



ICSI431/ICSI531 Data Mining

Lecture 5

Clustering

Feng Chen

fchen5@albany.edu

<http://www.cs.albany.edu/~fchen/course/2018-ICSI-431-531>

Slides adapted from David Sontag, Luke Zettlemoyer, Vibhav Gogate, Carlos Guestrin, Andrew Moore, Dan Klein, Pang-Ning Tan, Michael Seinbach, Vipin Kumar

Outline

- Introduction
- Cluster Methods
 - K-Means
 - Agglomerative (Hierarchical) Clustering
- Cluster Validation

Clustering

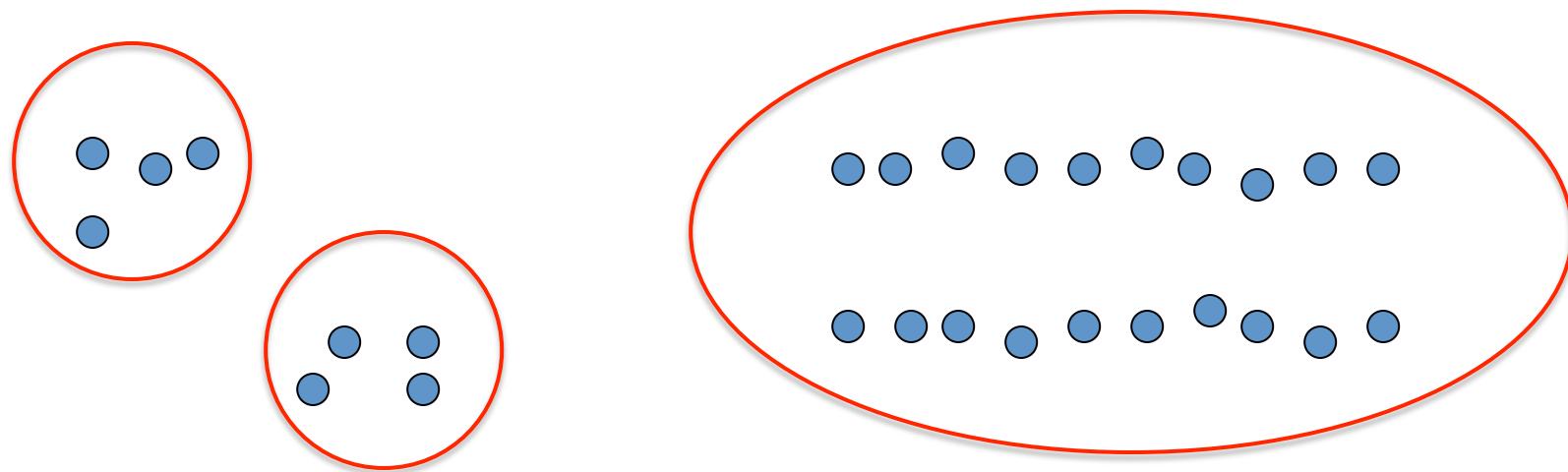
Clustering:

- Unsupervised learning
- Requires data, but no labels
- Detect patterns e.g. in
 - Group emails or search results
 - Customer shopping patterns
 - Regions of images
- Useful when don't know what you're looking for
- But: can get gibberish



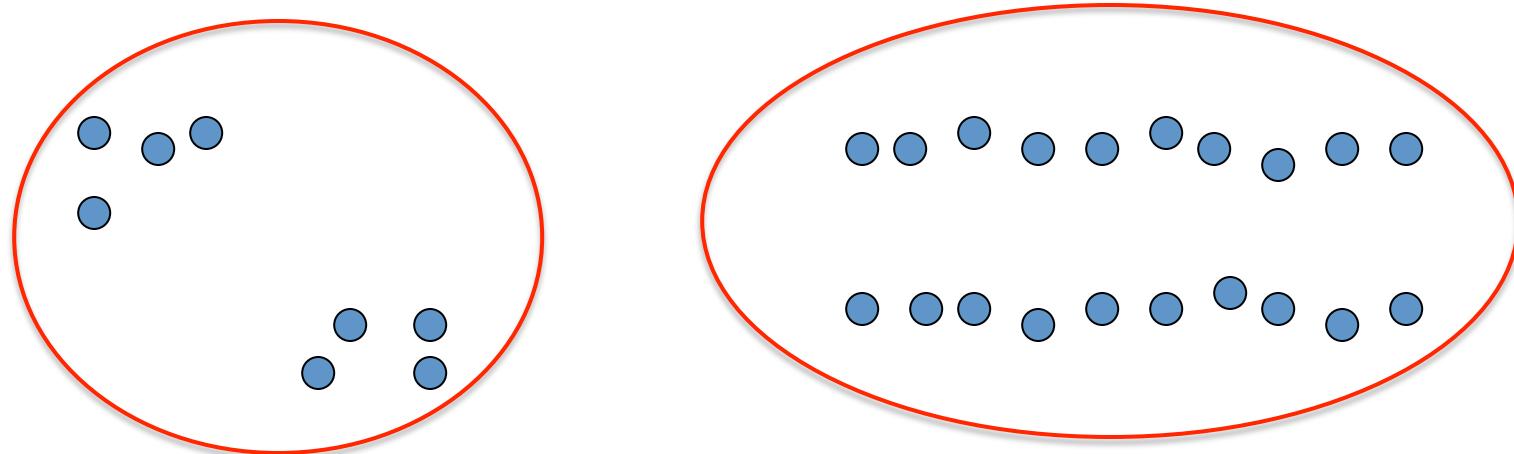
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



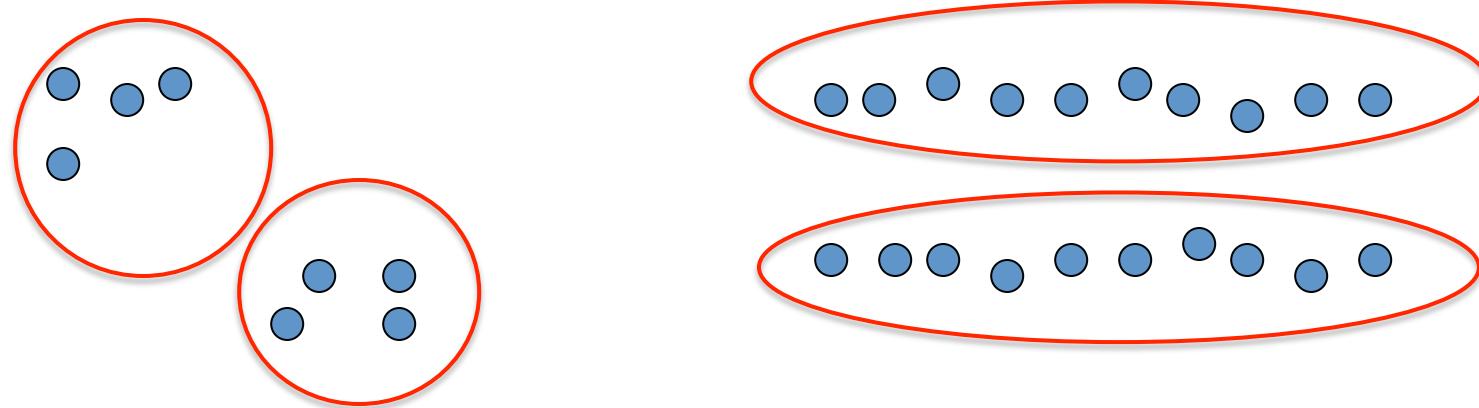
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

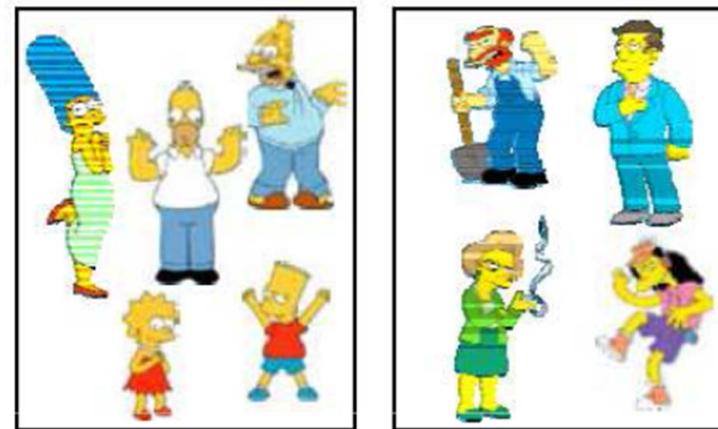
- Basic idea: group together similar instances
- Example: 2D point patterns



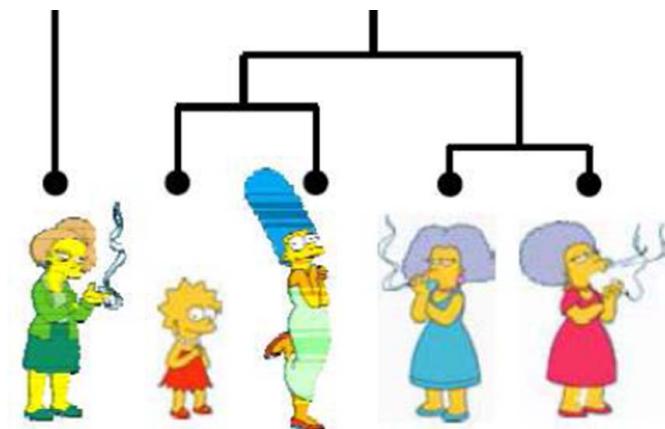
- What could “similar” mean?
 - One option: small Euclidean distance (squared)
$$\text{dist}(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|_2^2$$
 - Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered

Clustering algorithms

- Partition algorithms (Flat)
 - K-means
 - Mixture of Gaussian
 - Spectral Clustering



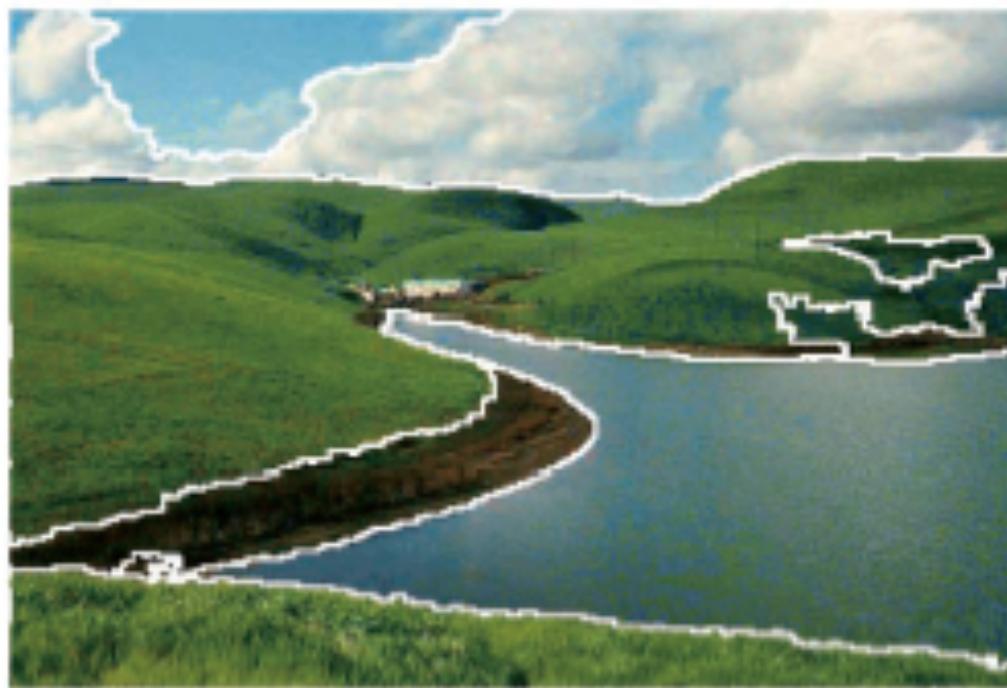
- Hierarchical algorithms
 - Bottom up – agglomerative
 - Top down – divisive



Clustering examples

Image segmentation

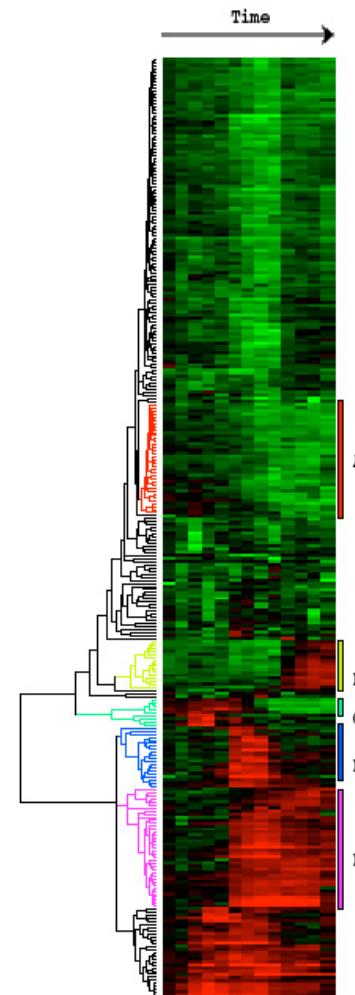
Goal: Break up the image into meaningful or perceptually similar regions



[Slide from James Hayes]

Clustering examples

Clustering gene expression data



Eisen et al, PNAS 1998

Cluster news articles

Clustering examples

Google News

News U.S. edition Classic

Top Stories

Boston Red Sox
Apple Inc.
Angela Merkel
Nokia Lumia
Bashar al-Assad
Republican Party
Facebook
Pets
Katy Perry
Bushfires in Australia
New York, New York

Recommended

U.S.
World
Sci/Tech
Business

More Top Stories

Health
Spotlight
Elections
Entertainment
Sports
Technology
Science

Top Stories

Teen suspect saw movie moments after allegedly killing beloved Massachusetts ...

Fox News - 8 minutes ago The 14-year-old student who authorities say murdered a beloved math teacher at a Massachusetts high school admitted to police that he slashed her throat with a box cutter, a source told MyFoxBoston.

Colleen Ritzer, slain Danvers High School teacher, remembered as passionate ...

CBS News
14-Year-Old Charged in Brutal Murder of Massachusetts Teacher

New York Magazine

Boston.com
Opinion: **Heslam: Heartbroken friends say Colleen was born to teach** Boston Herald

In Depth: Student, 14, arraigned in murder of Mass. teacher USA TODAY

Wikipedia: **Danvers, Massachusetts**

[See realtime coverage »](#)

Obamacare contractors tell their stories at congressional hearing

CNN - 40 minutes ago Washington (CNN) -- [Breaking news update at 10:09 a.m.] [URGENT - Congress-Obamacare-Testing]. (CNN) -- A contractor on the problem-plagued government website for President Barack Obama's signature health care reforms said Thursday his ...

Hearing on health care website today to focus on blame

WXIA-TV
Contractors Point Fingers Over Health-Law Website AllThingsD

[See realtime coverage »](#)

EU leaders meet amid concern about US spying claims

CNN - 1 hour ago (CNN) -- European Union leaders are meeting Thursday in Brussels for a summit that may be overshadowed by anger about allegations that the United States has been spying on its European allies.

Germany summons US ambassador over spying claims

USA TODAY
Germany Summons US Envoy Over Alleged NSA Spying

ABC News

Highly Cited: Readout of the President's Phone Call with Chancellor Merkel of Germany

Whitehouse.gov (press release)
From Germany: Press Review: Outrage over NSA eavesdropping Deutsche Welle

Opinion: **The Handyüberwachung Disaster**

New York Times

In Depth: **US ambassador to Germany summoned in Merkel mobile row**

BBC News

[See realtime coverage »](#)

US jobless claims miss forecasts, trade deficit widens slightly

Reuters - 59 minutes ago WASHINGTON | Thu Oct 24, 2013 9:19am EDT. WASHINGTON (Reuters) - The number of Americans filing new claims for unemployment benefits fell less than expected last week, but a lingering backlog of applications in California makes it difficult to get a ...

Weekly Jobless Claims Fall to 350000

Fox Business
How States Fared on Unemployment Benefit Claims

ABC News

In Depth: **More Americans Than Forecast Filed Jobless Claims**

Businessweek

[See realtime coverage »](#)

Kennedy cousin gets new trial in 1975 killing of neighbor; victim's mother ...


ABC News

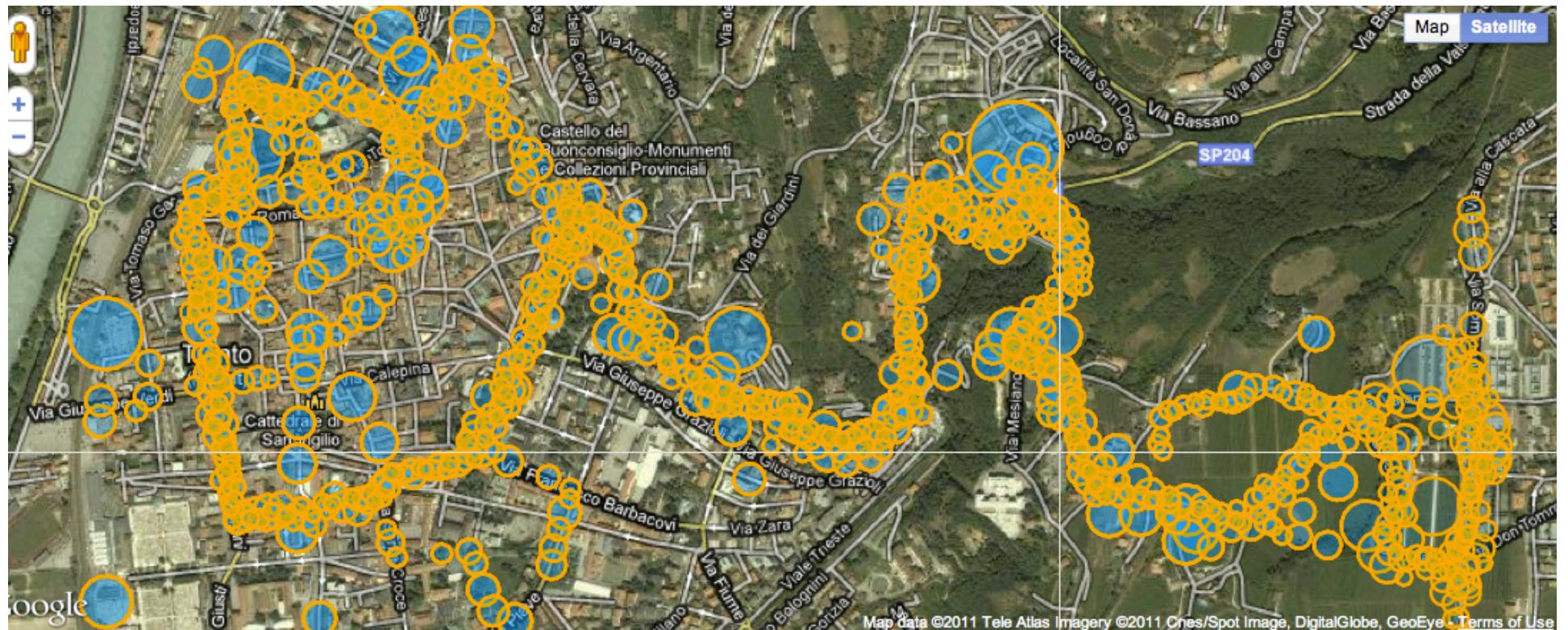

Wall Street Journal


National Post


The Olympian

Clustering examples

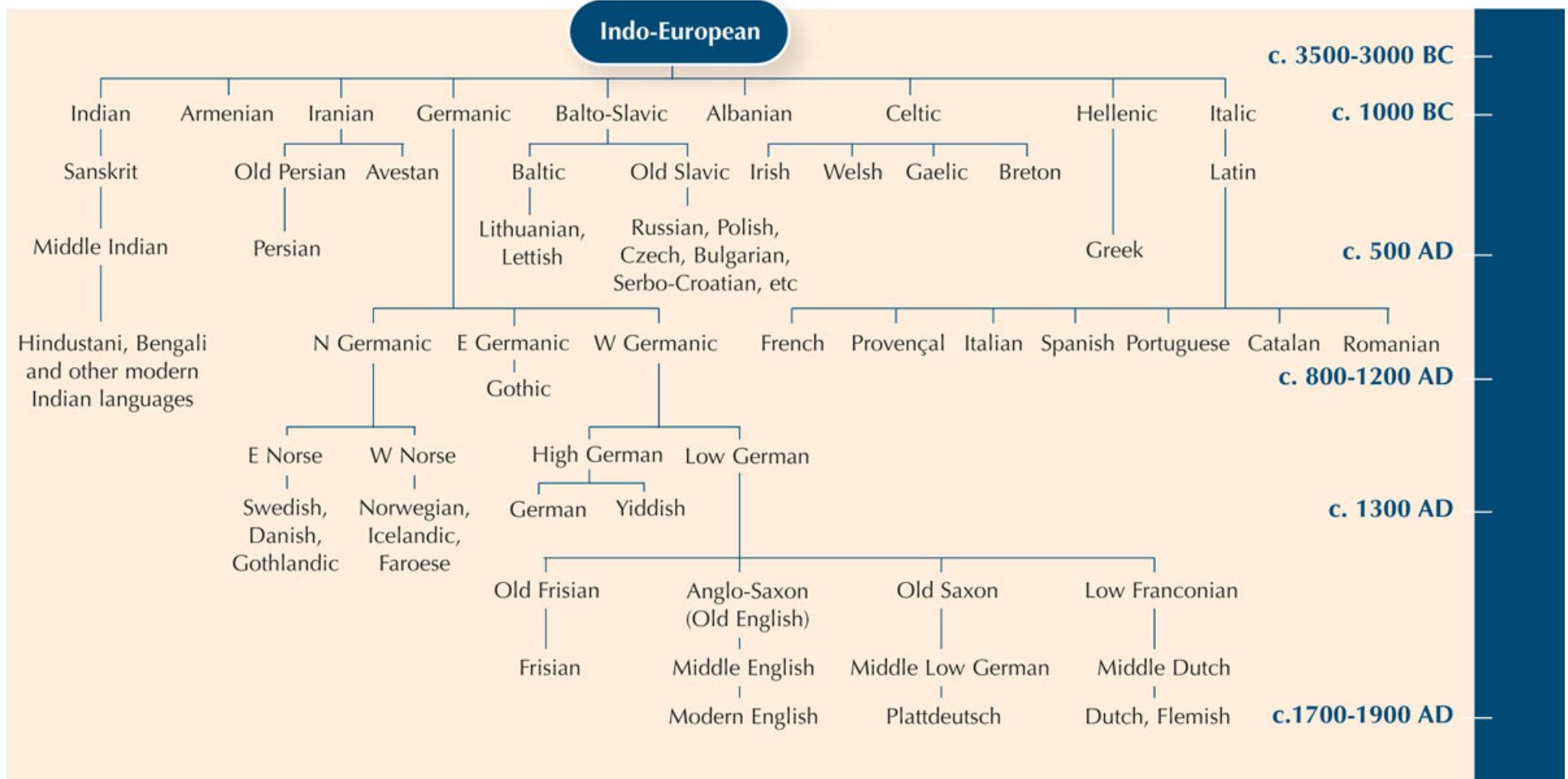
Cluster people by space and time



[Image from Pilho Kim]

Clustering examples

Clustering languages

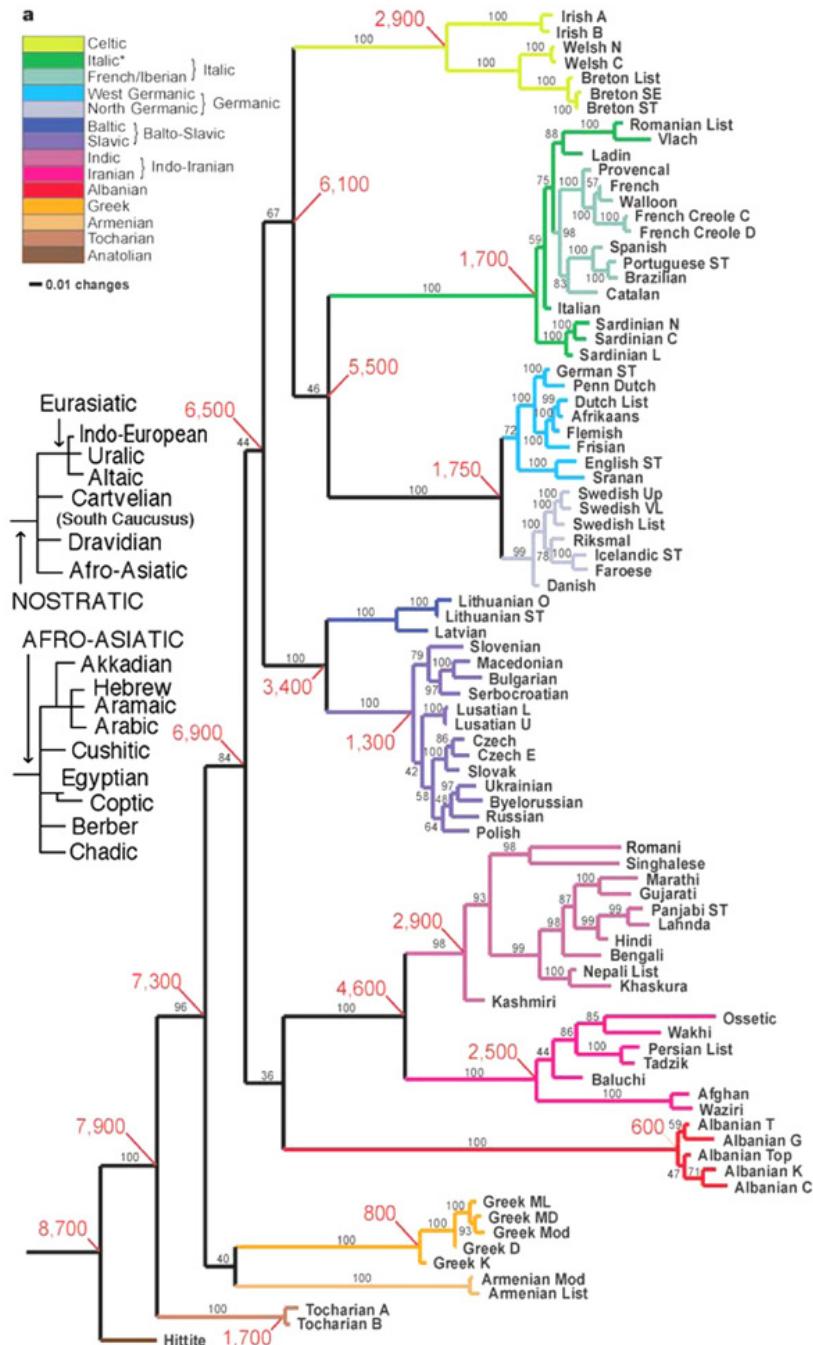


[Image from scienceinschool.org]

Clustering examples

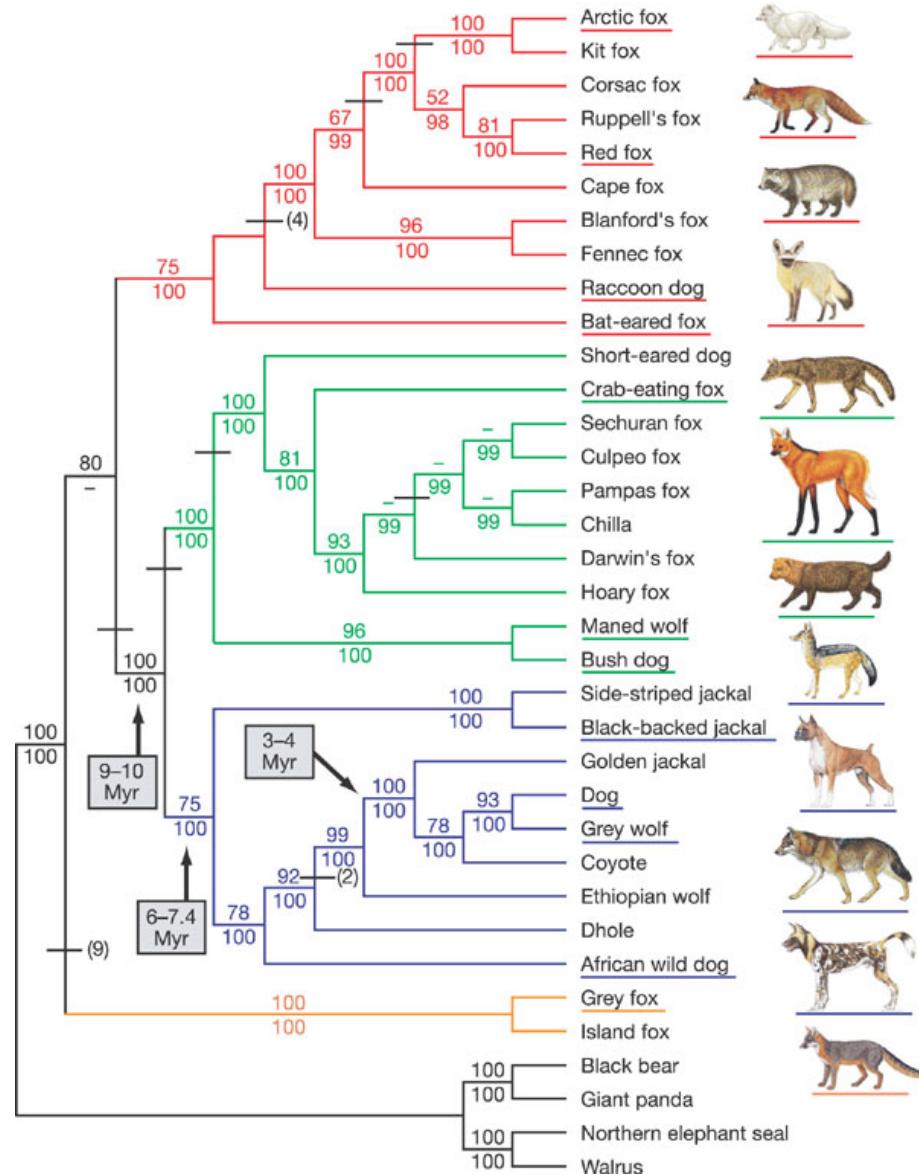
Clustering languages

[Image from dhushara.com]



Clustering examples

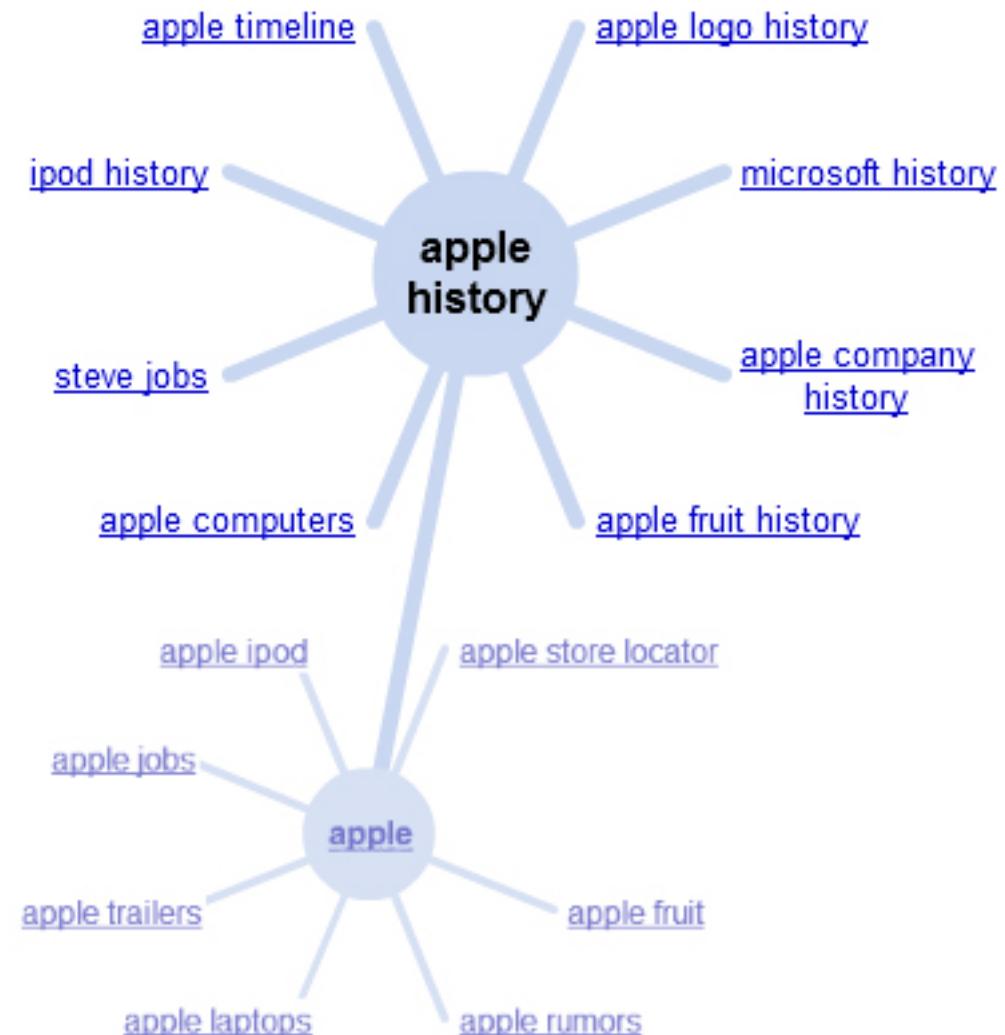
Clustering species
("phylogeny")



[Lindblad-Toh et al., Nature 2005]

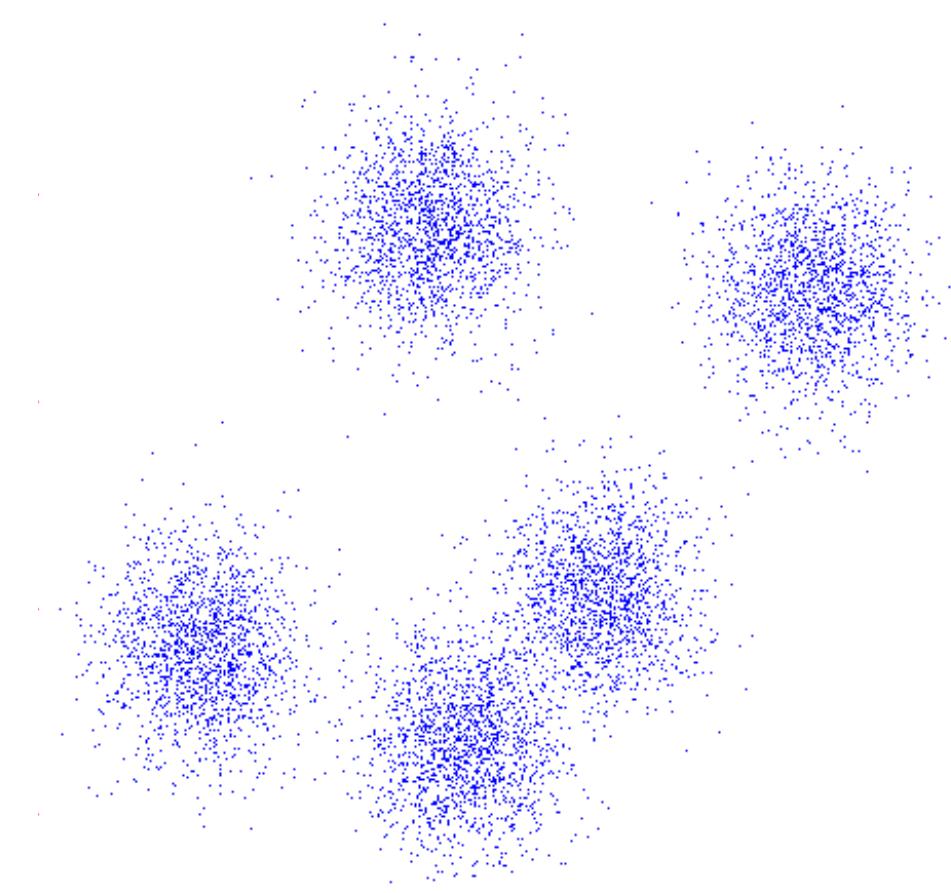
Clustering examples

Clustering search queries



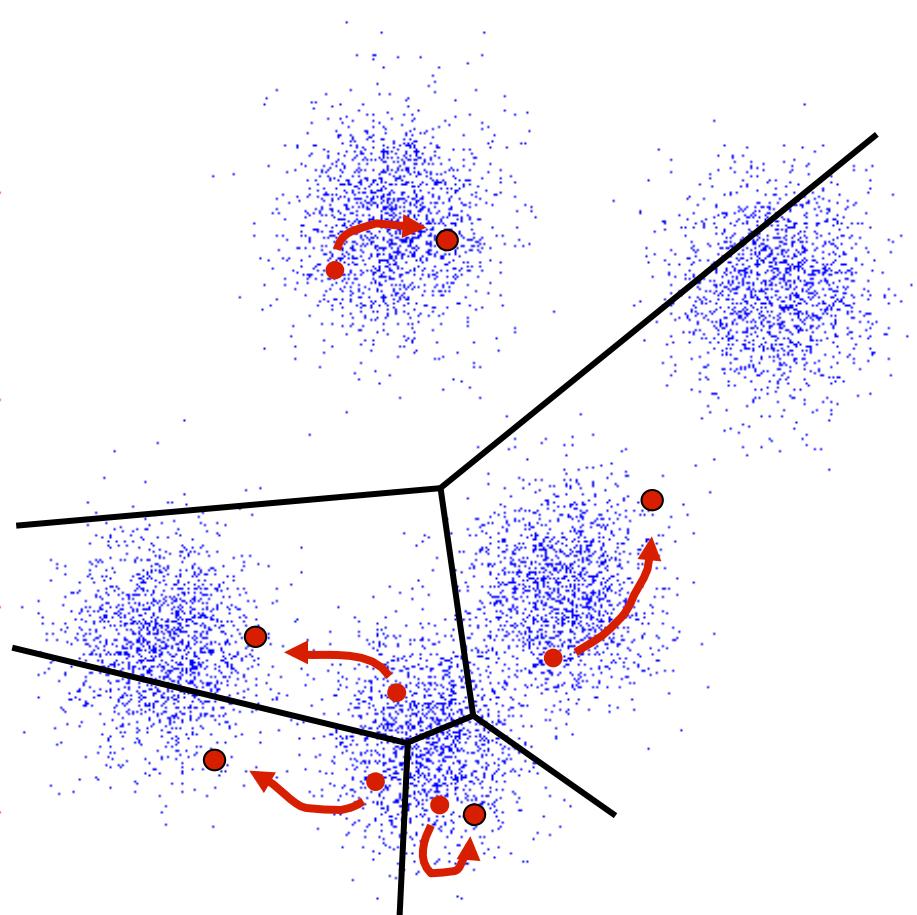
K-Means

- An iterative clustering algorithm
 - Initialize: Pick K random points as cluster centers
 - Alternate:
 1. Assign data points to closest cluster center
 2. Change the cluster center to the average of its assigned points
 - Stop when no points' assignments change

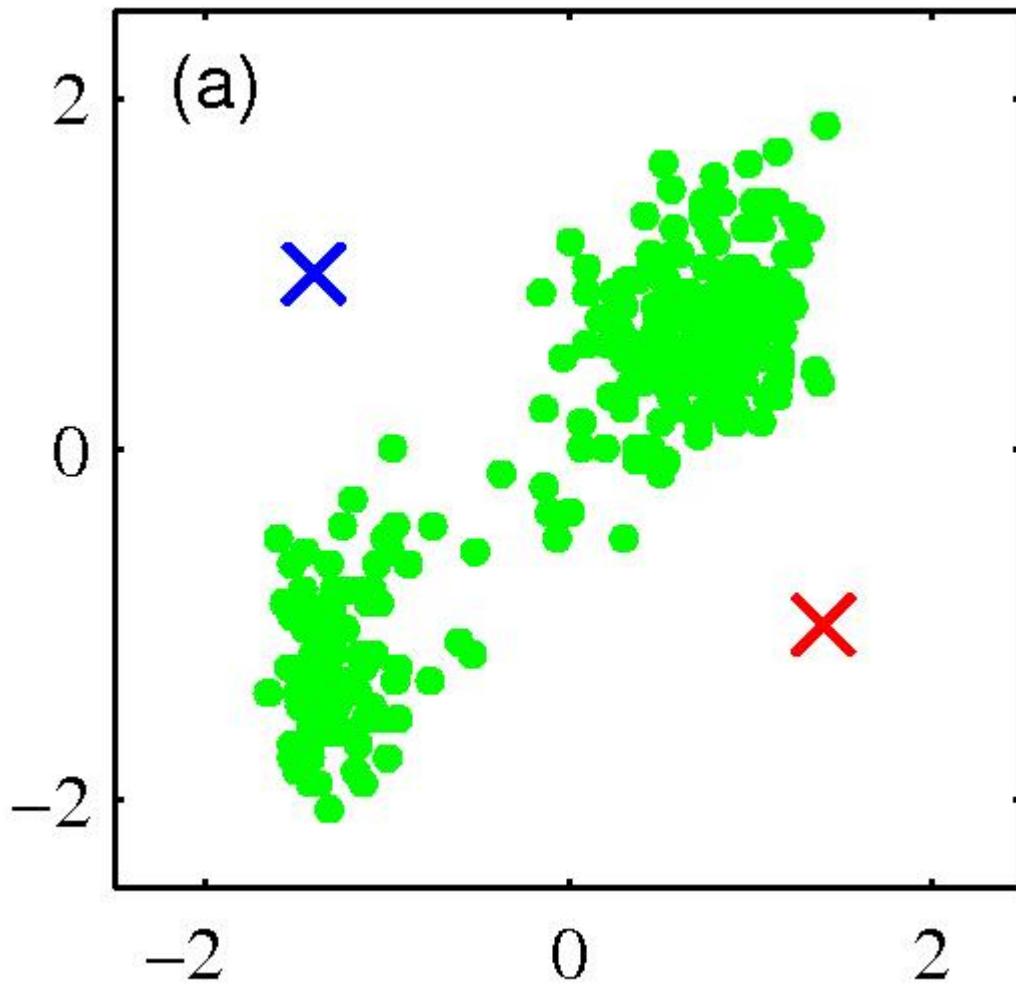


K-Means

- An iterative clustering algorithm
 - Initialize: Pick K random points as cluster centers
 - Alternate:
 1. Assign data points to closest cluster center
 2. Change the cluster center to the average of its assigned points
 - Stop when no points' assignments change



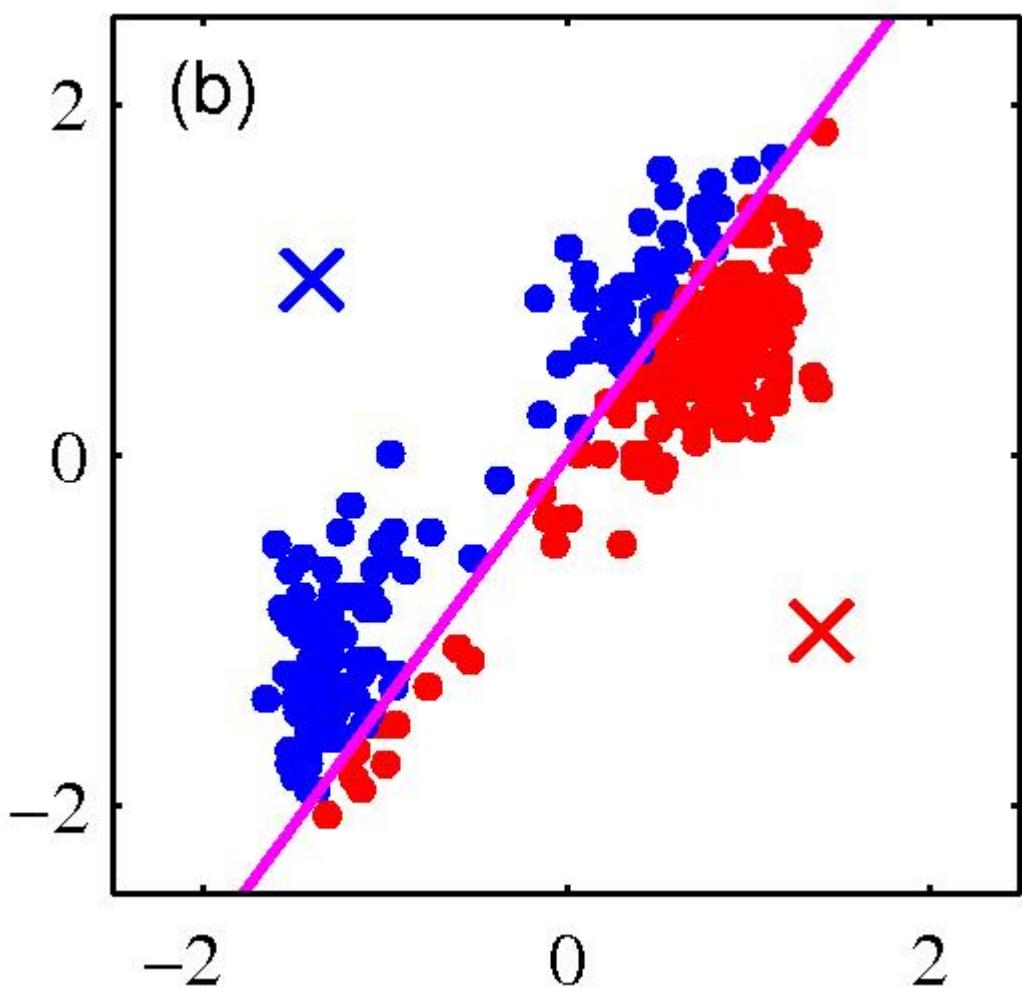
K-means clustering: Example



- Pick K random points as cluster centers (means)

Shown here for $K=2$

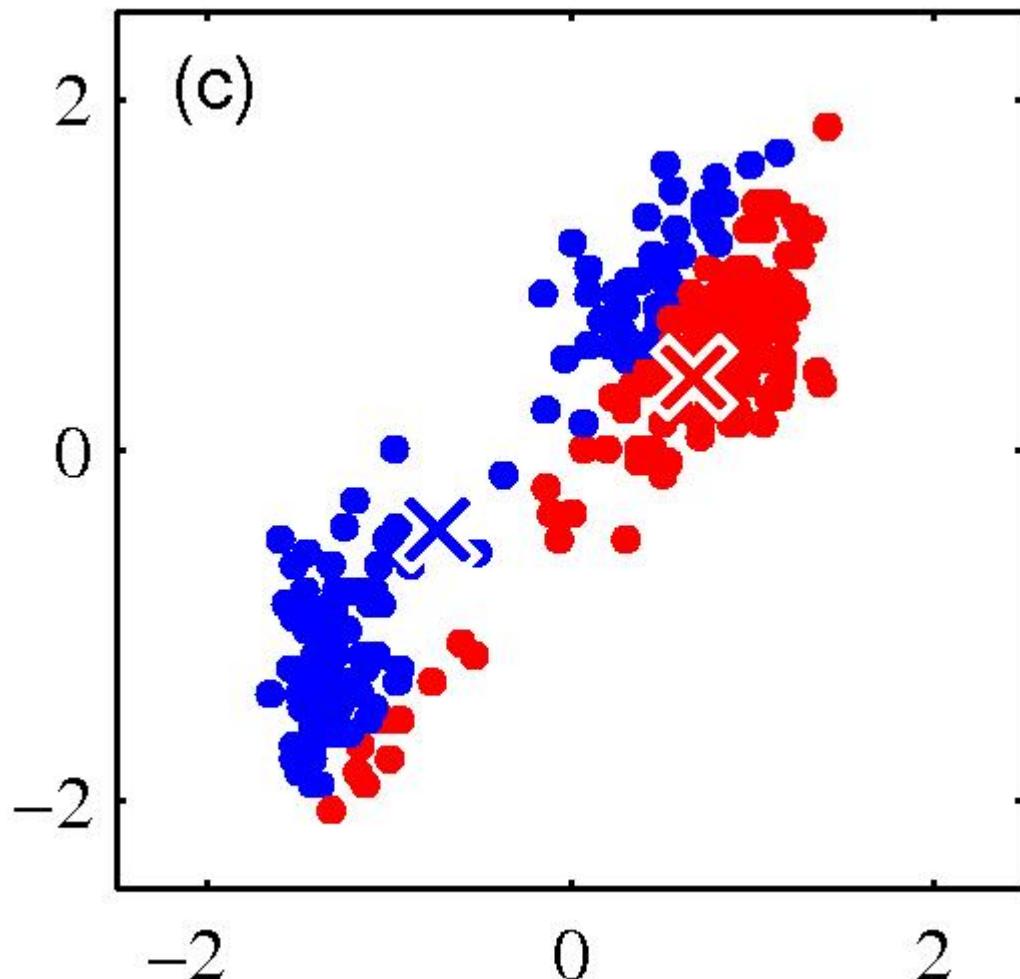
K-means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

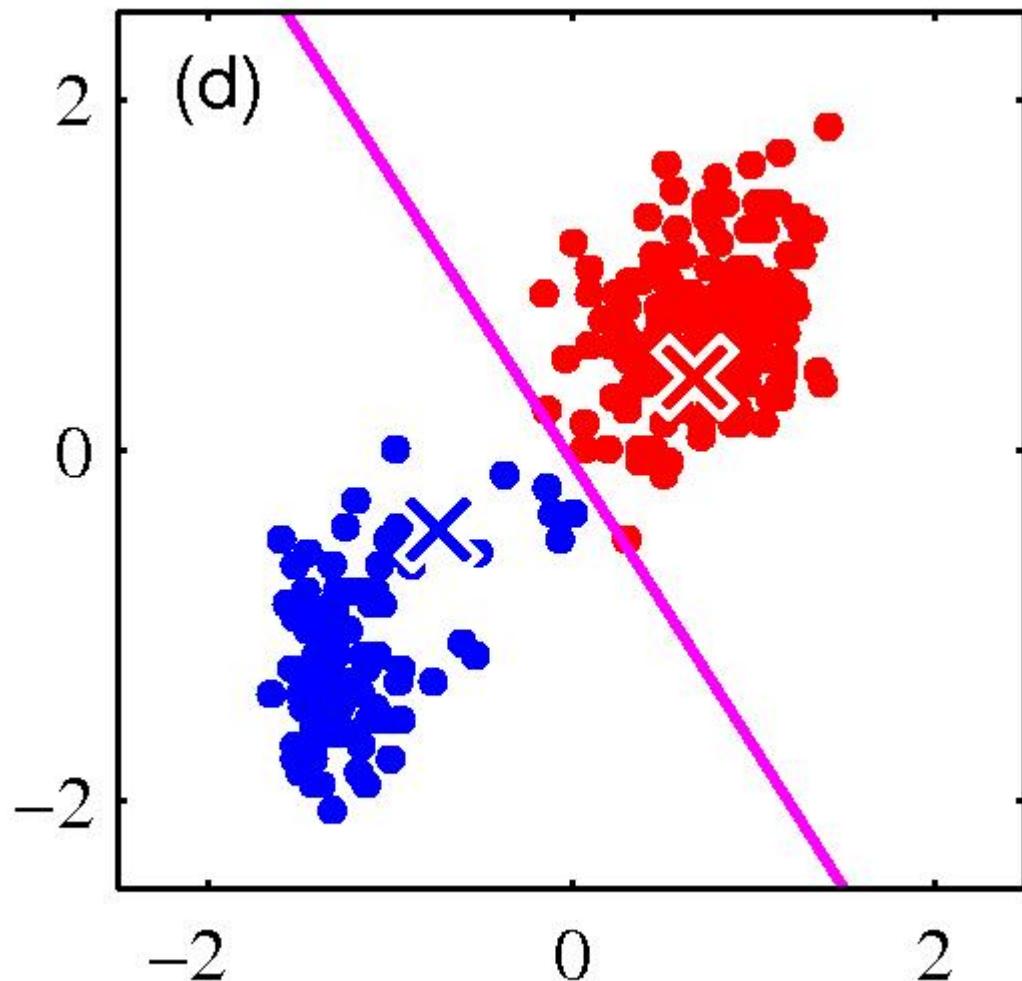
K-means clustering: Example



Iterative Step 2

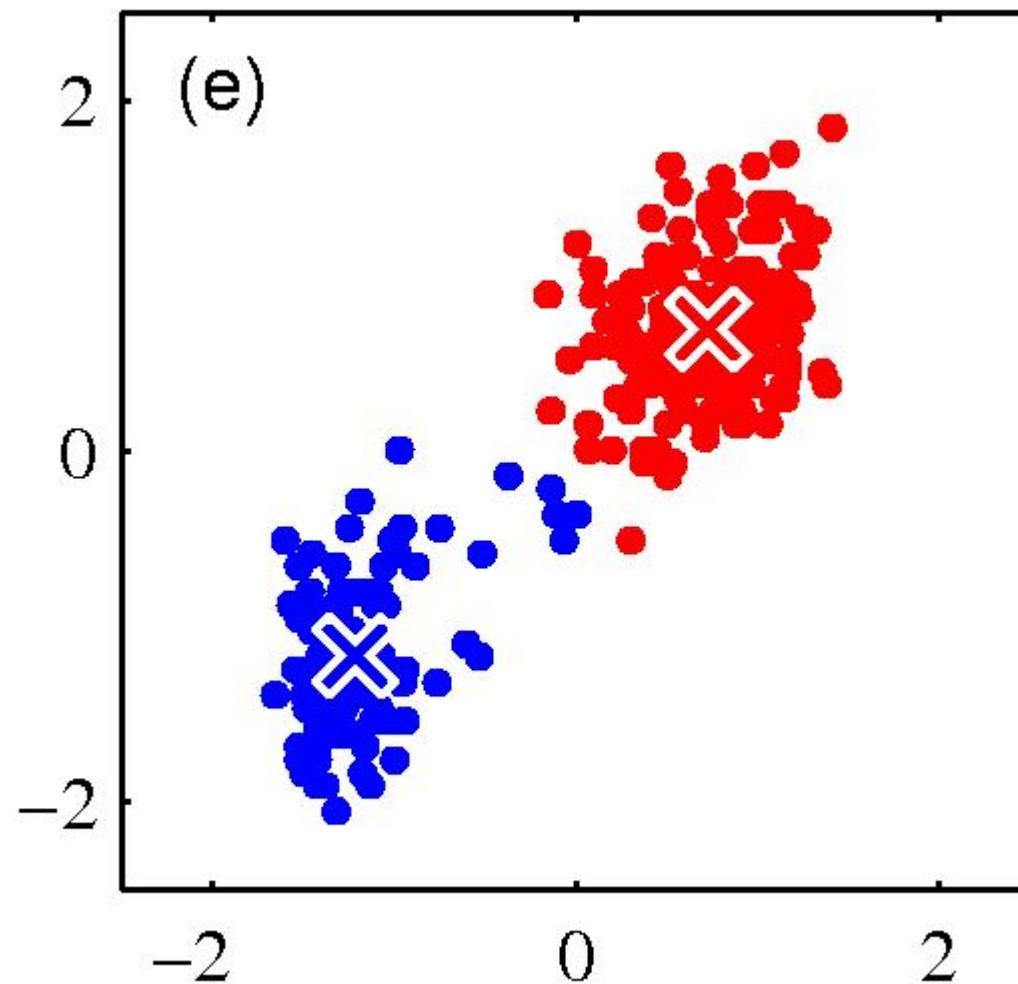
- Change the cluster center to the average of the assigned points

K-means clustering: Example

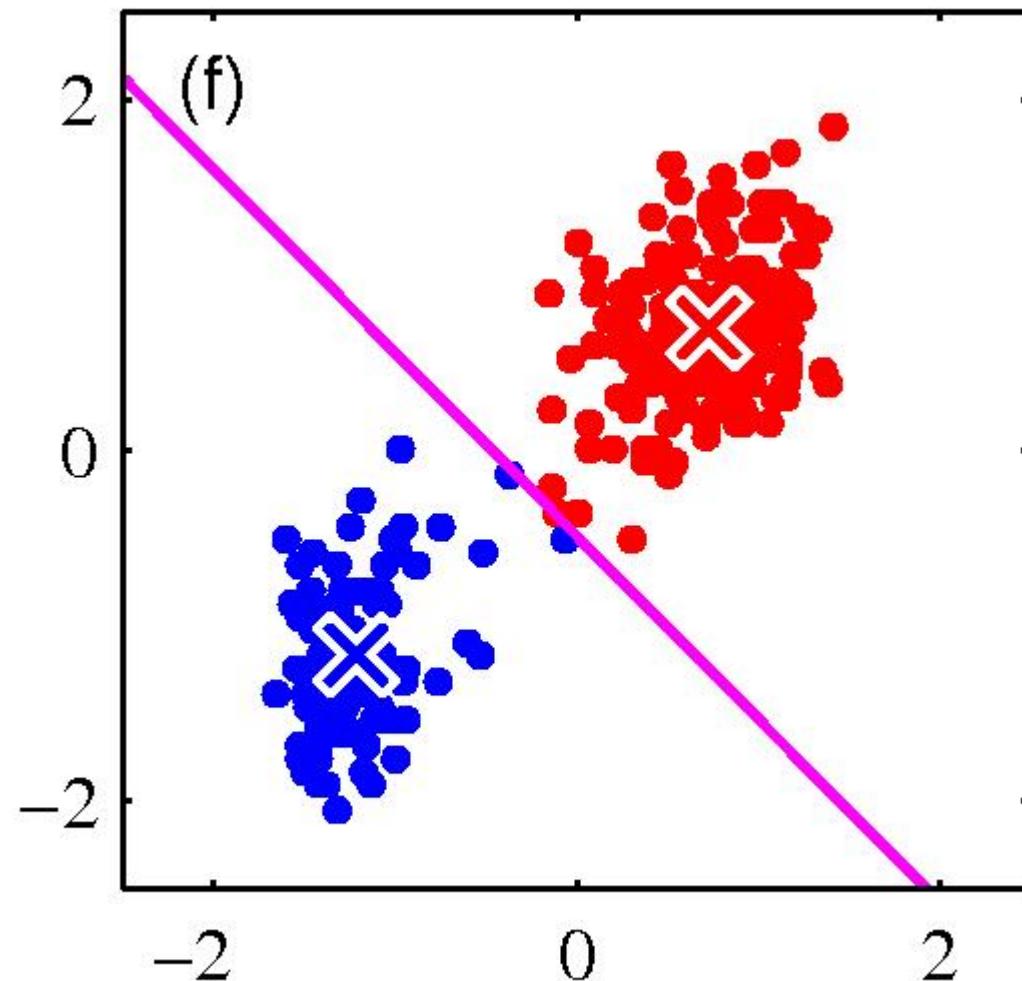


- Repeat until convergence

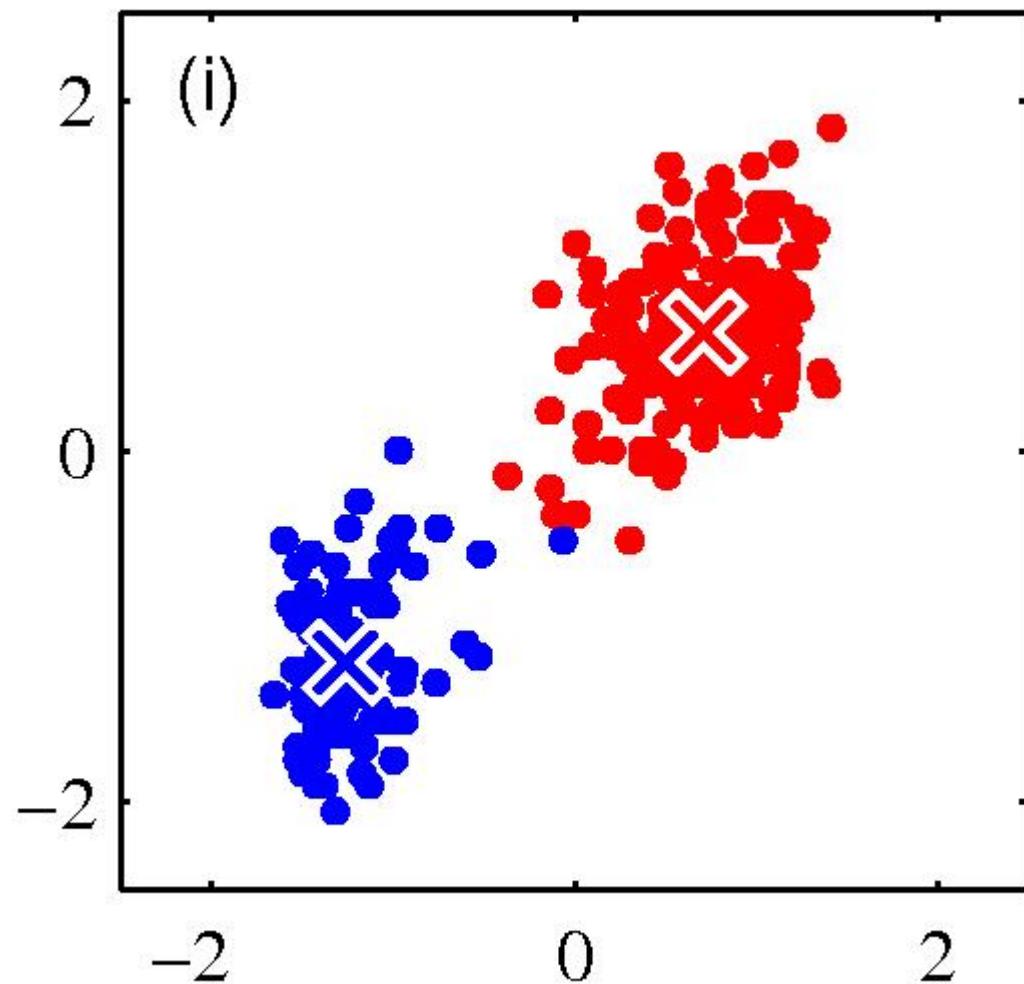
K-means clustering: Example



K-means clustering: Example



K-means clustering: Example



Properties of K-means algorithm

- Guaranteed to converge in a finite number of iterations
- Running time per iteration:
 1. Assign data points to closest cluster center
 $O(KN)$ time
 2. Change the cluster center to the average of its assigned points
 $O(N)$

Example: K-Means for Segmentation

K=2



Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Original



Example: K-Means for Segmentation

K=2



K=3



Original



Example: K-Means for Segmentation

K=2



K=3



K=10

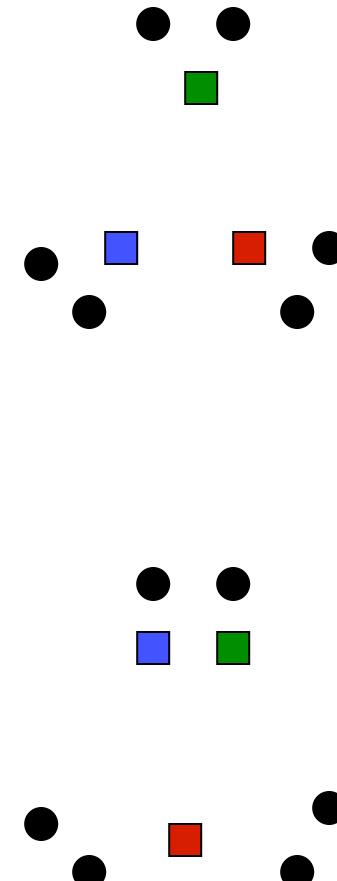


Original



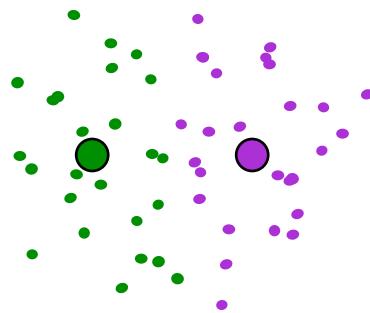
Initialization

- K-means **algorithm** is a heuristic
 - Requires initial means
 - It does matter what you pick!
 - What can go wrong?
 - Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics

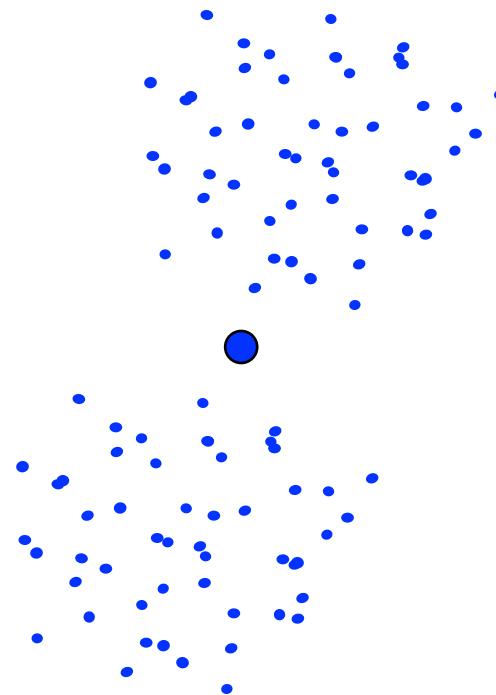


K-Means Getting Stuck

A local optimum:

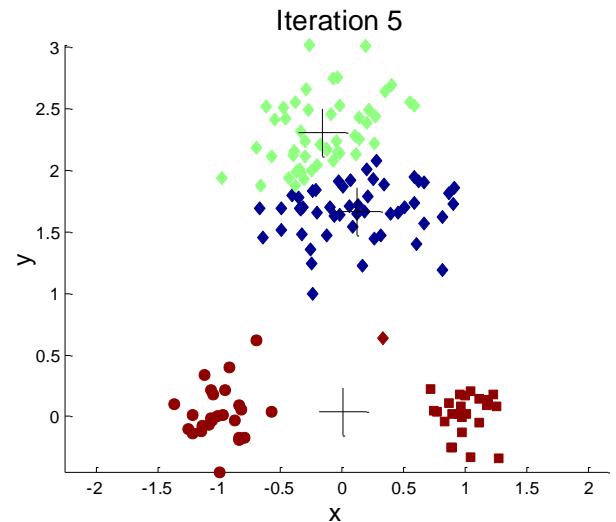
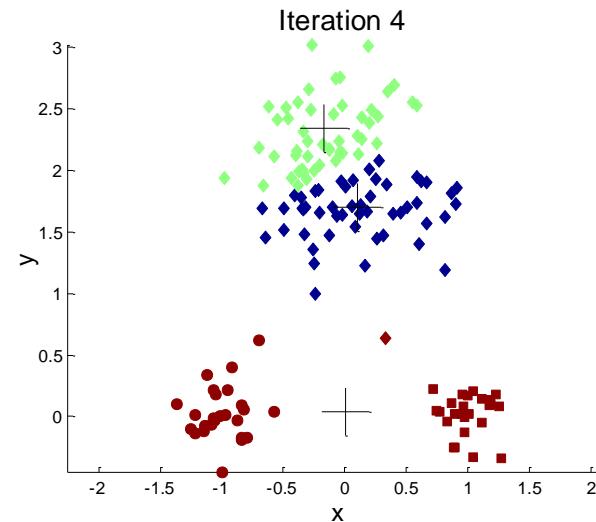
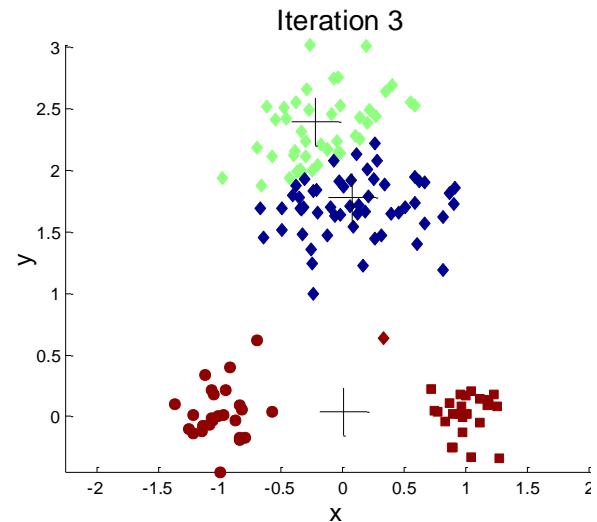
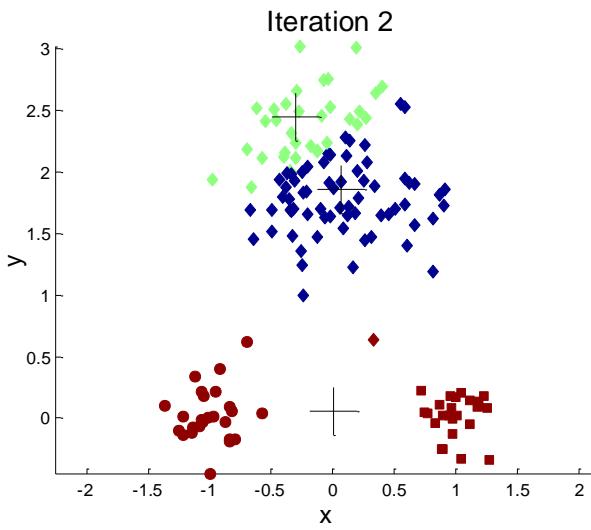
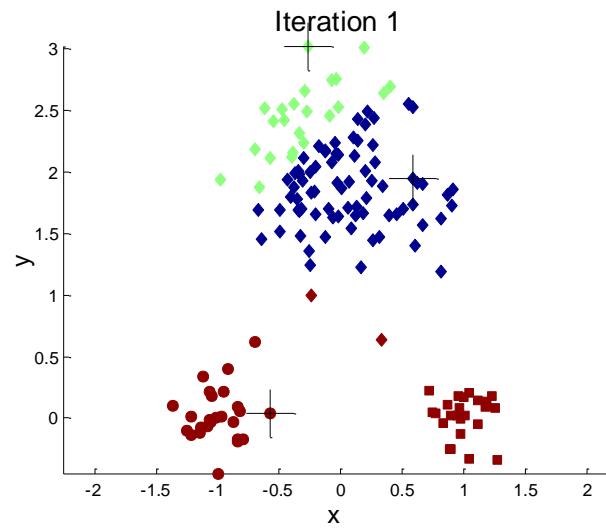


Would be better to have
one cluster here

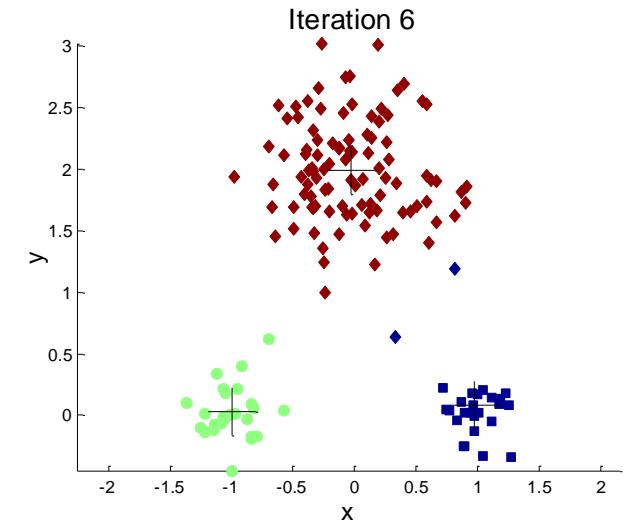
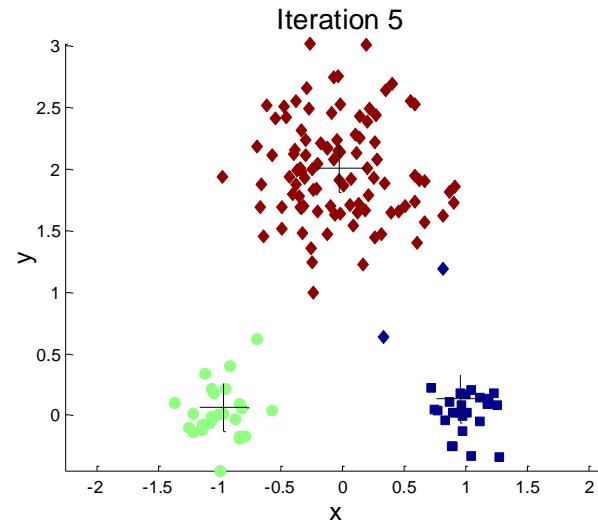
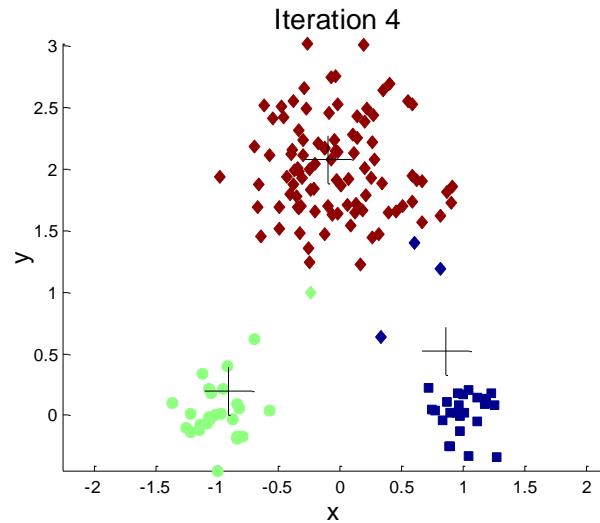
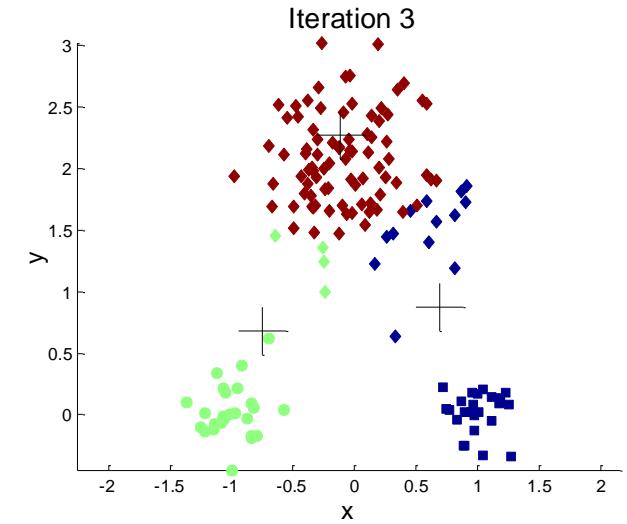
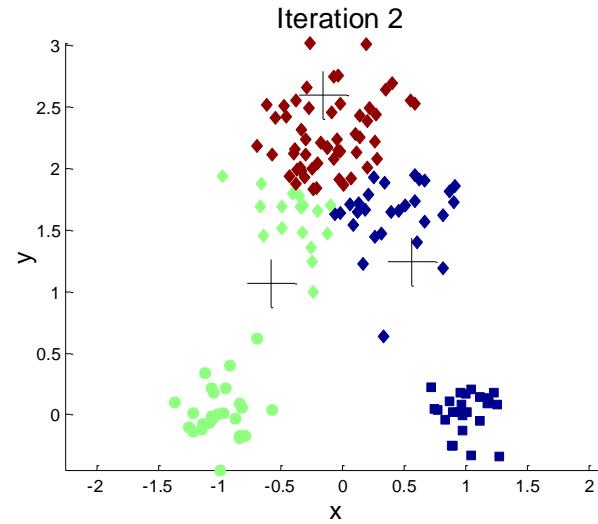
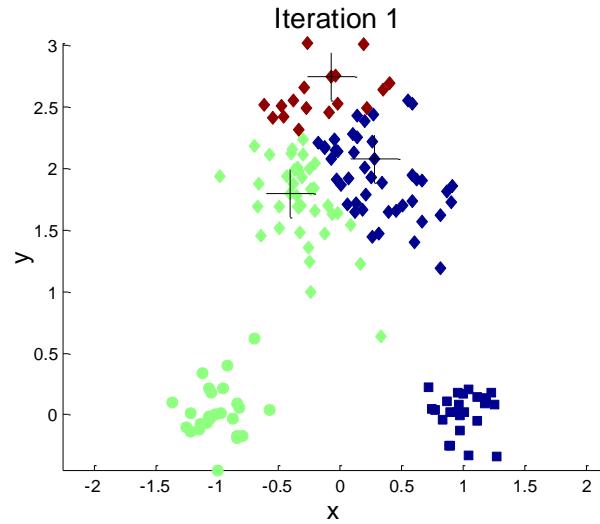


... and two clusters here

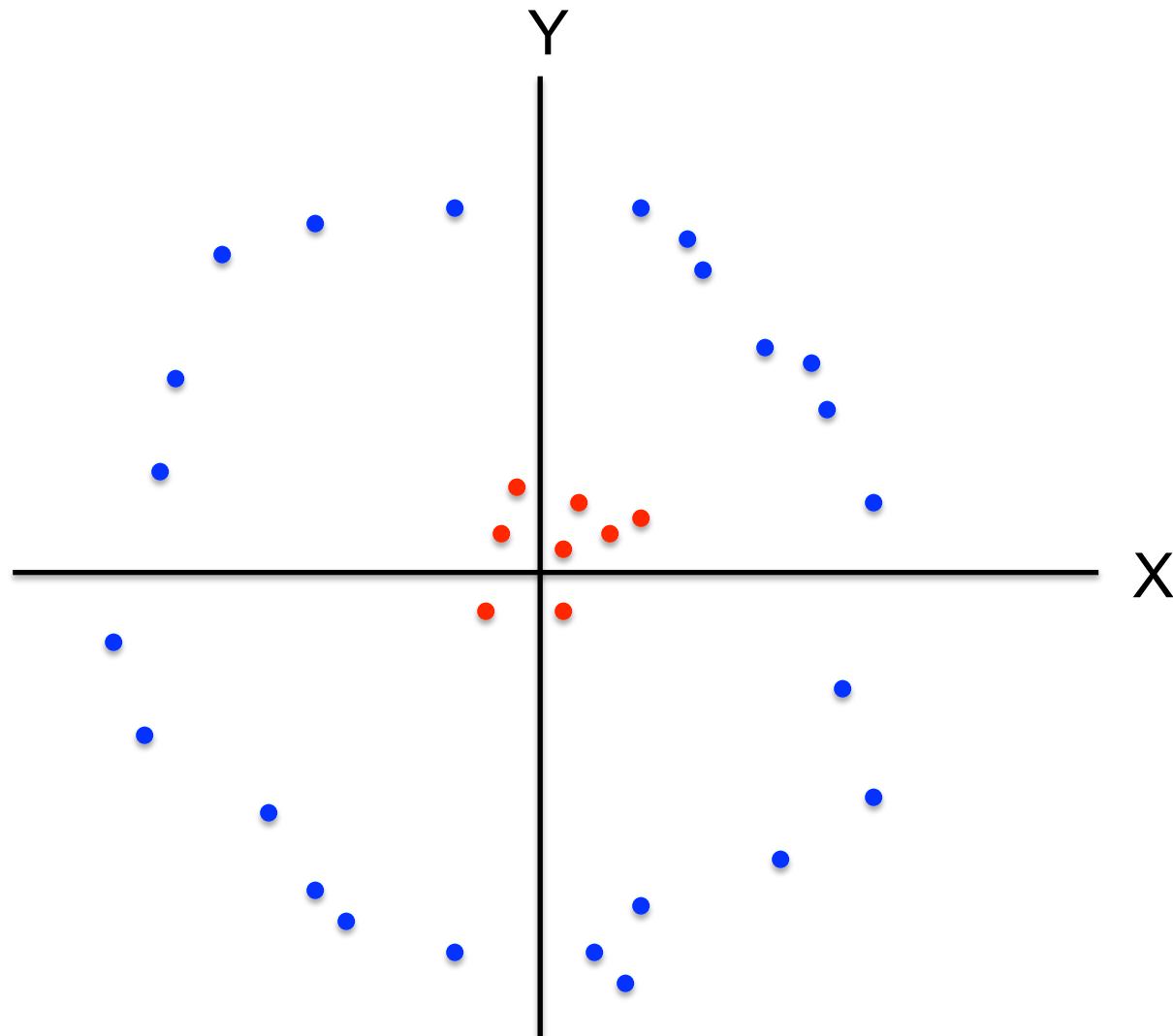
Example: bad initial centroids



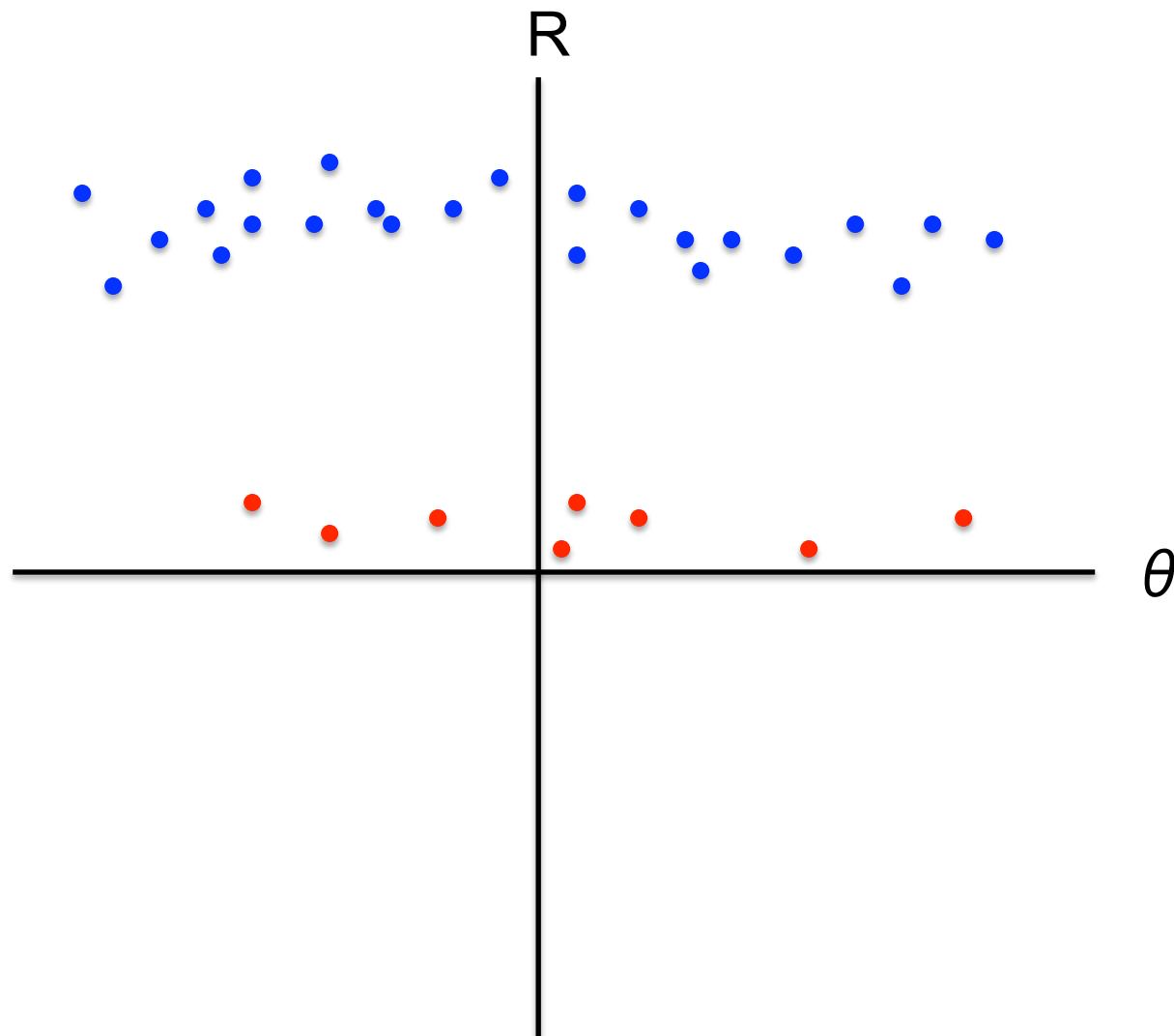
Example: good initial centroids



K-means not able to properly cluster

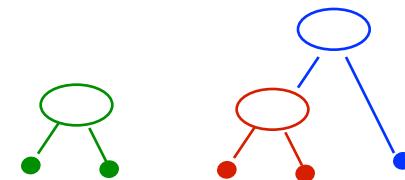
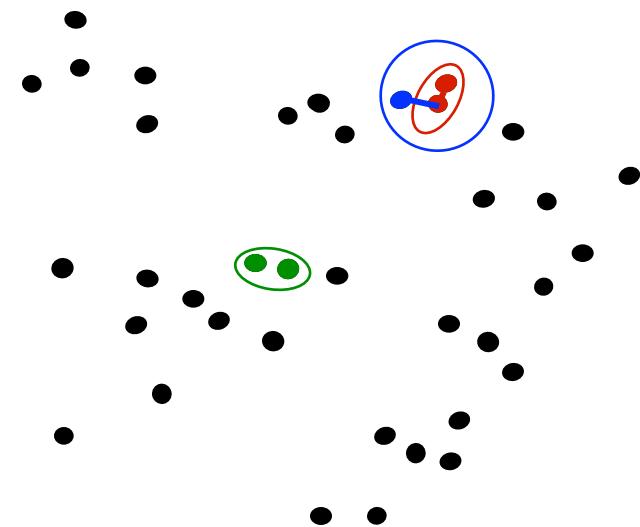


Changing the features (distance function)
can help



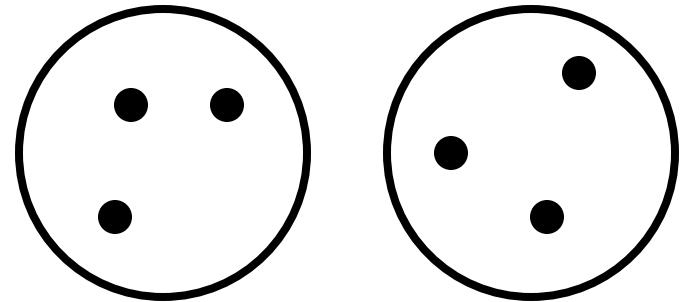
Agglomerative Clustering

- Agglomerative clustering:
 - First merge very similar instances
 - Incrementally build larger clusters out of smaller clusters
- Algorithm:
 - Maintain a set of clusters
 - Initially, each instance in its own cluster
 - Repeat:
 - Pick the two **closest** clusters
 - Merge them into a new cluster
 - Stop when there's only one cluster left
- Produces not one clustering, but a family of clusterings represented by a **dendrogram**



Agglomerative Clustering

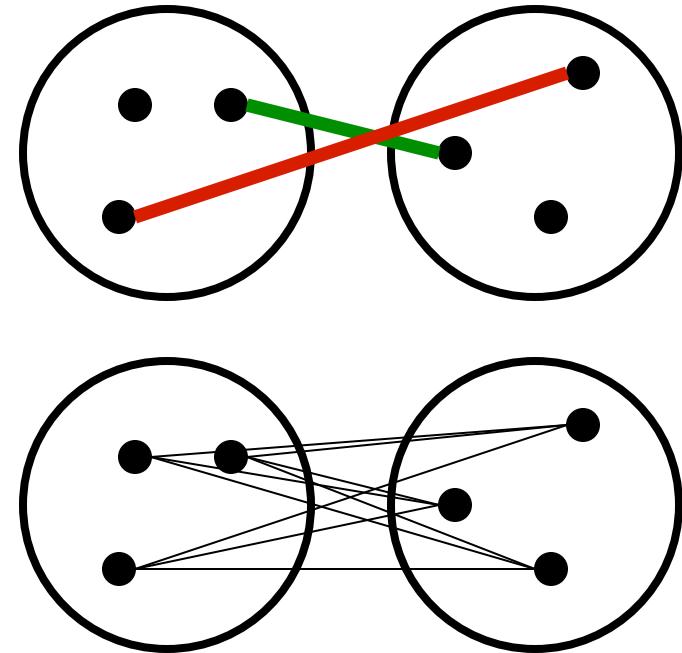
- How should we define “closest” for clusters with multiple elements?



Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?

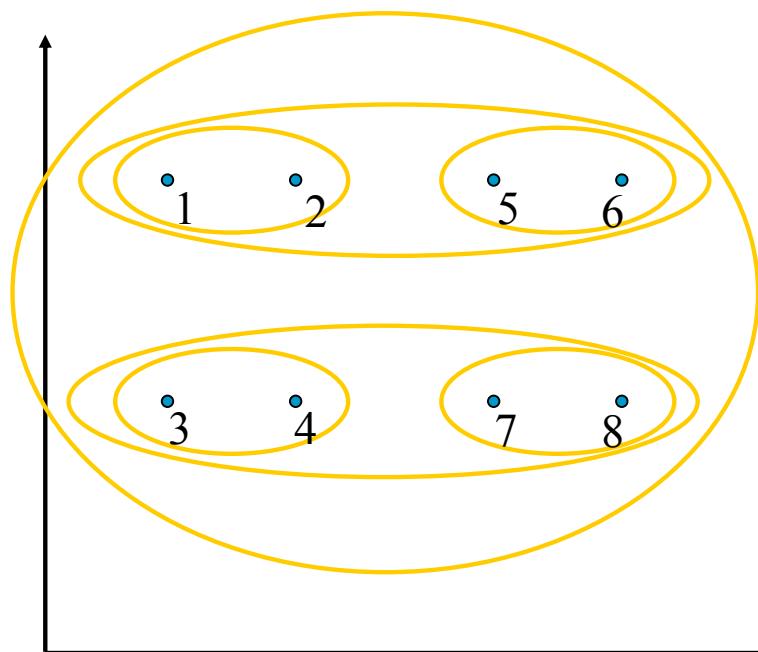
- Many options:
 - Closest pair
(single-link clustering)
 - Farthest pair
(complete-link clustering)
 - Average of all pairs
- Different choices create different clustering behaviors



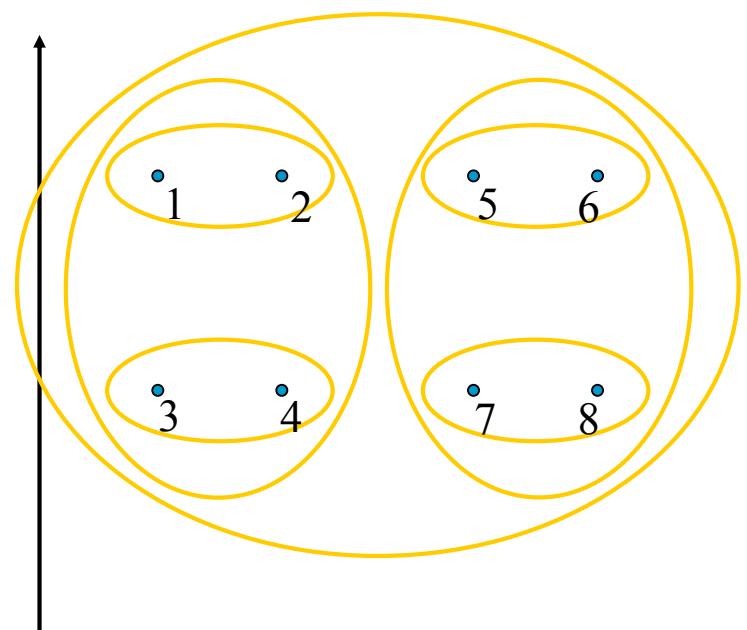
Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?

Closest pair
(single-link clustering)



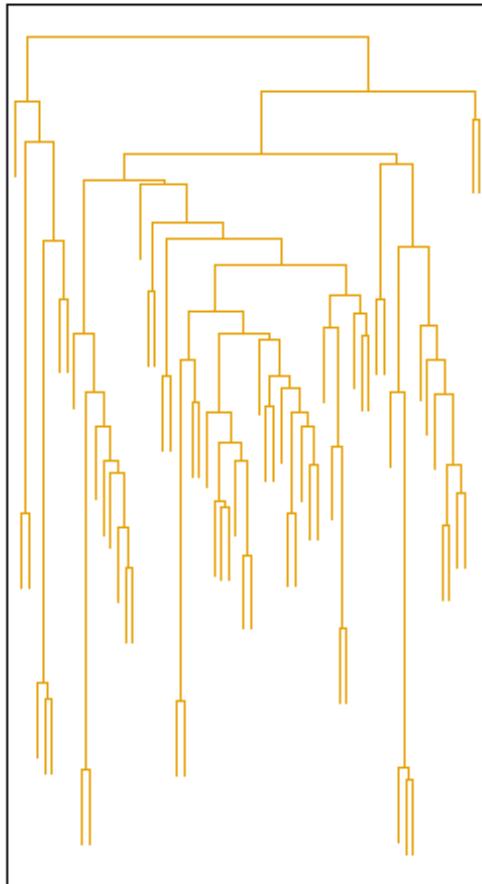
Farthest pair
(complete-link clustering)



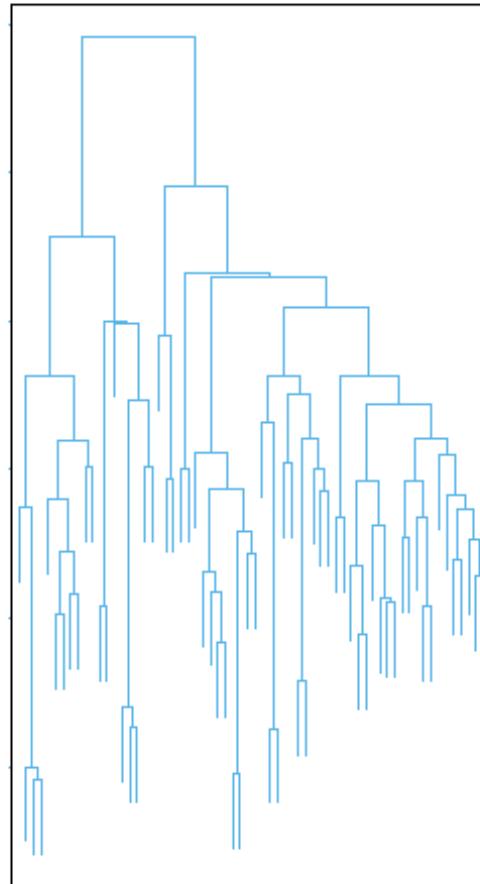
[Pictures from Thorsten Joachims]

Clustering Behavior

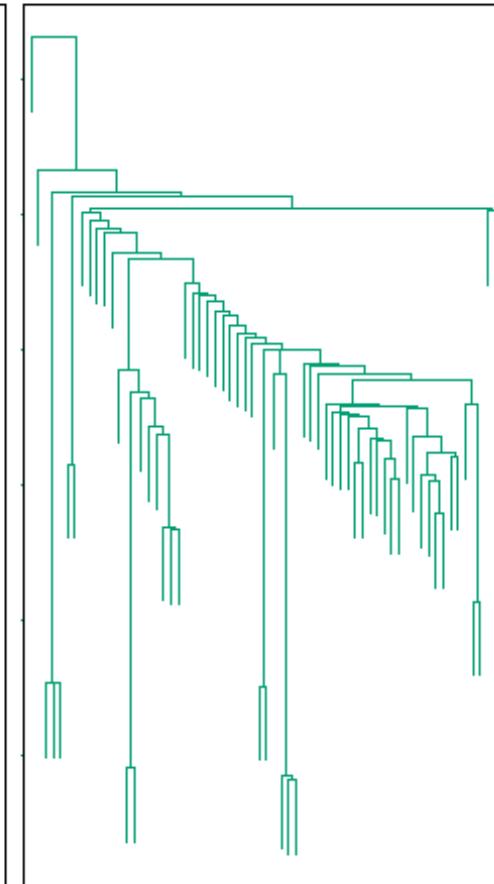
Average



Farthest



Nearest

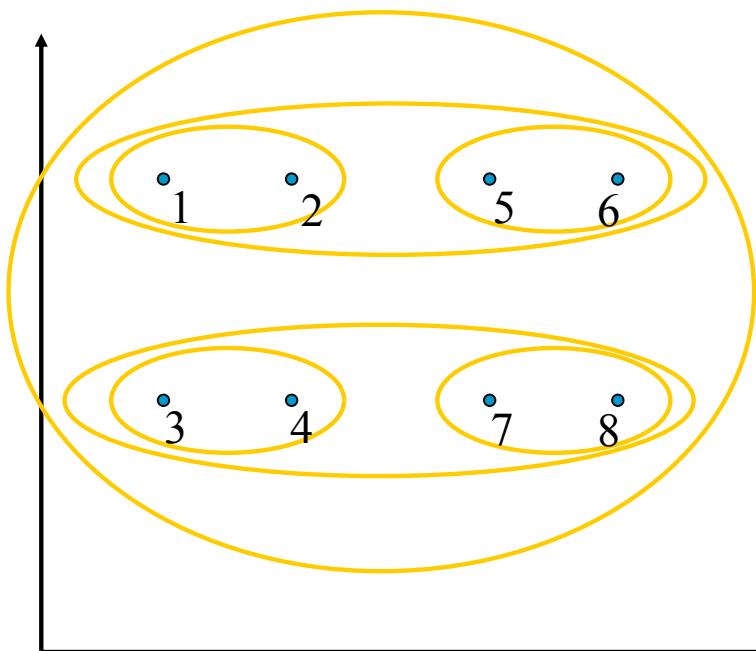


Mouse tumor data from [Hastie *et al.*]

Agglomerative Clustering

When can this be expected to work?

Closest pair
(single-link clustering)



Strong separation property:

All points are more similar to points in their own cluster than to any points in any other cluster

Then, the true clustering corresponds to some **pruning** of the tree obtained by single-link clustering!

Slightly weaker (stability) conditions are solved by average-link clustering

(Balcan et al., 2008)

Other cluster methods

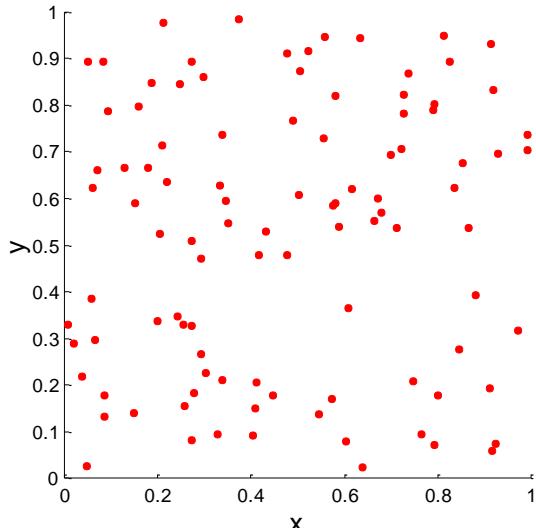
- DBSCAN
- Gaussian mixture model (GMM)
- Spectral clustering
- Biclustering
- Topic model
- Affinity propagation (Science-2007)
- DensityClust (Science-2014)

Cluster validation

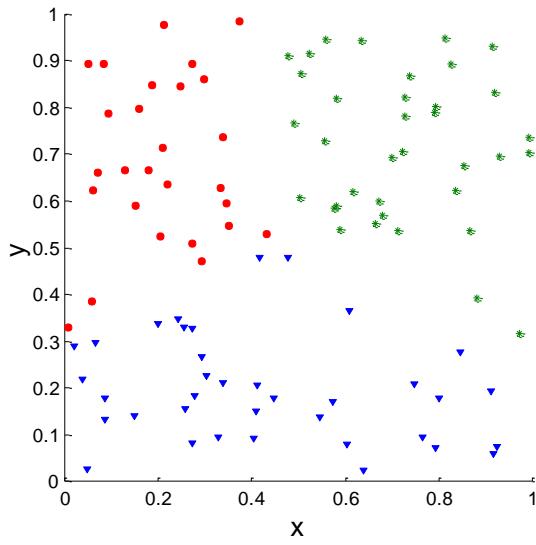
- For supervised classification we have a variety of measures to evaluate how good our model is
 - Accuracy, precision, recall
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
 - To avoid finding patterns in noise
 - To compare clustering algorithms
 - To compare two sets of clusters
 - To compare two clusters

Clusters found in random data

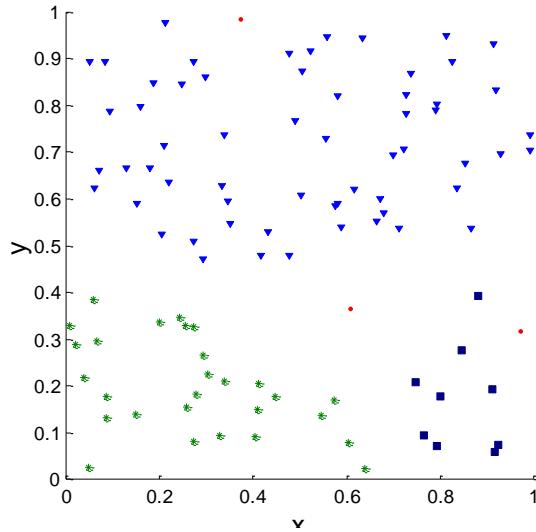
Random Points



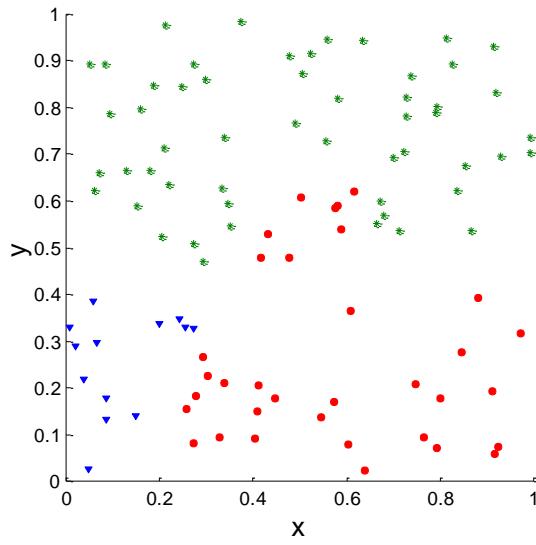
K-means



DBSCAN



Complete Link



Different aspects of cluster validation

- Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
- Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
- Evaluating how well the results of a cluster analysis fit the data without reference to external information
 - Use only the data
 - Determining the ‘correct’ number of clusters.

Measures of cluster validity

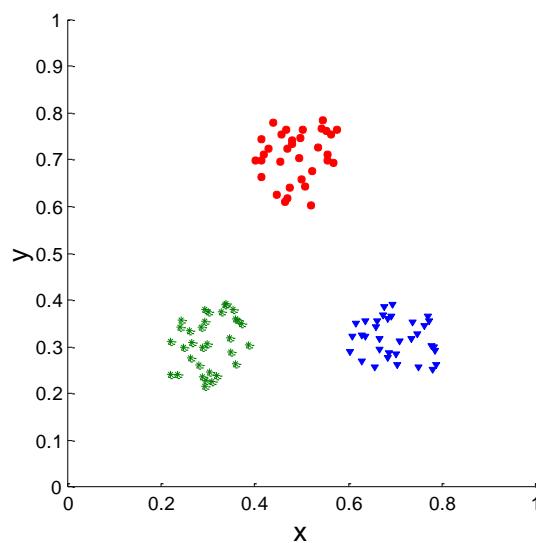
- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
 - **External Index:** Used to measure the extent to which cluster labels match externally supplied class labels.
 - Entropy
 - **Internal Index:** Used to measure the goodness of a clustering structure without respect to external information.
 - Sum of Squared Error (SSE)
 - **Relative Index:** Used to compare two different clusterings or clusters.
 - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as **criteria** instead of **indices**.
 - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Measures cluster validity via correlation

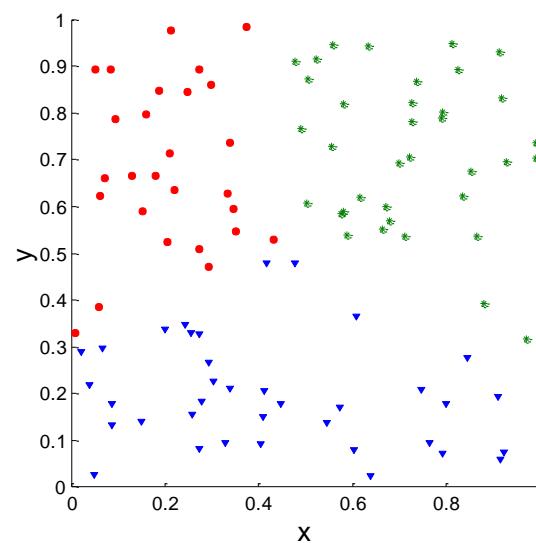
- Two matrices
 - Proximity matrix
 - “Incidence” matrix
 - One row and one column for each data point
 - An entry is 1 if the associated pair of points belong to the same cluster
 - An entry is 0 if the associated pair of points belong to different clusters
 - Computer the correlation between the two matrices
 - Since the matrices are symmetric, only the correlation between $n(n-1)/2$ entries needs to be calculated
- High correlation indicates that points that belong to the same cluster are close to each other.
- Not a good measure for some density or continuity based clusters

Measures cluster validity via correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two datasets



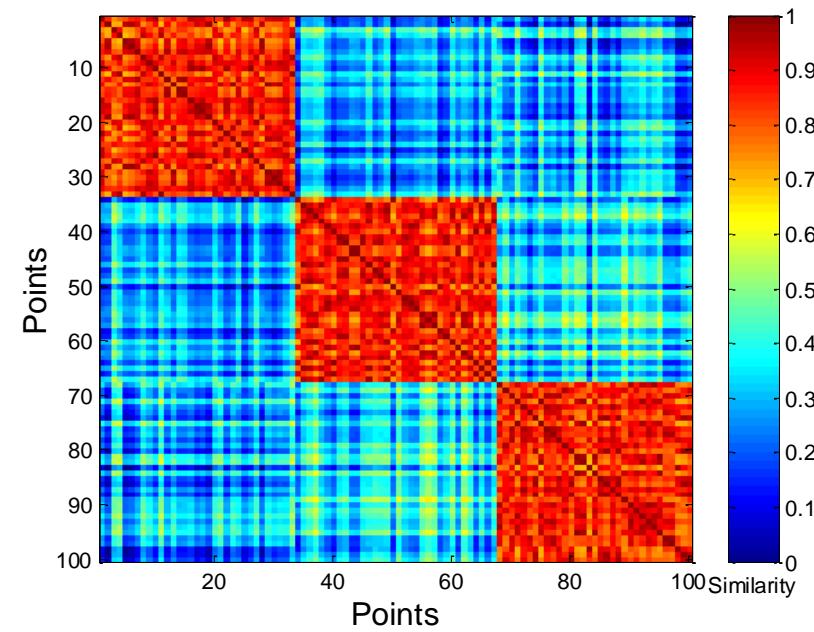
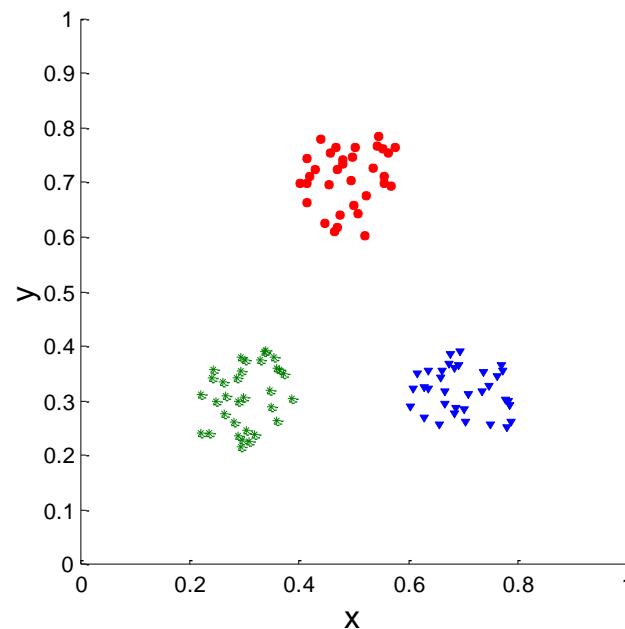
Corr = -0.9235



Corr = -0.5810

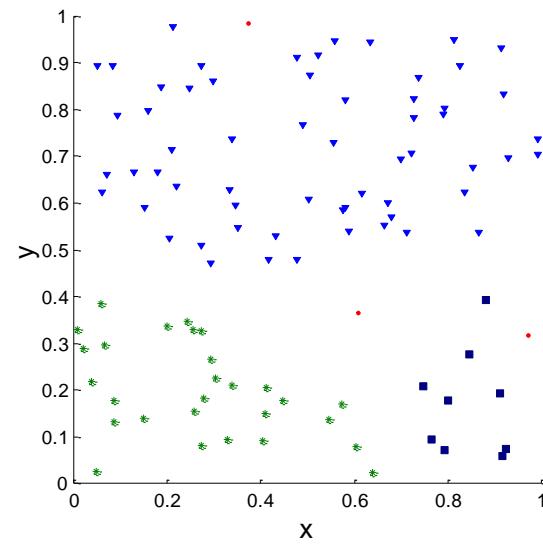
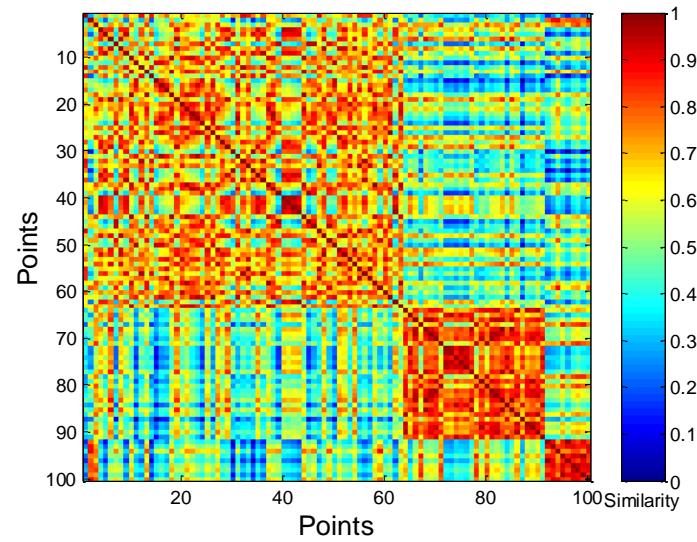
Using similarity matrix for cluster validation

- Order the similarity matrix with respect to cluster labels and inspect visually



Using similarity matrix for cluster validation

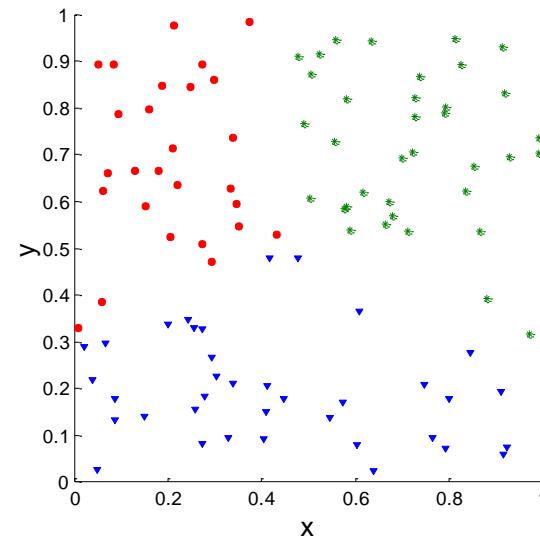
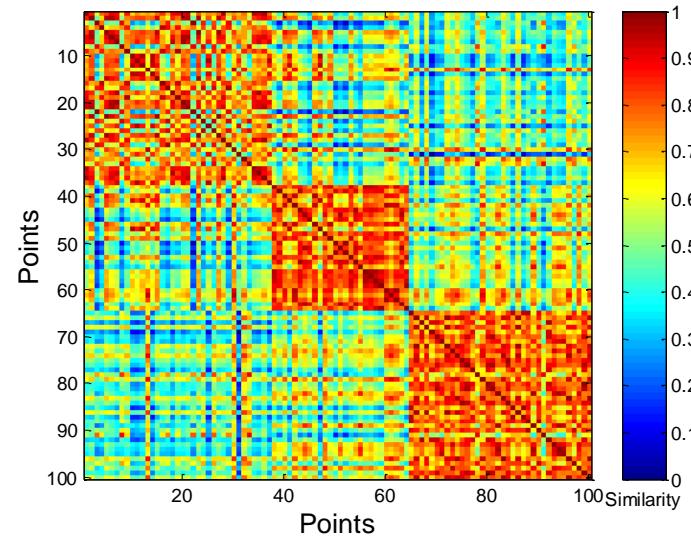
- Clusters in random data are not so crisp



DBSCAN

Using similarity matrix for cluster validation

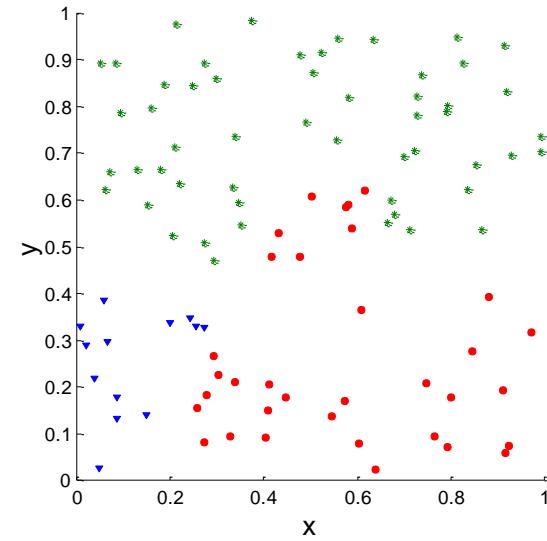
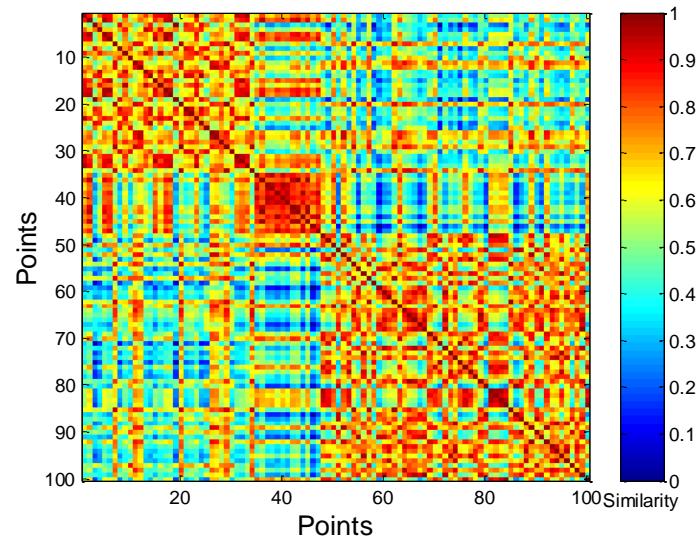
- Clusters in random data are not so crisp



K-means

Using similarity matrix for cluster validation

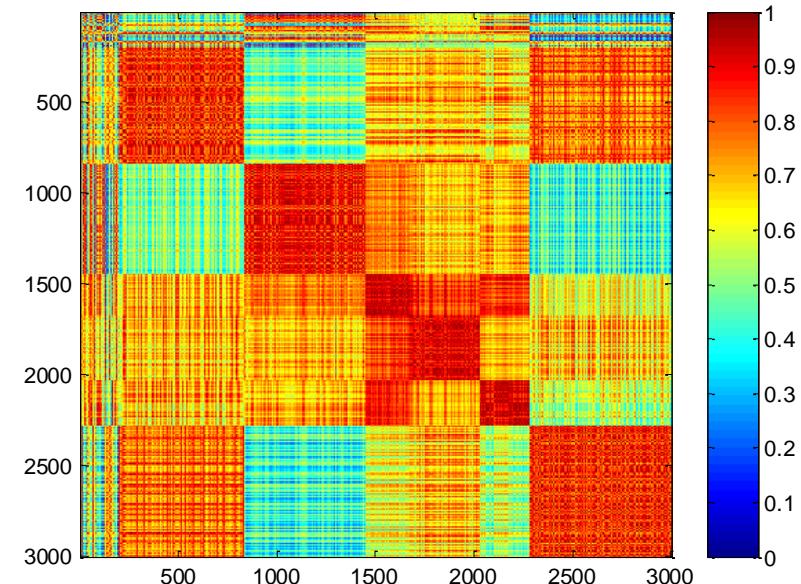
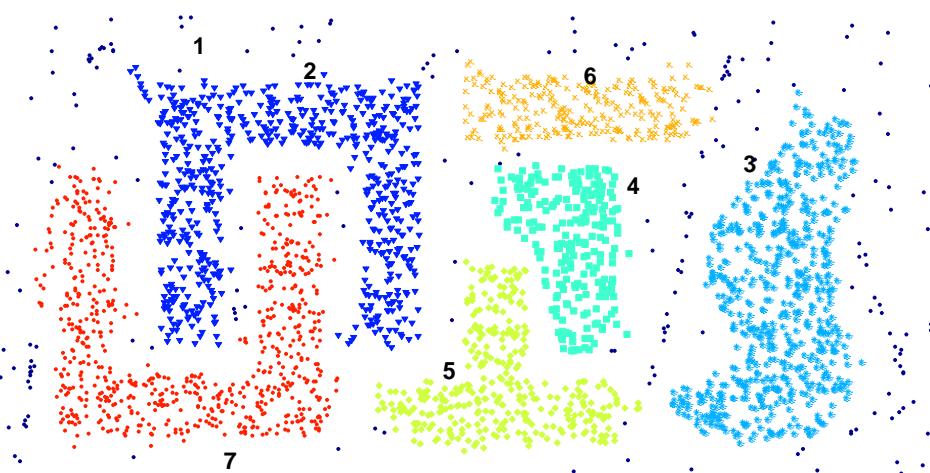
- Clusters in random data are not so crisp



Complete Link

Using similarity matrix for cluster validation

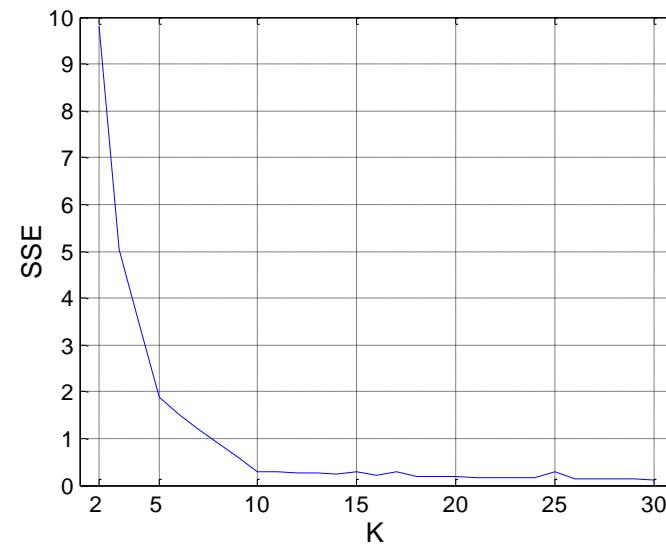
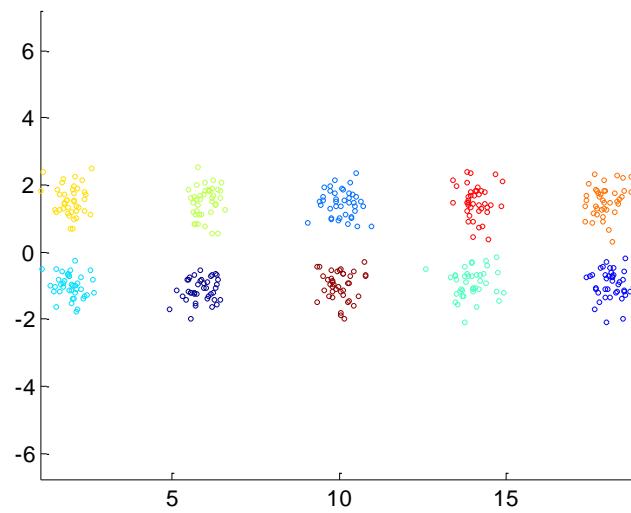
- Clusters in random data are not so crisp



DBSCAN

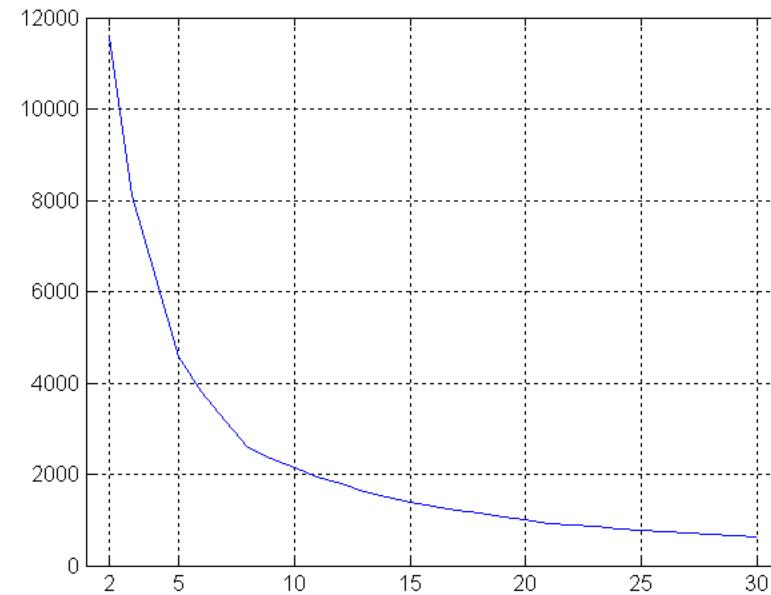
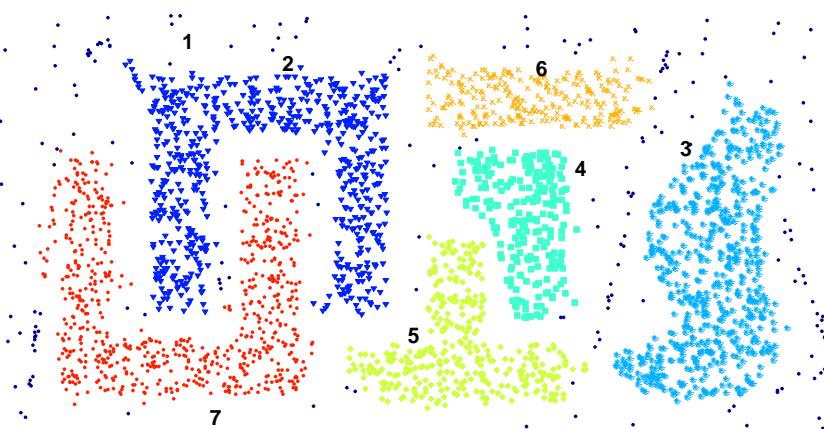
Internal Measures: SSE

- Clusters in more complicated figures aren't well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
 - SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters



Using similarity matrix for cluster validation

- SSE curve for a more complicated data set



SSE of clusters found using K-means

Final Comment on Cluster Validity

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes