Clustering

Types

Applications

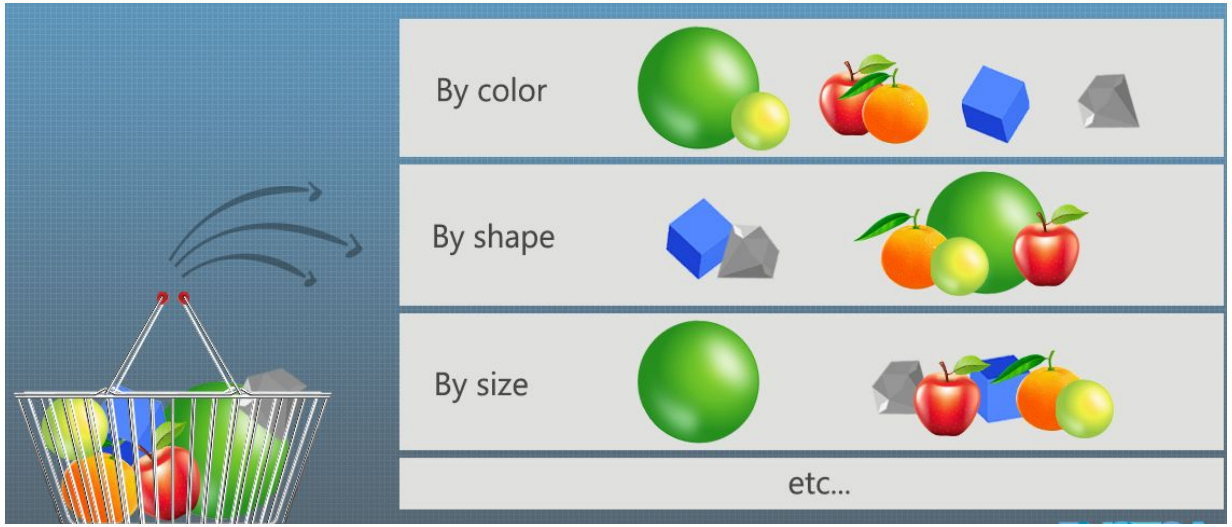
Similarity measures

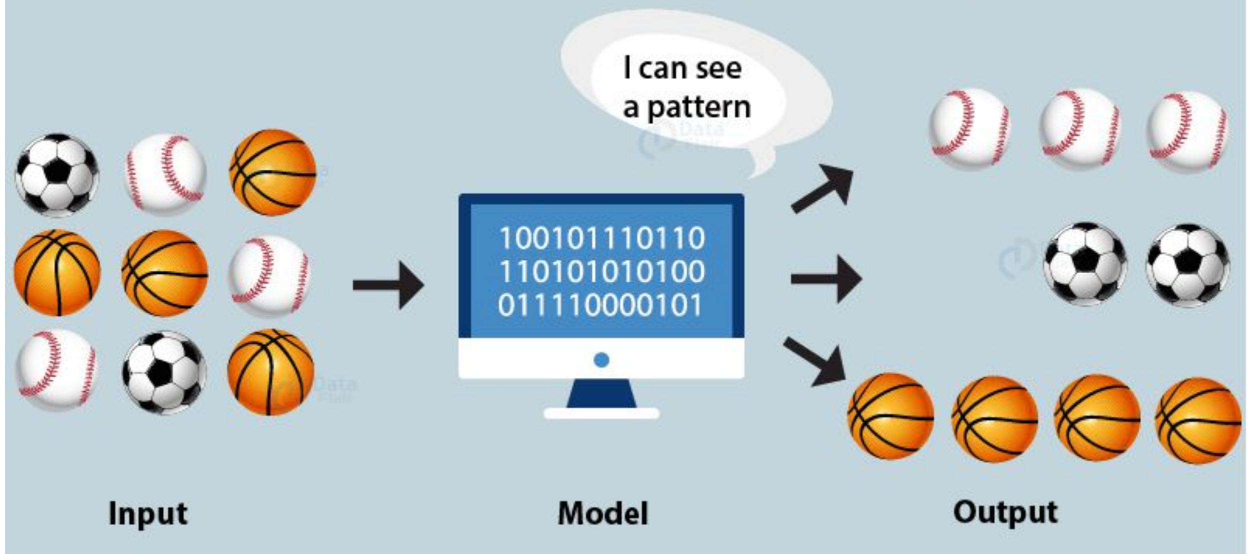
Clustering

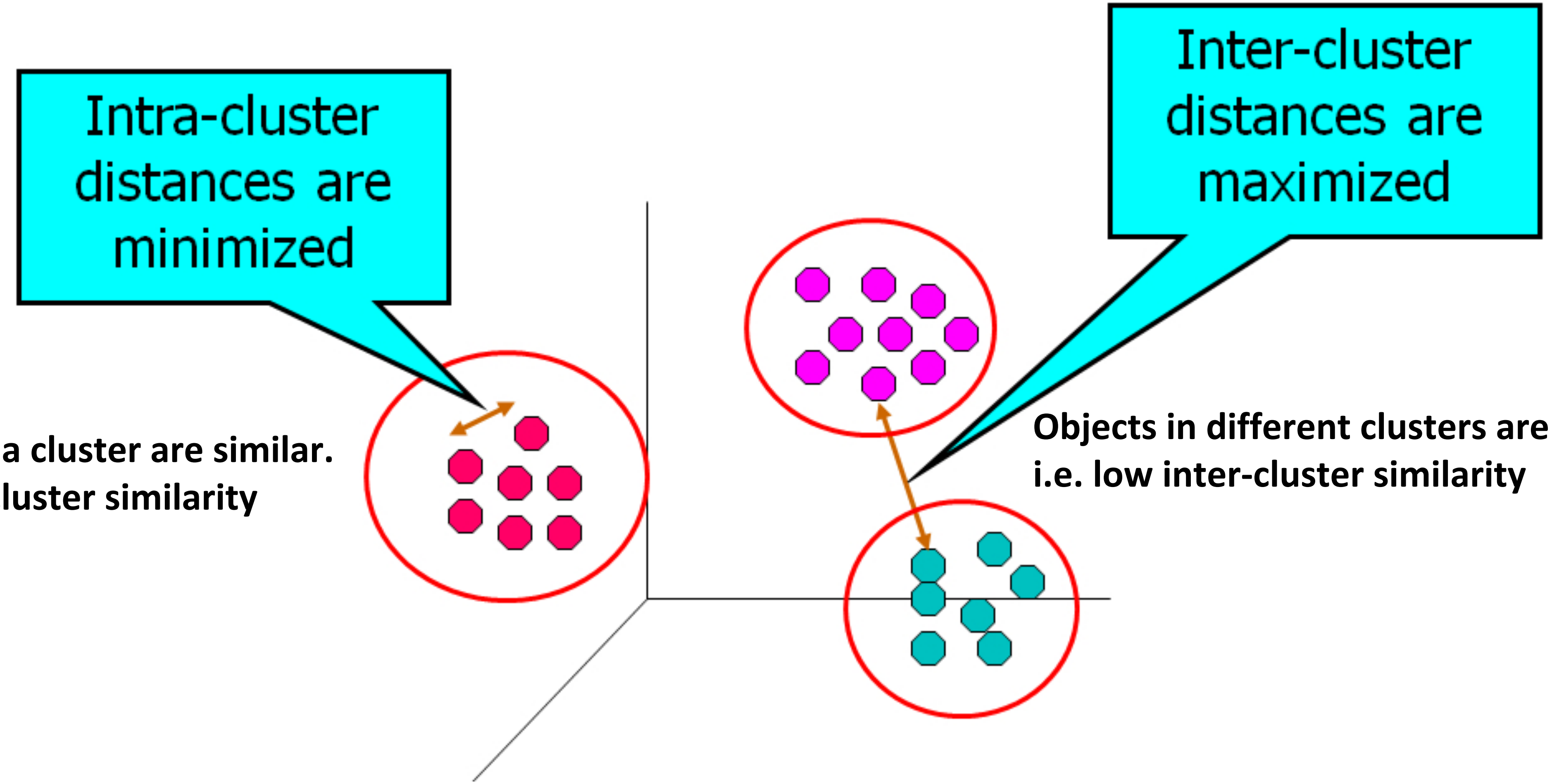
The machine learns through unlabelled data and discover outputs on its own

feature, characteristic, pattern

- Clustering is a method of unsupervised learning.
- It is a process of grouping similar objects together.
- The objects in one group should be similar to each other than to those in other groups.
- It finds a similar structure in a collection of unlabeled data.







Intra-cluster distances are minimized

Inter-cluster distances are maximized

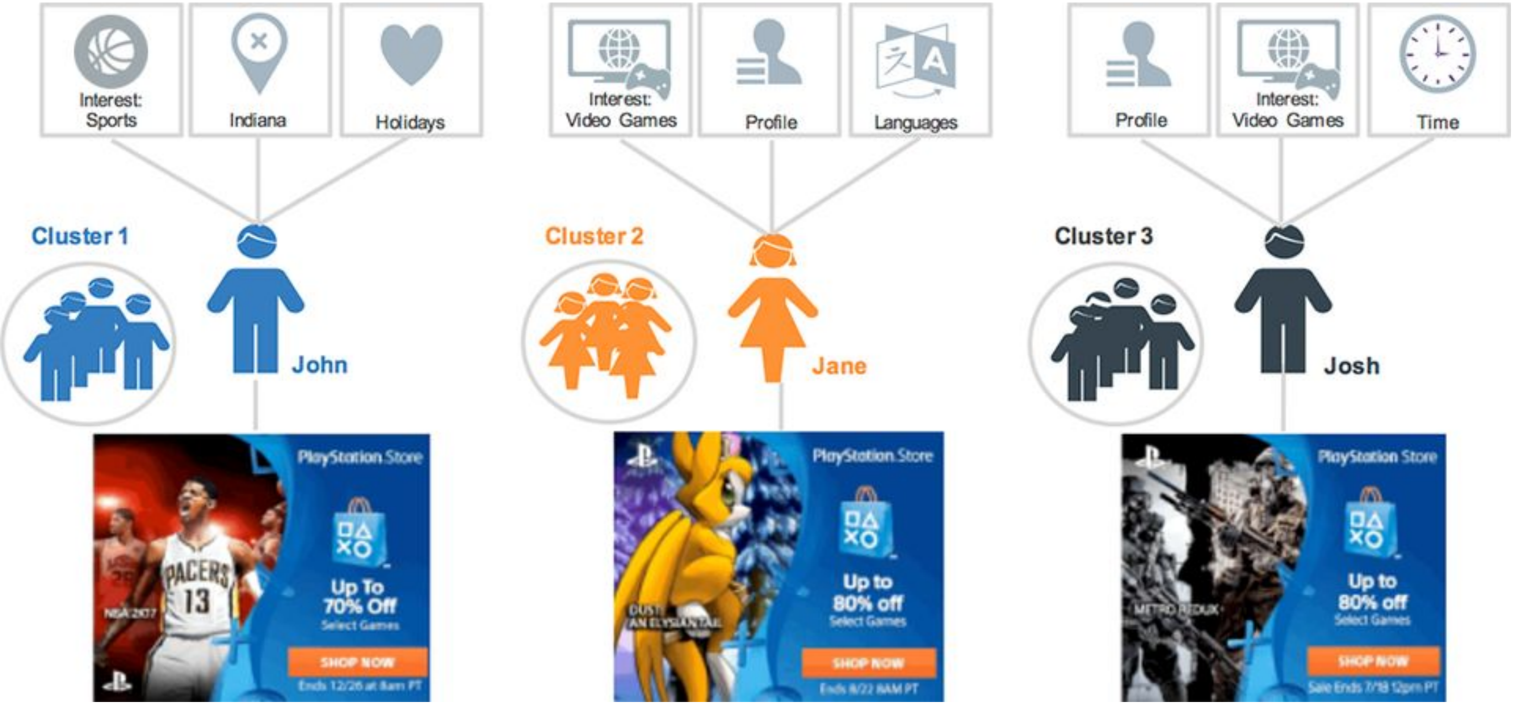
Objects within a cluster are similar.
i.e. high intra-cluster similarity

Objects in different clusters are different.
i.e. low inter-cluster similarity

Problem statement:

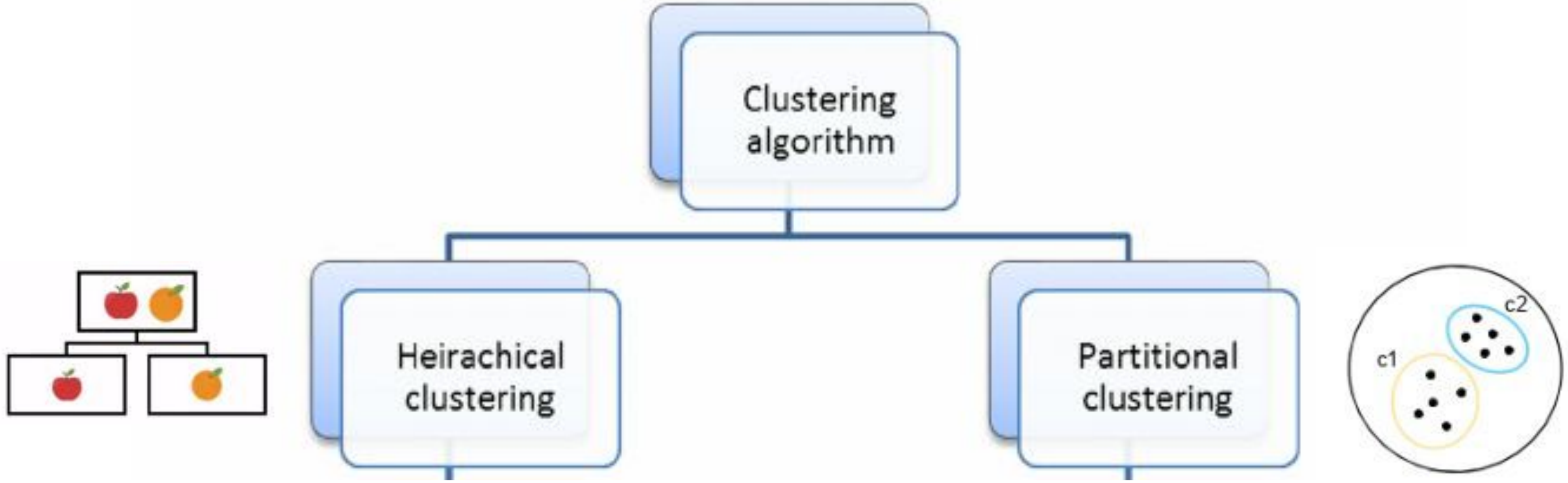
You are managing a PlayStation store and wish to understand preferences of your users and improve your business.

Is it possible for you to look at details of each user and devise a unique business strategy for each one of them?



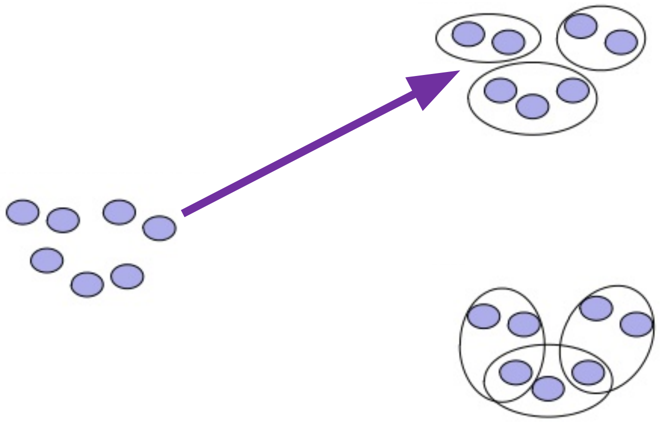
Solution:

Cluster all of your users into say 3 groups based on their usage habits and use a separate strategy for users in each of these groups.



- The data is organized into hierarchical structures
- The data can be either grouped in the bottom-up direction, or split in a top-down manner
- ***Eg. Agglomerative, Divisive clustering***

- Starts with a random partition of data and refine it iteratively
- These algorithms are called “flat” clustering
- ***Eg. K-means, Fuzzy c-means, Spectral clustering, Graph-based clustering***



Hard clustering:

Every object belongs to exactly one cluster

Eg. K-means

Soft clustering:

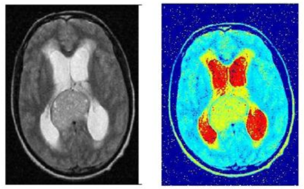
An object may belongs to more than one cluster

Eg. Fuzzy c-means

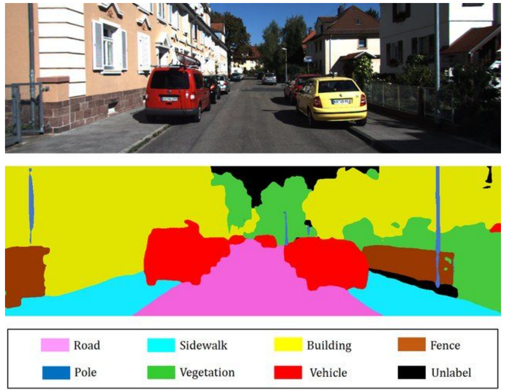
Applications of Clustering

- Recommendation engines
  - e-commerce websites recommending similar products and movie recommender sites
- Market segmentation
  - characterize their customer groups based on the purchasing patterns
- Social network analysis
  - recognize communities within large groups of people
  - Example: LinkedIn, Twitter, Instagram, Facebook, Youtube etc.

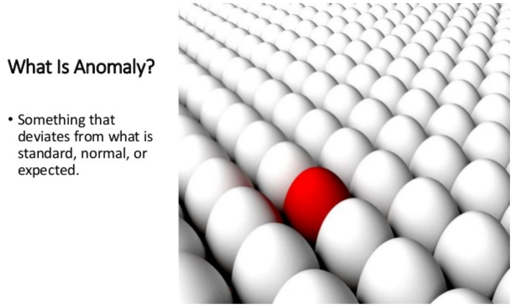
- Search result grouping
  - search engine creates a more relevant set of search results using ranking algorithms
- Medical imaging
  - locating tumors



- Image segmentation
  - partition an image into a number of segments



- Anomaly detection
  - identification of rare events. i.e detecting unknown network intrusions, fraud detection



What Is Anomaly?

- Something that deviates from what is standard, normal, or expected.

Similarity and distance measures

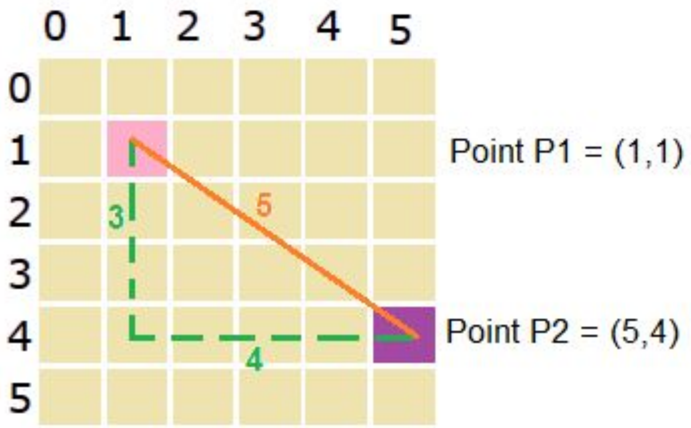
- Grouping requires some methods for computing the distance or the similarity between each pair of objects
- The classical methods for distance measures are Euclidean and Manhattan distances.

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

When q=1, Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

When q=2, Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

When q=inf, $$D_{\text{Chebyshev}}(x, y) := \max_i(|x_i - y_i|).$$



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2}$ = 5

Manhattan distance = |5-1| + |4-1| = 7