## Definition according to Physics

“A set of systems (. . .) used in *statistical mechanics to describe a single system.”[a]

[a] John Daintith, ed. *A dictionary of physics*. 6th ed. Oxford ; New York: Oxford University Press, 2009. 616 pp. ISBN: 978-0-19-923399-1.

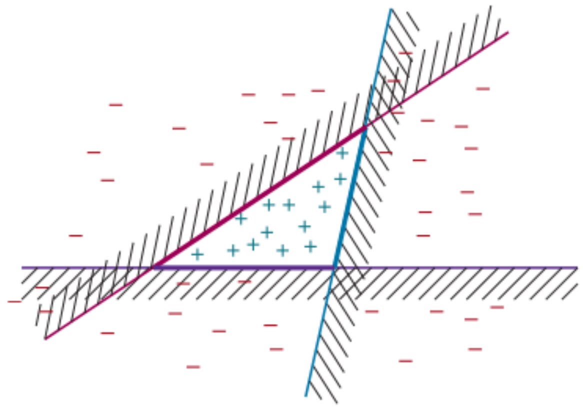
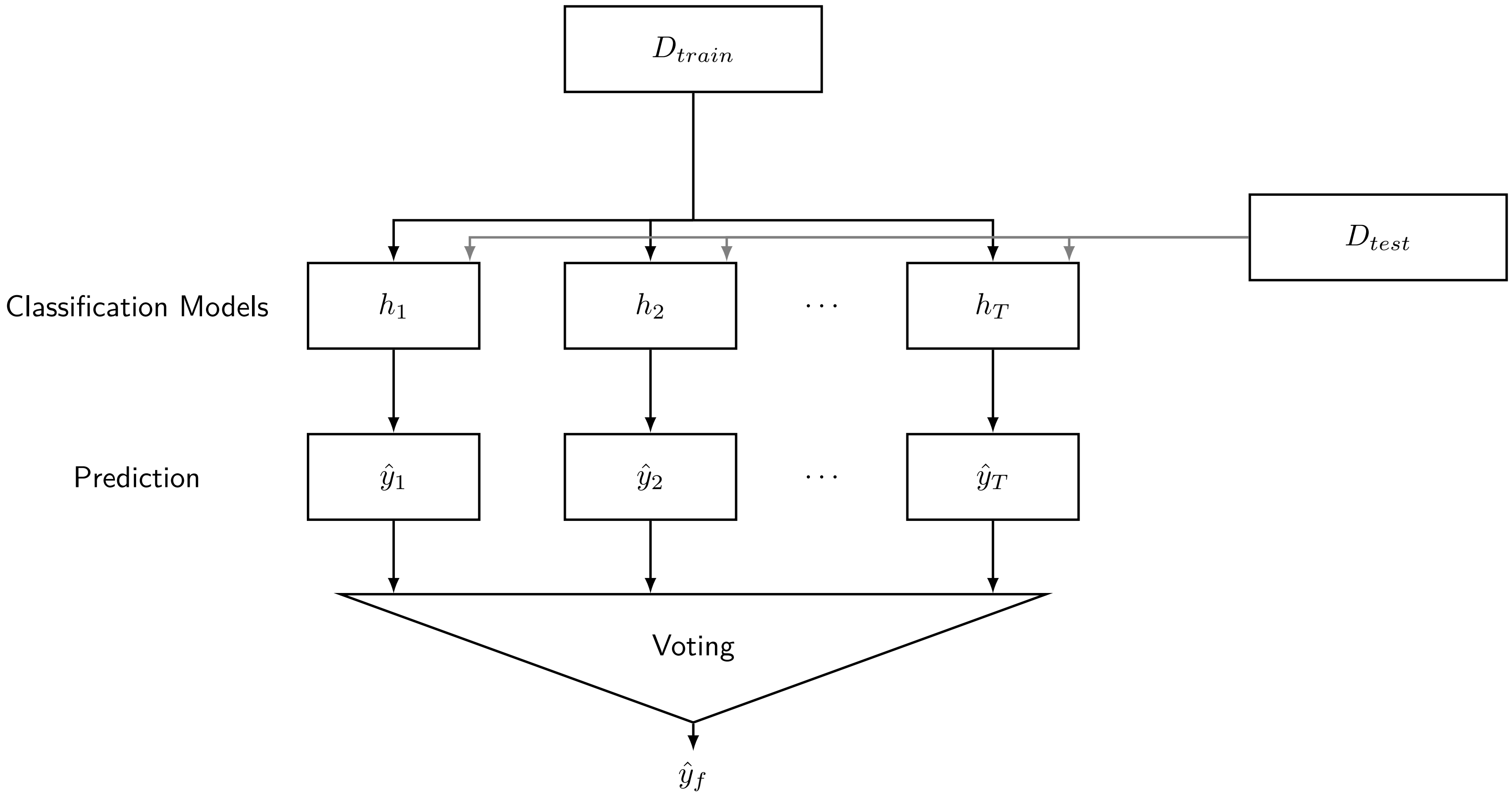


**Figure 19.23** Illustration of the increased expressive power obtained by ensemble learning. We take three linear threshold hypotheses, each of which classifies positively on the unshaded side, and classify as positive any example classified positively by all three. The resulting triangular region is a hypothesis not expressible in the original hypothesis space.

Figure: Motivation for Ensemble Methods. Taken from [RN21].

FH Salzburg



$D_{train}$
$D_{test}$
Classification Models
$h_1$
$h_2$
$\cdots$
$h_T$
Prediction
$\hat{y}_1$
$\hat{y}_2$
$\cdots$
$\hat{y}_T$
Voting
$\hat{y}_f$

# Stacking

## Intuition

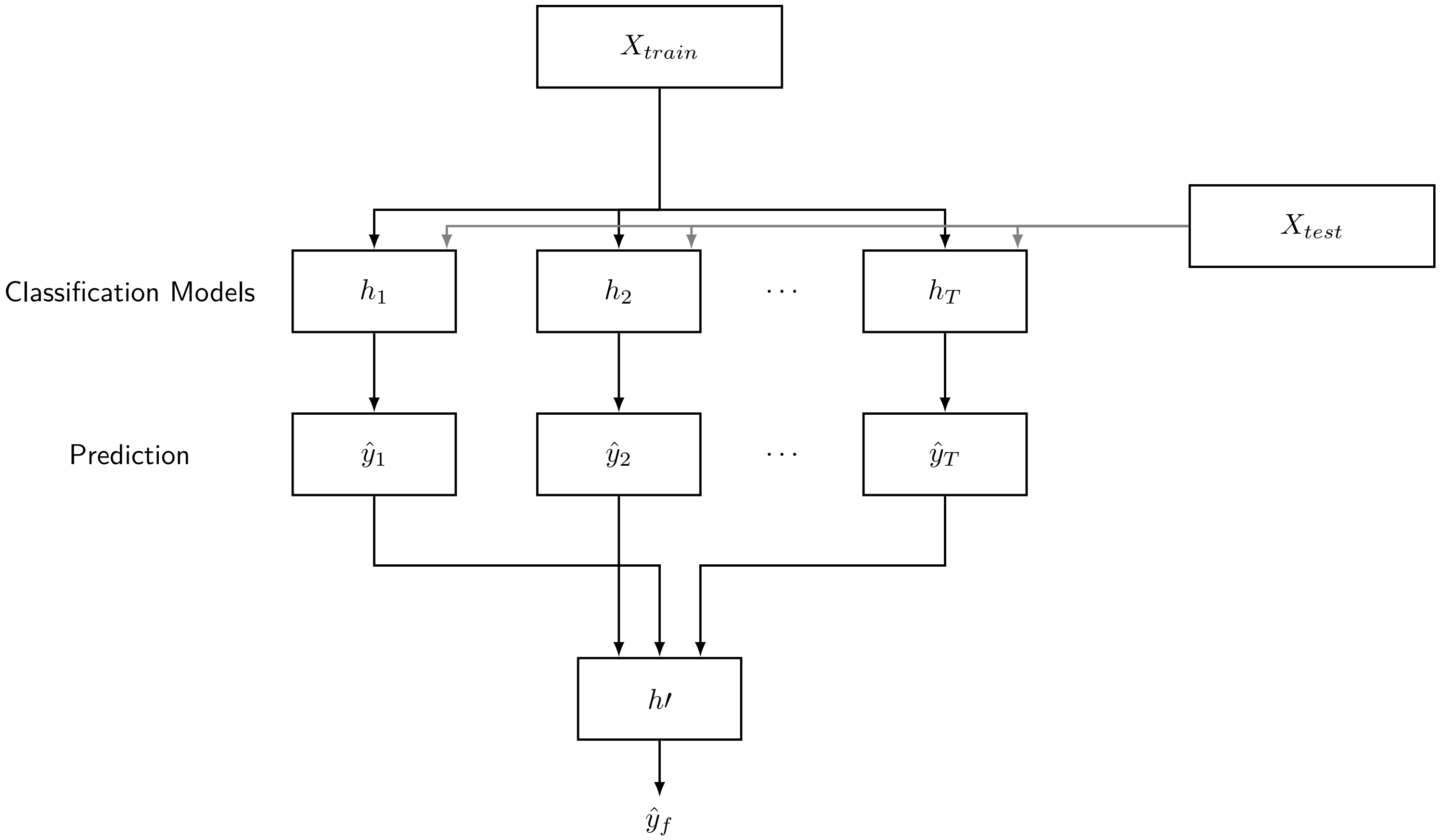
Learning on who to trust for a voting ensemble.[1]

[1] David H. Wolpert. “Stacked generalization”. In: *Neural Networks* 5.2 (Jan. 1, 1992), pp. 241–259. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80023-1.

FH Salzburg

# Stacking

## Stacked Classification



# Stacking

## Algortihm

**Algorithm 1:** Stacking

**Data:** $D = \{x_i, y_i\}_{i=1}^{n}$ $(x_i \in R^m, y_i \in \gamma)$

**Result:** Stacking Classifier $H$

\# step 1: learn first-level classifiers

**for** $t = 1$ *to* $T$ **do**

  Learn a base classifier $h_t$ based on $D$

\# step 2: Construct new dataset $D'$ from $D$

**for** $i = 1$ *to* $n$ **do**

  Construct a new dataset $D'$ that contains $\{x_i', y_i\}$,
  where $x_i' = \{h_1(x_i), h_2(x_i), \ldots, h_T(x_i)\}$

\# step 3: Learn a second-level classifier

Learn classifier $h'$ based on $D'$

**return** $H(x_i) = h'(h_1(x_i), h_2(x_i), \ldots, h_T(x_i))$

$n \ldots$ number of samples within the dataset

$m \ldots$ number of features within the dataset

$T \ldots$ number of first-level classifiers

FH Salzburg

Bagging - Boostrap Aggregating[2]

**Algorithm 3:** Bootstrap Aggregating

**for** $t = 1$ *to* $T$ **do**
- draw bootstrap sample $D^{(t)}$ from $D$ of size $n$
- train classifier $h_t$ on $D^{(t)}$

$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_T(x)\}$

$n \ldots$ size of the bootstrap sample

$D = \{x_i, y_i\}_{i=1}^{n}$ $(x_i \in R^m, y_i \in \gamma)$

FH Salzburg

[2] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00058655.

# Bagging - Bootrap Sampling I

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset $X$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Boostrap $X^{(1)}$ | 1 | 1 | 4 | 3 | 5 | 9 | 5 | 7 | 8 | 10 |
| Boostrap $X^{(2)}$ | 10 | 1 | 7 | 7 | 10 | 10 | 10 | 8 | 5 | 5 |
| Boostrap $X^{(3)}$ | 5 | 7 | 4 | 6 | 5 | 1 | 2 | 3 | 6 | 5 |

(Missing Labels)

| | | | | | |
|---|---|---|---|---|---|
| $X^{(1)}$ | 2 | 6 | | | |
| $X^{(2)}$ | 2 | 3 | 4 | 6 | 9 |
| $X^{(3)}$ | 8 | 9 | 10 | | |

FH Salzburg

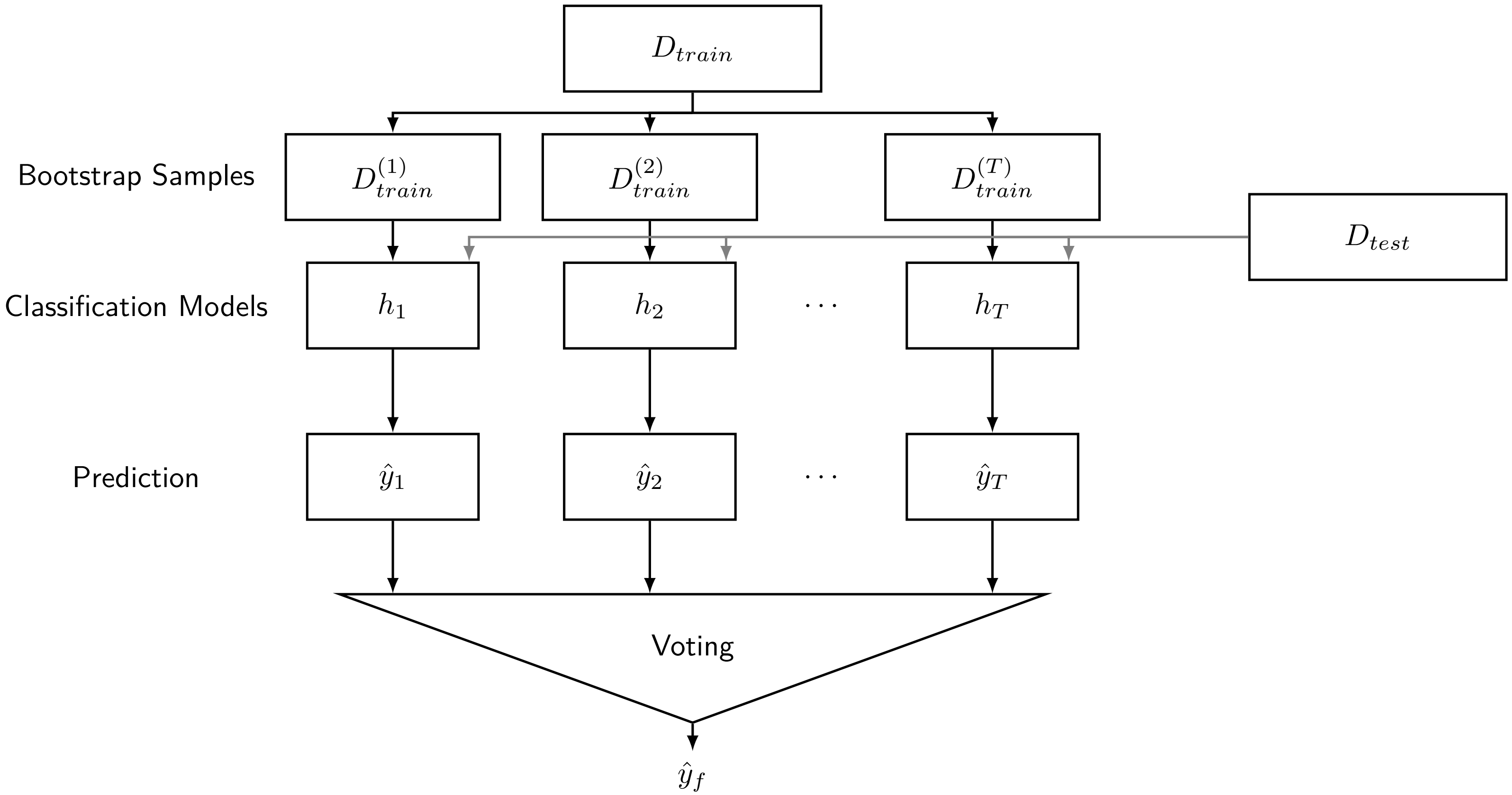
# Bagging - Bootrap Sampling II

| **Training example indices** | **Bootsrap 1** | **Bootsrap 2** | **...** | **Bootsrap $T$** |
|---|---|---|---|---|
| **1** | 2 | 7 | | 8 |
| **2** | 2 | 3 | | 9 |
| **3** | 1 | 2 | | 1 |
| **4** | 3 | 1 | | 9 |
| **5** | 7 | 1 | | 5 |
| **6** | 2 | 7 | | 6 |
| **7** | 4 | 7 | | 2 |

$h_1$ $h_2$ $h_T$

FH Salzburg

# Bagging - Bagging Classifier I



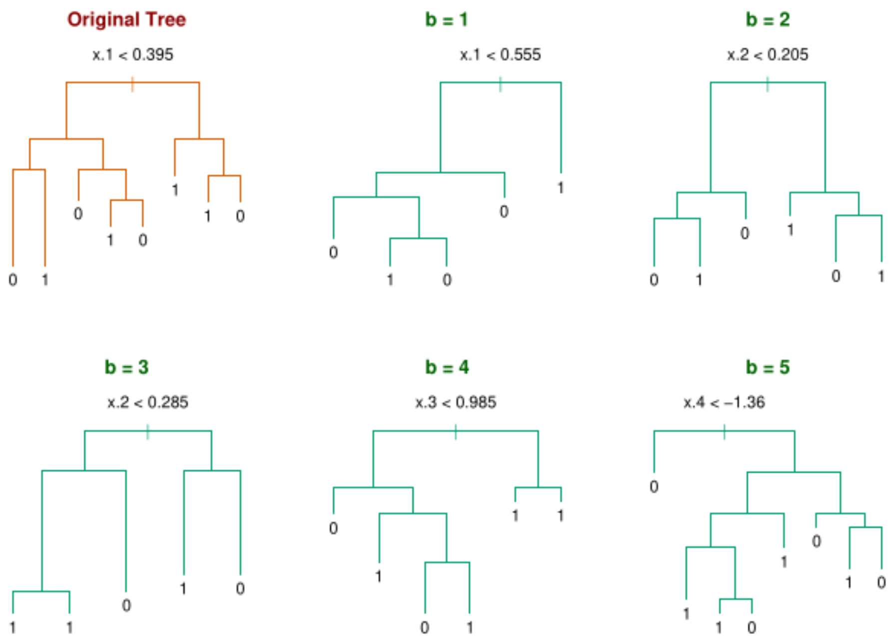



Figure: Comparisson of a single trained tree vs. an ensemble of several bootstrapped trees. Snippets taken from [HTF09].

FH Salzburg

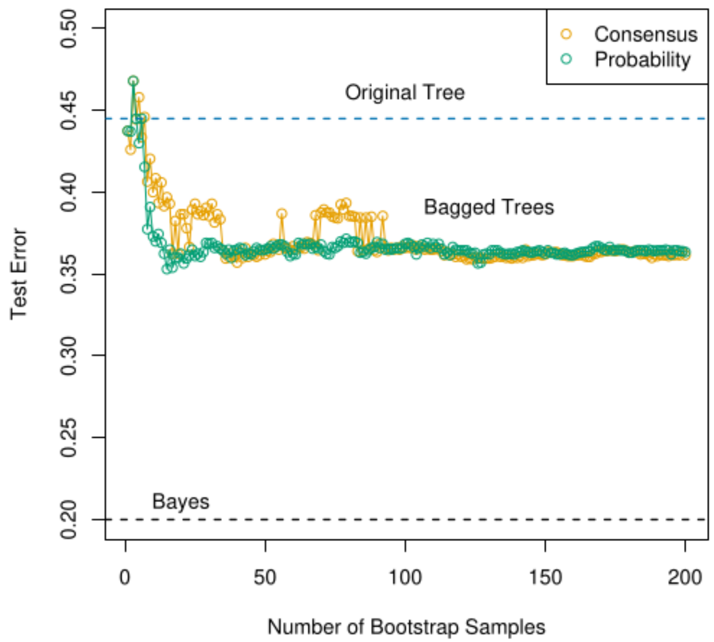
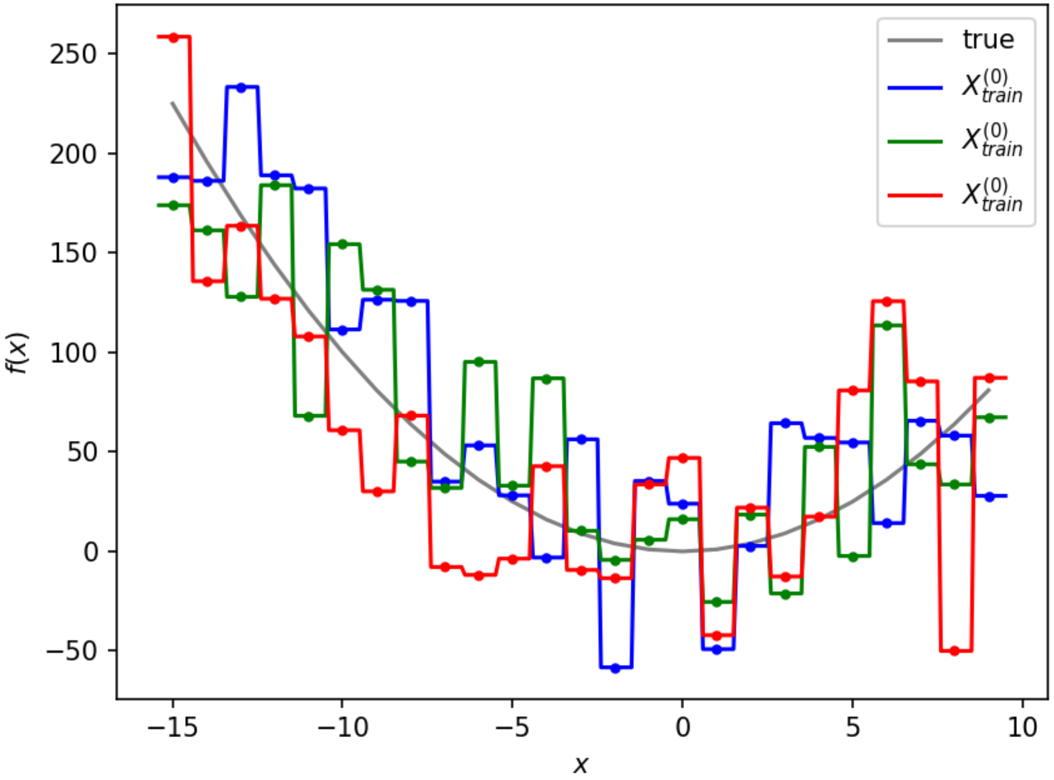


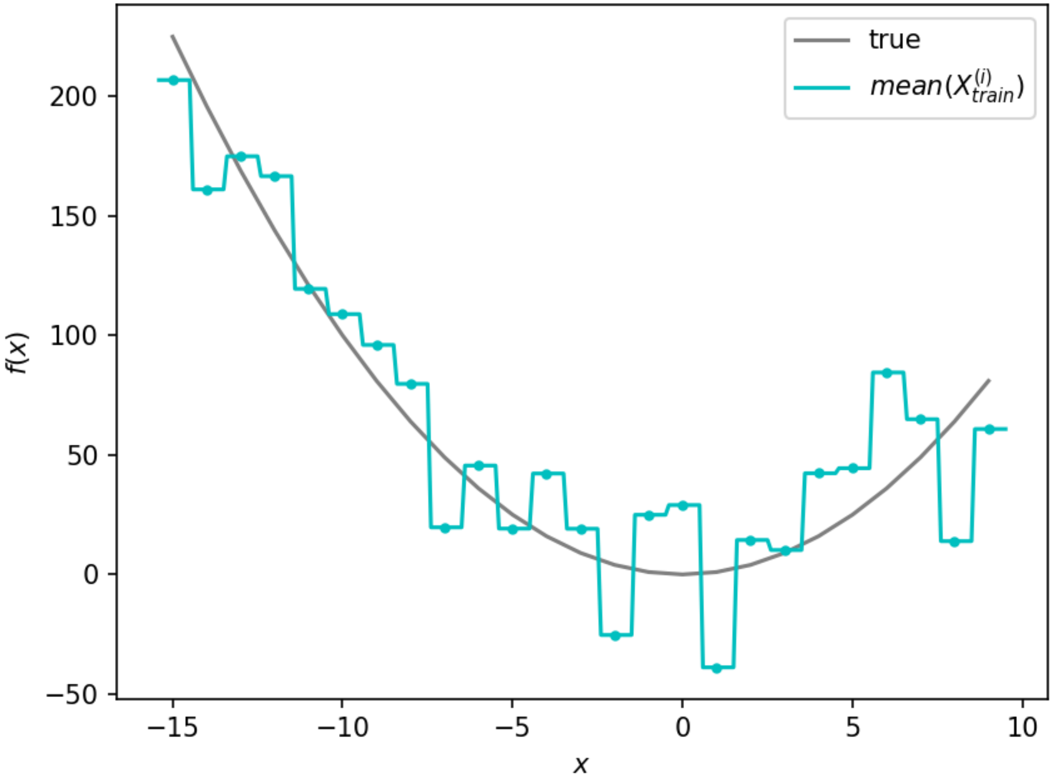

Figure: Experiment on number of Bootstrap Samples. Snippet taken from [HTF09].

# Bagging - Bias-Variance Decomposition VII

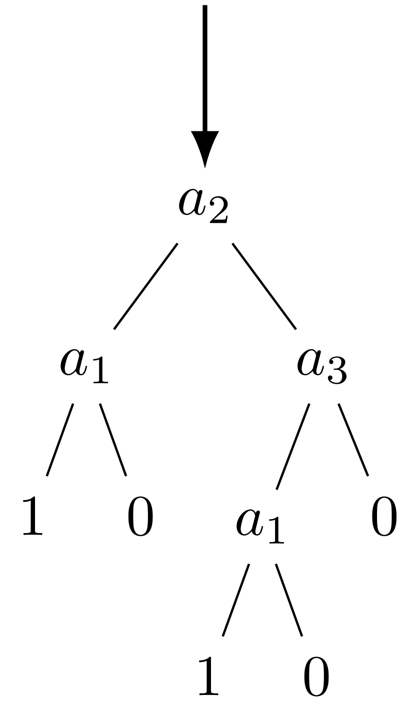



FH Salzburg

# Bagging - Bias-Variance Decomposition VIII




FH Salzburg




- one of the most used techniques
- easy to train
- low number of hyperparameters

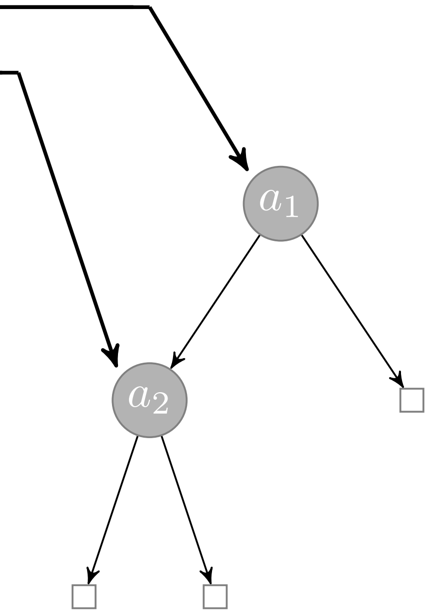
### Intuition

Random Forests = Bagging w. Trees + random feature subsets

# Random Forest

## Random Feature Subsets

| idx | $F1$ | $F2$ | $F3$ | $F4$ |
|---|---|---|---|---|
| **1** | 2 | 7 | 2 | 8 |
| **2** | 2 | 3 | 3 | 9 |
| **3** | 1 | 2 | 1 | 1 |
| **4** | 3 | 1 | 9 | 9 |
| **5** | 7 | 1 | 7 | 5 |
| ... | ... | ... | ... | ... |
| $n$ | 4 | 7 | 9 | 2 |




**Number of Features per Node**

$$\log_2(m) + 1,$$

where $m$ is the number of input features within the dataset[a].

[a] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324.

FH Salzburg

# Random Forest

## Random Feature Subsets - Tree vs. Node

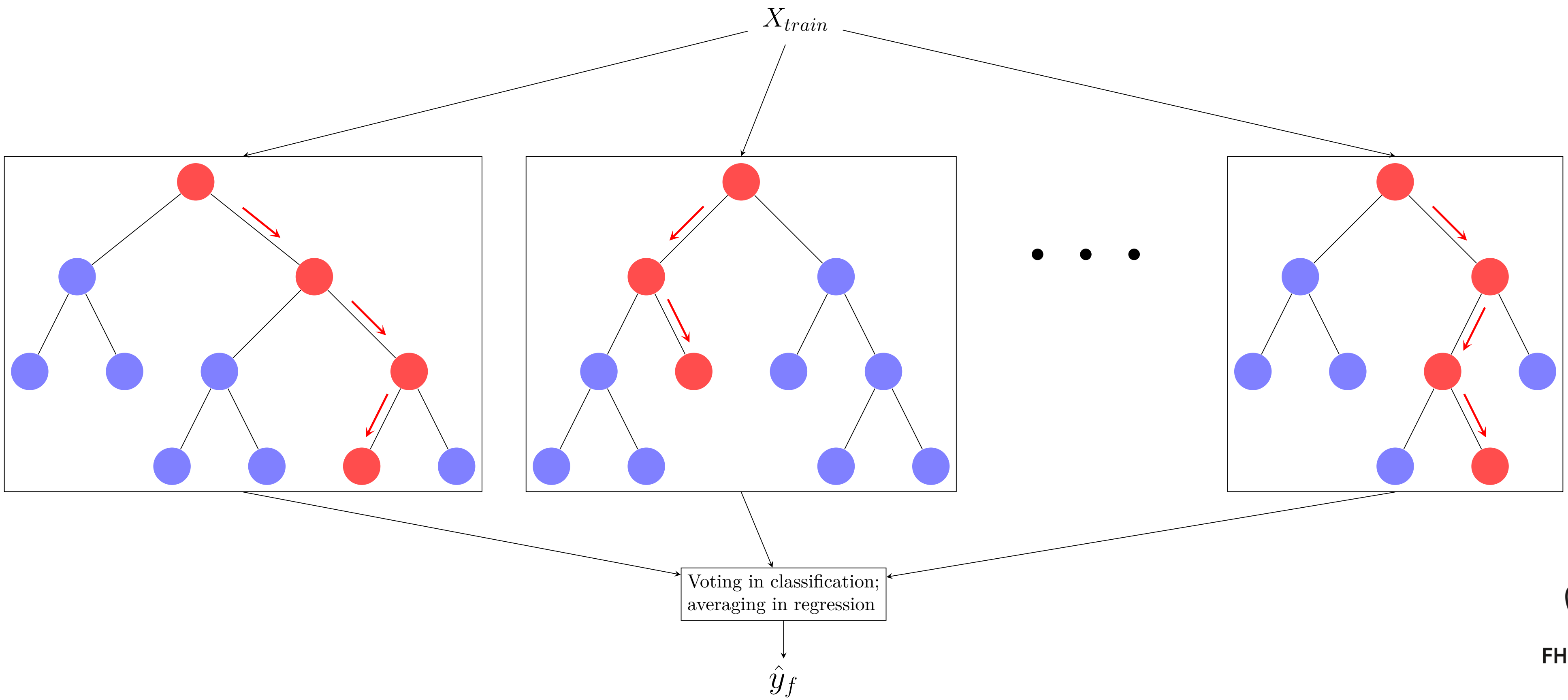
| | | |
|---|---|---|
| **Whole Tree** | *random subspace method* | "My method relies on an autonomous, pseudo-random procedure to select a small number of dimensions from a given feature space."[3] |
| **Single Node** | *random forests* | "The simplest random forest with random features is formed by selecting at random, at each node, a small group of input variables to split on"[4] |

[3] Tin Kam Ho. "The random subspace method for constructing decision forests". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8 (Aug. 1998), pp. 832–844. ISSN: 01628828. DOI: 10.1109/34.709601.

[4] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324.

FH Salzburg

# Random Forest
## Prediction




FH Salzburg

# Random Forest

## Feature Importance

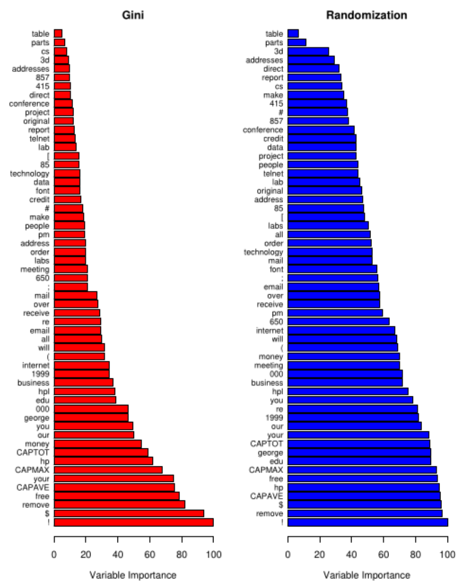



Figure: Variable importance plots for a classification random forest grown on the spam data. The left plot bases the importance on the Gini splitting index. The right plot uses oob randomization to compute variable importances, and tends to spread the importances more uniformly. Taken from[5].

[5] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer series in statistics. New York, NY: Springer, 2009. ISBN: 978-0-387-84857-0.

"In contrast to the original publication[6], the scikit-learn implementation combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class."[7]

## $\rightarrow$ Soft-Voting

[6] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324.

[7] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

# Random Forest

## Extremely Randomized Trees (Extra Trees)[8]

Random Forest random components:

1. ____________________
2. ____________________

ExtraTrees adds:

3. ____________________

FH Salzburg

[8] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine Learning* 63.1 (Apr. 2006), pp. 3–42. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-006-6226-1.