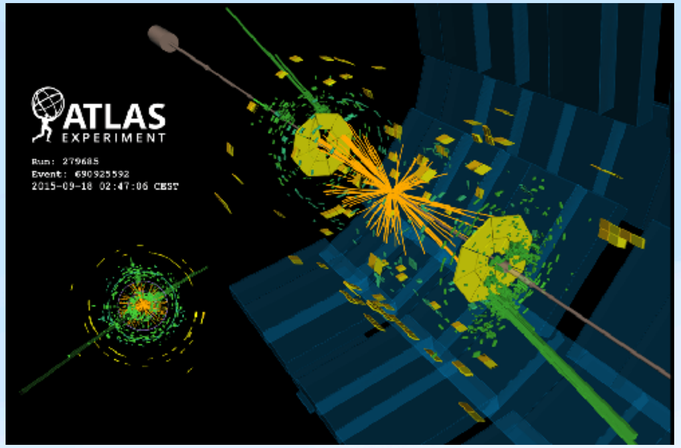
Antonin Vacheret

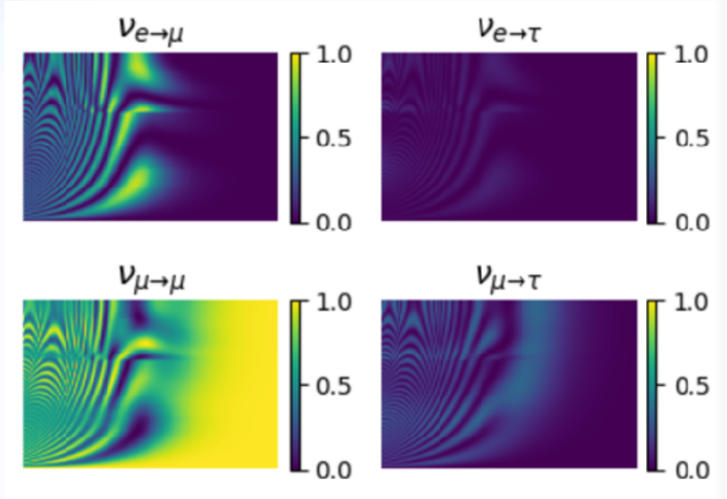
This Lecture

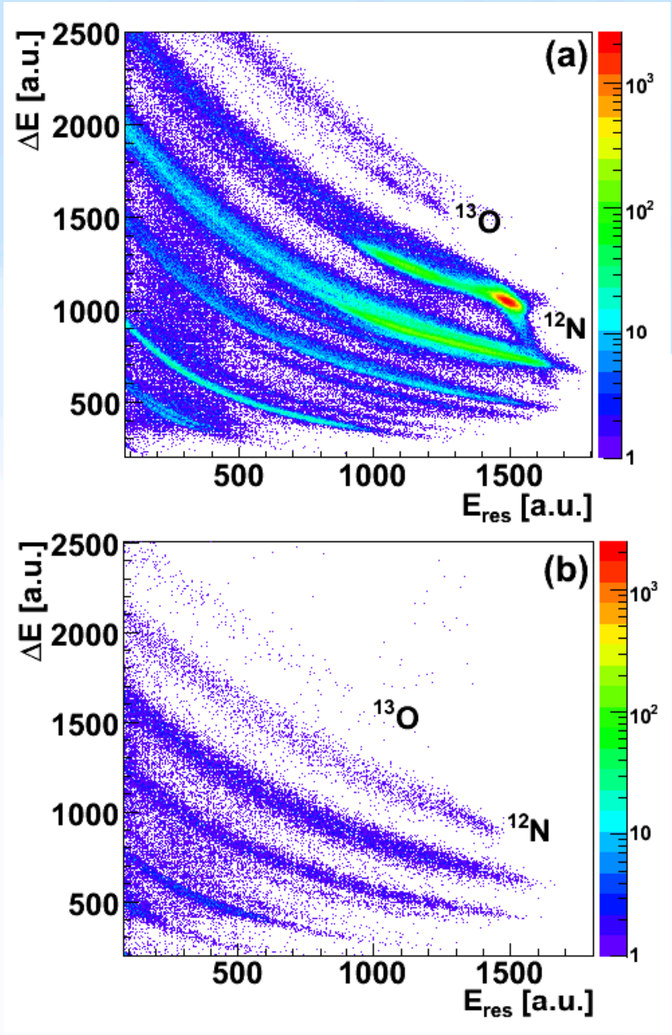
- Data & model dimensions
- Regularisation techniques
- Quantifying uncertainties

Data features and representation

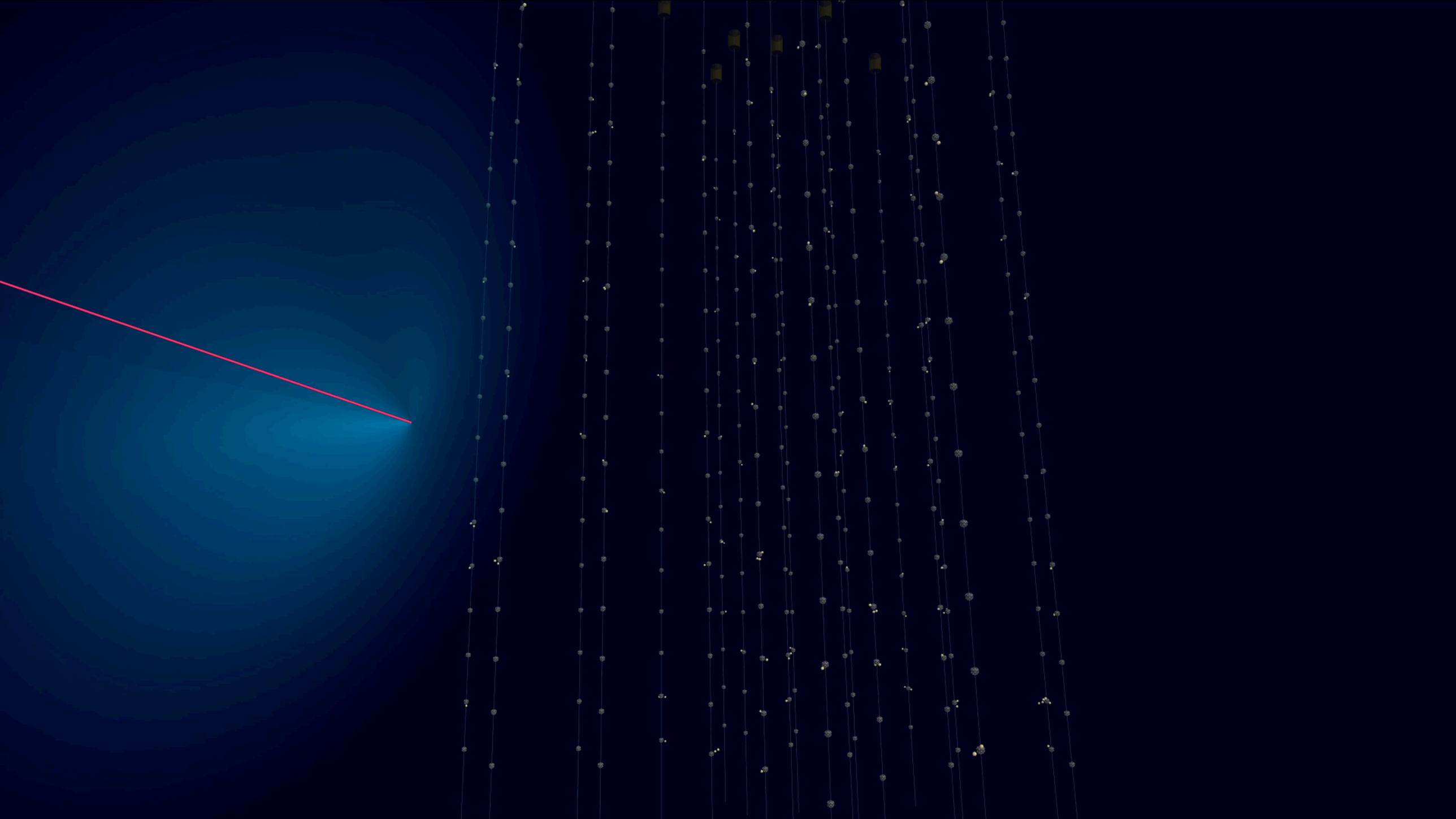
- Nuclear & Particle physics data is structured with features that directly comes from the underlying physics phenomenon
  - It can be sparse and multi-dimensional
  - It exhibits specific symmetries
    - It may be possible to compress it to better representation
- Estimating the number of features in a dataset is not trivial







Antonin Vacheret



How do I dimension my model based on my data ?

… for a fixed compute budget

- In Supervised learning we need training samples, but how many ?

How do I dimension my model based on my data ?

… for a fixed compute budget

- In Supervised learning we need training samples, but how many ?
- Well it depends on the model size
- What model size should I use ?

Antonin Vacheret

How do I dimension my model based on my data ?

… for a fixed compute budget

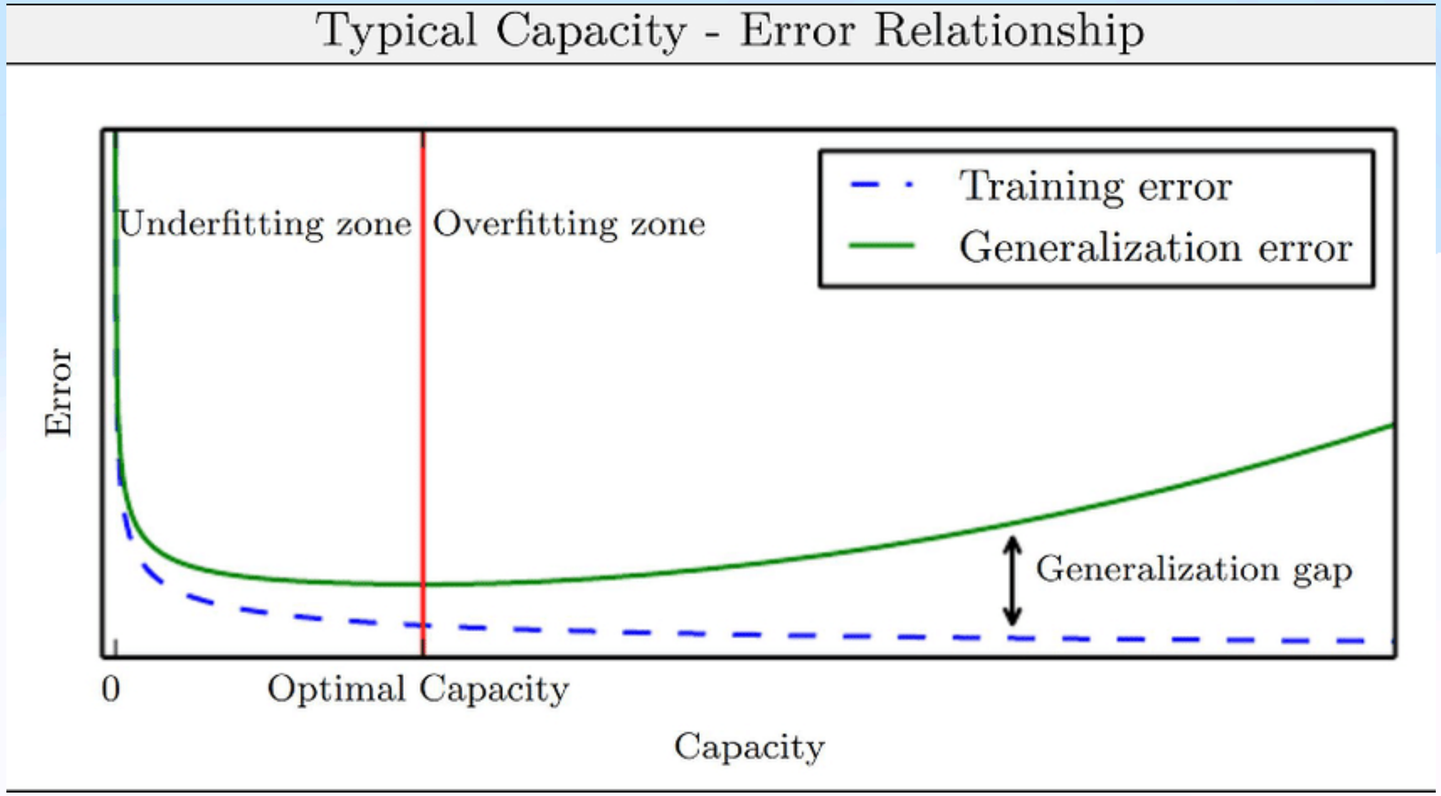
- In Supervised learning we need training samples, but how many ?
- Well it depends on the model size
- What model size should I use ?
- Well …

Antonin Vacheret

Understanding the training process

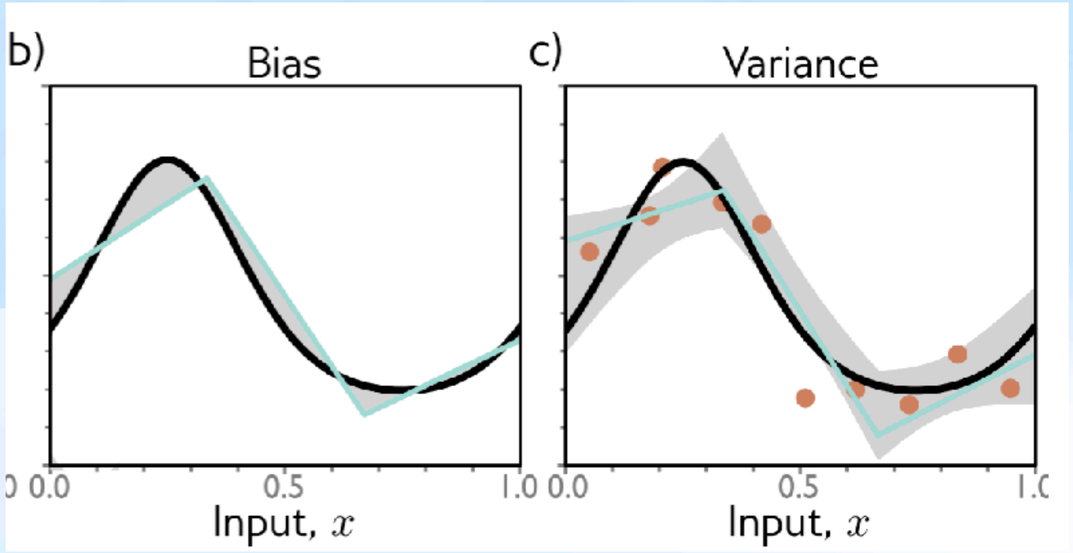
The bias-variance trade-off

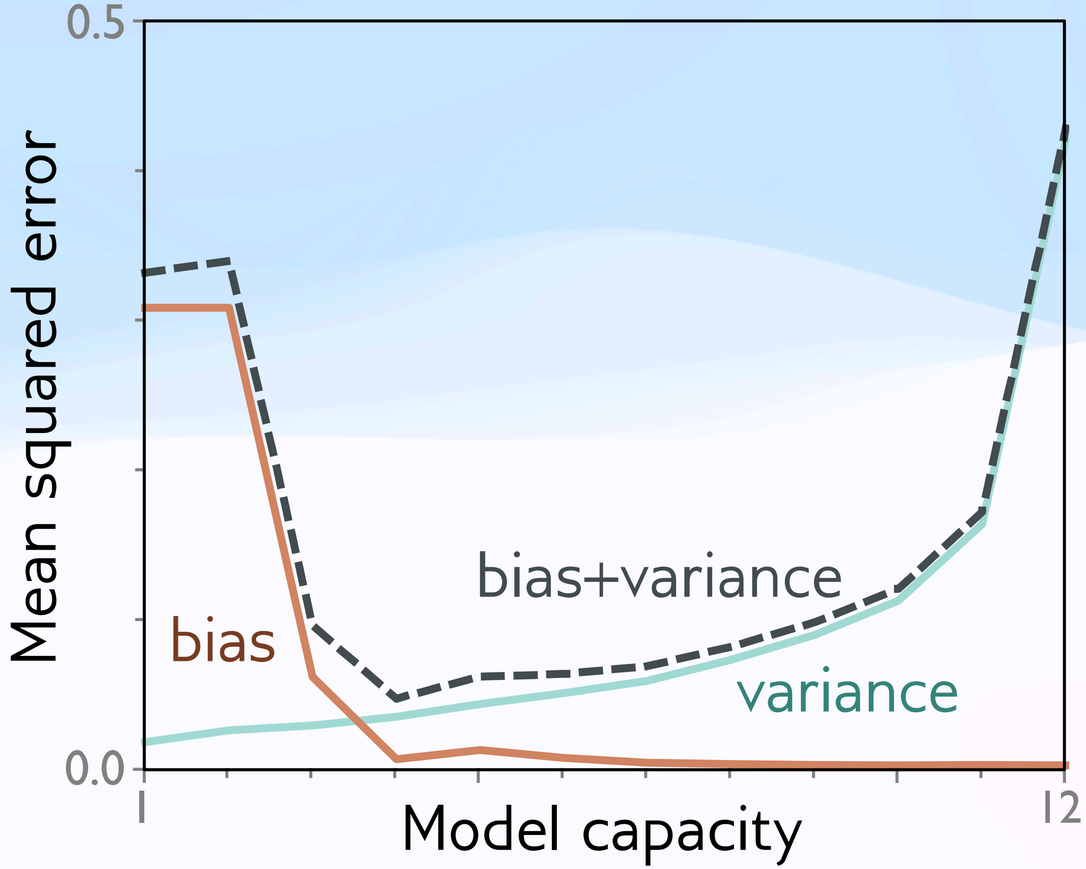
- Training a model is not just a minimisation process :
  - We want to model to learn to generalize on unseen data
  - Minimise the generalization gap
- This is the reason for splitting data in training, validation and test sets !



Understanding the training process

The bias-variance trade-off

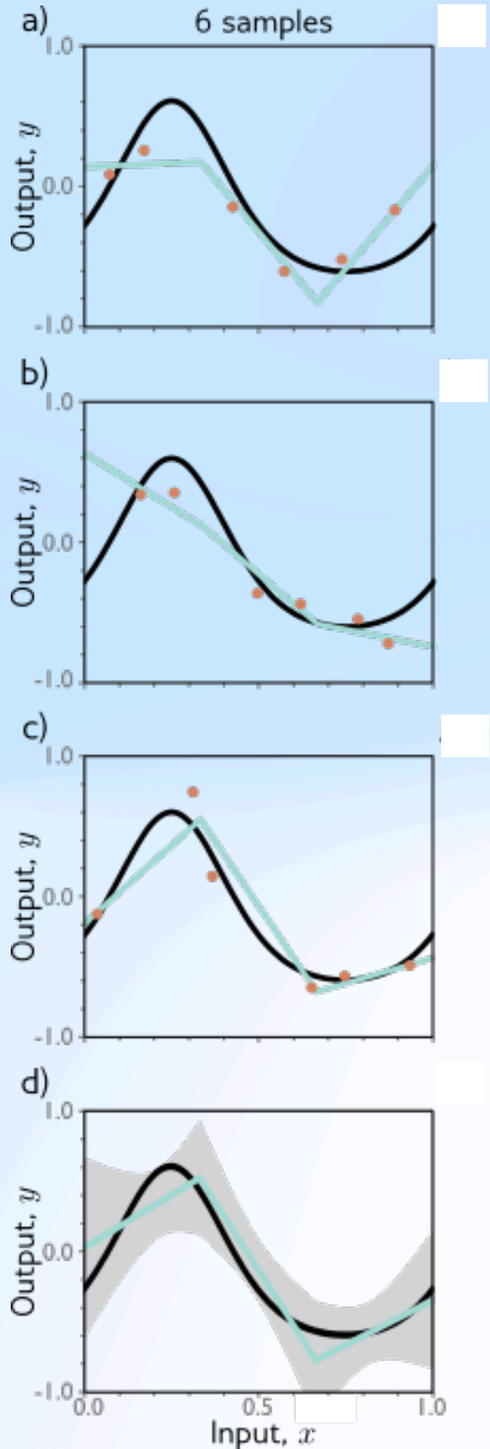




- Epistemic uncertainties : Biais and variance originated from the model training and generalisation process
- Can be reduced by model optimisation, more data, regularisation

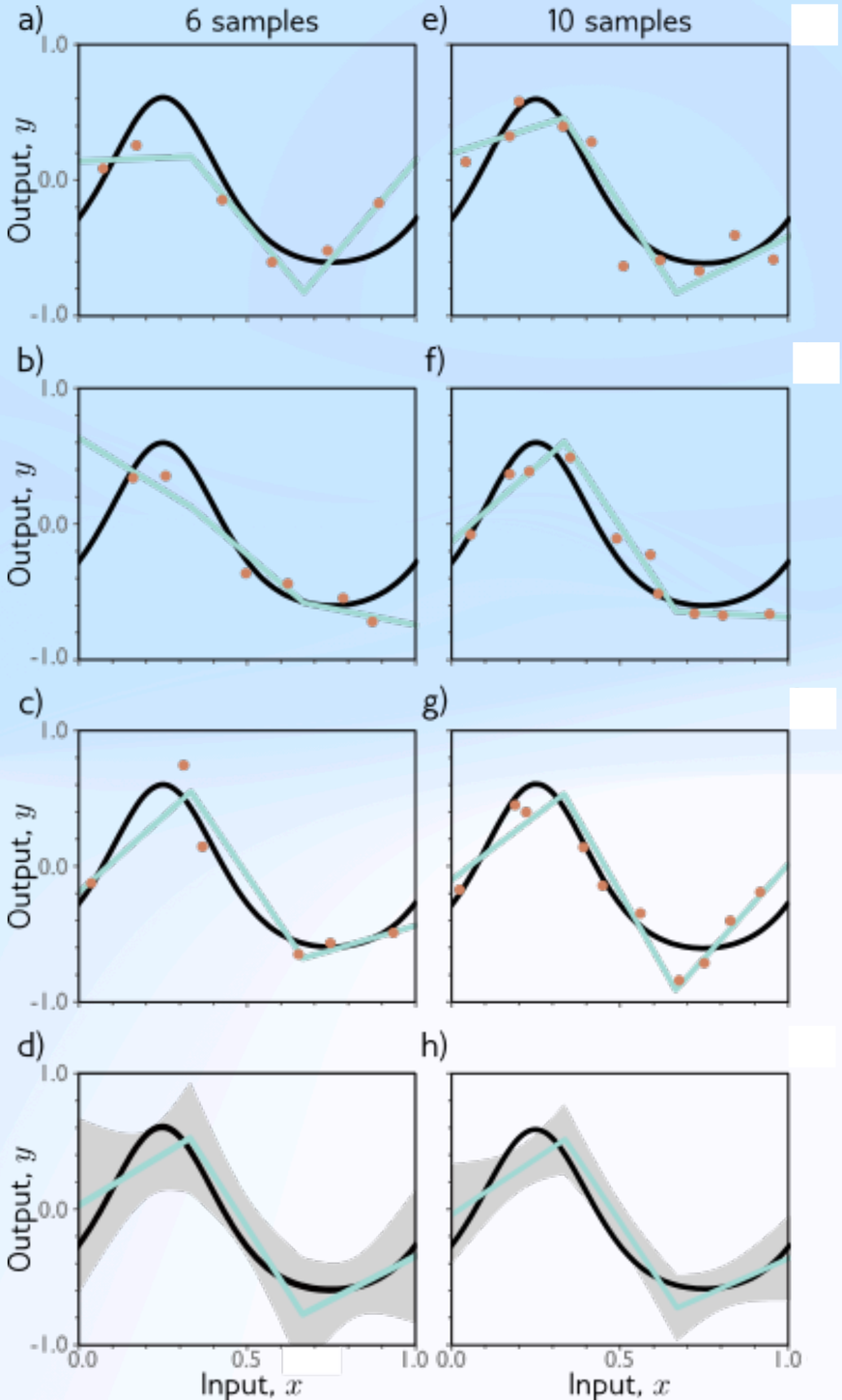
Antonin Vacheret

Variance



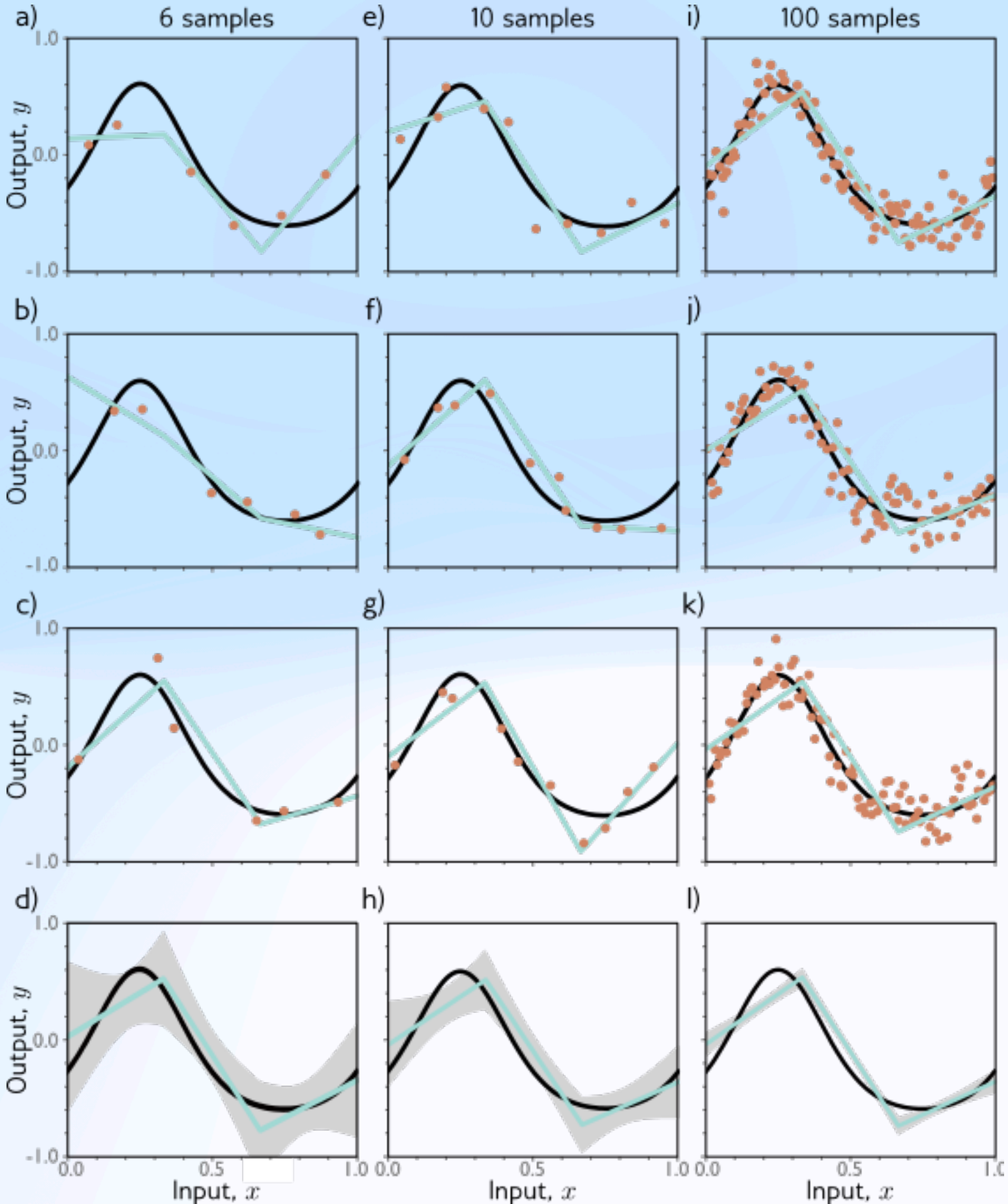
Antonin Vacheret

Variance



Antonin Vacheret

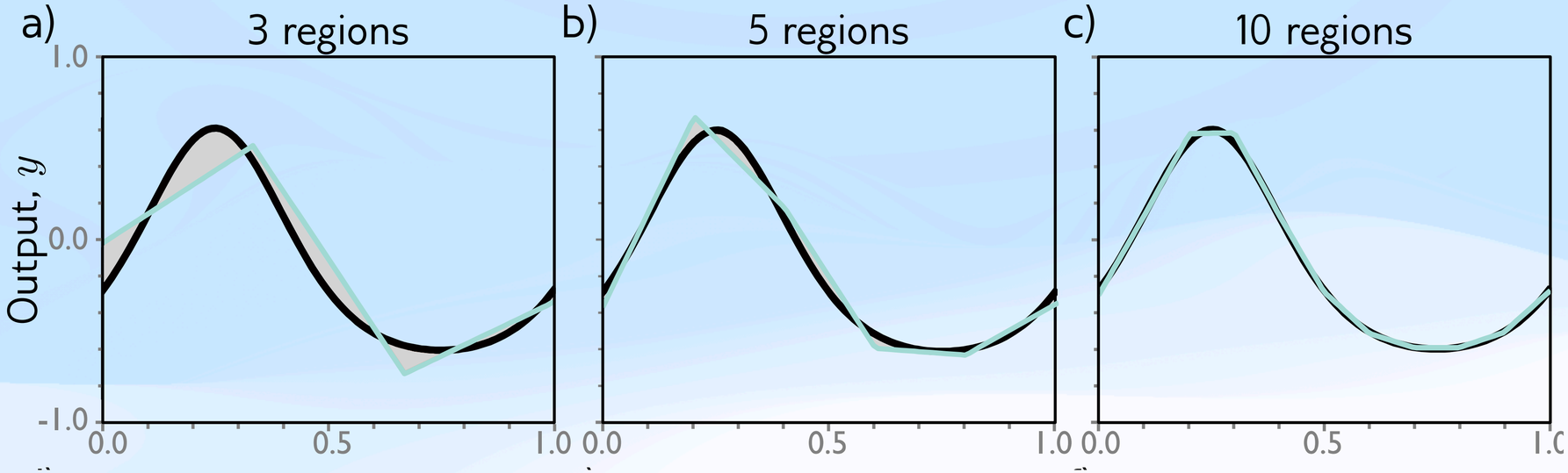
Variance



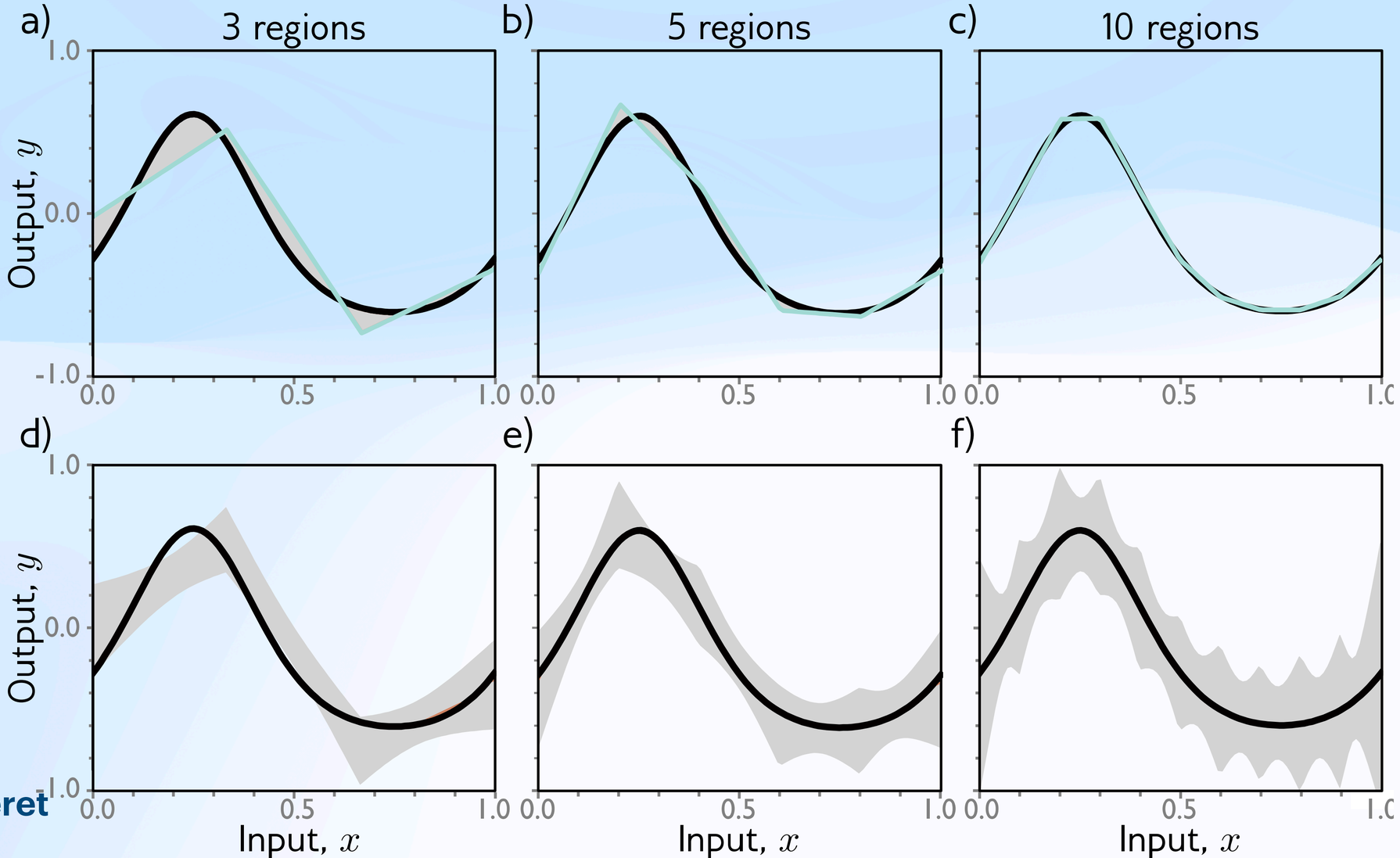
Can reduce variance by adding more samples

Antonin Vacheret

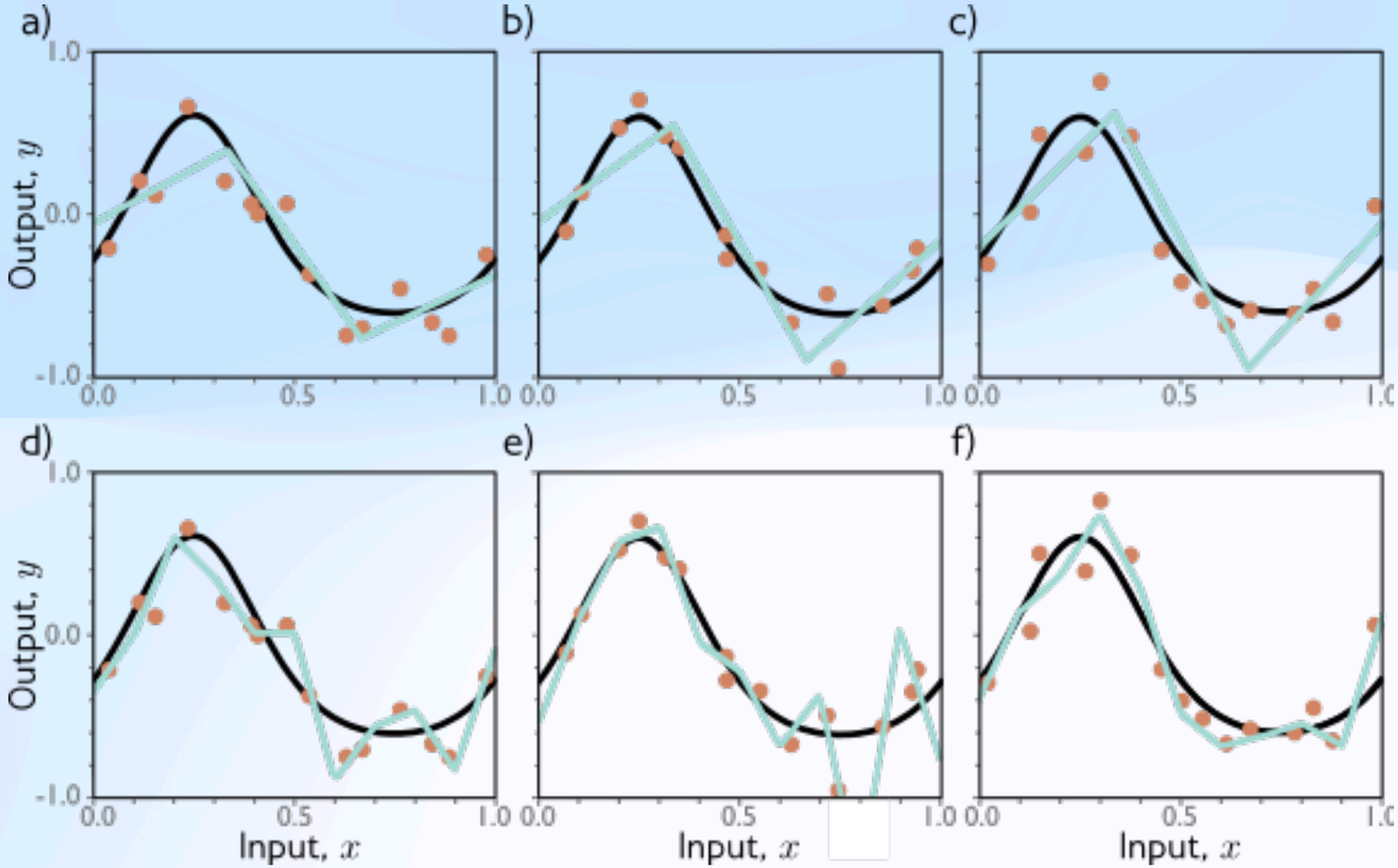
Reducing bias



Reducing bias



Why does variance increase? Overfitting



Antonin Vacheret

Describes the training data better, but not the true underlying function (black curve)

Understanding the training process

- During training the model uses more and more capacity to represent the data being presented
- More diverse the features, more capacity to encode the different cases and their details are needed
    - Capacity usage increases with training
        - We need to stop the training (or at least save the weights) at the minimum of the loss to get the optimal set of weights and the best performance
- Beyond, the model memorizes the data
- More data -> less variance, more depth -> better resolution i.e less bias
    - This implies we need to scale both in tandem to get the minimum loss value (or error)

Scaling laws

GPTs ArXiv:2001.08361v1

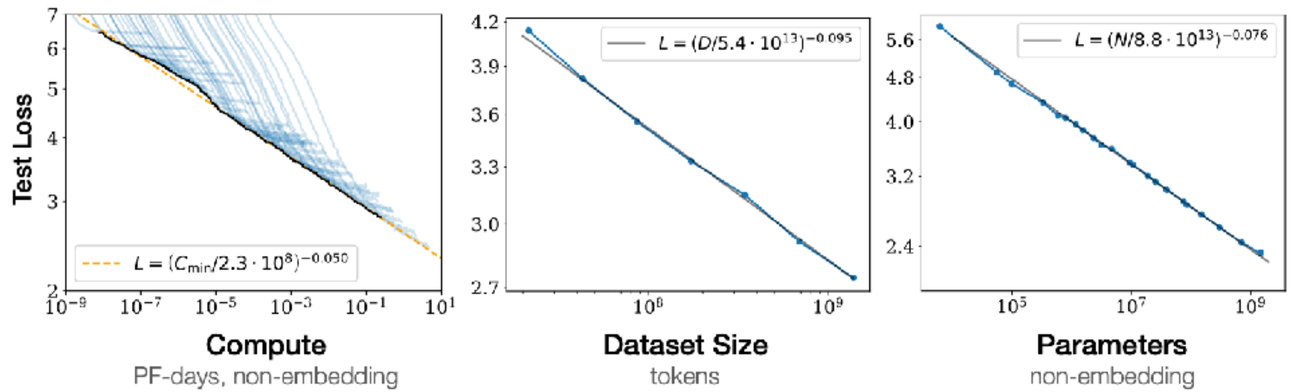
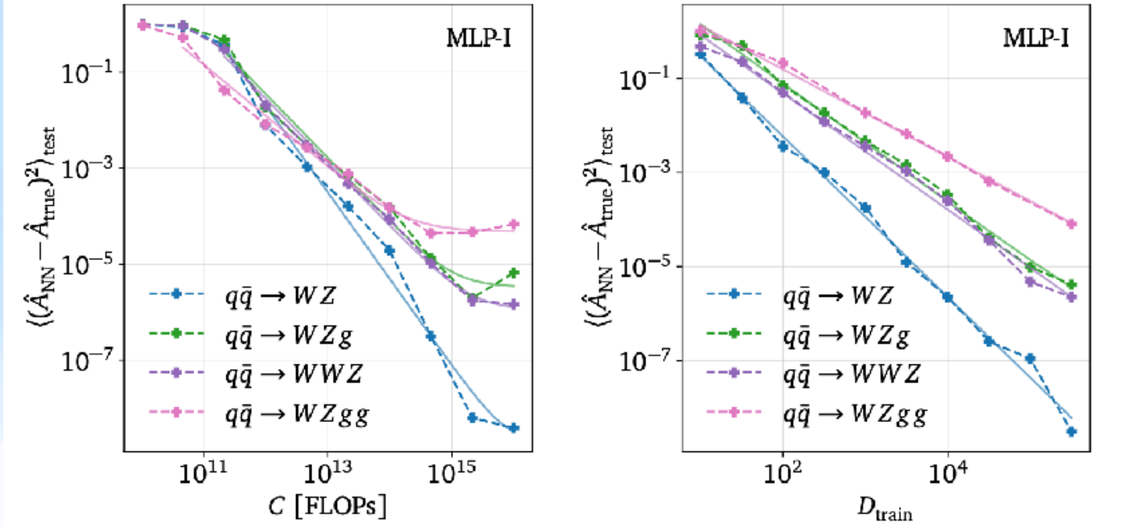


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

PP Amplitude surrogates ArXiv:2601.13308v1



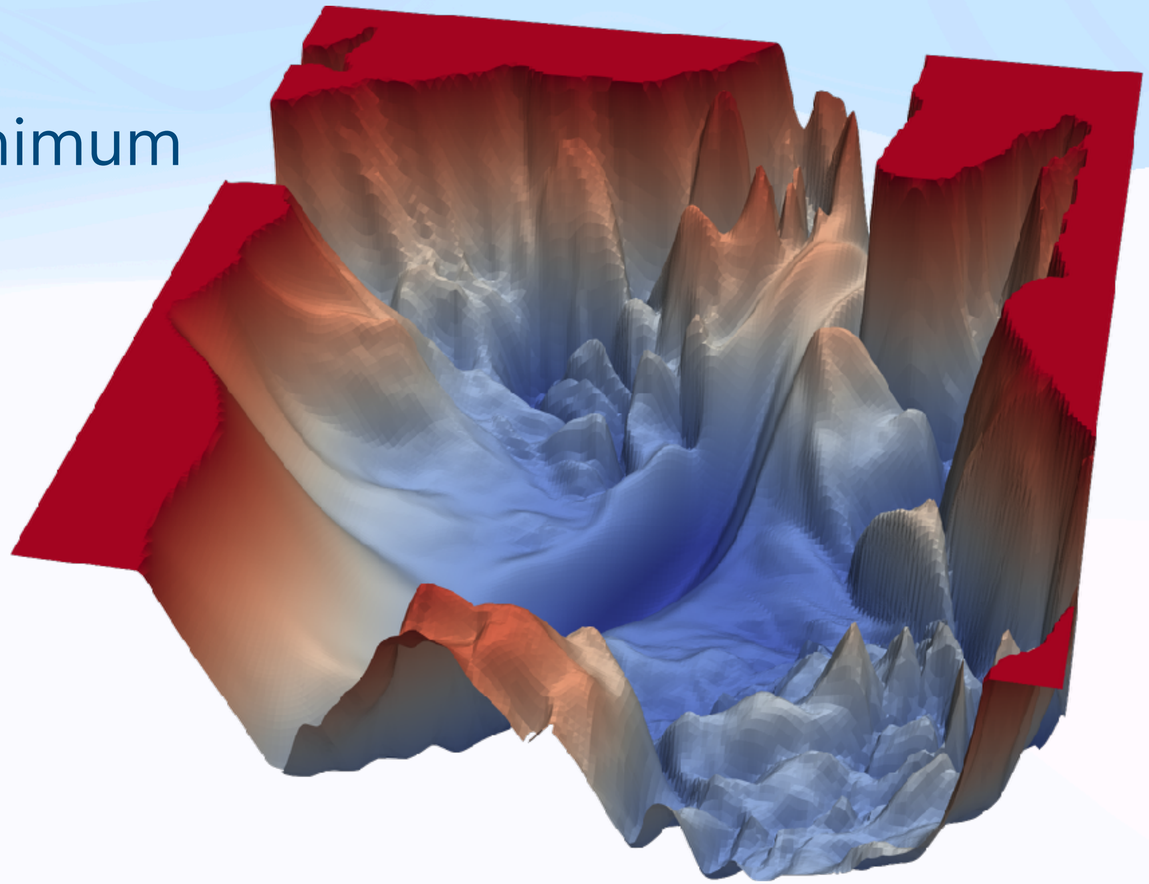
- Scaling laws have been studied for various architectures and extensively for Large language models :
  - Gives a relationship between Compute (FLOPS), Datasets size D and model parameters N.
- Precise scaling laws provides a way to do a systematic search for the best model and data sizes given a compute budget
- If you have complex datasets and able to make them large, a larger model will be needed to encode all of the data

Antonin Vacheret

Regularization

The need for regularization

- As model parameters increase, loss landscapes could become very difficult to navigate for the optimiser
  - Training will get stuck in a local minimum
  - Poor generalization will follow



The need for regularization

- **Regularization** : any method that reduces the generalization gap (and facilitates training)
- One explicit method is to modify the loss adding a penalizing term :

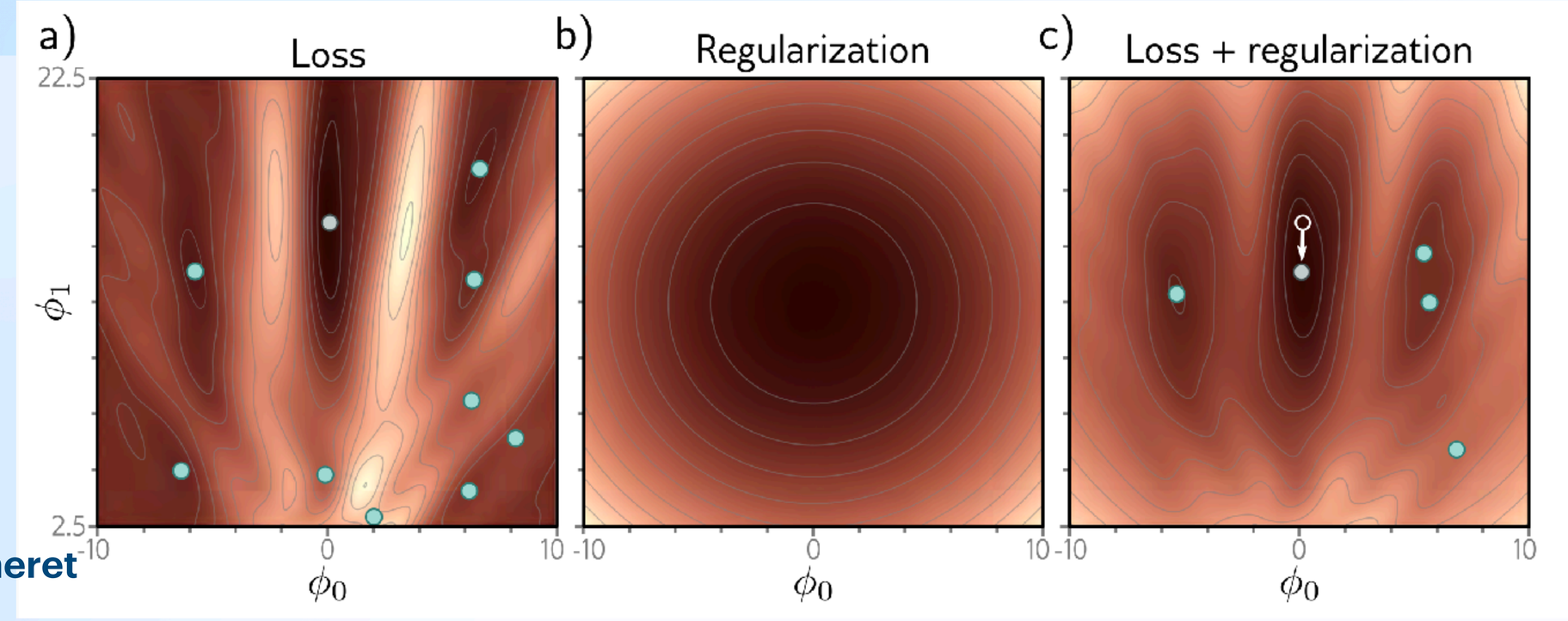
$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left[\sum_{i=1}^{I} \ell_i[\mathbf{x}_i, \mathbf{y}_i] + \lambda \cdot \mathrm{g}[\boldsymbol{\phi}] \right]$$

- For example using the *L2 norm* function :

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left[\sum_{i=1}^{I} \ell_i[\mathbf{x}_i, \mathbf{y}_i] + \lambda \sum_{j} \phi_j^2 \right]$$

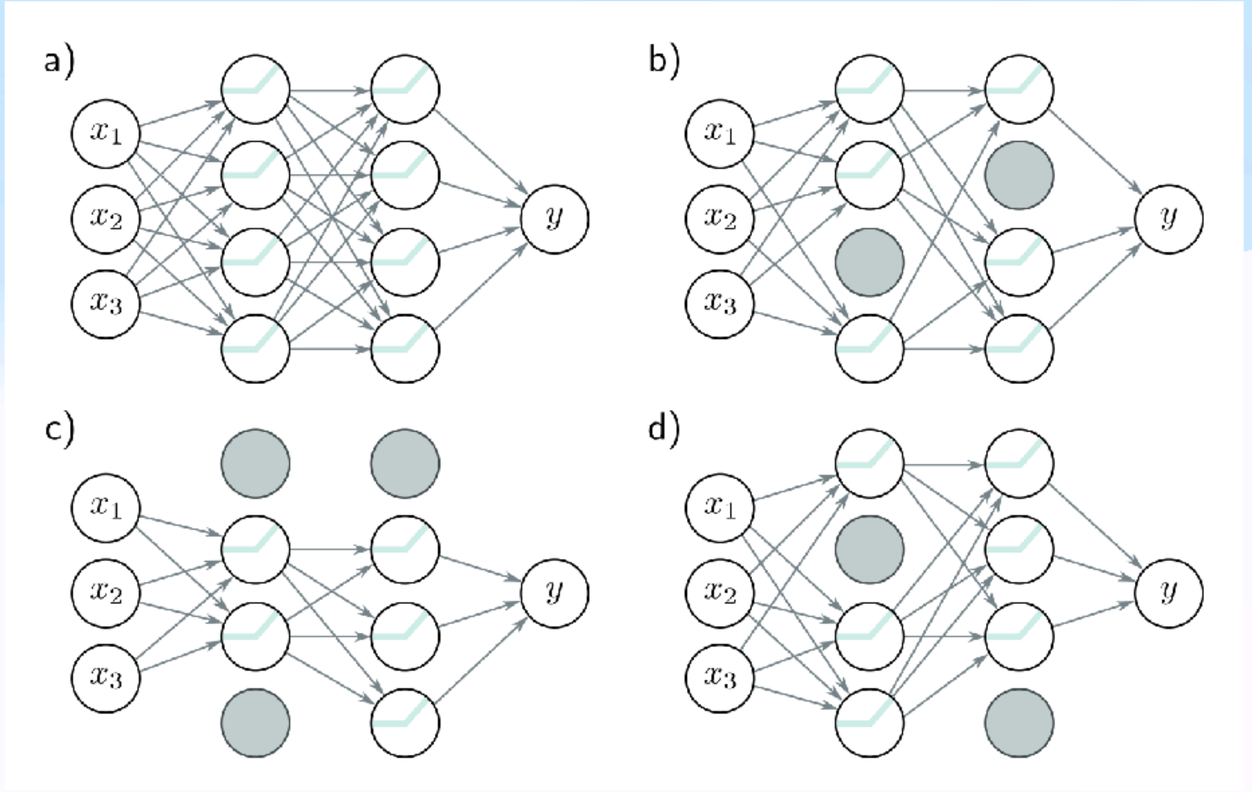
The need for regularization

- **Regularization** : any method that reduces the generalization gap (and facilitates training)
- Effect of the L2 norm on the loss :



Other regularization : Drop out

- Take a random fraction of weights and set them to zero during training to force the model to use all available units :
  - Addresses issues where model dependent on small number of hidden units and overfit early
  - Smooth out the training and provide better performance to deeper models
  - dropout() function readily available in pytorch



Quantifying Uncertainties

Motivation for Quantifying Uncertainties

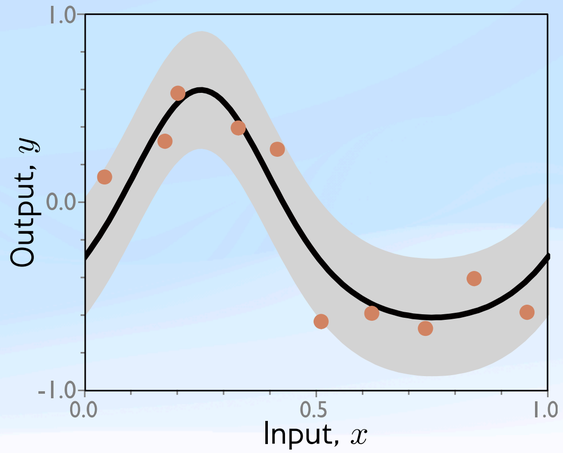
How confident is my model output ?

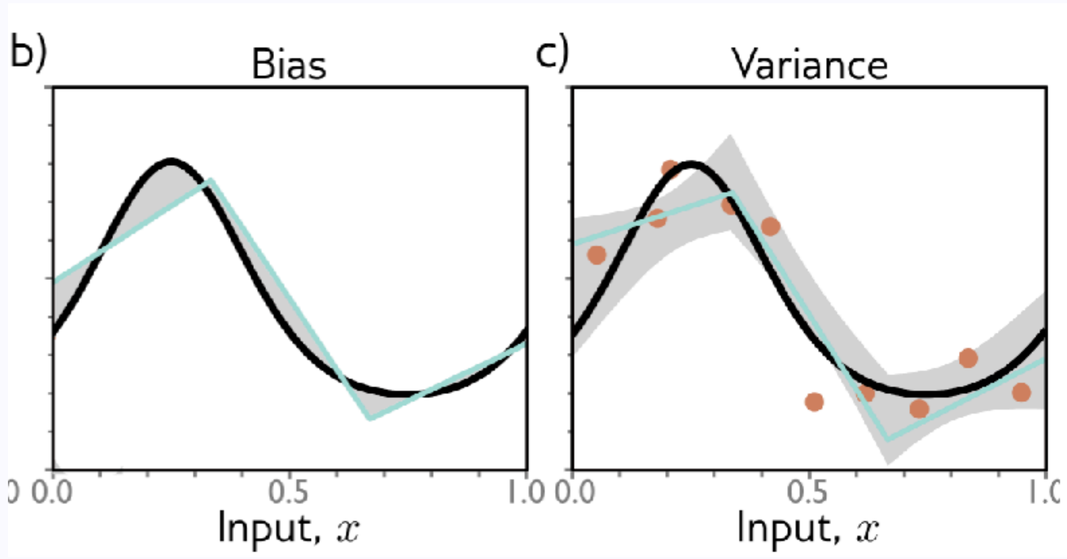
- Predictive models and output are usually **deterministic**
- Real world applications require **confidence estimates**
- **Reliability** of model prediction is as important as prediction themselves

- Different from model error:
  - **Error** : deviation between prediction and target (observed after the fact)
  - **Uncertainty** : Distribution over possible outcomes before knowing the ground truth

Type of uncertainties in ML

- **Aleatoric** : inherent randomness
  - Present in the training data
  - Irreducible with model optimisation
    - Sensor noise, coin flips...
- **Epistemic** : lack of knowledge
  - Reducible with more data, better model
- Out of distribution prediction





Antonin Vacheret

Uncertainties quantification methods

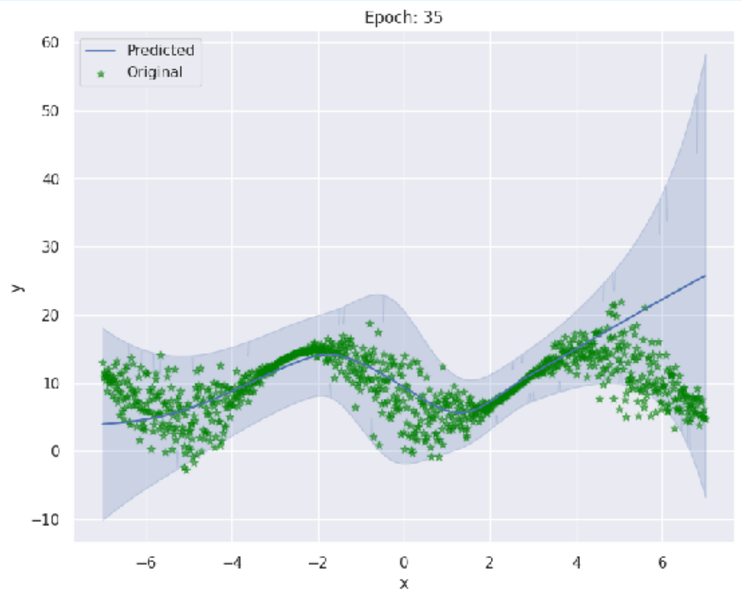
- Maximum likelihood estimation (MLE)
- Mixture Density networks (MDN)
- Quantile Regression
- DualAQD

- Epistemic
  - MC dropout : dropout as a Bayesian approximation
  - Deep Ensemble : train a set of models to obtain a distribution of the output

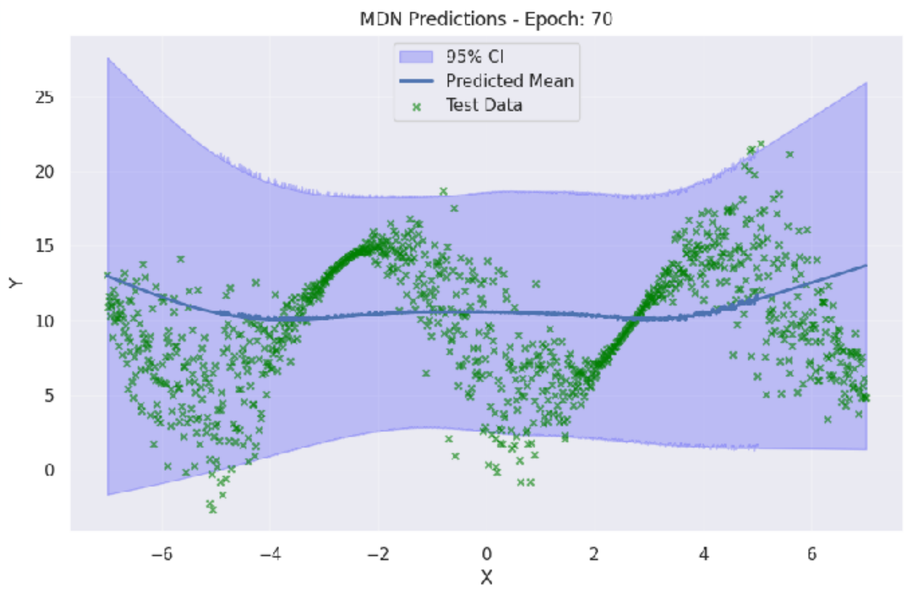
Uncertainties quantification methods

Example of regression

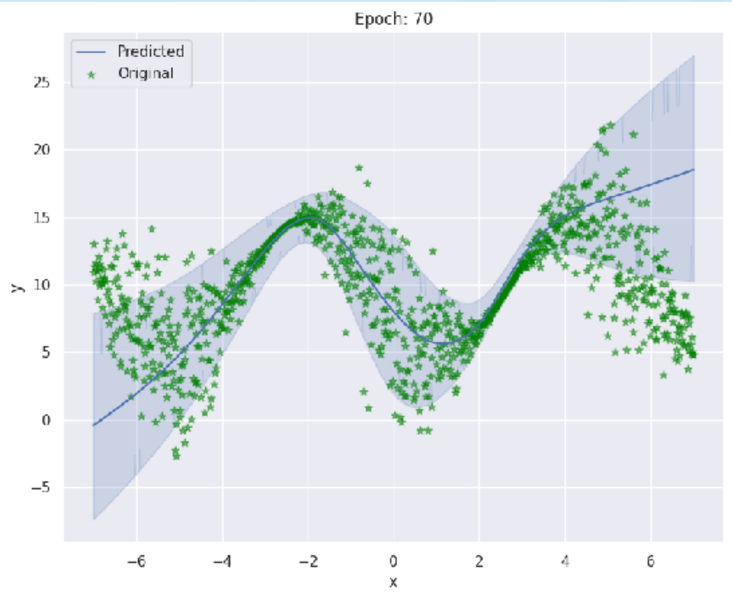
Maximum Likelihood Estimation



Mixture Density Network 1 Gaussian



Quantile Regression



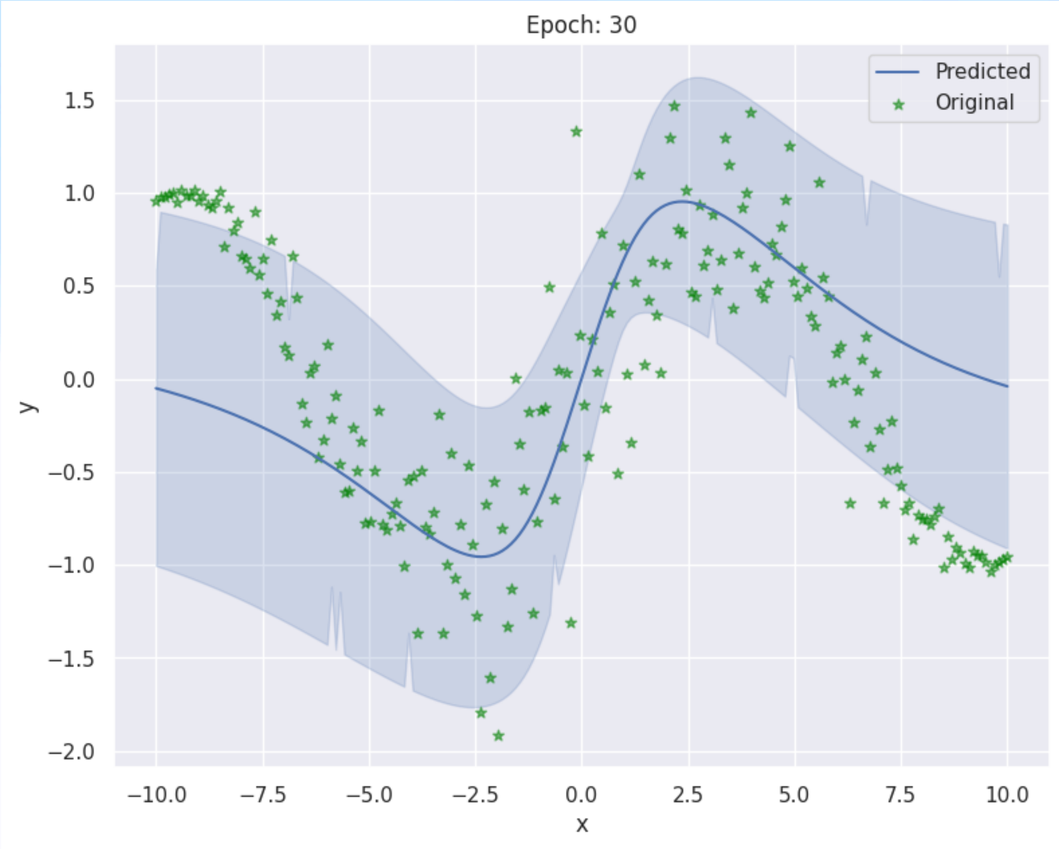
- Simple methods to understand total uncertainty from output within the training range and (to some level) cover also out of distribution cases.
  - Combination of aleatoric and epistemic uncertainties

Uncertainties quantification methods

Model uncertainty

MC drop out

Deep Ensemble

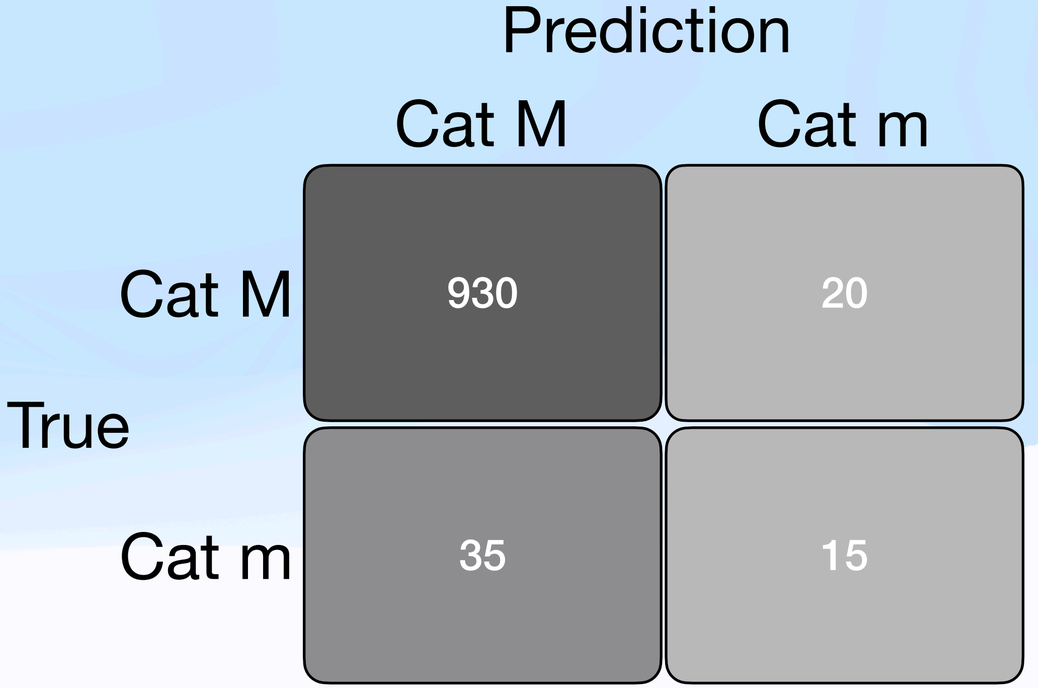


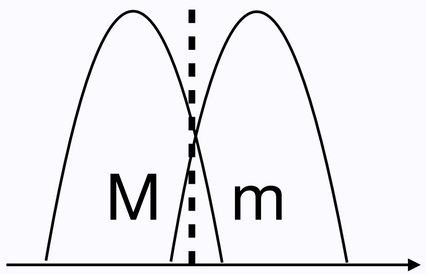
Biases in classification

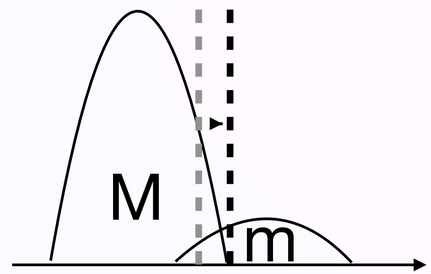
Antonin Vacheret

Data imbalance

- In a classification task machine learning model fits properly the data if equal examples of all categories are used during training.
- In practice, it is more often the case that we do have data imbalance
- What is the effect of data imbalance ?
  - Exemple of 1000 samples [950,50]



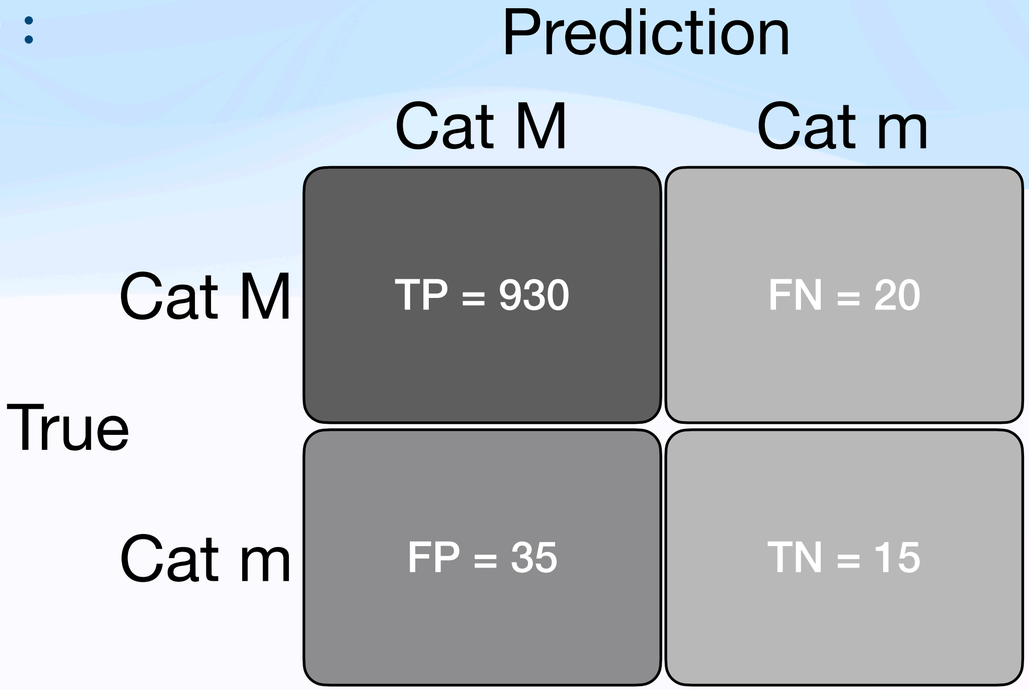




Antonin Vacheret

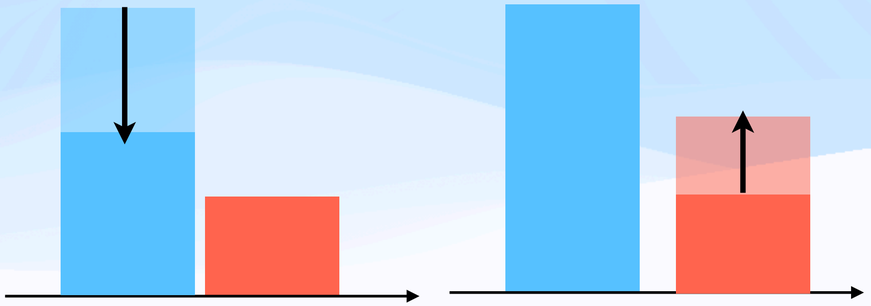
Data imbalance : other metrics

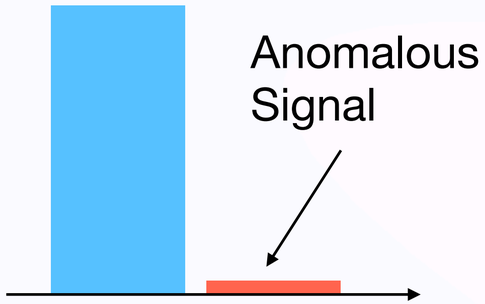
- Use other metrics to monitor imbalance :
  - **Precision** $= \frac{TP}{TP + FP}$
  - **Recall** $= \frac{TP}{TP + FN}$
  - **F-measure** $= 2 \times \frac{Prec \times Recall}{Prec + Recall}$



Data imbalance : Solutions

- Resampling techniques
  - Downsample majority class (randomly,...) or upsample minority one (SMOTE, data augmentation)
  - Could combine both for better effectiveness
- Weighted classes and use of specific loss fonction
  - Adjust the loss function by reducing importance to majority class based on fraction of each class in total dataset
  - Focal loss penalises FN more heavily, improving Recall
- Combine alternatives
- Reframing the problem : anomaly detection ?





Questions ?