individual projects

Arend Hintze

deliverables

- Progress session
- Final oral presentation: 10-25 minute talks about the project depending on group size summarizing the problem, methods, results
- Writeup: min. 1.5 - max. 3 pages, introduction, methods, result, discussion, including figures
- Code and Data needs to be submitted extra

Progress report

- Explain the project
- Explain where you are in the project, what you have accomplished, what needs to be done, and how you are going to solve those issues
- Possibly expose open questions, use the audience to get opinions and discuss problems
- questions the audience has gives you an insight about explanation deficits

Writeup

- Introduction explains what the problem is and why it is worth (or interesting) to solve

- methods explains the data source and structure and describes the tools or algorithms used

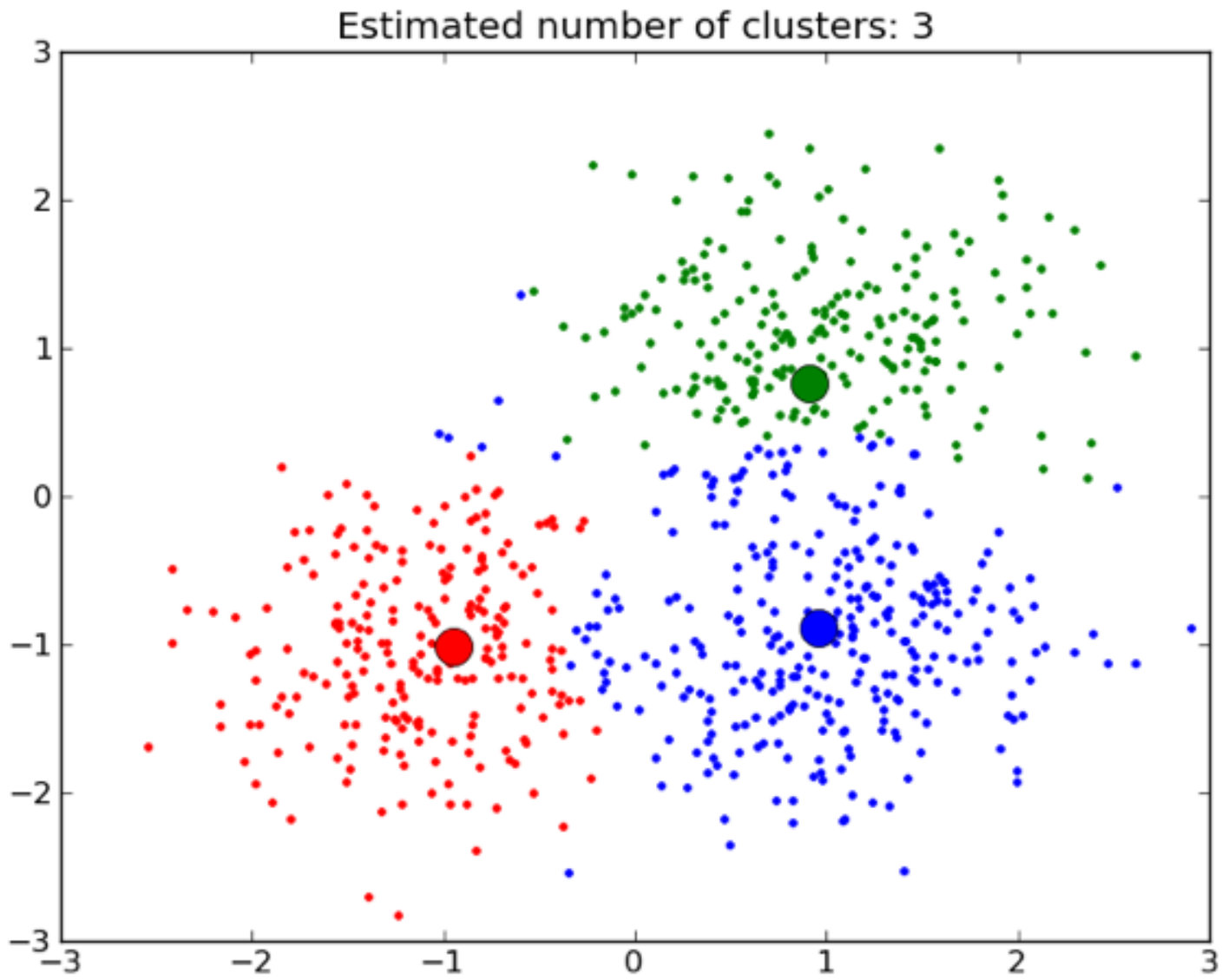
- results show the results and include figures or graphs visualizing the finding

- discussion elaborates in errors, mistakes, issues, and most importantly puts the results in perspective

clustering I

Arend Hintze

Cluster Analysis



problem definition

- given a collection of data instances
- where each instance is characterized by an attribute set X
- partition the data in such a way that instances in the same partition (cluster) are more similar to each other than to instances in other partitions (clusters)

examples

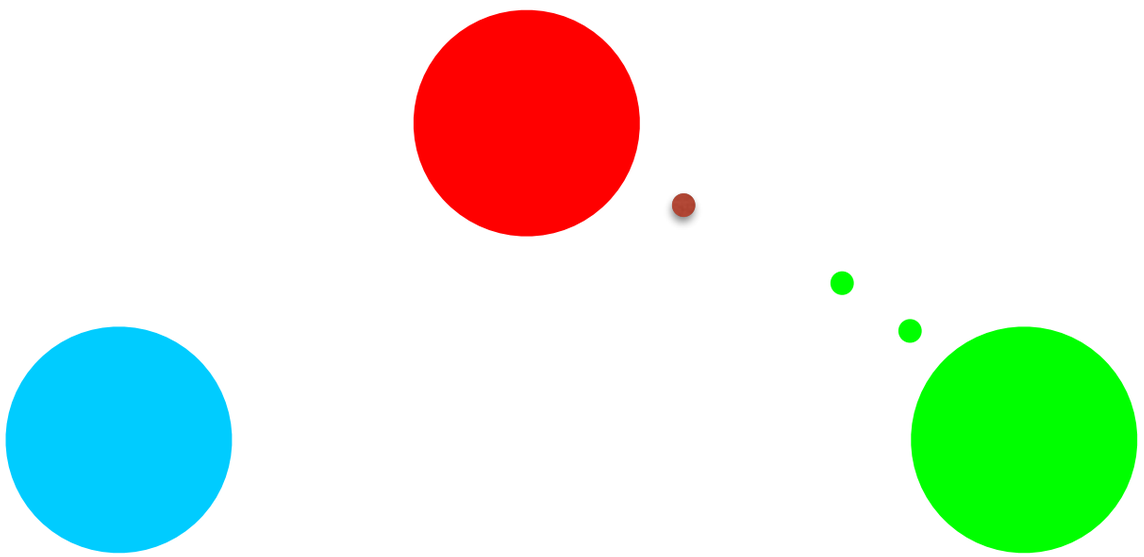
Instance	Attribute set, x	Clustering Task
Document	Words in documents	Group documents based on their similar topics
Customer	Demographic and purchase information	Group together similar customers
Location	GPS trajectories of mobile phone users	Finding hot spots frequently visited by users

different types of clusters and their definition

- Well-separated clusters
- center-based clusters
- contiguous clusters
- density-based clusters

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 that any point not in the cluster



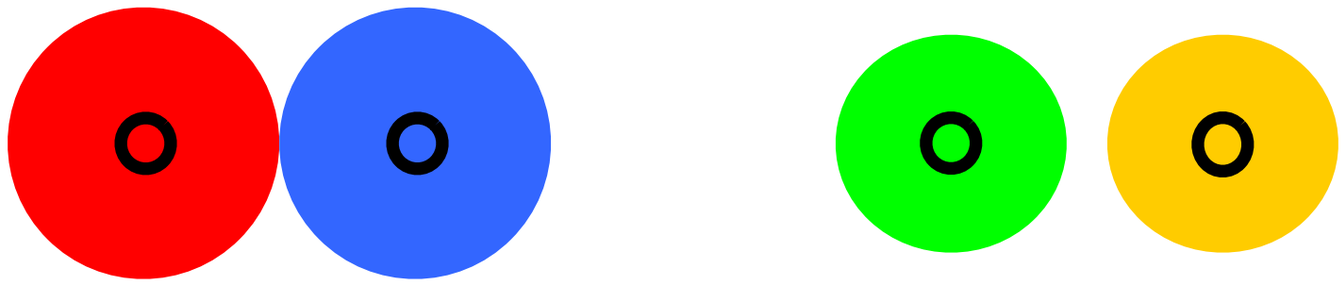
Center-Based Clusters

- A cluster is a set of objects such that an object in a cluster is closer to the “center” of it’s cluster, than to the center of any other cluster

- the center of a cluster is called “centroid”

- the average of all points in a cluster define the centroid

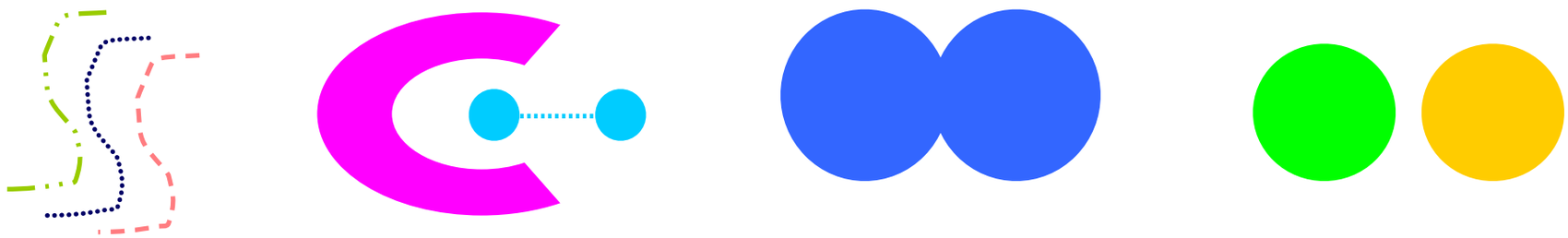
- “mediod” is the most representative point of a cluster



Contiguity-Based Clusters

- Nearest neighbor transitive

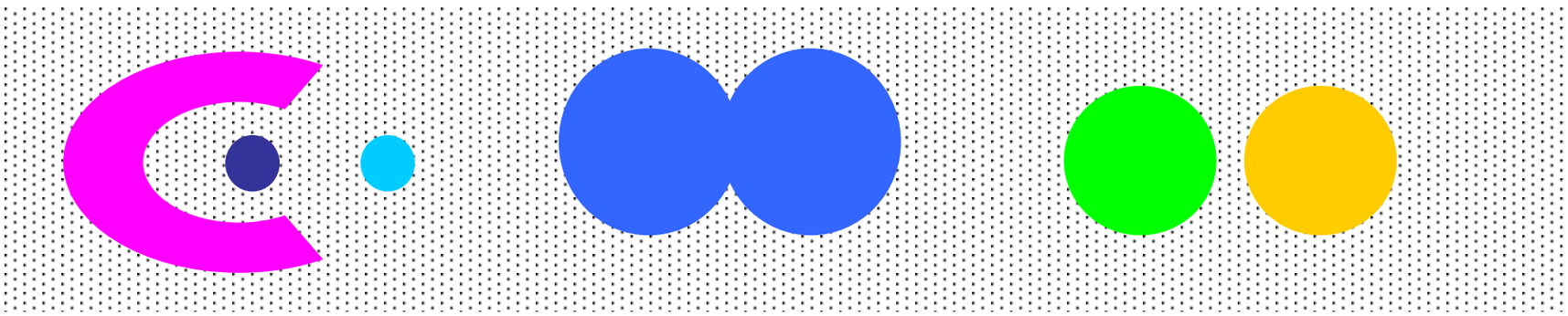
- A cluster is a set of points such that each point in a cluster is closer (or more similar) to ONE or more points it's cluster than to any other point not in the cluster



Density-Based Clusters

- a cluster is a set of points which all together are separated from other sets of points (clusters) by regions of low density

- preferred is data is noisy, shapes are irregular, or outliers are present



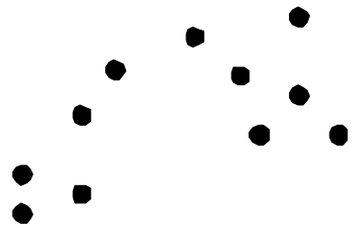
types of clusterings

- A clustering is a set of clusters (partitioning of the data)

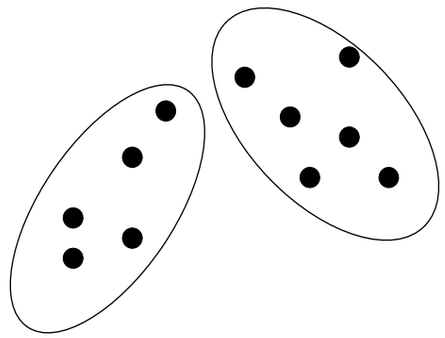
- hierarchical clustering algorithms create “trees” or hierarchies of clusters, where starting from a single cluster which consequently is split into further clusters creating a tree

- partitional clustering algorithms create a clustering that has no further organization

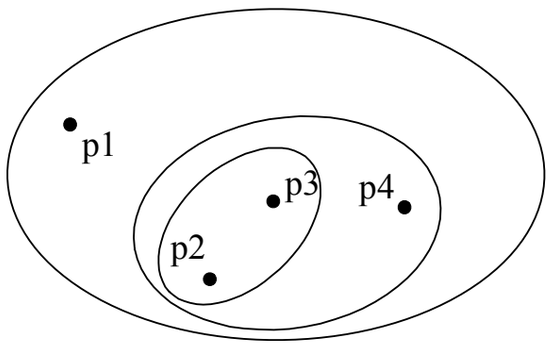
Partitional vs Hierarchical



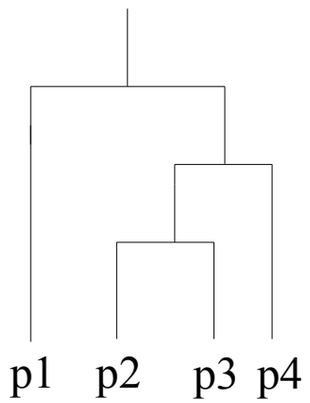
Original Points



A Partitional Clustering



Hierarchical Clustering



Dendrogram

Partitional vs Hierarchical

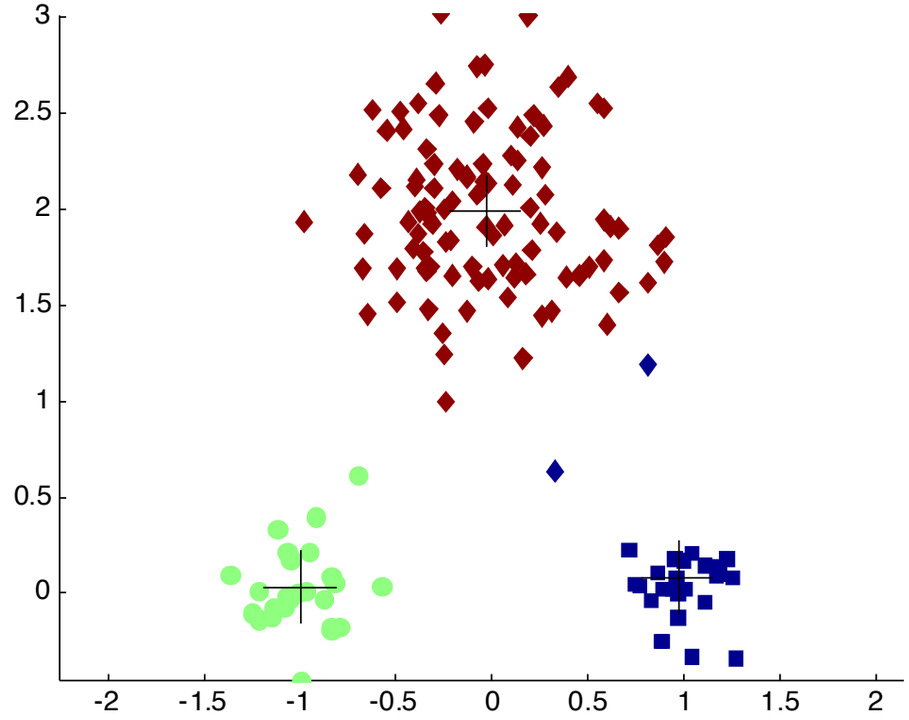
- partitional
    - k-means (self-organized maps, bisection k-means, fuzzy k-means, kernel k-means)
    - spectral clustering
    - support vector clustering
    - density based
    - ensemble clustering
- hierarchical
    - divisitive (MST)
    - agglomerative (single-link, complete-link, group average, Ward's method)

k-means clustering

- partitional clustering approach
- each cluster is associated with a centroid
- number of clusters (k) must be specified

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

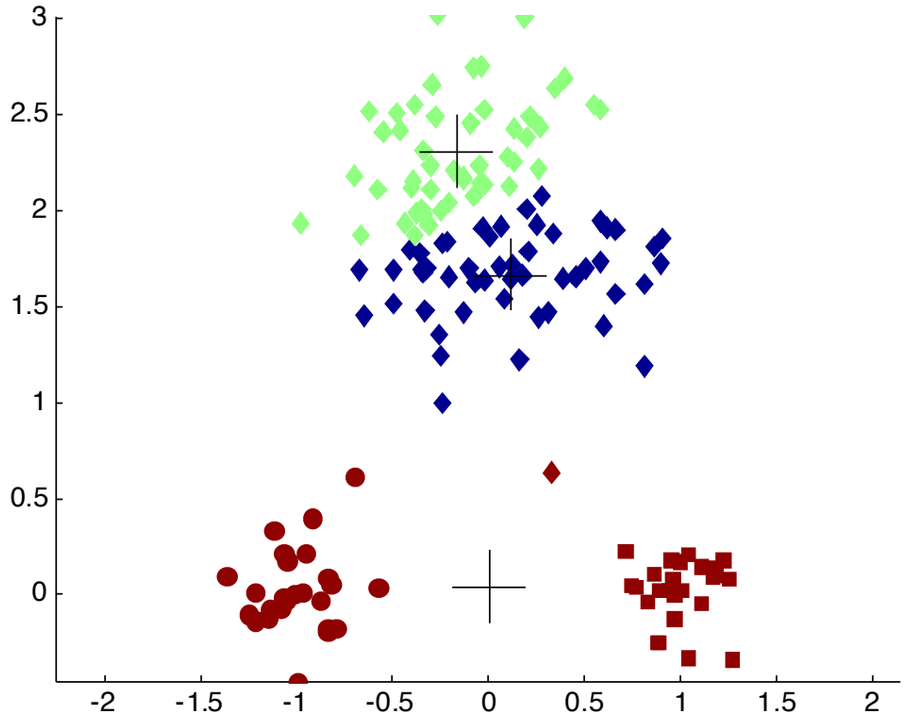
Illustration



Interpreting k-Means Results

- cluster membership of each data instance
- centroid vector
    - the centroid contains information that summarizes the characteristics of each set (cluster)

starting point limitations



further details k-means

- Initial centroids are chosen randomly -> not the same result every time

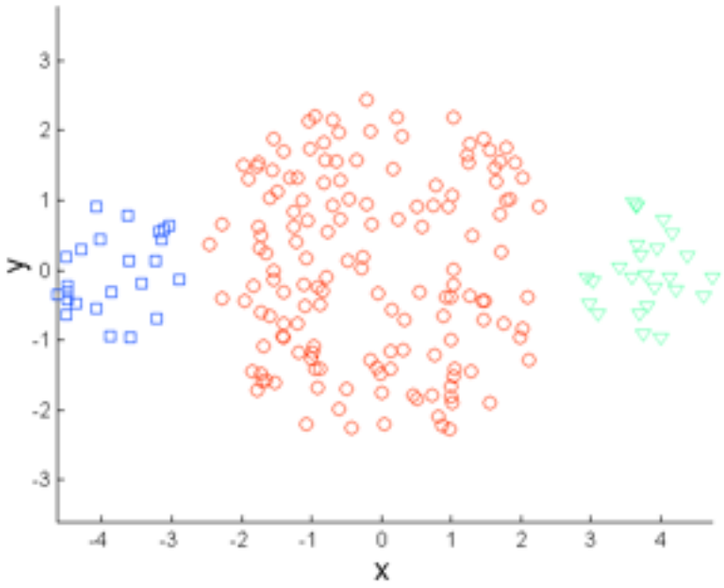
- closeness is a proximity measure and thus depends on the measure used (euclidean, cosine, correlation…)

- Complexity is O(n*k*I*d) [n=points,k=clusters,I=iterations,d=attributes]

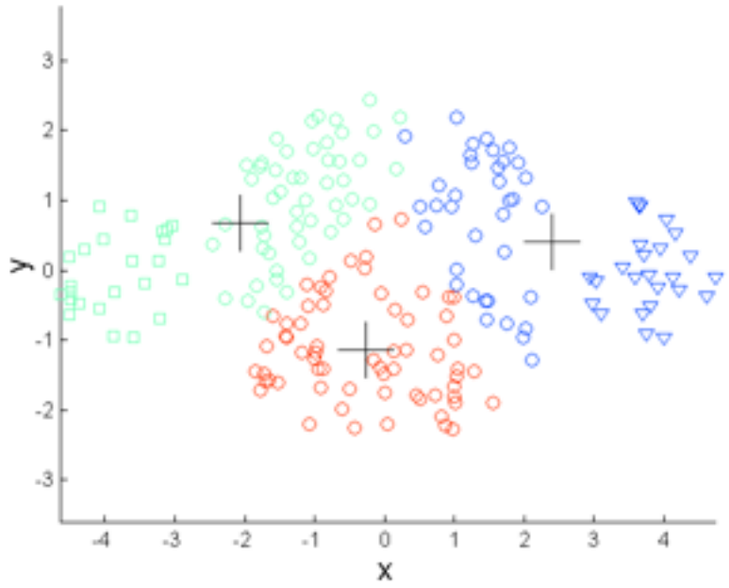
limitations of k-mean

- Sizes of clusters
- Densities
- non-globular shapes
- outliers

size issues

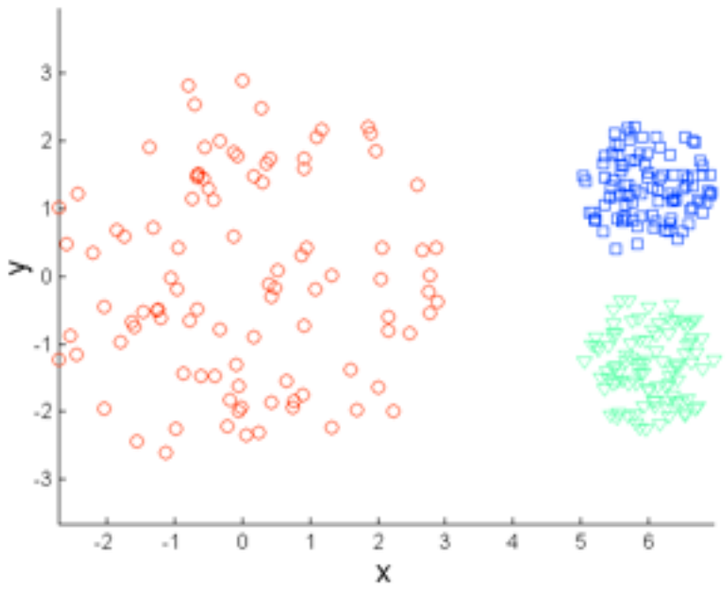


Original Points

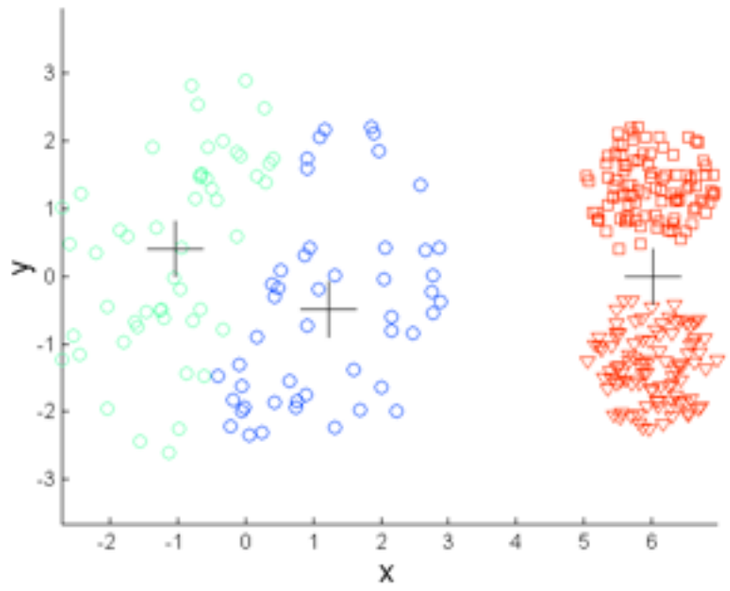


K-means (3 Clusters)

density issues

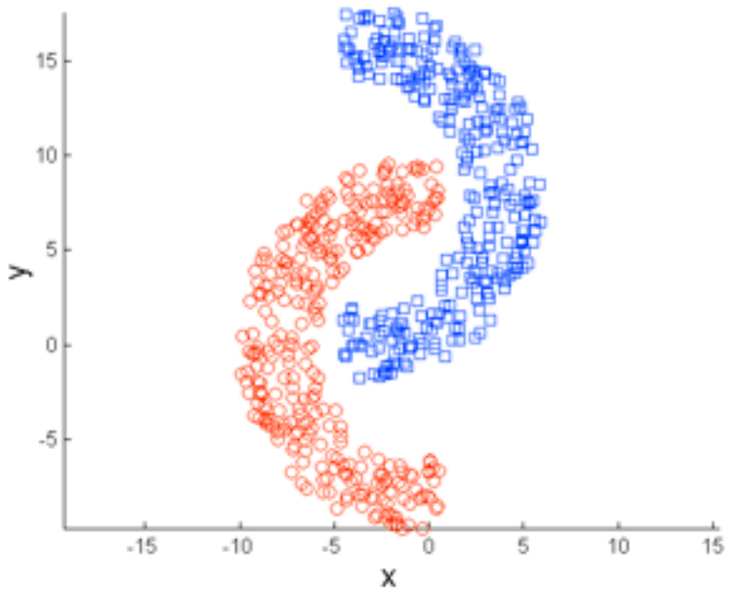


Original Points

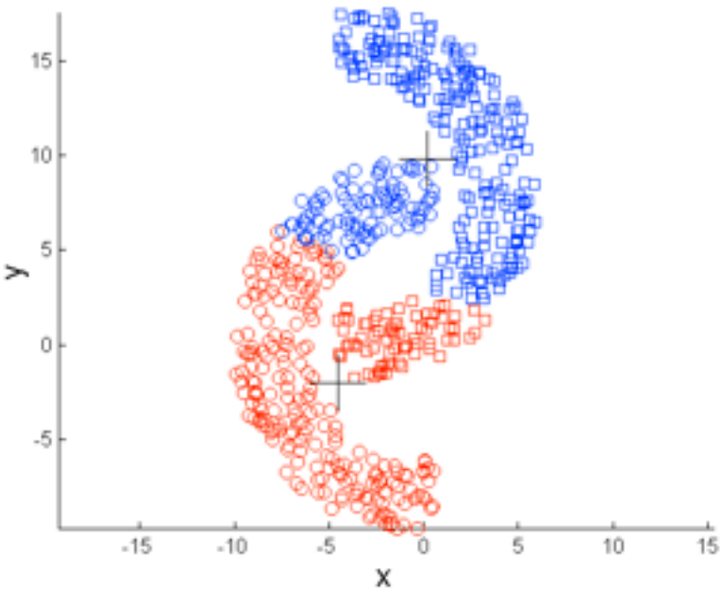


K-means (3 Clusters)

shape issues



Original Points

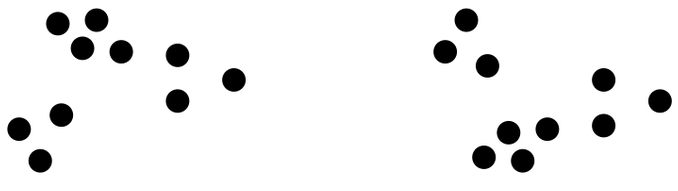


K-means (2 Clusters)

good vs. bad clustering - validity?

- classification methods have accuracy, precision, recall…
- for clustering methods the analogous question would be “how good are the clusters?”
- problem: “clusters are in the eye of the beholder”

find the cluster:



How many clusters?

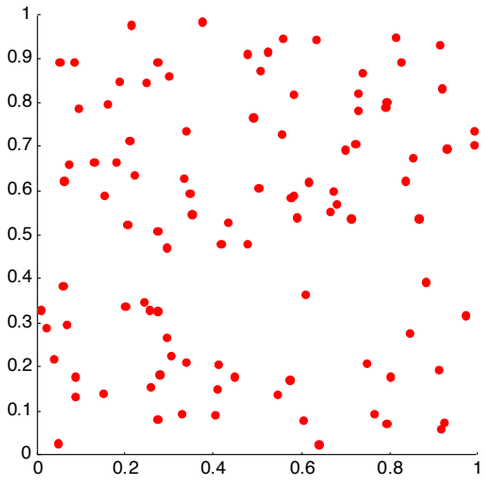
Six Clusters

Two Clusters

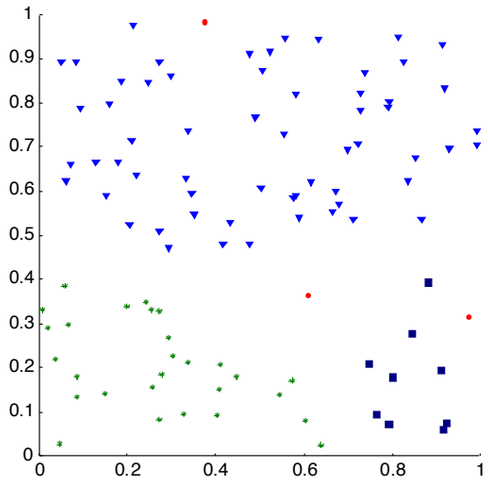
Four Clusters

clusters found in random data

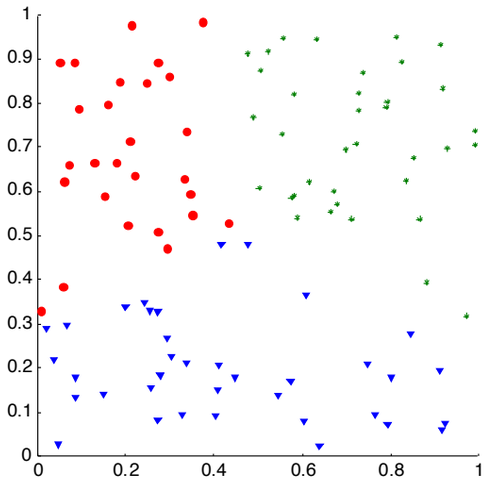
Random Points



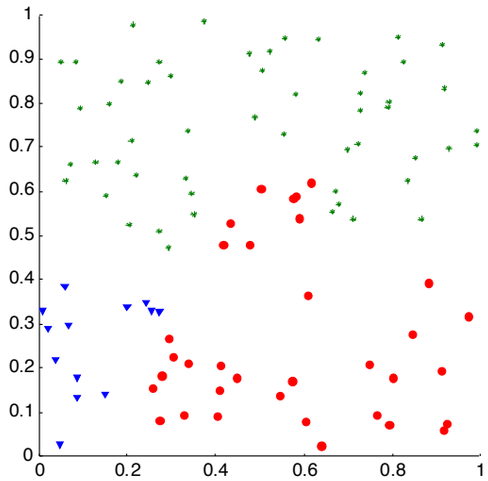
DBSCAN



K-means



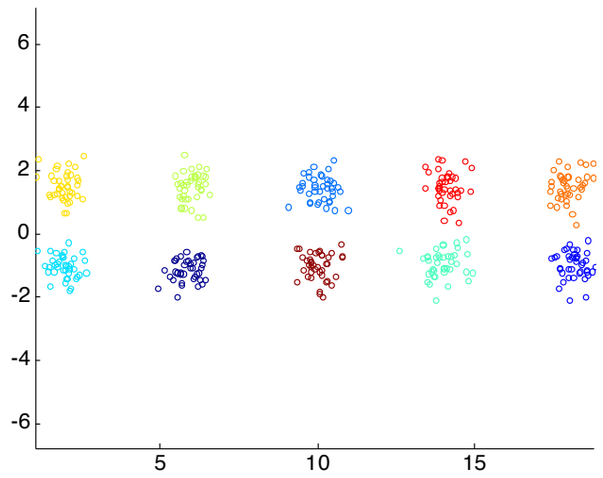
Complete Link

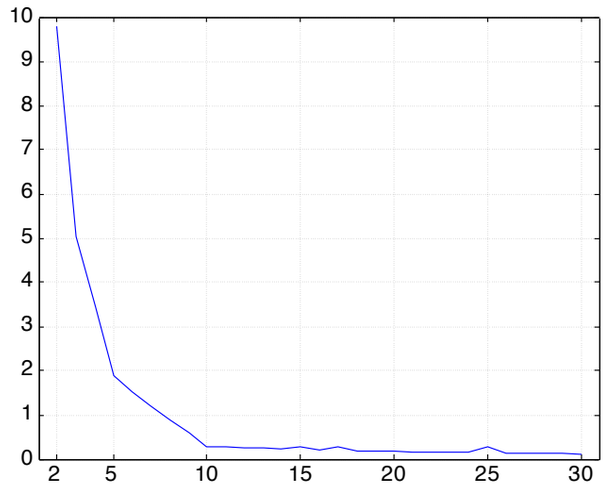


Sum of Squared Error (SSE)

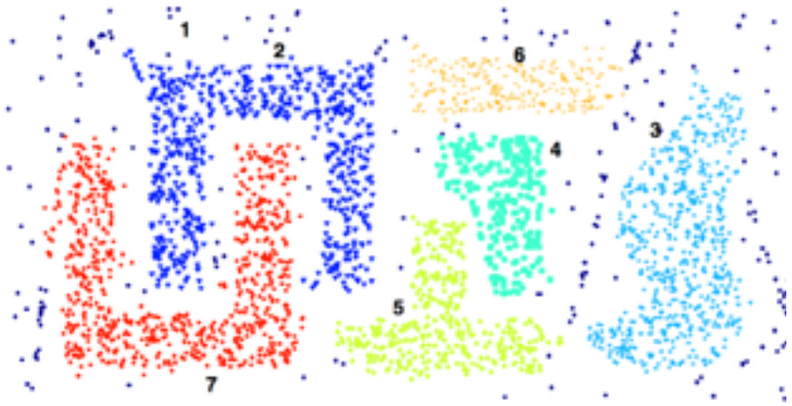
- sum the squares of the distances (either to each other or to the centroid)

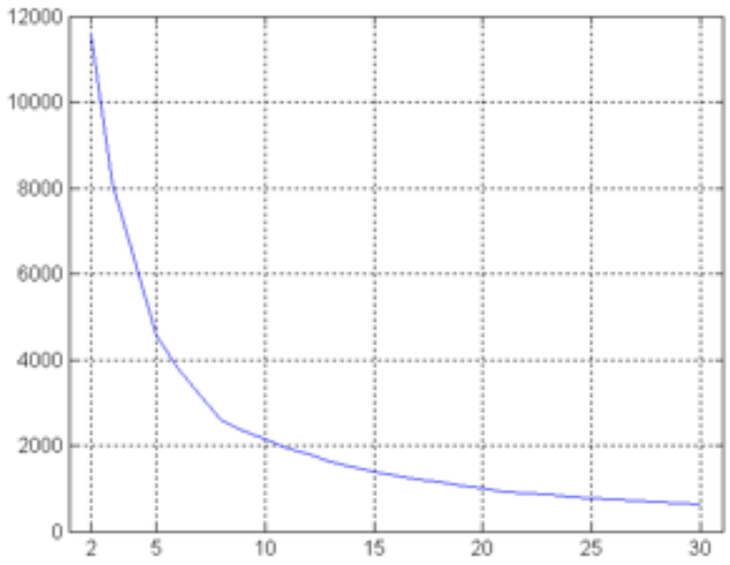
- can also be used to estimate the number of clusters





more complicated example





SSE of clusters found using K-means

Custer validity?

- Once you computed SSE, what does the value say in comparison?
- If you have 10 different results, say five times the same and also five different ones, which is the best result?

Bootstrapping

- compare the SSE of 0.005 (imaginary value) to random clustering and the distribution thereof