USN □□□□□□□□□□

13MCA442

Fourth Semester MCA Degree Examination, June/July 2017

Data Warehousing and Data Mining

Time: 3 hrs. Max. Marks:100

Important Note : 1. On completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages.
2. Any revealing of identification, appeal to evaluator and /or equations written eg, 42+8 = 50, will be treated as malpractice.

Note: ***Answer any FIVE full questions.***

1 a. What is a Data Warehouse? List down the differences between operational database systems and data warehouses. **(06 Marks)**

b. With a neat diagram, explain in detail about Three – tier data warehousing architecture. **(06 Marks)**

c. Discuss in detail about Star, Snowflake and Fact constellation schemas in detail. **(08 Marks)**

2 a. Define Data Mining. Explain the challenges that motivated the development of Data Mining. **(06 Marks)**

b. With a neat diagram, explain the Process of knowledge Discovery in Databases. **(04 Marks)**

c. Discuss Data Mining tasks in detail. **(10 Marks)**

3 a. Describe in detail about various types of data sets. **(10 Marks)**

b. What is Data Preprocessing? Explain the following techniques in detail : **(10 Marks)**
i) Sampling ii) Dimensionality Reduction iii) Discretization and Binarization.

4 a. Write down the Apriori principle and explain the Pseudo code for the frequent itemset generation part of the Apriori algorithm. **(07 Marks)**

b. Write a detailed note on Maximal frequent itemsets and Closed frequent itemsets. **(05 Marks)**

c. Discuss FP Growth algorithm in detail. **(08 Marks)**

5 a. What is a Decision Tree? Write an algorithm for Decision Tree induction. **(07 Marks)**

b. Explain Sequential Covering Algorithm in detail. **(06 Marks)**

c. Discuss K – nearest neighbour classification algorithm with characteristics of Nearest neighbour classifiers. **(07 Marks)**

6 a. Discuss in detail about various techniques for improving the accuracy of classification methods. **(07 Marks)**

b. Explain in detail about various evaluation criteria for classification methods. **(06 Marks)**

c. Describe Multiclass Problem in detail. **(07 Marks)**

7 a. What is Cluster Analysis? Explain Agglomerative clustering method in detail. **(06 Marks)**

b. Discuss K – means method in detail, with an example. **(08 Marks)**

c. Describe DBSCAN method in detail. **(06 Marks)**

8 a. What are Outliers? Explain statistical approaches in detail. **(10 Marks)**

b. Discuss Clustering – based approaches in detail. **(10 Marks)**

*****

VTU Question Paper- Even Semester 2017

Data Ware Housing and Data Mining 13MCA442

1a. What is a Data Warehouse? List down the differences between operational database systems and data warehouses. 06 Marks

Ans: *A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process.*

Important to note subject-oriented, integrated, and time-variant properties of a data warehouse.

Subject-oriented:

- A DW is organized around major subjects, such as student, degree, country.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations.

A DW provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated:

- A DW may be constructed by integrating information from multiple data sources e.g. multiple OLTP databases.
- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.

Time Variant:

- A DW usually has long time horizon, significantly longer than that of operational systems.
    - Operational database: current value data.
    - DW data: provide information from a historical perspective (e.g. past 5-10 years)
- Every key structure in the DW contains an element of time, explicitly or implicitly
- Operational data may or may not contain time element.

Non-volatile:

- A physically separate store of data transformed from the operational environment.
- No update of data
- Does not require transaction processing, recovery, and concurrency control mechanisms
- Requires only two operations in data accessing: *initial loading of data* and *access of data*.

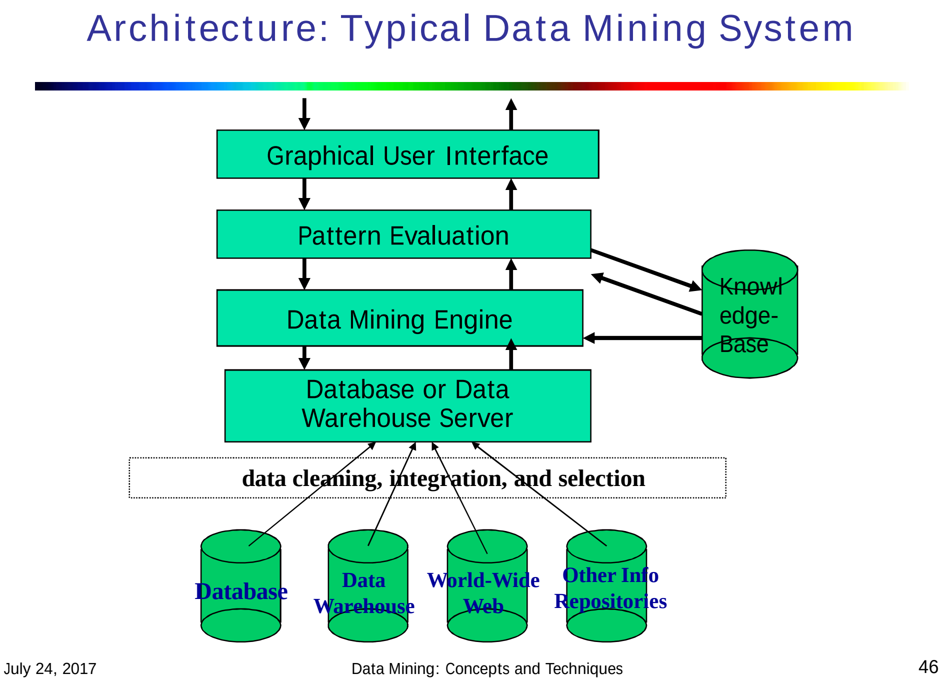
Difference between DW and ODS

ODS	**DW**
Data of high quality at detailed level and assured availability	Data may not be perfect, but sufficient for strategic analysis; data does not have to be highly available
Contains current and near-current data	Contains historical data
Real-time and near real-time data loads	Normally batch data loads
Mostly updated at data field level(even if it may be appended)	Data is appended, not updated
Typically detailed data only	Contains summarized and detailed data
Modeled to support rapid data updates(3NF)	Variety of modeling techniques used, typically multidimensional for data marts to

	optimize query performance
Transactions similar to those in OLTP systems	Complex queries processing larger volumes of data
Used for detailed decision making and operational reporting	Used for long-term decision making and management reporting
Used at the operational level	Used at the managerial level

1b. With a neat diagram, explain in detail about Three-tier data warehousing architecture. 06 Marks

Ans:



Architecture: Typical Data Mining System

July 24, 2017 Data Mining: Concepts and Techniques 46

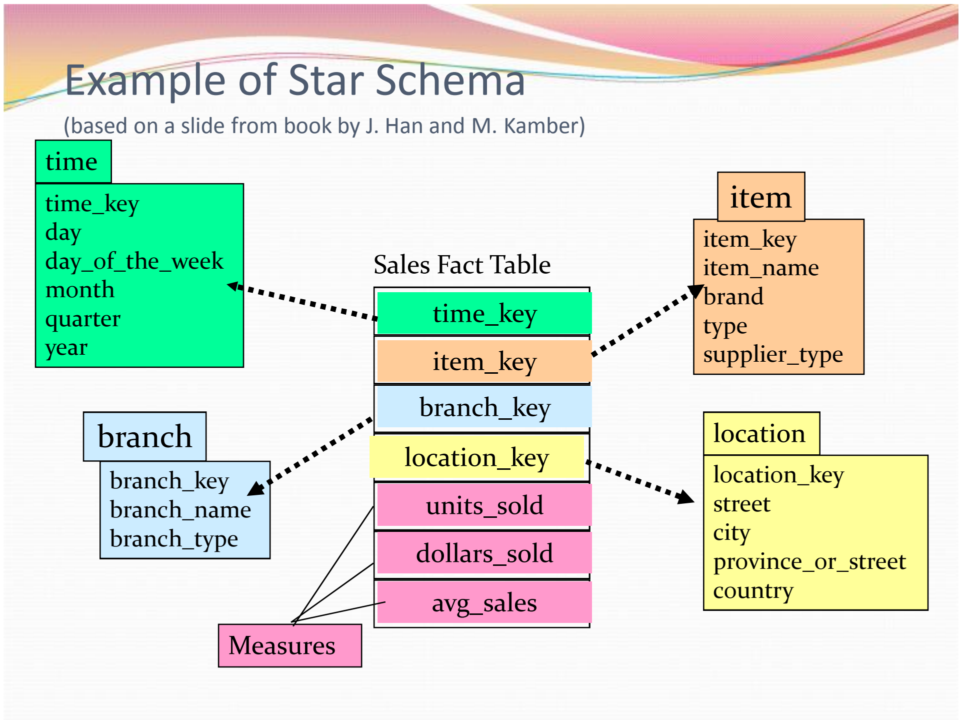
Data warehouses normally adopt three-tier architecture: 1. The bottom tiers is a warehouse database server that is almost always a relational database stsyem.Data from operational databases and from external sources are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to execute code. 2. The middle tier is an OLAP server that is typically implemented using a relational OLAP (ROLAP) model. 3. The top tier is a client, which contains query and and reporting tools, analysis tools and/or data mining tools. From the architecture point of view there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.

1c. Discuss in detail about Star, Snowflake and Fact constellation schemas in detail. 08 Marks

Ans: Data Warehouse Design:

One approach is the *star schema* to represent the multidimensional data model. The schema in this model consists of a large single fact table containing the bulk of the data, with no redundancy and a set of smaller tables called dimension table, one for each dimension.

Other models have been used. These include *snowflakes model* and *fact constellations model*.



Example of Star Schema

(based on a slide from book by J. Han and M. Kamber)

time
time_key
day
day_of_the_week
month
quarter
year

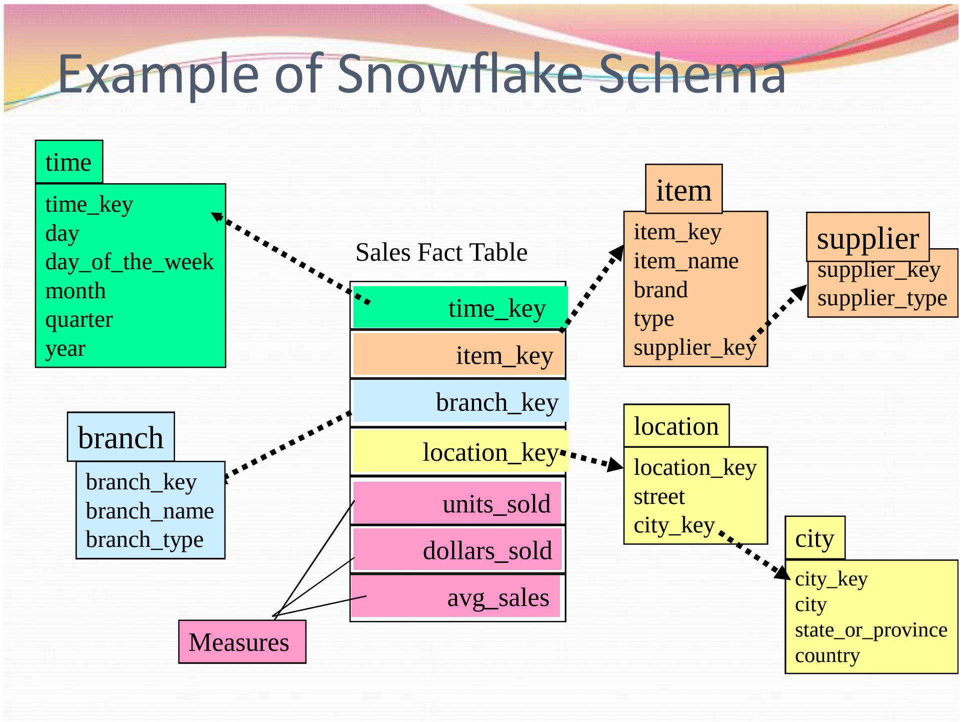
Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item
item_key
item_name
brand
type
supplier_type

branch
branch_key
branch_name
branch_type

location
location_key
street
city
province_or_street
country

Measures



Example of Snowflake Schema

time
time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item
item_key
item_name
brand
type
supplier_key

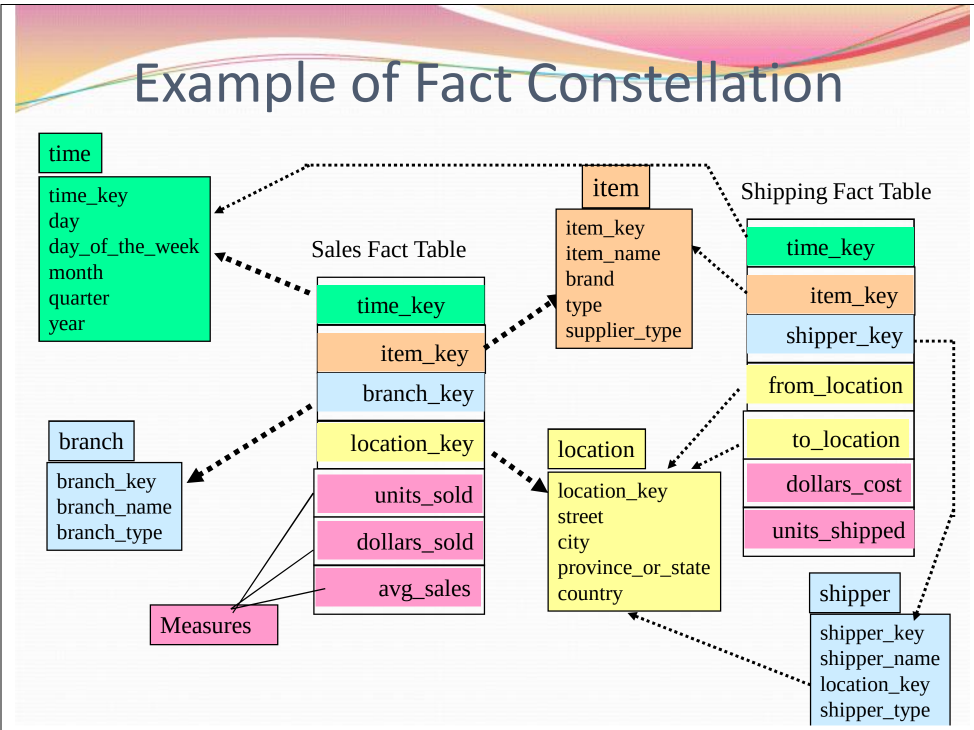
supplier
supplier_key
supplier_type

branch
branch_key
branch_name
branch_type

location
location_key
street
city_key

city
city_key
city
state_or_province
country

Measures



Example of Fact Constellation

time
time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

branch
branch_key
branch_name
branch_type

item
item_key
item_name
brand
type
supplier_type

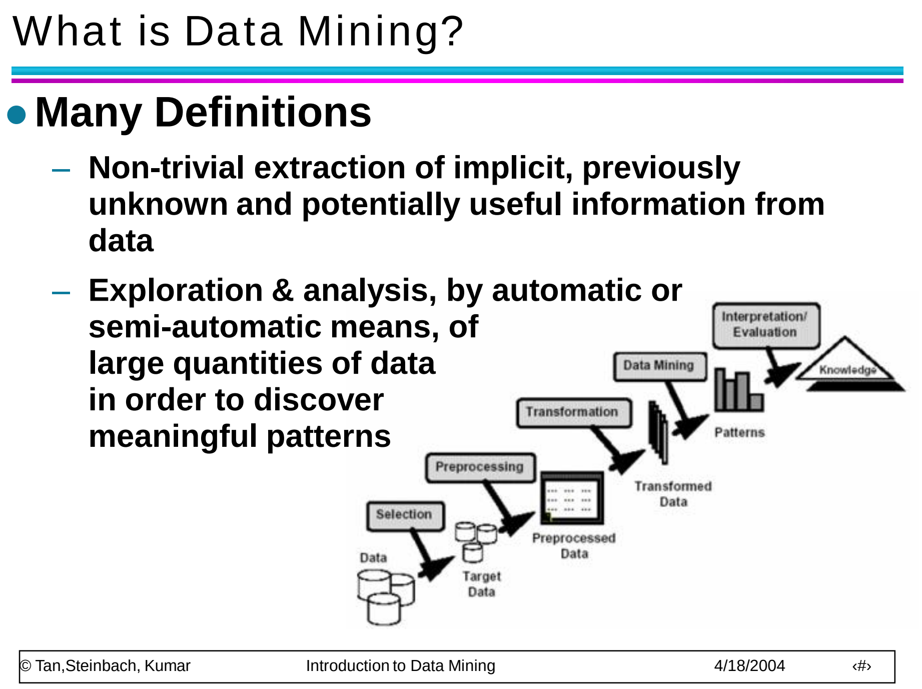
location
location_key
street
city
province_or_state
country

Shipping Fact Table
time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

shipper
shipper_key
shipper_name
location_key
shipper_type

2a. Define Data Mining. Explain the challenges that motivated the development of Data Mining. 06 Marks

Ans.



What is Data Mining?

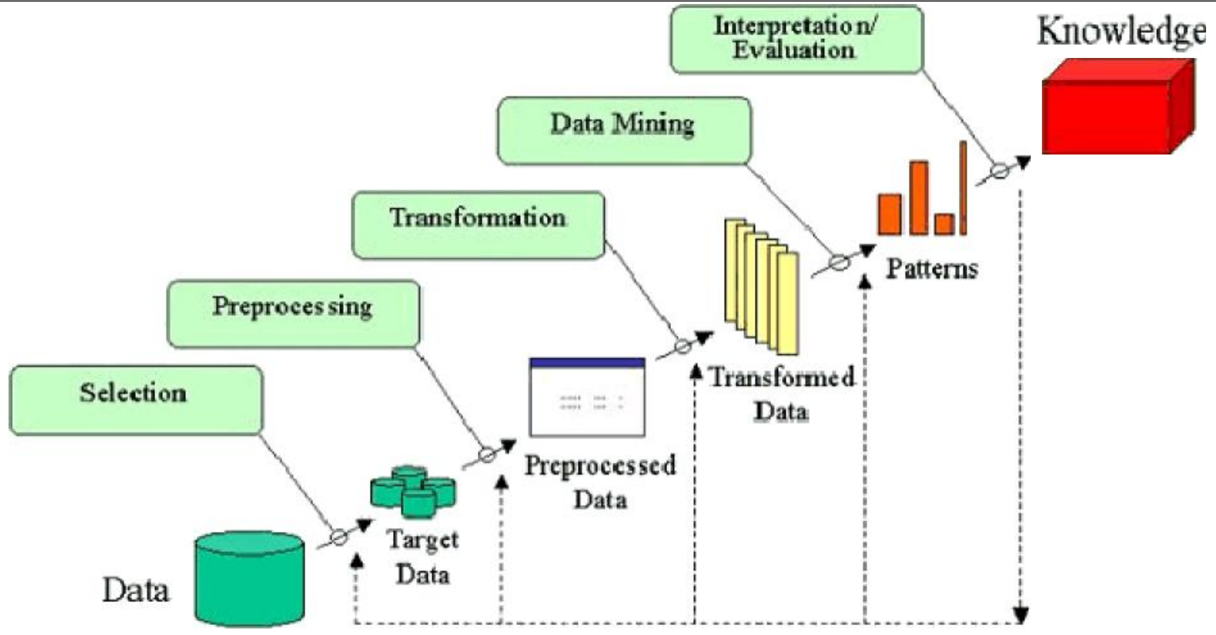
- **Many Definitions**
  - **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**
  - **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**

Interpretation/ Evaluation
Knowledge
Data Mining
Patterns
Transformation
Transformed Data
Preprocessing
Preprocessed Data
Selection
Data
Target Data

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

challenges that motivated the development of Data Mining:

1.Scalability

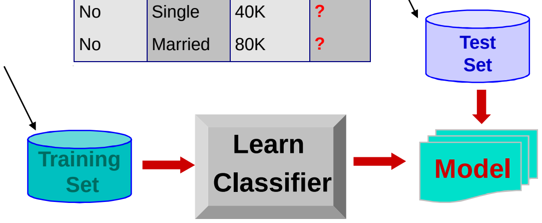
	If data mining algorithms are to handle these massive data sets, then they must be scalable. 2. High Dimensionality For some data analysis algorithms, the computational complexity increases rapidly as the dimensionality increases. 3. Heterogeneous and Complex Data Dealing with data with not the same type. 4. Data Ownership and Distribution Data is geographically distributed among resources belonging to multiple entities. 5. Non-traditional Analysis The traditional statistical approach is based on a hypothesize-and-test paradigm. Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.	
2b.	With a neat diagram, explain the process of knowledge Discovery in Databases.	04 Marks
Ans:	 Interpretation/ Evaluation Knowledge Data Mining Transformation Patterns Preprocessing Transformed Data Selection Preprocessed Data Target Data Data Here is the list of steps involved in the knowledge discovery process − • Data Cleaning − In this step, the noise and inconsistent data is removed.	

	• Data Integration − In this step, multiple data sources are combined. • Data Selection − In this step, data relevant to the analysis task are retrieved from the database. • Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. • Data Mining − In this step, intelligent methods are applied in order to extract data patterns. • Pattern Evaluation − In this step, data patterns are evaluated. • Knowledge Presentation − In this step, knowledge is represented.	
2c.	Discuss Data Mining tasks in detail.	10 Marks
Ans:	**Data Mining Tasks:** • Classification [Predictive] • Clustering [Descriptive] • Association Rule Discovery [Descriptive] • Sequential Pattern Discovery [Descriptive] • Regression [Predictive] • Deviation Detection [Predictive] **Classification: Definition:** • Given a collection of records (*training set*) o 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. o 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

	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	**?**



Clustering Definition

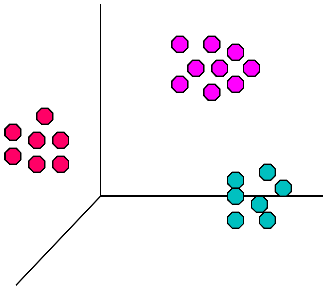
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - – Data points in one cluster are more similar to one another.
  - – Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - – Euclidean Distance if attributes are continuous.
  - – Other Problem-specific Measures.

Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Association Rule Discovery: Definition

- 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.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

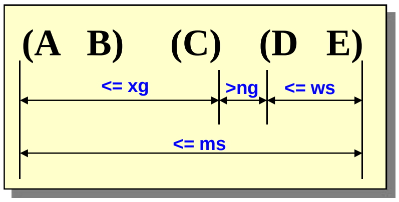
Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Sequential Pattern Discovery: Definition

- 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.

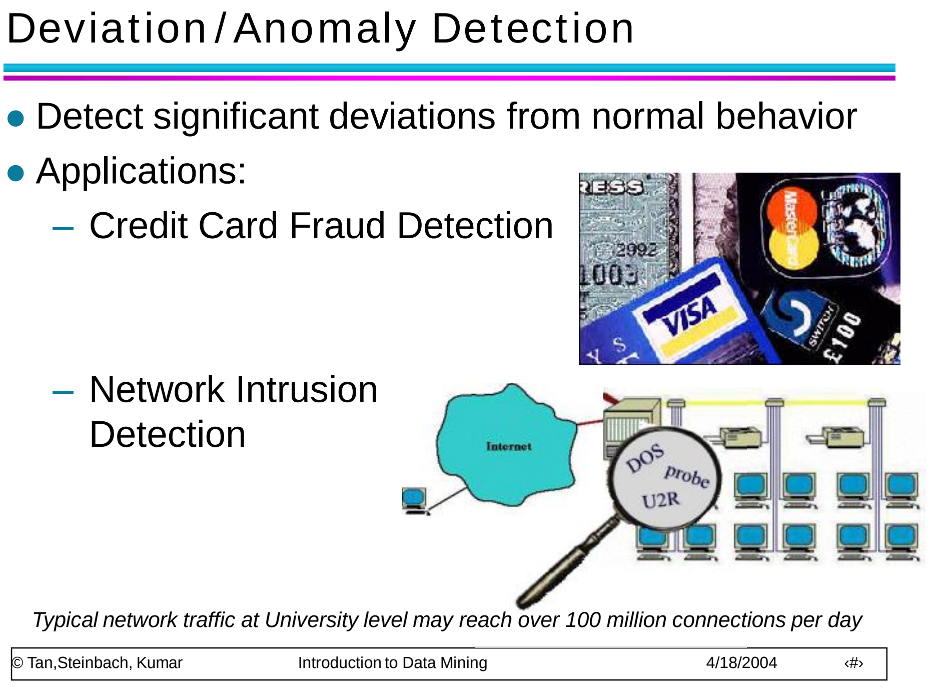
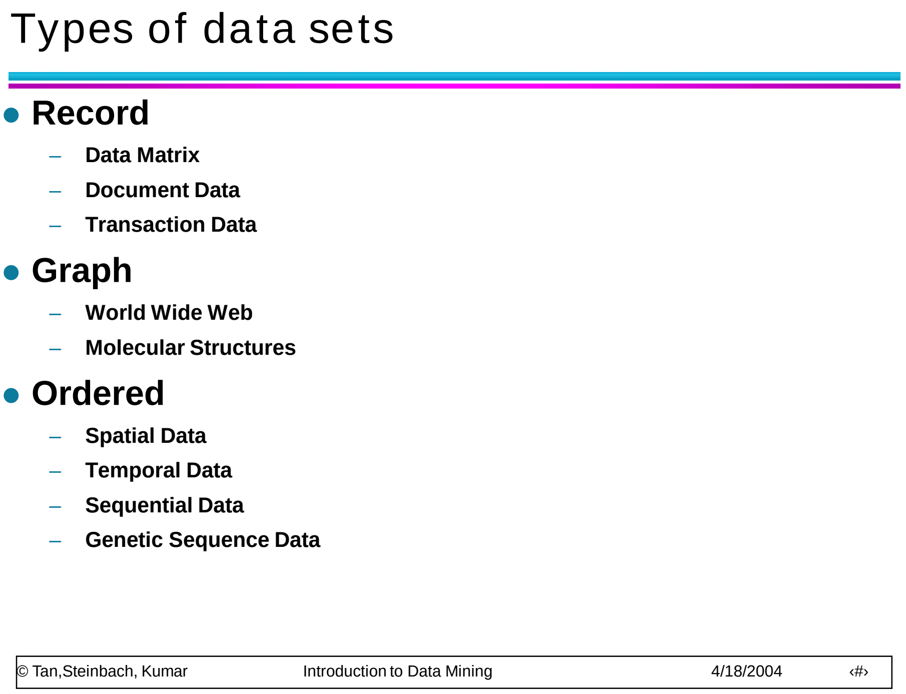
(A B) (C) ⟶ (D E)

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



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.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

	 Internet DOS probe U2R VISA MasterCard SWITCH £100 **Deviation/Anomaly Detection** ● Detect significant deviations from normal behavior ● Applications: – Credit Card Fraud Detection – Network Intrusion Detection *Typical network traffic at University level may reach over 100 million connections per day* © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
3a.	Describe in detail about various types of data sets.	10 Marks
Ans:	 **Types of data sets** ● **Record** – **Data Matrix** – **Document Data** – **Transaction Data** ● **Graph** – **World Wide Web** – **Molecular Structures** ● **Ordered** – **Spatial Data** – **Temporal Data** – **Sequential Data** – **Genetic Sequence Data** © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

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**

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a 'term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

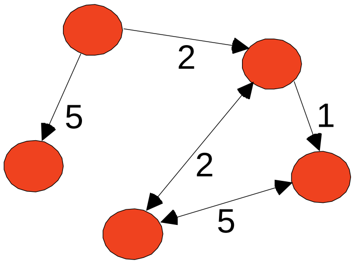
Transaction Data or Market Basket Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Graph Data

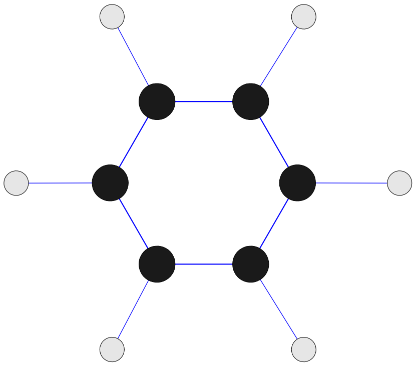
- Examples: Generic graph and HTML Links



```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

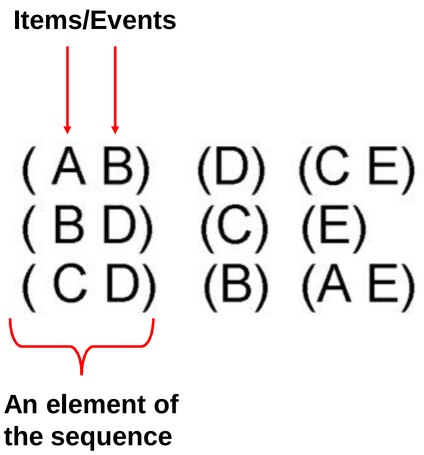
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- Sequences of transactions



Items/Events

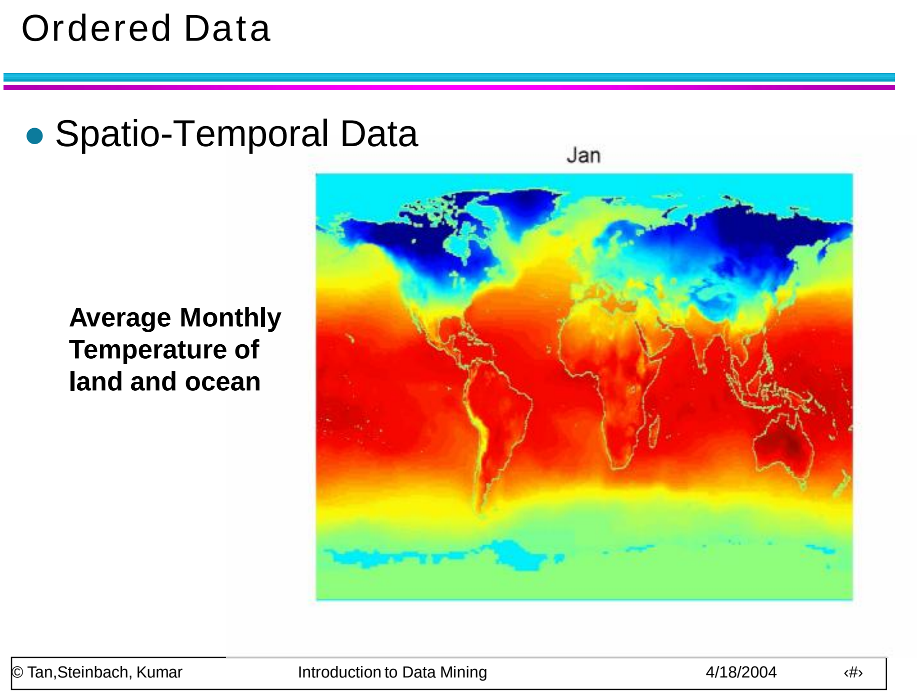
(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of
the sequence

Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

	 Ordered Data ● Spatio-Temporal Data Jan **Average Monthly Temperature of land and ocean** © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
3b.	What is data preprocessing? Explain the following techniques in detail: (i) Sampling (ii) Dimensionality Reduction (iii) Discretization and Binarization	10 Marks
Ans:	**Data Preprocessing:** **Data preprocessing** is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. **Data preprocessing** is a proven method of resolving such issues. Techniques involved: • Aggregation • Sampling • Dimensionality Reduction • Feature subset selection • Feature creation • Discretization and Binarization • Attribute Transformation **Sampling:** The key principle for effective sampling is the following: – using a sample will work almost as well as using the entire data sets, if the sample is representative – A sample is representative if it has approximately the same property (of interest) as the original set of data **Types of Sampling:** • Simple Random Sampling	

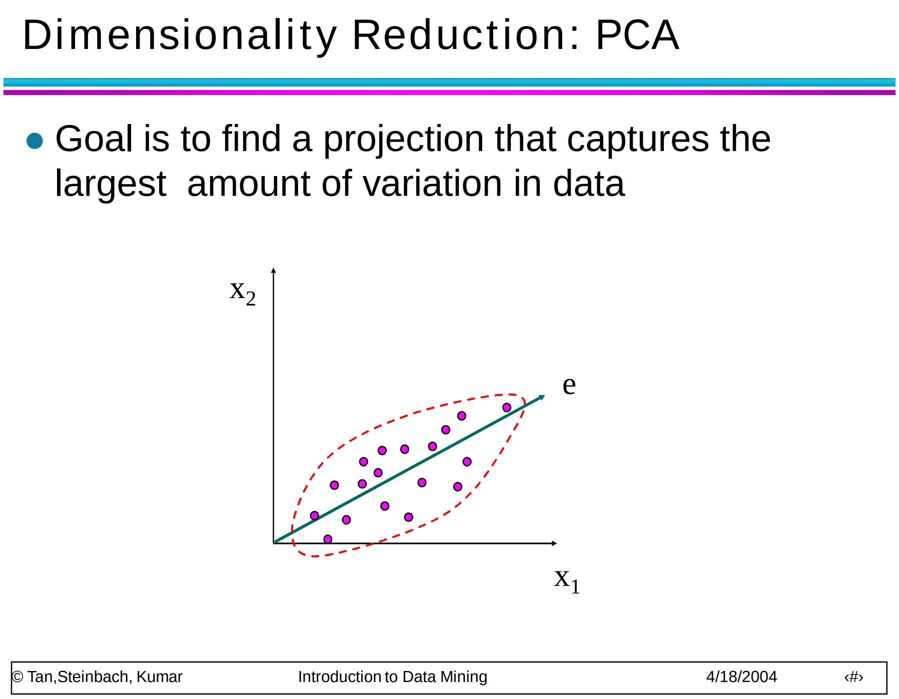
  - There is an equal probability of selecting any particular item
  - Two types
- Sampling without replacement
  - As each item is selected, it is removed from the population
      - Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
      - Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Dimensionality Reduction:

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

Techniques

  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques



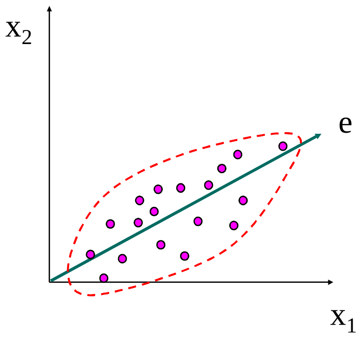
Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

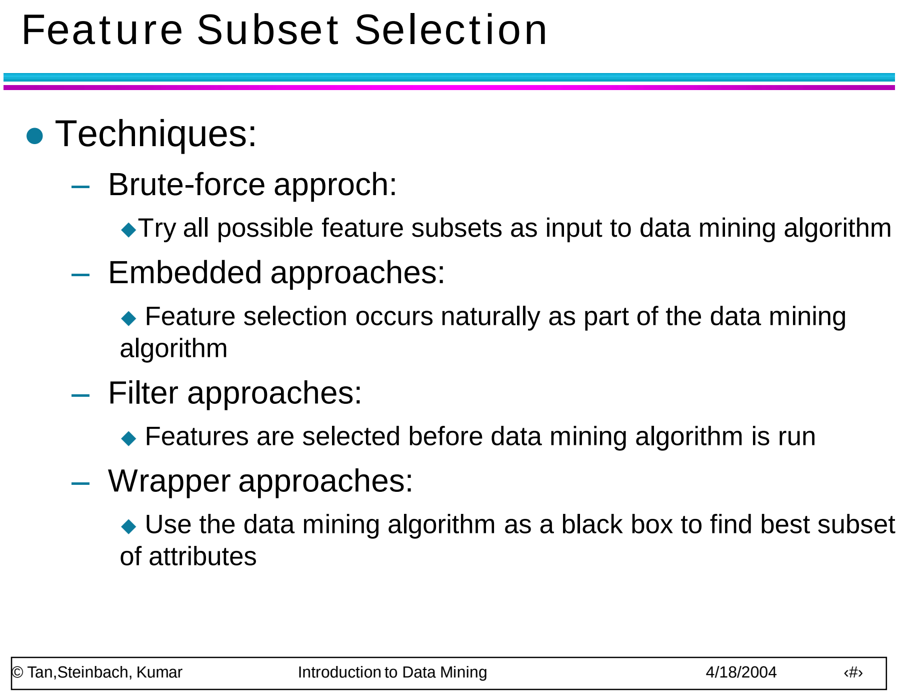
Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space



Feature Subset Selection

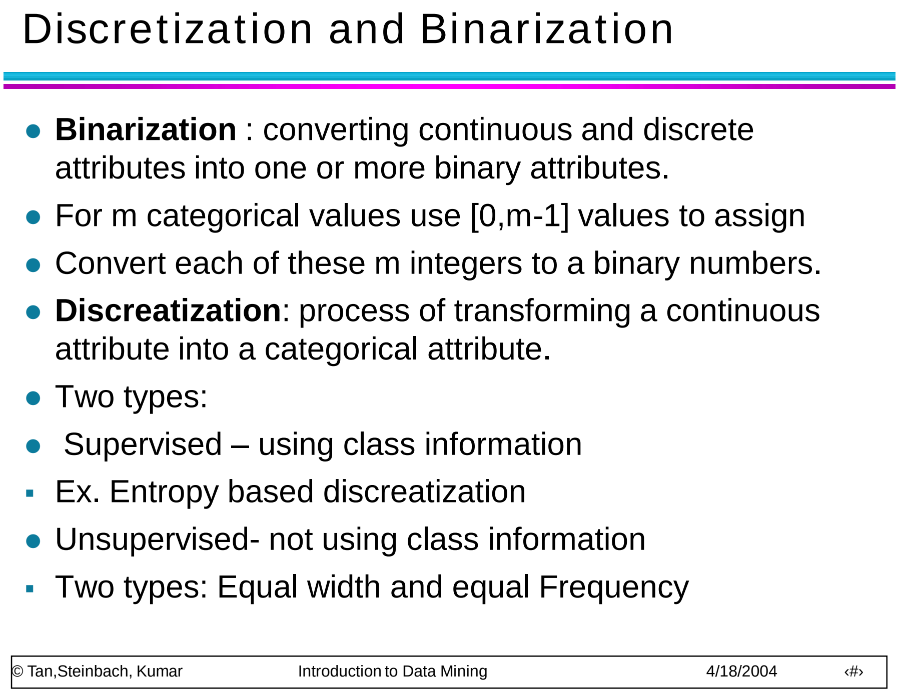
- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA



Feature Subset Selection

- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



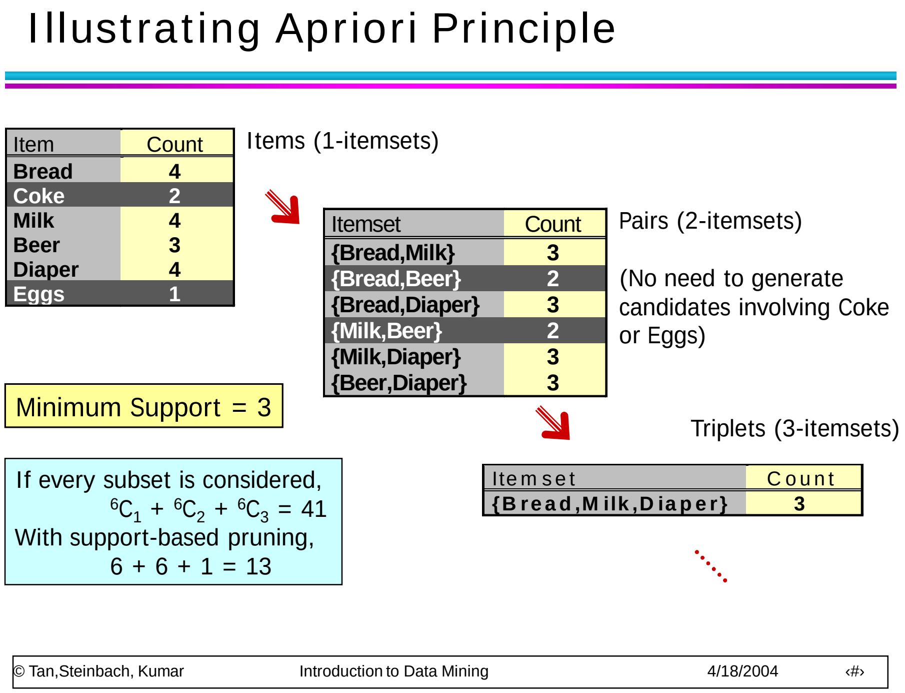
Discretization and Binarization

- **Binarization** : converting continuous and discrete attributes into one or more binary attributes.
- For m categorical values use [0,m-1] values to assign
- Convert each of these m integers to a binary numbers.
- **Discreatization**: process of transforming a continuous attribute into a categorical attribute.
- Two types:
- Supervised – using class information
- Ex. Entropy based discreatization
- Unsupervised- not using class information
- Two types: Equal width and equal Frequency

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

4a.	Write down the Apriori principle and explain the Pseudo code for the frequent itemset generation part of the Apriori algorithm.	07 Marks
Ans:	Apriori principle: – If an itemset is frequent, then all of its subsets must also be frequent Apriori principle holds due to the following property of the support measure: – Support of an itemset never exceeds the support of its subsets	

– This is known as the anti-monotone property of support



Illustrating Apriori Principle

Items (1-itemsets)

Item	Count
Bread	**4**
Coke	**2**
Milk	**4**
Beer	**3**
Diaper	**4**
Eggs	**1**

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Itemset	Count
{Bread,Milk}	**3**
{Bread,Beer}	**2**
{Bread,Diaper}	**3**
{Milk,Beer}	**2**
{Milk,Diaper}	**3**
{Beer,Diaper}	**3**

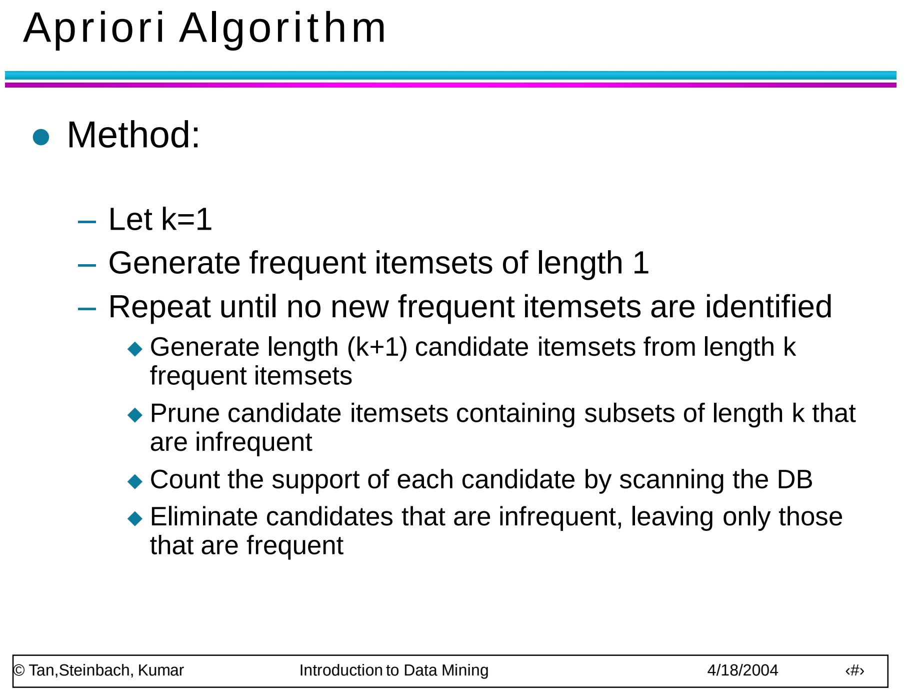
Minimum Support = 3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	**3**

If every subset is considered,
$^6C_1 + {}^6C_2 + {}^6C_3 = 41$
With support-based pruning,
6 + 6 + 1 = 13

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



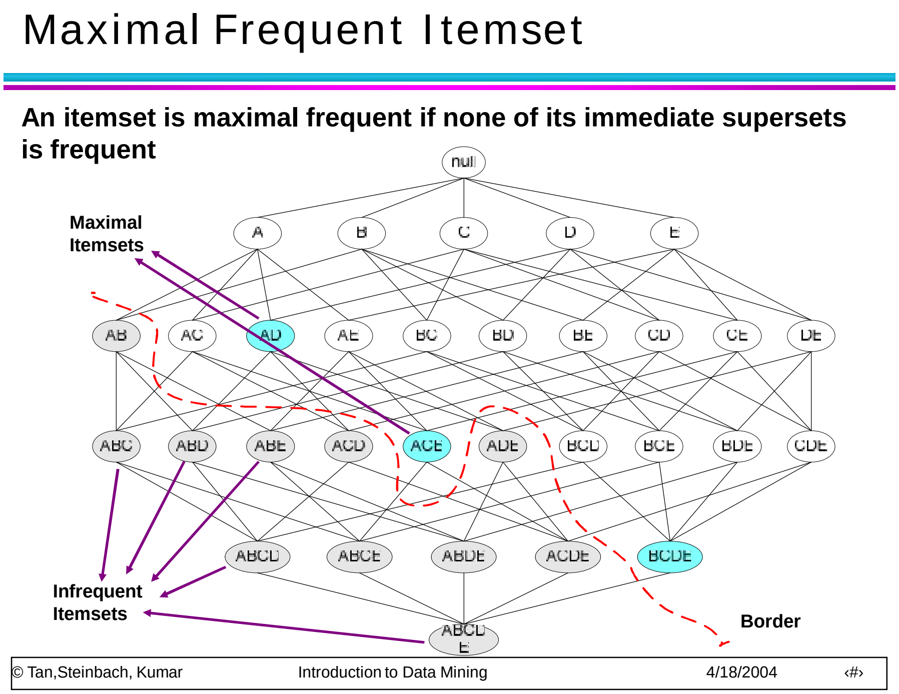
Apriori Algorithm

- Method:
  - Let k=1
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length (k+1) candidate itemsets from length k frequent itemsets
    - Prune candidate itemsets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

4b.	Write a detailed note on Maximal frequent itemsets and Closed frequent itemsets.	05 Marks

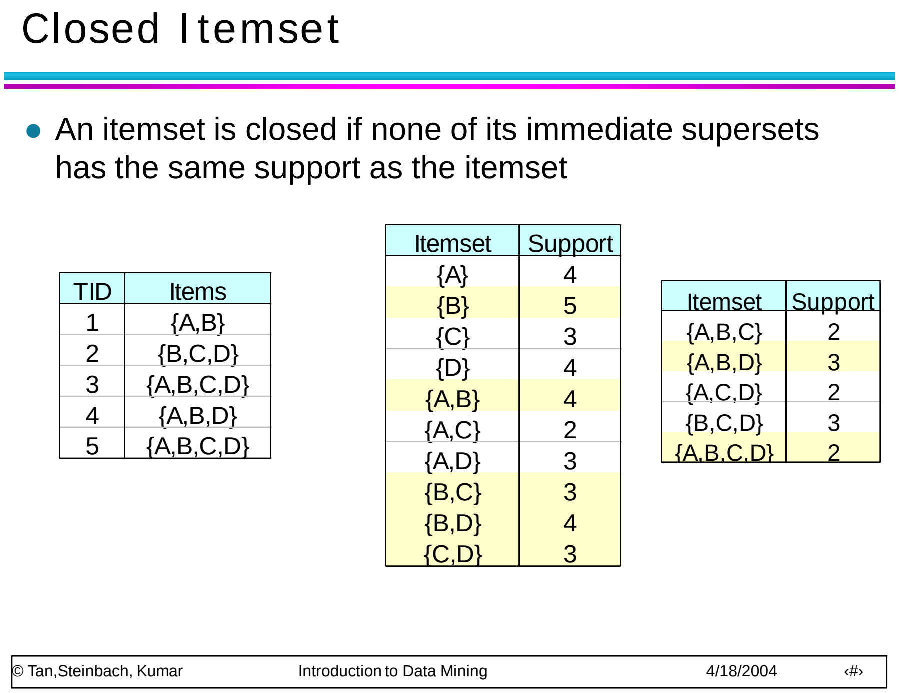
Ans:



Maximal Frequent Itemset

An itemset is maximal frequent if none of its immediate supersets is frequent

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the itemset

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

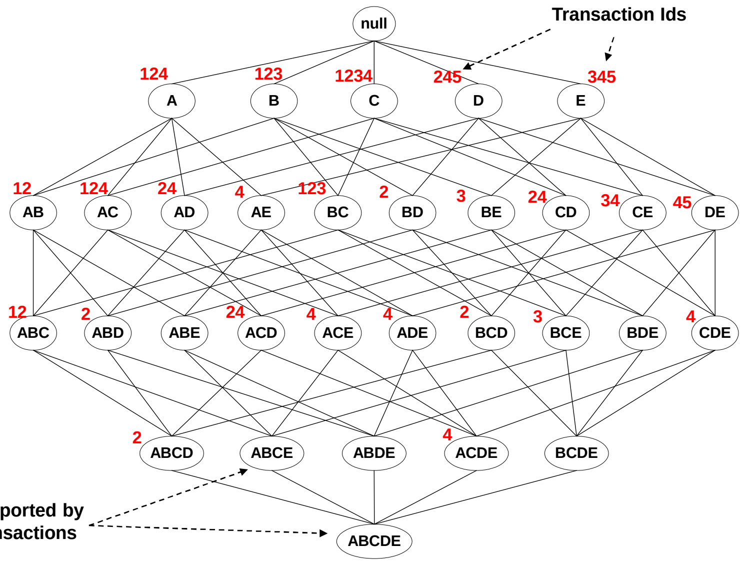
Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

Maximal vs Closed Itemsets

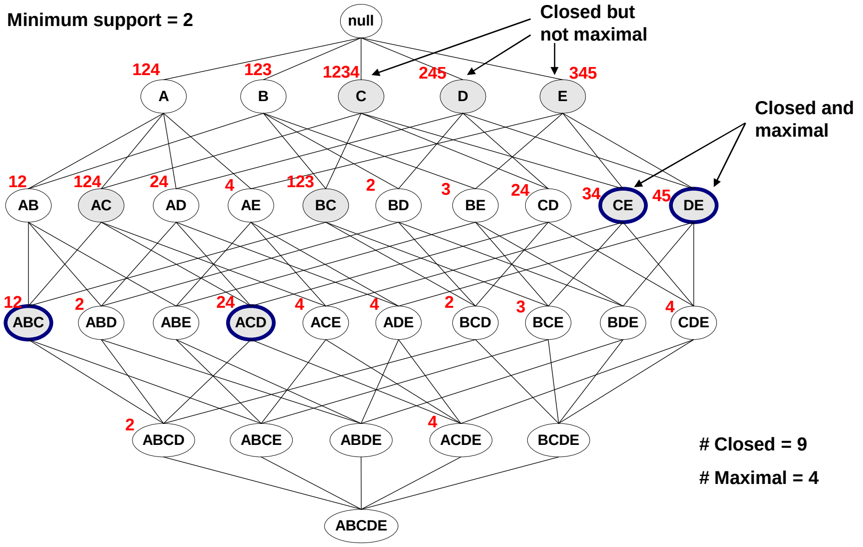
TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



Transaction Ids

Not supported by any transactions

Maximal vs Closed Frequent Itemsets



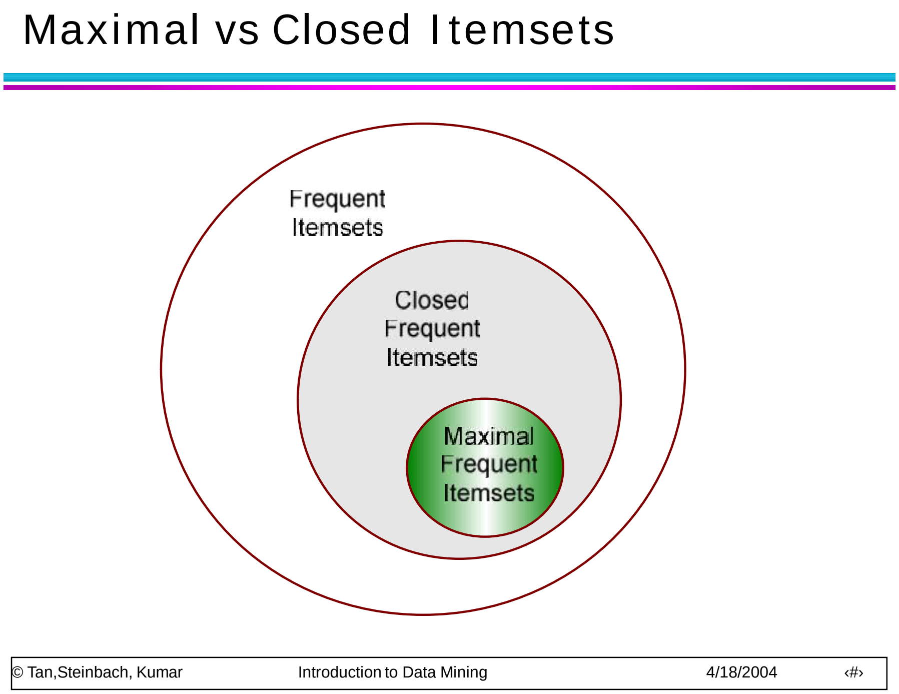
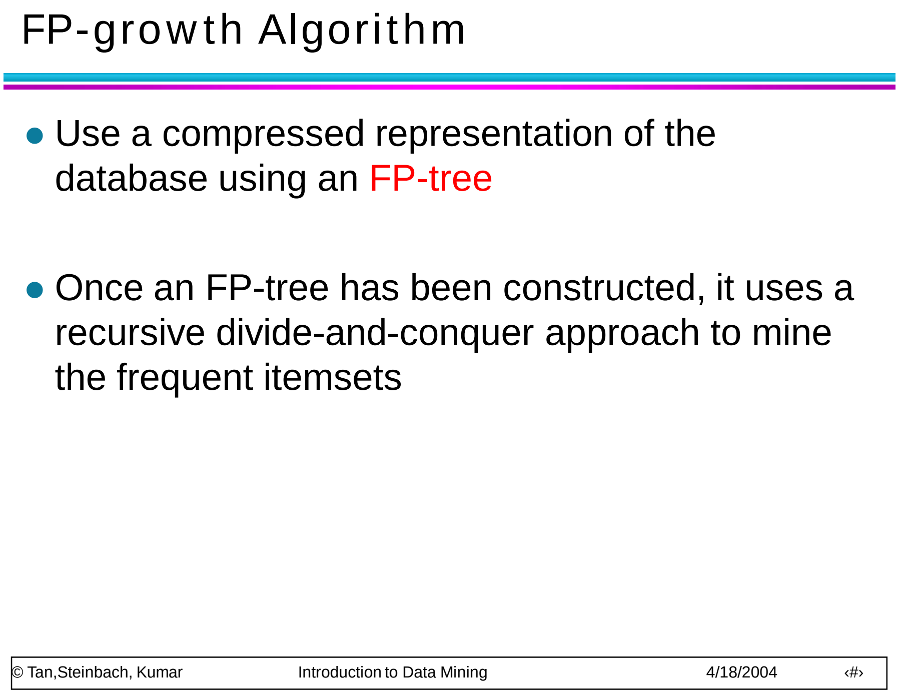
Minimum support = 2

Closed but not maximal

Closed and maximal

# Closed = 9

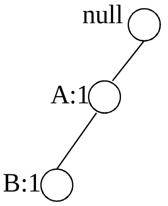
# Maximal = 4

	 Maximal vs Closed Itemsets Frequent Itemsets Closed Frequent Itemsets Maximal Frequent Itemsets © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
4c.	Discuss FP Growth algorithm in detail.	08 Marks
Ans:	 FP-growth Algorithm ● Use a compressed representation of the database using an FP-tree ● Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	

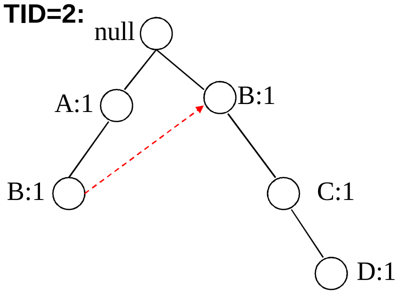
FP-tree construction

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

After reading TID=1:



After reading TID=2:



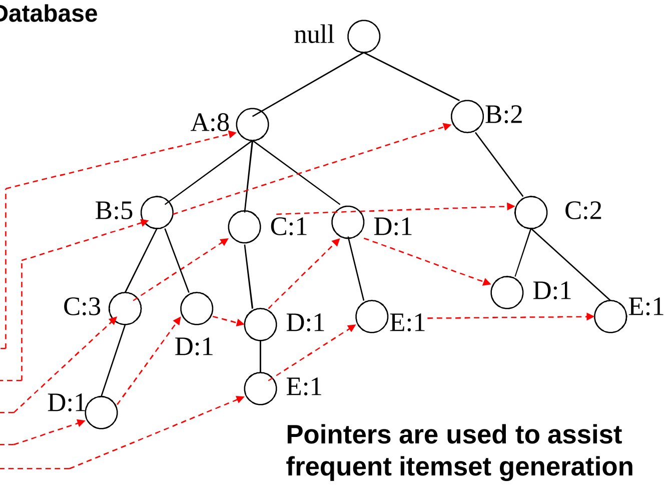
FP-Tree Construction

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{A}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

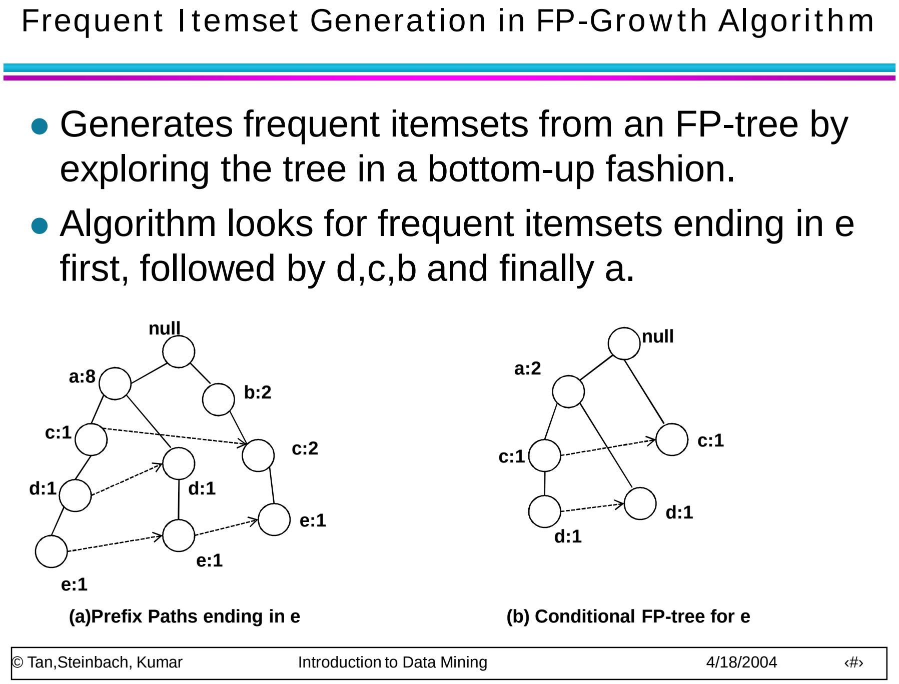
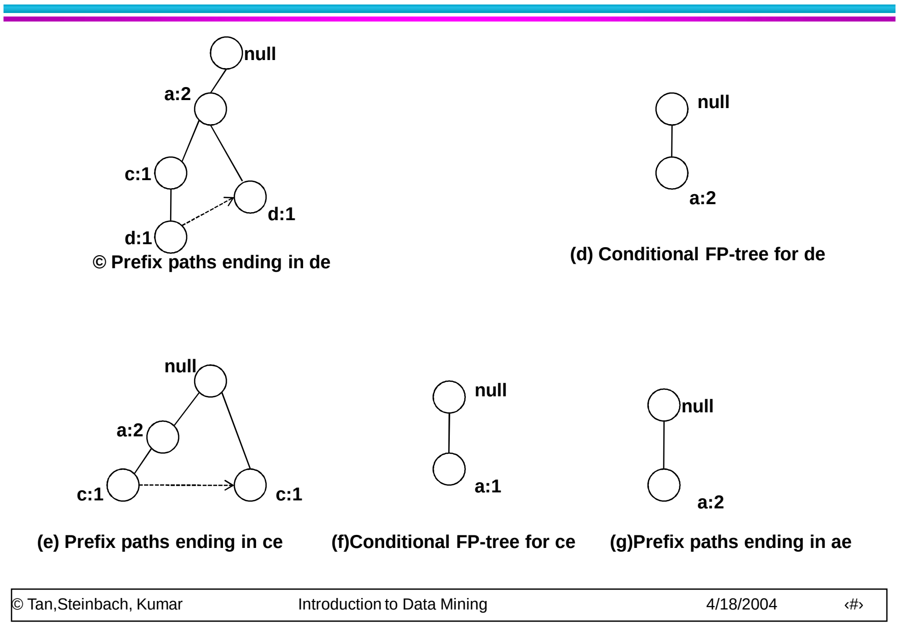
Transaction Database

Header table

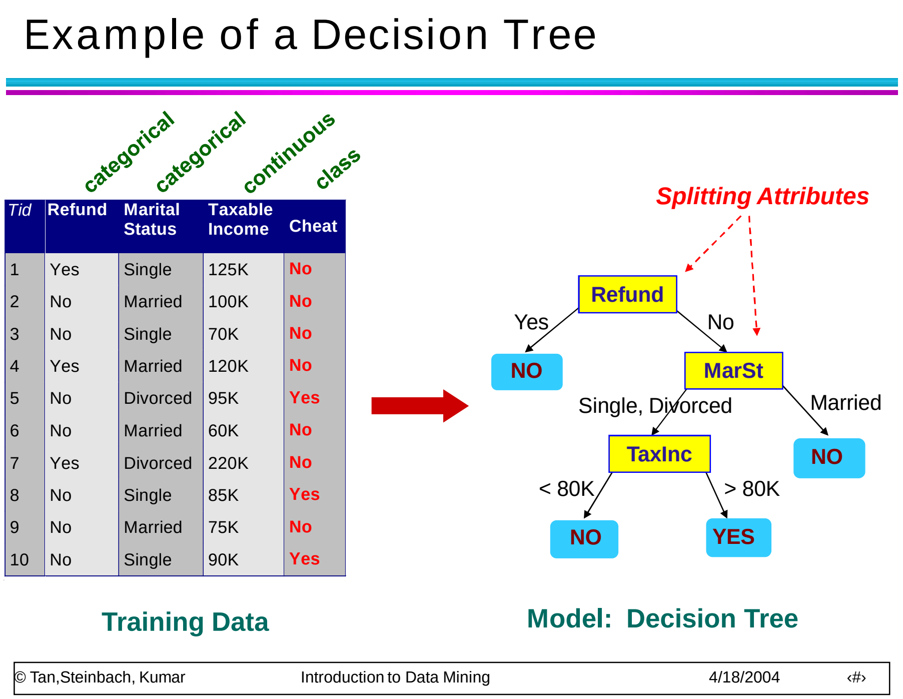
Item	Pointer
A	
B	
C	
D	
E	



Pointers are used to assist frequent itemset generation

	 Frequent Itemset Generation in FP-Growth Algorithm ● Generates frequent itemsets from an FP-tree by exploring the tree in a bottom-up fashion. ● Algorithm looks for frequent itemsets ending in e first, followed by d,c,b and finally a. null a:8 b:2 c:1 c:2 d:1 d:1 e:1 e:1 e:1 null a:2 c:1 c:1 d:1 d:1 (a)Prefix Paths ending in e (b) Conditional FP-tree for e © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›  null a:2 c:1 d:1 d:1 null a:2 © Prefix paths ending in de (d) Conditional FP-tree for de null a:2 c:1 c:1 null a:1 null a:2 (e) Prefix paths ending in ce (f)Conditional FP-tree for ce (g)Prefix paths ending in ae © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
5a.	What is a Decision Tree? Write an algorithm for Decision Tree induction.	07 Marks
Ans:	Decision Tree is a classification technique which is used to assign class for a previously unseen record.	

Example of a Decision Tree

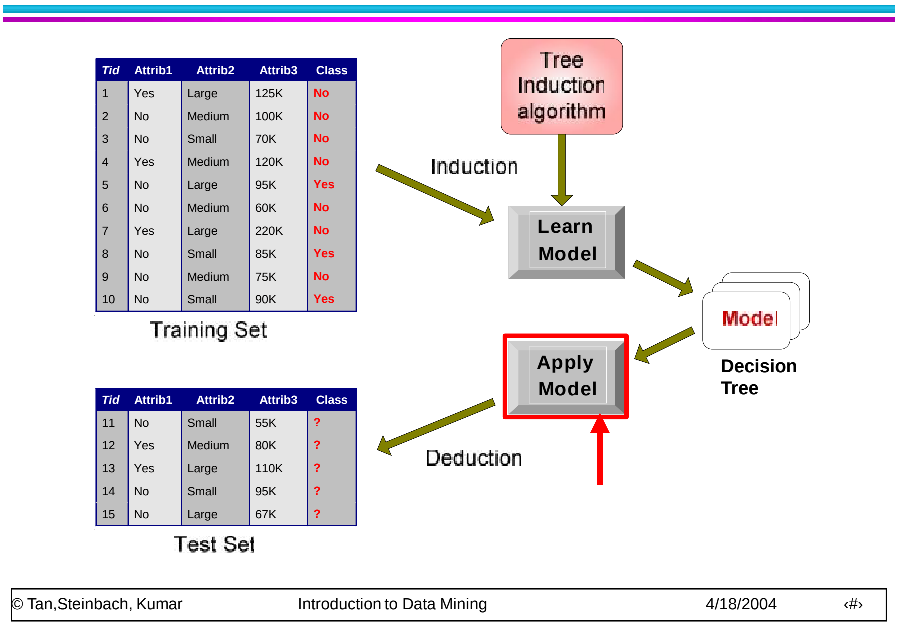


	categorical	categorical	continuous	class
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

Model: Decision Tree

Decision Tree Classification Task



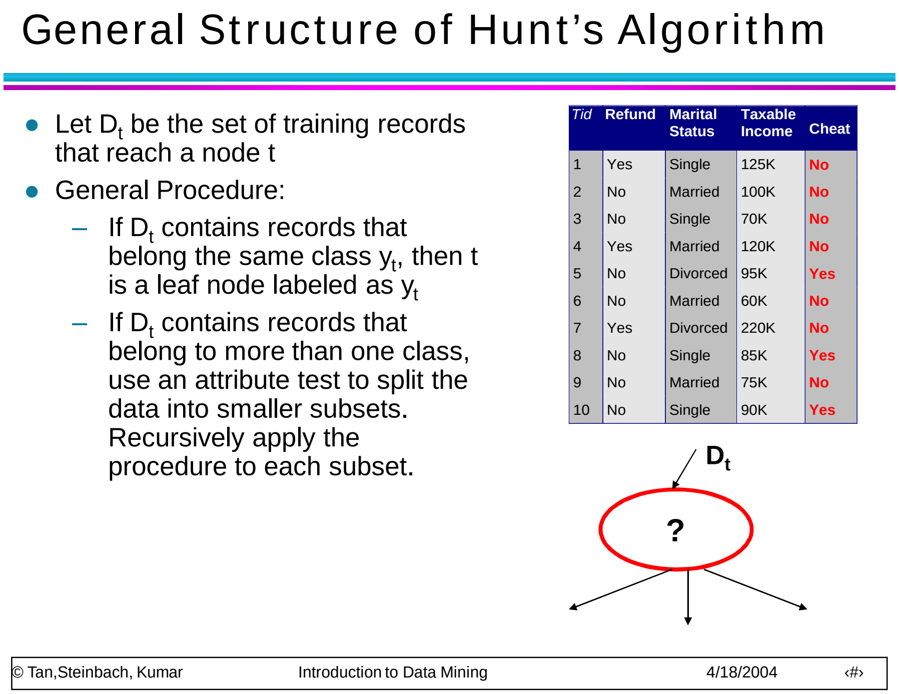
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

A decision tree induction algorithm:

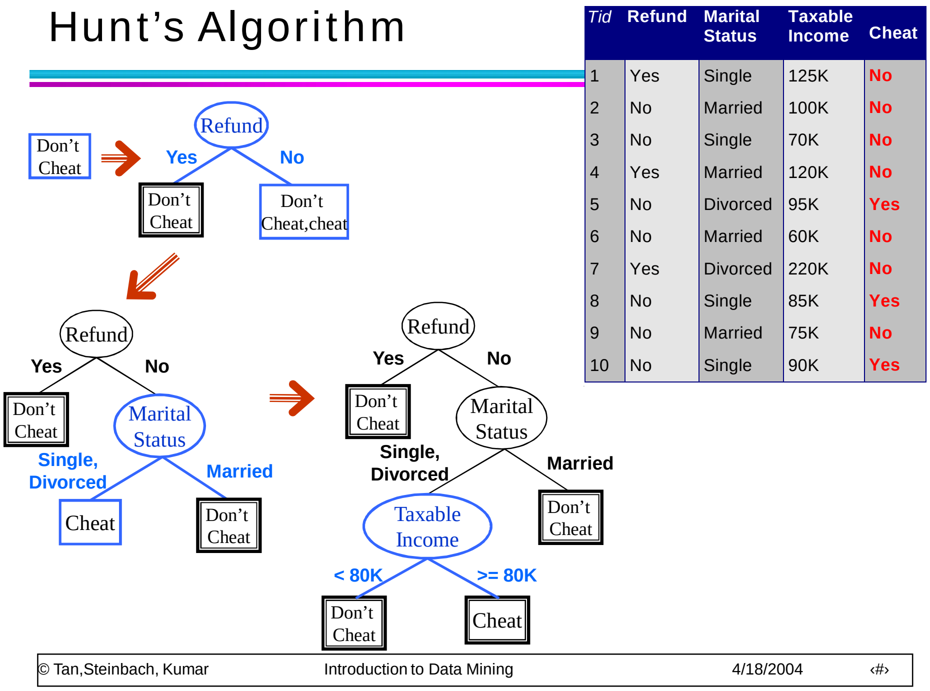


General Structure of Hunt’s Algorithm

- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

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

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

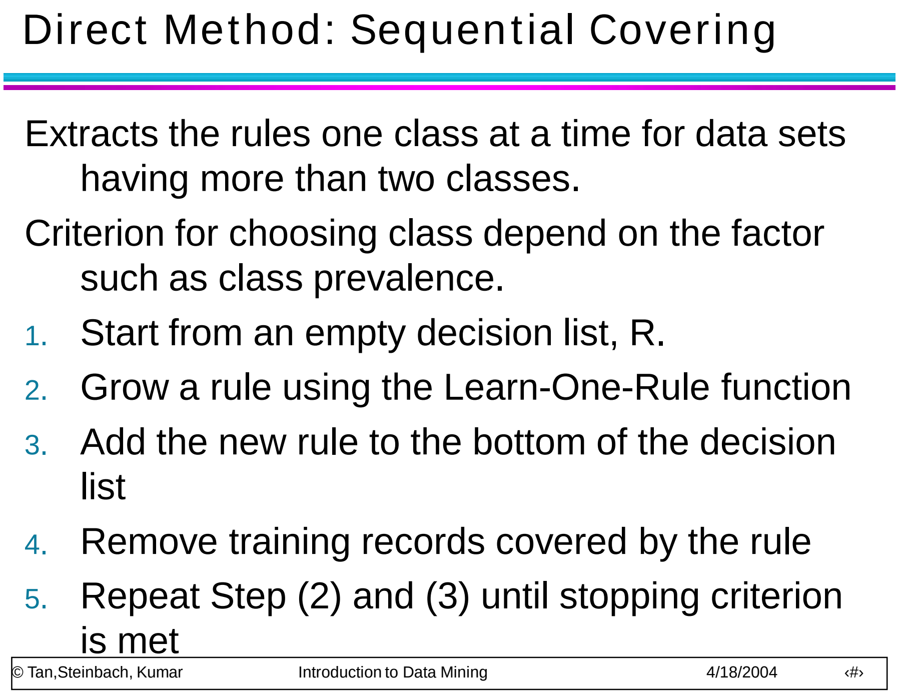


Hunt’s Algorithm

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

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

5b.	Explain Sequential Covering Algorithm in detail.	06 Marks
Ans:	Rule Generation :	



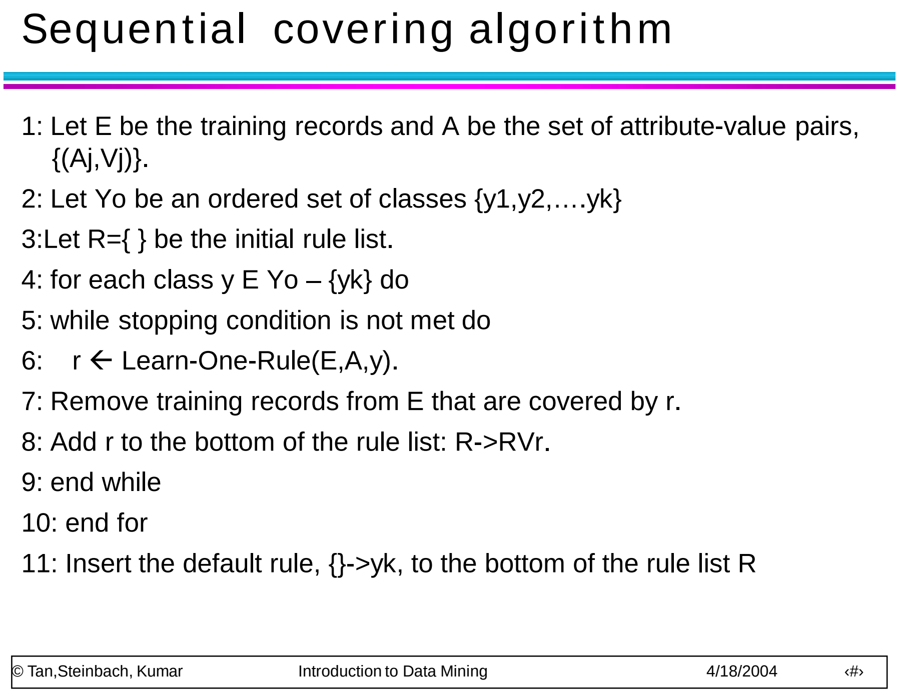
Direct Method: Sequential Covering

Extracts the rules one class at a time for data sets having more than two classes.

Criterion for choosing class depend on the factor such as class prevalence.

1. Start from an empty decision list, R.
2. Grow a rule using the Learn-One-Rule function
3. Add the new rule to the bottom of the decision list
4. Remove training records covered by the rule
5. Repeat Step (2) and (3) until stopping criterion is met

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



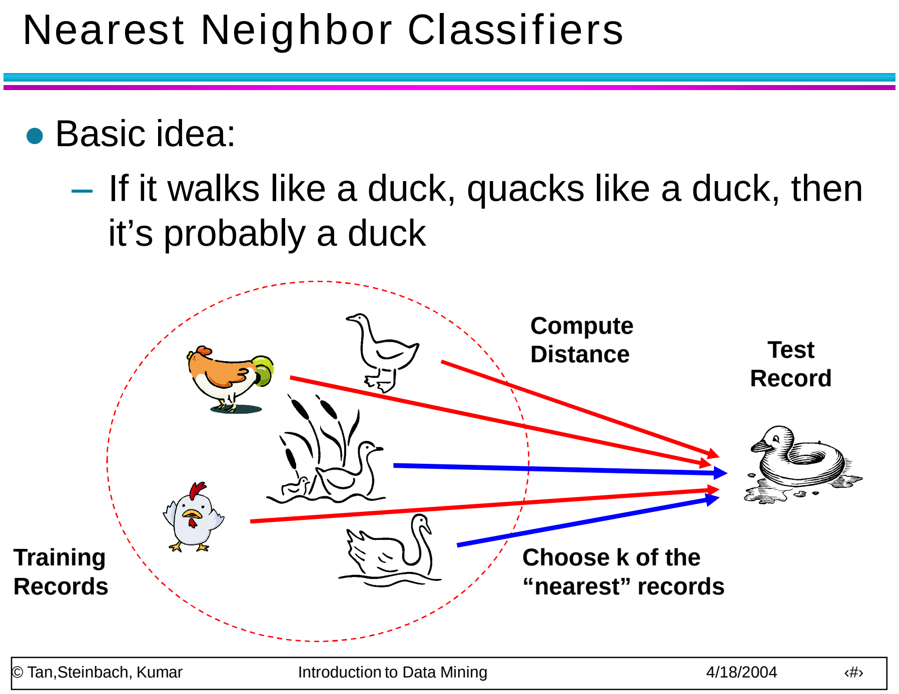
Sequential covering algorithm

```
1: Let E be the training records and A be the set of attribute-value pairs, {(Aj,Vj)}.
2: Let Yo be an ordered set of classes {y1,y2,….yk}
3:Let R={ } be the initial rule list.
4: for each class y E Yo – {yk} do
5: while stopping condition is not met do
6:    r ← Learn-One-Rule(E,A,y).
7: Remove training records from E that are covered by r.
8: Add r to the bottom of the rule list: R->RVr.
9: end while
10: end for
11: Insert the default rule, {}->yk, to the bottom of the rule list R
```

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

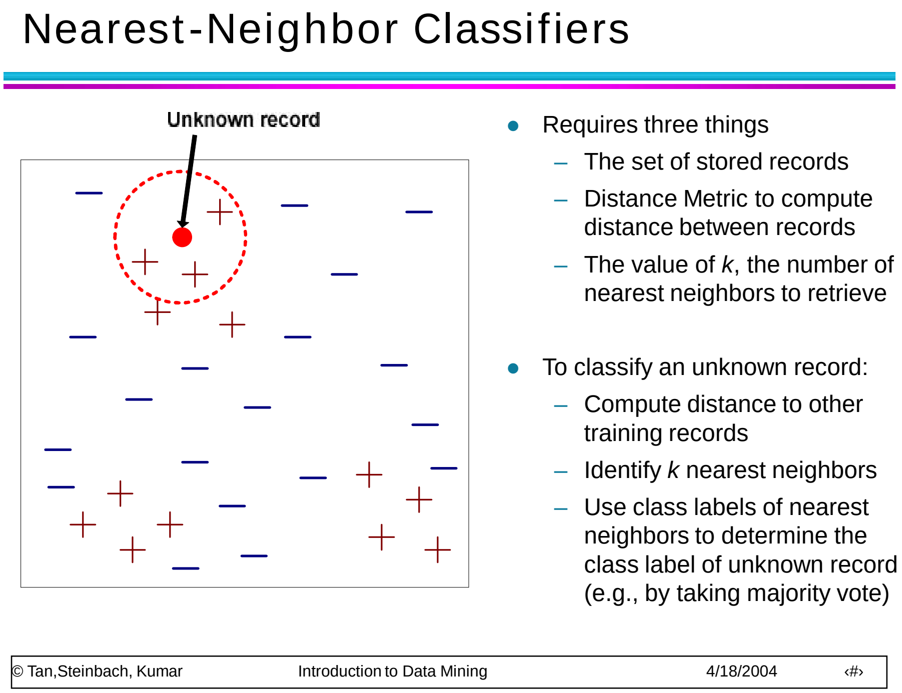
5c.	Discuss K-nearest neighbor classification algorithm with characteristics of Nearest neighbor classifiers.	07 Marks

Ans:



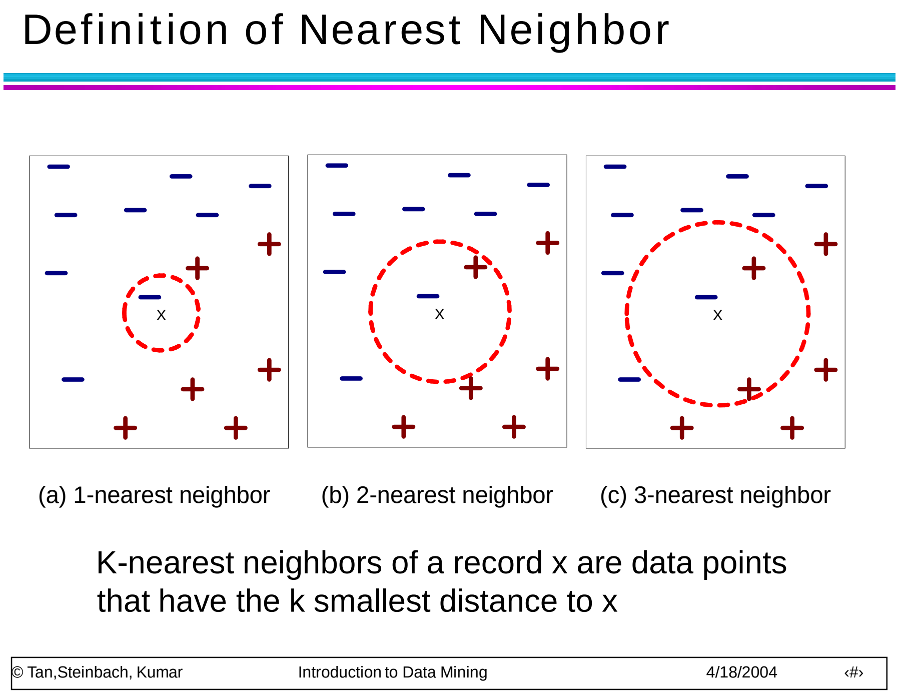
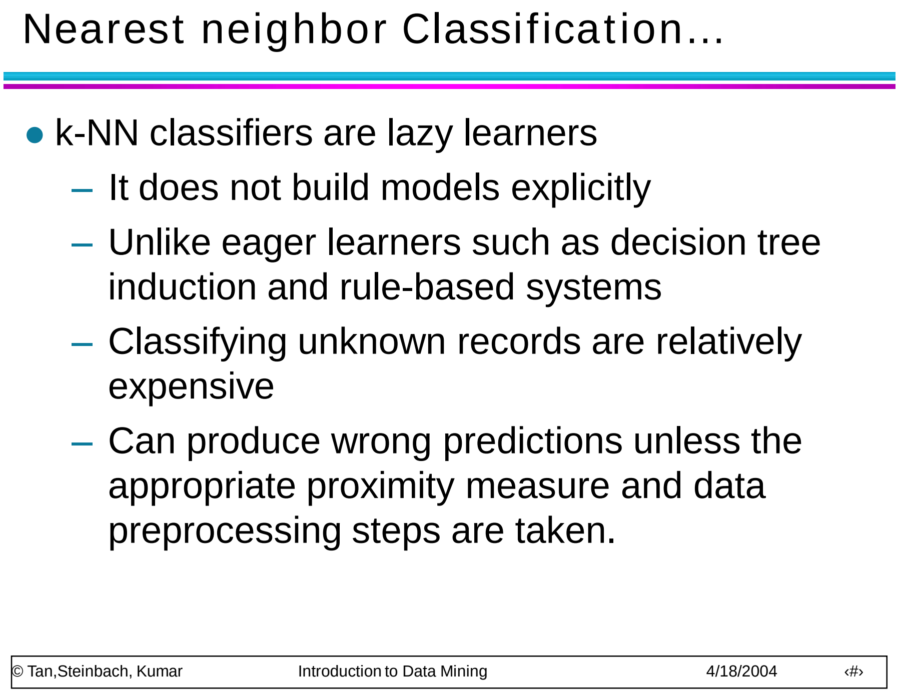
Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it’s probably a duck

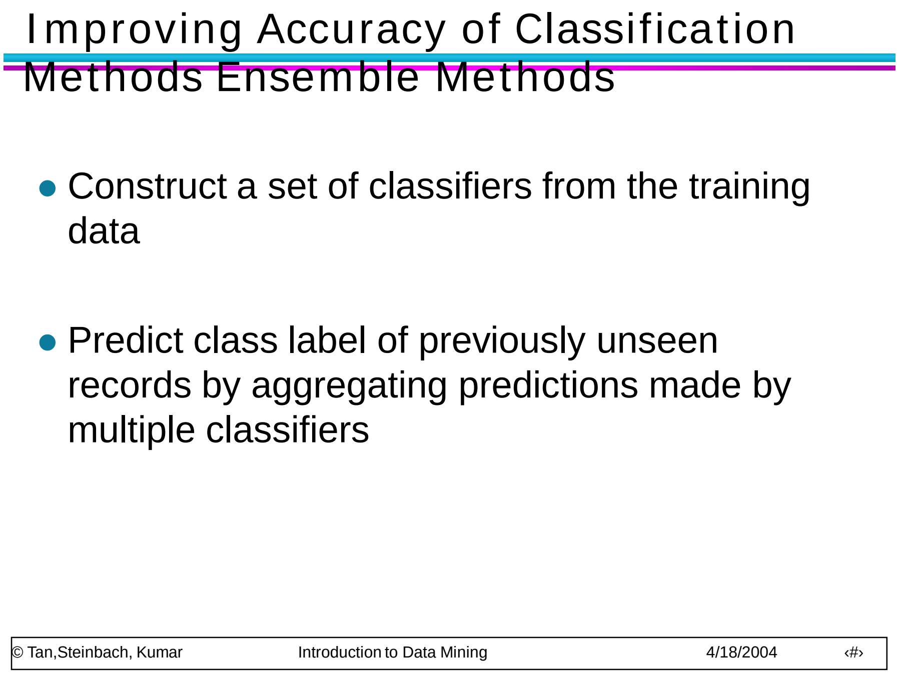


Nearest-Neighbor Classifiers

- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve
- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

	 Definition of Nearest Neighbor (a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor K-nearest neighbors of a record x are data points that have the k smallest distance to x © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›  Nearest neighbor Classification… ● k-NN classifiers are lazy learners – It does not build models explicitly – Unlike eager learners such as decision tree induction and rule-based systems – Classifying unknown records are relatively expensive – Can produce wrong predictions unless the appropriate proximity measure and data preprocessing steps are taken. © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
6a.	Discuss in detail about various techniques for improving the accuracy of classification methods.	07 Marks

Ans:

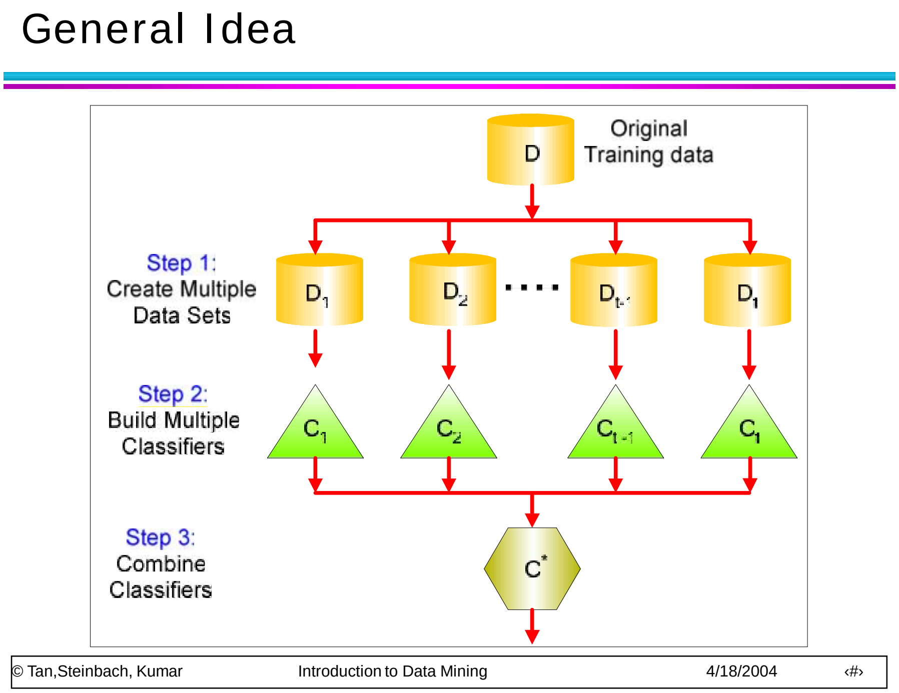


Improving Accuracy of Classification Methods Ensemble Methods

- Construct a set of classifiers from the training data

- Predict class label of previously unseen records by aggregating predictions made by multiple classifiers

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



General Idea

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

Examples of Ensemble Methods

- How to generate an ensemble of classifiers?
  - Bagging
  - Boosting

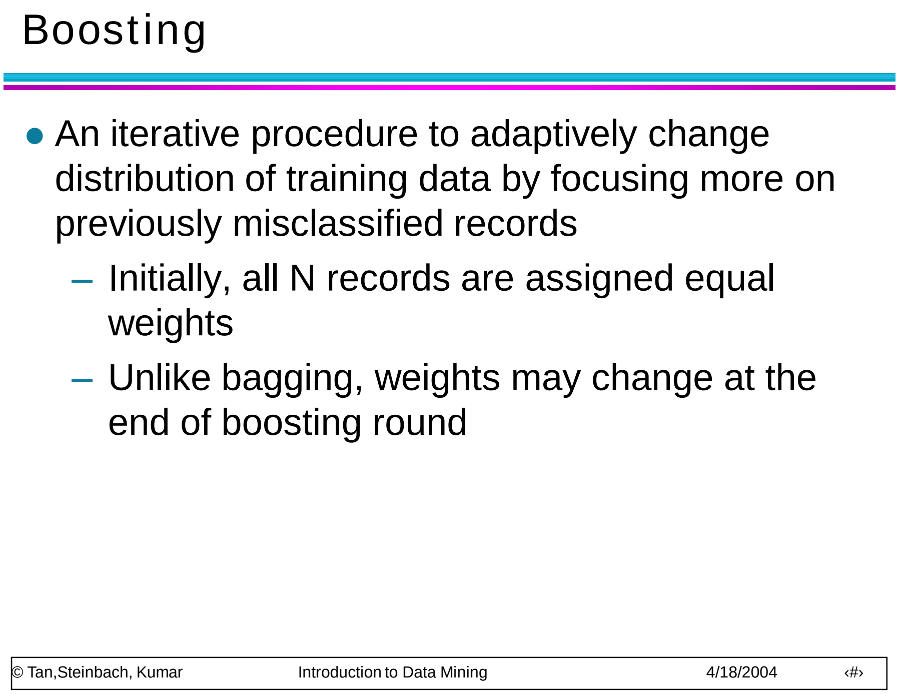
Bagging

- Sampling with replacement
- Each sample has the same size as the original data

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Some instances may appear several times while other may be omitted.
- Build classifier on each bootstrap sample

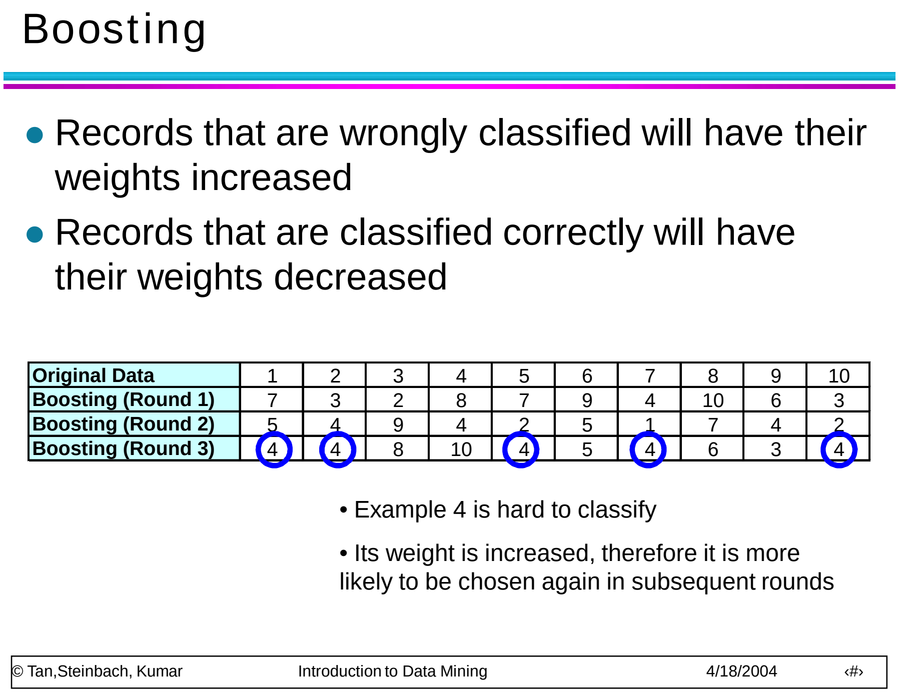
- Each sample has probability $(1 – 1/n)^n$ of being selected



Boosting

- An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records
  - Initially, all N records are assigned equal weights
  - Unlike bagging, weights may change at the end of boosting round

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



Boosting

- Records that are wrongly classified will have their weights increased
- Records that are classified correctly will have their weights decreased

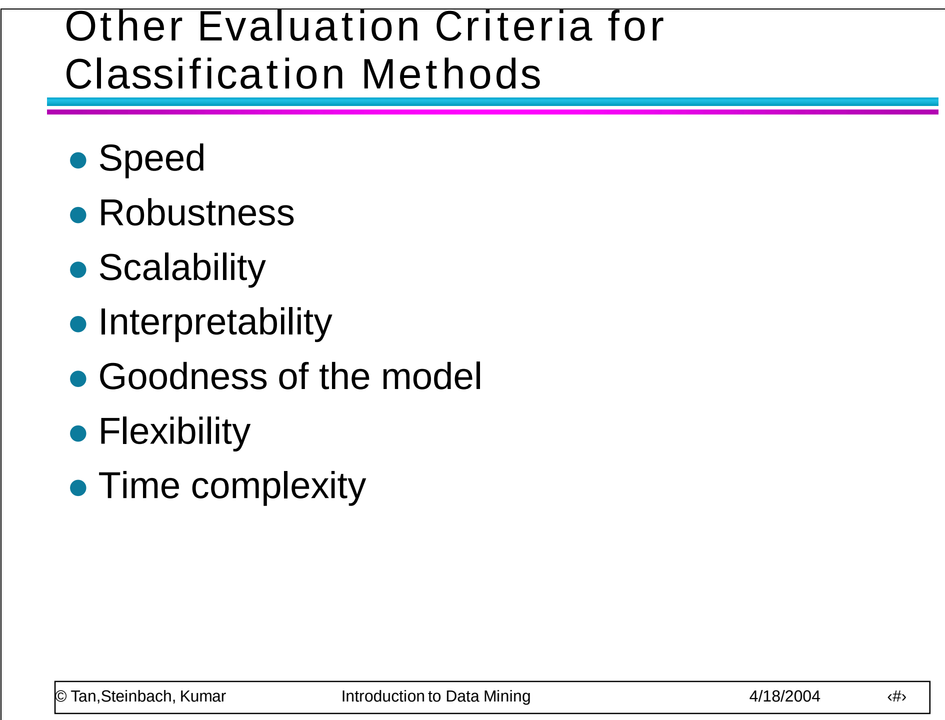
Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

- Example 4 is hard to classify
- Its weight is increased, therefore it is more likely to be chosen again in subsequent rounds

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

6b.	Explain in detail about various evaluation criteria for classification methods.	06 Marks

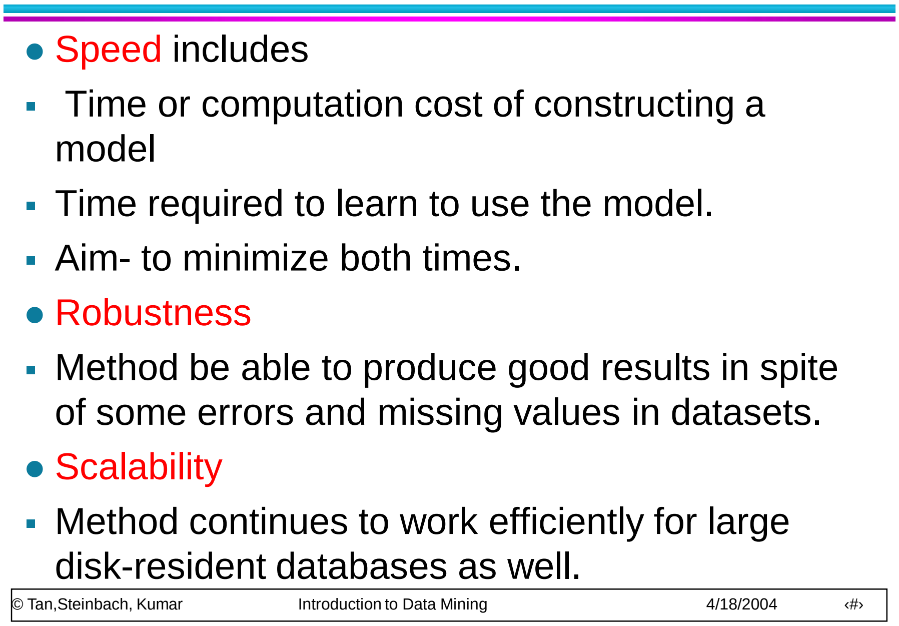
Ans:



Other Evaluation Criteria for Classification Methods

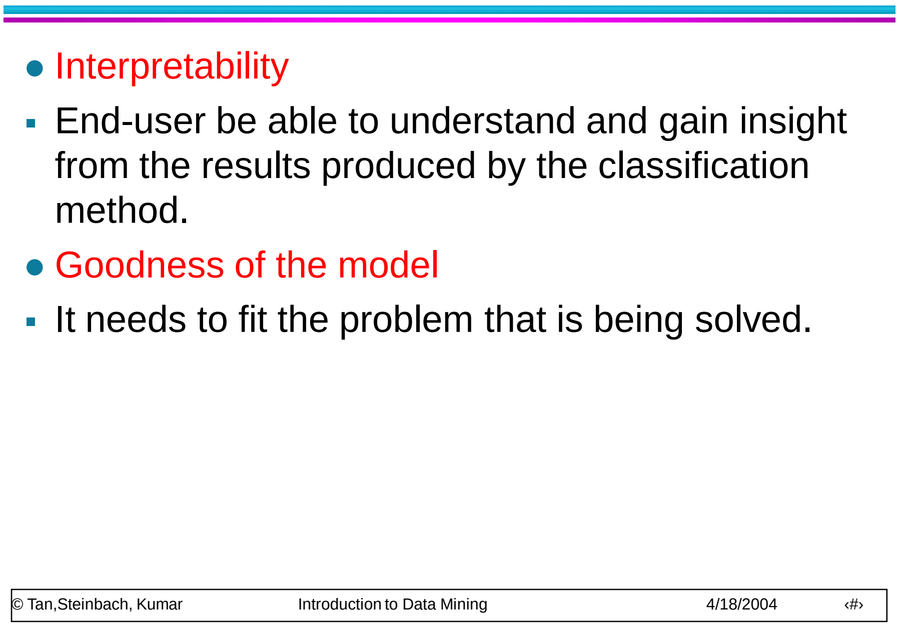
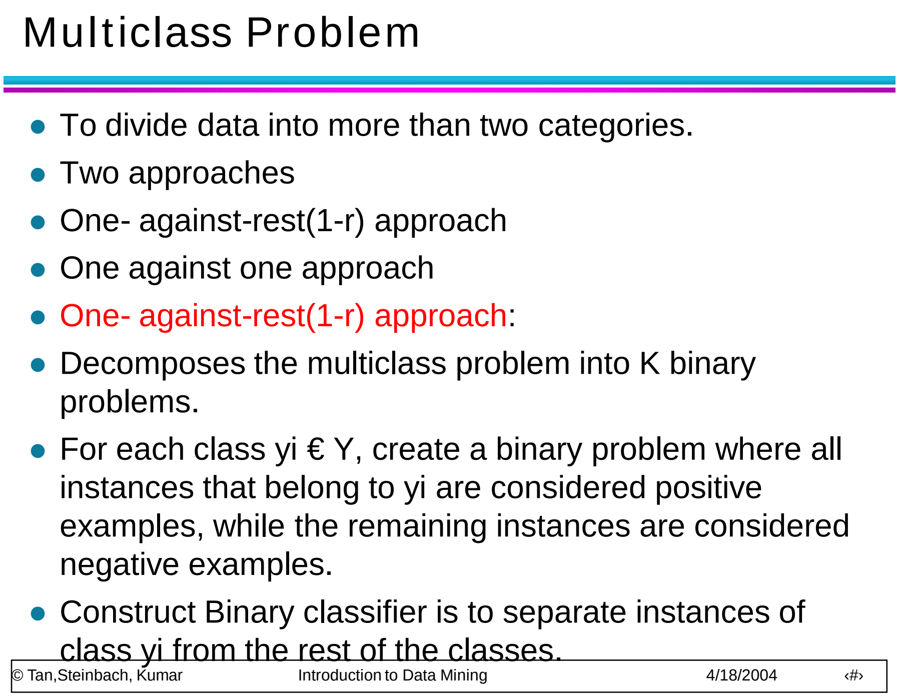
- Speed
- Robustness
- Scalability
- Interpretability
- Goodness of the model
- Flexibility
- Time complexity

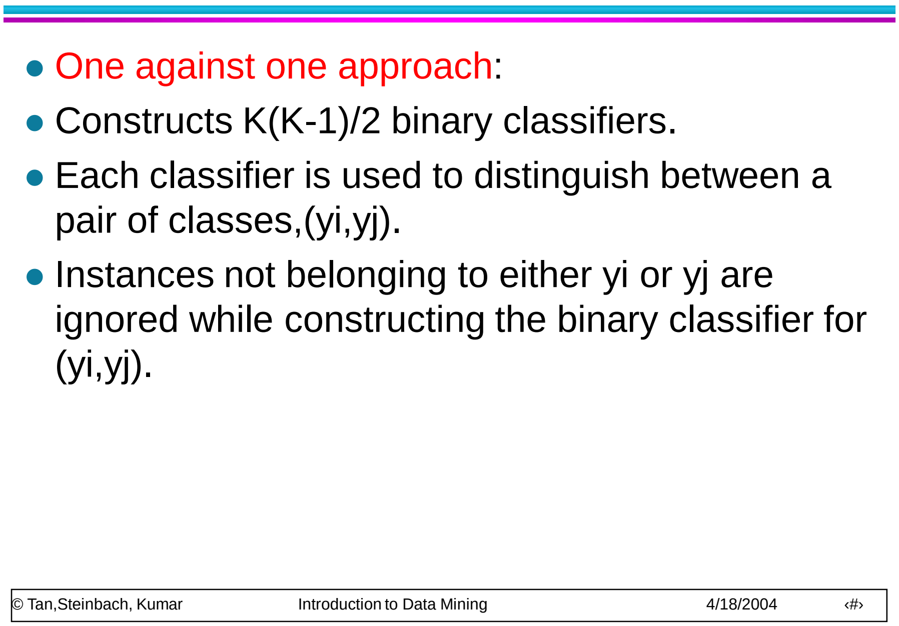
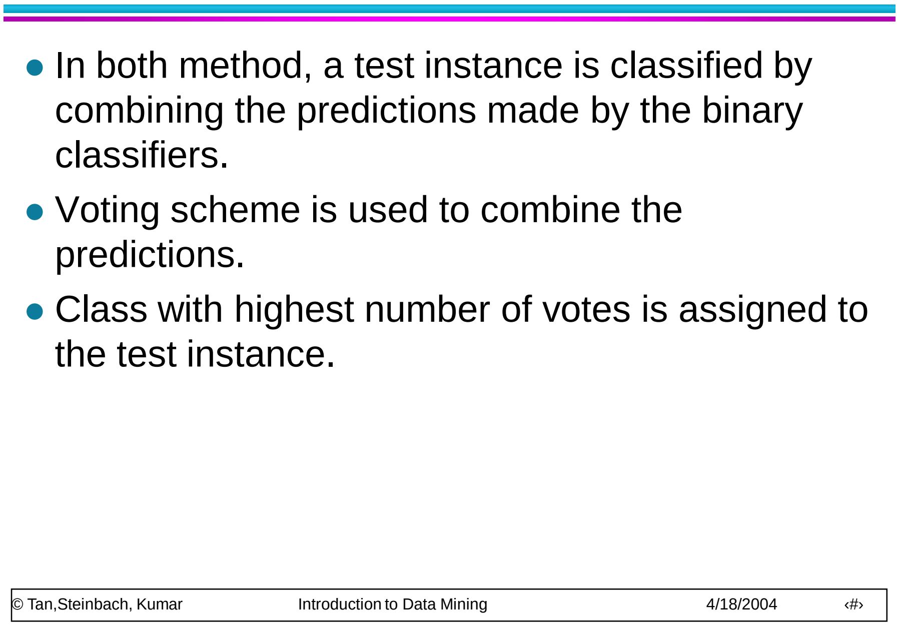
© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



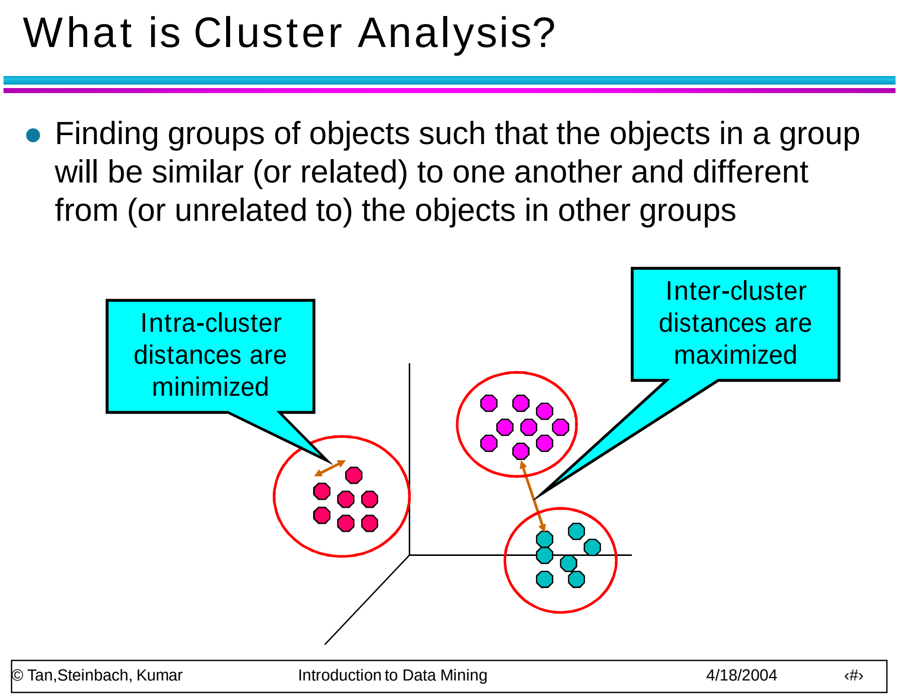
- Speed includes
  - Time or computation cost of constructing a model
  - Time required to learn to use the model.
  - Aim- to minimize both times.
- Robustness
  - Method be able to produce good results in spite of some errors and missing values in datasets.
- Scalability
  - Method continues to work efficiently for large disk-resident databases as well.

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

	 ● Interpretability ▪ End-user be able to understand and gain insight from the results produced by the classification method. ● Goodness of the model ▪ It needs to fit the problem that is being solved. © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
6c.	Describe Multiclass Problem indetail.	07 Marks
Ans:	 Multiclass Problem ● To divide data into more than two categories. ● Two approaches ● One- against-rest(1-r) approach ● One against one approach ● One- against-rest(1-r) approach: ● Decomposes the multiclass problem into K binary problems. ● For each class yi € Y, create a binary problem where all instances that belong to yi are considered positive examples, while the remaining instances are considered negative examples. ● Construct Binary classifier is to separate instances of class yi from the rest of the classes. © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	

	 - One against one approach: - Constructs K(K-1)/2 binary classifiers. - Each classifier is used to distinguish between a pair of classes,(yi,yj). - Instances not belonging to either yi or yj are ignored while constructing the binary classifier for (yi,yj). © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›  - In both method, a test instance is classified by combining the predictions made by the binary classifiers. - Voting scheme is used to combine the predictions. - Class with highest number of votes is assigned to the test instance. © Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›	
7a.	What is Cluster Analysis? Explain Agglomerative clustering method in detail.	06 Marks

Ans:

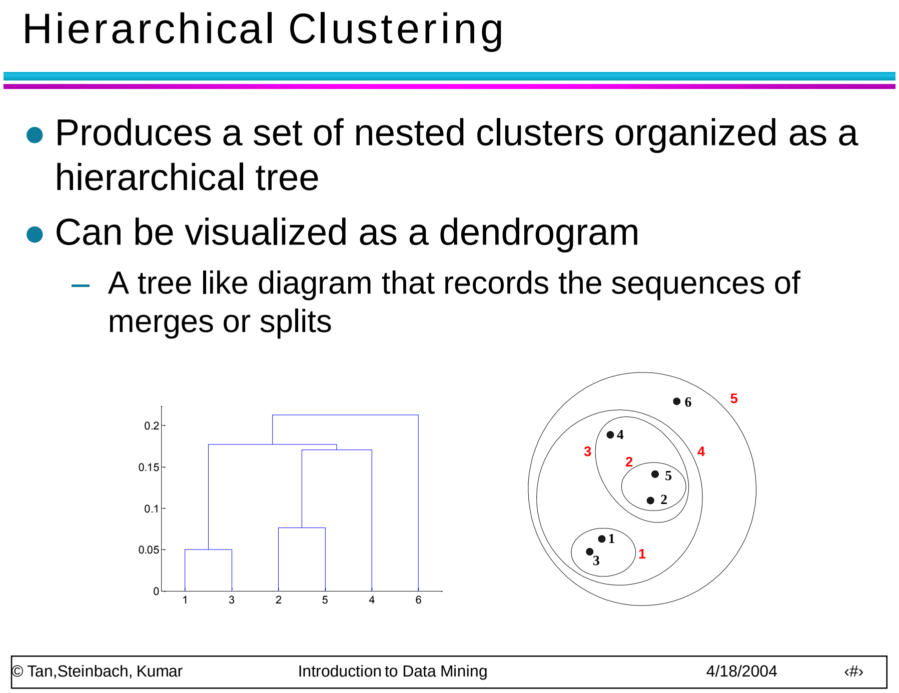


What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Intra-cluster distances are minimized
Inter-cluster distances are maximized

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits

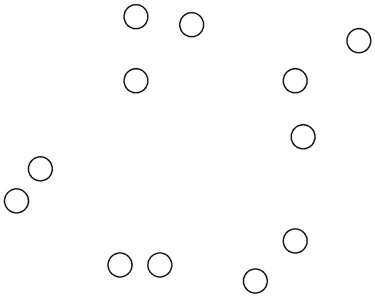
© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the proximity of two clusters
  – Different approaches to defining the distance between clusters distinguish the different algorithms

Starting Situation

- Start with clusters of individual points and a proximity matrix



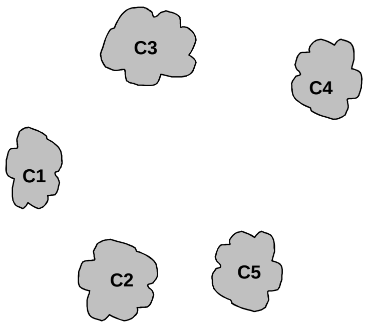
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

p1
p2
p3
p4
...
p9
p10
p11
p12

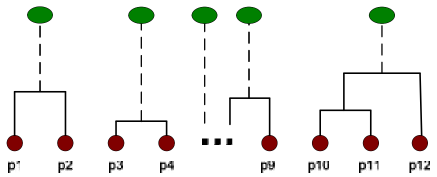
Intermediate Situation

- After some merging steps, we have some clusters



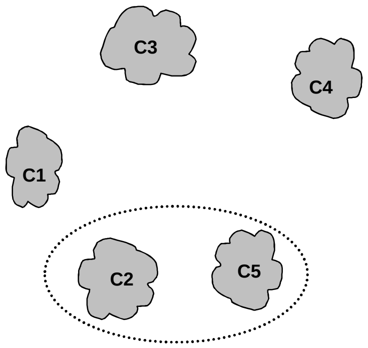
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



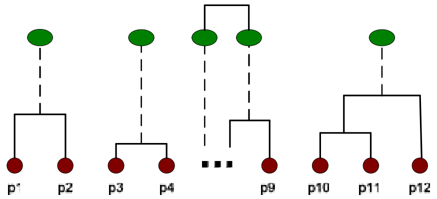
Intermediate Situation

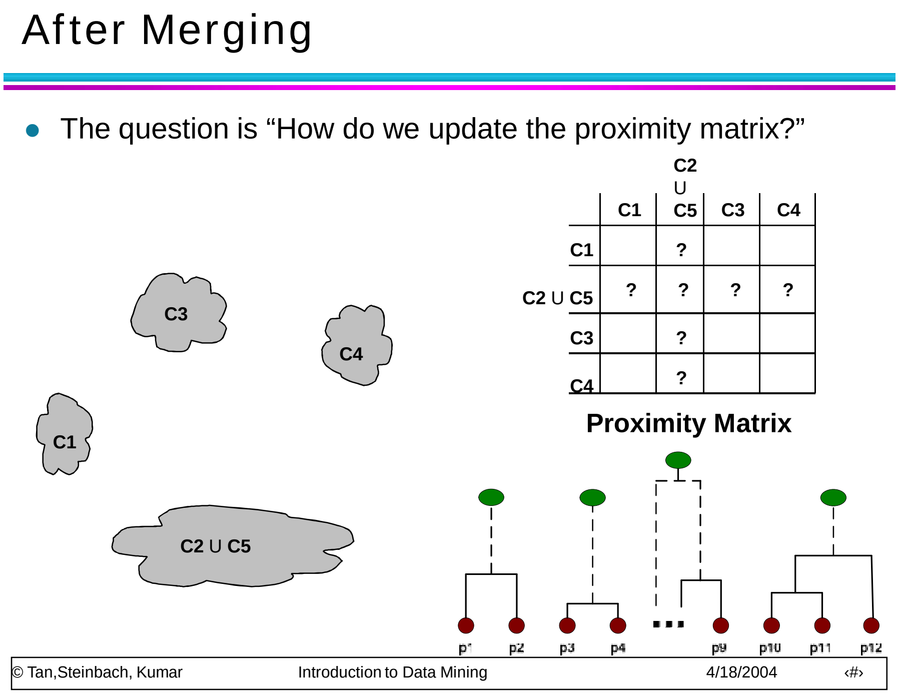
- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



	C1	C2	C3	C4	C5
C1		▨			▨
C2	▨	▨	▨	▨	▨
C3		▨			▨
C4		▨			▨
C5	▨	▨	▨	▨	▨

Proximity Matrix



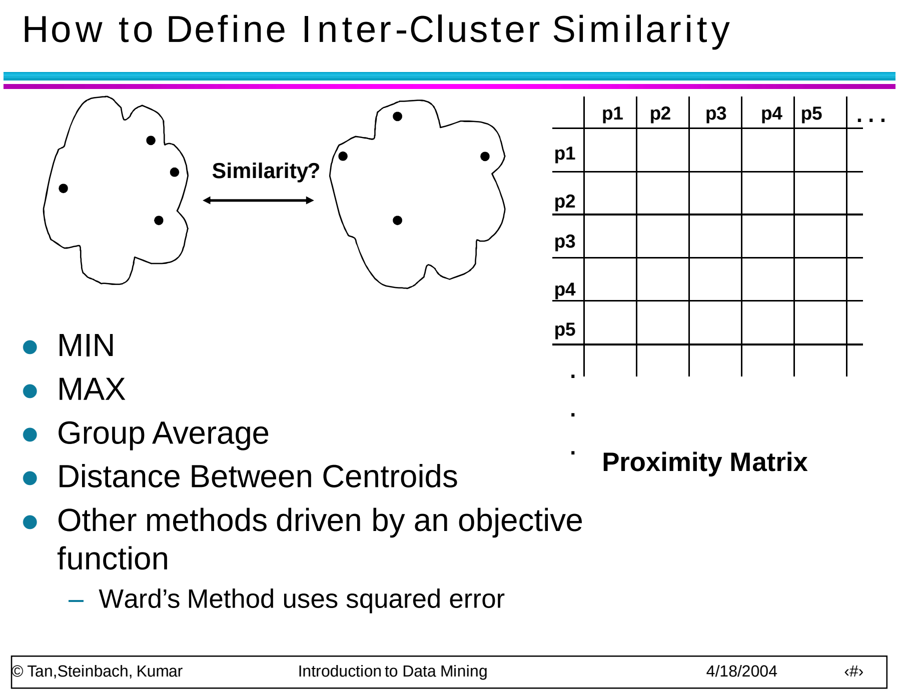


After Merging

- The question is “How do we update the proximity matrix?”

	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Similarity

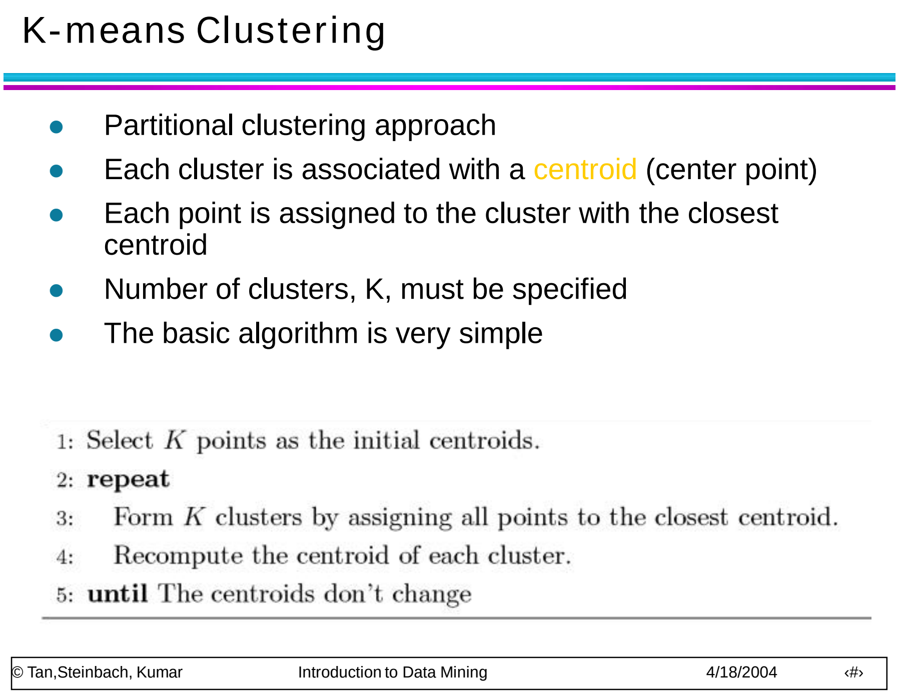
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward’s Method uses squared error

7b.	Discuss K- means method in detail, with an example.	08 Marks

Ans:

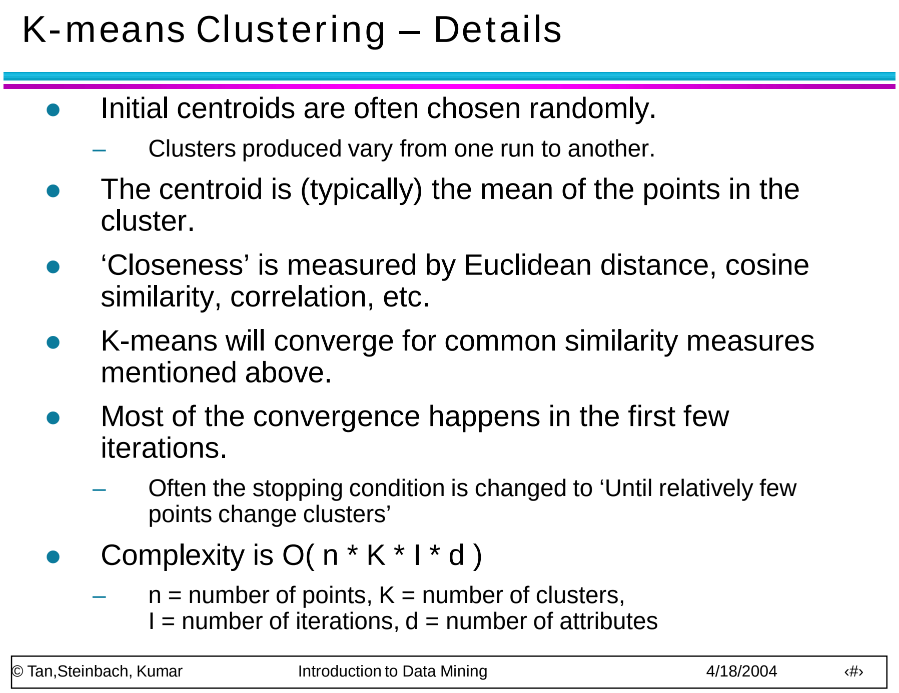


K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

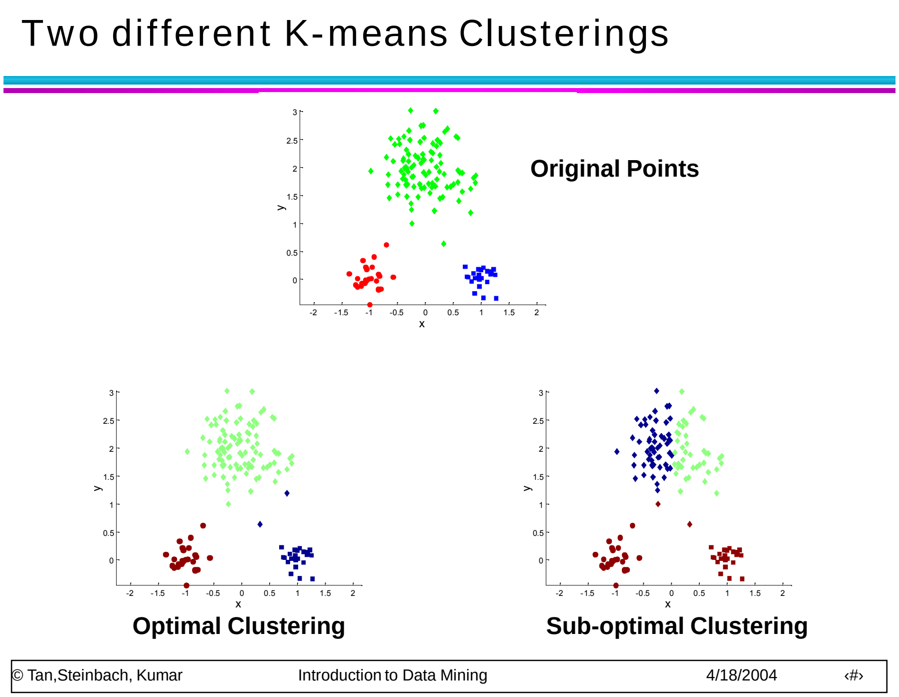
© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



K-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- 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

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



Two different K-means Clusterings

Original Points

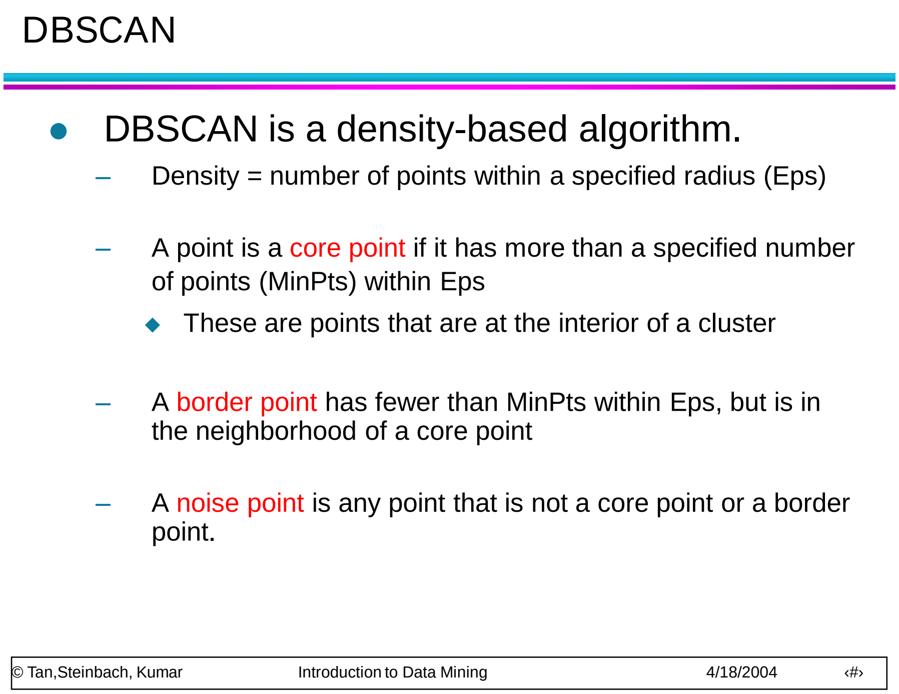
Optimal Clustering

Sub-optimal Clustering

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

7c. Describe DBSCAN method in detail. 06 Marks

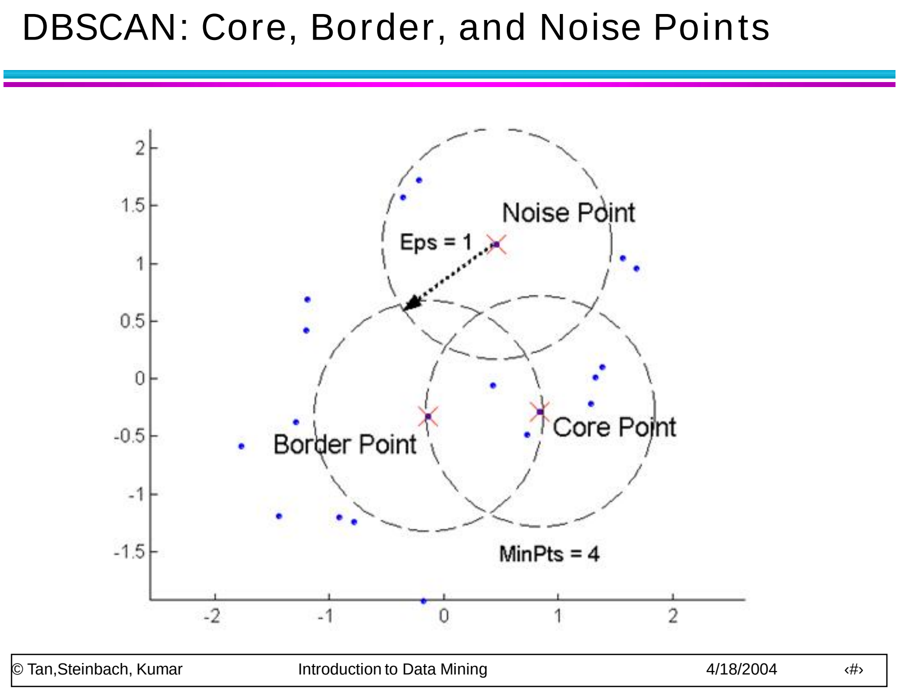
Ans:



DBSCAN

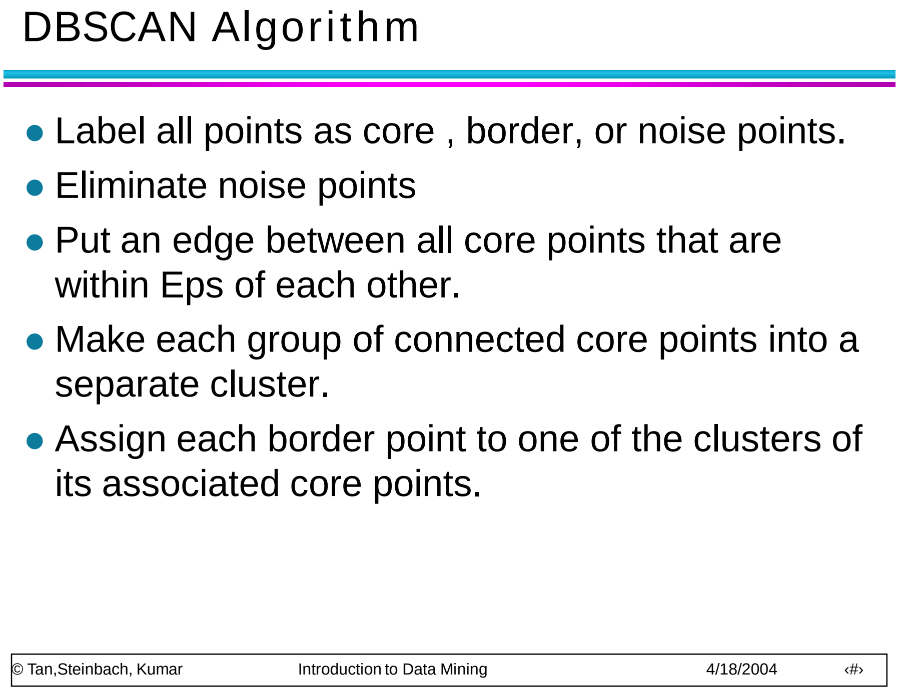
- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



DBSCAN: Core, Border, and Noise Points

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›



DBSCAN Algorithm

- Label all points as core , border, or noise points.
- Eliminate noise points
- Put an edge between all core points that are within Eps of each other.
- Make each group of connected core points into a separate cluster.
- Assign each border point to one of the clusters of its associated core points.

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

8a.	What are Outliers? Explain statistical approaches in detail.	10 Marks
Ans:	“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”	

Statistical-based – Likelihood Approach

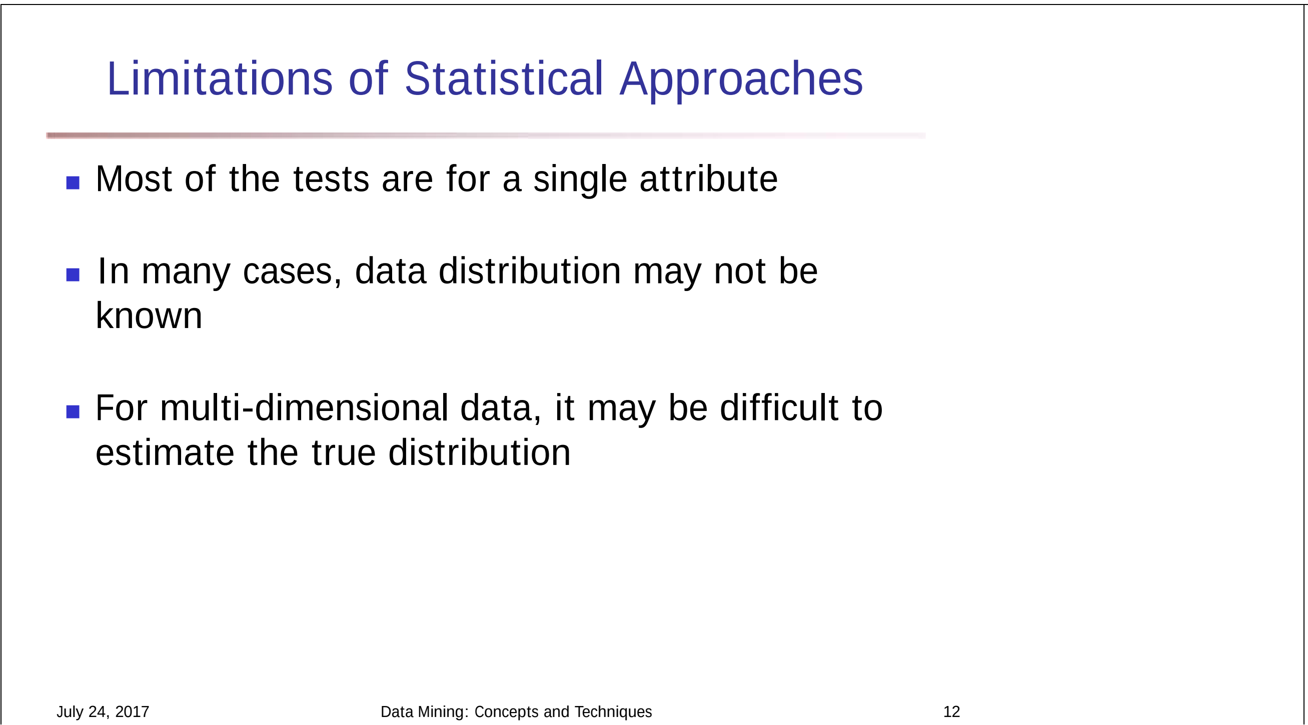
- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let L_{t+1} (D) be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, D = $(1 - \lambda)$ M + λ A
- M is a probability distribution estimated from data
  - Can be based on any modeling method
  - A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-a)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(a^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-a) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log a + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

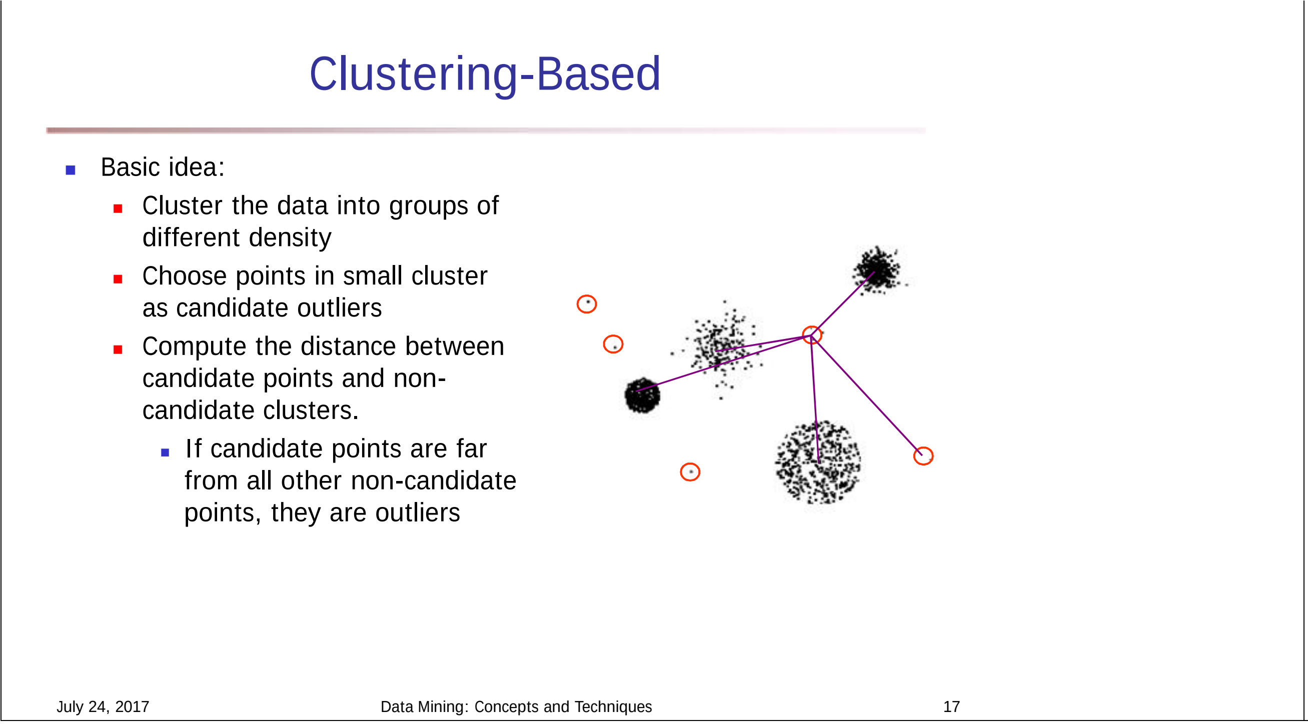


Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For multi-dimensional data, it may be difficult to estimate the true distribution

8b.	Discuss Clustering- based approaches in detail.	10 Marks

Ans:



Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, they are outliers

Outliers in Lower Dimensional Projections

- In high-dimensional space, data is sparse and notion of proximity becomes meaningless
  - Every point is an almost equally good outlier from the perspective of proximity-based definitions

- Lower-dimensional projection methods
  - A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density

Outliers in Lower Dimensional Projection

- Divide each attribute into ϕ equal-depth intervals
  - Each interval contains a fraction $f = 1/\phi$ of the records
- Consider a d-dimensional cube created by picking grid ranges from d different dimensions
  - If attributes are independent, we expect region to contain a fraction f^k of the records
  - If there are N points, we can measure sparsity of a cube D as:

$$S(\mathcal{D}) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

  - Negative sparsity indicates cube contains smaller number of points than expected
  - To detect the sparse cells, you have to consider all cells…. exponential to d. Heuristics can be used to find them…