Lecture 08

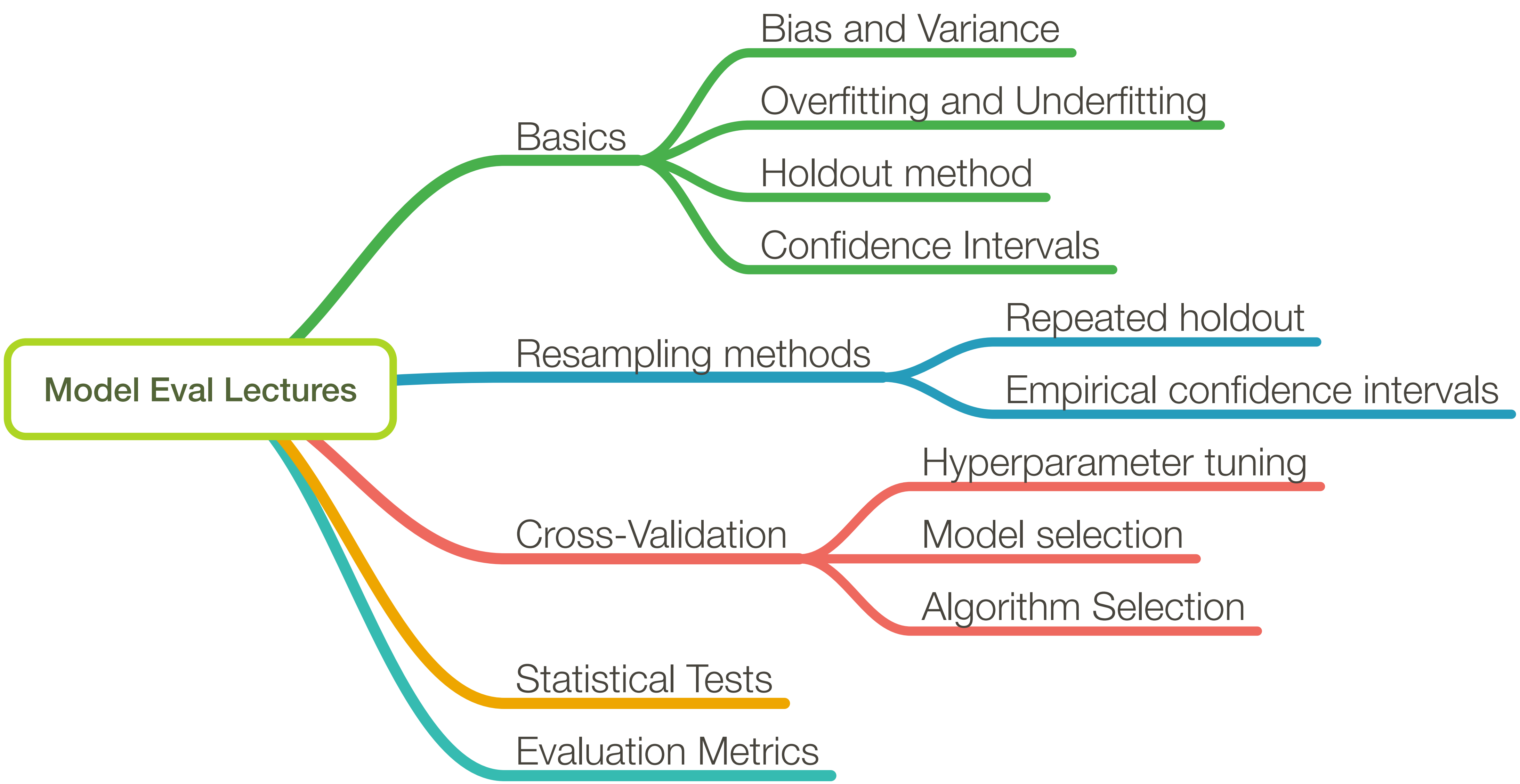
Model Evaluation 1:
Introduction to Overfitting and Underfitting

STAT 479: Machine Learning, Fall 2018

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

Overview



Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Evaluation Metrics

Overfitting and Underfitting

Overfitting and Underfitting

"Generalization Performance"

- Want a model to "generalize" well to unseen data
  ("high generalization accuracy" or "low generalization error")

Overfitting and Underfitting

Assumptions

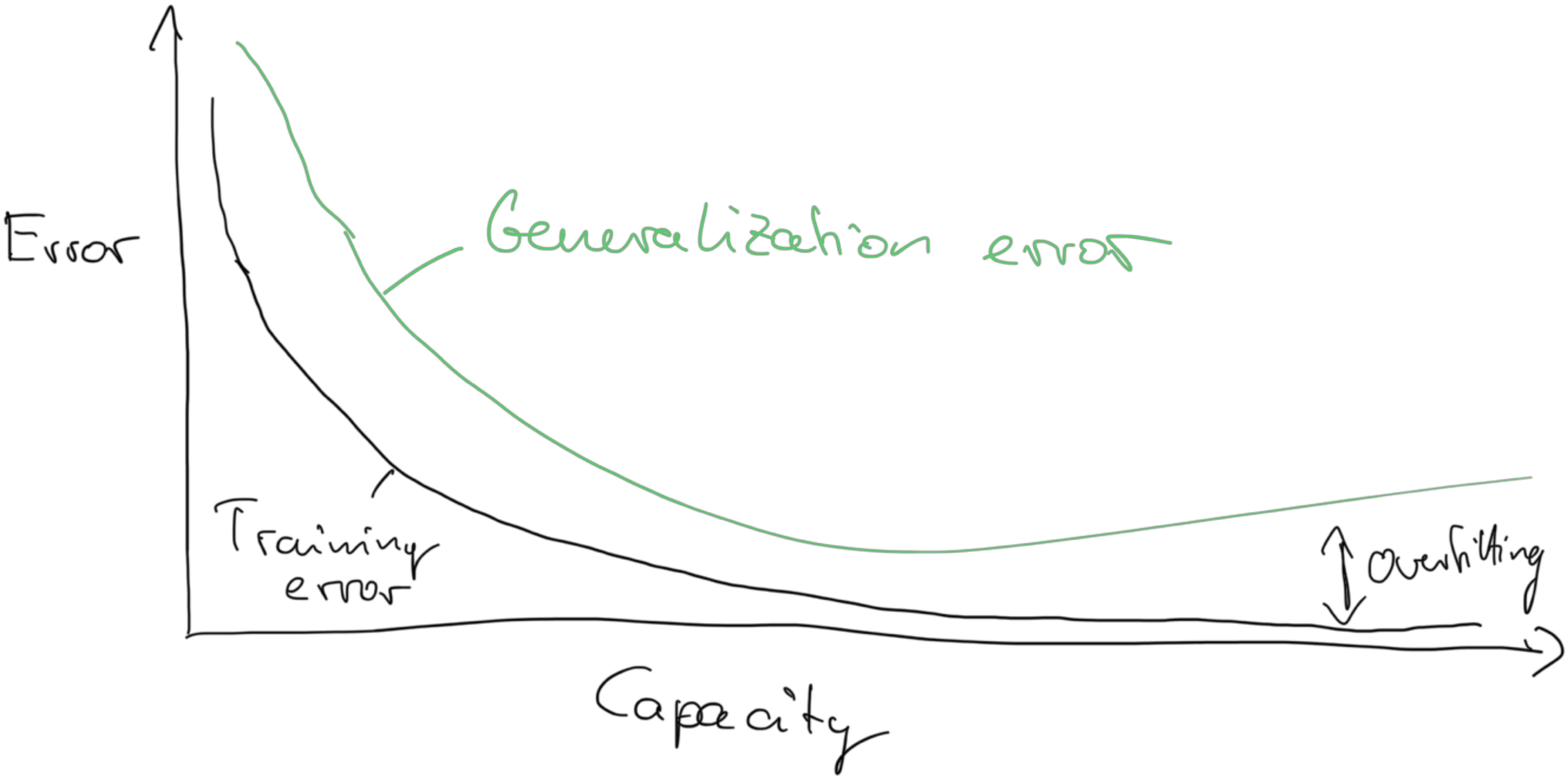
- i.i.d. assumption: inputs are independent, and training and test examples are identically distributed (drawn from the same probability distribution)
- For some random model that has not been fitted to the training set, we expect both the training and test error to be equal
- The training error or accuracy provides an (optimistically) biased estimate of the generalization performance

Overfitting and Underfitting

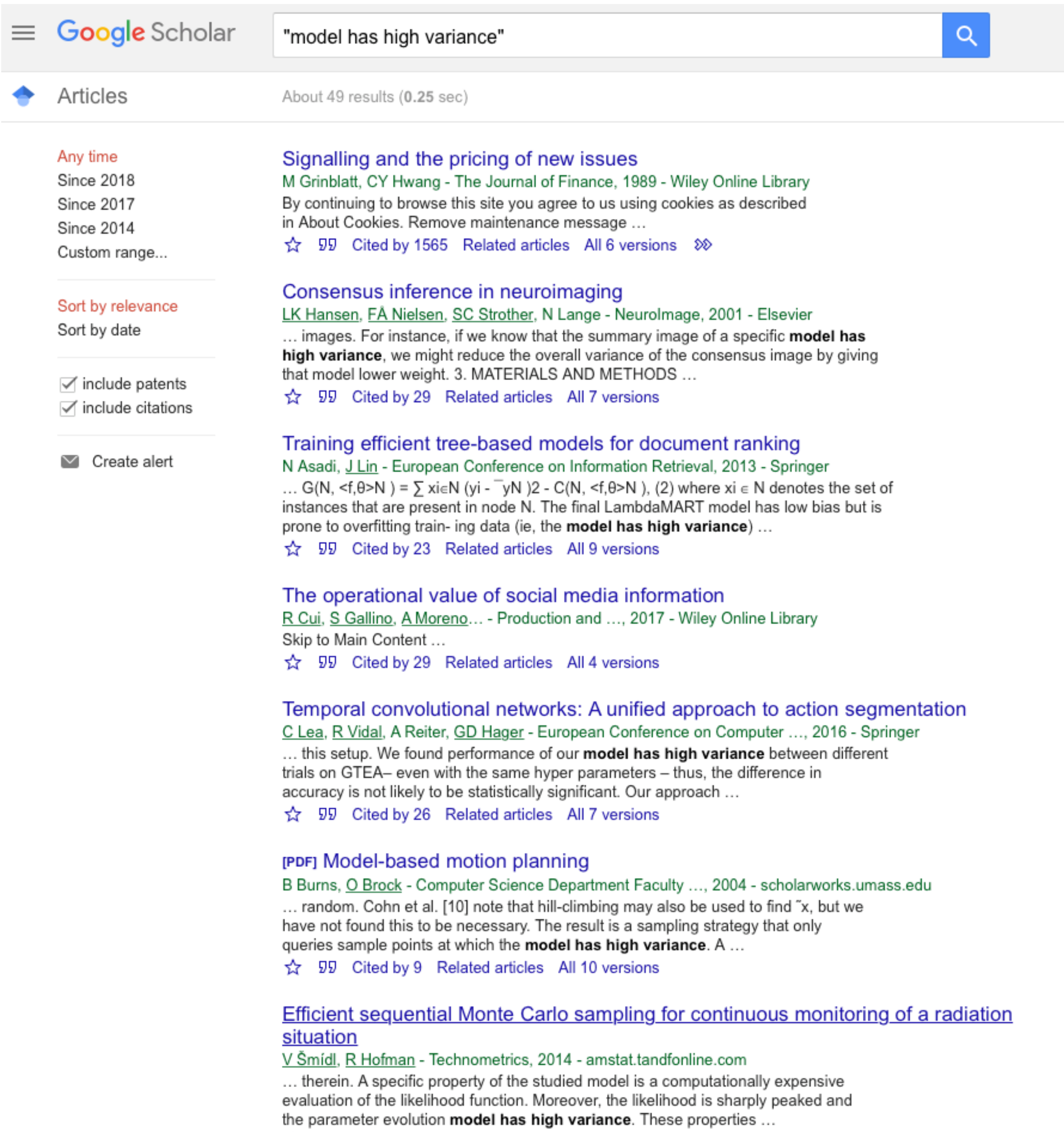
Model Capacity

- Underfitting: both training and test error are large
- Overfitting: gap between training and test error (where test error is higher)
- Large hypothesis space being searched by a learning algorithm
  -> high tendency to overfit

Overfitting and Underfitting



"[...] model has high bias/variance" -- What does that mean?



Google Scholar "model has high variance"

Articles — About 49 results (**0.25** sec)

Any time
Since 2018
Since 2017
Since 2014
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

Signalling and the pricing of new issues
M Grinblatt, CY Hwang - The Journal of Finance, 1989 - Wiley Online Library
By continuing to browse this site you agree to us using cookies as described in About Cookies. Remove maintenance message ...
Cited by 1565 Related articles All 6 versions

Consensus inference in neuroimaging
LK Hansen, FÅ Nielsen, SC Strother, N Lange - NeuroImage, 2001 - Elsevier
... images. For instance, if we know that the summary image of a specific **model has high variance**, we might reduce the overall variance of the consensus image by giving that model lower weight. 3. MATERIALS AND METHODS ...
Cited by 29 Related articles All 7 versions

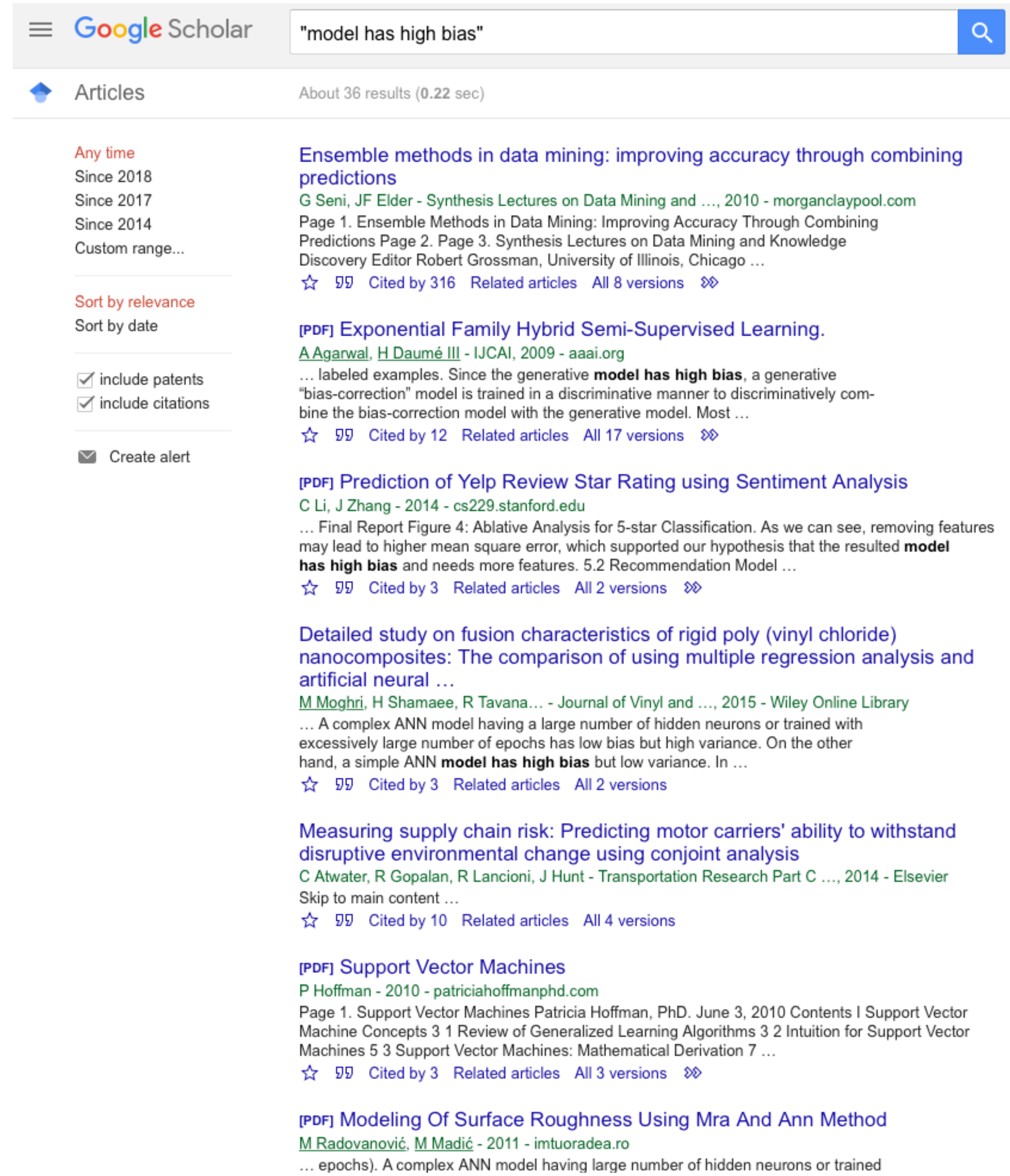
Training efficient tree-based models for document ranking
N Asadi, J Lin - European Conference on Information Retrieval, 2013 - Springer
... $G(N, \langle f,\theta\rangle_N) = \sum_{x_i \in N} (y_i - \bar{y}_N)^2 - C(N, \langle f,\theta\rangle_N)$, (2) where $x_i \in N$ denotes the set of instances that are present in node N. The final LambdaMART model has low bias but is prone to overfitting train- ing data (ie, the **model has high variance**) ...
Cited by 23 Related articles All 9 versions

The operational value of social media information
R Cui, S Gallino, A Moreno... - Production and ..., 2017 - Wiley Online Library
Skip to Main Content ...
Cited by 29 Related articles All 4 versions

Temporal convolutional networks: A unified approach to action segmentation
C Lea, R Vidal, A Reiter, GD Hager - European Conference on Computer ..., 2016 - Springer
... this setup. We found performance of our **model has high variance** between different trials on GTEA– even with the same hyper parameters – thus, the difference in accuracy is not likely to be statistically significant. Our approach ...
Cited by 26 Related articles All 7 versions

[PDF] **Model-based motion planning**
B Burns, O Brock - Computer Science Department Faculty ..., 2004 - scholarworks.umass.edu
... random. Cohn et al. [10] note that hill-climbing may also be used to find ¯x, but we have not found this to be necessary. The result is a sampling strategy that only queries sample points at which the **model has high variance**. A ...
Cited by 9 Related articles All 10 versions

Efficient sequential Monte Carlo sampling for continuous monitoring of a radiation situation
V Šmídl, R Hofman - Technometrics, 2014 - amstat.tandfonline.com
... therein. A specific property of the studied model is a computationally expensive evaluation of the likelihood function. Moreover, the likelihood is sharply peaked and the parameter evolution **model has high variance**. These properties ...



Google Scholar "model has high bias"

Articles — About 36 results (**0.22** sec)

Any time
Since 2018
Since 2017
Since 2014
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

Ensemble methods in data mining: improving accuracy through combining predictions
G Seni, JF Elder - Synthesis Lectures on Data Mining and ..., 2010 - morganclaypool.com
Page 1. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions Page 2. Page 3. Synthesis Lectures on Data Mining and Knowledge Discovery Editor Robert Grossman, University of Illinois, Chicago ...
Cited by 316 Related articles All 8 versions

[PDF] **Exponential Family Hybrid Semi-Supervised Learning.**
A Agarwal, H Daumé III - IJCAI, 2009 - aaai.org
... labeled examples. Since the generative **model has high bias**, a generative "bias-correction" model is trained in a discriminative manner to discriminatively com- bine the bias-correction model with the generative model. Most ...
Cited by 12 Related articles All 17 versions

[PDF] **Prediction of Yelp Review Star Rating using Sentiment Analysis**
C Li, J Zhang - 2014 - cs229.stanford.edu
... Final Report Figure 4: Ablative Analysis for 5-star Classification. As we can see, removing features may lead to higher mean square error, which supported our hypothesis that the resulted **model has high bias** and needs more features. 5.2 Recommendation Model ...
Cited by 3 Related articles All 2 versions

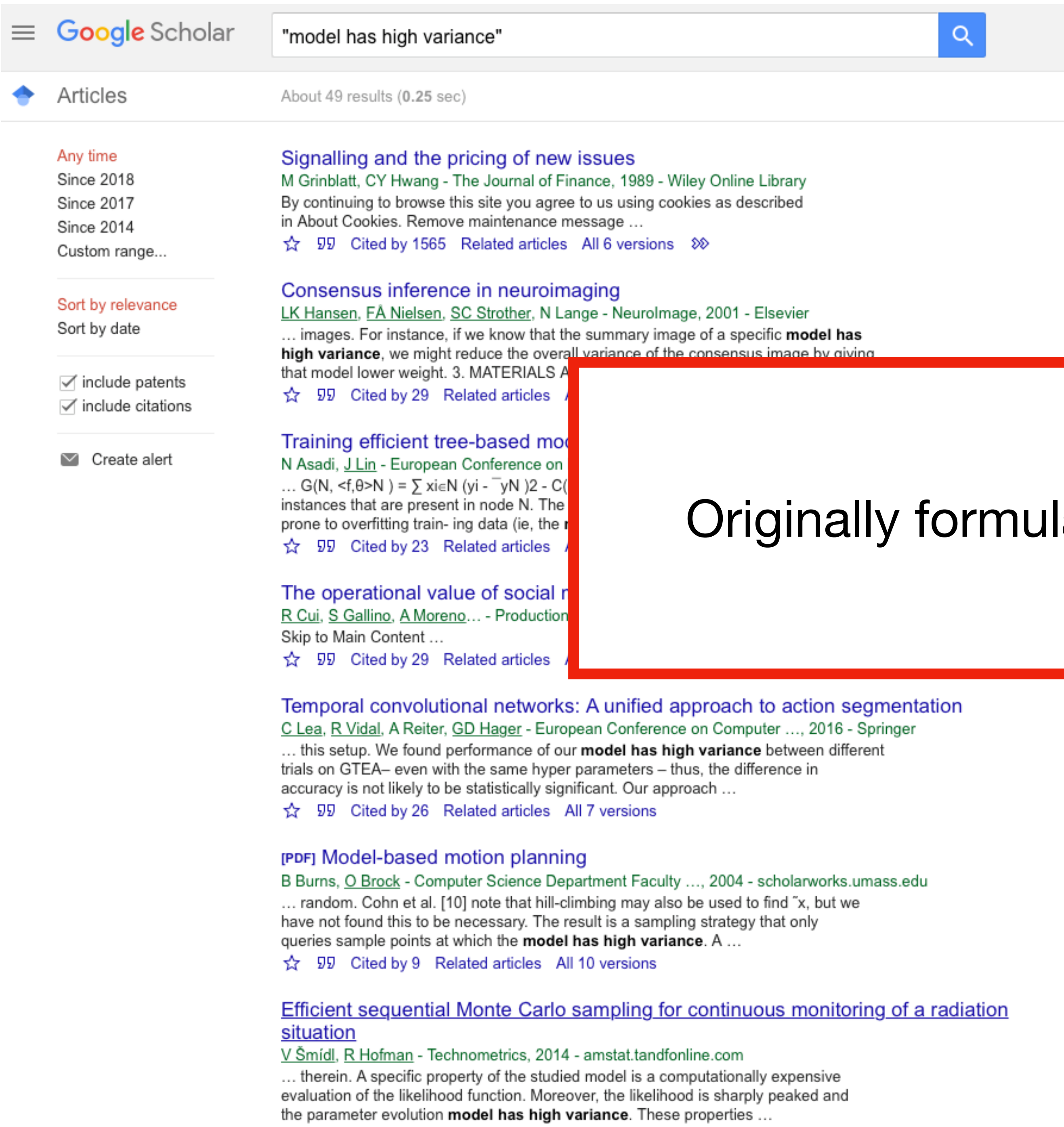
Detailed study on fusion characteristics of rigid poly (vinyl chloride) nanocomposites: The comparison of using multiple regression analysis and artificial neural ...
M Moghri, H Shamaee, R Tavana... - Journal of Vinyl and ..., 2015 - Wiley Online Library
... A complex ANN model having a large number of hidden neurons or trained with excessively large number of epochs has low bias but high variance. On the other hand, a simple ANN **model has high bias** but low variance. In ...
Cited by 3 Related articles All 2 versions

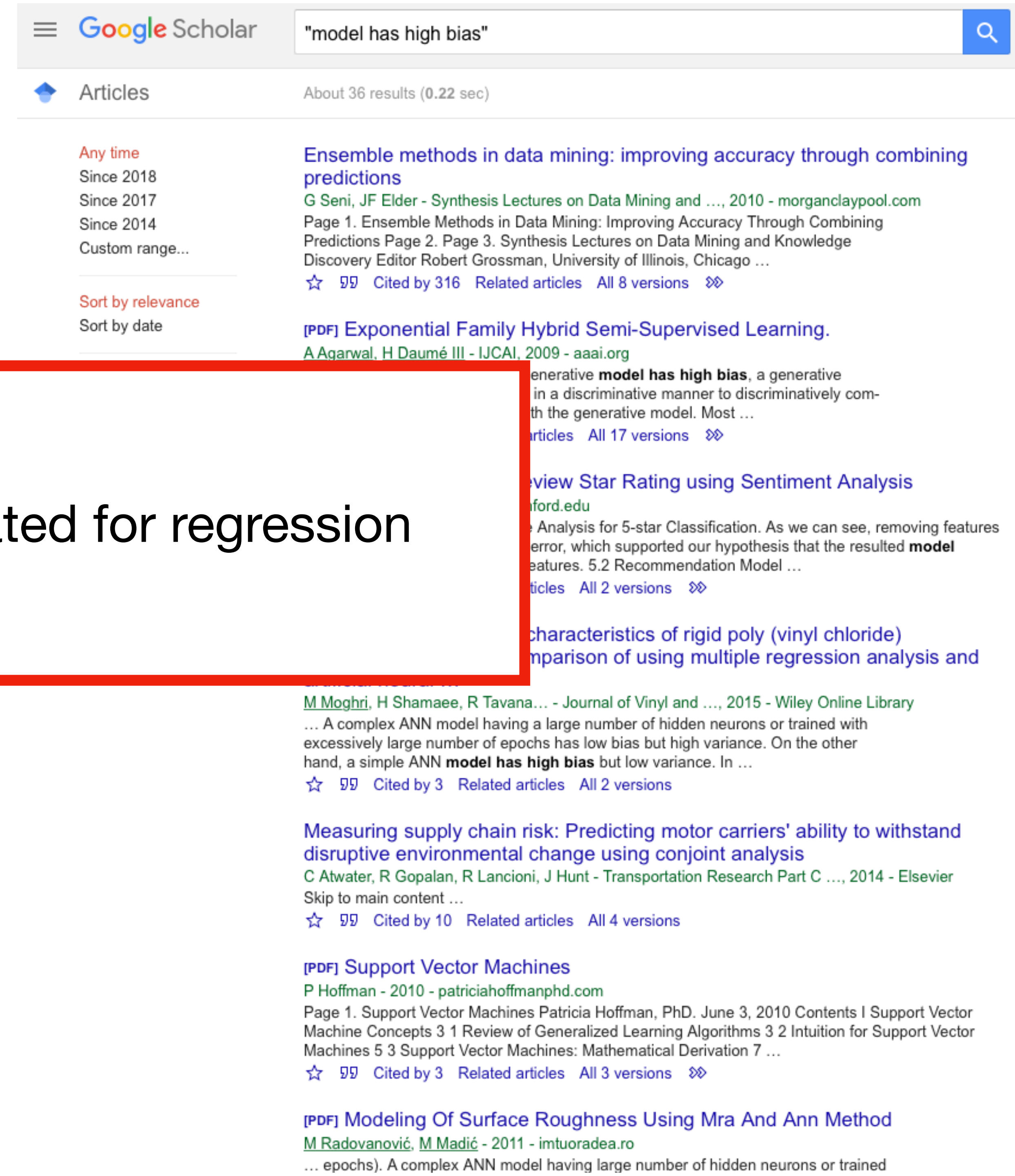
Measuring supply chain risk: Predicting motor carriers' ability to withstand disruptive environmental change using conjoint analysis
C Atwater, R Gopalan, R Lancioni, J Hunt - Transportation Research Part C ..., 2014 - Elsevier
Skip to main content ...
Cited by 10 Related articles All 4 versions

[PDF] **Support Vector Machines**
P Hoffman - 2010 - patriciahoffmanphd.com
Page 1. Support Vector Machines Patricia Hoffman, PhD. June 3, 2010 Contents I Support Vector Machine Concepts 3 1 Review of Generalized Learning Algorithms 3 2 Intuition for Support Vector Machines 5 3 Support Vector Machines: Mathematical Derivation 7 ...
Cited by 3 Related articles All 3 versions

[PDF] **Modeling Of Surface Roughness Using Mra And Ann Method**
M Radovanović, M Madić - 2011 - imtuoradea.ro
... epochs). A complex ANN model having large number of hidden neurons or trained

"[...] model has high bias/variance" -- What does that mean?





Originally formulated for regression

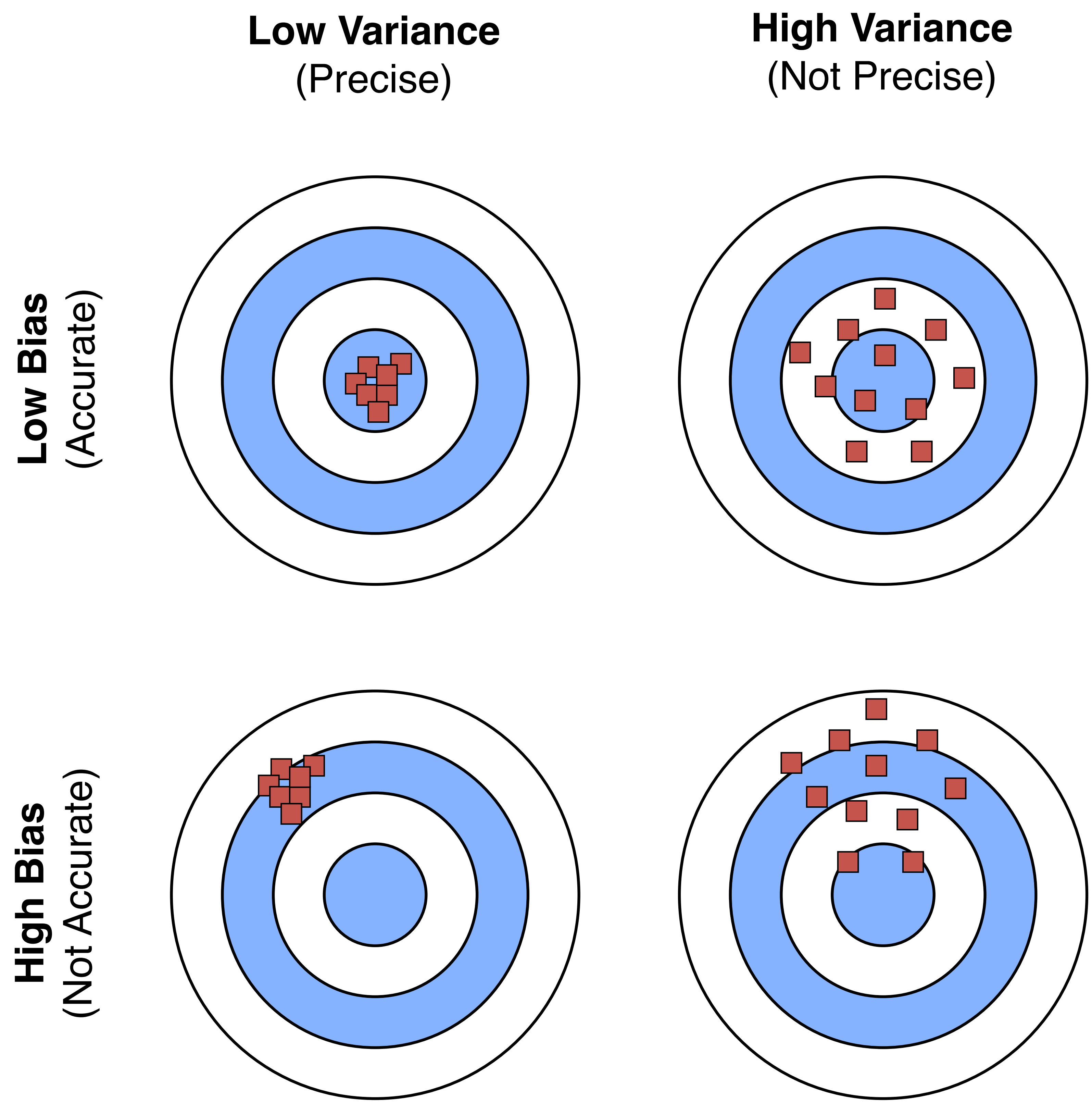
Bias-Variance Decomposition and Trade-off

Bias-Variance Decomposition

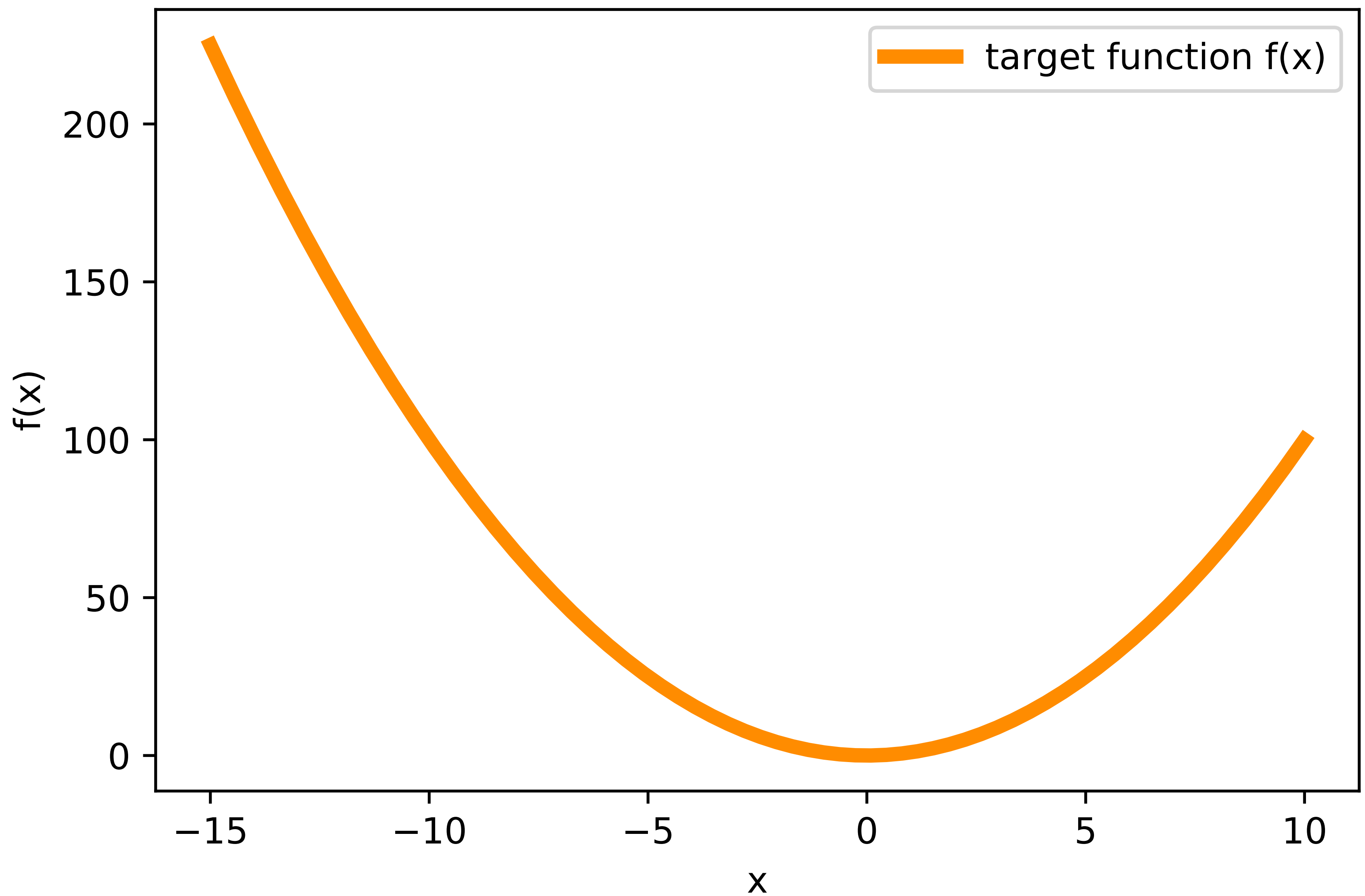
- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are correlated to underfitting and overfitting

- Helps explain why ensemble methods (last lecture) might perform better than single models

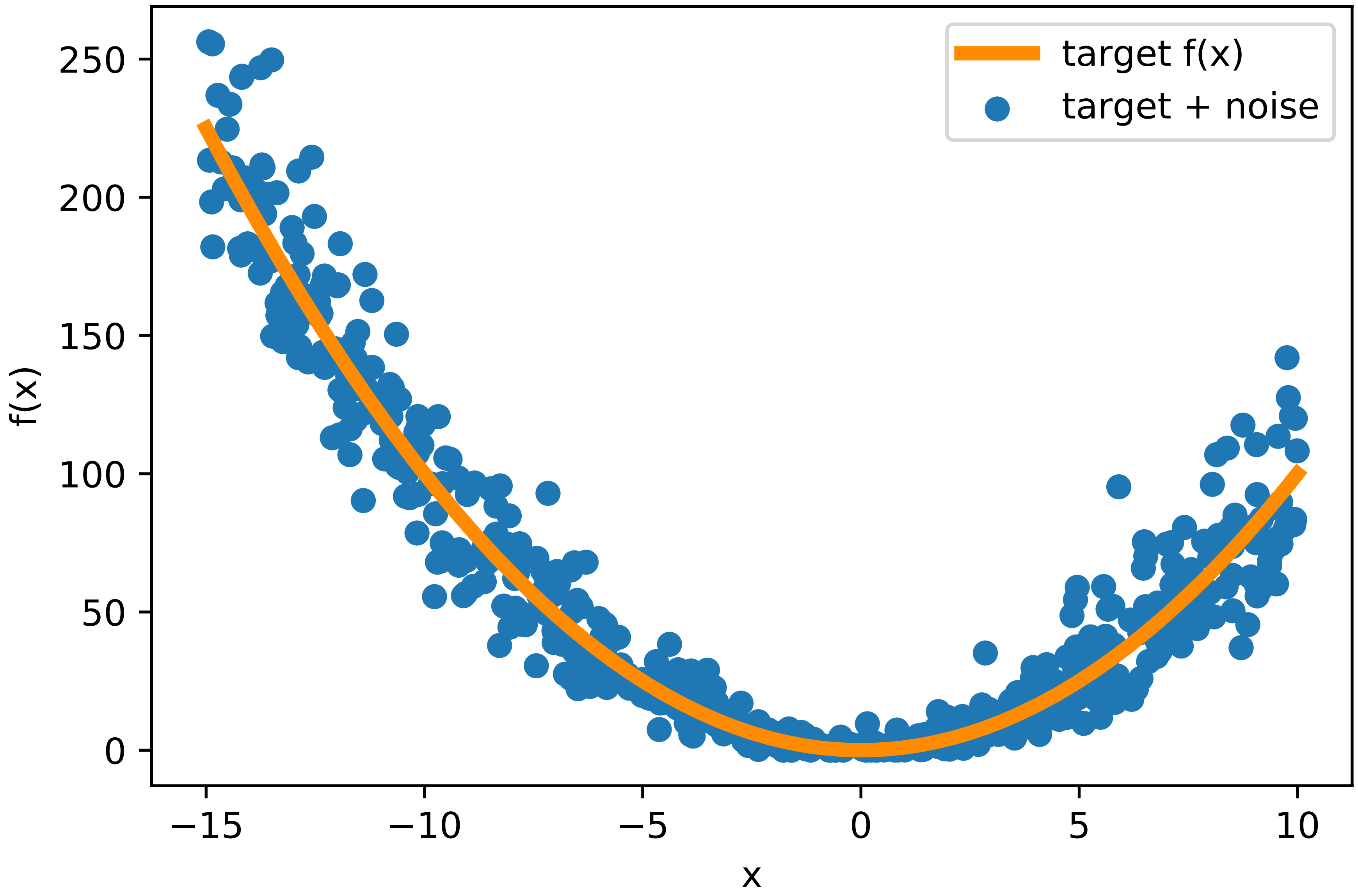
Bias-Variance Intuition



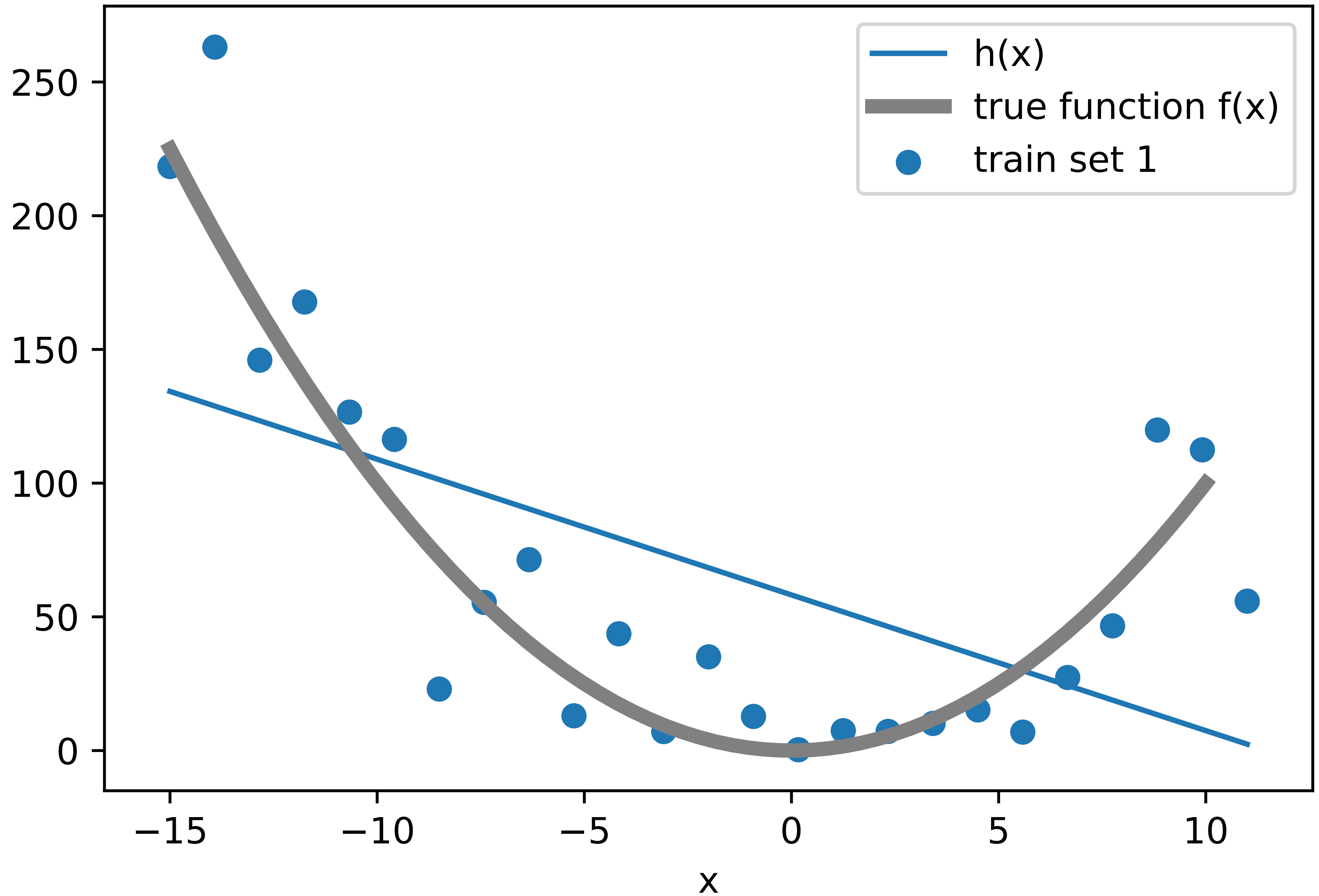
Bias and Variance Intuition



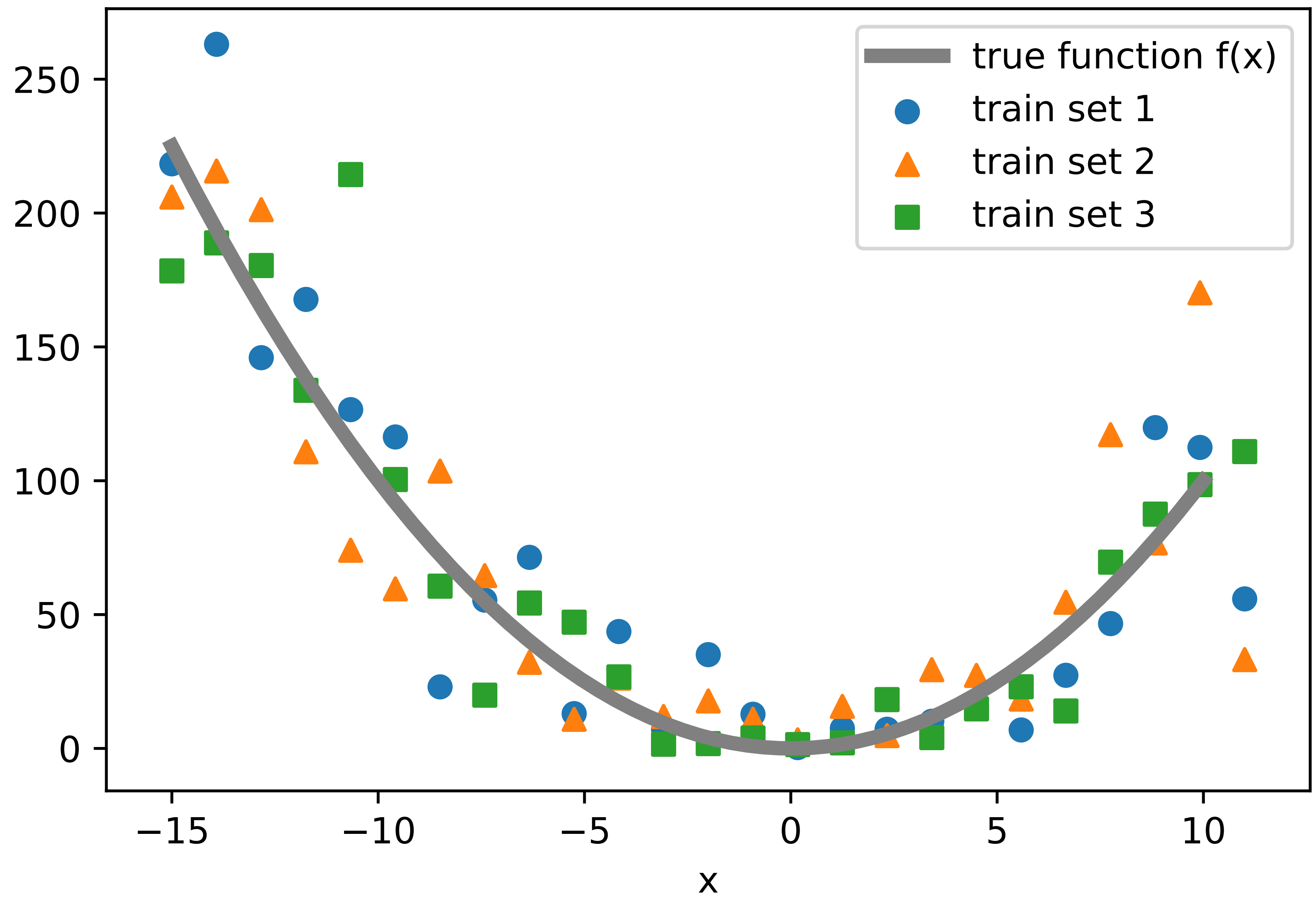
Bias and Variance Intuition



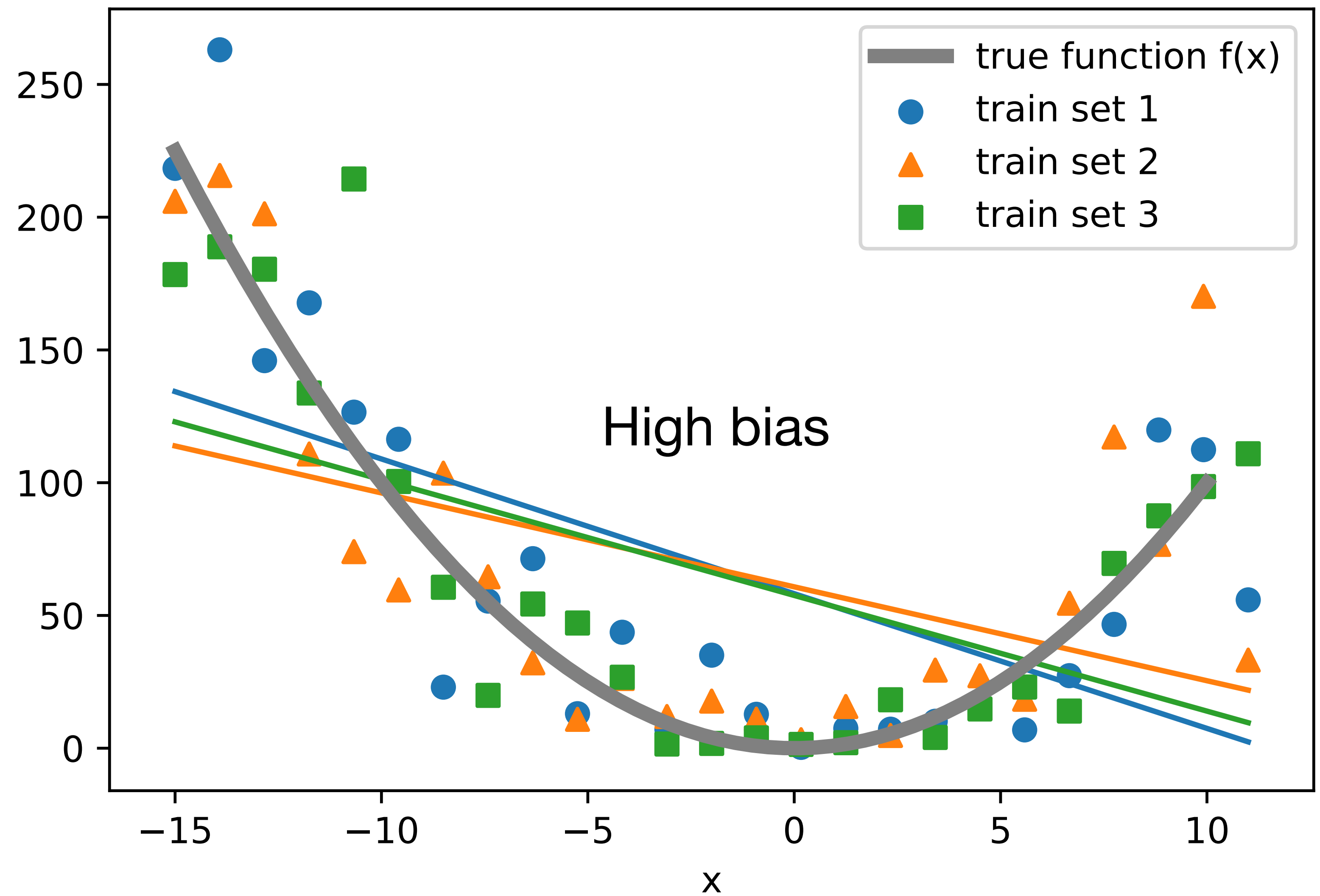
Bias and Variance Intuition



Bias and Variance Intuition

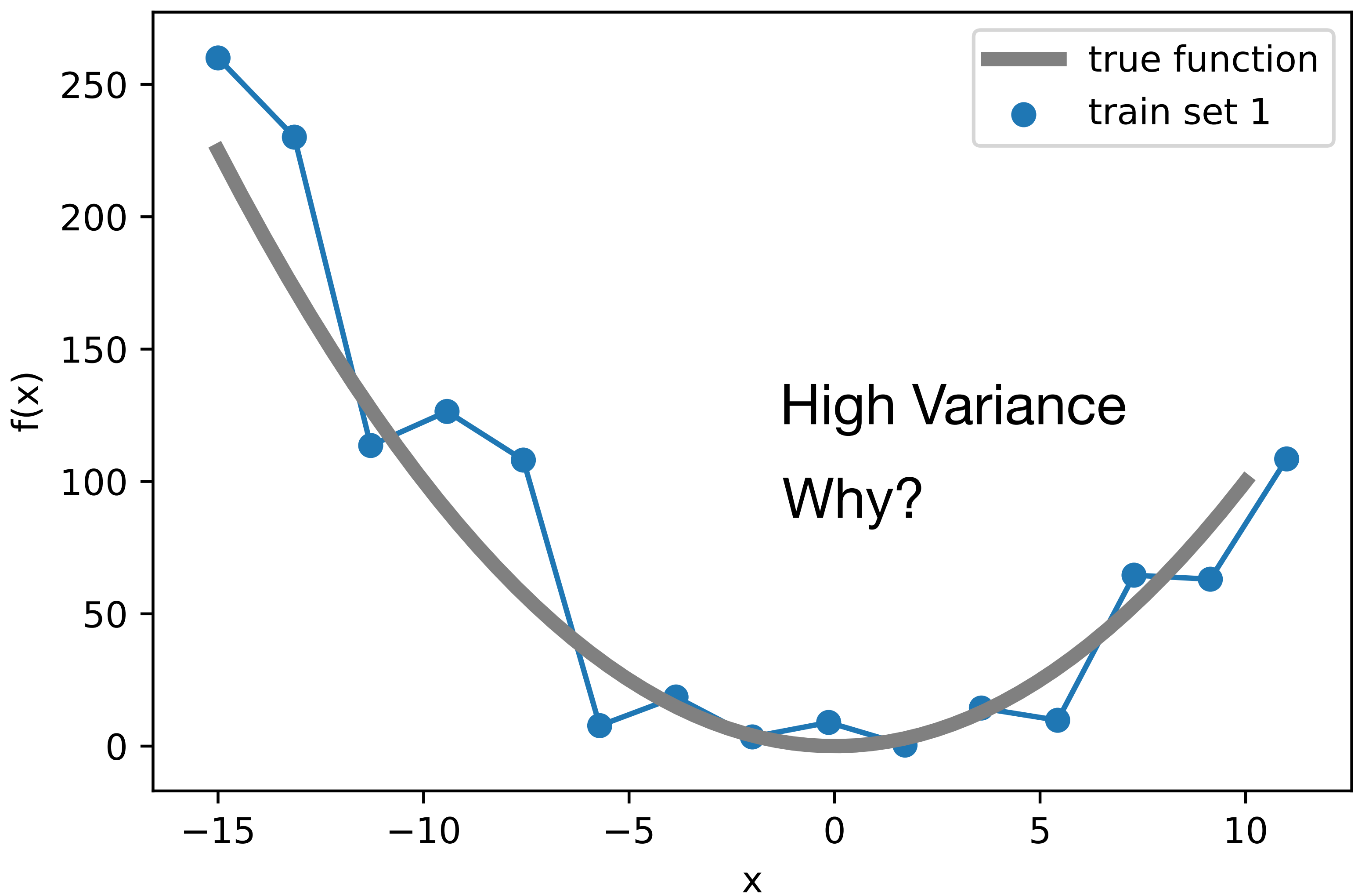


Bias and Variance Intuition



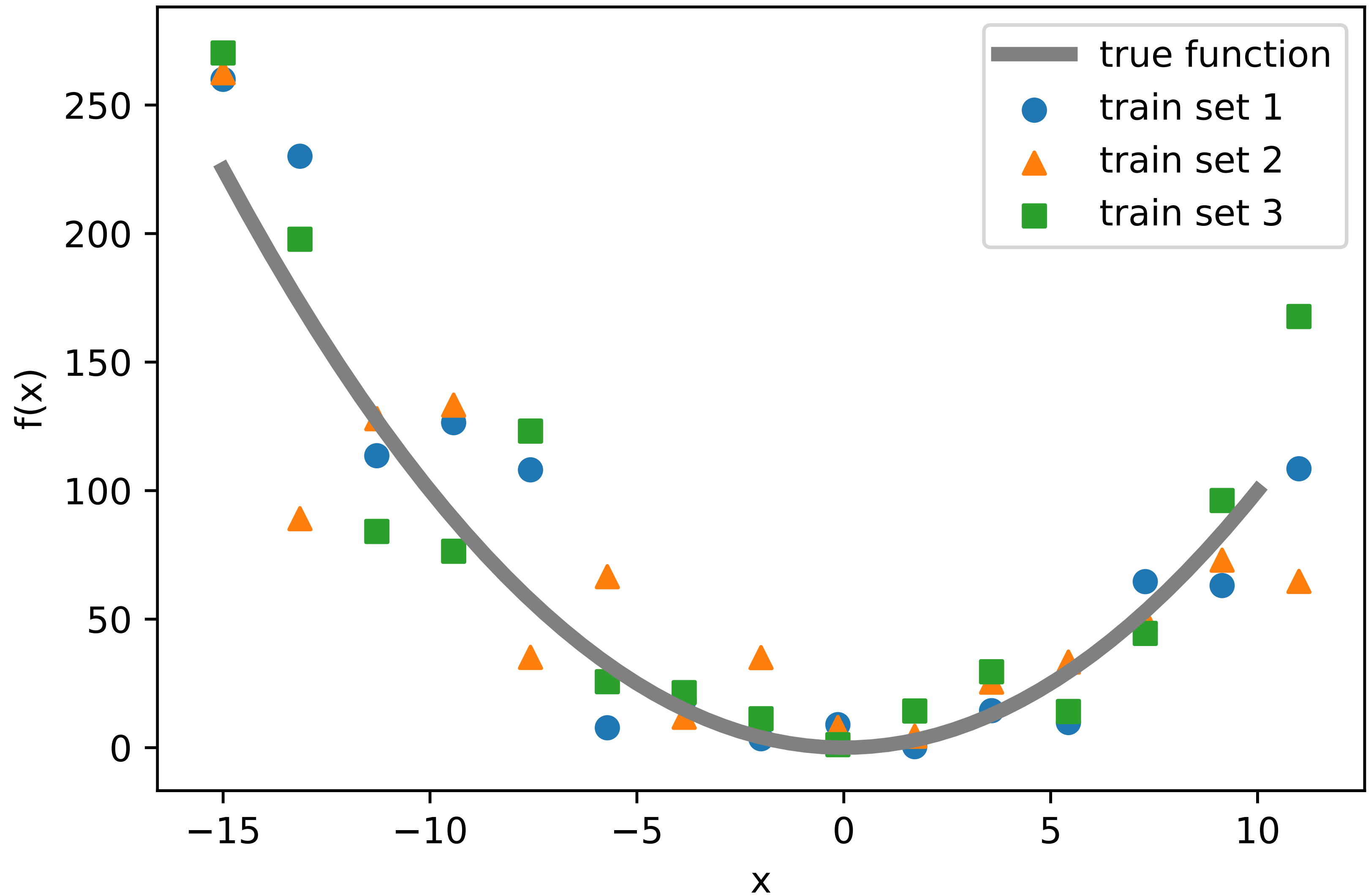
(There are two points where the bias is zero)

Bias and Variance Intuition



(here, I fit an unpruned decision tree)

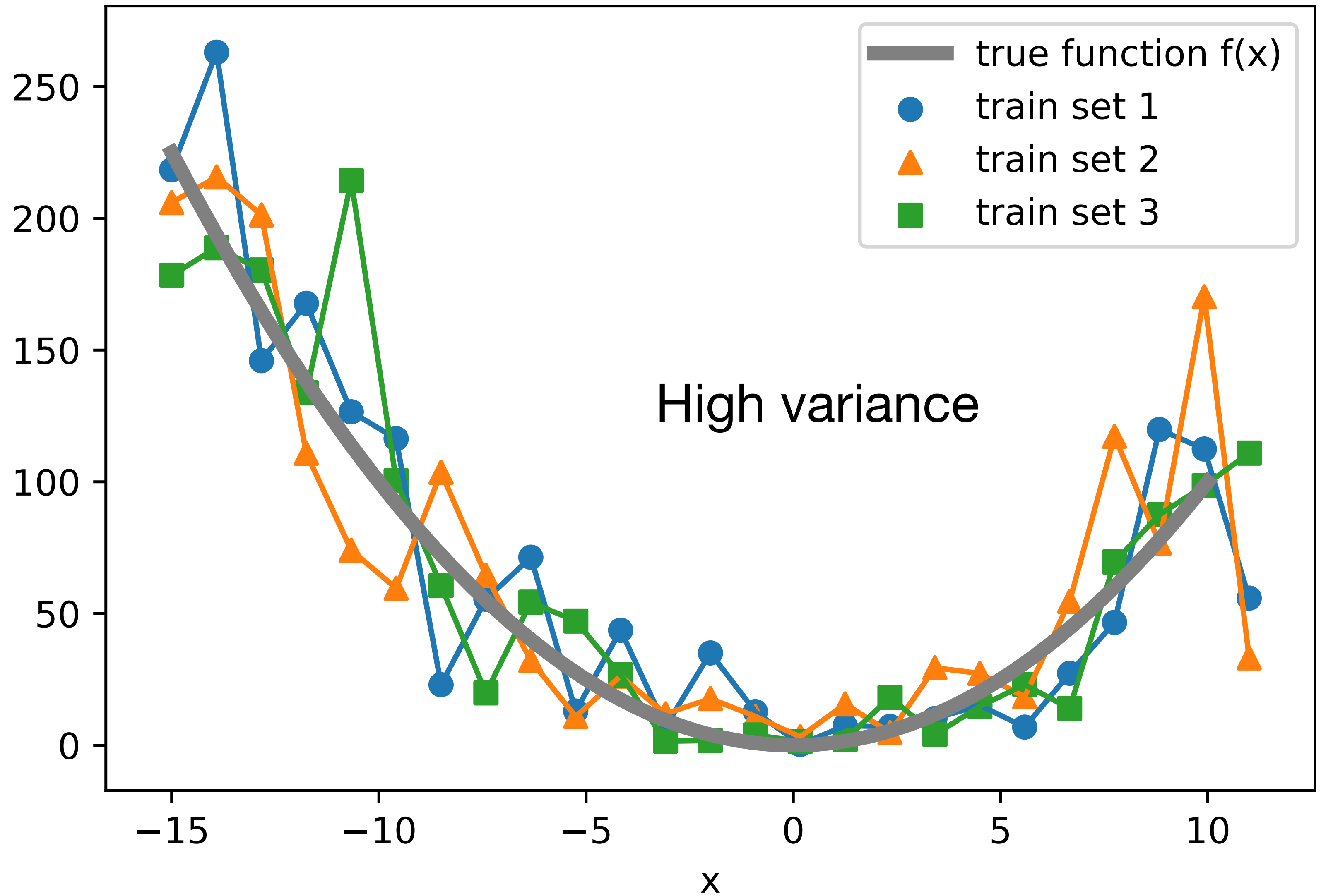
Bias and Variance Example



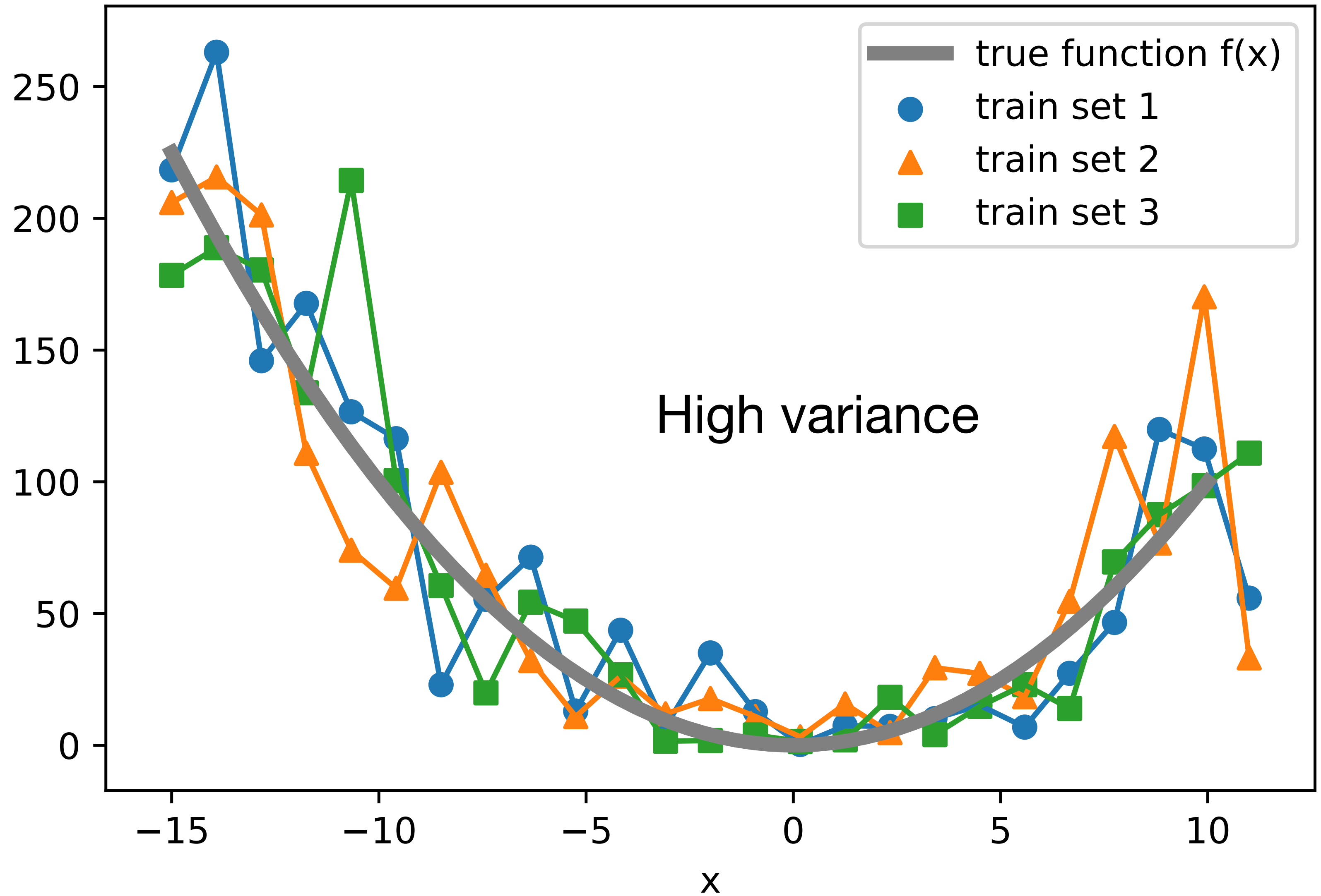
where f(x) is some true (target) function

suppose we have multiple training sets

Bias and Variance Example



Bias and Variance Example



What happens if we take the average?
Does this remind you of something?

Terminology

Point estimator $\hat{\theta}$ of some parameter θ

(could also be a function, e.g., the hypothesis is an estimator of some target function)

Terminology

Point estimator $\hat{\theta}$ of some parameter θ

(could also be a function, e.g., the hypothesis is an estimator of some target function)

Bias $= E[\hat{\theta}] - \theta$

Bias-Variance Decomposition

General Definition:

$$\textbf{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

$$\textbf{Var}(\hat{\theta}) = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$$

$$\textbf{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]$$

Bias-Variance Decomposition

General Definition:

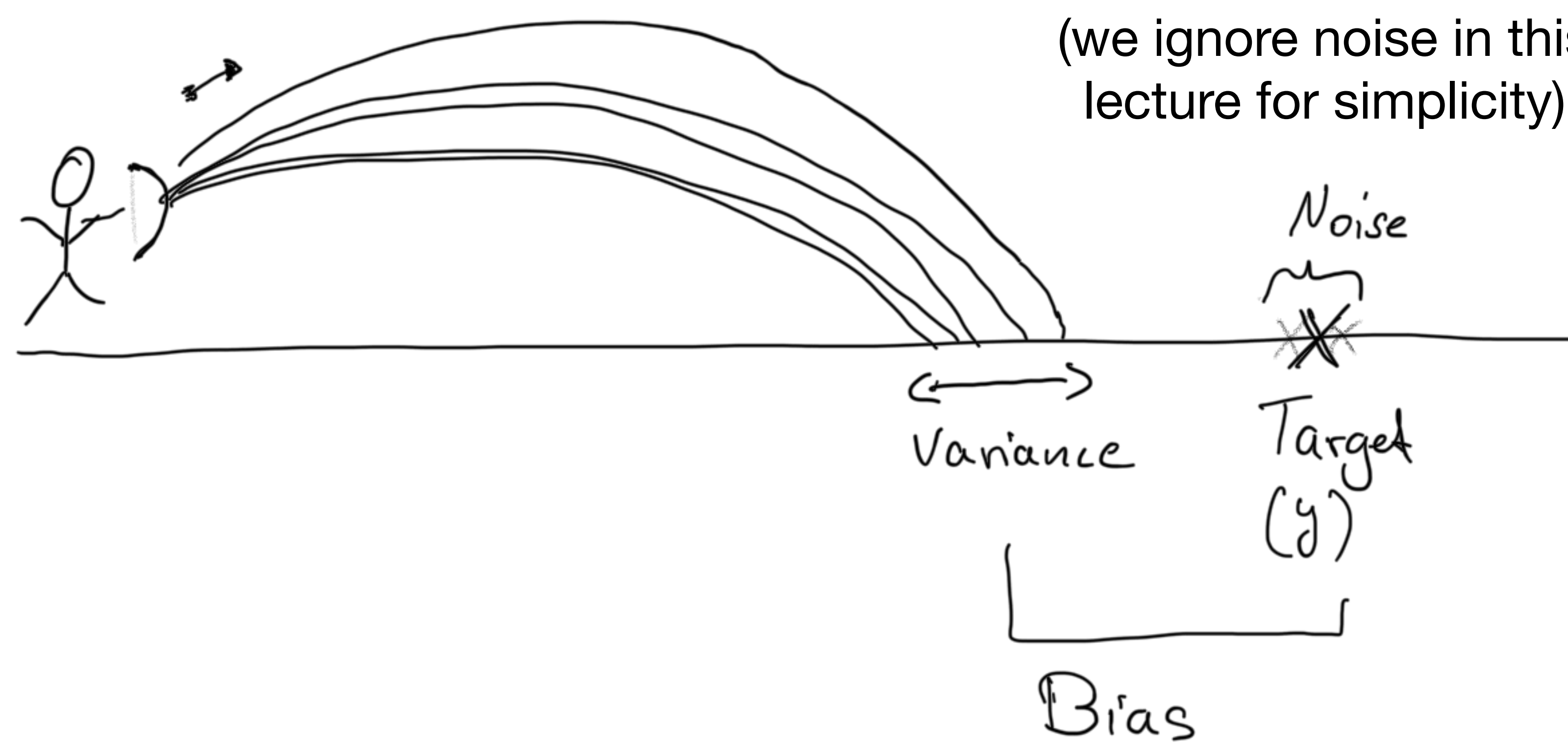
$$\textbf{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

$$\textbf{Var}(\hat{\theta}) = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$$

$$\textbf{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]$$

Intuition:

(we ignore noise in this lecture for simplicity)



Bias-Variance Decomposition of Squared Error

General Definition:

Intuition:

$$\textbf{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Bias is the difference between the average estimator from different training samples and the true value. (The expectation is over the training sets.)

$$\textbf{Var}(\hat{\theta}) = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$$

$$\textbf{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]$$

The variance provides an estimate of how much the estimate varies as we vary the training data (e.g,. by resampling).

Bias-Variance Decomposition

Loss = Bias + Variance + Noise

Bias-Variance Decomposition of Squared Error

General Definition:

$\textbf{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

$\textbf{Var}(\hat{\theta}) = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$

$\textbf{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]$

"ML notation" for the Squared Error Loss:

$y = f(x)$ (target, target function)

$\hat{y} = \hat{f}(x) = h(x)$

$S = (y - \hat{y})^2$

(For the sake of simplicity, we ignore the noise term in this lecture)

(Next slides: the expectation is over the training data, i.e, the average estimator from different training samples)

Bias-Variance Decomposition of Squared Error

"ML notation" for the Squared Error Loss:

$y = f(x)$ (target, target function)

$\hat{y} = \hat{f}(x) = h(x)$

(x is a particular data point e.g,. in the test set;
the expectation is over training sets)

$S = (y - \hat{y})^2$

$$S = (y - \hat{y})^2$$

$$(y - \hat{y})^2 = (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2$$

$$= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})$$

Bias-Variance Decomposition of Squared Error

$$S = (y - \hat{y})^2$$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})\end{aligned}$$

$$\begin{aligned}E[S] &= E[(y - \hat{y})^2] \\ E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] \\ &= \textbf{[Bias of the fit]}^2 + \textbf{Variance of the fit}\end{aligned}$$

(The expectation is over the training data, i.e, the average estimator from different training samples)

Bias-Variance Decomposition of Squared Error

$$S = (y - \hat{y})^2$$

???

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + \boxed{2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}\end{aligned}$$

$$\begin{aligned}E[S] &= E[(y - \hat{y})^2] \\ E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] \\ &= \textbf{[Bias]}^2 + \textbf{Variance}\end{aligned}$$

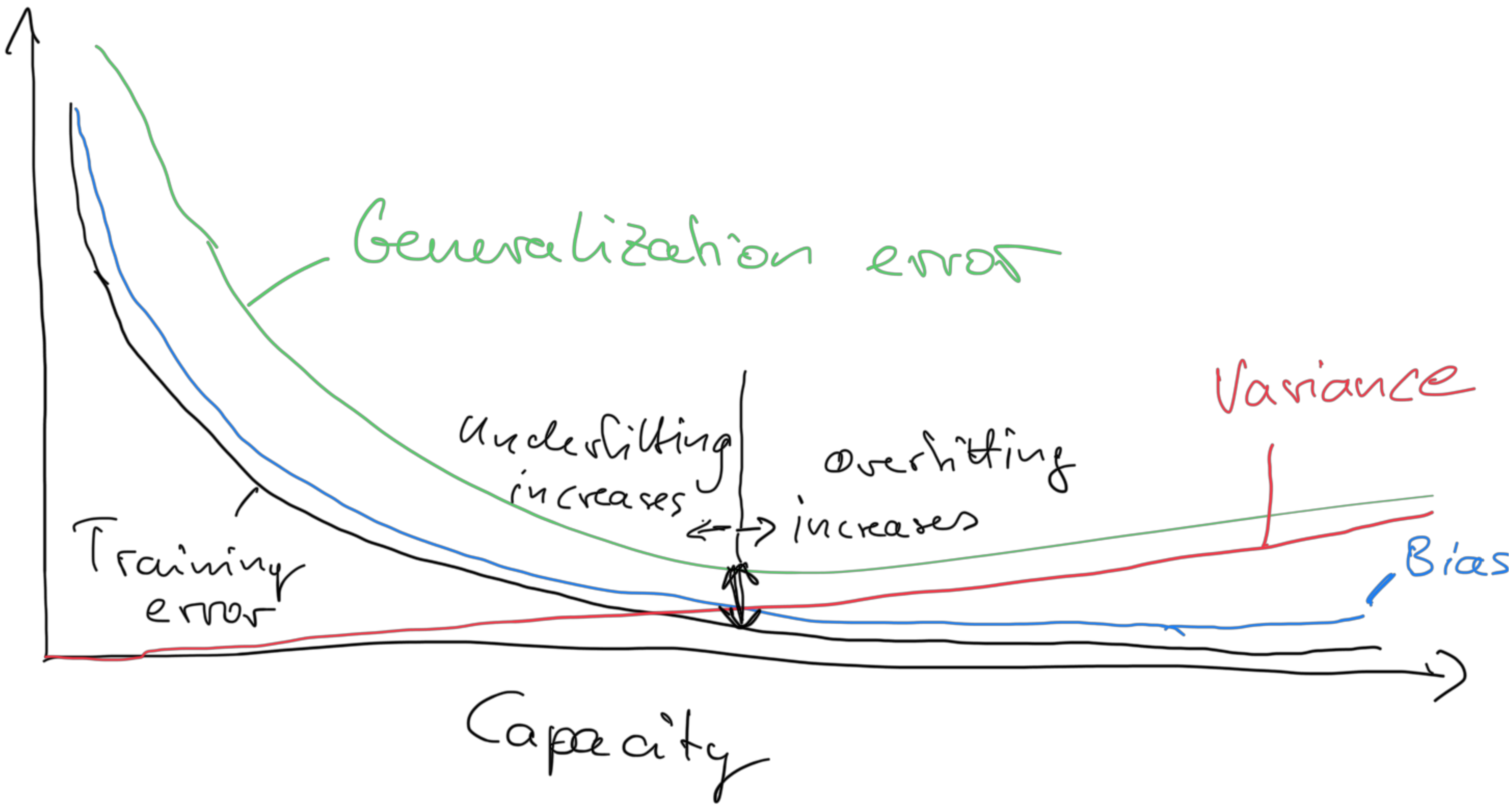
Bias-Variance Decomposition of Squared Error

$$S = (y - \hat{y})^2$$

???

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + \boxed{2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}\end{aligned}$$

$$\begin{aligned}E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] &= 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] \\ &= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})] \\ &= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}]) \\ &= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}]) \\ &= 0\end{aligned}$$


Generalization error
Variance
underfitting increases
overfitting increases
Bias
Training error
Capacity

Domingos, P. (2000). A unified bias-variance decomposition.
In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

"several authors have proposed bias-variance decompositions related to zero-one loss (Kong & Dietterich, 1995; Breiman, 1996b; Kohavi & Wolpert, 1996; Tibshirani, 1996; Friedman, 1997). However, each of these decompositions has significant shortcomings."

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

Squared Loss

$$(y - \hat{y})^2$$

$$E[(y - \hat{y})^2]$$

Generalized Loss

$$L(y, \hat{y})$$

$$E[L(y, \hat{y})]$$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

Squared Loss

$$(y - \hat{y})^2$$

$$E[(y - \hat{y})^2]$$

$$E[(y - \hat{y})^2] = (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2]$$

Bias^2 + Variance

Generalized Loss

$$L(y, \hat{y})$$

$$E[L(y, \hat{y})]$$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

Squared Loss	Generalized Loss
$(y - \hat{y})^2$	$L(y, \hat{y})$
$E[(y - \hat{y})^2]$	$E[L(y, \hat{y})]$

$$E[(y - \hat{y})^2] = \underbrace{(y - E[\hat{y}])^2}_{\text{Bias}^2} + \underbrace{E[(E[\hat{y}] - \hat{y})^2]}_{\text{Variance}}$$

	Squared Loss	Generalized Loss
Bias^2:	$(y - E[\hat{y}])^2$	$L(y, E[\hat{y}])$
Variance:	$E[(E[\hat{y}] - \hat{y})^2]$	$E[L(\hat{y}, E[\hat{y}])]$

Define "Main Prediction"

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

The main prediction is the prediction that minimizes the average loss

$$\bar{\hat{y}} = \underset{\hat{y}'}{\text{argmin}} \; E[L(\hat{y}, \hat{y}')]$$

For squared loss -> Mean

For 0-1 loss -> Mode

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

Squared Loss

$$(y - \hat{y})^2$$

$$E[(y - \hat{y})^2]$$

$$E[(y - \hat{y})^2] = (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2]$$

Bias^2 + Variance

Main prediction -> Mean

Bias^2: $(y - \boxed{E[\hat{y}]})^2$

Variance: $E[(E[\hat{y}] - \hat{y})^2]$

0-1 Loss

$$L(y, \hat{y})$$

$$E[L(y, \hat{y})]$$

Main prediction -> Mode

$$L(y, \boxed{E[\hat{y}]})$$

$$E[L(\hat{y}, E[\hat{y}])]$$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

Squared Loss

$$E[(y - \hat{y})^2]$$

Main prediction -> Mean

Bias²: $(y - \boxed{E[\hat{y}]})^2$

Variance: $E[(E[\hat{y}] - \hat{y})^2]$

0-1 Loss

$$E[L(y, \hat{y})]$$

$$P(y \neq \hat{y})$$

Main prediction -> Mode

$$L(y, \boxed{E[\hat{y}])}$$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

$$E[L(\hat{y}, E[\hat{y}])]$$

$$Variance = P(\hat{y} \neq \hat{\bar{y}})$$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = Bias + Variance = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

Loss = Variance = $P(\hat{y} \neq y)$

$$Variance = P(\hat{y} \neq \bar{\hat{y}})$$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = Bias + Variance = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \text{ if } y \neq \bar{\hat{y}} \\ 0 \text{ otherwise} \end{cases}$$

Loss = Variance = $P(\hat{y} \neq y)$

$$Variance = P(\hat{y} \neq \bar{\hat{y}})$$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \text{ if } y \neq \bar{\hat{y}} \\ 0 \text{ otherwise} \end{cases}$$

Loss = $P(\hat{y} \neq y) = 1 - P(\hat{y} = y)$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

Loss = $P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq \bar{\hat{y}})$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

Loss = $P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq \bar{\hat{y}})$

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

Loss = $P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq \bar{\hat{y}})$

Loss = Bias - Variance

Bias-Variance Decomposition of 0-1 Loss

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

0-1 Loss

Loss = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

Variance can improve loss!!
Why is that so?

Loss = $P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq \bar{\hat{y}})$

Loss = Bias - Variance

Bias-Variance Simulation of C 4.5

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

- simulation on 200 training sets with 200 examples each (0-1 labels)
  - 200 hypotheses

- test set: 22,801 examples (1 data point for each grid point)

- mean error rate is 536 errors (out of the 22,801 test examples)
  - 297 as a result of bias
  - 239 as a result of variance

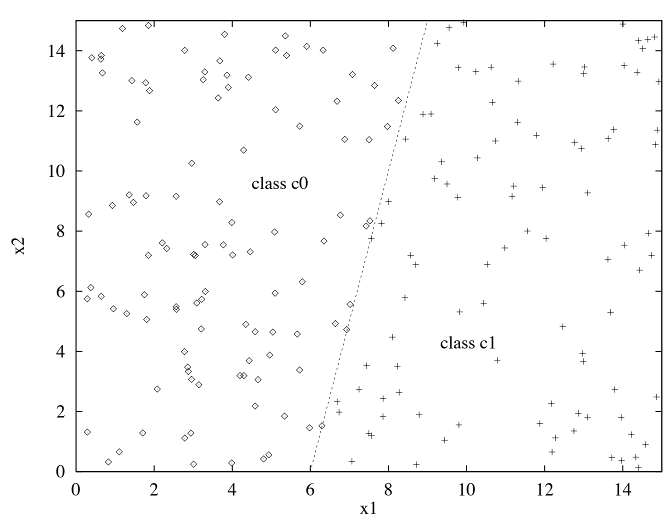


Figure 1: A two-class problem with 200 training examples.

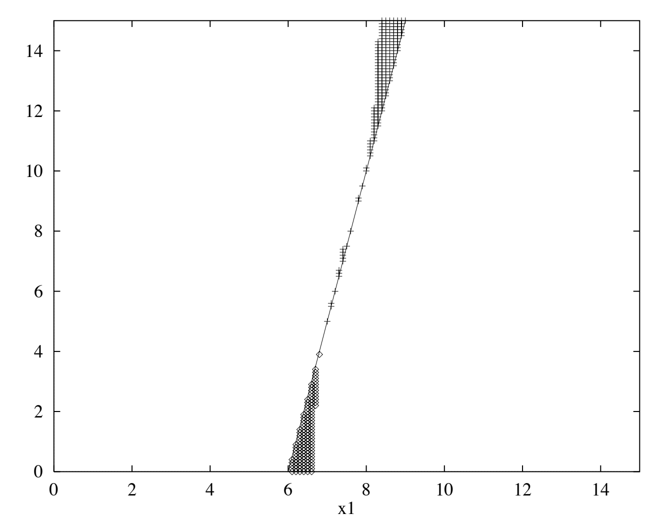


Figure 2: Bias errors of C4.5 on the problem from Figure 1.

(remember that trees use a "staircase" to approximate diagonal boundaries)

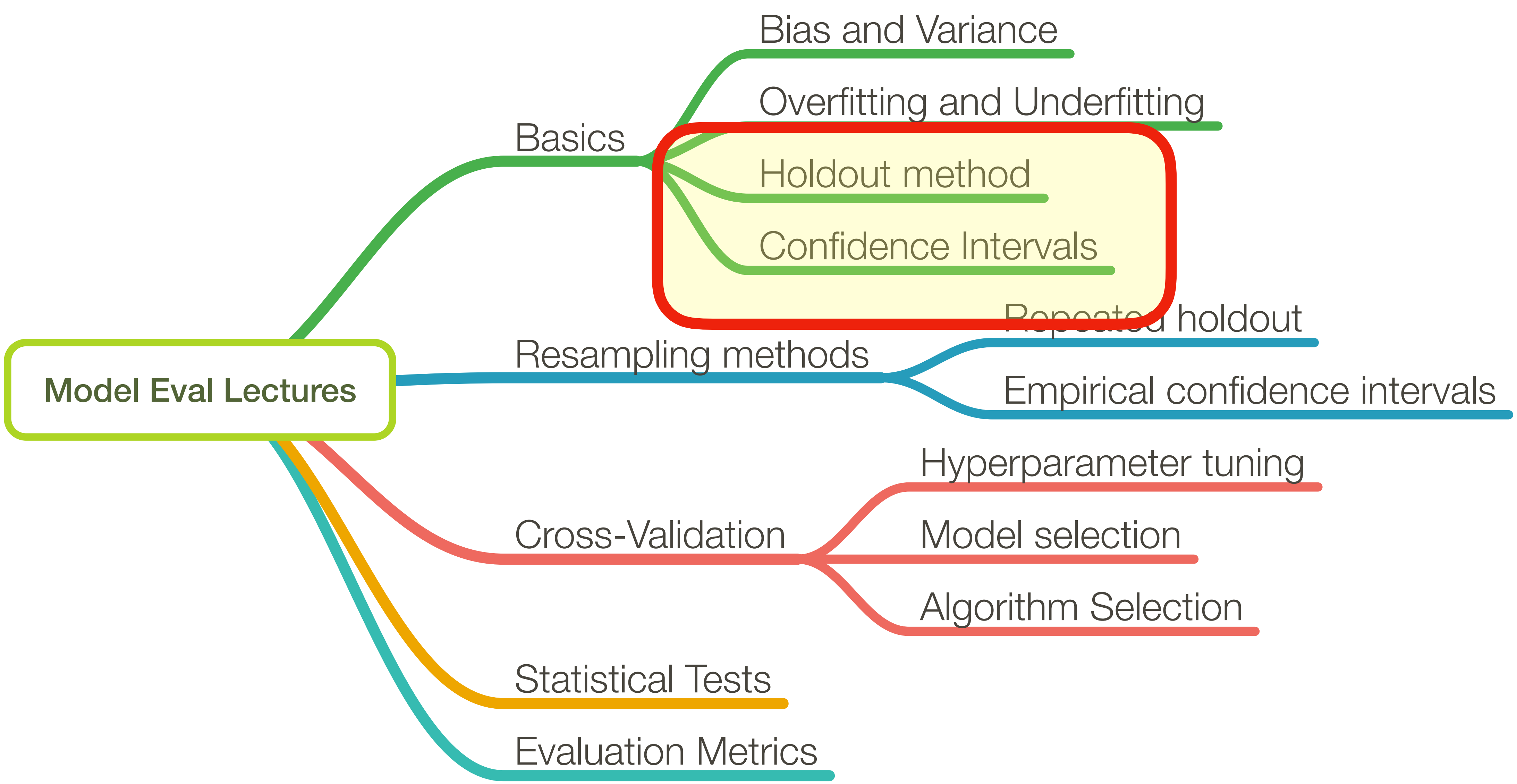
Recommended Reading Resources for Bias-Decomposition

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

0-1 loss

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

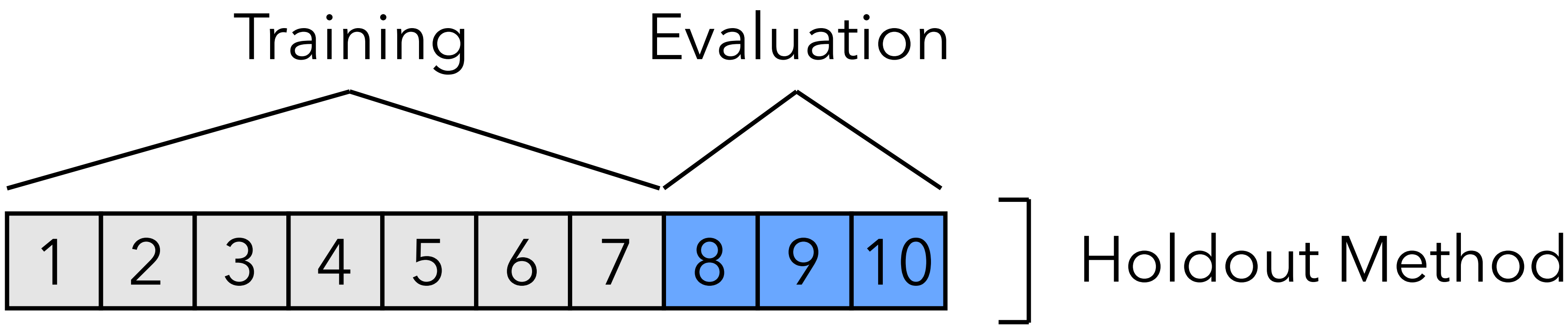
includes noise
and more general: Loss = Bias + *c* Variance


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Evaluation Metrics

- Training set error is an optimistically biased estimator of the generalization error

- Test set error is an unbiased estimator of the generalization error (test sample and hypothesis chosen independently)

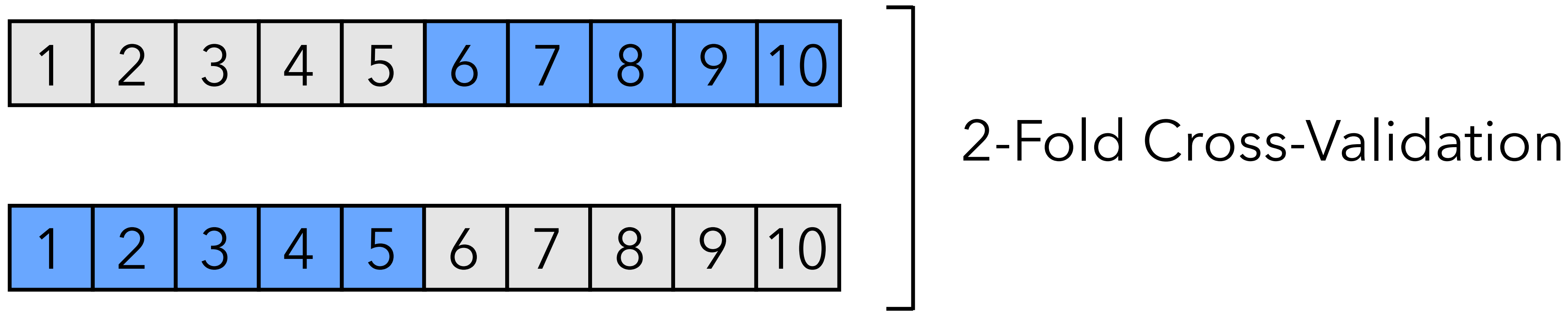
- (in practice, it is actually pessimistically biased; why?)


Training
Evaluation
1
2
3
4
5
6
7
8
9
10
Holdout Method

Often using the holdout method is not a good idea ...

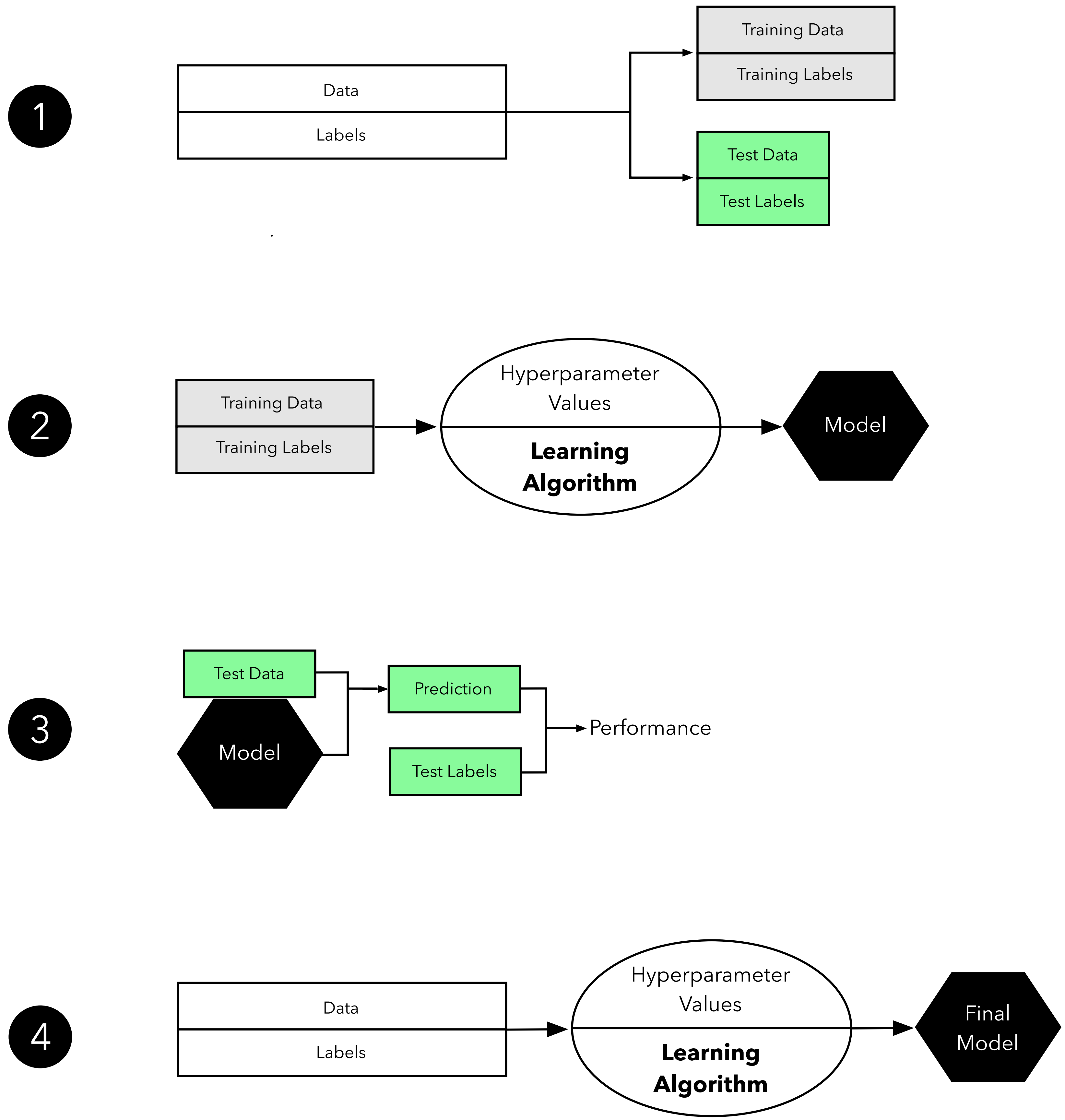
Often using the holdout method is not a good idea ...

- Pessimistically biased (not so bad)
- Does not account for variance in the training data (very bad)

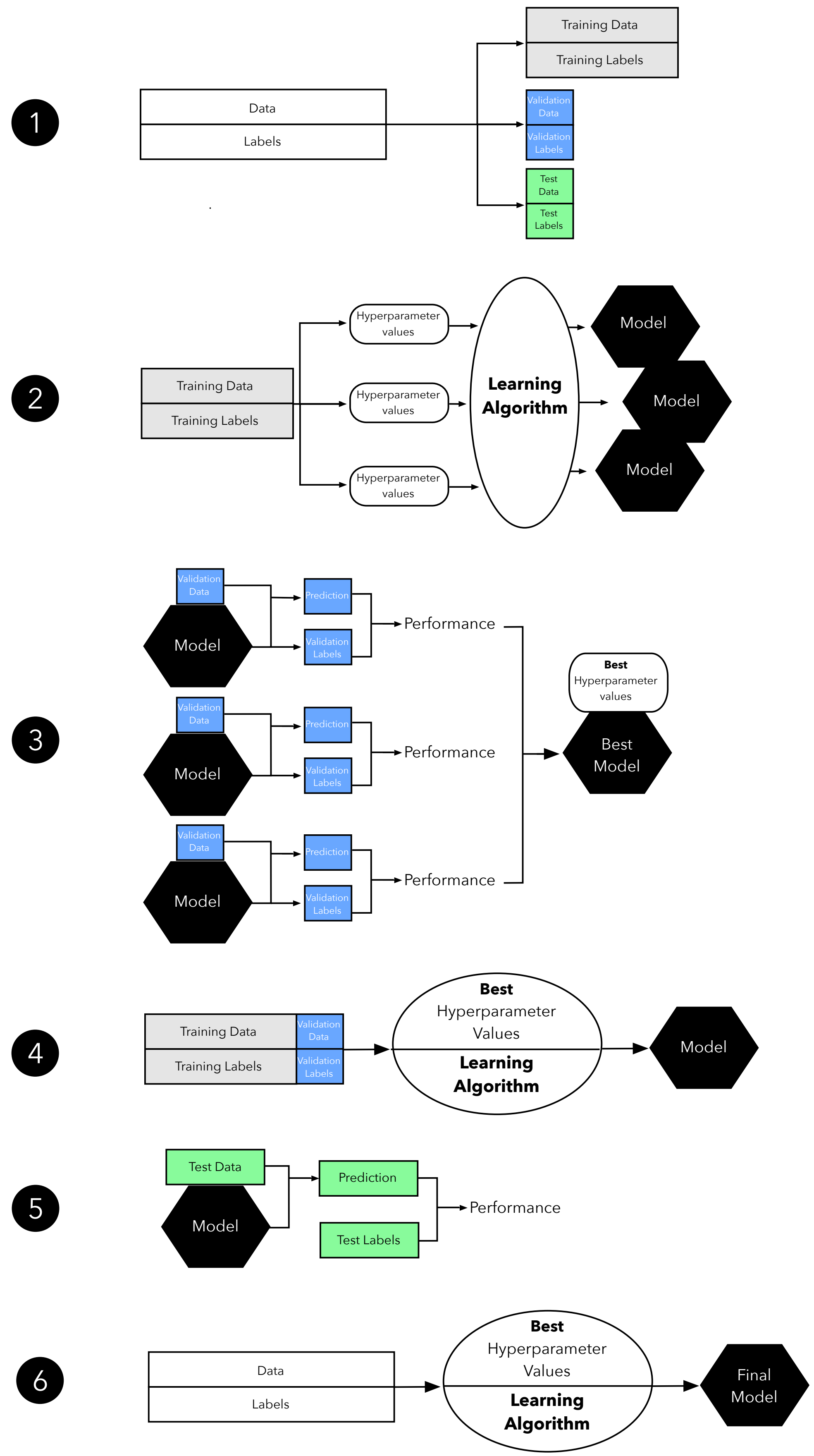


1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
2-Fold Cross-Validation

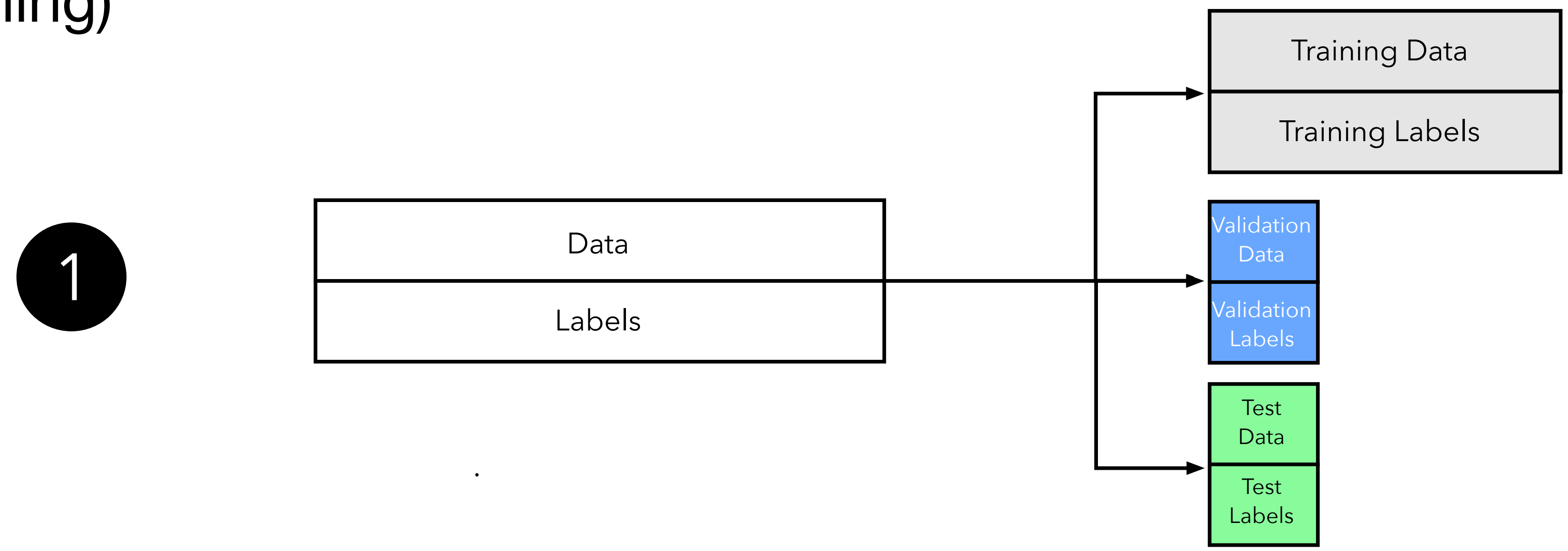
Holdout evaluation

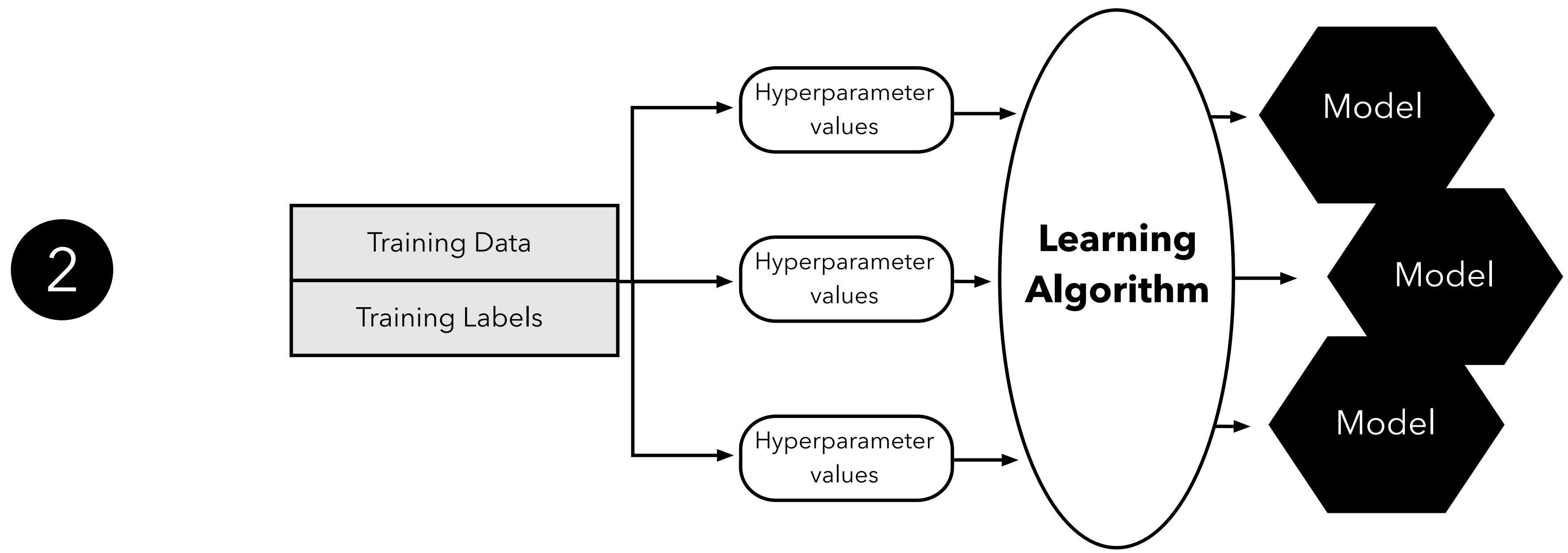


Holdout validation (hyperparam. tuning)

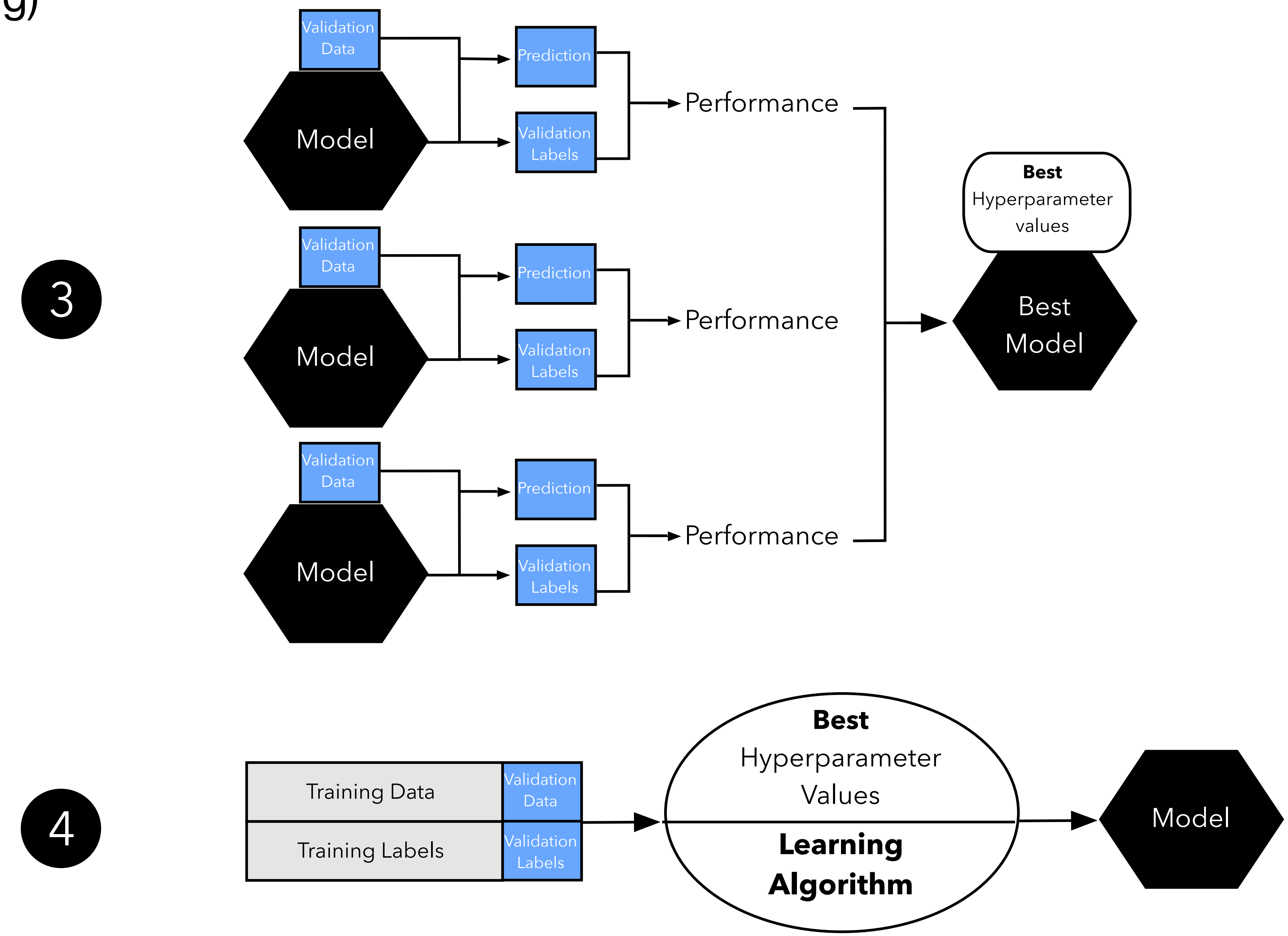


Holdout validation (hyperparam. tuning)



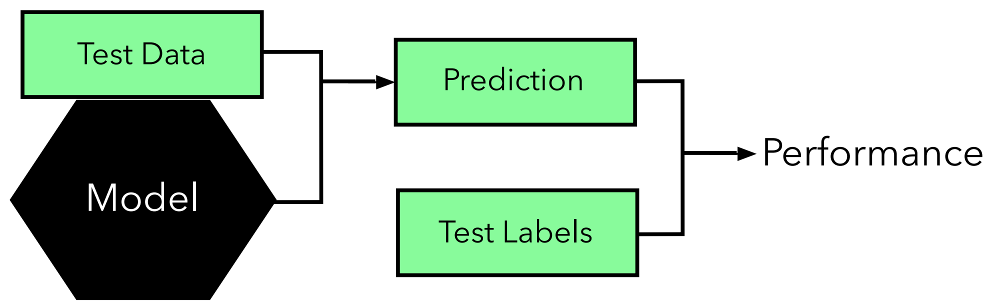


Holdout validation (hyperparam. tuning)

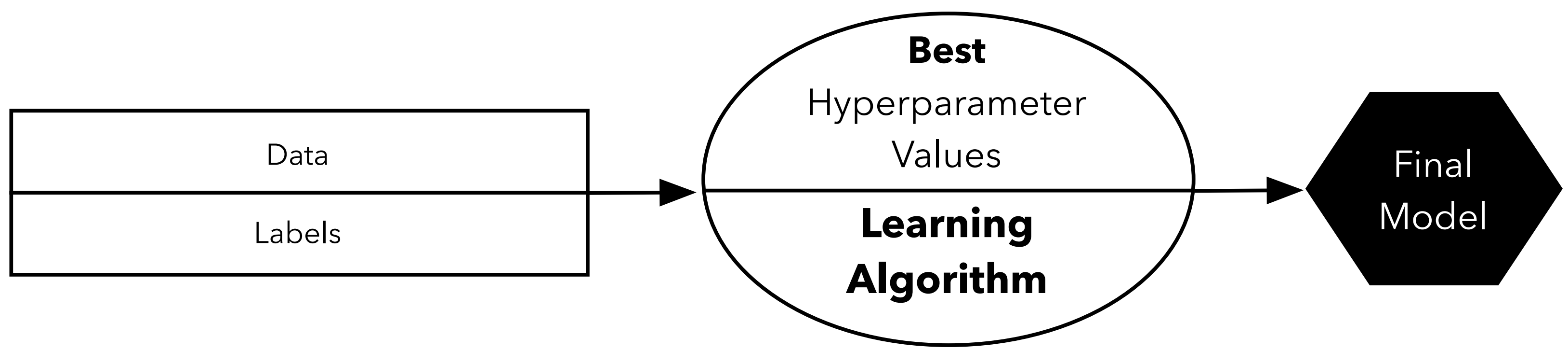


Holdout validation
(hyperparam. tuning)

5



6



Cross-Validation is generally better

... but ...

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, *5*(Sep), 1089-1105.

Bias of Estimators Example

Normal Distribution: $\mathcal{N}(\mu, \sigma^2)$

Probability density function: $f(x^{[i]}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x^{[i]} - \mu)^2}{\sigma^2} \right)$

Is the sample mean an unbiased estimator of the mean of the Gaussian?

$$\hat{\mu} = \frac{1}{n} \sum_i x^{[i]}$$

Bias of Estimators Example

Is the sample mean an unbiased estimator of the mean of the Gaussian?

$$\hat{\mu} = \frac{1}{n} \sum_i x^{[i]}$$

$$\begin{aligned} Bias(\hat{\mu}) &= E[\hat{\mu}] - \mu \\ &= E[\frac{1}{n} \sum_i x^{[i]}] - \mu \\ &= \frac{1}{n} \sum_i E[x^{[i]}] - \mu \\ &= \frac{1}{n} \sum_i \mu - \mu \\ &= \mu - \mu = 0 \end{aligned}$$

Bias of Estimators Example

Is the sample variance an unbiased estimator of the mean of the Gaussian

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x^{[i]} - \hat{\mu})^2$$

$$\begin{aligned} Bias(\hat{\sigma}^2) &= E[\hat{\sigma}^2] - \sigma^2 \\ &= E\left[\frac{1}{n} \sum_i \left(x^{[i]} - \hat{\mu}\right)^2\right] - \sigma^2 \\ &= \dots \\ &= \frac{m-1}{m}\sigma^2 - \sigma^2 \end{aligned}$$

Bias of Estimators Example

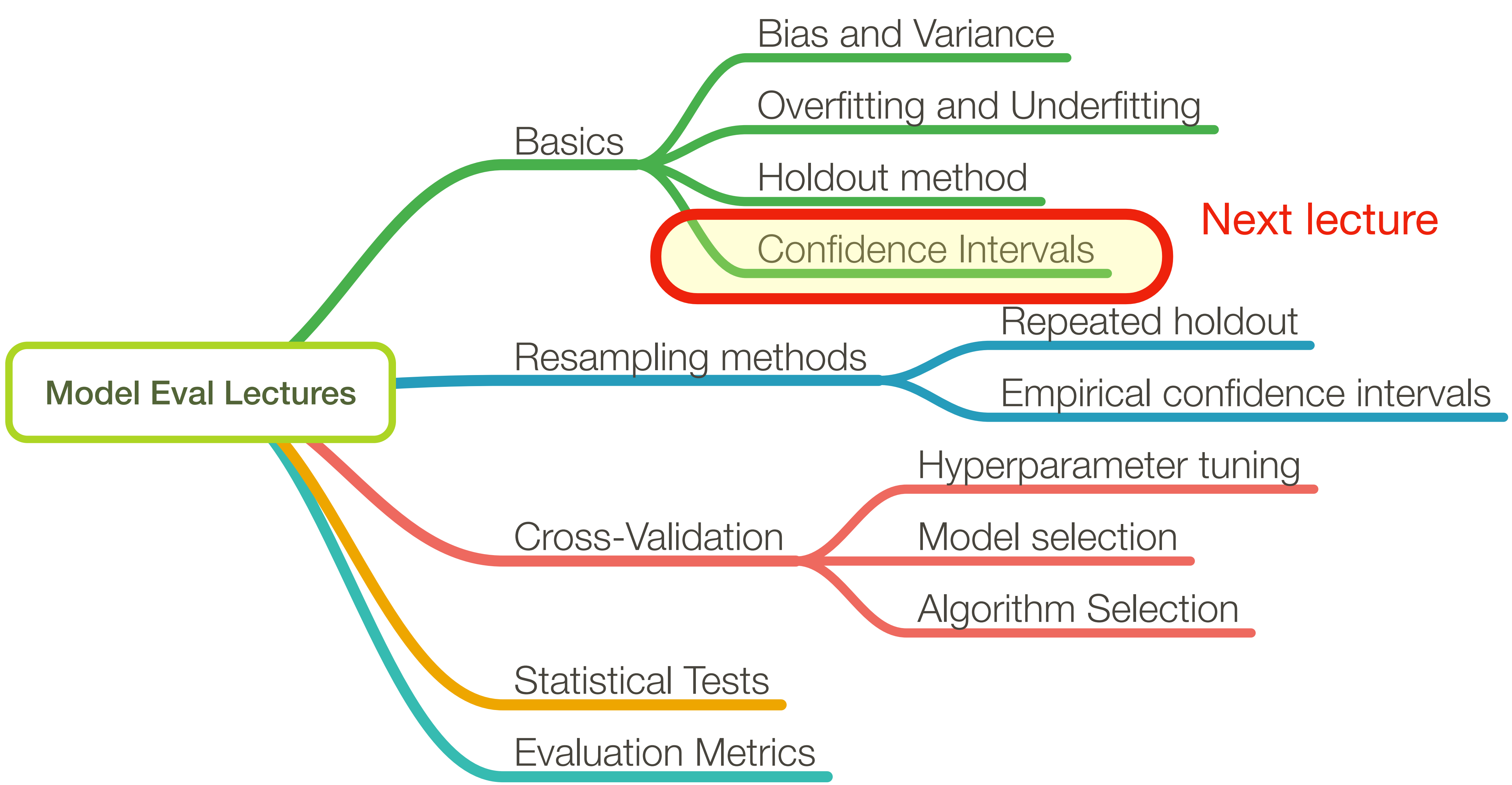
Is the sample variance an unbiased estimator of the mean of the Gaussian?

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x^{[i]} - \hat{\mu})^2$$

$$\begin{aligned} Bias(\hat{\sigma}^2) &= E[\hat{\sigma}^2] - \sigma^2 \\ &= E\left[\frac{1}{n} \sum_i \left(x^{[i]} - \hat{\mu}\right)^2\right] - \sigma^2 \\ &= \dots \\ &= \frac{m-1}{m} \sigma^2 - \sigma^2 \end{aligned}$$

The unbiased estimator is actually

$$\hat{\sigma}'^2 = \frac{1}{n-1} \sum_i (x^{[i]} - \hat{\mu})^2$$


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Next lecture
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Evaluation Metrics