新媒体大数据分析

New Media Big Data Analysis

第三章 数据建模

黄振亚，朱孟潇，张凯
Email: huangzhy@ustc.edu.cn，mxzhu@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/NM2025.html

助教：齐畅，朱家骏
bigdata_2025@163.com

11/26/2025

无监督学习

- 数据挖掘任务 —— 无监督学习
  - 无标签数据是现实中最常见的数据
    - 例如，拍摄的照片等

该图片有关联

这张照片是哪里？

无监督学习

- **数据挖掘任务 —— 无监督学习**
  - 无标签数据可以提炼何种规律？
    - **关注数据之间的关联性，如共现关系、距离、相似度等**

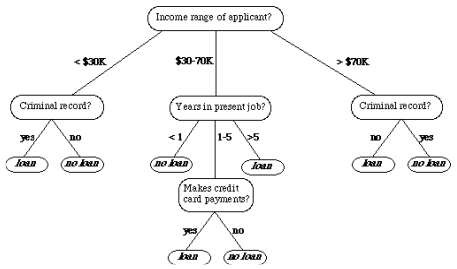
拍摄照片数据无标签

存在规律：科大校园

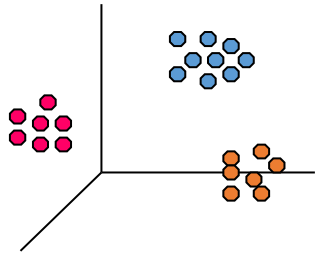
数据挖掘基础

- 数据挖掘——四个任务有哪些常用方法?

分类与预测



聚类



数据

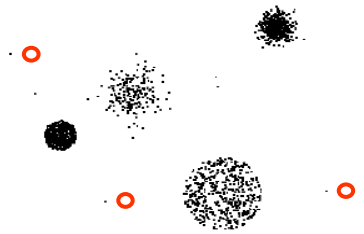
	T		H		P	
	L	H	L	H	L	H
J	-6.0	8.8	60	100	986	1044
F	-2.8	10.9	48	100	973	1025
M	-5.6	17.7	34	100	976	1037
A	-1.2	22.2	27	100	996	1036
M	-0.8	27.8	25	100	1003	1034
J	5.2	29.1	26	100	998	1030
J	9.8	30.6	23	99	997	1027
A	5.6	26.1	31	100	992	1029
S	5.2	24.8	35	100	998	1028
O	-0.4	21.3	42	100	990	1031
N	-7.6	17.3	55	100	963	1023
D	-10.4	9.2	53	100	987	1039

table 17a

2010 monthly weather variation, Cambridge (UK)

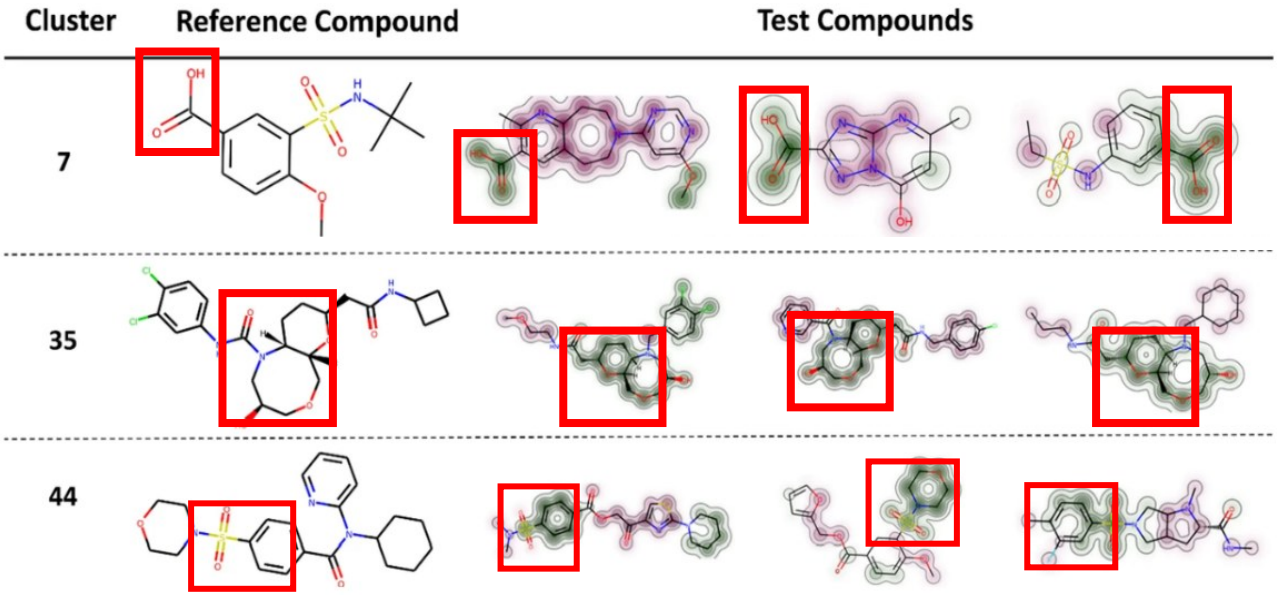
关联分析

异常检测



聚类分析: 应用实例

- 案例一：分子与药物分析
  - 输入：生物医药分子
  - **结构相似度更高**的分子被分配到一个聚簇



- 第7 簇中含有芳香族羧酸酯
- 第35簇中含有芳基卤化物
- 第44簇中含有磺胺

➢ Hadipour, H., Liu, C., Davis, R., Cardona, S.T., & Hu, P. (2022). Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. BMC Bioinformatics, 23.

聚类分析: 应用实例

- 案例二：传染病溯源
  - 输入：纽约市病例人群的信息：地理位置等
  - 对纽约市的冠状病毒病(COVID-19)**爆发场所**聚类， 定位的感染源

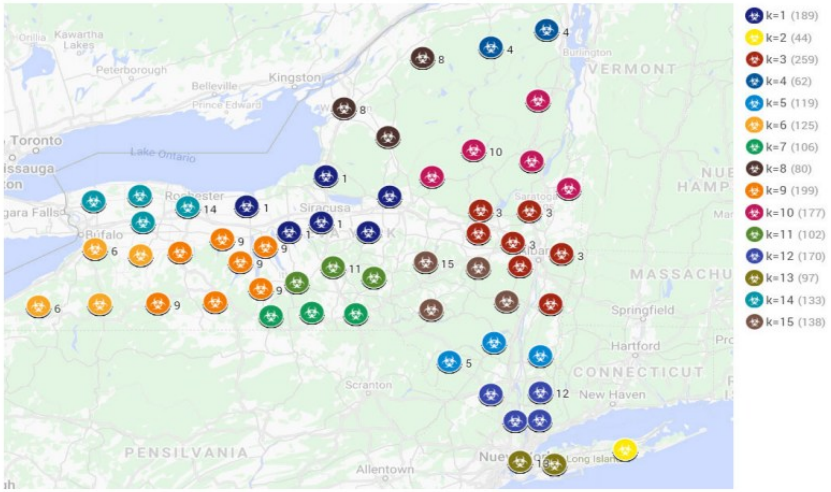


FIGURE 5. K-means clustering ($k = 15$) in New York state.

按病例的位置聚类，得到K个聚簇区域

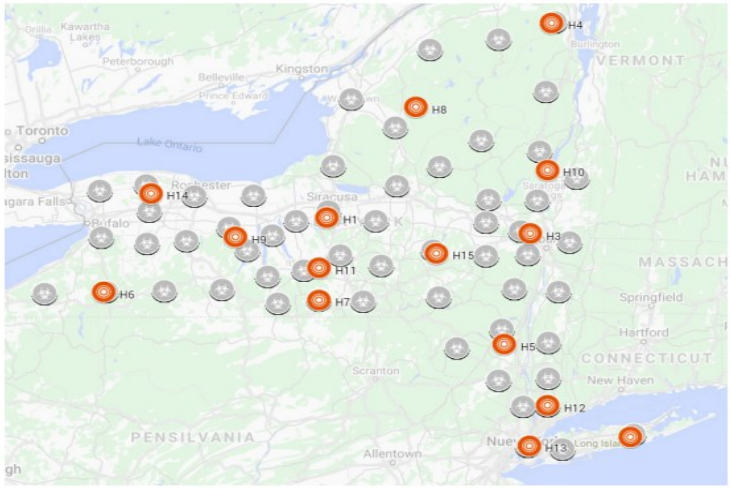


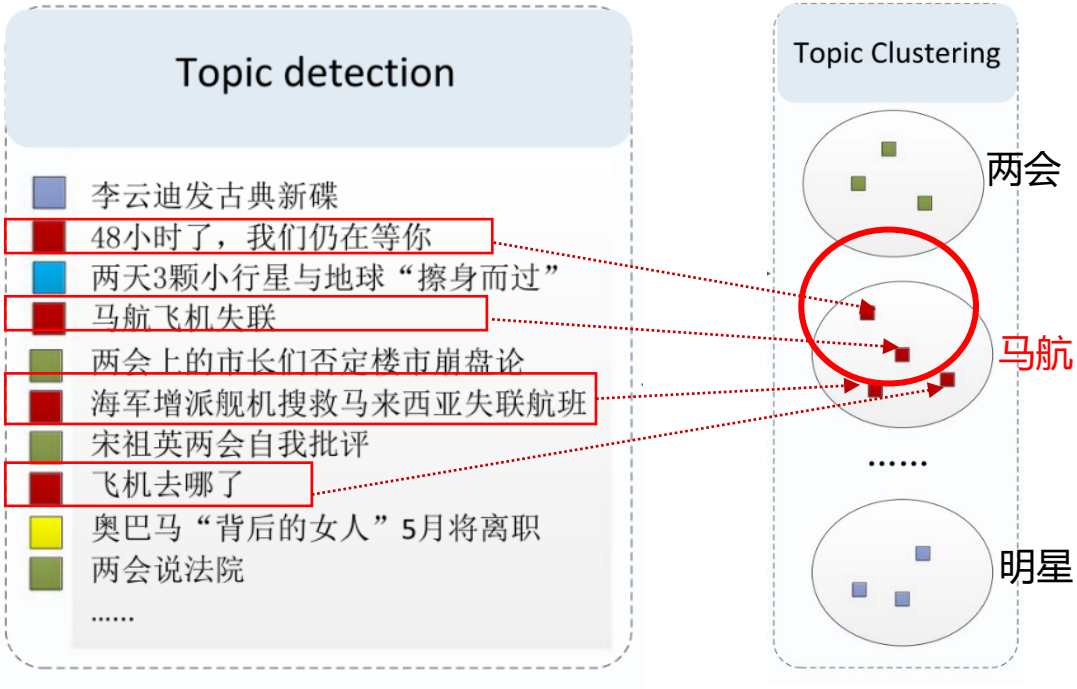
FIGURE 7. Hot spots H_k for each cluster in New York state (orange circles).

聚簇区域中心被视为感染源

- Guevara C, Peñas M S. Surveillance Routing of COVID-19 Infection Spread Using an Intelligent Infectious Diseases Algorithm[J]. Ieee Access, 2020, 8: 201925-201936.

聚类分析: 应用实例

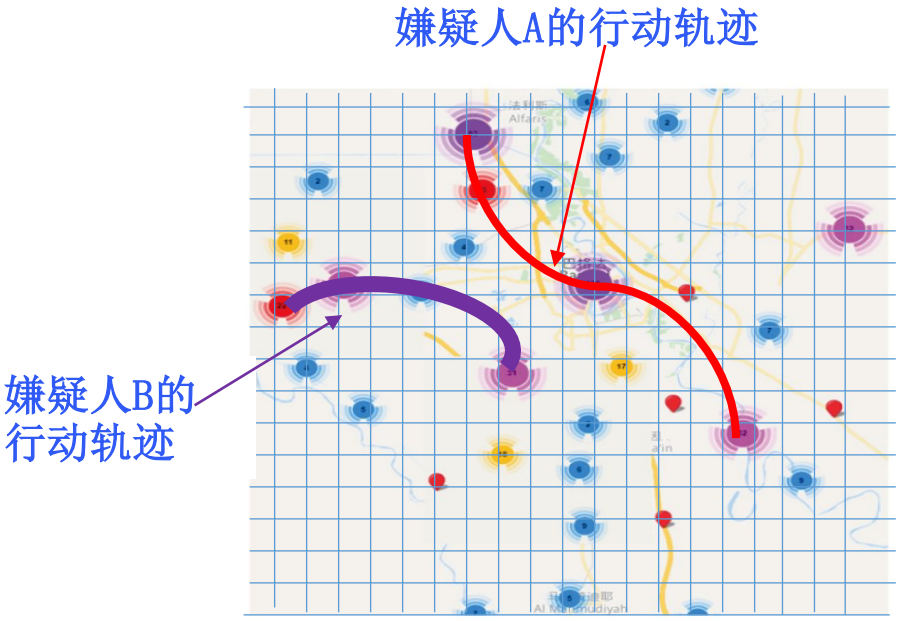
- 案例三：网络舆情分析-话题挖掘
  - 输入：社交媒体中的评论与话题
  - 话题聚类，同一话题簇中出现的**关键词相似**



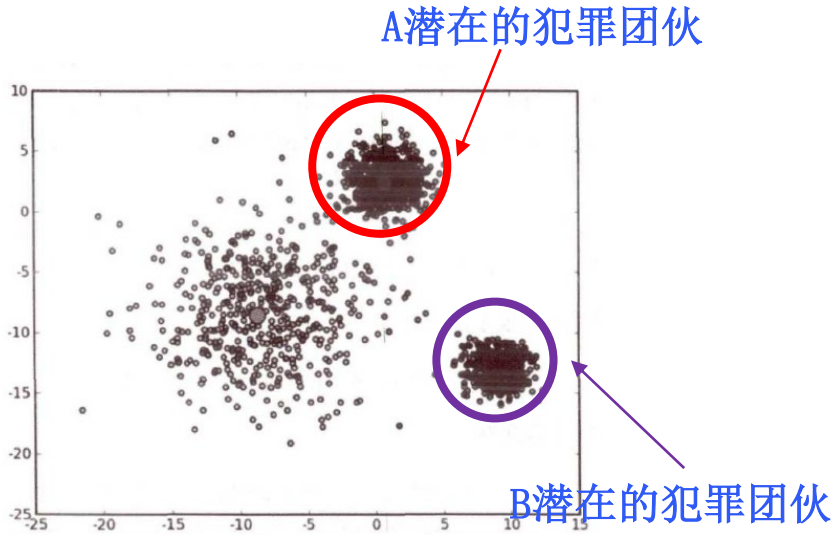
➢ Cai Y, Wu X, Xie X, et al. A topic mining method for multi-source network public opinion based on improved hierarchical clustering[C]//2019 IEEE DSC. IEEE, 2019: 439-444.

聚类分析: 应用实例

- 案例四：安防与维稳-犯罪团伙识别
  - 输入：人员轨迹时空数据：如网吧、酒店、车站等,
  - 对**嫌疑人的轨迹信息**进行聚类，找出犯罪团伙。



地理空间网格化



轨迹信息聚类结果

聚类分析: 应用实例

- 案例五：教育问题的聚类
  - 输入：数学应用题
  - 题目聚类，同类题目的解答模板一样

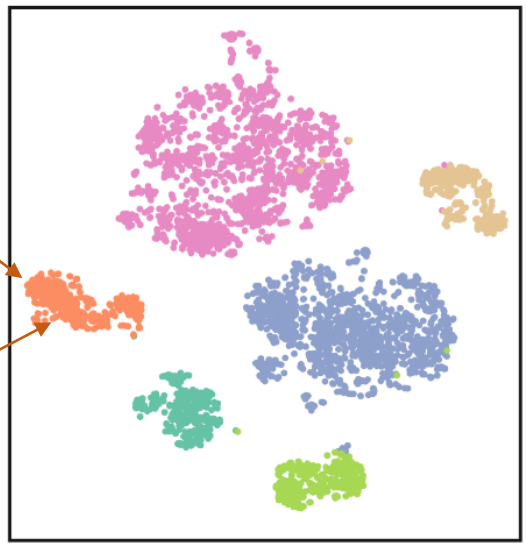
数学应用题

Prob. A: Norma has 88 cards. She loses 70. How many cards will Norma have ?

Eq: 88 - 70

Prob. B: Joyce starts with 75 apples. She gives 52 to Larry. How many apples does Joyce end with?

Eq: 75 - 52



基础运算的模式不同

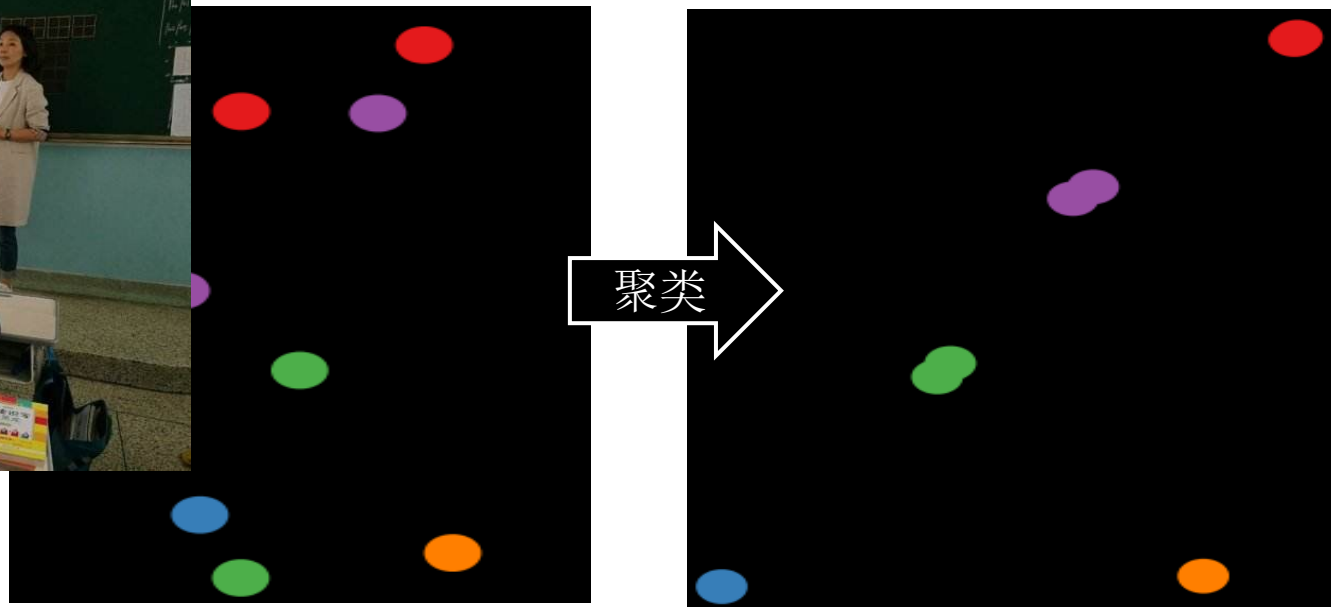
- $n_1 + n_2$
- n_1 / n_2
- $n_1 - n_2$
- $(n_1 + n_2) * n_3$
- $n_1 * n_2$
- $(n_1 + n_2) / n_3$

- Li, Z., Zhang, W., Yan, C., Zhou, Q., Li, C., Liu, H., & Cao, Y. (2022). Seeking Patterns, Not just Memorizing Procedures: Contrastive Learning for Solving Math Word Problems. ArXiv, abs/2110.08464.
- Huang, Z., Lin, X., Wang, H., Liu, Q., Chen, E., Ma, J., Su, Y., & Tong, W. (2021). DisenQNet: Disentangled Representation Learning for Educational Questions. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining.

聚类分析: 应用实例

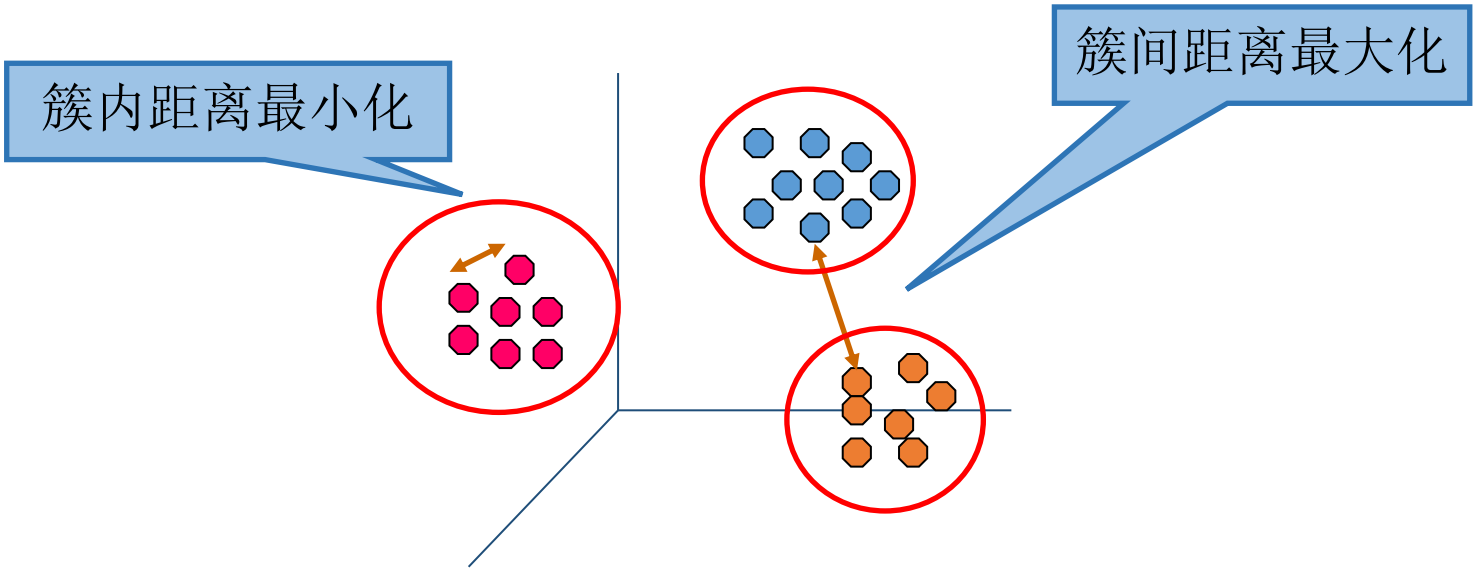
- 案例六：学业数据分析——优化教师教学
  - 输入：试验学校的学生考试数据
  - 聚类发现教师教学模式的规律

根据考试数据对班级进行简单聚类，根据聚类结果，发现70%的类里，两个班级是同位授课教师



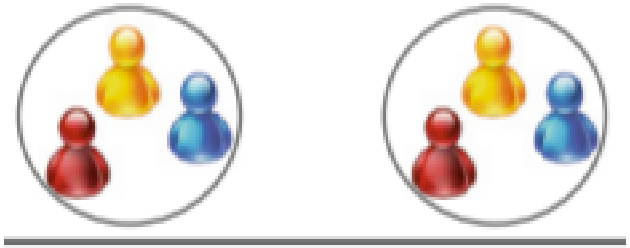
聚类分析

- 数据挖掘任务 —— 聚类(Clustering)
  - 目标：对数据进行**“群体性”分析**，将样本分为若干个**簇**（Clusters）
    - 其中，每个簇都由相似的样本所组成
  - 簇的特点：簇内相似（距离近），簇间相异（距离远）



聚类分析

- 聚类分析要解决三个问题
  - **1. 如何定义簇？：即，思考我们的目标（但具有主观性）**
    - “群体性”的依据：不同的“群体性”立场，可以得到不同的簇
    - 例如，学生分组应该考虑 **技能互补？** 还是 **能力相近**？
  - **2. 如何定义相关性？** 即，度量数据之间的相似性
    - 相似性度量往往存在一定局限性，未必反映聚类的真实意图
    - 例如，常用向量表征数据(人的爱好)， 但是否绝对相似？

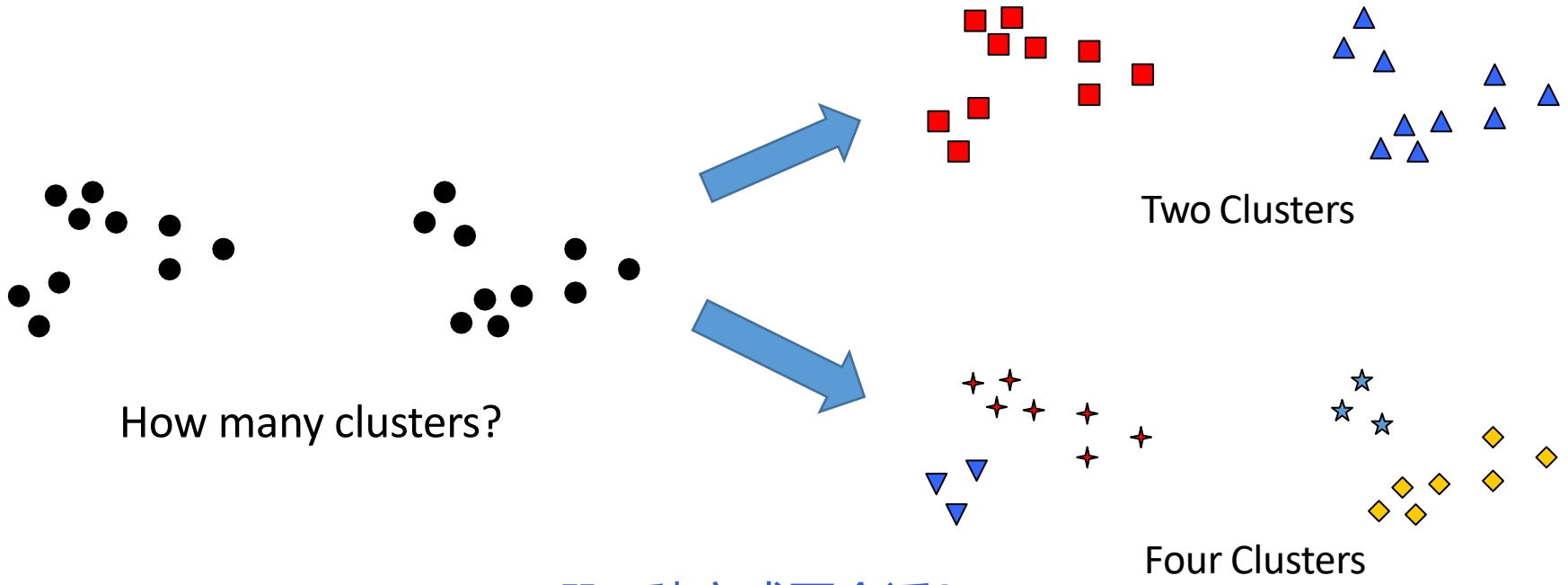


技能互补？ 能力相近？

图片本身相似，但代表的类别完全不同？

聚类分析

- 聚类分析要解决三个问题
  - **3. 如何决定簇的数量?** 即，选择合适数量的簇
    - 数据没有天然标签，簇的数量往往是个开放性问题
    - 避免过大或过小的簇，会导致失去代表性，但这未必可通过簇数调节



哪一种方式更合适?