Data Mining in the Data Warehouse Context

COMP323 Chapter 5

Outline

- Motivation
  - What is OLAP?
  - What is data mining?
  - OLAP vs. Data mining
  - Data warehouse as a data source
- Basic data mining techniques
  - Clustering
  - Decision tree
  - Association analysis

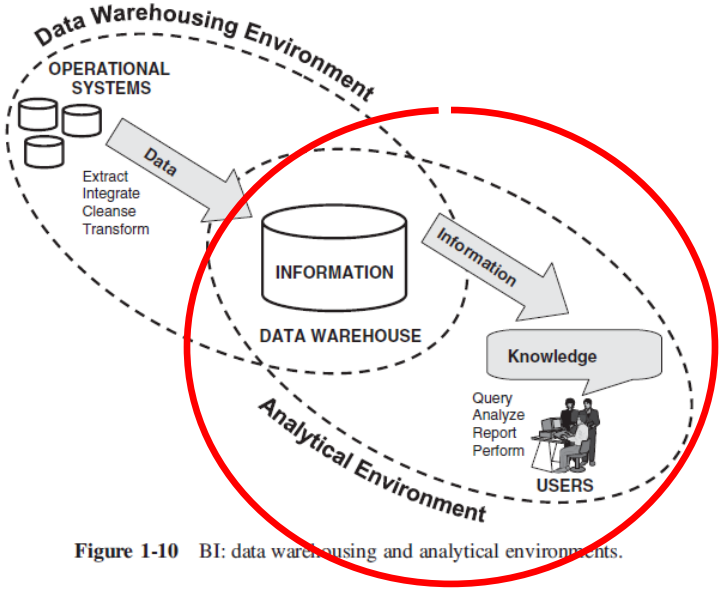


Figure 1-10 BI: data warehousing and analytical environments.

Online Analytical Processing

- Online Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user. (from OLAP council)
- Major features:
  - Multidimensional data analysis
  - Complex calculation
    - E.g. margin (sales – cost), percentage of parts to the whole, moving average, growth percentage, trend analysis using statistical methods
  - Speed-of-thought response
    - Critical to maintain the train of thought in interactive analysis session

A typical analysis session

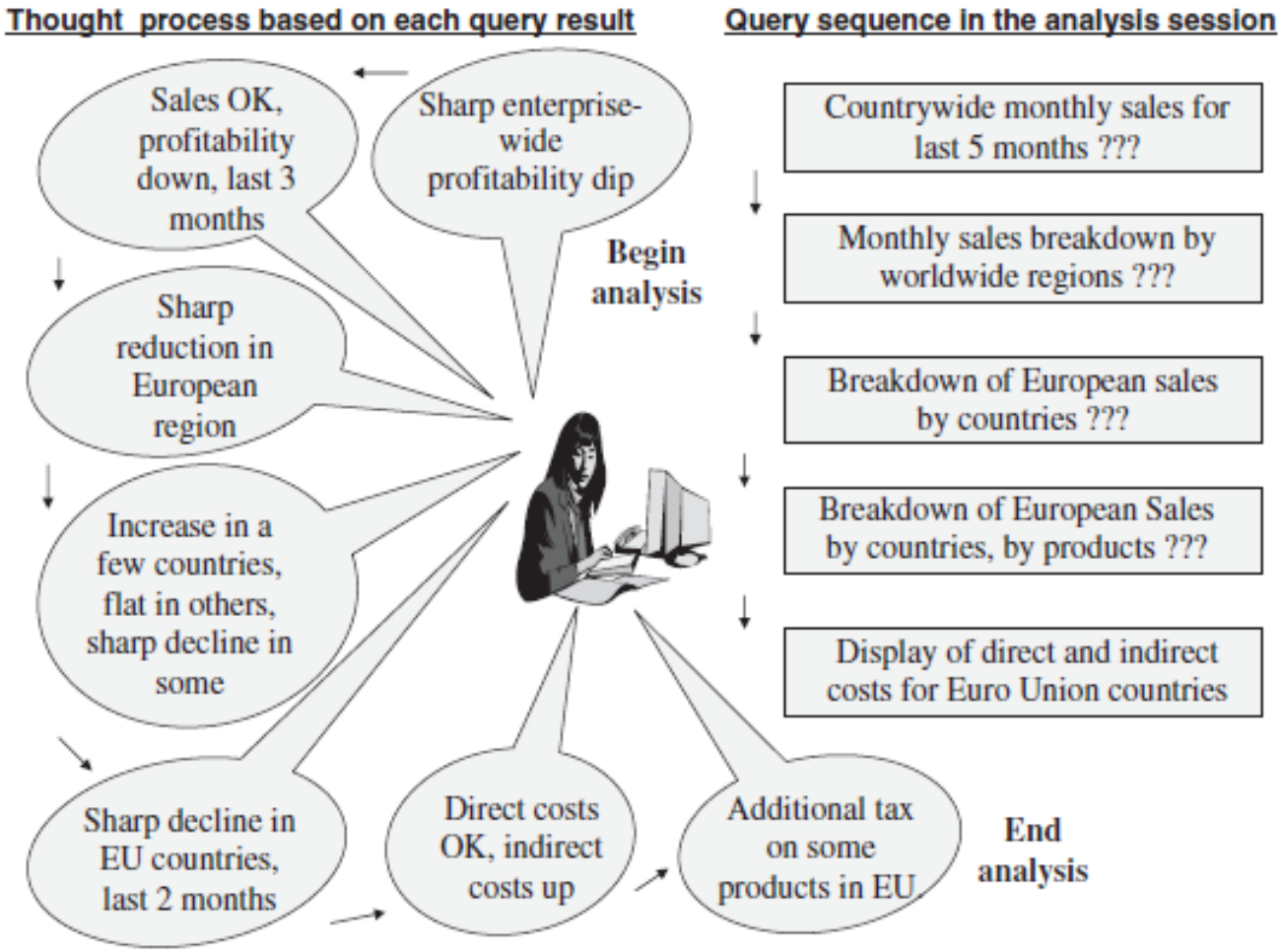


Figure 15-1 Query steps in an analysis session.

Queries in a typical analysis session

Why profitability dipped sharply in recent months in the entire enterprise?

- *What are the overall sales for the last five months for the entire company, broken down by individual months?*
- The sales do not show a drop, but there is a sharp reduction in profitability for the last three months.

- *Give me a breakdown of monthly sales by major worldwide regions.*
- The European region is responsible for the reduction in profitability.

- *Give me a breakdown of European sales by individual countries.*
- Profitability has increased for a few countries, decreased sharply for some other countries, and been stable for the rest.

- *Give me a breakdown of profitability for the European countries by country, month, and product.*
- Very sharp decline in profitability for the last two months for some products in the countries.

- *Display the direct and indirect costs for European countries of those products.*
- Direct costs (e.g. manufacturing) remain at the usual levels, but the indirect costs have shot up.

The decline is due to additional tax levies on some products in the EU.

Data Mining

- Data mining is the efficient discovery of valuable, non-obvious information from a large collection of data
- Usually, hidden information revealed by data mining is relationship / pattern, e.g.
  - A bank decides the possibility of bad debt given some customers information (e.g. income level, credit history, owned properties)
  - Products that are likely bought together (e.g. bread and butter) are placed in close proximity in supermarkets
  - A credit card company uses data mining to detect fraud in credit card usage.

Decision Support and Business Intelligence

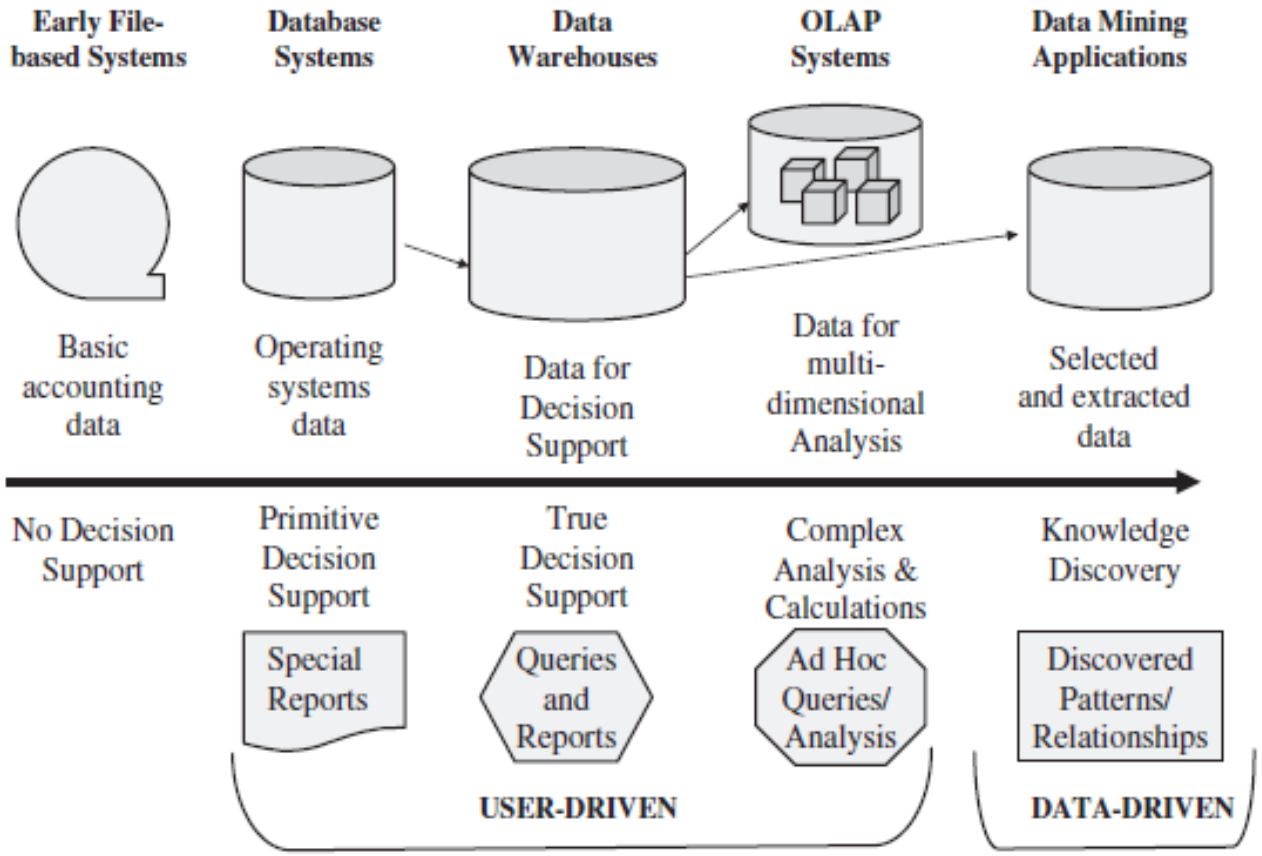


Figure 17-1 Decision support progresses to data mining.

Steps in Data Mining

- Define business objectives
- Prepare data:
  - Data selection, preprocessing and transformation
- Select data mining algorithms and perform data mining
- Evaluate results

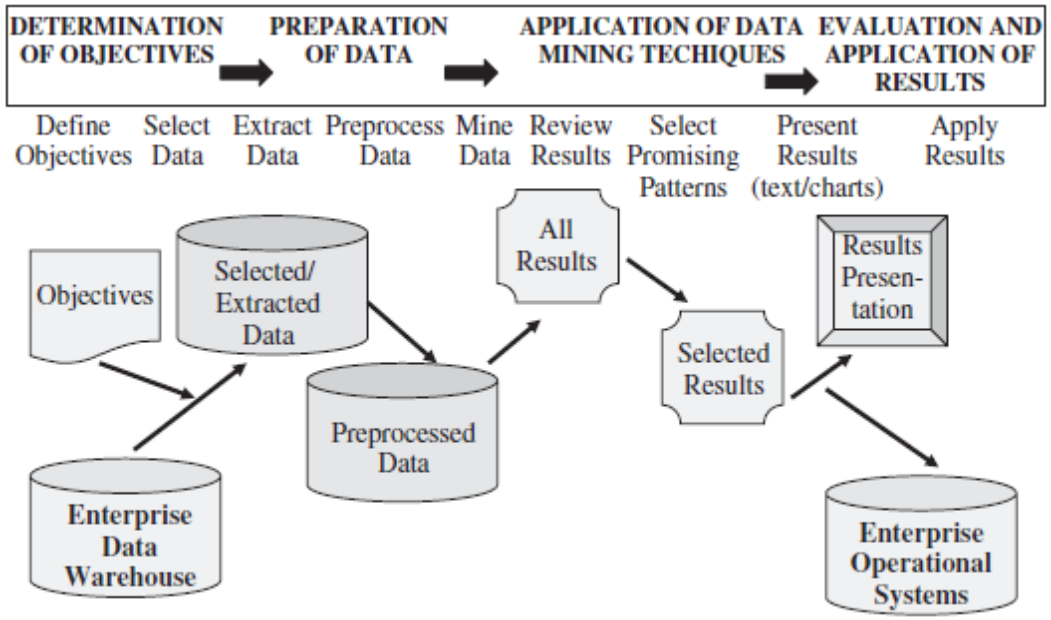


Figure 17-4 Knowledge discovery process.

OLAP vs. Data Mining

- OLAP users need to have some prior knowledge of what they are looking for. OLAP is user-driven and interactive.
- Data mining users has no prior knowledge of what the results are likely to be. Data mining is data-driven automatic knowledge discovery.

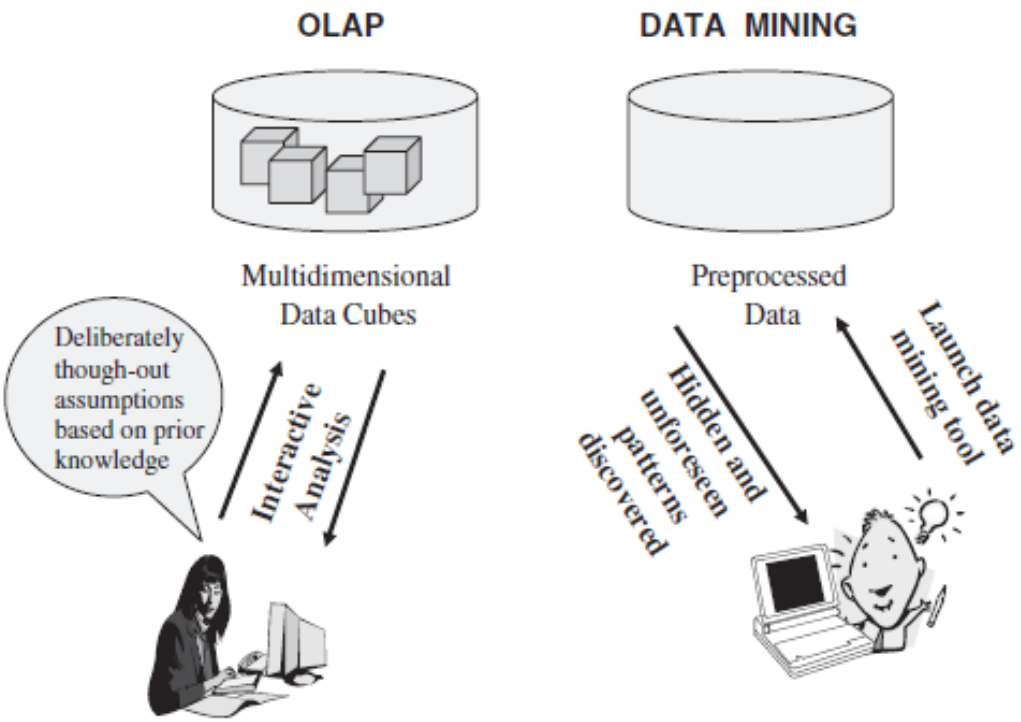


Figure 17-5 OLAP and data mining.

Differences between OLAP and data mining

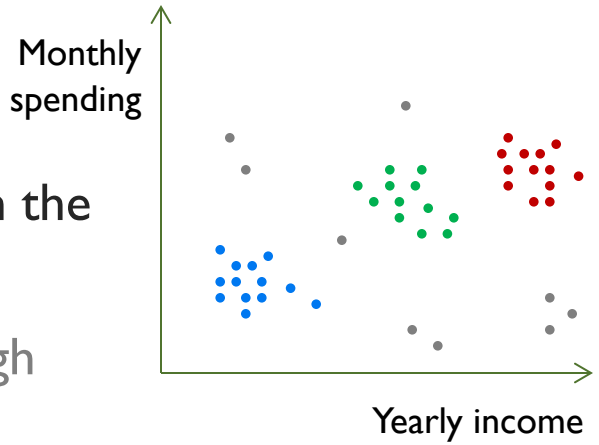
Features	OLAP	Data Mining
Motivation for information request	What is happening in the enterprise?	Predict the future based on why this is happening.
Data granularity	Summary data	Detailed transaction-level data
Number of business dimensions	Limited number of dimensions	Large number of dimensions
Number of dimension attributes	Small number of attributes	Many dimension attributes
Sizes of datasets for the dimensions	Not large for each dimension	Usually very large for each dimension
Analysis approach	User-driven, interactive analysis	Data-driven automatic knowledge discovery
Analysis techniques	Multidimensional, drill-down, and slice-and-dice	Prepare data, launch mining tool and sit back

Data warehouse and data mining

- Data warehouse provides an essential data source for data mining
  - Data mining algorithms need large amounts of data at the detailed level. Most data warehouses contain data at the lowest level of granularity.
  - Integrated and cleansed data are critical for the quality of data mining result. The ETL process of data warehouses provide such data.

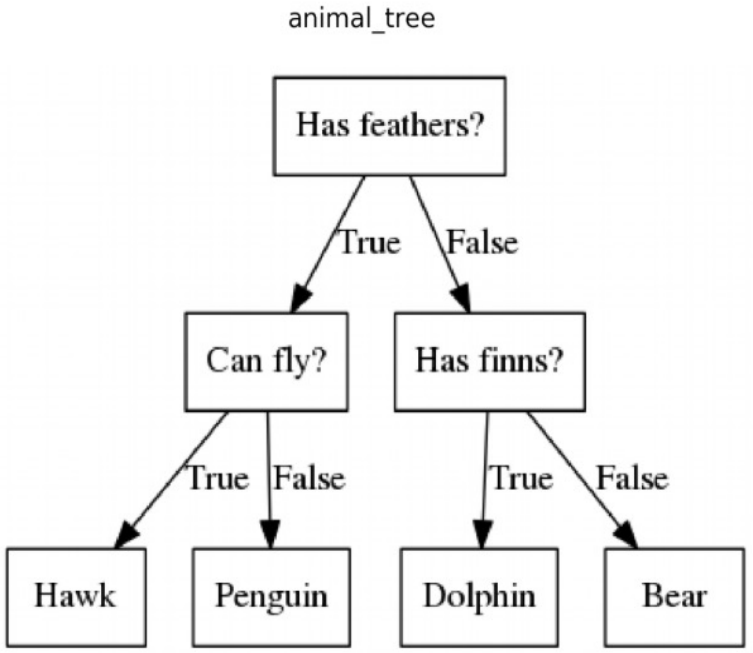
Clustering

- Clustering means identifying and forming groups. The objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.
- Clustering becomes more difficult when more variables are considered
  - E.g. age group, education, gender
- Application:
  - Market researchers use clustering to partition the general population of consumers into market segments.
    E.g. clustering may reveal customers with a high yearly income who live in certain cities prefer high-quality wine, or the sales of food by customers of some occupation becomes higher in weekends, whereas some other customers prefer to visit the supermarket during office hour.



Decision tree

- A decision tree is a model to classify objects based on observation
  - An internal node is a test on some attribute
  - A branch represents an outcome of the test
  - A leaf node represents a class label
- Application:
  - classify objects based on some training data. (e.g. given some attributes of past customers and their ability to repay loans, we can build a decision tree to predict bad debt)



Association analysis

- Association analysis finds combinations where the presence of one item suggests the presence of another.
- Application:
  - Which products are often bought together by customers? (E.g. bread and butter. A supermarket can then place these related products in near places.)
  - Association of past and future events.
    E.g. Purchase of window curtains
    is followed by purchase of living room furniture 50% of the time.

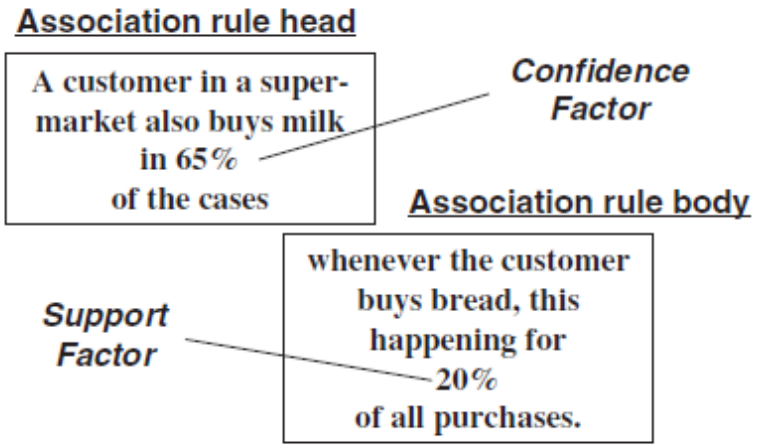


Figure 17-14 An association rule.

Decision Trees

Based on the slides of

- *Introduction to Data Mining*
  by Tan, Steinbach, and Kumar
- *Classification*
  by Lei

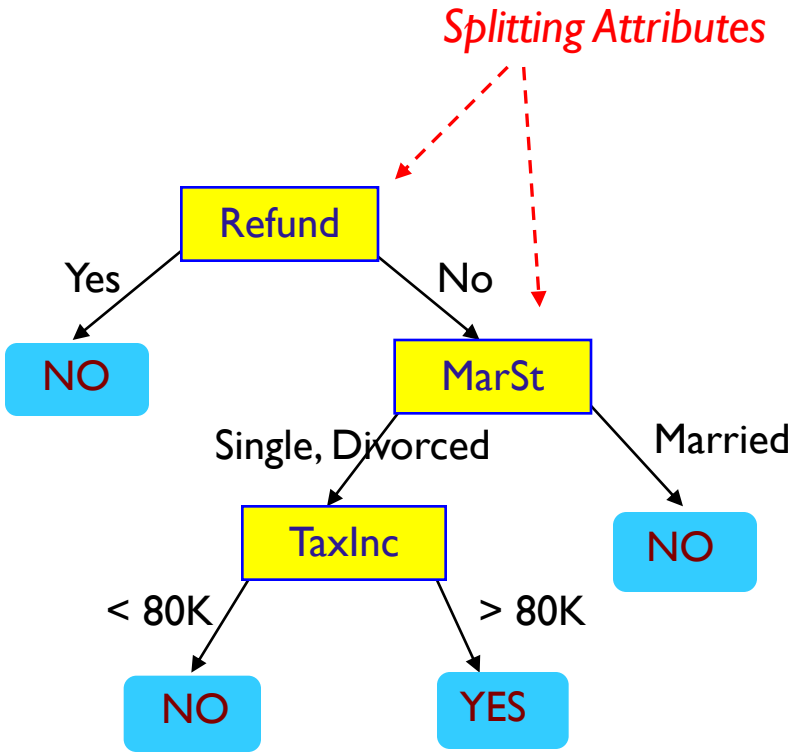
Definitions

- 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.

Example of a Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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

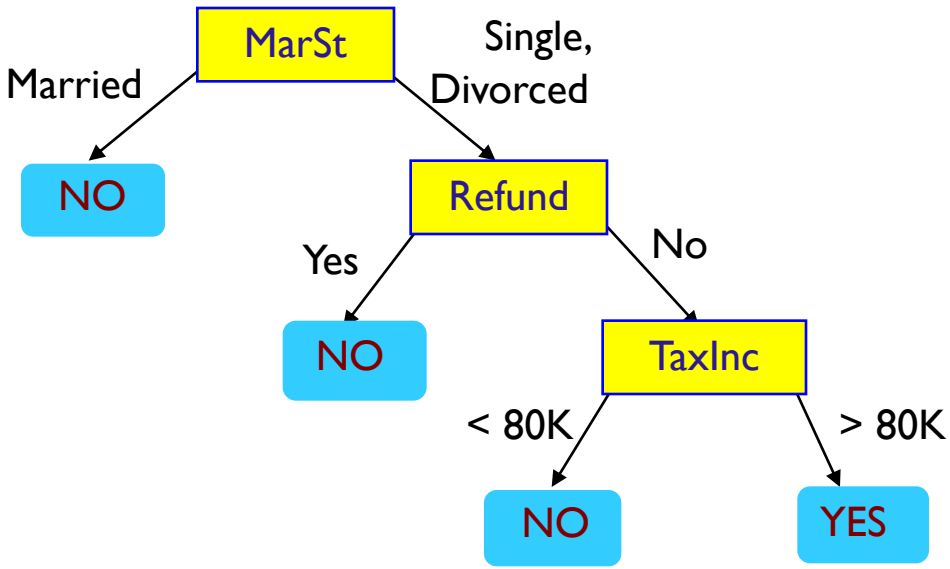
Training Data



Model: Decision Tree

Another Example of Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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



There could be more than one tree that fits the same data!

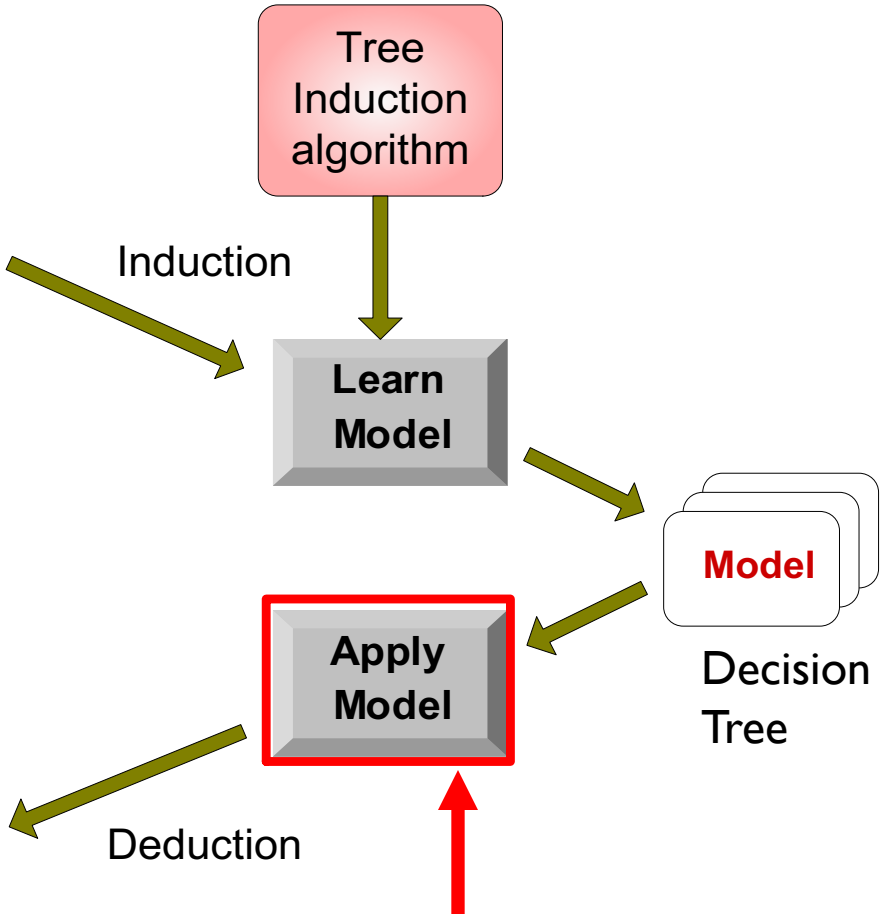
Decision Tree Classification Task

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

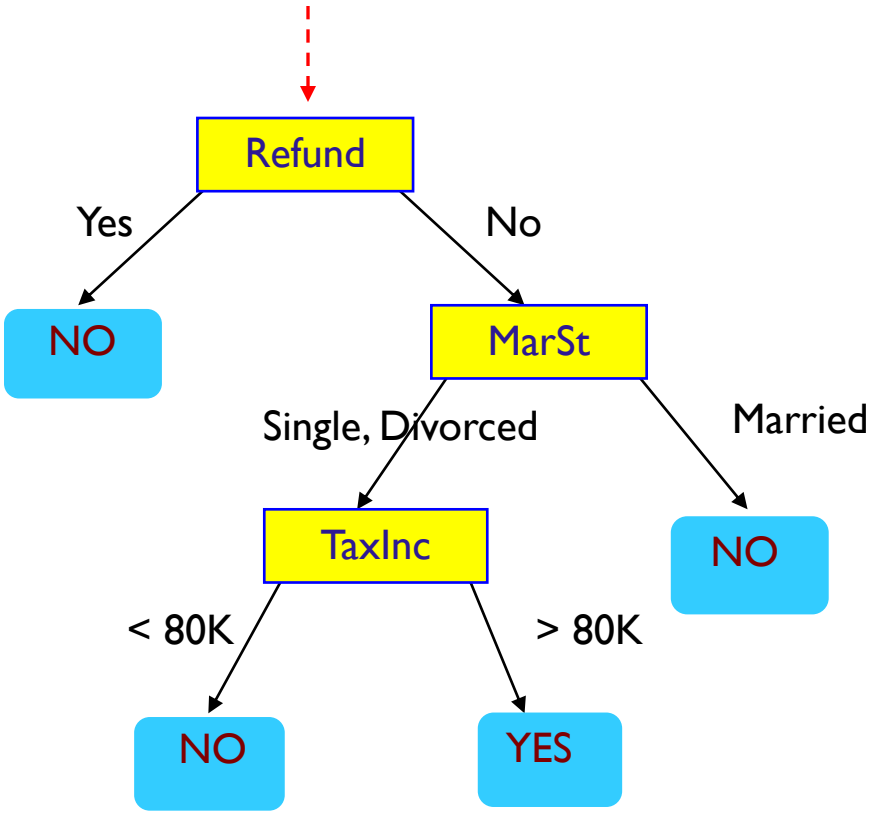
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Apply Model to Test Data

Start from the root of tree.



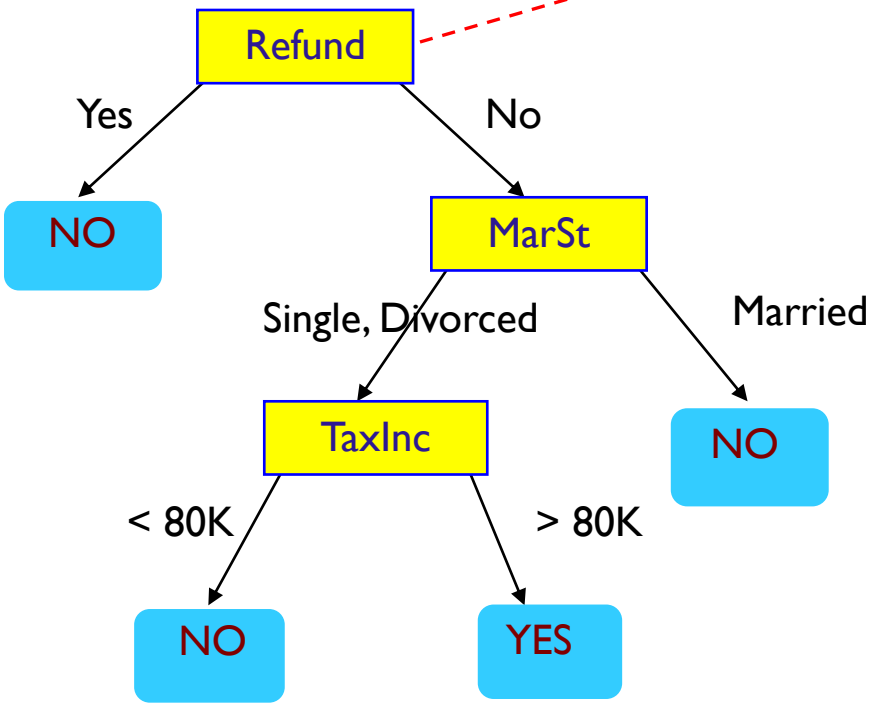
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

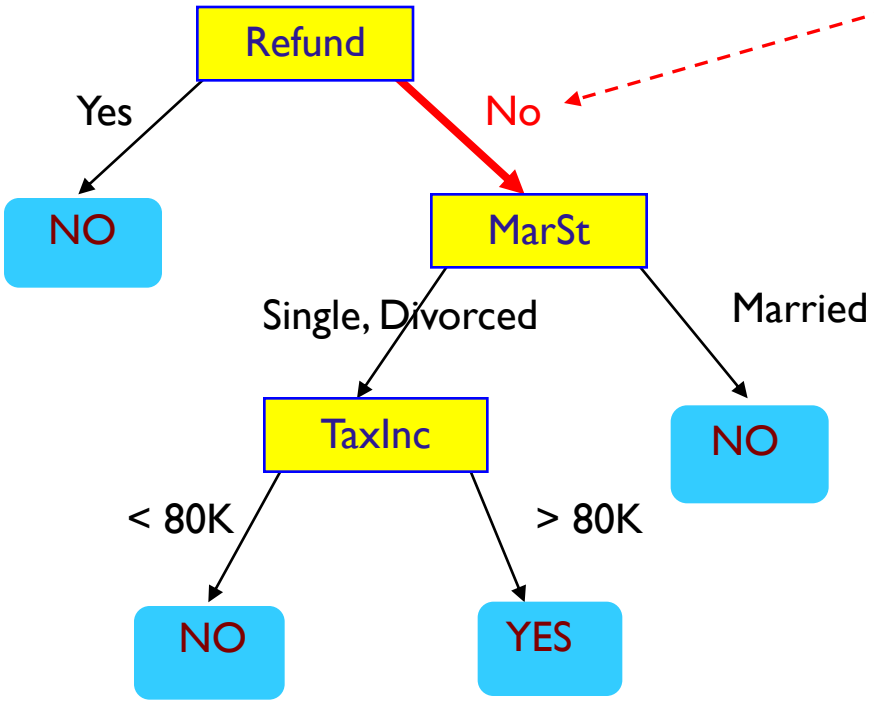
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

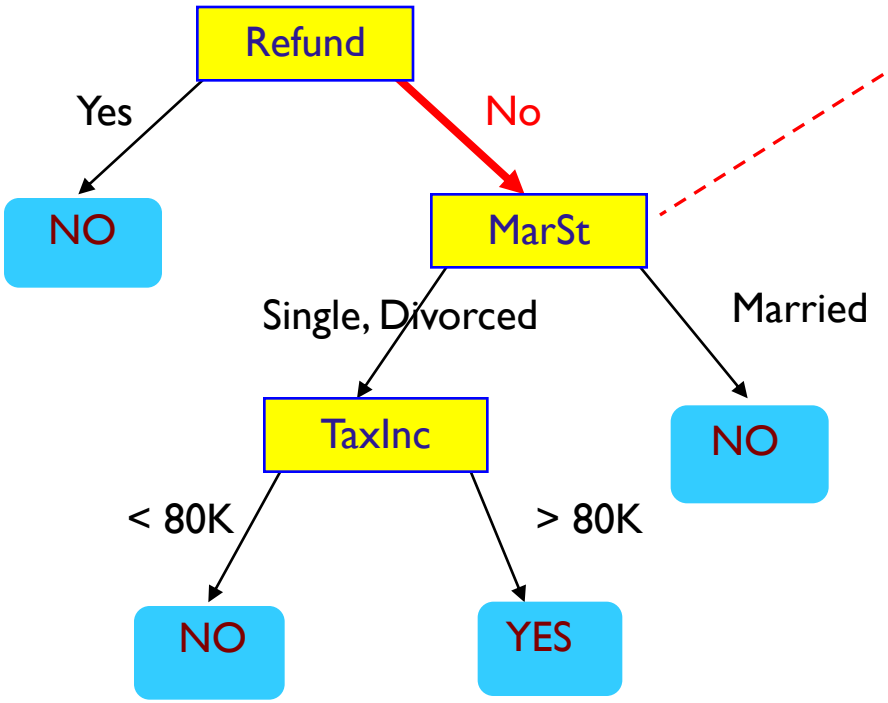
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

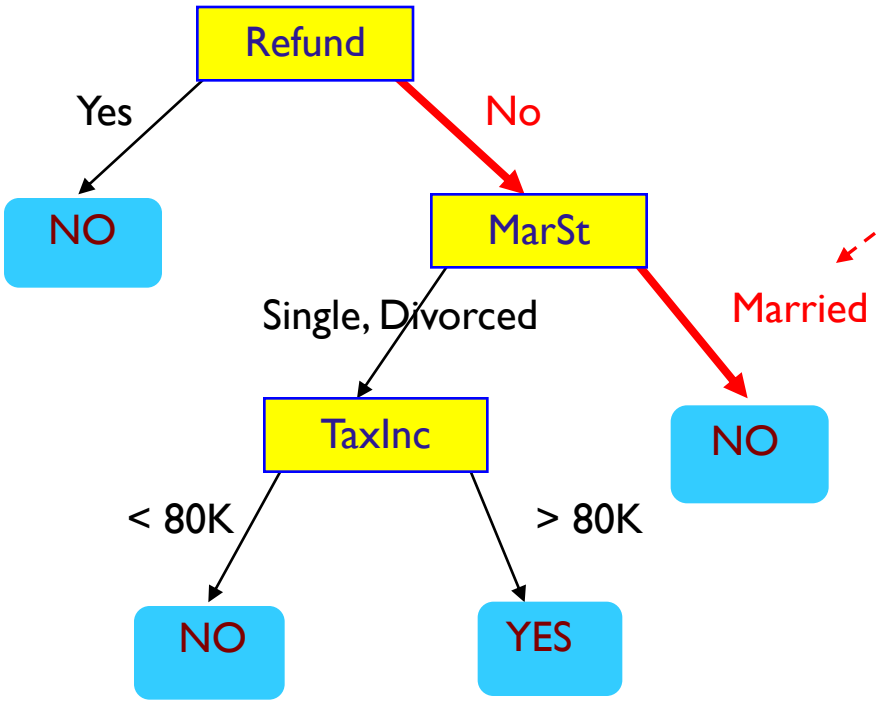
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

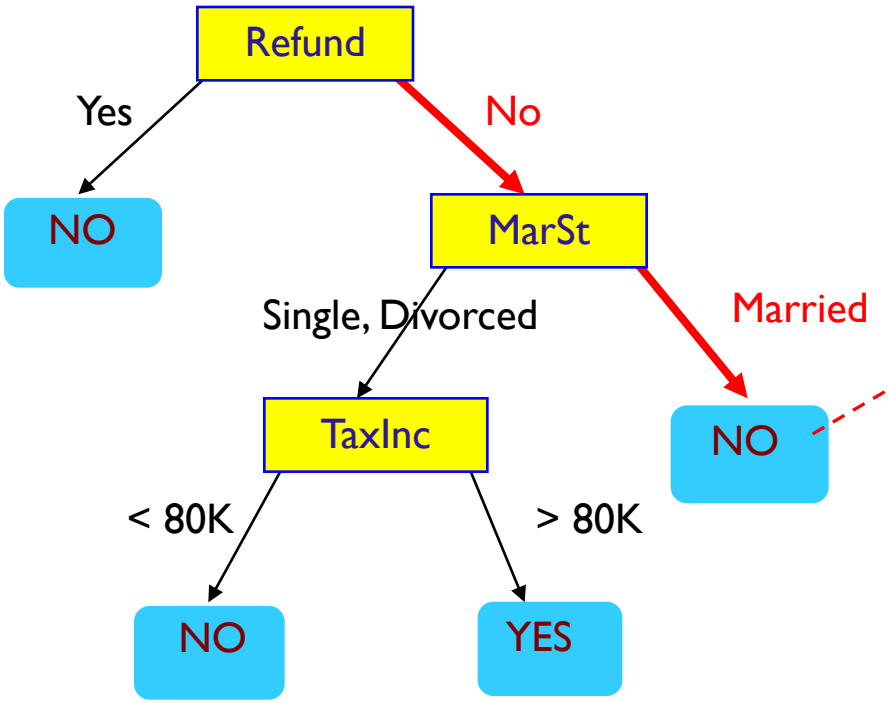
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to "No"

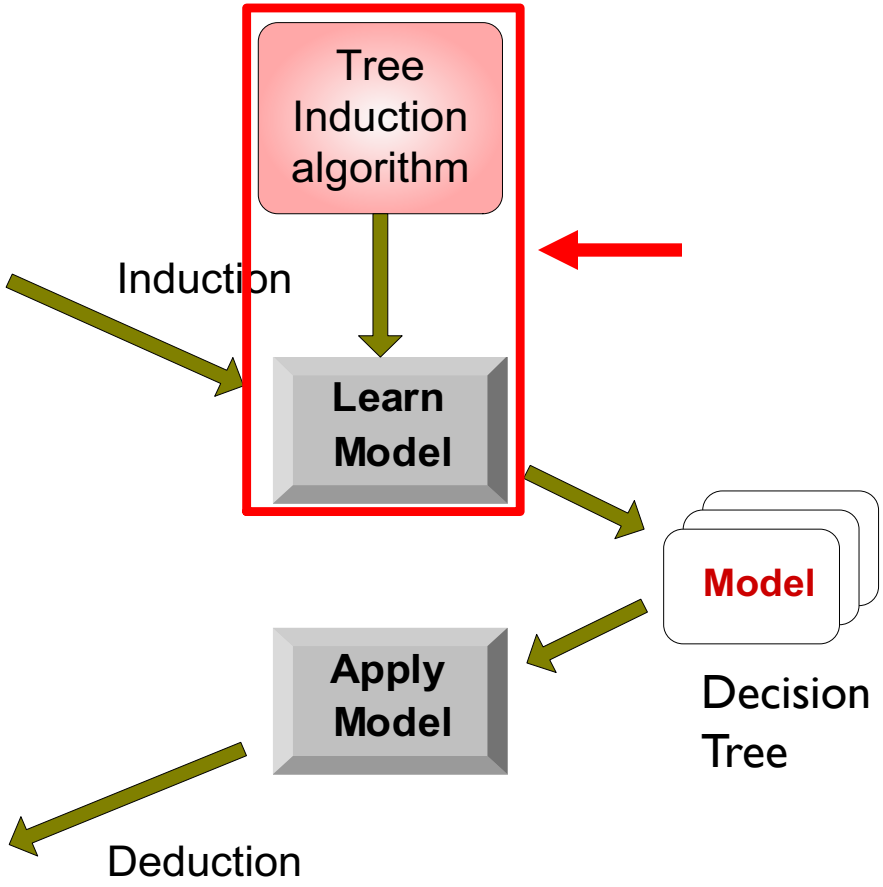
Decision Tree Classification Task

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	?

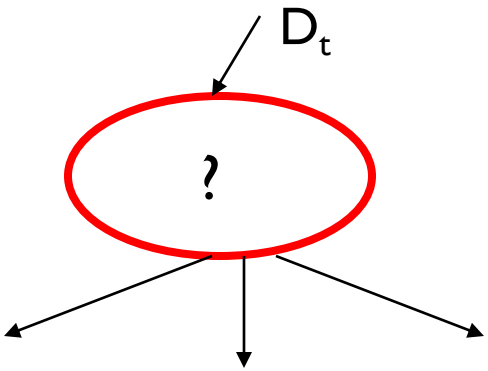
Test Set



Tree Induction – General Procedure

- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t contains records that belong to the same class y_t, then t is a leaf node labeled as y_t
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
  - 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.
    - If a node cannot be further split and it belongs to more than one class, use the class with more records

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



Tree Induction

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

Tree Induction

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

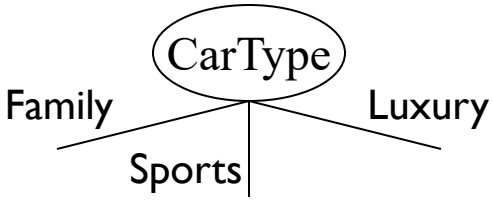
How to Specify Test Condition?

- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous

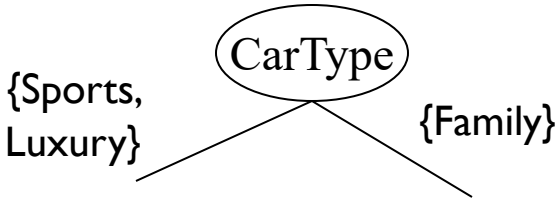
- Depends on number of ways to split
  - 2-way split
  - Multi-way split

Splitting Based on Nominal Attributes

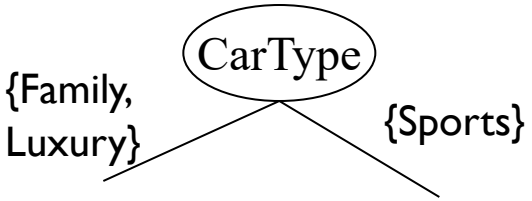
- Multi-way split: Use as many partitions as distinct values.



- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.

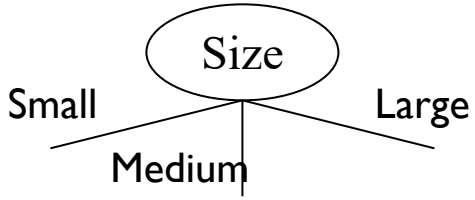


OR

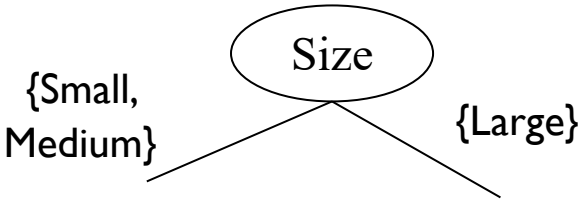


Splitting Based on Ordinal Attributes

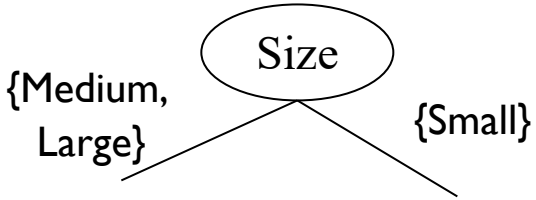
- Multi-way split: Use as many partitions as distinct values.



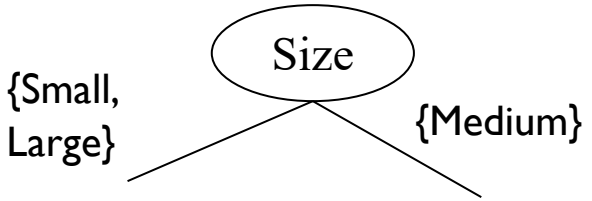
- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.



OR



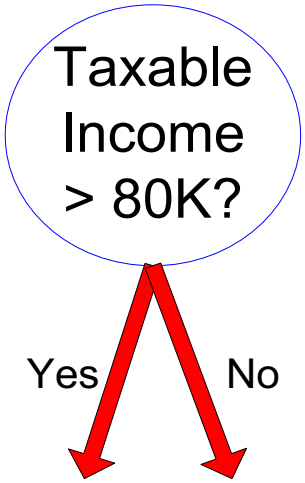
- What about this split?



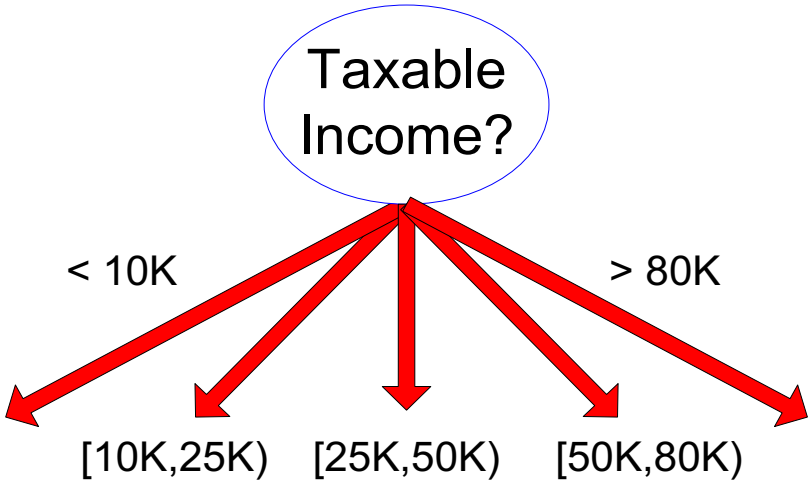
Splitting Based on Continuous Attributes

- Different ways of handling
  - Discretization to form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.
  - Binary Decision: $(A < v)$ or $(A \geq v)$
    - consider all possible splits and finds the best cut
    - can be more compute intensive

Splitting Based on Continuous Attributes



(i) Binary split



(ii) Multi-way split

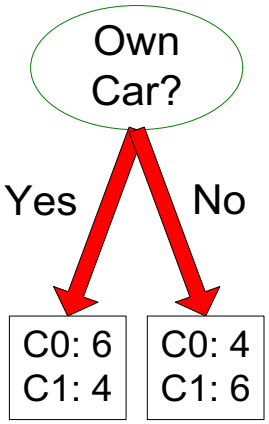
Tree Induction

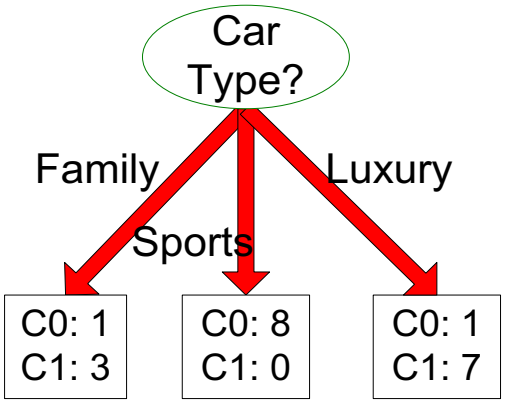
- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion

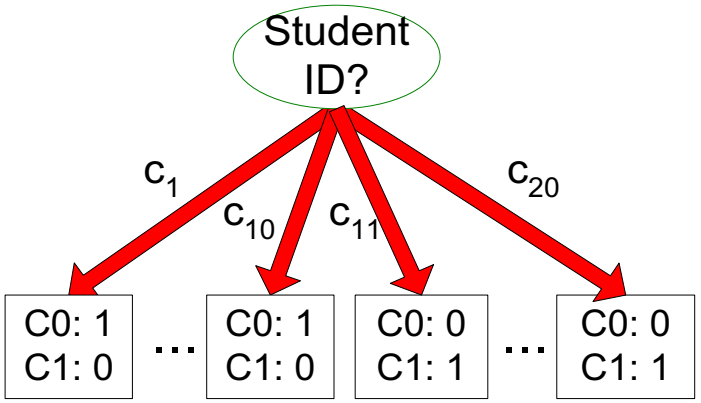
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

How to determine the Best Split

Before Splitting: 10 records of class 0,
10 records of class 1







Which test condition is the best?

How to determine the Best Split

- Greedy approach:
  - Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

C0: 5 C1: 5

Non-homogeneous,

High degree of impurity

C0: 9 C1: 1

Homogeneous,

Low degree of impurity

Measures of Node Impurity

- Gini Index (most basic)
- Entropy (not discussed in this module)
- Misclassification error (not discussed in this module)

Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Maximum ($1 - 1/n_c$) when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Examples for computing GINI

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

Splitting Based on GINI

- When a node p is split into k partitions (children), the quality of split is computed as,

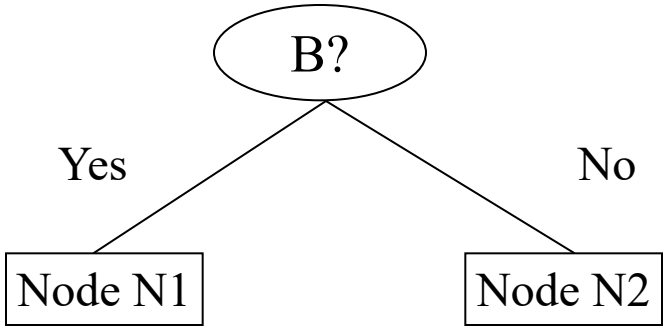
$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where, n_i = number of records at child i,

n = number of records at node p.

Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



Gini(N1)
= 1 – $(5/7)^2$ – $(2/7)^2$
= 0.408

Gini(N2)
= 1 – $(1/5)^2$ – $(4/5)^2$
= 0.32

	N1	N2
C1	5	1
C2	2	4
Gini=0.371		

Gini(Children)
= 7/12 * 0.408 +
5/12 * 0.32
= 0.371

Categorical Attributes: Computing Gini Index

- For each distinct value, gather counts for each class in the dataset
- Use the count matrix to make decisions

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split
(find best partition of values)

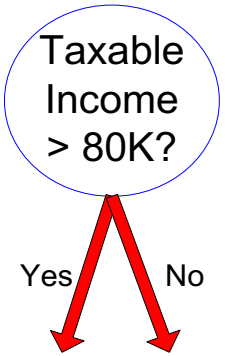
	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Continuous Attributes: Computing Gini Index

- Use Binary Decisions based on one value
- Several Choices for the splitting value
  - Number of possible splitting values = Number of distinct values
- Each splitting value has a count matrix associated with it
  - Class counts in each of the partitions, A < v and A ≥ v
- Simple method to choose best v
  - For each v, scan the database to gather count matrix and compute its Gini index
  - Computationally Inefficient! Repetition of work.

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



Continuous Attributes: Computing Gini Index...

- For efficient computation: for each attribute,
  - Sort the attribute on values
  - Linearly scan these values, each time updating the count matrix and computing gini index
  - Choose the split position that has the least gini index

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
Sorted Values →	60	70	75	85	90	95	100	120	125	220
Split Positions →	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	***0.300***	0.343	0.375	0.400	0.420

Tree Induction

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion

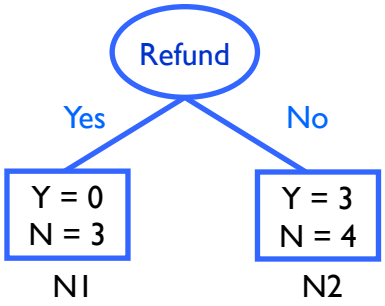
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

Stopping Criteria for Tree Induction

- Stop expanding a node when all the records belong to the same class
- Stop expanding a node when all the records have similar attribute values
  - Use the class of the majority
- Early termination (not the scope of this module)

Example – Root Node

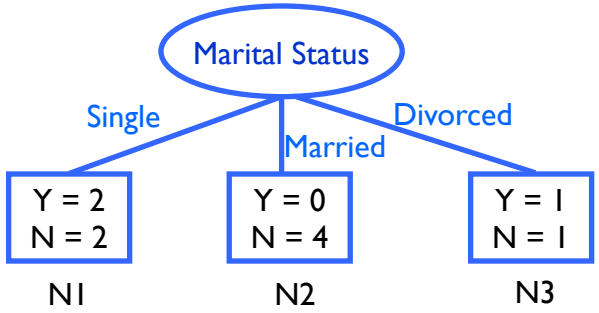
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Gini(N1) = 1 – $(0/3)^2$ – $(3/3)^2$
= 1 – 0.0 – 1.0 = 0.0

Gini(N2) = 1 – $(3/7)^2$ - $(4/7)^2$
= 1 – 0.184 – 0.327
= 0.489

Gini(Refund) = 3/10 * 0.0 + 7/10 * 0.489
= 0.3423

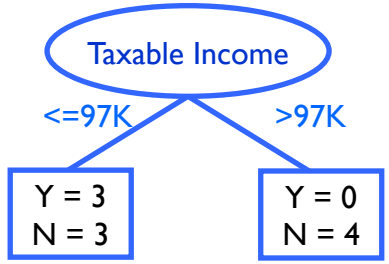


Gini(N1) = 1 – $(2/4)^2$ – $(2/4)^2$
= 1 – 0.25 – 0.25 = 0.5

Gini(N2) = 1 – $(0/4)^2$ - $(4/4)^2$
= 1 – 0.0 – 1.0 = 0.0

Gini(N3) = 1 – $(1/2)^2$ – $(1/2)^2$
= 1 – 0.25 – 0.25 = 0.5

Gini(MS) = 4/10 * 0.5 + 4/10 * 0.0 + 2/10 * 0.5
= 0.3

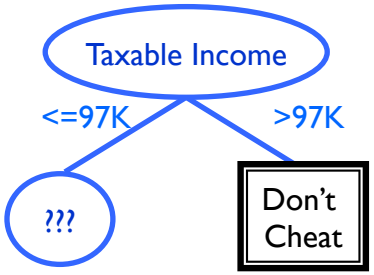


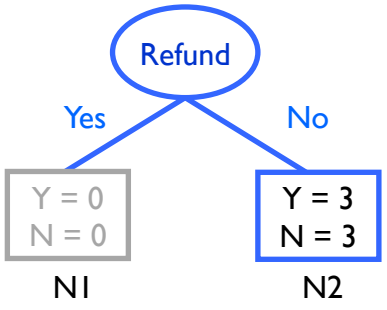
Gini(TI) = 0.3
* Please see slide p.45

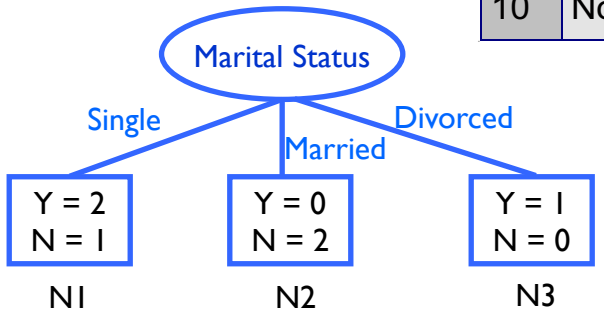
In this example, we proceed with **Taxable Income** as the root node.

Example – Level 1

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





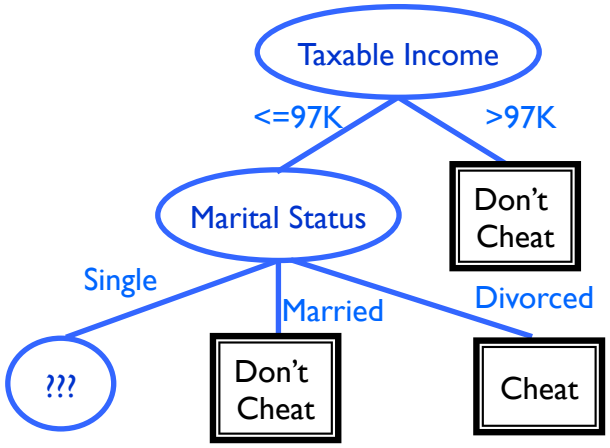


Gini(N2) = $1 - (3/6)^2 - (4/6)^2$
= 0.5
Gini(Refund) = 0.5

Gini(N1) = $1 - (2/3)^2 - (1/3)^2 = 0.444$
Gini(N2) = 0.0
Gini(N3) = 0.0
Gini(MS) = 3/6 * 0.444 0.0 + 0.0 = 0.222

Proceed with:
Martial Status

Example – Level 2

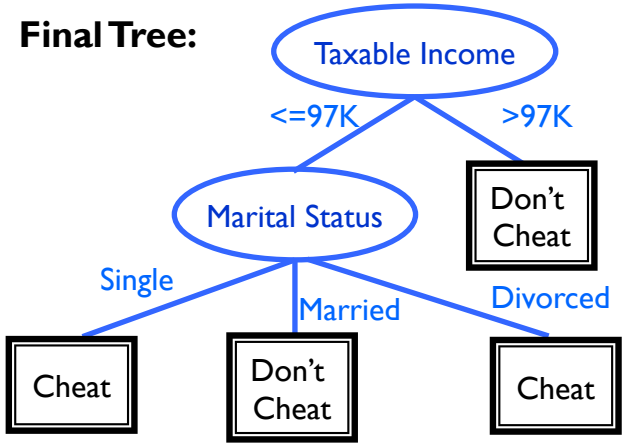


Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

For the rest of the records:

- "Refund" cannot differentiate those classes
- 2 Yes & 1 No remaining
- We assume "Single" leads to "Yes" in this case (Yes has more)

Final Tree:



Decision Tree Based Classification

- Advantages:
  - Inexpensive to construct
  - Extremely fast at classifying unknown records
  - Easy to interpret for small-sized trees
  - Accuracy is comparable to other classification techniques for many simple data sets

Exercise

- Please refer to the worksheet on Canvas

Association Analysis

Based on the slides of

- *Introduction to Data Mining*
  by Tan, Steinbach, and Kumar
- *Association Analysis*
  by Han and Pei

What Is Frequent Pattern Mining?

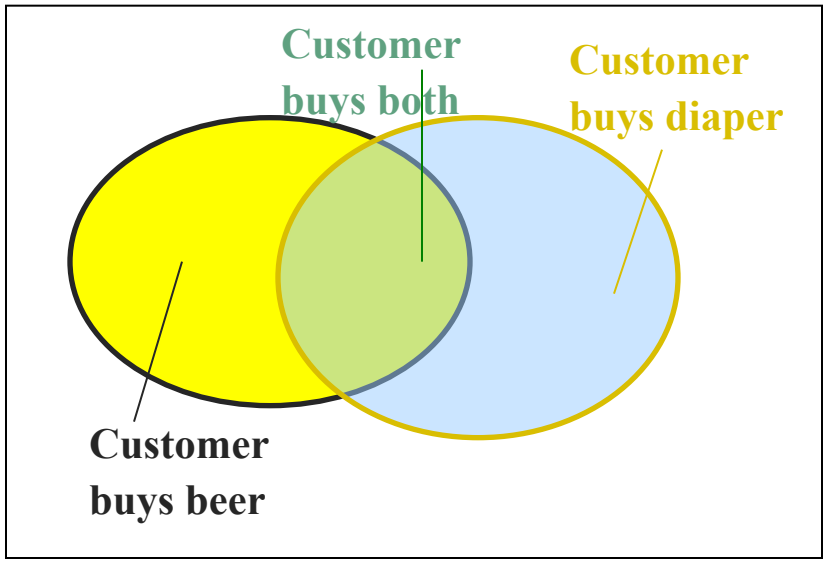
- Frequent pattern: pattern (set of items, sequence, etc.) that occurs frequently in a database
- Frequent pattern mining: finding regularities in data
  - What products were often purchased together?
  - What are the subsequent purchases after buying a PC?

Why Is Frequent Pattern Mining an Essential Task in Data Mining?

- Foundation for many essential data mining tasks
  - Association, correlation, causality
  - Sequential patterns, temporal or cyclic association, partial periodicity, spatial and multimedia association
  - Associative classification, cluster analysis, iceberg cube, fascicles (semantic data compression)
- Broad applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis
  - Web log (click stream) analysis, DNA sequence analysis, etc.

Basic Concepts: Frequent Patterns and Association Rules

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F



- Itemset $X=\{x_1, \ldots, x_k\}$
- Find all the rules $X \rightarrow Y$ with min confidence and support
  - support, *s*, probability that a transaction contains $X \cup Y$
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*.

Let min_support = 50%, min_conf = 50%:

$A \rightarrow C$ (50%, 66.7%)

$C \rightarrow A$ (50%, 100%)

Concept: Frequent Itemsets

Outlook	Temperature	Humidity	Play
sunny	hot	high	no
sunny	hot	high	no
overcast	hot	high	yes
rainy	mild	high	yes
rainy	cool	normal	yes
rainy	cool	normal	no
overcast	cool	normal	yes
sunny	mild	high	no
sunny	cool	normal	yes
rainy	mild	normal	yes
sunny	mild	normal	yes
overcast	mild	high	yes
overcast	hot	normal	yes
rainy	mild	high	no

- Minimum Support = 2
  - {sunny, hot, no}
  - {sunny, hot, high, no}
  - {rainy, normal}
  - …
- Minimum Support = 3
  - ?
- How strong is {sunny, no}?
  - Count =
  - Percentage =

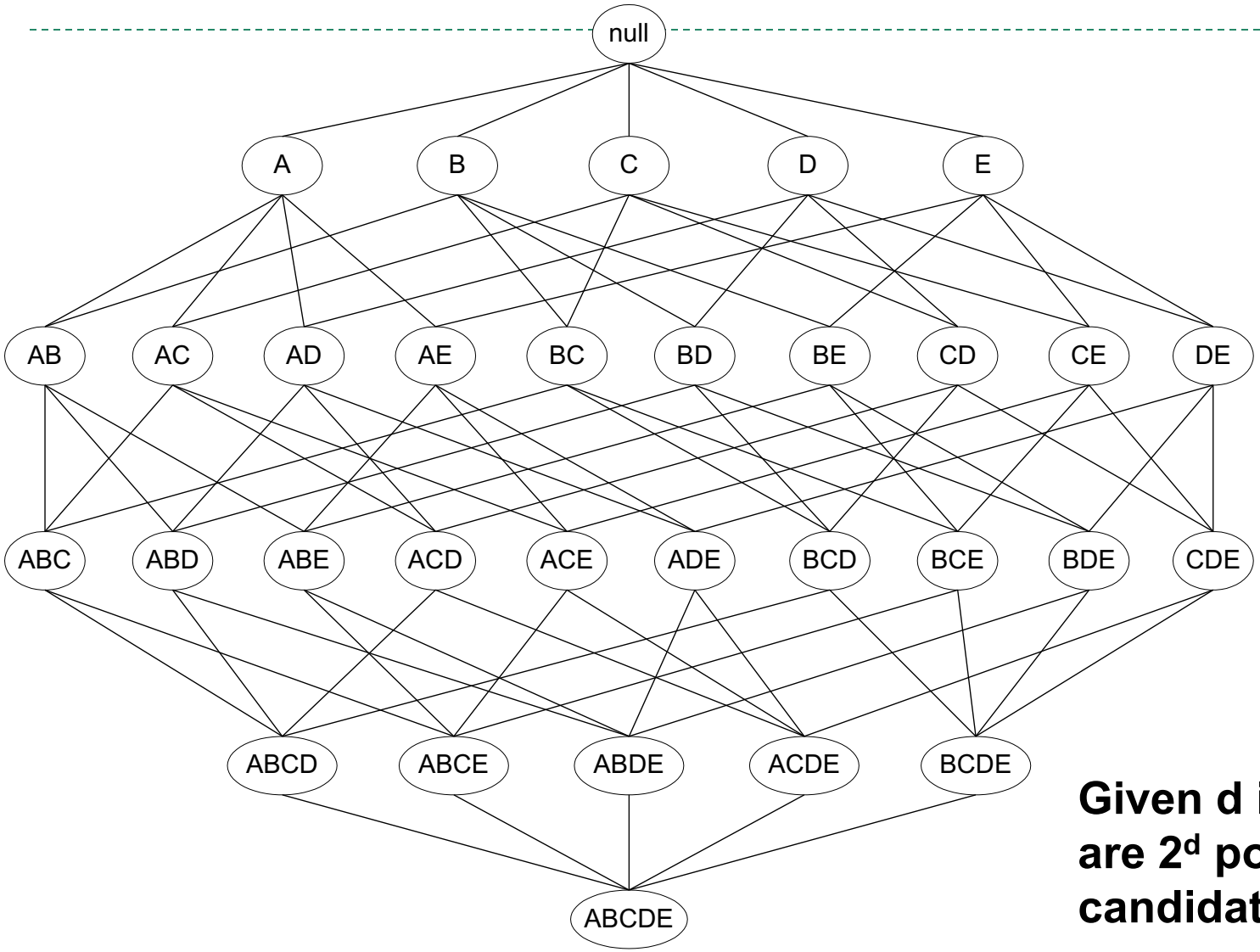
Concept: Itemset → Rules

- {sunny, hot, no} = {Outlook=Sunny, Temp=hot, Play=no}
- Generate a rule:
  - Outlook=sunny and Temp=hot ➔ Play=no
- How strong is this rule?
- Support of the rule
  - = support of the itemset {sunny, hot, no} = 2 = Pr({sunny, hot, no})
  - Either expressed in count form or percentage form
- Confidence = Pr(Play=no | {Outlook=sunny, Temp=hot})
  - In general LHS→ RHS, Confidence = Pr(RHS|LHS)
- Confidence
  - =Pr(RHS|LHS)
  - =count(LHS and RHS) / count(LHS)
- What is the confidence of Outlook=sunny➔Play=no?

Frequent Patterns

- Patterns = Item Sets
  - {i1, i2, … in}, where each item is a pair: (Attribute=value)
- Frequent Patterns (Frequent Itemsets)
  - Itemsets whose support >= minimum support
- Support
  - `count(itemset)/count(database)`

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Max-patterns

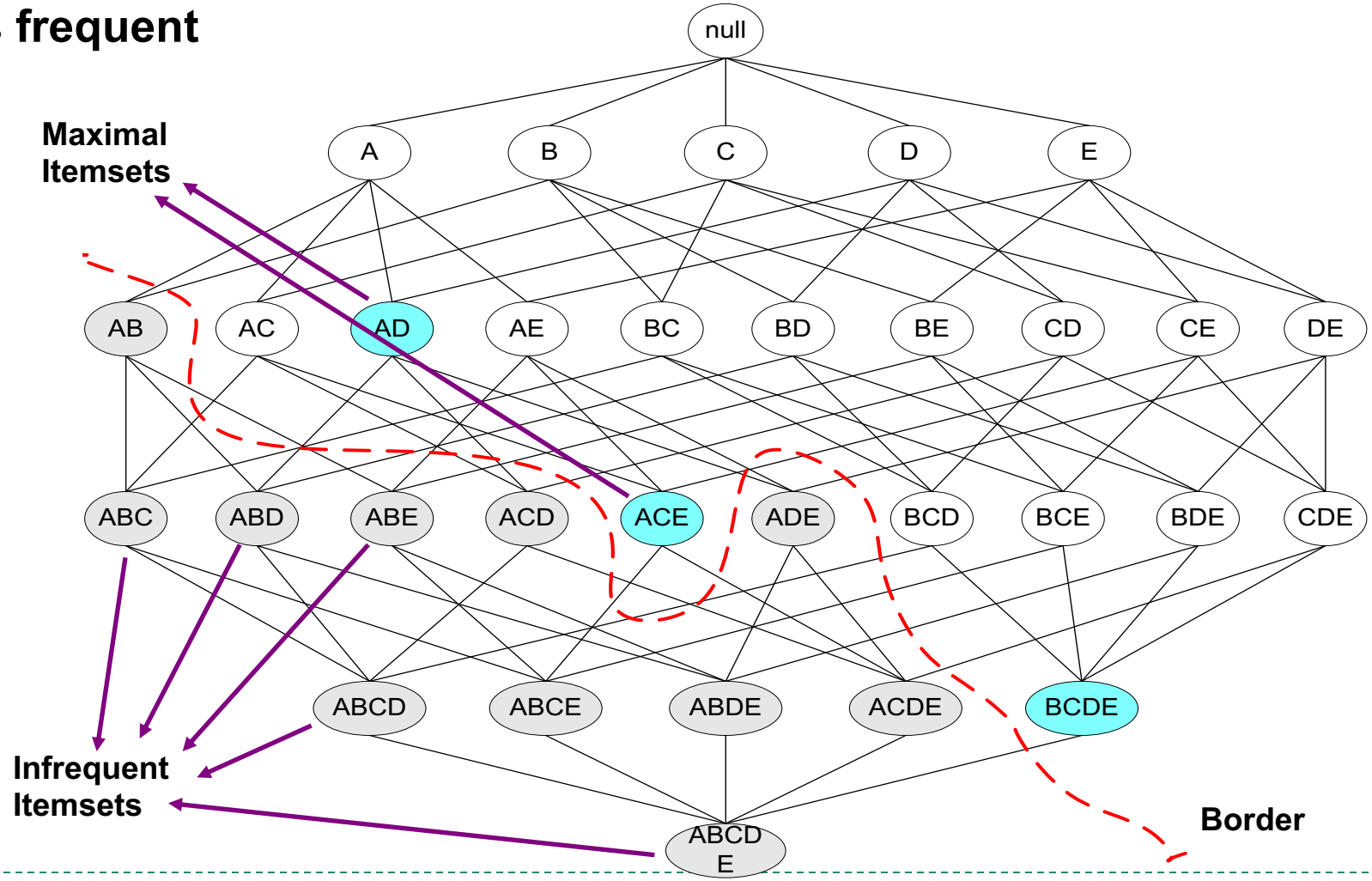
- Max-pattern: frequent patterns without proper frequent super pattern
  - BCDE, ACD are max-patterns
  - BCD is not a max-pattern

Min_sup=2

Tid	Items
10	A,B,C,D,E
20	B,C,D,E,
30	A,C,D,F

Maximal Frequent Itemset

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



Frequent Max Patterns

- Succinct Expression of frequent patterns
  - Let {a, b, c} be frequent
  - Then, {a, b}, {b, c}, {a, c} must also be frequent
  - Then {a}, {b}, {c}, must also be frequent
- By writing down {a, b, c} once, we save lots of computation
- Max Pattern
  - If {a, b, c} is a frequent max pattern, then {a, b, c, x} is **NOT** a frequent pattern, for any other item x.

Find Frequent Max Patterns

Outlook	Temperature	Humidity	Play
sunny	hot	high	no
sunny	hot	high	no
overcast	hot	high	yes
rainy	mild	high	yes
rainy	cool	normal	yes
rainy	cool	normal	no
overcast	cool	normal	yes
sunny	mild	high	no
sunny	cool	normal	yes
rainy	mild	normal	yes
sunny	mild	normal	yes
overcast	mild	high	yes
overcast	hot	normal	yes
rainy	mild	high	no

- Minimum support = 2
  - {sunny, hot, no} ??

Mining Association Rules — An Example

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

Min. support 50%
Min. confidence 50%

Frequent pattern	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

For rule $A \Rightarrow C$:

support = support($\{A\} \cup \{C\}$) = 50%

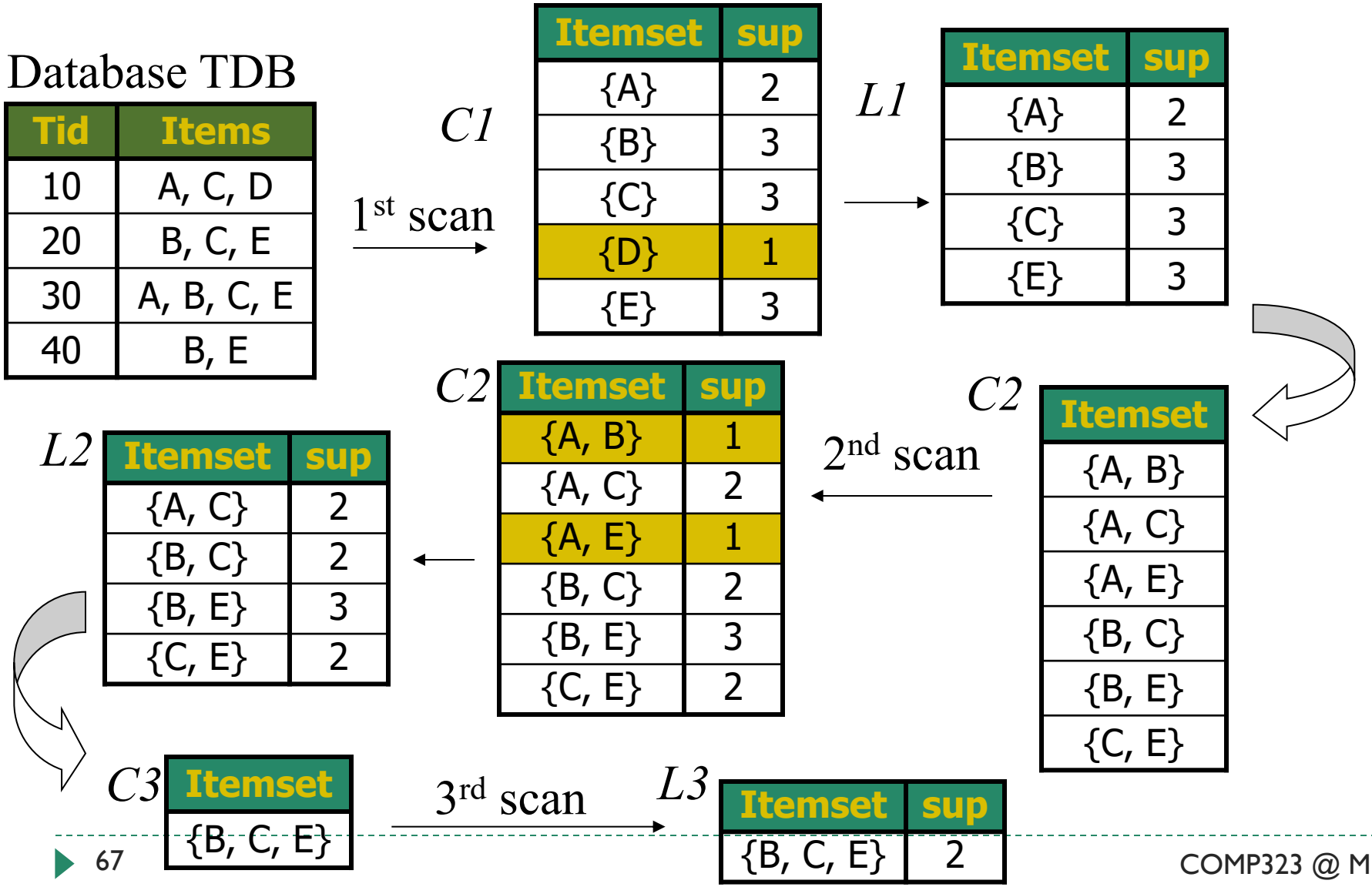
confidence = support($\{A\} \cup \{C\}$)/support($\{A\}$) = 66.6%

Apriori: A Candidate Generation-and-test Approach

- Any subset of a frequent itemset must be frequent
  - if **{beer, diaper, nuts}** is frequent, so is **{beer, diaper}**
  - Every transaction having {beer, diaper, nuts} also contains {beer, diaper}
- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!
- Method:
  - generate length (k+1) candidate itemsets from length k frequent itemsets, and
  - test the candidates against DB
- The performance studies show its efficiency and scalability

Agrawal & Srikant 1994, Mannila, et al. 1994

The Apriori Algorithm — An Example



Exercise

- Please refer to the worksheet on Canvas

Review questions

- Describe two aspects in which data warehouses serve as essential data source for data mining.
- Compare the different ways in which a business user applies OLAP and data mining.
- Compare and contrast the structure and size of data used in OLAP and data mining.
- Give an example to illustrate useful pattern that each of the following data mining techniques may detect for a [school / supermarket / credit card company / online shopping mall].
  - Clustering, decision tree and association analysis
- How to define a decision tree with the Gini approach?
- How to perform association analysis with the Apriori algorithm?