Technology and Application of Big Data

Qing LIAO(廖清)

School of Computer Science and Technology

HIT

Course Details

- Instructor:
  - Qing LIAO, liaoqing@hit.edu.cn
  - Rm. 303B, Building C
  - Office hours: by appointment
- Course web site:
  - liaoqing.me
- Reference books/materials:
  - Big data courses from University of California
  - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
  - Papers
- Grading Scheme:
  - Paper Report 30%
  - Final Exam 70%
- Exam:
  - 21st July(Friday), 14:00-16:00, A502

What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data
- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning would help
    - But not necessary

Previous Section: Machine Learning & Data Mining

Computer Algorithm

Process of Converting
Data & Experience
Into Knowledge

Computer Model

Previous Section: Machine Learning & Data Mining

- ML focuses more on algorithms
  - Typically more rigorous
  - Also on analysis (learning theory)
- DM focuses more on knowledge extraction
  - Typically uses ML algorithms
  - Knowledge should be human-understandable

Previous Section:
Machine Learning Algorithm → Data Mining Task (Classification)

$$w = (1,0,0,1,0,1)$$
$$b = 1.5$$

Training Set	Bag of Words	
SPAM!	(0,0,0,1,1,1)	$f(x\|w,b) = +1$
SPAM!	(1,0,0,1,0,0)	$f(x\|w,b) = +1$
NOT SPAM	(1,0,1,0,1,0)	$f(x\|w,b) = -1$
NOT SPAM	(0,1,1,0,1,0)	$f(x\|w,b) = -1$
SPAM!	(1,0,1,1,0,1)	$f(x\|w,b) = +1$
SPAM!	(1,0,0,0,0,1)	$f(x\|w,b) = +1$

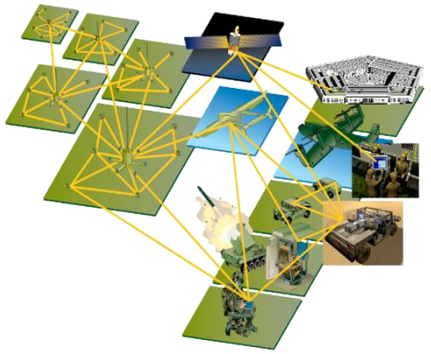
$$f(x|w,b) = sign(w^T x - b) = sign(w_1 * x_1 + \ldots w_6 * x_6 - b)$$

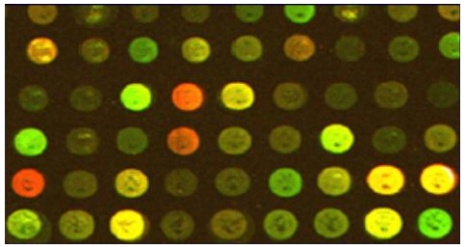
Why Mine Data? Commercial Viewpoint

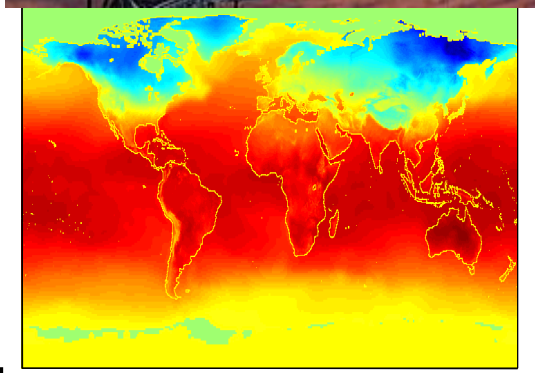
- Lots of data is being collected and warehoused
  - web data, e-commerce
  - purchases at department/ grocery stores
  - bank/credit card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

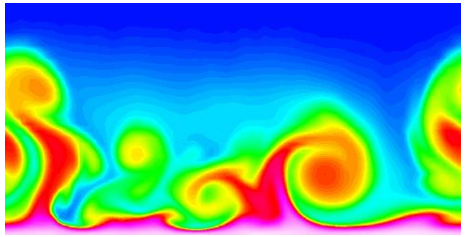
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - ➢ remote sensors on a satellite
  - ➢ telescopes scanning the skies
  - ➢ microarrays generating gene expression data
  - ➢ scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - ➢ in classifying and segmenting data
  - ➢ in hypothesis formation



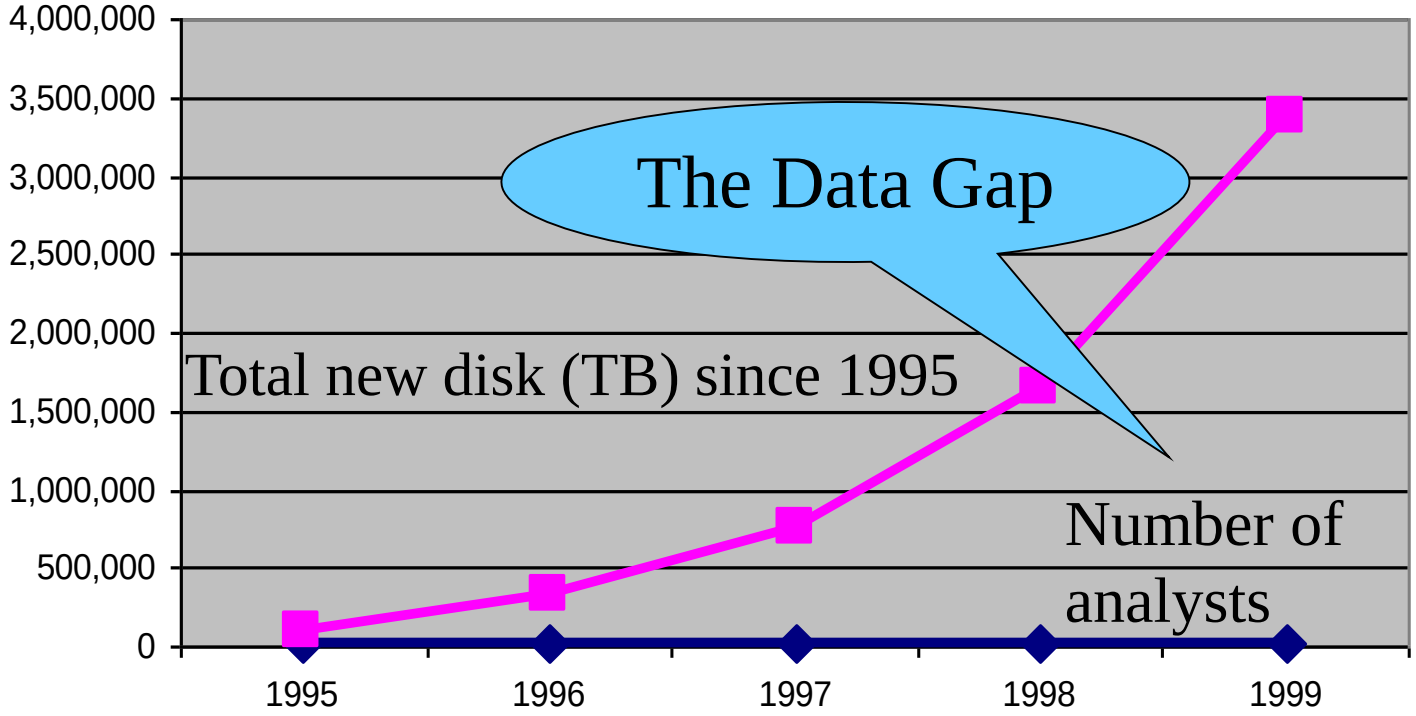






Mining Large Data Sets - Motivation

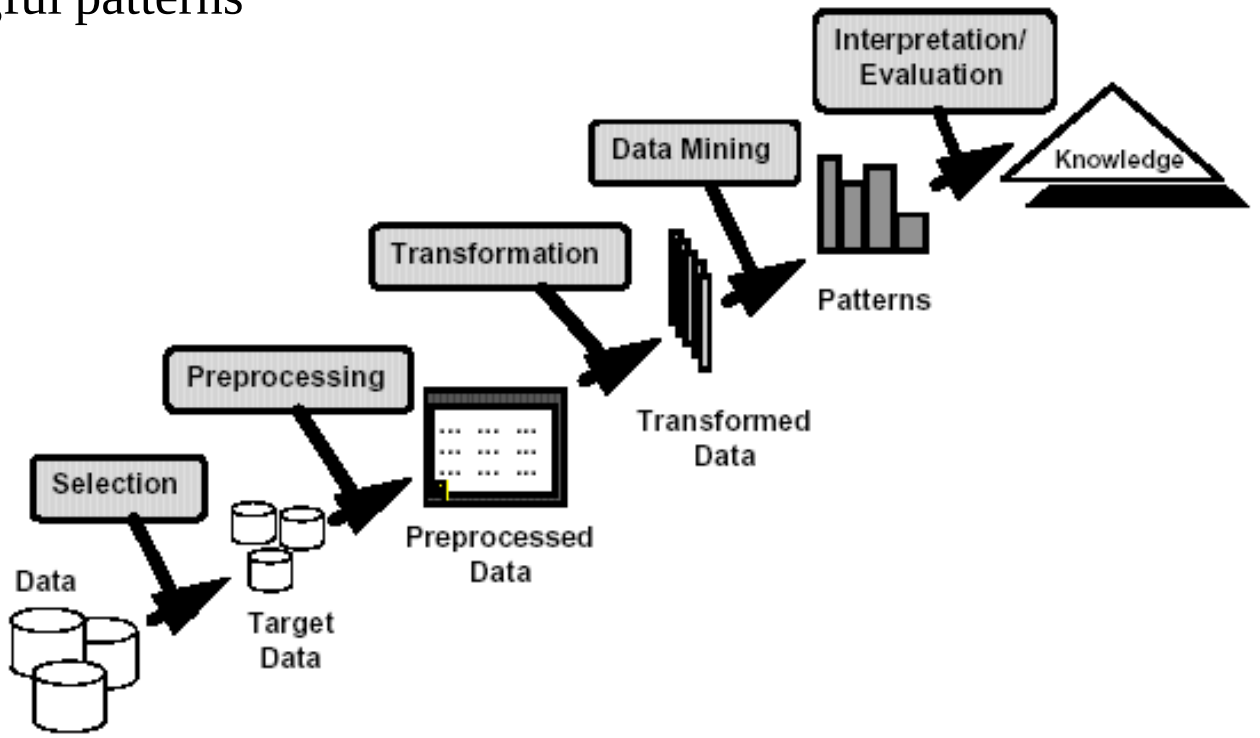
- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



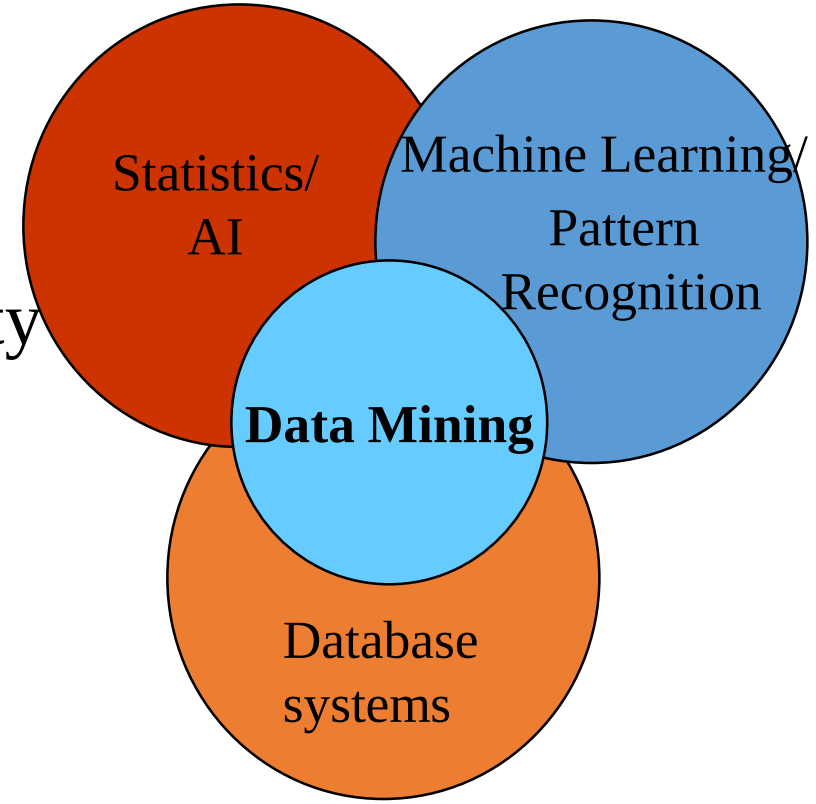
What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- What is Data Mining?
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g., Amazon.com,)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining Tasks

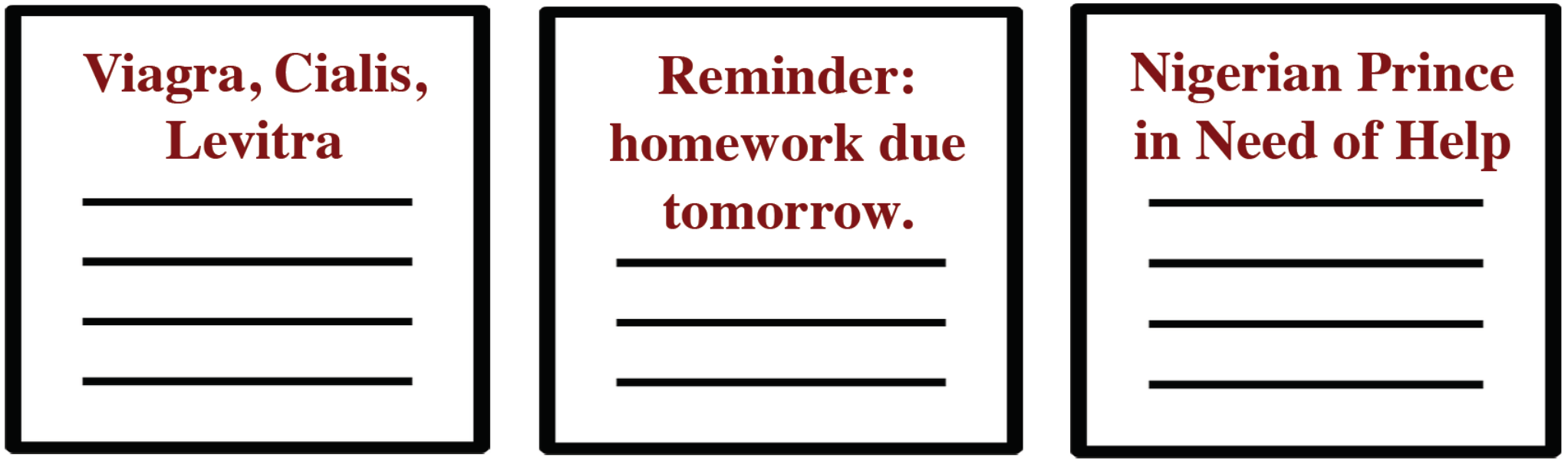
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Supervised Learning

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example: Spam Filtering

- Goal: write a program to filter spam.

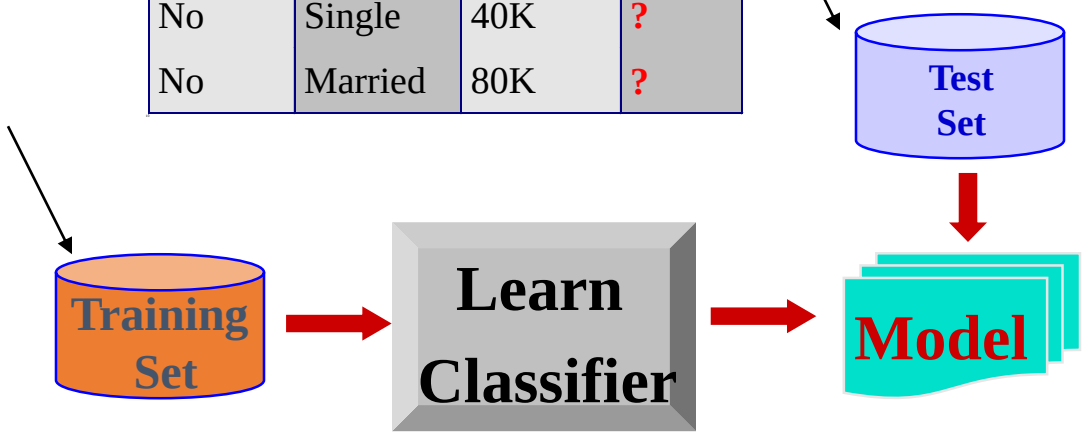


Classification Example: Tax Cheating

- Tax Income

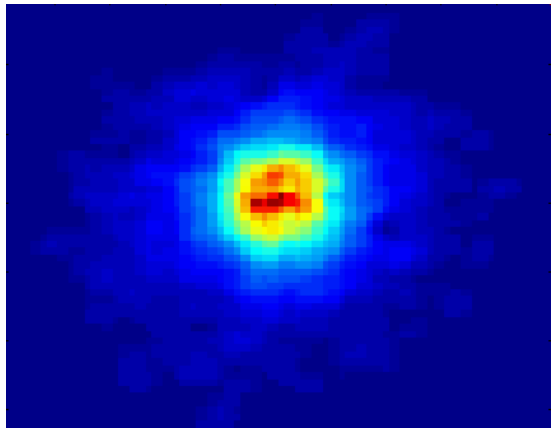
	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



Classification Example: Classifying Galaxies

Early



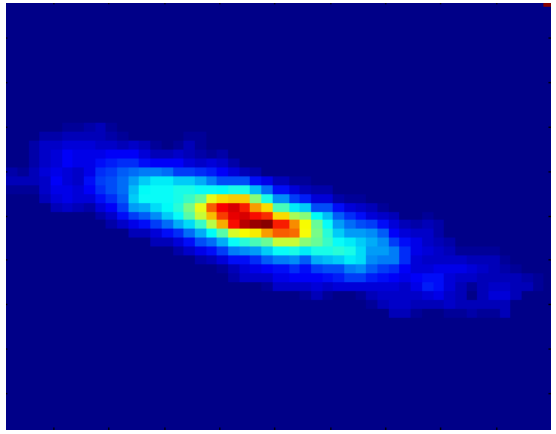
Class:

- **Stages of Formation**

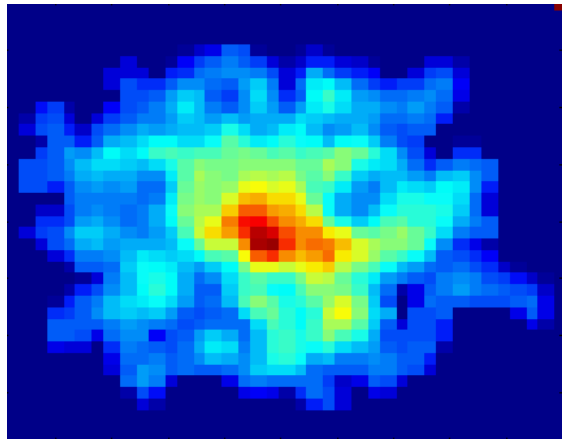
Attributes:

- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



Data Size:

- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Clustering: Unsupervised Learning

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that

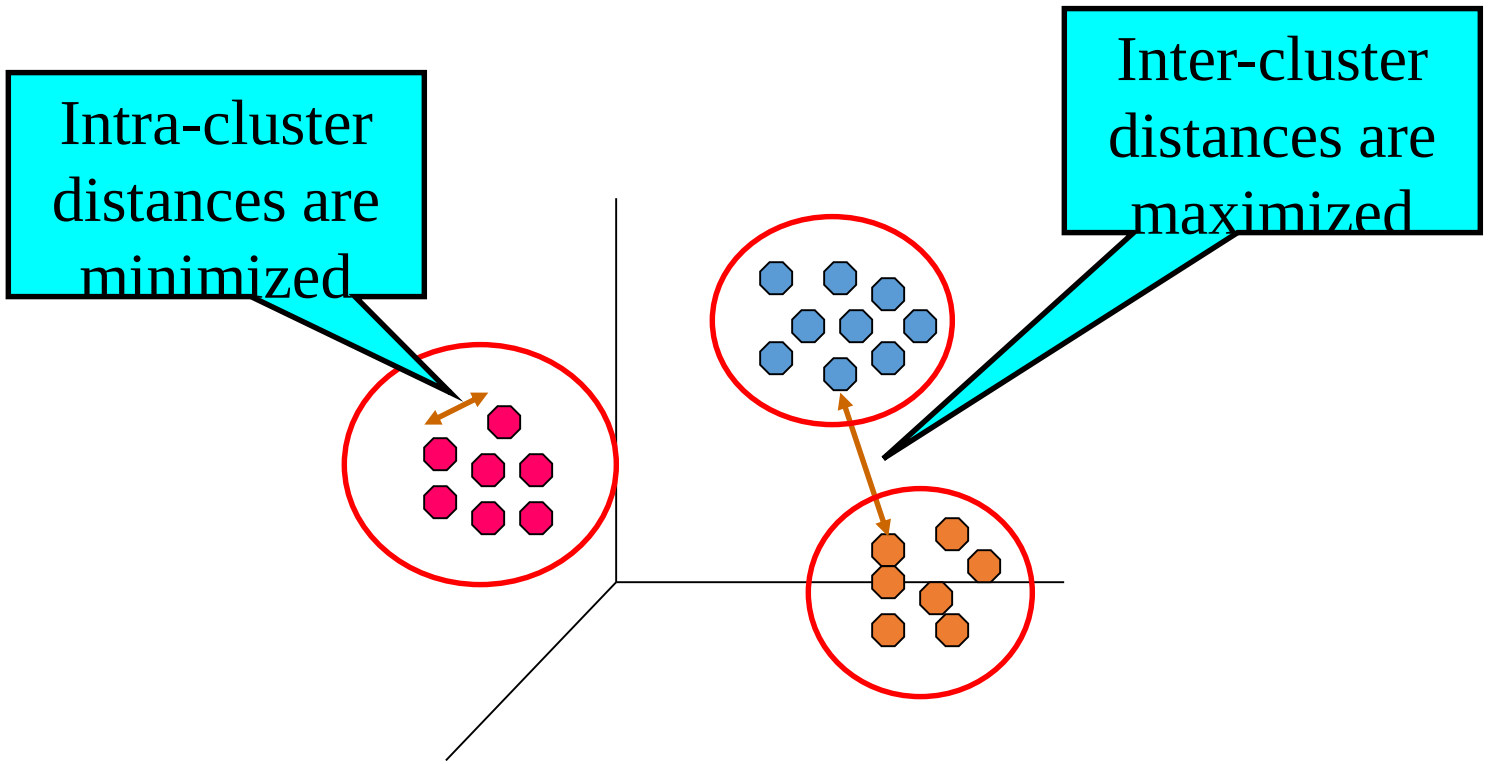
- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Similarity Measures:

- Euclidean Distance if attributes are continuous.
- Other Problem-specific Measures.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.



Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Similarity and Dissimilarity

Similarity

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range [0,1]

Dissimilarity

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

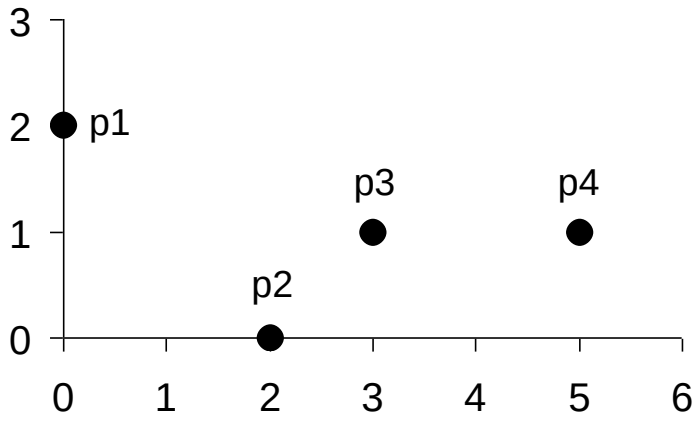
Euclidean Distance

$$\boldsymbol{dist} = \sqrt{\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} (\boldsymbol{p_k} - \boldsymbol{q_k})^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.

Standardization is necessary, if scales differ.

Euclidean Distance



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

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

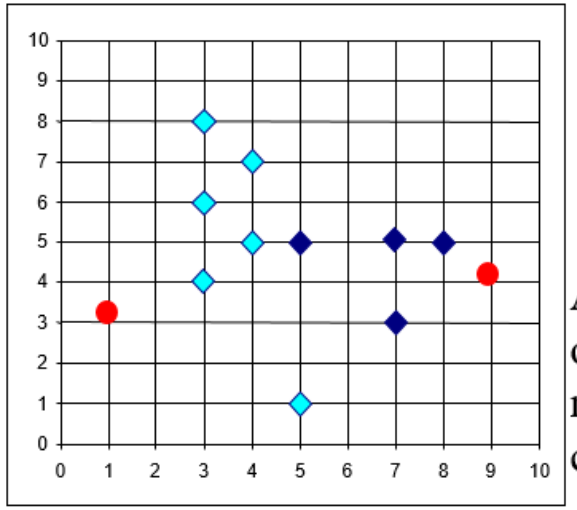
Distance Matrix

$dist_{p1,p3} = \sqrt{(0-3)^2 + (2-1)^2} = \sqrt{9+1} = 3.162$

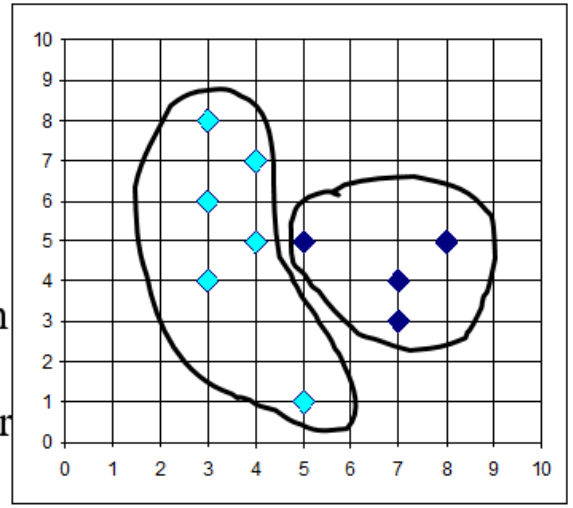
The K-Means Clustering Method: for numerical attributes

- Given k, the k-means algorithm is implemented in four steps:
  - Partition objects into k non-empty subsets
  - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., mean point, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when no more new assignment

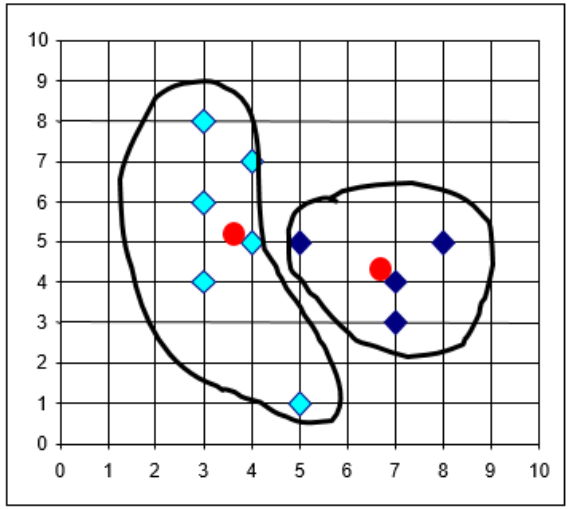
K-means Clustering



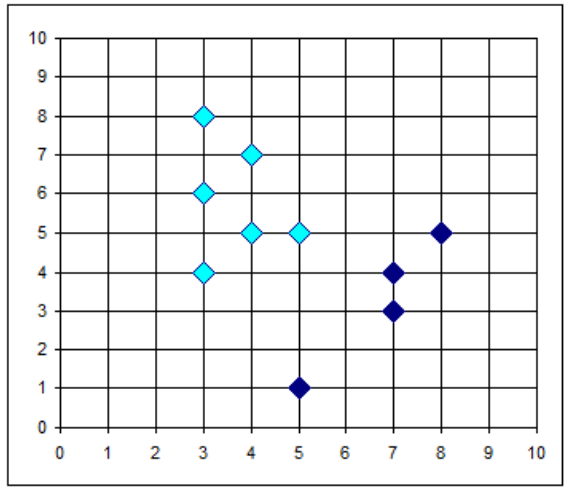
Assign each objects to most similar center



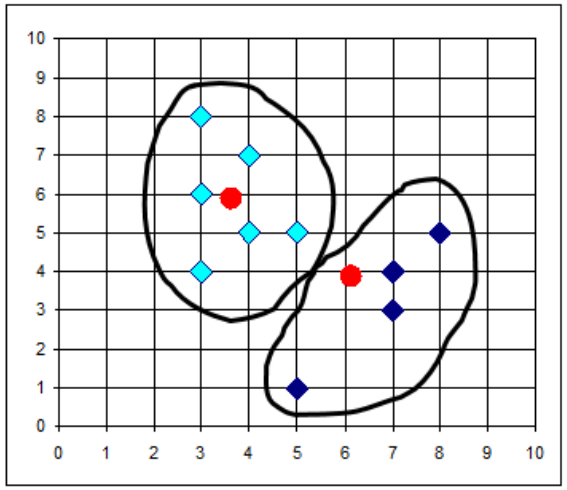
Update the cluster means



reassign



Update the cluster means



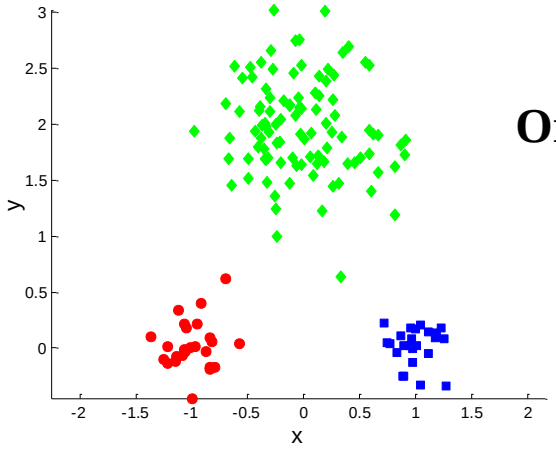
reassign

K=2

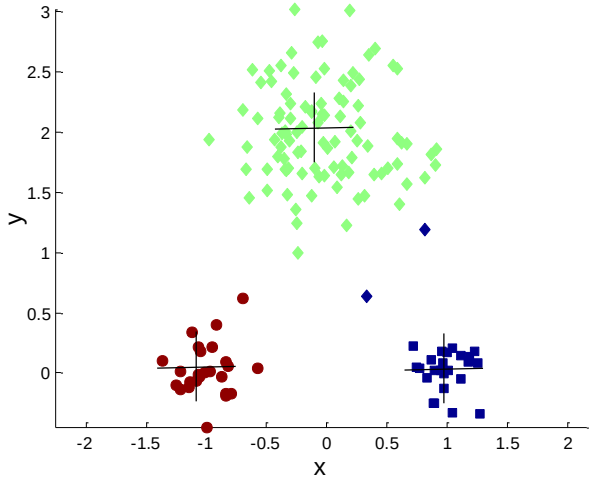
Arbitrarily choose K object as initial cluster center

The mean point can be a virtual point!

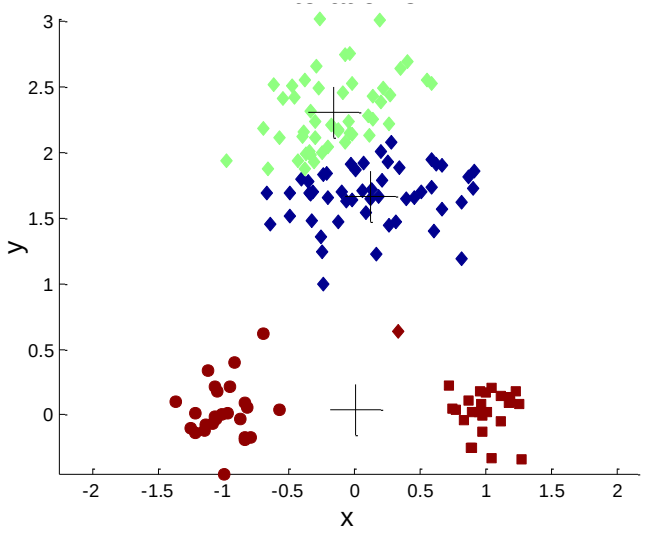
K-means Clustering



Original Points

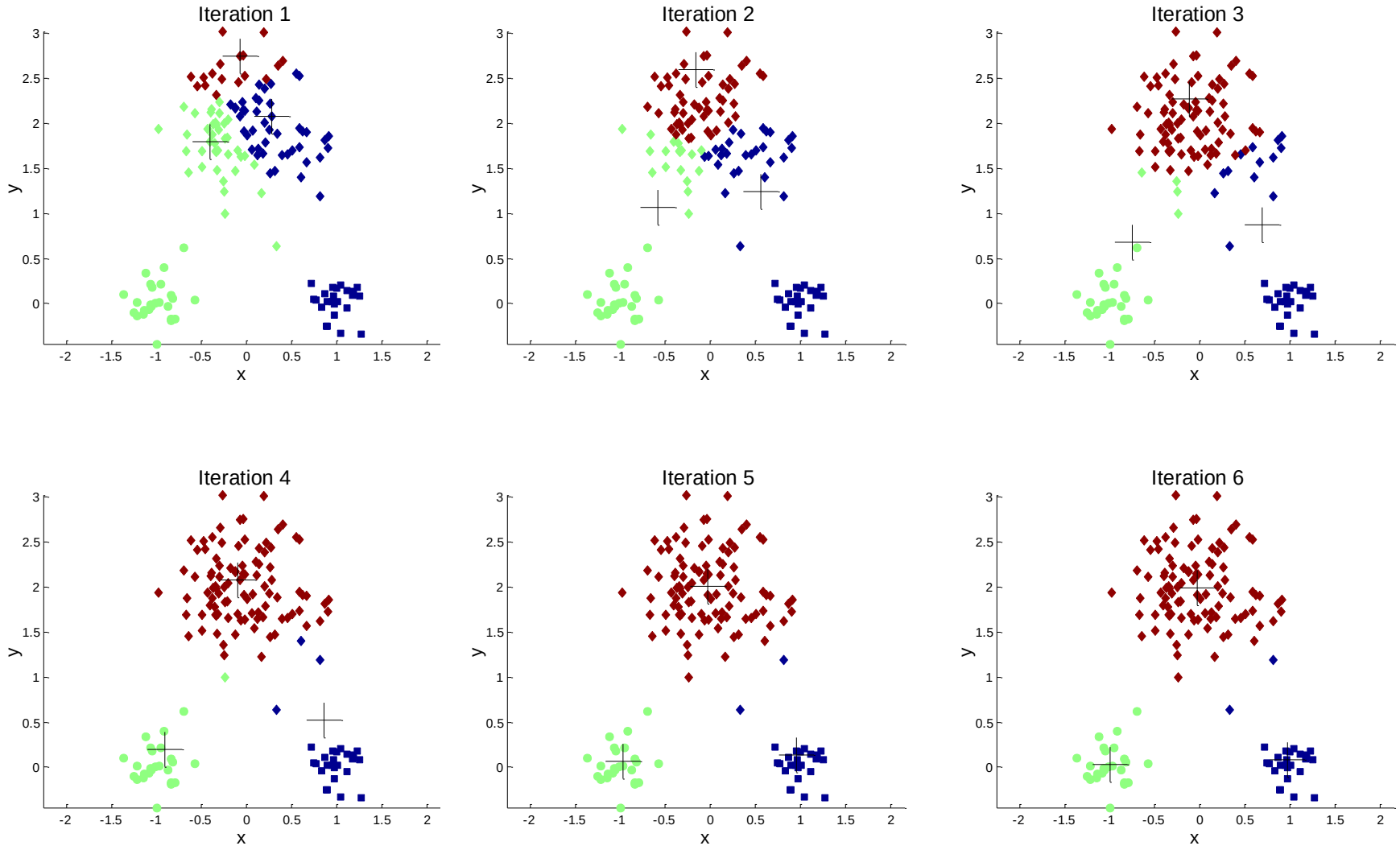


Optimal Clustering

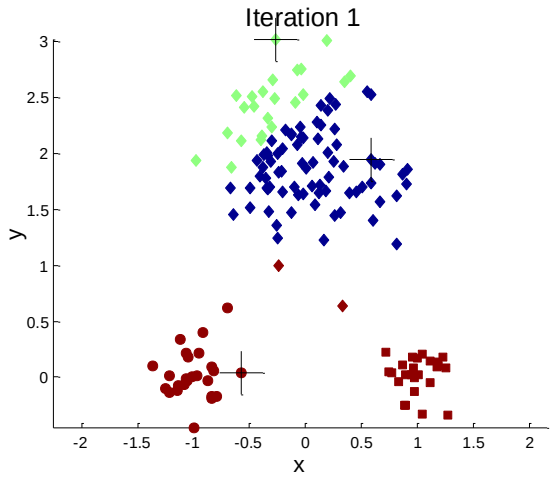


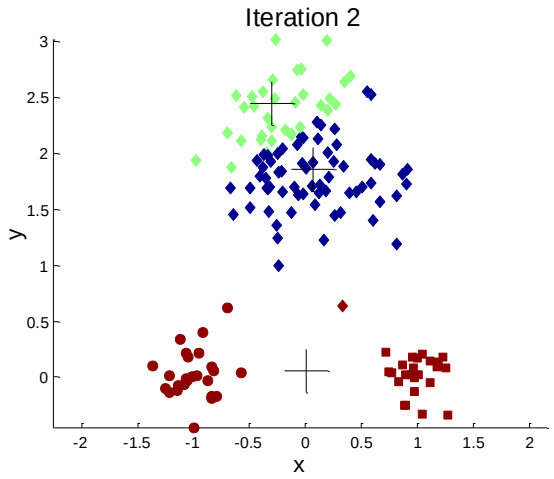
Sub-Optimal Clustering

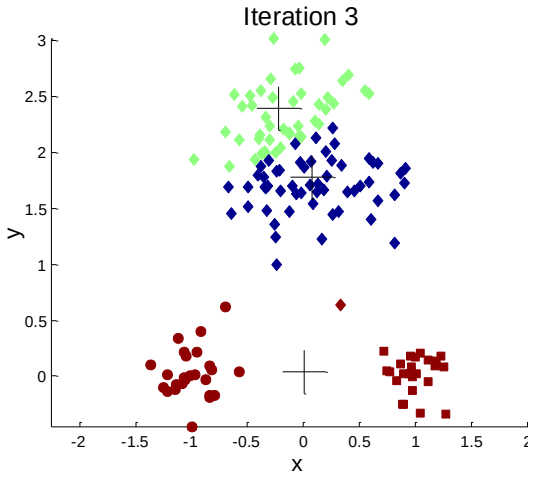
K-means Clustering: Importance of Choosing Initial Centroids

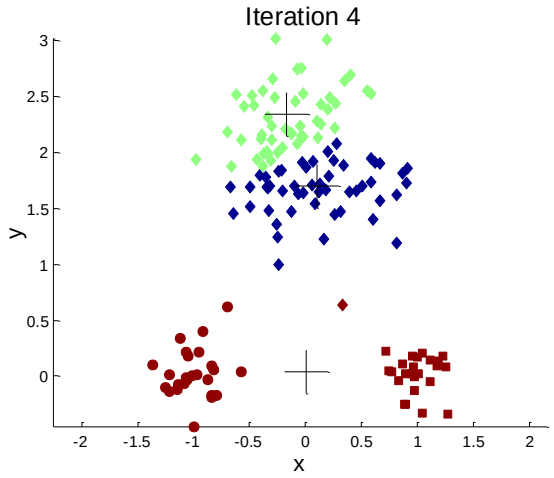


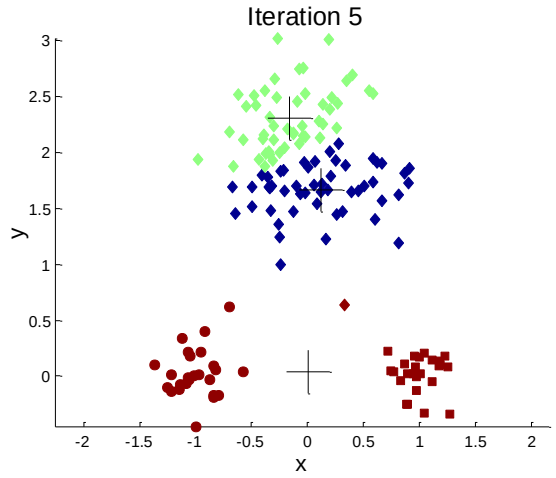
K-means Clustering: Importance of Choosing Initial Centroids





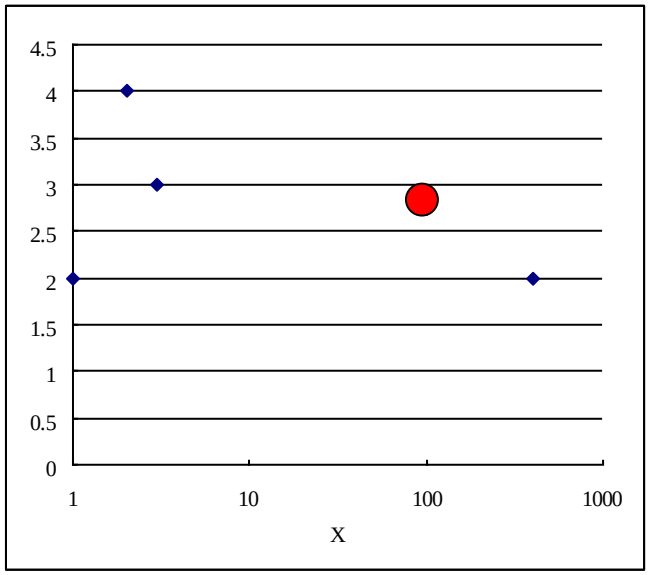






K-means Clustering Problem 1

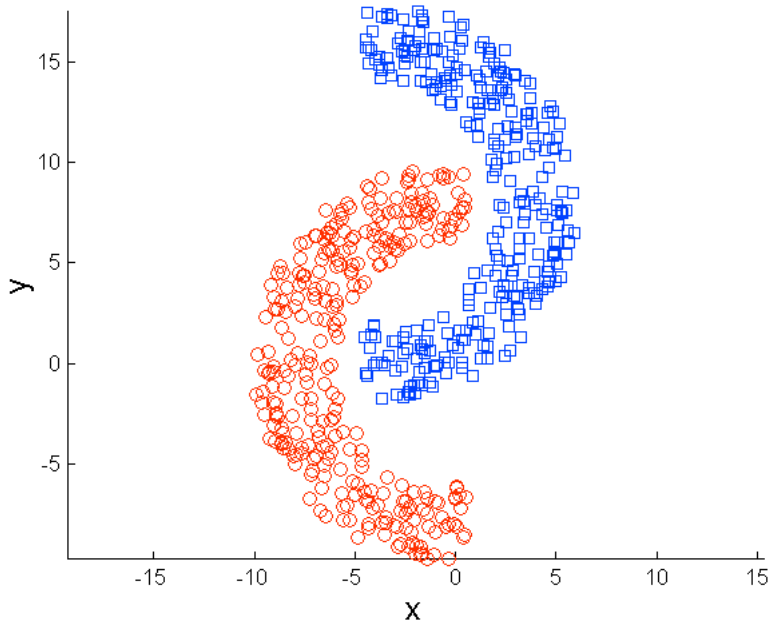
- The mean point can be influenced by an outlier



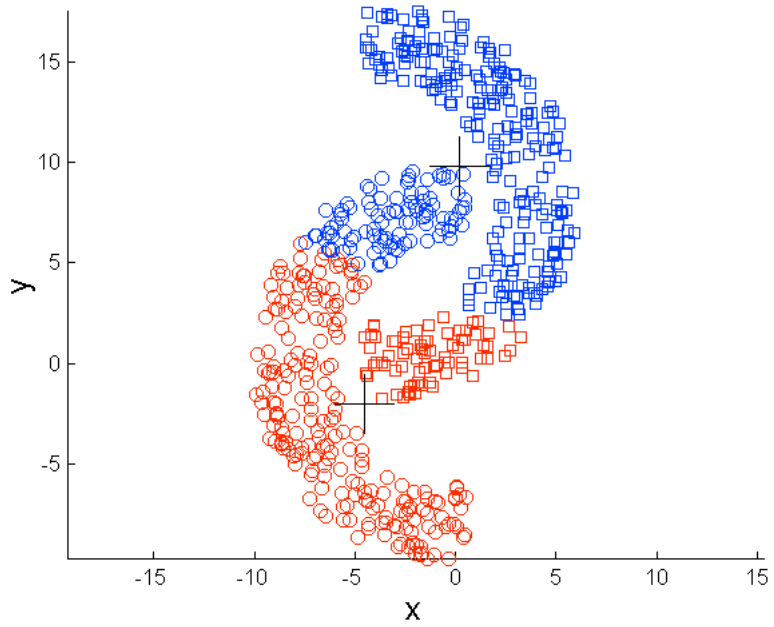
X	Y
1	2
2	4
3	3
400	2
101.5	2.75

K-means Clustering Problem 2

- Non-globular Shapes



Original Points



K-means (2 Clusters)

Clustering Example: Market Segmentation

- Goal: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering Example: Document Clustering

Classification vs. Clustering

- Supervised learning (classification)
  - training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning(clustering)
  - The class labels of training data is **unknown**
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data