Machine learning modelling for infectious disease

Patrick Njage

Researcher
Research Group for Genomic Epidemiology
DTU Food

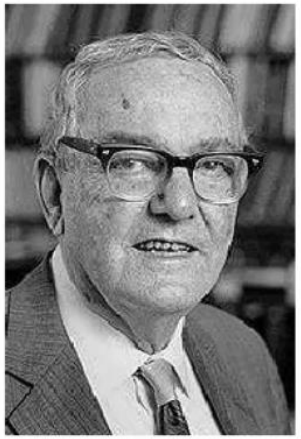
Week 8: Infectious Disease Bioinformatics course, 2022

Content

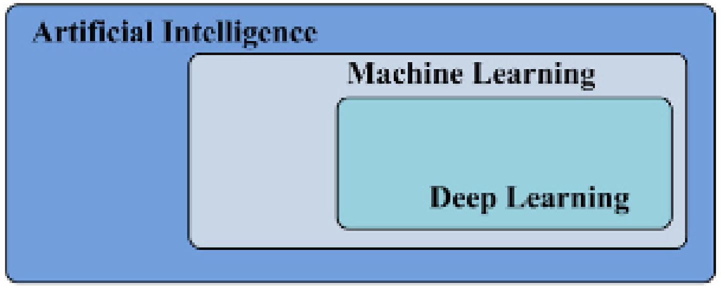
- Introduction machine learning
- Common types of machine learning algorithms
- Decision trees
- Properties of ensemble approaches
  - Boosting and bagging
  - Popular examples: Random forest and LogitBoost
- Deep learning
- Over and under-fitting
- Cross-validation
- Next generation sequencing data inputs

Machine learning

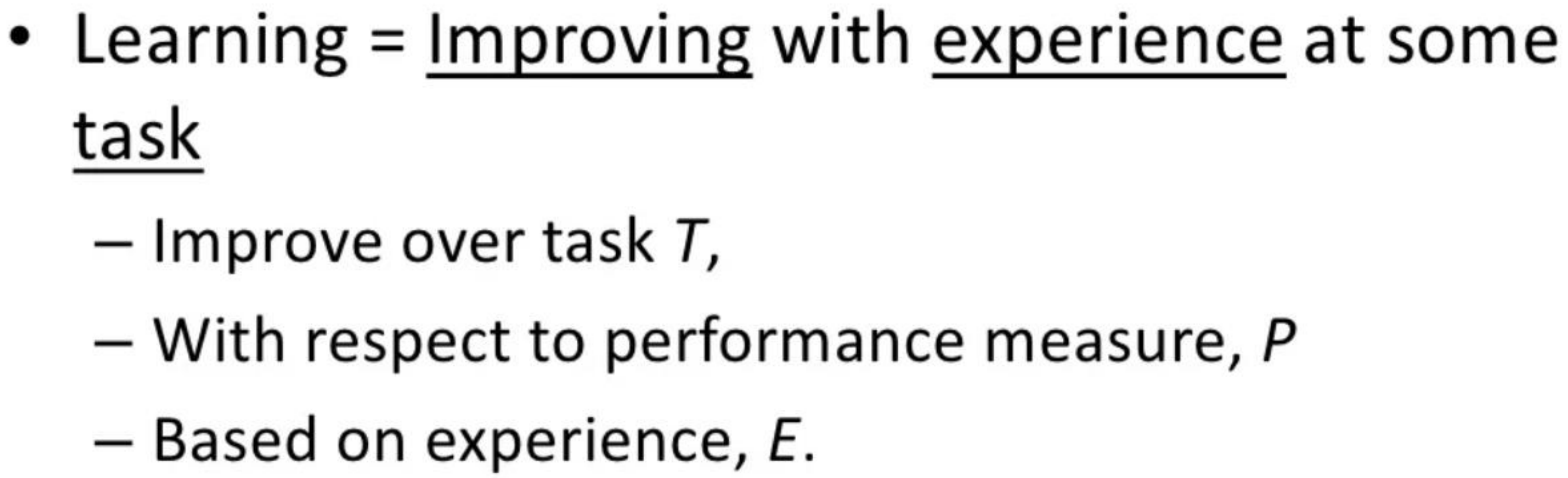
- **Herbert Alexander Simon**: "Learning is any process by which a system improves performance from experience."
- "Machine Learning is concerned with computer programs that automatically improve their performance through experience. "



Herbert Simon
Turing Award 1975
Nobel Prize in Economics 1978



Learning concept for machine learning



- Learning = Improving with experience at some task
  - Improve over task *T*,
  - With respect to performance measure, *P*
  - Based on experience, *E*.

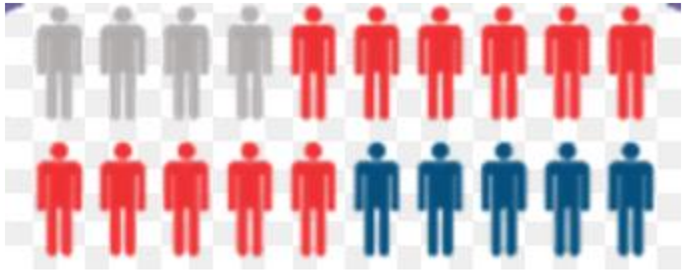
Example: spam filtering

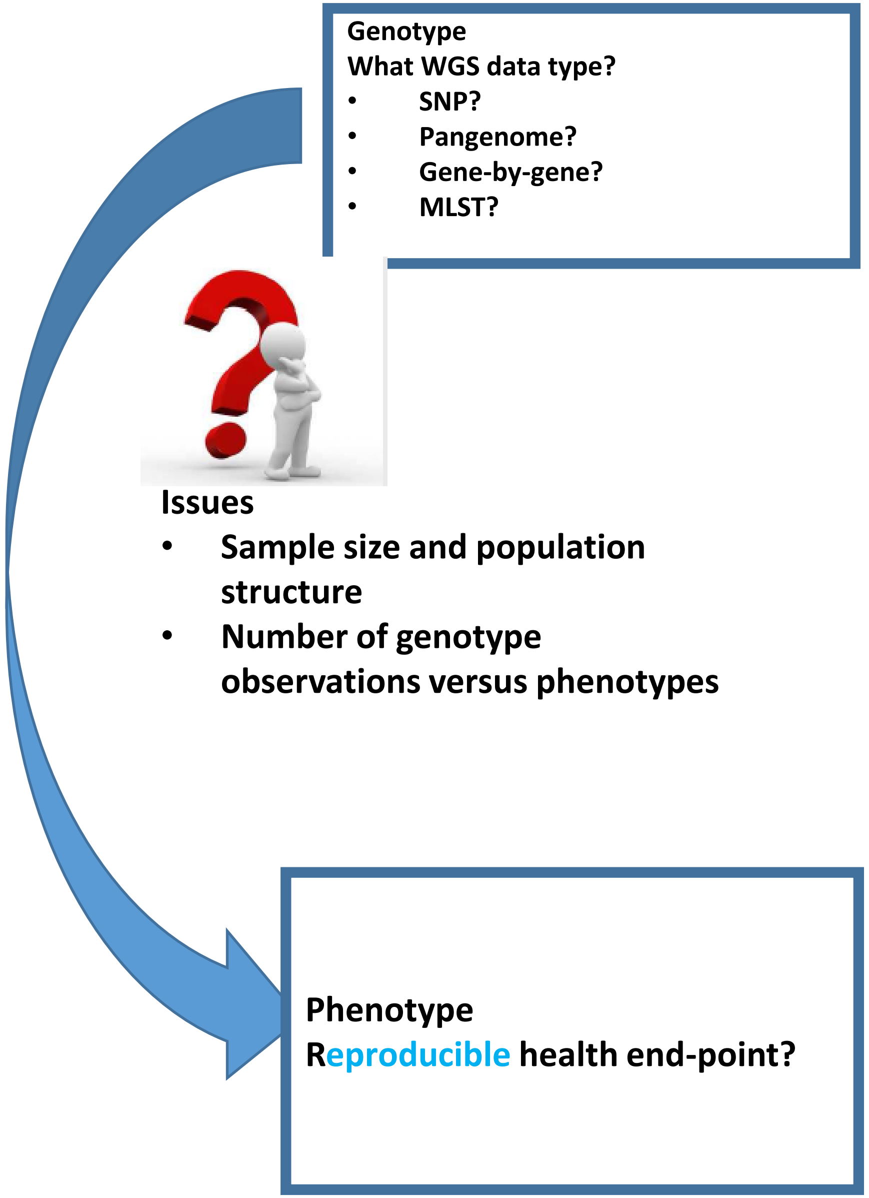
Spam - is all email the user does not want to receive and has not asked to receive

- *T*: Identify Spam Emails
- *P*:
  - % of spam emails that were filtered
  - % of ham/ (non-spam) emails that were incorrectly filtered-out
- *E*: a database of emails that were labelled by users

For infectious disease

Classical input data





Genotype
What WGS data type?

- SNP?
- Pangenome?
- Gene-by-gene?
- MLST?

Issues

- Sample size and population structure
- Number of genotype observations versus phenotypes

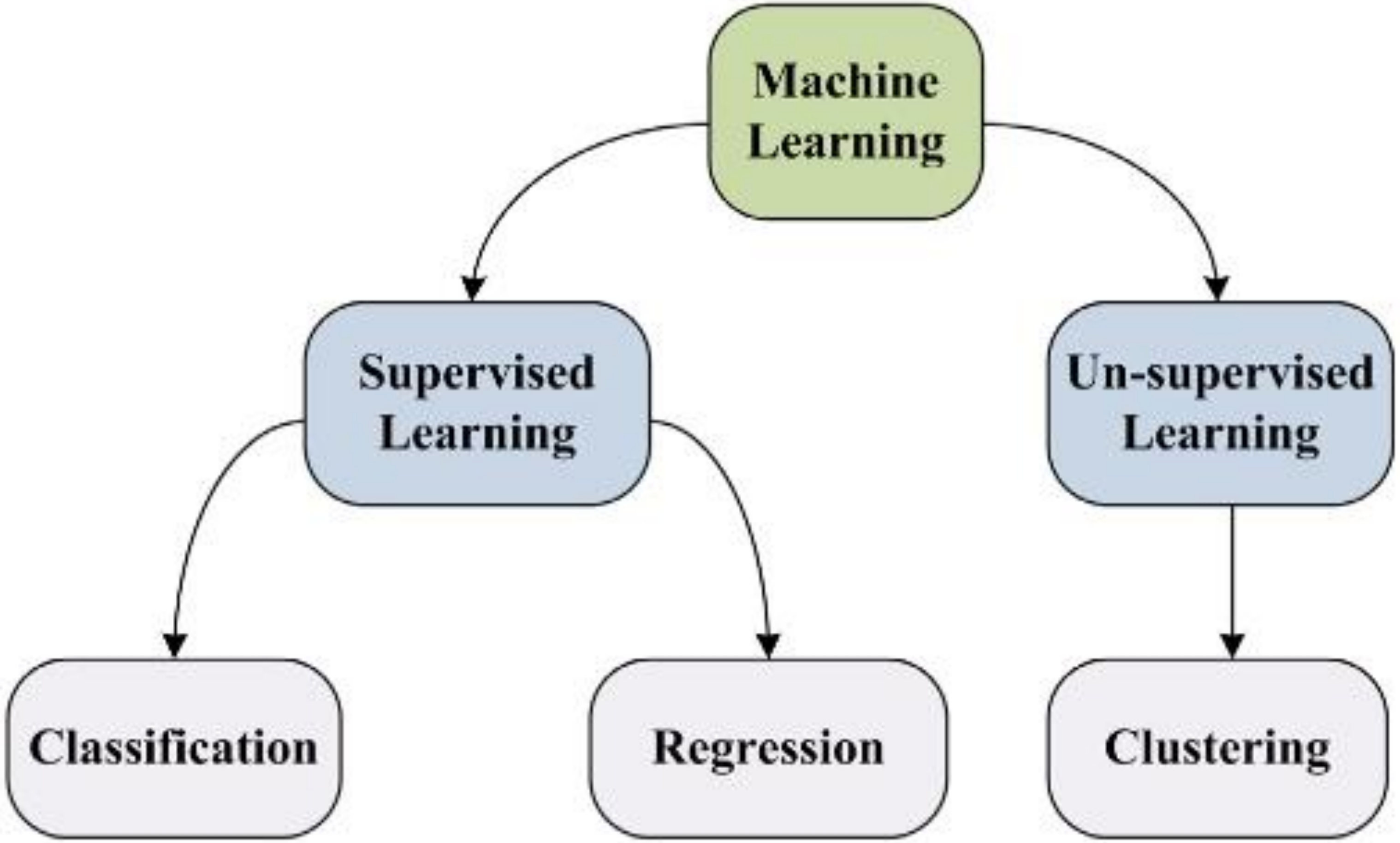
Phenotype
Reproducible health end-point?

Why machine Learning?

- Machine learning: "algorithms that improve with experience"
- Analysis of large, complex data sets
- Relevant "features" in a complex data set enable the ability to make a strong prediction
- Increase in data ad computational power
- Example applications

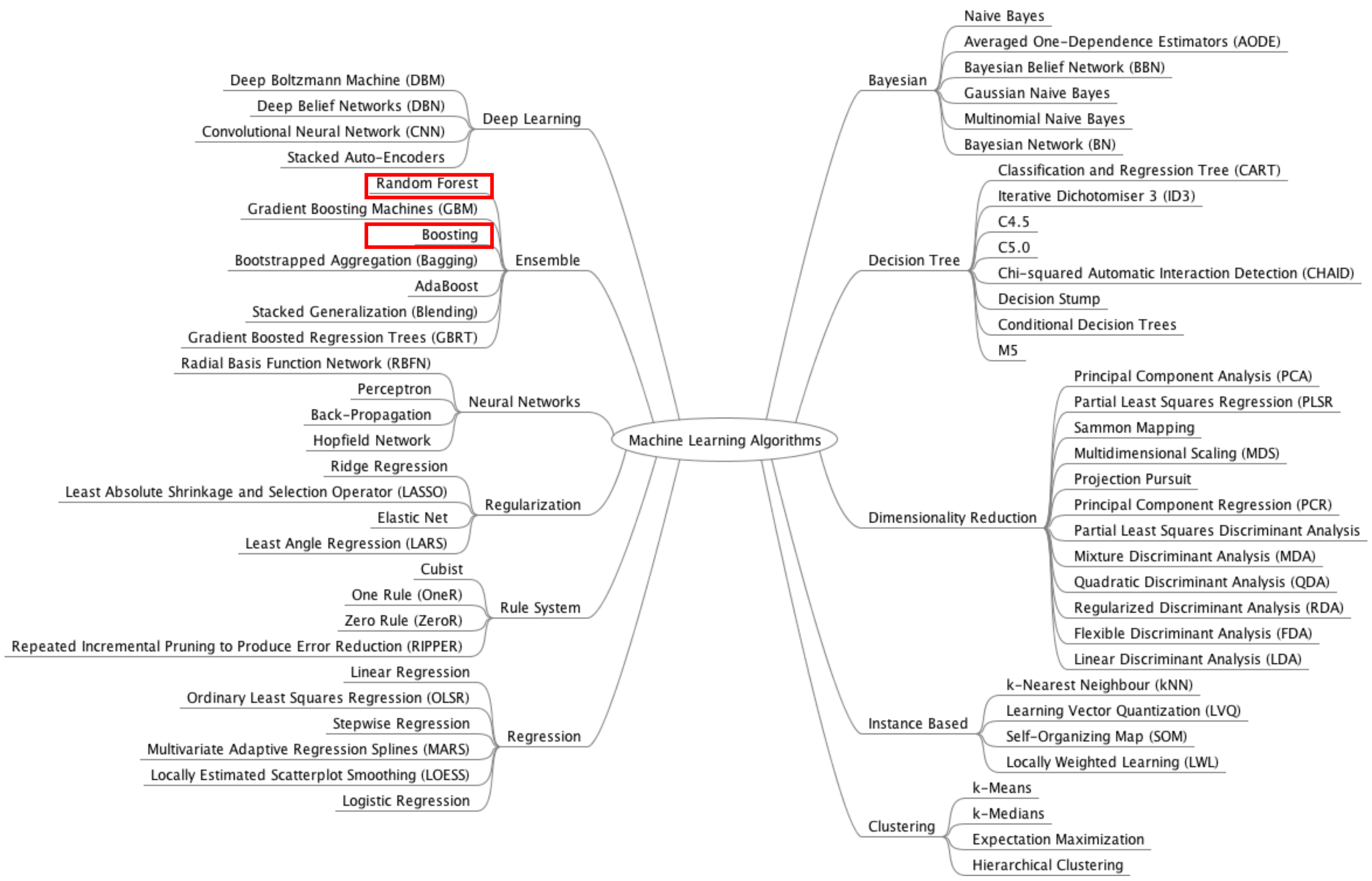
spam filtering, optical character recognition (OCR), search engines and computer vision

Machine learning categories



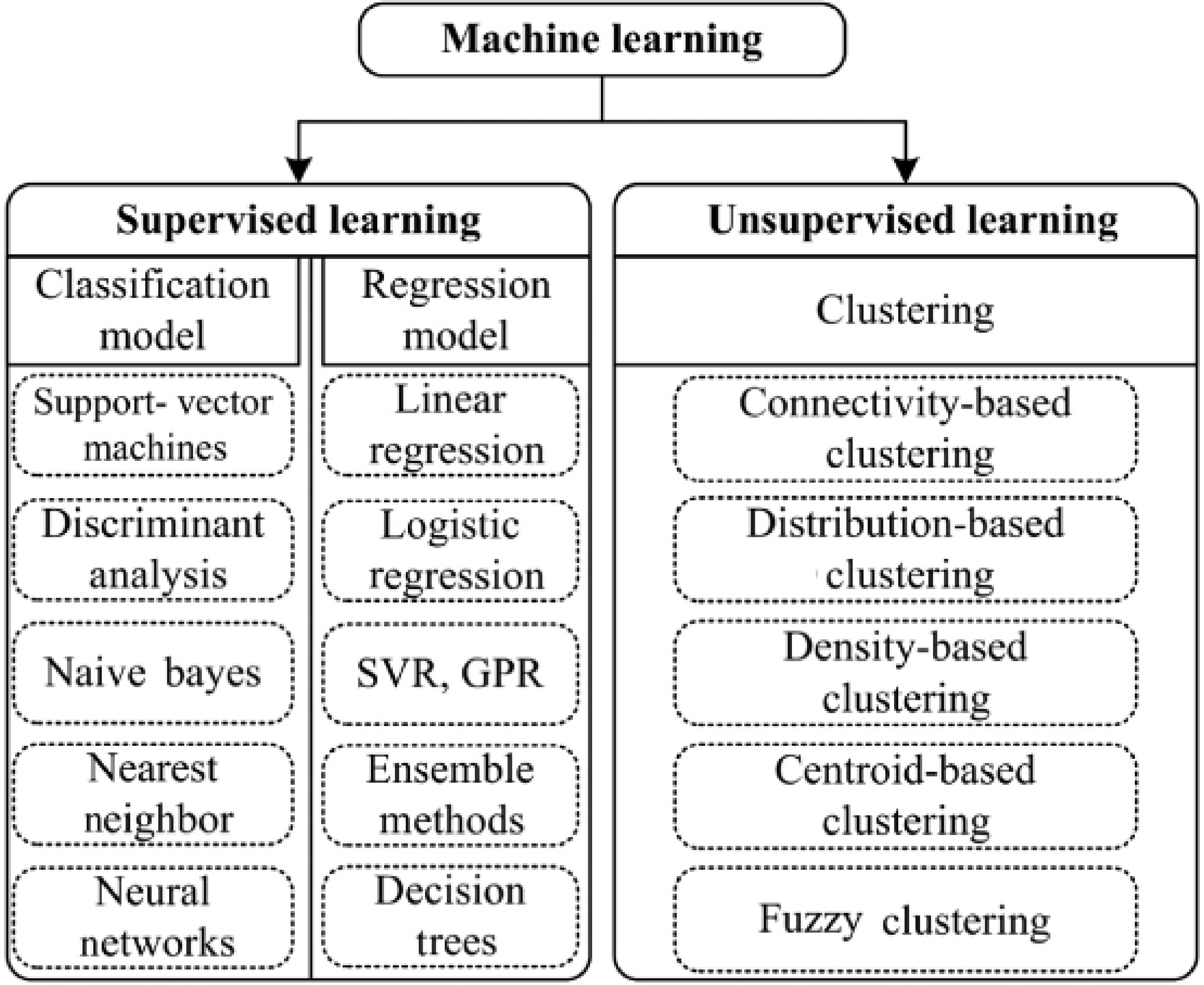
Quantitative outcome e.g. age- regression
Qualitative outcome e.g. disease type- classification

Machine Learning: Forest of algorithms
Which one do we choose?



- Machine Learning Algorithms
  - Deep Learning
    - Deep Boltzmann Machine (DBM)
    - Deep Belief Networks (DBN)
    - Convolutional Neural Network (CNN)
    - Stacked Auto-Encoders
  - Ensemble
    - Random Forest
    - Gradient Boosting Machines (GBM)
    - Boosting
    - Bootstrapped Aggregation (Bagging)
    - AdaBoost
    - Stacked Generalization (Blending)
    - Gradient Boosted Regression Trees (GBRT)
  - Neural Networks
    - Radial Basis Function Network (RBFN)
    - Perceptron
    - Back-Propagation
    - Hopfield Network
  - Regularization
    - Ridge Regression
    - Least Absolute Shrinkage and Selection Operator (LASSO)
    - Elastic Net
    - Least Angle Regression (LARS)
  - Rule System
    - Cubist
    - One Rule (OneR)
    - Zero Rule (ZeroR)
    - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
  - Regression
    - Linear Regression
    - Ordinary Least Squares Regression (OLSR)
    - Stepwise Regression
    - Multivariate Adaptive Regression Splines (MARS)
    - Locally Estimated Scatterplot Smoothing (LOESS)
    - Logistic Regression
  - Bayesian
    - Naive Bayes
    - Averaged One-Dependence Estimators (AODE)
    - Bayesian Belief Network (BBN)
    - Gaussian Naive Bayes
    - Multinomial Naive Bayes
    - Bayesian Network (BN)
  - Decision Tree
    - Classification and Regression Tree (CART)
    - Iterative Dichotomiser 3 (ID3)
    - C4.5
    - C5.0
    - Chi-squared Automatic Interaction Detection (CHAID)
    - Decision Stump
    - Conditional Decision Trees
    - M5
  - Dimensionality Reduction
    - Principal Component Analysis (PCA)
    - Partial Least Squares Regression (PLSR
    - Sammon Mapping
    - Multidimensional Scaling (MDS)
    - Projection Pursuit
    - Principal Component Regression (PCR)
    - Partial Least Squares Discriminant Analysis
    - Mixture Discriminant Analysis (MDA)
    - Quadratic Discriminant Analysis (QDA)
    - Regularized Discriminant Analysis (RDA)
    - Flexible Discriminant Analysis (FDA)
    - Linear Discriminant Analysis (LDA)
  - Instance Based
    - k-Nearest Neighbour (kNN)
    - Learning Vector Quantization (LVQ)
    - Self-Organizing Map (SOM)
    - Locally Weighted Learning (LWL)
  - Clustering
    - k-Means
    - k-Medians
    - Expectation Maximization
    - Hierarchical Clustering

Machine Learning: common options



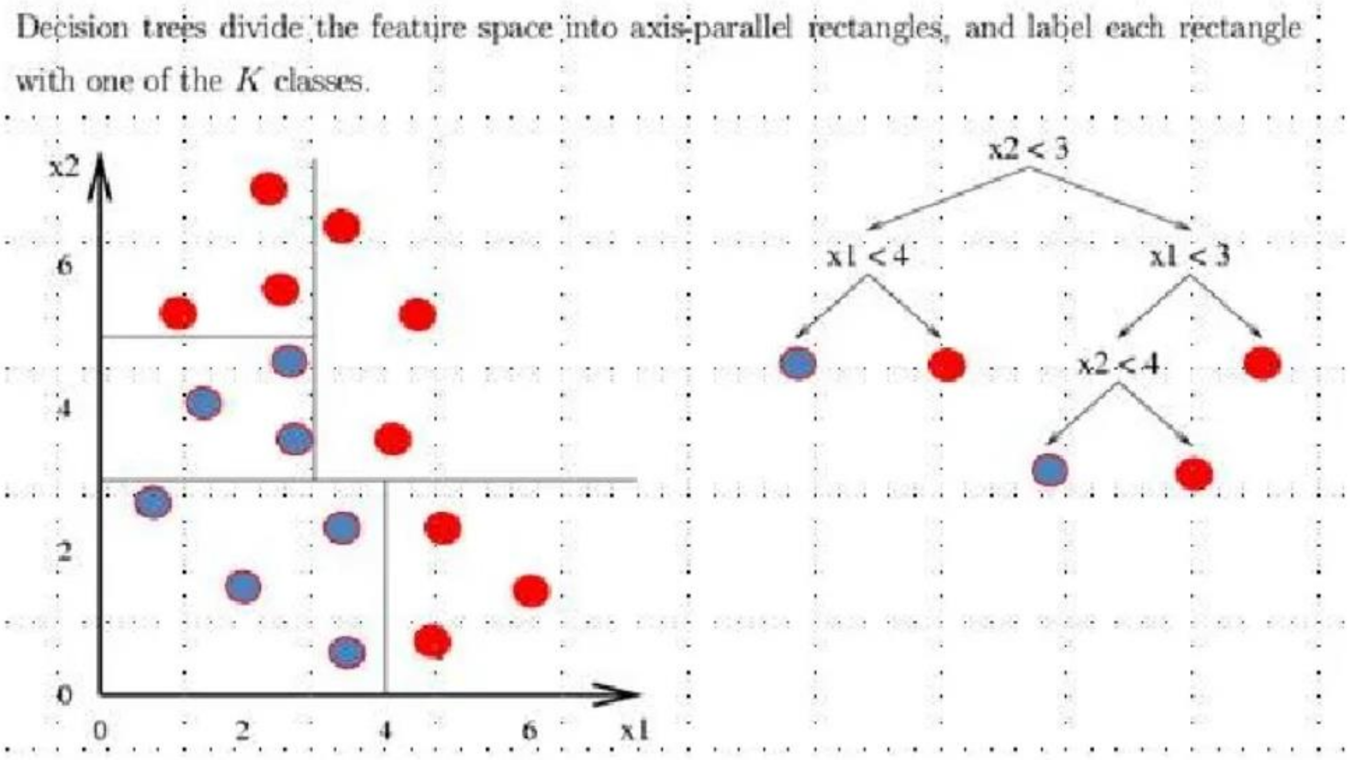
Monaco et al., 2021. Computational and Structural Biotechnology Journal, 19:4345-4359

Decision trees

- Decision trees aim: to partition the data into smaller and more **homogeneous groups**.
- **Homogeneity:** the nodes of the split are mode pure, defined e.g. by a Gini index (Kuhn & Johnson, 2013).

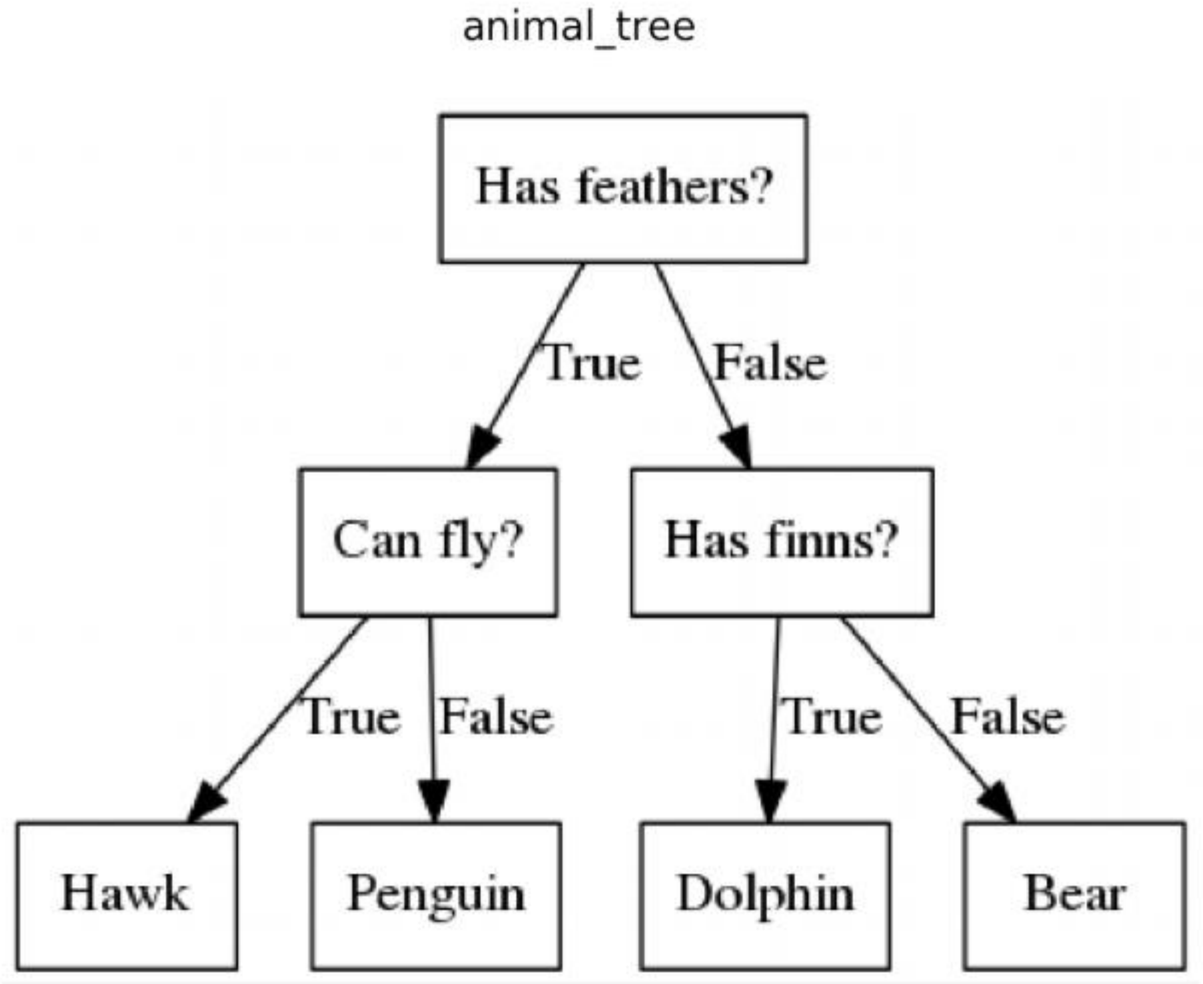
Decision trees

- A flow-chart-like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution



Decision trees divide the feature space into axis-parallel rectangles, and label each rectangle with one of the K classes.

Let’s look at a tree..and try to understand it



Let’s look at a tree..and try to understand it

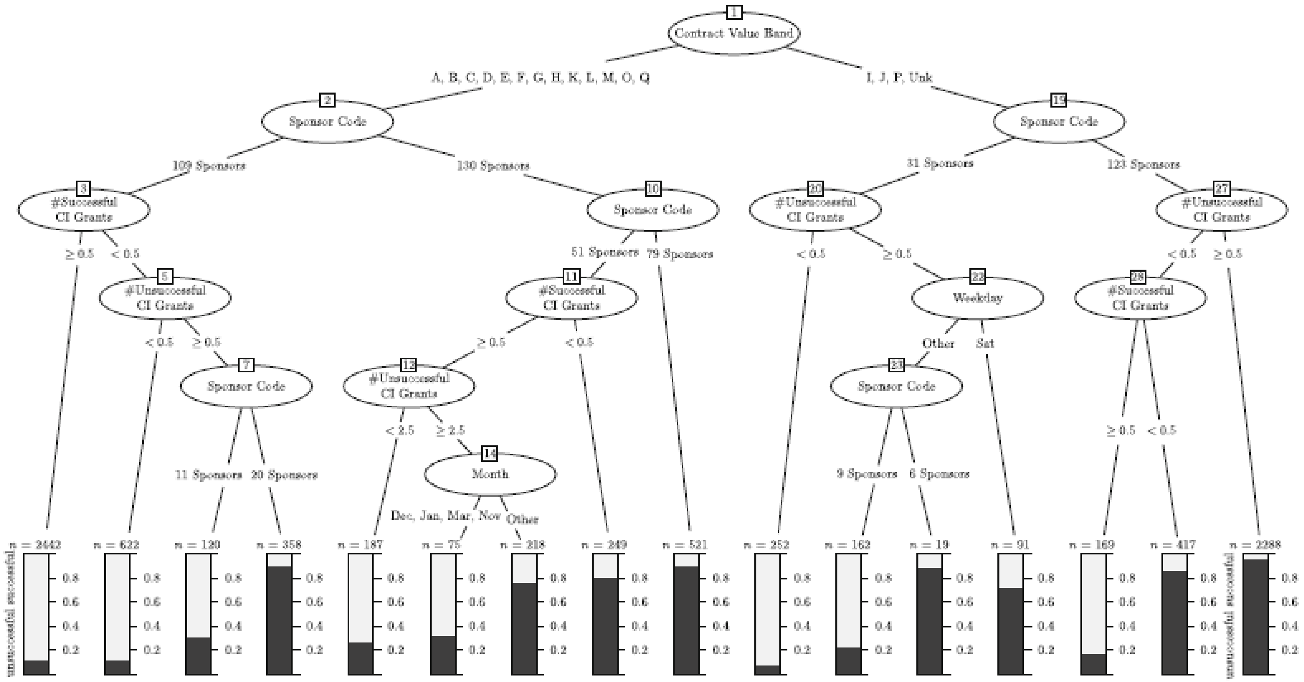


Fig. 14.3: The final CART model for the grant data using grouped category predictors

14.1 Basic Classification Trees

375

Kuhn & Johnson, 2013

- Aim: to partition the data into smaller and more **homogeneous groups**.
- **Homogeneity**: the nodes of the split are mode pure, defined e.g. by a Gini index.

Let’s look at some trees..and try to understand them

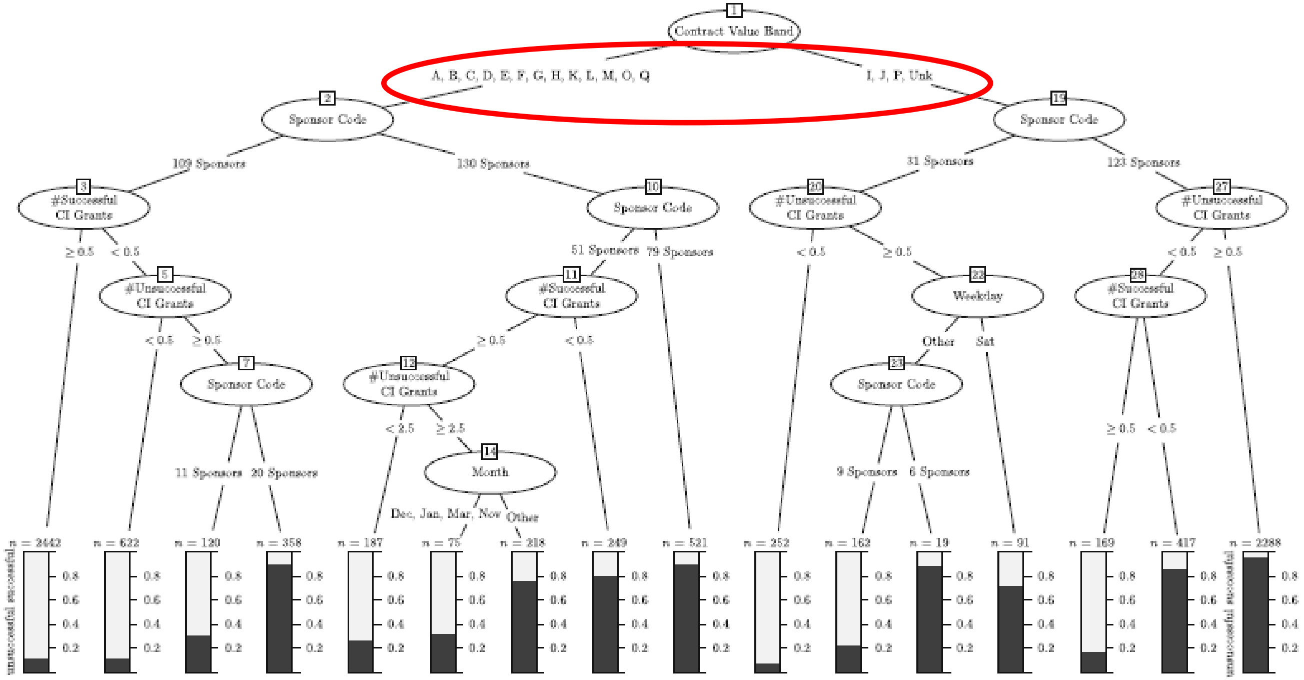


Fig. 14.3: The final CART model for the grant data using grouped category predictors

14.1 Basic Classification Trees

375

Split determined by the Gini koefficient

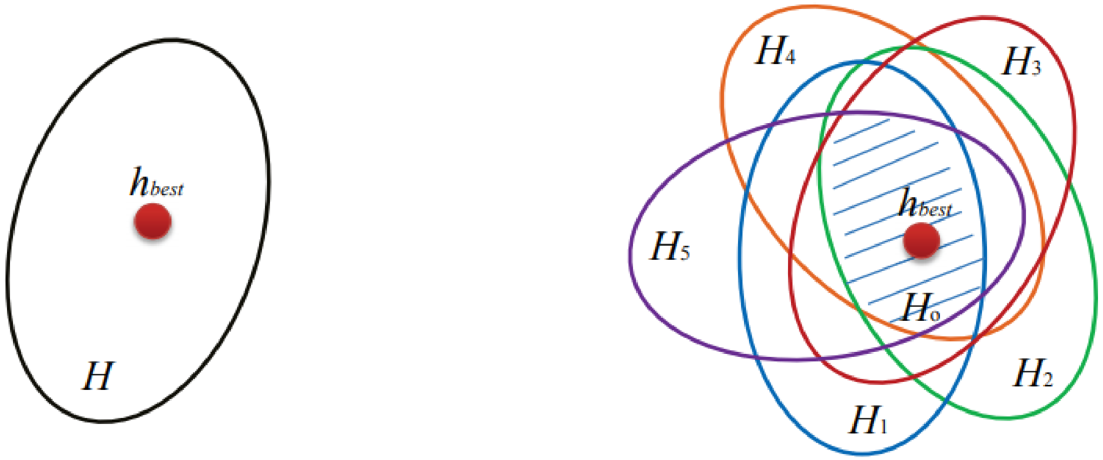
Kuhn & Johnson, 2013

Ensemble methods

- Methods that **combine multiple trees** or methods into one model that tends to **outperform the single models** (Kuhn & Johnson, (2013), Ren et al., (2016)

- Ensemble methods are applied in the field of bioinformatics where the **sample size is often low and number of features/predictors often very high** (Yang et al. (2010).
- A large number of ensemble methods have been applied to biological data analysis.
- Aim of ensemble methods: to achieve **more accurate classifications on training data** and **better generalization in predictions on unseen data**
- Include class of models such as bagging, boosting and random forest (Ren et al., (2016)
    - Random forests: handles high dimesionality data (Yang et al. (2010).
    - Bagging + boosting effective in dealing with data with low sample size (Yang et al. (2010).

Schematic illustration of hypothesis space for single classifier vs. ensemble of classifiers



(a) Hypothesis space of a single classifier

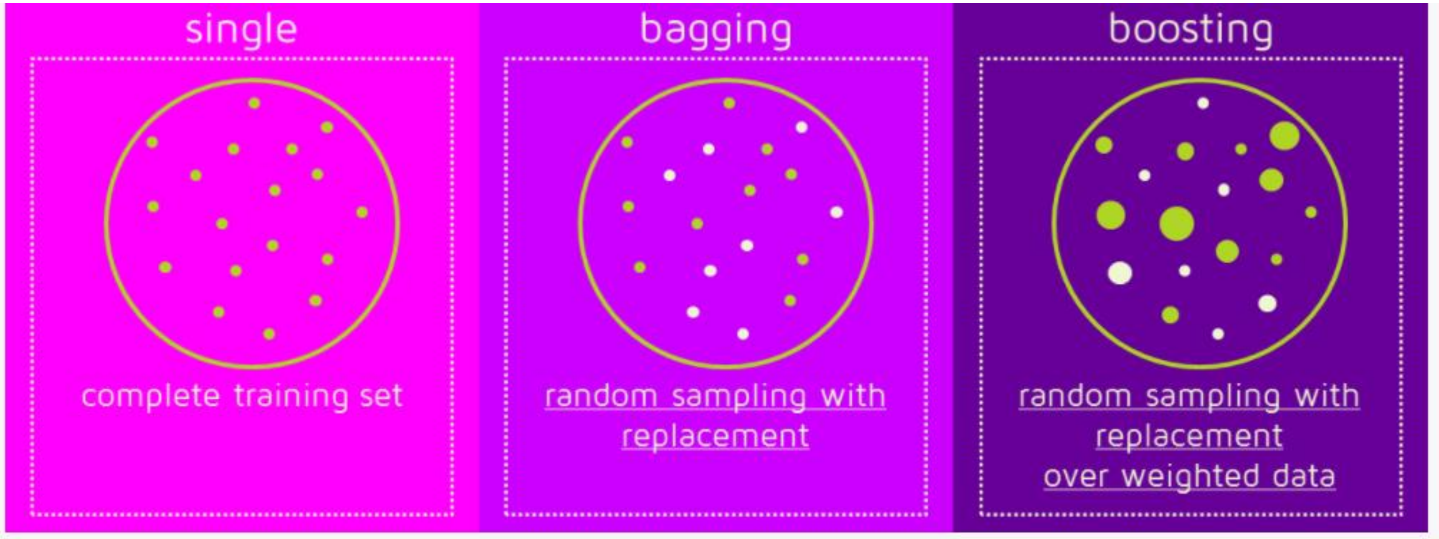
(b) Hypothesis space of an ensemble classifier

Fig. 2: A schematic illustration of hypothesis space partitioning with ensemble of classifiers. By combining moderate accurate base classifiers, we can approximate the best classification rule h_{best} with the increase of model complexity. This can be achieved by combining base classifiers with averaging or majority voting which takes advantage of the overlapped region.

Yang et al. (2010)

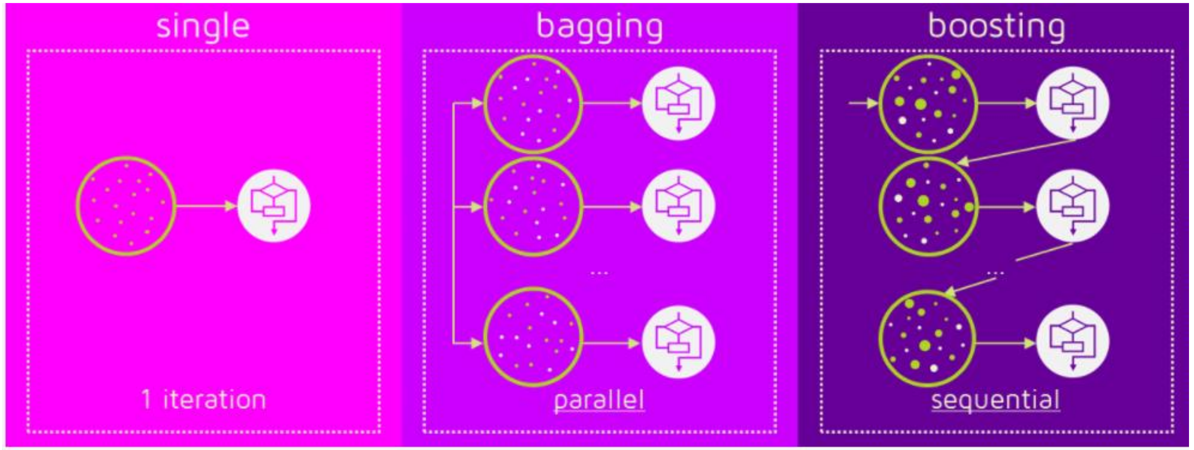
Ensemble methods

- N new training data sets are produced by **random sampling with replacement** from the original set.
- **Bagging**: any element has the same probability to appear in a new data set.


single
complete training set
bagging
random sampling with replacement
boosting
random sampling with replacement over weighted data

- **Boosting**: the observations are weighted and therefore some of them will take part in the new sets more often

Ensemble methods



Bagging

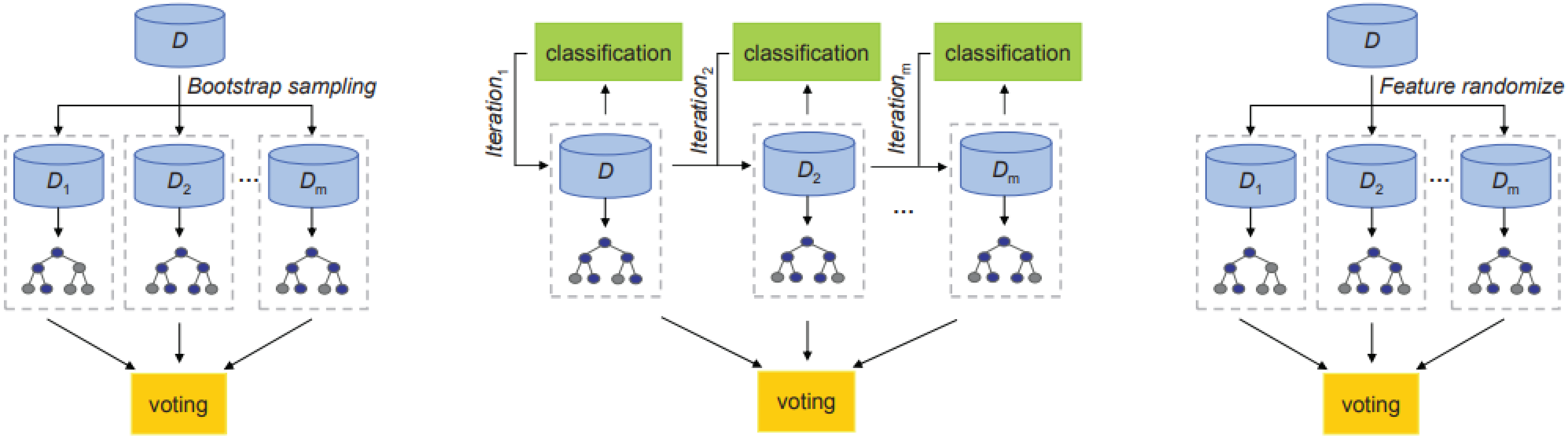
Training stage is parallel i.e., each model is built independently

Boosting:

- builds the new learner sequentially
- Each classifier is trained on data, taking into account the previous classifiers' success.
- After each training step, the weights are redistributed.
- **Misclassified data increases its weights** to emphasise the most difficult cases. In this way, subsequent learners will focus on them during their training.

Ensemble methods

Random forests are a special case of bagging



(a) bagging

(b) boosting

(c) random forests

Fig. 1: Schematic illustration of the three popular ensemble methods. Yang et al. (2010)

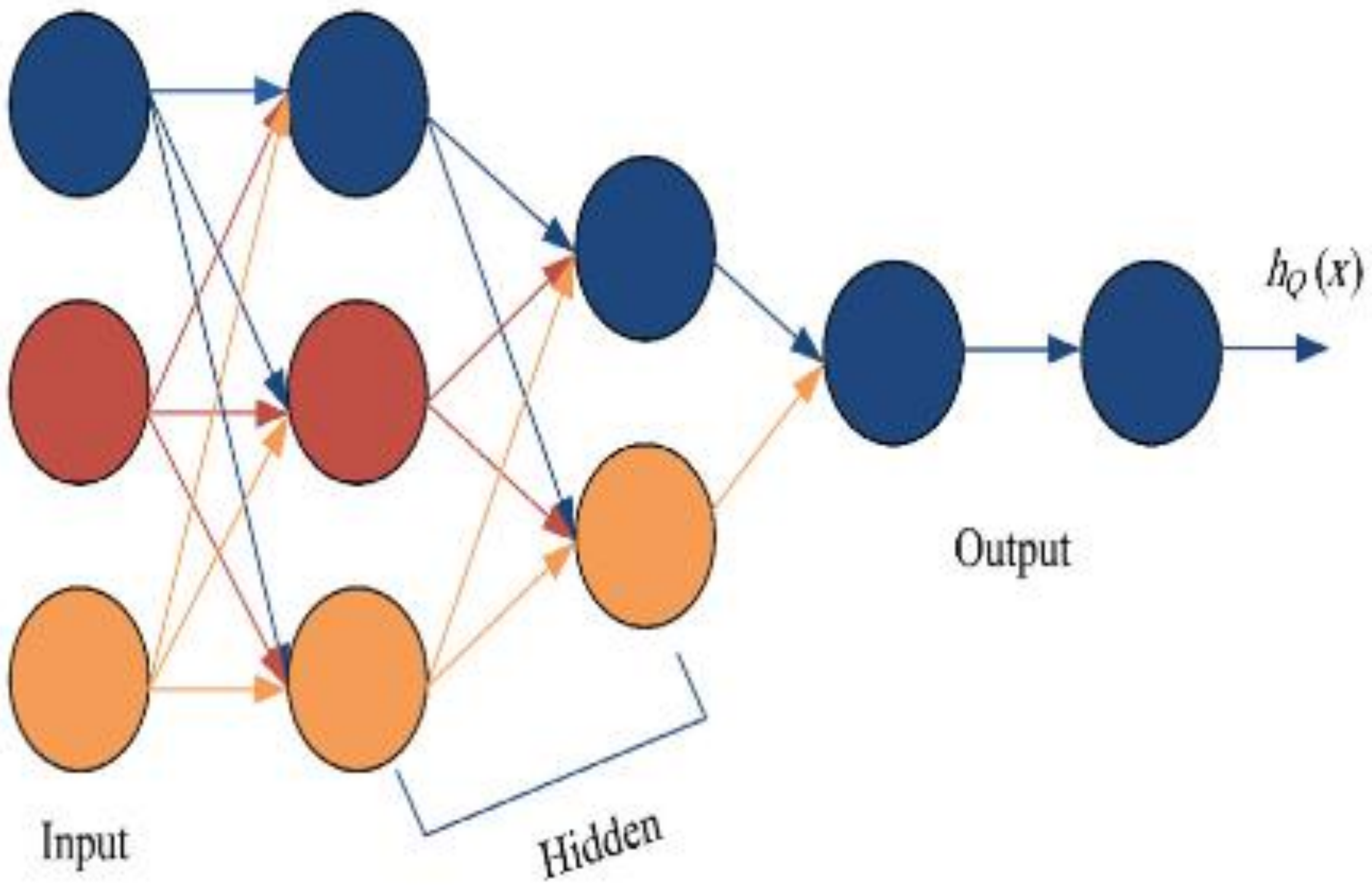
Random forest

- A collection of tree-structured classifiers {$h(\mathbf{x},\Theta_k)$, $k = 1, \ldots$} where the {Θk} are independent identically distributed random vectors and **each tree casts a unit vote** for the most popular class at input **x** (Breiman, L. 2001)
- Concept: each tree **cast a vote** for the classification of a new sample and the predicted probability vector is a result of **the most popular class** among the trees (Breiman, L. 2001)
- Prediction based on classification trees
- **Mtry, tuning parameter**:
  - **Number of ramdomly selected predictors** to choose from at **each split** in the tree.
  - Kept constant for each of the trees.
  - Model is however relatively insensitive to the values of mtry.
- Out-of-bag performance measures: accuracy, sensitivity, specificity, confusion matrices.

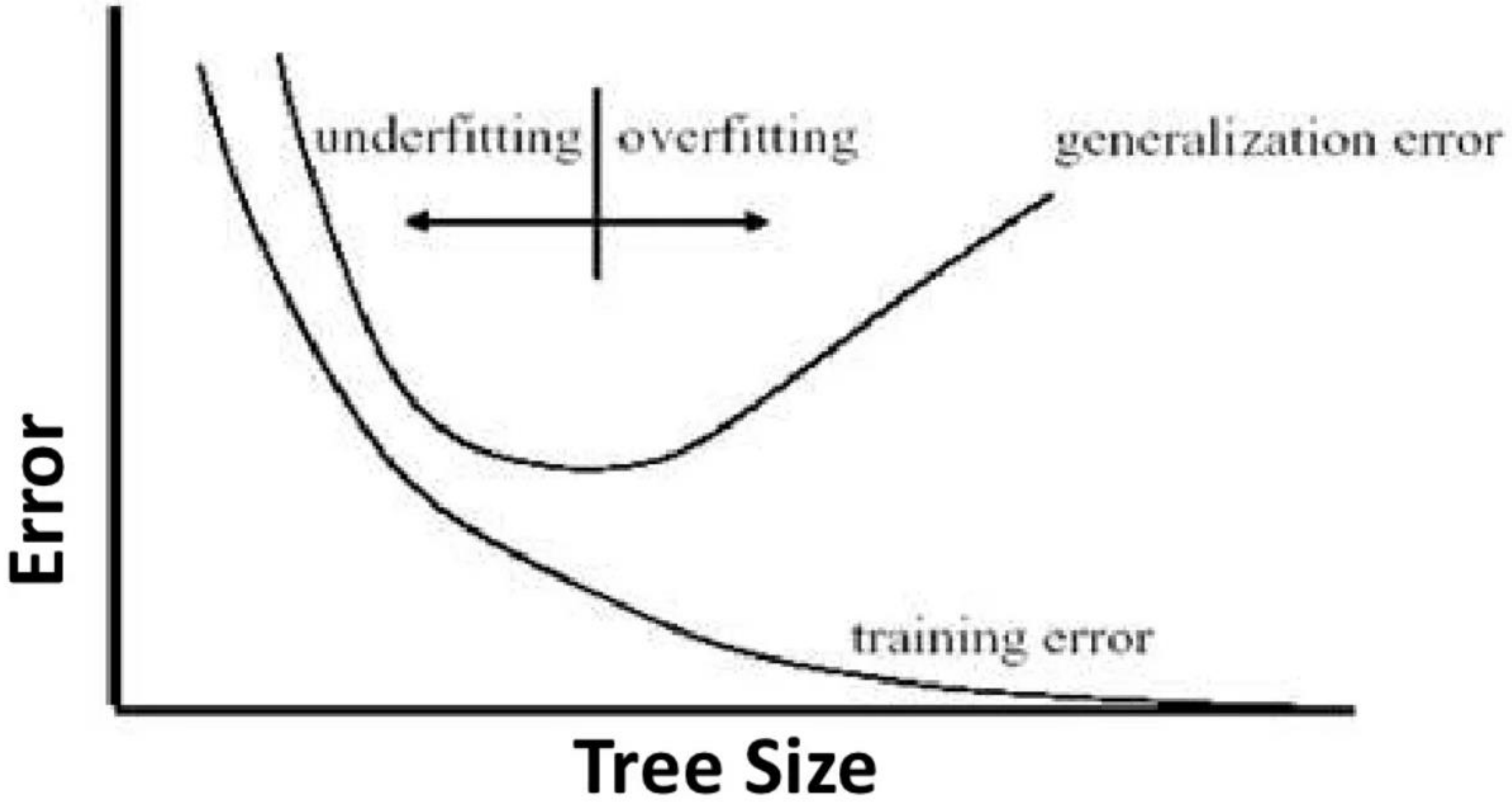
Gradient boosting machines (AdaBoost and LogitBoost)

- Boosting algorithms: additive models where **many weak classifiers are combined** (or boosted) into **a strong classifier** (Kuhn & Johnson 2013)
- Weak classifiers: trees that are **dependent on past trees** (in rf the trees are independent)
- Basic gradient boosting has **two tuning parameters**:
  - ➢ tree depth and number of interactions (Kuhn & Johnson 2013)
- LogitBoost: a boosted logistic regression
  - ➢additive on the logistic scale with the base learner providing the additive components (Friedman et al., 2000)

Deep learning: Neural networks



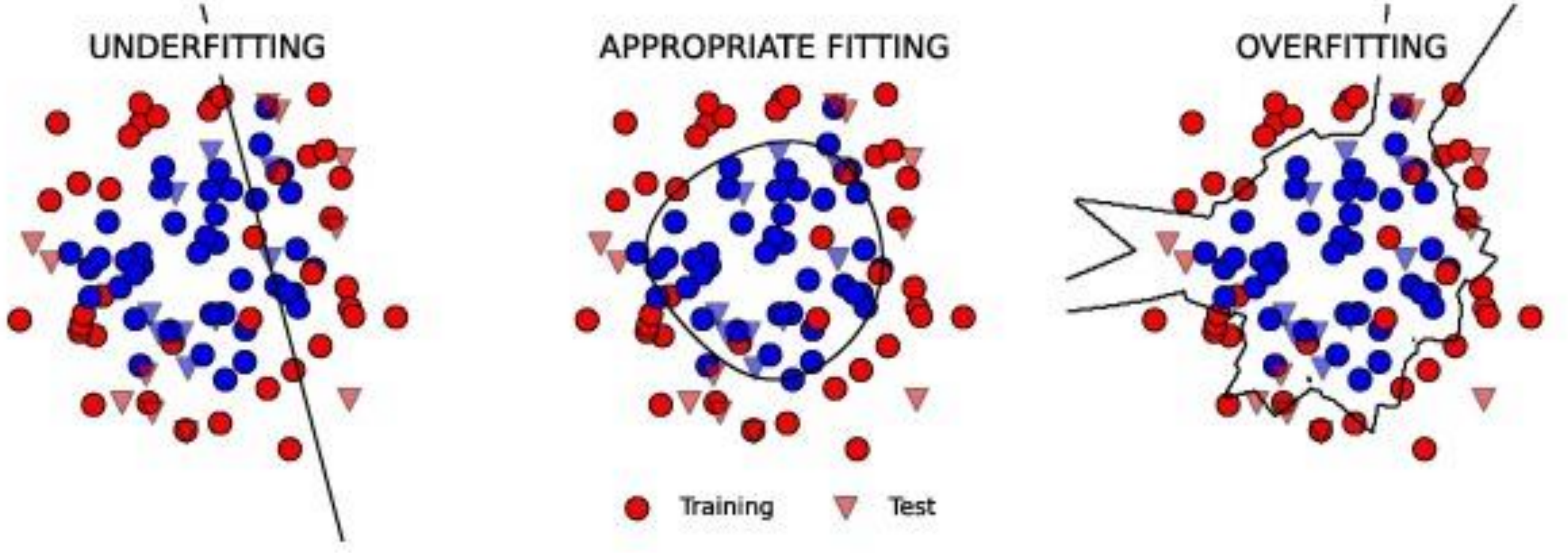
Overfitting and underfitting



Overtraining: means that it learns the training set too well – it overfits to the training set such that it performs poorly on the test set.

Underfitting: when model is too simple, both training and test errors are large

Overfitting and underfitting



Logistic Regression

Support Vector Machine with Gaussian kernel

K-NN

Monaco et al., 2021. Computational and Structural Biotechnology Journal, 19:4345-4359

Cross-validation

Resampling method that uses different portions of the data to test and train a model on different iterations.

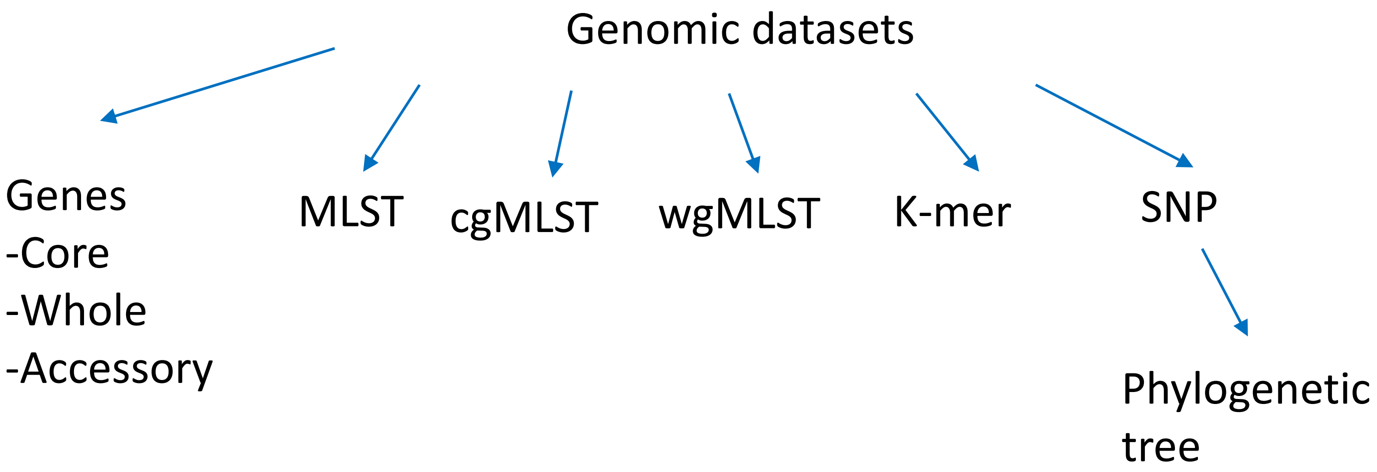
Training on *training dataset* and model tested on the validation dataset or *testing set*.

Goal: test the model's ability to predict new data that was not used in estimating it

Flags problems like overfitting or selection bias

NGS data for machine learning models

Bioinformatics



Genomic datasets

Genes
-Core
-Whole
-Accessory

MLST

cgMLST

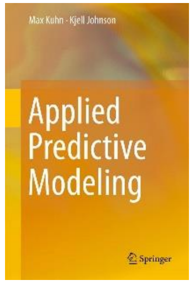
wgMLST

K-mer

SNP

Phylogenetic tree

References



Max Kuhn · Kjell Johnson

Applied Predictive Modeling

Springer

- 1. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer, 2013).
- 2. Yang, P., Hwa Yang, Y., B. Zhou, B. & Y. Zomaya, A. A Review of Ensemble Methods in Bioinformatics. *Curr. Bioinform.* **5**, 296–308 (2010).
- 3. Ren, Y., Zhang, L. & Suganthan, P. N. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. *IEEE Comput. Intell. Mag.* **11**, 41–53 (2016).
- 4. Friedman, J., Hastie, T. & Tibshirani, R. Special invited paper. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **28**, 337–374 (2000).
- 5. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).