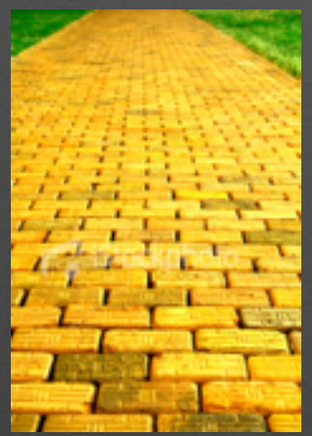
Advance XCMS Data Processing

H. Paul Benton

Reminder of what we're trying to do

Peak Detection

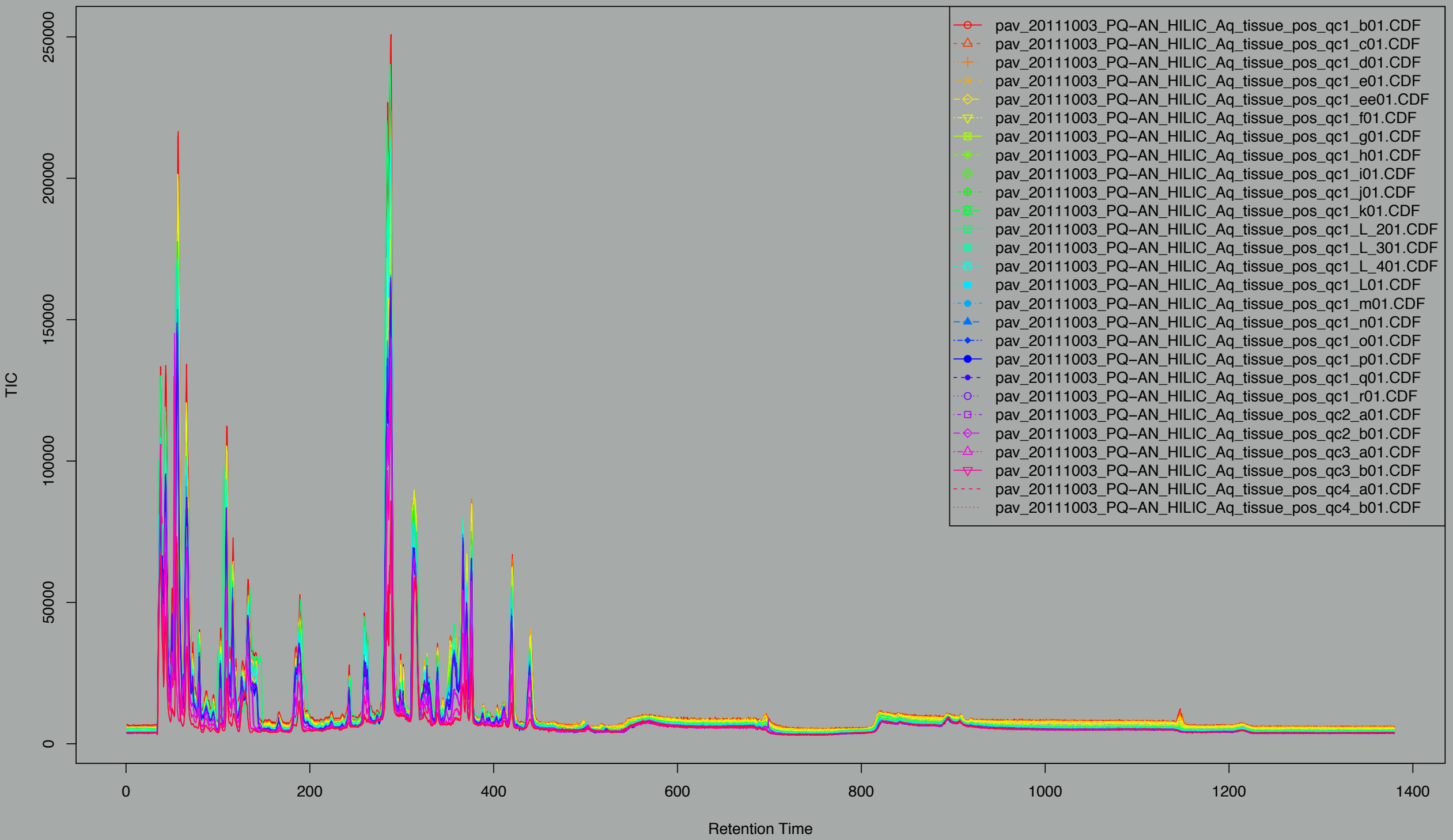


Grouping
Groups similar Peaks
across replicates

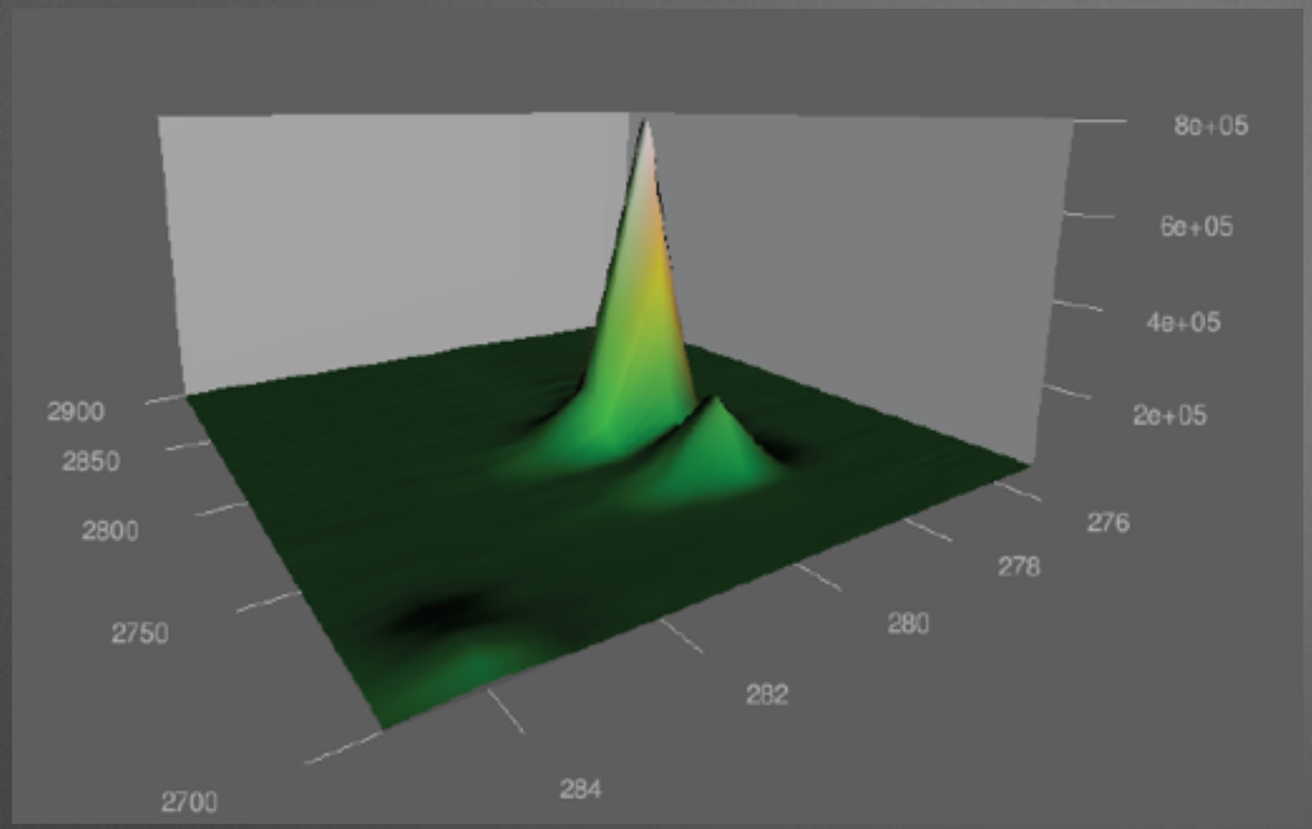
Retention Time
Alignment

Statistical Analysis
of Classes

Total Ion Chromatograms

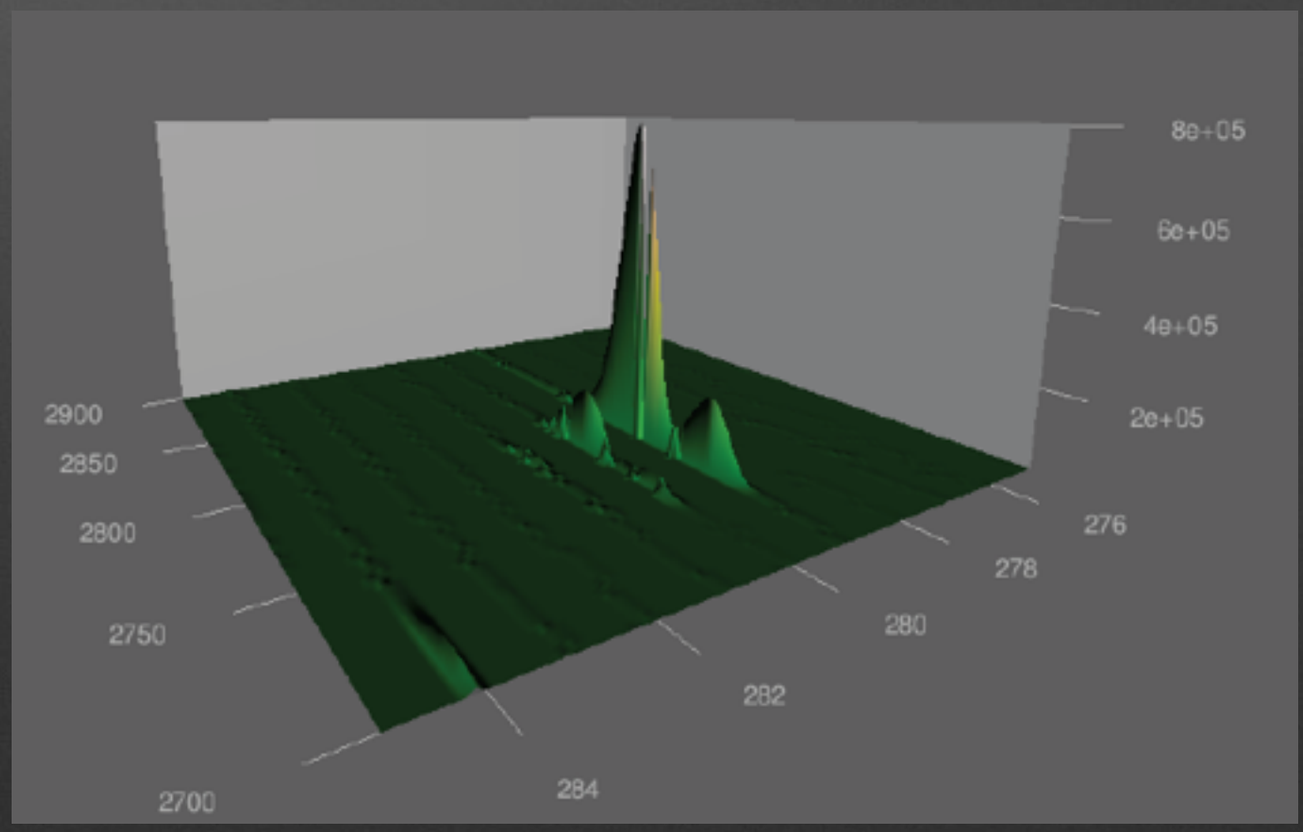


Parameters Matter !

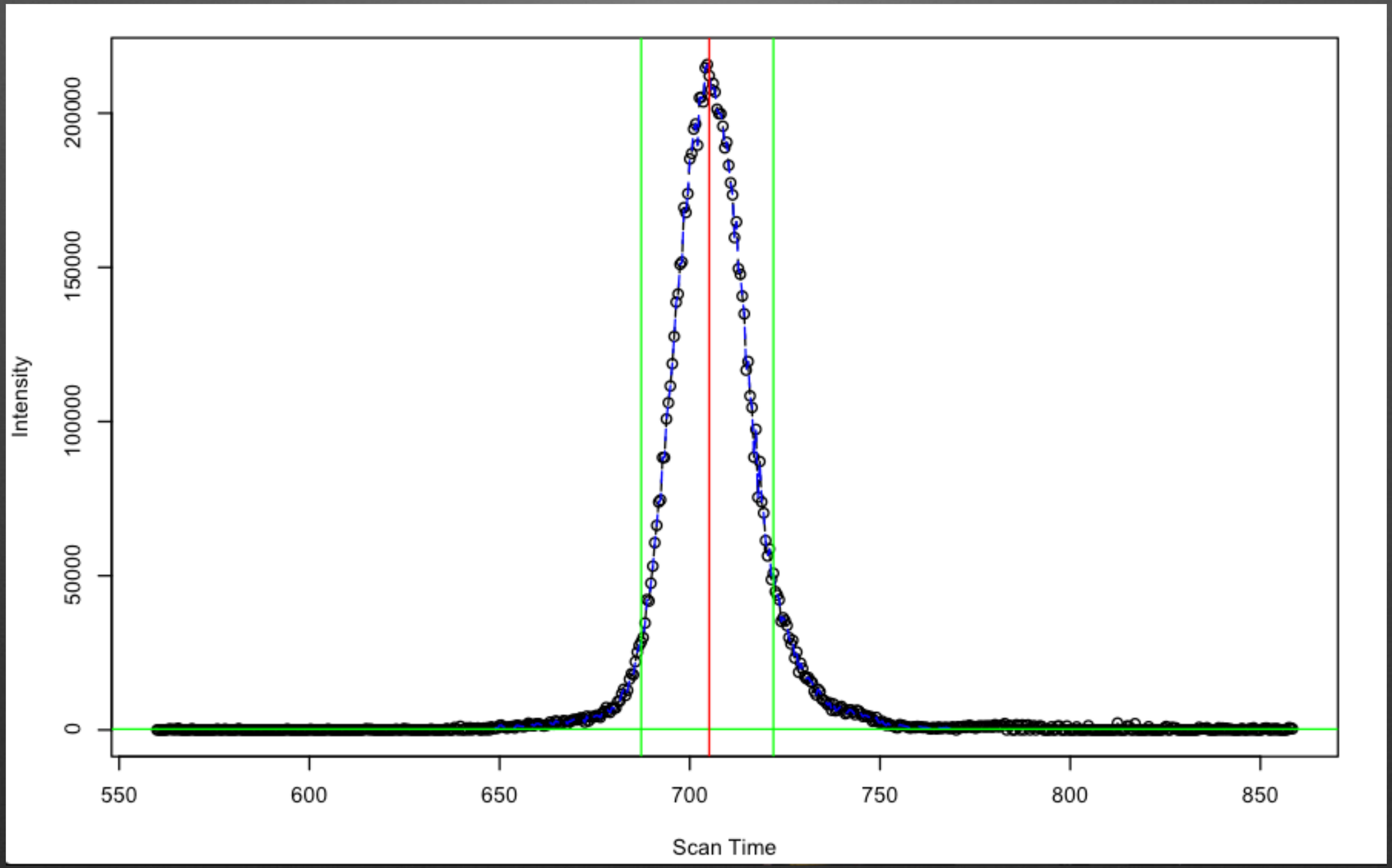


step = 1

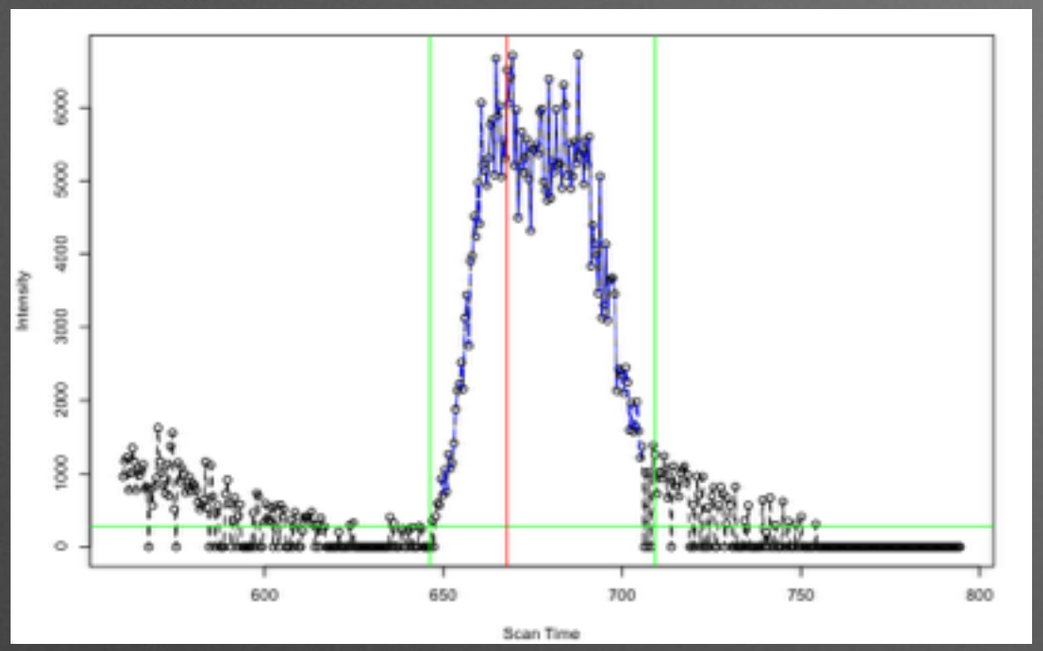
step = 0.1

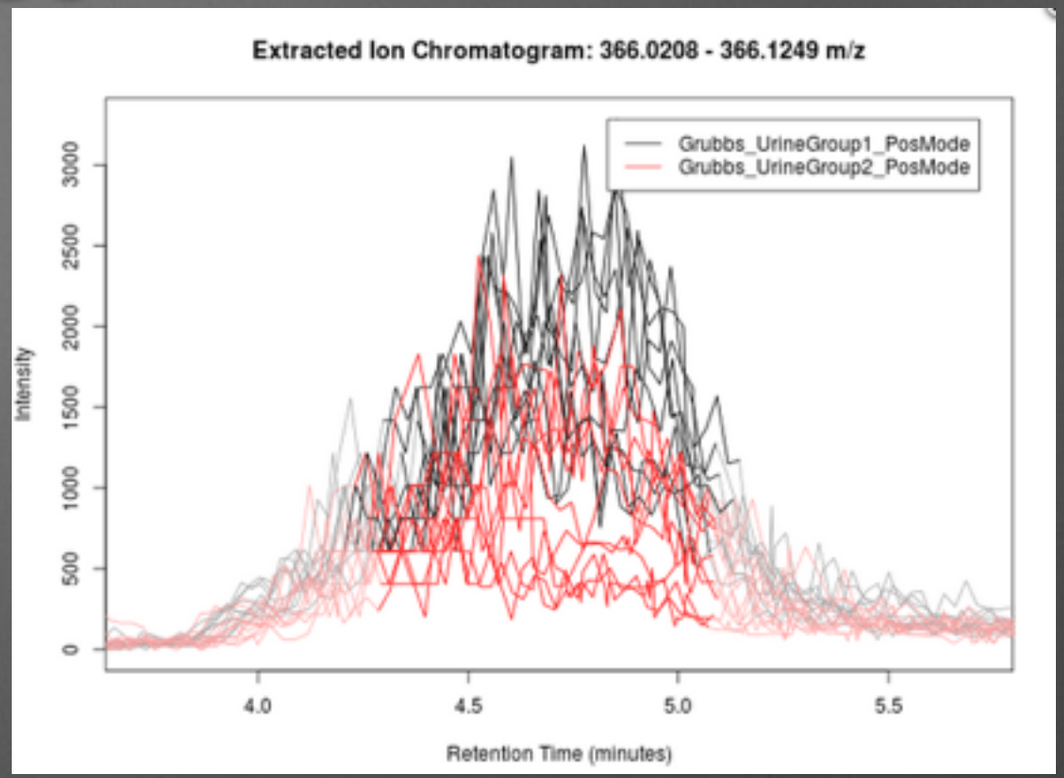


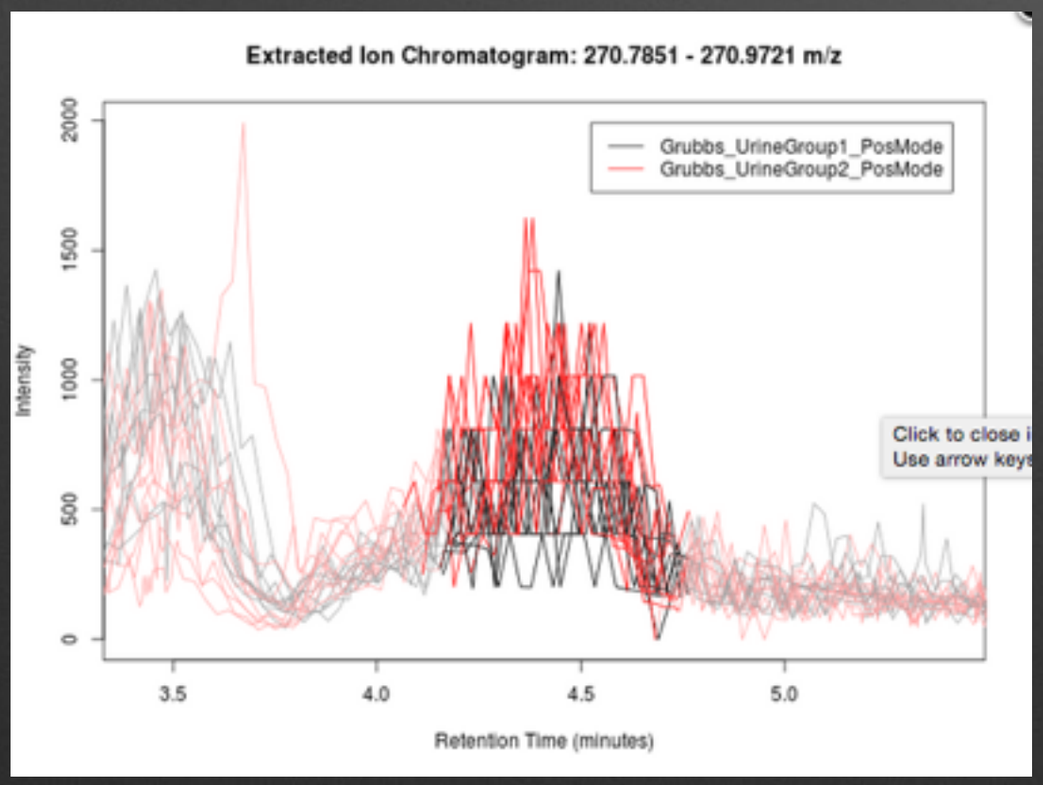
Peak Detection… Easy !

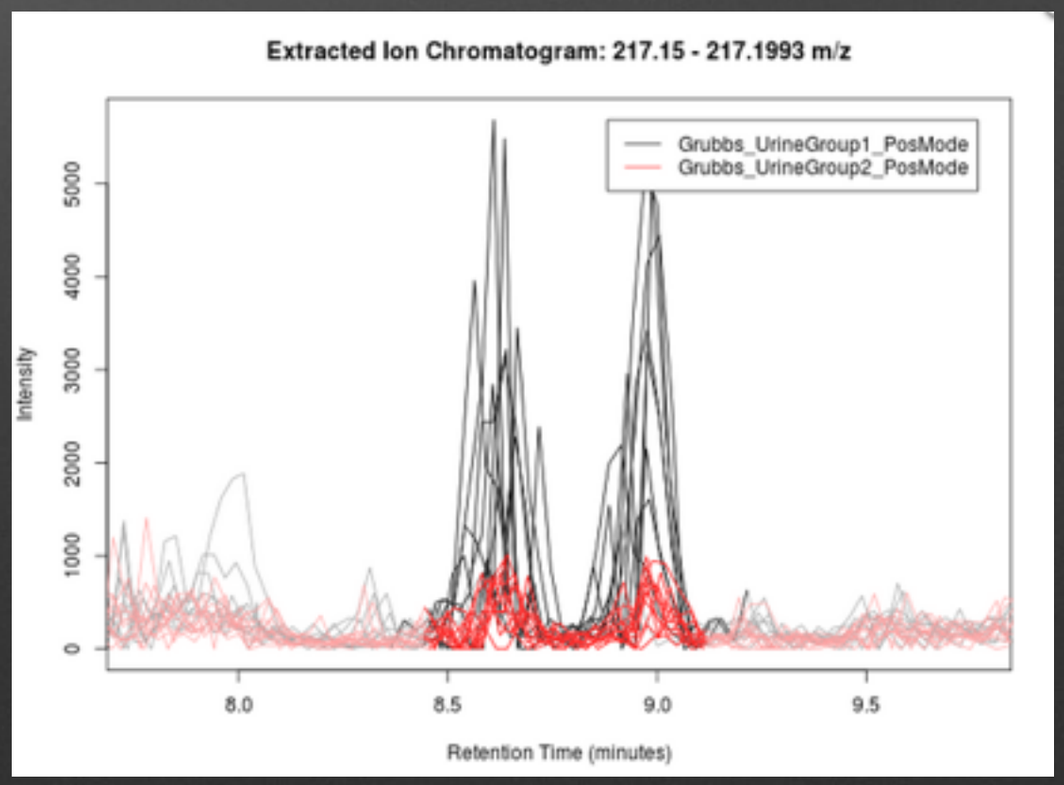


Right?


Intensity
0
1000
2000
3000
4000
5000
6000
600
650
700
750
800
Scan Time


Extracted Ion Chromatogram: 366.0208 - 366.1249 m/z
Grubbs_UrineGroup1_PosMode
Grubbs_UrineGroup2_PosMode
Intensity
0
500
1000
1500
2000
2500
3000
4.0
4.5
5.0
5.5
Retention Time (minutes)


Extracted Ion Chromatogram: 270.7851 - 270.9721 m/z
Grubbs_UrineGroup1_PosMode
Grubbs_UrineGroup2_PosMode
Intensity
0
500
1000
1500
2000
3.5
4.0
4.5
5.0
Retention Time (minutes)


Extracted Ion Chromatogram: 217.15 - 217.1993 m/z
Grubbs_UrineGroup1_PosMode
Grubbs_UrineGroup2_PosMode
Intensity
0
1000
2000
3000
4000
5000
8.0
8.5
9.0
9.5
Retention Time (minutes)

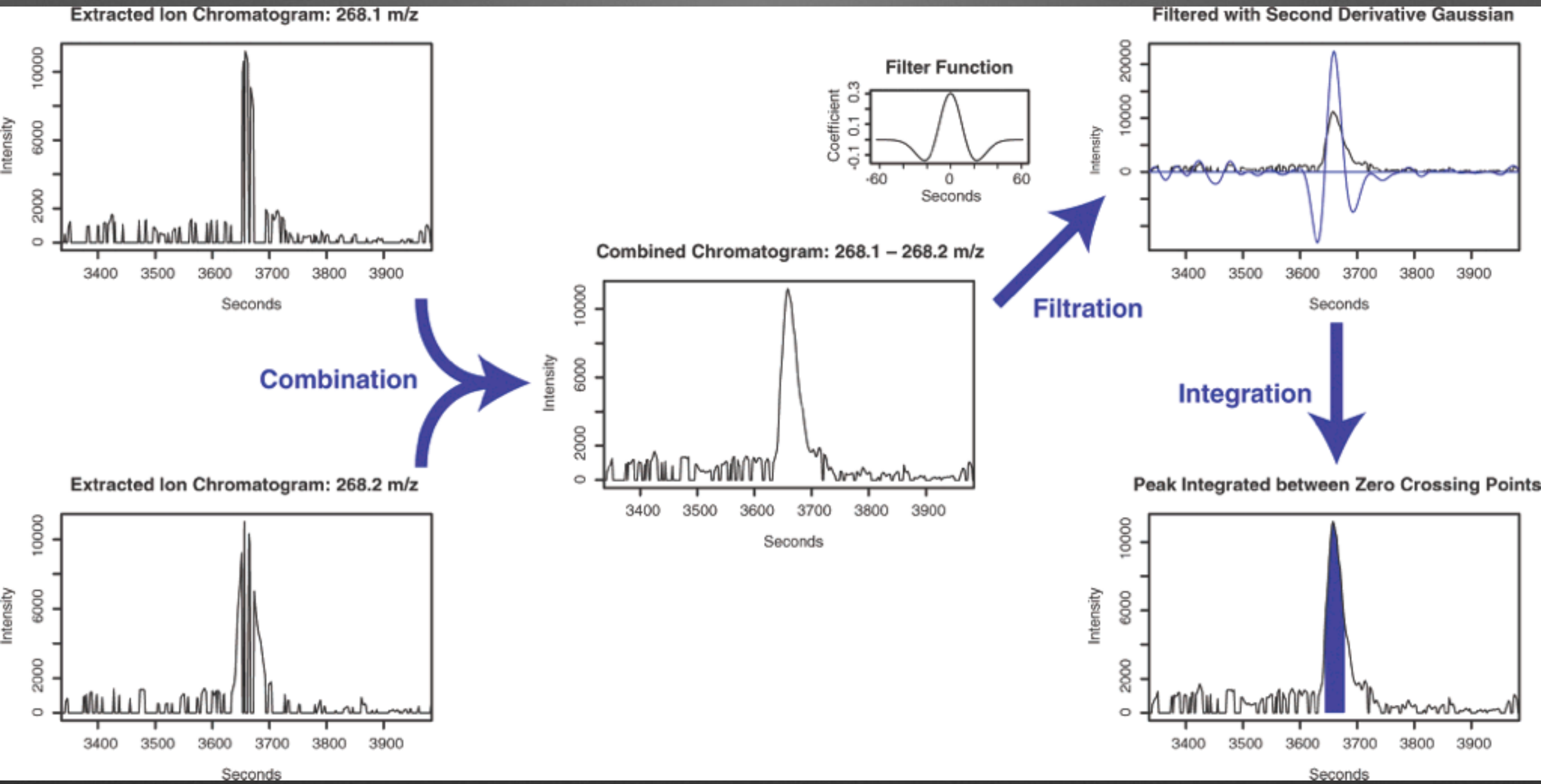
Peak detection

- Data comes in two types in MS : centroid & profile
- Generally high resolution or low resolution ~ high mass accuracy or low mass accuracy
- Two main choices in XCMS
  - MatchedFilter - profile low res
  - CentWave - centroid high res

Hat fitting

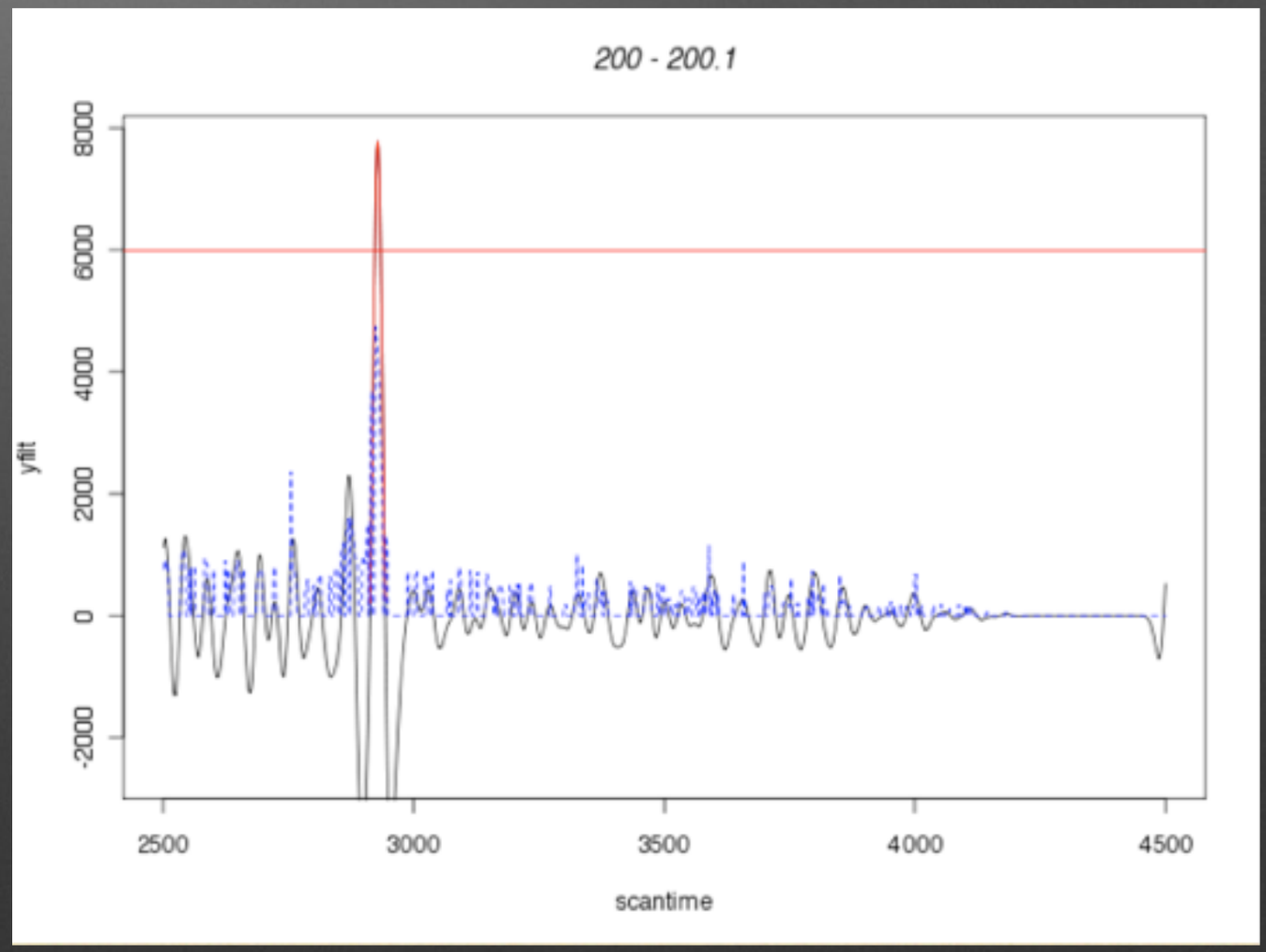
- Different hat for different heads (& faces apparently)
- A hat has to fit well so it must be sized

MatchedFilter



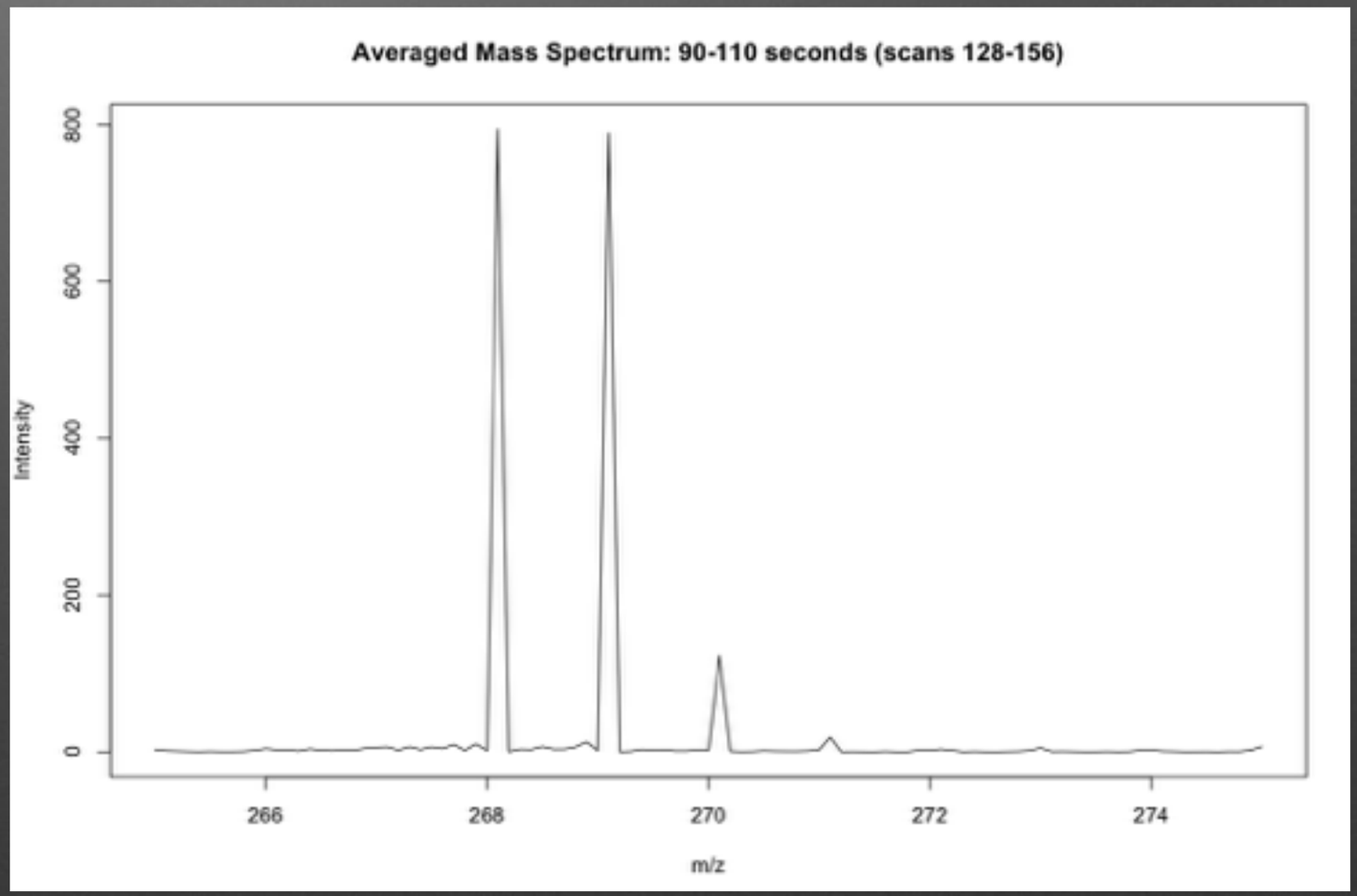
MatchedFilter : wearing hats

- Bin of each X m/z
- Apply a filter function to the data
- Any peak above a s/n ratio is selected
- Peak is selected to filter baseline



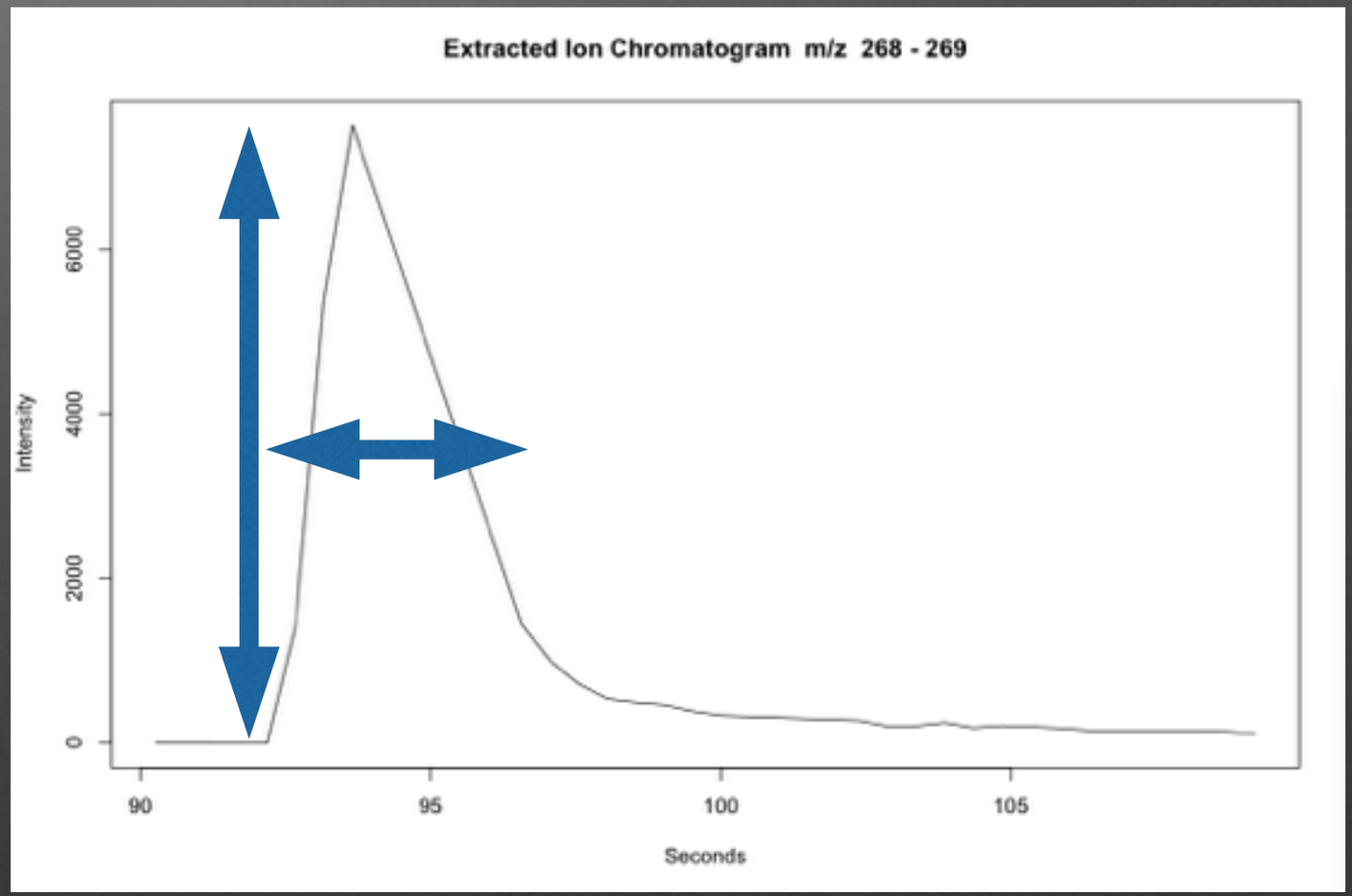
Matched Filter : Sizing the hats

- Profile Data - profMethod
  - binlinBase - profile data
  - bin - centroid data
- profStep - Bin Size
- Peak Width - FWHM



Matched Filter : Sizing the hats

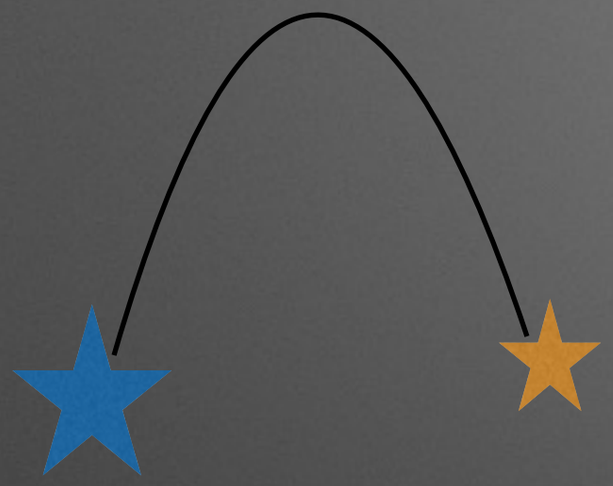
- Profile Data - profMethod
  - binlinBase - profile data
  - bin - centroid data
- profStep - Bin Size
- Peak Width - FWHM



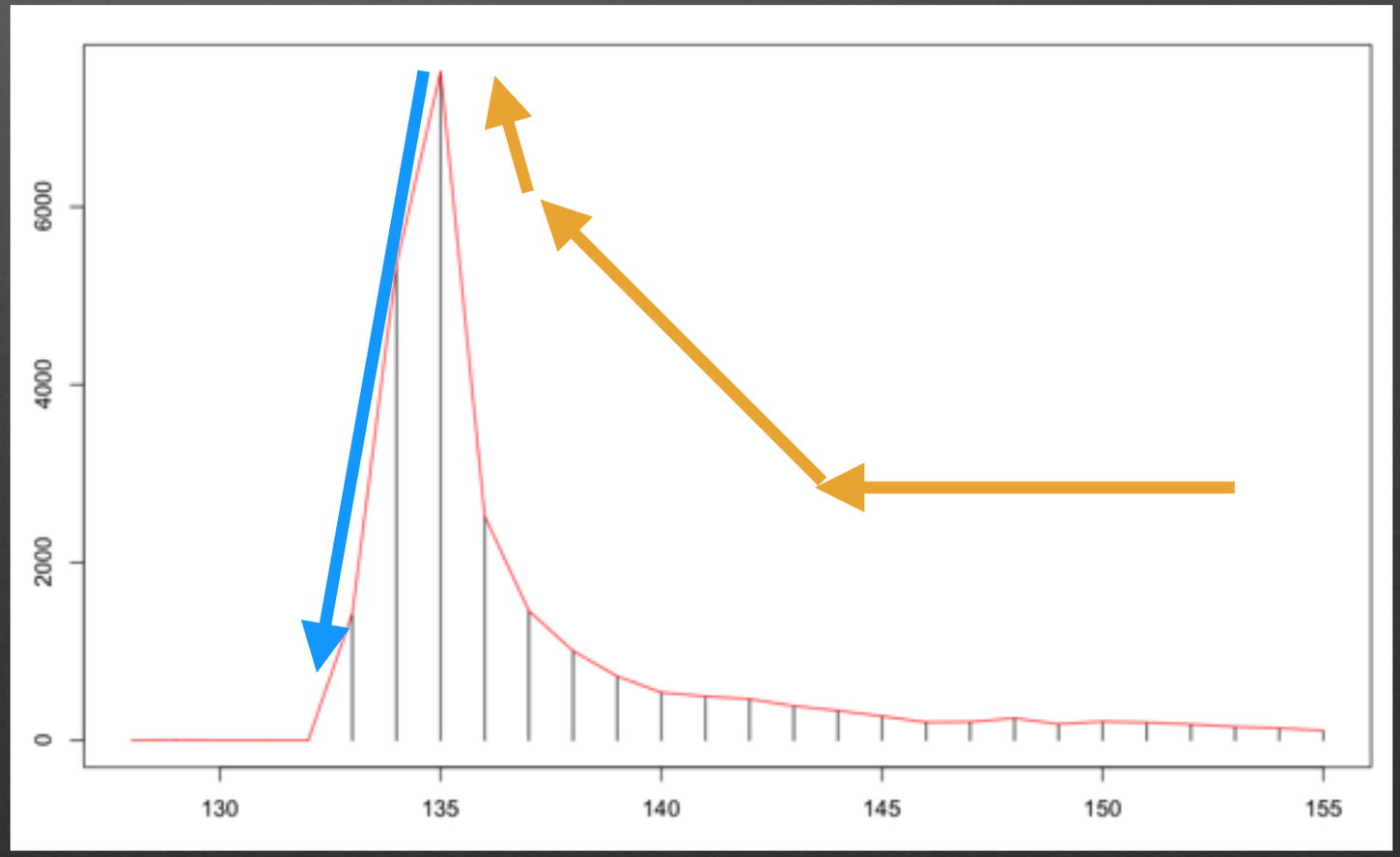
Missiles are like ions!

Kalman filtering

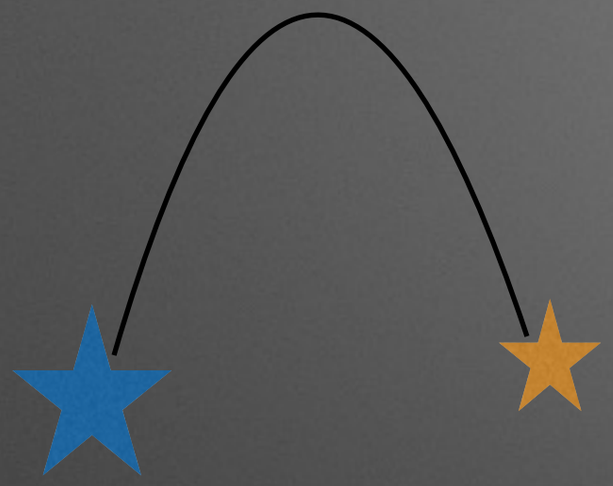
Tracking Missiles is
like tracking LC-MS traces



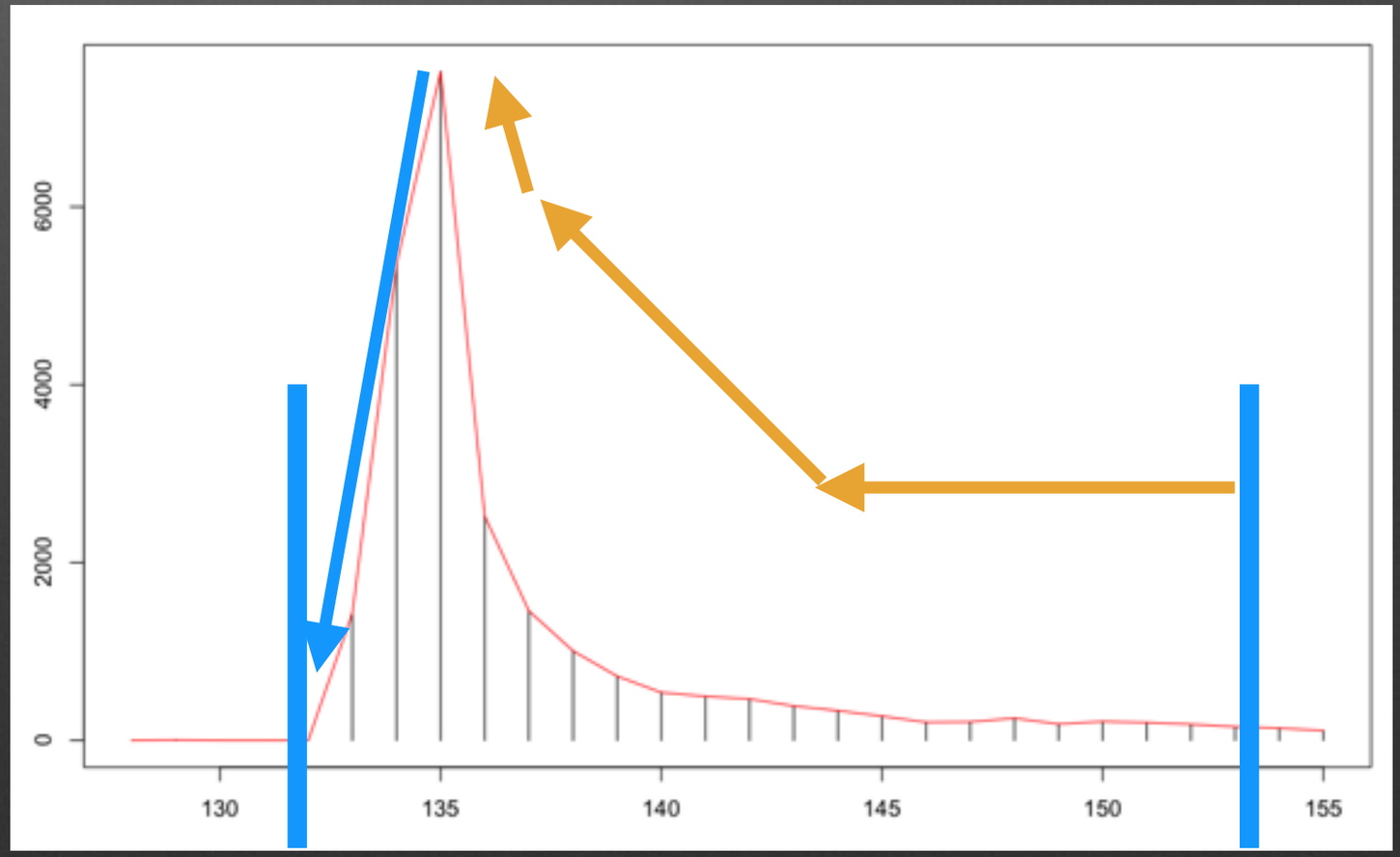
Trace backward along the trace
This will define the area of the ‘bin’



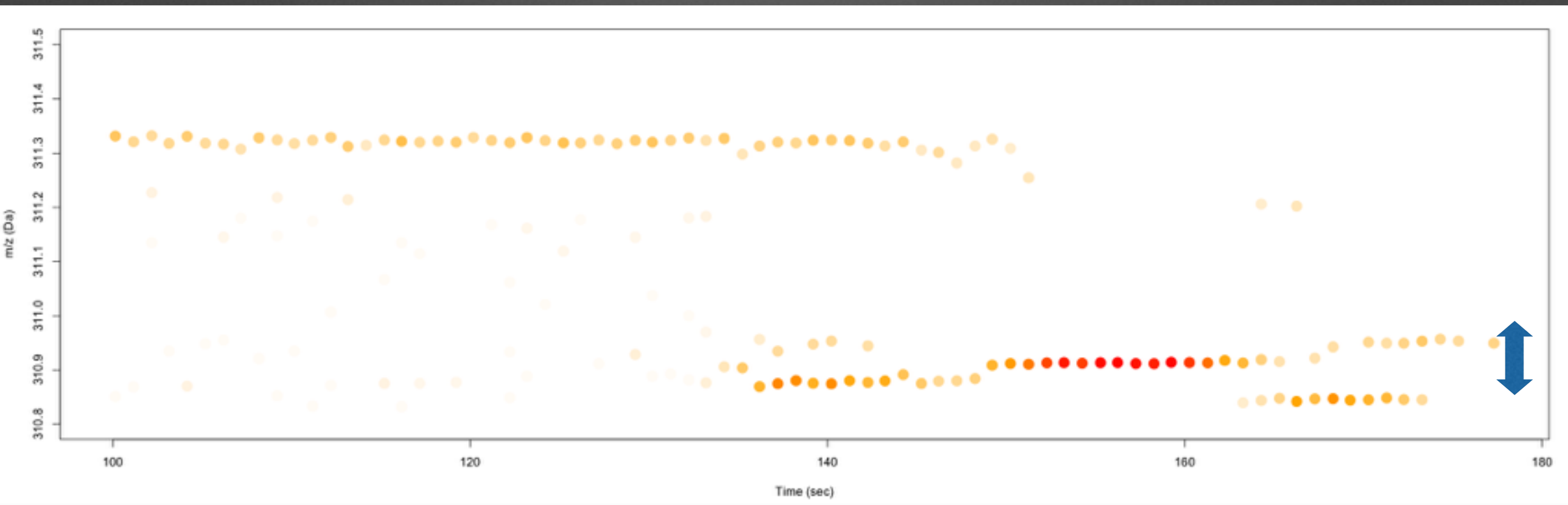
Tracking Missiles is
like tracking LC-MS traces



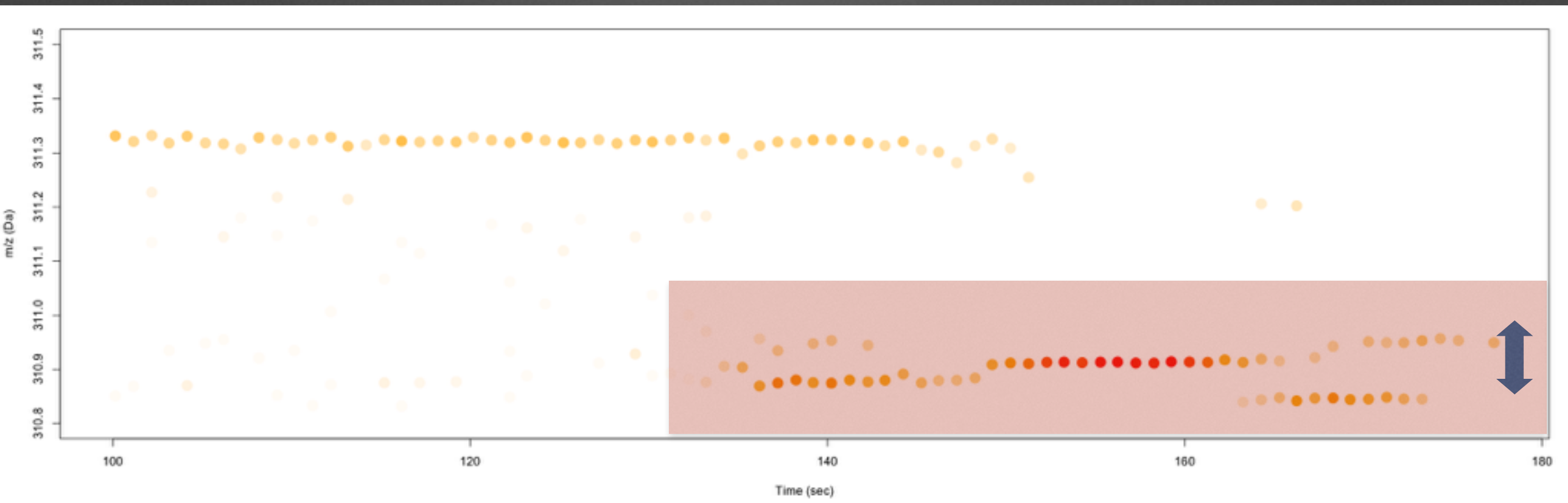
Trace backward along the trace
This will define the area of the ‘bin’



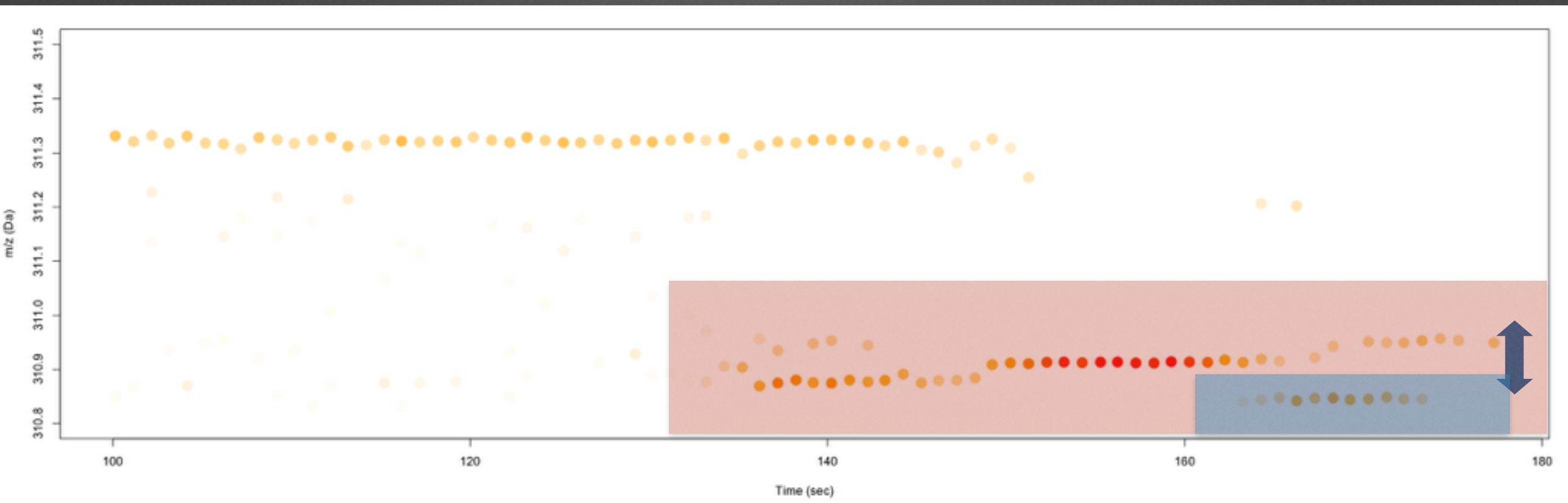
CentWave - ppm



CentWave - ppm

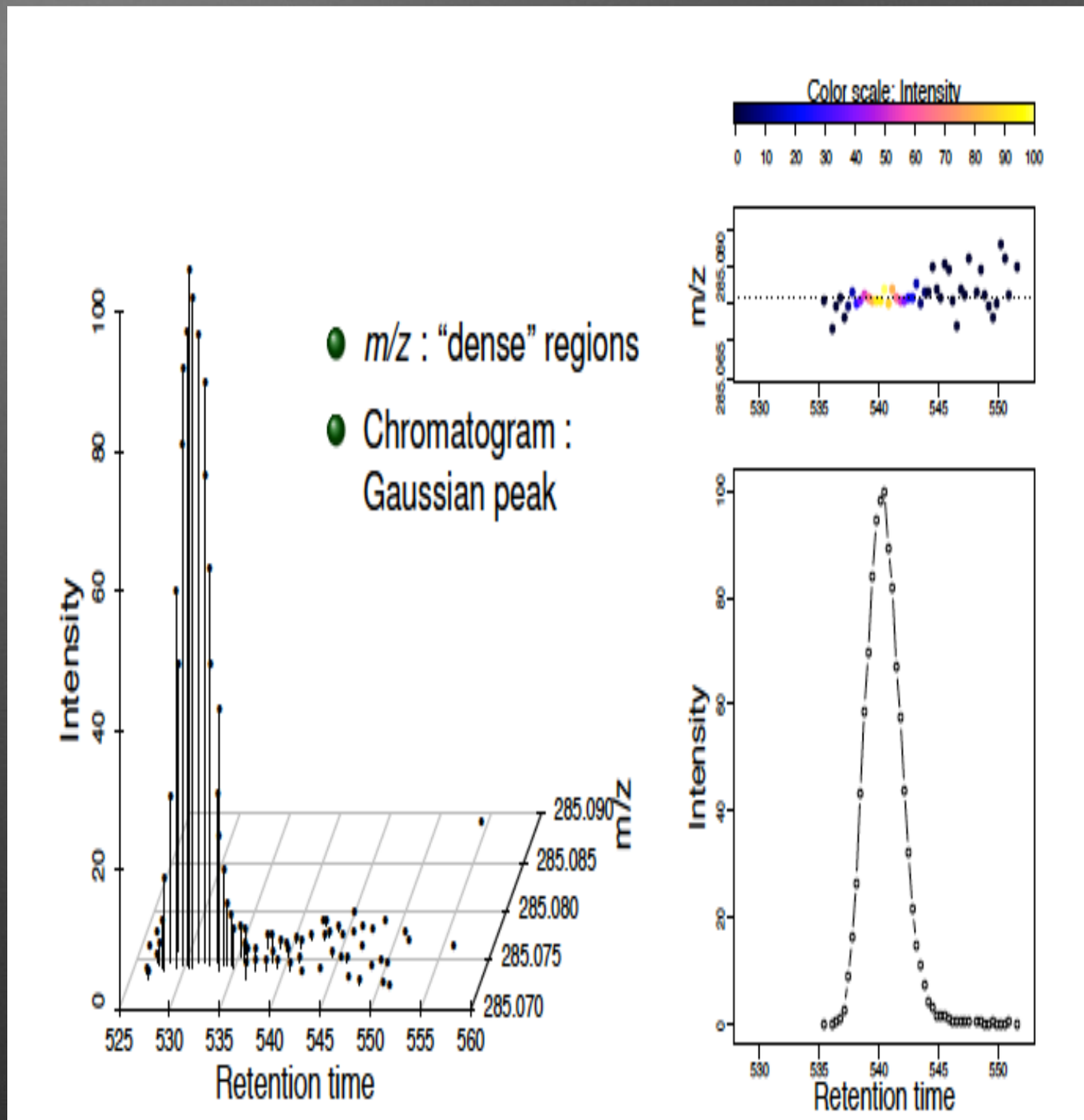


CentWave - ppm

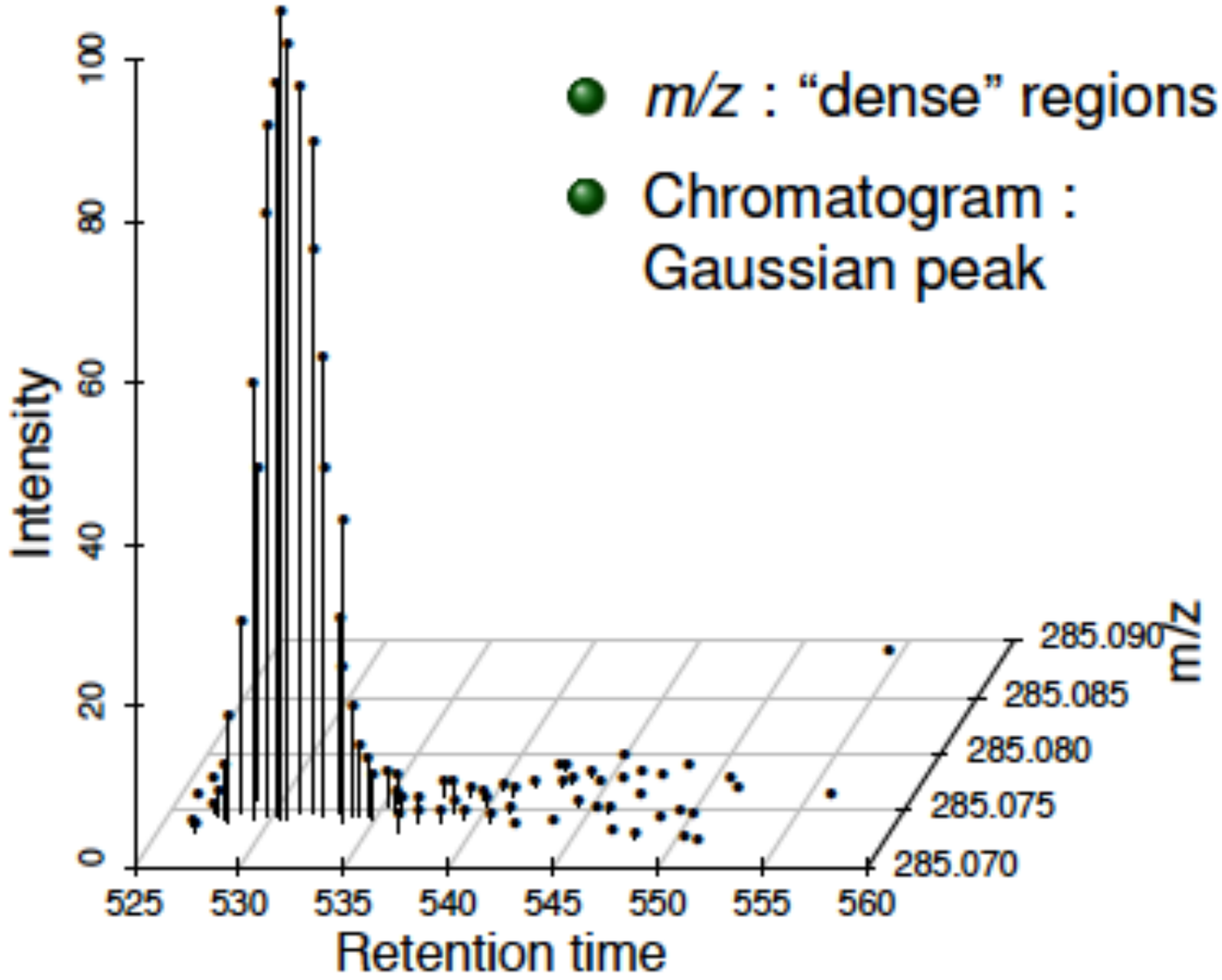


CentWave 2

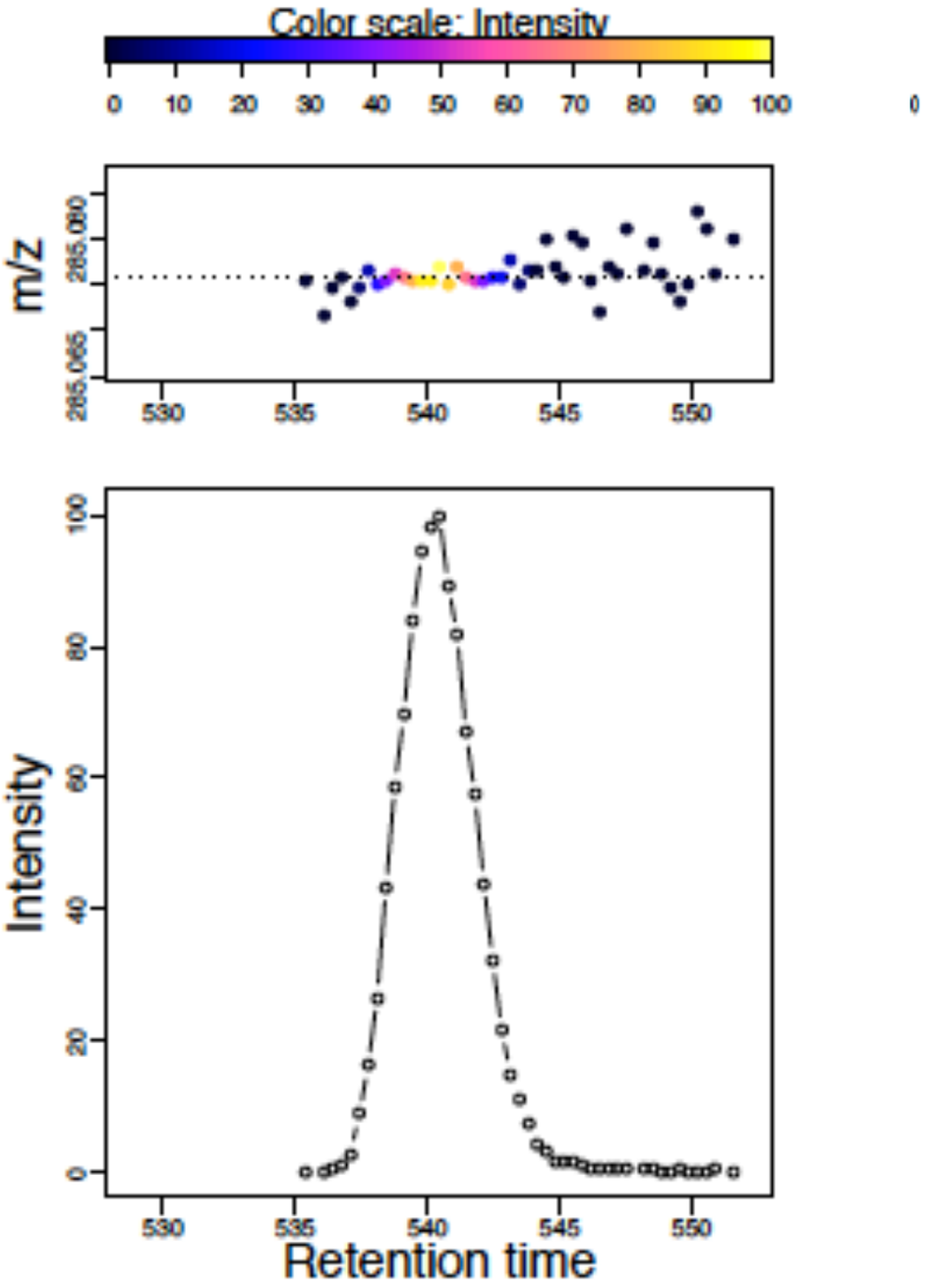
- Regions of Interest (ROI)
  - Found using Kalman Filter
  - Often over estimates



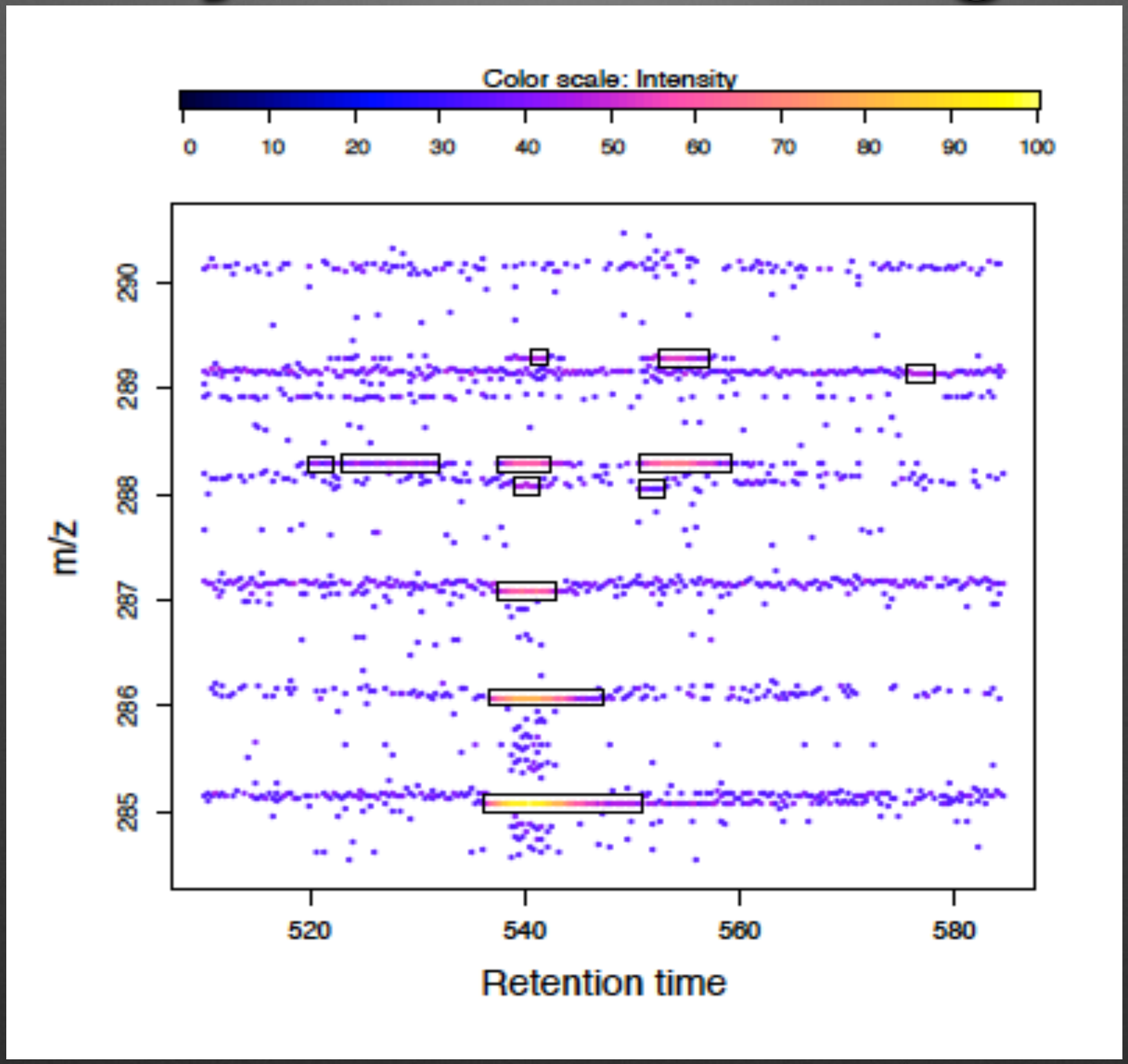
CentWave 2



- *m/z* : “dense” regions
- Chromatogram : Gaussian peak

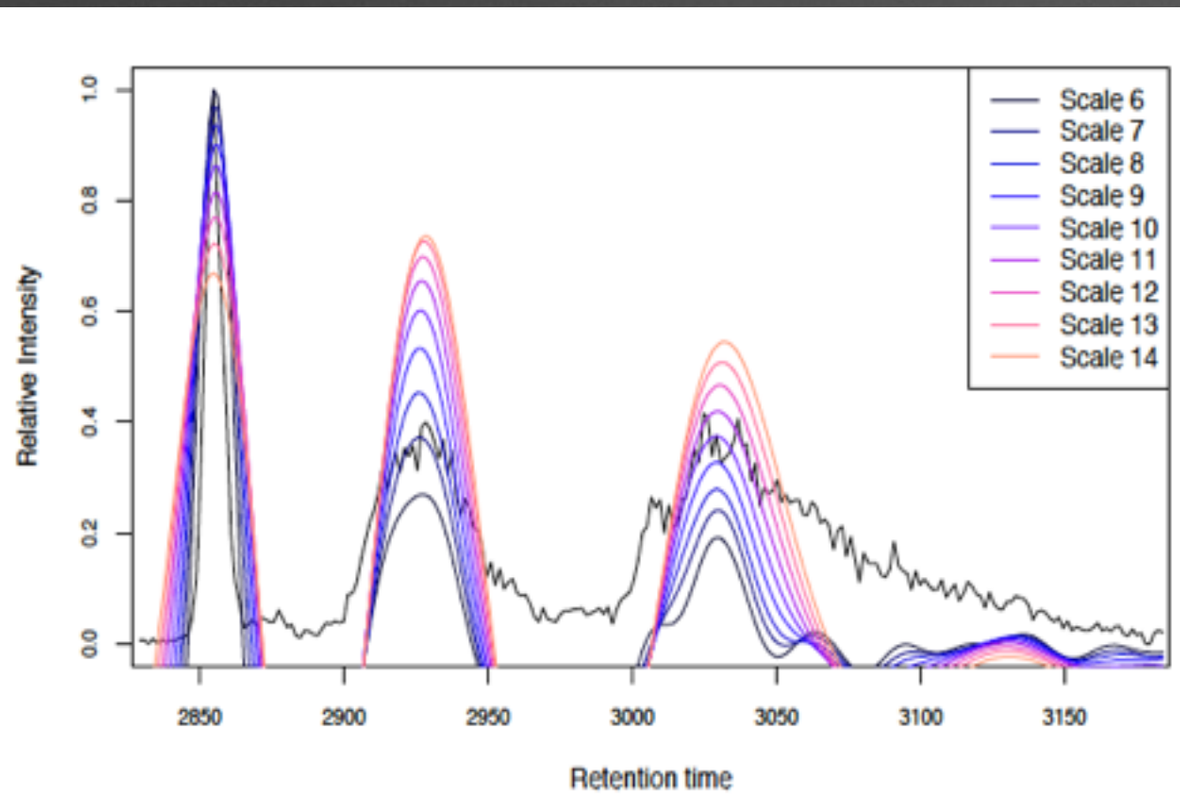


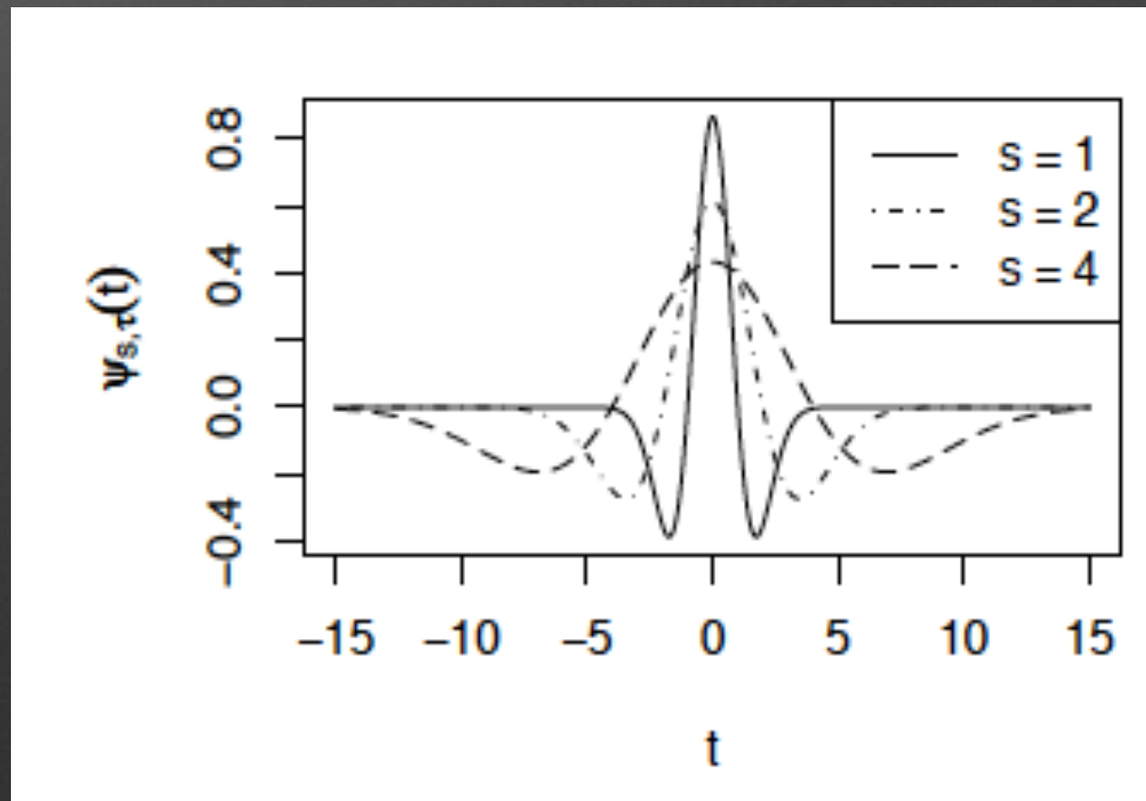
Dynamic Binning



Find and integrate the peak

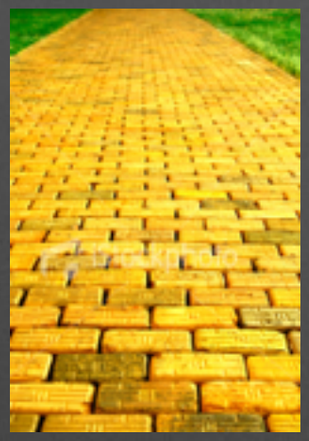
- Wavelet formation are then used over the ROIs to find the peak
- Several passes of wavelets are used until the correction ‘fit’ is found (mexican hat wavelets)





General Principals

Peak Detection



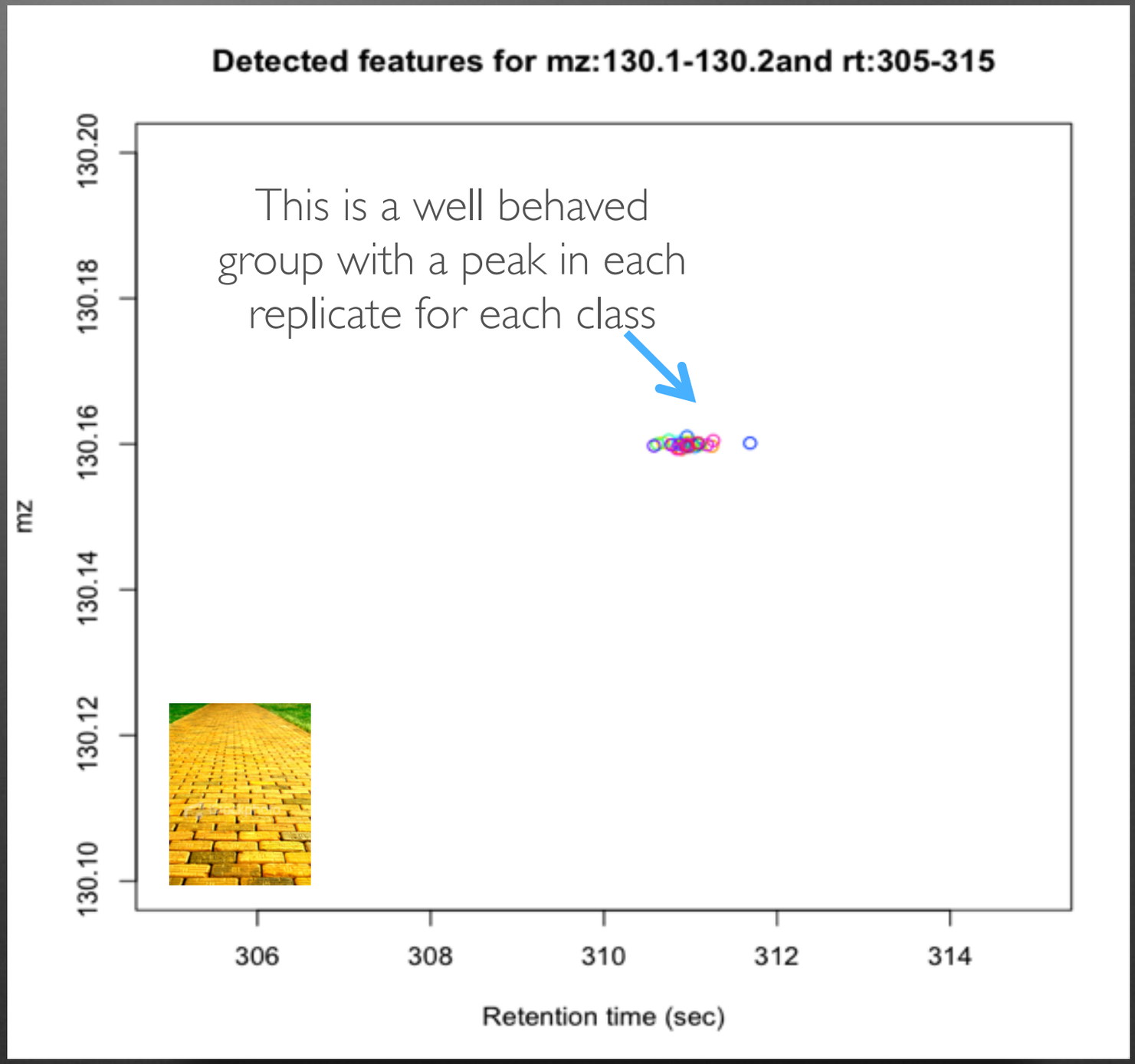
Grouping
Groups similar Peaks across replicates

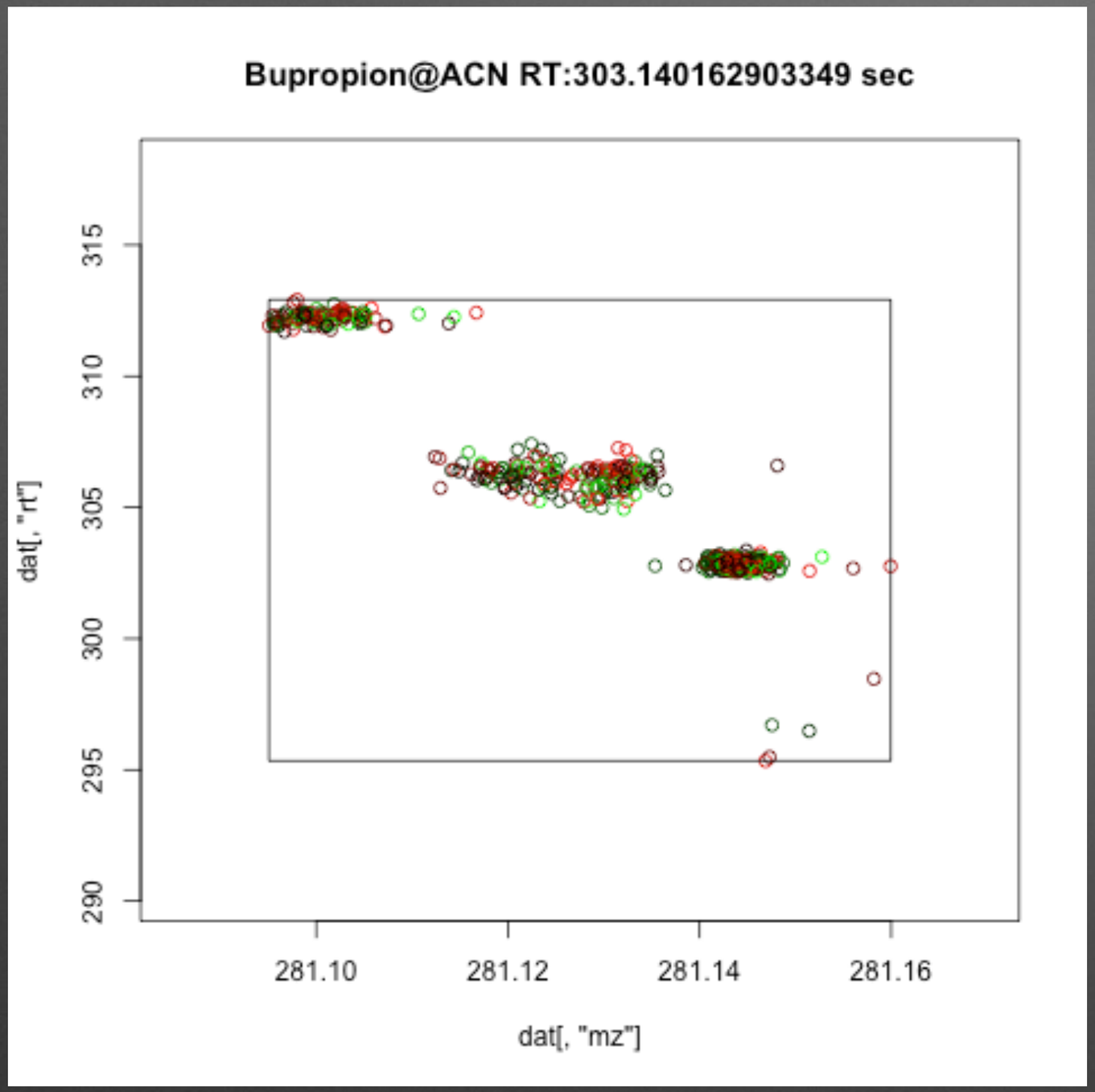
Retention Time Alignment

Statistical Analysis of Classes

Grouping

- First time using all of the files
- Looks for closely clustered/ dense peaks across multiple files.
  - Once peaks are grouped they’re know as a group or feature

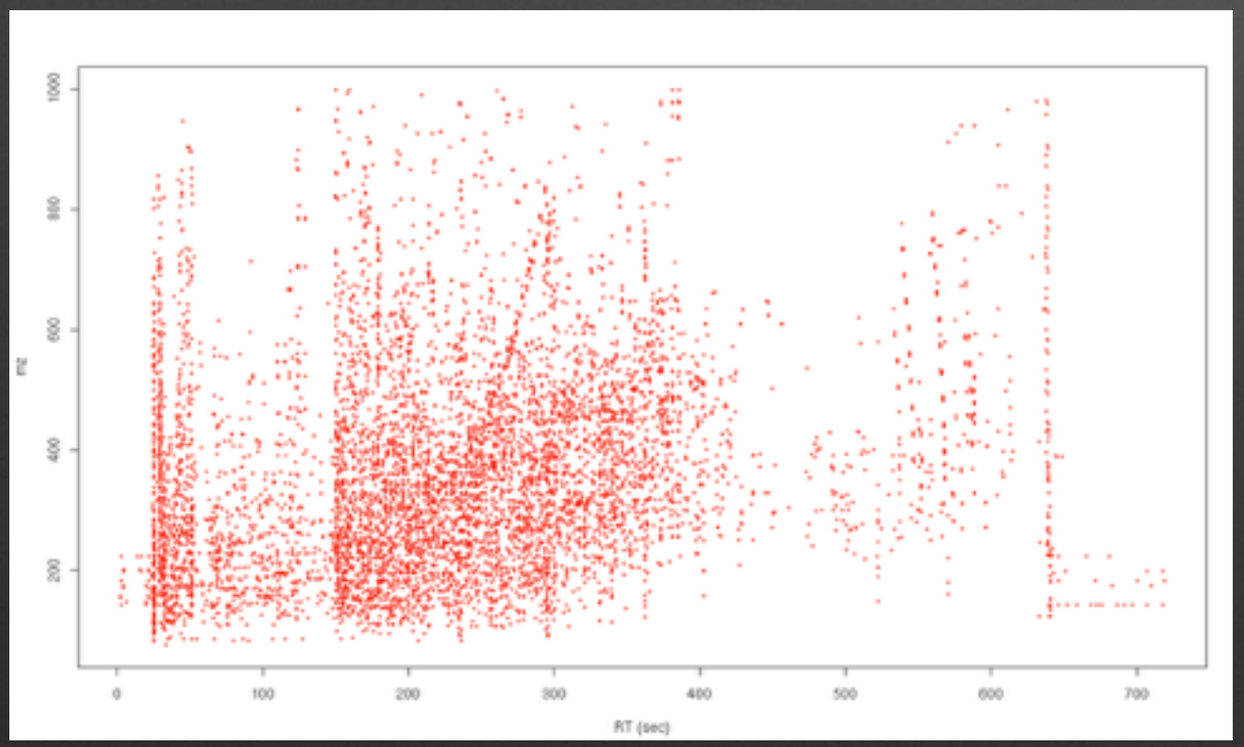


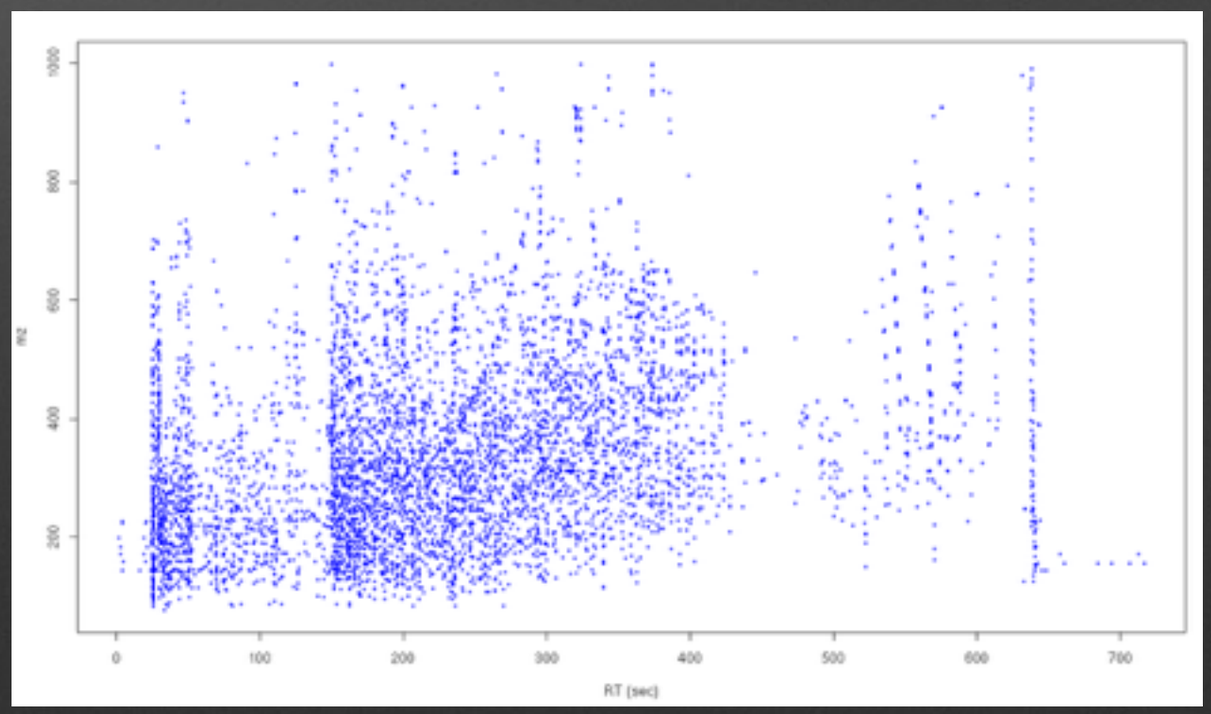


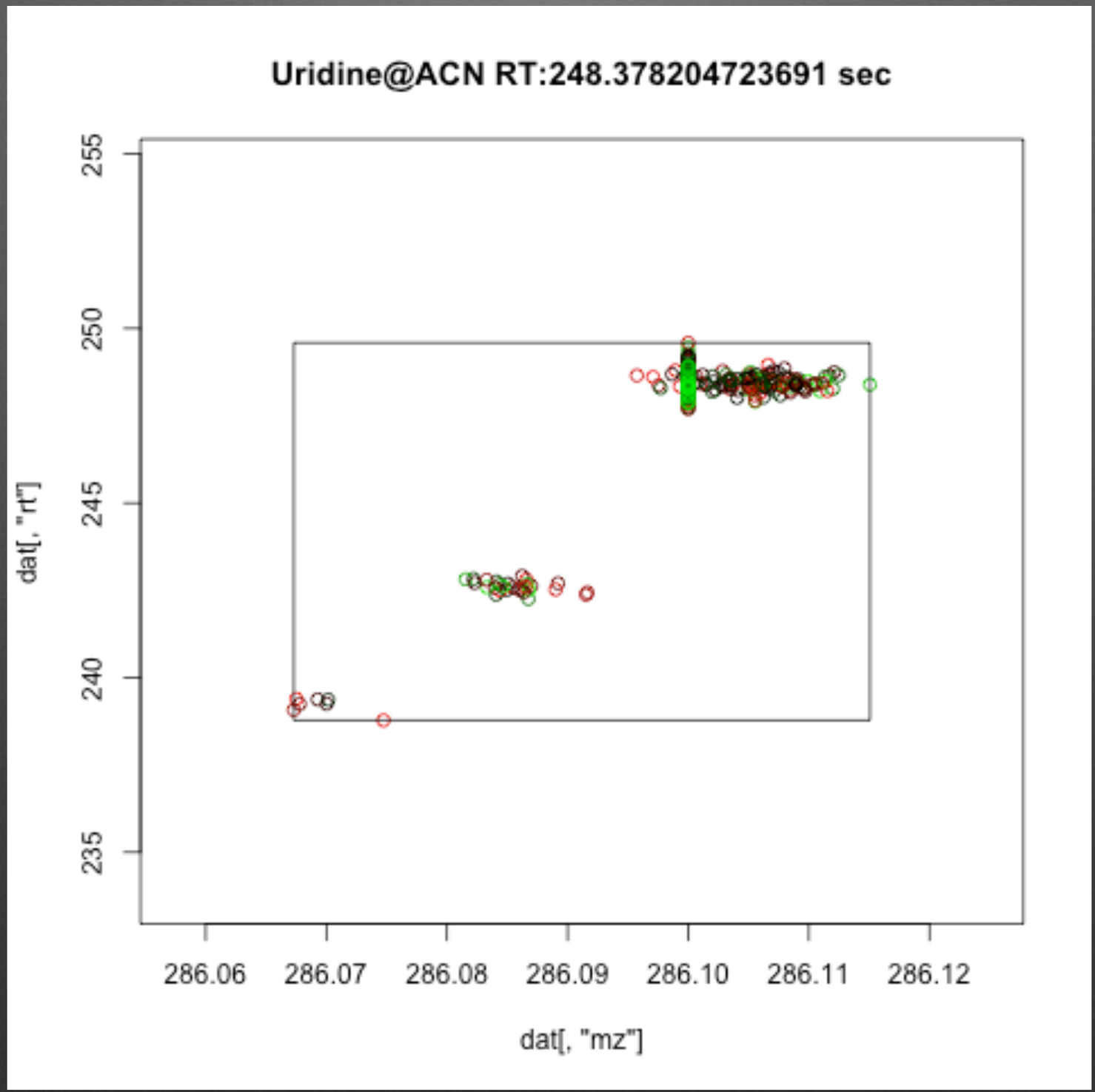
This would have ‘npeak’ > number of samples
You could play with parameters settings
Global parameters :(
Use different method :-)

Grouping = Nearest

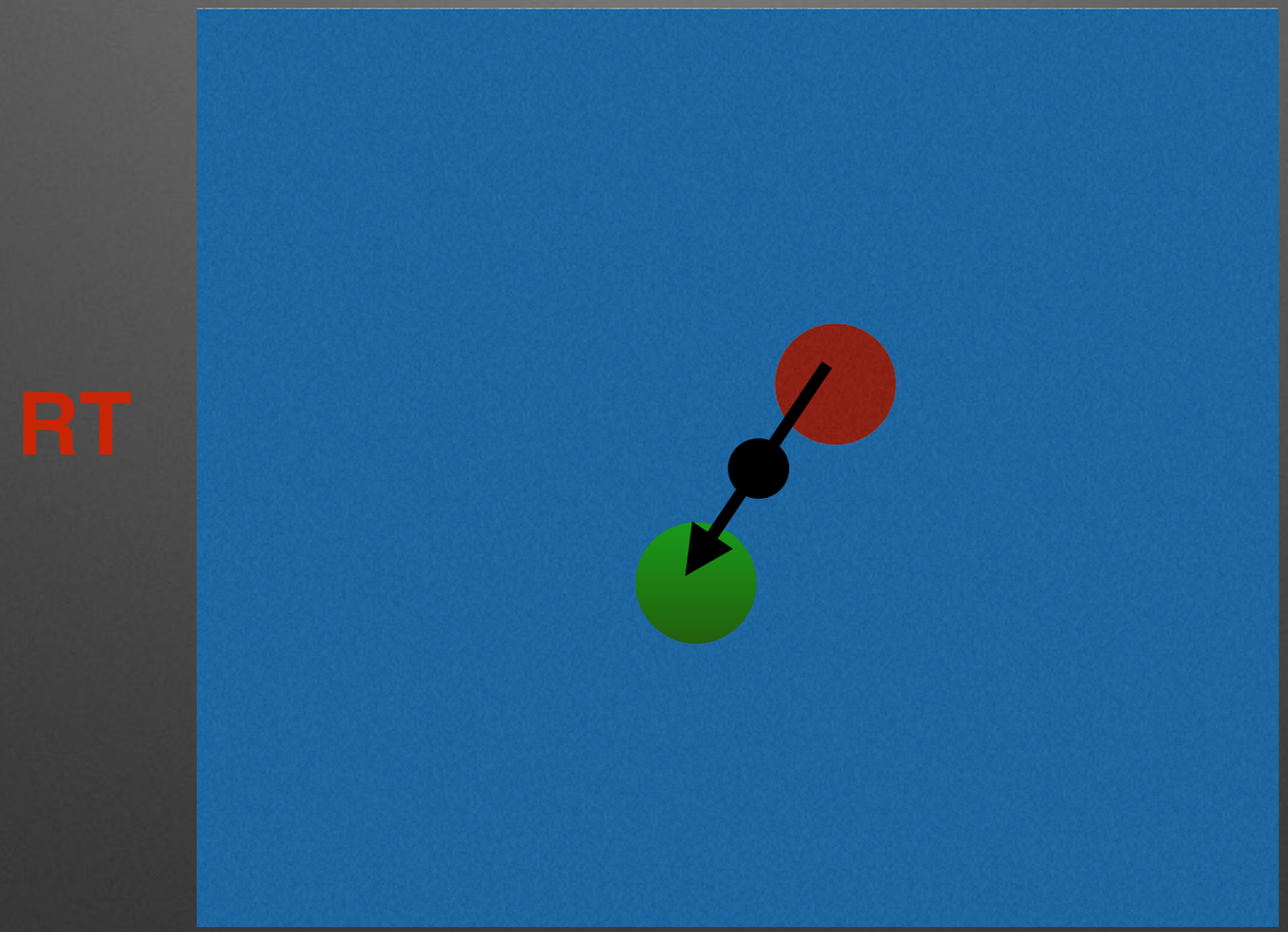
- Based on mzMine grouping/alignment algorithm
  - Uses nearest neighbor estimation.



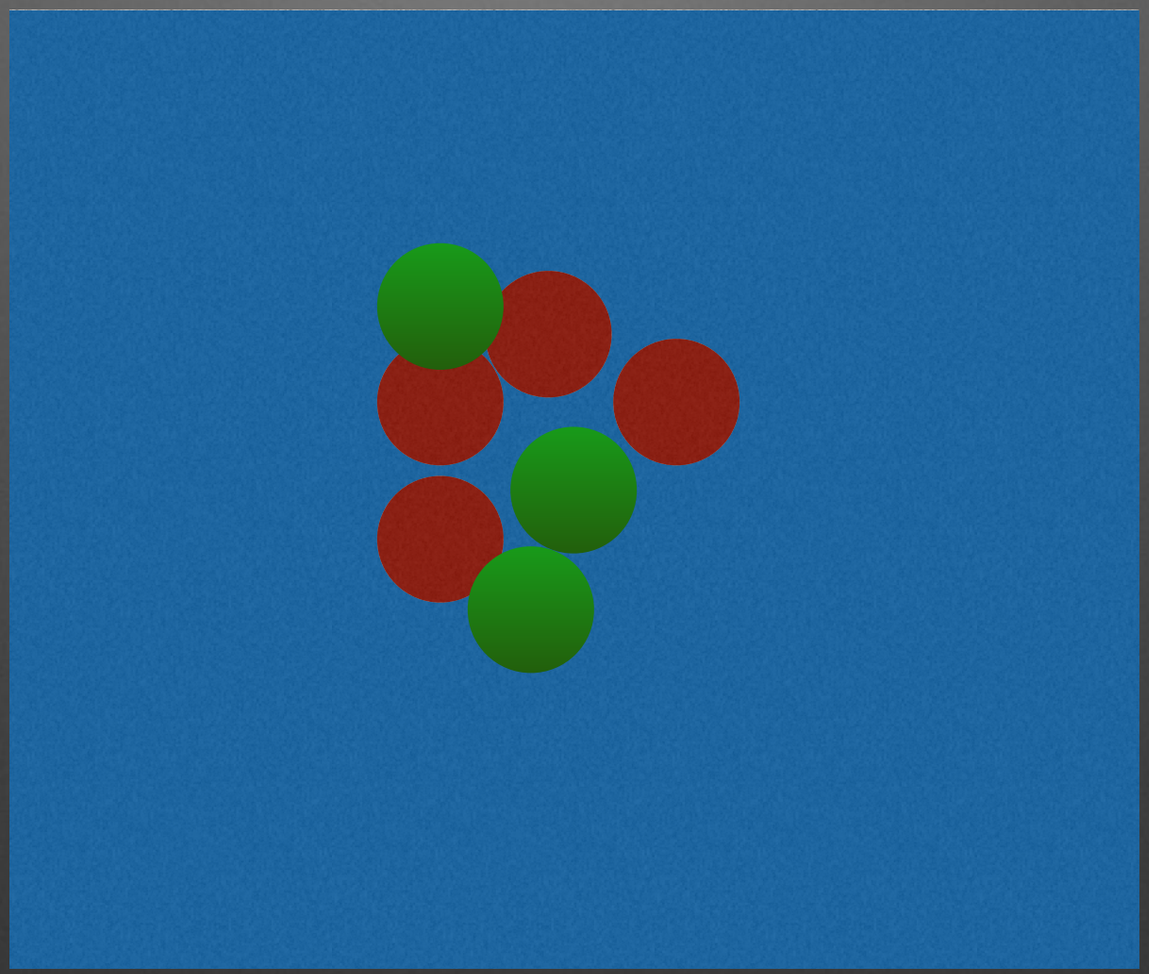



Uridine@ACN RT:248.378204723691 sec
255
250
245
240
235
dat[, "rt"]
286.06
286.07
286.08
286.09
286.10
286.11
286.12
dat[, "mz"]

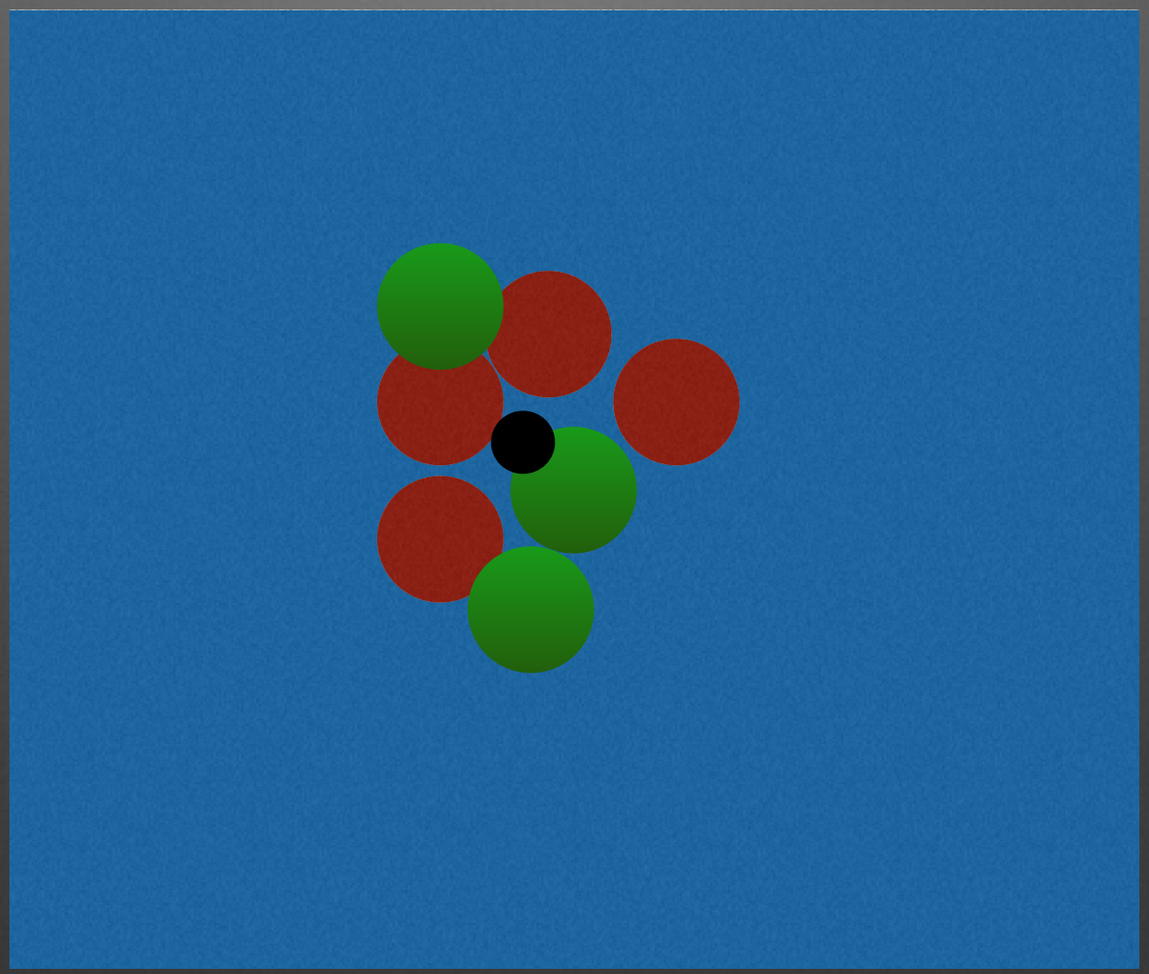
Group.nearest

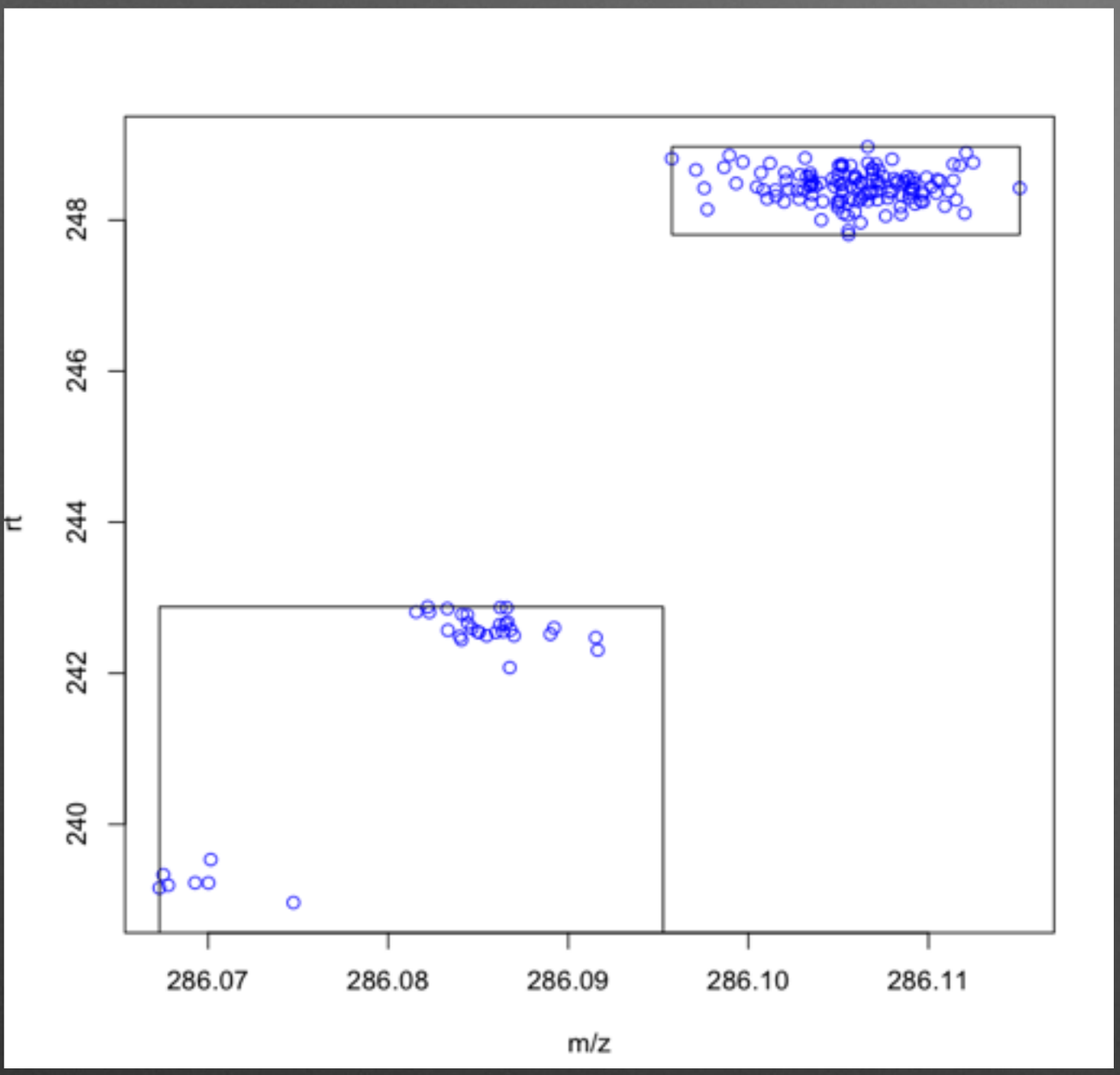


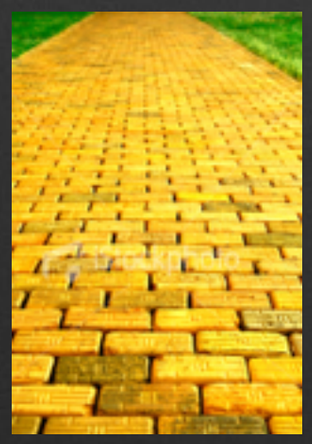
Group.nearest

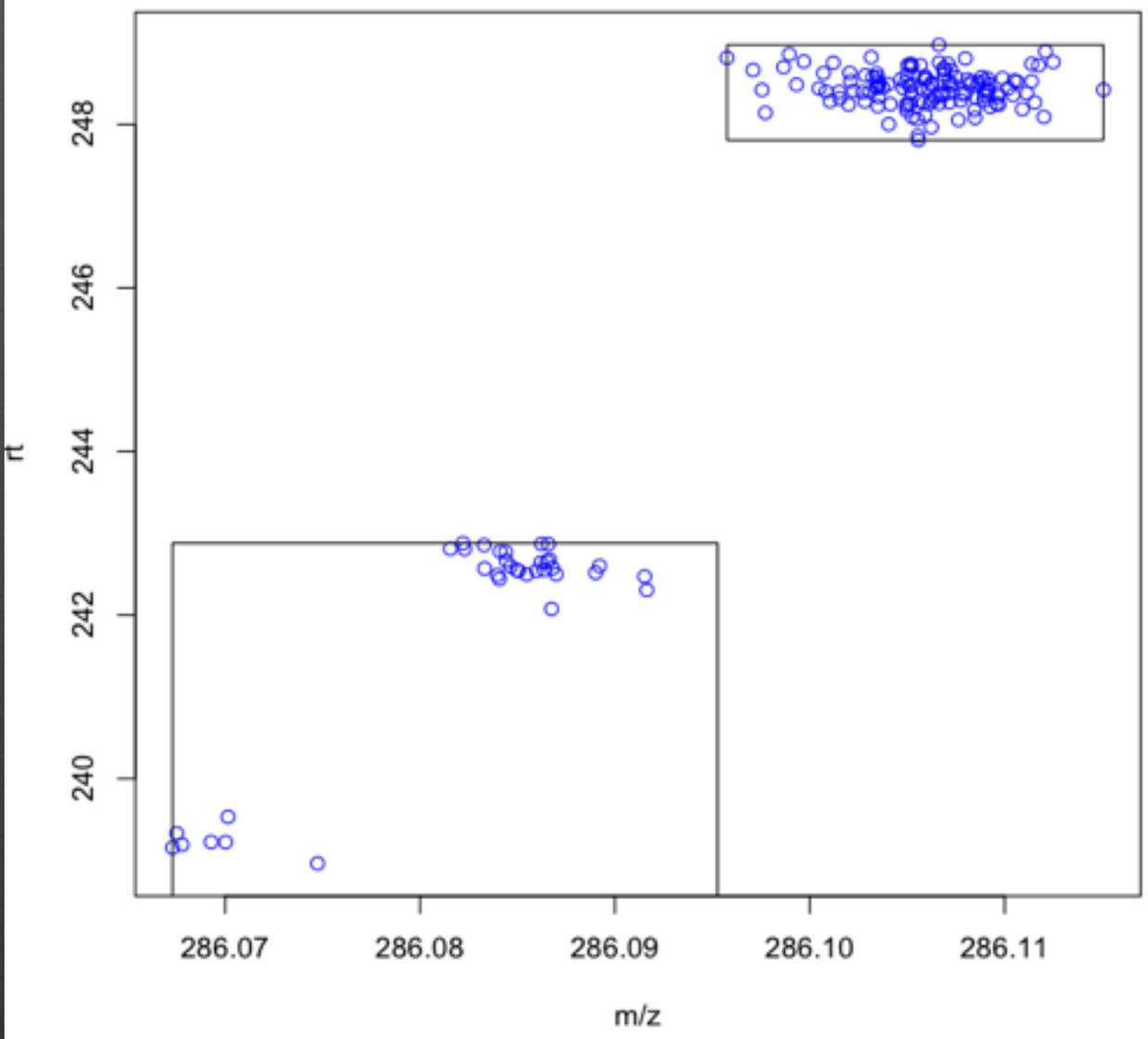


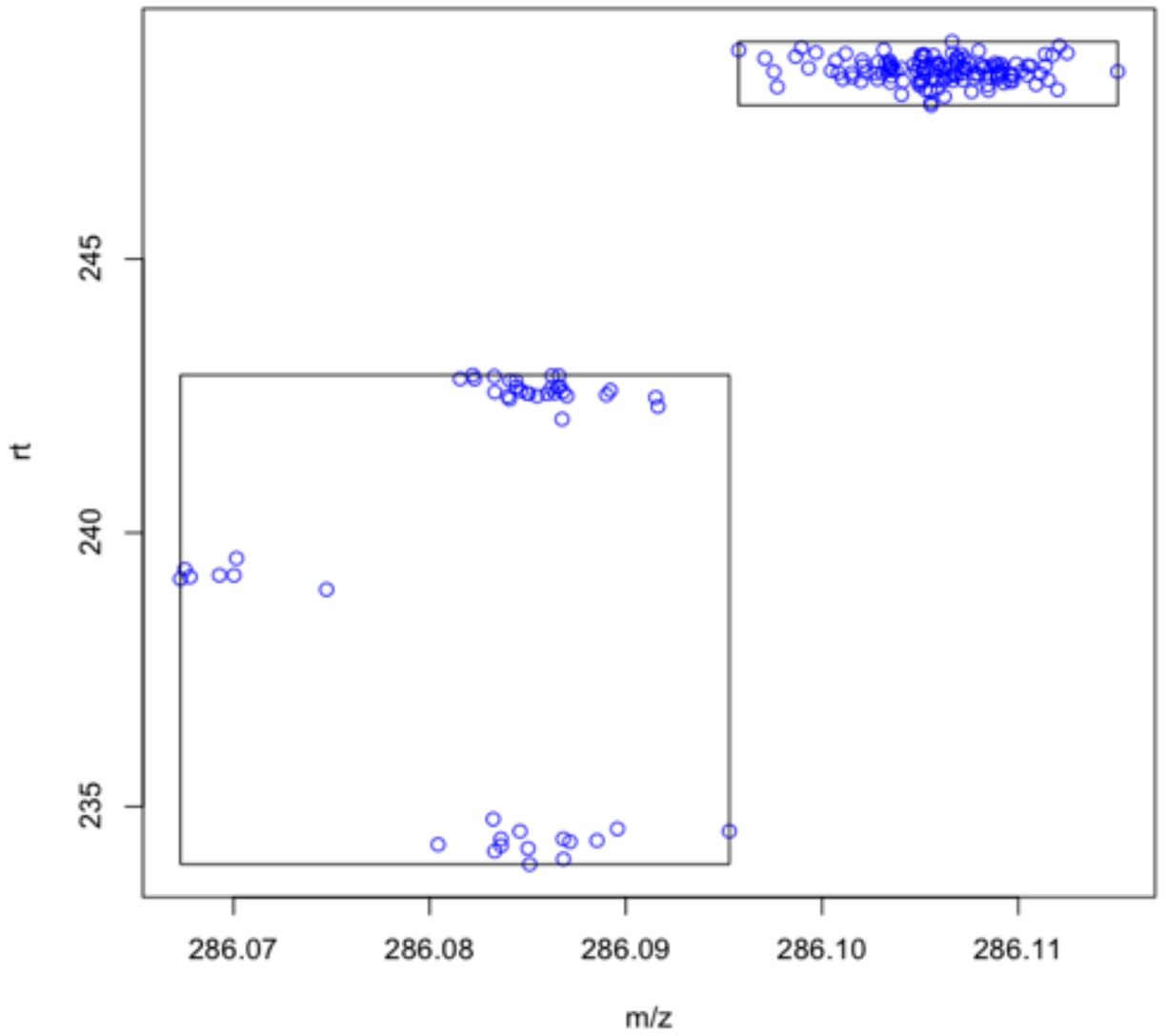
Group.nearest

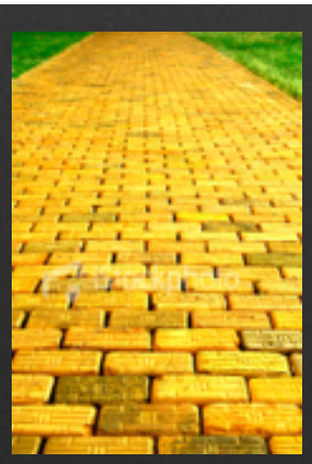



rt
248
246
244
242
240
286.07
286.08
286.09
286.10
286.11
m/z



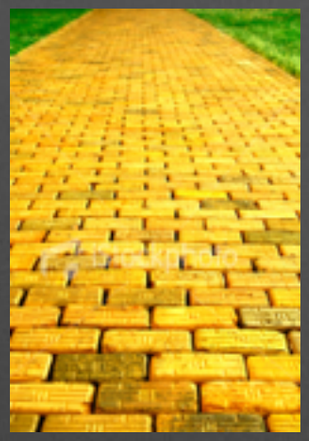

rt
248
246
244
242
240
286.07
286.08
286.09
286.10
286.11
m/z


rt
245
240
235
286.07
286.08
286.09
286.10
286.11
m/z



General Principals

Peak Detection



Grouping
Groups similar Peaks
across replicates

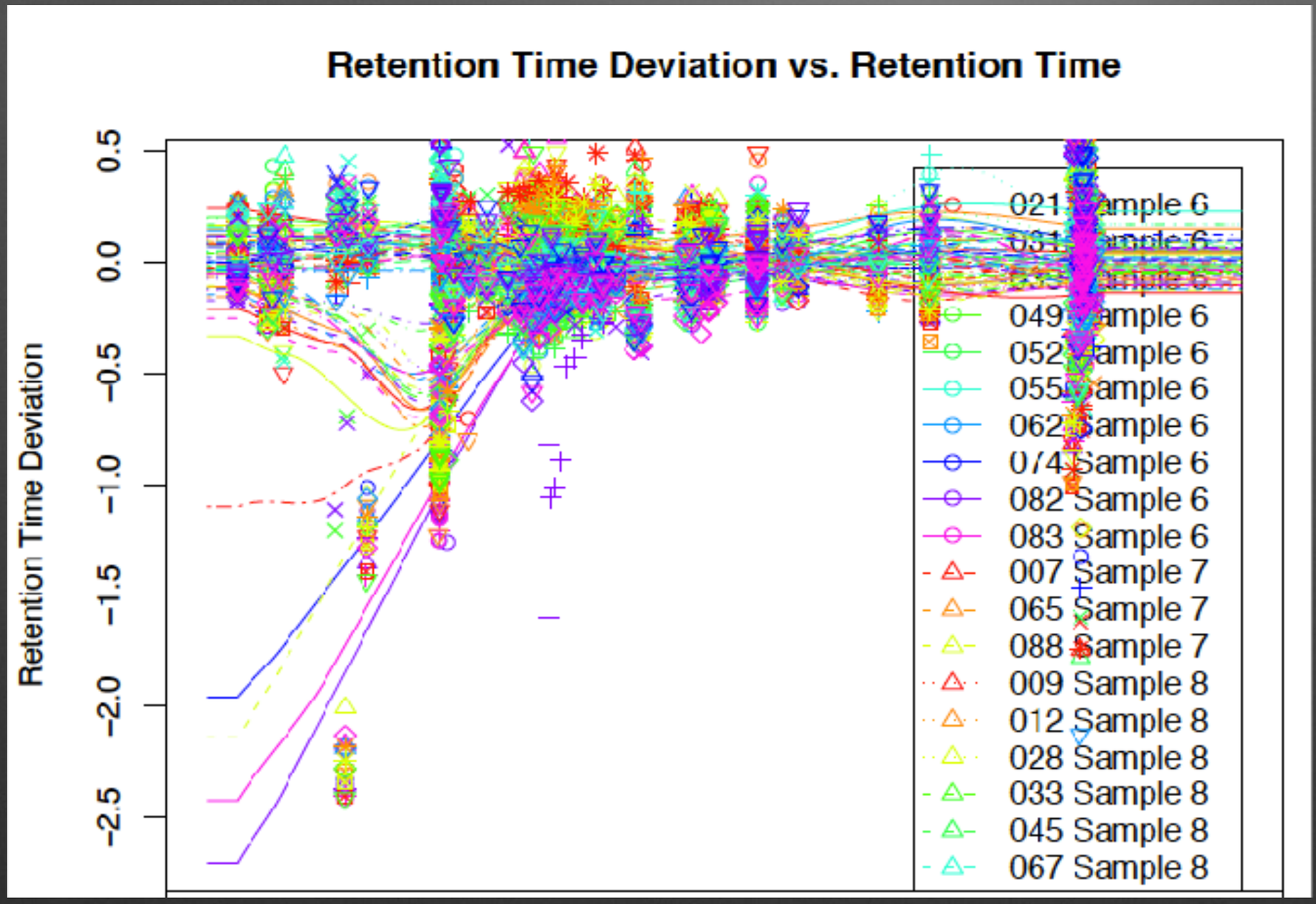
Retention Time Alignment

Statistical Analysis
of Classes

Retention Time alignment

- XCMS finds ‘well behaved groups’
  - These include group that have missing peaks, extra peaks or perfect groups (parameters)
    - Missing < n/2 !!
- Median found for each group
- Local regression used for each sample to find the deviation profile

Retention time alignment - loess



median rt of each ‘well behaved’ group
vs
rt of each file

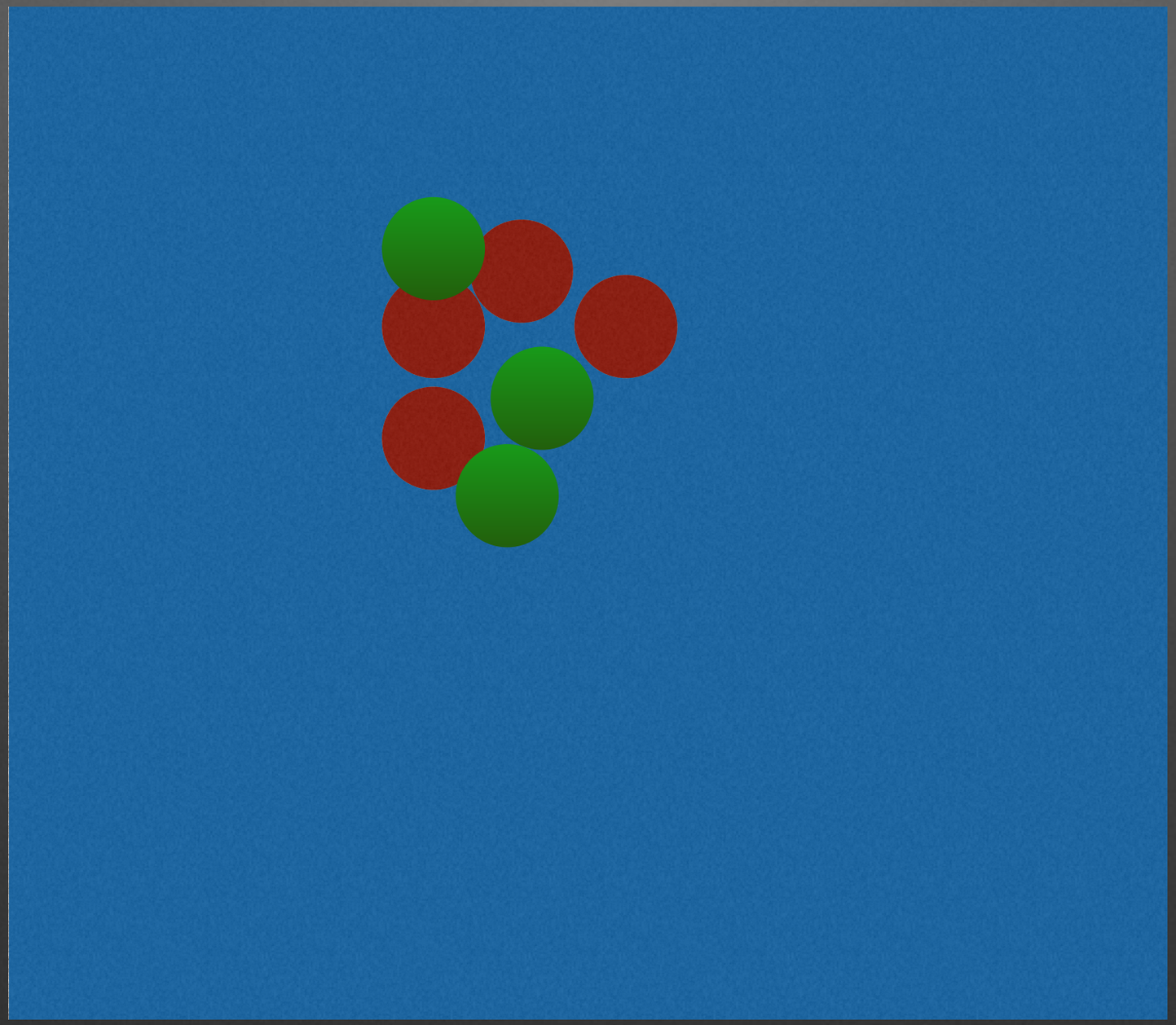
A good spread of anchors/‘well behaved peak groups’

Alignment

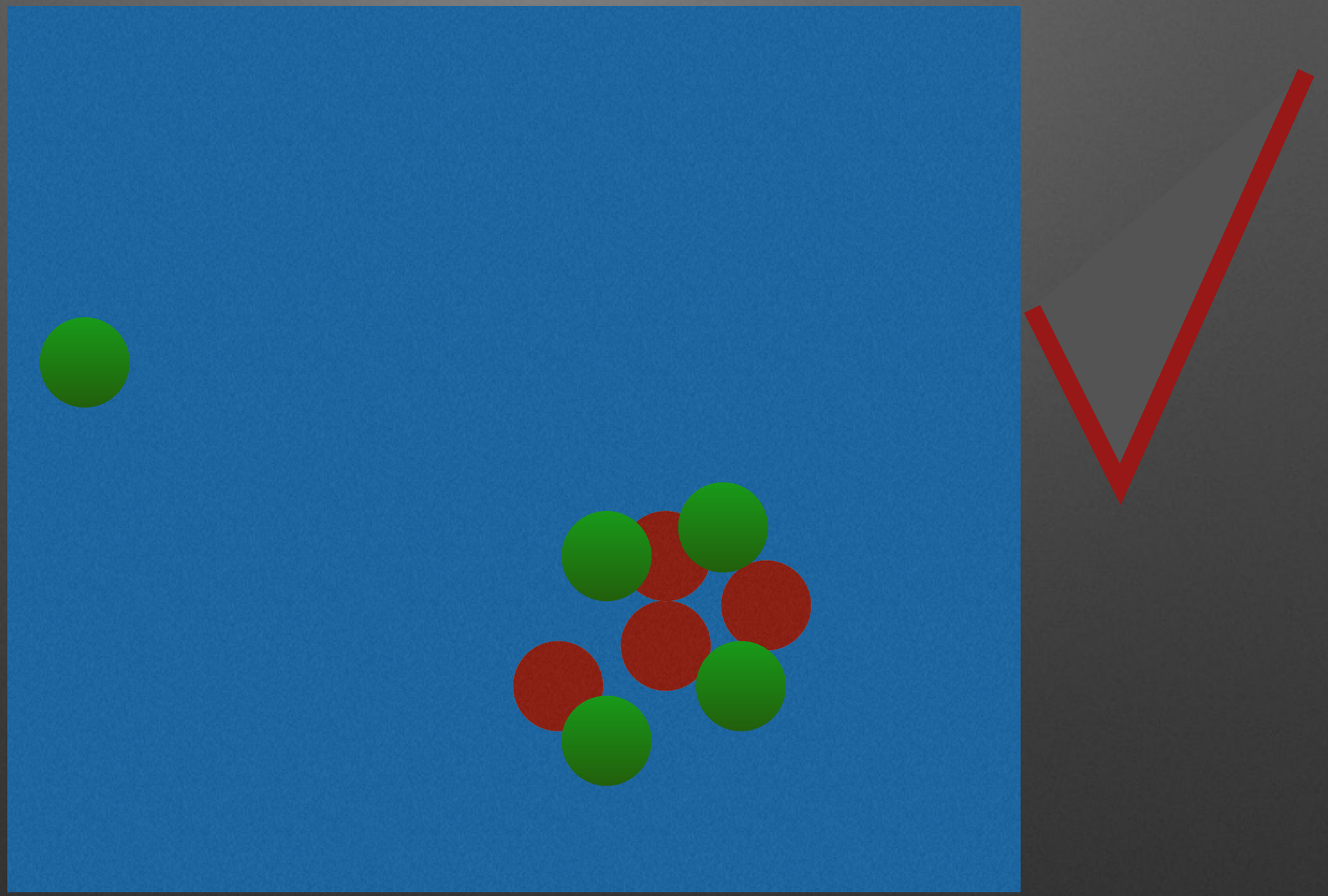
- Parameters:
  - missing = number of peaks removed from a ‘well behaved peak group’
  - extra = Number of additional peak in a ‘well behaved peak group’
  - span = Amount of smoothing in regression fitting ! Very sensitive! ~ smaller value more local alignment, larger more global alignment.

Missing = 1

4 samples each



Extra = 1



Retention time alignment obiwarp

John T. Prince and Edward M. Marcotte
Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping
Analytical Chemistry, 2060 78 (17), 6140-6152

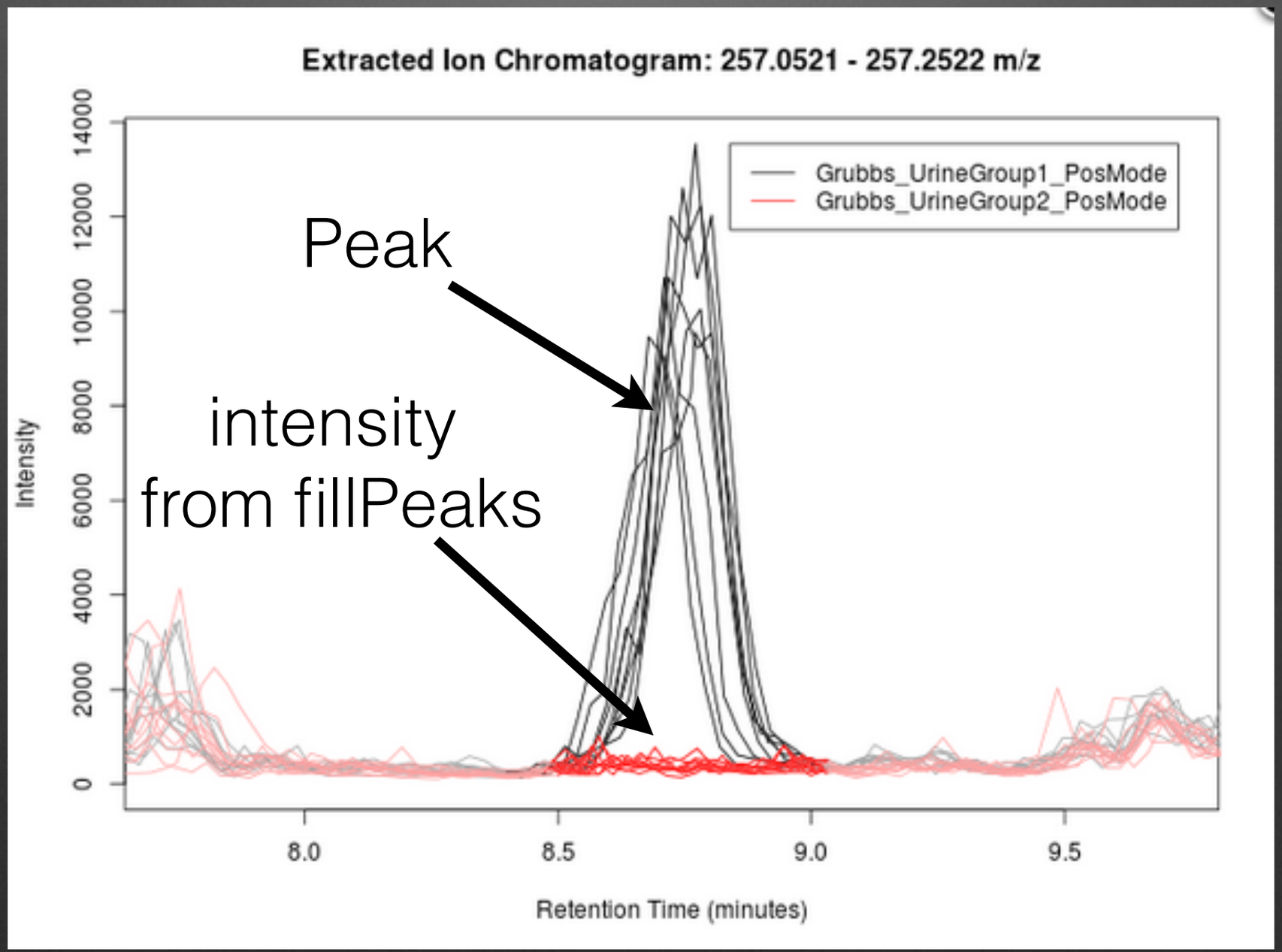
- obi-warp.sourceforge.net - original program
- Retention time correction based on spectra similarity
- Doesn’t rely on detected feature ~~ sort of
- No initial grouping needed

Retention time alignment obiwarp

- Uses a warping technique to warp data to a median chromatogram.
  - This acts as a mold which other spectra are warped to
- Uses a dynamic programming to find path of greatest similarity between median chromatogram and current chromatogram

FillPeaks

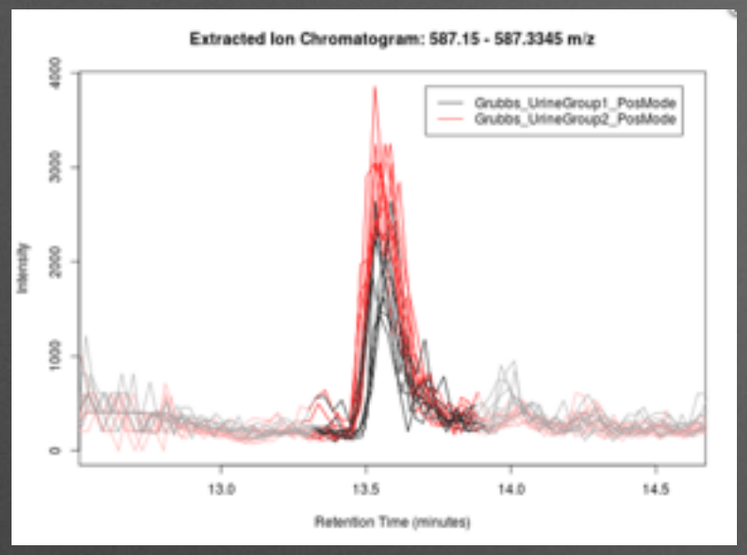
- Going back to each file to find any intensity that wasn't peak picked

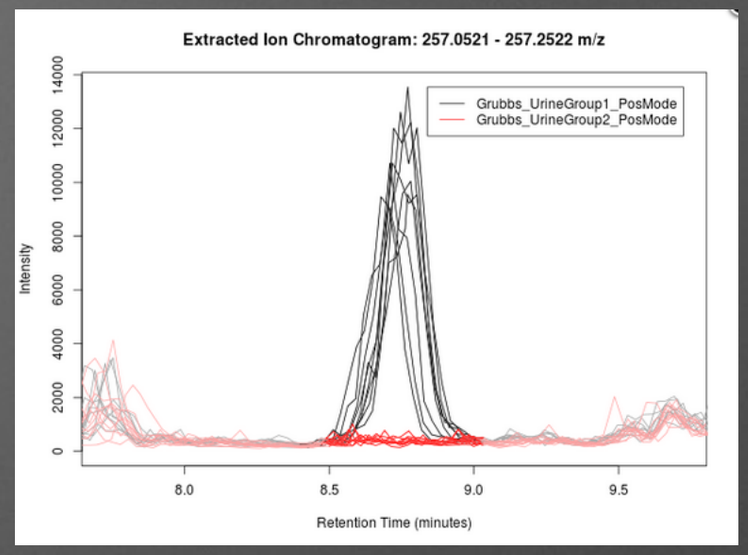


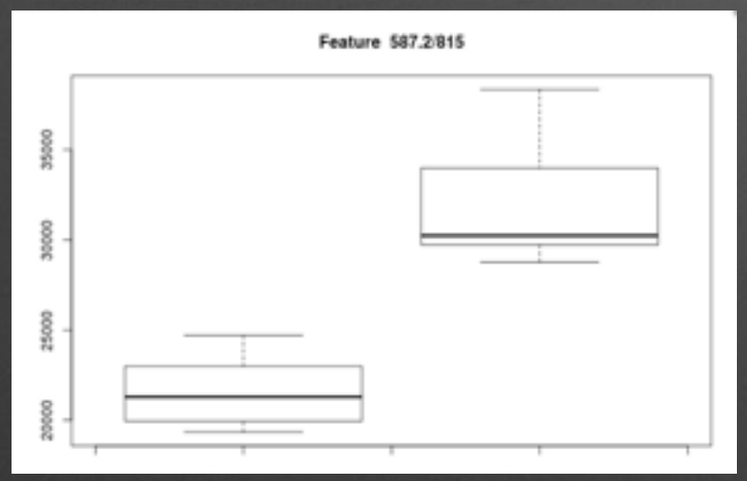
Finally !!

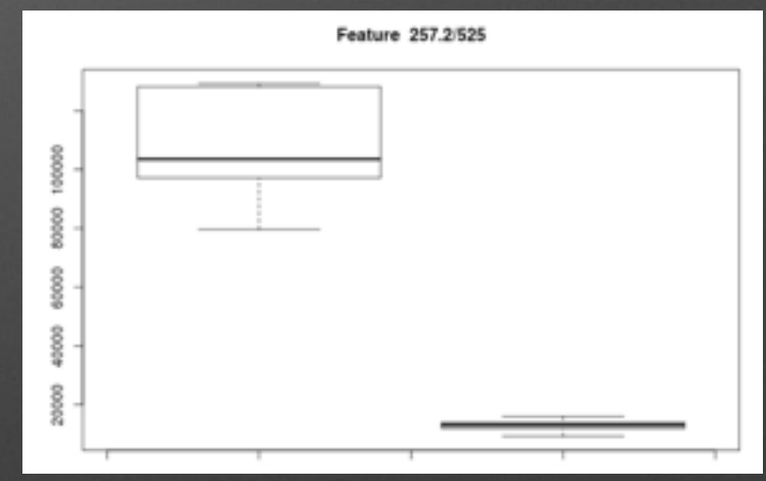
- We have all of our data corrected in a form we can use.
  - Lets look at some data processing:
    - heatmaps
    - PCA
    - Some Stats

WAIT !!!!







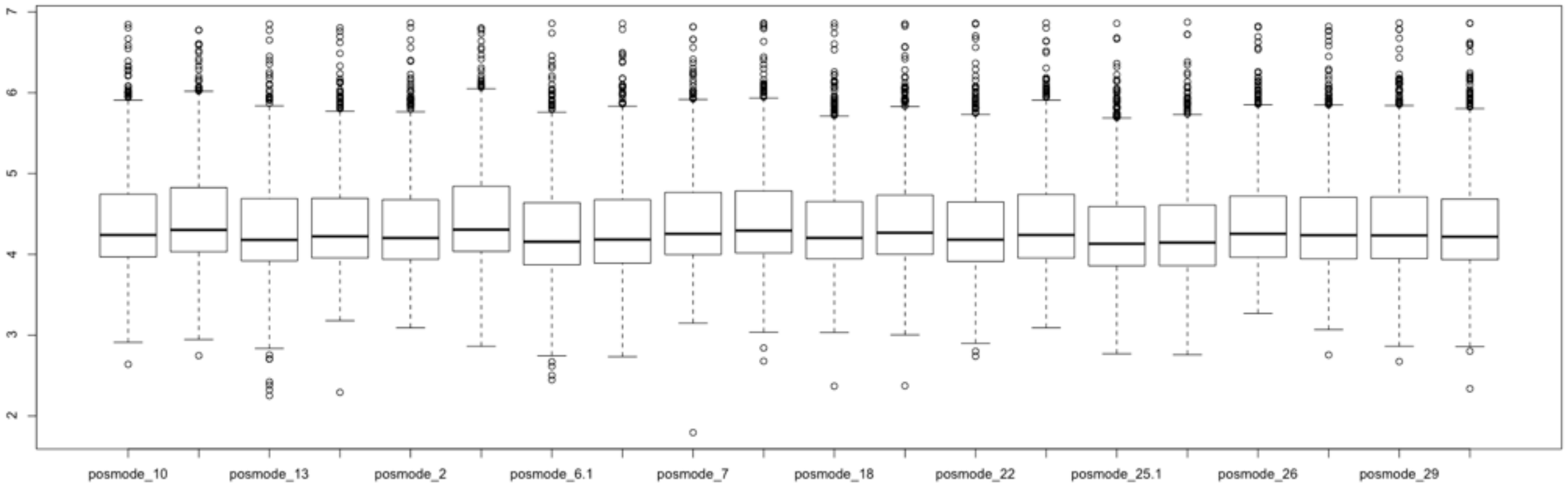


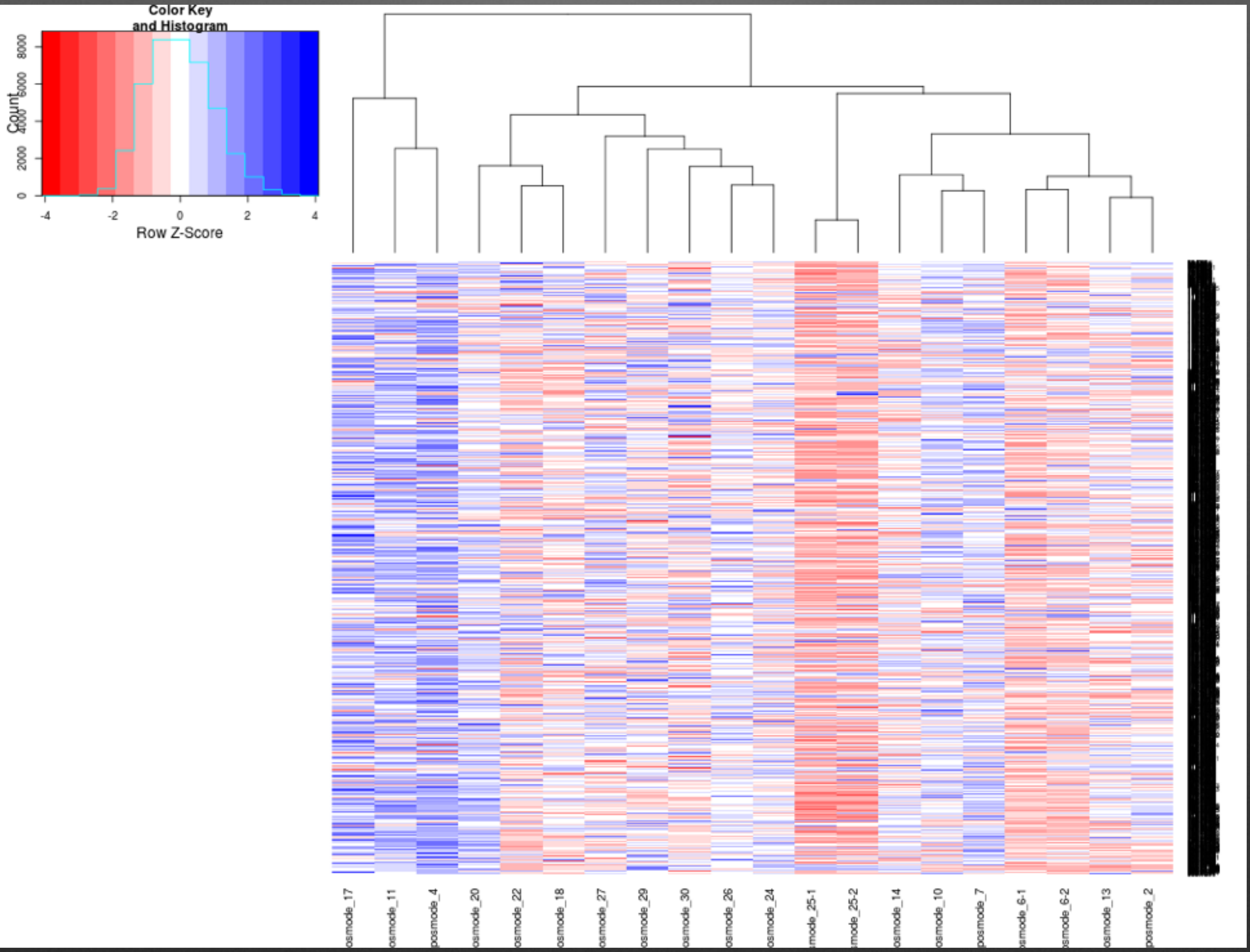
Job#1051415 : Grubbs_urine_pos_mmchg

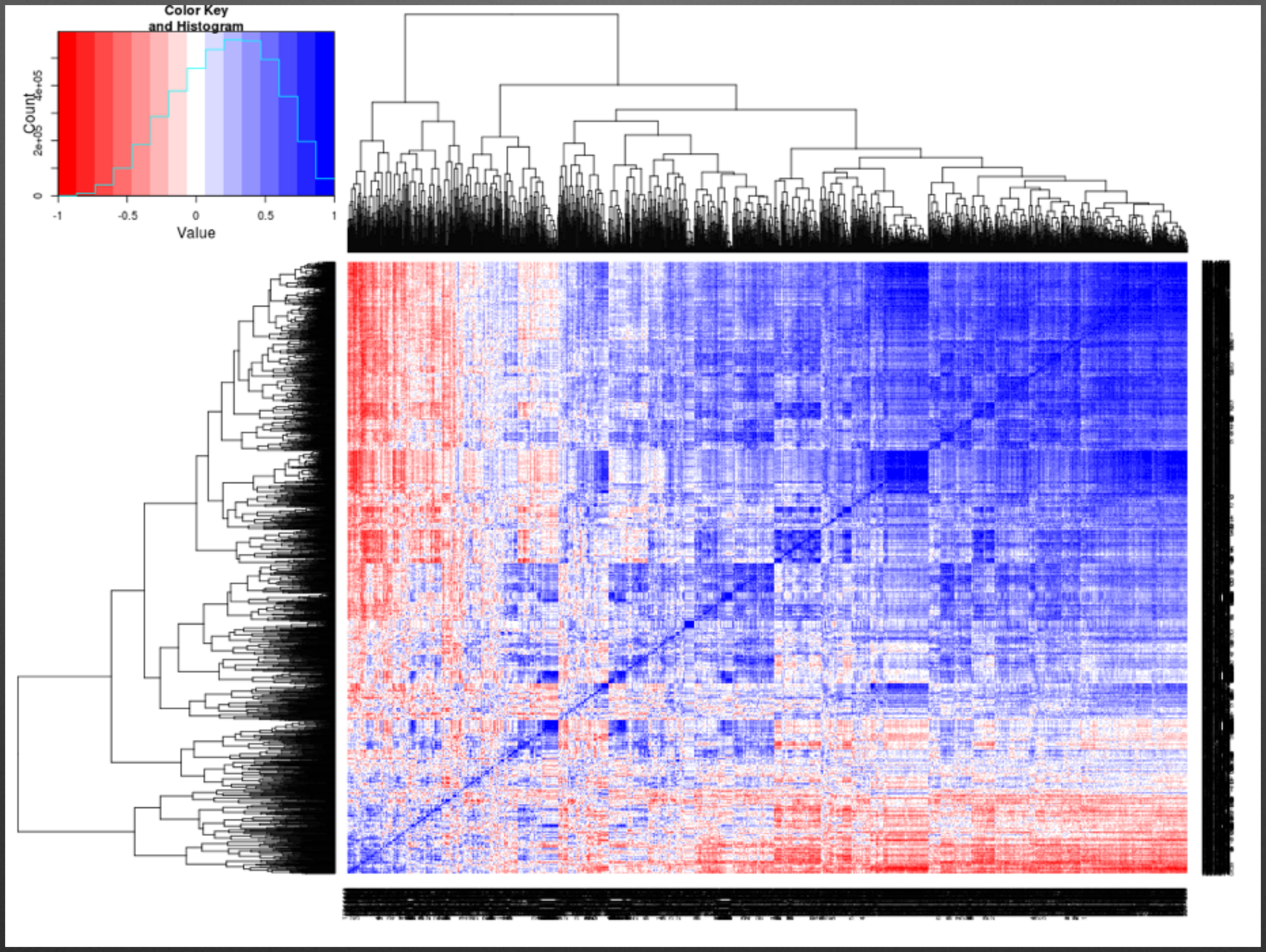
Columns | Show isotopic peaks | Page 1 of 19 | 100 | View 1 - 100 of 1,801

Feature	fold ch	p-value	q-value	m/z	retention tim	MaxInt	Ctrl(sd)	Ctrl(x̄)	Exp(sd)	Exp(x̄)	isotopes	adducts	feature g
1	3.0	5.33057e-8	0.00005	204.1446	11.55	8,367	8,385.9	61,606	3,635.3	20,590			35
2	1.5	6.94626e-8	0.00005	587.2266	13.59	3,858	1,914.6	21,603	3,067.4	31,738		[M+Na]+ 5	4
3	8.5	2.04304e-7	0.00008	257.1582	8.75	13,544	18,092.6	108,869	2,010.2	12,876		[M+H-H20]	53
4	3.8	2.44253e-7	0.00008	234.1663	11.74	8,264	14,440.9	71,894	10,301.1	19,157		[M+H-CH3	57
5	2.2	6.72076e-7	0.00018	345.1104	10.74	5,082	6,835.8	44,308	3,762.4	20,112			33
6	1.8	1.03879e-6	0.00023	377.1435	11.91	160,143	116,893	501,472	135,537.3	909,410	[69][M]+		18
7	1.3	2.79905e-6	0.00054	181.0589	11.24	148,137	70,401.6	925,013	48,407.5	715,125			14
8	1.5	4.03600e-6	0.00068	193.4786	13.59	4,468	2,356.9	20,927	4,613.7	32,018			4
9	2.5	4.56496e-6	0.00069	249.1814	19.42	3,787	13,536.2	68,040	6,310.9	26,914			136
10	1.8	5.57117e-6	0.00071	390.1744	12.87	3,106	7,077.7	43,763	4,776.8	23,898		[M+K+NH3	48
11	1.6	6.04996e-6	0.00071	425.1022	13.61	2,843	1,969.2	16,840	4,621.9	27,629		[2M+K]+ 1	4
12	2.3	6.40652e-6	0.00071	549.3642	16.16	2,285	2,938.8	14,390	2,200.5	6,283	[134][M]+	[3M+2Na-H	24

Normalisation needed?




Color Key
and Histogram
Count
0
2000
4000
6000
8000
-4
-2
0
2
4
Row Z-Score
posmode_17
posmode_11
posmode_4
posmode_20
posmode_22
posmode_18
posmode_27
posmode_29
posmode_30
posmode_26
posmode_24
mode_25-1
mode_25-2
posmode_14
posmode_10
posmode_7
posmode_6-1
posmode_6-2
posmode_13
posmode_2



Correlation heat map of the Bonferroni corrected ANOVA p-values

Summary

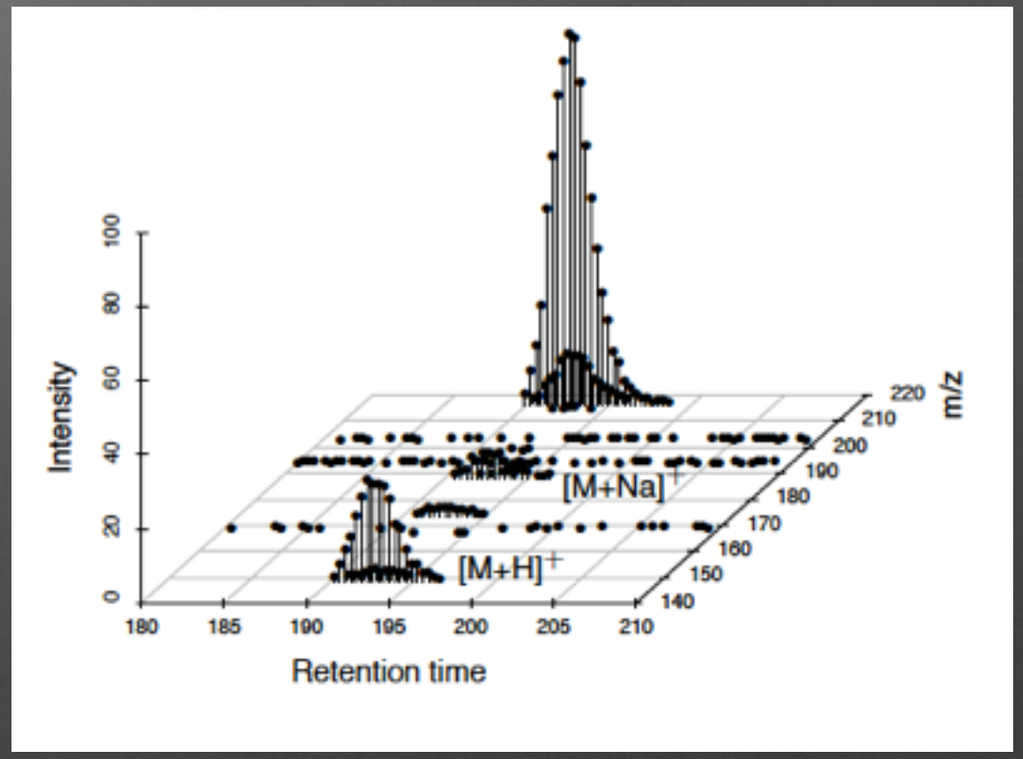
- XCMS processes LC-MS data and is complex
- XCMS processes LC-MS data and uses some simple algorithms. There are multiple algorithm for different jobs/data types.

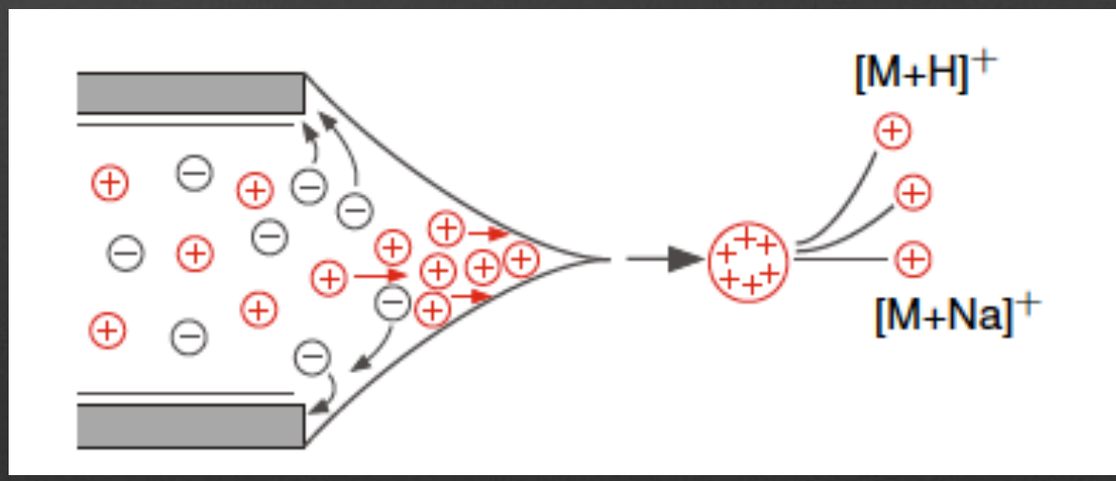
Boxes and Foxes

- XCMS is all about boxes
  - Boxes are sly and slippery and are the main problem in data analysis
  - If you’re having issues try changing alignment methods and thinking about how much deviation in m/z or RT the data has before and post alignment

CAMERA

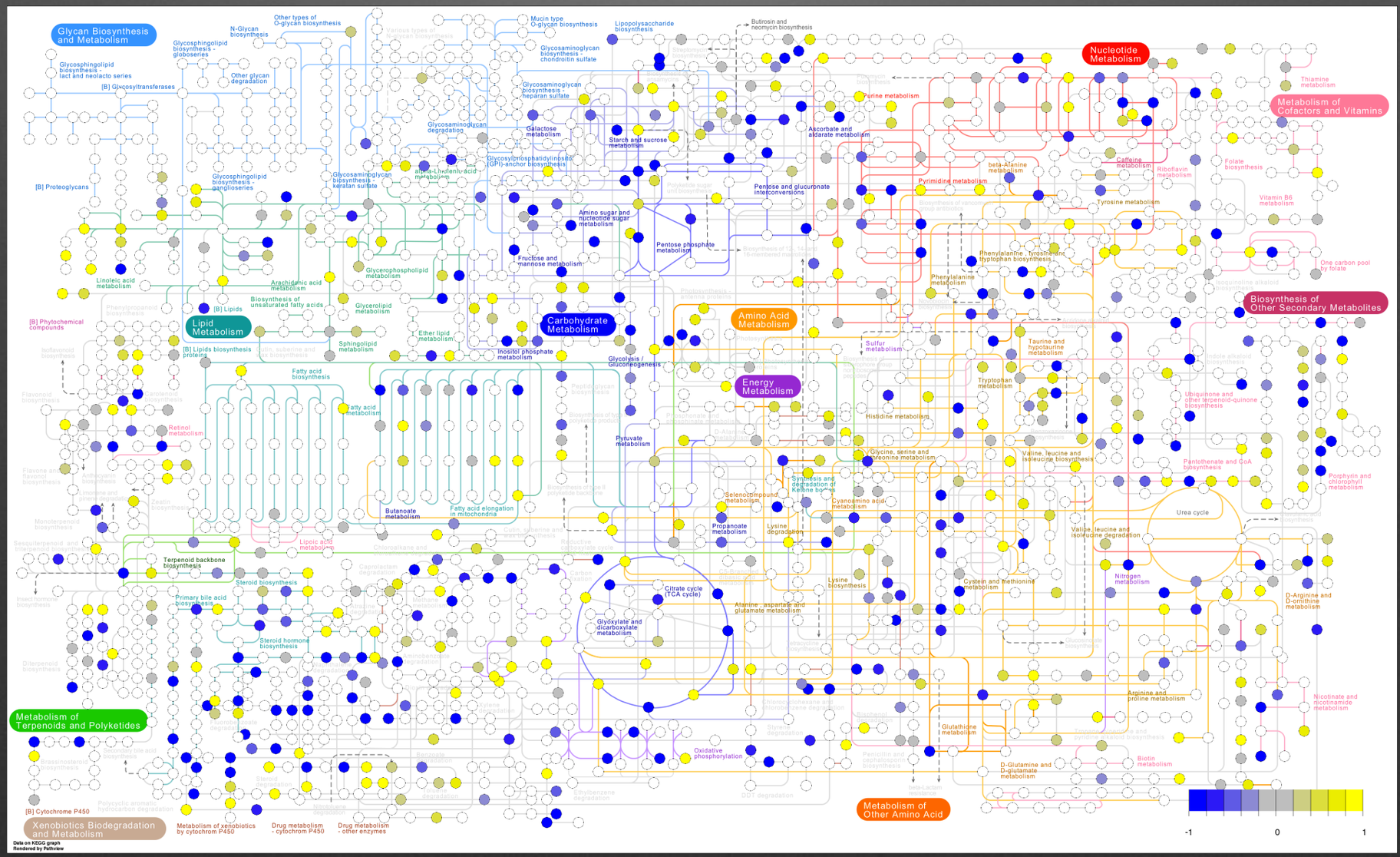
- Same compound should be at the same retention time
- Same compound should have a linear relationship
- Using linear correlation and RT windows adducts/ isotopes are labeled





On-wards to biology

- Network maps from related metabolites



Thank You!

- Questions?
- Many more updates coming soon including speed and more stats

Prof. Gary Siuzdak

Dr. Colin Smith

The whole xcms team