Time-of-flight Mass Spectroscopy and Disease Diagnosis: A Case Study

E. R. Tracy
Physics Department
William and Mary

Collaborators

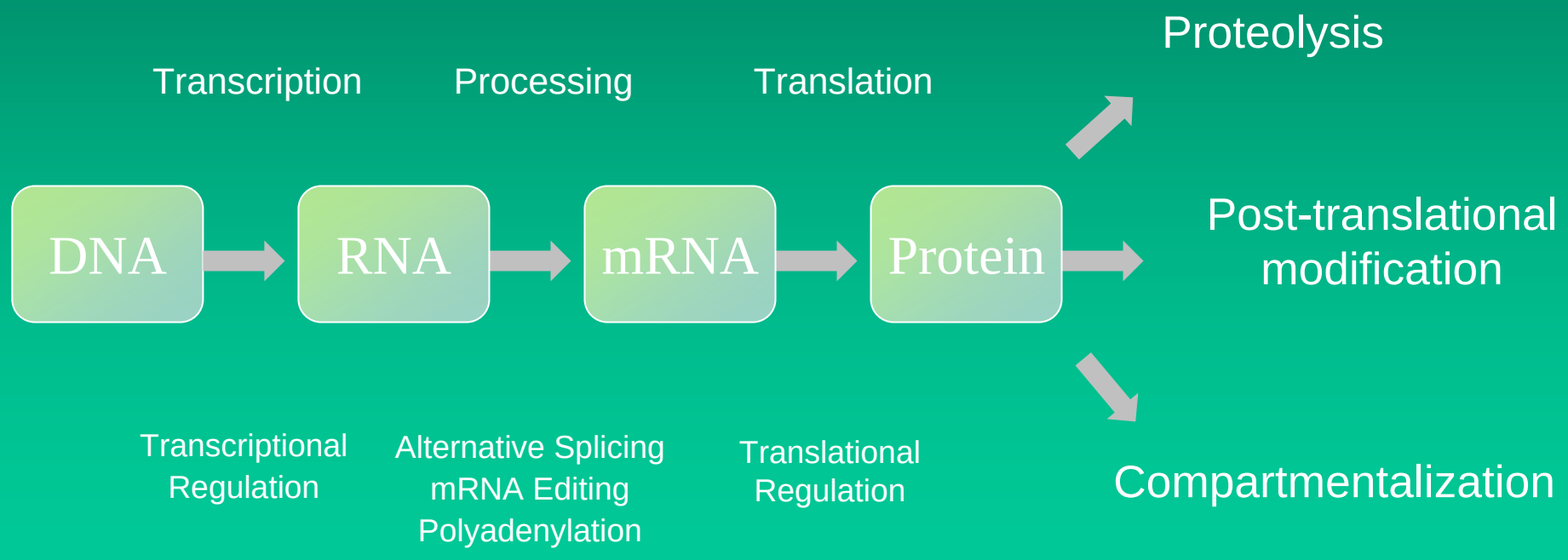
- William & Mary: H. Chen, W. Cooke, P. Harris, C. Hopkins, M. Lewis, D. Manos, D. Malyarenko, E. Tracy, M. Trosset, L. Wei
- INCOGEN: M. Sasinowski, H. Sasinowska, J. Miller, Y. Li
- EVMS: J. Semmes, R. Drake, G. Malik
- Applied Research Center: A. Wilkerson

Genome Sizes and Number of Genes

• HIV Virus:	9,750 bp	1,000 genes
• E. coli	4,600,000 bp	5,400 genes
• Yeast	12,000,000 bp	8,000 genes
• Fruitfly	180,000,000 bp	13,400 genes
• Chicken	1,200,000,000 bp	15-20,000 genes
• Human	3,400,000,000 bp	25-35,000 genes
• Gorilla	3,500,000,000 bp	~ same as human
• Maize	5,000,000,000 bp	?
• Amoeba	670,000,000,000 bp	?

Genes vs. Proteins

- ‘Proteomics’: the study/analysis of the entire protein complement in a given cell, tissue, or organism



One gene can ‘code’ for many proteins.

Protein Variation in Size

	Molecular weight	Number of residues	Number of polypeptide chains
Cytochrome *c* (human)	13,000	104	1
Ribonuclease A (bovine pancreas)	13,700	124	1
Lysozyme (egg white)	13,930	129	1
Myoglobin (equine heart)	16,890	153	1
Chymotrypsin (bovine pancreas)	21,600	241	3
Chymotrypsinogen (bovine)	22,000	245	1
Hemoglobin (human)	64,500	574	4
Serum albumin (human)	68,500	609	1
Hexokinase (yeast)	102,000	972	2
RNA polymerase (*E. coli*)	450,000	4,158	5
Apolipoprotein B (human)	513,000	4,536	1
Glutamine synthetase (*E. coli*)	619,000	5,628	12
Titin (human)	2,993,000	26,926	1

Mass Spec Proteomics

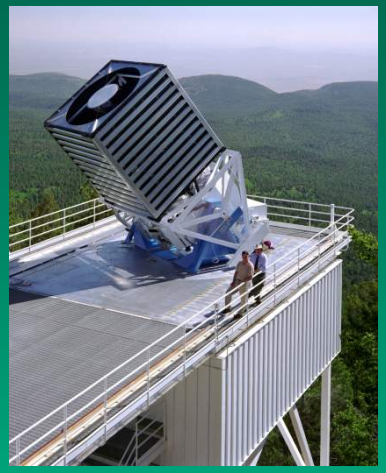
Use MS to identify protein ‘biomarkers’ that are indicative of disease.

- Survey: scan a wide range of masses and look for distinctive patterns.
- ID: once patterns are found, what proteins are they?
- Model/predict: once the relevant proteins are known, can we improve our understanding of the disease process? Can we use this insight to improve treatment (e.g. targeted drug therapy)?
- Of course, not only proteins, but genes, metabolites, and other information will be needed to fill out the picture.
- Reliable ID of proteins, and validation of the models, generated will require careful attention to data quality and characterization of uncertainties.

Data quality vs. data quantity

- Poincare: “One good measurement is worth more than a thousand poor ones.”
- The Chinese Emperor fallacy

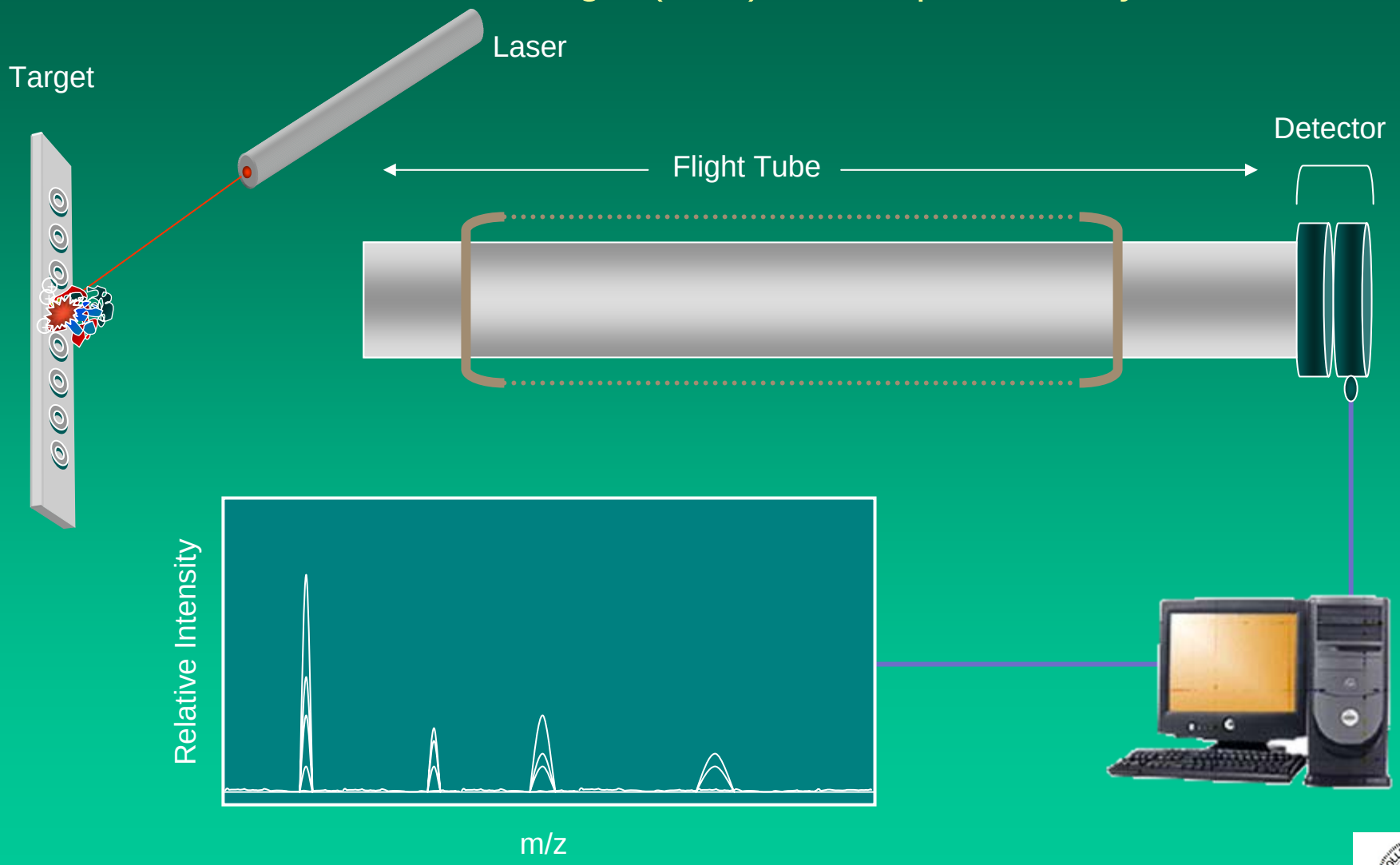
Survey vs. ID



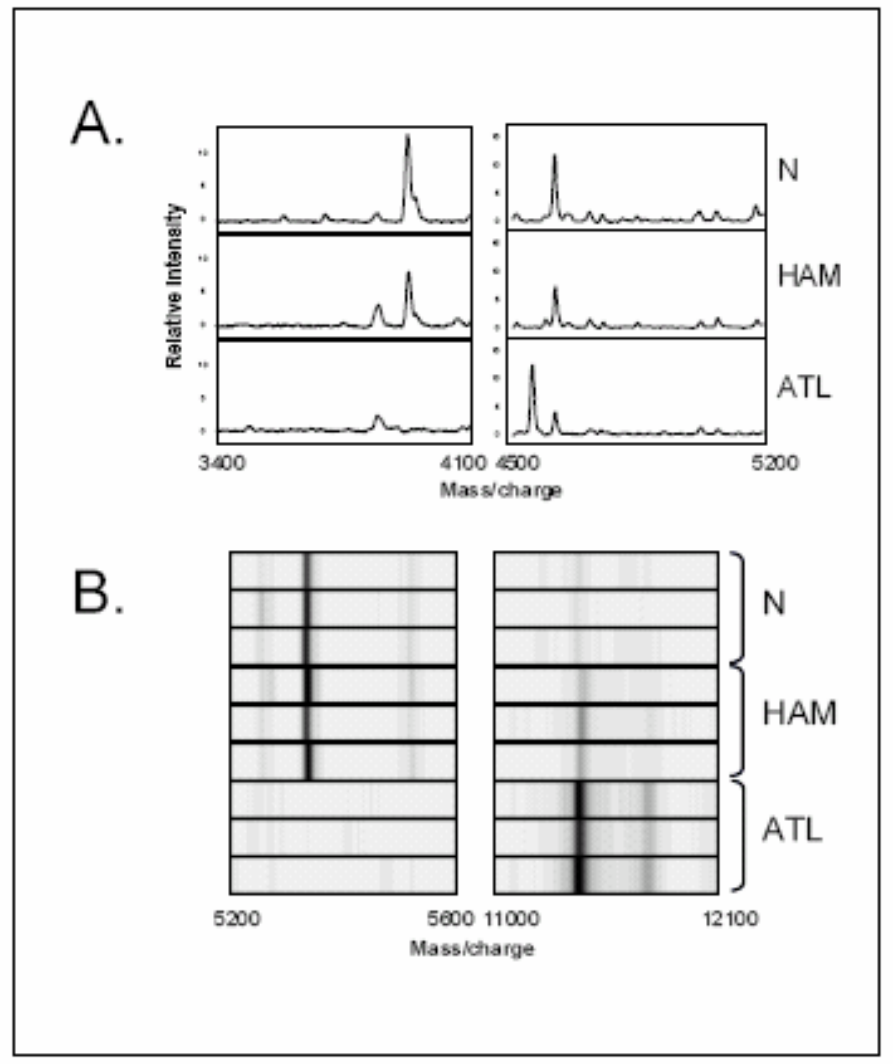
Sloan Sky Survey Telescope

Hubble Space Telescope

Laser Desorption/Ionization
Time-Of-Flight (TOF) Mass Spectrometry



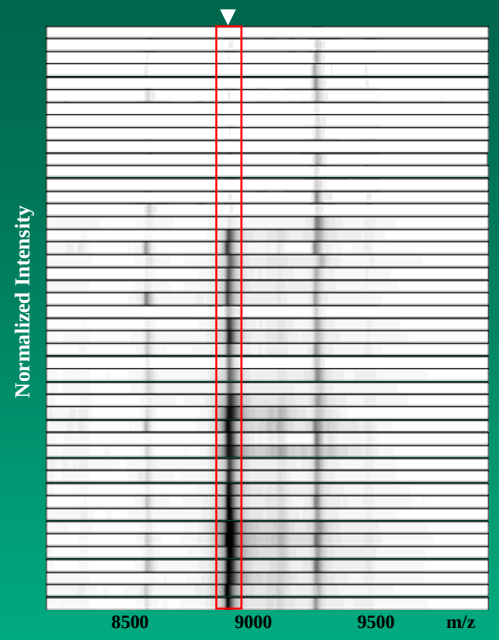
EASTERN VIRGINIA MEDICAL SCHOOL
ESTABLISHED 1973



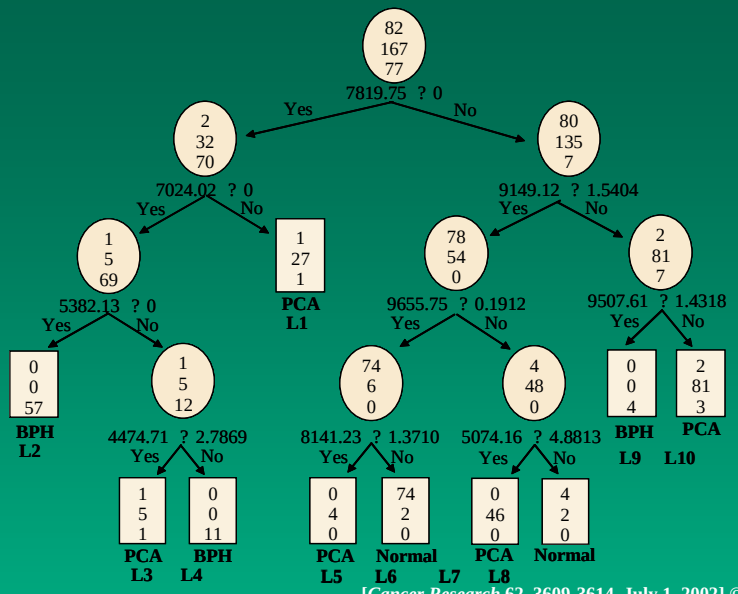
EASTERN VIRGINIA MEDICAL SCHOOL
ESTABLISHED 1973

Summary of Biomarker Discovery and Identification

SELDI-TOF



Classification and Regression Tree Analysis

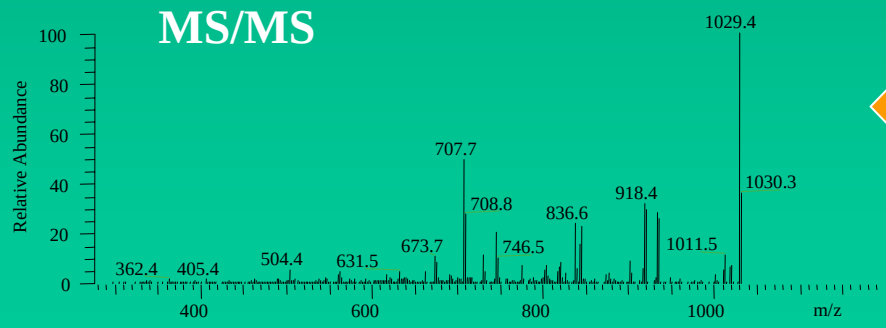


[*Cancer Research* 62, 3609-3614, July 1, 2002] © 2002
American Association for Cancer Research

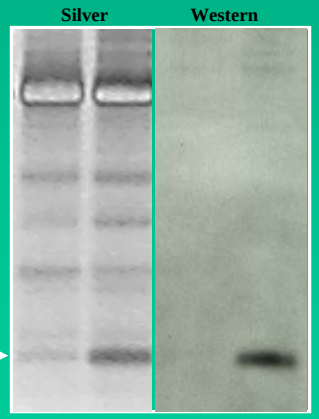
Identification

Apolipoprotein A2

MS/MS



Purification



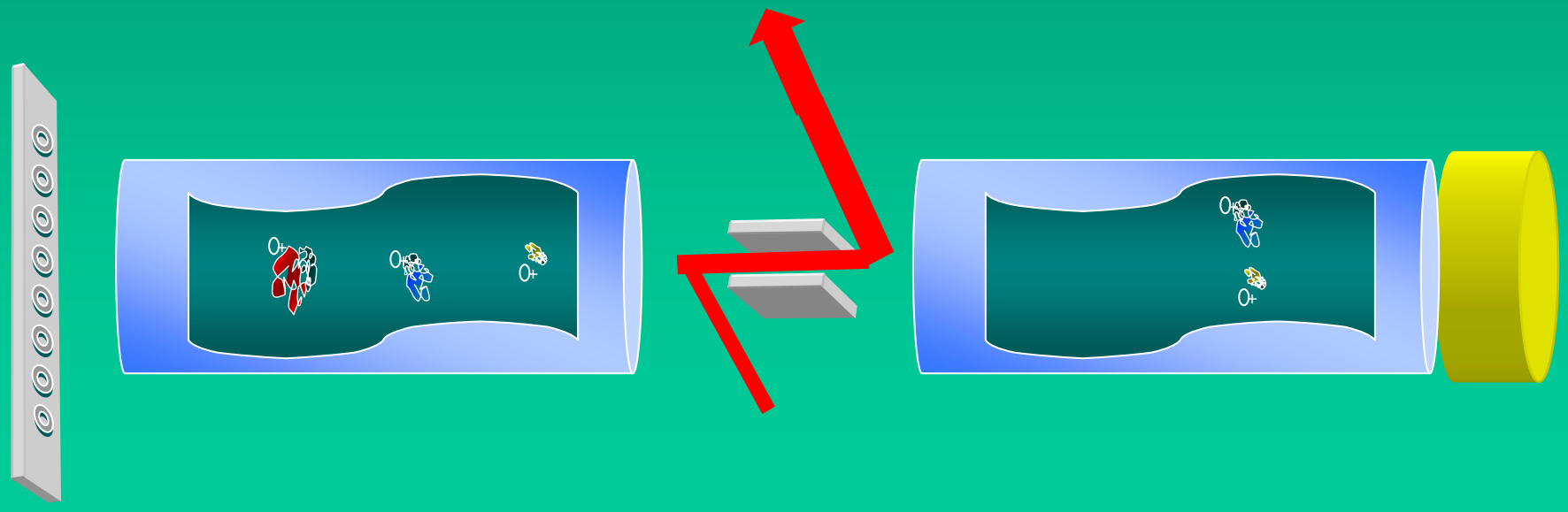
EASTERN VIRGINIA MEDICAL SCHOOL
ESTABLISHED 1973

Next talk by Bill Cooke

MS/MS for Using FEL Fragmentation/Ionization

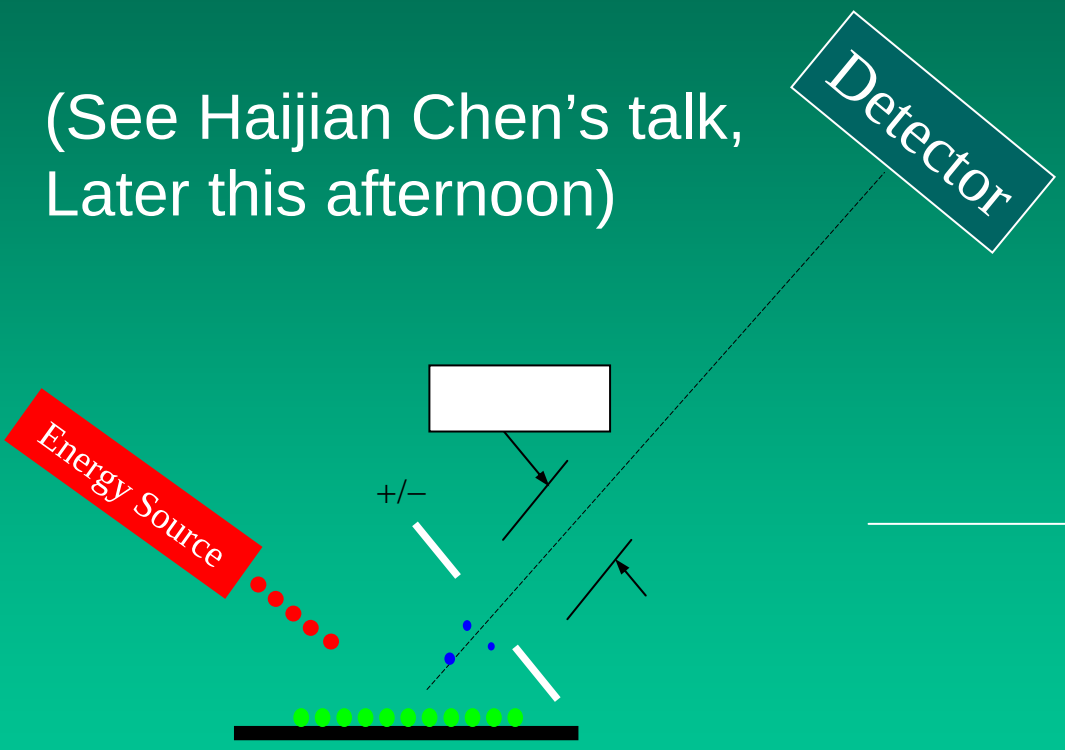
With FEL fragmentation/Ionization you can get more than one ion per parent!

- Timing determines the parent mass
- Vertical position determines the fragment mass



Example of a related project requiring automated analysis: TOF-SIMS peak picking

(See Haijian Chen's talk, Later this afternoon)

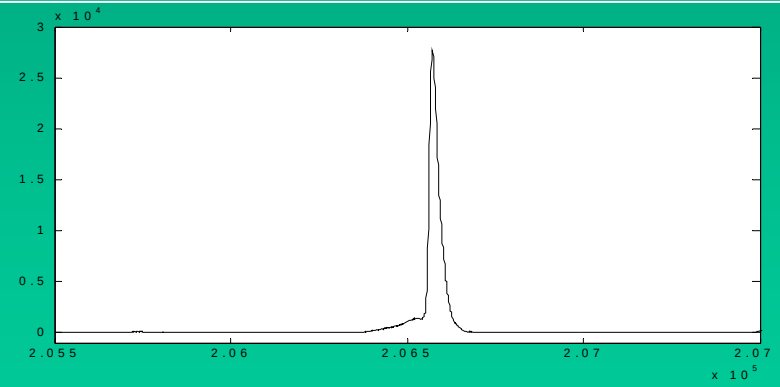


SIMS = Secondary-ion MS

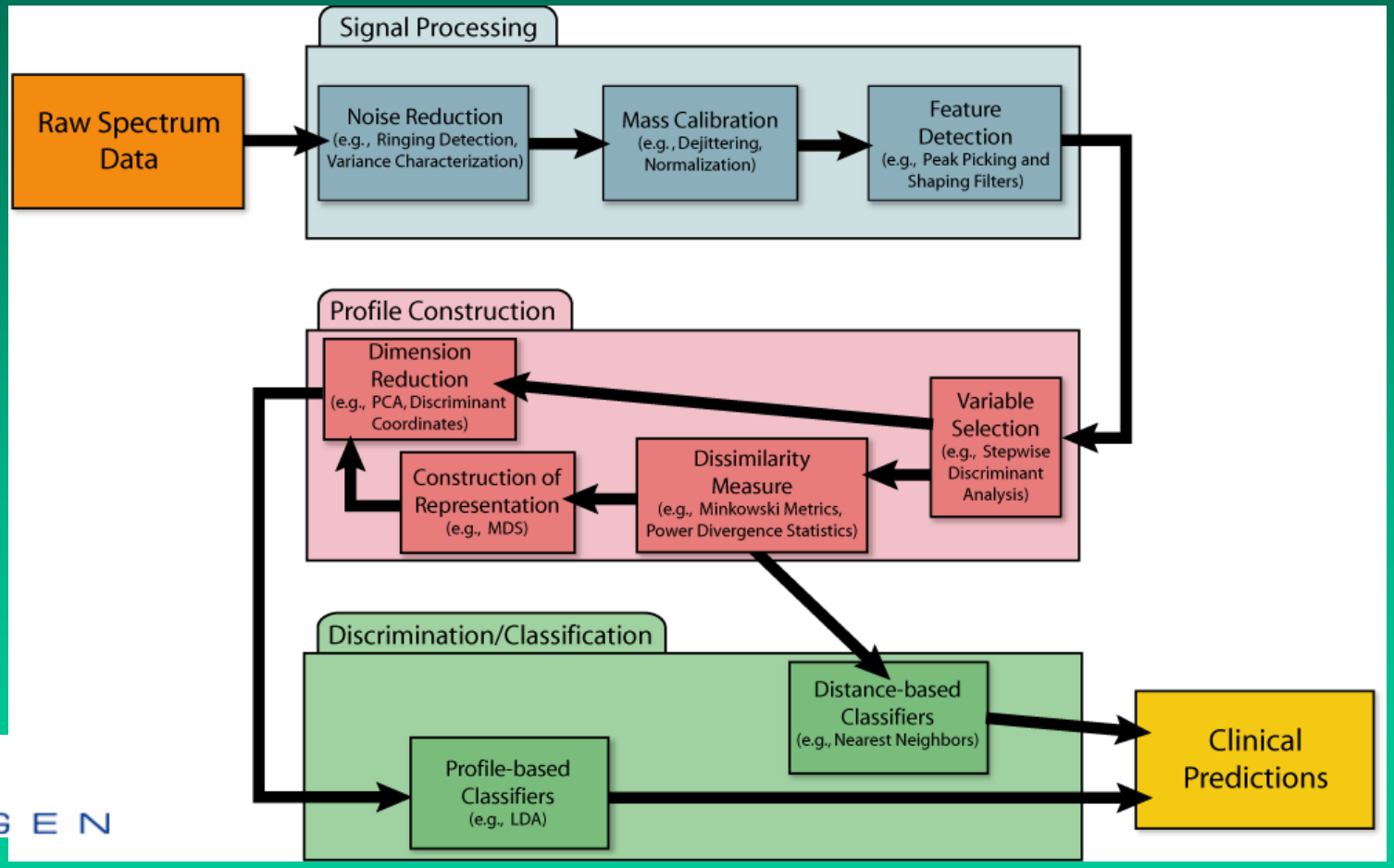
000000000000000000**1**000**1**00000000000**1**000000…
000000000000000000**1**00**1**000000000000000000…
000000000000000000**1**0000**1**000000000000000000…
000000000000000000000**1**000**1**0000000000000000…
000000000000000000000**1**00**1**00000000000000000…

⋮

000000000000000000000**1**00**1**00000000000000000…

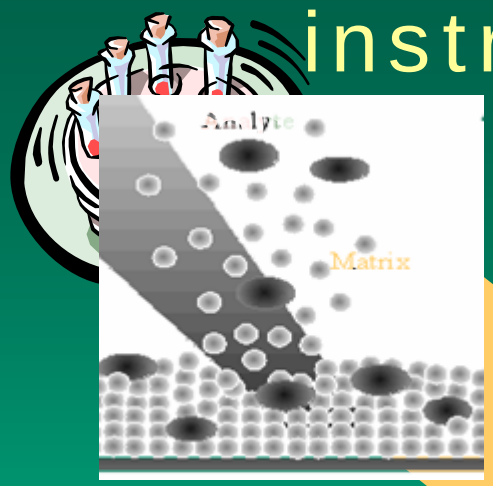


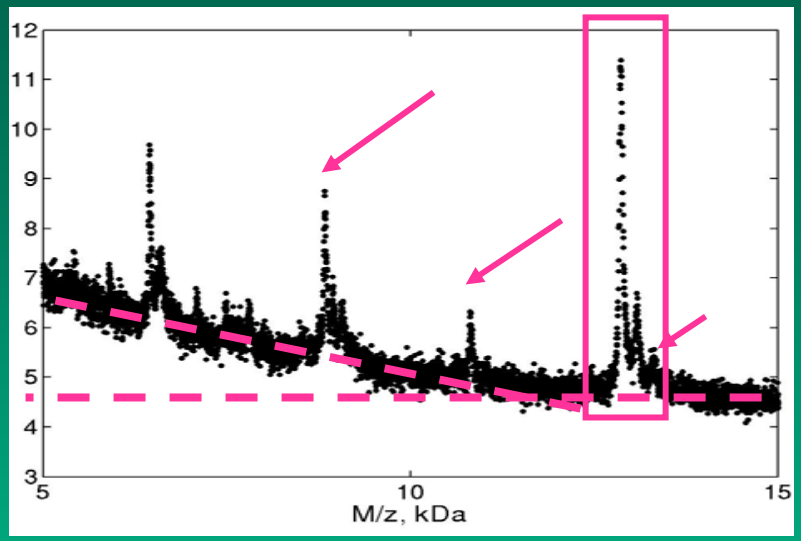
Software Framework for Classification of Mass Spectrometry Profiling Data (INCOGEN NIH Phase II SBIR)



While the logical flow of the analysis chain is very general, the details will be machine-specific.

SELDI Profiling: quick survey of some instrumental sources of error

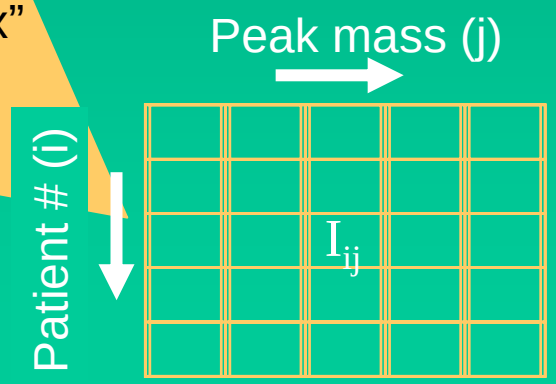




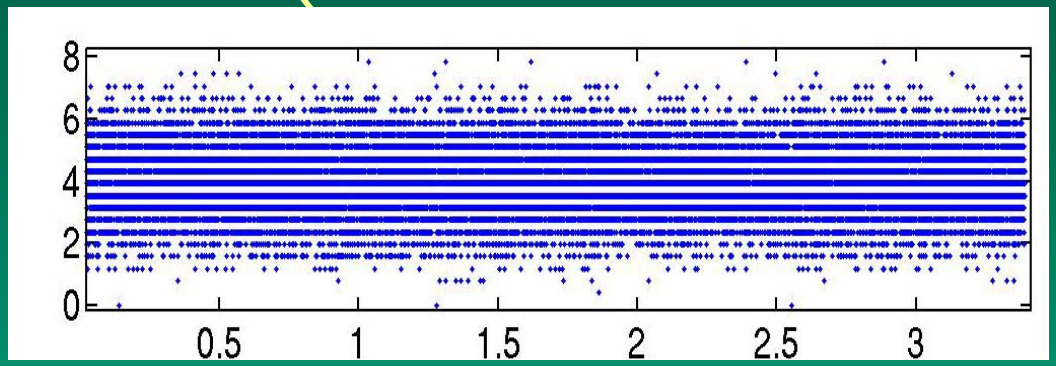
To insure data quality:
Need to characterize source and detector related uncertainties:

ΔI_{ij}, Δj ($\Delta M/z$)

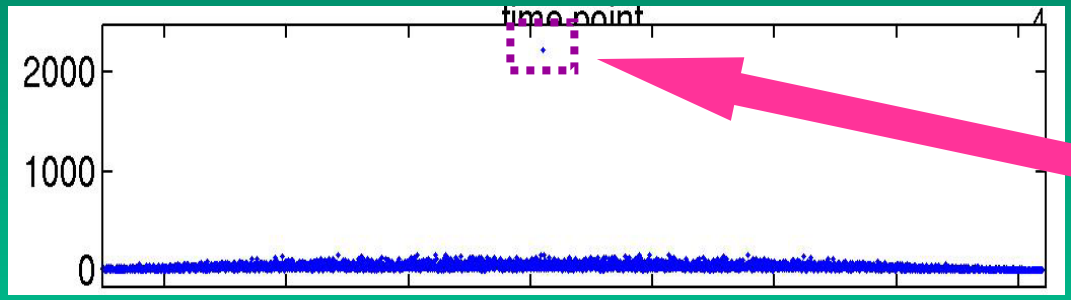
"Peak matrix"



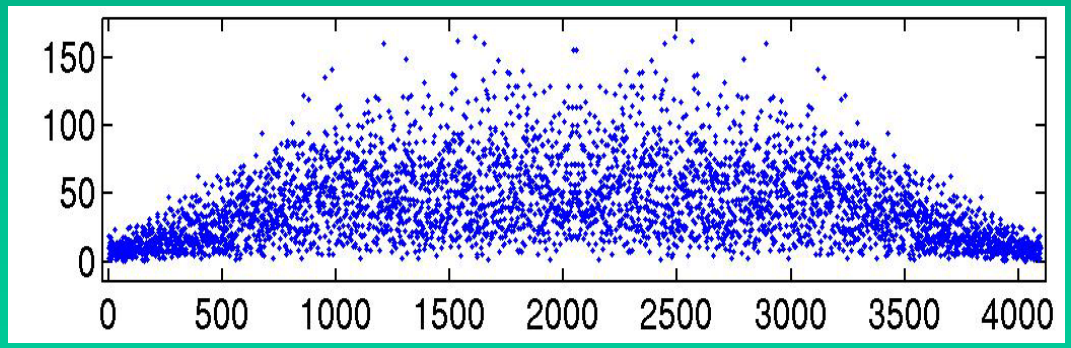
Single Shot Noise Characteristics (blank aluminum chip)



- 8-bit ADC *(max SNR = 30)*

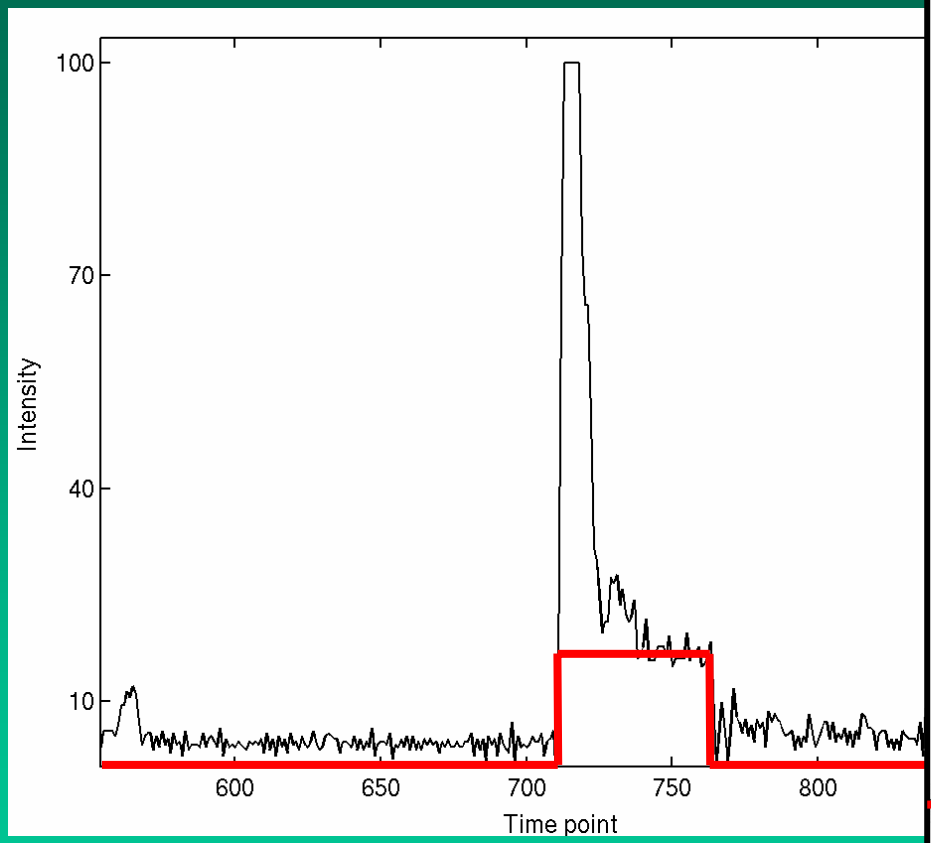


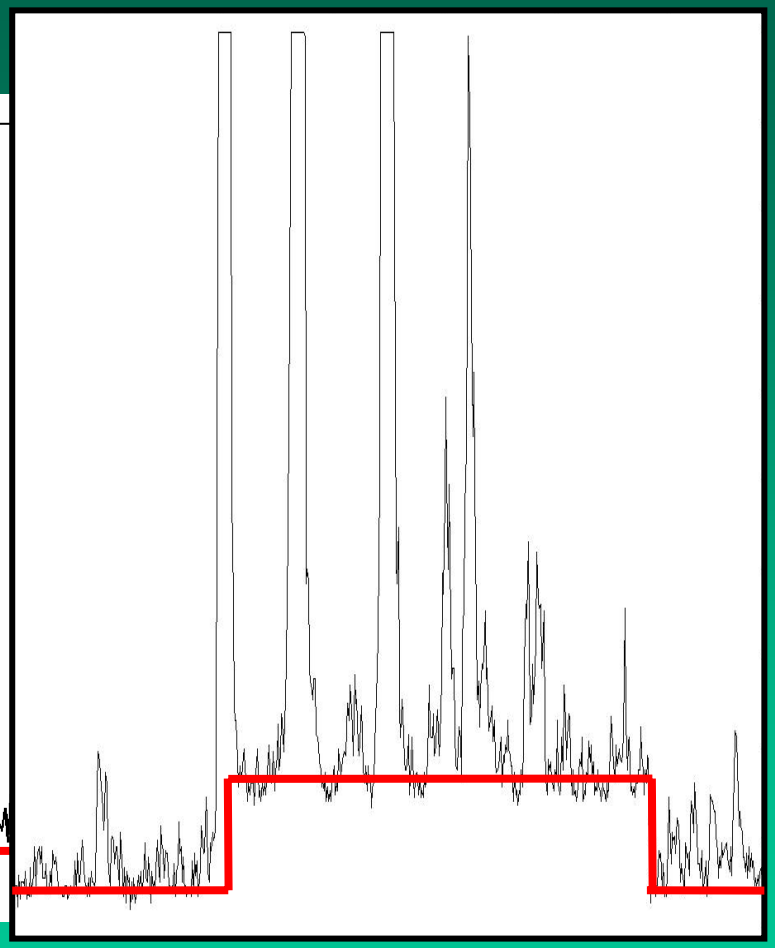
- period-2 of clock in FFT



- 2-channel differential amplifier *(colored noise)*

Overload Baseline



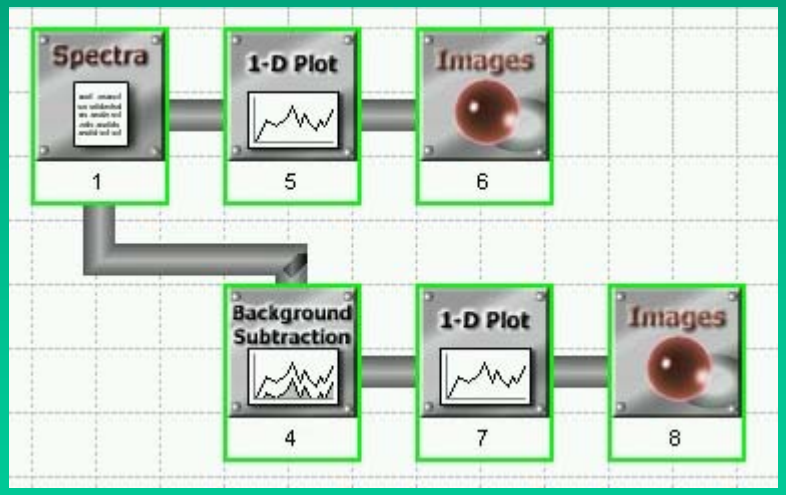


- Length is proportional to the number of overload points
- Single shot correction only
- ‘Counting’ experiments won’t have this problem

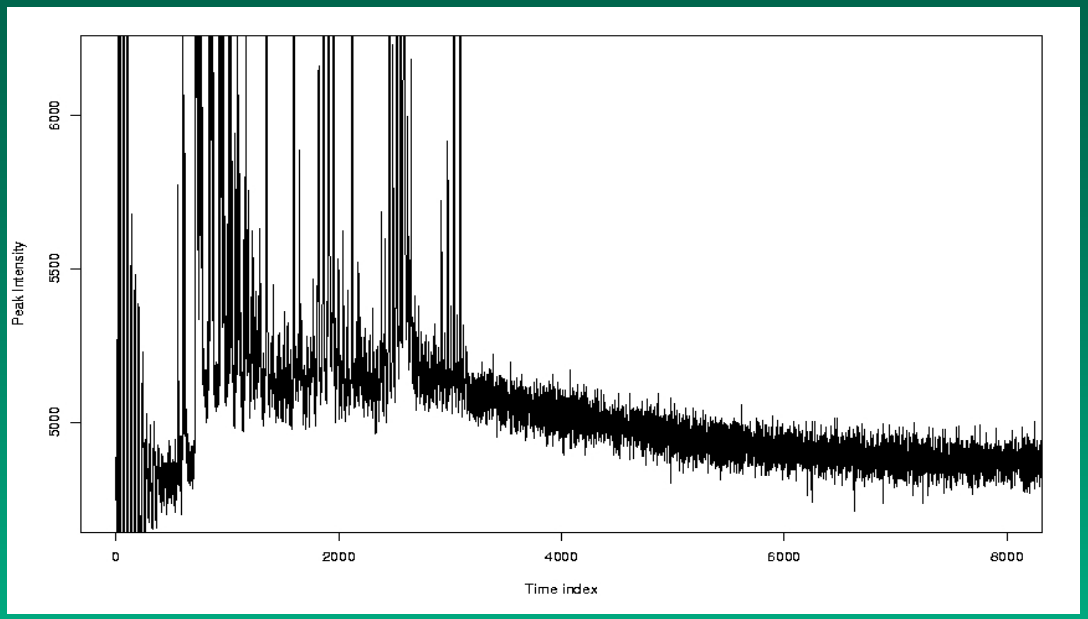
Charge Accumulation Baseline

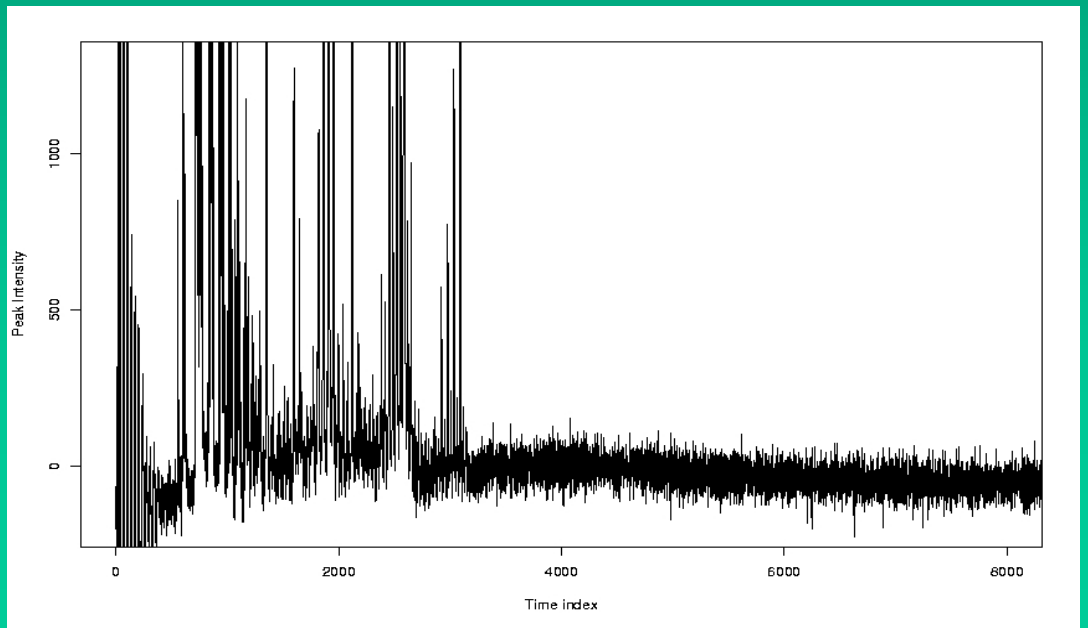
- Baseline is predictable from 3-parameter RC-modeling (Clin.Chem. 2005, **51**: 65-74)

- Can apply to single laser shot or average

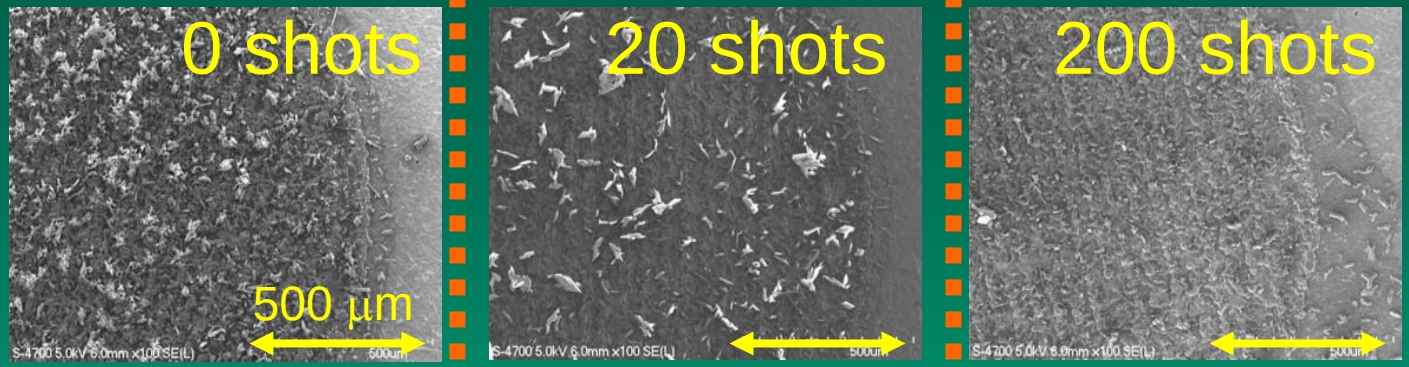


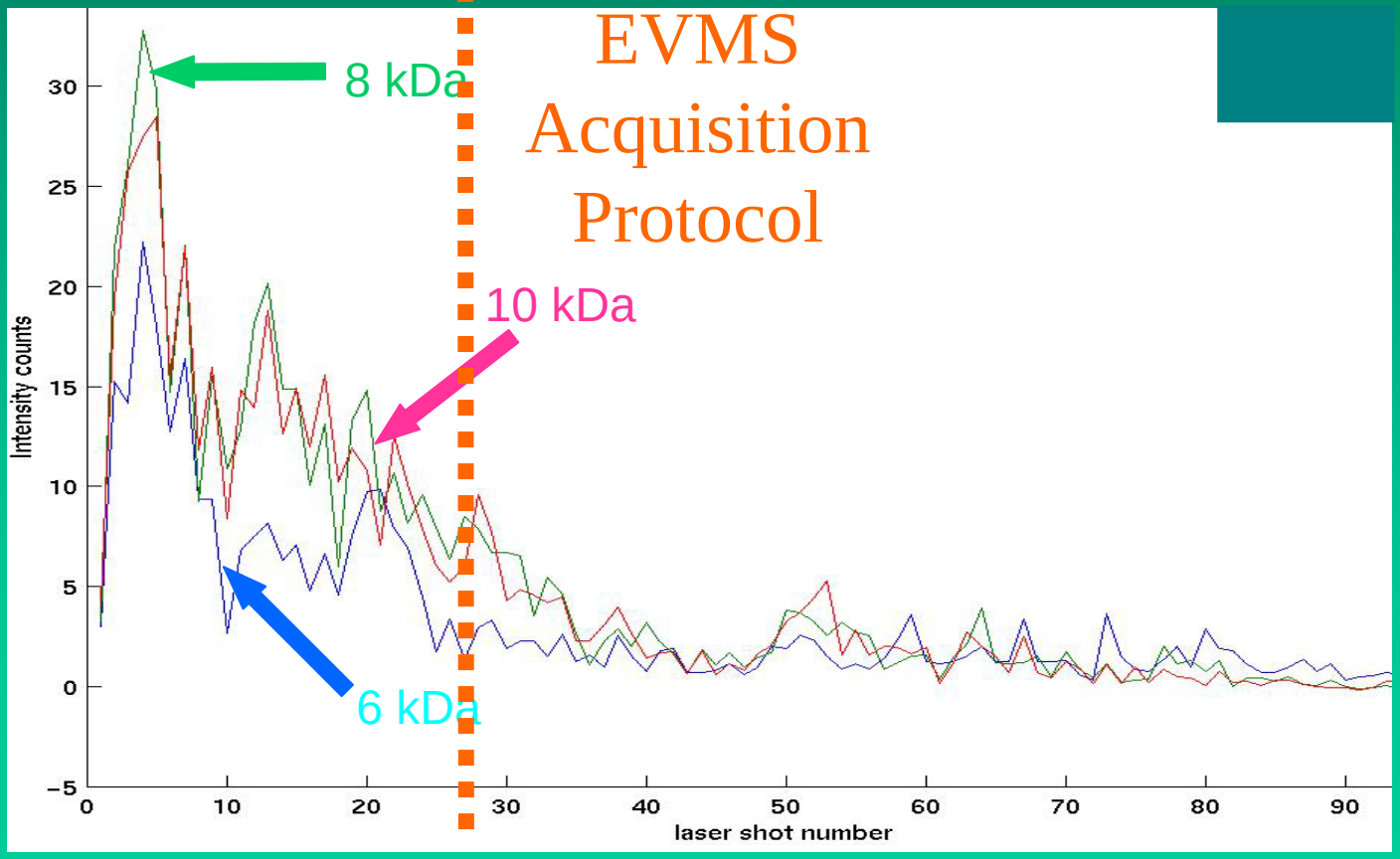
'Counting' experiments won't have this problem, either.





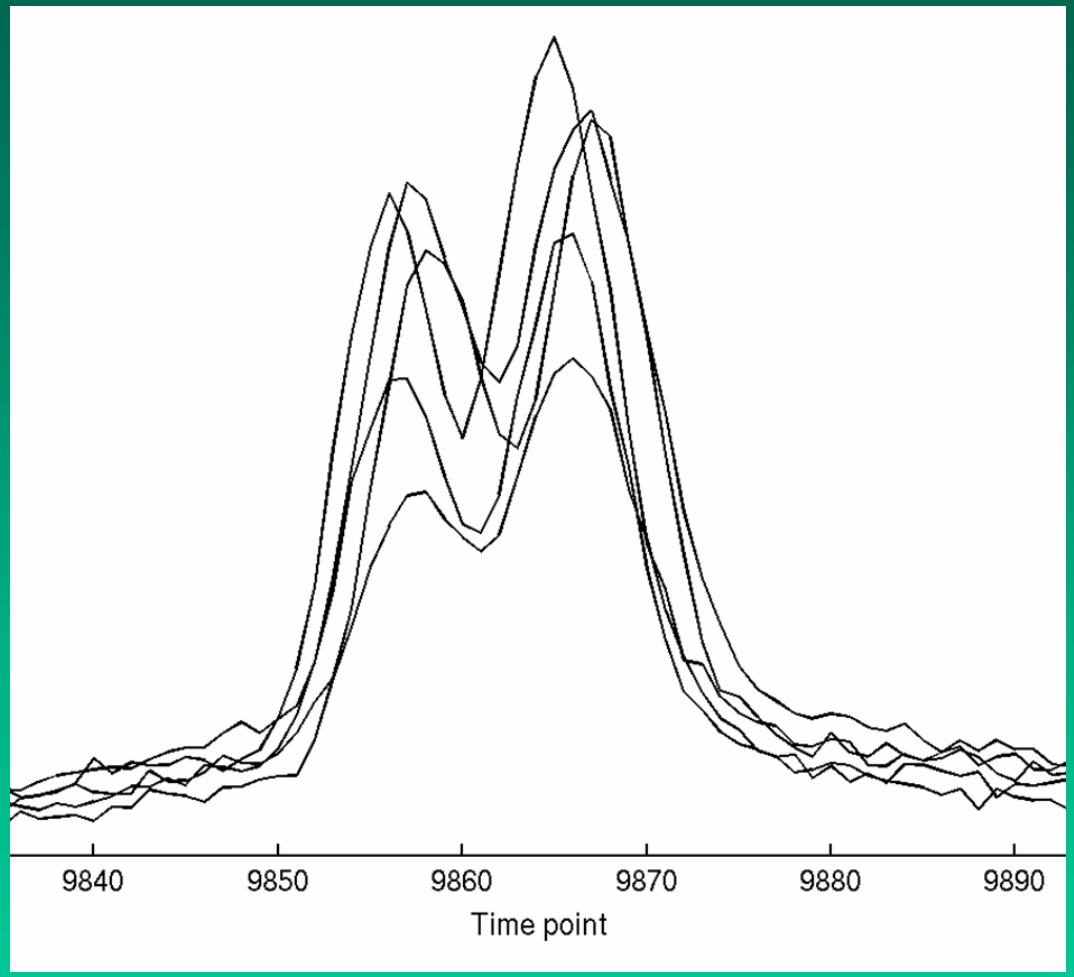
SELDI Surface Evolution

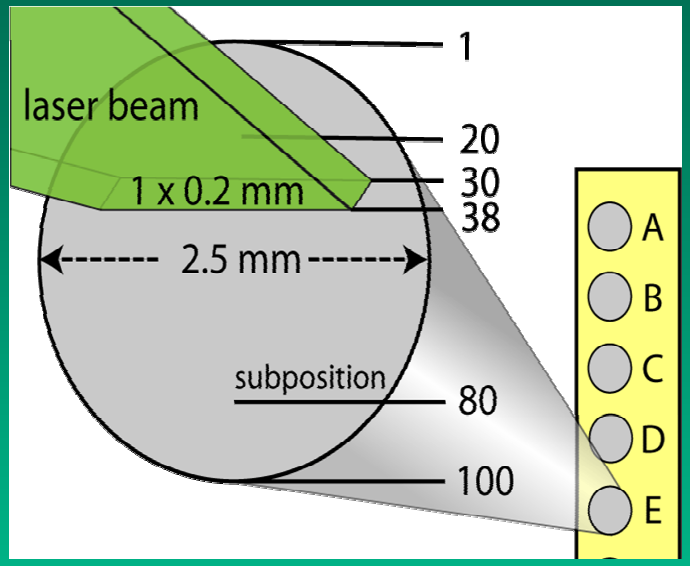




- Repeatable acquisition requires optimization of sample preparation
- Surface roughness contributes to peak ‘jitter’
- Ideally, should process individual shots

TOF-Jitter Within Spot



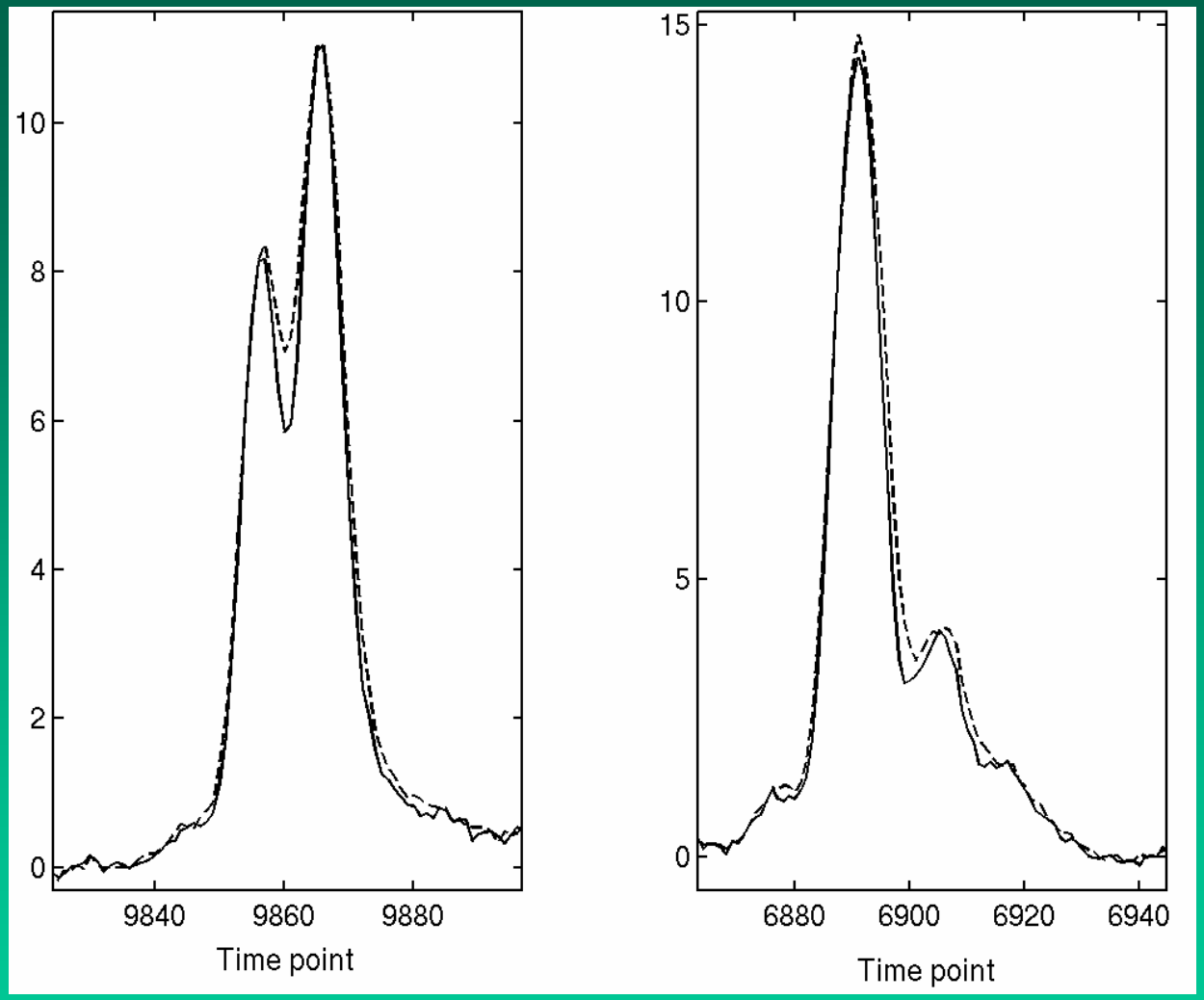


- TOF and intensity varies due to surface roughness

ΔTOF = 2 ns/μm (2 μm/dwell period)

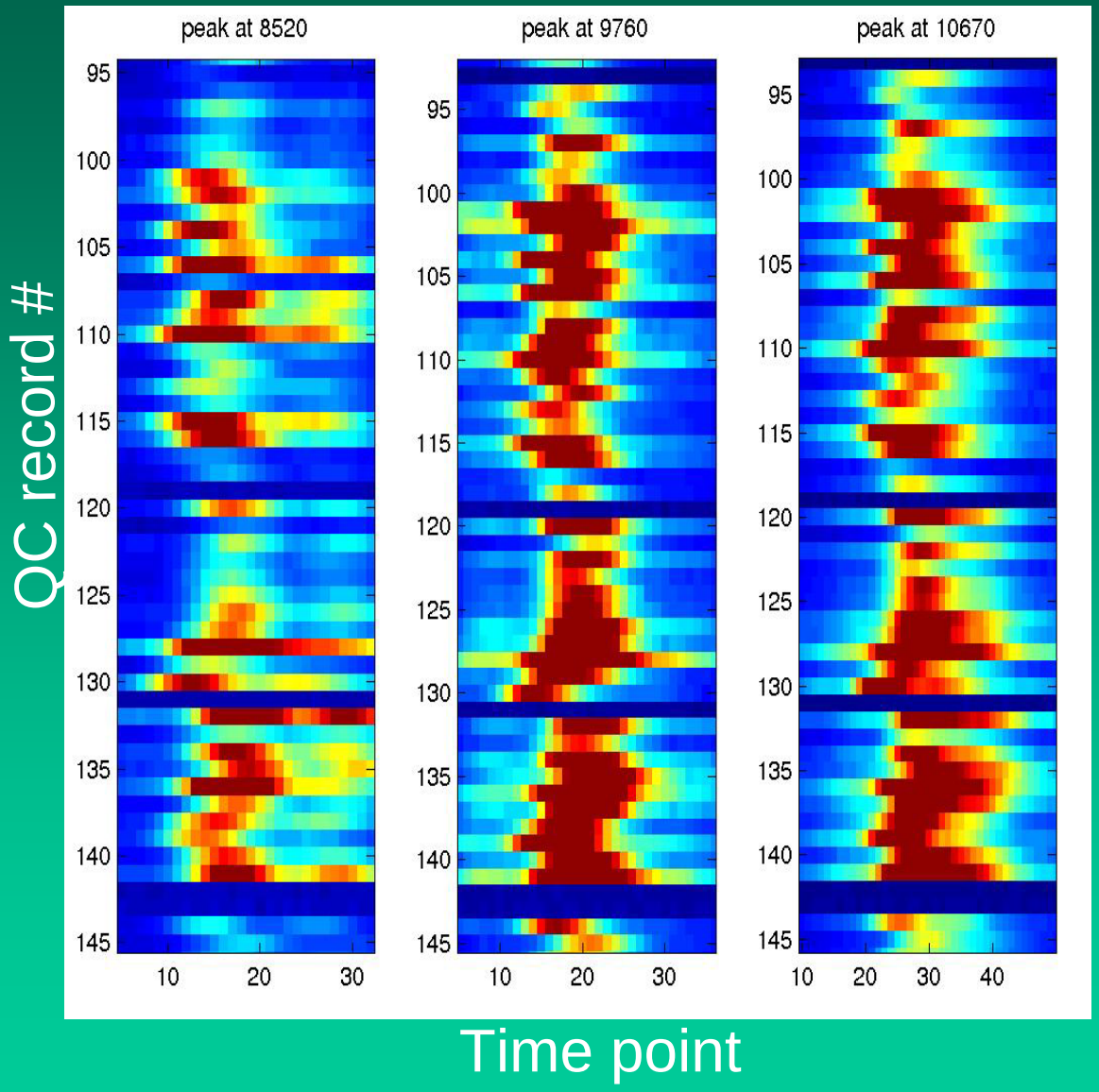
+/-10 μm → +/-5 time points

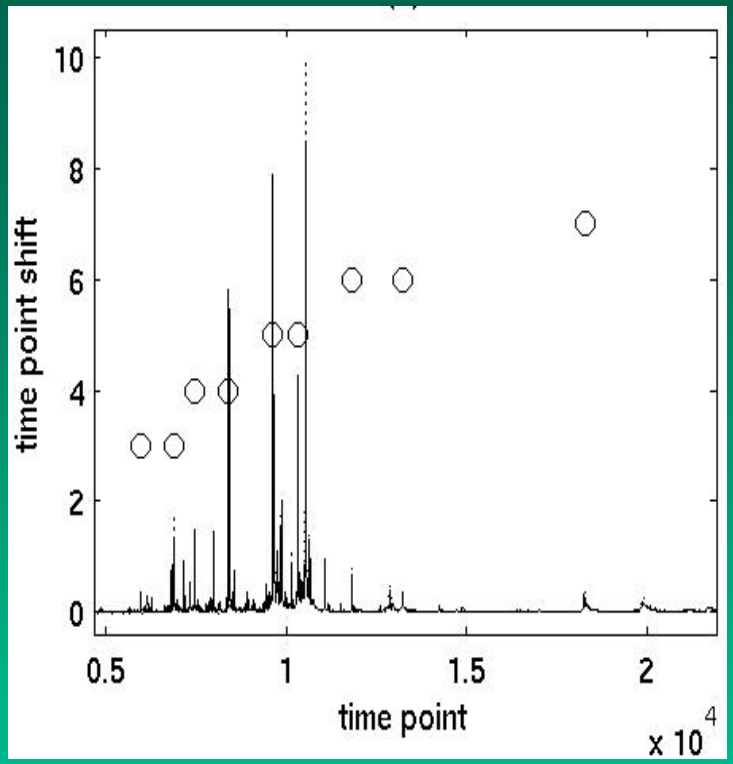
Subposition De-Jittering



- Correct by subposition de-jittering during acquisition
- Gain resolution (even from only 12 subpositions)

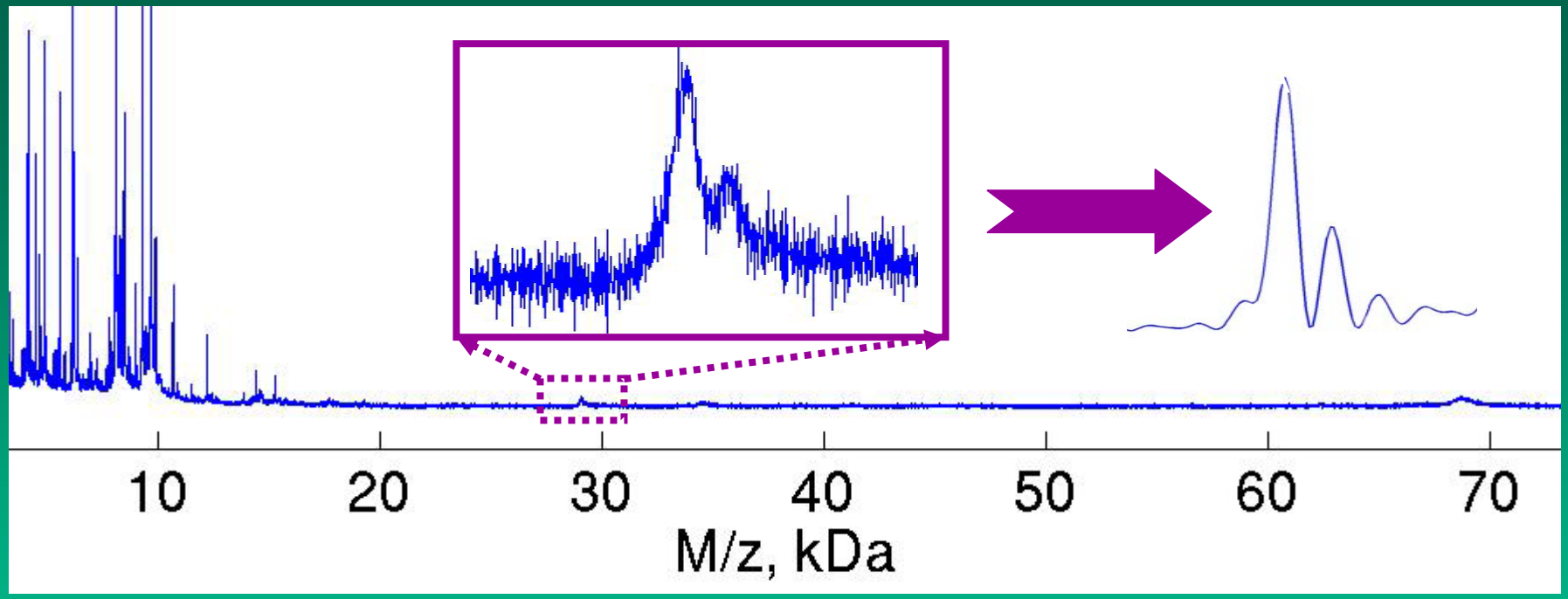
Peak Alignment Between Spots





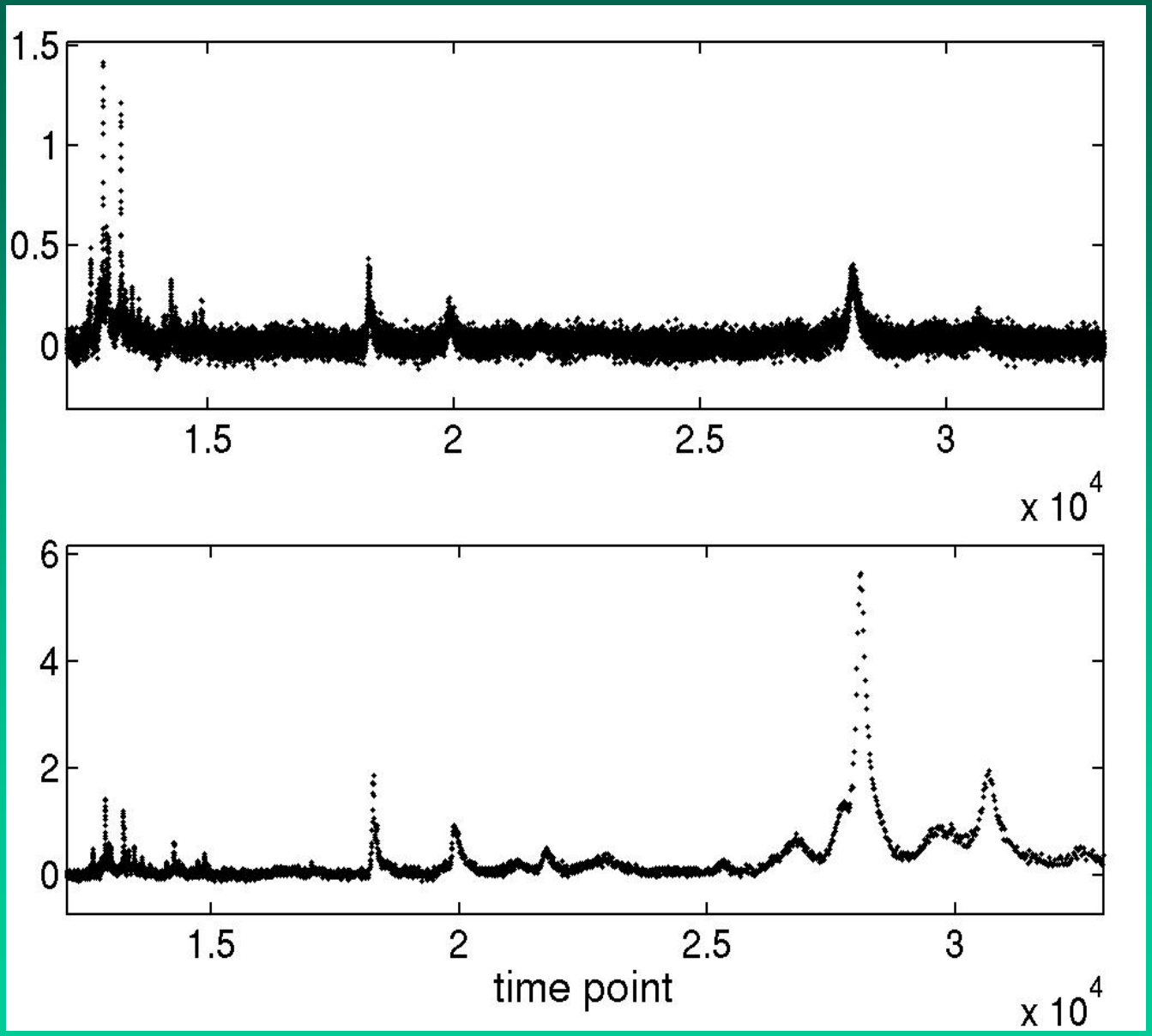
- Detect shifts *in time*
- Align *in time*

Enhancing Data Quality



- Background detector noise is stationary
- Peak lineshape is t-dependent (they get broader)
- Detector sensitivity falls at high mass
- High-mass peaks are broad and low amplitude
- ‘Signal’ is the integrated peak area: suggests resampling

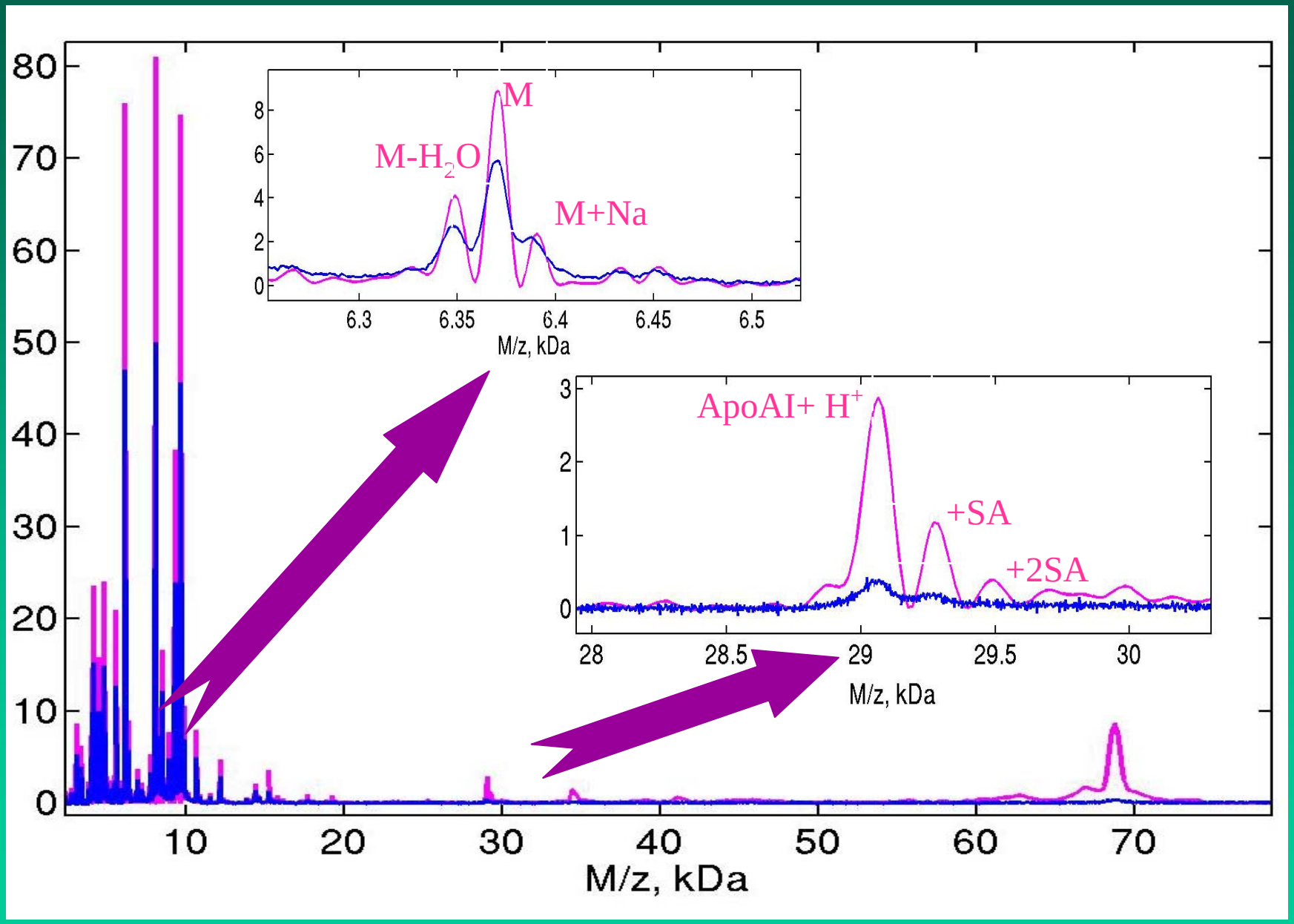
Effect of Resampling on TOF Signal



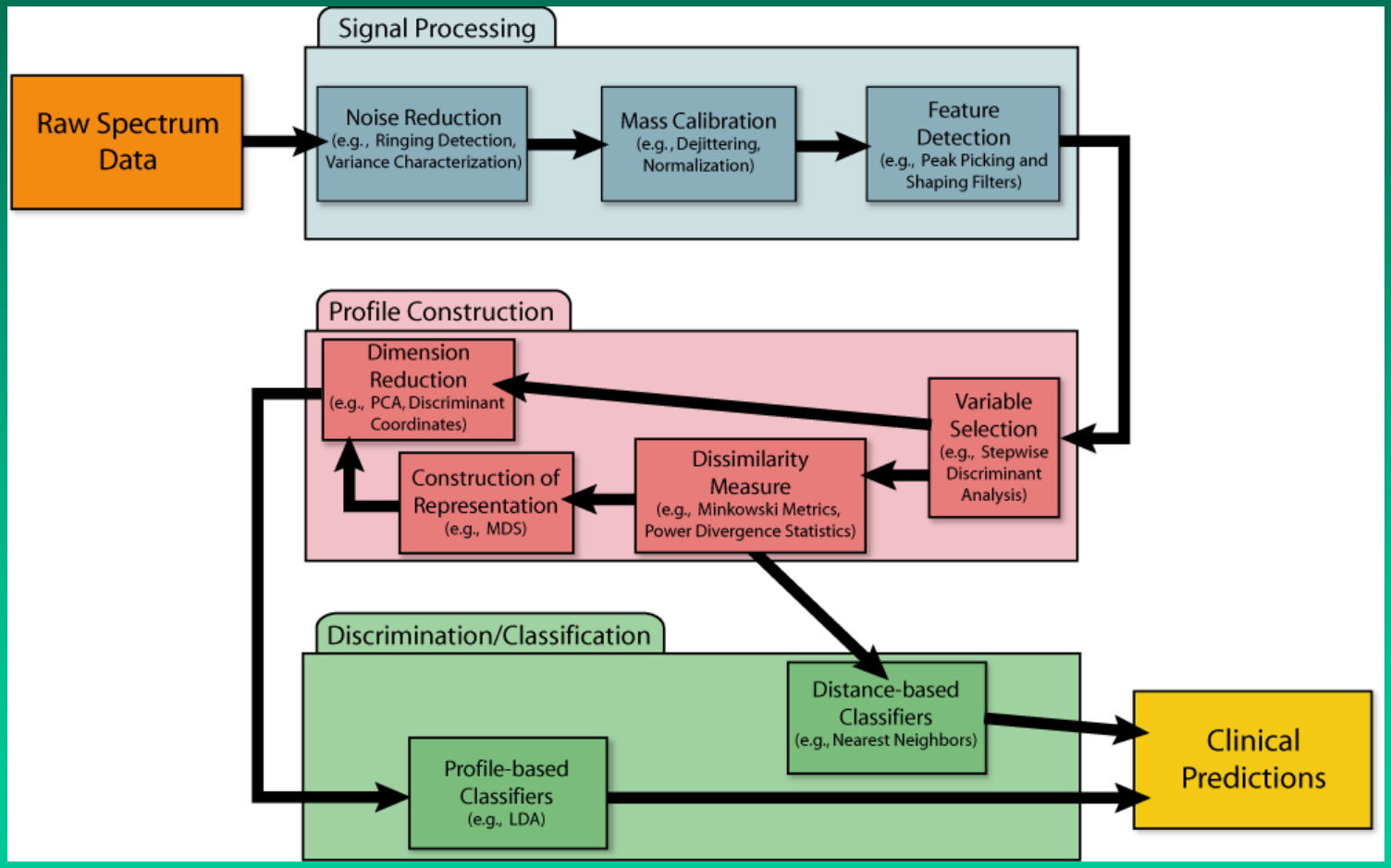
For all peaks
$\sigma_t = \text{const}(t)$

Enhanced
sensitivity
for heavy ions

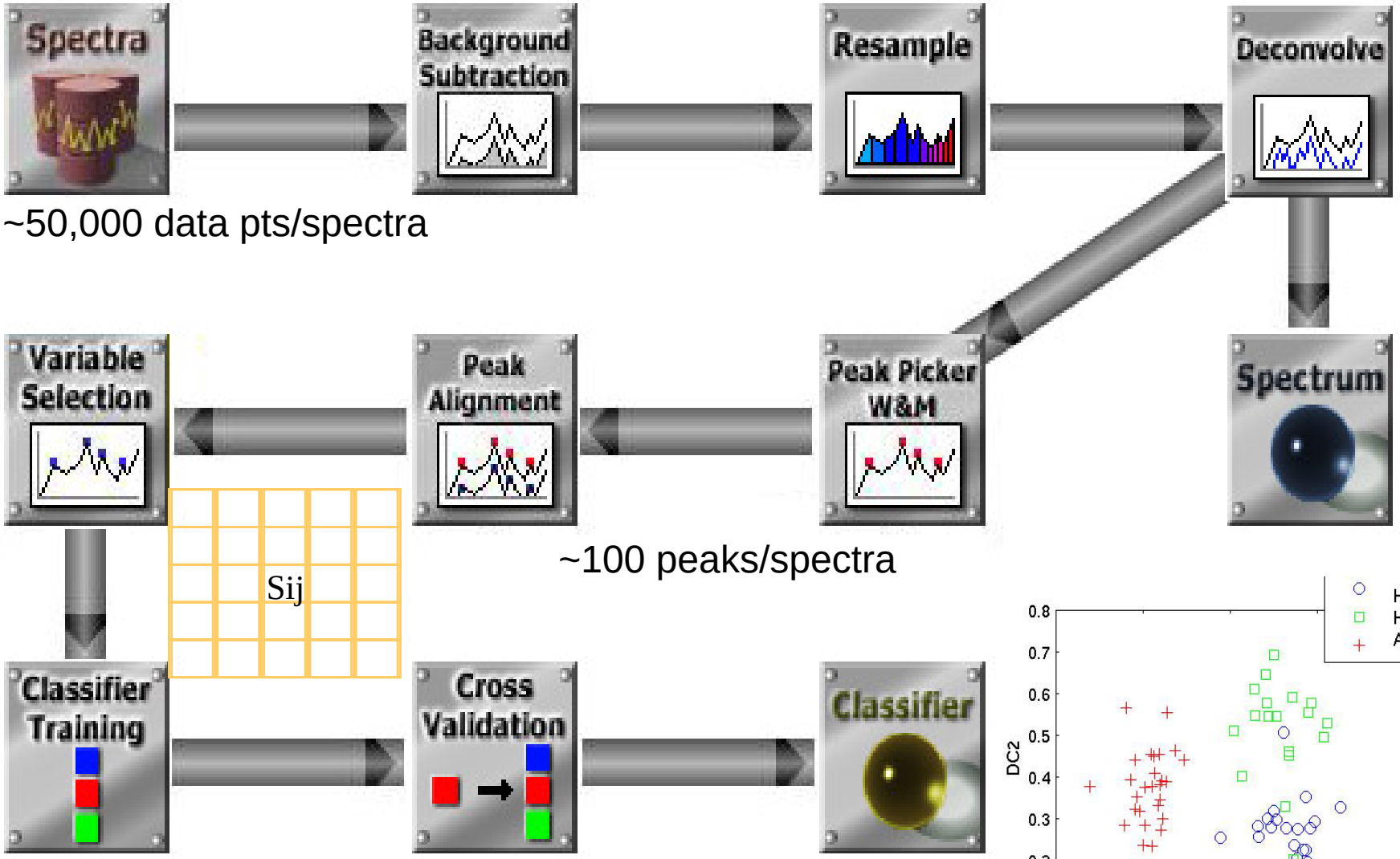
Broad Range Deconvolution

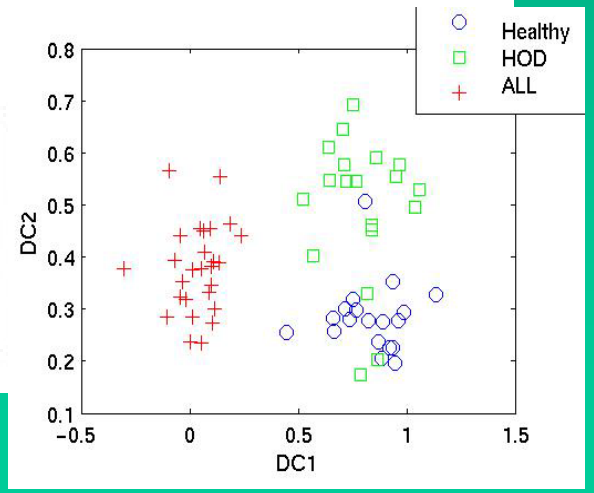


Recall logical flow of analysis



Typical VIBE-MS Analysis Pipeline





INCOGEN

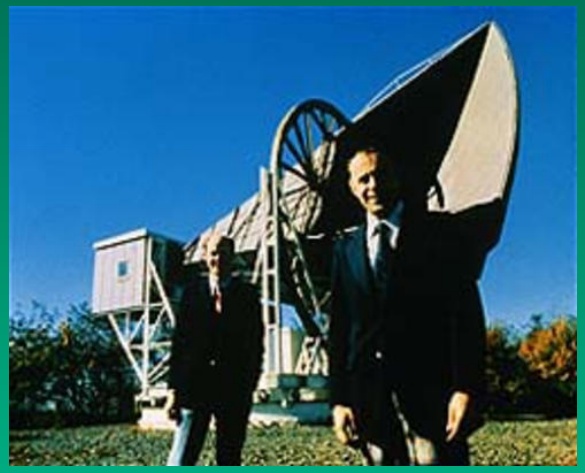
TANSTAFK!*

- The measuring instruments are also complex dynamical systems.
- The technology is rapidly evolving, hence early instrument prototypes will *not* be well characterized.
- Instrument characterization must be part of the *analysis* chain used for diagnosis.
- The first step in ‘dimension reduction’ must be the elimination of instrumental variation. Otherwise you run the risk of classifying on instrumental artifacts.
- The *details* are important, and naïve approaches will only work if we get lucky.
- “Denoising” is ill-defined until you understand the underlying processes.
- It’s *all* signal. ‘Noise’ is simply the part of the signal you aren’t interested in.

*There Ain’t No Such Thing As Free “Knowledge”!

A case where the ‘noise’ was really the ‘signal’: the 3K background from the Big Bang

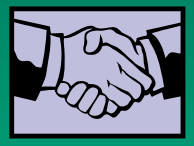
It’s safest to *understand* the noise process. This will lead to better de-noising strategies, and more confidence that something important hasn’t been thrown out (with the pidgeons).



Penzias and Wilson, Bell Labs

Acknowledgements

This work was supported by Virginia's Commonwealth Research Technology Fund #IN2003-04, and Phase II SBIR grant from the National Cancer Institute CA101479



We thank our collaborators and coauthors from EVMS, and INCOGEN for assistance with acquisition and analysis of calibration data for SELDI PBSII instrument, and Amy Wilkerson of the Applied Research Center for assistance with the TOF-SIMS instrument.

We are grateful to Dr. Stacy Moore and Dr. Scott Weinberger from Ciphergen, Inc. for their help in clarifying instrumental specifications and parameters

Coming soon

VIBE-MS Workshop (dimaly@wm.edu)