UNIVERSITY OF BIRMINGHAM

Metabolomics data analysis

LM Regulatory Science and Toxicology for the 21st Century

Dr. Ralf Weber (r.j.weber@bham.ac.uk)
Ossama Edbali (oxe410@student.bham.ac.uk)

UNIVERSITY OF BIRMINGHAM

Attendance code: **74095578**

Metabolomics and mass spectrometry

UNIVERSITY OF BIRMINGHAM

- **Metabolomics** is the large-scale study of small molecules, commonly known as metabolites, within cells, biofluids, tissues or organisms.
- **Mass spectrometry** (MS) is an analytical technique used to measure small molecules.
- **Direct infusion** is a mass spectrometry (MS) technique where a sample is introduced directly into the mass spectrometer without prior separation.

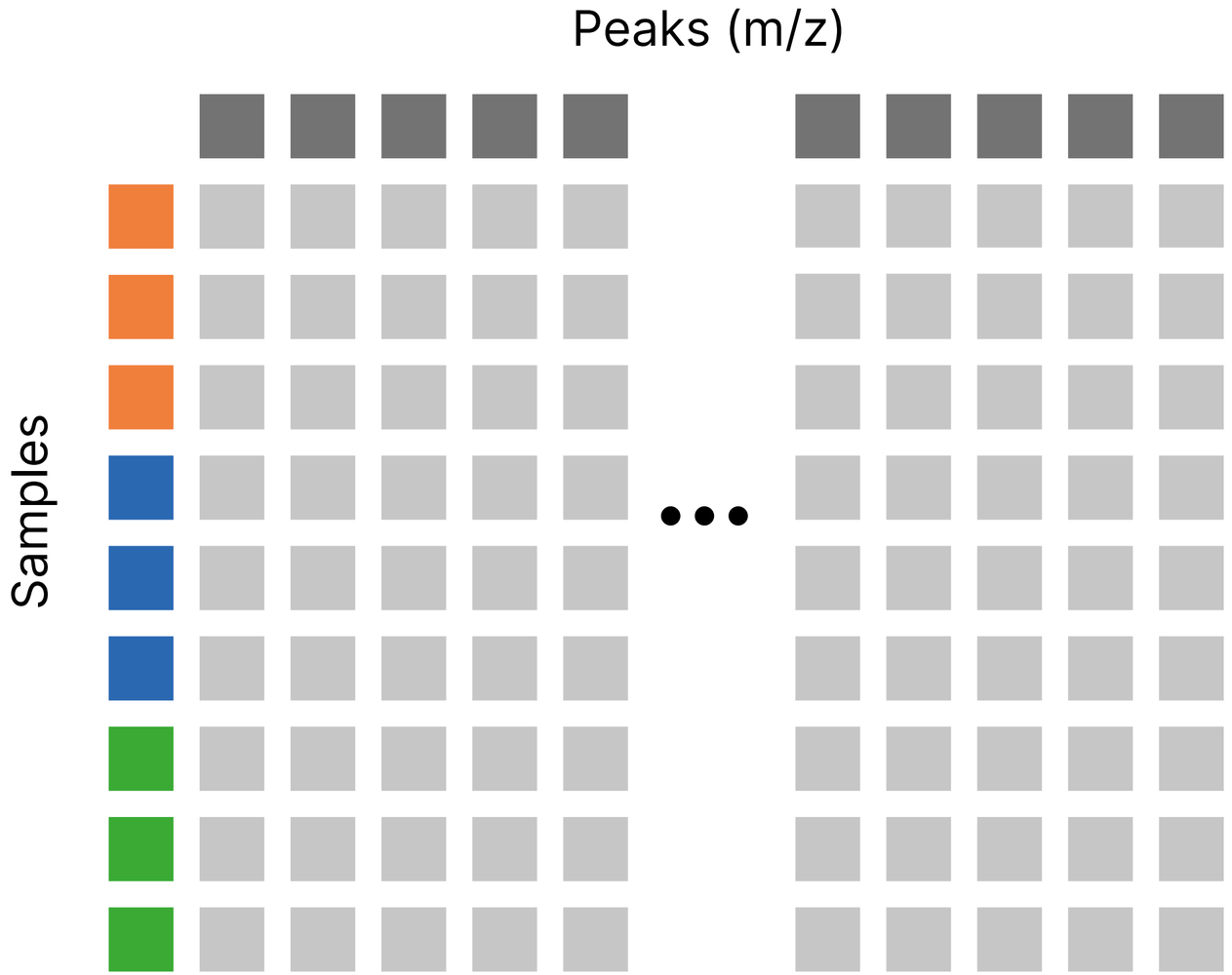
Sample types

UNIVERSITY OF BIRMINGHAM

- ***Blank***
  - What: consist of the exact solvent/buffer used to extract or dilute your samples, but without any biological material.
  - Why: Identify signals coming from the solvent, the plastic tubes, or the instrument itself.
- ***QC***
  - What: mixture created by taking a small aliquot from every single biological sample in your study (pooled).
  - Why: used for quality control.
- ***Biological***
  - What: actual experimental subjects
  - Why: answer the research question

DIMS data

UNIVERSITY OF BIRMINGHAM



Processing of DIMS data

UNIVERSITY OF BIRMINGHAM

Filtering

- Blank filtering
- Missing values filtering
- Low reproducibility filtering

- Missing value imputation

Normalisation

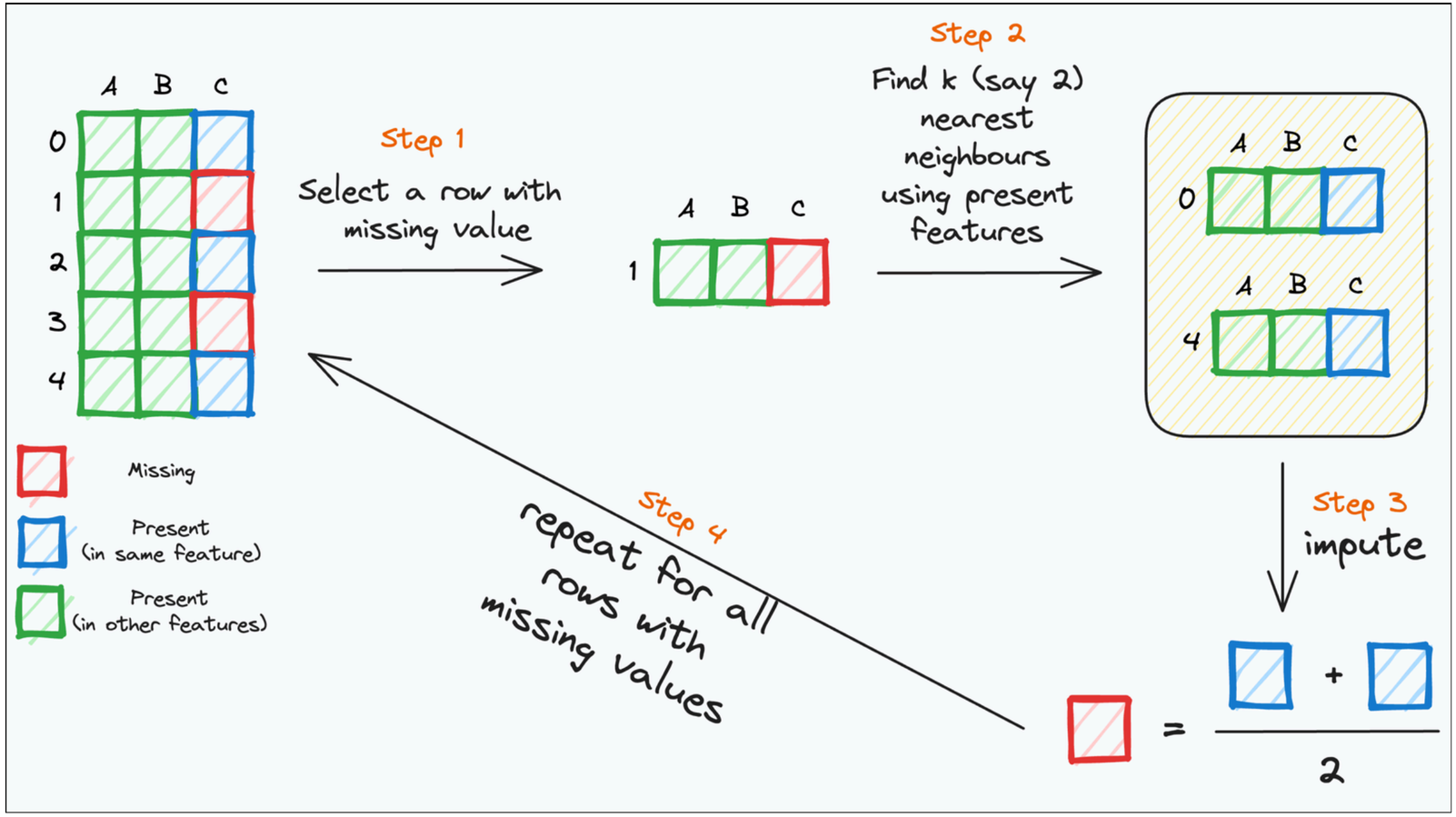
- Sample normalisation
- Data transformation
- Data scaling

Missing value imputation

UNIVERSITY OF BIRMINGHAM

- Missing values: a metabolite is truly absent (below the detection limit), or artifact miss.
- Before you run statistical methods like PCA, you have to impute missing values, as most algorithms can't handle empty cells.
- One of many methods: **k-Nearest Neighbors (kNN)** which estimates the missing value based on "similar" features/peaks or samples.

kNN missing value imputation (feature-wise)



Source: https://blog.dailydoseofds.com/p/the-most-overlooked-problem-with-768

Sample normalisation

UNIVERSITY OF BIRMINGHAM

- **Problem**: some of your samples might be more "diluted" than others.
- **A solution (normalisation by sum)**: add up the total signal of everything in a sample, and then divide every individual value by that total.

Data transformation

UNIVERSITY OF BIRMINGHAM

- **Problem**: biological data often spans huge ranges (e.g., 1×10^2 vs. 1×10^7). Statistical tests struggle with this: large values dominate the results while small ones get drowned out.
- **A solution (log transformation)**: take the logarithm of every value. “Squashes” the huge numbers down while keeping the order the same.

Data scaling

UNIVERSITY OF BIRMINGHAM

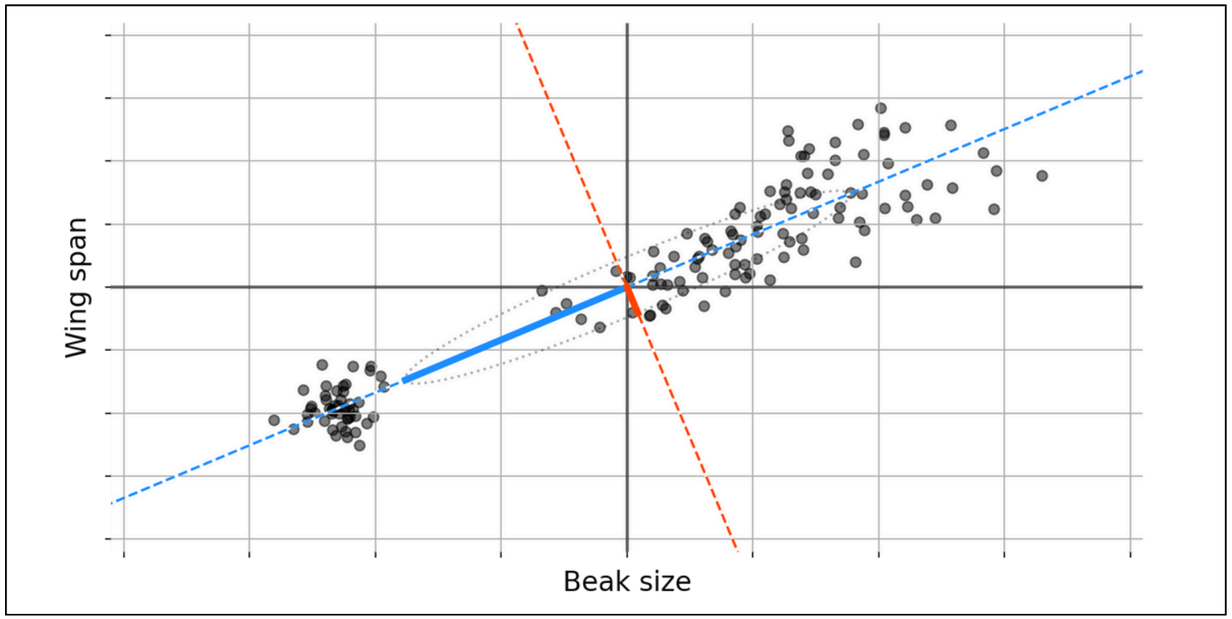
- **Problem**: the raw numbers are far away from a reference, making it hard for certain statistical methods to work.
- **A solution (mean centering)**: shifts each feature so its average is zero, removing scale offsets and letting all variables contribute equally.

Mean centering aligns the baseline, log transform compresses wide ranges

Principal Component Analysis (PCA)

UNIVERSITY OF BIRMINGHAM

Example from: https://gregorygundersen.com/blog/2022/09/17/pca/

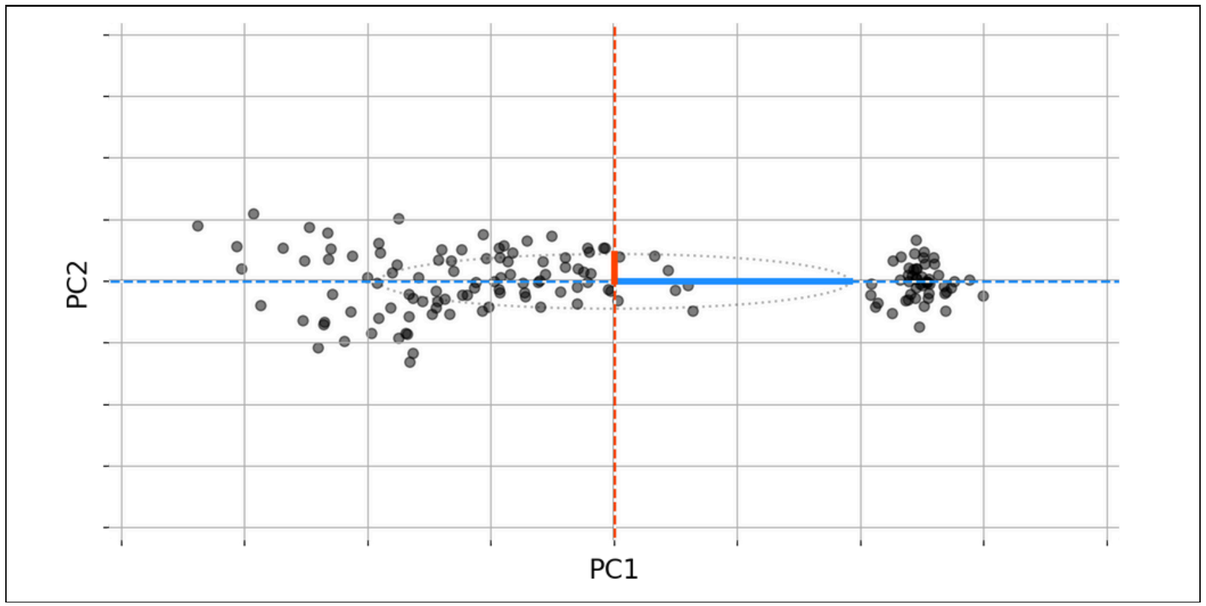


- The axis along which variation between birds is maximised is neither just beak size nor just wing span.
- Instead, the axis along which variation is maximised is some combination of both features, as represented by the blue dashed line (new composite feature).
- The direction of such feature is called principal component (PC)

Principal Component Analysis (PCA)

UNIVERSITY OF BIRMINGHAM

Example from: https://gregorygundersen.com/blog/2022/09/17/pca/



After finding the first PC, we find the next one by maximizing variance while requiring it to be orthogonal to the first, ensuring it captures a new, independent direction of variance.