Ordered Vector Quantization for the Integrated Analysis of Geochemical and Geoscientific Data Sets

Stephen Fraser[1] & Bruce Dickson[2]

"We are drowning in information and starving for knowledge."

Rutherford D. Roger

[1] CSIRO Exploration & Mining

[2]Dickson Research Pty Ltd

Background

Explorationists/Geochemists gather data faster than it can be interpreted.

A GIS enables data storage and display; but does not resolve the issue:

"How do we intelligently analyze and interpret the volumes of data we collect?"

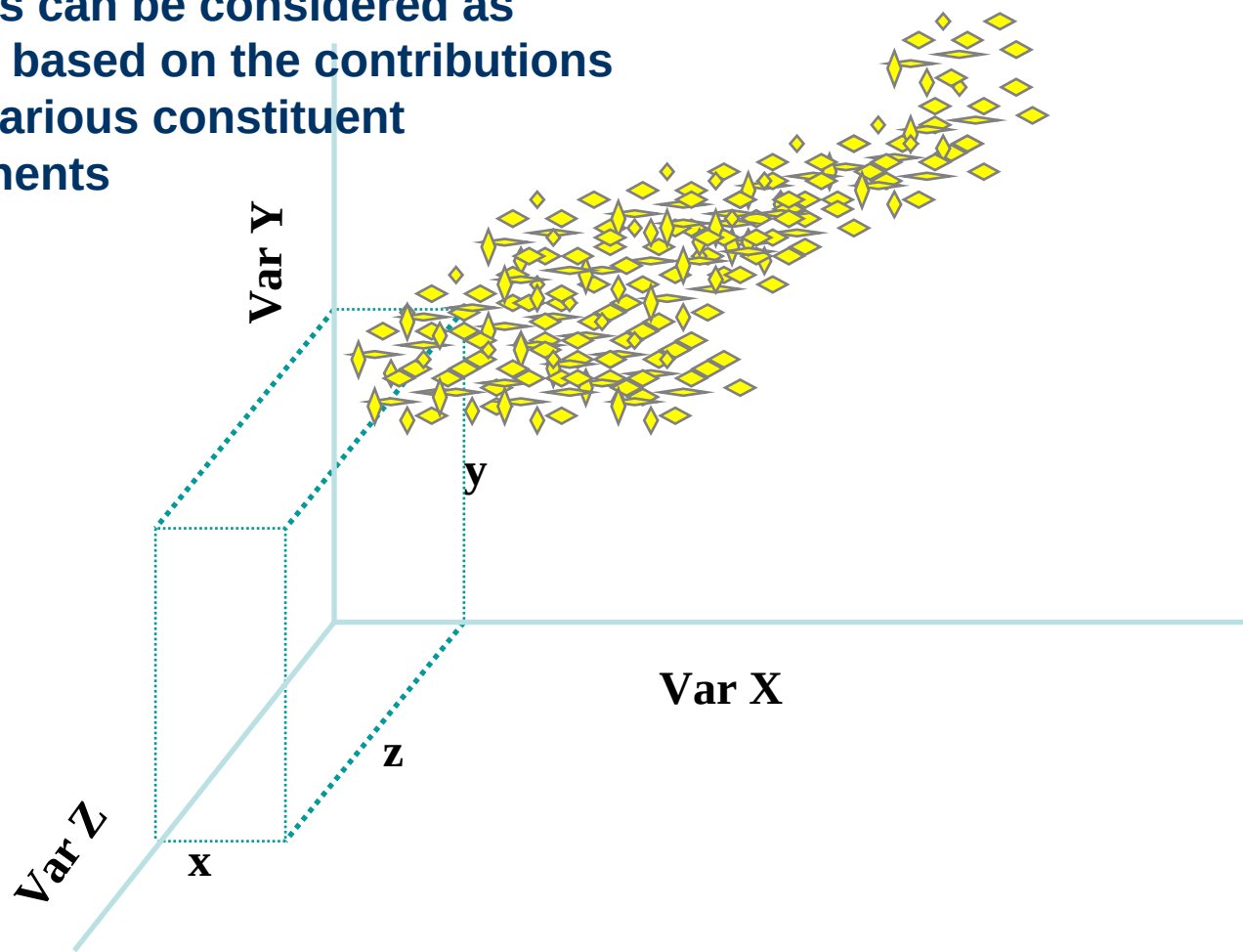
Clustering & Data Analysis: A brief history

- **Classical Statistical Approaches – Linear relationships with single or multiple Gaussian populations:**
  - Fisher's Discriminate Analysis, Least-Squares, Principal Components Analysis, Factor Analysis.
- **Modern Statistical Approaches – Flexible methods, that estimate within, and between class variances and probabilities:**
  - Nearest Neighbour, Projection Pursuit, Canonical Variate Analysis, Causal Networks, Classification And Regression Trees, Multivariate Adaptive Regression Trees,
- **Machine Learning – Computer Aided Methods :**
  - Artificial Intelligence, Expert Systems, Decision Trees, Neural Nets
    - (most Machine Learning methods are supervised!)
  - **"Ordered-Vector Quantization" – Self Organizing Maps**
    - **(Kohonen Nets ~1985)**

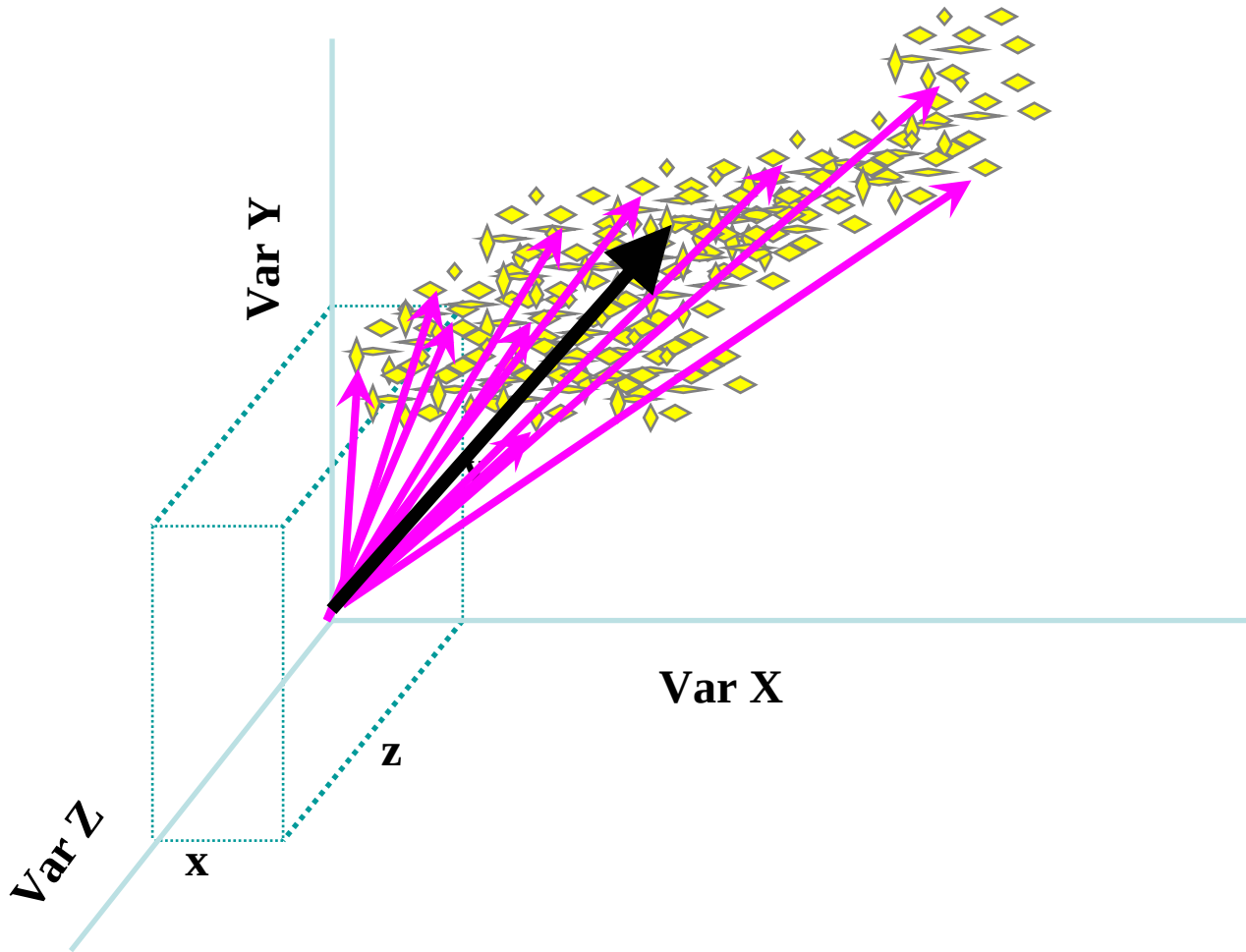
Scatter Plots: An Effective Tool for Conceptualizing Data Processing methods

Consider a grouping of similar/related samples in n-D space

Samples can be considered as vectors based on the contributions of the various constituent components



Scatter Plots:

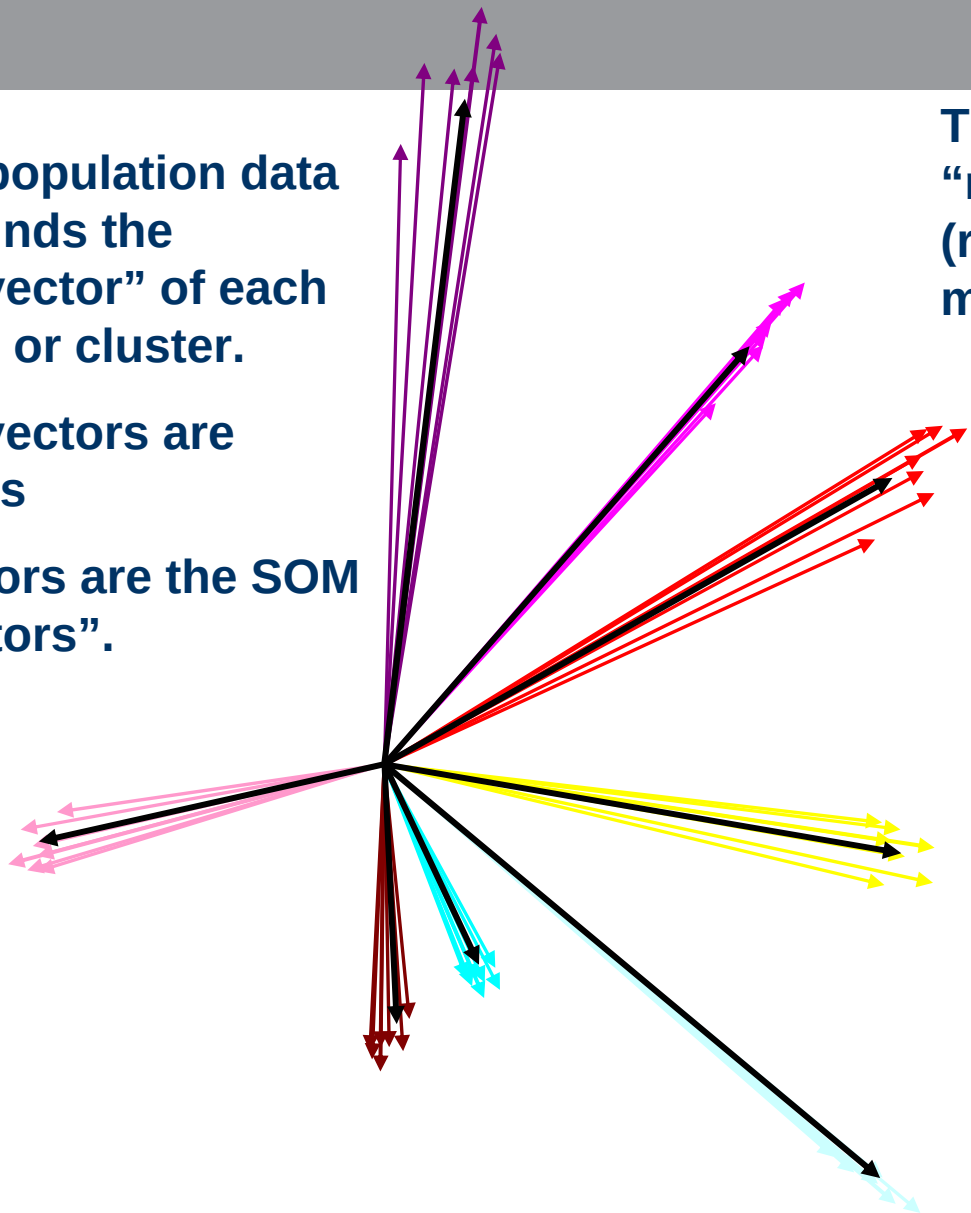


Scatter Plots: In a SOM analysis

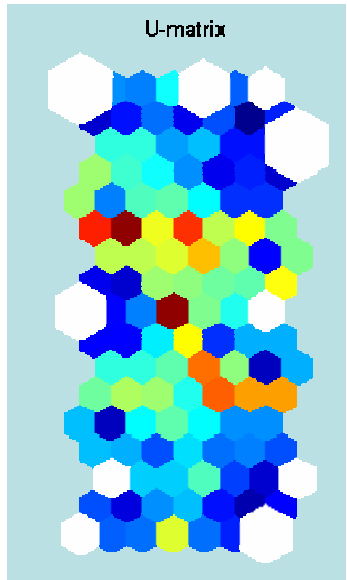
In a multi-population data set, SOM finds the "medium-vector" of each population or cluster.

Coloured vectors are populations

Black vectors are the SOM "code-vectors".



Then displays them as a "map", so that topology (relationships) is maintained



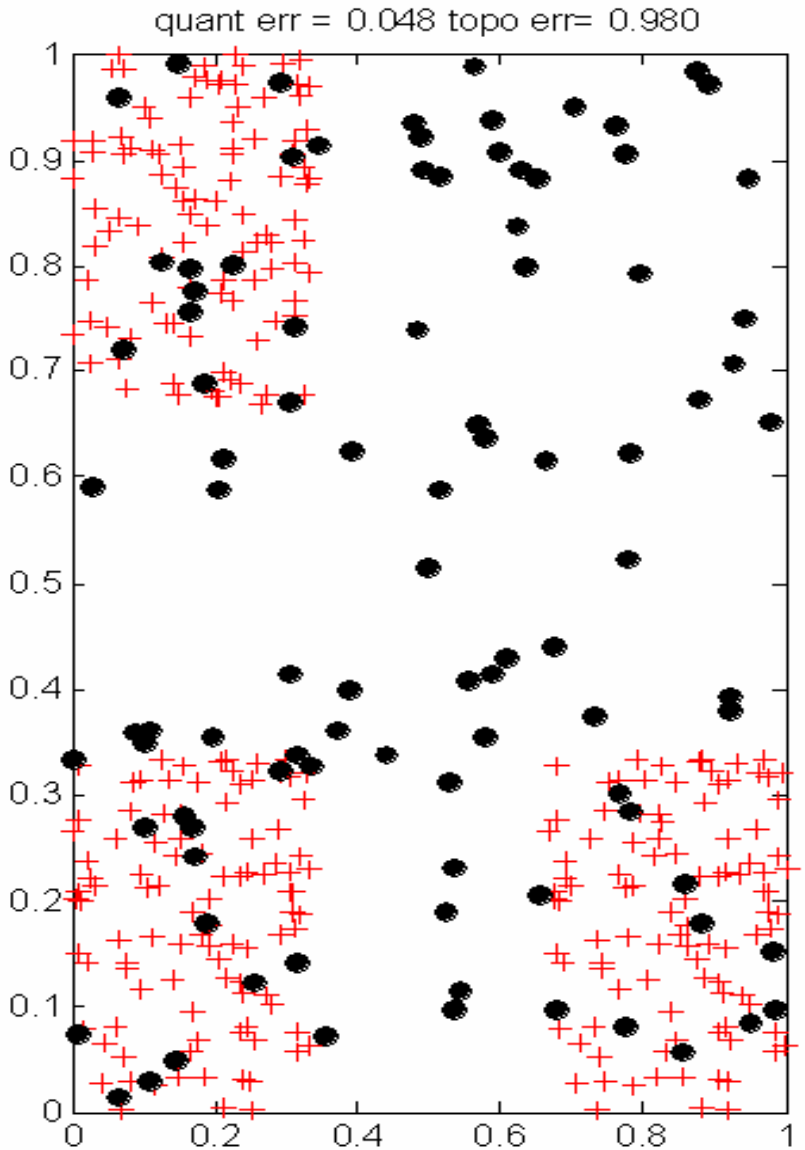
An Example Visualization in 2D – Step 1 Initialize the SOM

Red crosses represent data points in 2 dimensions

A SOM of 12x8 has been chosen

Begin by "randomizing" the SOM to cover the data space

Black circles are the random SOM "seed" vectors

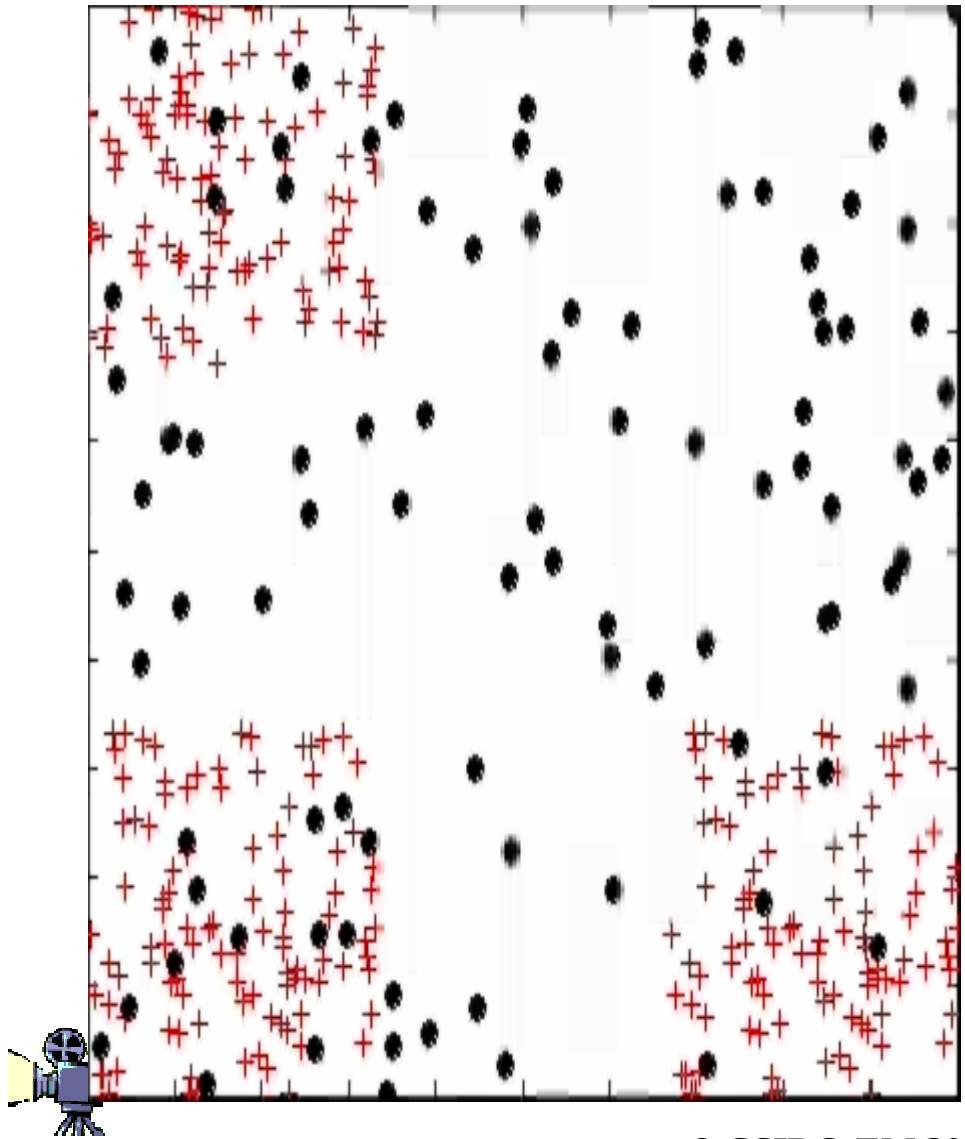


An Example Visualization in 2D – Step 2 Train the SOM

The training is based on two principles:

Competitive learning: the "seed" prototype vector most similar to a data vector is modified so that it it is even more similar to it. This way the map learns the position of the data cloud.

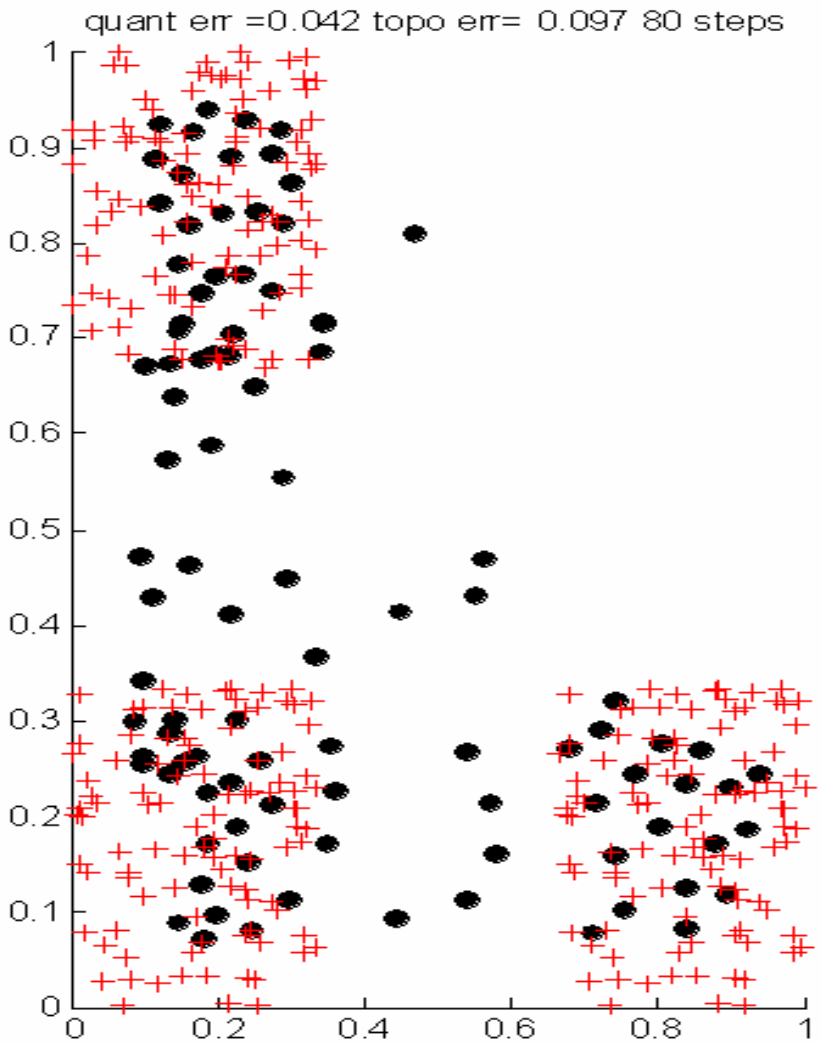
Cooperative learning: not only the most similar "seed" prototype vector, but also its neighbours on the map are moved towards the data vector. This way the map self-organizes.



The Trained SOM

Note that

- **SOM code-vectors are unevenly spaced;**
- **Some code-vectors are in space between data;**
- **Different runs will give different "looking" SOMs;**
- **Code-vectors are within the data clouds – not on the edges;**
- **Ideal stopping point is when system reaches a "steady-state".**



The Self-Organizing Map (SOM)
after Kohonen: http://www.cis.hut.fi/research/som-research/som.shtml)

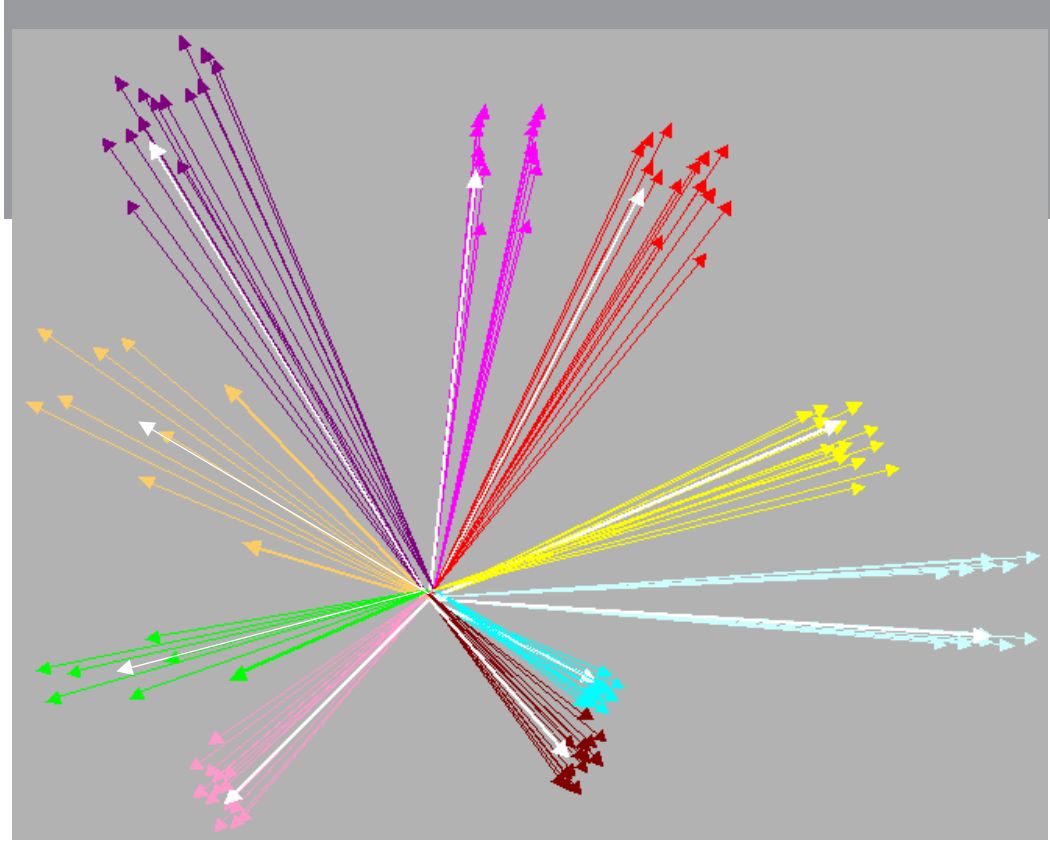
- Fitting of the model vectors is usually carried out by a sequential regression process, where *t* = 1,2,... is the step index: For each sample x*(t)*, first the winner index *c* (best match) is identified by the condition

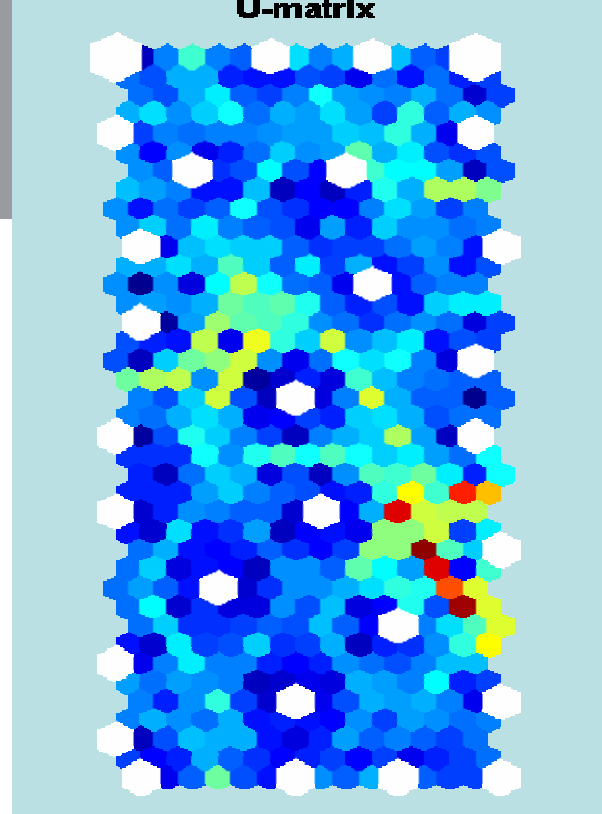
$$\forall i, \left|\left|\mathbf{x}(t) - \mathbf{m}_c(t)\right|\right| \leq \left|\left|\mathbf{x}(t) - \mathbf{m}_i(t)\right|\right|.$$

- After that, all model vectors or a subset of them that belong to nodes centered around node c = c(x) are updated as

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c(\mathbf{x}),i}(\mathbf{x}(t) - \mathbf{m}_i(t)).$$

- Here $h_{c(\mathbf{x}),i}$ is the ``neighborhood function'', a decreasing function of the distance between the i th and c th nodes on the map grid. This regression is usually reiterated over the available samples.



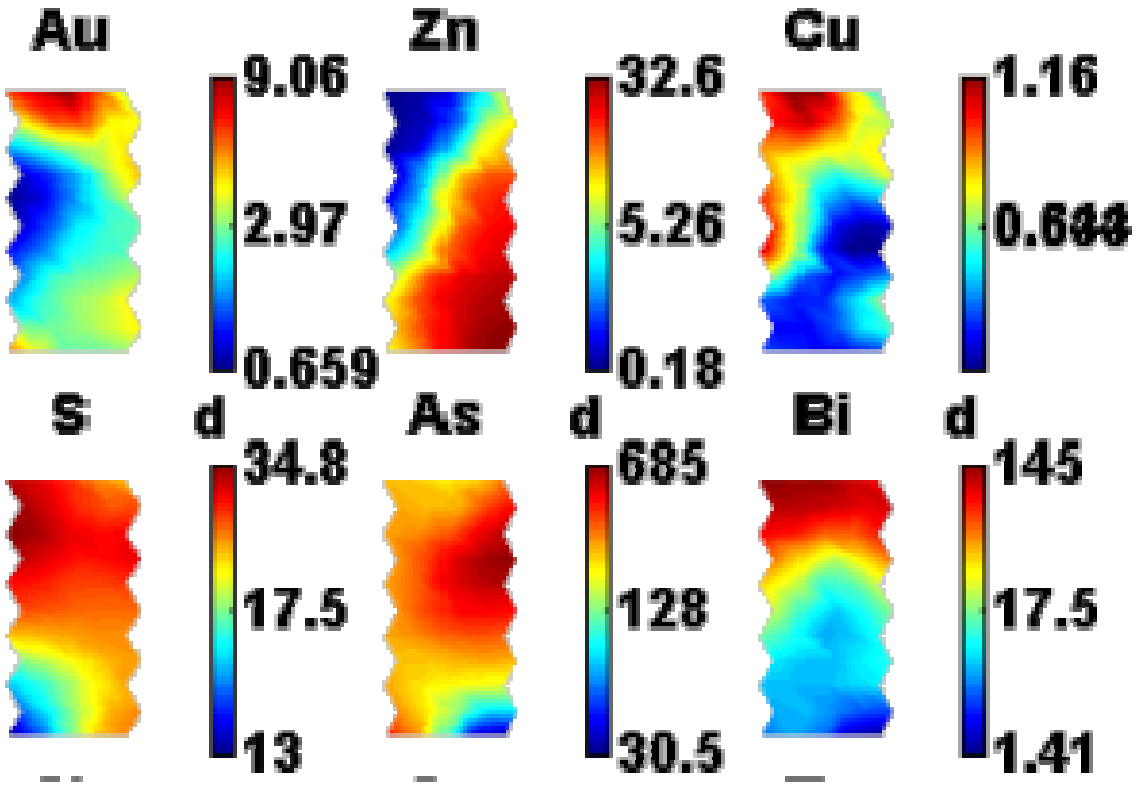


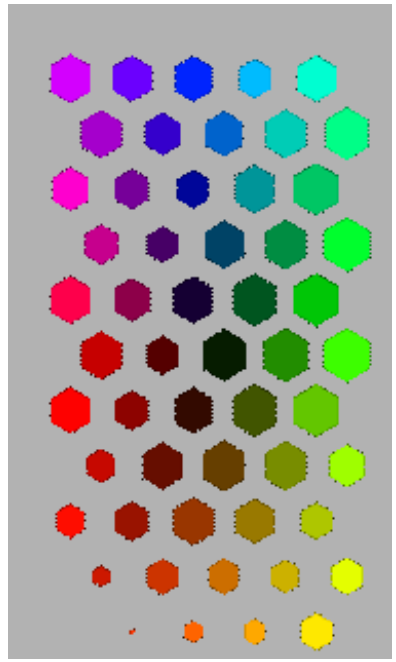
Code Vectors are then projected as "nodes" onto a surface (sheet, cylinder or toroid) "map" so as to maintain their n-D topology

Each "Code-Vector" sample can be described by its variables:
{c1,c2,c3,…cn ,d1, d2, d3…dn etc }

U-Matrix – adjacent "nodes" separated by cool colours are close in Euclidean space; adjacent "nodes" that are further apart will be coloured using a "'hotter-temperature" colour according to the degree of separation.

Self Organized Map – Representations: Component Plots & Colour Map LUT

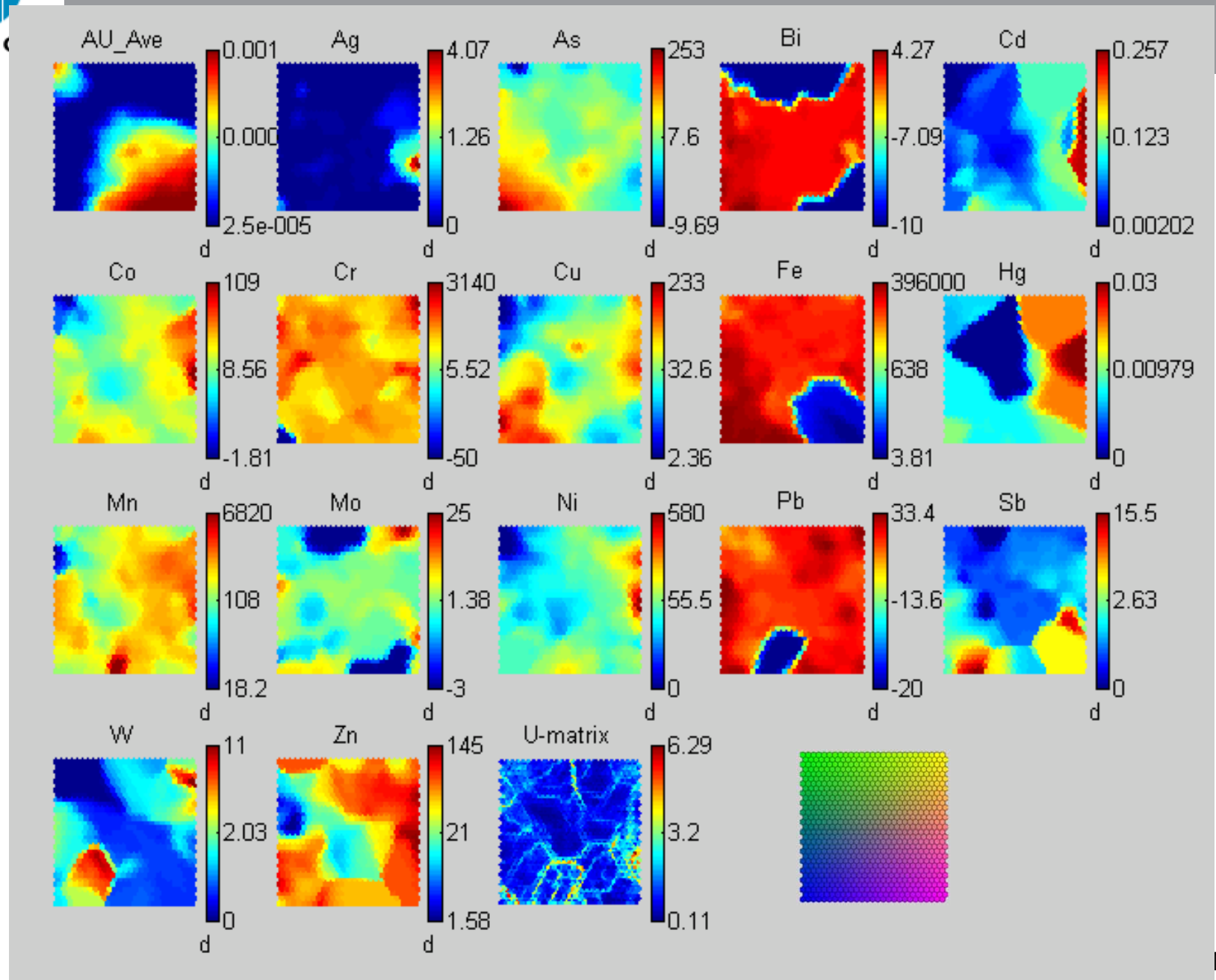




Component Plots show the contribution of each "component" to the "self-organized map"

The SOM "nodes" can be coloured for spatial display purposes;

Example of Component Plots and U-Matrix and Colour Map



A Tool to Assist : SiroSOM (CSIRO Self-Organizing Maps)

Data “Organization” (Clustering);

Dimensionality Reduction, and Visualization;

Based on Principles of Vector Quantization & Measures of Vector Similarity;

Can handle Non-linear, Linear, Like and Disparate data sets;

Can handle categorical data and “labels”;

Nulls (sparse data !) can be accommodated; and,

Errors are tracked throughout the process.

Applied SOM to raster, point, vector data

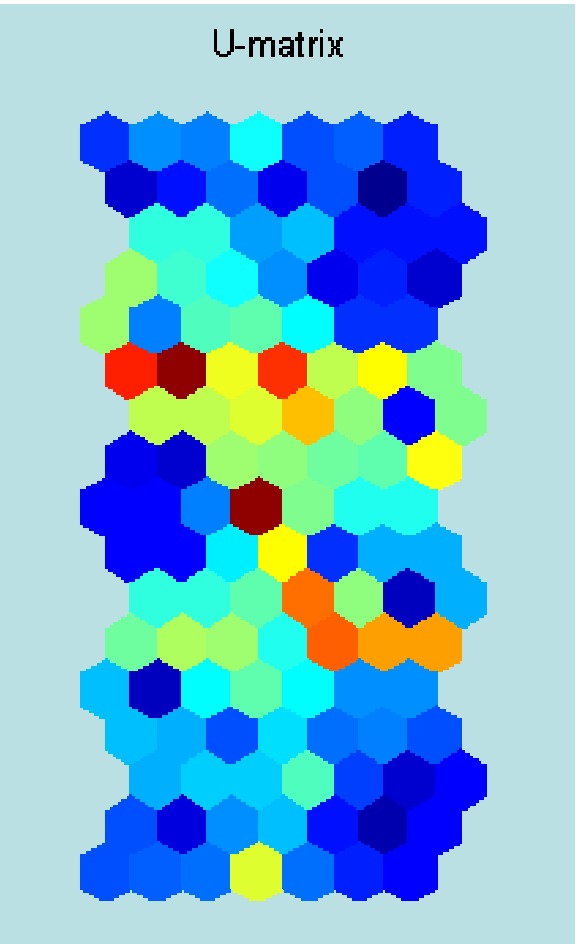
“Animals”

Animals - Categorical data example

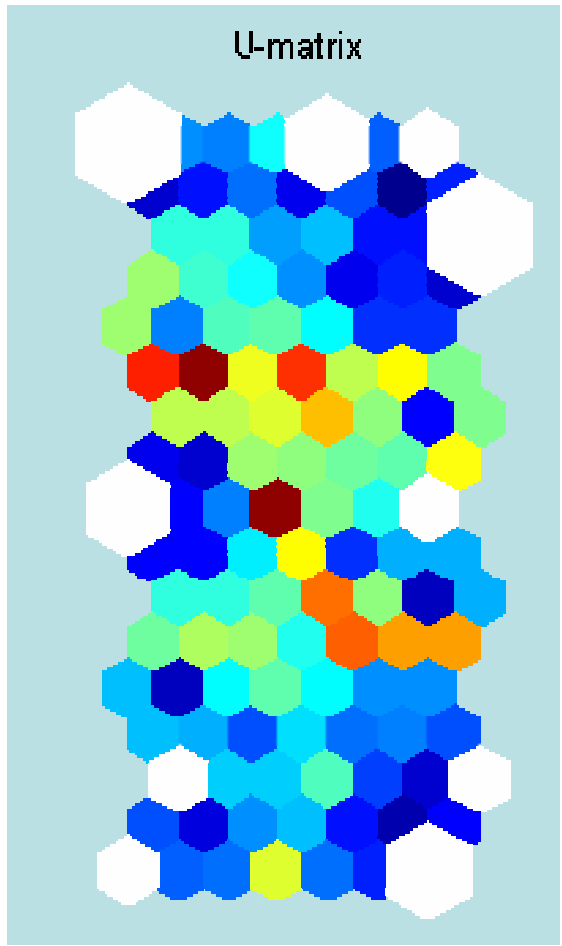
17 Animals (samples) 15 attributes (in 5 data fields)

sml	2lg	feather	fly	peck	dove
sml	2lg	feather	walk	peck	hen
sml	2lg	feather	fly	swim	duck*
sml	2lg	feather	fly	swim	goose*
sml	2lg	feather	fly	hunt	owl[1]
sml	2lg	feather	fly	hunt	hawk[1]
med	2lg	feather	fly	hunt	eagle
med	4lg	hair	run	hunt	fox[2]
med	4lg	hair	run	hunt	dog[2]
med	4lg	hair	run	hunt	wolf[2]
sml	4lg	hair	run	hunt	cat
big	4lg	hair	run	hunt	tiger[3]
big	4lg	hair	run	hunt	lion[3]
big	4lg	hair	run	hoov	horse[4]
big	4lg	hair	run	hoov	zebra[4]
big	4lg	hair	walk	hoov	cow
big	2lg	naked	run	hunt	man

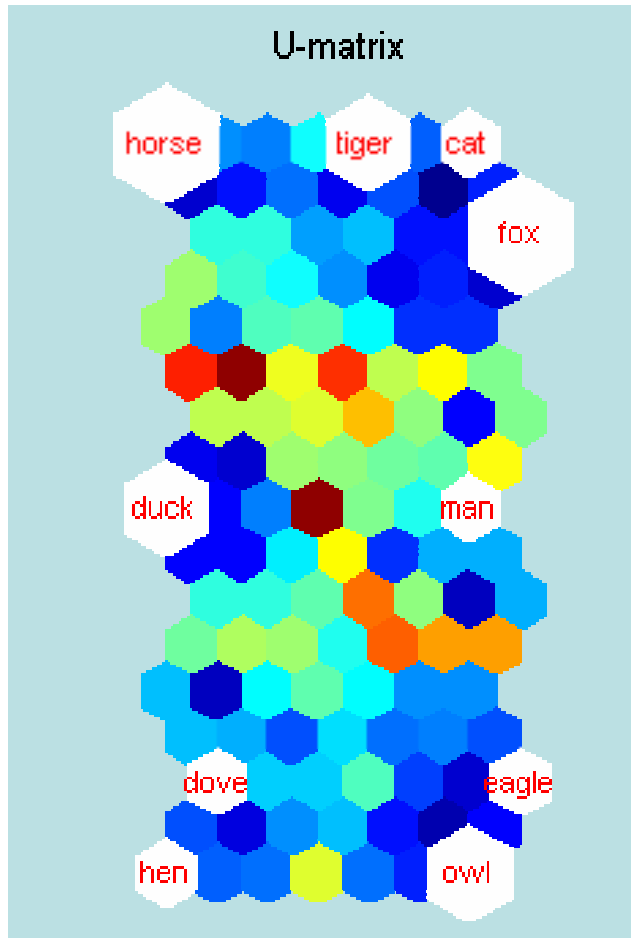
“Animals” U-matrix (9x4 Map)



U-Matrix



U-Matrix with “Hits”



U-Matrix with “Labels”

“Animals” 9x4 U-matrix Label Duplicates

Label sample 15 zebra does not match cell 1 horse has qer 1.1268

Label sample 16 cow does not match cell 1 horse has qer 2.5095

Label sample 4 goose does not match cell 5 duck has qer 0.10457

Label sample 13 lion does not match cell 19 tiger has qer 0.95733

Label sample 9 dog does not match cell 29 fox has qer 0.71732

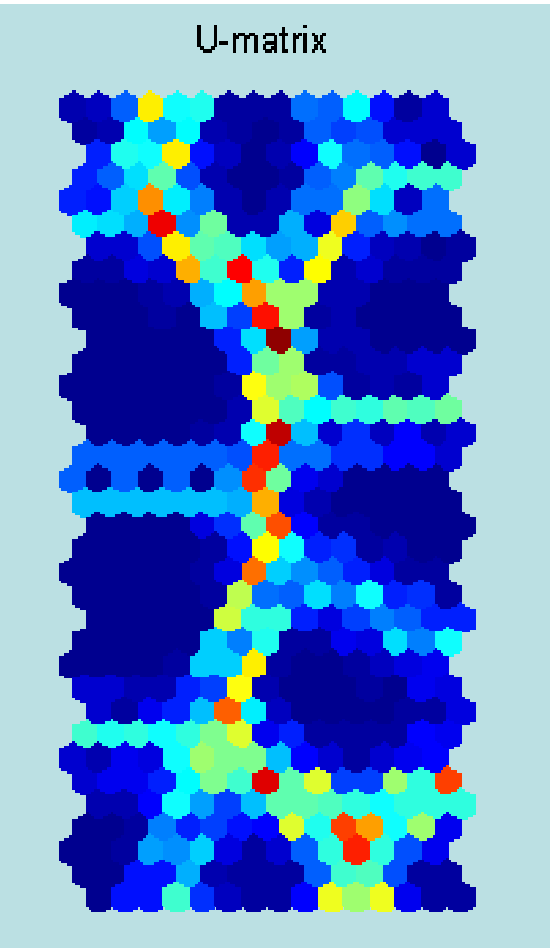
Label sample 10 wolf does not match cell 29 fox has qer 0.71732

Label sample 6 hawk does not match cell 36 owl has qer 0.69466

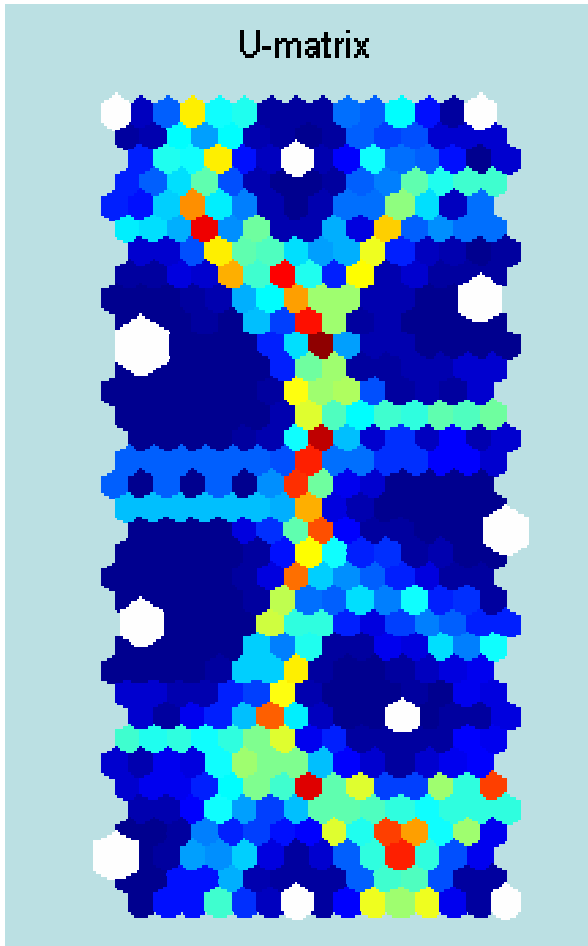
Quantization error is the distance between a sample vector and its best matching cell.

Small map condenses (clumps) data

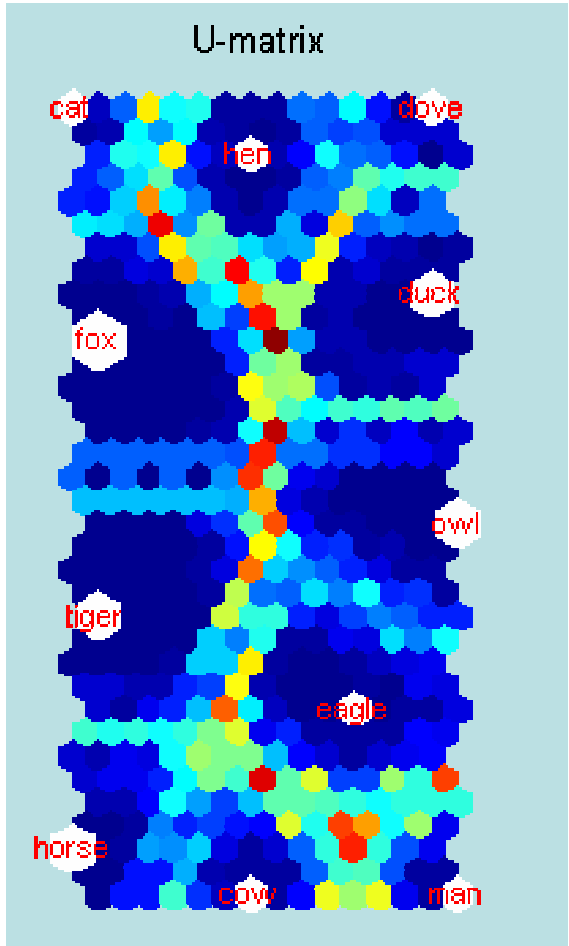
“Animals” U-matrix (18x 8 Map)



U-Matrix



U-Matrix with “Hits”



U-Matrix with “Labels”

“Animals” 18x8 Map Label Duplicates

Label sample 9 dog does not match cell 6 fox has qer 0.00012003

Label sample 10 wolf does not match cell 6 fox has qer 0.00012003

Label sample 13 lion does not match cell 12 tiger has qer 0.00024108

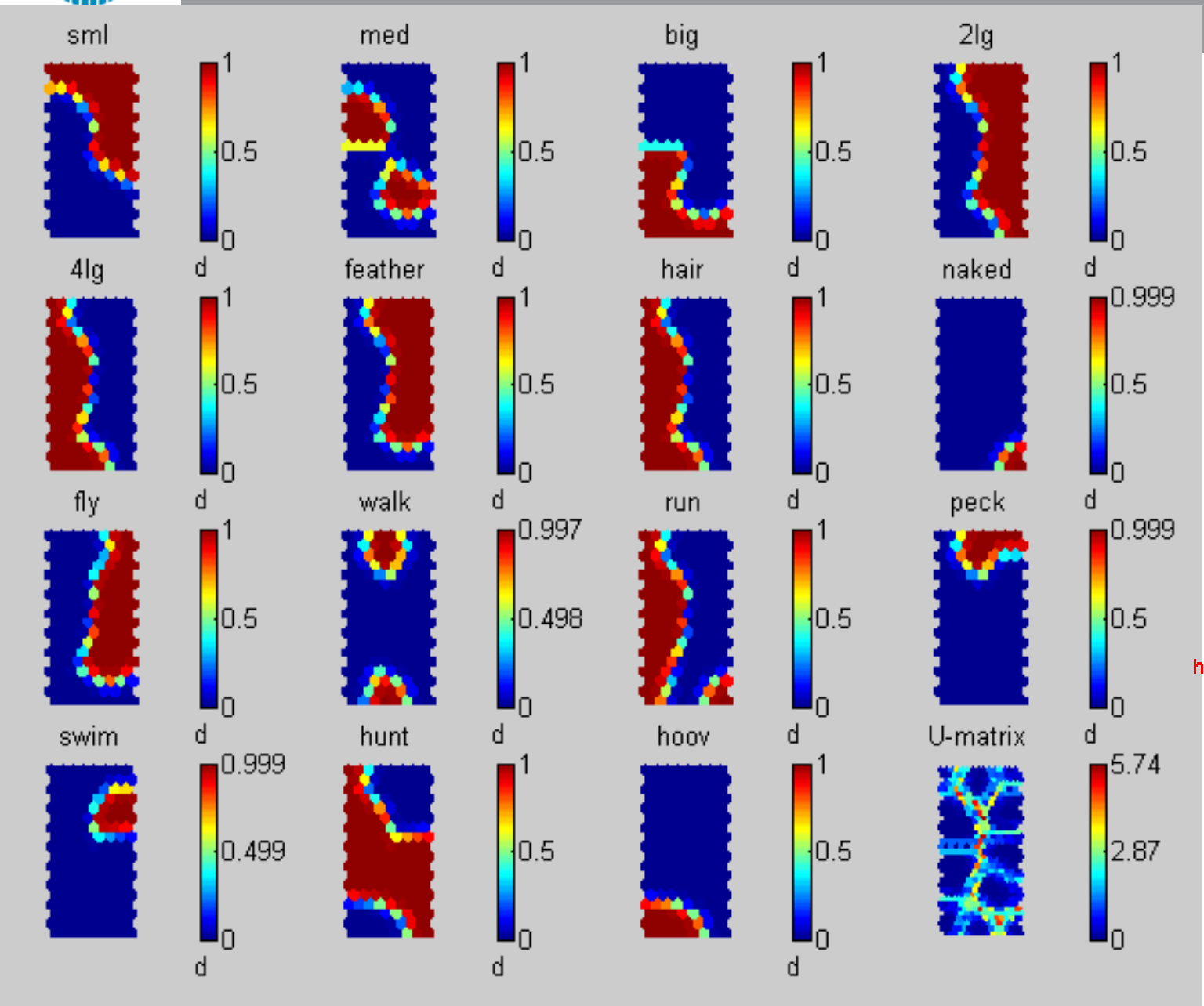
Label sample 15 zebra does not match cell 17 horse has qer 0.0027026

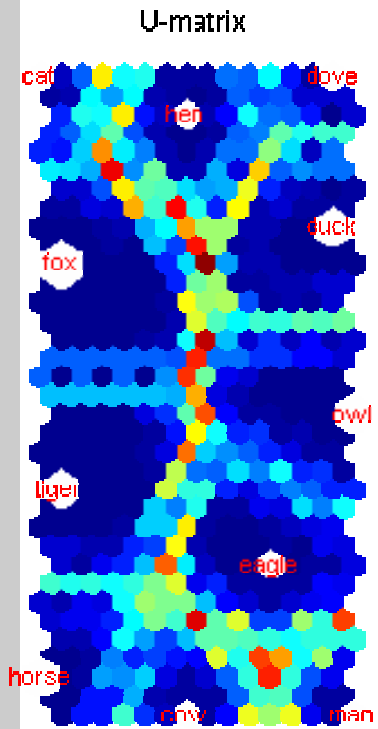
Label sample 4 goose does not match cell 131 duck has qer 0.0055947

Label sample 6 hawk does not match cell 136 owl has qer 0.00057349

Larger map spreads (splits) data allowing more information (species differences) to be seen

Component Plots 18x8





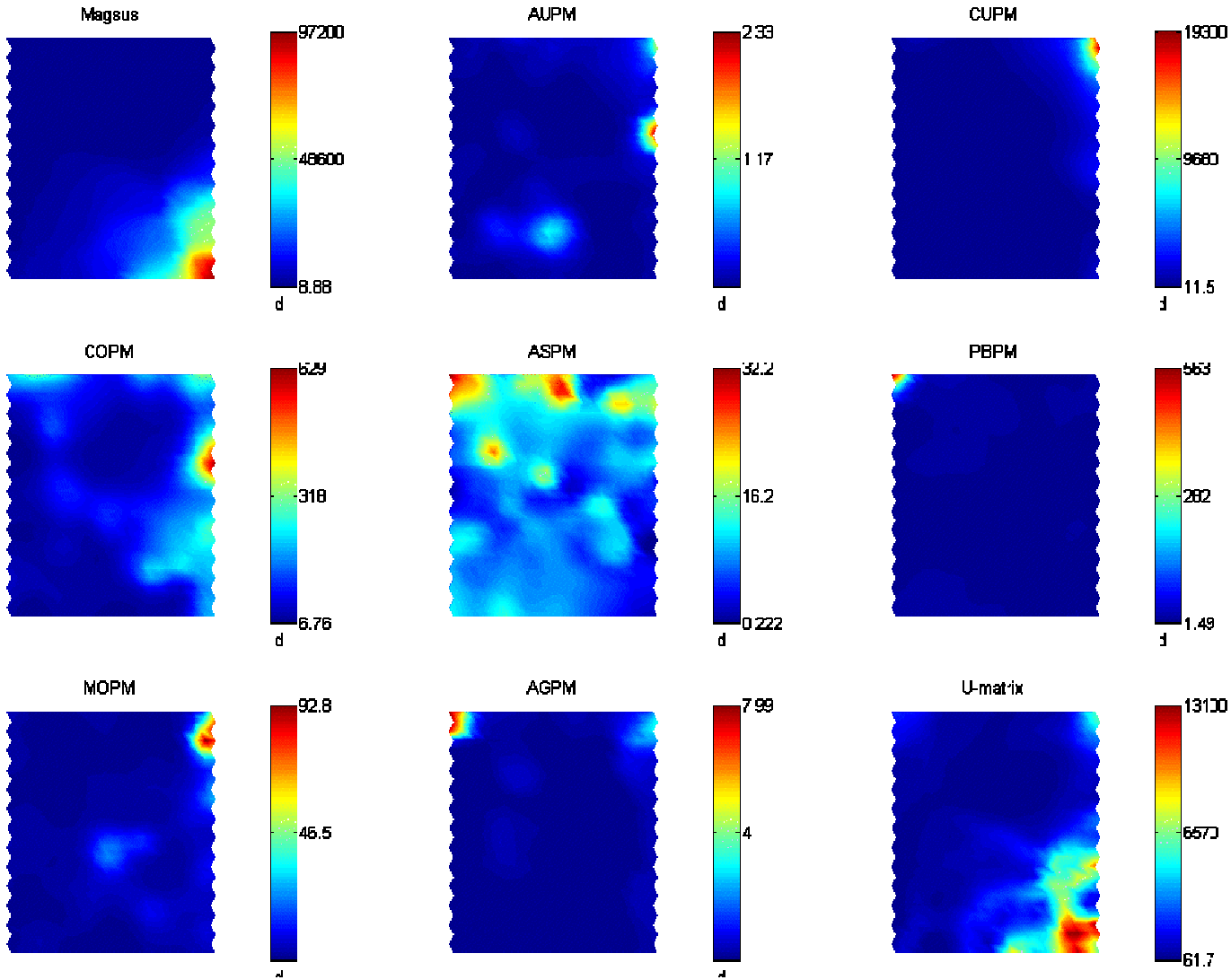
~ 15,000 RAB, RC, Air Core drill holes:

~ 40,000 located (XYZ) geochemical samples with up to 13 elements assayed:

~ 60% of data base is “empty”

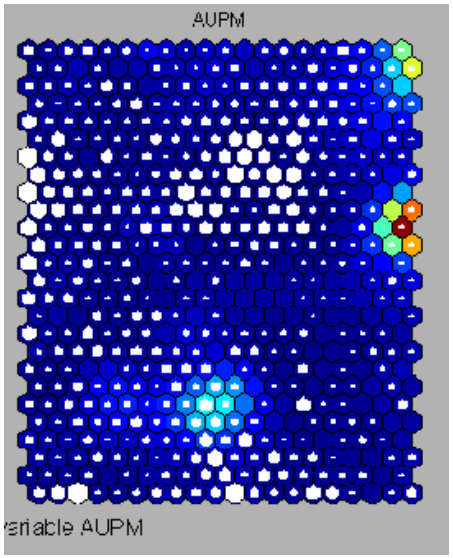
Area “A” Au deposit

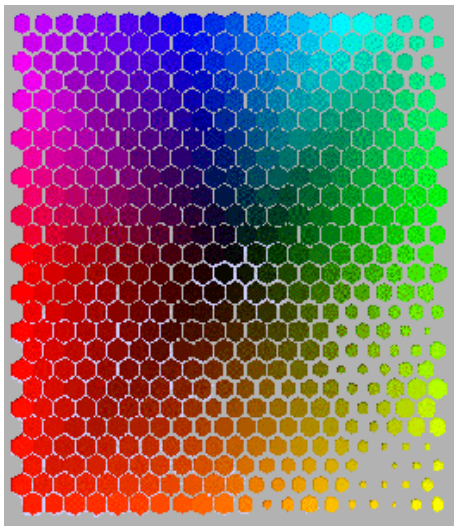
Area 1

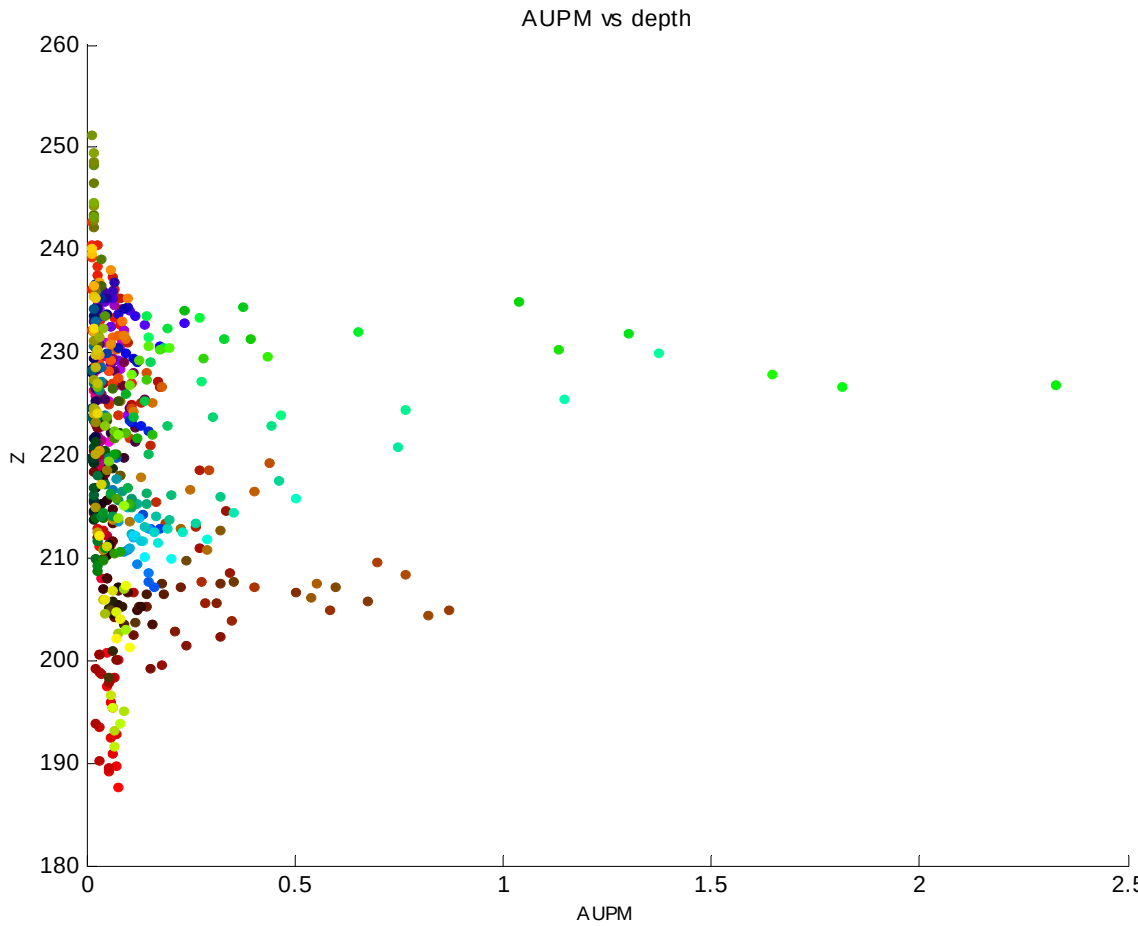


Component Plots

Area 1





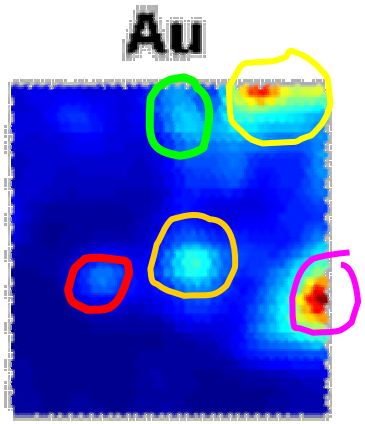


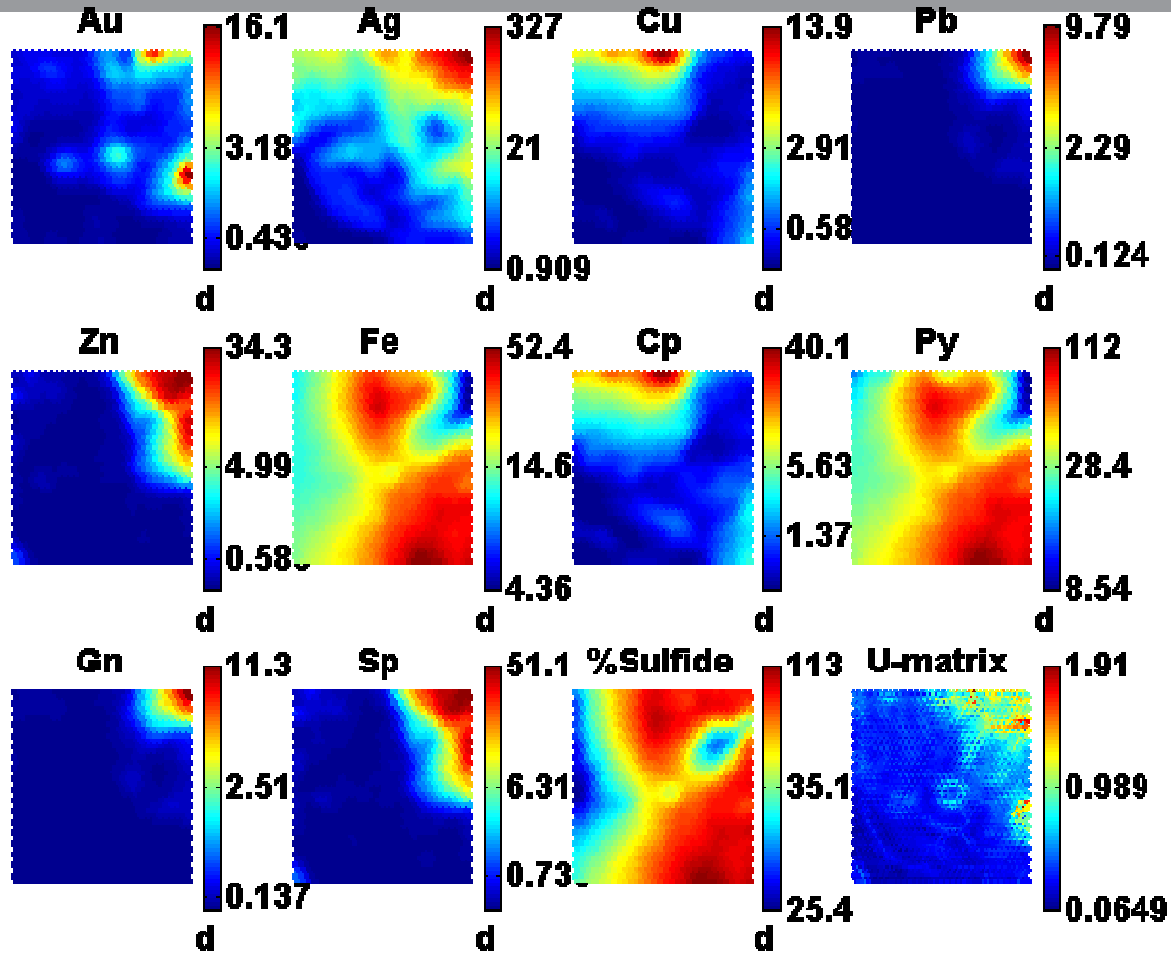
Au Vs Depth

3D Drill Hole Geochemistry & Mineralogy

Au Distribution in Mine
(10,000 samples x 10 variables)

Component Plots





5 main Au populations present:

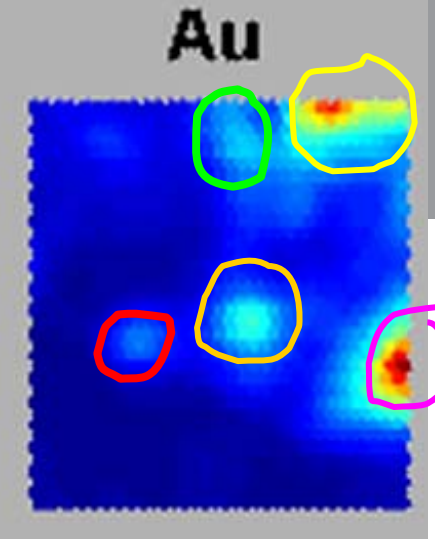
Top-Right – Au (s) with Ag(s), Pb(s) Zn(s)

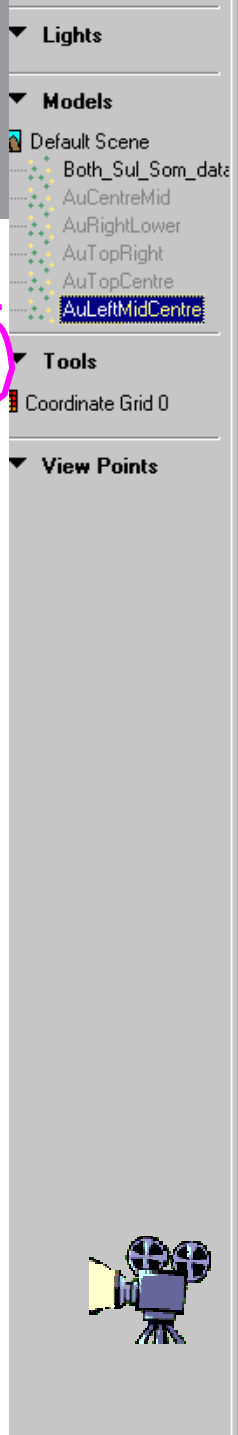
Lower-Right – Au(s) with Ag (m), Zn(w), Cu(w)

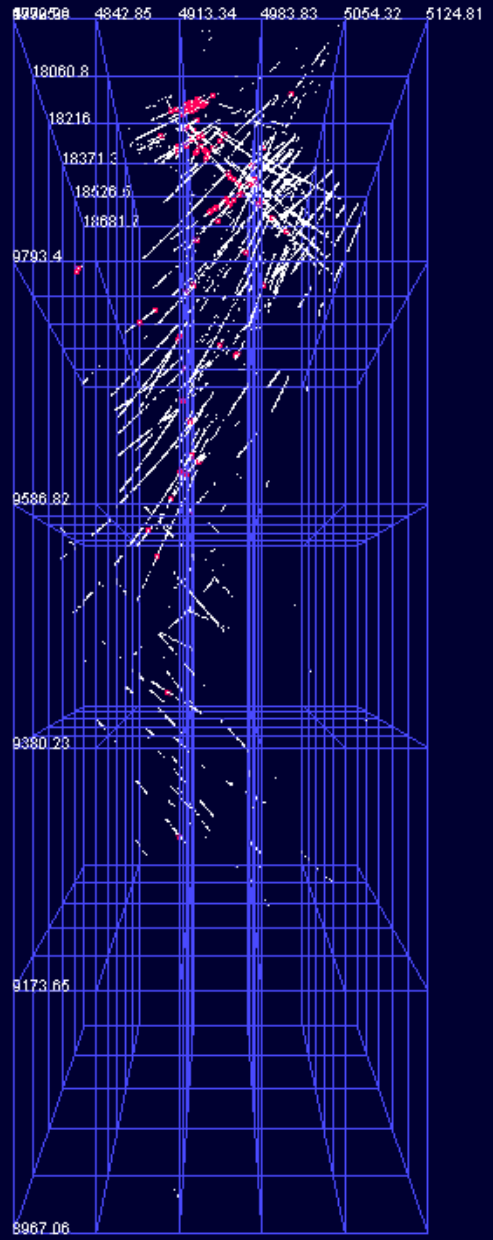
Top-Centre – Au(m), with Ag(m), Cu(s)

Mid-Centre – Au (m) with Ag(m),Cu(w)

Left-Mid-Centre – Au(m-w) with Ag (m-w) & Cu(m-w)







Au Section
Looking North

~10,000 samples

x 10 elements

R-G-Y-P-B

Area “C”

High sulphidation Au deposit

Geochemistry with “alteration” labels (not included in the processing, but carried through as labels)

Area “C”
Extract from Data Base

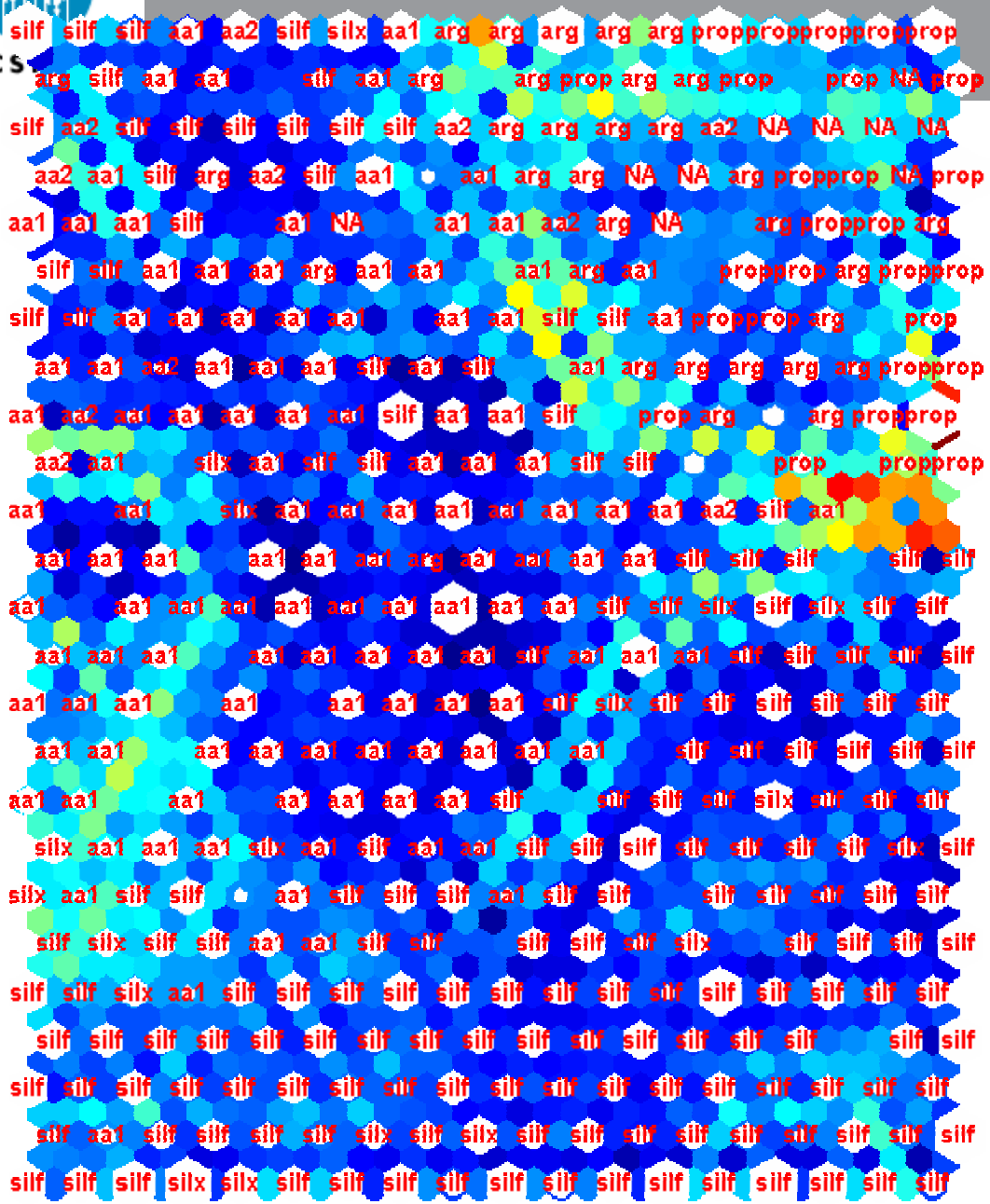
CAPCT	CDPPM	COPPM	CUPPM	FEPCT	KPCT	MGPCT	MNPPM	MOPPM	NAPCT	NIPPM	PBPPM	SBPPM	SNPPM	TLPPM	VPPM	ZNPPM	Alt
			25.0	0.430			37.0	5.0			1120.0	240.0				11.00	aa1
			17.0	1.770			20.0	2.5			1439.0	162.0				6.00	aa1
			10.0	0.670			17.0	2.5			1556.0	243.0				6.00	aa1
			16.0	1.260			17.0	2.5			2076.0	83.0				6.00	aa1

~2500 RC Chip & Core samples (2m composites)

~ 20 Elements & Alteration Label (propylitic – silica flooding)

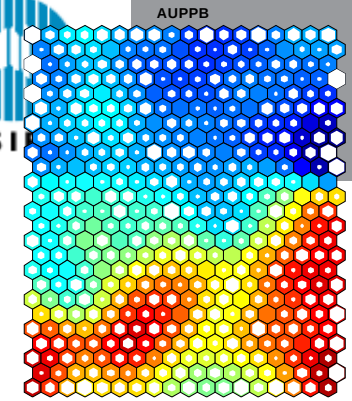
U matrix + crisp hits

U-matrix

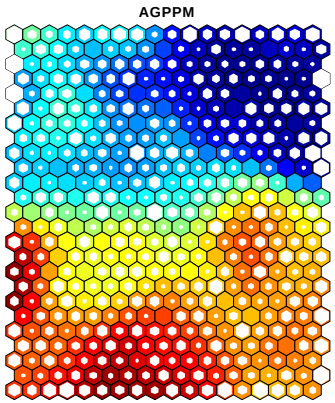


Can relate Alteration Mineralogy to Geochemistry

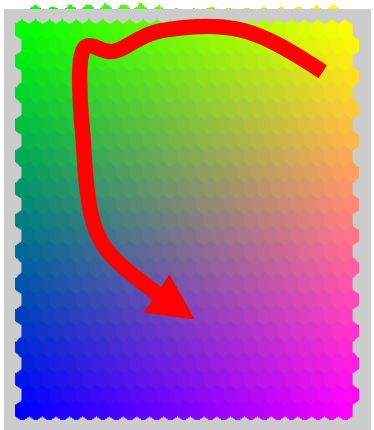
Area “C”
Au vs Ag Scatter Plot of BMU - Nodes

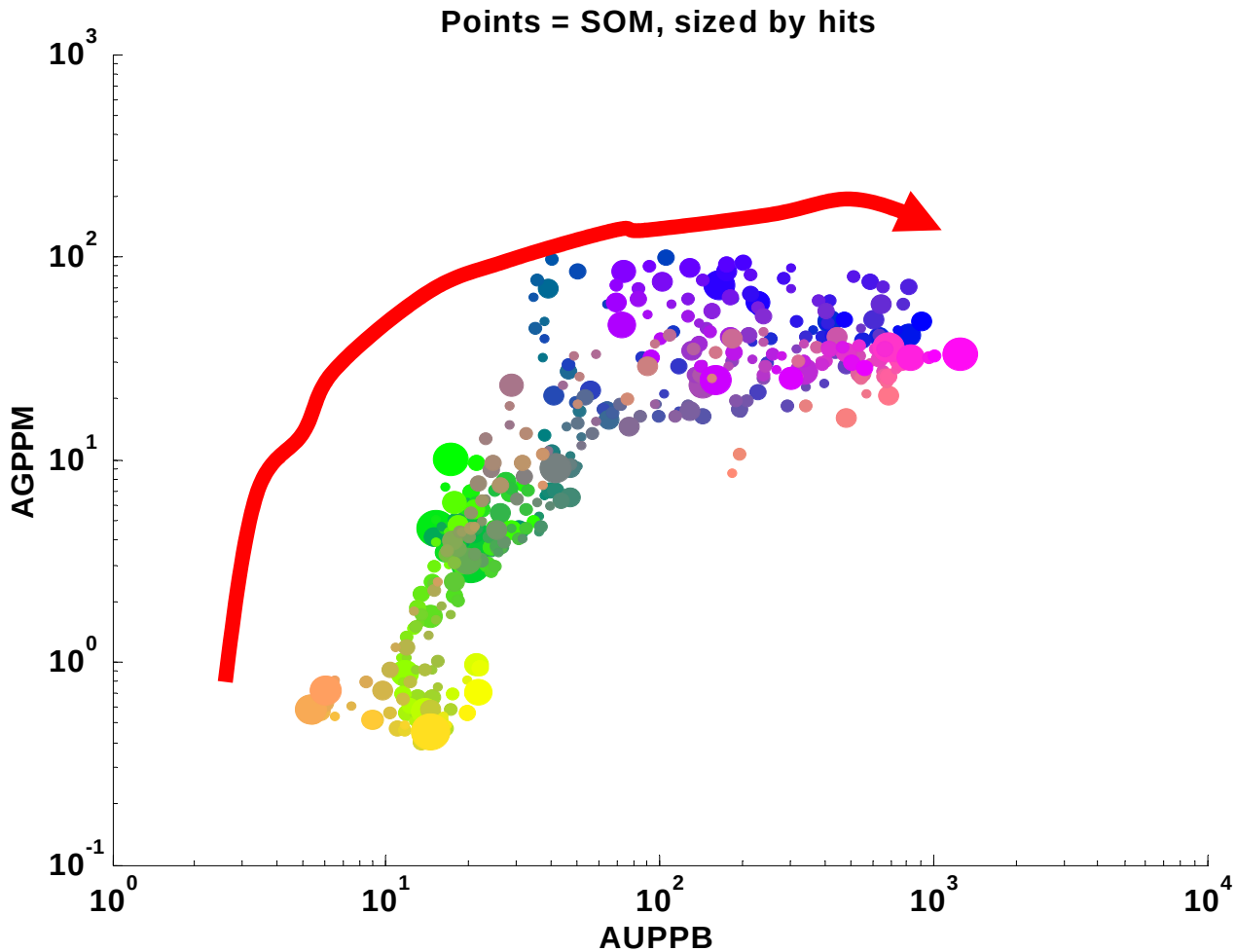


variable AUPPB



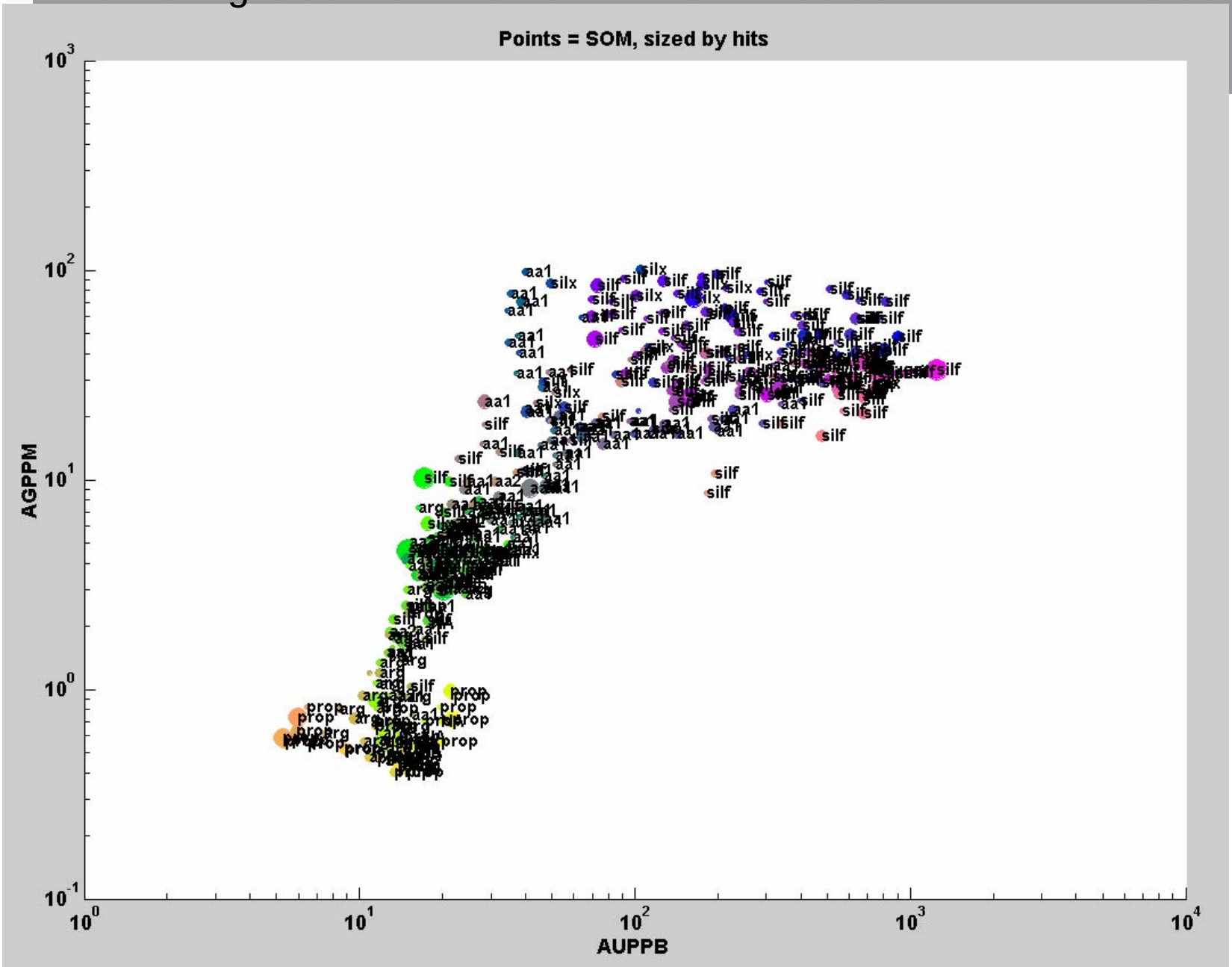
variable AGPPM



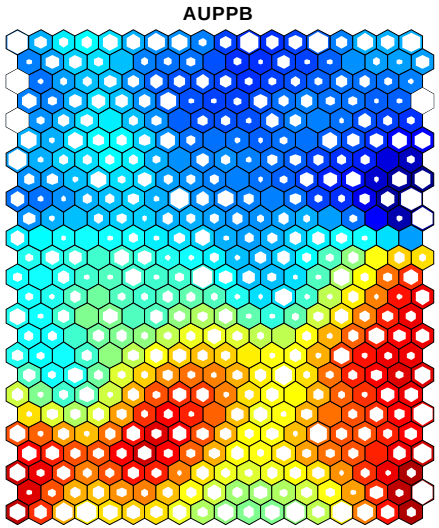


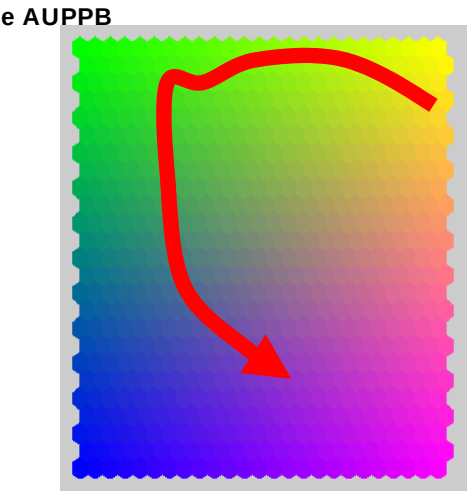
SOM can assist in showing the “process” of mineralization (ie, vectors to ore!!)

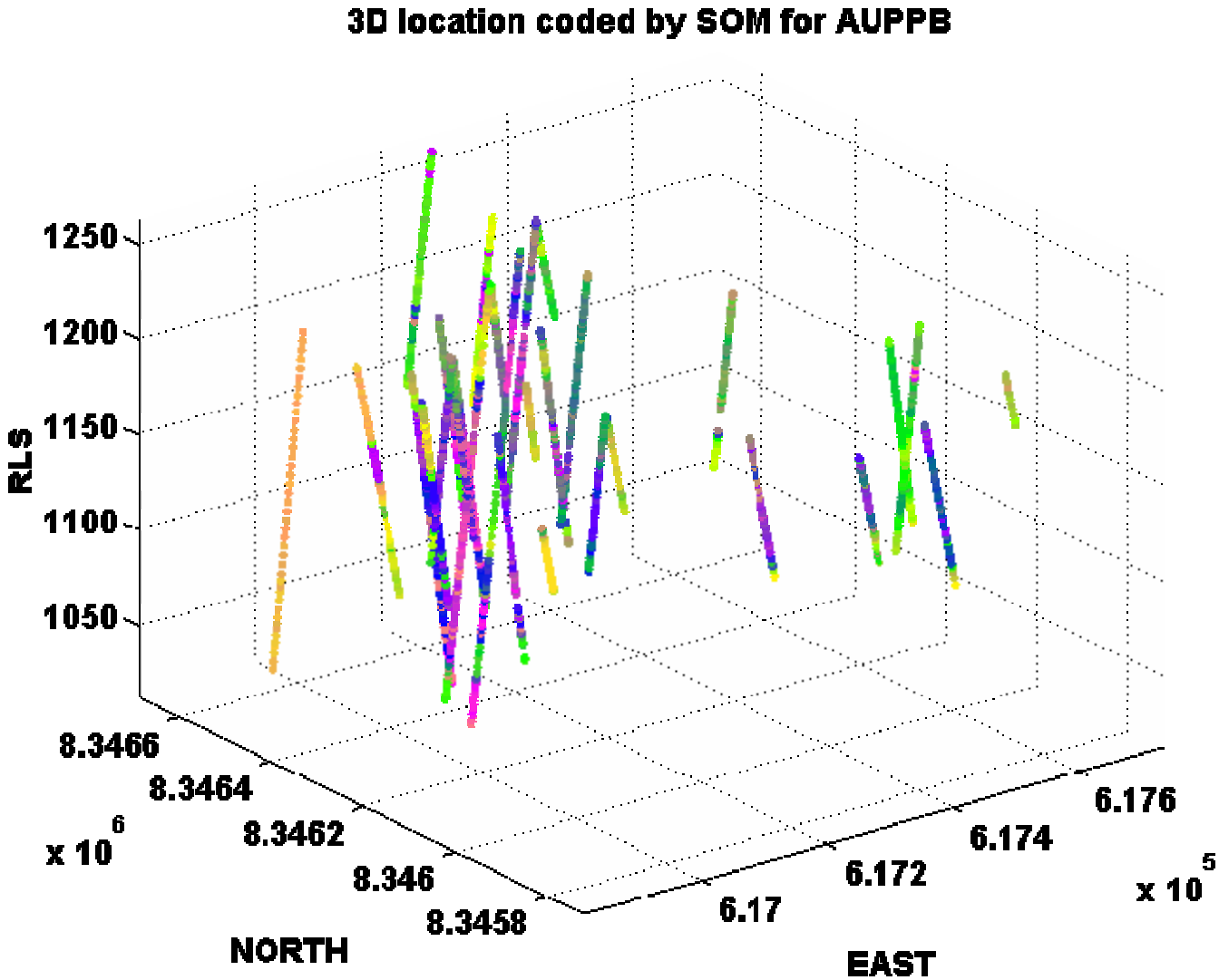
Au vs Ag Scatter Plot of BMU – Nodes with "Alt" labels

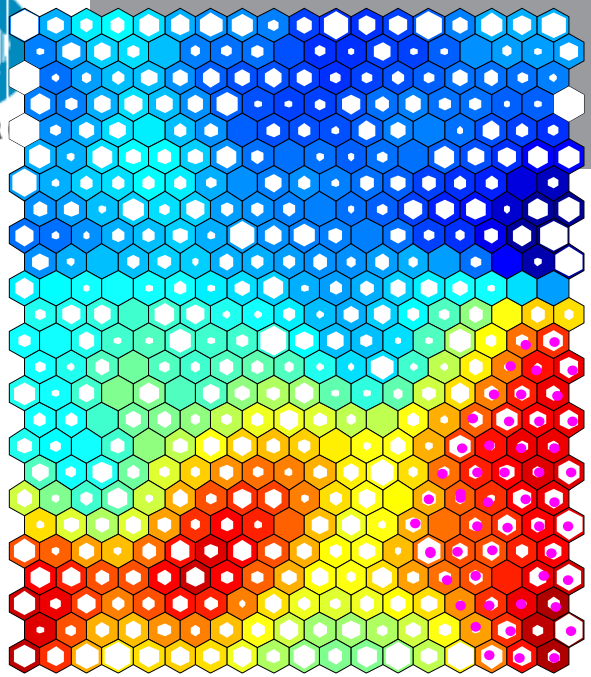


All Au Samples – Colour-coded by SOM Colour - LUT

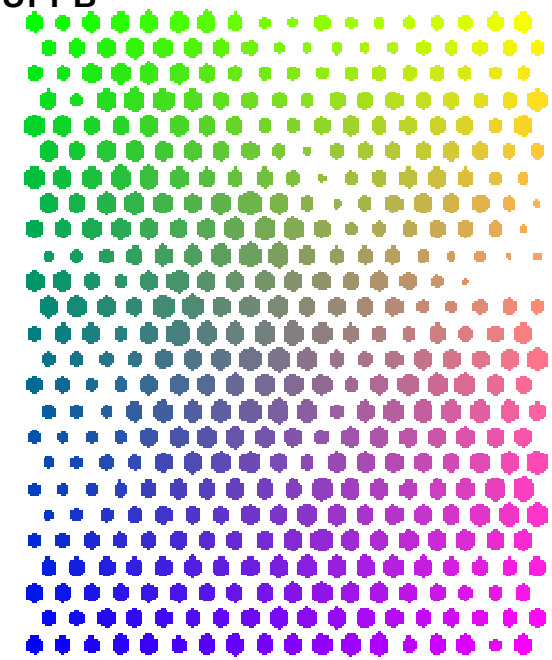




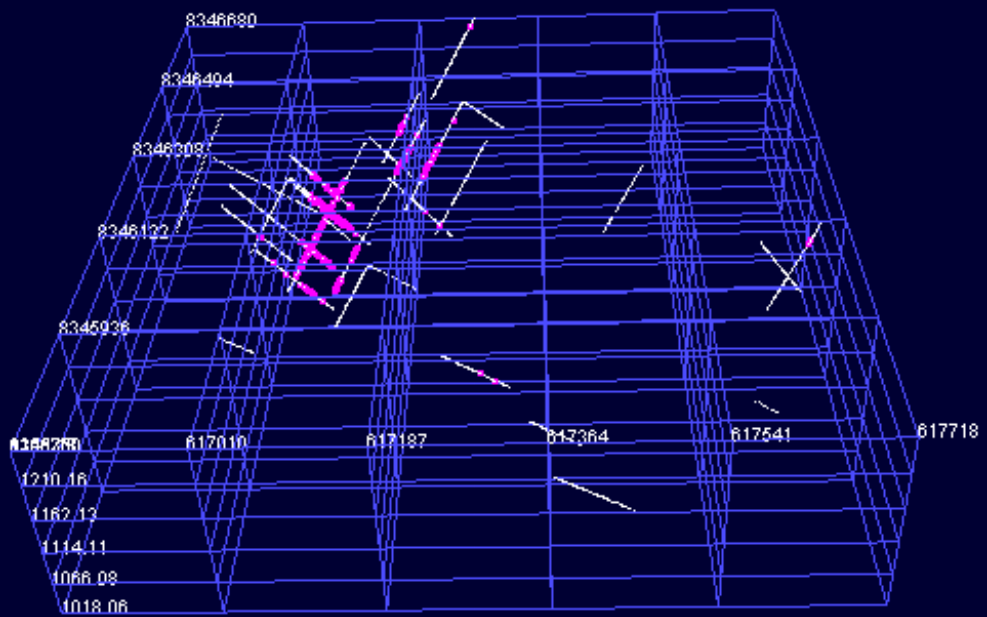




variable AUPPB



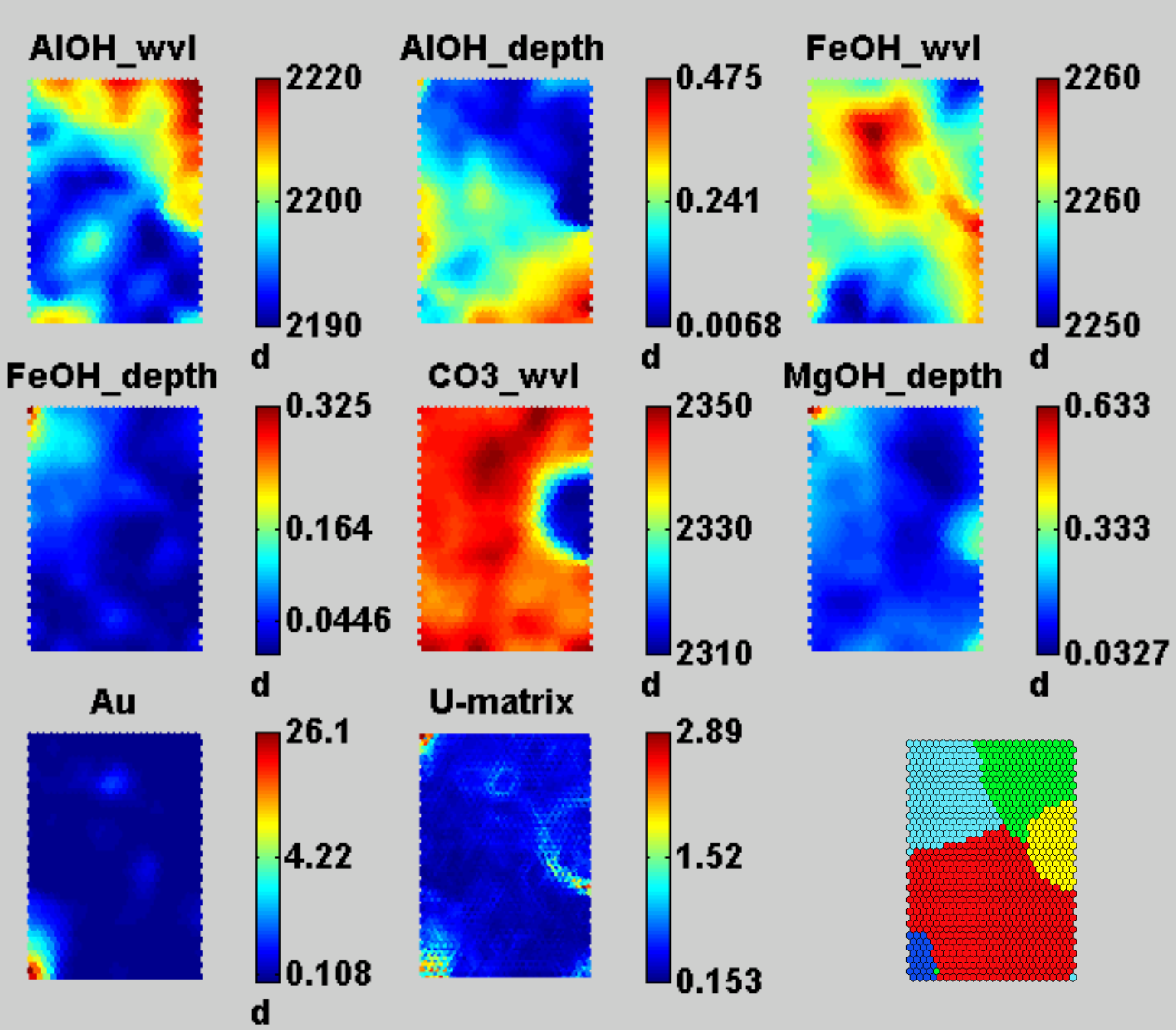
Bottom Right Au Samples – SOM-coded



FracView

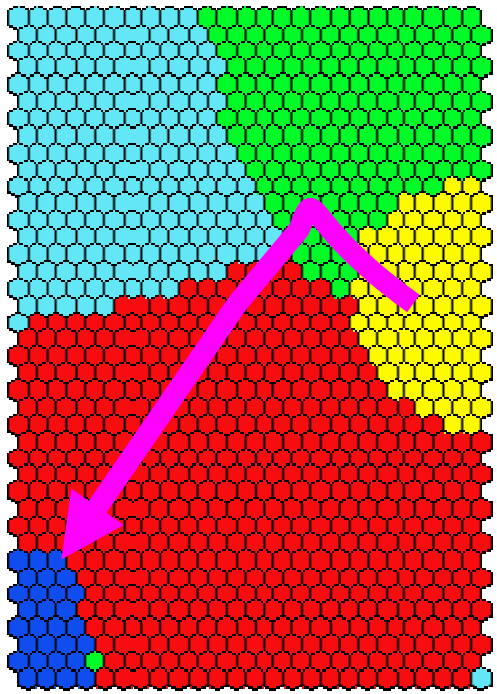
CSIRO Self Organizing Maps

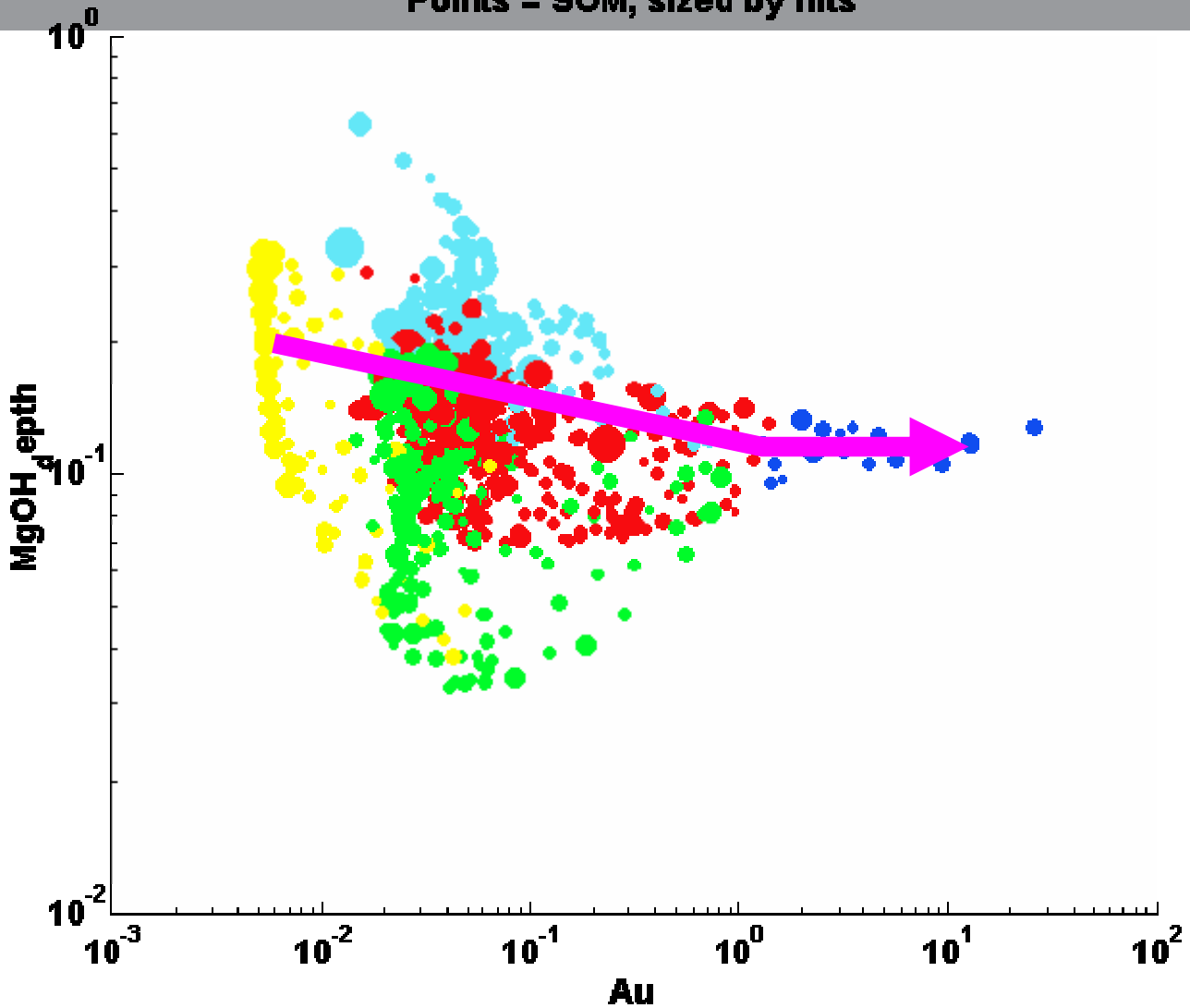
Core Spectra & Geochemistry



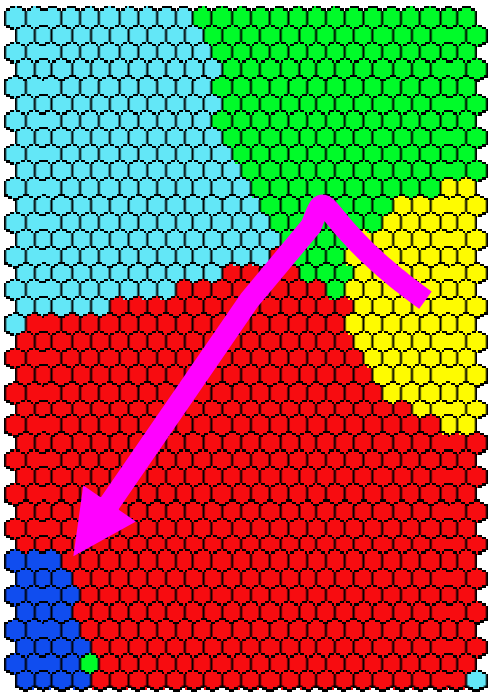
Component Plots – Extracted Spectral Features & Au

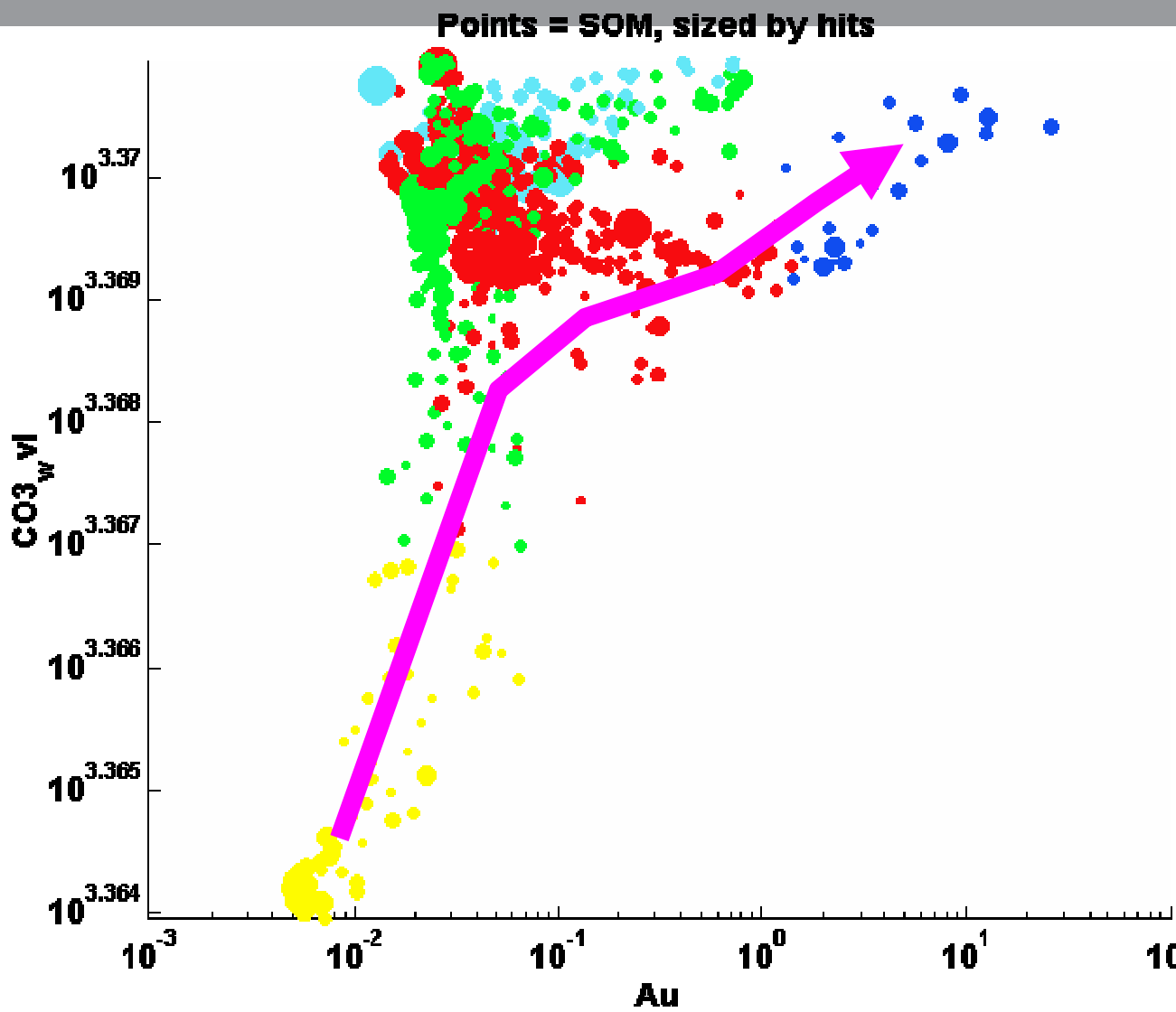
MgOH depth vs [Au] Scatter Plot of BMU - Nodes



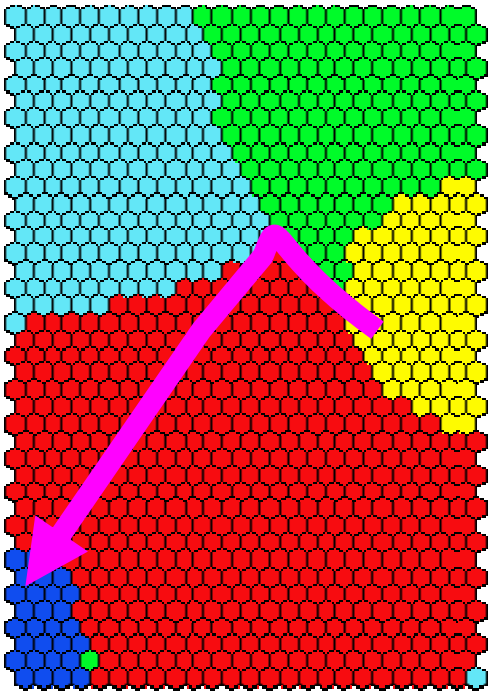


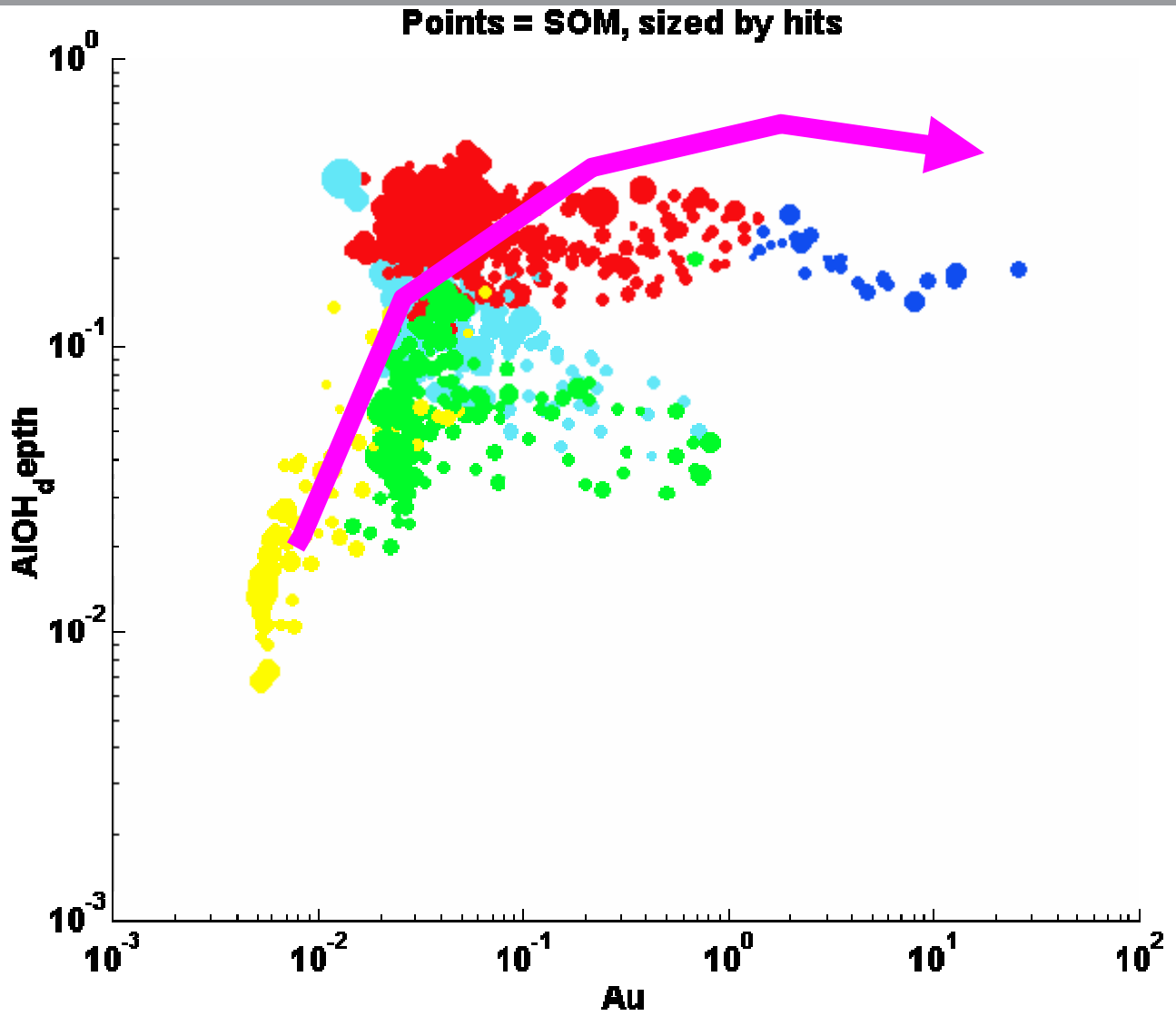
CO_3 wvl vs [Au] Scatter Plot of BMU - Nodes

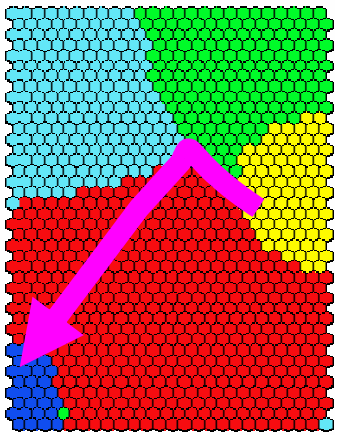




AlOH depth vs [Au] Scatter Plot of BMU - Nodes







SOMe General Conclusions

SOM is an unsupervised, data-driven, exploratory data analysis tool;

Non-traditional Non-Statistical approach to data analysis;

Ideal for “sparse” geological data

Opens the door to “Integrated Analysis and Interpretation of Disparate Data Types”;

The spatial coherence and juxtaposition of SOM “clusters” is important;

Scatterplots of SOM nodes highlight geological “process”;

Thank you for your time and interest

THE END

QUESTIONS ?