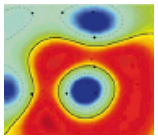
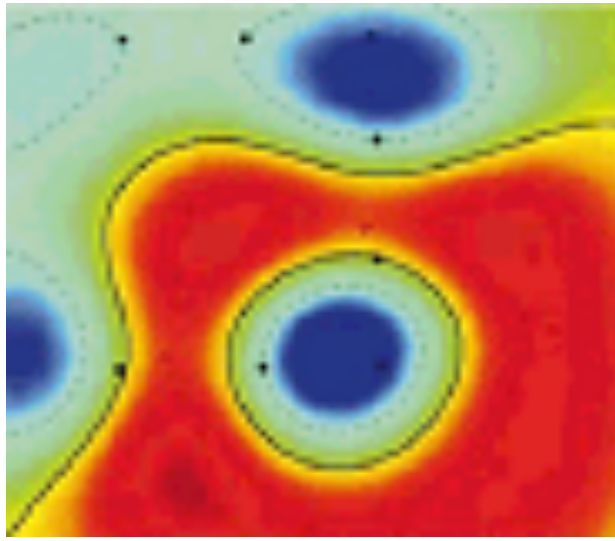


WiSe 2023/24

Machine Learning 1/1-X

Technische Universität Berlin

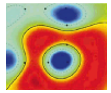


Lecture 7

Model Selection

Outline

- Model Selection
- Occam's Razor
  - Quantifying Model Complexity
- Popper's Prediction Strength
  - Holdout / Cross-Validation
- Limits of Holdout / Cross-Validation, The 'Clever Hans' Effect
- Bias-Variance Analysis

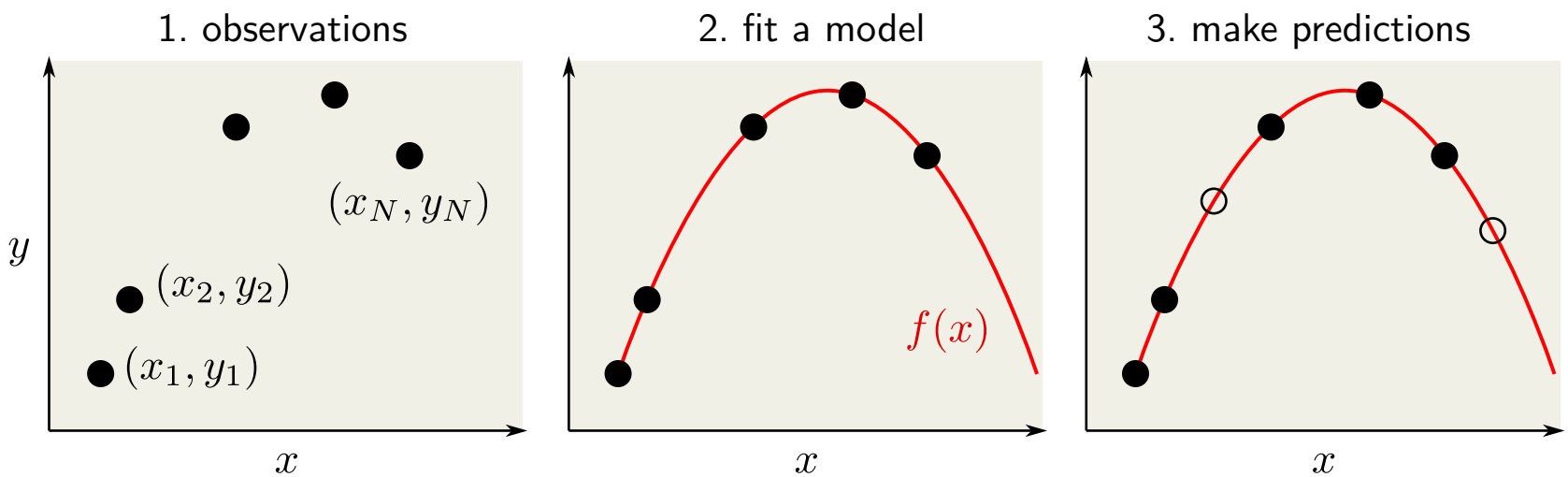


Learning a Model of the Data

Assume we have a few labeled examples

$$(x_1, y_1), \ldots, (x_N, y_N)$$

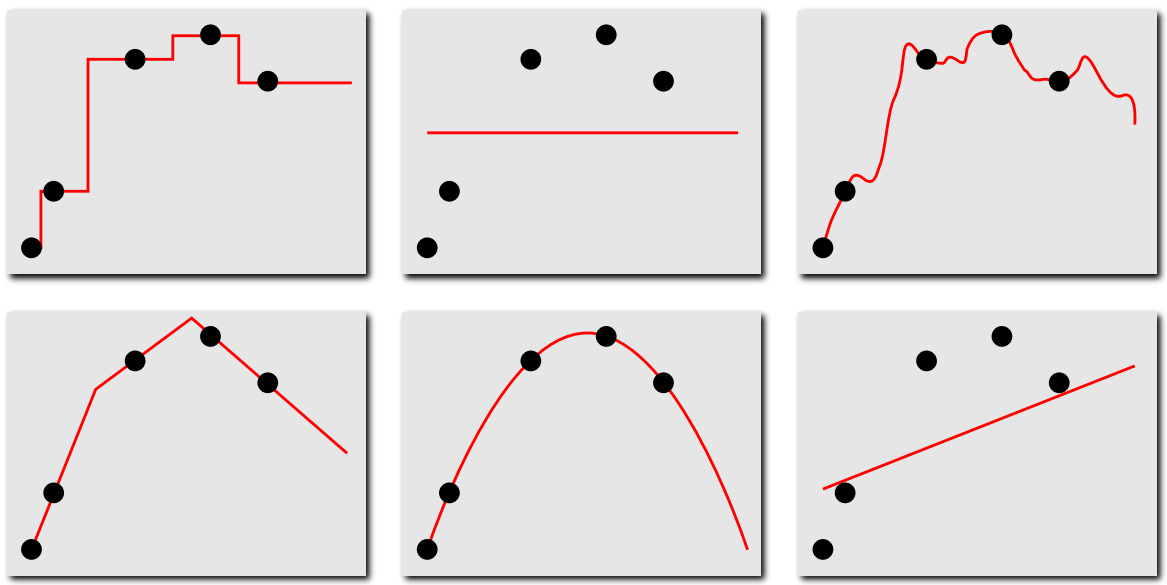
from some *unknown* true data distribution $p(x, y)$. We would like to learn a model $f(x)$ that not only predicts well $x_1, \ldots, x_N$ but also future data points drawn form $p(x)$.



Model Selection

Questions:

1. Among models that correctly predict the data, which one should be retained?
2. Should we always choose a model that perfectly fits the data?



Occam's Razor

William of Ockham (1287–1347)

> *"Entia non sunt multiplicanda praeter necessitatem"*

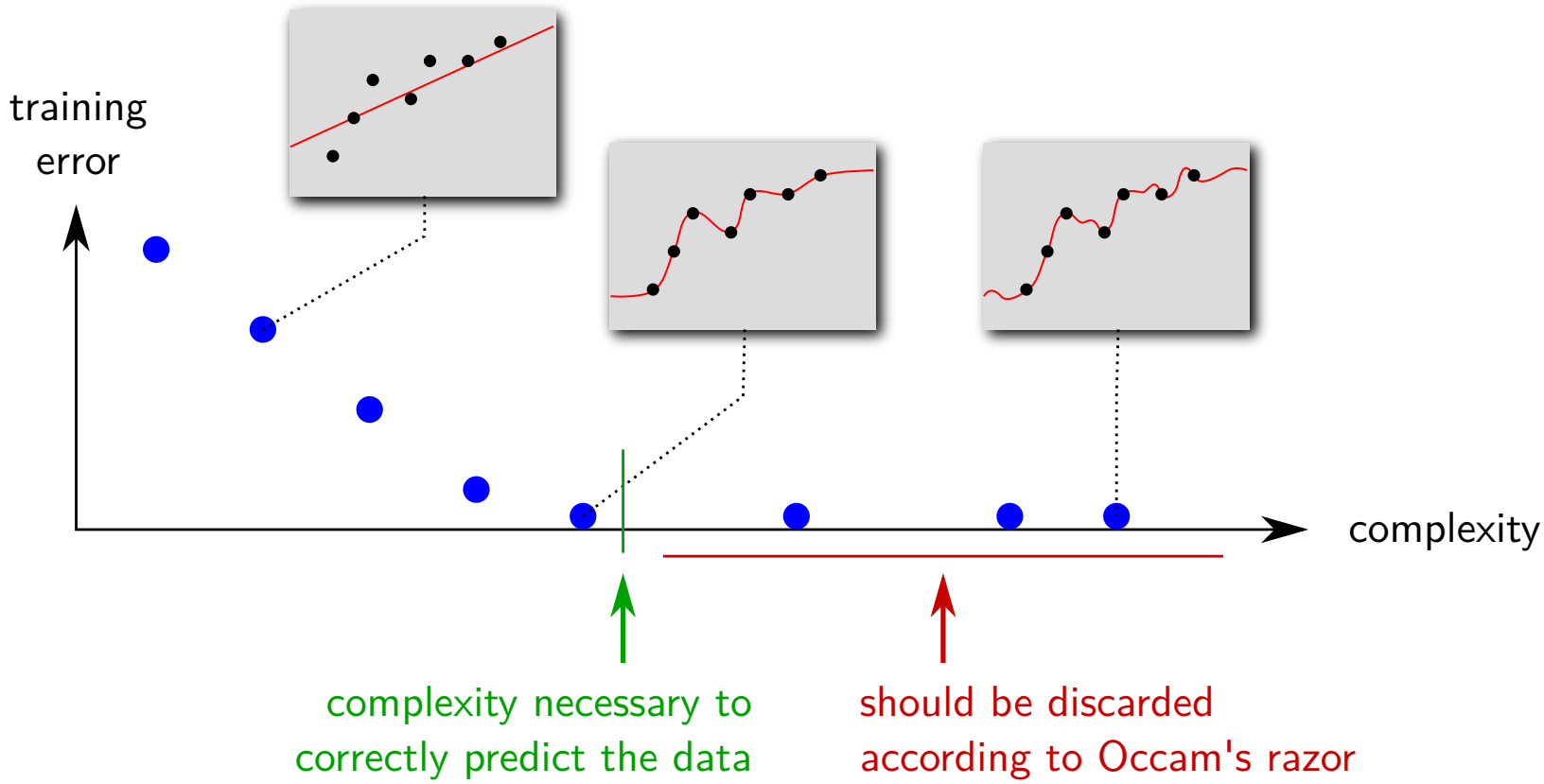
English translation

> *"Entities must not be multiplied beyond necessity."*

Machine learning reformulation:

> *"Model complexity must not be increased beyond what is necessary to correctly predict the data."*

Occam's Razor for Model Selection



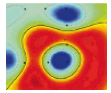
Two Interpretations of Occam's Razor

What do we gain from restricting model complexity?

1. If two models correctly predict the data, the least complex one should be preferred because *simplicity is desirable in itself*.
2. If two models correctly predict the data, the least complex one should be preferred because it is likely to *predict correctly new data points*.

In this lecture, we focus on (2).

Further reading: Domingos (1998) Occam's two Razors: The Sharp and the Blunt.



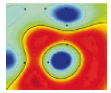
Quantifying Complexity

Quantifying complexity of a model is highly non-trivial and many proposal have been made:

Examples:

1. Counting the parameters of the model
2. Size of function class, aka. structural risk minimization (SRM)
3. Bayesian information criterion (BIC)
4. Minimum description length (MDL)
5. Smoothness / Lipschitzness
6. ...

In today's lecture, we discuss (1) and (2).



Approach 1: Counting the Parameters

Idea:

- If several models predict the data sufficiently well, prefer the one with the fewest parameters.

Examples of models:

Constant classifier $\quad g(\mathbf{x}) = \underbrace{C}_{1}$

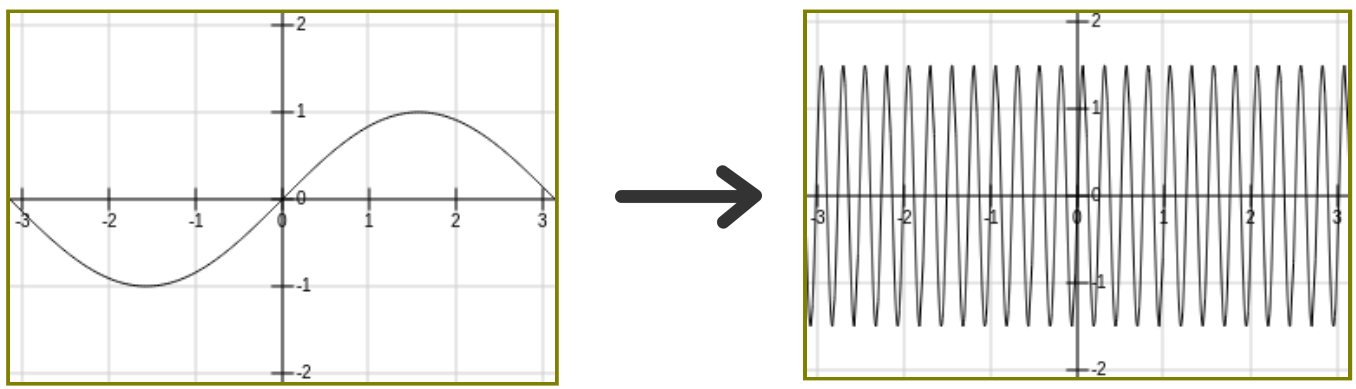
Means difference $\quad g(\mathbf{x}) = \mathbf{x}^\top \underbrace{(\mathbf{m}_1 - \mathbf{m}_2)}_{2d} + \underbrace{C}_{1}$

PCA + Fisher $\quad g(\mathbf{x}) = \underbrace{\mathrm{PCA}(\mathbf{x})^\top}_{k \cdot d} \underbrace{S_W^{-1}}_{k^2} \underbrace{(\mathbf{m}_1 - \mathbf{m}_2)}_{2k} + \underbrace{C}_{1}$

Approach 1: Counting the Parameters

Counter-example:

- The model $g(x) = a\sin(\omega x)$ has only two parameters but can fit almost *any* finite dataset in $\mathbb{R}$. This can be achieved by setting ω very large.

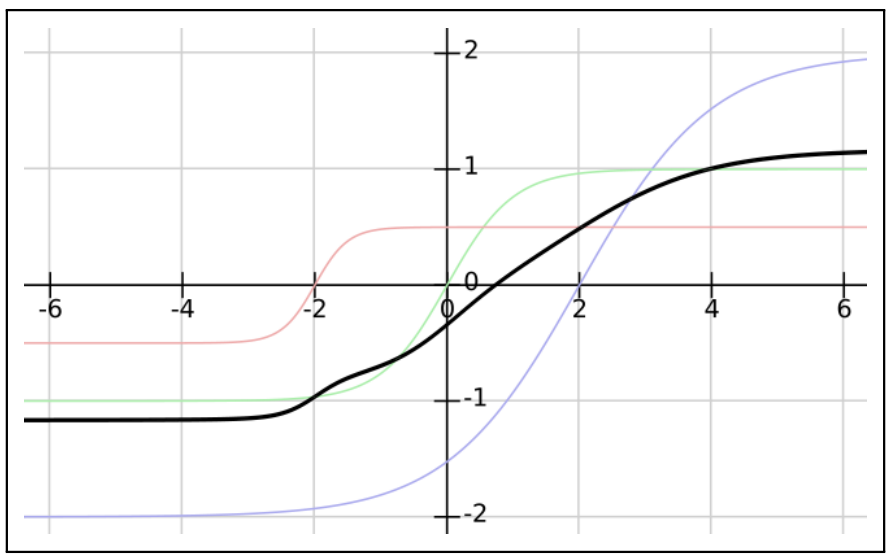


- However, the resulting high-frequency sinusoid will not work well at predicting new data points.

Approach 1: Counting the Parameters

Another counter-example:

- The model $g(x) = \frac{1}{K} \sum_{k=1}^{K} \alpha_k \tanh\left(\frac{x-\theta_k}{\alpha_k}\right)$ has a large number of parameters $(2 \cdot K)$ but is much more rigid than the sinusoid function (e.g. function never exceeds a slope of 1, and can only increase).

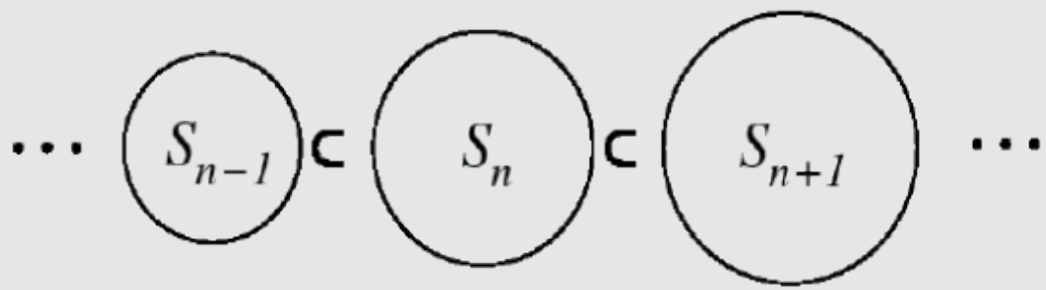


Approach 2: Structural Risk Minimization (SRM)

Structural risk minimization (Vapnik and Chervonenkis, 1974) is another approach to measure complexity and perform model selection.

SRM Idea:

- Structure the space of solutions into a nesting of increasingly large regions.



- If two solutions fit the data, prefer the solution that also belongs to the smaller regions.

Particular choices of $\cdots \subseteq S_{n-1} \subseteq S_n \subseteq S_{n+1} \subseteq \ldots$ lead to upper-bounds on the generalization error of a model (cf. next lecture).

Approach 2: Structural Risk Minimization (SRM)

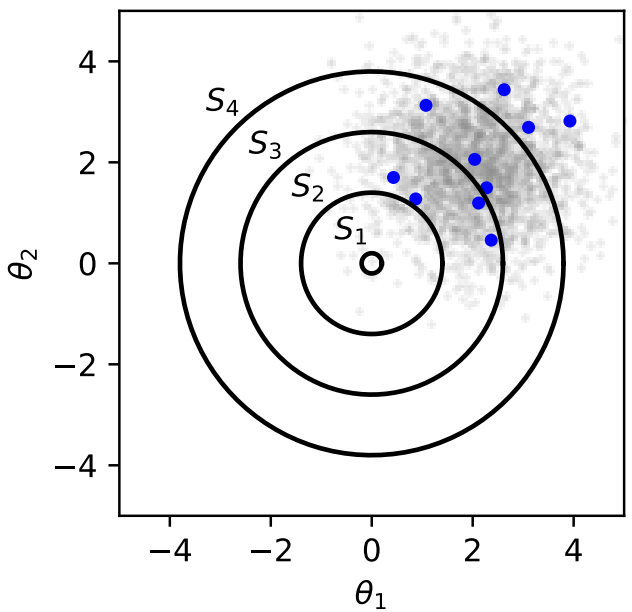
Example:

- Assume you would like to build an estimator $\widehat{\mu}$ of the mean $\mu \in \mathbb{R}^d$ of your distribution, based on some iid. sample $x_1, \ldots, x_N \in \mathbb{R}^d$.
- To apply the SRM principle, we first create a nested sequence

$$\cdots \subseteq S_{n-1} \subseteq S_n \subseteq S_{n+1} \subseteq \ldots$$

e.g. where

$$S_n = \left\{\theta \in \mathbb{R}^d : \|\theta\| \leq C_n\right\}$$



Approach 2: Structural Risk Minimization (SRM)

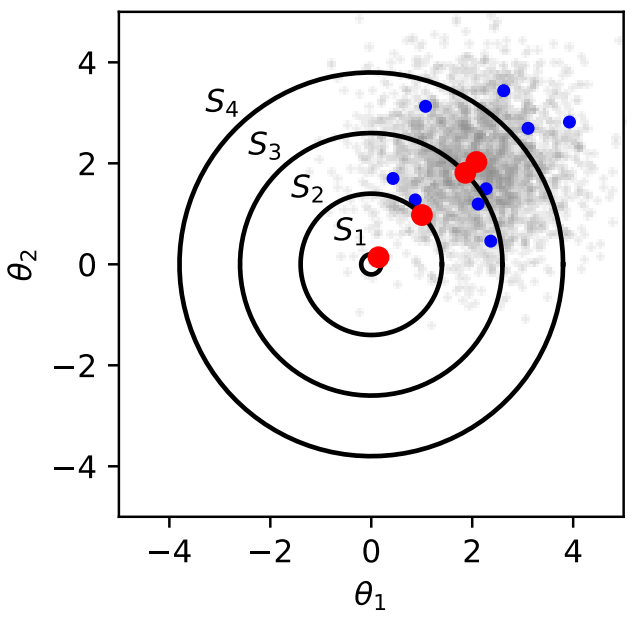
Example (cont.):

- The maximum likelihood estimator belonging to the set S_n can be found by solving

$$\widehat{\mu} = \arg\min_{\theta} \sum_{i=1}^{N} \|\theta - x_i\|^2$$
$$\text{s.t.} \quad \theta \in S_n.$$

  (shown in the figure as a red dot for various choices of S_n).

- Choosing $\widehat{\mu}$ associated to smaller sets S_n leads to better estimators of the true parameter μ (cf. the James-Stein estimator discussed later).



From Occam's Razor to Popper

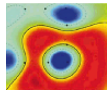
Occam's Razor

> *"Entities must not be multiplied beyond necessity."*

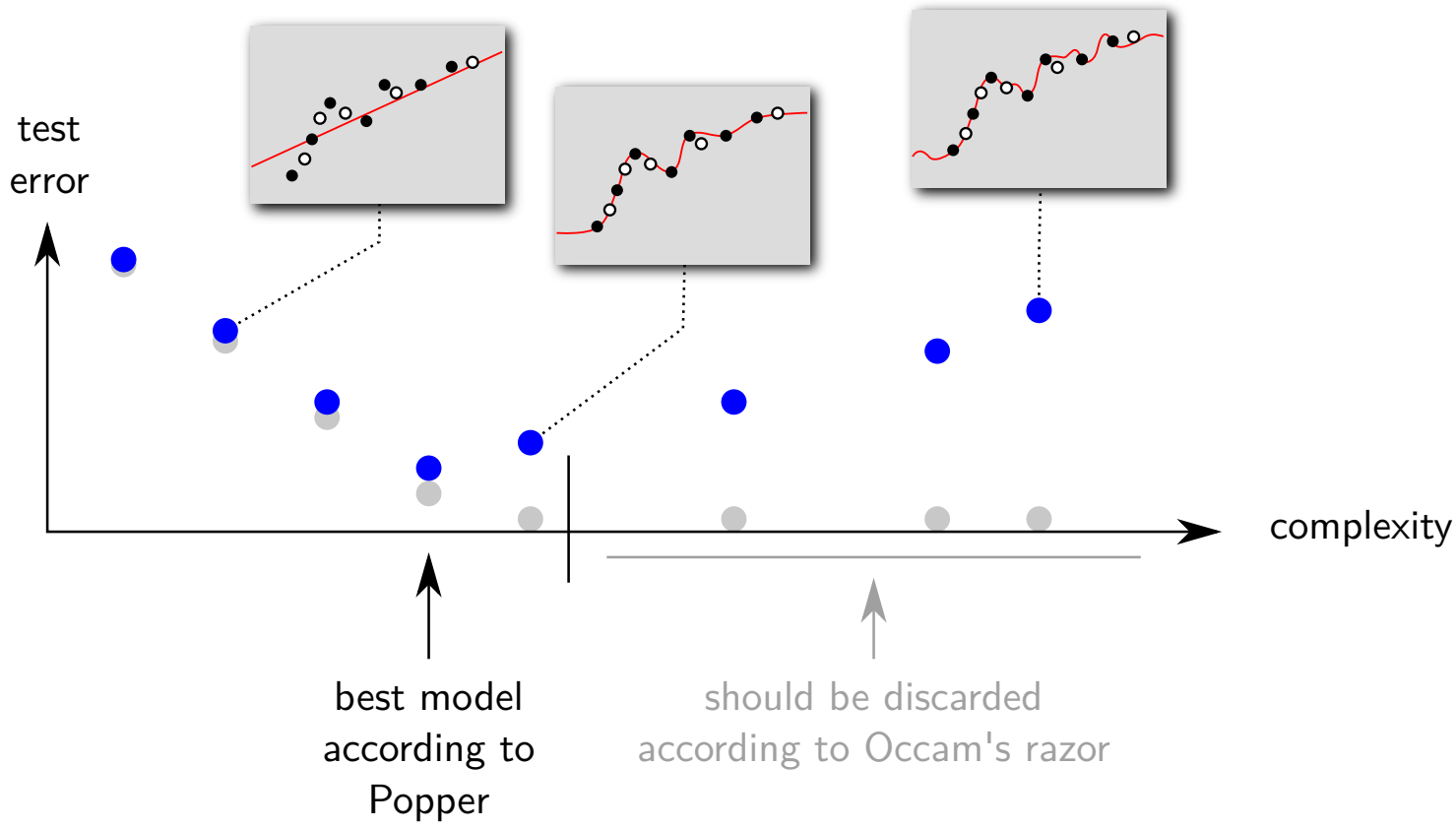
Falsifiability/prediction strength (S. Hawking, after K. Popper)

> *"[a good model] must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements,* **and** *it must make definite predictions about the results of future observations."*

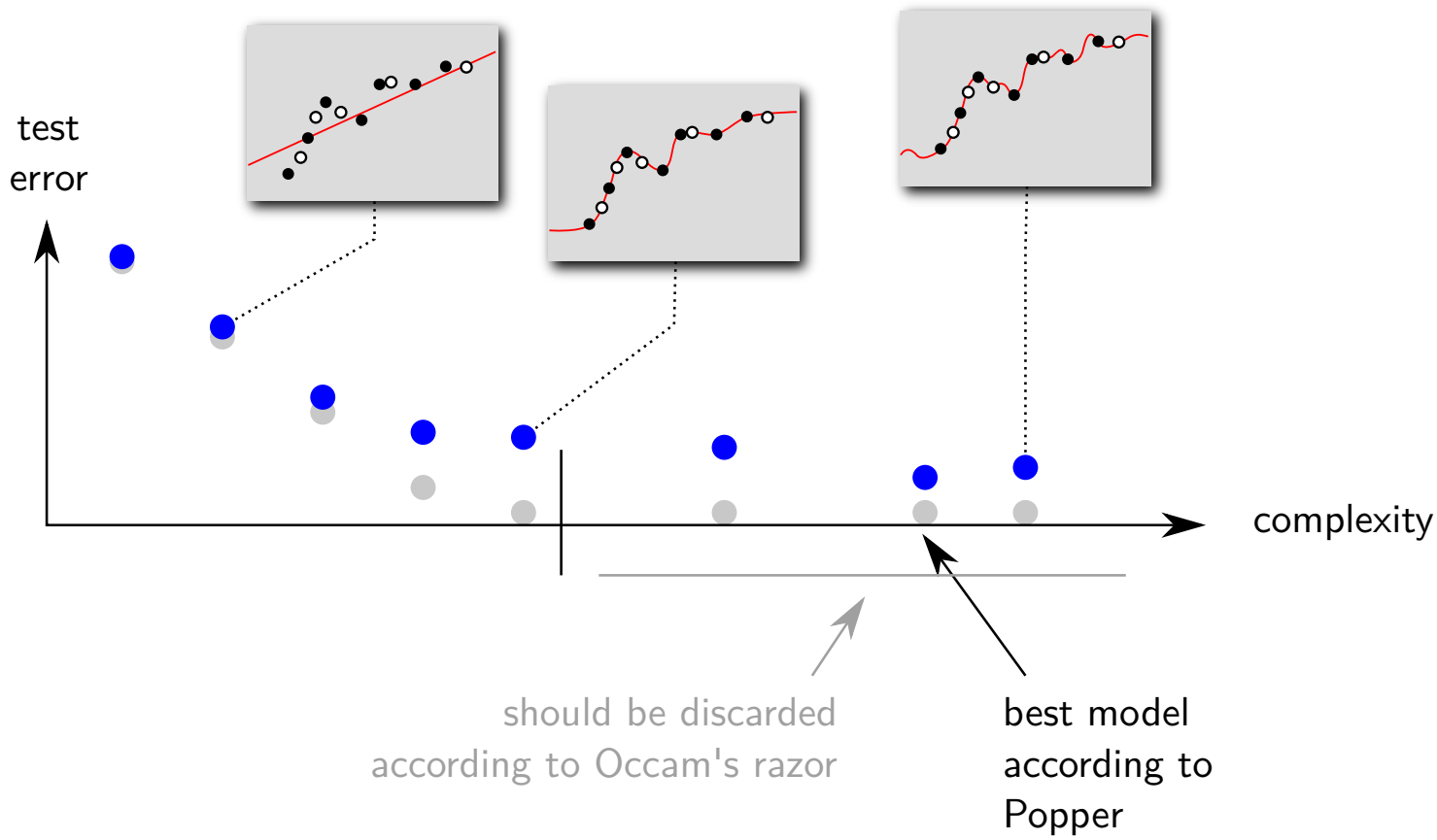
In other words, the model with lowest generalization error is preferable.



From Occam's Razor to Popper

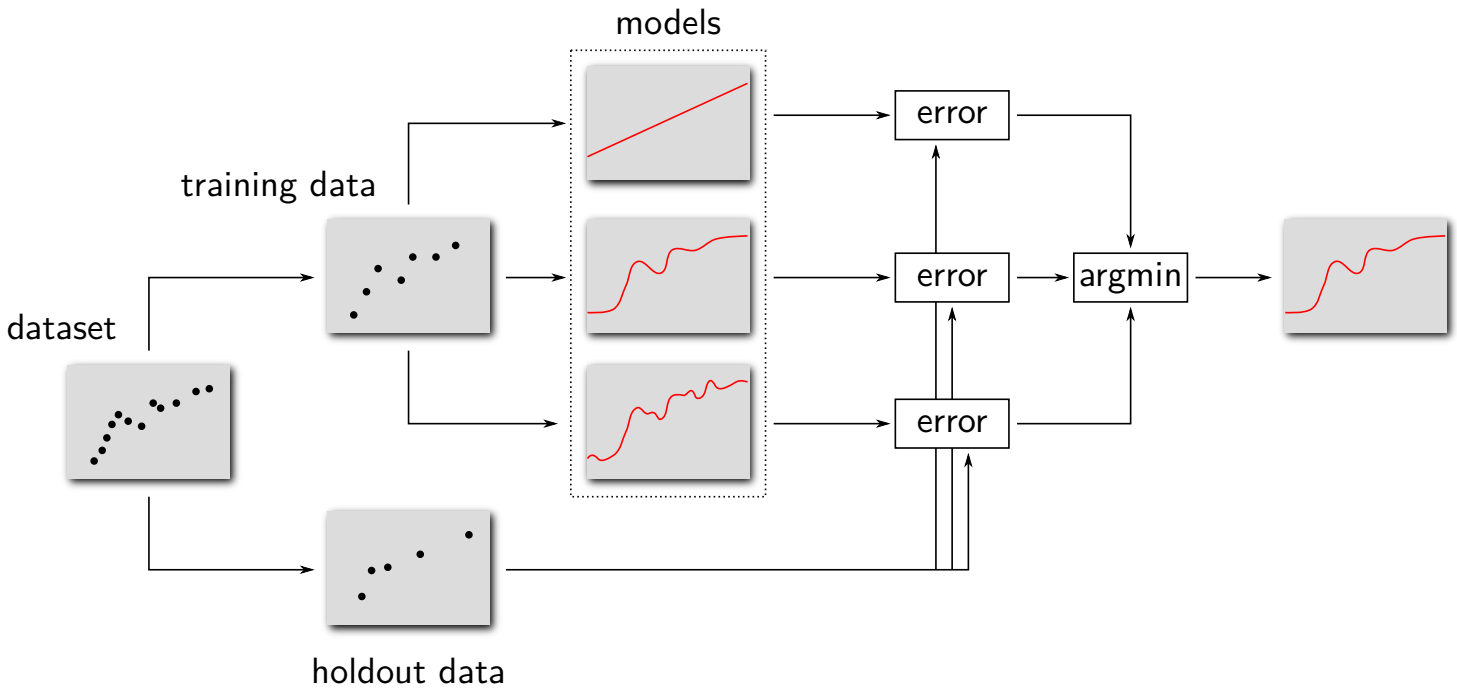


From Occam's Razor to Popper



The Holdout Selection Procedure

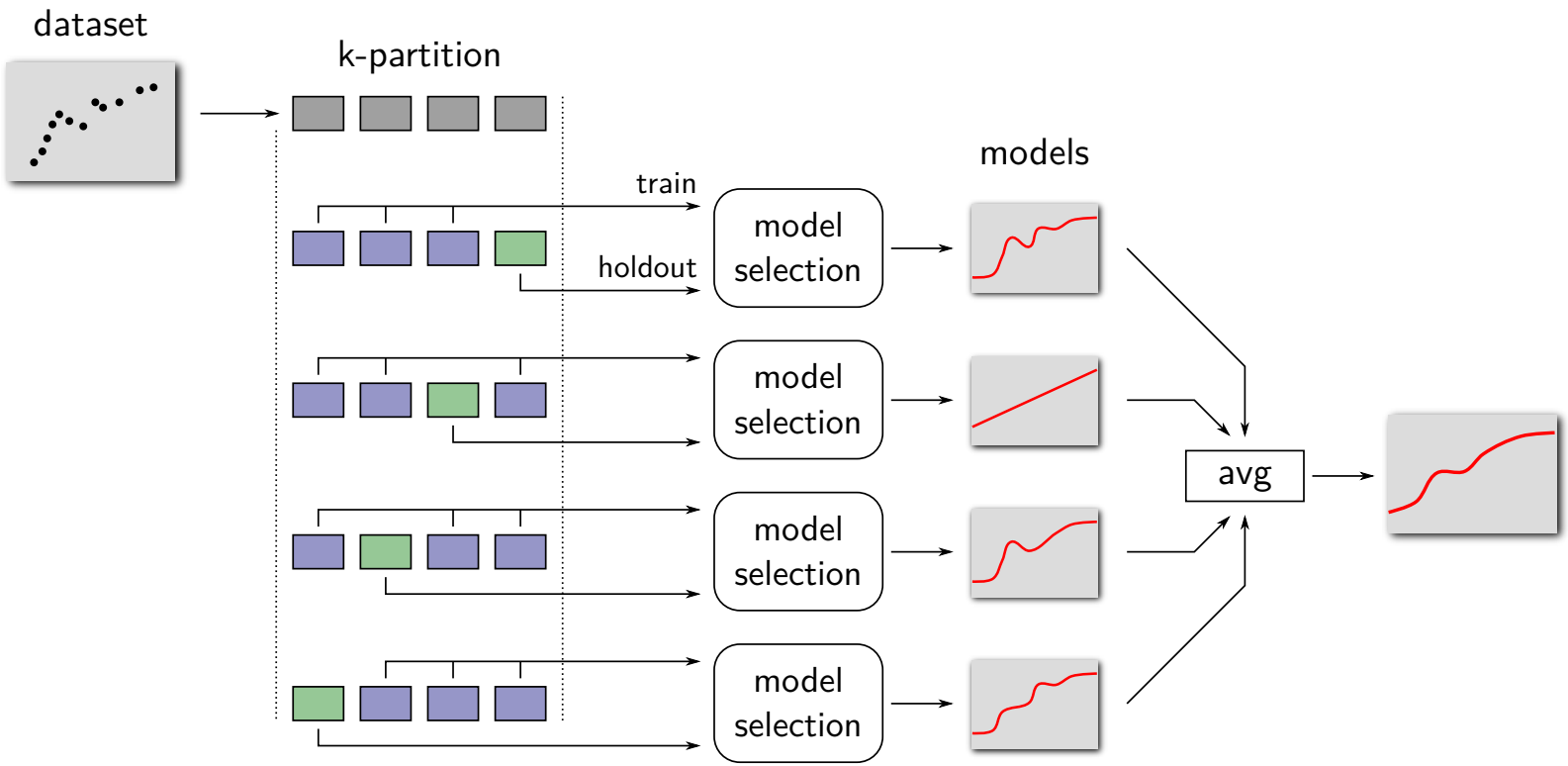
Idea: Predict out-of-sample error by splitting the data randomly in two parts (one for training, and one for estimating the error of the model).



Problem: The more data we use to accurately estimate the prediction error of each model, the less data is available for training an accurate model in the first place.

Cross-Validation (k-Fold Procedure)

Idea: Retain more data for training, and make up for the lower amount of holdout data by repeating the model selection procedure over multiple data splits (and averaging the selected models):



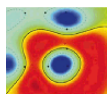
Holdout / Cross-Validation

Advantages:

- The model can now be selected directly based on simulated future observations (implements Popper's principle for model selection).

Limitations:

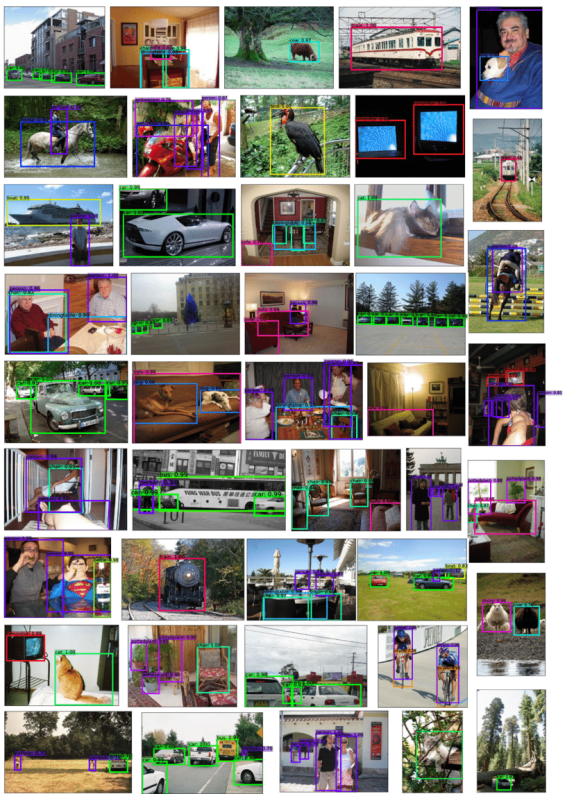
- For a small number of folds k, the training data is reduced significantly, which may lead to less accurate models. For k large, the procedure becomes computationally costly.
- This technique assumes that the available data is representative of the future observations (not always true!).



Limits of Holdout / Cross-Validation

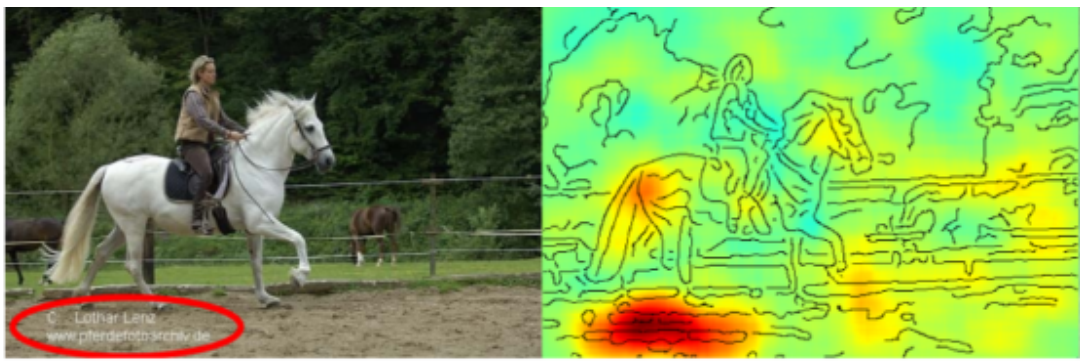
Example: *Pascal VOC 2007 dataset*

- Standard benchmark for image classification in the early 2010s.
- Large collection of images annotated with bounding boxes associated to twenty different object categories (airplane, boat, horse, person, ...).
- Holdout / cross-validation commonly used on this dataset for model selection and evaluation. *But does a low predicted error guarantee that the model will truly perform well?*



Limits of Holdout / Cross-Validation

- Explainable AI (presented later this semester) highlights features (e.g. pixels) used by the model to support its decision.



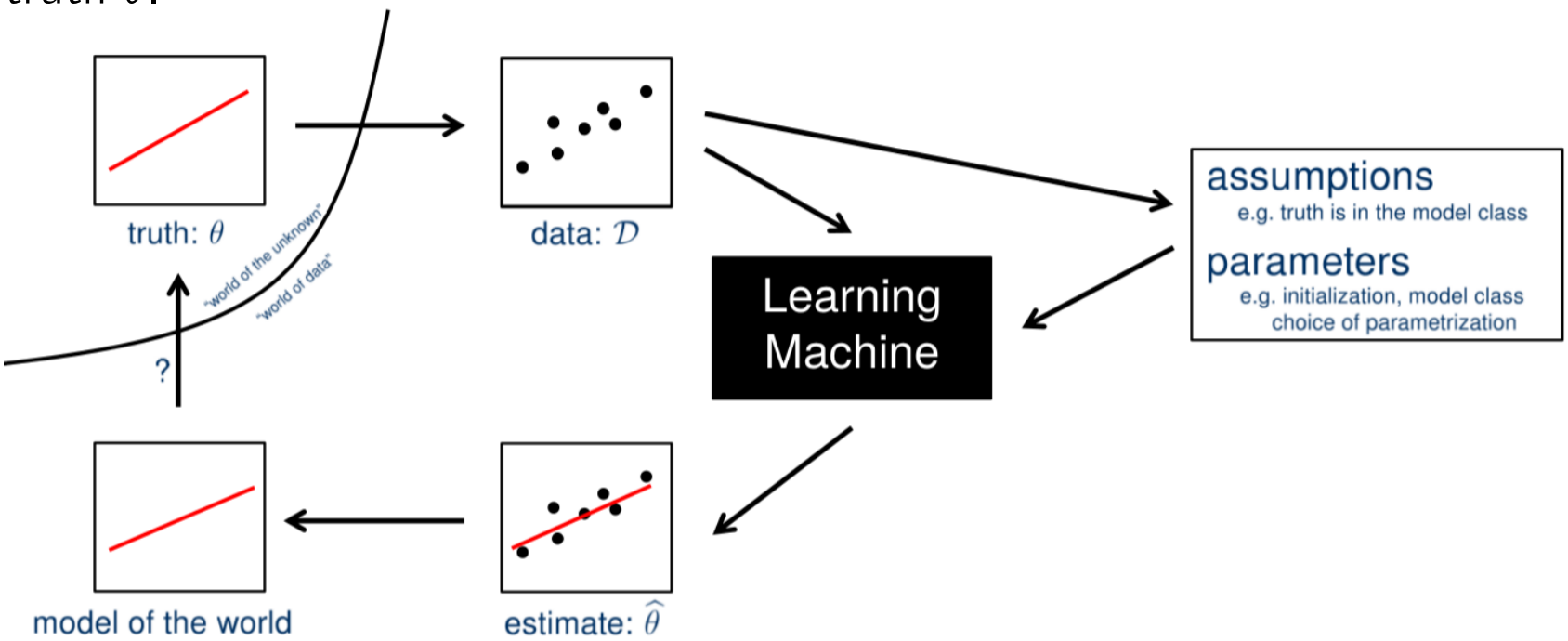
- Explanation reveals that the classifier at hand exploits a spurious correlation between the class horse and the presence of a copyright tag in the corner of the horse images (aka. Clever Hans effect).
- Spurious correlation only exists in the dataset, but not in real-world.
- Holdout / cross-validation cannot detect this flawed strategy and can result in poor model selection.

Further reading: Lapuschkin et al. Unmasking Clever Hans predictors and assessing what machines really learn, *Nature Communications*, 2019.

Part II. Bias-Variance Analysis of ML Models

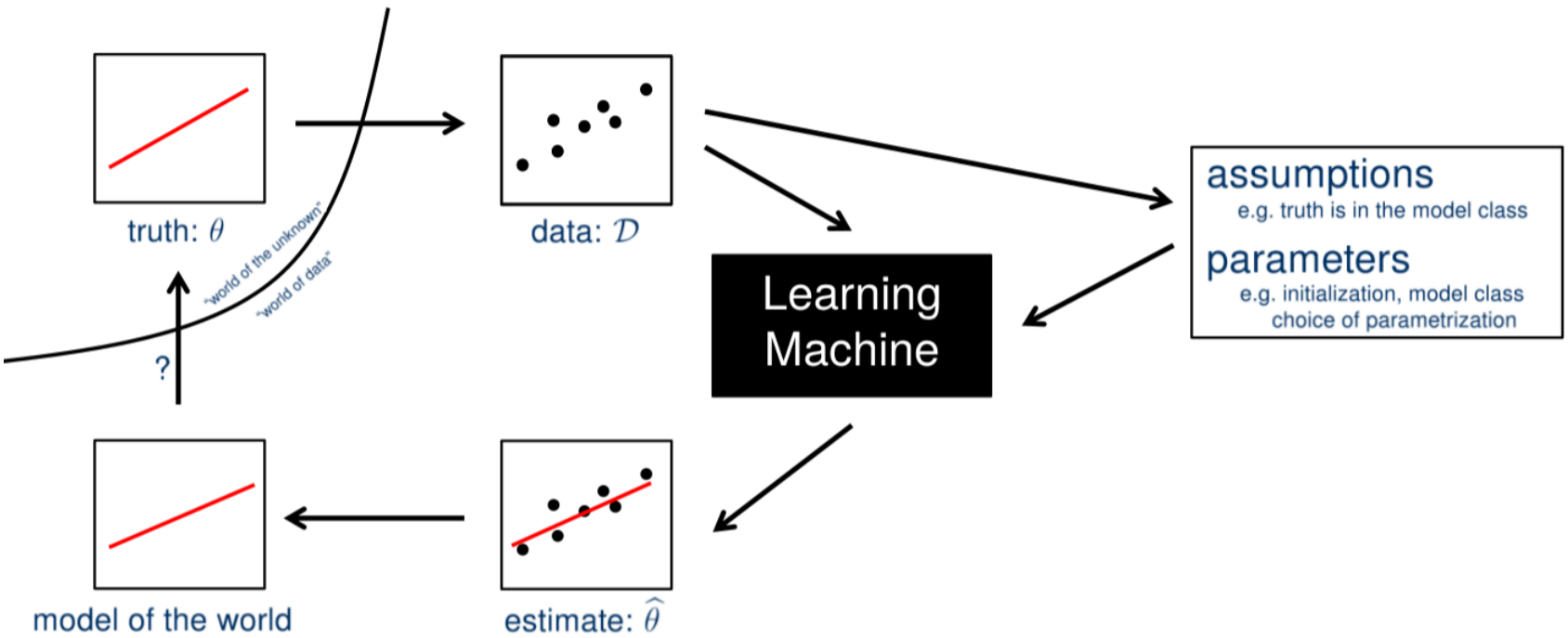
Machine Learning Models

Machine learning models are learned from the data to approximate some truth θ.



A learning machine can be abstracted as a function that maps a dataset $\mathcal{D}$ to an estimator $\hat{\theta}$ of the truth θ.

ML Models and Prediction Error



A good learning machine is one that produces an estimator $\hat{\theta}$ close to the truth θ. Closeness to the truth can be measured by some error function, e.g. the square error:

$$\text{Error}(\hat{\theta}) = (\theta - \hat{\theta})^2.$$

Bias, Variance, and MSE of an Estimator

Parametric estimation:

- θ is a value in $\mathbb{R}^h$
- $\widehat{\theta}$ is a function of the data $\mathcal{D} = \{X_1, \ldots, X_N\}$, where X_i are random variables producing the data points.

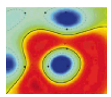
Statistics of the estimator:

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta] \qquad \text{(measures expected deviation of the mean)}$$

$$\text{Var}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2] \qquad \text{(measures scatter around estimator of mean)}$$

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2] \qquad \text{(measures prediction error)}$$

Note: for $\theta \in \mathbb{R}^h$, we use the notation $\theta^2 = \theta^\top \theta$.

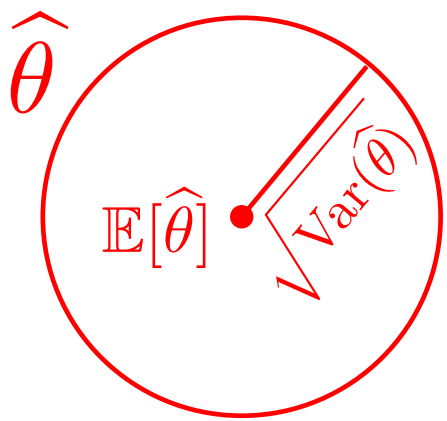


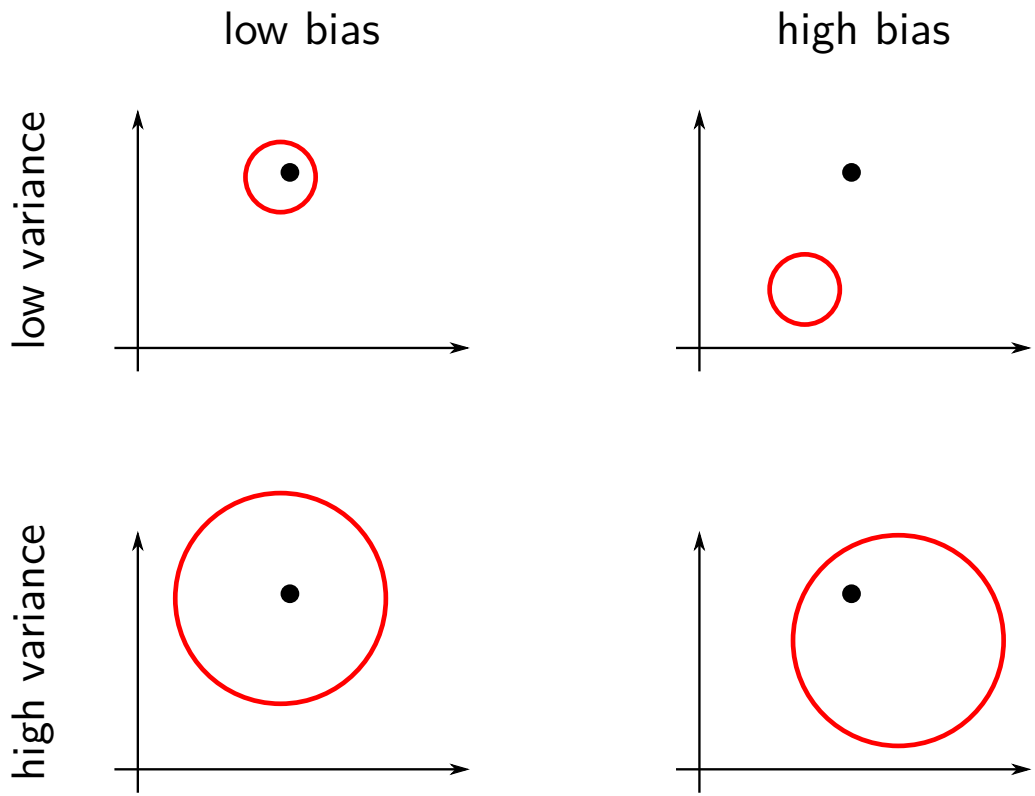
Visualizing Bias and Variance

True parameter

θ .

Parameter estimator





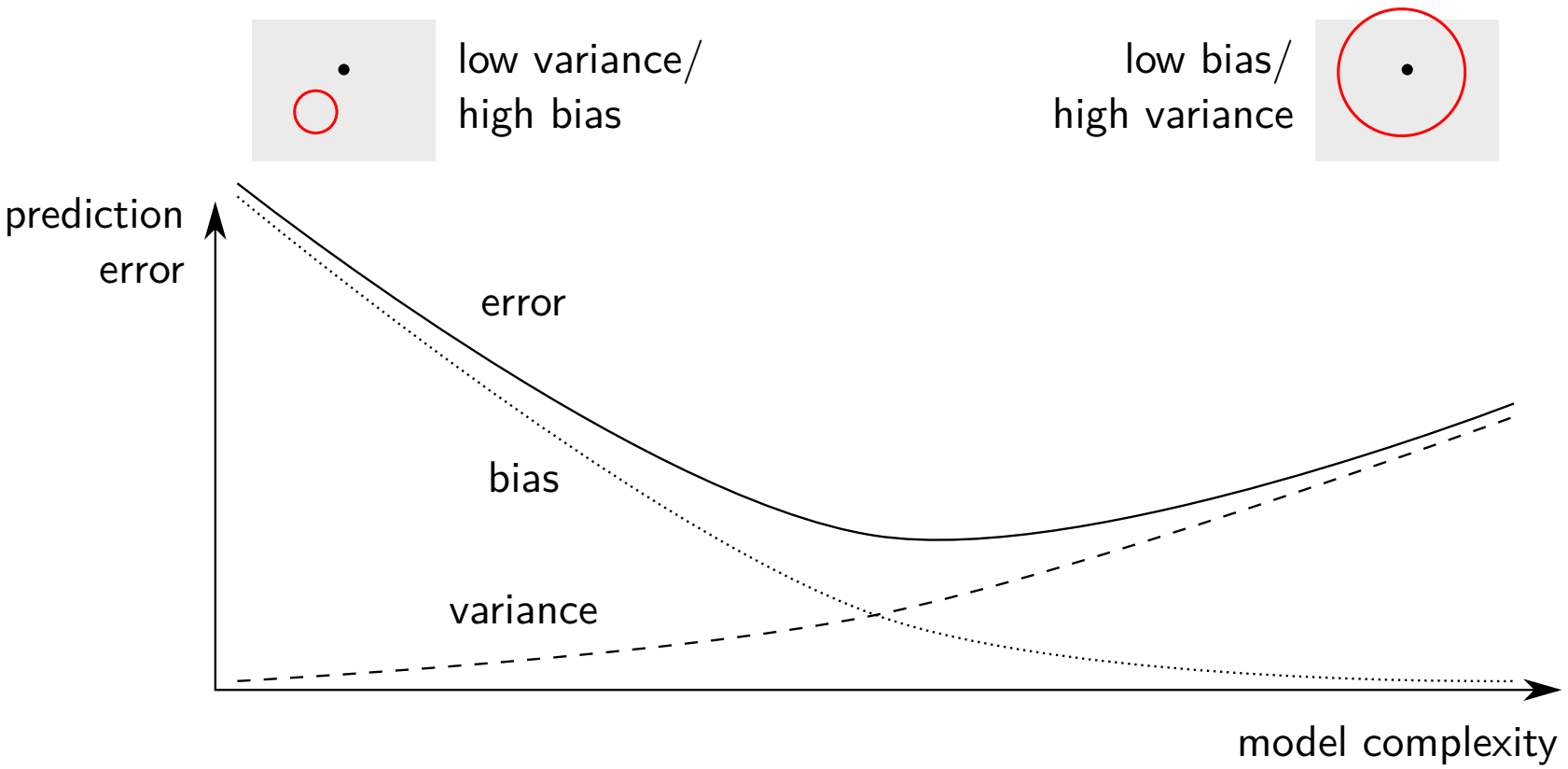
Bias-Variance Decomposition

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta], \quad \text{Var}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2], \quad \text{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2]$$

Exercise: Show that $\boxed{\text{MSE}(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})}$.

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \mathbb{E}[(\widehat{\theta} - \theta)^2] \\
&= \mathbb{E}[((\widehat{\theta} - \mathbb{E}[\widehat{\theta}]) + (\mathbb{E}[\widehat{\theta}] - \theta))^2] \\
&= \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2 + (\mathbb{E}[\widehat{\theta}] - \theta)^2 + 2(\widehat{\theta} - \mathbb{E}[\widehat{\theta}]) \cdot (\mathbb{E}[\widehat{\theta}] - \theta)] \\
&= \underbrace{\mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2]}_{\text{Var}(\widehat{\theta})} + \underbrace{(\mathbb{E}[\widehat{\theta} - \theta])^2}_{\text{Bias}(\widehat{\theta})^2} + 2\underbrace{\mathbb{E}[\widehat{\theta} - \mathbb{E}[\widehat{\theta}]]}_{0} \cdot (\mathbb{E}[\widehat{\theta}] - \theta)
\end{aligned}$$

Visualizing Bias and Variance



Example: Parameters of a Gaussian

parametric estimation:

θ is a value in $\mathbb{C}^n$ (e.g. $\theta = (\mu, \Sigma)$ for Gaussians)

$\widehat{\theta}$ is function in the data $\mathcal{D} = \{X_1, \ldots, X_N\}$

(X_i are random variables giving back data points)

e.g. mean estimator $\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i$

covariance estimator $\widehat{\Sigma} = \frac{1}{N-1} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$

Example: Parameters of a Gaussian

Exercise: Show that the bias and variance of the mean estimator $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ are given by

$$\boxed{\text{Bias}(\widehat{\mu}) = 0} \quad \text{and} \quad \boxed{\text{Var}(\widehat{\mu}) = \sigma^2/N}.$$

$$\begin{aligned}
\text{Bias}(\widehat{\mu}) &= \mathbb{E}[\widehat{\mu} - \mu] \\
&= \mathbb{E}[(\tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i) - \mu] \\
&= (\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{E}[X_i]) - \mu \\
&= (\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mu) - \mu \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\widehat{\mu}) &= \text{Var}[\tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i] \\
&= \tfrac{1}{N^2}\text{Var}[\textstyle\sum_{i=1}^{N} X_i] \\
&= \tfrac{1}{N^2}\textstyle\sum_{i=1}^{N} \text{Var}[X_i] \\
&= \tfrac{1}{N^2}\textstyle\sum_{i=1}^{N} \sigma^2 \\
&= \tfrac{1}{N^2} N\sigma^2 \\
&= \sigma^2/N
\end{aligned}$$

The James-Stein Estimator

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

"natural" estimator

$$\text{Bias}(\widehat{\mu}) = 0$$

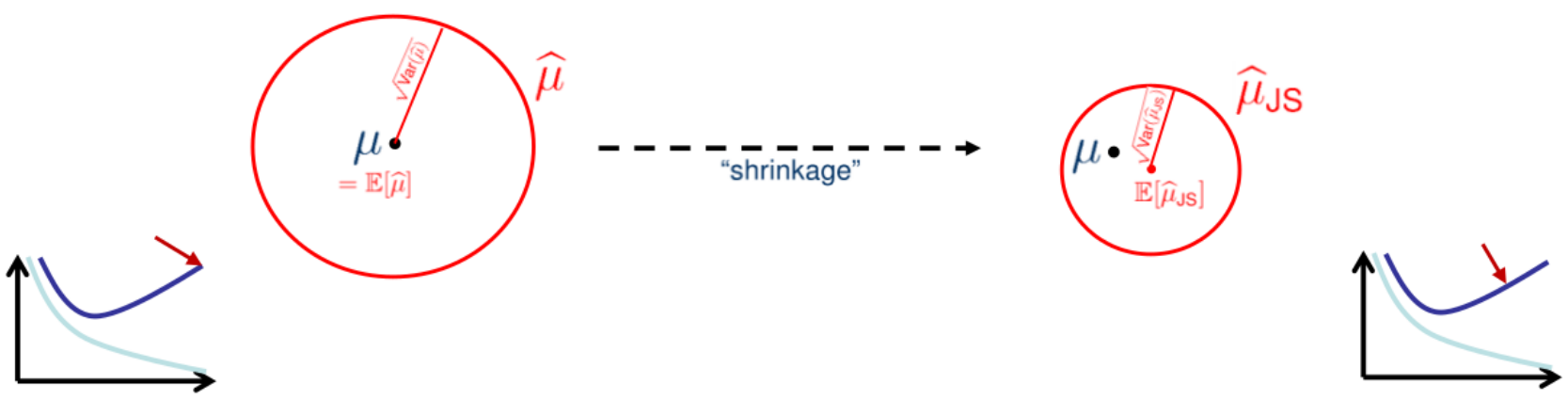
$$\text{MSE}(\widehat{\mu}) = \text{Var}(\widehat{\mu}) = \frac{\sigma^2}{N}$$

$$\widehat{\mu}_{\text{JS}} = \widehat{\mu} - \frac{(n-2)\sigma^2}{\widehat{\mu}^2}\widehat{\mu}$$

James-Stein estimator

$$\text{Bias}(\widehat{\mu}_{\text{JS}}) > 0$$

$$\text{MSE}(\widehat{\mu}_{\text{JS}}) < \text{MSE}(\widehat{\mu})$$



Estimator of Functions

supervised learning:

training data $\mathcal{D}$ is $X_1, \ldots, X_N$ with labels $Y_1, \ldots, Y_N$
(e.g. in regression, $X_i \in \mathbb{R}^n, Y_i \in \mathbb{R}$)

parameter θ "is" a generative function $f = f_\theta$:

$$Y_i = f(X_i) + \varepsilon_i$$

ε_i is error with $\mathbb{E}[\varepsilon_i] = 0$

Learning Machine learns approximation $\widehat{f} = f_{\widehat{\theta}}$
such that $Y_i \approx \widehat{f}(X_i)$

Example (Linear Regression):

$$f(x) = \beta^\top x + \alpha, \quad \theta = (\alpha, \beta)$$

Bias-Variance Analysis of the Function Estimator (locally)

supervised learning:

training data $\mathcal{D}$ is $X_1, \ldots, X_N$ with labels $Y_1, \ldots, Y_N$
(e.g. in regression, $X_i \in \mathbb{R}^n, Y_i \in \mathbb{R}$)

parameter θ "is" a generative function $f = f_\theta$:

$$Y_i = f(X_i) + \varepsilon_i$$

bias of $\widehat{f}$ at X_i: $\quad \mathrm{Bias}(\widehat{f}|X_i) = \mathbb{E}_Y[\widehat{f}(X_i) - f(X_i)]$

variance of $\widehat{f}$ at X_i: $\quad \mathrm{Var}(\widehat{f}|X_i) = \mathbb{E}_Y\left[(\widehat{f}(X_i) - \mathbb{E}_Y[\widehat{f}(X_i)])^2\right]$

MSE of $\widehat{f}$ at X_i: $\quad \mathrm{MSE}(\widehat{f}|X_i) = \mathbb{E}_Y\left[(\widehat{f}(X_i) - Y_i)^2\right]$

Proposition: $\mathrm{MSE}(\widehat{f}|X_i) = \mathrm{Var}(\varepsilon_i) + \mathrm{Bias}(\widehat{f}|X_i)^2 + \mathrm{Var}(\widehat{f}|X_i)$

Summary

- **Occam's Razor:** Given two models that have sufficiently low training error, the simpler one should be preferred.
  - **Measuring Complexity:** Counting the parameters can be misleading. Structured risk minimization (SRM) is more reliable in practice.
- **Popper's View:** How to make sure that a model predicts well? By testing it on out-of-sample data.
  - **Holdout and Cross-Validation:** Common practical procedures to simulate out-of-sample prediction behavior. Has some limitations (e.g. assume that the current dataset is representative of the true distribution).
- **Bias-Variance Decomposition:** The error of a predictive model can be decomposed into bias and variance. Best models can often be interpreted as finding a good tradeoff between the two terms.