Multivariate Classification

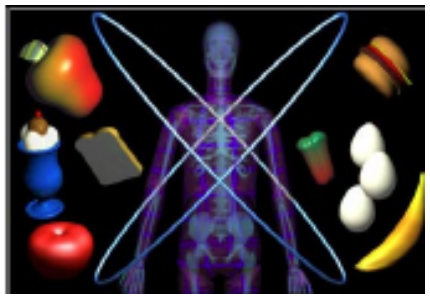
Dr Maria Anastasiadi
(m.anastasiadi@cranfield.ac.uk)

13th January 2025

www.cranfield.ac.uk

Experimental Workflow

1. Study



- Biological question
- Experimental Design

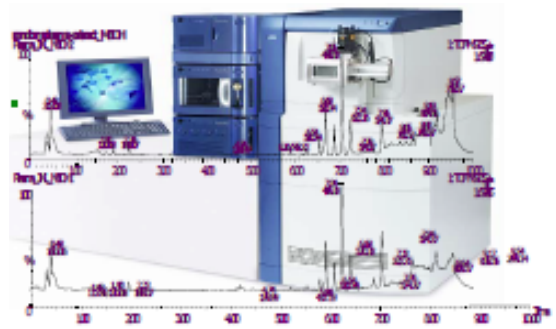
2. Sampling

- Sample collection
- Sample Storage

3. Sample preparation

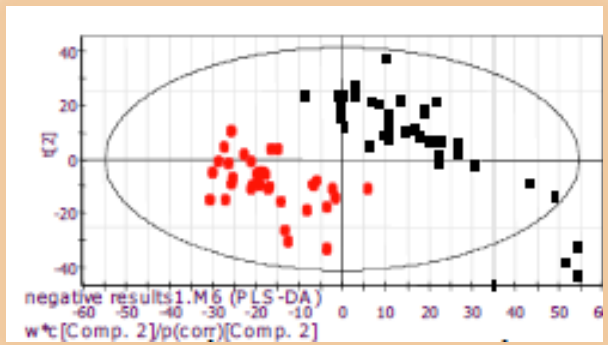
- Sample preparation
- Pre-treatment

4. Omics profiling



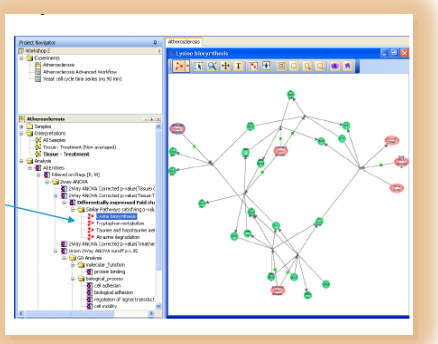
- Data acquisition
- Data pre-processing

5. Data mining



- Data analysis
- Modelling
- Biomarker discovery

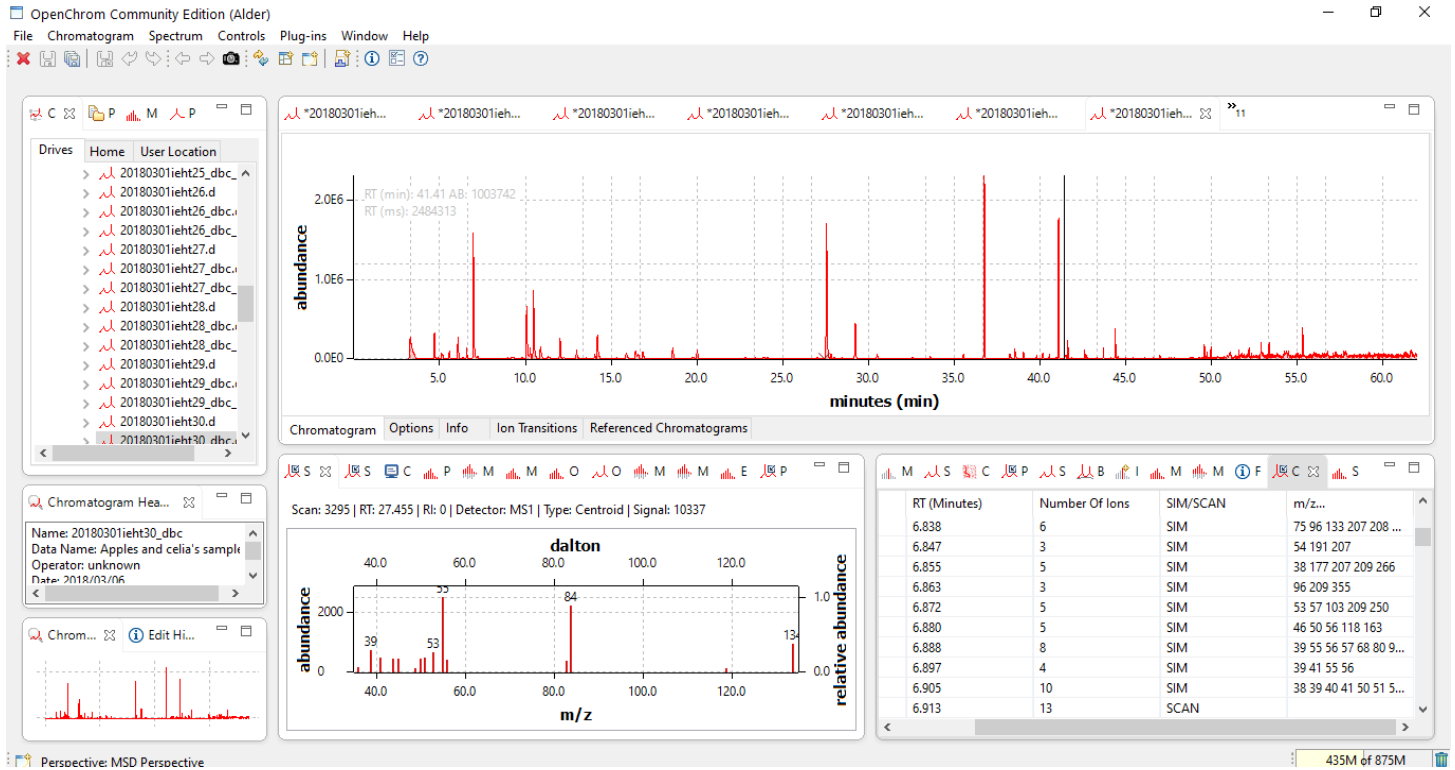
6. Biological interpretation



- Database search
- Metabolic pathways

Data handling

Most modern instrumental techniques generates a plethora of data.



Example: GC/MS files:

- Huge binary files (e.g. .DAT) storing large amounts of data.
- Dedicated programs required to extract the data and generate text files (e.g.CSV files).

Data Pre-treatment

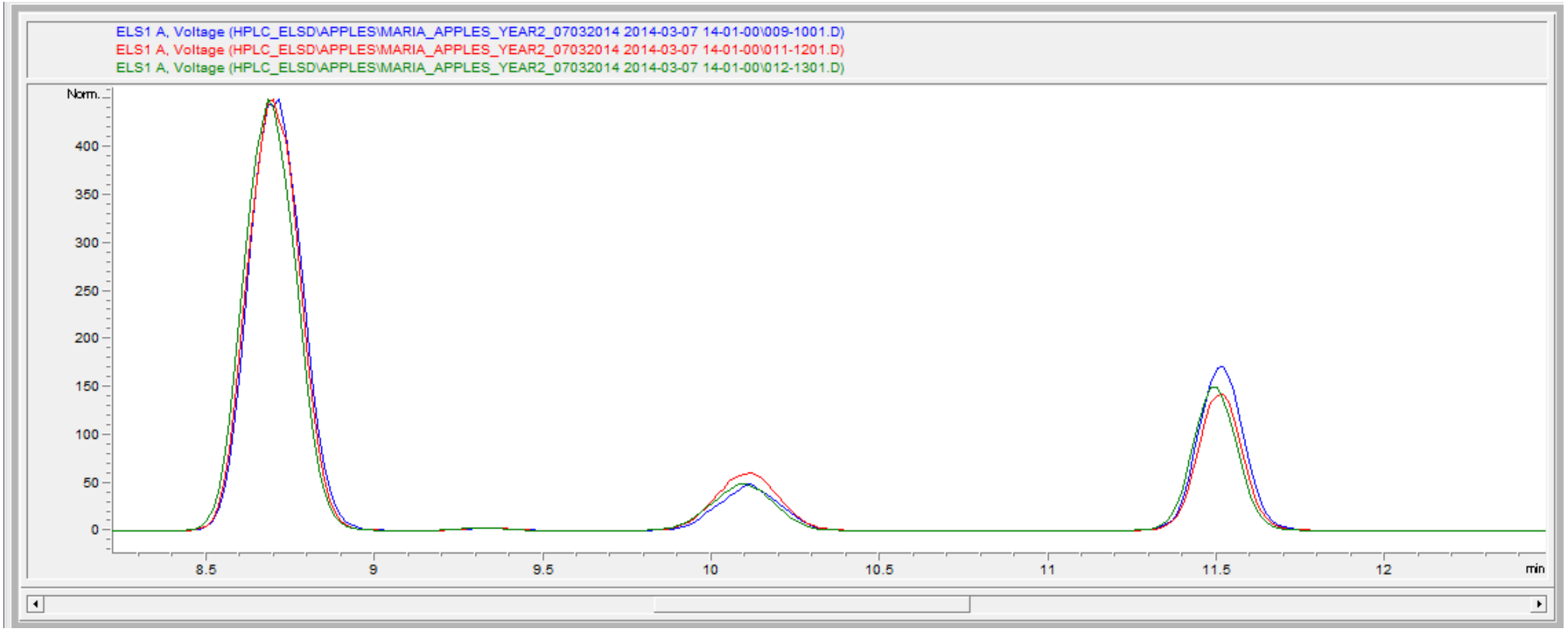
Types of pre-treatment:

- Background subtraction;
- Alignment;
- Smoothing/Filtering;
- Scaling;
- Combination of the above.

Commonly employed scaling techniques:

- Mean-centring;
- Auto-scaling;
- Range-scaling.

Data Pre-treatment: Alignment



Unsupervised Exploratory Data Analysis

Exploratory data analysis (EDA) is used to answer questions such as:

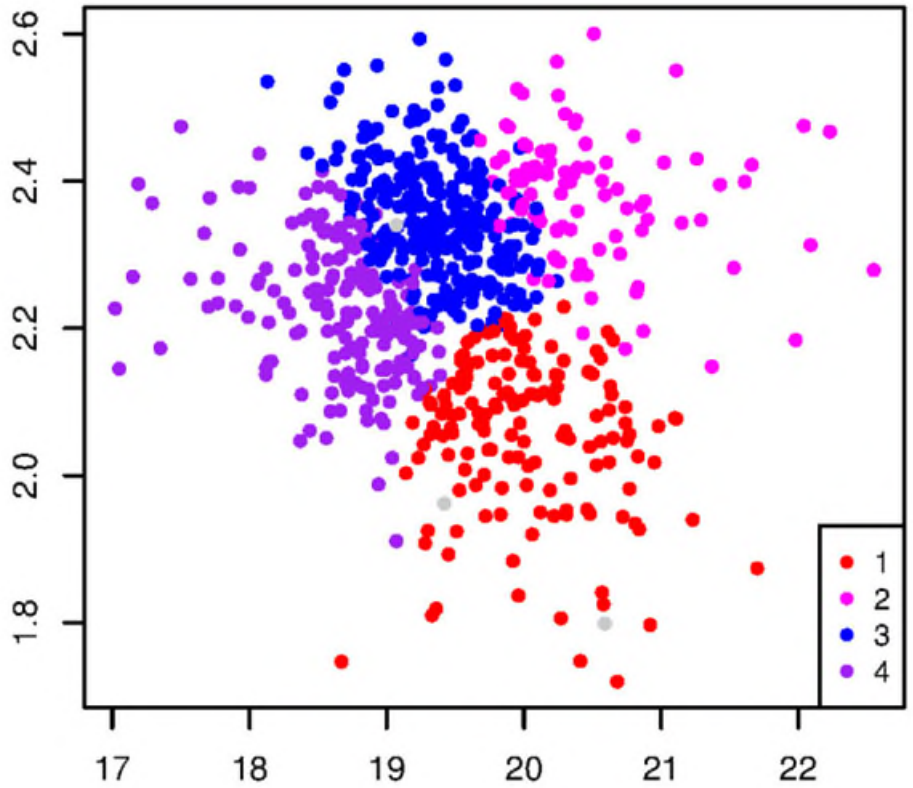
- Can we discriminate between different sample groups?
- Are there any rogue results (outliers)?
- How much variation is there between samples of the same type?
- They can also be used as data pre-treatment methods.

Methods for unsupervised data analysis include:

PCA, ICA, HCA, k-means clustering

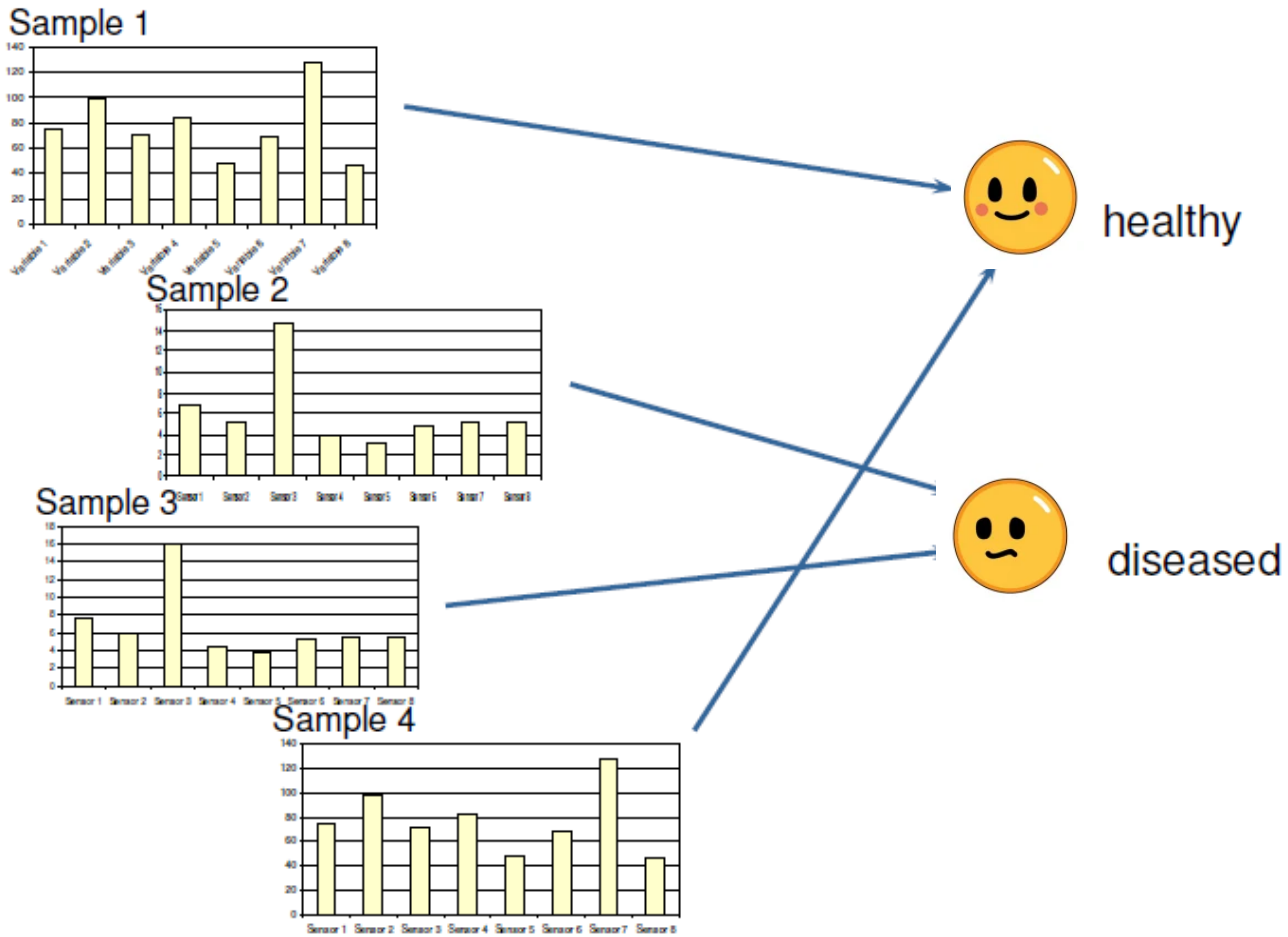
Predictive Analysis: Multivariate Classification

- With EDA we are looking to identify patterns in the data.
- But what if we want to use the multivariate data to predict the class of unknown samples?
- This process is known as **multivariate classification**.



Multivariate Classification

Aim of Multivariate classification is to assign each sample to a defined class.



Multivariate Classification

Multivariate classification is a *supervised* approach, i.e. *we need a certain number of known examples to begin with.*

Multivariate classification has two main approaches:

- **Statistical methods.**

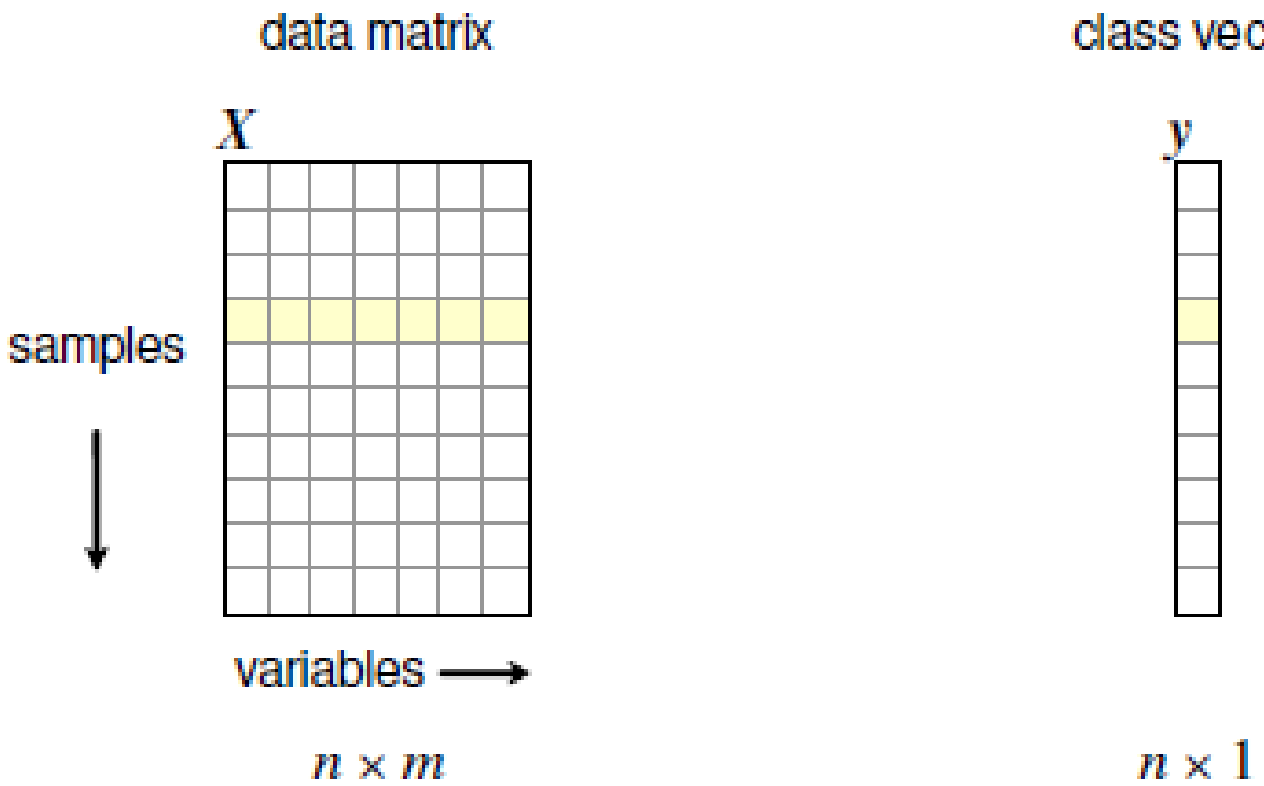
A subfield of mathematics which deals with finding relationships between variables to predict an outcome.

- **Machine learning methods.**

A subfield of artificial intelligence (AI) which deals with building systems that can learn from data, instead of explicitly programmed instructions.

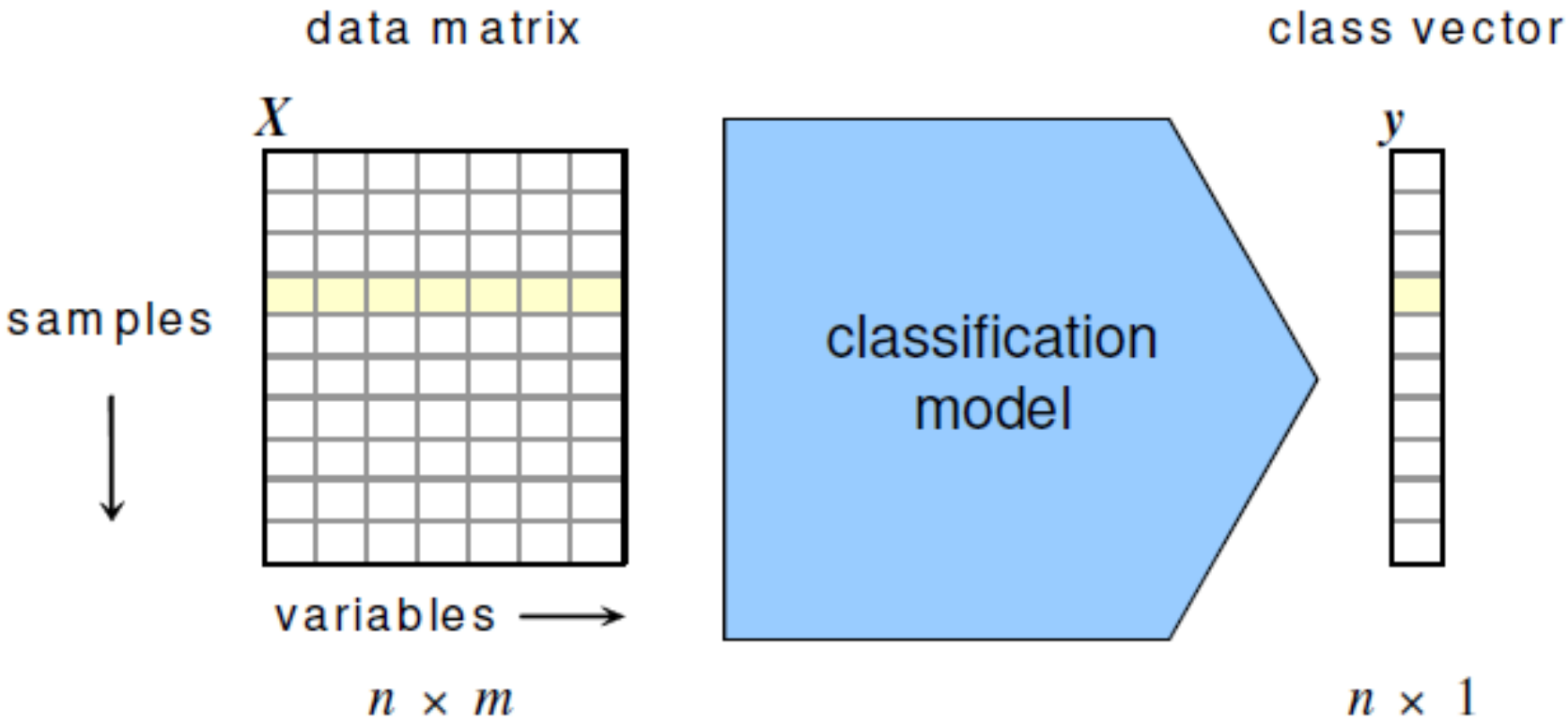
Multivariate Classification - Workflow

Step 1. We start off with a data matrix (input), and a corresponding class vector (output) which indicates the class of each sample. This is the training set.



Multivariate Classification

Step 2. Training: using our chosen method, we build a classification model.

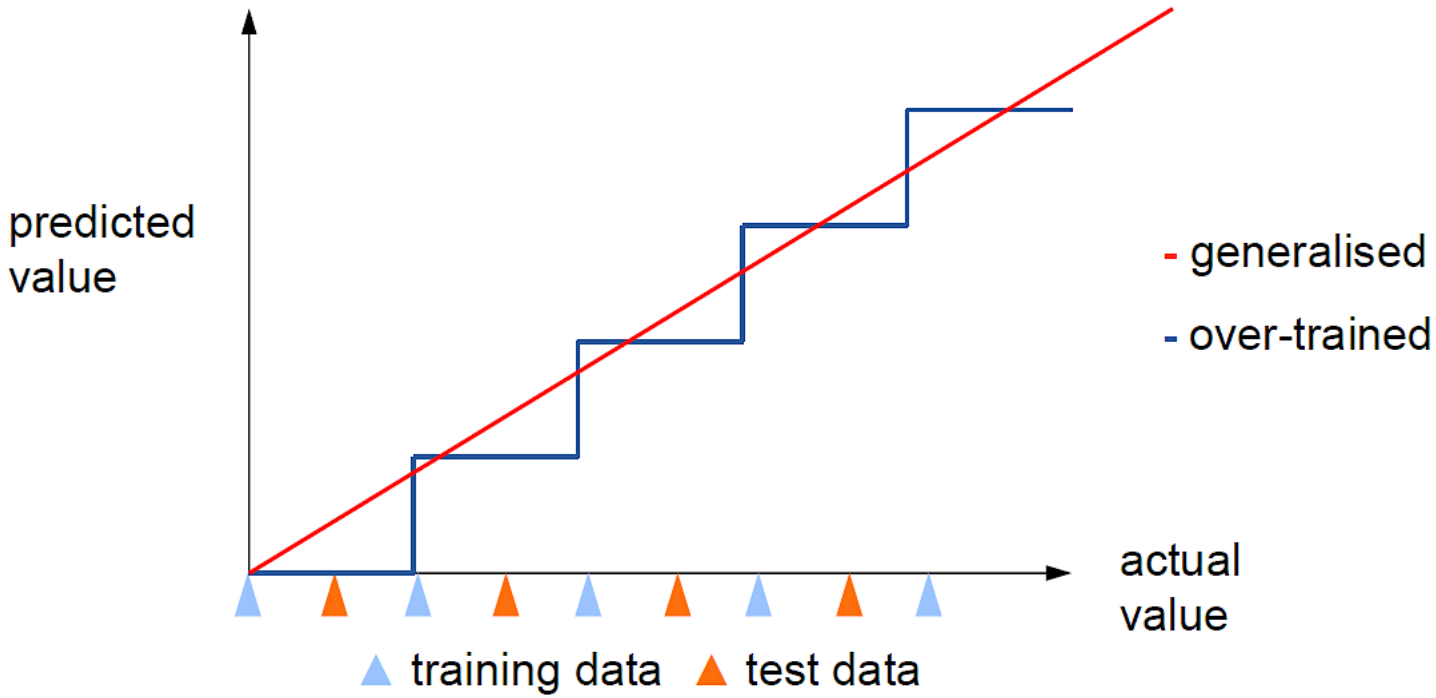


Multivariate Classification

Step 3. Model optimisation: apply cross validation to get a better estimate of the prediction error of the model. Helps avoid over-training pitfall.

Most common cross-validation methods include:

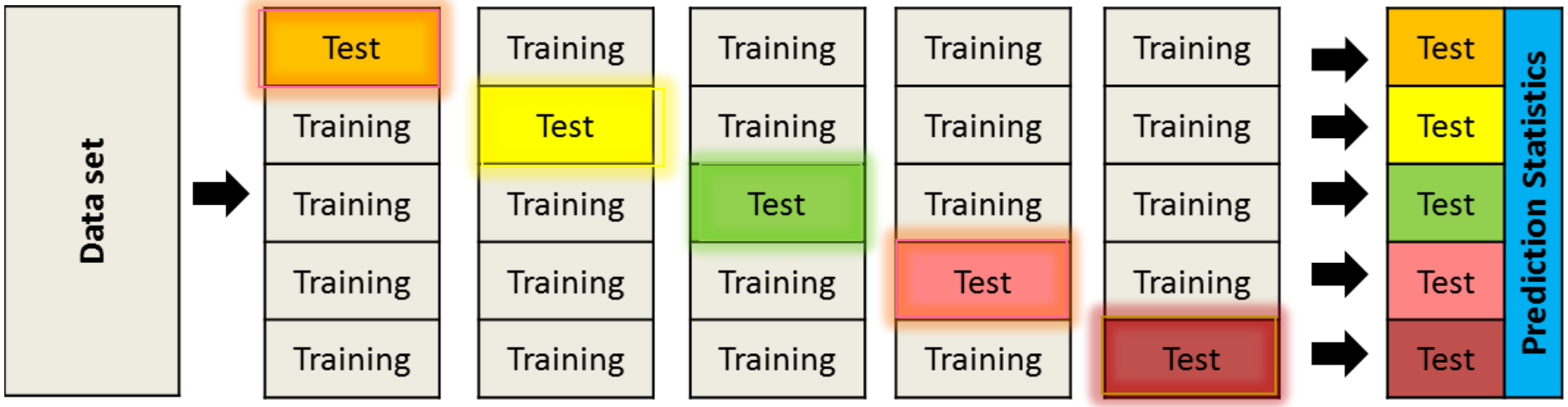
- k-fold cross-validation
- Bootstrapping



Multivariate Classification: Model Optimisation

k-fold cross-validation

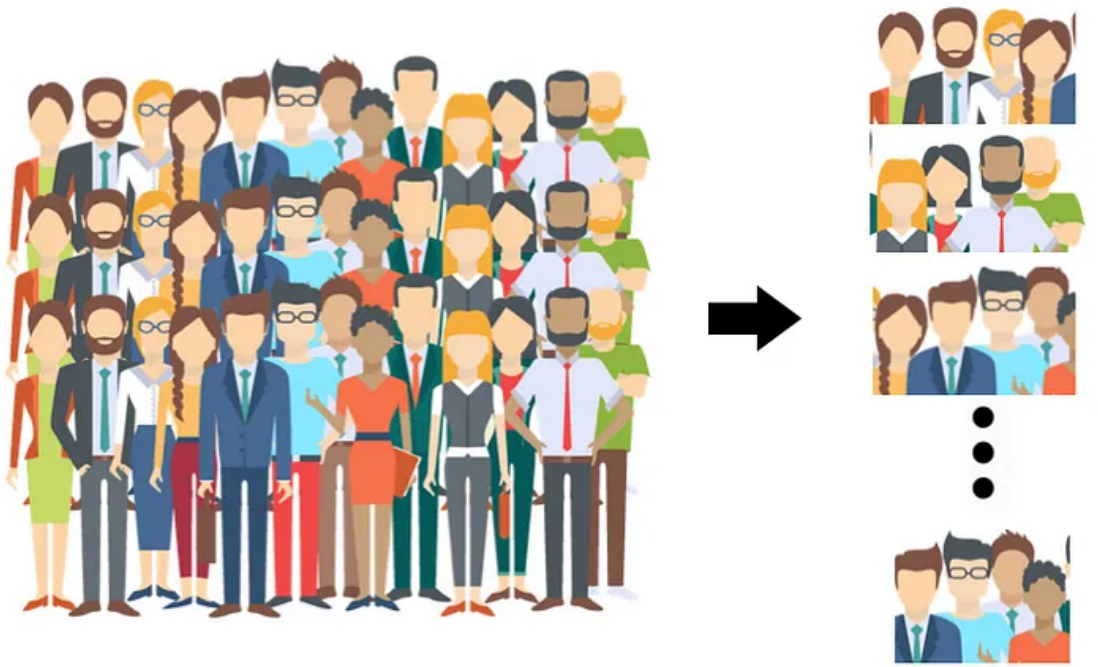
1) The training data set is portioned into equal size **k subsets**.
2) One subset k is used as a validation set and is thus omitted.
3) The remaining k-1 subsets are used to build a classification model.
4) The process is repeated for all k-folds.
5) Prediction error is calculated as average of the k prediction errors.
6) The root mean square error of prediction (RMSEP) is obtained.



Multivariate classification: model optimisation

Bootstrapping – Making use of resampling

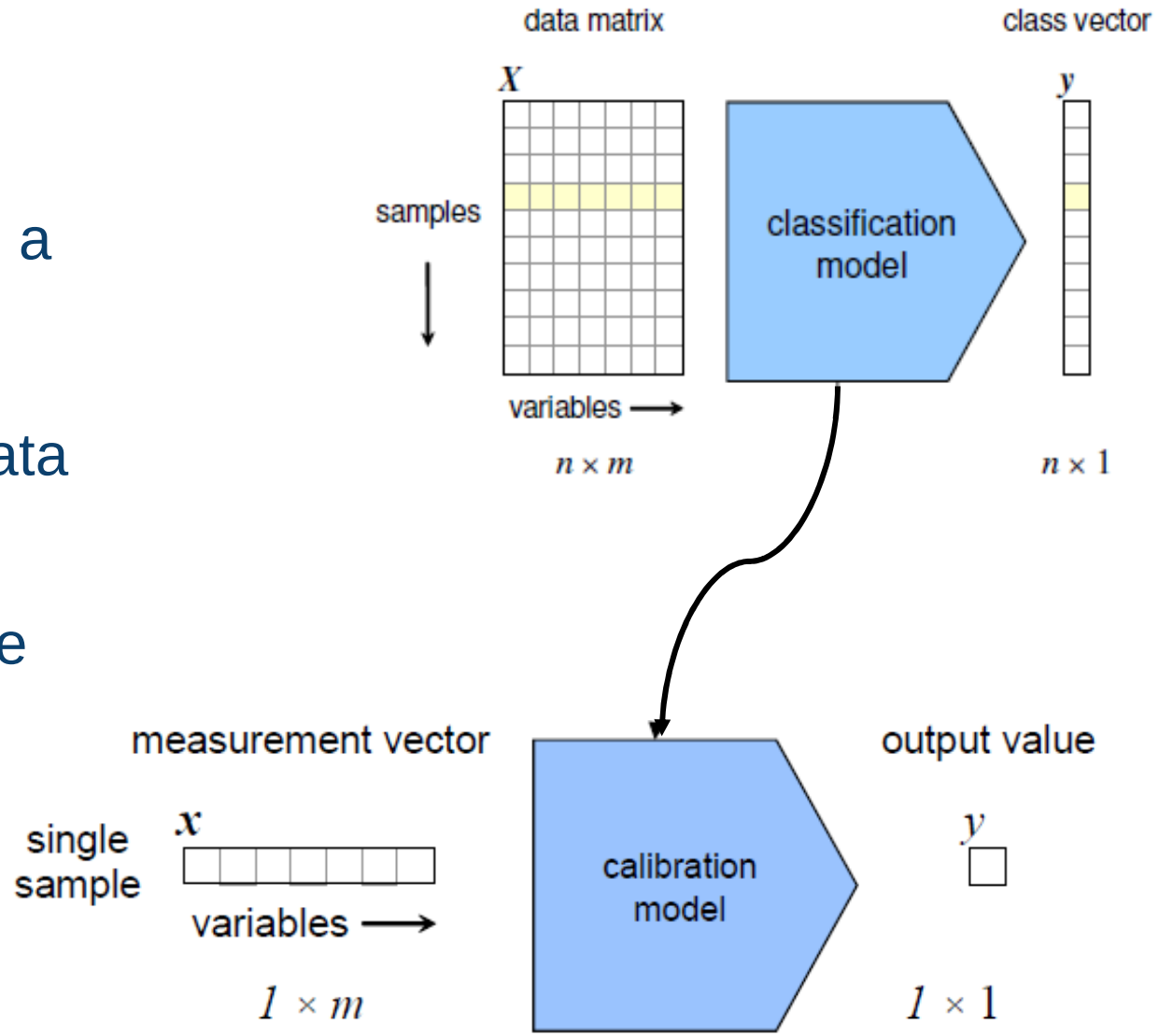
- A random selection of k samples is taken from the data set and used as a test set.
- Samples are selected with replacement.
- Random selection procedure is repeated multiple times, e.g. 100-200 times.
- The result is estimated by averaging the errors of the individual procedures.



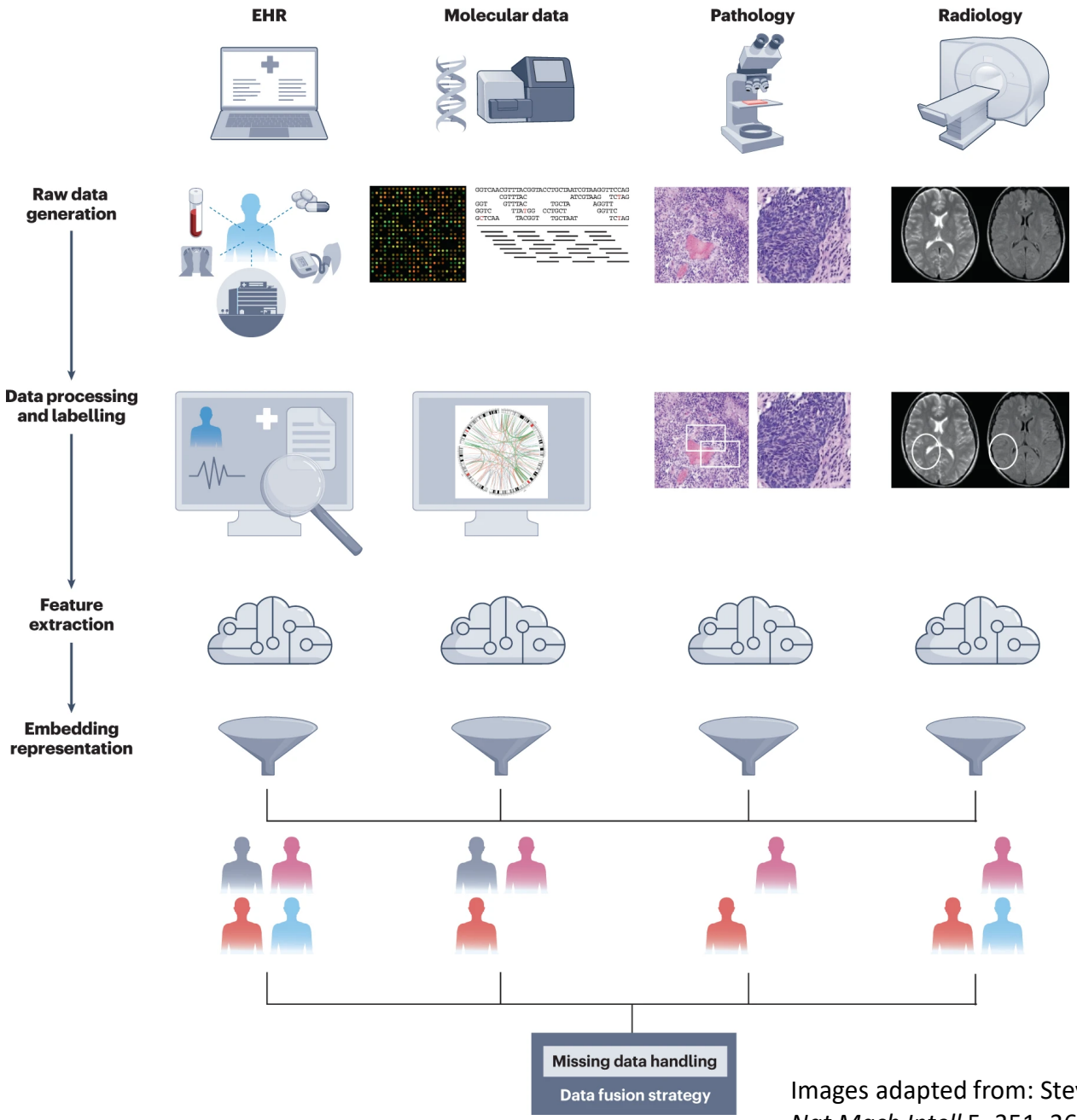
Multivariate Classification

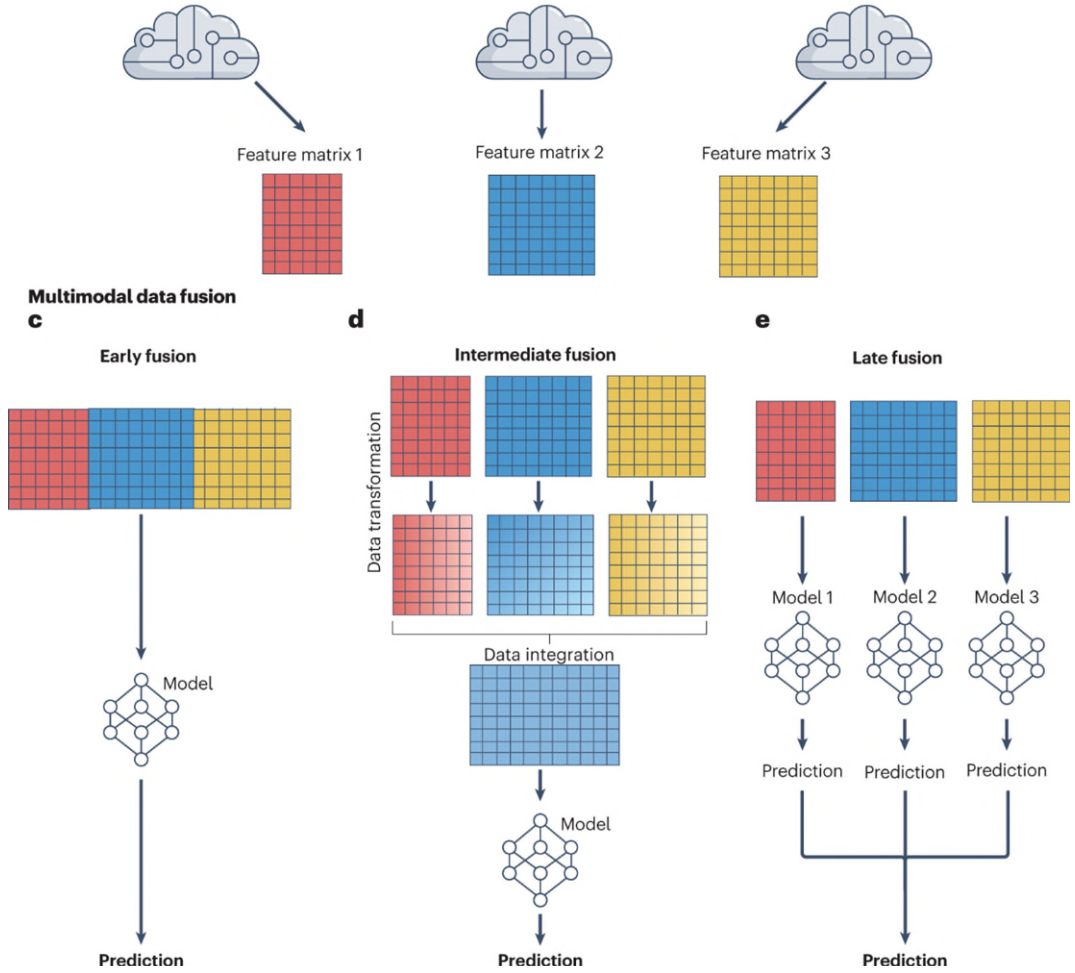
Step 4. Model validation:

- Once we've built and optimised a calibration model from known data, we can apply optimised calibrated model to unknown data (independent test set).
- A measure of the success of the model is the percentage of correctly classified validation samples.



Data Fusion





Images adapted from: Steyaert, S., et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat Mach Intell* 5, 351–362 (2023).

Multivariate Classification

Step 5. Model Evaluation

A confusion matrix:

Class	Actual Class		
	Condition	Healthy	Disease
Predicted Class	Healthy	Correct True Negative (TN)	Incorrect False Negative (FN)
	Disease	Incorrect False Positive (FP)	Correct True Positive (TP)

Multivariate Classification

Model Evaluation Metrics

$$Accuracy = \frac{(TP + TN)}{(Total)}$$

Correct predictions over the total number of samples

$$Specificity = \frac{TN}{(TN + FP)}$$

When it's actually healthy, how often does it predict so?

$$Sensitivity\ (Recall) = \frac{TP}{(TP + FN)}$$

When it's actually disease, how often does it predict so?

Multivariate Classification

Example

N = 165

H = 95

D = 70

Class	Actual Class		
	Condition	Healthy	Disease
Predicted Class	Healthy	TN = 80	FN = 5
	Disease	FP = 15	TP = 65

Accuracy: (TP+TN)/total = (65+80)/165 = 145/165 = 0.88

Error rate: 1 - accuracy = 0.12

Specificity: TN/actual healthy = 80/95 = 0.84

Sensitivity: TP/actual disease = 65/70 = 0.93

False Positive Rate: 1-specificity = 0.16

Multivariate classification : Model Evaluation Metrics

$$Accuracy = \frac{(TP + TN)}{(Total)}$$

What happens when we have class imbalance?

$$Precision = \frac{TP}{TP + FP}$$

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

F1 score combines precision and recall using their harmonic mean

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

F1 score can also be represented by this formula

Multivariate Classification

Model Evaluation Metrics

The Kappa statistic: The kappa statistic **adjusts accuracy** by accounting for the possibility of a correct prediction by chance alone.

Cohen's kappa coefficient

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where :

Pr(a) = the proportion of the actual agreement between the classifier and the true values.

and

Pr(e) = the proportion of the expected agreement between the classifier and the true values.

Multivariate Classification

Both Pr(a) and Pr(e) can be obtained from the confusion matrix.

$$\boldsymbol{Pr(a) = Accuracy}$$

Pr(e) = the probability that by chance alone the predicted and actual values match and can be obtained as follows:

P(actual type is negative) P(predictive type is negative)

$$\boldsymbol{Pr(e) = \frac{(FN+TP)}{(Total)} * \frac{(FP+TP)}{(Total)} + \frac{(TN+FP)}{(Total)} * \frac{(TN+FN)}{(Total)}}$$

P(actual type is positive) P(predicted type is positive)

Multivariate Classification

The kappa statistic is often considered a more robust representation of the model's performance especially in cases with high class imbalance.

Cohen's Kappa statistic (κ)	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.20–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Multivariate Classification

Multivariate Statistical Methods

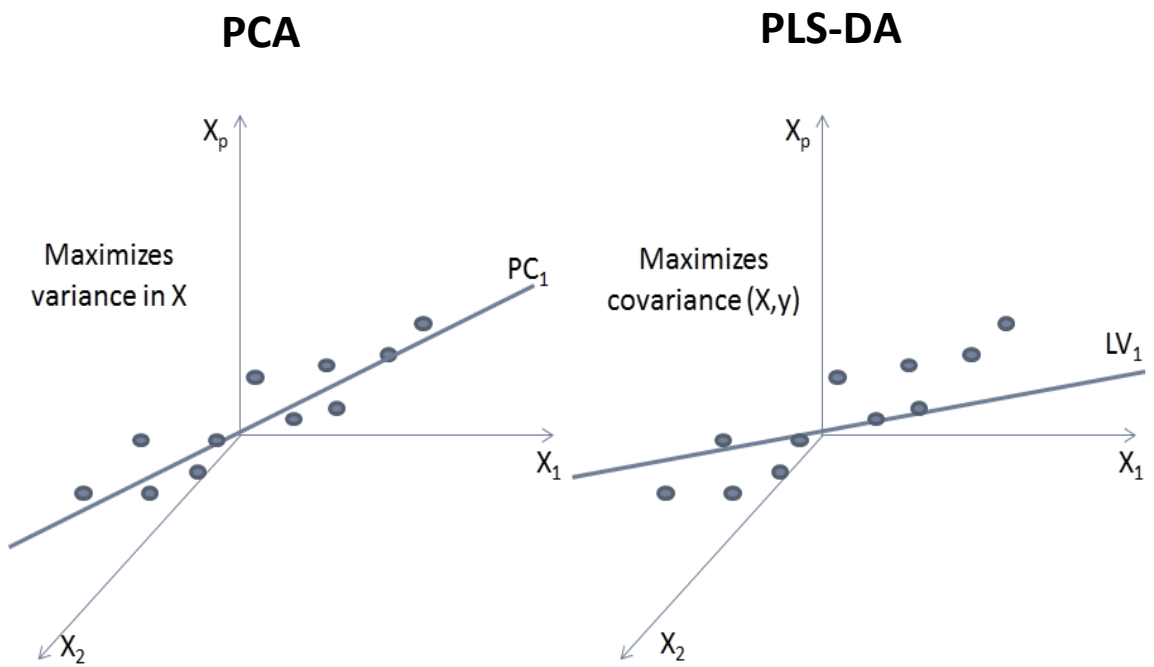
Building mathematical models to relate data to specific patterns of interest.

Examples of methods applied include:

- **Soft Independent Modelling of Class Analogies (SIMCA).**
- **Linear Discriminant analysis (LDA).**
- **Partial Least Squares Discriminant Analysis (PLS-DA).**

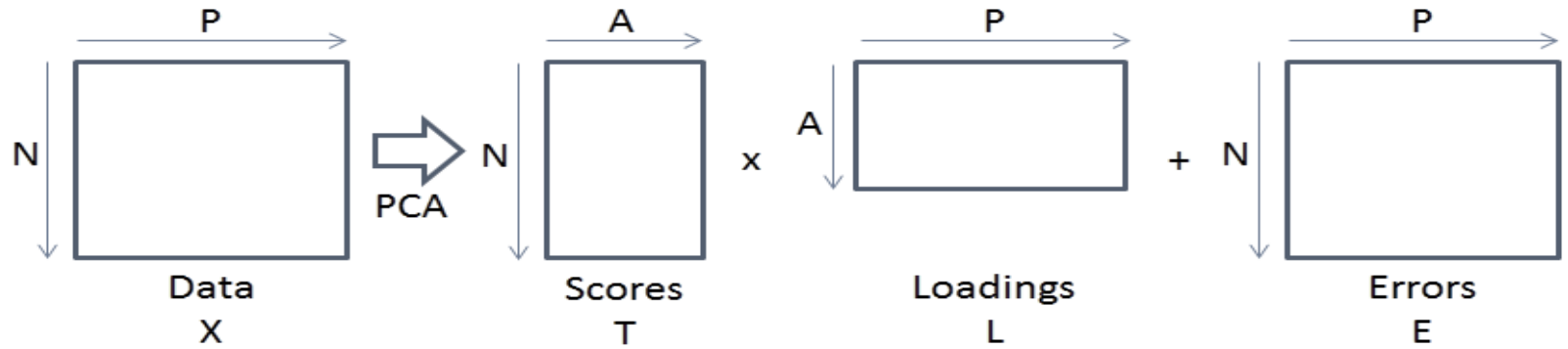
Partial Least Squares Discriminant Analysis (PLS-DA)

- In PLS, X and Y variables are modelled simultaneously, to find the **latent variables** (LVs) in X that will predict the LVs in Y.
- Similar to PCA but instead of maximising variance maximises the **covariance** between X and Y.



Multivariate Classification: Methods

2. Partial Least Squares Discriminant Analysis (PLS-DA)



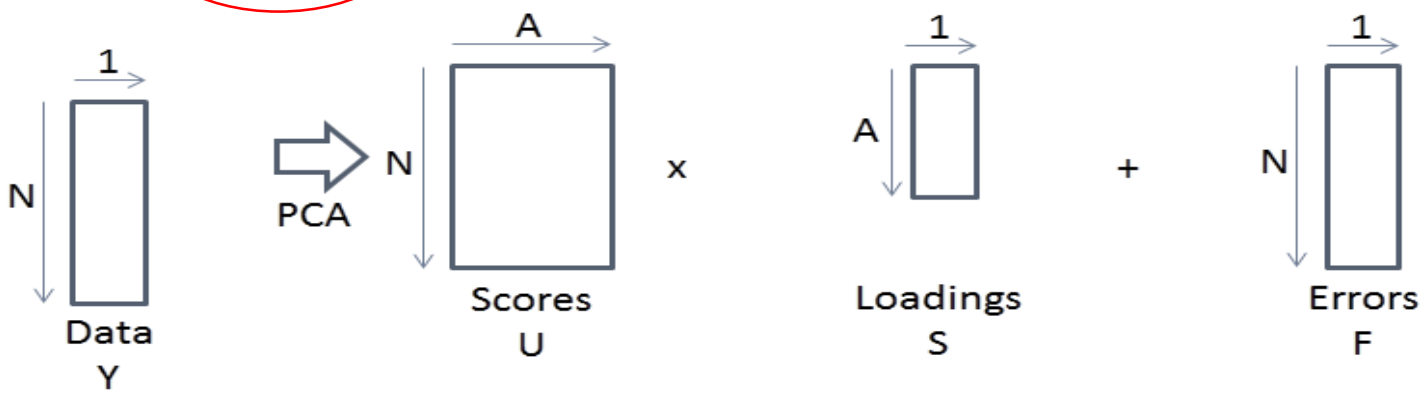
Inner correlation

PLS model

$$Y = TS' + F$$

$$Y = XW^{*}S' + F$$

$$Y = XB + F$$



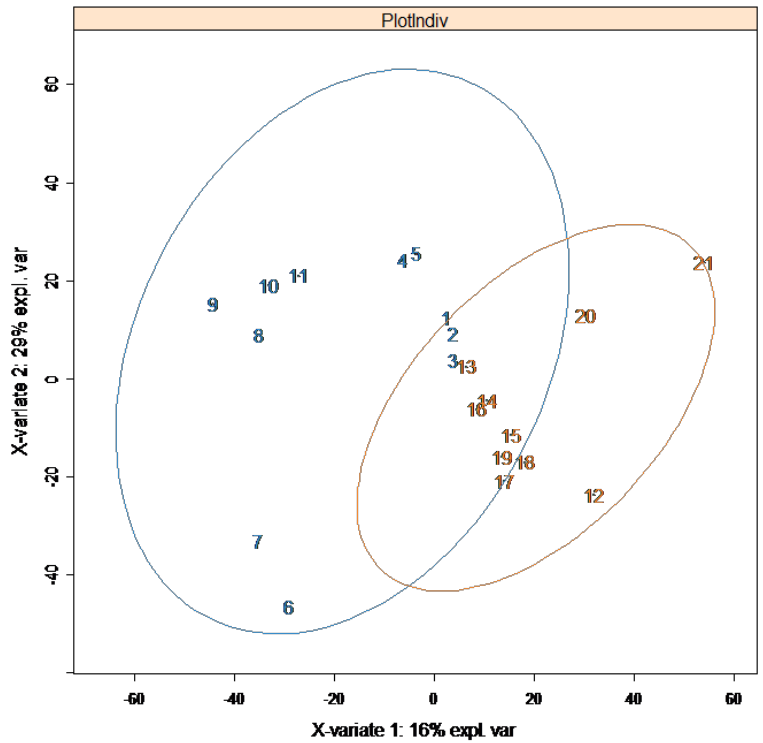
$$Y = XB + F$$

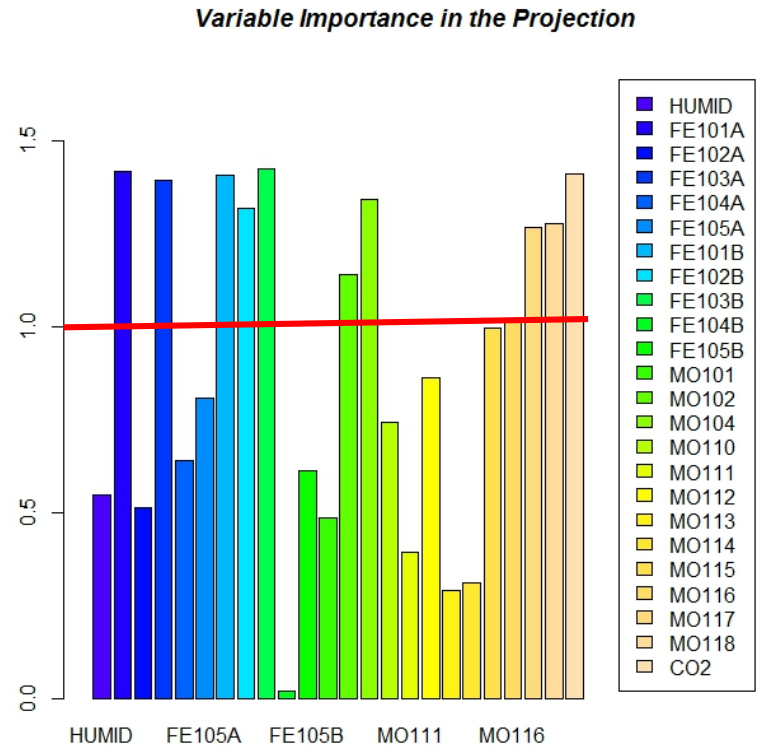
Manfredi & Marengo (2014)

Multivariate Classification: Methods

PLSDA permits candidate biomarker identification through calculation of the ***VIP*** *(variable importance for the projection)* statistic.

The VIP provides a summary of the importance of an X variable for both Y and X.





Biomarker Discovery

A biomarker is a metabolite/protein etc within a particular biological matrix that is indicative of a particular biological state (e.g. disease).

A biomarker can be used as the basis of a diagnostic test. For instance:

- Bilirubin in urine – a marker of liver dysfunction.
- Glucose in blood – used as a marker to manage diabetes.

Molecular biomarkers are potential targets for new pharmaceutical developments and disease diagnosis.

Biomarker Discovery: Identification *vs* Profiling

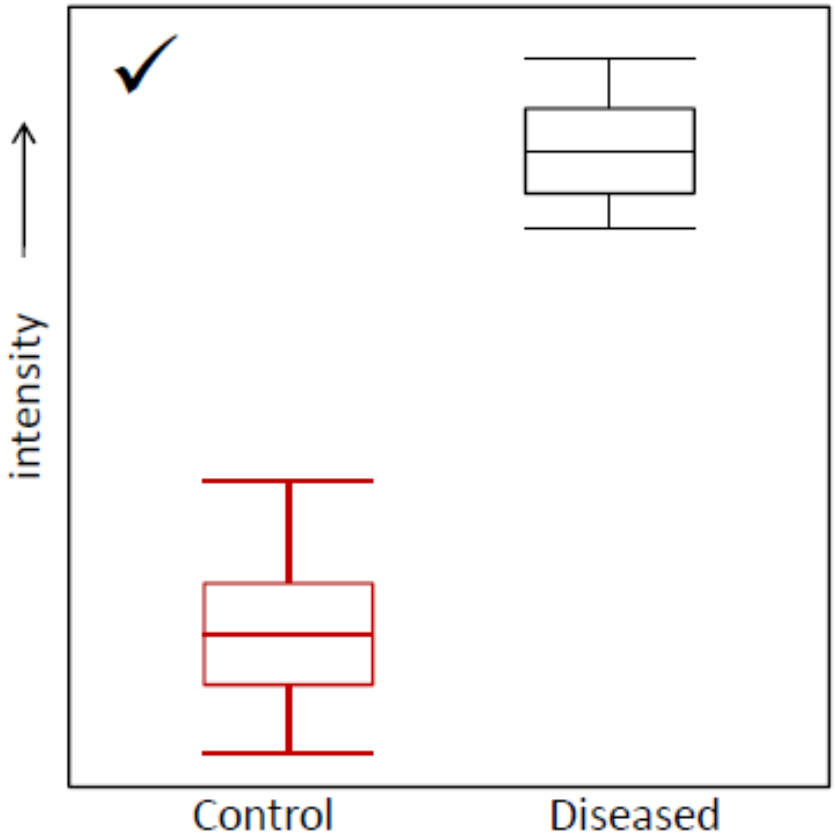
Biomarker discovery can be done on identified molecules, or on “raw” data.

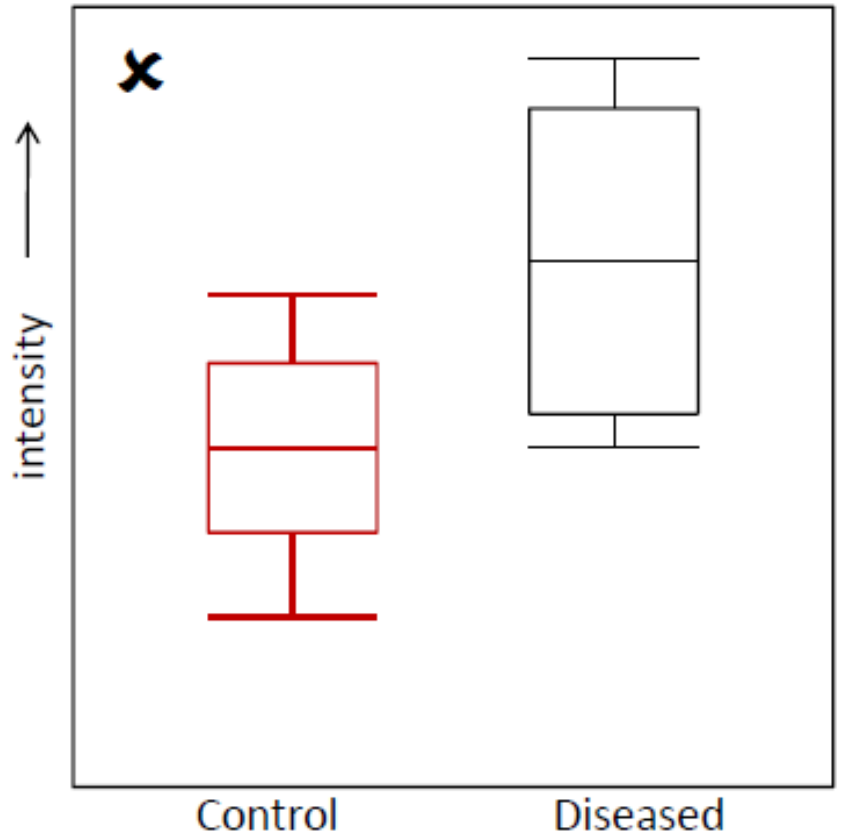
- **Identification** (Targeted approach):
- Resulting markers are molecules – good for diagnostics and drug targets.
- Much data is lost because we can’t identify most features within a given data set.

- **Profiling** (Untargeted approach):
- Many features – nothing is missed.
- Many features – a lot of data to wade through.
- Possible to link features back to molecules, but challenging.
- Only option for relatively unexplored species when no libraries or databases are available.

Biomarker Discovery

A reliable marker needs to be repeatable and significantly different between biological states under study.

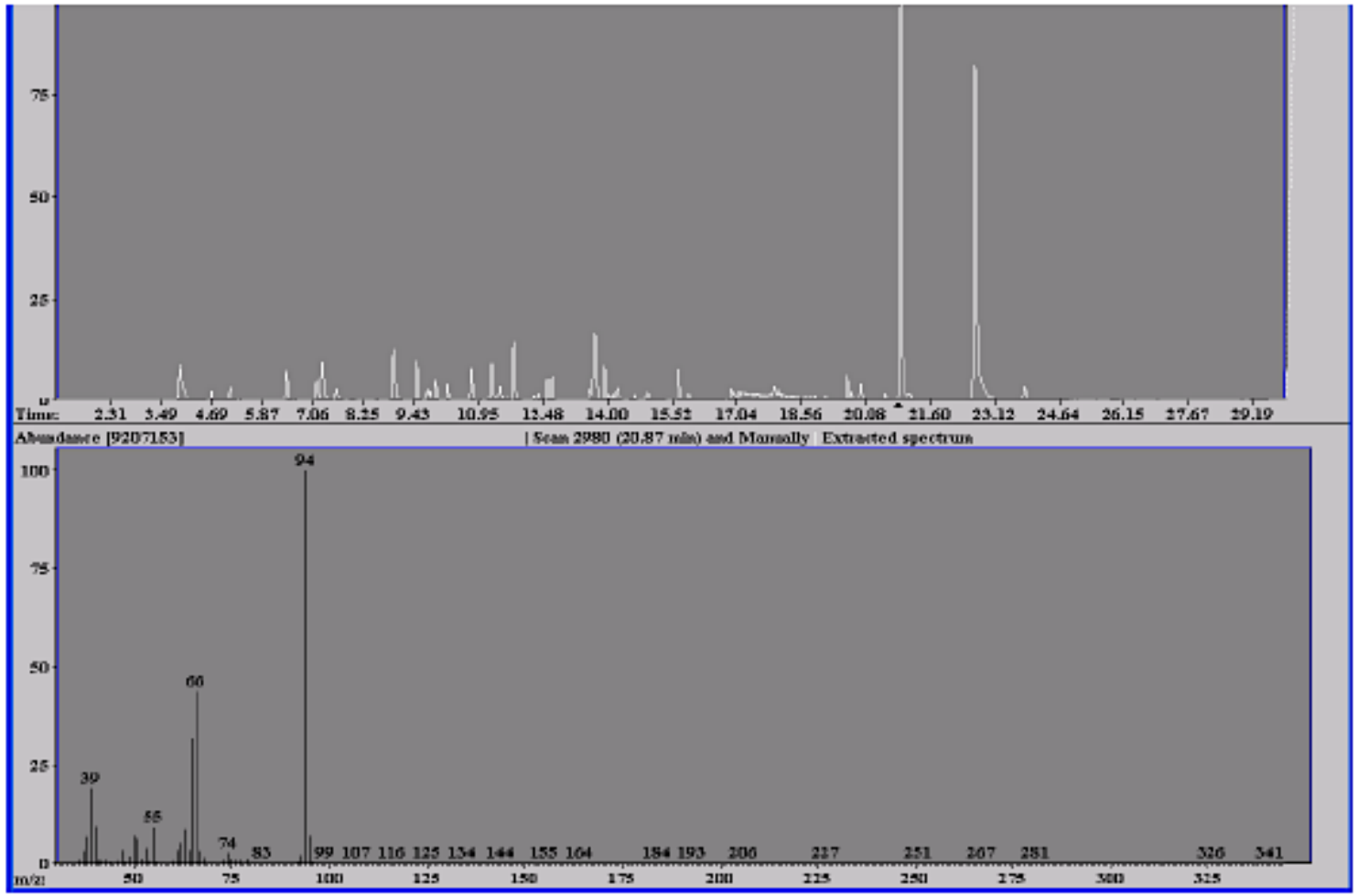




Biomarker Discovery

AMDIS/NIST

AMDIS:

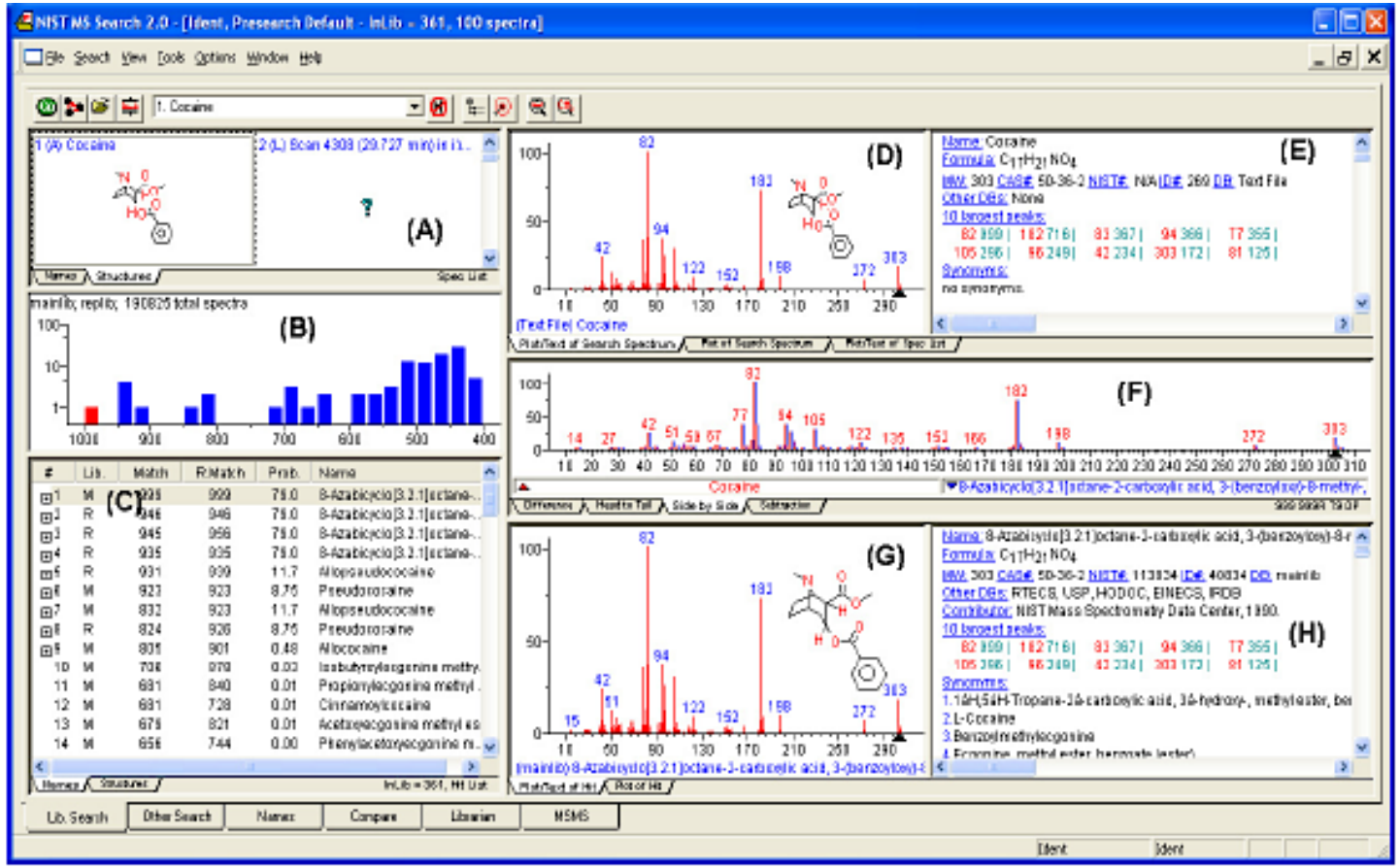


Automated Mass-spectral deconvolution and identification system

Biomarker Discovery

AMDIS/NIST

NIST:



(A) List of all spectra
(B) Histogram of selected hit
(C) List of hits matching the database
(D) Line plot of searched spectrum
(E) Data regarding the searched spectrum
(F) Mass spectrum to compare
(G) Line plot of selected spectrum
(H) Data regarding the selected spectrum

Biomarker Discovery

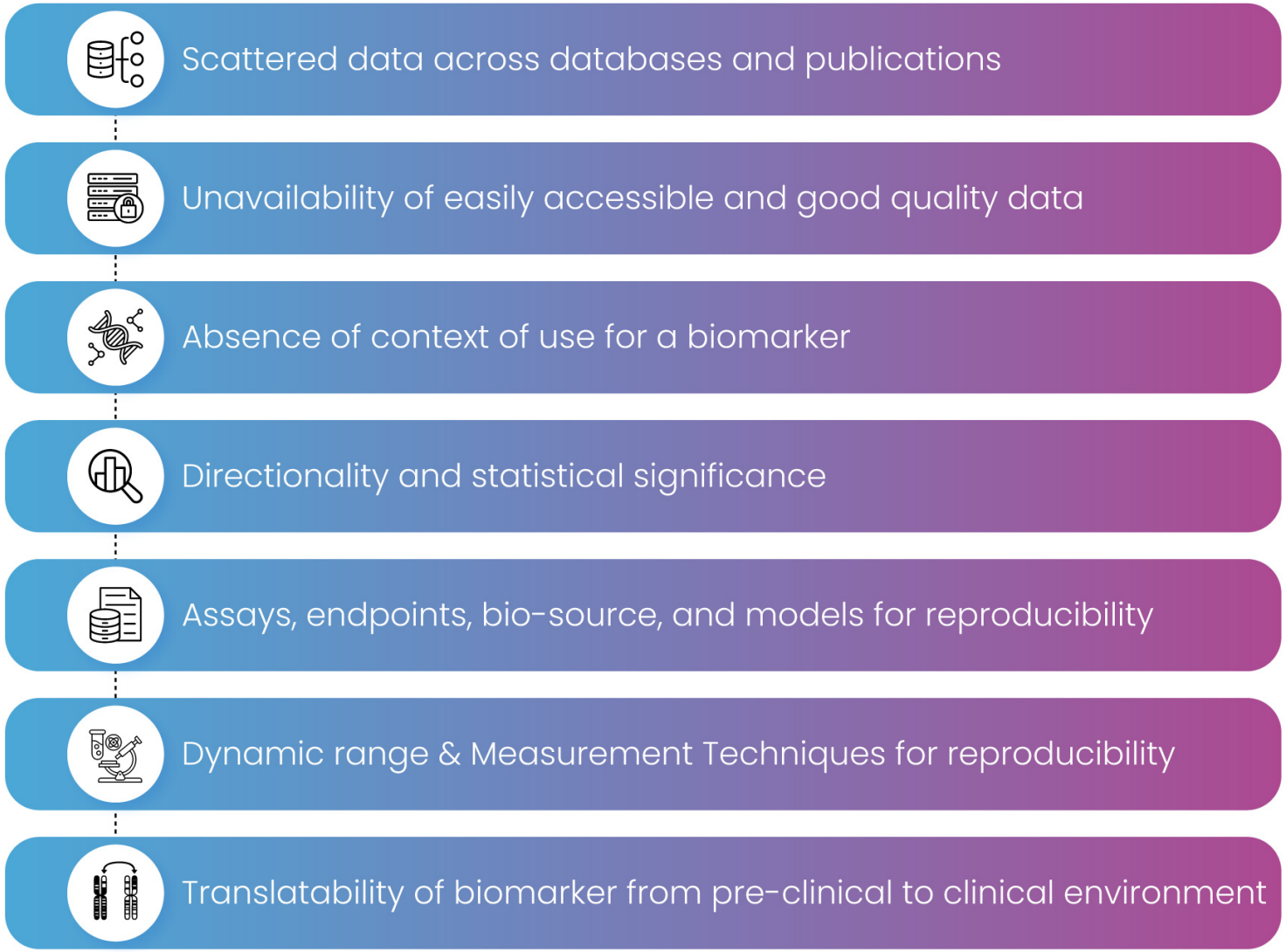
XCMS

- An open-source package available in R via BioConductor
- Offers the ability to analyse GC-MS, HPLC-MS, UPLC-MS and SFC-MS data
- Able to import data from different sources: NetCDF, mzXML, mzData and mzMl
- Permits the processing of the data: peak detection, alignment, filtration, peak matching
- System tends to run in the form of a workflow
    1. Import the data
    2. Group the samples
    3. Align the peaks
    4. Regroup the samples
    5. Check for missing peaks
    6. Generate results and reports*

* Calls the METLIN database

XCMS online http://xcmsonline.scripps.edu

Biomarker Discovery: Challenges



Summary

- Data pre-treatment.
- Exploratory data analysis using unsupervised methods:

  PCA, ICA, HCA, k-means.
- Multivariate Classification using statistically based methods (PLSDA)
- Biomarker discovery.