Peak identification and alignment

Yutaka Yasui, Ph.D.
Dale McLerran, M.S.

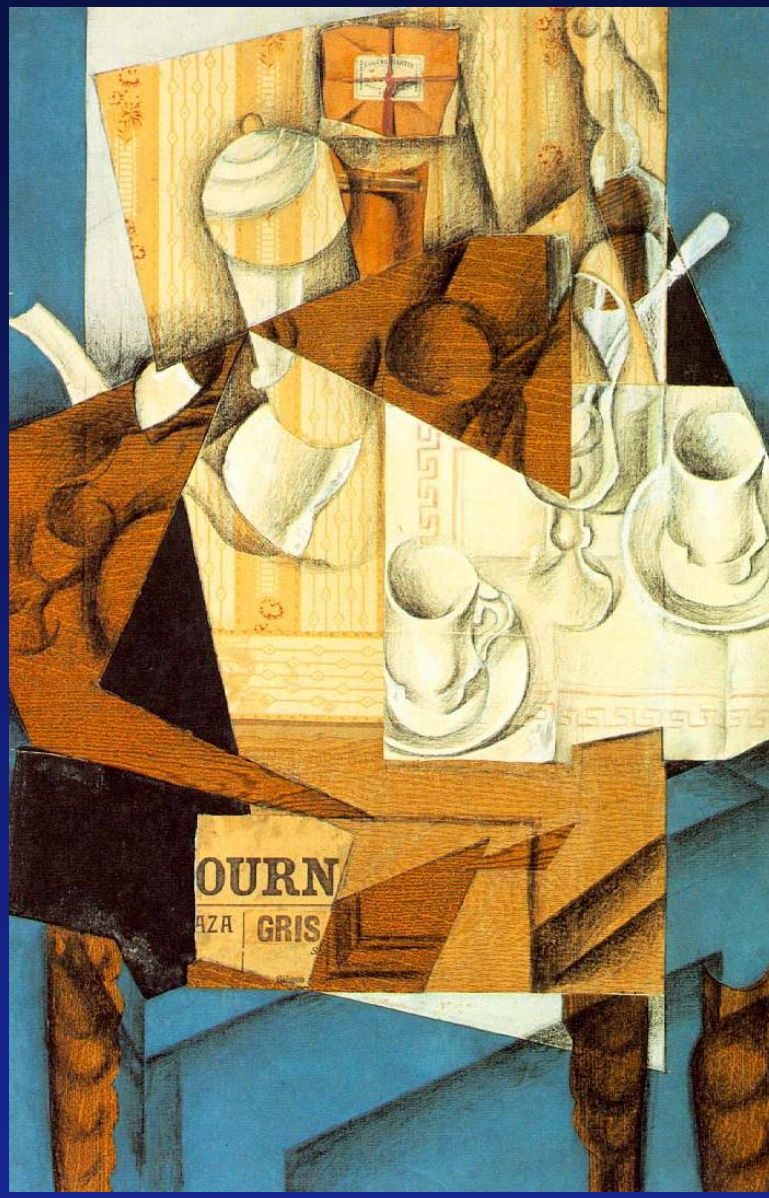
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center

Collaborators

NCI Early Detection Research Network
Sudhir Srivastava
Eastern Virginia Medical School
John Semmes
George Wright, Jr.
Bao-Ling Adam

Fred Hutchinson Cancer Research Center
Bree Mitchell
Phil Gafken
Paul Lampe
Johanna Lampe

Dale McLerran
Margaret Pepe
Tim Randolph
Yinsheng Qu
Li Hsu
Mary Lou Thompson
Mark Thornquist
Yan Liu
Toana Kawashima
Marcy Winget
John Potter
Ziding Feng

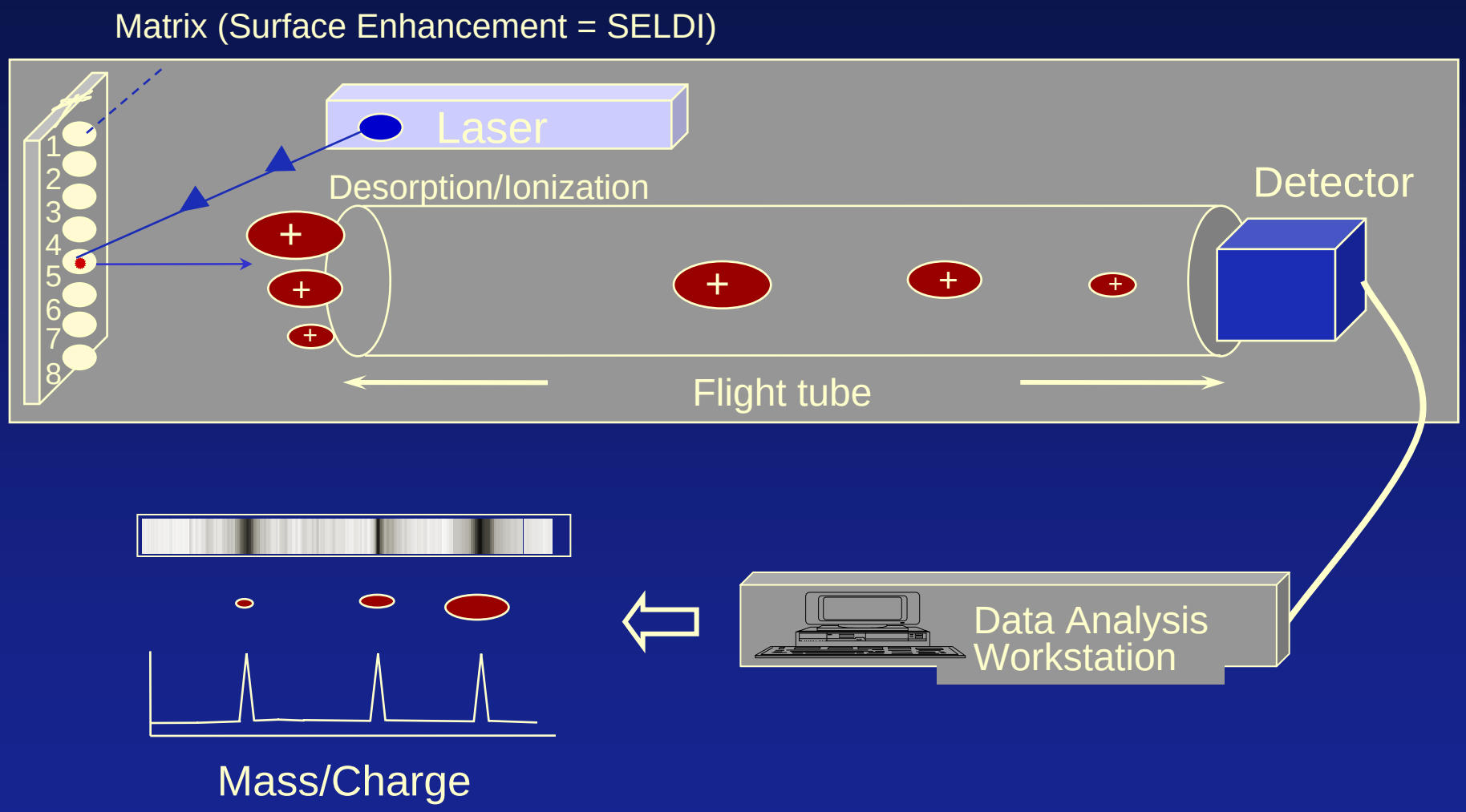


Recognition of a problem
in the current approach

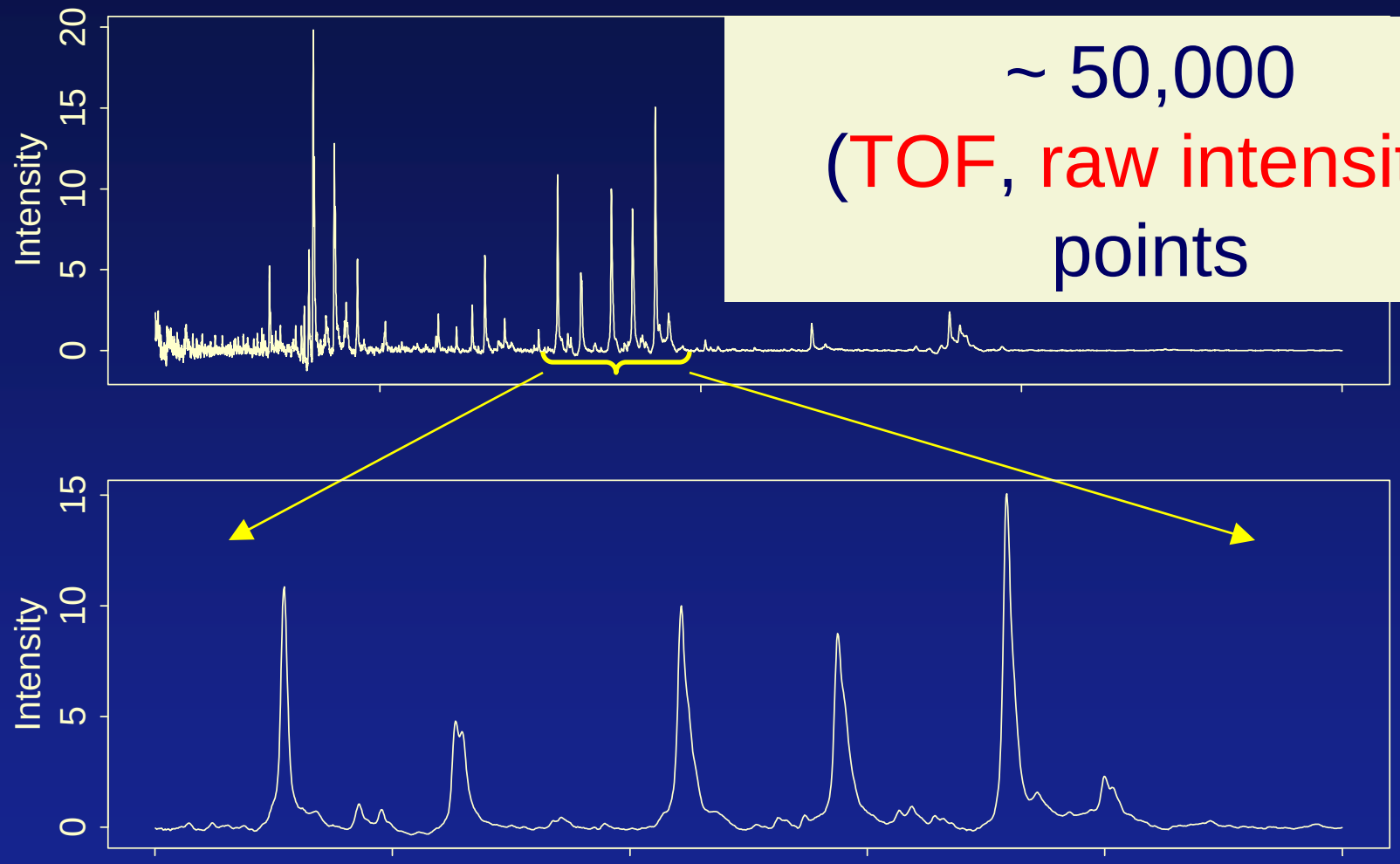
↓

Solution of the problem
via
modifications / new developments

Mass-spectrometry technology (MALDI, SELDI)



An example of SELDI output



~ 50,000
(TOF, raw intensity)
points

Analysis Steps

- Calibration from TOF to mass/charge (M/Z)
- Baseline subtraction / Normalization
- Peak identification
- Peak Alignment
- Search for signature profiles

1. Calibration

Conversion of TOF to mass/charge

$$\frac{\mathbf{m/z + pm}}{\mathbf{Voltage}} = \alpha(\mathbf{TOF} - \beta)^2 + \gamma$$

- Measure TOF of 7 (or 5) peptides with known m/z values
- Fit the above equation and estimate the parameters (α, β, γ)
- Apply the derived equation to convert TOF to m/z

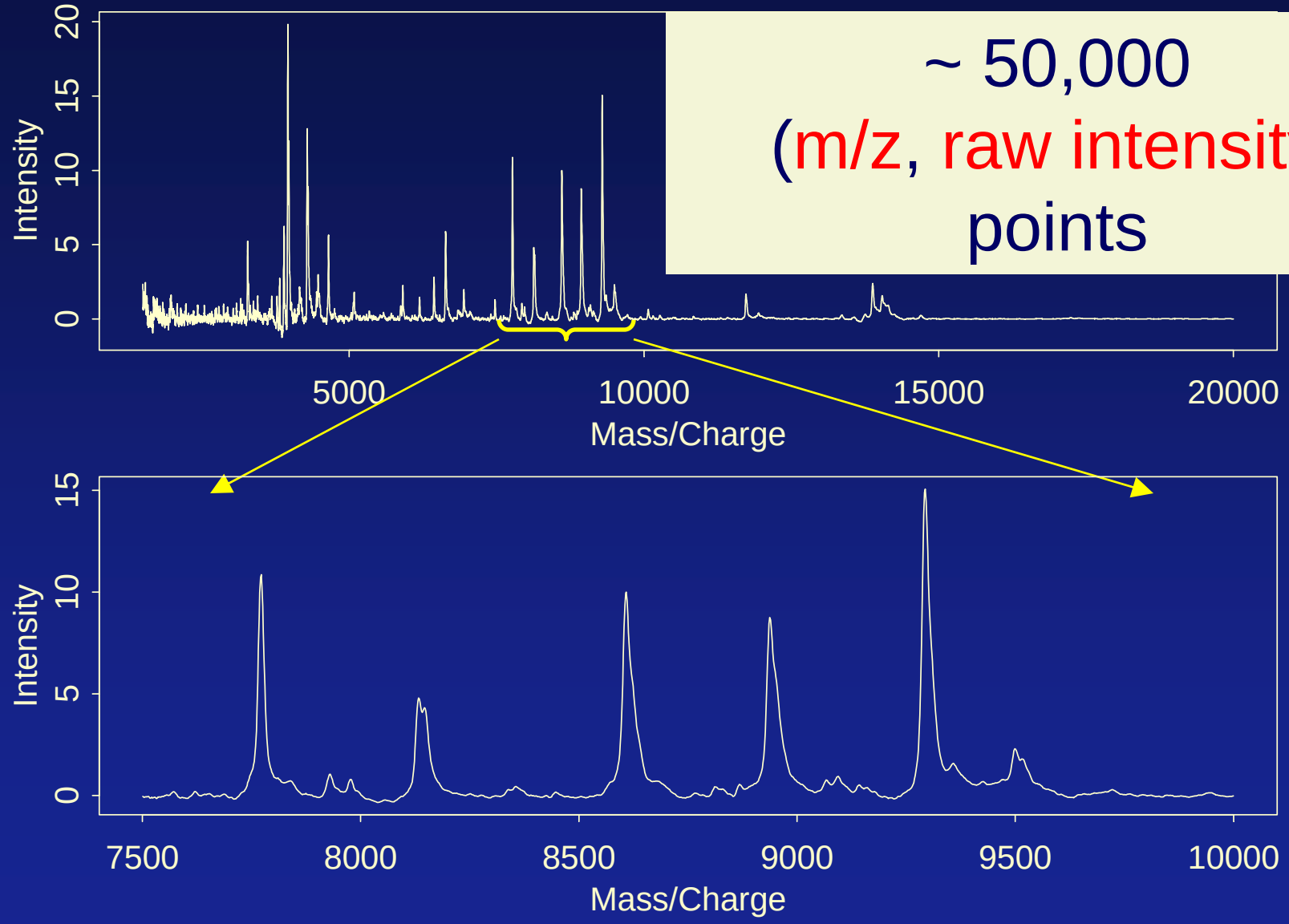
Calibration Issues

Goodness of fit with the 7 (or 5) standard peptides

Check the goodness of fit by eliminating one standard peptide: identify any “bad” standard peptide(s)

Over the course of an experiment,
what is the optimal schedule of calibration?
(once, multiple times, everyday, …)

After Calibration



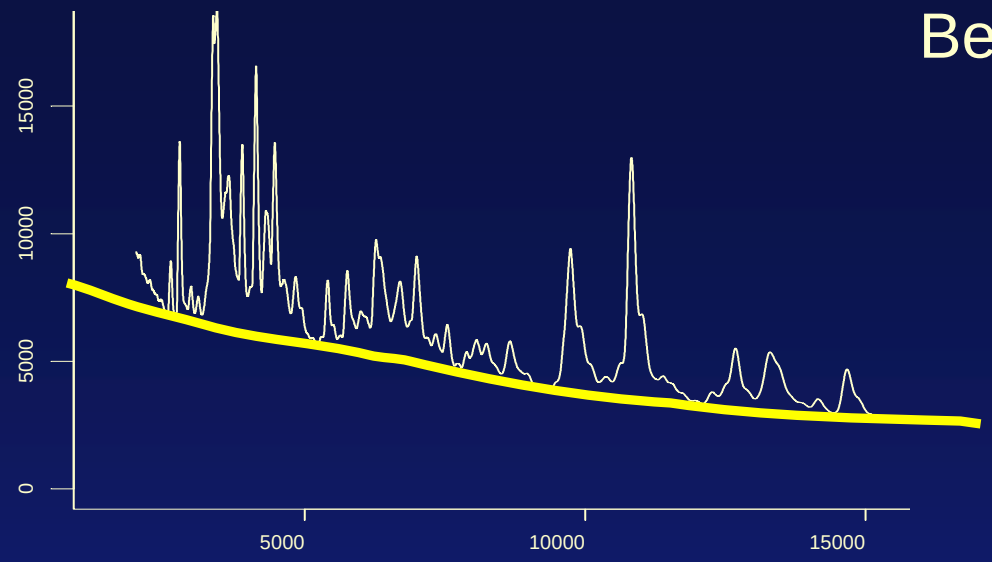
~ 50,000
(m/z, raw intensity)
points

2. Baseline subtraction & Normalization

Subtract the amount of intensity inflated by matrix

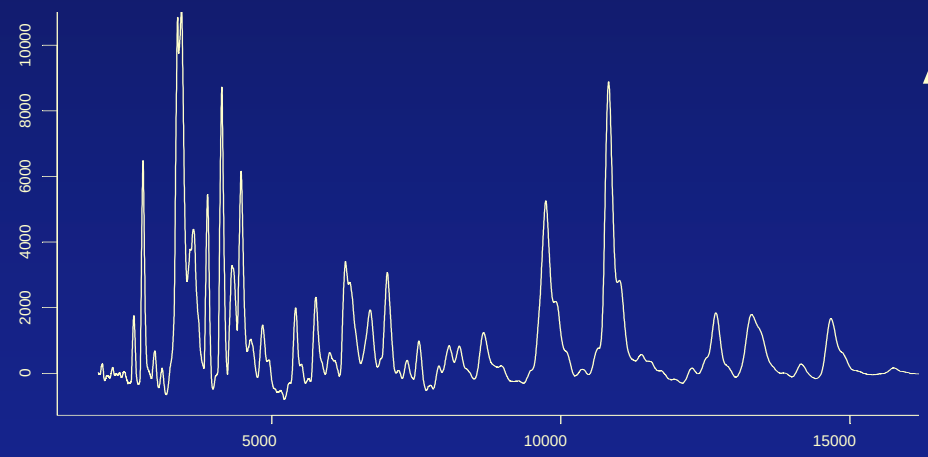
Scale the intensity to normalize spectra (total ion current)

Before baseline subtraction

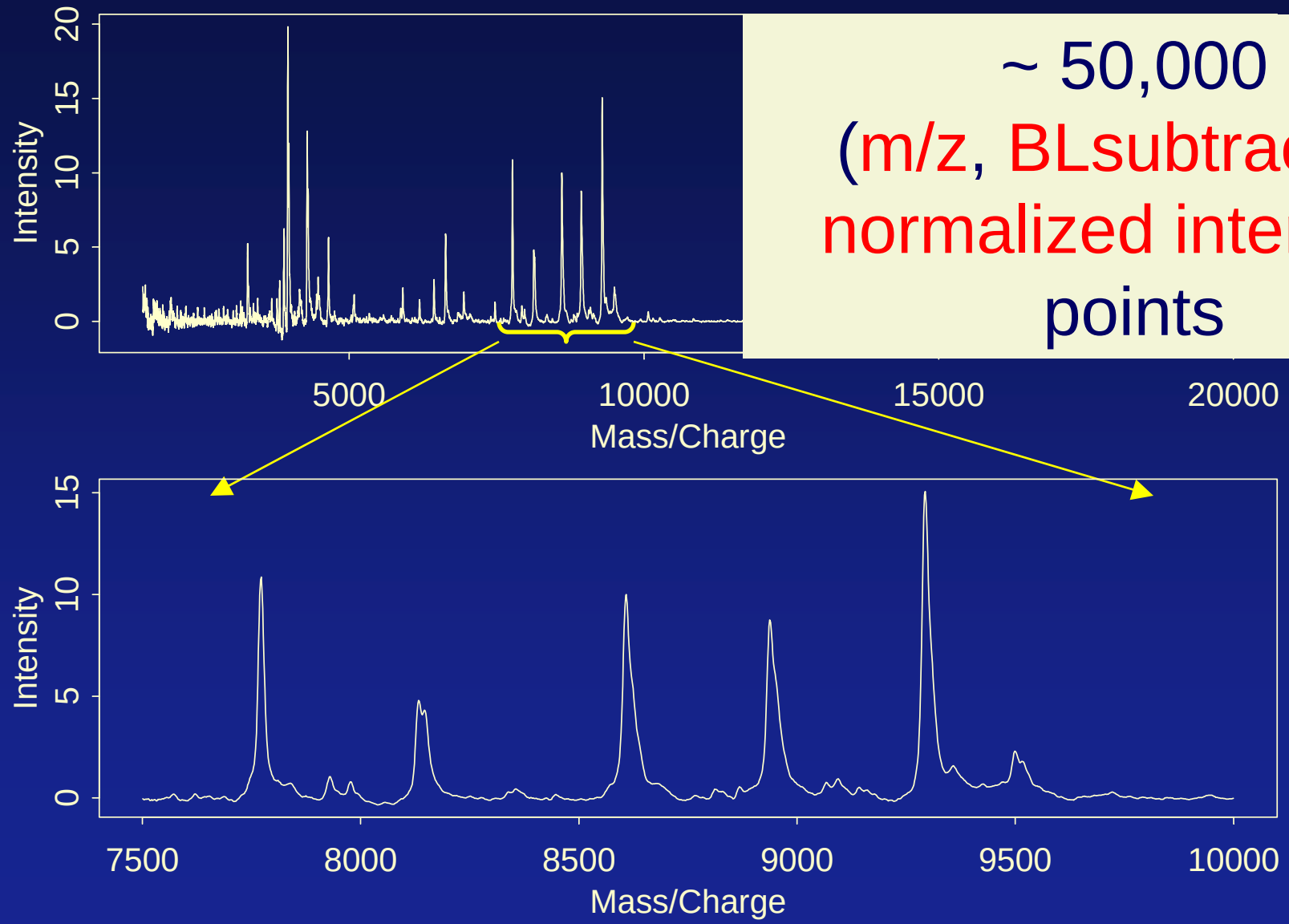


Baseline intensity due to matrix

After baseline subtraction



After Baseline Subtraction & Normalization



~ 50,000
(m/z, BLsubtracted-normalized intensity)
points

3. Peak Identification

A mathematical definition of peak locations

A critical issue in SELDI/MALDI-TOF analyses

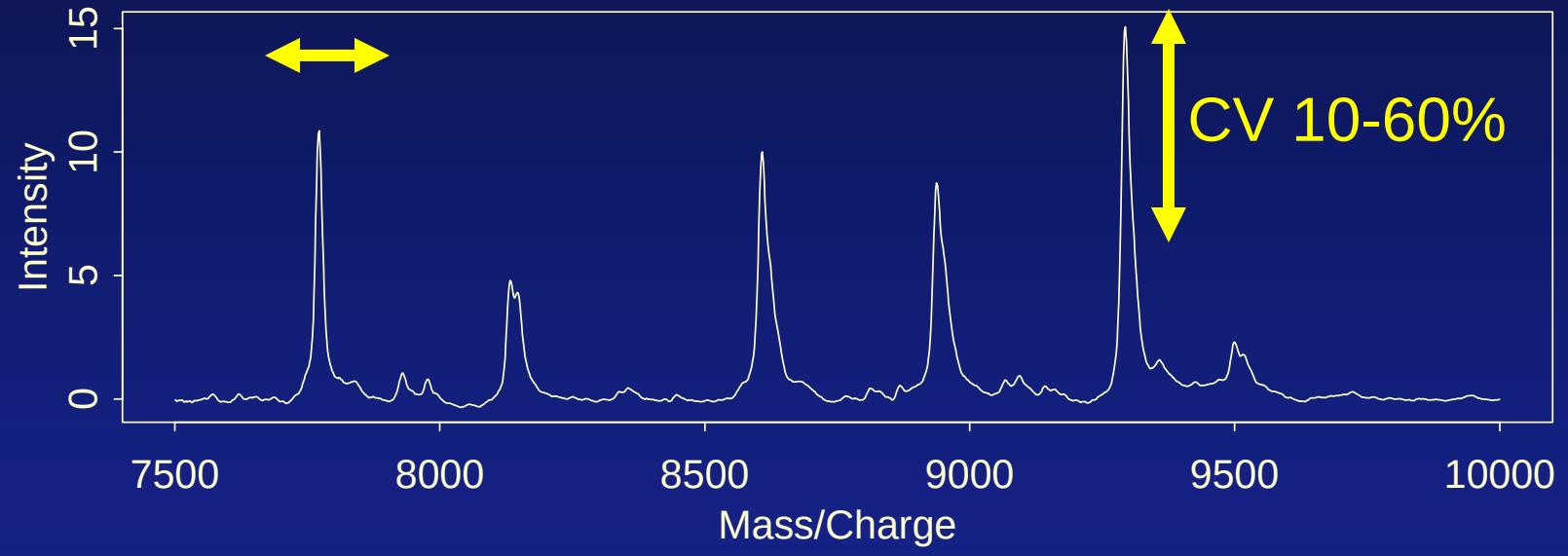
Two imprecision problems

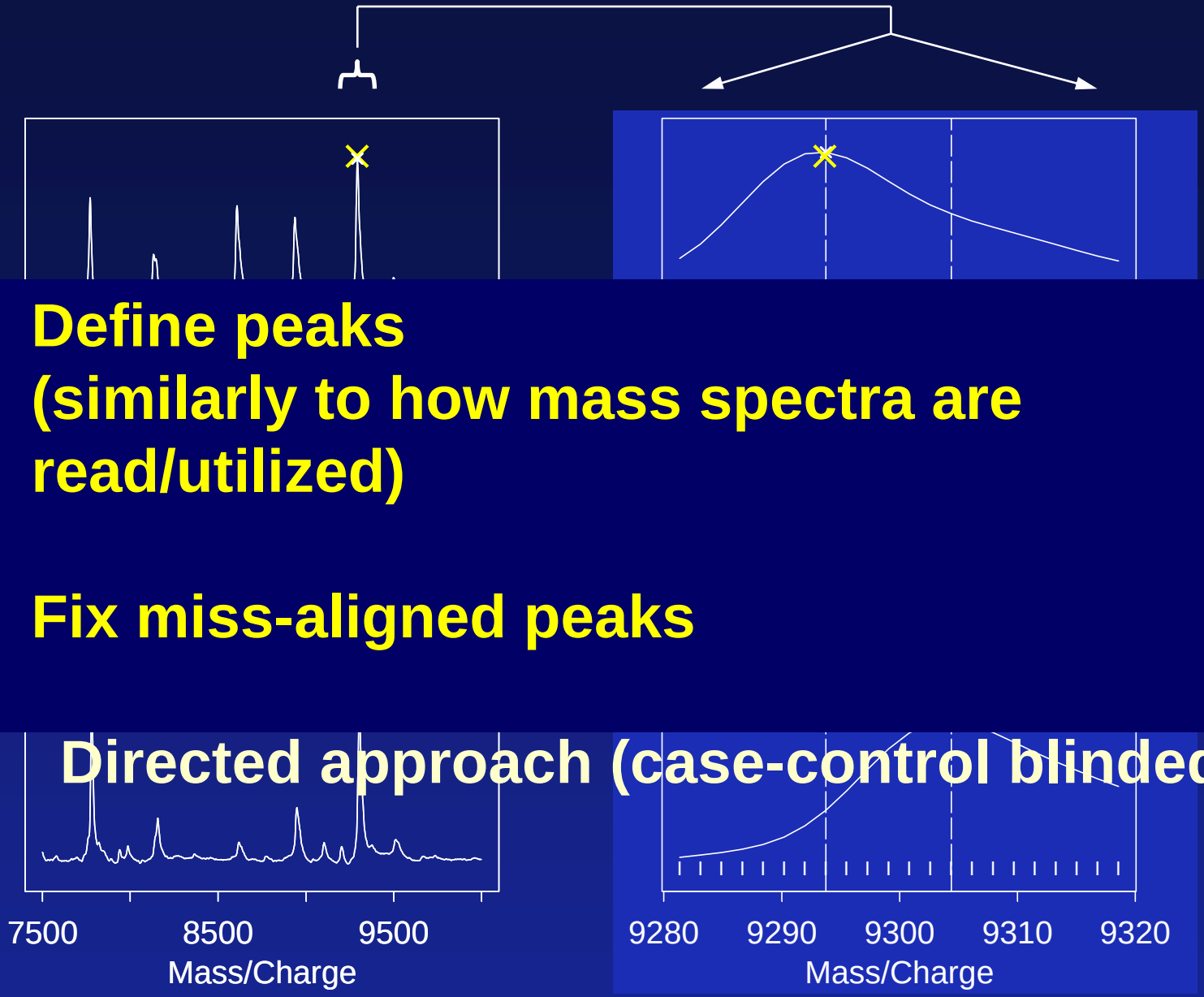
1. Imprecise measurements of mass/charge values (X-axi

2. Imprecise measurements of intensity values (Y-axis)

Properties of SELDI / MALDI-TOF output

Shift ± 0.1-0.2% of m/z (QC, not uniform)





1. **Define peaks (similarly to how mass spectra are read/utilized)**

2. **Fix miss-aligned peaks**

Directed approach (case-control blinded)

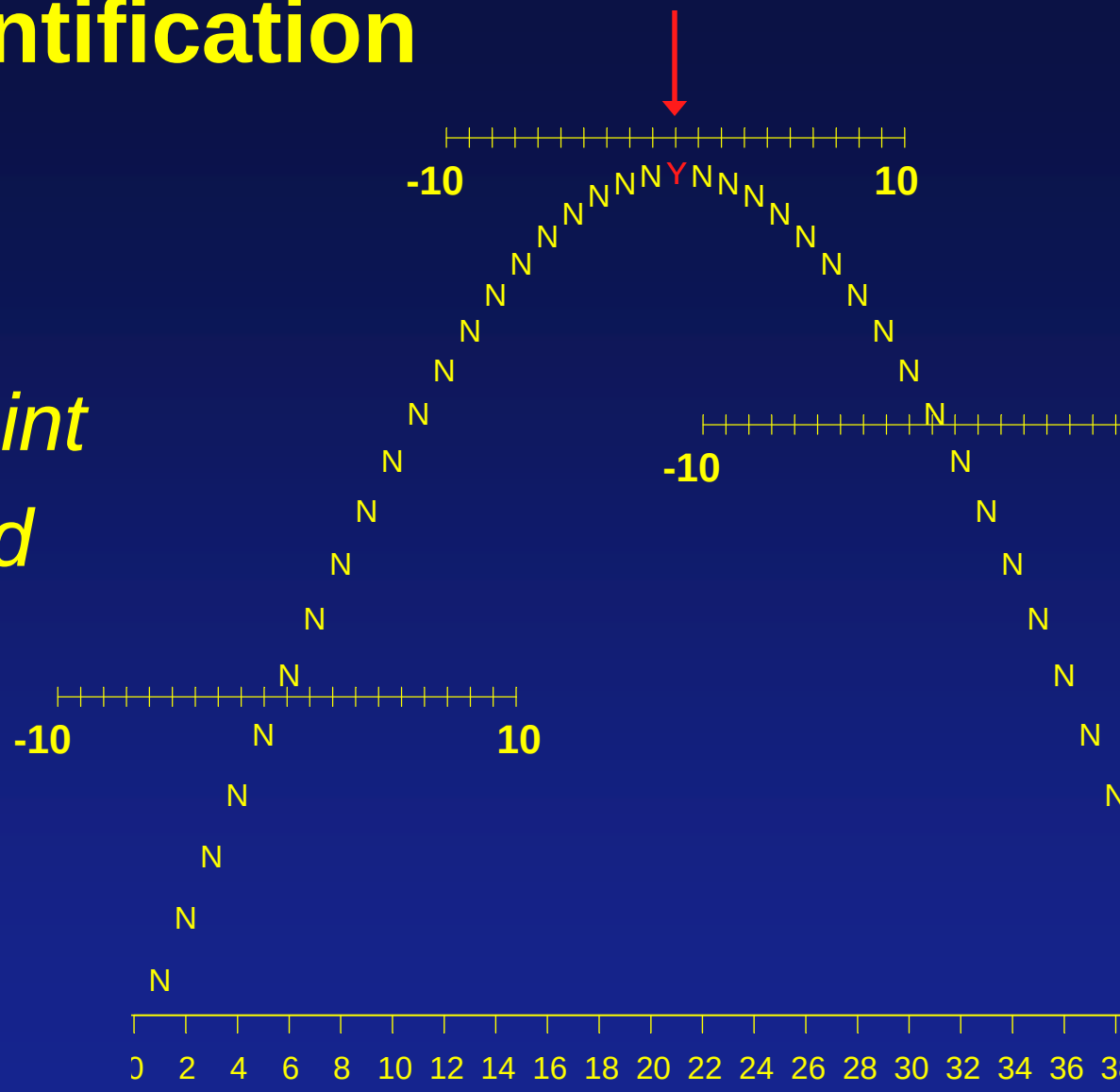
Peak identification

Ask at each point:

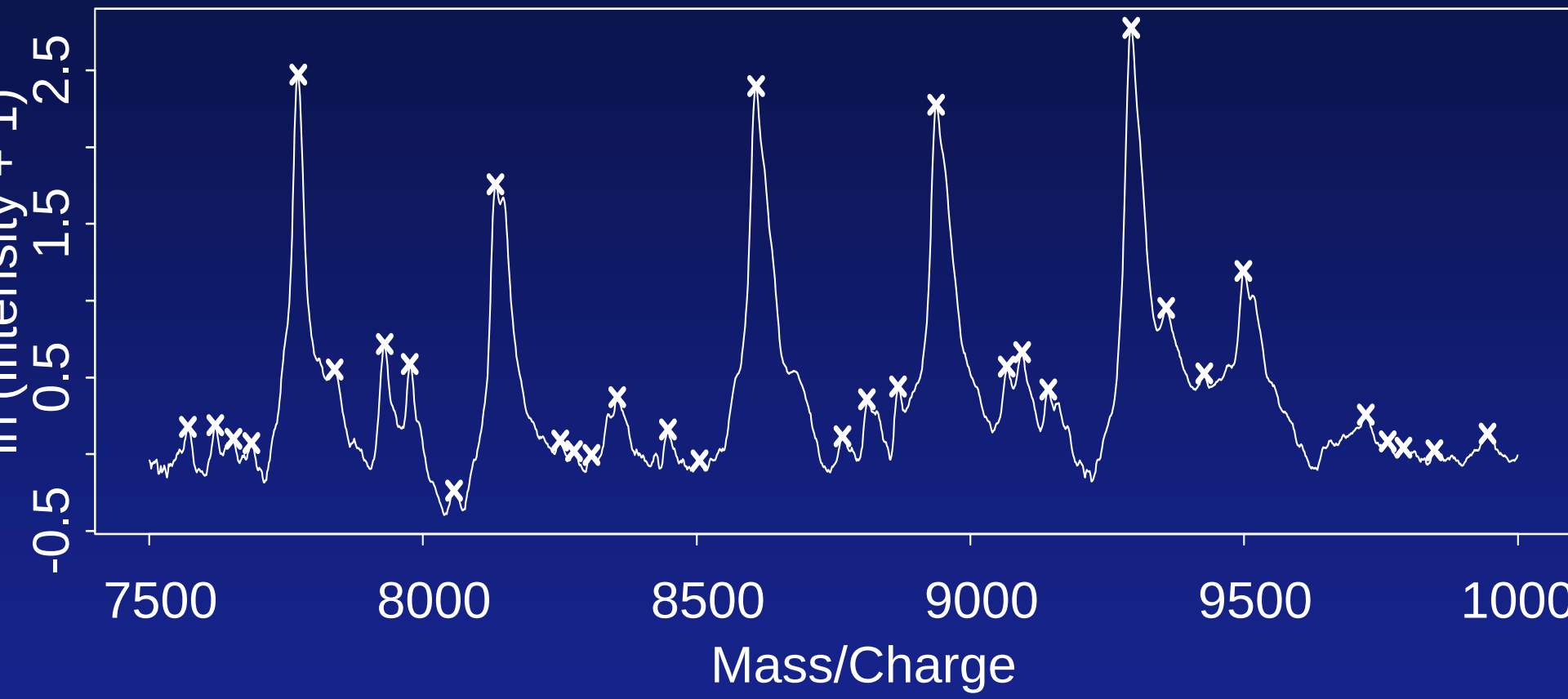
"Is it the highest point in the neighborhood of ±N points?"

YES = a peak

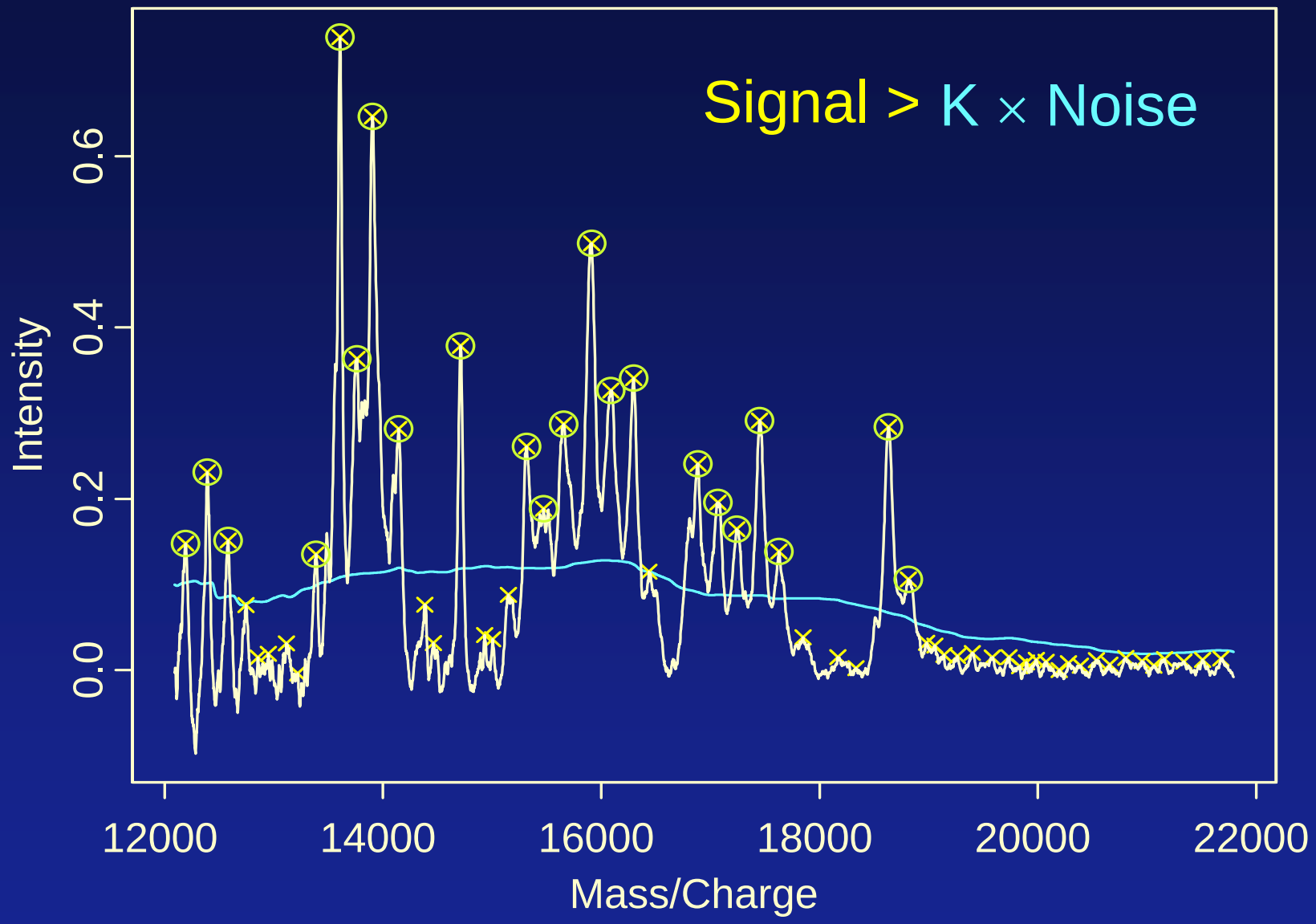
NO = not a peak



Peak identification results



Peak refinement



1. Moving median smoother with a wide span to obtain typical intensity values {(m/z, s)} locally

2. Moving median smoother with a wide span to obtain typical absolute deviation {(m/z, d)} locally

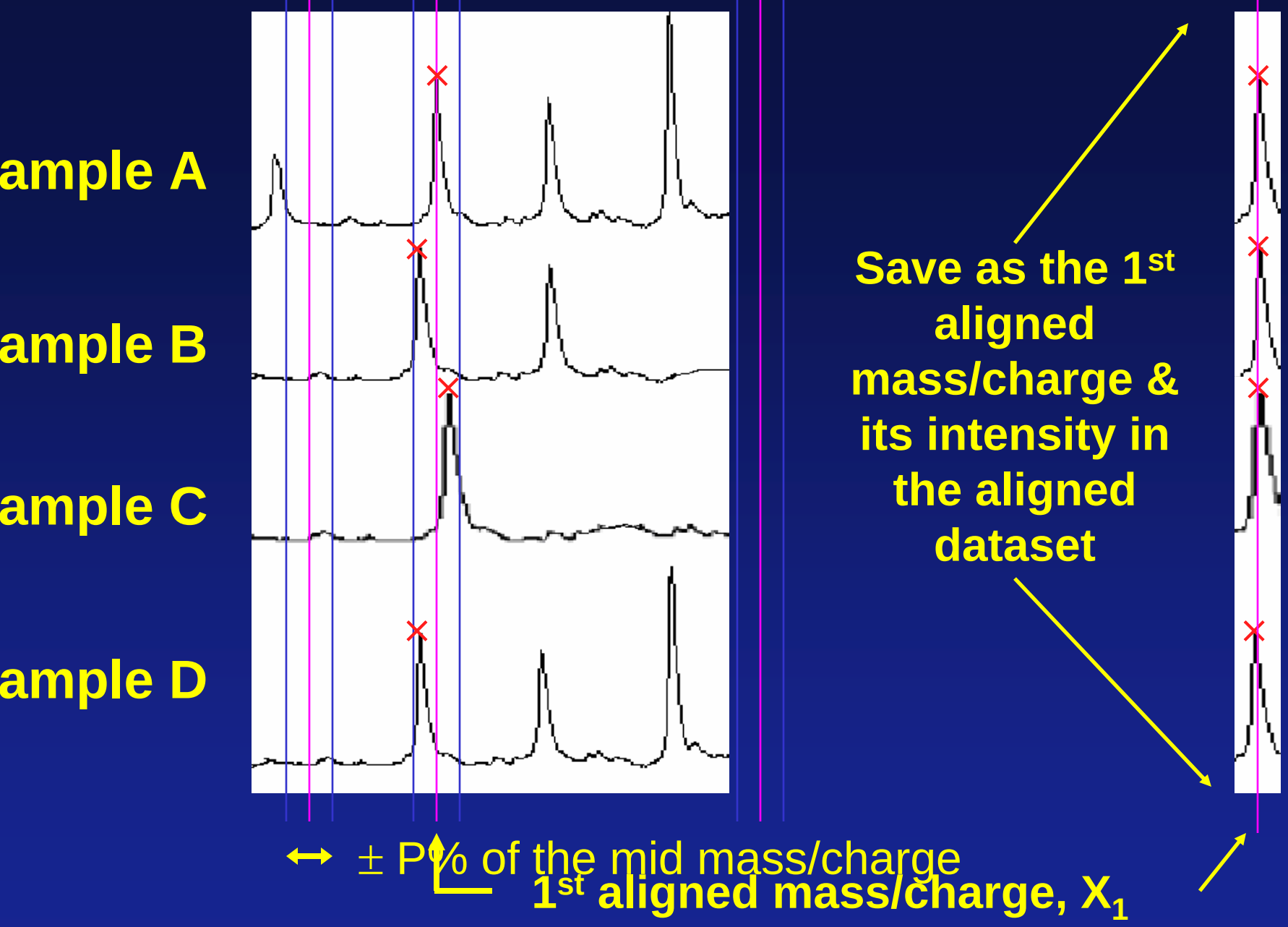
3. Define a peak if intensity > s + K * d

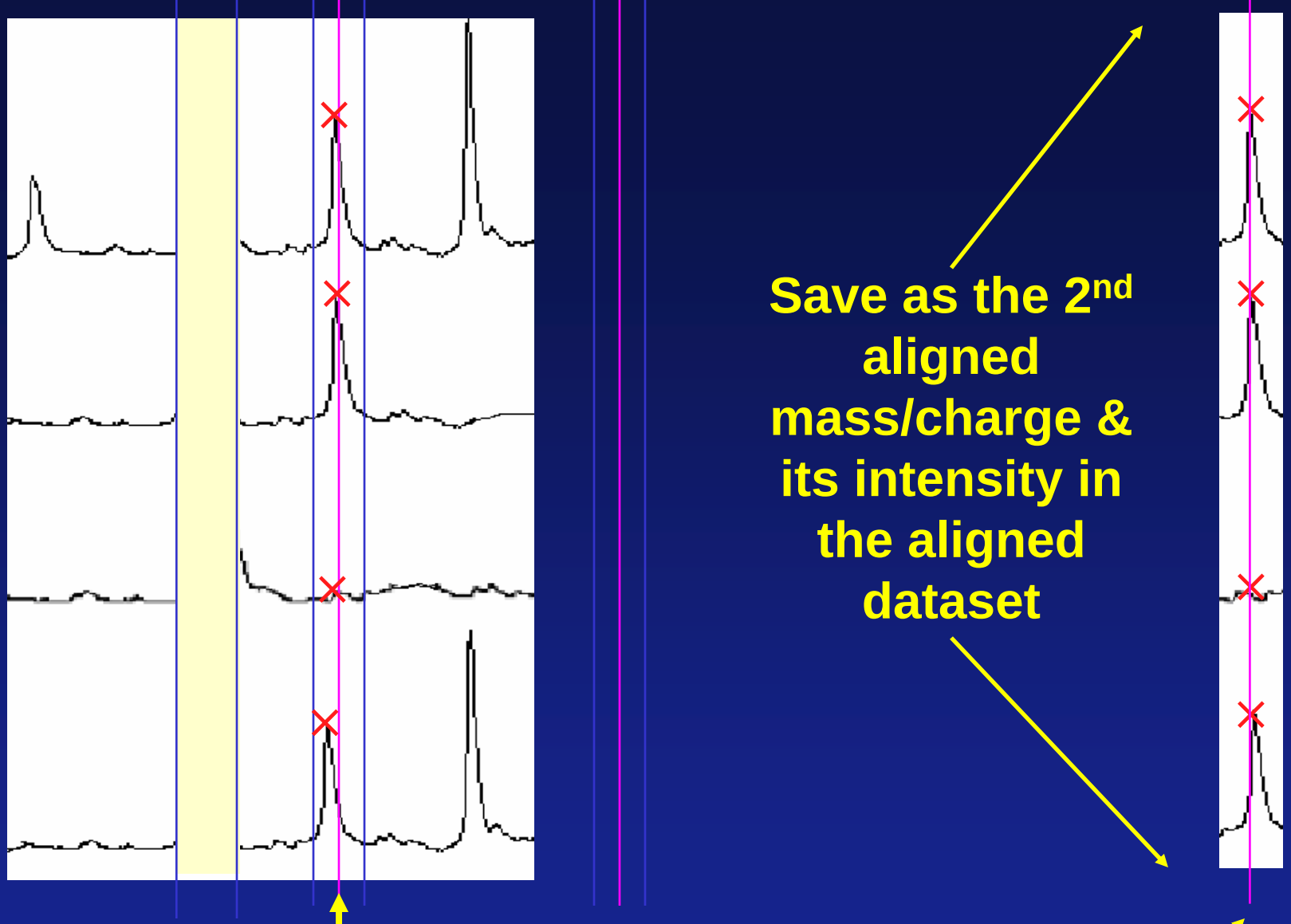
Peak identification

- Identify peaks in each spectrum
- The number of peaks per sample is ~1,000 (~ 2% of the original 50,000 points)
- High- and low-intensity peaks (vs. ~150 peaks by Coombes)
- Now, align peaks across samples (Alignment)

4. Peak Alignment

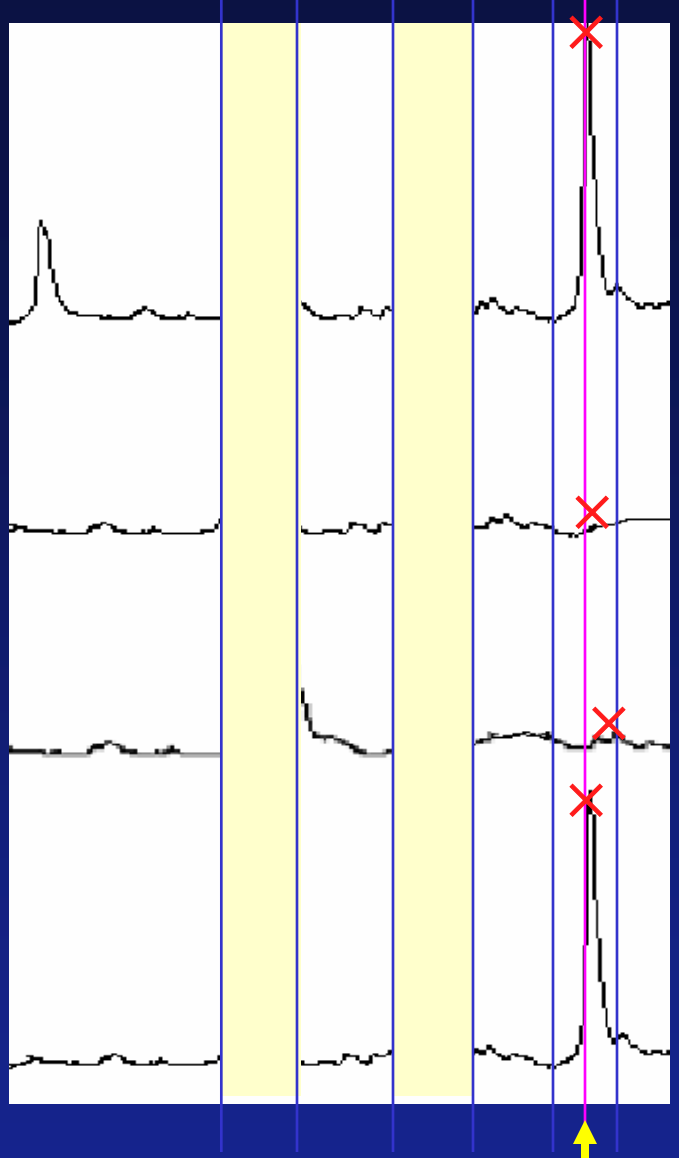
Correction of miss-aligned peaks across spectra


ample A
ample B
ample C
ample D
Save as the 1st aligned mass/charge & its intensity in the aligned dataset
± P% of the mid mass/charge
1st aligned mass/charge, X1

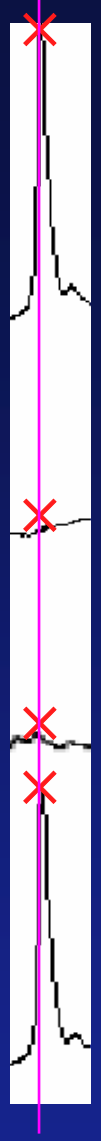


Save as the 2nd aligned mass/charge & its intensity in the aligned dataset

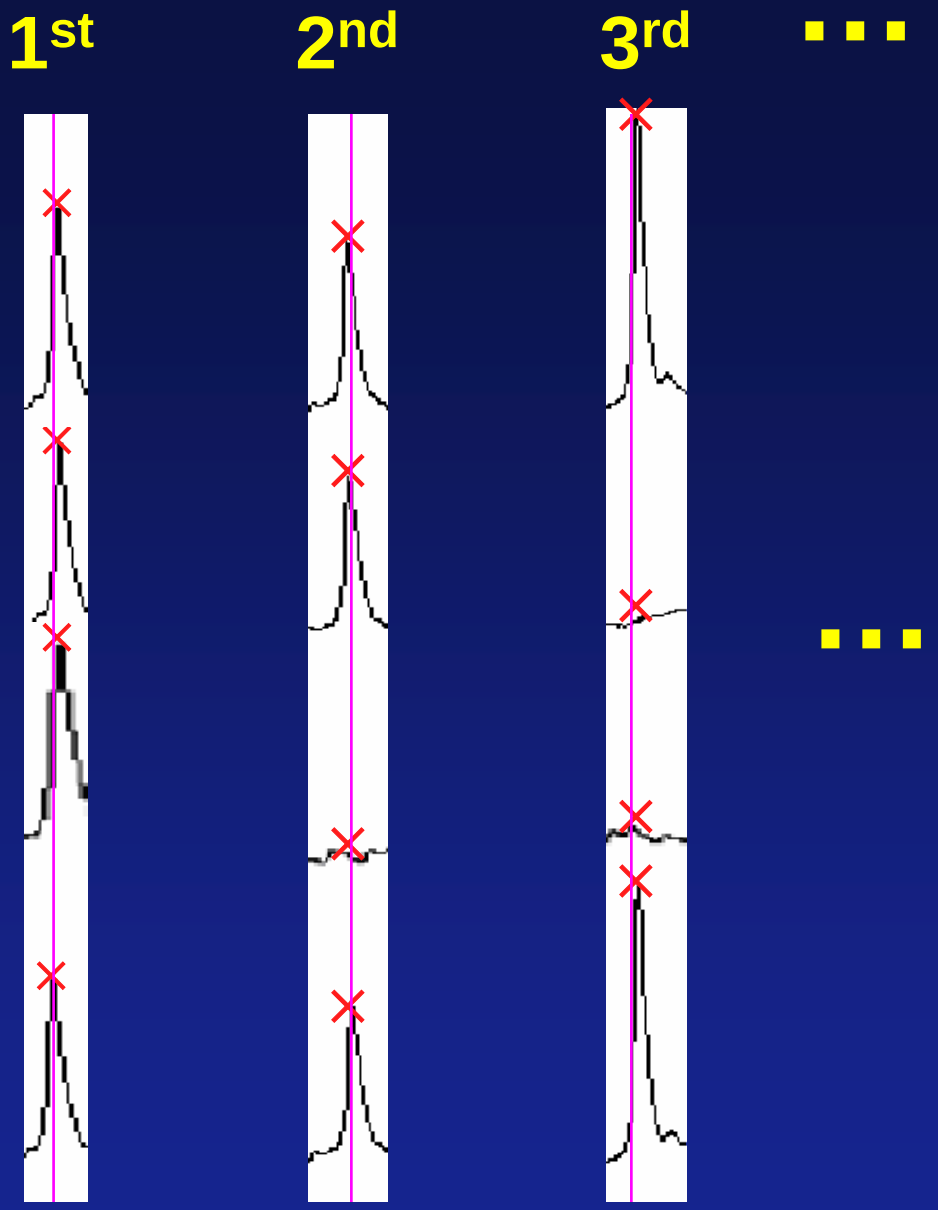
2nd aligned mass/charge, X_2



Save as the 3rd aligned mass/charge & its intensity in the aligned dataset



3rd aligned mass/charge, X_3



The aligned datase
for
searching
signature markers
profiles

Completion of pre-
analysis processing

Yasui et al. J. Biomed. & Biotech
(Special Issue on Proteomics) 20

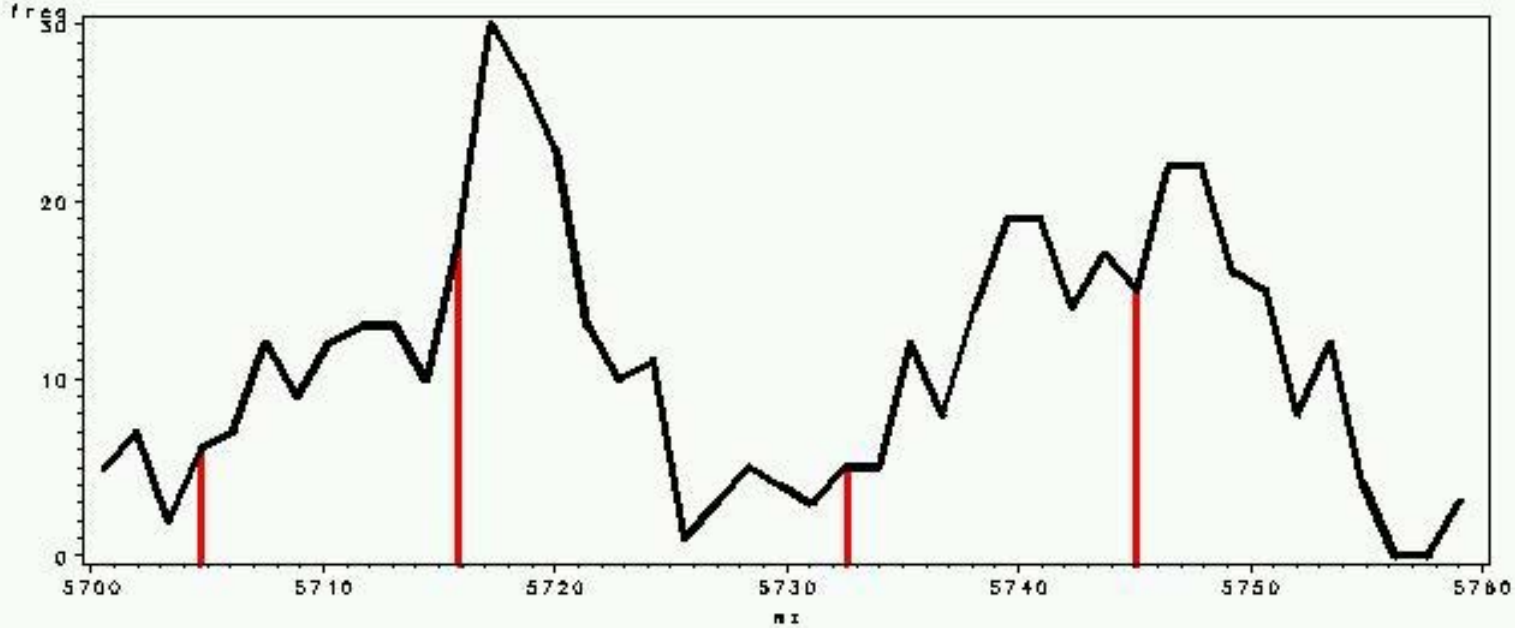
A new modification of our alignment algorithm by Dale McLerran

Instead of using the number of peaks,

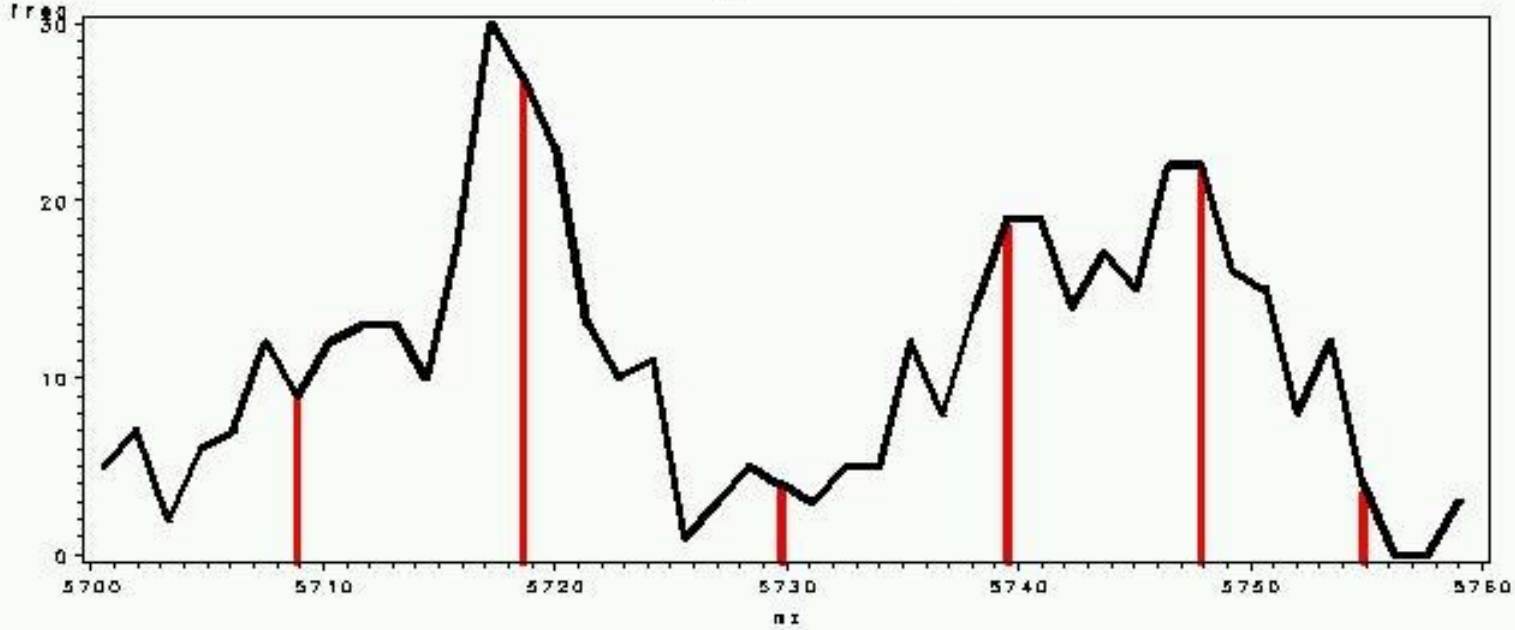
use a weighted sum (ws) of peak counts in a window

- If $b < \min(a, c)$ then $ws = 0$
- If $b > \max(a, c)$ then $ws = a + \lambda b + c$ where $\lambda = b / \max(a$
- Otherwise $ws = a + b + c$

Original alignment method peak locations



Overweight centers



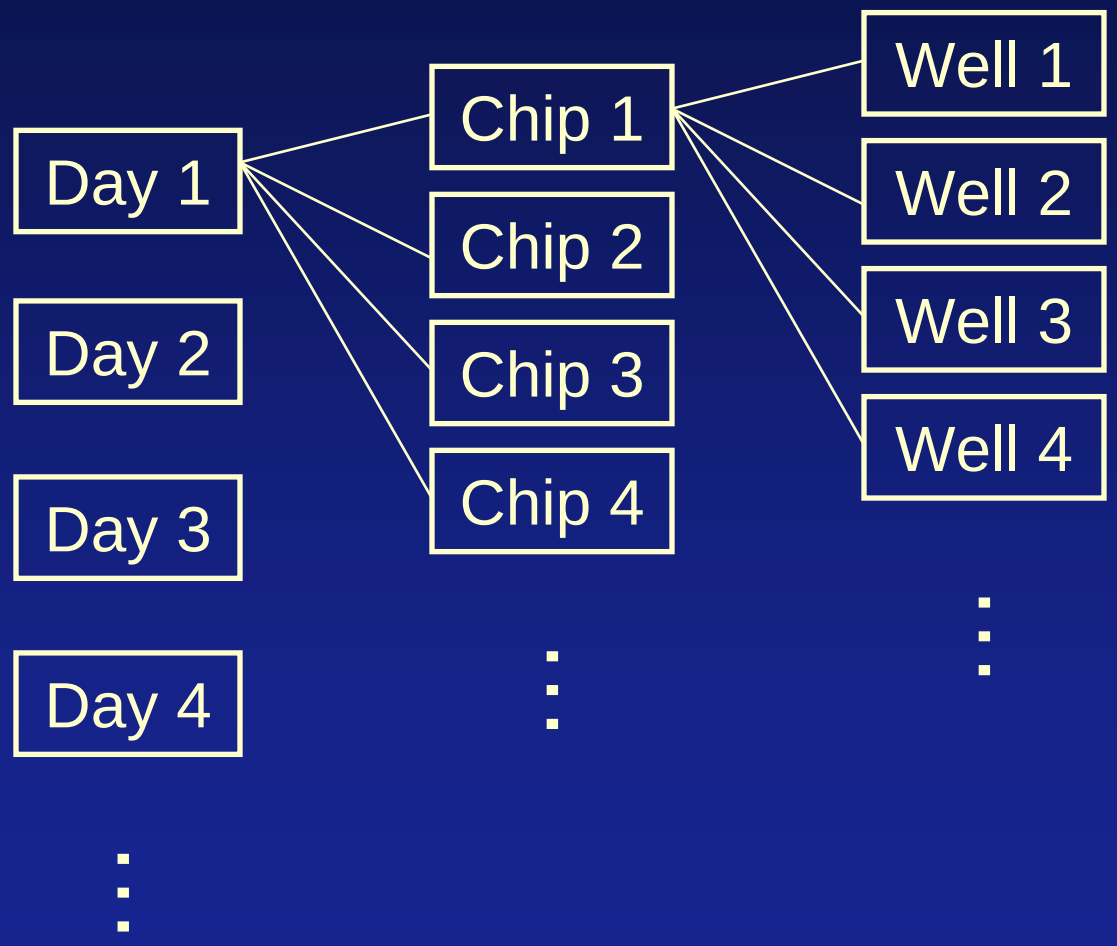
Imprecise measurements of intensity values (Y-axis)

Identify and quantitate sources of variations

Replicate the measurements at high-variation sources

Variance components assessment

Repeated measurements of a single QC sample



Sources of intensity variation

V_{day} = Day-to-Day

V_{chip} = Chip-to-Chip

V_{well} = Well-to-Well

Reduction of CV by averaging 3 replicates

$$CV = (\text{Variance of intensity})^{1/2} / \text{Mean}$$

$$= (V_{day} + V_{chip} + V_{well})^{1/2} / \text{Mean}$$

If spotted on 3 wells of a chip,

$$CV_3 = (V_{day} + V_{chip} + 1/3 \times V_{well})^{1/2} / \text{Mean}$$

Reduction of CV by averaging 3 replicates

If measured on 3 different days,

$$CV_3 = \{1/3 \times (V_{day} + V_{chip} + V_{well})\}^{1/2} / \text{Mean}$$

$$= CV / \sqrt{3}$$

(e.g., CV = 20% then CV_3 = 11.5%)

Summary

- Proper calibration/normalization is critical
- Imprecise measurements of m/z values necessitate the identification and alignment of peaks
- Simple algorithms have been developed and available
- Further refinements and alternative approaches are possible (need quick developments even if not optimal)
- Replication alleviates the imprecision problem of intensit


P. Gauguin 88