18th Annual Conference of the Metabolomics Society

METABOLOMICS 2022

Valencia, Spain | JUNE 19-23

Pre-Conference Workshops

Spectra processing, functional integration & covariate adjustment of global metabolomics data using MetaboAnalyst 5.0

Jeff Xia, Associate Professor

jeff.xia@mcgill.ca | www.xialab.ca

McGill University, Montreal, QC Canada

Your TAs

Qiang

Jessica

Yao

Schedule

Part I: 12:00 PM – 2:00 PM

12:00 – 12:30: Opening lecture (Jeff)

12:30 – 12:45: Logistics

12:50 – 1:10: Section 1: LC-MS spectral processing and functional analysis (Qiang)

1:10 – 2:00: Interactive protocol exercise

Part II: 2:15PM – 4:15PM

2:15 – 2:30: Section 2: multi-omics integration using pathways and networks (Yao)

2:30 – 3:10: Interactive protocol exercise

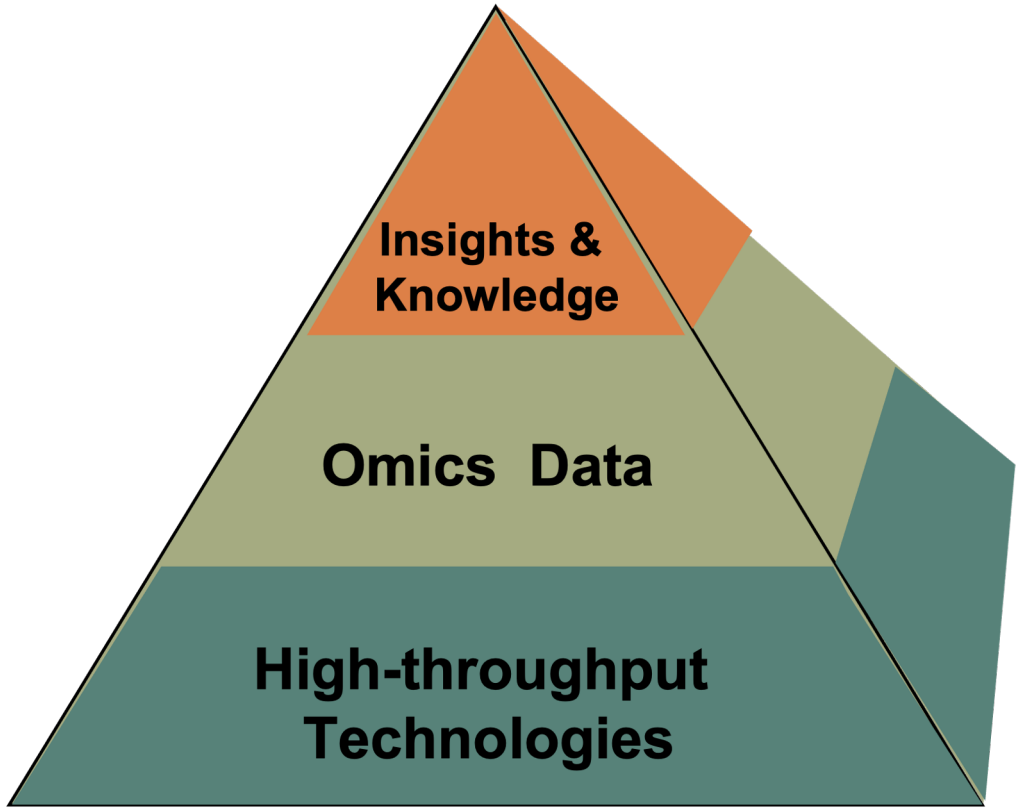
3:10 – 3:30: Section 3: Complex meta-data lecture (Jessica)

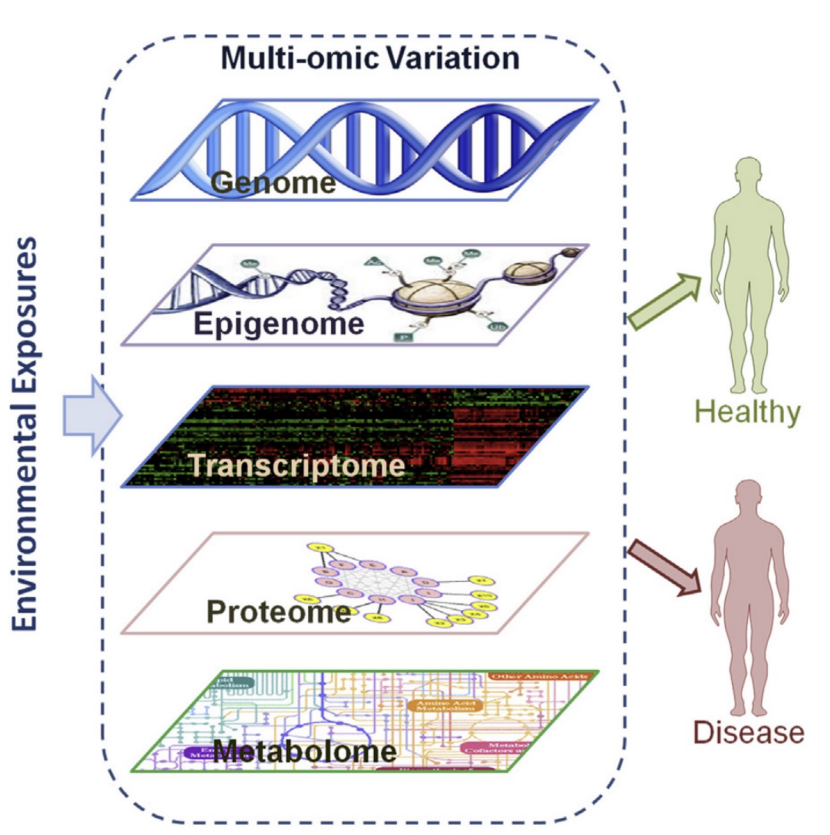
3:30 – 4:00: Interactive protocol exercise

4:00 – 4:15: Summary (Jeff)

A Gentle Introduction to Key Concepts in ‘Omics Data Analysis

Omics & multi-omics era





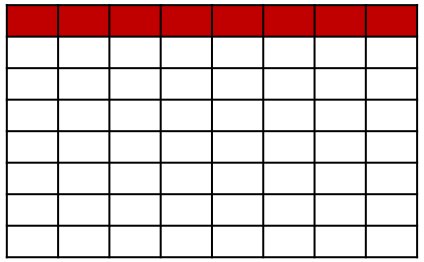
Two distinct challenges

Size challenge (raw data)

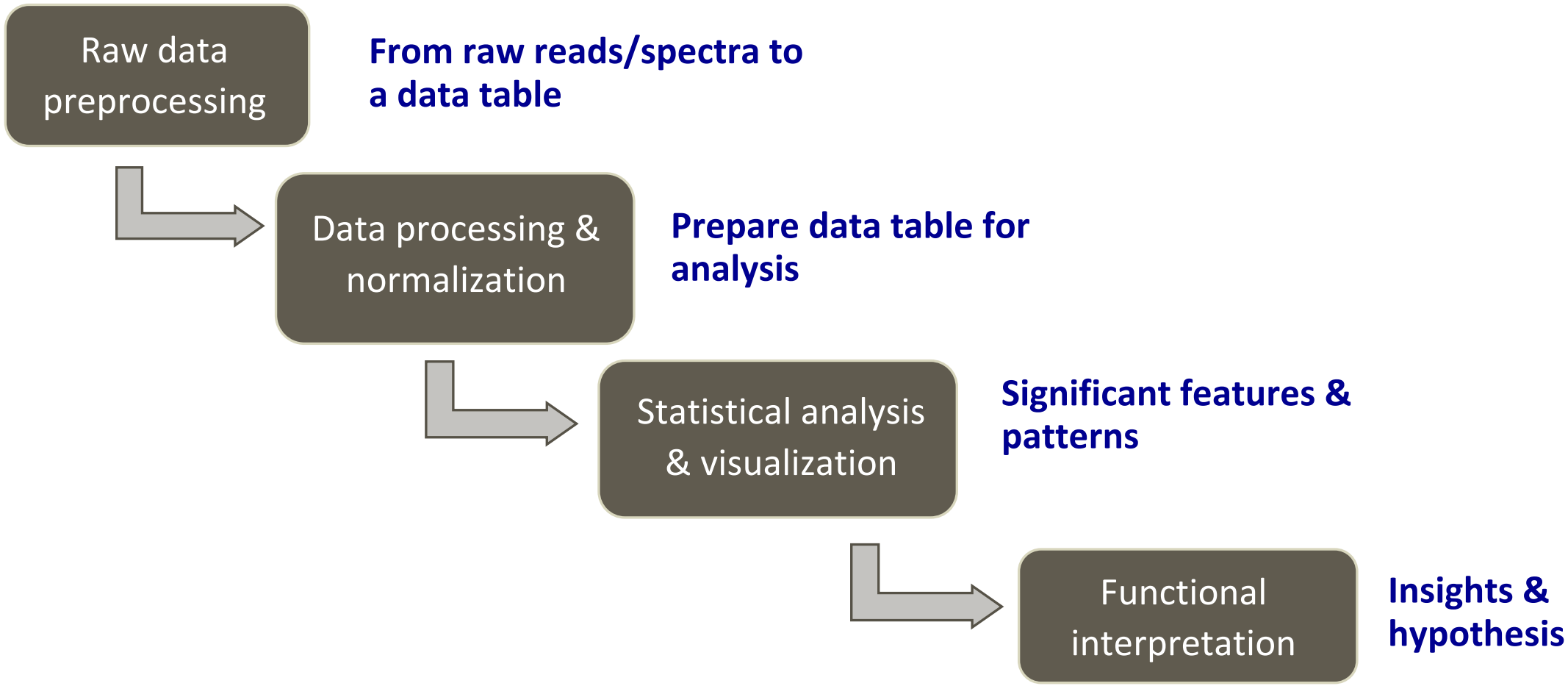
- Raw reads, spectra, images
- Large (GB ~TB)
- Large storage and computing resources

Complexity challenge (feature table)

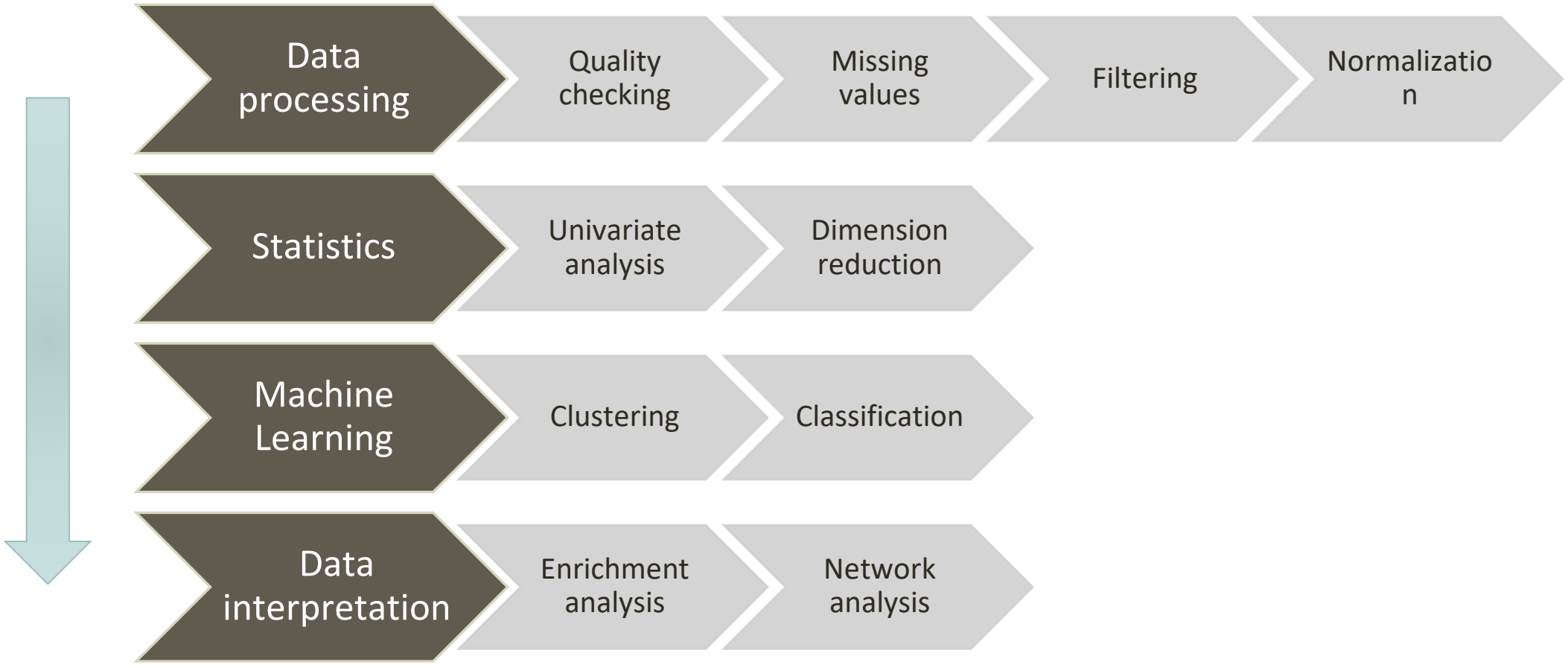
- Feature table (abundance, intensities)
- Small (100s KB ~ MB)
- High-dimensional, missing values
- Data integration usually starts here



Omics Data Analysis (in a nutshell)



Common steps in omics workflow



DATA PROCESSING

-- prepare data for main analysis

Data processing ….

General steps

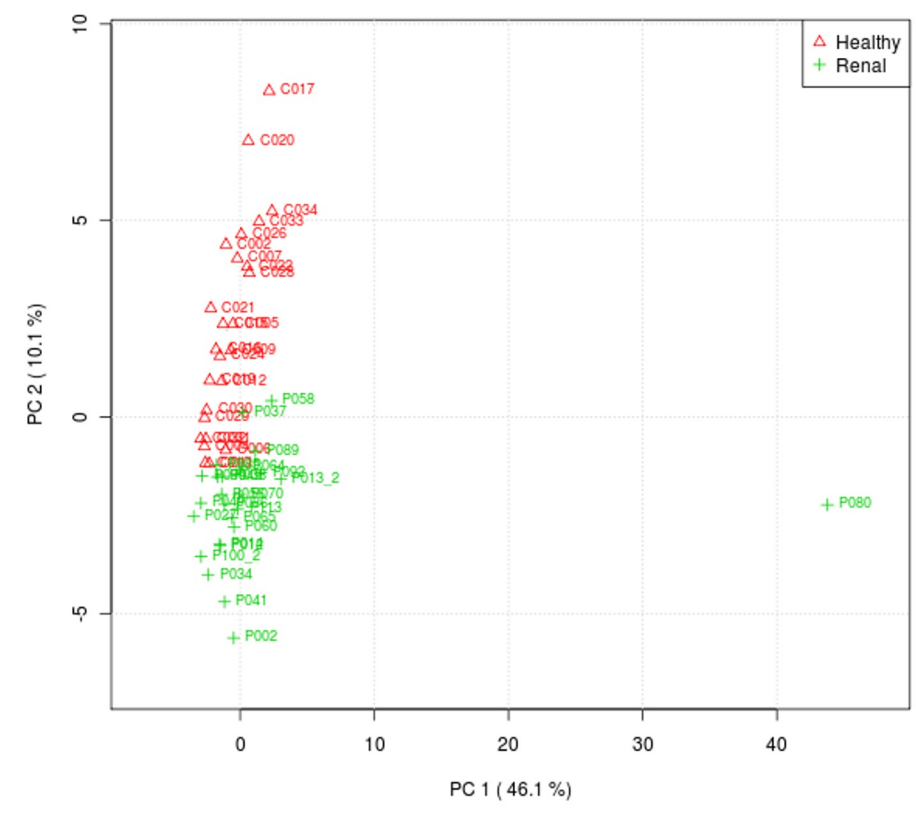
1. (Samples) Quality checking
2. (Features) Missing value imputation
3. (Features) Data filtering
4. (Both) Normalization

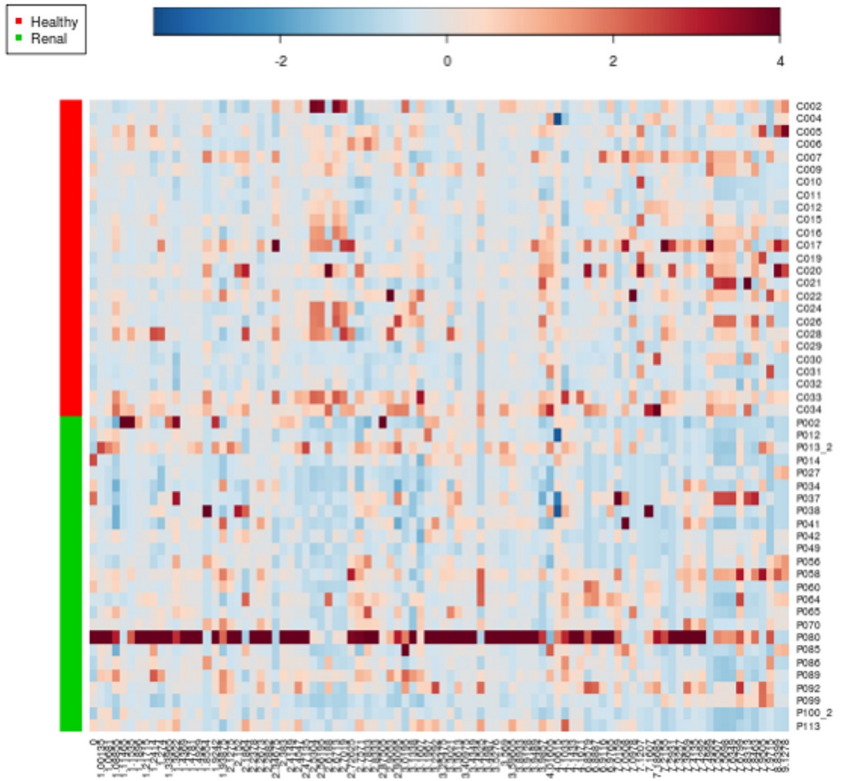
Quality checking

- The first & most critical step before analysis
  - Garbage in and garbage out
- Depending on
  - Good experimental design
  - Good laboratory practice
- Pay attention to
  - Outliers
  - Batch effects

Outliers (I)

- Relative to the majority

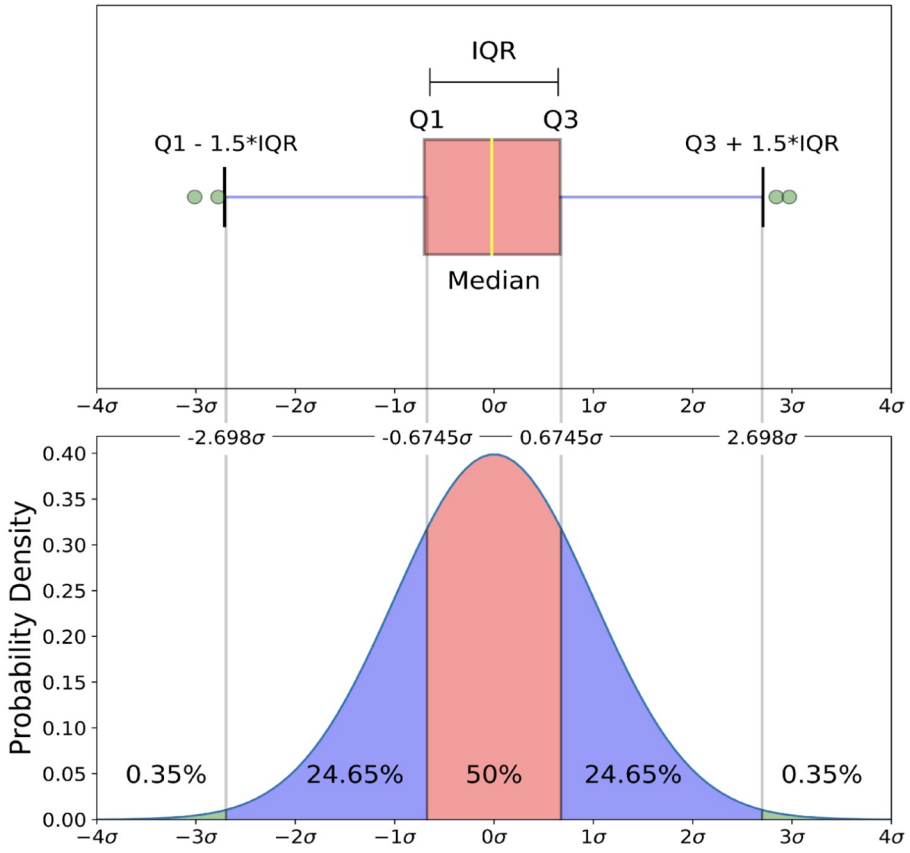




Outliers (II)

Mainly concerns **sample** outliers, not on feature space

- Features are measured en masse in omics experiments
- Features are dealt with in next steps: missing value imputation, filtering, normalization …..
- Statistical feature outliers could be our target of interest

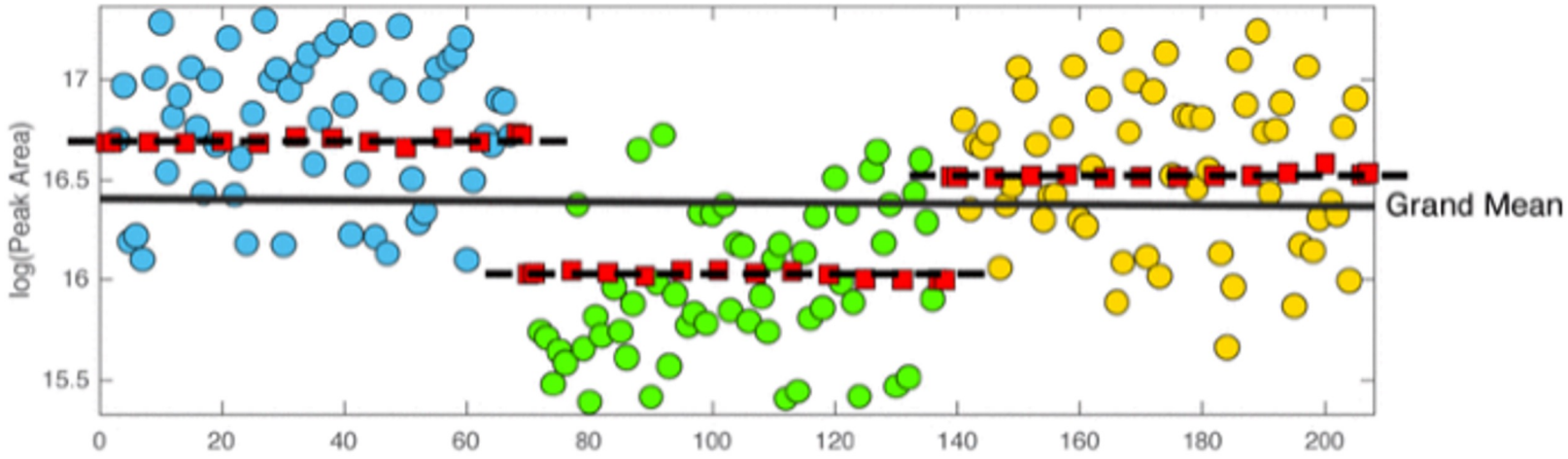


Source: towardsdatascience

Batch effects (I)

More relevant in large-scale studies when samples are measured across multiple days, multiple labs ….. (i.e., in different batches)

- Display overall or systematic differences
- Technical (not biological) reasons



Metabolomics (2018) 14:72

Batch effects (II)

Batch effect removal methods perform batch effect correction prior to statistical analysis;

- Internal standards
- Computational estimation

Batch effect adjustment methods include batch variables within the model for statistical analysis, such that differences associated with batch are accounted for during analysis.

Metabolomics (2018) 14:72

The sva package for removing batch effects and other unwanted variation in high-throughput experiments

JT Leek, WE Johnson, HS Parker, AE Jaffe… - …, 2012 - academic.oup.com

Heterogeneity and latent variables are now widely recognized as major sources of bias and variability in high-throughput experiments. The most well-known source of latent variation in genomic experiments are batch effects—when samples are processed on different days, in different groups or by different people. However, there are also a large number of other variables that may have a major impact on high-throughput measurements. Here we describe the sva package for identifying, estimating and removing unwanted sources of …

☆ Save 99 Cite Cited by 2763 Related articles All 34 versions Web of Science: 1788

Try to reduce batch effect during data collection

Missing values (I)

- Common in omics data. Can be introduced during data collection, or by algorithms during raw data pre-processing (i.e. peak picking)
- Most algorithms will complain if input contains missing values

1.4781	2	1.05	1.84	0.89	1.33	1.94	1.43	0.85	1.52	1.48	2.52	2.22
1.4929	2.03	1.06	1.86	0.88	1.33	1.13	1.46	0.88	0.97	1.47	1.88	2.19
1.8554	NA	0.47	0.83	NA	1.31	NA	NA	NA	NA	0.6	NA	0.79
1.9242	1.82	1.59	1.45	1.73	3.13	1.91	1.79	1.58	1.77	1.99	3.26	3.35
1.93875	NA	0.76	1.1	0.83	2.62	0.8	1.53	1.45	0.94	1.8	NA	3.36
2.1275	1.19	0.72	NA	NA	2.88	NA	1.68	NA	0.94	1.1	NA	3.16
2.152	NA	2.25	2.9	1.25	8.75	5.02	1.09	1.91	1.33	3.13	2.84	4.18
2.1864	NA	1.3	2.7	0.8	3.47	2.84	NA	0.9	NA	1.98	2.05	2.35
2.2378	1.58	0.61	1.03	1.75	0.84	1.51	0.72	0.77	1.15	0.85	0.79	0.96

Missing value (II)

- Missing value estimation could have a large impact on the downstream analysis, when a large portion of features are missing
- Why missing?
  - Missing completely at random
  - Missing at random
  - Missing not at random
- Be cautious not to introduce bias, and base our conclusion on high-quality data

Default in MetaboAnalyst 5.0,

1. Remove features with a high-level (adjustable) missing values
2. Fill the remaining NAs using detection limits (LoD) - 1/5 of the lowest positive values reported for individual features

MissForest—non-parametric missing value imputation for mixed-type data

DJ Stekhoven, P Bühlmann - Bioinformatics, 2012 - academic.oup.com

Motivation: Modern data acquisition based on high-throughput technology is often facing the problem of missing data. Algorithms commonly used in the analysis of such large-scale data often depend on a complete set. Missing value imputation offers a solution to this problem. However, the majority of available imputation methods are restricted to one type of variable only: continuous or categorical. For mixed-type data, the different types are usually handled separately. Therefore, these methods ignore possible relations between variable types. We …

☆ Save 99 Cite Cited by 2433 Related articles All 20 versions Web of Science: 1441

Feature filtering (I)

- Not all features are informative
- There are redundancies in omics data for most features
- Filtering non-informative features before statistical analysis can often significantly improve the power

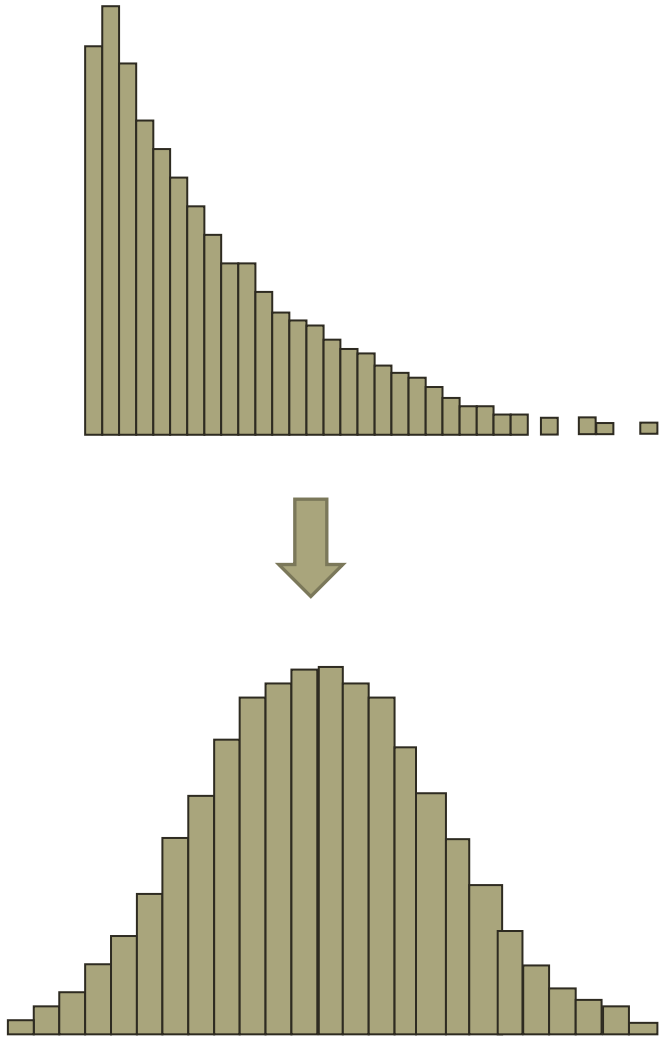
Feature filtering (II)

- **Low quality**
  - Too many missing values
  - Hard to measure: low repeatability based on QC
- **Low abundance**
  - Variables of very small values (close to baseline or detection limit).
- **Low variance**
  - Variables that are near-constant values throughout the experiment conditions (housekeeping or homeostasis)

DO NOT filter features based on their p-values or fold changes

Normalization (I)

- Most statistical methods works best when variables are normally distributed
- Variable abundance levels can vary across several magnitudes - the changes of abundant variables will dominate analysis (i.e., PCA) if unadjusted
  - More abundant features do not mean they are more important
- Adjust other effects:
  - Dilutions, tissue volumes, etc

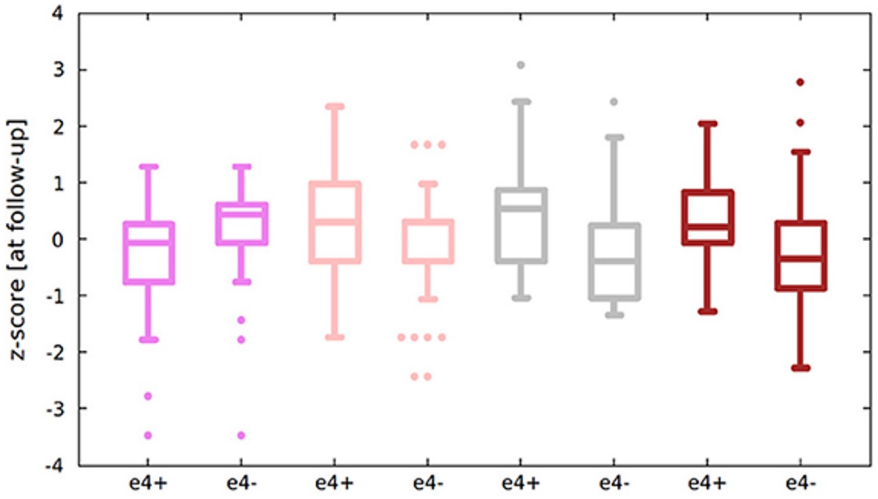


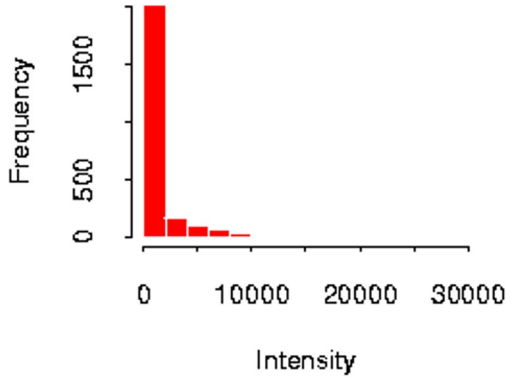
Normalization (II)

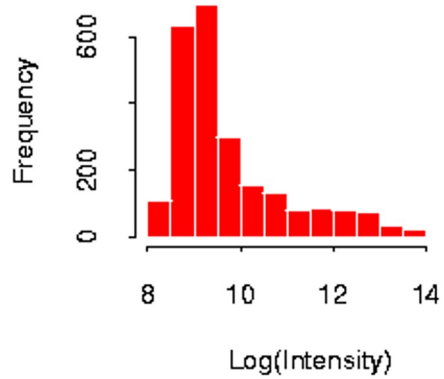
But ….. normalization often makes data difficult to interpret

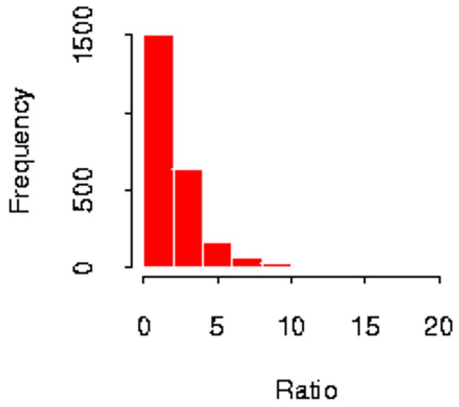
Try simple methods first

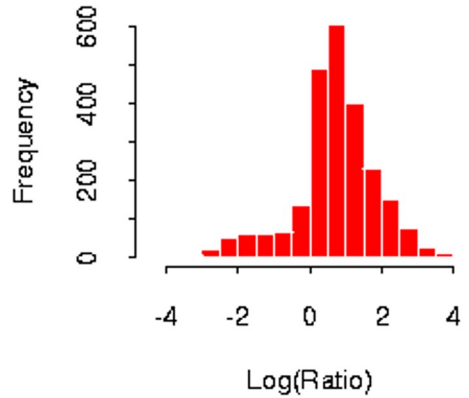
- Most physiological measures are log-normal
- Auto-scale (unit transformation, or Z-score)











Normalization (III)

Many methods are available

- Centering
- Scaling
- Transformation

There is NO guarantee of global normal distribution in omics data

Method	Formula	Unit	Goal	Advantages	Disadvantages
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	0	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{\max}} - x_{i_{\min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	0	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log 0	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√0	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

BMC Genomics. 2006; 7: 142

STATISTICAL ANALYSIS

-- identify significant features & patterns

Univariate analysis

Test each feature individually (ignore their potential correlations)

1. T-tests
   - Compare the means between 2 conditions
2. ANOVA & post-hoc analysis
   - One factor with more than 2 levels (One-way ANOVA)
   - Two factors (Two-way ANOVA)
3. Linear modeling (i.e., limma): more flexible analysis
   - Multiple factors
   - Time series
   - Covariates analysis

All these approaches are now available in MetaboAnalyst 5.0

P-value & multiple testing issue

1. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed if the null hypothesis is true
2. One "rejects the null hypothesis" when the p-value is less than the significance level α which is often 0.05 or 0.01
3. When the null hypothesis is rejected, the result is said to be statistically significant

Performing T-tests on typical metabolomic data might result in performing ~10000 separate hypothesis tests. If we use a standard p value cut-off of 0.05, we would see 500 (10000*0.05) features to be deemed “significant” by chance!

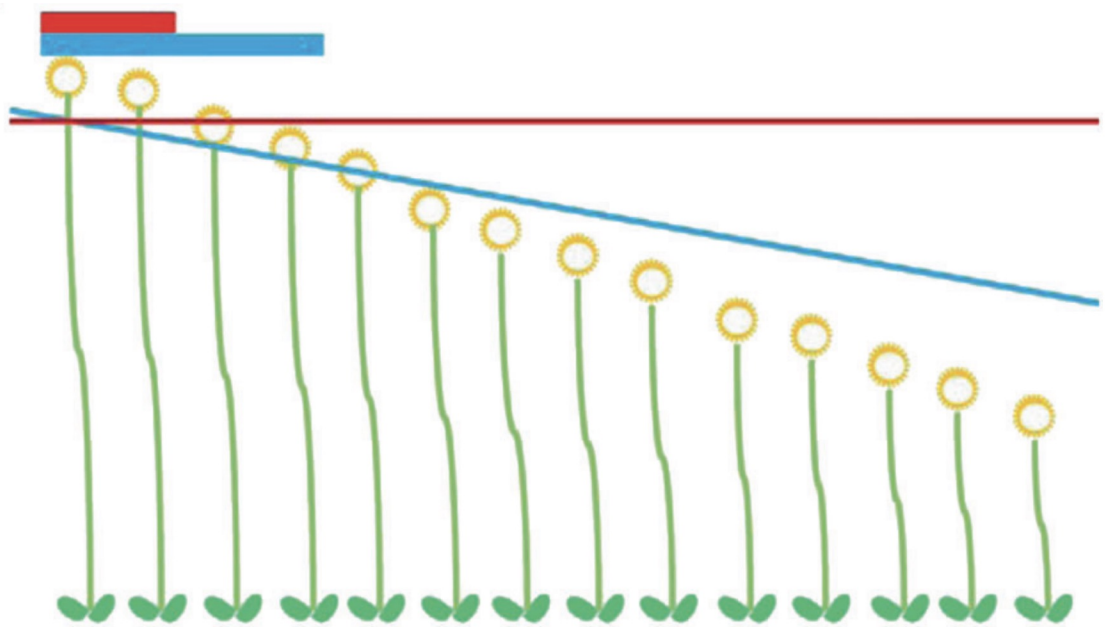
Multiple testing correction

Family-wise error rate (FWER)

- Apply the same cut-off value (red line) to all tests
- Such as Bonferroni correction

False-discovery rate (FDR)

- Apply different cut-off values (blue line) depending on the rank of the feature
- In 10,000 features, after the top 1 (the smallest p values) considered significant; the total number of null hypothesis become 9999, and the threshold is adjusted accordingly
- FDR of 0.05 means that 5% among the significant metabolites are expected to be false positives



Journal of Human Genetics (2021) 66:93–102

Empirical P-values

Previously mentioned p-values are based on well defined models (i.e., normal distributions)

What if we don’t know the distribution?

- The only thing we know is that the data does NOT follow a normal distribution
- Poor performance using a normal distribution-based model

We can find out the null distribution from the data itself (data needs to be relatively big), then calculate the p-value (also known as empirical p-value)

Basic steps

1. Under the null hypothesis, all data comes from the same distribution
2. We calculate our statistic, such as the mean difference, from the original data
3. We then shuffle the data with respect to group labels and recalculate the statistic (mean difference)
   - group labels do not matter under H_0
4. Repeat step 3 many many times
5. Find out where our statistic lies in comparison to the null distribution

A simple example

To find out whether there is a mean difference between original case vs. control groups

Mean difference .541

	case		control
1	-0.49274	10	1.471227
2	-0.30228	11	0.612679
3	0.093007	12	-0.47886
4	0.715722	13	0.746045
5	1.272872	14	0.871994
6	-1.37599	15	0.985237
7	-0.14798	16	-0.44421
8	-1.22195	17	0.246393
9	1.2812	18	0.68246
Mean	-0.01979		0.52144

Permutation #1

Re-assign samples randomly to different groups

- Note how the different labels have been swapped for the permutation

Mean difference = .329

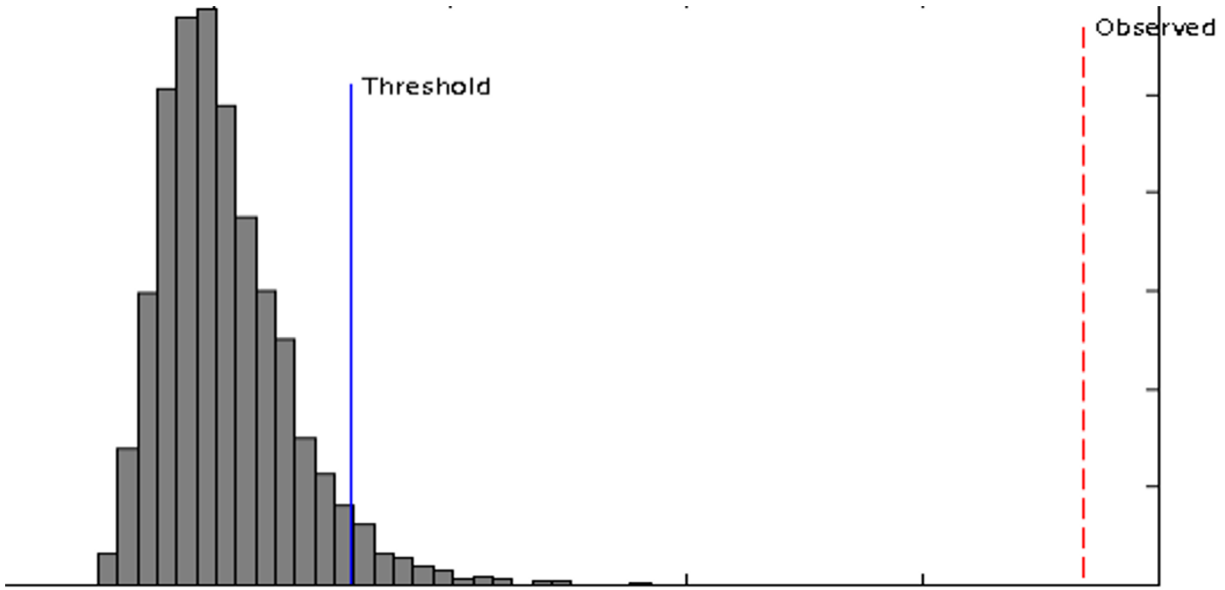
	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
16	-0.44421	6	-1.37599
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Permutations continued ….

	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Repeat many many times (i.e. 1000 times)

Compute empirical p-values



In 1000 times permutations

- There are three times the permuted data given large difference
  ➔ p = 0.003
- None of the permuted mean difference is bigger than the original one
  ➔ p < 0.001 or (1/1001) # prevent p-value equal to zero

From univariate to multivariate analysis

So far, our discussed analyses dealing with a single variable with regard to different experimental design

1. First analyze a single variable, and then apply the procedure to all variables, finally do multiple test adjustment
2. Visualization are limited to 2/3 dimensions

How can we analyze & visualize high-dimensional data simultaneously?

Principal Component Analysis (PCA)

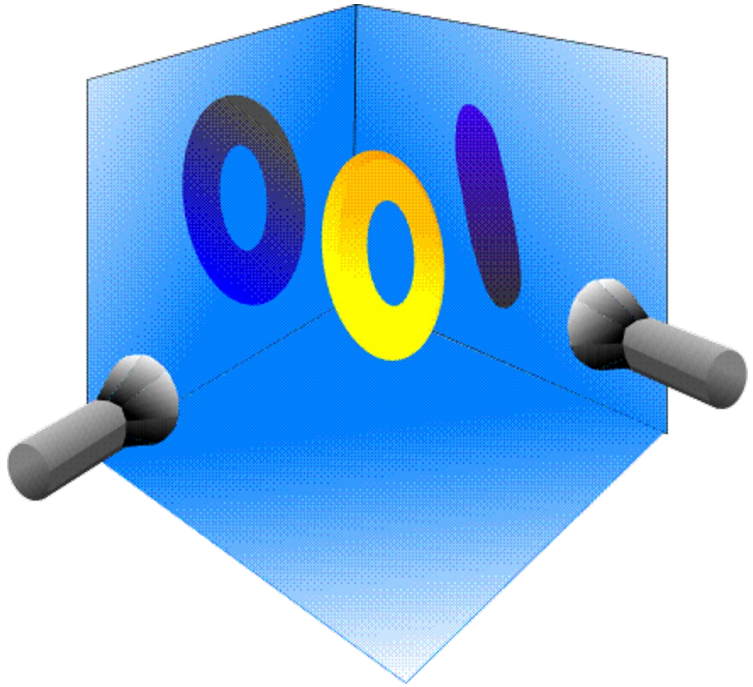
Project high-dimensional data into lower dimensions that capture the most variance of the data

Assumption:

Main directions of variance

≈ major data characteristics

PCA – projecting data to lower dimension

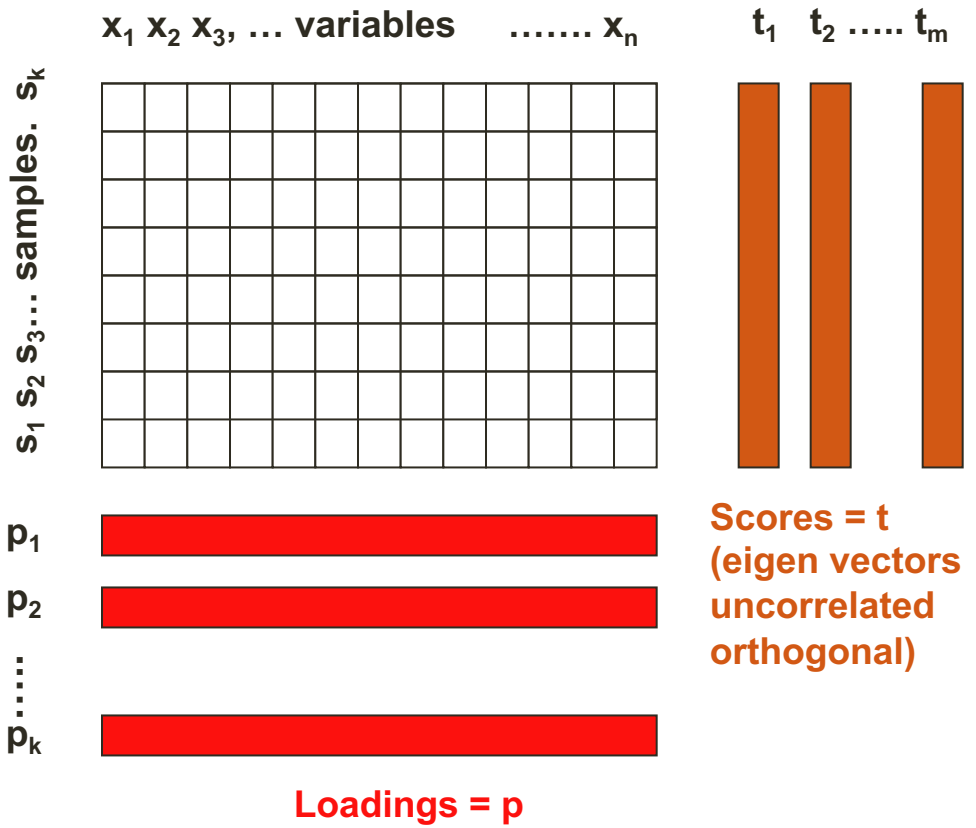


PCA of a “bagel”

- One projection produces a weiner (hotdog)
- Another projection produces an “O”
- The “O” projection captures most of the variation and has the largest eigenvector (PC1)
- The weiner projection is PC2 and gives depth info

PCA – linear transformation

- PCA is an orthogonal linear transformation
- PCA transforms data to a new coordinate system so that the greatest variance of the data comes to lie on the first coordinate (1st PC), the second greatest variance on the 2nd PC etc.

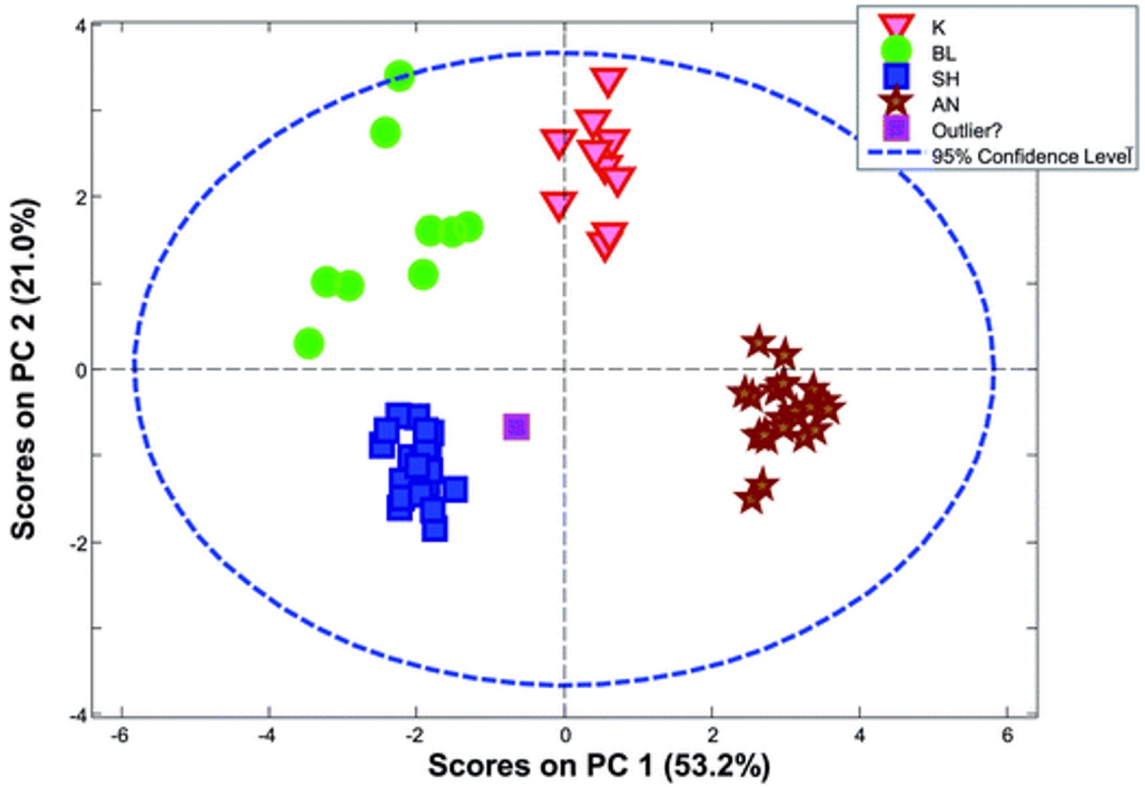


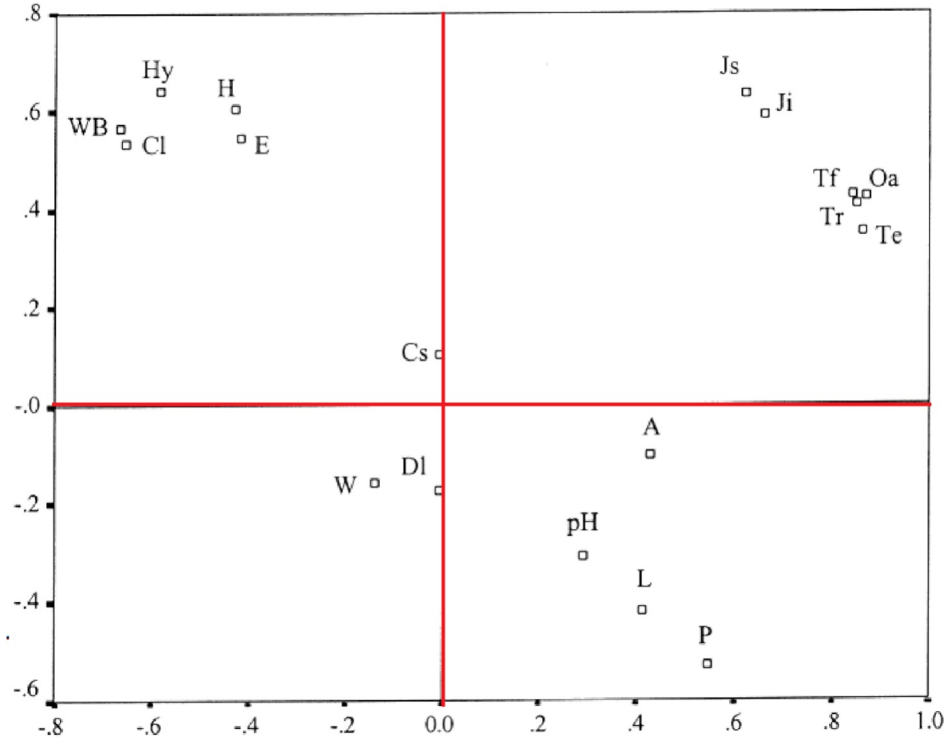
Intuitive interpretation

Scores = Loadings x data

$t_1 = p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n$

.....





Sample patterns (scores) are directly related to feature patterns (loadings)

PCA summary

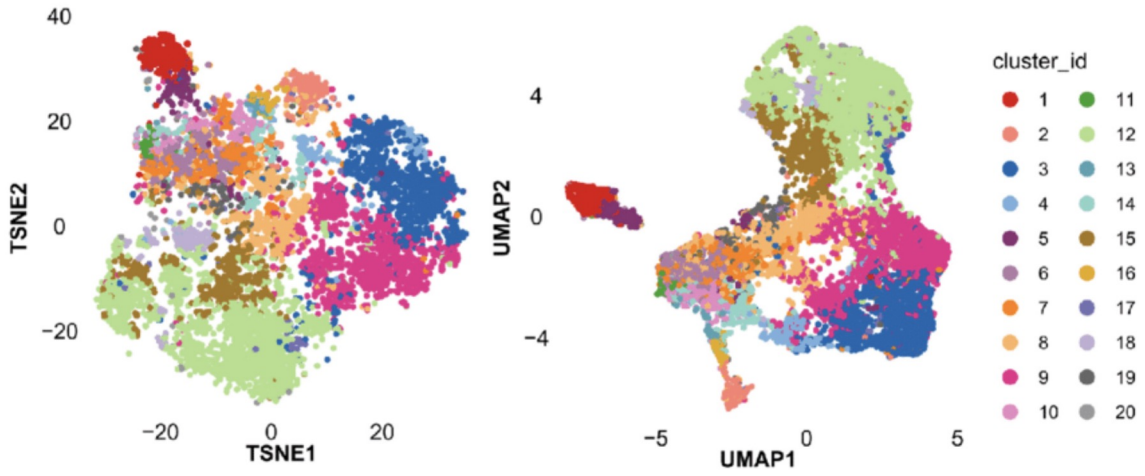
PCA rotates multivariate dataset into a new configuration which is easier to interpret. It is widely used for:

- Data overview
- Outlier detection
- Find out relationships between variables

PCA is a linear method for dimension reduction.

There are non-linear methods

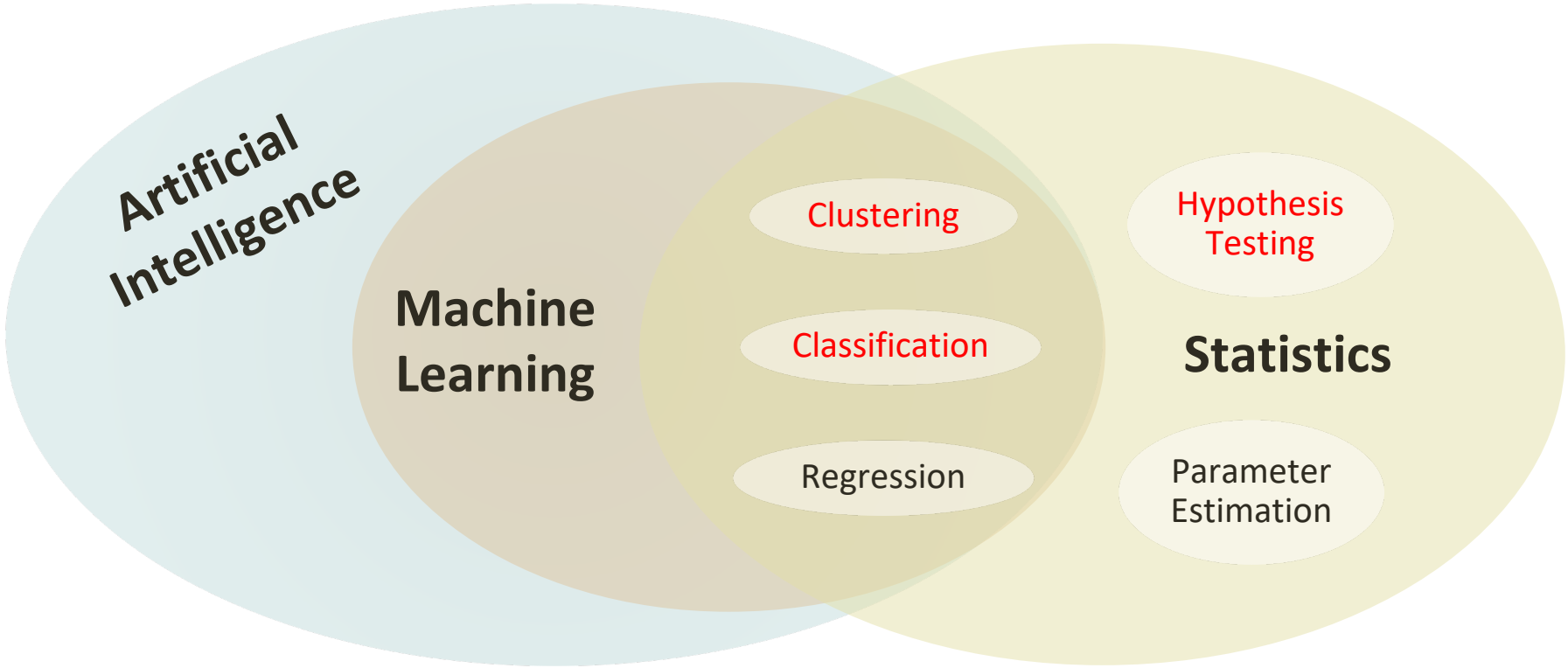
- t-SNE, UMAP, etc



MACHINE LEARNING

-- clustering and classification

Big Data Analytics



Machine Learning

Unsupervised learning: explore the data to find some intrinsic structures in them (disregard whether they are related to the class labels or not)

Supervised learning: discover patterns in the data that relate data attributes with related to a target (class) attribute.

- These patterns can then be utilized to predict the values of the target attribute in future data instances

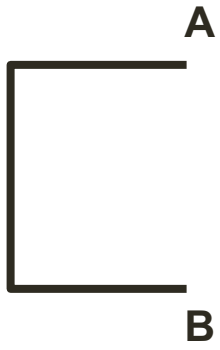
Clustering concept

A process by which objects that are logically similar in characteristics are grouped together

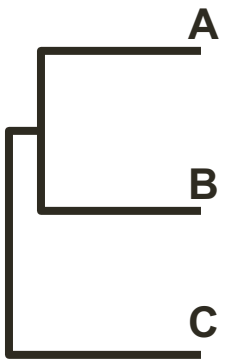
- Organize the 10000s of variables into blocks (clusters)
- Variables in each block are more similar to each other (i.e. homogenous), and we can treat each block as a single unit.
  - For instance, if in average, each block contains 200 variables, then 1000 compounds can be reduced to five blocks (five conceptual new variables)

We can then focus on understanding each cluster / block

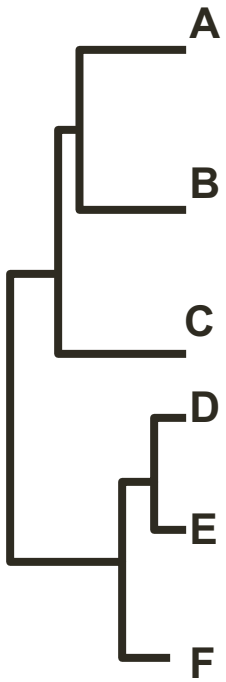
Hierarchical clustering & heatmap



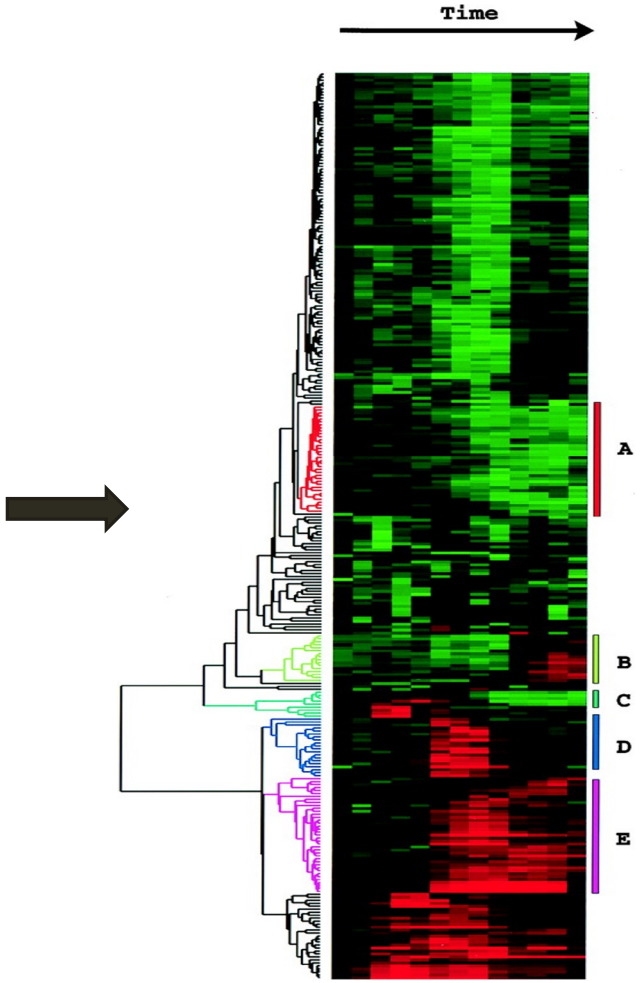
Find two most similar metabolite concentration levels



Find the next closest pair of levels or curves

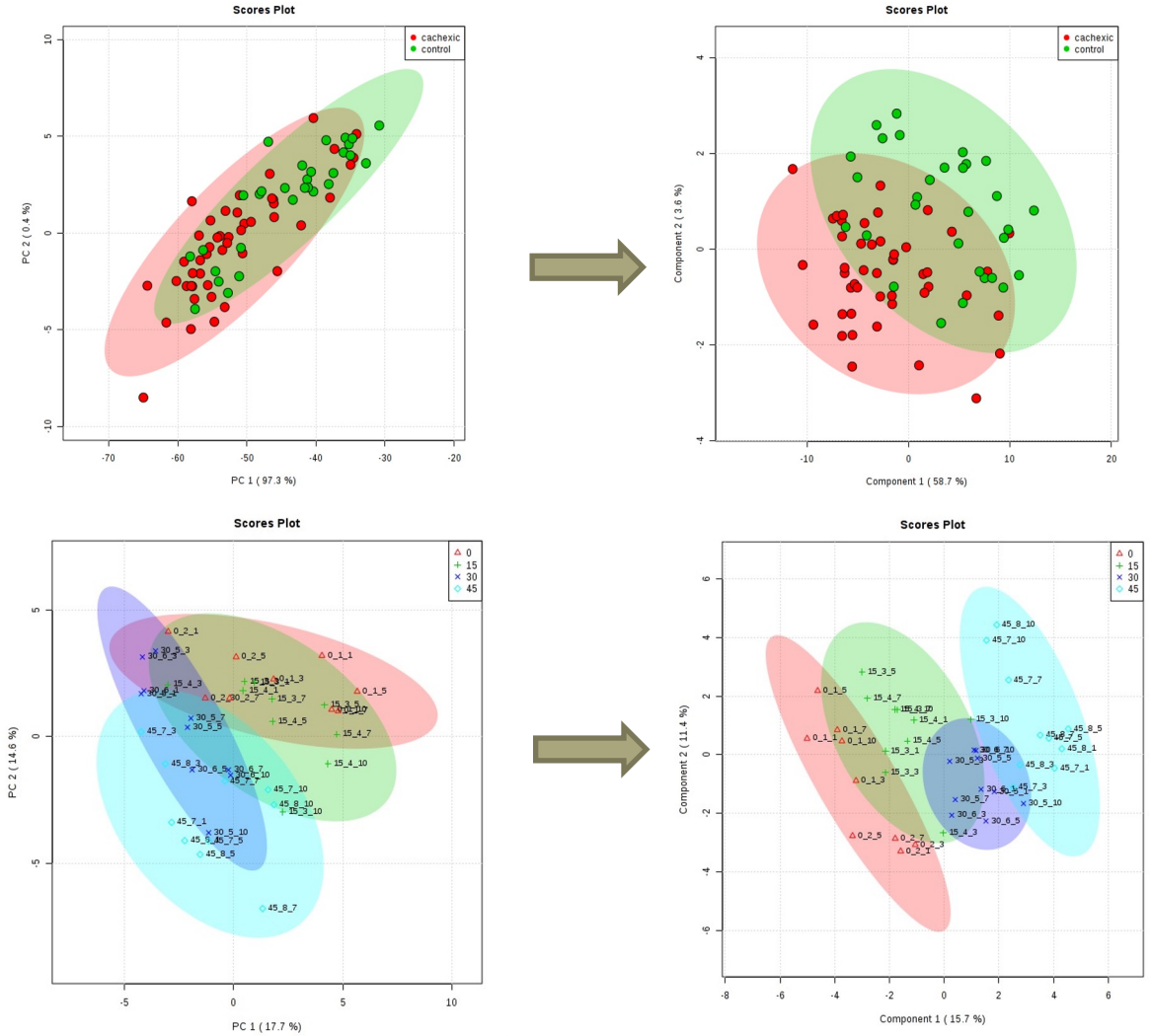


Iterat e



From unsupervised to supervised classification

PCA vs.
PLS-DA



Partial least squares-discriminant analysis (PLS-DA)

When the experimental effects are subtle or moderate, PCA will not show good separation patterns

- PCA is unsupervised (it does not consider group labels)
- PLS-DA is a supervised method, it is calculated by maximizing the co-variance between the data matrix (X) and the class labels (Y)

Caution! PLS-DA always produces certain separation patterns with regard the conditions

Use PLS-DA with caution

PLS-DA is based on regression

It first converts the class labels into **numbers** and then performs PLS regression between the data matrix and numerical Y

- Which means different group labels could lead to different separation pattern!
- See the next page

PLS-DA is susceptible to over-fitting by producing patterns of separation even for data randomly drawn from the same population

- Need for cross validations
- Need to do permutation tests

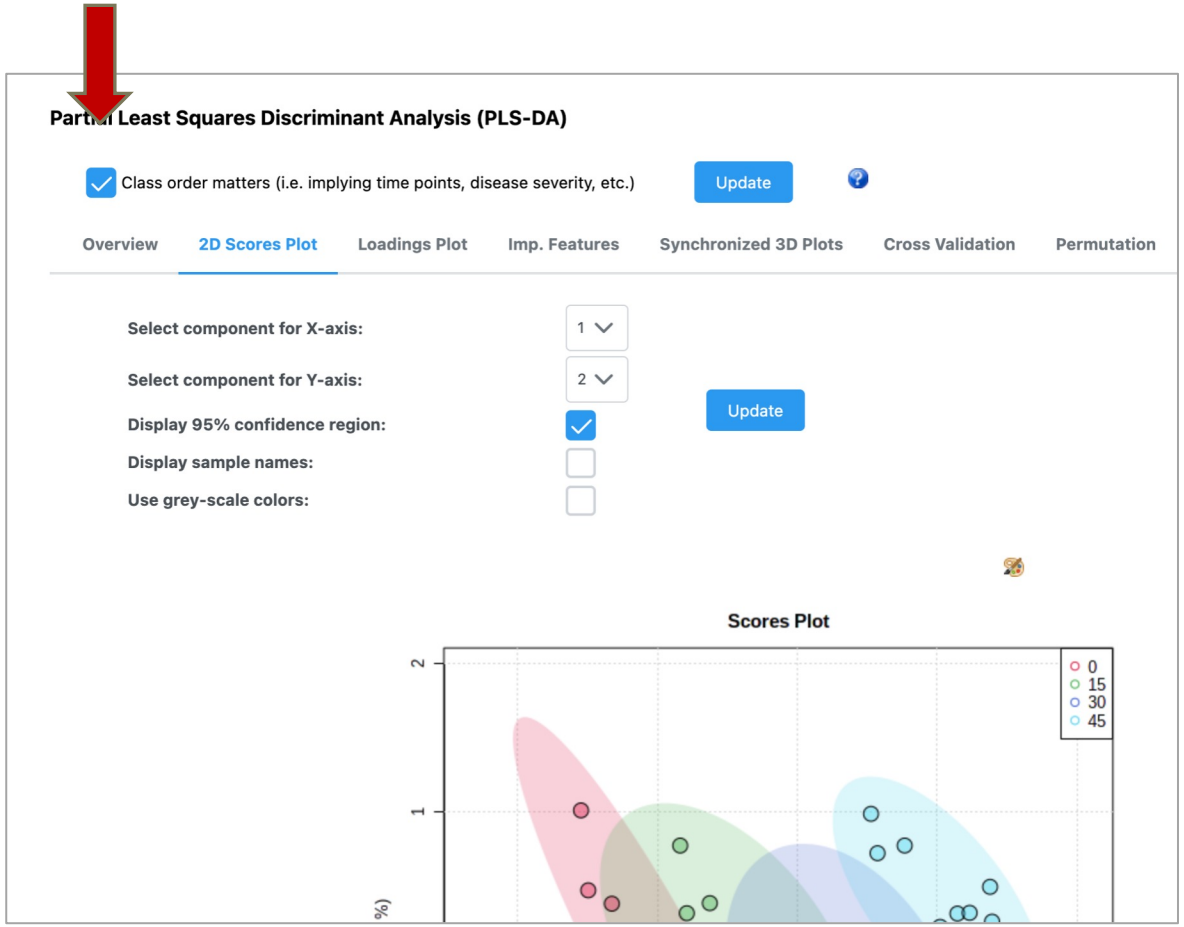
Use PLS-DA for multiple groups

PLS-DA first converts the class labels into numbers and then perform PLS regression between the data matrix X and numerical Y.

- "A, B, C" will be 1, 2, 3; "low, medium, high" will be 2, 3, 1

Using different class labels does **NOT** affect the classification accuracy, but may lead to different separation patterns derived from regression

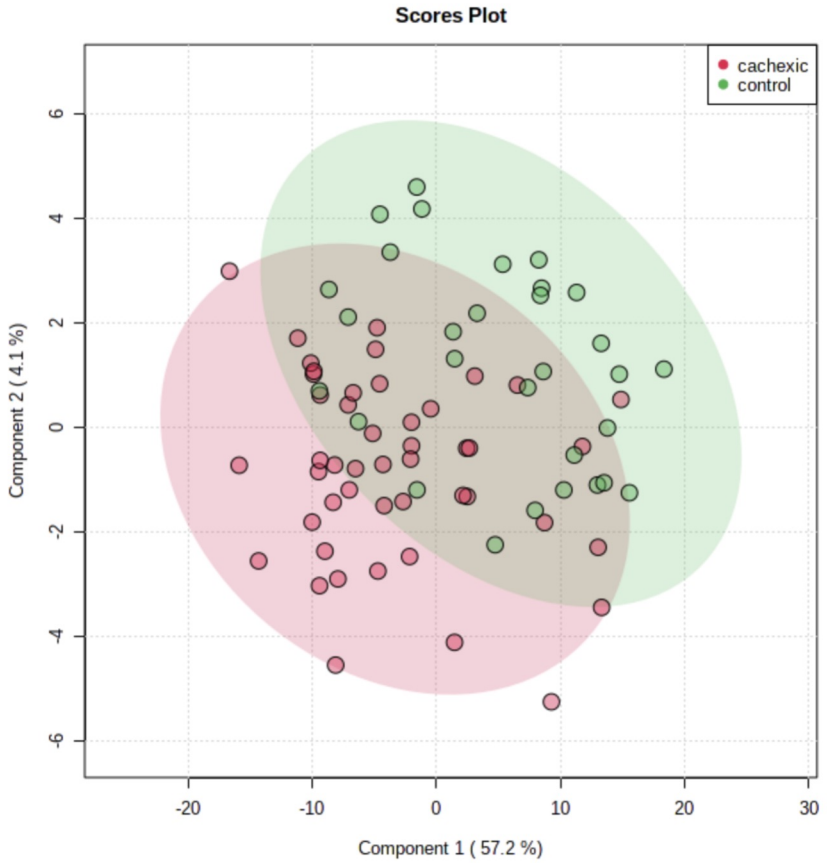
- Using **consistent class labels** if you compare different analysis results OR
- **Uncheck** the default option in PLS-DA in MetaboAnalyst 5.0 (red arrow)



PLS-DA (variance vs. covariance)

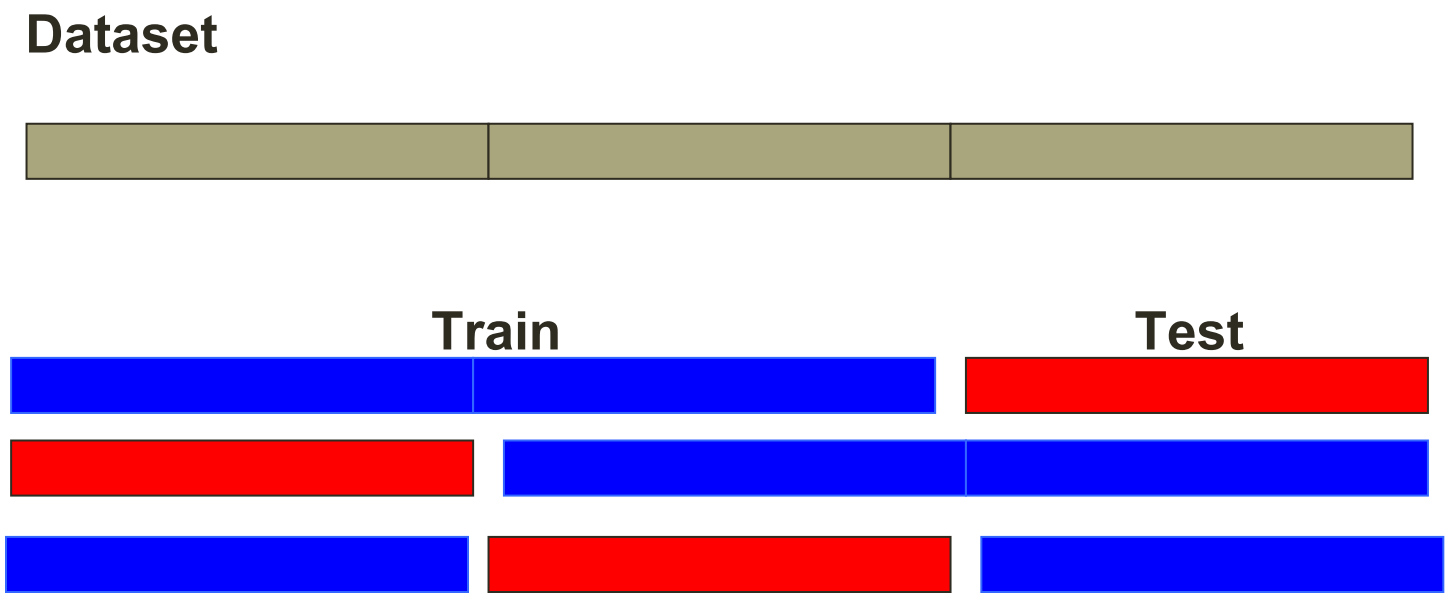
- PLS-DA maximizes the **covariance** between X (data) and Y (group). The variance displayed in the plot above is the **explained variance for X**.

- Covariance and x-variance may not agree with each other in some cases. For instance, the 1st component explains most co-variance, but it may not explain more X-variance than the 2nd component sometimes.



Cross validations (CV)

- Goal: test whether your model can predict class labels for new samples



PLS-DA performance measures

PLS-DA is susceptible to over-fitting, and require more rigorous validation

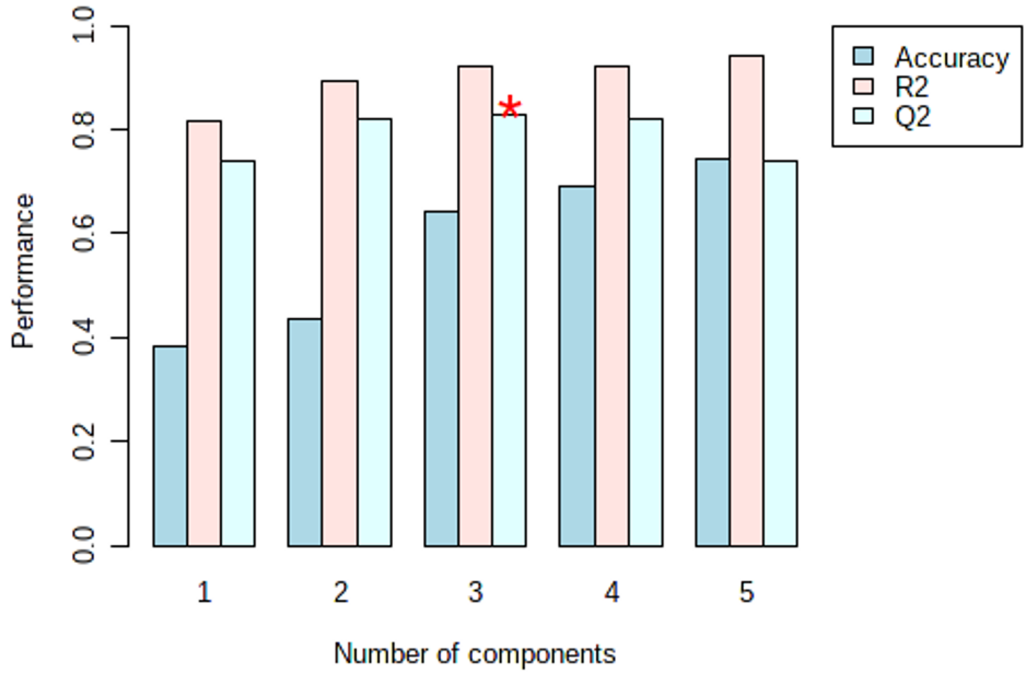
1. **Cross validation** – whether the model can predict on new events
   - Sum of squares captured by the model (R^2)
   - Cross-validated R^2 (also known as Q^2)
   - Prediction accuracy
2. **Permutation tests** – whether the model captures real signals compared to null

PLS-DA (R^2 & Q^2)

Q2 is calculated via cross-validation to compute Predicted Residual Sum of Squares (PRESS).

For convenience, the PRESS is divided by the initial sum of squares and subtracted from 1 to resemble the scale of the R2.

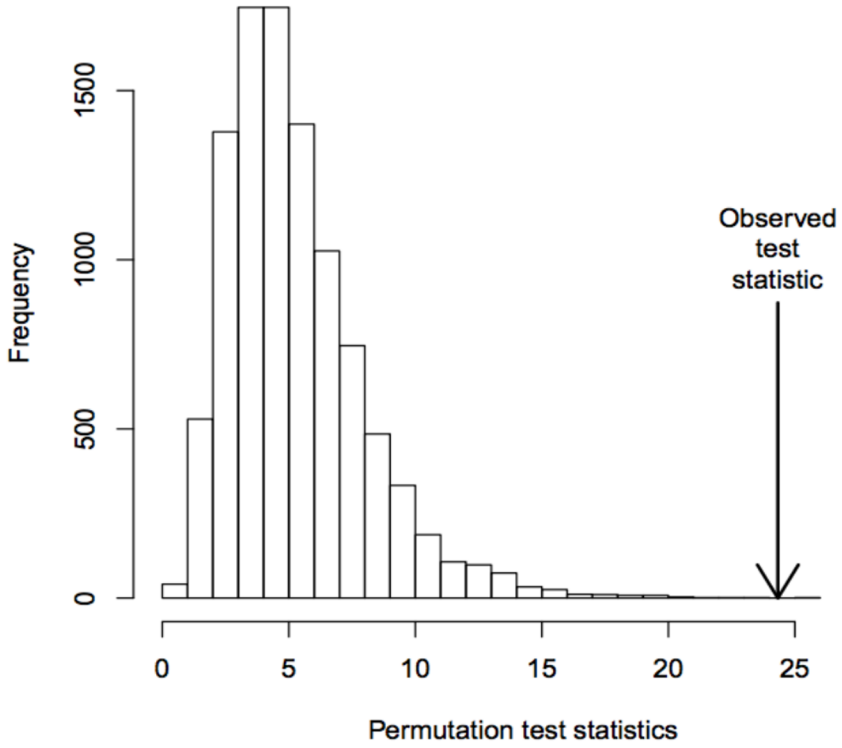
Good predictions will have low PRESS or high Q2. Low or even **negative Q2** means that your model is not at all predictive or is overfitted.



Permutation Tests

To test whether your model is significantly different from the null models

1. Randomly shuffle the class labels (y) and build the (null) model between new y and x;
2. Test whether there is still the similar patterns of separation;
3. We can compute empirical p values
   - If the result is similar as the permuted results (i.e. null model), then we can not say y and x is significantly correlated



Other Supervised Methods

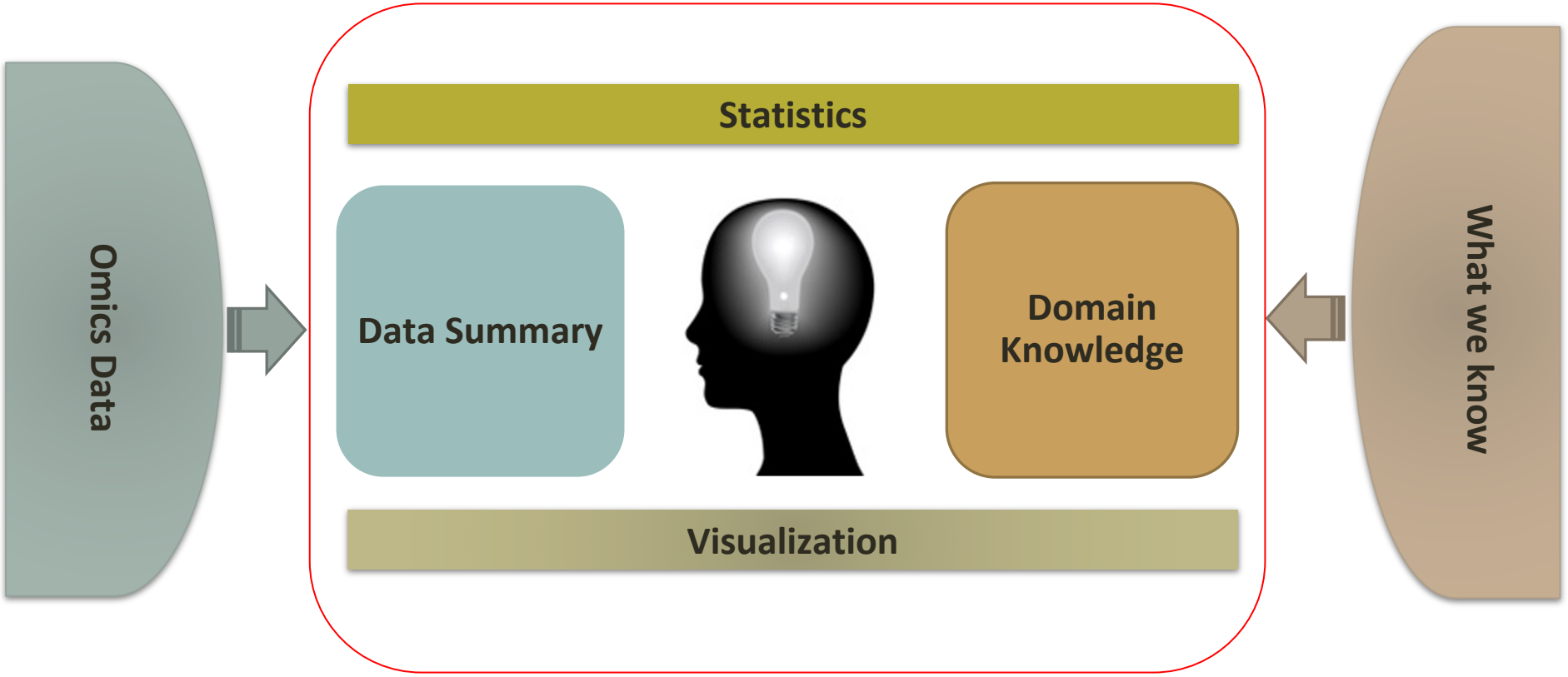
- Soft Independent Modeling of Class Analogy (SIMCA)
- Orthogonal Projection of Latent Structures (OPLS-DA)
- Support Vector Machines (SVM)
- Random Forest

Not everything was covered

1. Biomarker analysis
2. Spectral processing
3. Meta-analysis
4. Power-analysis
5. Integration with other omics
6. Co-variate analysis

…….

Don’t forget biology & context



Logistics

Workshop Materials & Resource

Instructions & Materials:

https://github.com/xia-lab/Metabolomics_2022

We have set up multiple servers:

- https://www.metaboanalyst.ca
- https://dev.metaboanalyst.ca
- https://new.metaboanalyst.ca
- https://genap.metaboanalyst.ca

Cautions & Recommendations

Page display may be **slow** due to bandwidth limitation in this room

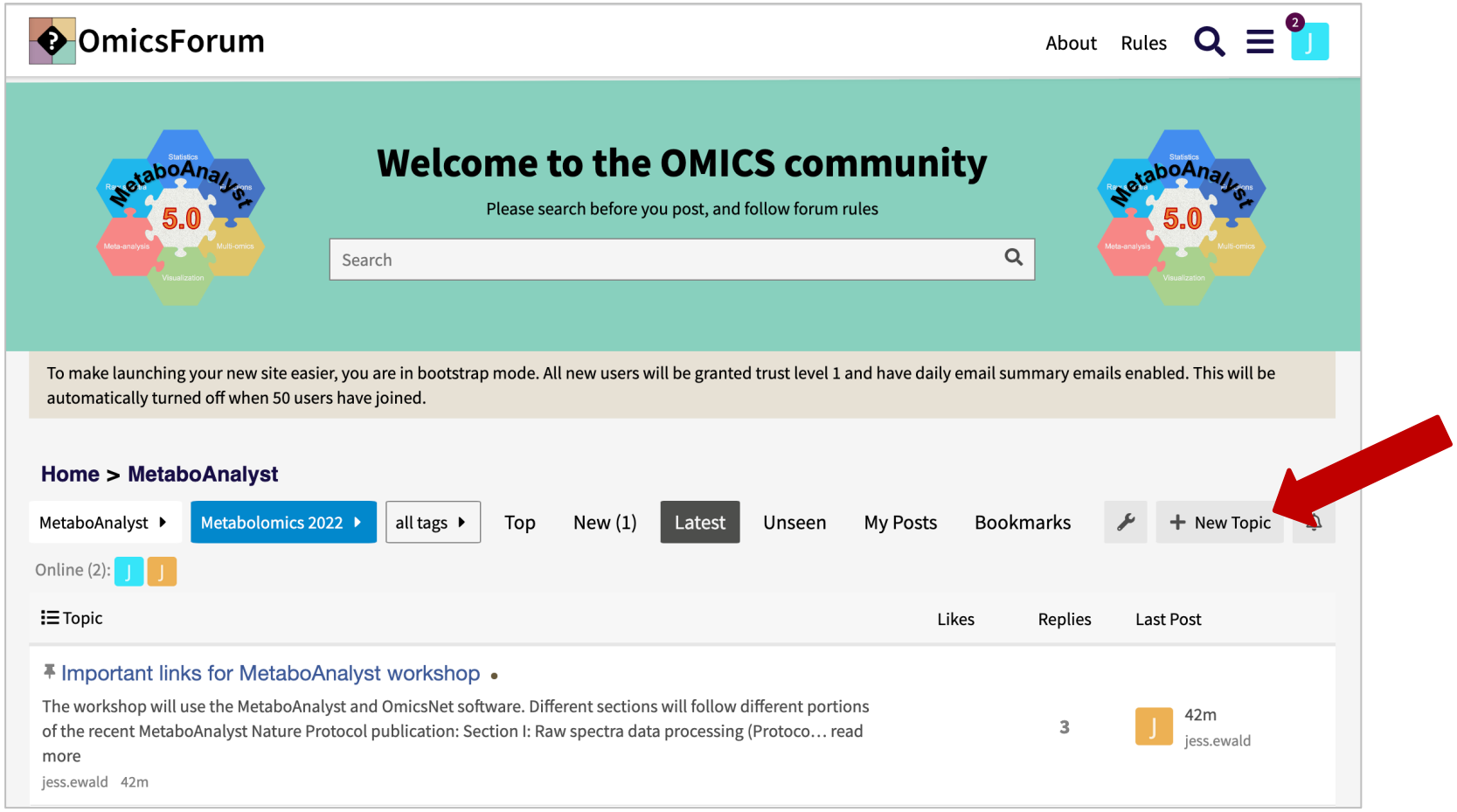
- Be patient
- Do not open multiple tabs to MetaboAnalyst (results will overwrite each other!)
- Form a group and share computers (see next page)

Group forming

- 2~3 people per group
  - Make new friends
  - Help each other and reduce stress

- Share laptops
  - One for displaying tutorial, and one for using MetaboAnalyst
  - Reduce bandwidth consumption

OmicsForum.ca



- Current topics in MetaboAnalyst: **116**
- Please search topics or open a new topic


GenomeCanada


GenomeQuébec


NIH
The Common Fund


NSERC
CRSNG


Canada Research Chairs
www.chairs-chaires.gc.ca


McGill
UNIVERSITY

Questions?