Learning Workflow

Overview

Decide on model purpose

Choose training data
- Proxy quantity?
- Balance quality against availability

Choose model(s)
- Parametric vs non-parametric
- Loss function

Detrend
- Improve data efficiency

Train models

Model Purpose

Necessity

- Only way to cover problem size
- Still open to systematic evaluation
- Often used as prefiltering step

- Complicated chemistry
- Tricky / error-prone reference calculations

Convenience

- Can be done more accurately
- Uneconomical/cumbersome reference method
- Often used as direct but optional substitute

- Standard energy calculations of well-behaved systems
- Semi-emipirical level sufficient

Data Preparation

Machine learning exploits correlation

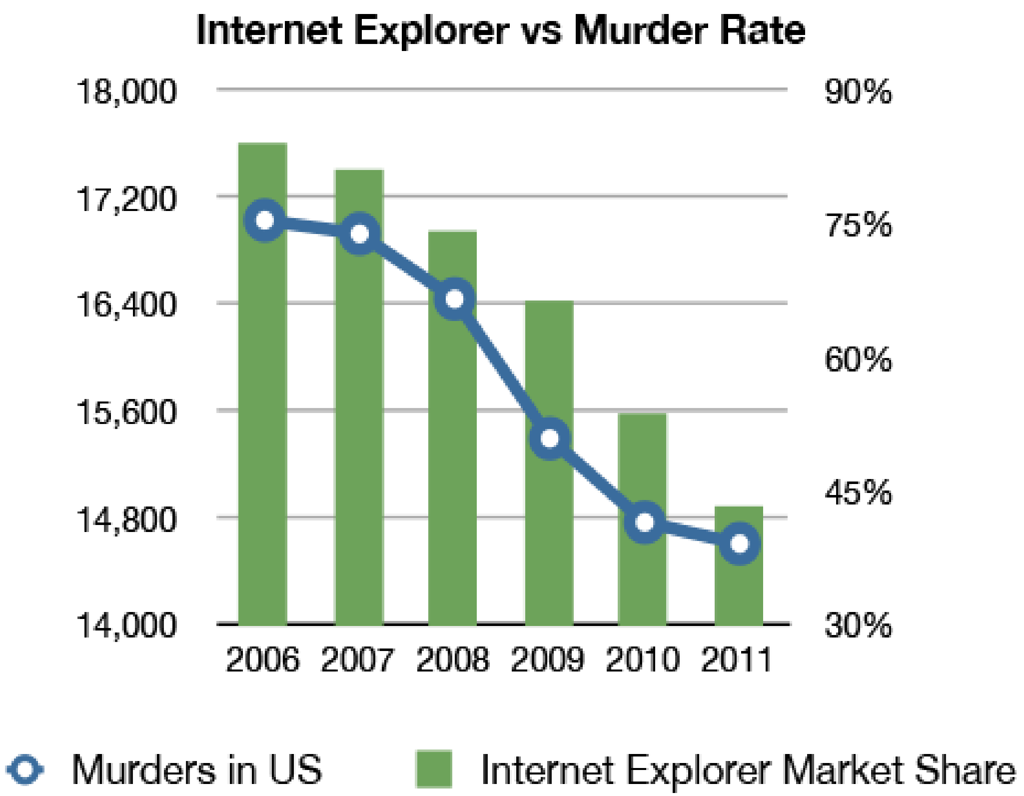
- Correlations necessary but not sufficient
- Spurious cases:
  - Chance (just zoom in / transform random data)
  - Common cause

$$A \to B, A \to C \neq B \to C$$

  - Identities

$$A \propto \epsilon A$$

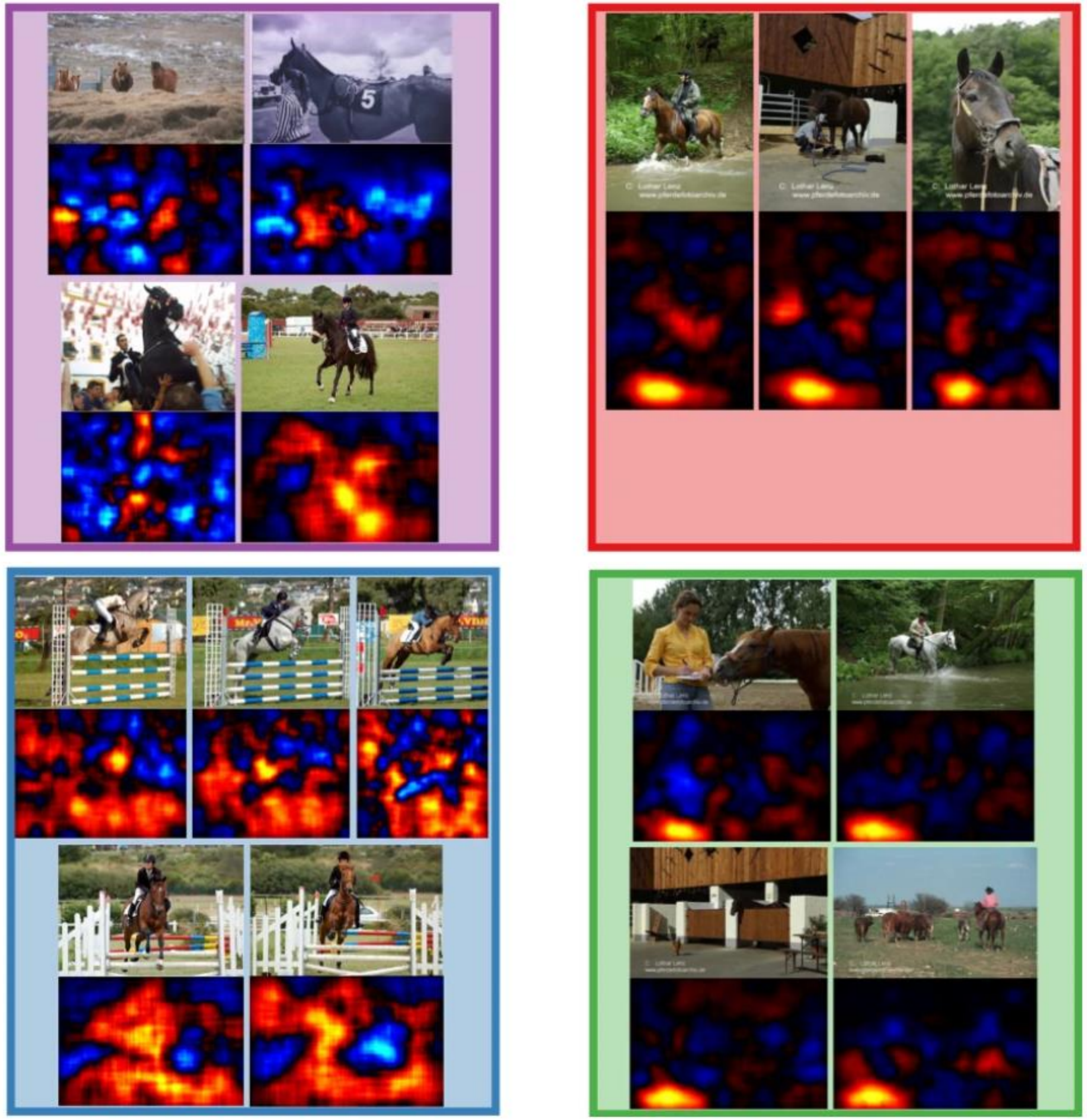
  - Selection bias



Data Preparation

Training data
- Representative
- Accurate
- Comparable (or labelled)

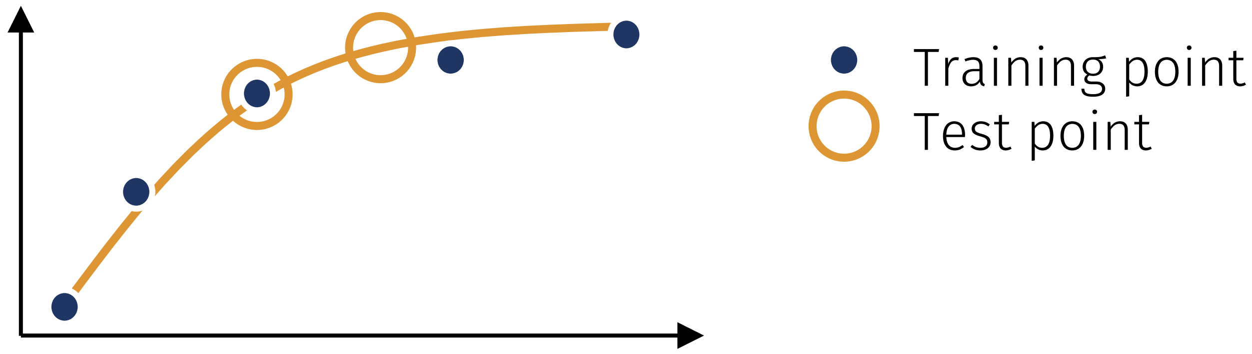
Objective function
- Sensitive to the right features
- Balanced over data set
- Proxy for usefulness



Lapuschkin et al., *Nature Comm,* 2019.

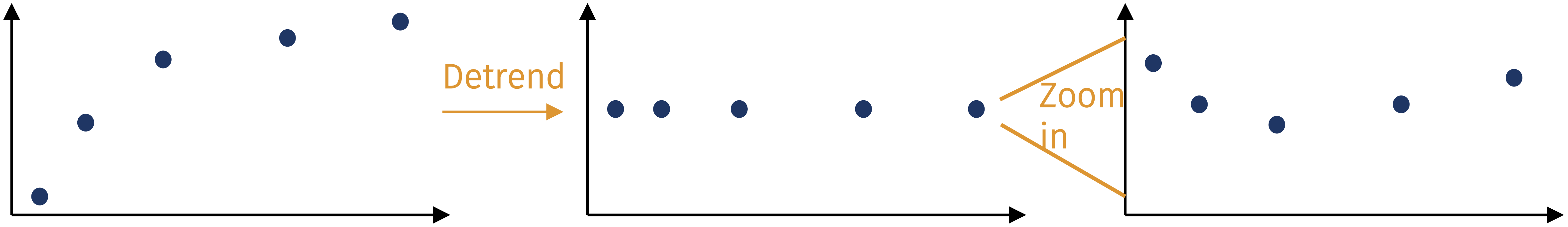
Data Preparation

Remove duplicates
- Can inflate model performance
- Over-emphasizes one point/region

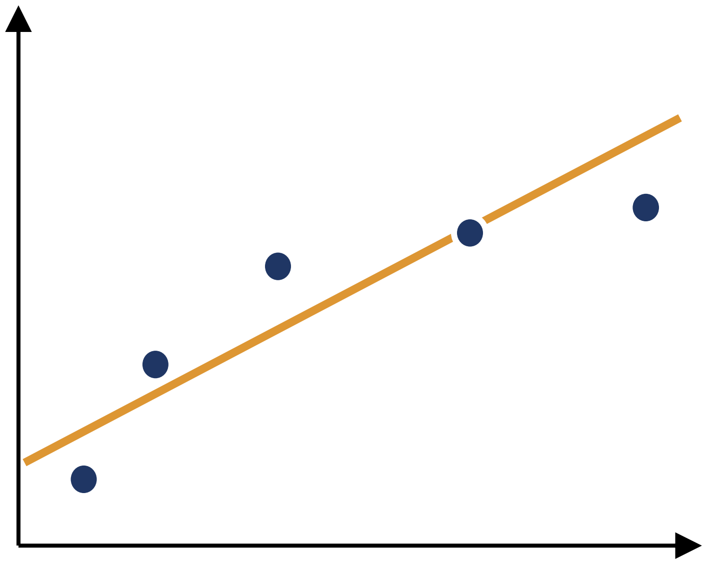


Exclude missing data
- Artificially reduces training or test set

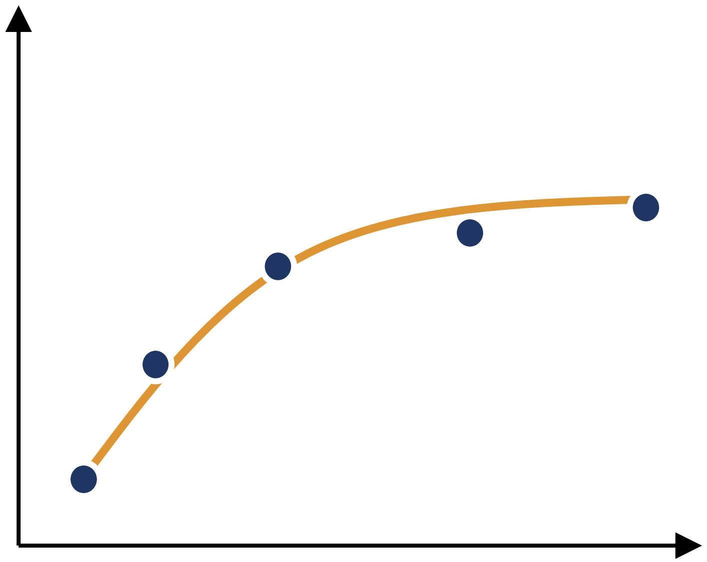
„Zoom in“ by detrending
- Focus on „hard“ part of the problem, e.g. atomization energy rather than total energy



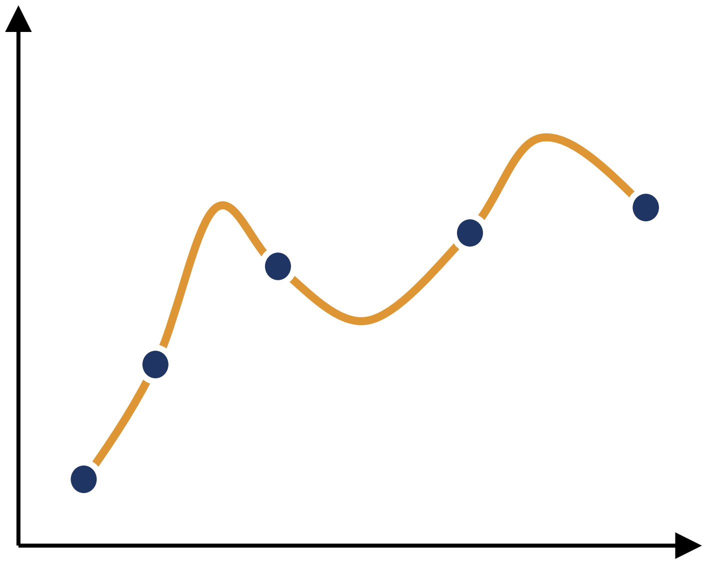
Model Flavours



Underfitting
- Simplistic model
- Over-regularised
- Missing flexibility
- “Classical fitting”

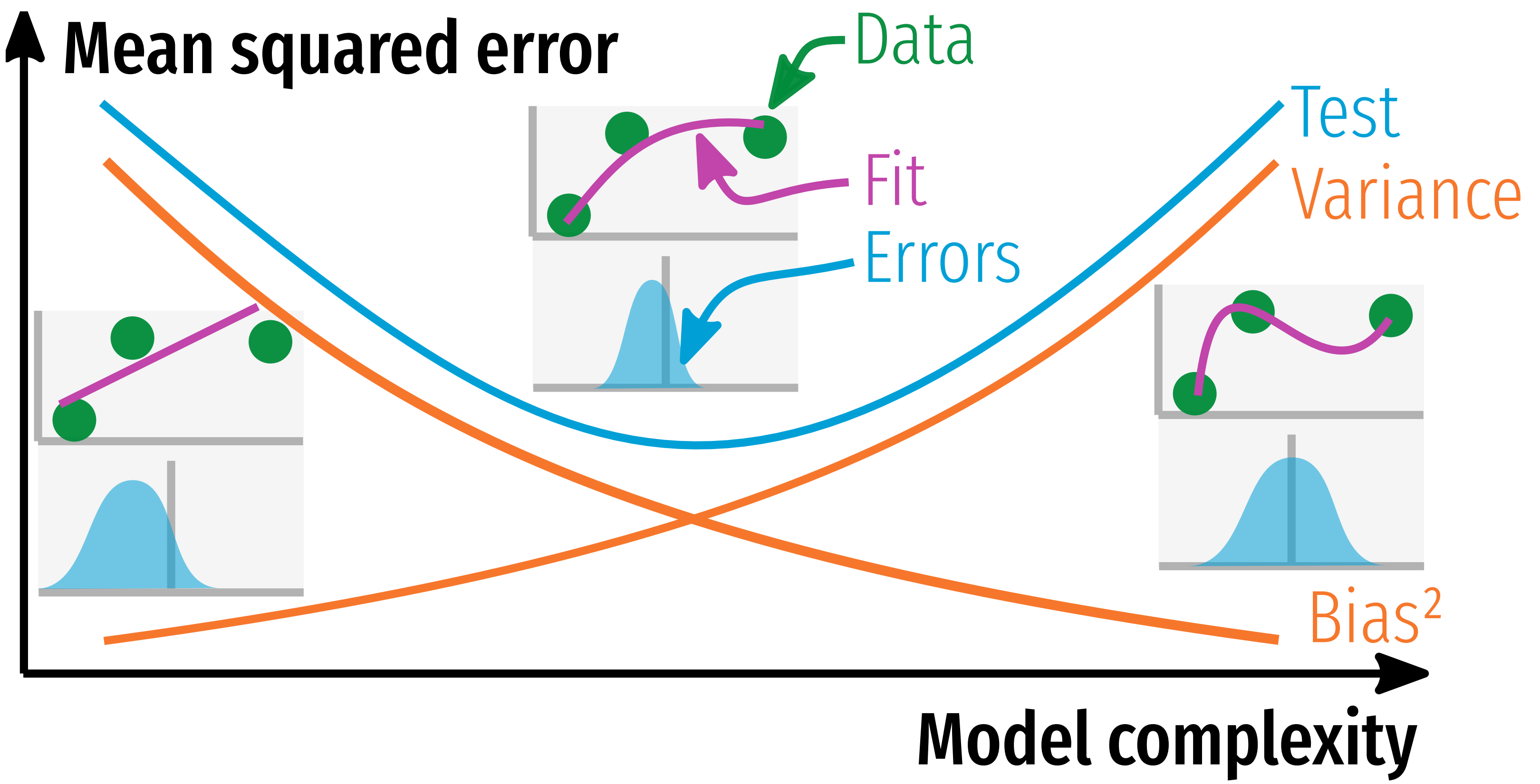


Adequate
- Captures dominating trend
- Forgiving for data points



Overfitting
- Match data points no matter the cost
- Little predictive power

Expressiveness



"Low expressiveness"

"High expressiveness"

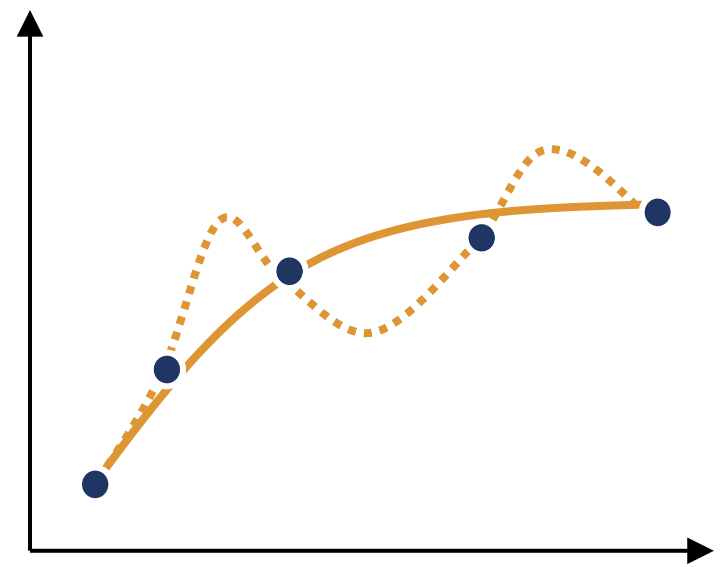
Regularization

Aim

- Procedure to choose „well-behaved“ models
- Avoid overfitting, but allow for flexibility

Common: Norm of coefficients

- Ridge regression or Tikhonov regularization



Training points

Labels

$$\min_{\mathbf{w}} \sum_j |f(x_j|\mathbf{w}) - y_j| \longrightarrow \min_{\mathbf{w}} \sum_j |f(x_j|\mathbf{w}) - y_j| + \lambda \|\mathbf{w}\|_2^2$$

Model parameters (“weights”)

Regularization strength

Summary Learning Workflow

- Decide every design aspect based on final research question
- Representative training data required, possibly use proxy properties
- Training data needs to be cleaned to obtain reliable and comparable models
- Detrending helps data efficiency
- Expressiveness: how many function classes can be represented in theory
- Flexible but regularized models helpful, i.e. constraints on „simple yet expressive models“
- Main difference to classical fitting: fewer parameters is not better