JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

Ciro Donalek (Caltech)

Clustering: SOM

Jet Propulsion Laboratory

Keck INSTITUTE FOR SPACE STUDIES

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

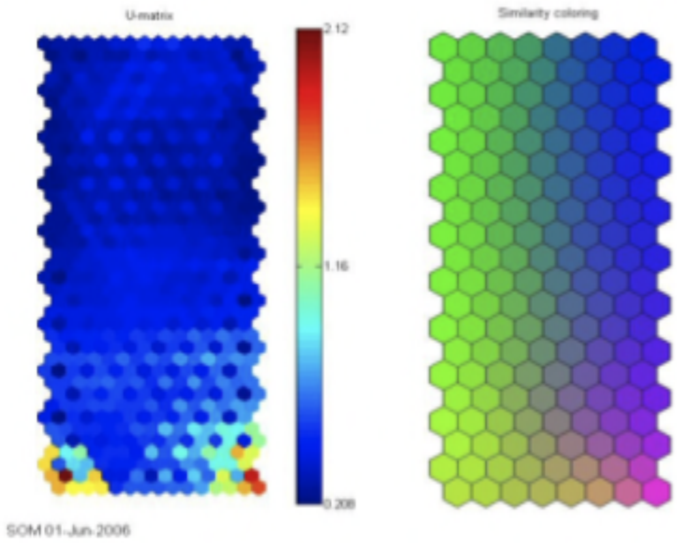
Caltech

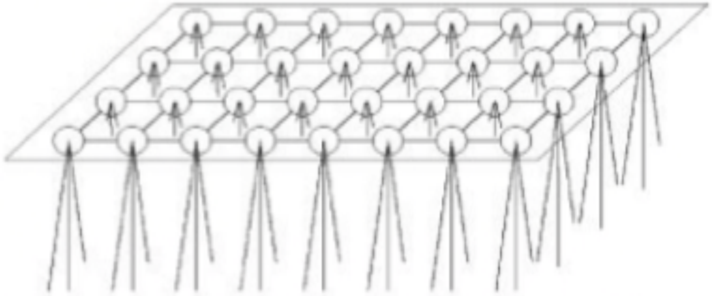
Self-Organizing Maps

Self Organizing Maps learn to recognize groups of similar input vectors in such a way that neurons physically near each other in the neuron layer respond to similar input vectors.

They project high dimensional data into a low dimensional output space.

Used for clustering (or classification), data visualization, modeling, probability density estimation.

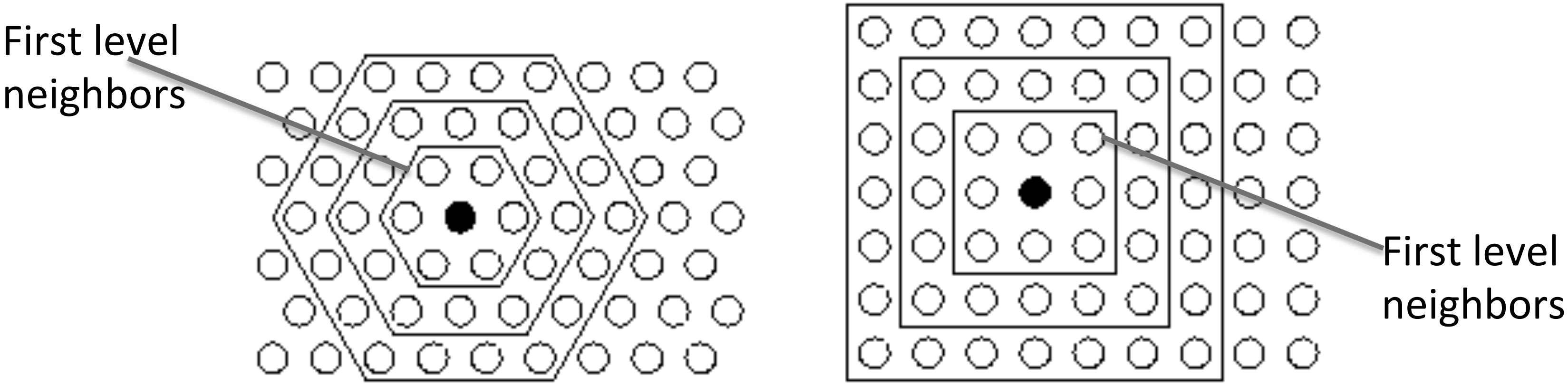




The architecture of a SOM with a two dimensional grid architecture and three inputs fed to all the neurons.

SOM Topology

- A SOM consists of neurons organized on a regular low dimension grid. High dimensional grids are possible but harder to visualize.
- Neurons can be organized on a rectangular or hexagonal lattice.
- The neurons are connected to adjacent neurons by a neighborhood relation dictating the structure of the map.



First level neighbors

First level neighbors

(a) Hexagonal grid

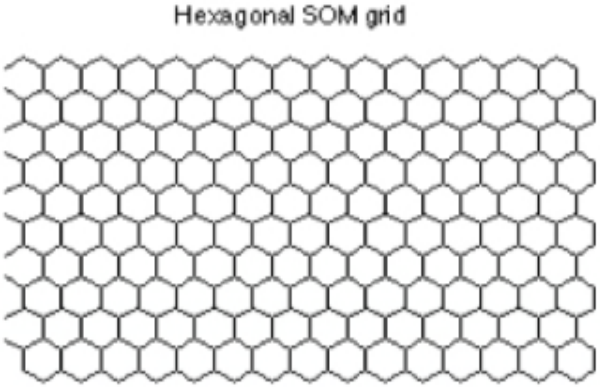
(b) Rectangular grid

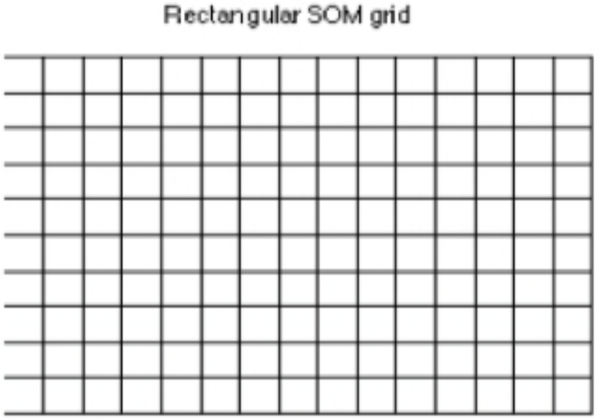
Images credits: Vesanto, University of Helsinki

SOM Prototypes

Each neuron is represented by a d-dimensional weight vector:

$\mathbf{m}_i$=[m_{i1}, ... , m_{id}]   where d=dim(input_vector)





Each map unit can be thought as having two sets of coordinates:
- in the input space: the prototype vectors;
- in the output space: the position on the map.

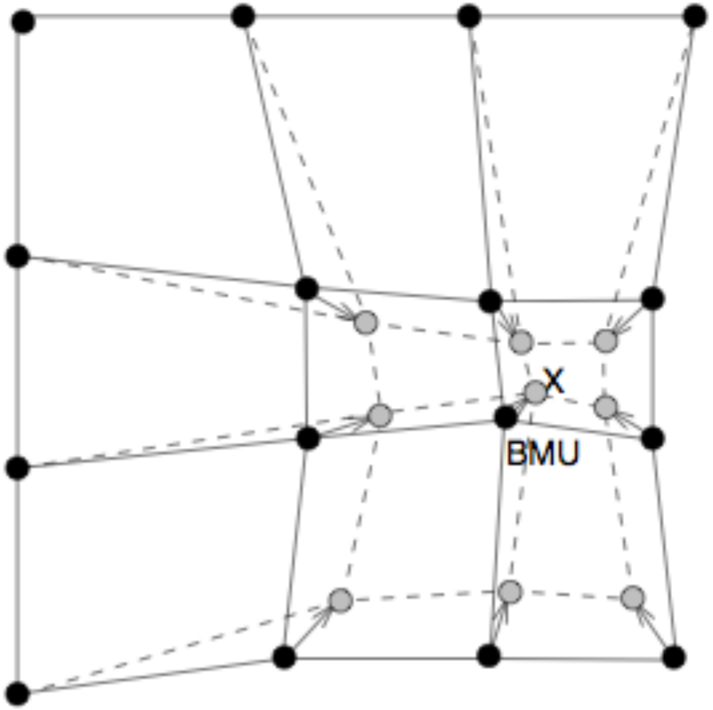
SOM Training

In each training step, one sample **x** from the input data set is chosen; the distances between **x** and all the weight vectors of the SOM are computed;
the neuron whose weight vector is closest to the input vector is called the **Best Matching Unit** (BMU, $\mathbf{m}_c$):

$$\left\| x - m_c \right\| = \min_i \left\| x - m_i \right\|$$

After finding the BMU, the weight vectors are updated so that the BMU is moved closer to the input vector in the input space.

Also the topological neighbors of the BMU are treated similarly.



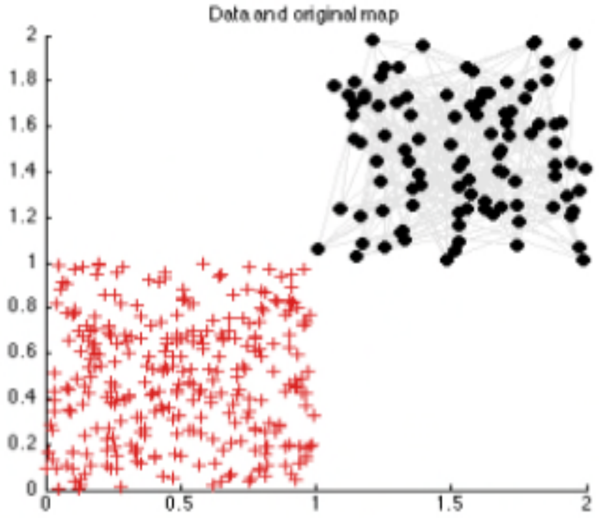
Cooperative and Competitive Learning

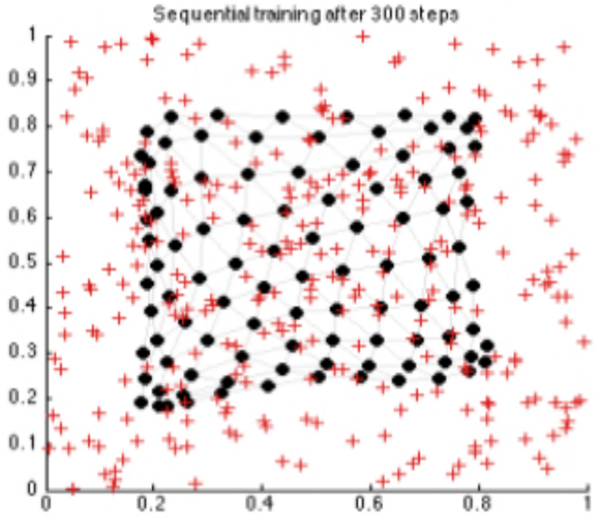
Competitive learning

The prototype vector most similar to a data vector is modified so that it is even more similar to it.

Cooperative learning

Not only the most similar prototype vector, but also its neighbors on the map are moved towards the data vector.





SOM Update Rule

The SOM update rule for the weight vector of unit *i* is:

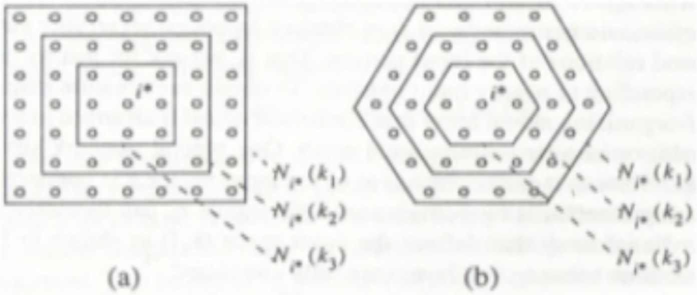
$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$$

where:

- x(t) input vector chosen at time t;
- $h_{ci}(t)$ is the neighborhood function that define the kernel around the winner unit (BMU) c:

$$h_{ci}(t) = \exp(-\|r_c - r_i\|^2 / 2\sigma^2(t))$$

- σ(t) is the neighborhood radius at time t



where $k_1<k_2<k_3$

- α(t) is the learning rate at time t

  linear .α(t)= α_0(1-t/T) where α_0=initial learning rate; T=training length

  inversely proportional to time α(t)= A/(t+B) with A,B suitable constants

Parameters

- Map size and Topology
  - Default: 5*sqrt(n) where n = # of samples
- Weights initialization
  - random: with small random values
  - sample: with random samples drawn from the input
  - linear: along the linear subspace spanned by the two principal eigenvectors of the data set
- Batch and Sequential training
  - sequential: sample are presented to the map one a time and the algorithm gradually moves the weight vectors toward them
  - batch: the data set is presented as a whole and the new weight vectors are weighted averages of the data vectors
- Learning rate, neighborhood function, radius.

Data Mining using SOMs

How to find clusters?

Which components are the most important in discriminating among the classes?

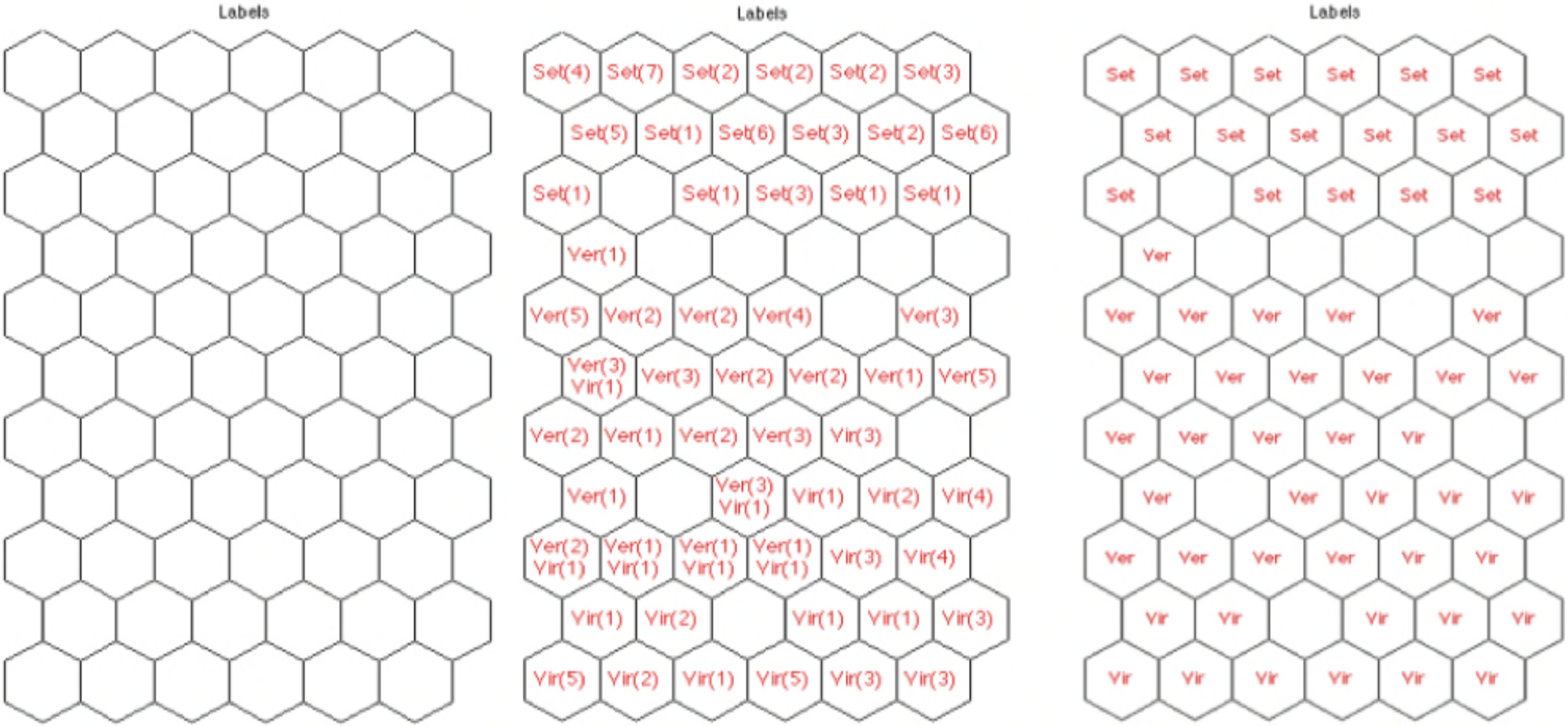
How do the parameters relate to the clusters?

With the SOM we have several ways to visually (and not only) answer to the these question.

“Cool” visualizations looks good in the papers too!

SOM Labeling

The BMU of each sample is found from the map, and the species label can be given to each map unit.



Labels

Labels

Set(4)	Set(7)	Set(2)	Set(2)	Set(2)	Set(3)
Set(5)	Set(1)	Set(6)	Set(3)	Set(2)	Set(6)
Set(1)		Set(1)	Set(3)	Set(1)	Set(1)
Ver(1)					
Ver(5)	Ver(2)	Ver(2)	Ver(4)		Ver(3)

Ver(3)
Vir(1)
	Ver(3)	Ver(2)	Ver(2)	Ver(1)	Ver(5)

Ver(2)	Ver(1)	Ver(2)	Ver(3)	Vir(3)	

Ver(1)		Ver(3)
		Vir(1)
			Vir(1)	Vir(2)	Vir(4)

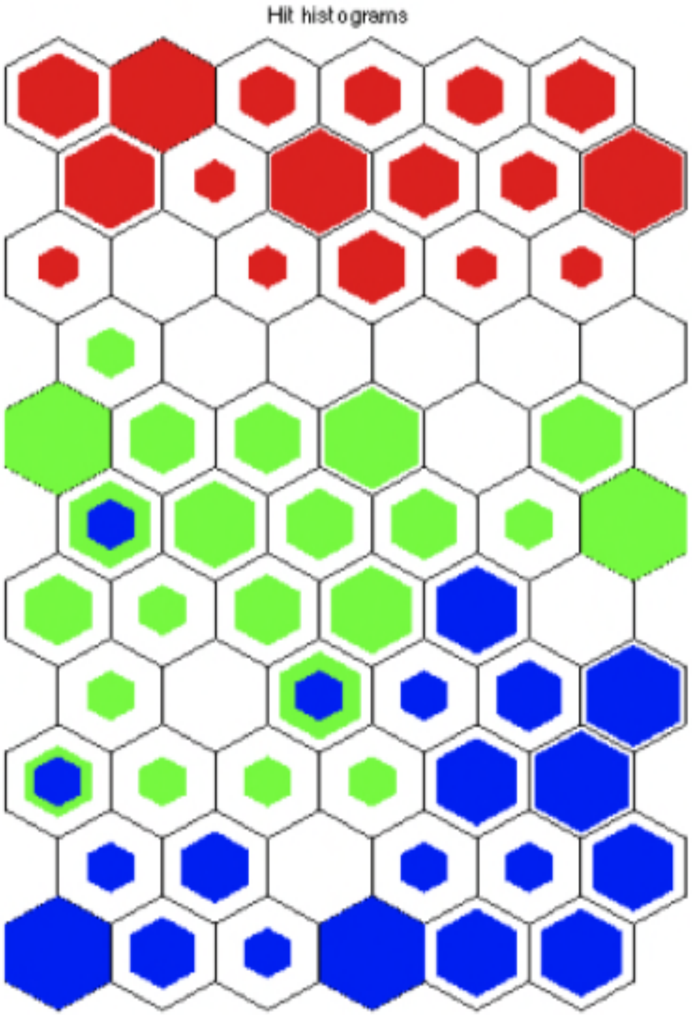
Ver(2)
Vir(1)
	Ver(1)
	Vir(1)
		Ver(1)
		Vir(1)
			Ver(1)
			Vir(1)
				Vir(3)	Vir(4)

Vir(1)	Vir(2)		Vir(1)	Vir(1)	Vir(3)
Vir(5)	Vir(2)	Vir(1)	Vir(5)	Vir(3)	Vir(3)

Labels

Set	Set	Set	Set	Set	Set
Set	Set	Set	Set	Set	Set
Set		Set	Set	Set	Set
Ver					
Ver	Ver	Ver	Ver		Ver
Ver	Ver	Ver	Ver	Ver	Ver
Ver	Ver	Ver	Ver	Vir	
Ver		Ver	Vir	Vir	Vir
Ver	Ver	Ver	Ver	Vir	Vir
Vir	Vir		Vir	Vir	Vir
Vir	Vir	Vir	Vir	Vir	Vir

Localizing Data

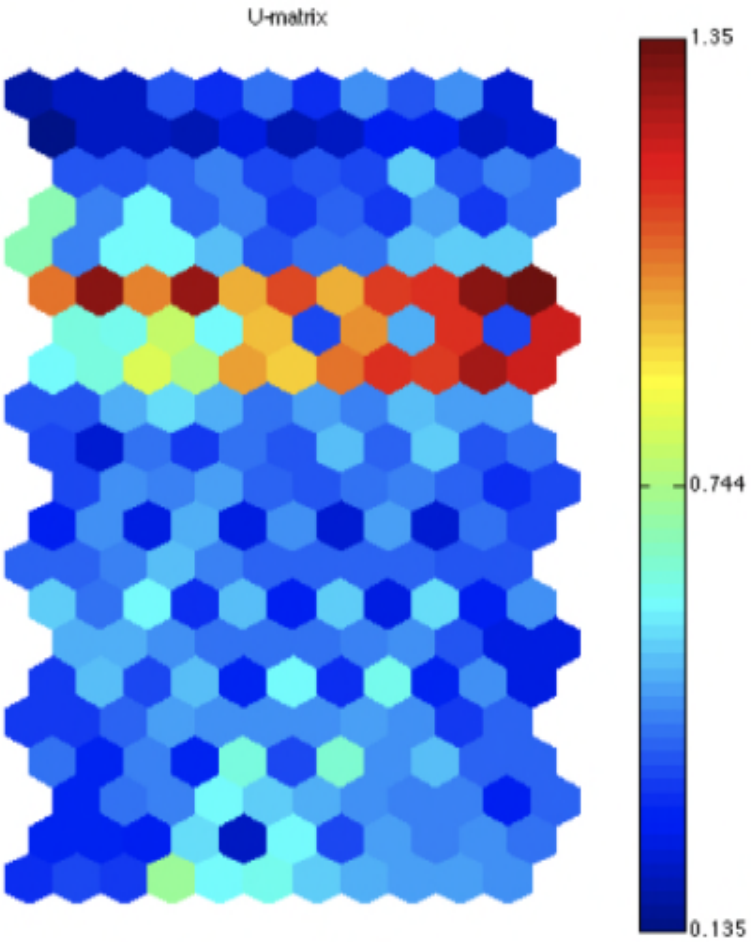


An important tool in data analysis using SOM are the so called **hit histogram**.

The hit histogram shows the distribution of the data set on the map.

They are formed by taking a data set, finding the BMU of each data sample from the map, and increasing a counter in a map unit each time it is the BMU.

Cluster Structure



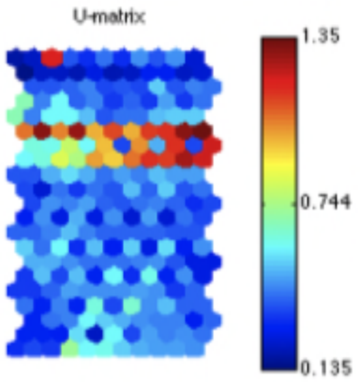
The **U-matrix** shows distances between neighboring units and thus visualizes the cluster structure of the map.

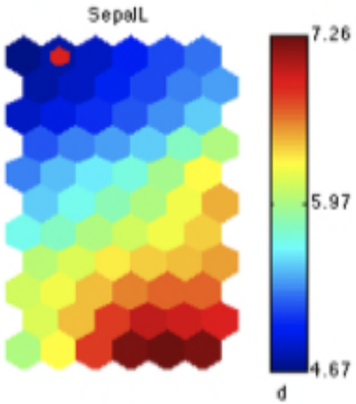
The U-matrix visualization has much more hexagons that the “real” map because distances between map units are shown.

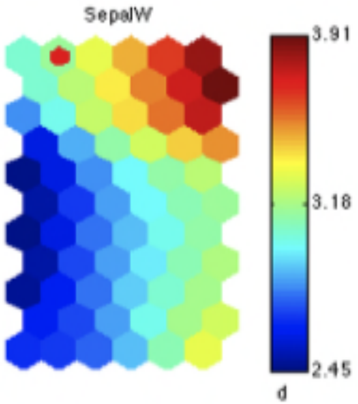
High values on the U-matrix mean large distance between neighboring map units.

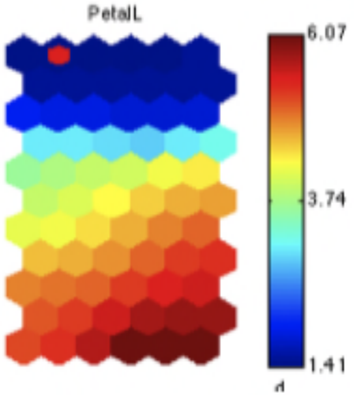
If we have five units 5x1: m(1), m(2), m(3), m(4), m(5)
U-Matrix is a 9x1 vector:
u(1), u(1,2), u(2), u(2,3), u(3), u(3,4), u(4), u(4,5), u(5)
Where:
u(i,j) is the distance between m(i) and m(j);
u(k) is the mean: u(k)=(u(k-1,k)+u(k,k+1))/2.

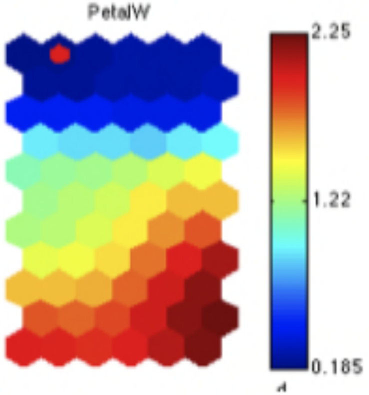
Component Planes









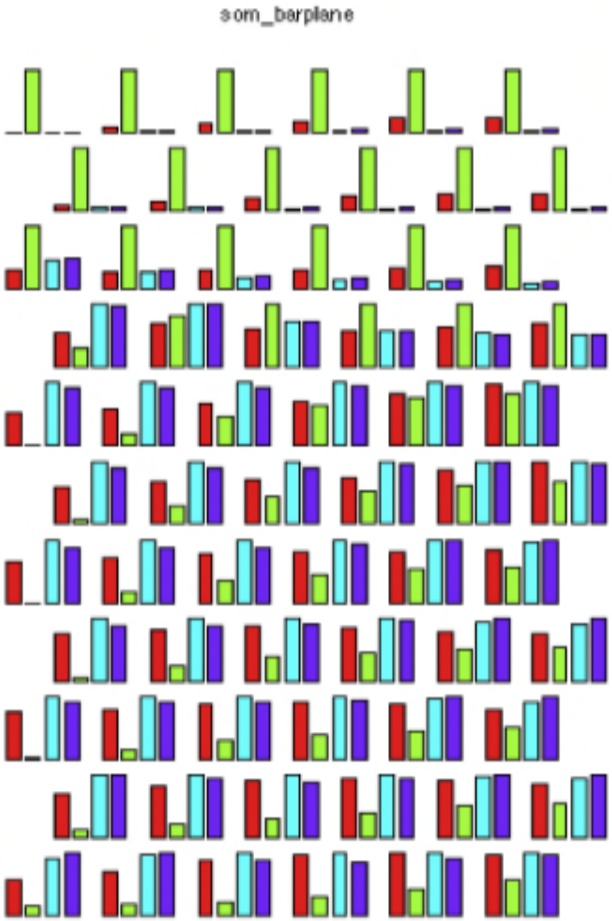


Visual inspection:

- PetalL and PetalW are highly correlated;
- SepalL is ~ correlated to PetalL and PetalW;
- Setosa has small petals and short wide sepals;
- the 2nd cluster is Virginica /Versicolor;
- Versicolor has smaller leaves than Virginica.

The **component planes** show the values of the prototype vectors for each parameter.
Can be used for correlation hunting.

Relative Importance



Which components are the most important in discerning among the classes?

The significance of the components with respect to the clustering is usually harder to visualize.

One indication of importance is that on the borders of the clusters, values of important variables change very rapidly.

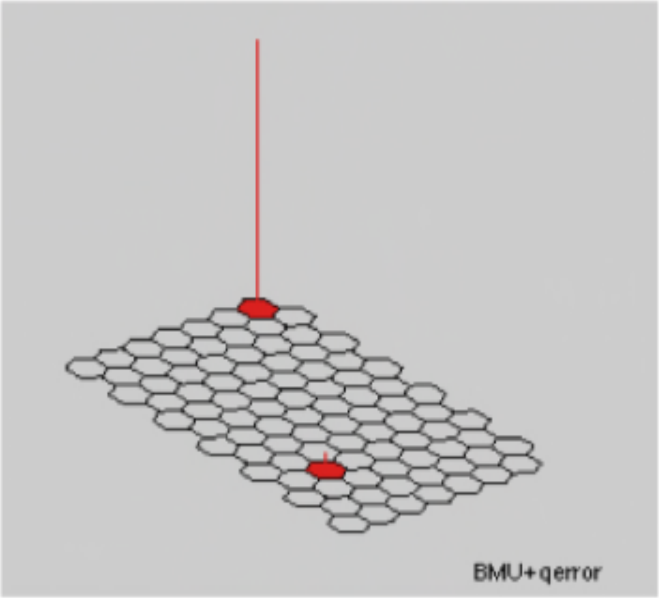
Each chart shows the relative importance of each variable in each map unit.

How accurate is your Clustering

We have seen how to locate a sample on the map. But how accurate is that localization?

We can compute and plot how much a data sample is far from its BMU, its 2nd-BMU and WMU and so on.

Errors are also computed and are useful in determining the quality of the clustering.

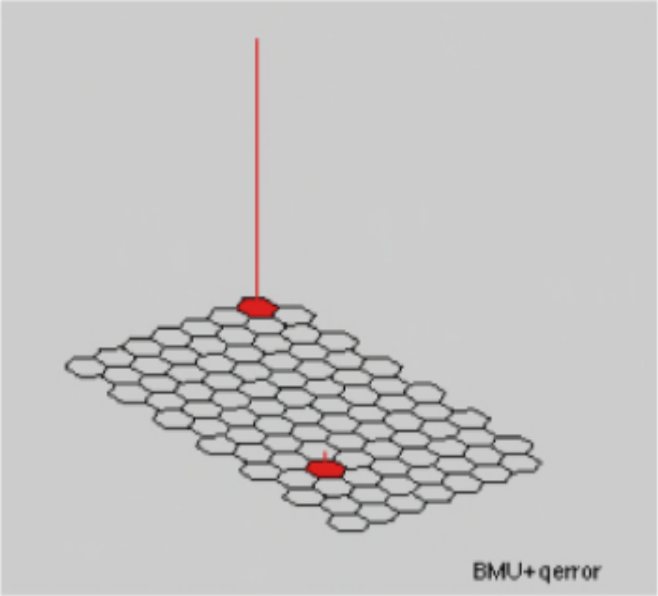


How accurate is your Clustering

We have seen how to locate a sample on the map. But how accurate is that localization?

We can compute and plot how much a data sample is far from its BMU, its 2nd-BMU and WMU and so on.

Errors are also computed and are useful in determining the quality of the clustering.

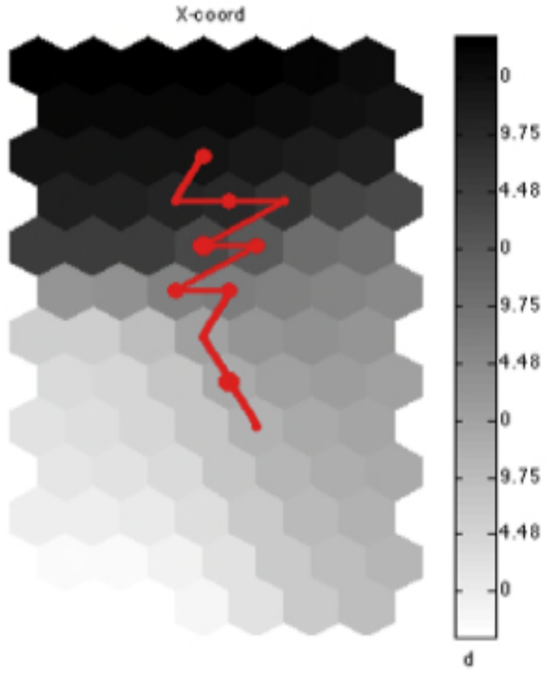


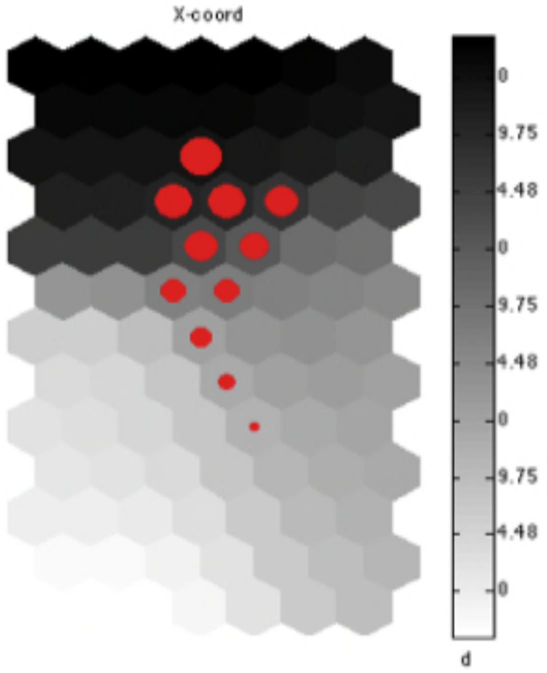
Average quantization error: measures the distance from each data vector to its BMU.

Topographic error measure: percentage of data vectors for which the BMU and the second-BMU are not adjacent units.

Trajectories

A special data mapping technique is trajectory. If the samples are ordered, forming a time-series for example, their response on the map can be tracked.





Summary

- Self-Organizing Maps
- Find clusters
- Locate new objects
- Find the most discriminative parameters
- Accuracy
- Trajectories