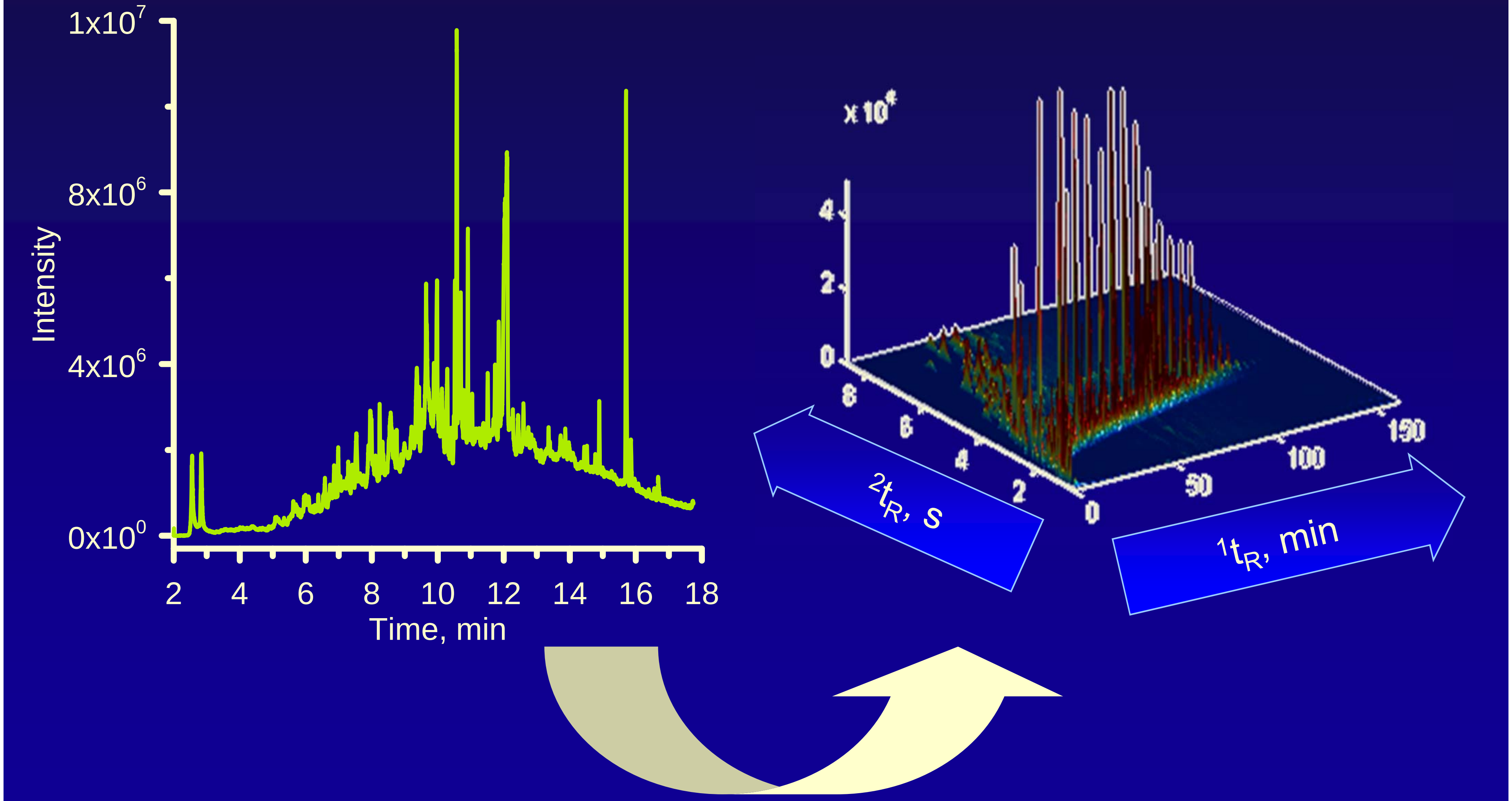
Data analysis in GCxGC

Gabriel Vivó Truyols
Analytical-chemistry group
Van ‘t Hoff Institute for Molecular Sciences
University of Amsterdam
g.vivotruyols@uva.nl

From 1D chromatography to 2D chromatography: what does change?

Introduction



The complexity of the data changes

The changes from 1D to 2D

Instruments can be classified according to the order of the tensor of data used to represent a single experiment:

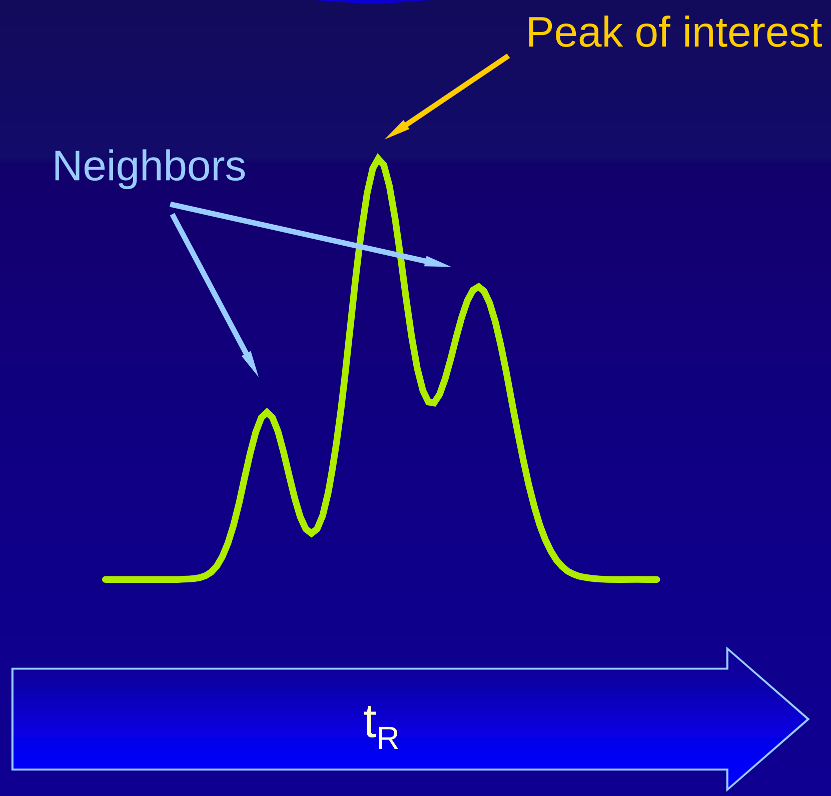
Zero-order instruments	Produce	Zero-order tensor (e.g. a number)	Example	pH - meter
First-order instruments	Produce	First-order tensor (e.g. a vector)	Example	UV-VIS spectrometer
Second-order instruments	Produce	Second-order tensor (e.g. a matrix)	Example	LC-MS, GCxGC
Third-order instrument	Produce	Third-order tensor (e.g. a “cube” of data)	Example	GCxGC-MS

nth-order instrument exist, but they are rare

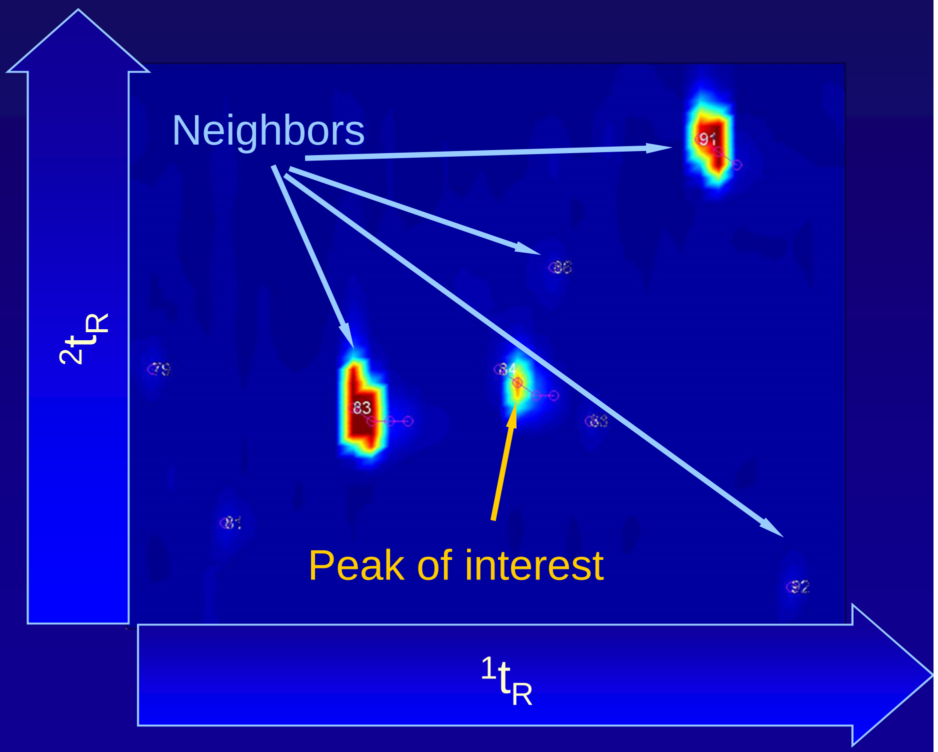
The concept of "peak vicinity" changes

The changes from 1D to 2D

One-dimensional



Two-dimensional



S. Peters, G. Vivó-Truyols, P. Marriott, P.J. Schoenmakers, J. Chromatogr. A. 1146 (2007), 232-241.

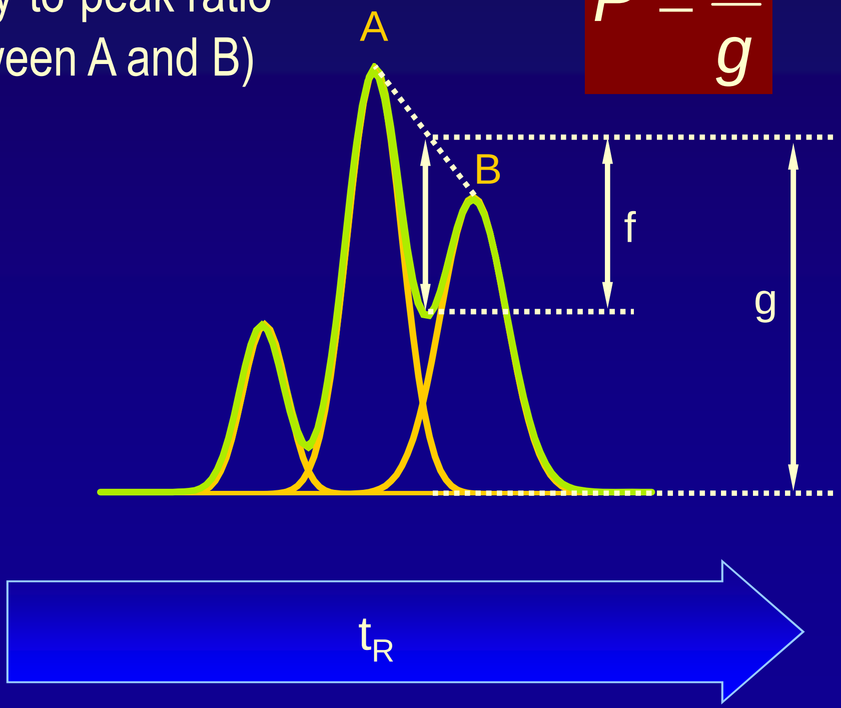
The concept of “peak resolution” is different

The changes from 1D to 2D

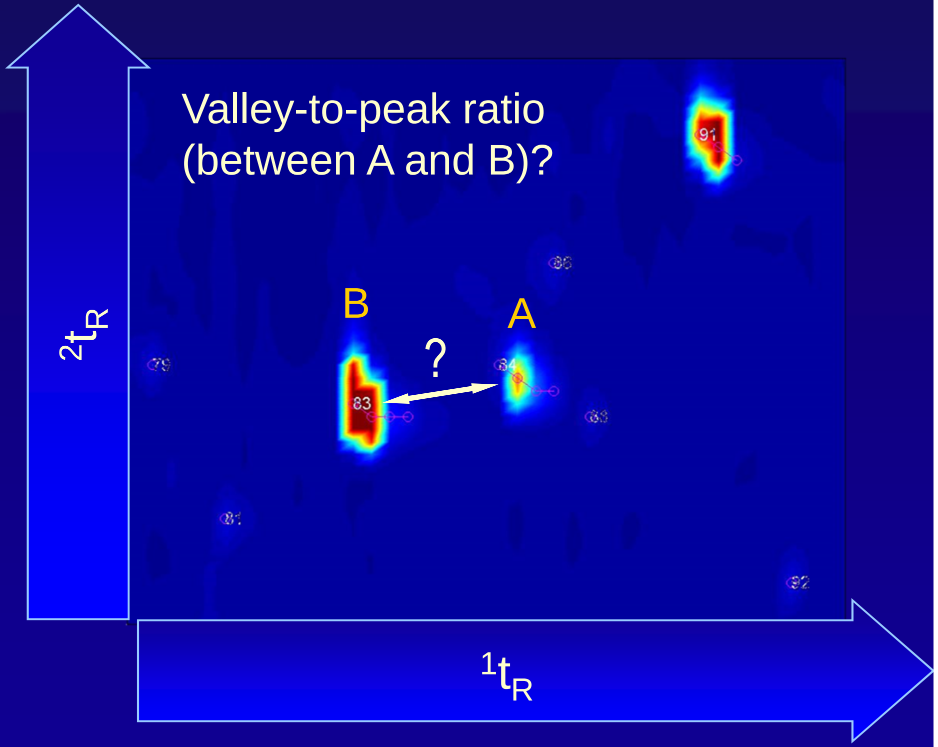
One-dimensional

Valley-to-peak ratio
(between A and B)

$$P = \frac{f}{g}$$



Two-dimensional

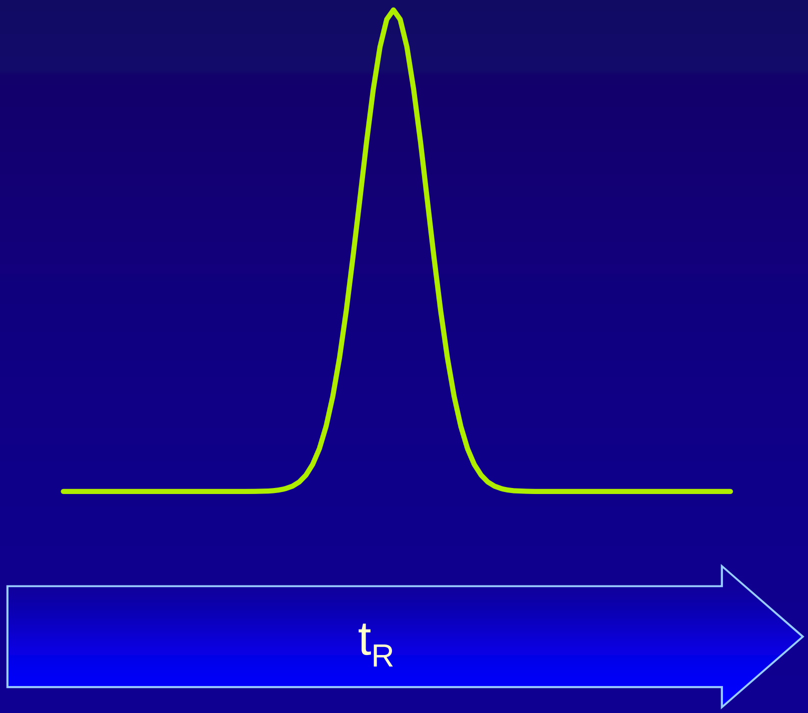

Valley-to-peak ratio
(between A and B)?

S. Peters, G. Vivó-Truyols, P. Marriott, P.J. Schoenmakers, J. Chromatogr. A. 1146 (2007), 232-241.

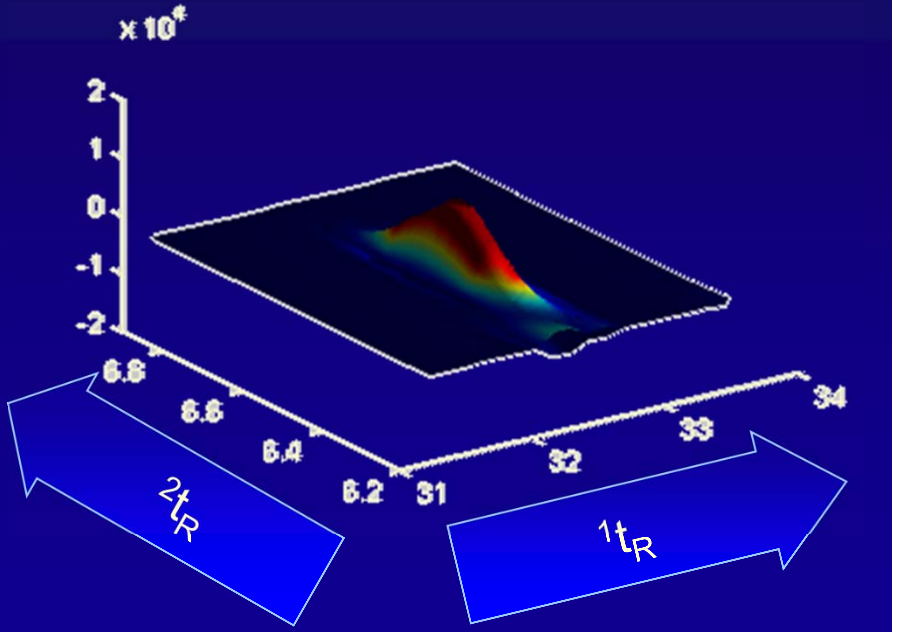
The concept of "peak" changes

The changes from 1D to 2D

One-dimensional

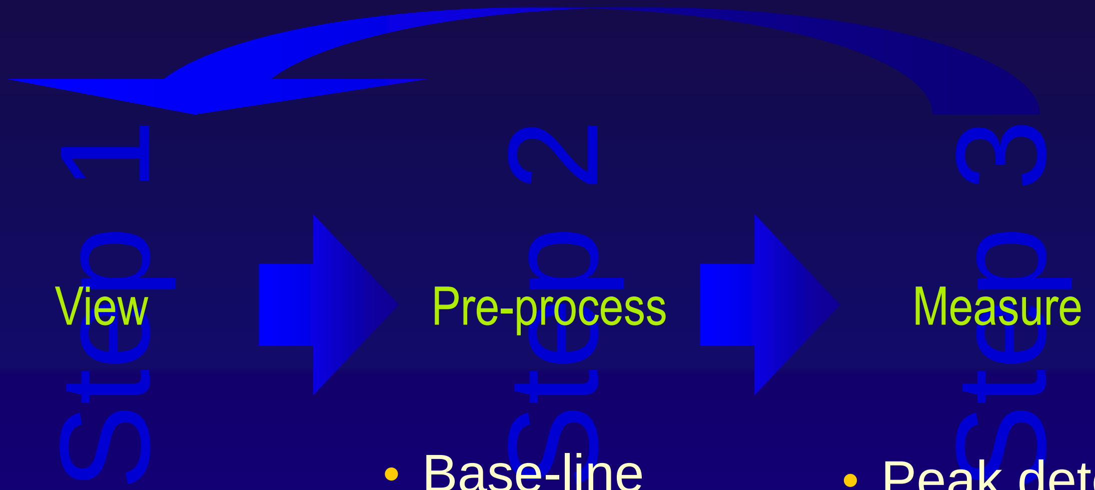


Two-dimensional



Do the main steps in data processing change?

The changes from 1D to 2D



Step 1 View

Step 2 Pre-process

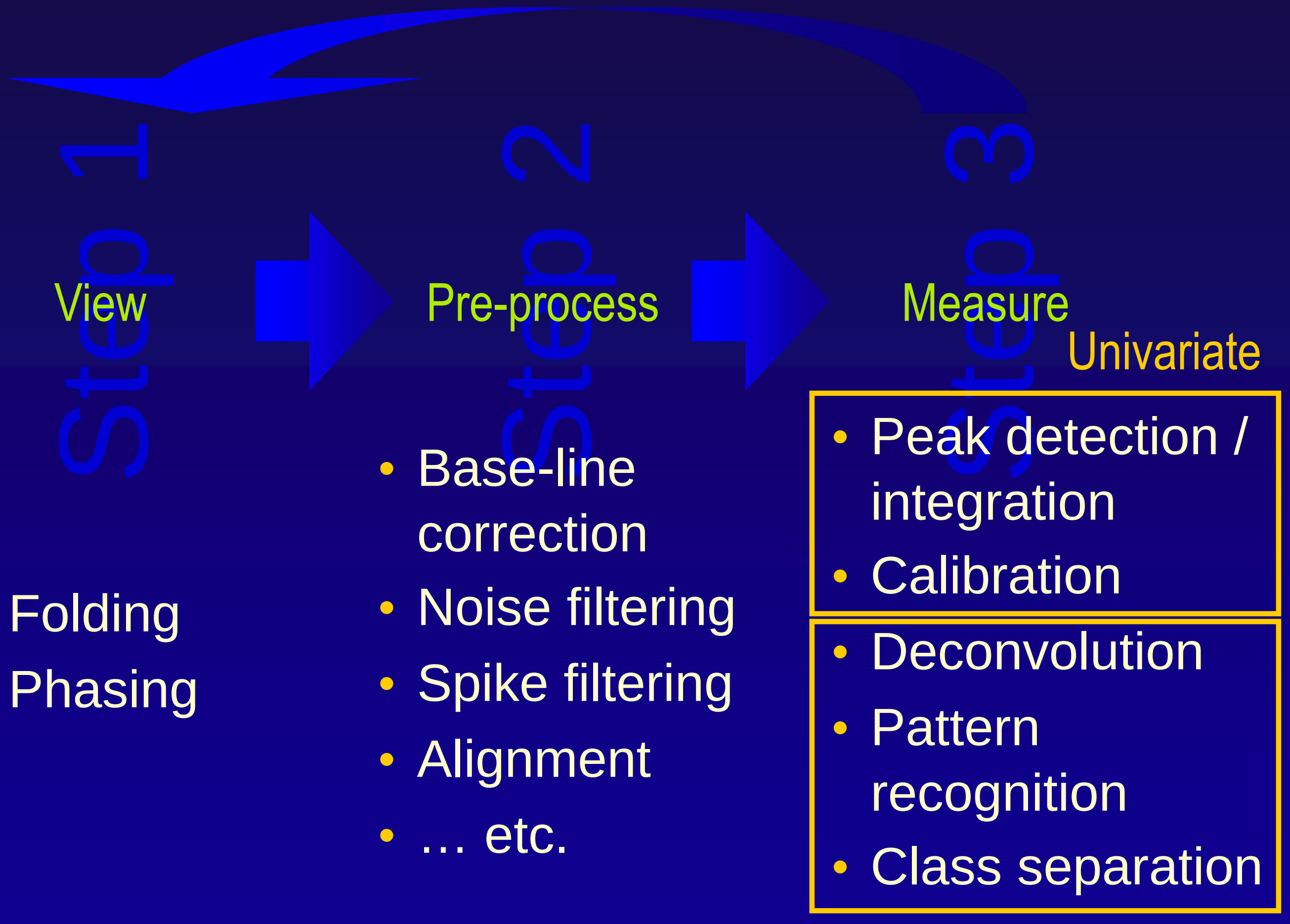
Step 3 Measure

- Base-line correction
- Noise filtering
- Spike filtering
- Alignment
- … etc.

- Peak detection / integration
- Calibration
- Deconvolution
- Pattern recognition

Do the main steps in data processing change?

The changes from 1D to 2D



View

- Folding
- Phasing

Pre-process

- Base-line correction
- Noise filtering
- Spike filtering
- Alignment
- … etc.

Measure

Univariate

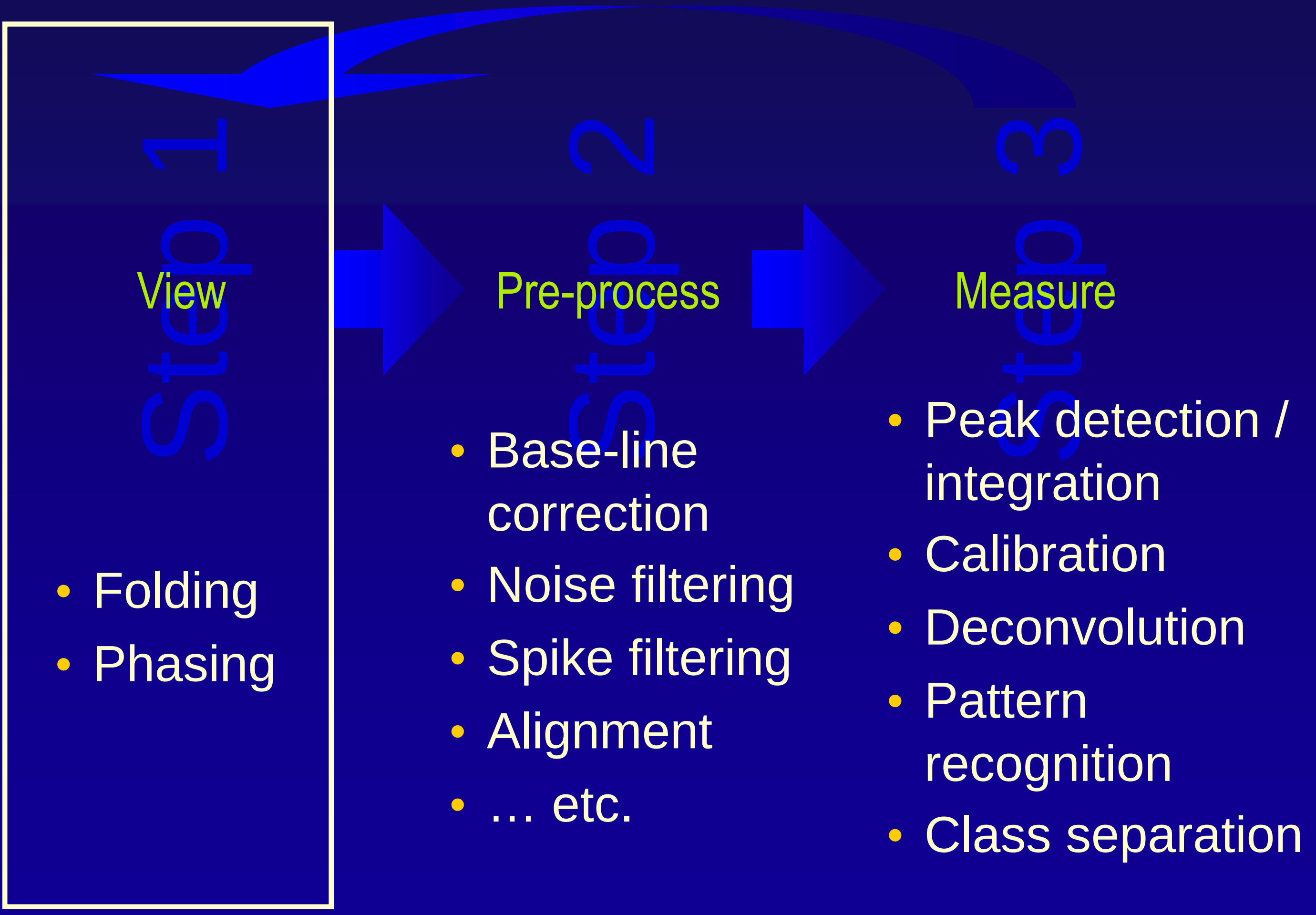
- Peak detection / integration
- Calibration

- Deconvolution
- Pattern recognition
- Class separation

Multivariate

Basically, the steps are the same, but the algorithms used for each step may be different.

First step: visualization



View

- Folding
- Phasing

Pre-process

- Base-line correction
- Noise filtering
- Spike filtering
- Alignment
- … etc.

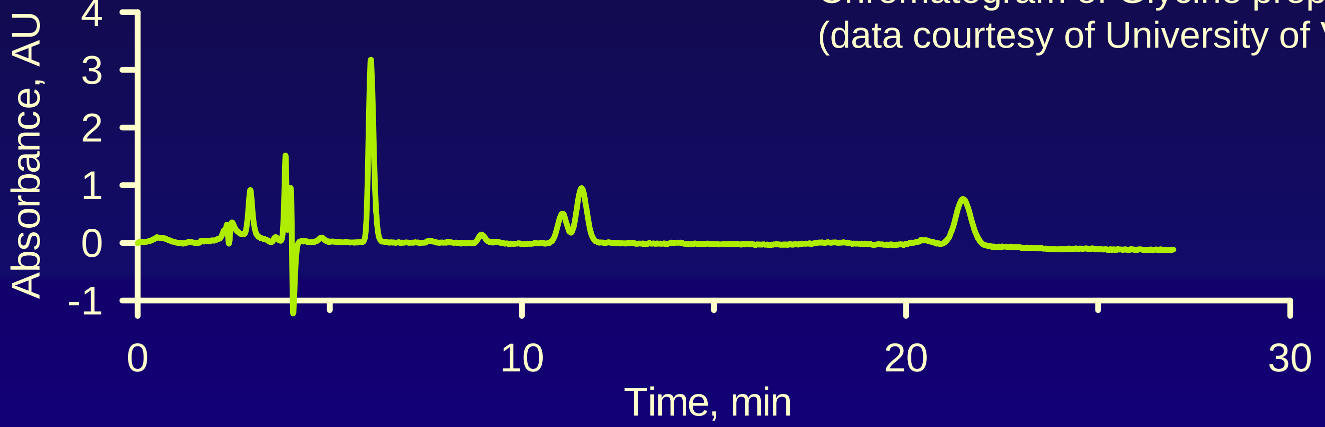
Measure

- Peak detection / integration
- Calibration
- Deconvolution
- Pattern recognition
- Class separation

Raw data in 2D chromatography

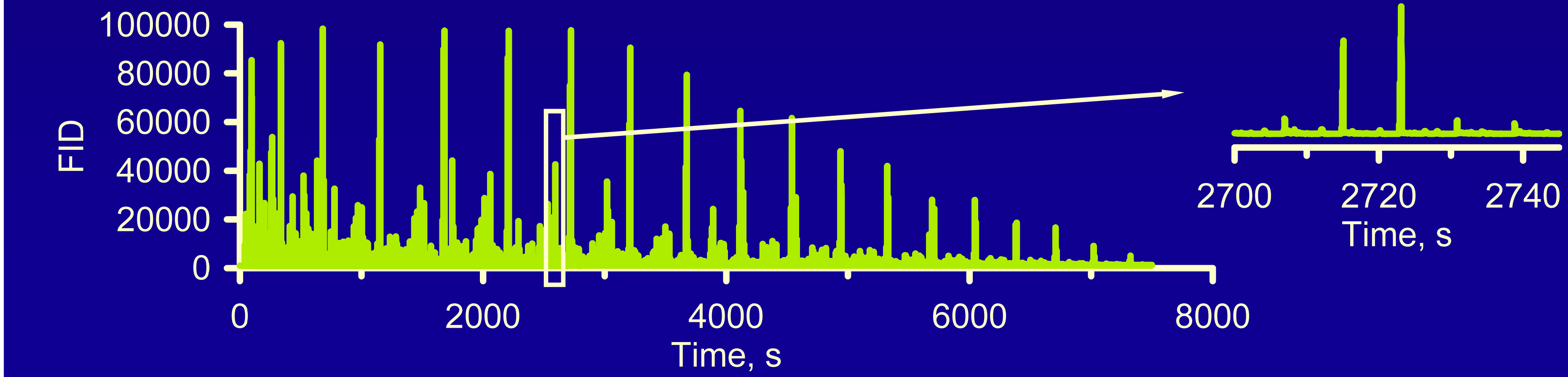
Visualisation

Chromatogram of Glycine preparation (254nm)
(data courtesy of University of Valencia)



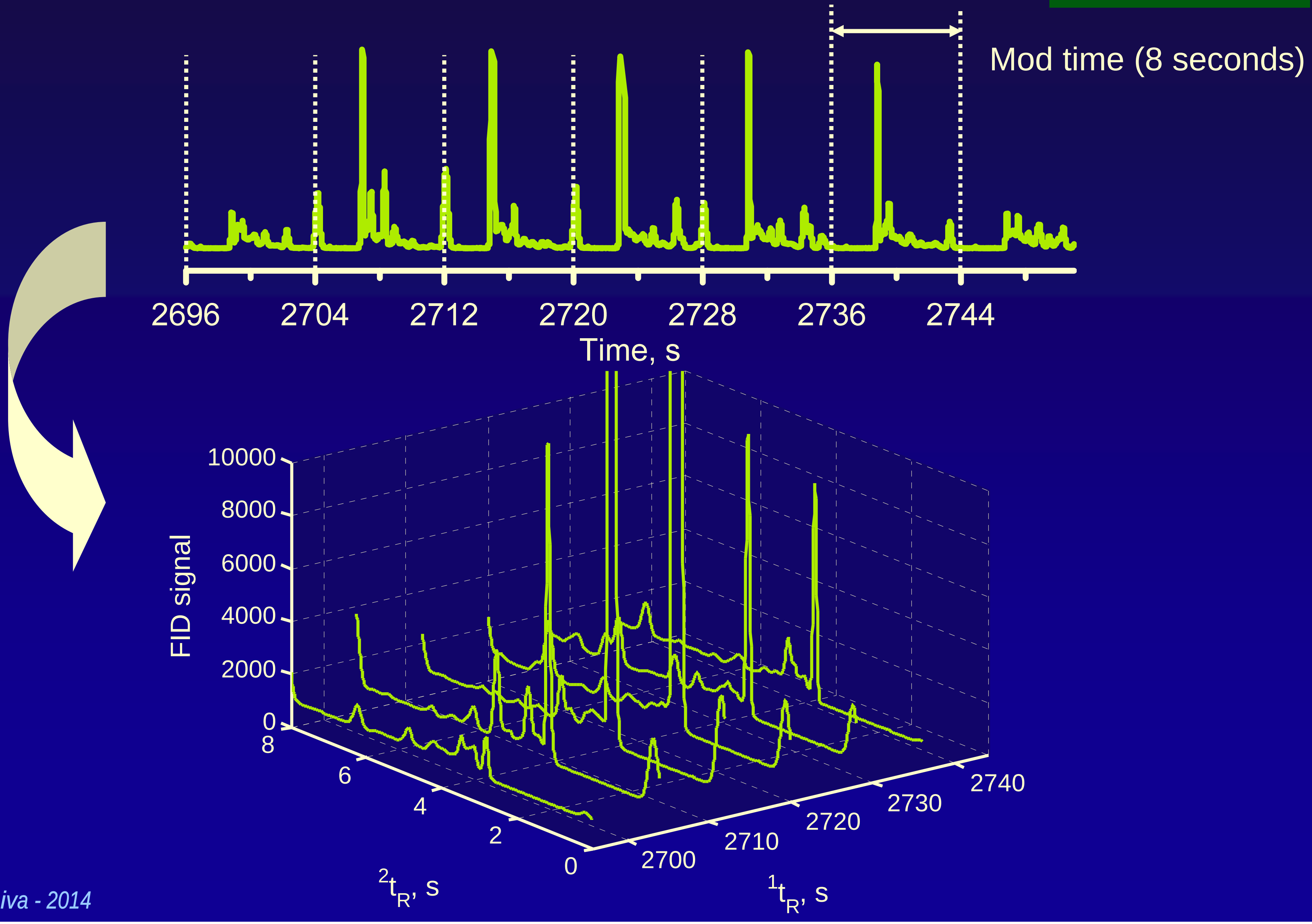
Two-dimensional chromatography

GCxGC chromatogram of diesel (FID detector)
First dimension: non-polar; Second dimension: polar



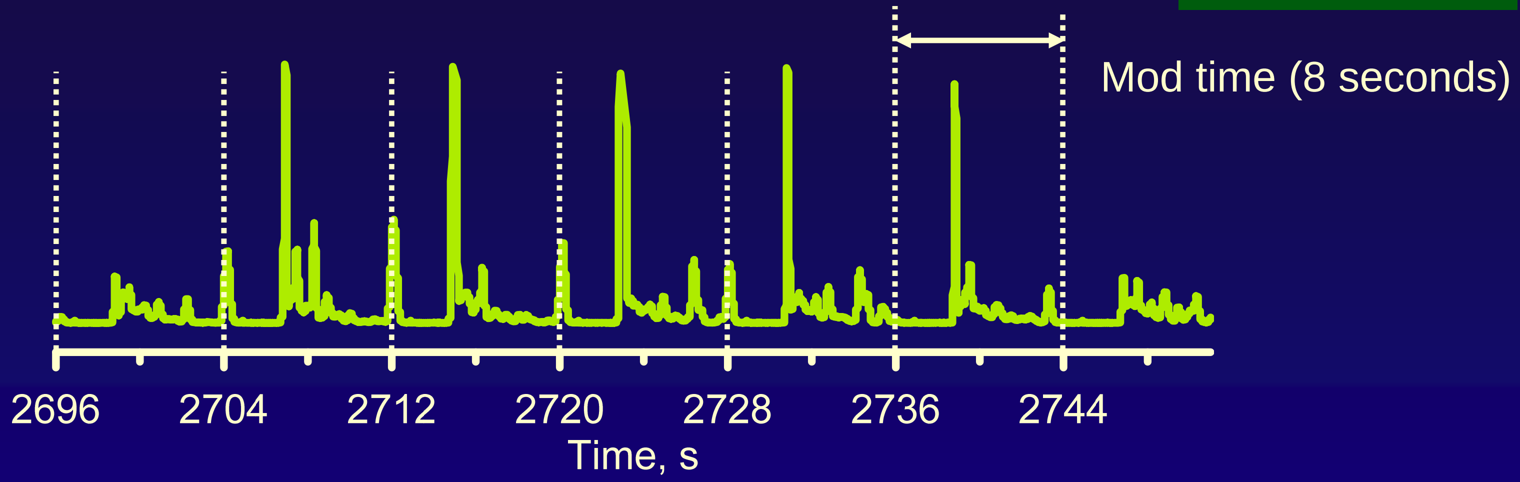
"Folding" the chromatogram

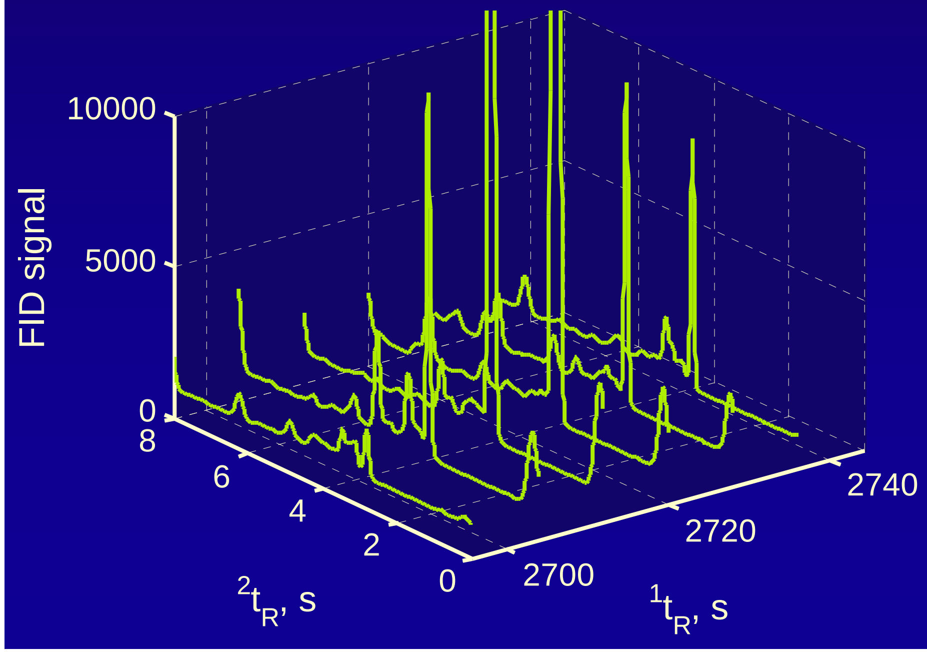
Visualisation

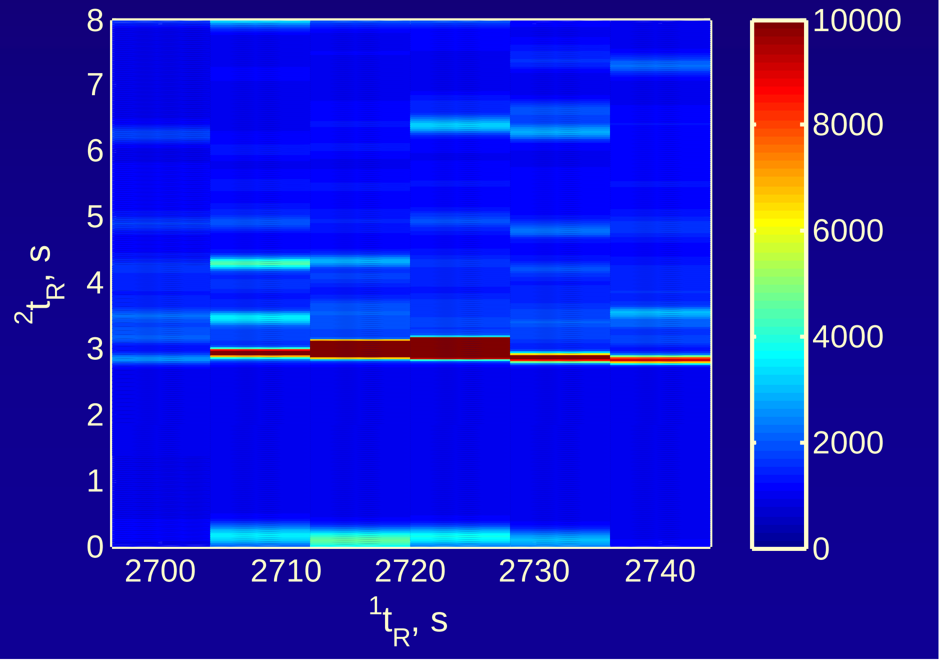


“Folding” the chromatogram

Visualisation



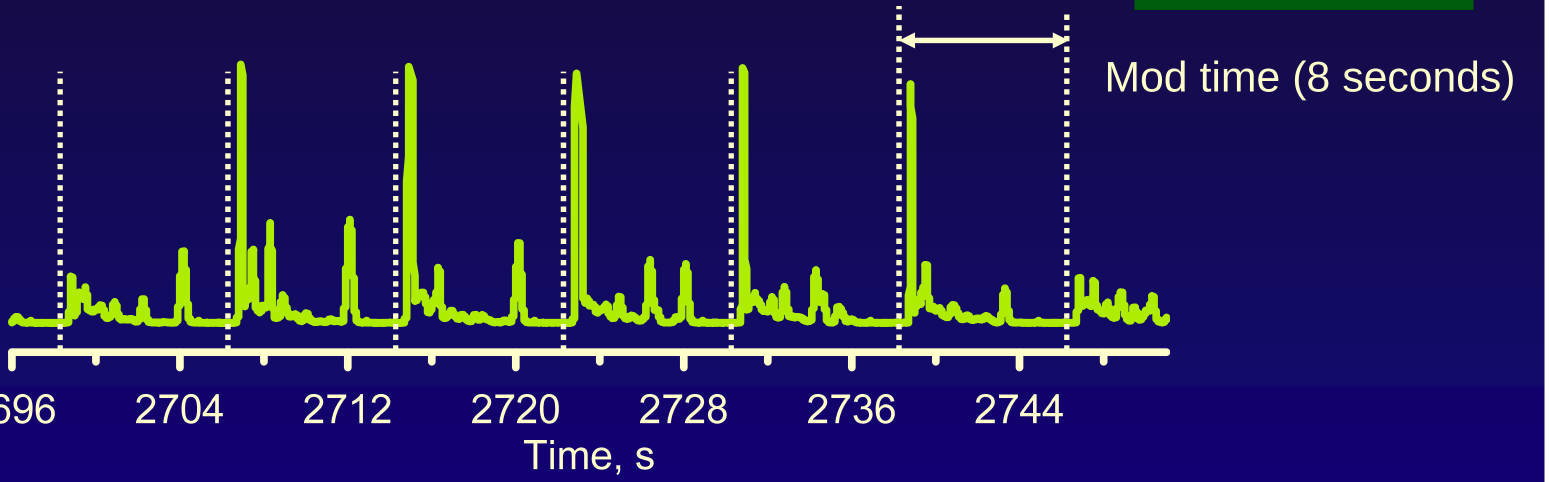


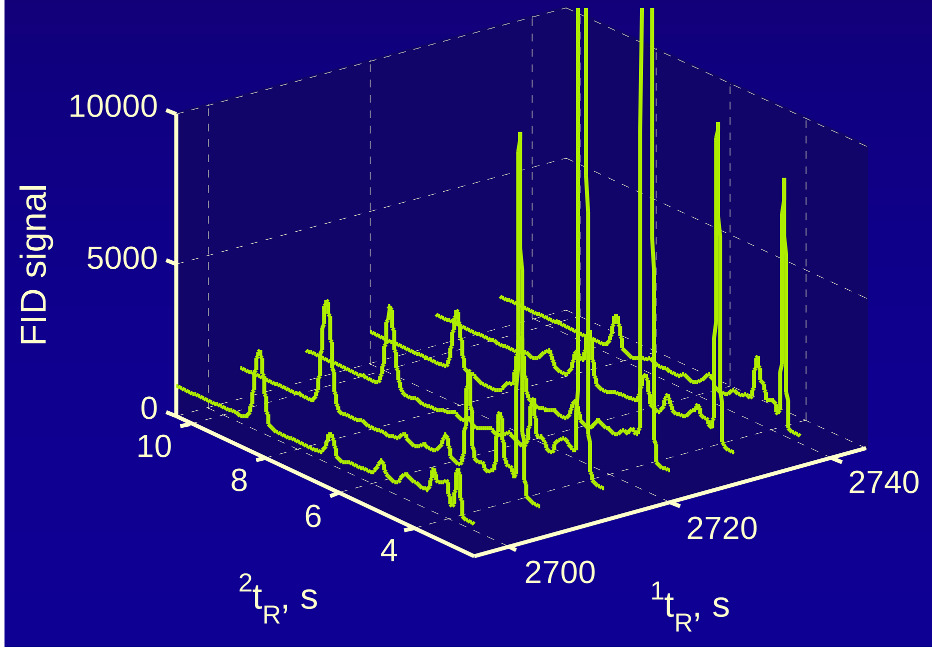


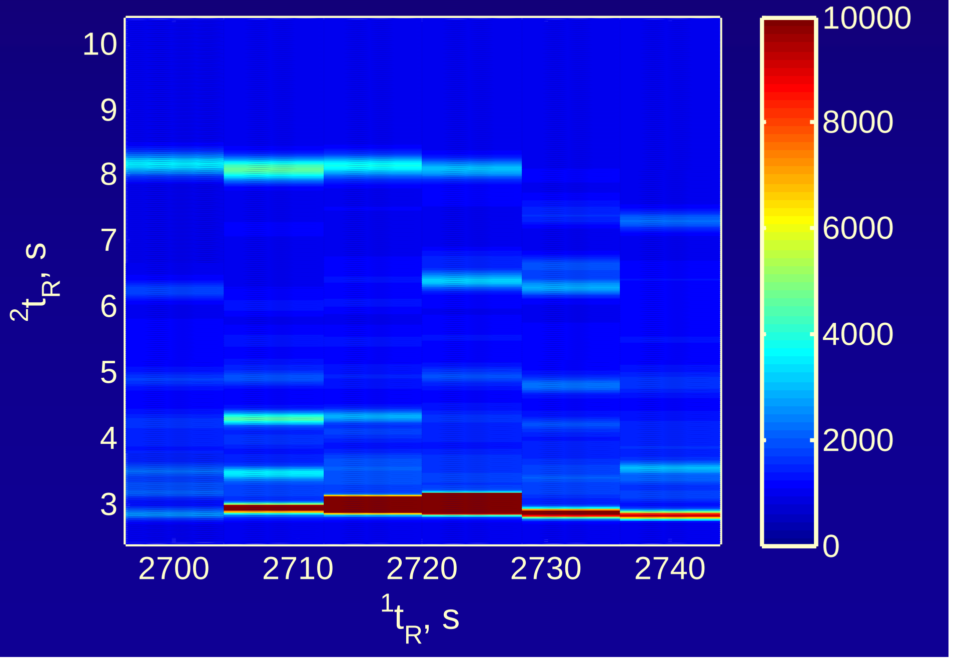
"Folding" the chromatogram: phasing

Visualisation

Phase = 0.3



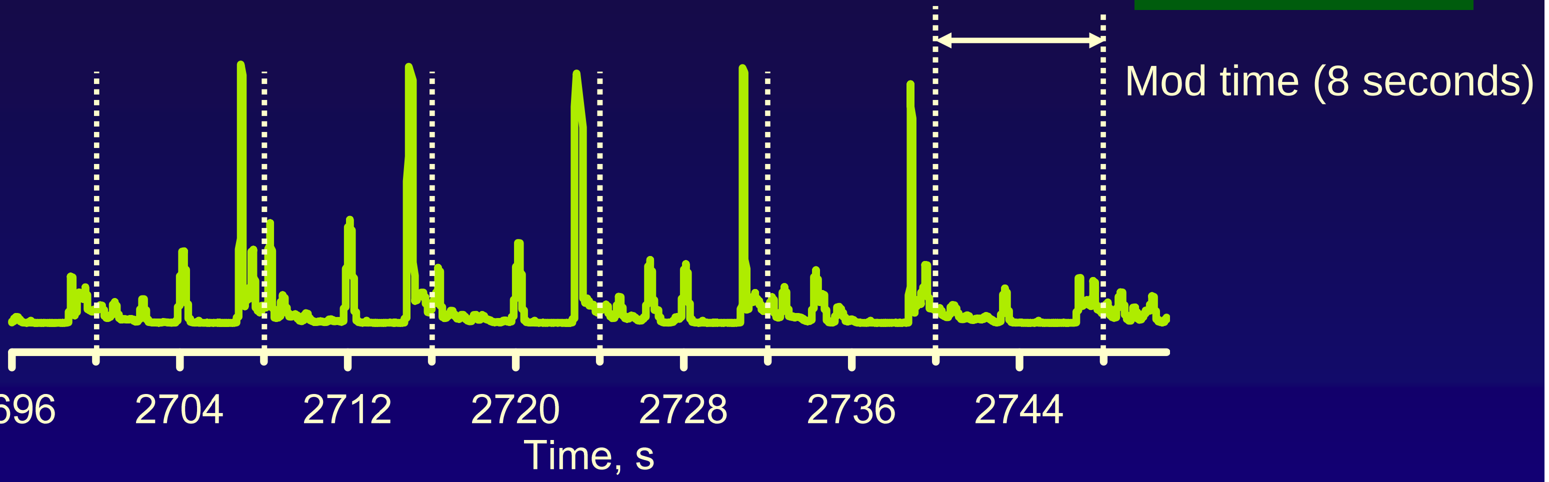


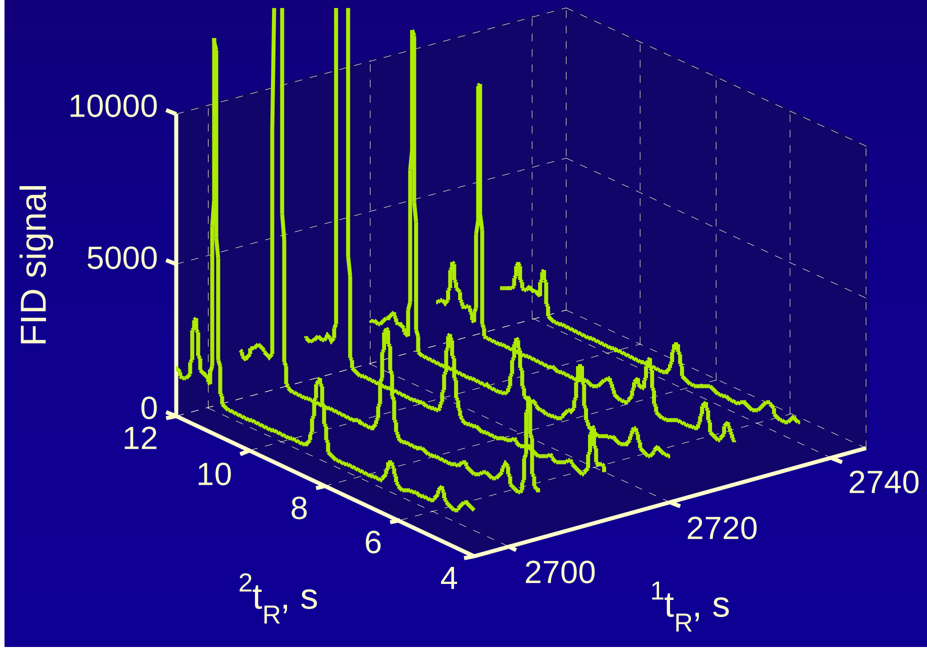


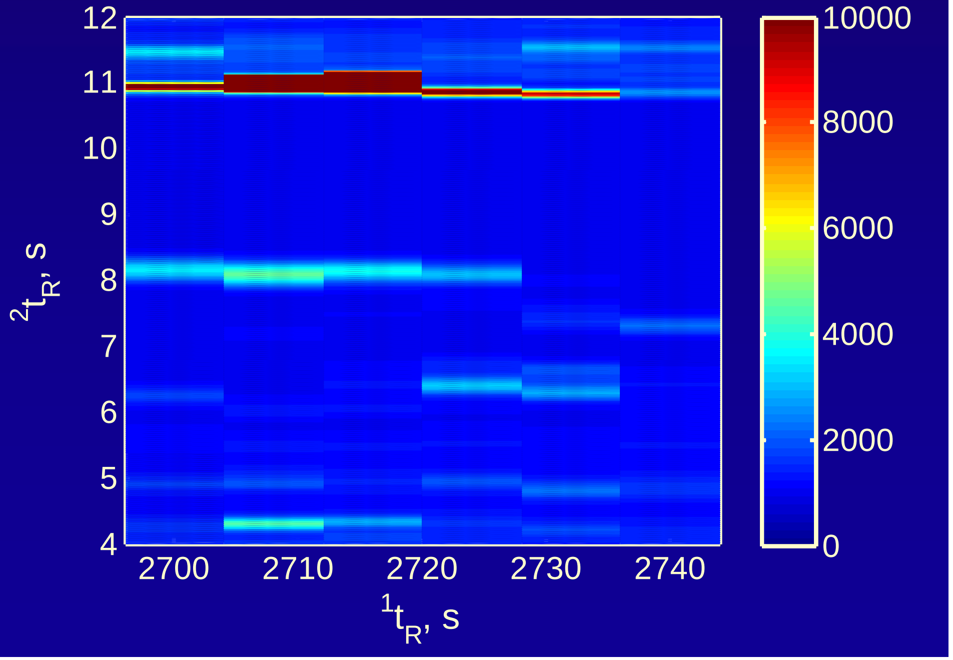
“Folding” the chromatogram: phasing

Visualisation

Phase = 0.5



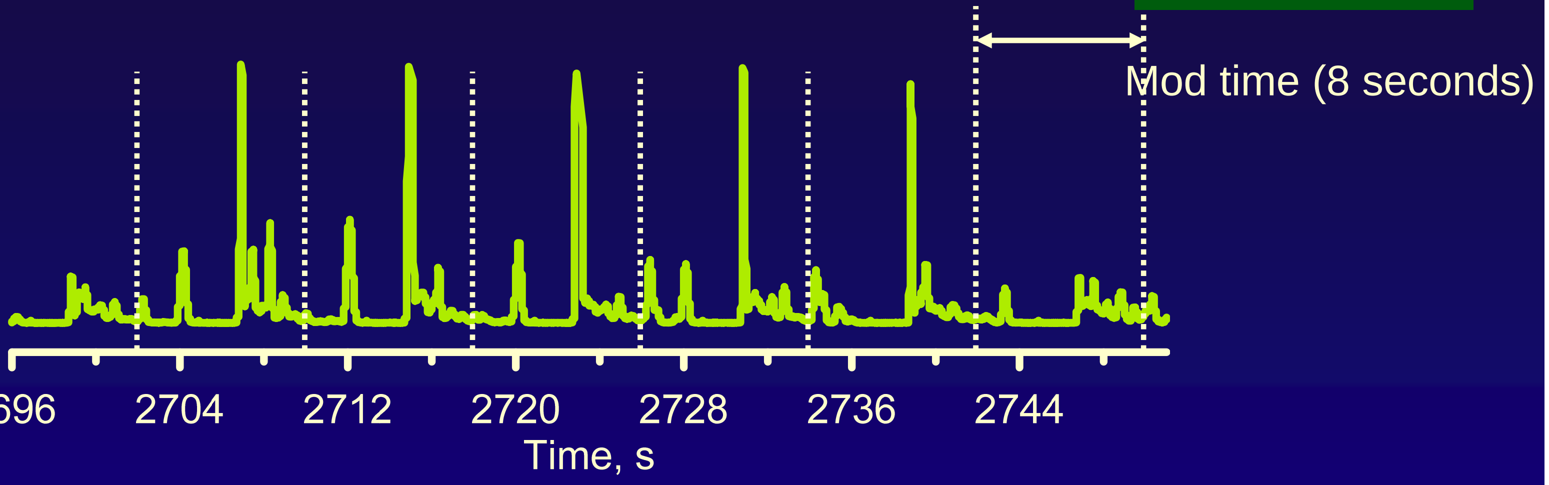


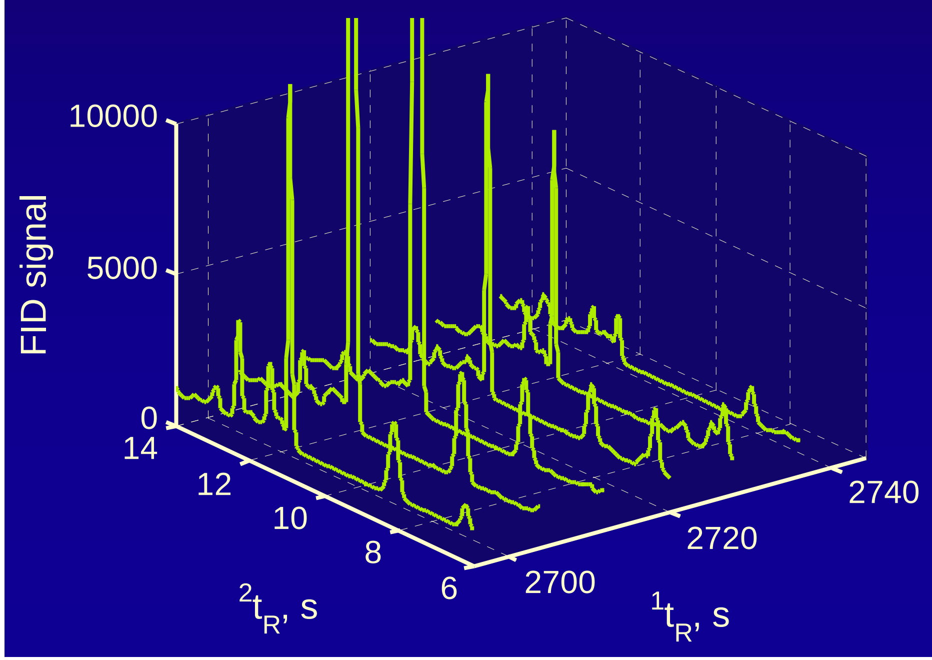


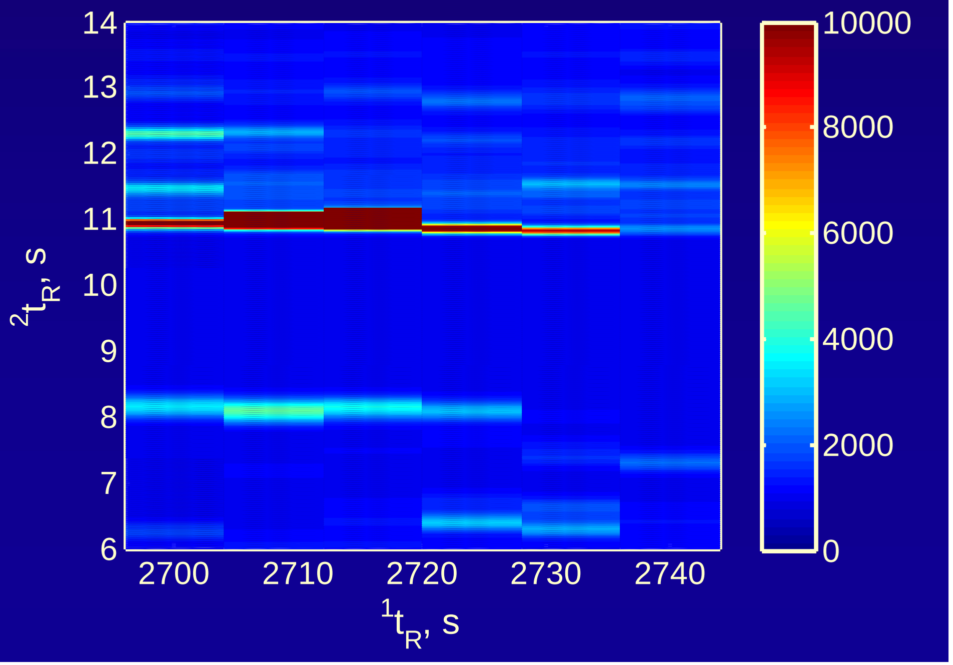
“Folding” the chromatogram: phasing

Visualisation

Phase = 0.75

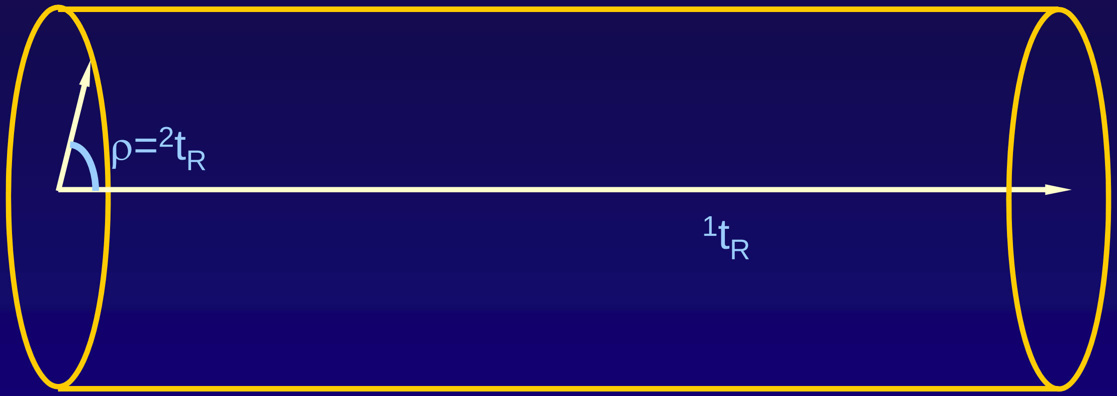


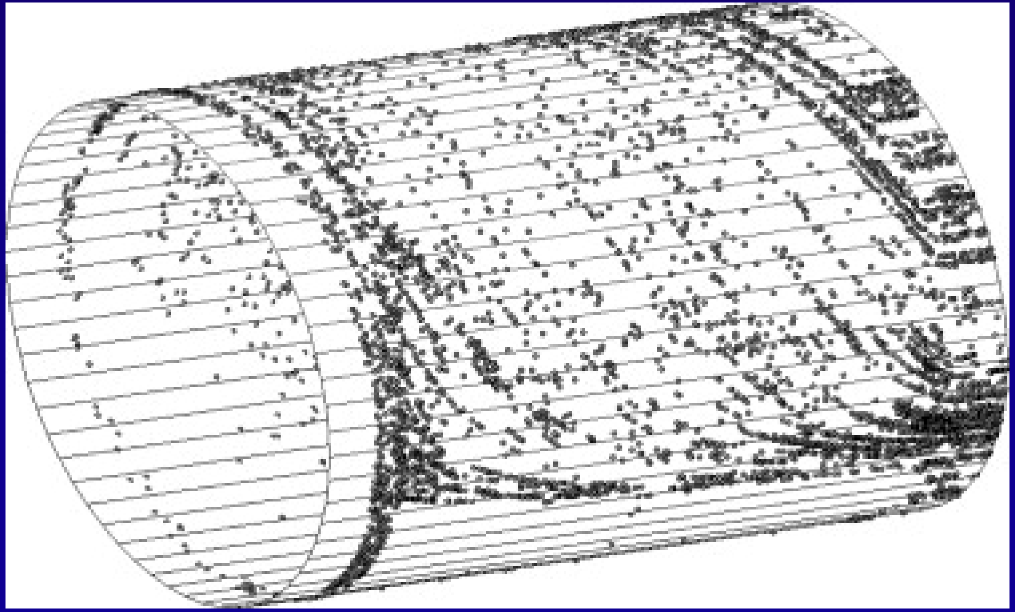




Cylindrical coordinates. An alternative way to represent the data.

Visualisation

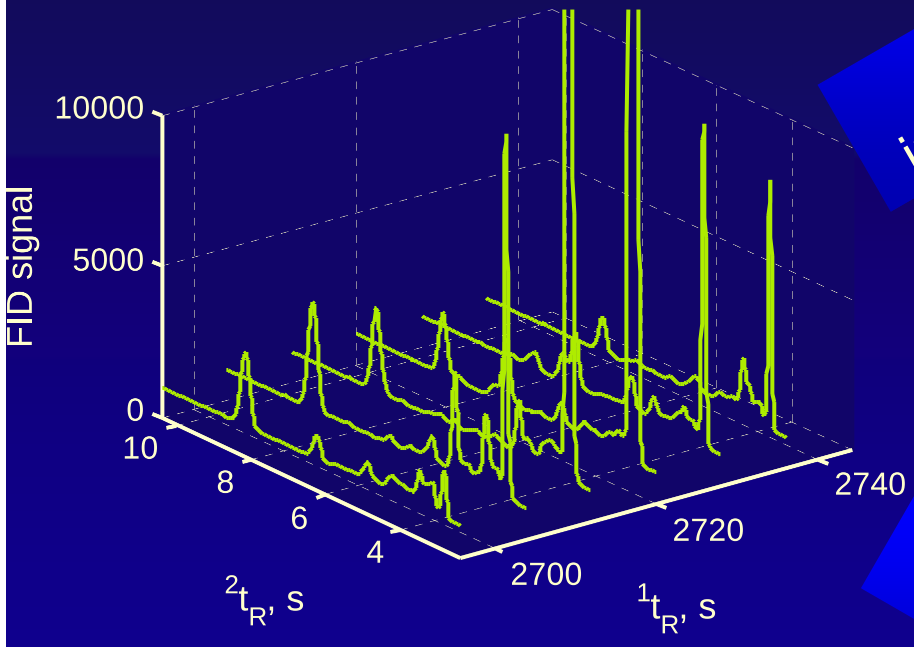




J.J.A.M. Weusten, E.P.P.A. Derks , J.H.M. Mommers, S. van der Wal, Anal. Chim. Acta 726 (2012), 9

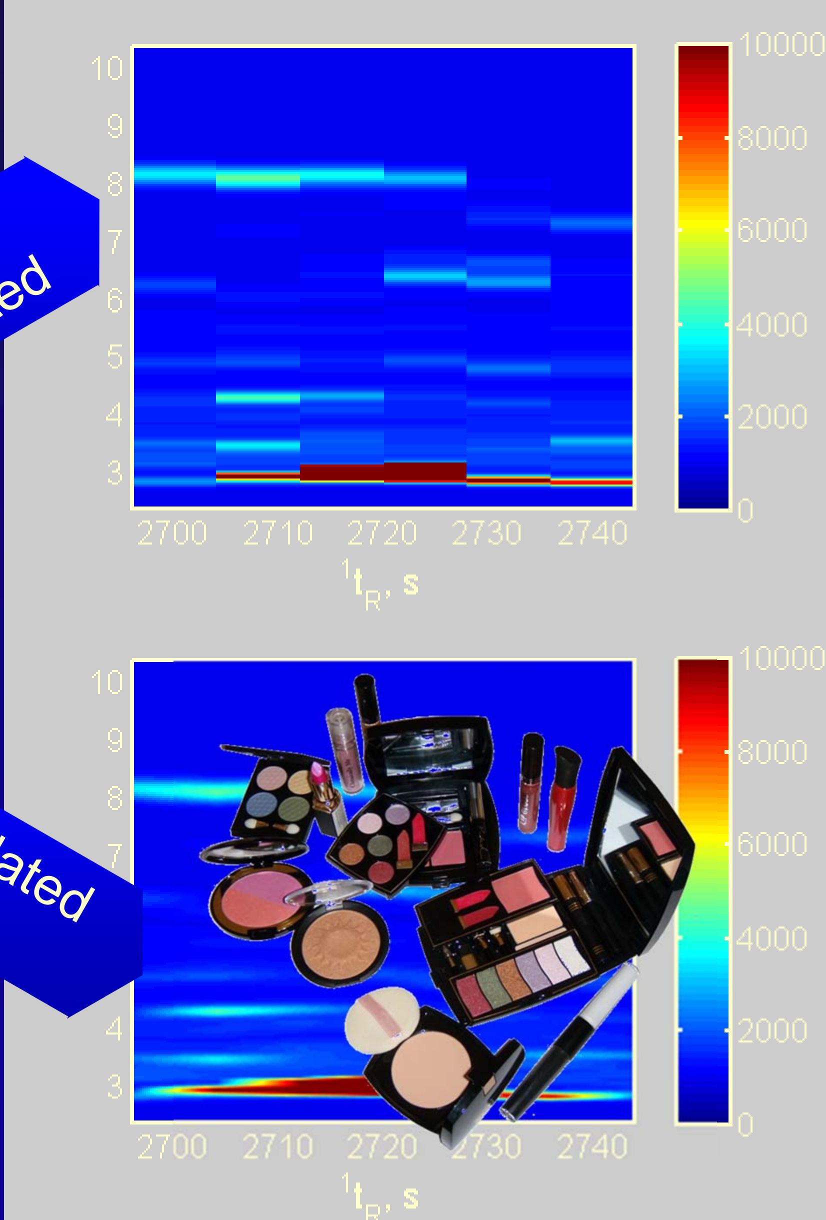
Interpolation

Visualisation



Not interpolated

Interpolated



Interpolation

Visualisation

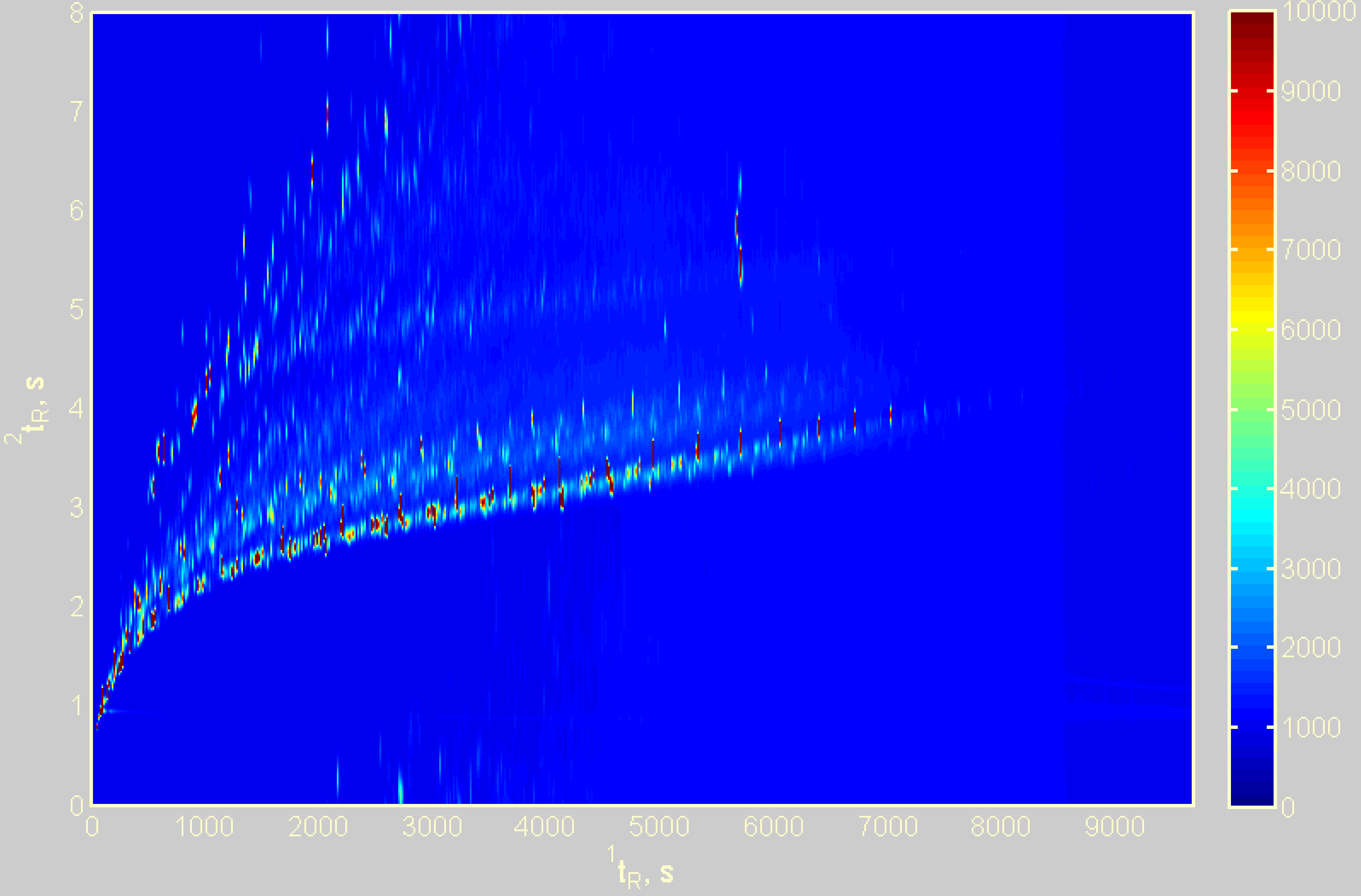
Welcome to the magic world of chemometrics!

+

=

“Folding” the chromatogram: final result

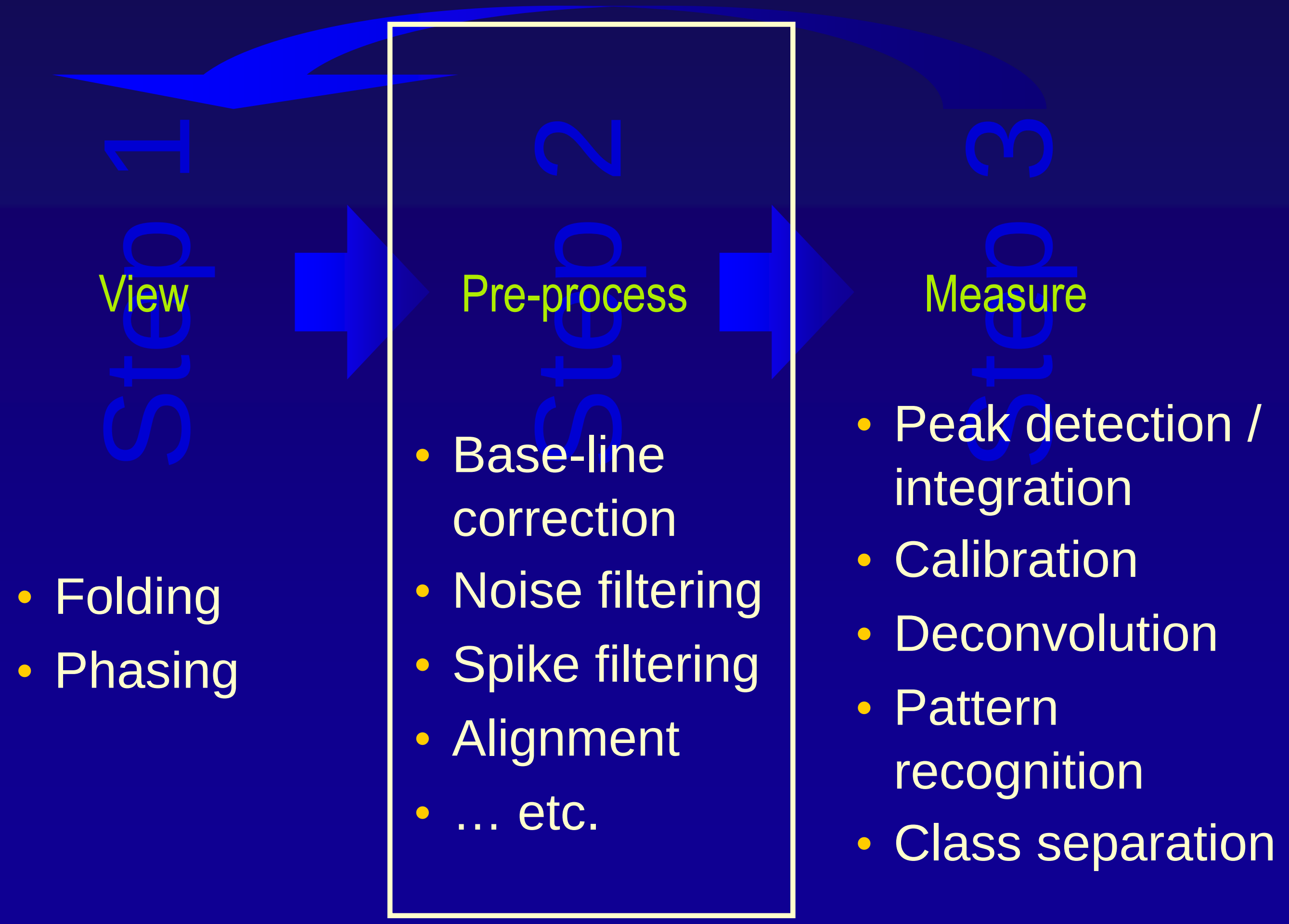
Visualisation



Conclusions

- Visualising is simple, and gives a lot of information.
- Folding (one-dimensional) data into (2D) image introduces discontinuities in the edges. Other visualization methods (cylindrical coordinates) possibloe.
- Phasing can be of great help.
- Careful with “cosmetic” effects!

Second step: Pre-processing



Step 1

View

- Folding
- Phasing

Step 2

Pre-process

- Base-line correction
- Noise filtering
- Spike filtering
- Alignment
- … etc.

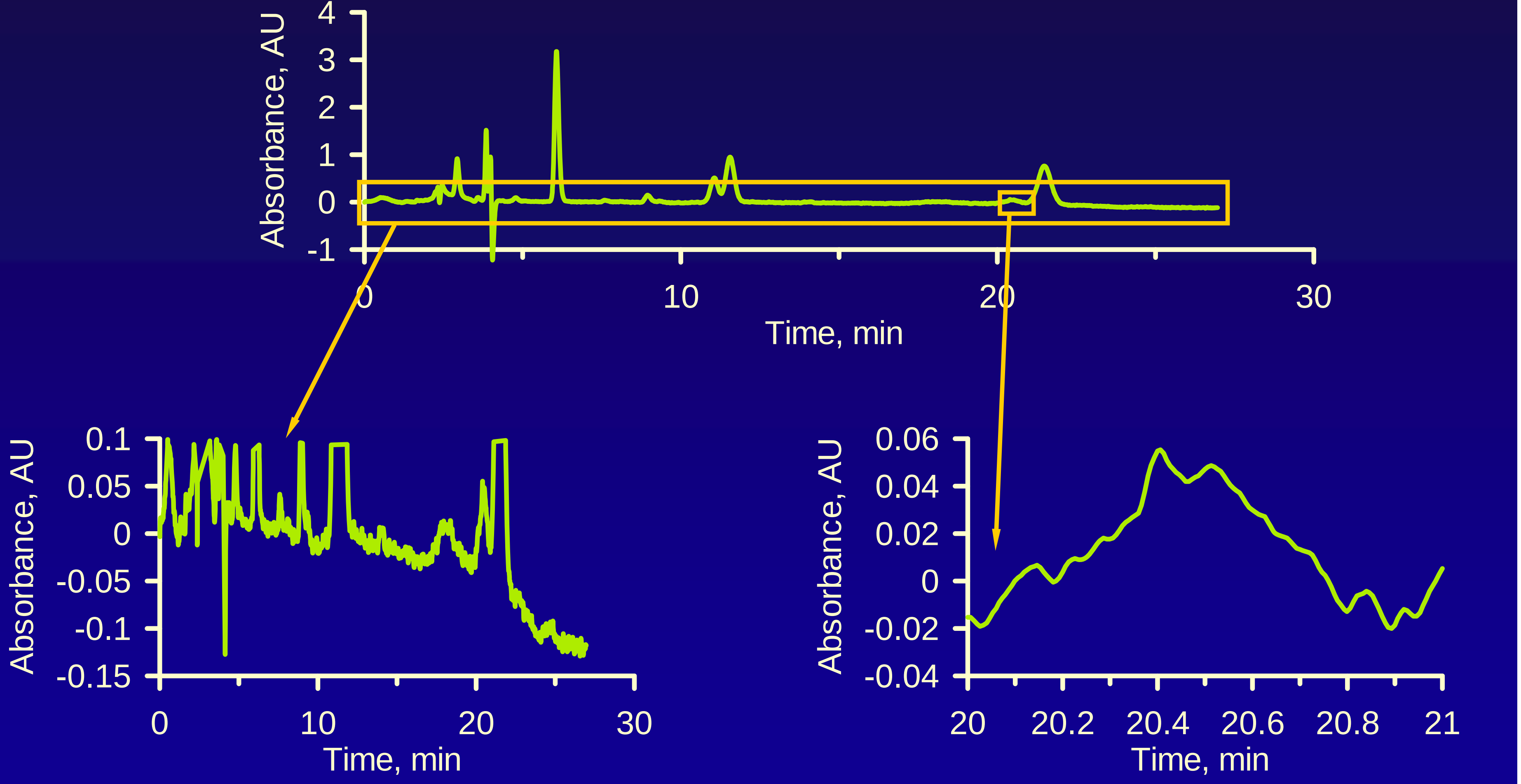
Step 3

Measure

- Peak detection / integration
- Calibration
- Deconvolution
- Pattern recognition
- Class separation

Typical problems: base-line drifts and noise

Pre-processing



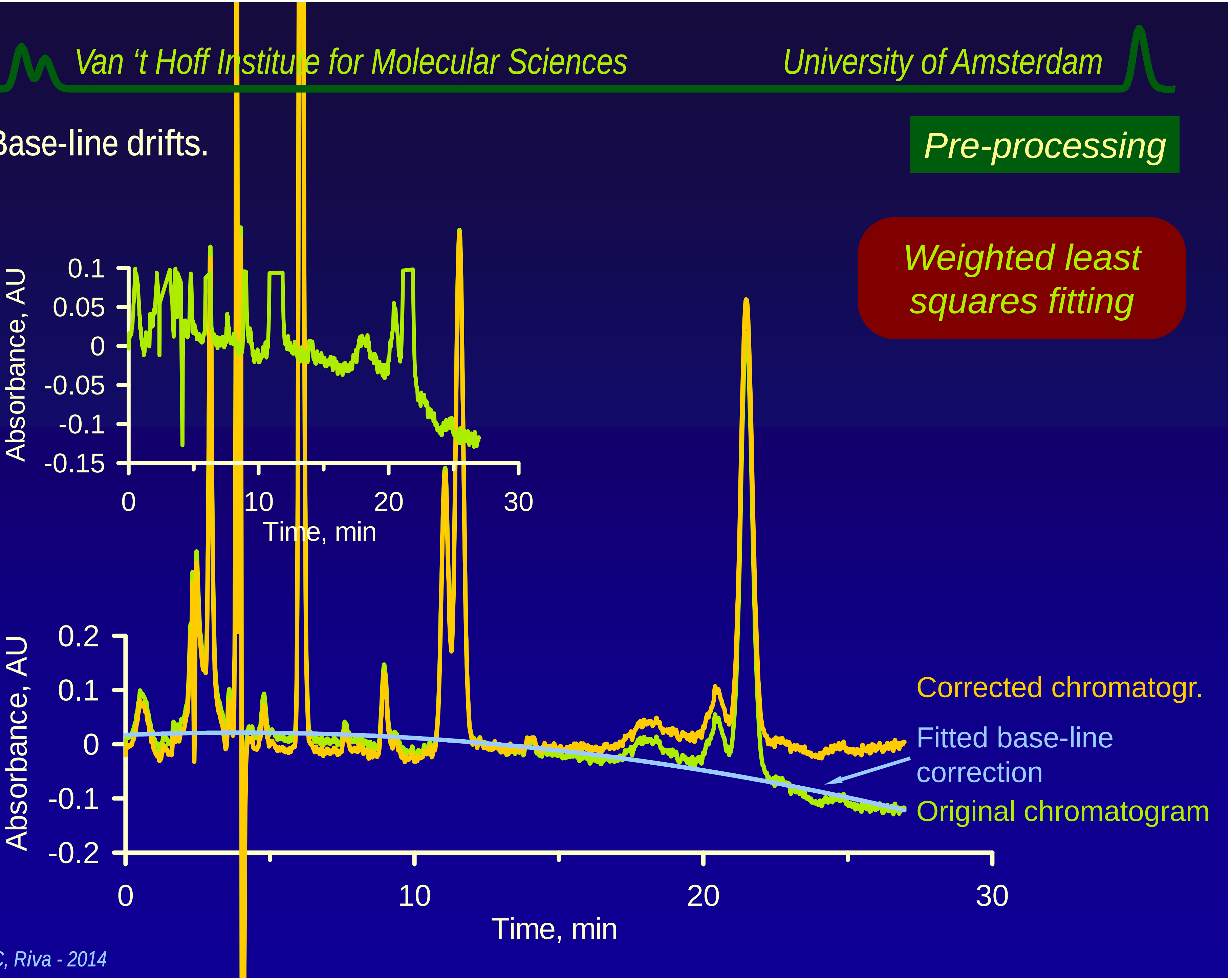
Base-line drifts

Noise

Base-line drifts.

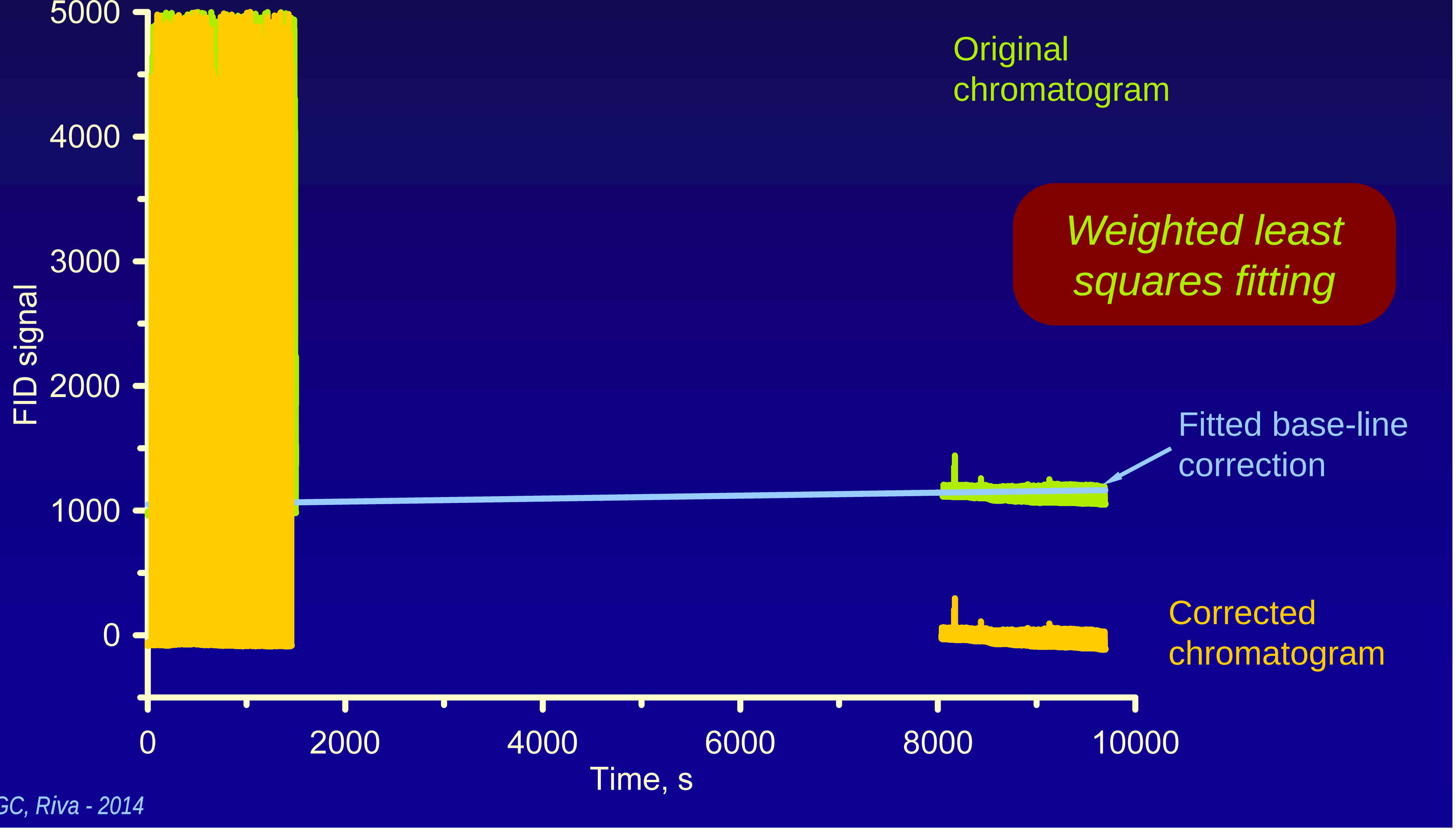
Pre-processing

Weighted least squares fitting



Base-line drifts.

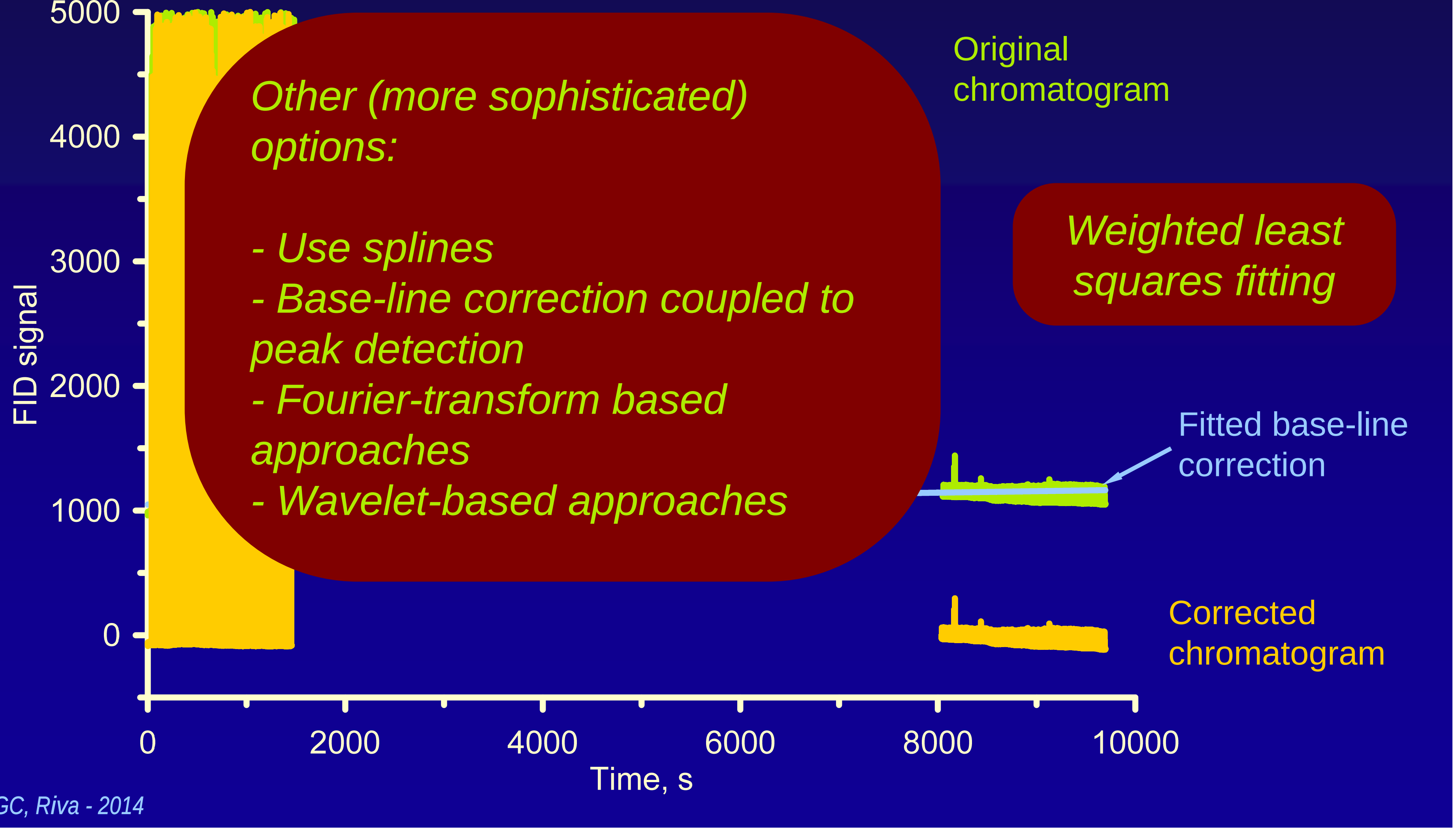
Pre-processing



Weighted least squares fitting

Base-line drifts.

Pre-processing



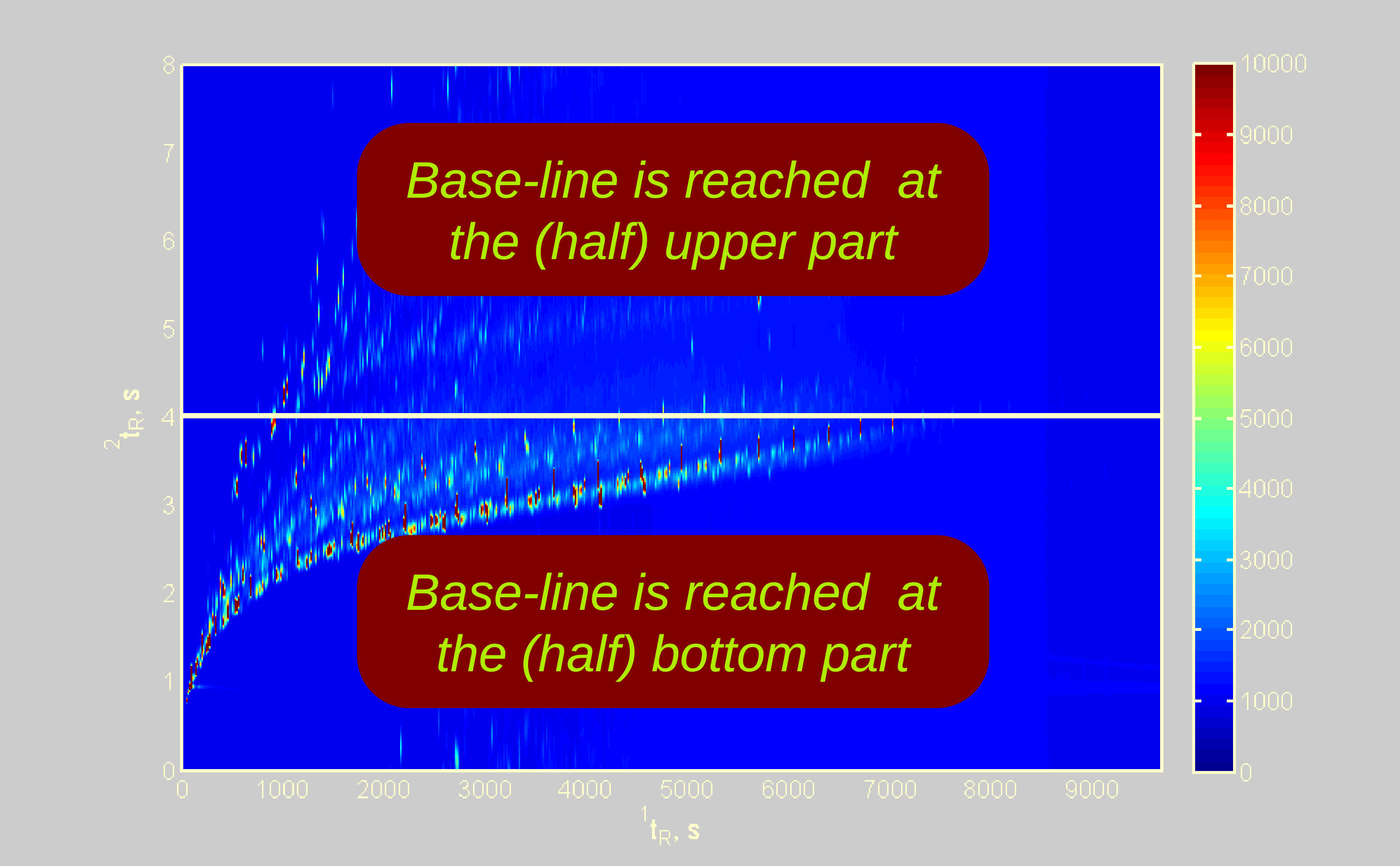
Other (more sophisticated) options:

- Use splines
- Base-line correction coupled to peak detection
- Fourier-transform based approaches
- Wavelet-based approaches

Weighted least squares fitting

Base-line drifts.

Pre-processing



Base-line is reached at the (half) upper part

Base-line is reached at the (half) bottom part

S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford Jr., J. Chromatogr. A, 985 (2003) 47 - 56

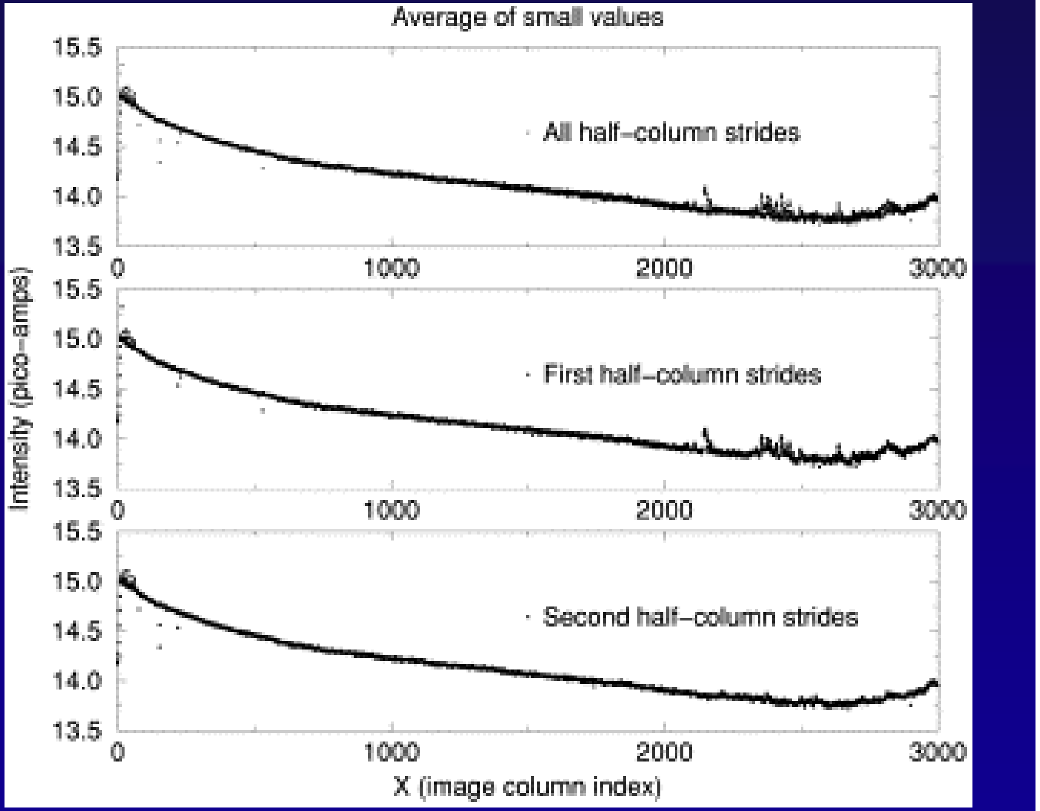
Base-line drifts.

Pre-processing

Consider the positions with the smallest values in each half

Estimate local background parameters using neighboring values

Interpolate the main background trend and substract it



S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford Jr., J. Chromatogr. A, 985 (2003) 47 - 56

Noise removal. Smoothing and derivatives.

Pre-processing

Savitzky-Golay filter is the most common method

Two parameters should be optmisized

- Window size
- Polynomial degree

These parameters govern how much correlated noise is removed

- Large window sizes and low polynomial degree

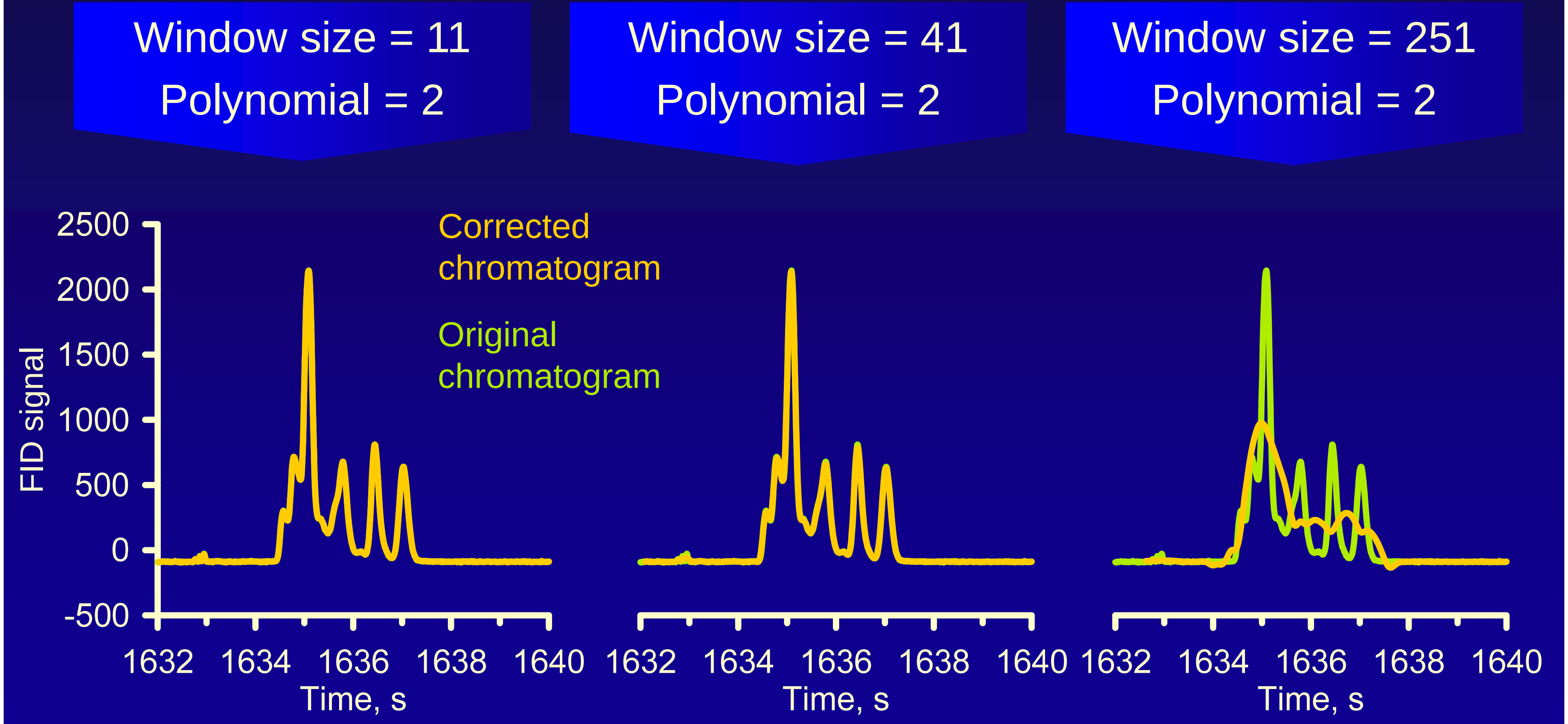
Too much noise is removed (chromatograms appear deformed)

- Small window sizes and large polynomial degrees

Too much noise remains

Noise removal. Smoothing and derivatives.

Pre-processing

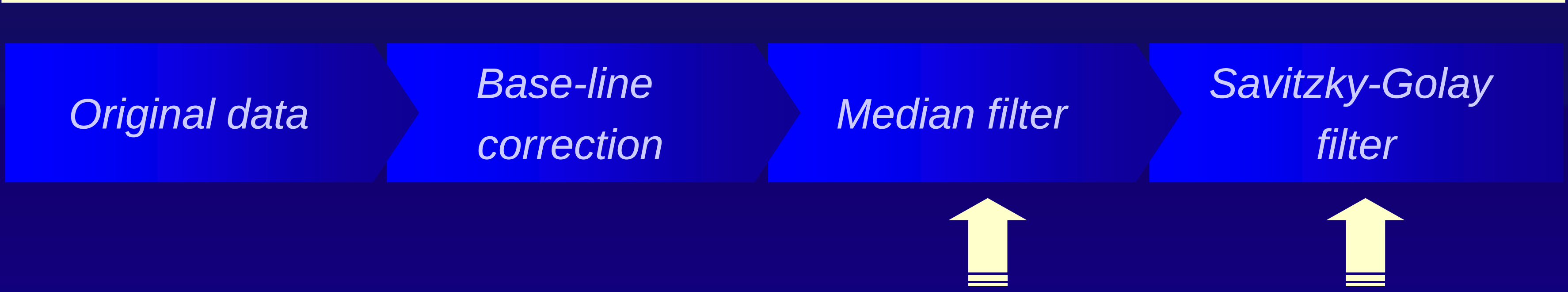


G. Vivó-Truyols, P.J. Schoenmakers, "Automatic selection of optimal Savitzky-Golay smoothing", Anal. Chem. 78 (2006) 4598-4608.

Noise removal. Spikes.

Pre-processing

A good way of removing spikes consists of passing a median filter (before the Savitsky-Golay filter)



Parameter to tune: window size

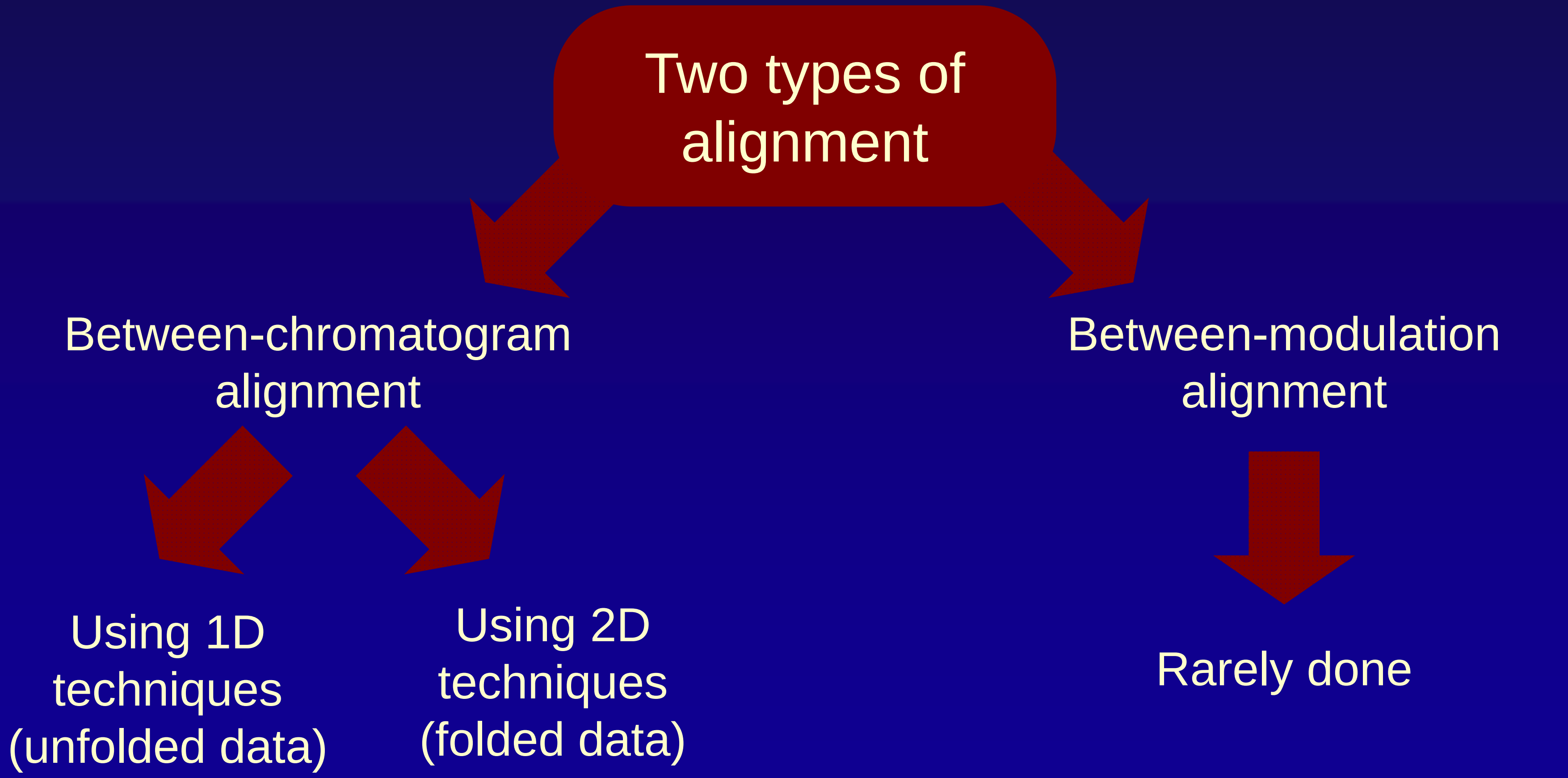
Parameters to tune: window size and polynomial degree

Optimising three parameters

Alignment.

Pre-processing

Alignment is not always necessary, depending on the final objective of the analysis



Two types of alignment

Between-chromatogram alignment

Using 1D techniques (unfolded data)

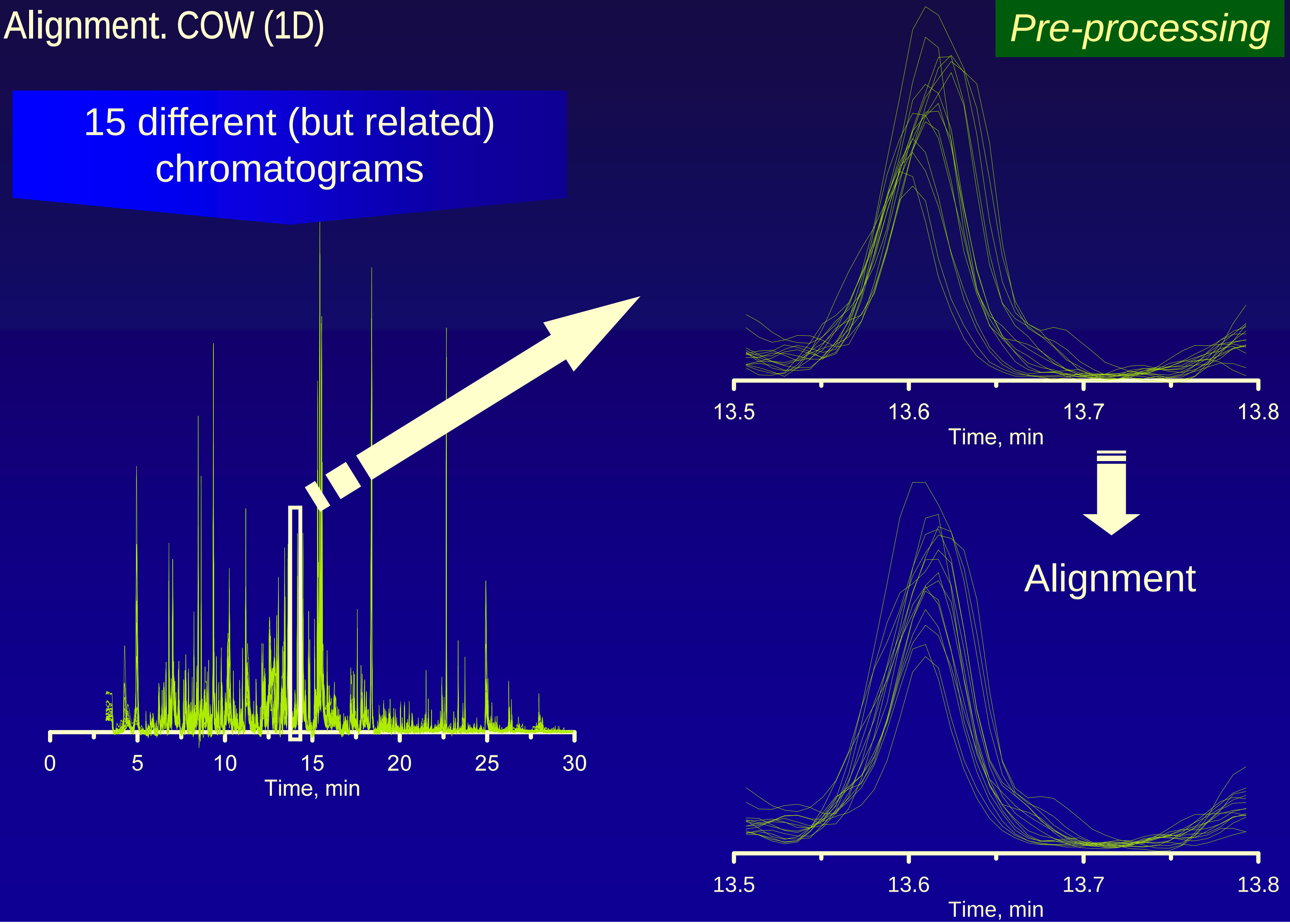
Using 2D techniques (folded data)

Between-modulation alignment

Rarely done

Alignment. COW (1D)

Pre-processing



15 different (but related) chromatograms

Alignment

Alignment. Using (truly) 2D algorithms

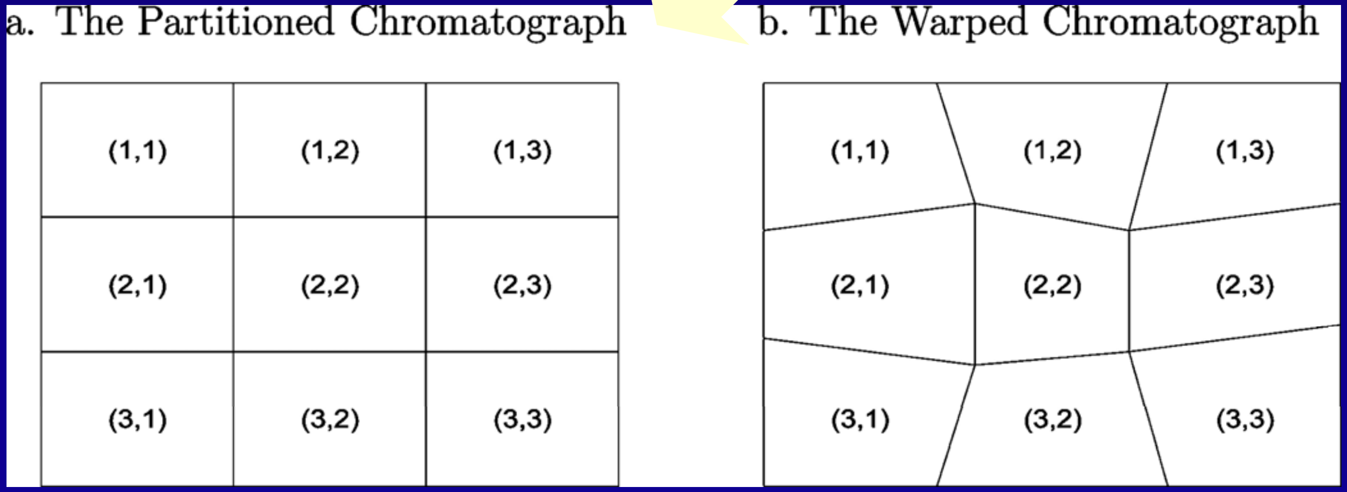
Pre-processing

Score alignment in GCxGC-MS

S. Castillo, I. Mattila, J. Miettinen, M. Orešič, T.Hyötyläinen, Anal. Chem. 83 (2011) 3058–3067

COW-adapted GCxGC-MS (using single channel)

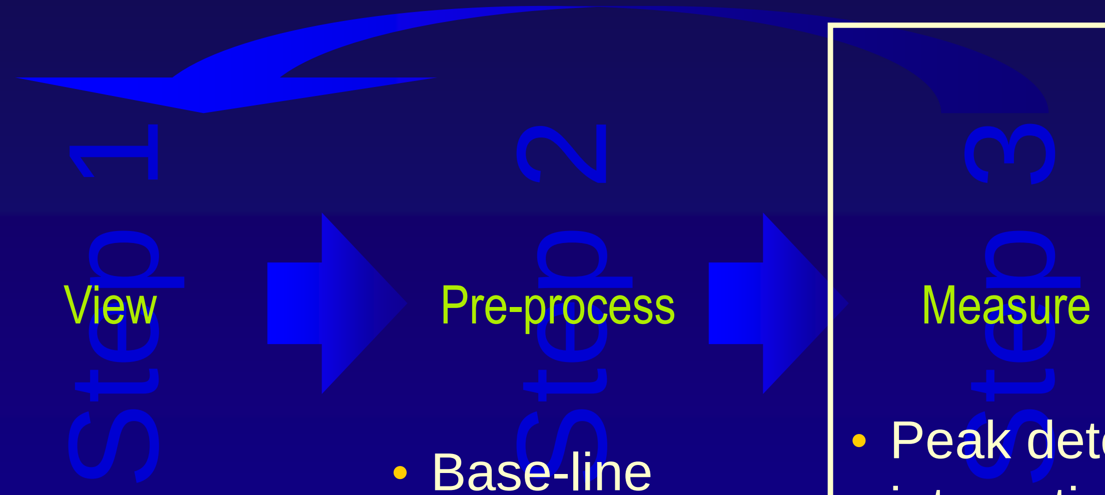
D. Zhang, X. Huang, F.E. Regnier,M. Zhang, Anal. Chem., 80 (2008) 2664–2671



Conclusions

- Pre-processing methods are almost the same: one-dimensional = two-dimensional. Normally done in the (pre-folded) raw data.
- Every case needs a particular solution (it always exists, but some care should be taken!)

Third step: measure



Step 1

View

Step 2

Pre-process

Step 3

Measure

- Folding
- Phasing

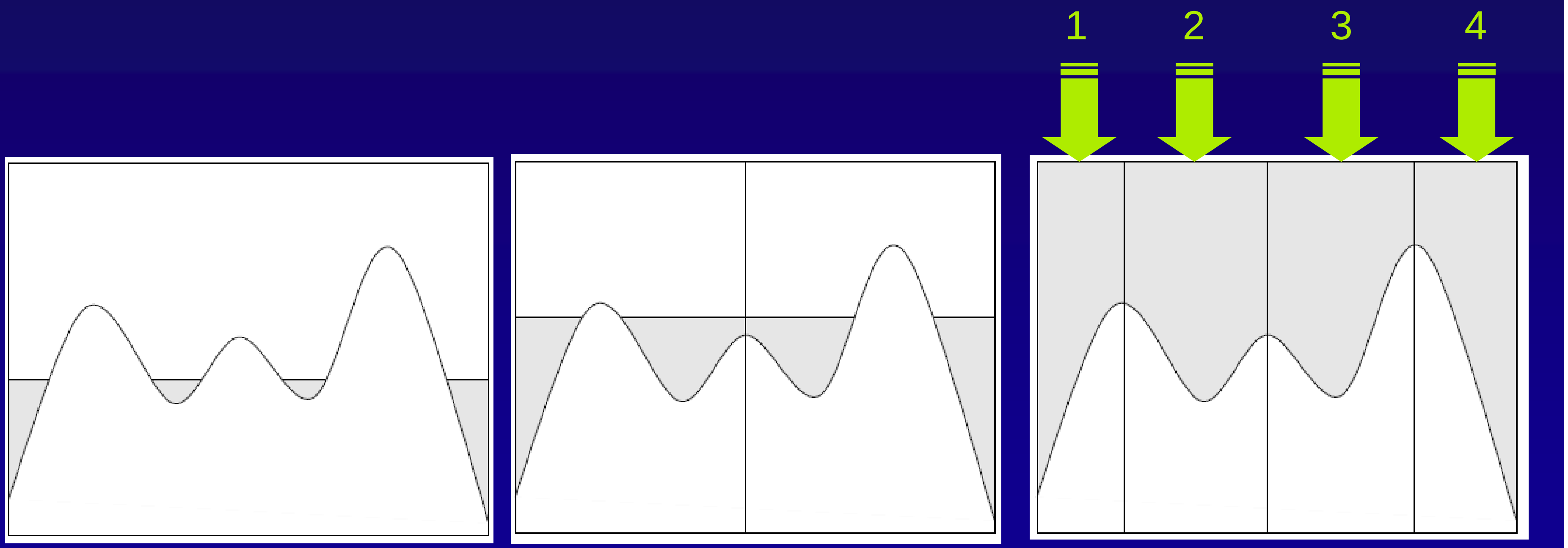
- Base-line correction
- Noise filtering
- Spike filtering
- Alignment
- … etc.

- Peak detection / integration
- Calibration
- Deconvolution
- Pattern recognition
- Class separation

Peak detection in one step: the watershed algorithm

Peak detection

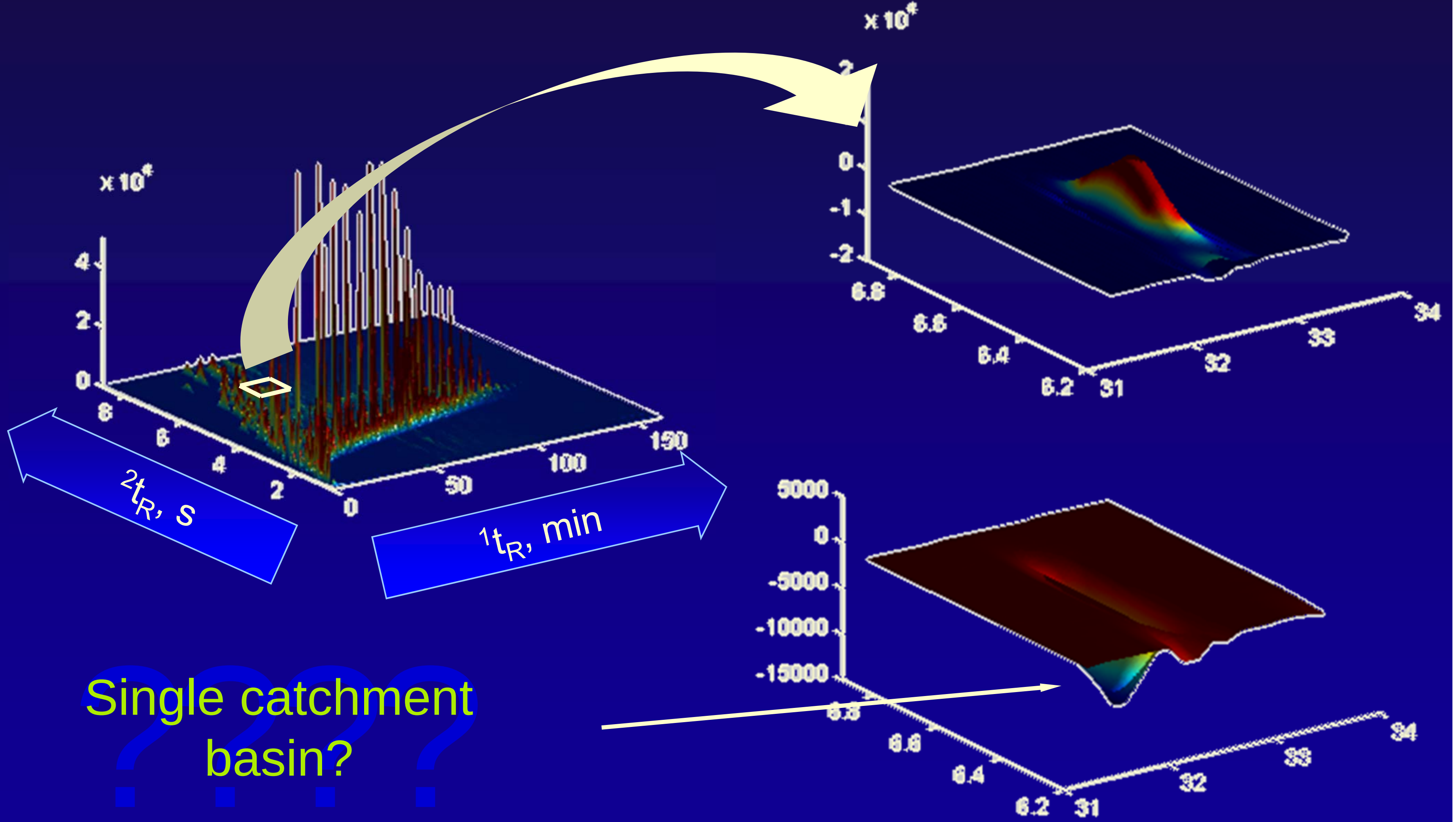
Most common software programs use the watershed algorithm to detect peaks in 2D chromatography:



J. De Bock et al., doi 10.1007/11558484

Peak detection in one step: the watershed algorithm

Peak detection



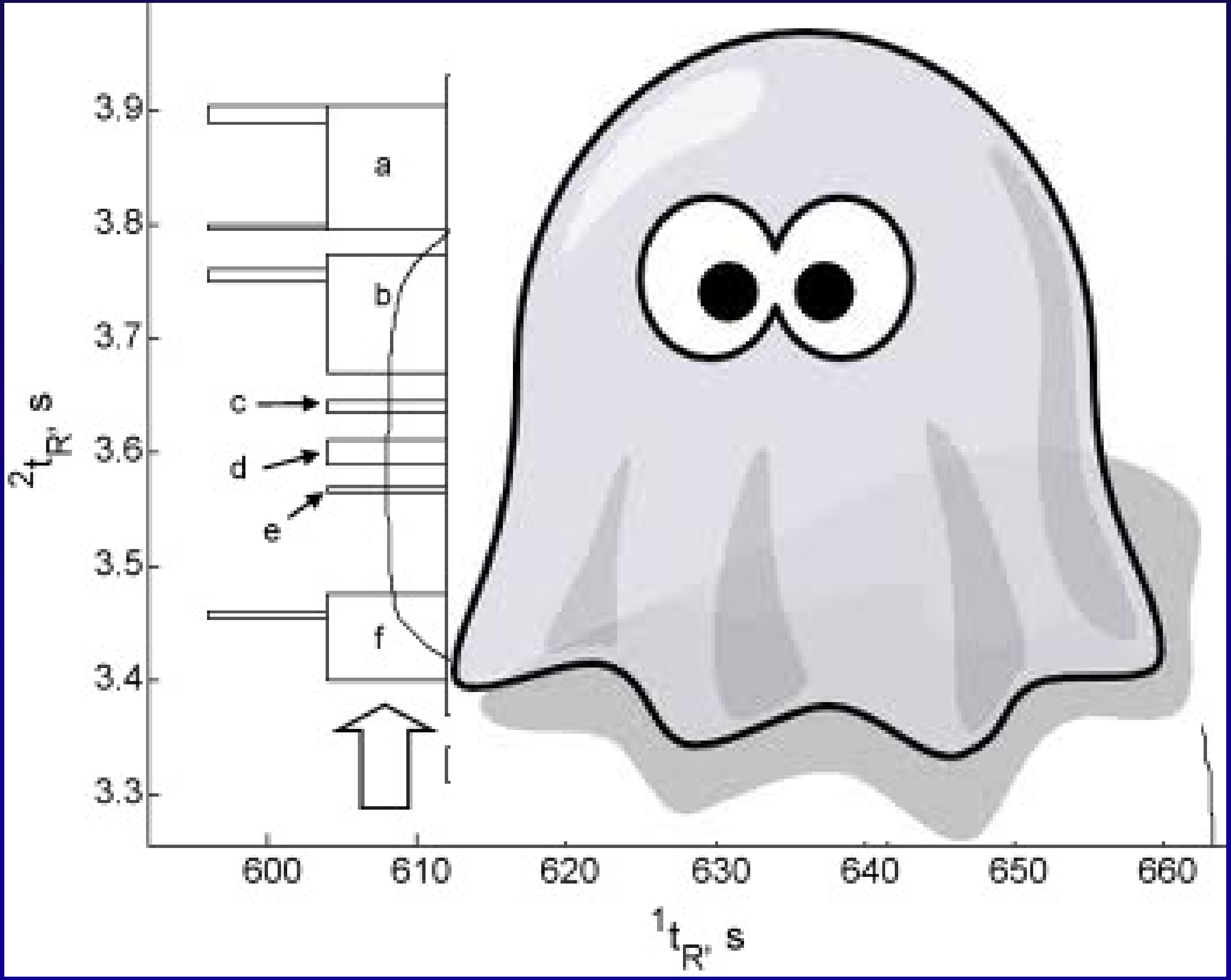
Single catchment basin?

G. Vivó-Truyols, H.G. Janssen, J. Chromatogr. A, doi:10.1016/j.chroma.2009.12.063

Peak detection in one step: the watershed algorithm

Peak detection

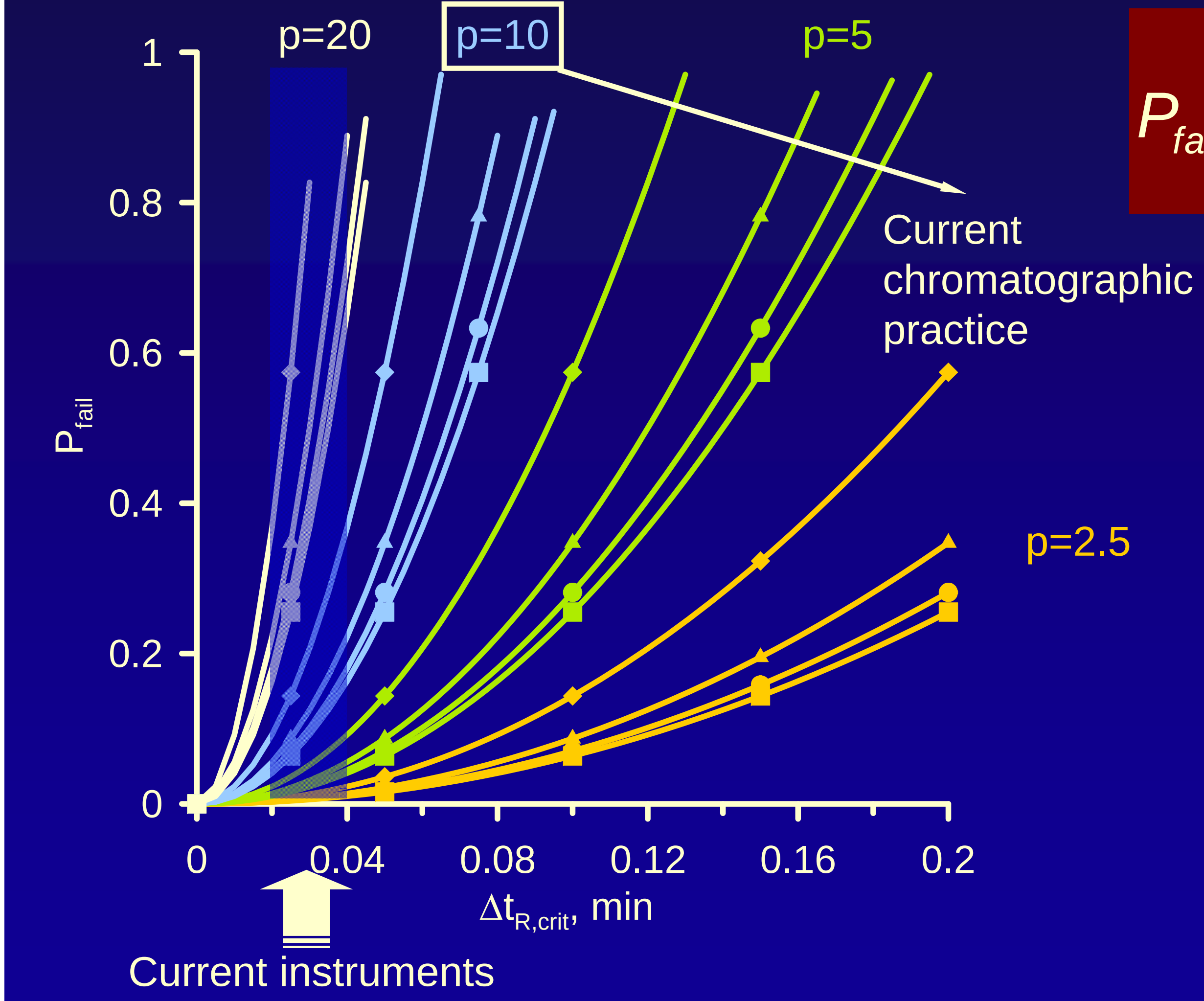
The first problem using the watershed algorithm.



Peak detectionin one step: the watershed algorithm

Peak detection

Critical variability in second-dimension retention time:



$$P_{fail} = (\Delta t_{crit})^2 \left(\frac{1}{\alpha + \beta q} p \right)^2$$

$$p = \frac{{}^1\sigma}{{}^2\sigma m} \qquad q = \frac{m}{{}^1\sigma}$$

◆ q=0.5 (eight cuts per peak)

▲ q=1 (four cuts per peak)

● q=2 (two cuts per peak)

■ q=4 (1 cut per peak)

G. Vivó-Truyols, H.G. Janssen, J. Chromatogr. A, doi:10.1016/j.chroma.2009.12.063

Peak detectioni in two steps.

Peak detection

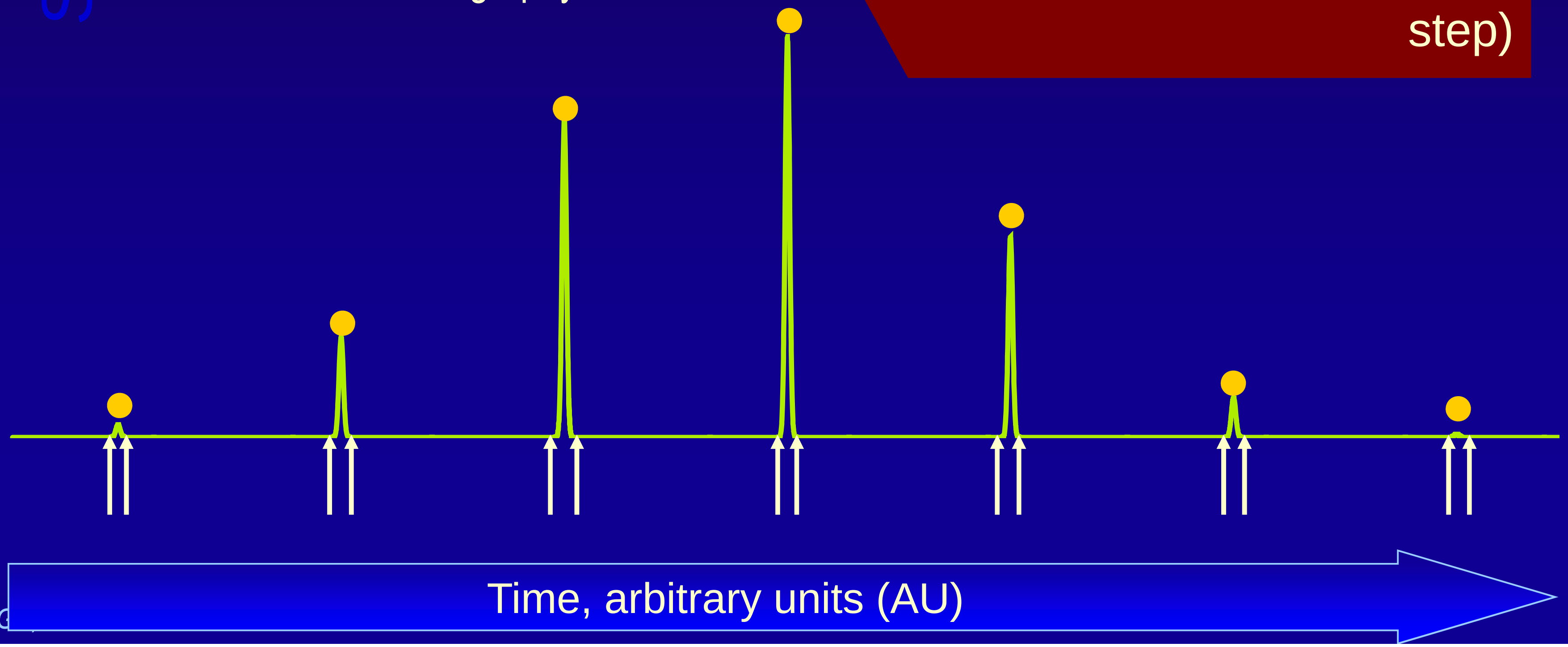
S. Peters, G. Vivó-Truyols, P.J. Marriott and P.J. Schoenmakers, J. Chromatogr. A 1156 (2007) 14.

E.J.C. van der Klift, G. Vivó-Truyols, F.W. Claassen, F.L. van Holthoon, T.A. van Beek, J. Chromatogr. A, 1178 (2008) 43.

Step 1

Detect peaks as in one-dimensional chromatography

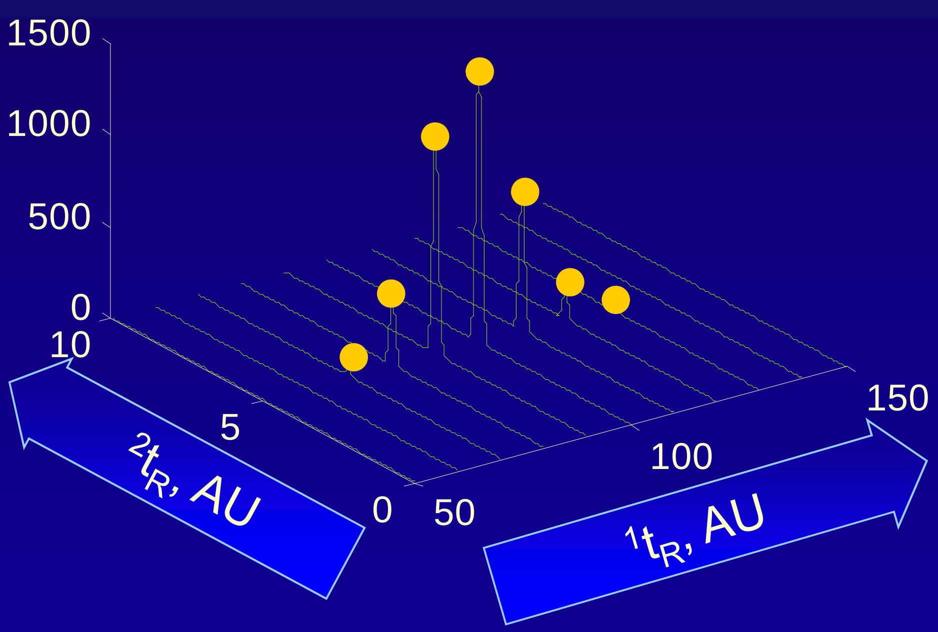
Use information from derivatives (pre-processing step)



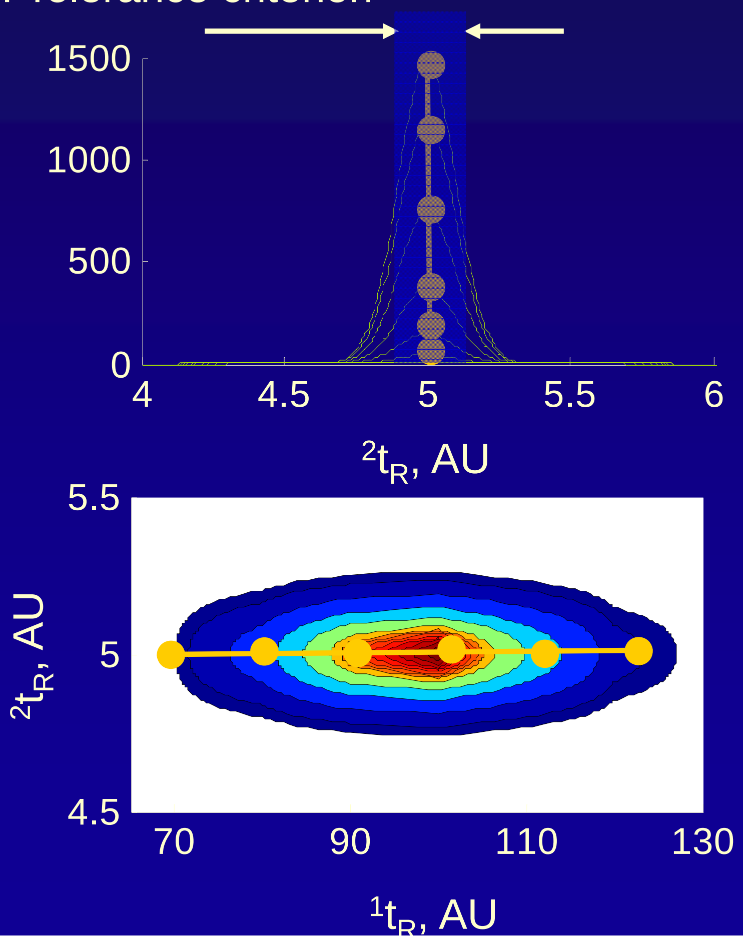
Peak detectioni in two steps.

Peak detection

step 2 Merge peaks that belong to the same compound according to 2nd-dimension retention time differences



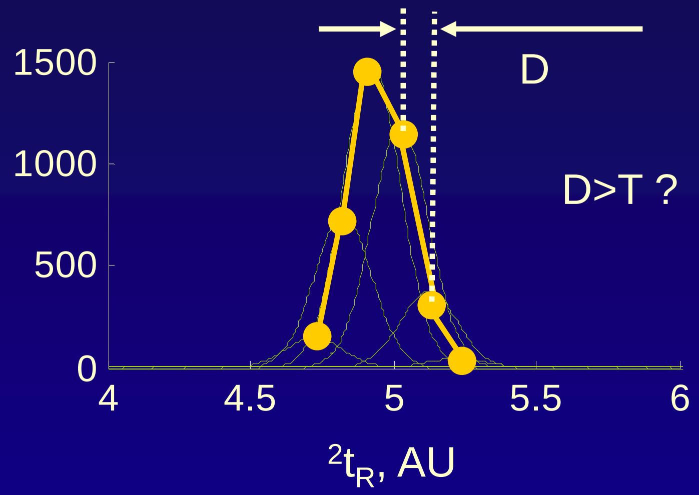
T: Tolerance criterion

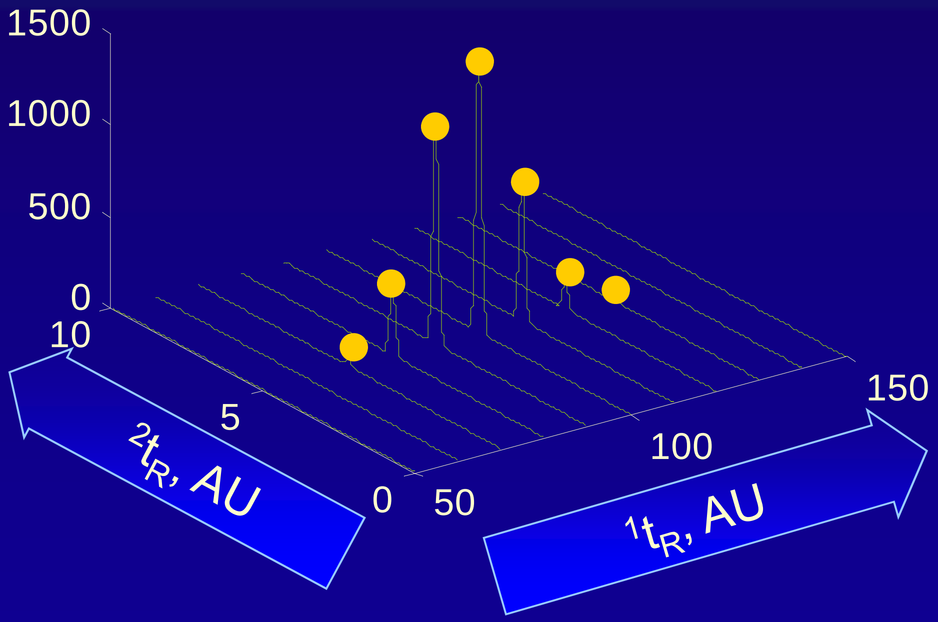


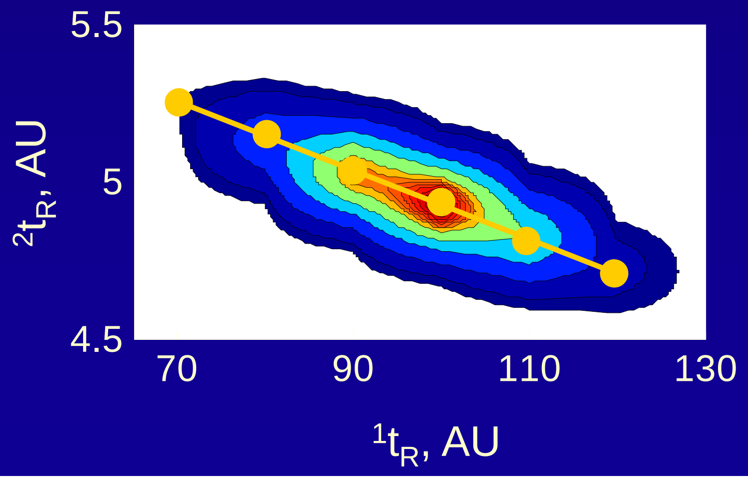
Peak detectioni in two steps.

Peak detection

step 2 Merge peaks that belong to the same compound according to 2nd-dimension retention time differences



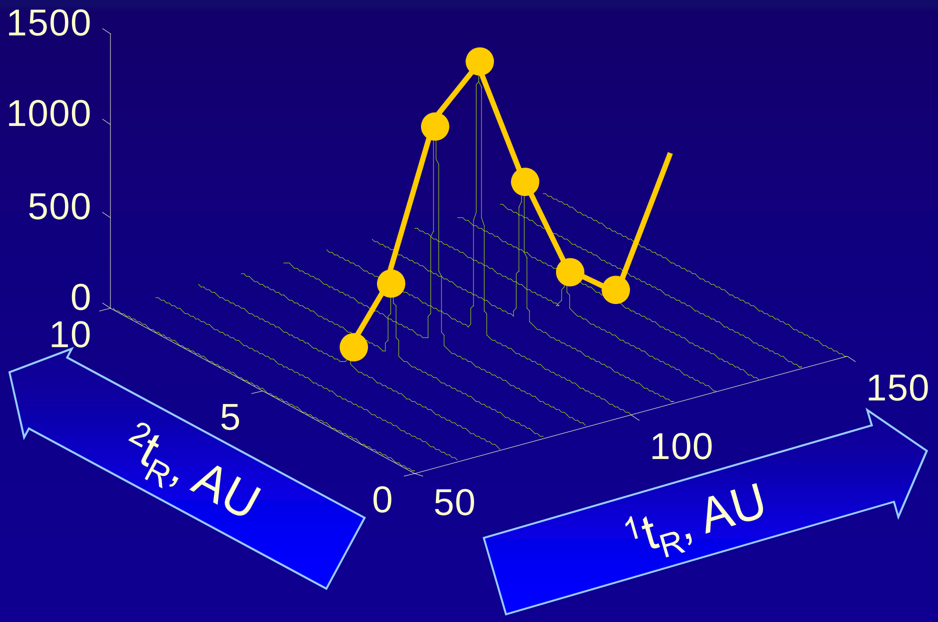


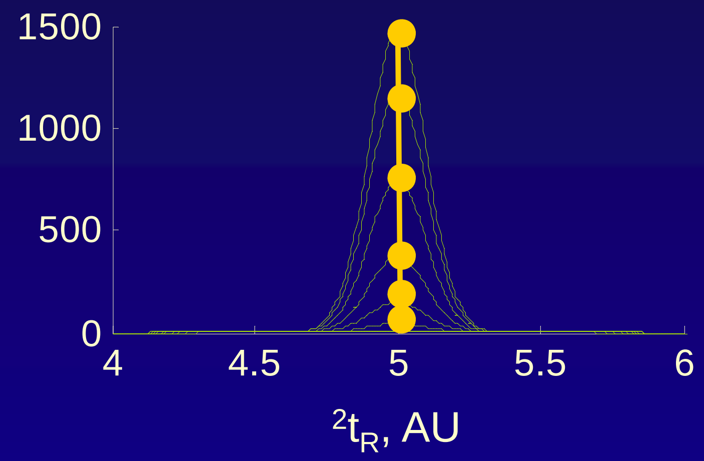


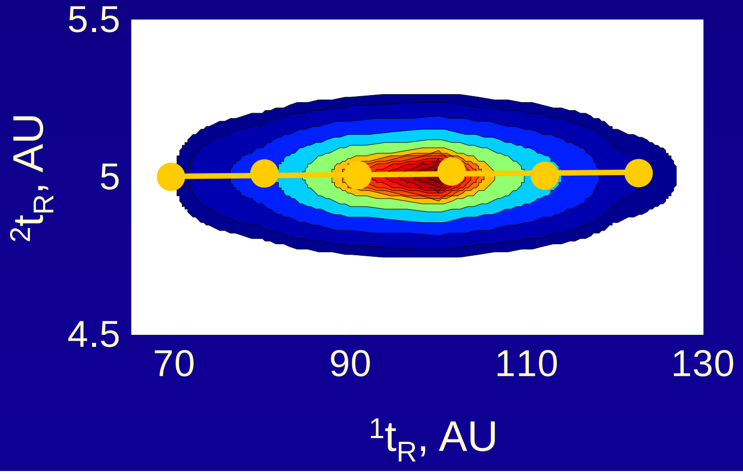
Peak detectioni in two steps.

Peak detection

step 3 Check unimodality





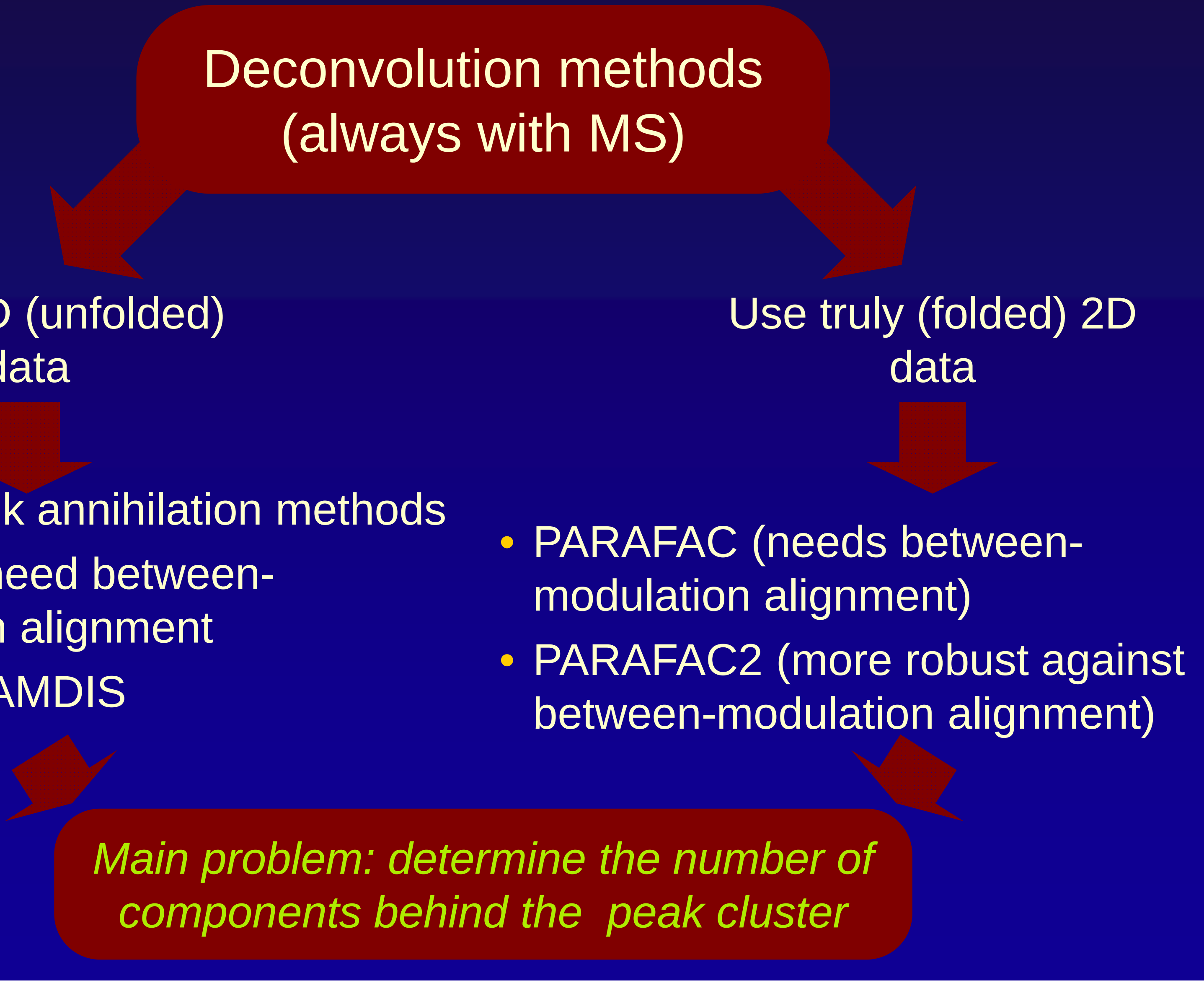


Conclusions

Peak detection

- Two methods available: (inverse) watershed, and two-step peak detection process.
- Peak detection seems to be still a subject of discussion.

Deconvolution mehtods.



Deconvolution methods (always with MS)

Using 1D (unfolded) data

- ALS or rank annihilation methods
- Does not need between-modulation alignment
- Similar to AMDIS

Use truly (folded) 2D data

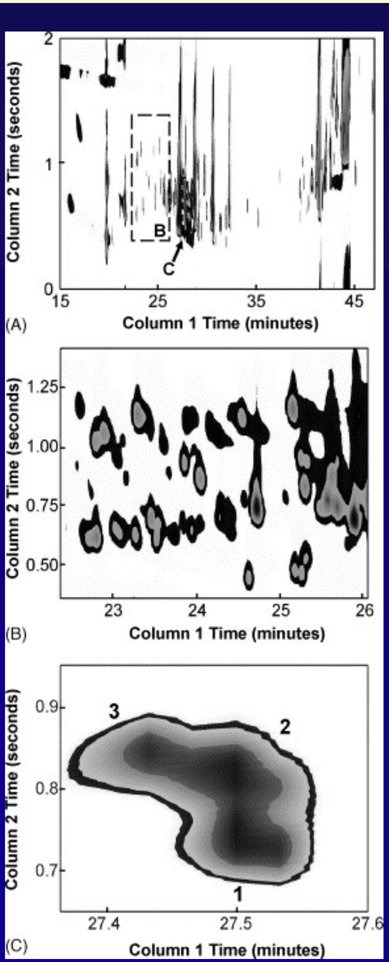
- PARAFAC (needs between-modulation alignment)
- PARAFAC2 (more robust against between-modulation alignment)

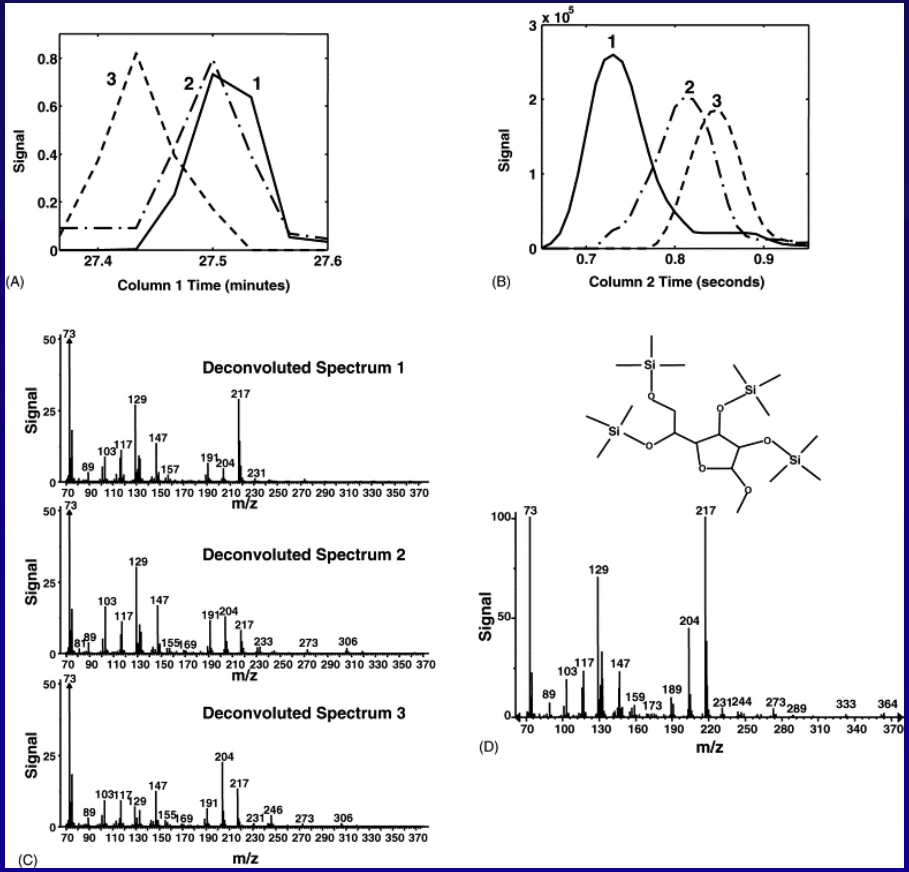
Main problem: determine the number of components behind the peak cluster

Deconvolution mehtods.

Deconvolution

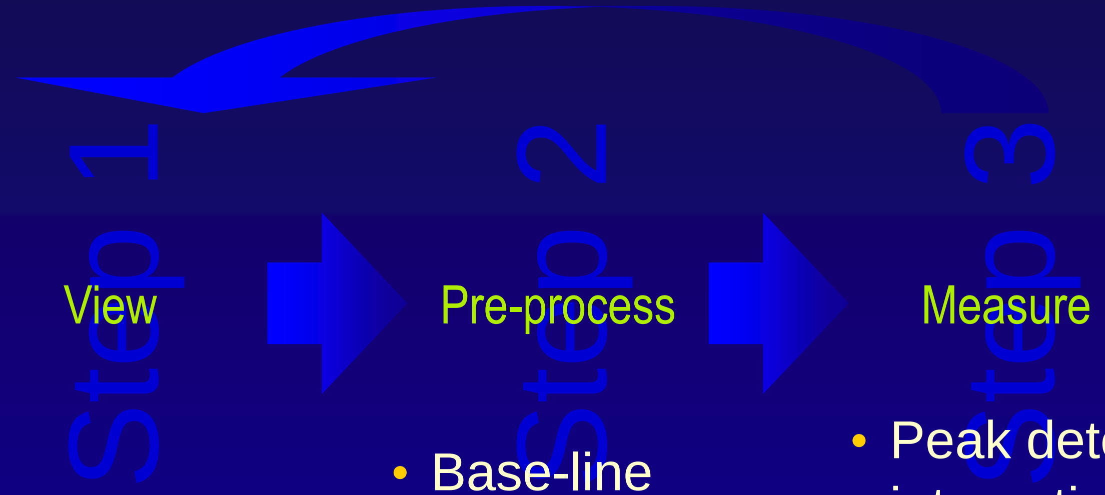
An example of PARAFAC





A.E. Sinha, J.L. Hope, B.J. Prazen, C.G. Fraga, E.J. Nilsson, R.E. Synovec, J. Chromatogr. A, 1056 (2004) 145 - 154

Third step: measure



View

Pre-process

Measure

- Folding
- Phasing

- Base-line correction
- Noise filtering
- Spike filtering
- Alignment
- … etc.

- Peak detection / integration
- Calibration
- Deconvolution
- Pattern recognition
- Class separation

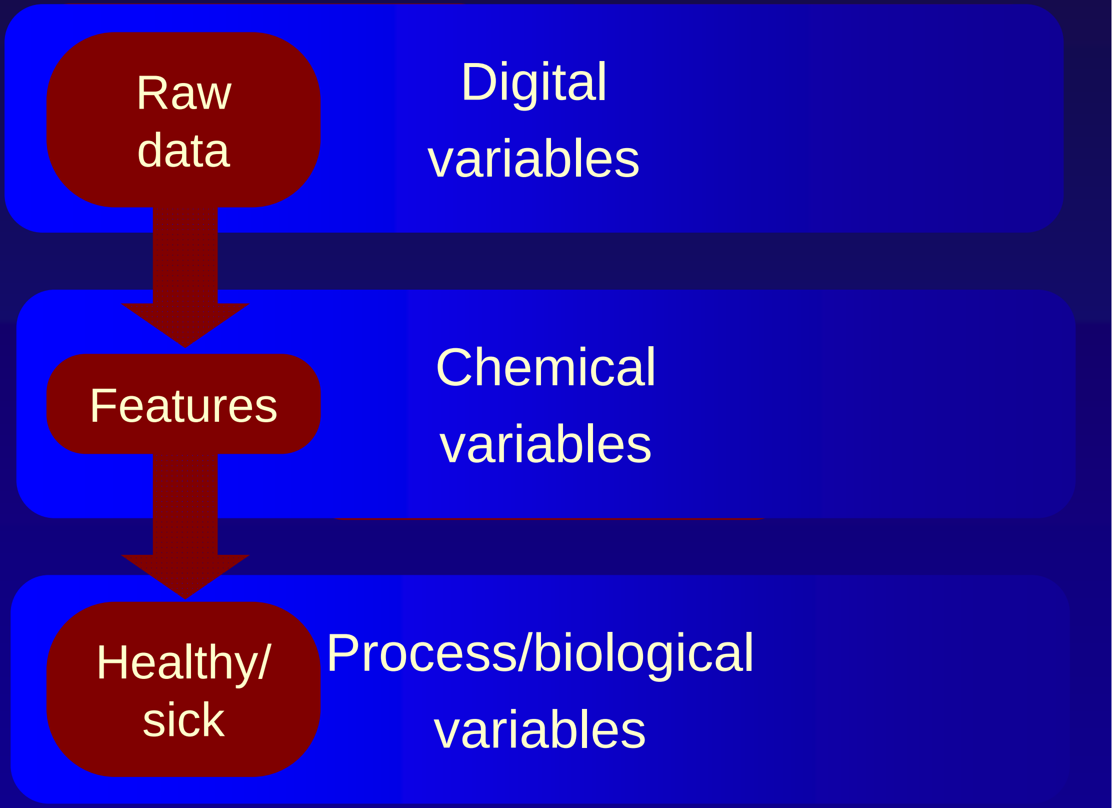
Pattern recognition as a variable reduction

step 1 Obtain chromatogram(s)

step 2 Peak detection

step 3 Pattern recognition

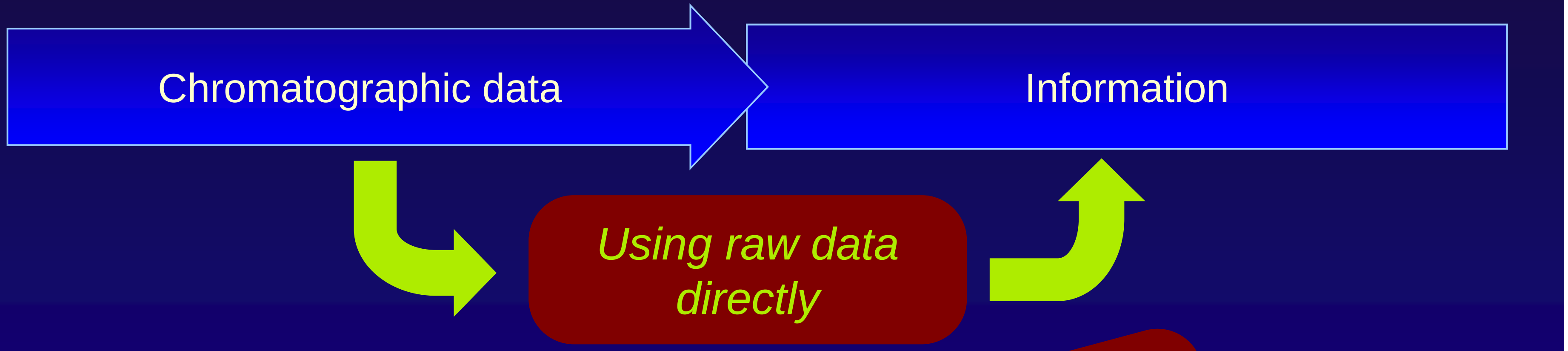
step 4 Unknown sample characterized



From digital variables (chromatogram) to process variables

Is peak detection necessary?

Pattern recognition



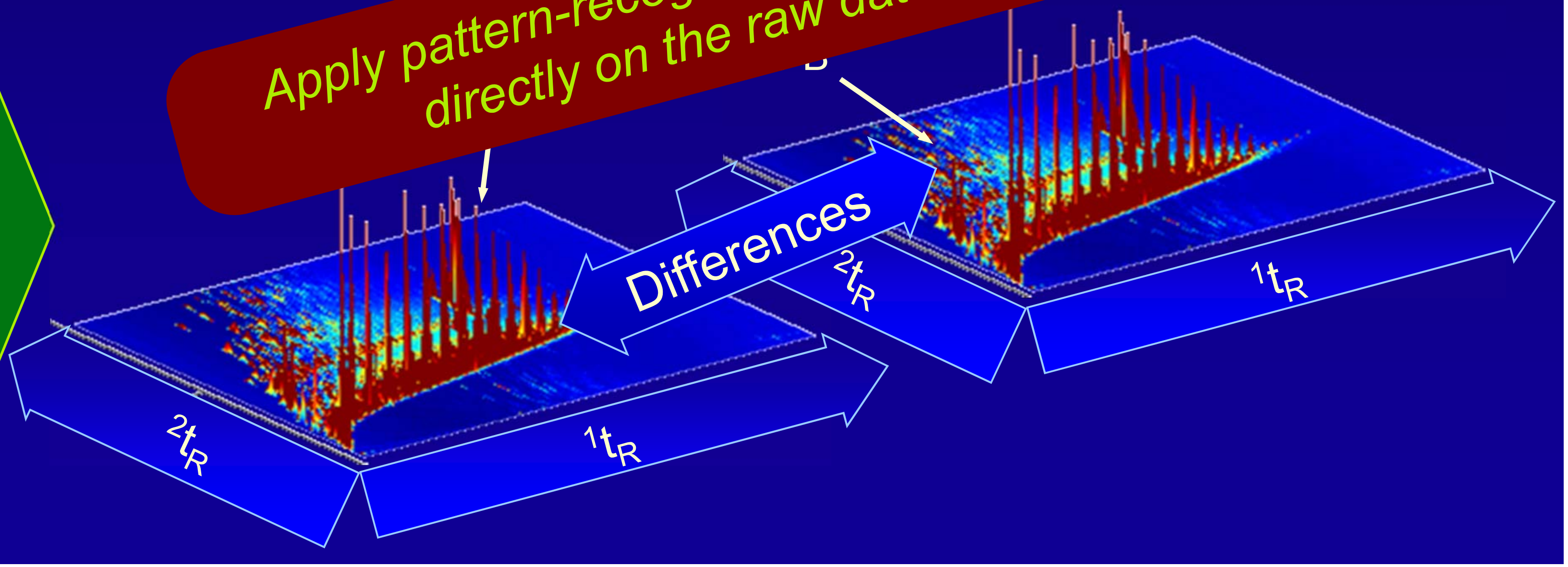
Chromatographic data

Information

Using raw data directly

… for example:

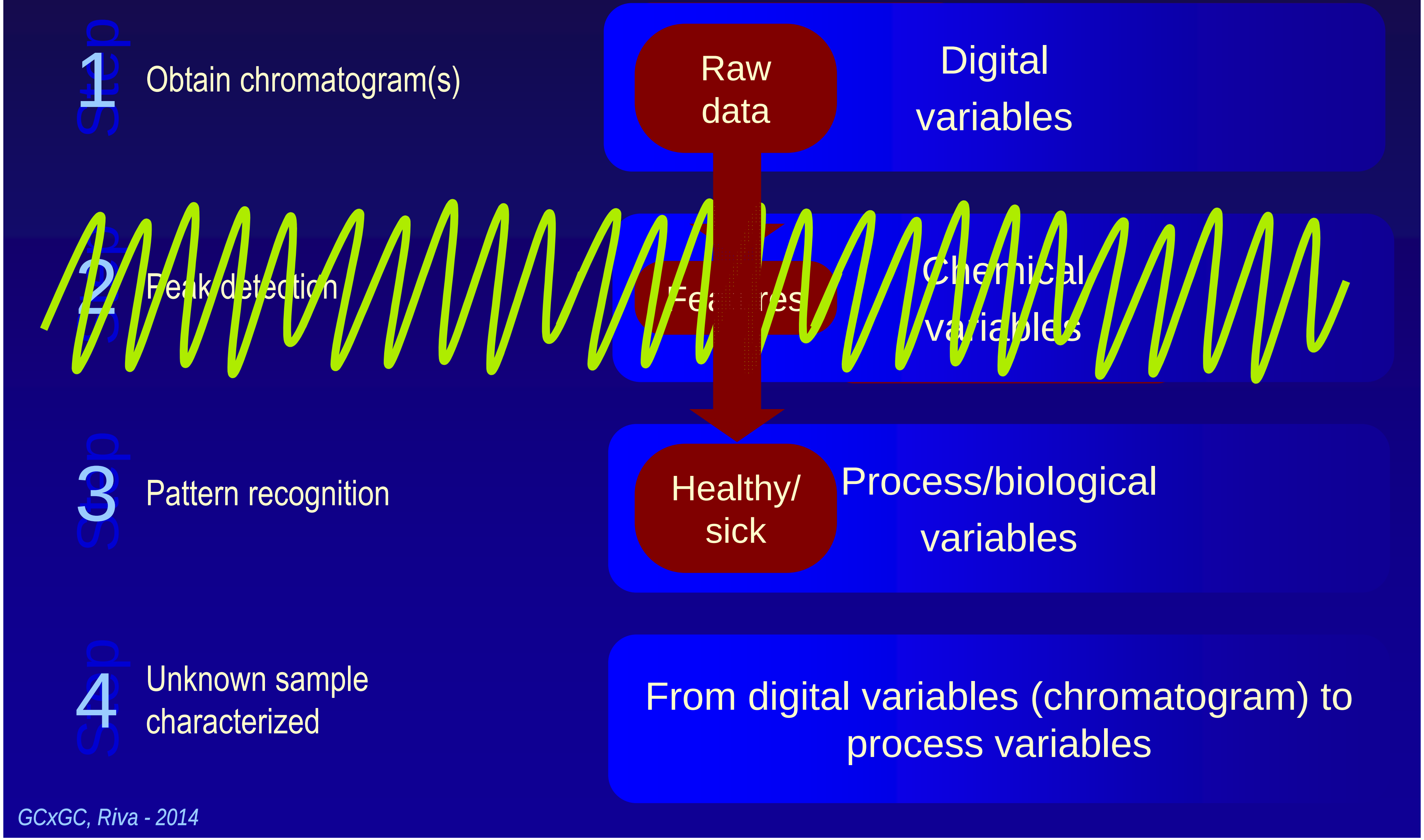
Analysis of flame accelerants using GCxGC



Apply pattern-recognition techniques directly on the raw data

Pattern recognition as a variable reduction

Pattern recognition



step 1 Obtain chromatogram(s)

Raw data

Digital variables

step 2 Peak detection

Features

Chemical variables

step 3 Pattern recognition

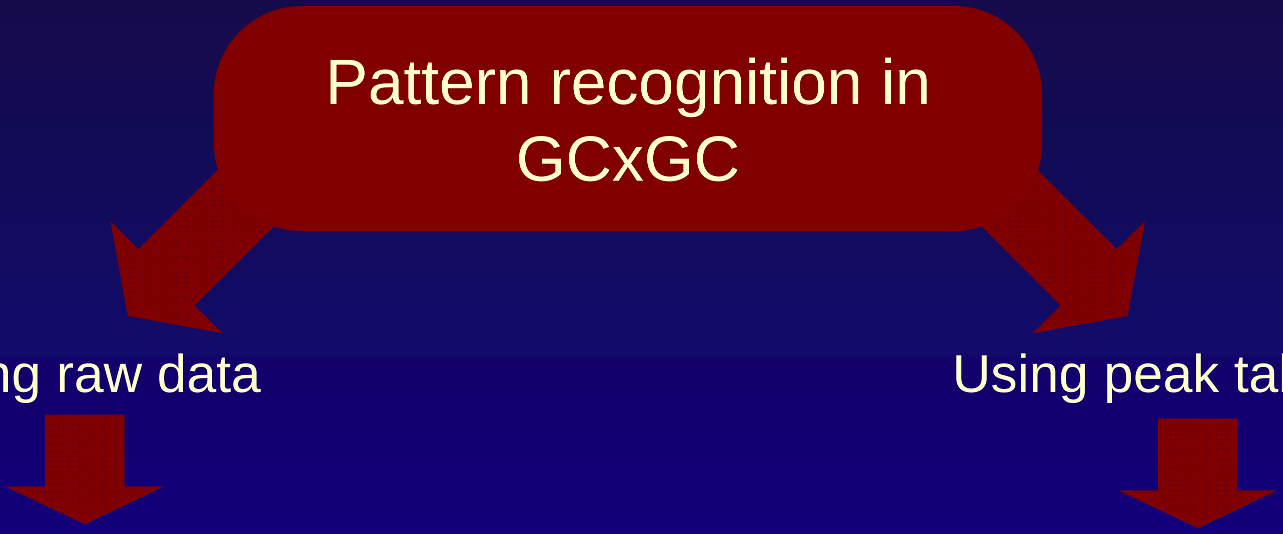
Healthy/ sick

Process/biological variables

step 4 Unknown sample characterized

From digital variables (chromatogram) to process variables

Pattern recognition in GCxGC. The options



Pattern recognition in GCxGC

Using raw data

Using peak tables

- Alignment is critical
- Less chance to miss important compounds
- Normally done with the unfolded (raw) data, but not always (e.g. N-PLS)

- Alignment not important, but peak tracking is essential (normally MS should be present)
- Chance to miss important compounds (close to the S/N)
- (Truly) 1D method

Pattern recognition in GCxGC. Supervised methods.

Pattern recognition

In supervised pattern recognition of GCxGC, a tremendous reduction of variables is performed (form millions to a few tens/hundreds)

Any method will be prone to overfitting

Any variable pre-reduction (e.g. using Fisher ratios) should be done within a cross-validation loop

Otherwise the results will be optimistic (a method that seems to work, when in fact it only works for that data)

Pattern recognition in GCxGC. Example of a wrong strategy

Pattern recognition

Objective: discovering metabolites responsible for cancer tumor

step 1

Obtain GCxGC chromatograms for sick (50) and healthy (50)

Probably align GCxGC data

step 2

Variable selection: Fisher ratio on the raw data

Keep only variables with a FR>threshold

step 3

Supervised pattern recognition: PLS-DA to separate sick from healthy

Use only the selected variables from step 2

step 4

Consider the coefficients from PLS-DA as indicators of potential metabolites

Hurrah! I have a collection of interesting metabolites!

????

Aren't you Overfitting?

No, I've been cross-validating the PLS-DA

… but the variable pre-selection has been done with the full data set!!

Pattern recognition in GCxGC. Example of a ~~wrong~~ correct strategy

Pattern recognition

Objective: discovering metabolites responsible for cancer tumor

Step 1

Obtain GCxGC chromatograms for sick (50) and healthy (50)

Step 2

Variable selection: Fisher ratio on the raw data

Step 3

Supervised pattern recognition: PLS-DA to separate sick from healthy

Step 4

Consider the coefficients from PLS-DA as indicators of potential metabolites

Probably align GCxGC data

Keep only variables with a FR>threshold

Use only the selected variables from step 2

Hurrah! I have a collection of interesting metabolites!

????

Aren’t you Overfitting?

No, I’ve been cross-validating the variable seletion and the PLS-DA

Pattern recognition in GCxGC. Example of a correct strategy

“model” = “variable pre-selection (Fisher ratio) + PLS-DA”

Step 1

Divide the original data set in k subsets (randomly)

Step 2

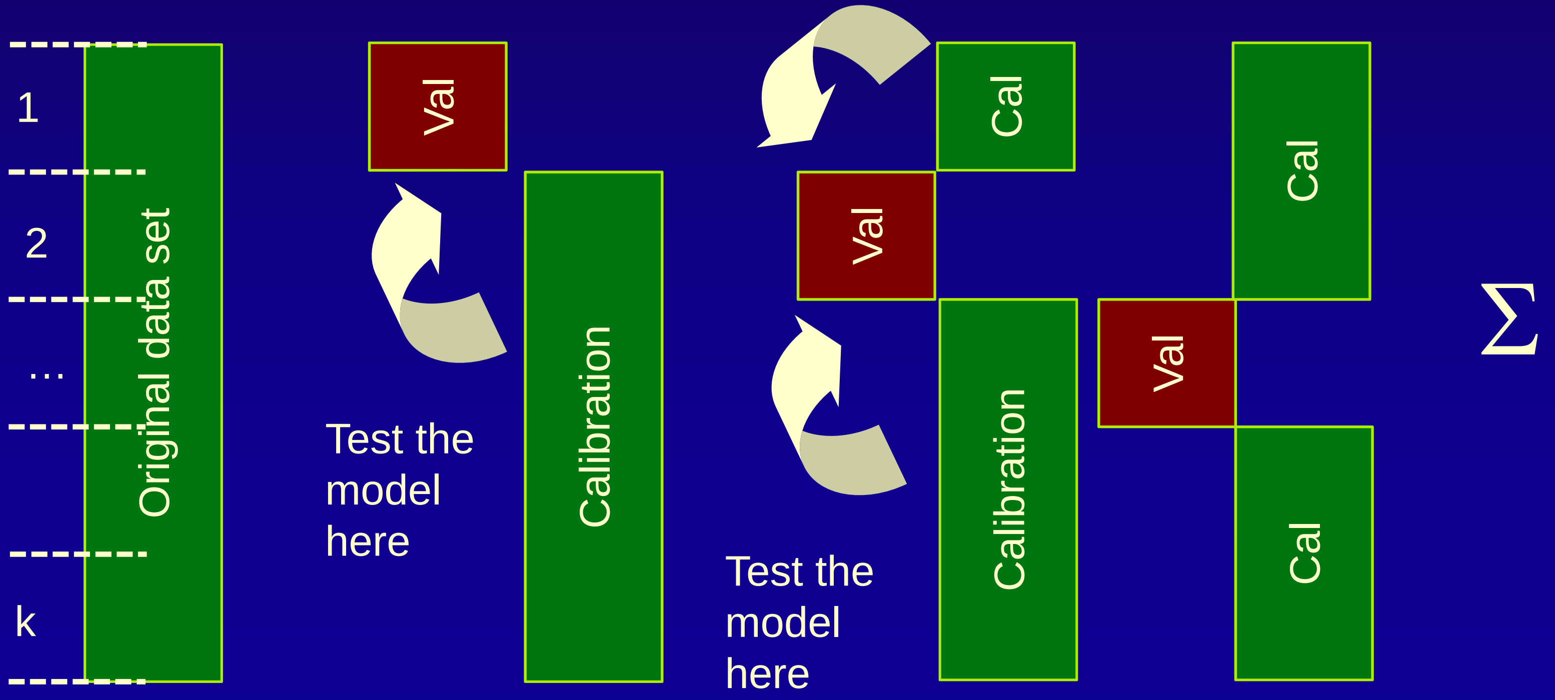
Withdraw subset 1, fit the model with the rest, and test the model with subset 1

Step 3

Do the same for the other k sections

Step 4

Sum up the error of the model in all validation sets



Conclusions

Multivariate methods

- Deconvolution: normally done with the unfolded data (less problems with between-modulation alignment)
- Deconvolution: problem to establish the number of compounds (normally done in a manual way)
- Two ways for pattern recognition: with raw data (normally preferred) or with peak table.
- Careful with validation of supervised pattern recognition. Variable pre-selection should be included in the validation loop.