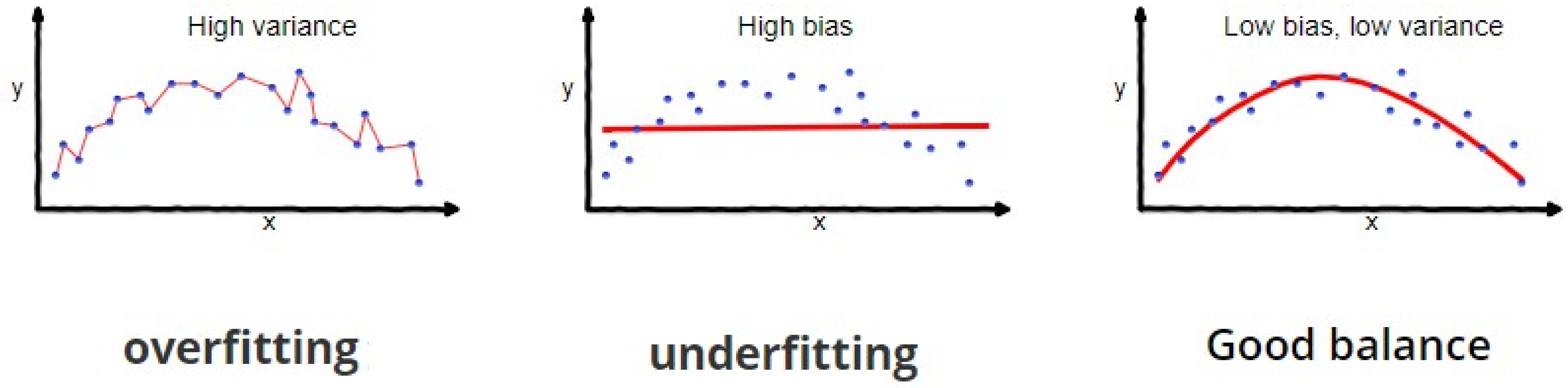
Martin Mundt, Dr. Iuliia Pliushch, Prof. Dr. Visvanathan Ramesh

Pattern Analysis & Machine Intelligence Praktikum: MLPR-SS21

Week 3: Random Forests

Recap: bias-variance trade-off

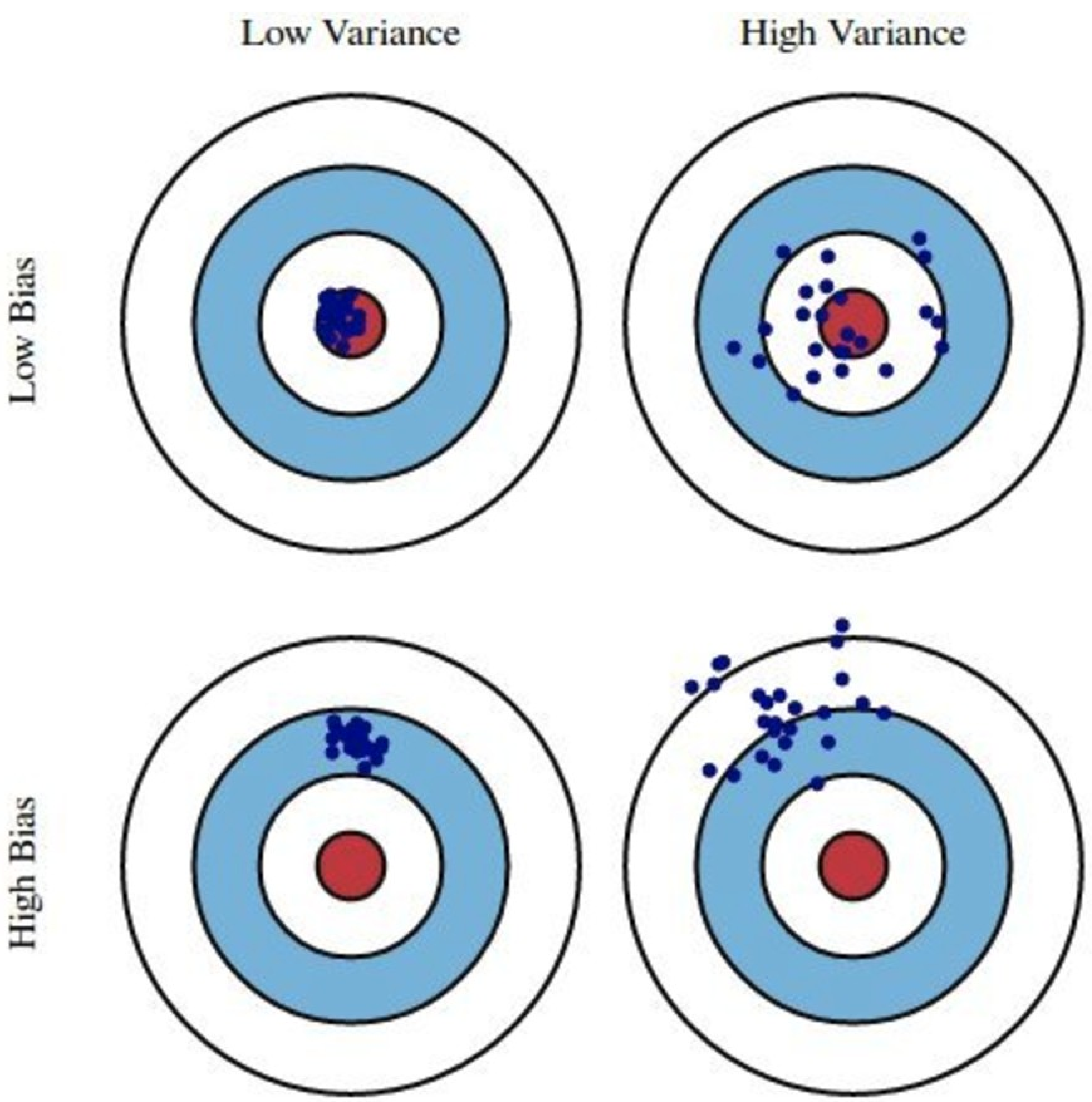
L



https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Recap: bias-variance trade-off

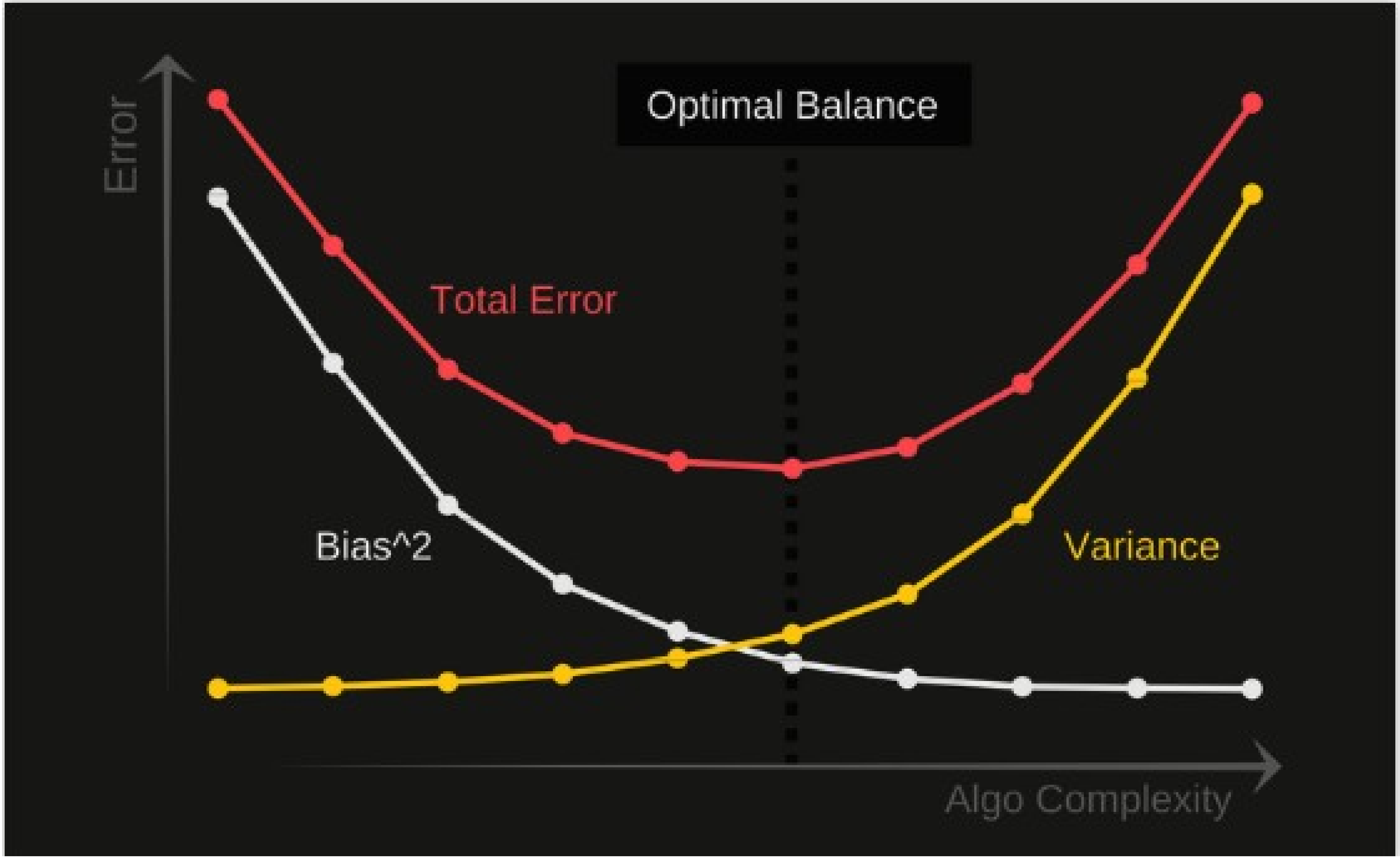
L



https://becominghuman.ai/machine-learning-bias-vs-variance-641f924e6c57

Recap: bias-variance trade-off

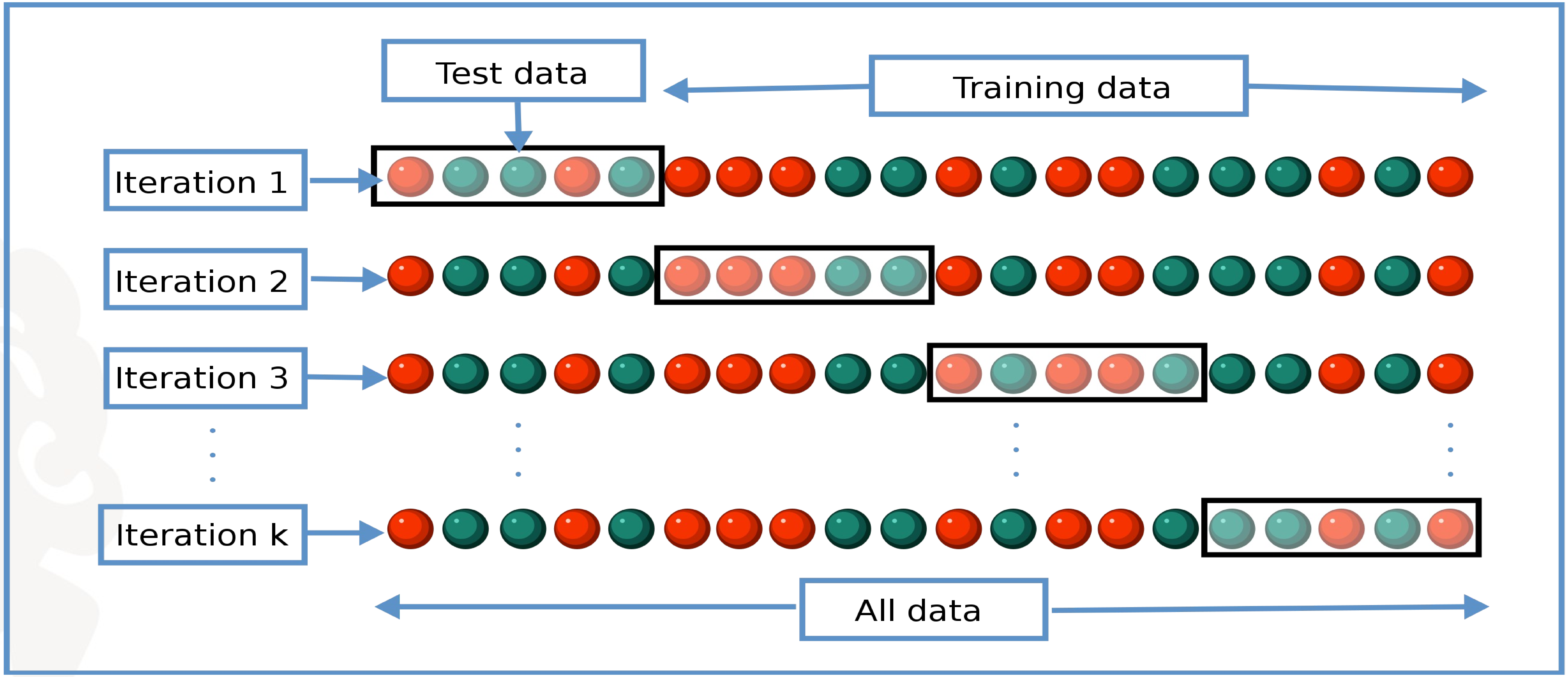
L



https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

K-fold cross-validation

L

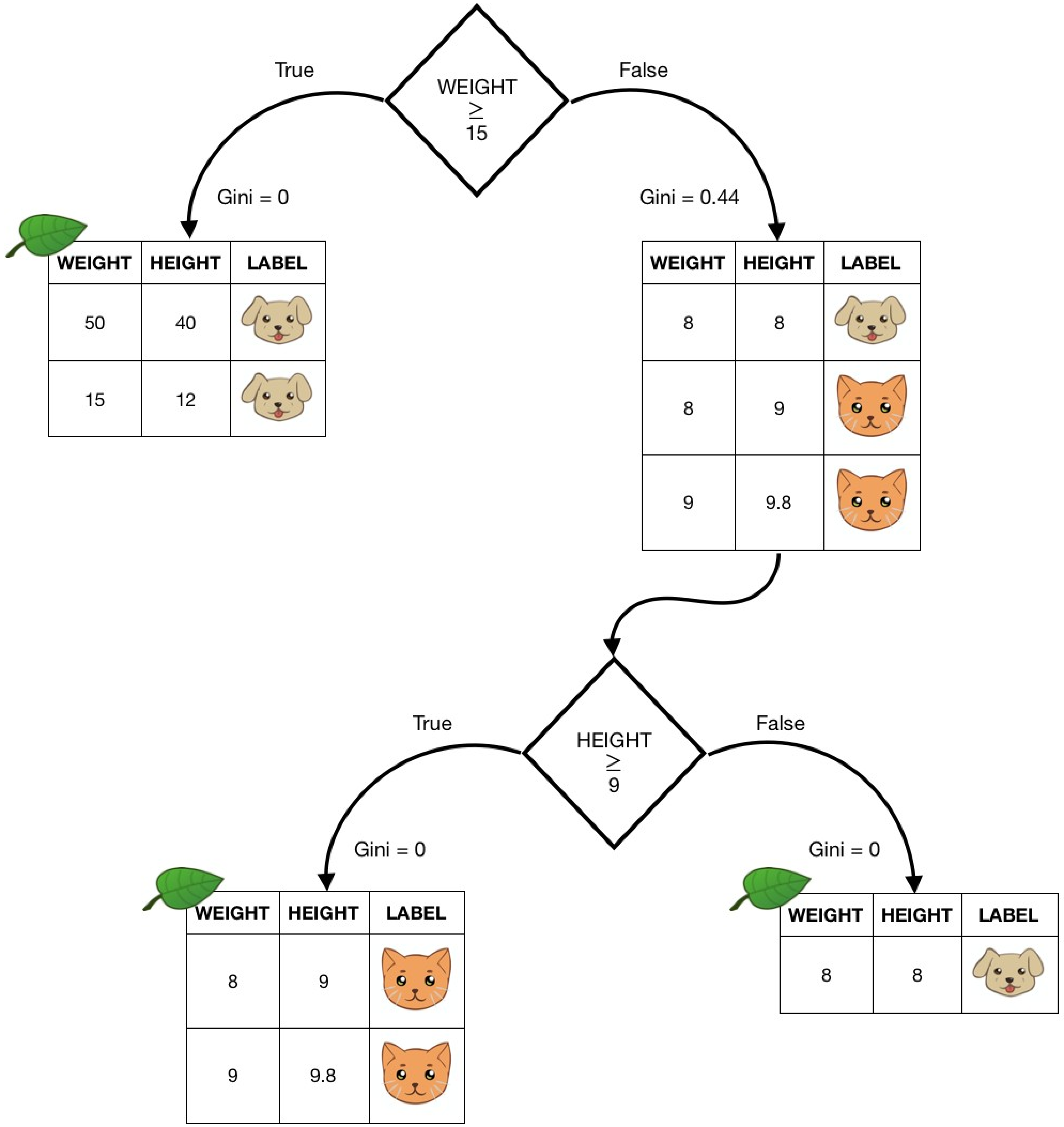


By Gufosowa - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=82298768

Decision trees and random forests

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- **Decision tree** is a machine learning algorithm for classification and regression
- **Random forests** is an **ensemble** learning algorithm which uses **multiple** decision trees for classification and regression



https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

Decision trees and random forests

- How to **split** the data for the **trees**?
- How to **combine** the results of the **trees**?

Input data:
Set of features

Numerical?
(continuous or discrete)
Categorical?
(map to discrete)
Missing values?

Random Forest

Decision tree 1

Decision tree 2

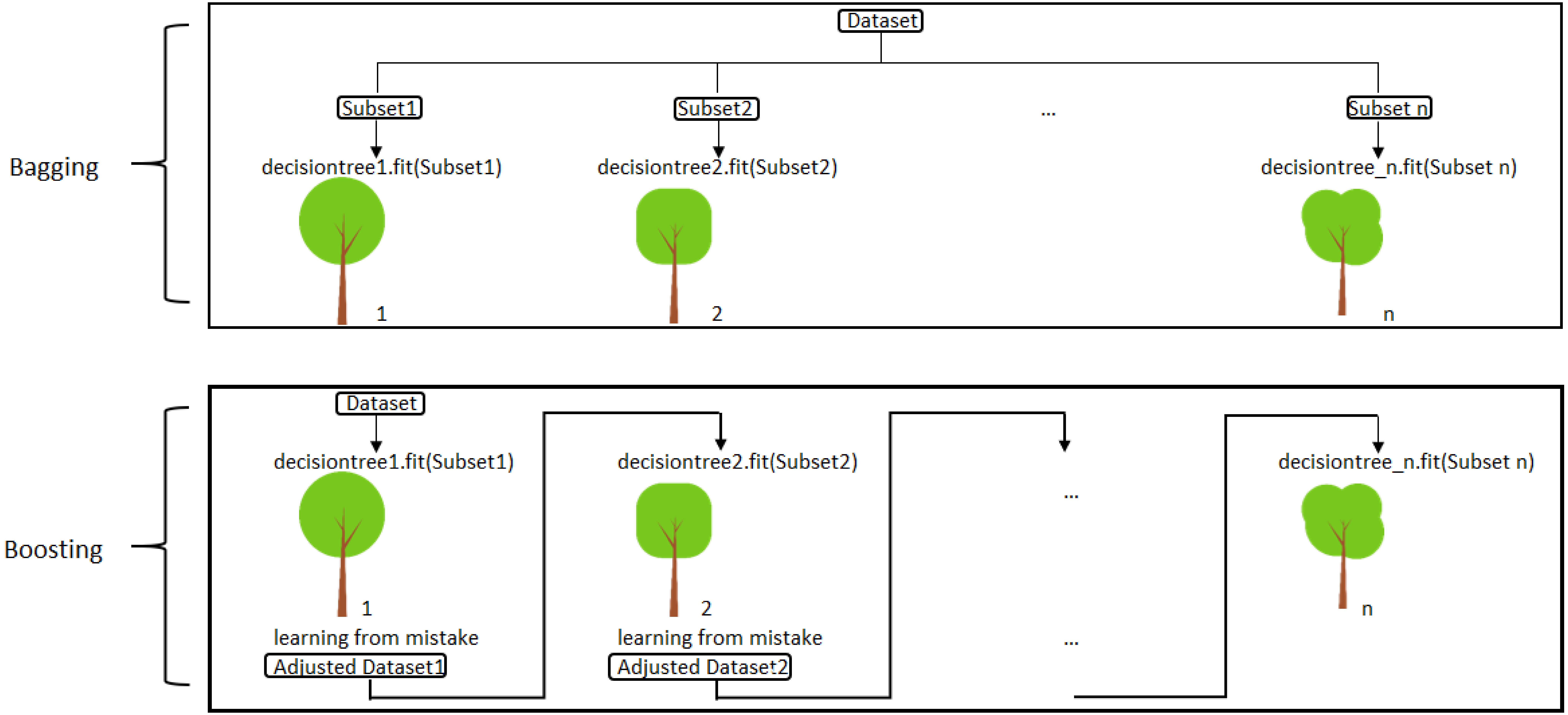
Decision tree n

Output:

A **category**
or **regressed value**
for a data instance

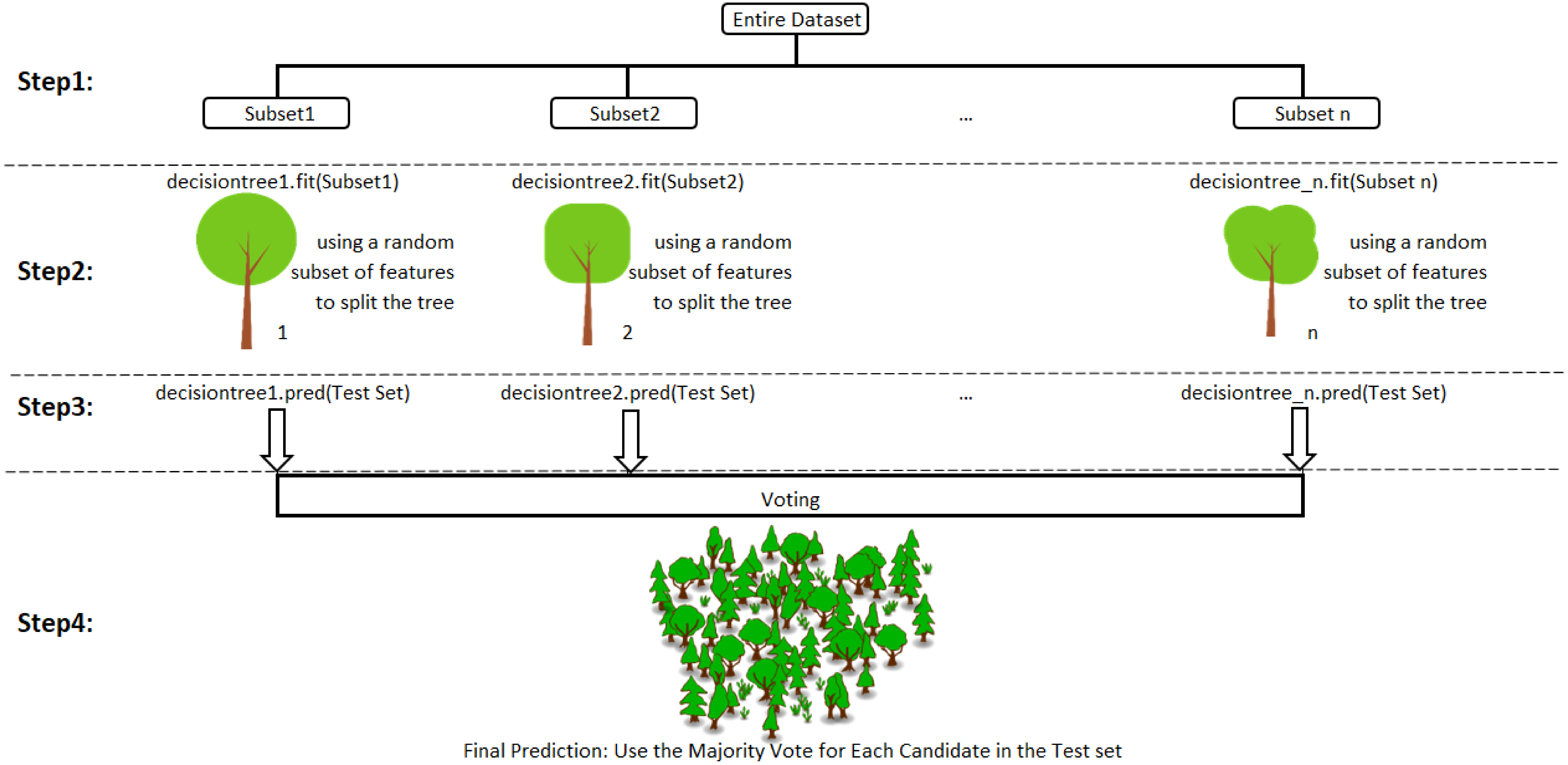
- **Pre- (while growing) and Postpruning** of trees as a means to avoid **overfitting**

Ensemble methods



https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Random forest (bagging)



https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Decision tree algorithms

- **ID3** (developed in 1986 by Ross Quinlan):
  - **categorical** features and targets
  - splitting criterion: **Information Gain**

- **C.5** (Quinlan) – commercial version of C4.5

- **C4.5** (Quinlan, 1993):
  - partitions the **continuous** features into a **discrete** set of intervals
  - suports missing values
  - splitting criterion: **Gain Ratio**

- **CART** (Classification and Regression trees):
  - similar to C4.5
  - supports **numerical target** variables (regression)
  - splitting criterion: **Gini-Index** for Classification**,** **Sum-of-Squares** for Regression

https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://scikit-learn.org/stable/modules/tree.html

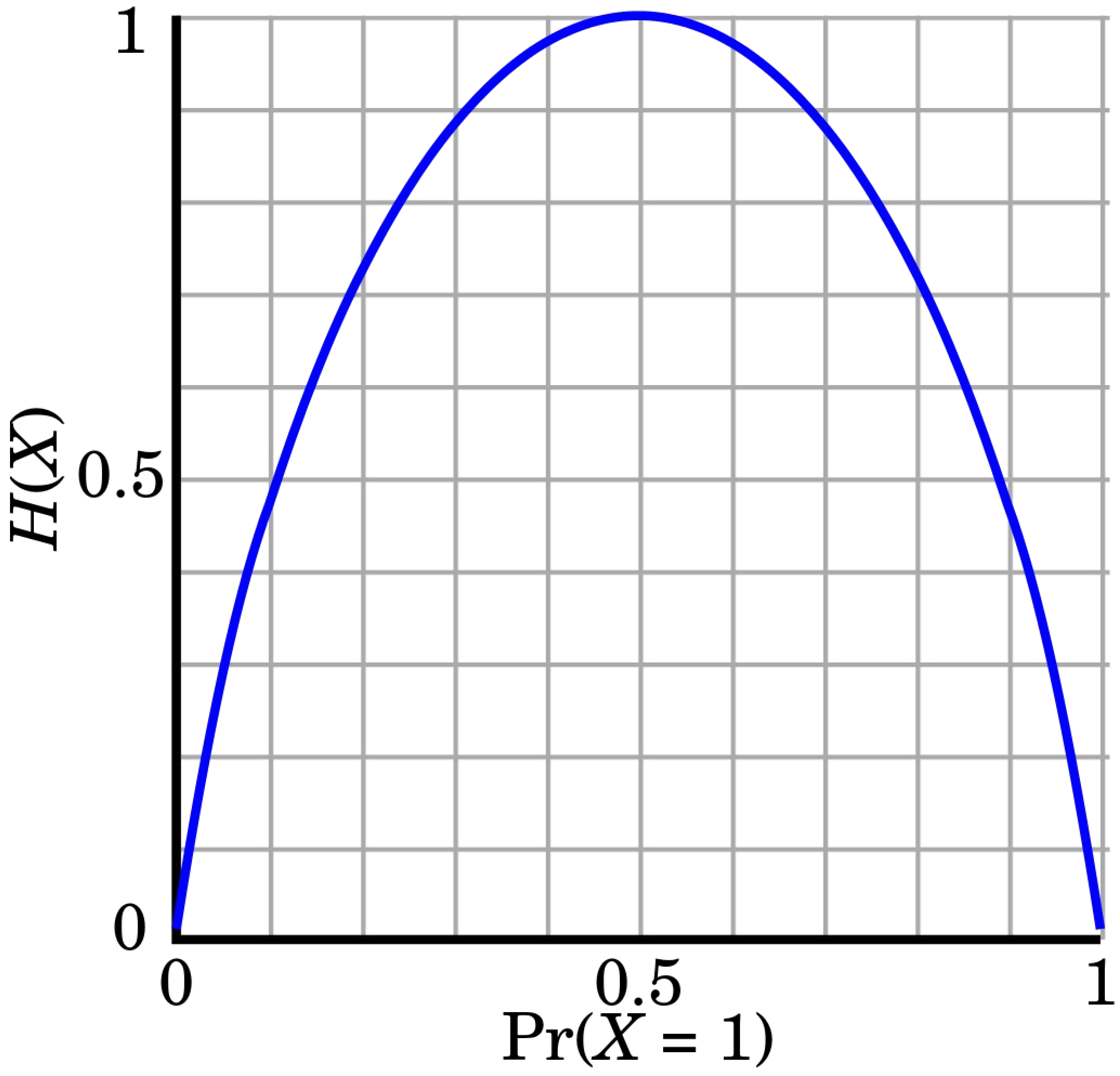
Splitting criteria: Entropy

- **Binary:**

$$-p_0 * \log_2(p_0) - p_1 * \log_2(p_1)$$

- **Multiclass:**

$$-\sum_{i \in Classes} p_i * \log_2(p_i)$$



By Brona and Alessio DamatoNewer version by Rubber Duck (☮ • ✍) - original work by Brona, published on Commons at Image:Binary entropy plot.png. Converted to SVG by Alessio Damato, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1984868

https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

Splitting criteria

$$H(parent) = -\frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{3}{5} * \log_2\left(\frac{3}{5}\right) = 0.97$$

- **Entropy Gain:**

$$Gain(S, A) = H(S) - \sum_i \frac{|S_i|}{|S|} H(S_i)$$

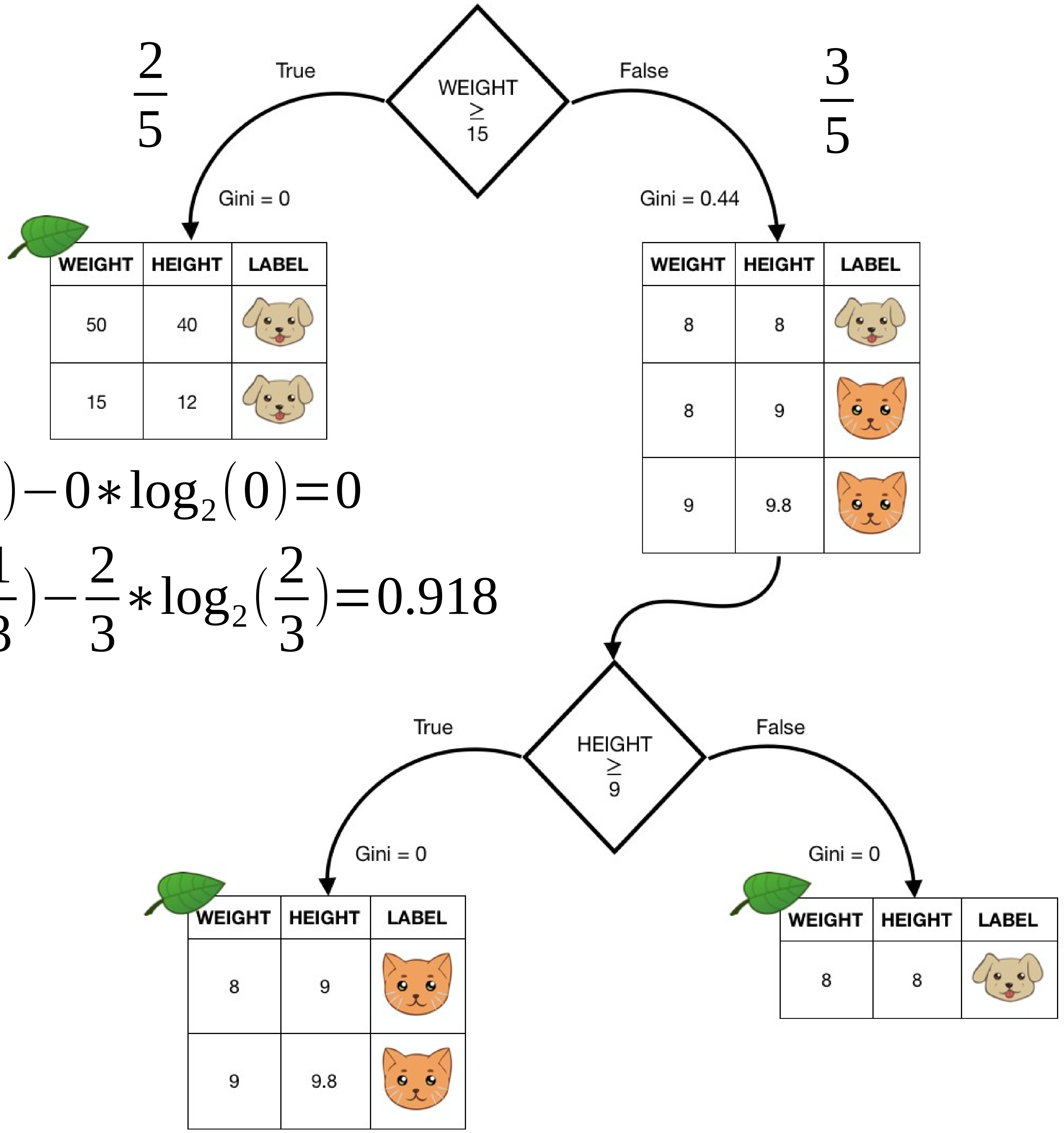
- **Intrinsic Information:**

$$H(leftchild) = -\frac{2}{2} * \log_2\left(\frac{2}{2}\right) - 0 * \log_2(0) = 0$$

$$H(rightchild) = -\frac{1}{3} * \log_2\left(\frac{1}{3}\right) - \frac{2}{3} * \log_2\left(\frac{2}{3}\right) = 0.918$$

$$IntI(S, A) = -\sum_i \frac{|S_i|}{|S|} \log_2\left(\frac{|S_i|}{|S|}\right)$$

- **Gain Ratio:** $\dfrac{Gain(S, A)}{IntI(S, A)}$



https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

Splitting criteria (CART)

- **Gini (impurity measure)**
  – for classification

$$Gini(S) = 1 - \sum_{i \in Classes} p_i^2$$

$$Gini(S, A) = \sum_i \frac{|S_i|}{|S|} Gini(S_i)$$

- **MSE (Mean Squared Error)**
  – for regression

$$MSE(S) = \frac{1}{N} \sum_{i \in Ndata} (y_i - y_{i\,estimated})^2$$

https://www.ke.tu-darmstadt.de/lehre/ws-18-19/mldm/dt.pdf

https://en.wikipedia.org/wiki/Variance

https://web.stanford.edu/class/stats202/content/lec19.pdf

Splitting criteria (summary)

- For classification: **Information gain** (entropy-based) vs. **Gini** (impurity)
  - Mainly computational difference (logarithmic function)

- For regression**: Mean squared error**
  - If the estimator is unbiased, **MSE** is equal to the **variance**

$$MSE(\hat{\theta}) = Var_{\theta}(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$$

Raileanu, Laura Elena, and Kilian Stoffel. "Theoretical comparison between the gini index and information gain criteria." Annals of Mathematics and Artificial Intelligence 41.1 (2004): 77-93.

https://en.wikipedia.org/wiki/Mean_squared_error
http://people.missouristate.edu/songfengzheng/teaching/mth541/lecture%20notes/evaluation.pdf

https://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf

Inductive biases in DT and RF

- Trees that place high information gain attributes close to the root are preferred over those that do not.

- Shorter trees are preferred over longer trees.

- The choice of the **splitting criterion** introduces its own inductive bias

- **Averaging** (potentially) non-smooth interpolating trees – solution with higher degree of smoothness (better than solutions of individual trees)

http://www.lauradhamilton.com/inductive-biases-various-machine-learning-algorithms

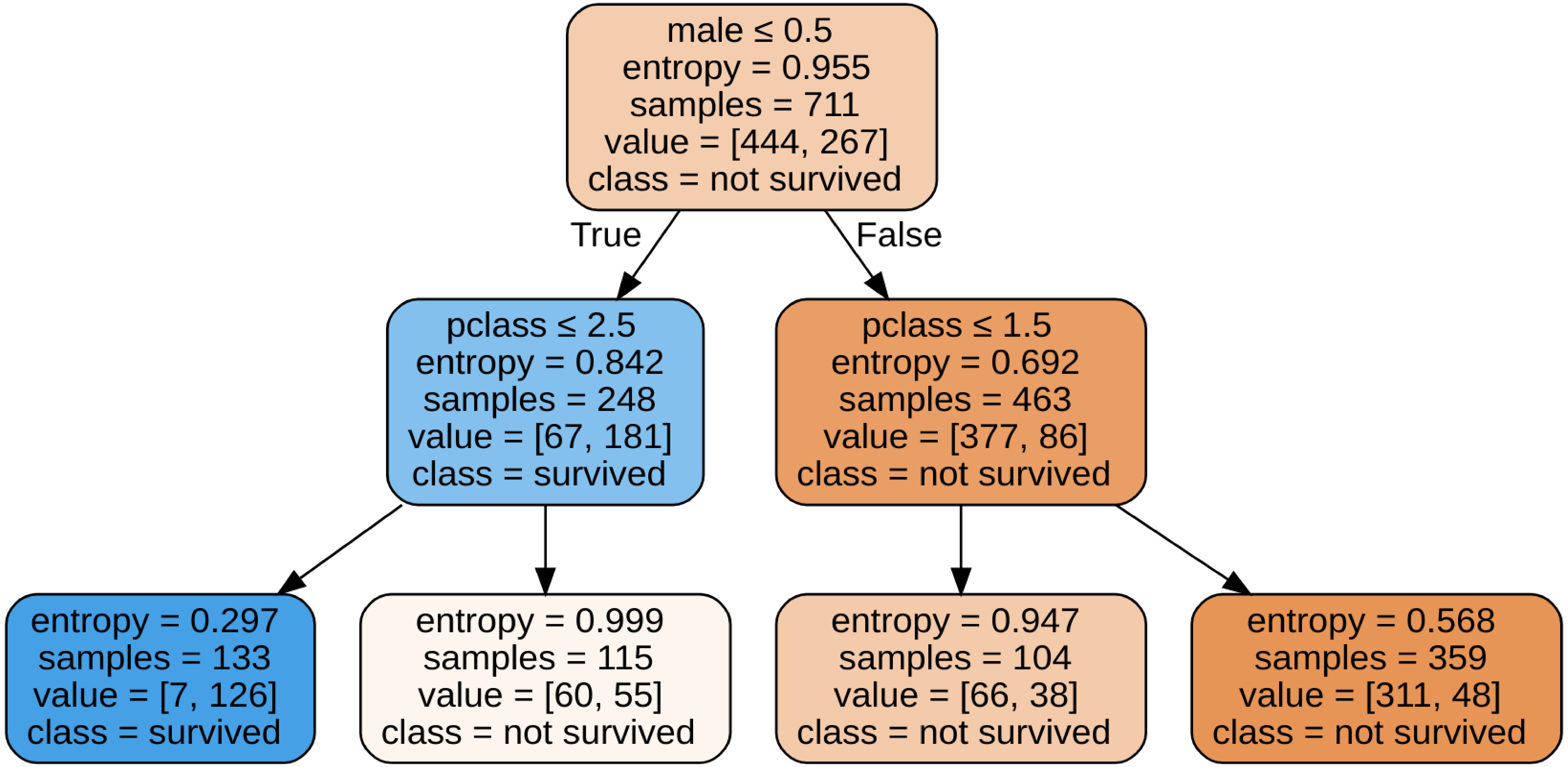
https://www.cs.columbia.edu/~djhsu/papers/biasvariance-arxiv.pdf

Decision tree

- Input: Set of features, class to predict

- 1. Create a (root) node
- 2. If termination criteria are met, make it a **leaf**
- 2. Select the best **feature** to split the data according to **criterion (loop over selected features)**
- 3. Split the **data** accordingly
- 4. Create subtrees for each **data subset (RECURSION!)**

Titanic dataset



San Francisco Crime Challenge

https://www.kaggle.com/c/sf-crime

- Predict a specific **crime category** on the basis of time of day, day of the week, city district, address and other attributes
- **Logarithmic loss** (logistic loss or cross-entropy) is used as the evaluation criterion, because **accuracy** is low (think why?)
- The aim of this exercise:

  1. practice data preprocessing

  2. see how to apply log loss (introduced in the context of neural networks) in a random forest scenario

  3. practice K-fold cross-validation

http://www.lauradhamilton.com/inductive-biases-various-machine-learning-algorithms