

# Liquid Chromatography Mass Spectrometry (LC-MS) based Metabolomics

Preprocessing

**Adrie D. Dane**

Department of Epidemiology and Data Science (EDS)

Bioinformatics Laboratory

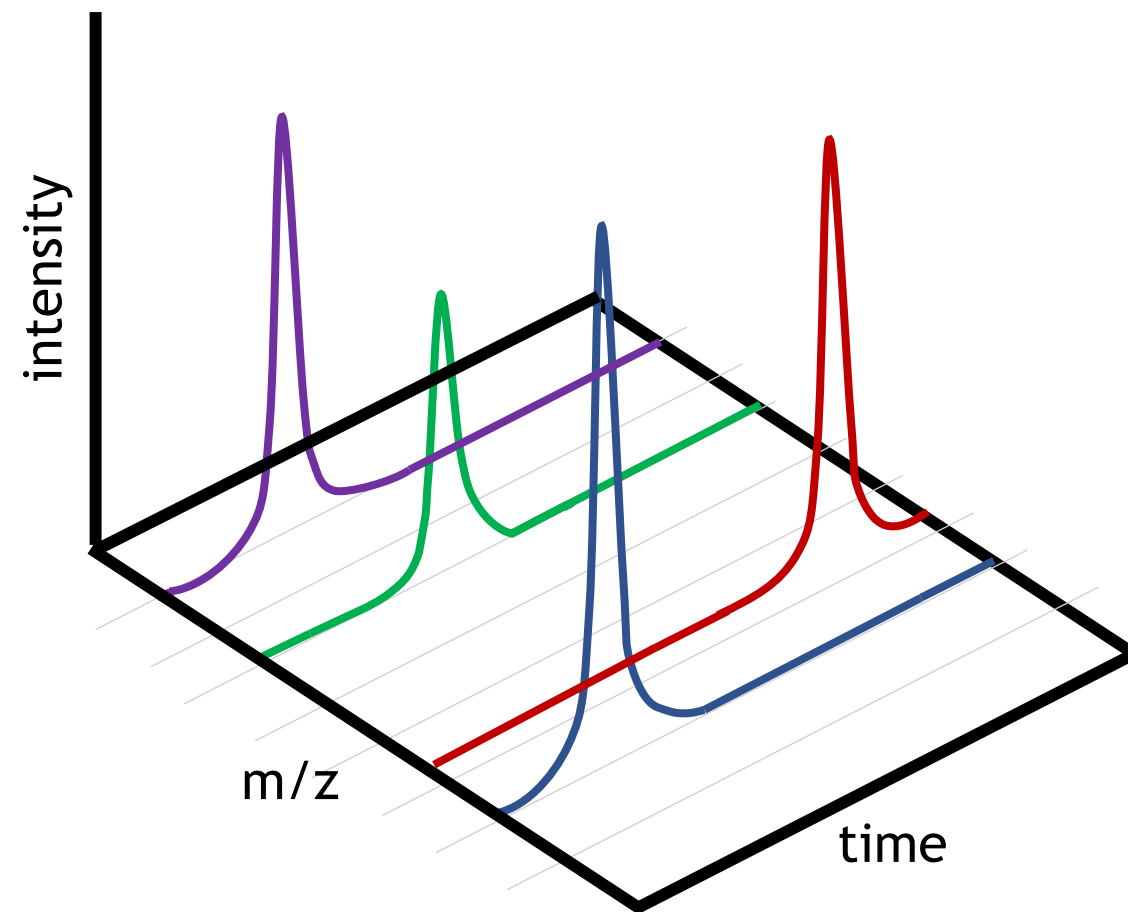
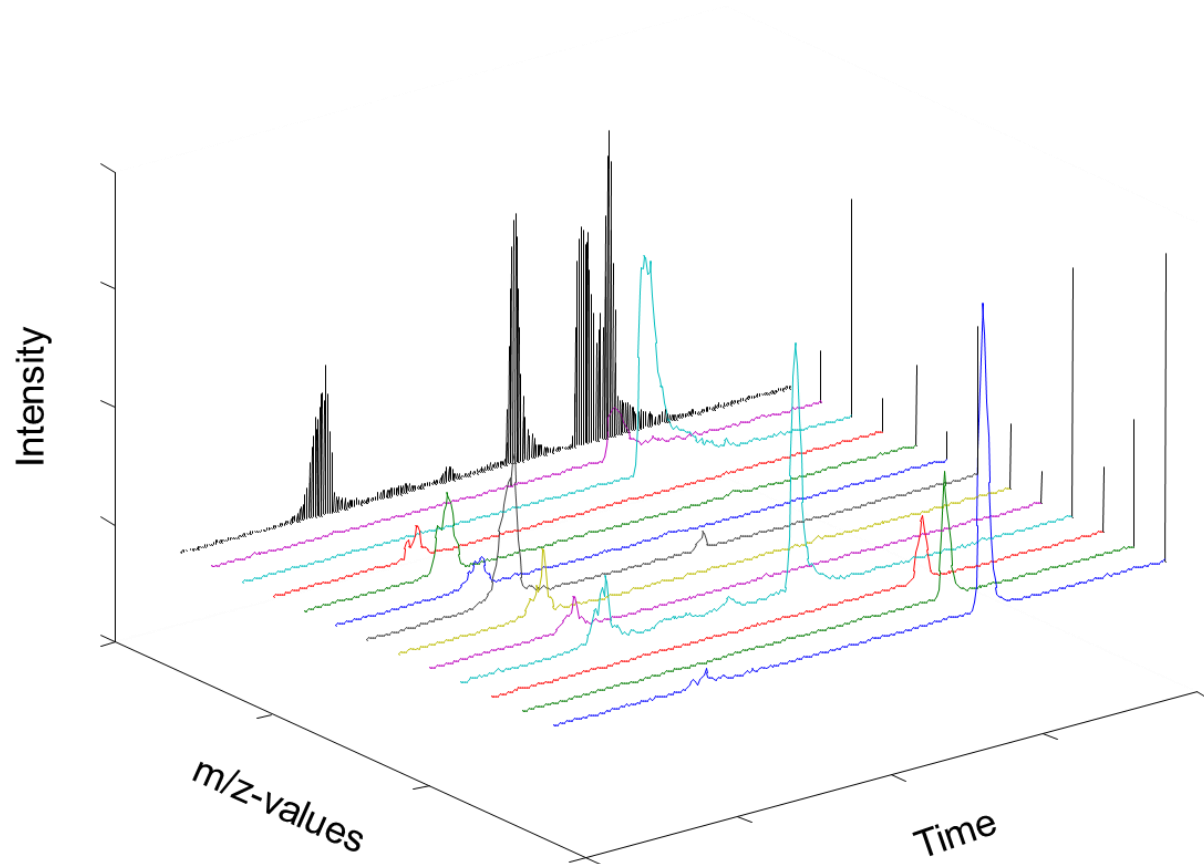
[www.amsterdamumc.nl](http://www.amsterdamumc.nl) | [bioinformaticslaboratory.eu](http://bioinformaticslaboratory.eu) | [cfmetabolomics.nl](http://cfmetabolomics.nl)

# Preprocessing

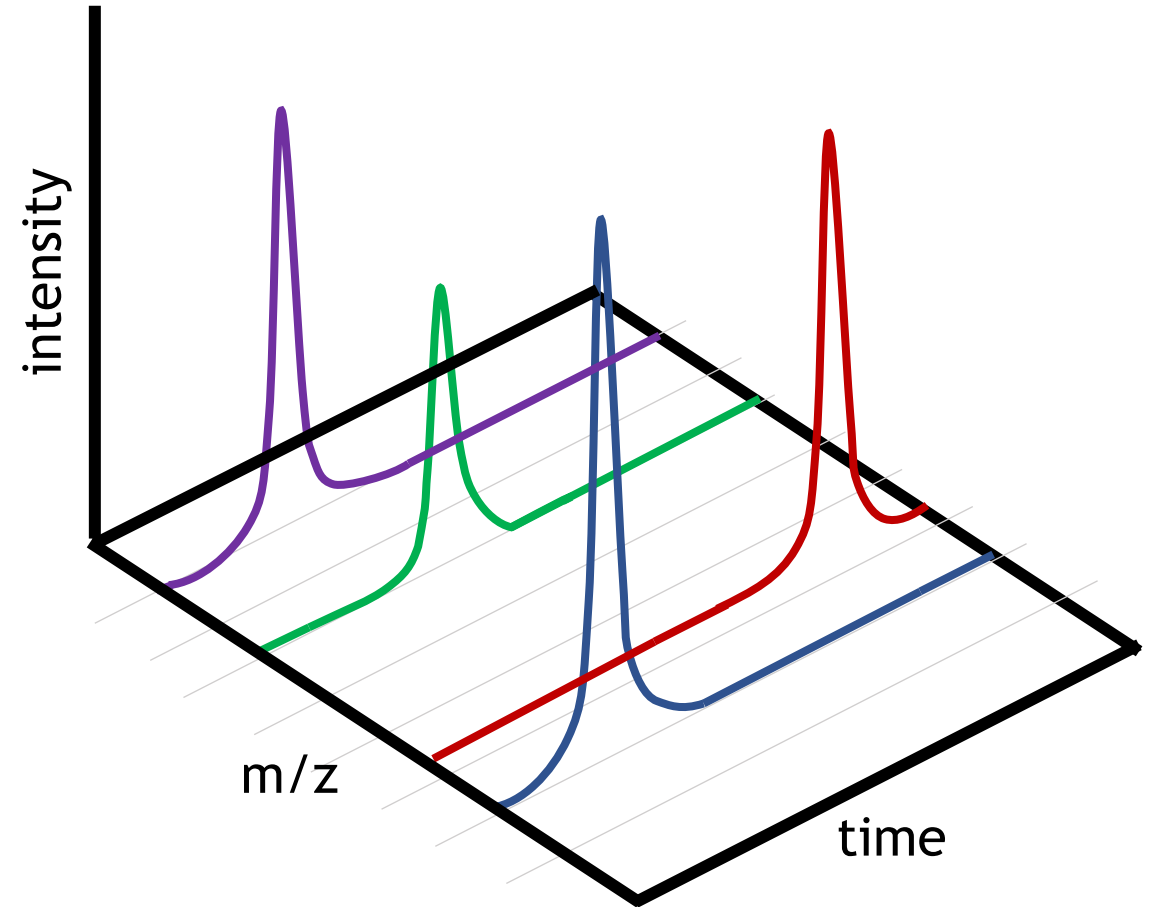


1. Import LC-MS Data
  - Extract Ion Chromatograms EIC
2. Find peaks
  - Filter data
  - Identify peaks
  - Integrate
3. Match peaks across samples
  - Retention time correction
4. Fill in Missing Peak data

# Simplified data



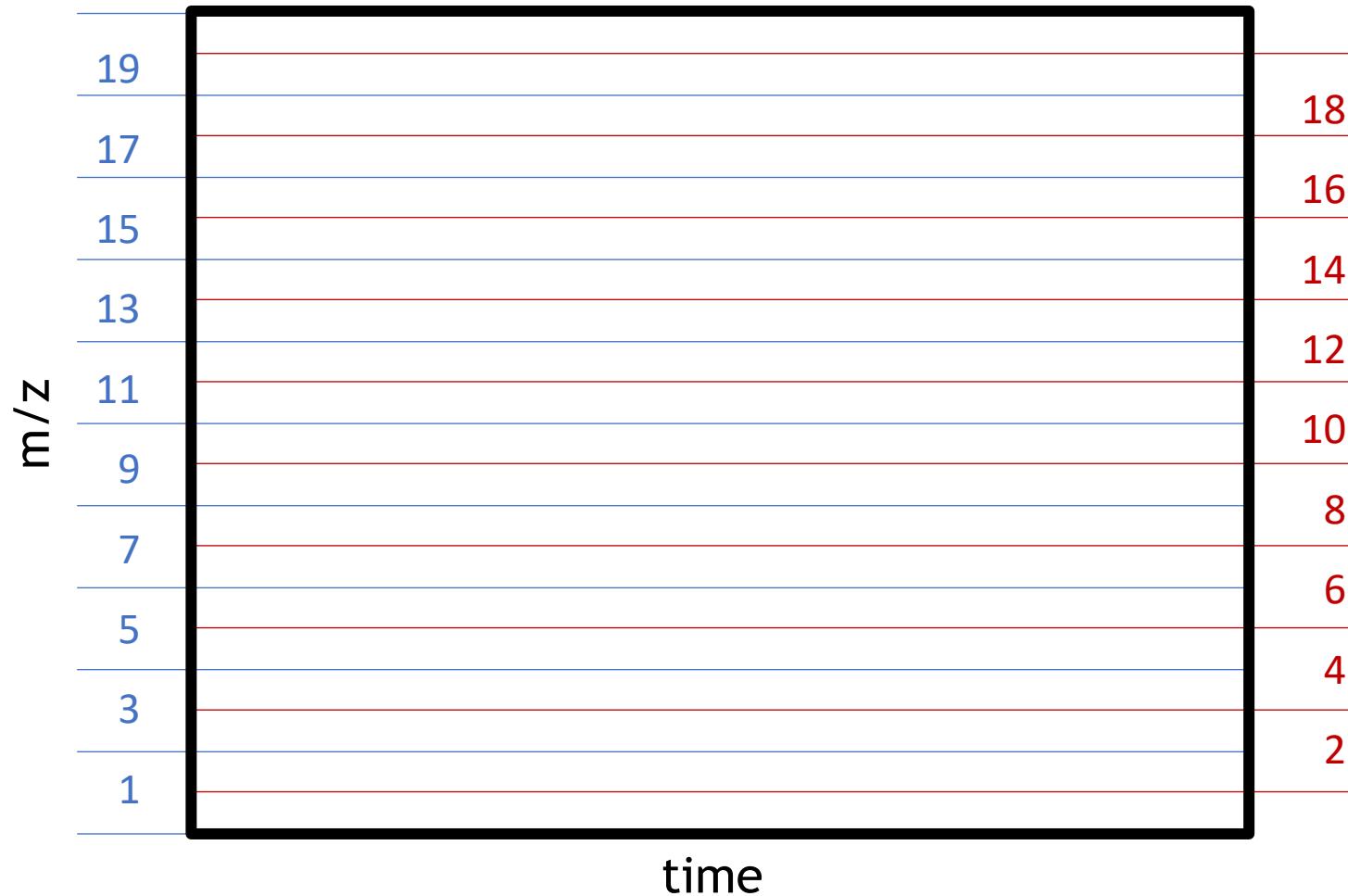
# Preprocessing 1: Extract ion chromatograms



# Binning: construct profile matrix



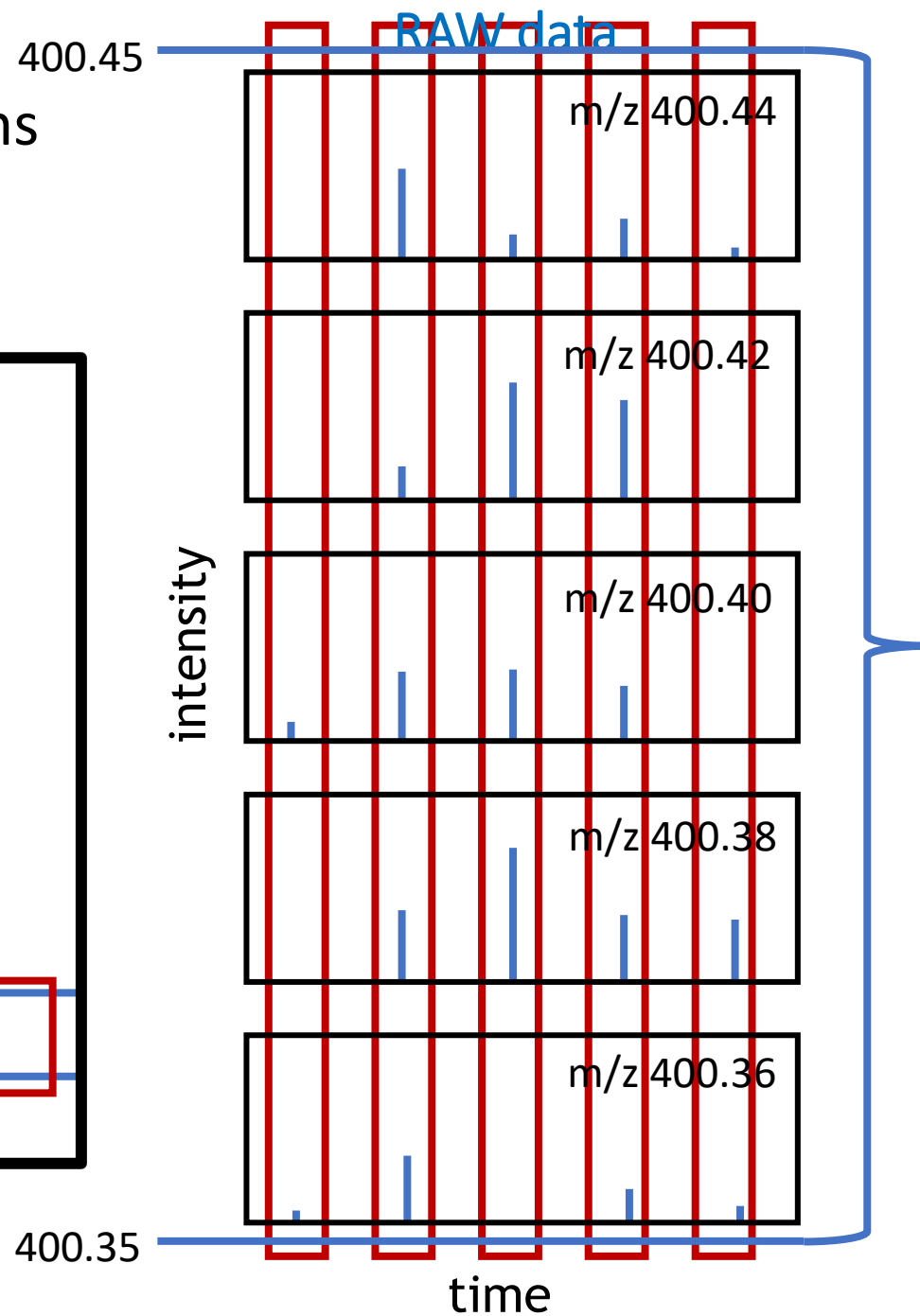
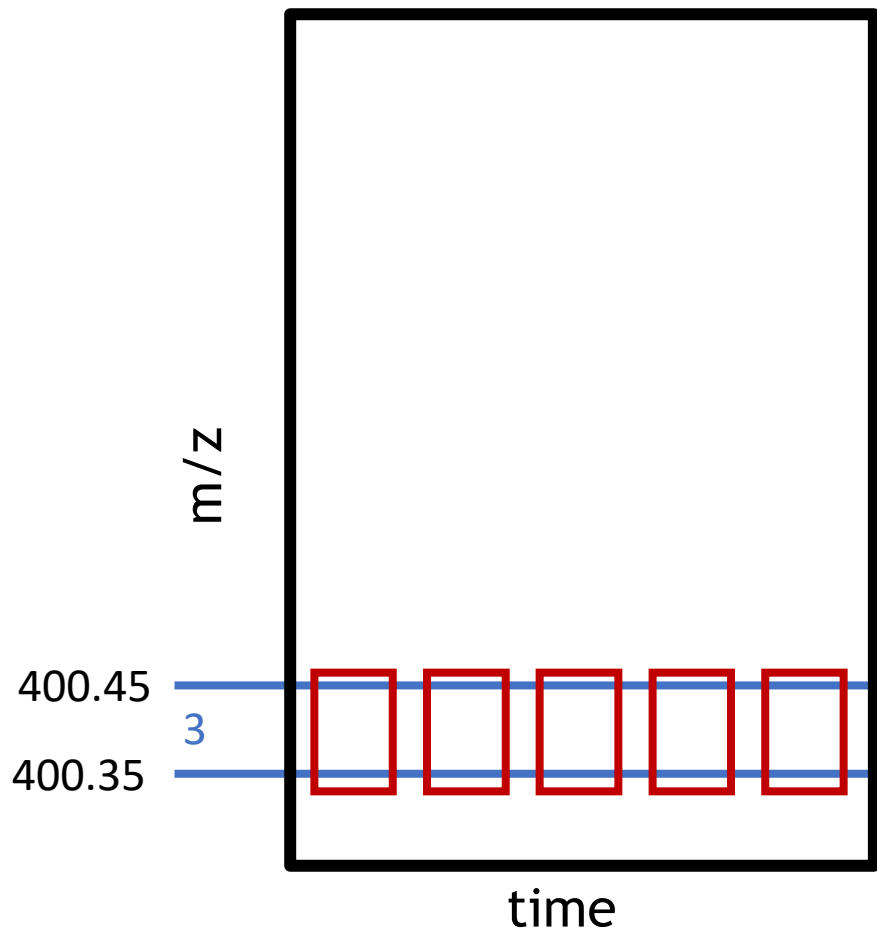
- Extract ion chromatograms



## Extract Ion Chromatograms

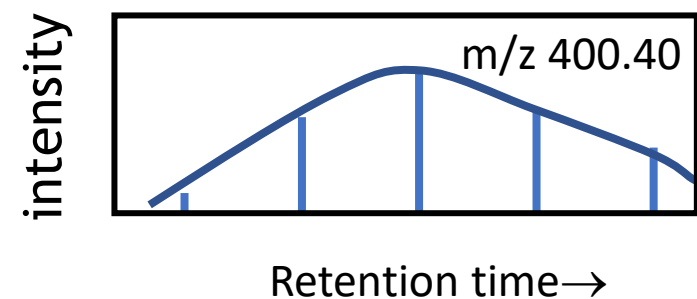
- Binning

- construct profile matrix



Take maximum at each time point

Extracted Ion Base Peak  
Chromatogram EIBPC



Extracted Ion Chromatogram EIC

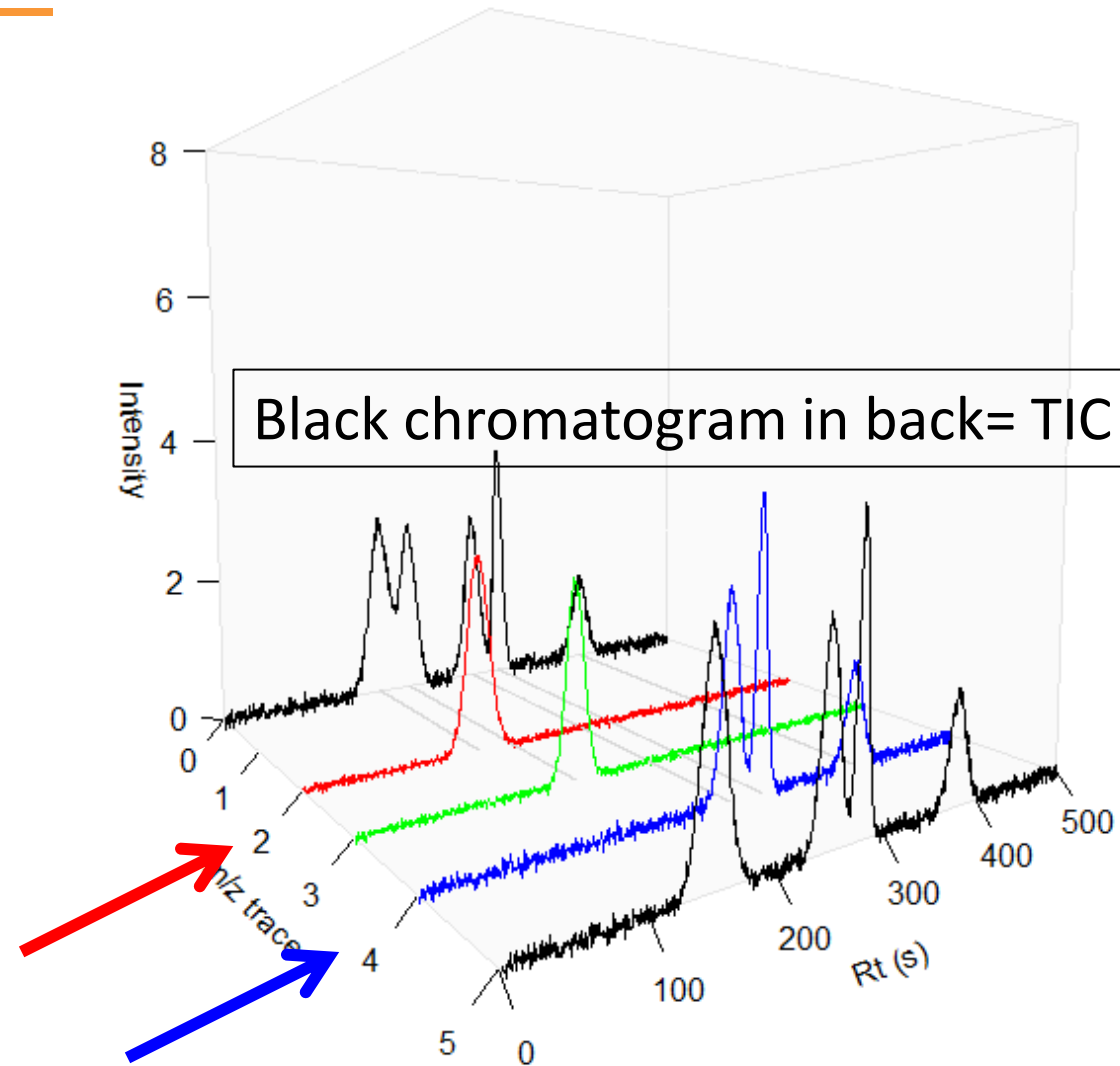
# Extracted Ion Chromatogram (EIC)

LC-MS spectrum



Reconstruct chromatogram  
for specific  $m/z$  value(s)

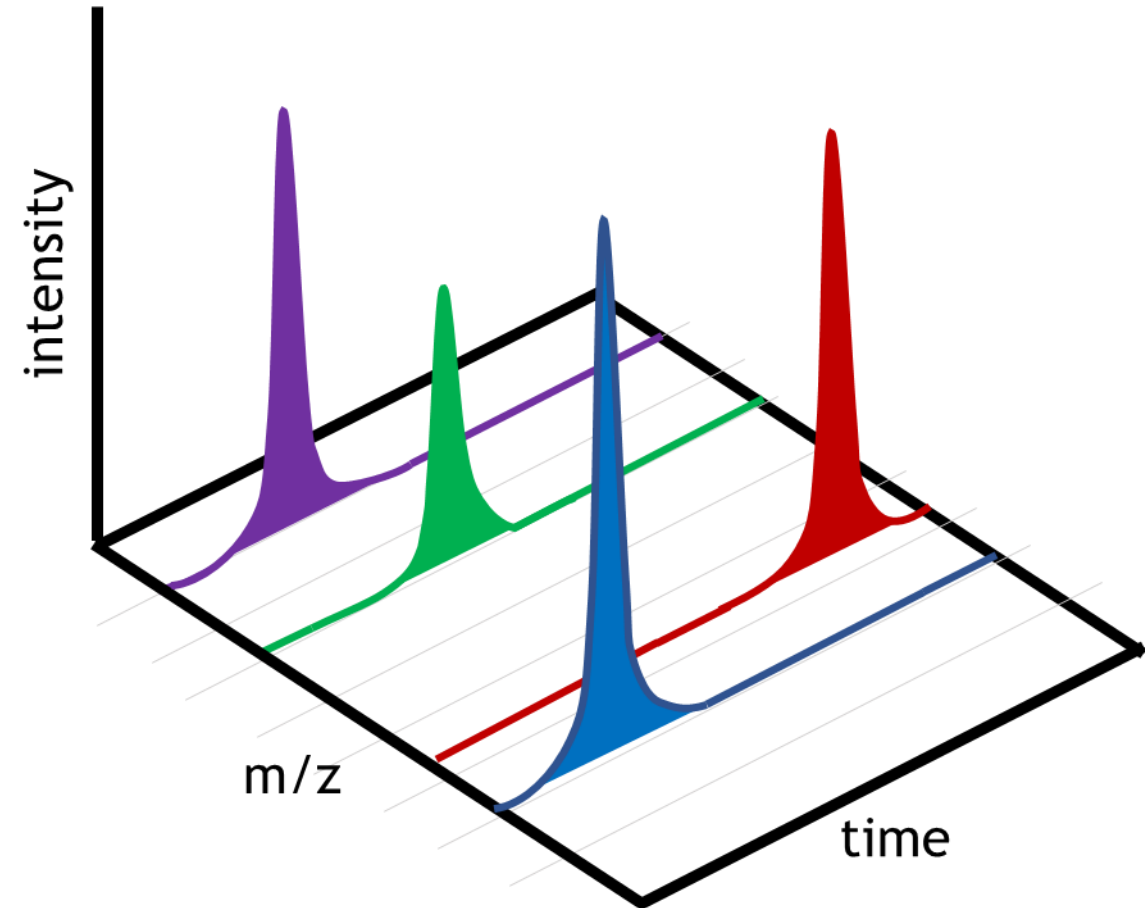
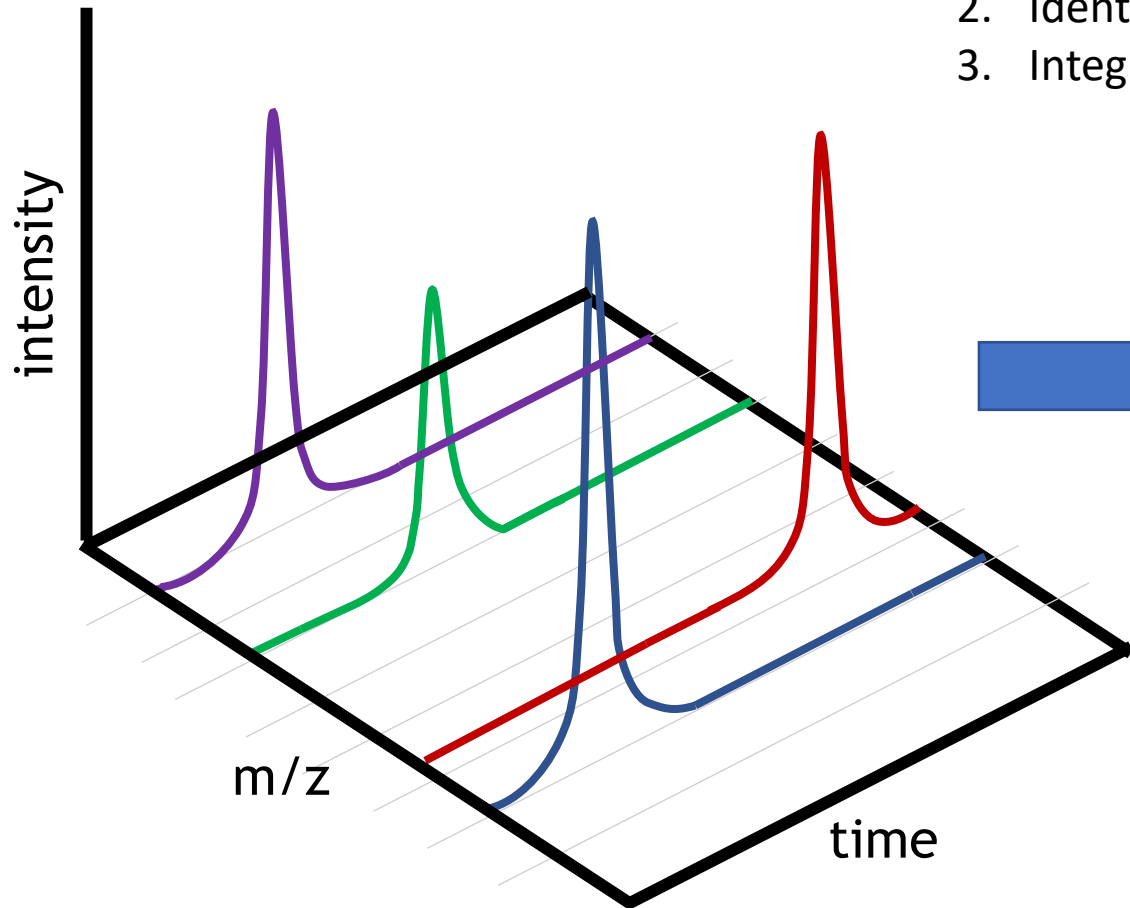
Example:  
Use RED and BLUE  $m/z$  values



# Preprocessing 2: Find Peaks



1. Filter data
2. Identify peaks
3. Integrate

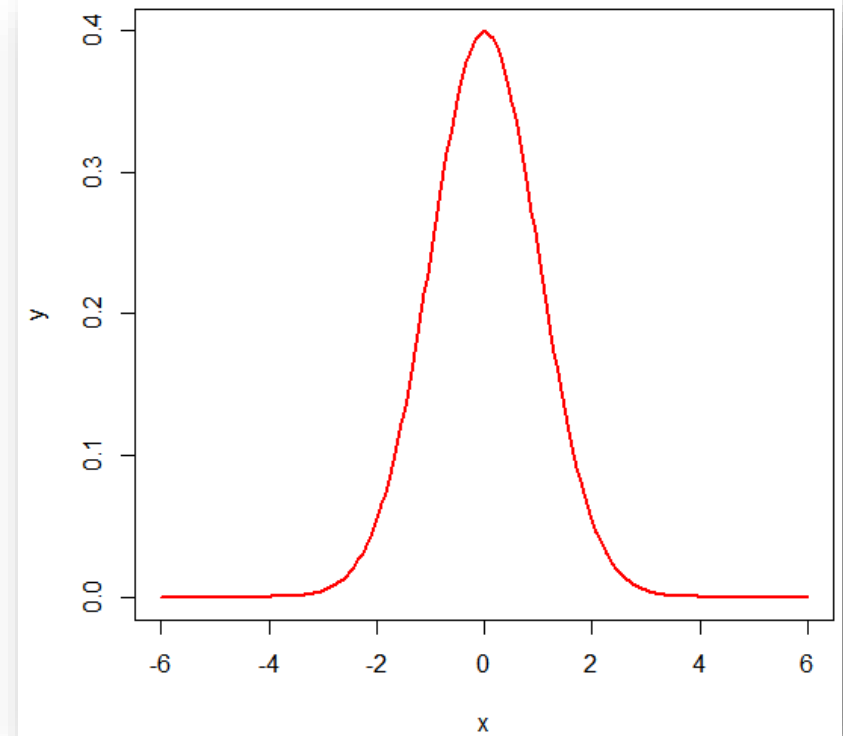




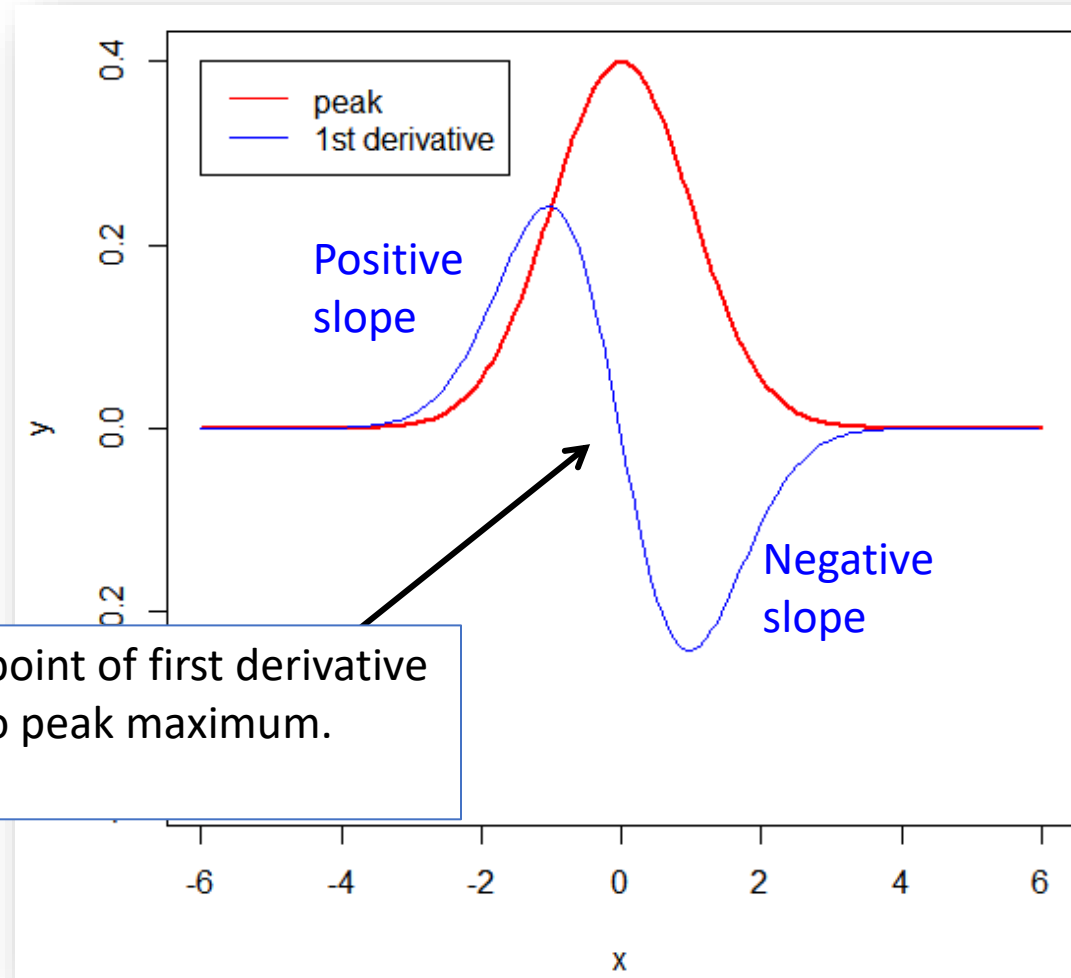
# Peak detection



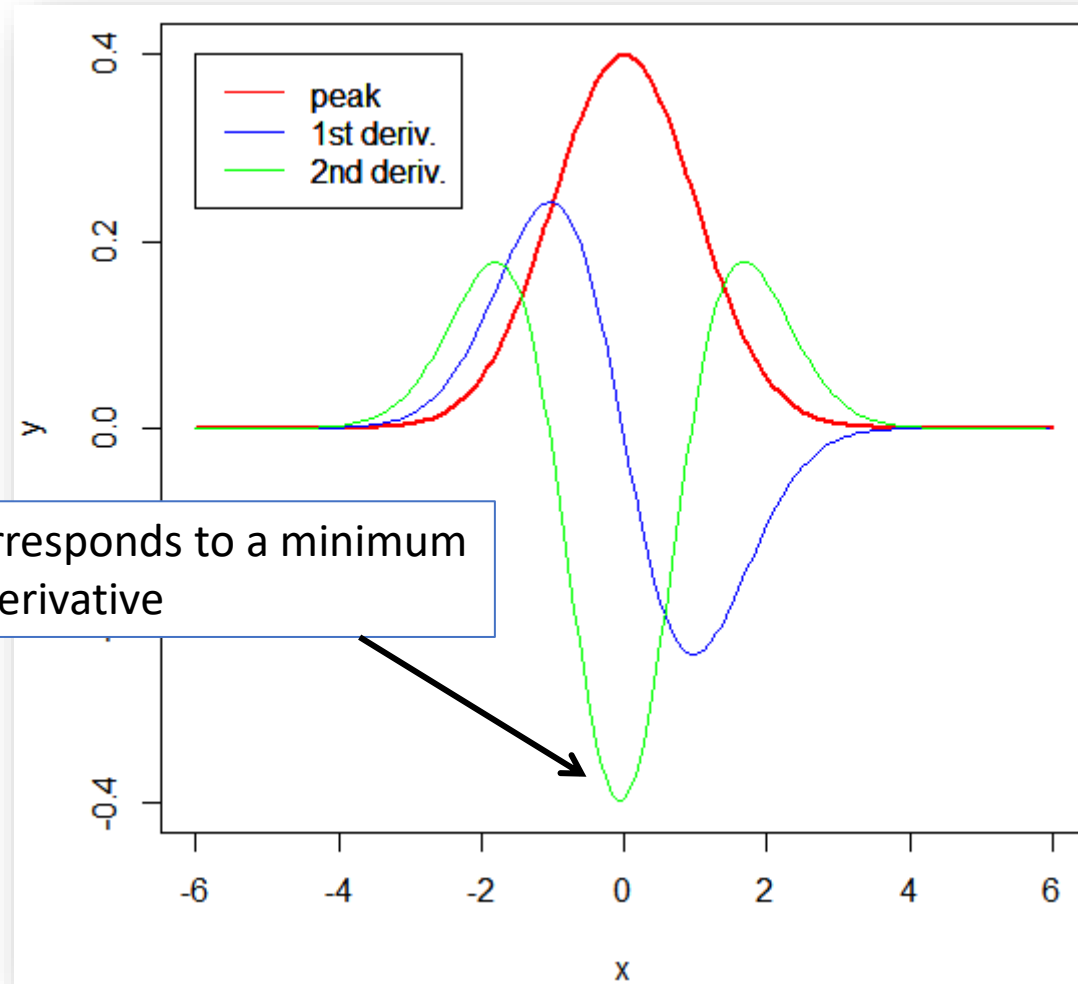
- Simple case
- How would you detect such peak?
  - Position
  - Start / End
  - Shape
- In general: peak detection is difficult.



# First derivative to find peak position



# Second derivative

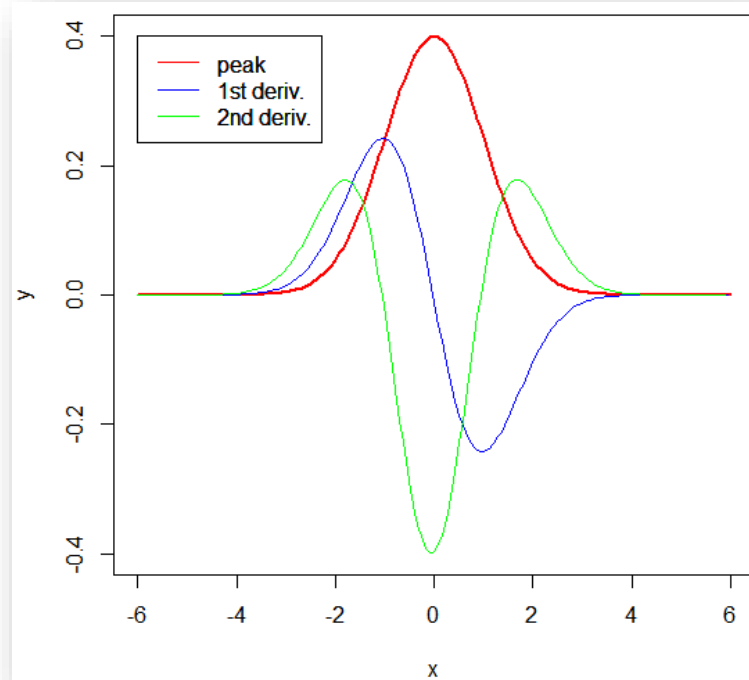


A maximum corresponds to a minimum in the second derivative

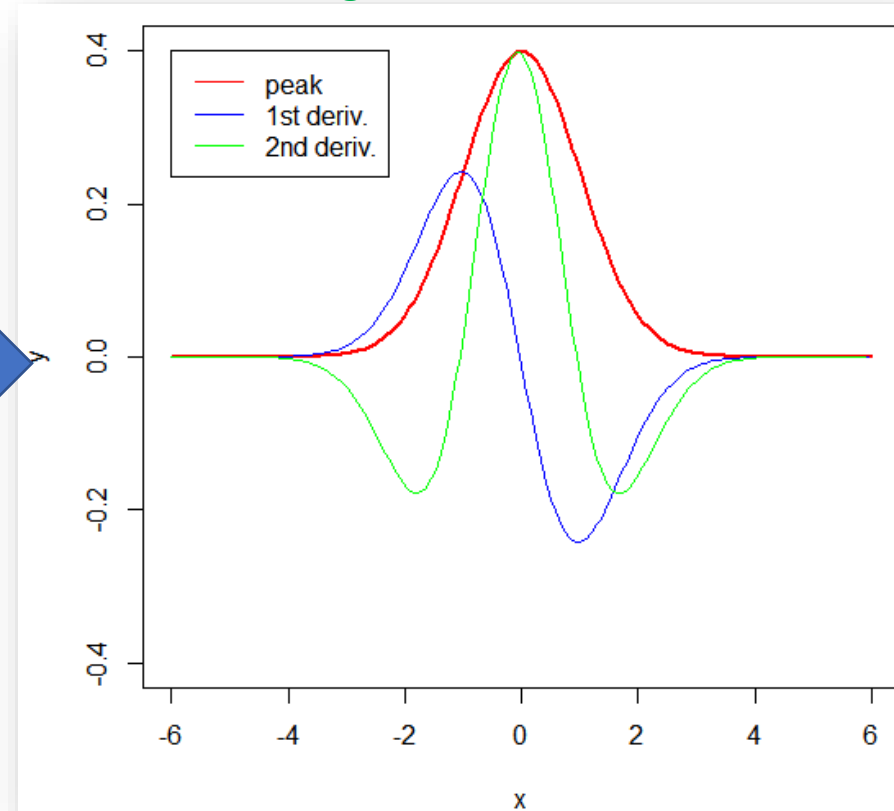
# Negative second derivative



Second derivative



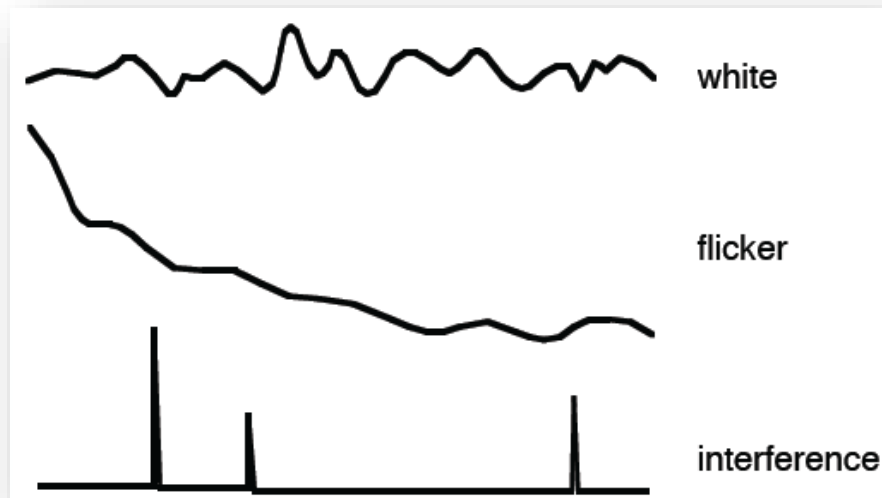
Negative second derivative



# Peak detection can be difficult



- Due to peak shapes, overlapping peaks, .....
- Due to artifacts
  - **White noise:** Random background
  - **Flicker (drift):** changes in response with changing operation or conditions
  - **Interference:** noise spikes of random occurrence and intensity



# Assignment



We have done a LC run for 11 minutes and observe the following peaks:

- 3 minutes; ion intensity = 200
- 5 minutes; ion intensity = 100
- 8 minutes; ion intensity = 220
- 10 minutes; ion intensity = 11

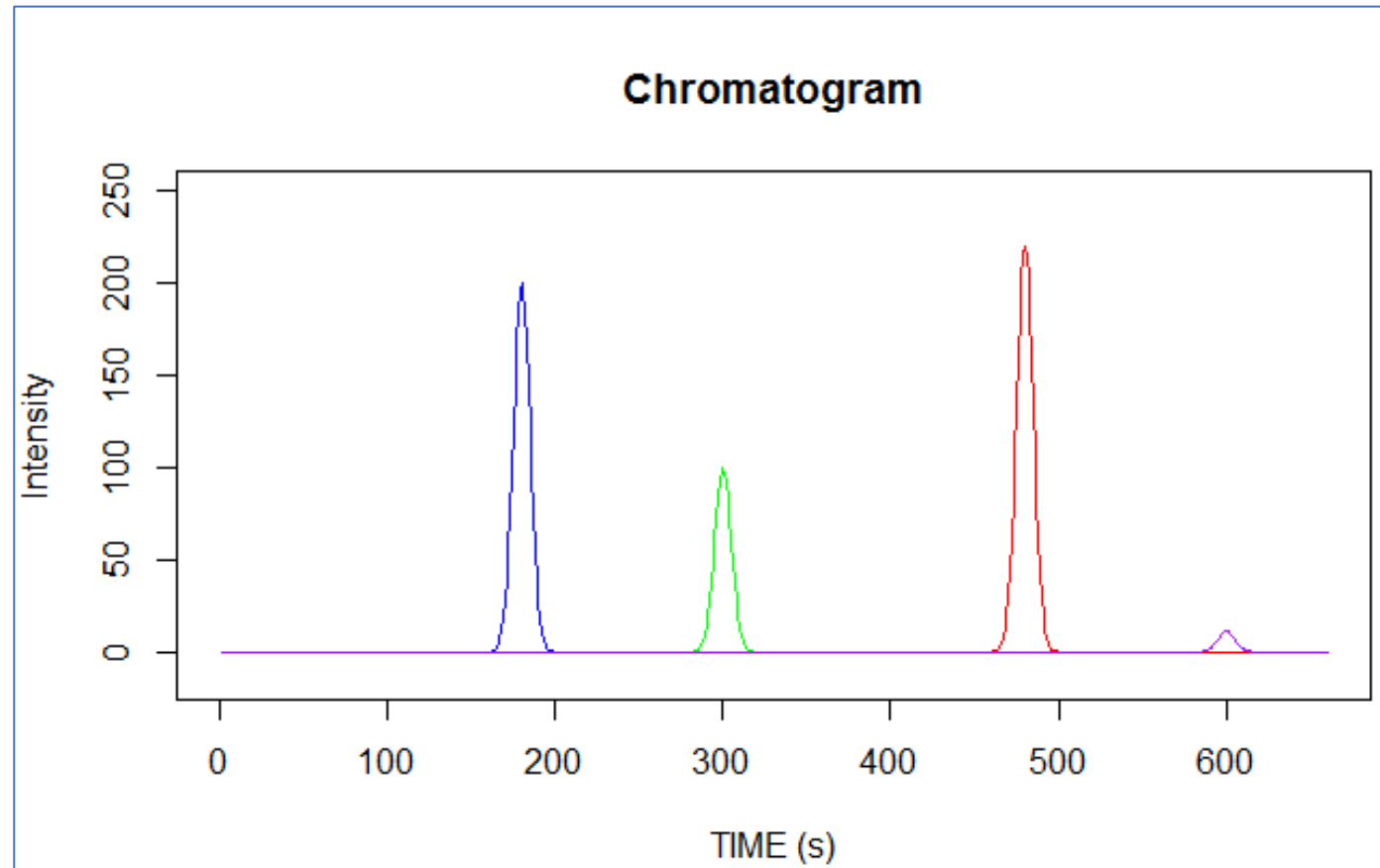
Draw the chromatogram

# Answer



We have done a LC run for 11 minutes and observe the following peaks:

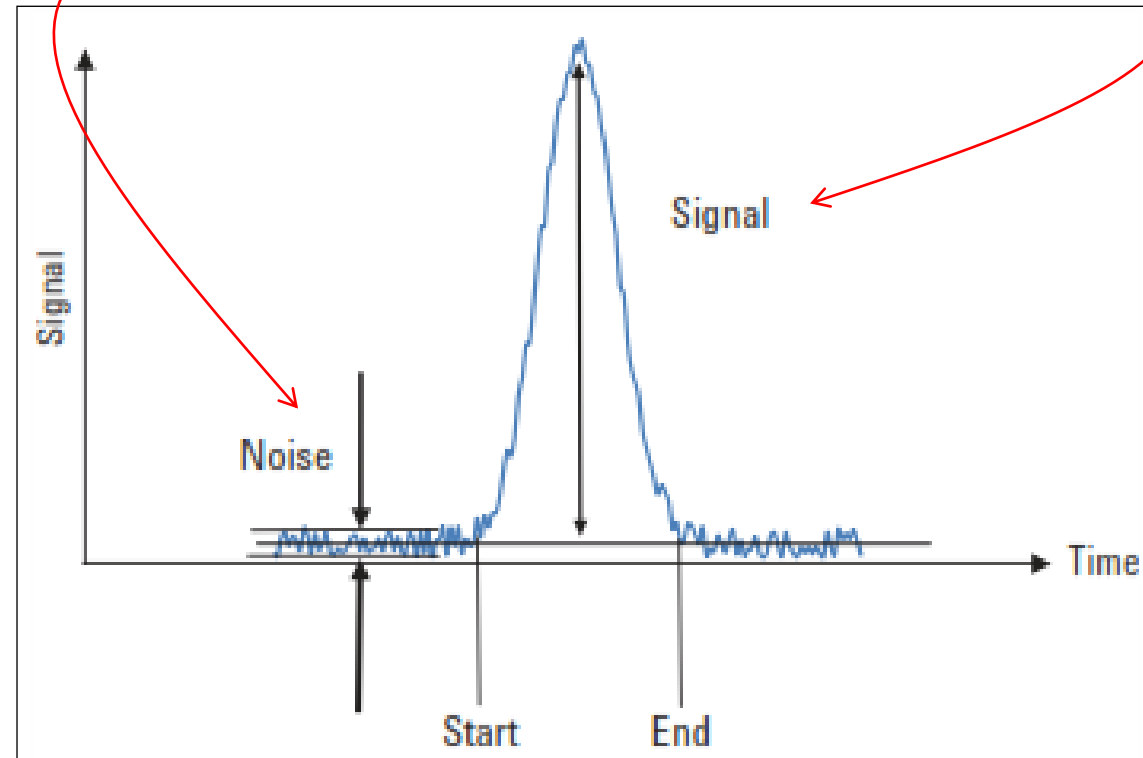
- 3 minutes; ion intensity = 200
- 5 minutes; ion intensity = 100
- 8 minutes; ion intensity = 220
- 10 minutes; ion intensity = 11



# Signal-to-noise ratio



$$S / N = \frac{\mu_S}{\sigma_N}$$





# Assignment



We have done a LC run for 11 minutes and observe the following peaks:

- 3 minutes; ion intensity = 200
- 5 minutes; ion intensity = 100
- 8 minutes; ion intensity = 220
- 10 minutes; ion intensity = 11

Constant noise fluctuations (-12 to +12)

~ standard deviation = 4

Draw the chromatogram

# Answer



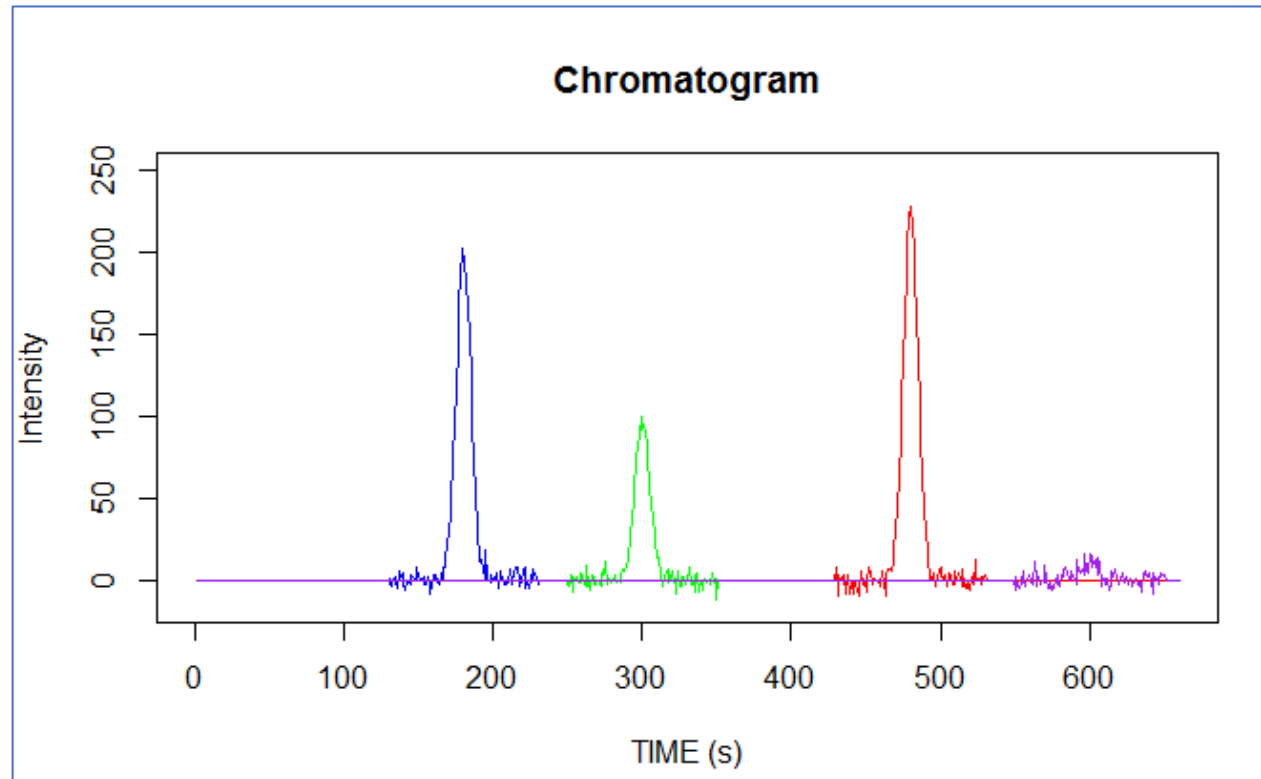
We have done a LC run for 11 minutes and observe the following peaks:

- 3 minutes; ion intensity = 200
- 5 minutes; ion intensity = 100
- 8 minutes; ion intensity = 220
- 10 minutes; ion intensity = 11

Constant noise fluctuations (-12 to +12)

~ standard deviation = 4

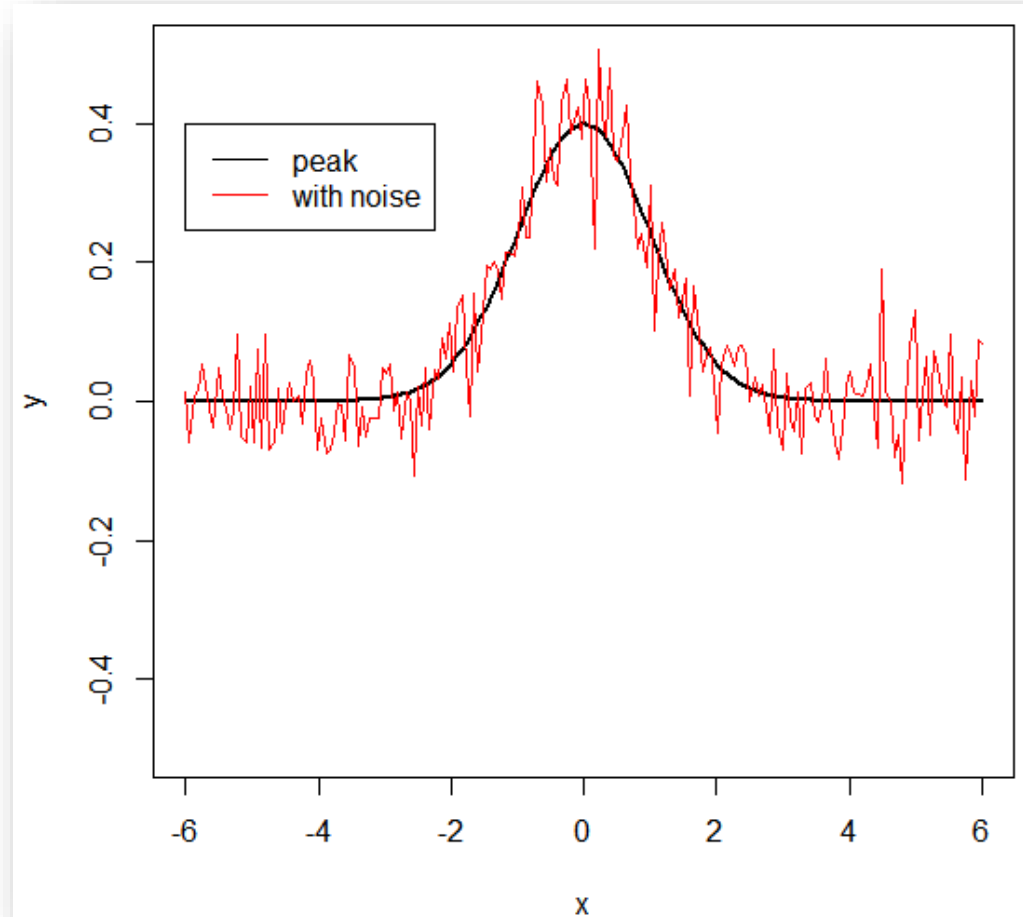
Notice: difficult to distinguish last peak from noise



# Our peak (black) with some noise added (red)



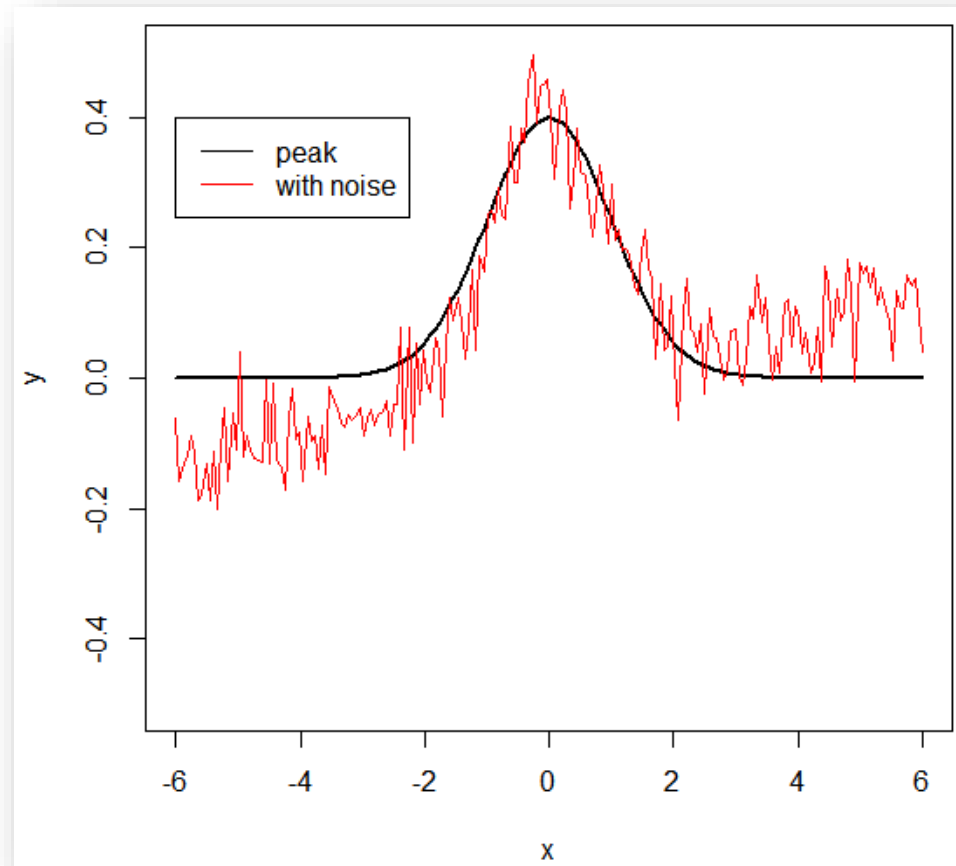
- Original peak without noise (black)
- Peak+Noise (red)



# Our peak (black) with some noise and baseline added (red)



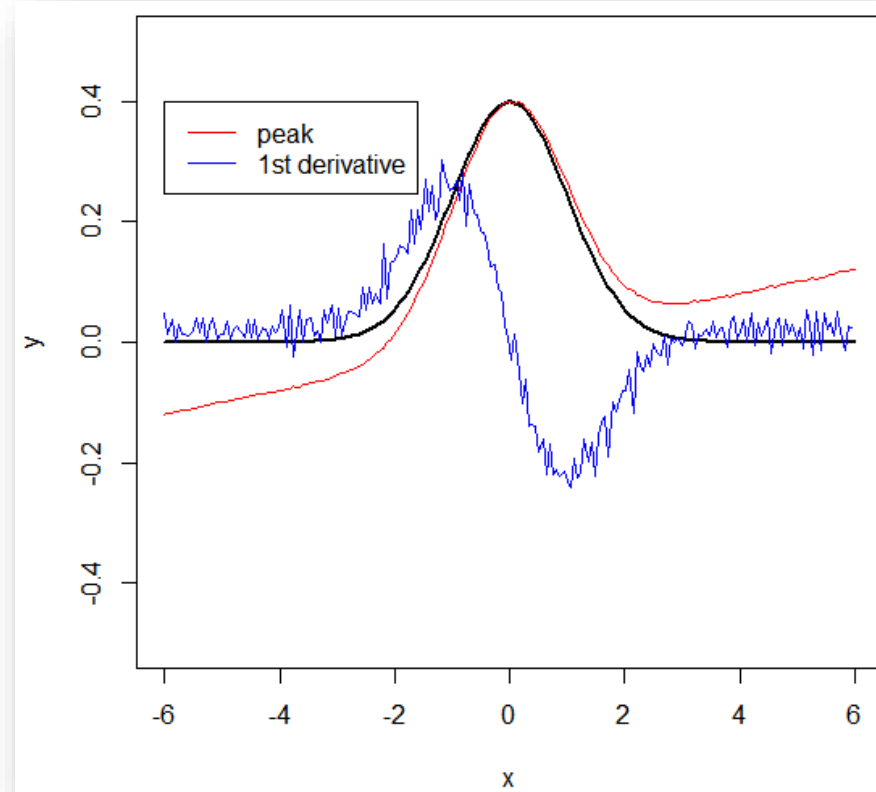
- Original peak without noise (black)
- Peak+Noise+Baseline (red)



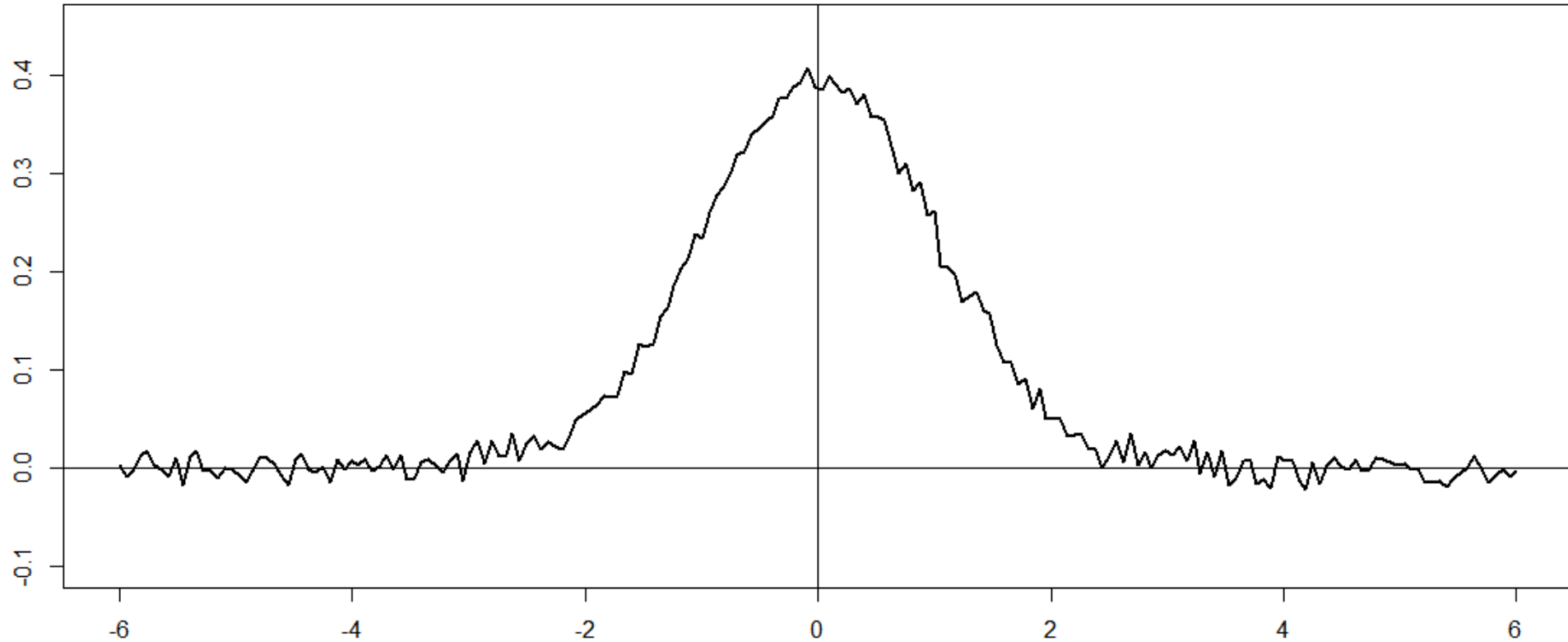
# Derivatives are sensitive to noise



- A small amount of noise present on the red line
- Has large effect on the first derivative which becomes much more noisy.
- First derivative also corrects the linear baseline



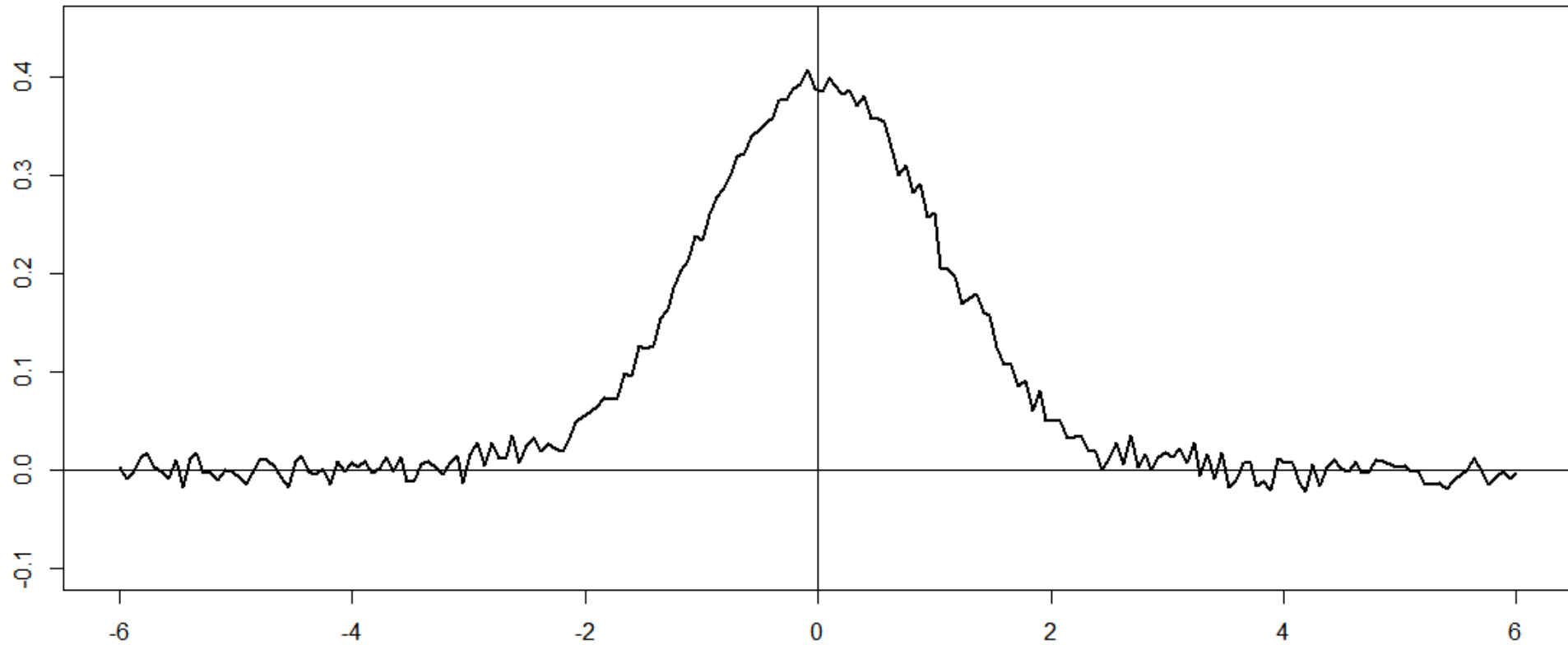
# How would you remove the noise on this peak?



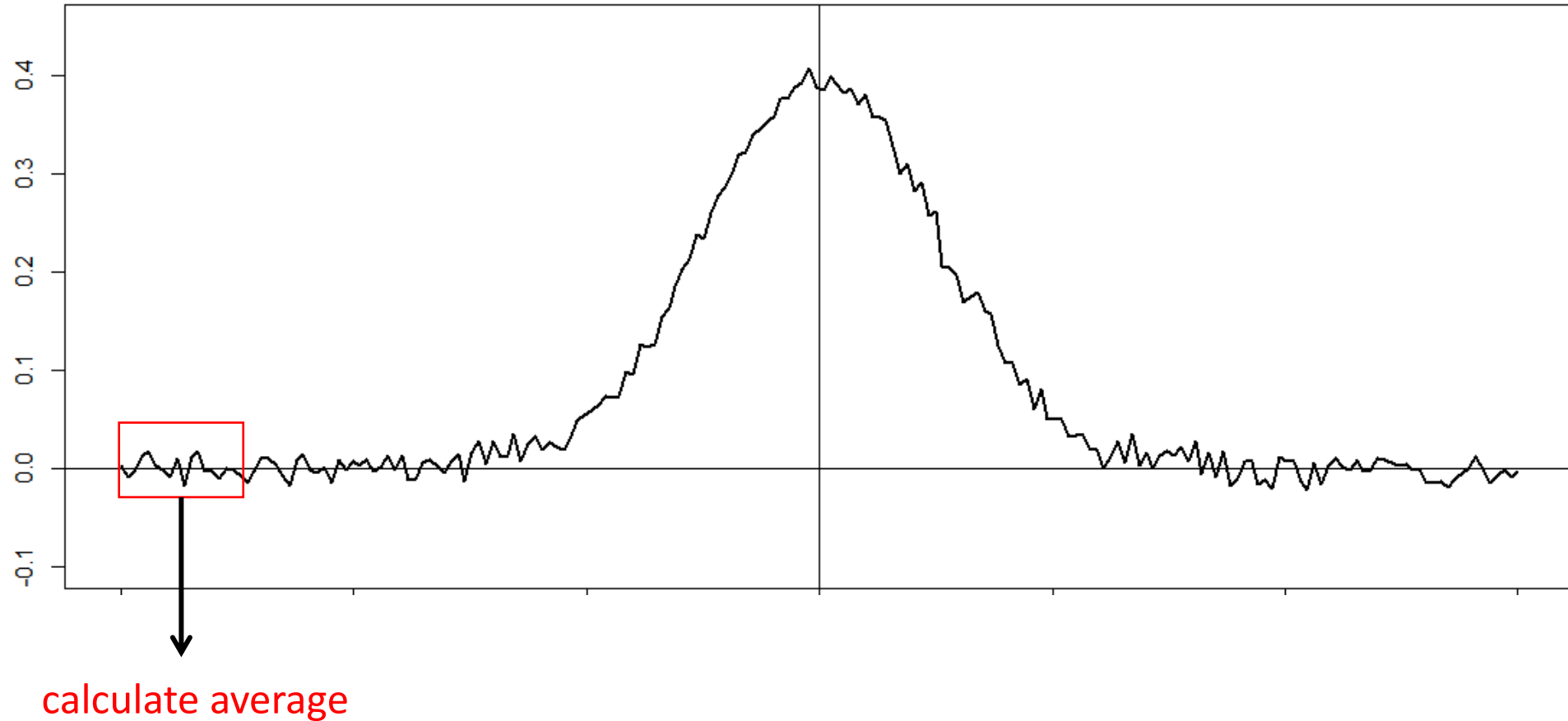
# Use smoothing



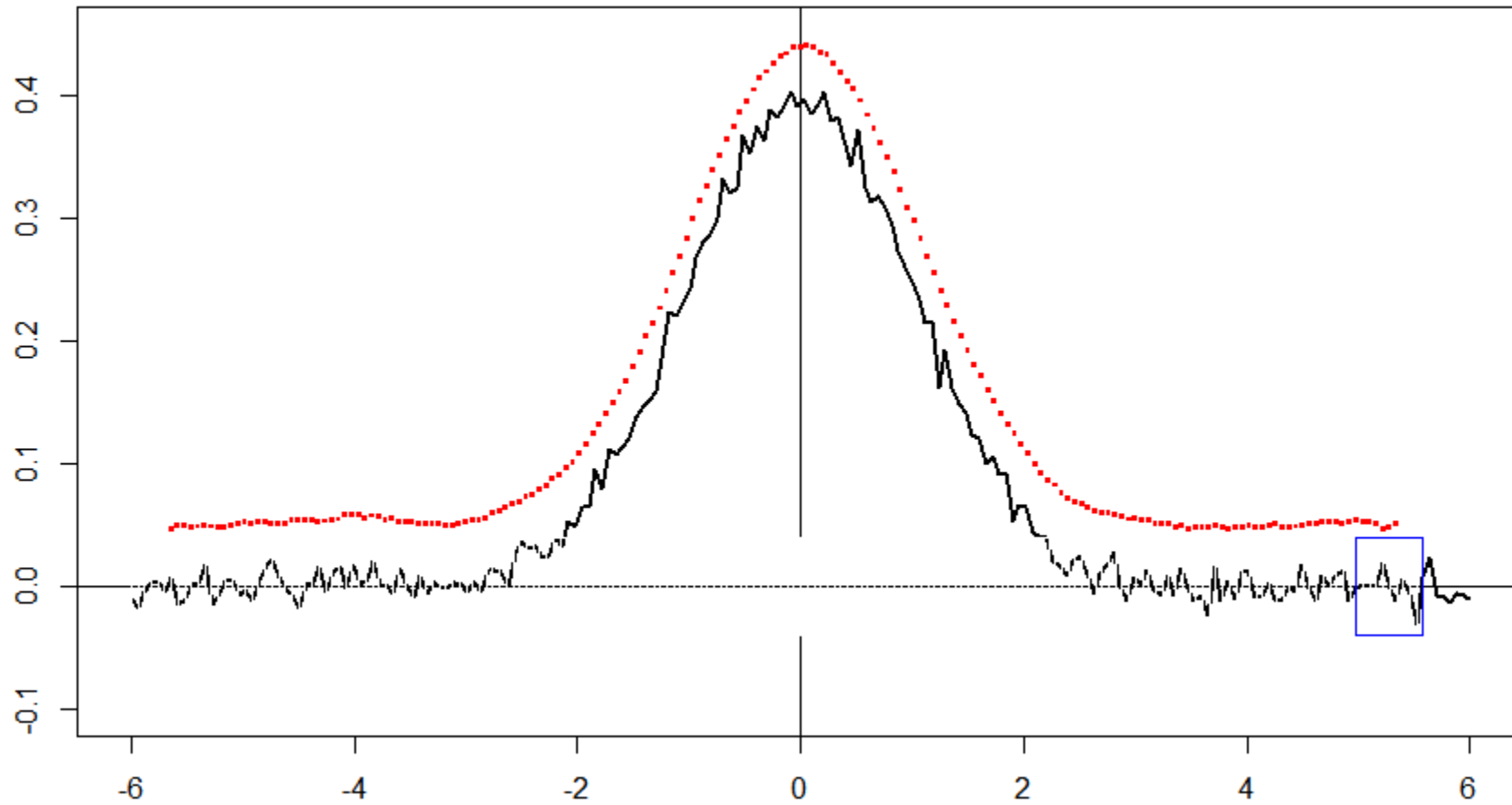
**Moving average** is one approach towards smoothing



# Moving average





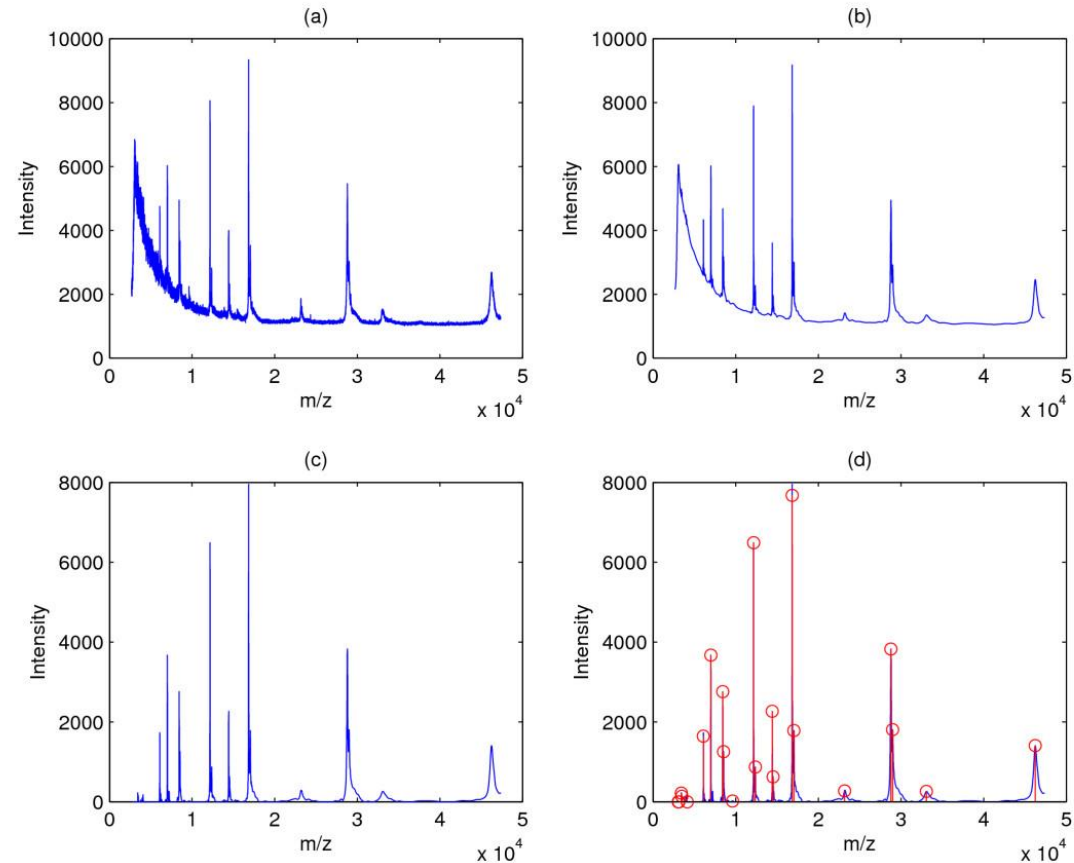


- Next you can differentiate the smoothed curve.

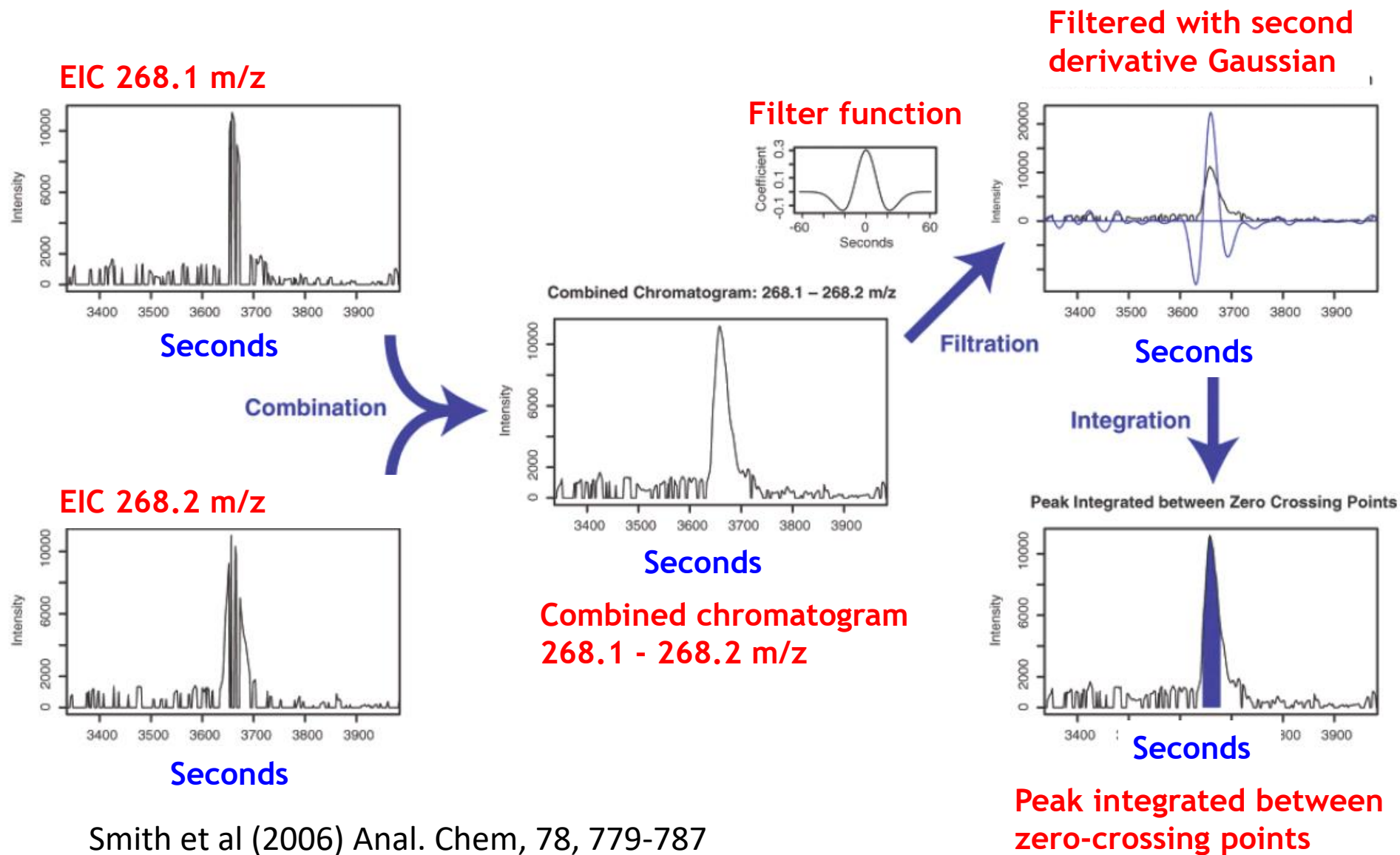
# Peak detection steps



- Raw data
- Smoothed data
- Baseline corrected and smoothed data
- Detected peaks



# Peak detection with xcms: overview



# Result of peak detection



	mz	mzmin	mzmax	rt	rtmin	rtmax	into	intf	maxo	maxf	i	sn
[1,]	200.1000	200.1	200.1	2928.610	2912.961	2942.695	147887.53	290506.9	9687	15899.054	1	13.40257
[2,]	201.0638	201.0	201.1	2531.112	2515.463	2549.892	204572.42	280386.0	7726	13300.725	1	12.13748
[3,]	205.0000	205.0	205.0	2784.635	2770.550	2800.284	1778568.94	3610059.7	84280	195026.431	1	70.14981
[4,]	205.9819	205.9	206.0	2786.200	2772.115	2800.284	237993.62	448580.0	10681	23860.099	1	31.89939
[5,]	207.0821	207.0	207.1	2712.647	2698.562	2726.731	380873.05	730980.9	18800	40065.736	1	23.89175
[6,]	208.0671	208.0	208.1	2640.659	2625.009	2656.308	96070.72	150033.4	4112	7560.078	1	12.67485
[7,]	208.1201	208.1	208.2	2711.082	2698.562	2726.731	67967.10	126124.7	2878	7029.903	1	12.18258
[8,]	219.0848	219.0	219.1	2518.593	2504.508	2534.242	235544.92	422133.1	11588	22499.345	1	16.51047
[9,]	229.1000	229.1	229.1	2515.463	2502.943	2529.547	87236.08	198958.4	6649	11782.670	1	10.18295
[10,]	233.0390	233.0	233.1	3019.378	3005.293	3033.462	399145.34	749332.4	19752	41554.242	1	27.47936

m/z

Retention time

Peak intensities  
raw / filtered data

Peak area  
raw / filtered data

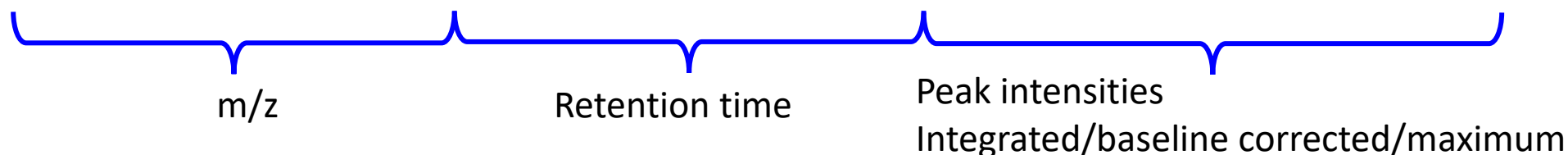
# Result of peak detection High Resolution Data



- Long list of peaks for each individual sample with following information:

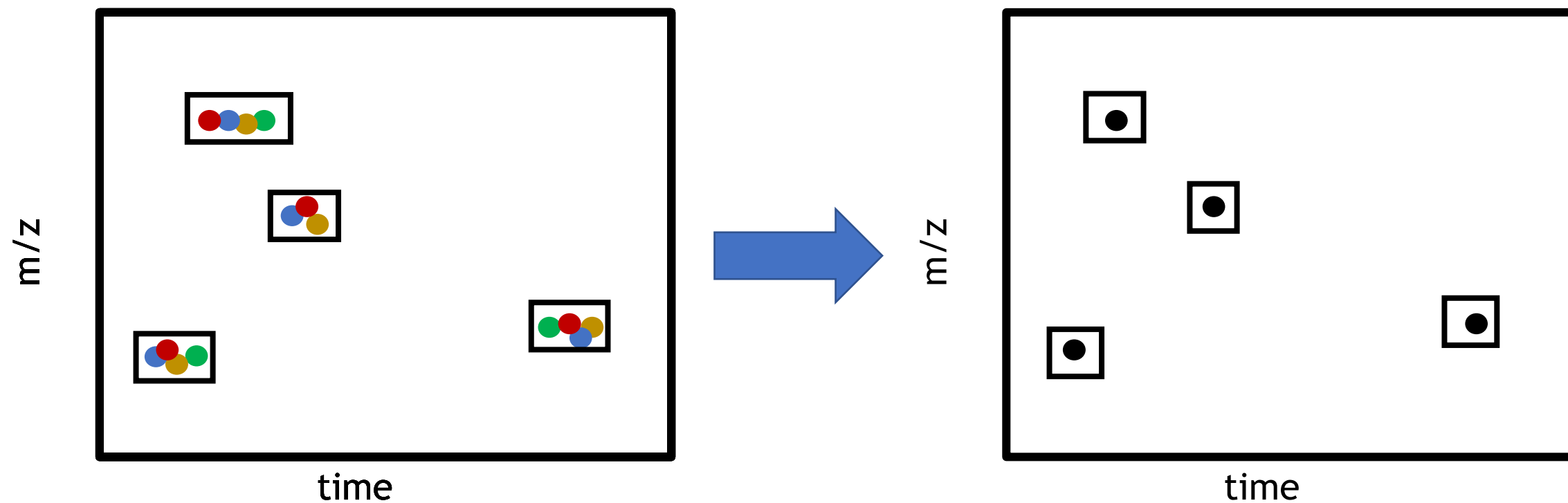
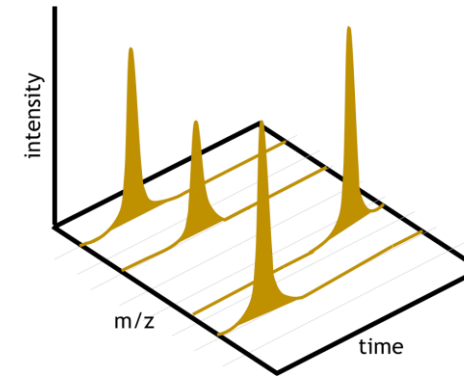
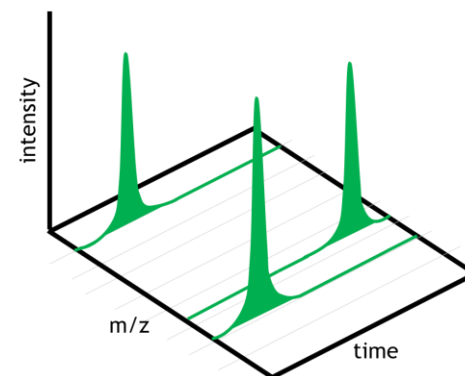
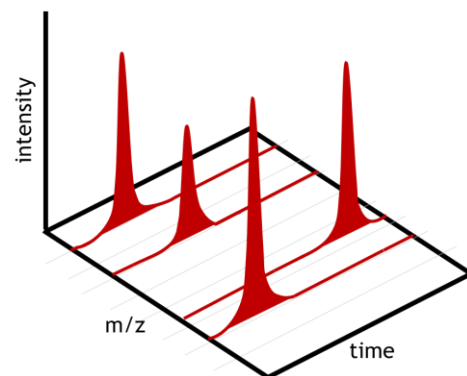
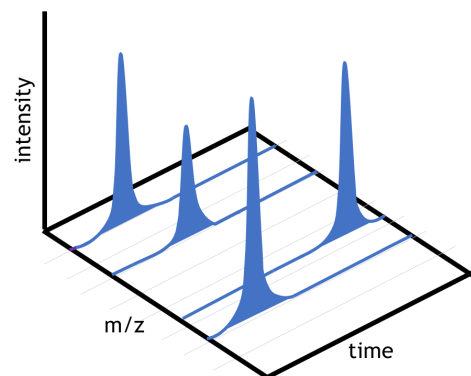
- mz, mzmin, mzmax
- rt, rtmin, rtmax
- peak intensities = areas
- s/n ratio

	mz	mzmin	mzmax	rt	rtmin	rtmax	into	intb	maxo	sn	sample
[1,]	204.1136	204.1134	204.1138	64.6471	62.46690	67.9266	3.373765e+04	3.329312e+04	1.128980e+04	19	1
[2,]	221.0808	221.0805	221.0809	66.8313	62.46690	90.7837	1.607109e+05	1.409400e+05	5.519782e+04	23	1
[3,]	228.0085	228.0081	228.0085	63.5487	62.46690	67.9266	2.836644e+05	2.836601e+05	2.428473e+05	242846	1
[4,]	230.1864	230.1861	230.1864	63.5487	62.46690	67.9266	1.219324e+05	1.219280e+05	9.490152e+04	264	1
[5,]	233.1900	233.1898	233.1902	64.6471	62.46690	73.3863	8.568772e+04	8.315198e+04	2.883581e+04	24	1
[6,]	236.0740	236.0737	236.0741	63.5487	62.46690	72.3009	3.093353e+05	3.060179e+05	2.188165e+05	105	1
[7,]	236.1493	236.1490	236.1494	64.6471	62.46690	71.2131	7.780719e+04	7.182568e+04	3.336154e+04	16	1
[8,]	238.1550	238.1548	238.1552	65.7395	62.46690	74.4725	5.730793e+04	5.484125e+04	1.150864e+04	9	1
[9,]	239.0586	239.0583	239.0587	64.6471	62.46690	67.9266	1.032282e+05	1.032239e+05	4.870609e+04	48705	1
[10,]	240.2322	240.2319	240.2323	66.8313	62.46690	91.8599	1.473217e+05	1.375371e+05	5.350293e+04	27	1

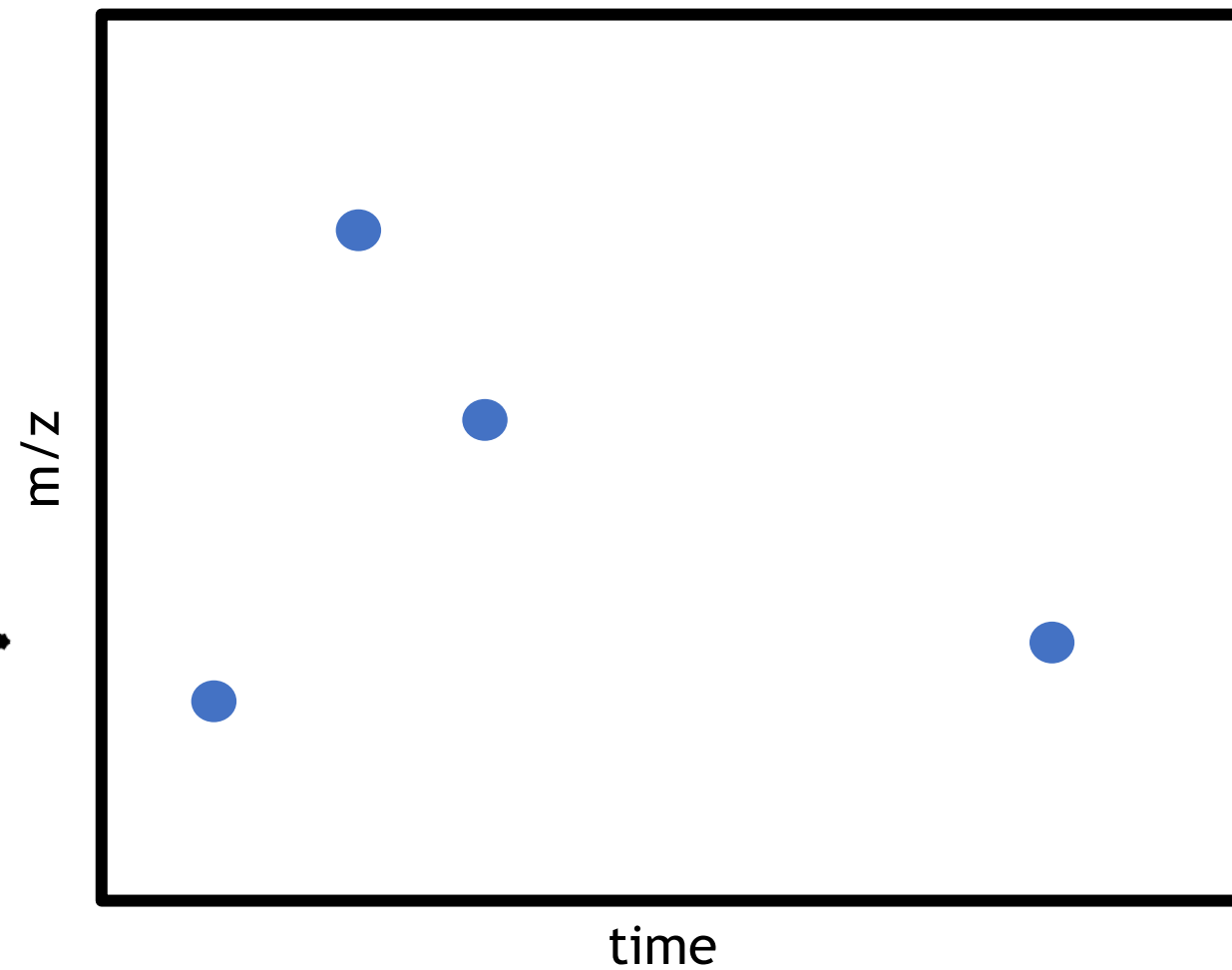
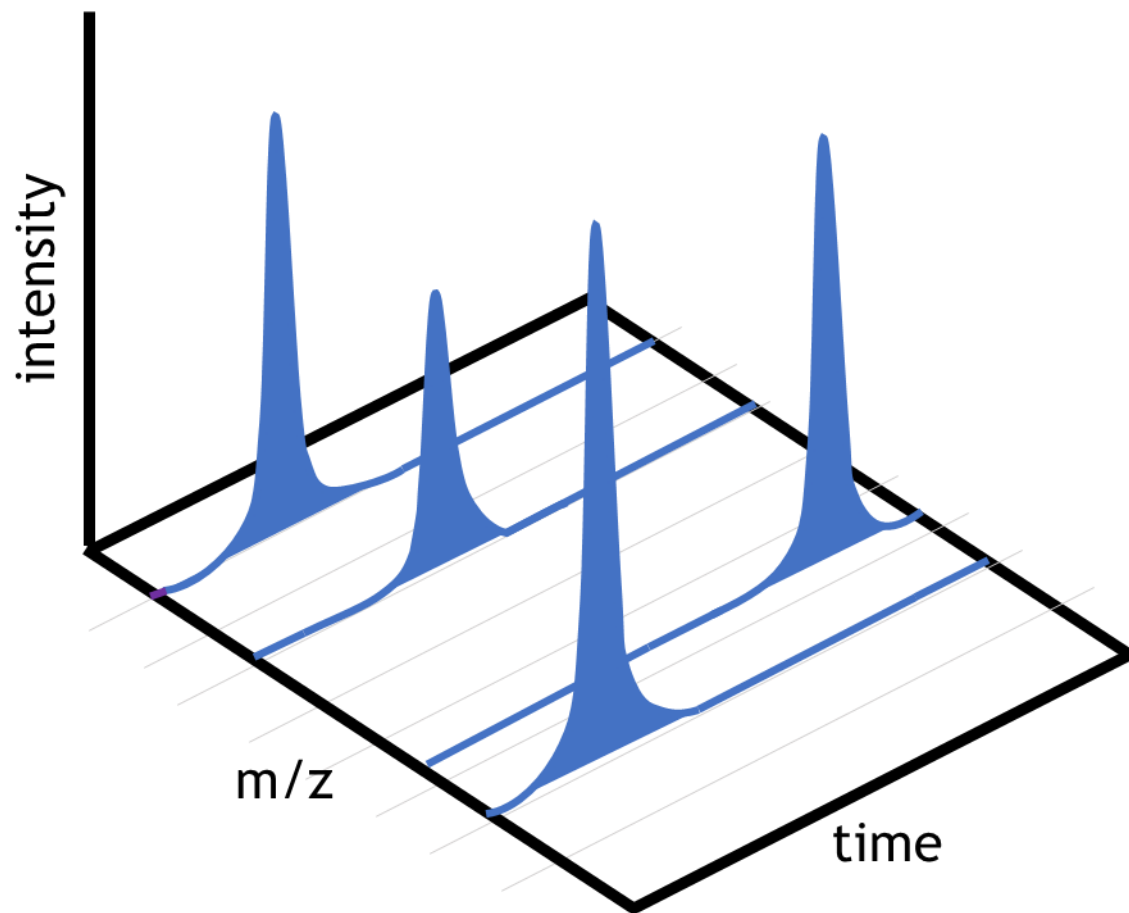


- Note: we obtain a peak table for every individual sample
  - xcms stores them in one big indexed list

# Preprocessing 3: Match Peaks / RT correction



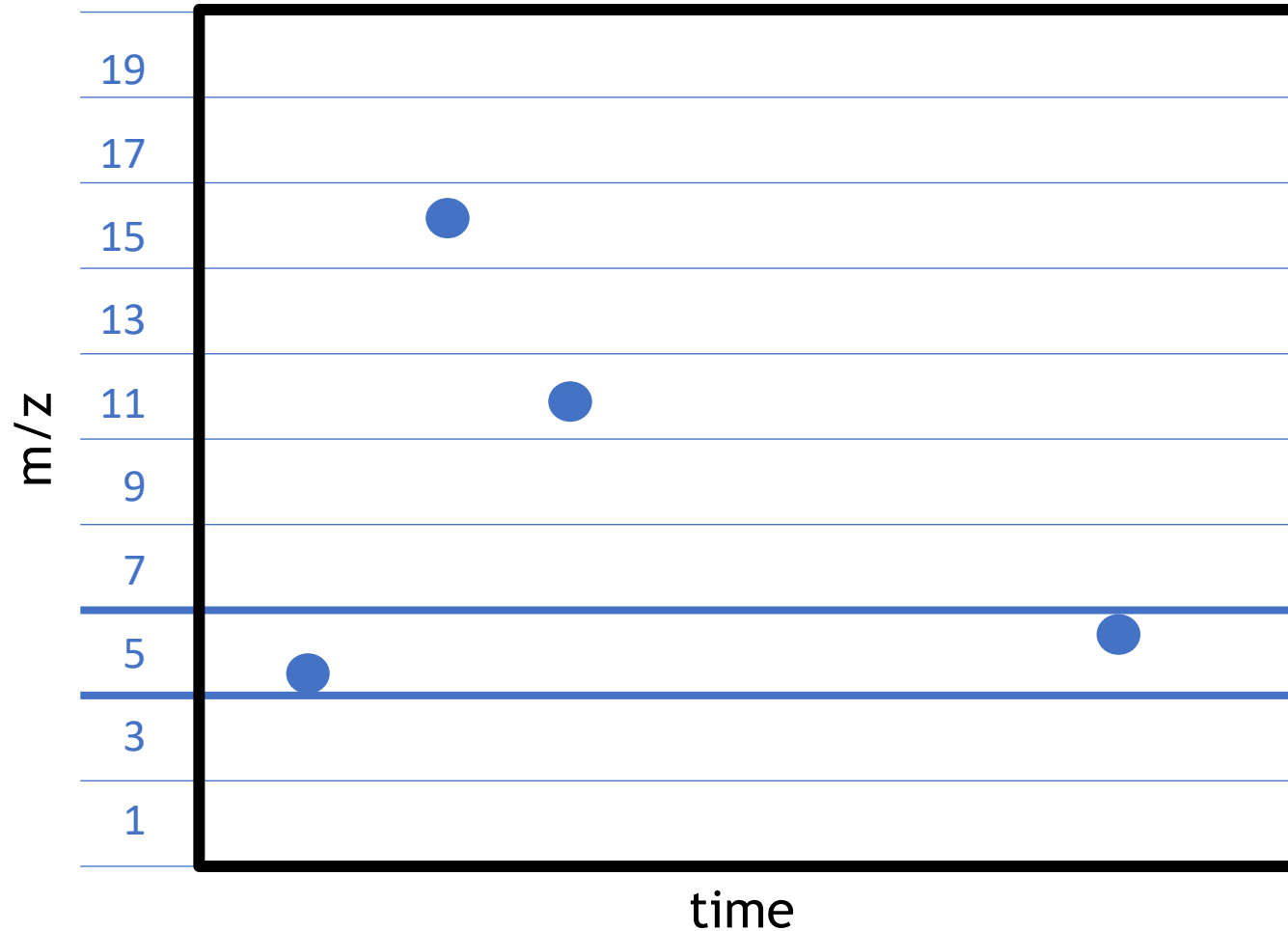
# Top view: only peak positions



# Matching $m/z$ peaks of across samples



bin 5 contains two metabolites with about the same  $m/z$  but at two different retention times

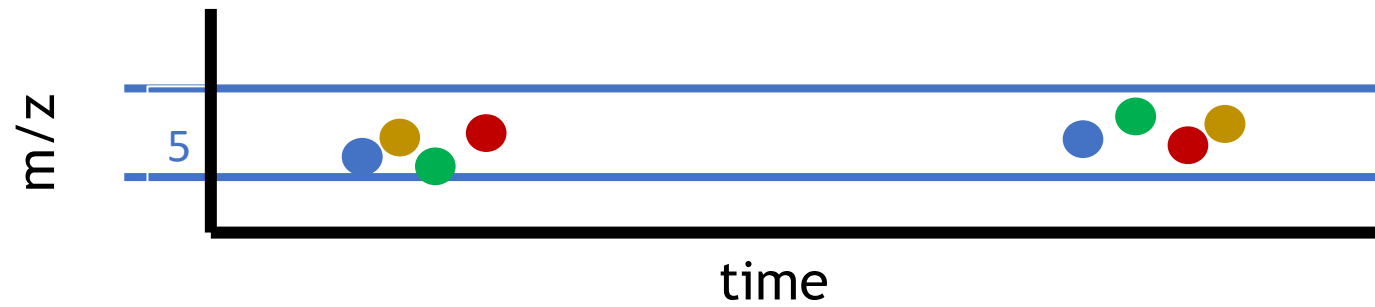




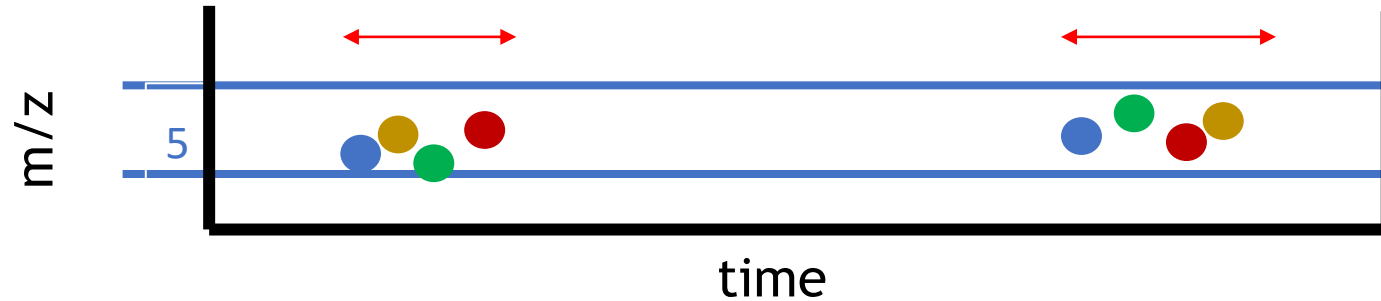
# Matching $m/z$ peaks across samples



Showing only bin 5 for 4 samples



# Matching $m/z$ peaks across samples



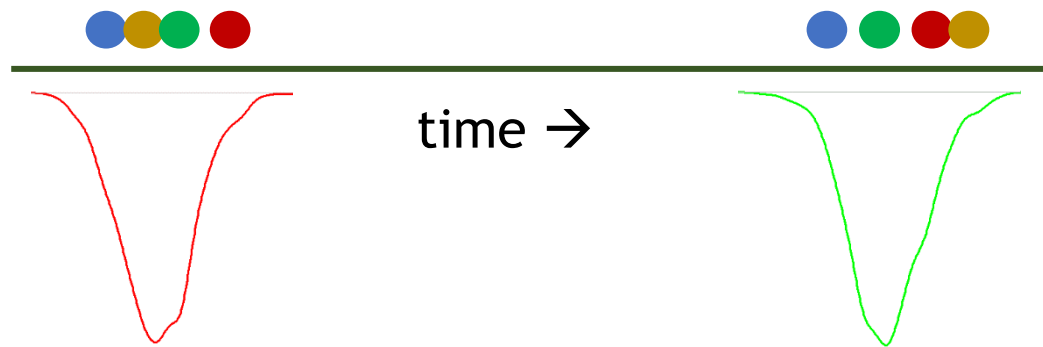
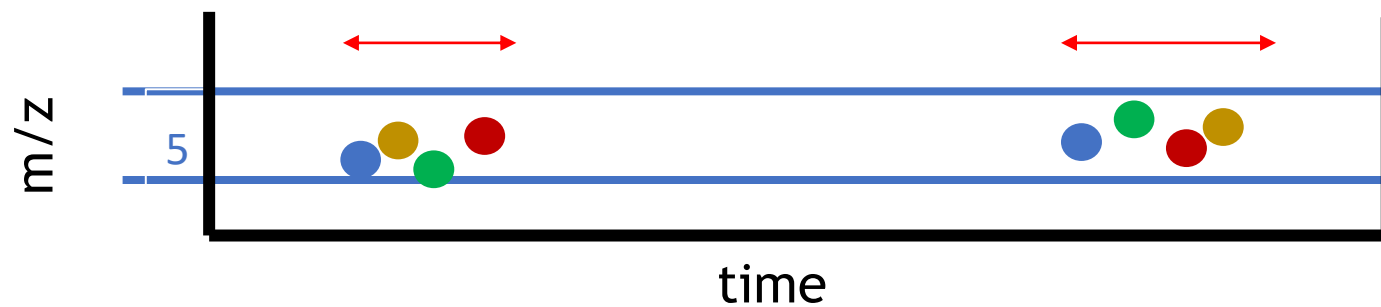
Bin 5 contains two metabolites with about the same  $m/z$  but at two different retention times

Within the bin there is a small variation of the mass

Also the retention times are slightly variable (for both the grey and yellow metabolites)

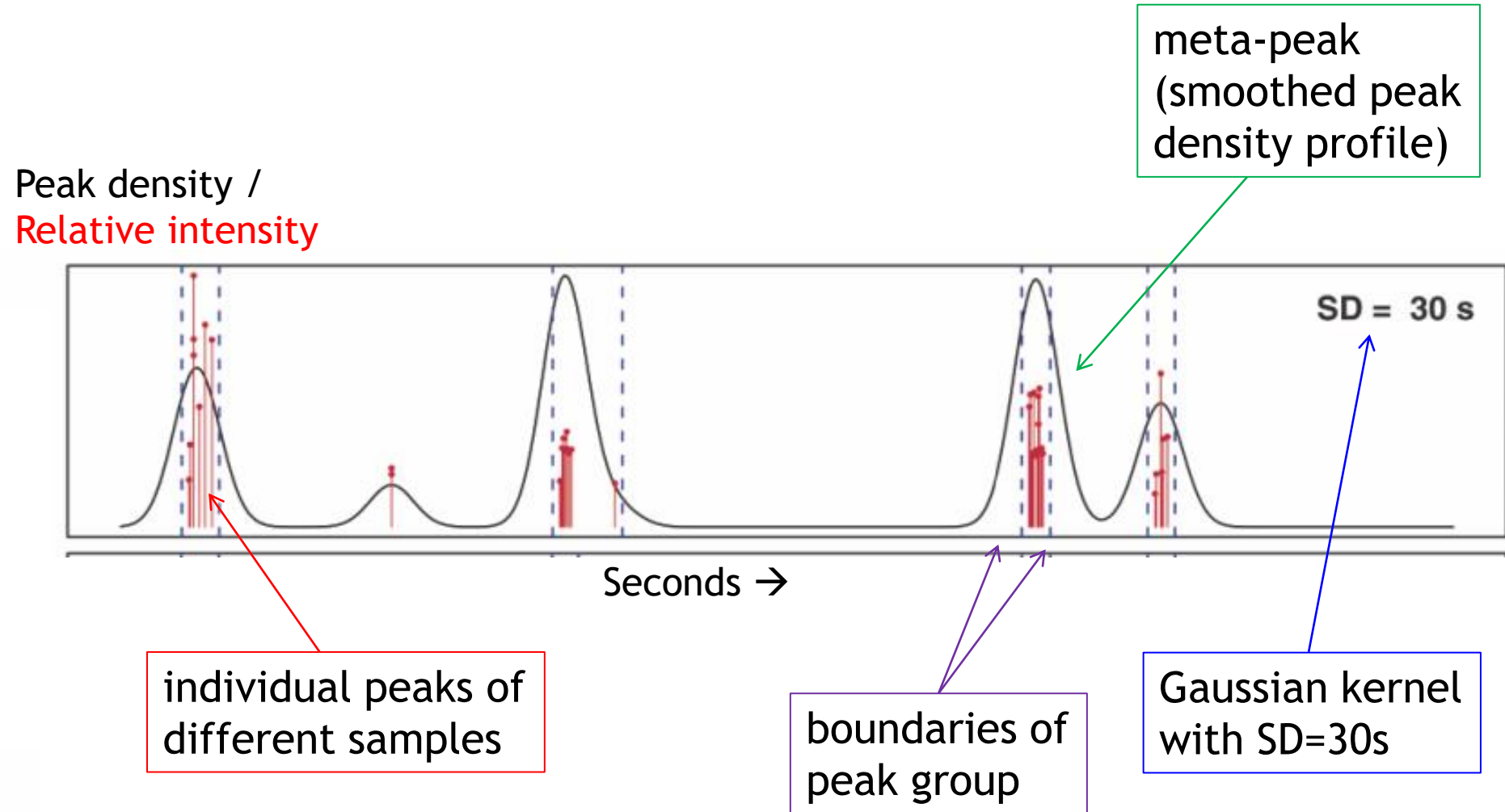
In the first step these peaks are grouped together

# Matching $m/z$ peaks across samples

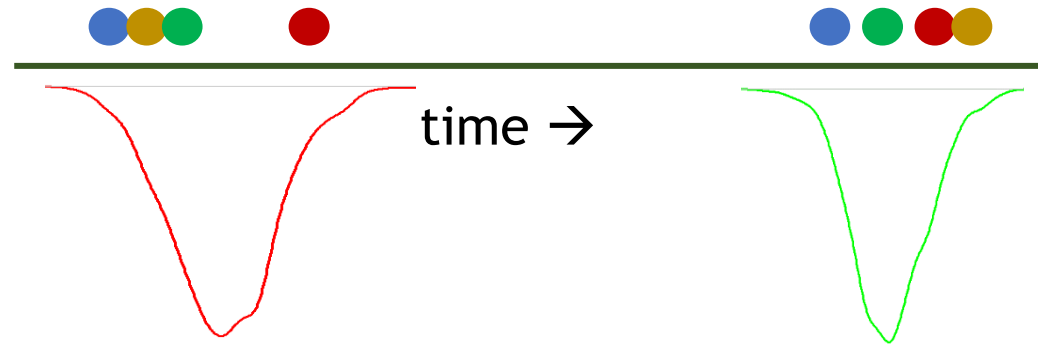
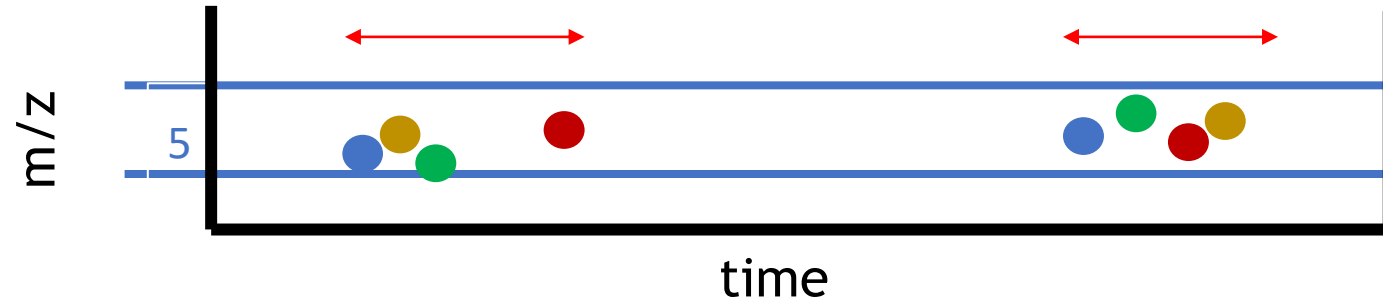


kernel density estimation

# Peak group boundaries

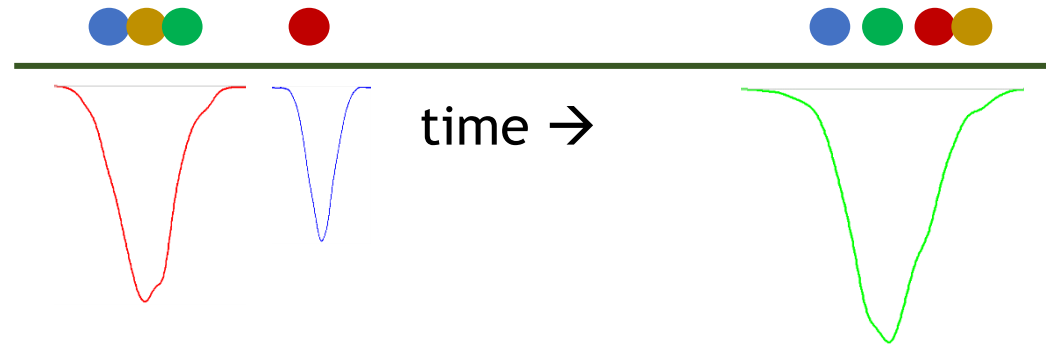
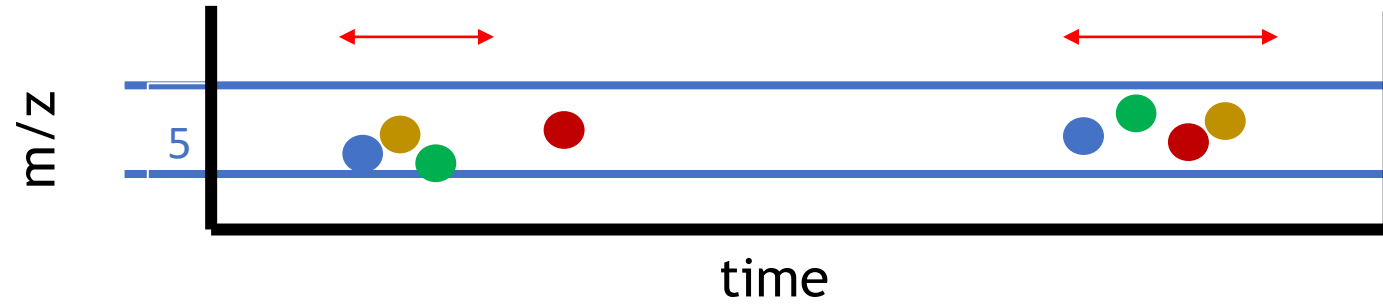


# When is peak an outlier?



Should the red point still be part of peak group?

# When is peak an outlier?

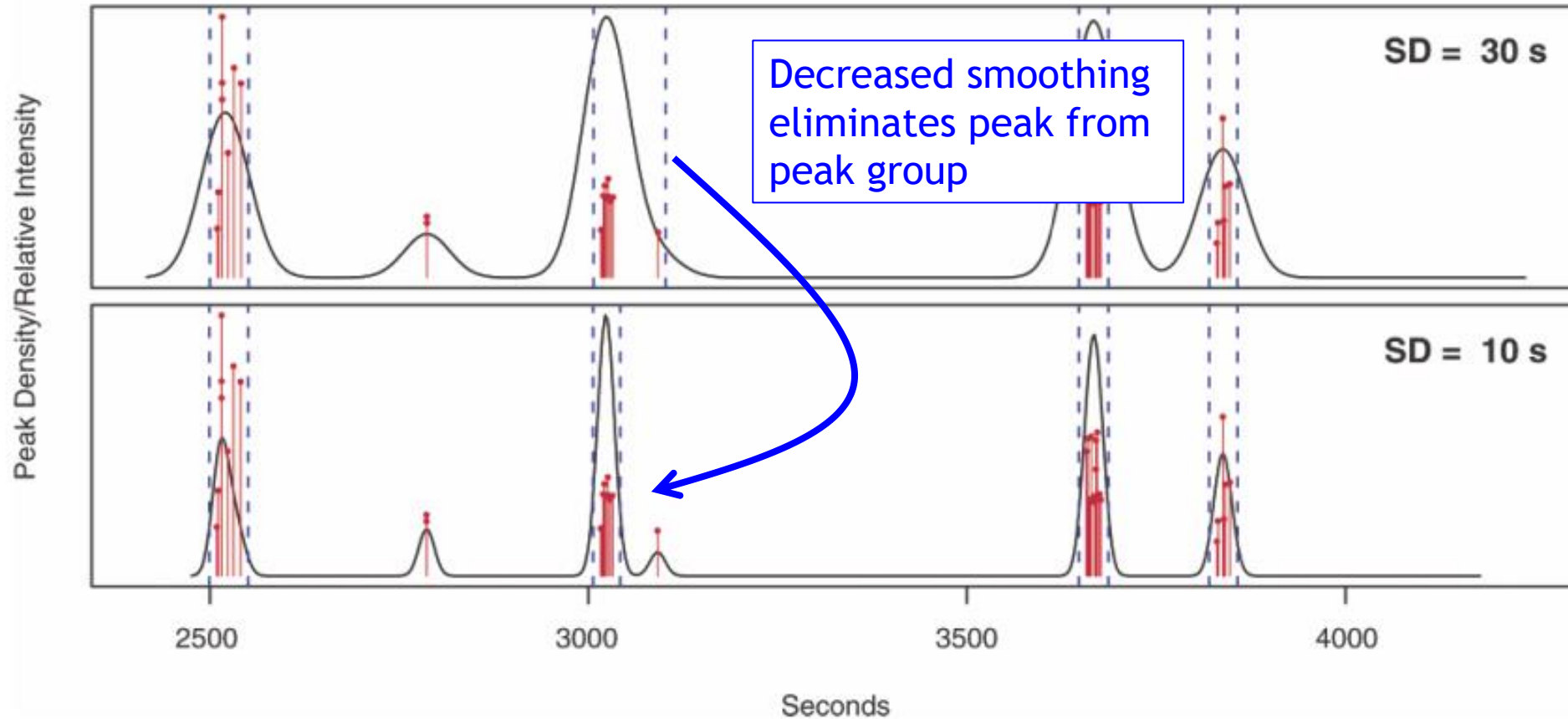


Or is it a separate peak group?

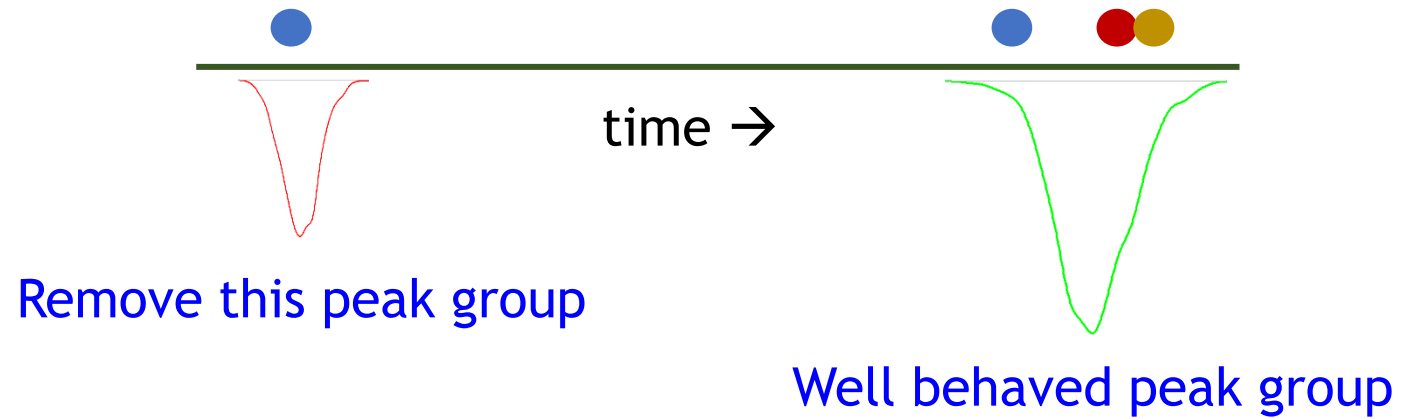
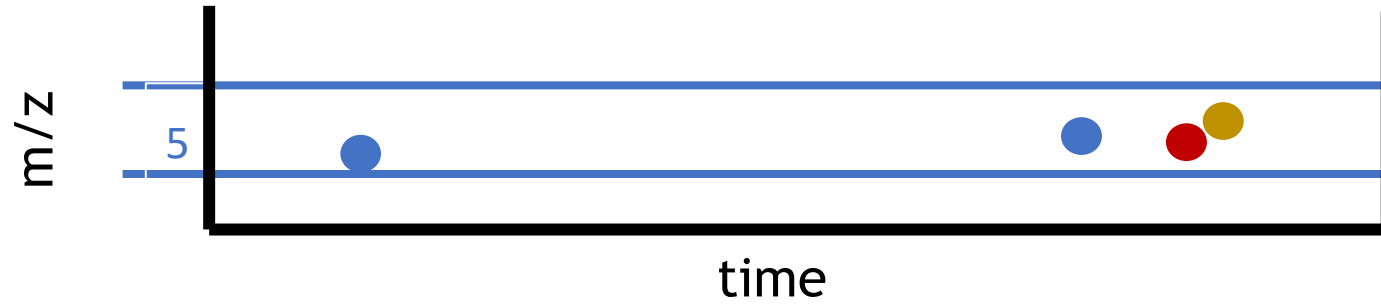
# Smoothing affects number of peak groups found



Peak density /  
Relative intensity

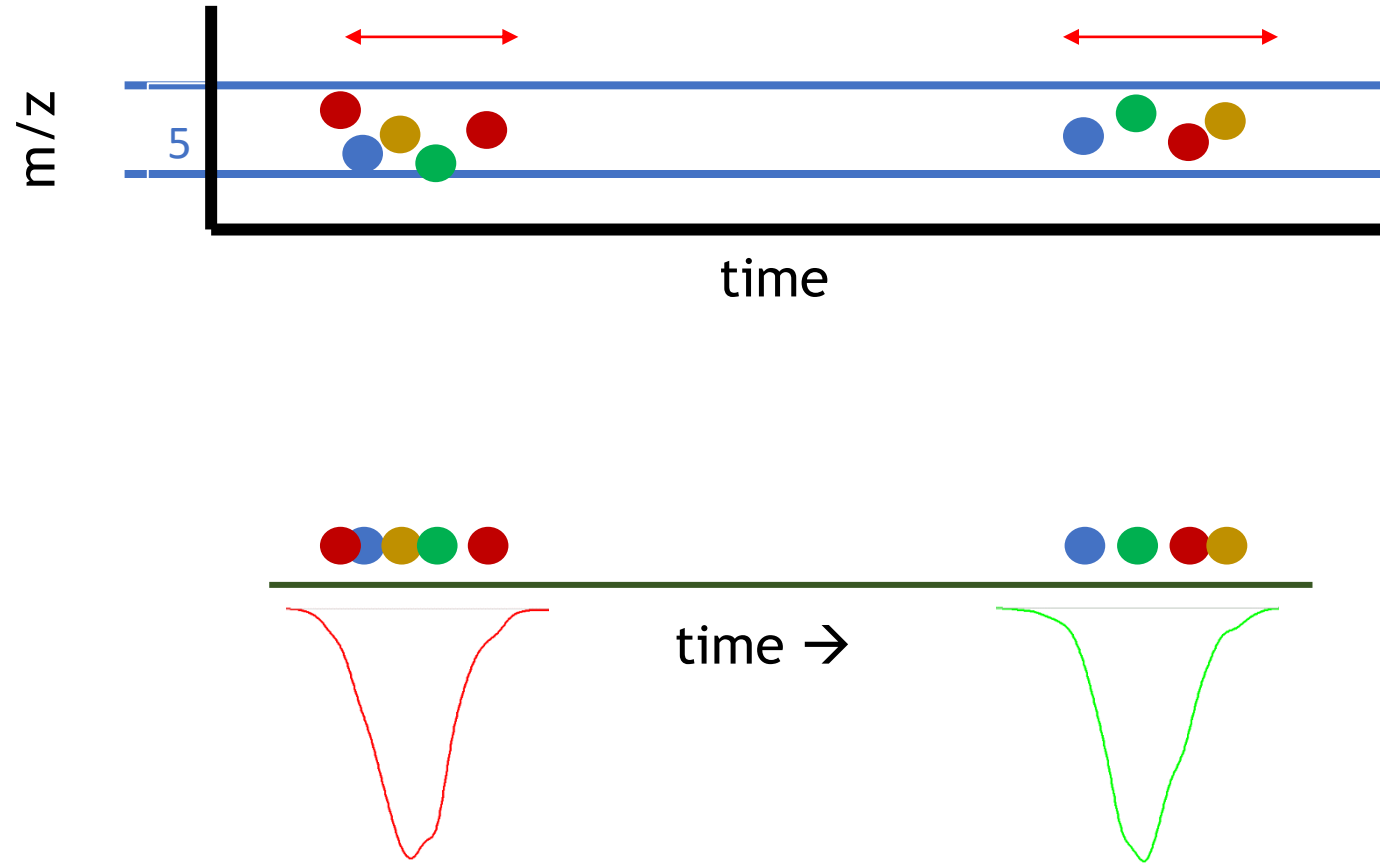


# When is peak an outlier?





# Multiple peaks in peak group for specific sample

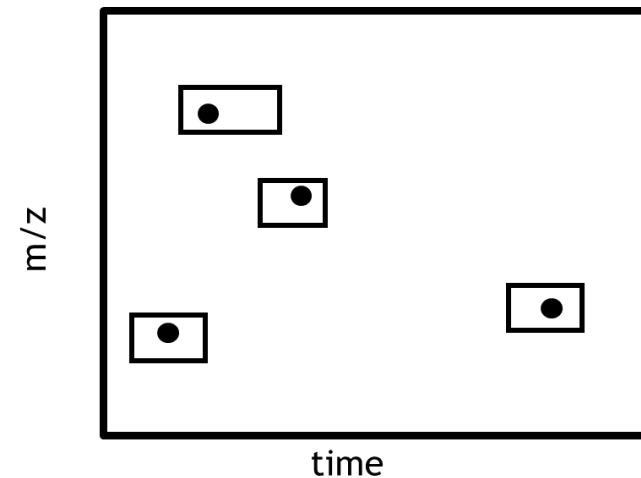


One sample with 2 peaks in one peak group

# Result of peak matching



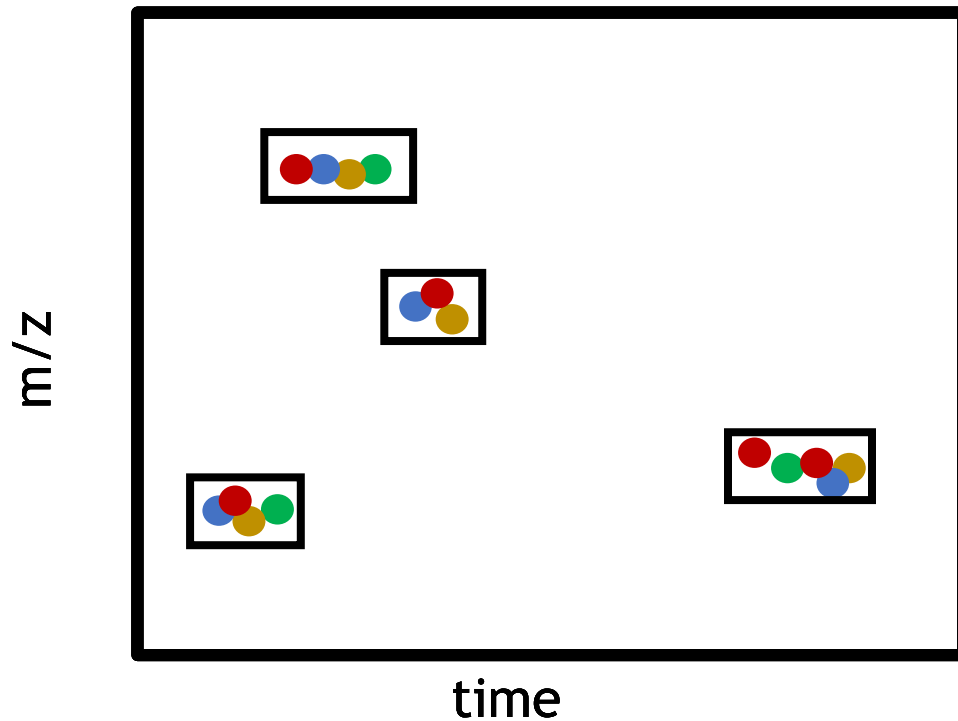
	mzmed	mzmin	mzmax	rtmed	rtmin	rtmax	npeaks	KO	WT
[1,]	200.1000	200.1000	200.1000	2925.480	2876.967	2931.740	9	4	5
[2,]	205.0000	205.0000	205.0000	2790.894	2784.635	2795.591	12	6	6
[3,]	205.9927	205.9786	206.0023	2790.112	2784.635	2795.591	12	6	6
[4,]	207.0850	207.0440	207.1000	2718.906	2712.647	2726.731	12	6	6
[5,]	219.0848	219.0488	219.1000	2524.852	2518.592	2529.547	9	4	5
[6,]	231.0236	231.0000	231.0812	2517.029	2509.202	2535.807	6	3	3
[7,]	233.0371	233.0128	233.0652	3022.507	3016.247	3077.281	13	6	6
[8,]	234.0277	234.0133	234.1000	3020.943	3019.378	3028.769	5	2	3
[9,]	235.0550	235.0000	235.0938	2692.302	2667.264	2720.471	6	3	2
[10,]	236.1018	236.0678	236.1188	2523.287	2518.592	2529.547	10	5	5



# Peak groups



- Result of 'peak matching' are 'peak groups'
- A peak group accounts for the variability in  $R_t$  and  $m/z$  for corresponding compounds across samples.



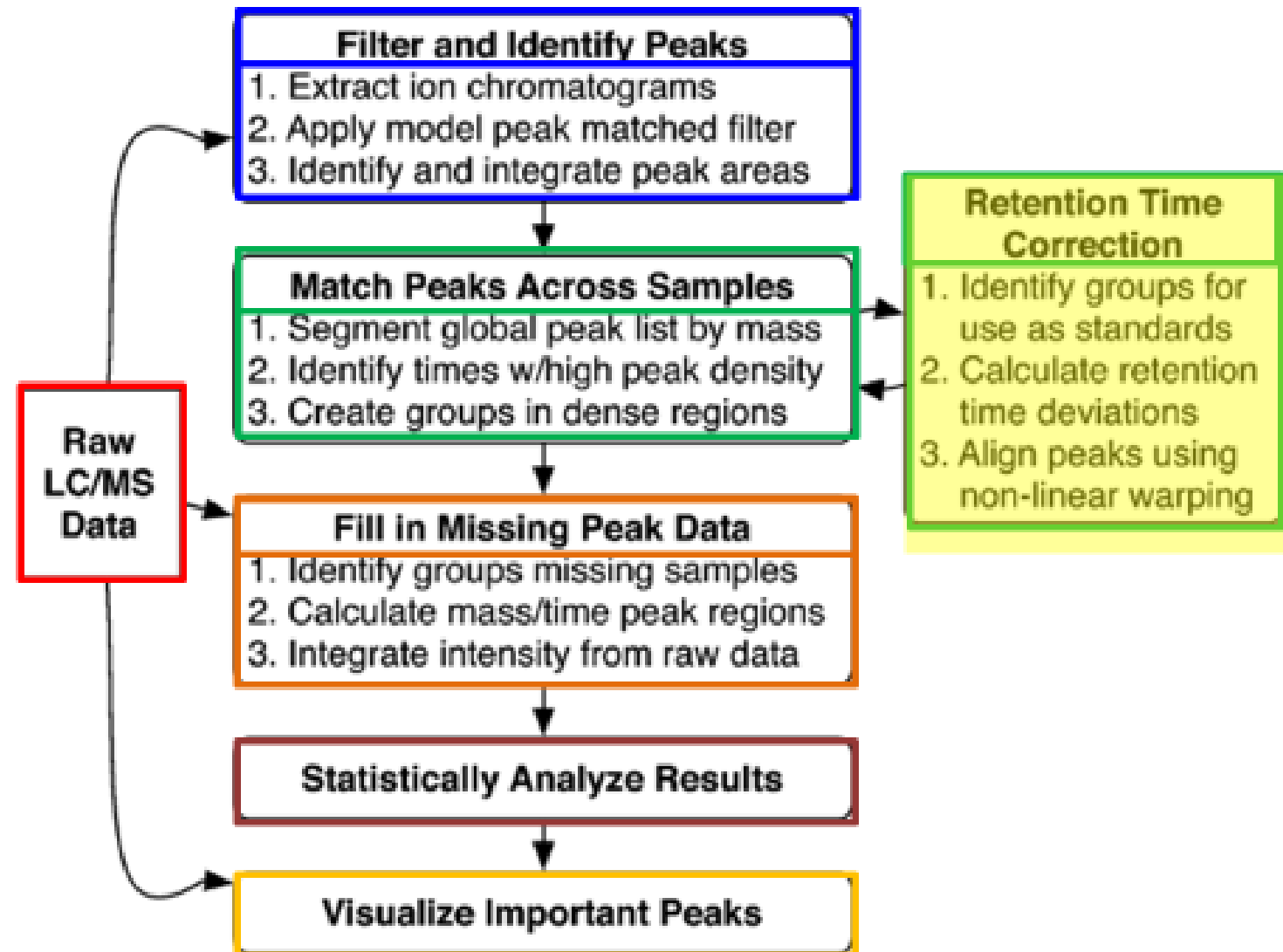
● peak in sample

Suppose we have 4 samples

Overall each peak group shows 4 peaks  
in 4 samples

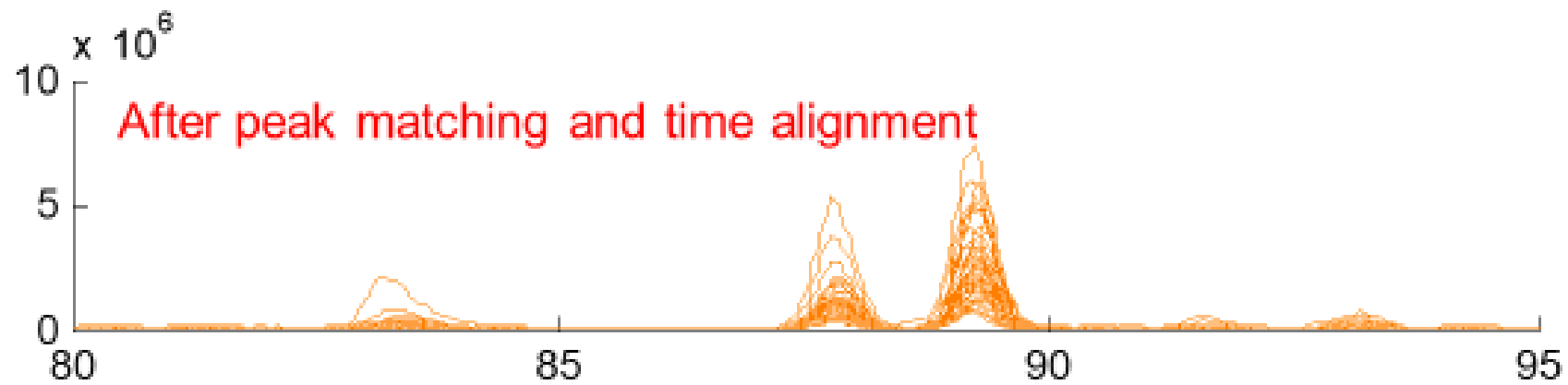
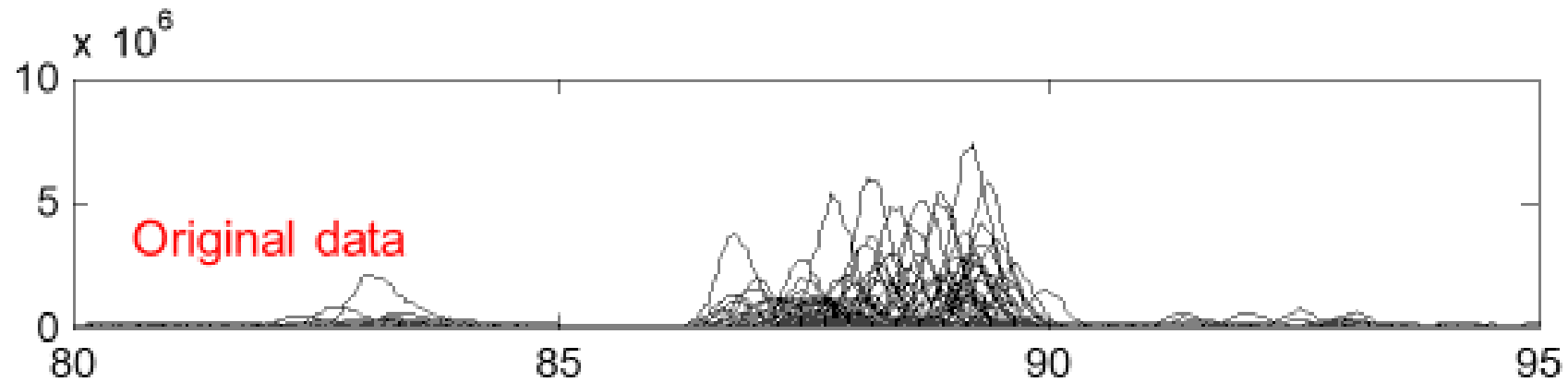
We see peak groups with too many  
and two few peaks

## Retention time correction



## Why is time alignment necessary?

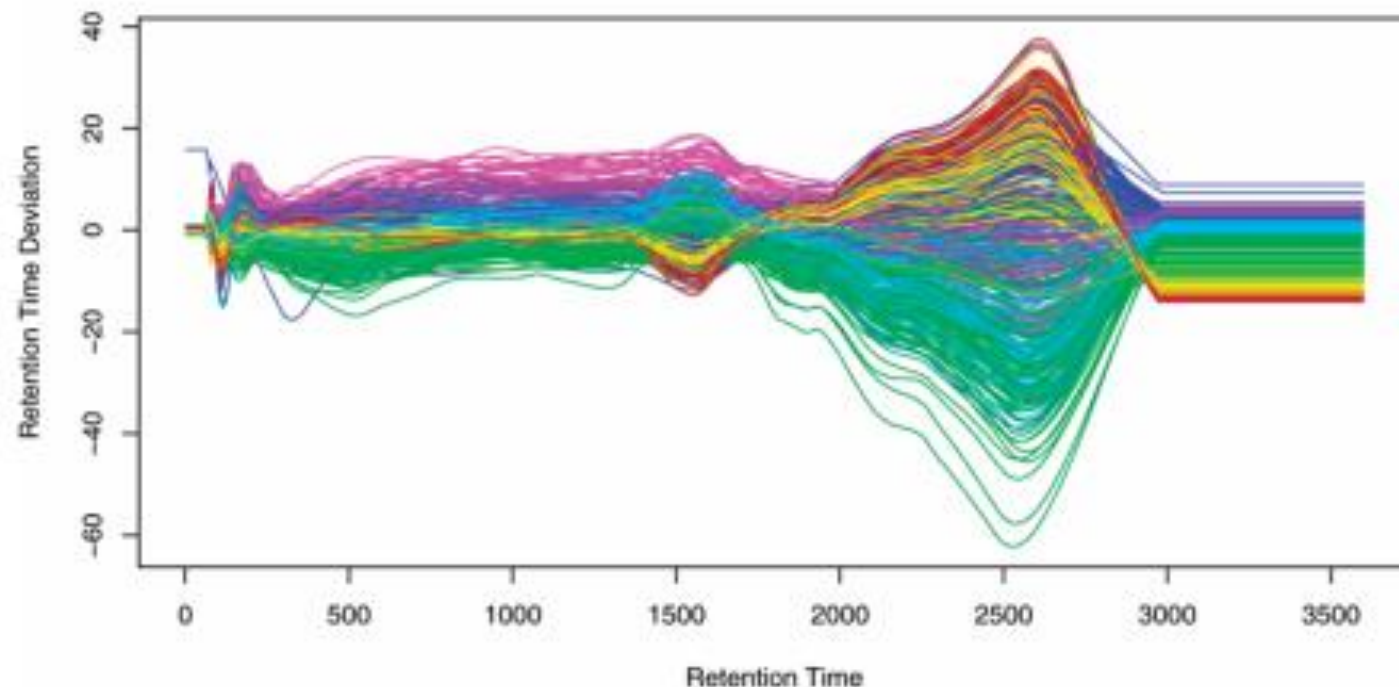
Extracted Ion Chromatograms (EICs) from 19 LC-MS runs



## Retention time deviation profiles

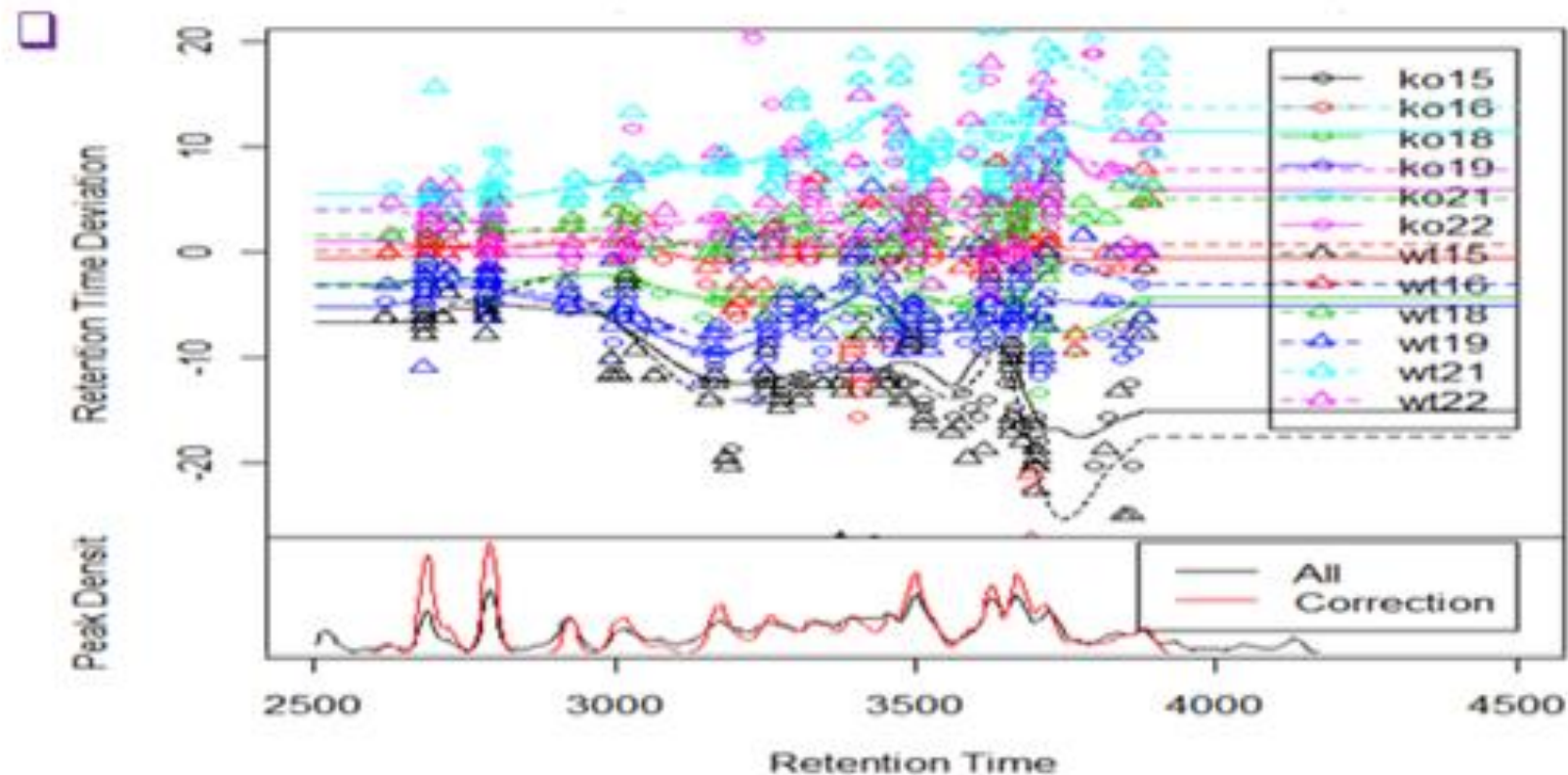
- ❑ 476 LC/MS analysis from serum samples
- ❑ **Positive deviation:** sample elutes after median retention time
- ❑ **Negative deviation:** sample elutes before median retention time
- ❑ Sample profiles are coloured in a rainbow by the order in which they were run, with red being the first samples and violet being the last samples run

Retention Time Deviation vs. Retention Time



## Retention time vs deviation

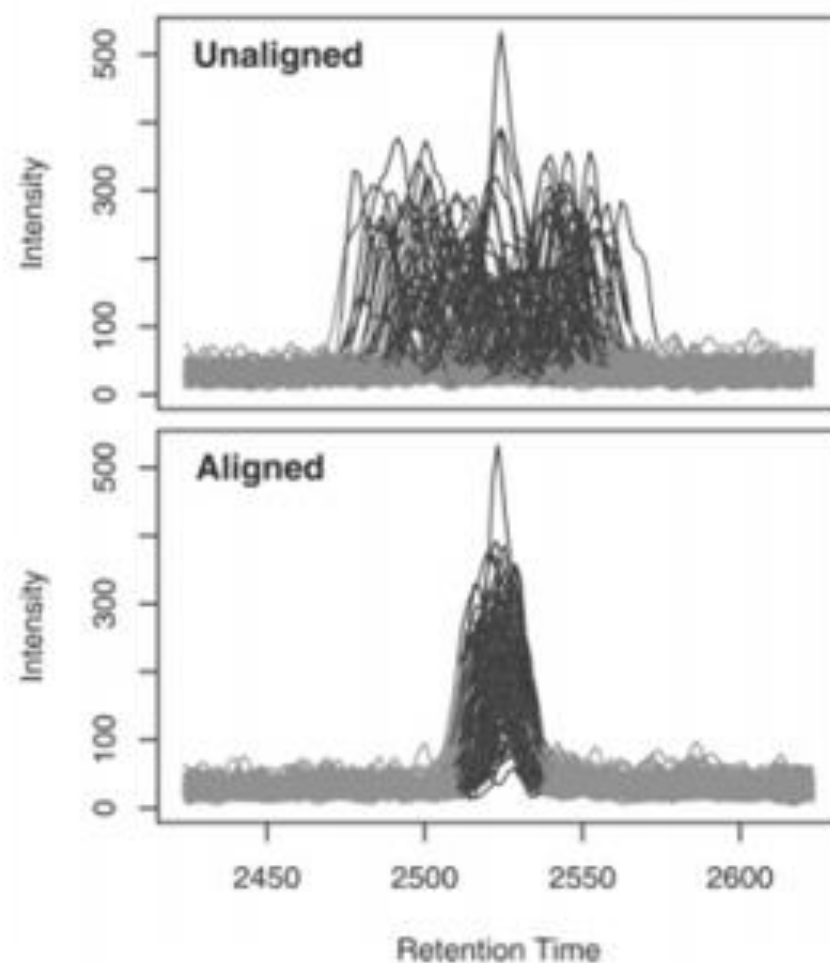
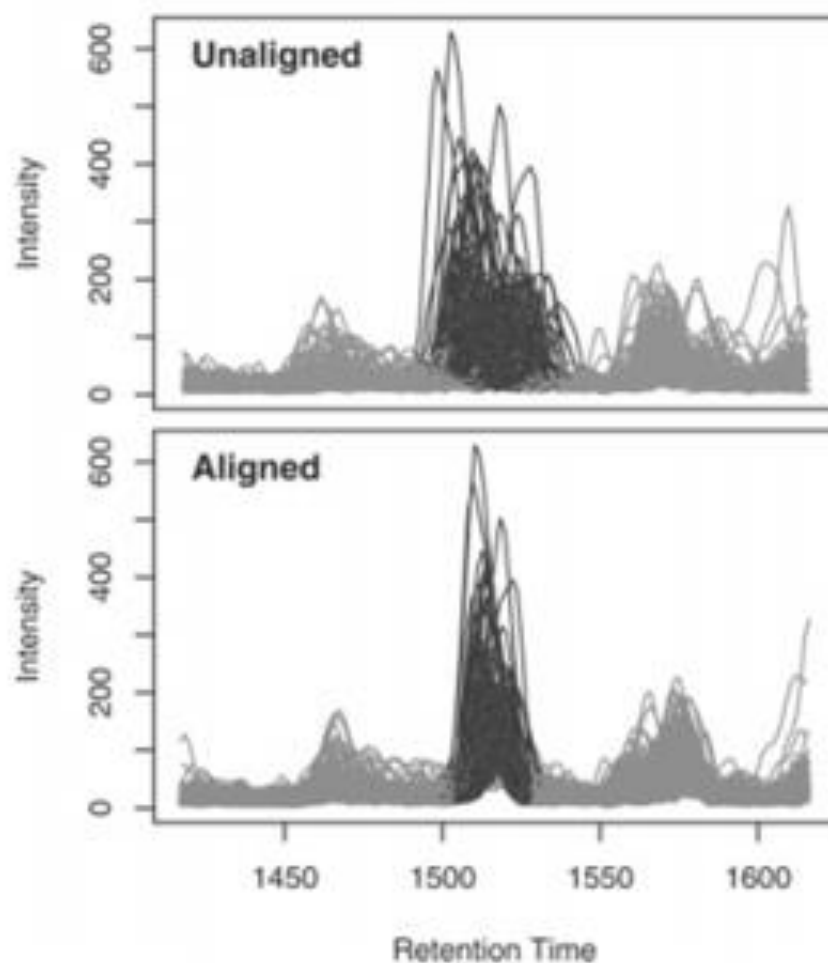
- Example: 6 WT and 6 KO samples
- Well-behaved peak groups are shown
  - Density plot shows that the distribution of all peaks and the well-behaved peaks (peaks used for correction) is similar



## Result: before and after retention time correction

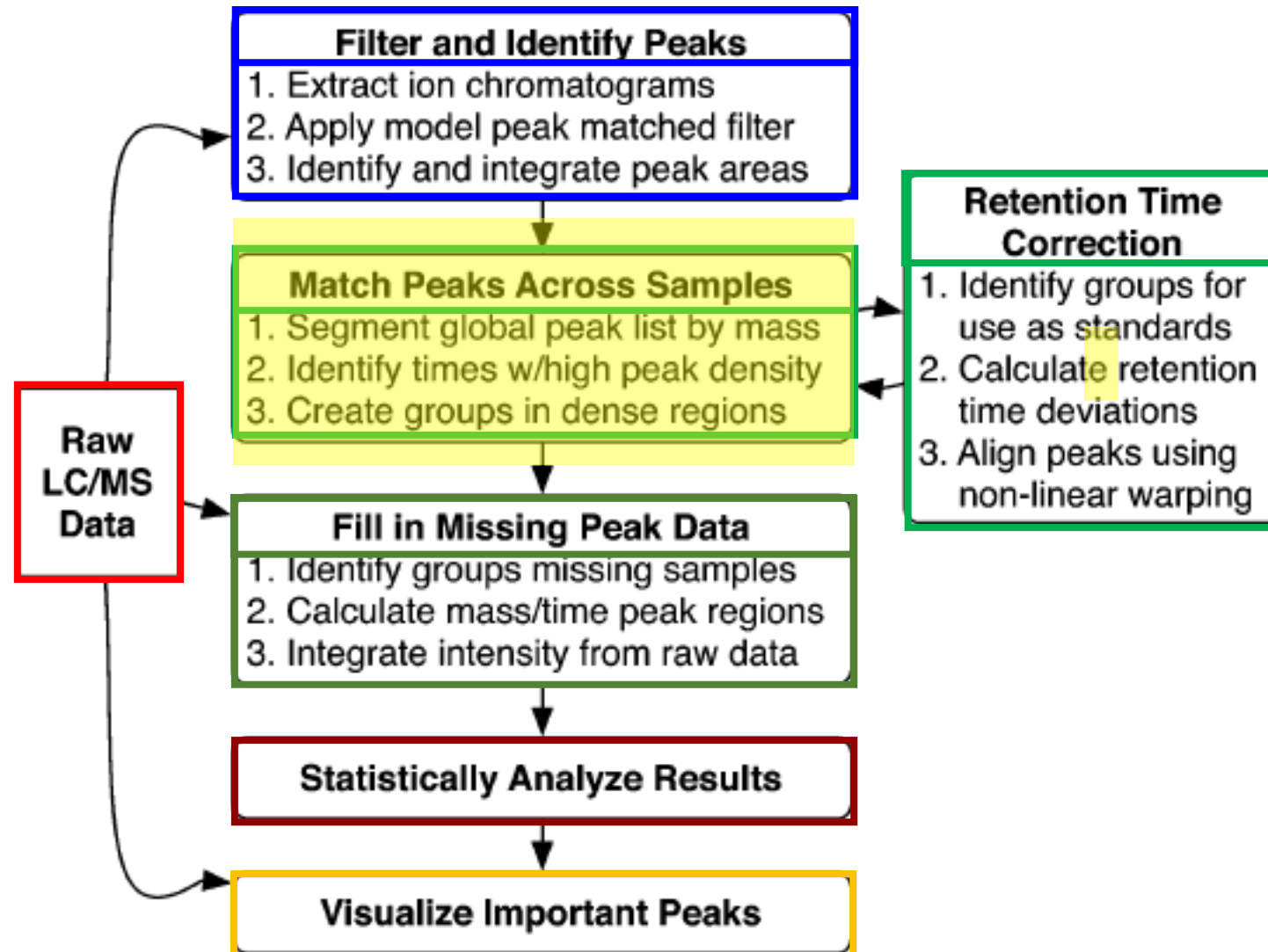
Extracted Ion Chromatogram: 550.2 - 550.5 m/z

Extracted Ion Chromatogram: 320.1 - 320.4 m/z

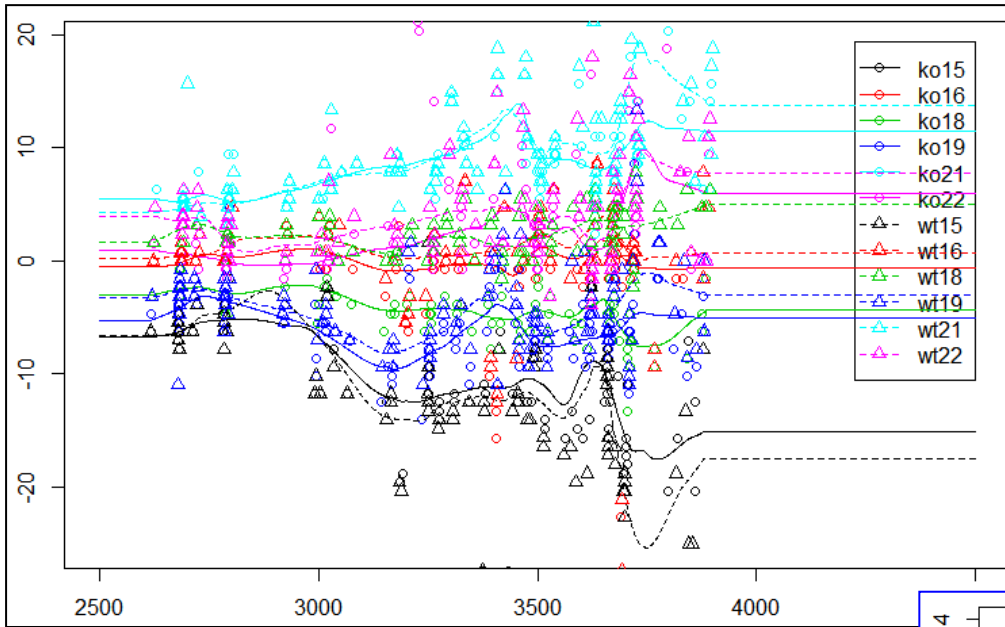




# Peak matching and Retention time correction can be iterated to get the desired result

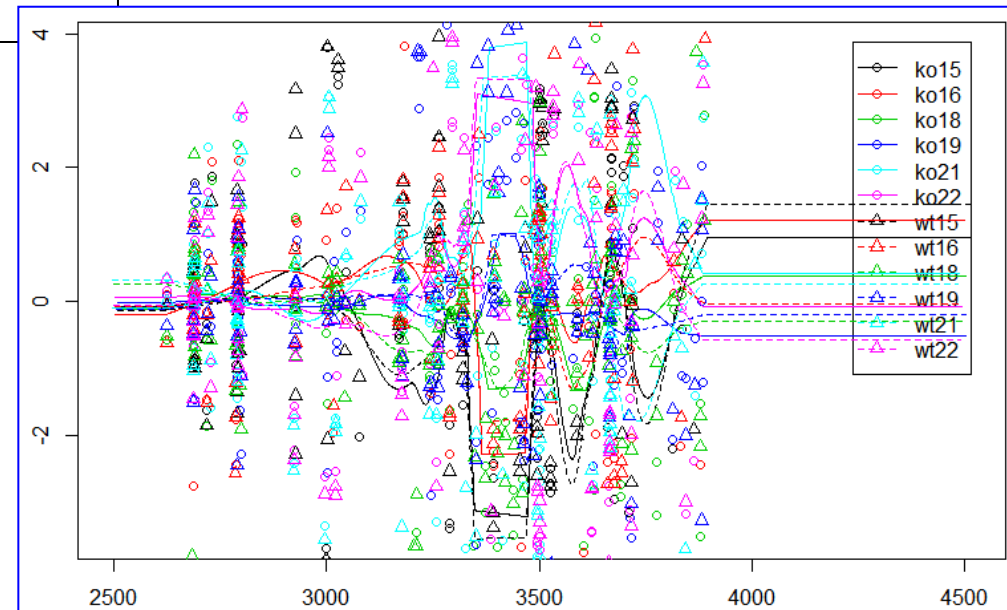


# Two iterations of time alignment

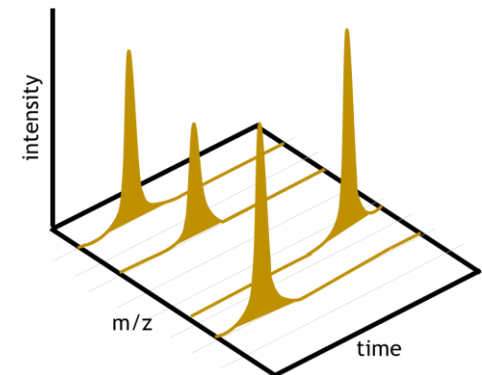
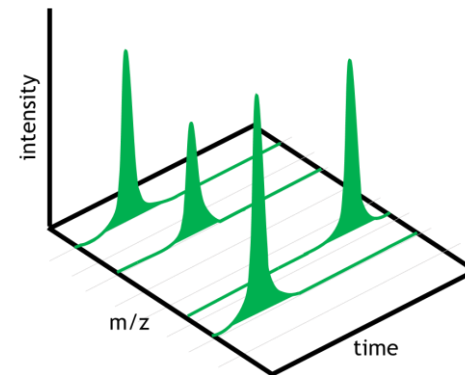
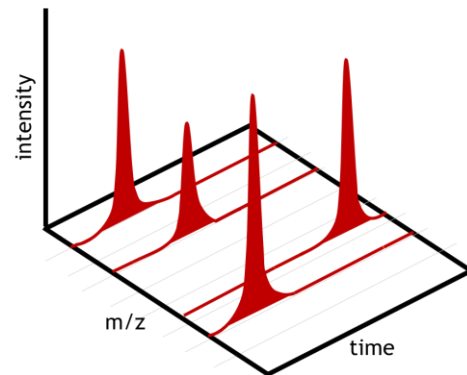
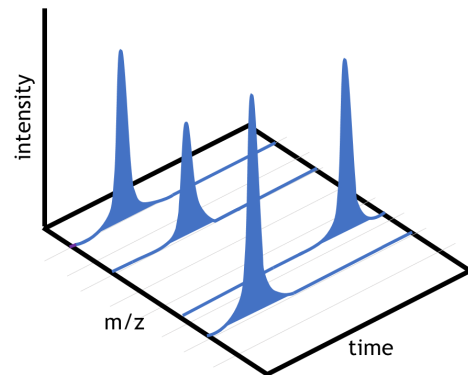
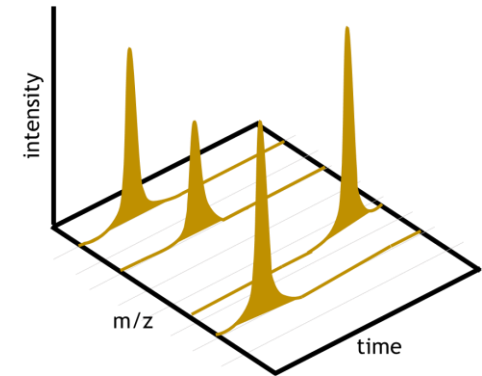
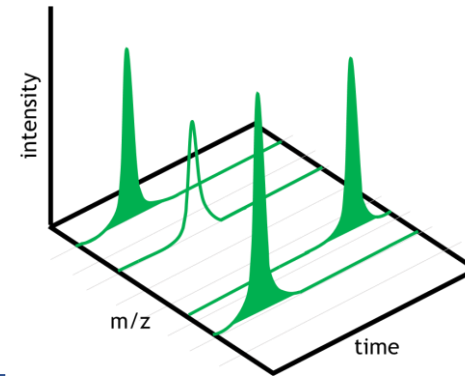
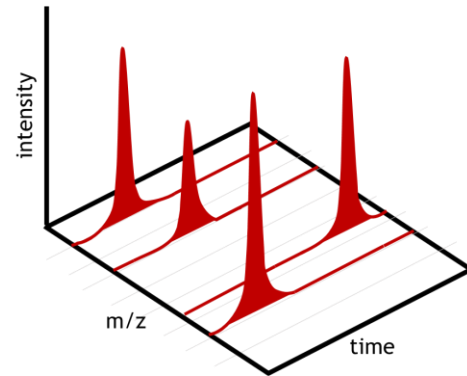
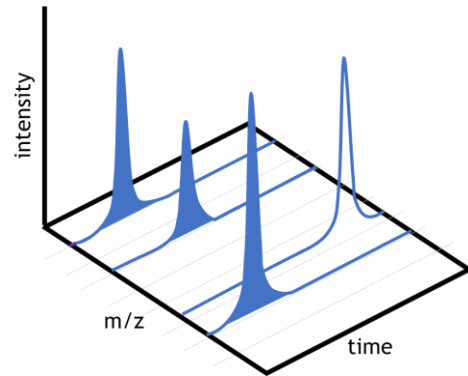


First iteration

Second iteration  
(not the y-scale)



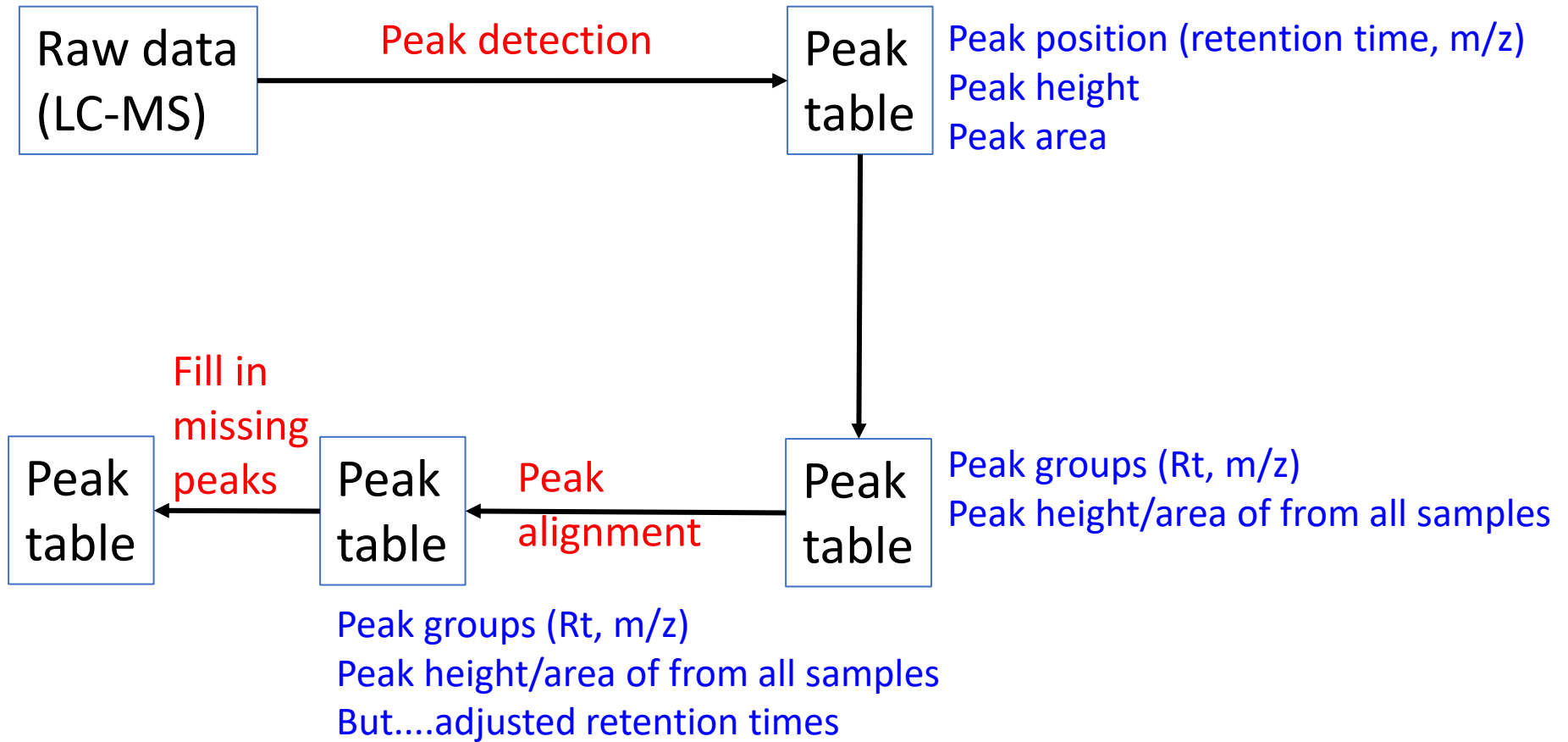
# Preprocessing 4: Fill missing peak data



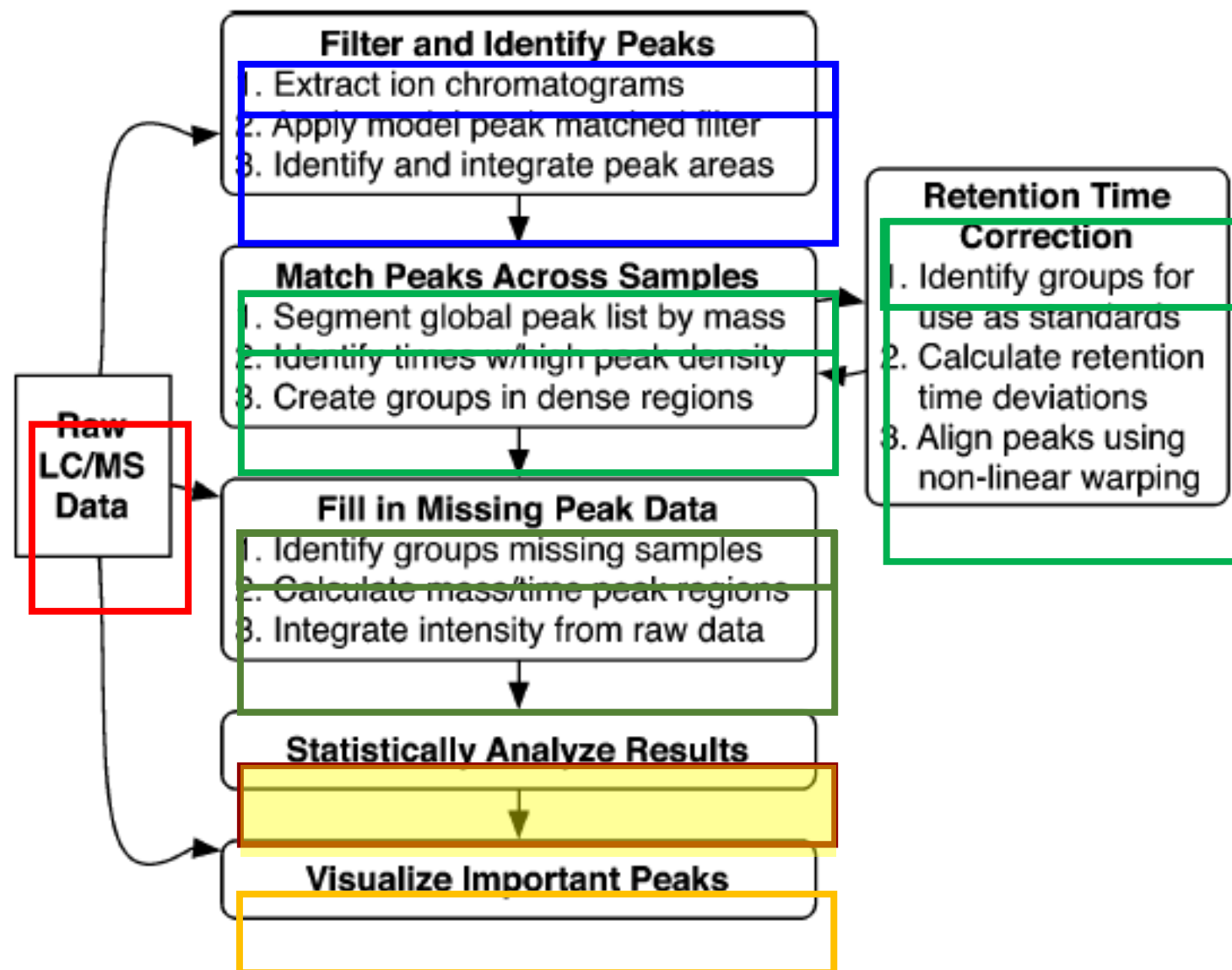
# Fill in missing peak data



- Determines which samples are missing from each peak group.
  - Thus, no peak detected for those samples.
- Using information from peak detection about where peaks begin and end, and aligned retention times for each sample
  - Then integrate the raw LC/MS data to fill in intensity values for each of the missing data points.
- A significant number of potential peaks can be missed during peak detection.
- The step of filling in missing peak data is necessary for robust statistical analysis.



Peak groups (Rt, m/z)  
Peak height/area of from all samples  
But....adjusted retention times and without missing peaks



# Statistical analysis of peaks

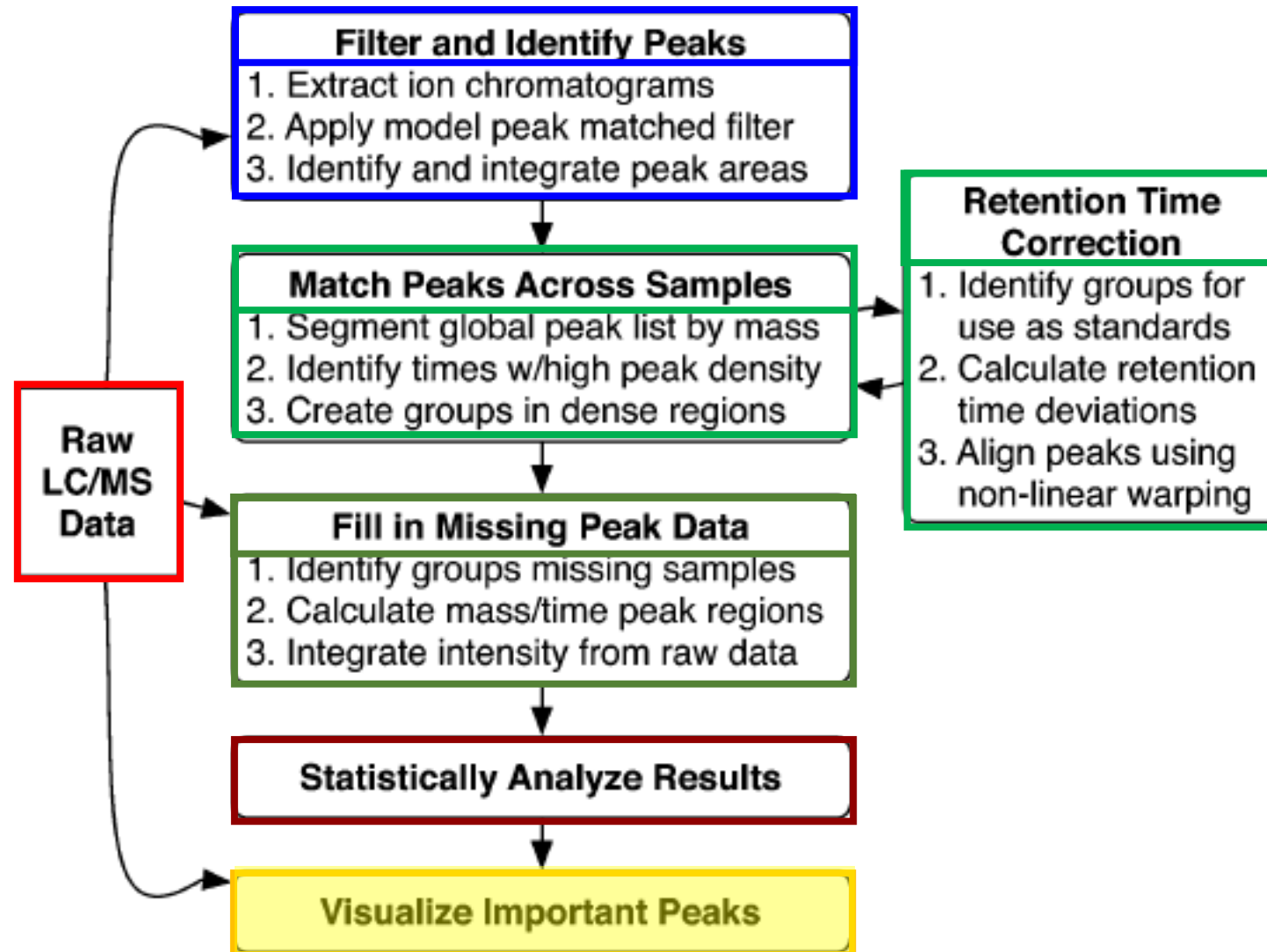


- Comparison of two conditions

name	fold	tstat	pvalue	mzmed	mzmin	mzmax	rtmed	rtmin	rtmax	npeaks	KO	WT
M300T3390	5.69359	14.4437	5.03E-08	300.19	300.171	300.2	3390.32	3386.76	3396.33	12	6	6
M301T3390	5.87659	15.5757	6.71E-08	301.188	301.166	301.195	3389.63	3386.76	3392.1	7	6	1
M298T3187	3.87092	11.9389	3.31E-07	298.151	298.105	298.159	3186.8	3184.12	3191.31	4	4	0
M491T3397	24.9757	16.8399	4.46E-06	491.2	491.188	491.206	3397.16	3367.12	3424.68	6	6	0
M348T3288	9.00507	17.2627	5.03E-06	348.162	348.123	348.174	3288.25	3284.66	3294.13	6	6	0
M423T3257	6.24617	10.8189	4.71E-05	423.15	423.106	423.157	3256.61	3254.85	3261.53	6	6	0
M327T3419	26.7484	11.0506	9.44E-05	327.199	327.168	327.2	3419.47	3412.92	3427.23	6	6	0
M326T3417	15.5444	10.4239	0.00013	326.2	326.17	326.2	3417.01	3411.31	3425.62	12	6	6
M410T3938	6.78734	9.11467	0.00017	410.266	410.212	410.299	3937.7	3932.87	3946.14	9	6	3

ko15	ko16	ko18	ko19	ko21	ko22	wt15	wt16	wt18
4534354	4980914	5290739	4564263	4733236	3931593	349661	491793	64552
962353	1047934	1109303	946943	984787	806171	86450.4	120097	14300
180781	203927	191016	190627	156869	220289	16269.1	43677.8	54739
432037	332159	386967	334951	294816	373578	7643.14	10519.9	26472
165831	183665	150845	134637	136452	167008	24302.8	16631.4	19213
236250	255169	212711	180691	191747	152861	29530.2	17037.2	35133
1108851	950127	674223	677091	772290	1013978	58898.7	21991	2764
4809521	3931305	2913712	2819101	3284987	4346410	259229	314154	16190
080012	1310876	1375015	040310	800086	700403	137451	202674	22254

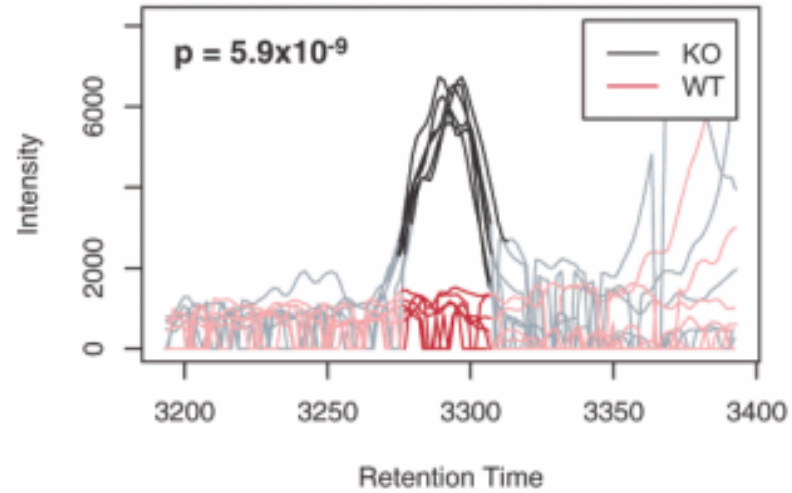
# Peak matching and Retention time correction can be iterated to get the desired



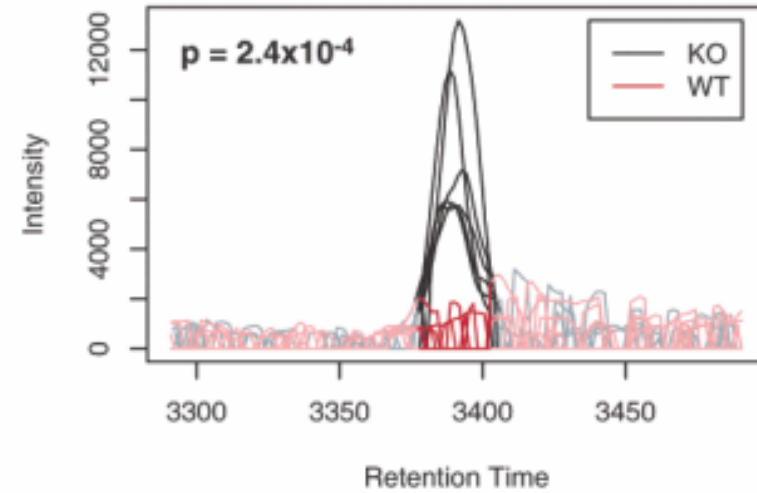


# Data visualization

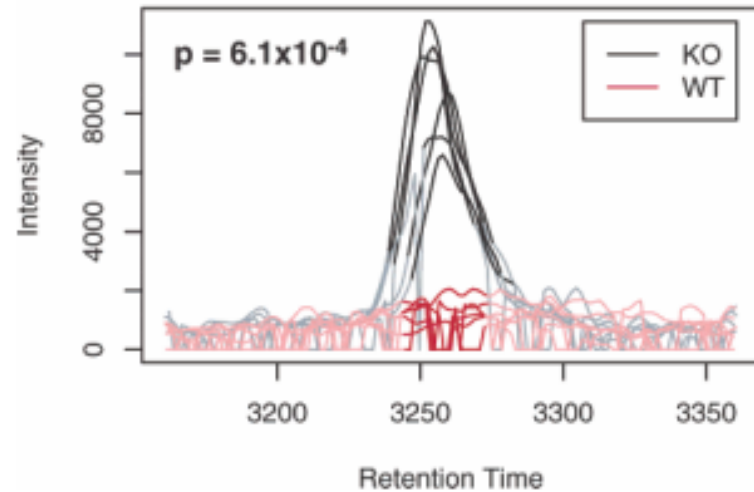
Extracted Ion Chromatogram: 449.1 - 449.2 m/z



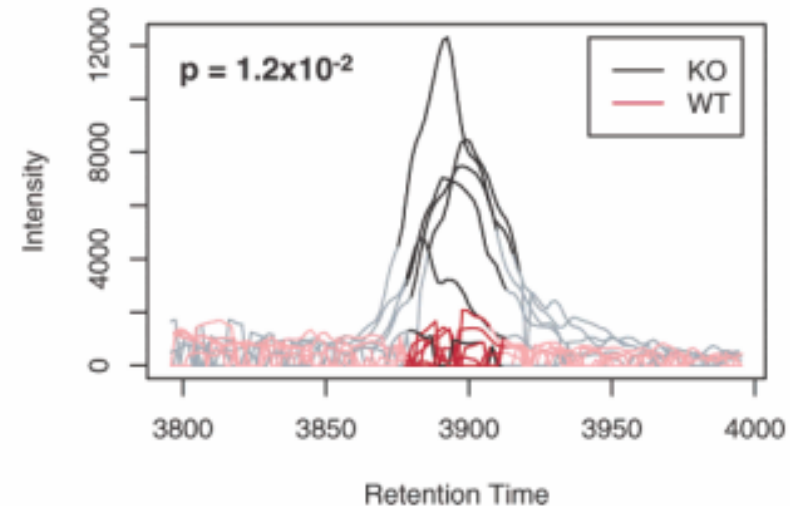
Extracted Ion Chromatogram: 621.3 - 621.4 m/z



Extracted Ion Chromatogram: 423.1 - 423.2 m/z



Extracted Ion Chromatogram: 533.2 - 533.4 m/z

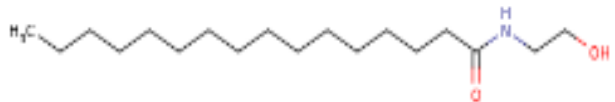




Alternatively, try identifying the mass using the Human Metabolome Database at <http://www.hmdb.ca/>. Use menu option Search >> MS search. Input mass 299.19, Ionization Neutral and Molecular Weight Tolerance 0.1 Da or use the mass for the positive ion given in the mzmed column: 300.19, Ionization Positive. Towards the end of the list of possible annotations you'll find Palmitoylethanolamine.

# Compound Identification (Metlin)

- [http://metlin.scripps.edu/metabo\\_list.php?mass\\_min=299.04&mass\\_max=299.34](http://metlin.scripps.edu/metabo_list.php?mass_min=299.04&mass_max=299.34)

43210		<b>Palmitoyl Ethanolamide</b> <i>Formula:</i> $C_{18}H_{37}NO_2$ <i>CAS</i> : 544-31-0	YES	
-------	--	--	-----	---

This information is added to the peak table

# Now you are ready for down-stream analysis!

