Advances in Self-Organizing Maps

Dr. Lutz Hamel
Dept. of Comp. Sci & Stats
University of Rhode Island

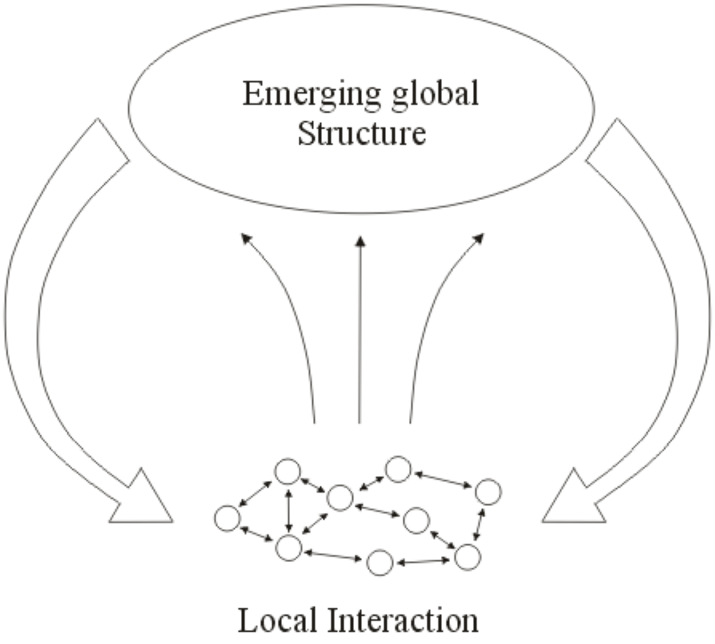
Overview

- Self-Organization
- Basic SOM Algorithm
- Applications of SOM we have worked on
- Model "Goodness of Fit"
  - Standard Approaches, e.g., Quantization error
  - New Approach: Convergence Test with 2-Sample test.
- New Approaches to SOM Visualization
  - Connected Components
  - Cartograms

Self-Organization and Learning

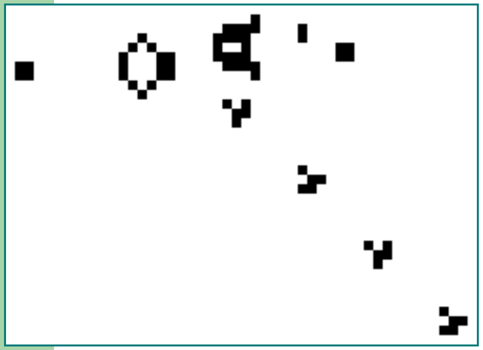
- **Self-organization** refers to a process in which the internal organization of a system increases automatically without being guided or managed by an outside source.
- This process is due to local interaction with simple rules.
- Local interaction gives rise to global structure.

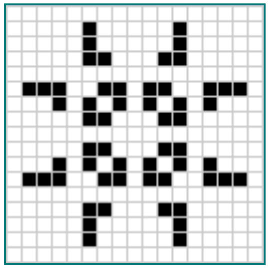
☞ We can interpret emerging global structures as learned structures.
☞ Learned structures appear as clusters of similar objects.



Complexity : Life at the Edge of Chaos, Roger Lewin, University Of Chicago Press; 2nd edition, 2000

Game of Life





- Most famous example of self-organization - Game of Life
- Simple local rules:
  - Any live cell with fewer than two live neighbours dies, as if caused by under-population.
  - Any live cell with two or three live neighbours lives on to the next generation.
  - Any live cell with more than three live neighbors dies, as if by overcrowding.
  - Any dead cell with exactly three live neighbors becomes a live cell, as if by reproduction.

Source: http://en.wikipedia.org/wiki/Conway's_Game_of_Life

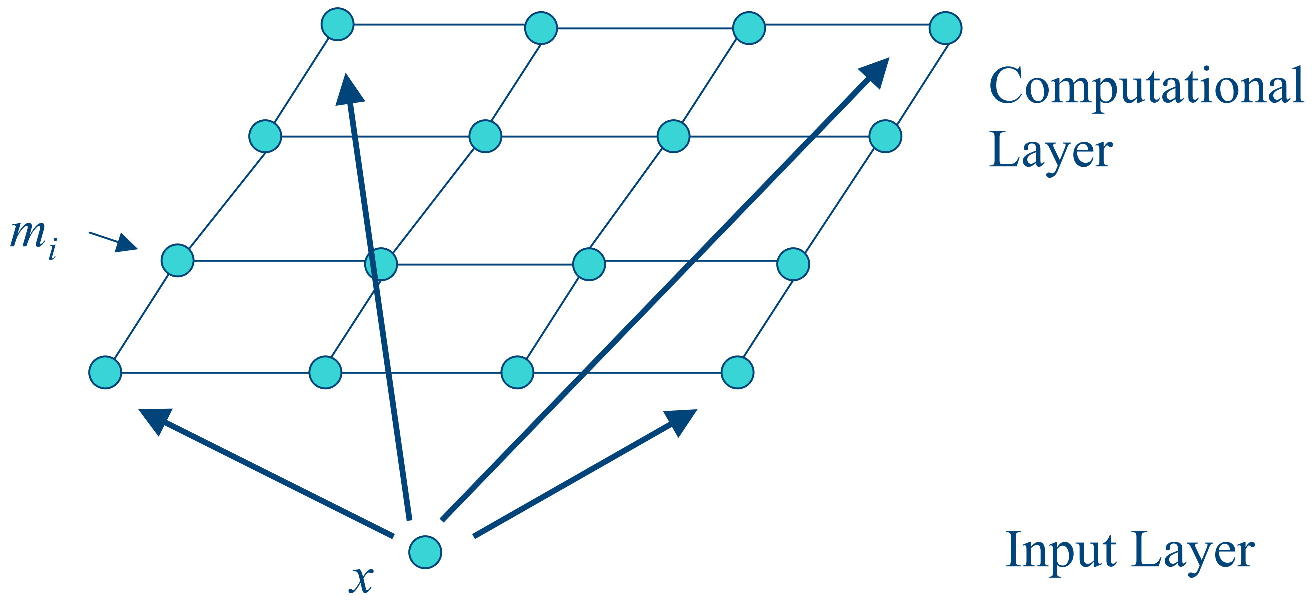
Supervised vs.Unsupervised Learning

- In supervised learning we train algorithms with predefined concepts and functions based on labeled data, e.g.
  $D = \{ (x, y) \mid x \in X, y \in \{yes,no\}.$
- In unsupervised learning we are given a set of instances *X* (without labels) and we let the algorithm discover interesting properties of this set.
- Most unsupervised learning algorithms are based on the idea of discovering *similarities* between elements in the set *X.*

SOM Architecture

- A feed-forward neural network architecture based on competitive learning invented by Teuvo Kohonen in 1981.
- Does not depend on *a priori* selection of number of clusters to search for – will find the appropriate number of clusters for given the set of instances.
- Sometimes is considered a 2D projection of clusters in high-dimensional space.

SOM Architecture

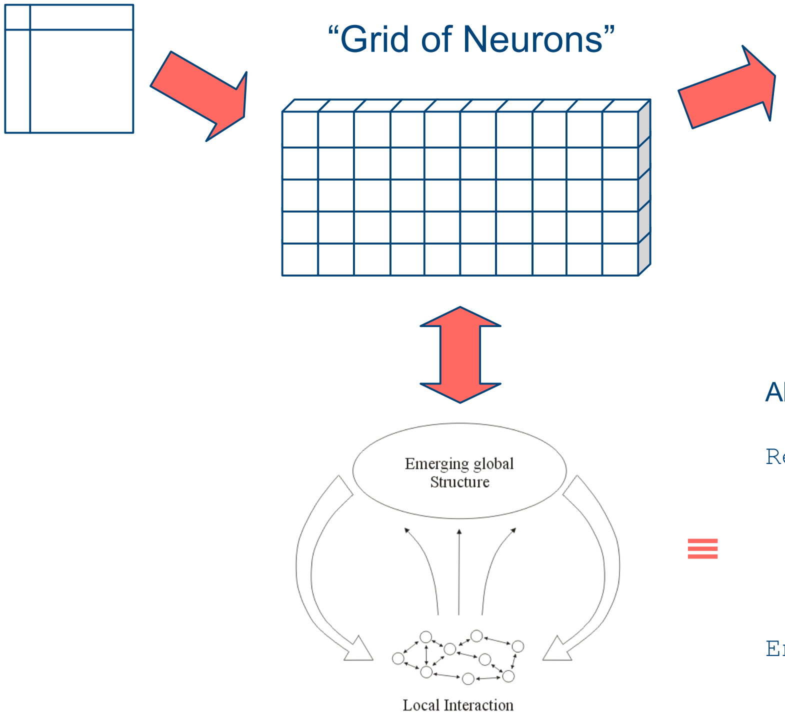


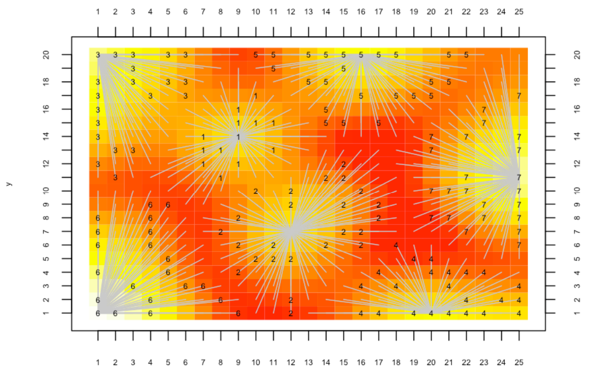
- SOM has a feed-forward structure with a single computational layer arranged in rows and columns.
- Each neuron is fully connected to the input node in the input layer.
- The goal is to organize the neurons in the computational layer into clusters/regions associated with patterns in the instance set *X.*

Self-Organizing Maps

Data Table

“Grid of Neurons”





Visualization

≡

Algorithm:

```
Repeat until Done
  For each observation in Data Table Do
    Find the neuron that best describes the observation.
    Make that neuron look more like the observation.
    Smooth the immediate neighborhood of that neuron.
  End For
End Repeat
```

Feature Vector Construction

In order to use SOMs we need to describe our objects

- Feature Vectors

small	medium	big	Twolegs	Fourlegs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
1	0	0	1	0	0	0	0	1	0	0	0	1

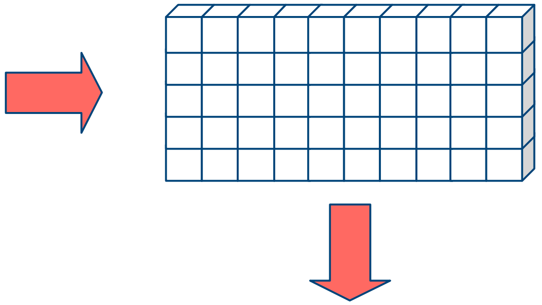
small	medium	big	Twolegs	Fourlegs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
0	0	1	0	1	1	1	0	0	0	1	0	0

Training a SOM

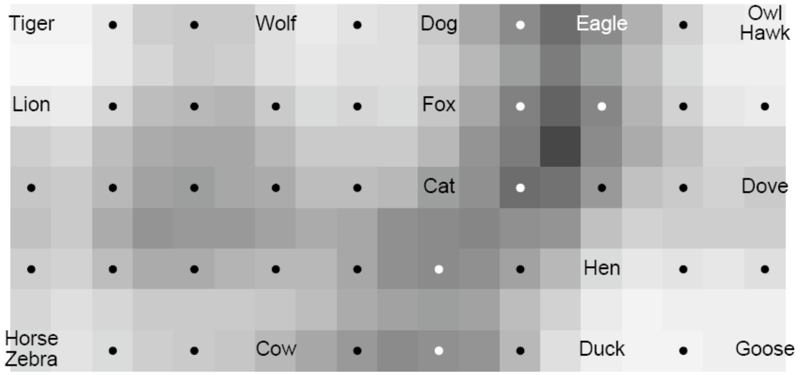
	small	medium	big	Two legs	Four legs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
Dove	1	0	0	1	0	0	0	0	1	0	0	1	0
Hen	1	0	0	1	0	0	0	0	1	0	0	0	0
Duck	1	0	0	1	0	0	0	0	1	0	0	0	1
Goose	1	0	0	1	0	0	0	0	1	0	0	1	1
Owe	1	0	0	1	0	0	0	0	1	1	0	1	0
Hawk	1	0	0	1	0	0	0	0	1	1	0	1	0
Eagle	0	1	0	1	0	0	0	0	1	1	0	1	0
Fox	0	1	0	0	1	1	0	0	0	1	0	0	0
Dog	0	1	0	0	1	1	0	0	0	0	1	0	0
Wolf	0	1	0	0	1	1	0	1	0	1	1	0	0
Cat	1	0	0	0	1	1	0	0	0	1	0	0	0
Tiger	0	0	1	0	1	1	0	0	0	1	1	0	0
Lion	0	0	1	0	1	1	0	1	0	1	1	0	0
Horse	0	0	1	0	1	1	1	1	0	0	1	0	0
Zebra	0	0	1	0	1	1	1	1	0	0	1	0	0
Cow	0	0	1	0	1	1	1	0	0	0	0	0	0

Table of Feature Vectors

“Grid of Neurons”



Visualization



SOMs Sample the Data Space

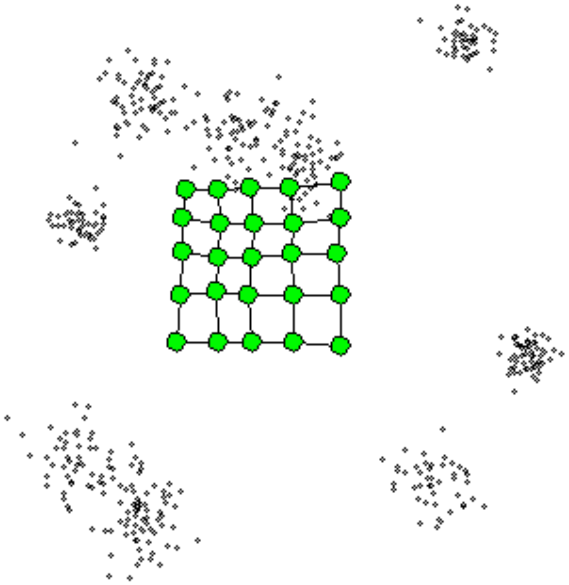


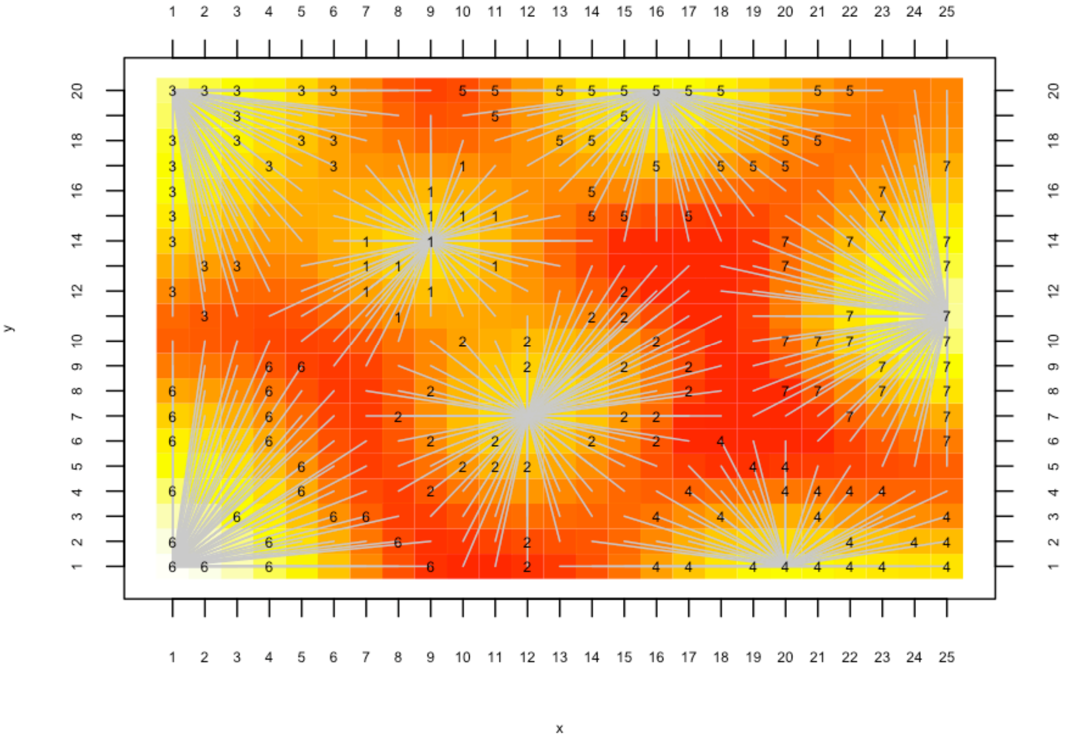
Image source: www.peltarion.com

- Given some distribution in the data space, SOM will try to construct a sample that looks like it was drawn from the same distribution.

Algorithm:

```
Repeat until Done
  For each observation in Data Table Do
    Find the neuron that best describes the observation.
    Make that neuron look more like the observation.
    Smooth the immediate neighborhood of that neuron.
  End For
End Repeat
```

SOM Visualization



Visualization of Seven clusters using SOM

Comparison

- Pros:
  - K-means - SOM does not need an *a priori* estimate of the number of clusters to look for.
  - Hierarchical Clustering - SOM can deal with ambiguity, assignment of points to multiple clusters.
- Cons:
  - Training time can be substantial, especially for large maps with lots of training data.

Applications of SOM

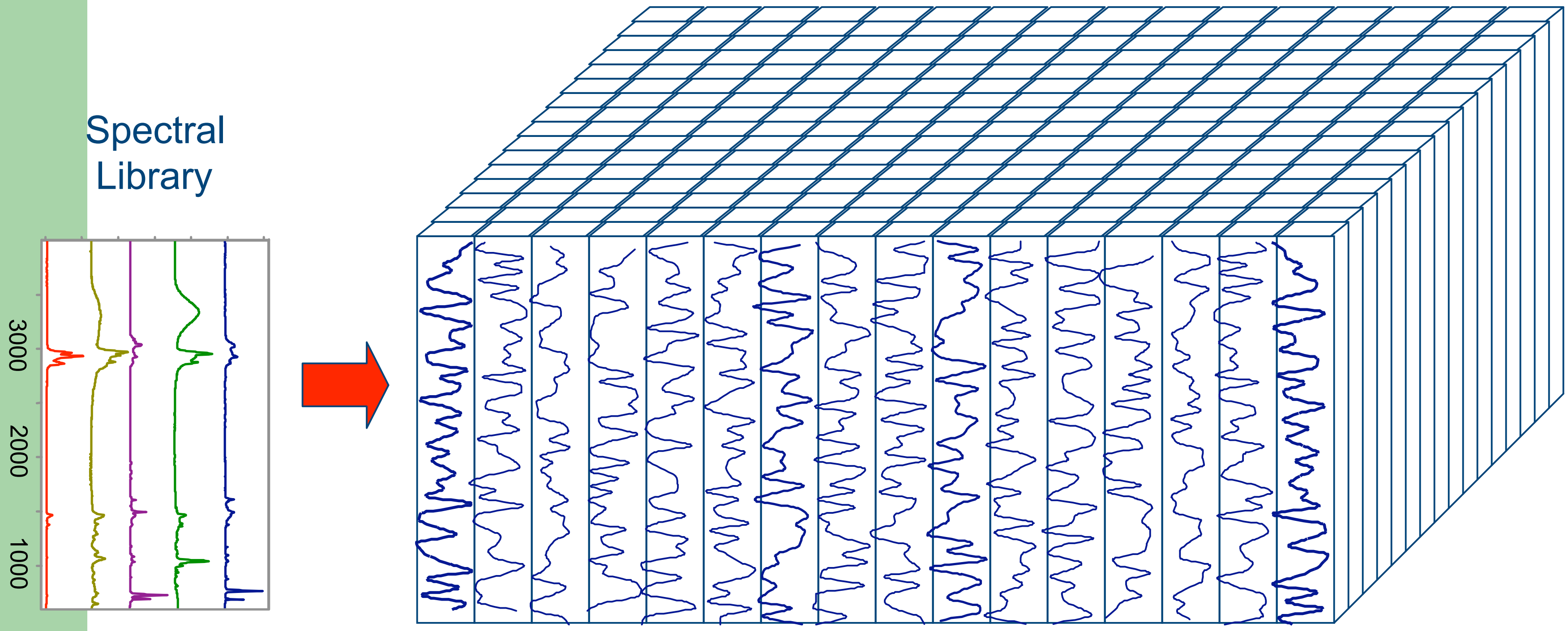
- Infrared Spectroscopy
  - Goal: to find out if compounds are chemically related without performing an expensive chemical analysis.
  - Each compound is tested for light absorbency in the infrared spectrum.
  - Specific chemical structures absorb specific ranges in the infrared spectrum.
  - This means, each compound has a specific “spectral signature”.

Sensitivity of Raman Spectra to Chemical Functional Groups, Kevin Judge, Chris W. Brown, and Lutz Hamel. Appl Spectrosc. 2008 Nov;62(11):1221-5.

Sensitivity of Infrared Spectra to Chemical Functional Groups, Kevin Judge, Chris W. Brown, and Lutz Hamel. Anal. Chem., 80 (11), 4186-4192, 2008.

Training SOM with Spectra

Grid of N

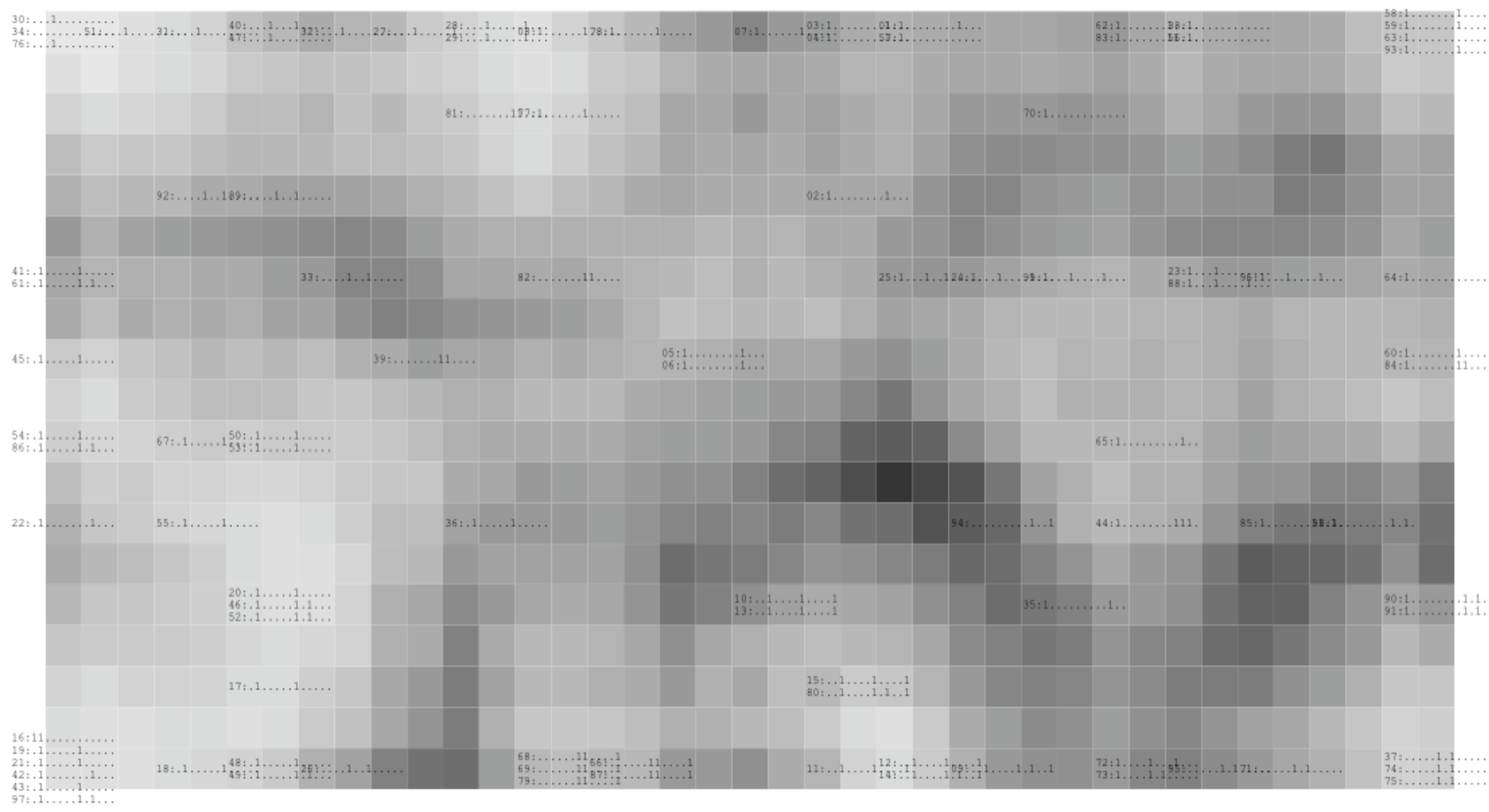


Spectral Library

Random Number Spectra

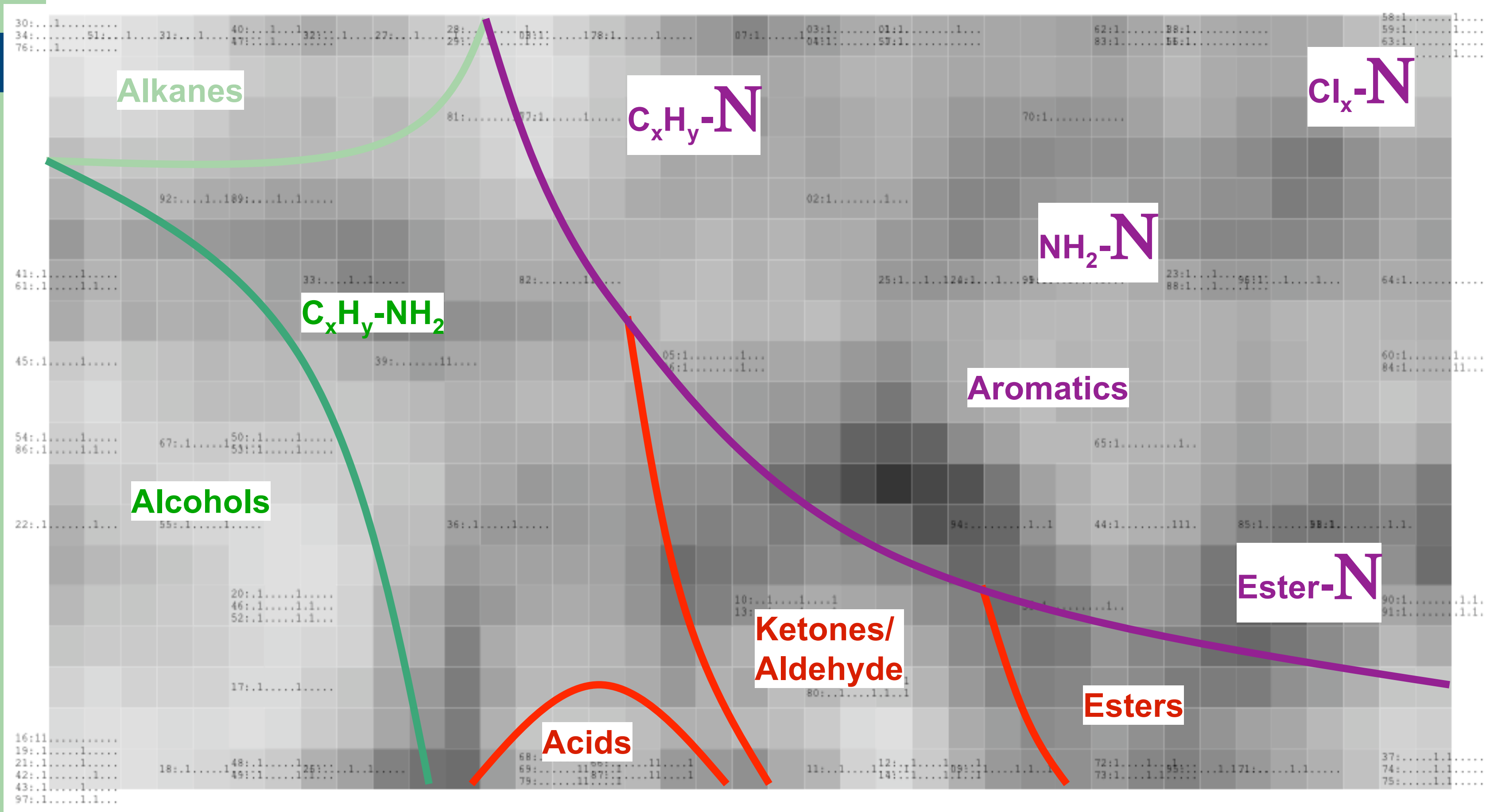
Self-Organizing-Map

MIR Spectra



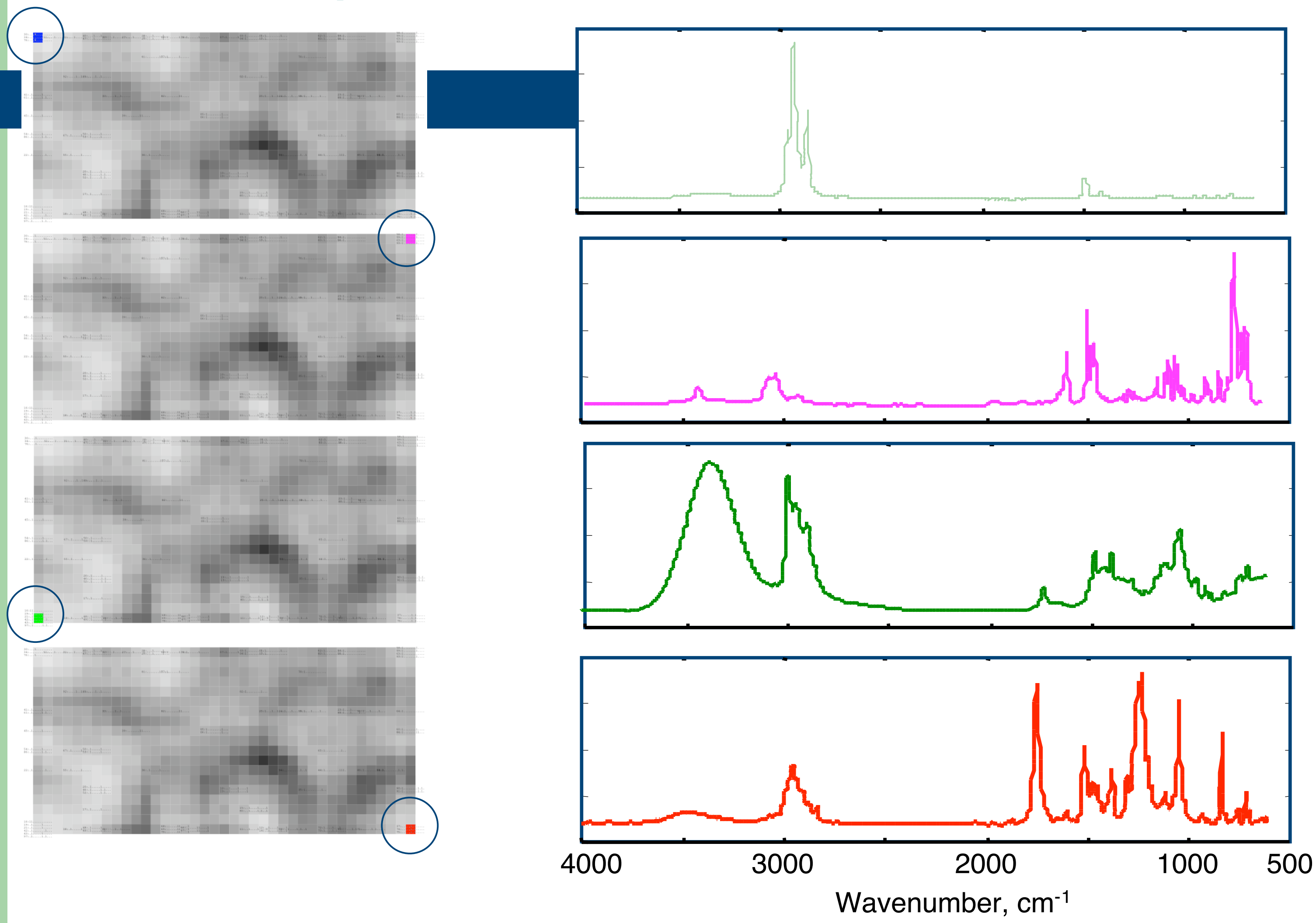
MIR SOM

Functional Groups



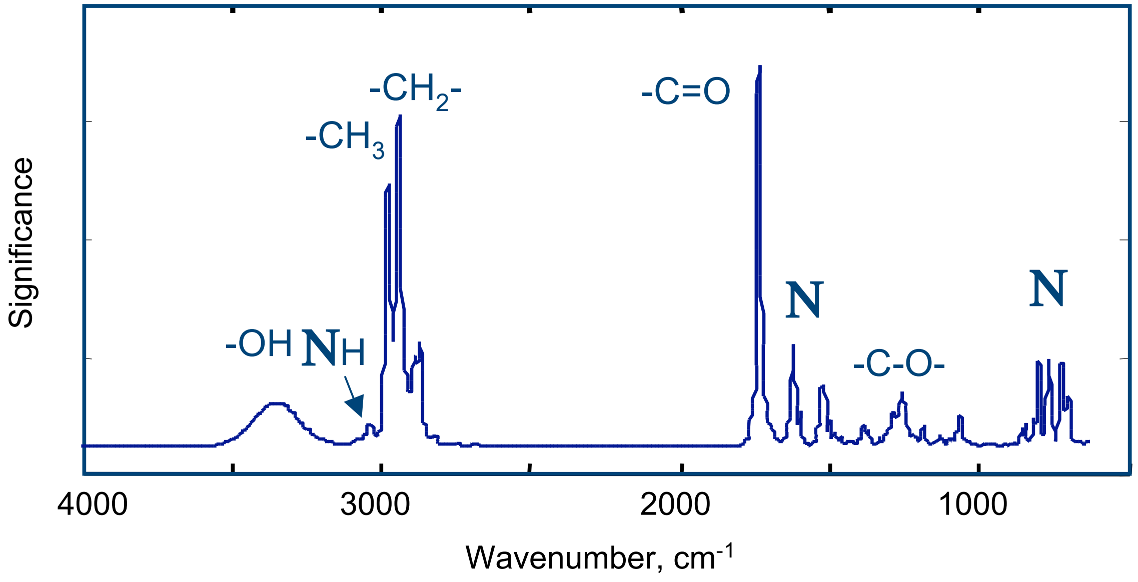
MIR

Centroid Spectra

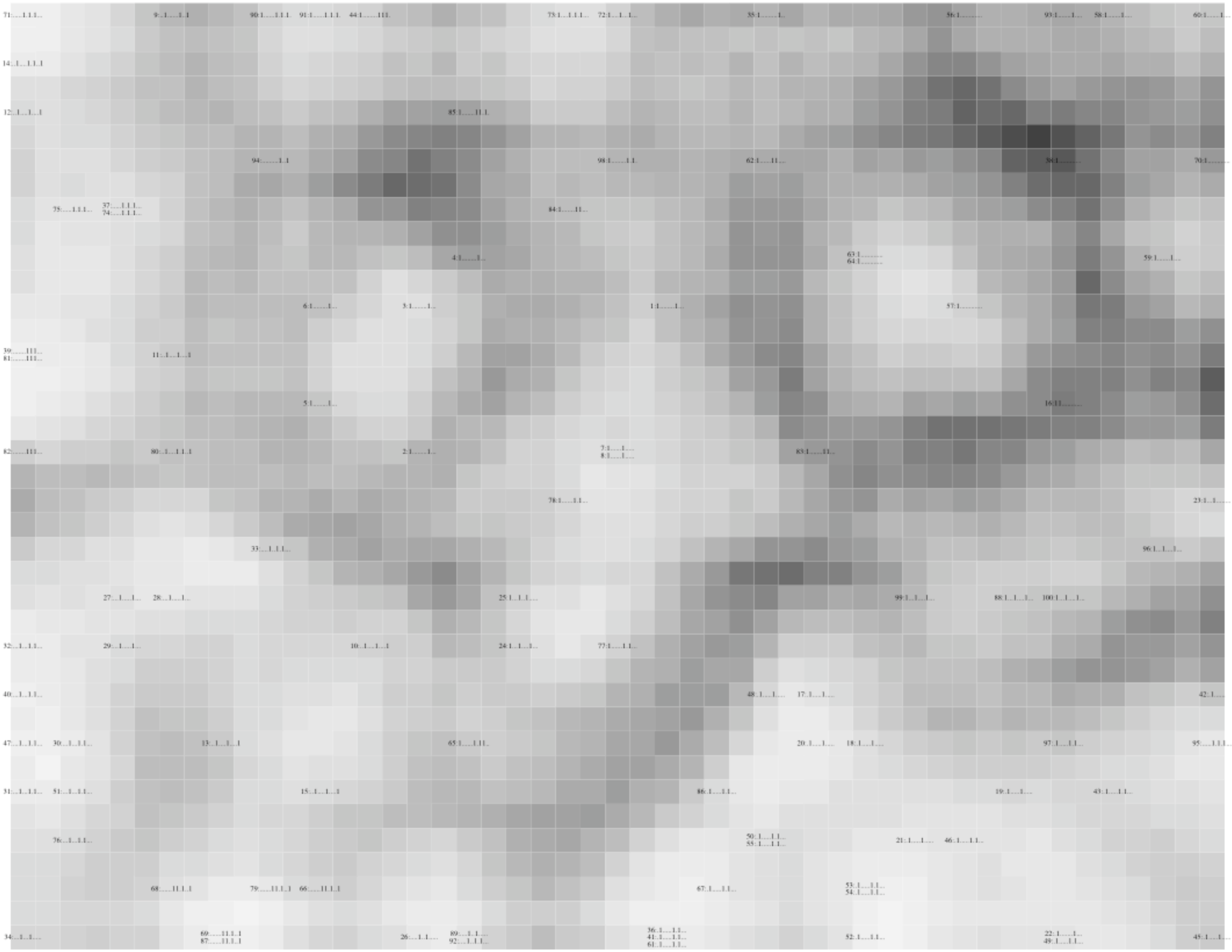


MIR

Significance Spectrum

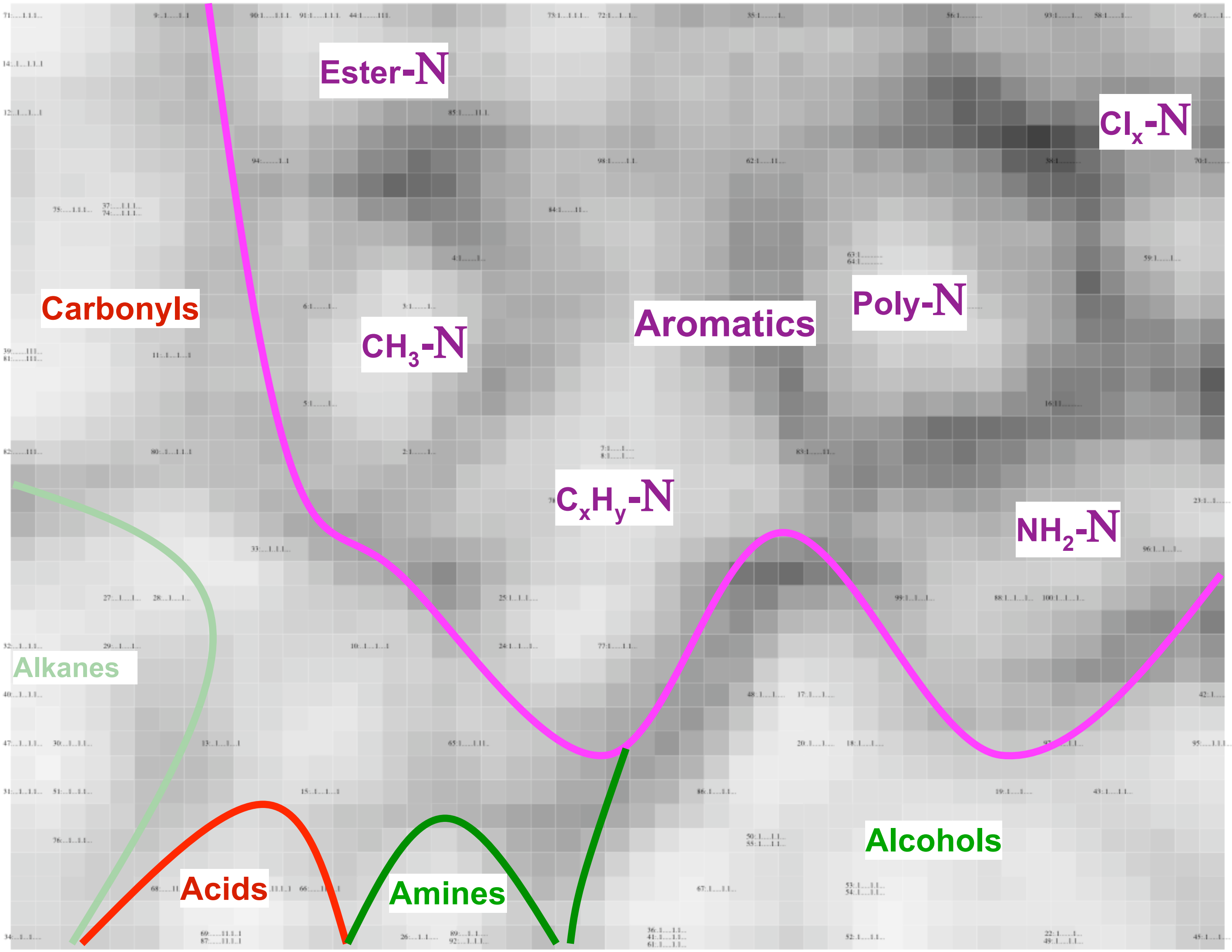


NIR SOM


71:.....1.1.1...
9:..1.....1..1
90:1.....1.1.1.
91:1.....1.1.1.
44:1......111.
73:1...1.1.1...
72:1....1...1...
35:1.......1..
56:1..........
93:1......1...
58:1......1...
60:1......1...
14:..1....1.1.1
12:..1....1....1
85:1......11.1.
94:.........1..1
98:1........1.1.
62:1......11....
58:1..........
70:1..........
75:.....1.1.1...
37:.....1.1.1...
74:.....1.1.1...
84:1......11...
4:1.......1...
63:1..........
64:1..........
59:1......1...
6:1.......1...
3:1.......1...
1:1.......1...
57:1..........
39:.......111...
81:.......111...
11:..1....1....1
5:1.......1...
16:11..........
82:.......111...
80:..1....1.1..1
2:1.......1...
7:1......1.....
8:1......1.....
83:1......11...
78:1......1.1...
23:1...1........
33:.....1..1.1...
96:1...1....1...
27:...1.....1...
28:...1.....1...
25:1...1..1.....
99:1...1....1...
88:1...1....1...
100:1...1....1...
32:...1...1.1...
29:...1.....1...
10:..1....1....1
24:1...1....1...
77:1......1.1...
40:...1...1.1...
48:..1.....1.....
17:..1.....1.....
42:..1.....
47:...1...1.1...
30:...1...1.1...
13:..1....1....1
65:1......1.11..
20:..1.....1.....
18:..1.....1.....
97:..1.....1.1...
95:......1.1.1...
31:...1...1.1...
51:...1...1.1...
15:..1....1....1
86:..1......1.1...
19:..1.....1.....
43:..1.....1.1...
76:...1...1.1...
50:..1......1.1...
55:..1......1.1...
21:..1.....1.....
46:..1.....1.1...
68:......11.1..1
79:......11.1..1
66:......11.1..1
67:..1......1.1...
53:..1......1.1...
54:..1......1.1...
34:...1....1.....
69:......11.1..1
87:......11.1..1
26:......1..1.....
89:.....1...1.....
92:.....1...1.1...
36:..1......1.1...
41:..1......1.1...
61:..1......1.1...
52:..1......1.1...
22:..1.......1...
49:..1......1.1...
45:..1......1.....

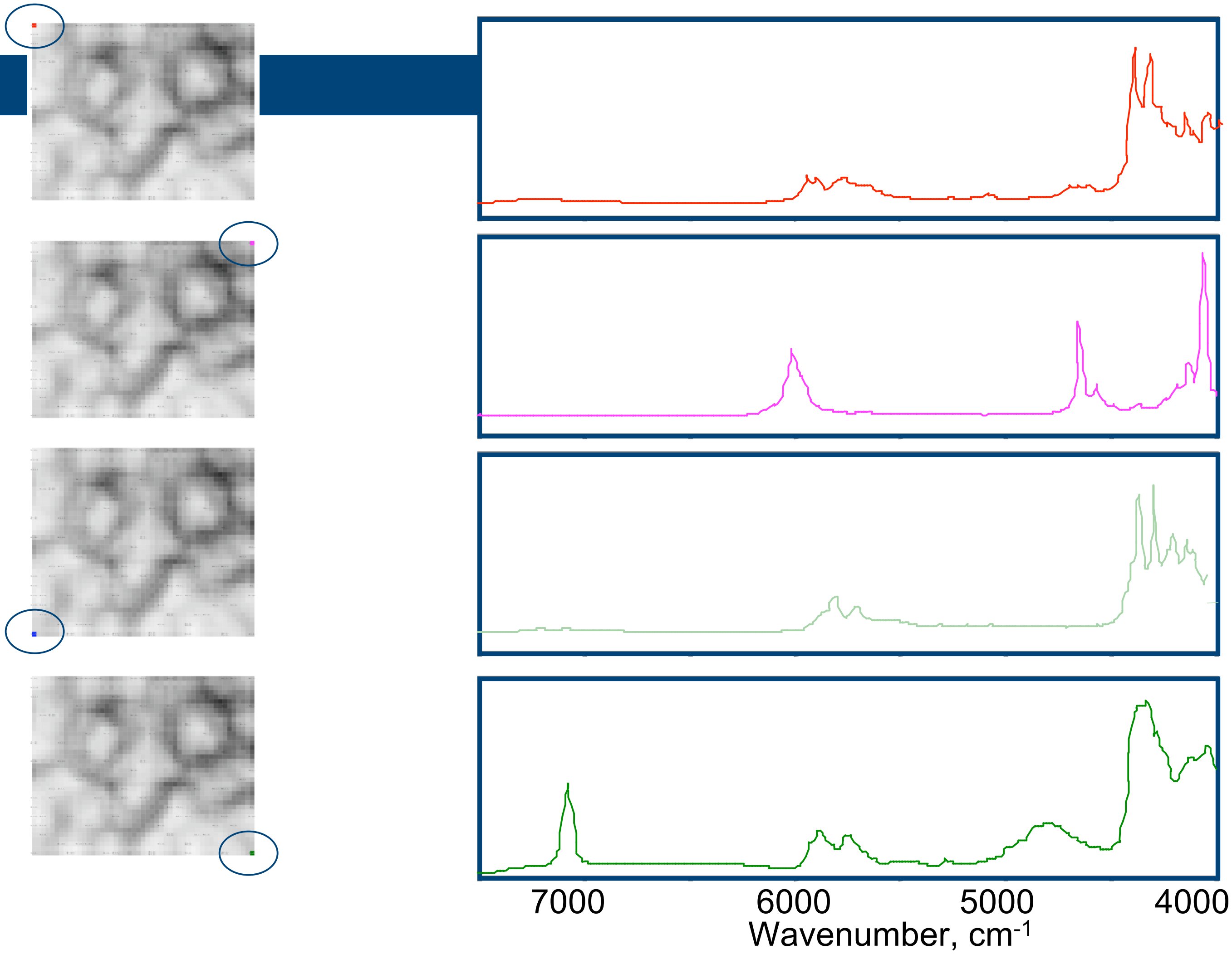
NIR SOM

Functional Groups



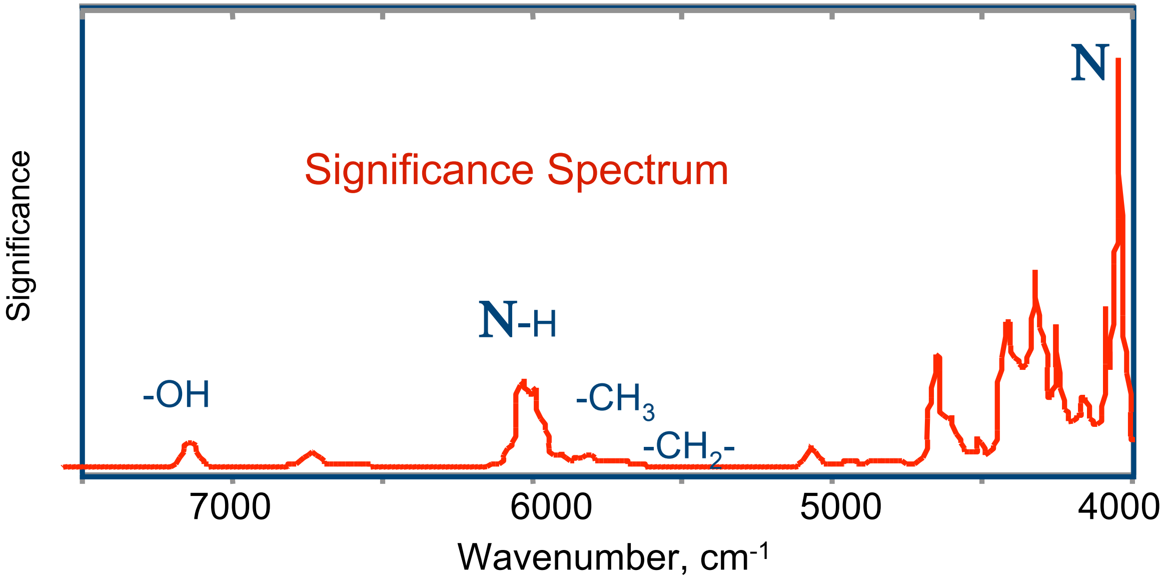
NIR

Centroid Spectra



NIR

Significance Spectrum



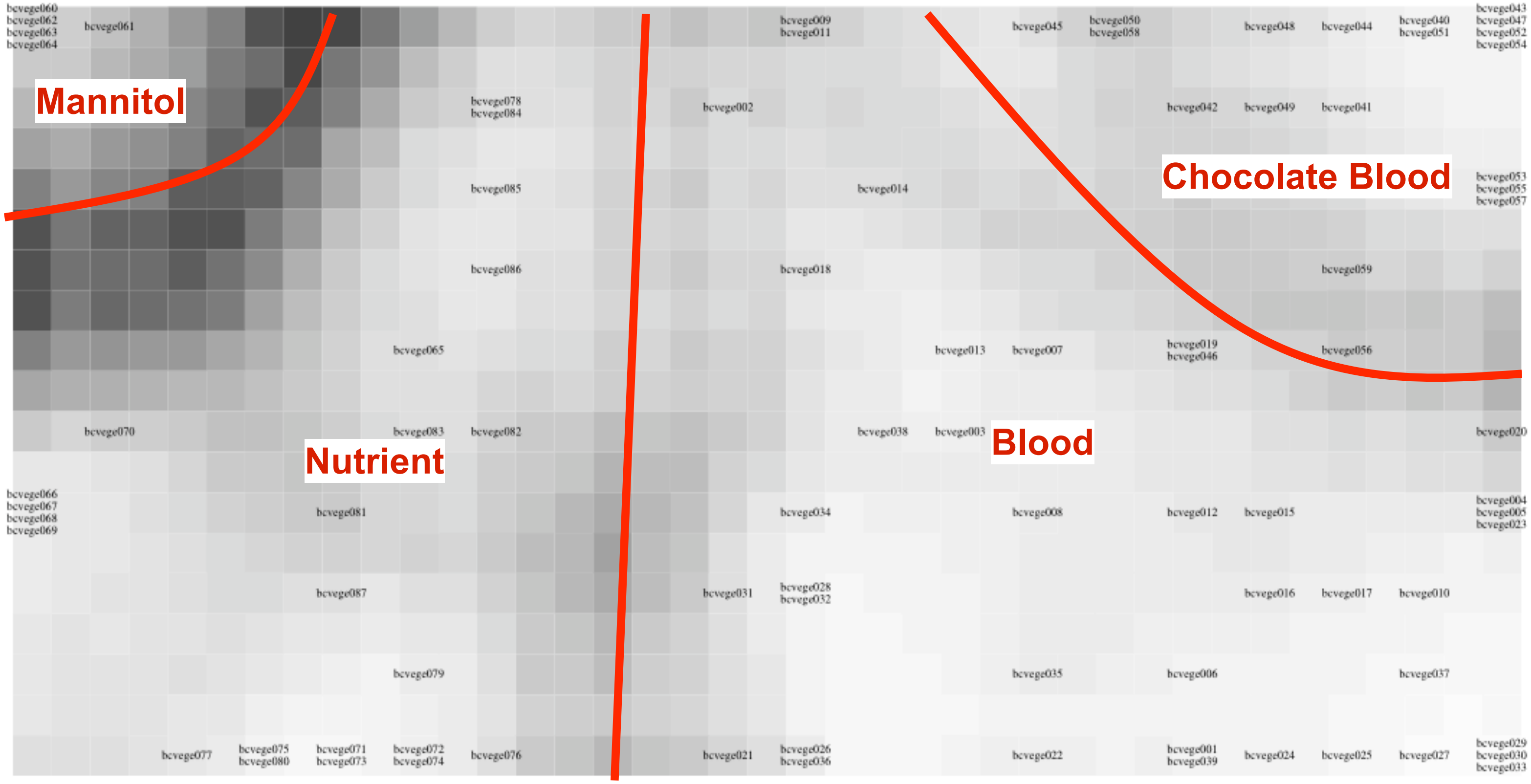
Applications of SOM

- We investigated bacteria using spectroscopy:
  - Can we detect spectroscopic differences between bacteria metabolizing different sugars?
  - Can we detect spectroscopic differences between the different stages of a bacterium’s existence?
  - Can we detect spectroscopic differences between Gram-Positive and Gram-Negative bacteria?

Bayesian Probability Approach to Feature Significance for Infrared Spectra of Bacteria, Lutz Hamel, Chris W. Brown, Applied Spectroscopy, Volume 66, Number 1, 2012.

SOM

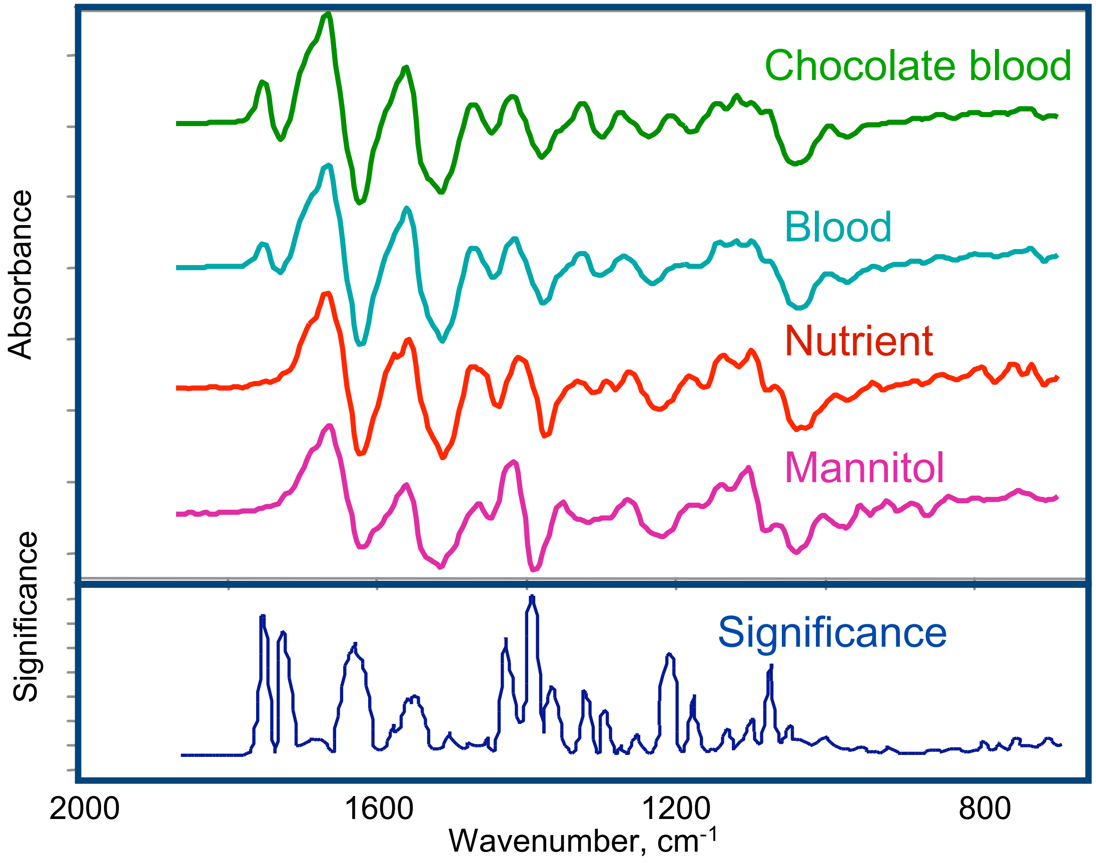
Bacterium *b-cereus* on different agars



"You are what you eat!"

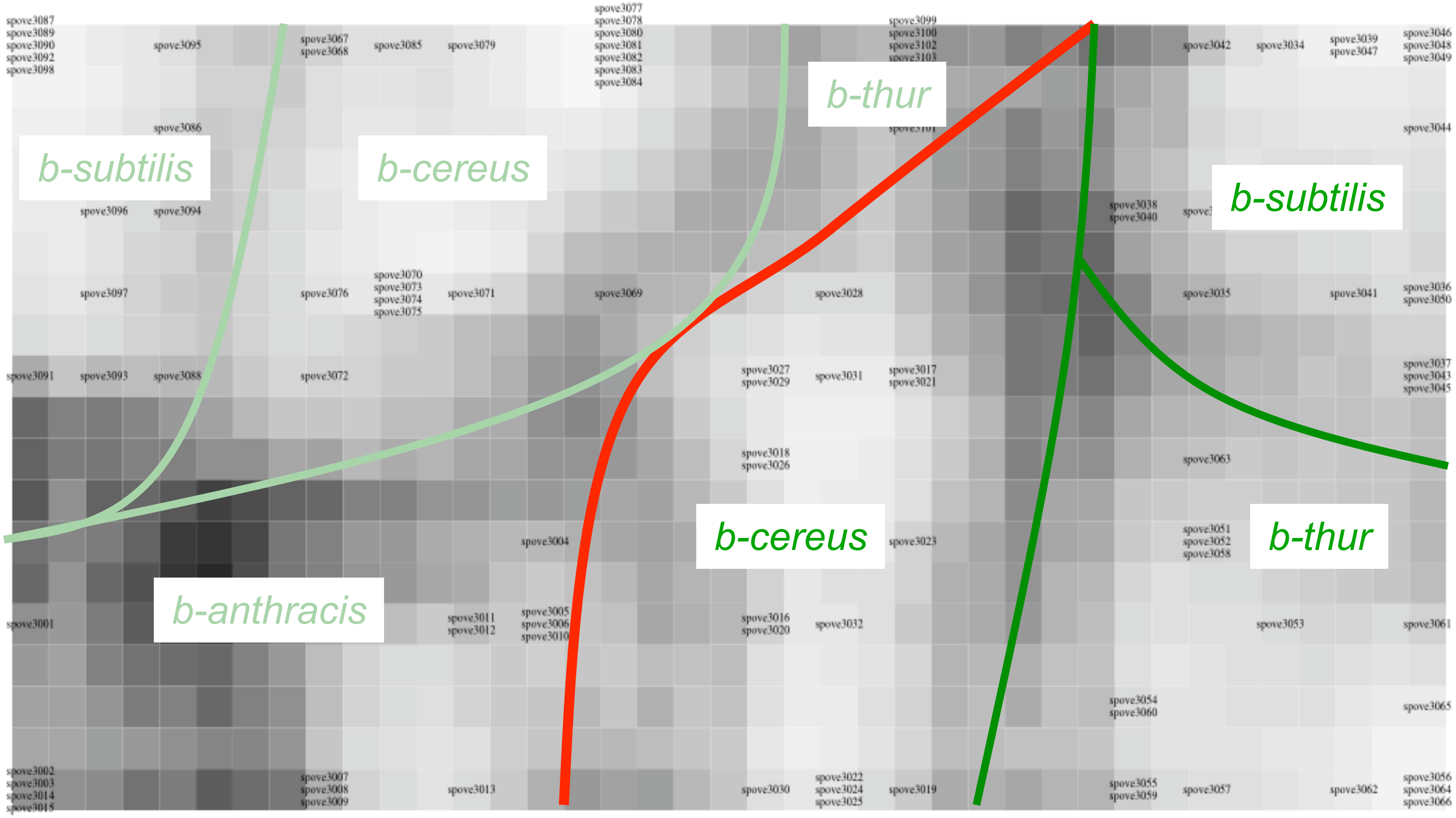
Significance Spectrum

b-cereus on different agars



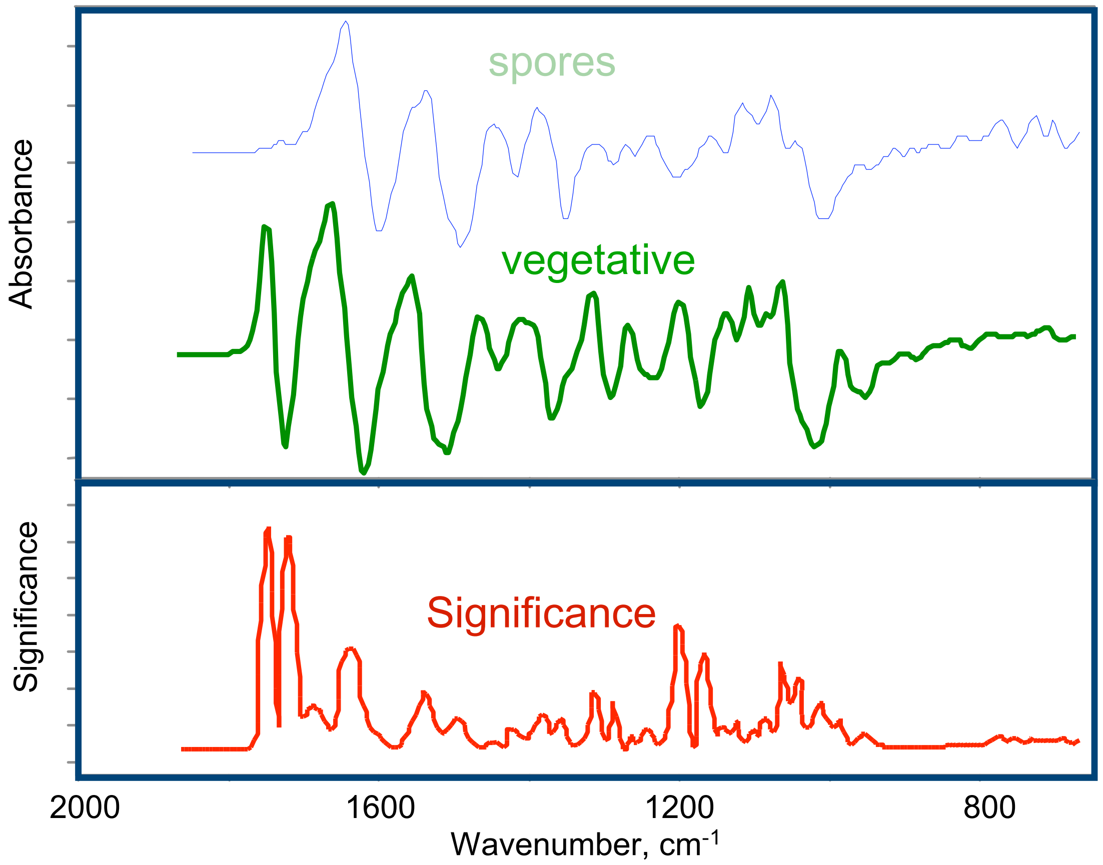
SOM

Bacteria Spectra

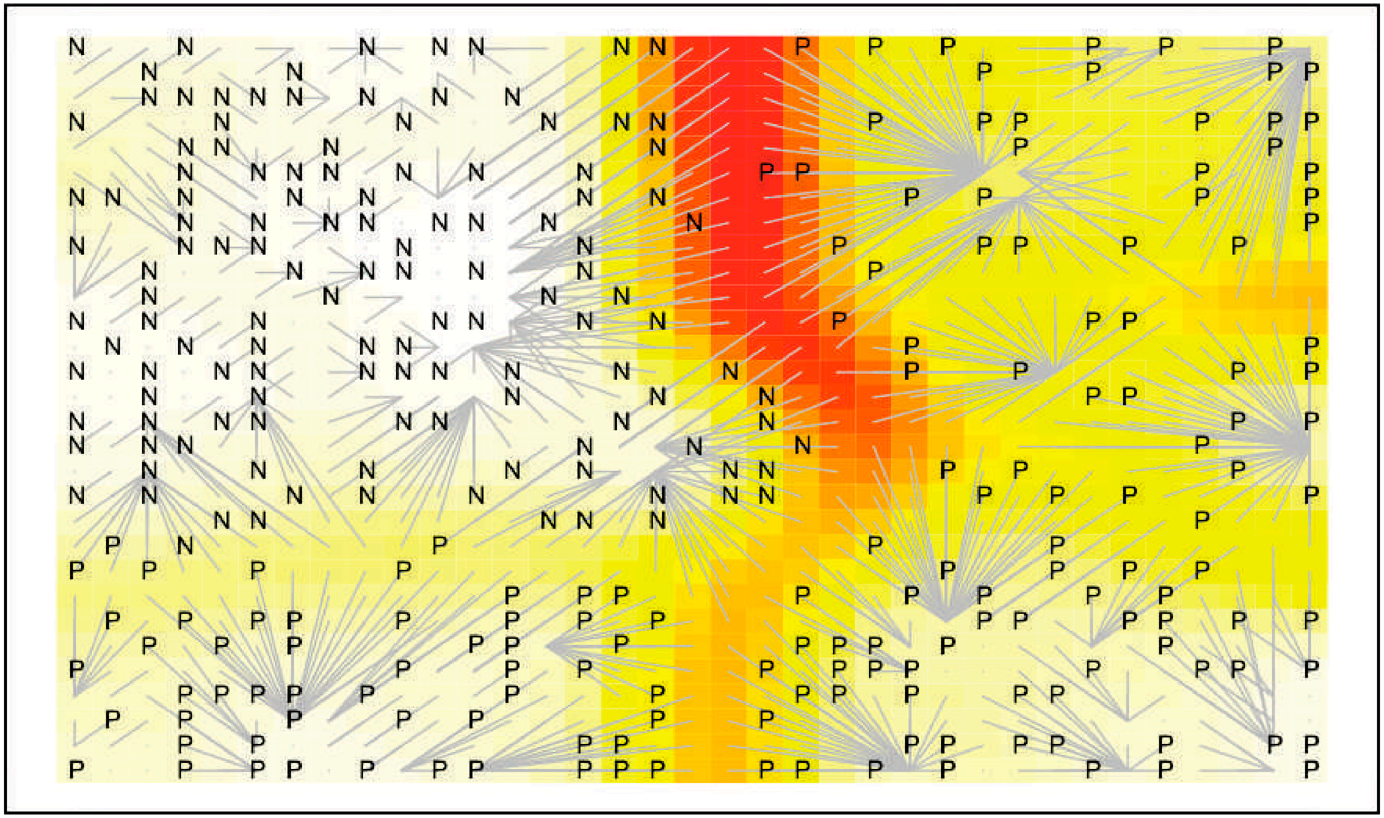


🡐spores / vegetative🡒

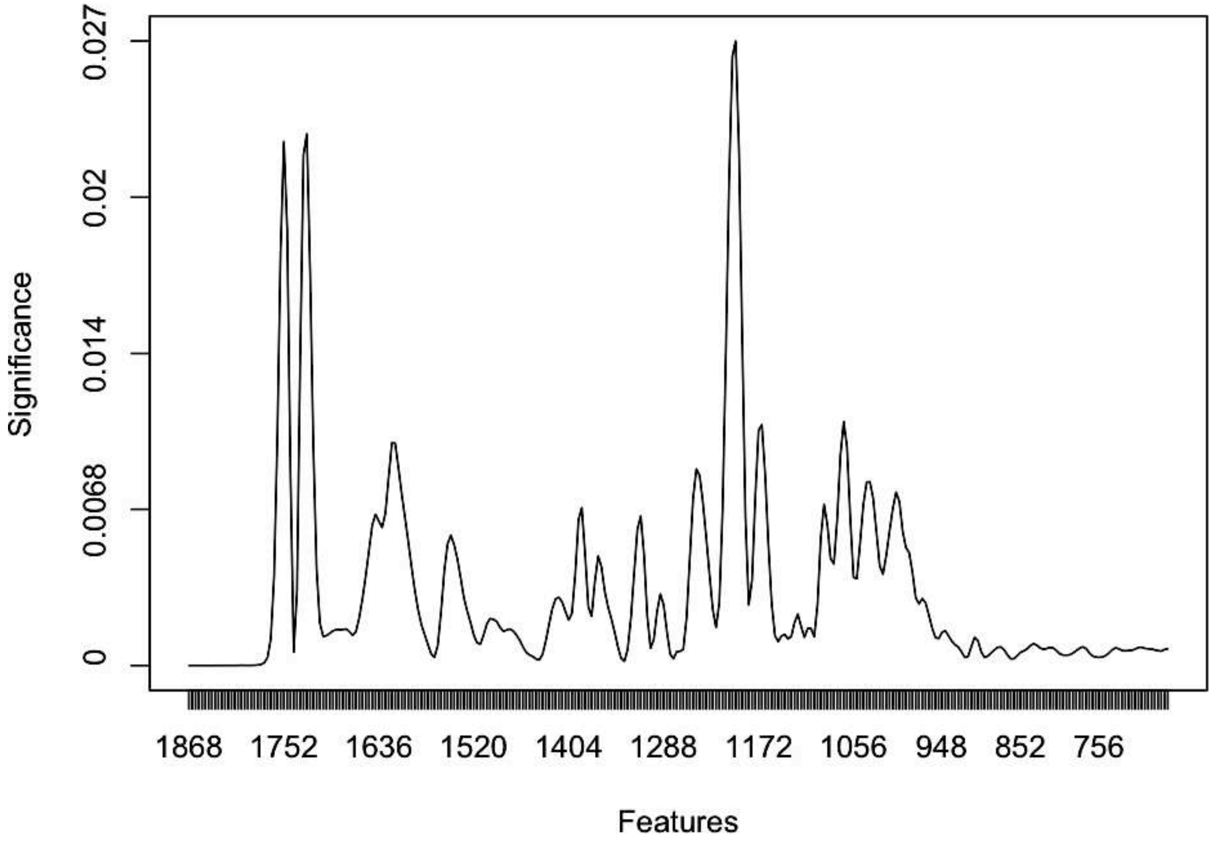
Significance Spectrum vs *b-subtilis* 1^{st} Derivative Spectra



Gram-Pos. vs. Gram-Neg.



Significance Spectrum



Applications of SOM

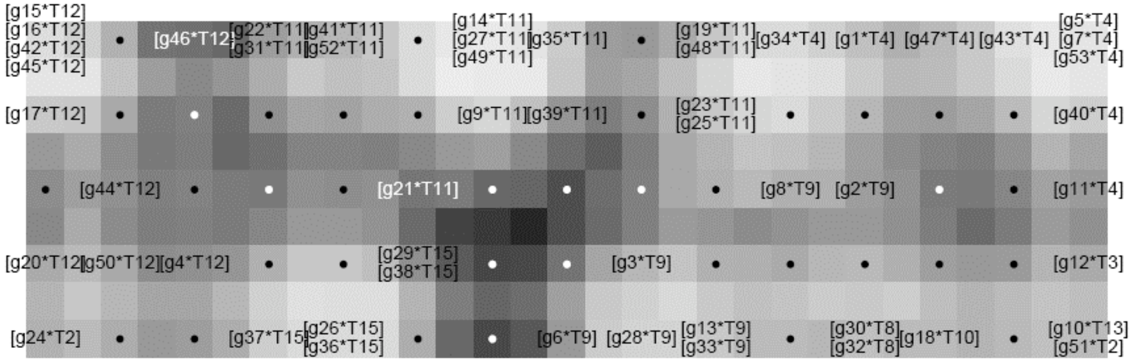
- Genome Clustering
  - Goal: trying to understand the phylogenetic relationship between different genomes.
  - Compute bootstrap support of individual genomes for different phylogentic tree topologies, then cluster based on the topology support.

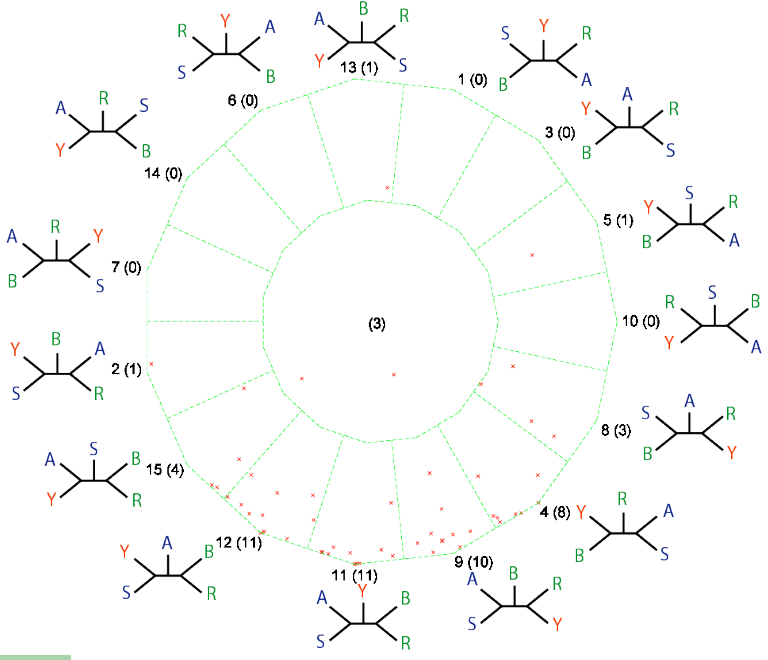
Unsupervised Learning in Detection of Gene Transfer, Lutz Hamel, Neha Nahar, Maria S. Poptsova, Olga Zhaxybayeva, and J. Peter Gogarten. Journal of Biomedicine and Biotechnology, vol. 2008, Article ID 472719, 7 pages, 2008. doi:10.1155/2008/472719

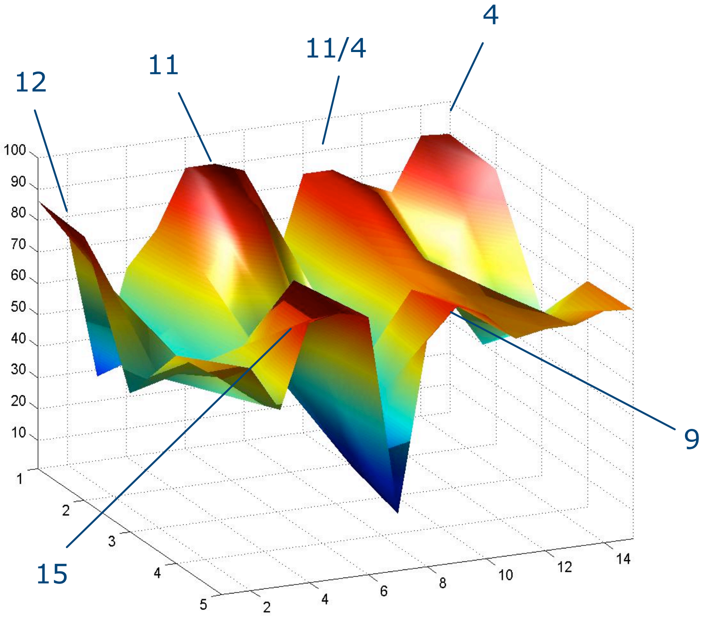
PentaPlot: A Software Tool for the Illustration of Genome Mosaicism, Lutz Hamel, Olga Zhaxybayeva, and J. Peter Gogarten. BMC Bioinformatics, 2005 6:139, http://www.biomedcentral.com/1471-2105/6/139

Visualization of the phylogenetic content of five genomes using dekapentagonal maps, Olga Zhaxybayeva, Lutz Hamel, Jason Raymond and J Peter Gogarten. Genome Biology, 2004 5:R20, http://genomebiology.com/2004/5/3/R20

Phylogenetic Visualization with SOMs







GPX

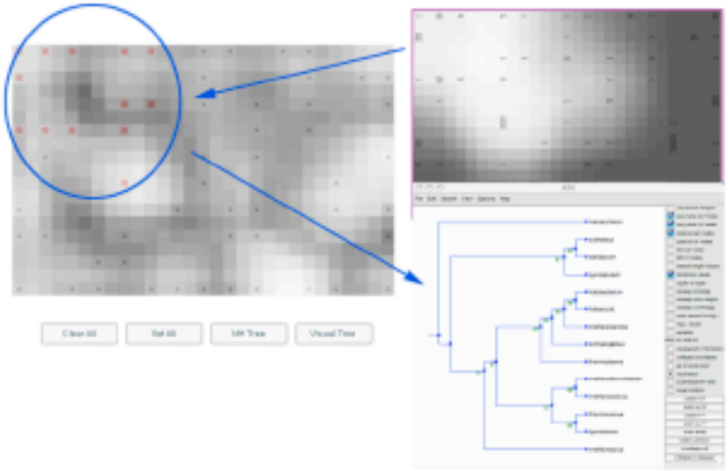


Figure 8. Tree reconstructed from the selected clusters (red dots on the left map) that fell into white areas on the bipartition superposition map (on the right).

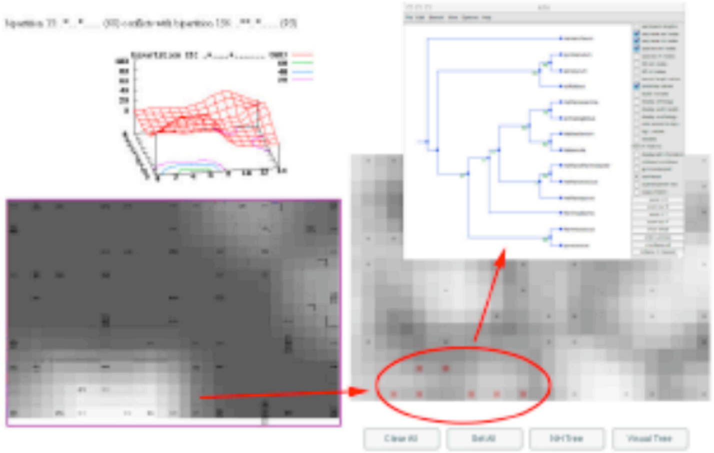


Figure 9: Analysis of the conflicting bipartition (see text for explanation).

GPX: A Tool for the Exploration and Visualization of Genome Evolution, Neha Nahar, Maria S. Poptsova, Lutz Hamel, and J. Peter Gogarten. Proceedings of the IEEE 7th International Symposium on Bioinformatics & Bioengineering (BIBE07), Oct 14th-17th 2007, Boston, pp1338 - 1342, IEEE Press, ISBN 1-4244-1509-8.

Applications of SOM

- Clustering Proteins based on the architecture of their activation loops.
  - Align the proteins under investigation.
  - Extract the functional centers.
  - Turn 3D representation into 1D feature vectors.
  - Cluster based on the feature vectors.

Toward Protein Structure Analysis with Self-Organizing Maps, Lutz Hamel, Gongqin Sun, and Jing Zhang, IEEE 2005 Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp506-513, La Jolla, CA, IEEE, 2005, ISBN 0-7803-9387-2.

Processing of Protein Structures

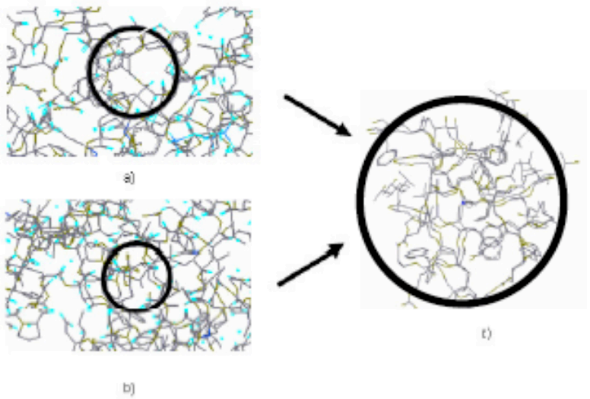


Fig.3: Alignment of active site structures in proteins; a) active site of cAMP-dependent protein-kinase (1ATP), b) active site of glycogen synthase kinase-3β (1GNG), c) the extracted and locally aligned structures surrounding the active sites are shown.

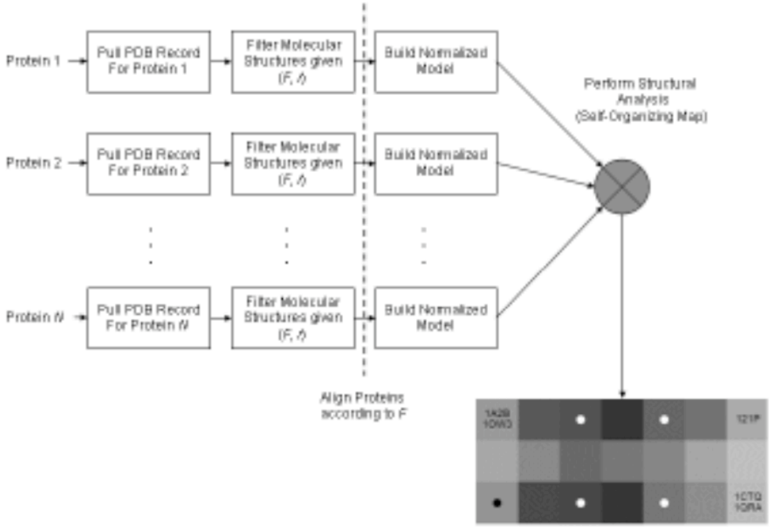

Protein 1
Pull PDB Record For Protein 1
Filter Molecular Structures given (F, I)
Build Normalized Model
Protein 2
Pull PDB Record For Protein 2
Filter Molecular Structures given (F, I)
Build Normalized Model
Protein N
Pull PDB Record For Protein N
Filter Molecular Structures given (F, I)
Build Normalized Model
Align Proteins according to F
Perform Structural Analysis (Self-Organizing Map)
1A2B 1DW3
121P
1CTQ 1QRA

Fig. 5: Summary of Workflow.

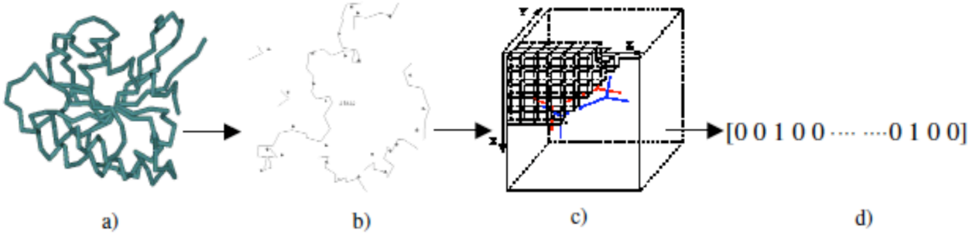
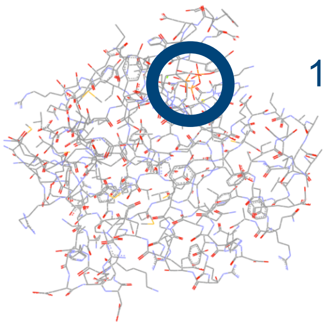


Fig. 4: Protein feature vector construction: a) the 3D structure of a protein without side-chains; b) the normalized structure of the functional center of the protein, the crosses pinpoint the normalized locations of the α-carbons representing our normalized model; c) encoding the normalized model by using cubic subunits; if there is a normalized α-carbon atom in a cubic subunit then the subunit is assigned a 1, otherwise it is assigned a 0; d) the 3D structure of the cubic subunits is unfolded giving rise to a one dimensional feature vector describing the structure of the protein; each position in the feature vector describes the state of a single subunit of the original 3D structure.

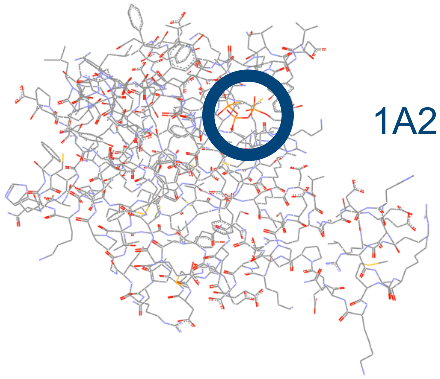
Structural Classification of GTPases

Can we structurally distinguish between the Ras and Rho subfamilies?

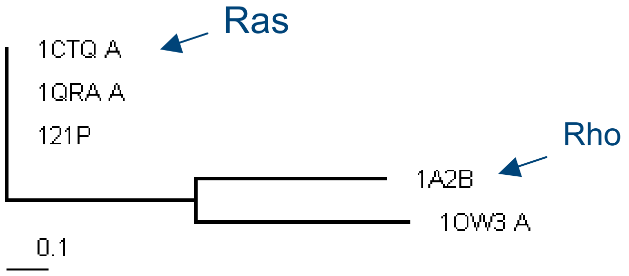
- Ras: 121P, 1CTQ, and 1QRA
- Rho: 1A2B and 1OW3
- F = p-loop, r = 10Å

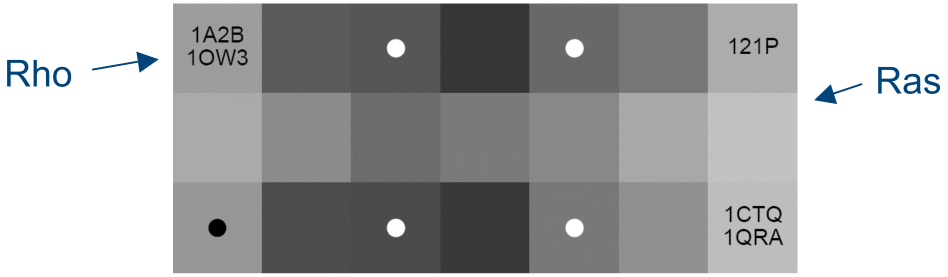


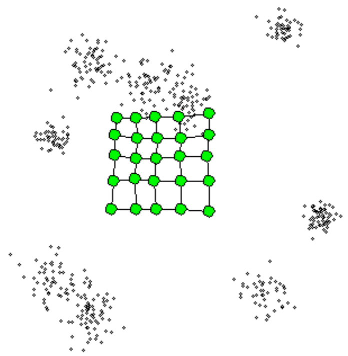
1CTQ



1A2B







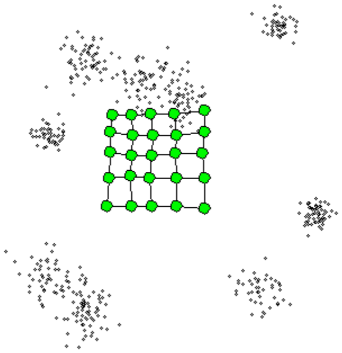
Model Fitting

- Standard approach is *minimizing the quantization error:*

$$E_{\mathbf{Y}}(t) = \sum_{c \in \mathbf{Y}} \sum_{\mathbf{x} \in \mathbf{X}_c(t)} \| \mathbf{w}_c(t) - \mathbf{x} \|^2$$

- However, there exists no statistical criterion that tells us when the quantization error is good enough!
- In the limit (enough neurons, enough time) the quantization error can always be reduced to ≈0

  ⇒ Overfitting!
- Therefore not very useful as a “Goodness of Fit” criterion.



Convergence

- Our approach is different: we treat the training data and the set of neurons as two individual populations.
- We say that a *maps has converged* if both populations appear to have been dawn from the same distribution.
- This is easily testable with appropriate 2-sample tests.

A Population Based Convergence Criterion for Self-Organizing Maps, Benjamin Ott and Lutz Hamel, submitted.

2-Sample Tests

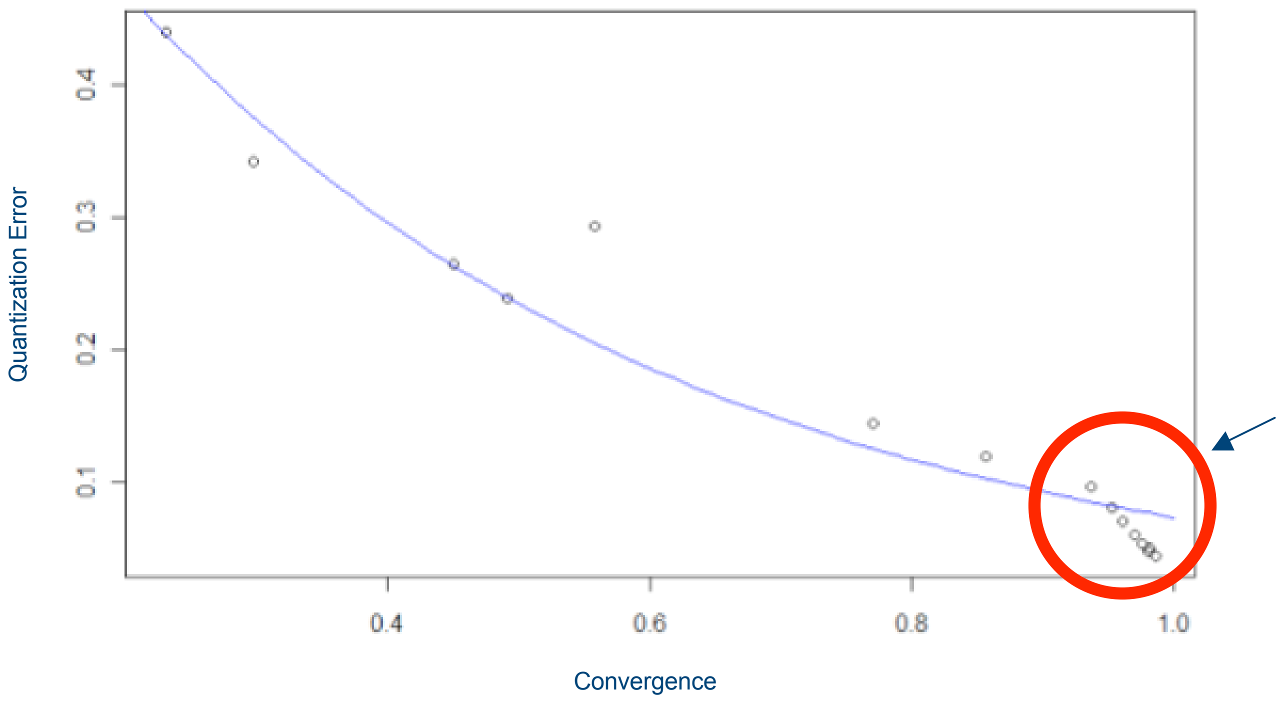
- Variance:

$$\frac{s_1^2}{{s_2}^2} \cdot \frac{1}{f_{\frac{\alpha}{2},n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{{s_2}^2} \cdot f_{\frac{\alpha}{2},n_1-1,n_2-1};$$

- Mean:

$$\begin{aligned} \mu_1 - \mu_2 &> (\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \\ \mu_1 - \mu_2 &< (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \end{aligned}$$

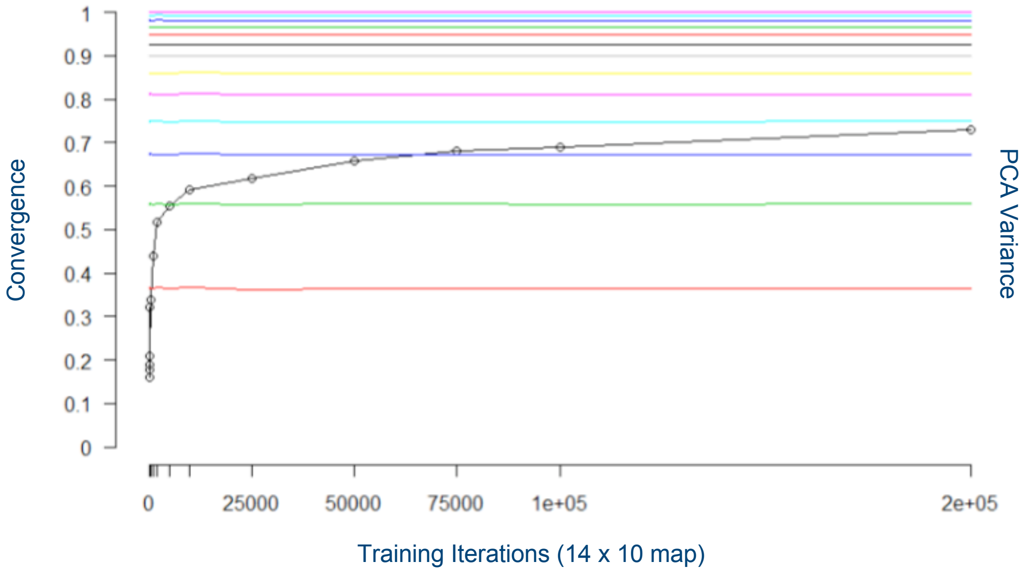
Error vs. Convergence



Convergence with non-zero quantization error - no overfitting!

Observations

- SOMs, in most applications, are severely undertrained and therefore do not represent the underlying structure reliably!



UCI Machine Learning Repository:
Wine Data Set

Visualizations

- We have developed two new SOM visualizations that assist in interpreting the structure of the data:
  - Starburst
  - Cartogram

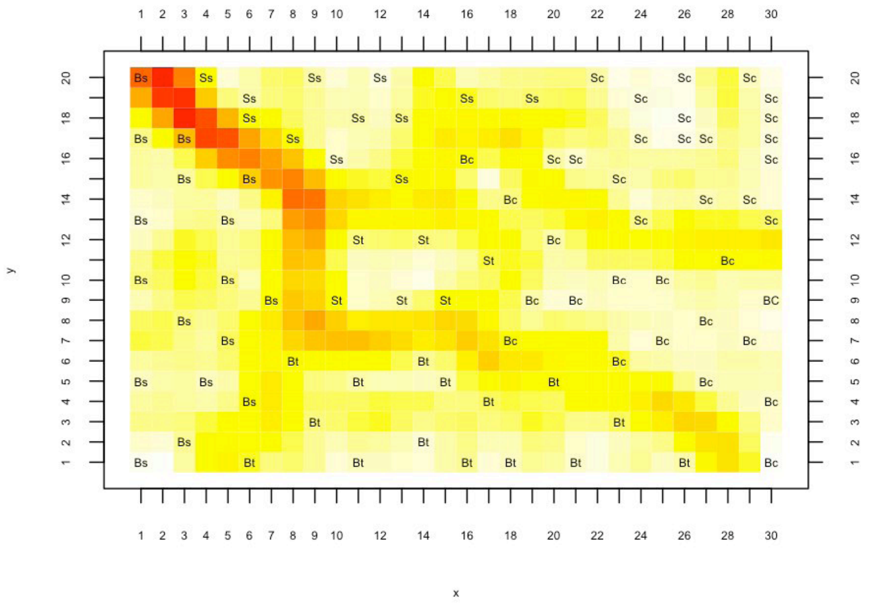
Improved Interpretability of the Unified Distance Matrix with Connected Components, Lutz Hamel and Chris W. Brown. Proceeding of the 7th International Conference on Data Mining, July 18-21, 2011, Las Vegas Nevada, USA, ISBN: 1-60132-168-6, pp338-343, CSREA Press, 2011.

Cartogram Data Projection for Self-Organizing Maps, David Brown and Lutz Hamel, submitted.

Starburst

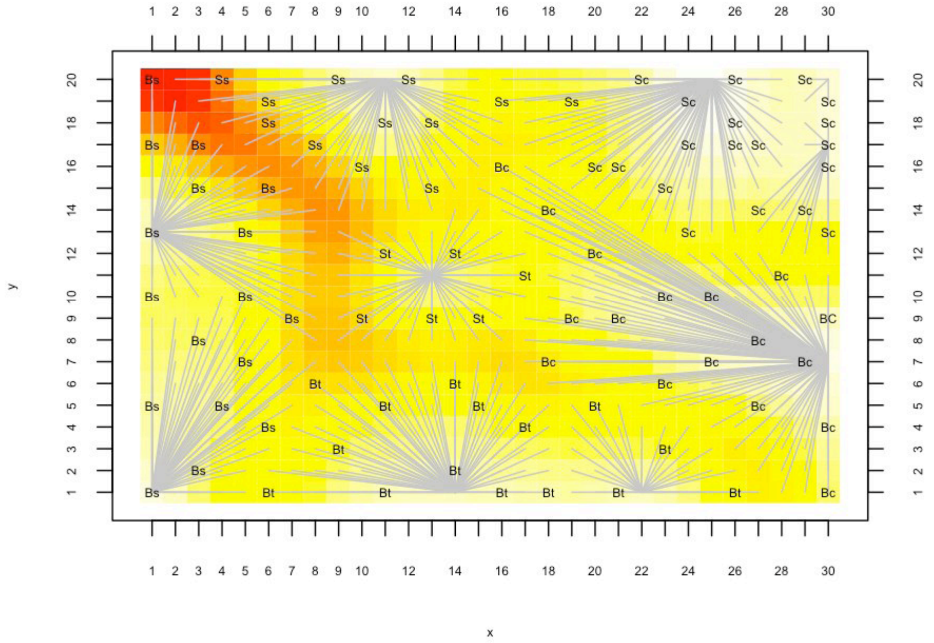
- Assists in identifying clusters on the SOM Unified-Distance map (Umat)
- Starbursts are constructed by
  - First following the steepest gradient on the Umat to the center of the cluster (the center of the cluster has a gradient of 0)
  - Then connecting all points who gradient vector point to a particular center to that center.

Starburst



Kind of difficult to see where the clusters are.

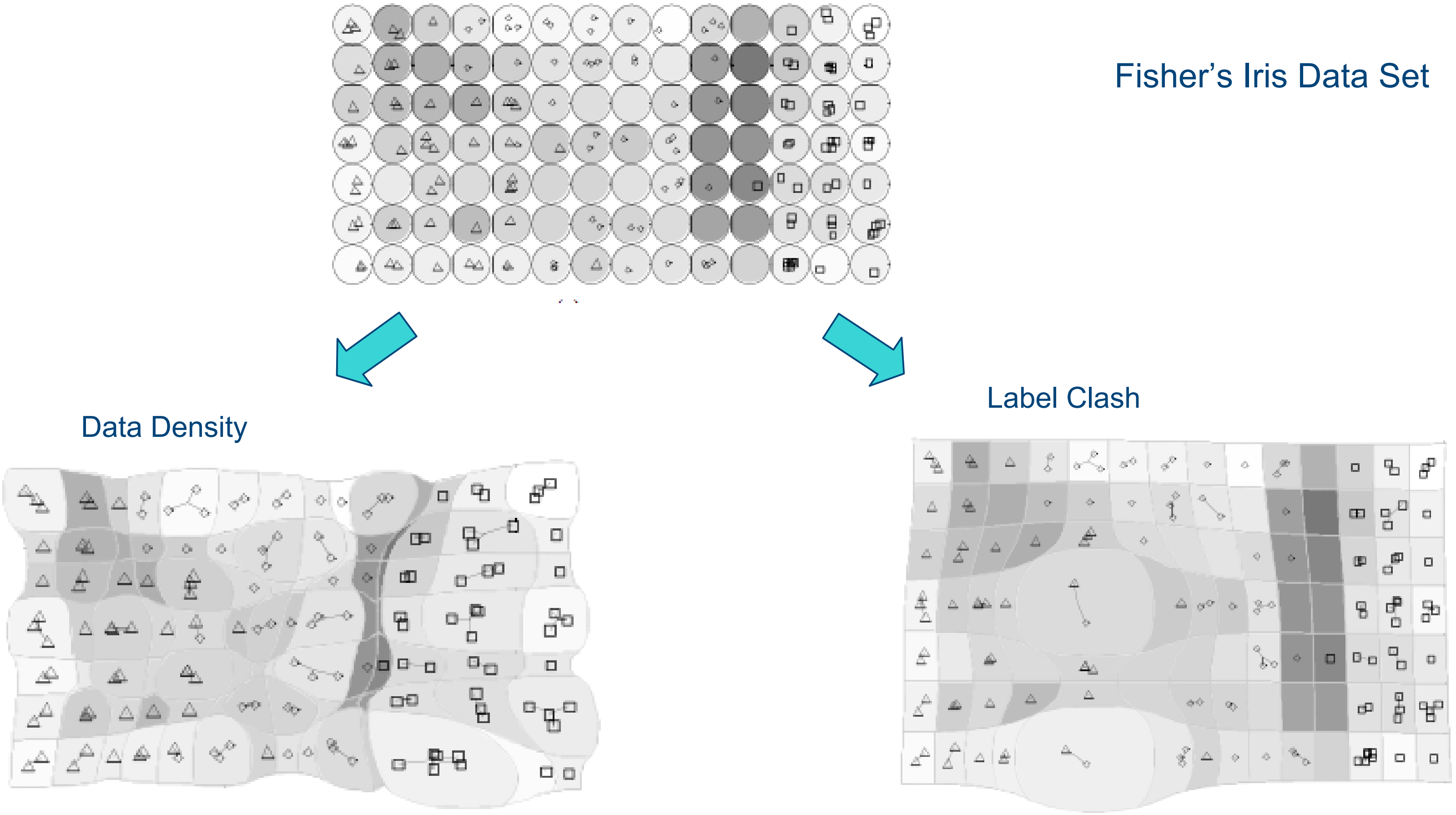
Now much easier



Cartogram Data Projection

- Technique borrowed from geographic map making
- Distort the SOM map to highlight features of interest:
  - Data density
  - Label clashes (if labels are available)
  - Risk factors, etc.
- Map training data back onto map in a meaningful fashion, I.e., it conveys the data distribution around the neuron in data space.

Cartogram



Fisher’s Iris Data Set

Data Density

Label Clash

Cartogram

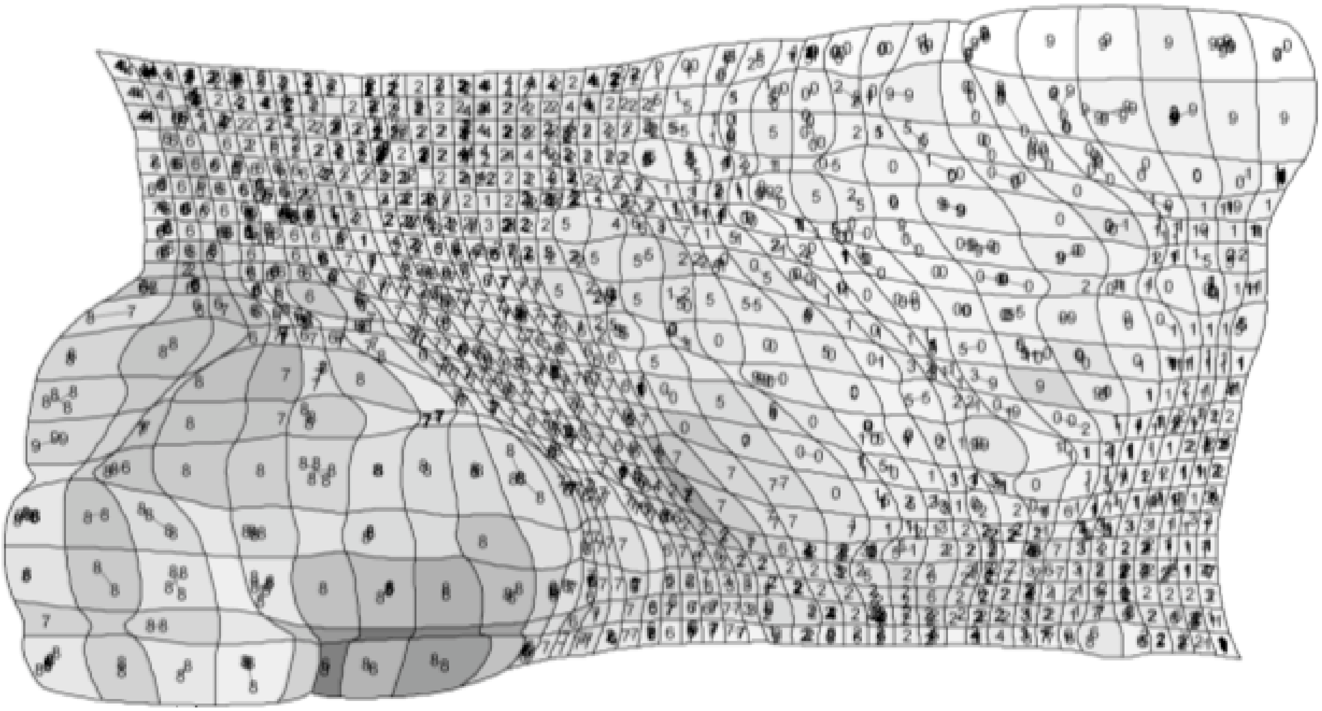


Figure 7. SOM of the cardiotocography data set using the expert assessment of risk as cell size.

D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de-Sa, and L. Pereira-Leite, “SisPorto 2.0: a program for automated analysis of cardiotocograms.,” *J Matern Fetal Med*, vol. 9, no. 5, pp. 311-8, 2000.

Conclusions

- SOMs are powerful tools for data visualization and discovery
- Our new convergence criterion puts SOM training on a solid statistical foundation
- Our new visualization techniques help interpreting the map generated by the SOM algorithm

Thank You!

- Questions?

www.cs.uri.edu/~hamel