中国传媒大学

深度学习与类脑计算

（五）

曹立宏

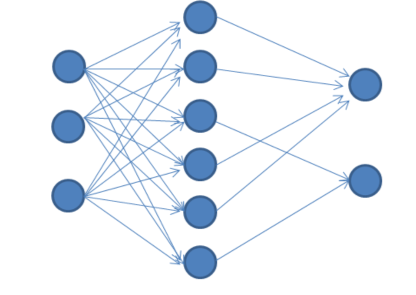
脑科学与智能媒体研究院

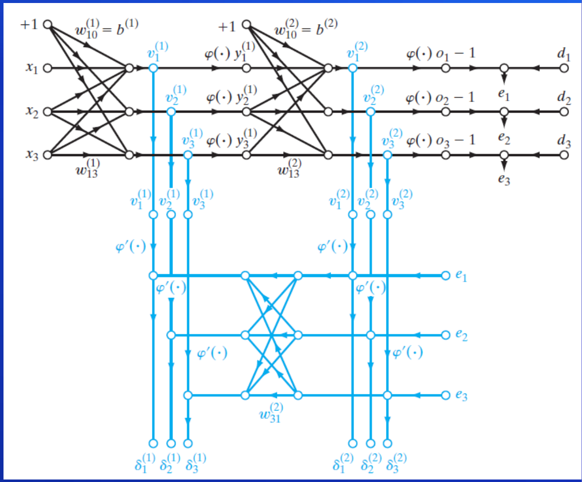
Review

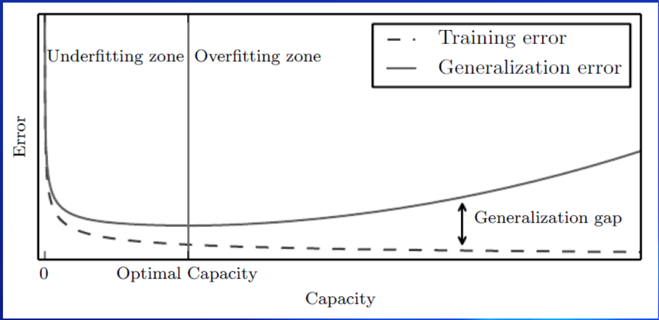
- Universal Approximation Theorem
- MLP with BP

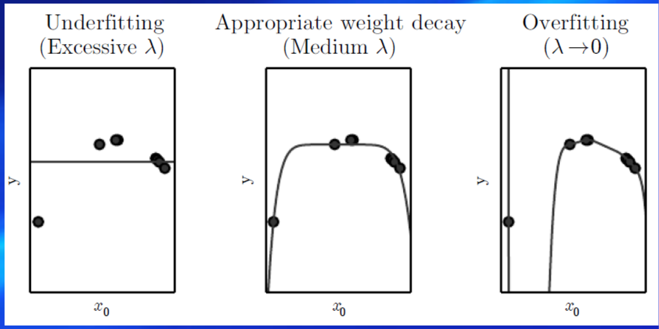
Overfitting Underfitting Curse of Dimensionality

- Training error – Testing error – MSE
- Unbiased - variance tradeoff
- Regularization
- Maximum Likelihood - MSE
- Maximum A Posteriori (Baysian)
- Lagrange Multiplier - KKT



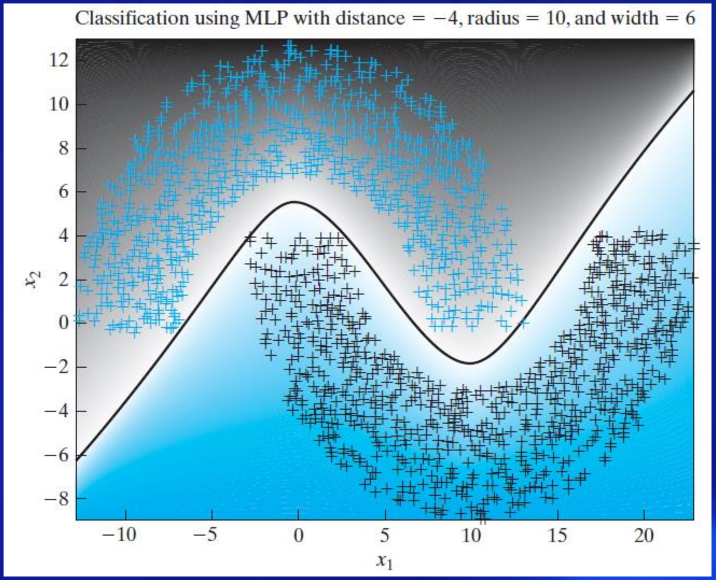

$+1$
$w_{10}^{(1)} = b^{(1)}$
$+1$
$w_{10}^{(2)} = b^{(2)}$
x_1
x_2
x_3
$w_{13}^{(1)}$
$w_{13}^{(2)}$
$v_1^{(1)}$
$v_2^{(1)}$
$v_3^{(1)}$
$\varphi(\cdot)\ y_1^{(1)}$
$\varphi(\cdot)\ y_2^{(1)}$
$\varphi(\cdot)\ y_3^{(1)}$
$v_1^{(2)}$
$v_2^{(2)}$
$v_3^{(2)}$
$\varphi(\cdot)\ o_1\ -1$
$\varphi(\cdot)\ o_2\ -1$
$\varphi(\cdot)\ o_3\ -1$
d_1
d_2
d_3
e_1
e_2
e_3
$\varphi'(\cdot)$
$w_{31}^{(2)}$
$\delta_1^{(1)}\ \delta_2^{(1)}\ \delta_3^{(1)}$
$\delta_1^{(2)}\ \delta_2^{(2)}\ \delta_3^{(2)}$

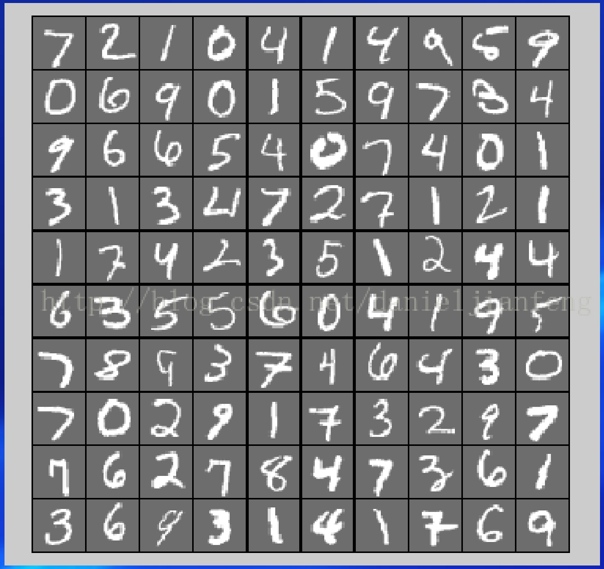




Homework

- Program BP algorithm in Python
- Pattern classification (ref. p150-153 on Haykin)
- Change the activation function of hidden layers to ReLU
- The MNIST Database （陈雯婕）





Homework

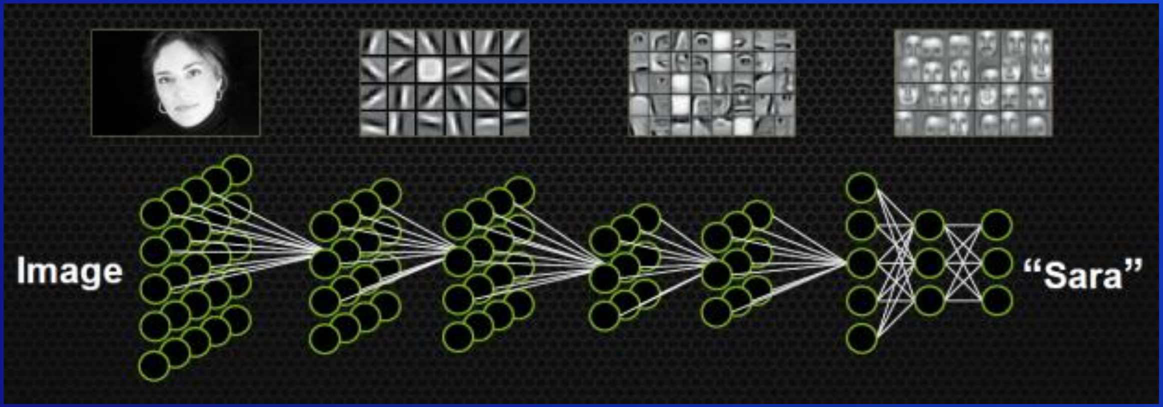
- Finishing up the last homework
- How to write a technical report?
- Prove: Bayesian or MAP estimation for linear regression is equivalent to regularization of MSE
- We（陈雯婕）will give you a dataset RegData2D

$$\{x_i, y_i \mid_{i=1,2,\ldots,N,N+1,\ldots N+M}\}$$

use whatever you have learnt to find the best relationship between x and y. Write a report, and specifically pay attention to underfitting and overfitting and how you come up with your best solutions.

- (optional) We（陈雯婕）will give you another dataset RegData3D, $\{x_i, y_i, z_i \mid_{i=1,2,\ldots,N,N+1,\ldots N+M}\}$ do the same.

From Regression to Categorization



- Final Layer for categorization
- Discrete responses
- Supervised Learning
  - Logistic Regression
  - SVM
- Unsupervised Learning
  - Dimension Reduction
  - K-mean Clustering
  - Kernel Method
  - RBF Network

Categorization – 分类

- 相对图像维度
  - 自然界的物体种类似乎很有限！
  - 概念也很有限！

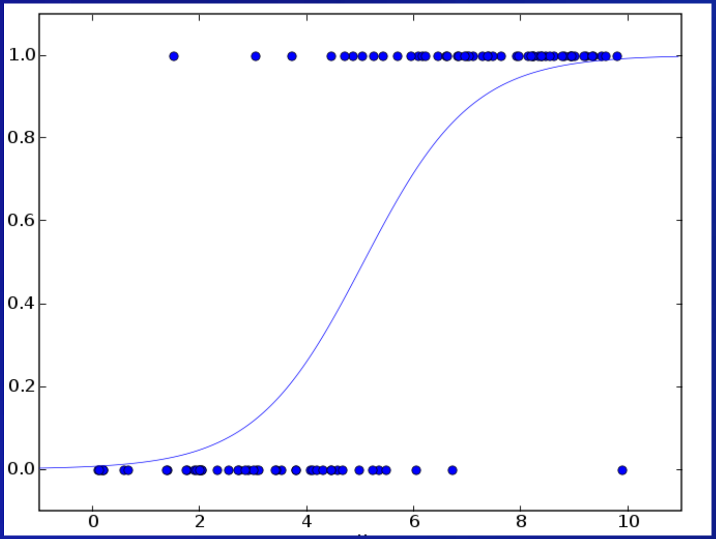
离散与连续的本质区别

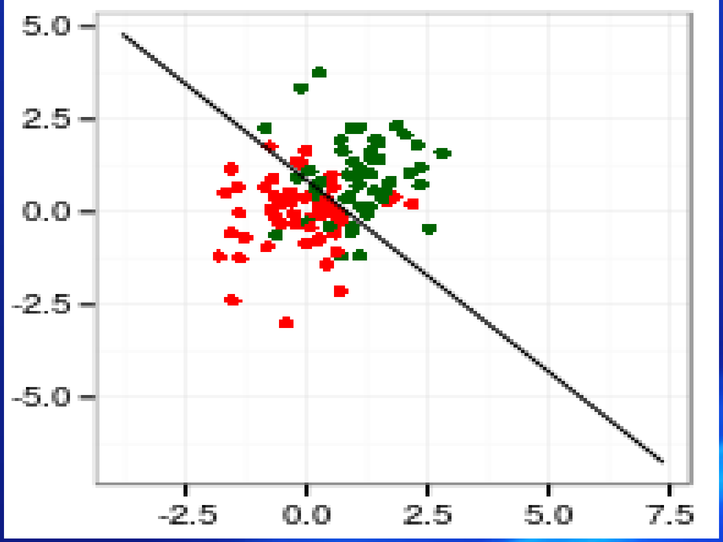
- 离散没有次序和程度可言
- 回归的本质是连续性假设

Which does not belong?

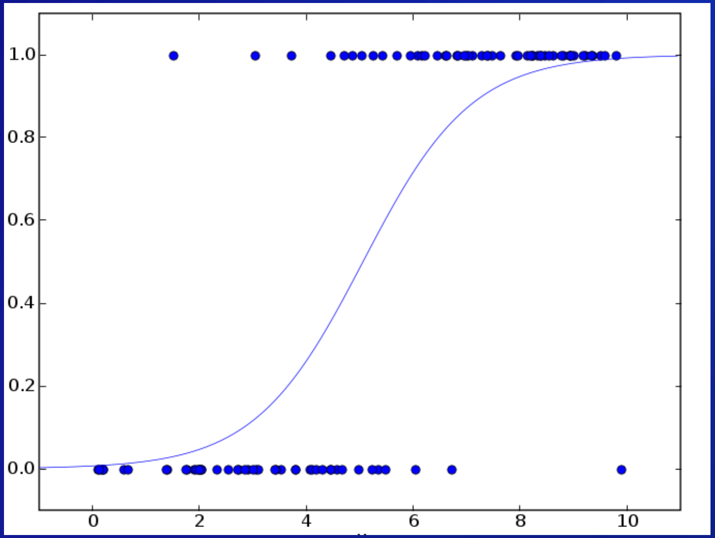
Logistic Regression

- Sex predicted by Weight
- Sex predicted by Weight and Height





Logistic Regression



$$p(x|y=1) = \frac{1}{1+e^{-x}}$$

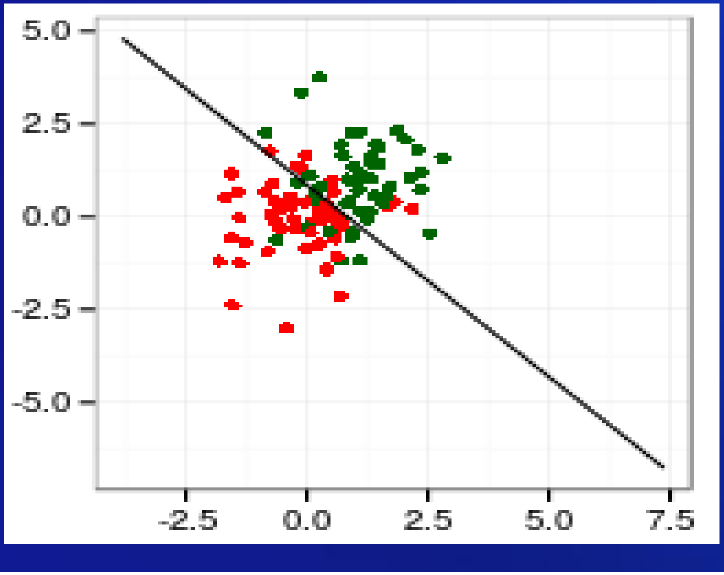
$$p(x|y=1) = \frac{1}{1+e^{-(ax+b)}}$$

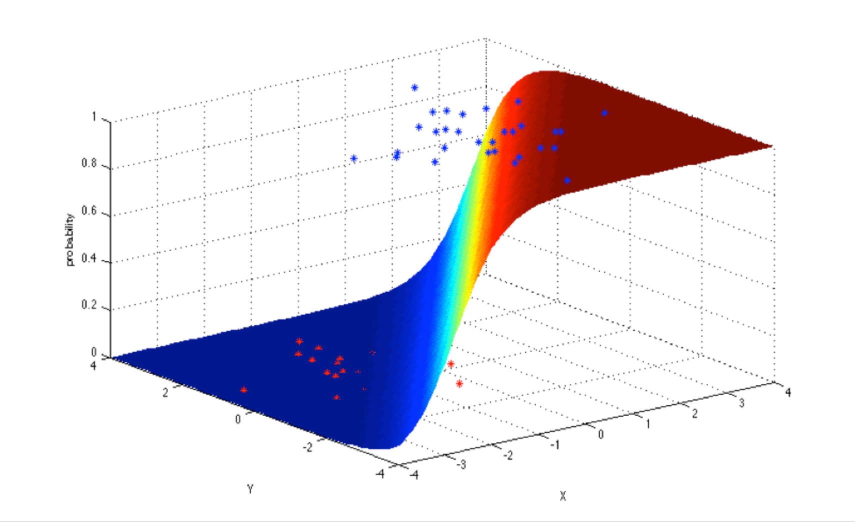
$$p(x|y=0) = \frac{e^{-(ax+b)}}{1+e^{-(ax+b)}}$$

ML Estimation:

MAP Estimation:

Logistic Regression –多变量



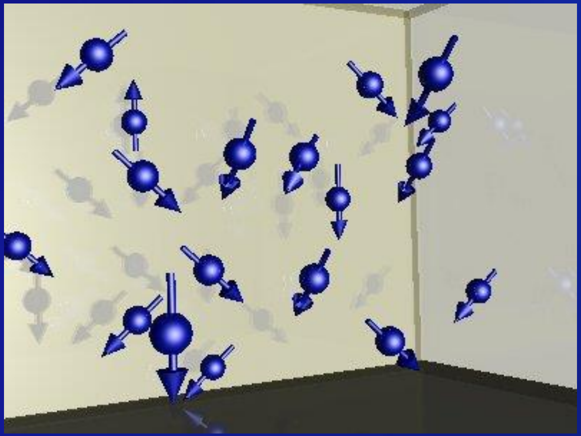


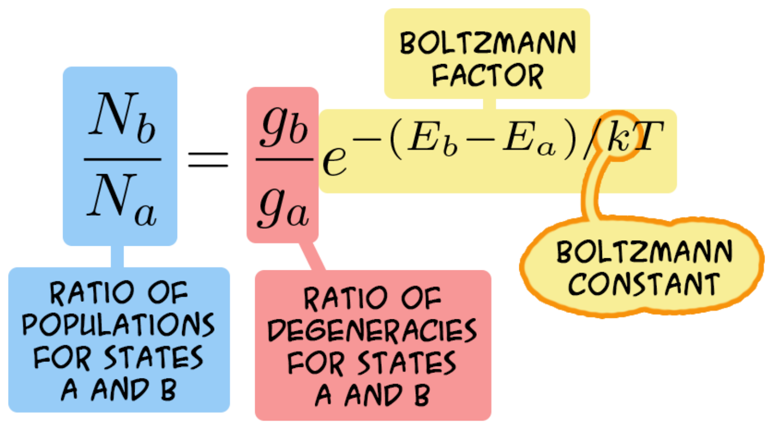
$$p(\vec{x}|y=1)=\frac{1}{1+e^{-\vec{w}\vec{x}}}$$

$$p(\vec{x}|y=0)=\frac{e^{-\vec{w}\vec{x}}}{1+e^{-\vec{w}\vec{x}}}$$

Logistic Regression-多态

Boltzmann Theory



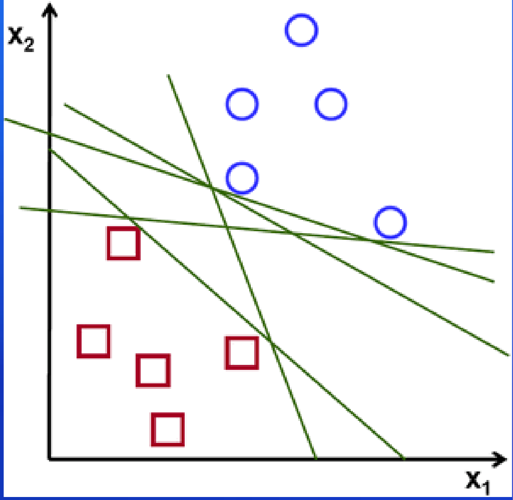

BOLTZMANN FACTOR
$\frac{N_b}{N_a} = \frac{g_b}{g_a} e^{-(E_b - E_a)/kT}$
BOLTZMANN CONSTANT
RATIO OF POPULATIONS FOR STATES A AND B
RATIO OF DEGENERACIES FOR STATES A AND B

$$p(\vec{x}|y = \mathrm{i}) \propto \pi_i e^{-\vec{w}_i \vec{x}}$$

$$\vec{w}_i \vec{x} = E_i$$

归一化条件

SVM-Support Vector Machine



- Given labeled data

$$\{\vec{x}_i, d_i, i = 1, \dots, N, d_i = \pm 1\}$$

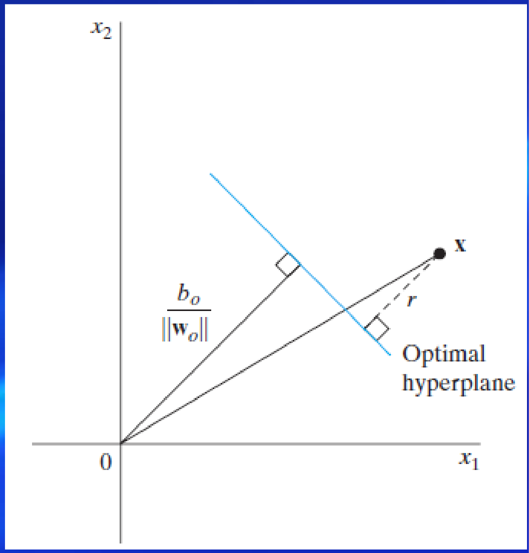
- find separation Line/Plan

$$\vec{w}^T \vec{x} + \mathrm{b} = 0$$

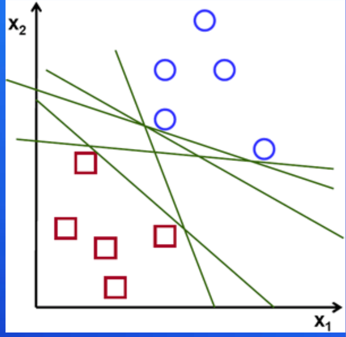
- Optimal one

$$\vec{w_0}^T \vec{x} + b_0 = 0 \qquad \vec{w_0}^T \vec{x}_{support} + b_0 = \pm 1$$

- Distance from Point to the Line/Plan



$$r(\vec{x}) = \frac{\vec{w_0}^T \vec{x} + b_0}{\|\vec{w_0}\|} \qquad r(\vec{x}_{support}) = \frac{\pm 1}{\|\vec{w_0}\|}$$



SVM-Support Vector Machine

- Optimization Goal:

  Minimization: $||\vec{w}_0||$

  Under conditions: $d_i({\vec{w}_0}^T \vec{x}_i + b_0) \geq 1$

- Lagrangian function:

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\vec{w}^T\vec{w} - \sum_{i=1}^{N} \alpha_i[d_i(\vec{w}^T\vec{x}_i + b) - 1] \qquad \alpha_i \geq 0$$

$$\frac{\partial L(\vec{w}, b, \alpha)}{\partial \vec{w}} = 0 \Longrightarrow \vec{w} - \sum_{i=1}^{N} \alpha_i d_i \vec{x}_i = 0$$

$$\frac{\partial L(\vec{w}, b, \alpha)}{\partial x} = 0 \Longrightarrow \sum_{i=1}^{N} \alpha_i d_i = 0$$

$$L(\vec{w}, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_i d_i d_i {\vec{x}_i}^T \vec{x}_j \qquad \alpha_i \geq 0$$

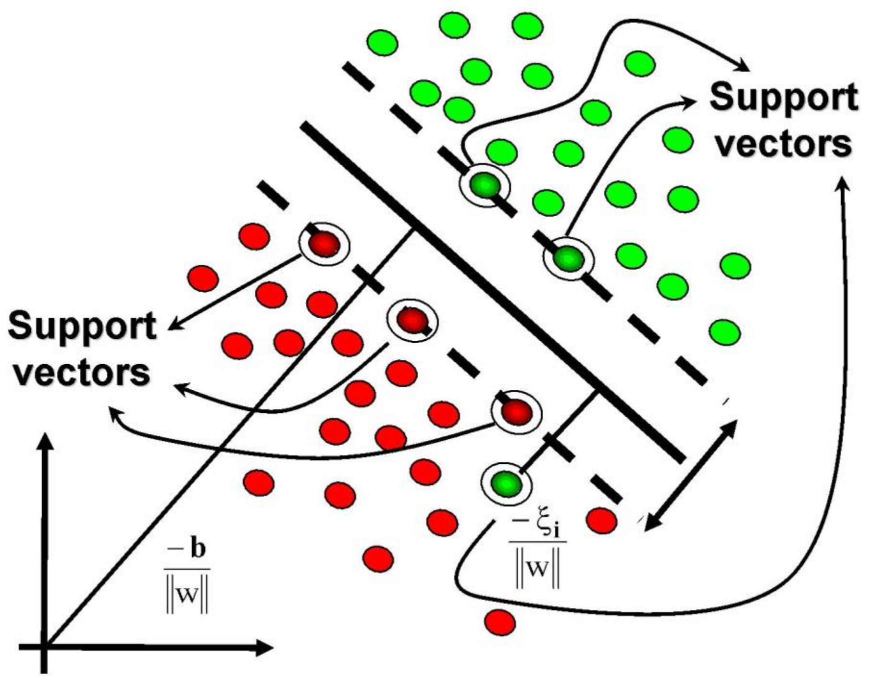
SVM-Support Vector Machine

- Separation Line/Plan
  - Only support vectors contribute
  - Don’t care about non-support data

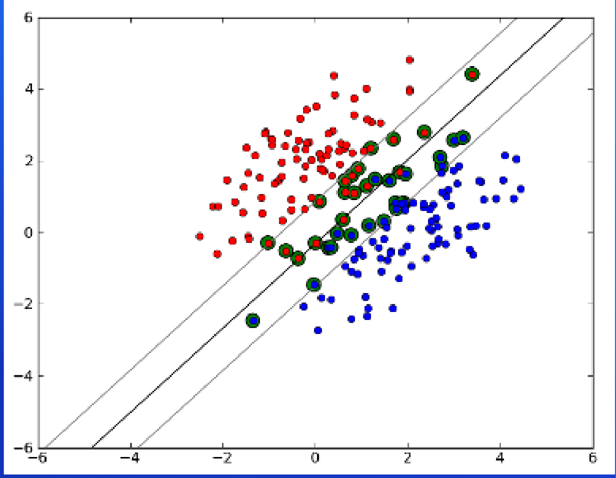
$$\vec{w} - \sum_{i=1}^{N} \alpha_i d_i \vec{x}_i = 0$$

$$\vec{w}_0 = \sum_{i=1}^{N} \alpha_{i,0} d_i \vec{x}_i$$

$$\alpha_{i,0} > 0$$

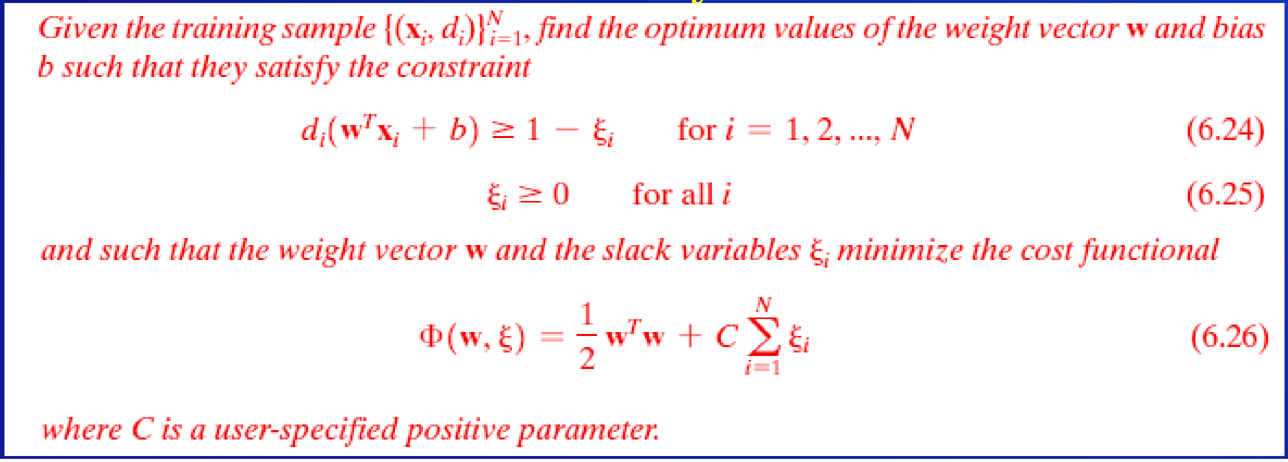


SVM
for nonseparable data



- Introducing *slack variables* $\xi_i \geq 0$

$$d_i(\vec{w}_0{}^T\vec{x}_i + b_0) \geq 1 - \xi_i$$



Given the training sample $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, find the optimum values of the weight vector $\mathbf{w}$ *and bias b such that they satisfy the constraint*

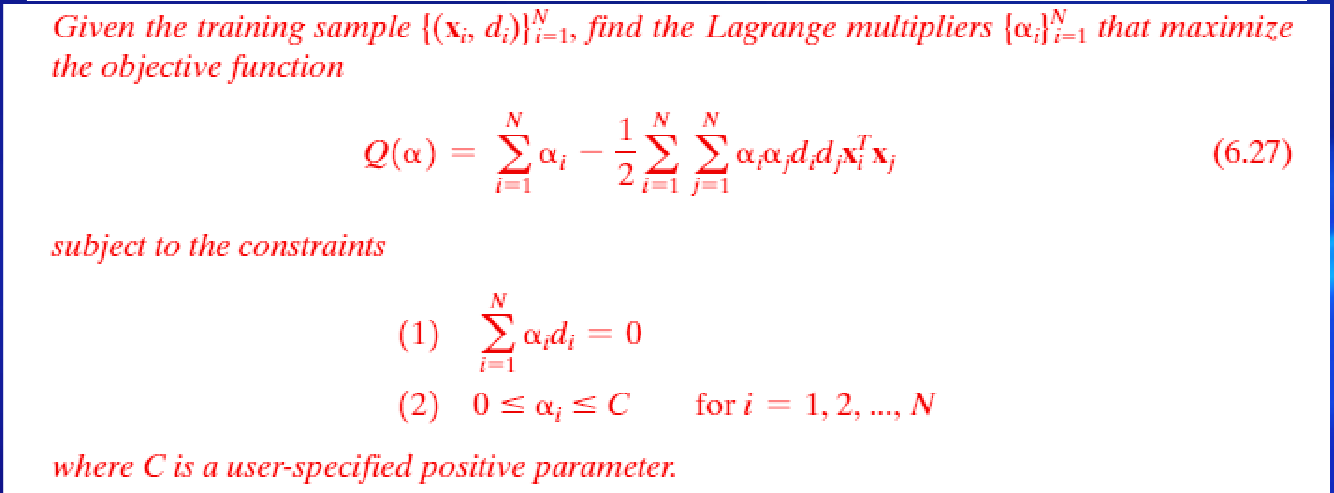
$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \qquad \text{for } i = 1, 2, \ldots, N \tag{6.24}$$

$$\xi_i \geq 0 \qquad \text{for all } i \tag{6.25}$$

and such that the weight vector $\mathbf{w}$ *and the slack variables* ξ_i *minimize the cost functional*

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \tag{6.26}$$

where C is a user-specified positive parameter.



Given the training sample $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, find the Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$ that maximize the objective function

$$Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j \tag{6.27}$$

subject to the constraints

(1) $\sum_{i=1}^{N}\alpha_i d_i = 0$

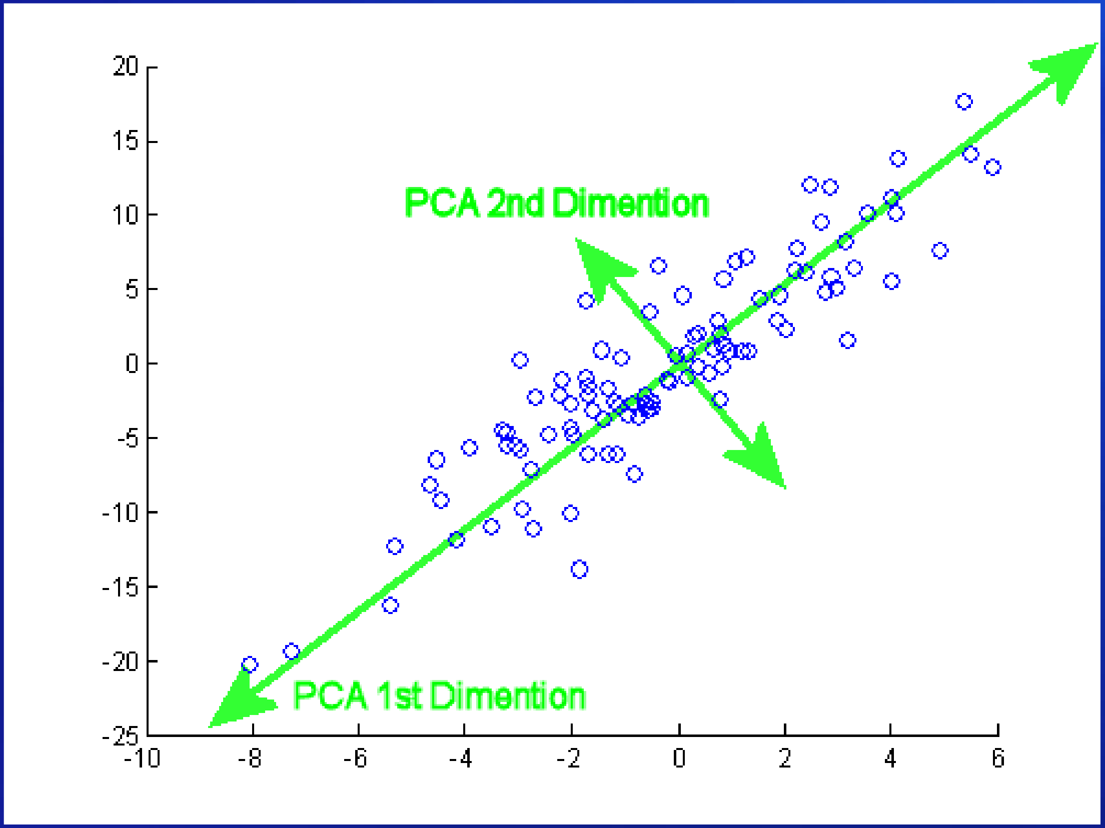
(2) $0 \leq \alpha_i \leq C \qquad \text{for } i = 1, 2, \ldots, N$

where C is a user-specified positive parameter.

Unsupervised vs supervised

- Explore Data is more important
- Data Visualization – Dimension Reduction
  - PCA
  - MDS
  - Manifold Learning
- Clustering
  - K-Mean Clustering
  - Hierarchical Clustering

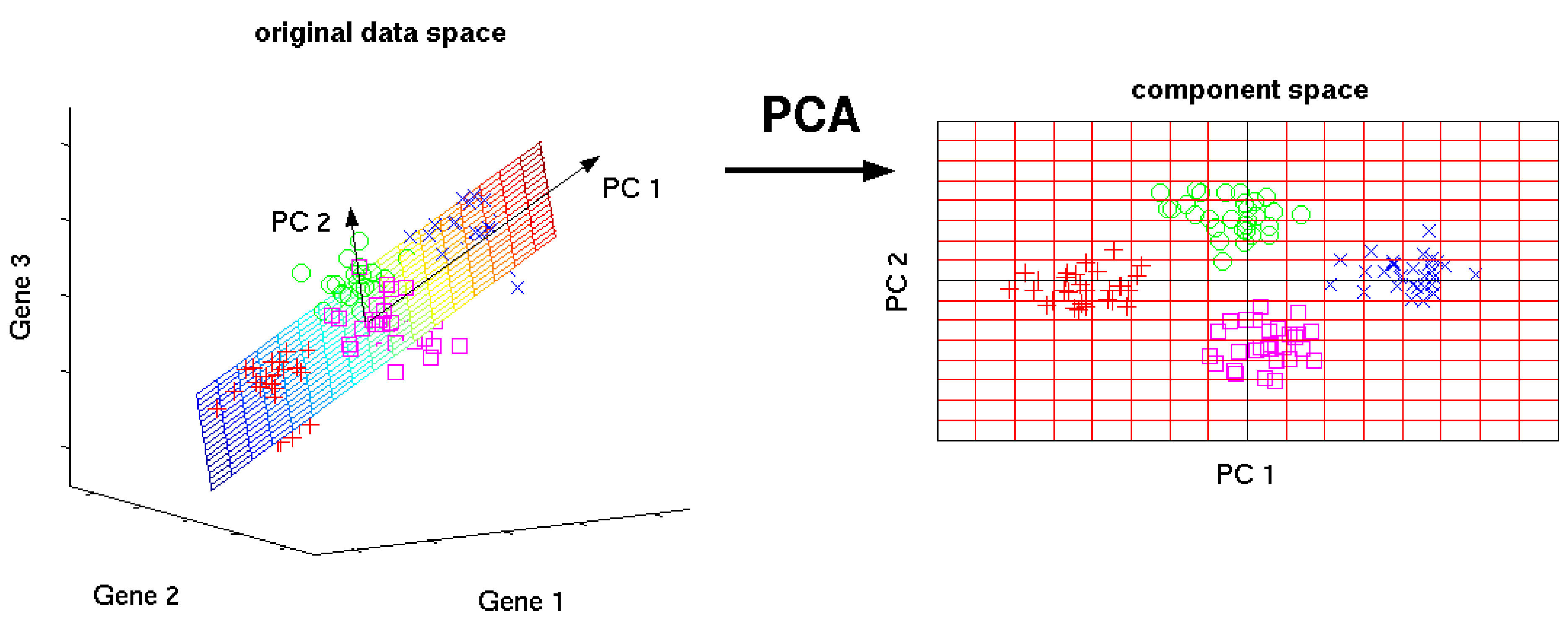
PCA-Principle Components Analysis



1，找线性变换，使得变换后的数据的弥散度最大 － 最主要成分

2，寻找一组线性变换，使得变换后的数据的第一个分量的弥散度最大、第二个分量的弥散度次大、如此类推。

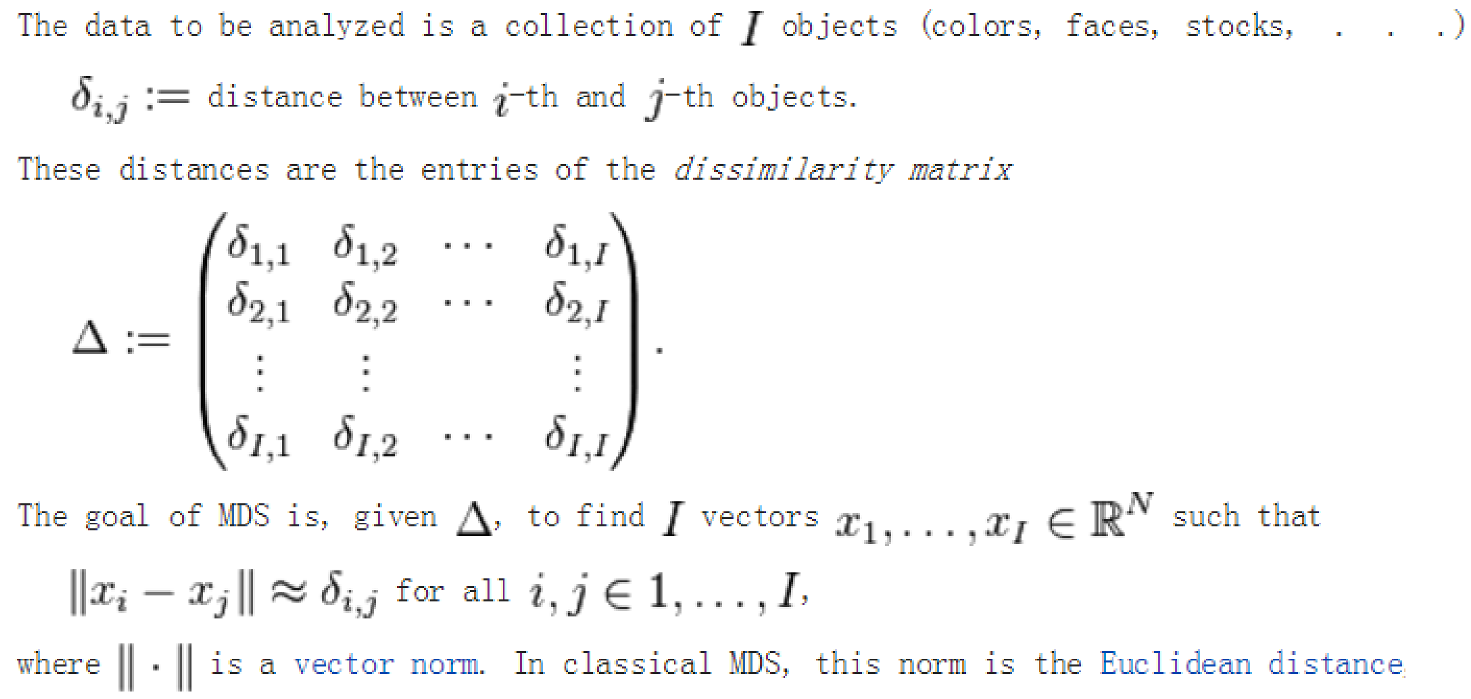
PCA-Principle Component Analysis



从高维到低维有可能揭示出真实规律 （我们很多时候在摸大象）

MDS-Multidimensional Scaling

找到从高维到低维的投影，使得两两之间的距离变化最小。



The data to be analyzed is a collection of I objects (colors, faces, stocks, . . .)

$\delta_{i,j} :=$ distance between i-th and j-th objects.

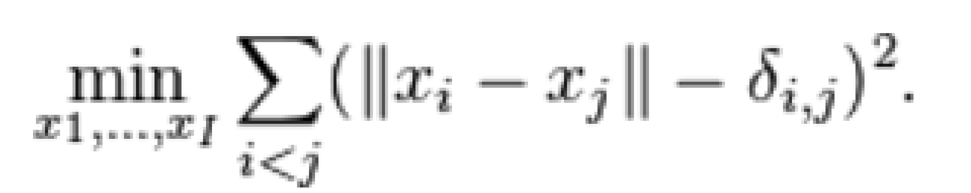
These distances are the entries of the *dissimilarity matrix*

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}.$$

The goal of MDS is, given Δ, to find I vectors $x_1, \ldots, x_I \in \mathbb{R}^N$ such that

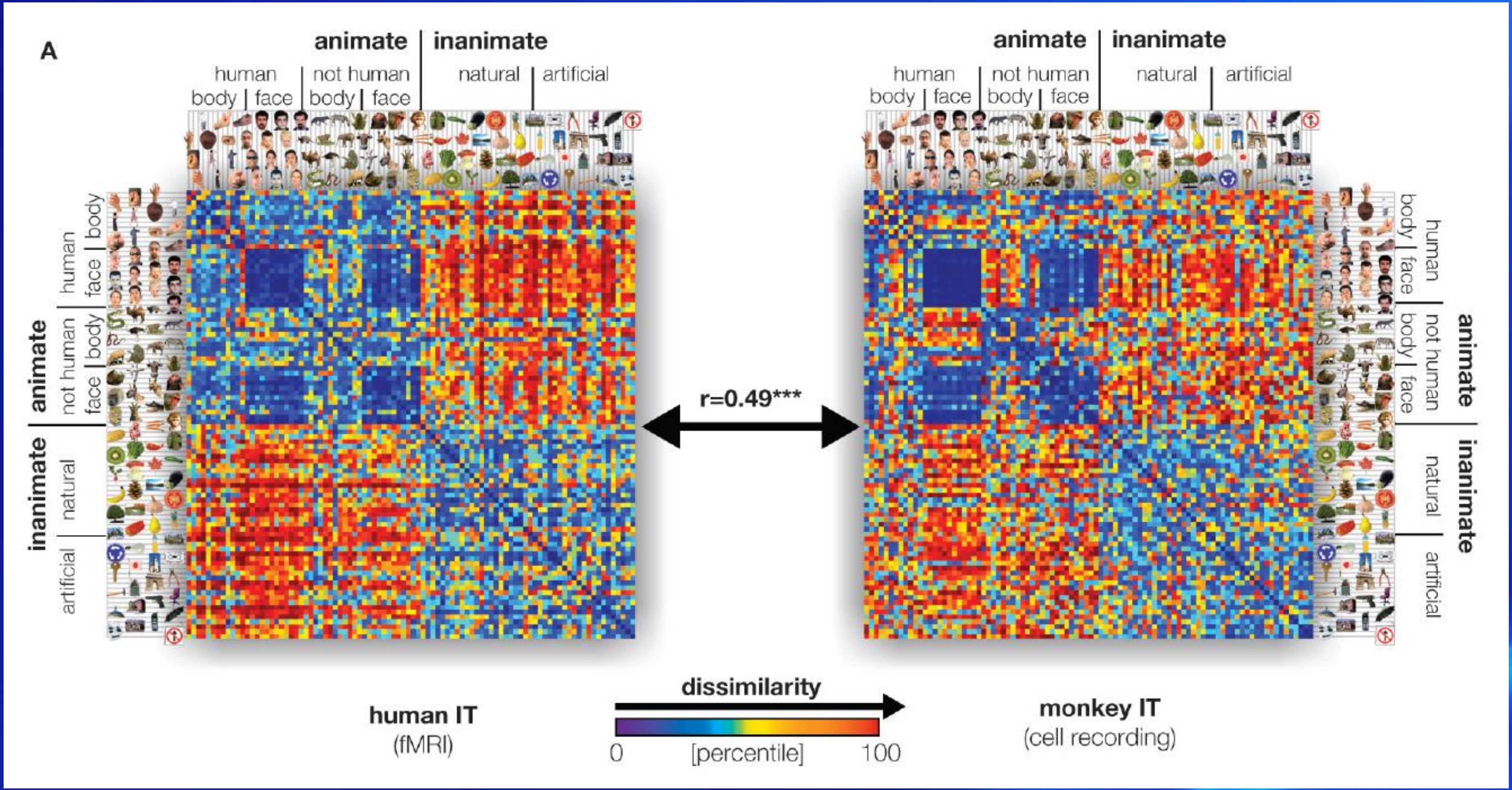
$\|x_i - x_j\| \approx \delta_{i,j}$ for all $i, j \in 1, \ldots, I$,

where $\|\cdot\|$ is a vector norm. In classical MDS, this norm is the Euclidean distance

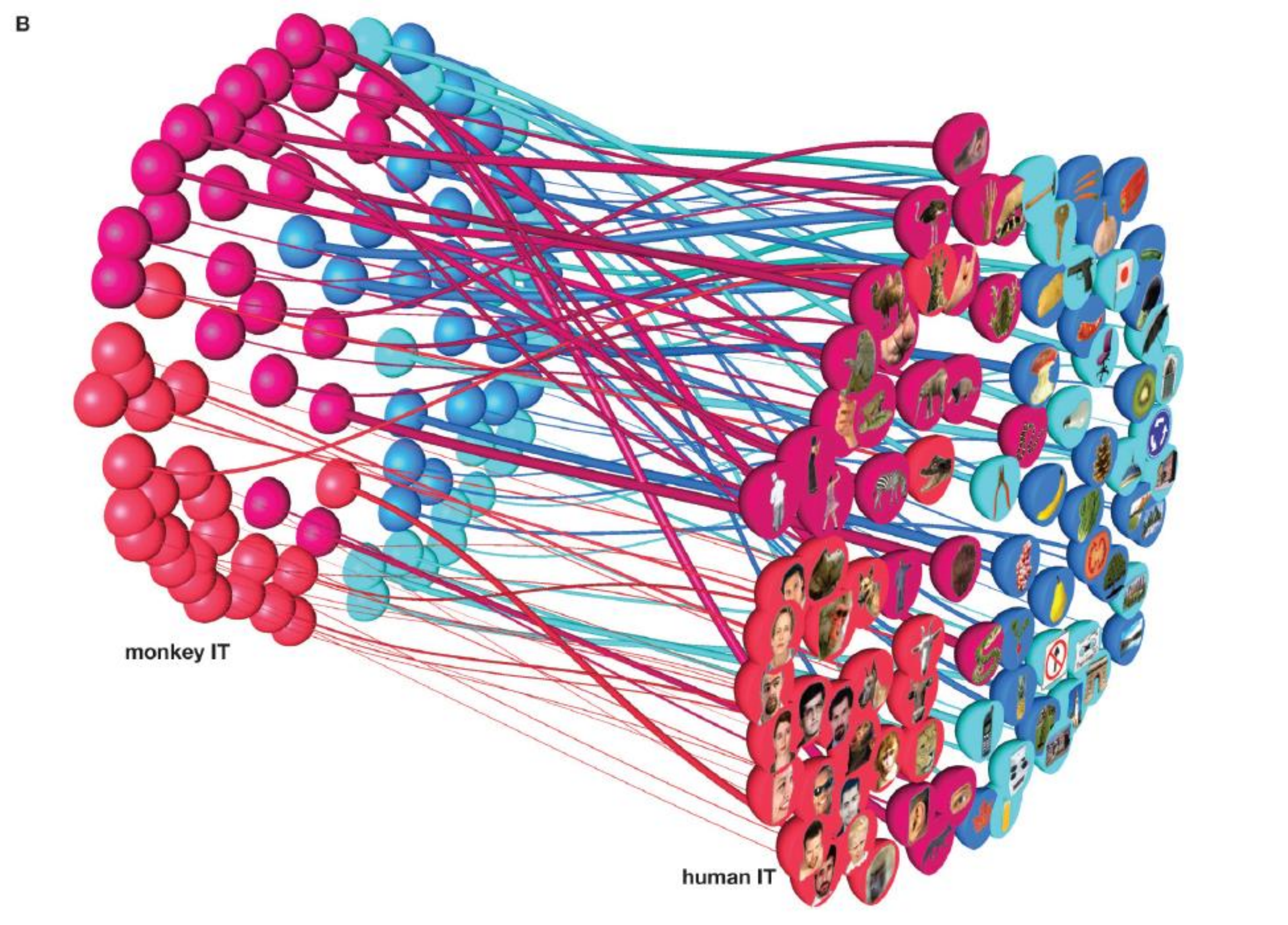


$$\min_{x_1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2.$$

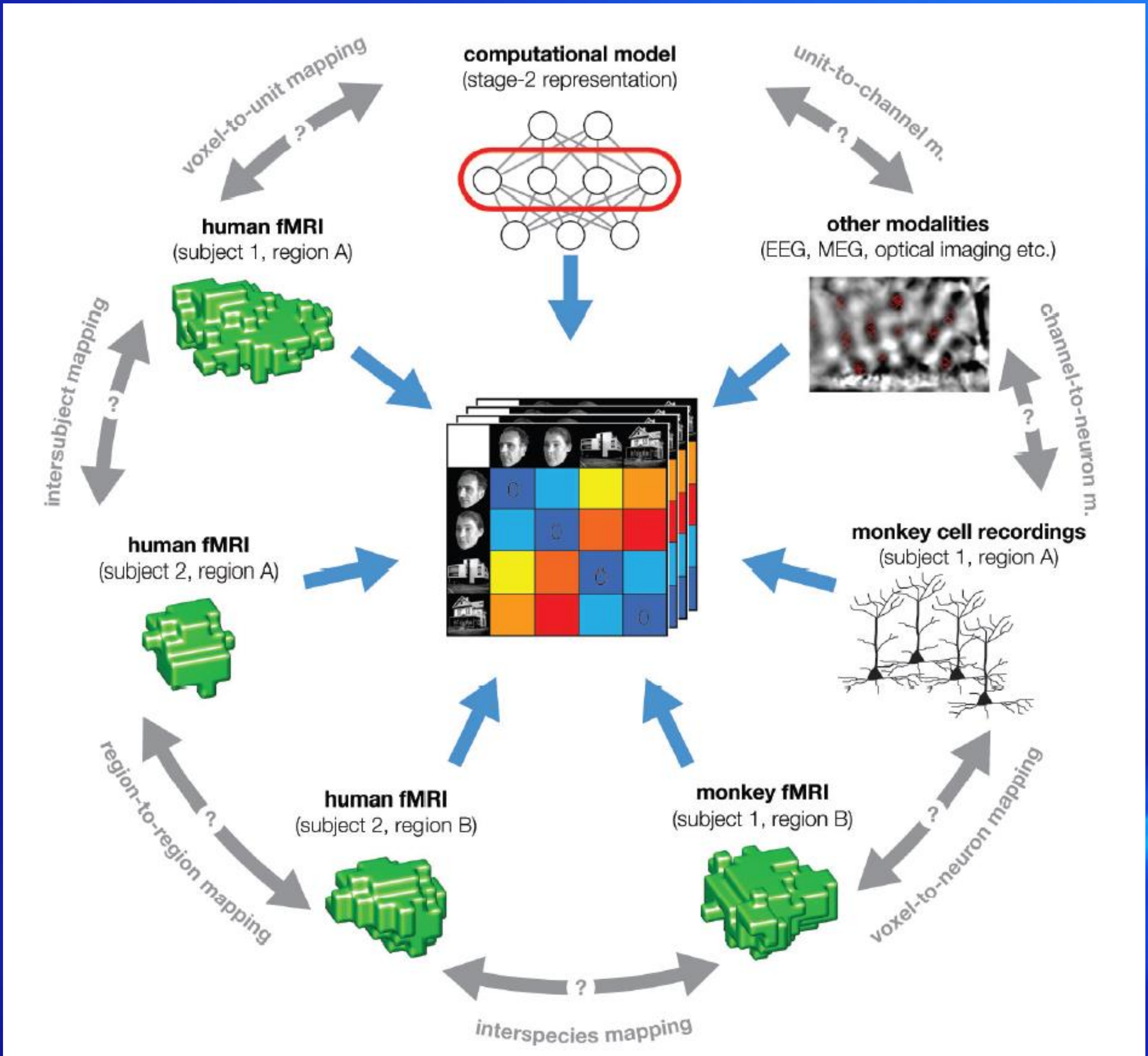
Distance Metrix in Brains



Human data is from 316 bilateral inferior temporal voxels selected by their visual-object response in an independent data set. Monkey data is from 674 IT single cells isolated in two monkeys (Kiani et al., 2007).



2009， Kriegeskorte



computational model
(stage-2 representation)

other modalities
(EEG, MEG, optical imaging etc.)

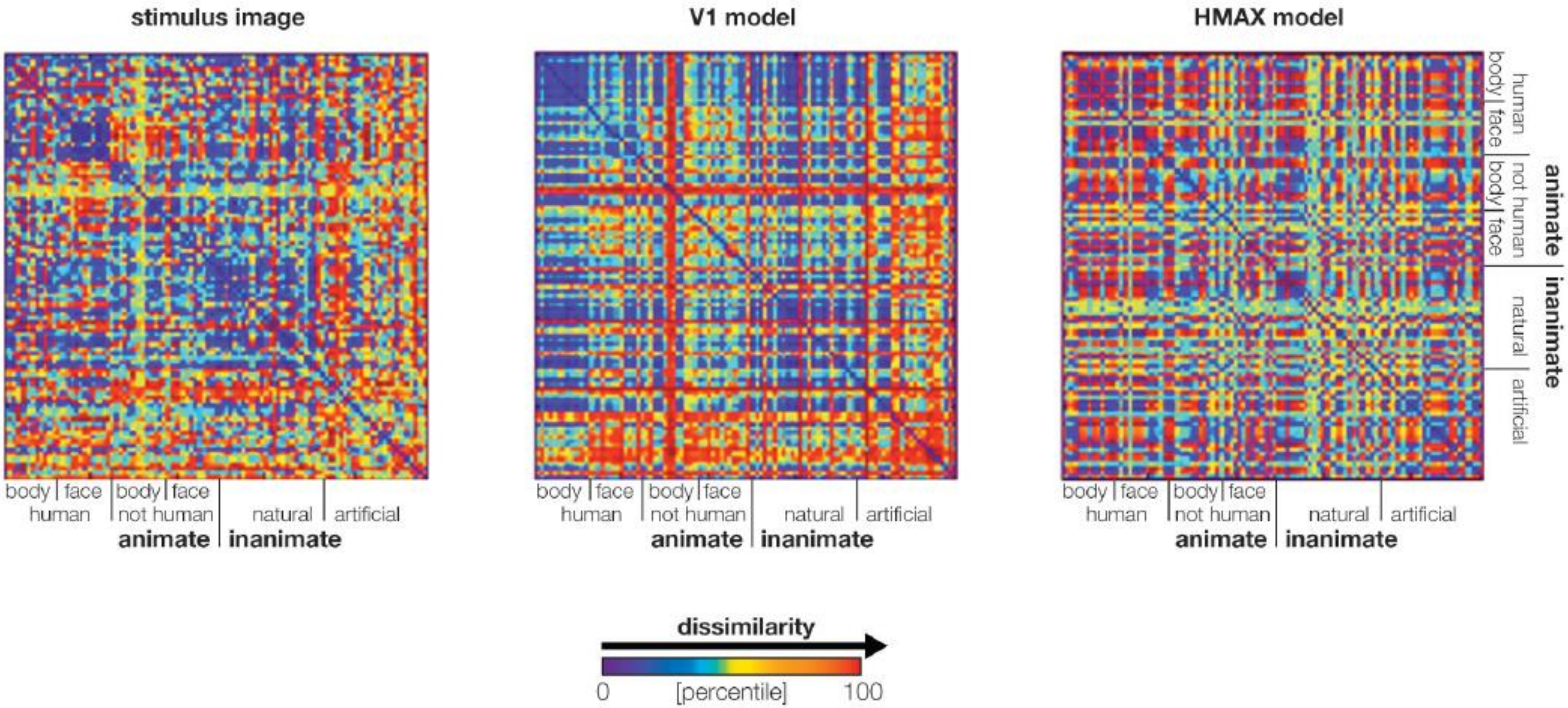
human fMRI
(subject 1, region A)

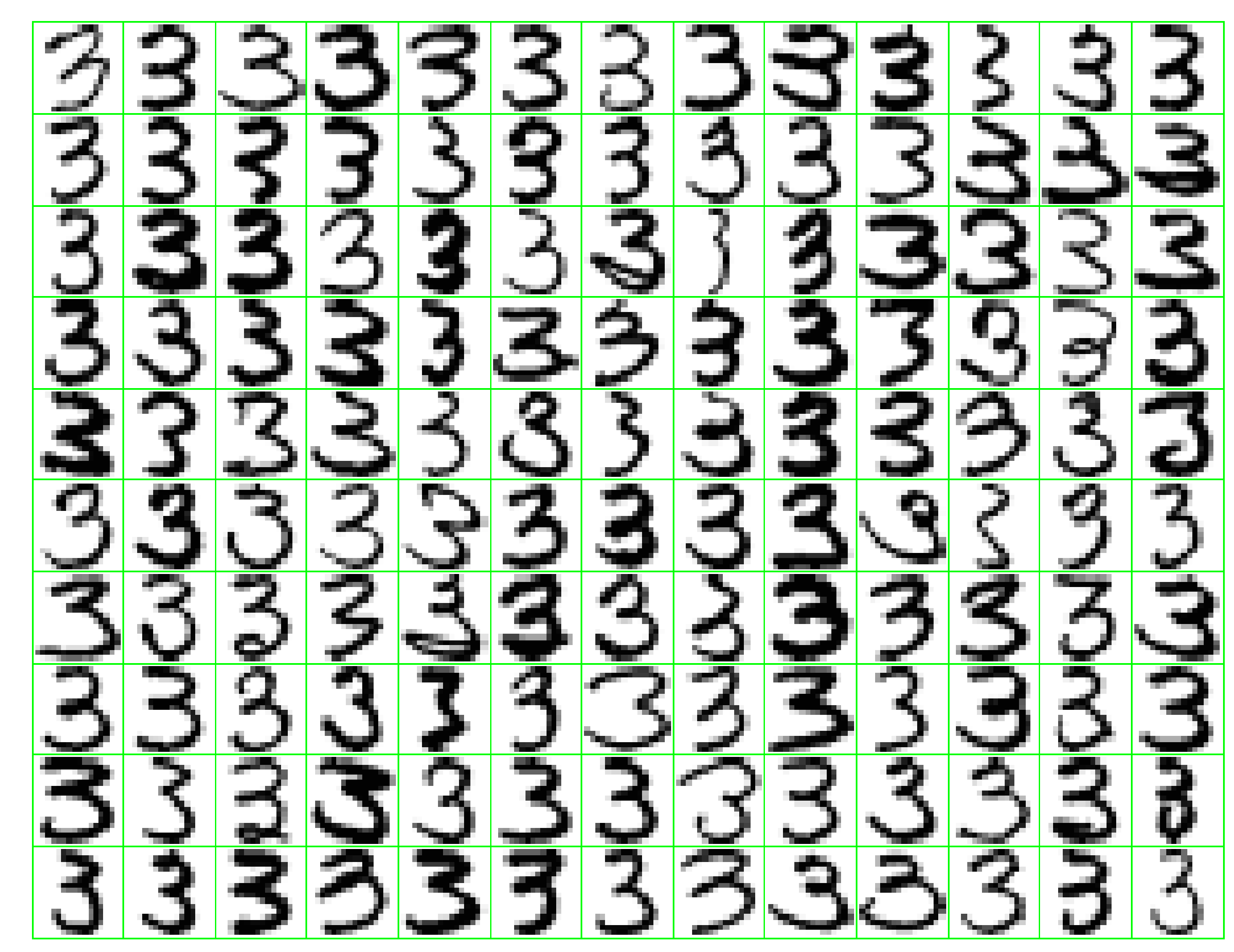
human fMRI
(subject 2, region A)

monkey cell recordings
(subject 1, region A)

human fMRI
(subject 2, region B)

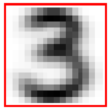
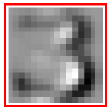
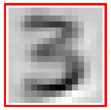
monkey fMRI
(subject 1, region B)

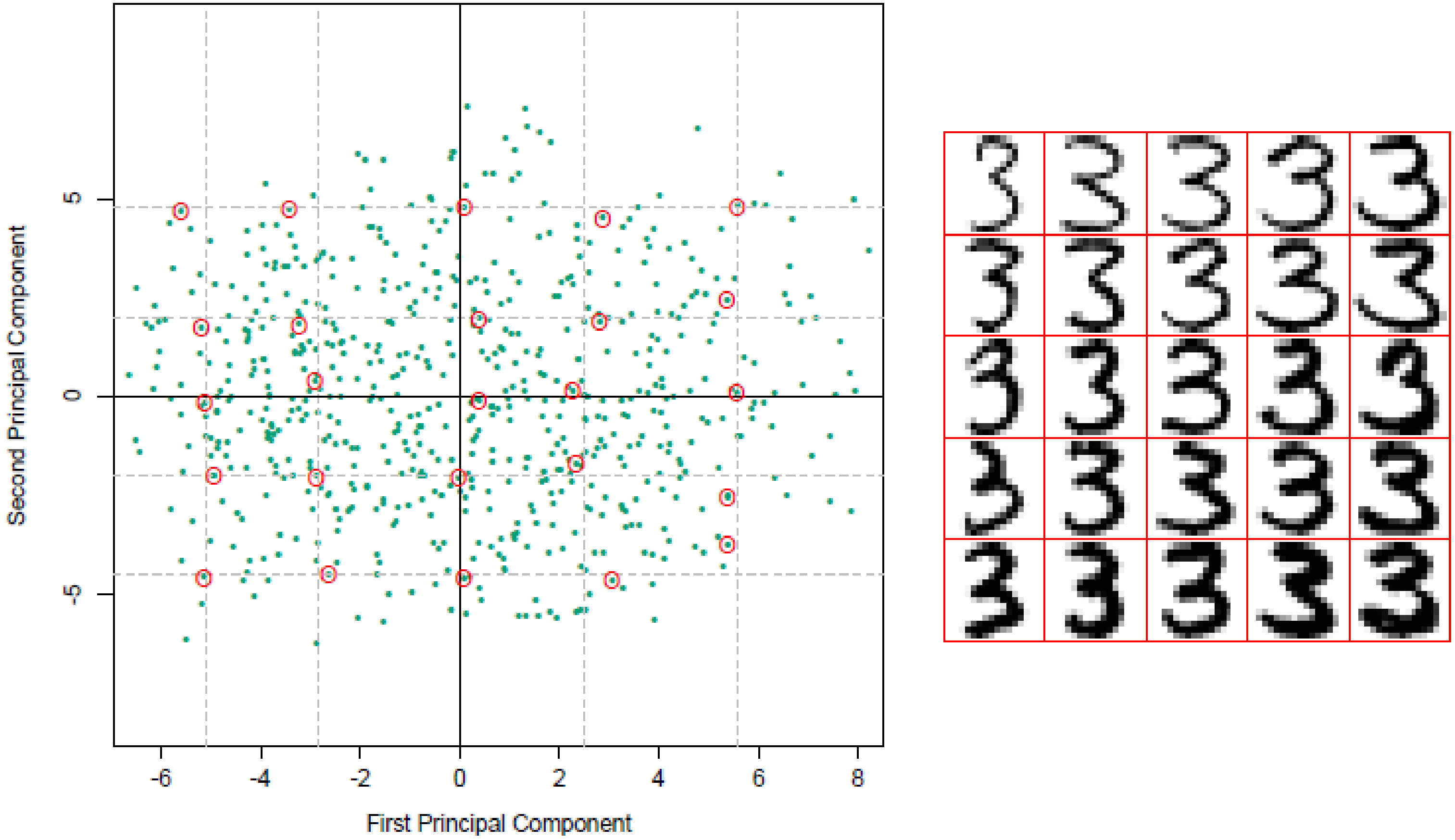

stimulus image
V1 model
HMAX model
body | face
human
body | face
not human
natural | artificial
animate | inanimate
dissimilarity
0
[percentile]
100



a sample of 130 handwritten 3’ s, each a digitized 16 × 16 grayscale image

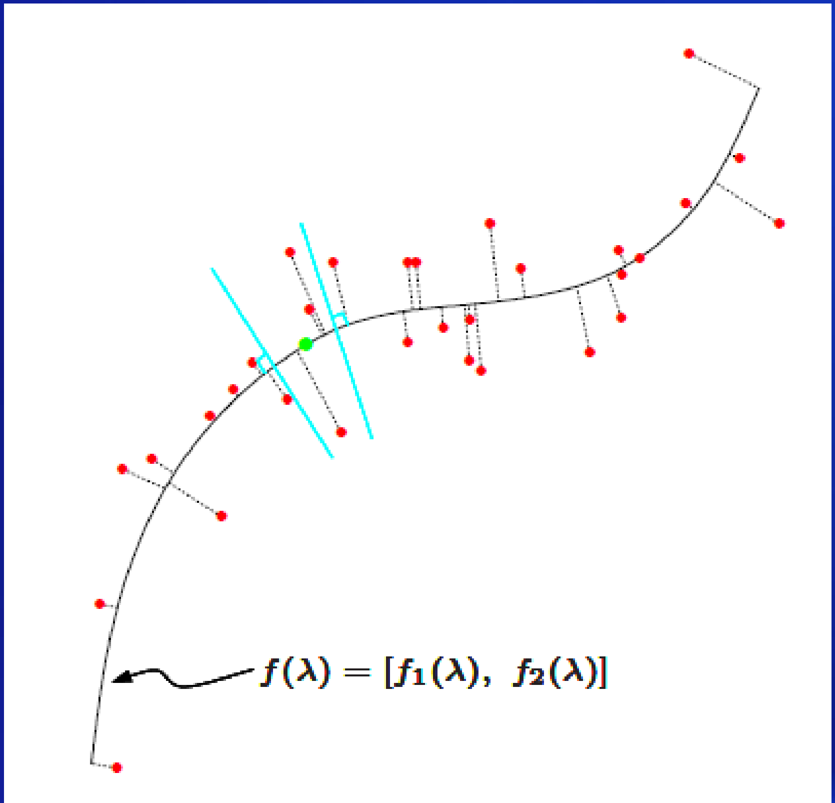
$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$=$  $+ \lambda_1 \cdot$  $+ \lambda_2 \cdot$  .

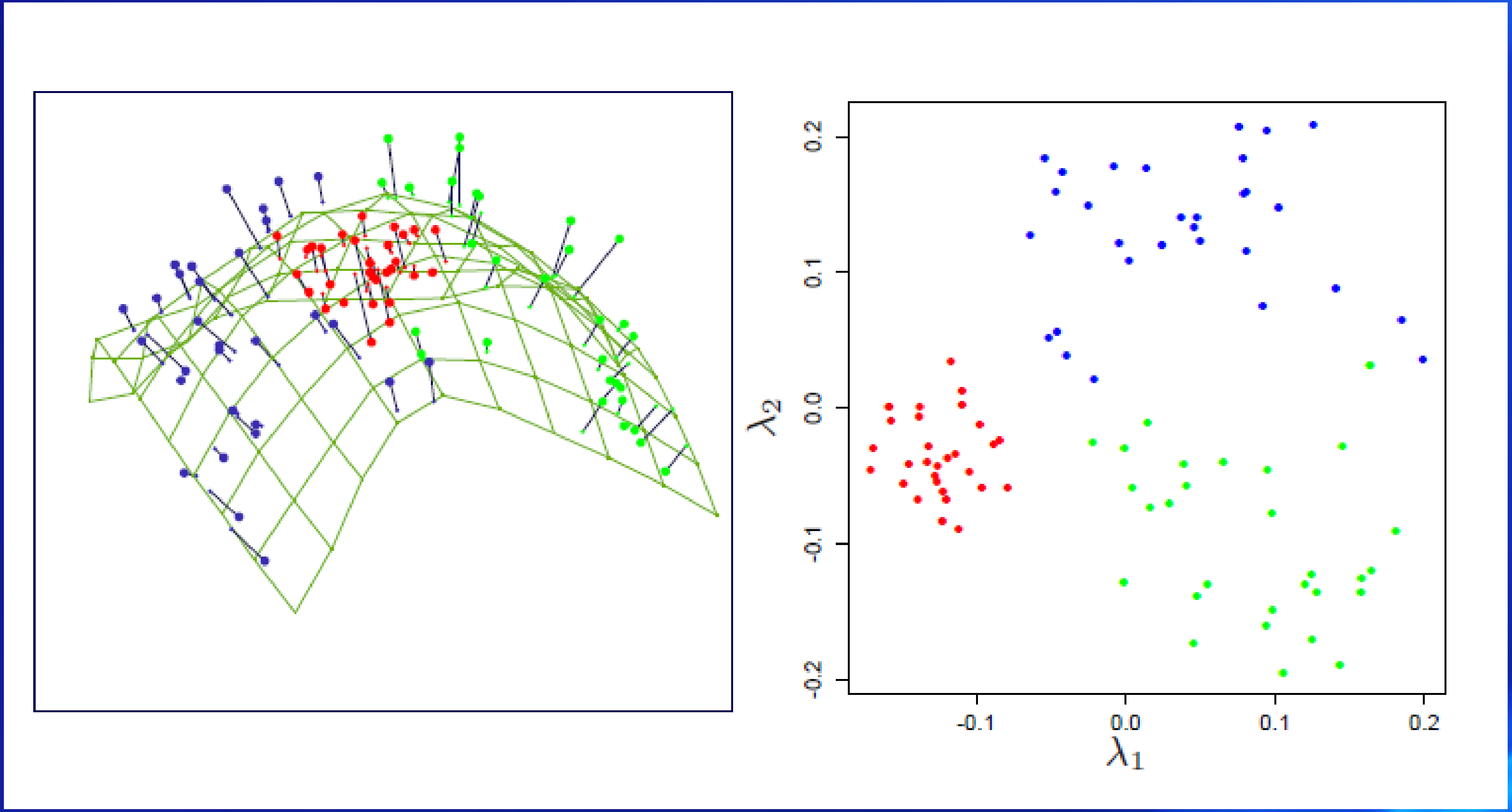


PCA-Principle Curves

线性变换 =〉非线性变换



PCA-Principle Surfaces



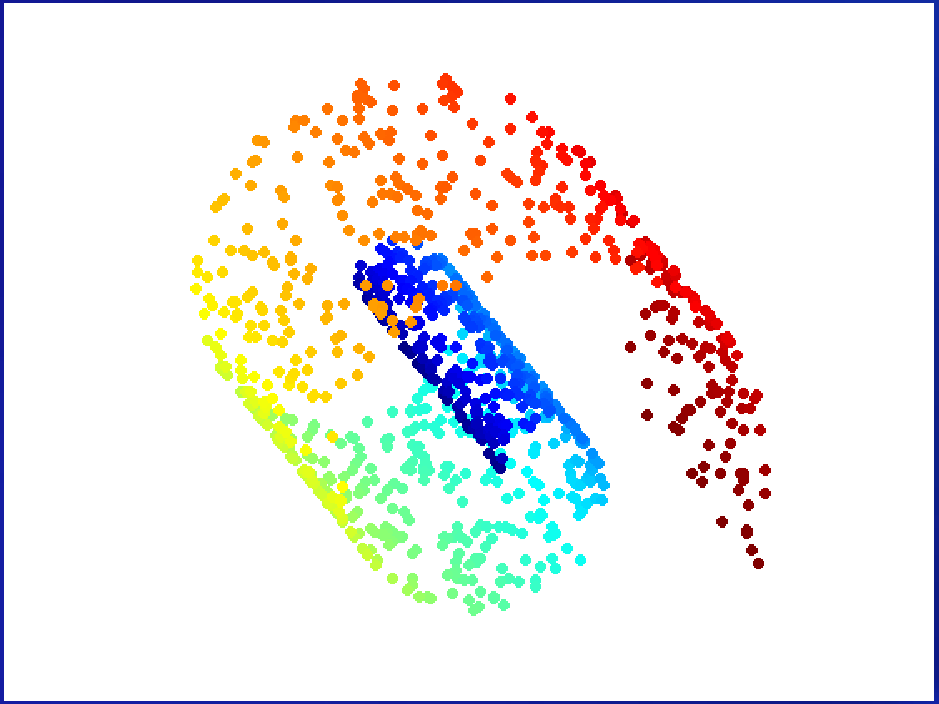
What about the surfaces are folded?

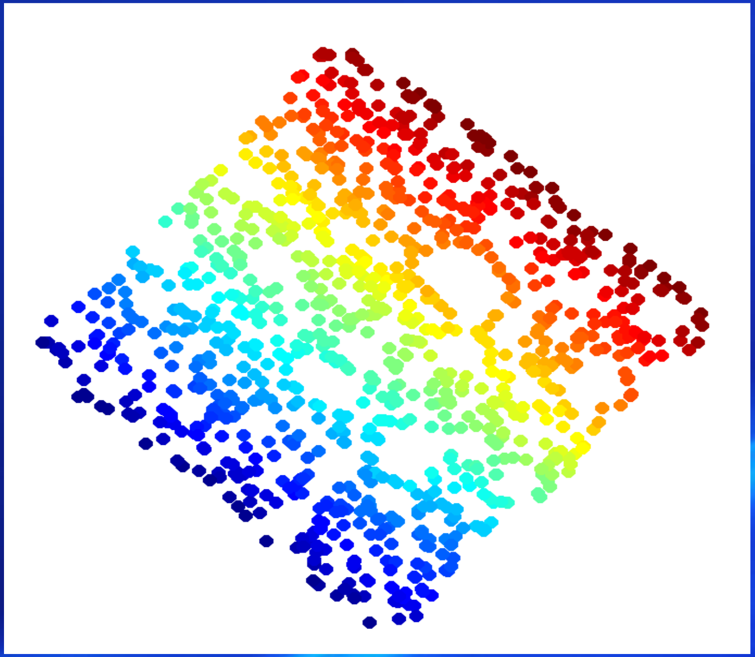
流型学习 (Manifold Learning)

- Data may live on low-dim space, but we observe them from too many different angles

流型学习

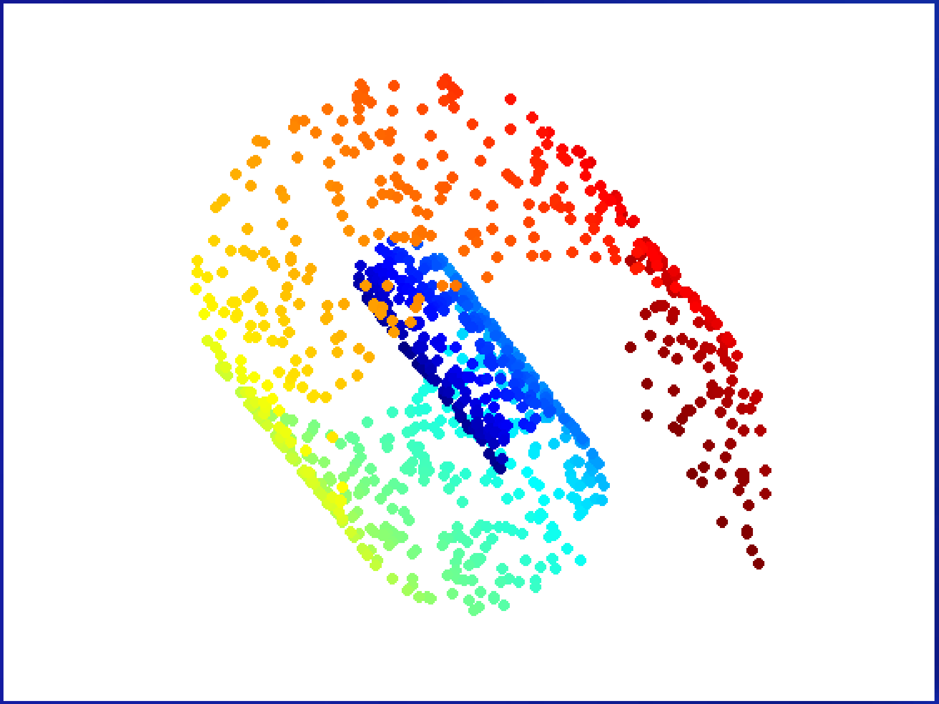
- Data distributed around a manifold form
- Find a way a flat them out – Reducing Dimension

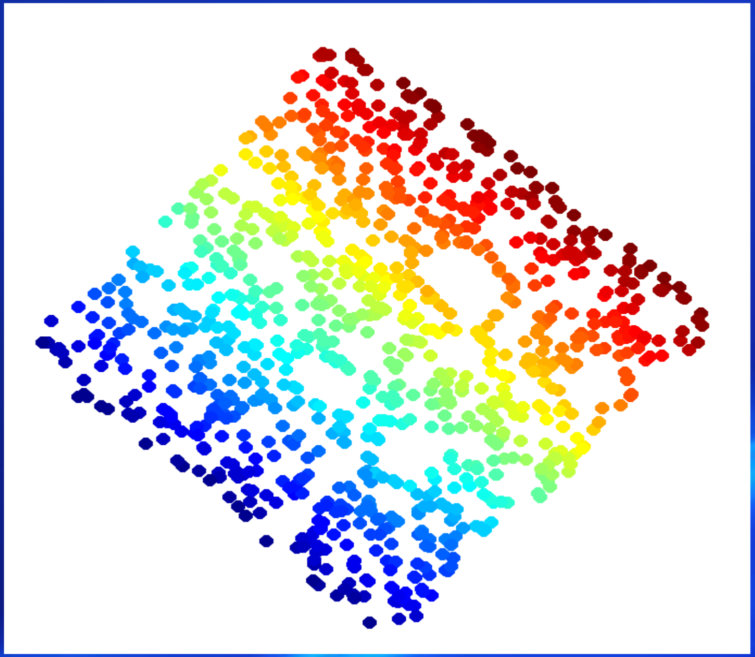




流型学习

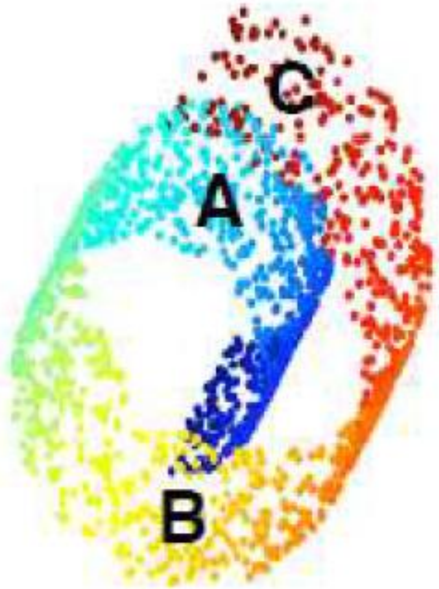
- Data distributed around a manifold form
- Find a way a flat them out – Reducing Dimension



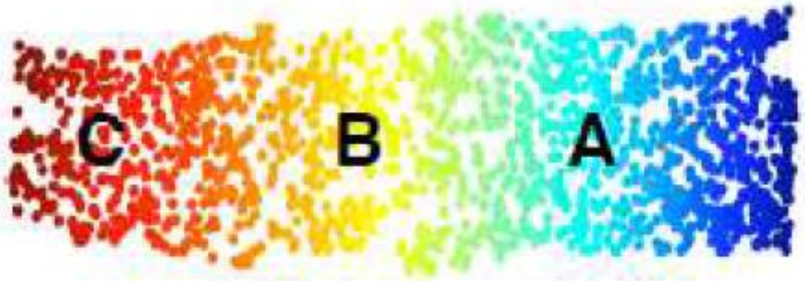


Non-metric MDS for manifolds?

Rank ordering of Euclidean distances is NOT preserved in “manifold learning”.

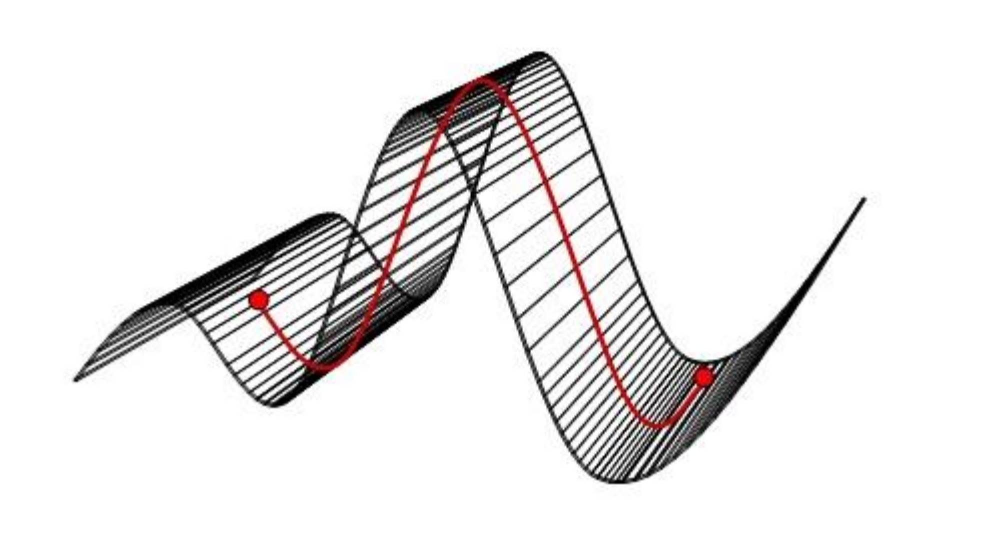


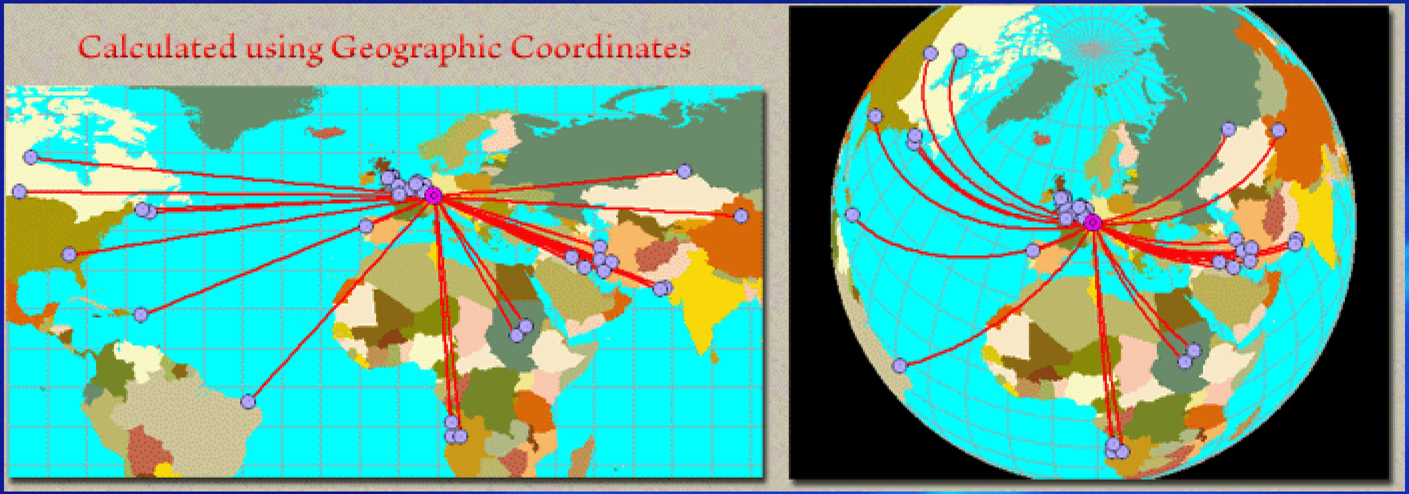
$d(A,C) < d(A,B)$



$d(A,C) > d(A,B)$

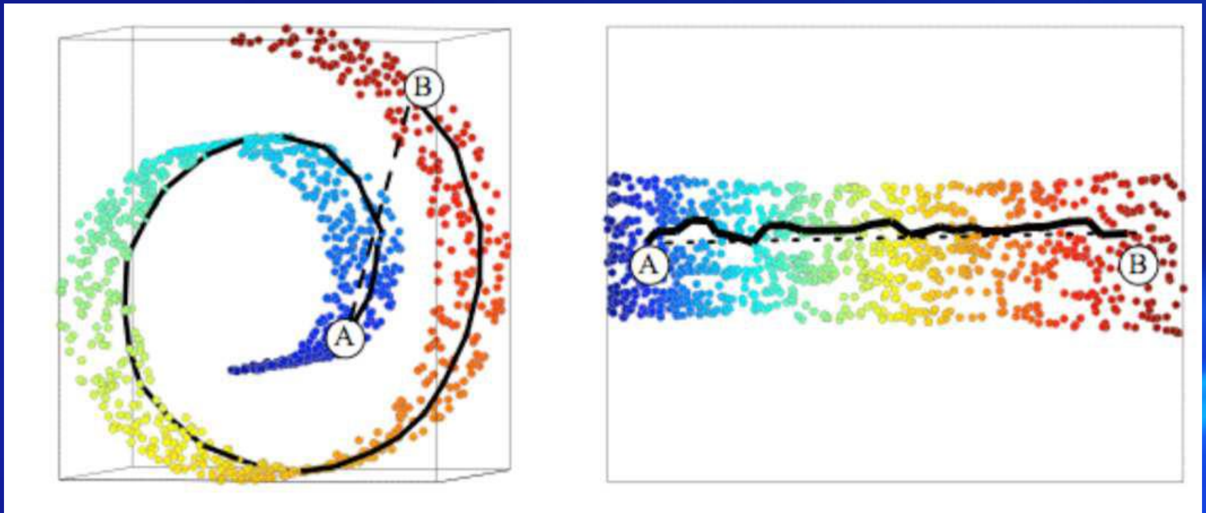
The geodesic distance





Isomap - Key Idea

- Local distance can use Euclidean distance.
- Build locally connected Graph
- Build Long geodesic distance through Graph
- Once we have all paired distances, we can use MDS



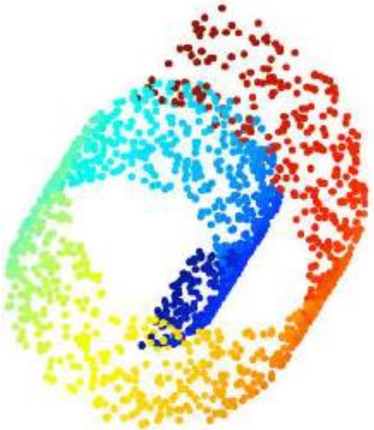
Building the Graph

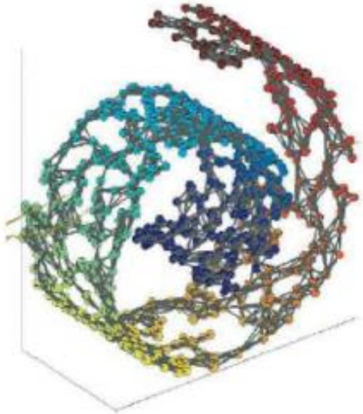
- **Adjacency graph**

  Vertices represent inputs. Undirected edges connect neighbours.

- **Neighbourhood selection**

  Many options: k-nearest neighbours, inputs within radius r, prior knowledge.





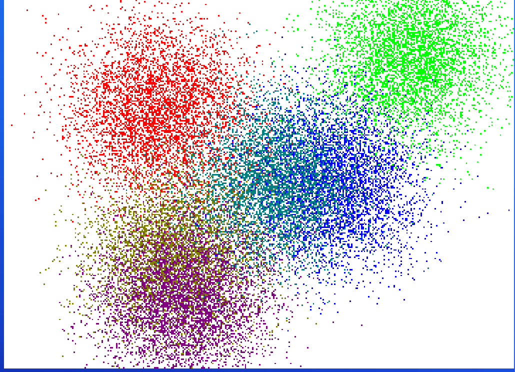
Graph is discretized approximation of submanifold.

手写体的识别

- 联机识别比较容易，已经商用
- 中文脱机识别仍较困难？！

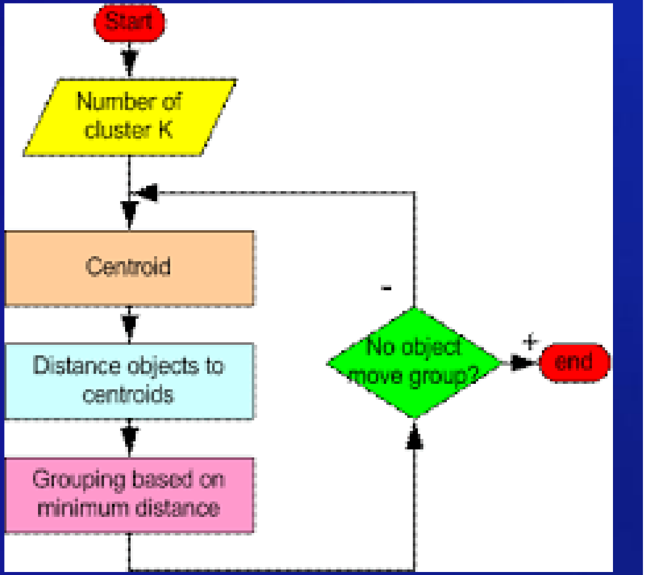
采用流型学习的方法把数据从Image变成Graph可能是很好的方法

Clustering

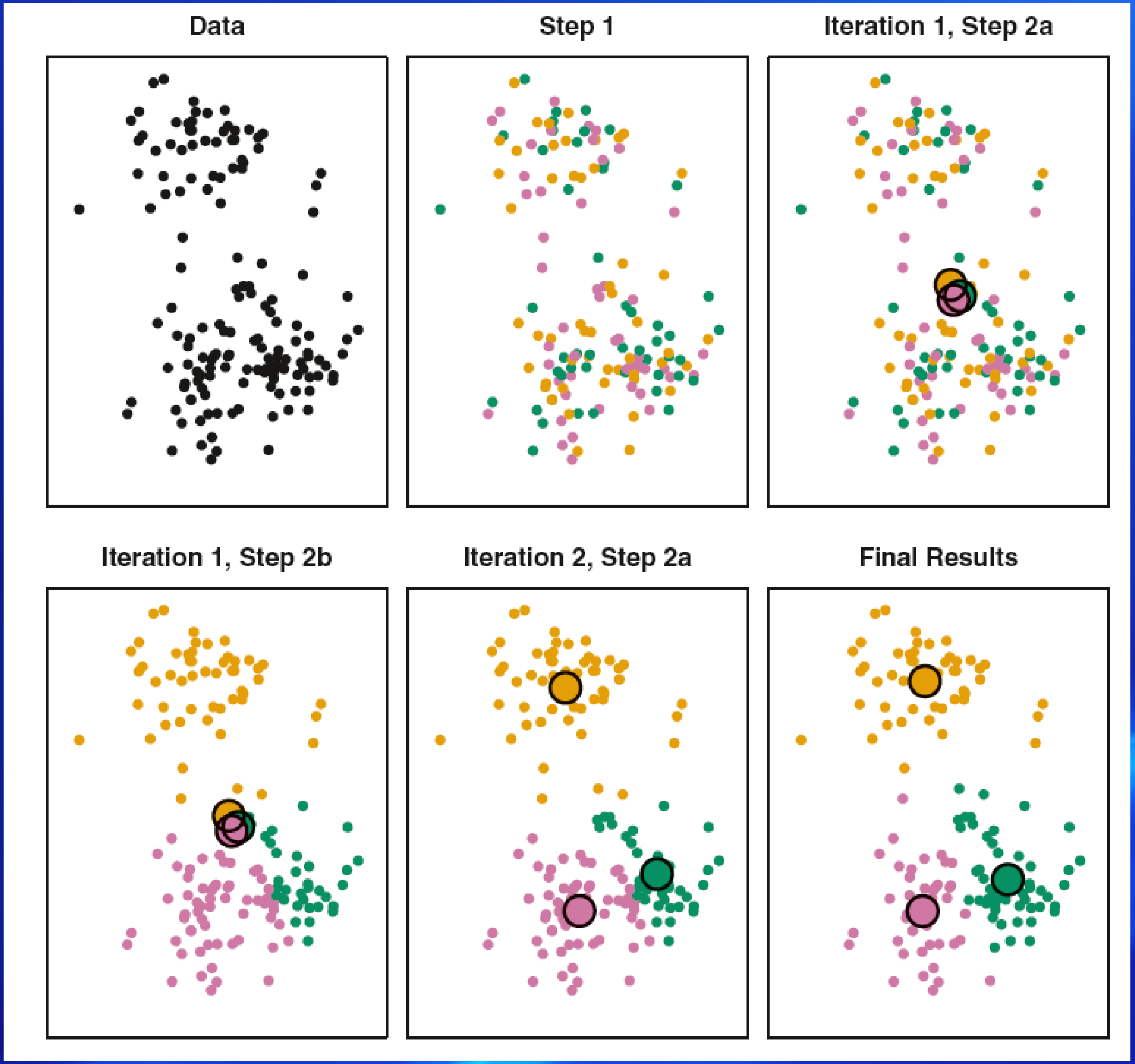


- Minimize Within Group Variance
- Maximize Between Group Variance
- Too many possibilities to consider!
- Strategy:
  - Looking for local and fast solutions
  - Try some and get the best solution

Clustering – K-mean


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
-
No object move group?
+
end

$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j\|^2$$


Data

Step 1

Iteration 1, Step 2a

Iteration 1, Step 2b

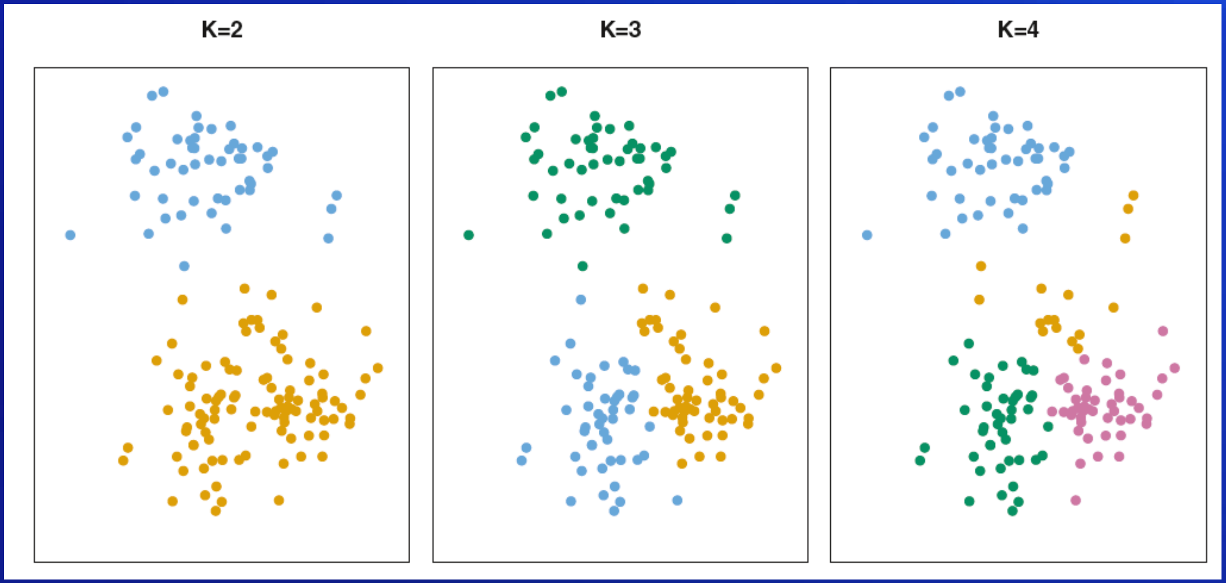
Iteration 2, Step 2a

Final Results

An Introduction to Statistical Learning

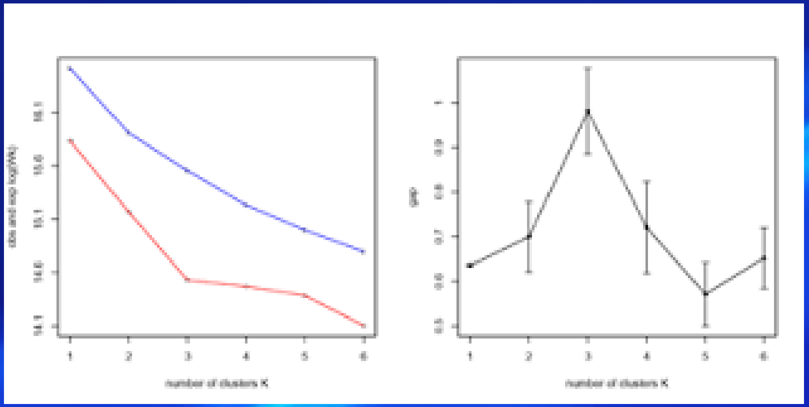
Clustering – K-mean

How do you know about K?



$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j\|^2$$

Gap Statistic

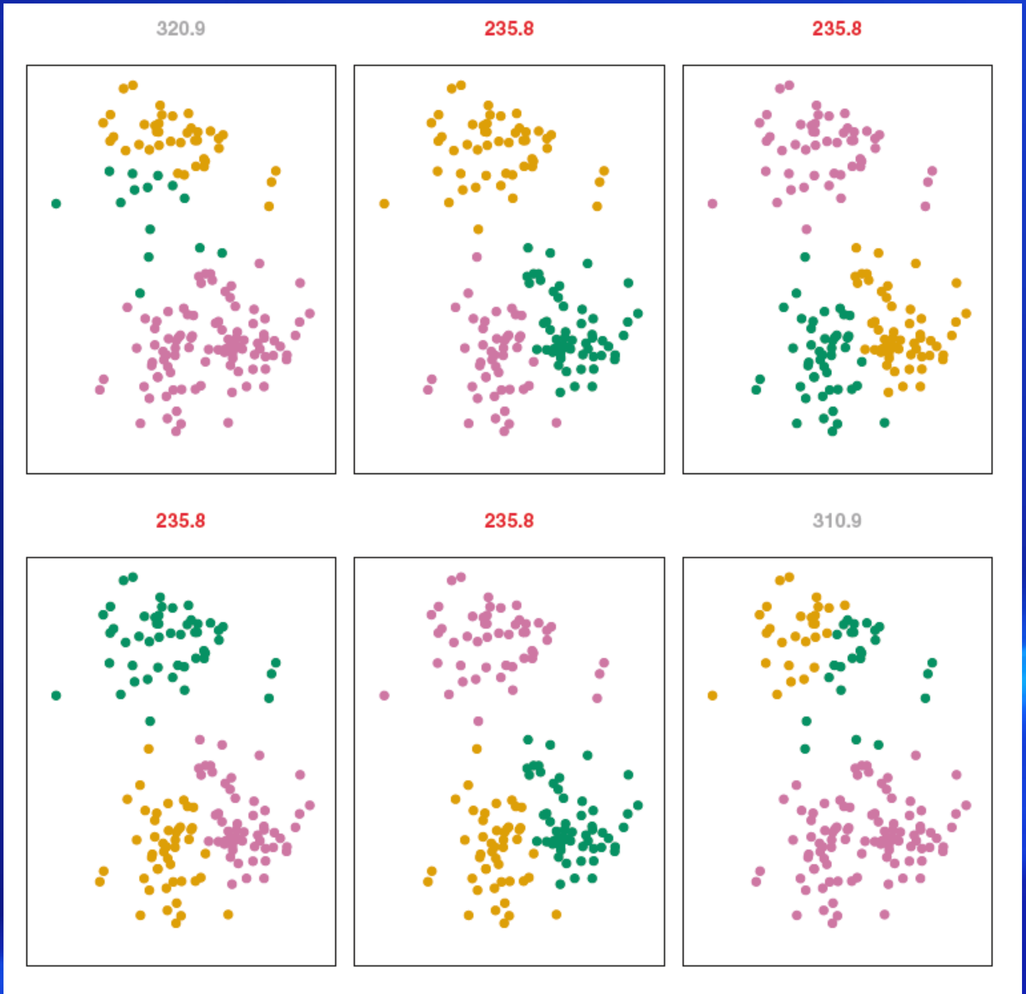


Clustering – K-mean

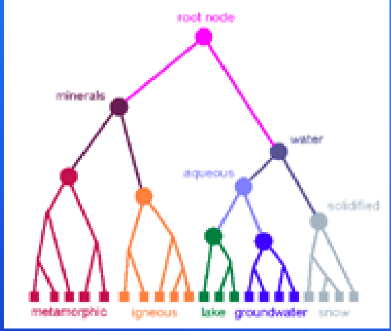
How do you know optimum solution?

Multiple Trials

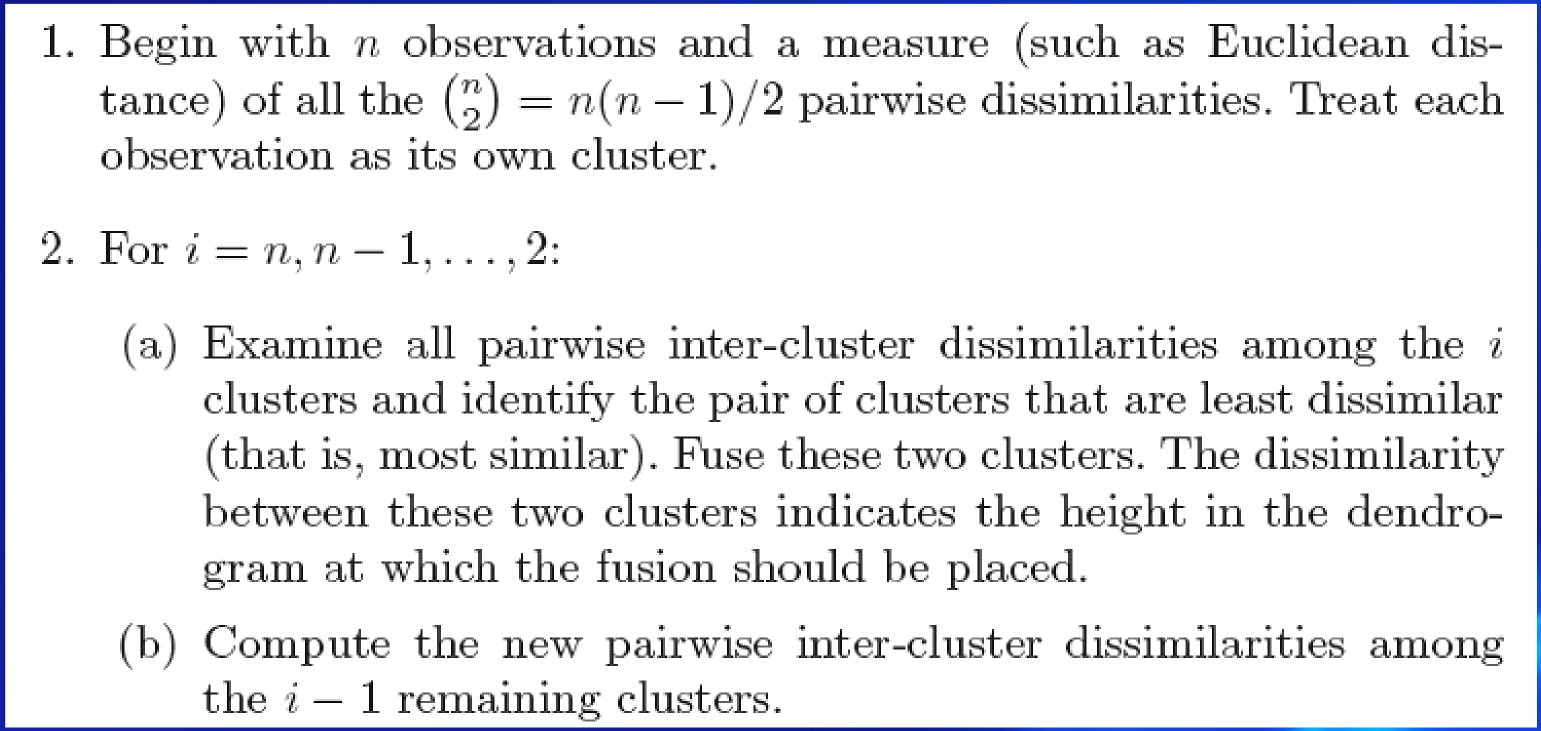
$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j\|^2$$



Hierarchical Clustering



Bottom-up Approach

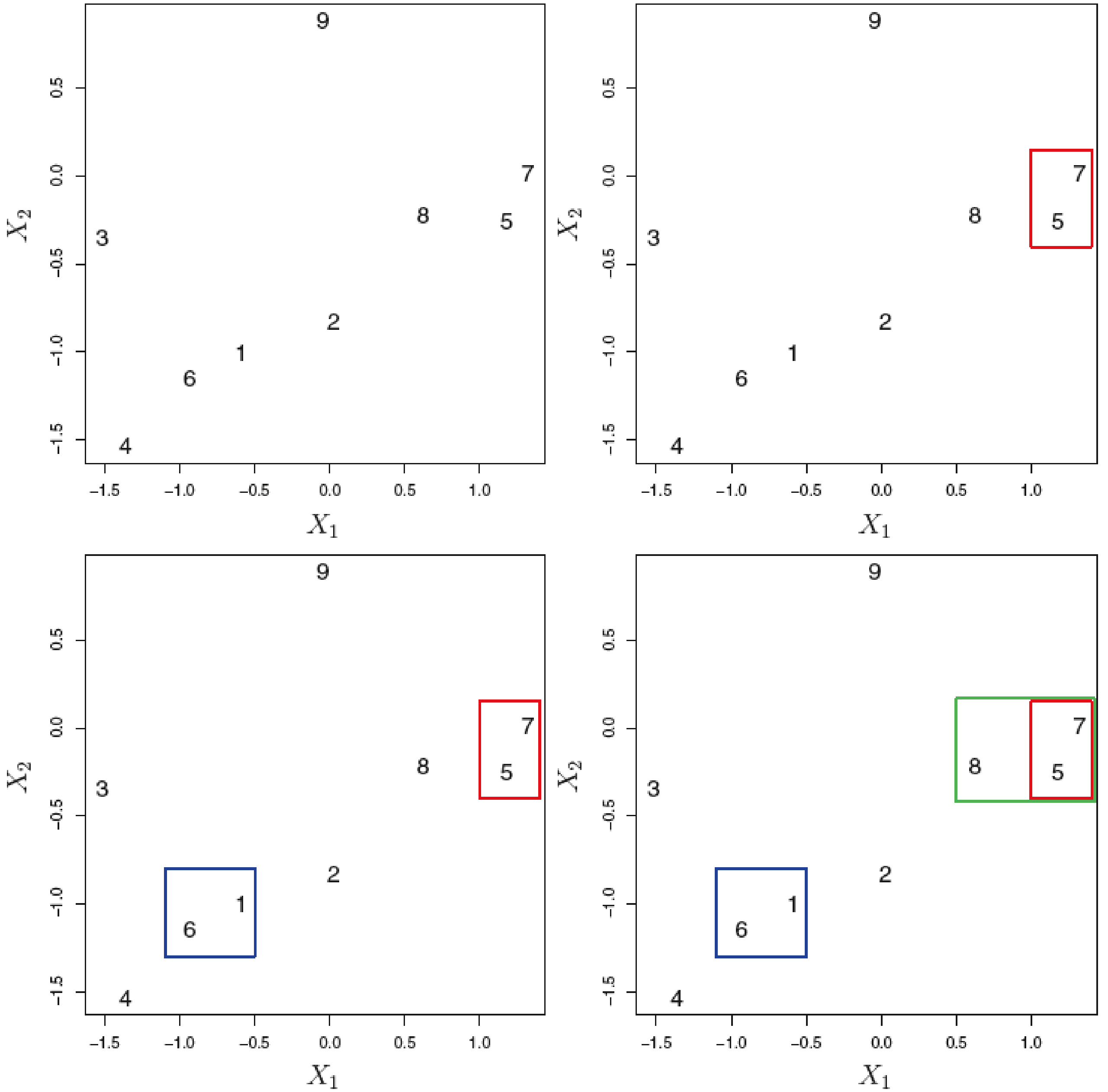


1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

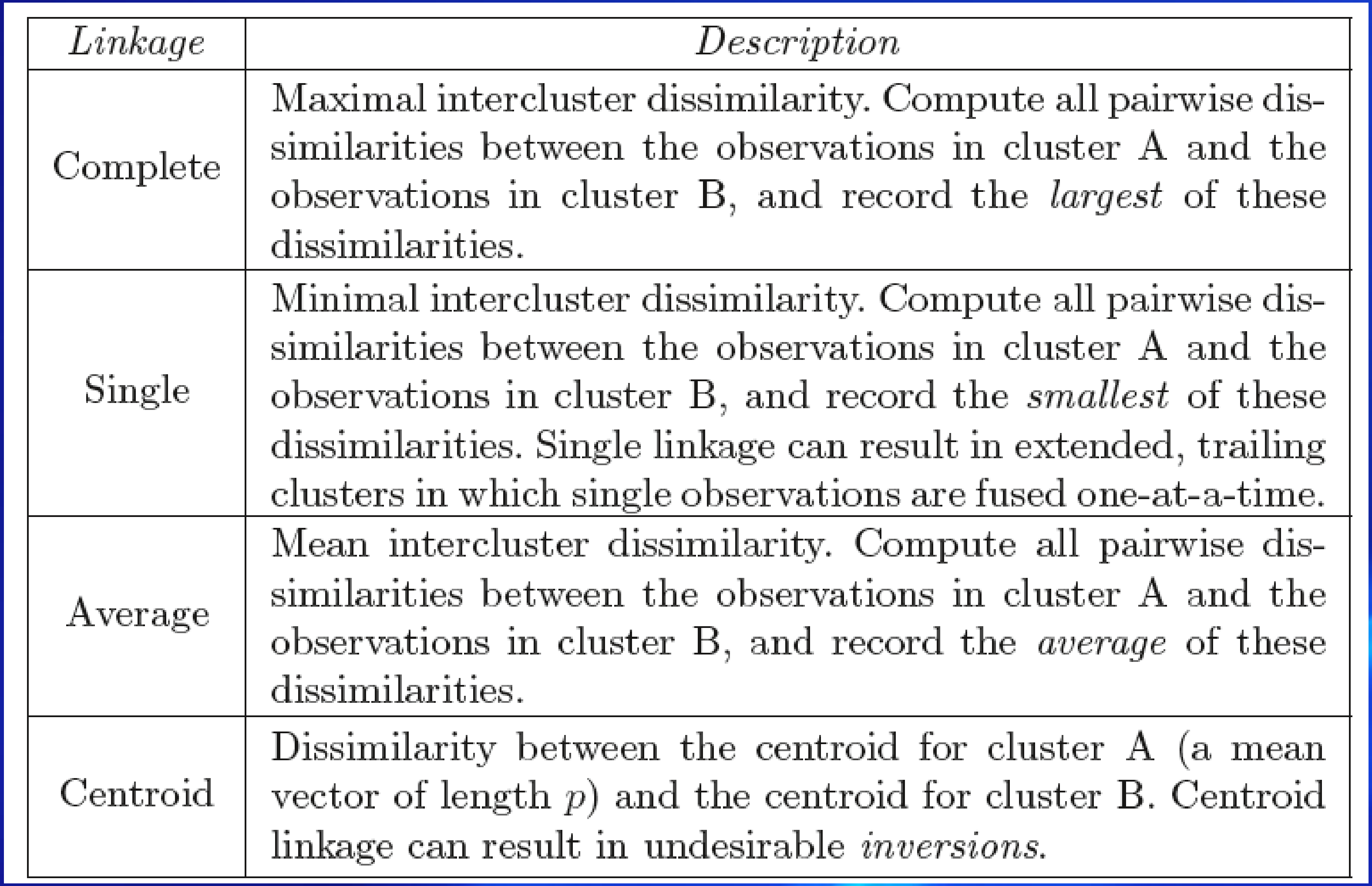
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.



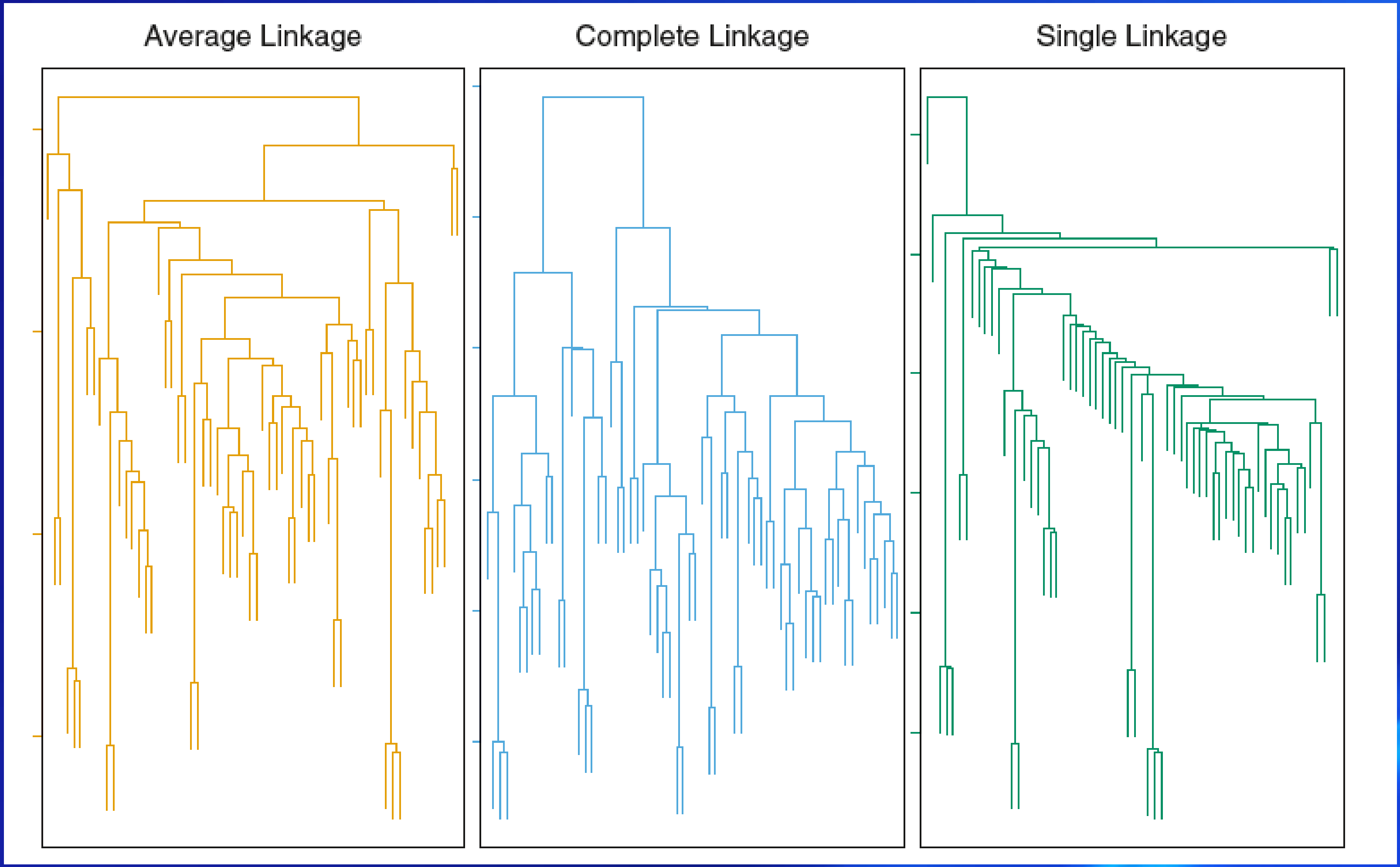
9
7
8
5
3
2
1
6
4
X_1
X_2
-1.5
-1.0
-0.5
0.0
0.5
1.0

Clustering - Hierarchical Clustering

- *Linkage* - the dissimilarity between two groups of observations.



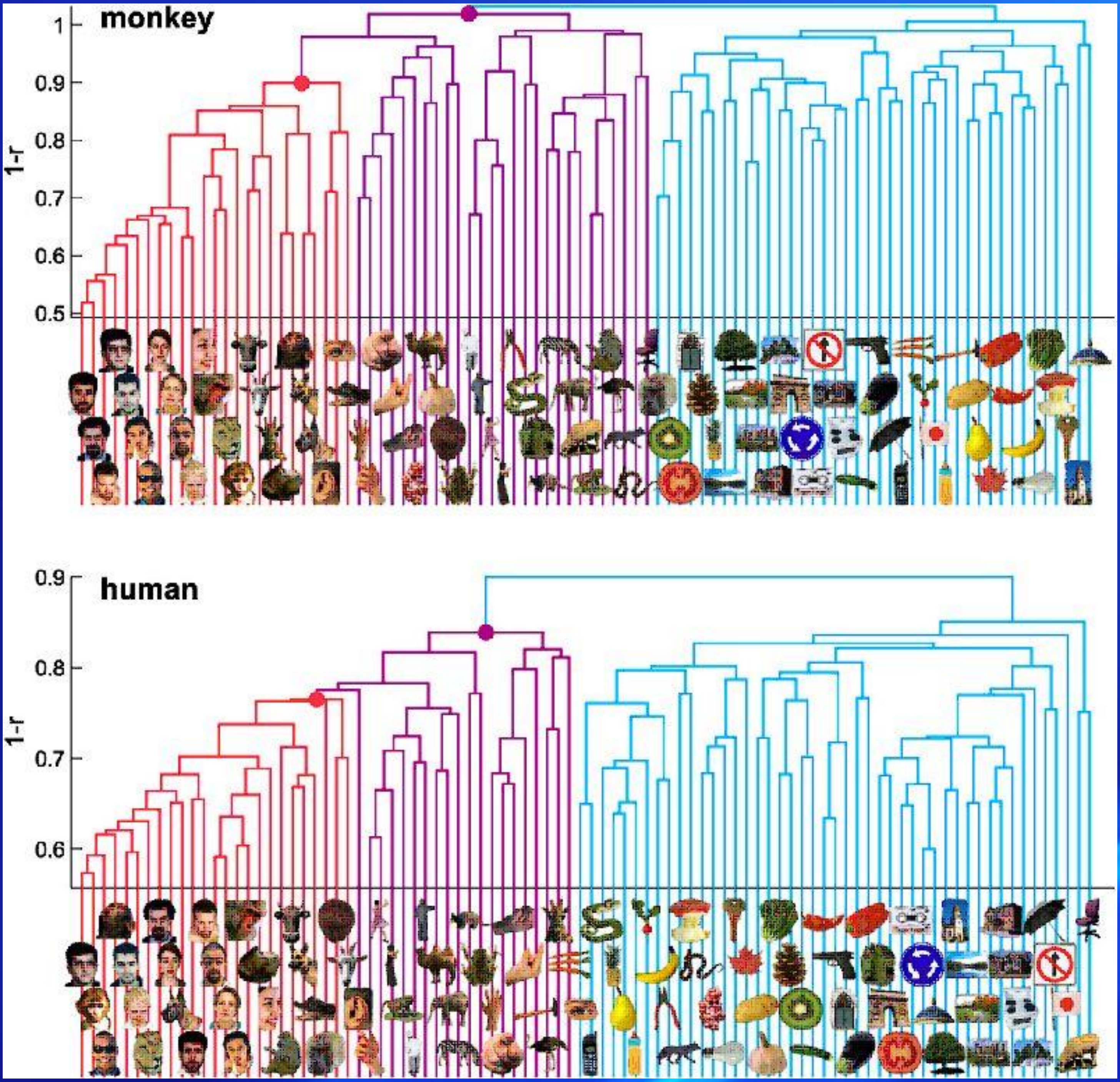
Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.



Average Linkage

Complete Linkage

Single Linkage


monkey
1
0.9
0.8
0.7
0.6
0.5
1-r
human
0.9
0.8
0.7
0.6
1-r

总结

- 分类问题
  - Supervised
    - Logistic Regression
    - SVM
  - Unsupervised
    - Dimension Reduction
      - CPA
      - MDS
      - IsoMap
    - Clustering
      - K-Mean
      - Hierarchical

Homework

- Theory
  - 请叙述并推导Logistic Regression中参数的MAP估计表达。
  - 请叙述并推导SVM在不可分数据情况下的最优解表达。
  - 请叙述并推导线性PCA中的主成分表达式。
- Practice
  - （本题可以组队完成）收集中传男女学生身高体重信息，
    - 分别采用1维和2维Logistic Regression的方法实现ML和MAP估计。
    - 用SMV的方法，依据中传学生身高体重信息对性别分类。
    - 比较以上两种方法的结果。
  - 数据Cat4D3Groups是4维观察数据，
    - 请先采用MDS方法降维到3D，形成Cat3D3Groups数据，显示并观察。
    - 对Cat3D3Groups数据采用线性PCA方法降维到2D，形成Cat2D3Groups数据，显示并观察。
    - 对Cat2D3Groups数据采用K-Mean方法对数据进行分类并最终确定K，显示分类结果。
    - 对Cat2D3Groups数据采用Hierarchical分类法对数据进行分类，并显示分类结果。

在Python中使用最优化算法可以参考scipy.optimize
http://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html