人工智能
无监督学习
+Cluster Analysis

中山大学 计算机学院

中山大學
SUN YAT-SEN UNIVERSITY

Unsupervised Learning

- Why?
- How?

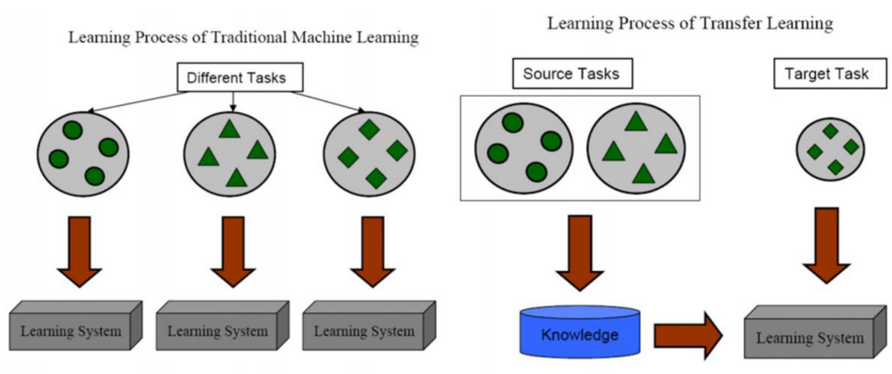
Pre-trained Models

- **For DNN, one of challenges： data hungry**
  - overfit
  - poor generalization ability

- **One solution: construct high-quality datasets**
  - expensive
  - time-consuming

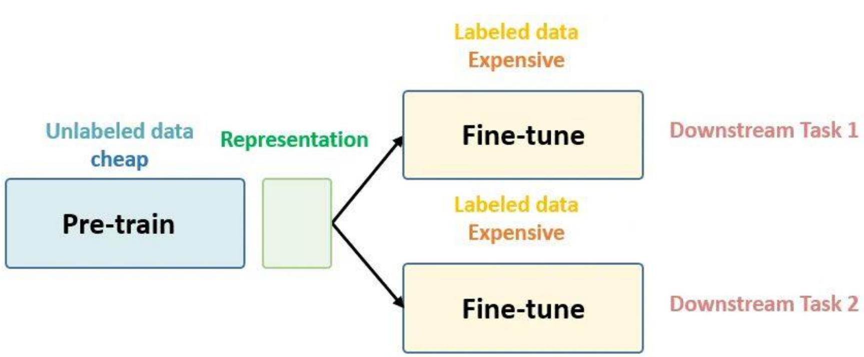
Key: train effective deep neural models with limited human-annotated data

- **An Intuitive idea : Transfer learning**



(a) Traditional Machine Learning (b) Transfer Learning

- **Two stage :**
  - **1. Pre-training on source tasks**
  - **2. Fine-tuning on target tasks**



- **The history of pre-trained models**

1. **Vision models(based on CNNs)**

   AlexNet, VGG, ResNet

2. **Language models(based on Transformer)**

   GPT, BERT, UNILM

3. **Vision models(based on Vision Transformer)**

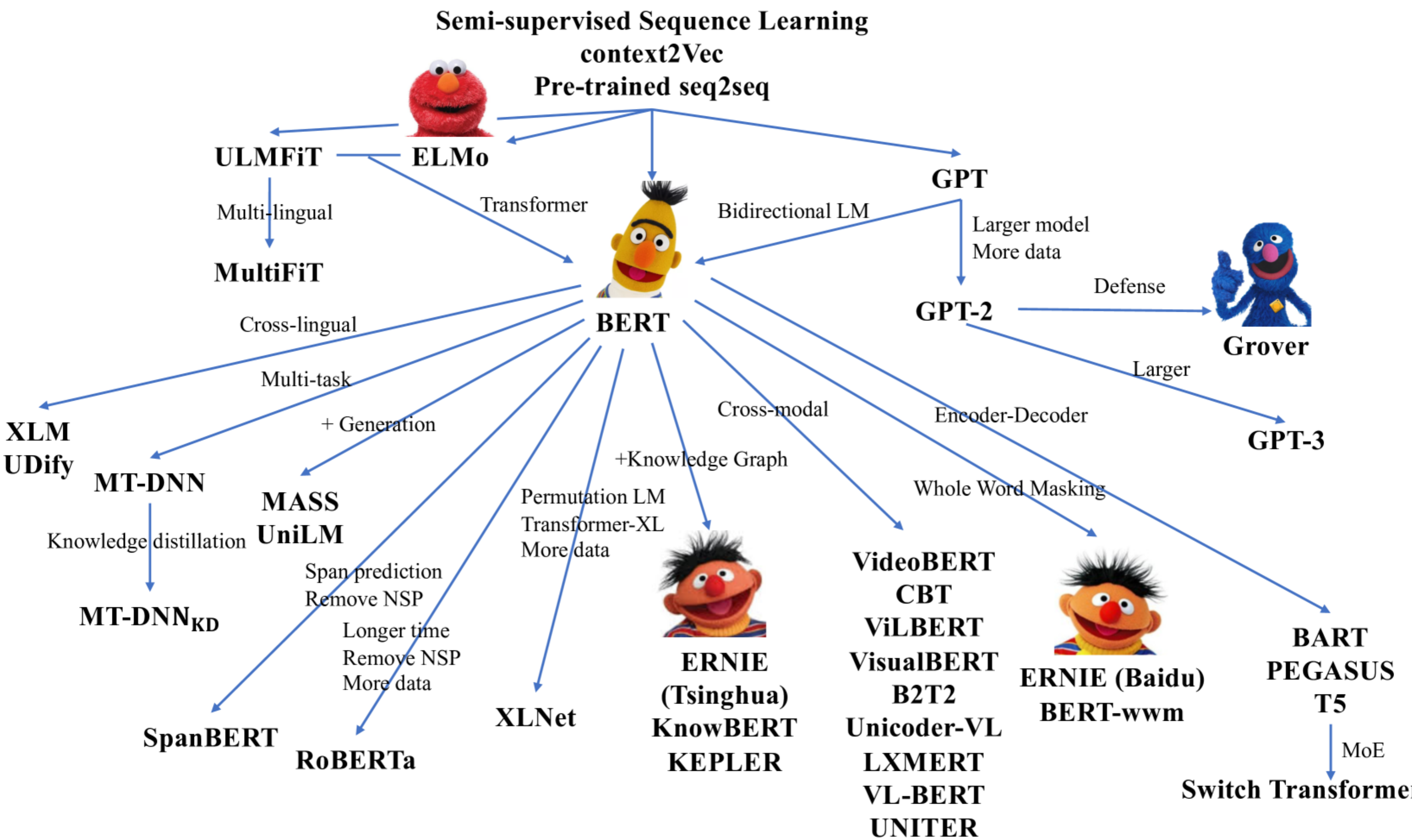
   MAE, MoCo, BEiT, simMIM

4. **Multimodal models**

   CLIP, BLIP

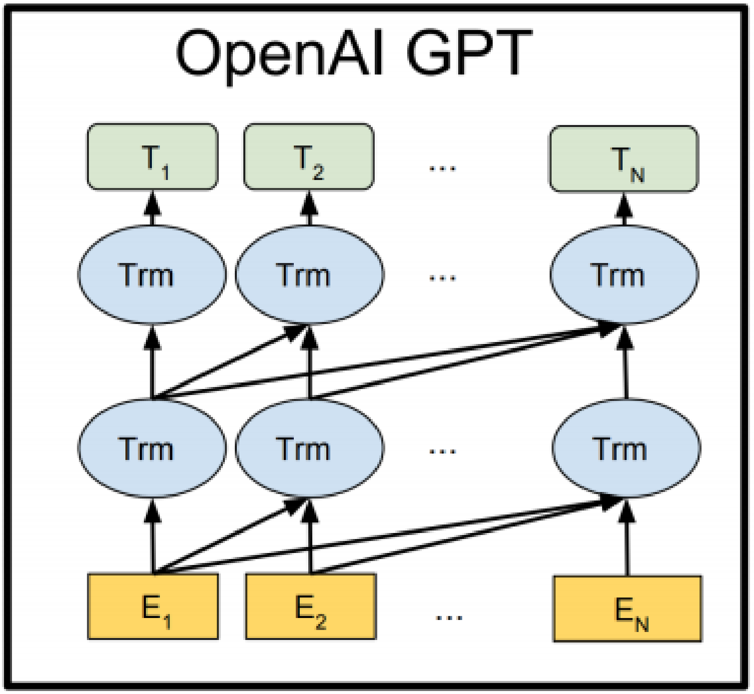
Pre-trained Models

- Overview: Pre-trained models



GPT

- First SOTA pre-trained language model
- From **OpenAI**
- Later: GPT-2, GPT-3, ……



GPT

- **Unsupervised pre-training**
  - Unsupervised pre-training, maximizing the log-likelihood

$$L_1(\mathcal{U}) = \sum_i \log P(u_i \mid u_{i-k}, \ldots, u_{i-1}; \Theta)$$

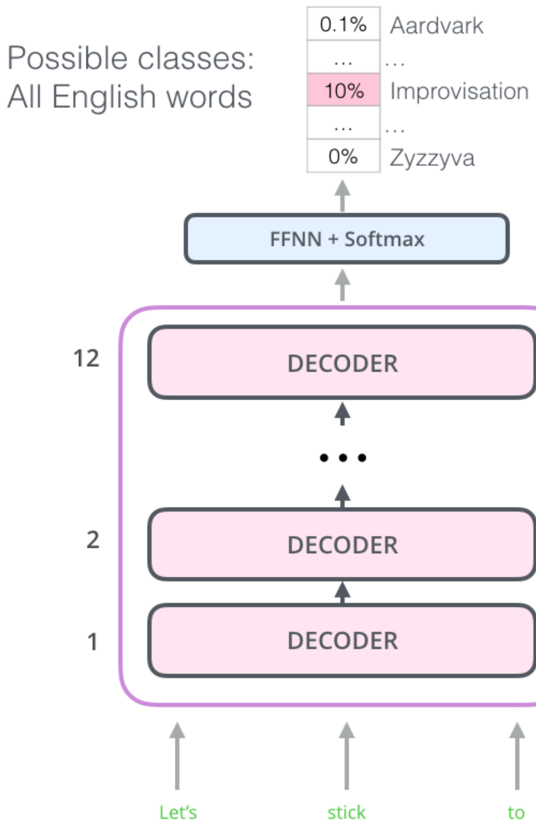
  - where $\mathcal{U} = \{u_1, \ldots, u_n\}$ is an unsupervised corpus of tokens, k is the size of context window, P is modelled as a neural network with parameters Θ.

$$h_0 = UW_e + W_p$$
$$h_l = \text{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \text{softmax}(h_n W_e^T)$$

  - where U is one-hot representation of tokens in the window, n is the total number of transformer layers, transformer_block() denotes the decoder of the Transformer model (*multi-headed self-attention and position-wise feedfoward layers*).

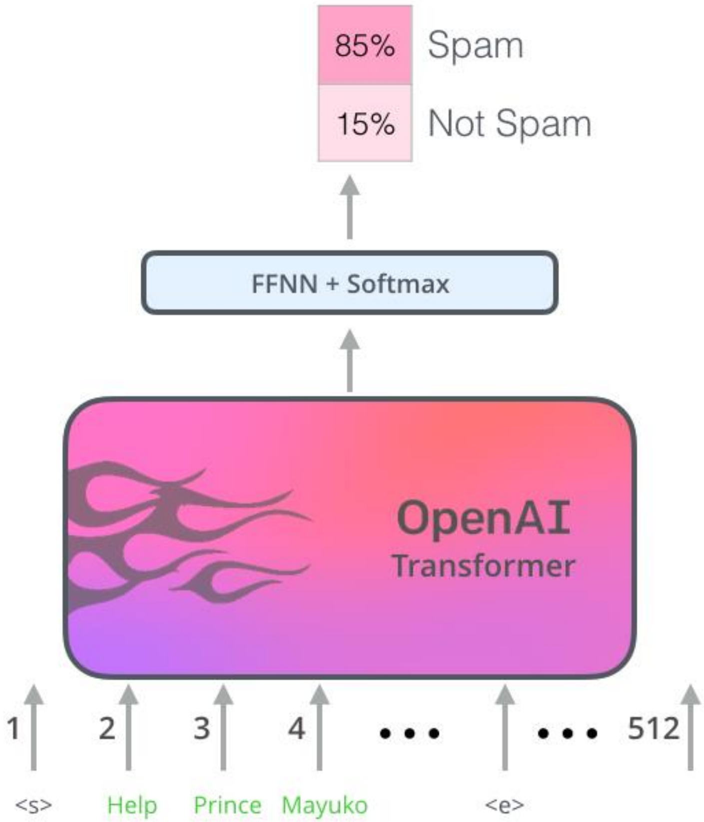
GPT

- **Pre-training dataset:**
  - 7000+ books



GPT

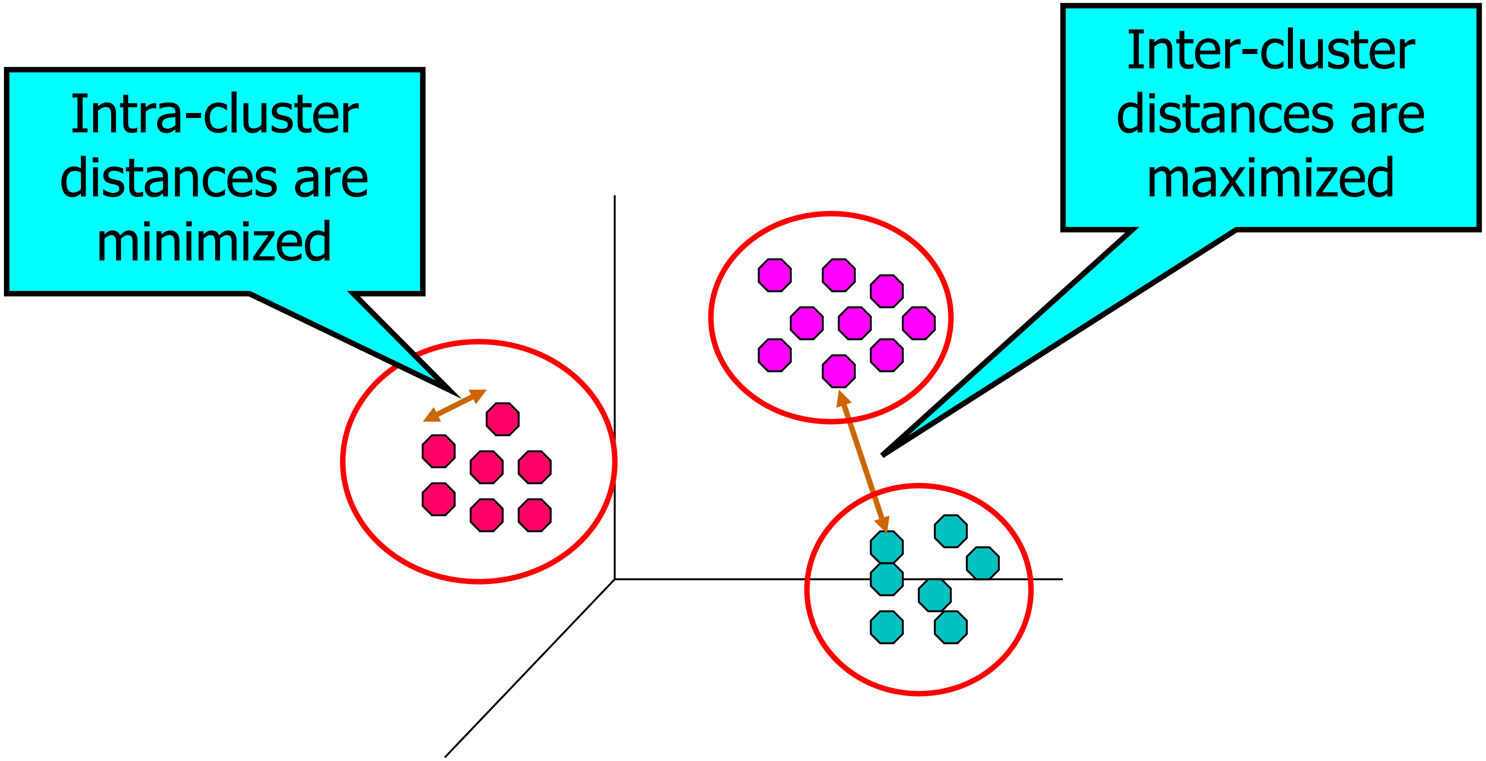
- **An example:** Transfer Learning to Downstream Tasks



Cluster Analysis

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



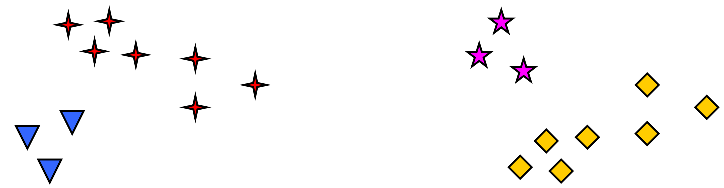
Clustering in Social Network

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters



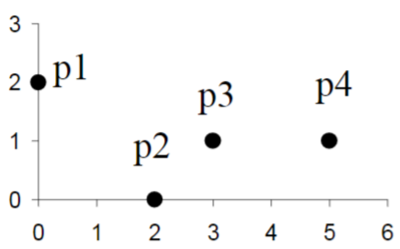
Four Clusters

Notion of a Cluster can be Ambiguous

What is a natural grouping among these objects?

Types of Clusters: Objective Function …

- Map the clustering problem to a different domain and solve a related problem in that domain
  - Proximity（亲近度） matrix defines a weighted graph, where the nodes are the data points, and the weighted edges represent the proximities (similarity or dissimilarity) between data points



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0.000	2.828	3.162	5.099
p2	2.828	0.000	1.414	3.162
p3	3.162	1.414	0.000	2.000
p4	5.099	3.162	2.000	0.000

Figure Four points and their corresponding data and proximity (distance) matrices.

  - Similarity? Dissimilarity?

Euclidean Distance / Manhattan Distance / Chebyshev Distance/

Cosine Similarity /…..

Similarity=1-Dissimilarity

Bonus: Some Distance Metrics

- $L_p(x_1, x_2) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}}$
- $L_1(x_1, x_2) = \sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|$ Manhattan Distance
- $L_2(x_1, x_2) = \sum_{l=1}^{n} \sqrt{\left(x_i^{(l)} - x_j^{(l)}\right)^2}$ Euclidean Distance
- $L_\infty(x_1, x_2) = \sqrt[\infty]{\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^\infty} = max(|x_i - x_j|)$

  Chebyshev Distance
- $L_{-\infty}(x_1, x_2) = \sqrt[-\infty]{\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^{-\infty}} = min(|x_i - x_j|)$
- Cosine Distance: $D(x_1, x_2) = 1 - \cos(x_1, x_2) = 1 - \frac{x_1 \cdot x_2}{||x_1||_2 ||x_2||_2}$

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

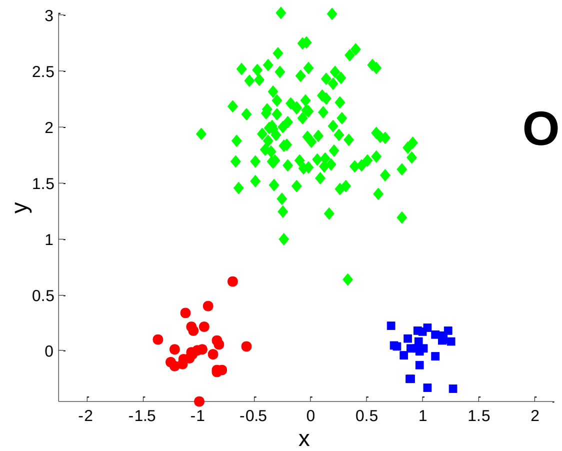
---

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

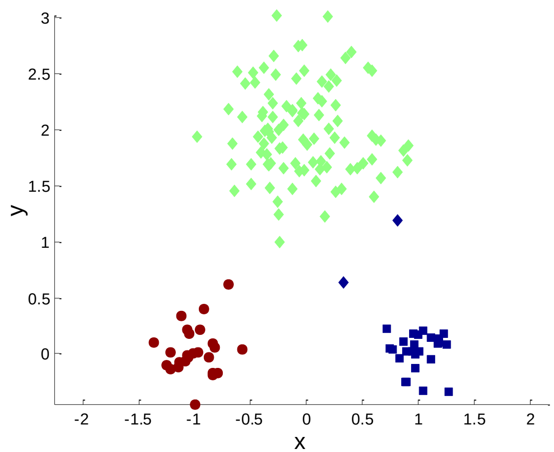
K-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters'
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

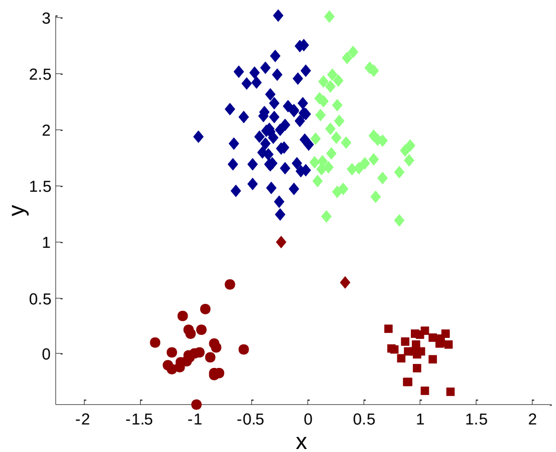
Two different K-means Clusterings



Original Points

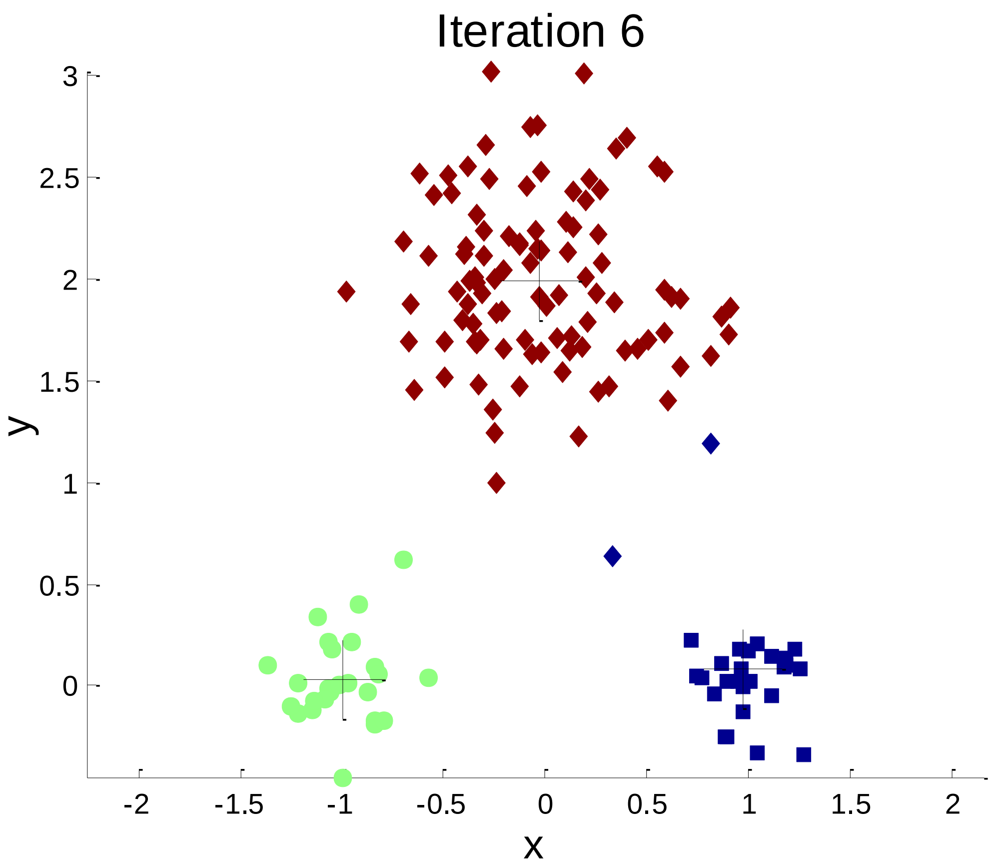


Optimal Clustering

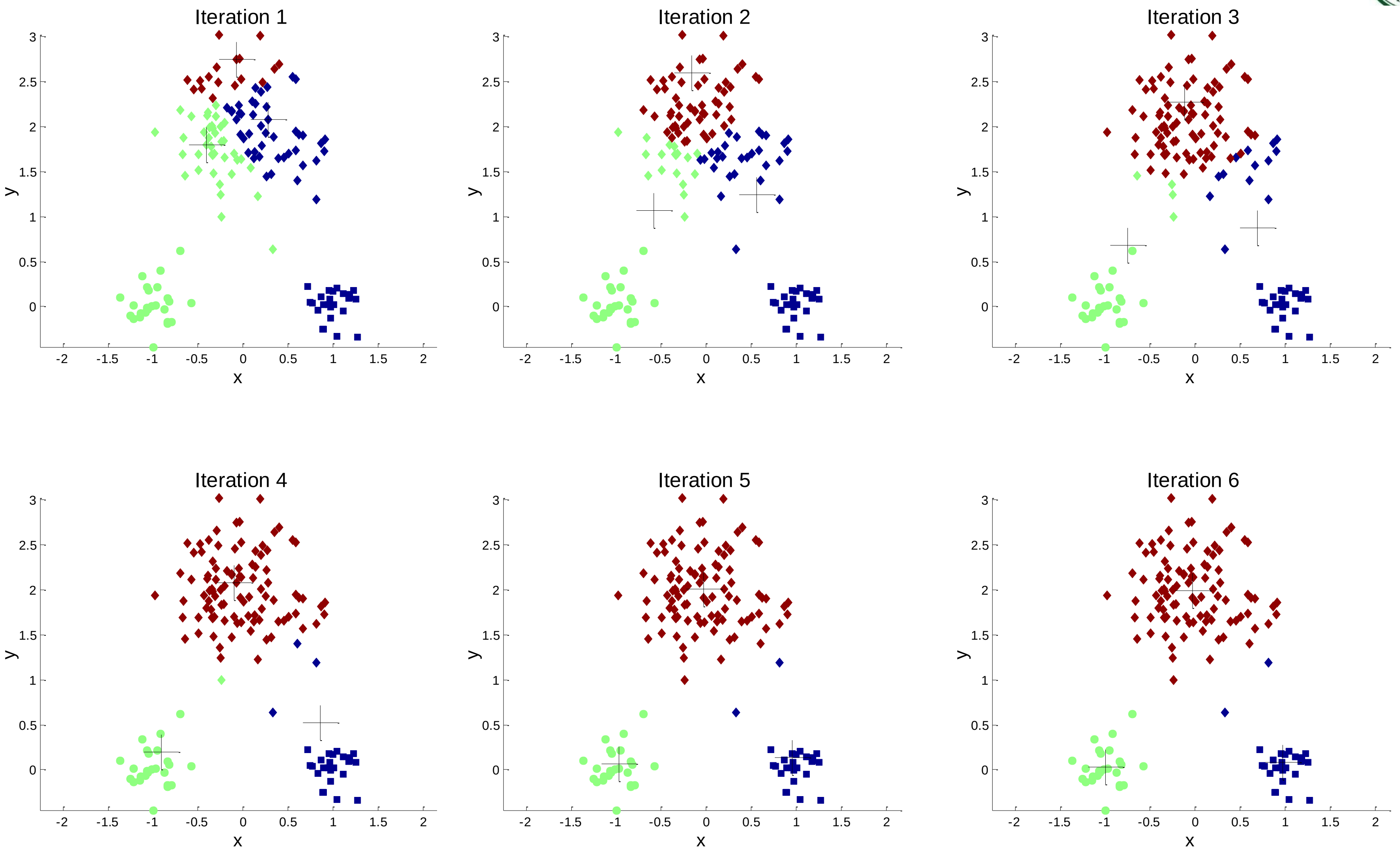


Sub-optimal Clustering

Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids



Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, \mathbf{x})$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Limitations

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

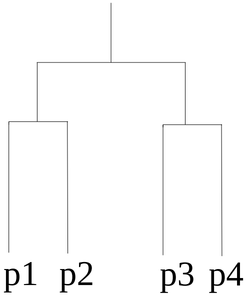
- K-means also has problems when the data contains outliers.

Discussion

Q：**假如**灭霸一个响指就能灭掉一半生命，如何挑选消灭对象?

Method 1：Hierarchical Clustering

将全宇宙的生命按照基因相似度构造Dendrogram，确定聚类数目后，在每两个相邻的聚类簇中随机挑选一个进行消灭，以保证宇宙生物多样性。



Discussion

Method 2：K-means Clustering

宇宙包含“九大国度”，为防止某个国度灭绝，在每个国度选取一个聚类中心，通过K-means得到聚类簇，再在每个聚类中随机选取一半生物进行消灭，宇宙生物多样性得到进一步保证。