Unsupervised and supervised machine learning

Introduction to multi-omics data analysis workshop

14. January 2022

Matti Ruuskanen

Postdoctoral researcher,

UNIVERSITY OF TURKU

Topics of the lecture

- What is machine learning?
- Unsupervised learning
  - Dimensionality reduction
  - Clustering
- Supervised learning
  - Classification
  - Regression
- Data curation
- Cross-validation
- Evaluation
- Issues & solutions

What is machine learning?

- In statistics, a project-specific probability model is fitted and quantitative measures of confidence are calculated

- In machine learning, general purpose learning-algorithms are applied on data to find patterns and perform predictions

What is machine learning?

- Machine learning methods are best suited for cases where we have more input variables than samples
- Often both (frequentist) statistics and ML produce comparable results
- Different types of analyses compliment each other in microbiome data science

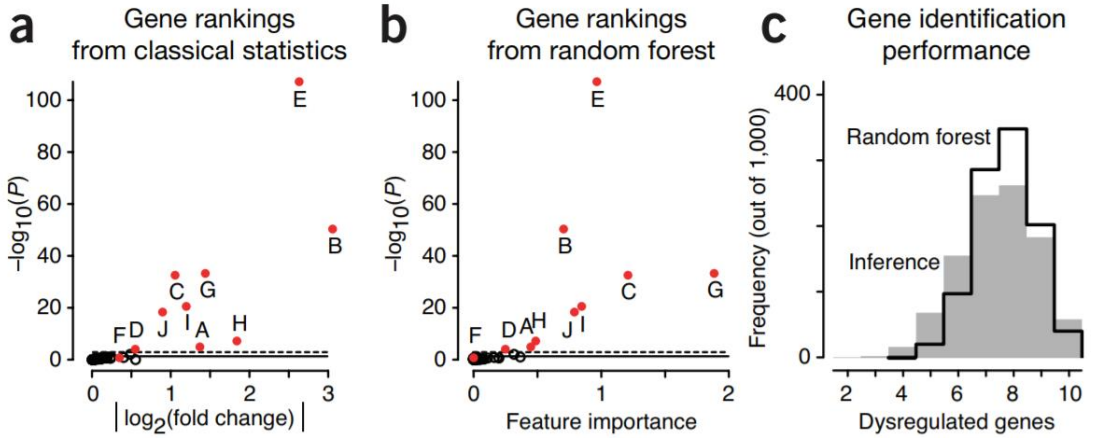


Figure 2 | Analysis of gene ranking by classical inference and ML. (**a**) Unadjusted log-scaled *P* values from statistical differential expression analysis as a function of effect size, measured by fold change in expression. (**b**) Log-scaled *P* values from **a** as a function of gene importance from random forest classification. In **a** and **b**, red circles identify the ten differentially expressed genes from **Figure 1**; the remaining genes are indicated by open circles. (**c**) Distribution of the number of dysregulated genes correctly identified in 1,000 simulations by inference (gray fill) and random forest (black line).

Bzdok et al. (2018); https://www.nature.com/articles/nmeth.4642

What is machine learning?

Variable	Supervised learning	Unsupervised learning
Discrete	Classification or categorization	Clustering
Continuous	Regression	Dimensionality reduction

What do we already know of unsupervised machine learning?

Go to www.menti.com and use the code 6517 2821

What is unsupervised machine learning?

- Unsupervised learning is used to detect patterns in data
- “Data mining”
- Can be divided into
  - Dimensionality reduction / ordination
    - ➢ Summarizing data in lower dimensions
  - Clustering
    - ➢ Finding groups

Dimensionality reduction / ordination

- Reduces dimensions of the data
- Preserves relationships between samples as well as possible
  - Projection of the data into lower dimensions
  - Aim is the “compression” of the variables
- New dimensions can be used as (proxy) variables
- Other variables can be fit in the ordination space to examine their relationship with the compressed data

Commonly used ordination methods

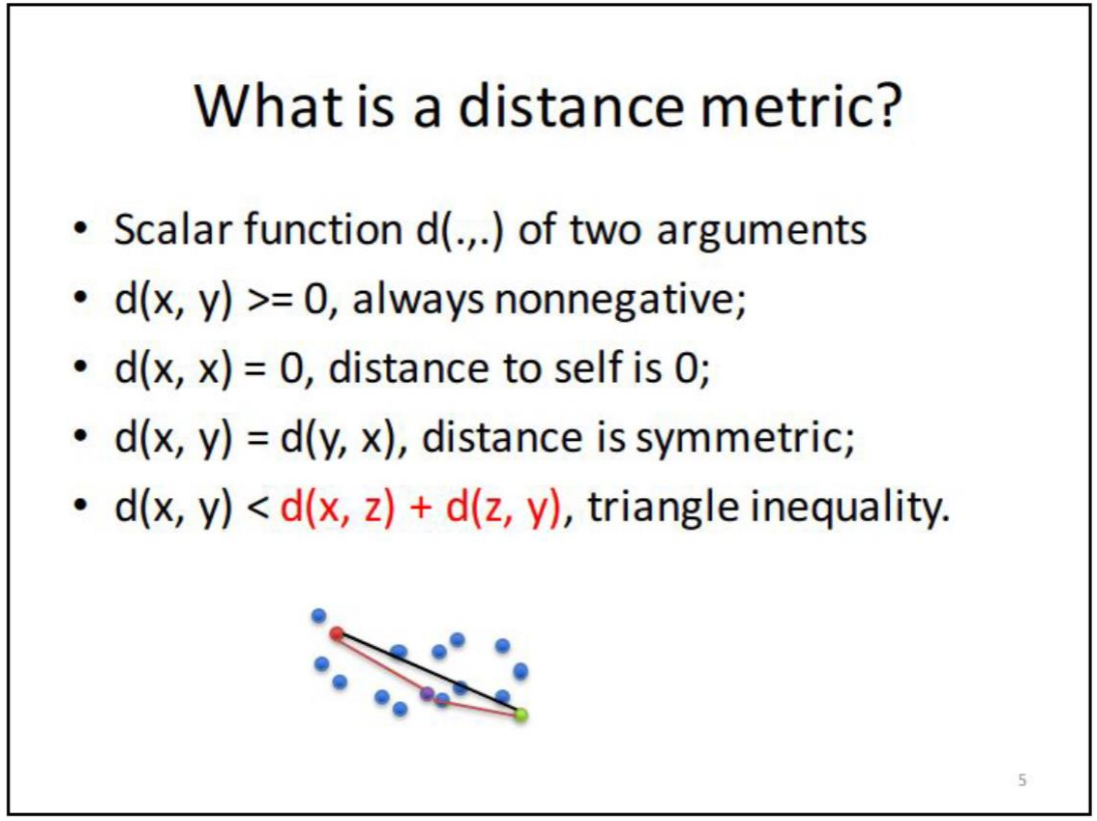
- Correspondence Analysis (CA)
  - Based on chi-square distance

- Multidimensional Scaling
  - Based on any (dis)similarity / distance
    - ➢ Many metrics can be used for microbiome data
  - Principal Coordinates Analysis (PCoA)
    - Tries to capture most of the variation in the (dis)similarity matrix in the first few axis
  - Non-metric MultiDimensional Scaling (NMDS)
    - Non-parametric rank-based method (very robust)
    - Tries to represent the pairwise dissimilarity most accurately usually in 2-d space

Commonly used ordination methods

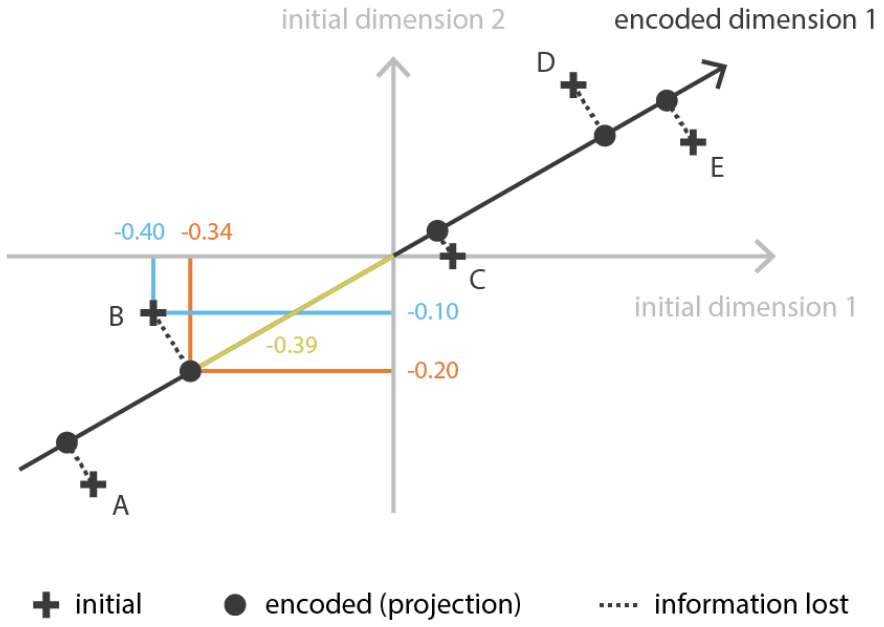
- Correspondence Analysis (CA)
  - Based on chi-square distance

- Multidimensional Scaling
  - Based on any **(dis)similarity / distance**
    - ➢ Many metrics can be used for microbiome data
  - Principal Coordinates Analysis (PCoA)
    - Tries to capture most of the variation in the (dis)similarity matrix in the first few axis
  - Non-metric MultiDimensional Scaling (NMDS)
    - Non-parametric rank-based method (very robust)
    - Tries to represent the pairwise dissimilarity most accurately usually in 2-d space



Principal Component Analysis

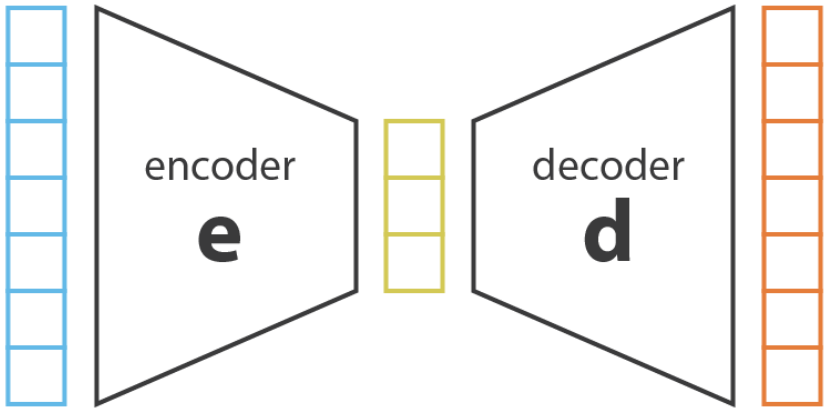
- Principal Component Analysis (PCA)
  - Based on linear combinations of variables between samples – a 'rotation' of the data
  - Tries to maximize variability in the first dimensions (which can still be visualized)
  - Linearity is preserved, so the dimensions can be used in further analyses



Point	Initial	Encoded	Decoded
A	(-0.50, -0.40)	-0.63	(-0.54, -0.33)
B	(-0.40, -0.10)	-0.39	(-0.34, -0.20)
C	(0.10, 0.00)	0.09	(0.07 0.04)
D	(0.30, 0.30)	0.41	(0.35, 0.21)
E	(0.50, 0.20)	0.53	(0.46, 0.27)

https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Autoencoders

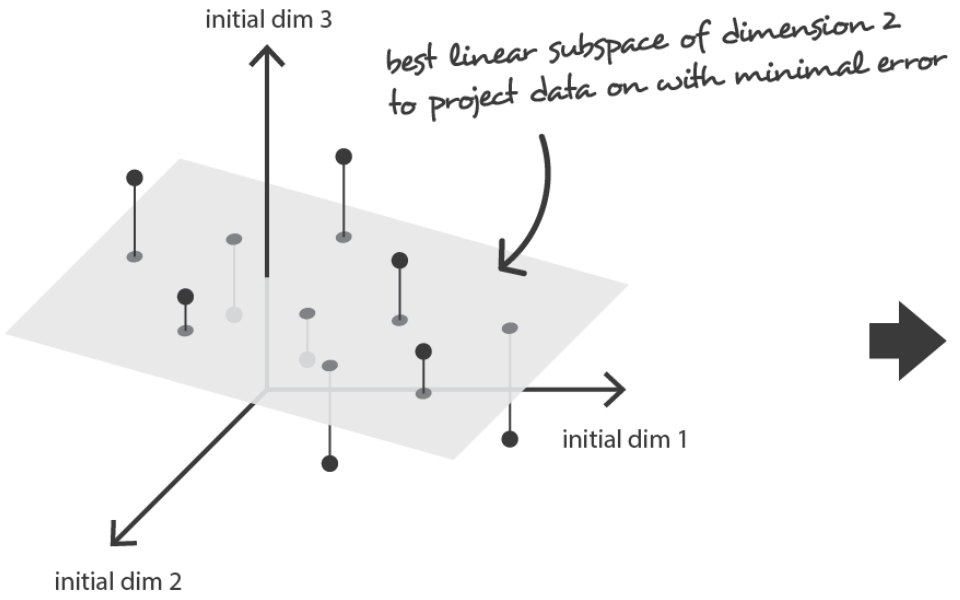


$x = d(e(x))$ ➡ **lossless encoding**
no information is lost when reducing the number of dimensions

$x \neq d(e(x))$ ➡ **lossy encoding**
some information is lost when reducing the number of dimensions and can't be recovered later

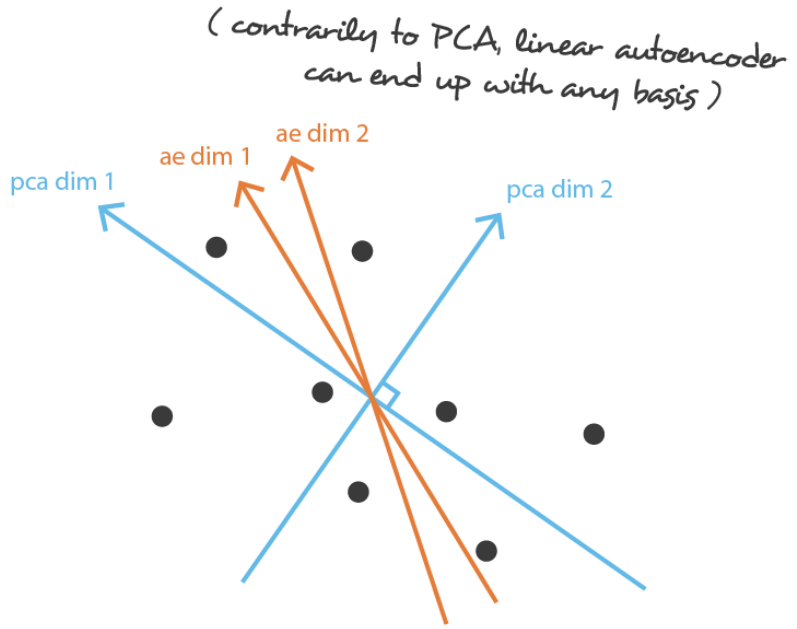
https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Autoencoders



Data in the full initial space

In order to reduce dimensionality, PCA and linear autoencoder target, in theory, the same optimal subspace to project data on...



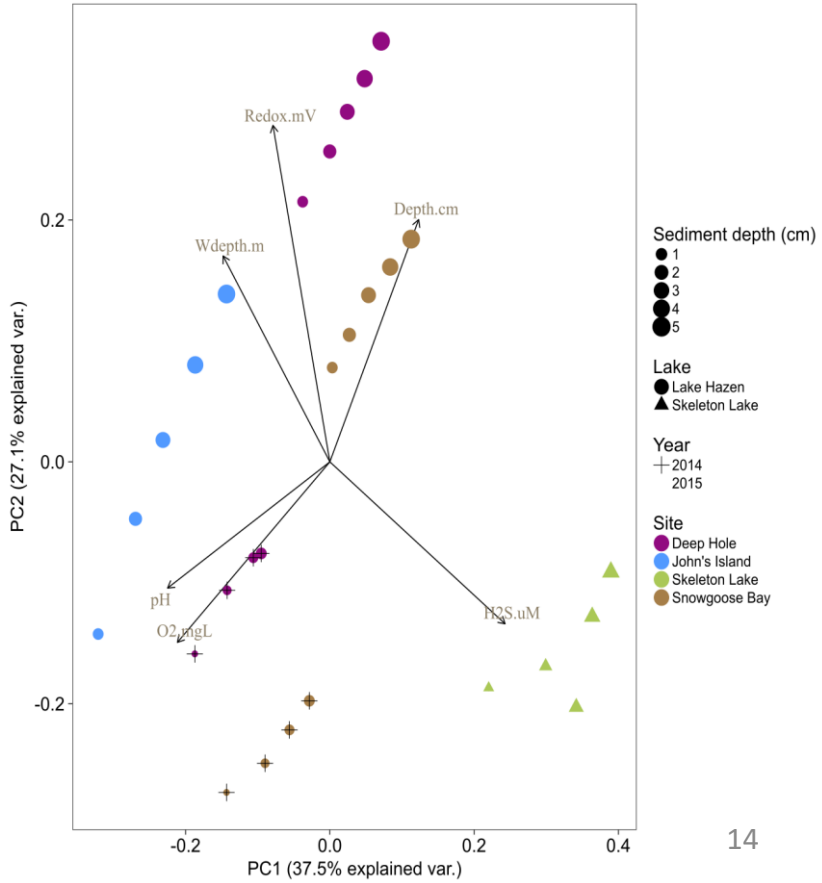
Data projected on the best linear subspace

... but not necessarily with the same basis due to different constraints (in PCA the first component is the one that explains the maximum of variance and components are orthogonal)

https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

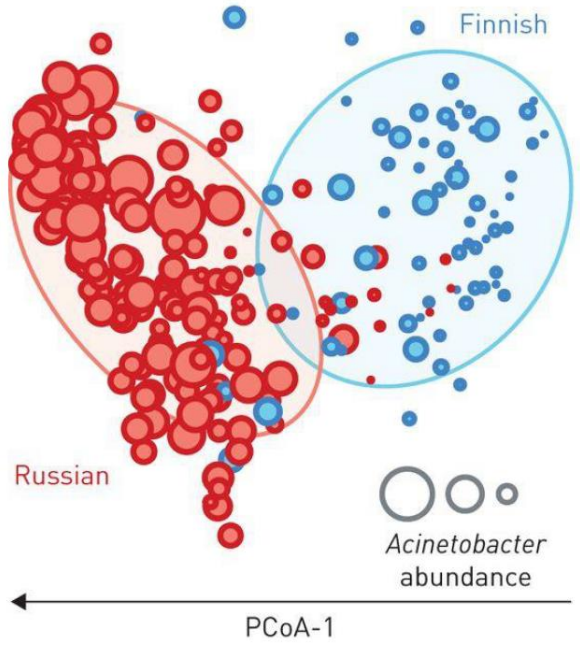
Use examples of dimensionality reduction

PCA of variables in
lake sediment samples

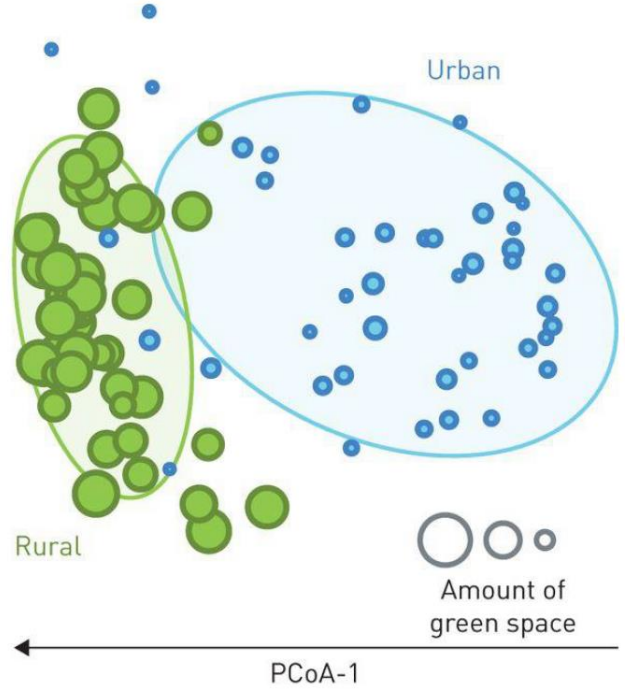


Ruuskanen et al. (2018)

a) Differences in nasal microbiota between adolescents from Finnish and Russian Karelia



b) Differences in skin microbiota between urban and rural children



European Respiratory Journal 2017 49: 1700481; **DOI:** 10.1183/13993003.00481-2017

Ruokolainen et al. (2017)

tSNE and UMAP

- Non-linear transformations of the data to minimize distance between similar points and maximize distance between groups
- Data is projected into lower dimensions
- Highly stochastic – result is dependent on hyperparameters
- Demo: https://pair-code.github.io/understanding-umap/

Clustering

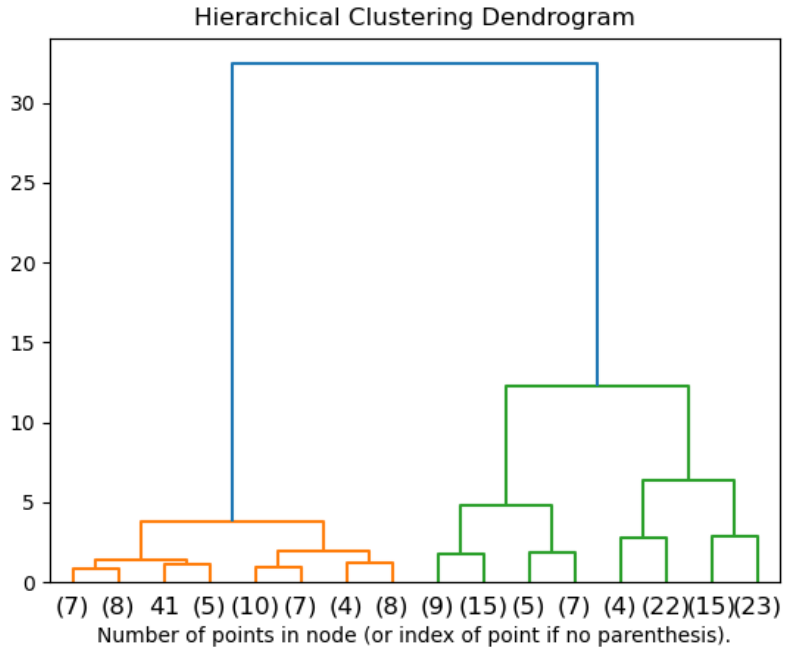
- Aims to find groups of similar features or samples
- Different feature or sample sets can be used for clustering
- Can be used to *e.g.,* detect biologically meaningful patterns
  - Cell types clustered by gene expression
  - Groups of microbial taxa associated with a disease

Hierarchical clustering

- Agglomerative clustering
  - Observations start as individual clusters, which are gradually merged
- Divisive clustering
  - Observations start as one single cluster, which is gradually divided

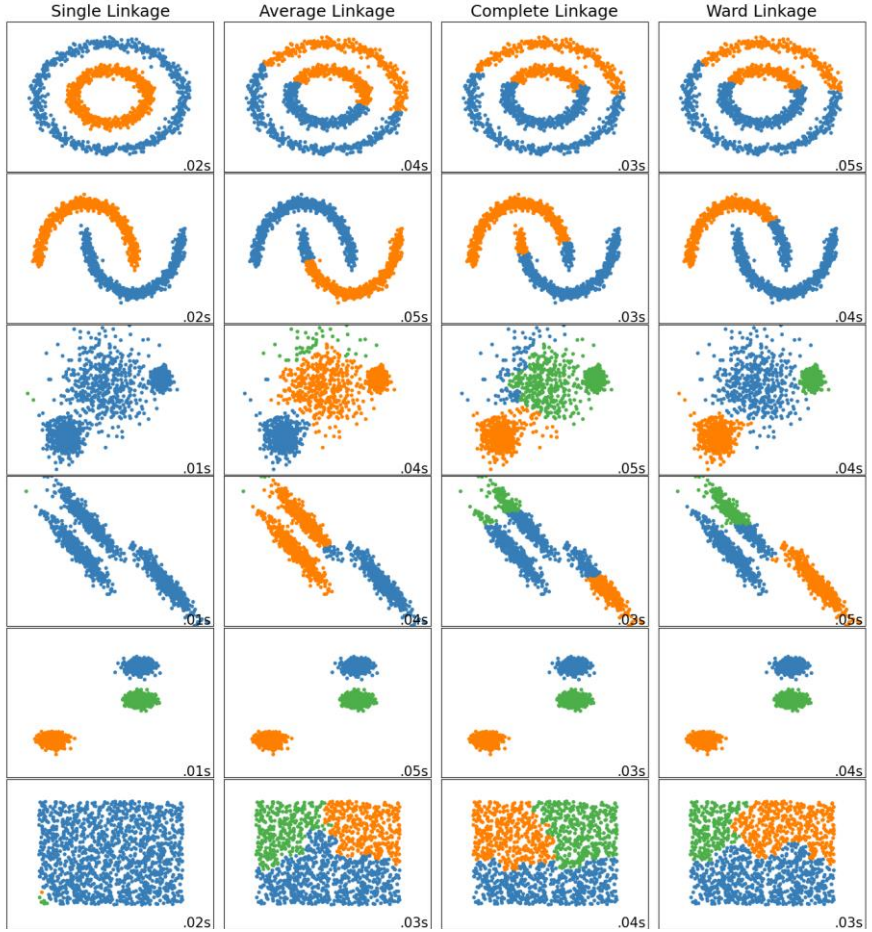
Hierarchical clustering

Agglomerative clustering with Ward linkage (min SS)



https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html

Results of different linkage functions vary



https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html

Biclustering

- Both rows and columns of a data matrix are clustered simultaneously

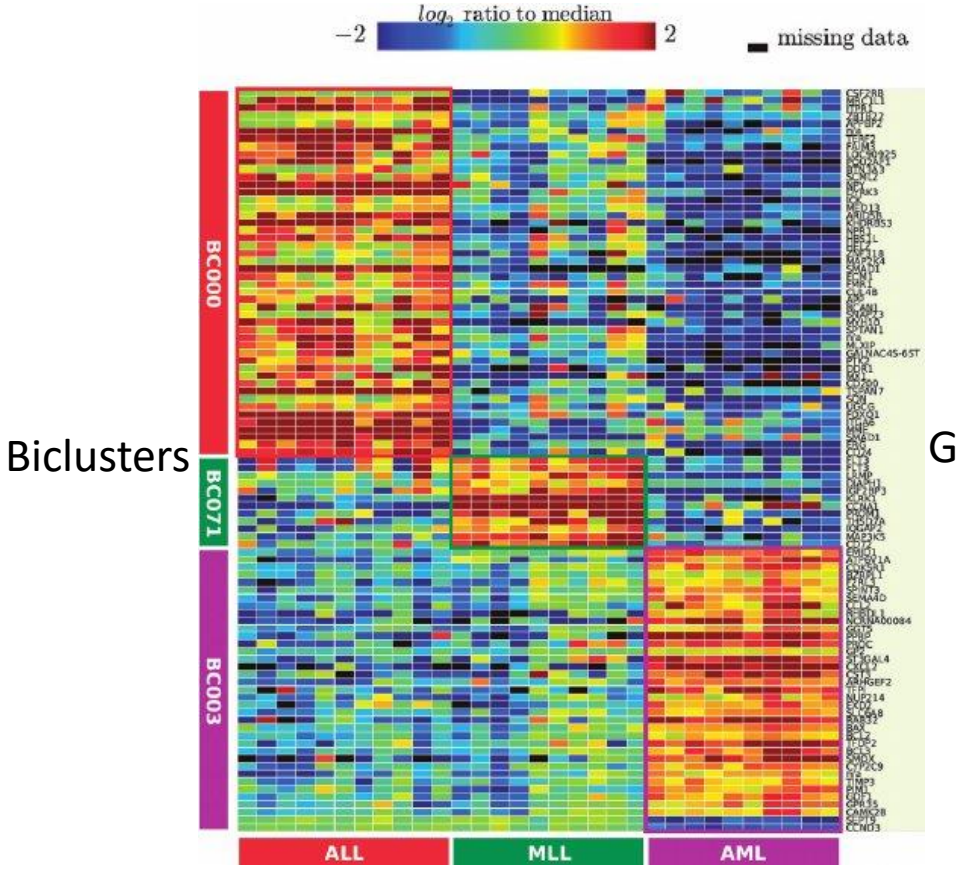
- Aim is to find groups of co-occuring features (or biclusters)
  - ➢ Which metabolites co-occur with which microbial taxa?

- "Biclusters are subspaces where a subset of rows … exhibit a correlated pattern over a subset of columns …"
  (Henriques & Madeira, 2015; https://doi-org/10.1109/TCBB.2014.2388206)

- Various algorithms exist, which re-organize the rows and columns to form biclusters

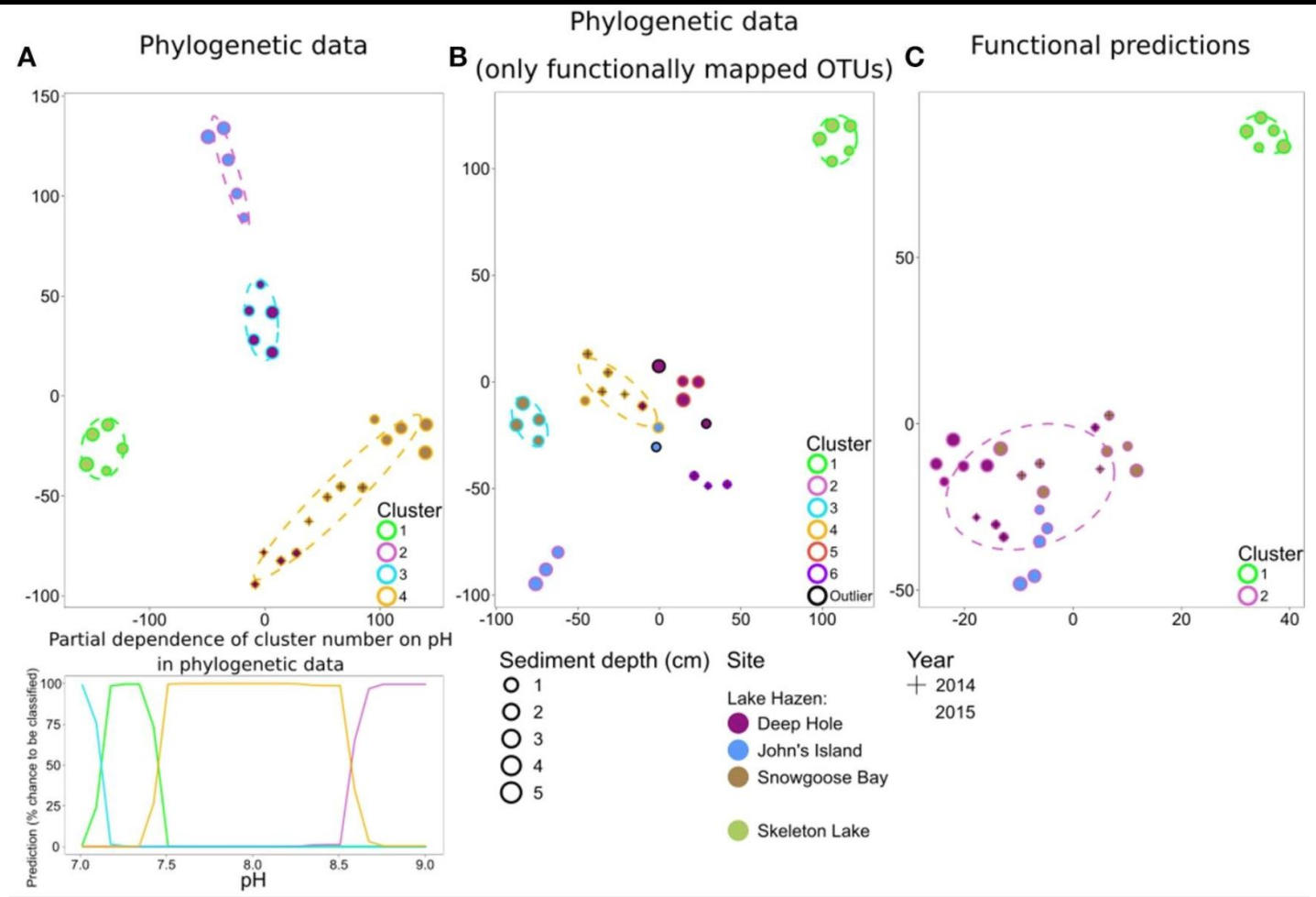
Biclustering

- Both rows and columns of a data matrix are clustered simultaneously
- Aim is to find groups of co-occuring features (or biclusters)
  - ➢ Which metabolites co-occur with which microbial taxa?
- "Biclusters are subspaces where a subset of rows … exhibit a correlated pattern over a subset of columns …"
  (Henriques & Madeira, 2015; https://doi-org/10.1109/TCBB.2014.2388206)
- Various algorithms exist to re-organize the rows and columns to form biclusters



Li et al. (2009); http://dx.doi.org/10.1093/nar/gkp491

Use examples of clustering



Ruuskanen et al. (2018); https://www.frontiersin.org/articles/10.3389/fmicb.2018.01138

Evaluation of clustering - Silhouettes

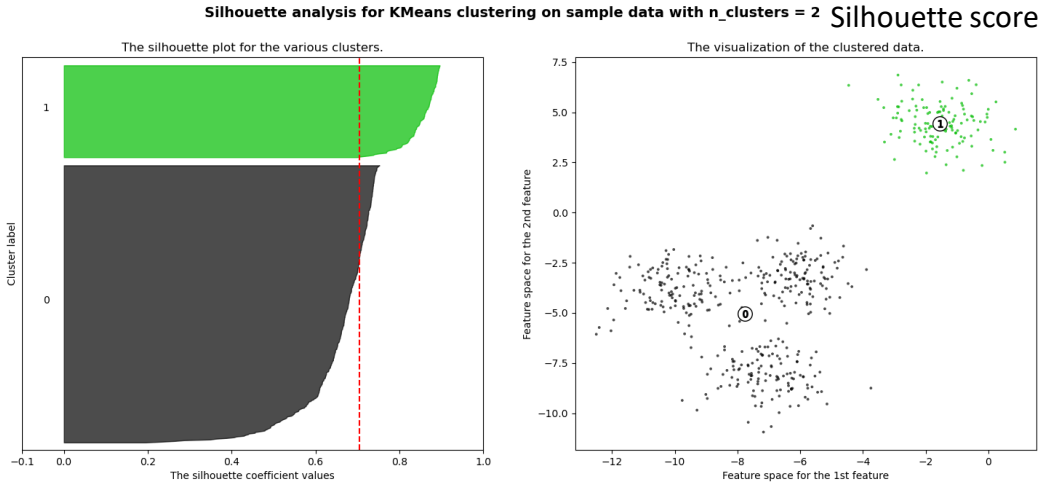
- Silhouette plots and Silhouette coefficients can be used to estimate the quality of the clustering

- Silhouette coefficient *s(i)*:

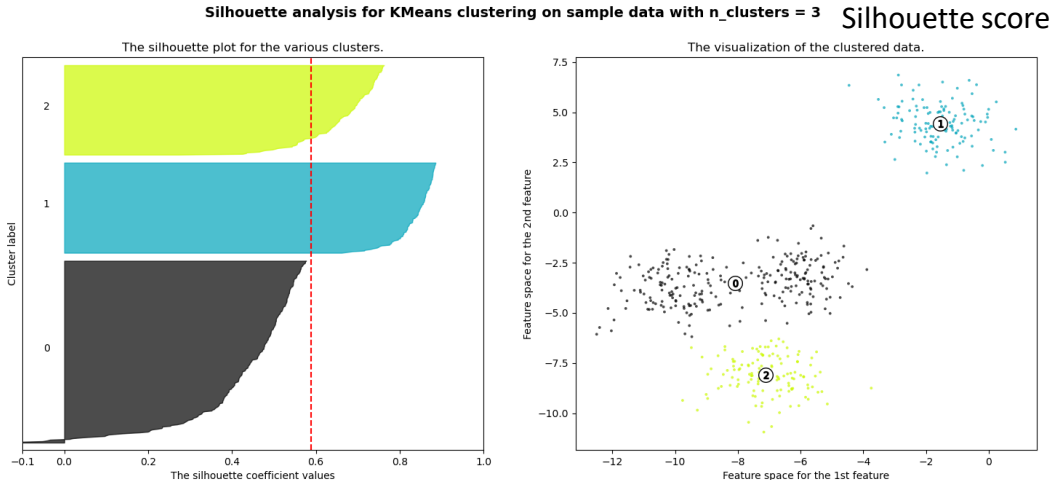
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- *i* = data point in a cluster
- *a(i)* = average distance of *i* to all points in the **same** cluster
- *b(i)* = average distance of *i* to all points in the **nearest** cluster

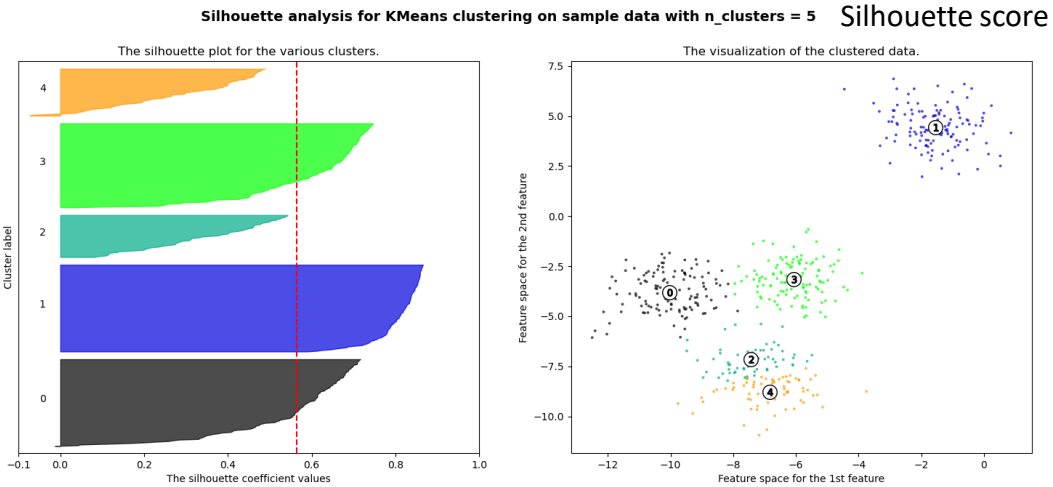
Evaluation of clustering - Silhouettes



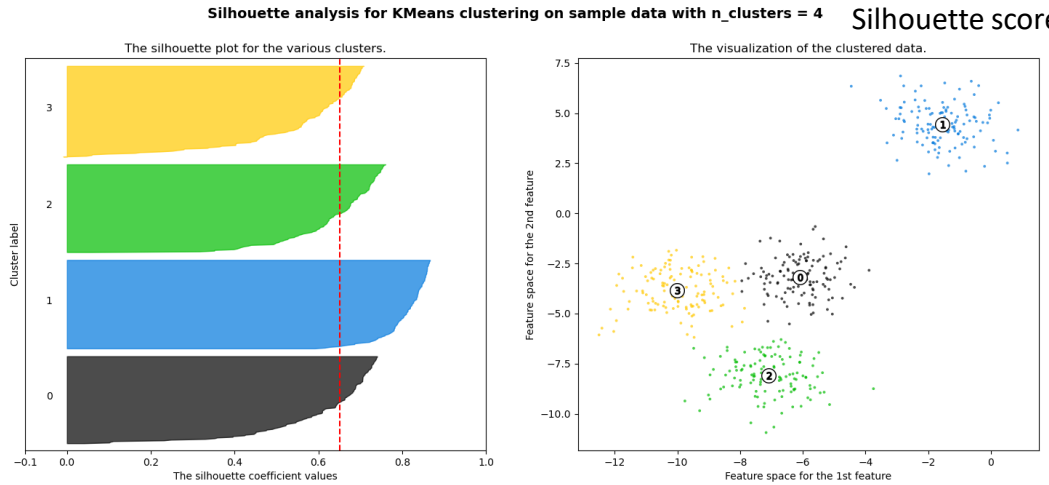
Silhouette score is : 0.70



Silhouette score is : 0. 65



Silhouette score is : 0. 59



Silhouette score is : 0. 56

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Where does supervised ML fit?

Variable	Supervised learning	Unsupervised learning
Discrete	**Classification or categorization**	Clustering
Continuous	**Regression**	Dimensionality reduction

What do we already know of supervised machine learning?

Go to www.menti.com and use the code

97 10 17 3

What is supervised machine learning?

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y, given x,
  or: $y = f(x)$

- Supervised ML methods are:
  - Often nonparametric -> flexible
  - Able to take interactions between features into account

Problems suited for supervised ML

- Labeled data
  - Examples where the true values of *y* are known
- Big data
  - Nearly all high-throughput sequencing data
  - Image and textual data (with applications in microbiology)
- Complex interactions
  - Data from microbial communities
- For example:
  - Spatial and temporal patterns
  - Disease diagnosis and prediction
  - Modeling of environmental interactions

“Doing” supervised machine learning

1. Know your system and data
   - Biological understanding?

2. Split to train and test / validation sets
   - Models are validated by predicting on previously unseen data
   - For example 70% train / 30% test

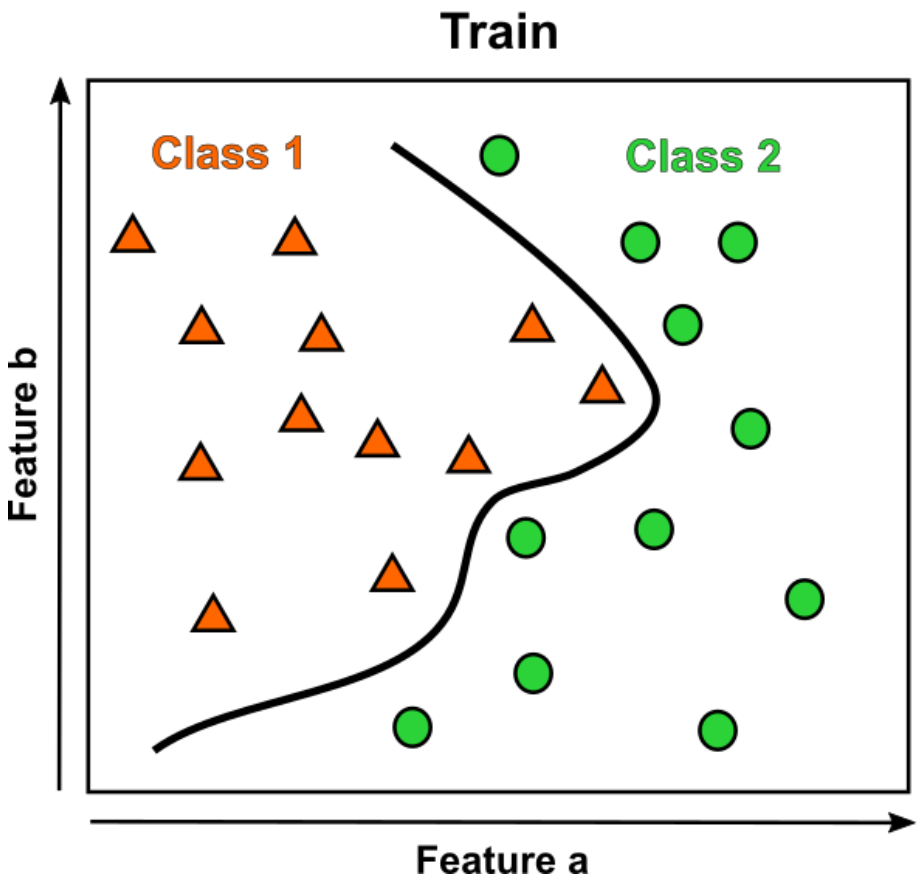
3. Selection and training of the first model

4. Adjusting and selecting features, model architectures and hyperparameters

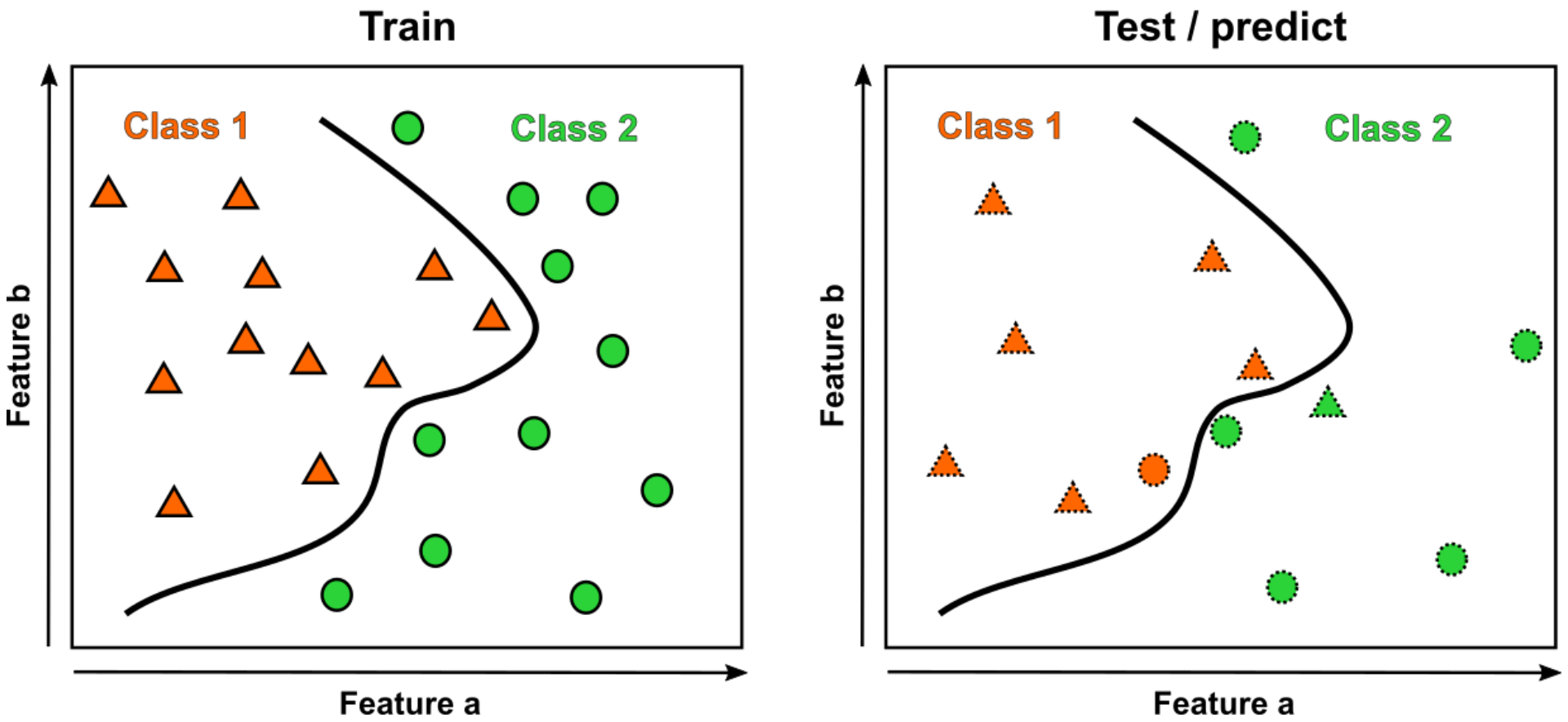
5. Final evaluation of the model with predictions on test data
   - Various performance metrics are available

6. (optional) Deployment of the model to production

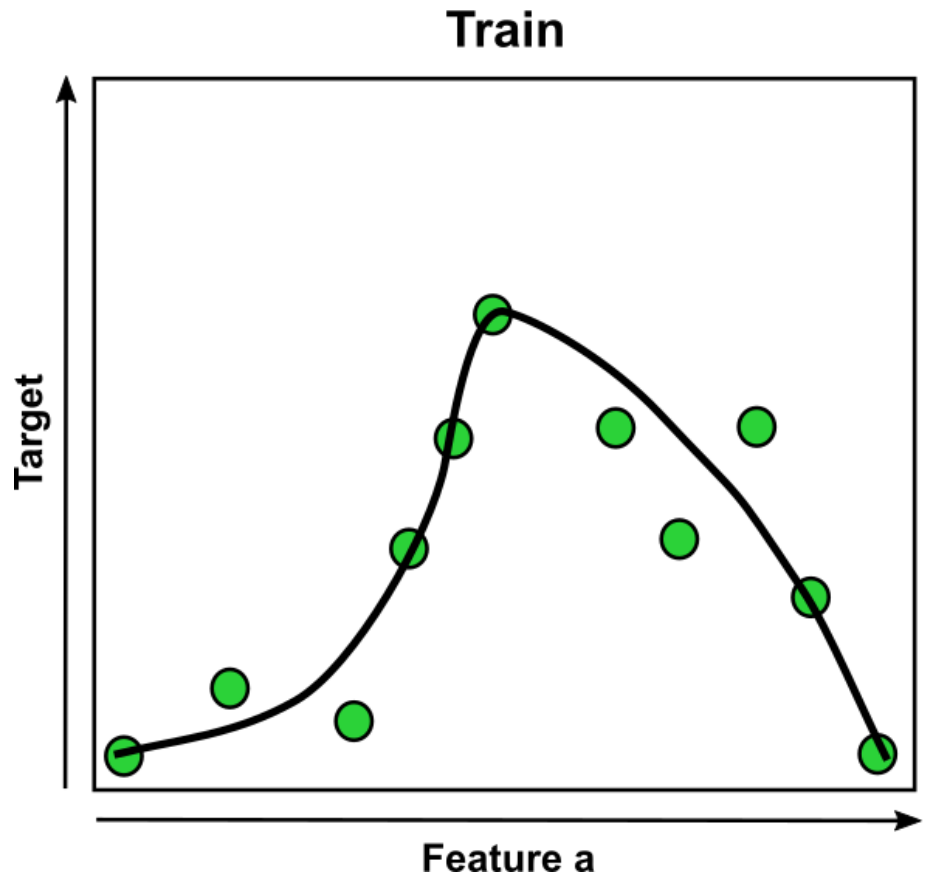
Classification



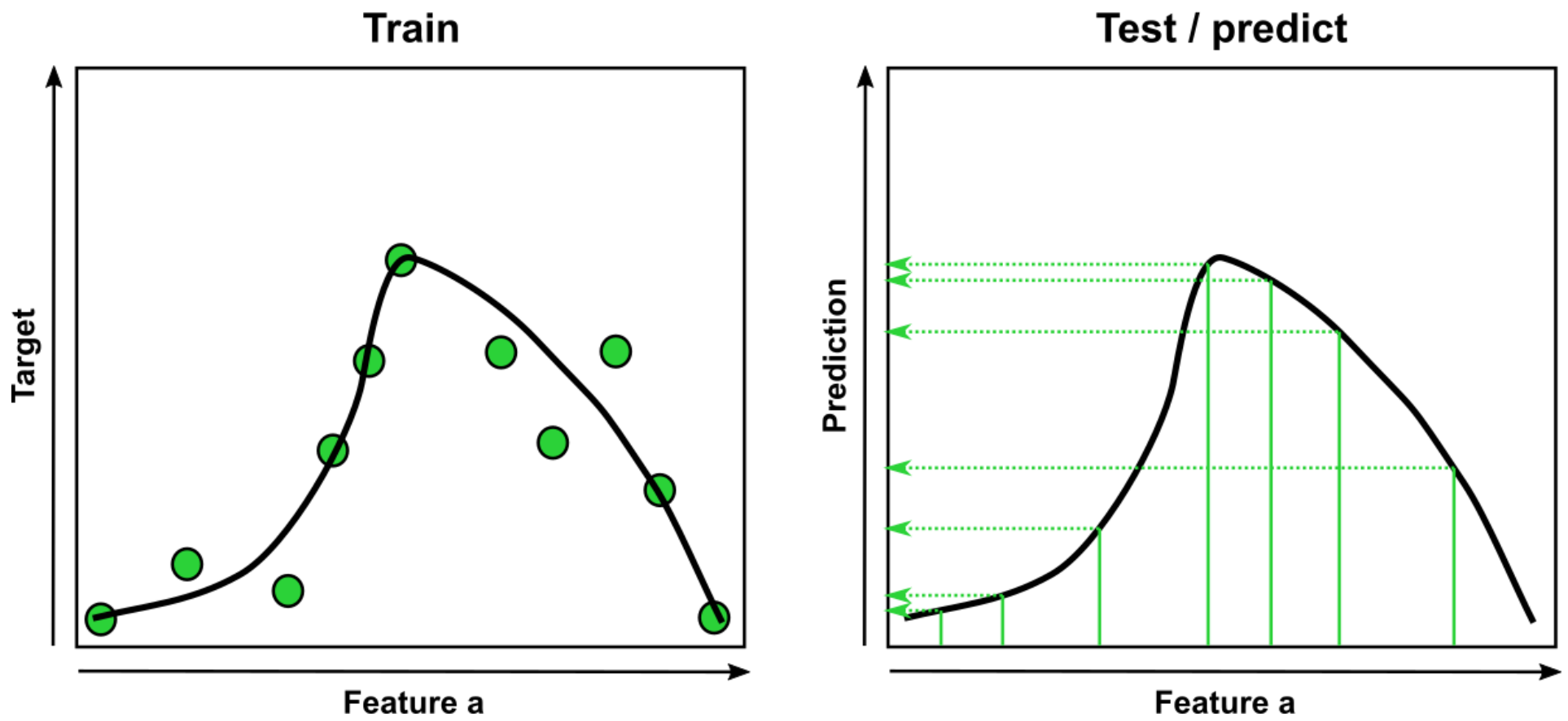
Classification



Regression



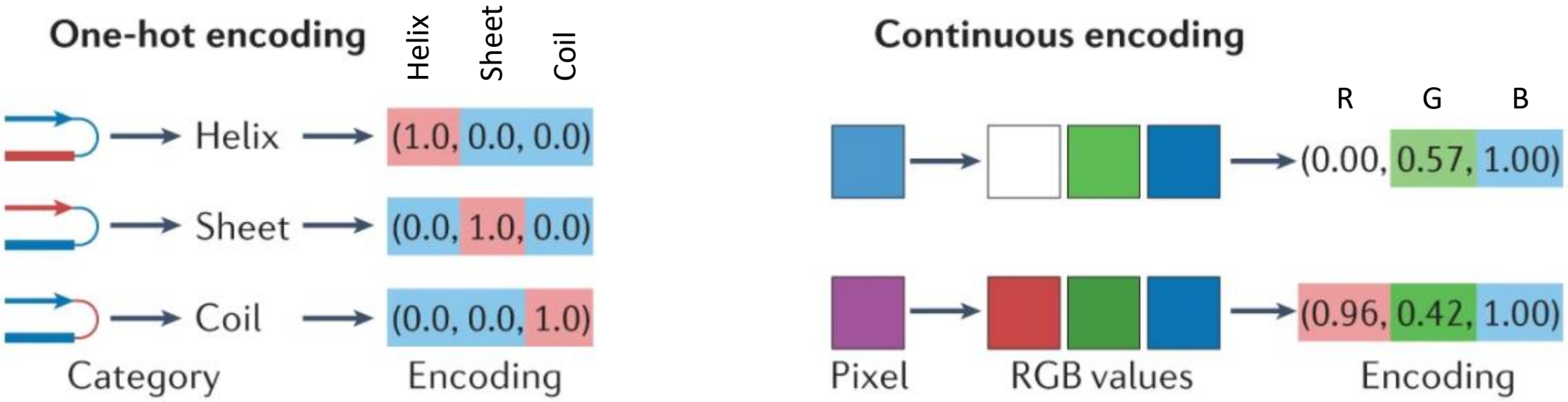
Regression



Data curation and annotation

- ‘garbage in, garbage out’, was first noted well over a century ago (Babbage, 1864); *Passages from the Life of a Philosopher.*
- Can the data be trusted?
- Labels need to be as accurate as possible
  - Manual annotation, or at least curation is often required
- Recoding and encoding of variables?
- In microbiome data, counts need to be compositionally transformed
  - For example, centered log-ratio, phylogenetic ILR…

Encoding



Adapted from Greener et al. (2021); https://www.nature.com/articles/s41580-021-00407-0

Cross-validation and evaluation

- How do we know if our model has learned something from the data?
- We want the model to focus on relevant features and **not fit to noise** (overfit)
- We are making a model for ***y = f(x)***, thus we can compare the true or known values of *y* against predicted values of *y*
- If we would test with the same data we trained with, the flexible models would be able to give perfect accuracy!

Cross-validation

- Testing / validation **needs to be done with new data**
  - Testing is conducted both during training and at final evaluation
- Multiple ways to conduct cross-validation (CV):
- Holdout: complete separation of train and test sets
  - For example, 70%/30% or 80%/20%
- Leave-p-out CV
- K-fold CV
- Spatial or temporal CV

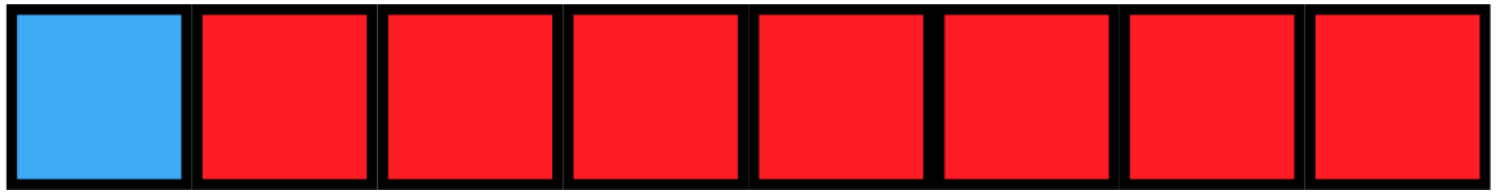
Leave-p-out cross-validation

$p = 1$
$n = 8$

Test

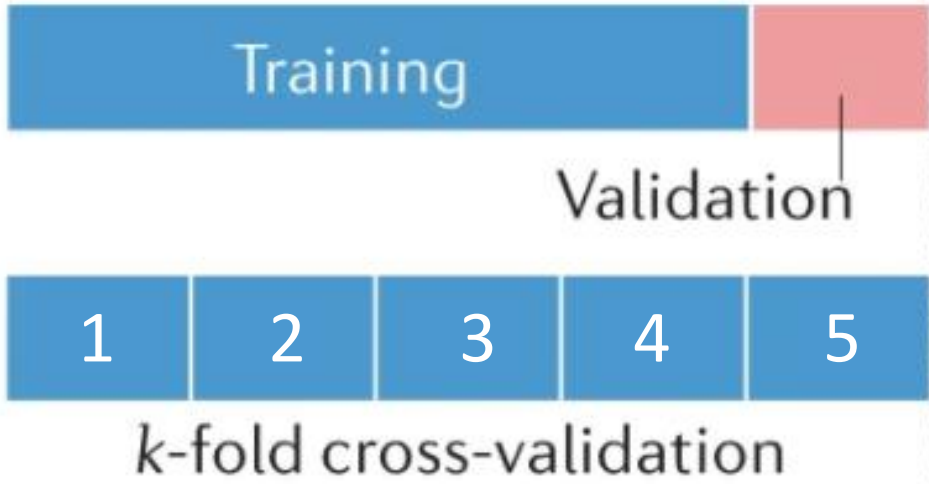
Train

Model 1



Banuelos (2020); https://commons.wikimedia.org/w/index.php?curid=87684543

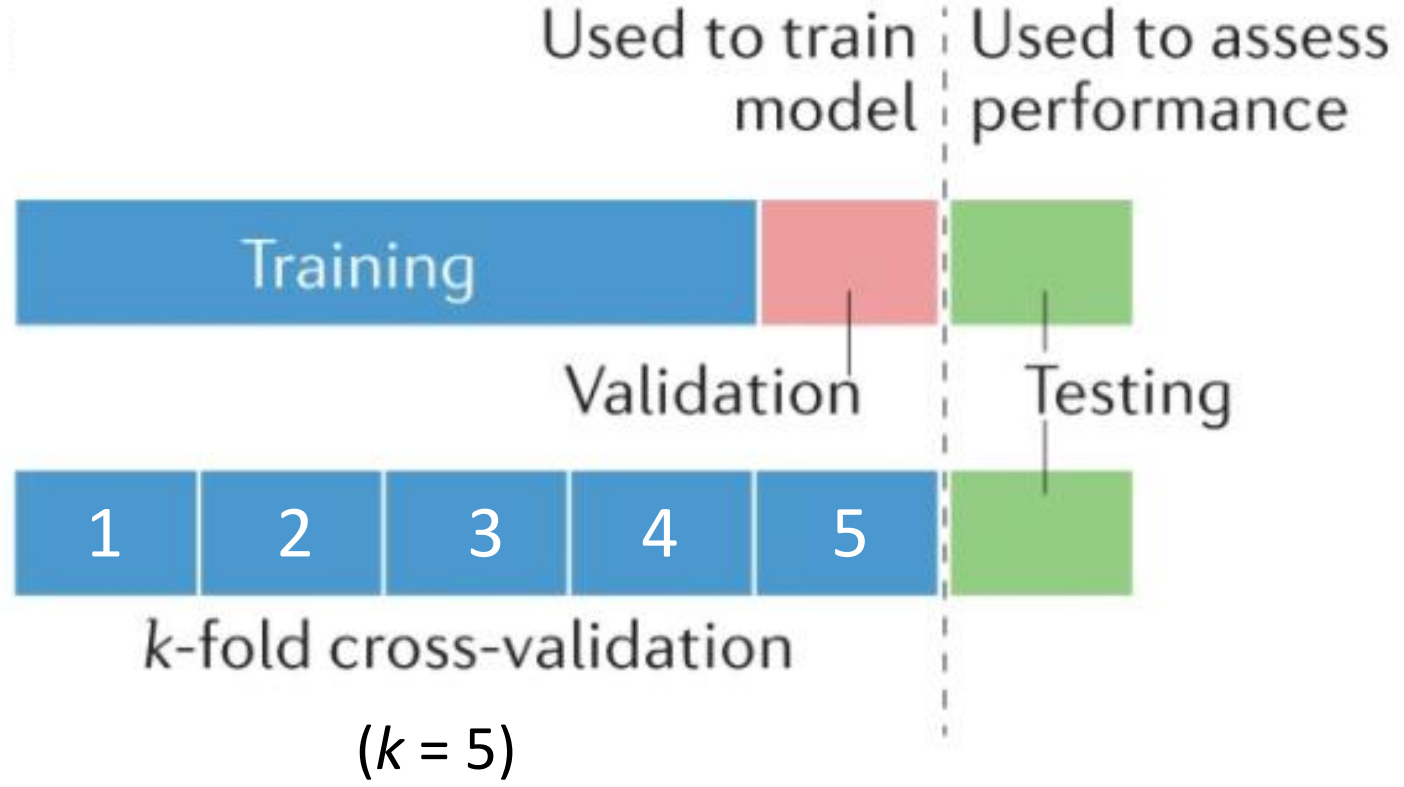
Simple holdout and k-fold CV



(k = 5)

Adapted from Greener et al. (2021); https://www.nature.com/articles/s41580-021-00407-0

Nested cross-validation



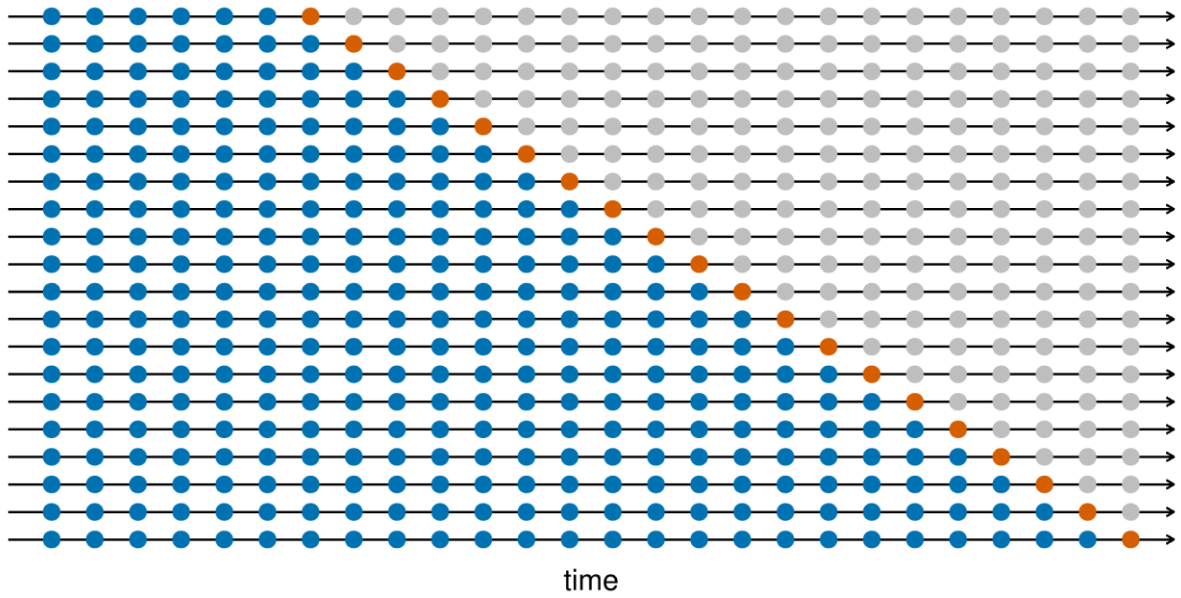
Adapted from Greener et al. (2021); https://www.nature.com/articles/s41580-021-00407-0

Spatial cross-validation

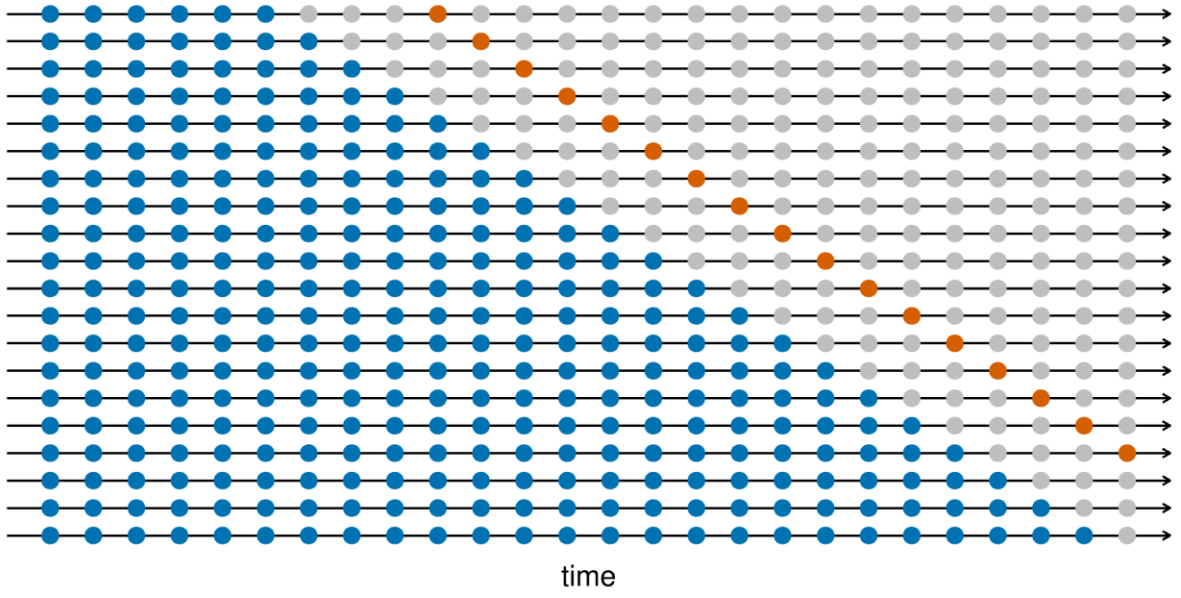
Lovelace et al. (2021); https://geocompr.robinlovelace.net/spatial-cv.html

Temporal cross-validation

One-step forecasting



Four-step forecasting



Hyndman & Athanasopoulos (2021); https://otexts.com/fpp3/tscv.html

Evaluation - Classification

Confusion matrix for binary classification

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP)	False Positive (FP) Type I error
	Negative	False Negative (FP) Type II error	True Negative (TN)

Evaluation - Classification

- Simple classification metrics:

1. Accuracy = Correct predictions / All cases
   - TP + TN / TP + FP + TN + FN

2. Sensitivity, or Recall = Correct positives / All actual positive cases
   - TP / TP + FN

3. Specificity = Correct negatives / All actual negative cases
   - TN / TN + FP

4. Precision = Correct positives / All predicted positive cases
   - TP / TP + FP

- Increasing *precision* reduces *recall* / Increasing *recall* reduces *precision*!

Evaluation - Classification

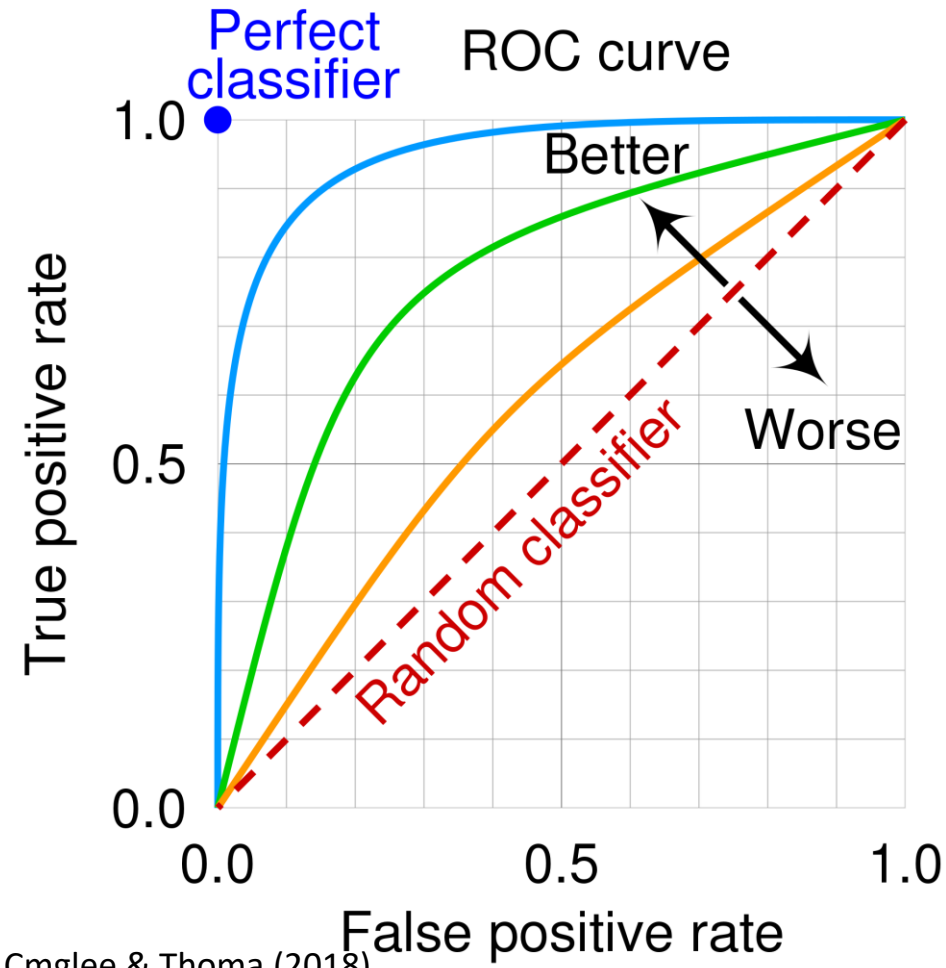
- Dealing with precision/recall tradeoff
- Some classification methods output probabilities between 0 and 1
- F1 score = Harmonic mean between precision and recall

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

  - Not sensitive to extremely large values in either one
  - Balances both metrics in a single figure
  - Does not account for true negatives (which might be important)

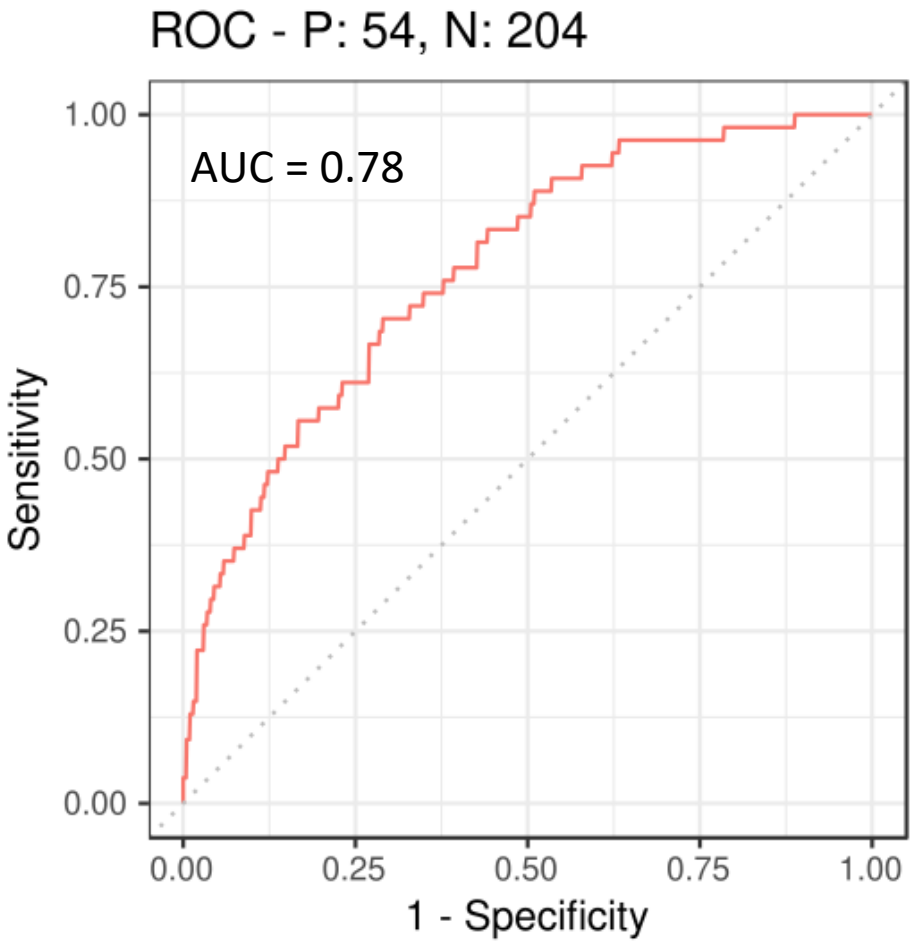
Evaluation - Classification

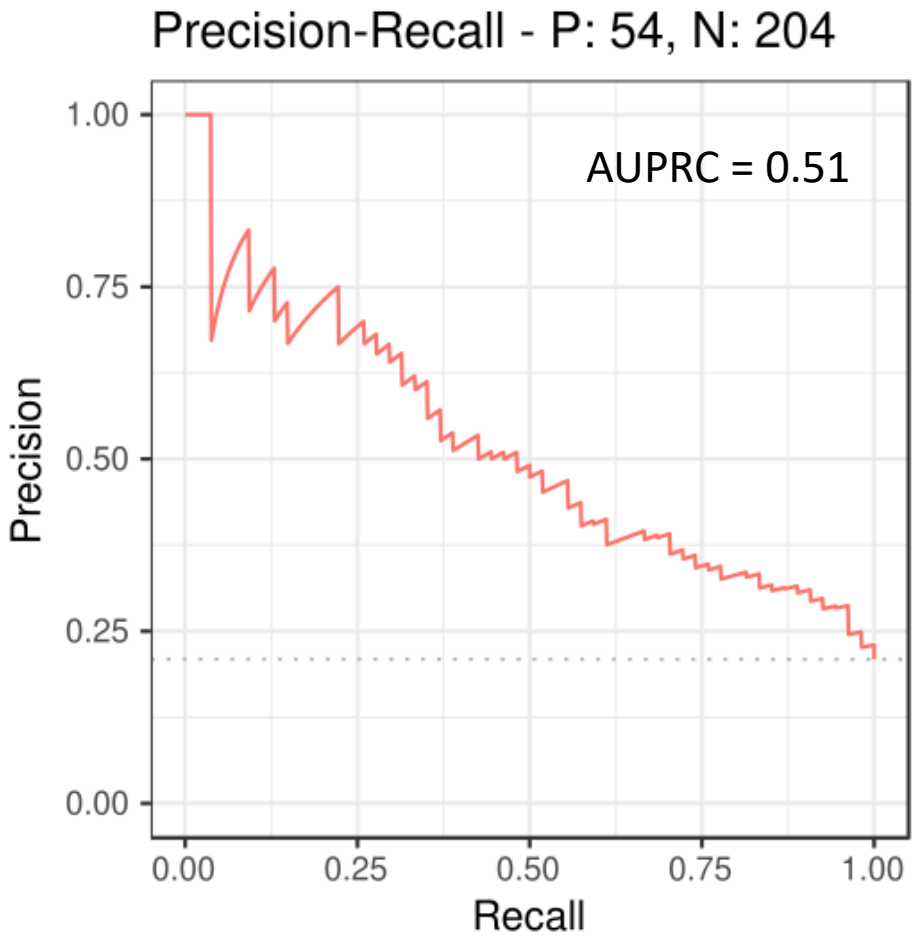
- Receiver operator characteristic curve (ROC)
- **False positive rate** (FPR): Fraction of actual negative cases incorrectly classified as positive
  - Also, **1 - specificity**
- **True positive rate** (TPR): Fraction of actual positive cases correctly classified as positive
  - Also, **recall / sensitivity**
- Area under ROC (AUC) summarizes ROC in a single number
  - 0.5 = random classifier
  - 1.0 = perfect classifier



Cmglee & Thoma (2018)
https://commons.wikimedia.org/w/index.php?curid=109730045

Evaluation - Classification





Ruuskanen et al. (2021); https://doi.org/10.1080/19490976.2021.1888673

Evaluation - Regression

- Multiple ways to evaluate model error with numeric *y*
  - y_i = true value
  - $\widehat{y_i}$ = predicted value

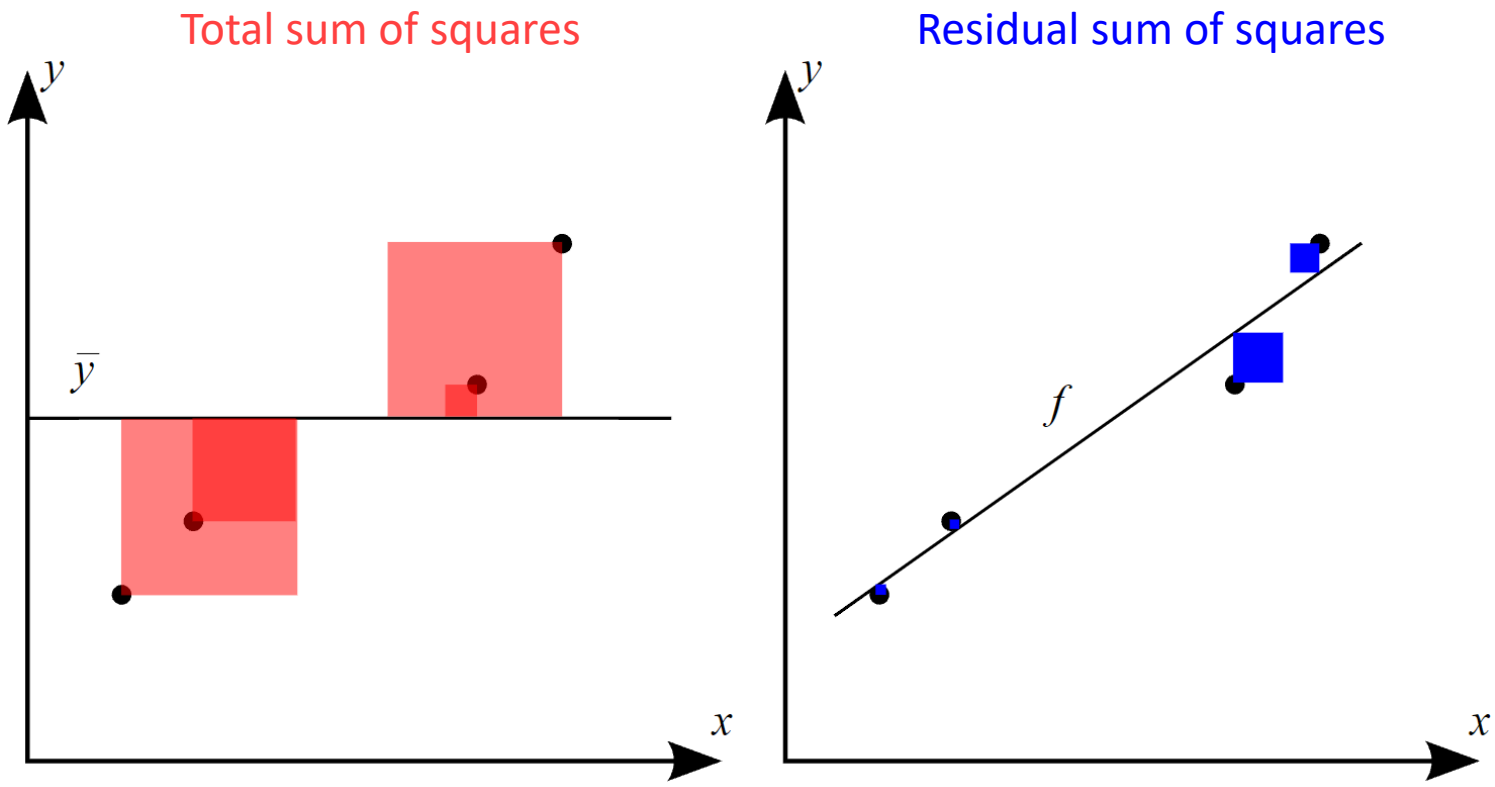
- Mean absolute error (MAE)
  - $MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n}$
  - Same unit as *y*
  - Robust to outliers

- Mean squared error (MSE)
  - $MSE = \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n}$
  - Squared units of *y*
  - Not robust to outliers

- Root mean squared error (RMSE)
  - $RMSE = \sqrt{MSE}$
  - Same unit as *y*
  - Not robust to outliers

Evaluation - Regression

- R squared (R^2)
  - $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
  - Ranges between 0 and 1
  - Units are % of variance explained



Orzetto (2010); https://commons.wikimedia.org/wiki/File:Coefficient_of_Determination.svg

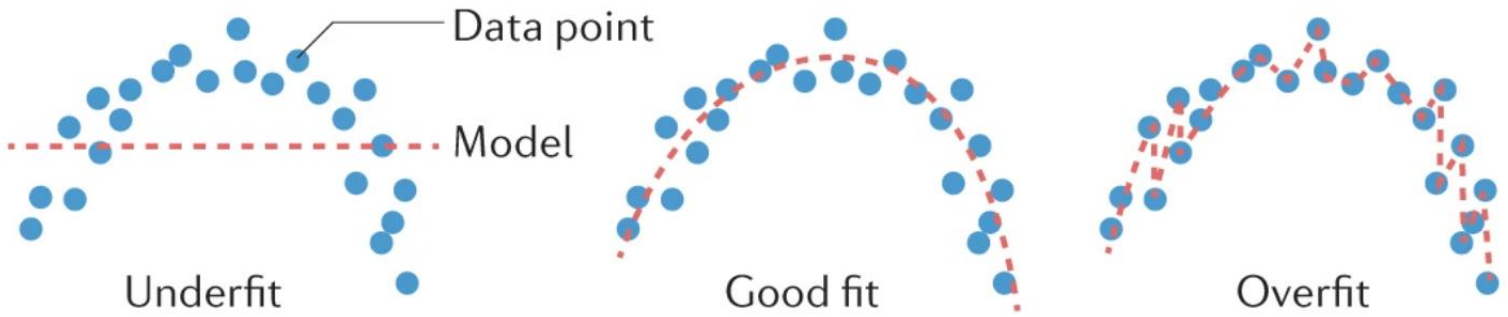
Issues & Solutions

- The data scientist applying supervised ML should be aware of several issues:

- Model performance
  - Over- and underfitting
  - Overestimation of performance (data leakage)

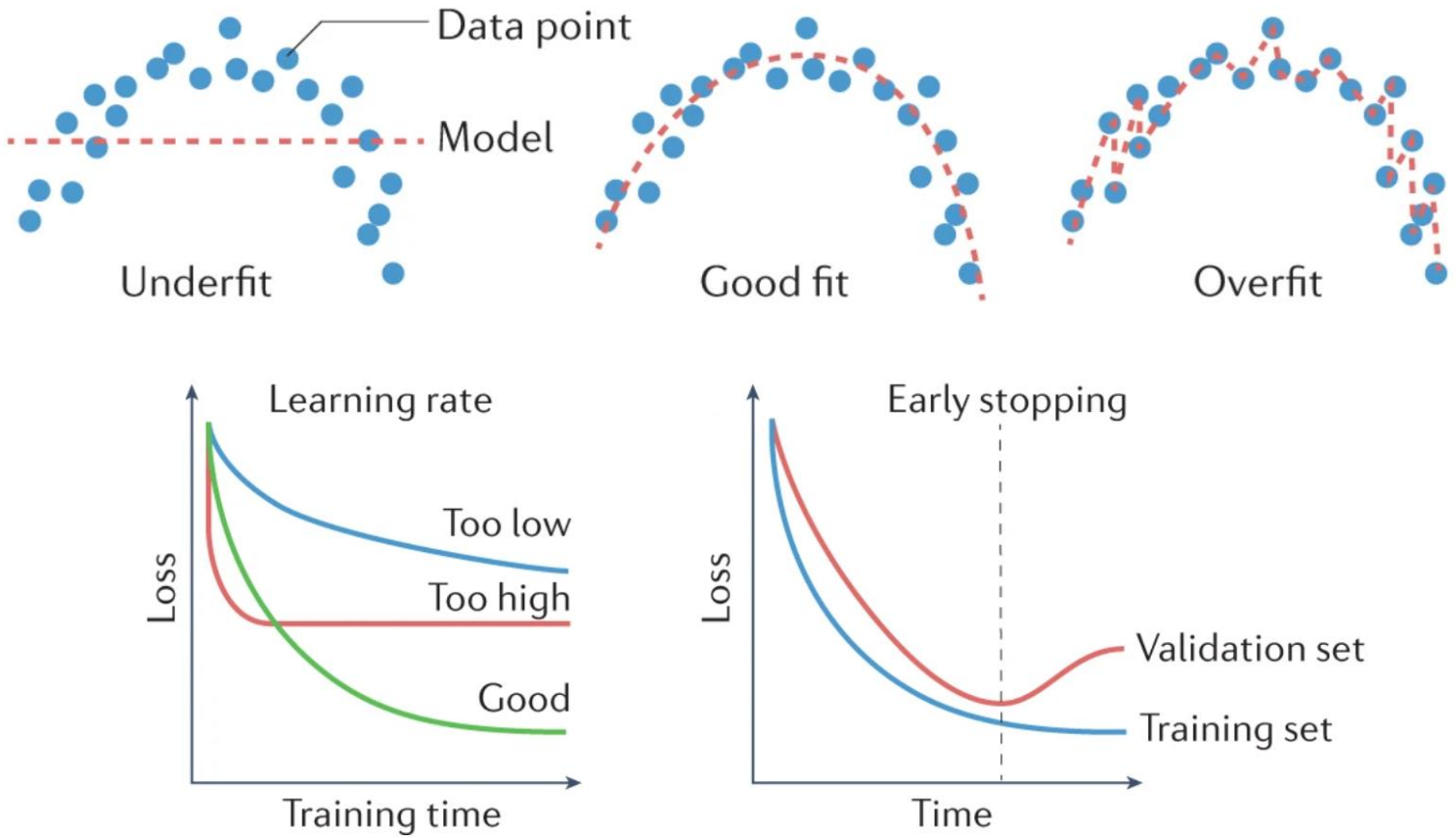
- Model applicability
  - Computational costs
  - Interpretability

Over- and underfitting



Greener et al. (2021); https://www.nature.com/articles/s41580-021-00407-0

Over- and underfitting



Greener et al. (2021); https://www.nature.com/articles/s41580-021-00407-0

Overestimation of performance

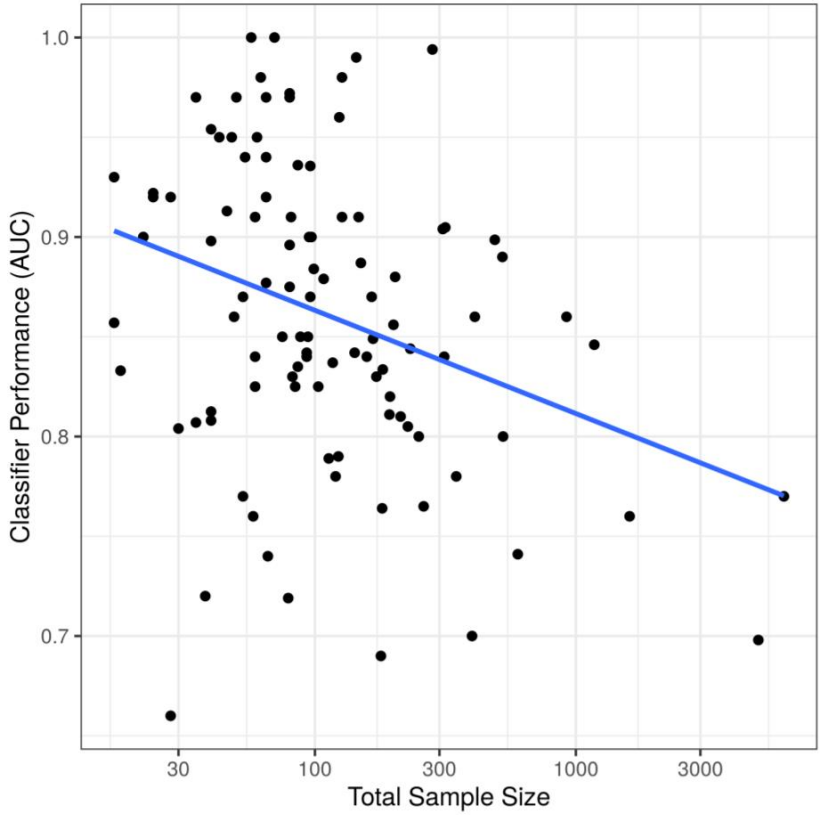


Figure 4: A scatter plot of the Abstract-reported AUC (y-axis) as a function of sample size (x-axis). Studies with larger sample sizes tended to report lower AUCs ($\rho = -0.31$; $p = 0.0013$).

Quinn (2021); https://arxiv.org/abs/2107.03611

Overestimation of performance

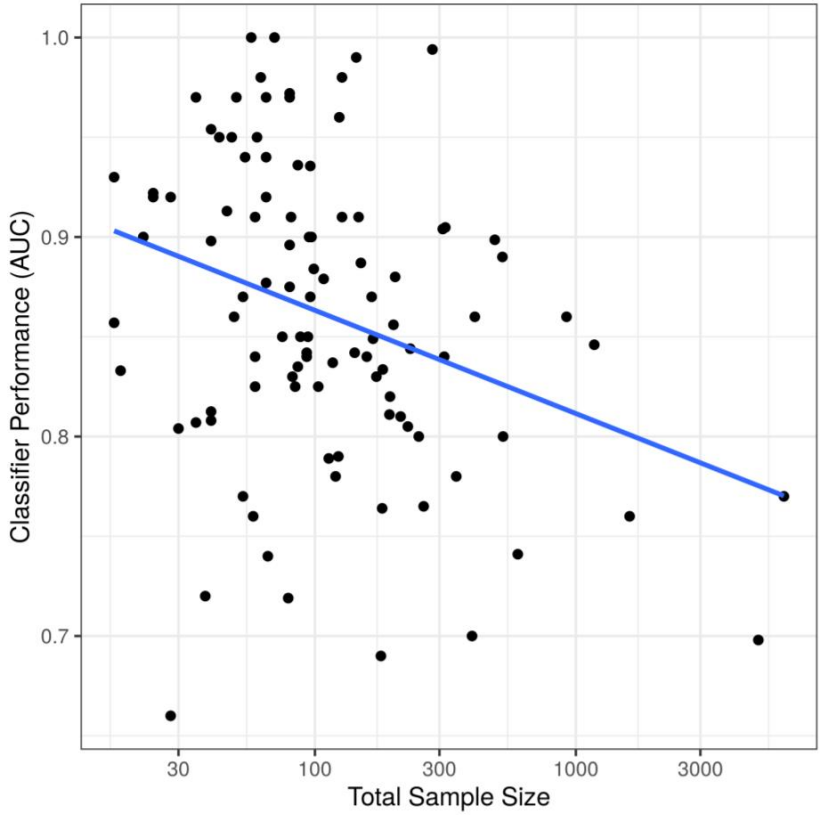
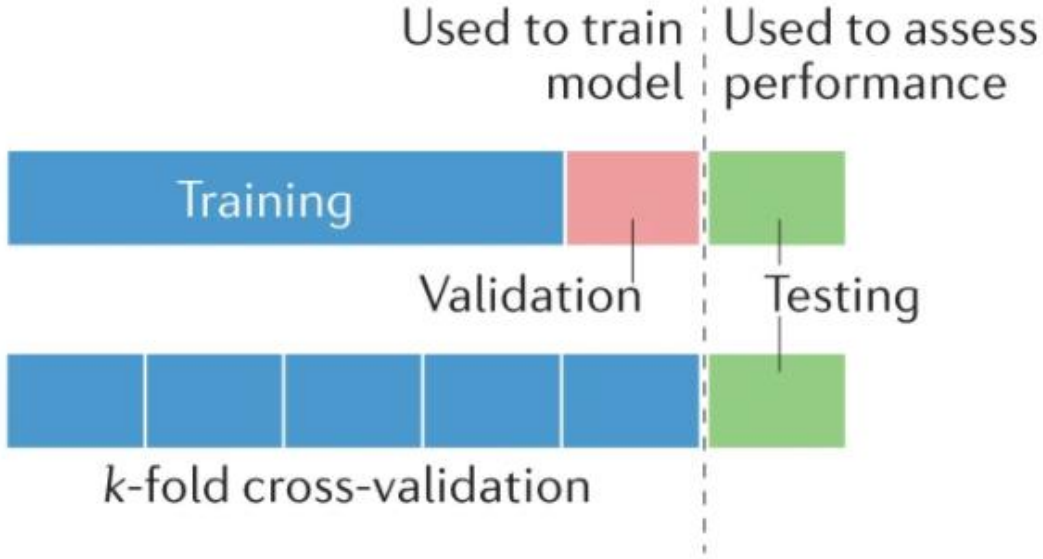


Figure 4: A scatter plot of the Abstract-reported AUC (y-axis) as a function of sample size (x-axis). Studies with larger sample sizes tended to report lower AUCs ($\rho = -0.31$; $p = 0.0013$).

Quinn (2021); https://arxiv.org/abs/2107.03611



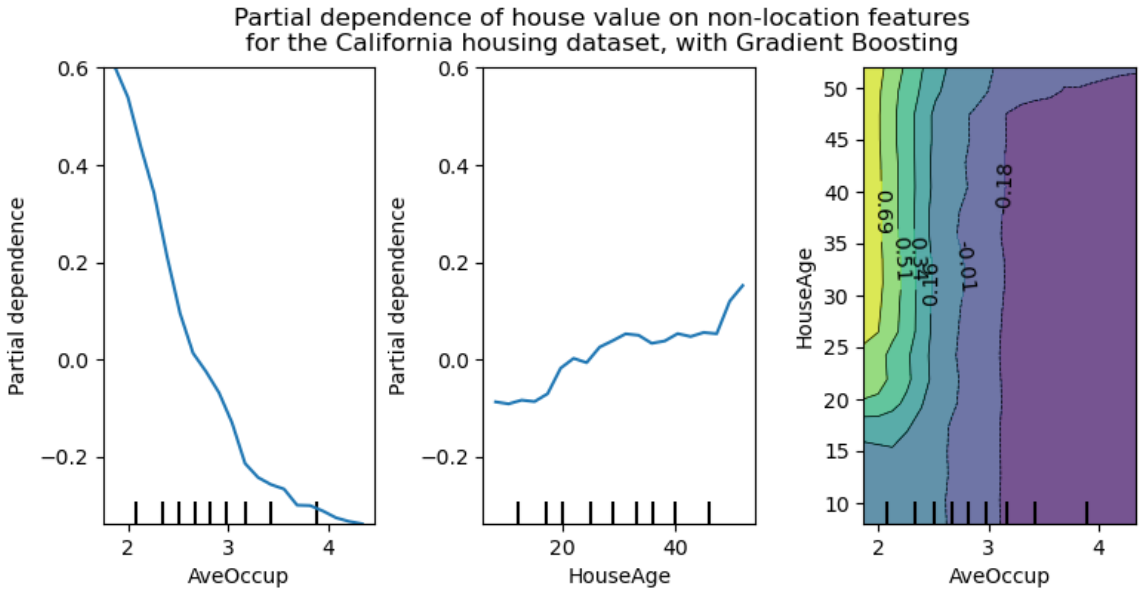
Greener et al. (2021); https://www.nature.com/articles/s41580-021-00407-0

- Data leakage?
  - Same or related samples
  - Features unavailable in new data

Interpretability

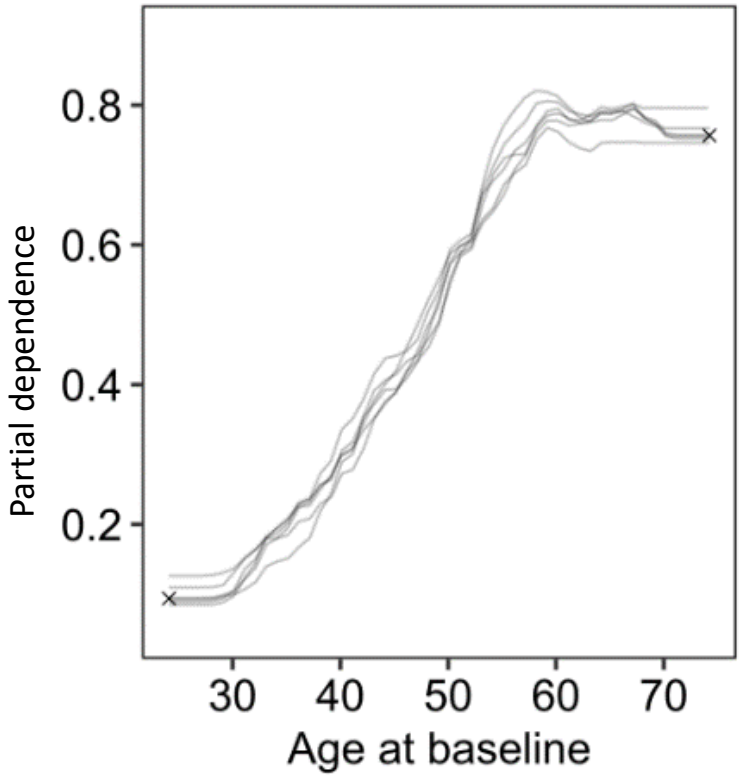
- Depending on the study question, a more easily interpretable model might be more useful than a better predicting one
- Parametric models are easier to directly interpret than complex ensembles
- Partial dependence of $\hat{y}$ on individual features and their combinations can be examined even in 'black box' models
  - A partial dependence plot shows the marginal effect of one or more input features on the model prediction

Partial dependence plots



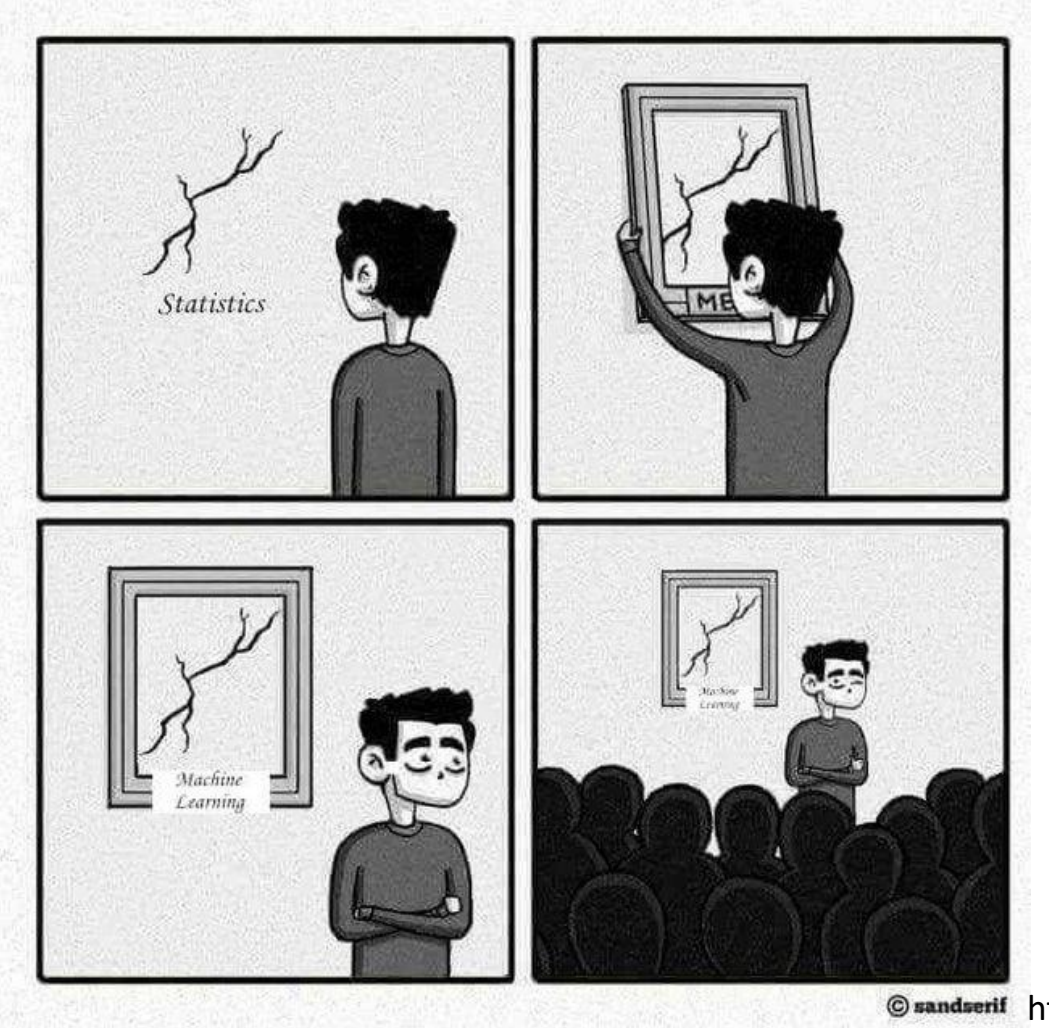
Scikit-learn developers (2021);
https://scikit-learn.org/stable/modules/partial_dependence.html

Partial dependence of high fatty liver risk group on baseline age, with 6 spatially cross-validated XGBoost models.



Ruuskanen et al. (2021); https://doi.org/10.1080/19490976.2021.1888673

Comments? Questions?



https://instagram.com/sandserifcomics