Advanced Data Processing Requirements for Fast Chromatography
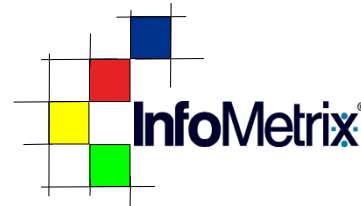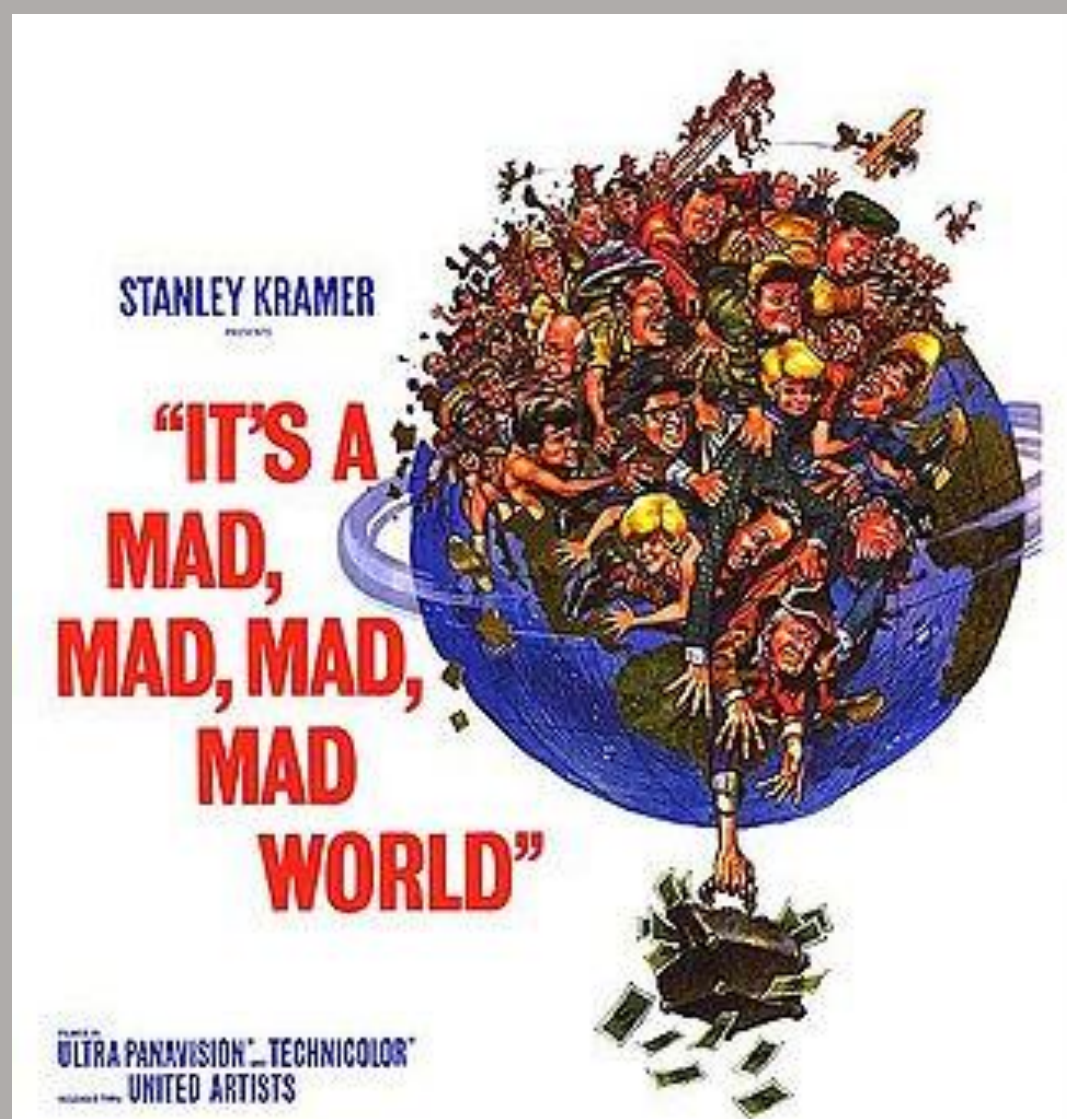
Automation, Automation, Automation

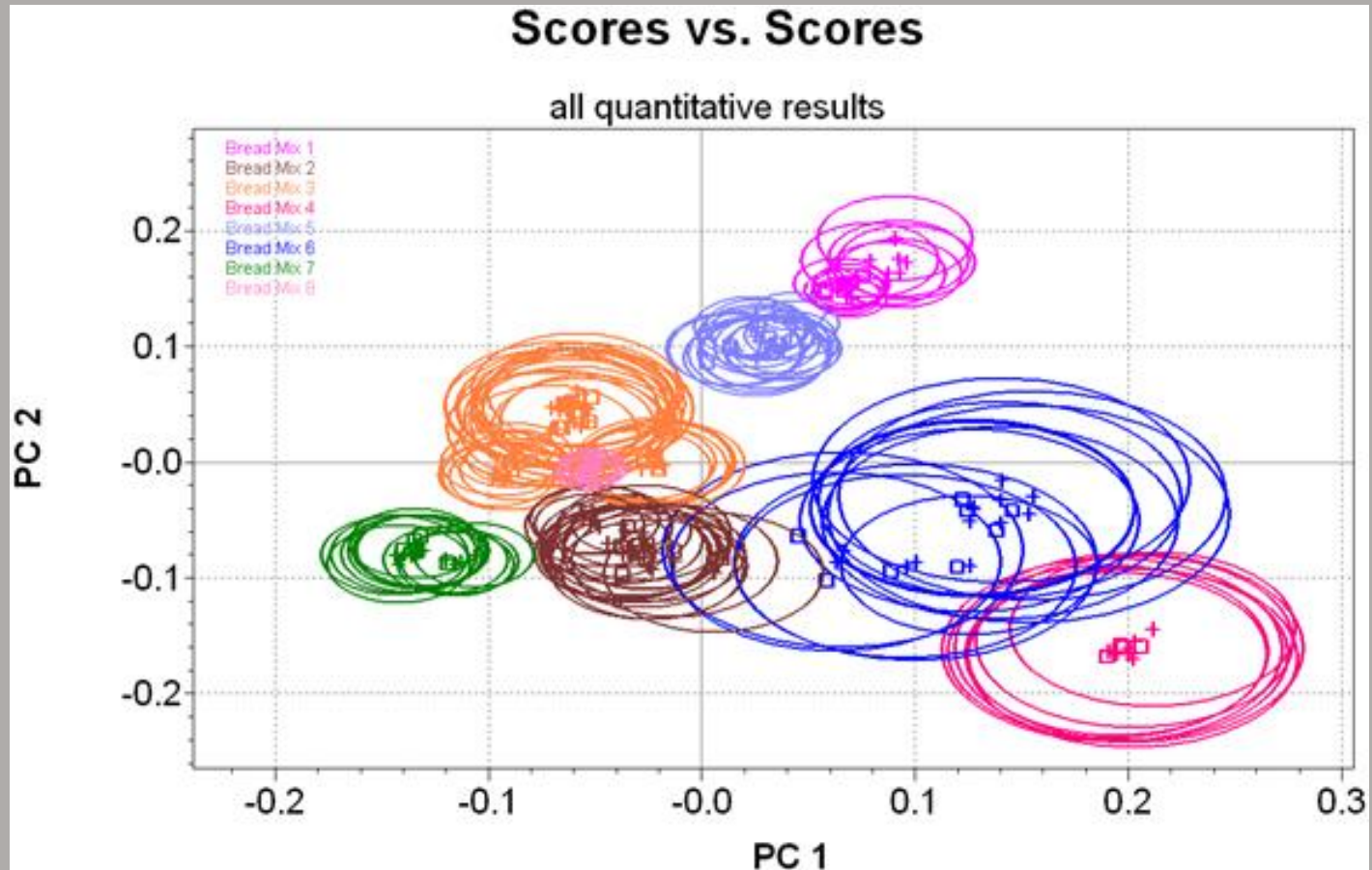**Brian Rohrback**

Gulf Coast Conference

October 16, 2013

Infometrix

As chromatography is brought into the sub-five-minute realm, we need to rethink the software that is used to collect, store and interpret the chromatographic traces.

Let's Borrow a Page from Spectroscopy's Playbook …

Infometrix

3

Two Ways to Use Chromatography

1. Quantitative Analysis
   - Provide a means of accurately quantitating a small number of compounds.
   - Predicting a physical property or system parameter
   - Unbundling a mixture
2. Qualitative Analysis
   - Evaluate a pattern of components to determine if the mixture is within specifications.

Infometrix

Another role for chemometrics

With the increase in speed, we need to automate the assessment of the chromatographic data such that samples behaving normally are accepted, but any problem is noted whether it be

- – a raw material input deficiency,
- – a process problem, or
- – an instrument problem

Infometrix

Why Chemometrics?

	X	Y
1	0.230	0.390
2	0.218	0.340
3	0.223	0.359
4	0.217	0.335
5	0.229	0.385
6	0.220	0.348
7	0.225	0.370
8	0.226	0.375
9	0.216	0.328
10	0.214	0.321
11	0.226	0.374
New	0.219	0.378
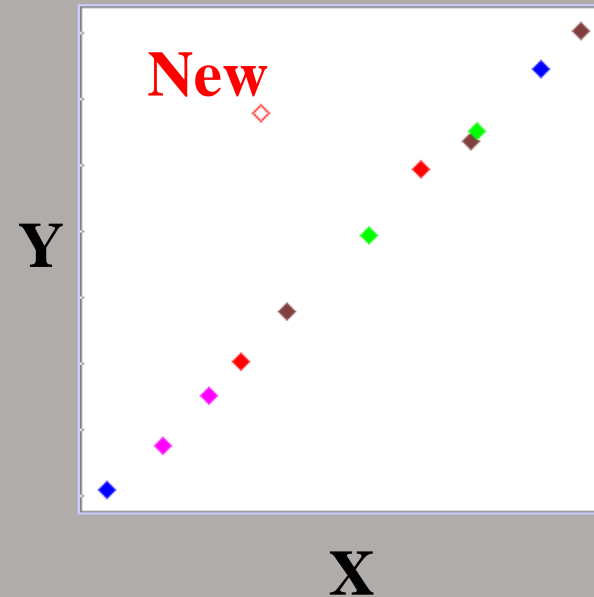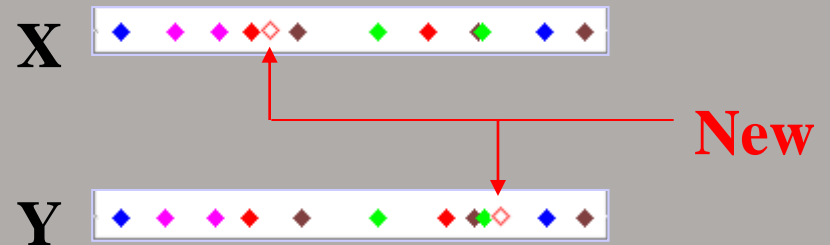
**Is this sample good?**

X

Y

New

New

Y

X

6

# Visualization – Why?

4 different data sets containing 2 measurements with apparently identical statistics

Anscombe's Quartet								
	Set I		Set II		Set III		Set IV	
	x	y	x	y	x	y	x	y
smp1	10	8.04	10	9.14	10	7.46	8	6.58
smp2	8	6.95	8	8.14	8	6.77	8	5.76
smp3	13	7.58	13	8.74	13	12.74	8	7.71
smp4	9	8.81	9	8.77	9	7.11	8	8.84
smp5	11	8.33	11	9.26	11	7.81	8	8.47
smp6	14	9.96	14	8.1	14	8.84	8	7.04
smp7	6	7.24	6	6.13	6	6.08	8	5.25
smp8	4	4.26	4	3.1	4	5.39	19	12.5
smp9	12	10.84	12	9.13	12	8.15	8	5.56
smp10	7	4.82	7	7.26	7	6.42	8	7.91
smp11	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
CorrCoef	0.82		0.82		0.82		0.82	
Line	slope	intercept	slope	intercept	slope	intercept	slope	intercept
	0.50	3.00	0.50	3.00	0.50	3.00	0.50	3.00

(F.J. Anscombe, "Graphs in Statistical Analysis," American Statistician, 27 [February 1973], 17-21)

Infometrix

# Anscombe's Quartet
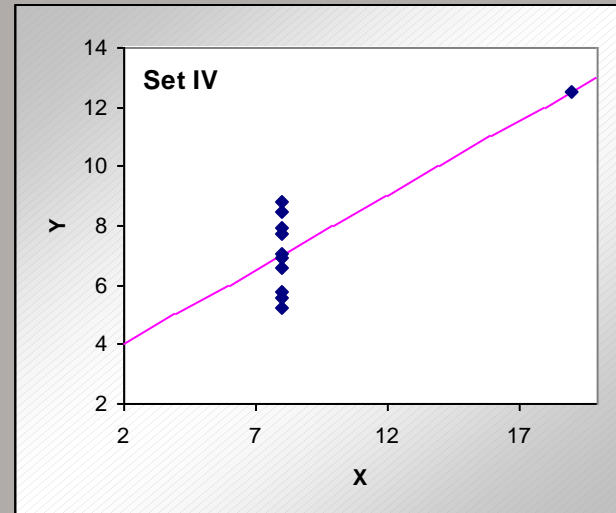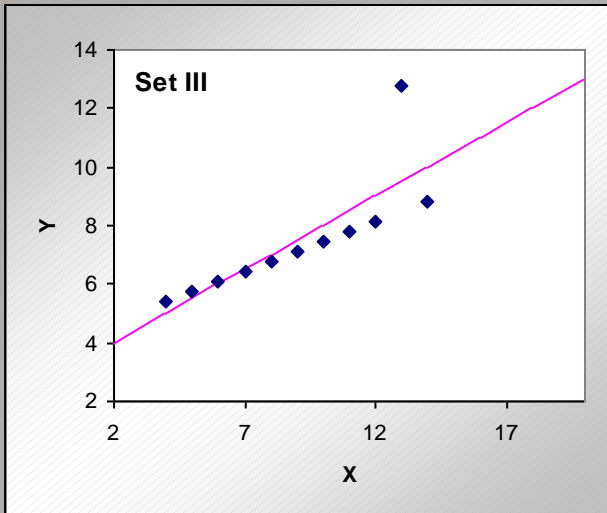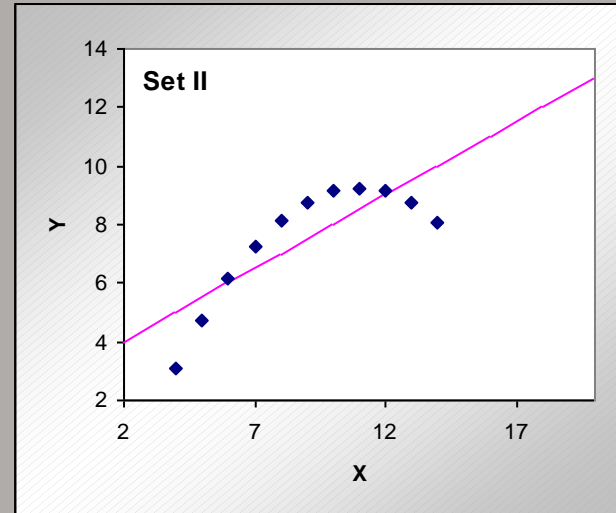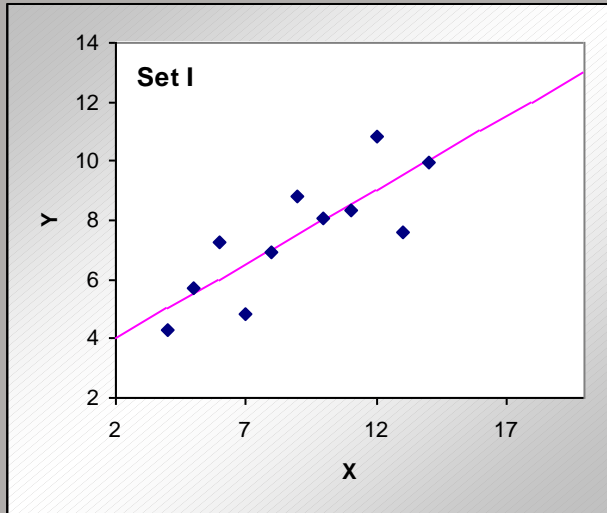
Delivering Information

Just having the measurements does not translate into control

- There are not enough skilled technicians to handle even the current workload.

- Chemometrics solves the information processing problem with two technologies:
    - Alignment enables us to sell instruments that have vastly-lower calibration requirements.
    - Interpretation algorithms automates the generation and the qualification of the information derived from the raw data.
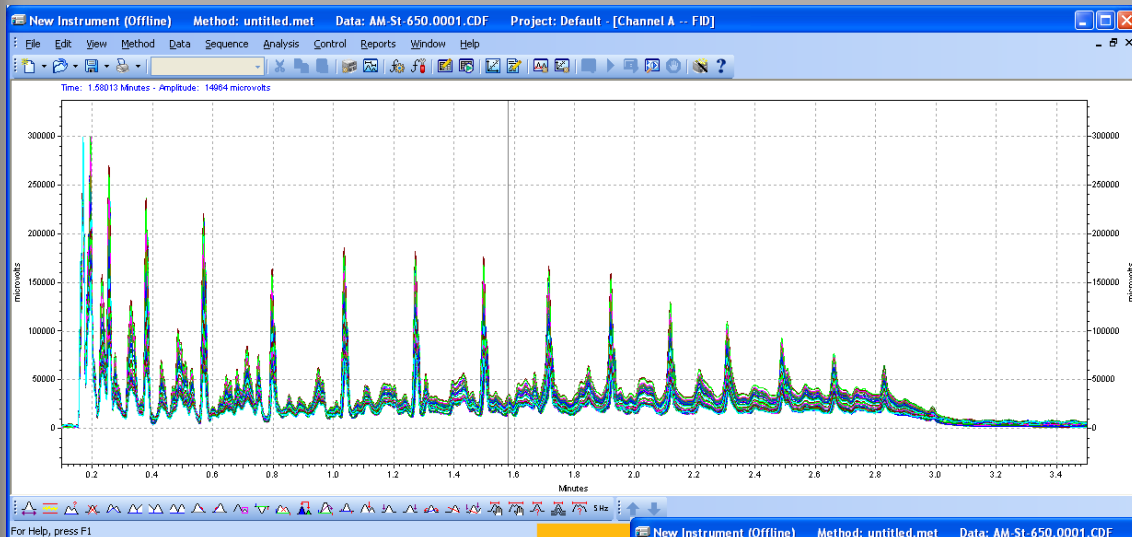
And if we can make all of our instruments look as much alike as possible.

Interchangeability
Common interpretive base

Gas Chromatography: 2 Instruments
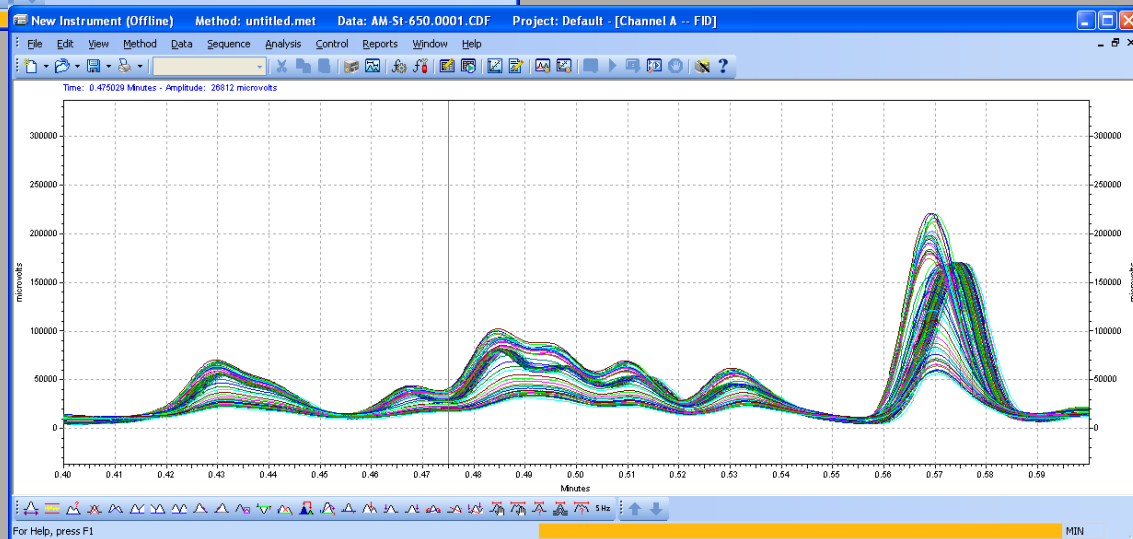
Raw

Aligned
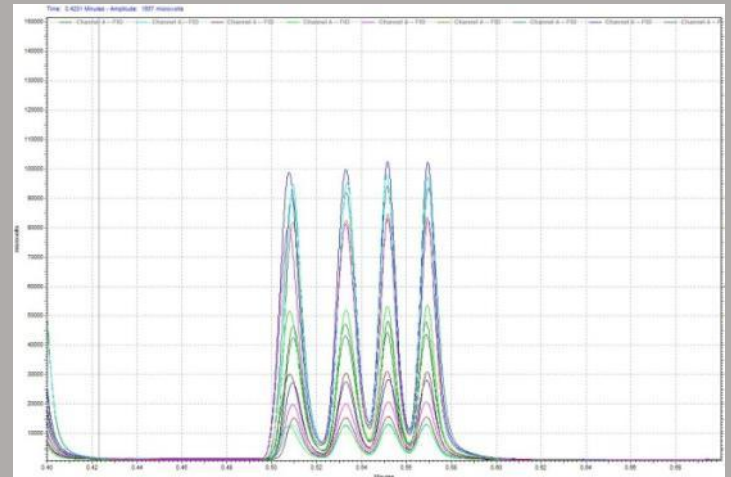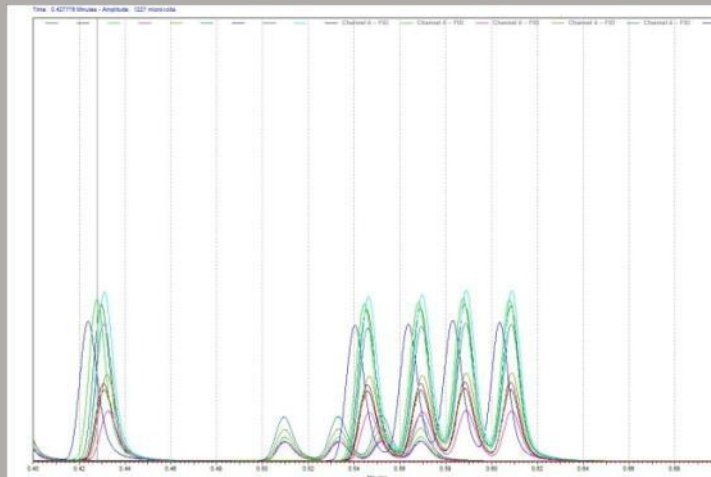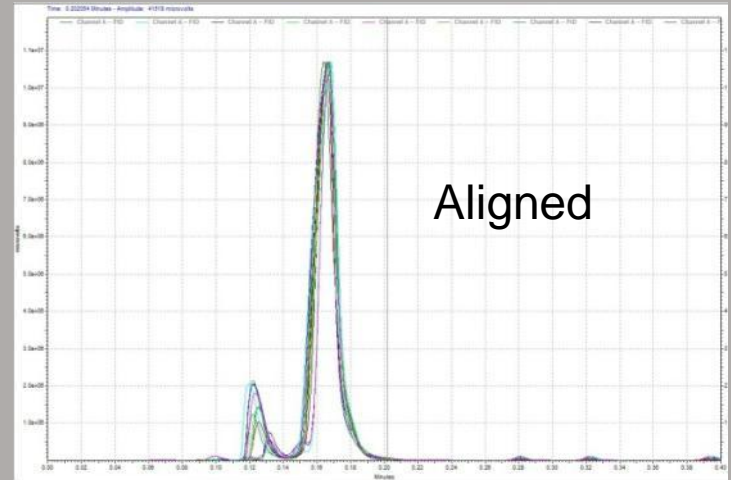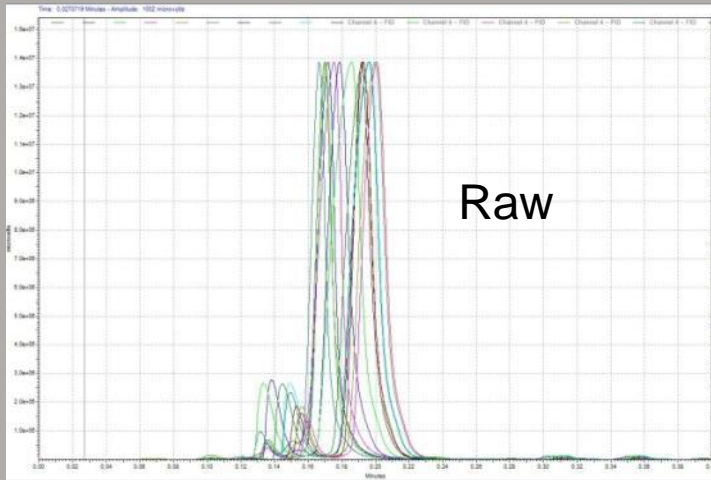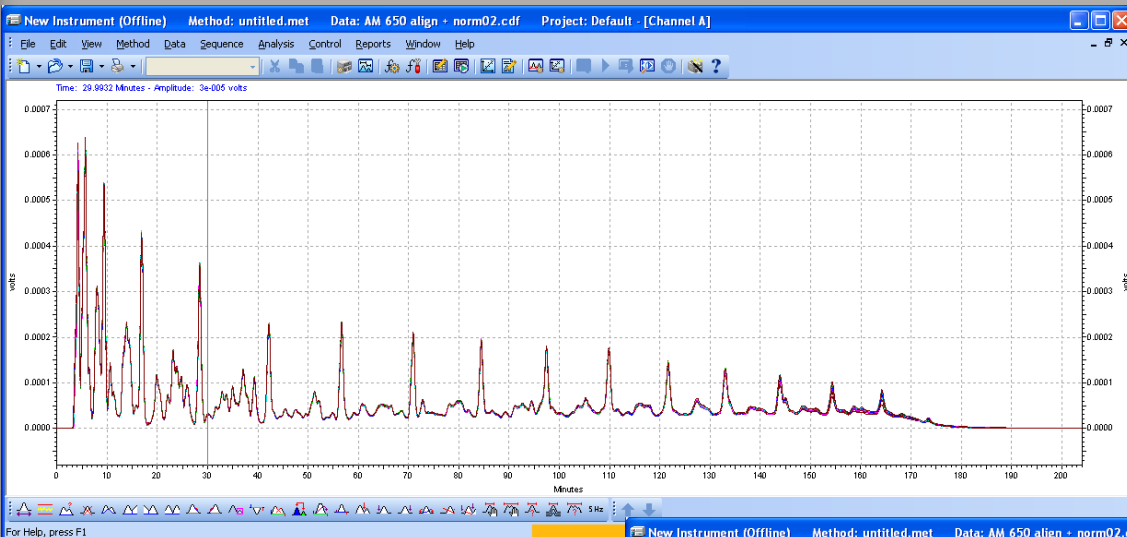
Repeatability of a Process-Based Micro GC

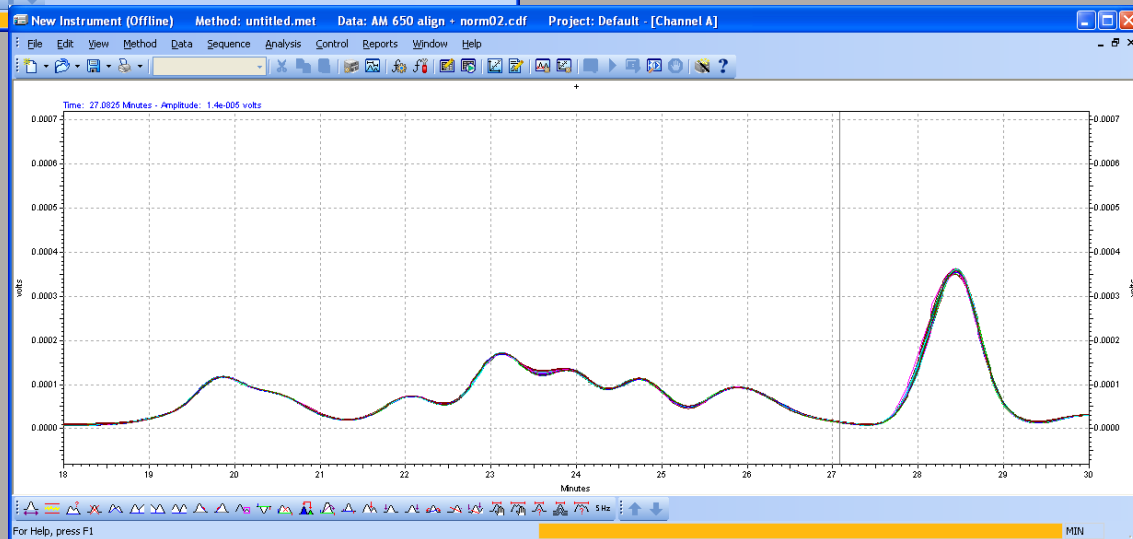Start to 650°F Sample of a Conventional Crude Oil

Above: 3.5 min run

Right: expansion from 0.4 to 0.6 min.

Infometrix

Repeatability of a Process-Based Micro GC

Substantial Reduction in Sample to Sample Variability

Above: 3.5 min run

Right: expansion from 0.4 to 0.6 min.

Infometrix

12

Two Approaches

There are two ways of handling chromatographic data which can be done separately or in tandem:

1. Peak Tables

   Peak tables are simple and small and contain concentration information about the primary components in the mixture. They miss unexpected peaks.

2. Raw Chromatograms

   Treating the chromatogram as if it were a spectrum means that both the expected and the unexpected are covered. This approach is sensitive to variations in retention time and forces the system to deal with 100 to 1000 times as much data.
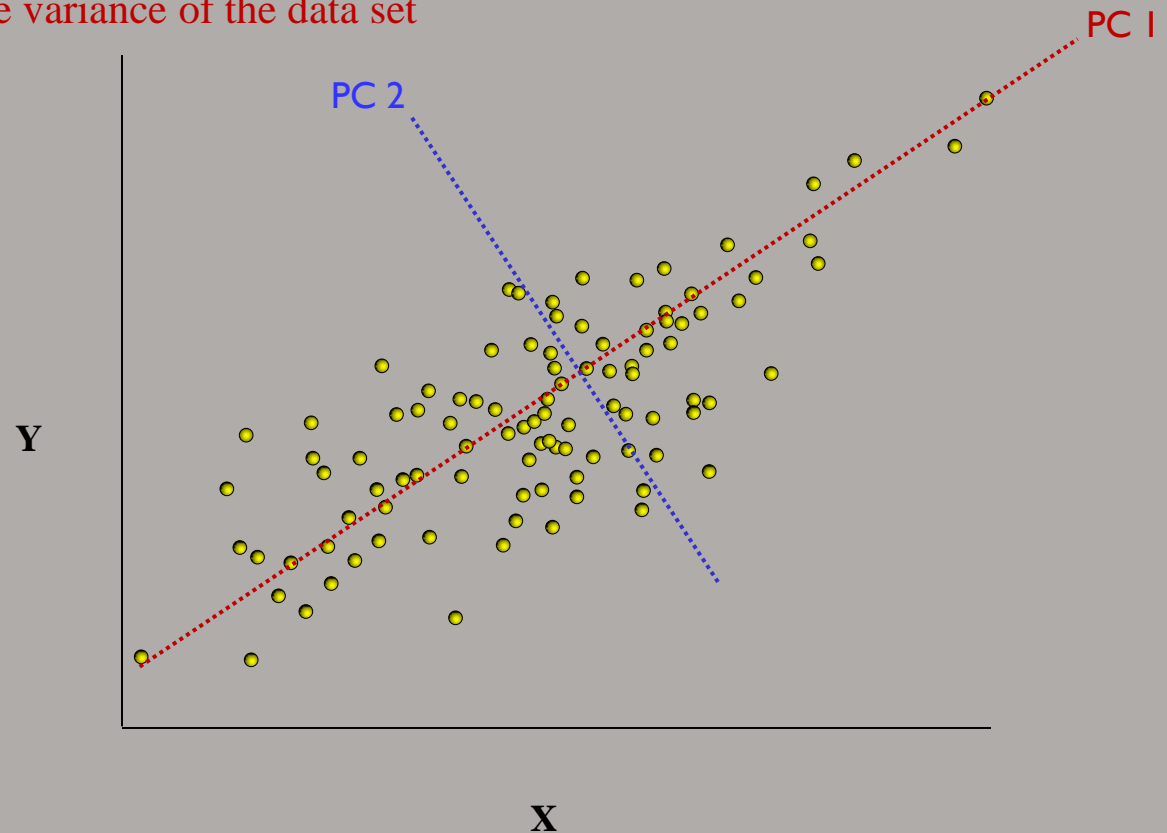
3 points represent 3 chromatograms
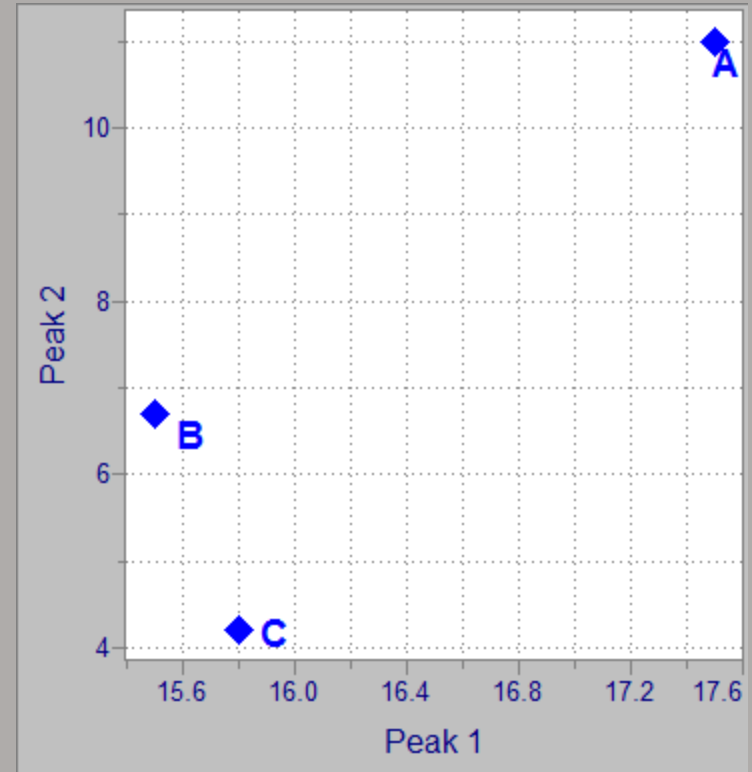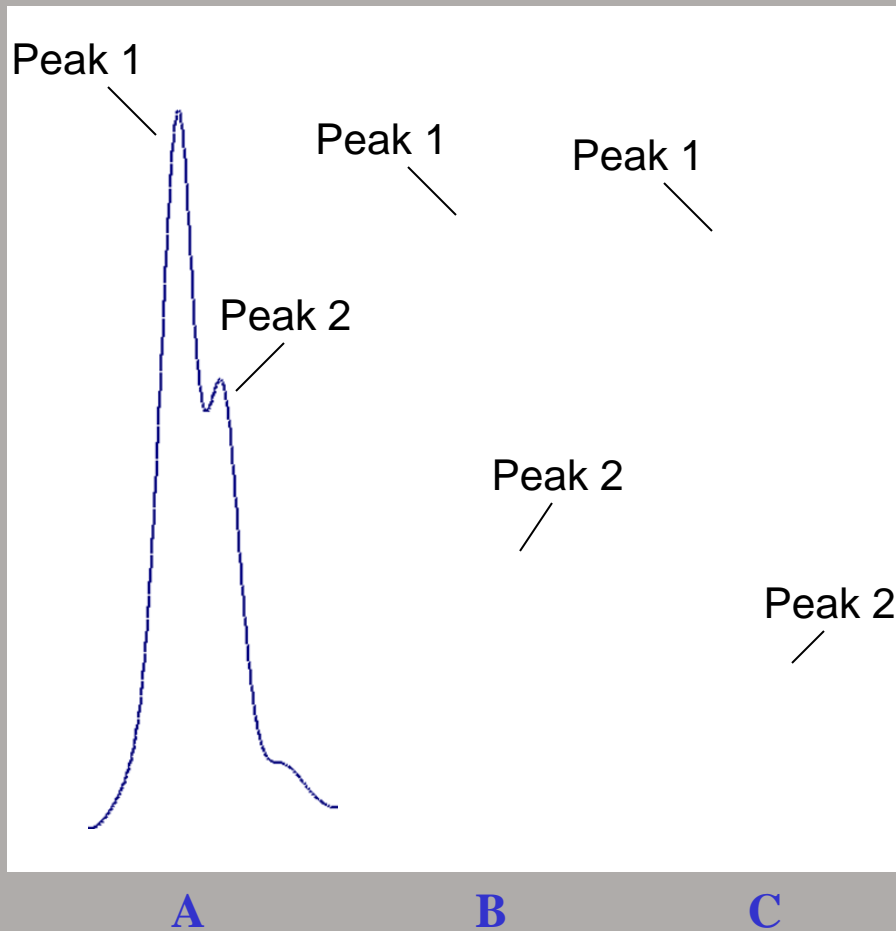
PCA method

First Principal Component (or Factor)

• Describes most of the variance of the data set

PC 1

PC 2

Y

X

Second Principal Component (or Factor)
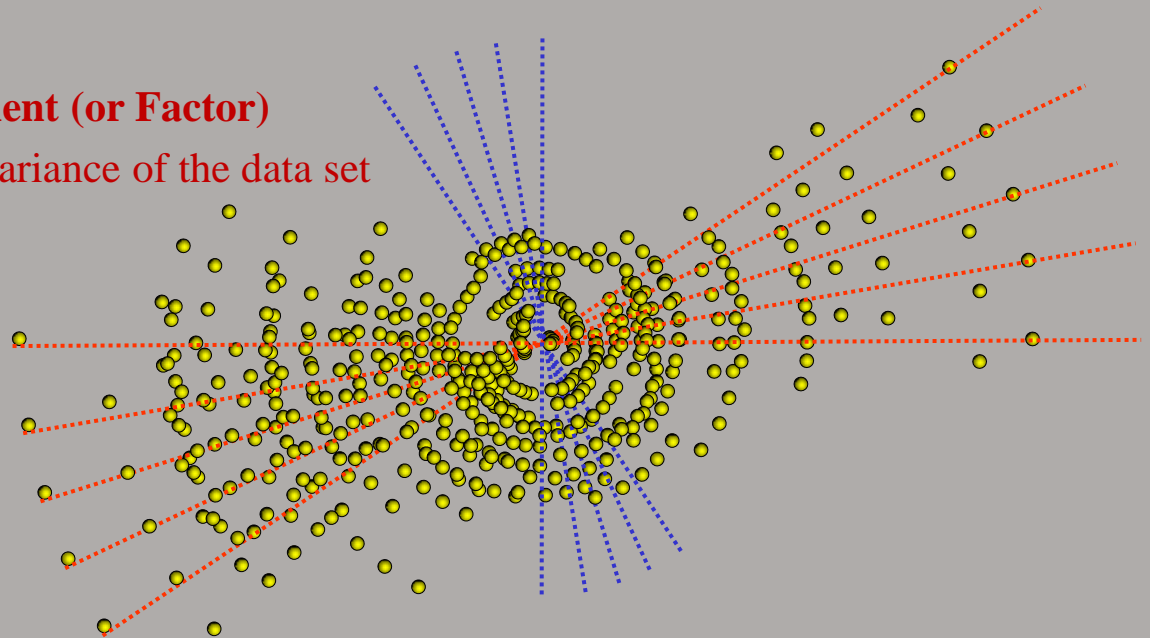
• Orthogonal to the First Principal Component

• Describes more of the variance of the data set not described by the First PC

Infometrix

15

PCA method

Rotation
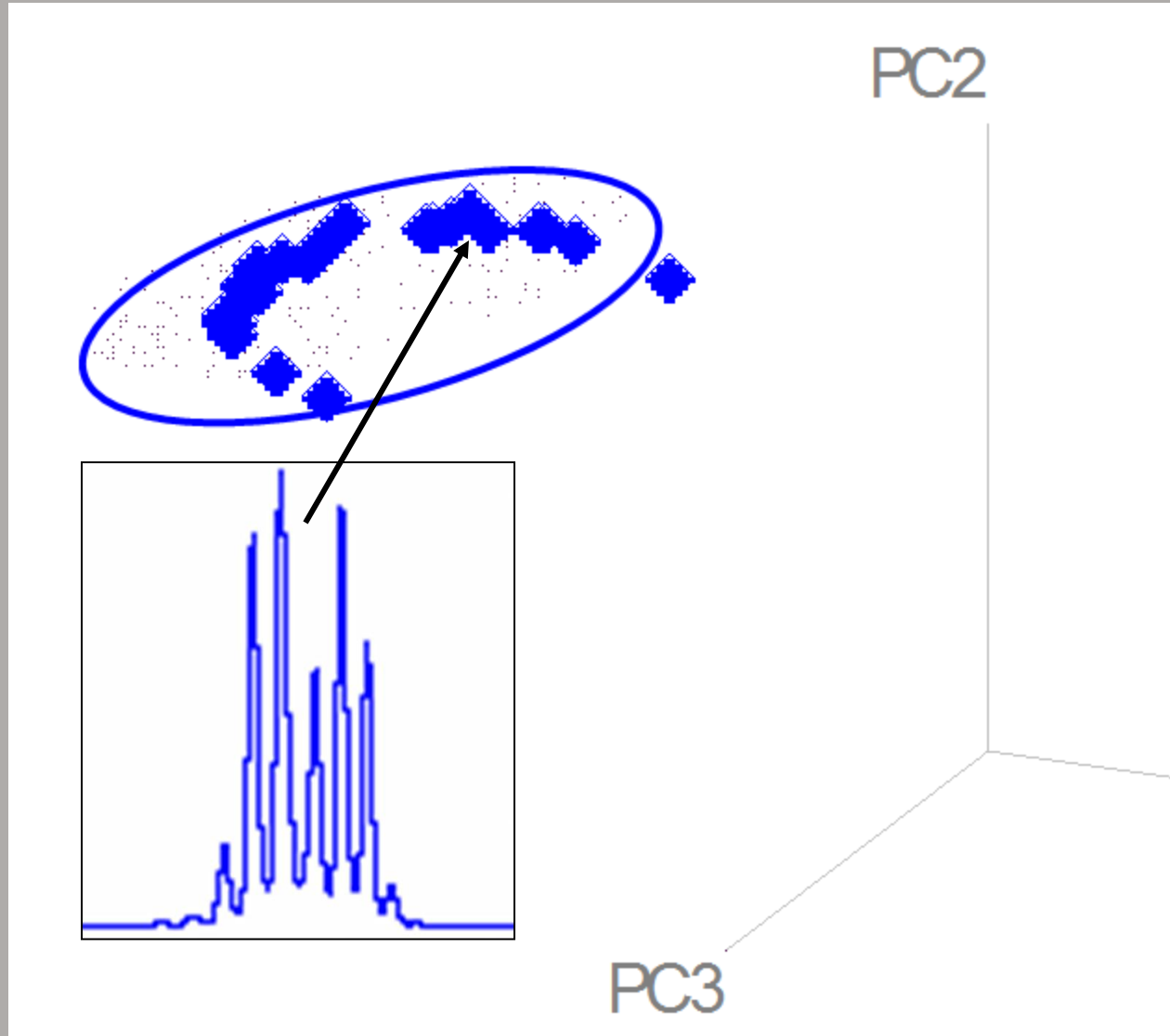
First Principal Component (or Factor)

• Describes most of the variance of the data set

Second Principal Component (or Factor)

• Orthogonal to the First Principal Component

• Describes more of the variance of the data set not described by the First PC

Infometrix

A chromatogram is a point in PCA space
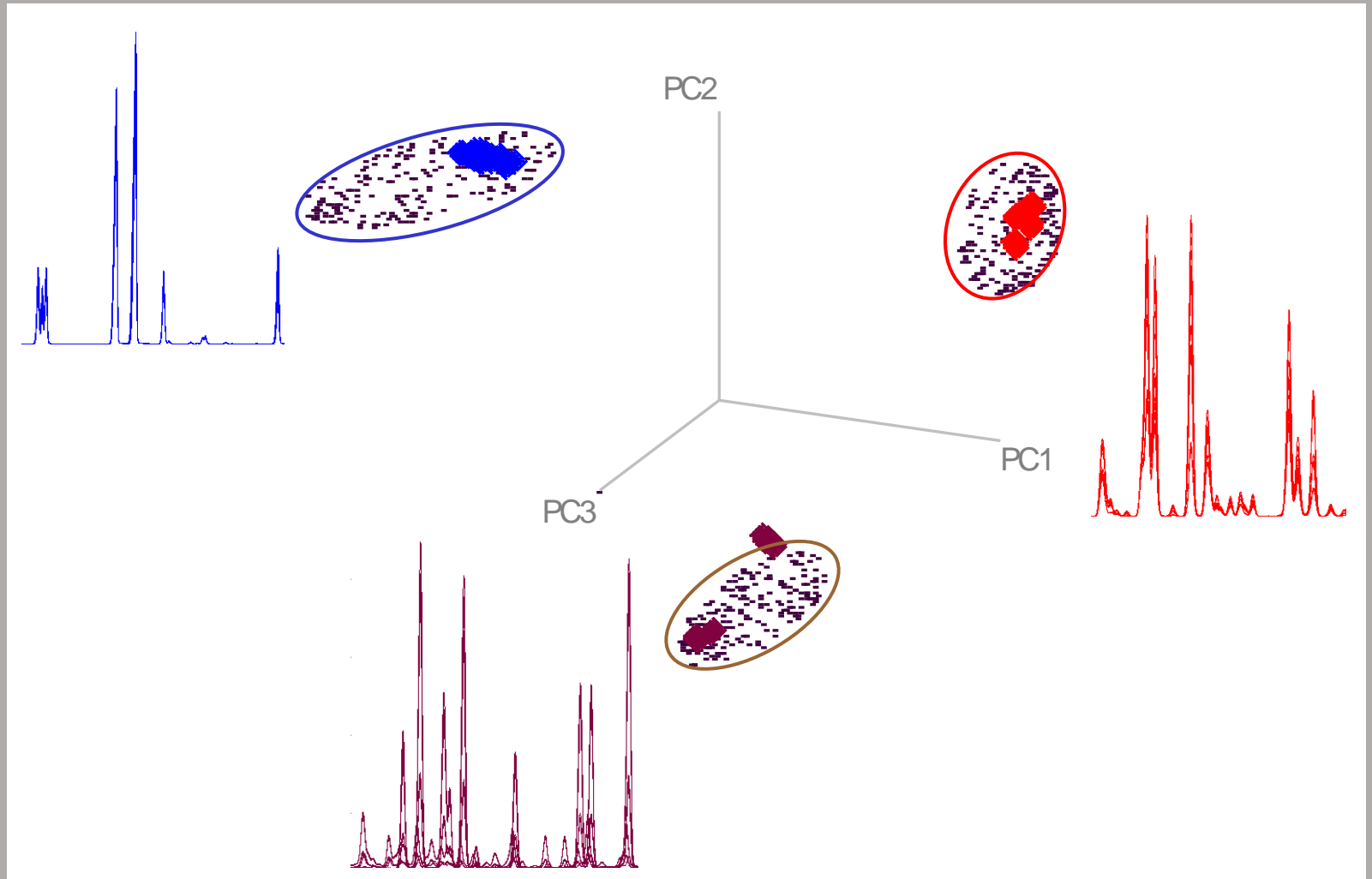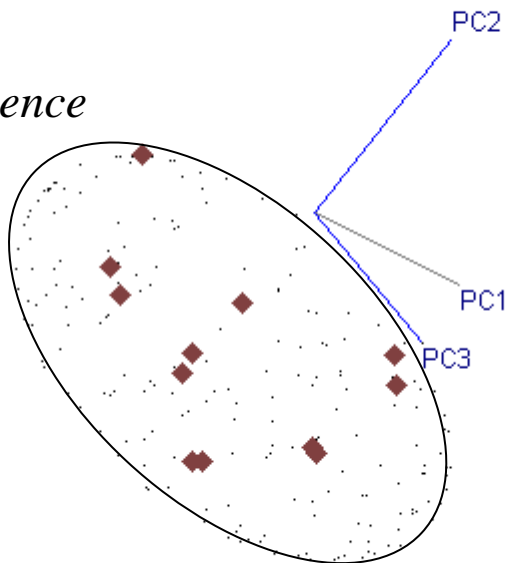
A basis for interpretation

Table example: DHA summary report

Composite report				Winter gasoline			
Total by group type & carbon number							
(in volume percent)							
	n-Paraffins:	i-Paraffins:	Olefins:	Naphtenes:	Aromatics:	Oxygenates	Total:
C1:	0	0	0	0	0	0	0
C2:	0	0	0	0	0	0	0
C3:	0.424	0	0	0	0	0	0.424
C4:	11.108	4.513	1.1	0	0	0	16.721
C5:	0.427	2.867	0.928	0.029	0	0	4.251
C6:	0.197	2.112	0.467	0.348	0.198	0	3.323
C7:	1.054	14.459	17.395	0.411	10.5	0	43.819
C8:	0.404	3.686	0	0.827	10.715	0	15.631
C9:	0	0.938	0.052	1.667	6.04	0	8.698
C10:	0.141	1.915	0	0.208	2.24	0	4.505
C11:	0.024	0.951	0	0	0.125	0	1.101
C12:	0.035	0.068	0.025	0.026	0	0	0.155
C13:	0	0	0	0	0	0	0
C14:	0	0	0	0	0	0	0
Total:	13.816	31.509	19.966	3.517	29.818	0	98.626
Total C14+:	0						
Total unknowns:	1.374						
Grand total	100						

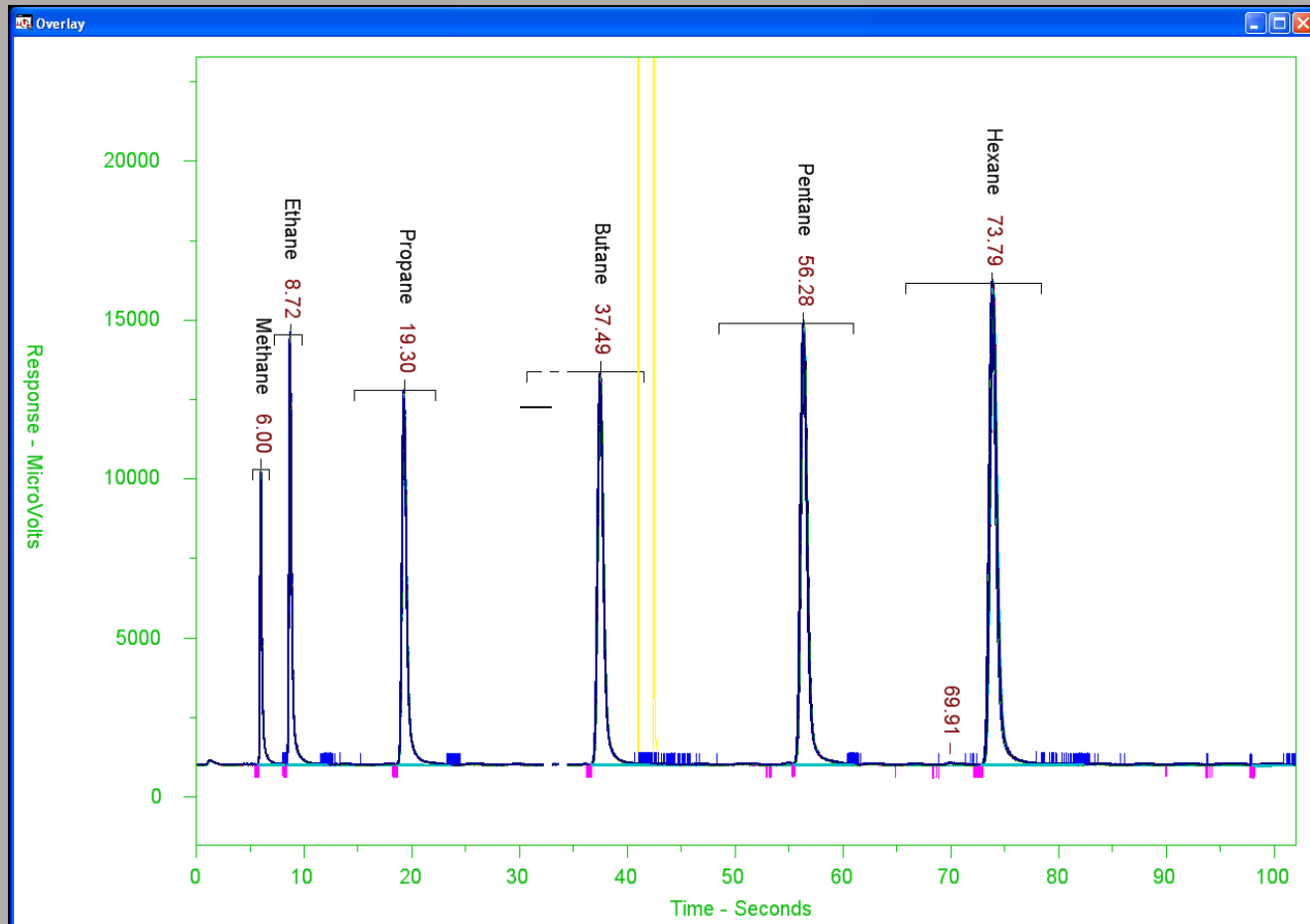Infometrix

Build a classification model

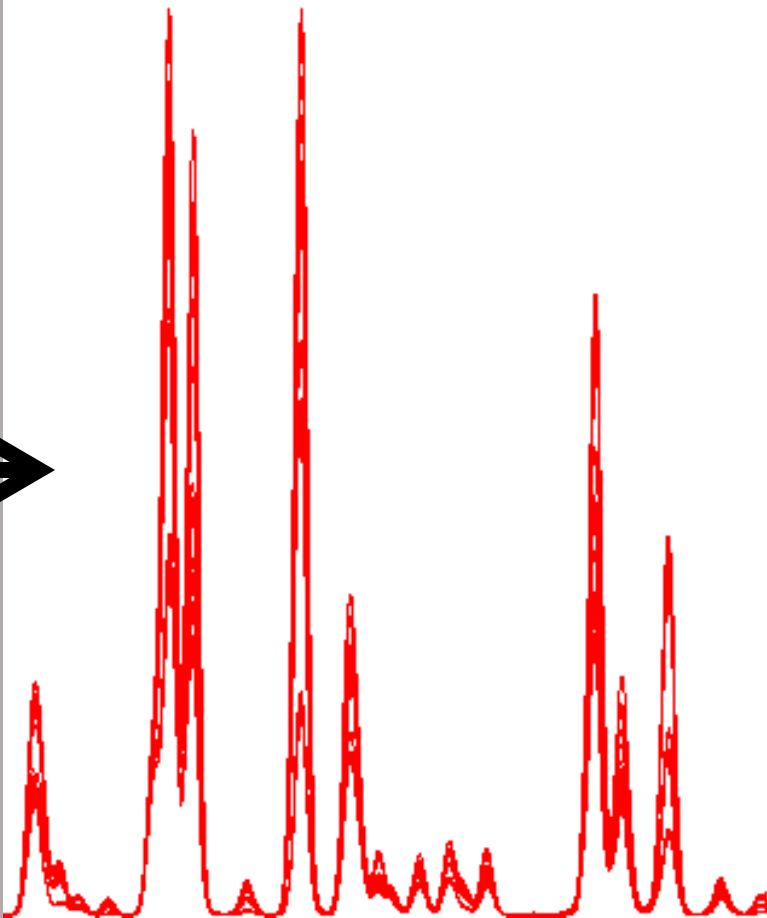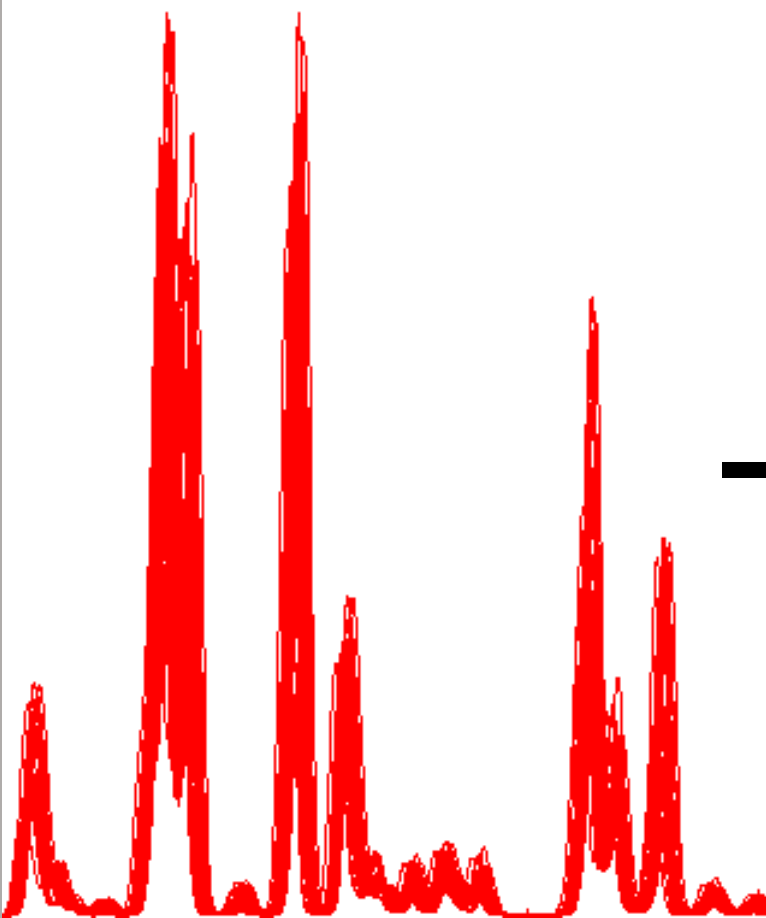# Identify cause for outlier groupings and trends
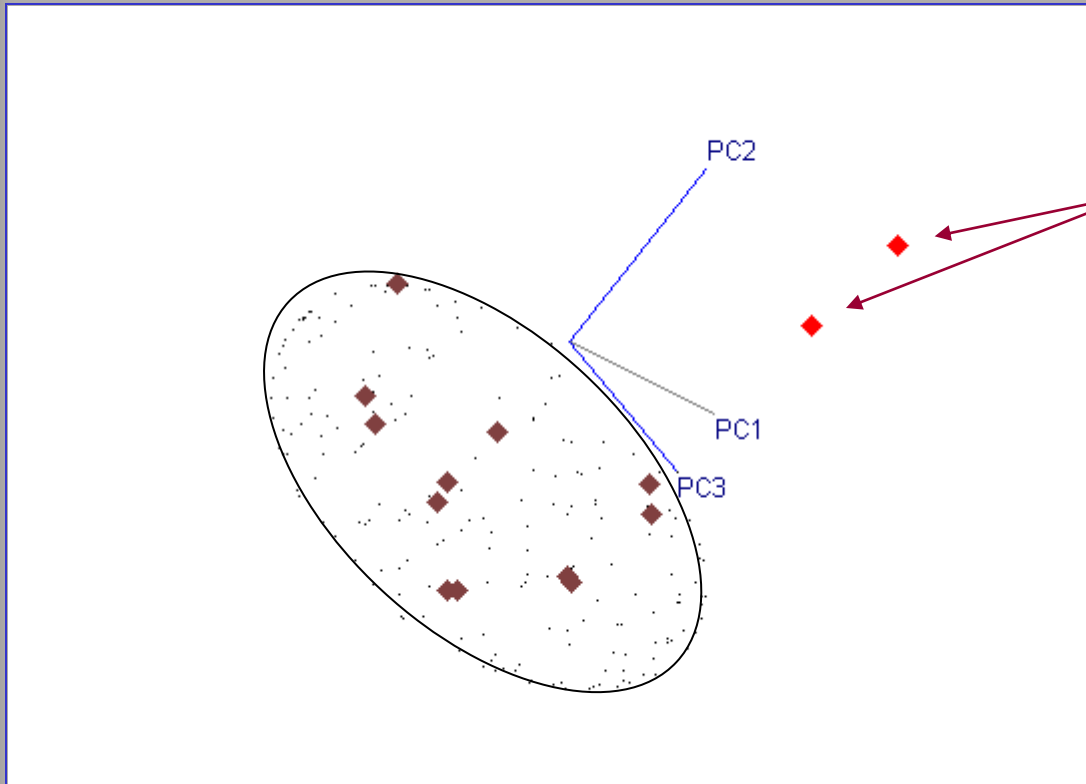
Outliers, because of:

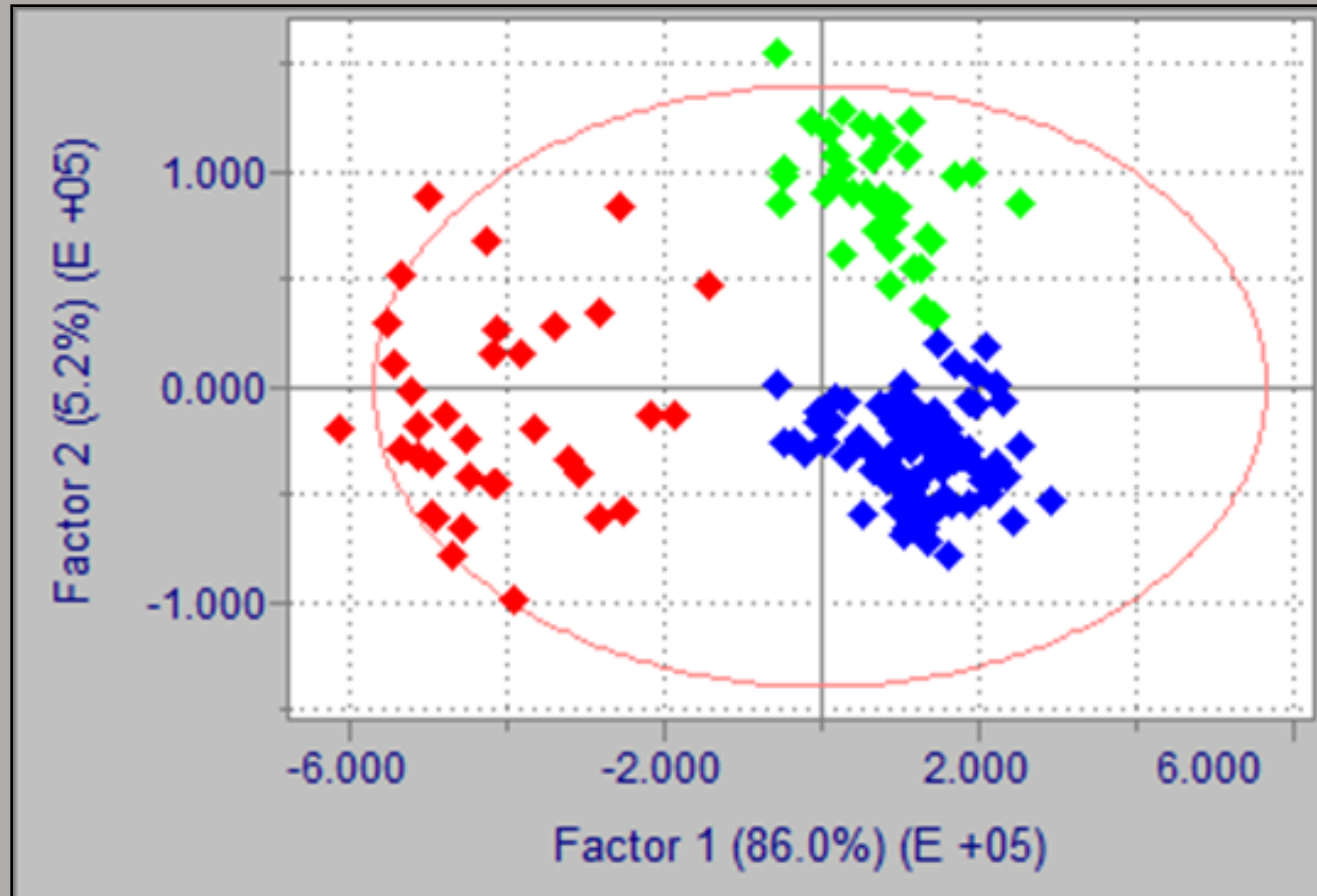- Instrument problem?
- Process upset?

Consistent GC data

22

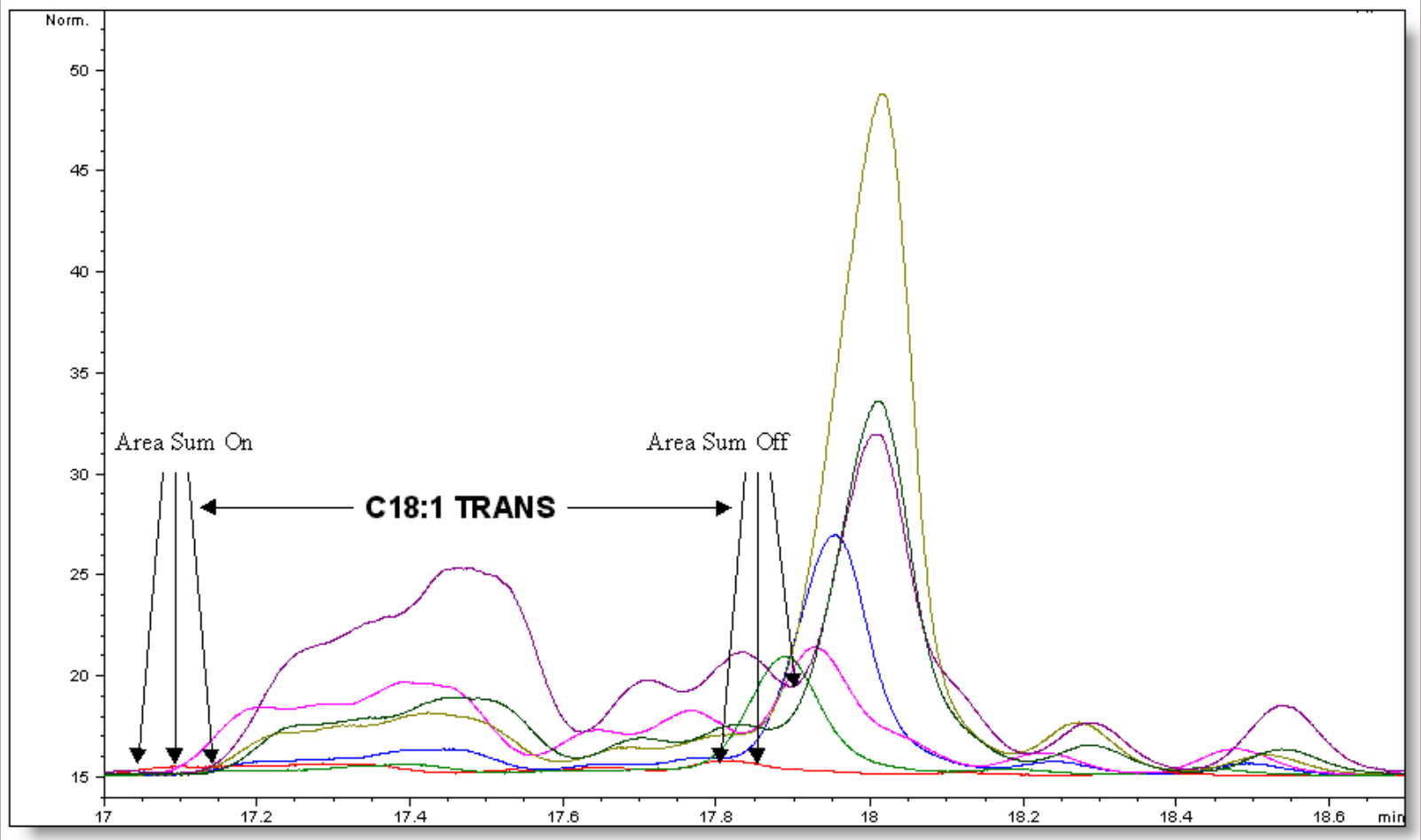# Eliminate residual misalignment

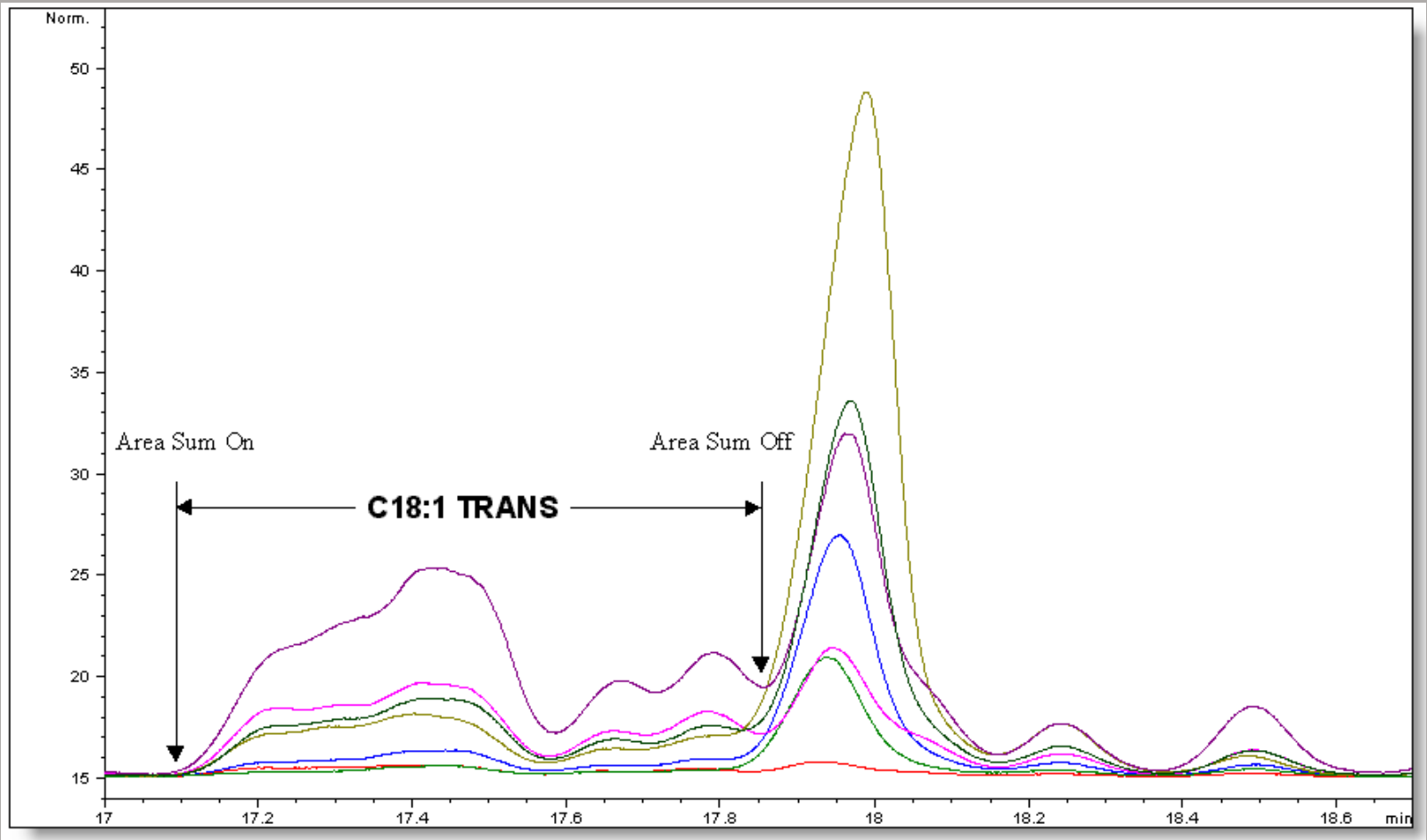# Build a classification model using aligned chromatograms

We selected 160 representative good-quality chromatograms to make a model from the 3600 chromatograms supplied.

Gating Problem

Gating Problem Solved

Complex Mixtures

- Pulp mill using hog-fuel boiler
 - Operated for 60+ years
  - Emission controls upgraded several times
  - Odor complaints history
  - A state clean-up site—status: in process
- Many prior analyses
  - Prior data interpretation not multivariate
  - Different levels of quality in GC work
  - Multiple agencies involved
- Current study objectives
  - Characterize dioxin distribution
  - Apportion sources

Infometrix

# MCR-ALS Source Contributions

- Majority of samples dominated by Source 3 pattern

- Four samples related to pattern of Source 2

- Several samples show contribution from a Source 1 pattern
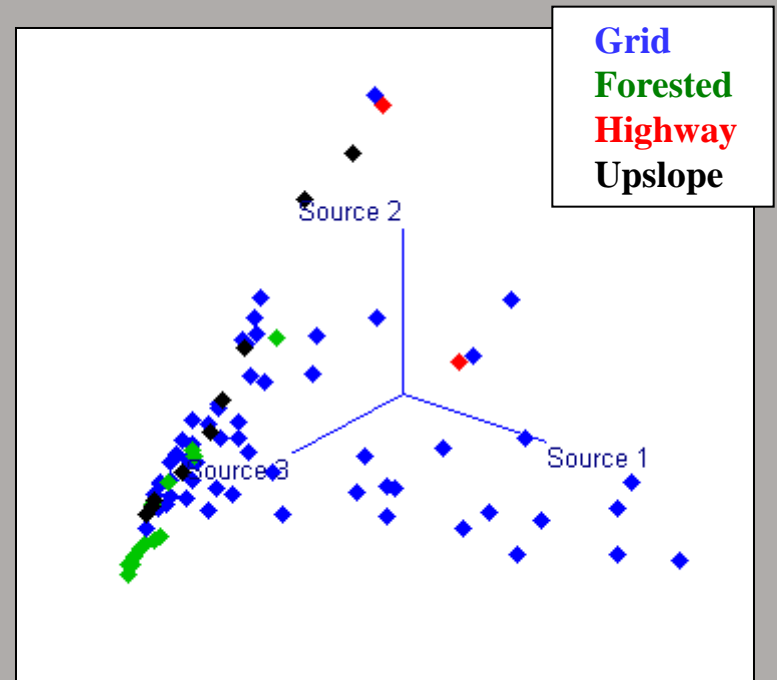
Spatial Distribution of Source Contributions

- Contributions of each proposed source pattern to study locations

- Magnitude of source contribution proportional to color intensity

Boiler Site

Source 3
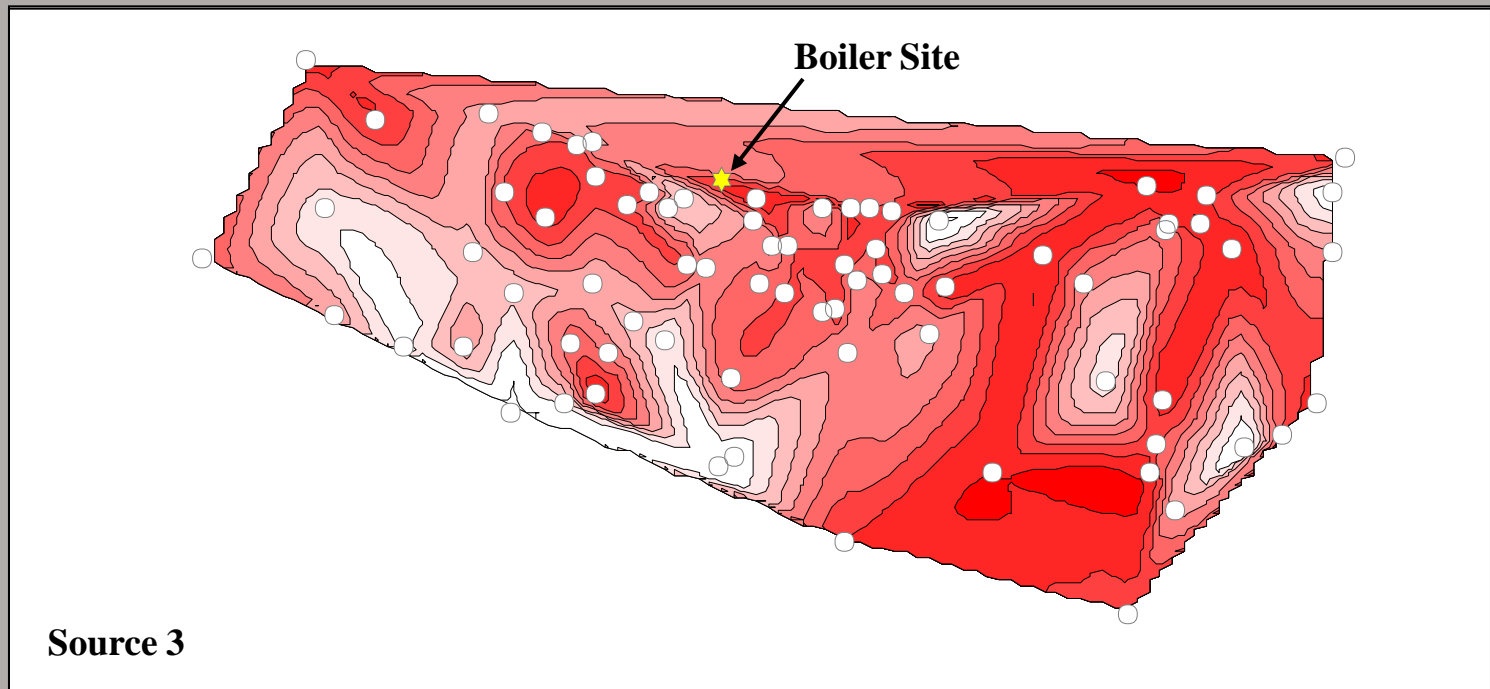
Continuous validation of a multivariate instrument

We can correct retention times to match an application-specific relevant sample

You can use this to make all instruments performing a similar task to look identical (Plug and Play)

Common regression and classification algorithms can be applied automatically to infer physical properties or characteristics

This allows us to bring more complex analyses into on-line use

Chemometrics for instrumentation: the value proposition

Anything you can do to improve precision of the multivariate measurements collected by the instrument will allow you to tighten the control – essentially for free.

We use the signal processing aspect of chemometrics to reduce instrument-derived variability

Within an instrument (e.g., noise reduction)

Between instruments (i.e., transfer of calibration)

This creates the ability to construct an application-specific, objective evaluation system

Infometrix