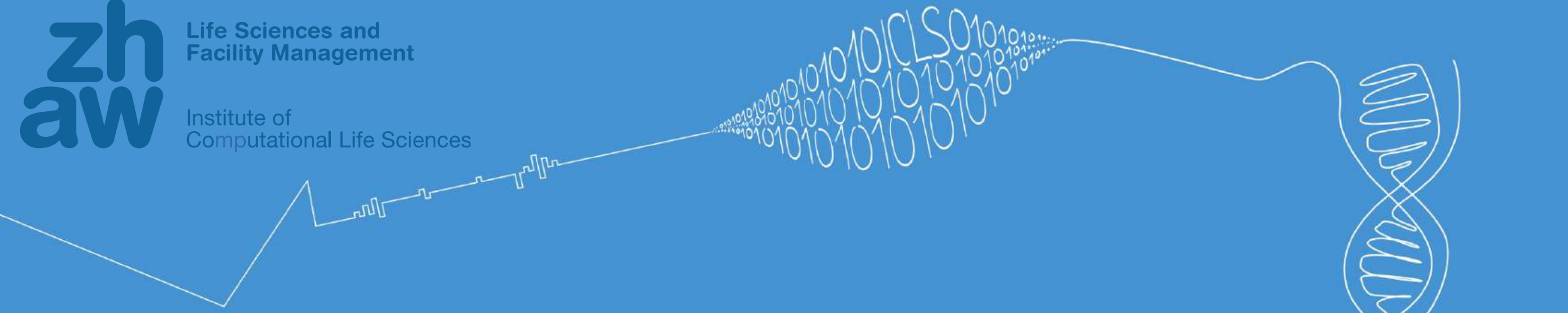
Life Sciences and
Facility Management

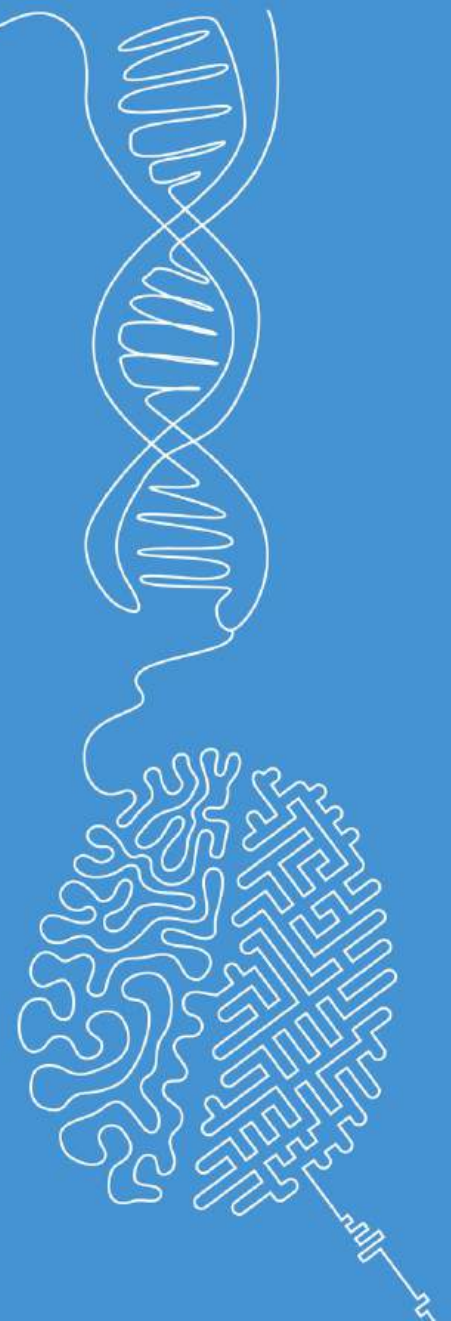
Institute of
Computational Life Sciences



Fundamental Aspects of ML Problems

Machine Learning

Norman Juchler



Learning objectives

- Why is there no free lunch for ML models?
- How can we measure model complexity?
- Why is inductive bias key to effective ML solutions
- The counterintuitive properties of high-dimensional spaces

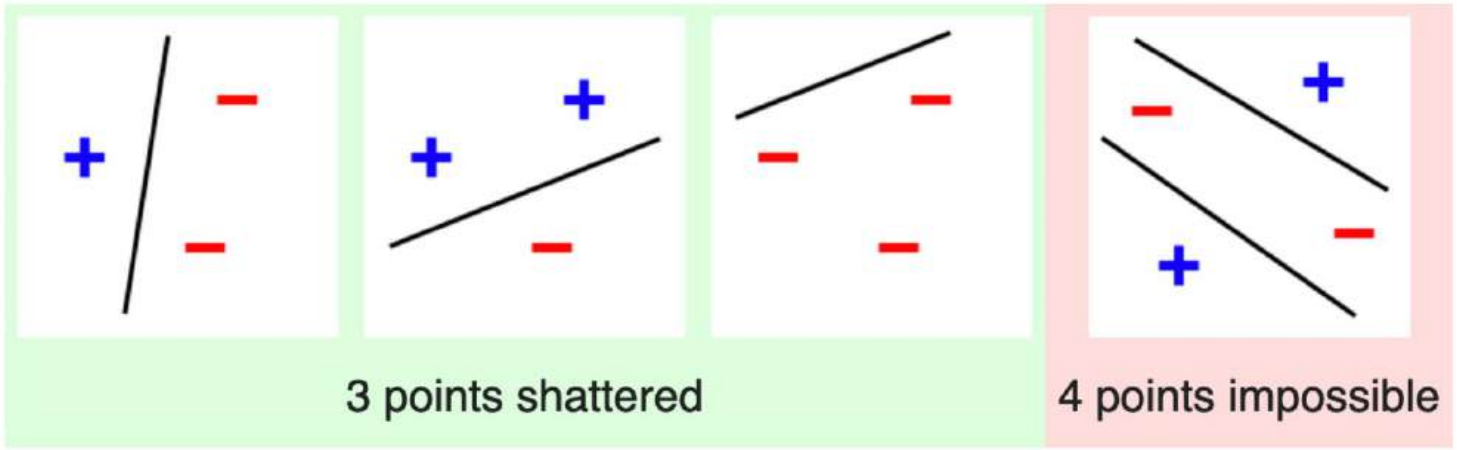
The No Free Lunch Theorem

> No Free Lunch Theorem:
> "All optimization algorithms perform equally well when their performance is averaged across all possible problems."

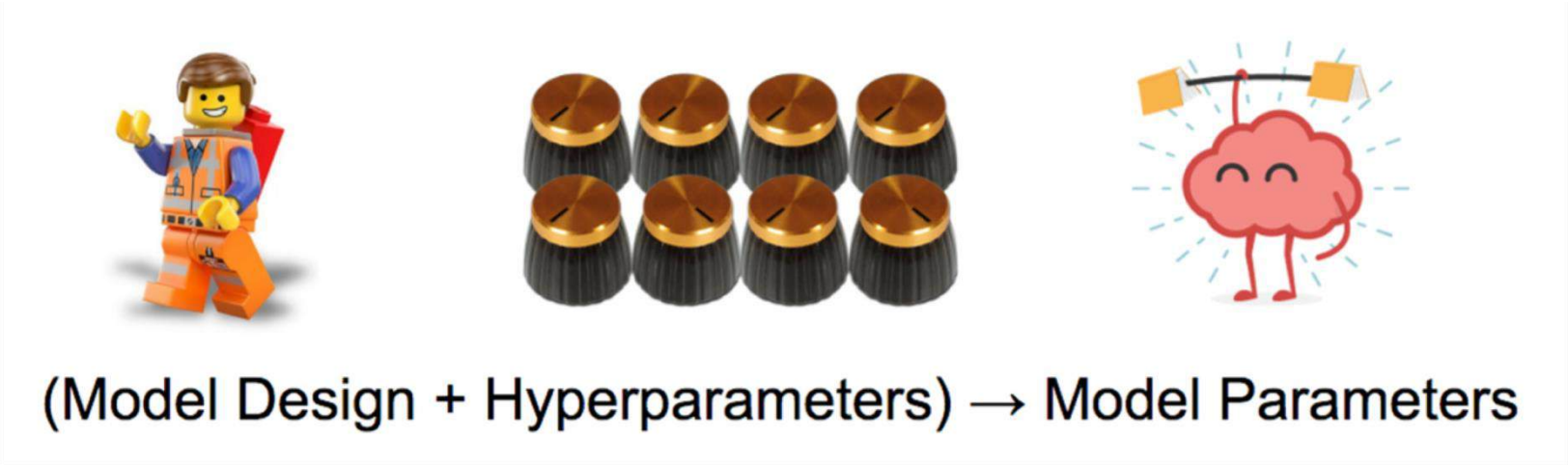
- There is no generally best algorithm.
- There is no generally valid answer to the question “What is the best algorithm for this problem?”.
- The answer depends on the training set that is available.

Model Complexity

- Model complexity refers to how fine-grained a model's decision boundaries can be in the feature space.
- **VC dimension** (Vapnik–Chervonenkis dimensions):
  - A measure for model complexity
  - VC dimension: Number of samples for which a model can learn a perfect classifier (for any labeling of the input samples).
  - Example for straight line classifier (e.g., a Perceptron): VC Dimension = 3



The three levels of degrees of freedom in modeling



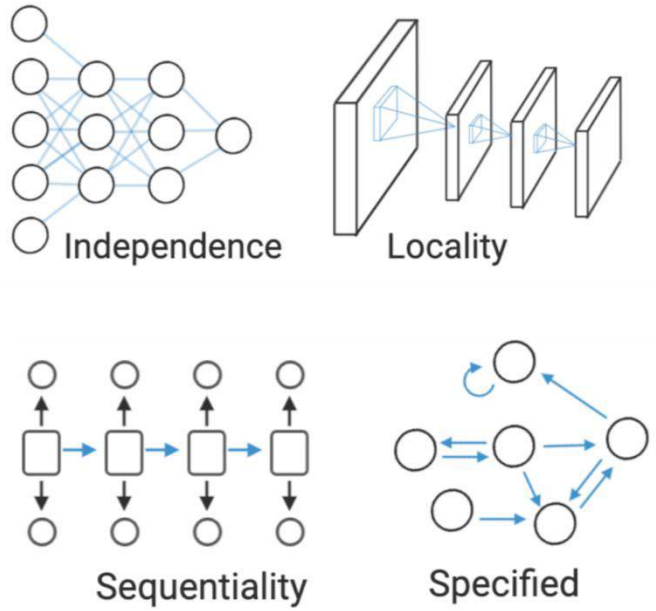
(Model Design + Hyperparameters) → Model Parameters

Examples

Function class (e.g. cosine or a polynomial)	Degree of the polynomial	The polynomial's parameters (a_0, a_1, etc.)

Inductive bias

- **Problem**: It is often difficult to obtain a training dataset that prepares a model for all future inputs.
- **Solution**: Use inductive bias to guide the model training.
  - Inductive bias: Assumptions a learning algorithm makes to learn the target function more effectively and to generalize beyond the training data.
  - Helps controlling the model complexity and can be used to embed prior expert knowledge into the model
- **Examples**:
  - Type of ML algorithm (e.g., linear regression model)
  - ML model architecture (e.g., structure of a neural network)
  - Types of features used (e.g., features to symmetries in the data)

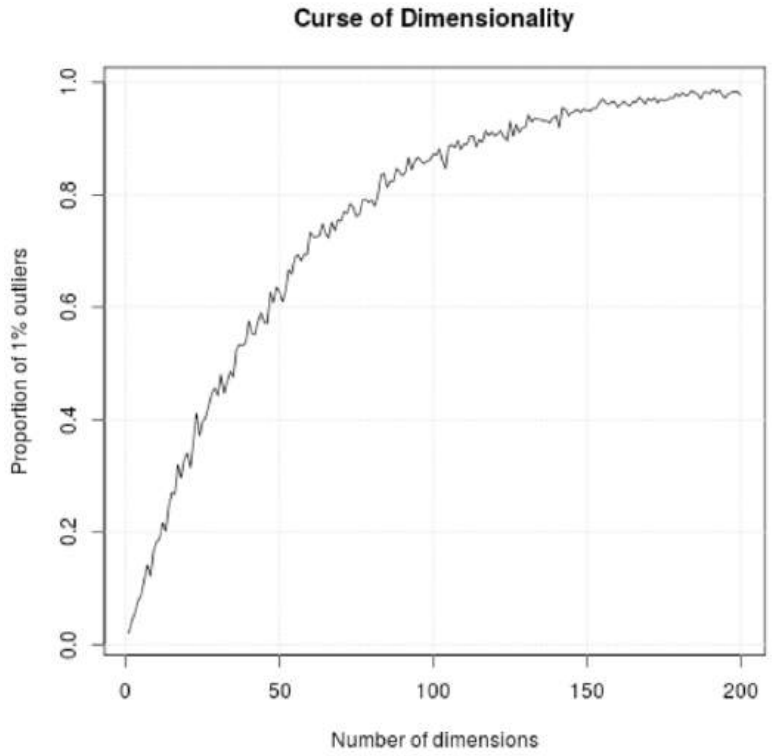


Examples of inductive biases engraved in neural networks or graph models.

Choice of good inductive biases is key to effective ML solutions!

The curse of dimensionality

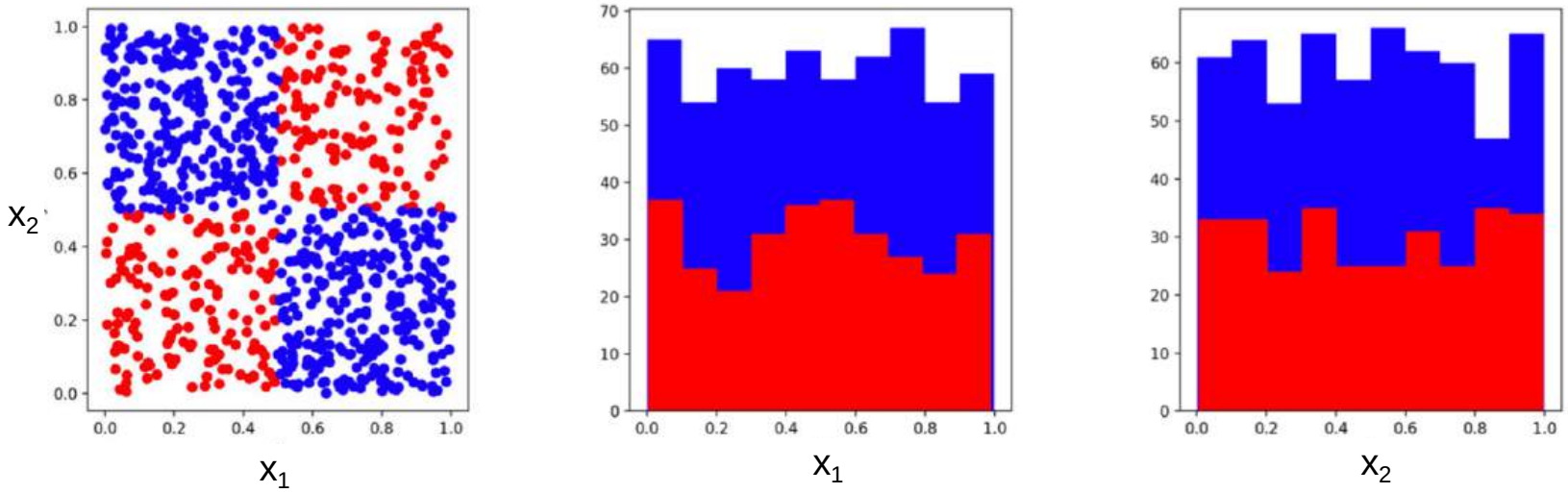
- High-dimensional data can cause problems
  - **Data sparsity**: data points are more spread out in high-dimensional spaces → it's more difficult to find patterns.
  - **Increased computational cost**: The amount of data needed to adequately represent the feature space increases exponentially with the dimensionality.
  - **Distance measures become less informative**: It becomes hard to distinguish close and distant points (e.g., relevant for k-NN clustering).
  - **Overfitting risk**: Models may learn noise rather than true patterns, Random patterns may appear significant in high dimensions → poor generalization, overfitting
- Solutions:
  - Dimensionality reduction
  - Feature selection
  - More data...



Proportion of randomly sampled points that sit close to the boundary of a hypercube, as a function of the dimensionality (=number of features). See corresponding Jupyter notebook!

Importance of features correlations

- High dimensional spaces can have many counterintuitive features
- In most ML problems most of the information is contained in the correlations!
- This includes spatial patterns, as demonstrated below:

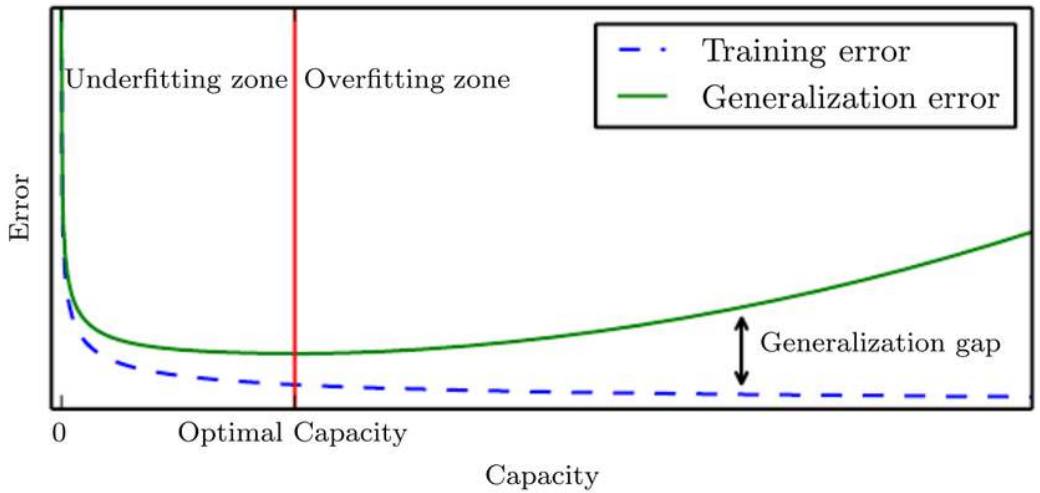


Example of two feature distributions x_1 and x_2 that are separable in the 2D plane (left plot): We can easily find a function that classifies red and blue dots. However, the problem is not separable if we look at the individual distributions of the features (middle and right plot).-

Generalization

- Generalization is the aim of ML!
- We want our model to perform well on as many inputs as possible, also on unseen future data

$$\text{generalization gap} = \varepsilon_{\text{production}} - \varepsilon_{\text{test}}$$



- Possible reasons for a generalization gap:
  - Insufficient training data
  - Production and test sets have different distributions
  - Too little inductive bias present in the model