



Interplay of Chemometrics and Large-Scale Databases

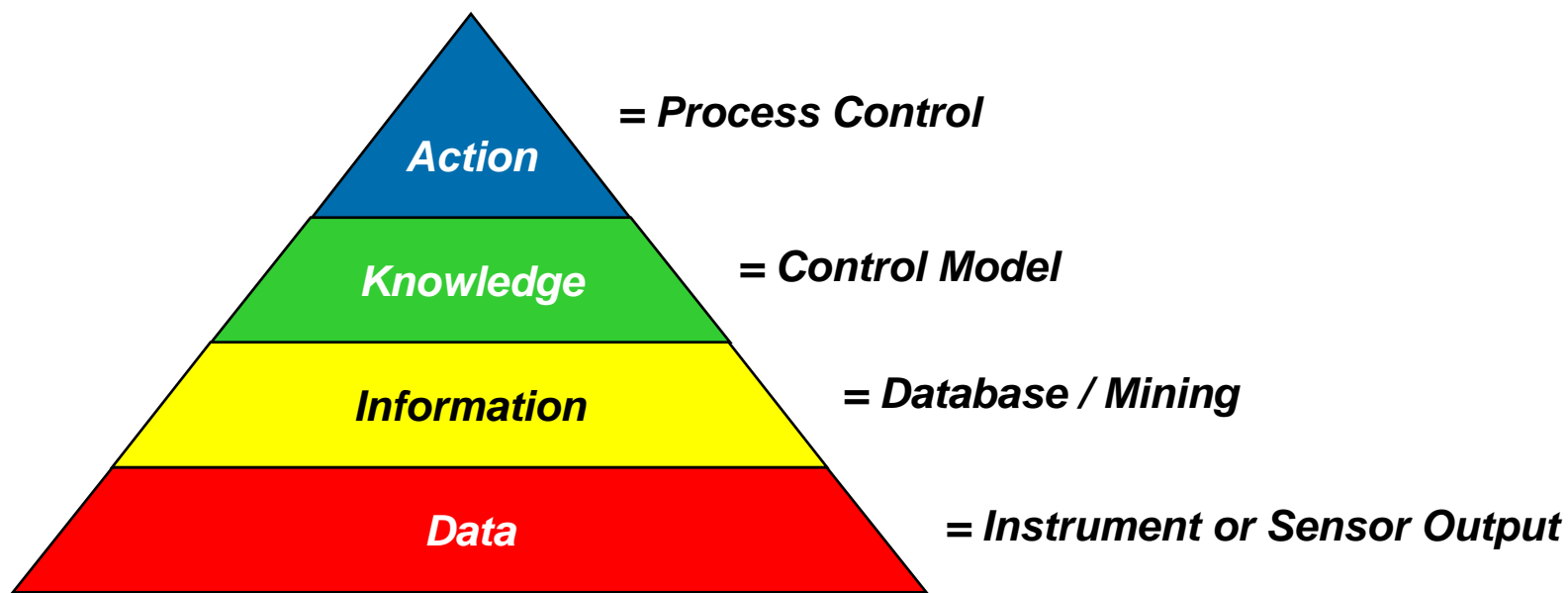
*Scott Ramos & Brian Rohrbach, Infometrix, Inc.,
Gregory Banik, BioRad Laboratories Informatics Division*

Agenda

The chemometrics role

- Compress the data to fit into a process historian structure better
- Reduce the number of chromatograms or spectra required to manage a process effectively
- Evaluate a possibly aberrant chromatogram to see if a similar trace has been seen in any plant at any time in the past
- Execute a database query to identify the cause of off-target features

Data-information Value Hierarchy



Each layer above is a refinement of the layer below.

Situation

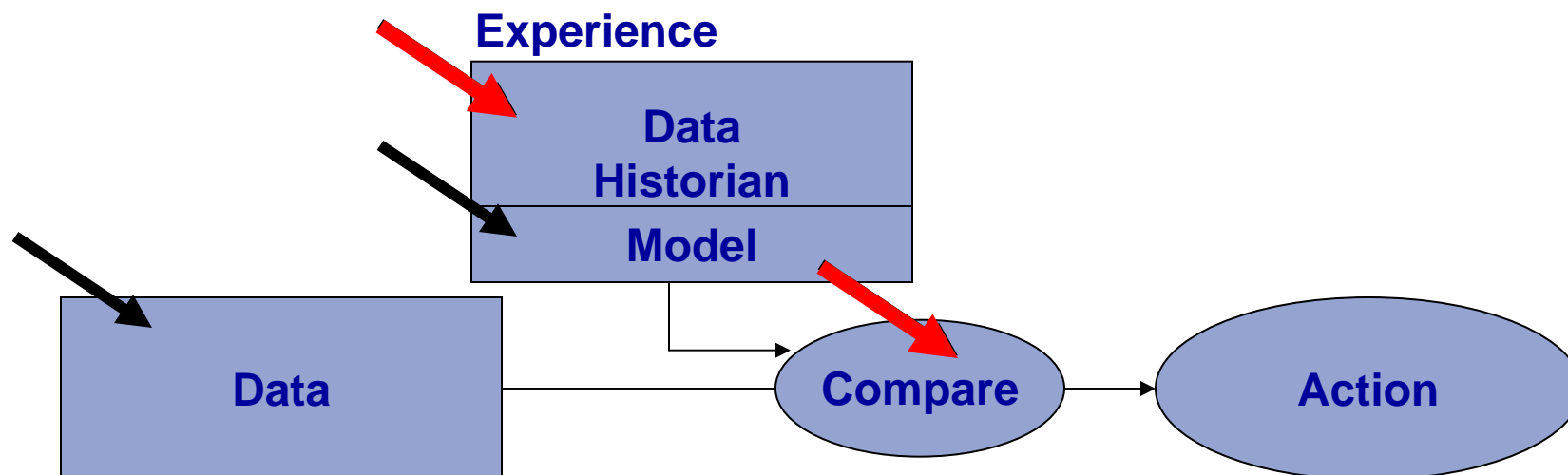
PAT initiative has focused attention on use of multivariate instruments on-line

- Addition of spectrometers and chromatographs to arena of control
 - yields advantage in monitoring and in understanding of critical quality attributes
- But
 - size of data files and number of spectra and chromatograms overwhelming

For process quality improvement

- Store routine spectra as well as data used to generate the multivariate model in the first place

Chemometrics Roles in the Process



- ✓ Signal improvement
 - ✓ Modeling and optimization
 - ✓ Routine data processing
 - ✓ Database optimization

Database Optimization

Incorporation of multivariate instruments puts pressure on process archival activity, leading to choices

- Grab the “PLS” numbers, no spectra ✘
- Take only the occasional spectrum

How to decide?

- Compress the data
- Employ an alternate spectral storage medium

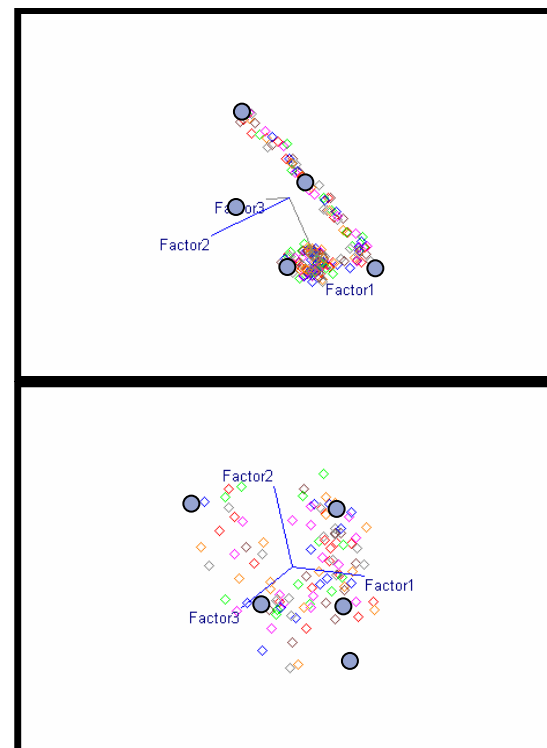
Another technology?

Selecting Samples

Kennard & Stone

Find samples most dispersed across the data space.

- Euclidean distances are calculated among samples.
- Select first two samples with largest intersample distance.
- Add samples to the list by two criteria:
 - for all samples not yet in the list, determine its nearest neighbor among the current selections;
 - select that sample whose nearest neighbor is of the largest distance.

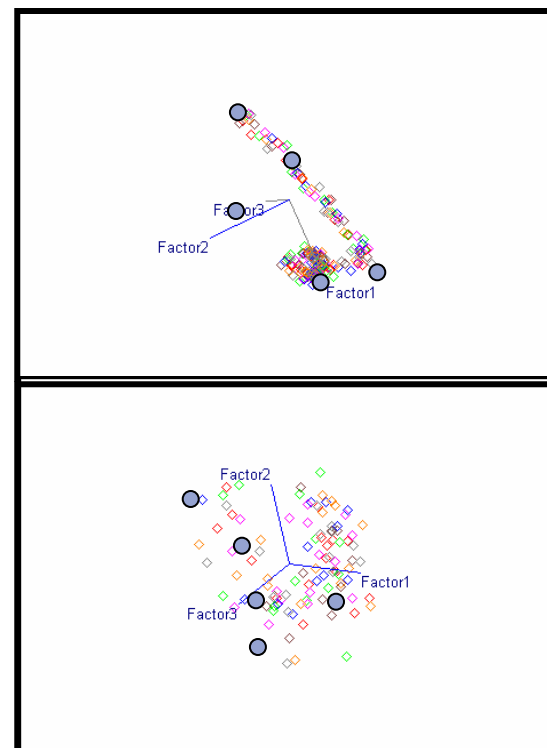


Selecting Samples

Orthogonal Leverage

Find samples of greatest influence within the data space.

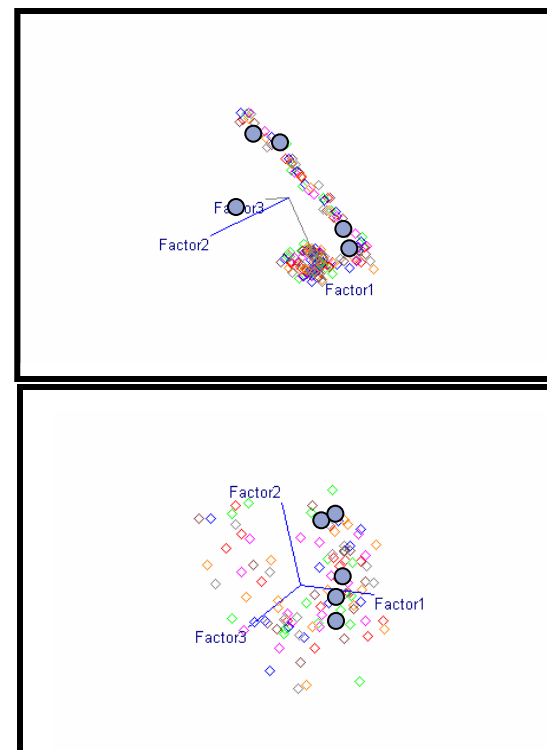
- Leverages are computed among all samples; sample of greatest leverage is chosen
- Remaining samples are orthogonalized against previous and new sample chosen
- Process is repeated until desired number of samples are selected.



Selecting Samples

PCA Hypergrid

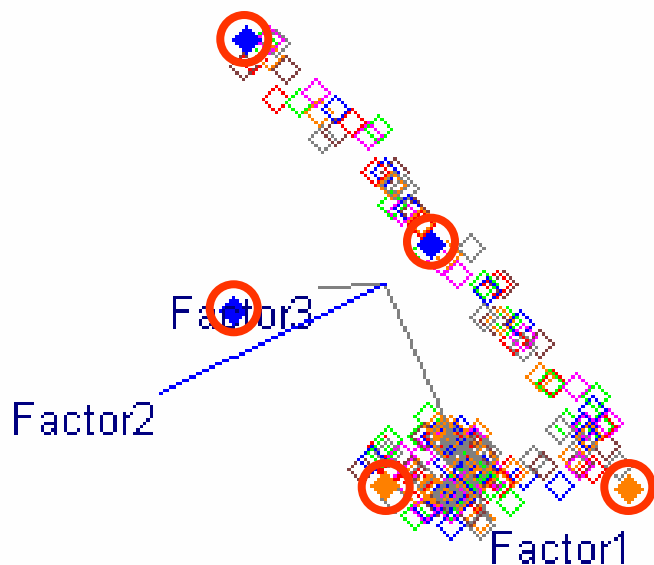
- Find samples that are most uniformly distributed in a reduced factor space
- PCA is run on sample set, factors are trimmed, reducing dimensionality
 - In trimmed scores space, a hypergrid is formed by dividing each factor dimension proportionally
 - Samples are selected by choosing one sample nearest the center of each block formed from the hypergrid



Selecting samples: Kennard & Stone

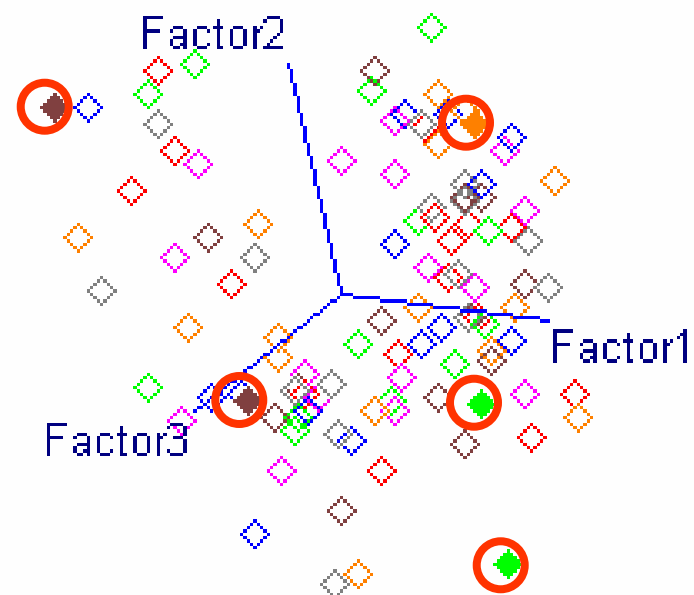
Non-uniform distribution

187 chromatograms

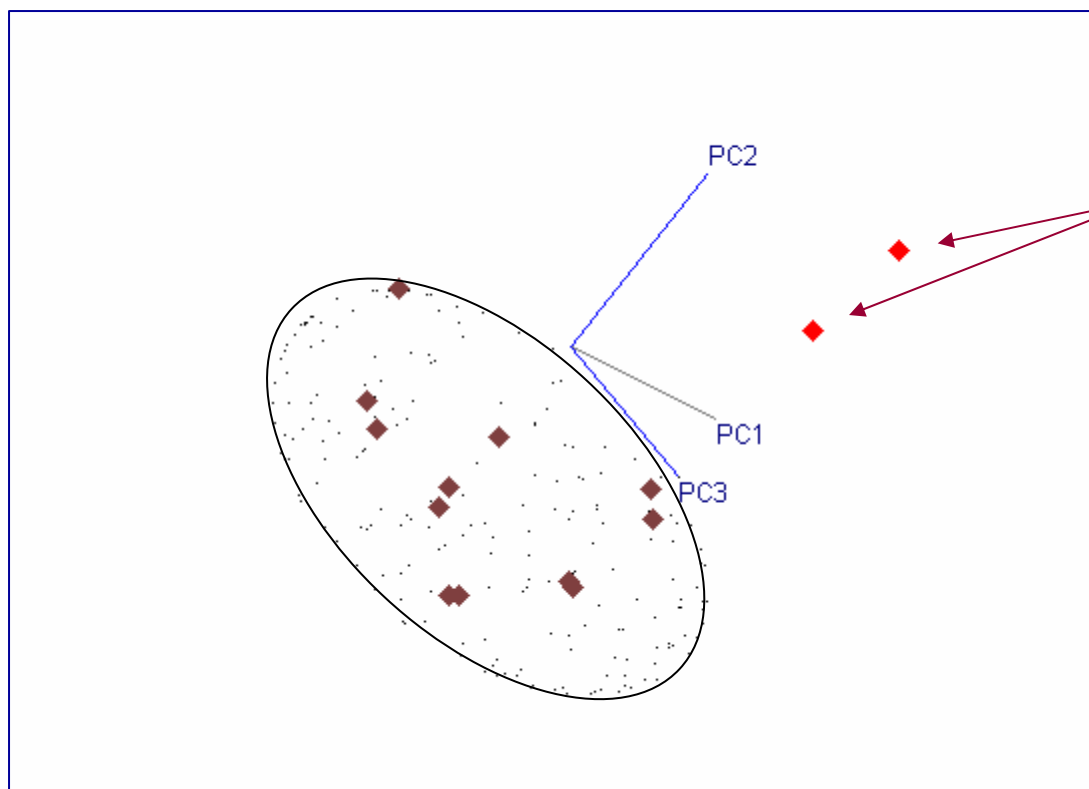


Uniform distribution

108 chromatograms



During routine evaluation



Outliers, because of:

- Instrument problem?
- Ingredient change?
- Process upset?

Or,

- Good samples, just low frequency?

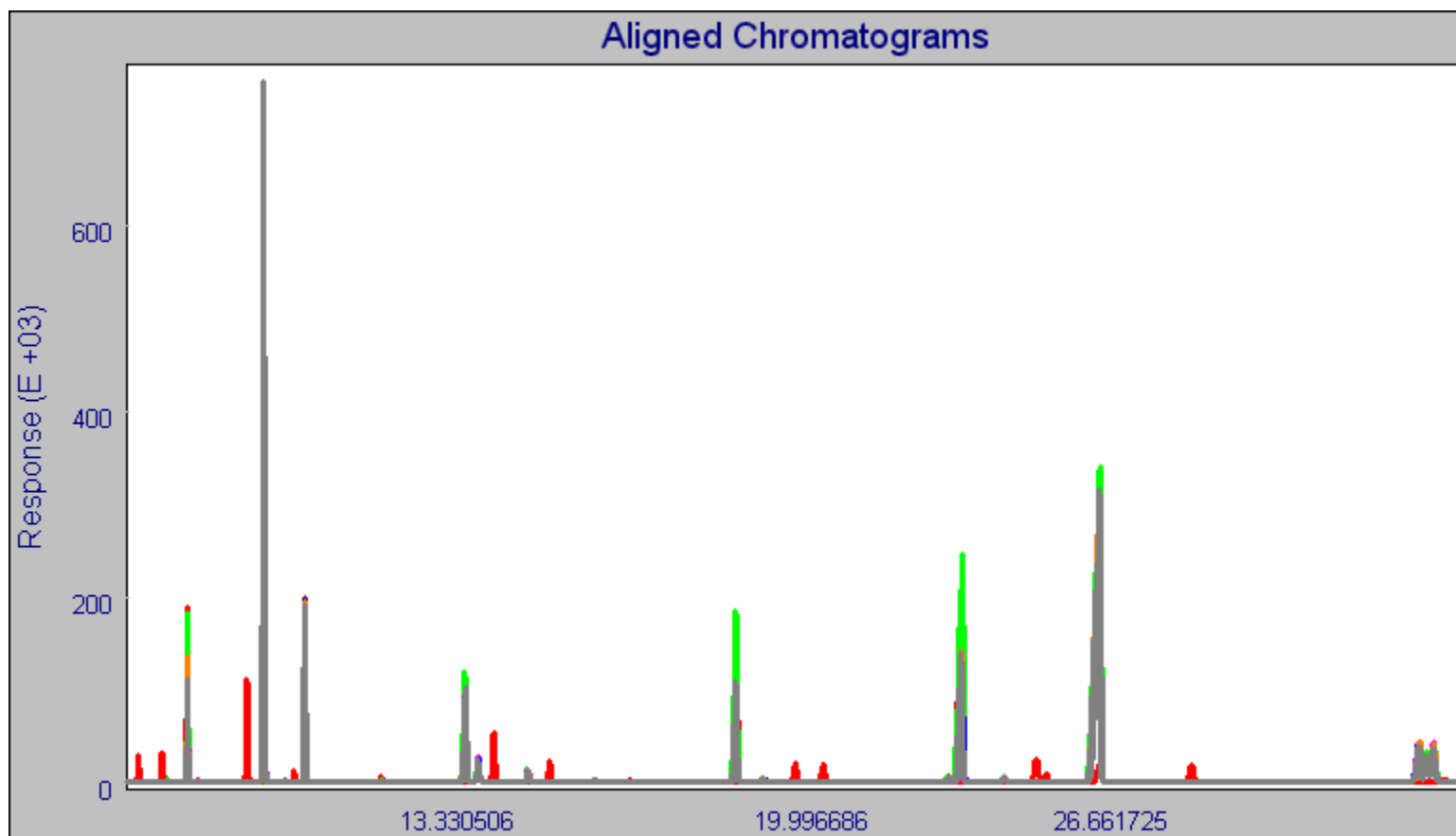
Data compression

PCA data compression allows reducing dimensionality of the data to a size manageable within a historian

- Faster searches
- Reduced storage requirements
- Can reconstruct original data

- MVA on compressed 'scores'

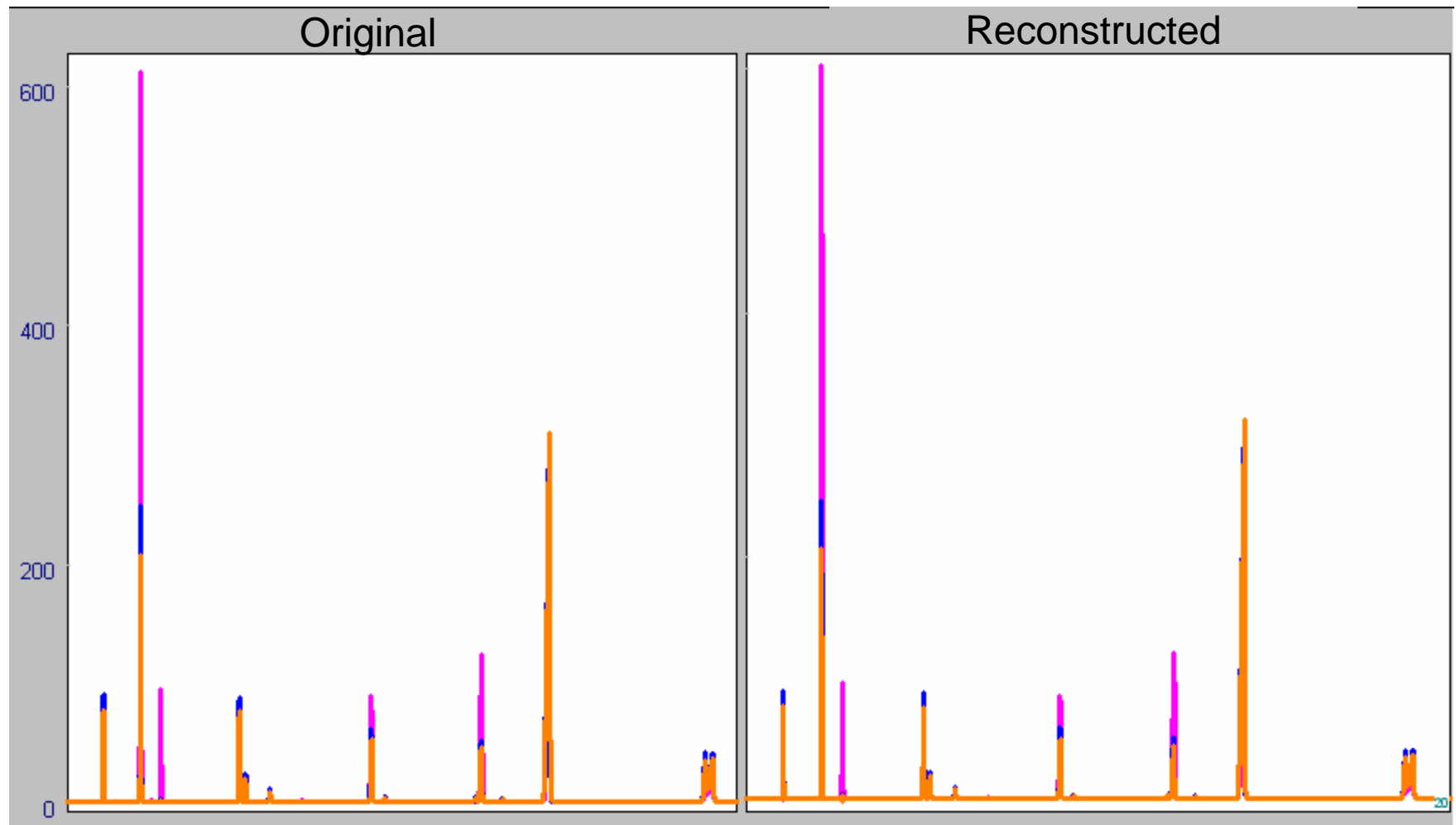
Examining alkylate chromatograms



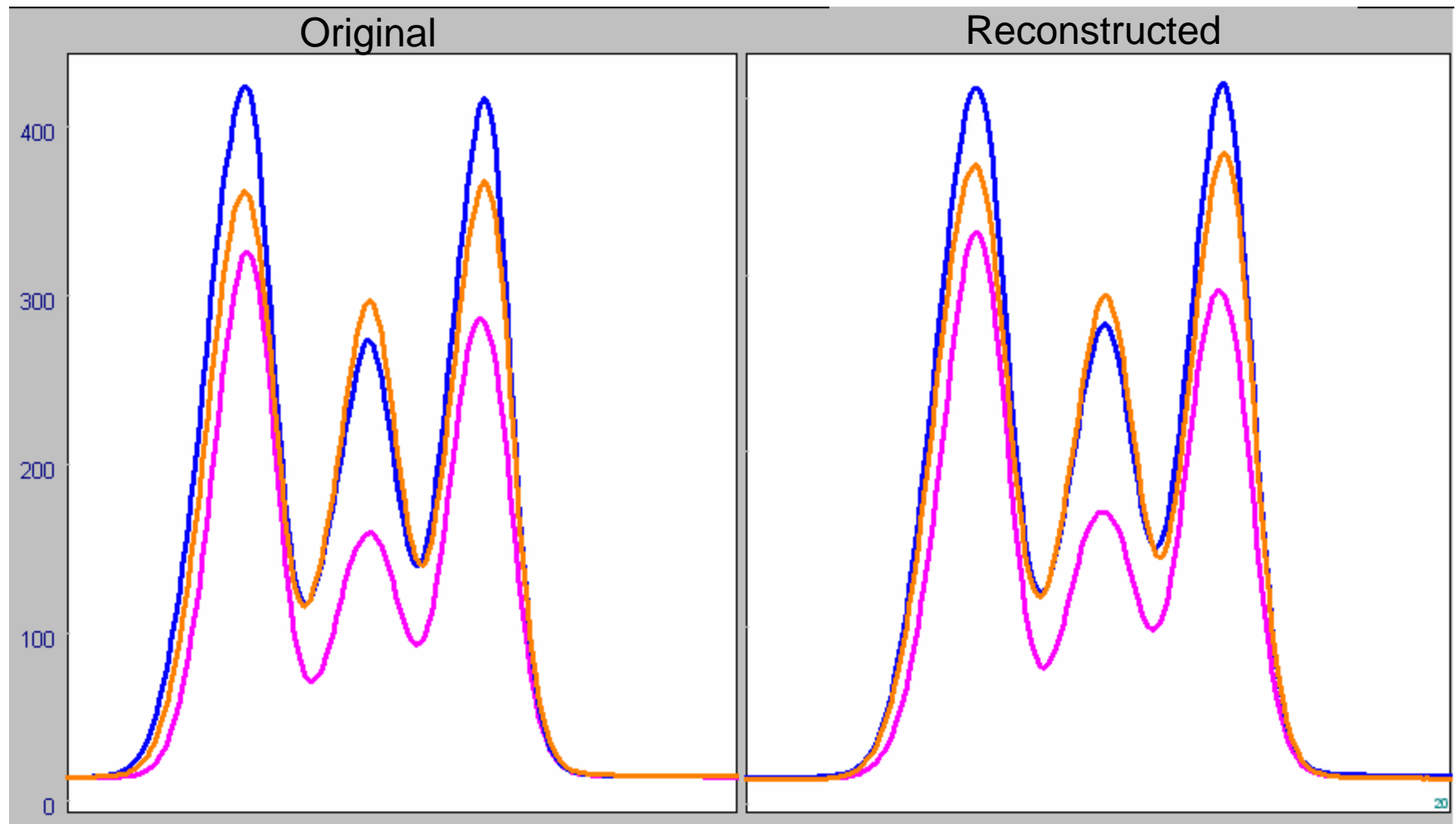
PCA of alkylates

	<u>Percent</u>	<u>Cumulative</u>
Factor1	35.64	35.64
Factor2	29.20	64.84
Factor3	19.46	84.30
Factor4	6.96	91.26
Factor5	2.73	93.99
Factor6	1.98	95.96

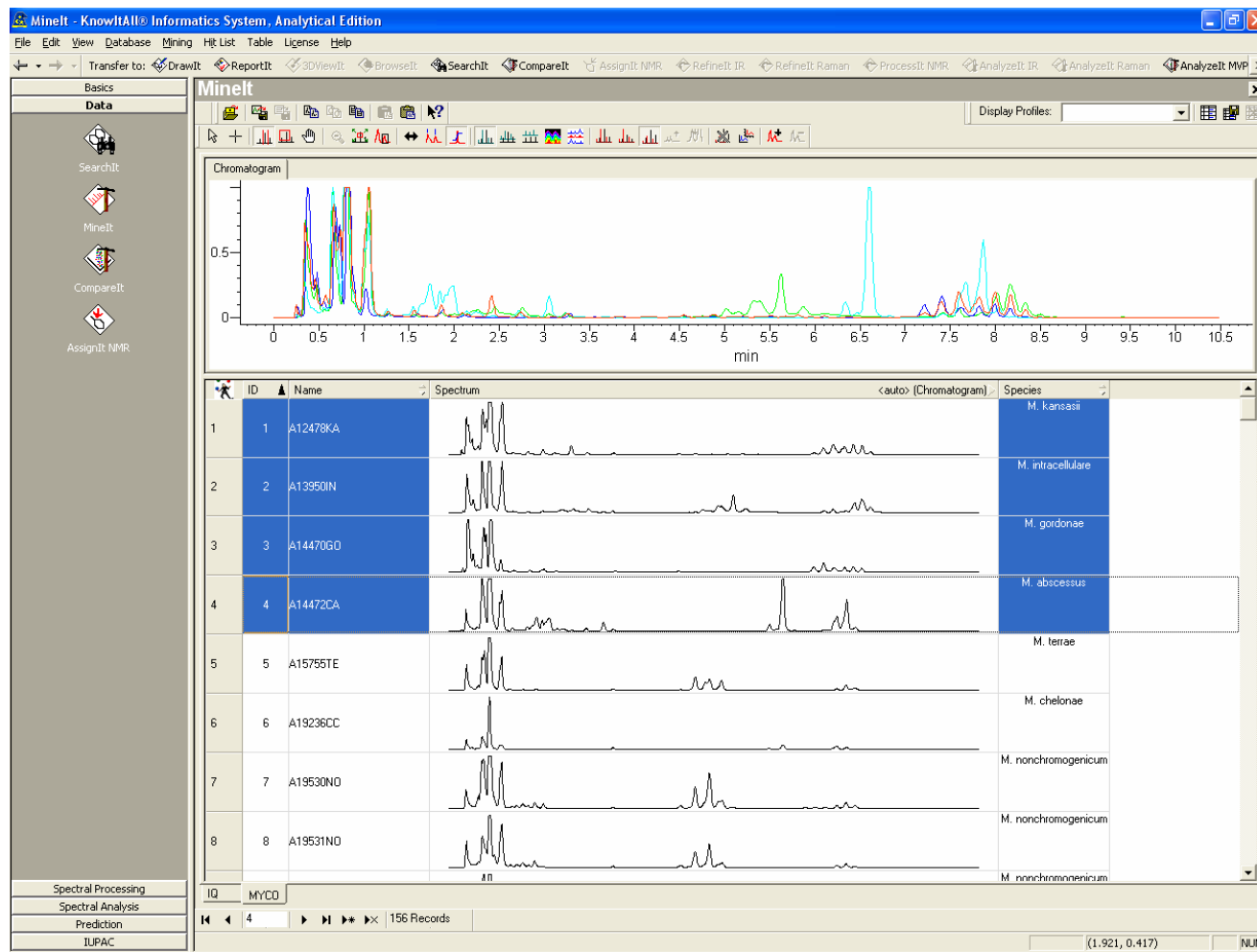
Original vs. 20-factor model



Original vs. 20-factor model

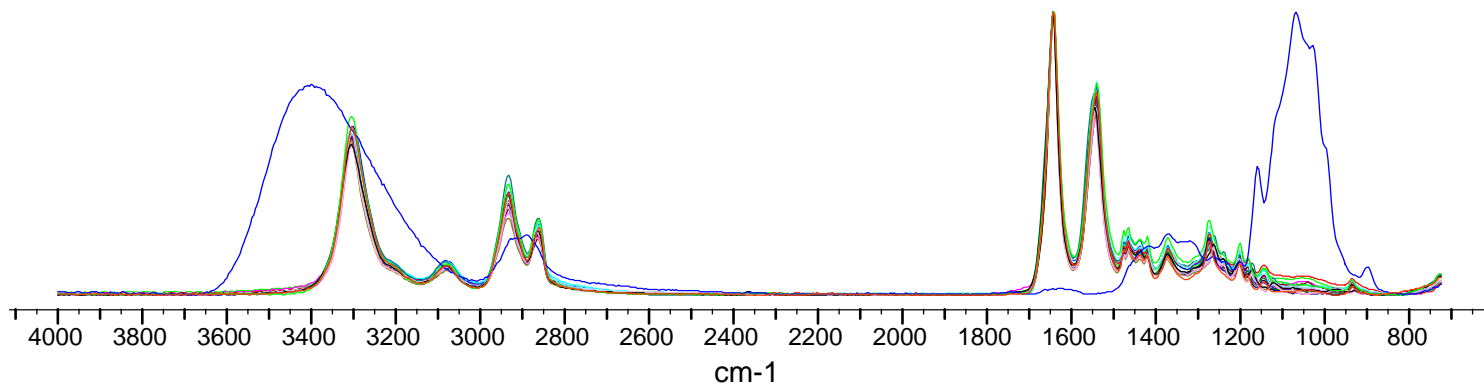


Spectral library software



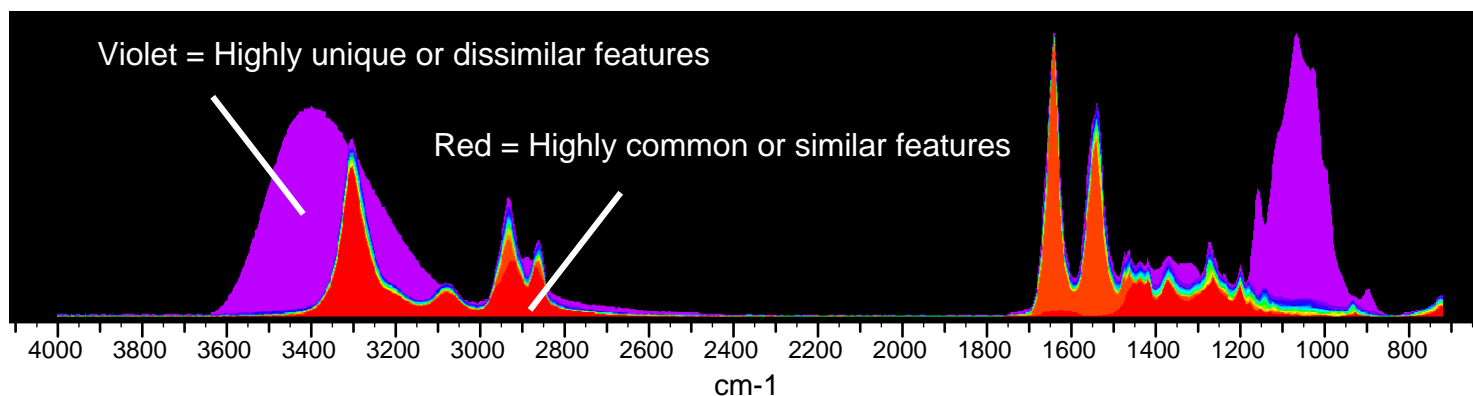
Spectral Presentation

Overlay Representation of Overlapped Spectra

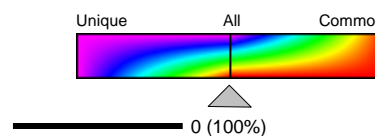


Spectral Presentation

Overlap Density Heatmap - OD Level = 0

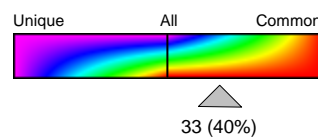
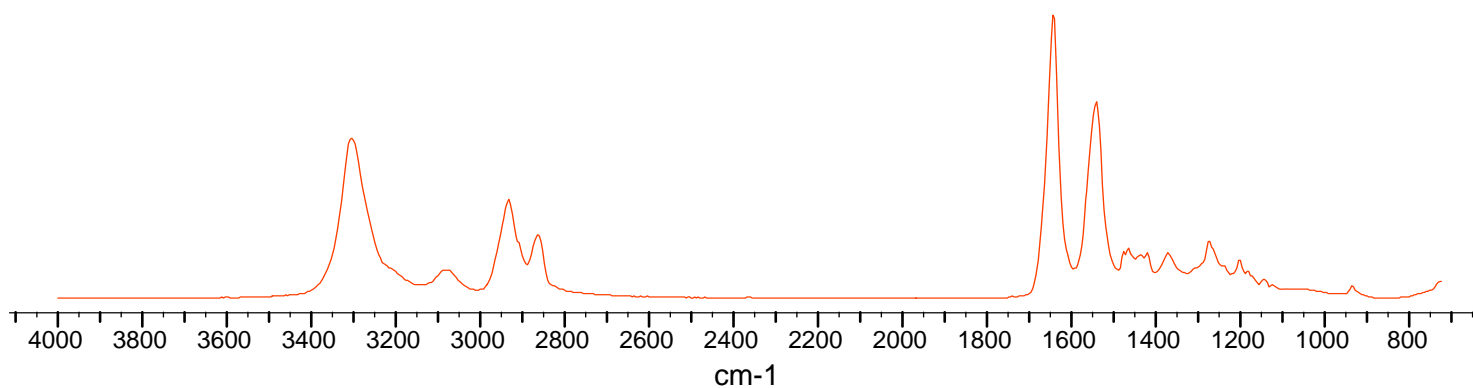


OD Level – Arbitrary scale where 0 shows all overlap density levels, 100 shows areas common to all spectra, and -100 shows areas unique to only one spectrum.

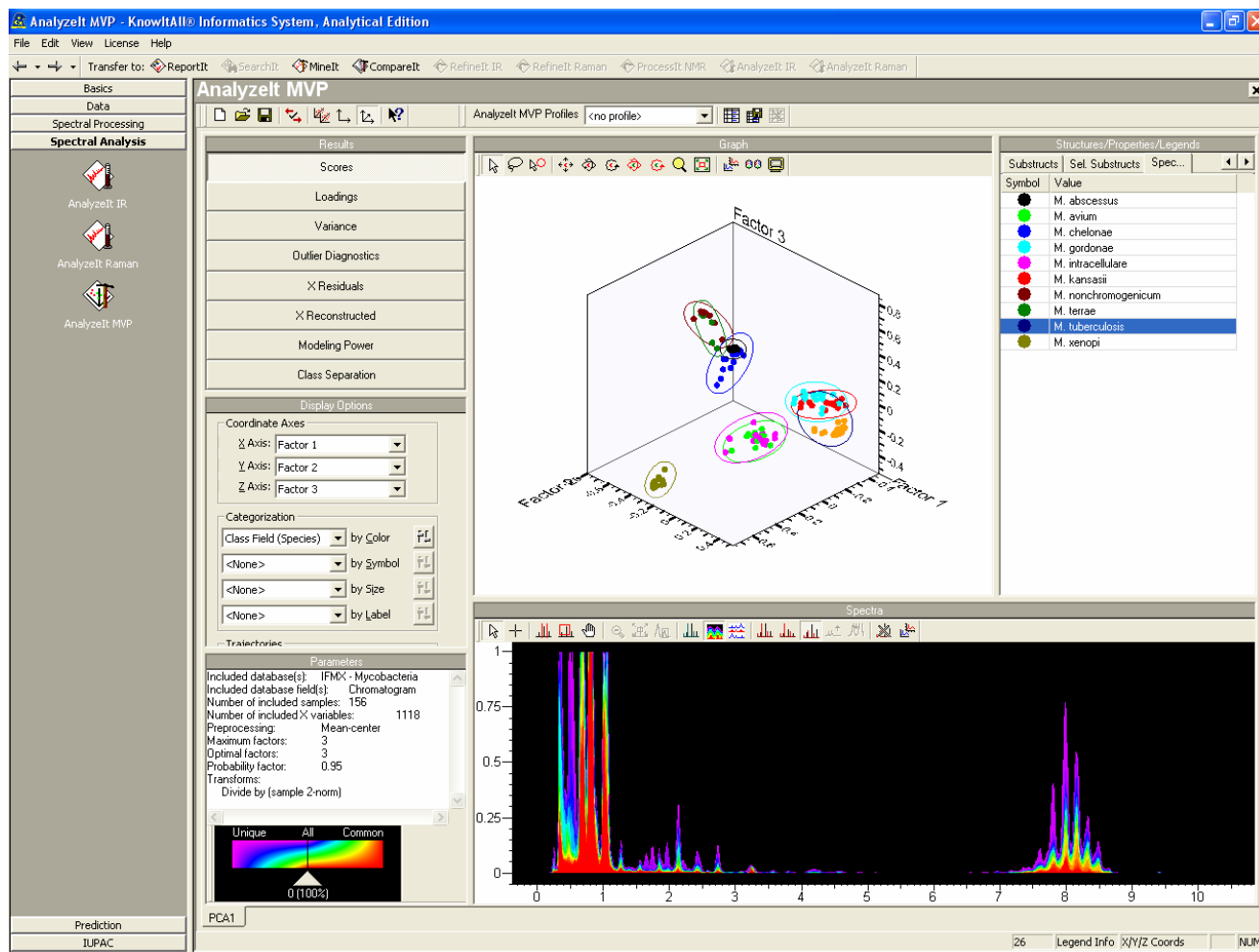


Spectral Presentation

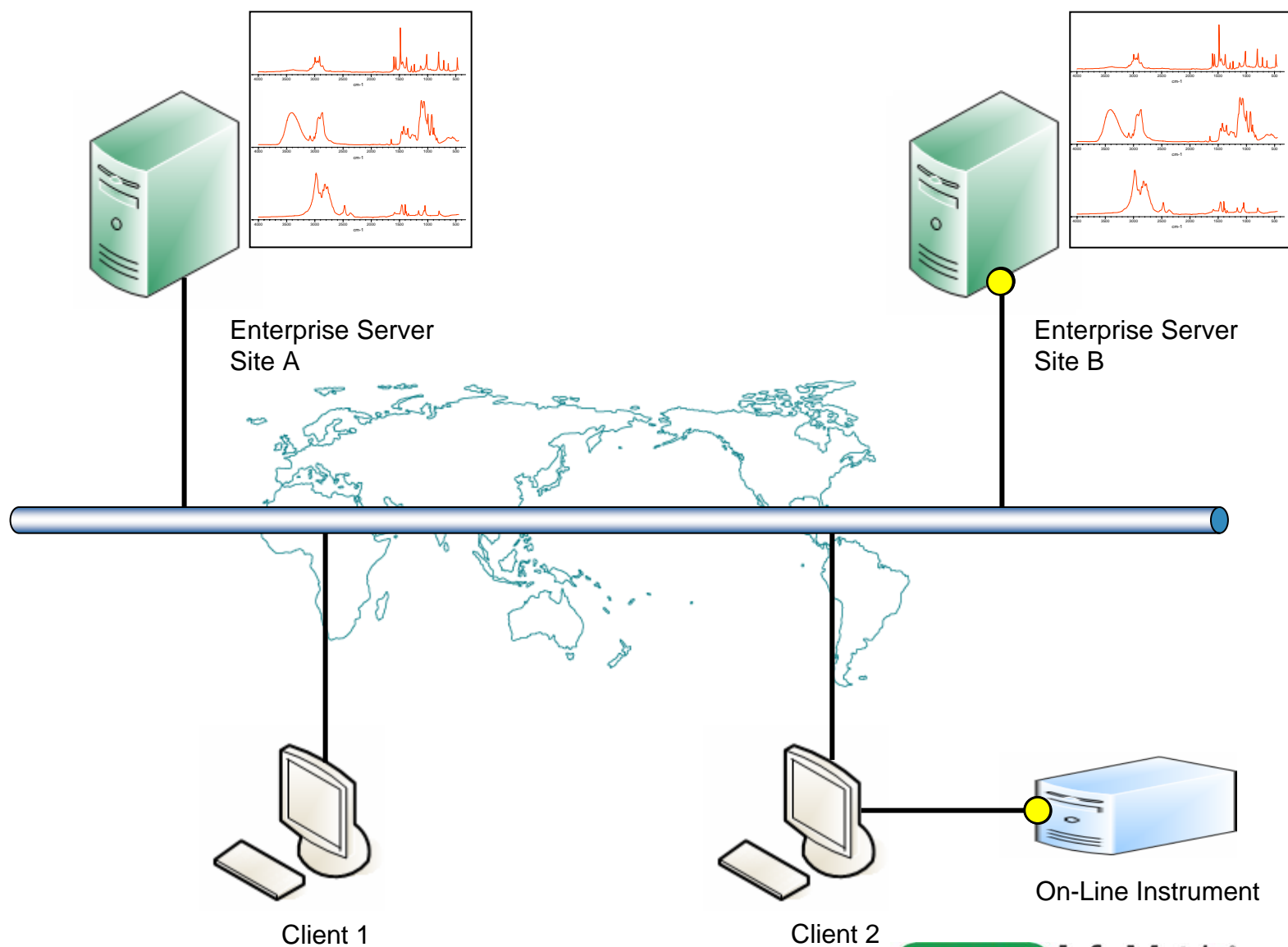
Consensus spectrum - OD Level = 33



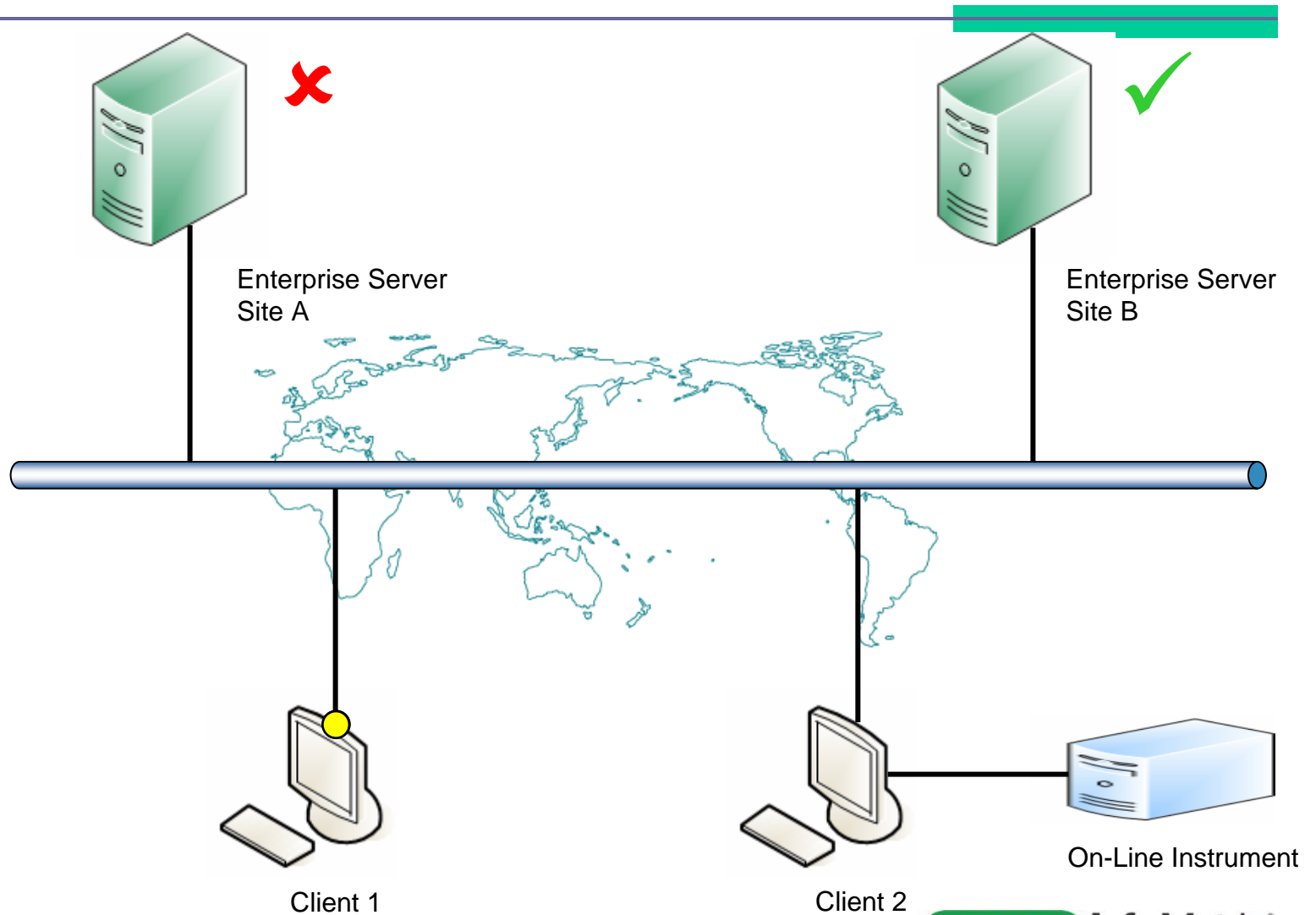
KnowItAll Spectral Library



Row-level data replication



System latency detection



Agenda

The Chemometrics Role

- Compress the data to fit better into a process historian structure

PCA

- Reduce the number of chromatograms or spectra required to manage a process effectively

Kennard & Stone

- Evaluate a potentially errant chromatogram to see if a similar trace has been seen in any plant at any time in the past

MV models, global database

- Execute a database query to identify the cause of off-target features.

Enterprise spectral servers

