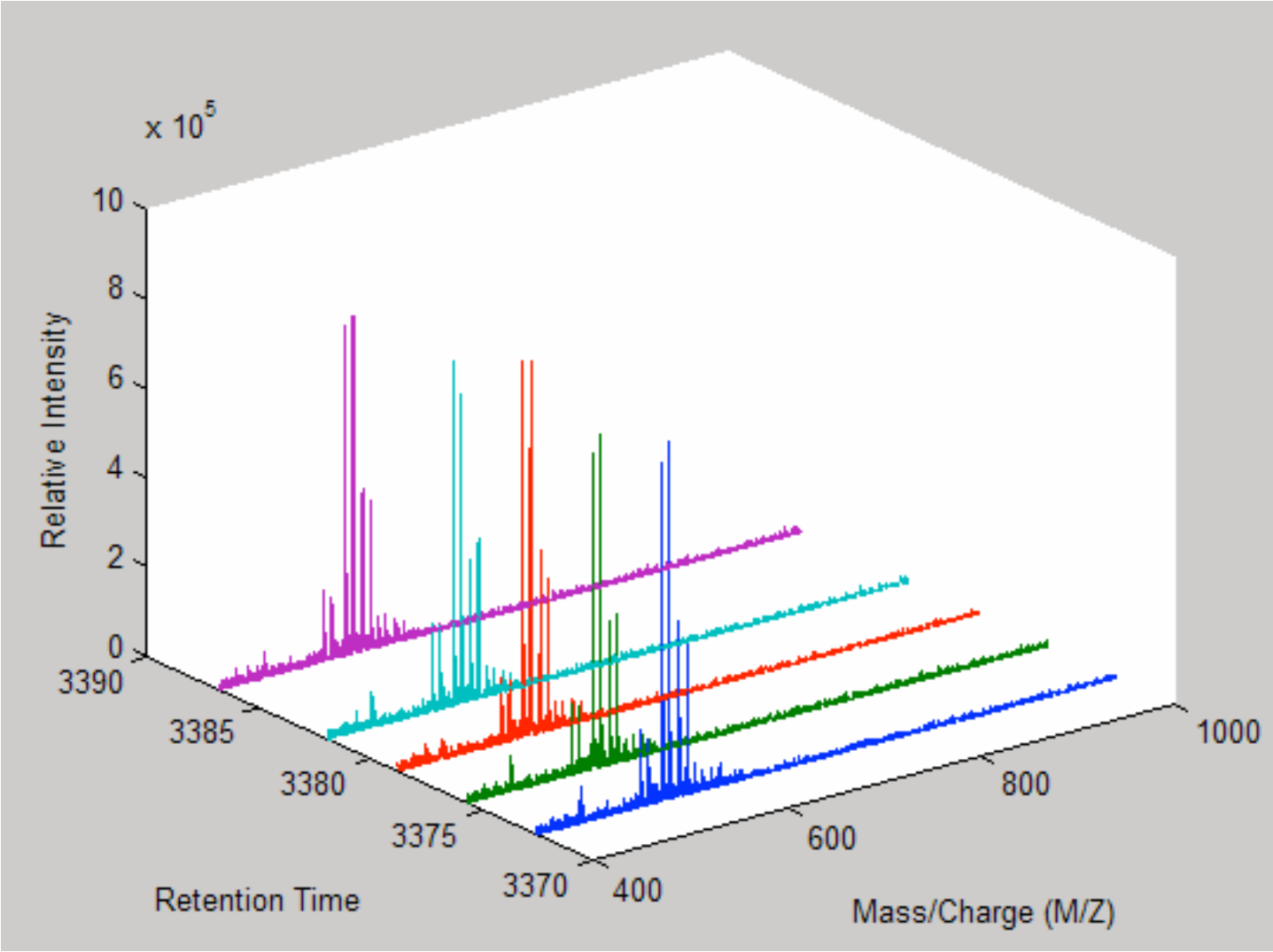
Principles of Pre-processing LC-MS-Based Untargeted Metabolomics Data

Xiuxia Du
Department of Bioinformatics & Genomics
University of North Carolina at Charlotte

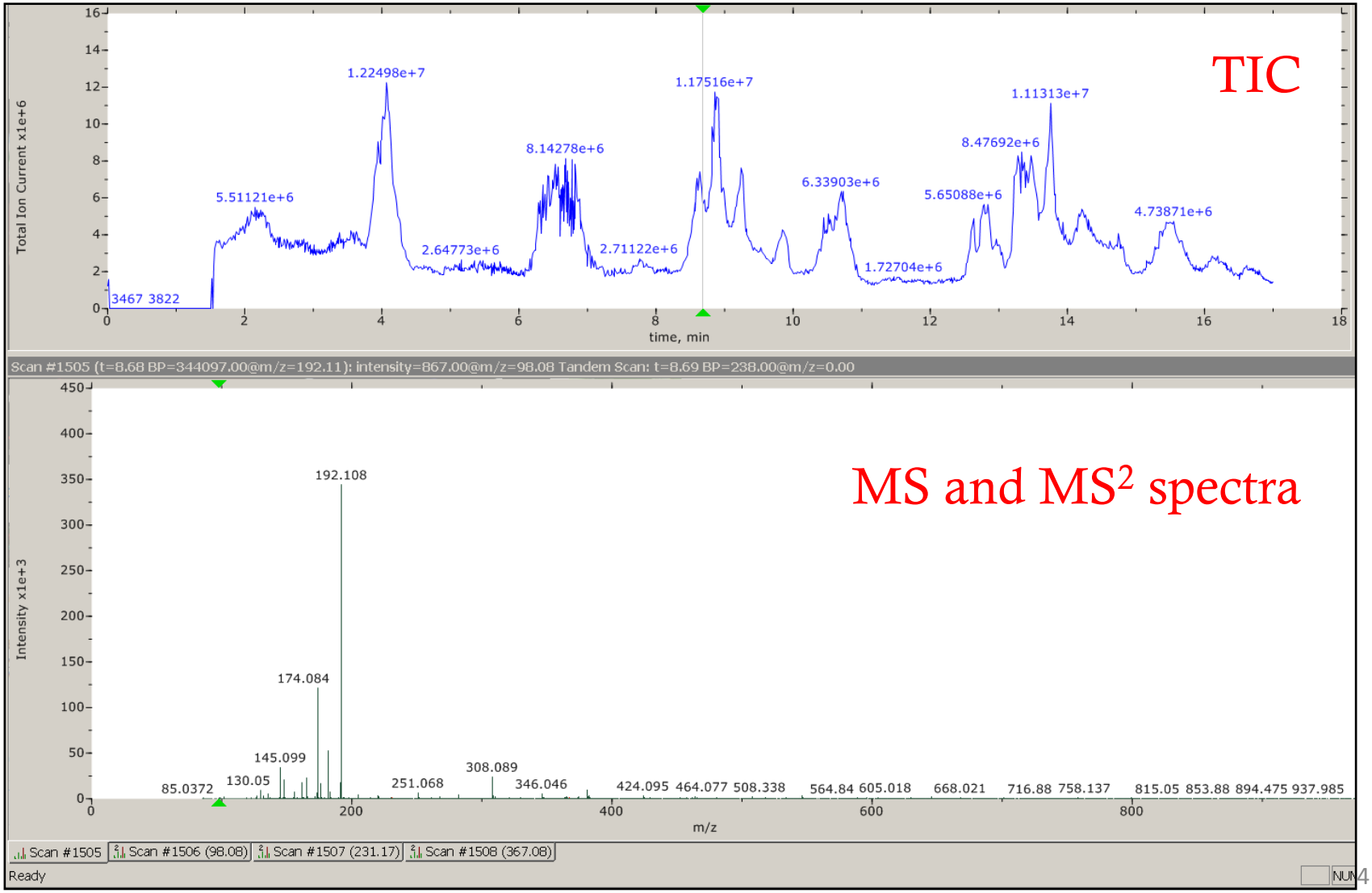
Goals of pre-processing

- Extract qualitative and quantitative information of possible metabolites
  - Determine the identity
  - Estimate the relative abundance
- Align samples to correct retention time shifts
- Produce a table of possible metabolites with their quantitative information for subsequent statistical analysis

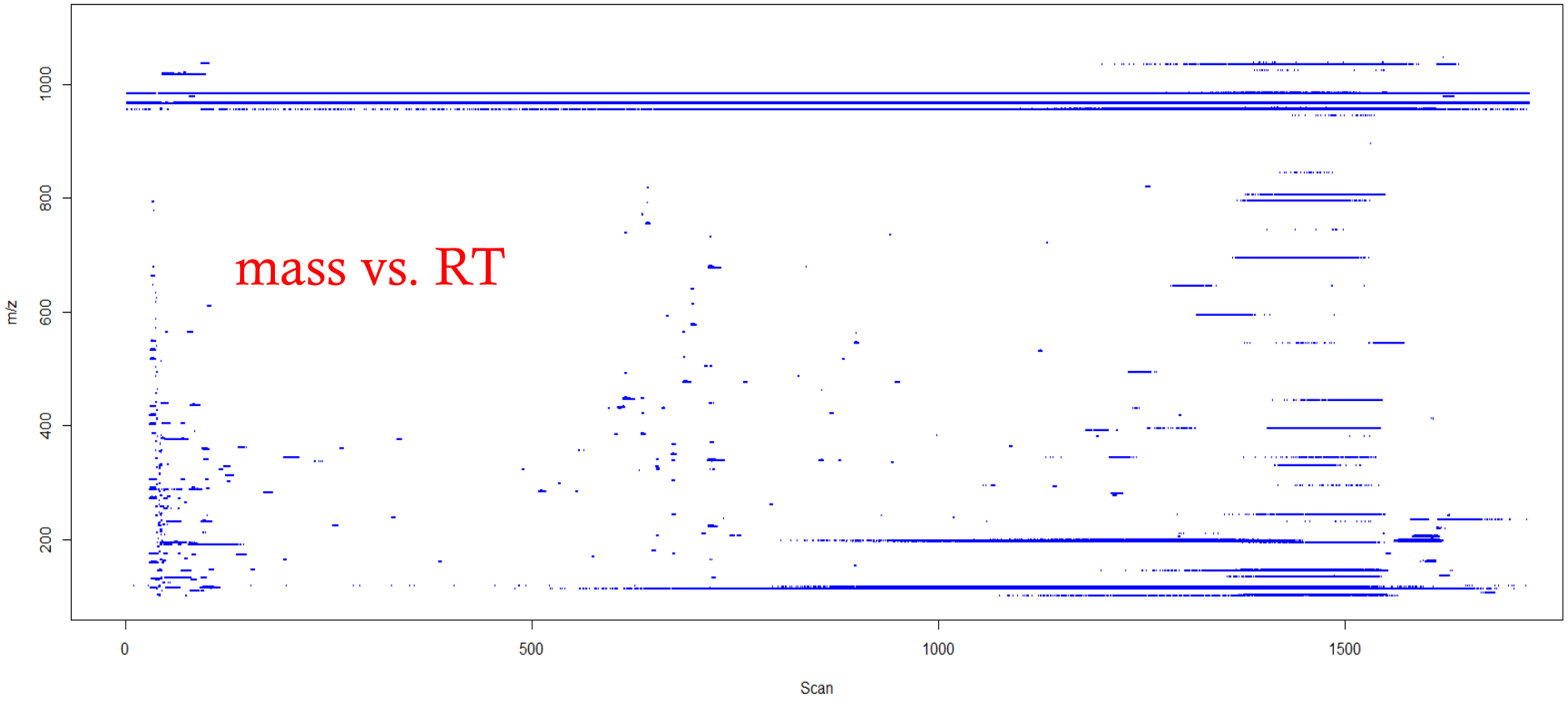
Raw LC-MS data (I)



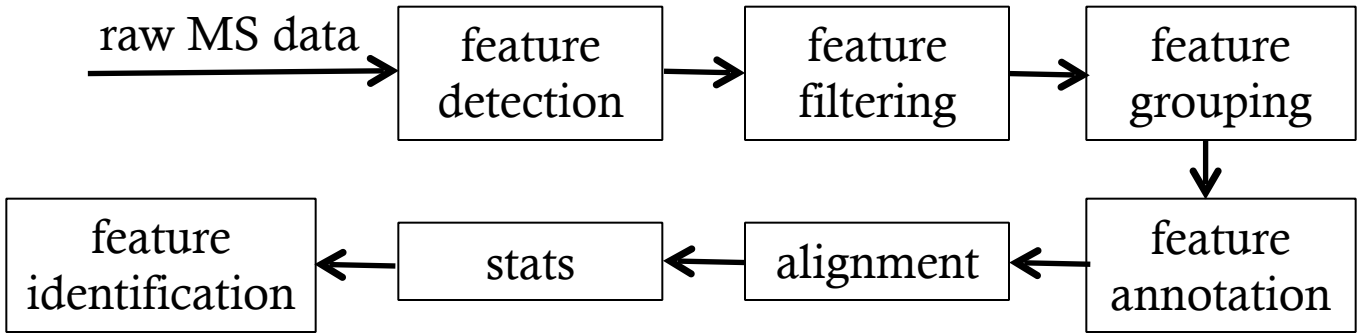
Raw LC-MS data (II)



Raw LC-MS data (III)

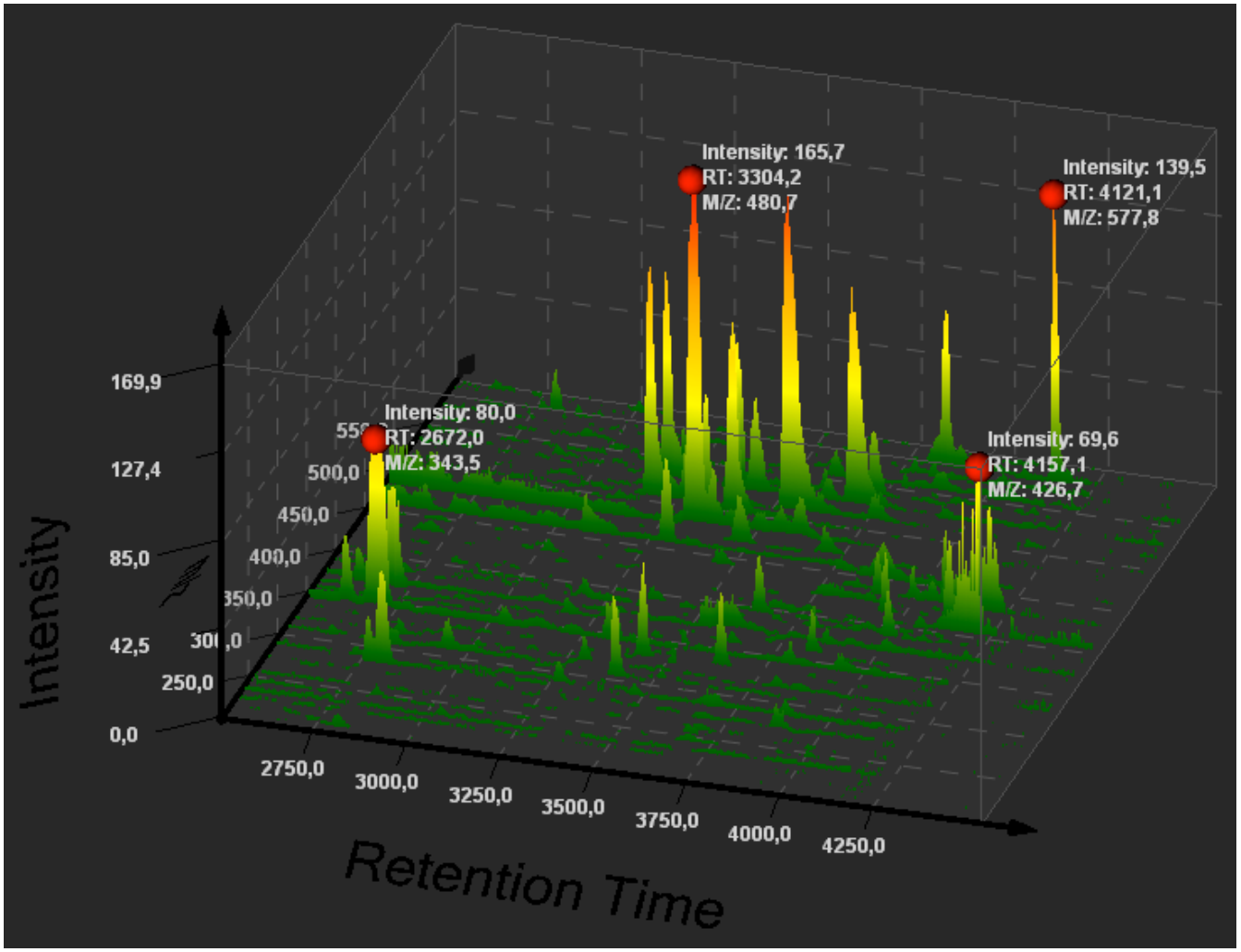


Work flow



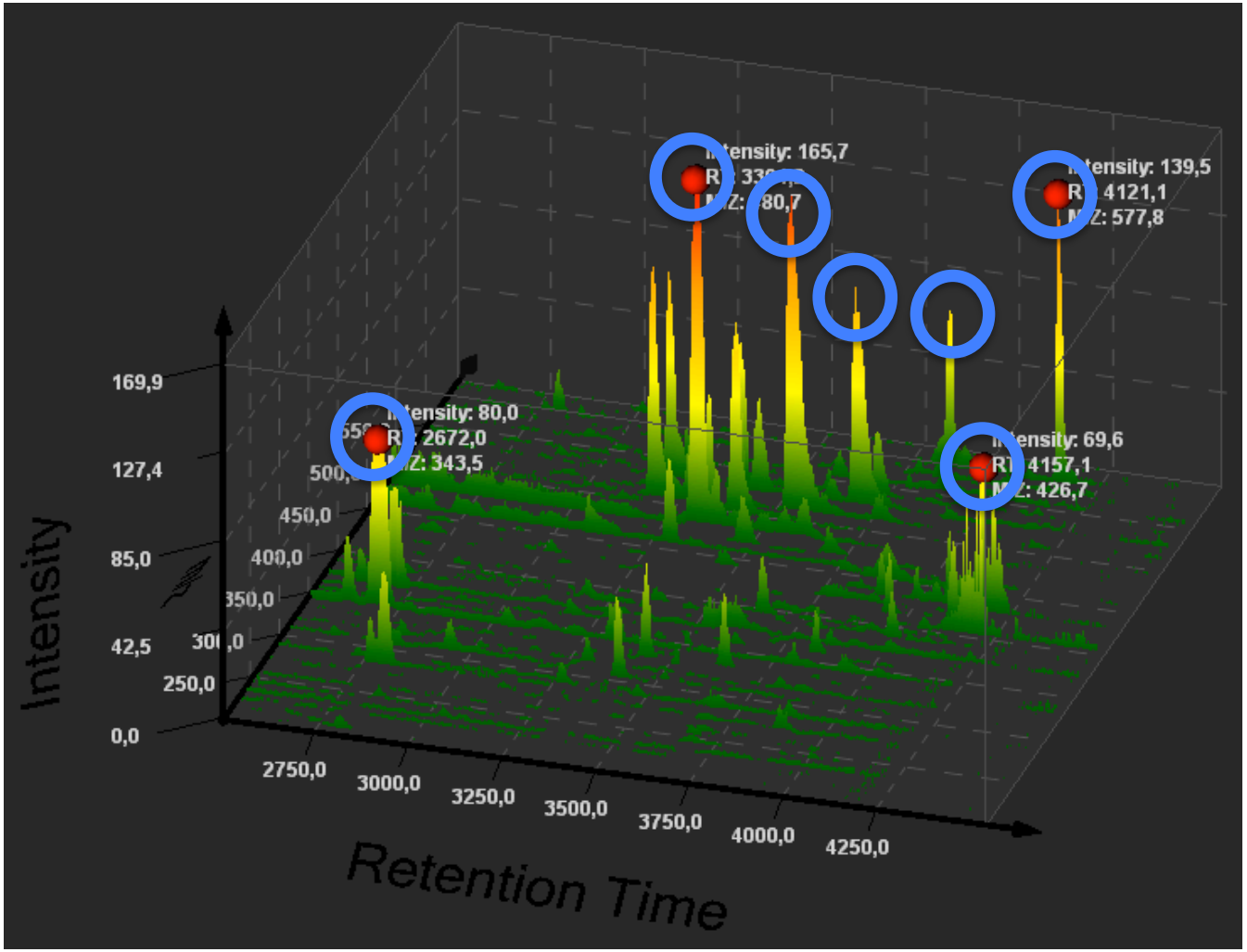
- Feature: a 3D signal induced by a single ion species (e.g. $[M+H]^+$ or $[M-H]^-$ of a compound)

Features (I)



http://www.spectrolyzer.com/fileadmin/manual/html/Manual233.png

Features (II)


Intensity: 165,7
Intensity: 139,5
RT: 4121,1
MZ: 577,8
Intensity: 80,0
RT: 2672,0
MZ: 343,5
Intensity: 69,6
RT: 4157,1
MZ: 426,7
169,9
127,4
85,0
42,5
0,0
Intensity
500,0
450,0
400,0
350,0
250,0
2750,0
3000,0
3250,0
3500,0
3750,0
4000,0
4250,0
Retention Time

Feature detection

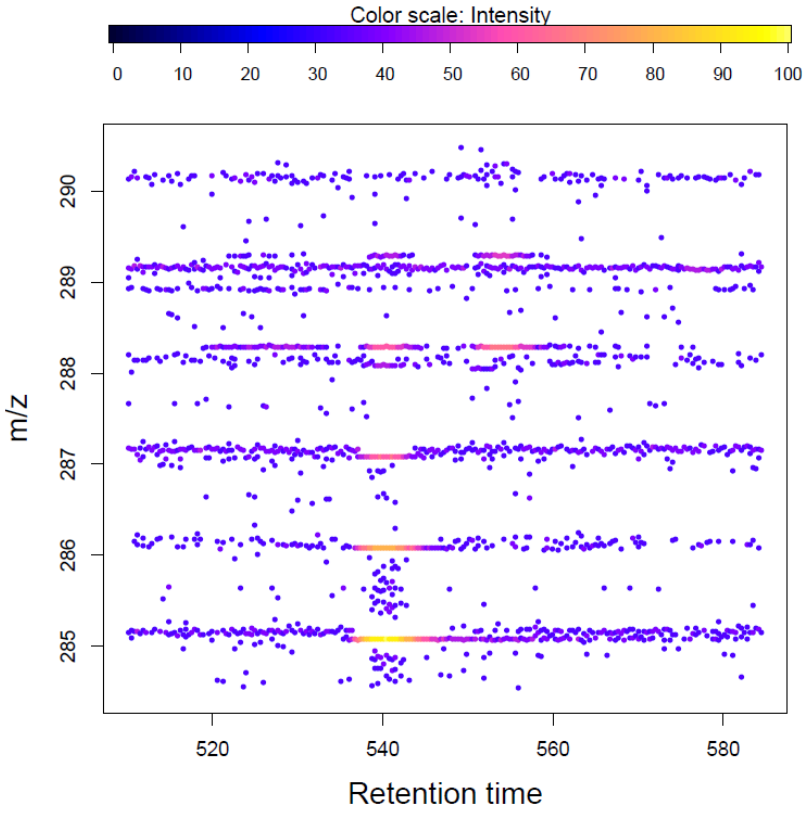
- Two steps
  - Detect mass traces
    - Binning
    - Region of interest (ROI)
  - Detect chromatographic features

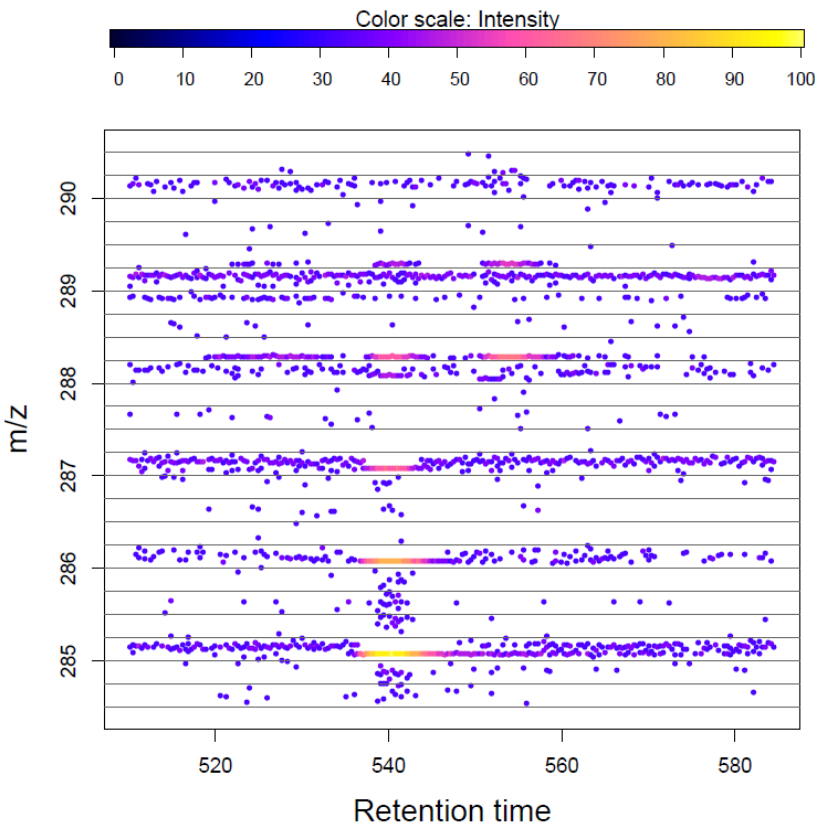
- Binning
  - Partition the mass *vs.* RT map into bins of fixed width
  - Difficult to estimate optimal bin width

    Too small → split features

    Too wide → possible feature merging

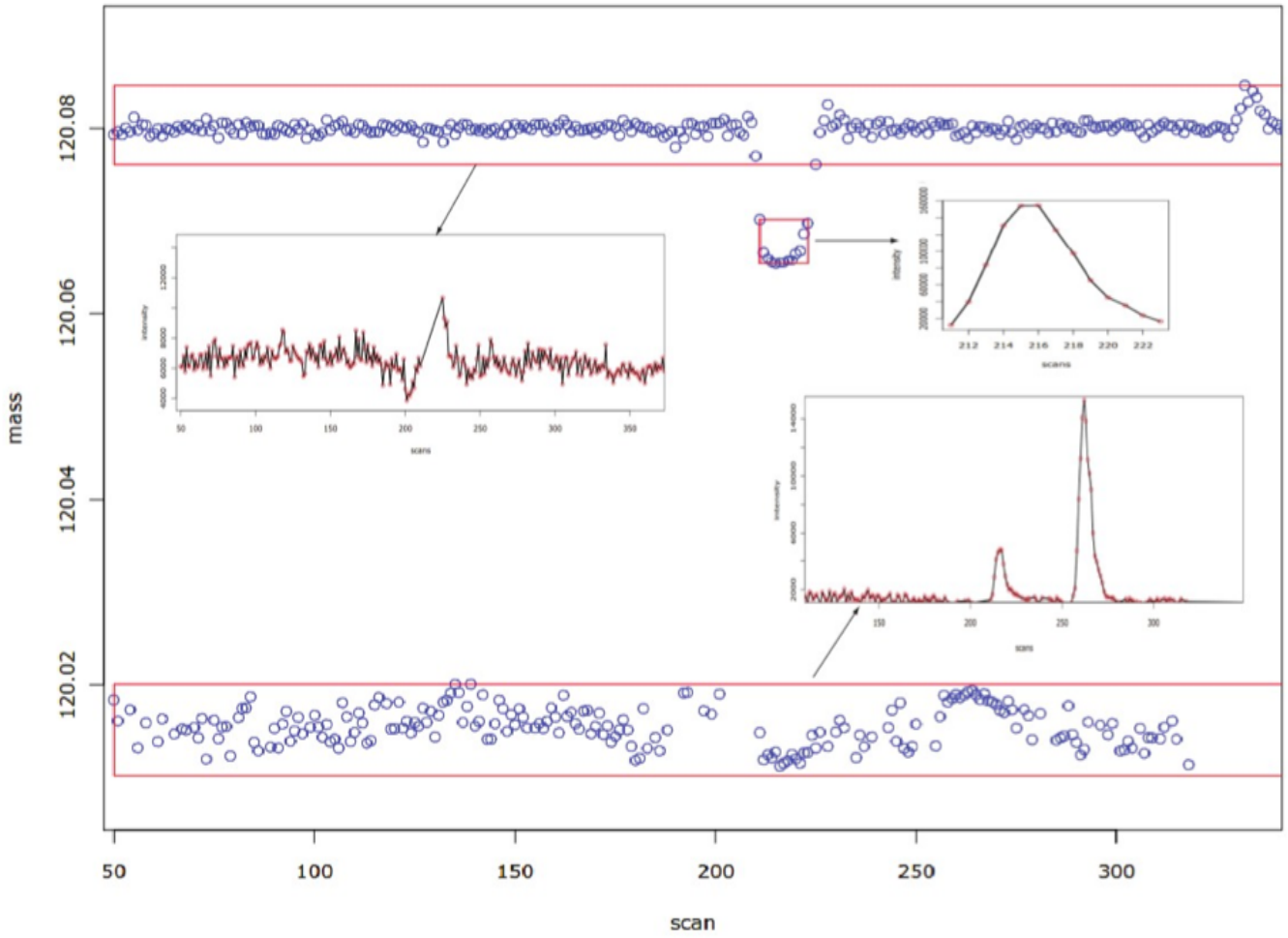
Binning





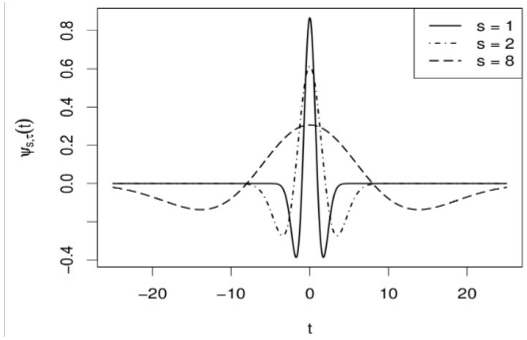
Tautenhahn, R.; Bottcher, C.; Neumann, S., Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics* **2008,** *9*, 504.

ROI

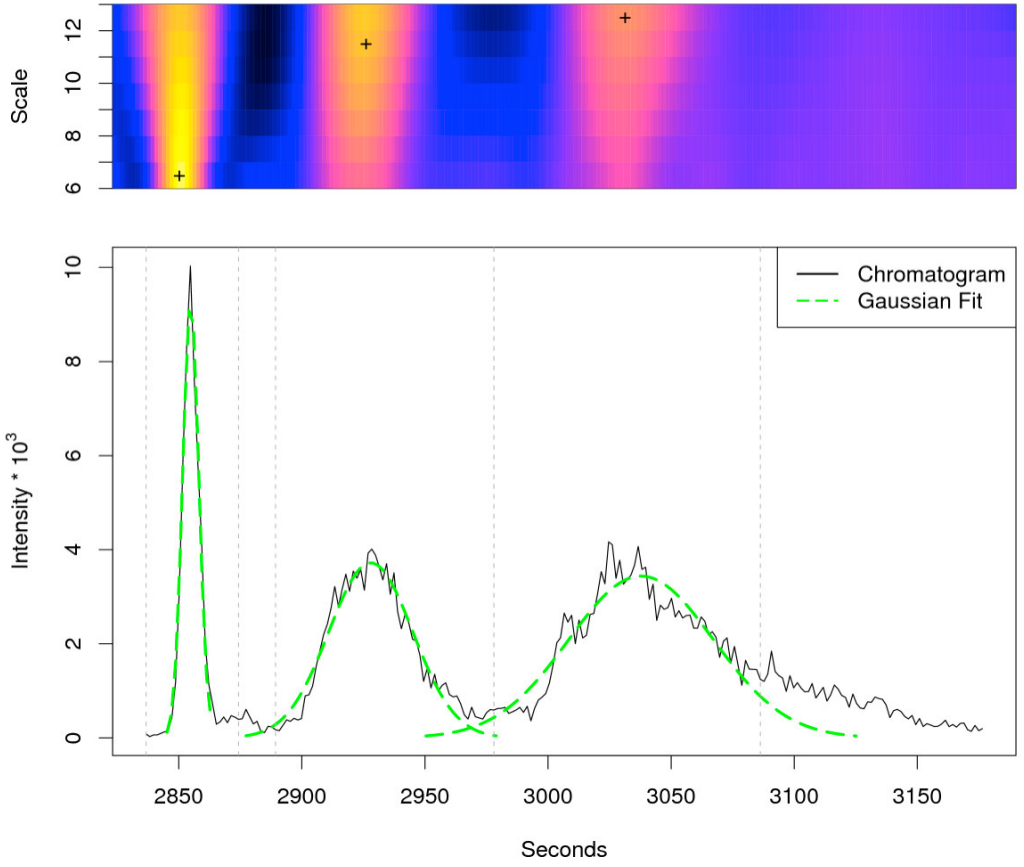

mass
120.08
120.06
120.04
120.02
50
100
150
200
250
300
scan
intensity
scans

Detect chromatographic peaks

- Use wavelet transform



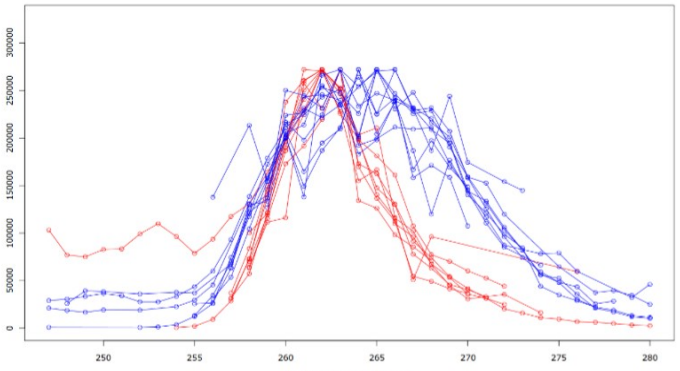
mexican hat wavelet



Feature filtering and grouping

- Measures of feature quality
  - S/N
  - Feature width
  - Abundance

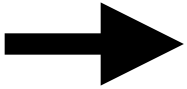
- Feature grouping
  - Similarity measure: normalized dot product



Feature annotation

adduct selection

Formula	N	Mass shift
[M+H]+	1	1.007276
[M+2H]+	1	2.014552
[M+3H]+	1	3.021828
[M+Na]+	1	22.98977
[M+K]+	1	38.963708
[M-C_3H_9N]+	1	-59.073499
[M+2Na-H]+	1	44.96563
[2M+Na]+	2	22.98977
[M+H-NH_3]+	1	-16.01872
[2M+H]+	2	1.007276
[M-OH]+	1	-17.0028

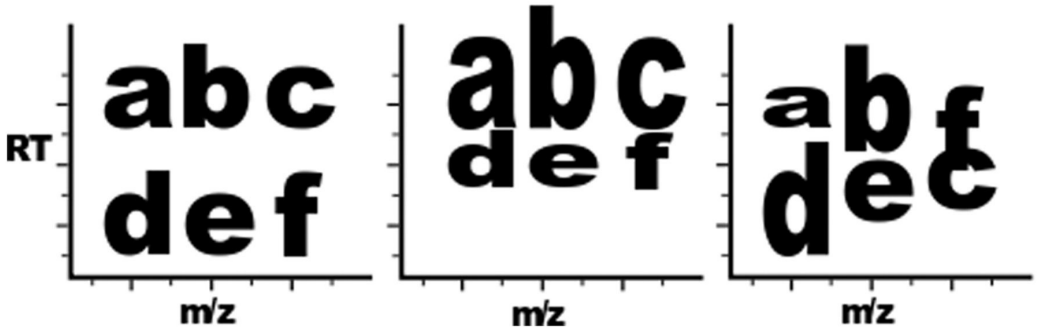


annotated features

id	mz	rt	isotopes	adduct	pc
65	176.04	280.09			4
76	136.05	280.43	[14][M+1]1+		5
77	135.05	280.43	[14][M]1+		5
74	153.06	280.43		[M+H]+ 152.05437	5
75	175.04	280.43		[M+Na]+ 152.05437	5
73	197.02	280.76		[M+2Na-H]+ 152.05437	5
78	377.74	286.15			6
79	732.5	286.49			6
83	488.32	286.82		[M+Na]+ 465.33205	7
82	466.34	286.82		[M+H]+ 465.33205	7
...					

Alignment

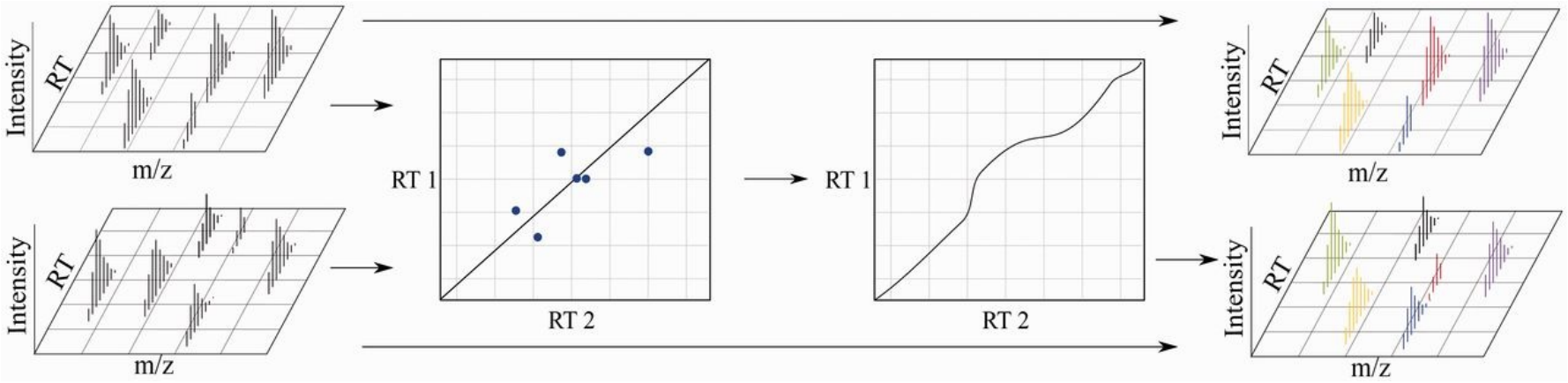
- Goal: Correct retention time shift from run to run



- Approaches
  - Warping
  - Direct match

Alignment approach: warping

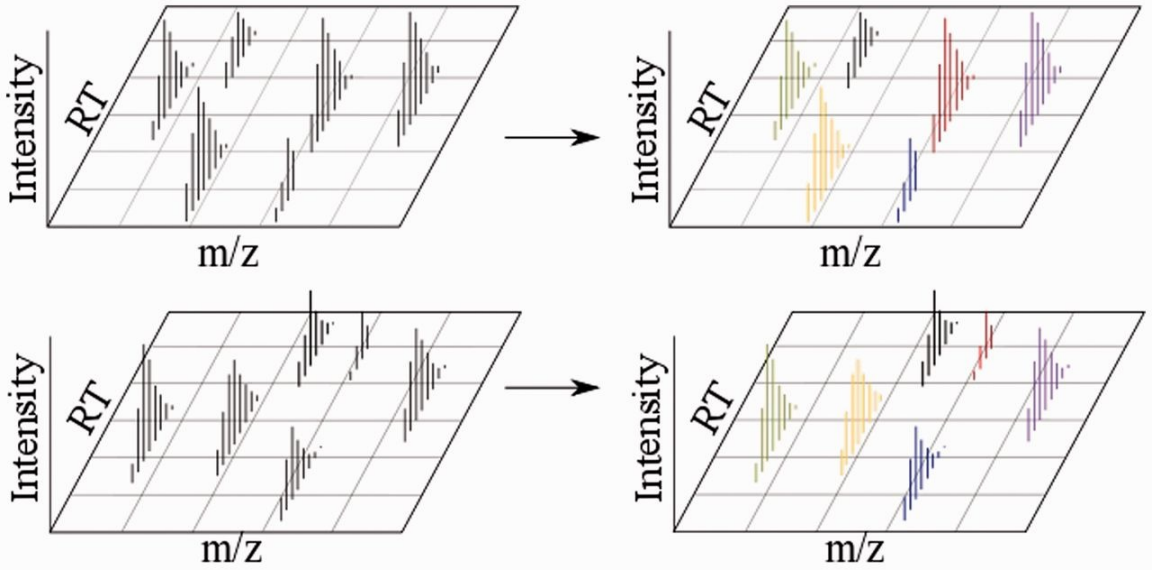
- Principle: Models systematic RT shift
- Limitation: the warping functions required for alignment are incapable of capturing component-level variation.



Smith, R.; Ventura, D.; Prince, J. T., LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* **2013**.

Alignment approach: direct matching

- Principle: analytes are matched directly based on factors such as elution time, charge state, and isotopic envelope characteristics.



Smith, R.; Ventura, D.; Prince, J. T., LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* **2013**.

Correspondence

- Limitations of warping
  - Warping models systematic shifting.
  - Warping functions are monotonic and cannot capture component-level variation.
  - Warping incorrectly assumes that elution order is preserved across runs.

- The problem should be casted as a correspondence problem.
  - Mapping of identically sourced features across all runs

Smith, R.; Ventura, D.; Prince, J. T., LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* **2013**.

Result of pre-processing

DB	Name	Mass	RT	platform	IN1	IN2	IN3	IN4	IN5	IN6
HMDB	1-Phenylethylamin	122.09745	24.97845	ES-	0.12862	0.1421305	0.1301326	0.1247924	0.1200045	0.1053275
HMDB	2-Ethylacrylic acid	101.06421	17.811575	ES-	0.0332025	0.0174262	0.0158166	0.0179326	0.0143742	0.0064953
HMDB	Canavanine	177.09653	10.338581	ES-	0.0141136	0.0134146	0.0182777	0.0193855	0.0245958	0.0011908
HMDB	Diketogulonic acid	193.03069	4.7050639	ES-	0.0209463	0.0203901	0.0165056	0.0189088	0.0137482	0.017231
HMDB	Iso-Valeraldehyde	87.080171	11.164359	ES-	0.6558109	0.2742277	0.2651933	0.3093793	0.2101024	0.0541026
in-house	3,4-Dehydro-Dprol	114.04431	3.5491023	ES-	0.2900544	0.287811	0.2290651	0.2754269	0.2314117	0.2061301
in-house	4-hydroxy-proline	132.05326	3.5958634	ES-	0.5584389	0.7353401	0.5273908	0.4412898	0.5074794	0.5423602
in-house	Malic acid	133.01996	3.9406386	ES-	0.0555016	0.0461576	0.0290383	0.0390783	0.0380952	0.0308288
in-house	2,3,4-Trihydroxybu	135.04472	3.5763487	ES+	0.0223984	0.0146371	0.0150894	0.0097238	0.0116862	0.0116129
in-house	2,3-Diaminopropic	105.07016	3.3202935	ES+	0.024859	0.0207034	0.0225235	0.0201288	0.0226763	0.0226569
in-house	4-Methy2-oxovaler	129.07306	16.624045	ES+	0.1341287	0.2458095	0.2138968	0.2383272	0.1646037	0.2156238
in-house	5-Aminopentanoic	116.0542	3.9125471	ES+	0.015214	0.0157145	0.0152048	0.0139855	0.0148445	0.0151512
in-house	Acetylcarnitine	204.12263	3.8790521	ES+	0.503742	0.4063954	0.3690539	0.3346704	0.1894332	0.267591
HMDB	11-beta-hydroxyan	483.25453	21.64161	ES+	0.0352862	0.0143528	0.0117155	0.0149876	0.0110671	0.003493
HMDB	13-Hydroperoxylin	313.23515	21.000715	ES+	0.012489	0.0124697	0.0117186	0.0120185	0.0129048	0.0116153
HMDB	17-Hydroxylinolen	295.22749	19.925457	ES+	0.0141132	0.0156397	0.0151444	0.0142477	0.0153367	0.015173
HMDB	2,4-Diaminobutyri	119.0844	3.8790898	ES+	0.0636478	0.0838566	0.0635174	0.067999	0.0942851	0.0625007
HMDB	2,6 dimethylheptar	302.23203	18.02586	ES+	0.0031349	0.0042189	0.0027814	0.0082044	0.002749	0.0032303
HMDB	2-Ethylhydracrylic	119.07199	15.226531	ES+	0.0236145	0.0239315	0.0242947	0.0237831	0.0239368	0.0242611
HMDB	2-Ketohexanoic ac	131.07027	3.7353582	ES+	0.0038071	0.0051703	0.0041894	0.0056894	0.0057567	0.0036369

Thank you!