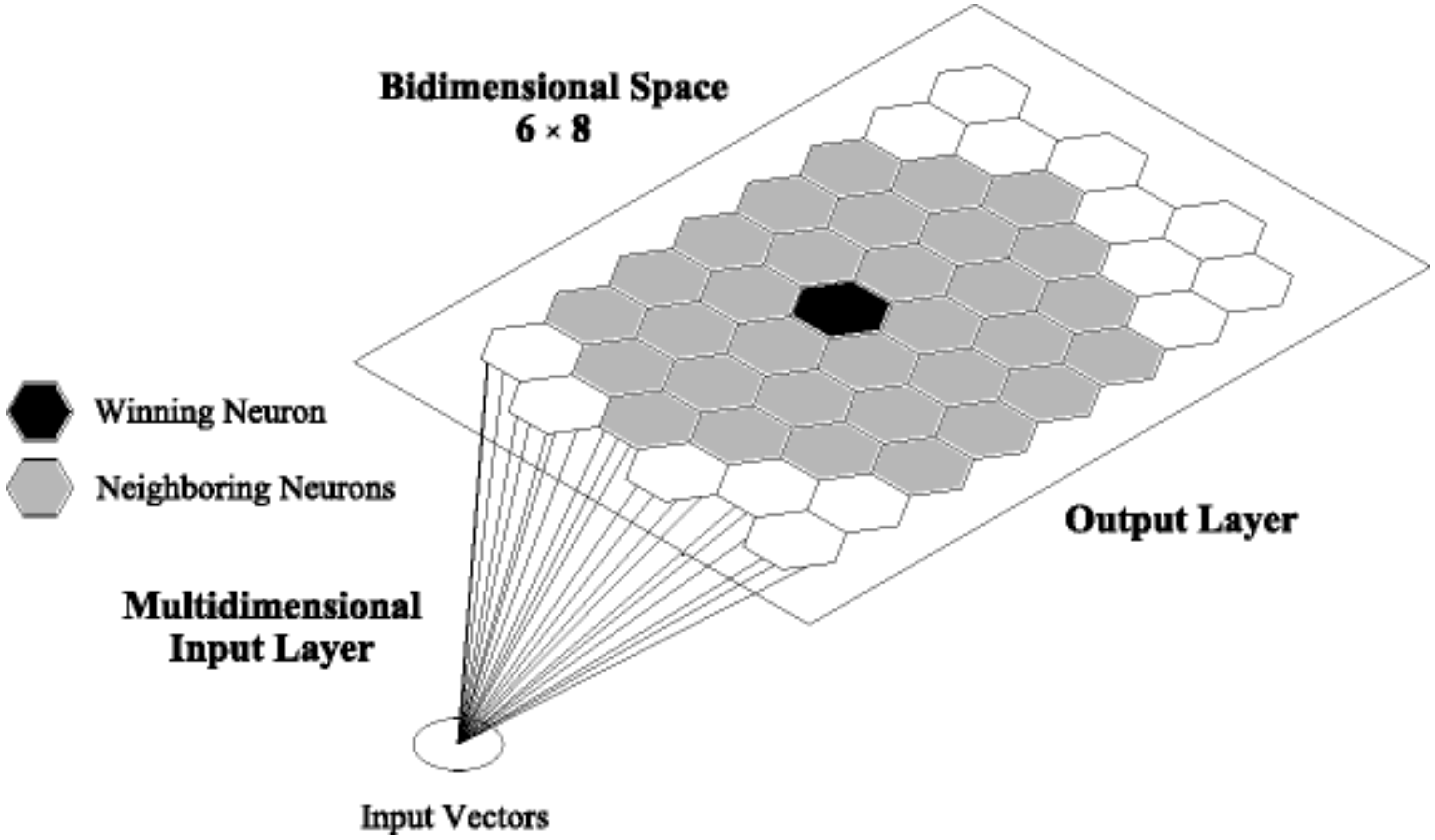
The Self-Organizing Map (SOM)

- A neural network approach to unsupervised machine learning.
- Appealing visual presentation of learned results as a 2D map.
- Learned result represents a partitioning of the input space according to a similarity metric.

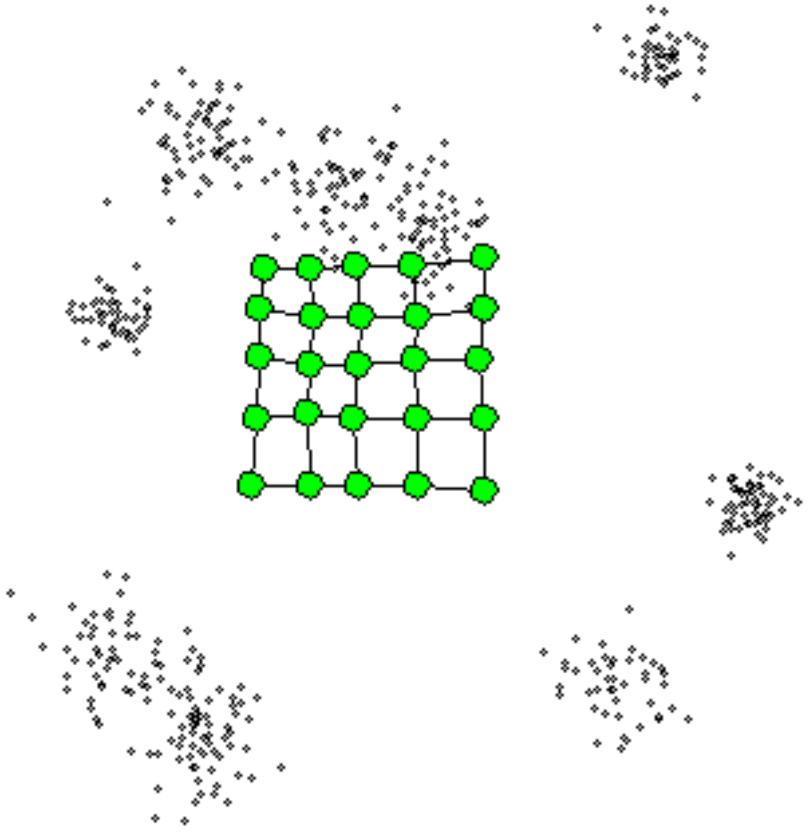
T. Kohonen, *Self-organizing maps*, 3rd ed. Berlin ; New York: Springer, 2001.

SOM Architecture

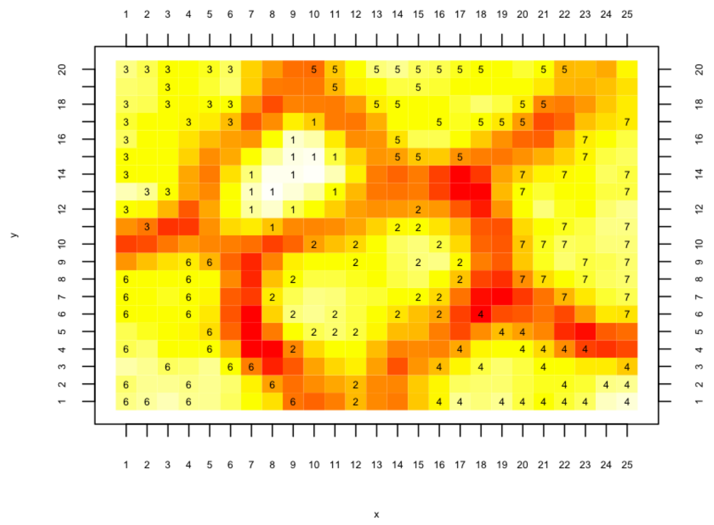


Source: http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40677-015-0022-9/MediaObjects/40677_2015_22_Fig3_HTML.gif

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*.



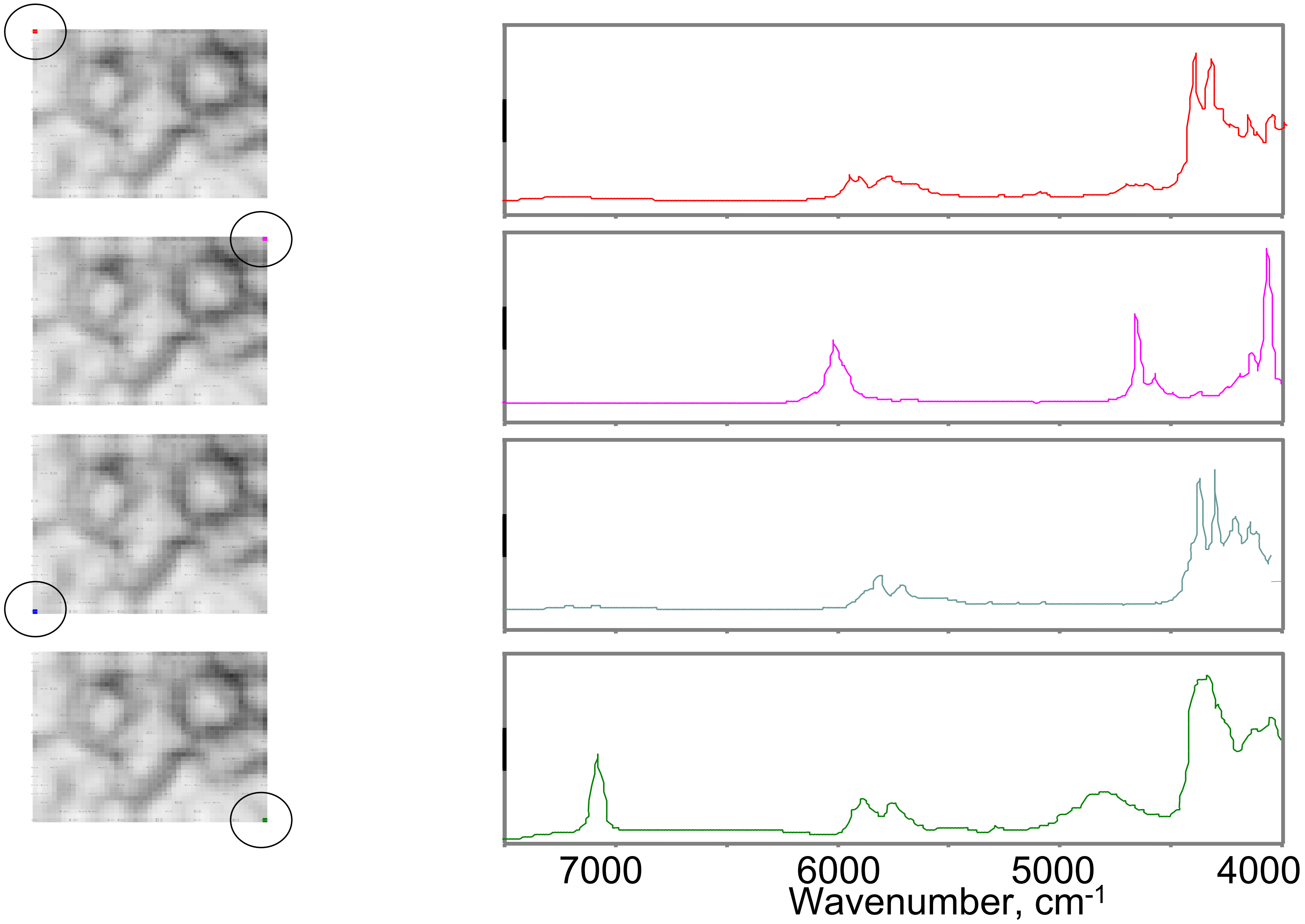
Algorithm:

```
Repeat until Done
  For each training observation Do
    Find the neuron that best matches the observation.
    Make that neuron look more like the observation.
    Smooth the immediate neighborhood of that neuron.
  End For
End Repeat
```

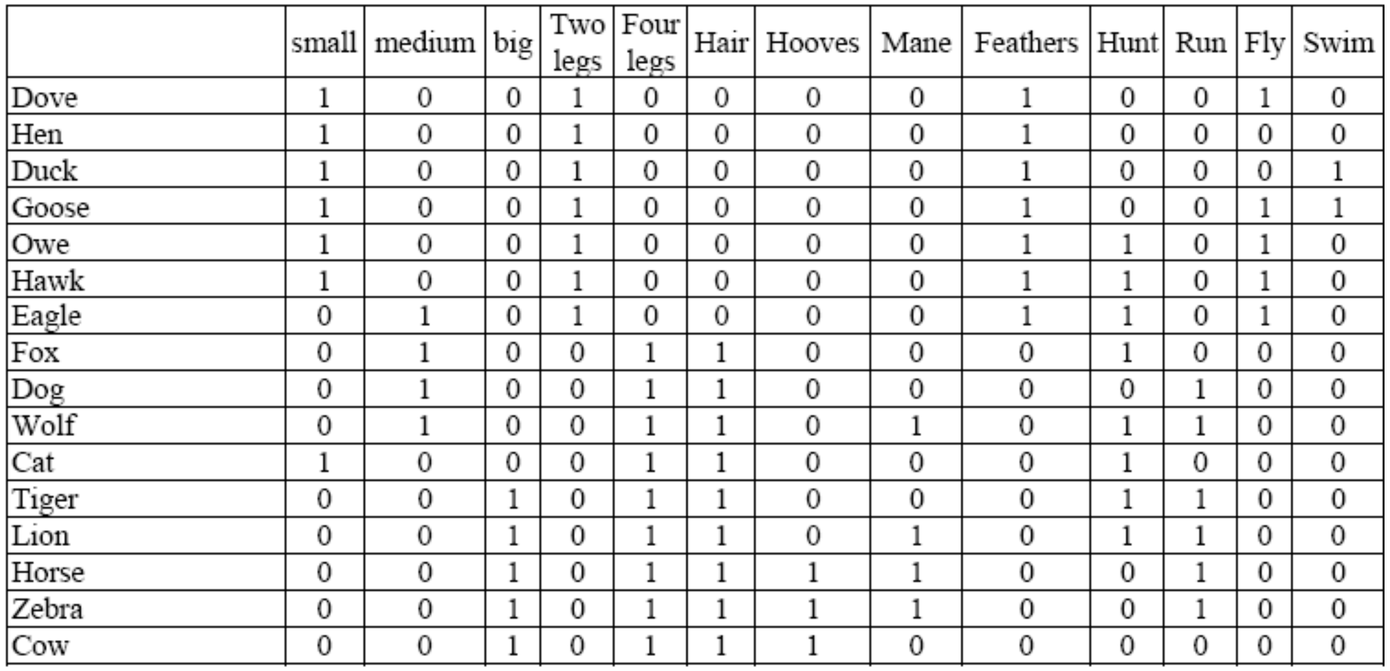
Animation source: www.peltarion.com

SOM Neurons

SOM neurons capture information about the input data – similar to the centroids in k-means.



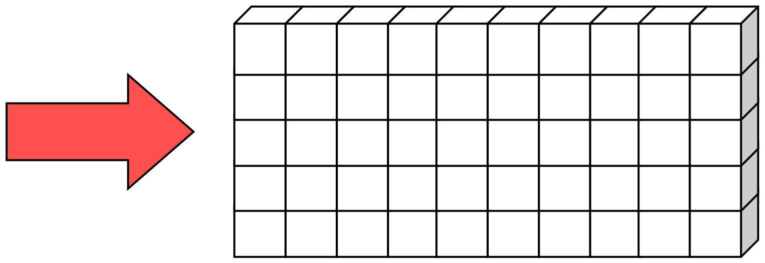
Training a SOM



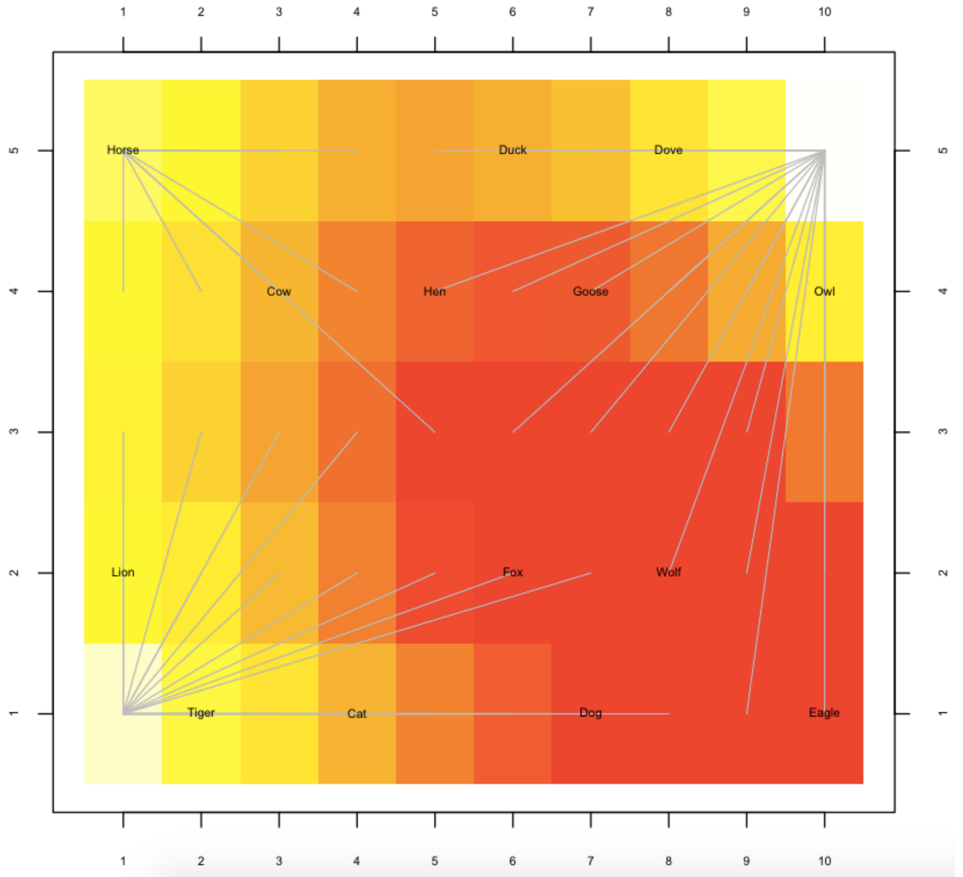
	small	medium	big	Two legs	Four legs	Hair	Hooves	Mane	Feathers	Hunt	Run	Fly	Swim
Dove	1	0	0	1	0	0	0	0	1	0	0	1	0
Hen	1	0	0	1	0	0	0	0	1	0	0	0	0
Duck	1	0	0	1	0	0	0	0	1	0	0	0	1
Goose	1	0	0	1	0	0	0	0	1	0	0	1	1
Owe	1	0	0	1	0	0	0	0	1	1	0	1	0
Hawk	1	0	0	1	0	0	0	0	1	1	0	1	0
Eagle	0	1	0	1	0	0	0	0	1	1	0	1	0
Fox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



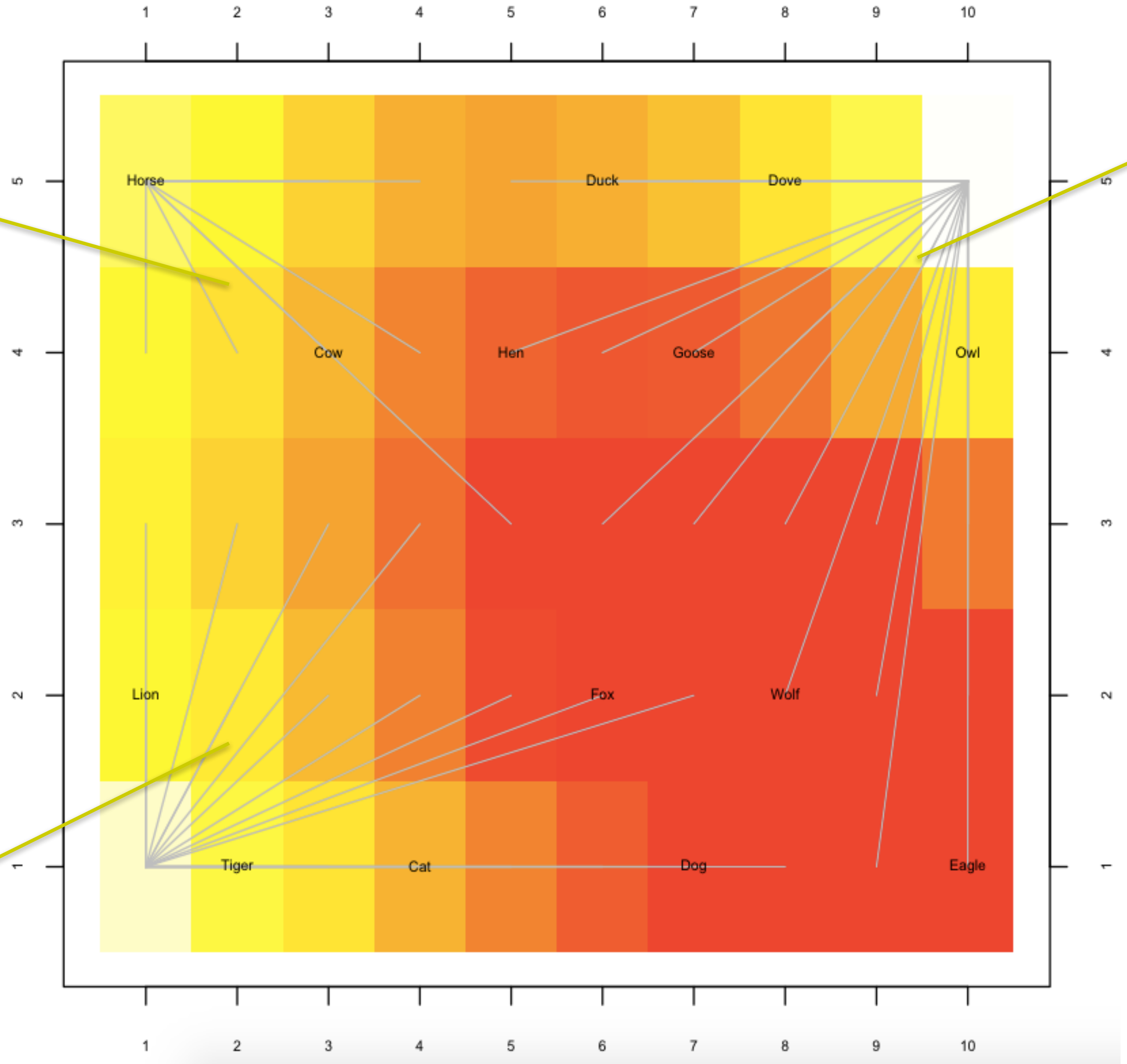
Visualization



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 Visualization



Ungulate Cluster

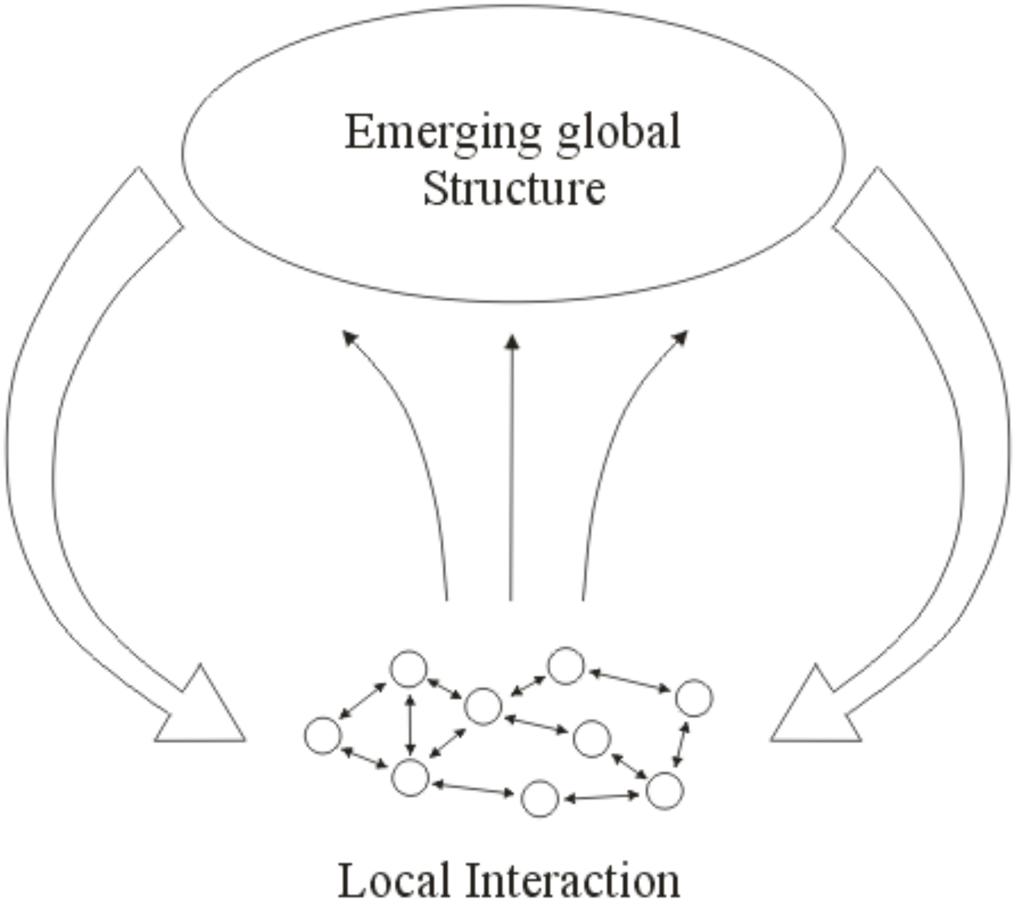
Bird Cluster

Hunter Cluster

Self-Organization and Learning

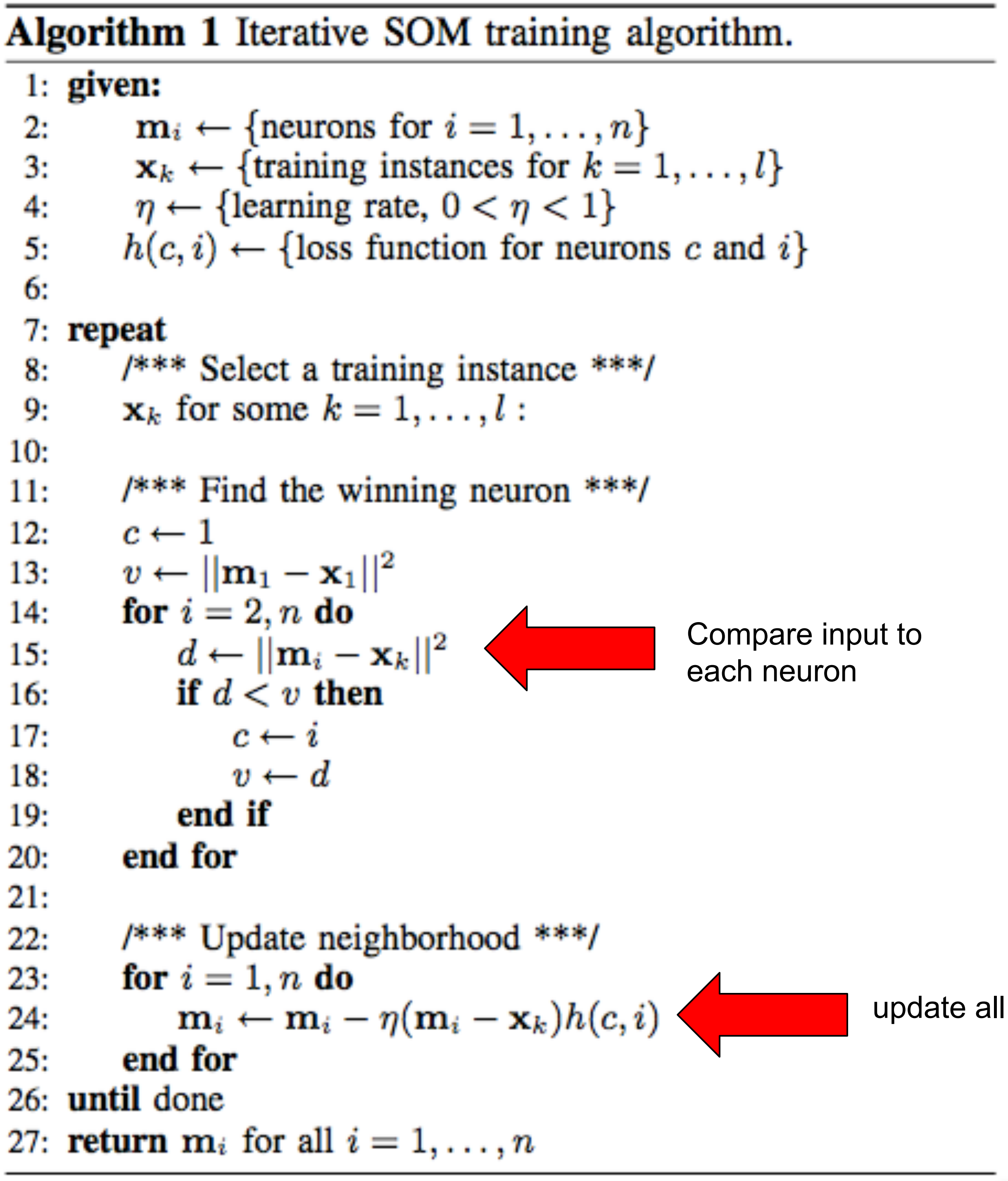
- **Self-organization** refers to a process in which the internal organization of a system increases automatically without being guided or managed by an outside source.
- This process is due to local interaction with simple rules.
- Local interaction gives rise to global structure.

☞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

Iterative Training: SOM



Algorithm 1 Iterative SOM training algorithm.

```
given:
    $\mathbf{m}_i \leftarrow$ {neurons for $i = 1, \dots, n$}
    $\mathbf{x}_k \leftarrow$ {training instances for $k = 1, \dots, l$}
    $\eta \leftarrow$ {learning rate, $0 < \eta < 1$}
    $h(c, i) \leftarrow$ {loss function for neurons $c$ and $i$}

repeat
    /*** Select a training instance ***/
    $\mathbf{x}_k$ for some $k = 1, \dots, l$ :

    /*** Find the winning neuron ***/
    $c \leftarrow 1$
    $v \leftarrow ||\mathbf{m}_1 - \mathbf{x}_1||^2$
    for $i = 2, n$ do
        $d \leftarrow ||\mathbf{m}_i - \mathbf{x}_k||^2$
        if $d < v$ then
            $c \leftarrow i$
            $v \leftarrow d$
        end if
    end for

    /*** Update neighborhood ***/
    for $i = 1, n$ do
        $\mathbf{m}_i \leftarrow \mathbf{m}_i - \eta(\mathbf{m}_i - \mathbf{x}_k)h(c, i)$
    end for
until done
return $\mathbf{m}_i$ for all $i = 1, \dots, n$
```

Compare input to each neuron

update all neurons

The Loss Function:

$$h(c, i) = \begin{cases} 1 & \text{if } i \in \Gamma(c), \\ 0 & \text{otherwise,} \end{cases}$$

The Neighborhood Function (time dependent):

$$\Gamma(c)|_{t=0} = \{1, 2, \dots, n\}.$$

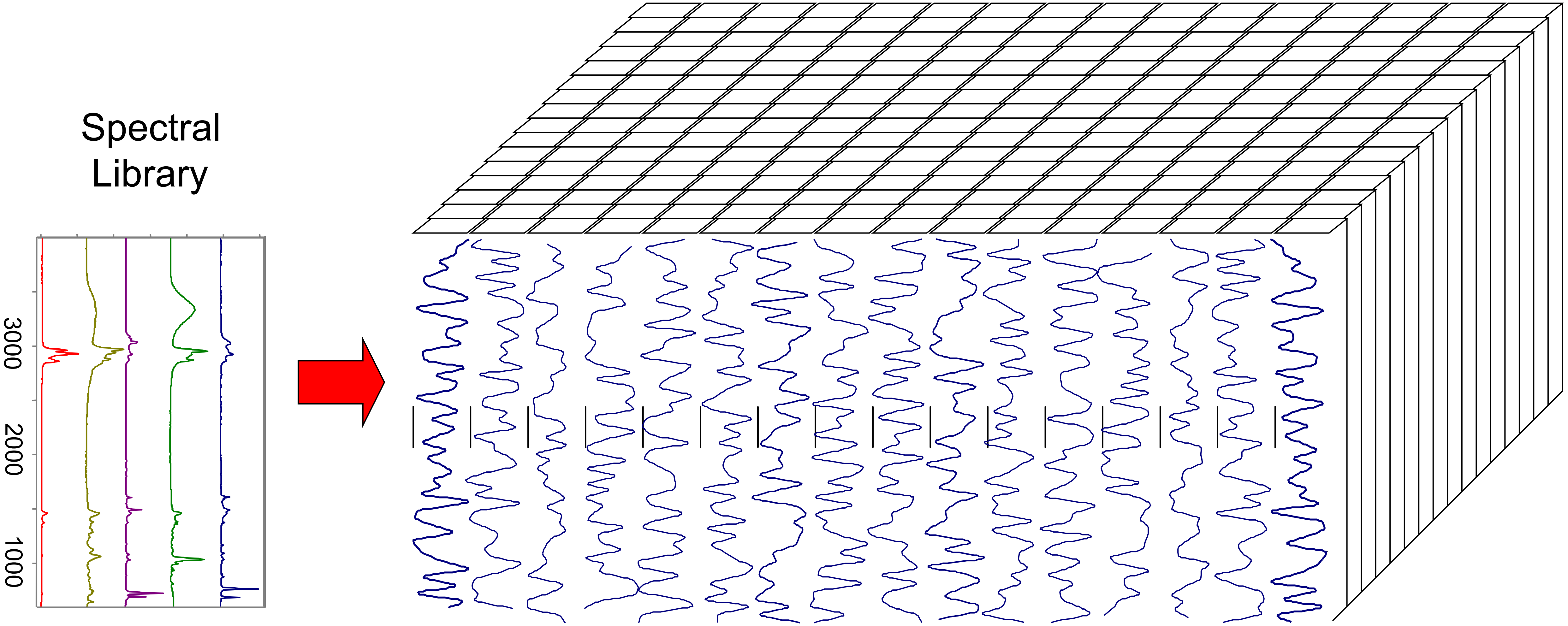
$$\Gamma(c)|_{t \gg 0} = \{c\}.$$

This is the "standard" training algorithm for SOMs implemented in most packages.

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



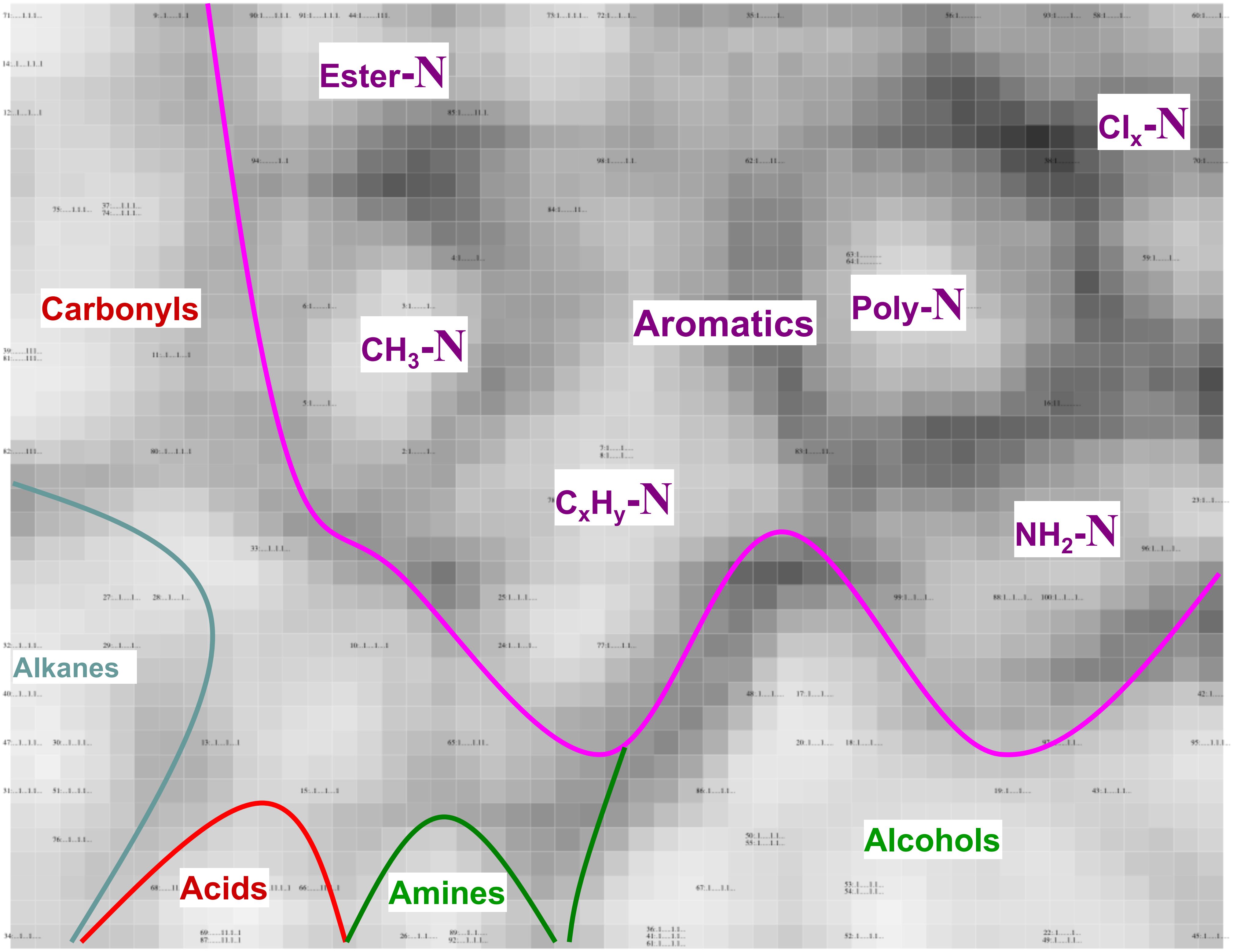
Grid of Neurons

Spectral Library

Random Number Spectra

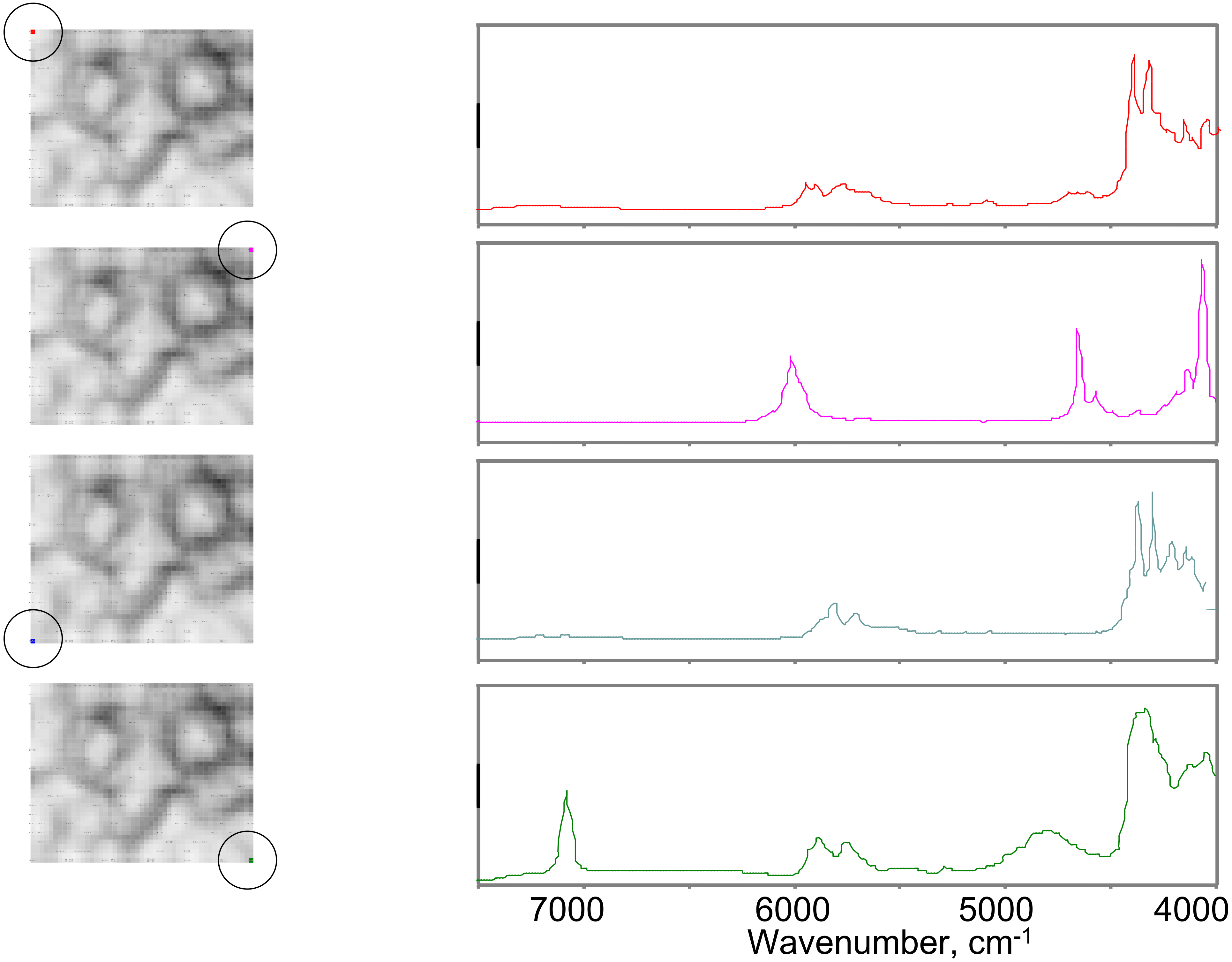
NIR SOM

Functional Groups



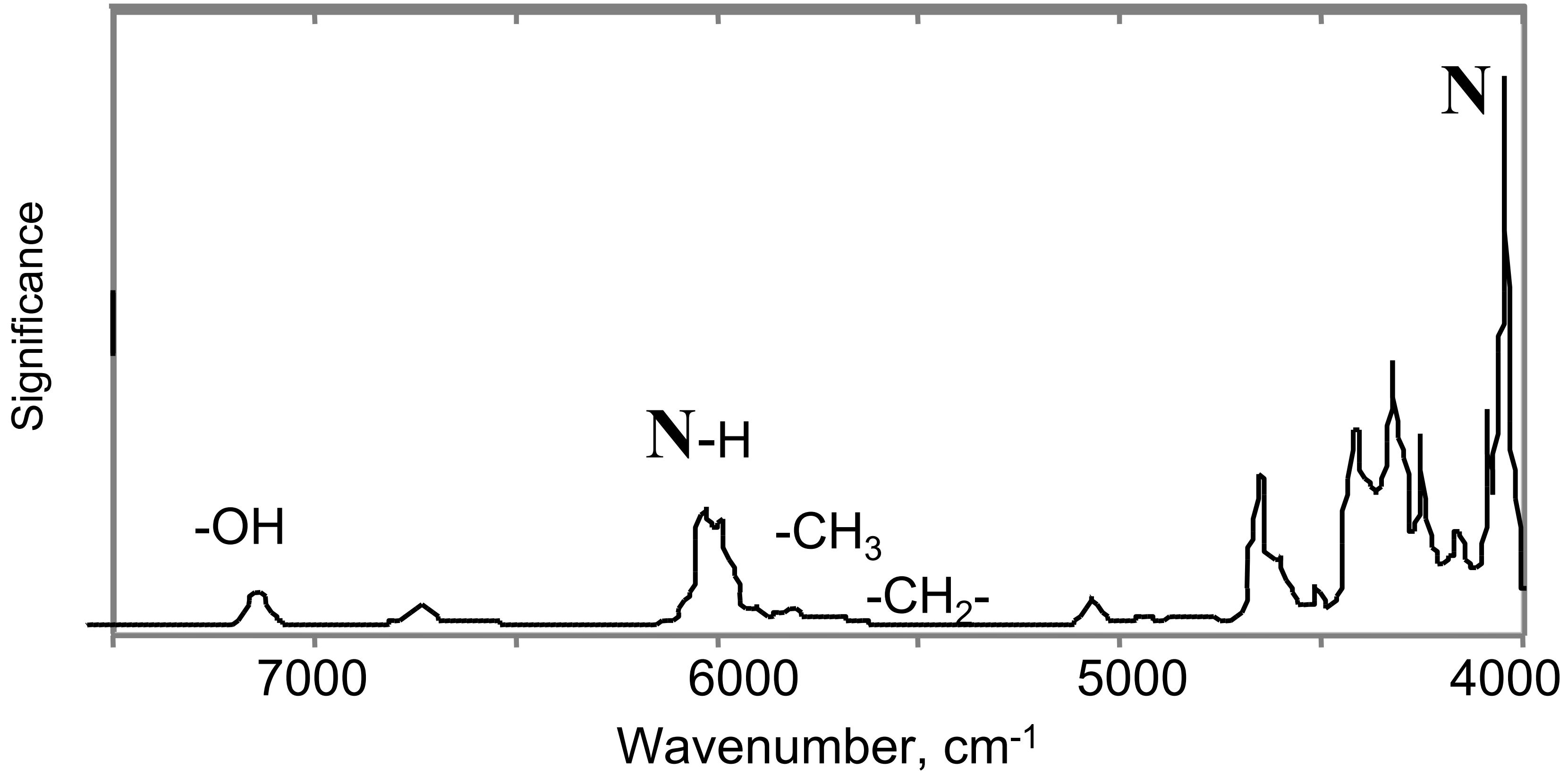
NIR

Centroid Spectra



NIR

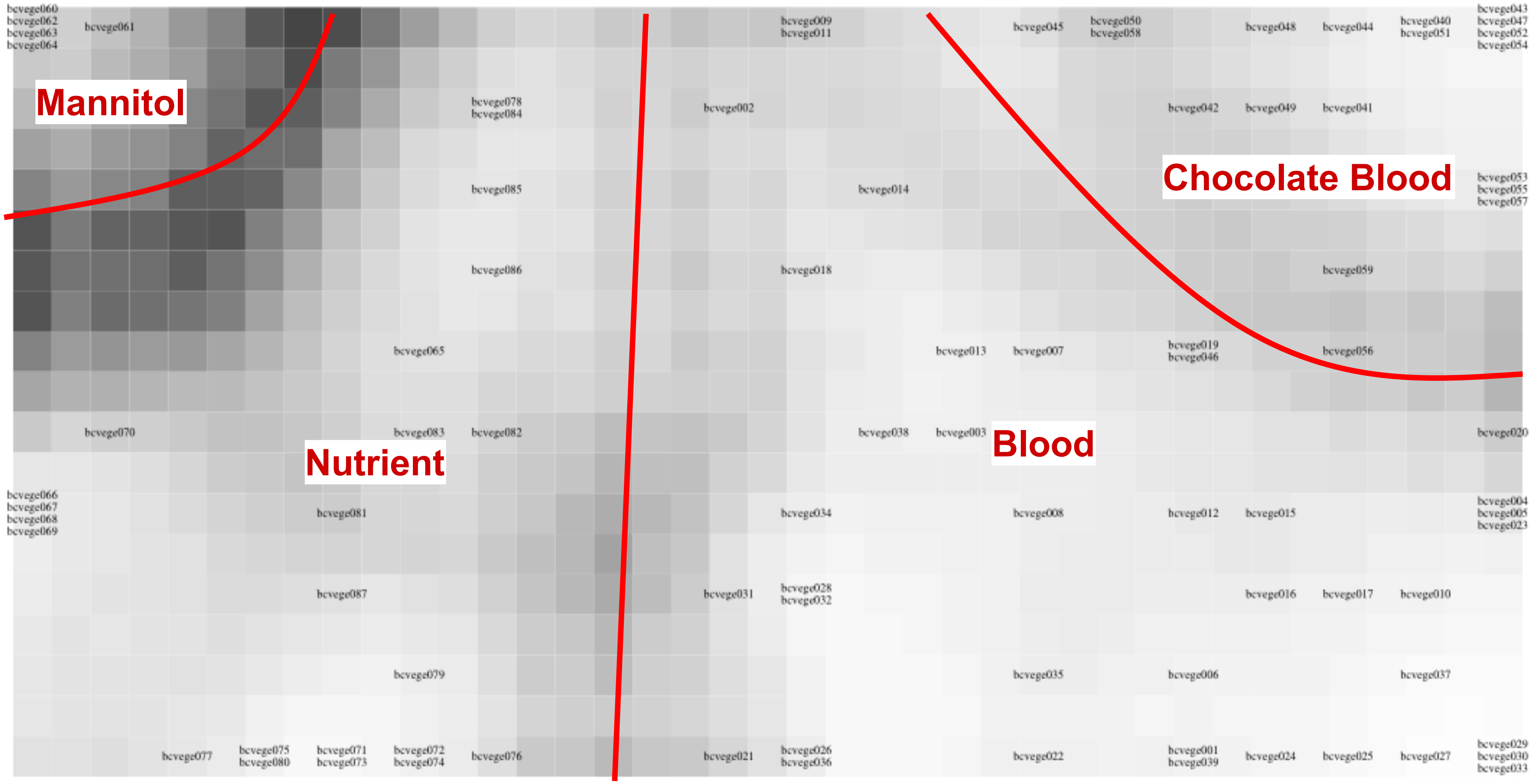
Significance Spectrum



SOM

Bacterium *b-cereus* on different agars

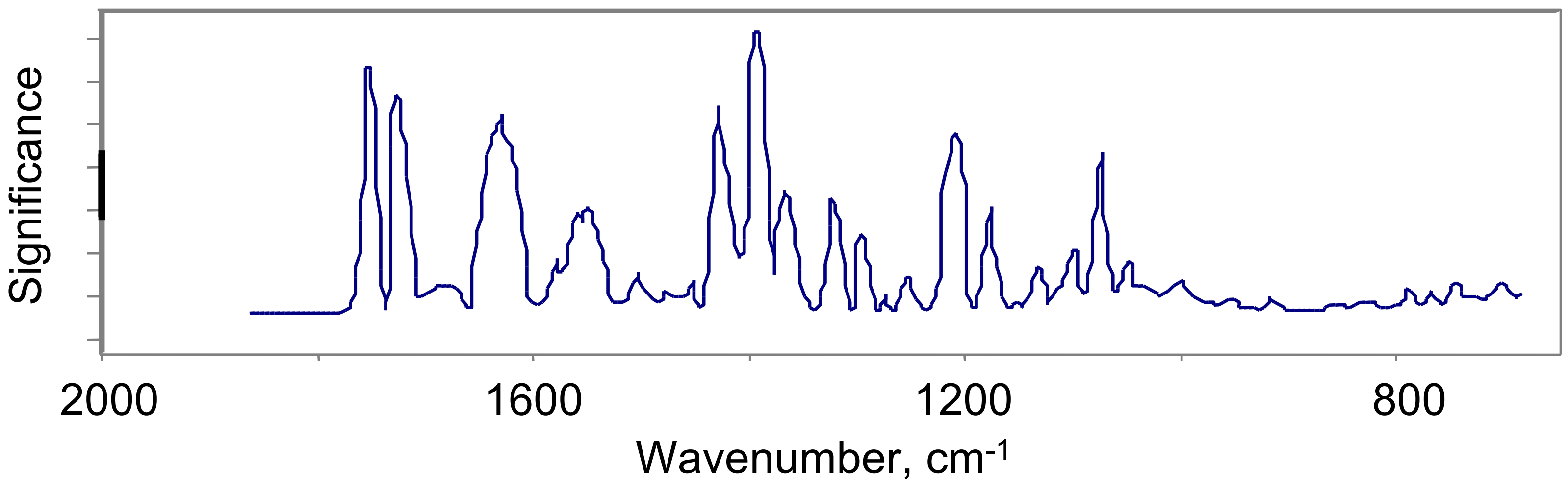
“You are what you eat!”



“You are what you eat!”

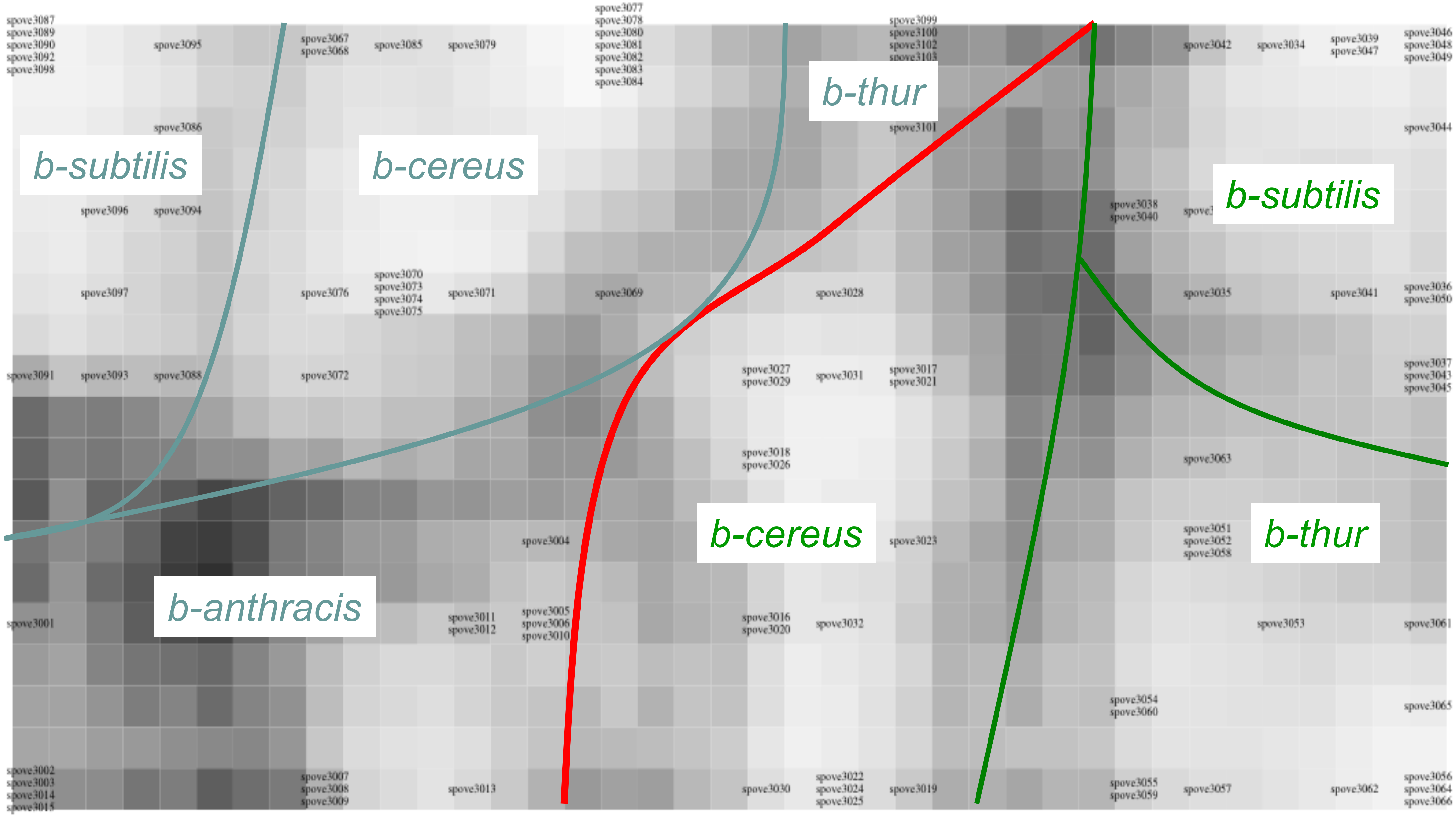
Significance Spectrum

b-cereus on different agars



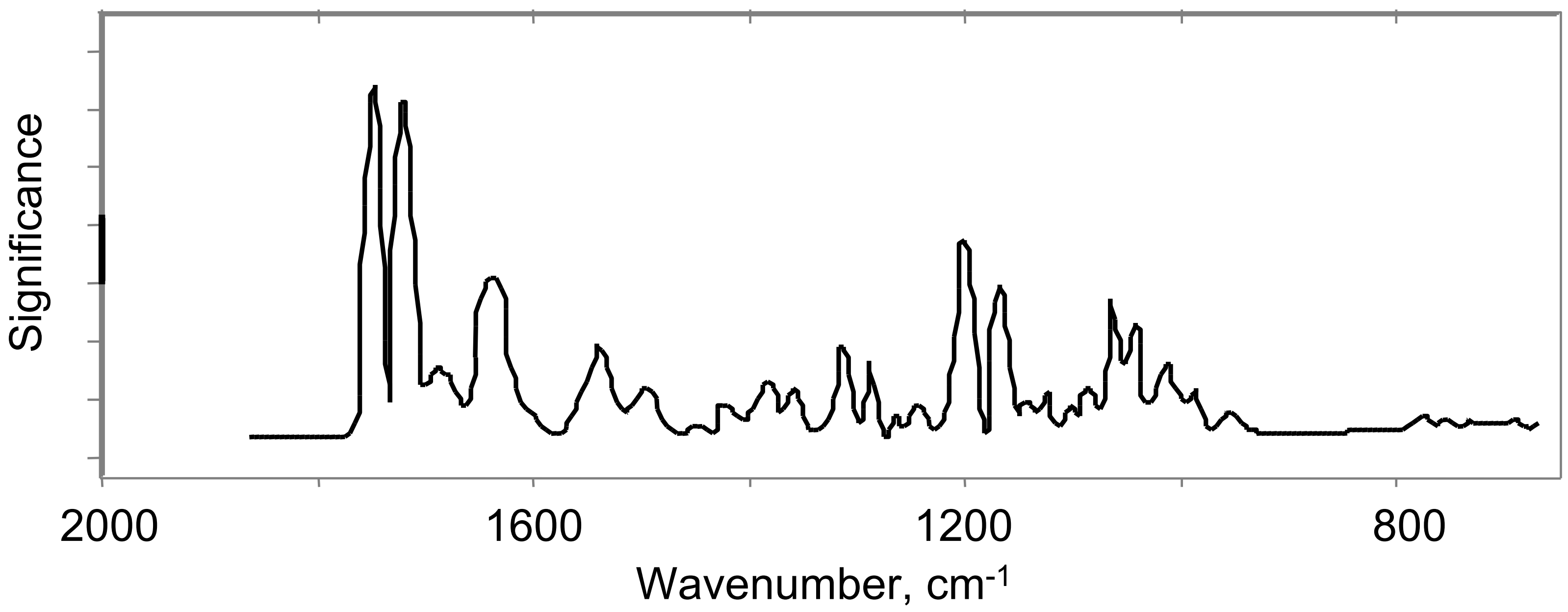
SOM

Bacteria Spectra



← spores / vegetative →

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



Applications of SOM

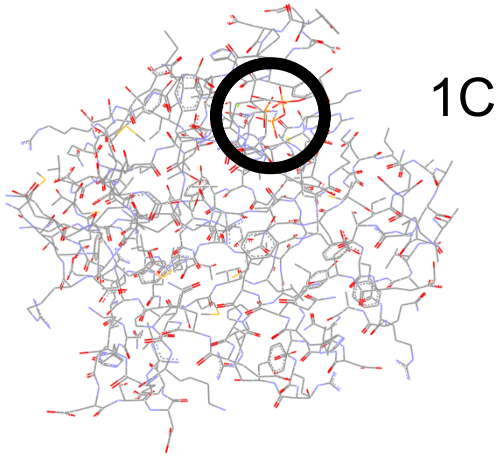
- Clustering Proteins based on the architecture of their activation loops.
  - Align the proteins under investigation.
  - Extract the functional centers.
  - Turn 3D representation into 1D feature vectors.
  - Cluster based on the feature vectors.

Joint work with Dr. Gongqin Sun, Dept. of Cell and Molecular Biology, URI

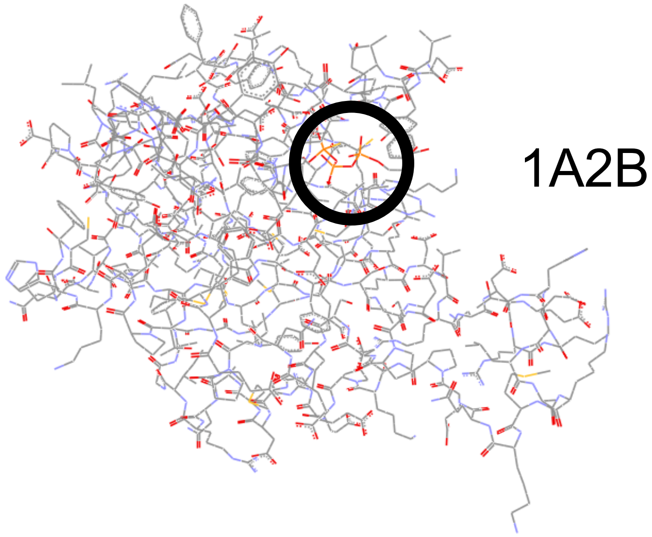
Structural Classification of GTPases

Can we structurally distinguish between the Ras and Rho subfamilies?

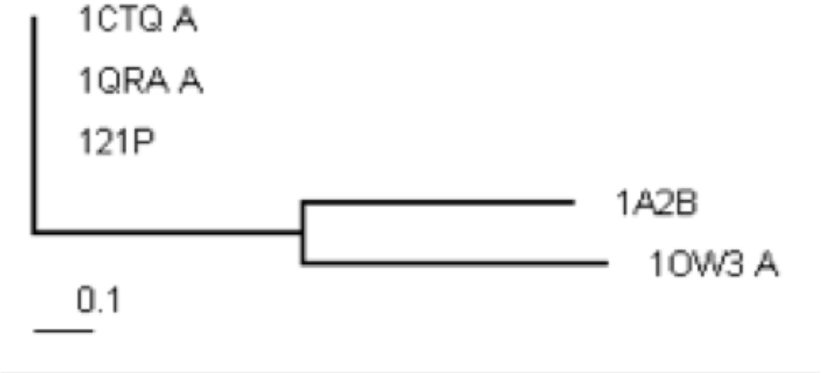
- Ras: 121P, 1CTQ, and 1QRA
- Rho: 1A2B and 1OW3
- F = p-loop, r = 10Å

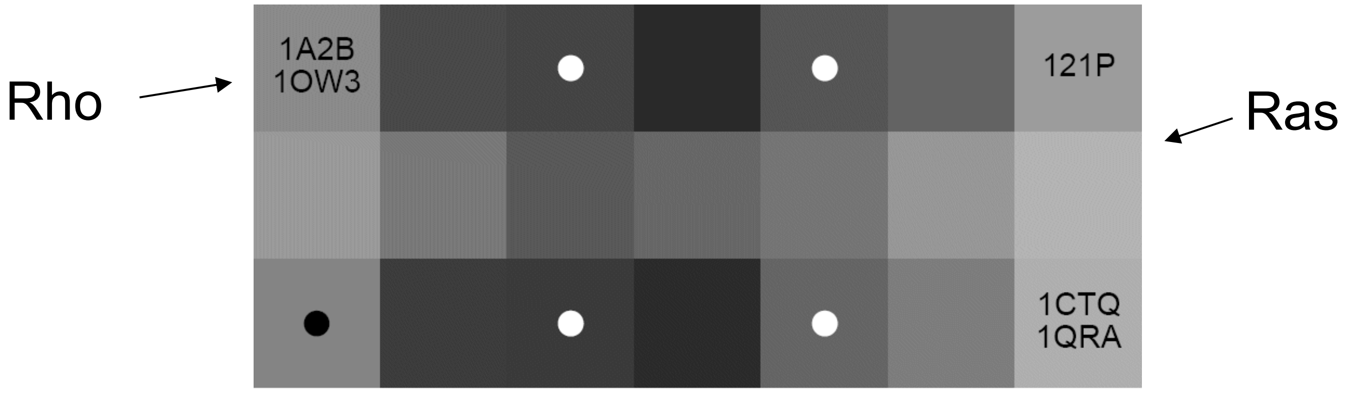


1CTQ



1A2B





Demo

- popsom package in R
  - Very fast implementation
  - Statistically motivated convergence criterion

The End