Processing LC-MS Metabolomics Data with xcms

Pietro Franceschi

pietro.franceschi@fmach.it

What is xcms

Framework for processing and visualization of chromatographically separated and single-spectra mass spectral data.

xcms

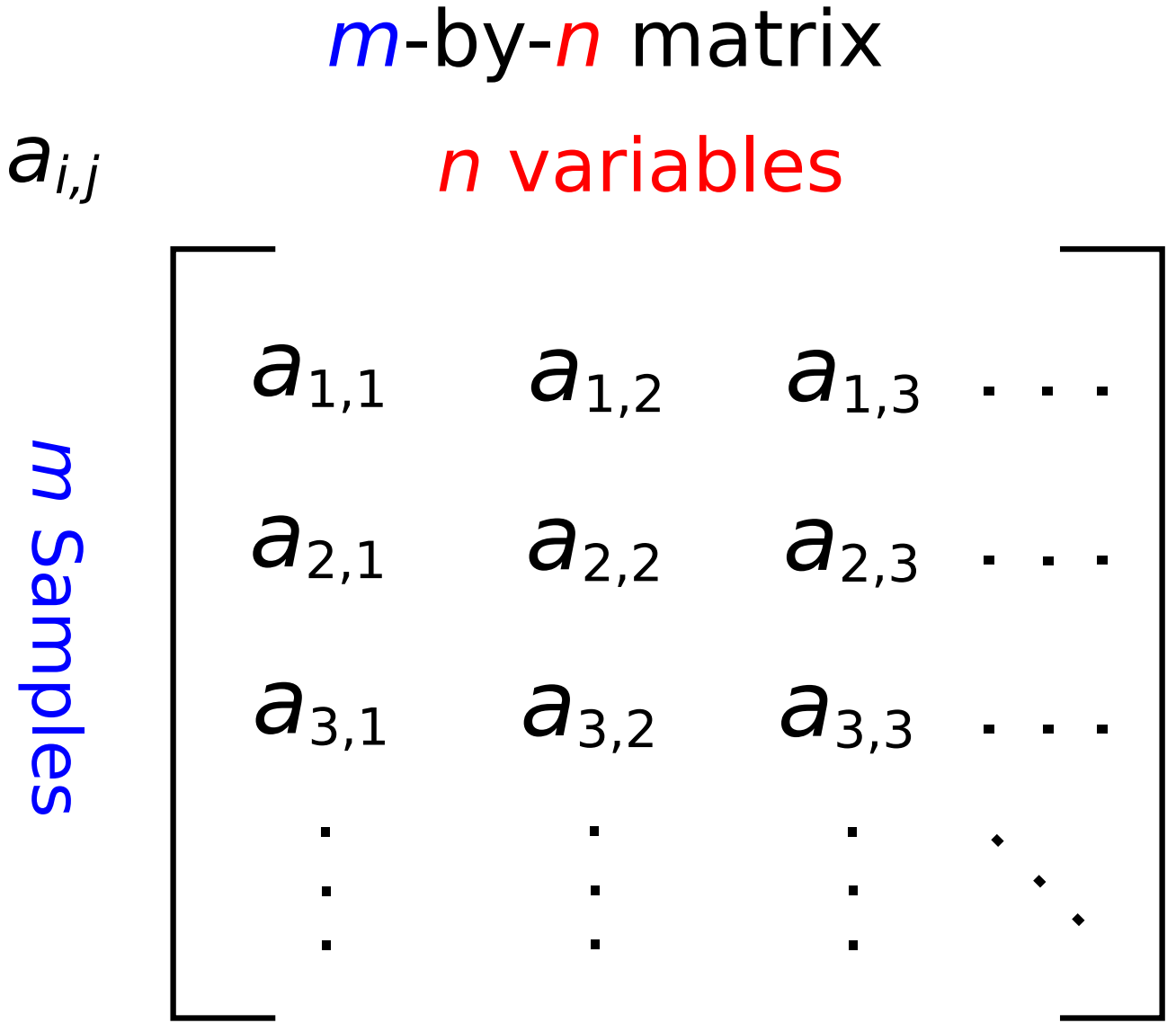
and still growing …

RforMassSpectrometry

Outline

- Data analysis, organization and data matrices
- Some thoughts on validation
- Preprocessing and analytical variability
- MS for Dummies
- LC-MS data handling
- Demo & DIY
- Peak Picking in `xcms`
- Demo & DIY
- Retention time correction and features definition
- Demo & DIY
- Dealing with Fragmentation esperiments
- Demo & DIY

The data matrix

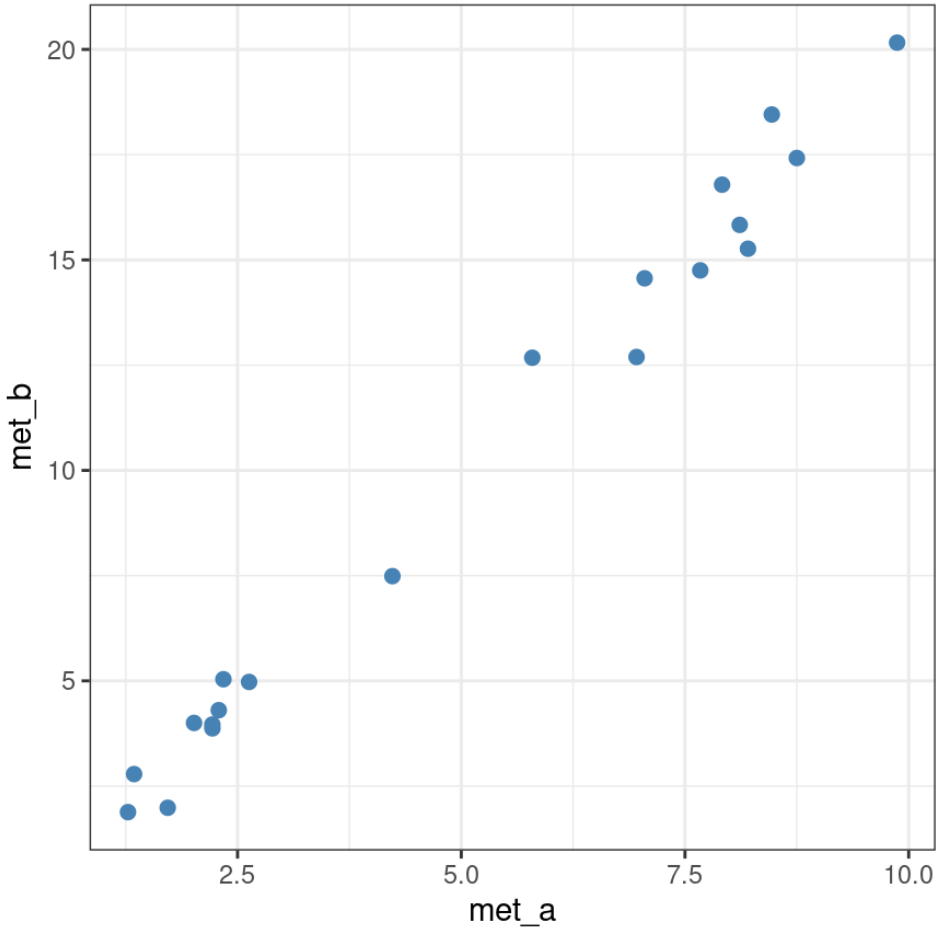


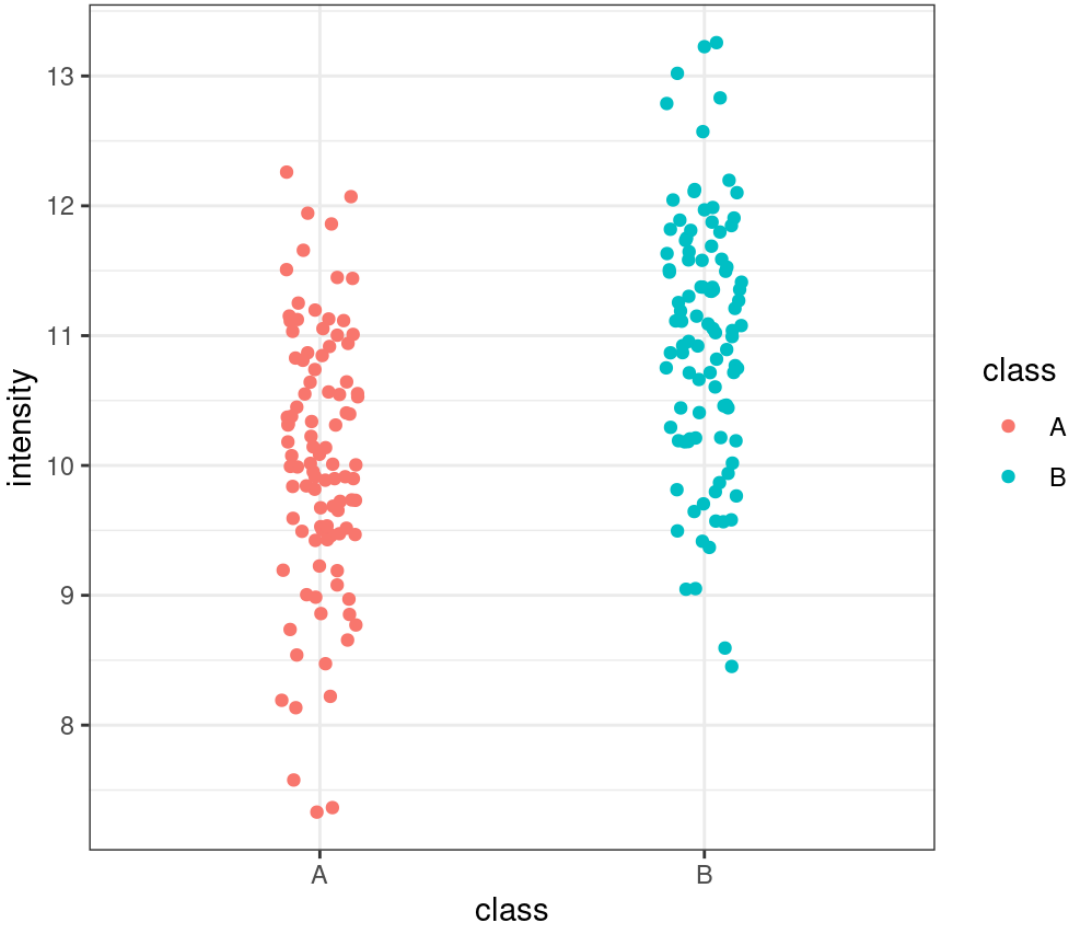
The role of Data Analysis

Statistics, Bionformatics, Machine Learning, Chemometrics, …, provide the tools to:

- make science shared and reproducible … ;-)
- process and organize **big data** into the matrix
- identify the presence of **organization** in the data matrix
- assess the confidence that our result is true “at the population level”

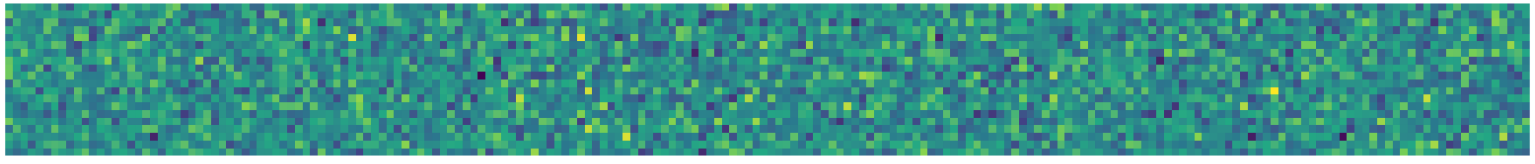
Examples of Organization





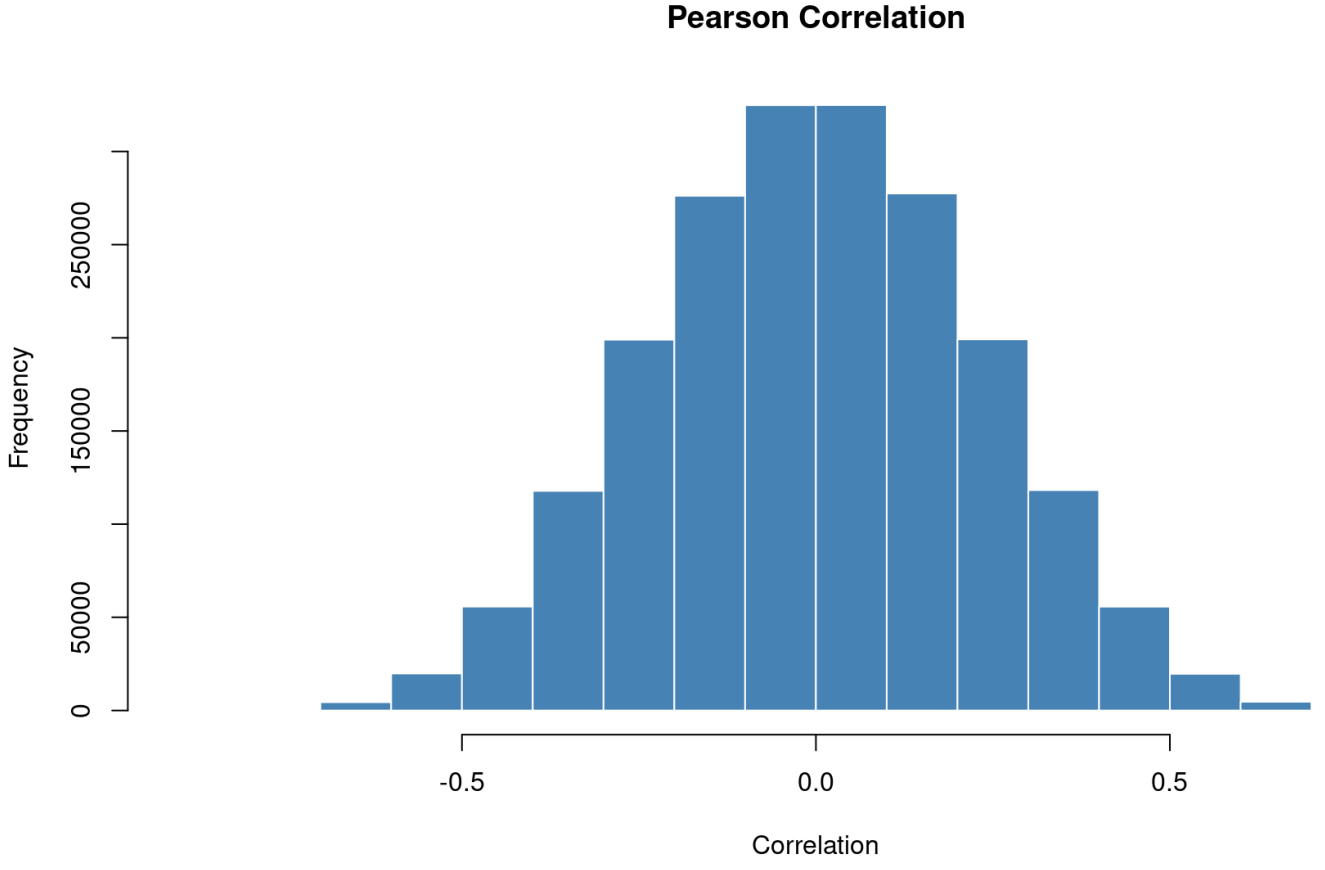
Fat Data Matrices

The typical metabolomics data matrix looks like this:



Let’s take 20 samples and 2000 variables … and fill the matrix of random numbers

Correlation coefficients



Range of correlations: -0.87, 0.86

We have relatively large values. Why?

False Positives

- Organization can show up only by chance
- These results are *true*, but the hold only for the data we are analyzing now
- Organization is not necessarily science
- Variability causes this
- We need to *validate* our outcomes

On Validation

- **Statistical Validation**: get brand new samples and see if what we get is still there
- **Domain Validation**: is what I'm getting in keeping with the domain specific body of knowledge? Could I design an experiment to check my hypothesis?

Do we always need statistics?

By the way …

- What are the variables measured in a *targeted metabolomics assay?*
- What are the variables measured in an *untargeted LC-MS metabolomics experiment?*
- What are the variables measured in an *untargeted NMR metabolomics experiment?*

Data hygiene

- Go for a scripting language and forget Excel
- … at least use a gui pipeline or a web based solution
- Organize data and metadata
- Avoid as much as possible *manual curation*
- Share your data, your scripts, your results
- Go open source

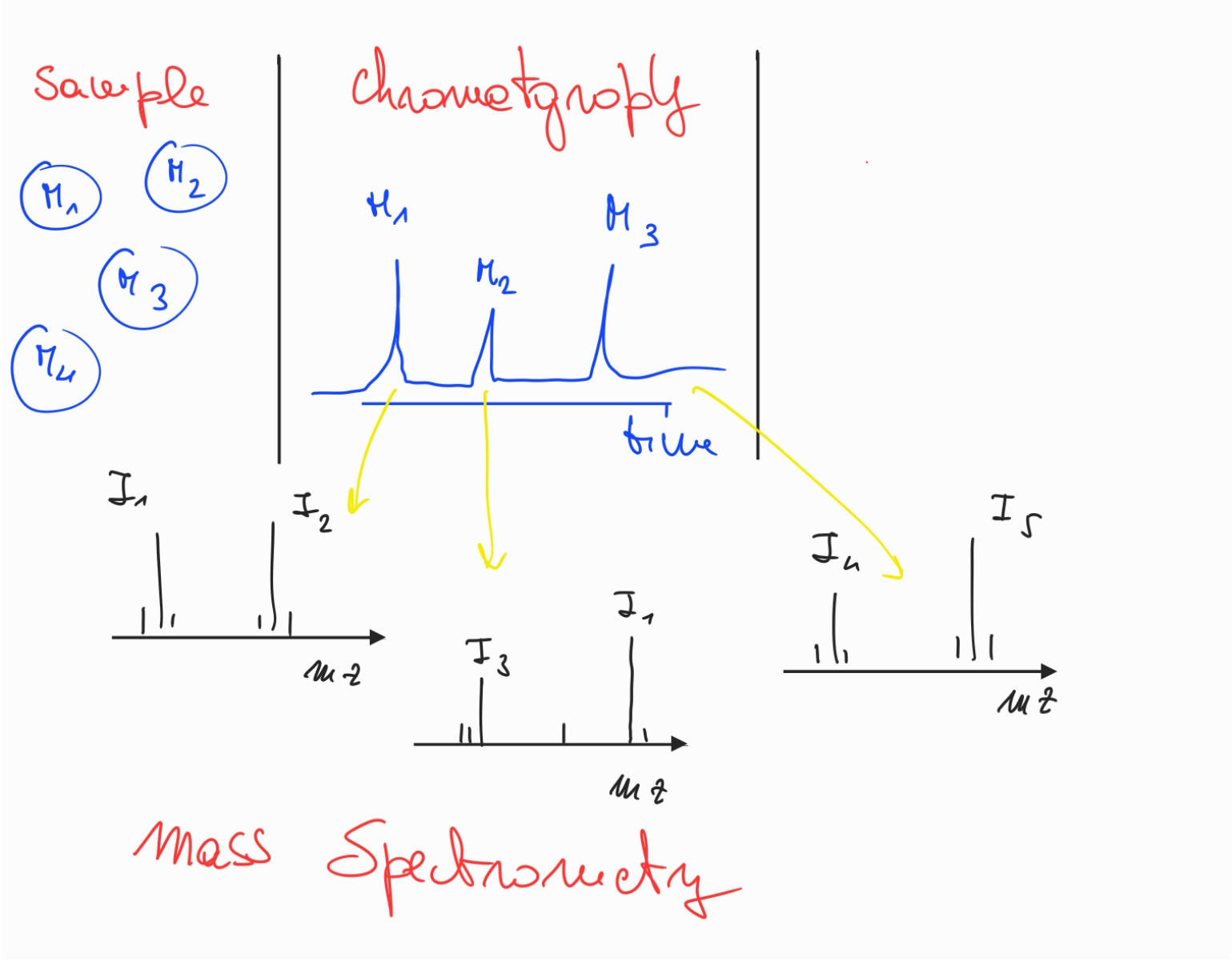
Get out your data

- Metabolomics data are always stored in “formats” which are specifically developed by instrument vendors
- In the case of MS data several open source standards are available (cdf, mzML, mzIML, …)

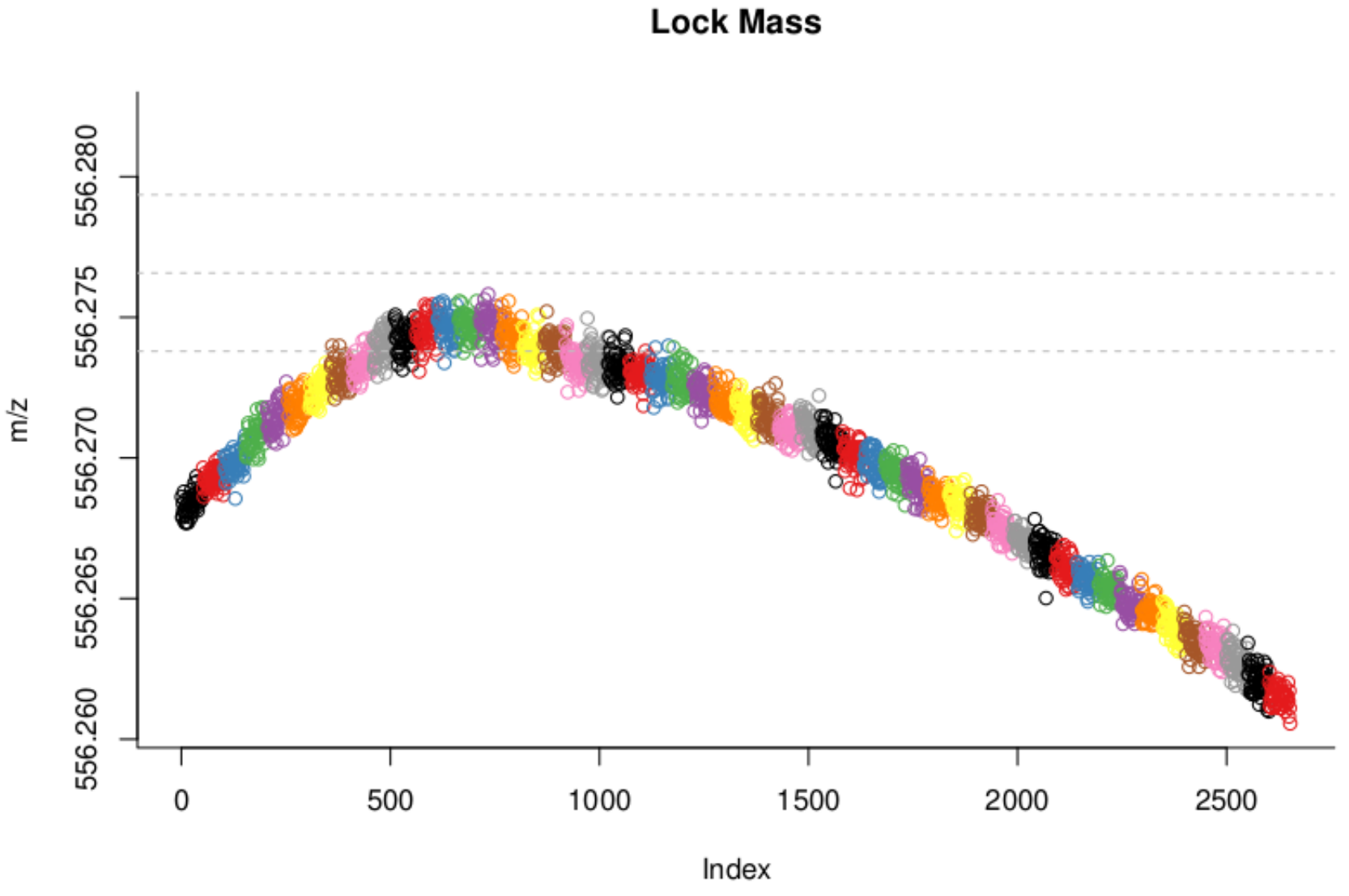
Proteowizard

- command line tool
- gui application
- docker with proprietary libraries

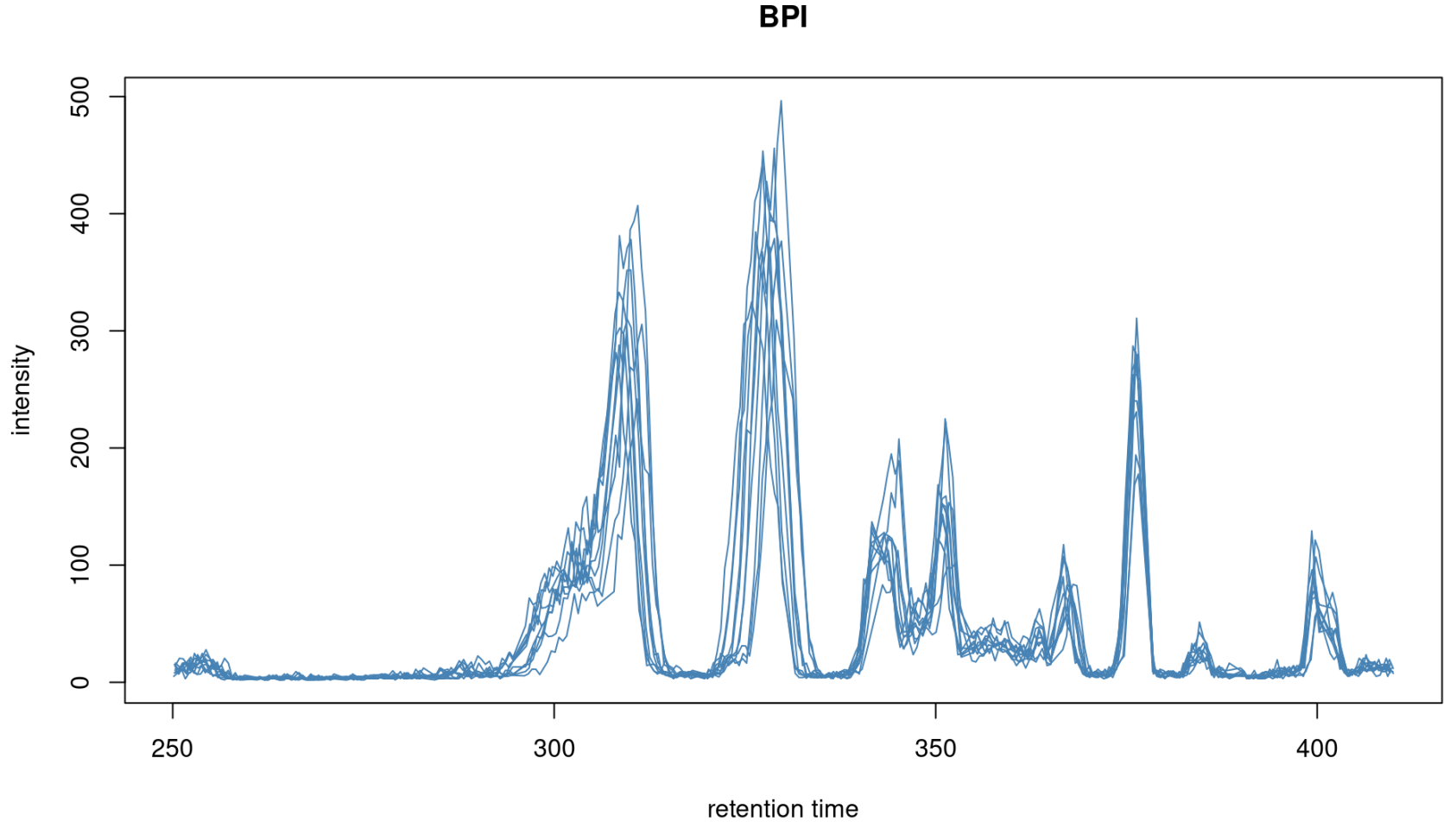
LC-MS For Dummies



Analytical Variability in LC-MS: mass



Analytical Variability in LC-MS: retention time



Analytical Variability in LC-MS: intensity

Preprocessing

- I call **preprocessing** all the data carpentry steps I do to go from the raw experimental data to the data matrix
- The aim of this process is to compensate for *analytical variability* being able to reliably build a data matrix
- *QC samples* play a big role on that because they are sensitive only to analytical variability

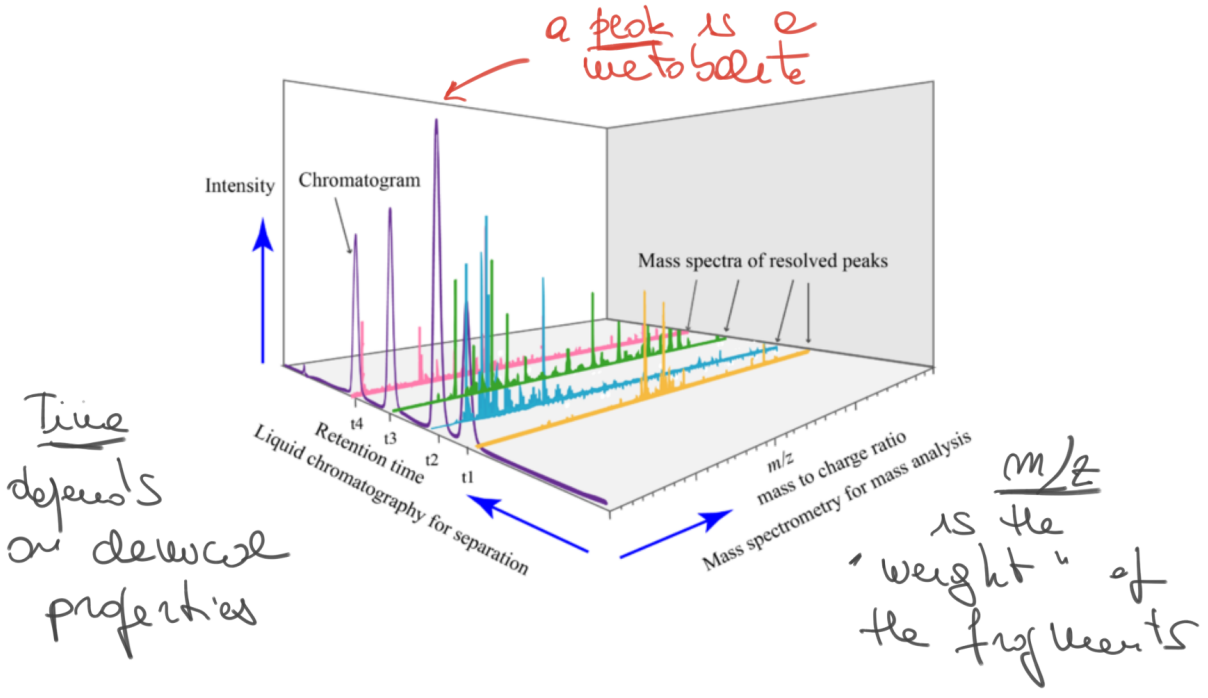
Uses of QCs

QCs should be representative of the chemical complexity of your samples

- correct for retention time shifts
- identify reliable variables:
  - variance in QC should be smaller than in samples
  - they should decrease during dilution
  - …
- help in correcting for bath effects …

http://www.freepik.com, Designed by

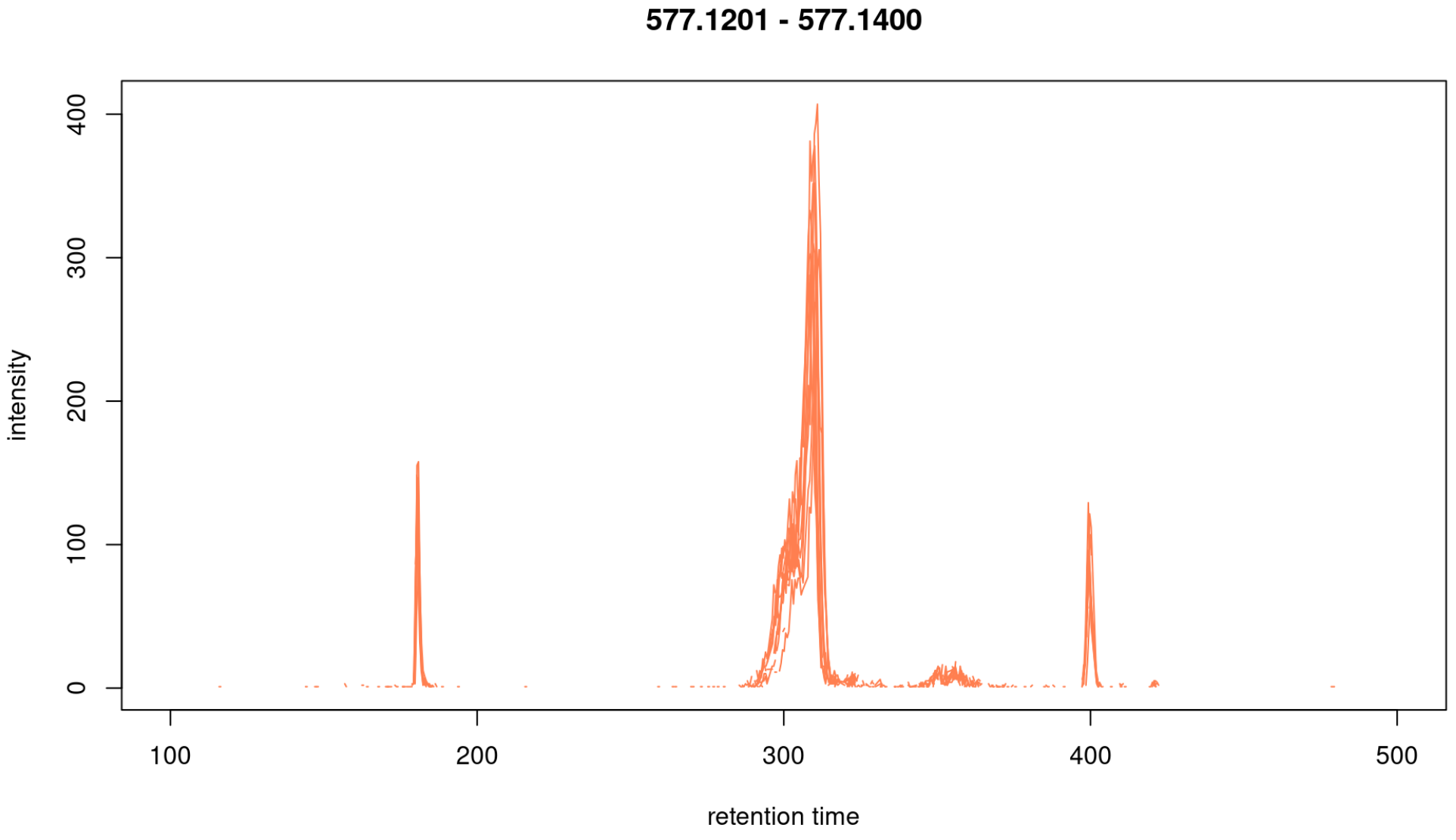
LC-MS produces 3D data (rt,mz,I)



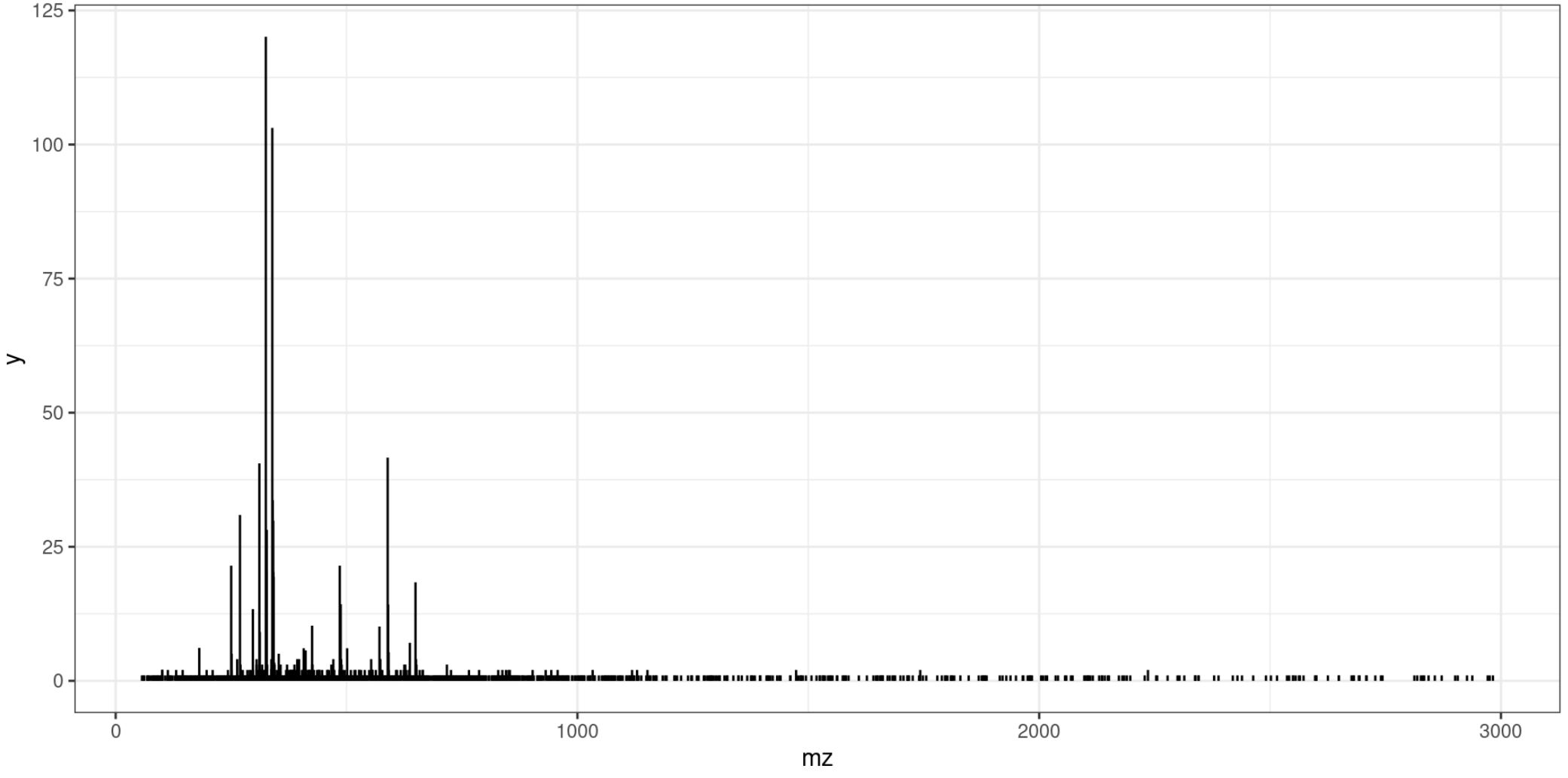
Things too look at

- Extracted Ion Trace/Current (EIT/EIC)
- Mass Spectra

Extracted ion traces



Mass Spectra



Back to Raw data

Always check your results on the raw data

- problems in preprocessing
- bad peaks
- biomarkers
- results hidden in noise

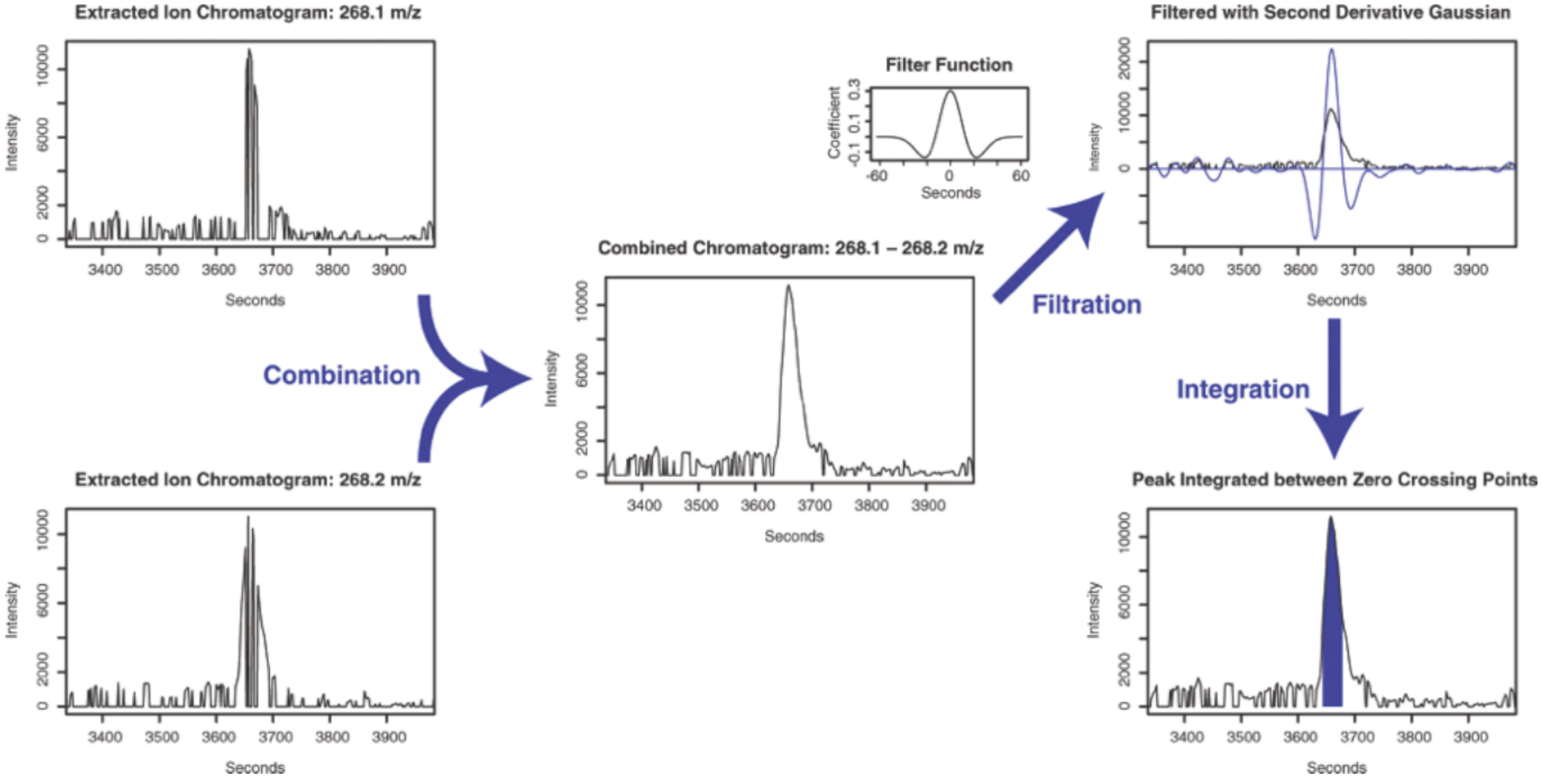

R
Studio®

Peak Picking

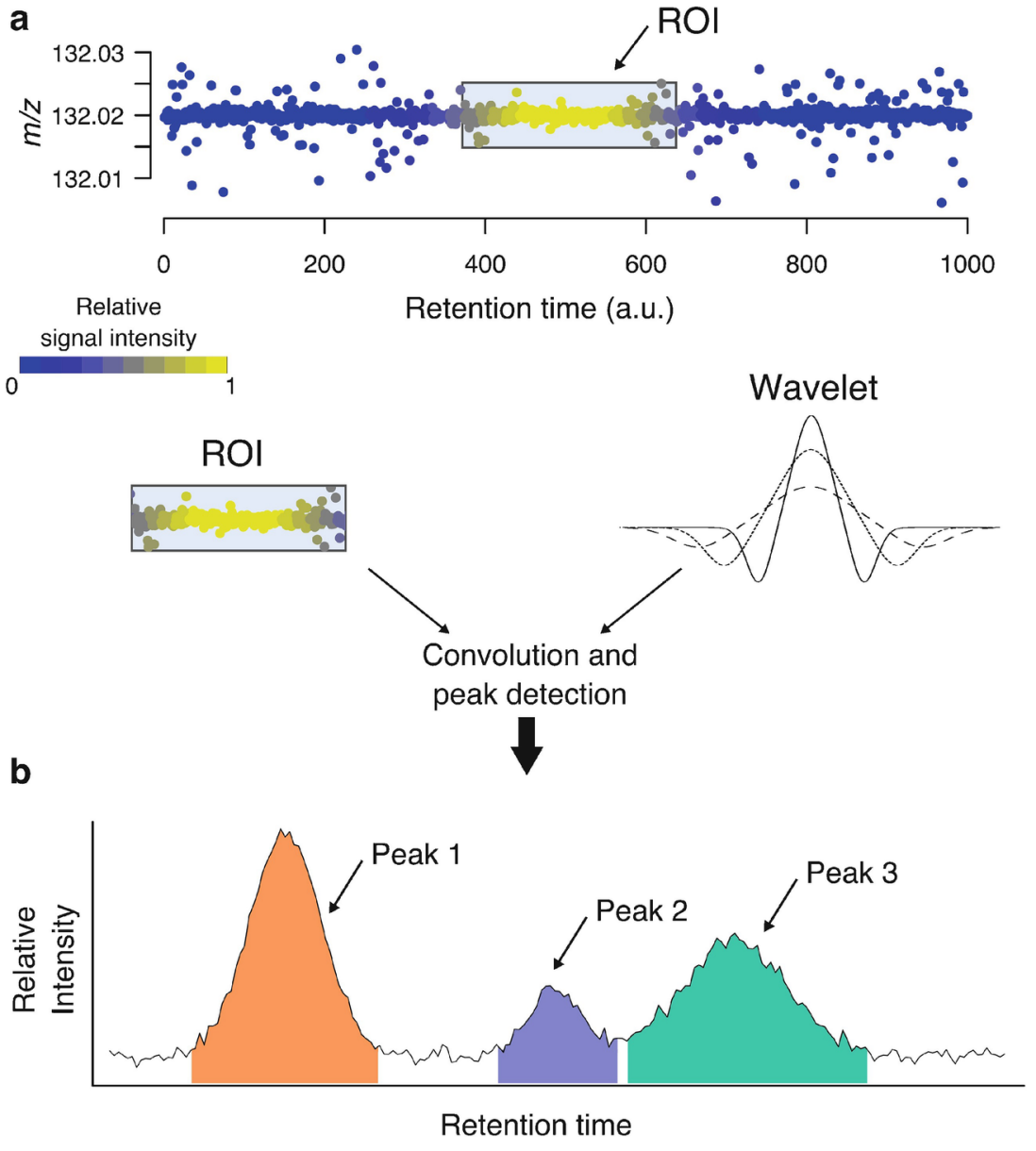
Peaks and metabolites: facts

- A metabolite produces peaks in the extracted ion traces of its associated ions
- Different peaks in the same ion chromatograms are associated to different metabolites
- Peaks are not metabolites
- The same peak can slightly move across the injections

MatchedFilter

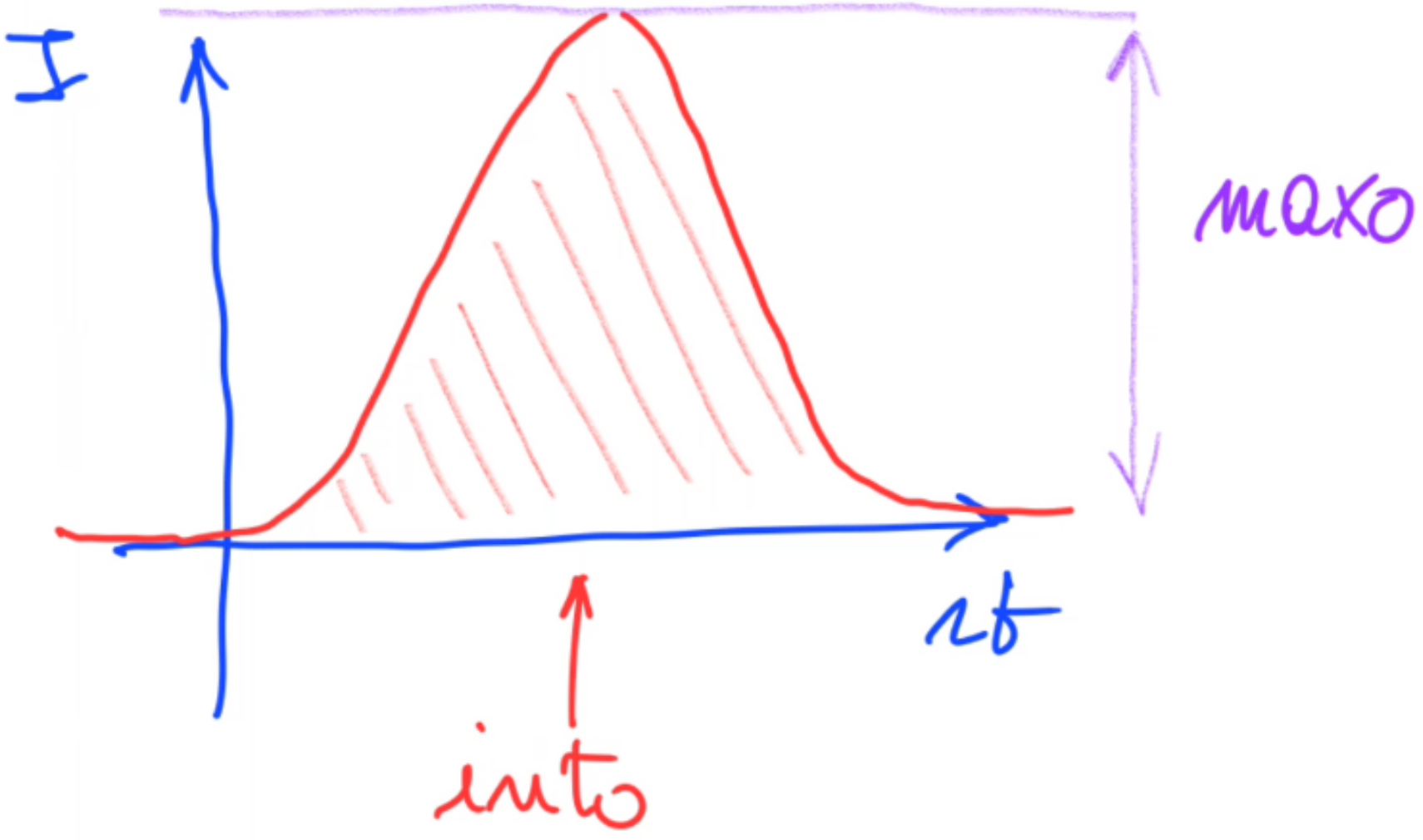


Cent Wave



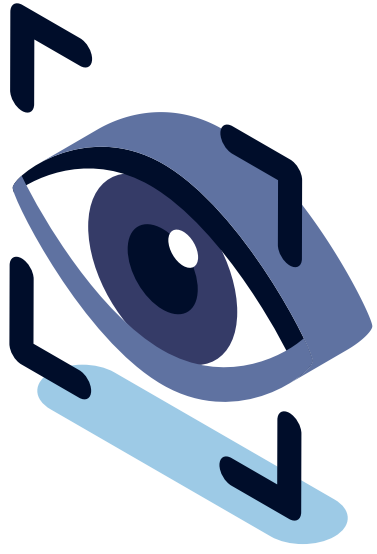
pietro.franceschi@fmach.it

Peak Intensity: `into` and `maxo`



Things to always consider

- Real peaks can be really badly shaped
- You are better than an algorithm … maybe AI will do well
- Every algorithm has parameters to tune!
- Look to the data!
- Know how the instrument works
- Check what happens to metabolites you know should be there



http://www.freepik.com" Designed by rawpixel.co


R
Studio®

Retention time correction and feature definition

… Just a recap

1. We converted the data files in an open source format (here mzML)
2. We optimized the peak picking parameters working on a representative sample (Qc)
3. We have been running peak picking on the full set of samples
4. We have been saving the output somewhere, just to avoid re-starting from scratch ;-)

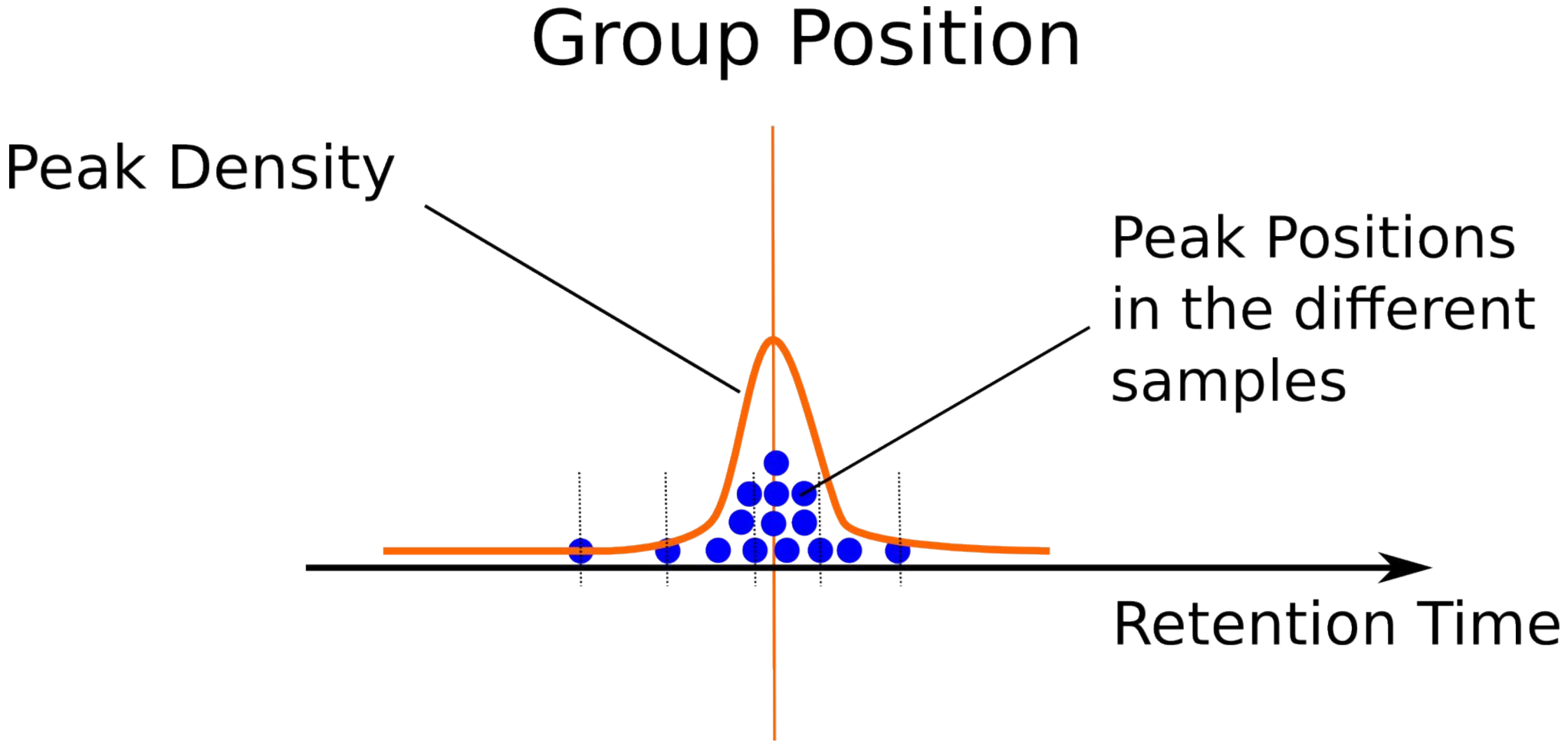
What Next

We have to merge the lists of chromatographic peaks into a consensus list of features peaks, which will be the columns of our data matrix

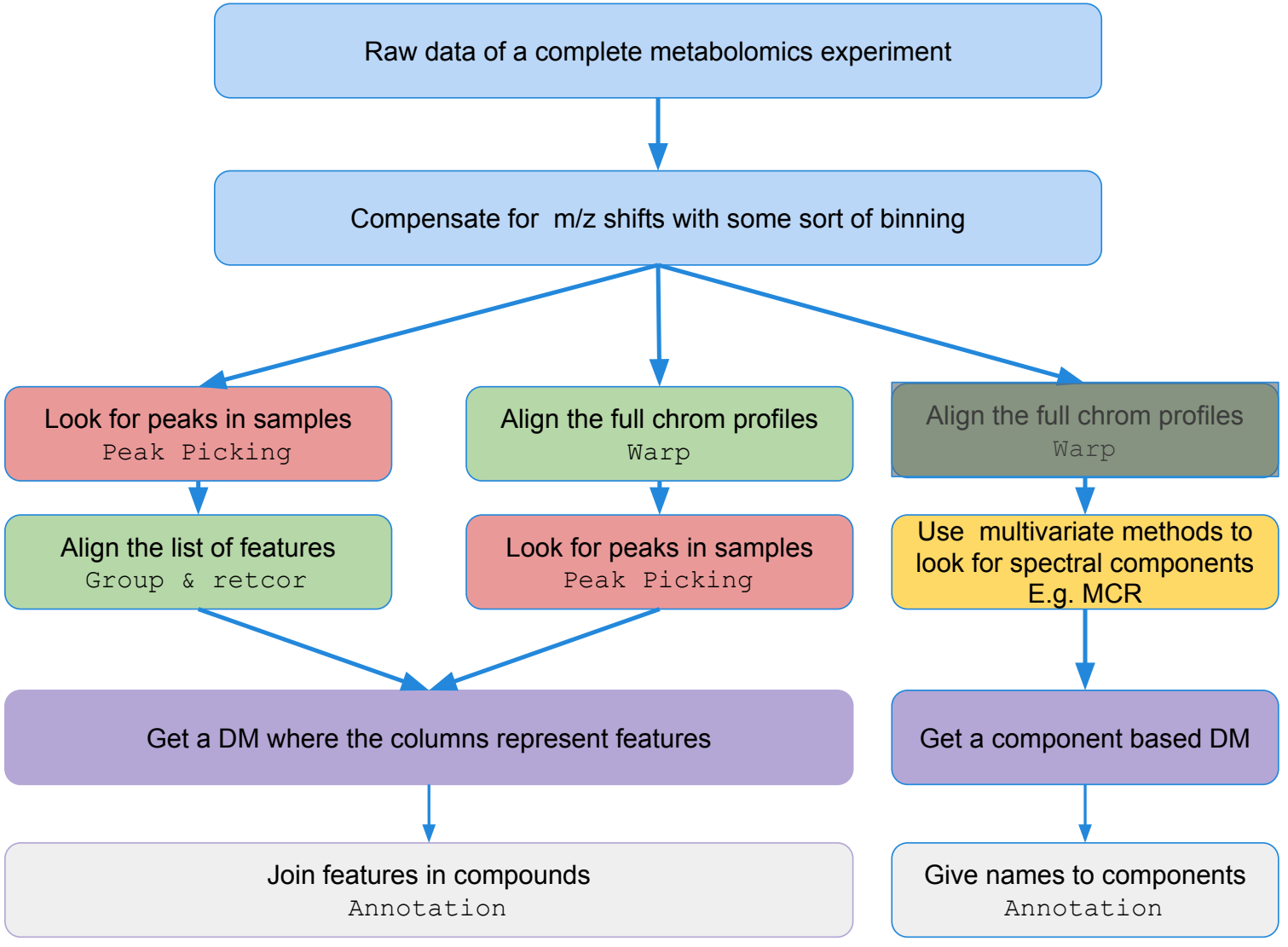
- *chromatographich peaks* what was detected in the individual samples (mz,rt,intensity)
- *features* consensus variables which are *grouping* several peaks coming from the different injections (mz,rt, intensity)

Grouping

For each m/z slice ...



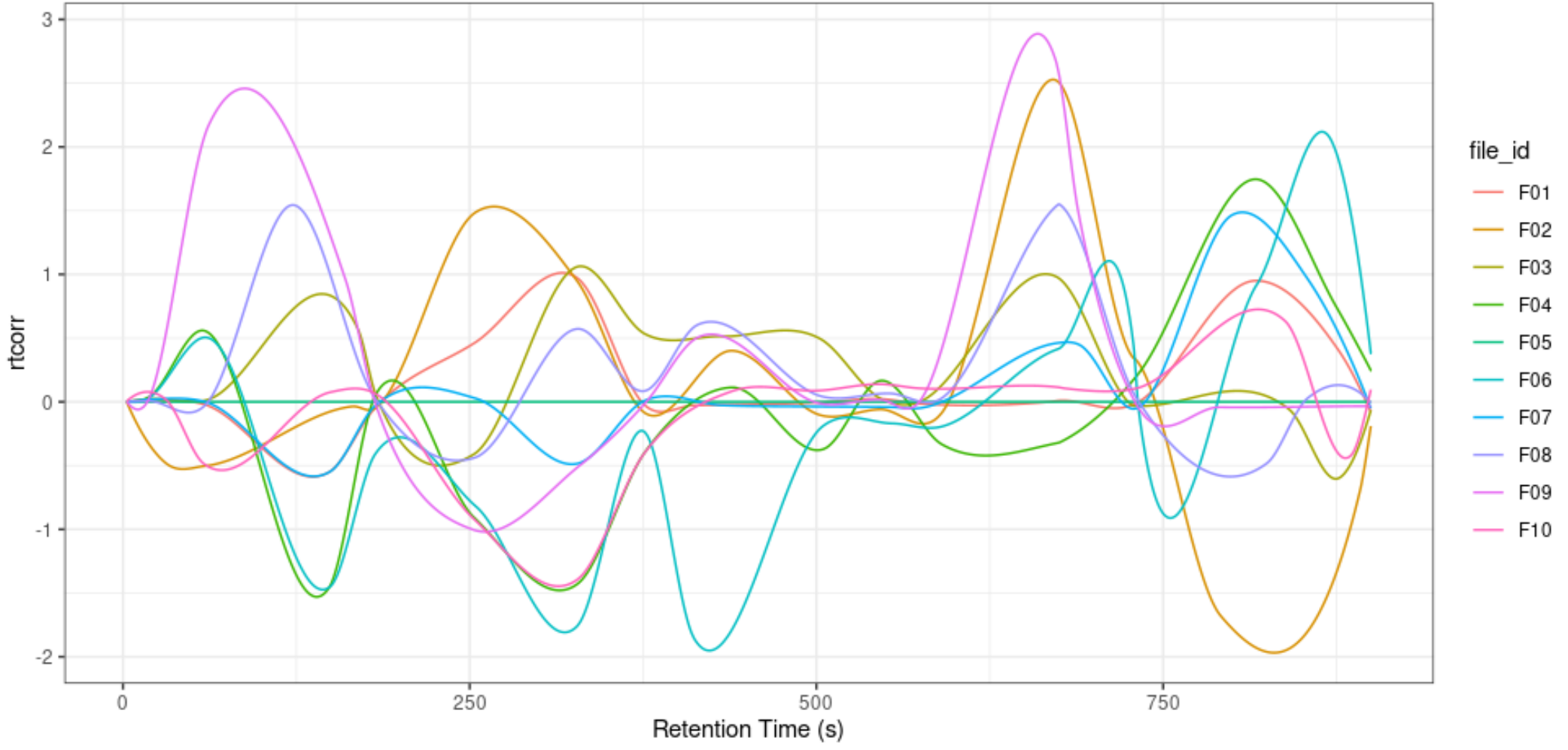
Retention time correction



Dynamic Time Warping

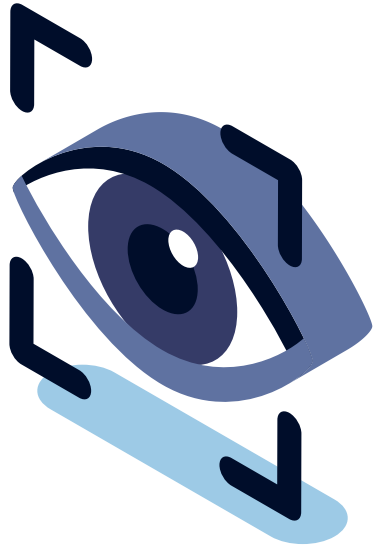
Available in `xcms` trough `obiwarp`

Mind the power of warping ...


rtcorr
3
2
1
0
-1
-2
0
250
500
750
Retention Time (s)
file_id
F01
F02
F03
F04
F05
F06
F07
F08
F09
F10

Things to always consider

- Aligning samples and not QCs can be tricky
- Some metabolites could not be present in pooled QC (dilutions)
- Sometimes chromatographic peaks are missed
- Always check the data and the known peaks!
- Parameters are easier to tune if you know how the analytics works



http://www.freepik.com" Designed by rawpixel.co

NAsss NAsss

Even if you do everything well your final data matrix will be full of missing values:

- errors in peak picking
- “absence” of a metabolite in one or more samples (biology)
- that metabolite is below the detection limit (analytics)

- missing at random
- missing not at random


R
Studio®

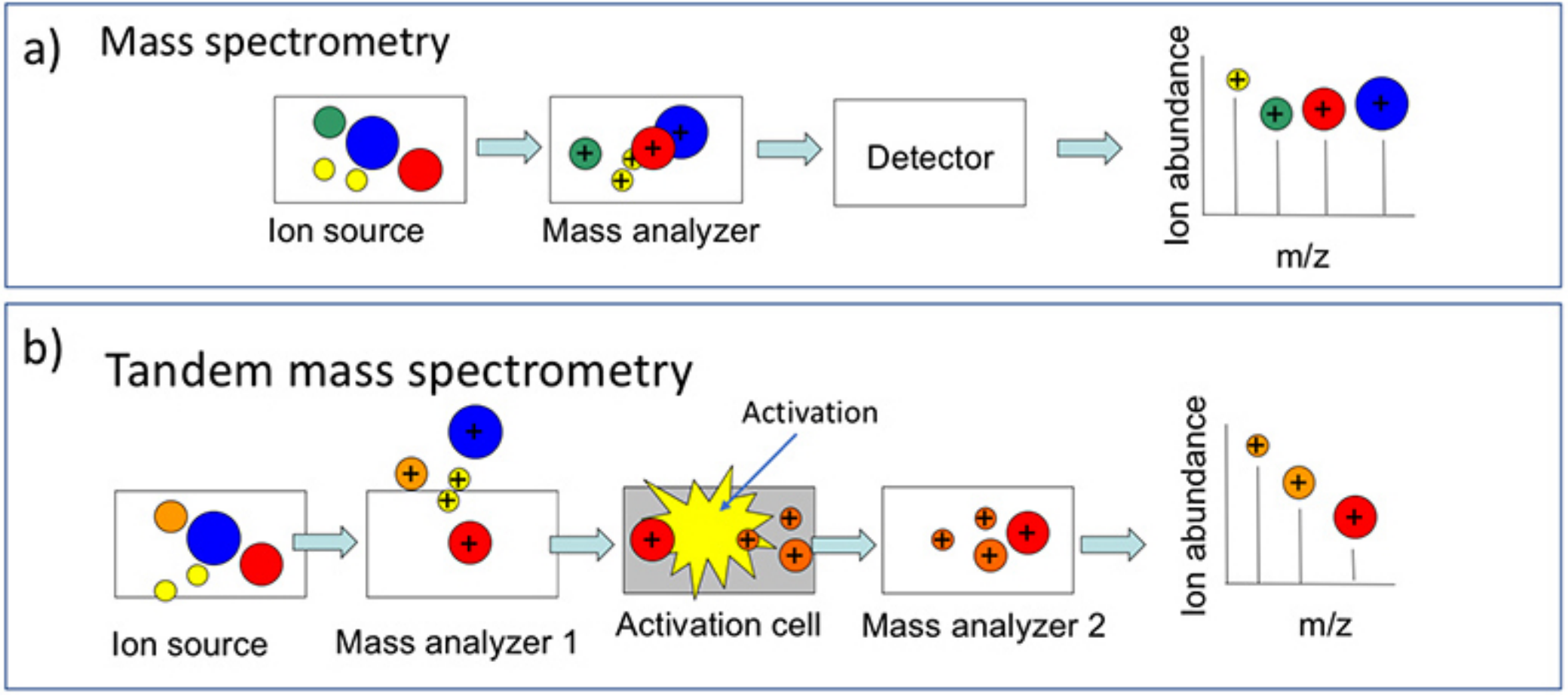
Bonus Section: fragmentation data

Annotation and MS

http://www.freepik.com, Designed by

MS/MS and DDA

DDA: data dependent acquisition




R
Studio®