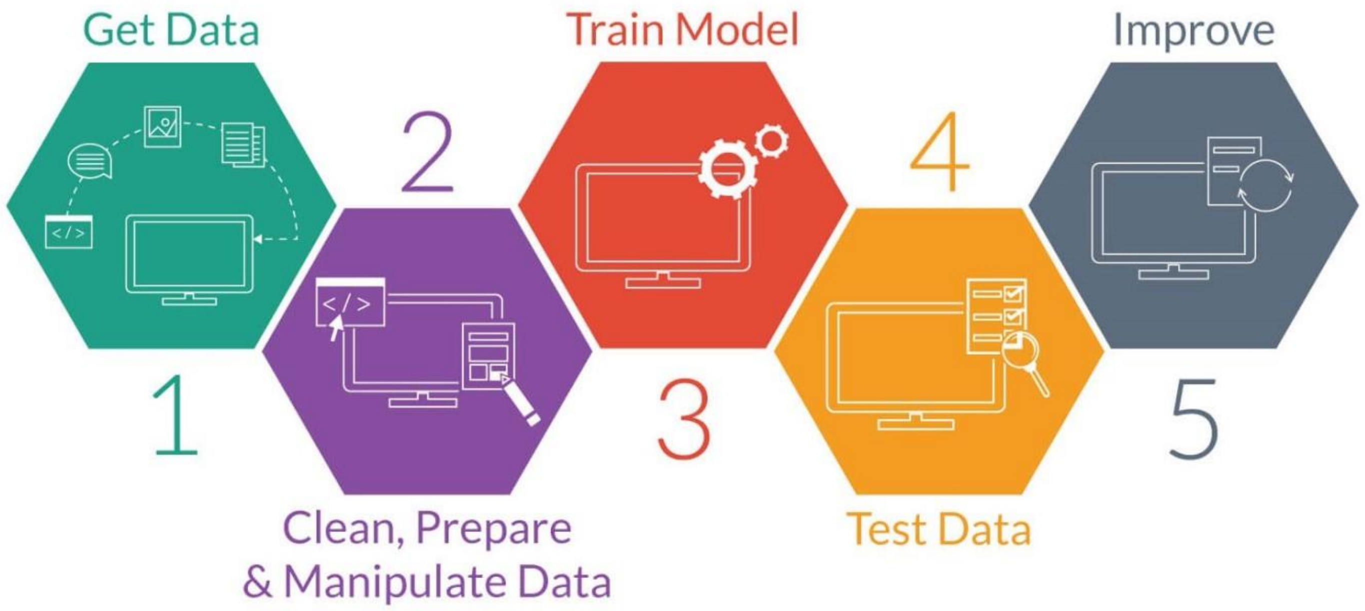
機器學習於材料資訊的應用 Machine Learning on Material Informatics

陳南佑(NAN-YOW CHEN)

nanyow@narlabs.org.tw

楊安正(AN-CHENG YANG)

acyang@narlabs.org.tw



使用軟體產生資料

LAMMPS

檔案處理

特徵萃取

NumPy

SciPy

MD ANALYSIS

ASE

建立網路

分群演算法

scikit learn

用測試資料檢驗演算法

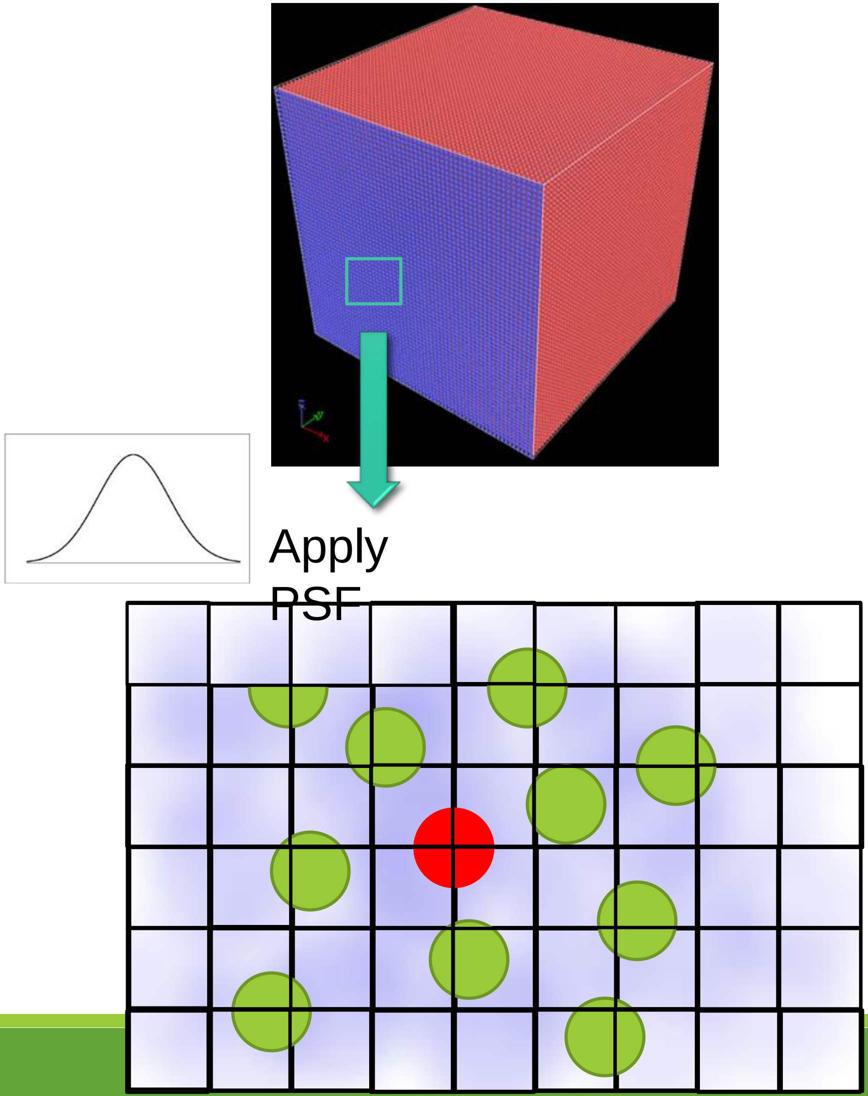
調整萃取特徵方法

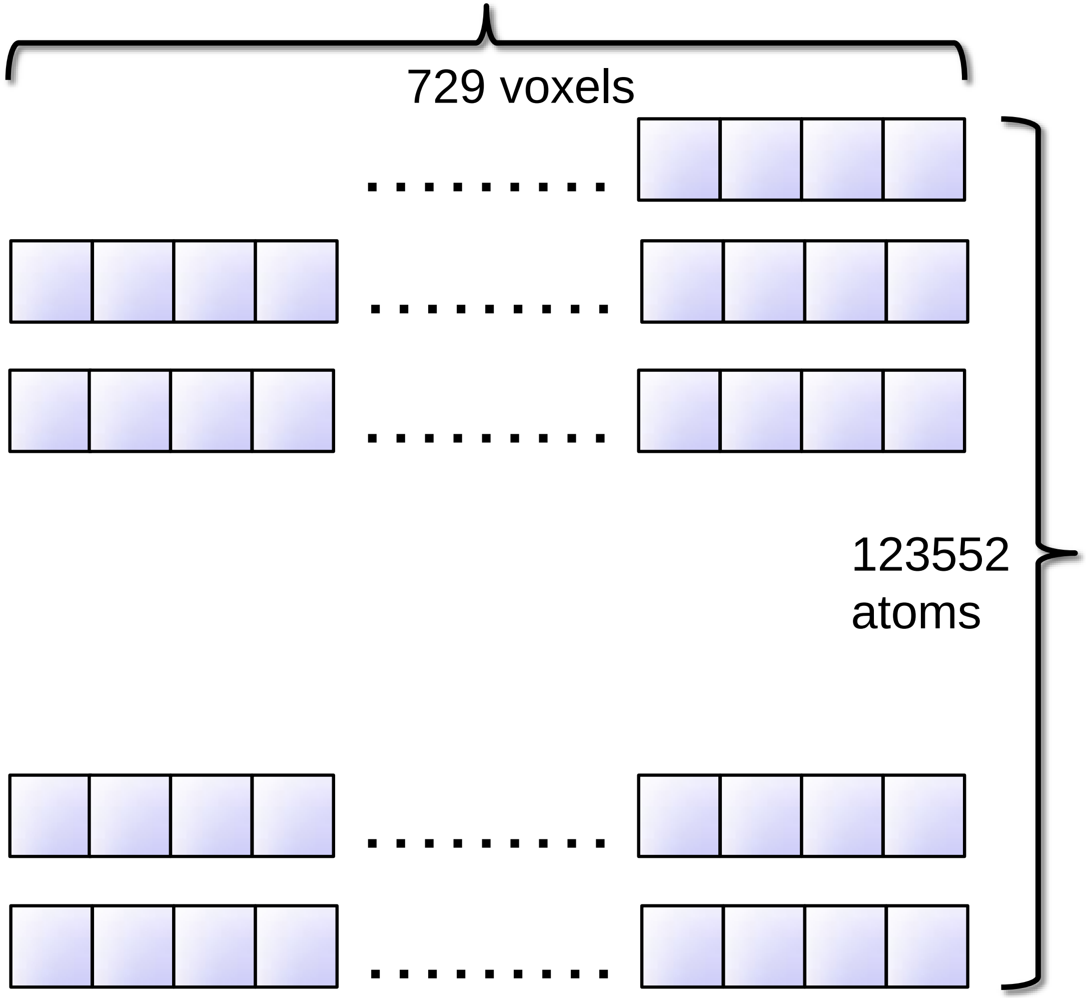
檔案處理&特徵萃取

- LAMMPS可以輸出dumpfile(cfg)，xyz，trajectory file(dcd)，自己打造Parser的話不用特別考慮。
  - dumpfile(cfg):在模擬過程中標準輸出檔案，ascii檔，人可以直接讀和編輯。 (MDANALSIS不能讀，需要用ovito轉成data檔)
  - xyz:輸出另一種檔案形式，ascii檔，人可以直接讀和編輯。(MDANALSIS可以讀，但是缺了mass info。)
  - trajectory file:輸出另一種檔案形式，binary檔，人不能直接解讀。(MDANALSIS可以讀，ovito讀不了。)
- 特徵萃取:結構分類問題是局部的，考慮一個原子的特徵，需要從與周遭原子的關係下手。
  - Voxelize local region (This class)
  - Local environment
  - Nearest neighbor

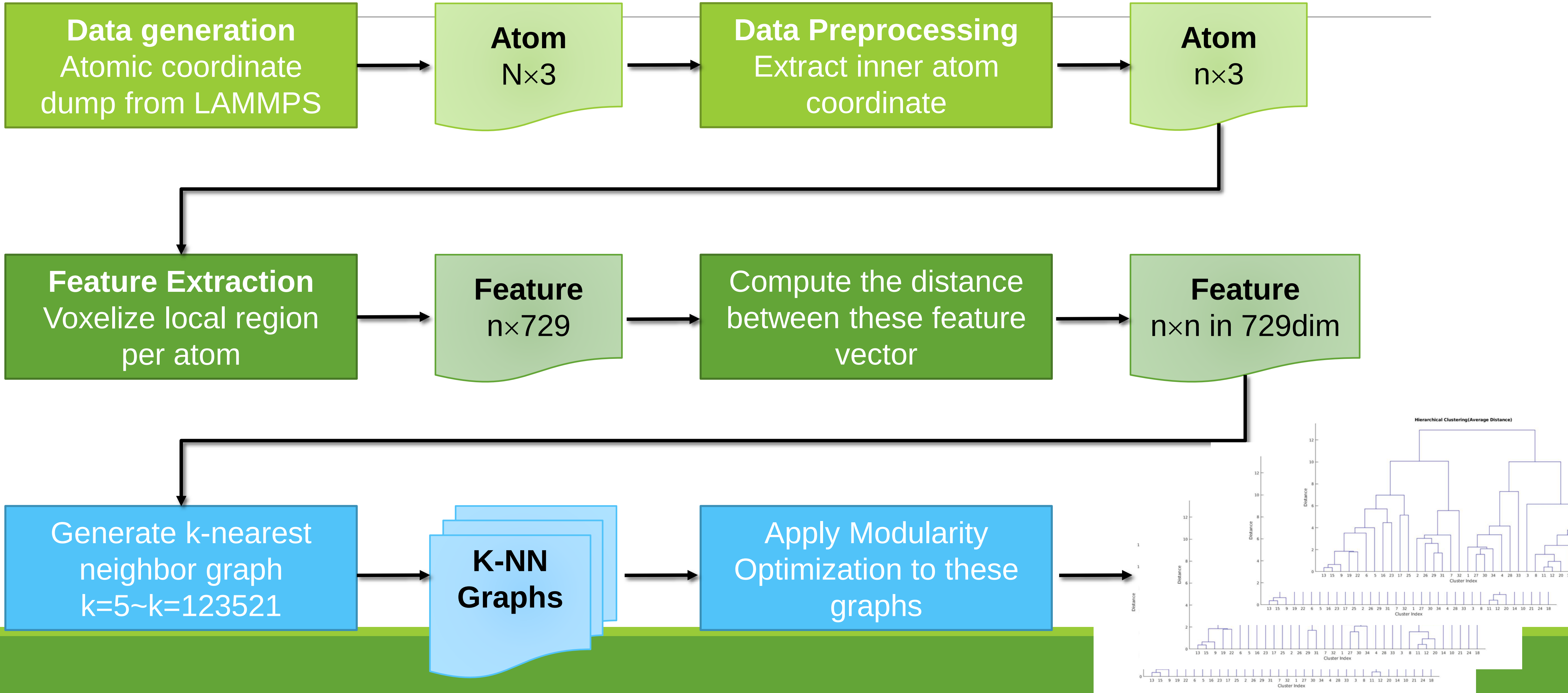
Feature of local environment

```
>python feature_engineering.py
```





Flowchart



Data generation
Atomic coordinate dump from LAMMPS

Atom
N×3

Data Preprocessing
Extract inner atom coordinate

Atom
n×3

Feature Extraction
Voxelize local region per atom

Feature
n×729

Compute the distance between these feature vector

Feature
n×n in 729dim

Generate k-nearest neighbor graph
k=5~k=123521

K-NN Graphs

Apply Modularity Optimization to these graphs

More on descriptors

- Atom-centered Symmetry Functions (ACSF)
- Smooth Overlap of Atomic Positions (SOAP)
- Gaussian descriptor
- Behler type Symmetry function(目前最多人採用)
- …

分群演算法

- 原子結構的分類，避免直接從原子座標(卡式座標系統)進行分類，反而是從原子的其他座標系統去挑選特徵(座標系統的基底)，來描述原子的local environment。
- 有了原子的local environment，便可以在這些座標系統進行原子的分類，scikit-learning已提供多種現成的分群演算法可以使用。
  - K-means
  - Affinity Propagation
  - Hierarchical clustering
  - DBSCAN
- 雖然已經有多種分群演算法可以直接使用，但無可避免的是這些方法都還是需要人來挑選演算法參數，容易淪為先射箭再畫靶。
- 所以我們嘗試導入網路分析(network analysis)中的Modularity方法來進行微結構分類，最大的不同在於Modularity是非監督式學習，不需要人工決定分群的參數，比起其他方法要來的客觀。

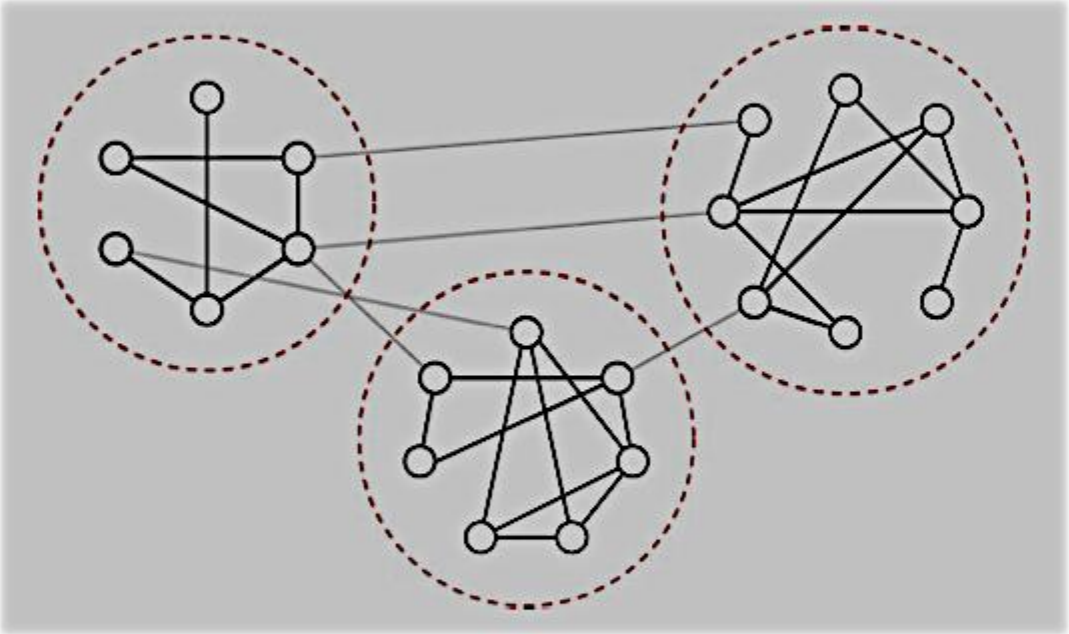
https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

Modularity of networks

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

- Partitioning a ne are similar to each other and are as differe

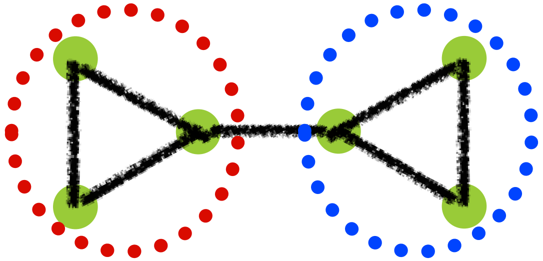
- In order to desc ne a similarity measure and we also need a s



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

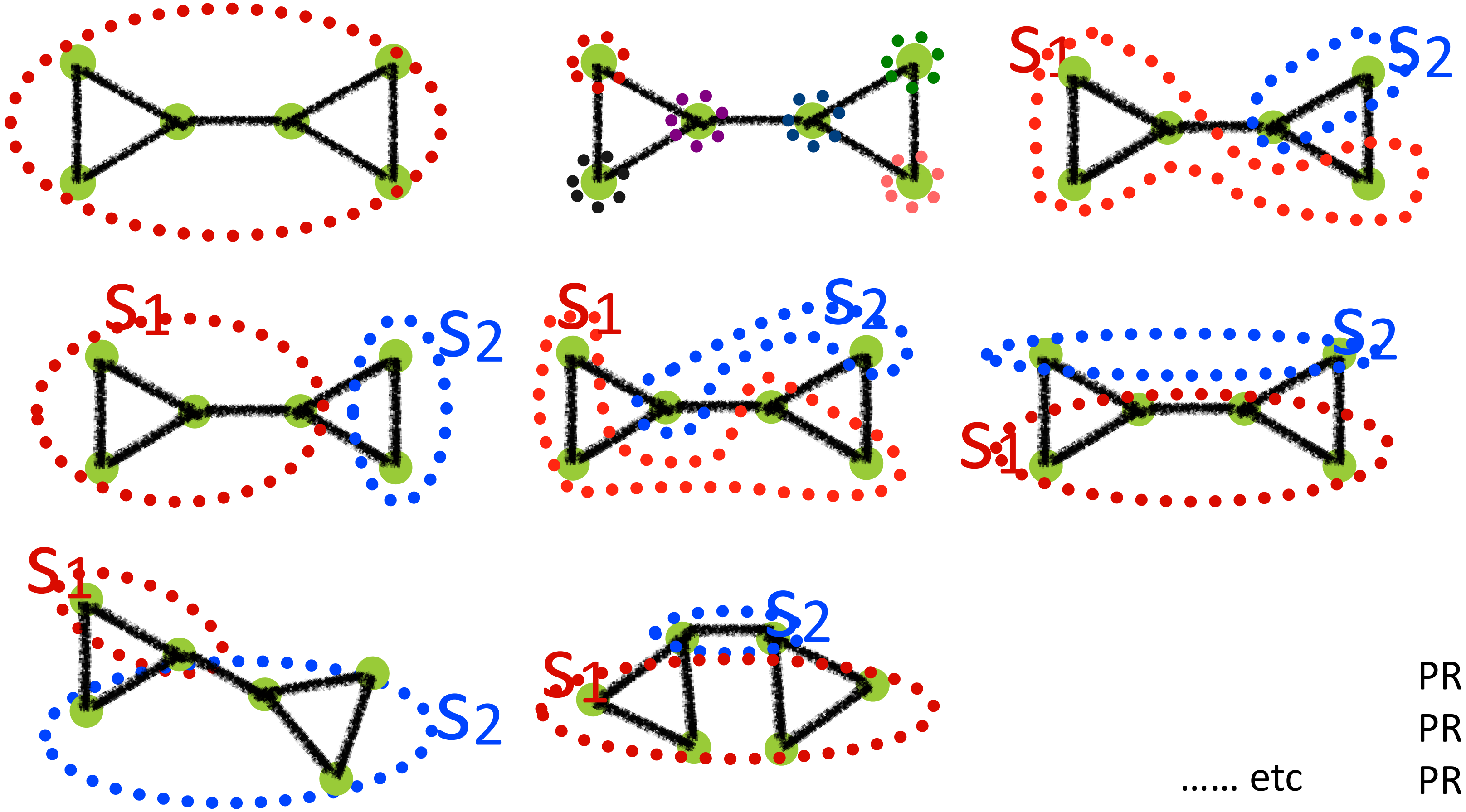
d_1=7, d_2=7

L=7

- Q = 0.357

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

Modularity of networks



...... etc

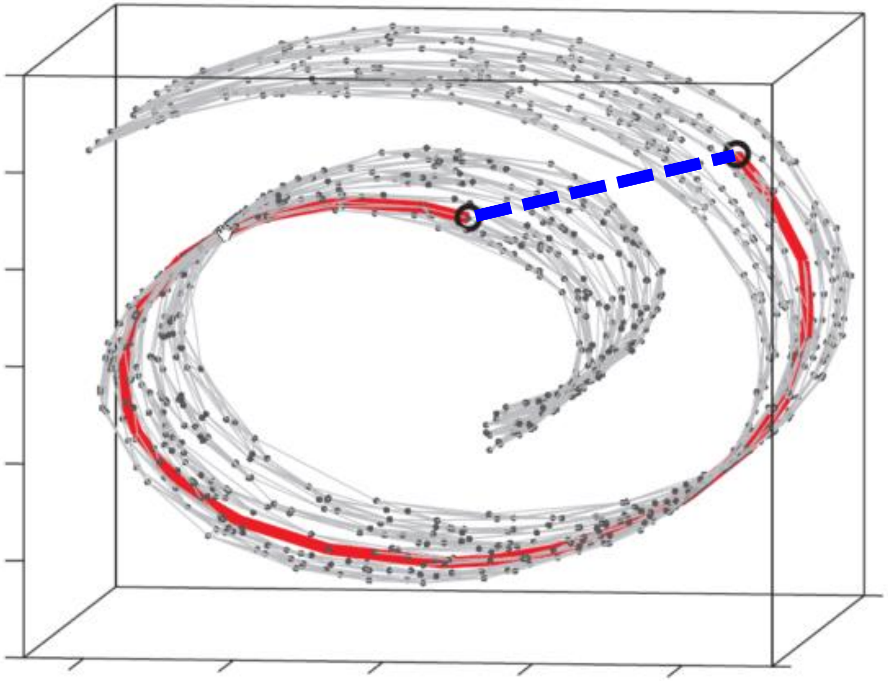
PRE **68**, 026121 (2003)
PRE **69**, 026113 (2004)
PRE **70**, 025101 (2004)

建立網路

- Modularity是基於網路分析的方法，所以比起其他分群演算法，需要多一個建立資料點的網路關係。
- 網路式建立於資料點的空間，不是原本問題的卡式座標。
- 距離的定義有許多種，歐式距離、曼哈頓距離、 Dijkstra distances …
- 網路連通的定義也要選擇。

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.



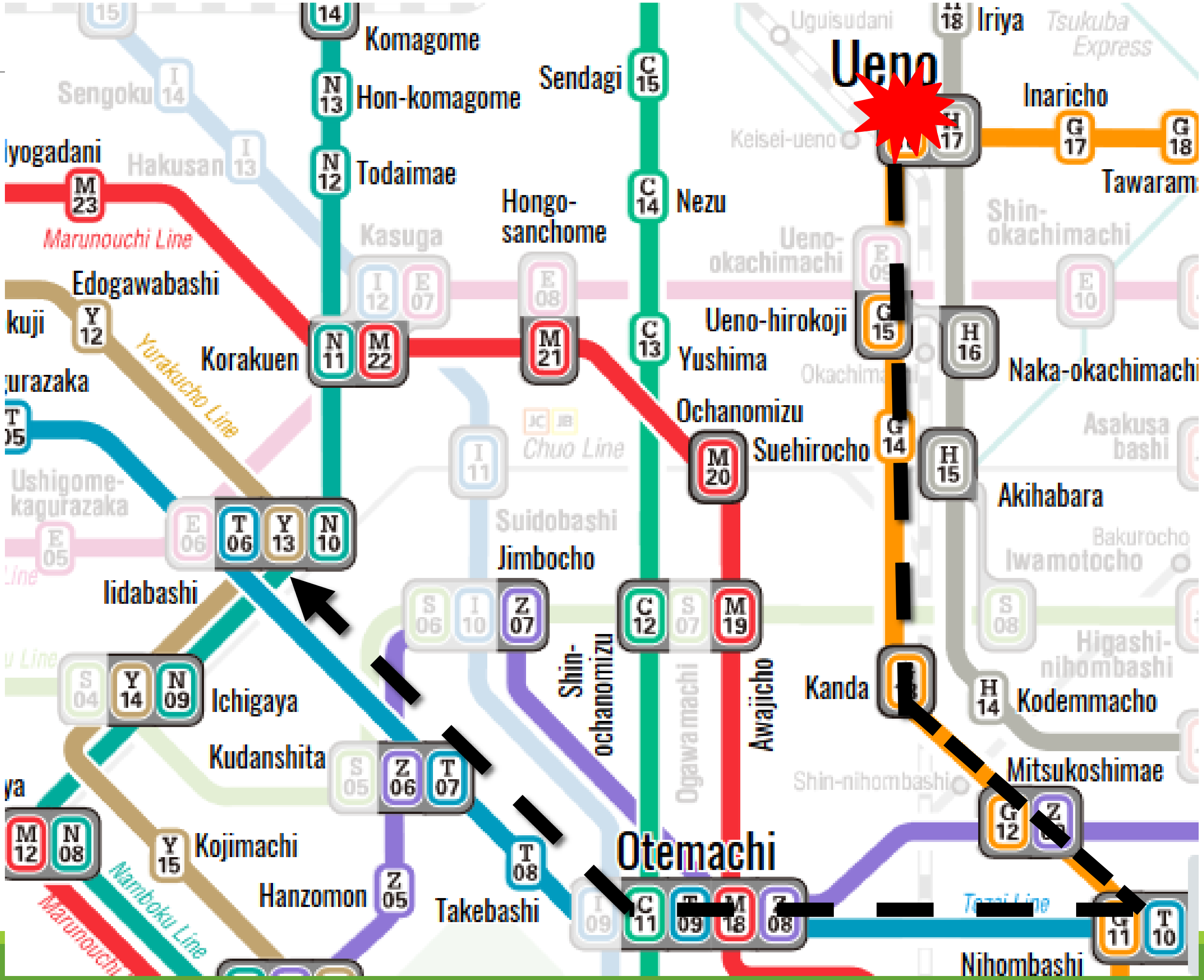
For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high dimensional input space (length of blue dashed line) may not accurately reflect their intrinsic similarity.

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.



Ex. Ueno and Iidabashi

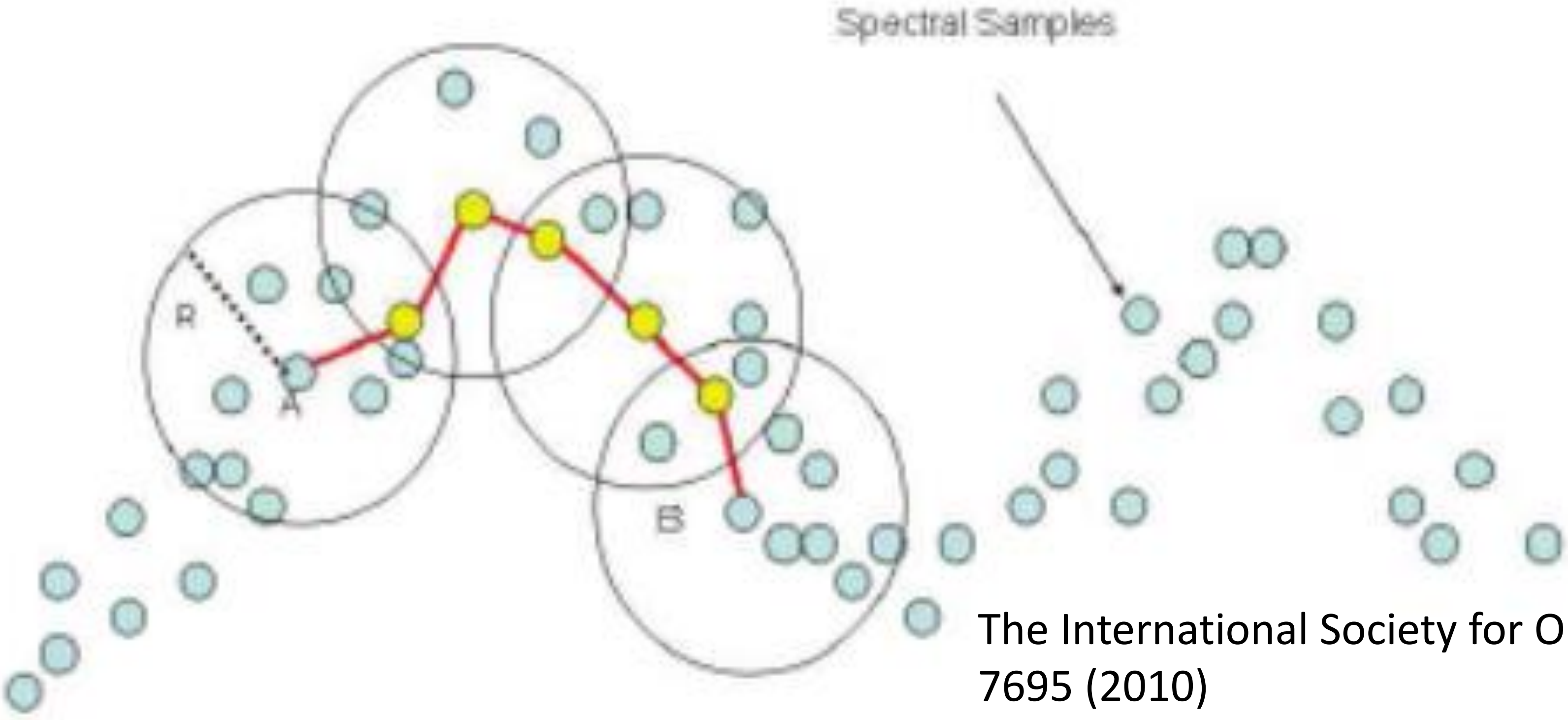
- Euclidean distances : 3.2km
- Dijkstra distances : 3.4+3.7=7.1km

(Ueno → Nihonbashi→Iidabashi)

Construct Networks

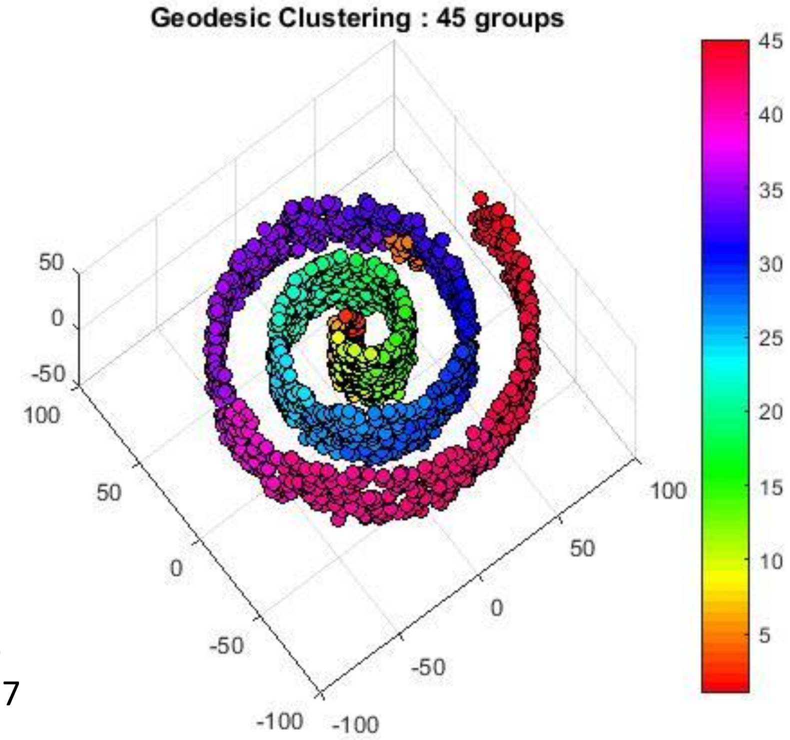
□ Steps:

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



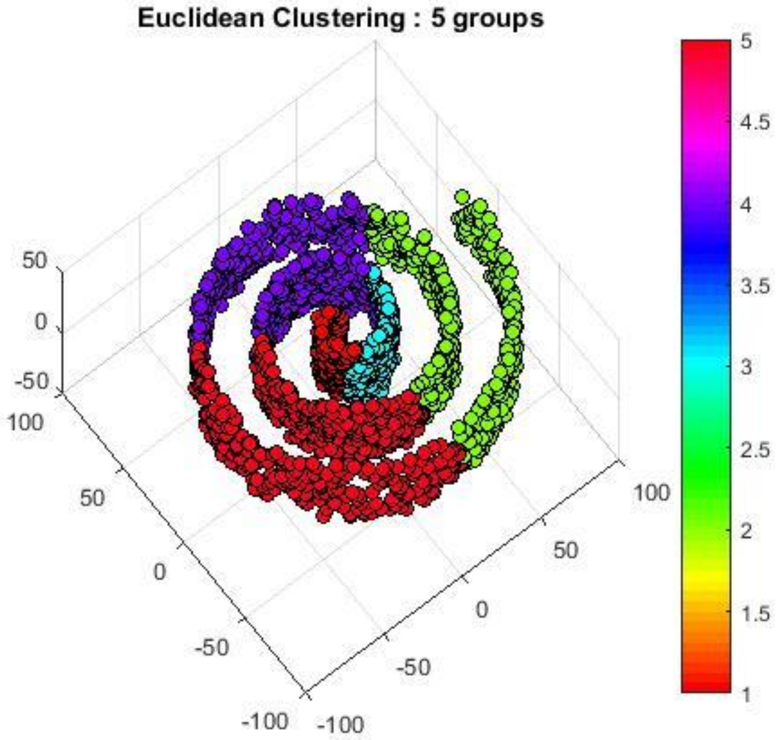
The International Society for Optical Engineering 7695 (2010)

Testing case - swiss roll manifold



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)