

Data Mining

COIS 4400H / AMOD 5440H

Introduction

Goals

at the end of the term, you should know

- how to prepare data for data mining
- how the most relevant/most popular data-mining algorithms work
- know the major pitfalls in applying these algorithms and how to avoid them
- be able to evaluate the information you extract from the data
- where to look for help or additional information
- have gained familiarity with one particular data-mining tool (and have analyzed a few datasets yourself)

What is data mining?

“The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.”

(Piatetsky-Shapiro)

“ The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. ”

(Hand)

Data mining is a combination of:

Machine learning

Databases

Visualization

Application domain

Statistics

Data mining vs statistics

datasets used in data mining typically are

- samples
- larger
- noisier, incomplete, heterogeneous (contains many different types)

statistics

- often deals with the whole population
- often is concerned with hypothesis testing

Example applications

Science: bioinformatics, discovery of drugs, astronomy

Government: law enforcement, income tax, anti-terror

Business: Market basket analysis, targeted marketing

Engineering: Satellite navigation

Some questions we can answer with data mining techniques:

Is this object a star or a galaxy?

Are customers likely to buy bread together with milk?

Which customers are likely to buy DVD's after buying a DVD player the week before?

Does the traffic we currently see in our network contain any malicious packets?

What is the value for a particular stock going to be tomorrow? Next month? Next year?

How many groups of customers are there in the data we collected?

Would it make sense to send flyers to one or more of those groups to increase our sales

How can we characterize those groups of customers?

How are those customers in a group similar to each other?

How are they different from other groups?

What book is this customer likely to buy?

Are there additional books we should recommend?

Which movies is this customer going to rent in the next month?

Main data-mining tasks

Classification: assign a category to each object in the data

Visualization: what does the data look like?

Clustering: can we determine groups of objects in the data?

Association Rule Discovery: which objects belong together?

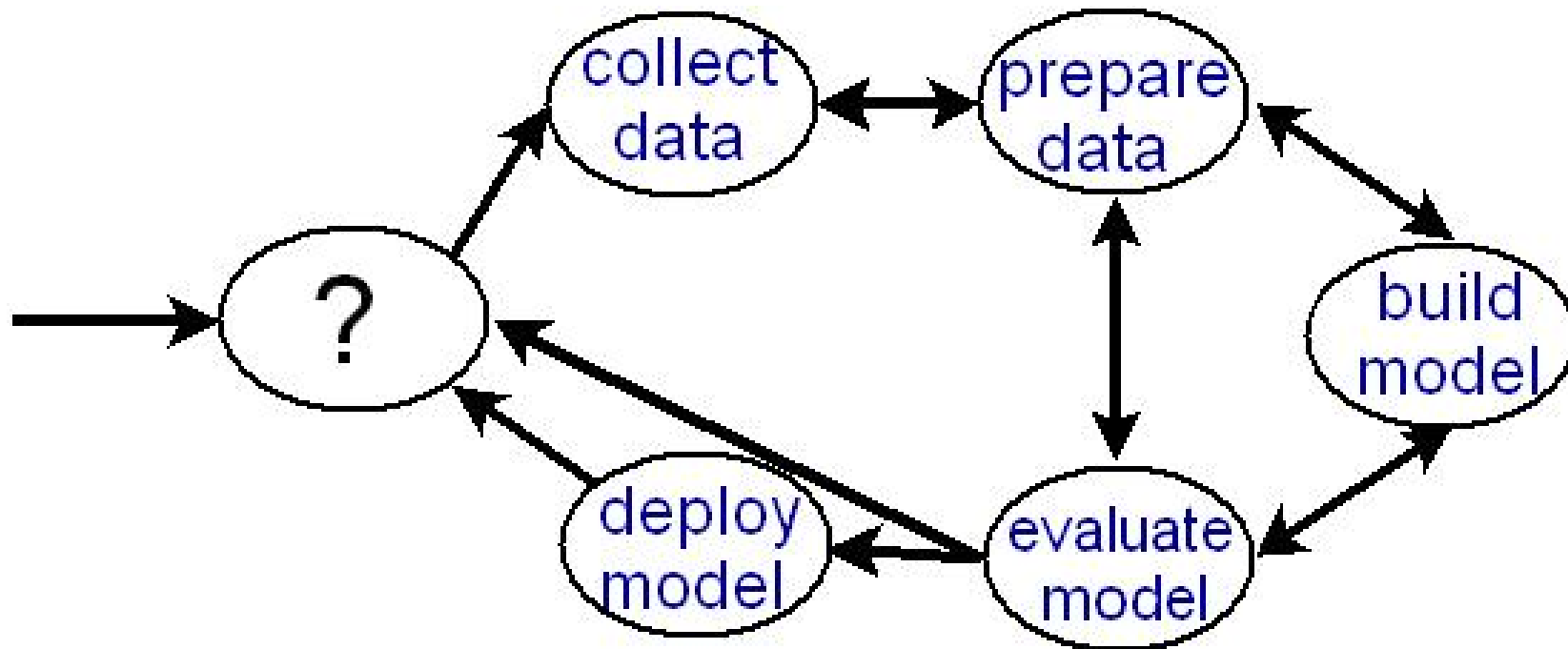
Outlier Detection: which objects do not belong with the rest?

Sequential Pattern Discovery: what happens over time?

Regression: assign a numerical value to each object in the data

The data-mining process

(Knowledge discovery in databases, data dredging, data fishing...)



Preparing the data

- Neural networks like data to be scaled
- Decision trees do not care about scaling, but work better with discrete attributes that have small numbers of possible values
- Neural networks can handle irrelevant or redundant attributes, while they may lead to large decision trees
- Neural networks do not like noisy data, especially for small datasets, while decision trees do not care about noise much
- Nearest-neighbour approaches can handle noise if a certain parameter is adjusted
- Distance-based approaches do not work well if the attributes are not equally weighted, and typically work with numerical data only
- Expectation Maximization approaches can deal with missing data, but k-means techniques require substitution of missing data
- ...

The data pre-processing steps vary with the data and the data-mining techniques applied to the data

Types of data-mining techniques

Data- mining techniques can be

predictive (supervised)

predict (discrete or continuous) class attribute based on other attribute values. This is like “*learning from a teacher*”.

descriptive (unsupervised)

discover structure of data without prior knowledge of class labels

Main data-mining tasks (for this course)

Classification [Predictive]	Is this object a star or a galaxy?
Visualization [Descriptive]	How many groups of customers are there in the data we collected? What book is this customer likely to buy? Are there additional books we should recommend?
Clustering [Descriptive]	How many groups of customers are there in the data we collected? What book is this customer likely to buy? Are there additional books we should recommend?
Association Rule Discovery [Descriptive]	Are customers likely to buy bread together with milk?
Outlier Detection [Predictive or Descriptive]	Does the traffic we currently see in our network contain any malicious packets?

Predictive data-mining techniques

Predictive techniques are subdivided into:

- classification
try to predict a categorical value
- regression
try to predict a numerical value

object ID	_RAJ2000	_DEJ2000	distance	flags	x size	y size	U-B	error	Bar?	class
134633	00 03 09.1	+21 57 34	398	A	1629	1654	14.4	low	no	Irr
3555432	00 03 48.8	+07 28 45	113	D	939	1332	14	medium	yes	Spiral
3432223	00 03 58.6	+20 45 07	835	A	1713	2219	12.7	low	no	Ell
124123	00 05 53.0	+22 32 14	398	A	1092	1400	0	low	no	Irr
333456	00 06 21.4	+17 26 03	398	A	1121	1419	15.1	low	no	Irr
3355478	00 07 16.7	+27 42 31	398	A	1343	1810	13.4	high	no	Spiral
875	00 07 16.1	+08 18 03	879	A	1095	1281	14.6	medium	yes	Spiral
33378	00 08 10.7	+27 00 15	578	A	1154	1493	14.4	high	no	Irr
569433	00 08 20.5	+40 37 54	398	A	1661	1683	0	low	no	Irr
3321347	00 09 54.3	+25 55 28	778	A	1961	2180	12.5	low	no	Spiral
5464648	00 10 47.7	+33 21 18	79	B	929	1359	13.5	high	no	Ell
454345476	00 12 49.9	+77 47 44	398	A	1393	1671	0	low	no	Irr
4646788	00 13 27.5	+17 29 16	398	A	1141	1573	14.2	medium	yes	Spiral

Regression

Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

Examples:

- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices.

Classification: Definition

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.

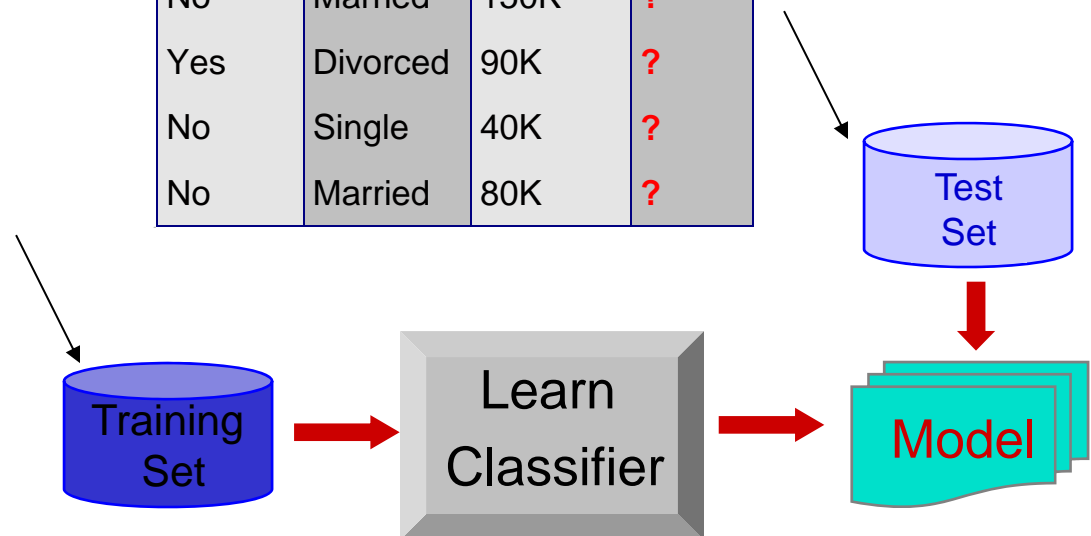
Goal: previously unseen records should be assigned a class as accurately as possible.

Classification Example

categorical
categorical
continuous
class

<i>Tid</i>	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: Application 1

Direct Marketing. Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.

Approach:

- Use the data for a similar product introduced before.
- We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
- Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
- Use this information as input attributes to learn a classifier model.

Classification: Application 2

Fraud Detection. Goal: Predict fraudulent cases in credit card transactions.

Approach:

- Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
- Label past transactions as fraud or fair transactions. This forms the class attribute.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

Customer Attrition/Churn. Goal: To predict whether a customer is likely to be lost to a competitor.

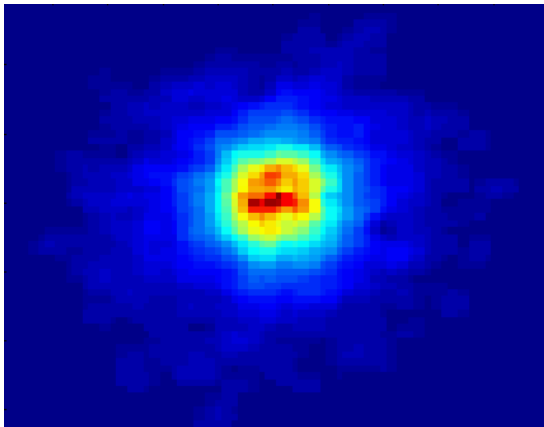
Approach:

- Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day called most, financial status, marital status, etc.
- Label the customers as loyal or disloyal.
- Find a model for loyalty.

Classifying Galaxies

Courtesy: <http://aps.umn.edu>

Early



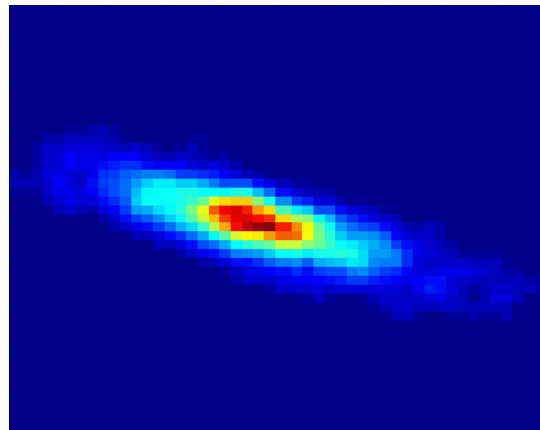
Class:

- Stages of Formation

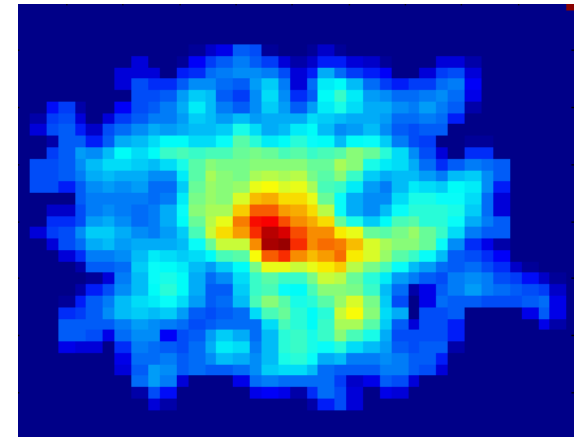
Attributes:

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

Intermediate



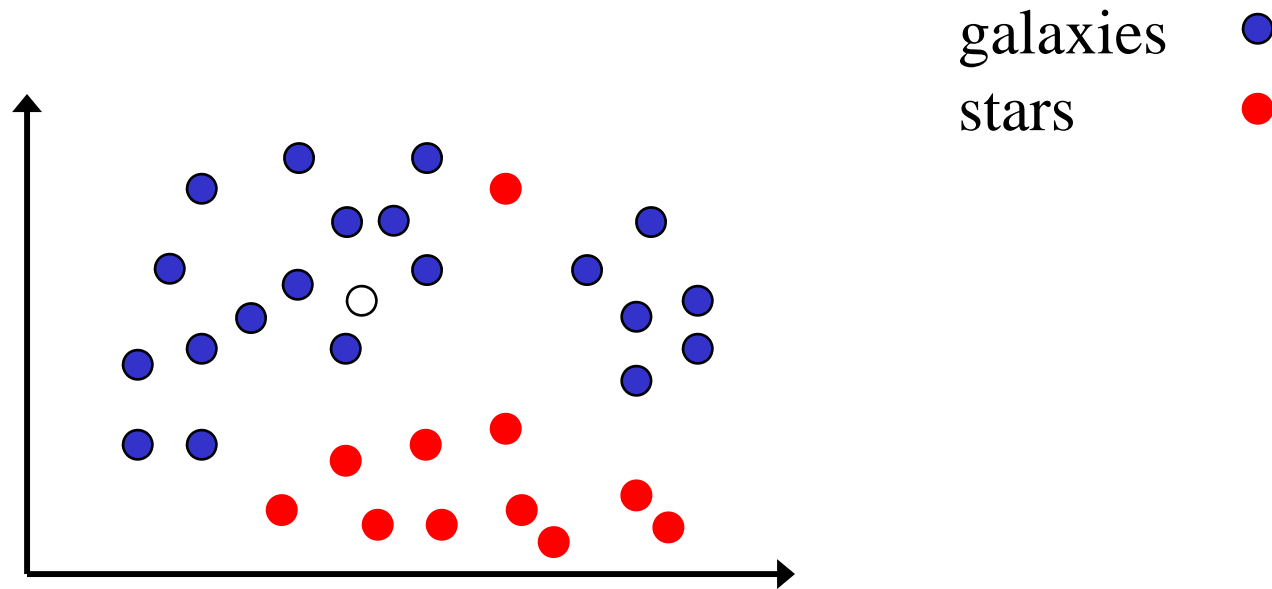
Late



Data Size:

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

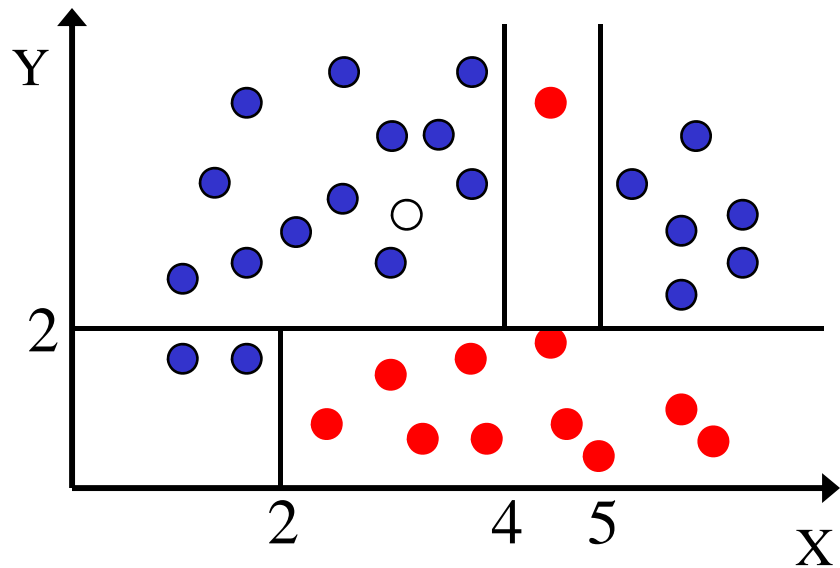
Predictive data mining: Learn a model to predict categories or numerical values



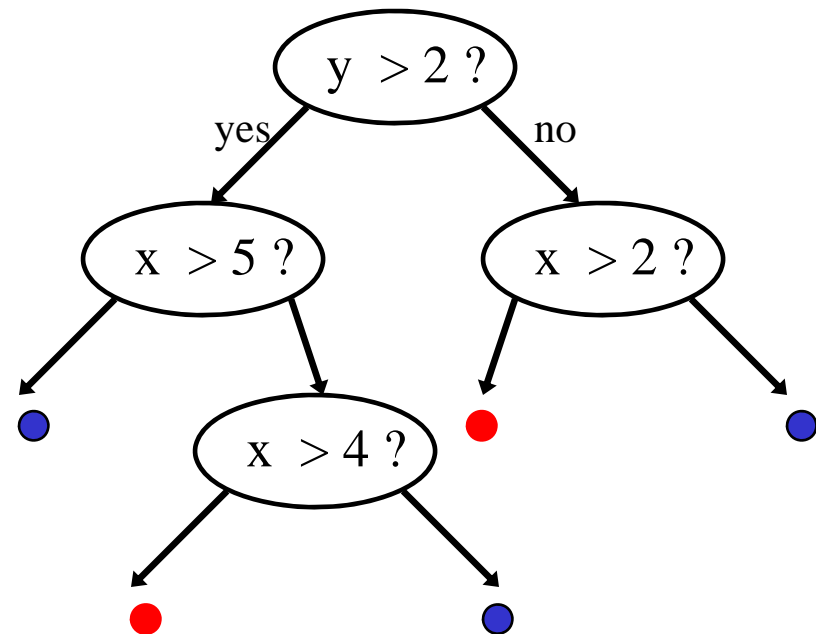
Given a set of points from classes ● ●
what is the class of new point ○?

Is the new point a star or a galaxy?

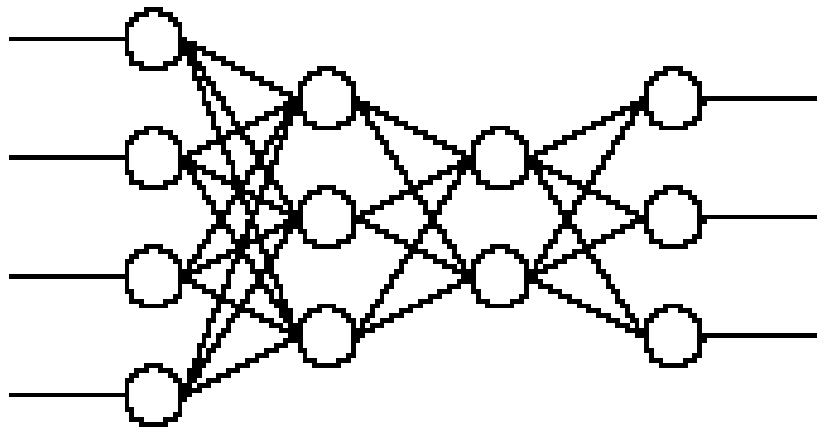
Predictive data mining: Decision Trees



if $y > 2$ then
 if $x > 5$ then blue
 else
 if $x > 4$ then red
 else blue
else
 if $x > 2$ then red
 else blue



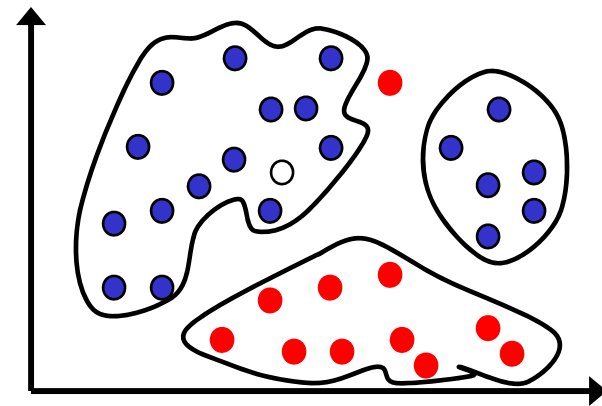
Predictive data mining: Neural Networks



input
layer

hidden
layers

output
layer



- more complex borders
- more accurate
- may *overfit* the data

Clustering

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters (= groups of things) such that:

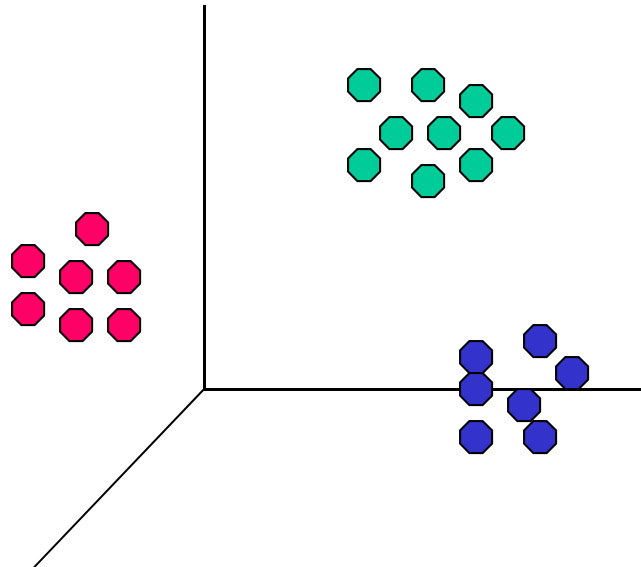
- things in one cluster are more similar to one another
- things in separate clusters are less similar to one another

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracuster distances
are minimized

Intercluster distances
are maximized



Clustering: Application 1

Market Segmentation. Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

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: Application 2

Document Clustering. Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.

Approach:

- To identify frequently occurring terms in each document.
- Form a similarity measure based on the frequencies of different terms.
- Use it to cluster

Illustrating Document Clustering

Clustering Points: 3204 Articles of Los Angeles Times.

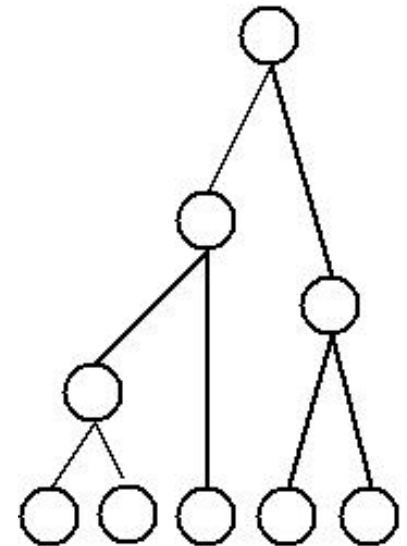
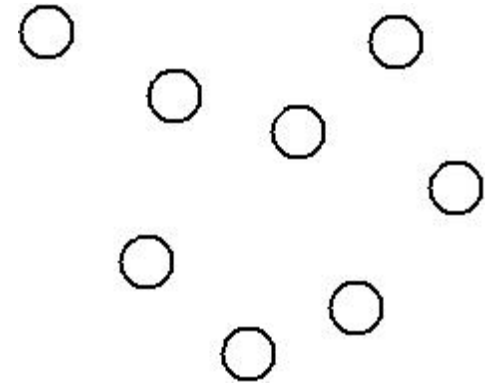
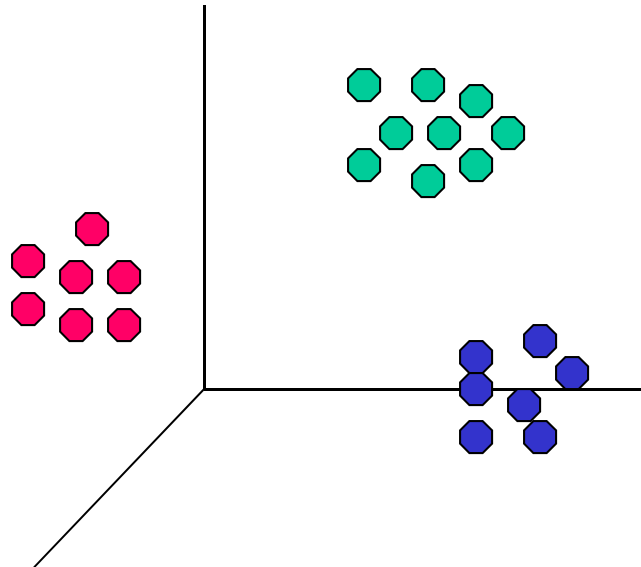
Similarity Measure: How many words are common in these documents (after some word filtering).

<i>Category</i>	<i>Total Articles</i>	<i>Correctly Placed</i>
<i>Financial</i>	555	364
<i>Foreign</i>	341	260
<i>National</i>	273	36
<i>Metro</i>	943	746
<i>Sports</i>	738	573
<i>Entertainment</i>	354	278

Clustering Example

which algorithm should I use?

when are objects similar? or dissimilar?



Descriptive data mining: k-means algorithm (clustering)

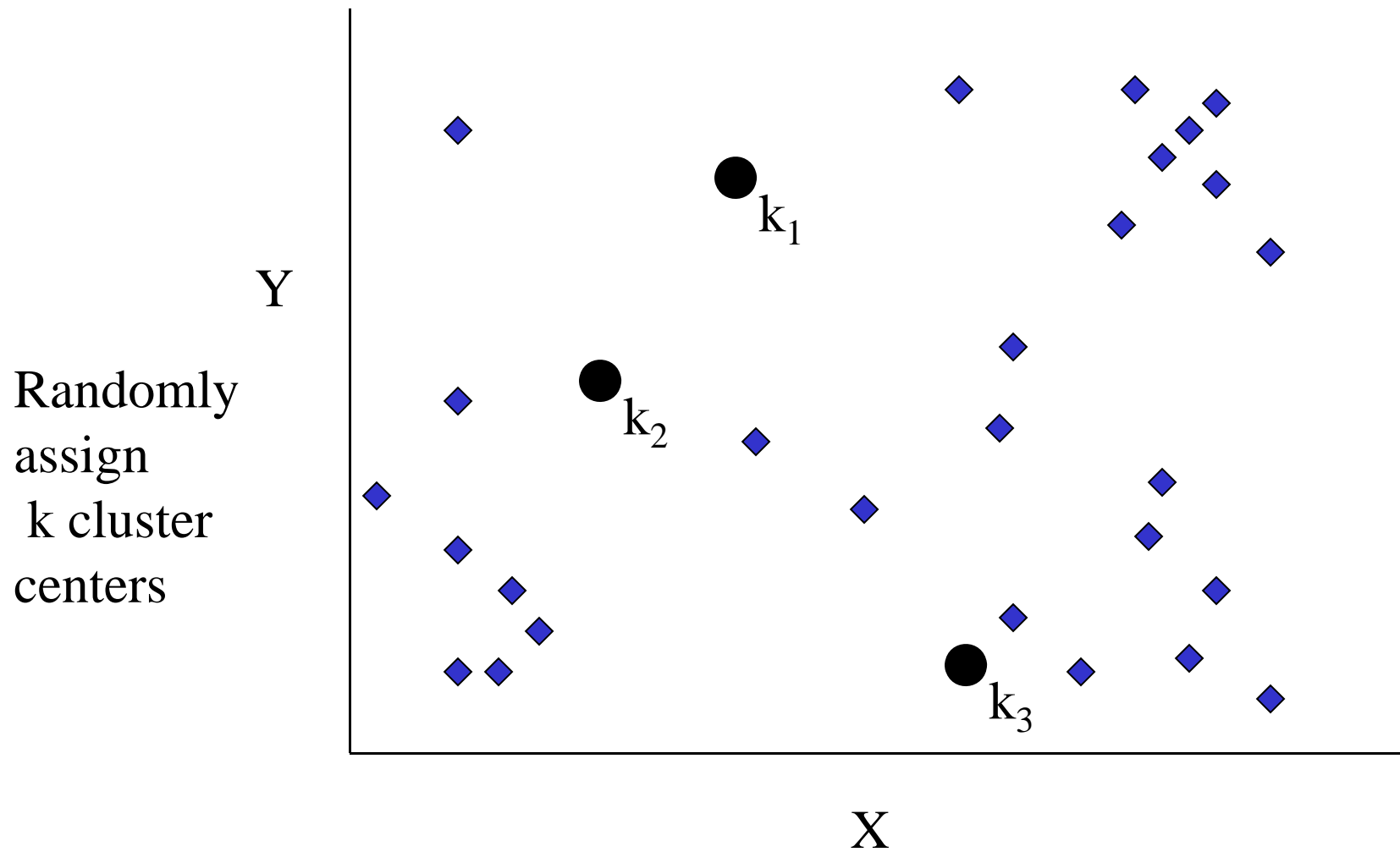
Step 1: Randomly pick k cluster centers

Step 2: Assign every object to its nearest cluster center

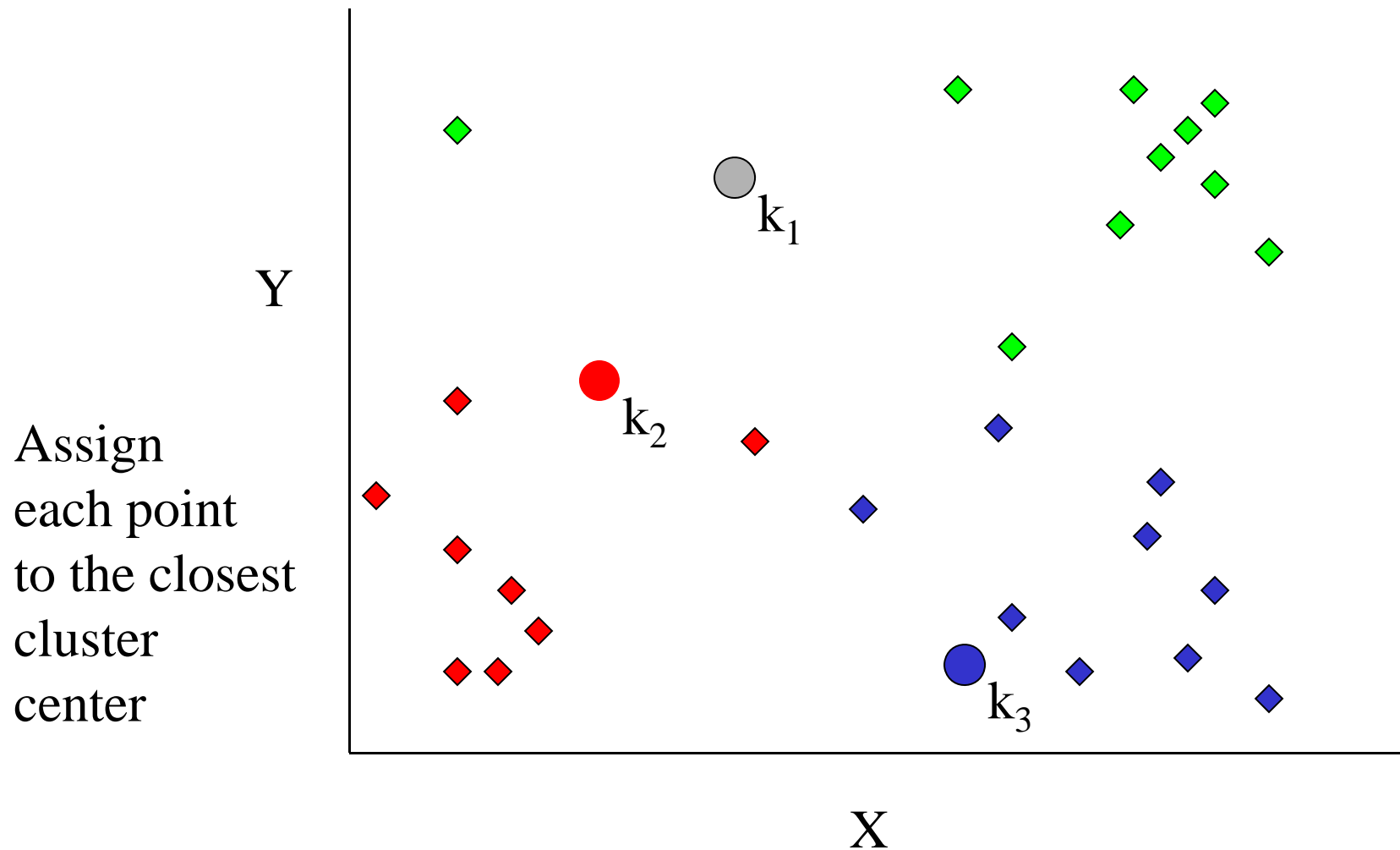
Step 3: Move each cluster center

Step 4: Repeat steps 2,3 until stopping criterion is satisfied

K-means, Step 1

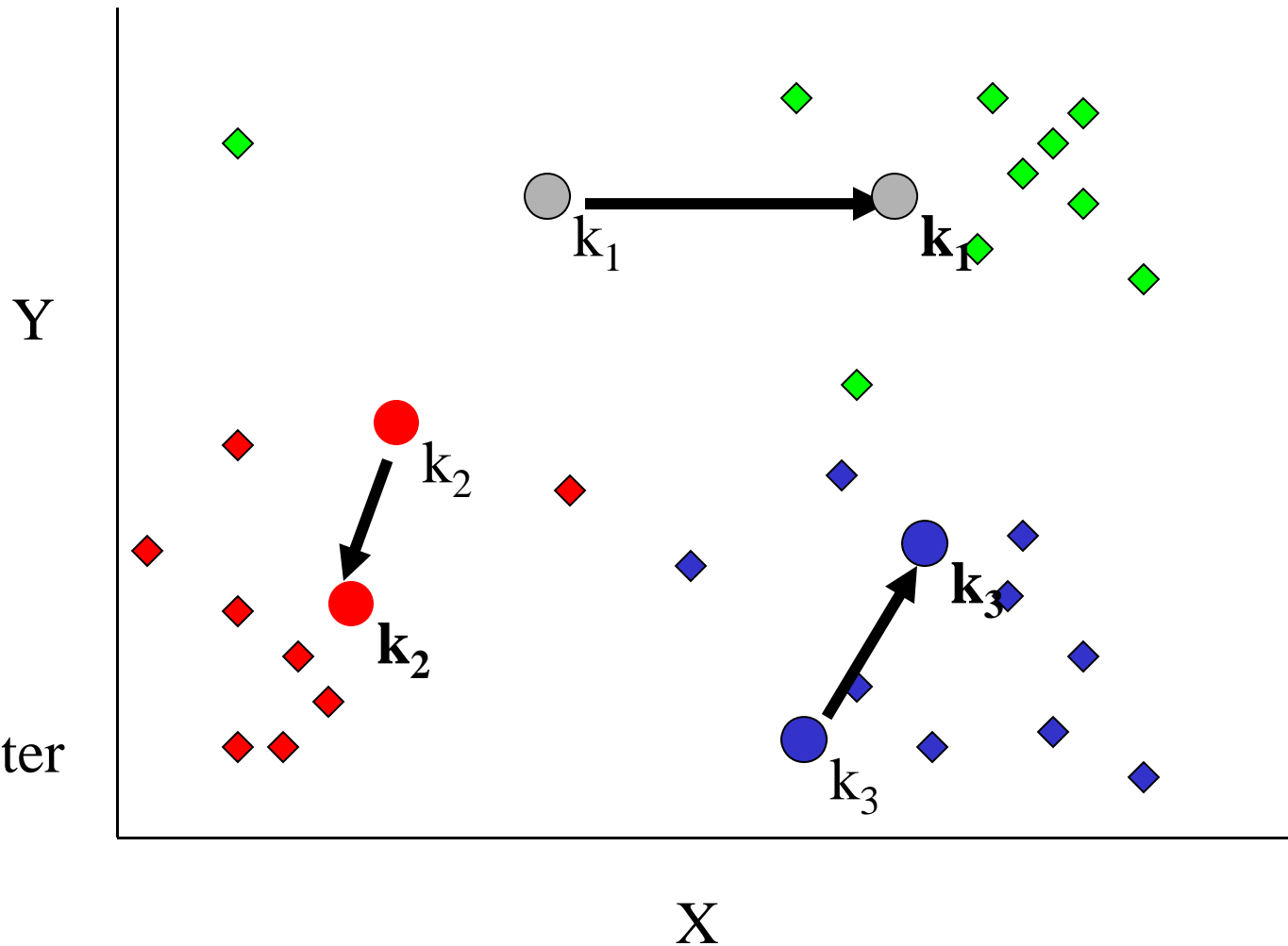


K-means, Step 2



K-means, Step 3

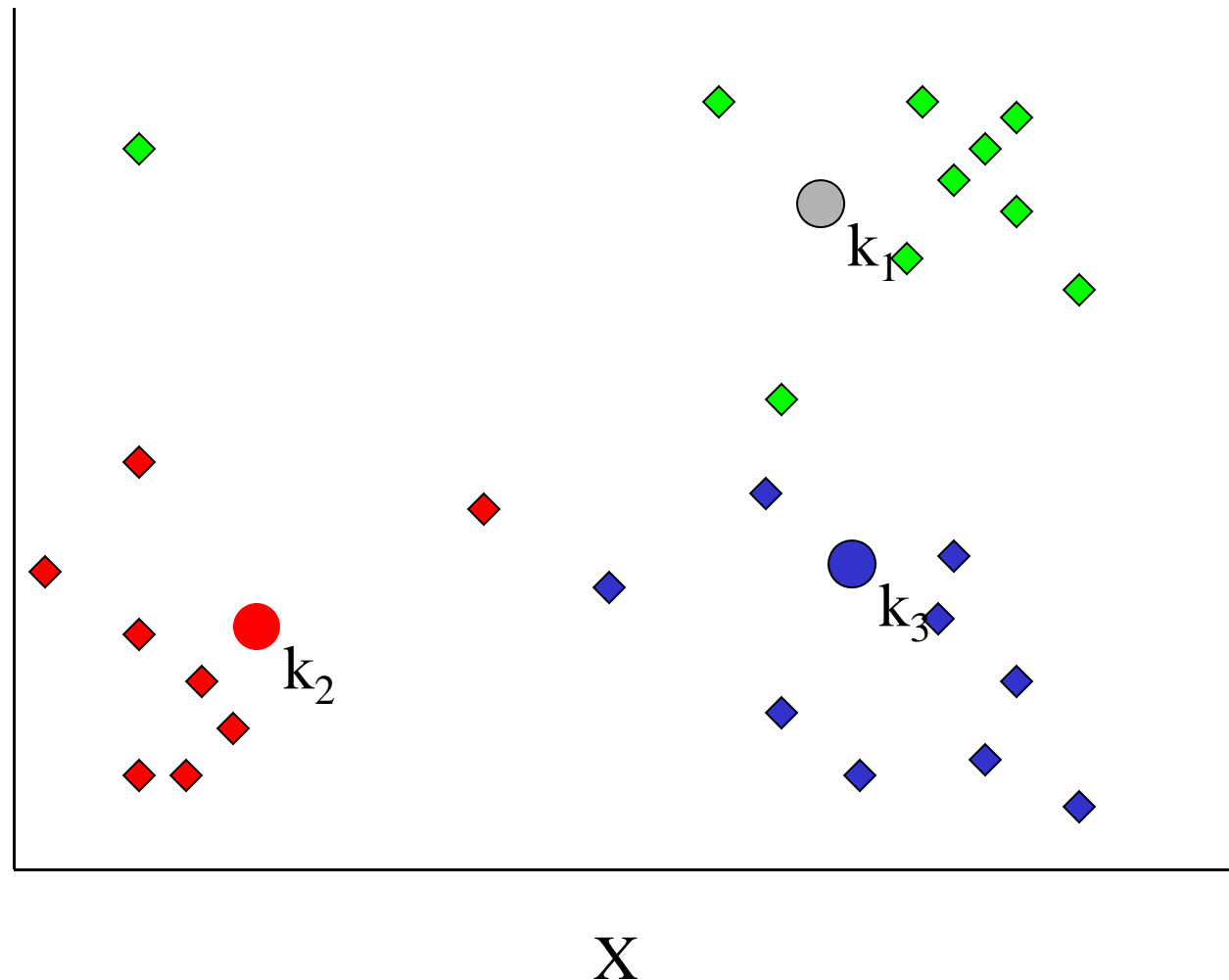
Move
each cluster
center
to the mean
of each cluster



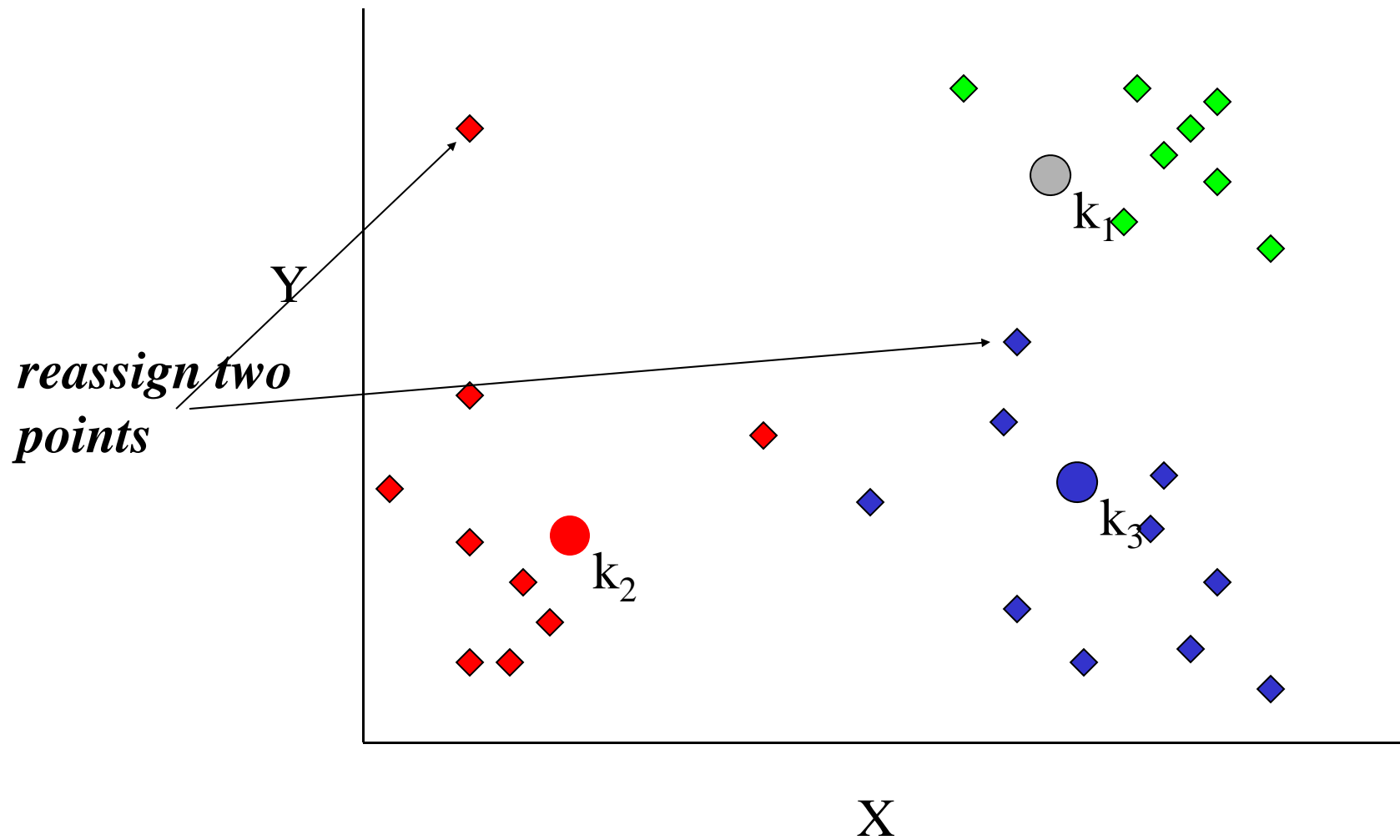
K-means, Step 4

Reassign
points
closest to a
different new
cluster center

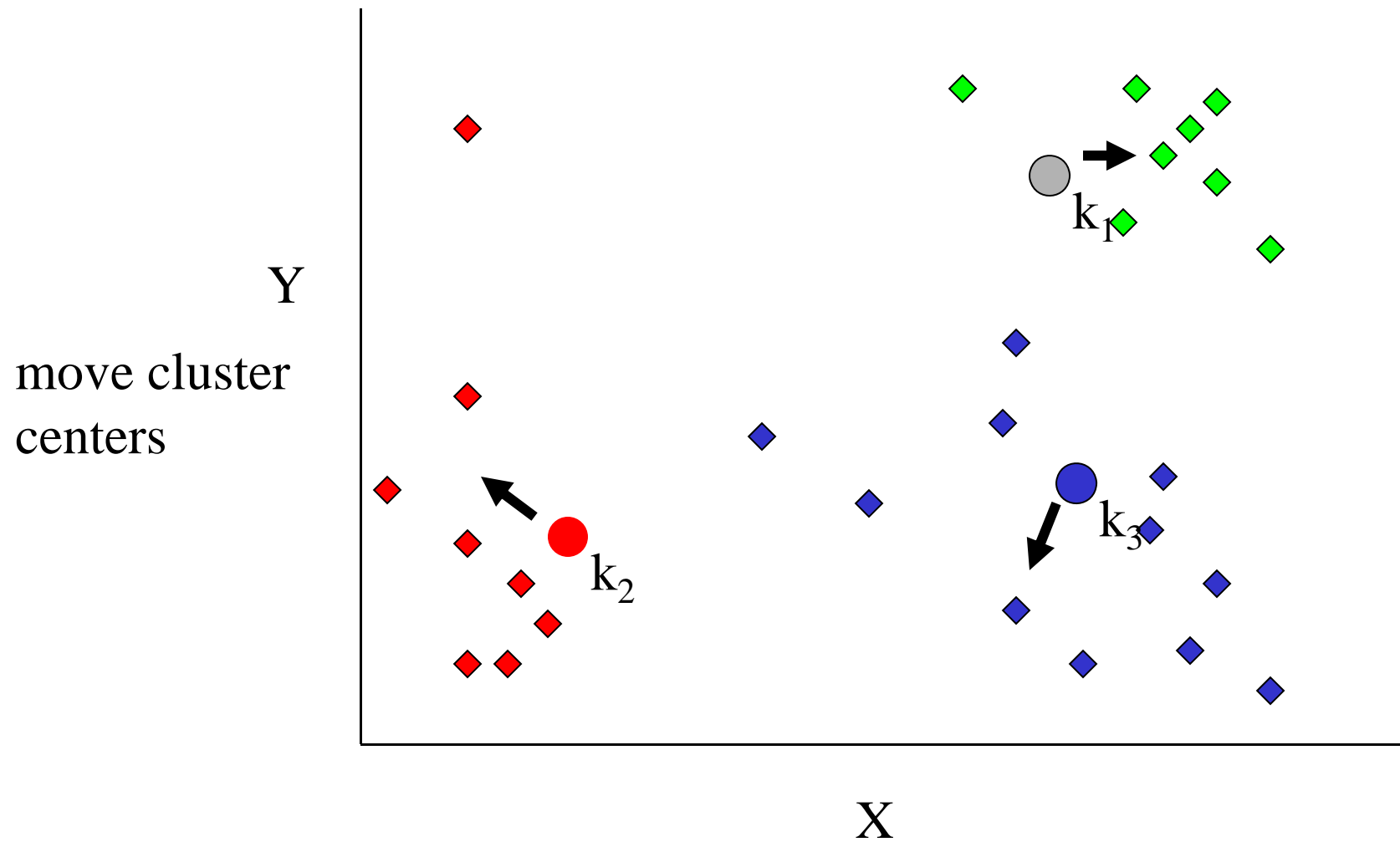
*Q: Which
points are
reassigned?*



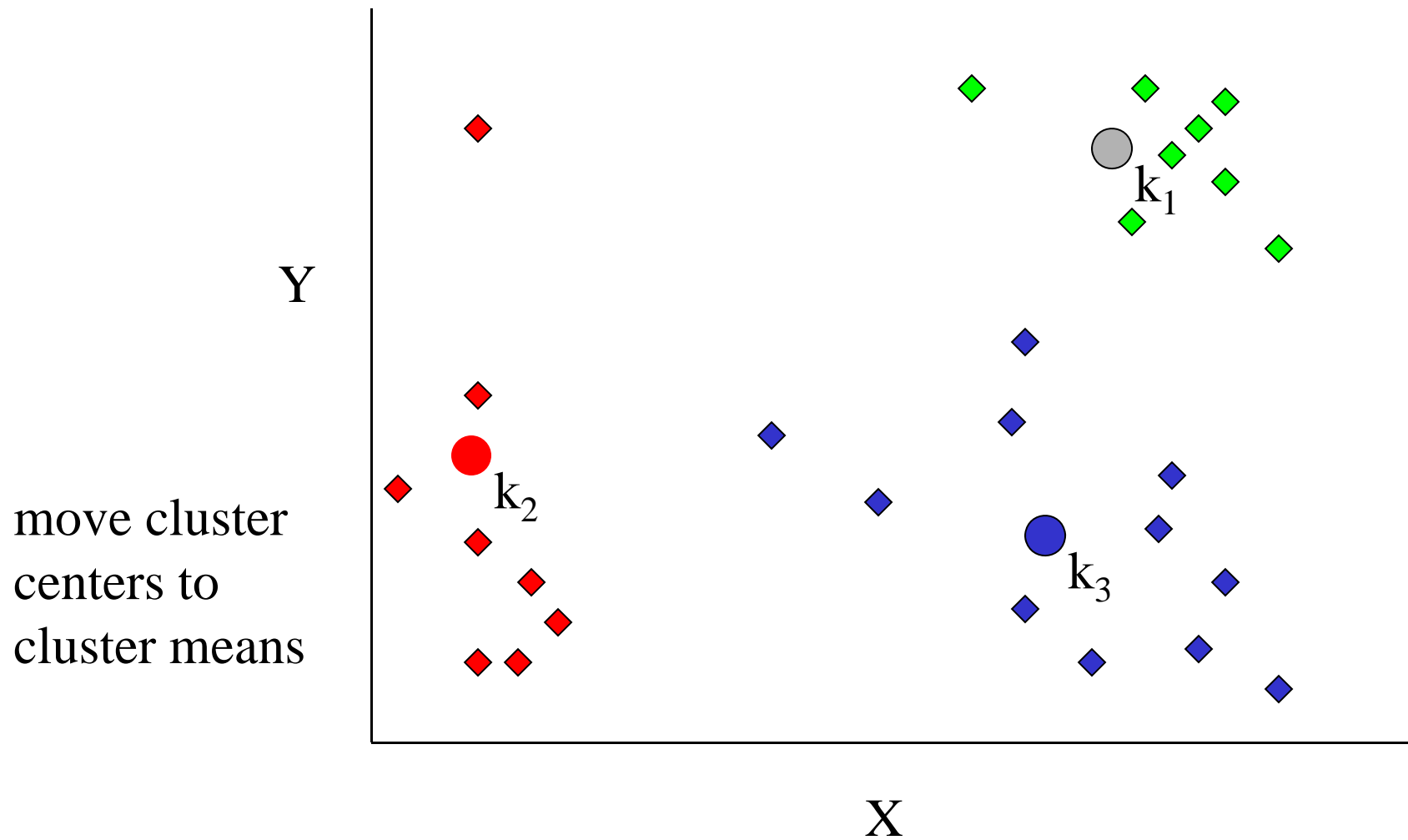
K-means, Step 5



K-means, Step 6



K-means, Step 7



Association rule discovery: definition

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Given a set of records each of which contain some number of items from a given collection : Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

Association Rule Discovery for Marketing and Sales Promotion

Let the rule discovered be

{Bagels, ... } --> {Potato Chips}

- Potato Chips on Right-Hand Side => Can be used to determine what should be done to boost its sales.
- Bagels on the Left-Hand Side => Can be used to see which products would be affected if the store discontinues selling bagels.
- Bagels on left *and* Potato chips on right => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

Supermarket shelf management. Goal: To identify items that are bought together by sufficiently many customers.

Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.

— A classic rule:

- If a customer buys diaper and milk, then very likely to buy beer.
- So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

Inventory Management. Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.

Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery

Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong **sequential dependencies** among different events.

$$(A \ B) \ (C) \longrightarrow (D \ E)$$

Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.

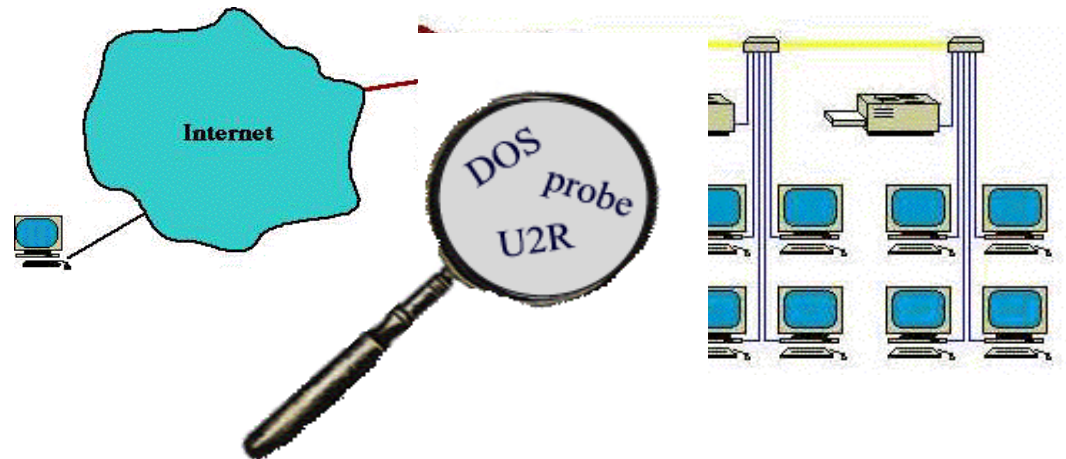
Deviation/Anomaly Detection

Detect significant deviations from normal behavior

- Credit Card Fraud Detection



- Network Intrusion Detection



(Some) Data-mining concerns:

How do we deal with missing data?

What about data that is stored in different physical locations?

Can we always find the best solution? Do we have to?

What about the quality of the data...is everything we know correct?

How can we determine whether what we found out is going to be useful?

Do we have to keep the information we extracted confidential?

Is there a standard approach to data mining that could be used as a step-by-step guide?

Can we somehow analyze data that takes on different forms (eg text or images) as a whole?

How can we cope with the large amount of data available? How do we determine what part of that data is relevant/useful for data mining?



<http://www.eecs.uc.edu/~mazlack/dbm.sp2007/dbm.cartoons.html>