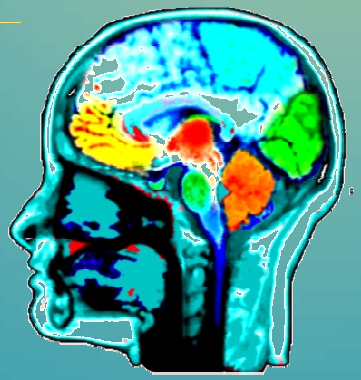
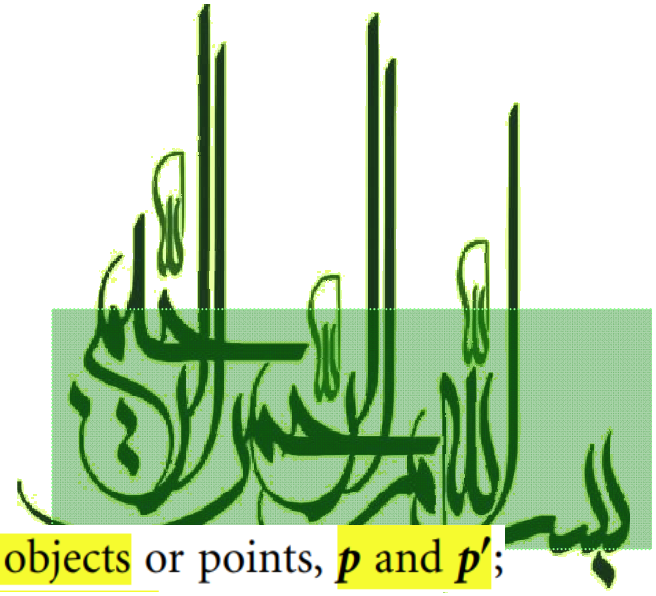
Introduction To Data Mining

Isfahan University of Technology (IUT)
Esfand1401



Clustering

Dr. Hamidreza Hakim
hamid.hakim.u@gmail.com



$\boldsymbol{m_i}$ is the mean for cluster, C_i; distance between two objects or points, $\boldsymbol{p}$ and $\boldsymbol{p'}$;

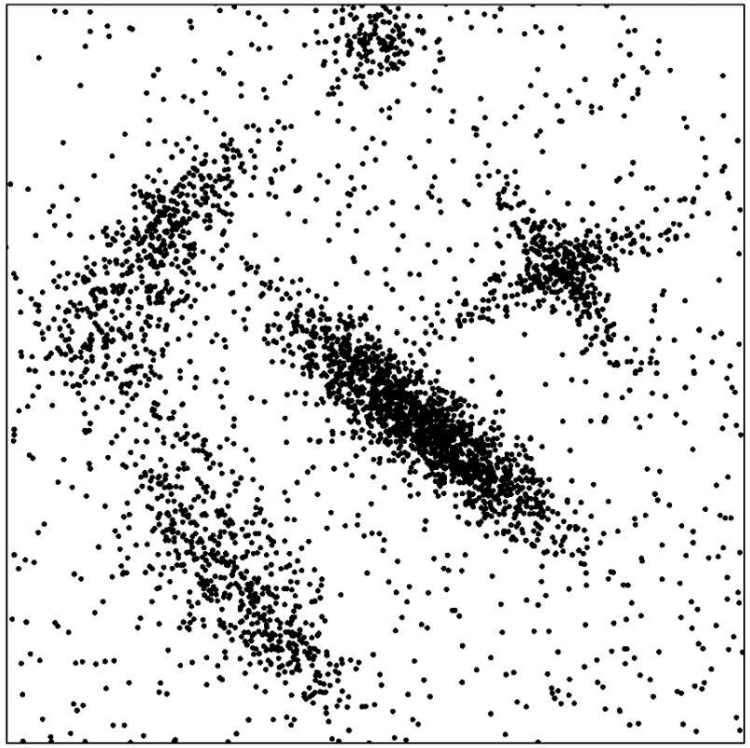
Minimum distance: $dist_{min}(C_i, C_j) = \min_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \{|\boldsymbol{p} - \boldsymbol{p'}|\}$

Maximum distance: $dist_{max}(C_i, C_j) = \max_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \{|\boldsymbol{p} - \boldsymbol{p'}|\}$

Mean distance: $dist_{mean}(C_i, C_j) = |\boldsymbol{m_i} - \boldsymbol{m_j}|$

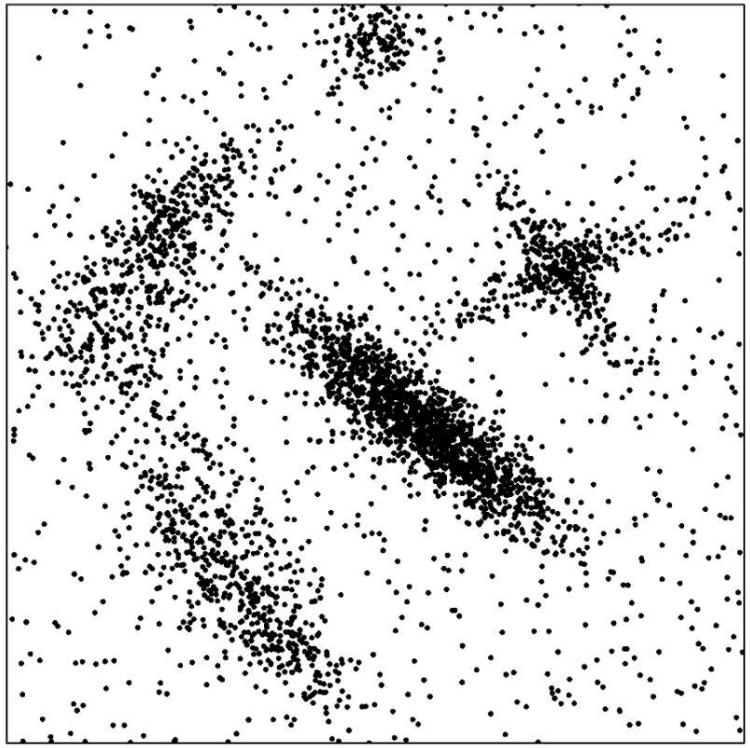
Average distance: $dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} |\boldsymbol{p} - \boldsymbol{p'}|$

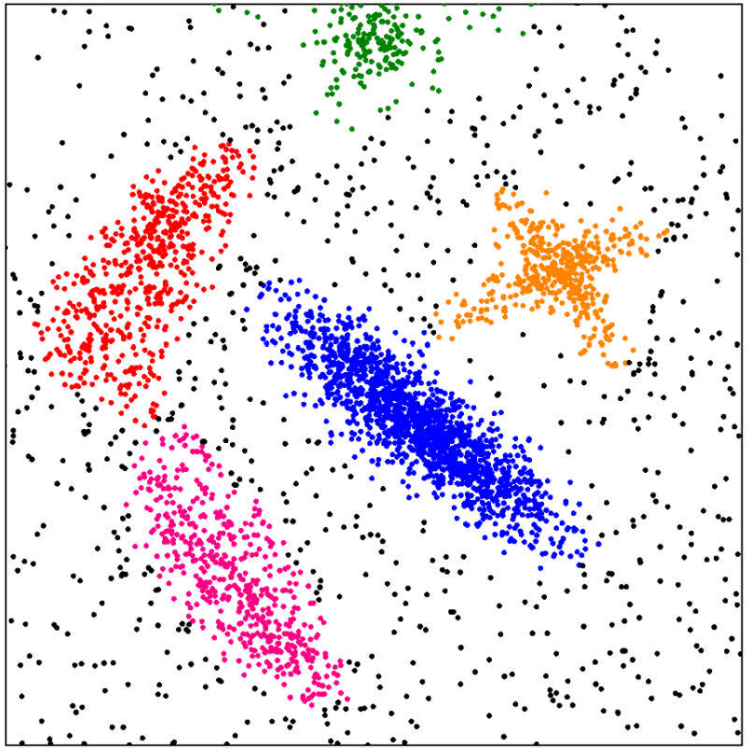
What is Cluster Analysis?



میشه بگی هرکدوم از این رکوردها
توی چه دسته هایی قرار دارند؟
چندتا دسته داریم کلا؟
براساس توده ها دسته بندی میکنیم

What is Cluster Analysis?



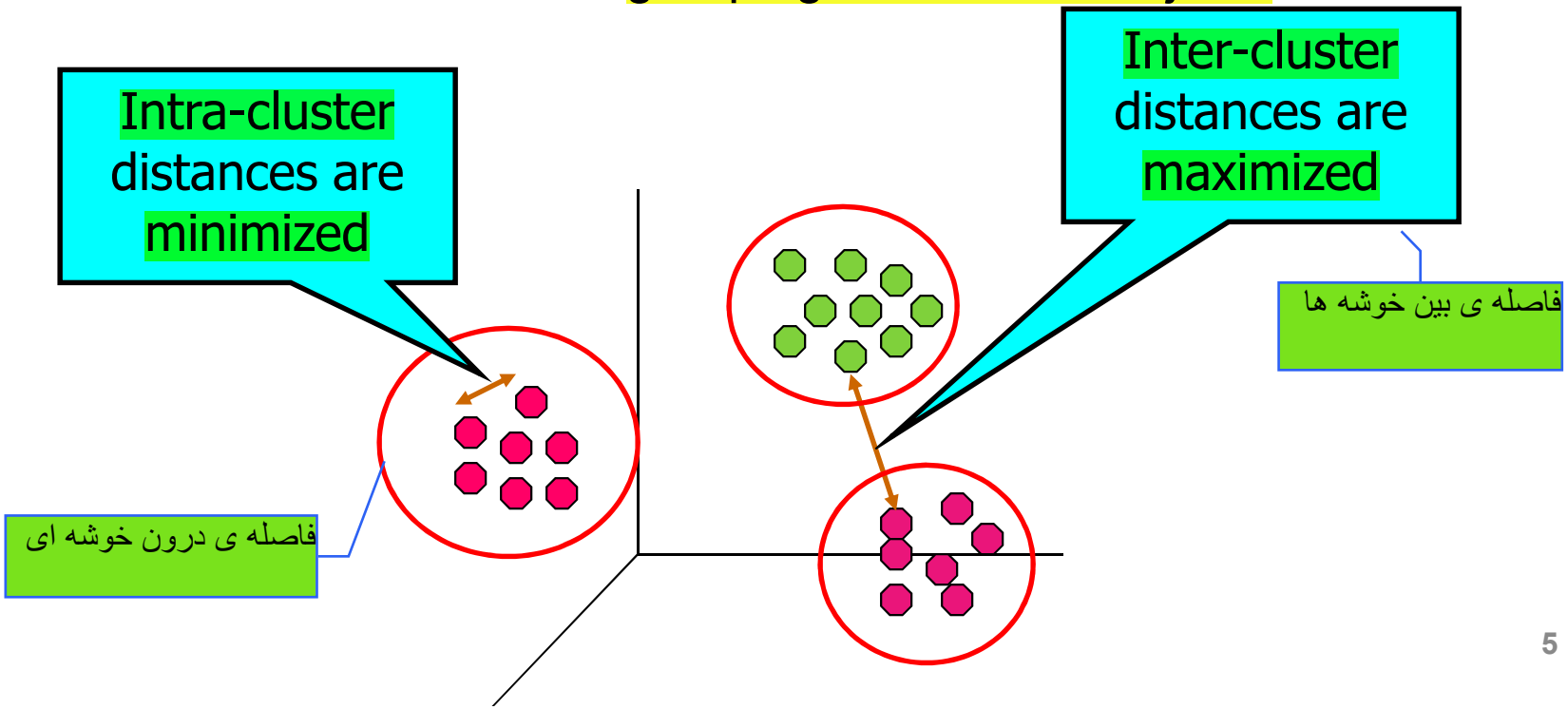


پس یه مشت داده میدن و میگن به یه تعداد دسته تقسیم کن و هرکدوم از رکوردها برای کدوم دسته میشه را باید بگیم.

What is Cluster Analysis?

کلاستر: مجموعه ای از ابجکت ها که دوتا ویژگی مهم دارند: ۱. این مجموعه از ویژگی ها توی گروهشون شبیه به هم هستند.
۲.نسبت به بقیه ی گروه ها متفاوت هستند.

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters



What is Cluster Analysis?

- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

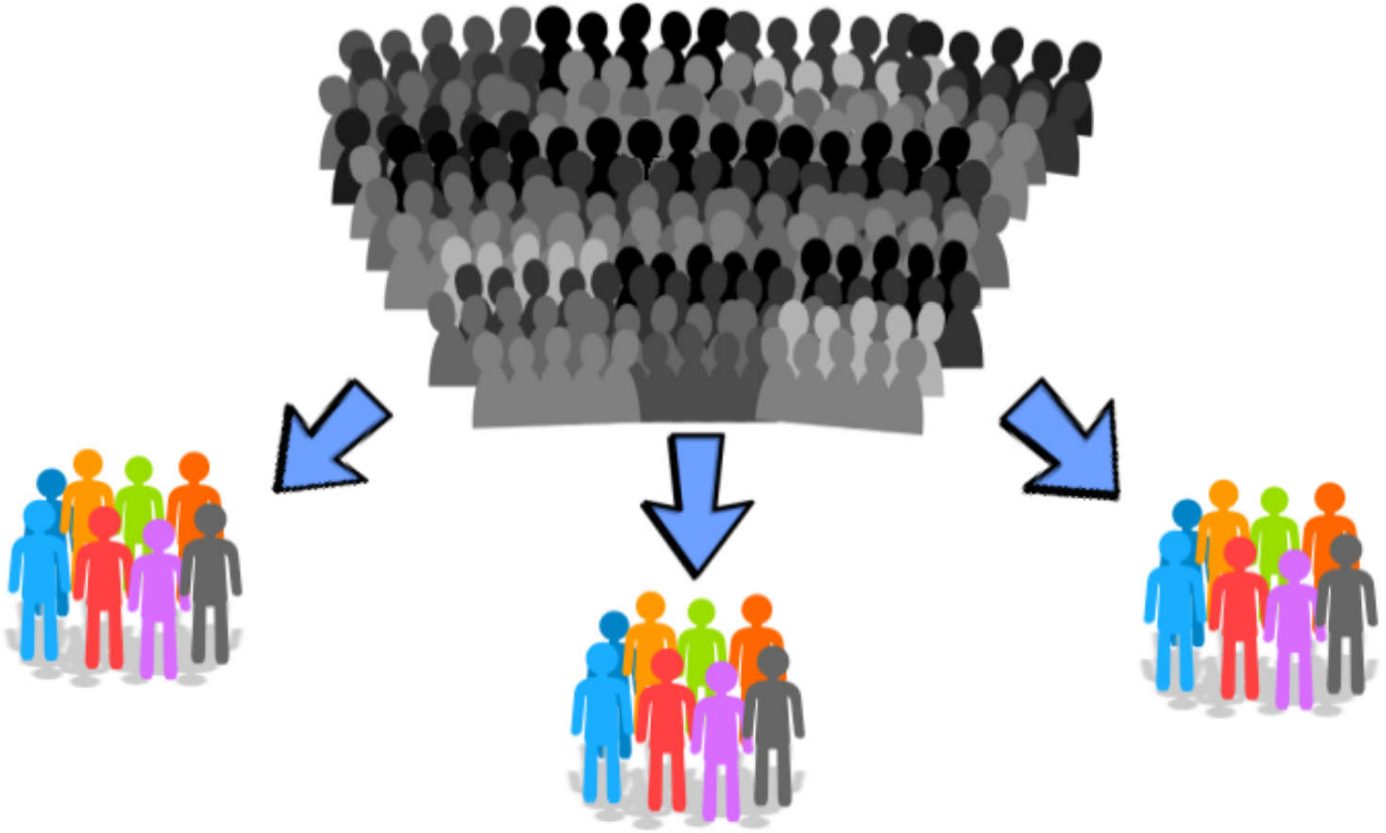
وقتی یه دیتایی رو بهمون میدن دیگه نتیجشو بهمون نمیدن مدل باید از اختلاف ها و تشابه ها تشخیص بده

مثال از کمک کلاسترینگ در مبحث پری پراسسینگ:
اگر به جای ۲میلیون رکورد بخاهیم ۱۰تا رکورد برداریم از داده هامون و بااونها کار کنیم کدوم ۱۰تا را برداریم تا نماینده ی خوبی از داده ها باشند؟
مثلا زمانی که ابعاد داده ها زیاد است بهتر است تعداد رکورد کمتری انتخاب کنیم

اماده کردن یه برچسب اولیه از داده ها

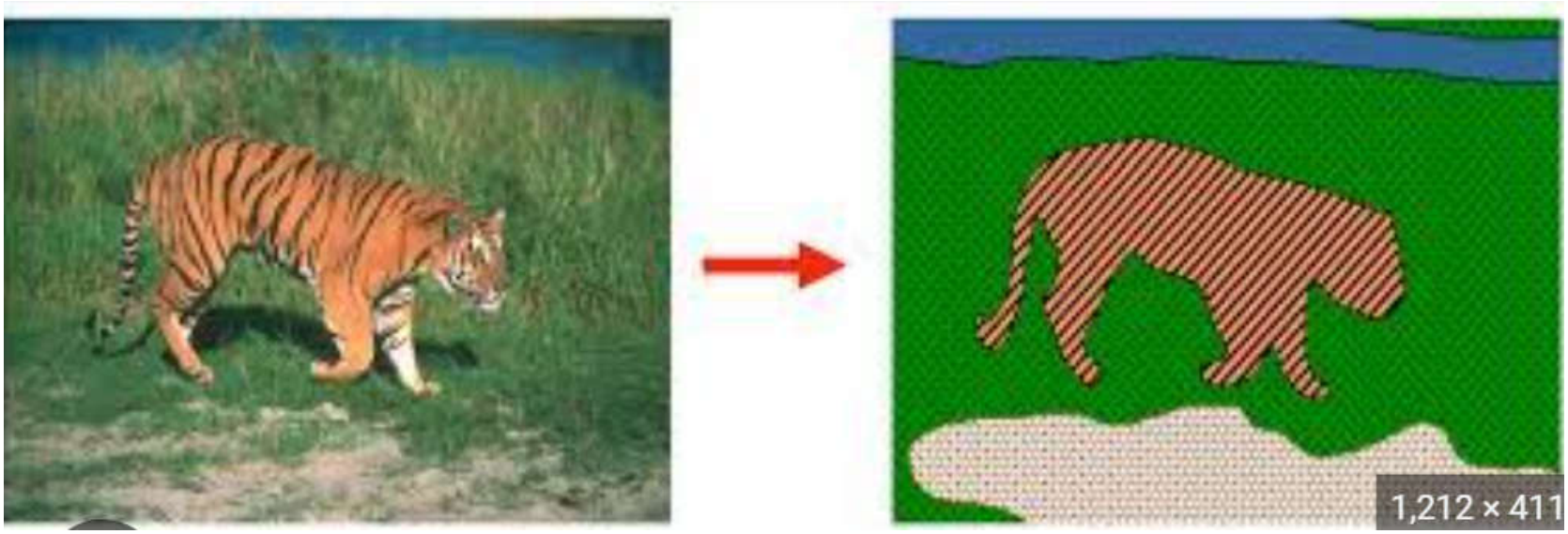
Example: customer segmentation

به دلیل تنوع بالای مشتری ها از کلاسترینگ استفاده میکنیم
خیلی خوب میشه که دسته بندی کنیم مشتری ها را تا برای هردسته یه استراتژی جداگانه درنظر بگیریم.



Example: image segmentation

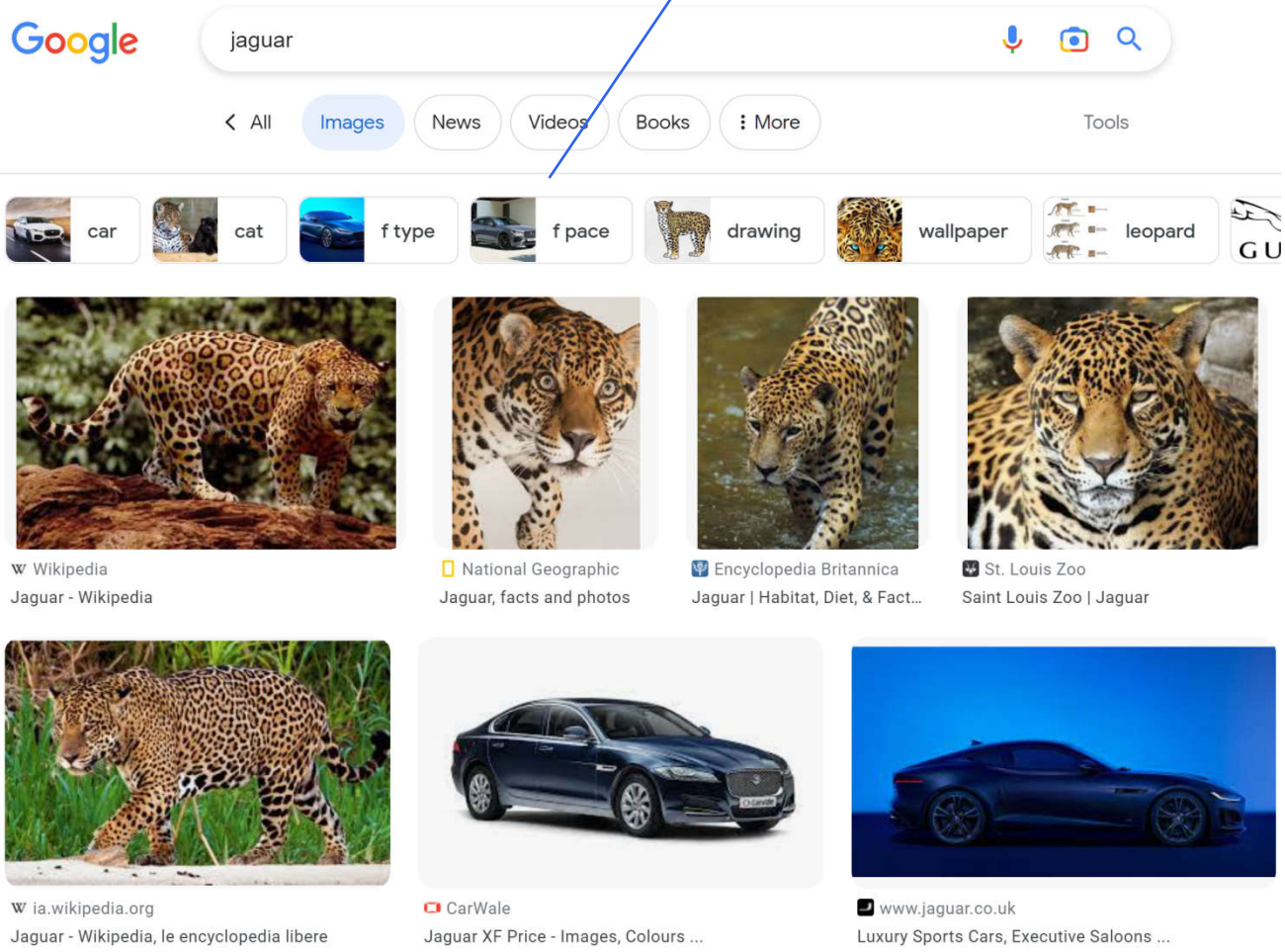
در پردازش تصویر خیلی مهم میشه که اطلاعات مهممون کجای عکسه؟
مثلا جداکردن بکگراند از فورگراند.
مثلا در مثال زیر جداکردن حیوان از سبزه ها



Example: information retrieval

الان مهمه که موتور جستجو هردوتا موضوع جگوار را به ما پیشنهاد ده هم ماشین هم حیوانش را

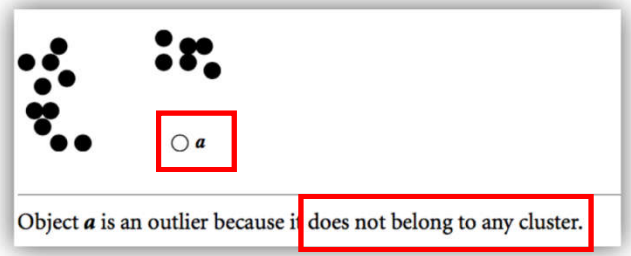
کلاسترهای مرتبط با موضوع سرچ شده

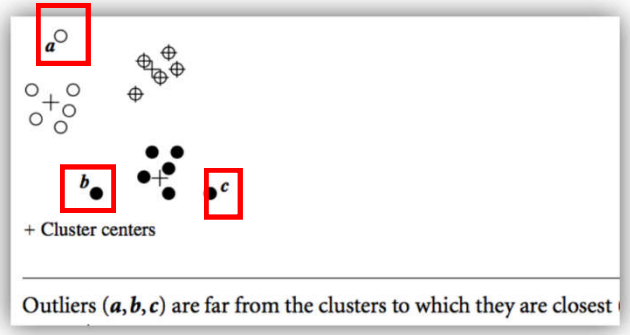


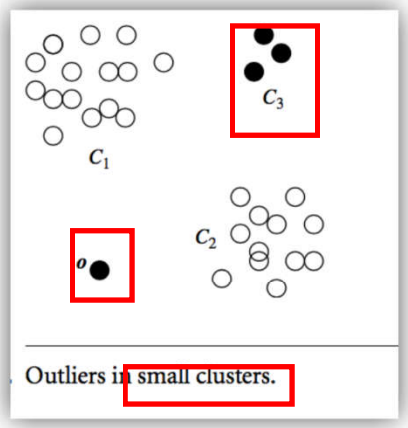
Example: outlier detection

- What are outliers?
- The set of objects are considerably dissimilar from the remainder of the data
- Outlier detection is useful in fraud detection applications such as creditcard fraud detection.
- It is also useful as a preprocessing tool.
- One way for outlier detection is using clustering:
  - Objects that do not belong to any cluster
  - Objects far from other objects in the same cluster
  - Clusters of very small cardinality

Example: outlier detection



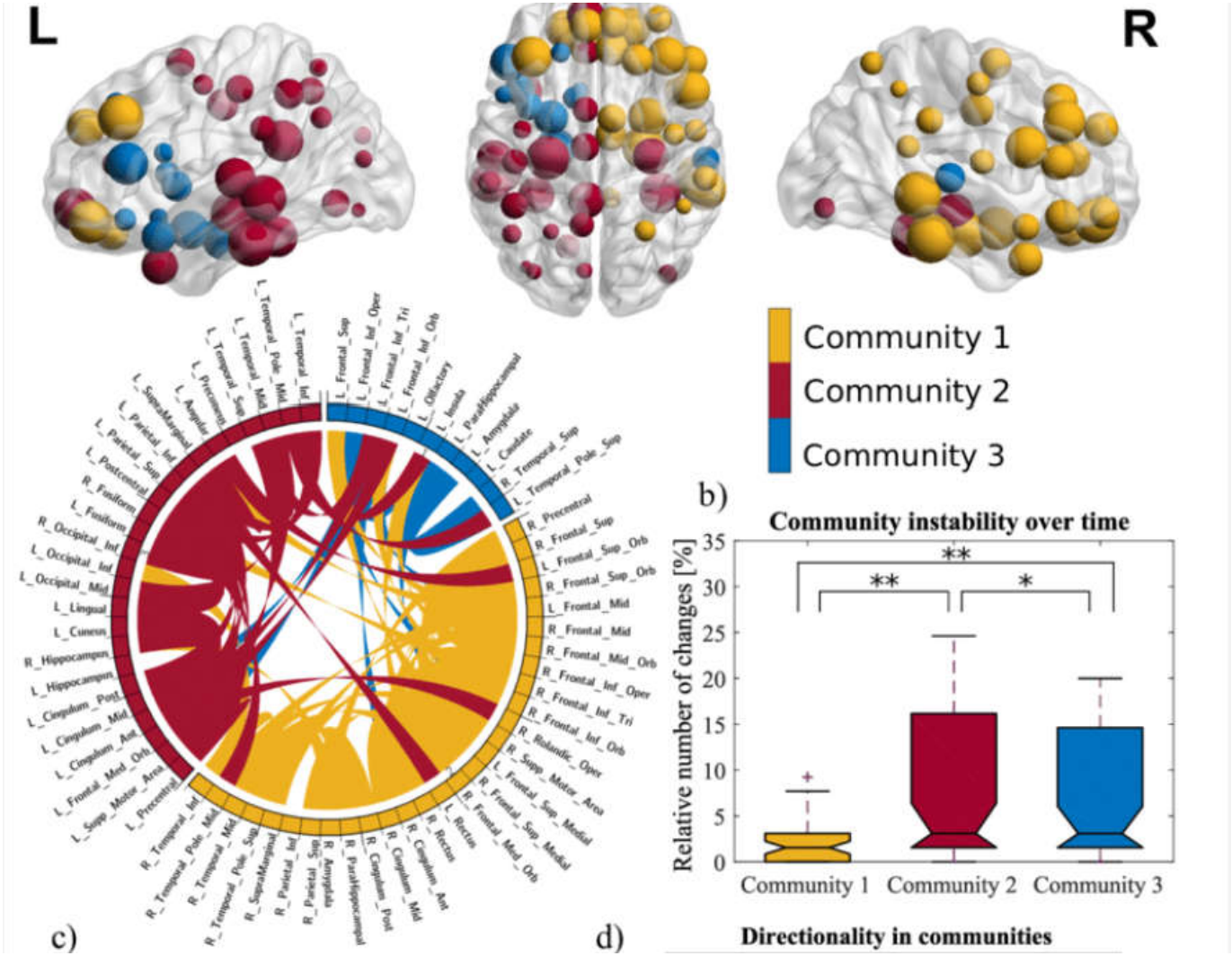




Refer to chapter 12 of the third edition of "Data Mining: Concepts and Techniques" for more information.

Example: community detection

کاربرد در شبکه های اجتماعی
پیداکردن علایق افراد، شغل افراد،
بررسی حال فرد براساس سلول های عصبی



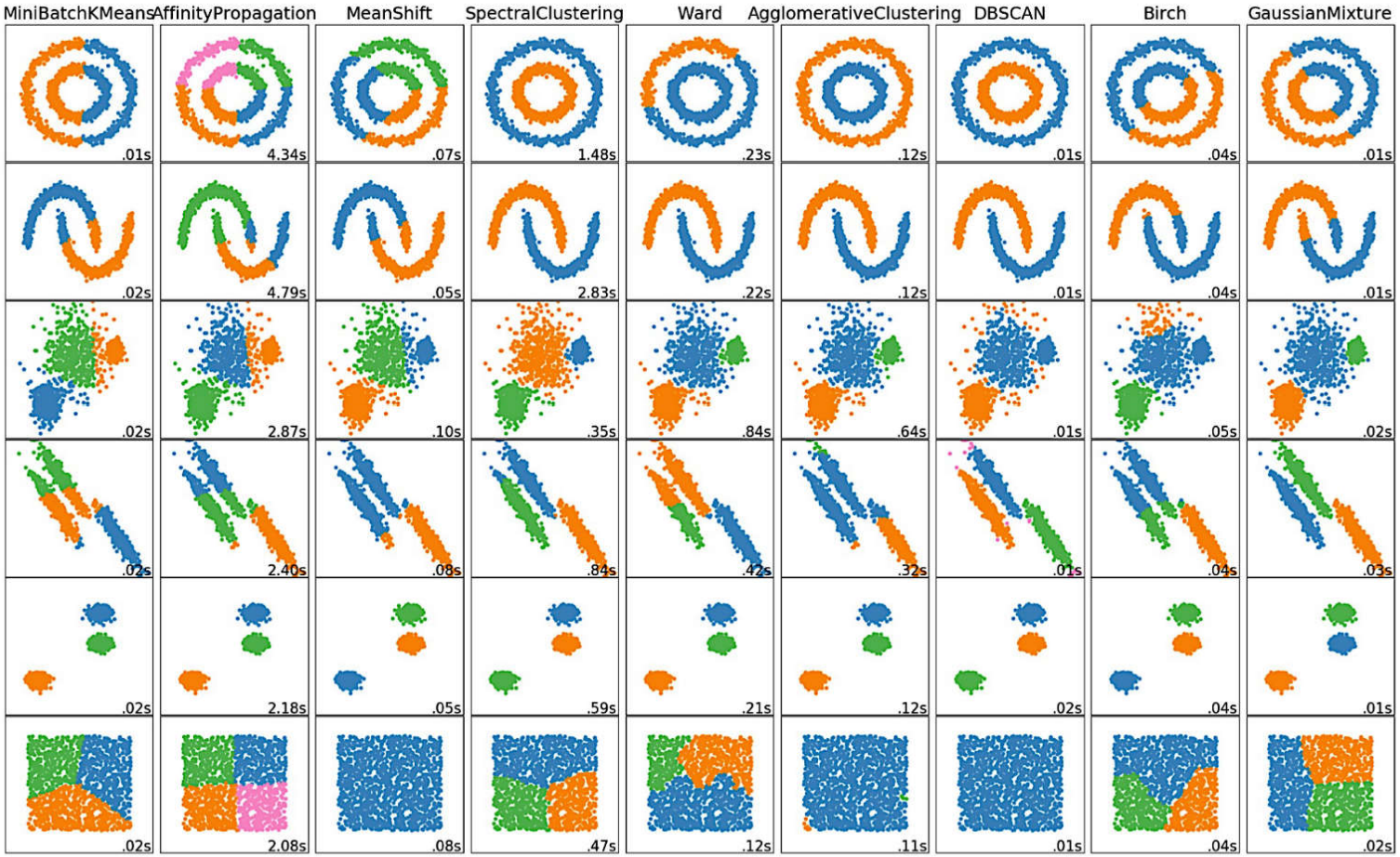
Clustering for Data Understanding and Applications

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market resarch

Clustering as a Preprocessing Tool (Utility)

- Summarization:
  - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
  - Image processing: vector quantization
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection
  - Outliers are often viewed as those “far away” from any cluster

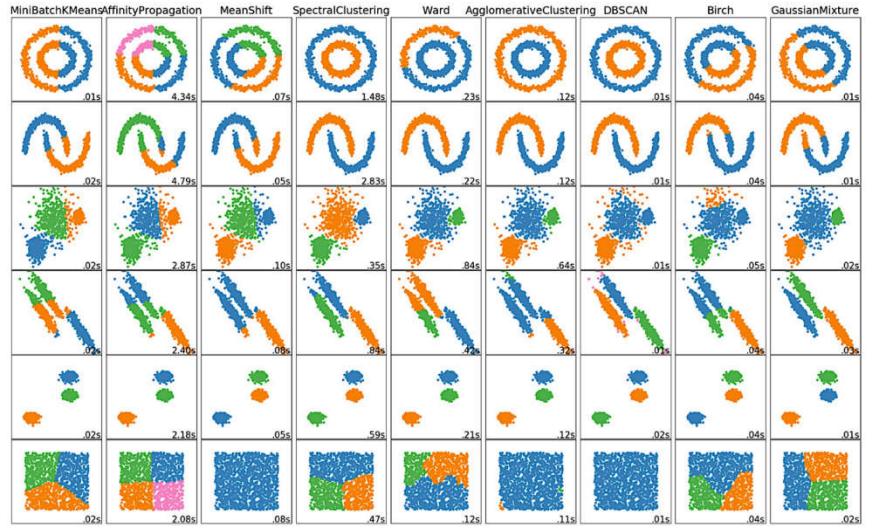
Clustering Toy Data



http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Clustering Toy Data

- These are toy 2D datasets.
- The last dataset is an example of a 'null' situation for clustering.
- With the exception of the last dataset, the parameters of each of these datasetalgorithm pairs has been tuned to produce good clustering results.
- Note that the intuitions might not apply to very high dimensional data.



http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

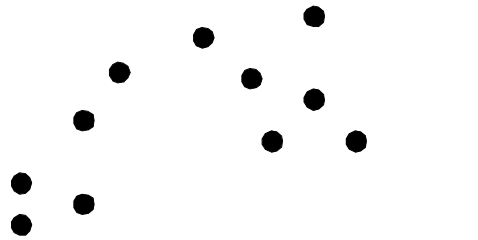
Types of Clusterings

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
  - Partitional Clustering
    - A division of data objects into non-overlapping subsets (clusters)
  - Hierarchical clustering
    - A set of nested clusters organized as a hierarchical tree

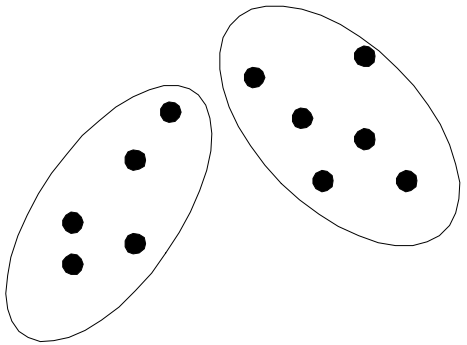
۱. تکه تکه کردن یا پارتیشن کردن کلاسترها
۲. کلاسترینگ سلسله مراتبی

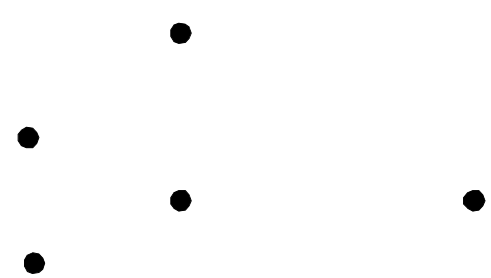
یه خوشه میتواند عضو یه خوشه بزرگتر باشه

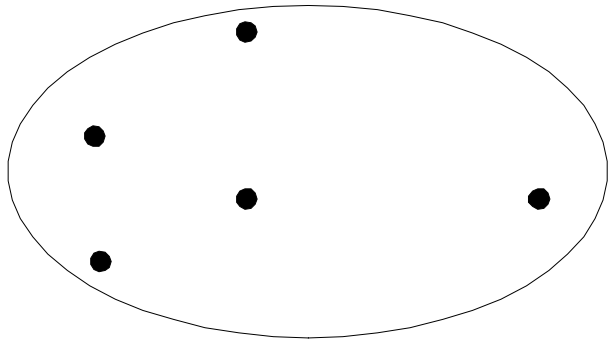
Partitional Clustering



Original Points

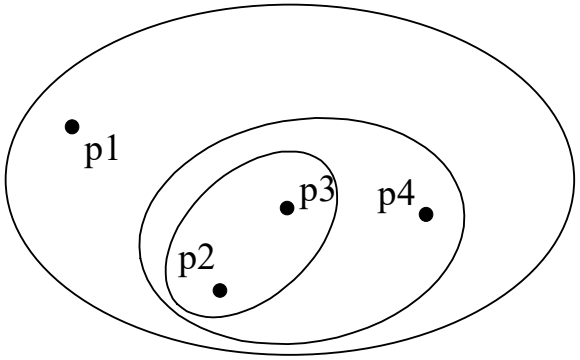




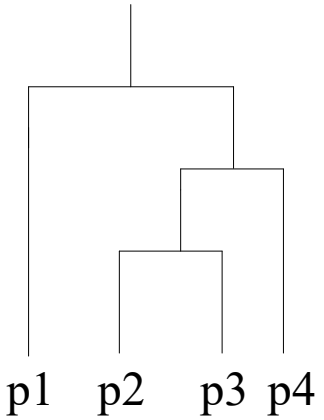


A Partitional Clustering

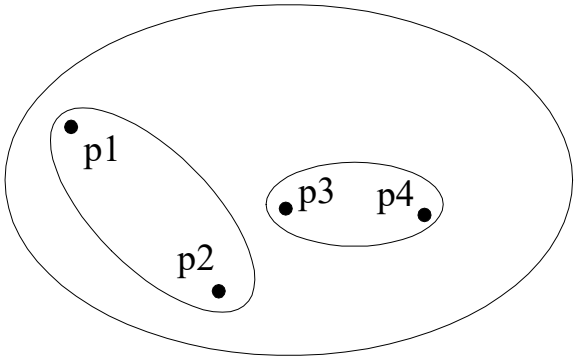
Hierarchical Clustering



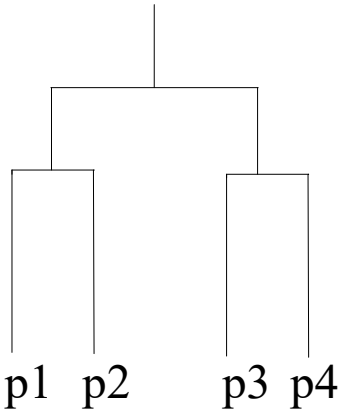
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Other Distinctions Between Sets of Clusters

- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
    - Can belong to multiple classes or could be 'border' points
  - Fuzzy clustering (one type of non-exclusive)
    - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
    - Weights must sum to 1
    - Probabilistic clustering has similar characteristics

آیا نمونه ای که داریم بررسی میکنیم قطعا رای یک خوشه است یا نه؟ نسبی است؟ یه نمونه ای با یک حتمالی متعلق به یک خوشه است

- Partial versus complete
  - In some cases, we only want to cluster some of the data

تکنیک خوشه بندی روی کل داده ها اعمال میشه یا روی یه بخشی از داده ها؟

Types of Clusters

انواع خوشه ها
خوشه ها در فضا چگونه نسبت به هم قرار میگیرند؟

- Well-separated clusters

خوشه ها کاملا از هم جدا هستند یعنی داده ها به گونه ای هستند که همه ی رکوردهای یک خوشه از یک خوشه دیگر فاصله دارند

- Prototype-based clusters

خوشه ها نسبت به یک نمونه ی اولیه از هم گسسته شدند.

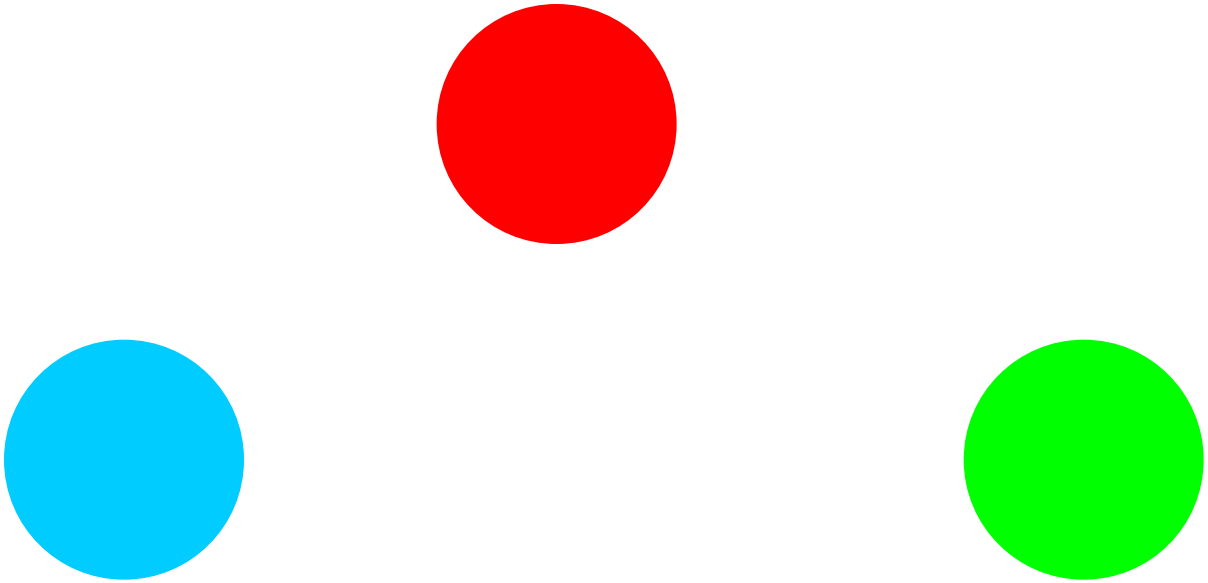
- Contiguity-based clusters
- Density-based clusters

خودمون تعریف کنیم که یک خوشه چگونه شکل میگیره؟

- Described by an Objective Function

Types of Clusters: Well-Separated

- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters: Prototype-Based

فاصله نقاط نسبت به یک نمونه
مثل مرکز متفاوت میشه

- Prototype-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the prototype or "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster

4 center-based clusters

مبتنی بر مجاورت

Types of Clusters: Contiguity-Based

پیوسته

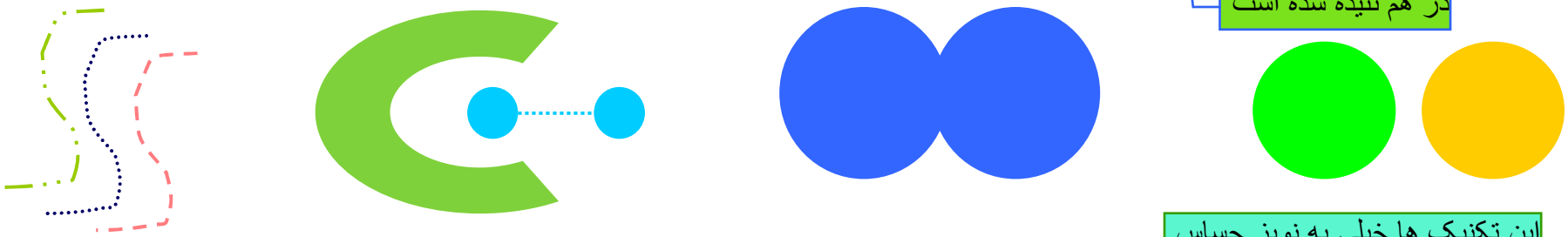
- Contiguous Cluster (Nearest neighbor or Transitive)

متعدی

یک درجه ای از نزدیک بودن و مجاورت را تعریف میکنیم نمونه های مجاور هم توی یک خوشه قرار میگیرند.

  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.
  - Used when the clusters are irregular or intertwined(non Noise)

در هم تنیده شده است



8 contiguous clusters

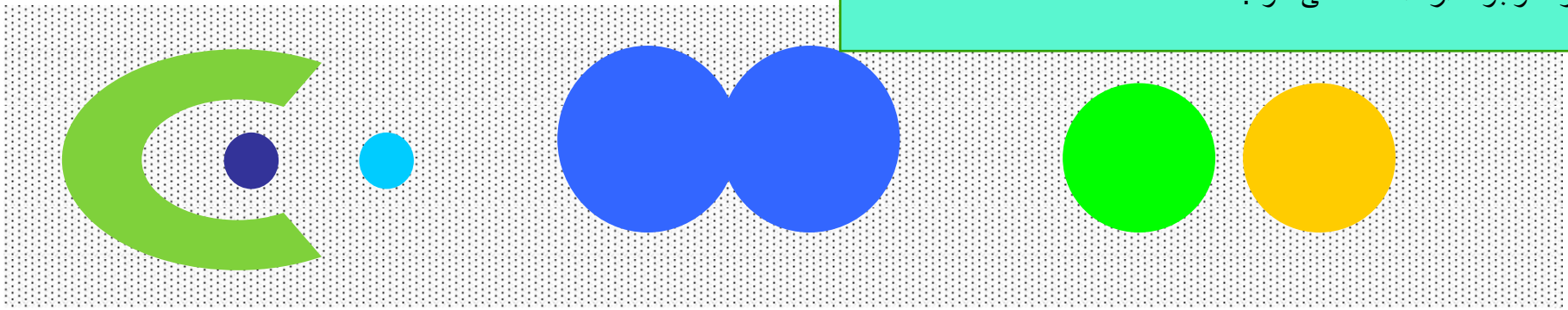
این تکنیک ها خیلی به نویز حساس هستند نویز مثل اینکه یه سری داده های تصادفی بریزیم توی داده های اصلی
پس برای داده هایی که نویز دارند از این روش ها استفاده نمیکنیم.

زمانی استفاده می شود که خوشه ها نامنظم یا در هم تنیده باشند (بدون نویز)

Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.

بر اساس تراکم
- خوشه ناحیه متراکمی از نقاط است که توسط مناطق کم تراکم از سایر مناطق با تراکم بالا جدا می شود.
- در مواقعی که خوشه ها نامنظم یا در هم تنیده هستند و زمانی که نویز و نقاط پرت وجود دارد استفاده می شود.



6 density-based clusters

برای داده های نویزی خیلی خوبه
نگاه میکنیم به چگالی داده ها

Clustering Algorithms

- K-means and its variants
- Hierarchical clustering
- Density-based clustering

از الگوریتم های پارتیشن بندی یعنی فضا را پارتیشن بندی میکنه

شکل کروی

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

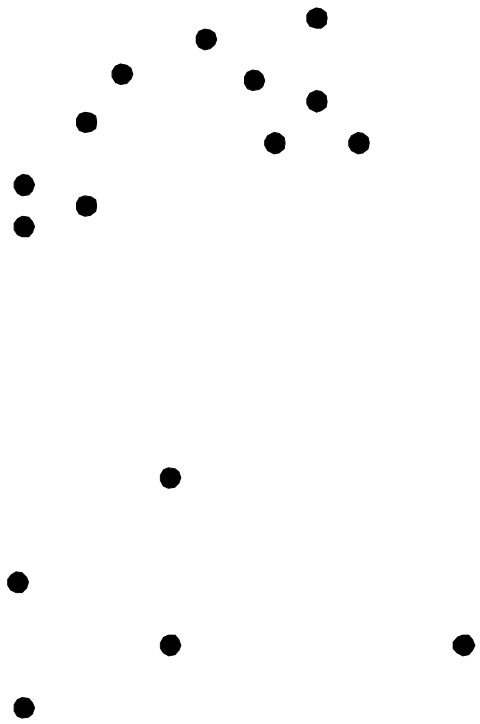
Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

K-MEANS CLUSTERING

Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters,



Partitioning Algorithms: Basic Concept

هدف : بهینه کردن یک تابع

- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters,

such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

تابعی تعریف میکنیم که فاصله از میانگین را تعریف میکنه تابع را طوری تعریف میکنیم که زمانی که خوشه بندی خوب انجام بشه مقداری که تابع میده کمینه میشه

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

نمونه هایی که توی هرخوشه هستند چقدر به مرکز هاشون نزدیک هستند؟

Partitioning Algorithms: Basic Concept

- Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

تعداد کلاسترها: k

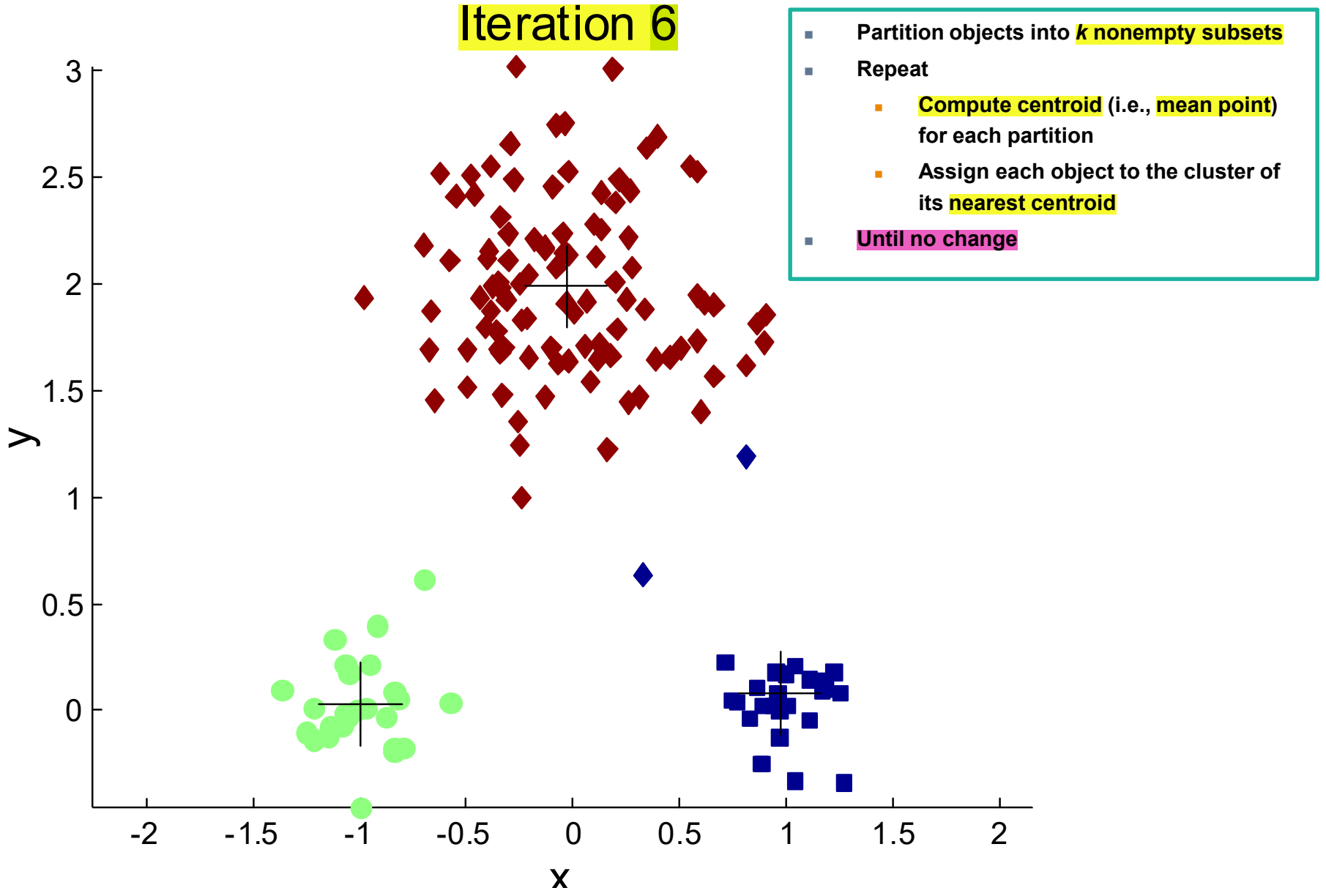
$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
  3. Assign each object to the cluster with the nearest seed point
  4. Go back to Step 2, stop when the assignment does not change

در ابتدا به صورت تصادفی n تا ابجکت را به k تا خوشه تقسیم بندی میکنیم.
بعدش توی این کا تا مجموعه میایم میانگین هر خوشه را حساب میکنیم.
حالا تک تک ابجکت هامون را براساس فاصله از این میانگینی که بدست اوردیم مجدداد به خوشه ها اساین میکنیم (خوشه ها را اپدیت میکنیم)

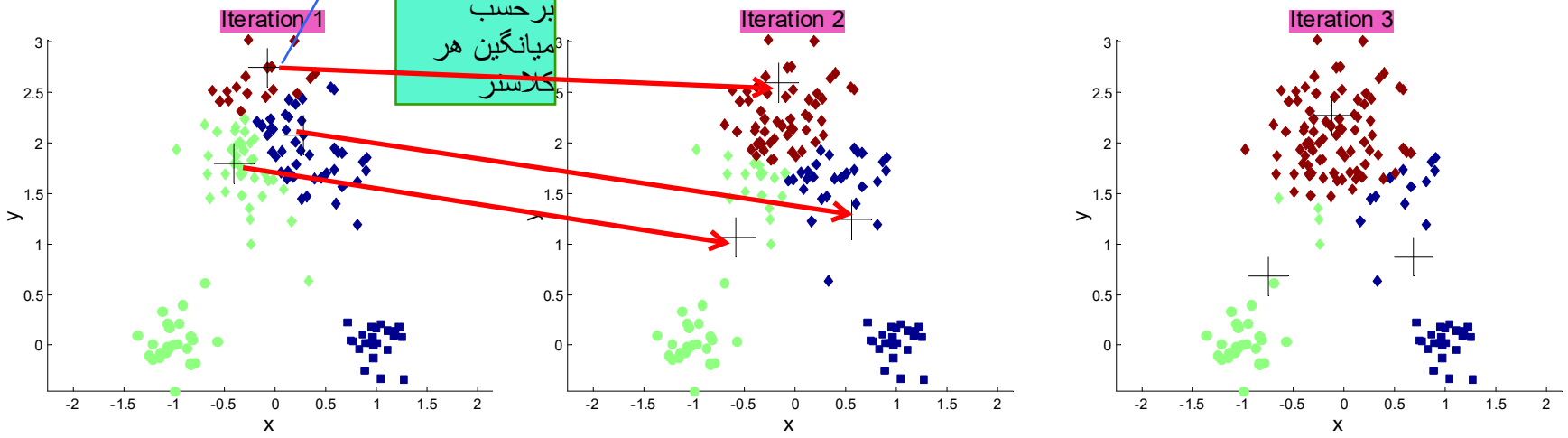
Example of K-means Clustering

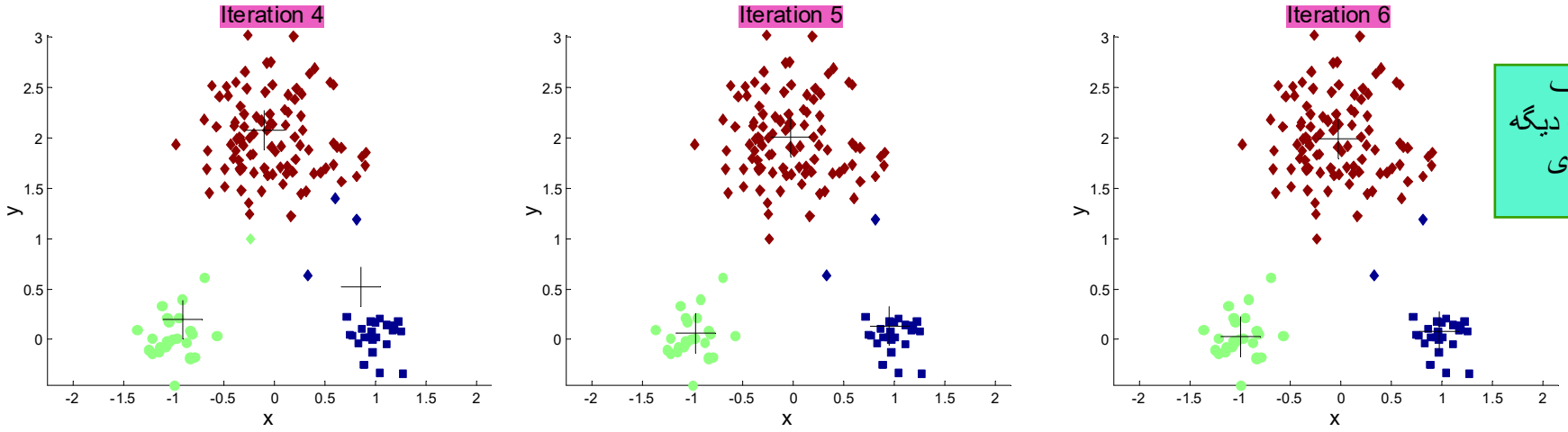


centroid of cluster

Example of K-means Clustering

اپدیت کردن برچسب ها برحسب میانگین هر کلاستر





چه زمانی متوقف میشیم؟ وقتی که دیگه برچسب نمونه ای تغییر نمیکنه

K-means Clustering – Details

- Simple iterative algorithm.
  - Choose initial centroids;
  - repeat {assign each point to a nearest centroid; re-compute cluster centroids}
  - until centroids stop changing.
- Initial centroids are often chosen randomly.
  - Clusters produced can vary from one run to another
- The centroid is (typically) the mean of the points in the cluster, but other definitions are possible (see Table 5.2).

K-means Clustering – Details

- K-means will converge for common proximity measures with appropriately defined centroid (see Table 5.2)
- 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

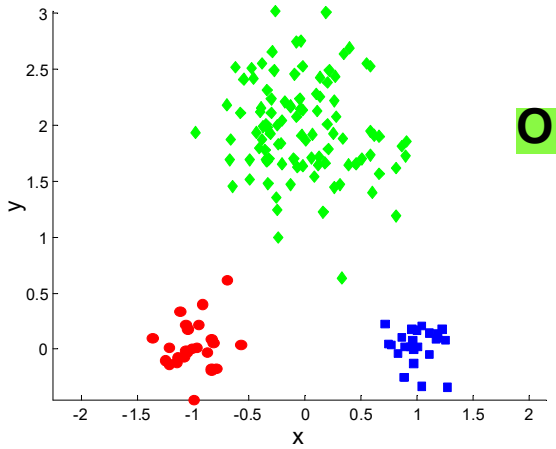
K-means Objective Function

- A common objective function (used with Euclidean distance measure) is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster center
  - To get SSE, we square these errors and sum them.

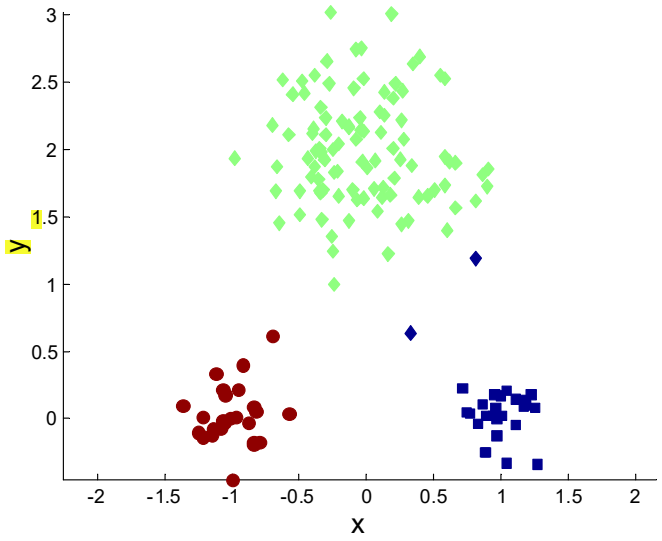
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the centroid (mean) for cluster C_i
  - SSE improves in each iteration of K-means until it reaches a local or global minima.

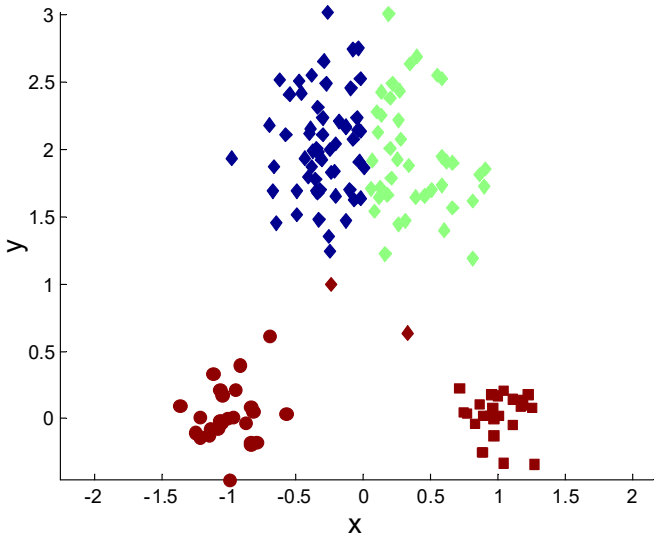
Two different K-means Clusterings



Original Points

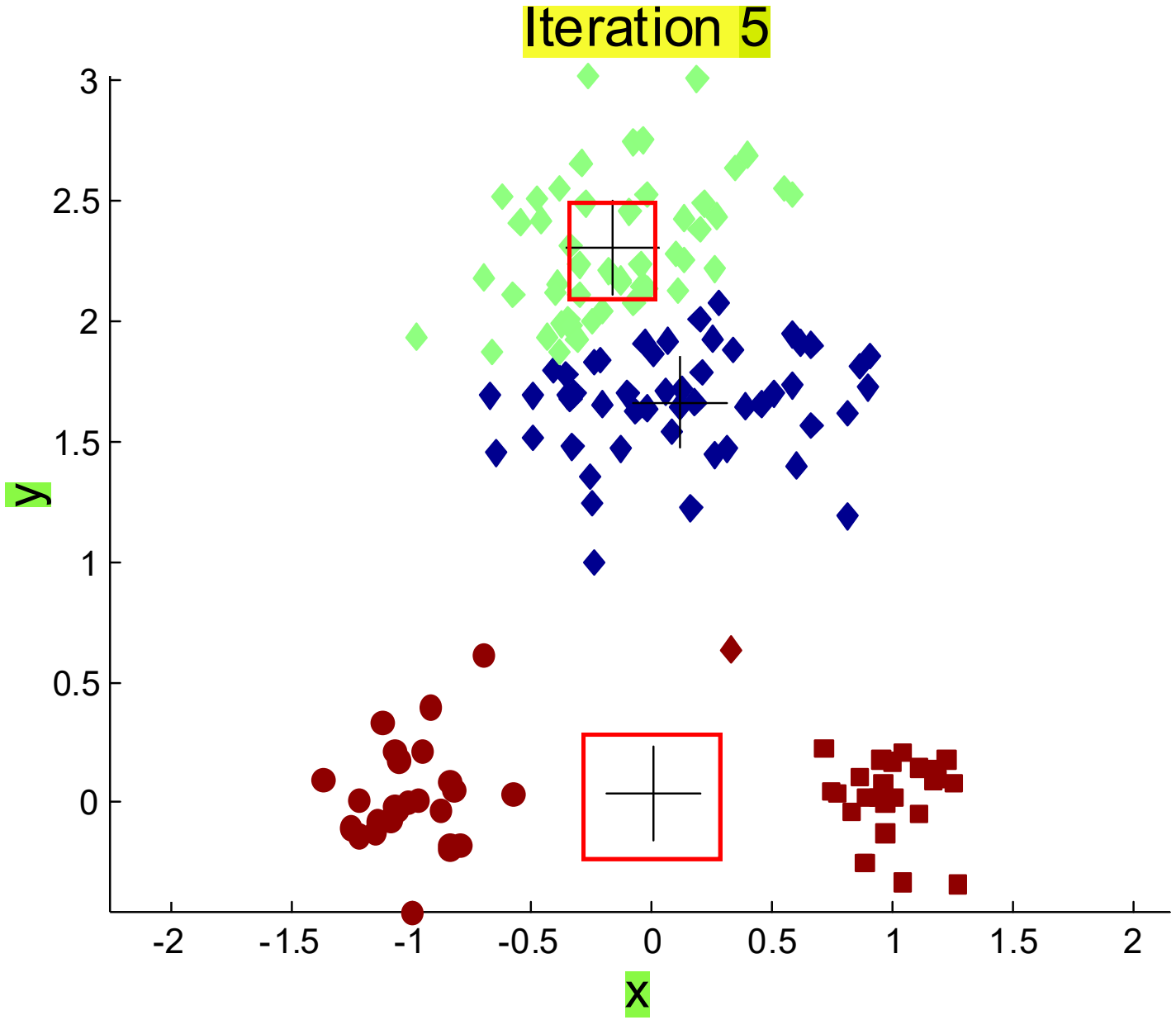


Optimal Clustering



Sub-optimal Clustering

Importance of Choosing Initial Centroids ...



Importance of Choosing Initial Centroids ...