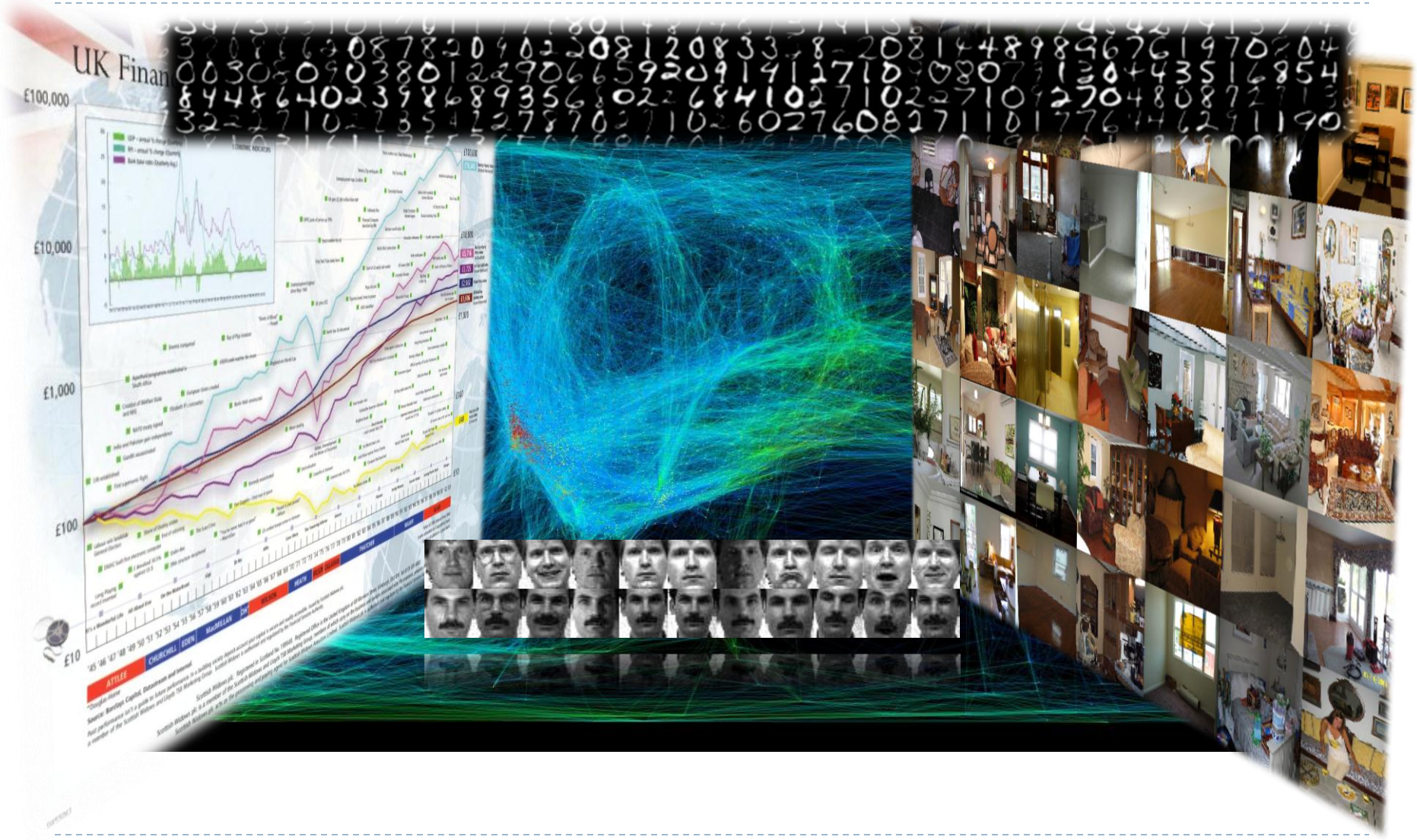
流形学习

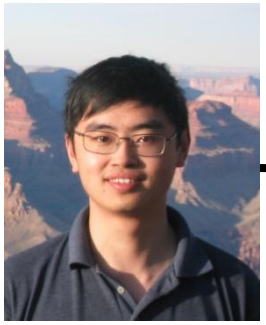
何晓飞
浙江大学

信息时代


UK Finan
£100,000
£10,000
£1,000
£100
£10
ATTLEE
CHURCHILL
EDEN
MacMILLAN
WILSON
HEATH
CALLAGHAN
THATCHER
MAJOR

机器学习问题

信息
(训练集)



f

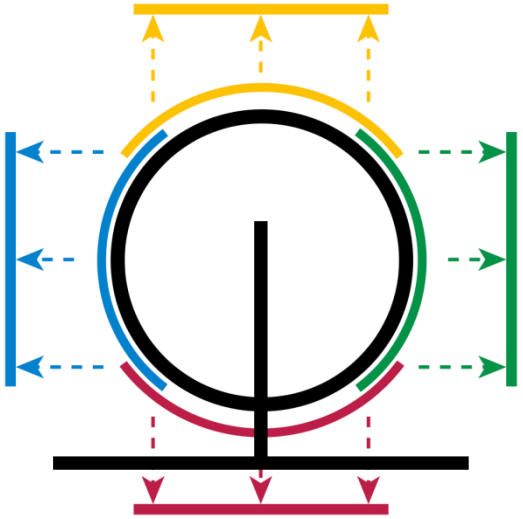
何晓飞

$f: X \to Y$

我们考虑的X和Y往往是欧氏空间

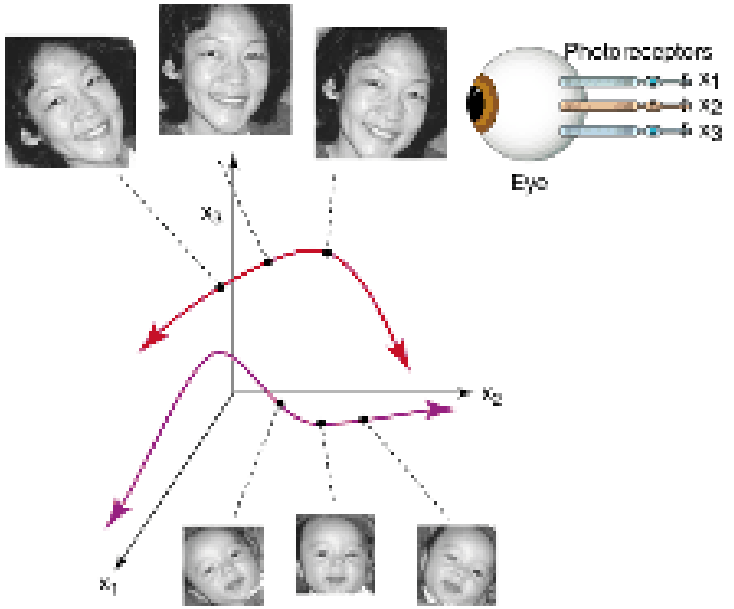
流形（Manifold）

- Manifold = Many + Fold, 很多曲面片的叠加
- 叠加但不是拼接，不自交
- 欧氏空间属于流形
- 任何一个流形都可以嵌入到足够高维度的欧氏空间中(Whitney嵌入定理)



流形假设

- 真实数据是怎样的？
- 外围欧氏空间的维度很高
- 数据存在一定的低维内在结构
- 我们假设数据是位于一个低维子流形上

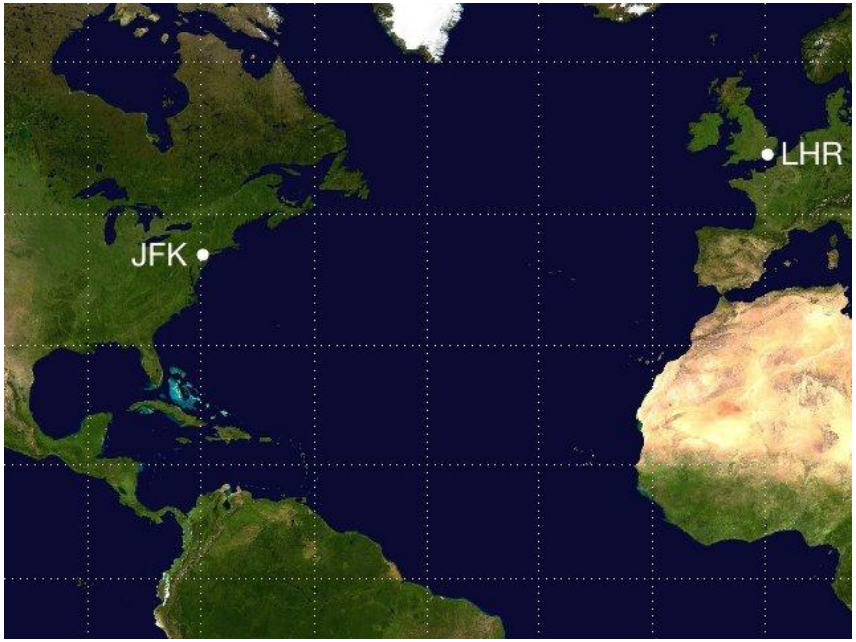


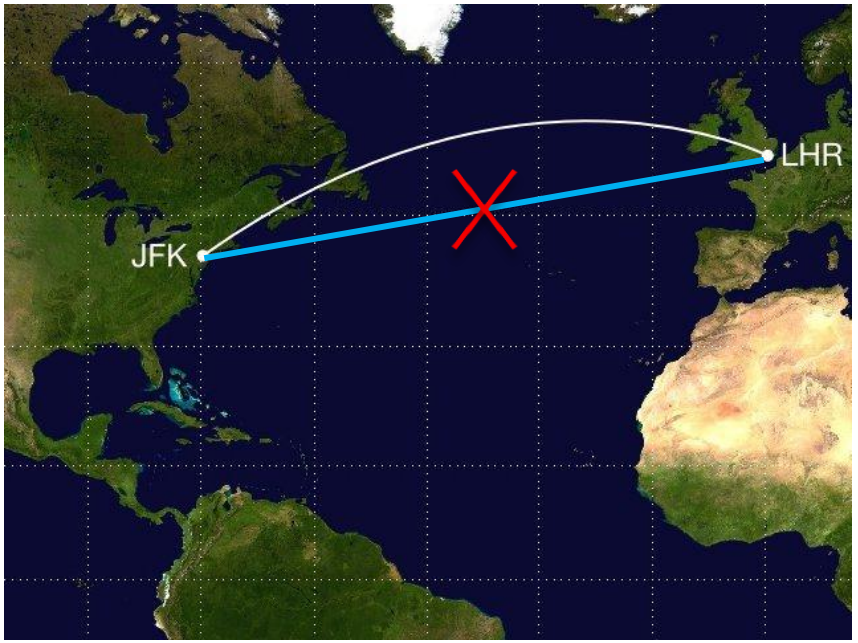
流形

我们的地球表面是二维流形　　局部可以认为是欧氏空间

测地线

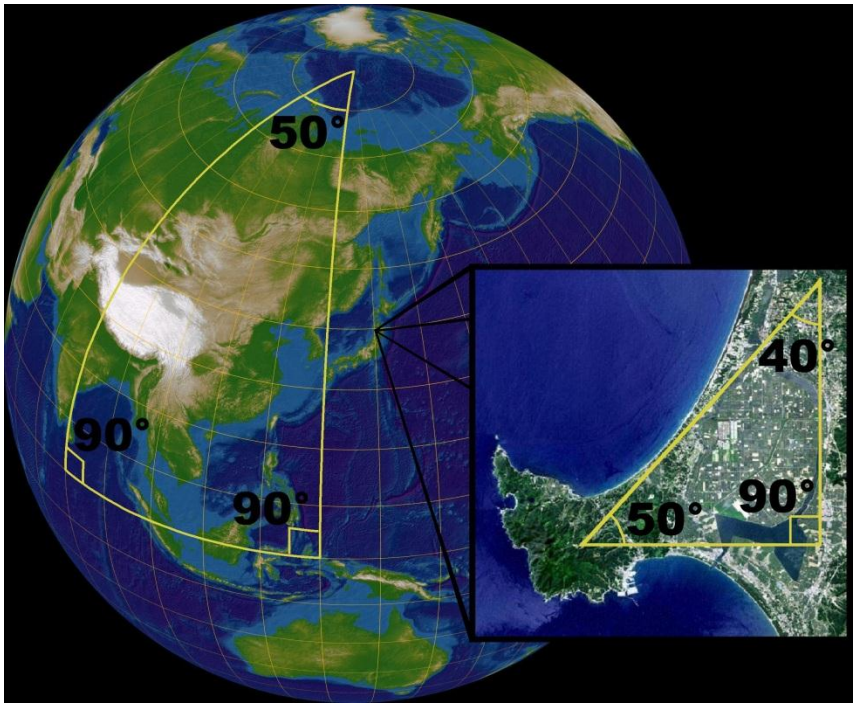
- 弯曲的'直线'，用来计算流形上两点的最短距离
- 球面上的测点线不是直线，因为直线根本不在球面上





流形的特殊性质

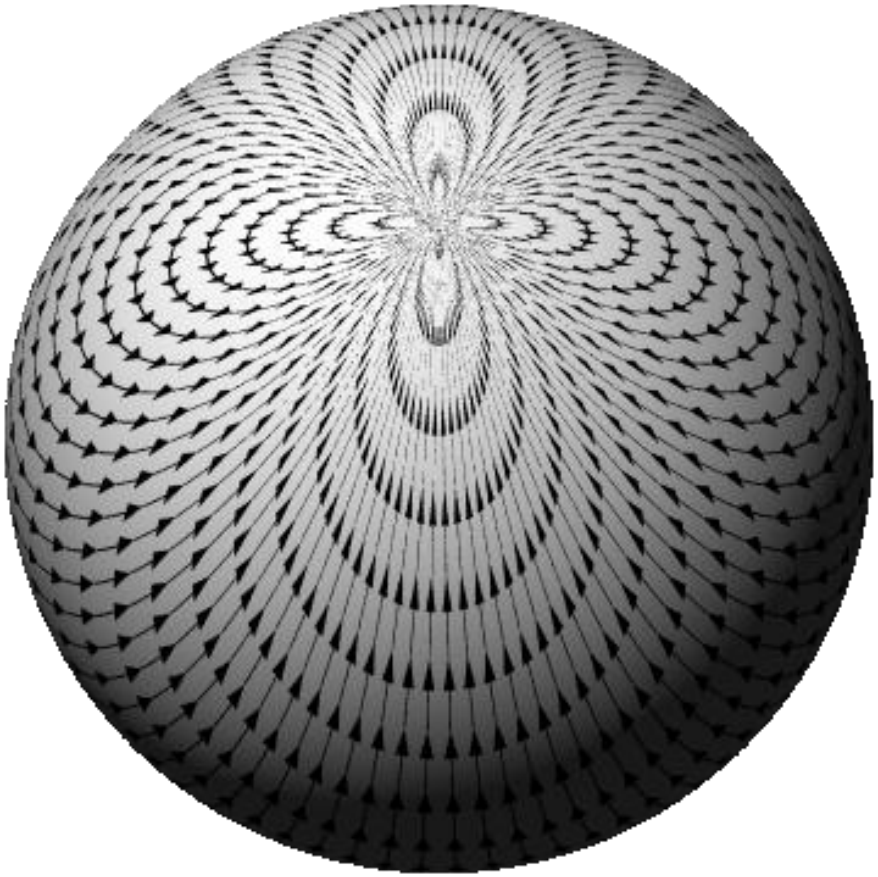
- 不满足平行公设：存在过空间中任意两点的平行直线(测地线)
- 球面：任意两条测地线(大圆弧)都相交
- 测地三角形的内角和不一定等于180度



1818年至1826年间，高斯曾测量在Harz山脉中由Inselberg、Brocken和Hoher三地形成的三角形，看看其内角和是否等于180度。

流形的特殊性质

- 地球上总有一点是风平浪静的



基本概念：拓扑空间

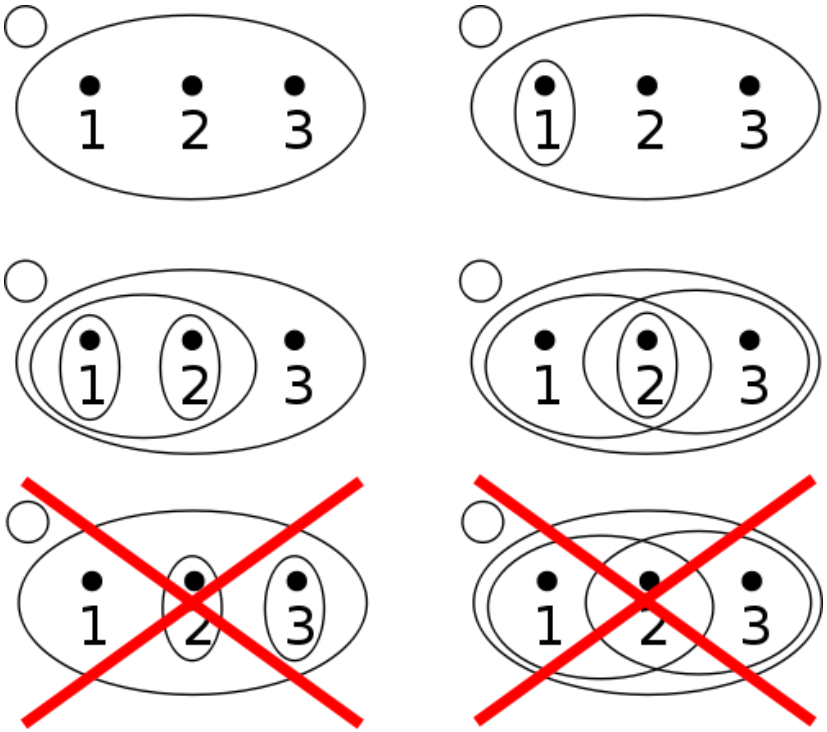
Topological spaces are mathematical structures that allow the formal definition of concepts such as convergence, connectedness, and continuity.

- A **topological space** is a set X together with τ (a collection of subsets of X) satisfying the following axioms:
  - The empty set and X are in τ.
  - τ is closed under arbitrary union.
  - τ is closed under finite intersection.

拓扑空间

- The collection τ is called a **topology** on X. The elements of X are usually called *points*, though they can be any mathematical objects. The sets in τ are called the open sets, and their complements in X are called closed sets. A subset of X may be neither closed nor open, either closed or open, or both.
- A function between topological spaces is called **continuous** if the inverse image of every open set is open.

拓扑空间

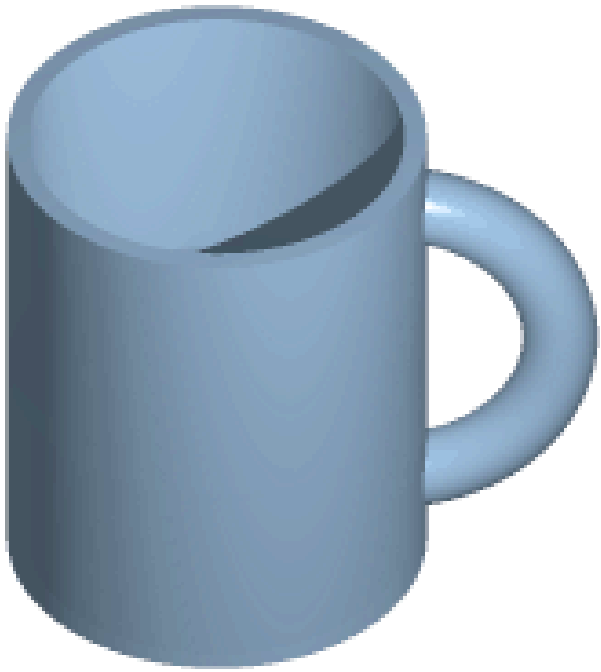


- Four examples and two non-examples of topologies on the three-point set {1,2,3}. The bottom-left example is not a topology because the union of {2} and {3} [i.e. {2,3}] is missing; the bottom-right example is not a topology because the intersection of {1,2} and {2,3} [i.e. {2}], is missing.

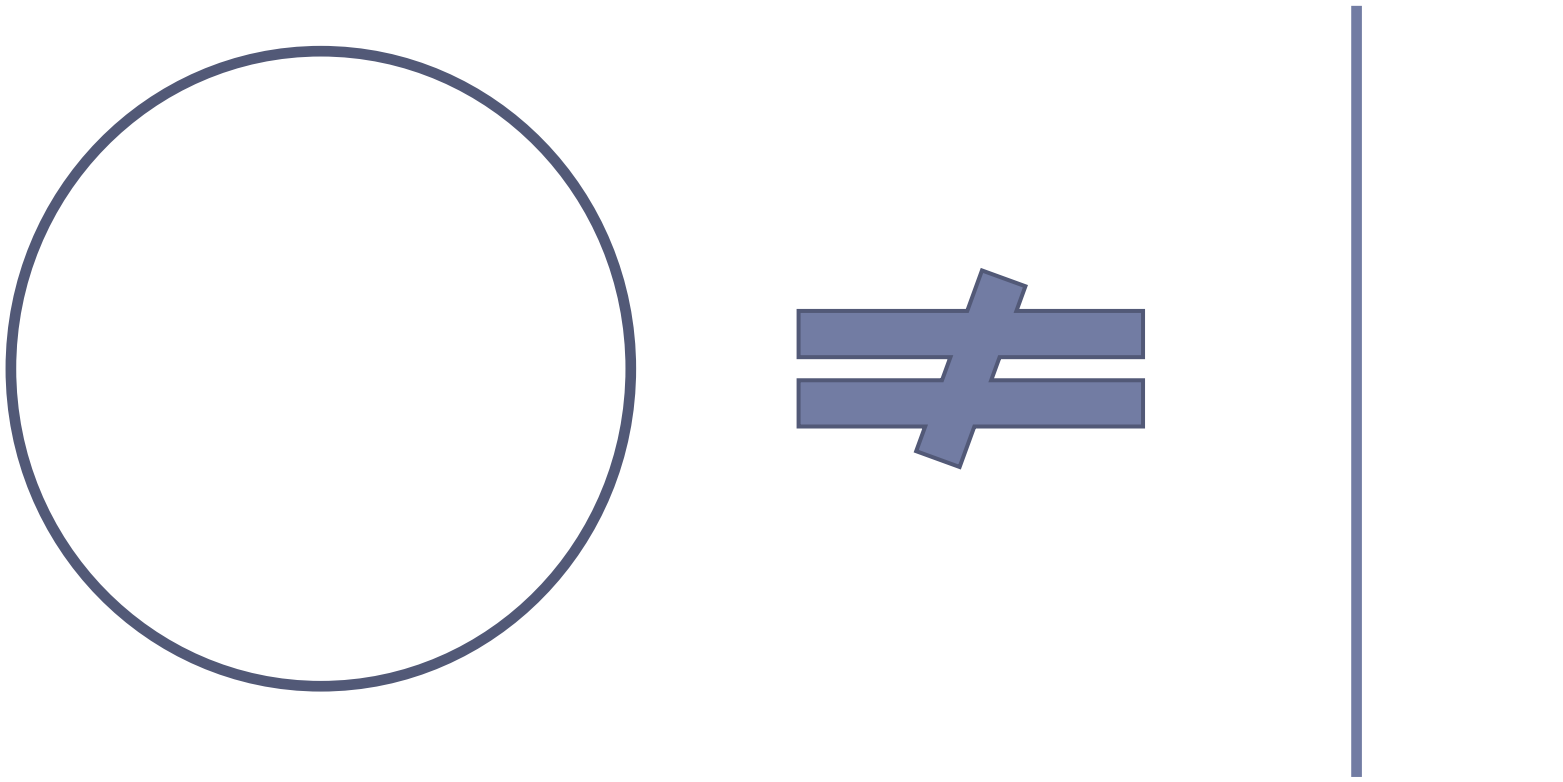
同胚（homeomorphism）

In the mathematical field of topology, a homeomorphism or topological isomorphism or bicontinuous function is a continuous function between two topological spaces that has a continuous inverse function. Homeomorphisms are the isomorphisms in the category of topological spaces—that is, they are the mappings that preserve all the topological properties of a given space. Two spaces with a homeomorphism between them are called homeomorphic, and from a topological viewpoint they are the same.

同胚（homeomorphism）



同胚（homeomorphism）



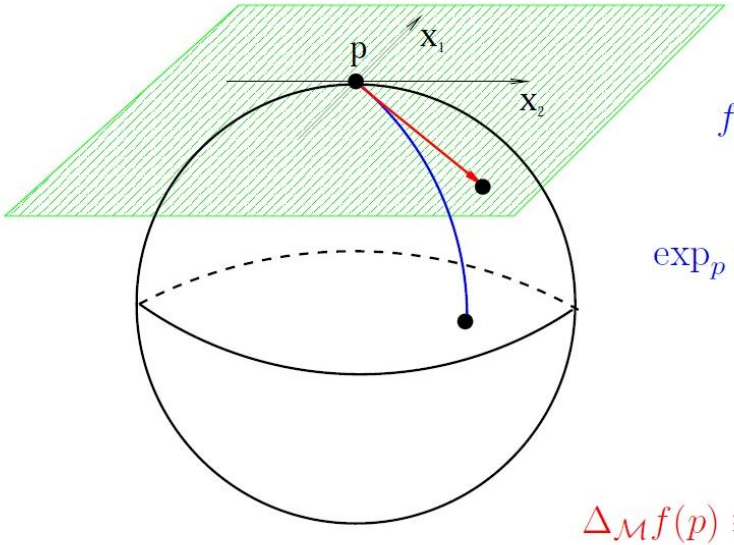
流形上的拉普拉斯算子

- 流形上的学习问题往往是用微分算子表示的微分方程问题
- 微分几何中最重要的微分算子，表示为L或者Δ
- 度量了流形上函数的光滑性

R^3中的拉普拉斯算子：

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

一般流形上的拉普拉斯算子：



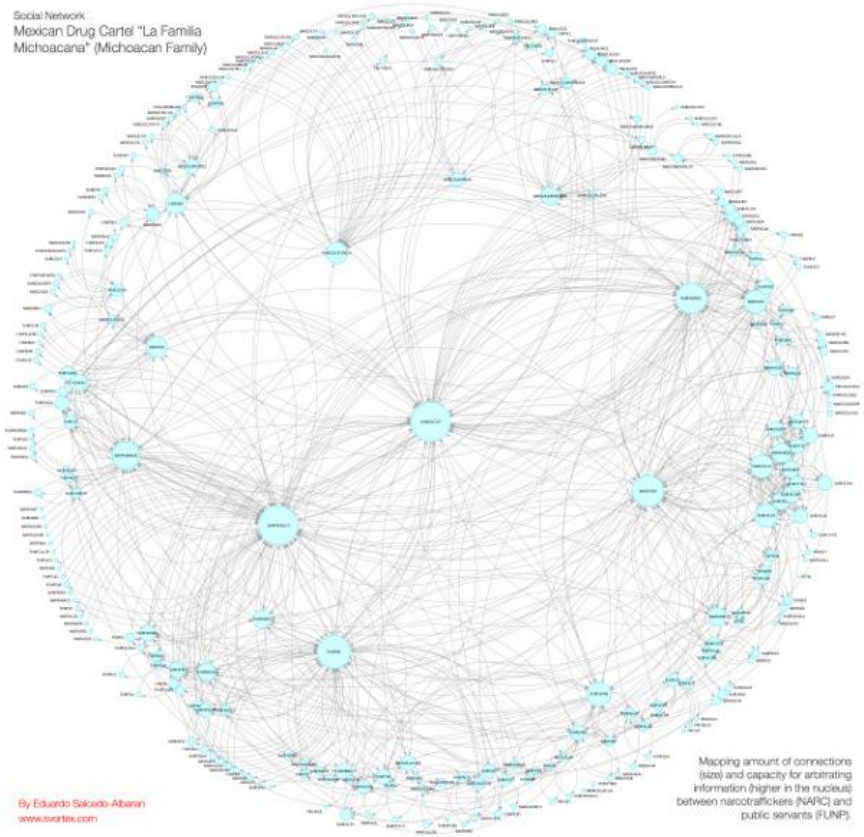
$$f : \mathcal{M}^k \to \mathbb{R}$$

$$\exp_p : T_p\mathcal{M}^k \to \mathcal{M}^k$$

$$\Delta_{\mathcal{M}} f(p) \equiv \sum_i \frac{\partial^2 f(\exp_p(x))}{\partial x_i^2}$$

表示流形的离散数学模型

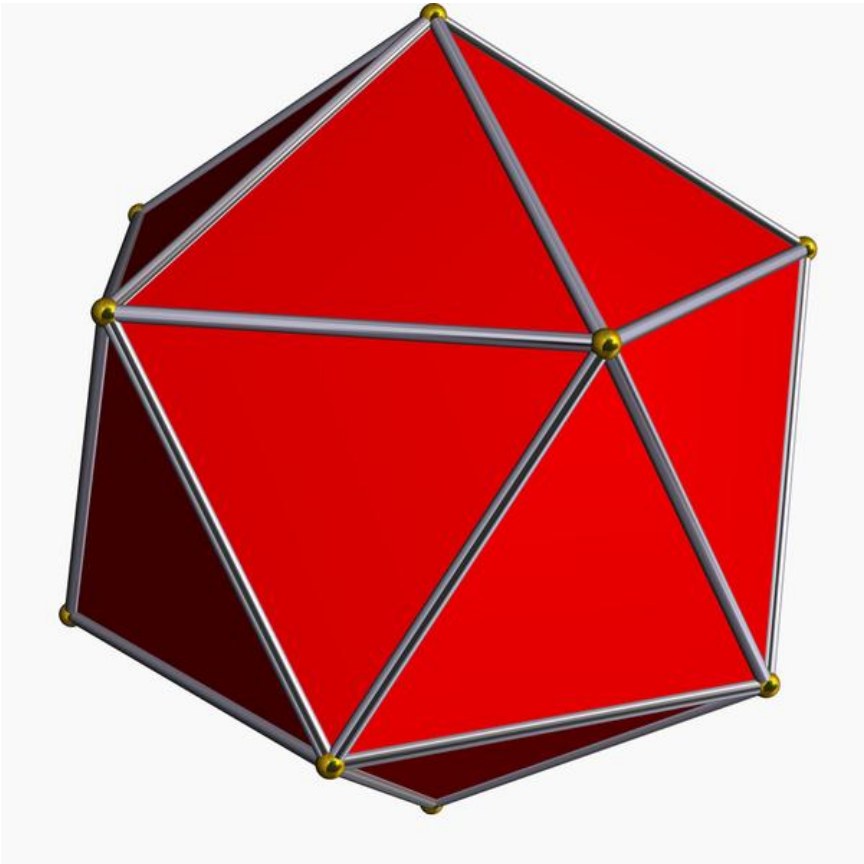
- 图



社会网络

表示流形的数学模型

- 单纯复形：曲面片的拼接
- 图是1维的单纯复形



模型的选择

- 简单模型，对数据要求低，对流形的描述不太准确
- 复杂模型，对数据要求高，对流形的描述很准确
- 目前的流形学习基本上都是基于图模型
- 研究拓扑结构的时候要用到单纯复形，图模型不能刻画高维拓扑

流形学习

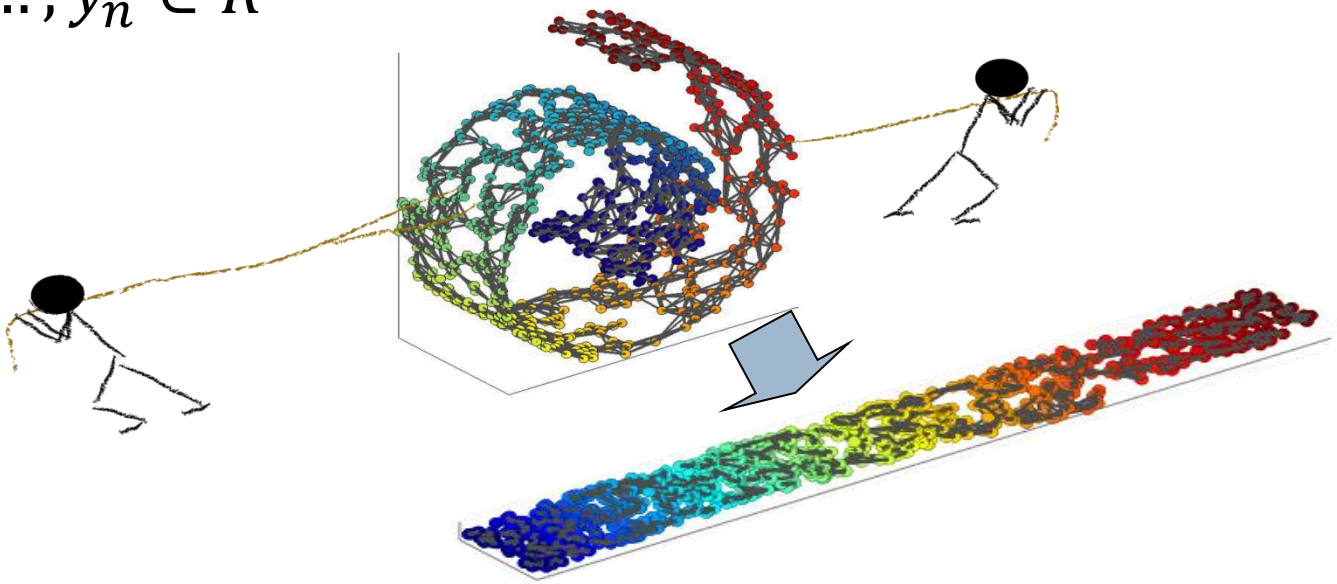
- 研究在流形假设下的机器学习问题，数据流形$M \subset R^N$
- 聚类$f: M \to \{1, \ldots, k\}$
  - 例子：图像分割，社会网络分析，数据挖掘等等
- 分类/回归$f: M \to \{-1, +1\}$ 或者$f: M \to R$
  - 例子：语音识别，手写体识别，文本分类等等
- **降维:$f: M \to R^n, n \ll N$**
  - 例子：可视化，应用于后续学习

流形学习热点问题

- 研究数据流形的几何和拓扑
  - 流形学习的中心问题
- 降维：找一个映射从流形到欧氏空间
  - 经典算法: ISOMAP, LLE和LE
- 根据流形结构进行学习
  - 半监督学习
  - 主动学习

降维

- Unfold a manifold，展开一个流形
- 保持流形的几何结构，理想情况下希望能够保持测地距离
- 问题描述：给定数据点$x_1, \ldots, x_n \in M \subset R^N$，求$y_1, \ldots, y_n \in R^d$



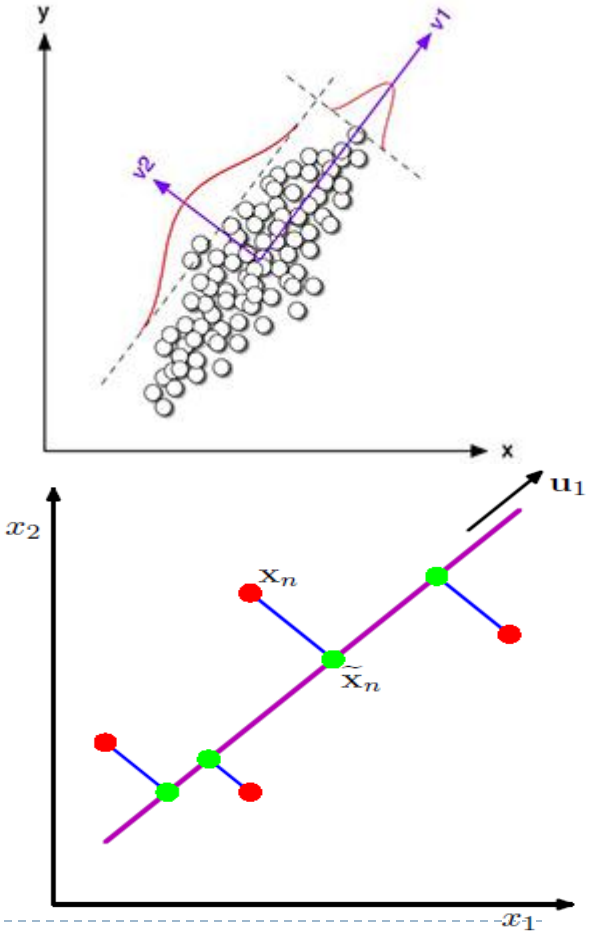
PCA：传统降维方法

- Principal Component Analysis 采用线性投影的方法进行降维，它的目的是使得数据在给定的方向上投影会得到最大的方差。

$$\max_{\|w\|=1} Var\{w^T X\}$$

- 也等价于点 x_n 和投影之后得到的点 $\tilde{x}_n$ 之间的距离最小。

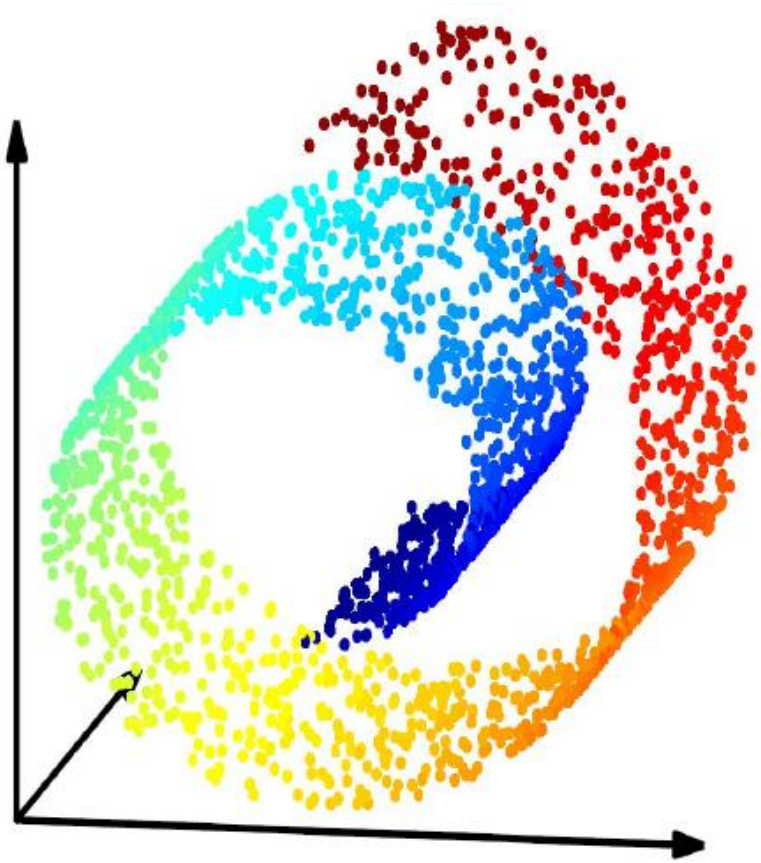
$$\min_{W^T W=I} \sum_{n=1}^{N} \|x_n - WW^T x_n\|^2$$

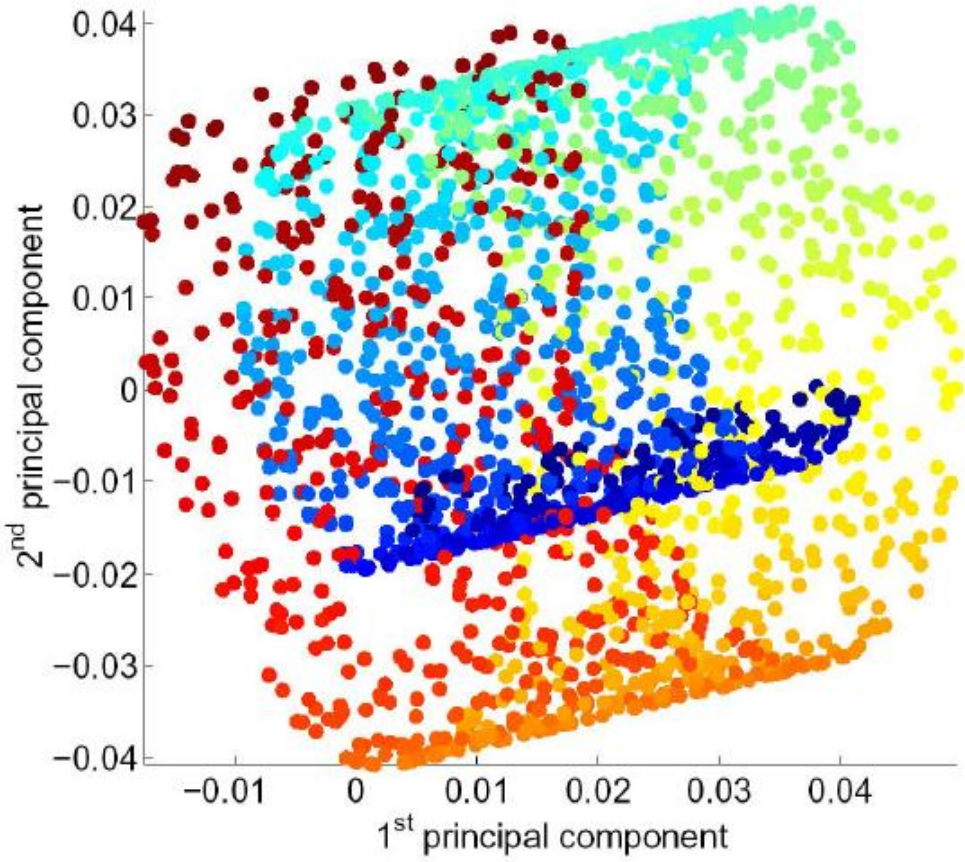


PCA 的地位和重要性

- PCA 是到目前为止应用最为广泛的一个降维算法
  - 在模式识别、金融、生物信息学等各个领域均得到广泛应用
  - 在机器学习本身的众多场景中也通常被用作数据预处理的首要方法
  - 发展出了各种变种（例如 Sparse PCA、Online PCA、Robust PCA、Probabilistic PCA 等）和扩展工作
- 当流形是一个线性流形时，PCA 得到的结果是最优的

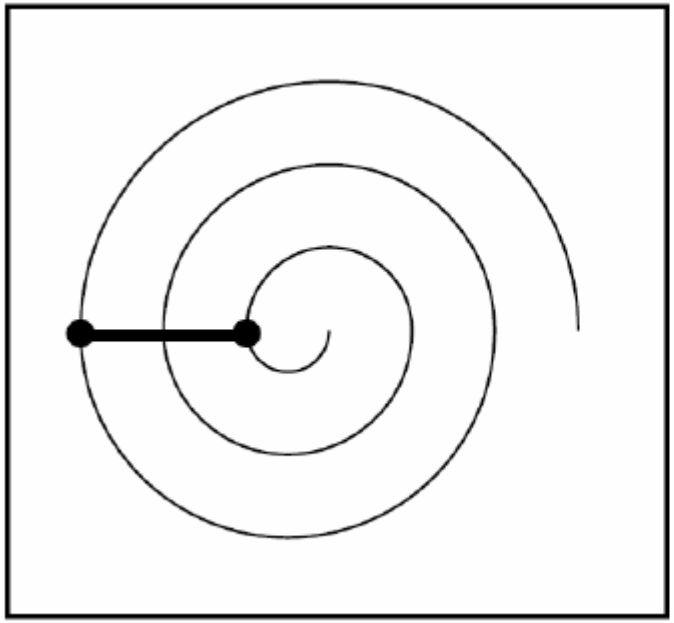
PCA 无法处理非线性流形



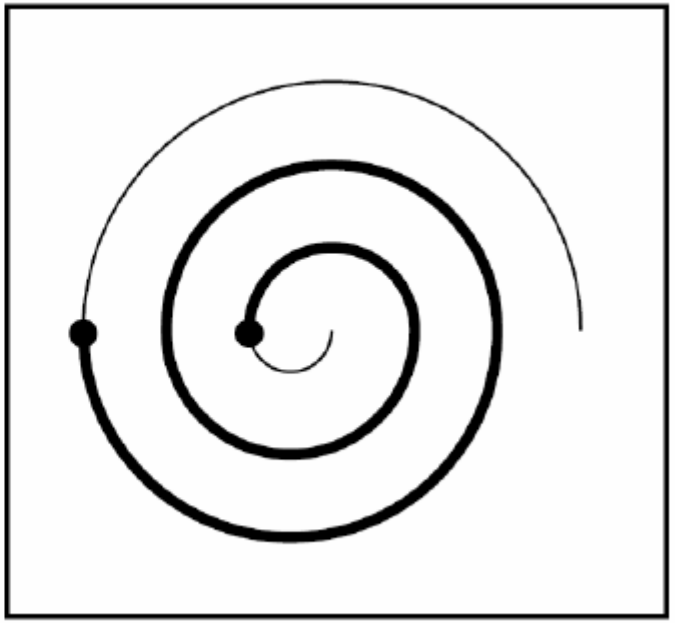


ISOMAP

- Isomap 希望在映射过程中保持流形上测地线的距离



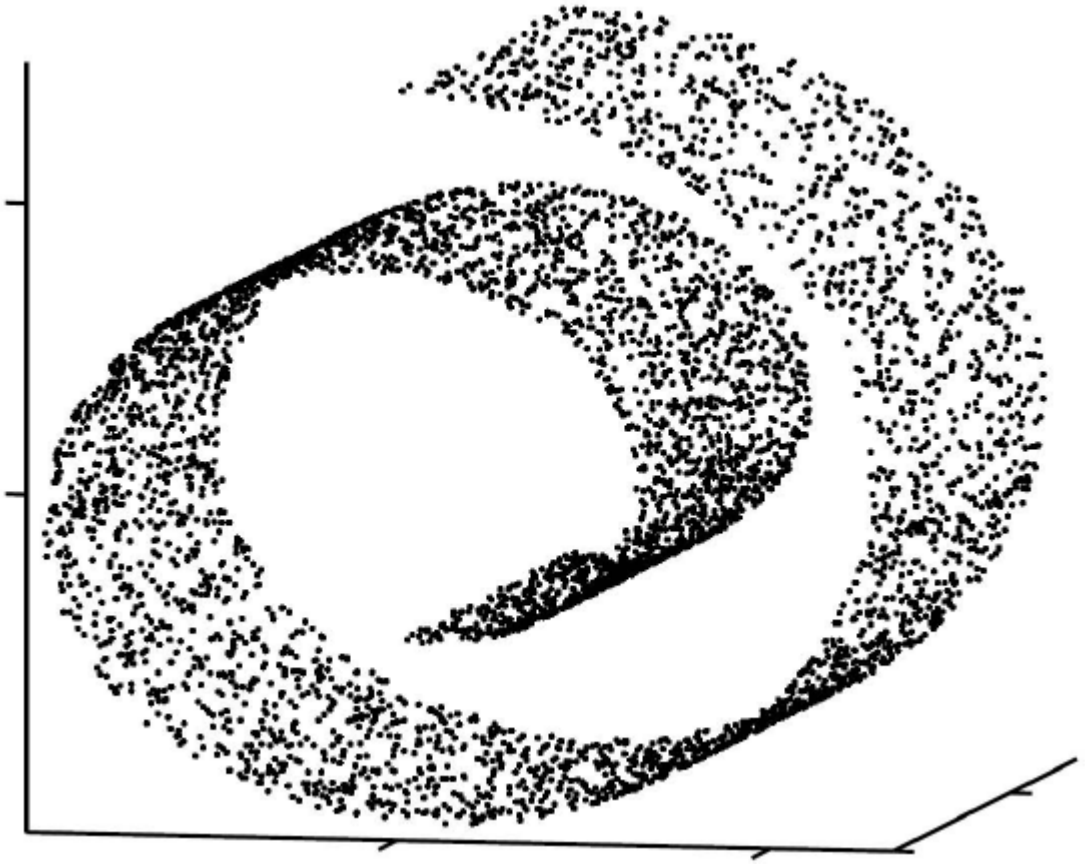
两点间的欧氏距离



两点间的测地距离

ISOMAP：测地线的计算

- 在流形结构未知的情况下，如何根据有限的数据采样来估算流形上的测地线？

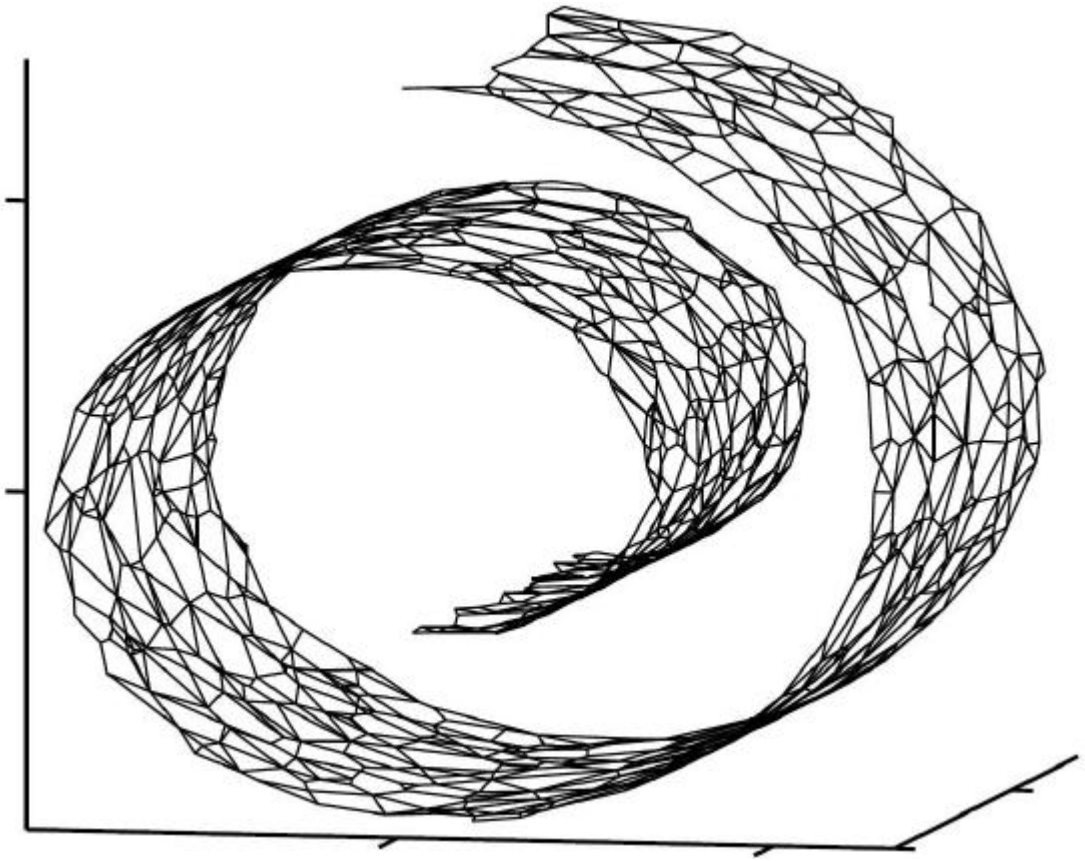


ISOMAP：测地线的计算

- 在流形结构未知的情况下，如何根据有限的数据采样来估算流形上的测地线？

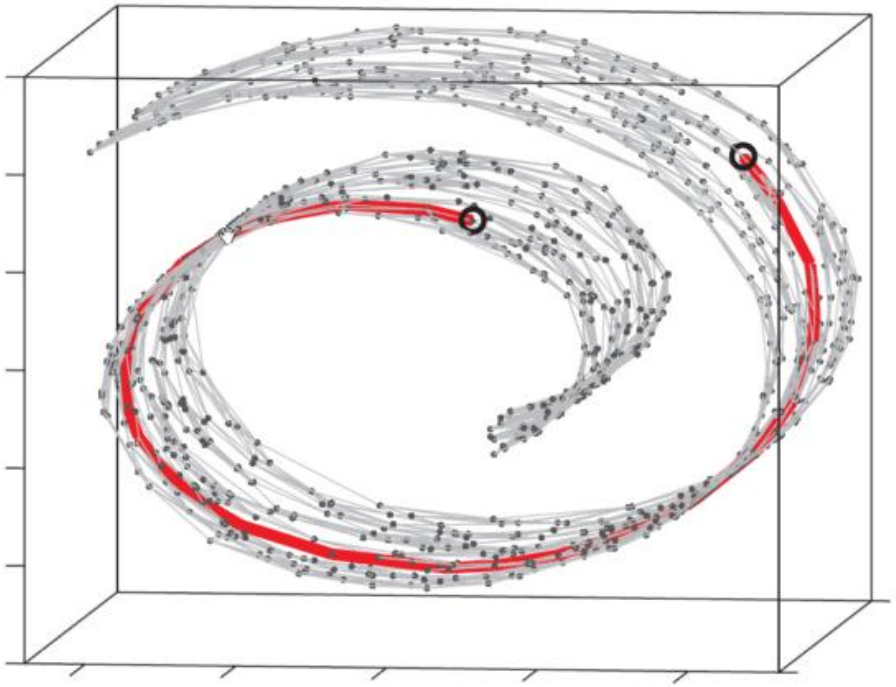
构造邻接图（Graph），用图上的最短距离来近似测地线。

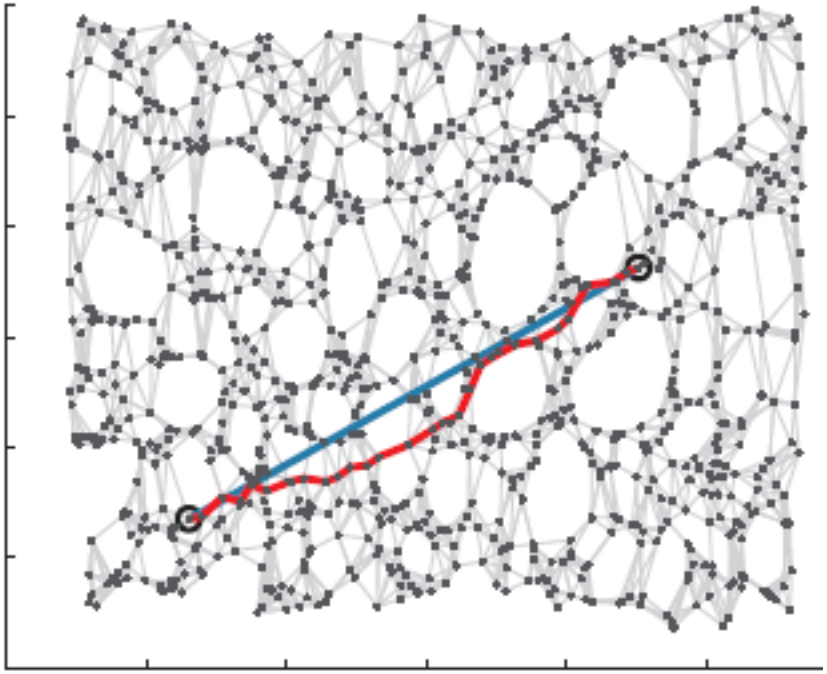
$$W_{ij} = \begin{cases} 1 & \|x_i - x_j\| < \epsilon \\ 0 & otherwise \end{cases}$$



ISOMAP：测地线的计算

- 图上两点之间的最短路径（可以用Dijkstra或者Floyd算法来计算）对应于流形上测地线距离的一个近似值。当数据点趋向于无穷多时，这个估计趋向于真实的测地线距离。

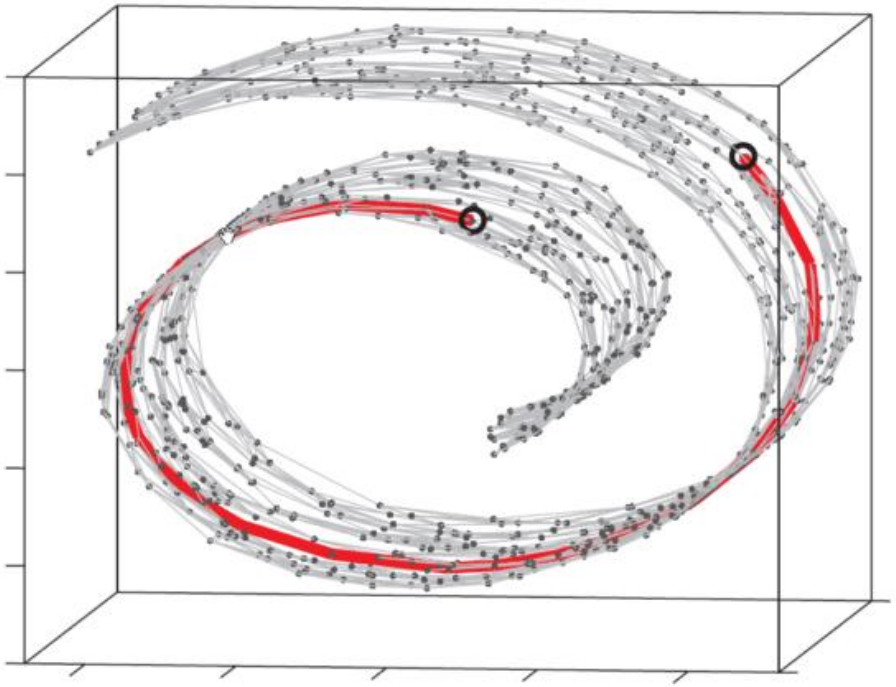


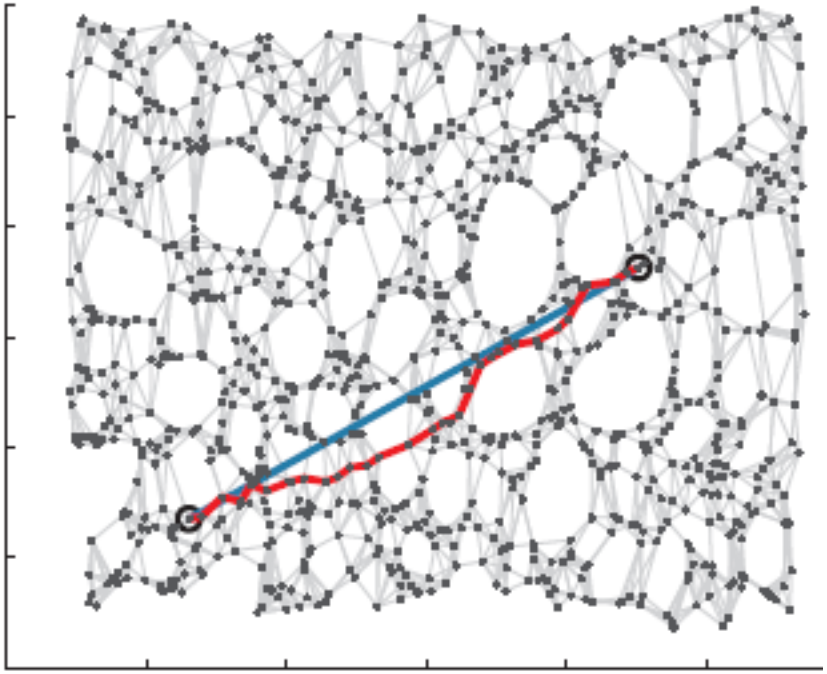


ISOMAP：降维坐标的计算

- Isomap使用 MDS 计算映射后的坐标 y，使得映射坐标下的欧氏距离与原来的测地线距离尽量相等

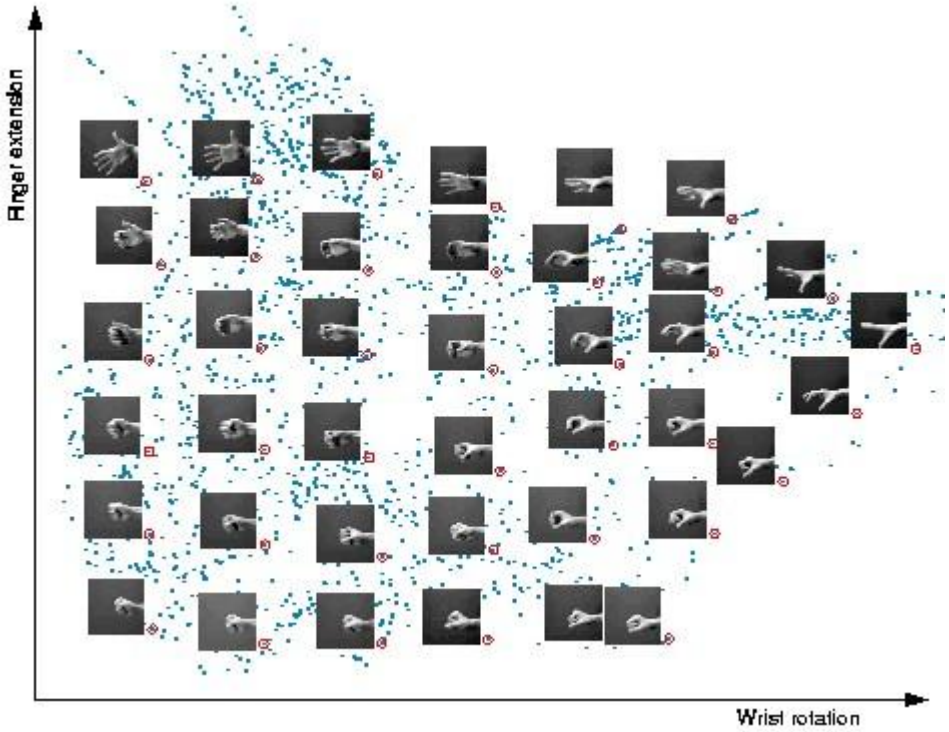
$$\min_{y} \sum_{i,j} \left(d_{\mathcal{M}}(x_i, x_j) - \|y_i - y_j\| \right)^2$$





ISOMAP：直观示例

- 将一张图片看成一个数据点，每个像素是一个维度，一张 $n \times m$ 的图像就是一个 nm 维欧氏空间中的一个点
- 另一方面，数据集中的手的图像只有“张开”和“闭合”以及旋转角度两个自由度，所以这些图片其实分布在一个二维流形上。



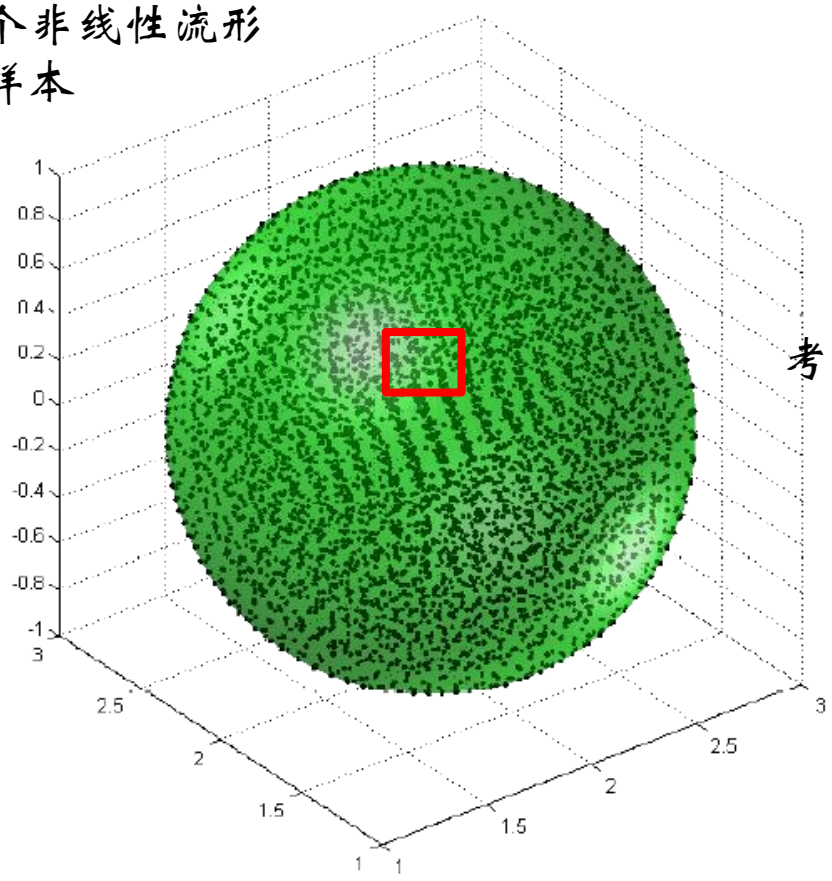
Isomap 在保持流形测地线距离的前提下将数据点映射到二维欧氏空间中，如图所示。其中在各个位置选取了一些代表性的点（红圈标出）将它所对应的原始图片画在旁边。可以看到手的图像在两个自由度上的相似性（距离）得到了保持。

LLE

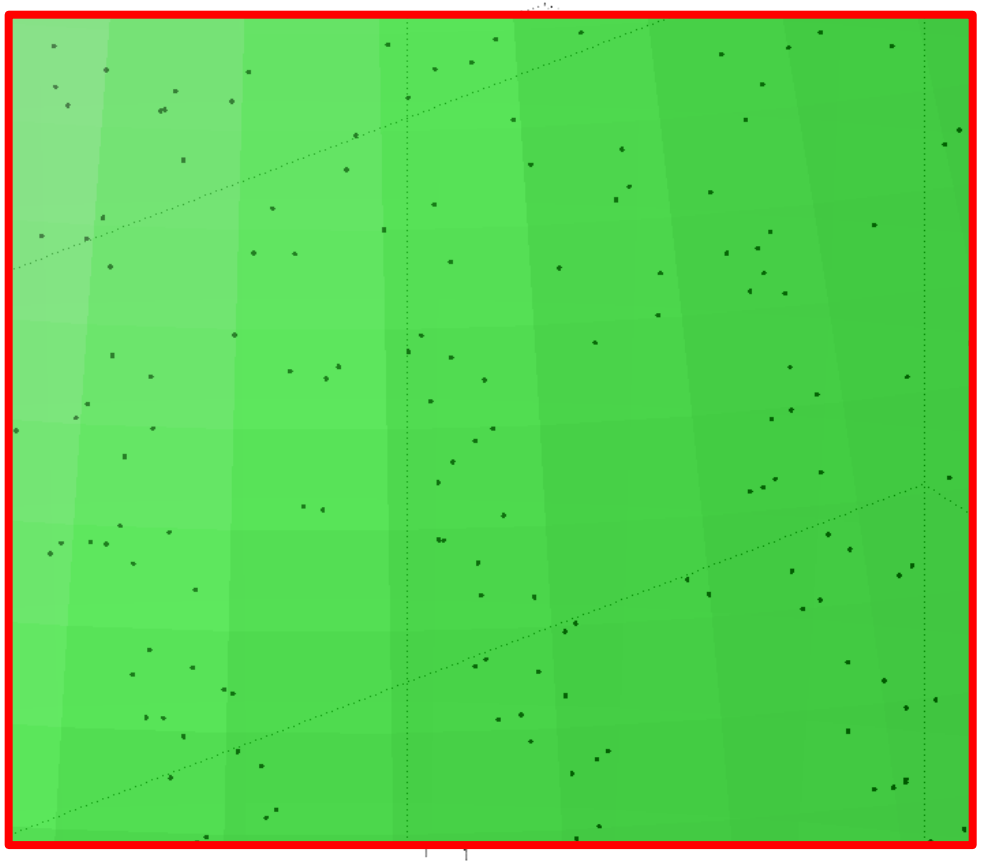
- Isomap 试图通过保持任意两点之间的测地线距离来保持流形的全局几何结构；而 LLE 则从局部来进行分析。
- “流形在局部可以近似等价于欧氏空间”便是 LLE 分析方法的出发点。

下面来看一个例子

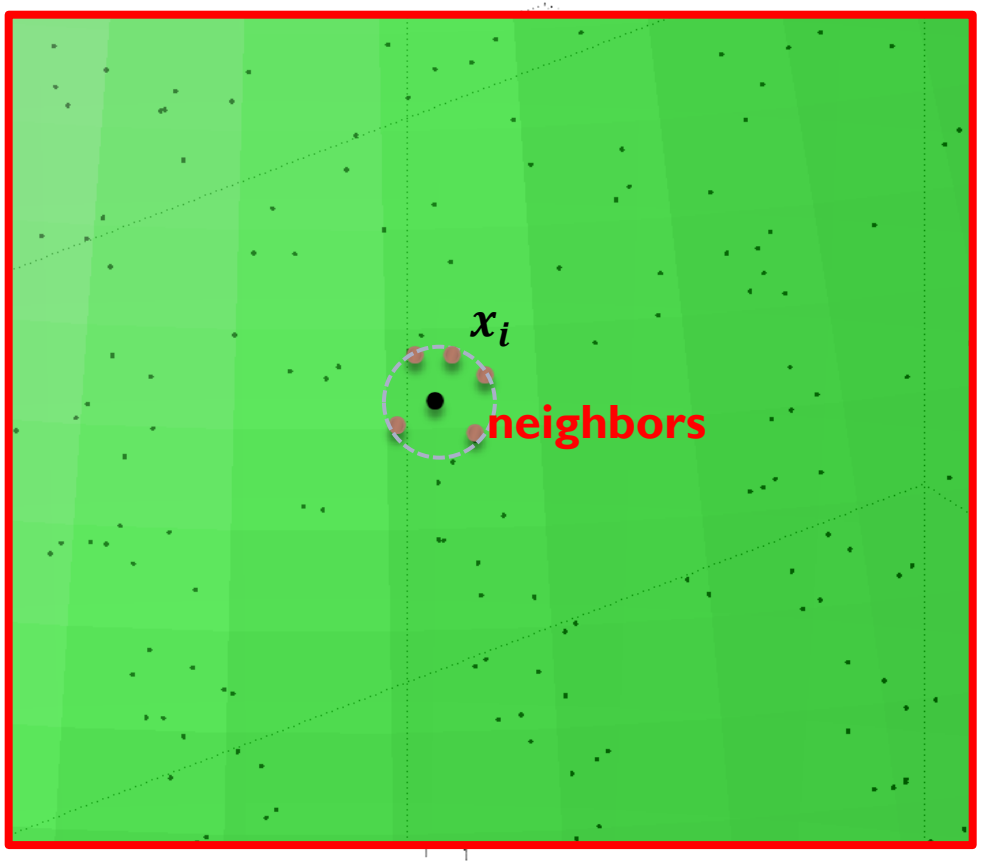
分布在一个非线性流形上的数据样本



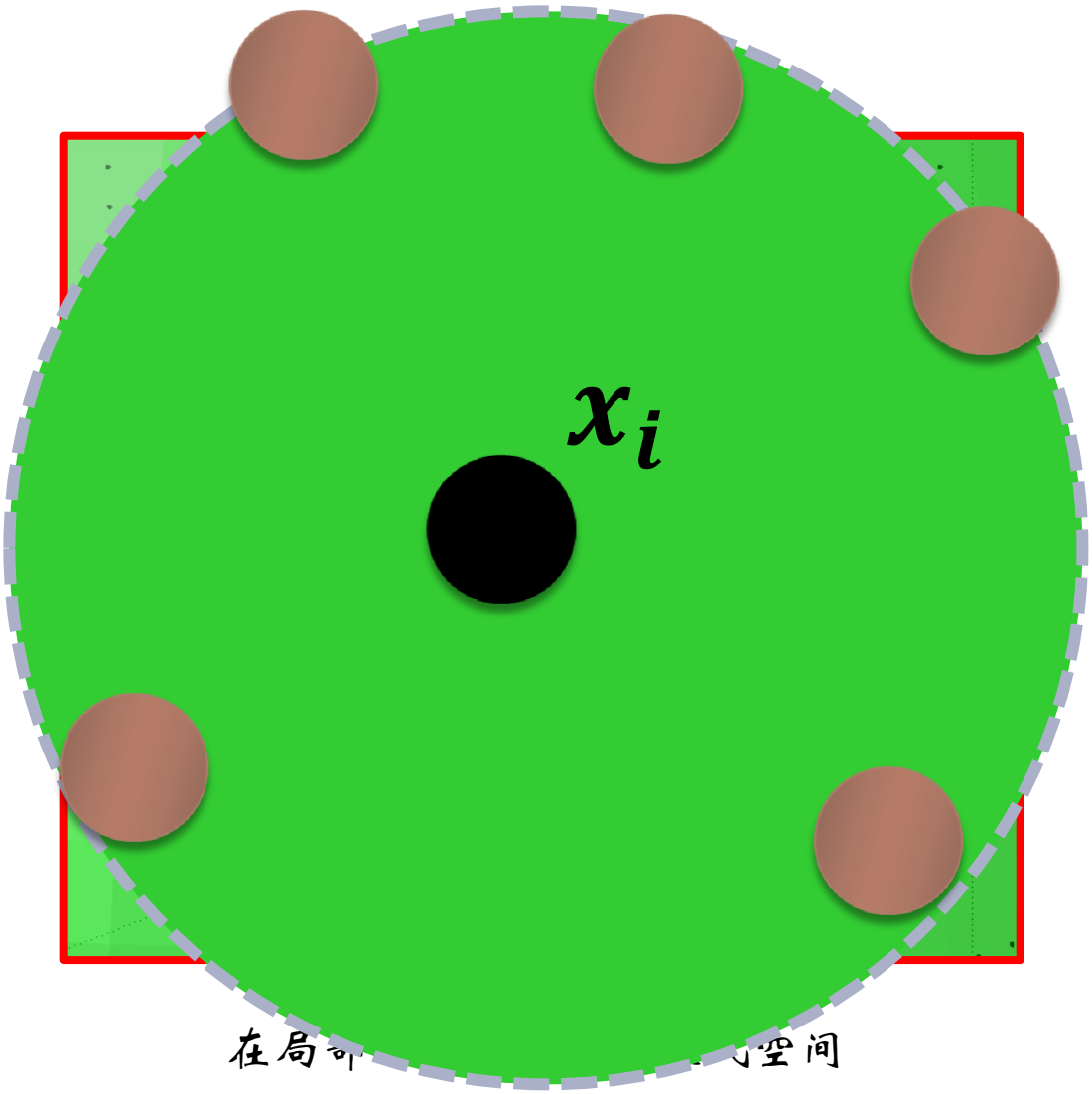
考虑一个局部领域



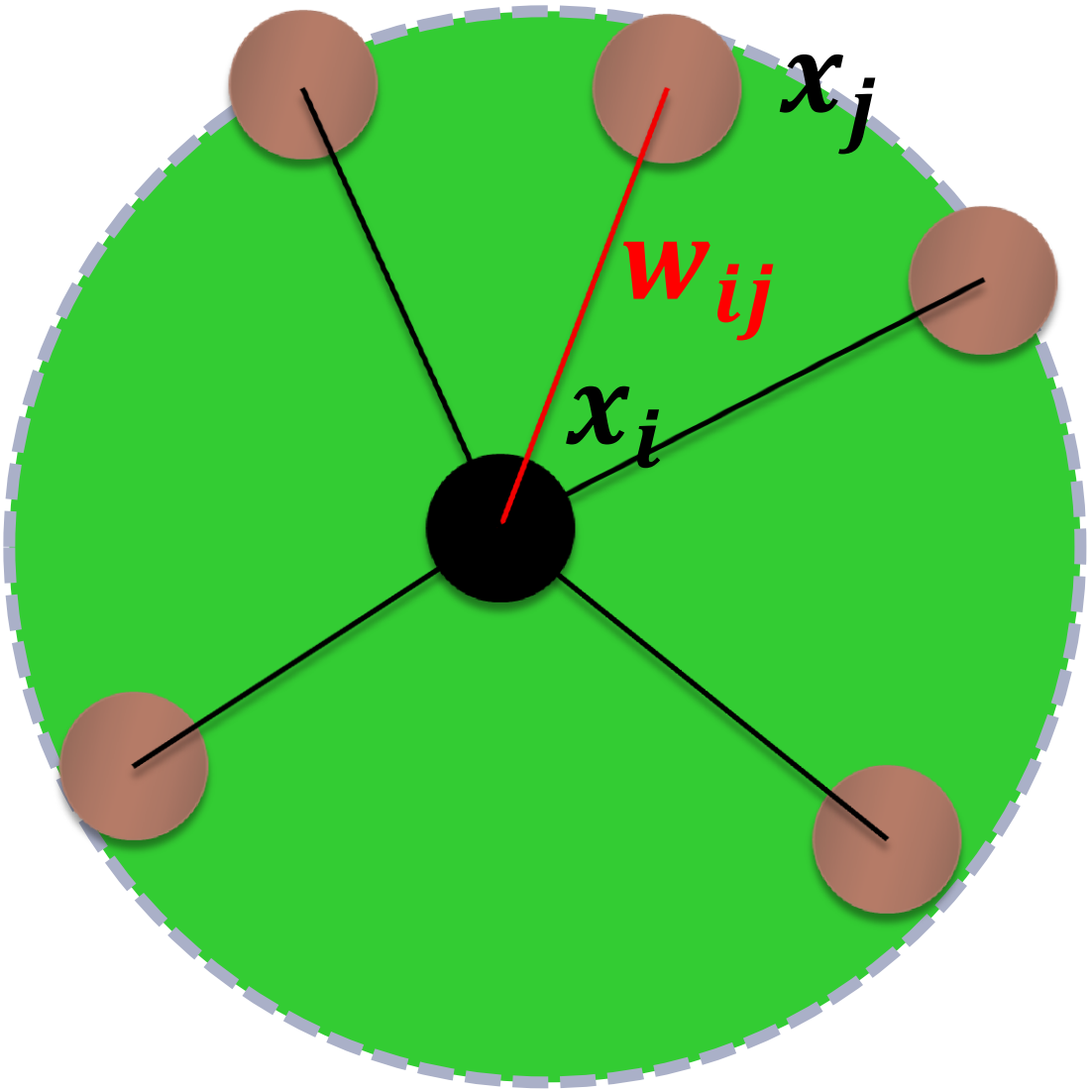
在局部上近似于一个欧氏空间



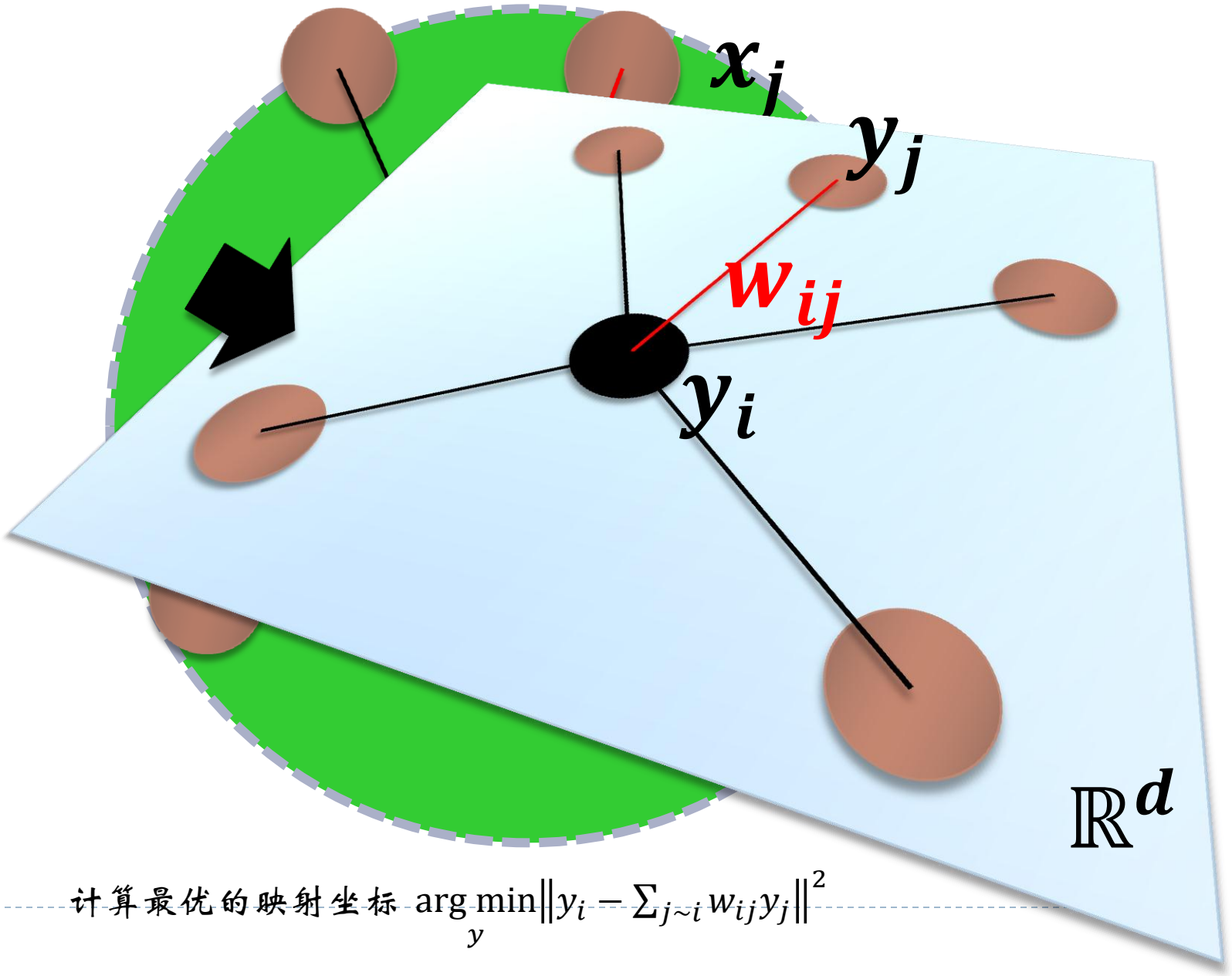
在局部上近似于一个欧氏空间



在局部
空间



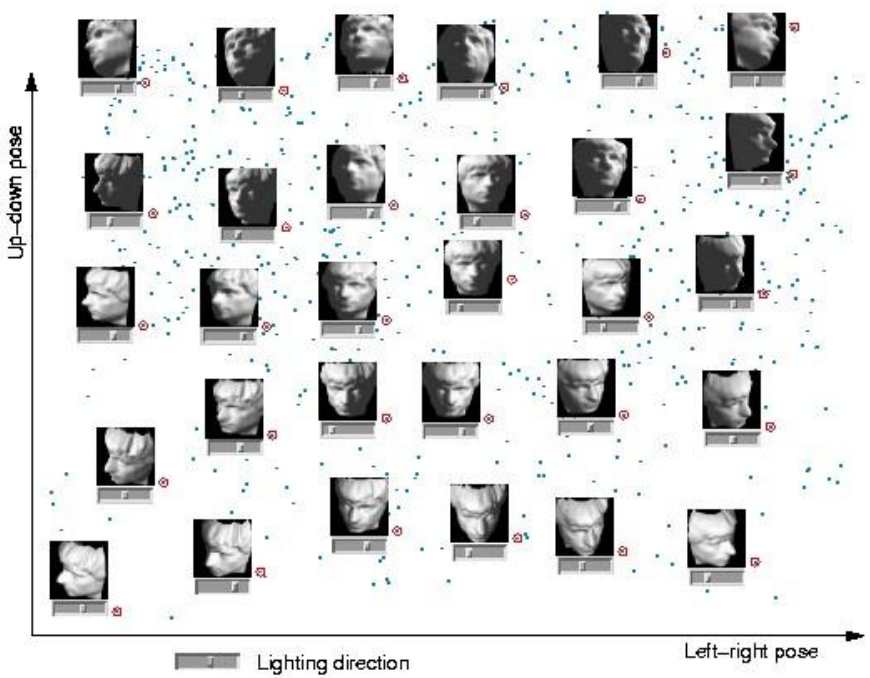
计算最优的重构权重 $\arg\min_{w_i} \left\| x_i - \sum_{j \sim i} w_{ij} x_j \right\|^2$

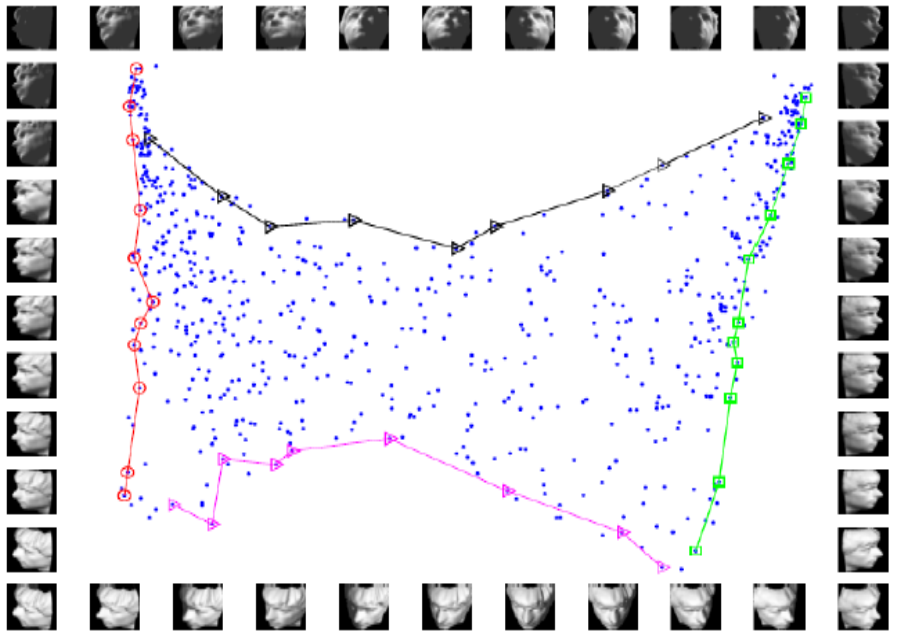


计算最优的映射坐标 $\arg\min_{y} \left\| y_i - \sum_{j \sim i} w_{ij} y_j \right\|^2$

ISOMAP vs. LLE

- Isomap 和 LLE 从不同的出发点来实现同一个目标，它们都能从某种程度上发现并在映射的过程中保持流形的几何性质。



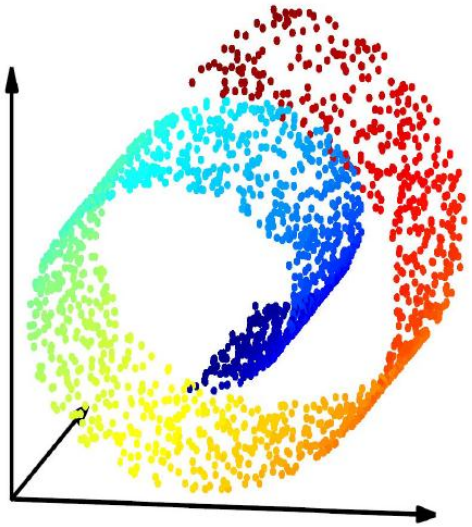


Isomap

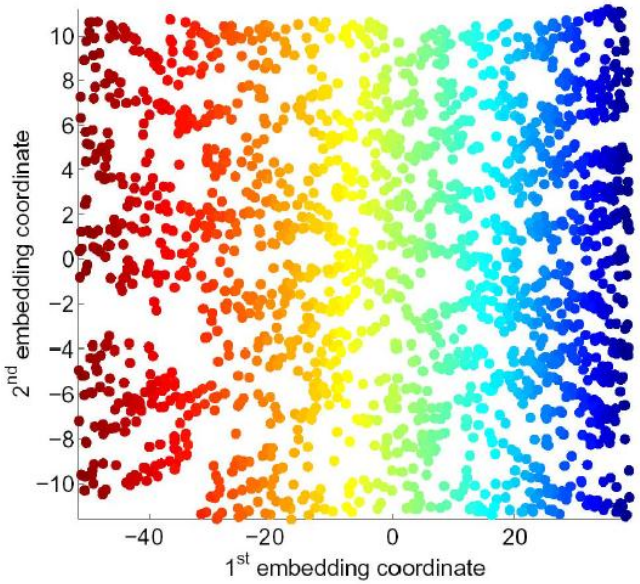
LLE

ISOMAP vs. LLE

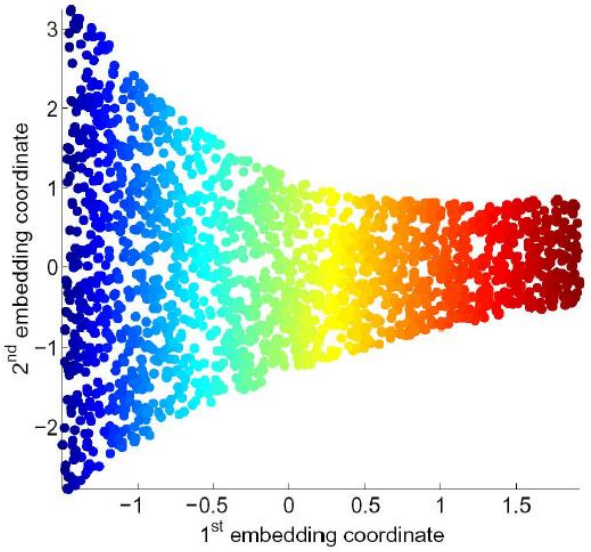
- Isomap 希望保持任意两点之间的测地线距离；LLE 希望保持局部线性关系。
- 从保持几何的角度来看，Isomap保持了更多的信息量



采样数据



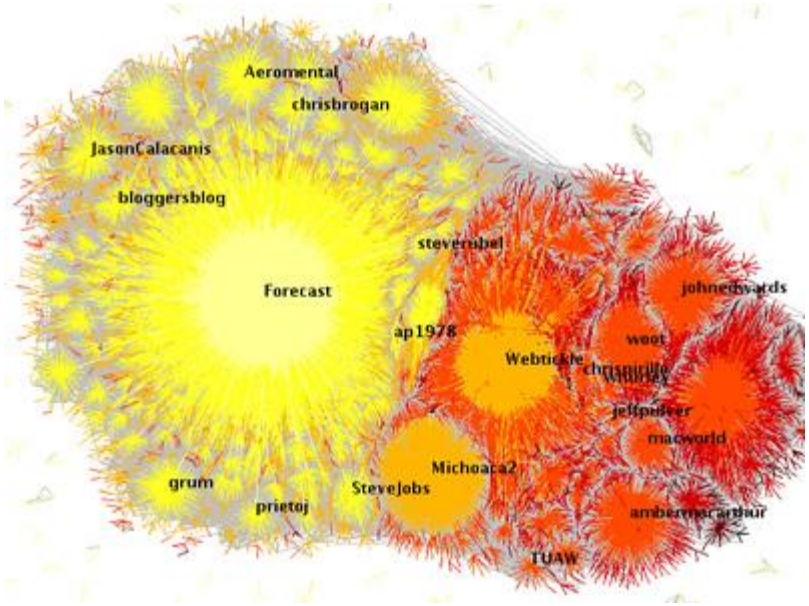
Isomap



LLE

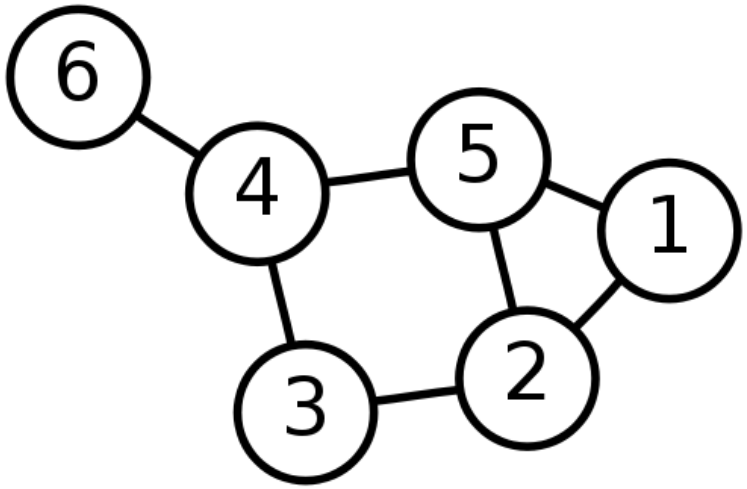
Global vs. Local

- 然而 Isomap 的全局方法有一个很大的问题就是要考虑任意两点之间的关系，这个数量将随着数据点数量的增多而爆炸性增长，从而使得计算难以负荷。
- 另一方面，随着互联网的发展，我们所面临的数据规模正变得越来越大，例如图中所示的 twitter 社交网络于2008年的一个子集所构成的一个图，包含大约2万个节点、25万条边。Twitter 现在的用户数量已经超过一亿，并且还在飞速增长。诸如此类的巨型结构使用全局方法进行分析正在变得越来越不切实际。
- 因此，以 LLE 为开端的局部分析方法的变种和相关的理论基础研究逐渐受到更多的关注。。



谱图理论

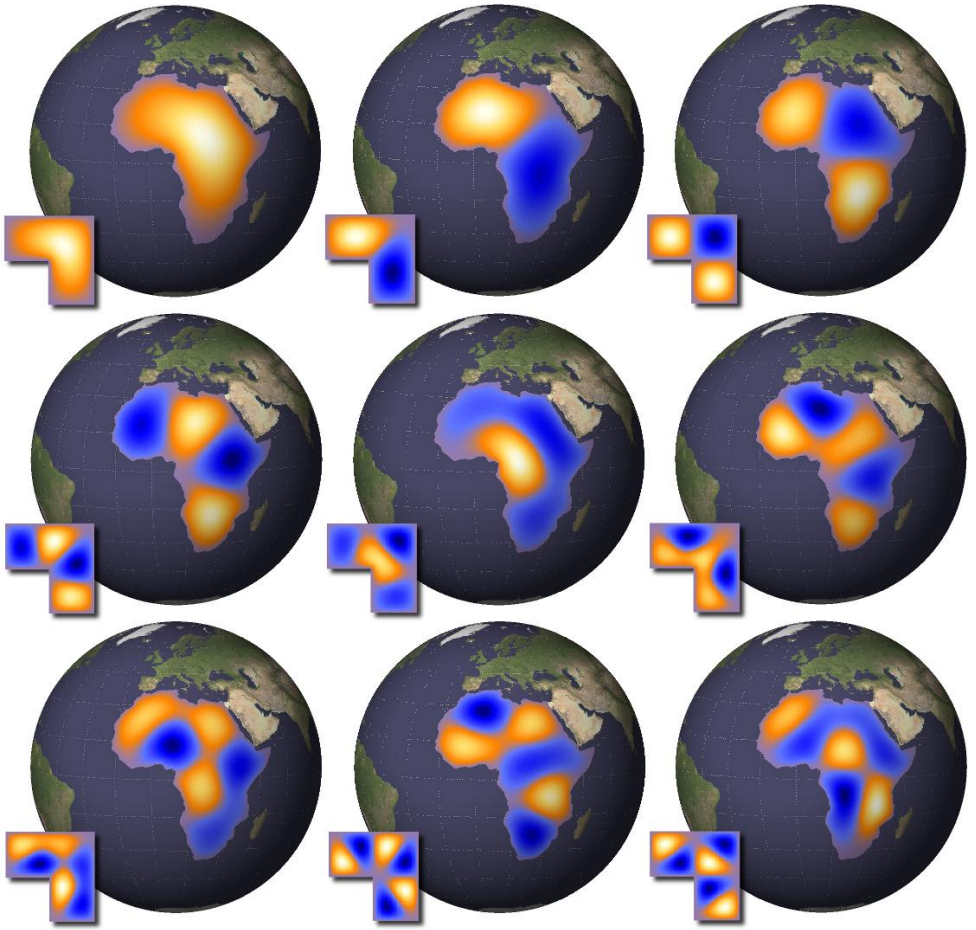
- 图上的拉普拉斯算子：拉普拉斯矩阵
- $L = D - W$，D是对角线元素为度数的对角阵，W是权重矩阵



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} \end{array}$$

谱图理论

- 拉普拉斯矩阵特征向量组成了一组正交向量基，大量应用于学习问题
- “非洲”的特征向量

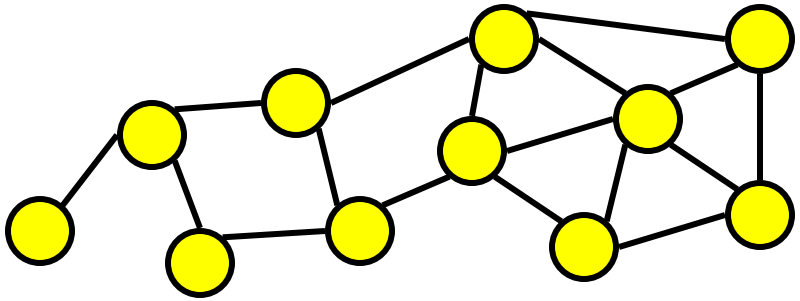


谱图理论

- 图上的拉普拉斯算子收敛到流形上的拉普拉斯算子
  - M. Belkin, P. Niyogi, Towards a theoretical foundations for Laplacian based manifold methods, COLT 2005.
  - M. Hein, et al., From graphs to manifolds – weak and strong pointwise consistency of graph Laplacian, COLT 2005.
  - A. Singer, From graph to manifold Laplacian: the convergence rate, Applied and Computational Harmonic Analysis, 2006
- 拉普拉斯矩阵的特征向量收敛到流形上拉普拉斯的特征函数
  - U. von Luxburg, et al., Consistency of spectral clustering, Max Planck Institute technique report, 2004
  - M. Belkin, P. Niyogi, Convergence of Laplacian eigenmaps, NIPS 2006.

LE

- 希望保持流形的近邻关系：将原始空间中相近的点映射成目标空间中相近的点
- 目标函数：$\mathrm{E}(\mathrm{y}) = w_{ij}(y_i - y_j)^2$
- 约束条件：$y^T y = 1$，去除任意的缩放
- 转化成一个特征向量问题：$Ly = \lambda y$
- LE的求特征向量问题对应于连续的时候求拉普拉斯特征函数问题

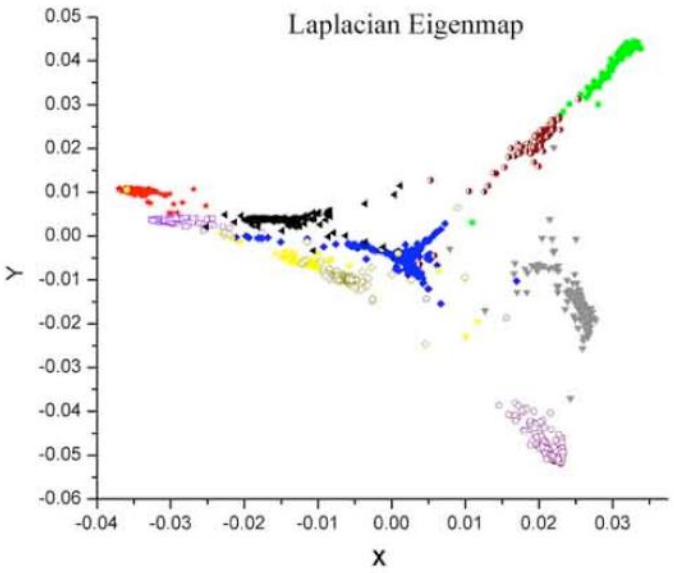


LE算法

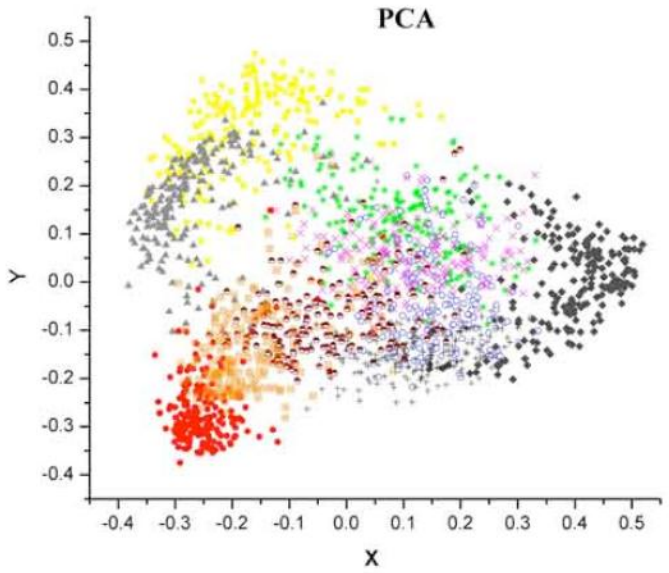
- 第一步：构建近邻图
- 第二步：计算每条边的权重(不相连的边权重为0)
  - 热核权重：$w_{ij} = \exp\left(\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$
  - 0-1 权重：$w_{ij} = 1$
- 第三步：求解特征向量方程，$Ly = \lambda y$，将点x_i映射到$(y_1(i), \ldots, y_d(i))$

LE: 数字上的聚类结果

- LE

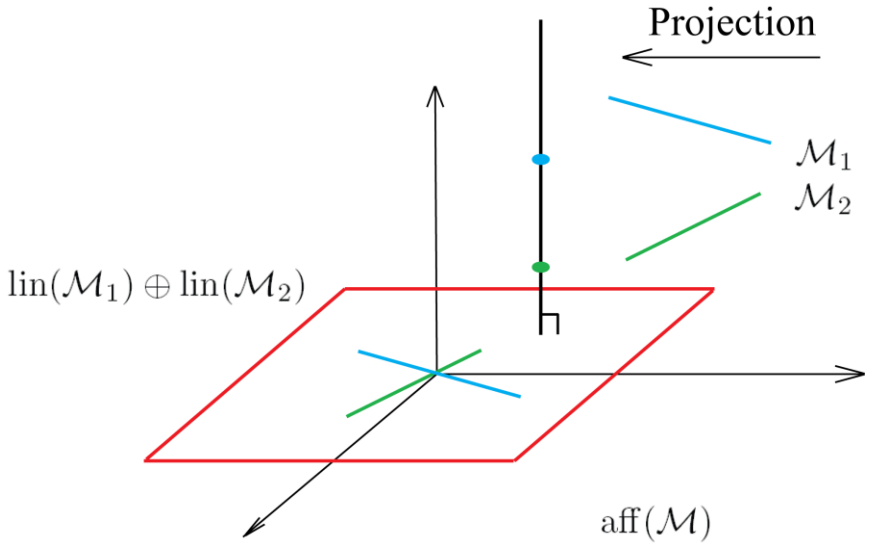


- PCA



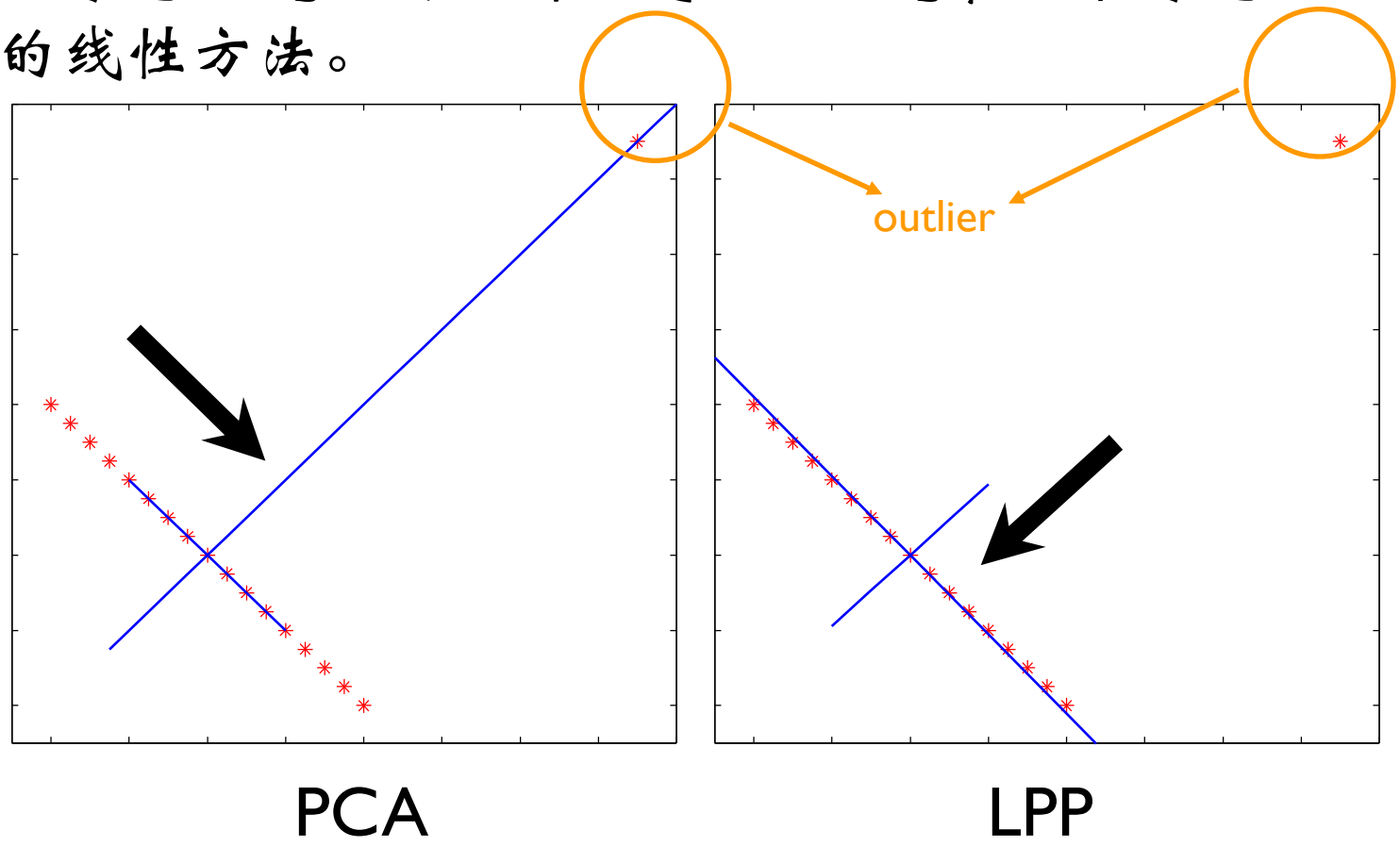
Locality Preserving Projections (LPP)

- 跟LE一样的准则，但是限制函数是外围欧氏空间的线性函数：$f(x) = \boldsymbol{a}^T x, x \in R^N$
- 最终转化成求解如下特征向量问题：$XLX^T\boldsymbol{a} = \lambda XDX^T\boldsymbol{a}$，其中$X = (x_1 \ \ldots x_n)$是数据矩阵。
- LPP可以将位于平行仿射凸包的流形分开(Binbin Lin, 2010)。



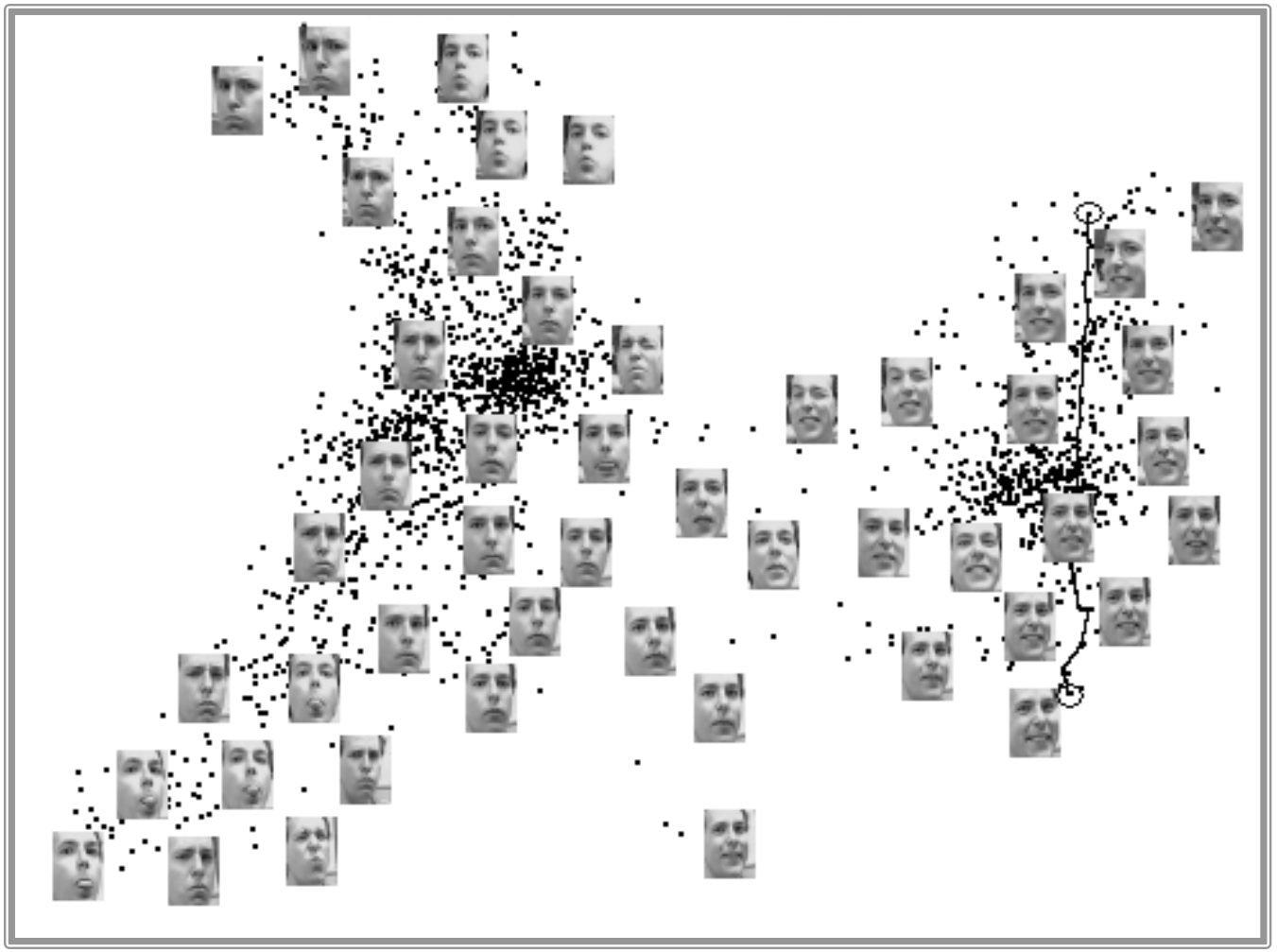
LPP vs. PCA

- PCA考虑的是全局统计信息，LPP是第一个考虑流形结构的线性方法。



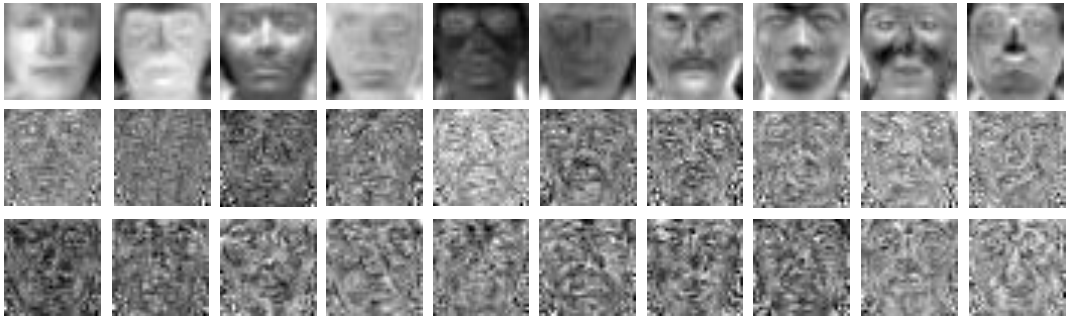
蓝色的线段表示投影的两个基向量。长的线段表示第一个基向量，短的线段表示第二个基向量。

人脸数据流形



人脸识别

- Eigenface
- Fisherface
- Laplacianface (LPP)



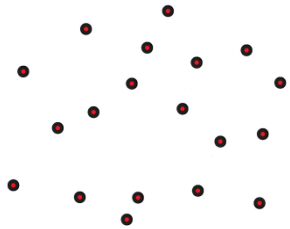
YALE人脸数据库中计算的前十个Eigenface(第一行), Fisherfaces (第二行) and Laplacianfaces (第三行).

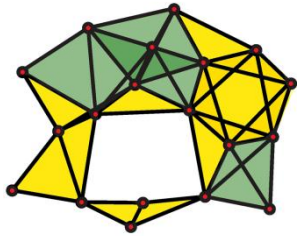
在三个数据库中的人脸识别率比较

	YALE	PIE	MSRA
Eigenfaces	74.7%	79.4%	64.6%
Fisherfaces	80%	94.3%	73.5%
Laplacianfaces	84%	95.4%	91.8%

流形学习的展望和挑战

- 研究数据的几何和拓扑，对于人们认识数据和处理数据具有本质意义。
- 挑战：现有方法对于数据的要求比较高，对噪音的情况处理能力不够



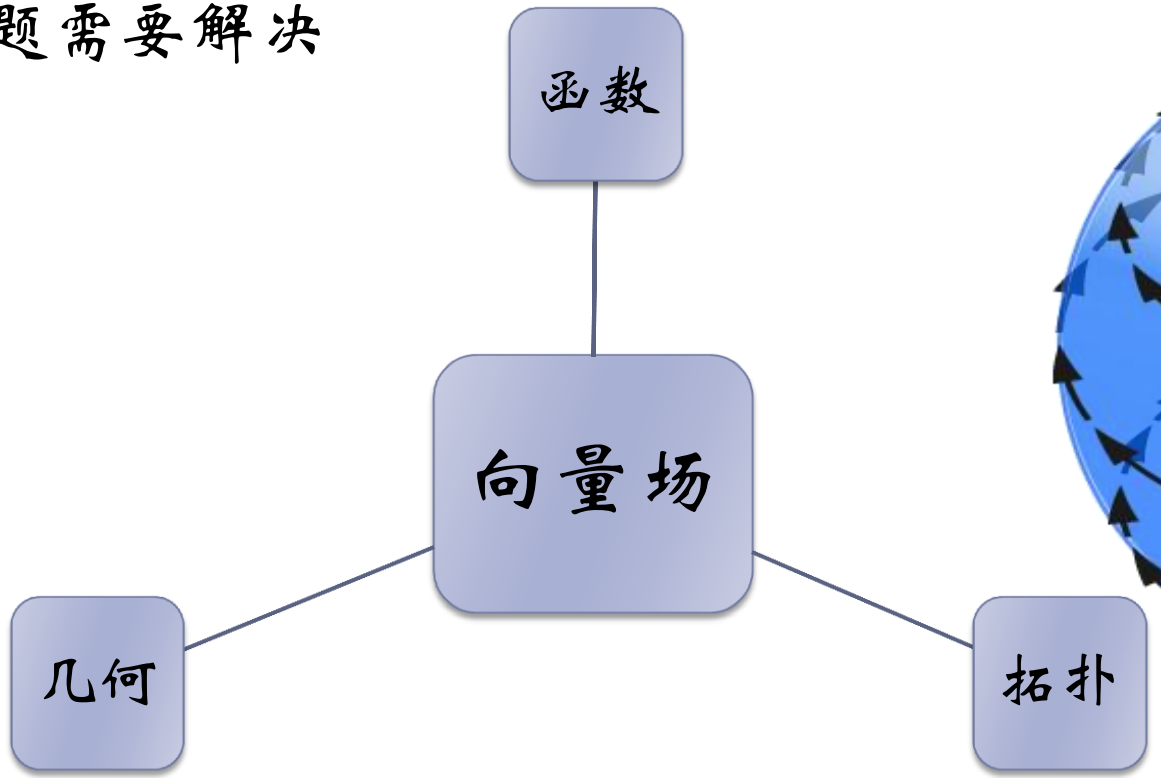


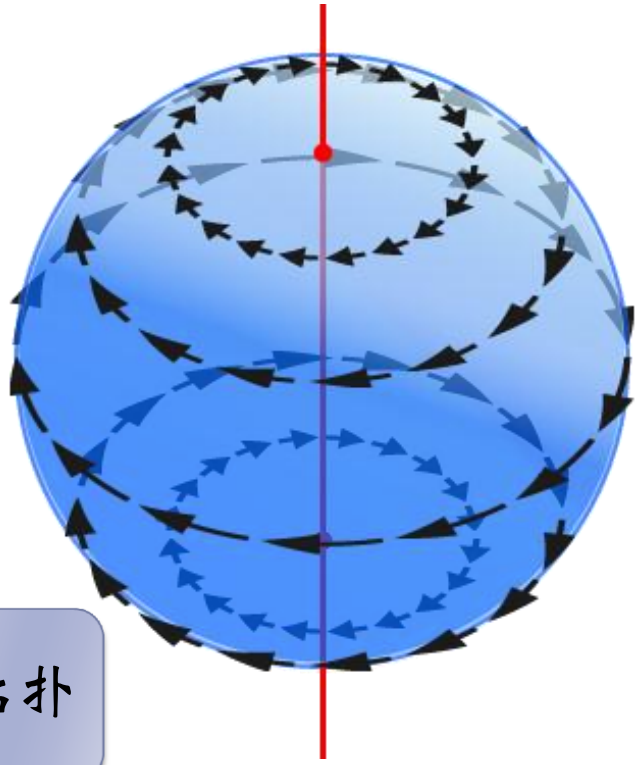
维度
连通分支
不同维度的洞

样本点 → 单纯复形 → 同调群

流形学习的展望和挑战

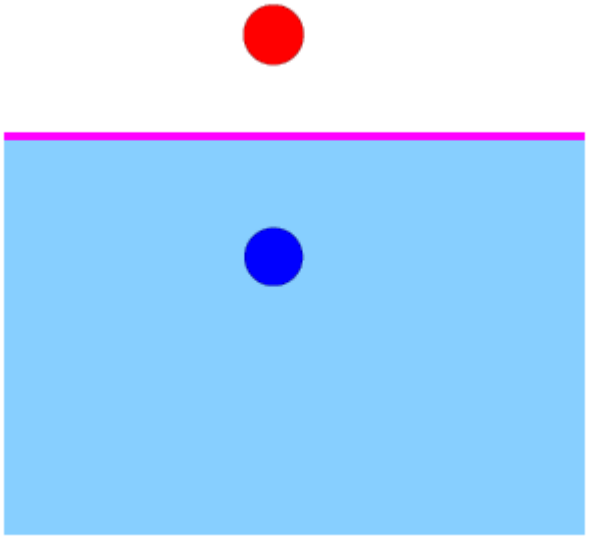
- 特殊的向量场反应流形的几何和拓扑，同时跟流形上的函数密切相关。
- 挑战：证明向量场方法的优势，有很多基础的理论问题需要解决

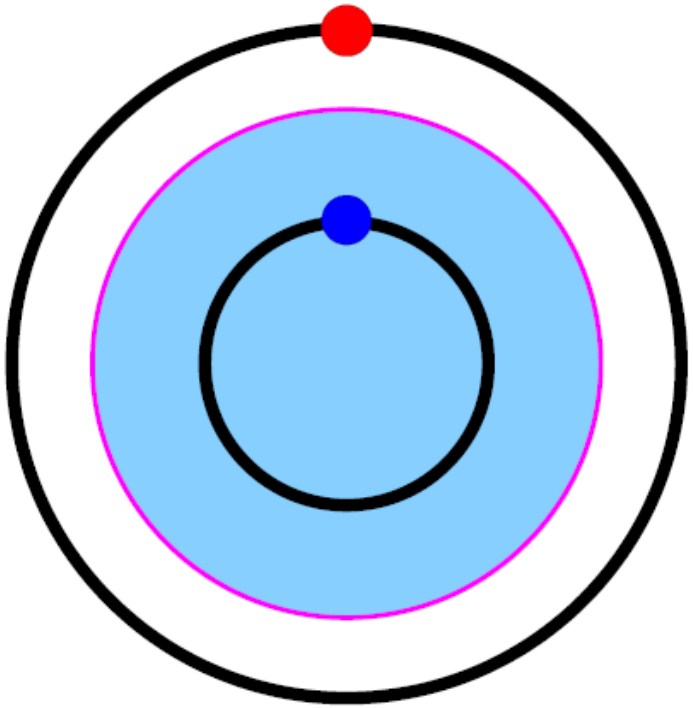




流形学习的展望和挑战

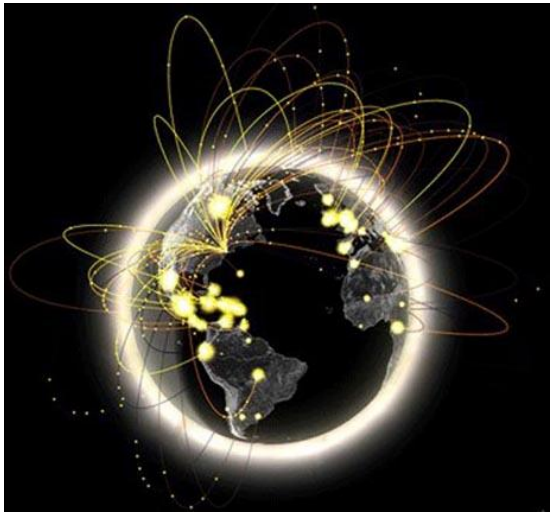
- 结合流形结构的学习问题
- 挑战：需要在流形上发展相应的统计概念以及学习理论





流形学习的展望和挑战

- 互联网时代，大规模数据的流形学习
- 挑战：流形学习算法往往是要求一个整体的矩阵分解，很难处理大规模数据



以YouTube为例：

1. 每日的视频播放量为1亿次；
2. 每日新增65000段视频；
3. 60%的视频是在线观看的；
4. 视频总量大小至少是45TB。

谢谢！