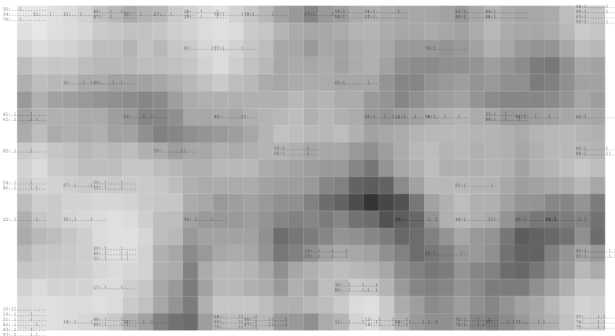




Practical Tools for Self-Organizing Maps



Lutz Hamel

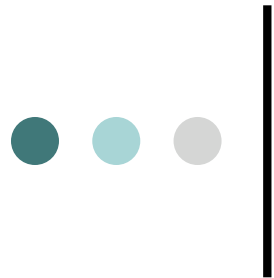
Dept. of Computer Science &
Statistics

URI

Chris Brown

Dept. of Chemistry

URI



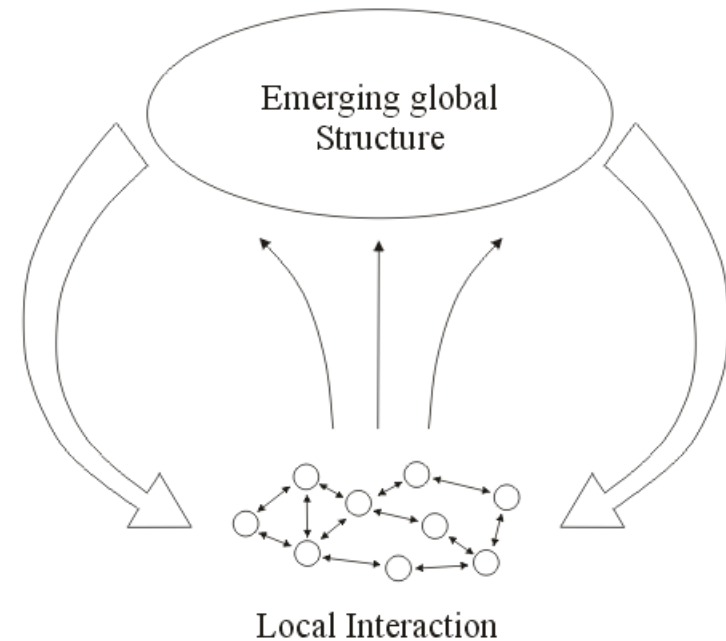
Self-Organizing Maps (SOMs)

- A neural network approach to unsupervised machine learning.
- Appealing visual presentation of learned results as a 2D map.

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

● ● ● | Feature Vector Construction

In order to use SOMs we need to describe our objects

- Feature Vectors



small	medium	big	Two legs	Four legs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
1	0	0	1	0	0	0	0	1	0	0	0	1



small	medium	big	Two legs	Four legs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
0	0	1	0	1	1	1	0	0	0	1	0	0

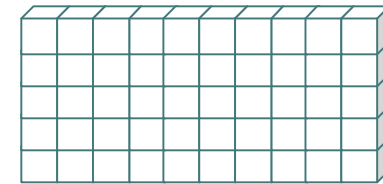


Training a SOM

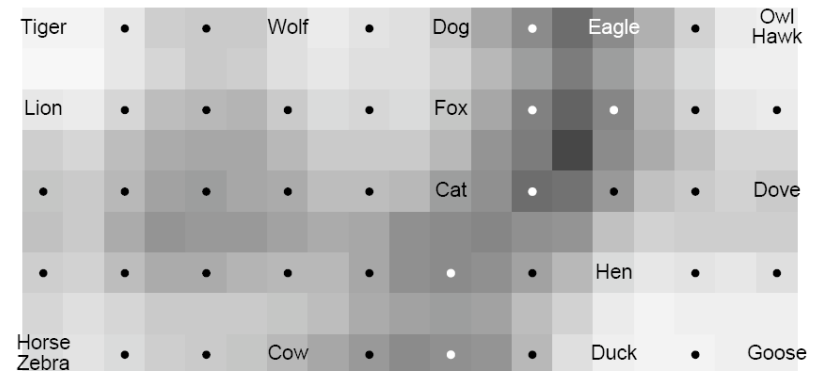
	small	medium	big	Two legs	Four legs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
Dove	1	0	0	1	0	0	0	0	1	0	0	1	0
Hen	1	0	0	1	0	0	0	0	1	0	0	0	0
Duck	1	0	0	1	0	0	0	0	1	0	0	0	1
Goose	1	0	0	1	0	0	0	0	1	0	0	1	1
Owe	1	0	0	1	0	0	0	0	1	1	0	1	0
Hawk	1	0	0	1	0	0	0	0	1	1	0	1	0
Eagle	0	1	0	1	0	0	0	0	1	1	0	1	0
Fox	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	1	0	0	0	0	0

Table of Feature Vectors

“Grid of Neurons”



Visualization



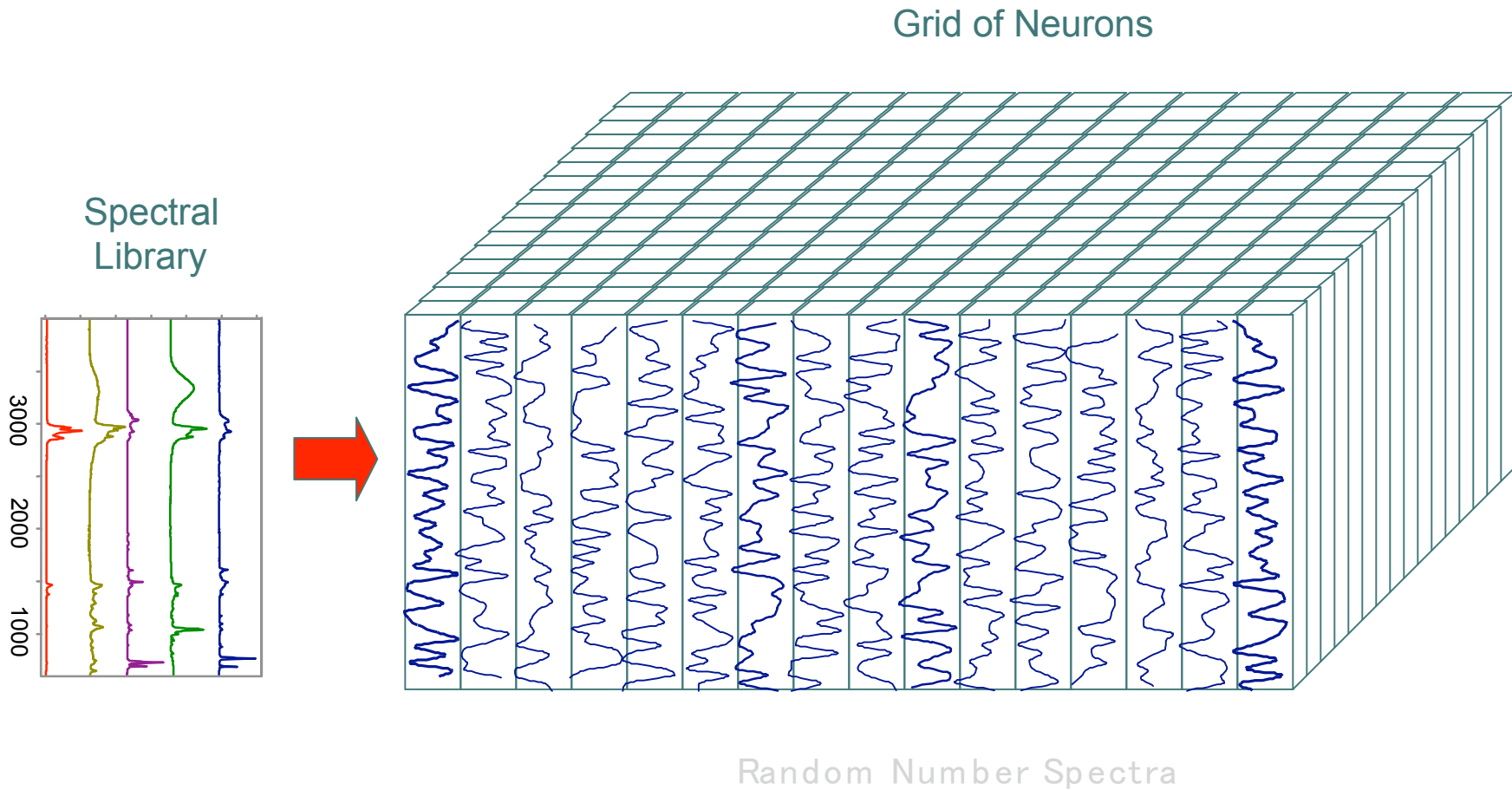


Applications of SOM

- 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”.
 - We can use SOMs to investigate similarity.



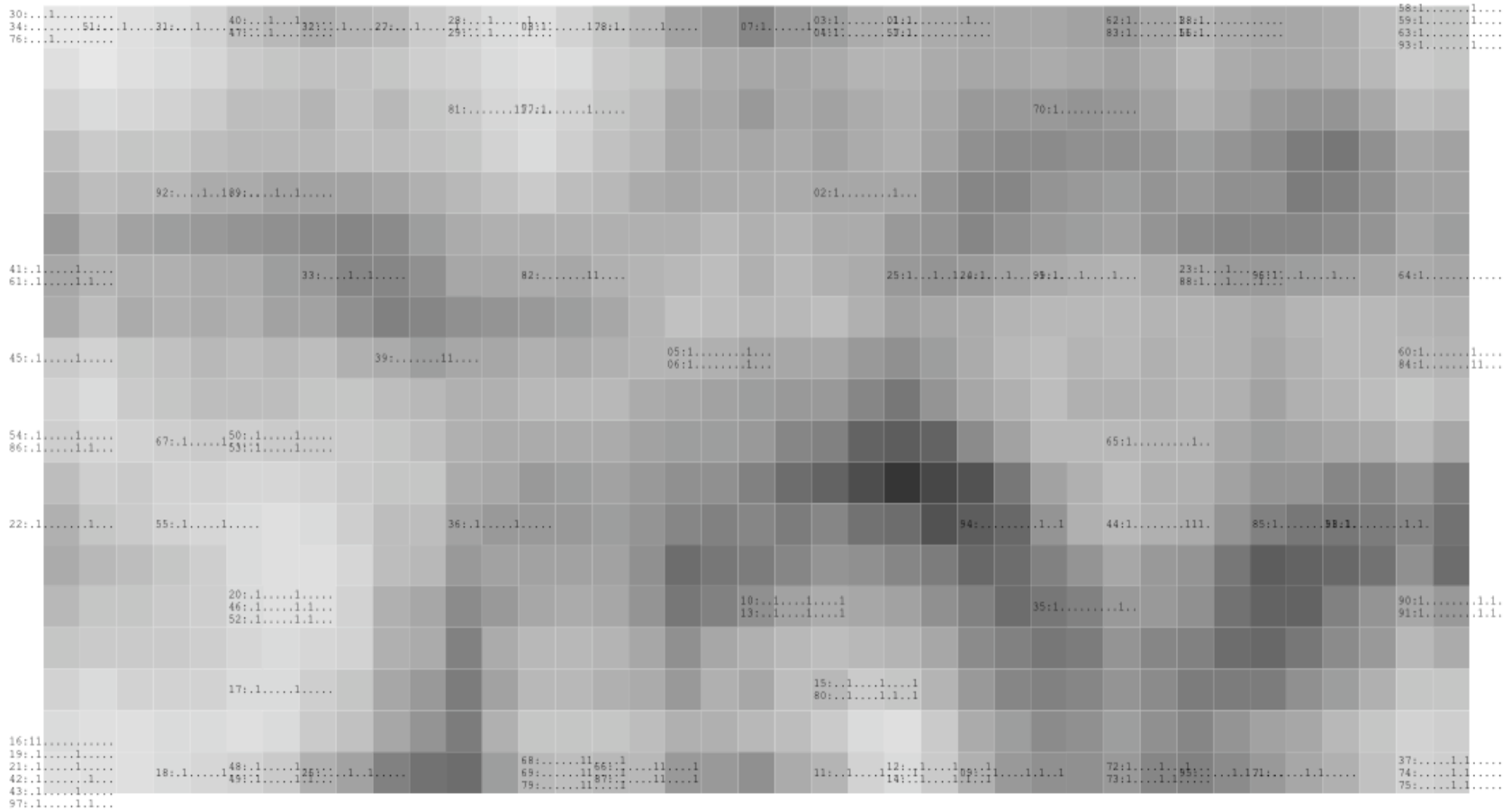
Training SOM with Spectra





Self-Organizing-Map

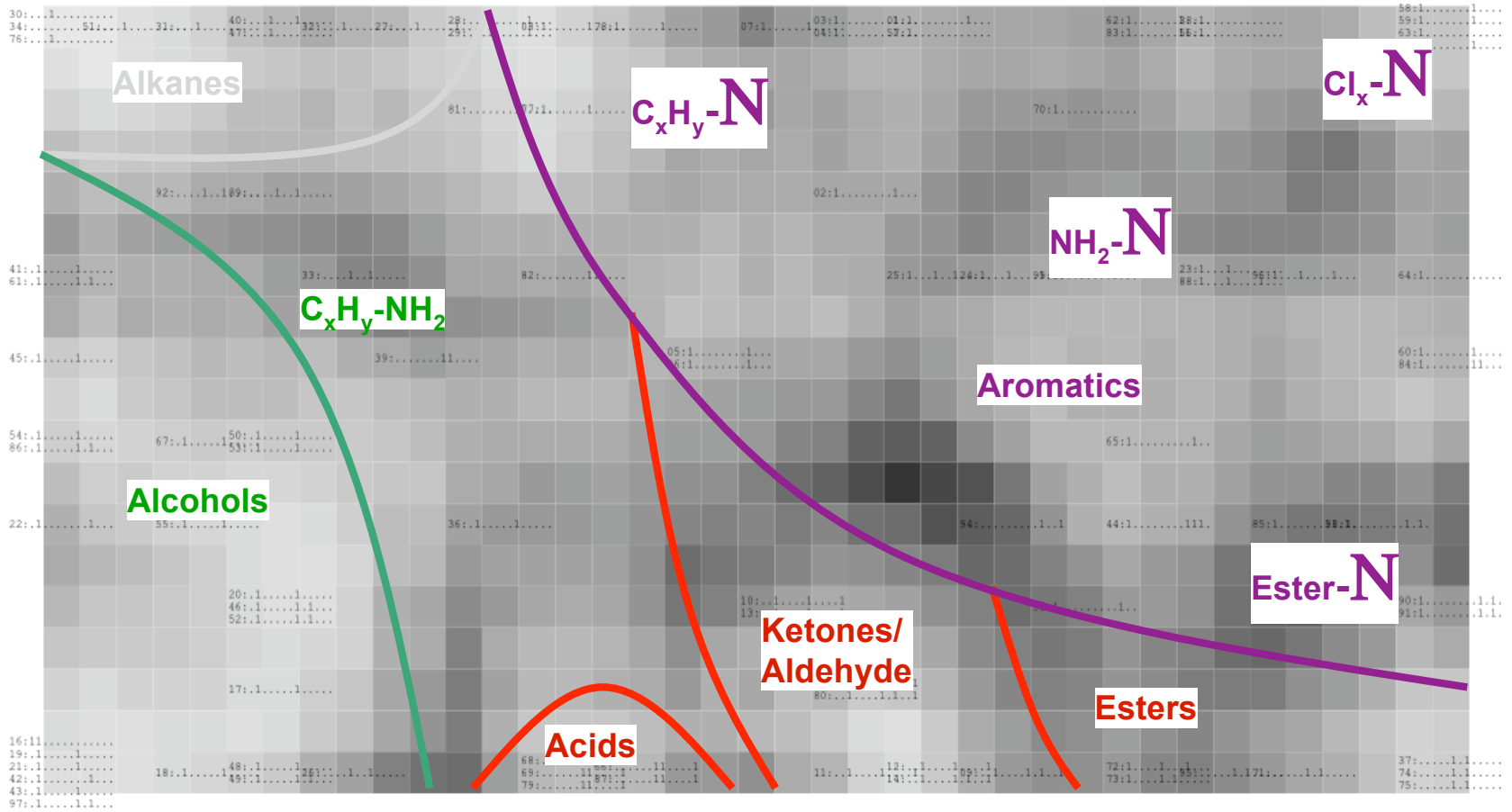
MIR Spectra





MIR SOM

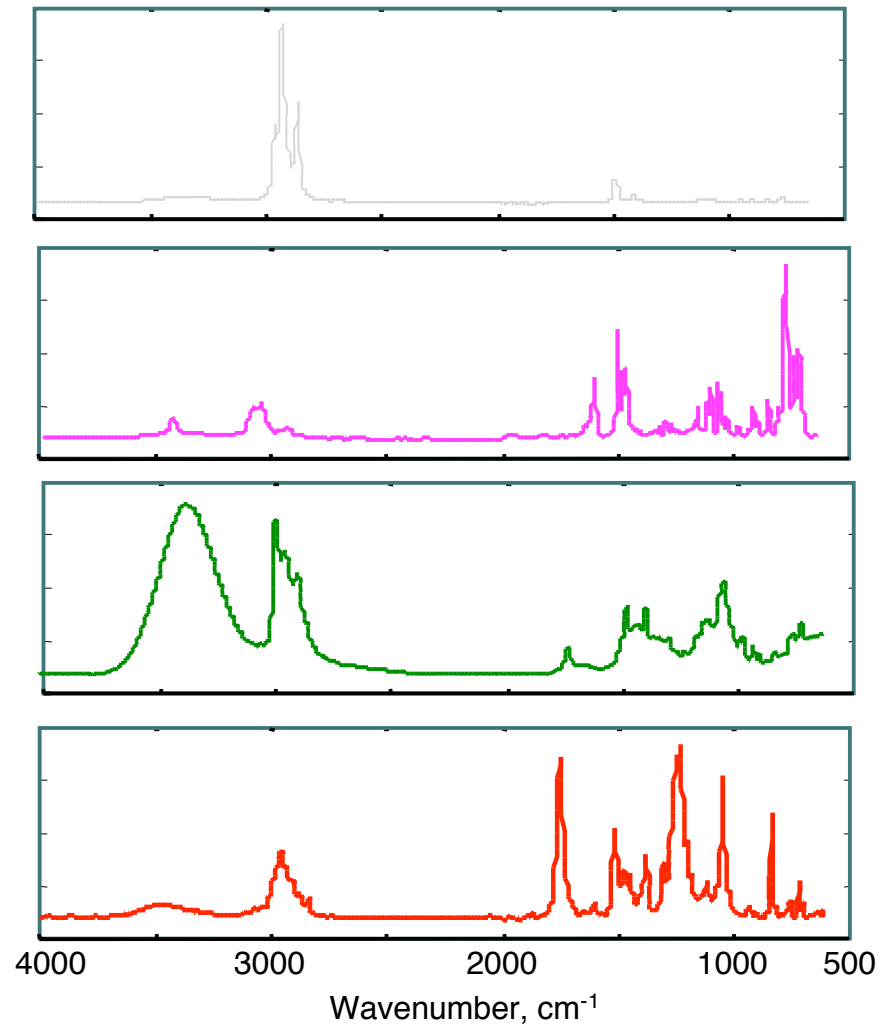
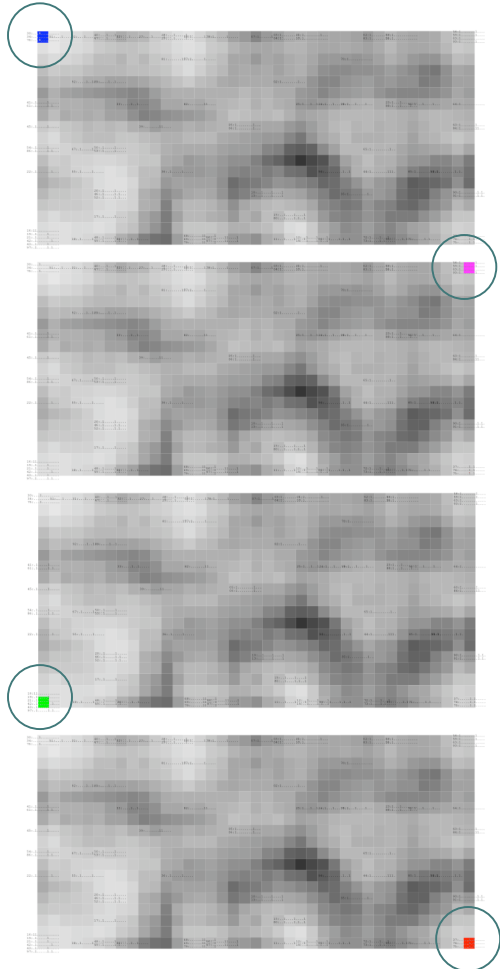
Functional Groups





MIR

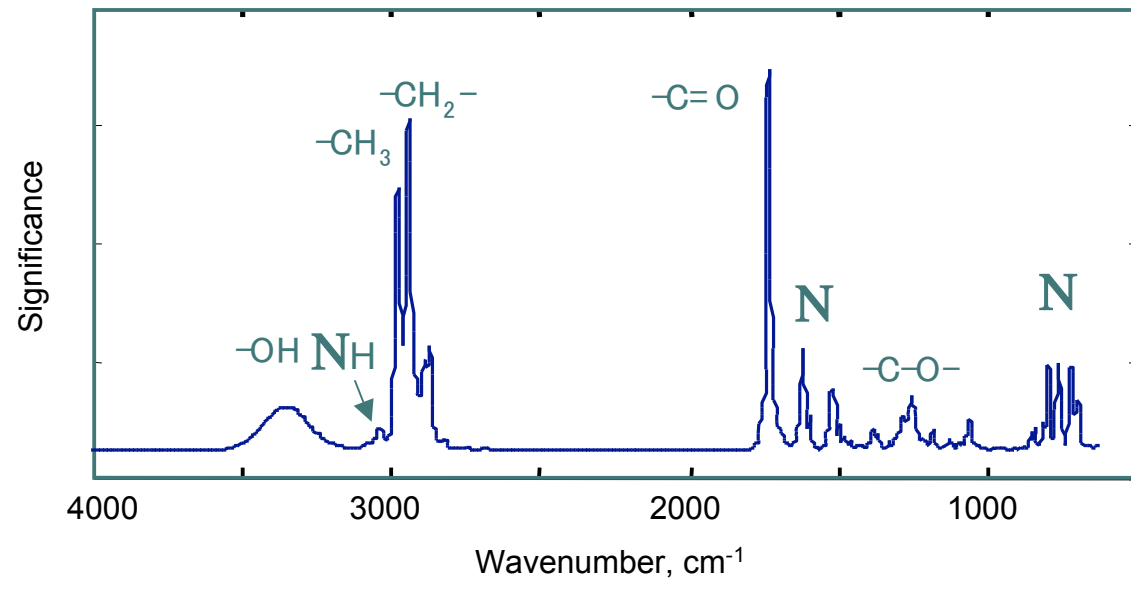
Centroid Spectra





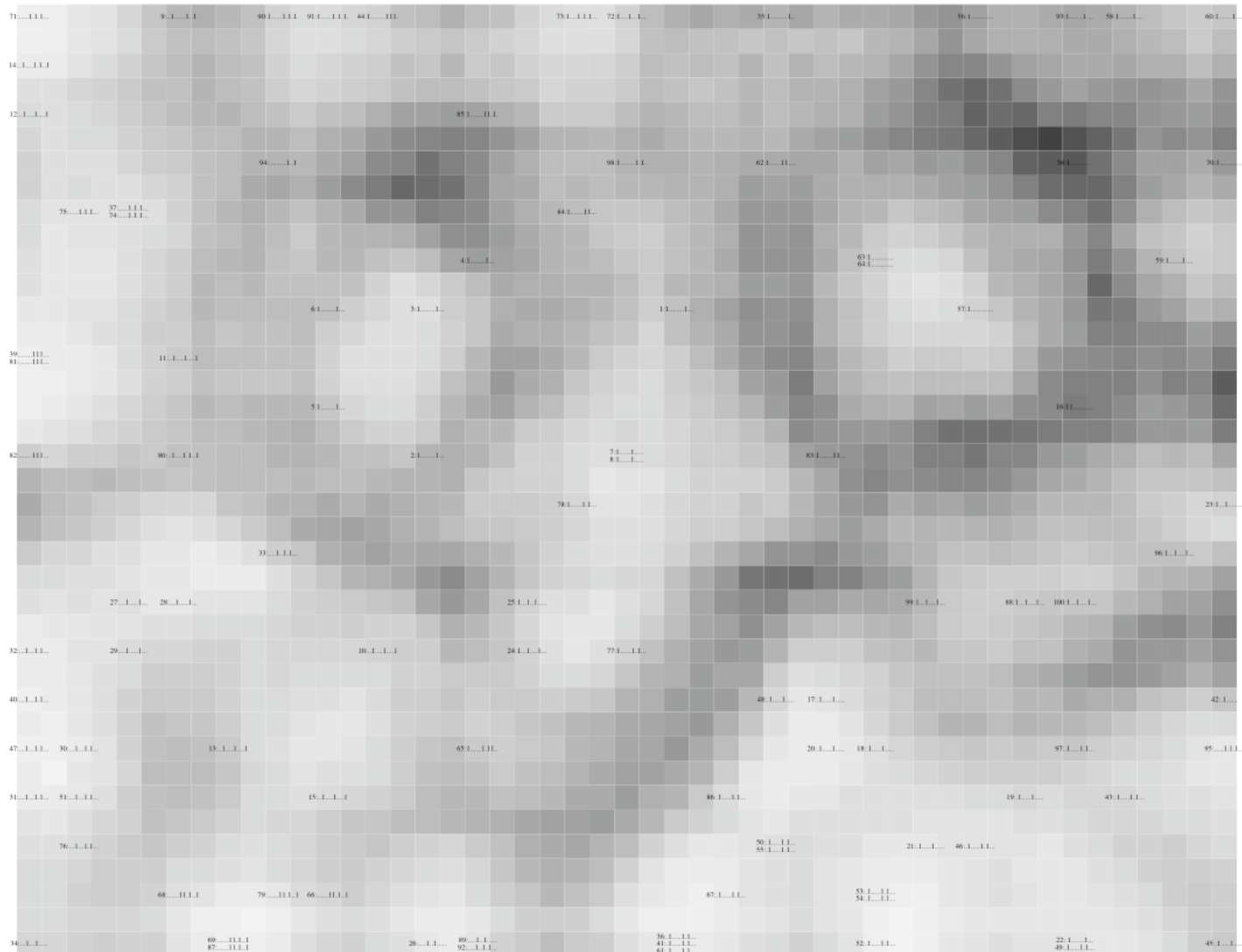
MIR

Significance Spectrum





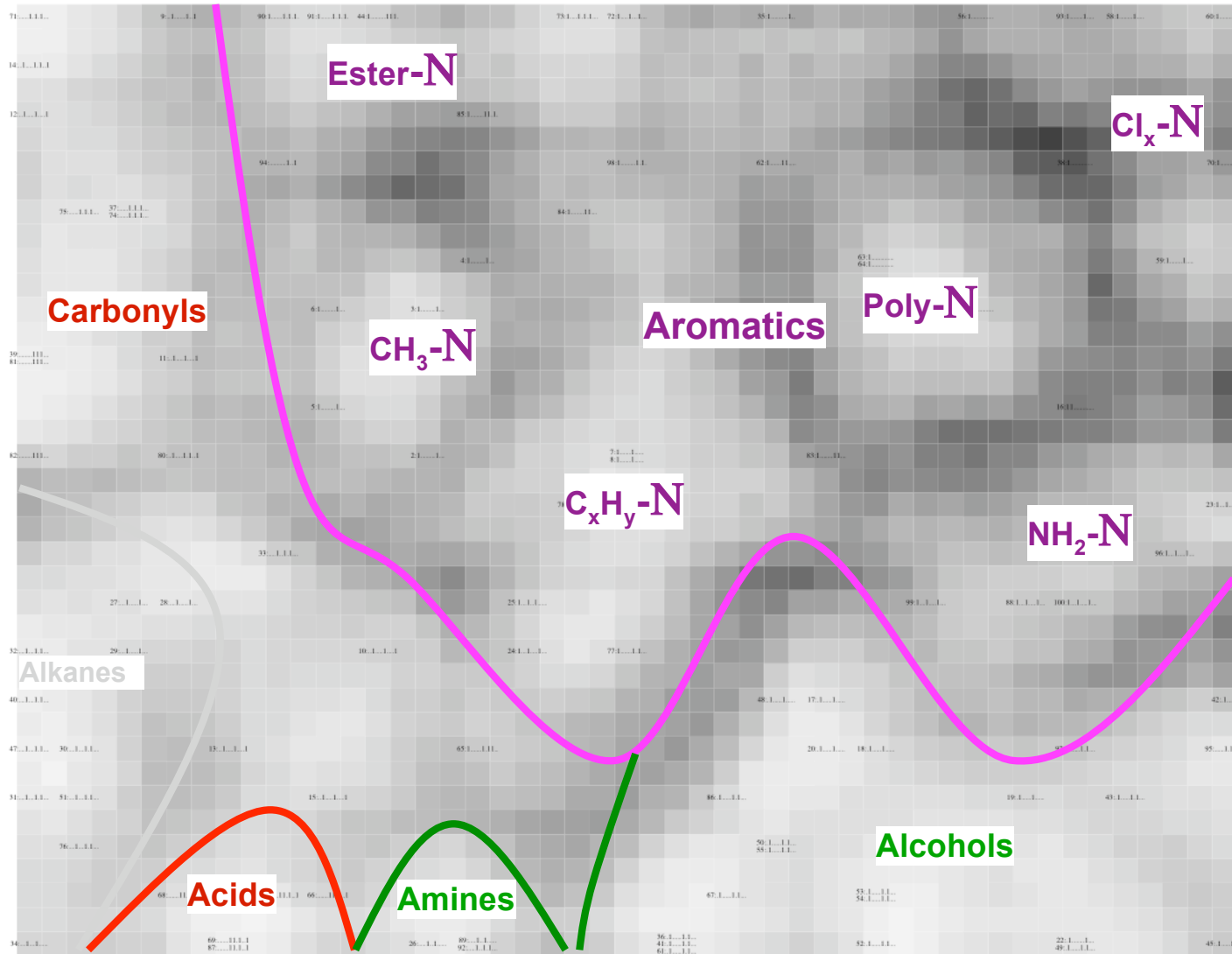
NIR SOM





NIR SOM

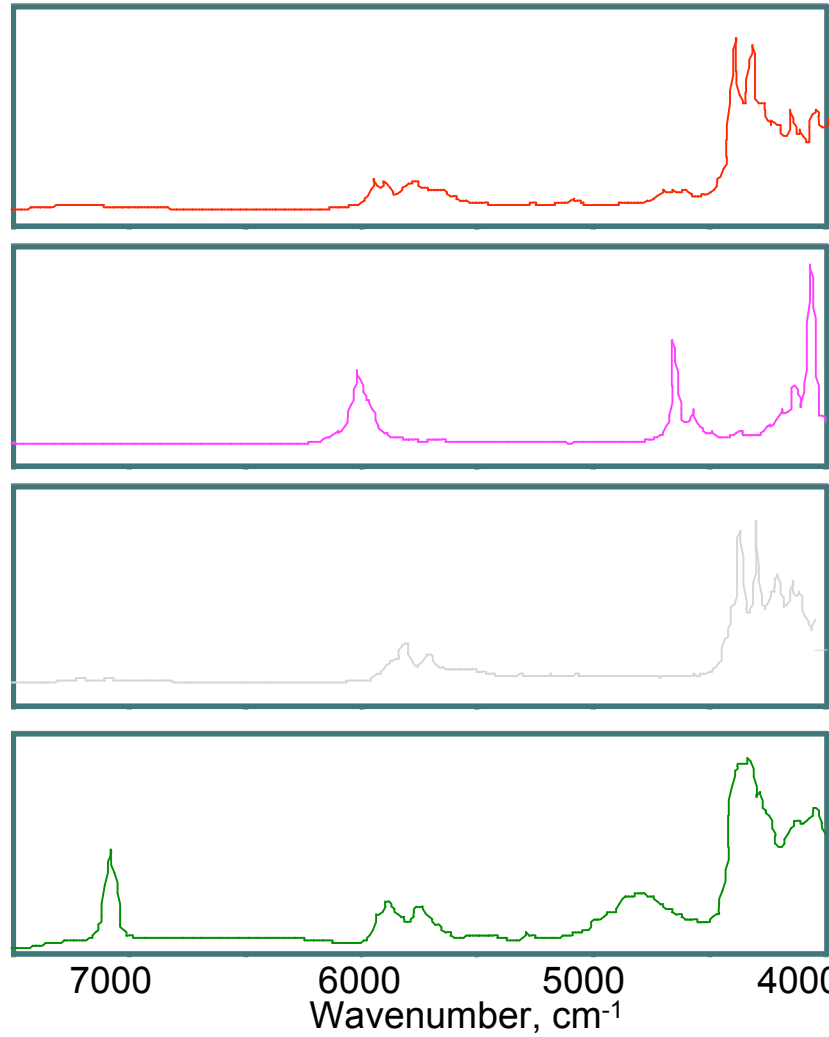
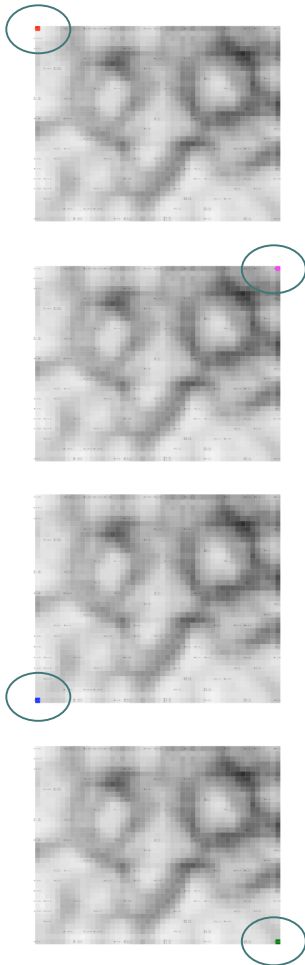
Functional Groups





NIR

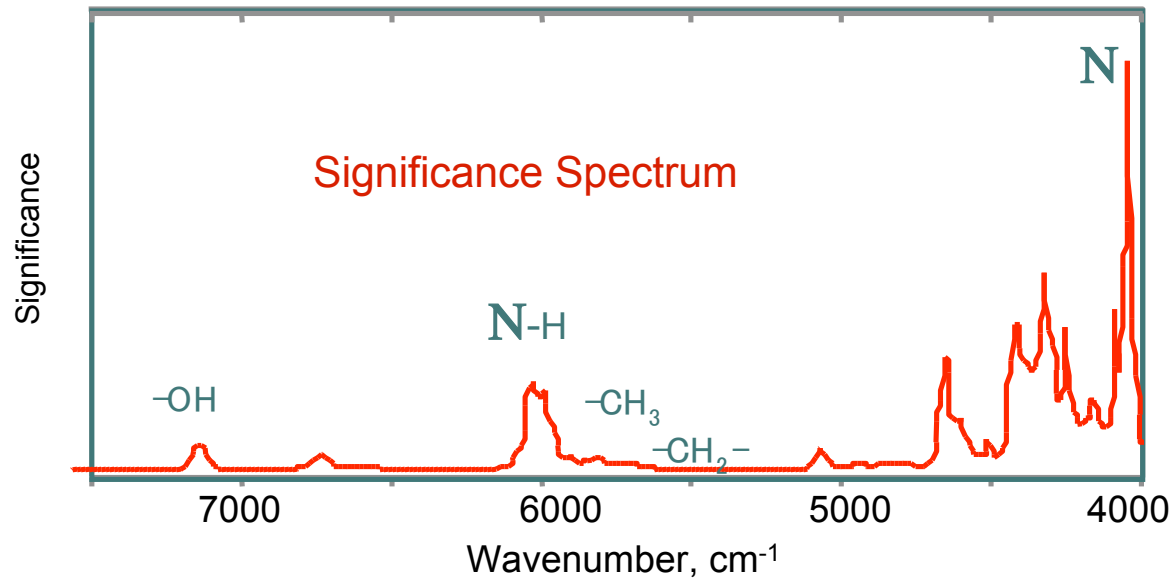
Centroid Spectra





NIR

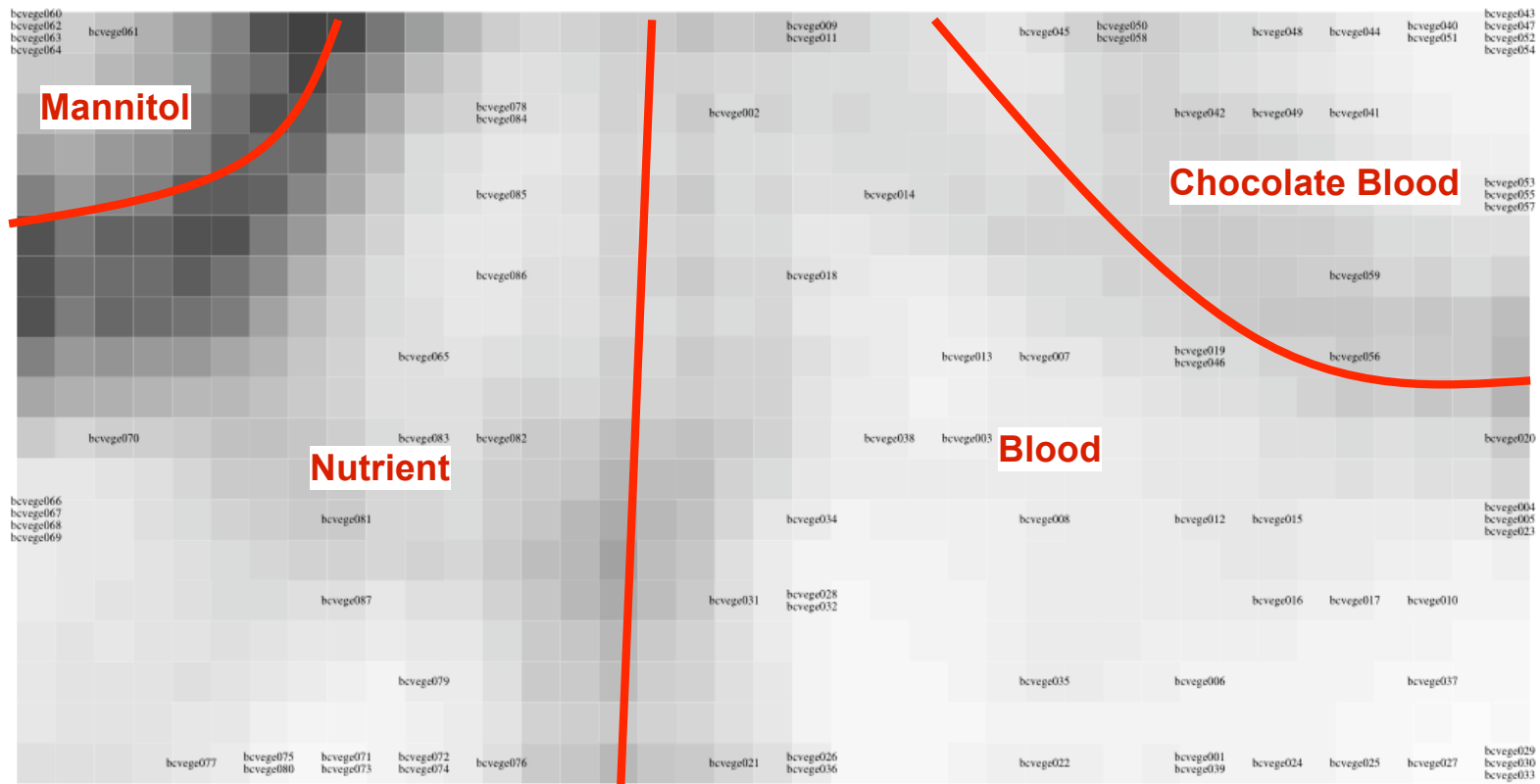
Significance Spectrum





SOM

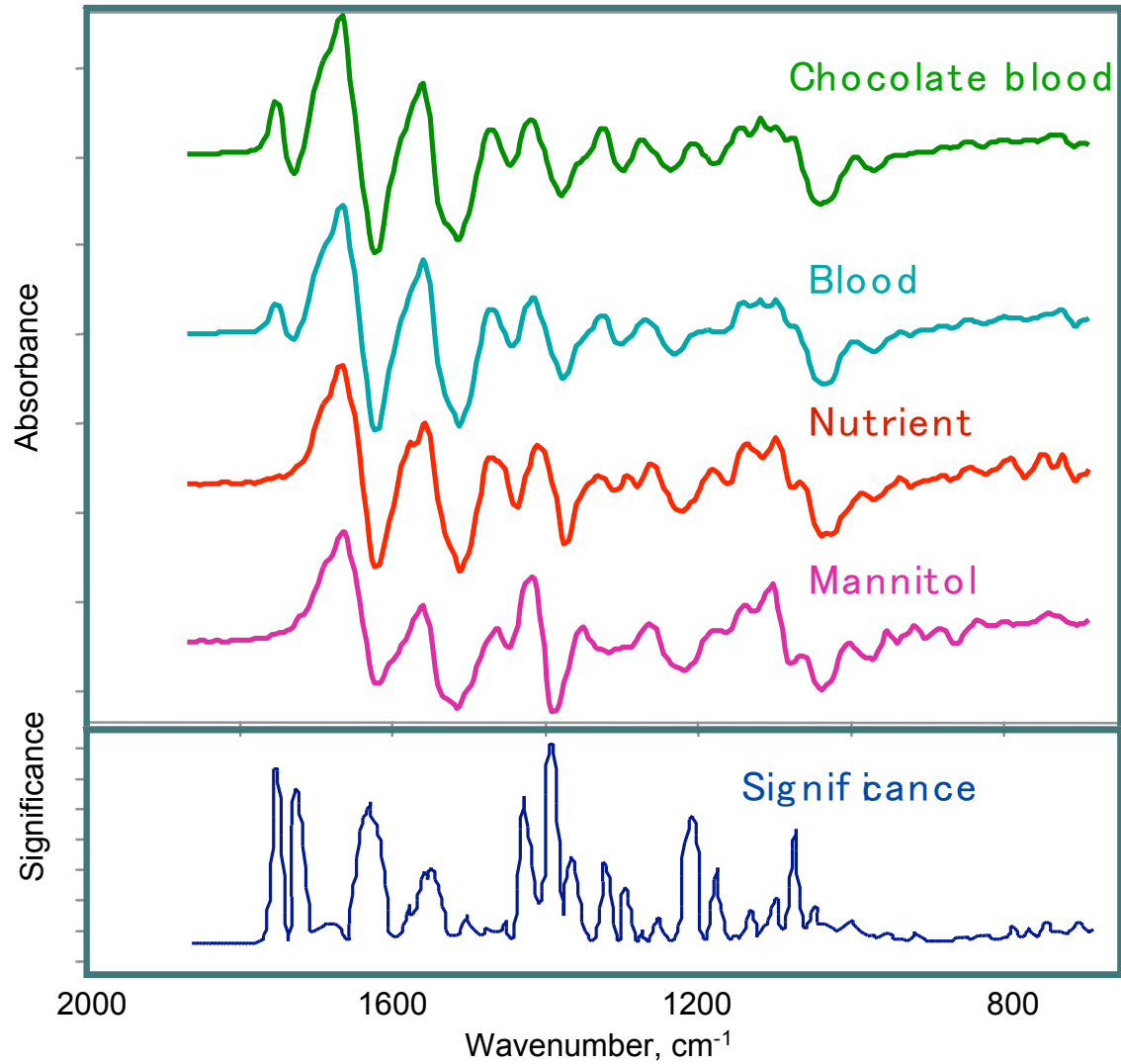
Bacterium *b-cereus* on different agars



"You are what you eat!"

Significance Spectrum

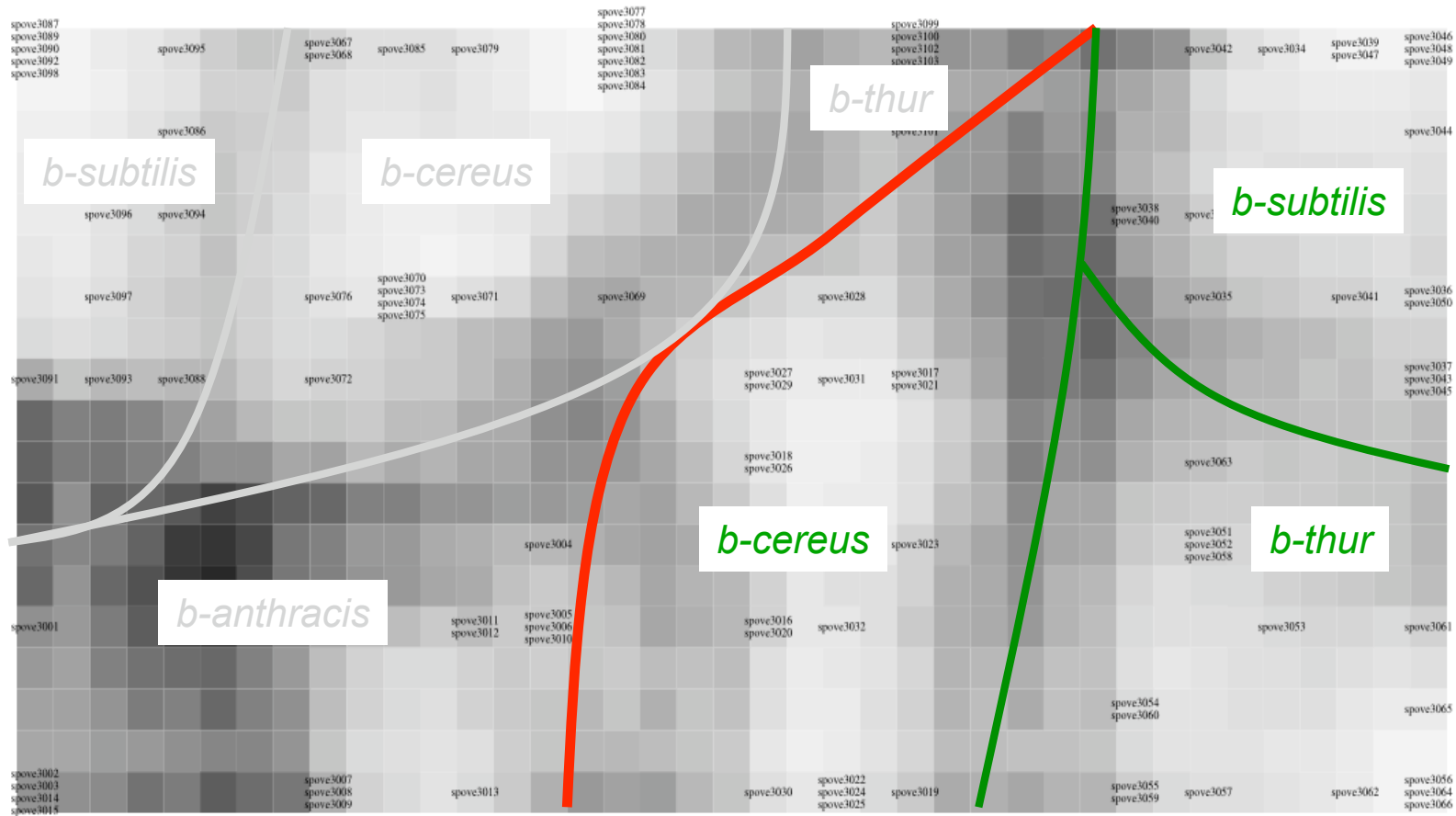
b-cereus on different agars





SOM

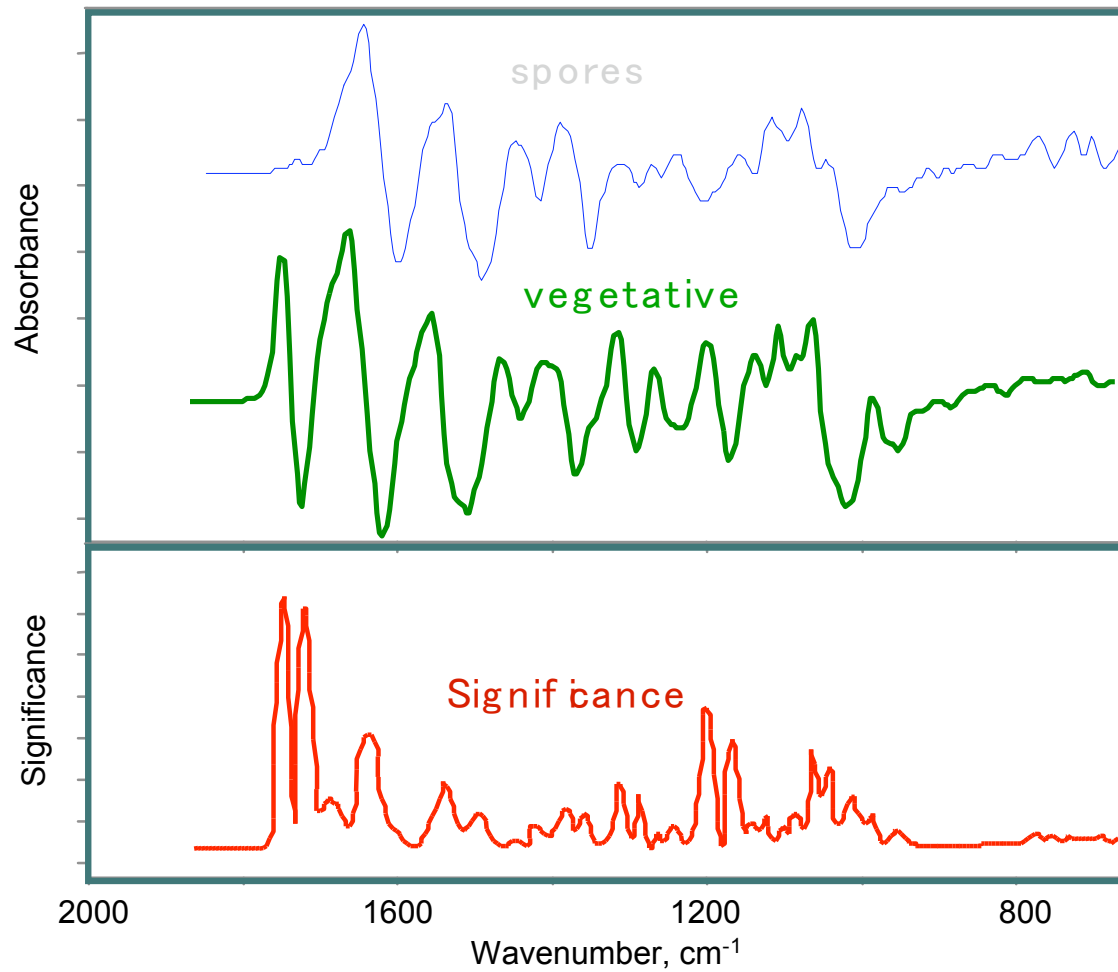
Bacteria Spectra



← spores / vegetative →



Significance Spectrum vs *b-subtilis* 1st Derivative Spectra



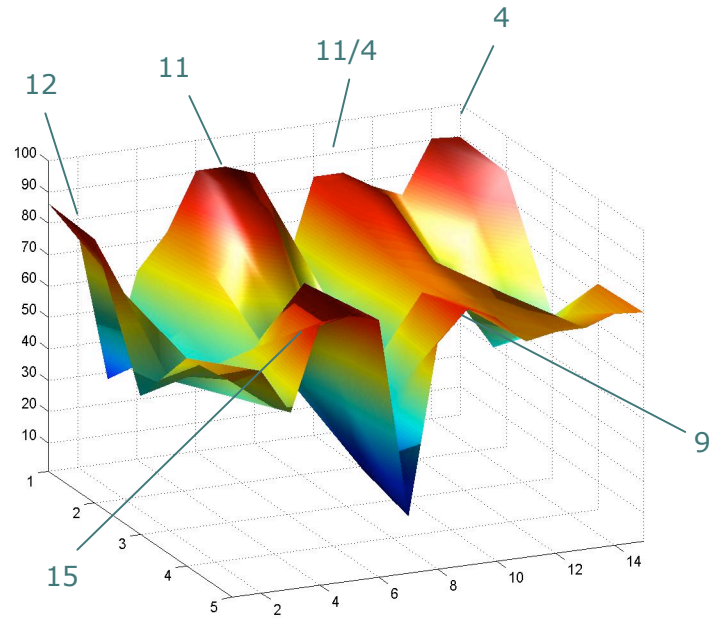
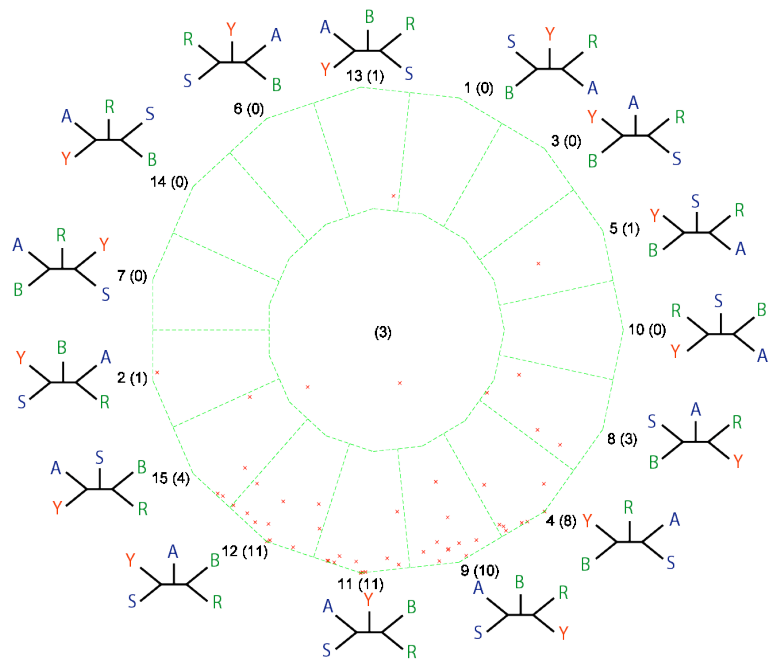
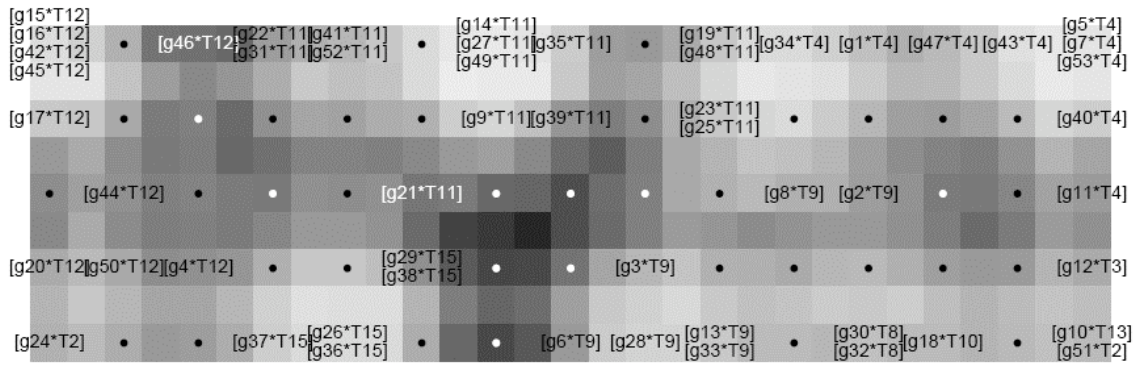


Applications of SOM

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



Phylogenetic Visualization with SOMs





Applications of SOM

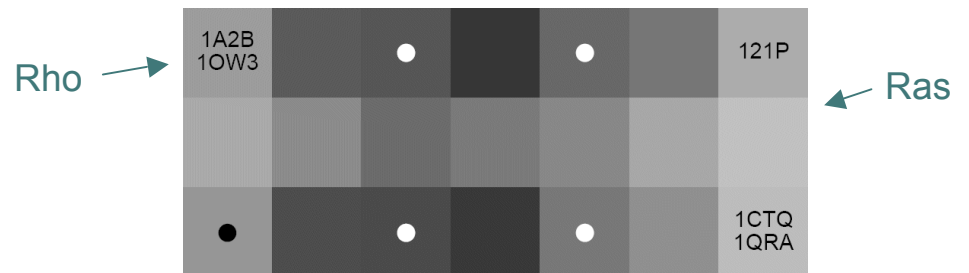
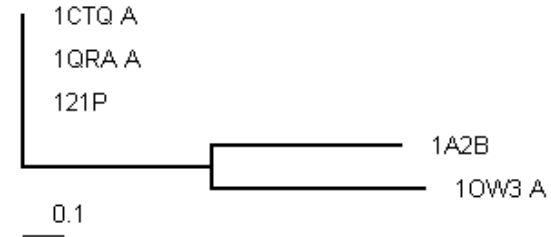
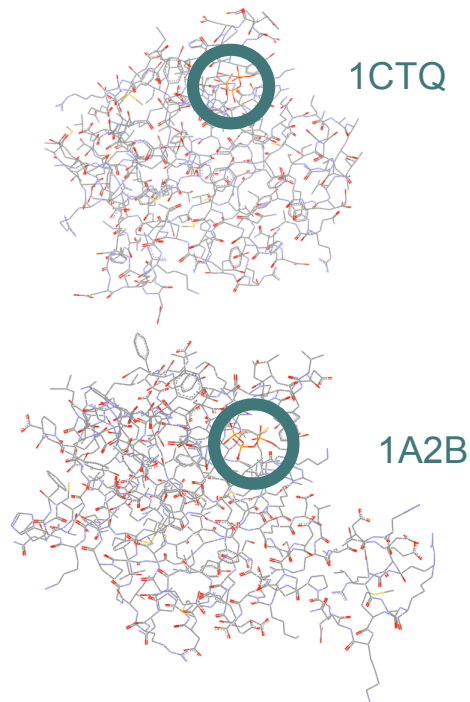
- Clustering Proteins based on the architecture of their activation loops.
 - Align the proteins under investigation.
 - Extract the functional centers.
 - Turn 3D representation into 1D feature vectors.
 - Cluster based on the feature vectors.



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



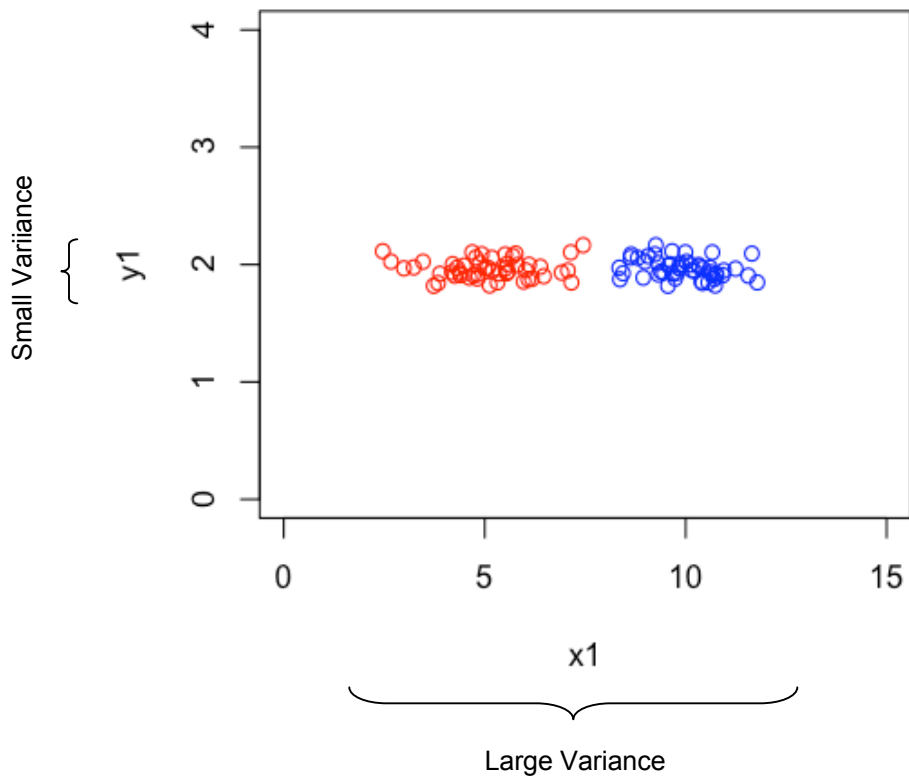


Two Central Questions

- Which features are the most important ones for clustering?
- How good is the map?



Variance Matters!



- Features with large variance have a higher probability of showing structure than features with small variance.
- Therefore, features with large variance tend to be more *significant* to the clustering process than features with small variance.



Bayes Theorem

- Using Bayes theorem we turn the observed variances (observed significances) into significance probabilities:

$P(A_i | +)$ \equiv observed significance of feature A_i

$P(+ | A_i)$ \equiv probability that feature A_i is significant (significance)

$P(A_i)$ \equiv prior of feature A_i

$$P(+ | A_k) = \frac{P(A_k | +)P(A_k)}{\sum_i P(A_i | +)P(A_i)} \quad \text{Significance of feature } A_k$$

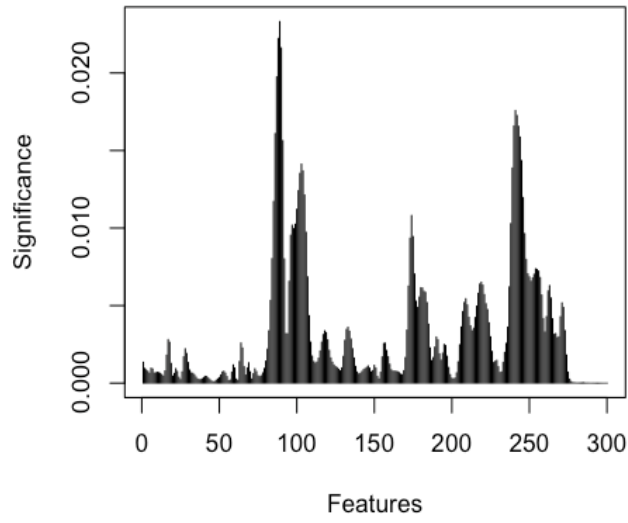


Probabilistic Feature Selection

- Given the significance probability of each feature of a data set we can ask interesting questions:
 - How many features do we need to include in our training data in order to be 95% sure that we included all the significant features?
 - What is the significance level of the top 10% of my features?

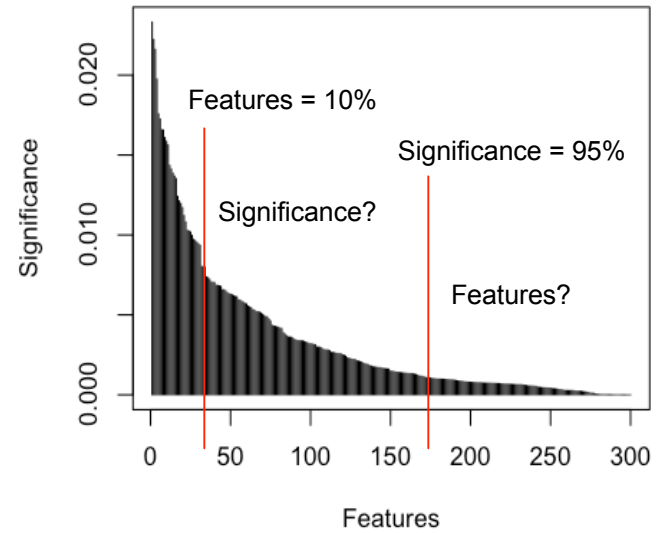


Feature Selection



Significance Plot

Probability Distribution

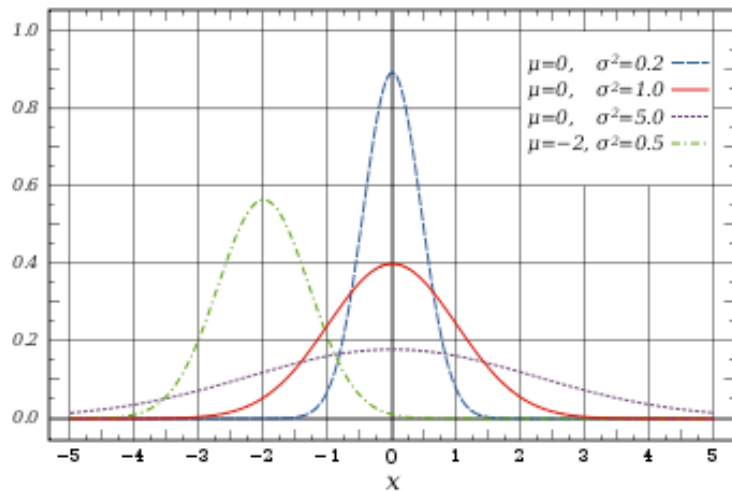




Evaluating a SOM

- The canonical performance metric for SOMs is the *quantization error*.
 - Very difficult to relate to the training data (e.g., how small is the optimal quantization error?)
- Here we take a different approach: we view a map as *non-parametric, generative model*.
- This gives rise to a new model evaluation criterion via the classical *two sample problem*.

Generative Models

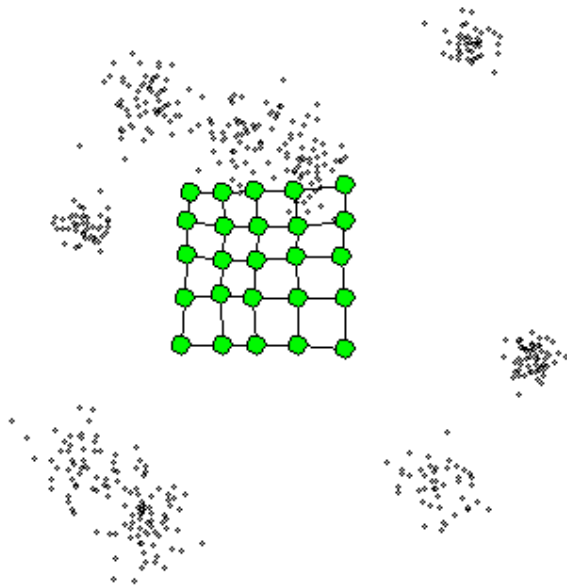


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

Notation: $N(\mu, \sigma^2), \forall x$

- A generative model is a model that we can sample and compute new values of the underlying input domain.
- The classical generative model is the Gaussian function, once we have fitted the function through our known samples, then we can compute the probability of any sample of the input domain.
- However, the model is *parametric*; it is governed by the mean μ and the standard deviation σ .

Insight: SOMs Sample the Data Space



- Given some distribution in the data space, SOM will try to construct a sample that looks like it was drawn from the same distribution.
- It will construct the sample using interpolation (neighborhood function) and constraints (the map grid).
- We can then measure the quality of the map using a *statistical two sample approach*.

Algorithm:

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




SOM as a Non-parametric Generative Model

Let D be our training set drawn from a distribution $N(\mu, \sigma^2)$, then $N(\mu_D, \sigma_D^2)$ is a good approximation to the original distribution if D is large enough,

$$N(\mu, \sigma^2) \approx N(\mu_D, \sigma_D^2).$$

Now, let M be the set of samples SOM constructs at its map grid nodes, then we say that SOM is *converged* if the mean μ_M and the variance σ_M^2 of the model samples appear to be drawn from the same underlying distribution $N(\mu, \sigma)$ as the training data,

$$N(\mu, \sigma^2) \approx N(\mu_M, \sigma_M^2)$$



SOM as a Non-parametric Generative Model

Now, the distribution $N(\mu, \sigma^2)$ is unknown, but we have a good approximation to it as our training set D , $N(\mu_D, \sigma_D^2)$.

Therefore, in order to test for convergence we have to show that,

$$N(\mu_M, \sigma_M^2) \approx N(\mu_D, \sigma_D^2),$$

or "*we test that the model samples and training samples were drawn from the same distribution*".

This is an application of the classical *statistical two sample test*; we use the *student - t test* to test that the means μ_M and μ_D are due to the same distribution and we use the *F - test* to show that the variances σ_M^2 and σ_D^2 are due to the same distribution.



SOM as a Non-parametric Generative Model

- Observations:
 - The SOM model is non-parametric (or distribution free) since there are no distribution parameters to fit.
 - We can sample from a SOM model using linear interpolation on the node grid.
 - A converged model is a good fitting model, it models the underlying distribution very well.



Conclusions

- SOMs have a wide range of applications.
- We have developed two statistical tools that allow us to evaluate SOMs very effectively:
 - Probabilistic feature selection
 - Goodness of fit.
- In the future we need to address the reliance on normal distributions in our tests...resampling techniques (e.g. bootstrap)