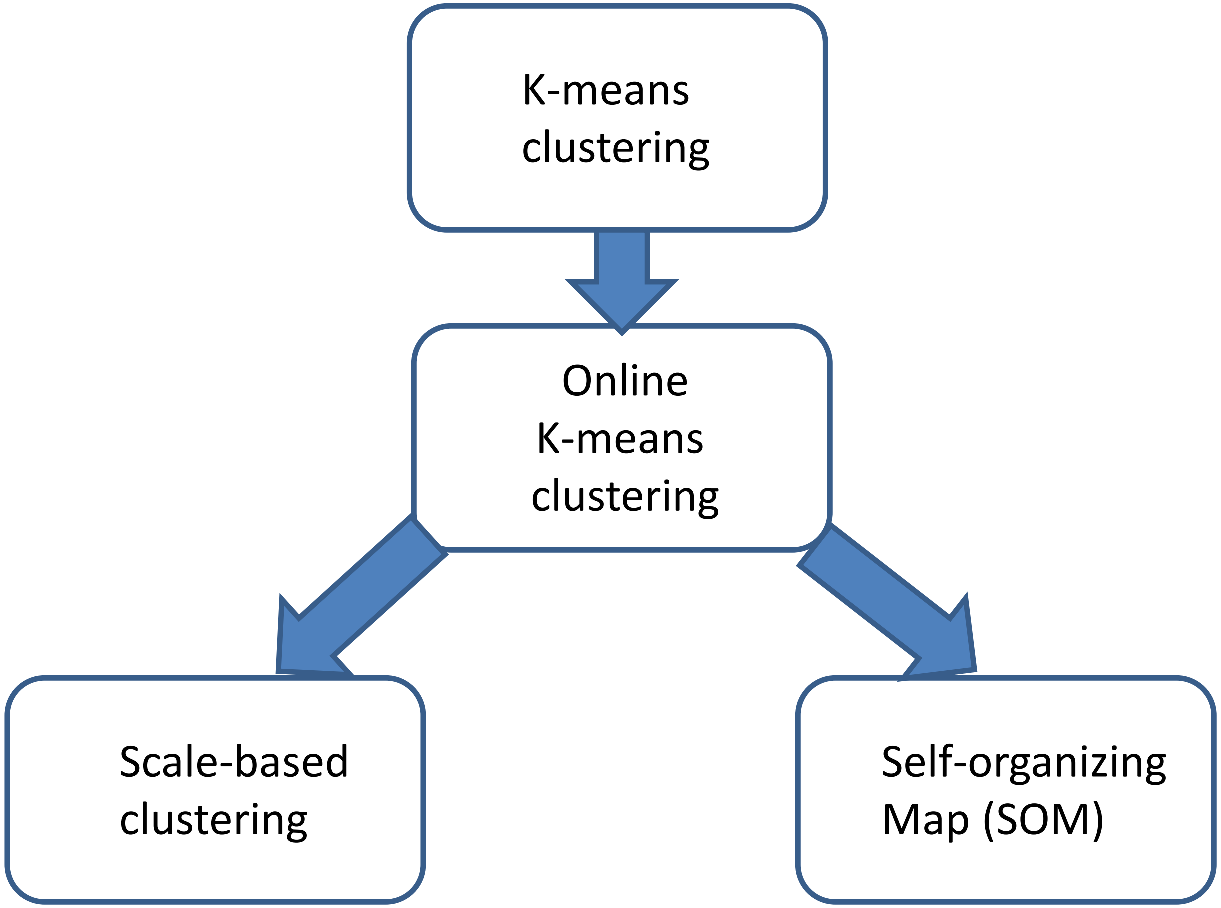
# ONLINE CLUSTERING, SCALE-BASED CLUSTERING AND SELF-ORGANIZING MAP



K-means clustering

Online K-means clustering

Scale-based clustering

Self-organizing Map (SOM)

# Online clustering

$$E = \sum_{j=1}^{N} \left\| (m_{i(j)} - x_j) \right\|^2$$

- $m_{i(j)}$ is the centroid closest to $x_j$

- E is SSE

Finding $m_i$ by gradient descent:

$$\Delta m_i \propto -\nabla_{m_i} E$$

$$\Delta m_i = -\eta \nabla_{m_i} E$$

$$\Delta m_{i(j)} = \eta (x_j - m_{i(j)})$$

# Scale-based Clustering

- Clustering is done at a “scale”
- An answer to the question of “how many clusters”
- Best clusters tend to live over the longest range of scales

# Algorithm

- Start with a large number of clusters
- Initialize by selecting from data set
- Initialize “sigma” to a small value
- Update all centroids
- Eliminate duplicate centroids whenever there is a merger
- Increase sigma by a constant factor
- If there are more than 1 unique centroid continue update of centroids
- Stop only when a single unique centroid remains

# Data set

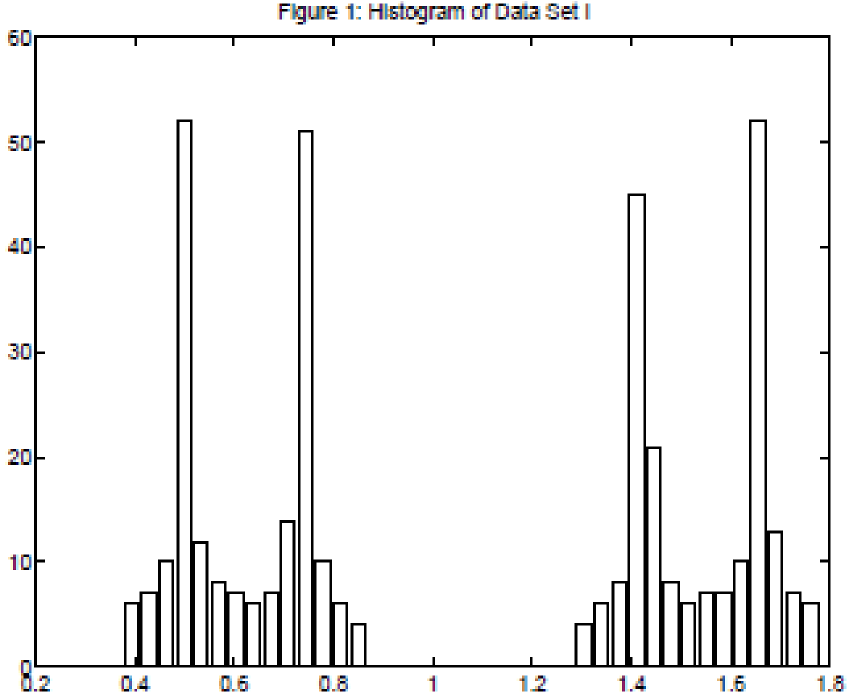


Figure 1: Histogram of Data Set I with 400 pts.

# Clustering result: Evolution of the centroids

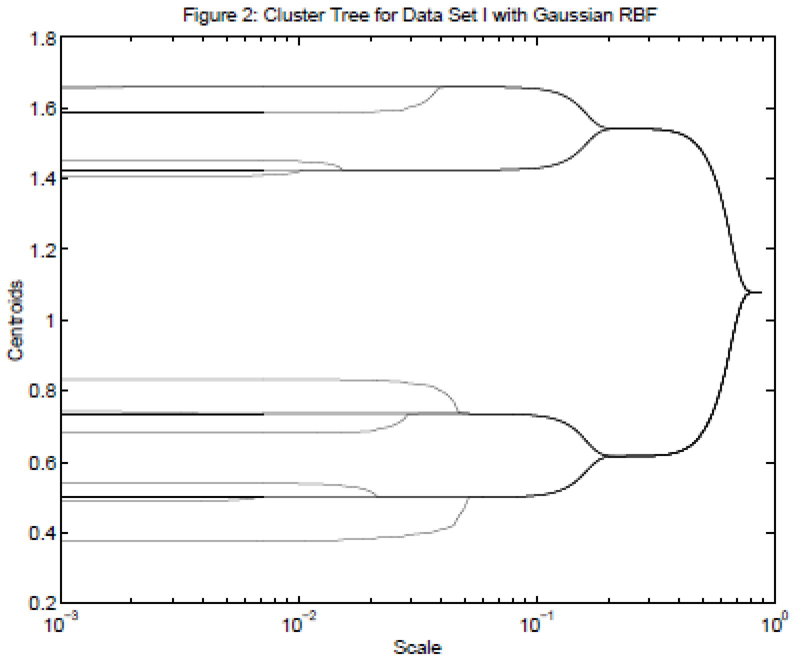
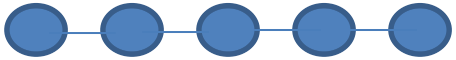


Figure 2: Cluster Tree for Data Set I with 14 RBF nodes. Only 13 branches seem to be present even at the lowest scale because the topmost "branch" is actually two branches which merge at $\sigma = 0.002$.

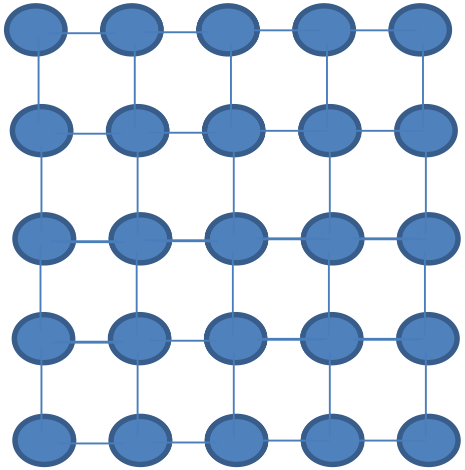
# THE SELF-ORGANIZING MAP

# SOM topology

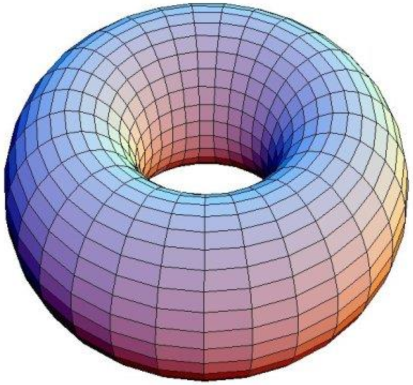


1D SOM

2D SOM

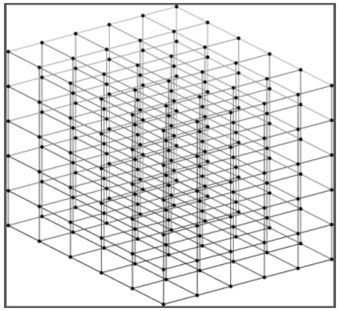


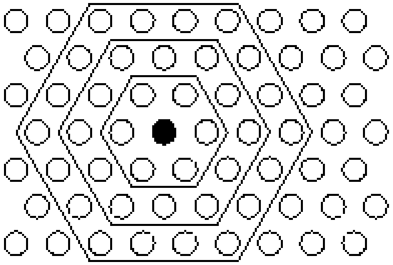
(rectangular)



Toroid SOM

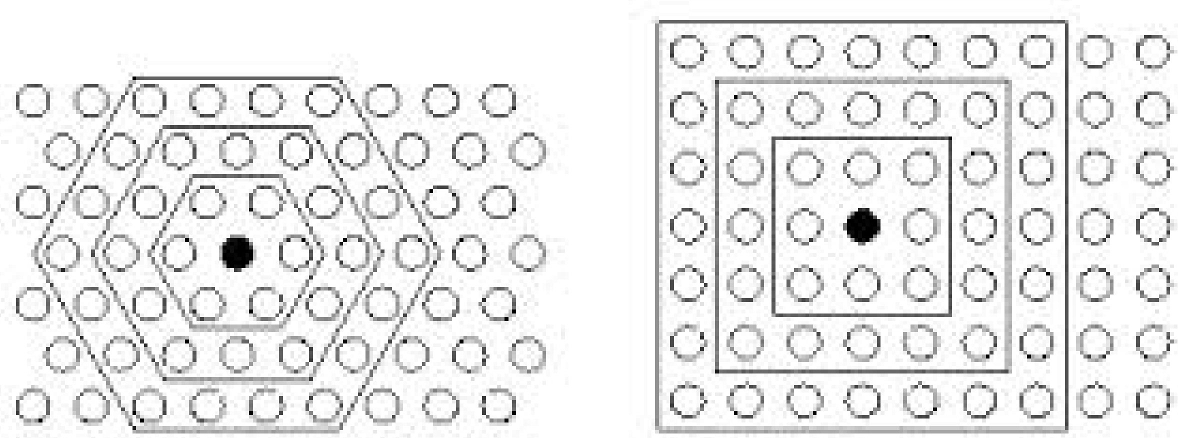
3D SOM





(hexagonal)

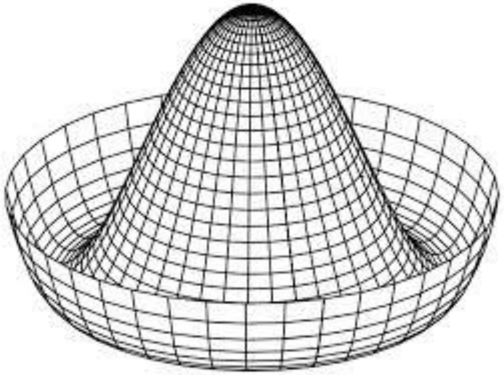
# Neighborhood



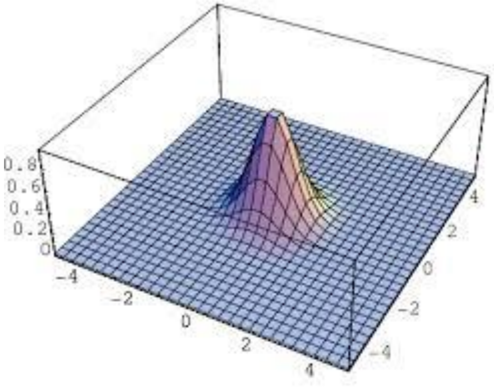
(a) Hexagonal grid

(b) Rectangular grid

# Neighborhood functions ($\Lambda(r^*, r)$)



Mexican Hat function



Gaussian function

# SOM algorithm

- Randomly initialize the weights from the training data set, X
- Begin Loop
  - Present xp to all neurons and find the Winner (r*)
  - Update the weights of the winner and its neighbors
- End Loop (when weights converge)

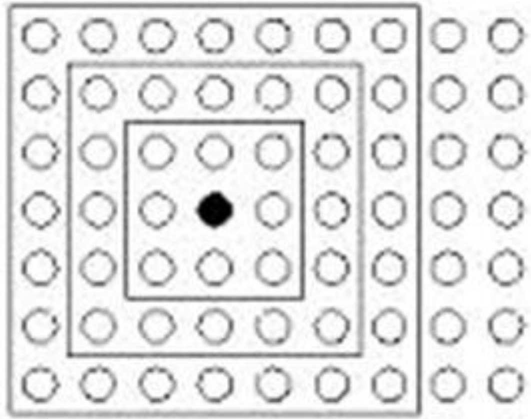
$$\Delta w_{r*} = \eta(x_p - w_{r*})$$

Move the “winner” towards xp

$$For \;\; r \in N$$

$$\Delta w_r = \eta\Lambda(r, r*)(x_p - w_r)$$

Move the neighbors of the “winner” also towards xp, but to a lesser extent



(b) Rectangular grid

Neighborhood, ℕ

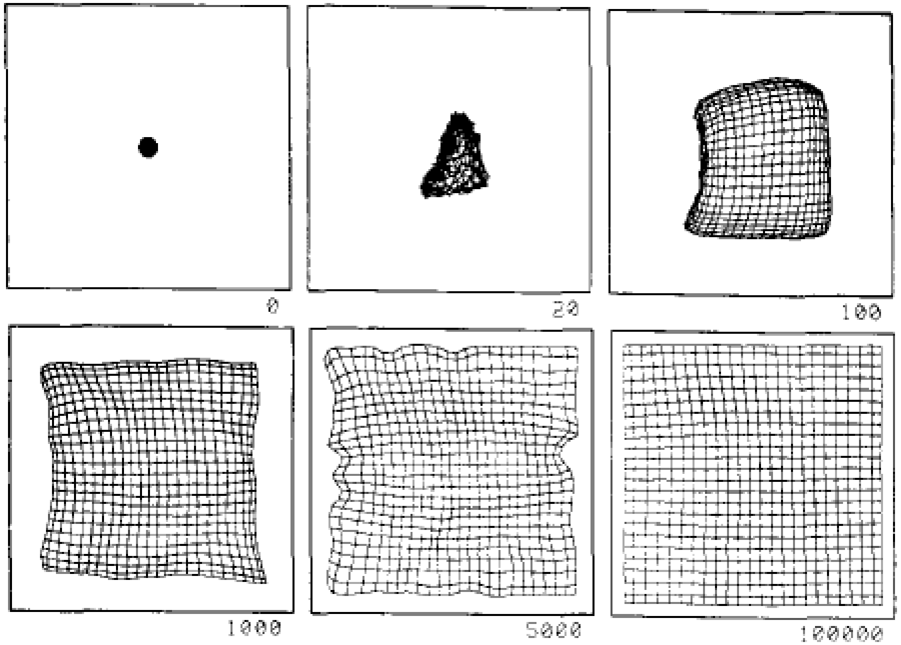
# Annealing

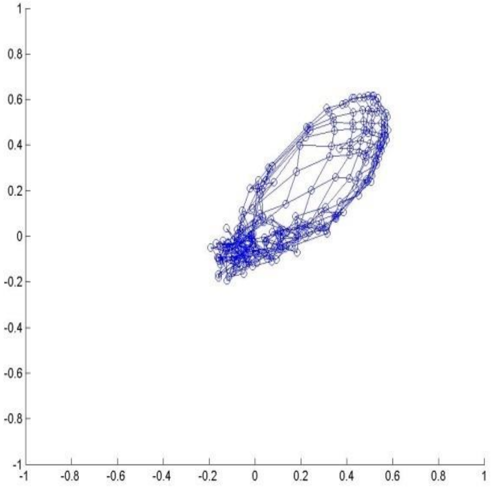
- as training proceeds reduce
  - neighborhood size
  - Learning rate

# Learning stages

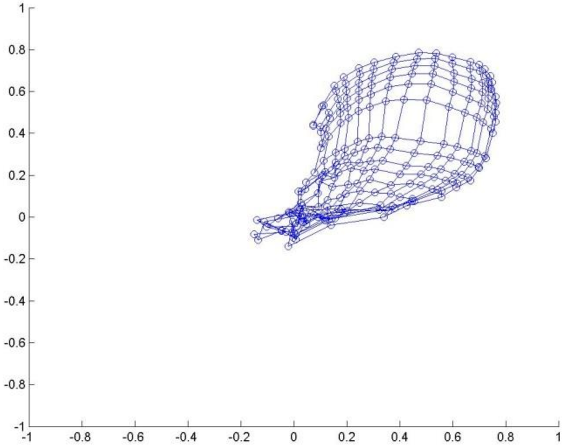
- Ordering phase
- Settling phase:

# Ordering and Settling

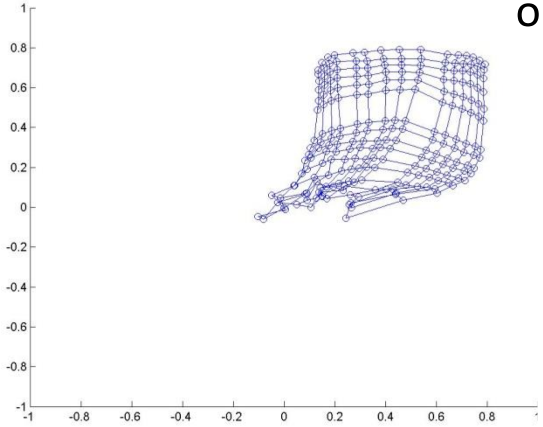




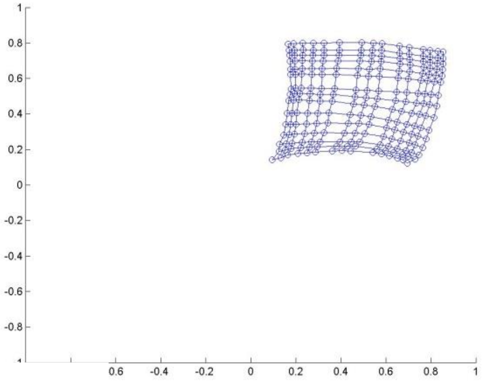
ordering



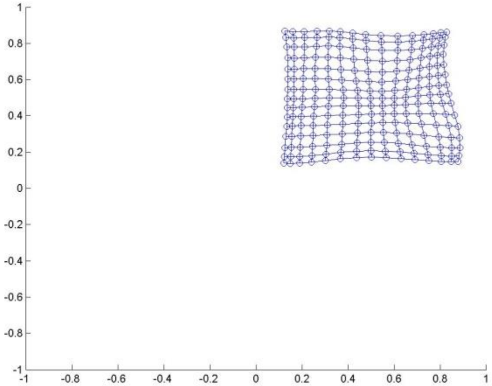
ordering



ordering

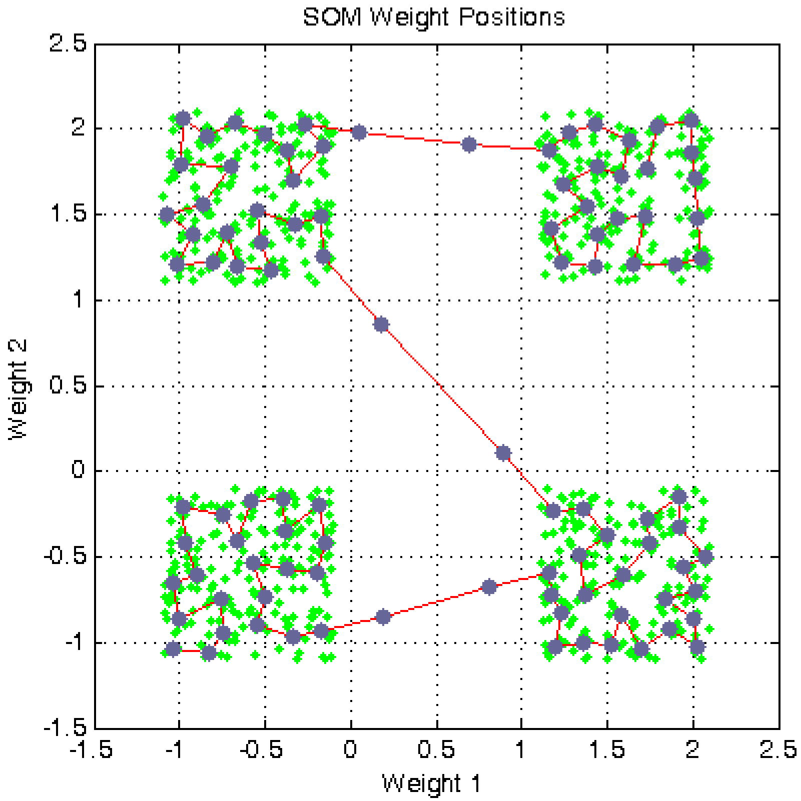


settling



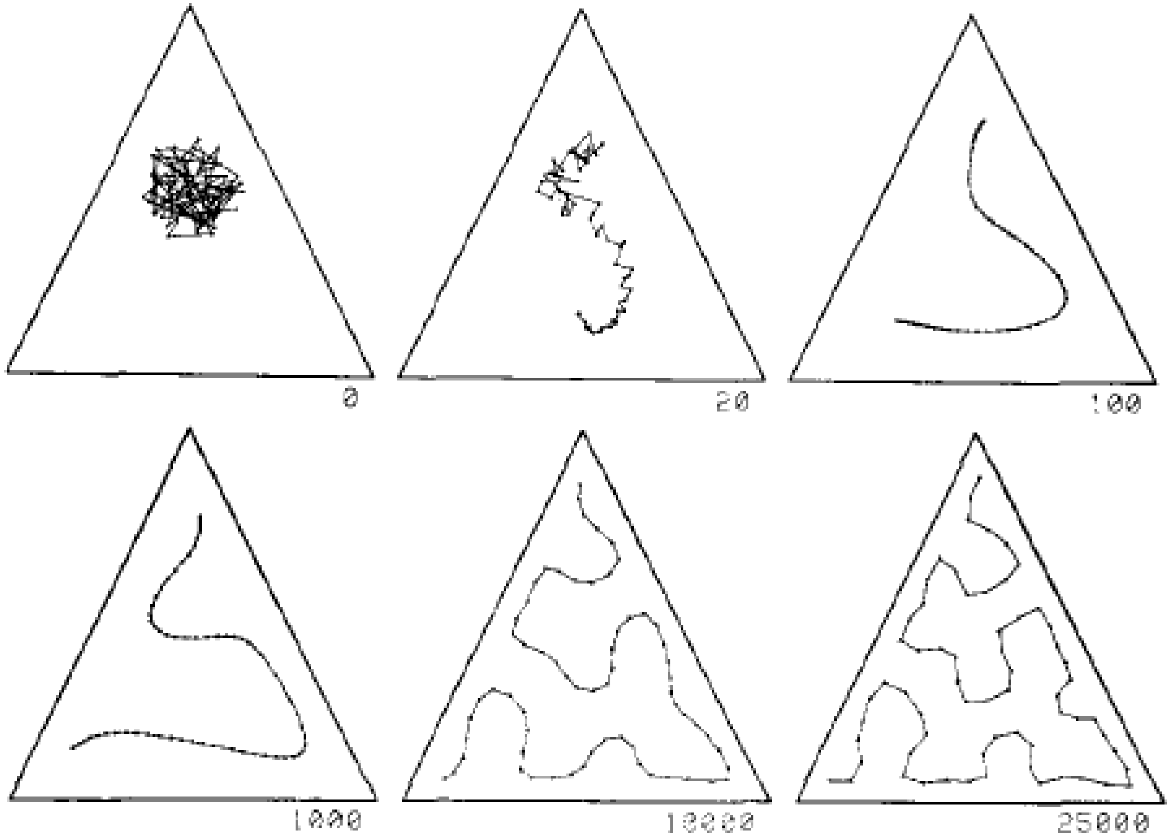
settling

# Simple examples (2D data; 1D SOM)



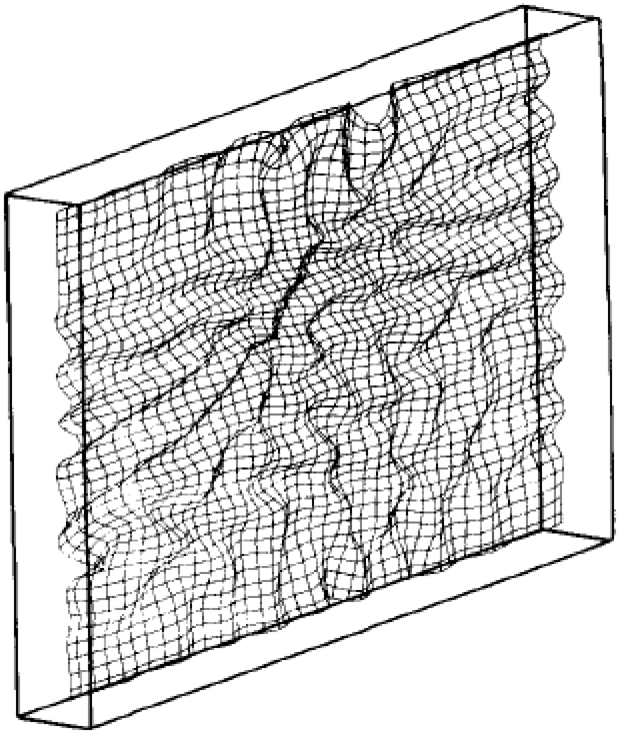
# 2D data; 1D SOM

Peano's curve
Or
Space-filling
curve

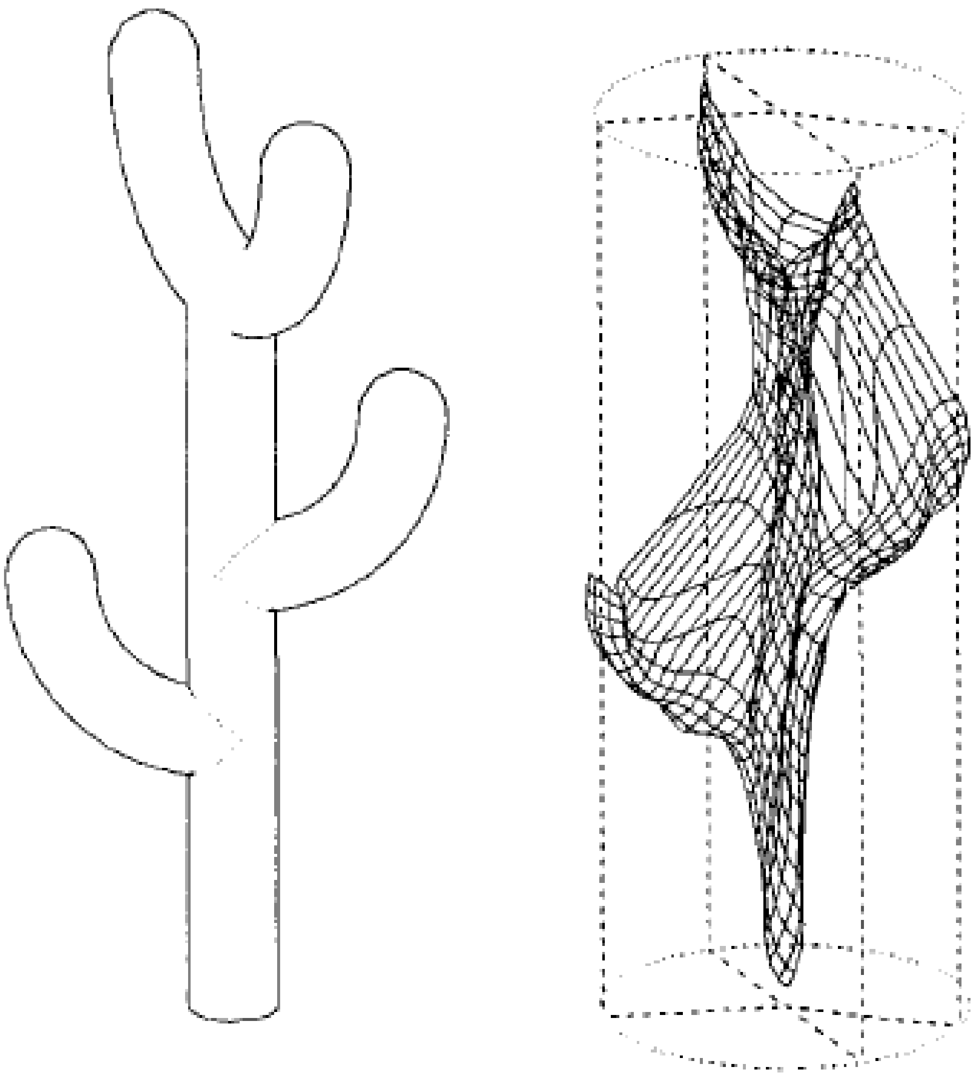


(Kohonen 1990)

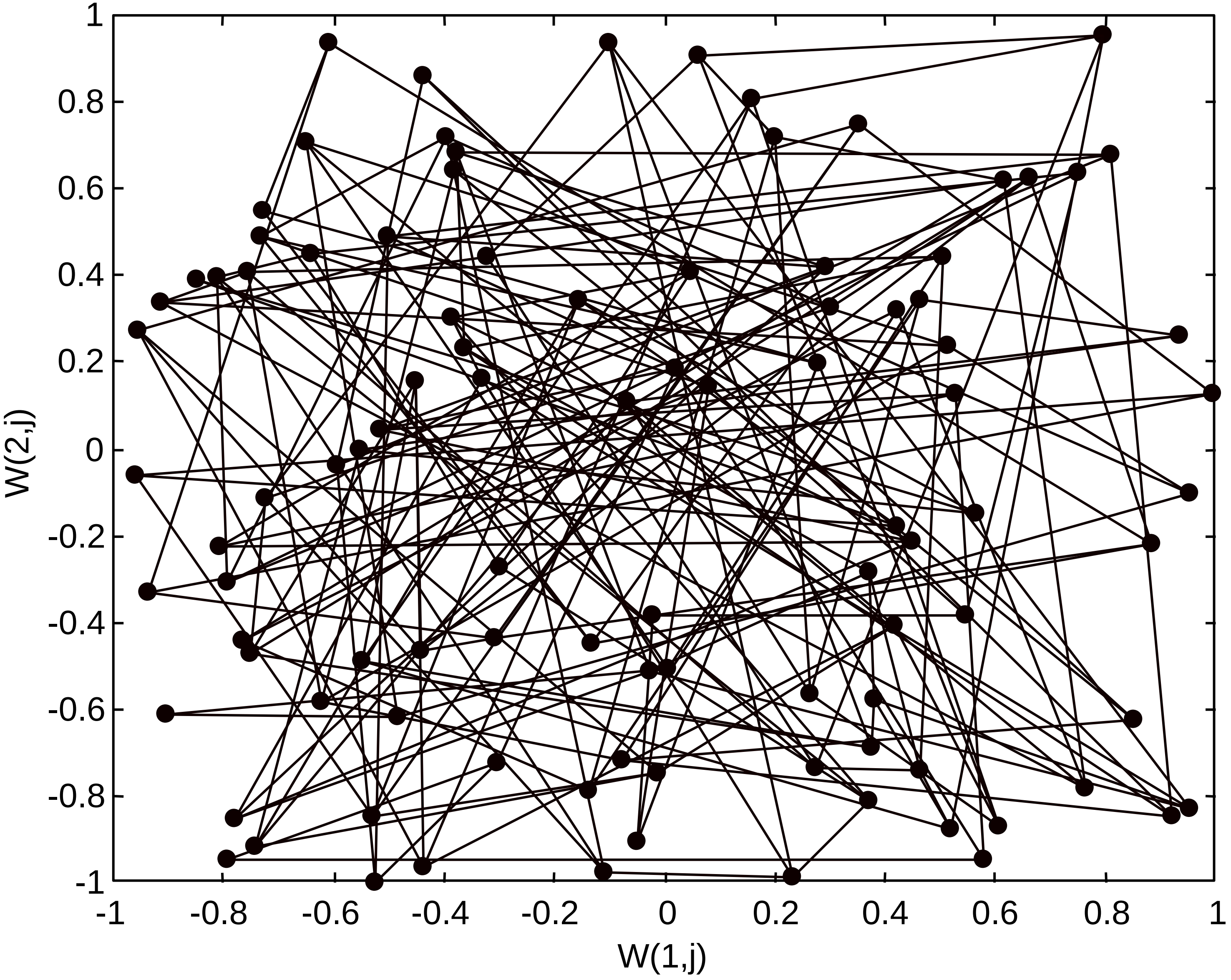
# 3D data; 2D SOM



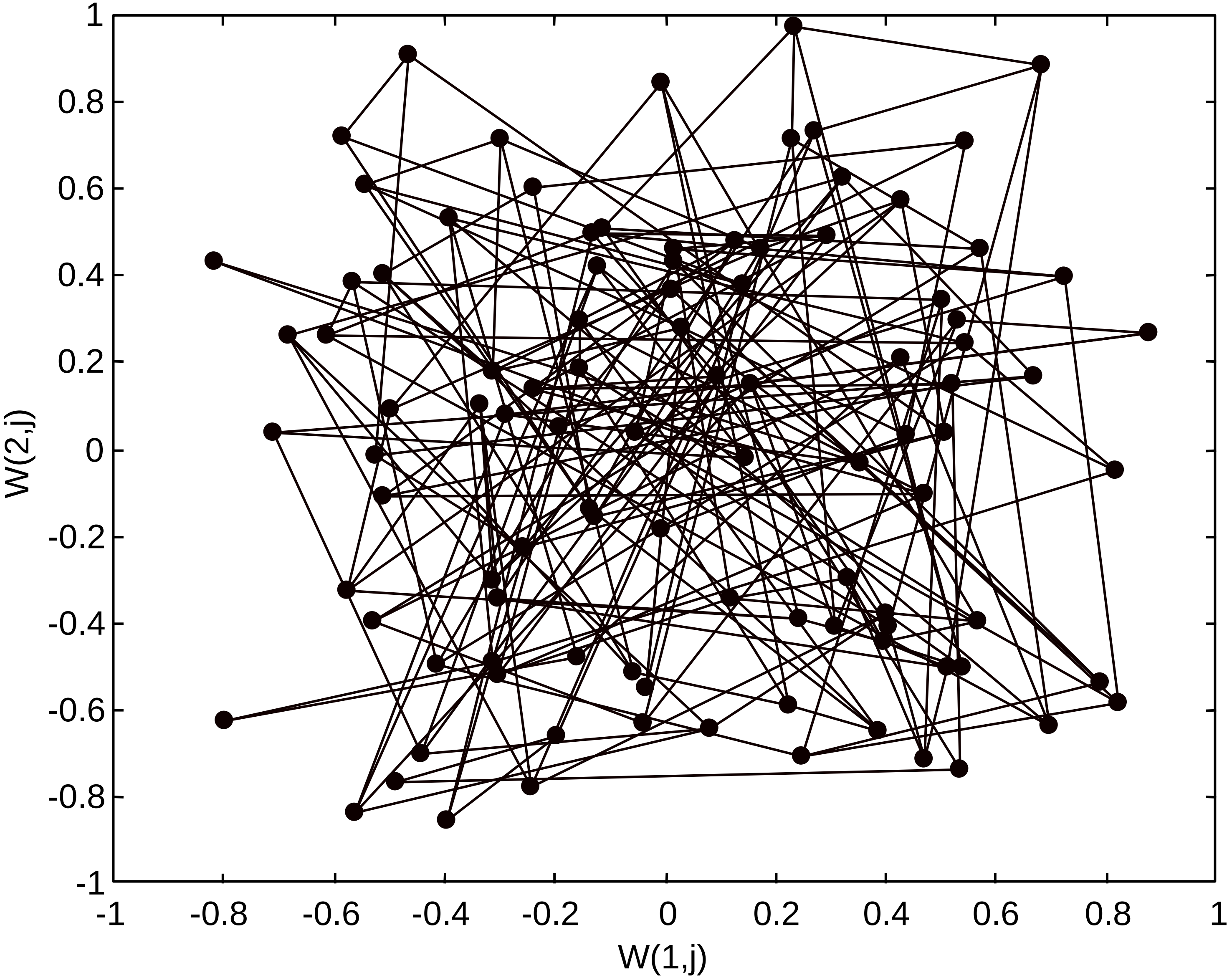
# 3D data; 2D SOM



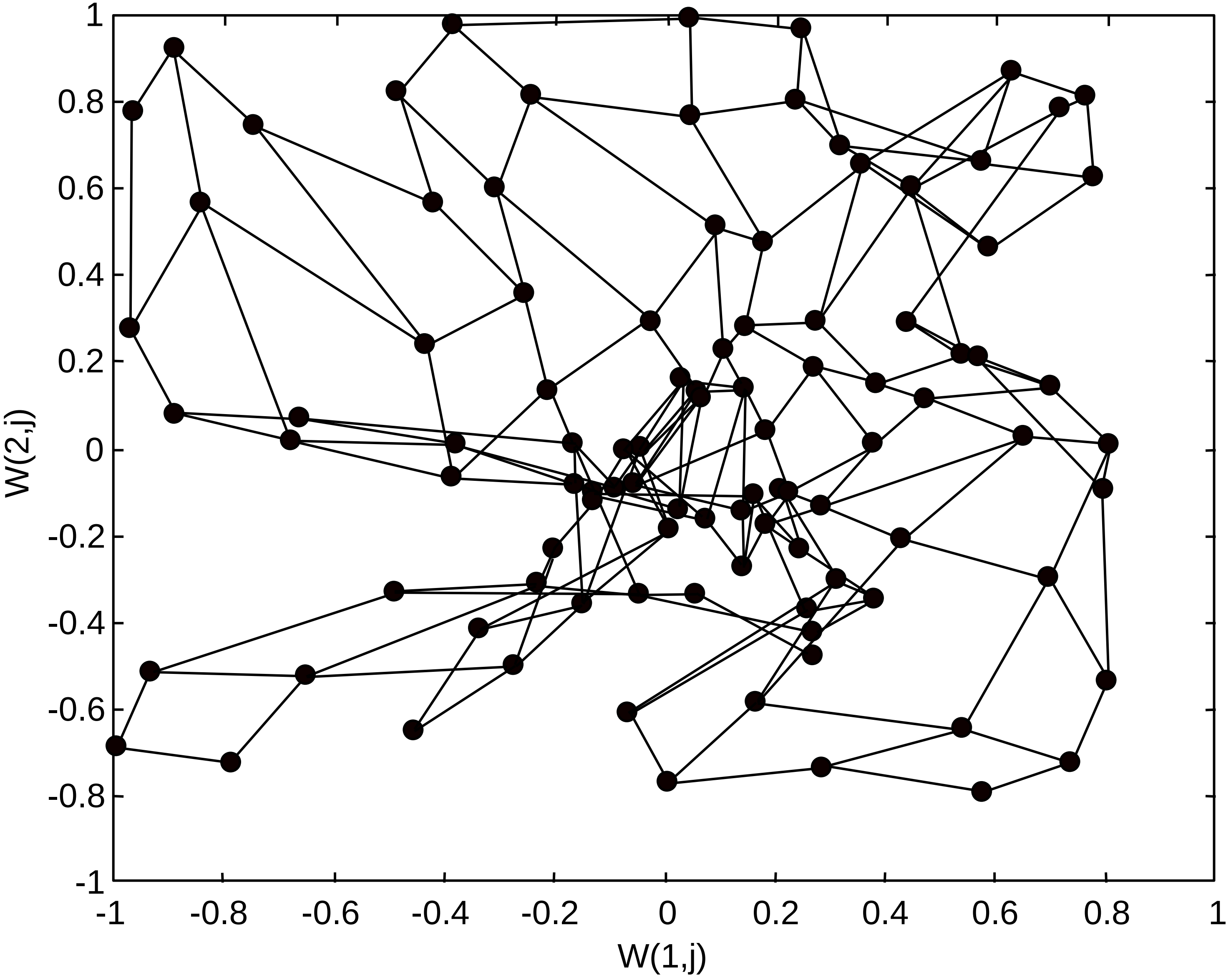
# Initial random weights



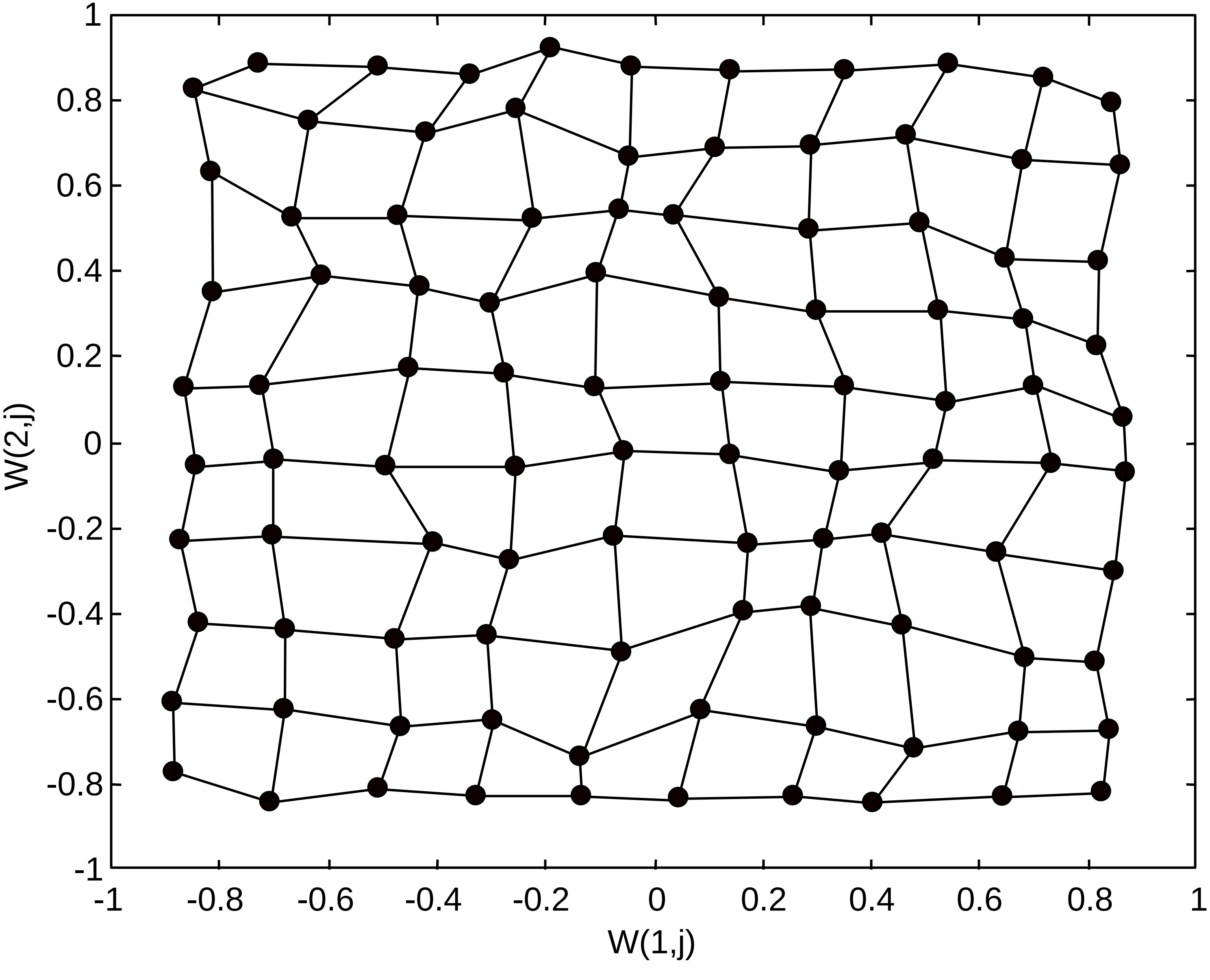
# Network after 100 iterations



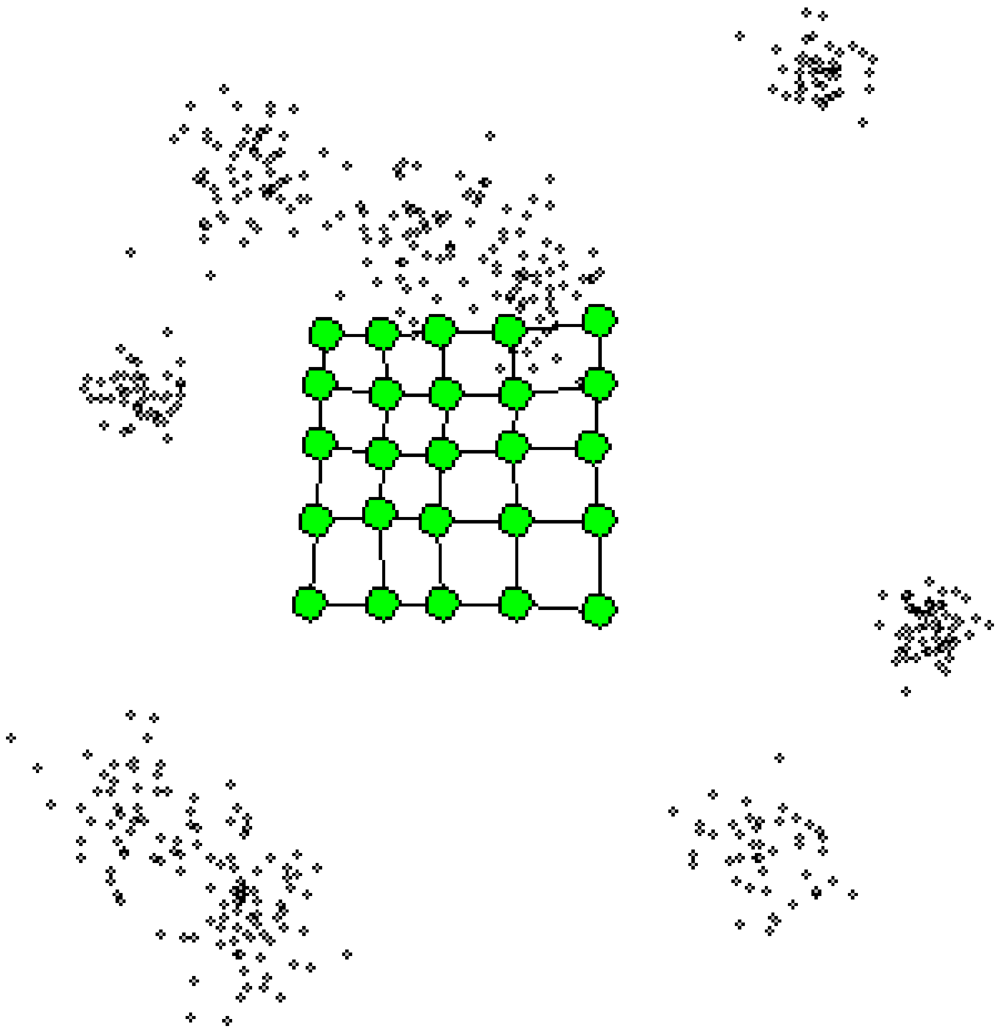
# Network after 1000 iterations



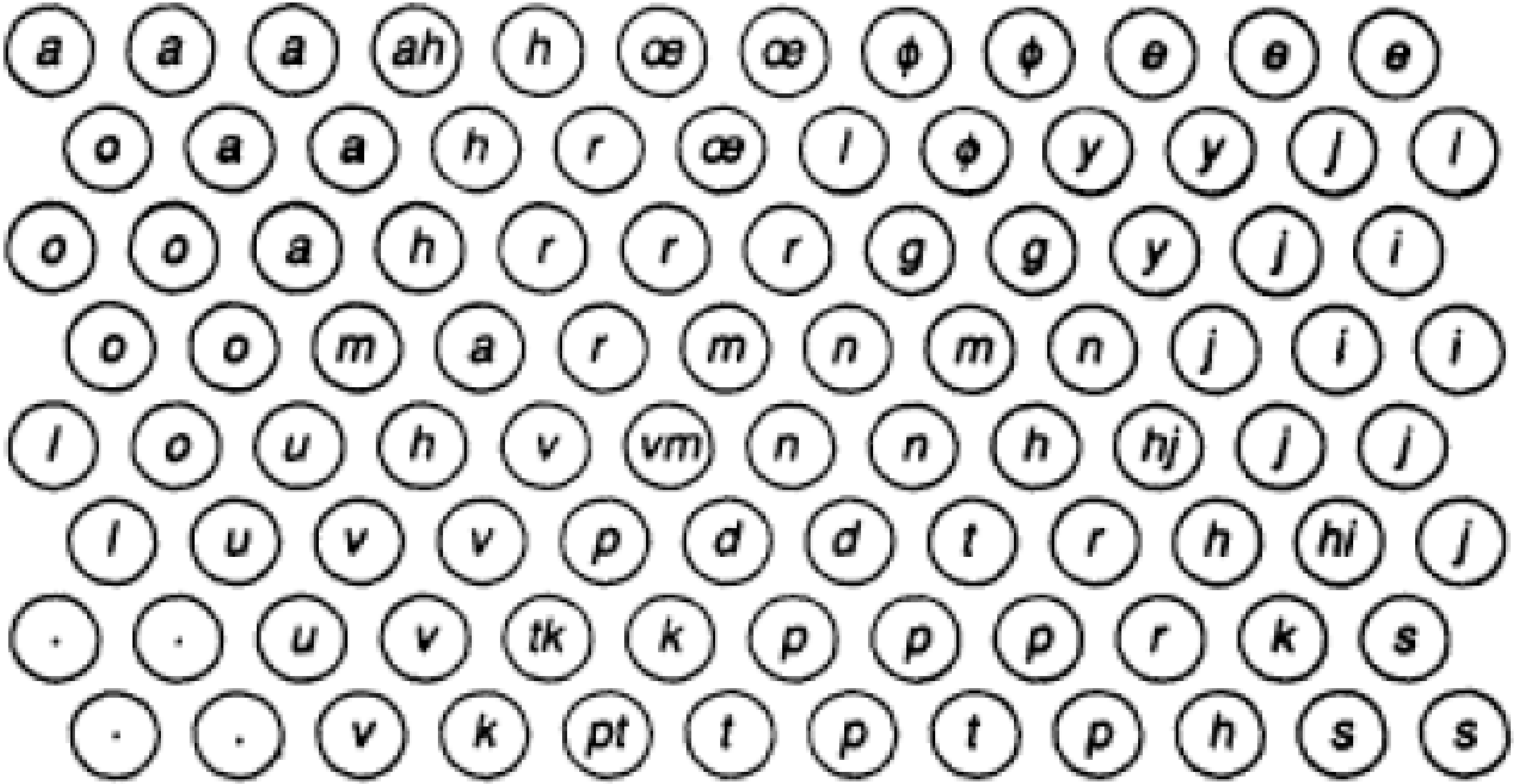
# Network after 10,000 iterations



# Another example



# Phonetic Keyboard



a a a ah h œ œ ϕ ϕ e e e

o a a h r œ l ϕ y y j l

o o a h r r r g g y j i

o o m a r m n m n j i i

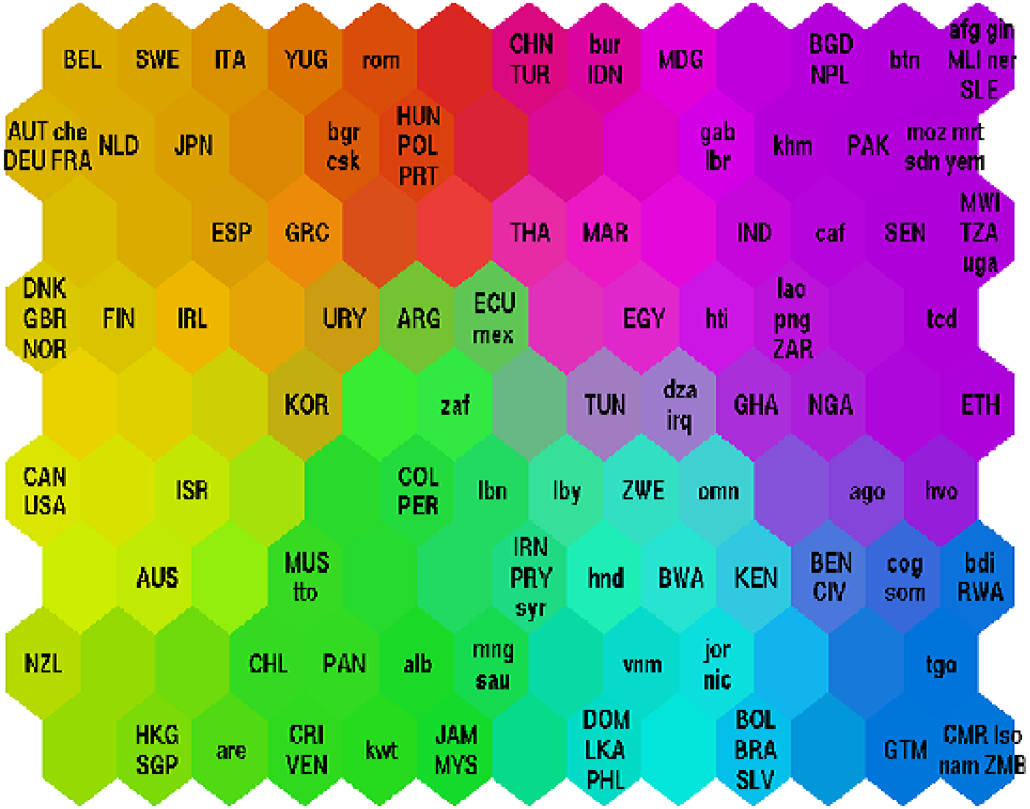
l o u h v vm n n h hj j j

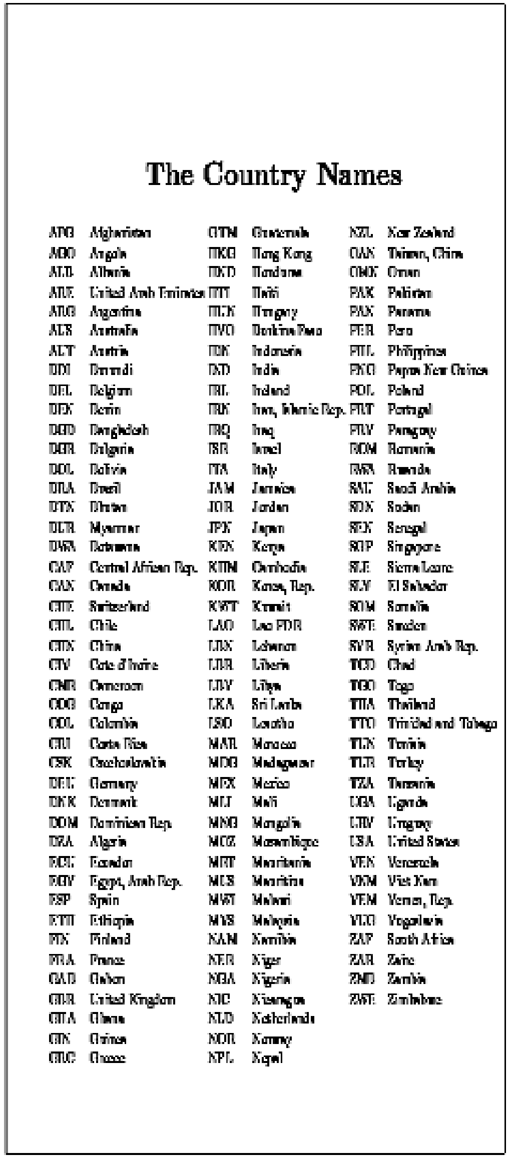
l u v v p d d t r h hi j

· · u v tk k p p p r k s

· · v k pt t p t p h s s

# Mapping World Poverty

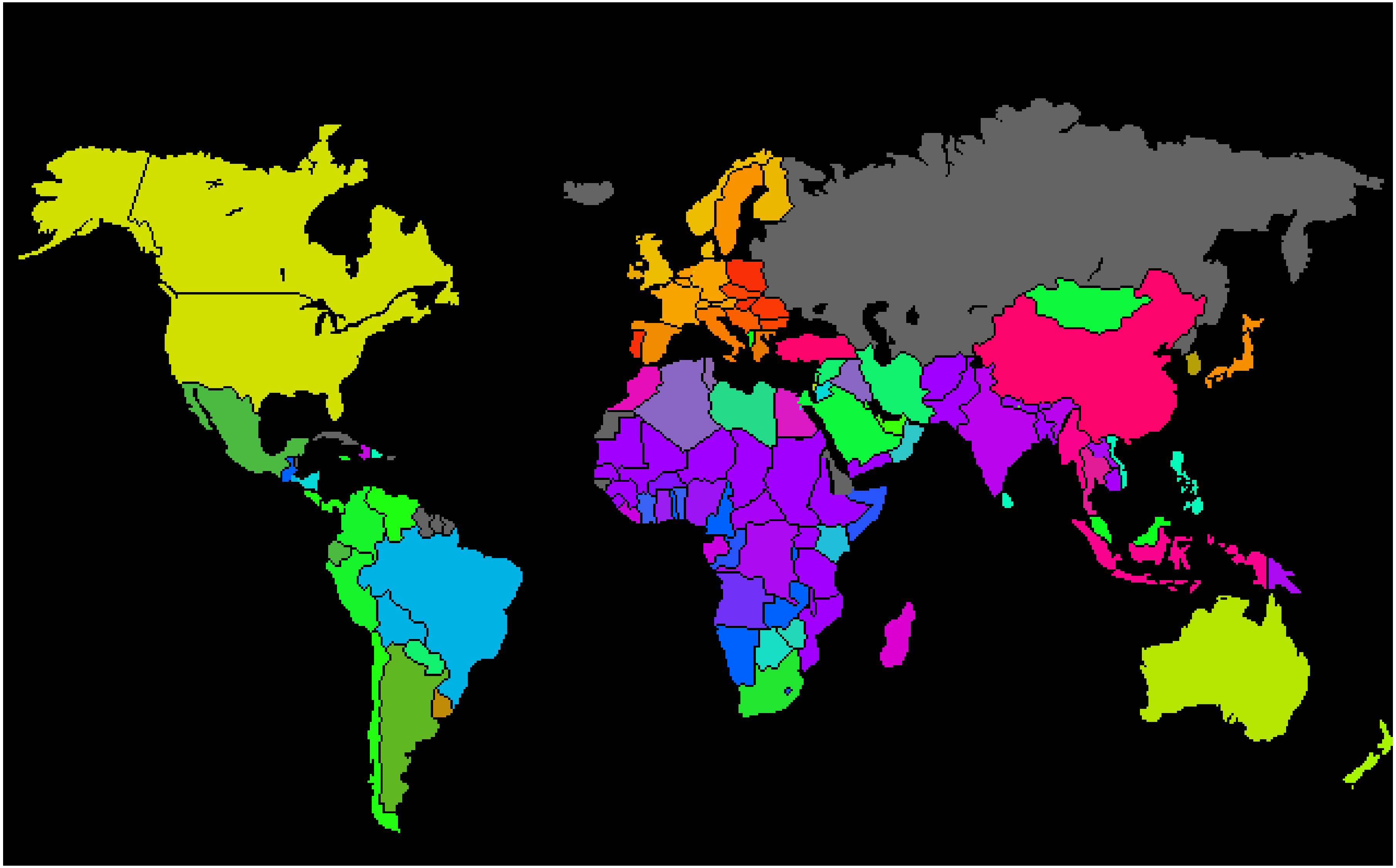




## The Country Names

| | | | | | |
|---|---|---|---|---|---|
| AFG | Afghanistan | GTM | Guatemala | NZL | New Zealand |
| AGO | Angola | HKG | Hong Kong | OAN | Taiwan, China |
| ALB | Albania | HND | Honduras | OMN | Oman |
| ARE | United Arab Emirates | HTI | Haiti | PAK | Pakistan |
| ARG | Argentina | HUN | Hungary | PAN | Panama |
| AUS | Australia | HVO | Burkina Faso | PER | Peru |
| AUT | Austria | IDN | Indonesia | PHL | Philippines |
| BDI | Burundi | IND | India | PNG | Papua New Guinea |
| BEL | Belgium | IRL | Ireland | POL | Poland |
| BEN | Benin | IRN | Iran, Islamic Rep. | PRT | Portugal |
| BGD | Bangladesh | IRQ | Iraq | PRY | Paraguay |
| BGR | Bulgaria | ISR | Israel | ROM | Romania |
| BOL | Bolivia | ITA | Italy | RWA | Rwanda |
| BRA | Brazil | JAM | Jamaica | SAU | Saudi Arabia |
| BTN | Bhutan | JOR | Jordan | SDN | Sudan |
| BUR | Myanmar | JPN | Japan | SEN | Senegal |
| BWA | Botswana | KEN | Kenya | SGP | Singapore |
| CAF | Central African Rep. | KHM | Cambodia | SLE | Sierra Leone |
| CAN | Canada | KOR | Korea, Rep. | SLV | El Salvador |
| CHE | Switzerland | KWT | Kuwait | SOM | Somalia |
| CHL | Chile | LAO | Lao PDR | SWE | Sweden |
| CHN | China | LBN | Lebanon | SYR | Syrian Arab Rep. |
| CIV | Cote d'Ivoire | LBR | Liberia | TCD | Chad |
| CMR | Cameroon | LBY | Libya | TGO | Togo |
| COG | Congo | LKA | Sri Lanka | THA | Thailand |
| COL | Colombia | LSO | Lesotho | TTO | Trinidad and Tobago |
| CRI | Costa Rica | MAR | Morocco | TUN | Tunisia |
| CSK | Czechoslovakia | MDG | Madagascar | TUR | Turkey |
| DEU | Germany | MEX | Mexico | TZA | Tanzania |
| DNK | Denmark | MLI | Mali | UGA | Uganda |
| DOM | Dominican Rep. | MNG | Mongolia | URY | Uruguay |
| DZA | Algeria | MOZ | Mozambique | USA | United States |
| ECU | Ecuador | MRT | Mauritania | VEN | Venezuela |
| EGY | Egypt, Arab Rep. | MUS | Mauritius | VNM | Viet Nam |
| ESP | Spain | MWI | Malawi | YEM | Yemen, Rep. |
| ETH | Ethiopia | MYS | Malaysia | YUG | Yugoslavia |
| FIN | Finland | NAM | Namibia | ZAF | South Africa |
| FRA | France | NER | Niger | ZAR | Zaire |
| GAB | Gabon | NGA | Nigeria | ZMB | Zambia |
| GBR | United Kingdom | NIC | Nicaragua | ZWE | Zimbabwe |
| GHA | Ghana | NLD | Netherlands | | |
| GIN | Guinea | NOR | Norway | | |
| GRC | Greece | NPL | Nepal | | |

FV=39 indices of quality of life

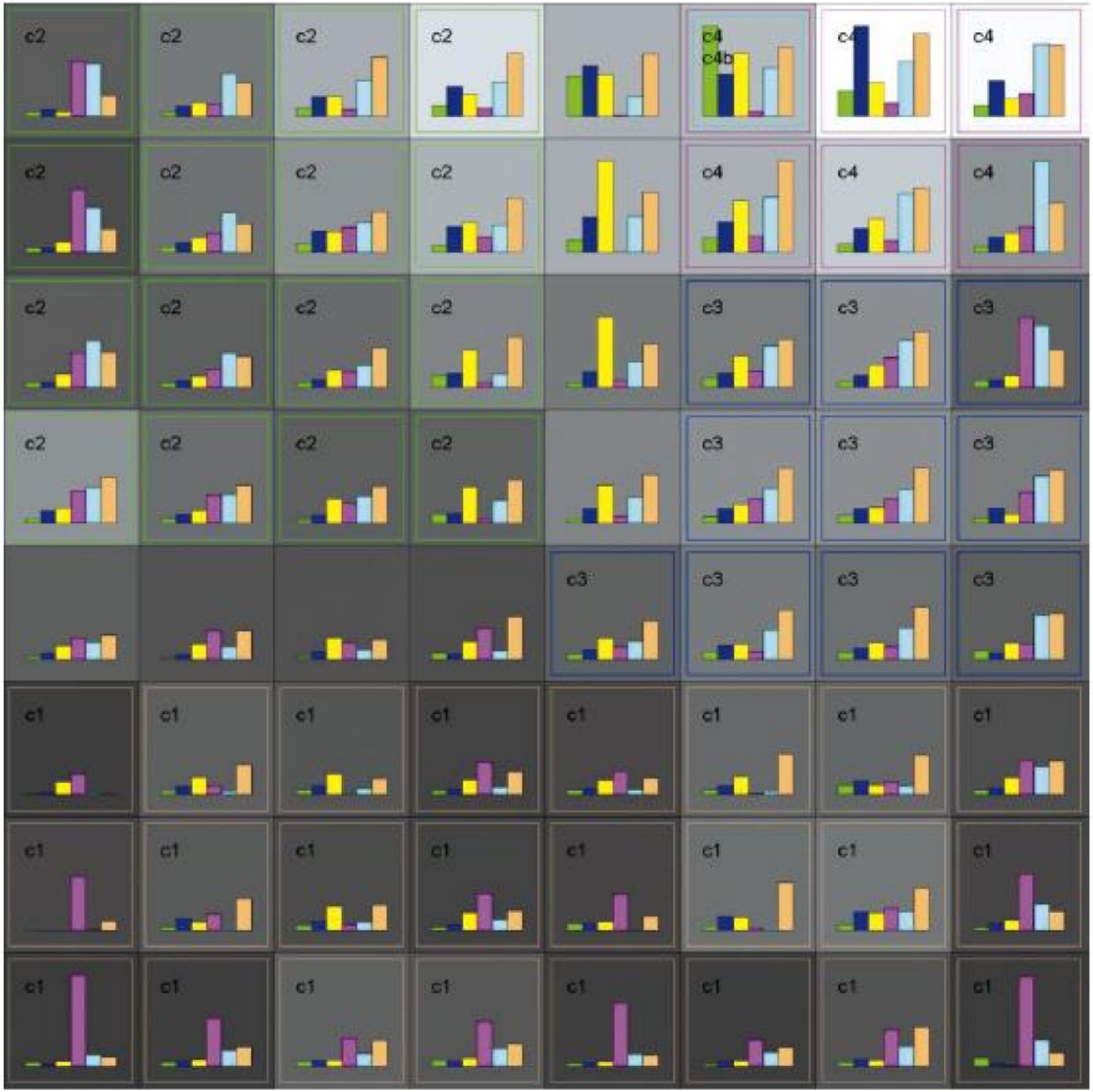


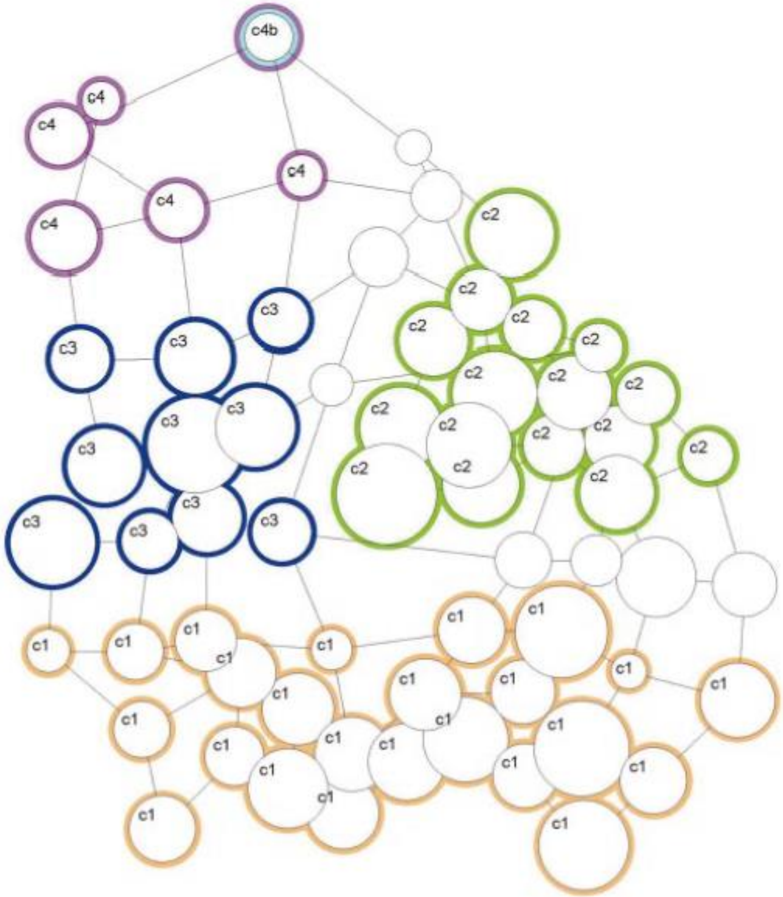
# SOM IN BIOLOGY

# Insulin resistance syndrome revisited: application of self-organizing maps

- Understand risk factors through clustering
- Variables considered:
  - Blood glucose
  - Serum insulin
  - Triglycerides
  - High density lipoprotein cholesterol
  - Systolic Blood pressure
  - Body mass index
  - Waist to hip ratio

- C1 – healthy controls
- C4 – hypertensive and compensatory hyperinsulinaemic subjects
- C4b – insulin resistant
- C2, C3 – intermediate groups


c2
c2
c2
c2
c4
c4b
c4
c4
c2
c2
c2
c2
c4
c4
c4
c2
c2
c2
c2
c3
c3
c3
c2
c2
c2
c2
c3
c3
c3
c3
c3
c3
c3
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1
c1



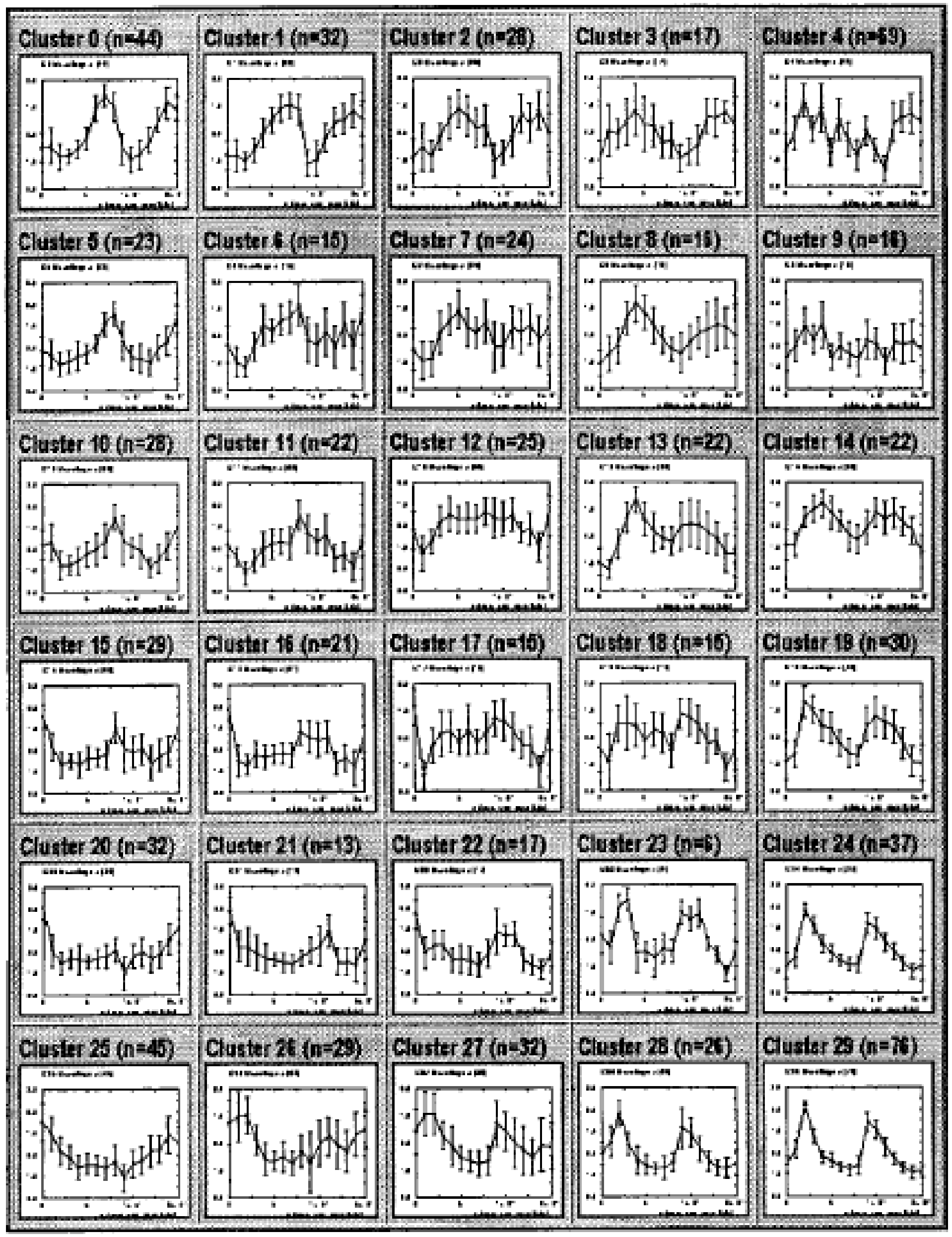
# Sammon's mapping

Each circle → a neuron in SOM
Circle size → # data points in that cluster
Inter-neuron dist → distance between
Centroids in the original n-d space

## Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation (Tamayo et al , PNAS,1999)

- Using SOMs to interpret gene expression data
- Hierarchical clustering is not quite suitable; more suitable when there is true hierarchical descent
- K-means produces "an unorganized collection of clusters"

a)



Yeast cell cycle SOM

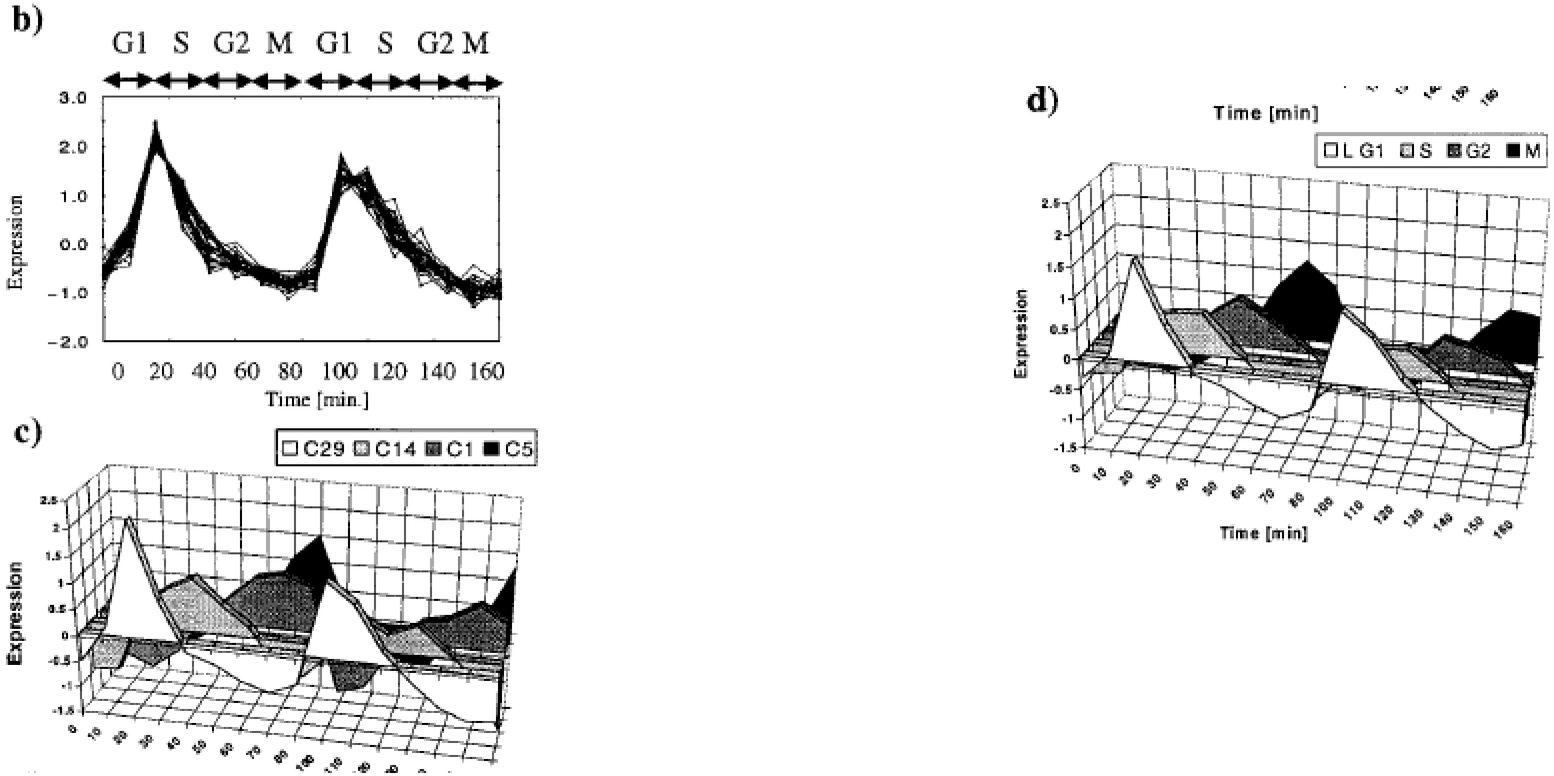


FIG. 2. Yeast Cell Cycle SOM. (*a*) 6 × 5 SOM. The 828 genes that passed the variation filter were grouped into 30 clusters. Each cluster is represented by the centroid (average pattern) for genes in the cluster. Expression level of each gene was normalized to have mean = 0 and SD = 1 across time points. Expression levels are shown on *y*-axis and time points on *x*-axis. Error bars indicate the SD of average expression. *n* indicates the number of genes within each cluster. Note that multiple clusters exhibit periodic behavior and that adjacent clusters have similar behavior. (*b*) Cluster 29 detail. Cluster 29 contains 76 genes exhibiting periodic behavior with peak expression in late $G_1$. Normalized expression pattern of 30 genes nearest the centroid are shown. (*c*) Centroids for SOM-derived clusters 29, 14, 1, and 5, corresponding to $G_1$, S, $G_2$ and M phases of the cell cycle, are shown. (*d*) Centroids for groups of genes identified by visual inspection by Cho *et al.* (4) as having peak expression in $G_1$, S, $G_2$, or M phase of the cell cycle are shown.

# Salient Features

- SOM picked cell cycled periodicity as a prominent feature
- Genes in the same cluster typically peak during the same phase (cluster 29, fig. b)
- Genes in neighboring clusters peak in nearby phases (fig. c, 24, 28, 29 have genes that peak in late G1 phase)