PXRD with RAMAN SPECTROSCOPY, DSC and IR DATA

Chris Gilmore, Gordon Barr and Gordon Cunningham

Classifying Data

- Statistics/data mining problem: put patterns (PXRD, spectra) into clusters where each cluster contains patterns which are most similar to each other.
- Not always a unique solution.
- Problems with:
    - Data quality.
    - Sample quality.
    - Data quantity.
    - Need for automation, and speed.

Correlation

- Forget peaks.
- Match *every* data point.
- Use correlation coefficients:
  - Pearson correlation coefficient (parametric).
  - Spearman correlation coefficient (non-parametric).

Pearson correlation coefficient

$$r_{Pearson} = \frac{\sum_{\text{all points}} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{\text{all points}} \left(x_i - \overline{x}\right)^2}\sqrt{\sum_{\text{all points}} \left(y_i - \overline{y}\right)^2}}$$

Spearman correlation coefficient

$$r_{Spearman} = \frac{\sum\limits_{\text{all points}} \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum\limits_{\text{all points}} \left(R_i - \overline{R}\right)^2}\sqrt{\sum\limits_{\text{all points}} \left(S_i - \overline{S}\right)^2}}$$

Ranks

Combine them

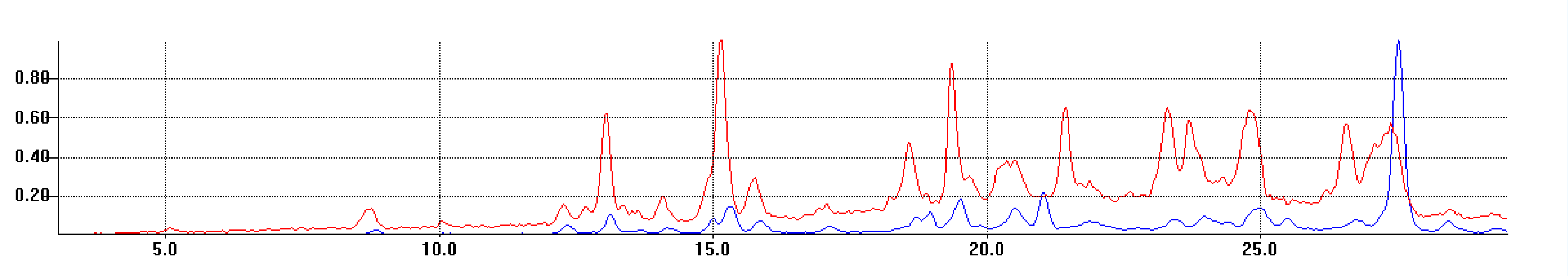
~~$r = w_1 r_{Pearson} + w_2 r_{Spearman}$~~

Use Fischer transforms:

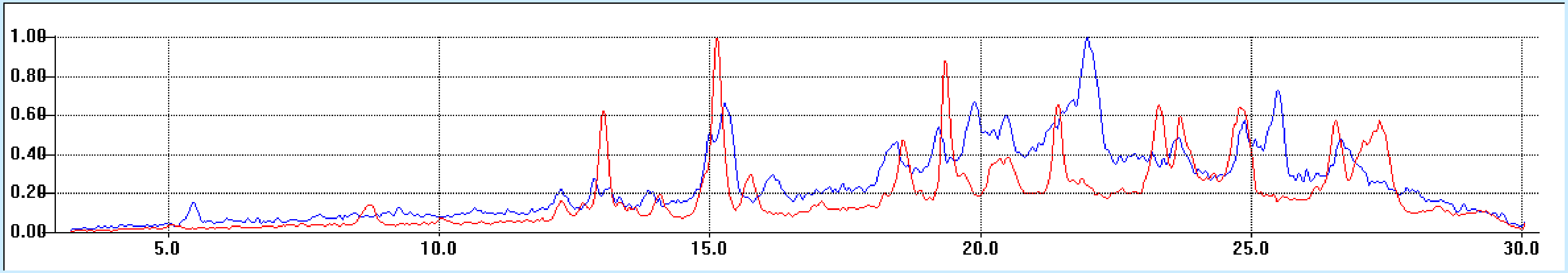
$$r = \tanh\left(\frac{\tanh^{-1}\left(r_{Pearson}\right) + \tanh^{-1}\left(r_{Spearman}\right)}{2} \right)$$

Correlation

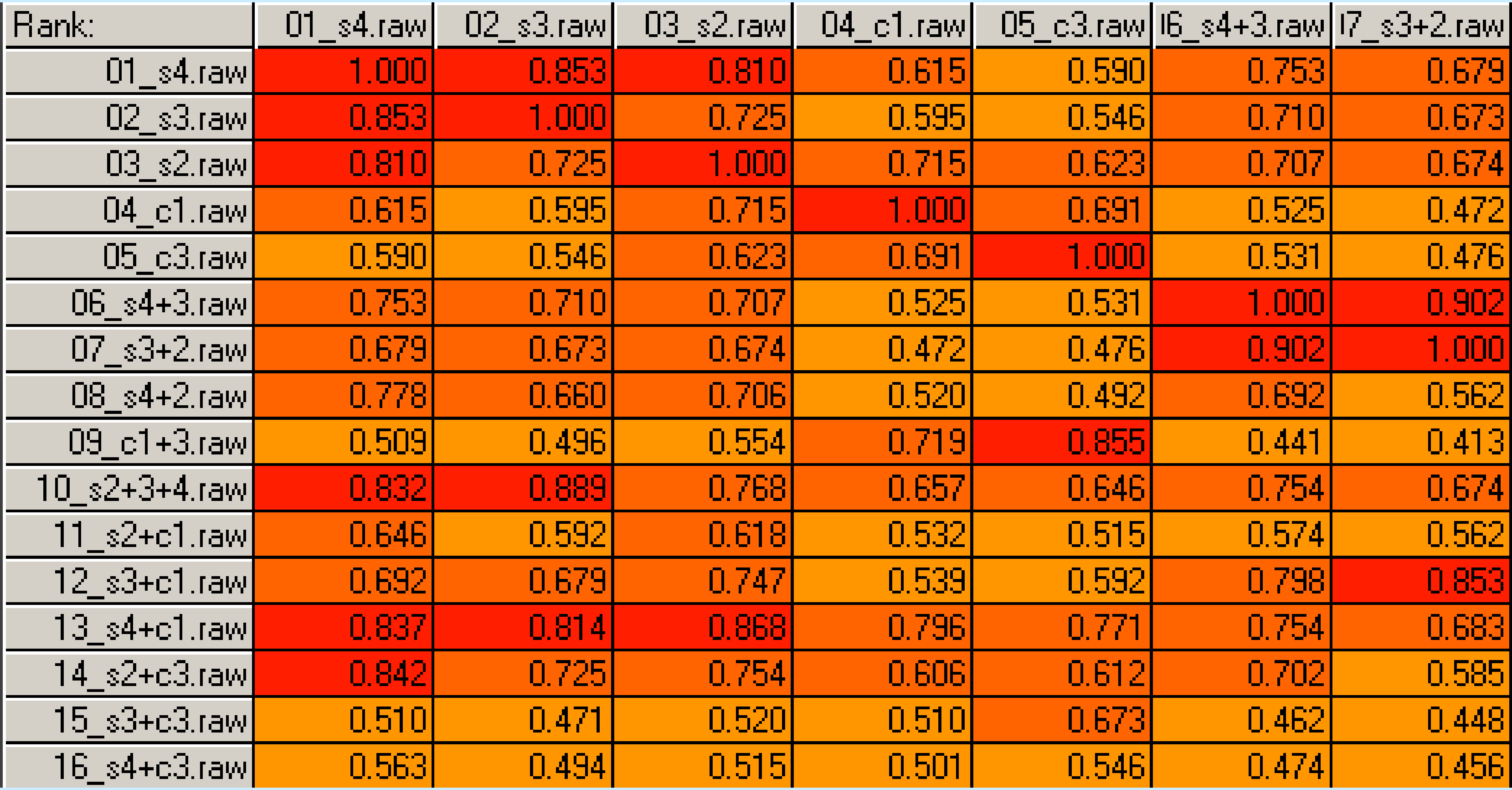
Mean correlation coefficient = 0.673



Mean correlation coefficient = 0.771



Correlation matrix



Rank:	01_s4.raw	02_s3.raw	03_s2.raw	04_c1.raw	05_c3.raw	I6_s4+3.raw	I7_s3+2.raw
01_s4.raw	1.000	0.853	0.810	0.615	0.590	0.753	0.679
02_s3.raw	0.853	1.000	0.725	0.595	0.546	0.710	0.673
03_s2.raw	0.810	0.725	1.000	0.715	0.623	0.707	0.674
04_c1.raw	0.615	0.595	0.715	1.000	0.691	0.525	0.472
05_c3.raw	0.590	0.546	0.623	0.691	1.000	0.531	0.476
06_s4+3.raw	0.753	0.710	0.707	0.525	0.531	1.000	0.902
07_s3+2.raw	0.679	0.673	0.674	0.472	0.476	0.902	1.000
08_s4+2.raw	0.778	0.660	0.706	0.520	0.492	0.692	0.562
09_c1+3.raw	0.509	0.496	0.554	0.719	0.855	0.441	0.413
10_s2+3+4.raw	0.832	0.889	0.768	0.657	0.646	0.754	0.674
11_s2+c1.raw	0.646	0.592	0.618	0.532	0.515	0.574	0.562
12_s3+c1.raw	0.692	0.679	0.747	0.539	0.592	0.798	0.853
13_s4+c1.raw	0.837	0.814	0.868	0.796	0.771	0.754	0.683
14_s2+c3.raw	0.842	0.725	0.754	0.606	0.612	0.702	0.585
15_s3+c3.raw	0.510	0.471	0.520	0.510	0.673	0.462	0.448
16_s4+c3.raw	0.563	0.494	0.515	0.501	0.546	0.474	0.456

Correlation, distance and Similarity

$$-1 \leq \rho \leq 1$$

Larger the value of ρ the closer the match.

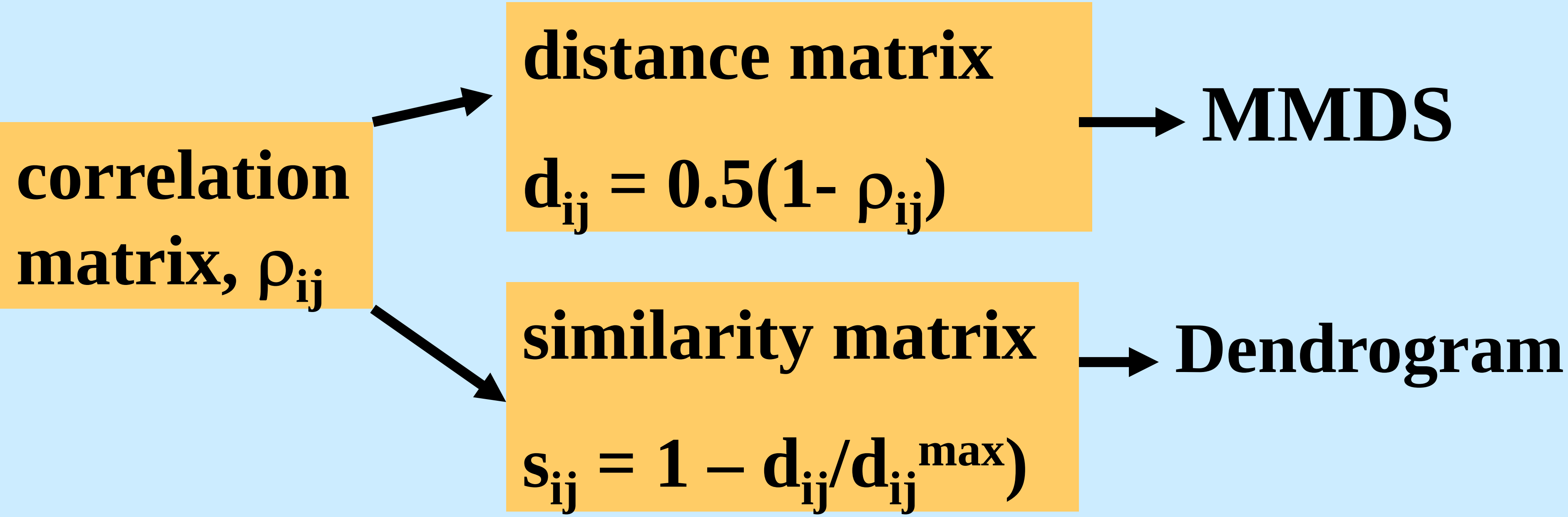
$$0 \leq d \leq 1$$

Smaller the value of d the closer the match.

$$0 \leq s \leq 1$$

Larger the value of s the closer the match.

Correlation, distance and Similarity



correlation matrix, ρ_{ij}

distance matrix

$d_{ij} = 0.5(1- \rho_{ij})$

MMDS

similarity matrix

$s_{ij} = 1 – d_{ij}/d_{ij}^{max})$

Dendrogram

Example: Phase Transitions in Ammonium Nitrate

Work with

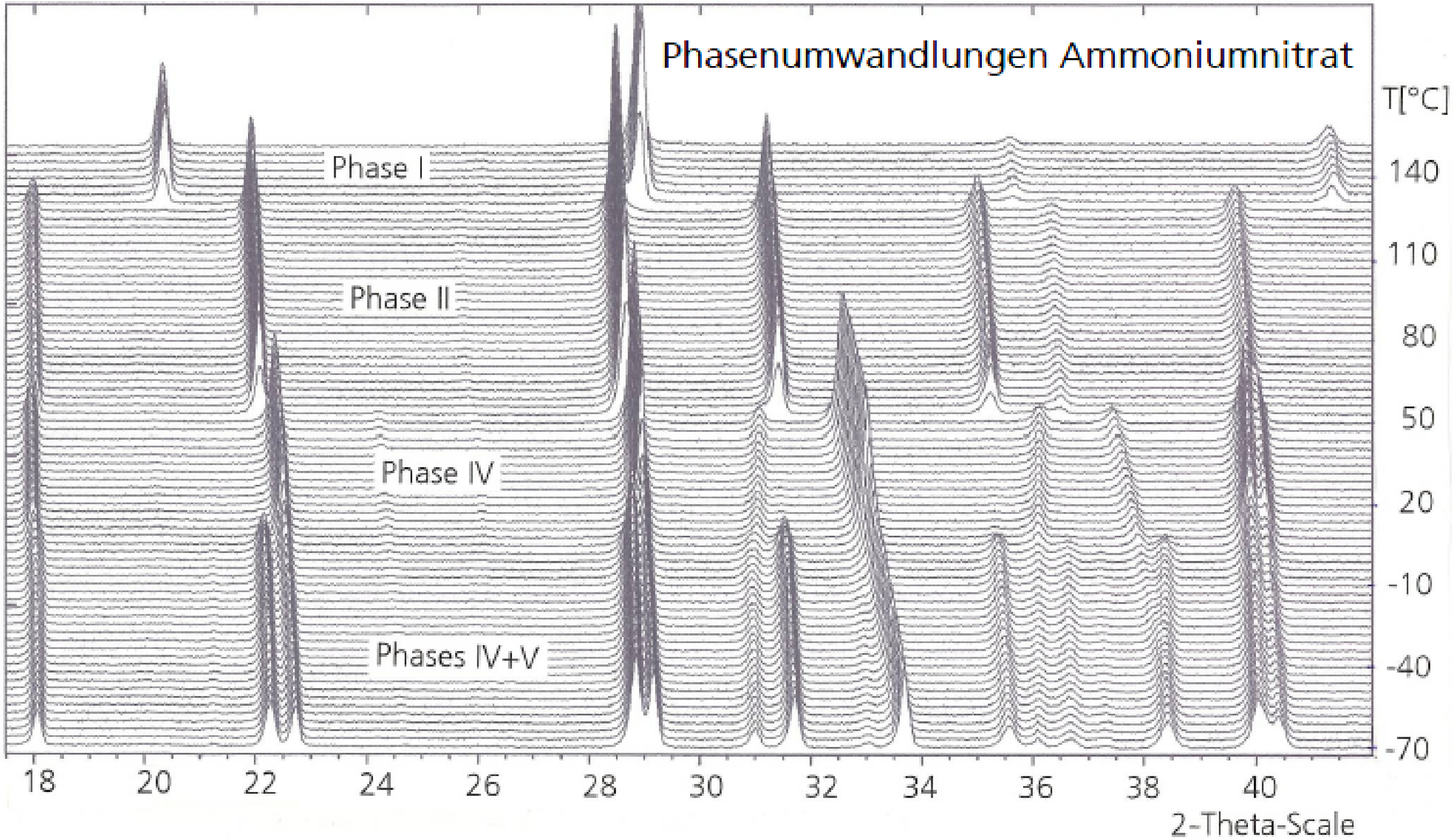
- **Michael Hermannn (Fraunhofer-Institut Chemische Technologie)**
- **Karsten Knorr (Bruker AXS, Karlsruhe)**

See Hermann & Engel ***Propellants, Explosives, Pyrotechnics*** **22**, 143-147 (1997).

Ammonium Nitrate

- 5 phases I – V
- Start with a mixture of IV + V that transforms to IV, II and I.

Ammoniumnitrat; Pfinztal 15. März 2007, M. Herrmann

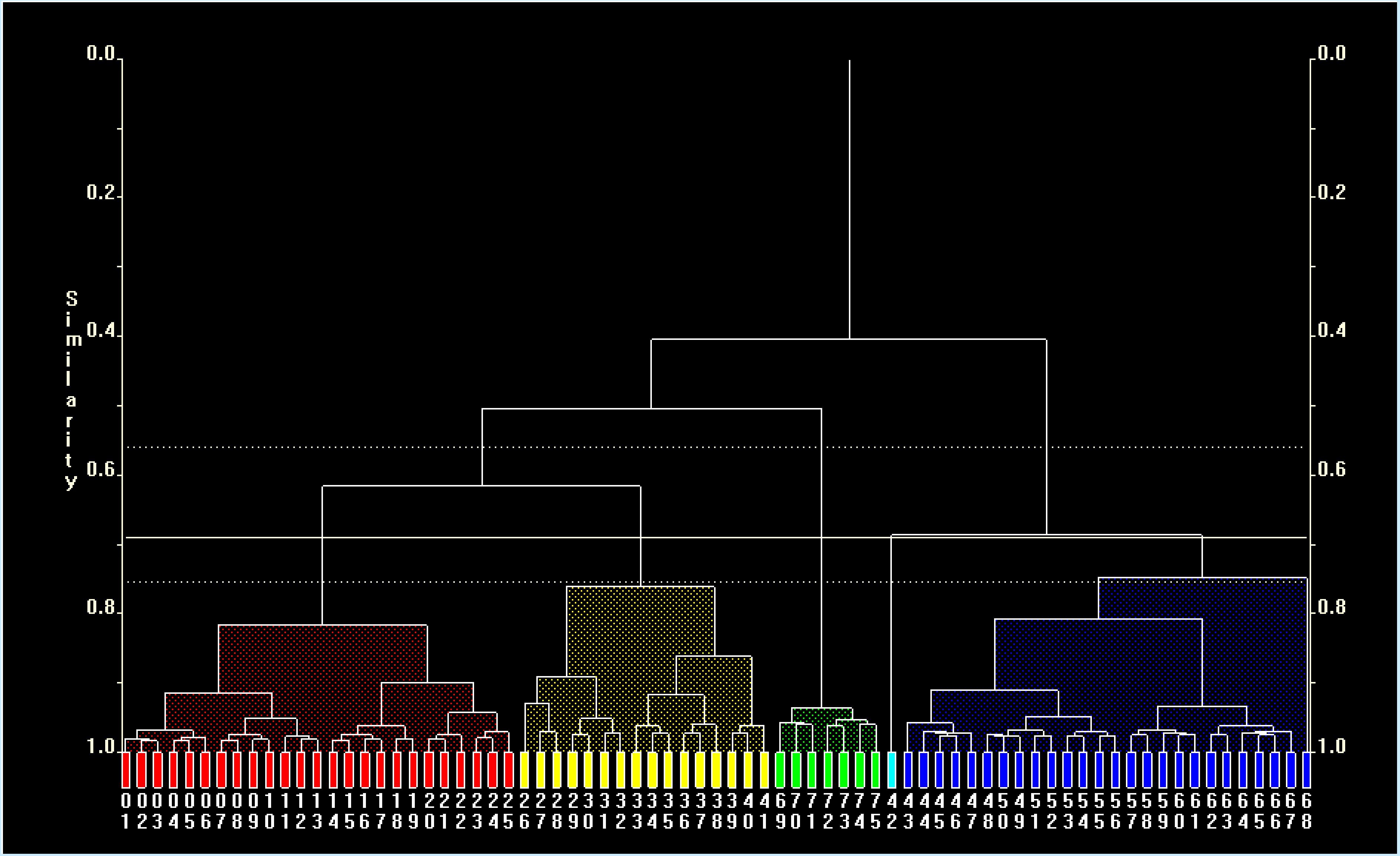


Fraunhofer Institut Chemische Technologie

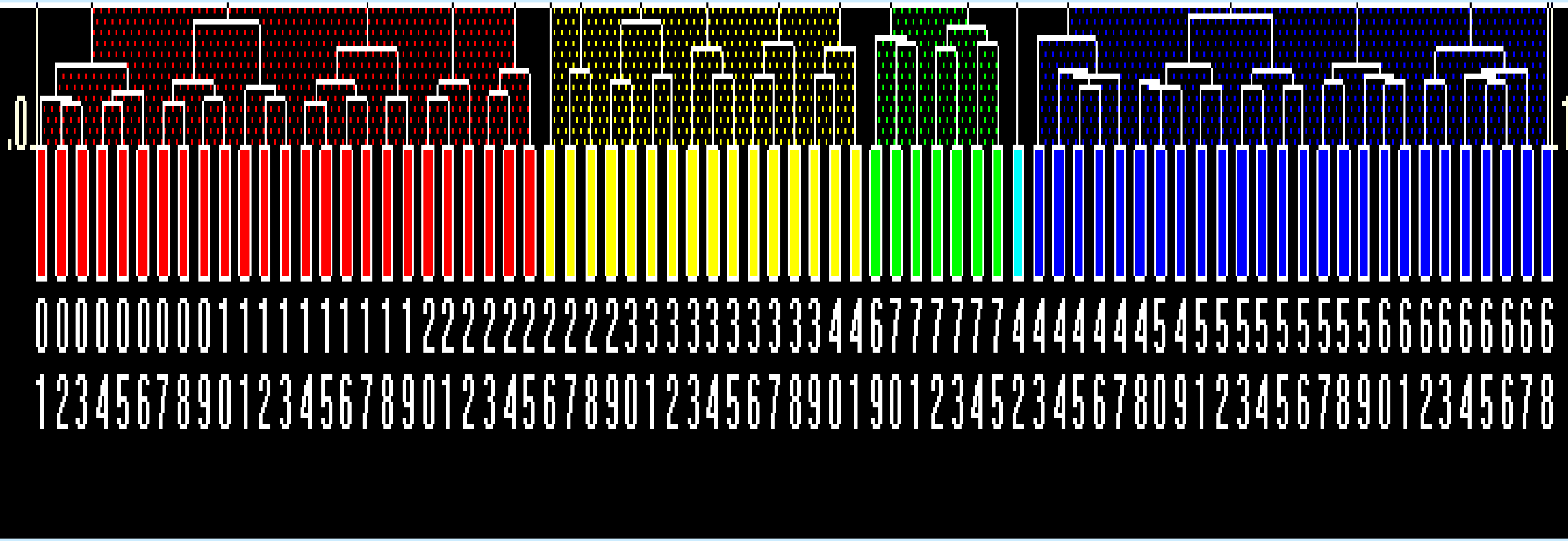
Ammonium Nitrate

- 75 PXRD data sets.
- How to visualize and interpret these data?

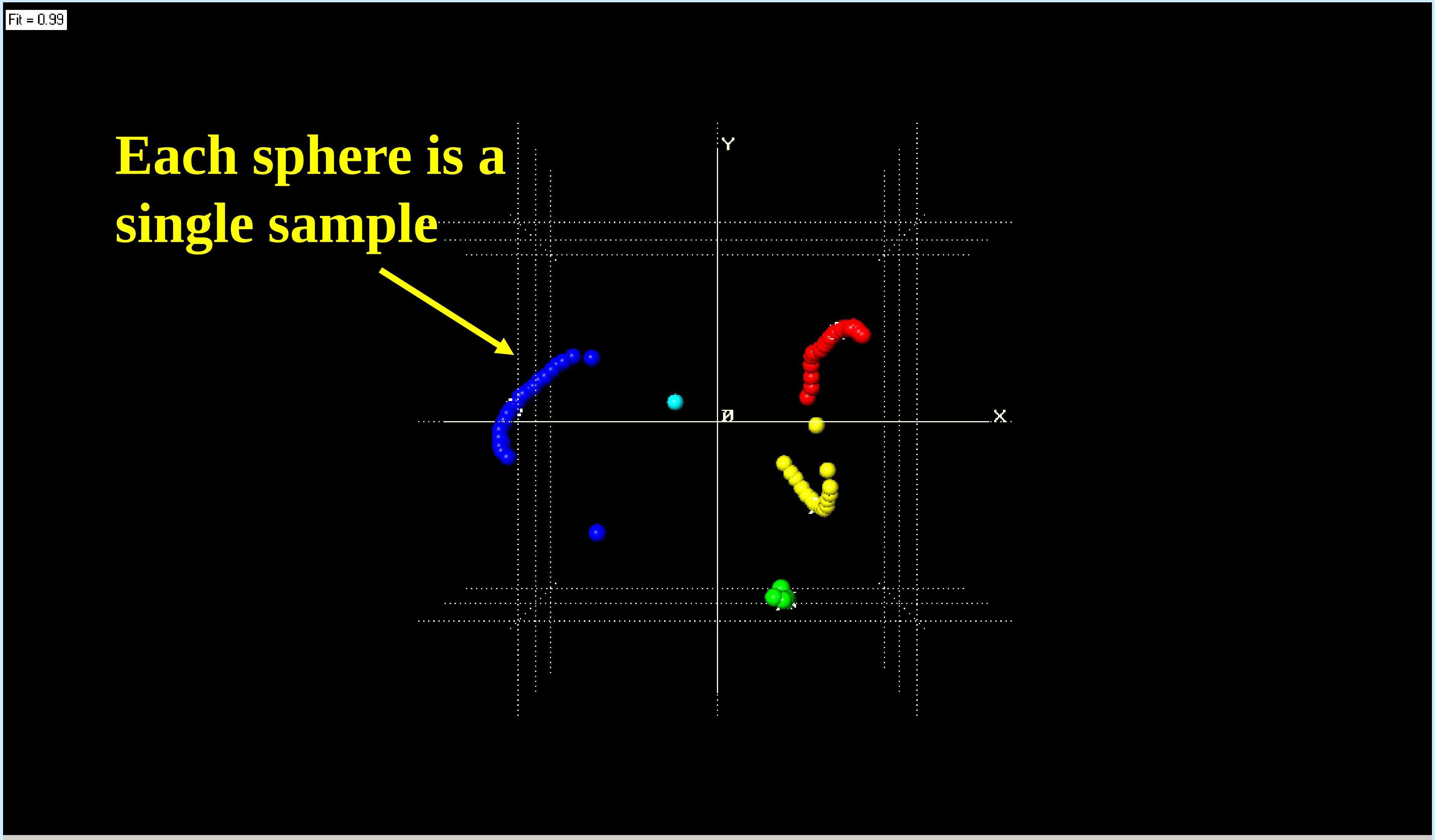
Dendrogram


Similarity
0.0
0.2
0.4
0.6
0.8
1.0
0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 6 7 7 7 7 7 4 4 4 4 4 4 4 5 4 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 9 0 1 2 3 4 5 2 3 4 5 6 7 8 0 9 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8

Numbering

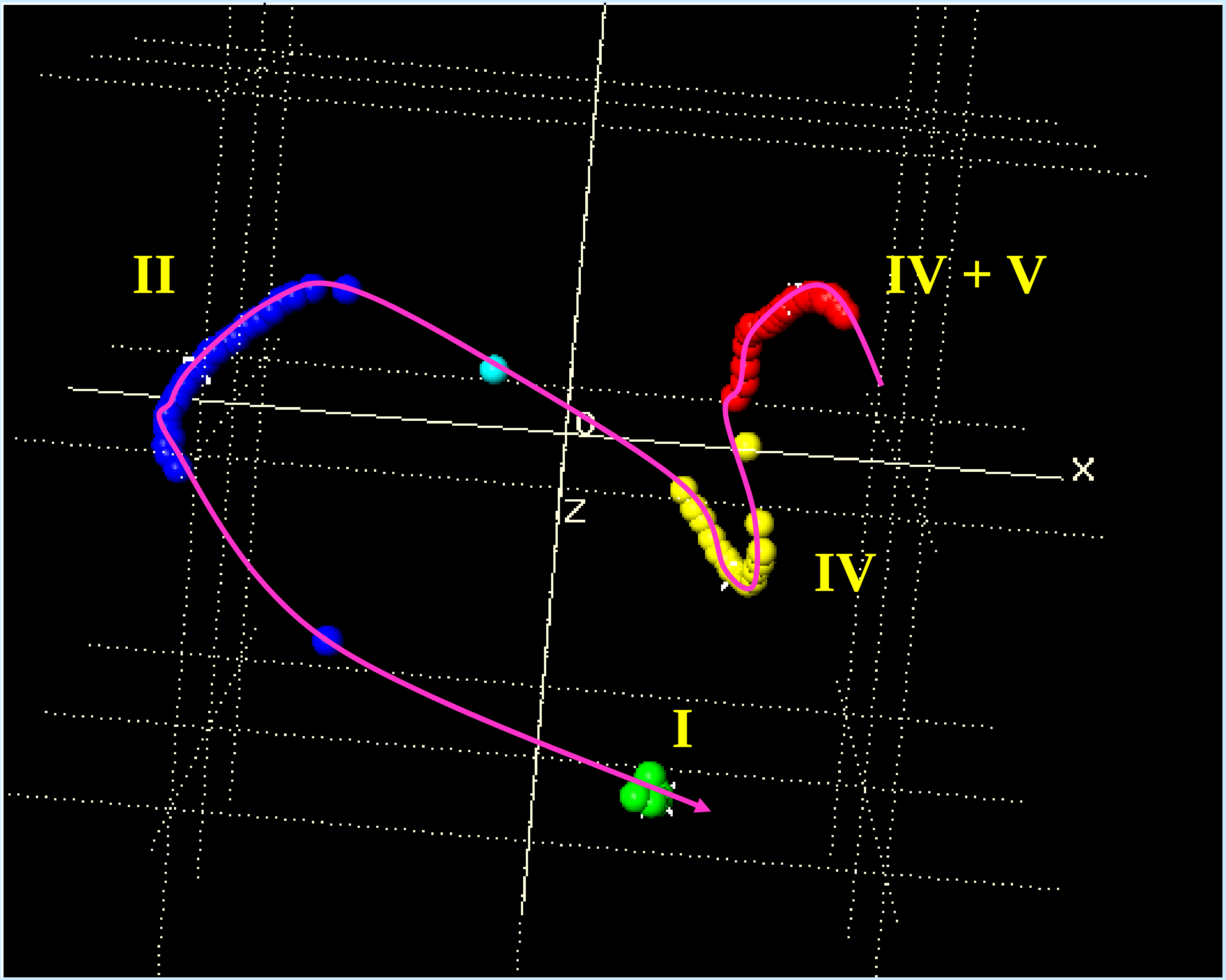

.0
0000000001111111111222222222233333333334467777774444444545555555556666666666
1234567890123456789012345678901234567890190123452345678091234567890123456789
1

Metric Multidimensional Scaling (MMDS)

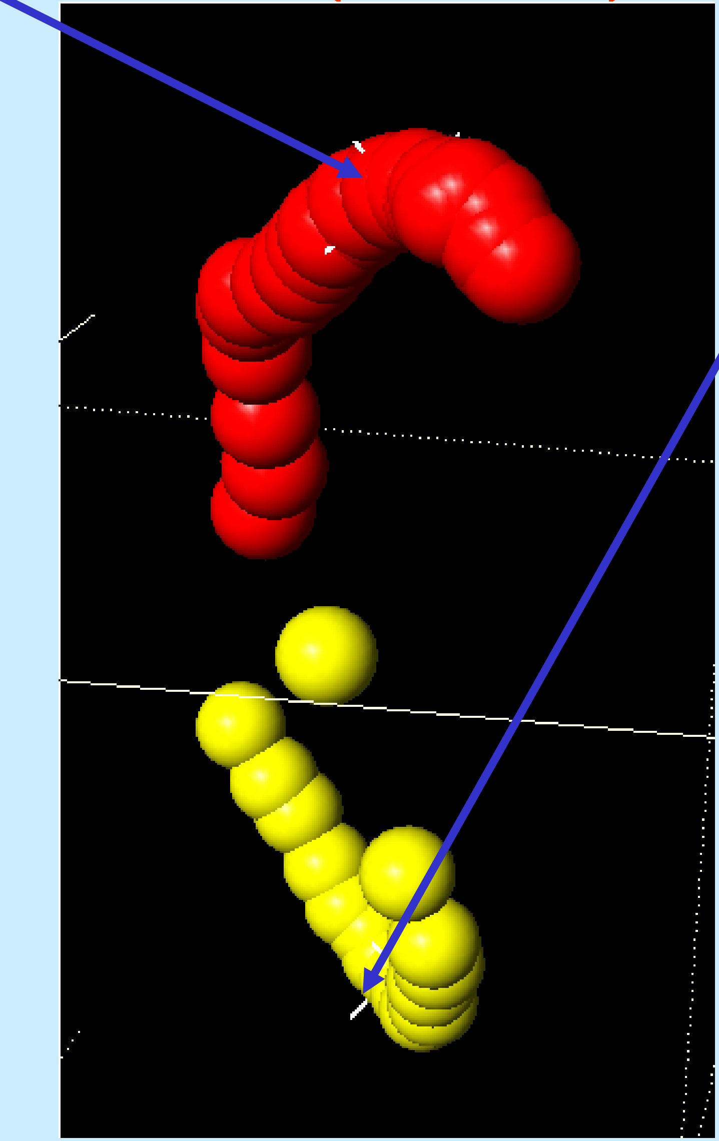


Each sphere is a single sample

Metric Multidimensional Scaling



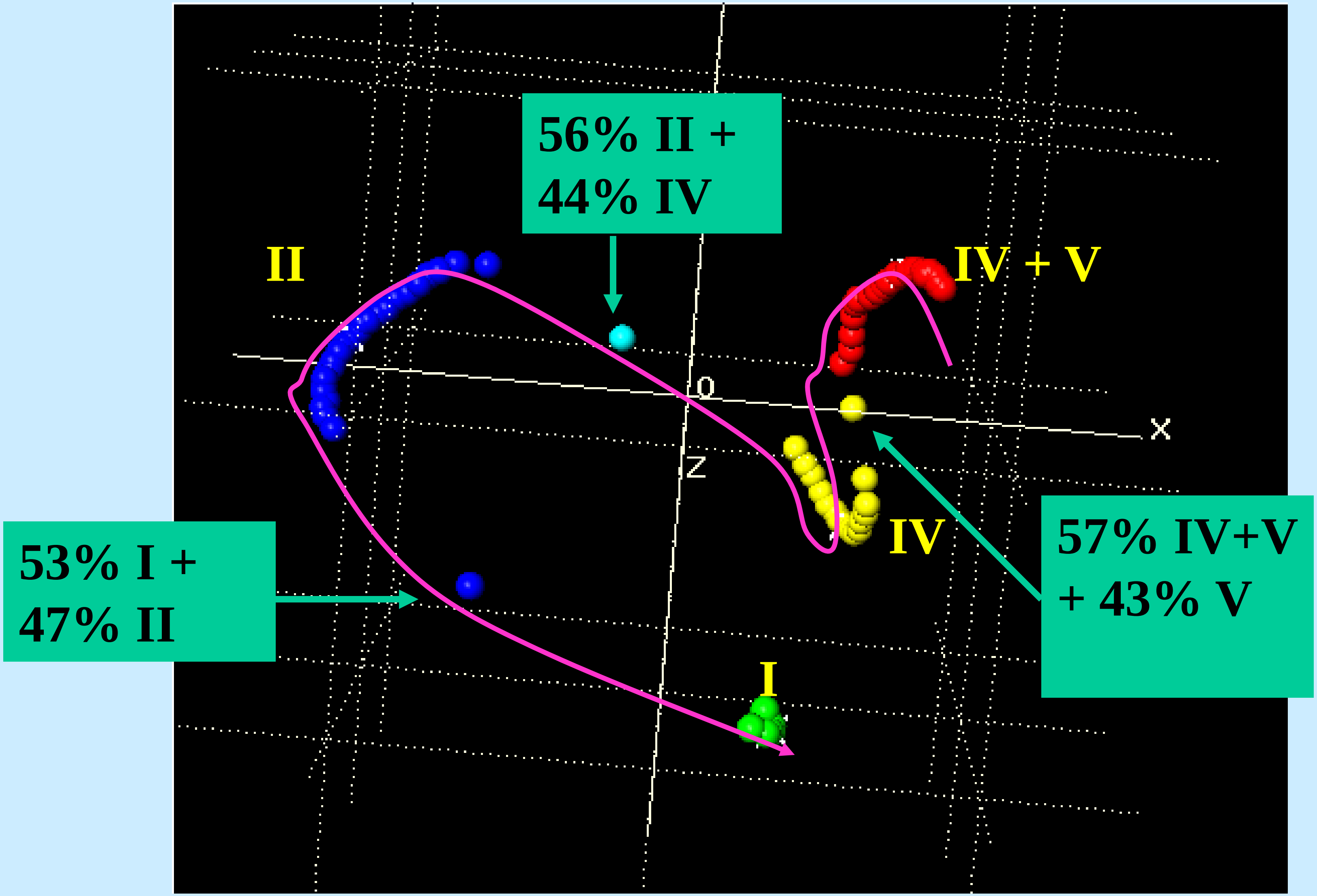
Most Representative Sample (MRS)



Simple Quantitative Analysis

- We know the sequence IV + V → IV → II → I
- Take MRS of IV+V, IV, II, I and use in quantitative analysis.

Simple Quantitative Analysis



56% II + 44% IV

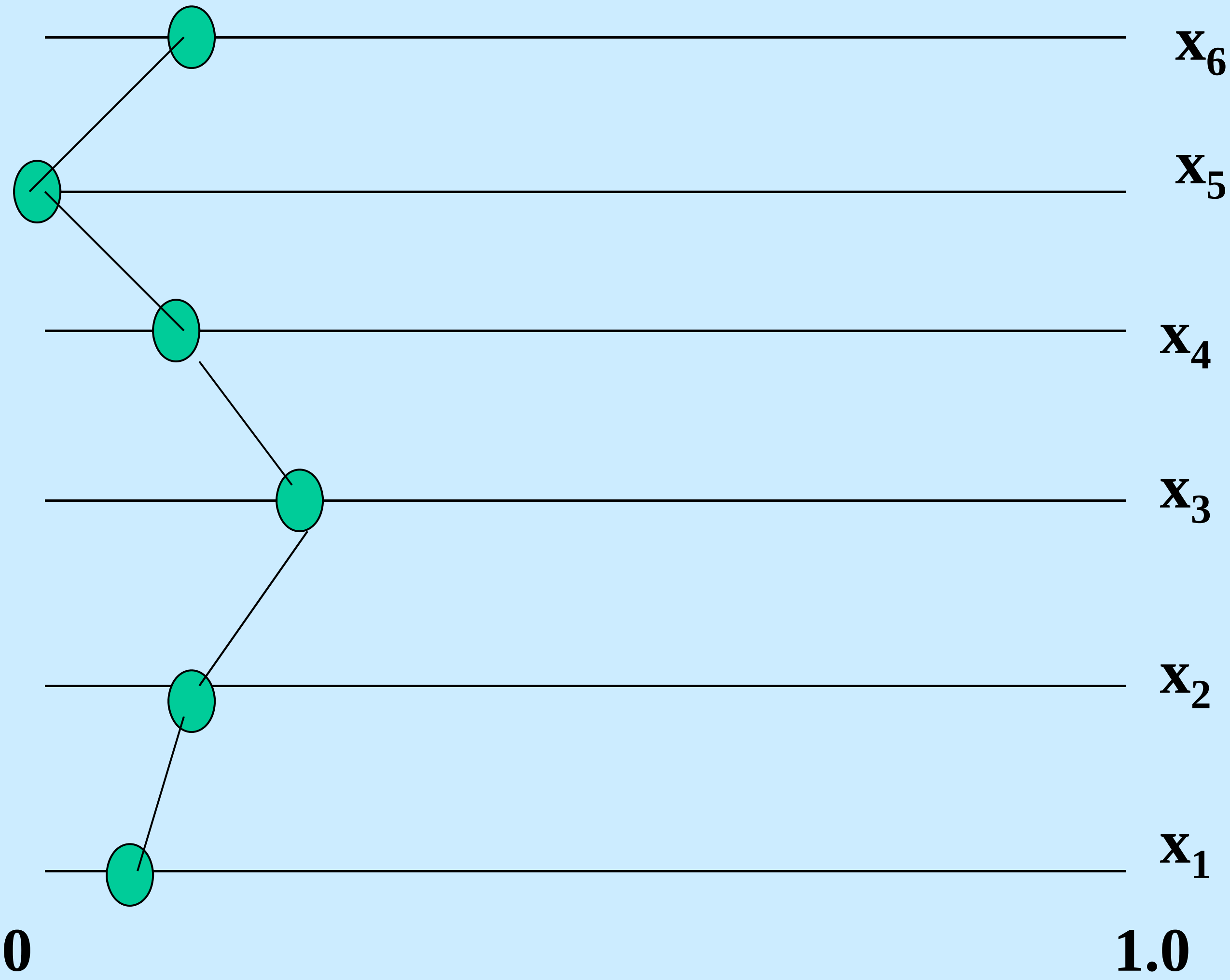
53% I + 47% II

57% IV+V + 43% V

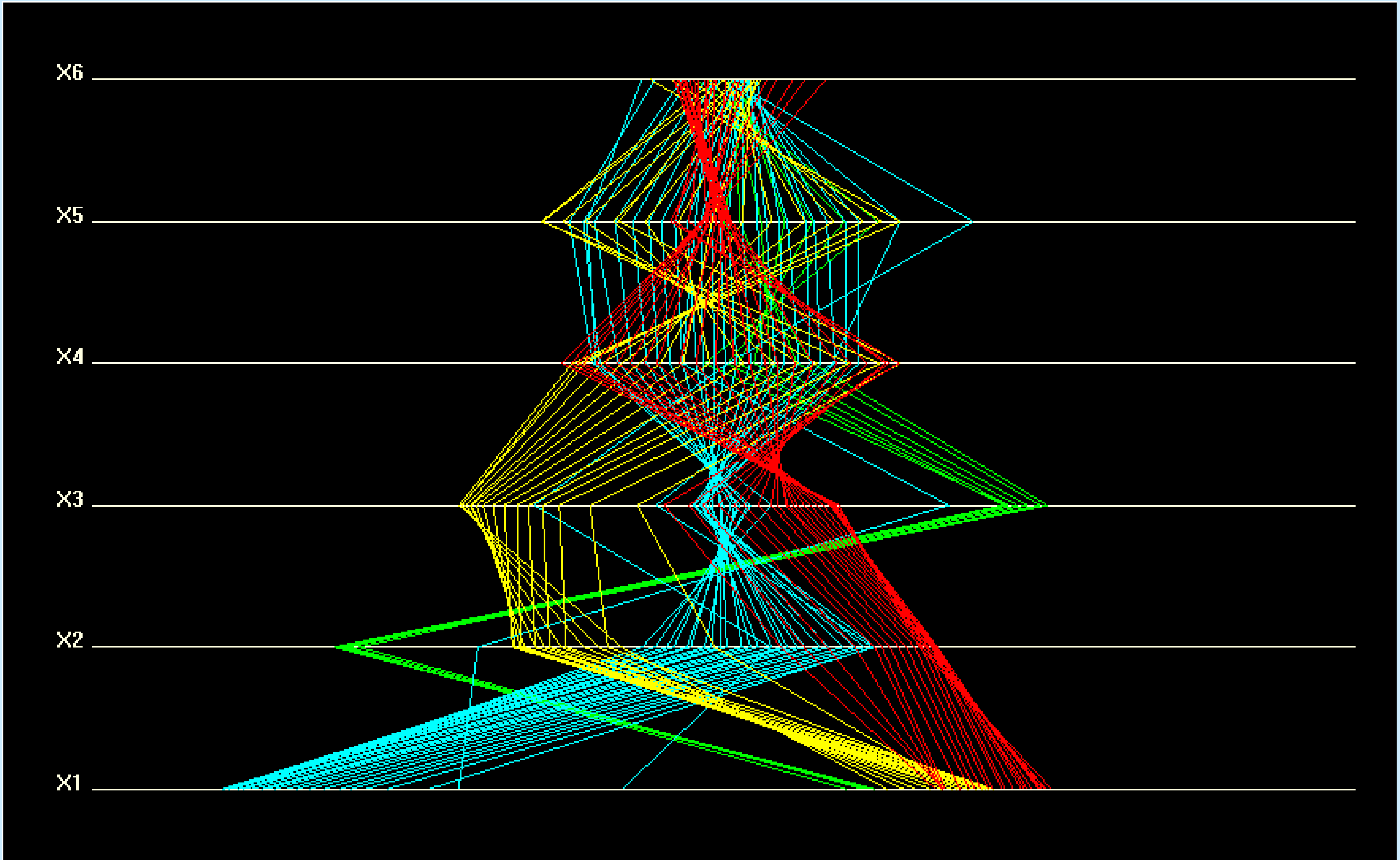
Parallel Coordinate Plots

- **The MMDS (or PCA) calculation gives results in > 3 dimensions, and we merely select the 1st 3 (e.g. the first 3 eigenvectors for each eigenvalue)**
- **Consider the 6 dimensional coordinates**

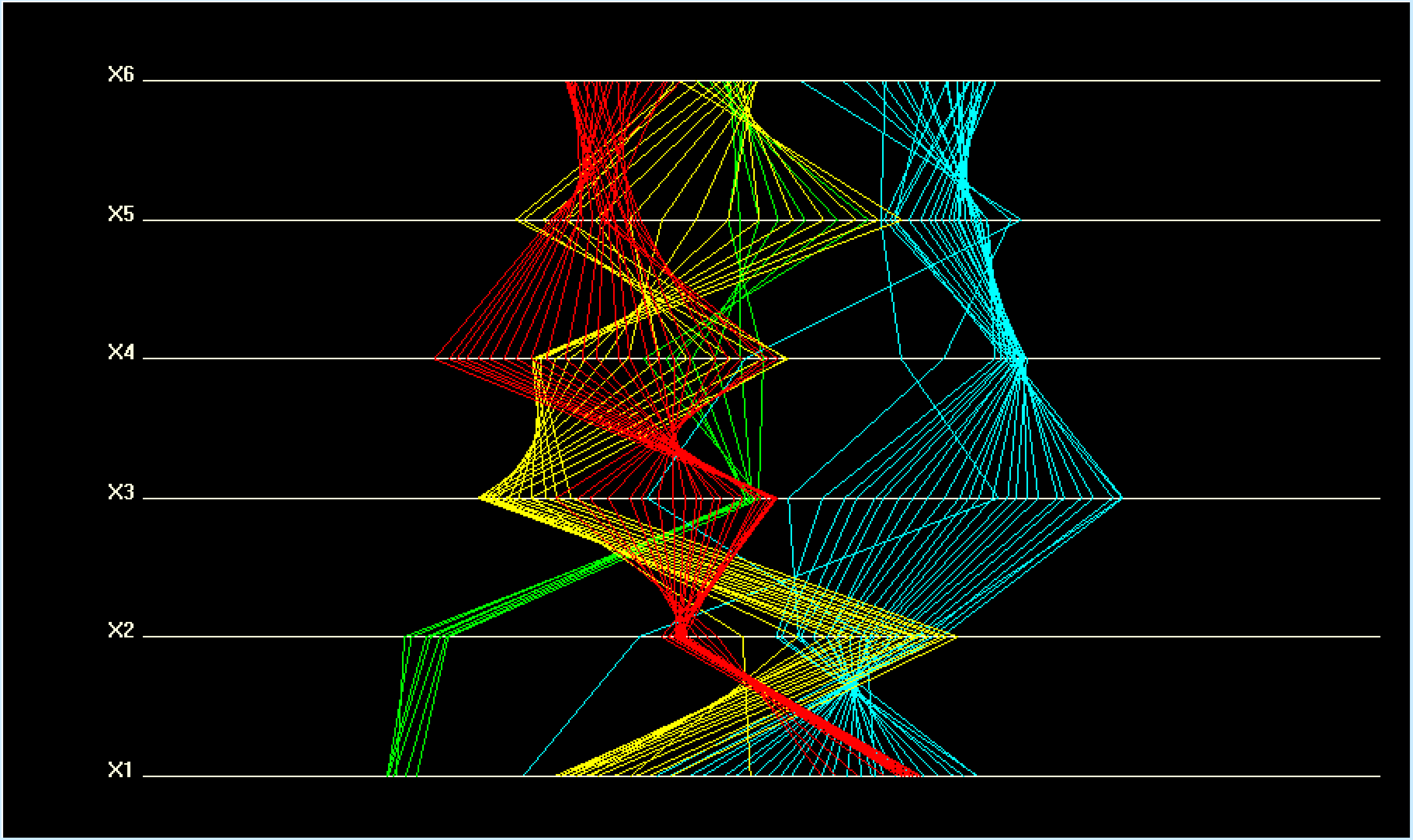
(0.1, 0.2, 0.3, 0.2, 0.0, 0.2) = (x_1, x_2, x_3, x_4, x_5, x_6)


x_6
x_5
x_4
x_3
x_2
x_1
0
1.0

Other Visualization Tools: Parallel Coordinate Plots – More than 3 Dimensions



Another View



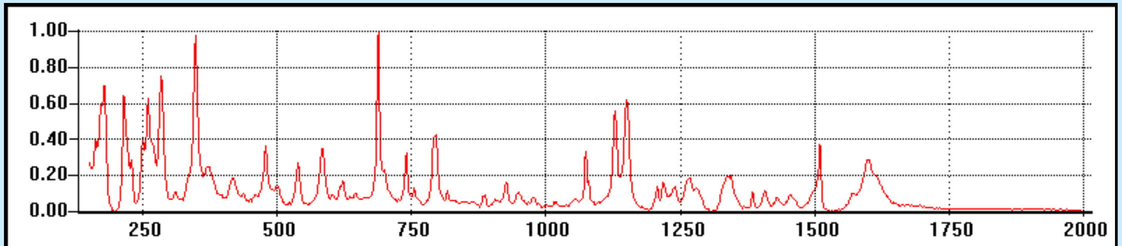
Is PXRD the Gold Standard?

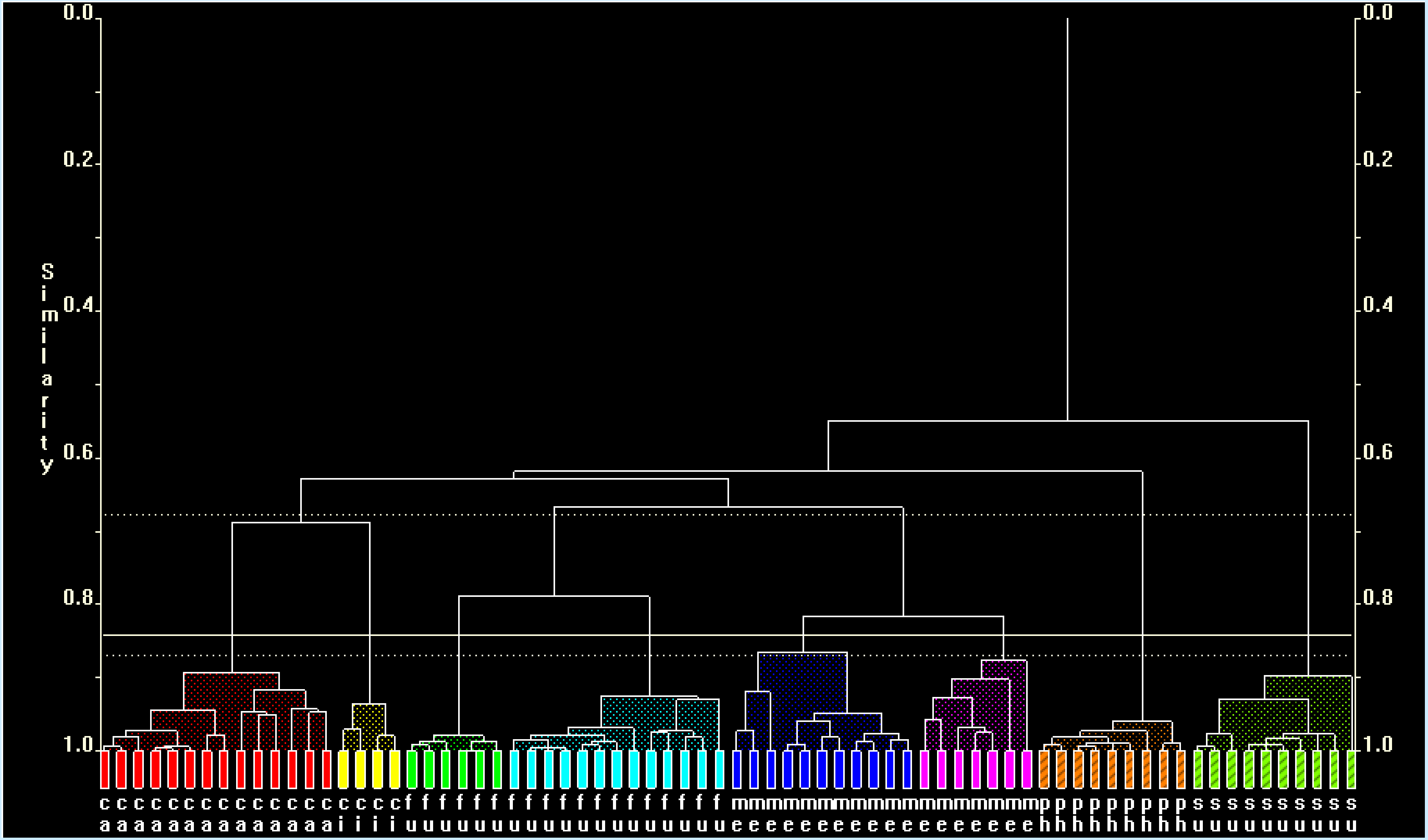
Powder X-ray diffraction is usually considered the gold standard in high throughput studies designed to identify polymorphs, salts, co-crystals *etc.*, but other techniques such as Raman and IR spectroscopy, or differential scanning calorimetry (DSC) can have a major role to play also especially with poorly characterised samples.

Raman Data

- **Raman spectroscopy is well suited to screening: good quality spectra can be collected in a few minutes, and sample preparation is straightforward and flexible, although the resulting spectra are not always as distinct as the PXRD equivalent**

High-throughput powder diffraction V: the use of Raman spectroscopy with and without X-ray powder diffraction data
Barr, Cunningham, Dong, Gilmore & Kojima *J. Appl. Cryst.* (2009). 42, 706–714





carbamazepine

furosemide

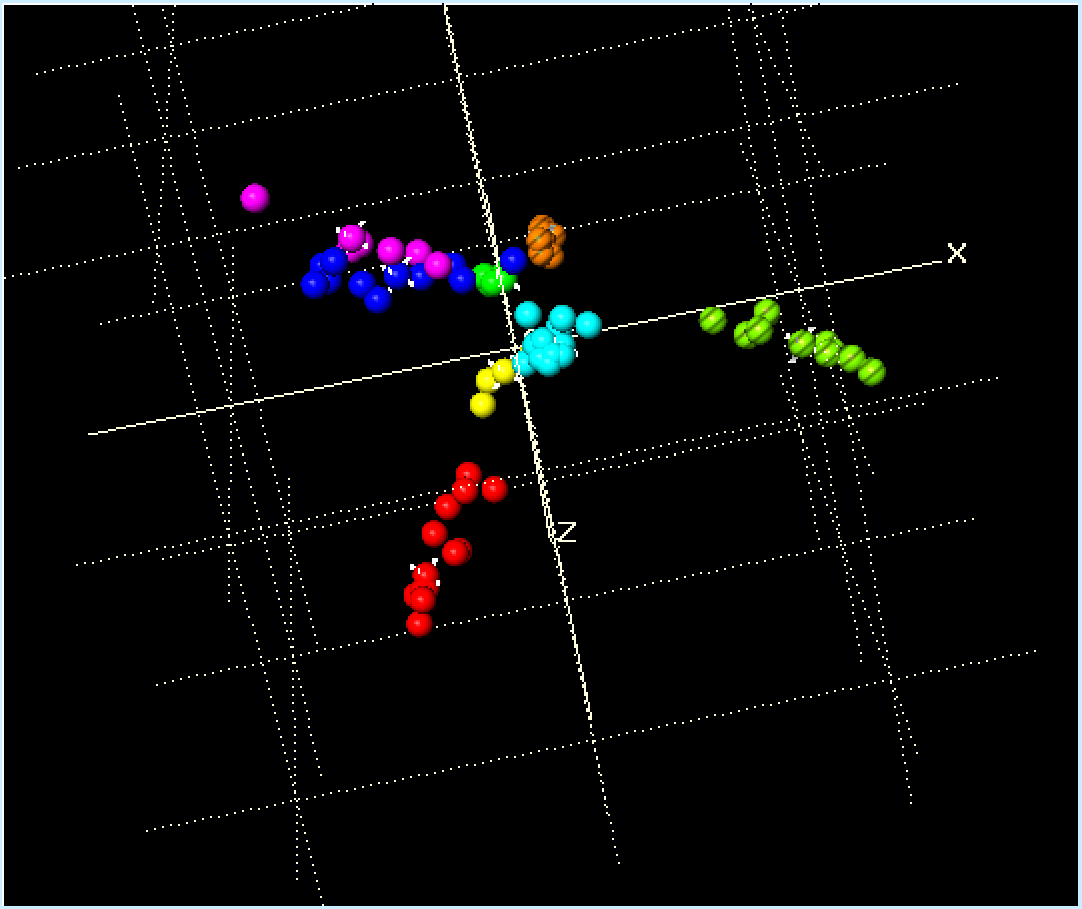
phenylbutazone

cimetidene

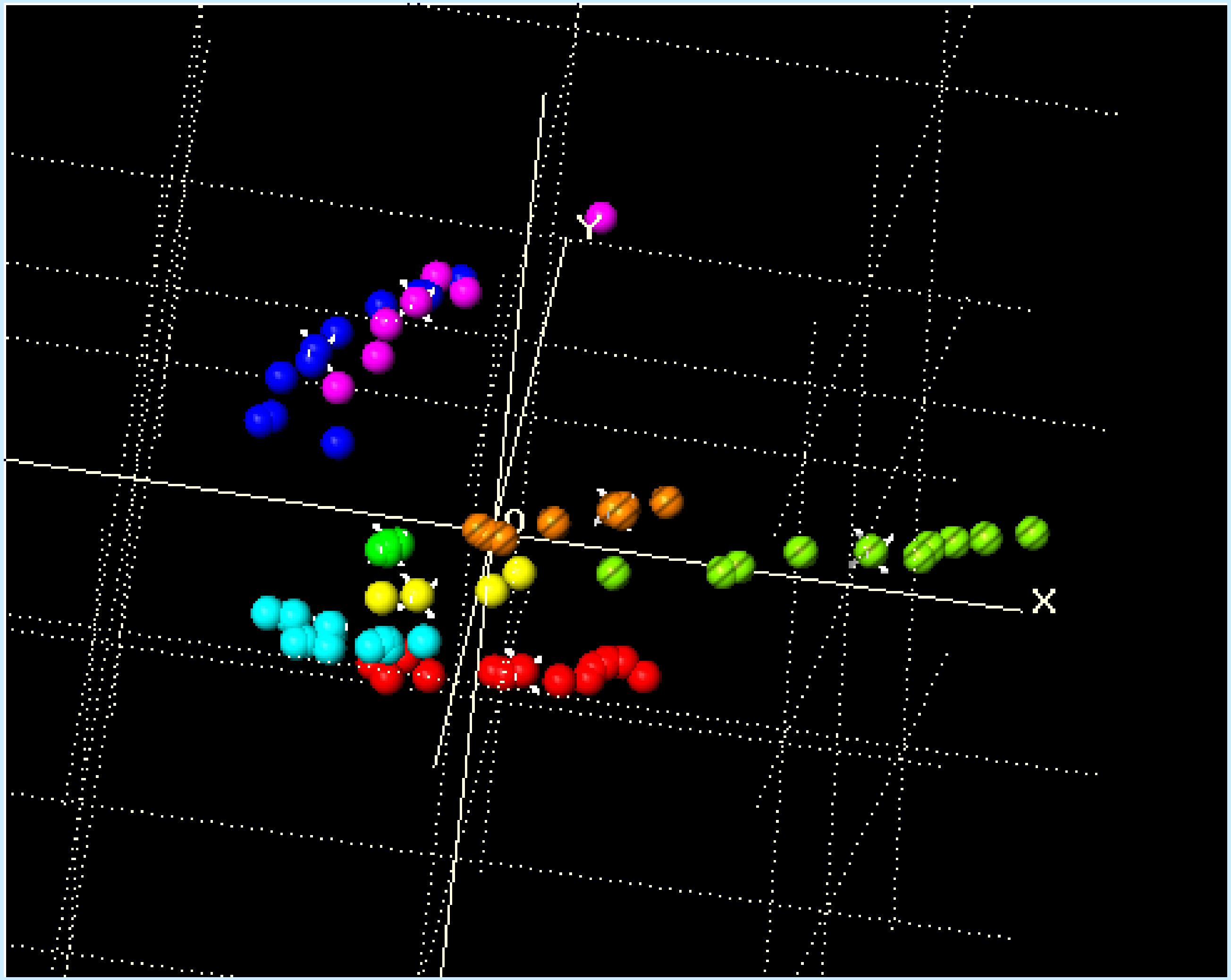
mefenamicacid

sulfamerazine

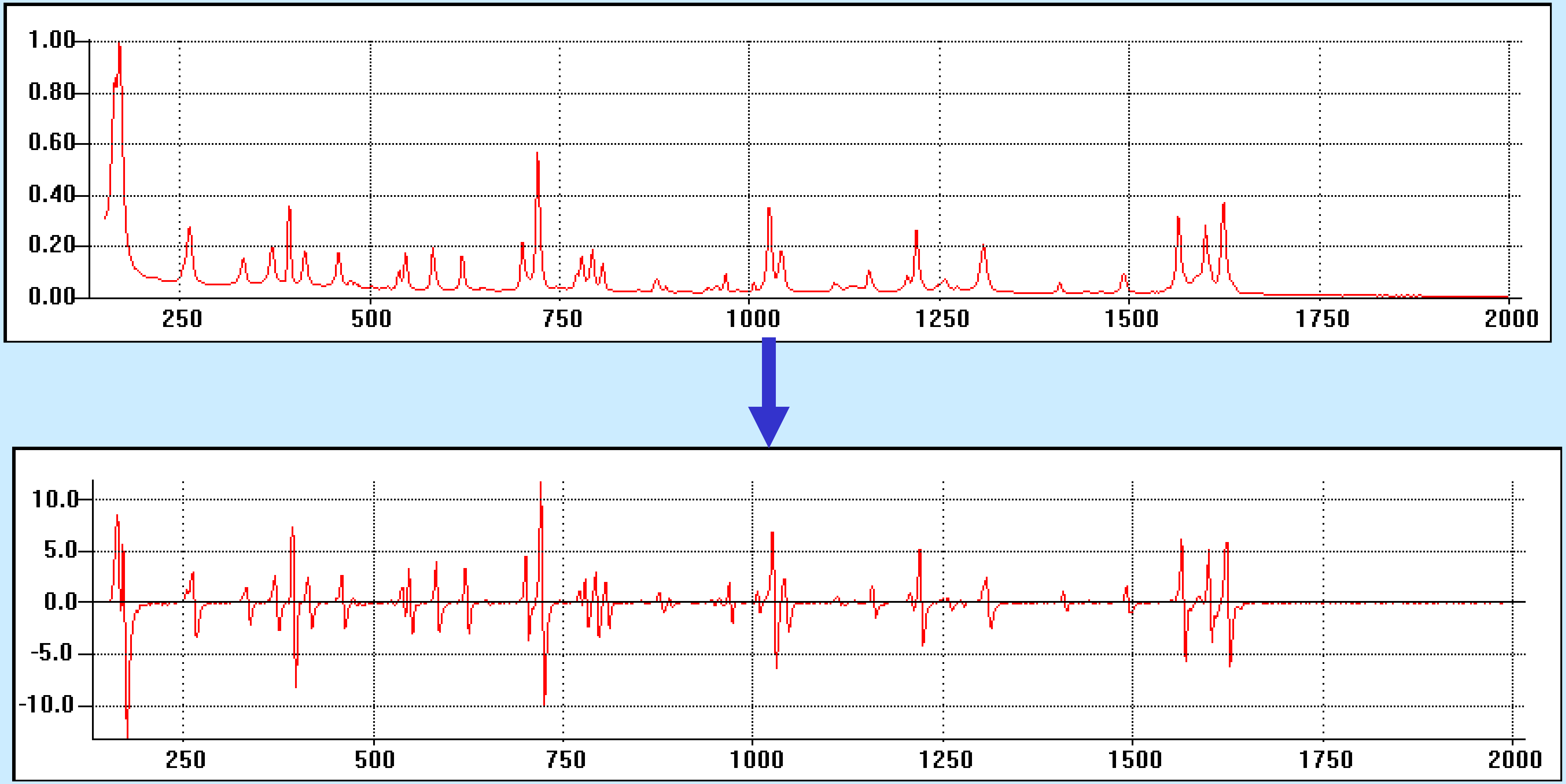
MMDS



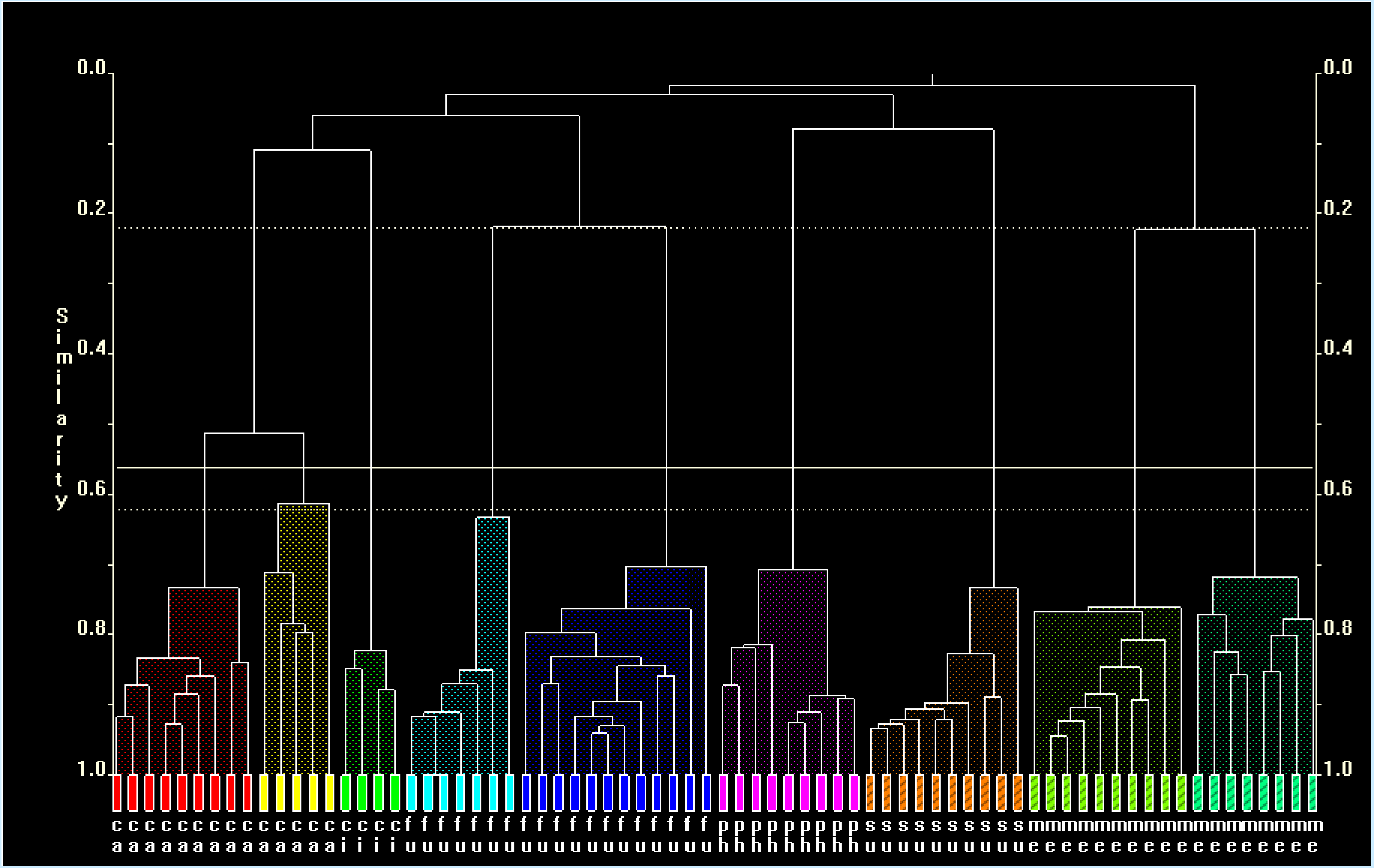
PCA – Be Wary!



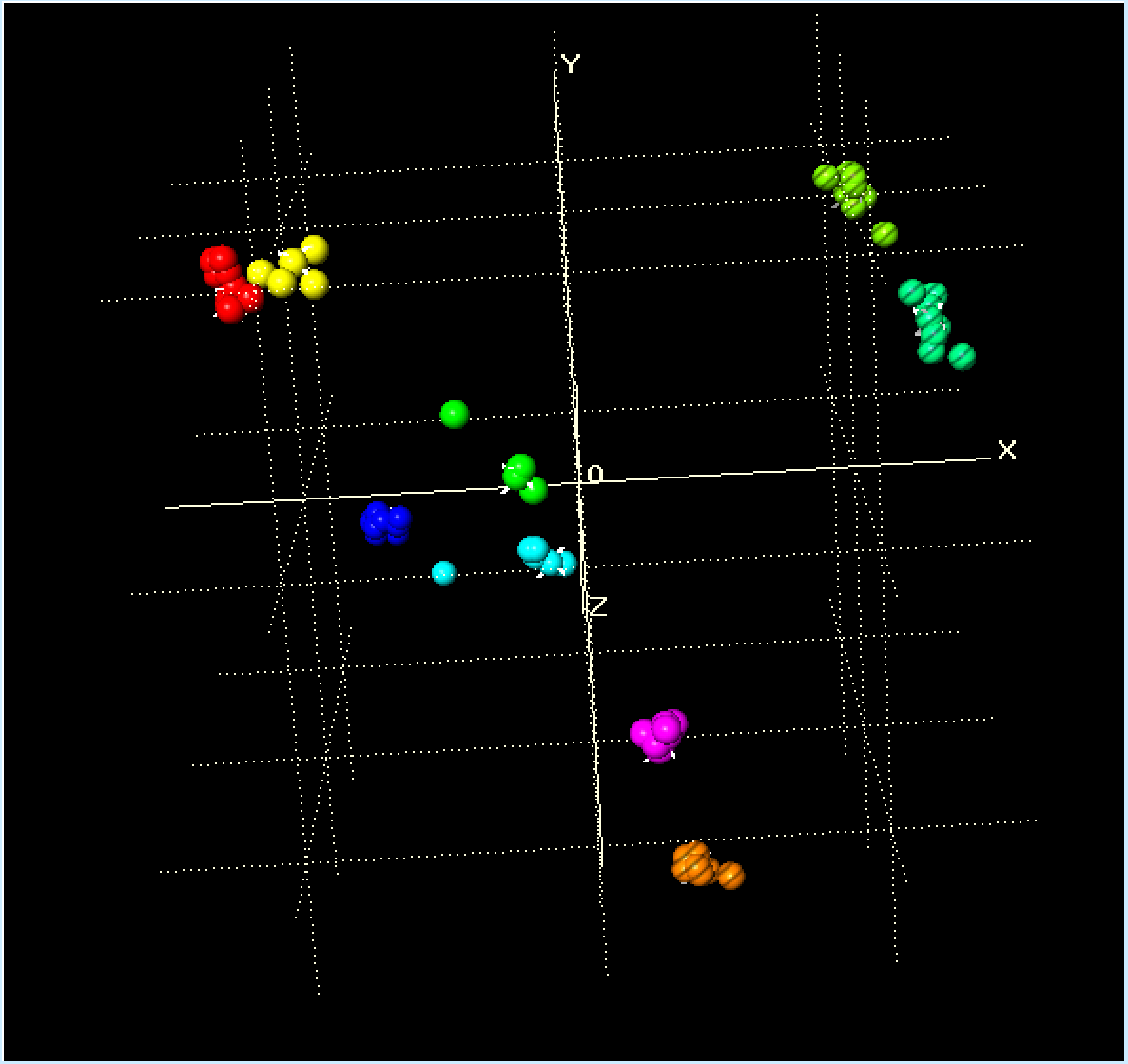
Use 1st Derivative Data



Dendrogram on Derivative Data



MMDS on Derivative Data



How Do You Combine Different Data Types?

- Combined XRPD + Raman instruments now available
- Applying multiple techniques to the same samples gives additional information
- How would we actually combine results from two (or more) such different techniques ?

Combining Data Types

- **Manual weighting:**
  - **Give a single weight to each dataset *e.g.* Powder 0.8, Raman 0.2**
  - **Use Fischer transforms.**

- **Dynamic weighting:**
  - **Automatically calculate optimal weighting for each entry in each dataset**
  - **Unbiased solution that scales the differences between individual distance matrices**

Dynamic weighting

- **Dynamic Weighting using INDSCAL:**

  Independent Scaling of Differences

  Carroll & Chang, (1970) *Psychometrica* 35, 283-319

- **Each data set has a 2-D distance matrix *d***
- **D_k is squared ($n \times n$) distance matrix for dataset *k***

  ***e.g.* we have Raman and XRPD data on 20 samples, so $k = 2$, n=20.**

- **We want a Group Average Matrix, G, to optimally describe our data**
- **Specify diagonal weight matrices W_k which can vary over the *k* datasets**

Dynamic Weighting

Matrices are matched to weighted form of G by minimising

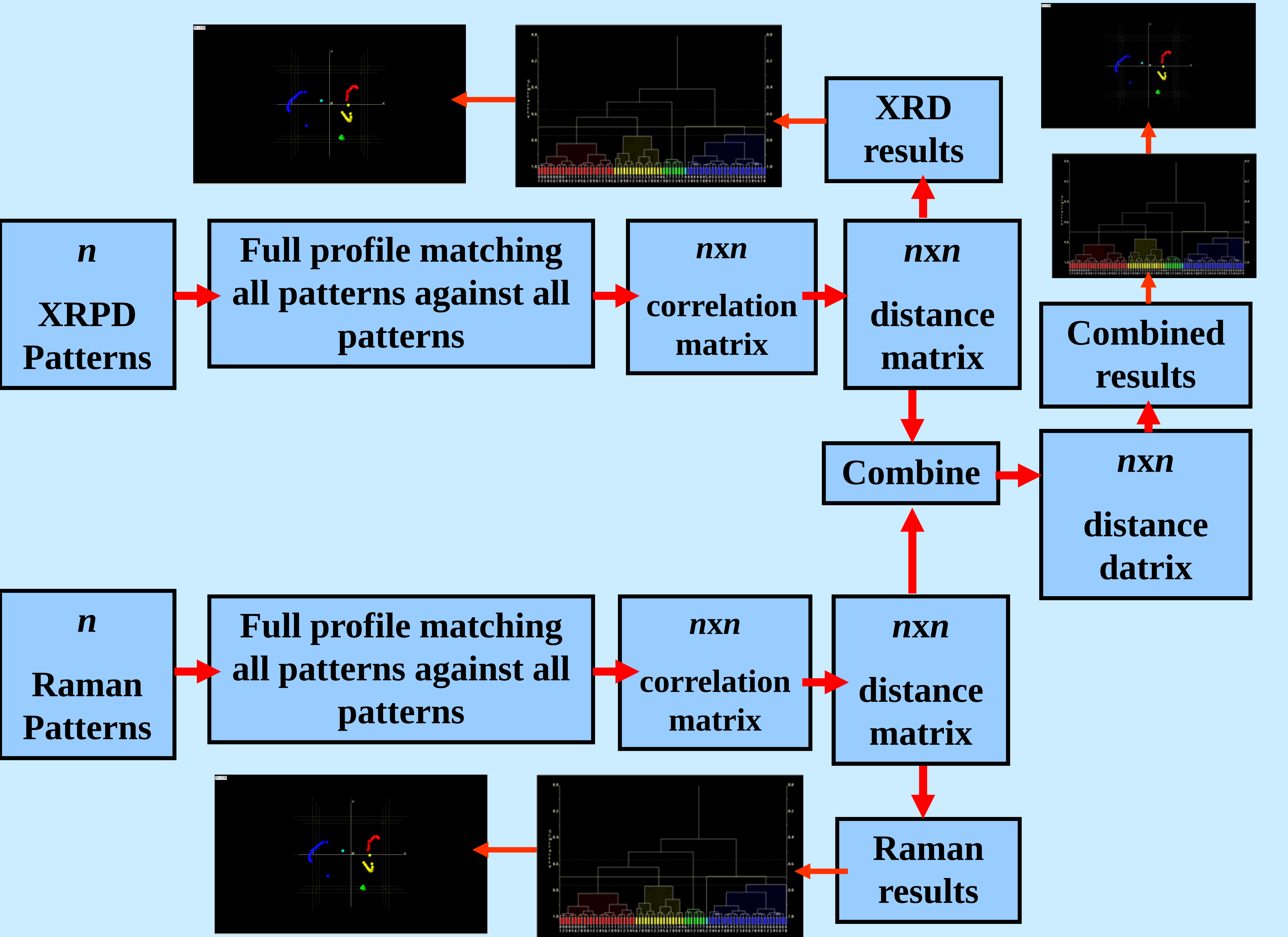
$$\sum_{k=1}^{K} \left\| \mathbf{B}_k - \mathbf{G}\mathbf{W}_k^2\mathbf{G}' \right\| \qquad (1)$$

Where

$$\mathbf{B}_k = -\frac{1}{2}\left(\mathbf{I} - \mathbf{N}\right)\mathbf{D}_k\left(\mathbf{I} - \mathbf{N}\right)$$

(a double-centering operation on D), and solve (1) to get best values for G and W

The resulting G matrix is then used as before



n

XRPD Patterns

Full profile matching all patterns against all patterns

nxn

correlation matrix

nxn

distance matrix

XRD results

Combine

nxn

distance datrix

Combined results

n

Raman Patterns

Full profile matching all patterns against all patterns

nxn

correlation matrix

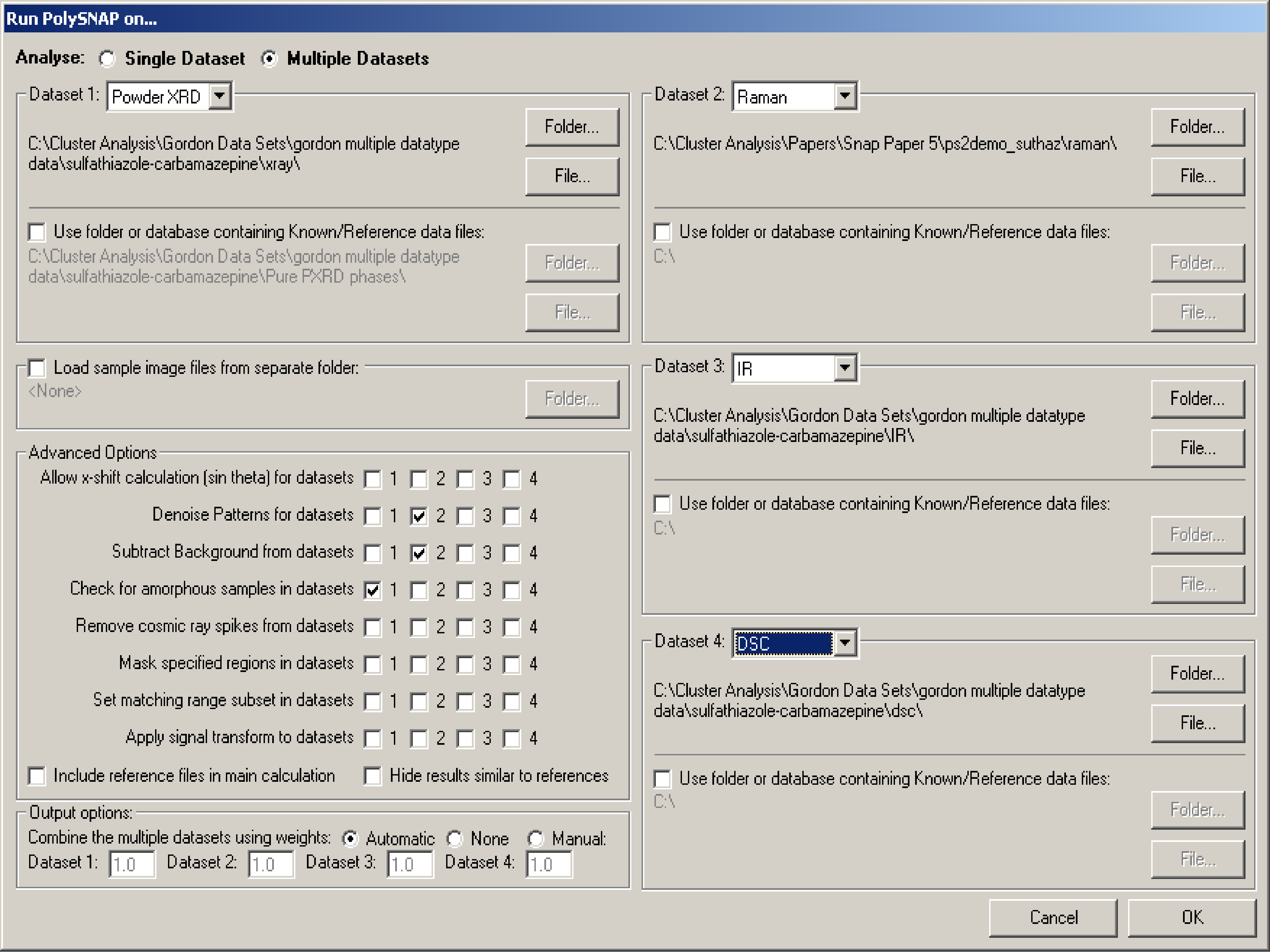
nxn

distance matrix

Raman results

PXRD + Raman

- Forms 2,3 and 4 of sulfathizole
- 48 samples, no mixtures.



Run PolySNAP on...

Analyse: ○ Single Dataset ● Multiple Datasets

Dataset 1: Powder XRD

C:\Cluster Analysis\Gordon Data Sets\gordon multiple datatype data\sulfathiazole-carbamazepine\xray\

□ Use folder or database containing Known/Reference data files:

C:\Cluster Analysis\Gordon Data Sets\gordon multiple datatype data\sulfathiazole-carbamazepine\Pure PXRD phases\

□ Load sample image files from separate folder:

<None>

Advanced Options

Allow x-shift calculation (sin theta) for datasets	□ 1	□ 2	□ 3	□ 4
Denoise Patterns for datasets	□ 1	☑ 2	□ 3	□ 4
Subtract Background from datasets	□ 1	☑ 2	□ 3	□ 4
Check for amorphous samples in datasets	☑ 1	□ 2	□ 3	□ 4
Remove cosmic ray spikes from datasets	□ 1	□ 2	□ 3	□ 4
Mask specified regions in datasets	□ 1	□ 2	□ 3	□ 4
Set matching range subset in datasets	□ 1	□ 2	□ 3	□ 4
Apply signal transform to datasets	□ 1	□ 2	□ 3	□ 4

□ Include reference files in main calculation □ Hide results similar to references

Output options:

Combine the multiple datasets using weights: ● Automatic ○ None ○ Manual:

Dataset 1: 1.0 Dataset 2: 1.0 Dataset 3: 1.0 Dataset 4: 1.0

Dataset 2: Raman

C:\Cluster Analysis\Papers\Snap Paper 5\ps2demo_suthaz\raman\

□ Use folder or database containing Known/Reference data files:

C:\

Dataset 3: IR

C:\Cluster Analysis\Gordon Data Sets\gordon multiple datatype data\sulfathiazole-carbamazepine\IR\

□ Use folder or database containing Known/Reference data files:

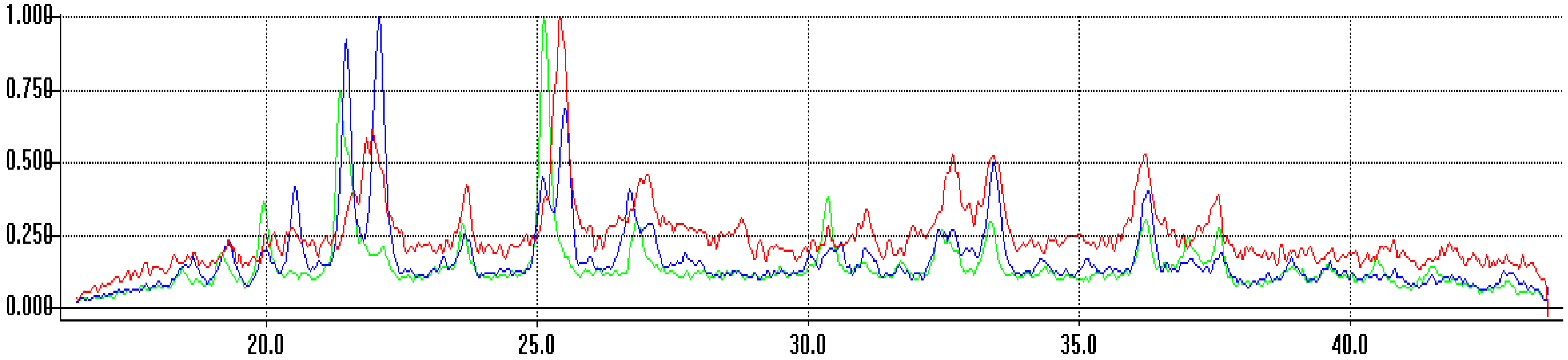
C:\

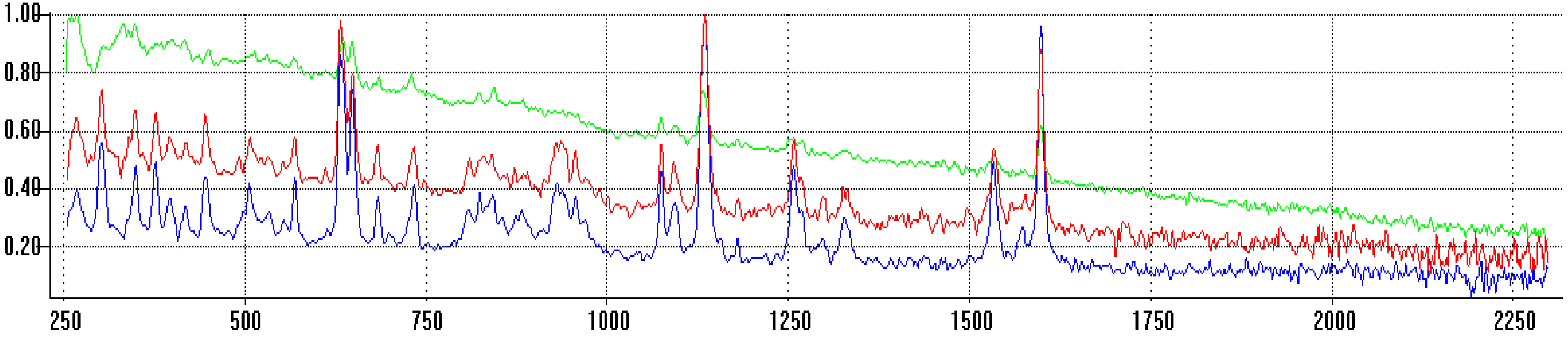
Dataset 4: DSC

C:\Cluster Analysis\Gordon Data Sets\gordon multiple datatype data\sulfathiazole-carbamazepine\dsc\

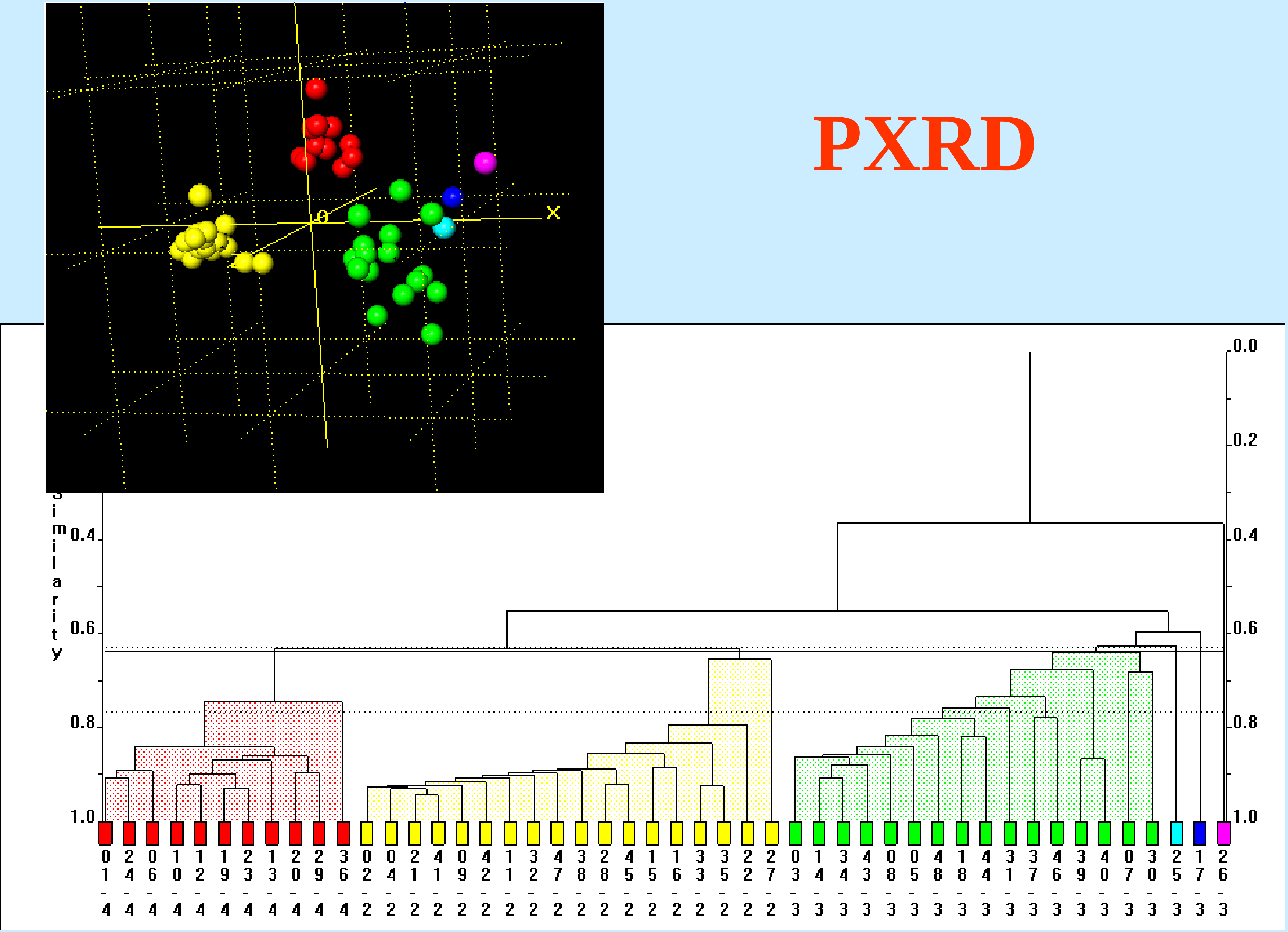
□ Use folder or database containing Known/Reference data files:

C:\

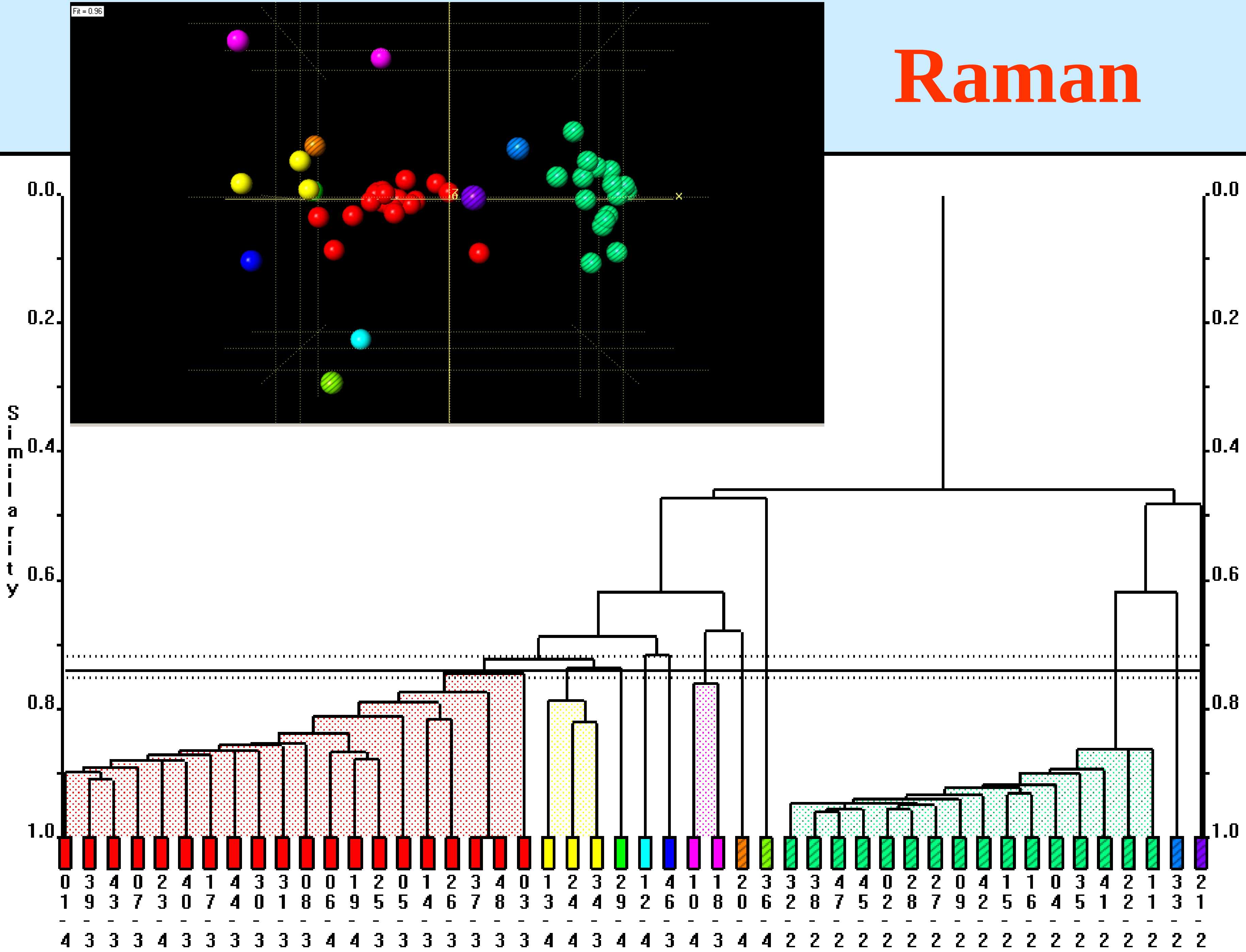

1.000
0.750
0.500
0.250
0.000
20.0
25.0
30.0
35.0
40.0


1.00
0.80
0.60
0.40
0.20
250
500
750
1000
1250
1500
1750
2000
2250

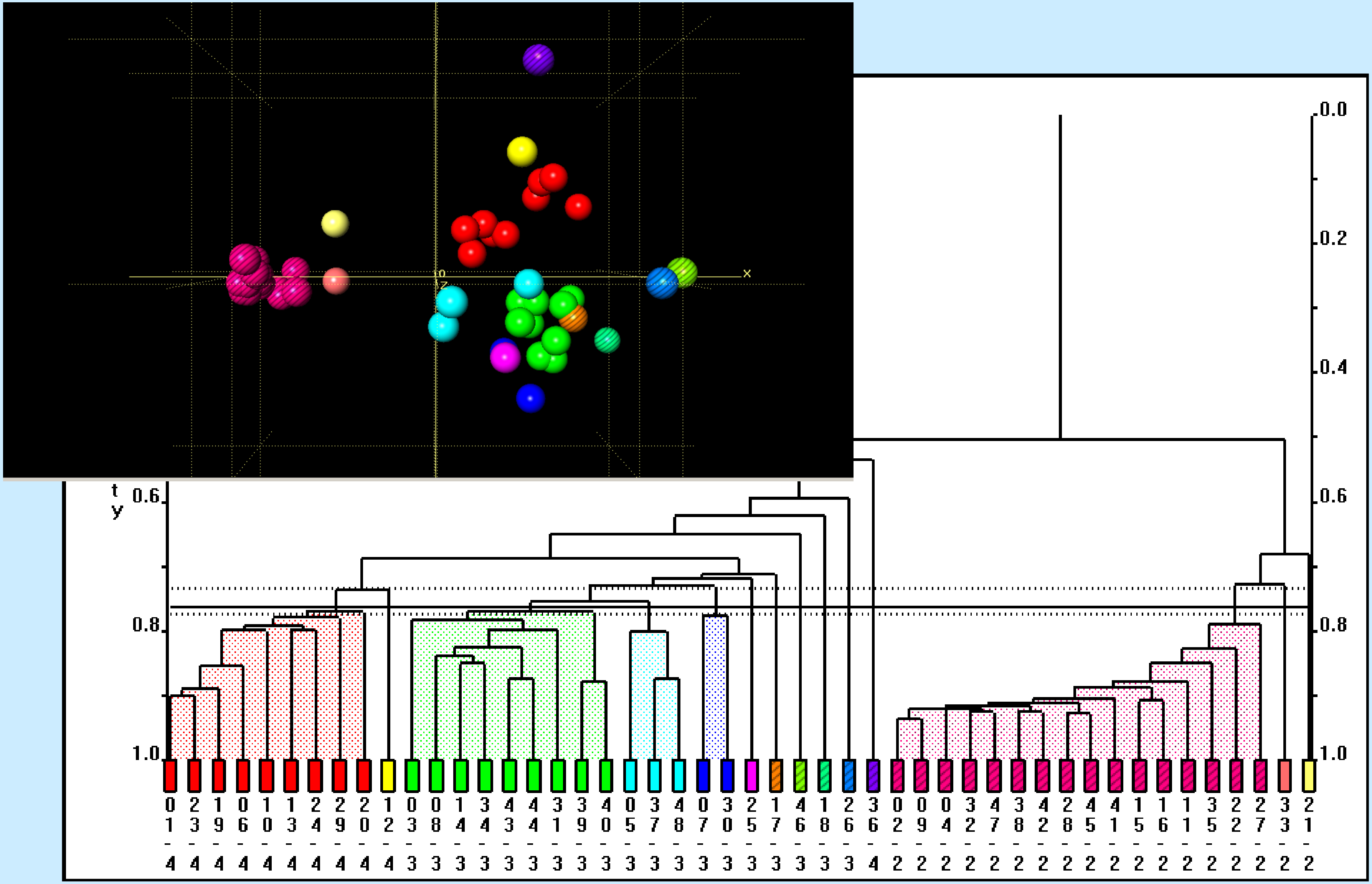
PXRD


X
O
Similarity
0.0
0.2
0.4
0.6
0.8
1.0
01-4
24-4
06-4
10-4
12-4
19-4
23-4
13-4
20-4
29-4
36-4
02-2
04-2
21-2
41-2
09-2
42-2
11-2
32-2
47-2
38-2
28-2
45-2
15-2
16-2
33-2
35-2
22-2
27-2
03-3
14-3
34-3
43-3
08-3
05-3
48-3
18-3
44-3
31-3
37-3
46-3
39-3
40-3
07-3
30-3
25-3
17-3
26-3

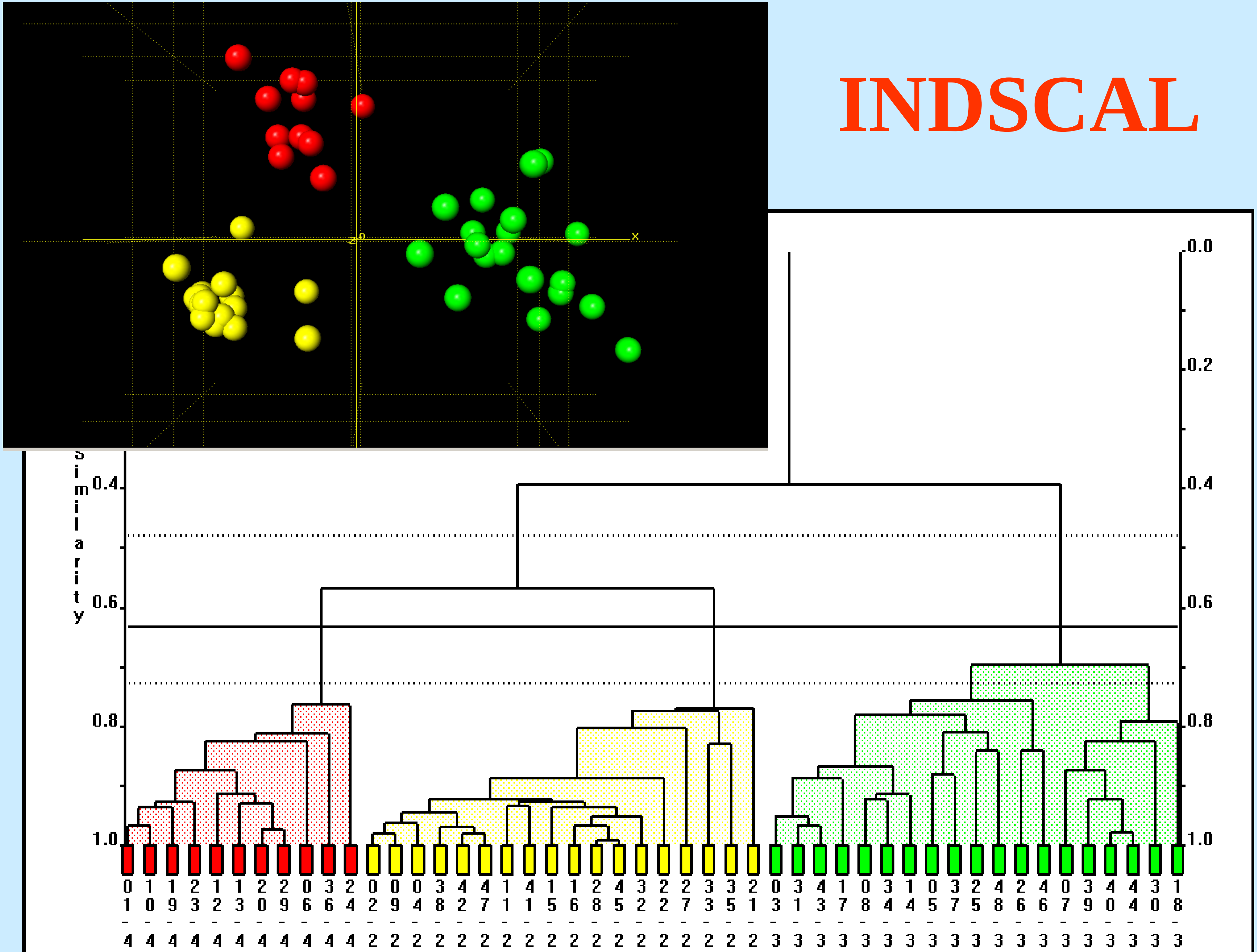
Raman


Fit = 0.96
Similarity
0.0
0.2
0.4
0.6
0.8
1.0
01-4
39-3
43-3
07-3
23-4
40-3
17-3
44-3
30-3
31-3
08-3
06-4
19-4
25-3
05-3
14-3
26-3
37-3
48-3
03-3
13-4
24-4
34-3
29-4
12-4
46-3
10-4
18-3
20-4
36-4
32-2
38-2
47-2
45-2
02-2
28-2
27-2
09-2
42-2
15-2
16-2
04-2
35-2
41-2
22-2
11-2
33-2
21-2

Raman and PXRD Averaged



INDSCAL


x
Similarity
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
01-4
10-4
19-4
23-4
12-4
13-4
20-4
29-4
06-4
36-4
24-4
02-2
09-2
04-2
38-2
42-2
47-2
11-2
41-2
15-2
16-2
28-2
45-2
32-2
22-2
27-2
33-2
35-2
21-2
03-3
31-3
43-3
17-3
08-3
34-3
14-3
05-3
37-3
25-3
48-3
26-3
46-3
07-3
39-3
40-3
44-3
30-3
18-3

Other Combinations

- **Raman + Raman derivative data.**
- **Different data collection protocols/apparatus on the same samples.**
- **You can include numeric data as a data type *e.g.* image analysis data.**

Numeric data – get the distance matrix directly.

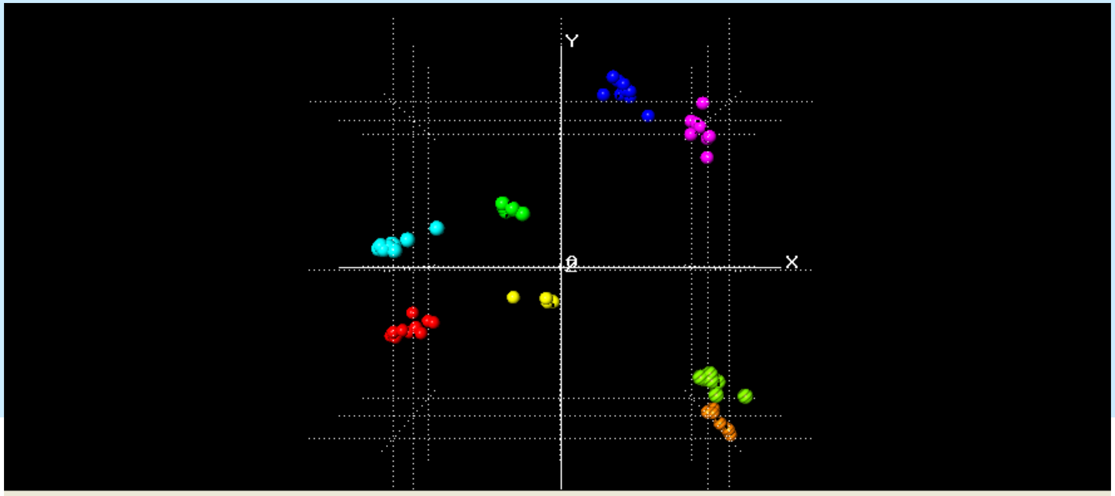
Sample 1: 113.431 58.531 155.845 … {x_{11}, x_{12} , x_{13}…….}

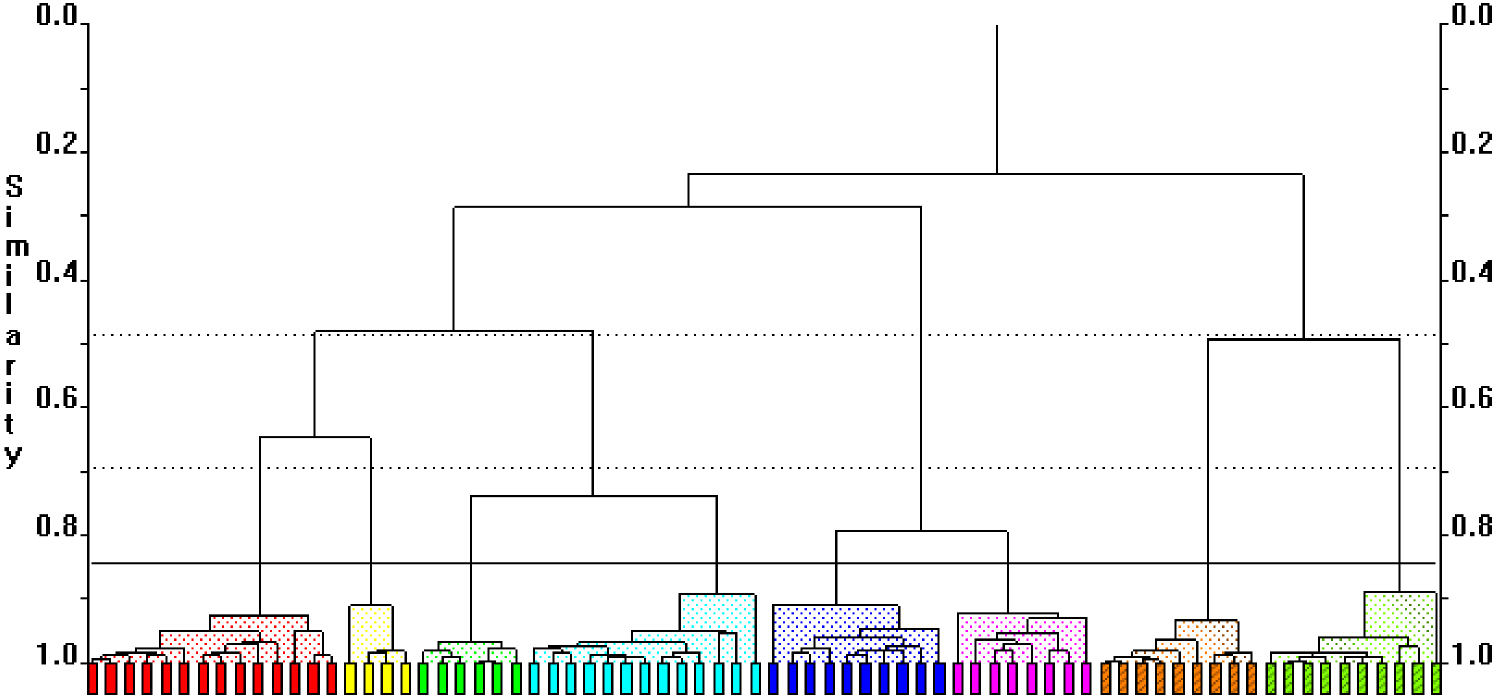
Sample 2: 113.44 58.328 153.602 … {x_{21}, x_{22} , x_{23}…….}

Sample 3: 117.873 60.117 93.686 … {x_{31}, x_{32} , x_{33}…….}

$$d_{ij} = \left(\sum_{k=1}^{m} w_k \left| x_{ik} - x_{jk} \right|^2 \right)^{\frac{1}{2}} \Rightarrow \mathbf{d}$$

Raman + Derivative Data





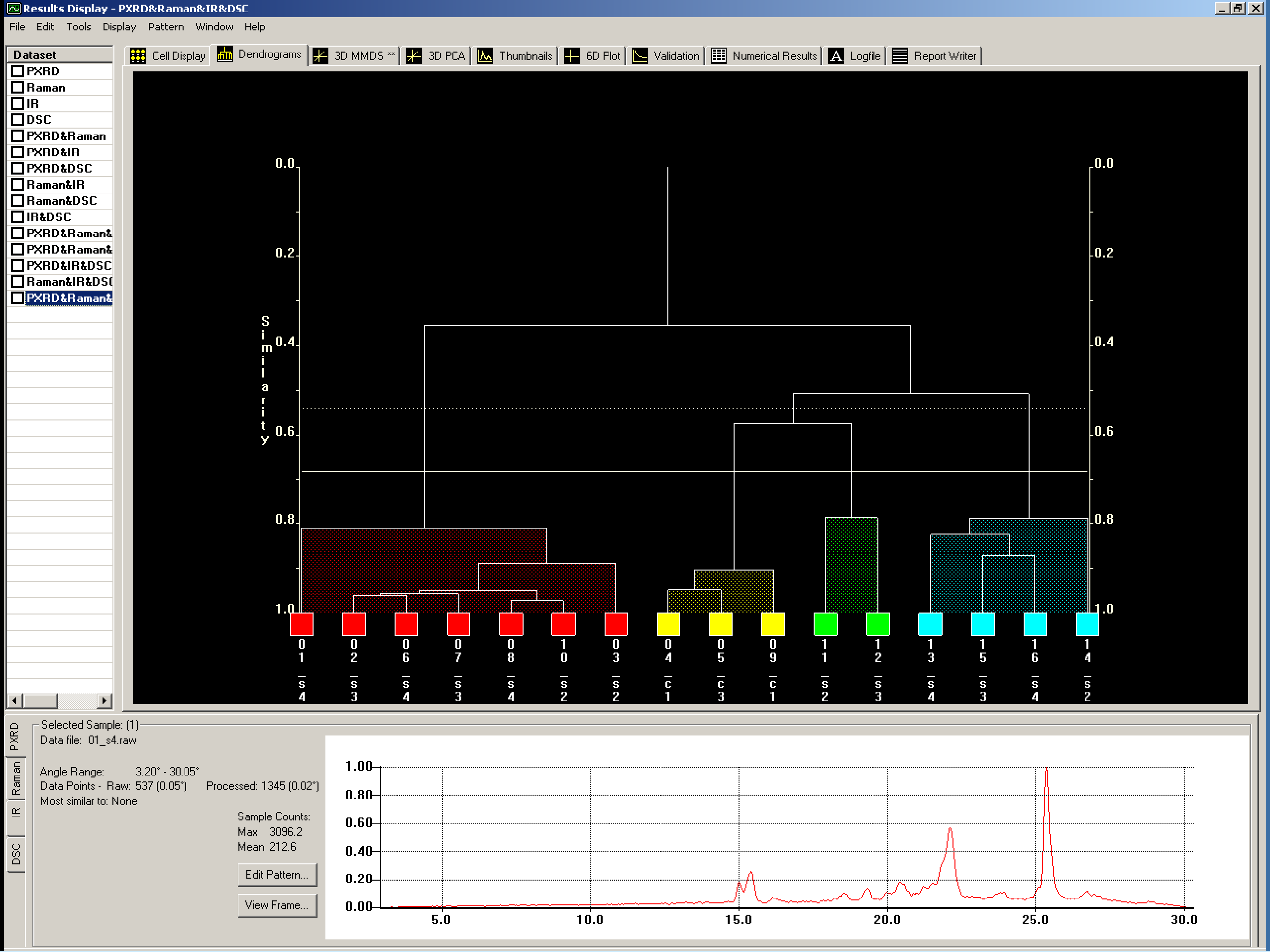
PXRD, Raman, IR + DSC

16 samples containing 3 forms of sulfathiazole and carbamazepine + mixtures:

- PXRD: Bruker C2 GADDS,
- IR: JASCO FT/IR 4100,
- DSC: TA instruments Q100.
- Raman: Renishaw inVia Reflex Spectrometer System.

PXRD, Raman, IR + DSC

- 15 different data sets and combinations!
- You need good software to explore all these options.

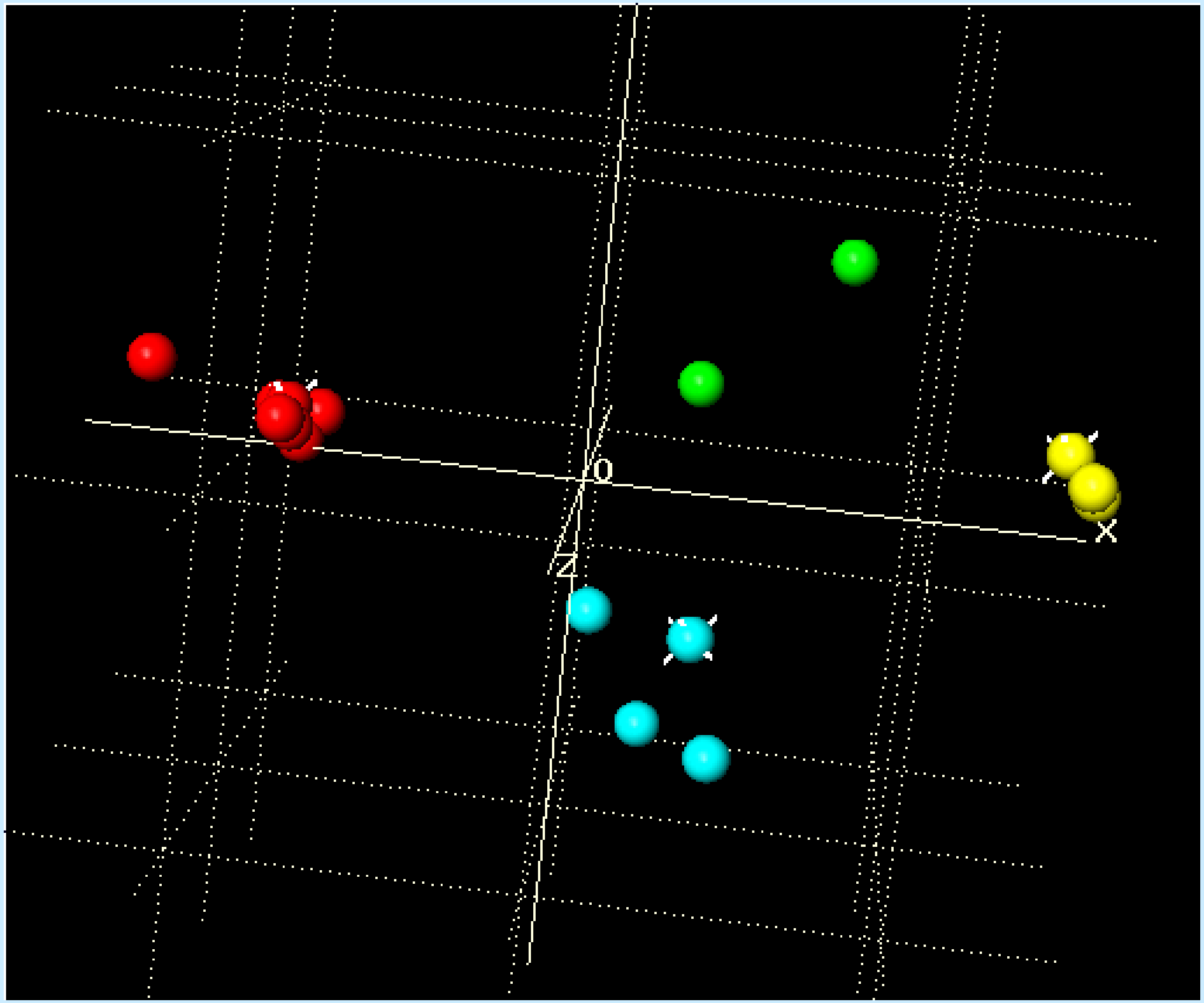


Selected Sample: (1)
Data file: 01_s4.raw

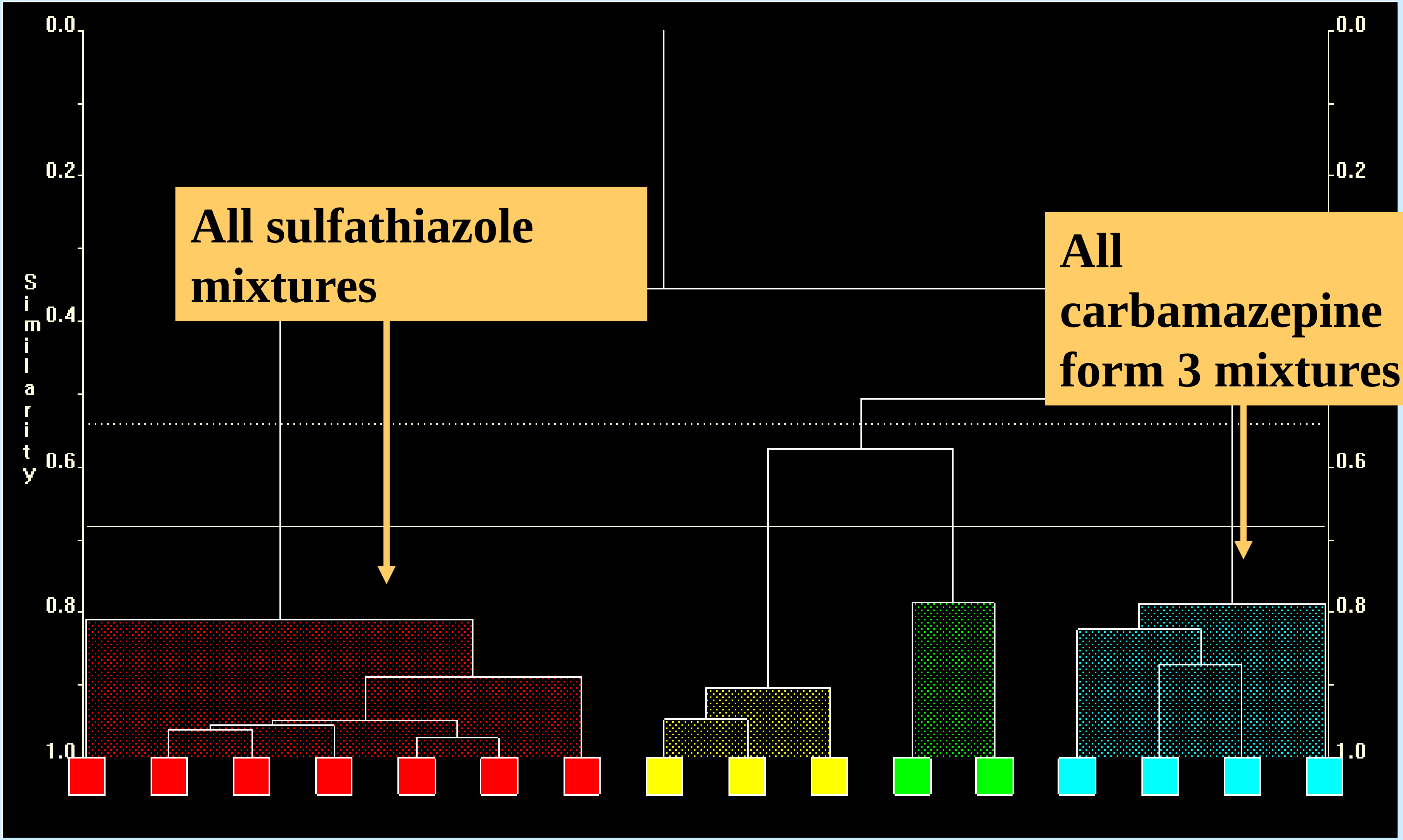
Angle Range: 3.20° - 30.05°
Data Points - Raw: 537 (0.05°) Processed: 1345 (0.02°)
Most similar to: None

Sample Counts:
Max 3096.2
Mean 212.6

PXRD + IR + Raman + DSC



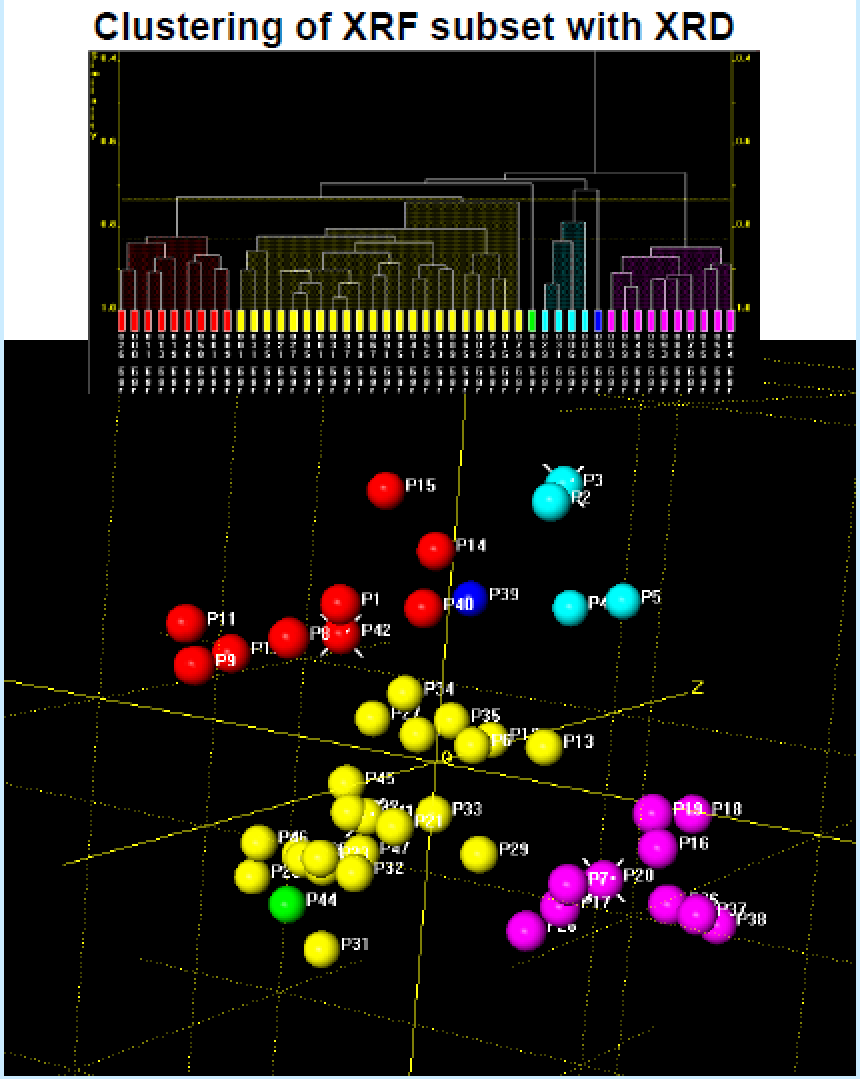
PXRD + IR + Raman + DSC



All sulfathiazole mixtures

All carbamazepine form 3 mixtures

PXRD + XRF



Lowe-Ma, et al.

Ford Motor Company

PolySNAP 3

You can do this with PolySNAP3

http://www.chem.gla.ac.uk/snap/PolySNAP_index.html

See:

Barr, Dong & Gilmore, *J. Appl. Cryst.*, (2009), 42, 965-974.

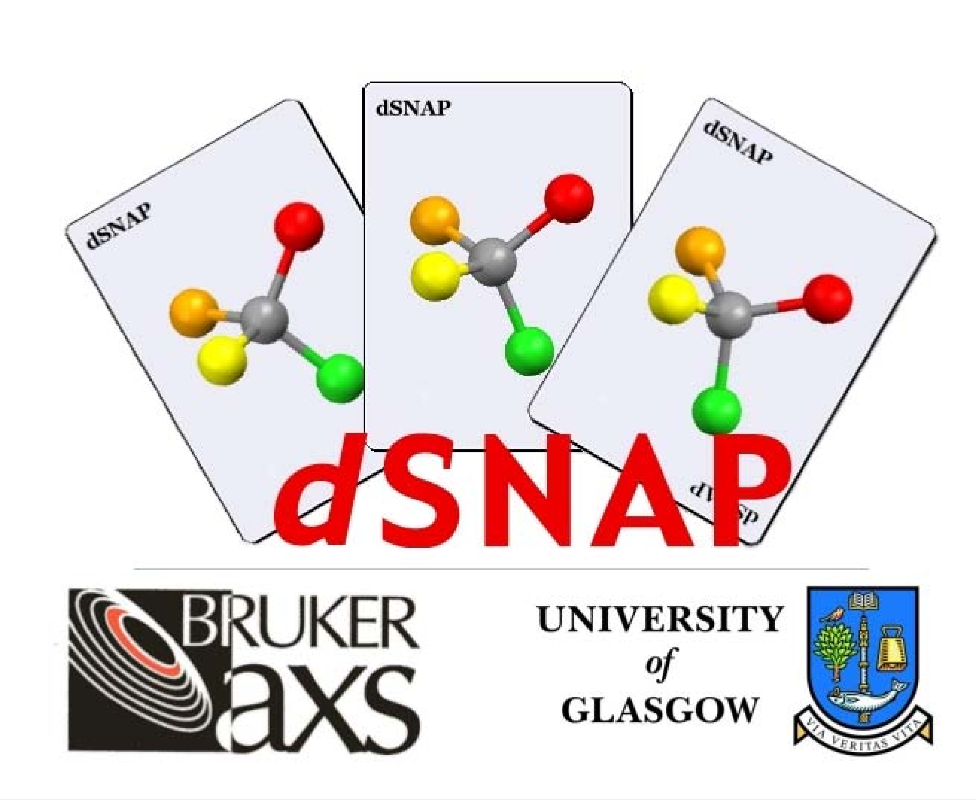
or

Use MATLAB/R/Sage

In progress

- Missing data.
- DSC processing
- Non numeric data

dSNAP: A New Way of Analysing the Results of Cambridge Data Base Searches

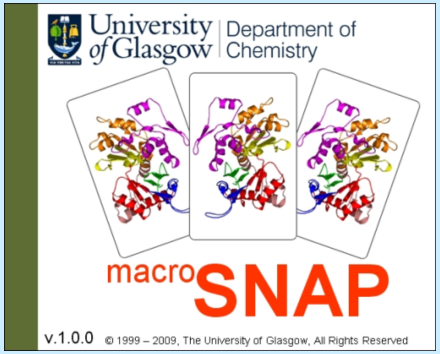


It’s free:

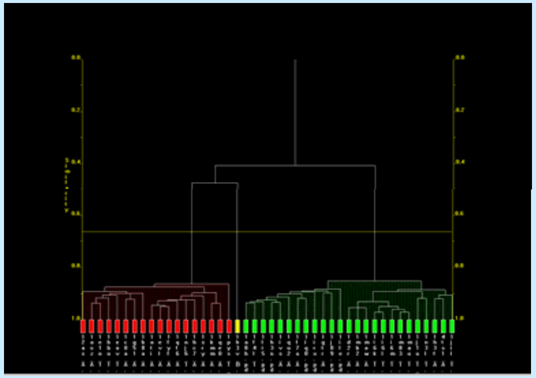
www.chem.gla.ac.uk/snap

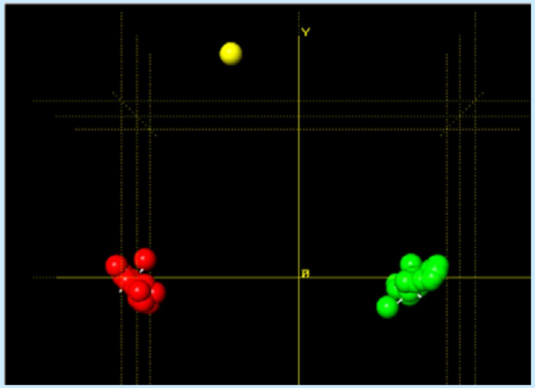
macroSNAP

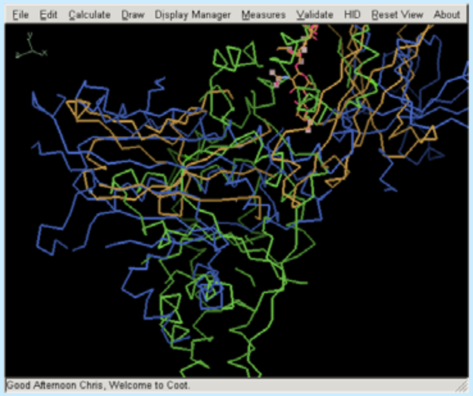
dSNAP for proteins – watch this space!



v.1.0.0 © 1999 – 2009, The University of Glasgow, All Rights Reserved







Good Afternoon Chris, Welcome to Coot.

Acknowledgements

- Michael Hermannn (Fraunhofer-Institut Chemische Technologie)
- Karsten Knorr & Arnt Kern (Bruker AXS, Karlsruhe)
- Chris Frampton & Susie Buttar, Pharmorphix