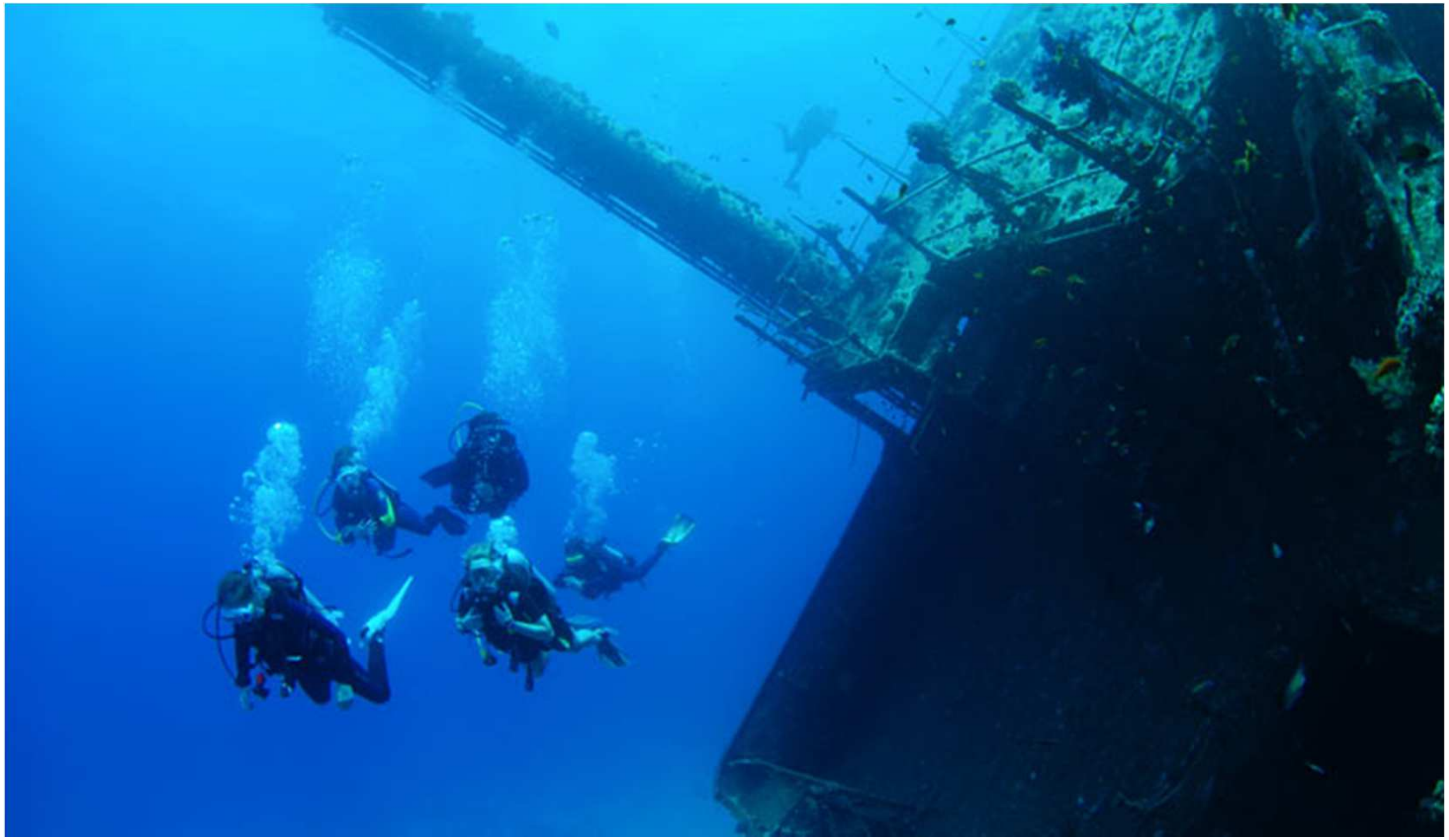
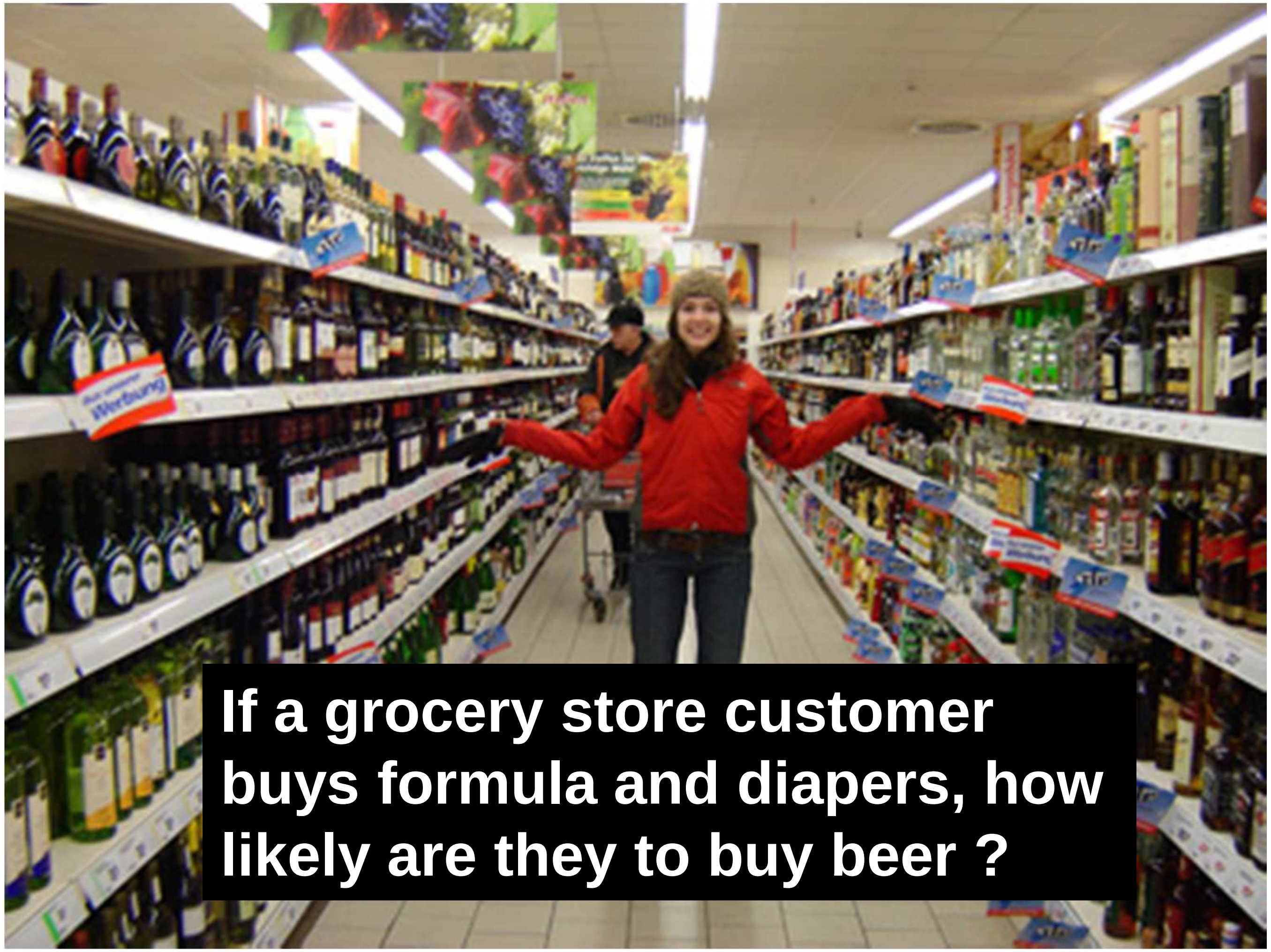


Diving For Treasure in Complex Data

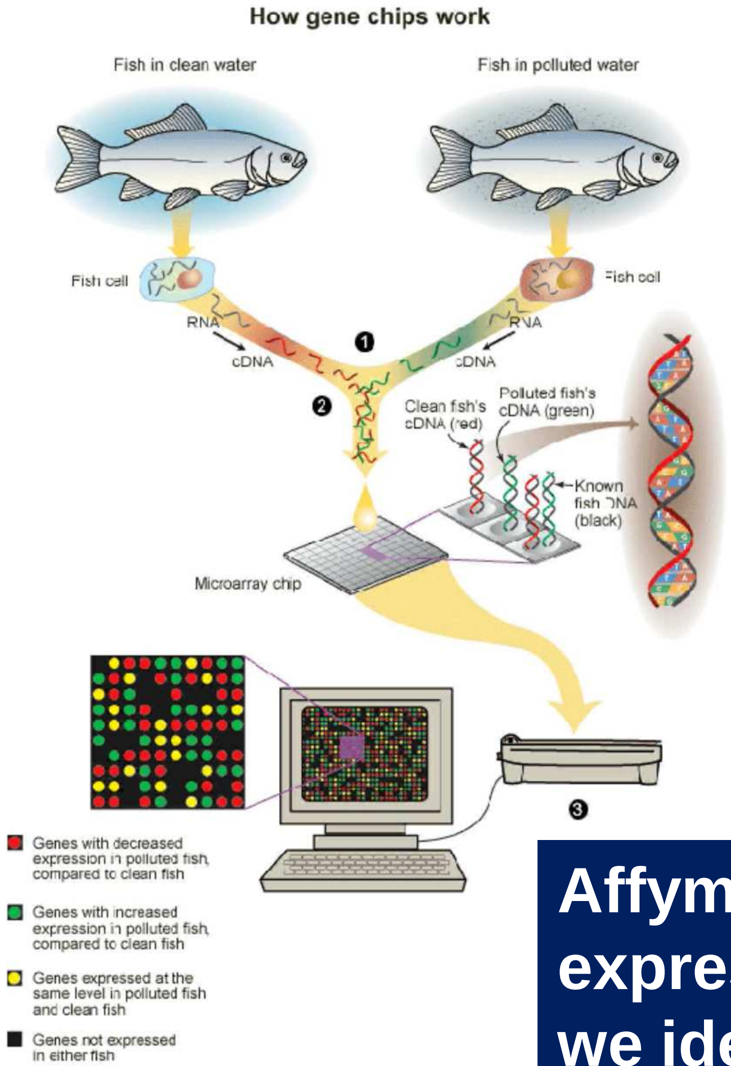
From Roman Urns To Mid-East Earthquakes

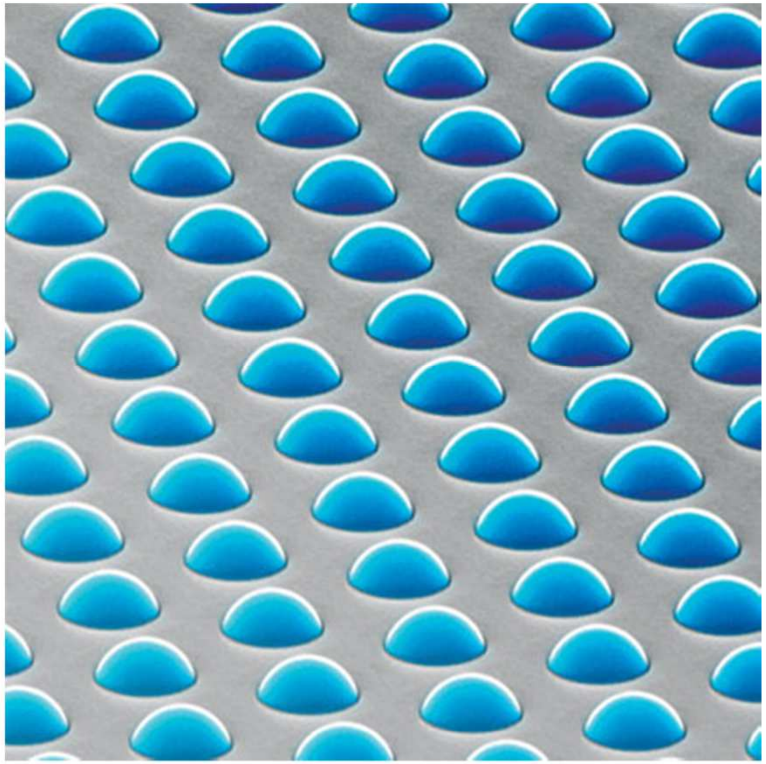
Marvin Weinstein and David Horn

What's The Problem ?



Biology and Medicine

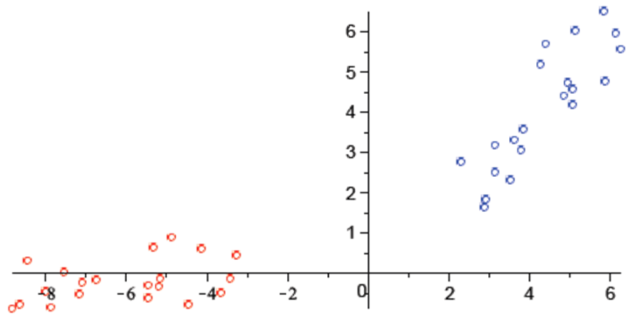




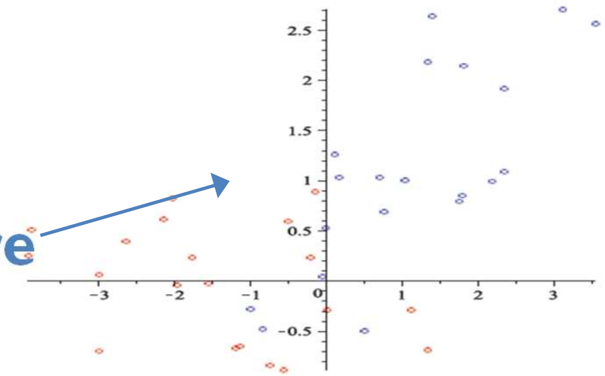
Affymetrix chip measures the expression of ~7000 genes. Can we identify types of Leukemia from gene expression alone ?

What Is Clustering ?

- Plot the data and look for places where there are "more" things near one another.
  - If clusters are obvious, then no problem !

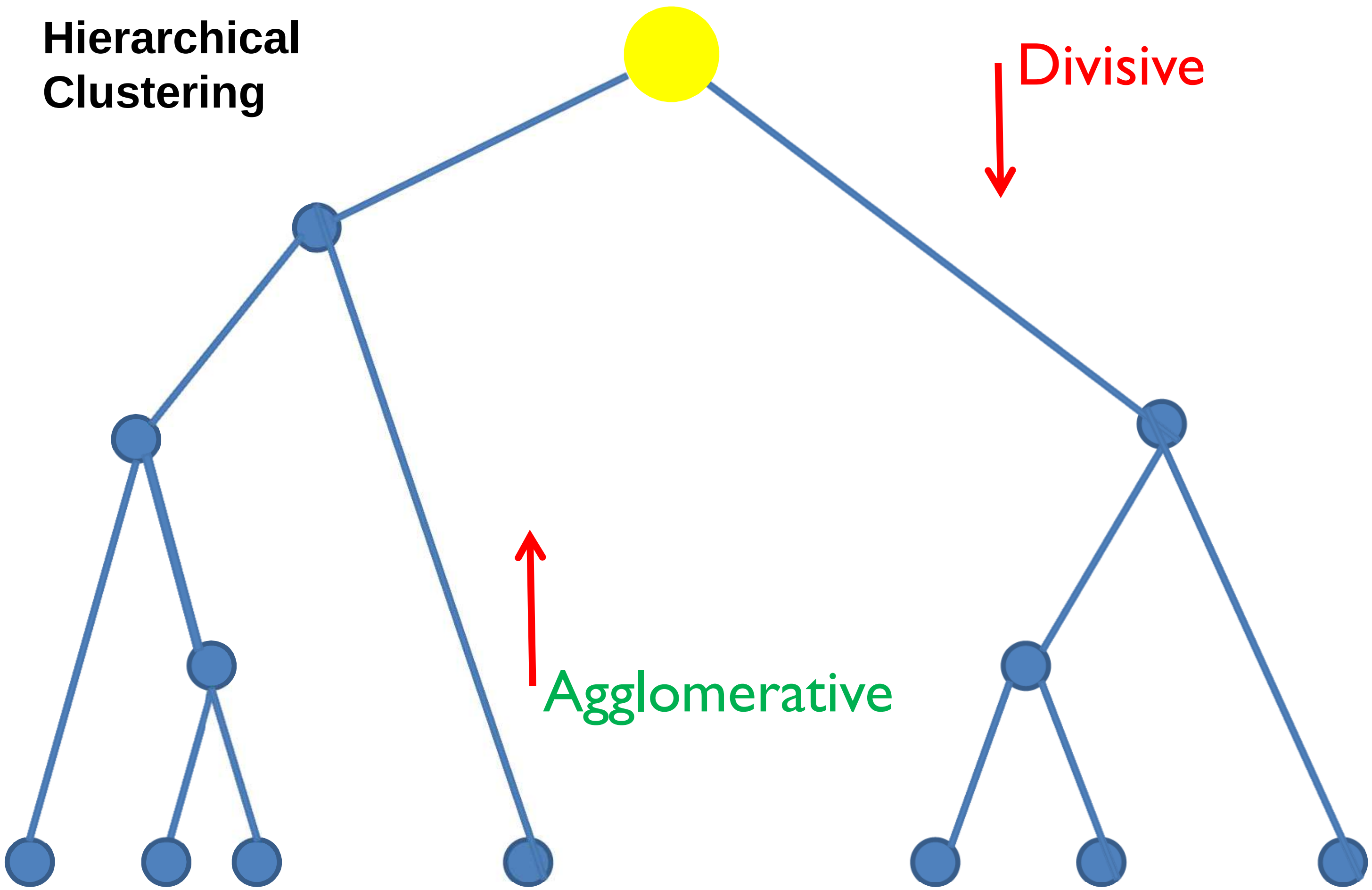


  - But usually things are more like this



  - Actually we need more than simple clustering algorithms! We now know complex data shows complex structures!!!

Most Clustering Algorithms Make Assumptions



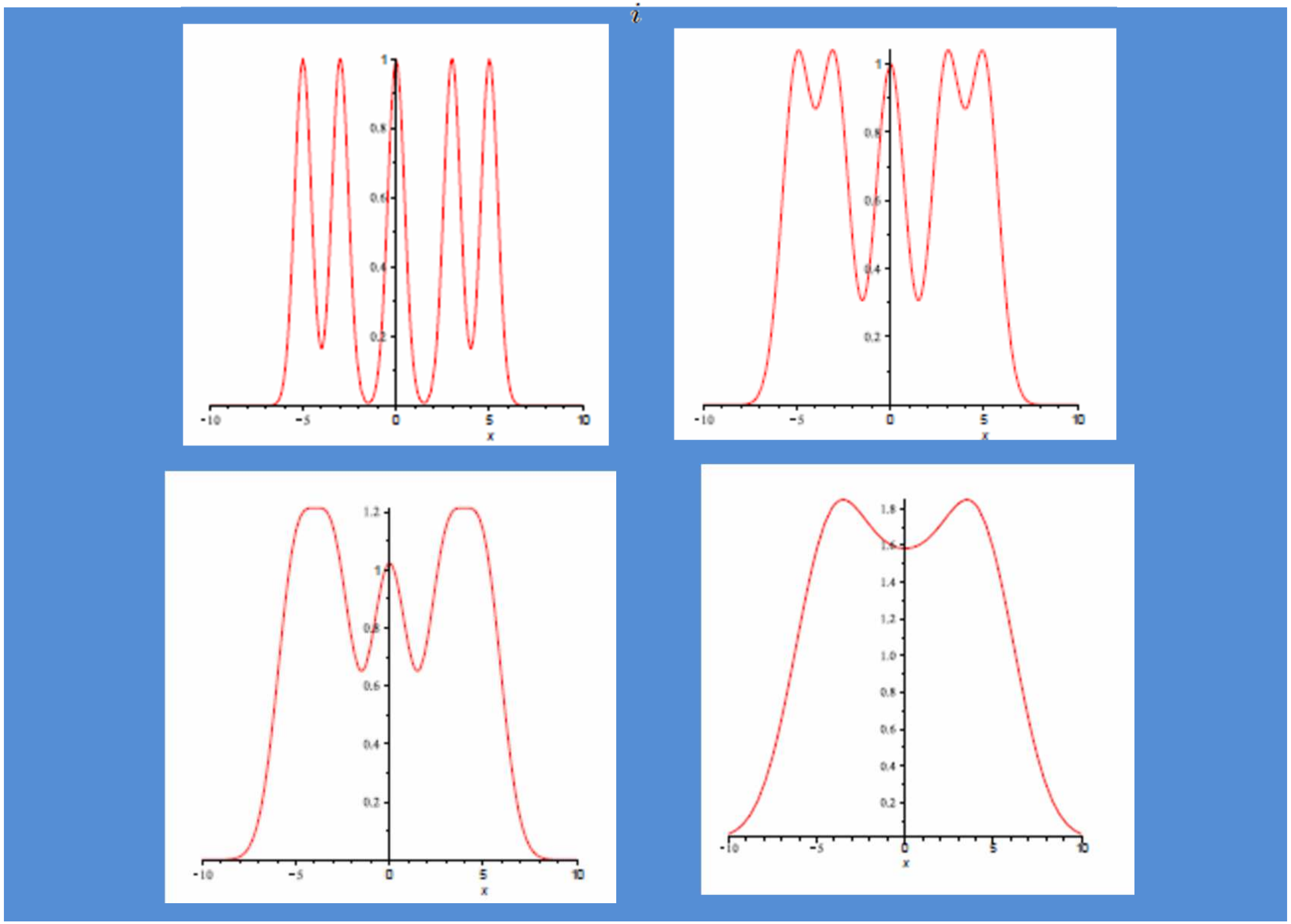
DQC – dynamic quantum clustering

- **DQC is very different from existing methods: it doesn't make assumptions!**

  - **1: DQC maps the problem into a problem in quantum mechanics**
  - **2: Then DQC uses the properties of the quantum problem to have both ordinary clusters and extended structures form dynamically (we find extended structures to be common in complex data).**
  - **3: DQC is highly visual and well suited to exploration and discovery**

Density Based Clustering
The Parzen Window Estimator

$$|\Psi\rangle = \sum_i e^{-\frac{1}{2}\gamma(x-x_j)^2}$$



A Quantum Potential As A Proxy For Density

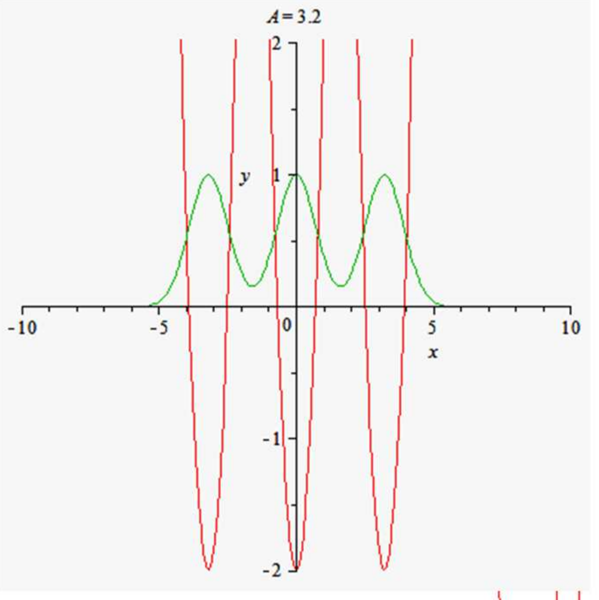
- **Amplifying peaks and valleys…**

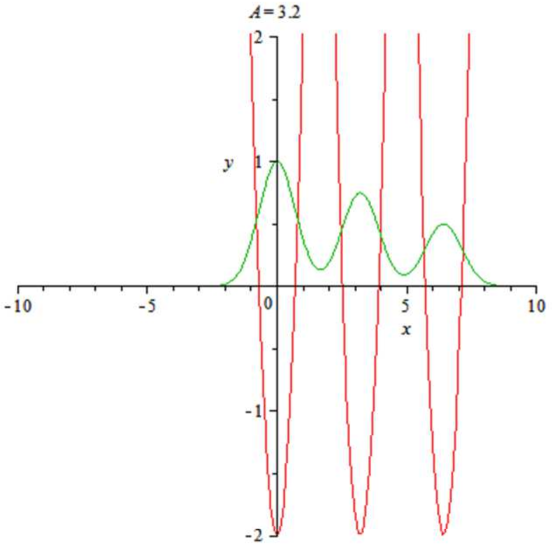
  - **Given: A wave-function that is a sum of Gaussians. Quest: Is there a potential for which it is the ground-state of the Schrodinger equation ?**
  - Yes!!!

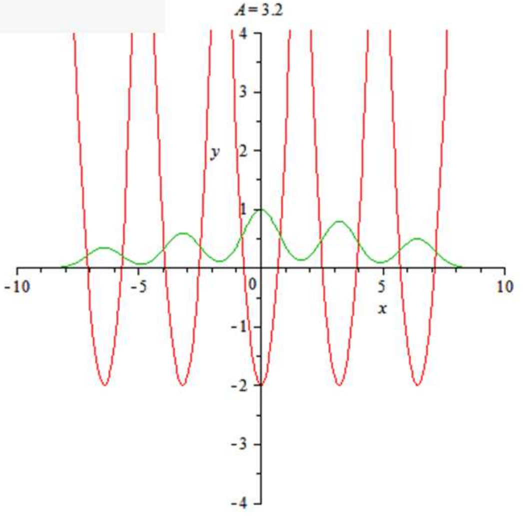
$$-\frac{1}{2}\nabla^2\psi(\vec{x}) + V(\vec{x})\,\psi(\vec{x}) = 0$$

$$V(\vec{x}) = \frac{\nabla^2\psi(\vec{x})}{2\,\psi(\vec{x})}$$

Parzen vs Potential







Green is the sum of Gaussians

Red is the potential

The Potential For Crabs In 2-d

- **The Crab problem – Once upon a time there was a museum which had a display case with 200 crab carapaces.**

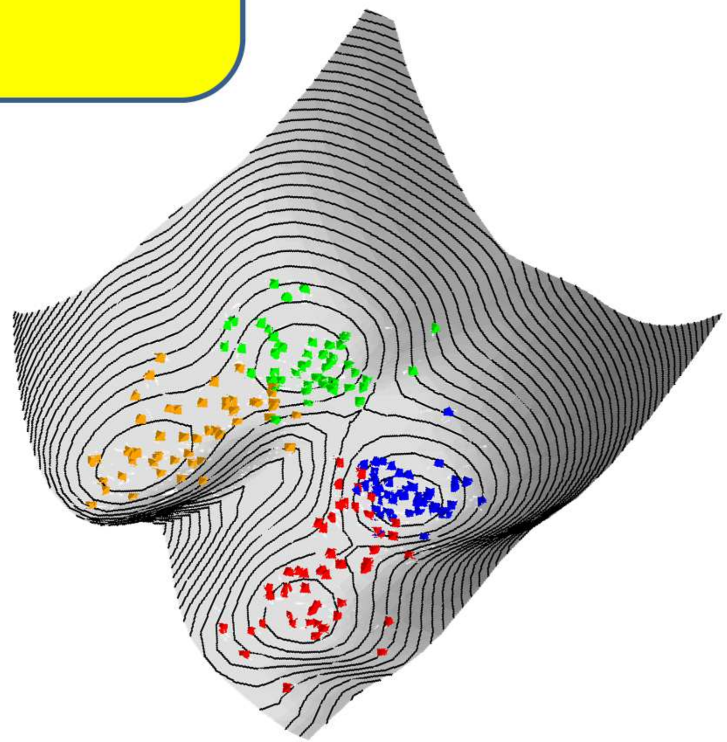
  - **The crabs were distinguished by color, male female and one of two species**
  - **But the shells sat in the sun for many years and faded**

  - **So, in an attempt to reclassify them, they made 5 measurements of size of shell and claws.**

The Potential For Crabs In 2-d

- Procedure:
  - **Create the Parzen window estimator (a sum of Gaussians in 2-dim)**
  - **Form the potential function**

- **Result: The minima capture the clusters to a high degree.**
  - **Problem: In high dimensions, finding minima and moving points is difficult.**



Strategy: Roll The Points Downhill

- **We have a potential function**
- $$V(\vec{x}) = \frac{\nabla^2 \psi(\vec{x})}{2\psi(\vec{x})}$$

- **Each component wave-function is centered on the original data point** $\vec{x}_\alpha$ **.**
  - **Thus** $\vec{x}_\alpha = \langle \psi | \vec{x} | \psi \rangle$

$$\left| \psi_t \right\rangle = e^{-itH} \left| \psi \right\rangle$$

$$\langle \vec{x} \rangle (t) = \left\langle \psi_t \right| \vec{x} \left| \psi_t \right\rangle$$

$$\frac{d^2 < \vec{x} > (t)}{dt^2} = - \left\langle \psi_t \right| \vec{\nabla} V(\vec{x}) \left| \psi_t \right\rangle$$

Key Equations

- The full set of equations

$$H = \frac{p^2}{2m} + V(x)$$

Note m

$$H_{ij} = \langle \psi_i | H | \psi_j \rangle$$

$$N_{ij} = \langle \psi_i | \psi_j \rangle$$ Metric

$$\vec{X}_{ij} = \langle \psi_i | \vec{x} | \psi_j \rangle$$

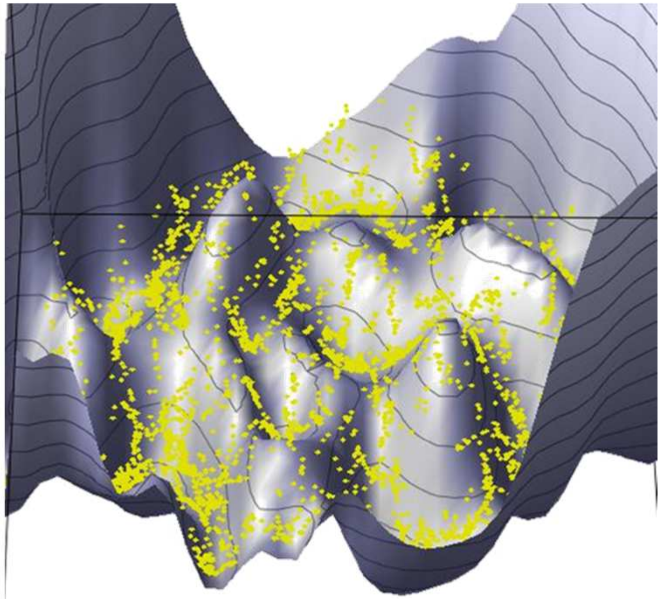
- Exponentiate the finite matrix in the orthonormal basis and thus compute the time evolution of the expectation values of the position operators.

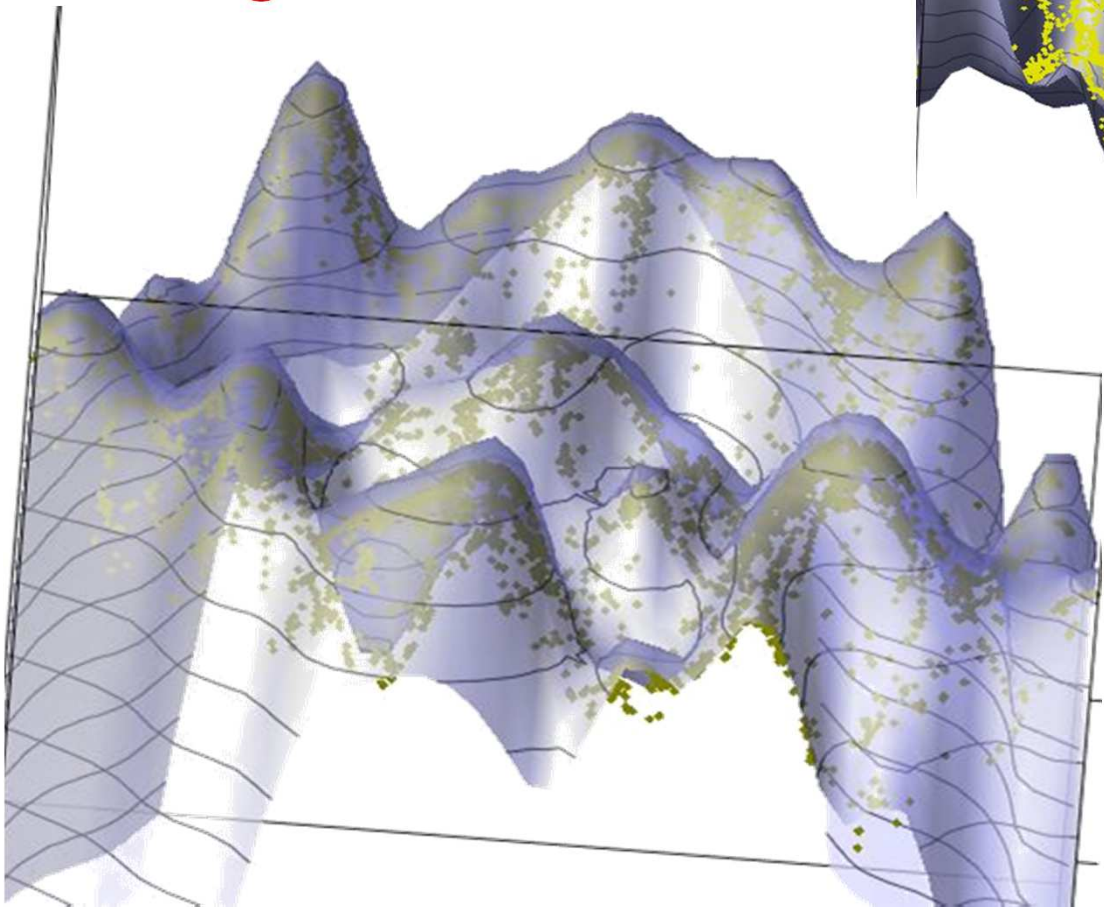
More Complex Data

- Let us take a look at data we understand:
  - We are looking at a slice of Sloan Digital Sky Data for a narrow range in the z coordinate (redshift mapped to z) and plotting the potential it creates
  - As with the crabs, we put the actual points on the potential

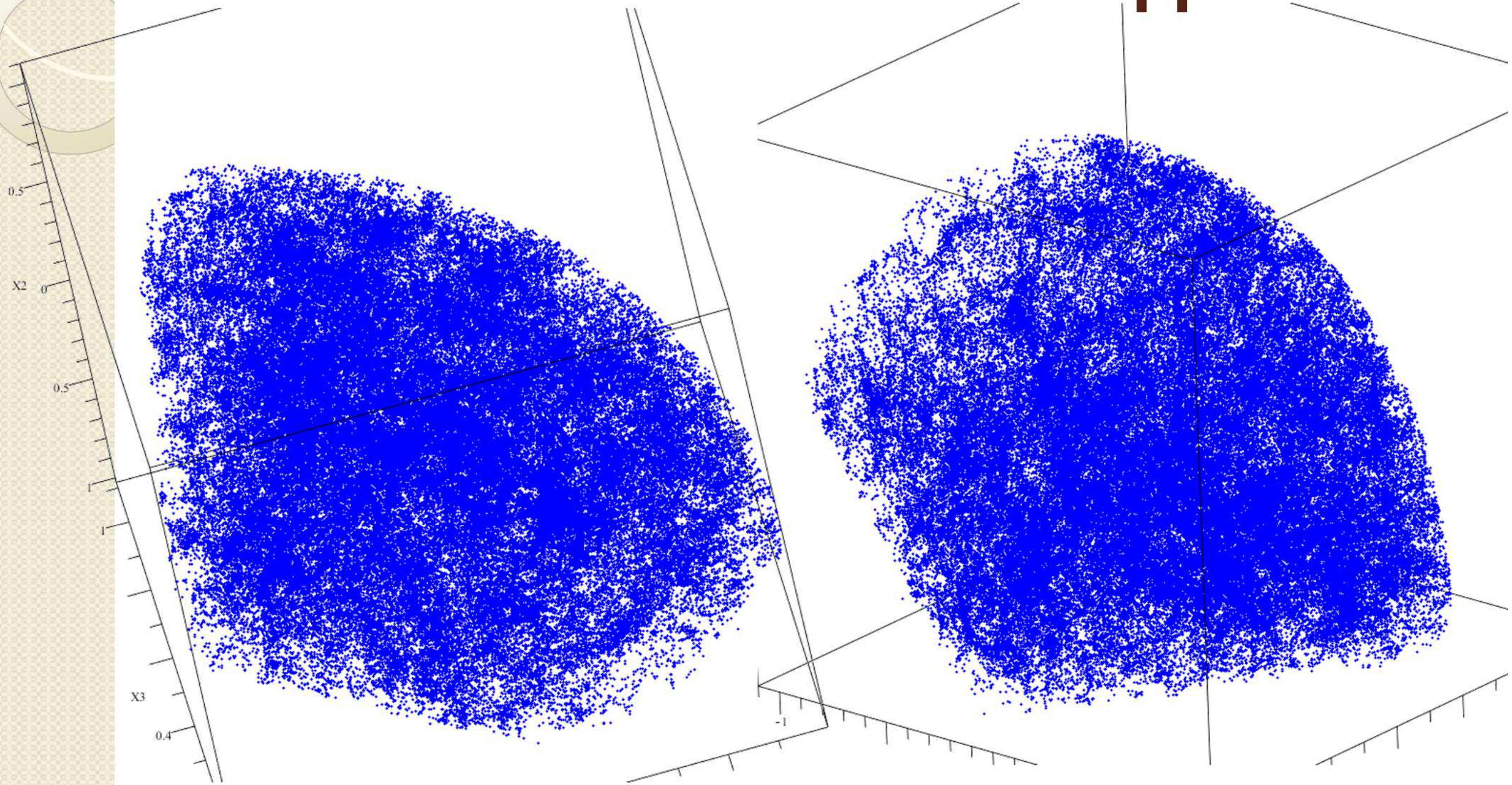
Sloan Data

- Take a thin slice in z
  - Sigma = 0.1





Making The High Density Regions Of The 3d Sloan Data More Apparent

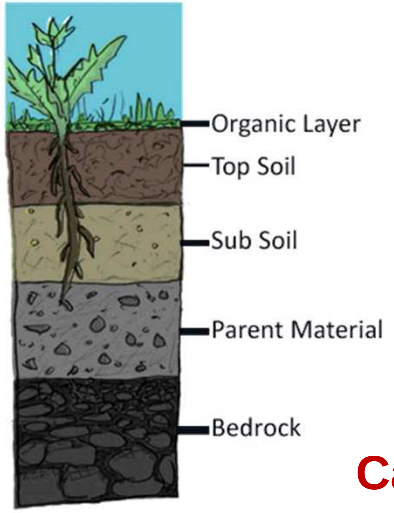


Beyond Simple Clustering
TXM Xanes meets DQC

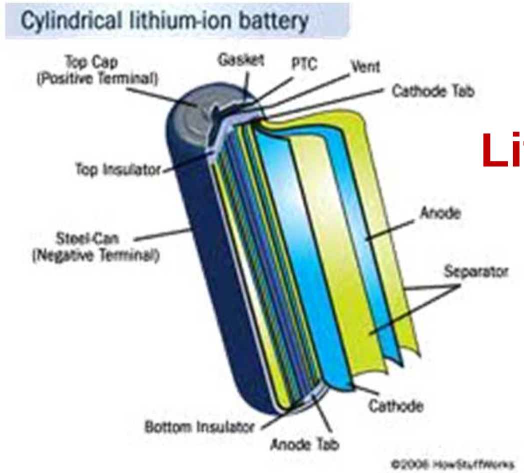
Marvin Weinstein
Apurva Mehta
Florian Meirer
Allison Hume (SULI student)

Hierarchically Heterogeneous Materials

Soils



Lithium Ion Batteries



Biomaterials - Bone

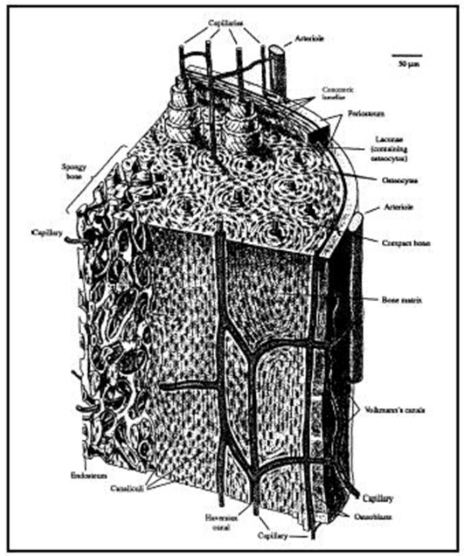
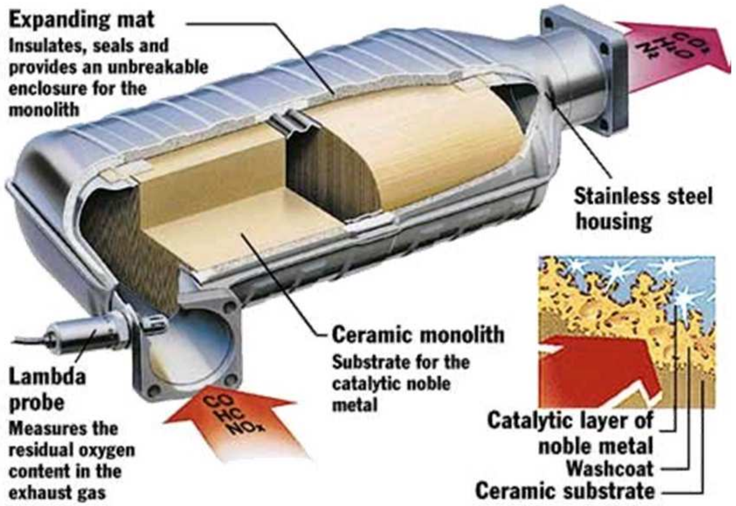


Fig. 8.23. Internal cellular structure of bone (redrawn from Guyton[863]).

Catalysts



Proto-Sigilatta: Black Gloss Pottery

Advanced Material of 2000 years ago

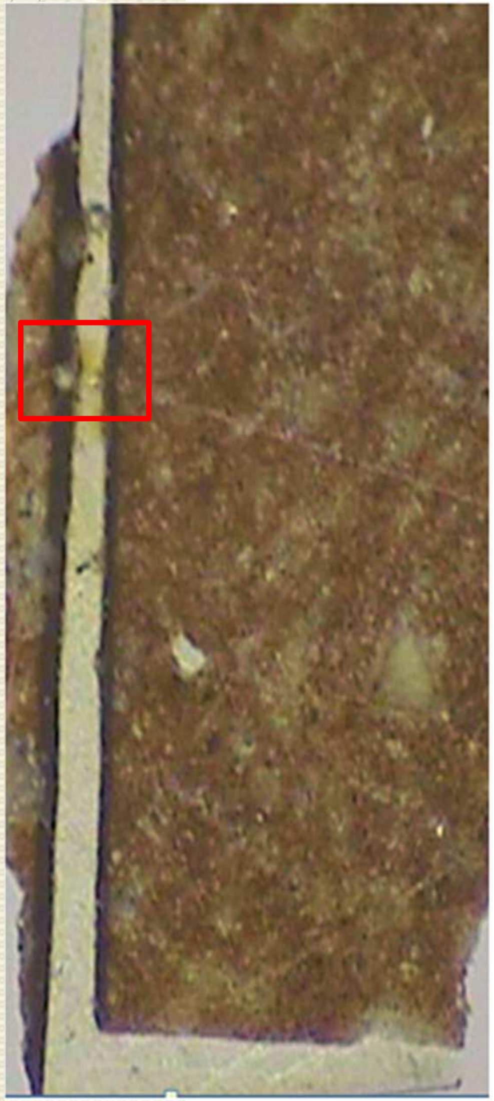
Red Body, Black Slip

Both colors from oxides of Fe

How does one keep the body Red/oxidized and the Slip Black/reduced?

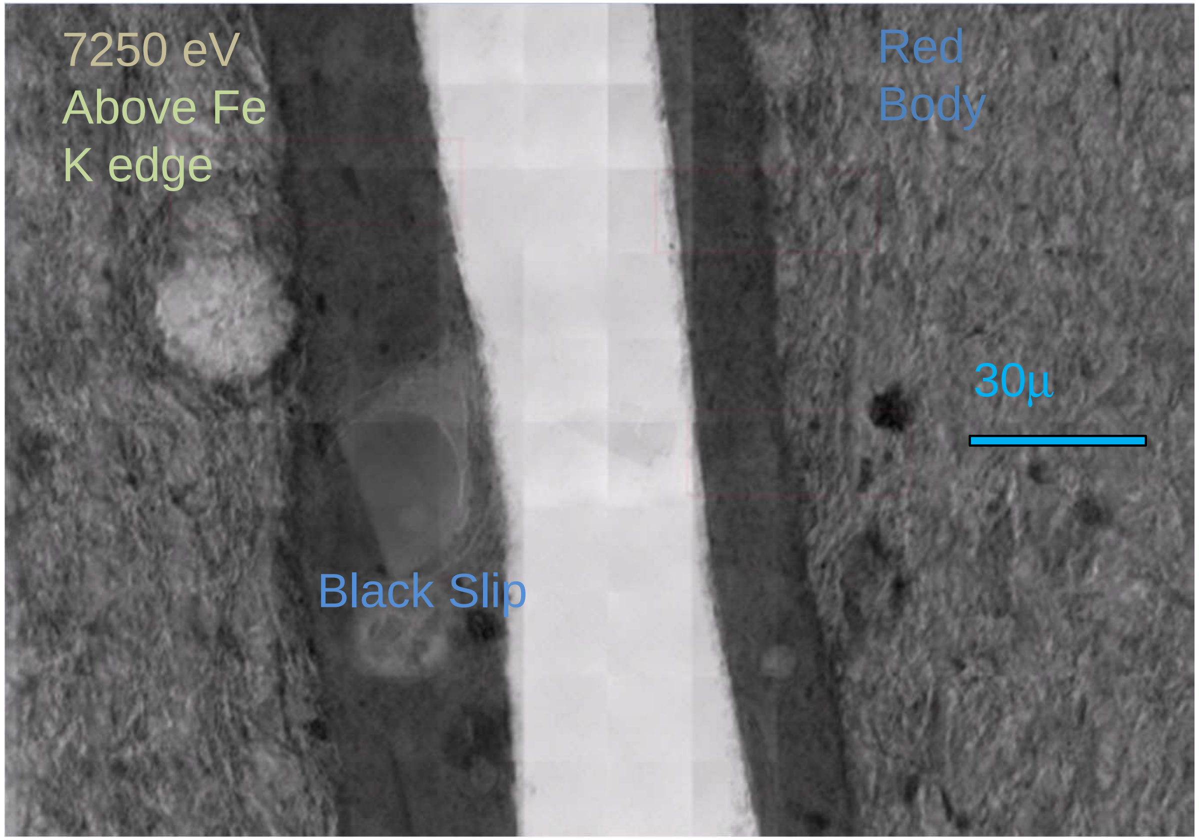
One of the Benchmark Technological Advancement

Transmission X-ray Microscopy
30 nm resolution



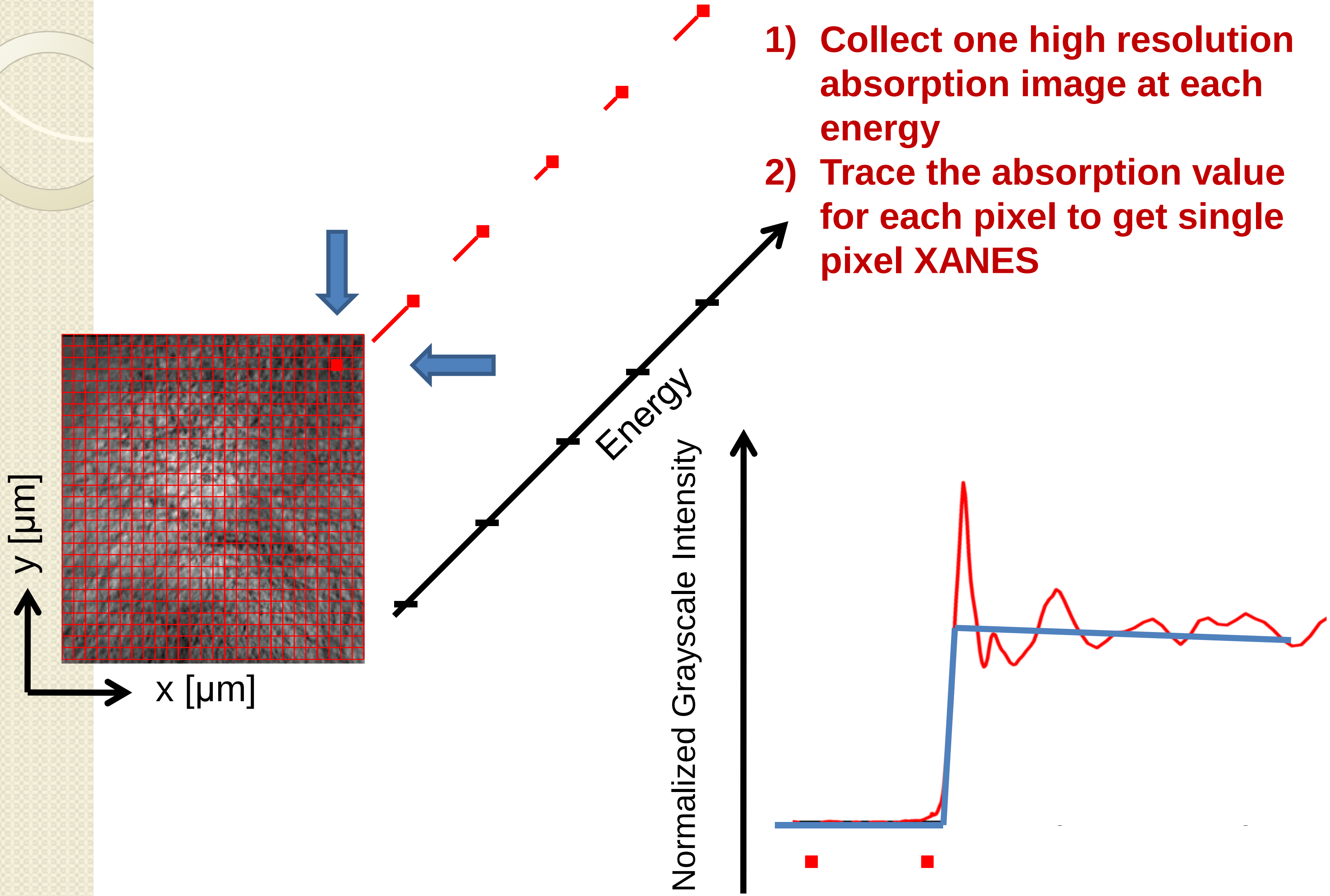
Southern Gaul, 1st cent BCE

From the La Graufesenque workshop

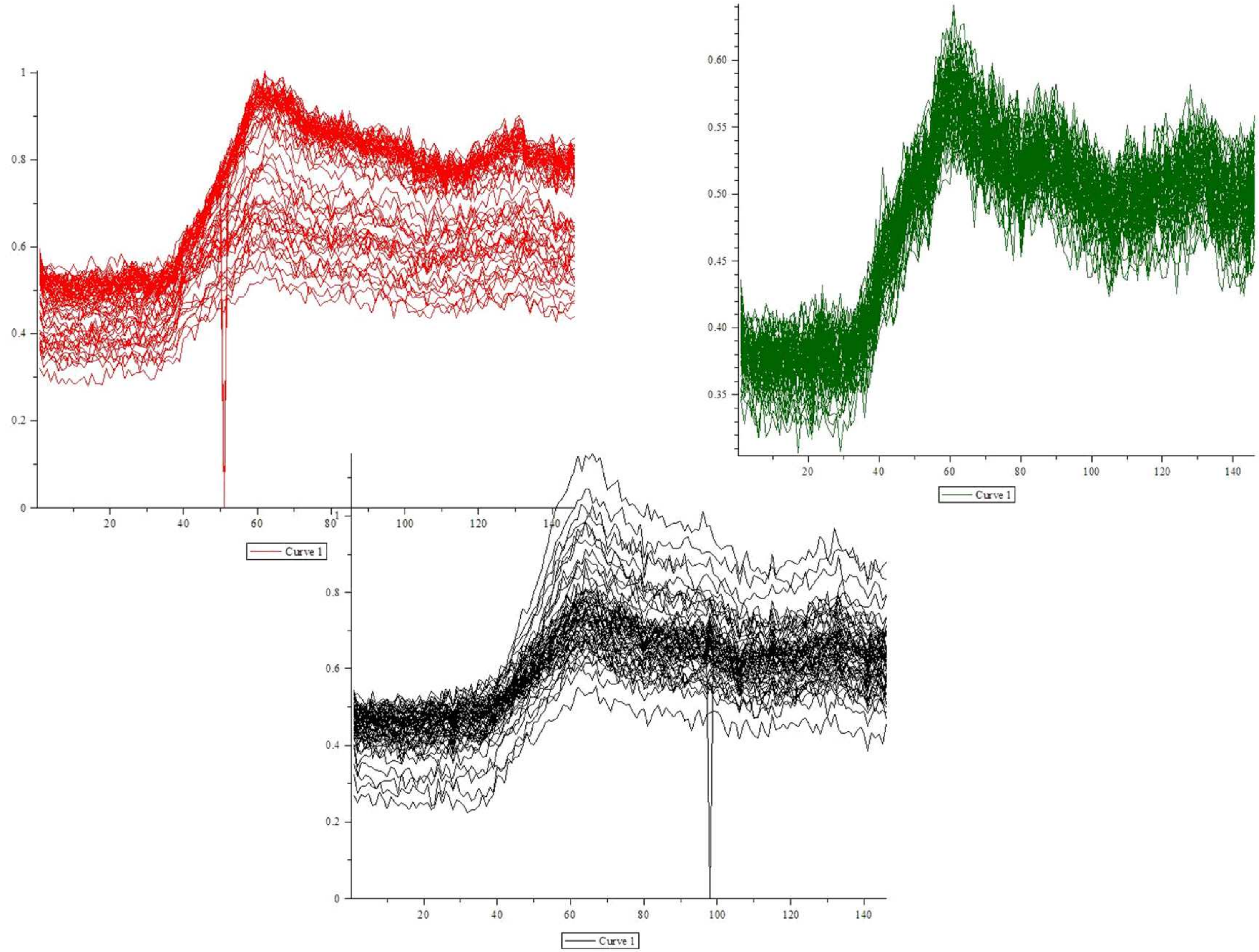


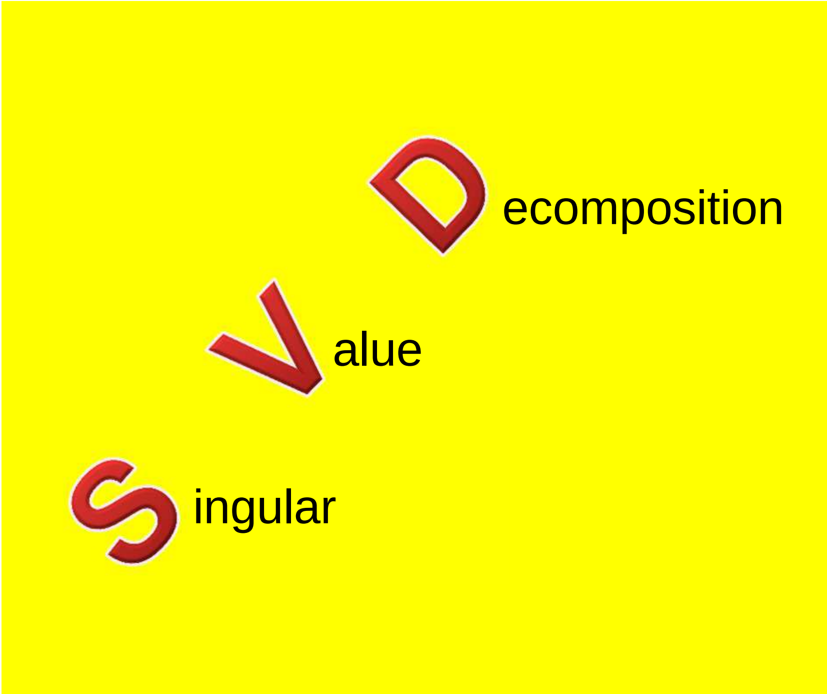
TXM-XANES: basic idea

1) Collect one high resolution absorption image at each energy
2) Trace the absorption value for each pixel to get single pixel XANES



What Spectra Look Like





The Swiss Army Knife of data mining
A trick to find good coordinates!

What is SVD?

- **What is the best way to look at data when there are *n* samples and *m* features measured ?**
  - **Consider an n x m – matrix M. The SVD decomposition of this matrix writes the matrix as:**

$$M_{ij} = U_{i\alpha} S_{\alpha} V^{tr}_{\alpha j}$$

  - **Where U is n x n**
    V is m x m
    S is n x m and only has non-vanishing stuff on diagonal

Data Compression

- **Consider a matrix M which is a picture**
- **The SVD decomposition lets us write M as**

$$M_{ij} = \sum_{\alpha=1}^{\min(m,n)} \lambda_\alpha T_{ij}^\alpha$$

$$\sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij}^2 = Tr(\, M M^{tr}) = \sum_{\alpha=1}^{\min(m,n)} \lambda_\alpha^2$$

- **Define approximation:**

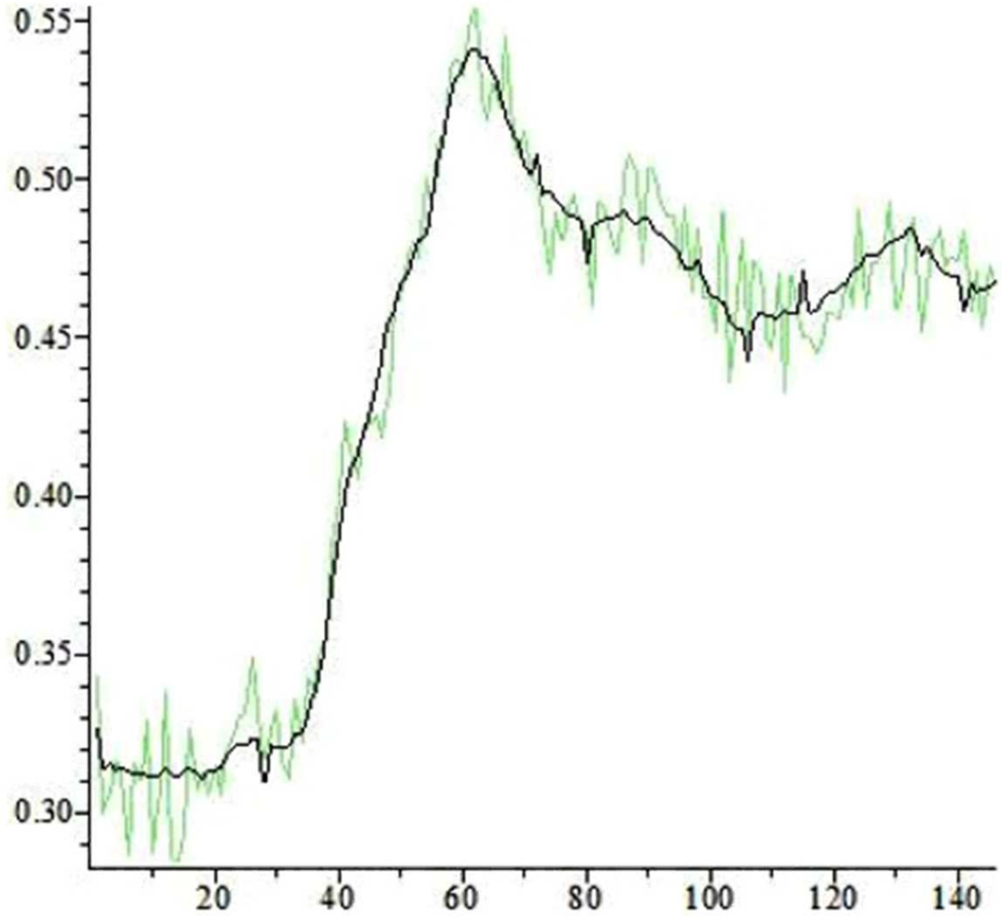
$$\tilde{M} = \sum_{\alpha=p}^{\min(m,n)} \lambda_\alpha T^\alpha$$

$$Tr\left((M - \tilde{M})^+ (M - \tilde{M}) \right) = \sum_{\alpha=p}^{\min(m,n)} \lambda_\alpha^2$$

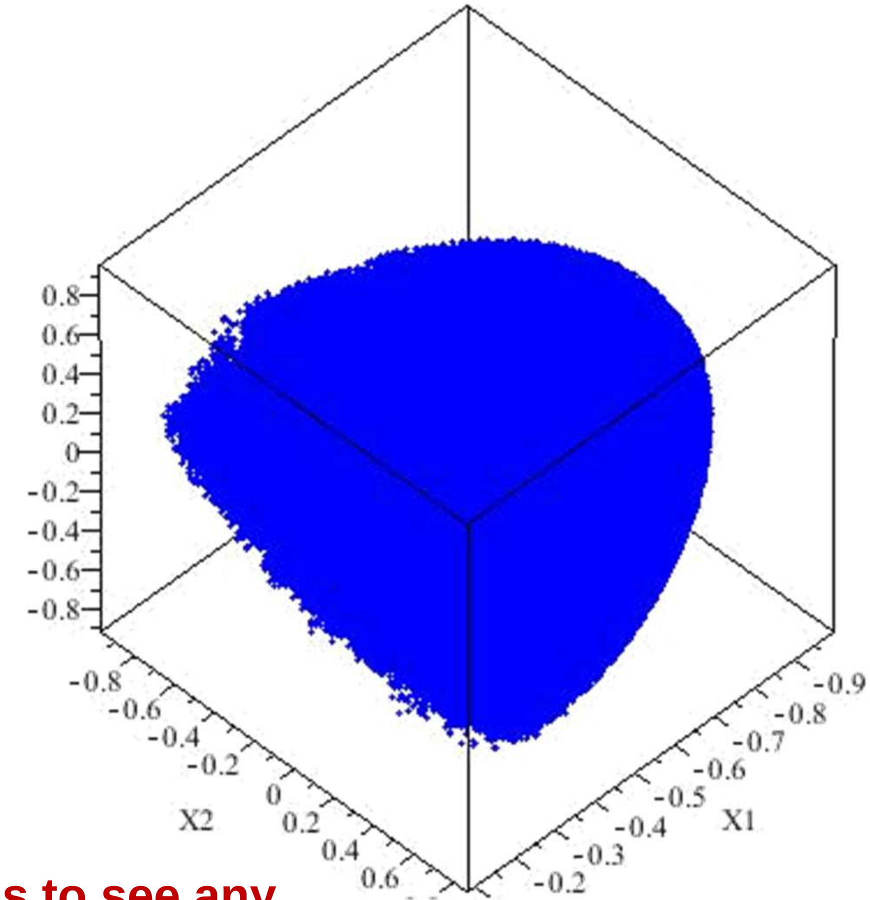
How well does this work ?

Any data matrix is a picture, or at least like a picture!

Sample Spectrum After SVD



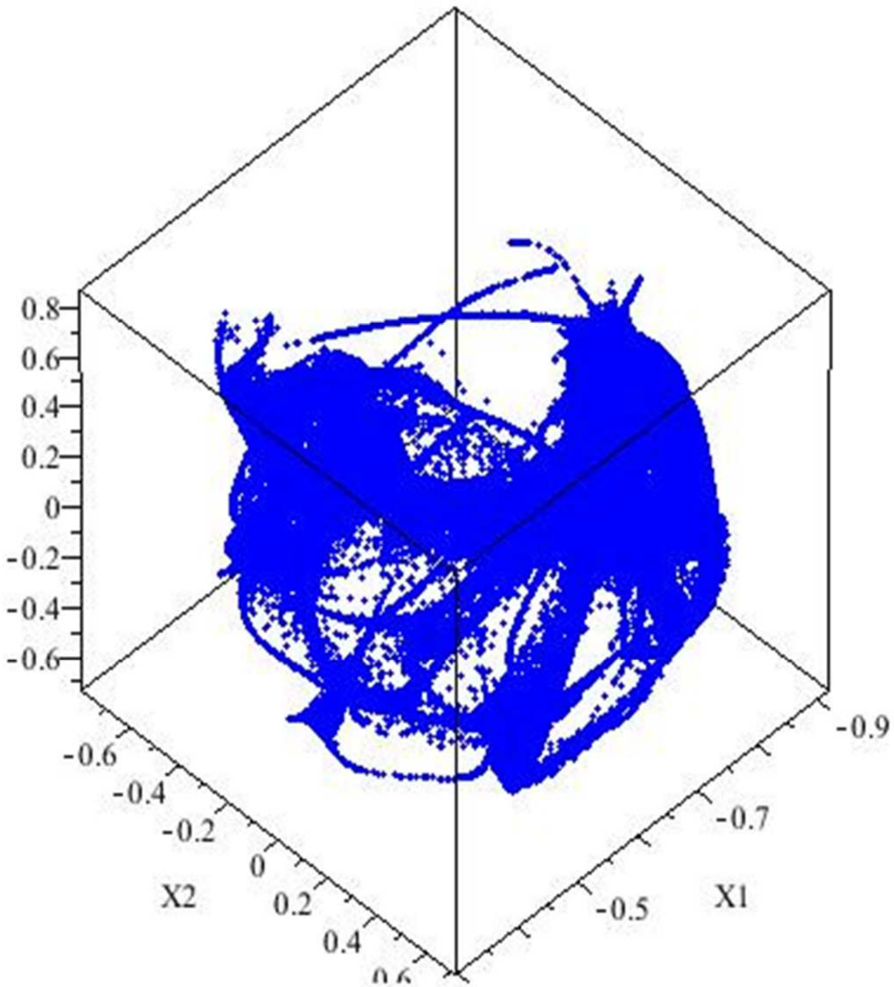
SVD Data At The Outset



K-Means fails to see any structure

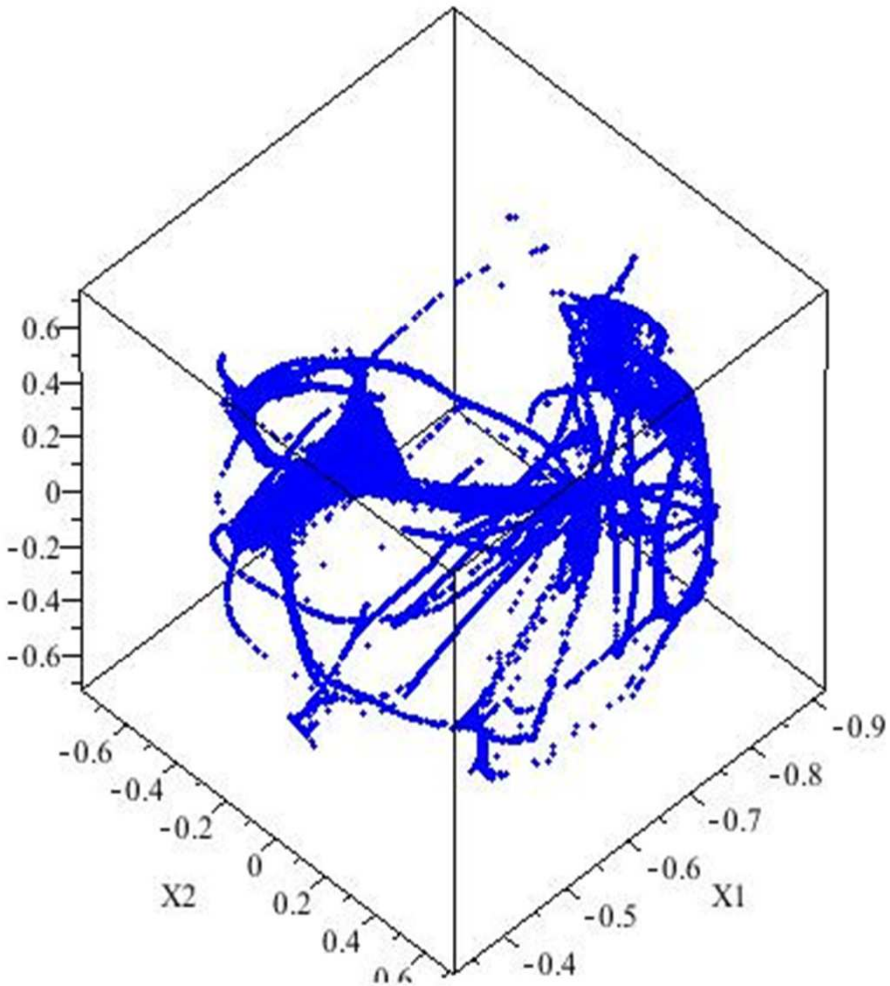
Clustering Process

Data collapses into clumps and strands



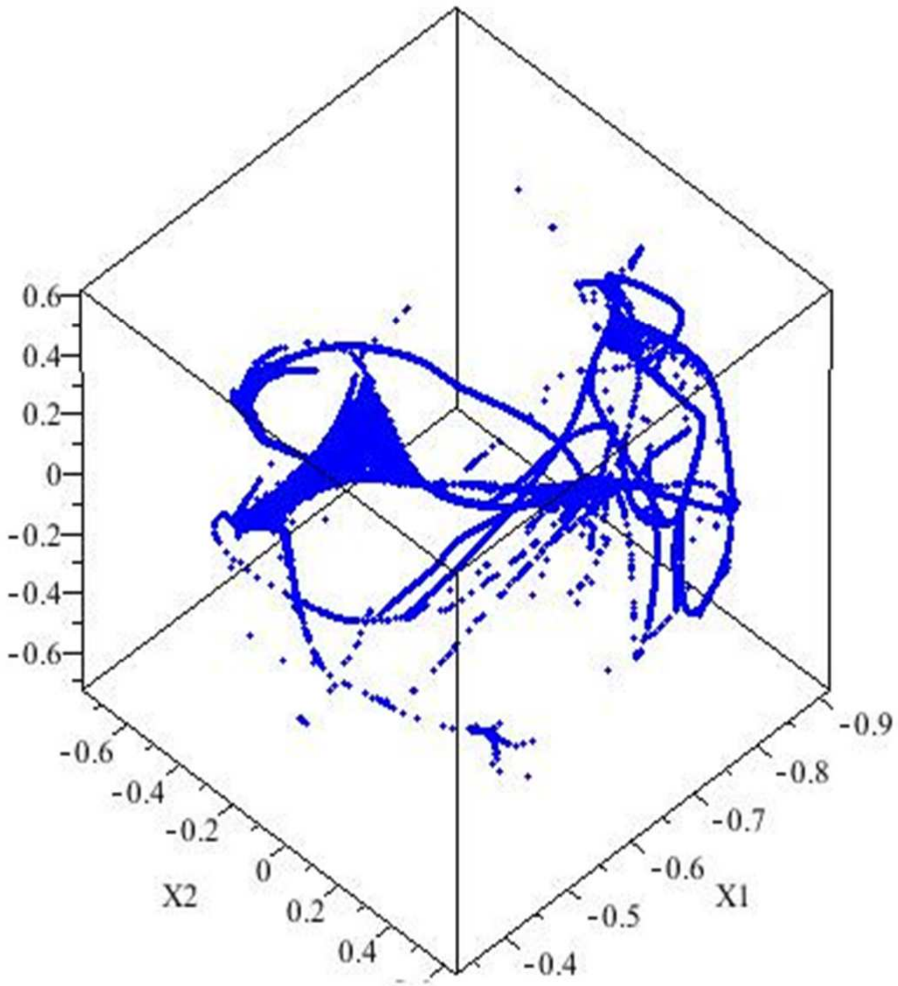
Clustering Process

Data collapses into clumps and strands



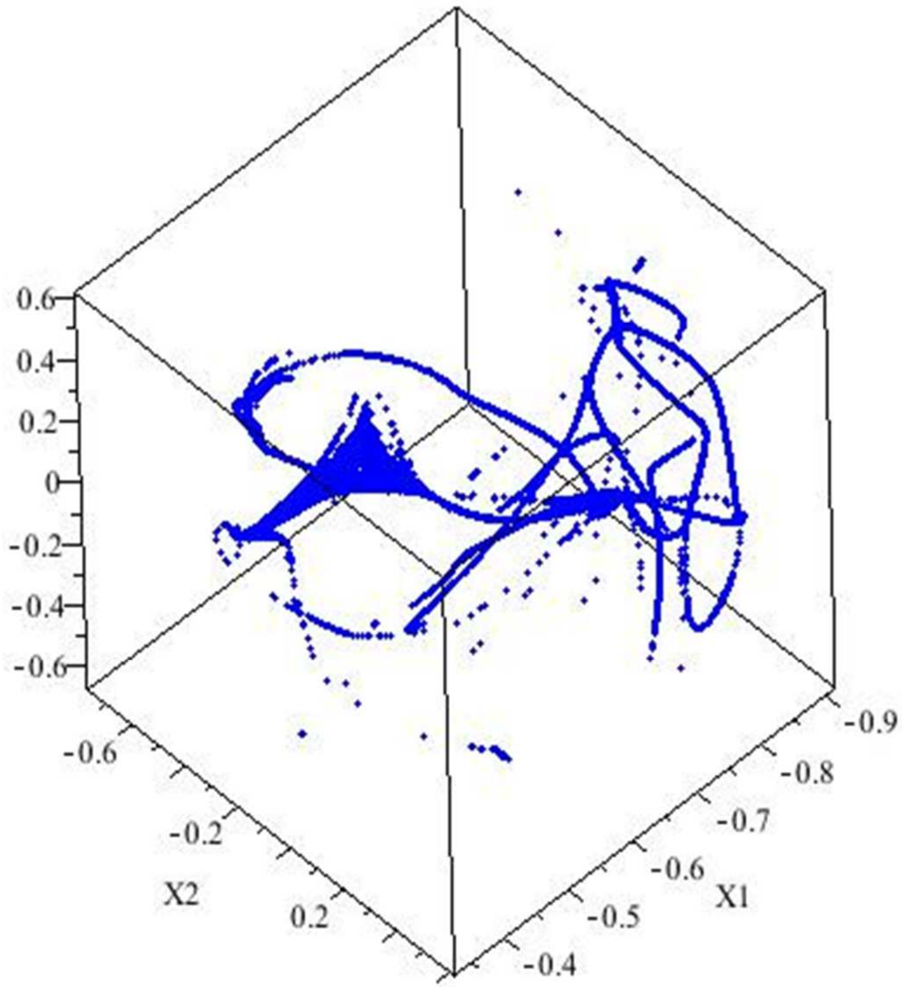
Clustering Process

Some strands collapse to points, others remain



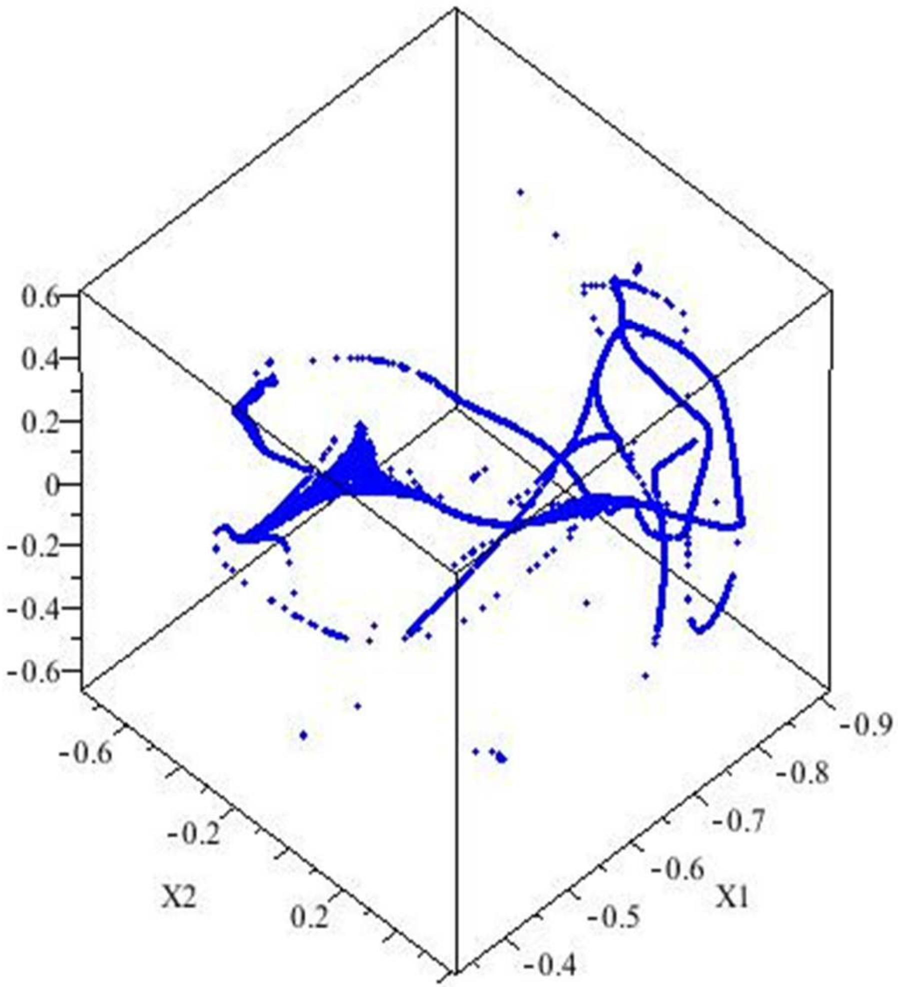
Clustering Process

Some strands collapse to points, others remain



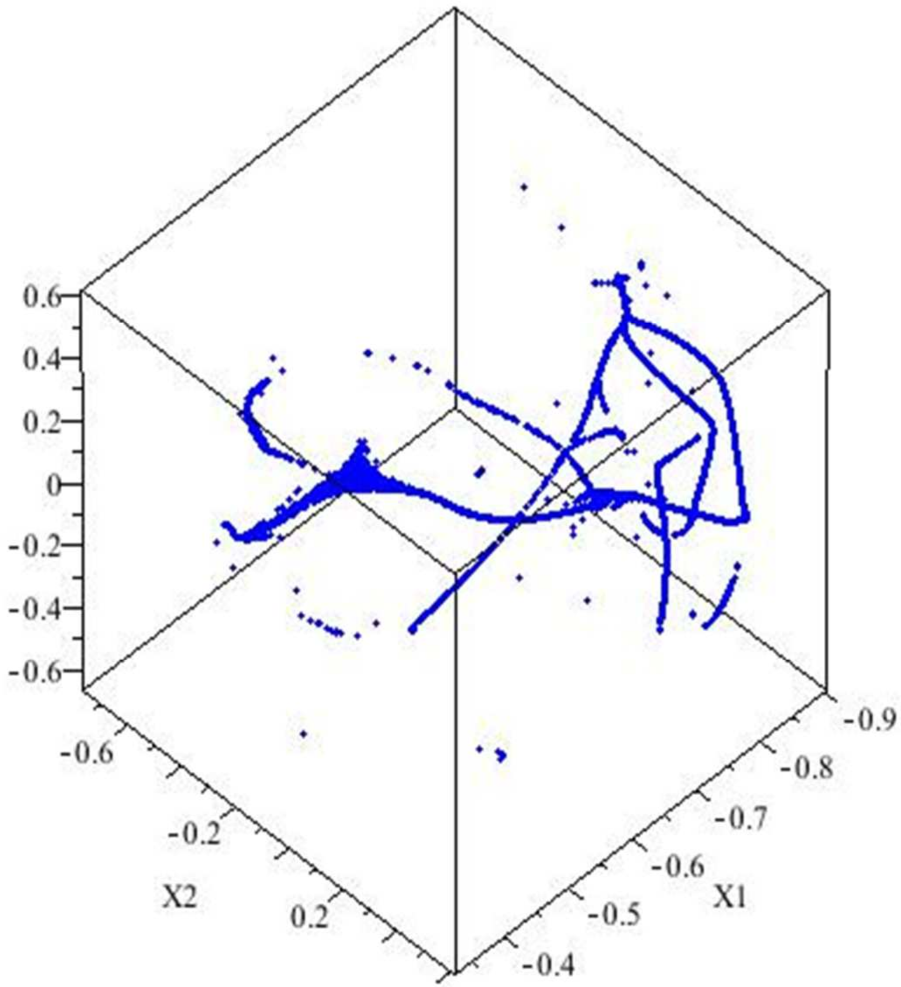
Clustering Process

Separation continues



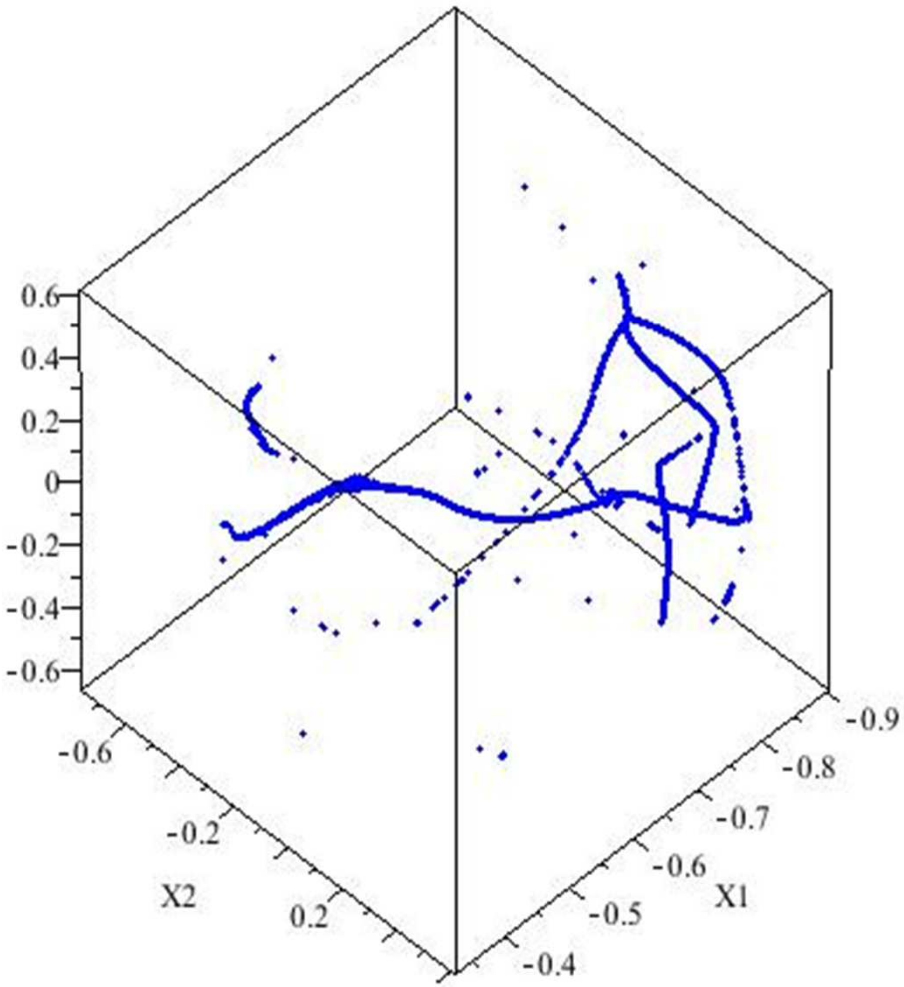
Clustering Process

Separation continues

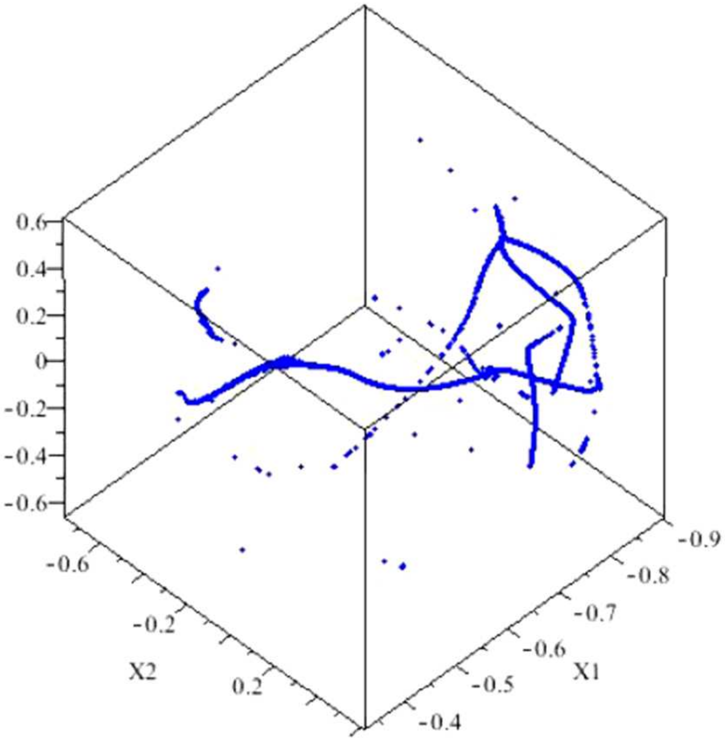


Clustering Process

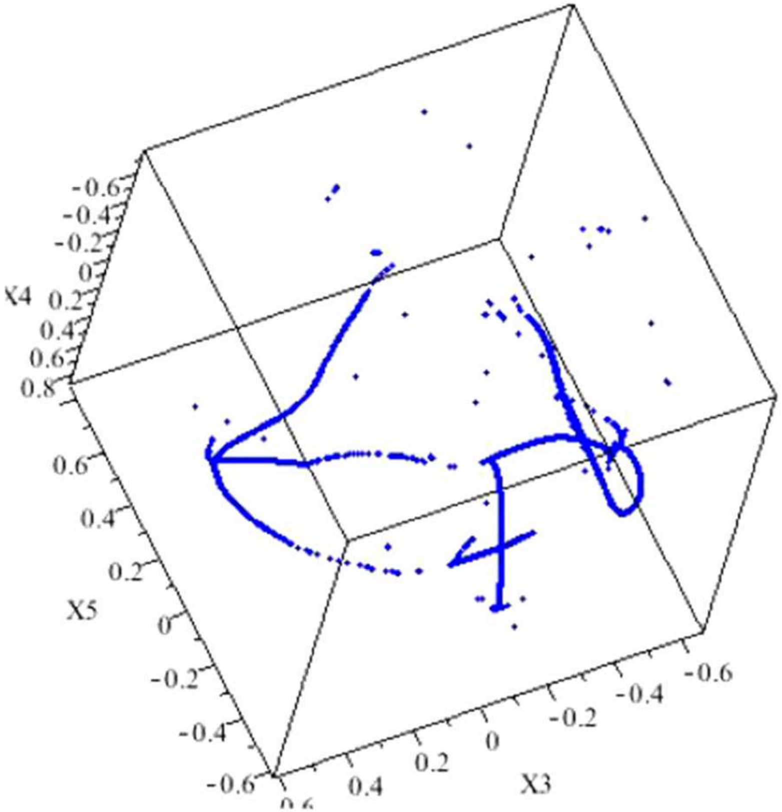
Separation continues



Finally The Clustered Data Looks Like

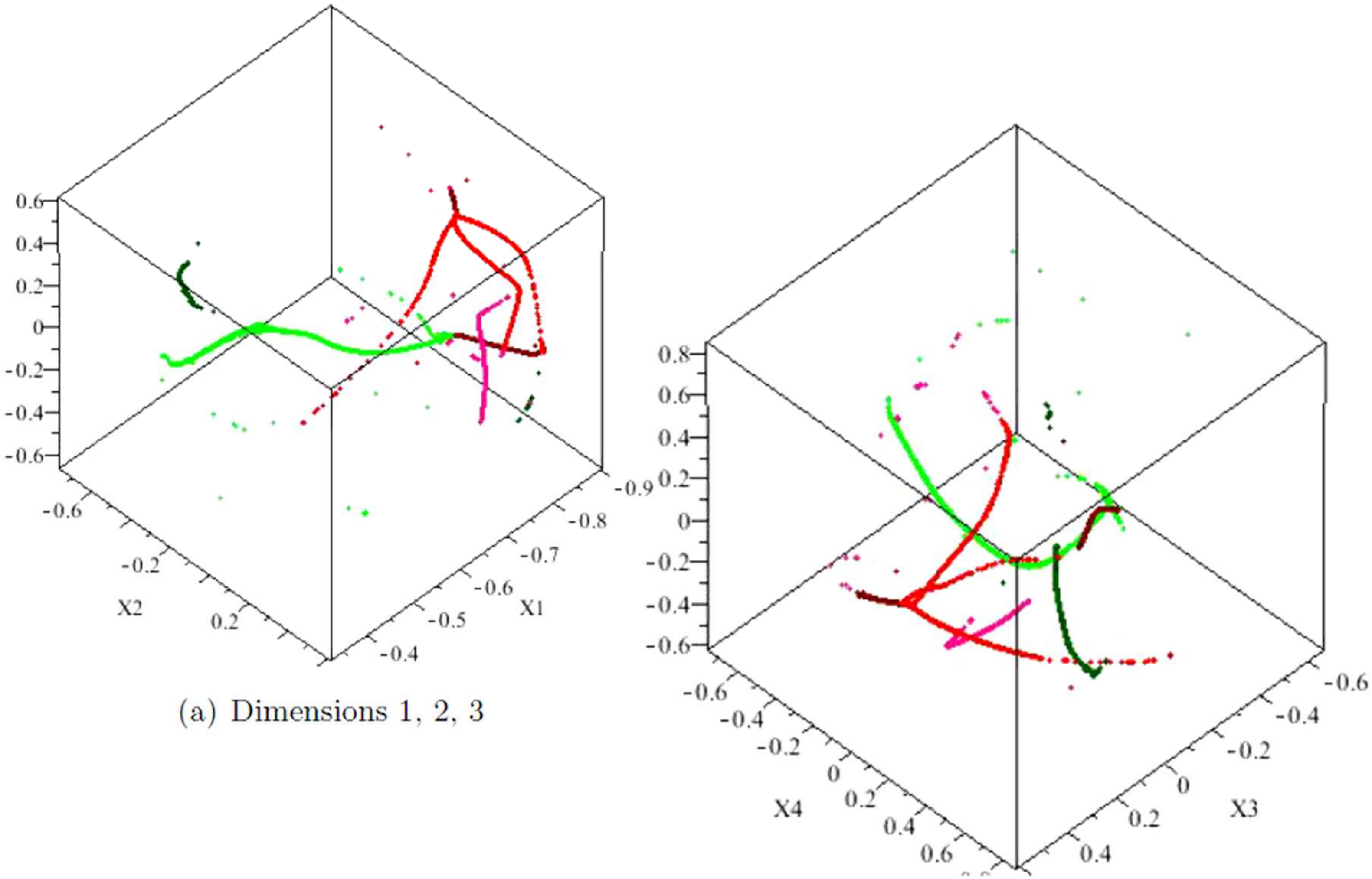


(a) Dimensions 1, 2, 3



(b) Dimensions 3, 4, 5

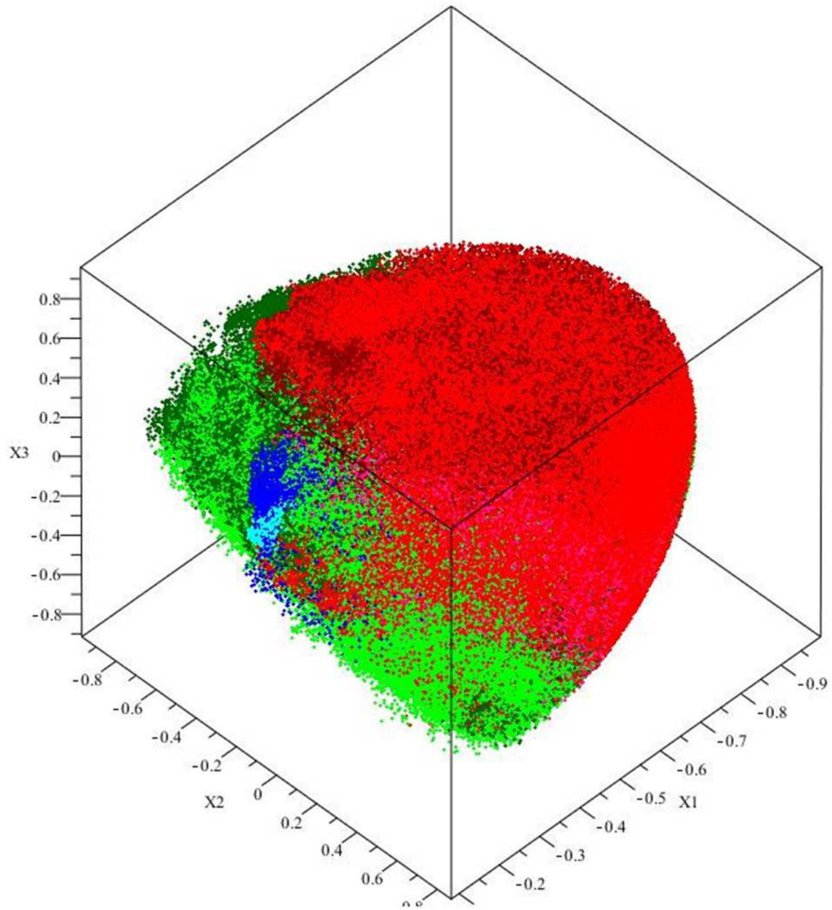
After Selecting Clusters and Coloring



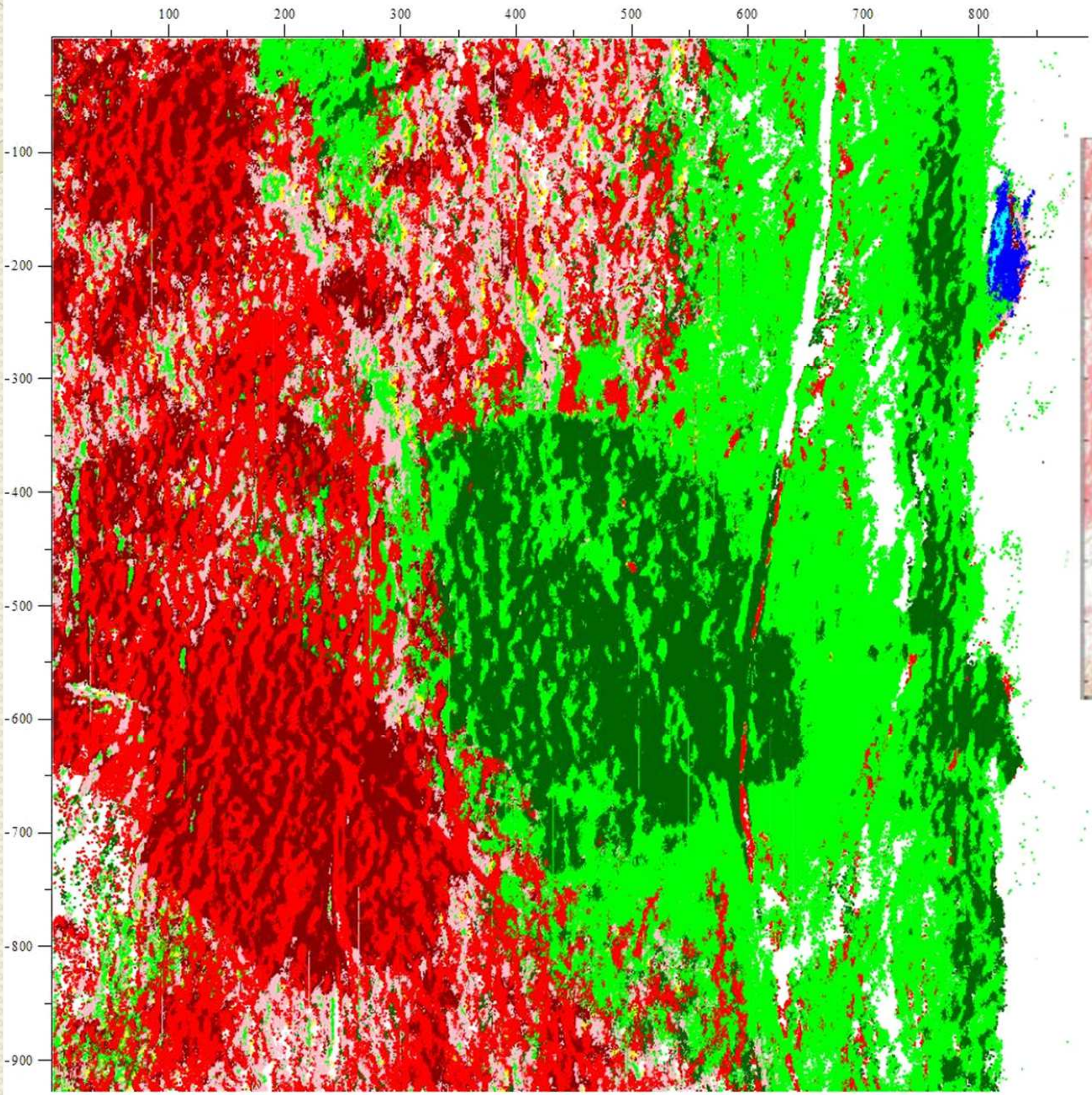
(a) Dimensions 1, 2, 3

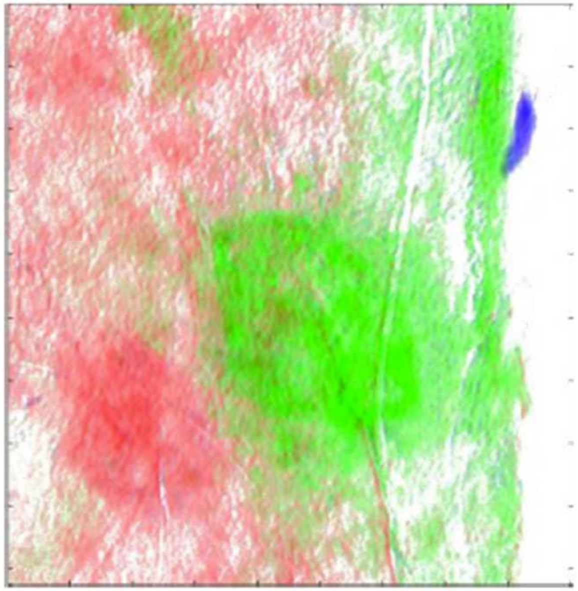
(b) Dimensions 3, 4, 5

SVD Data At The End



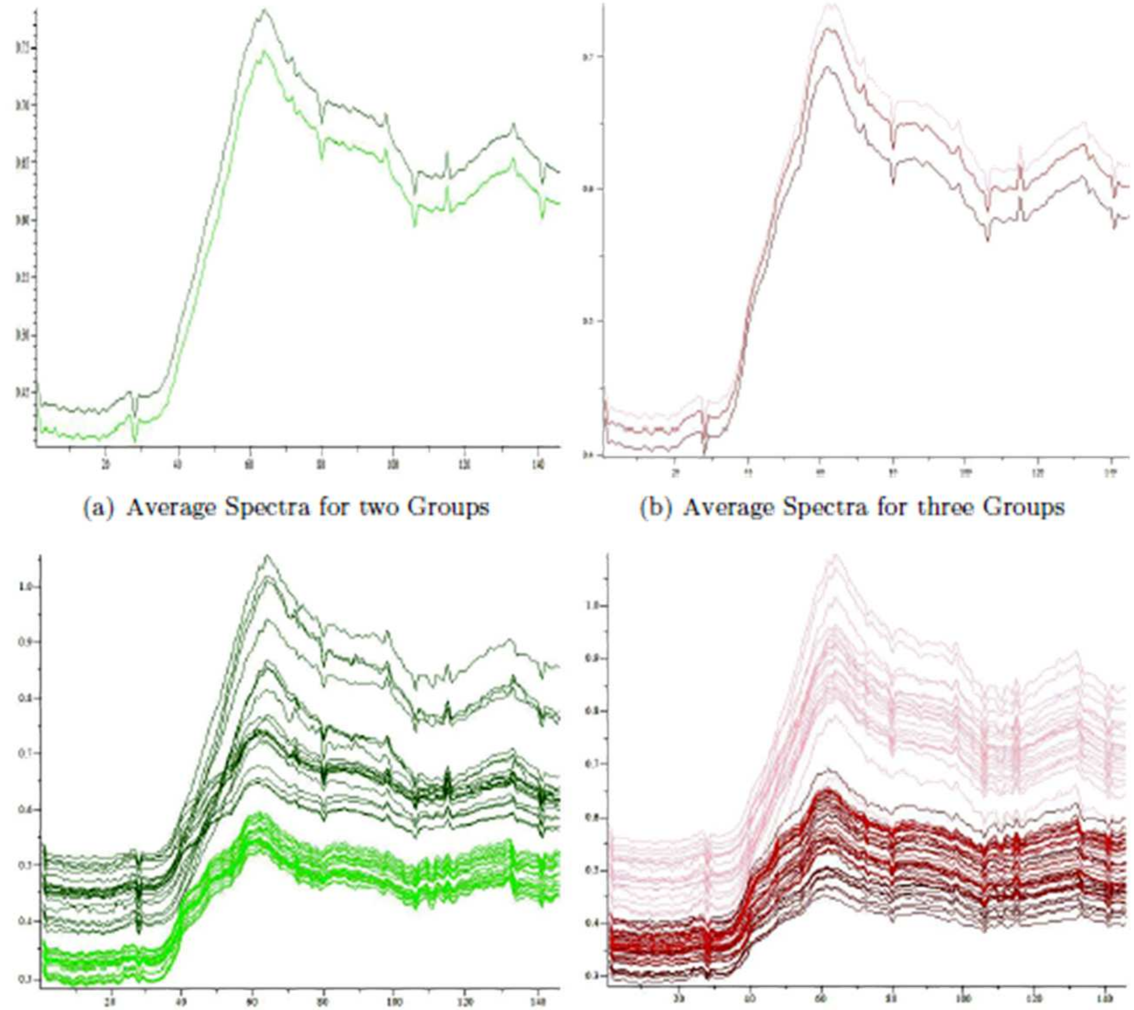
Putting It All Together





Putting the colored points back on the sample to see what is there

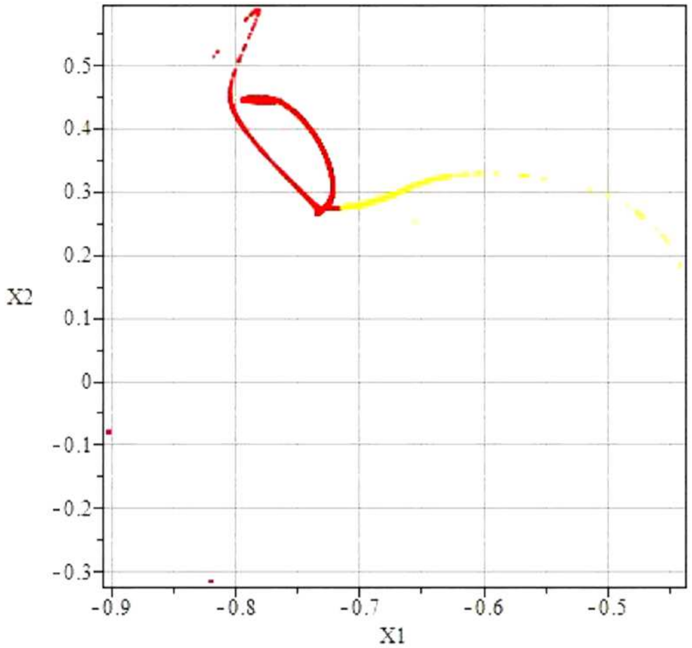
What Do Different Clusters Tell Us?



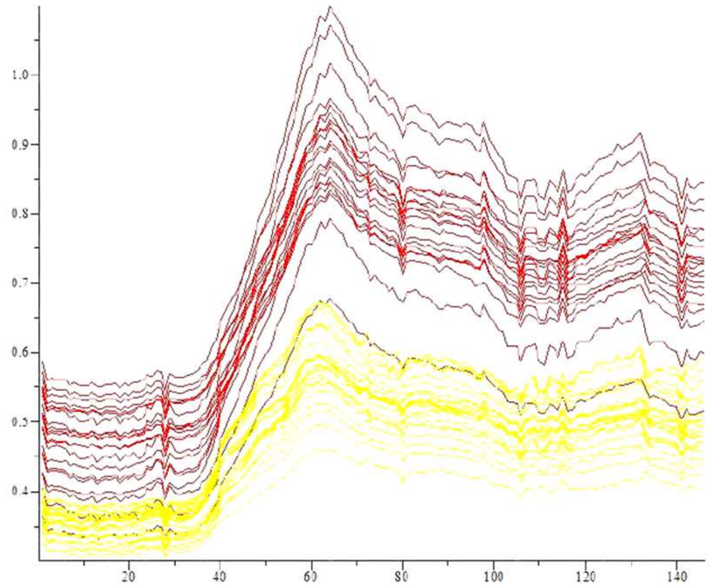
(a) Average Spectra for two Groups (b) Average Spectra for three Groups

(c) 100 Random Spectra for two Groups (d) 100 Random Spectra for three Groups

Finer Details

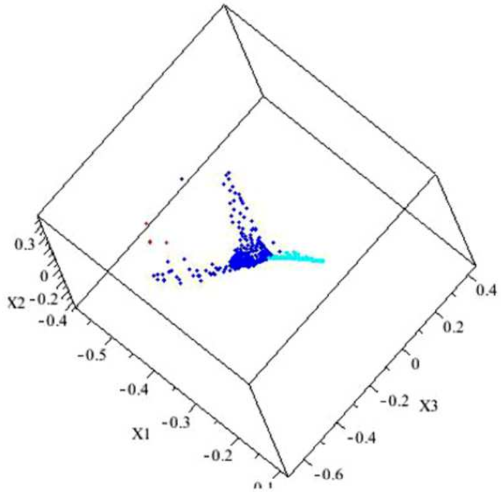


(a) Sub-structure of Red Cluster

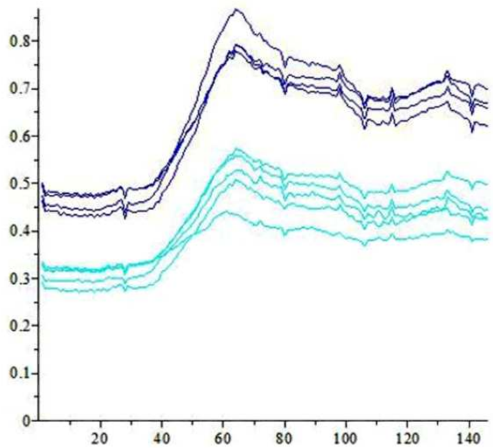


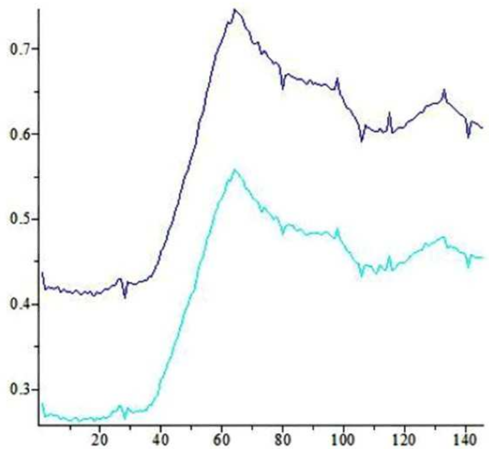
(b) Sub-groups of Red Spectra

More Details



(a) Sub-structure of Blue Cluster


(b) Sub-groups of Blue Spectra


(c) Averages of Sub-groups of Blue Spectra

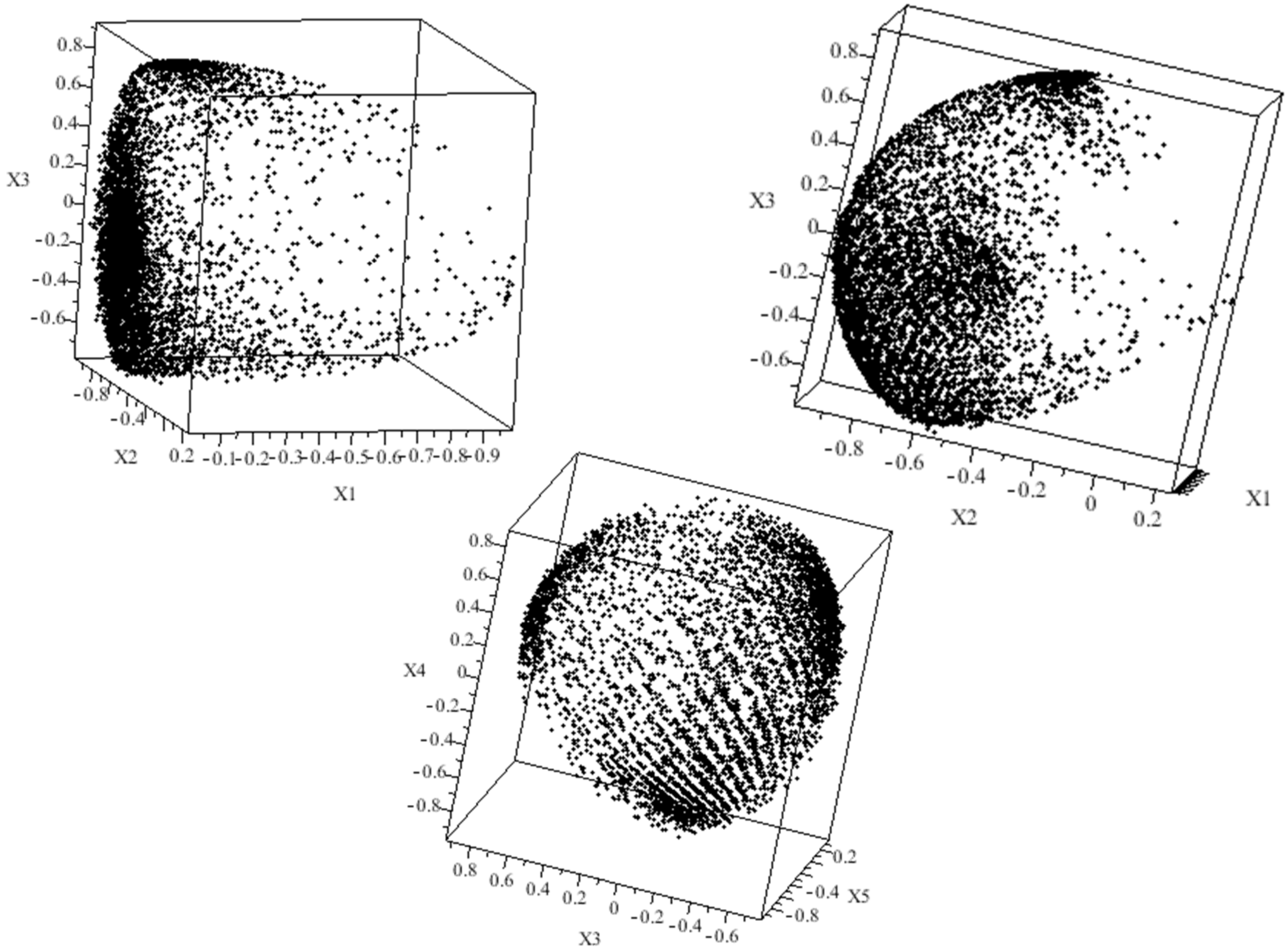
Lessons Learned

- **Large data sets *typically* reveal structures that are more complicated than simple clusters**
  - **These extended structures have features (limbs, arms, reefs, archipelagos..?) that are significant and, in cases studied to date, reveal a lot about the data**
  - **While it is easy to extract eventual simple clusters algorithmically, the extended structures are better explored with our visual approach**

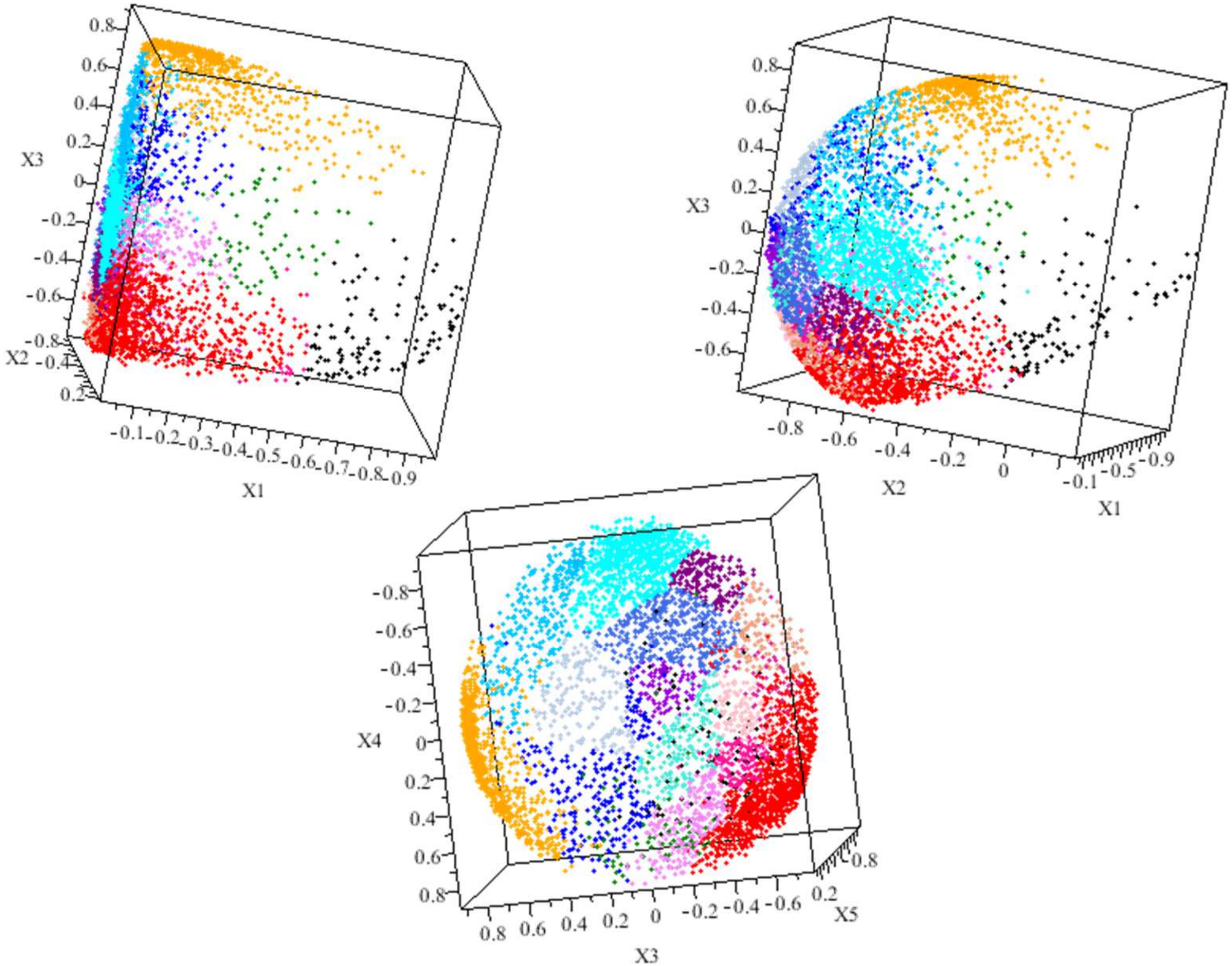
Earthquakes in the Middle East

- **Five quantities are measured (they are extracted from the seismic data)**
  - **Md – the magnitude of the earthquake**
  - **M0 – the moment of the earthquake**
  - **Stress**
  - **Radius**
  - **F0**

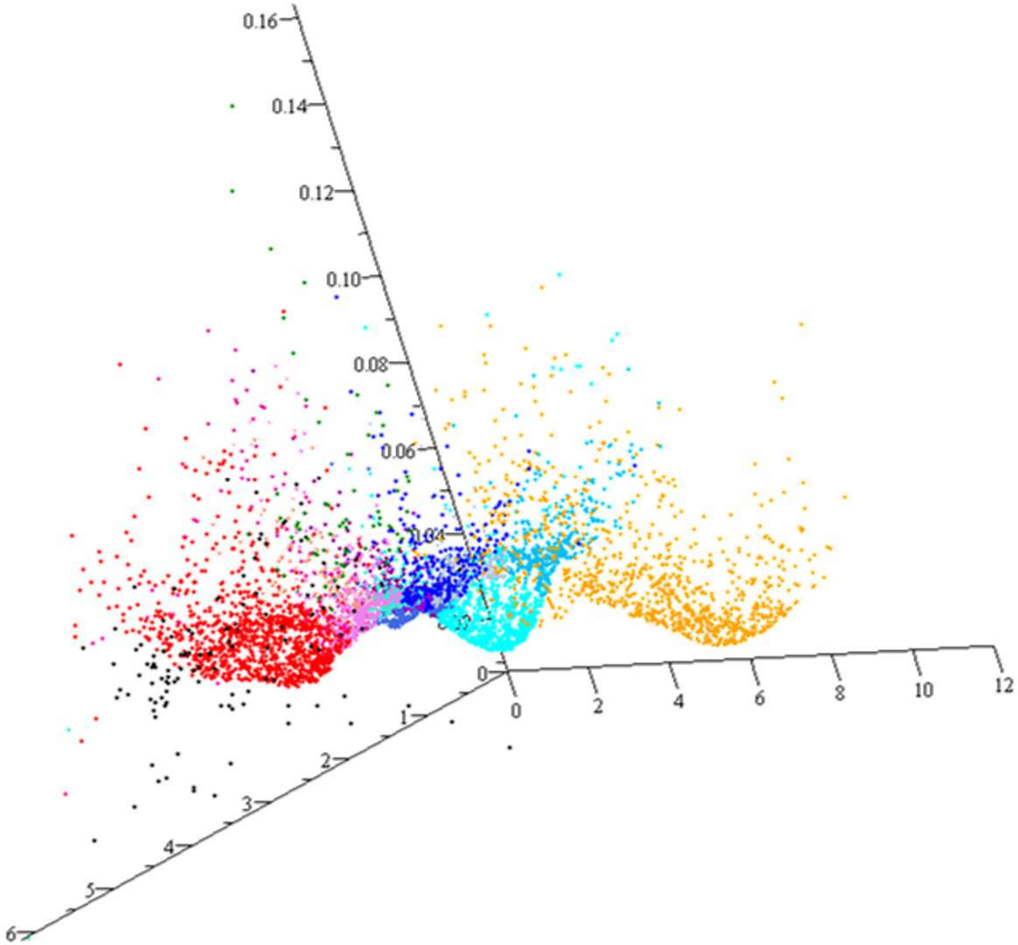
SVD of 5 Dimensional Data



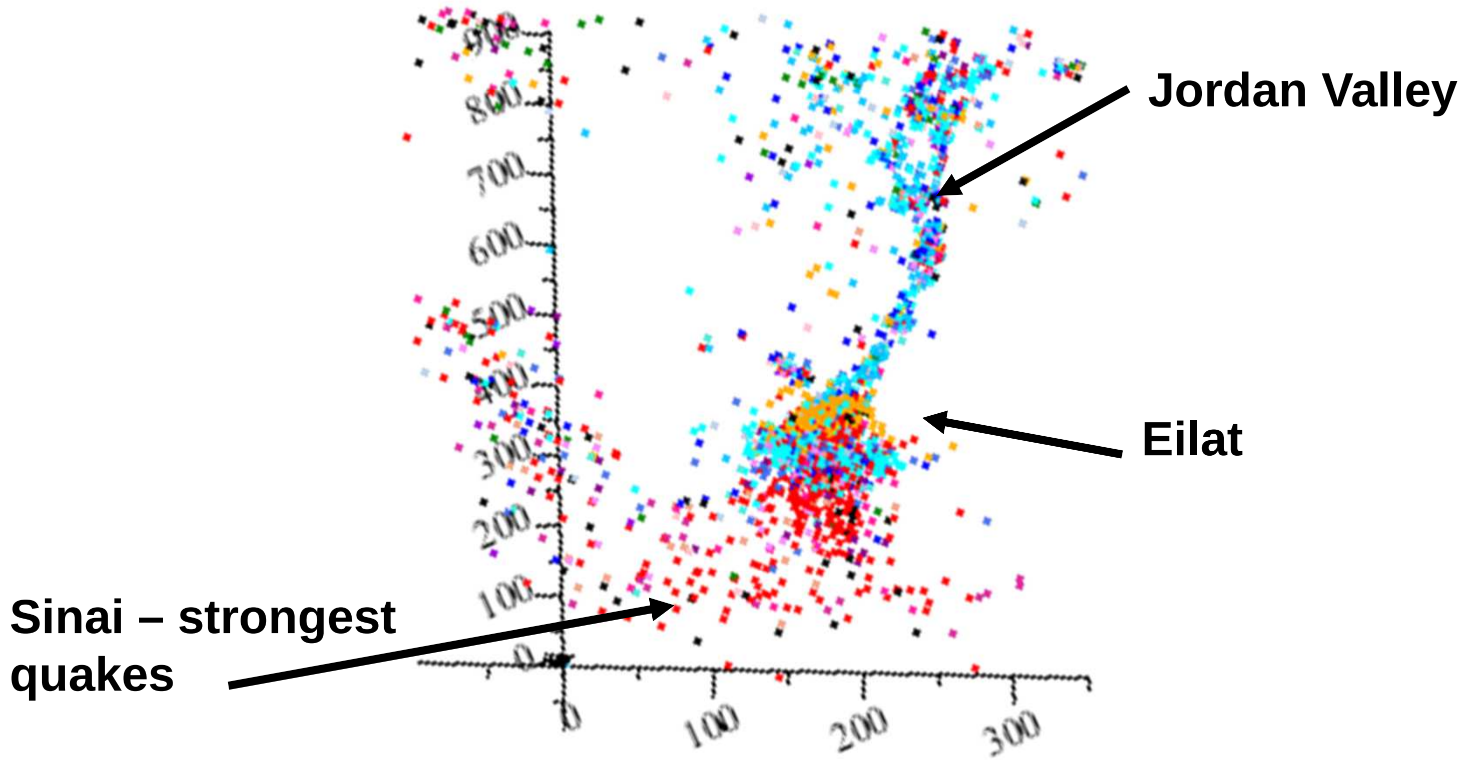
Coloring The Clusters



A 2-d Slice Of The Potential



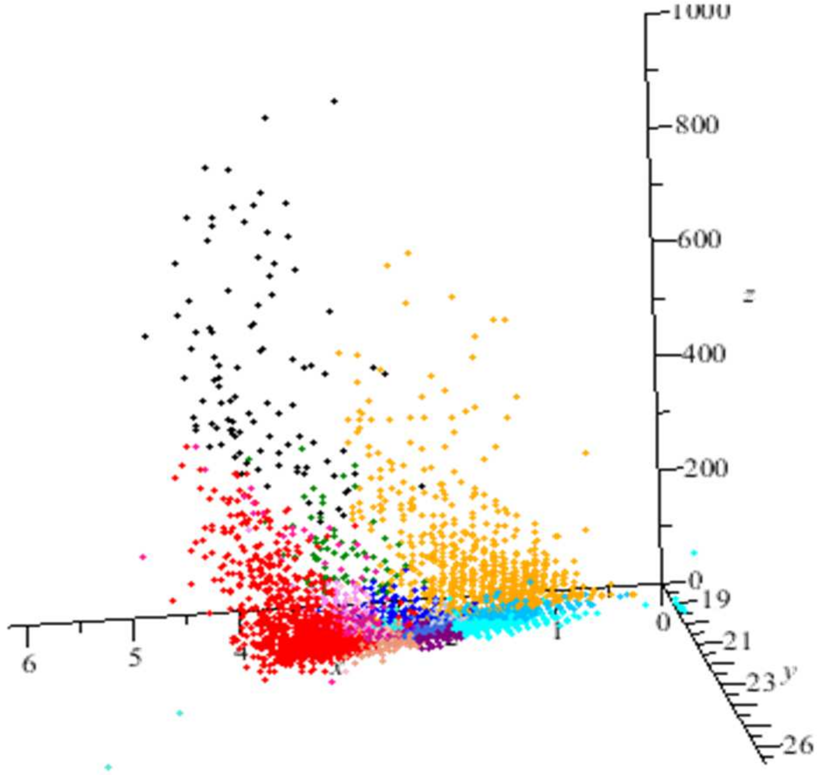
Putting It On A Map



The geophysicists didn’t know about the gold and lighter blue regions

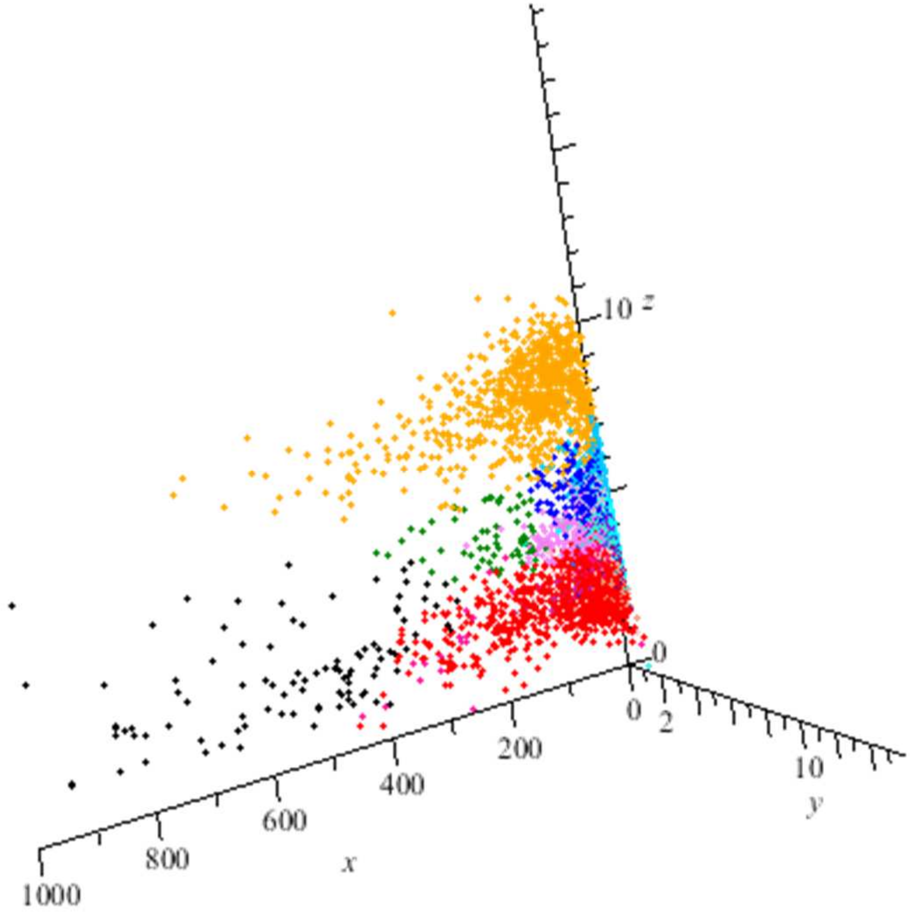
In The Original Coordinates

- **Data isn't separable**
  - **Variables Md(x), M0(y), Stress(z)**

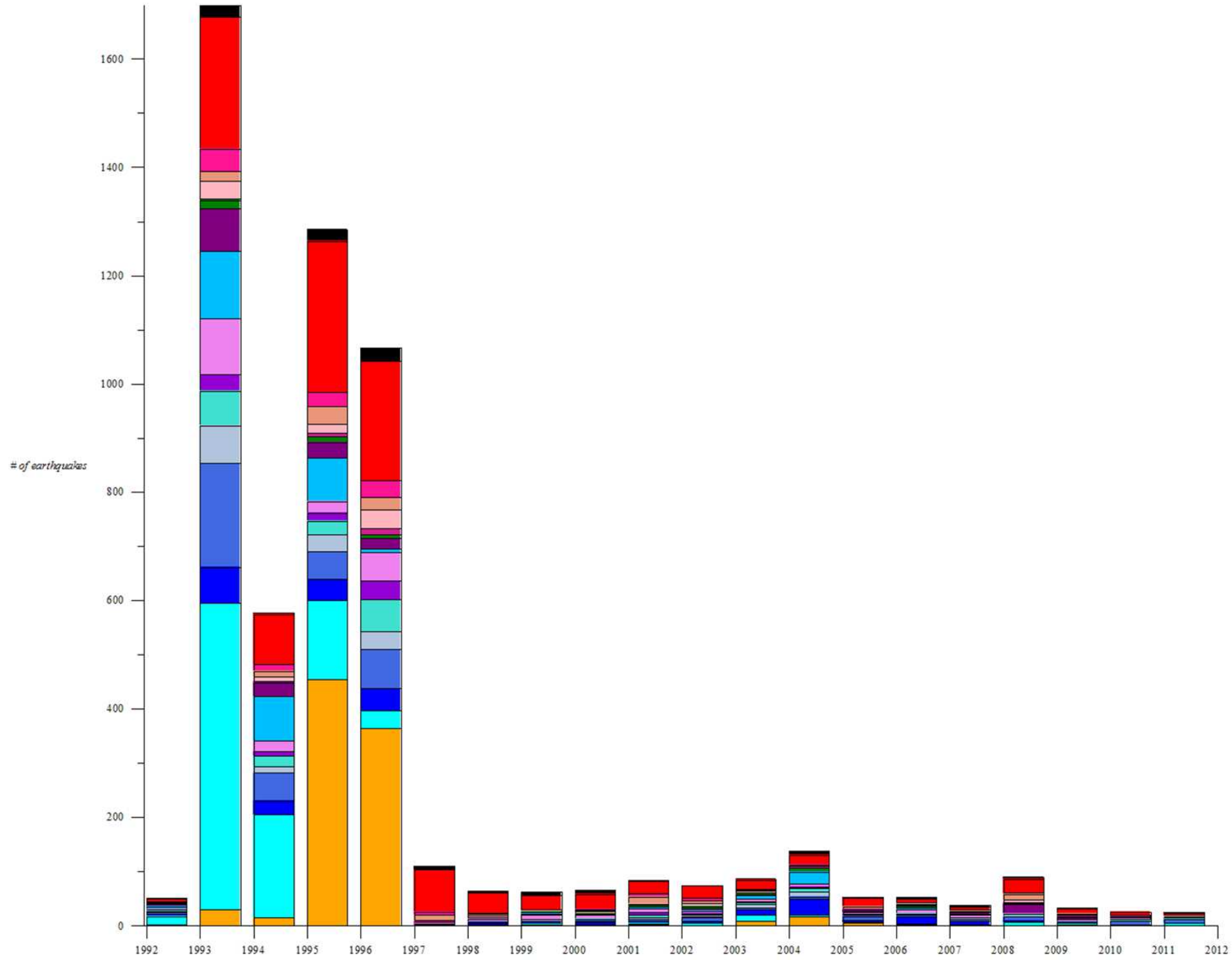


More Original Coordinates

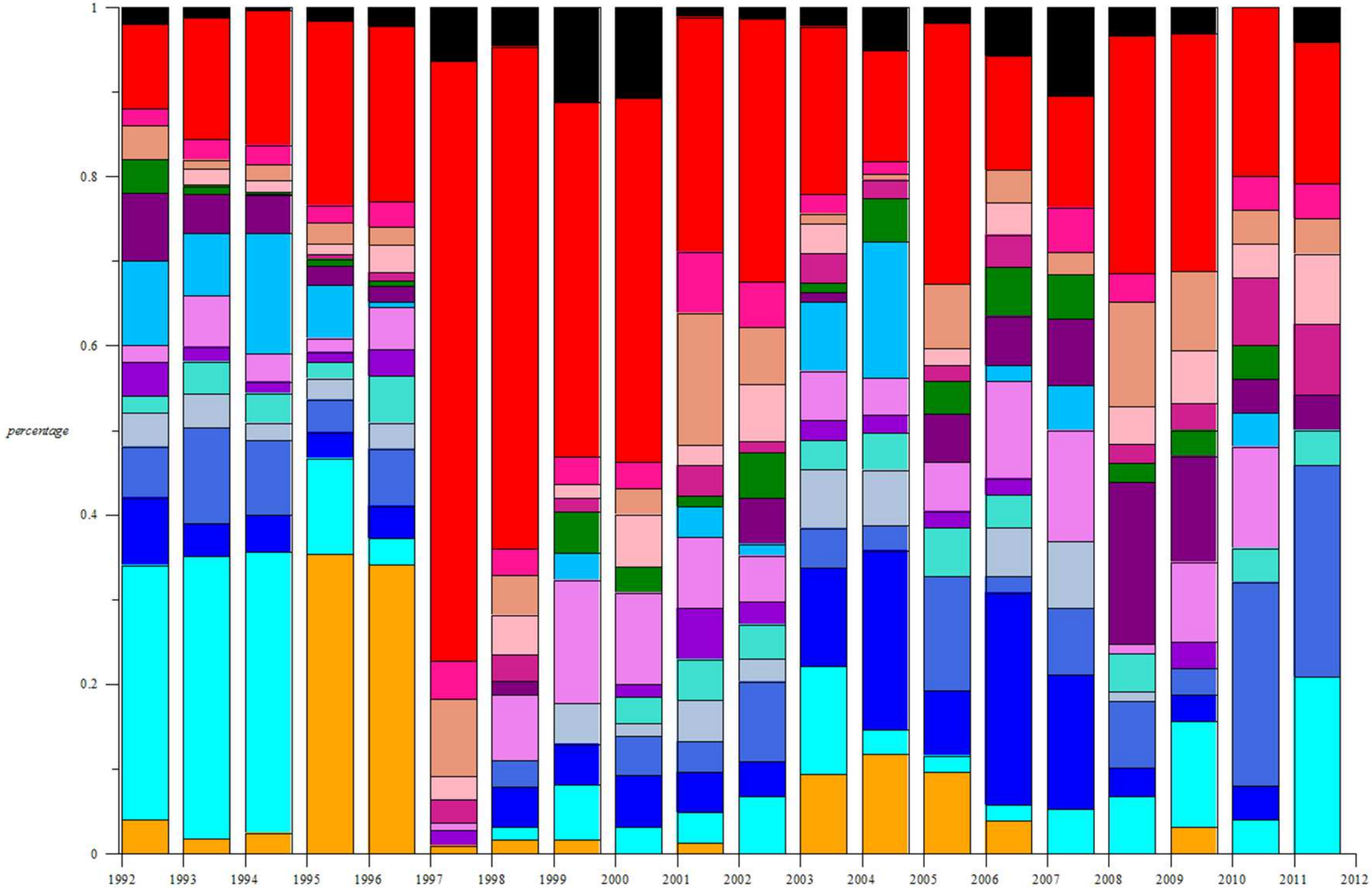
- **Stress(x), radius(y), f0(z)**



Earthquakes By Year And Cluster



Percentage of Cluster By Year



	Machine Learning	Hierarchical	Partitional e.g. k-means	Density Based	DQC
Assume number of cluster?	Hidden in choice of Training set	Number of clusters	Number of clusters	Sensitive to choice of 2 parameters	No One param low sensitivity to choice
Topological Structures	For SVC must assume many outliers	No	No	Somewhat	Yes
If High Dimensional Data	?	Many small clusters	Many small clusters	?	Linear Cost in Compute Time
Visual	No	No	No	Not really	Yes
Extract Clusters	Algorithmic	Algorithmic	Algorithmic	Algorithmic	Visual + Algorithmic

Based on information in
On Clustering Validation Techniques, *Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis,* Journal of Intelligent Information *Systems, 17:2/3, 107-145,2001*

Final Message

- **All of the large, complex, data sets we have looked at exhibit complicated structures in addition to simple clusters**
  - **DQC can find these structures when other data mining methods fail**
  - **Every strand in these clusters is meaningful**
  - **Complex structures can be explored using the visual approach.**
    - **After we know what they mean we can extract them algorithmically.**