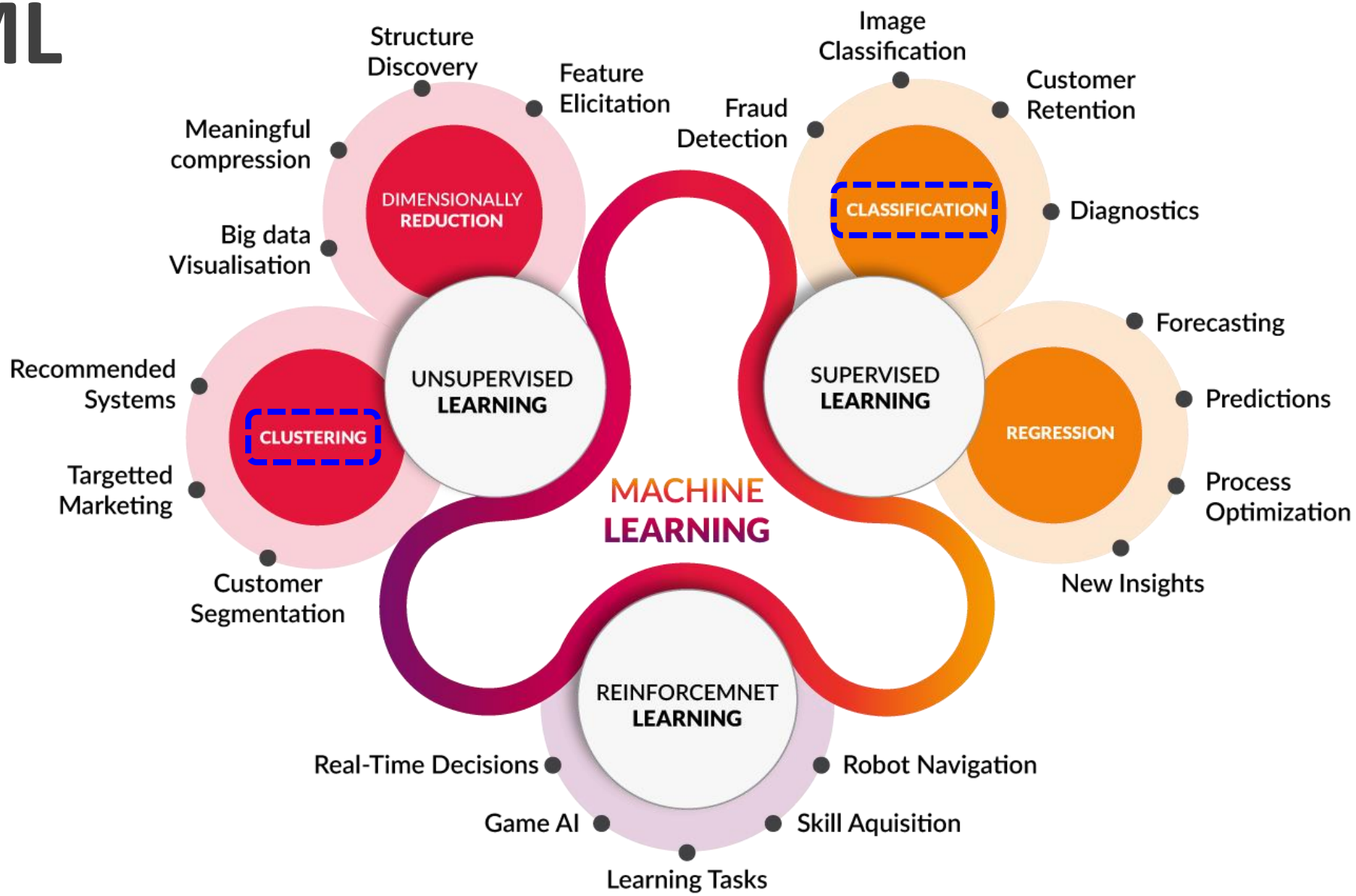
機器學習於材料資訊的應用
Machine Learning on Material Informatics

陳南佑(NAN-YOW CHEN)

楊安正(AN-CHENG YANG)

Types of ML



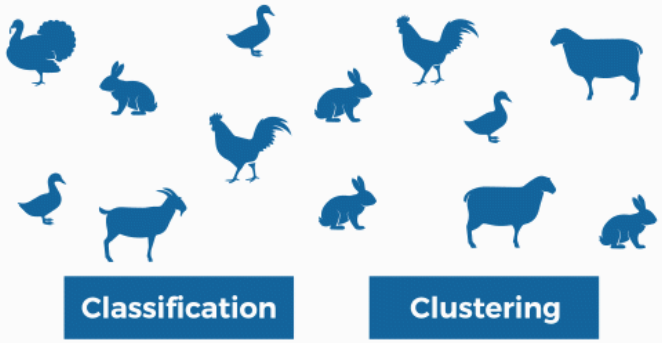
http://www.cognub.com/index.php/cognitive-platform/

Outline

- Classification vs clustering
- Hierarchical clustering
- K-means
- Modularity of networks
- Isomap and Dijkstra's algorithm
- Applications:
  - Morphological Structures Clustering for Antennal Lobe Local Neurons in Olfactory System of *Drosophila*
  - Morphological Structures Clustering for MD simulation results of Graphene
- Summary

Classification vs Clustering

- Classification is the process of learning a model that elucidate different predetermined classes of data. It is a two-step process, comprised of a learning step and a classification step. In learning step, a classification model is constructed and classification step the constructed model is used to prefigure the class labels for given data.
- Clustering is a technique of organising a group of data into classes and clusters where the objects reside inside a cluster will have high similarity and the objects of two clusters would be dissimilar to each other. Here the two clusters can be considered as disjoint. The main target of clustering is to divide the whole data into multiple clusters. Unlike classification process, here the class labels of objects are not known before, and clustering pertains to unsupervised learning.

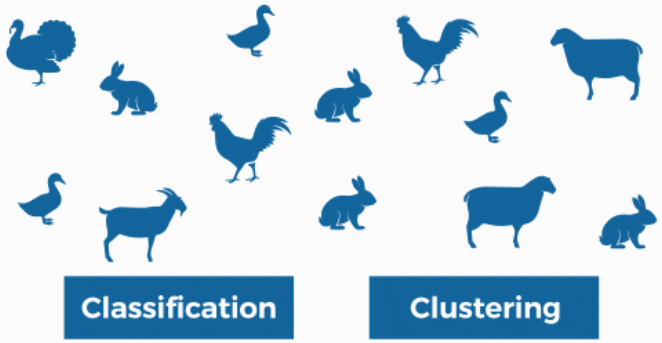


https://techdifferences.com/difference-between-classification-and-clustering.html
https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation

Classification vs Clustering

- In classification, you have a set of predefined classes and want to know which class a new object belongs to. On the other hand, clustering tries to group a set of objects and find whether there is some relationship between the objects.

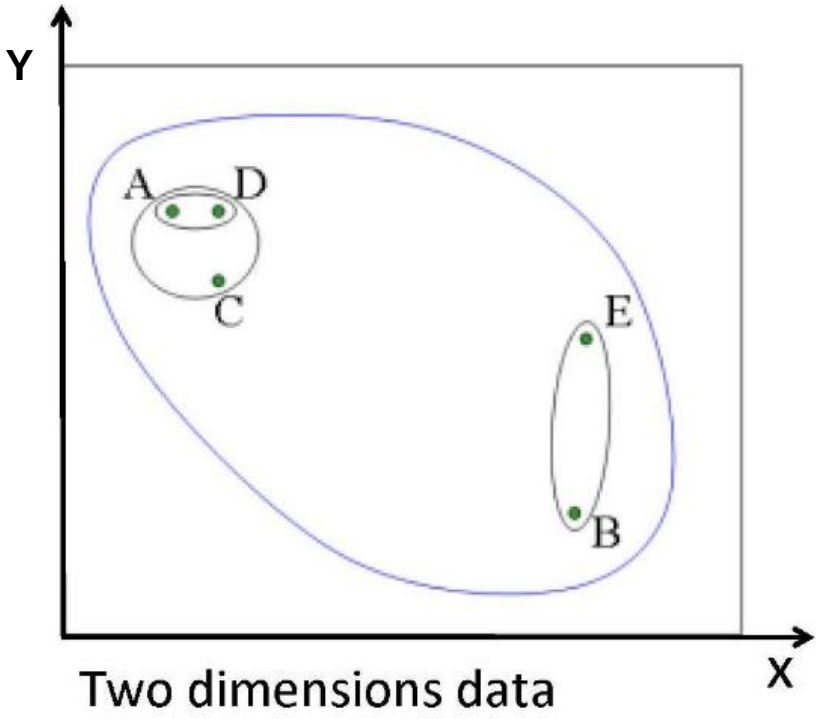
- The prior difference between classification and clustering is that classification is used in supervised learning technique where predefined labels are assigned to instances by properties. On the contrary, clustering is used in unsupervised learning where similar instances are grouped, based on their features or properties.



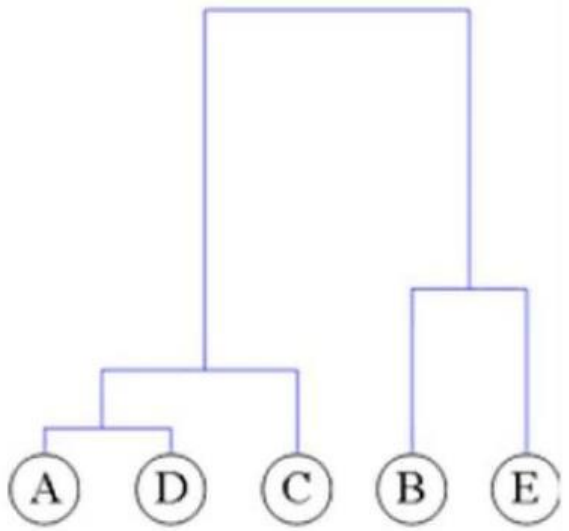
https://techdifferences.com/difference-between-classification-and-clustering.html
https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation

Hierarchical clustering

- Hierarchical clustering is an exploratory data analysis method for data classification to group similar characteristics together.



Two dimensions data

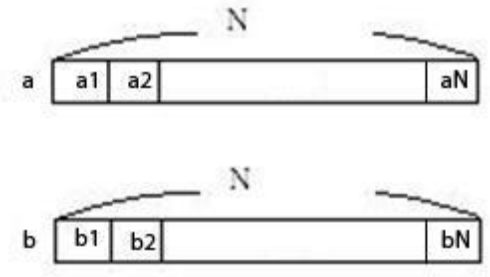


Dendrogram

https://en.wikipedia.org/wiki/Hierarchical_clustering

Hierarchical clustering

- Metric: distance measurement



Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the Covariance matrix

- Linkage criteria:

Names	Formula
Maximum or complete-linkage clustering	$\max\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Minimum or single-linkage clustering	$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{\|A\|.\|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b).$
Centroid linkage clustering, or UPGMC	$\|c_s - c_t\|$ where c_s and c_t are the centroids of clusters s and t, respectively.
Minimum energy clustering	$\frac{2}{nm} \sum_{i,j=1}^{n,m} \|a_i - b_j\|_2 - \frac{1}{n^2} \sum_{i,j=1}^{n} \|a_i - a_j\|_2 - \frac{1}{m^2} \sum_{i,j=1}^{m} \|b_i - b_j\|_2$

https://en.wikipedia.org/wiki/Hierarchical_clustering

Example of hierarchical clustering

- Distance between Italian cities in km using single linkage.

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0
	Bari	Firenze	Milano	Napoli	Roma	Torino

Example of hierarchical clustering

- Distance between Italian cities in km using single linkage.

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

Example of hierarchical clustering

- Distance between Italian cities in km using single linkage.

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

Example of hierarchical clustering

- Distance between Italian cities in km using single linkage.

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

Example of hierarchical clustering

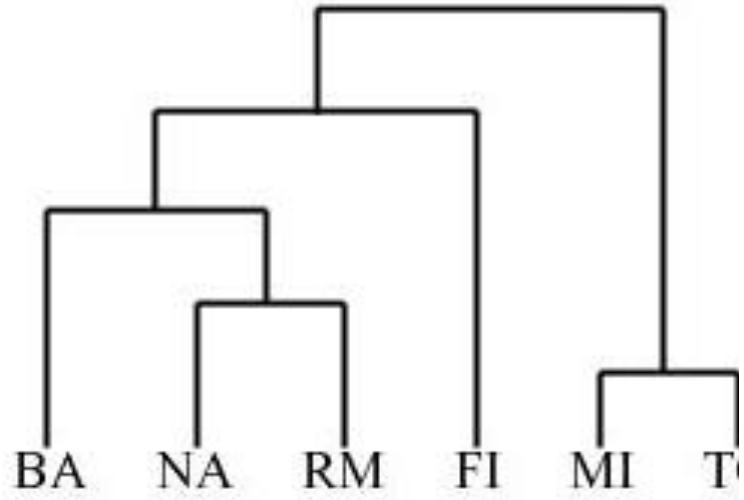
- Distance between Italian cities in km using single linkage.

	BA/NA/RM/FI	MI/TO
BA/NA/RM/FI	0	295
MI/TO	295	0

https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

Example of hierarchical clustering

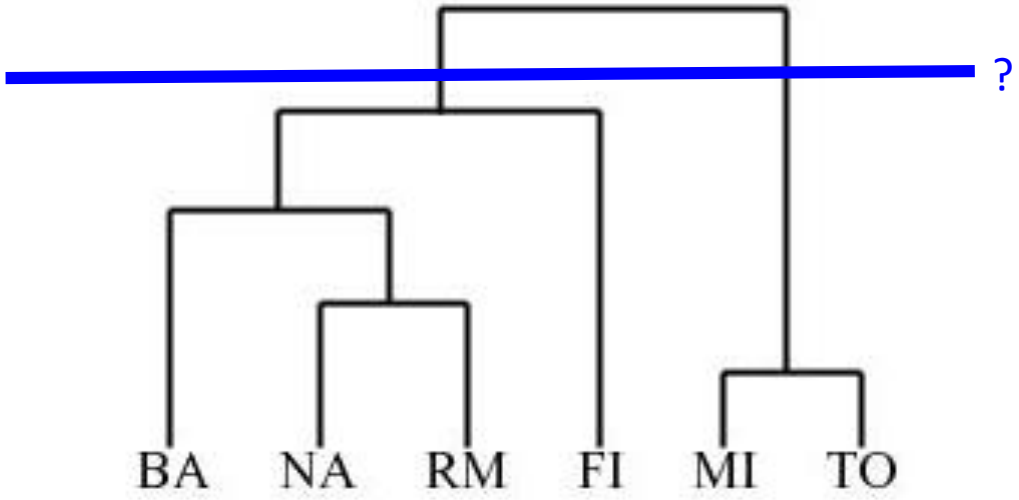
- The process is shown by the following dendrogram:



Example of hierarchical clustering

- The process is shown by the following dendrogram:

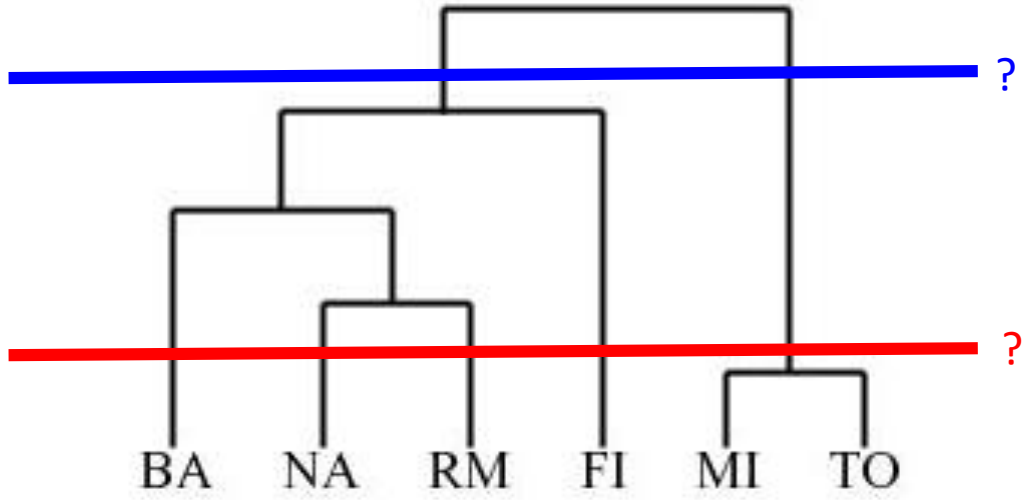
- Dilemma:
  - Large threshold
    - too few groups



Example of hierarchical clustering

- The process is shown by the following dendrogram:

- Dilemma:
  - Large threshold
    - too few groups
  - Small threshold
    - too many groups



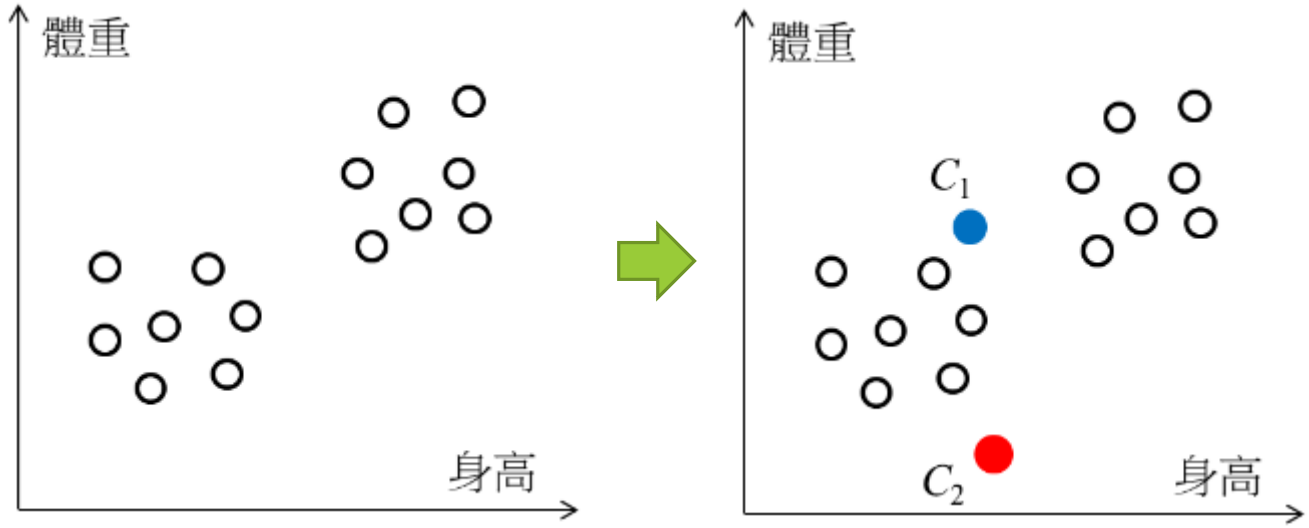
Example of hierarchical clustering

- The process is shown by the following dendrogram:

- Dilemma:
  - Large threshold
    - too few groups
  - Small threshold
    - too many groups

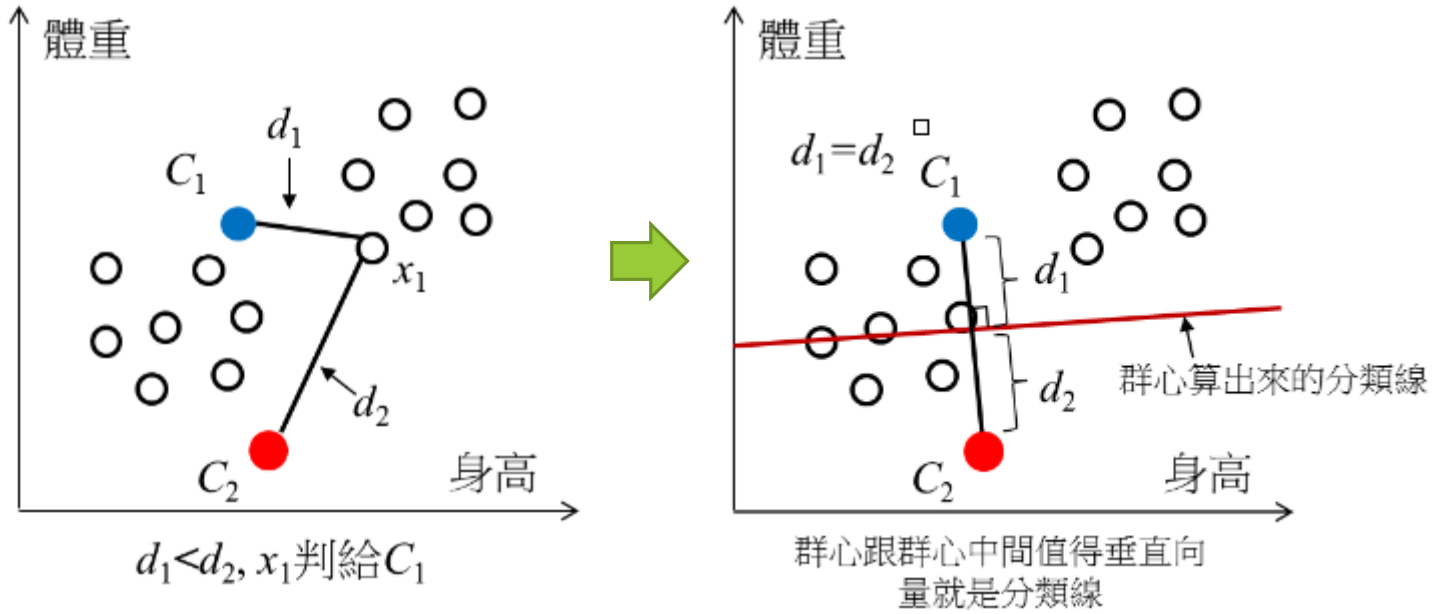
K-means

- Step 1: 設定好要分k群，並在資料空間中隨機取k個群心。



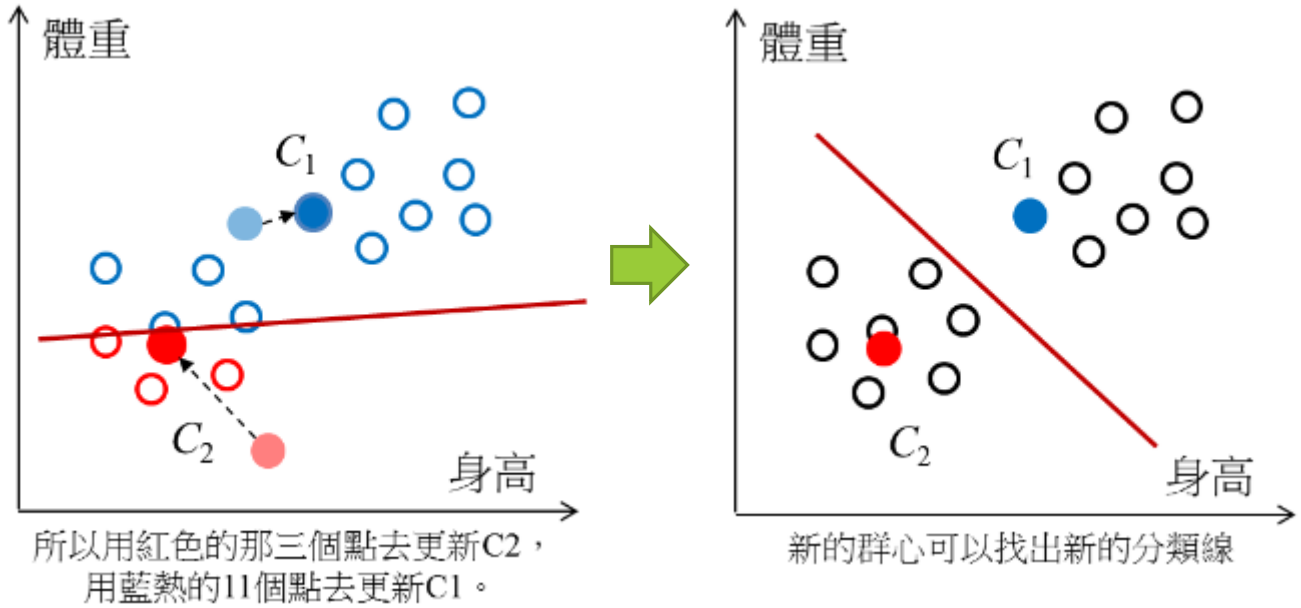
K-means

- Step 1: 設定好要分k群，並在資料空間中隨機取k個群心。
- Step 2: 每個資料點都跟這些群心計算距離，將資料點判給離最近的群心。



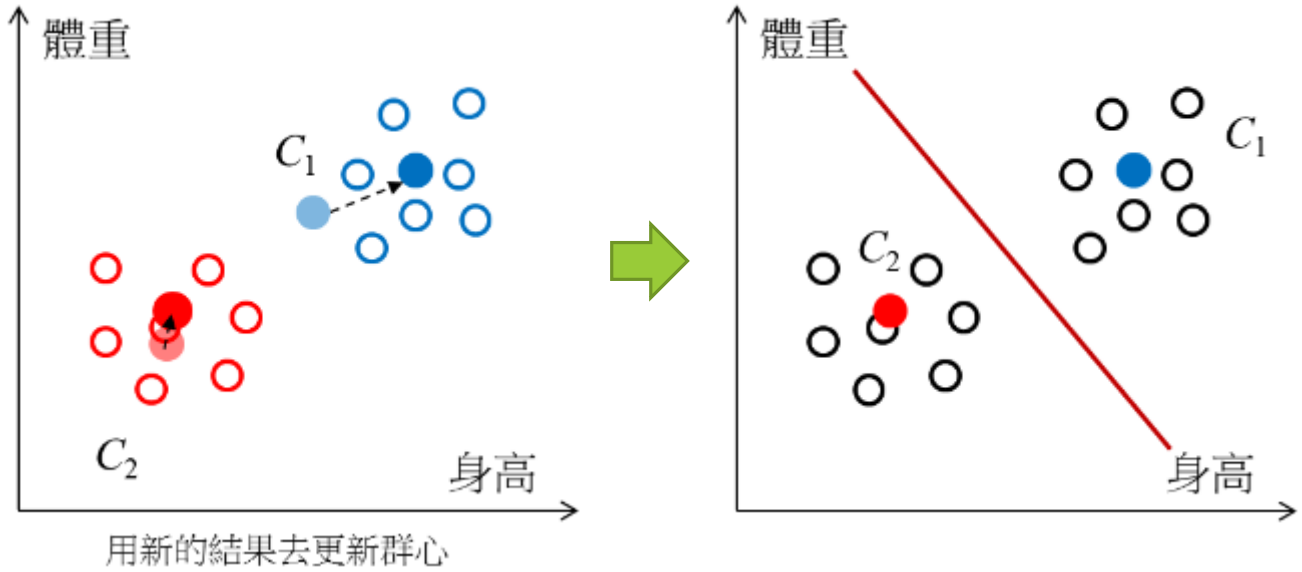
K-means

- Step 1: 設定好要分k群，並在資料空間中隨機取k個群心。
- Step 2: 每個資料點都跟這些群心計算距離，將資料點判給離最近的群心。
- Step 3: 每個群心都有被分過來的資料，重新計算這些資料的群心。



K-means

- Step 1: 設定好要分k群，並在資料空間中隨機取k個群心。
- Step 2: 每個資料點都跟這些群心計算距離，將資料點判給離最近的群心。
- Step 3: 每個群心都有被分過來的資料，重新計算這些資料的群心。
- Step 4: 重覆Step 2~4，直到所有群心不太變動為止。

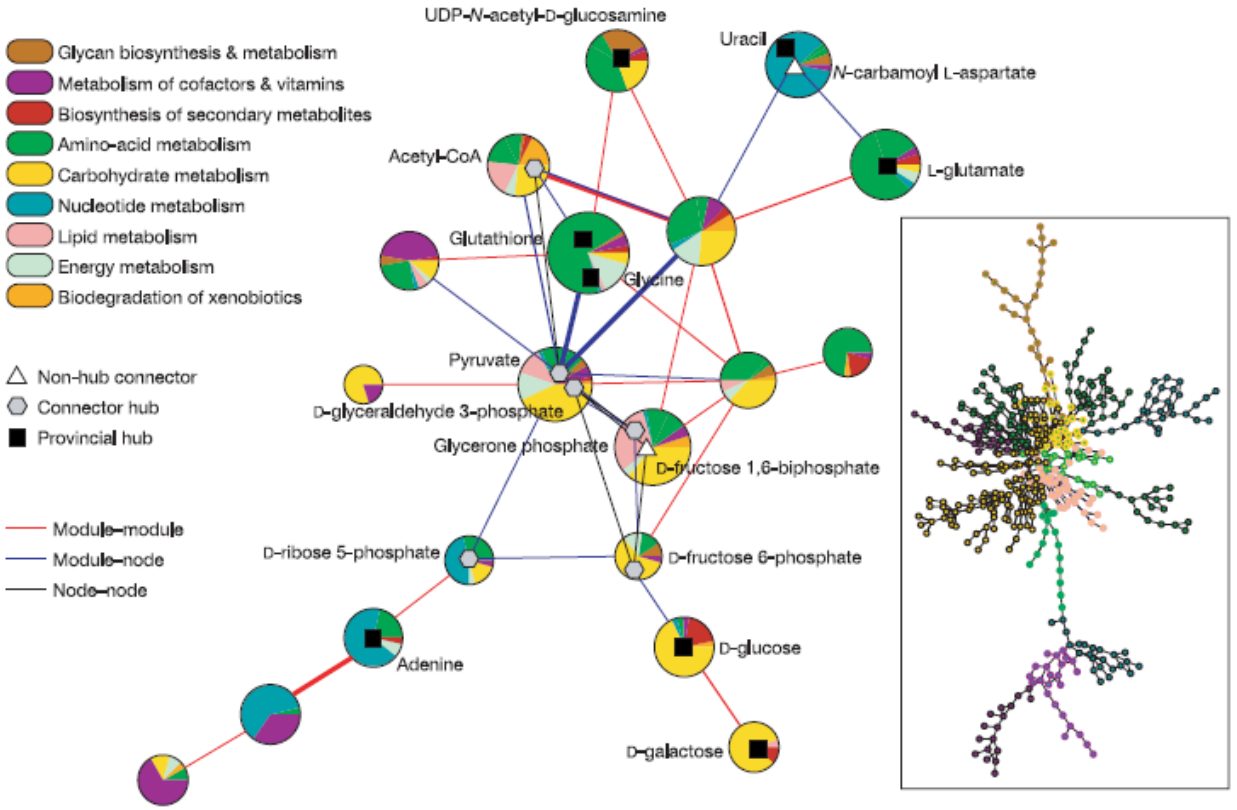


缺點：

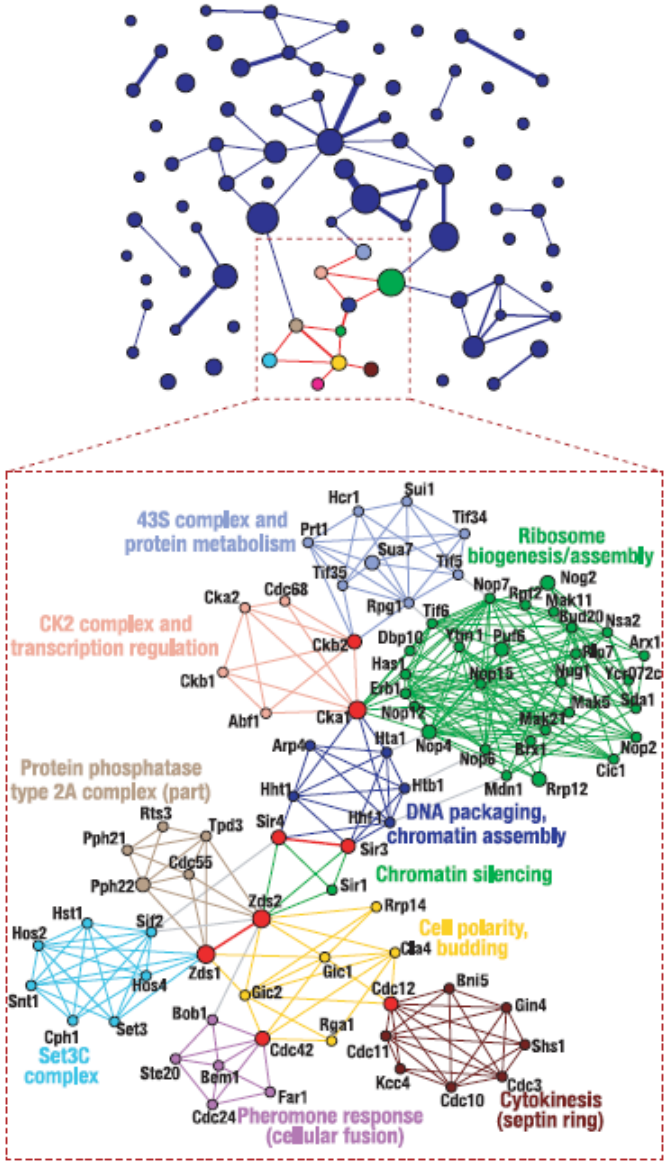
- K需要事先給定的；初始值選的不好，可能無法收斂。
- 當資料量非常大時，計算時間非常可觀。

https://medium.com/@chih.sheng.huang821/機器學習集群分析-k-means-clustering-e608a7fe1b43

Modularity of networks



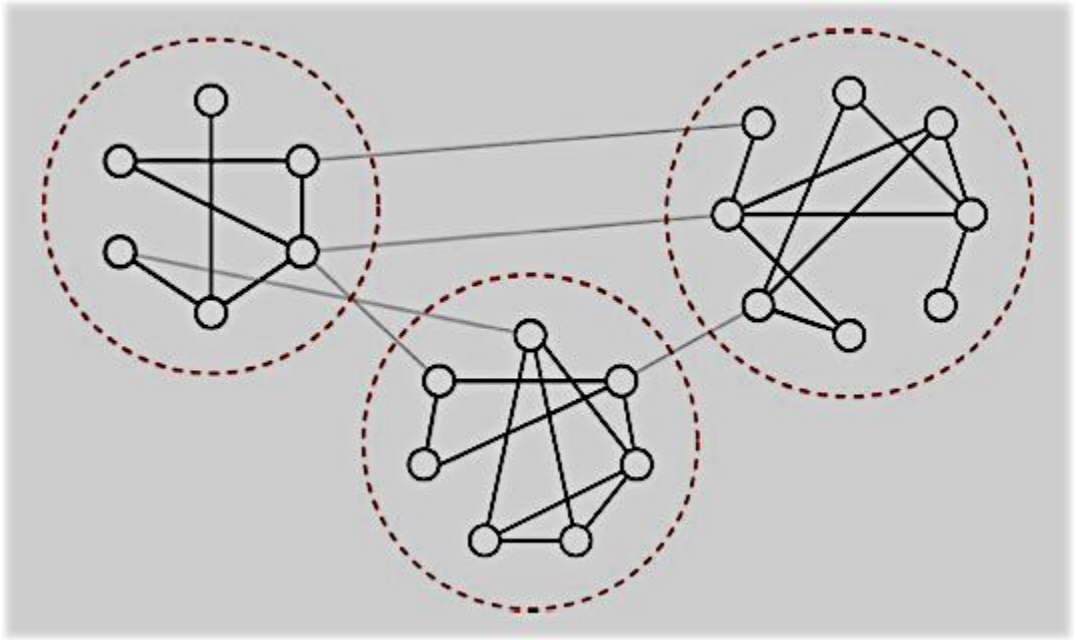
Nature **433**, 895 (2005)



Nature **435**, 814 (2005)

Modularity of networks

- Definition of a module: loosely linked island of densely connected nodes.



Modularity of networks

- Definition of a module: loosely linked island of densely connected nodes.
- Partitioning a network into modules so that nodes in one module are similar to each other and are as different as possible from the nodes in other modules.
- In order to describe two nodes are similar or not, we need to define a similarity measure and we also need a score function for our objectives.

Modularity of networks

- The goal of modularity is to find the best community structure, i.e., to maximize the intra module connections as many as possible and to minimize the inter module connections as few as possible.

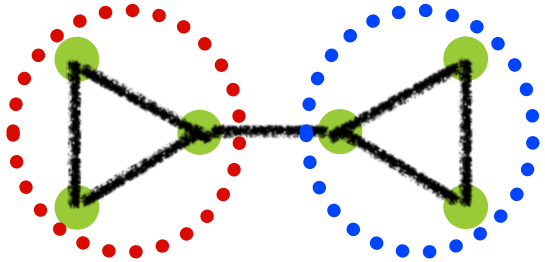
- This problem is the same as the partitioning task in parallel computing. In general, finding an exact solution to a partitioning task of this kind is believed to be an NP-hard (non-deterministic polynomial-time hard) problem, making it prohibitively difficult to solve exactly for large graphs.

- Optimization algorithms can be used to partition the nodes into modules with some given score functions.

Modularity of networks

- Definition of modularity:

$$Q = \sum_{s=1}^{N_M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right]$$



where

- N_M: number of modules in the network
- l_s: number of intra-modular links in module s
- d_s: sum of the degrees of the nodes in module s
- L: total number of links in the network

N_M=2

l_1=3, l_2=3

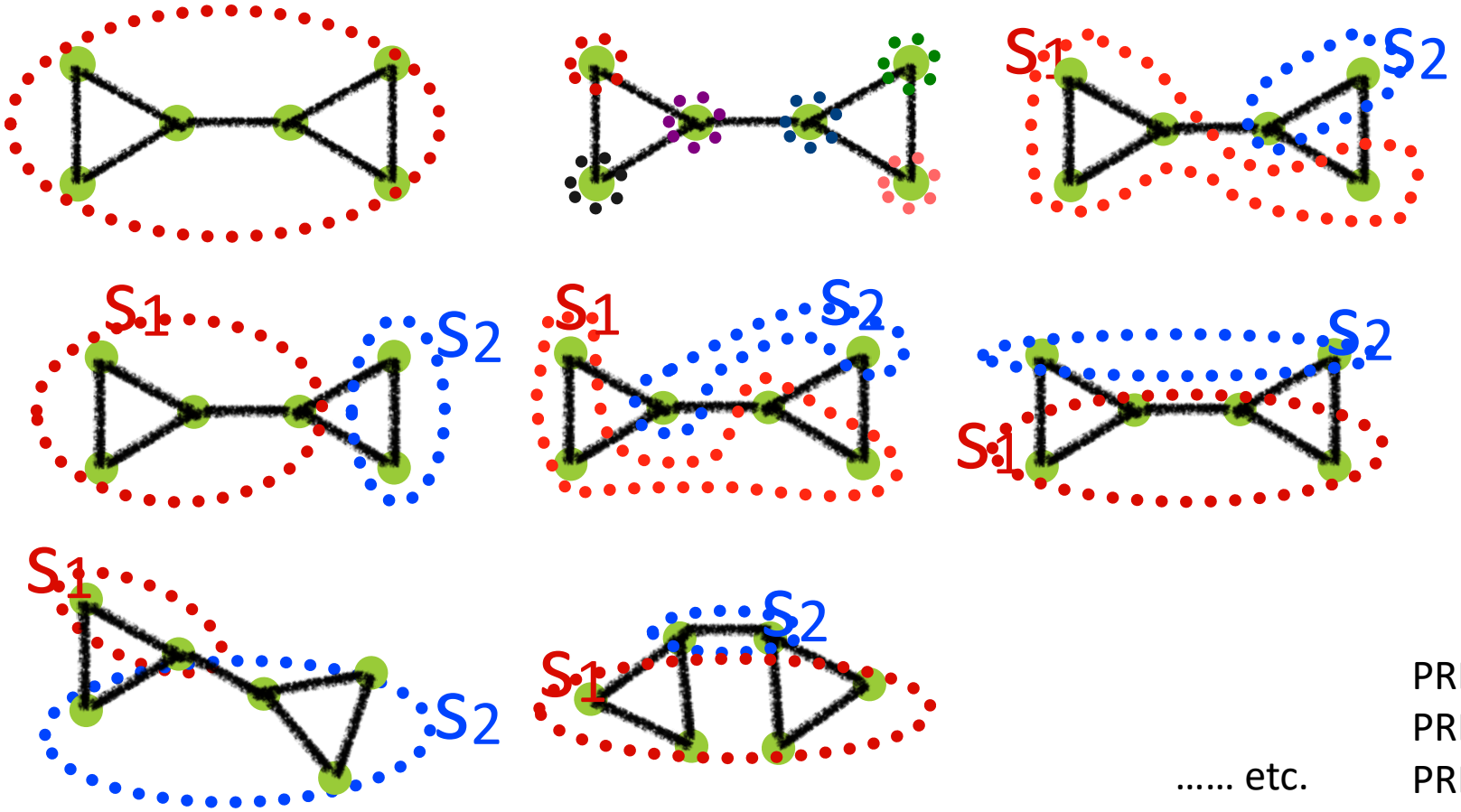
d_1=7, d_2=7

L=7

- Q = 0.357

Roger Guimerà, et al.: Nature **433**, 895 (2005)

Modularity of networks

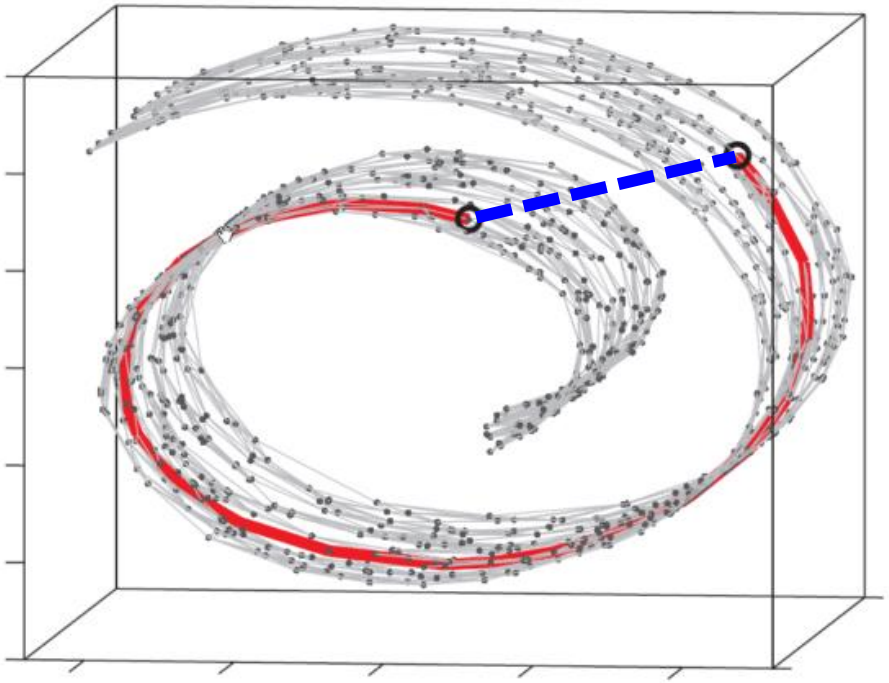


…… etc.

PRE **68**, 026121 (2003)
PRE **69**, 026113 (2004)
PRE **70**, 025101 (2004)

Isomap

- Isomap is an extension of multi dimensional scaling (MDS), where pairwise euclidean distances between data points are replaced by geodesic distance on a high-dimensional manifold which is constructed by these data points.



For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high dimensional input space (length of blue dashed line) may not accurately reflect their intrinsic similarity.

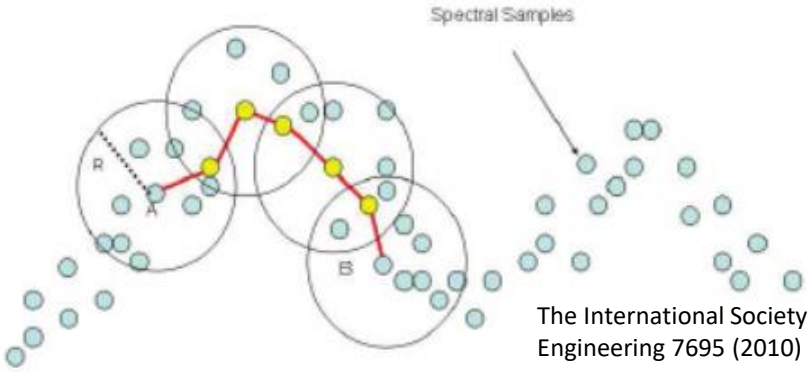
The red solid line is the geodesic distance (*i.e.* Dijkstra's distance) and the blue dashed line is the euclidean distance between two points, respectively.

Joshua B. Tenenbaum, et al.: Science **290**, 2319 (2000).

Dijkstra’s algorithm

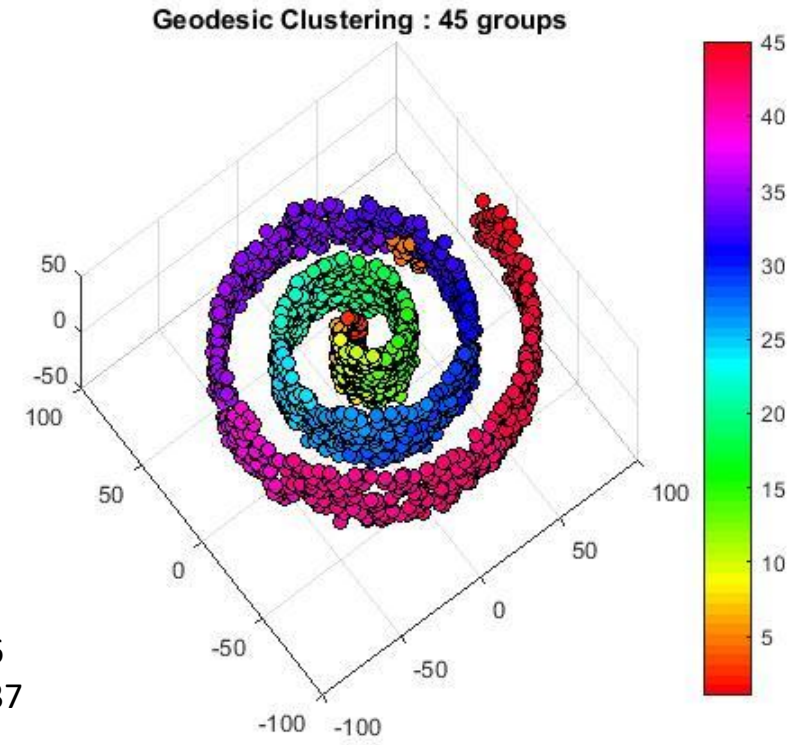
- Dijkstra's algorithm is an algorithm for finding the shortest paths between nodes in a graph, which may represent, for example, road networks. It was conceived by computer scientist Edsger W. Dijkstra in 1956.

- Steps:
  - Build graph with k-neighbors or ε-ball.
  - Weight graph with euclidean distance.
  - Compute pairwise geodesic distances by Dijkstra’s algorithm.



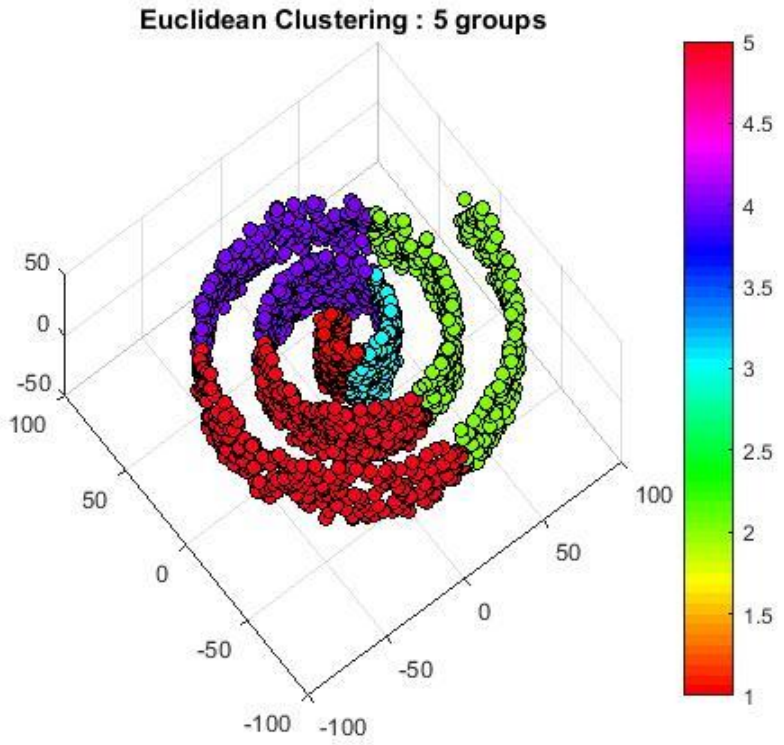
The International Society for Optical Engineering 7695 (2010)

Testing case - swiss roll



Nodes = 2000
Edges = 6036
Ave. Degree = 6.036
Modularity Q = 0.937

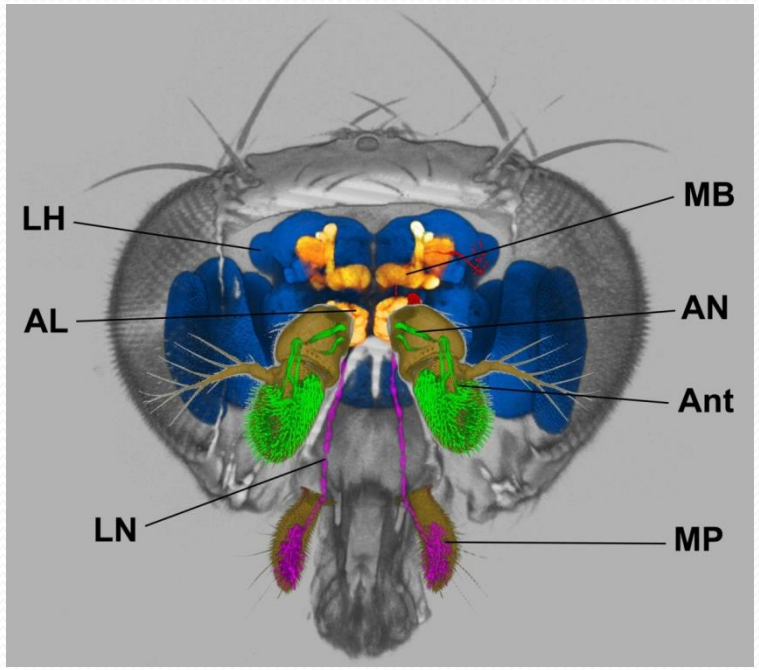
(5-neighbors)

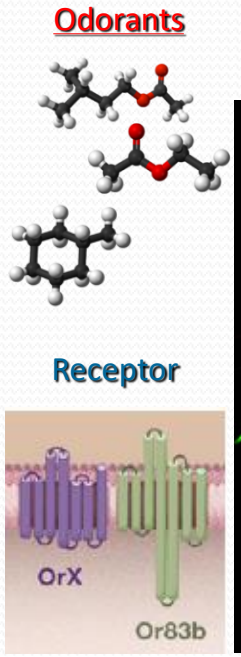


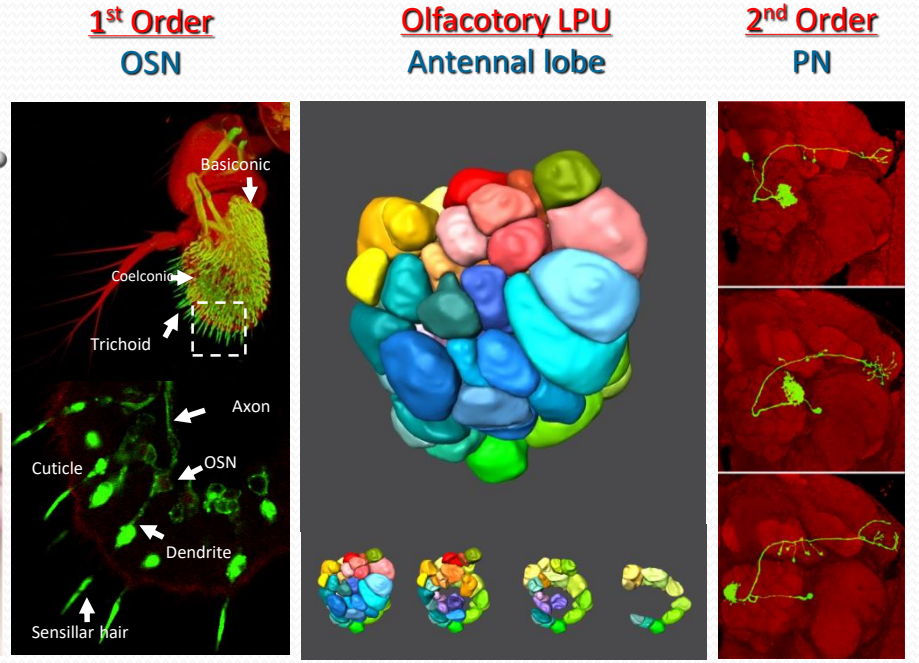
Nodes = 2000
Edges = 1999000
Ave. Degree = 1999
Modularity Q = 0.187

(1999-neighbors)

Adult fly olfactory system



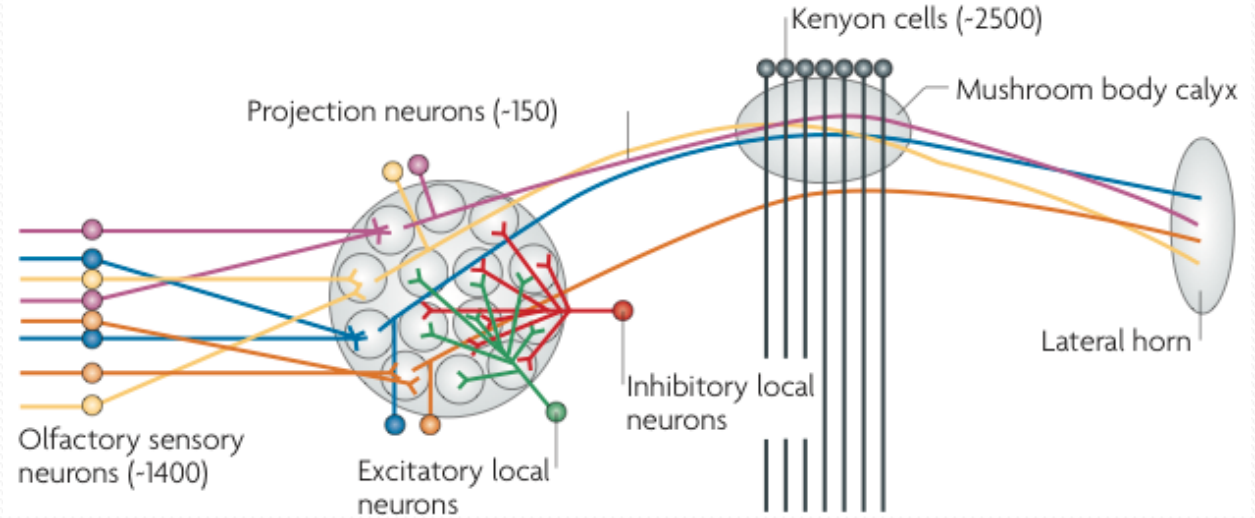




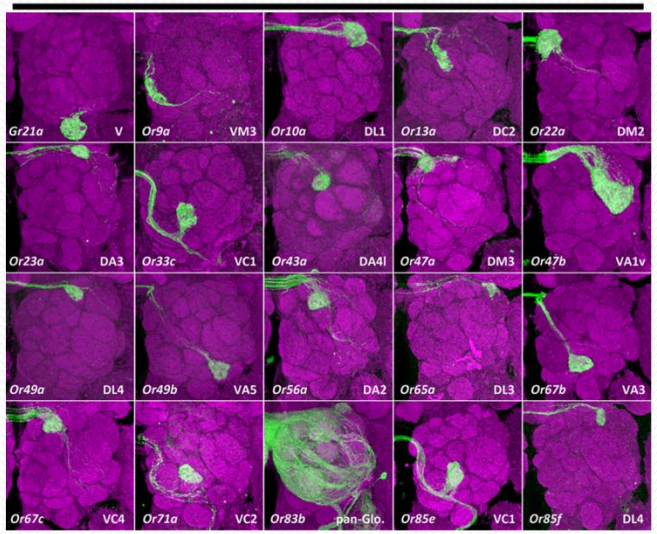
OSN: Olfactory Sensory Neurons
PN: Projection Neurons
LPU: Local Processing Units

http://www.sciencephoto.com/index.html

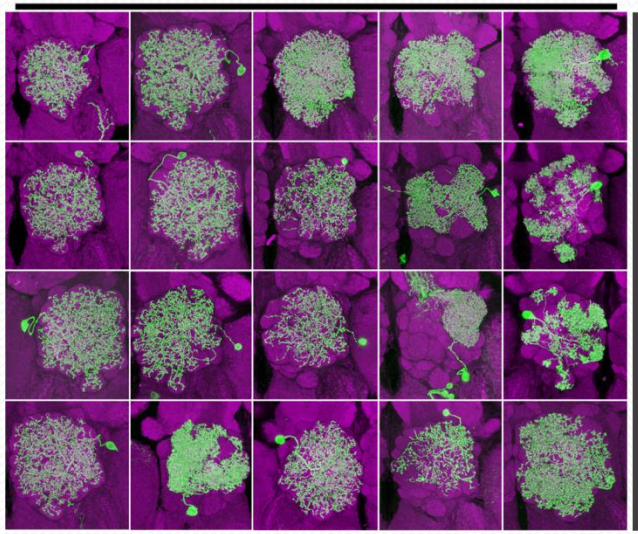
The principles for olfactory transmission



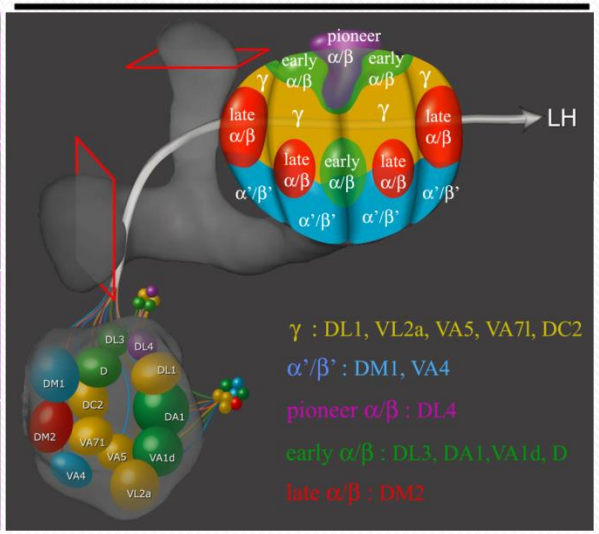
Stereotypy of OSNs



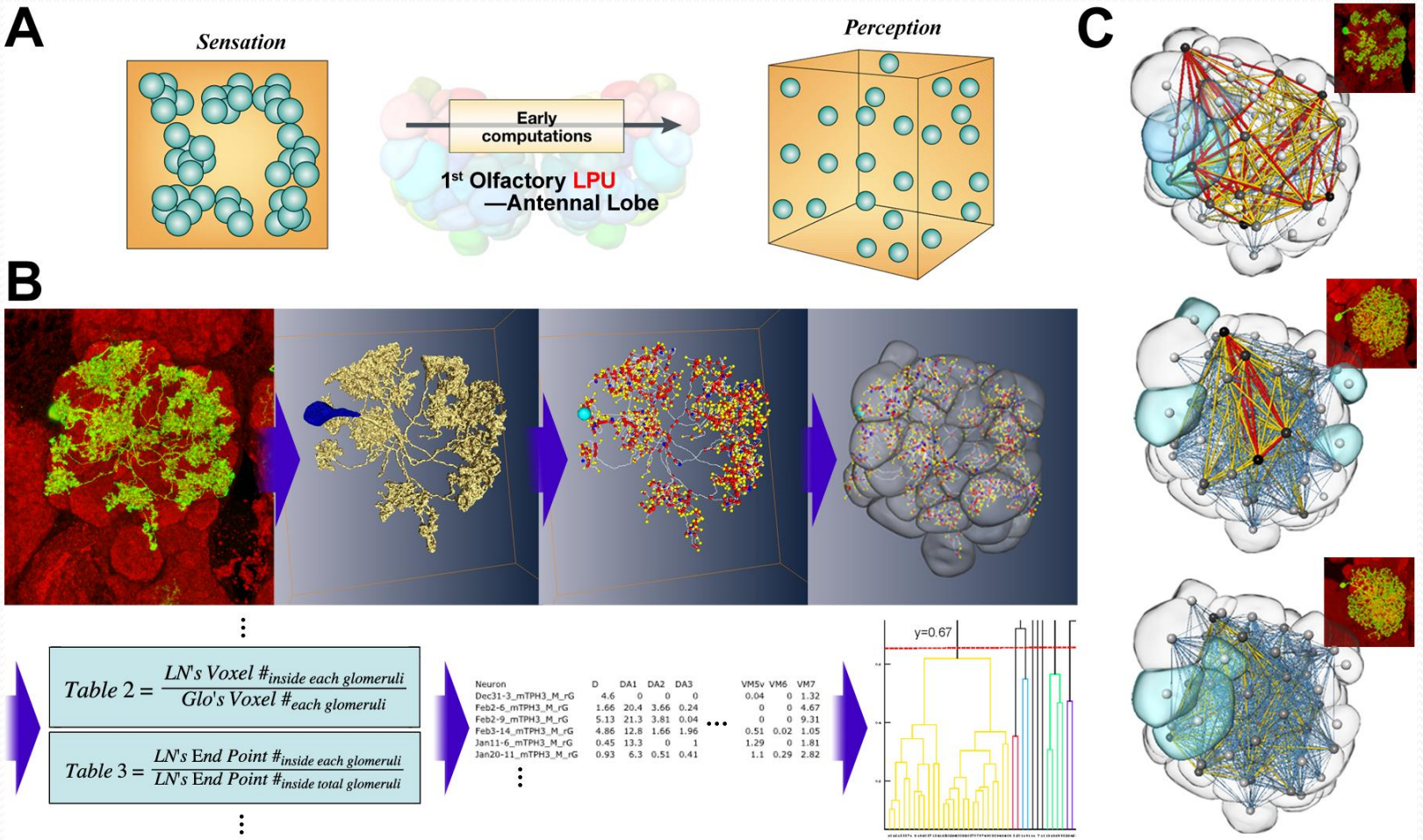
Excitatory & inhibitory local interneurons



Stereotypy of PNs

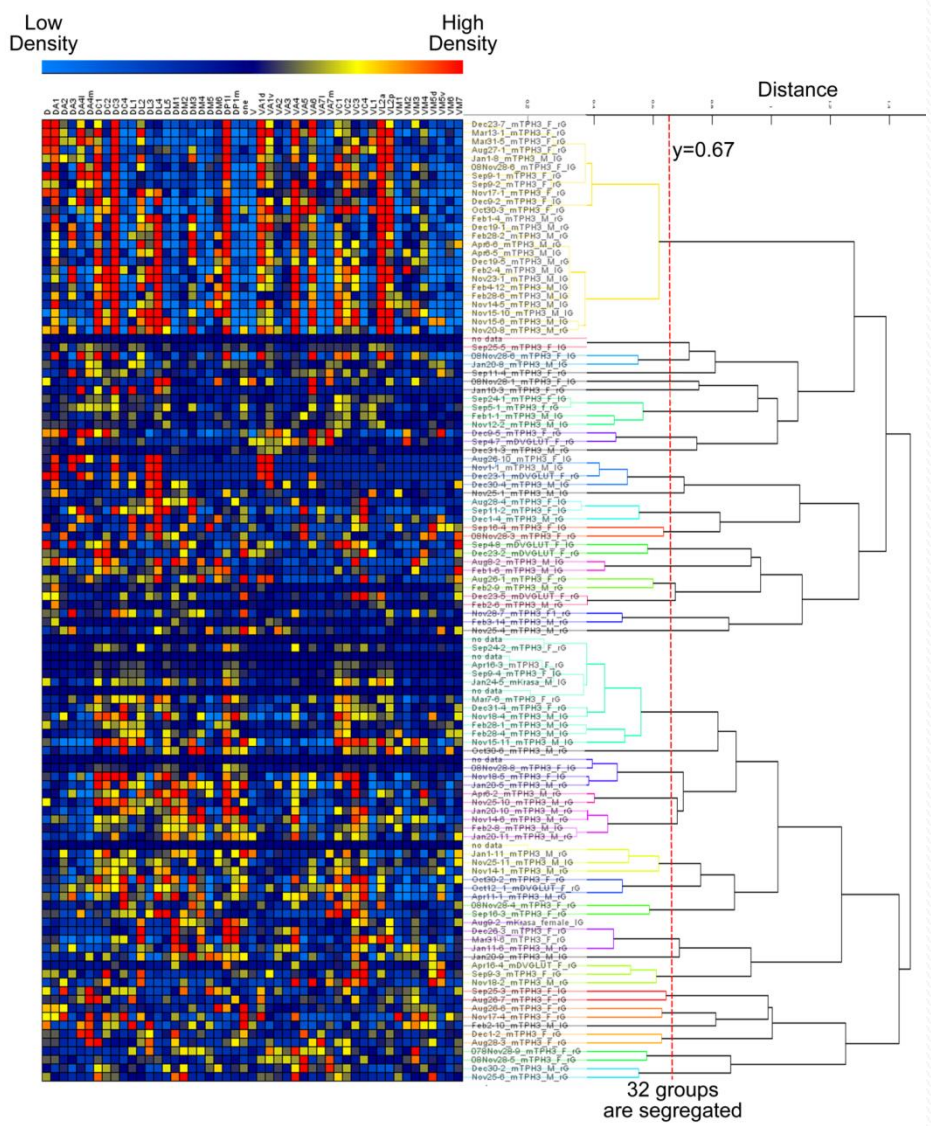


The operating procedure for hard wiring network analysis

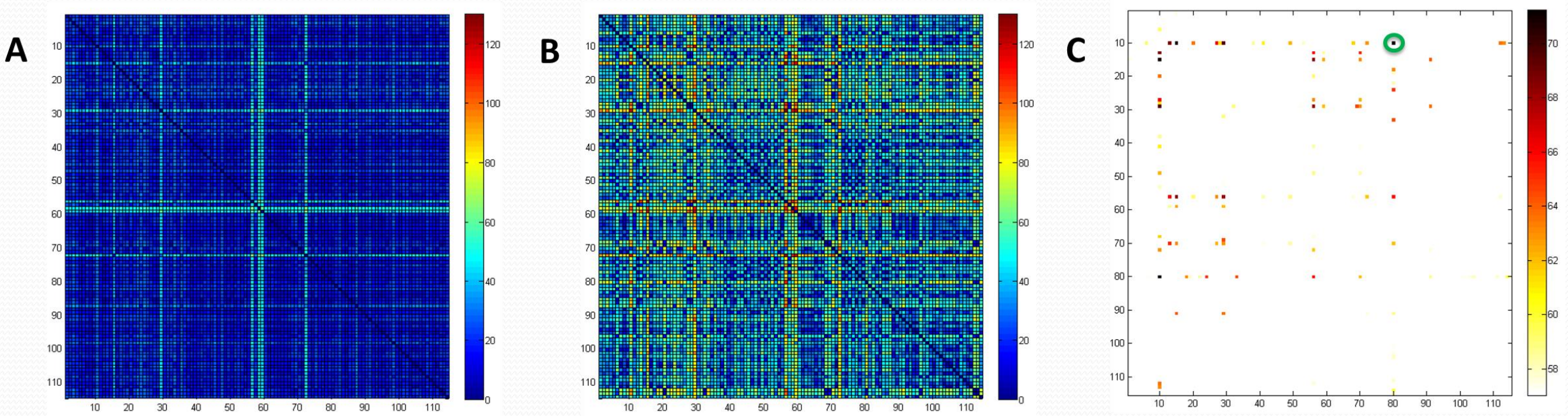


Hierarchical clustering results

A schematic diagram for innervation table and clustering results of 115 local neurons in olfactory system of *Drosophila* by hierarchical clustering.

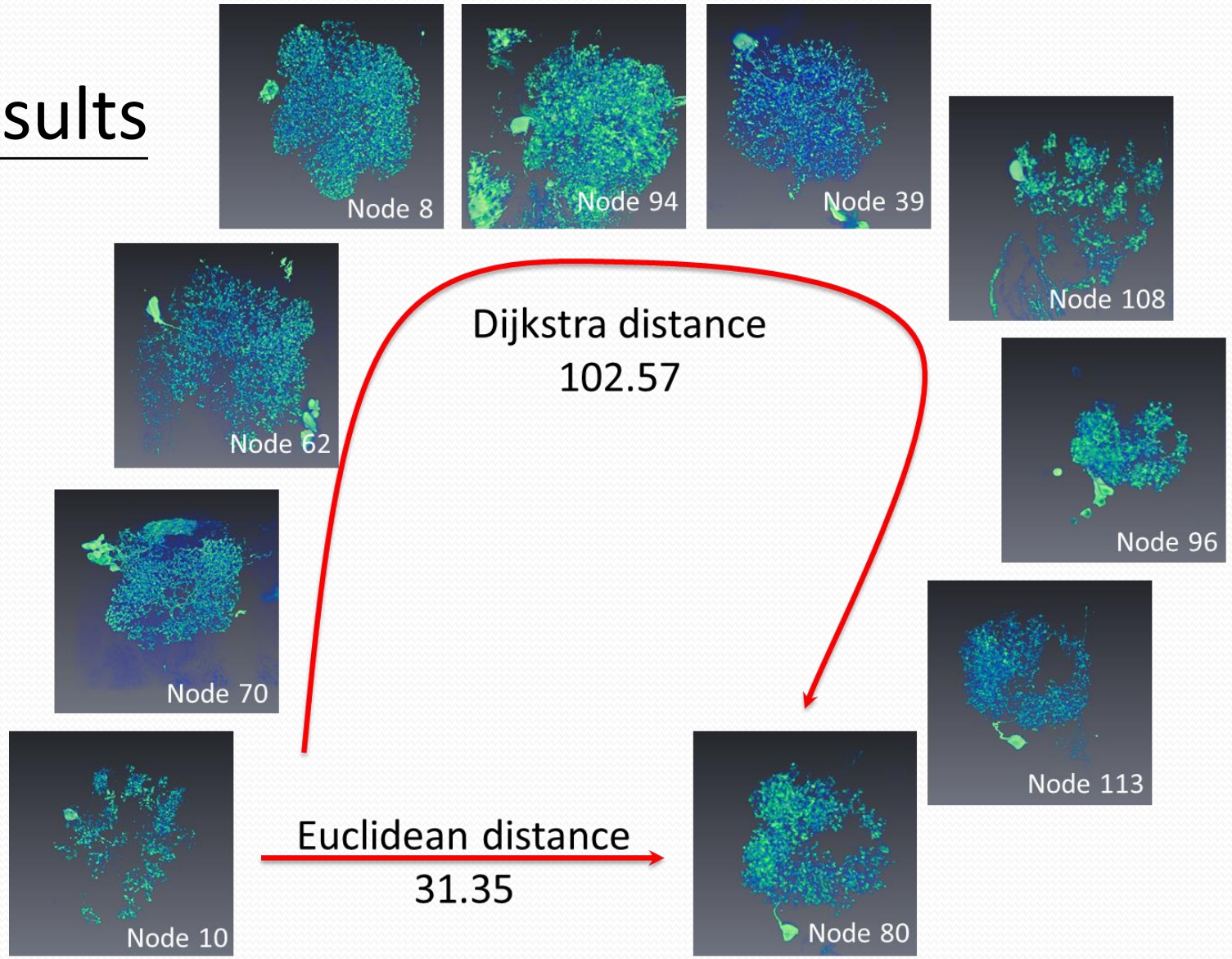


Isomap results



(A) Euclidean distance matrix. (B) Dijkstra distance matrix (3-neighbors). (C) Top 20% difference between (A) and (B). The max. difference in Euclidean and Dijkstra distance matrix is between Node 10 and Node 80.

Isomap results

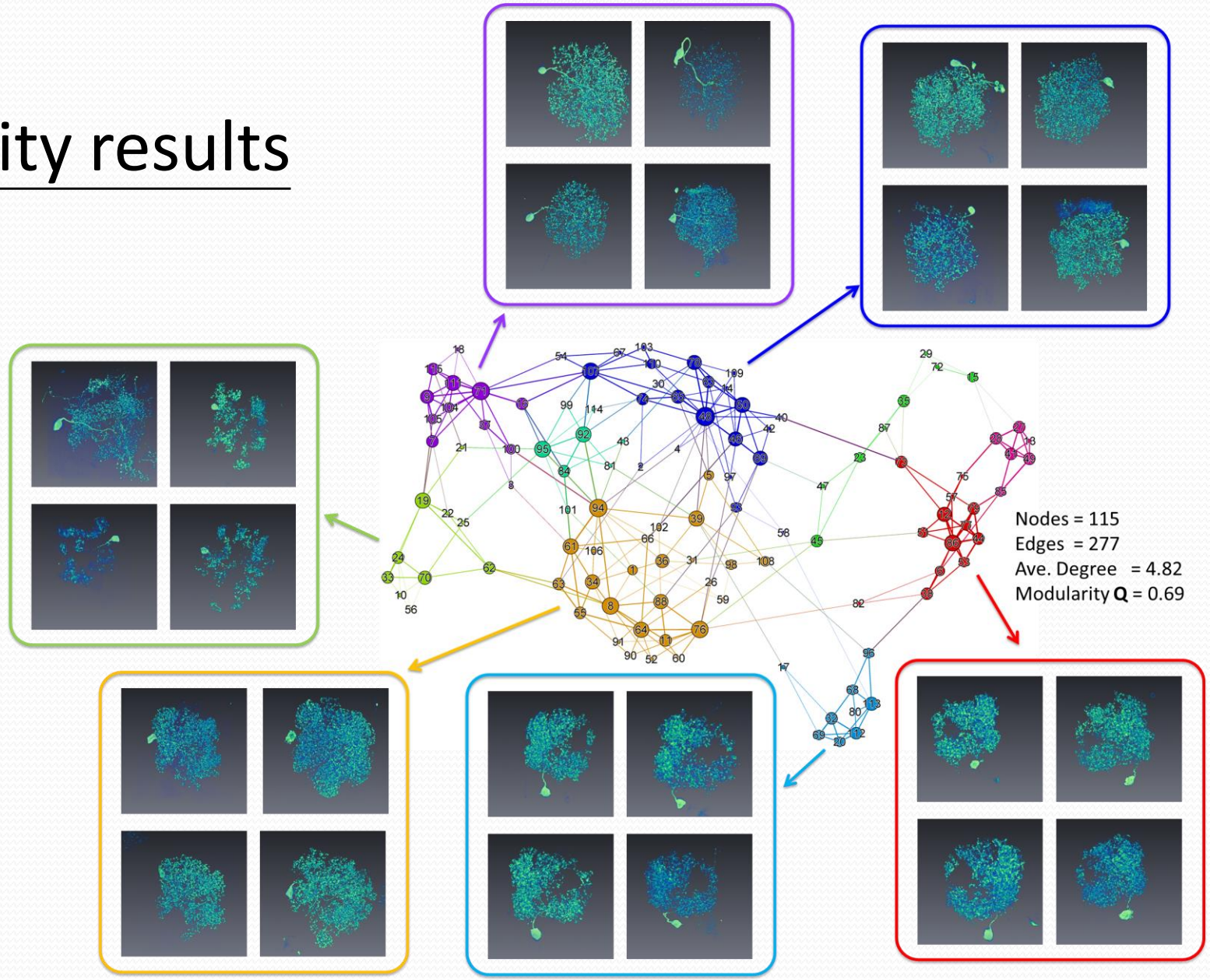


Images from Node 10 to Node 80.

Node 10 → Node 70 → Node 62 → Node 8 → Node 94 → Node 39 → Node 108 → Node 96 → Node 113 → Node 80

Euclidean distance = 31.35
Dijkstra distance = 102.57

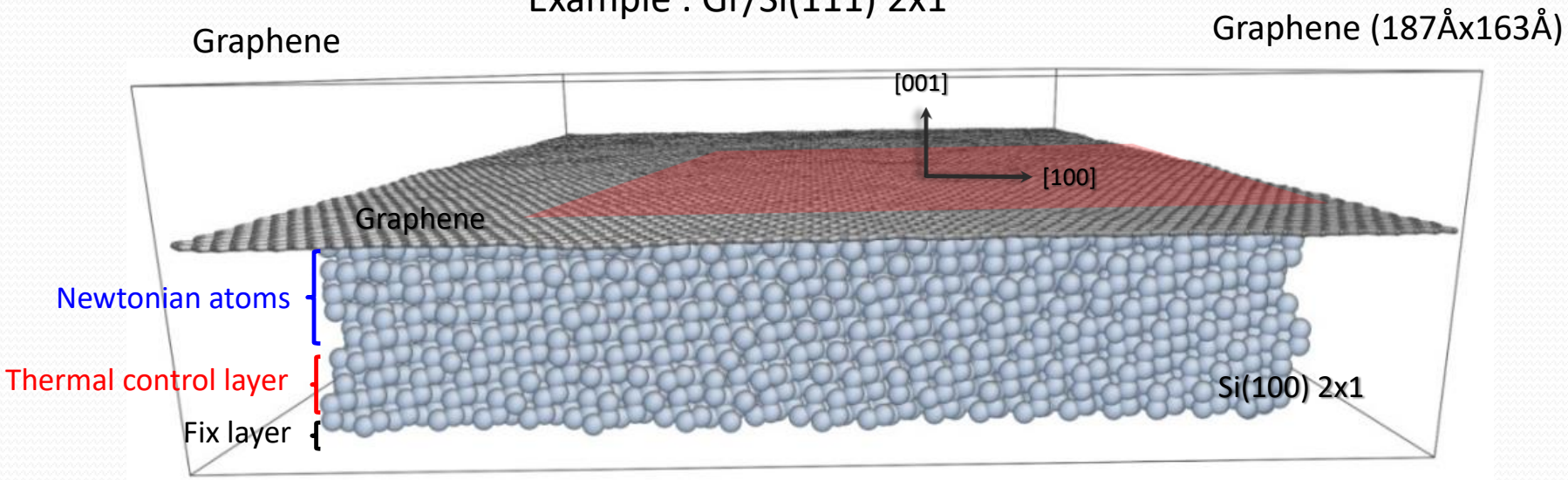
Modularity results



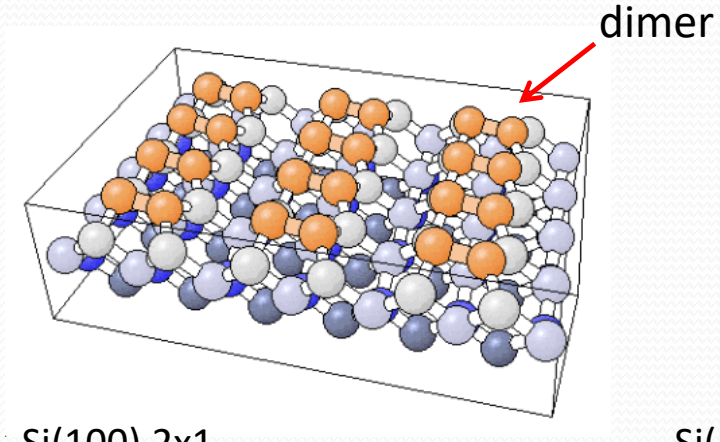
Simulation model

- Method: molecular dynamics
- Software: LAMMPS
- Potential function: C-C interaction ➔ AIREBO Potential
  Si-Si interaction ➔ Erhartand Albe Potential
  C-Si interaction ➔ Erhartand Albe Potential
- Temperature: from 10K to 1200K at an annealing rate of 5x1010 K/s.
- A strained Graphene flake cover onto the Si substrate.

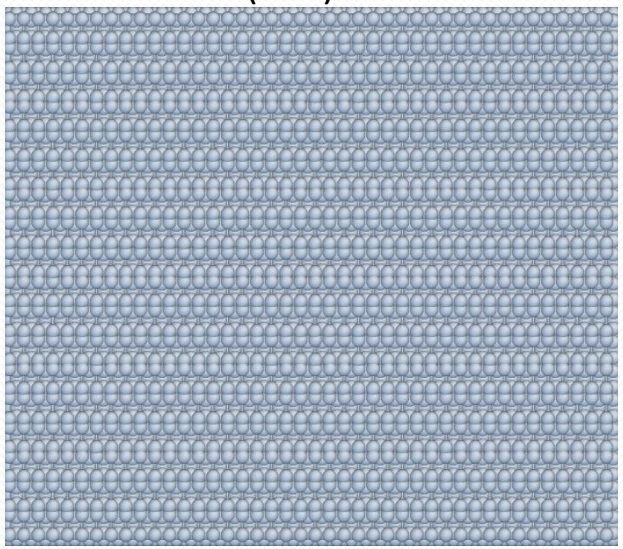
Example : Gr/Si(111) 2x1



Simulation model

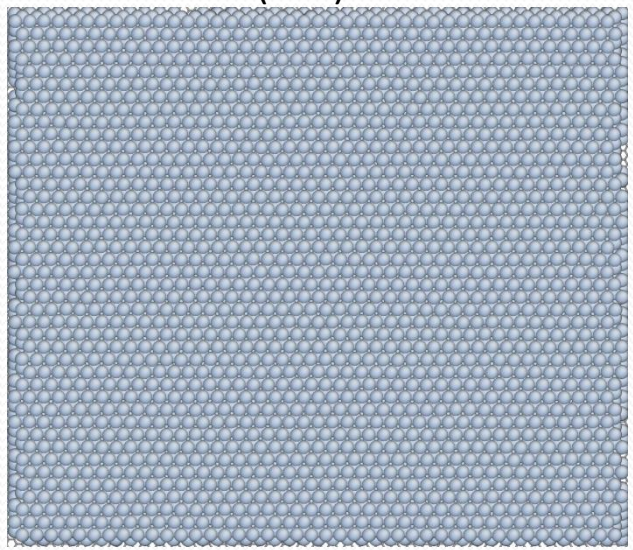


Si(100) 2x1

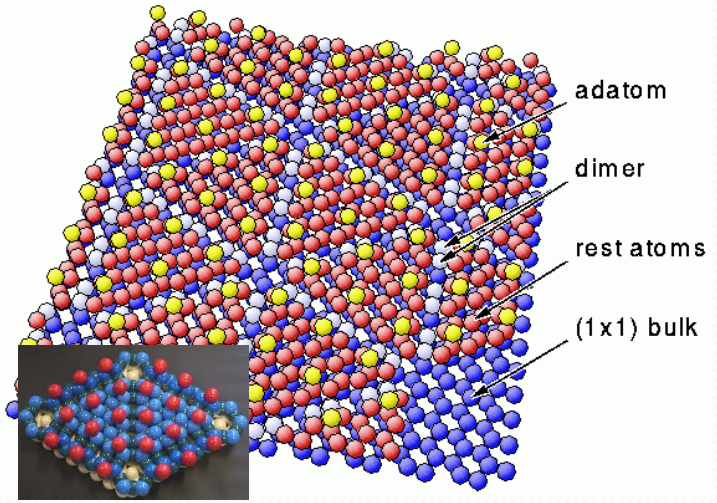


16.0x13.7 nm^2

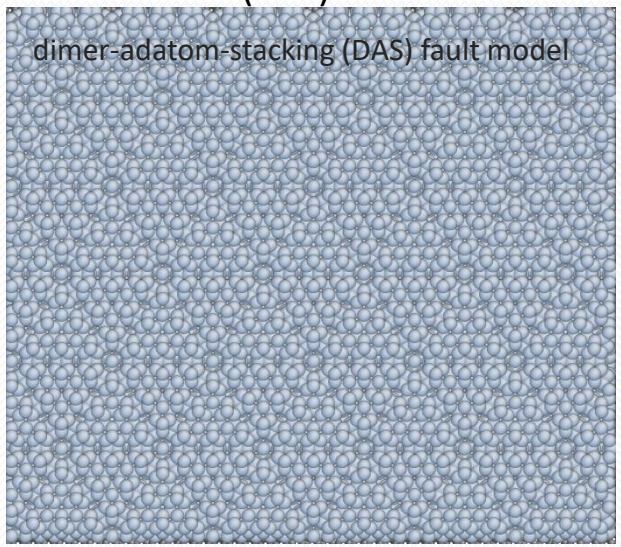
Si(111) 1x1



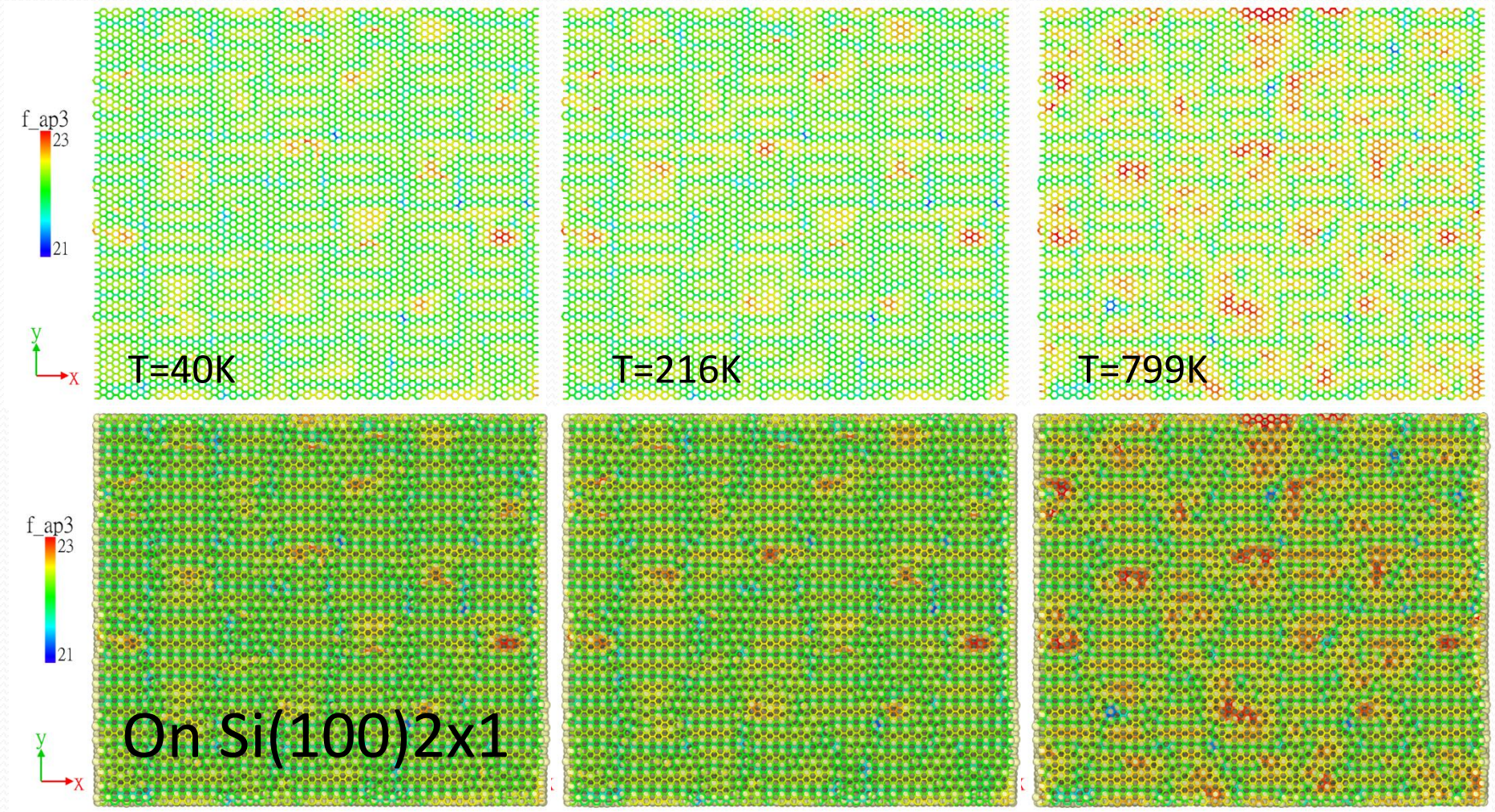
16.1x13.9 nm^2

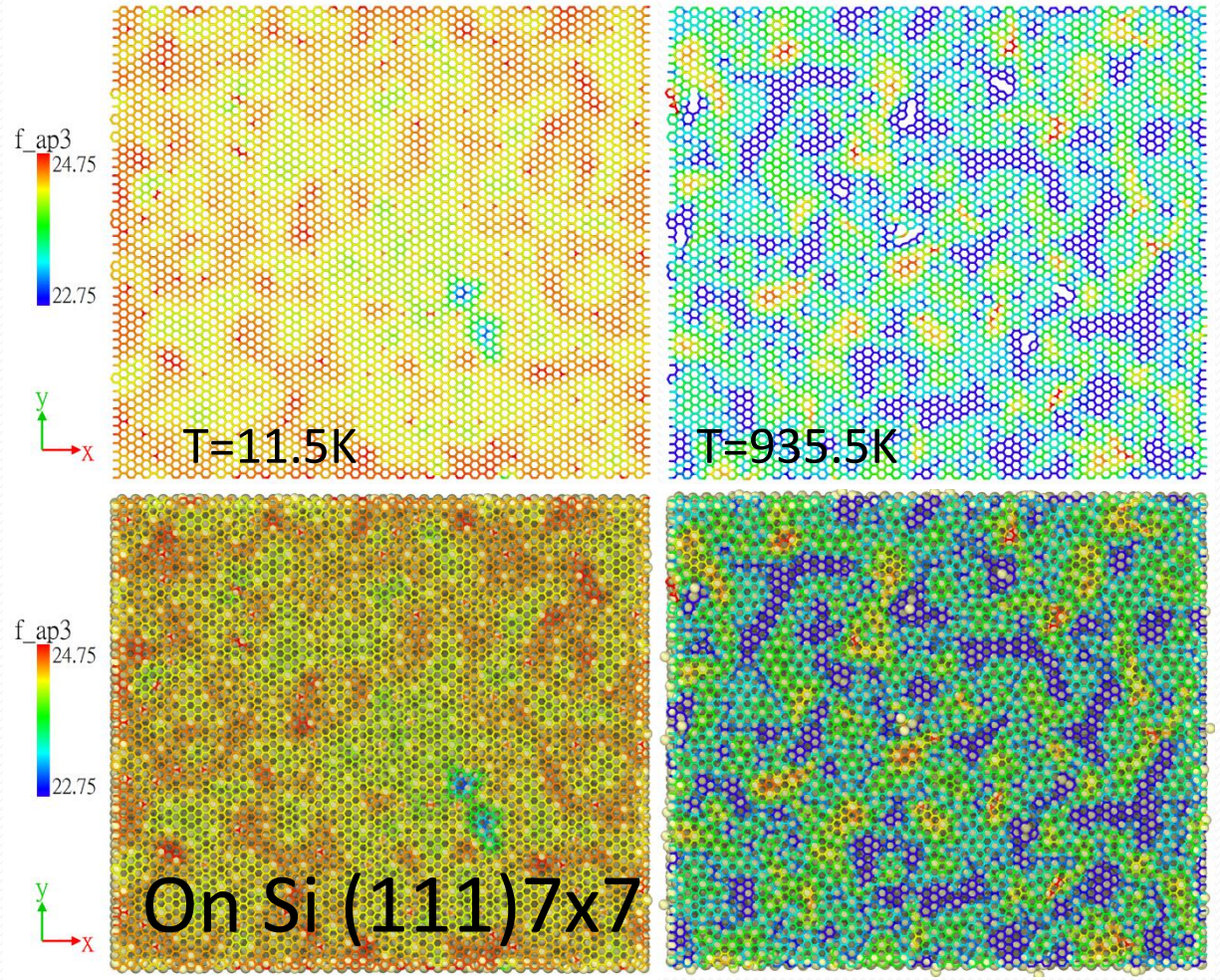


Si(111) 7x7



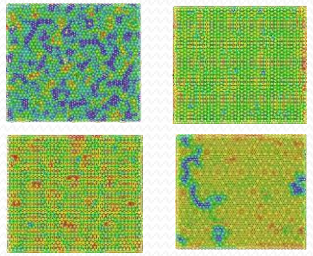
16.2x14.0 nm^2


f_ap3
23
21
y
x
T=40K
T=216K
T=799K
f_ap3
23
21
y
x
On Si(100)2x1

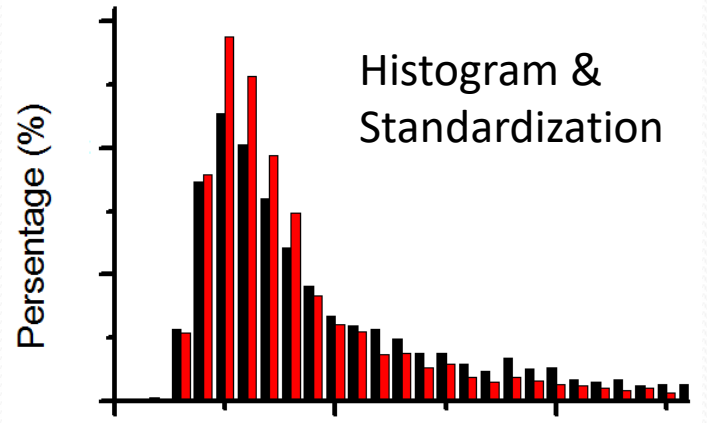

f_ap3
24.75
22.75
y
x
T=11.5K
T=935.5K
f_ap3
24.75
22.75
y
x
On Si (111)7x7

Vectorization

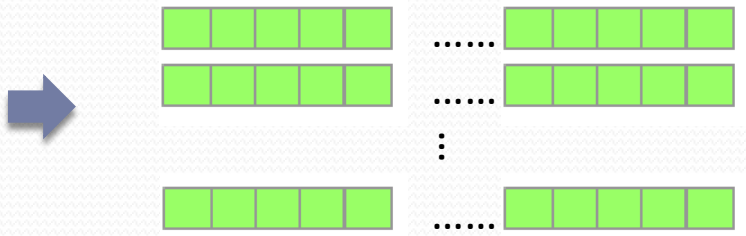
Simulation data set



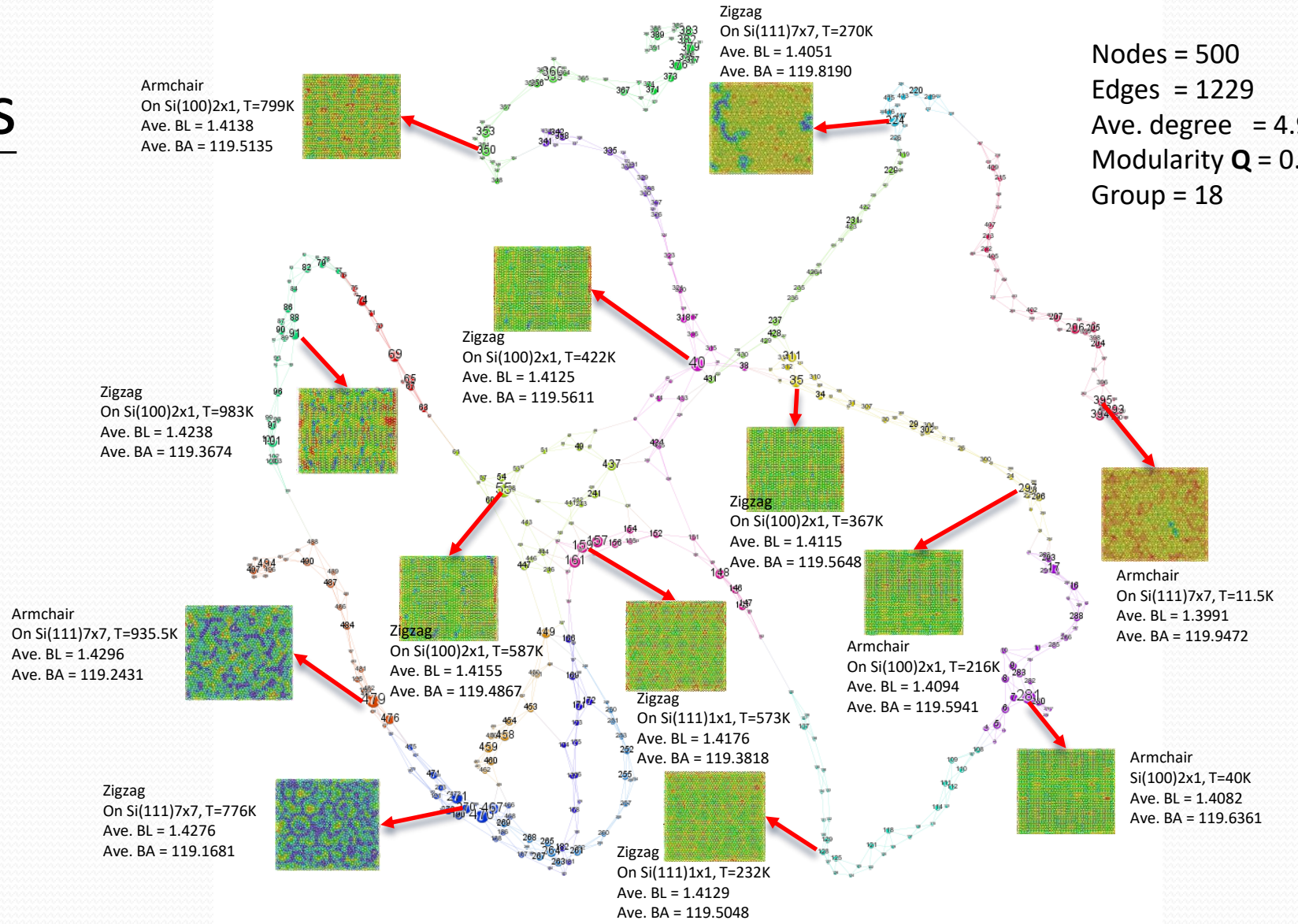
e.g. bond lengths, bond angles.



Characteristic Indexes vectors



Results



Nodes = 500
Edges = 1229
Ave. degree = 4.916
Modularity **Q** = 0.886
Group = 18

Summary

- The isomap method can defined the similarity between structures by geodesic paths in a high-dimensional manifold.
- The modularity method can find the best community structure of classification for structures by optimization method, i.e., to maximize the intra module connections as many as possible and to minimize the inter module connections as few as possible.
- Large-scale morphological structures, their annotations as well as quantified characteristics, and classifications can be facilely and reliably retrieved as useful data.

Next class

- CNN
- ……