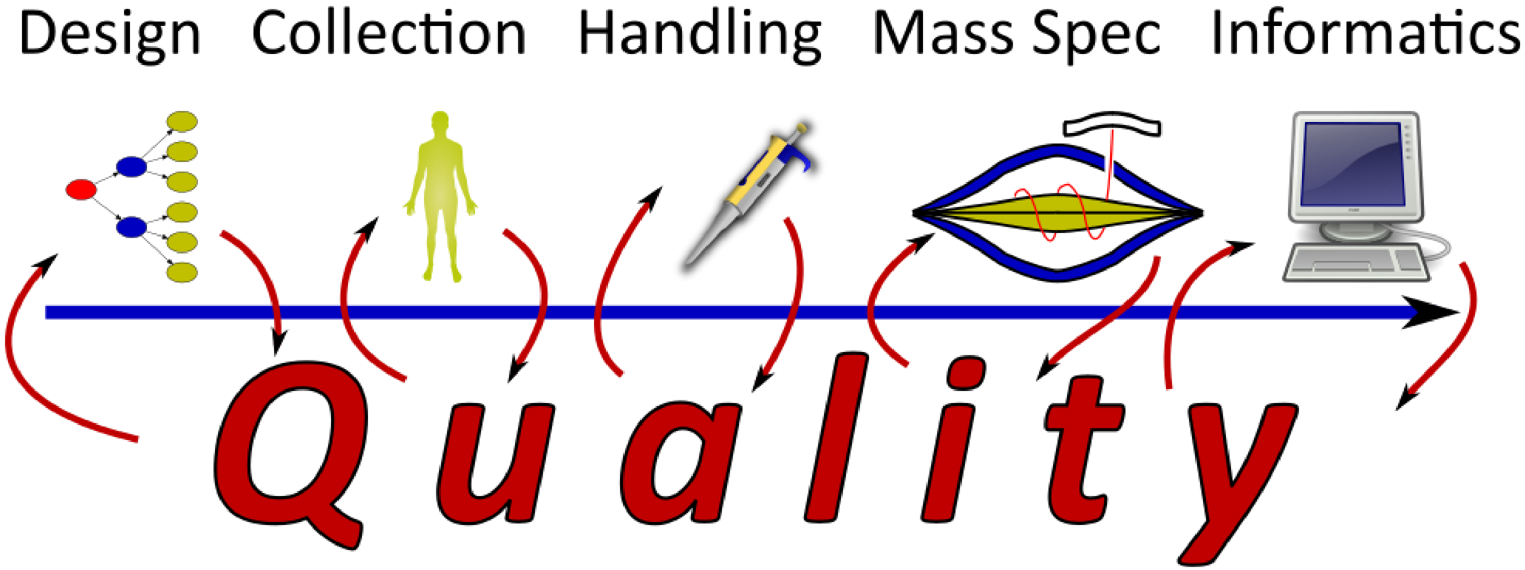
PCA-powered quality control for large-scale proteomics

DAVID L. TABB, PH.D.

OCTOBER 20, 2020

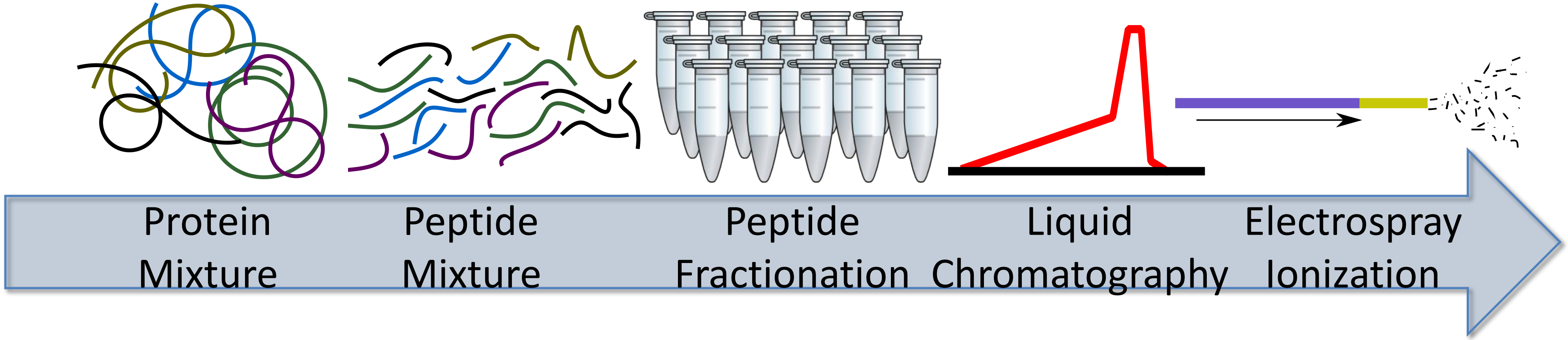


10.1016/j.clinbiochem.2012.12.003

Overview

- Proteomics: technical complexity accumulates variability
- Quality metrics quantify variation, but how to interpret?
- Principal Components Analysis: a brief introduction
- Outliers, batch effects, and analysis of variance, oh my!

Discovery Proteomics



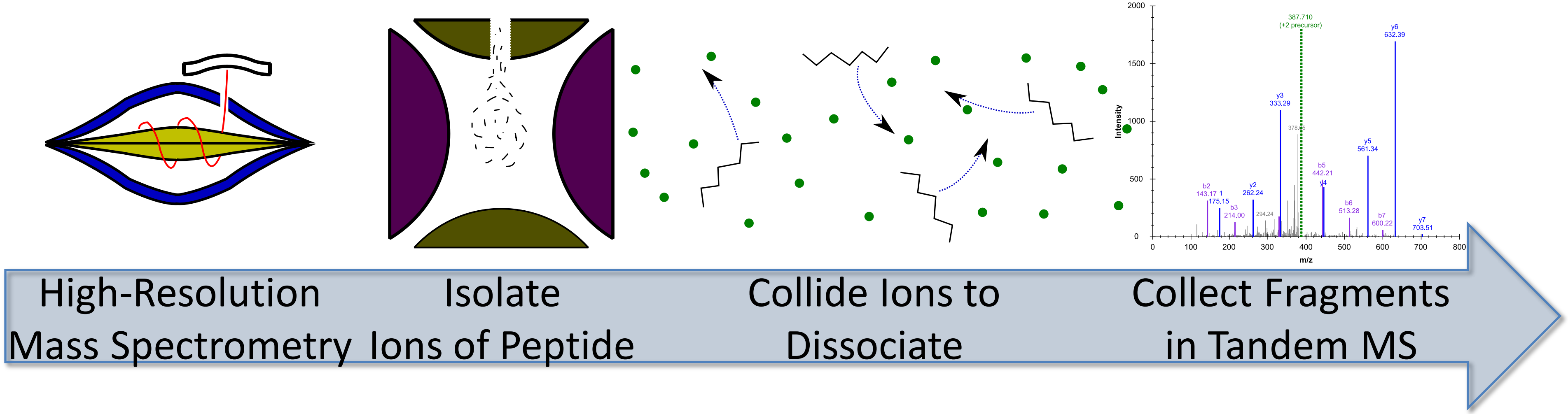
Protein Mixture

Peptide Mixture

Peptide Fractionation

Liquid Chromatography

Electrospray Ionization



High-Resolution Mass Spectrometry

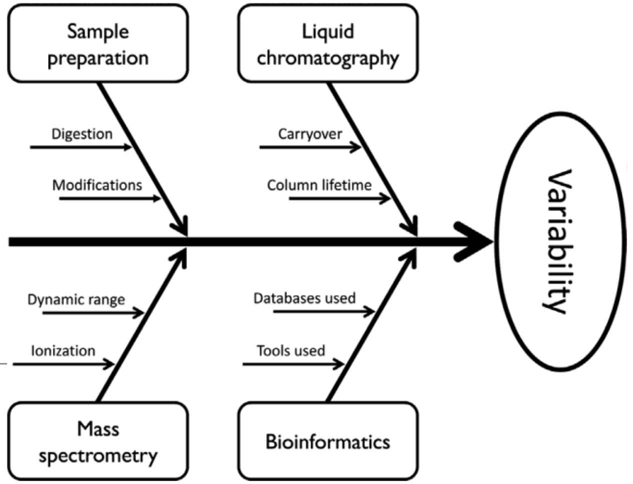
Isolate Ions of Peptide

Collide Ions to Dissociate

Collect Fragments in Tandem MS

M. Bantscheff et al. *Anal. Bioanal. Chem.* (2007) 389: 1017

Factors contributing variability

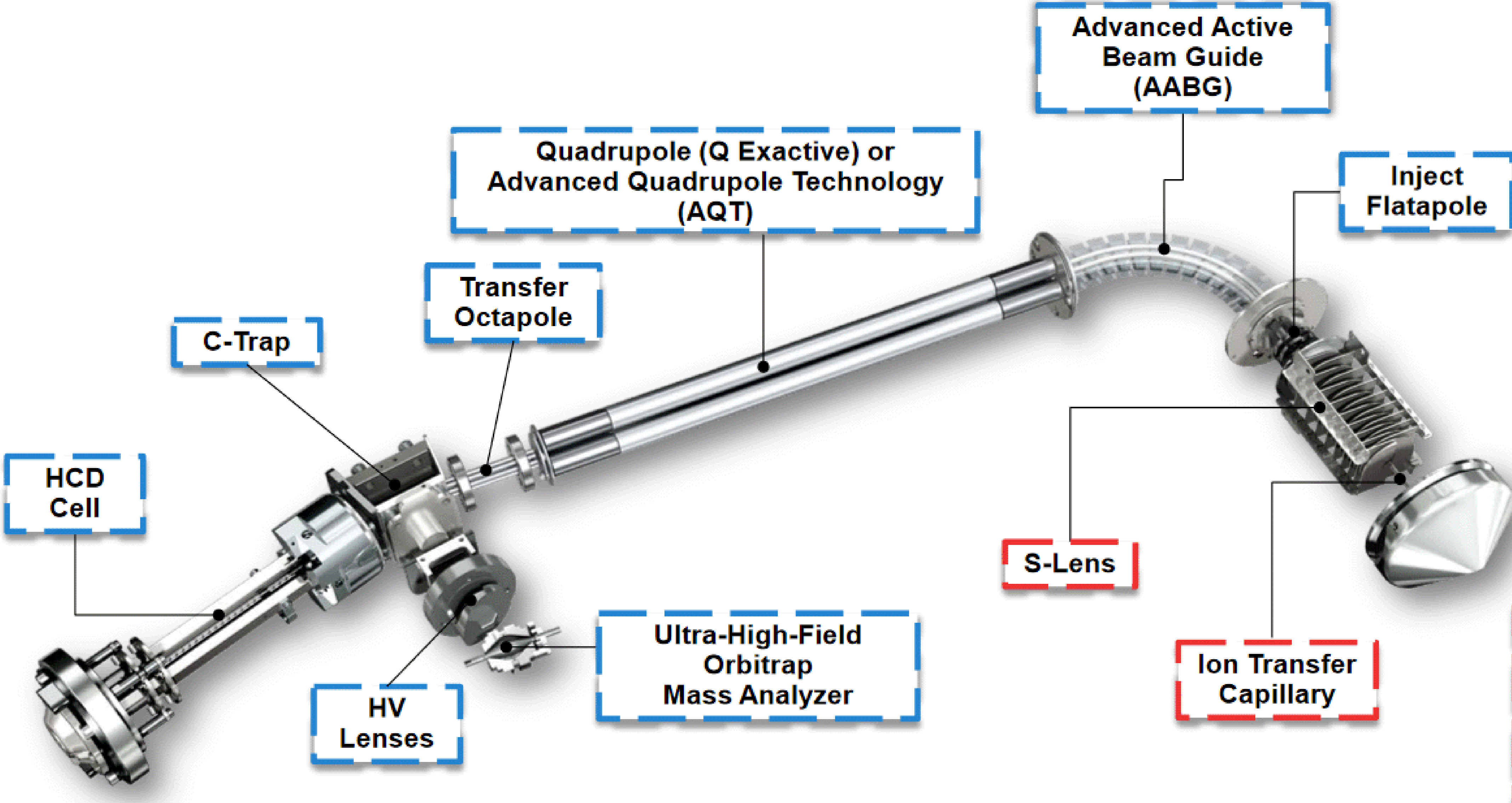


10.1002/mas.21544

- Enrichment (Co-IP or PTM)
- Depletion (from biofluids)
- Denaturation and digestion
- Fractionation
- Liquid chromatography
- Electrospray

- Tune response
- Mass calibration
- Source cleanliness
- Ion optics
- Vacuum pump efficiency
- Detector age

https://www.spectroscopyonline.com/view/ion-burn-and-dirt-mass-spectrometry



Advanced Active Beam Guide (AABG)

Quadrupole (Q Exactive) or Advanced Quadrupole Technology (AQT)

Inject Flatapole

Transfer Octapole

C-Trap

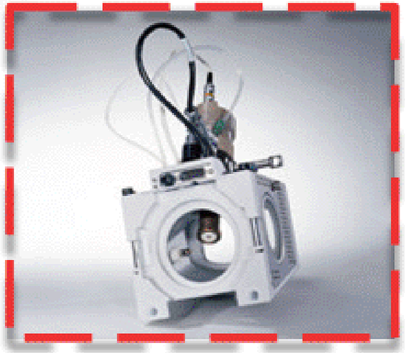
HCD Cell

S-Lens

Ion Transfer Capillary

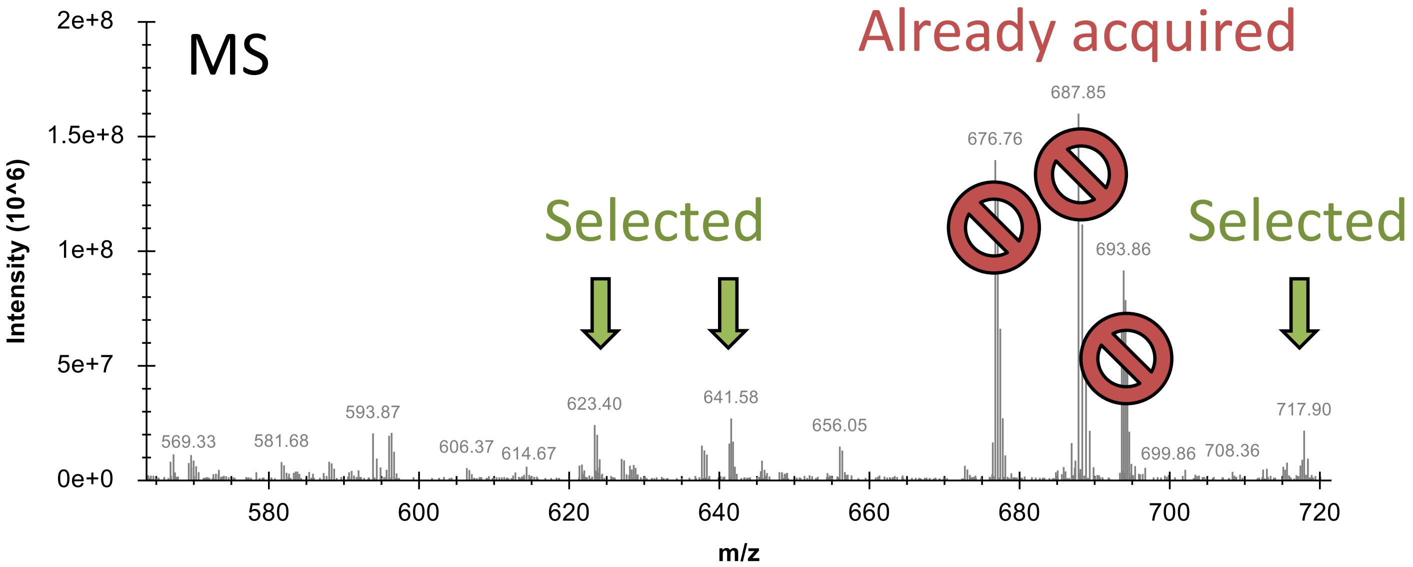
HV Lenses

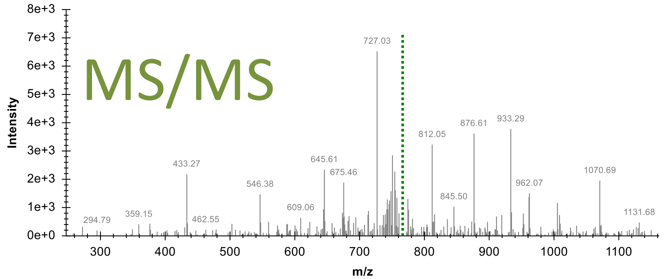
Ultra-High-Field Orbitrap Mass Analyzer

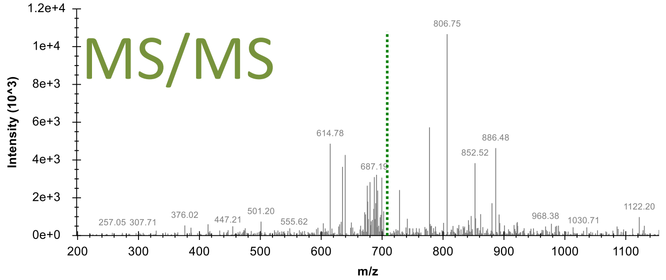


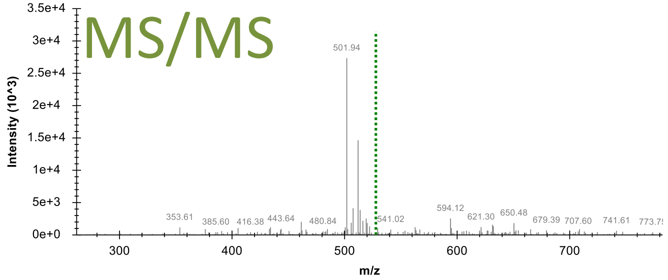
ThermoFisher: WS-MS-Q-Exactive-Calibration-Maintenance-iQuan2016-EN.pdf

MS/MS acquisition prioritizes intense precursor ions in MS1.







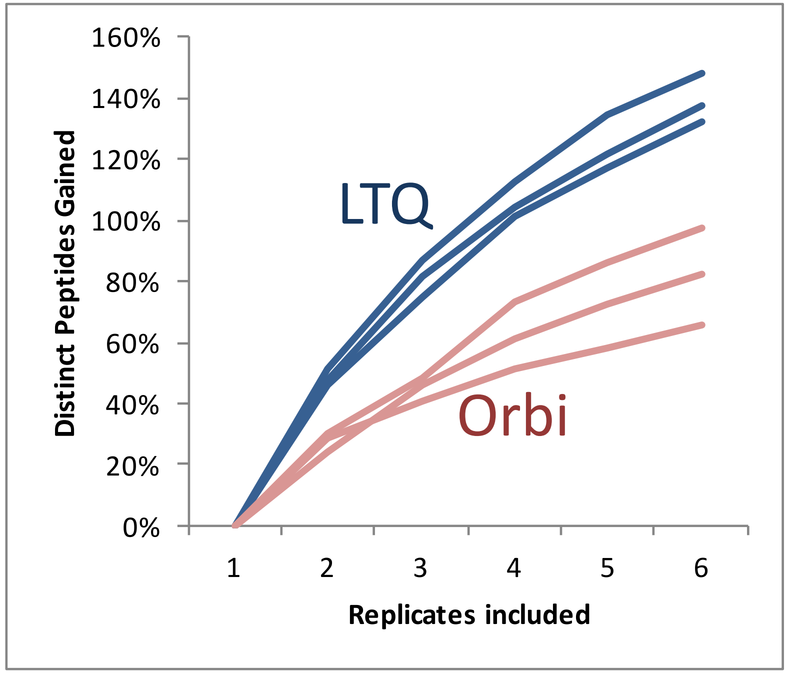


In "data-dependent acquisition," the computer chooses high-intensity m/z values from MS that have not recently been subjected to MS/MS.

DDA method leads to *stochastic* differences among experiments with the same sample.

CPTAC: Clinical Proteomic Technology Assessment for Cancer

DO WE SEE THE SAME PEPTIDES EACH TIME WE “SHOOT” YEAST?



HOW STABLE ARE SPECTRAL COUNTS WHEN A PROTEIN *DOESN’T* CHANGE?

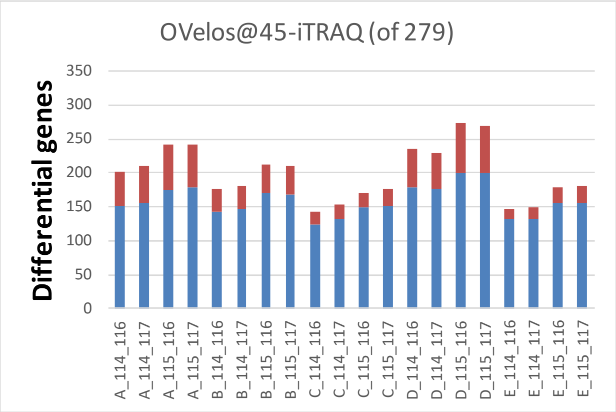
Typically stable spectral counts by instrument

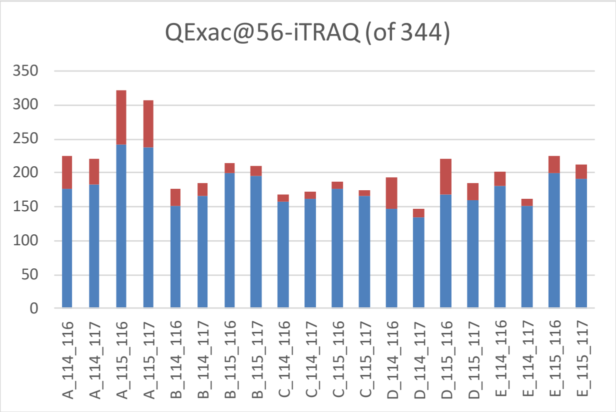
Instrument	Accession	MVH[1]	Rep1	Rep2	Rep3	Rep4	Rep5	Rep6
LTQ@73	YDR012W	1.05	28	27	29	23	25	18
LTQ2@95	YGL062W	0.90	2	2	2	2	5	3
LTQc@65	YLR441C	1.06	12	14	10	8	10	9
Orbi@86	YDL124W	0.64	5	3	3	4	2	3
OrbiP@65	YLR150W	0.41	4	4	4	4	3	2
OrbiW@56	YBR109C	0.54	3	4	3	4	2	5

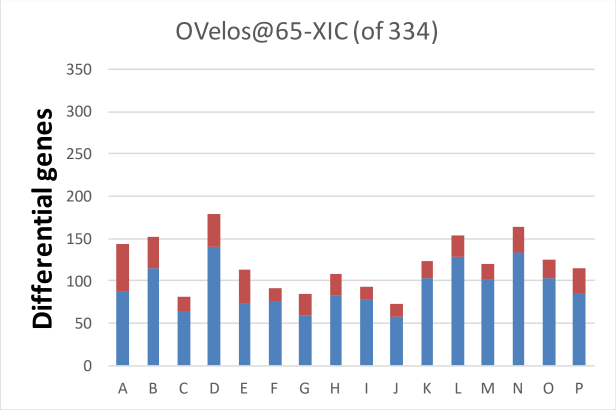
[1] Natural log of multivariate hypergeometric probability ratio: observed distribution versus expected distribution

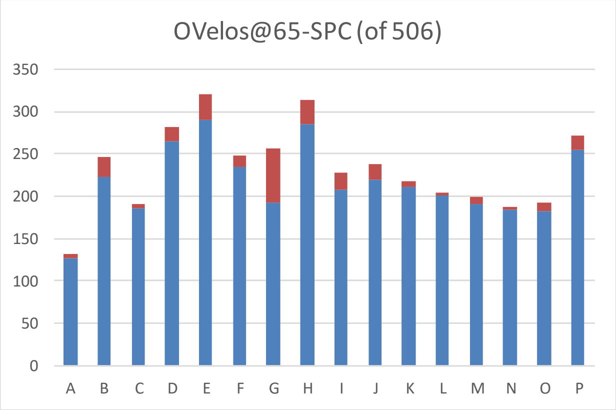
DL Tabb et al. *J. Proteome Res*. (2009) 9: 761-776.

Are the differences we claim reproducible?


OVelos@45-iTRAQ (of 279)


QExac@56-iTRAQ (of 344)


OVelos@65-XIC (of 334)


OVelos@65-SPC (of 506)

- Blue differential genes are found in common by another experiment in this instrument, while orange ones are unique to a single replicate.
- For iTRAQ, the confirmation was required to come from a different multiplex.

DL Tabb et al. *J. Proteome Res*. (2015) 15: 691-706.

National Institute of Standards and Technology MSQC metrics

Chromatography

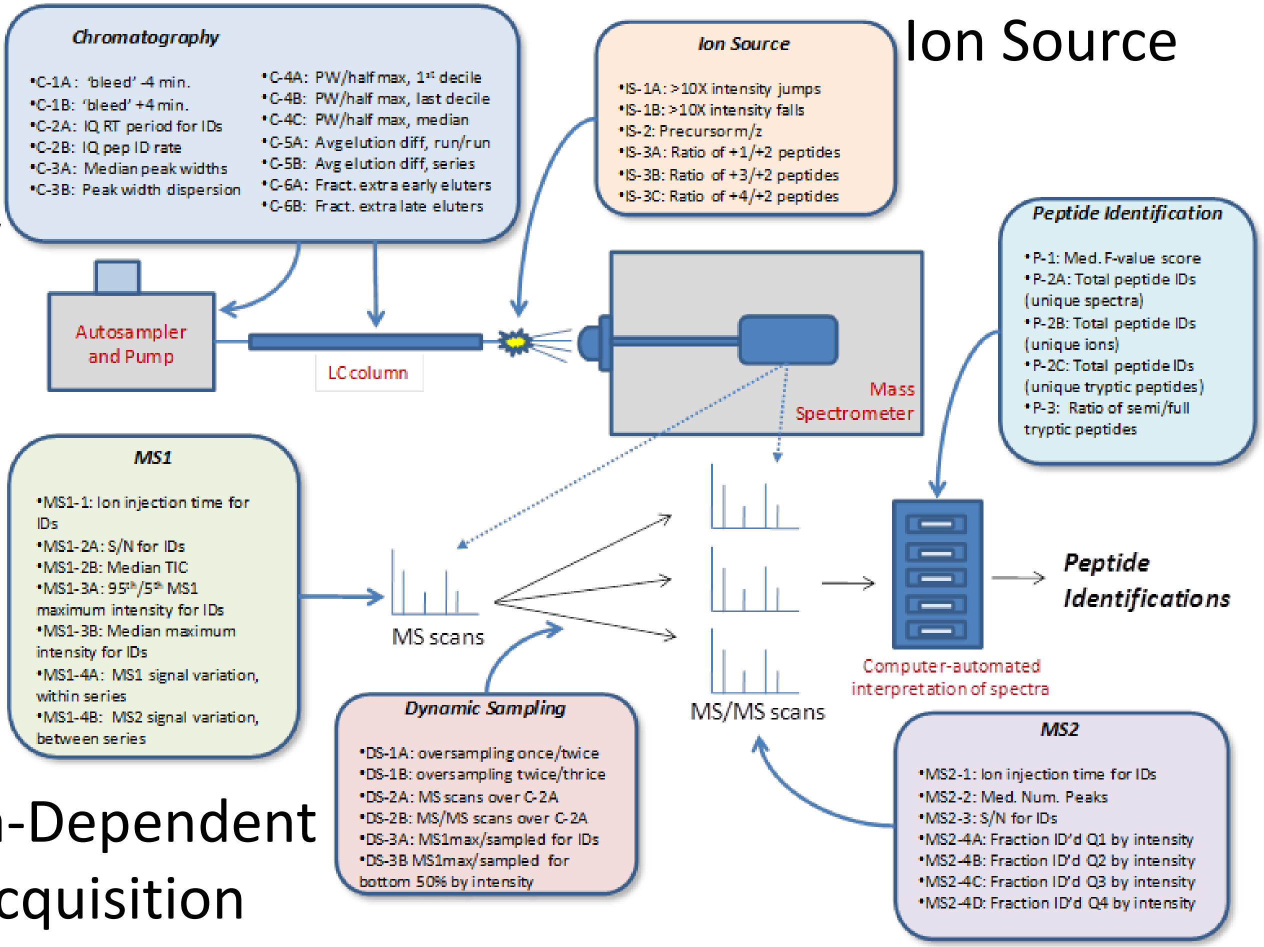
Ion Source

Peptide ID

MS1 Signals

MS2 Signals

Data-Dependent Acquisition



Chromatography

- C-1A: 'bleed' -4 min.
- C-1B: 'bleed' +4 min.
- C-2A: IQ RT period for IDs
- C-2B: IQ pep ID rate
- C-3A: Median peak widths
- C-3B: Peak width dispersion
- C-4A: PW/half max, 1st decile
- C-4B: PW/half max, last decile
- C-4C: PW/half max, median
- C-5A: Avg elution diff, run/run
- C-5B: Avg elution diff, series
- C-6A: Fract. extra early eluters
- C-6B: Fract. extra late eluters

Ion Source

- IS-1A: >10X intensity jumps
- IS-1B: >10X intensity falls
- IS-2: Precursor m/z
- IS-3A: Ratio of +1/+2 peptides
- IS-3B: Ratio of +3/+2 peptides
- IS-3C: Ratio of +4/+2 peptides

Peptide Identification

- P-1: Med. F-value score
- P-2A: Total peptide IDs (unique spectra)
- P-2B: Total peptide IDs (unique ions)
- P-2C: Total peptide IDs (unique tryptic peptides)
- P-3: Ratio of semi/full tryptic peptides

MS1

- MS1-1: Ion injection time for IDs
- MS1-2A: S/N for IDs
- MS1-2B: Median TIC
- MS1-3A: 95th/5th MS1 maximum intensity for IDs
- MS1-3B: Median maximum intensity for IDs
- MS1-4A: MS1 signal variation, within series
- MS1-4B: MS2 signal variation, between series

Dynamic Sampling

- DS-1A: oversampling once/twice
- DS-1B: oversampling twice/thrice
- DS-2A: MS scans over C-2A
- DS-2B: MS/MS scans over C-2A
- DS-3A: MS1max/sampled for IDs
- DS-3B MS1max/sampled for bottom 50% by intensity

MS2

- MS2-1: Ion injection time for IDs
- MS2-2: Med. Num. Peaks
- MS2-3: S/N for IDs
- MS2-4A: Fraction ID'd Q1 by intensity
- MS2-4B: Fraction ID'd Q2 by intensity
- MS2-4C: Fraction ID'd Q3 by intensity
- MS2-4D: Fraction ID'd Q4 by intensity

Rudnick et al. *Mol. Cell. Proteomics* (2010) 9: 225

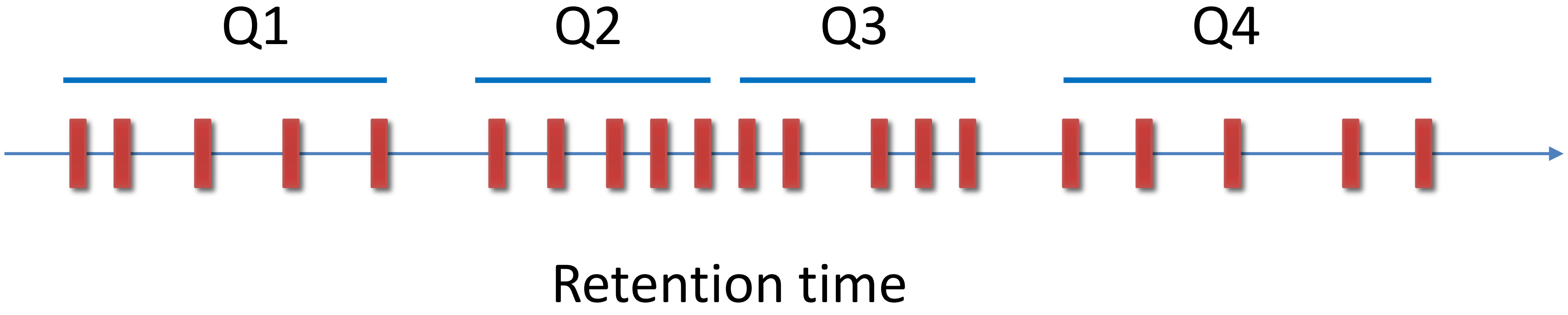
http://proteowizard.sourceforge.net/

QuaMeter “IDFree” metrics

- 40 metrics categorized into extracted ion chromatograms, retention times, mass spectrometry, and tandem mass spectrometry
- Metrics emphasize quartiles and log ratios to give robust behavior in extreme variation
- Example fields for rapid characterization:
  - XIC-FWHM-Q2: median of peak widths in time
  - RT-TIC-Q*x*: relative duration for TIC integration
  - MS1-Density-Q2: median of peak counts for MS
  - MS2-Freq-Max: Fastest rate of MS/MS acquisition

```
quameter.exe *.raw -MetricsType idfree -OutputFilepath metrics.tsv
```

RT-Duration	What is the highest scan time observed minus the lowest scan time observed?
RT-TIC-Q1	The interval when the first 25% of TIC accumulates divided by RT-Duration
RT-TIC-Q2	The interval when the second 25% of TIC accumulates divided by RT-Duration
RT-TIC-Q3	The interval when the third 25% of TIC accumulates divided by RT-Duration
RT-TIC-Q4	The interval when the fourth 25% of TIC accumulates divided by RT-Duration
RT-MS-Q1	The interval for the first 25% of all MS events divided by RT-Duration
RT-MS-Q2	The interval for the second 25% of all MS events divided by RT-Duration
RT-MS-Q3	The interval for the third 25% of all MS events divided by RT-Duration
RT-MS-Q4	The interval for the fourth 25% of all MS events divided by RT-Duration
RT-MSMS-Q1	The interval for the first 25% of all MS/MS events divided by RT-Duration
RT-MSMS-Q2	The interval for the second 25% of all MS/MS events divided by RT-Duration
RT-MSMS-Q3	The interval for the third 25% of all MS/MS events divided by RT-Duration
RT-MSMS-Q4	The interval for the fourth 25% of all MS/MS events divided by RT-Duration



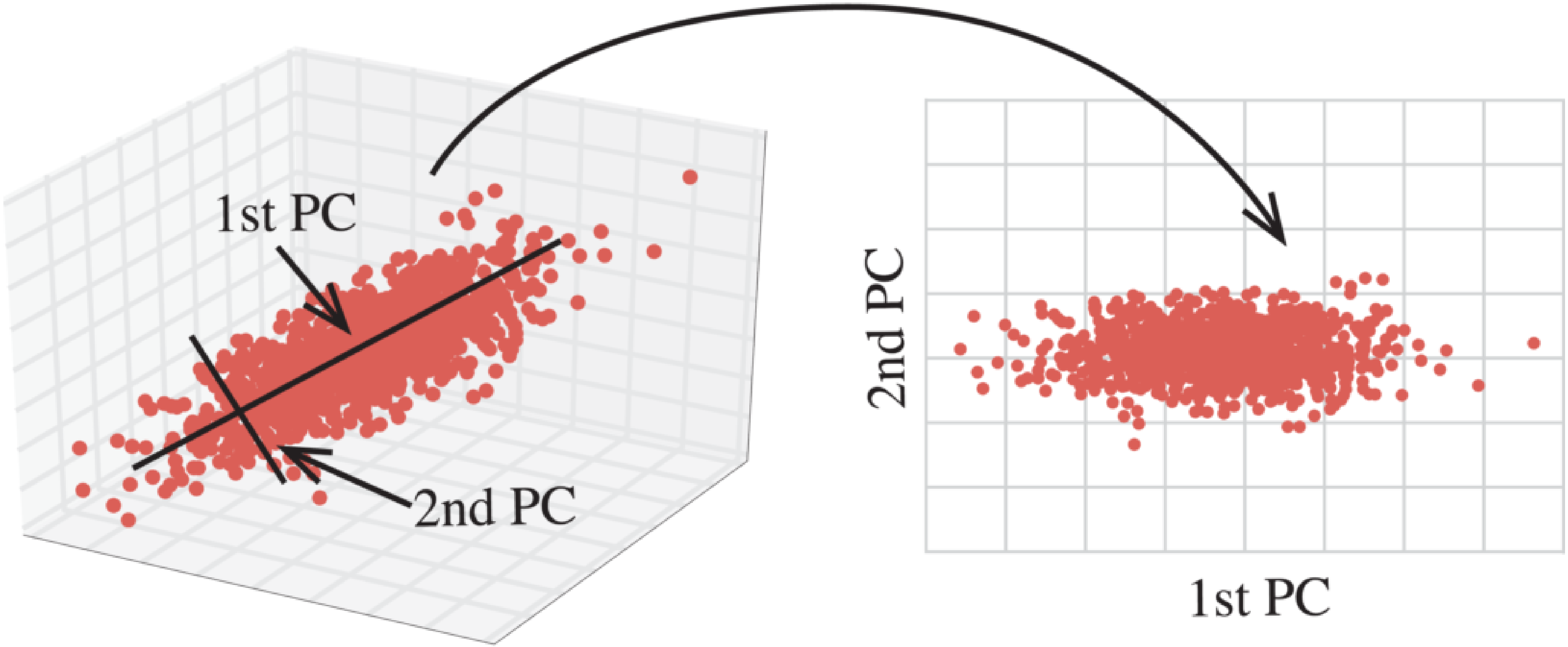
X Wang et al. *Anal. Chem*. (2014) 86: 2497-2509

A role for dimensionality reduction

- Measurements may be redundant. They may contain *mutual information* or be *correlated* with each other.
- Principal Components Analysis combines these metrics into components that account for observed variance.
  - "Component:" PCA accepts a table of *n* metrics for each sample; each component is a linear combination.
  - "Principal:" PCA prioritizes the components by the amount of observed variability that each explains.

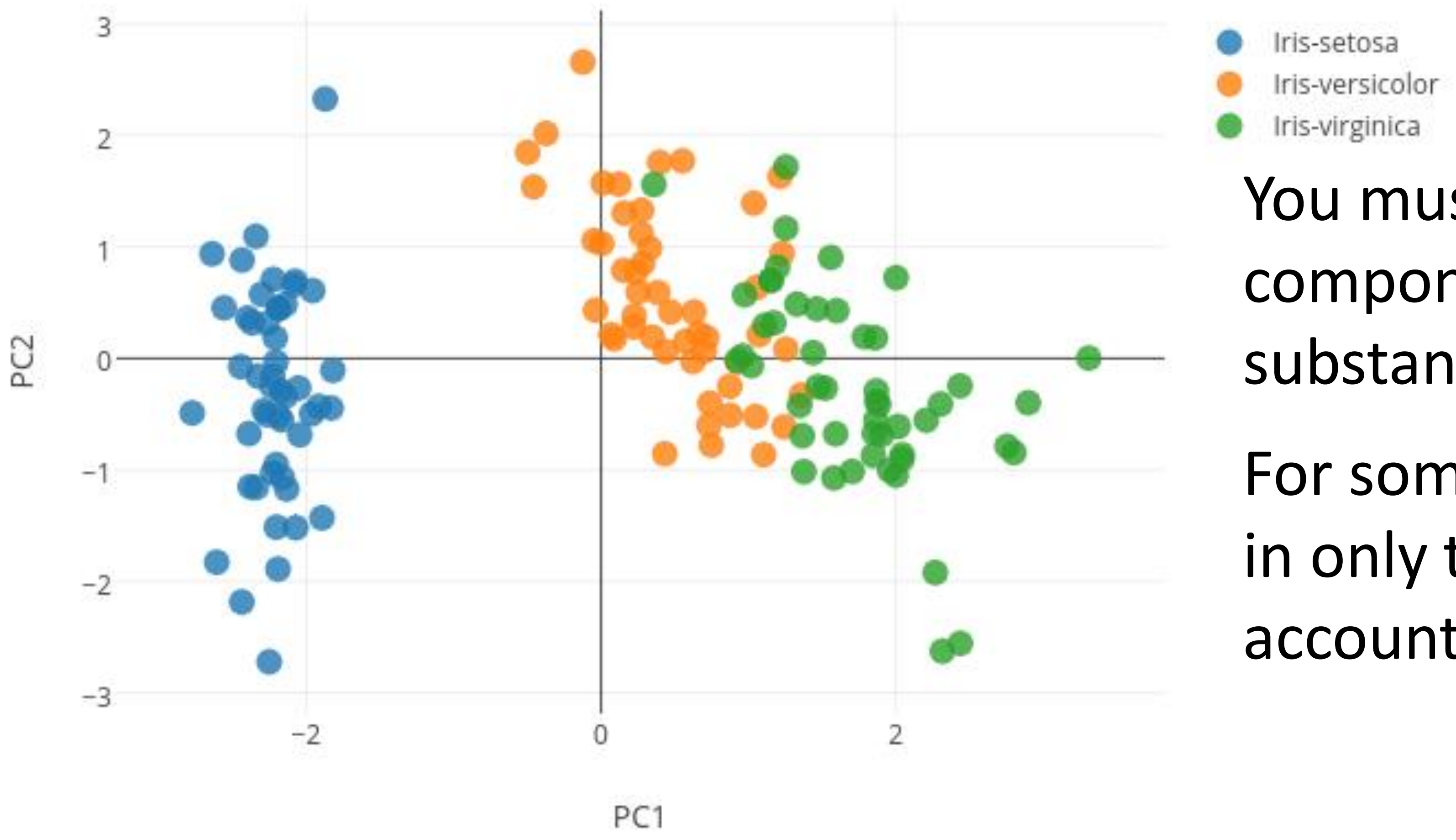
http://www.theanalysisfactor.com/tips-principal-component-analysis/

Projection pursuit: PCA rotates data in *n*-D space



http://innovaleur.com/exploring-the-structure-of-high-dimensional-data-with-hypertools-in-kaggle-kernels/

PCA plot typically shows first two components

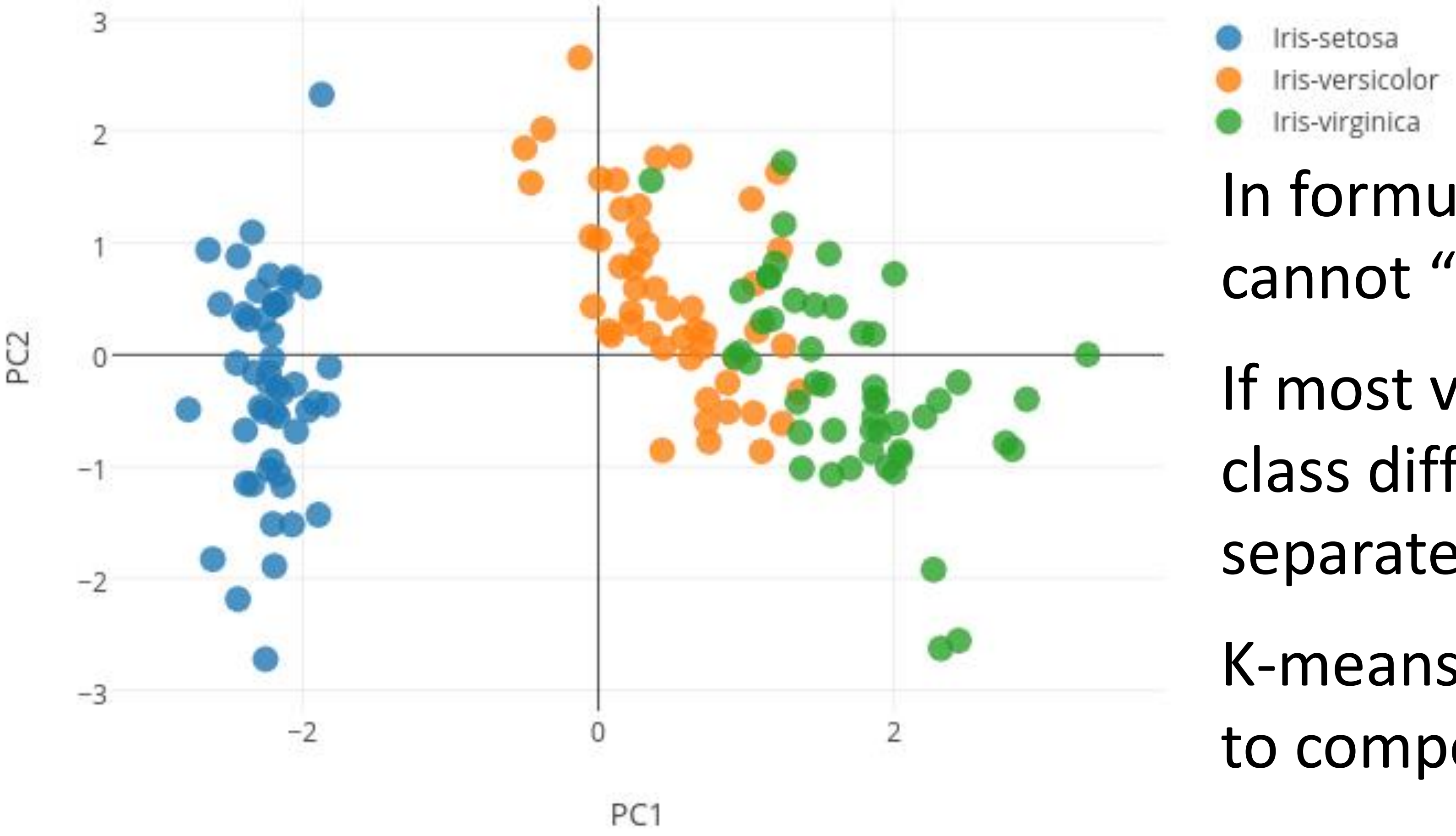


You must keep enough components to account for substantial variability.

For some experiments, visualizing in only two dimensions may not account for enough variability.

https://plot.ly/ipython-notebooks/principal-component-analysis/

Unsupervised learning: PCA may reveal classes



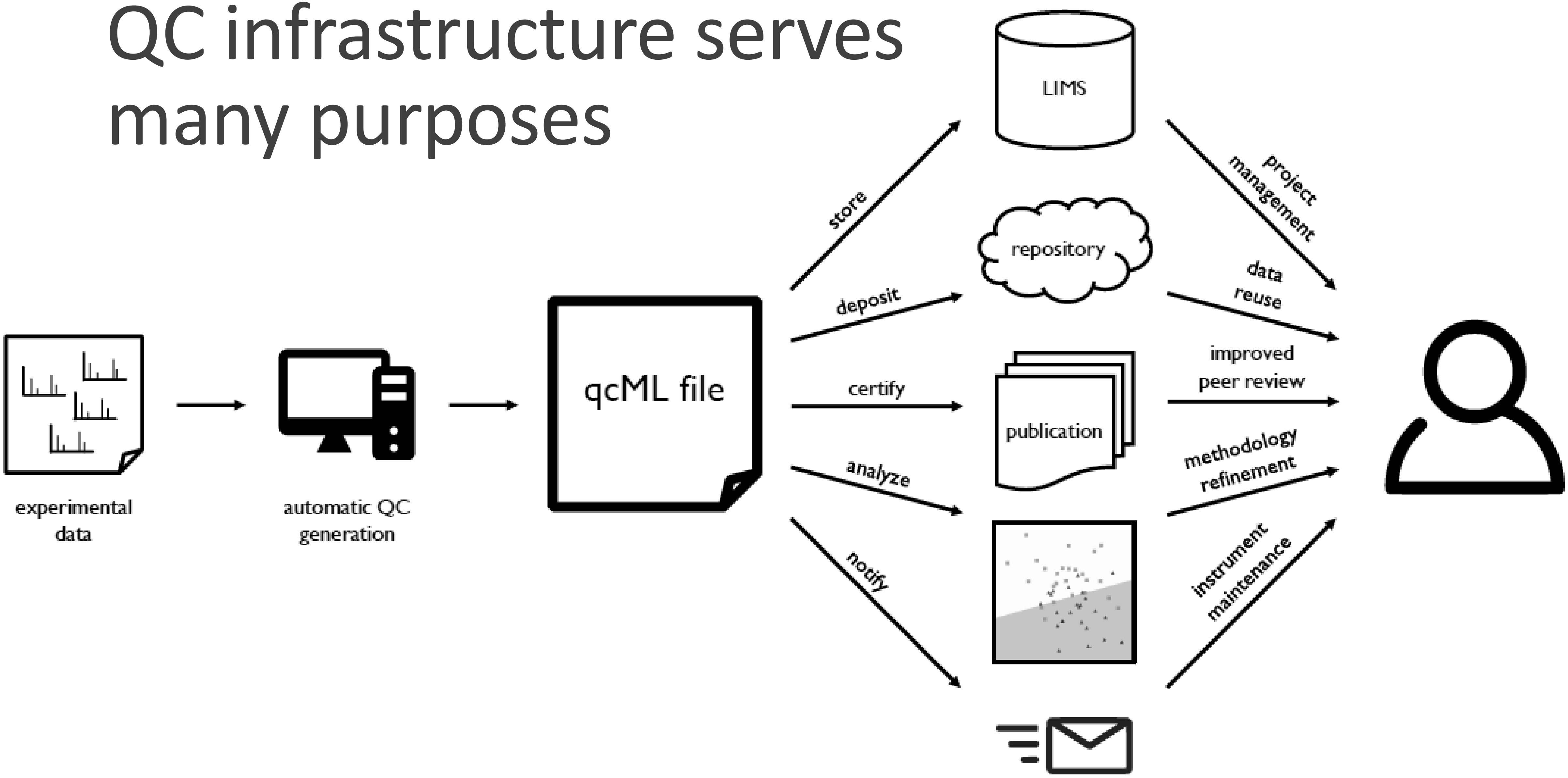
In formulating components, PCA cannot “see” class labels.

If most variance comes from inter-class differences, the classes will separate in PCA plot.

K-means clustering can be applied to components.

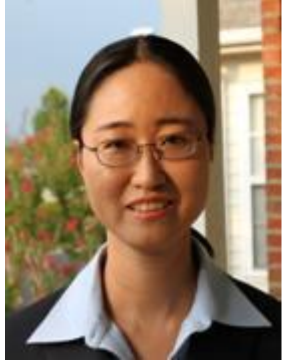
C. Ding and X. He. *ICML ’04: Proc. 21st Intl Conf. Machine Learn*

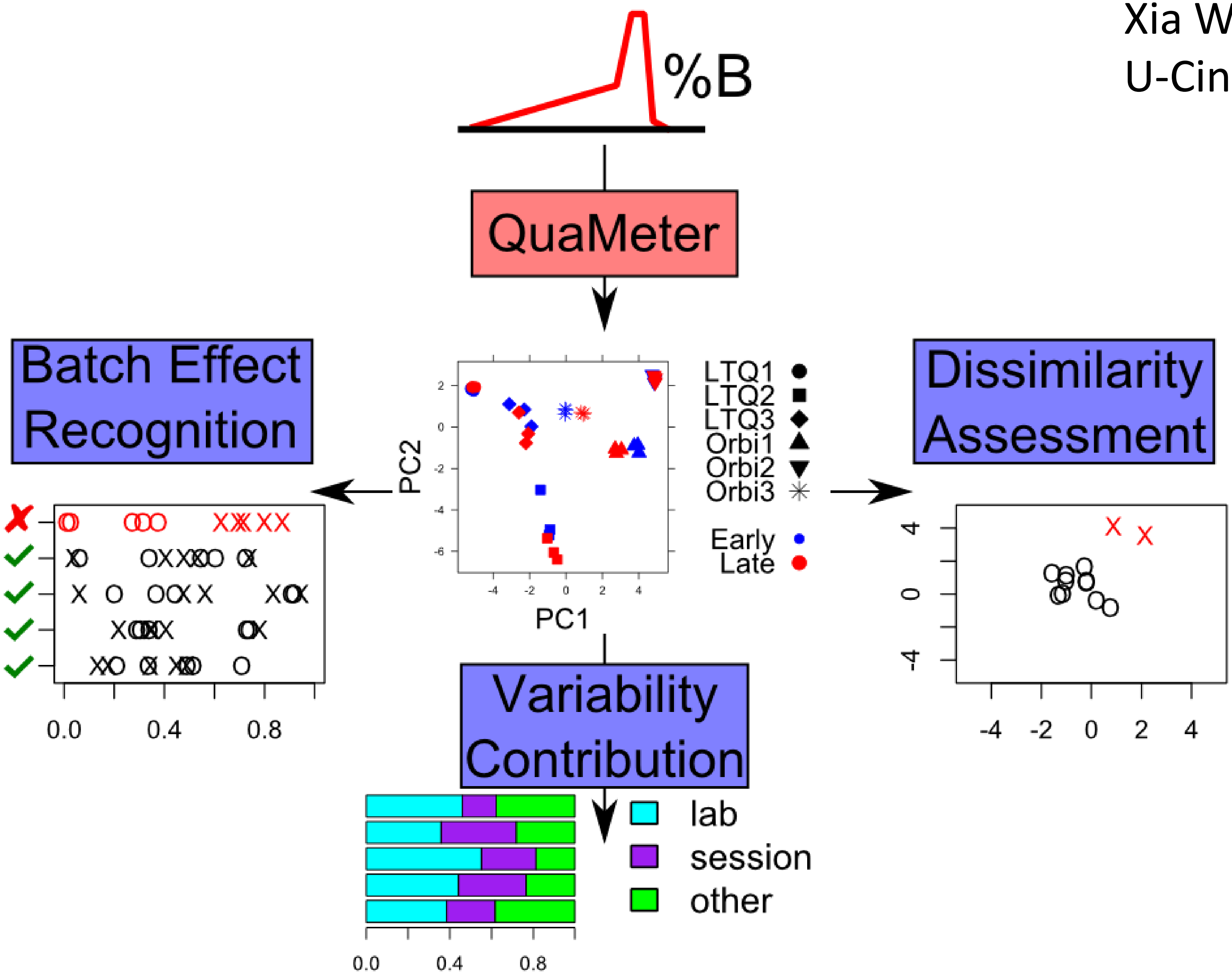
QC infrastructure serves many purposes



W. Bittremieux et al. *Anal Chem*. (2017) 89: 4474-4479

Xia Wang,
U-Cincinnati





QuaMeter

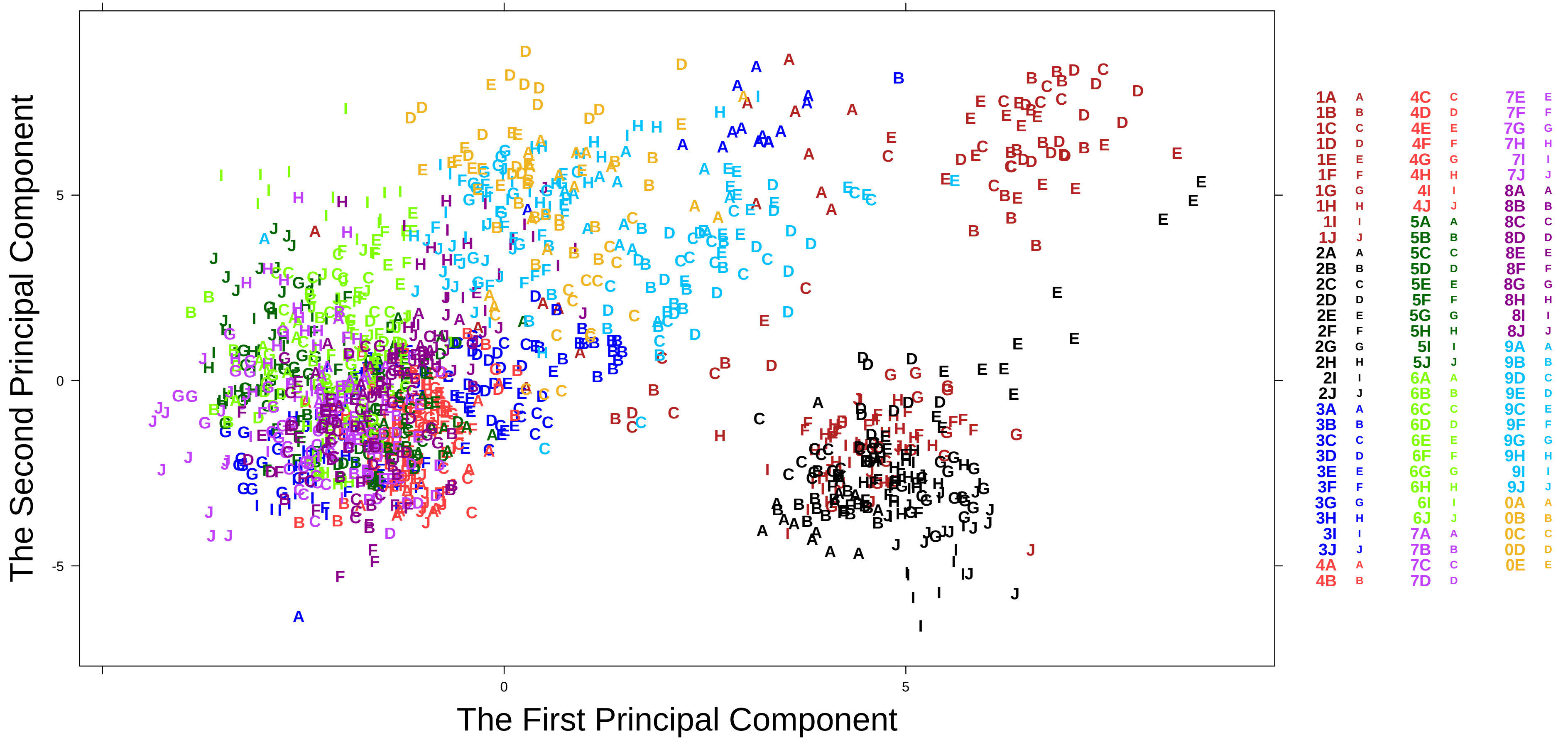
Batch Effect Recognition

Dissimilarity Assessment

Variability Contribution

Wang et al. (2014) *Analytical Chemistry* 86: 2497-2509.

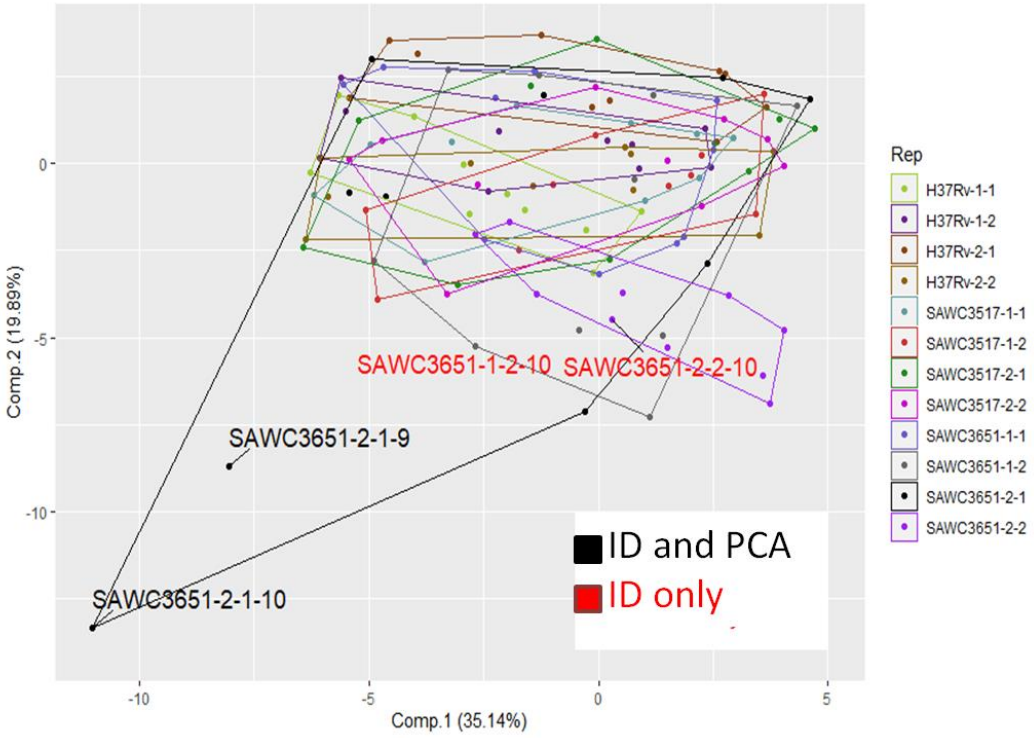
CPTAC Colon: 95 bRPLCs of 15 fxns over eight months of operation



Slebbos et al. *Scientific Data* (2015) 2: 150022

Marina Kriek, SUN

Outliers may dominate variance

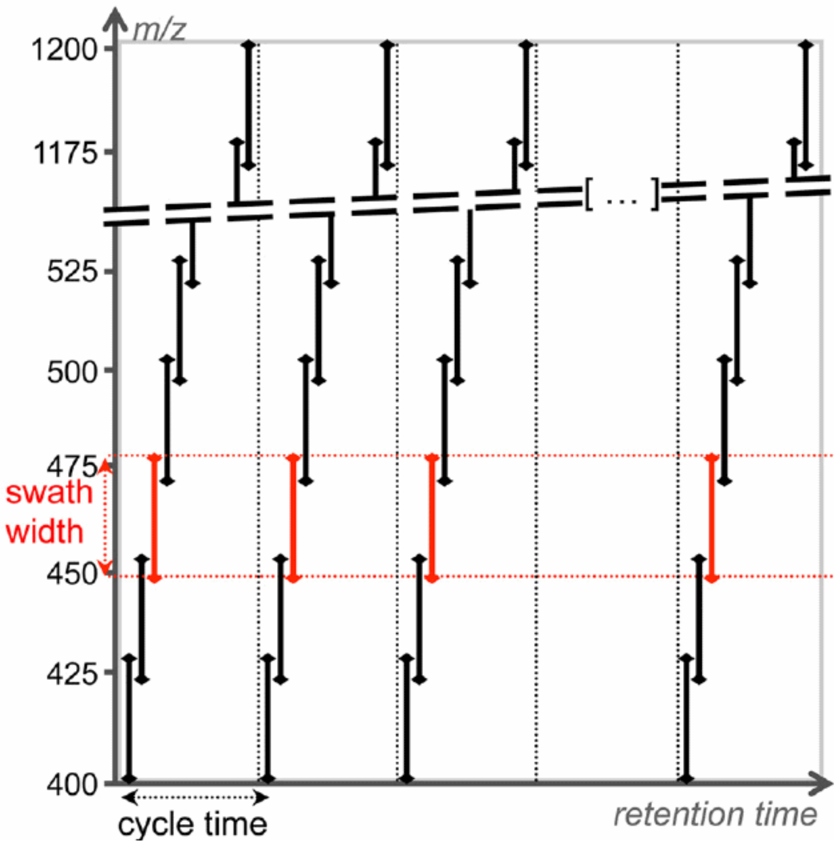


- Three *M. tuberculosis* strains were subjected to GeLC-MS.
- Two LC-MS/MS experiments were far from the “cloud.”
- We rediscovered that data production was temporarily halted after this pair.

M. Kriek et al. *Proteomics* (2020) 10.1002/pmic.201900382

Marina Kriek, SUN

Specializing QC metrics to DIA



The SwaMe project describes SWATH experiments with QC metrics:

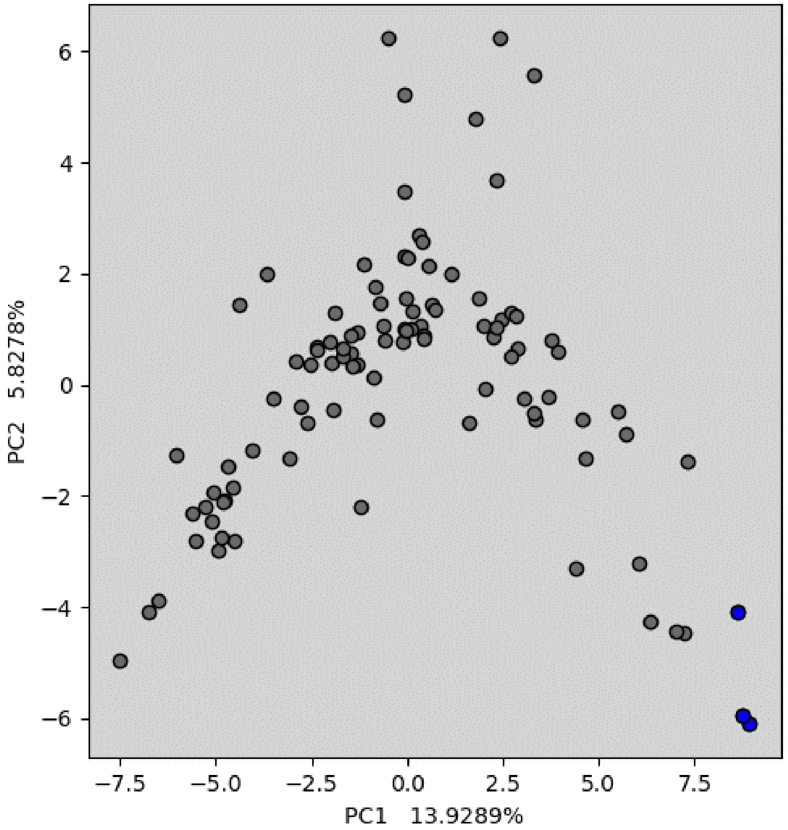
- Comprehensive metrics apply to complete experiment.
- Isolation window metrics detail each SWATH individually.
- RT metrics subdivide LC retention to different time intervals.

Image from Gillet et al. *Mol. Cell. Proteomics* (2012) 10.1074/mcp.O111.016717–3

Marina Kriek, SUN

Struggling with R? Assurance simplifies QC

PCA of the comprehensive QC metrics



- Assurance enables you to produce QuaMeter or SwaMe metrics and visualize them through a GUI.
  - PCA of metrics highlights outliers.
  - Images interrogate experiments for a particular metric.
  - Machine learning examines longitudinal series data.

https://github.com/marinaPauw/Assurance

mzQC: designing a HUPO-PSI standard for QC information

- Tab-separated text is fine for simple tables, but what if a single metric is a matrix of values?
- Just as mzML represents a text format for mass spec data, an mzQC represents a text format for quality data, easing passage from one tool to another.
- The use of JSON streamlines the size of mzQC documents.

https://github.com/HUPO-PSI/mzQC

Takeaways

- Proteomics QC software is growing in maturity. Bench researchers are no longer required to “eyeball it.”
- Statistical models based on QC metrics can recognize outliers, batch effects, and factors contributing variability.
- PCA is ubiquitous and valuable; we can all improve our understanding of how it works!