BITS Pilani

IS ZC415 – DATA MINING

Prof. Navneet Goyal
Department of Computer Science
BITS-Pilani, Pilani Campus

Importance of Data

"Data is the new 'oil' and there is a growing need for the ability to refine it,"
Dhiraj Rajaram, founder of Mu Sigma

BIG Data!!

Topics

What is Data Mining?

Data Mining Tasks

- Association Rules
- Clustering
- Classification & Prediction
- Sequence Discovery
- Regression
- Time-series Analysis

Applications

Necessity is the Mother of Invention

- **Data explosion**
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- **We are drowning in data, but starving for knowledge!**

Necessity is the Mother of Invention

- **We are drowning in data, but starving for knowledge!**

- **Solution**

  Data Mining

  Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data vs. Information

Society produces massive amounts of data

- business, science, medicine, economics, sports, …

Potentially valuable resource

Raw data is useless

- need techniques to automatically extract information
- Data: recorded facts
- Information: patterns underlying the data

What is NOT Data Mining?

Originally a “statistician” term

- *Overusing of data to draw invalid inferences*

Bonferroni's theorem warns us that if there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity.

Famous example: David Rhine, a “parapsychologist" at Duke in the 1950's tested students for extrasensory perception" by asking them to guess 10 cards - red or black. He found about 1/1000 of them guessed all 10, and instead of realizing that is what you'd expect from random guessing, declared them to have ESP. When he retested them, he found they did no better than average.

His conclusion: telling people they have ESP causes them to lose it!

What is NOT Data Mining?

Searching a phone number in a phone book

Searching a keyword on Google

Generating histograms of salaries for different age groups

Issuing SQL query to a database and reading the reply

Data Mining is NOT

- Data Warehousing
- (Deductive) query processing
  - SQL/ Reporting
- Software Agents
- Expert Systems
- Online Analytical Processing (OLAP)
- Statistical Analysis Tool
- Data visualization

What is Data Mining?

Discovery of useful summaries of data - Ullman

Extracting or “Mining” knowledge form large amounts of data

The efficient discovery of previously unknown patterns in large databases

Technology which predict future trends based on historical data

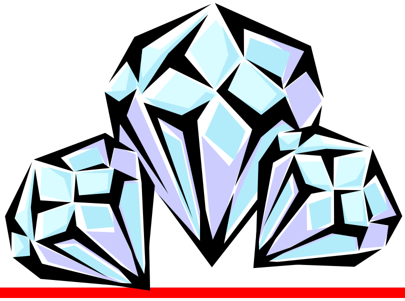
It helps businesses make proactive and knowledge-driven decisions

Data Mining vs. KDD

The name “Data Mining” a misnomer?

What Is Data Mining?

- Data mining:
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases



Data Mining

Programs that detect patterns and rules in the data

Strong patterns can be used to make non-trivial predictions on new data

Data Mining

Data mining is ready for application in the business & scientific community because it is supported by three technologies that are now sufficiently mature:

- **Massive data collection**
- **Powerful multiprocessor computers**
- **Data mining algorithms**

Data Mining:
On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

History of Data Mining

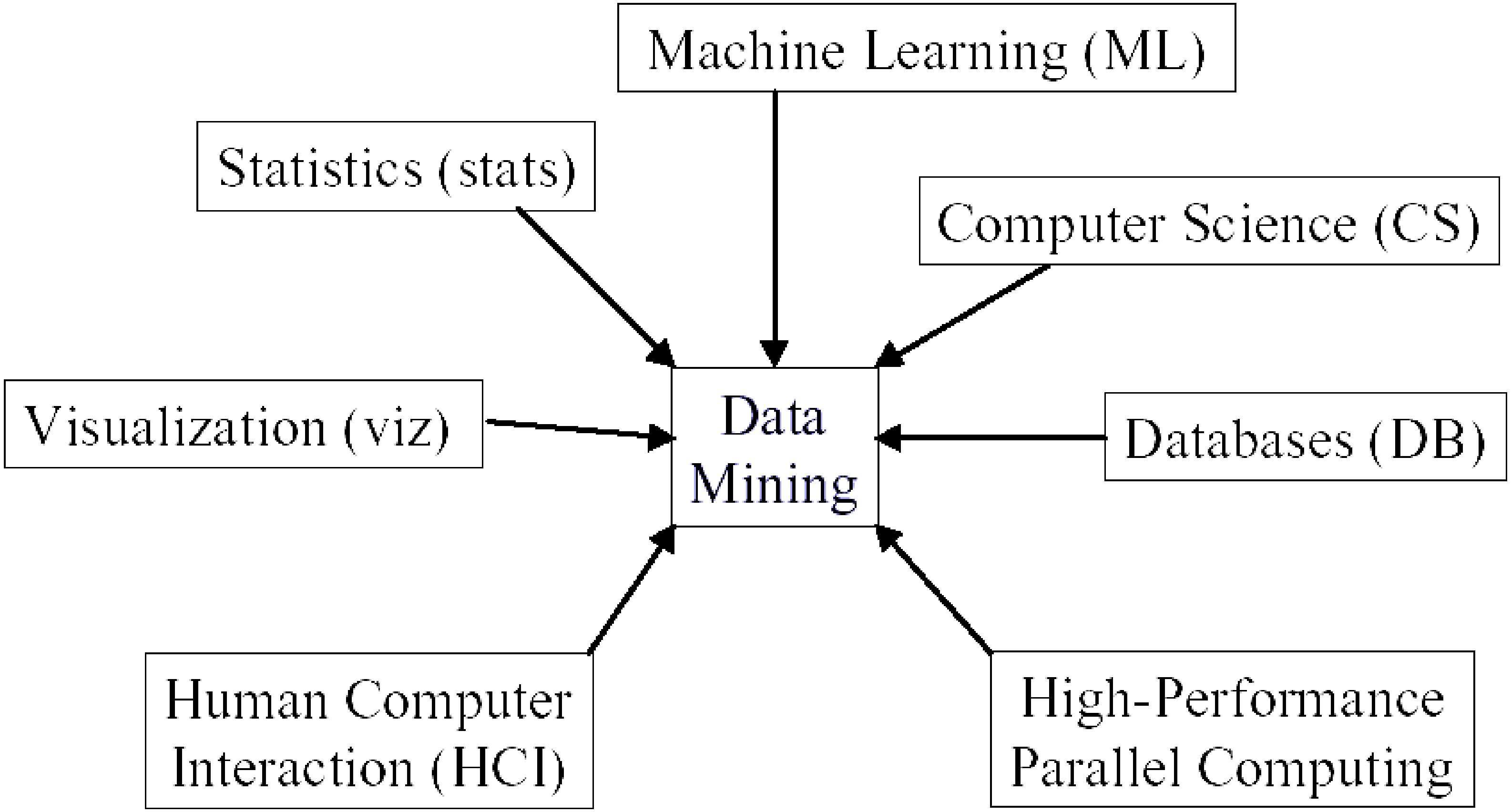
Emerged late 1980s

Flourished –1990s

Roots traced back along three family lines

- **Classical Statistics**
- **Artificial Intelligence**
- **Machine Learning**

Data Mining



Machine Learning (ML)

Statistics (stats)

Computer Science (CS)

Visualization (viz)

Data Mining

Databases (DB)

Human Computer Interaction (HCI)

High-Performance Parallel Computing

Some Humour

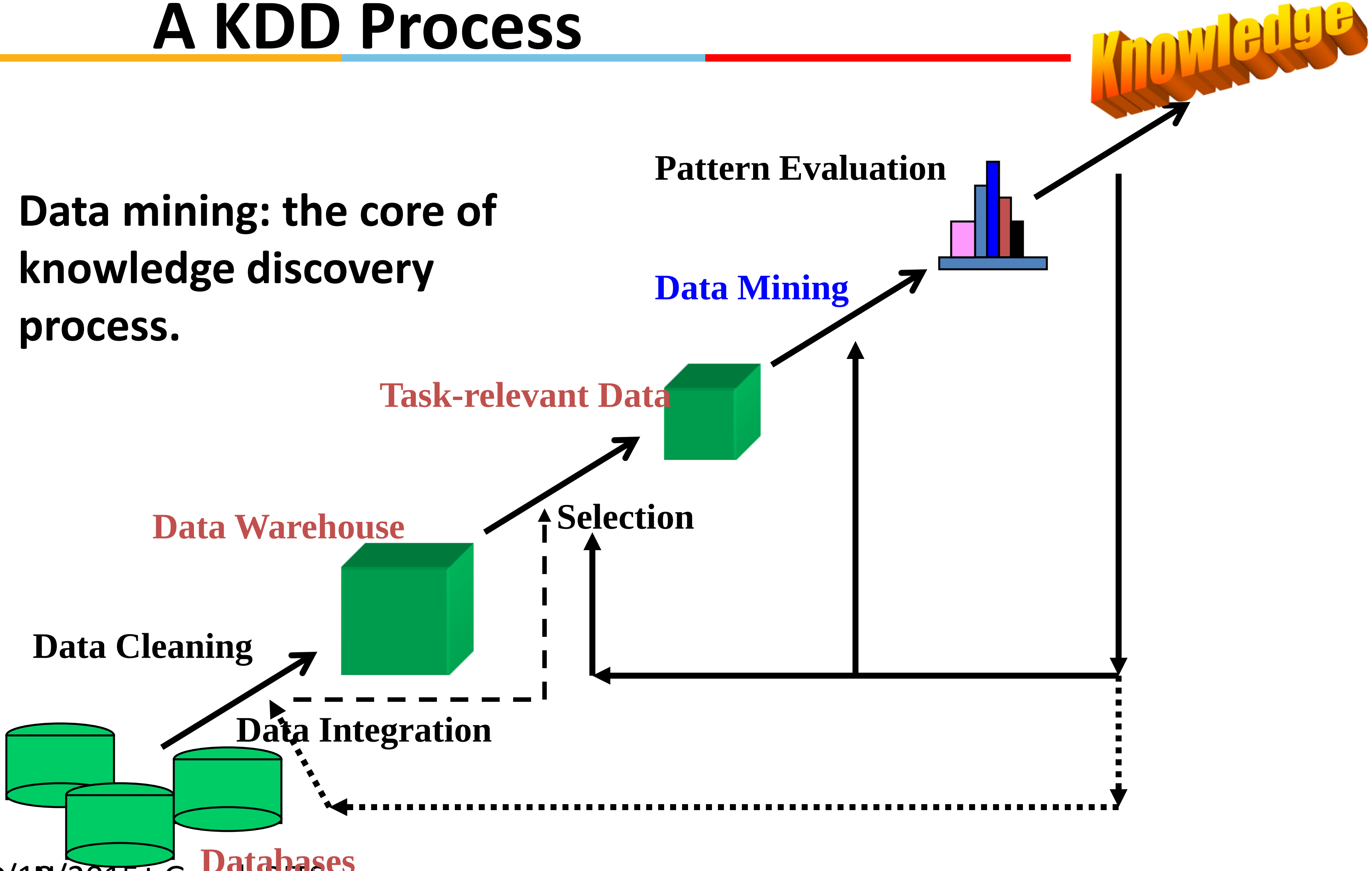
What is the difference between statistics, machine learning, AI and data mining?

- If there are up to 3 variables, it is statistics.
- If the problem is NP-complete, it is machine learning.
- If the problem is PSPACE-complete, it is AI.
- If you don't know what is PSPACE-complete, it is data mining.

Source – http://www.kdnuggets.com/2012/12/machine-learning-data-mining-humor.html

Data Mining: A KDD Process

Data mining: the core of knowledge discovery process.



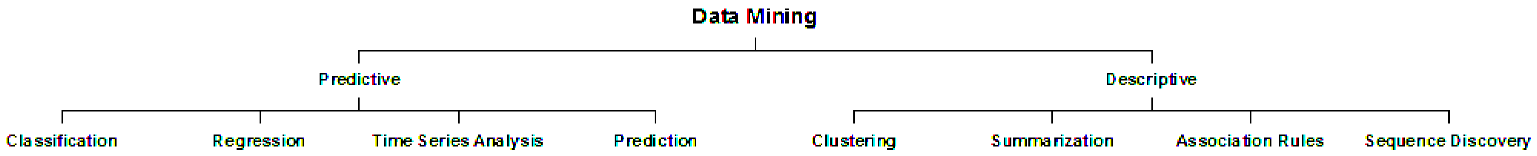
Stages of Data Mining Process

1. **Data gathering, e.g., data warehousing.**
2. **Data cleansing: eliminate errors and/or bogus data, e.g., patient fever = 125.**
3. **Feature extraction: obtaining only the interesting attributes of the data, e.g., “date acquired” is probably not useful for clustering celestial objects, as in Skycat.**
4. **Pattern extraction and discovery. This is the stage that is often thought of as “data mining” and is where we shall concentrate our effort.**
5. **Visualization of the data.**
6. **Evaluation of results; not every discovered fact is useful, or even true! Judgment is necessary before following your software's conclusions.**

Data Mining

- Many different algorithms for performing many different tasks
- DM algorithms can be characterized as consisting of 3 parts:
  - Model
  - Preference
  - Search
- Model could be
  - Predictive
  - Descriptive

Data Mining



Data Mining

Predictive

Descriptive

Classification

Regression

Time Series Analysis

Prediction

Clustering

Summarization

Association Rules

Sequence Discovery

Predictive Model

Making prediction about values of data using known results from different data

Example: Credit Card Company

Every purchase is placed in 1 of 4 classes

1. Authorize
2. Ask for further identification before authorizing
3. Do not authorize
4. Do not authorize but contact police

Two functions of Data Mining

1. Examine historical data to determine how the data fit into 4 classes
2. Apply the model to each new purchase

Descriptive Model

Identifies patterns or relationship in data

Example: Later

Two Important Terms

Supervised Learning

- Training Data Set
- Model is told to which class each training data belongs
- Learning by example
- Example CLASSIFICATION
- Similar to Discriminate Analysis in Statistics

Unsupervised Learning

- Class-label of training set is not known
- No. of classes also may not be known
- Learning by observation
- Example CLUSTERING

Data Mining Applications

Some examples of “successes":

1. **Decision trees constructed from bank-loan histories to produce algorithms to decide whether to grant a loan.**

2. Patterns of traveler behavior mined to manage the sale of discounted seats on planes, rooms in hotels,etc.

3. “Diapers and beer." Observation that customers who buy diapers are more likely to by beer than average allowed supermarkets to place beer and diapers nearby, knowing many customers would walk between them. Placing potato chips between increased sales of all three items.
More Recently – Polo Shirts and Barbie Dolls!

Data Mining Applications

Some examples of “successes":

4. Skycat and Sloan Sky Digital Sky Survey: clustering sky objects by their radiation levels in different bands allowed astronomers to distinguish between galaxies, nearby stars, and many other kinds of celestial objects.
(168 million records and some 500 attributes)
for details see http://www.sdss.org/dr1/

5. Comparison of the genotype of people with/without a condition allowed the discovery of a set of genes that together account for many cases of diabetes. This sort of mining has become much more important as the human genome has fully been decoded

Examples

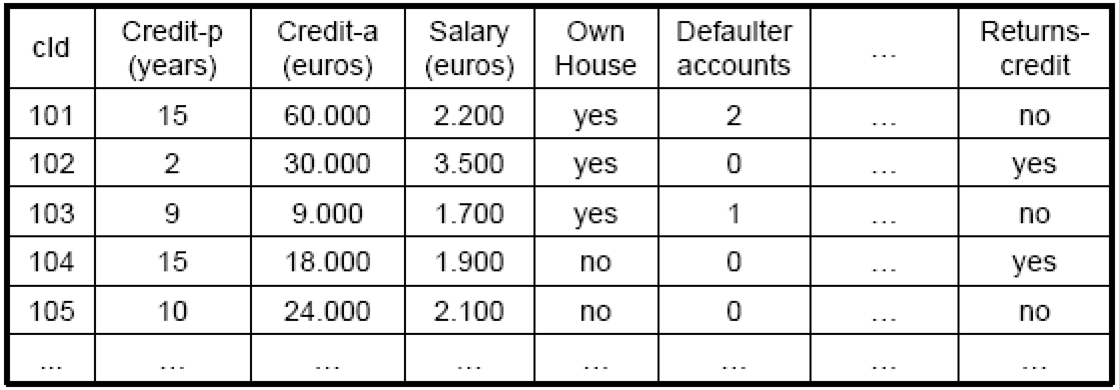
- **BANK AGENT:**
  - Must I grant a mortgage to this customer?
- **SUPERMARKET MANAGER:**
  - When customers buy eggs, do they also buy oil?
- **PERSONNEL MANAGER:**
  - What kind of employees do I have?
- **TRADER in a RETAIL COMPANY:**
  - How many flat TVs do we expect to sell next month?

Classification Example

- BANK AGENT:

Must I grant a mortgage to this customer?

Historical Data:



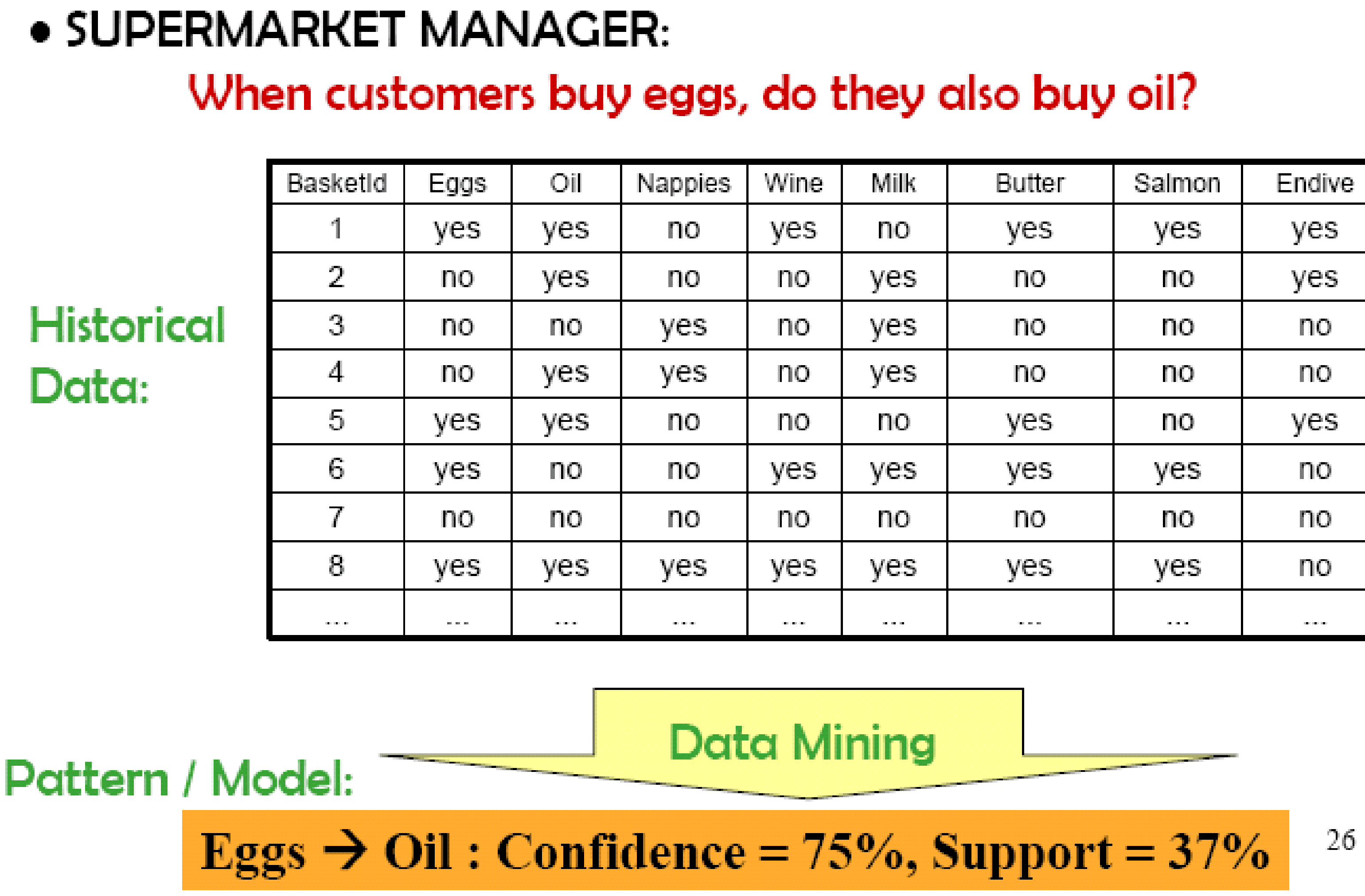
cId	Credit-p (years)	Credit-a (euros)	Salary (euros)	Own House	Defaulter accounts	…	Returns-credit
101	15	60.000	2.200	yes	2	…	no
102	2	30.000	3.500	yes	0	…	yes
103	9	9.000	1.700	yes	1	…	no
104	15	18.000	1.900	no	0	…	yes
105	10	24.000	2.100	no	0	…	no
…	…	…	…	…	…	…	…

Pattern / Model:

If Defaulter-accounts > 0 **then** Returns-credit = no

If Defaulter-accounts = 0 **and** [(Salary > 2.500) **or** (credit-p > 10)] **then** Returns-credit = yes

Association Rule: Example



- SUPERMARKET MANAGER:

When customers buy eggs, do they also buy oil?

Historical Data:

BasketId	Eggs	Oil	Nappies	Wine	Milk	Butter	Salmon	Endive	...
1	yes	yes	no	yes	no	yes	yes	yes	...
2	no	yes	no	no	yes	no	no	yes	...
3	no	no	yes	no	yes	no	no	no	...
4	no	yes	yes	no	yes	no	no	no	...
5	yes	yes	no	no	no	yes	no	yes	...
6	yes	no	no	yes	yes	yes	yes	no	...
7	no	no	no	no	no	no	no	no	...
8	yes	yes	yes	yes	yes	yes	yes	no	...
...	...	...	...	...	...	...	...	...	...

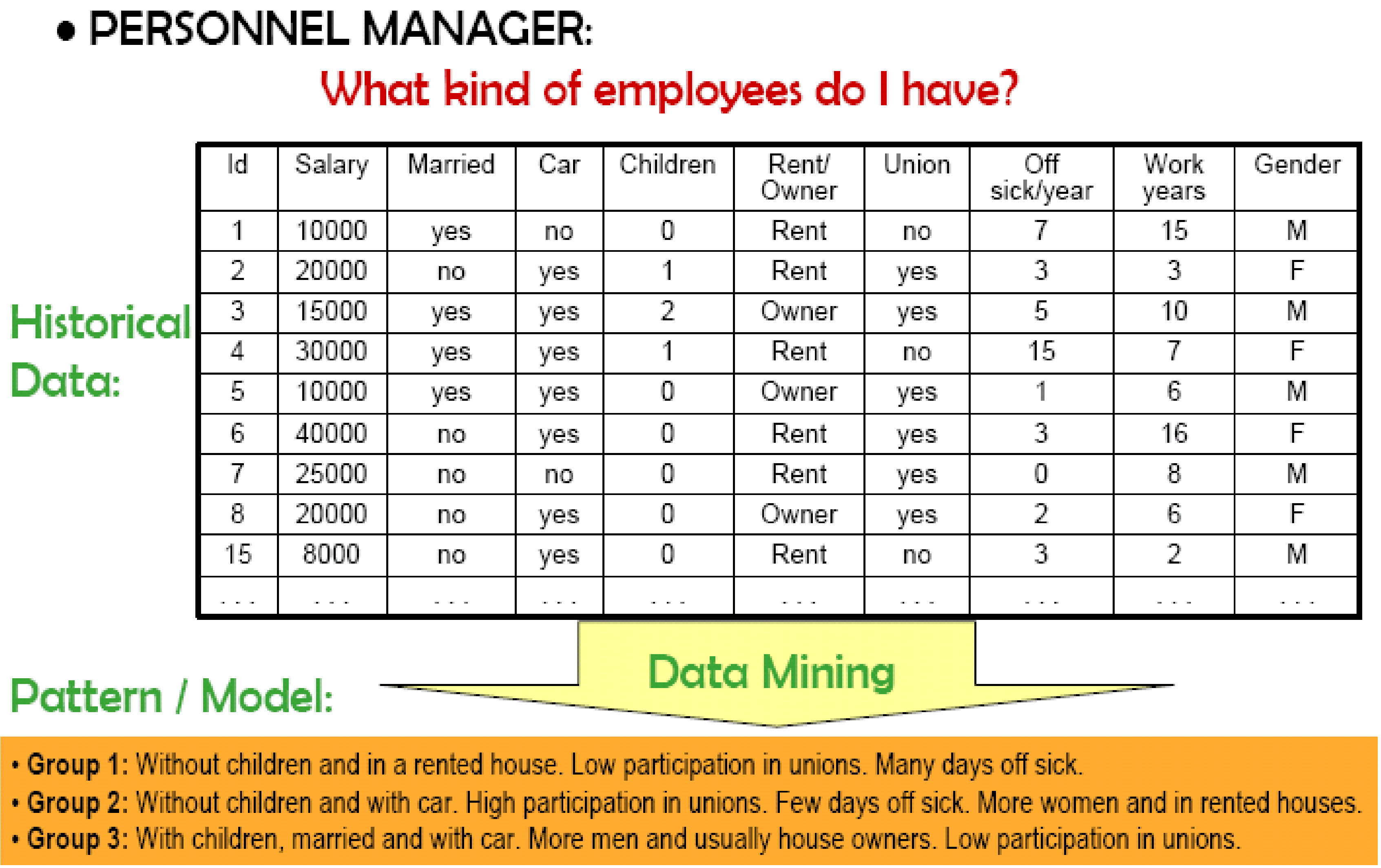
Data Mining

Pattern / Model:

Eggs → Oil : Confidence = 75%, Support = 37%

26

Clustering: Example



- PERSONNEL MANAGER:

What kind of employees do I have?

Historical Data:

Id	Salary	Married	Car	Children	Rent/ Owner	Union	Off sick/year	Work years	Gender
1	10000	yes	no	0	Rent	no	7	15	M
2	20000	no	yes	1	Rent	yes	3	3	F
3	15000	yes	yes	2	Owner	yes	5	10	M
4	30000	yes	yes	1	Rent	no	15	7	F
5	10000	yes	yes	0	Owner	yes	1	6	M
6	40000	no	yes	0	Rent	yes	3	16	F
7	25000	no	no	0	Rent	yes	0	8	M
8	20000	no	yes	0	Owner	yes	2	6	F
15	8000	no	yes	0	Rent	no	3	2	M
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Pattern / Model:

- **Group 1:** Without children and in a rented house. Low participation in unions. Many days off sick.
- **Group 2:** Without children and with car. High participation in unions. Few days off sick. More women and in rented houses.
- **Group 3:** With children, married and with car. More men and usually house owners. Low participation in unions.

Examples of Discovered Patterns

- Association rules
  - 98% of people who purchase diapers also buy beer
- Classification
  - People with age less than 25 and salary > 40k drive sports cars
- Similar time sequences
  - Stocks of companies A and B perform similarly
- Outlier Detection
  - Residential customers for telecom company with businesses at home

Association Rules & Frequent Itemsets

Market-Basket Analysis

Grocery Store: Large no. of ITEMS

Customers fill their market baskets with subset of items

98% of people who purchase diapers also buy beer

Used for shelf management

Used for deciding whether an item should be put on sale

Classification

Customer's name, age income_level and credit _rating known

Training Set

Use classification algorithm to come up with classification rules

If age between 31 & 40 and income_level= 'High', then credit_rating = 'Excellent'

New Data(customer): Sachin, age=31, income_level='High' implies credit_rating='Excellent'

Classifier Accuracy?

Hold-out, k-fold cross validation

Prediction vs Classification

Clustering

Given points in some space, often a high-dimensional space

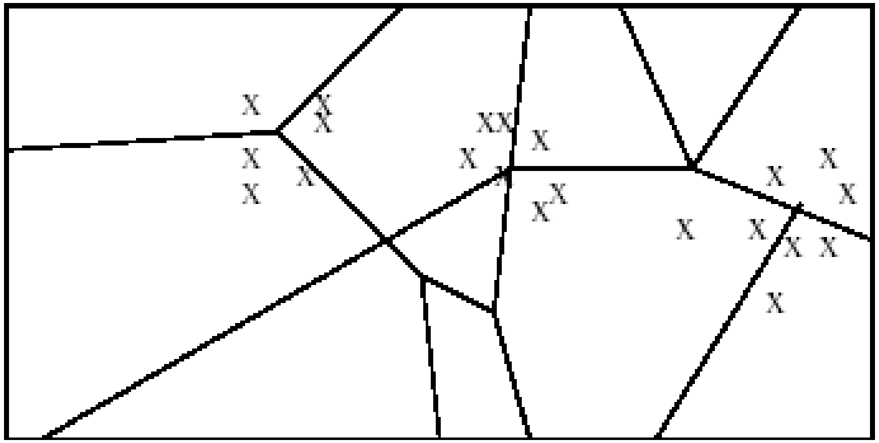
Group the points into a small number of clusters

Each cluster consisting of points that are “near” in some sense

Points in the same cluster are “similar” and are “dissimilar” to points in other clusters

Clustering: Examples

Cholera outbreak in London



Skycat clustered **$2x10^9$** sky objects into stars, galaxies, quasars, etc. Each object was a point in a space of 7 dimensions, with each dimension representing radiation in one band of the spectrum.

The Sloan Sky Survey is a more ambitious attempt to catalog and cluster the entire visible universe

Association Rules

Purchasing of one product when another product is purchased represents an AR

Used mainly in retail stores to

- Assist in marketing
- Shelf management
- Inventory control

Faults in Telecommunication Networks

Transaction Database

Item-sets, Frequent or large item-sets

Support & Confidence of AR

Association Rules

A rule must have some minimum user-specified *confidence*

1 & 2 => 3 has 90% confidence if when a customer bought 1 and 2, in 90% of cases, the customer also bought 3.

A rule must have some minimum user-specified *support*

1 & 2 => 3 should hold in some minimum percentage of transactions to have business value

AR X => Y holds with confidence T, if T% of transactions in DB that support X also support Y

Example

- Transaction Database

Transaction Id	Purchased Items
1	{1, 2, 3}
2	{1, 4}
3	{1, 3}
4	{2, 5, 6}

- For minimum support = 50%, minimum confidence = 50%, we have the following rules
  - 1 => 3 with 50% support and 66% confidence
  - 3 => 1 with 50% support and 100% confidence

Support & Confidence

I=Set of all items

D=Transaction Database

AR A=>B has support s if s is the %age of Txs in D that contain AUB

s(A=>B)=P(AUB)

AR A=>B has confidence c in D if c is the %age of Txs in D containing A that also contain B

c(A=>B)=P (B/A)=P(AUB)/P(A)

Mining Association Rules

2 Step Process

1. Find all frequent Itemsets is all itemsets satisfying *min_sup*
2. Generate strong ARs from frequent itemsets ie ARs satisfying *min_sup* & *min_conf*

Classification & Prediction

- What is Classification?
- What is Prediction?
- Any relationship between the two?
- Supervised or Unsupervised?
- Issues
- Applications
- Algorithms
- Classifier Accuracy

Classification & Prediction

- **Classification:**
  - **predicts categorical class labels**
  - **classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data**
- **Prediction:**
  - **models continuous-valued functions, i.e., predicts unknown or missing values**

Classification & Prediction

- **Given a database $D=\{t_1,t_2,\ldots,t_n\}$ and a set of classes $C=\{C_1,\ldots,C_m\}$, the *Classification Problem* is to define a mapping $f:D\rightarrow C$ where each t_i is assigned to one class.**

- ***Prediction* is similar, but may be viewed as having infinite number of classes.**

Applications

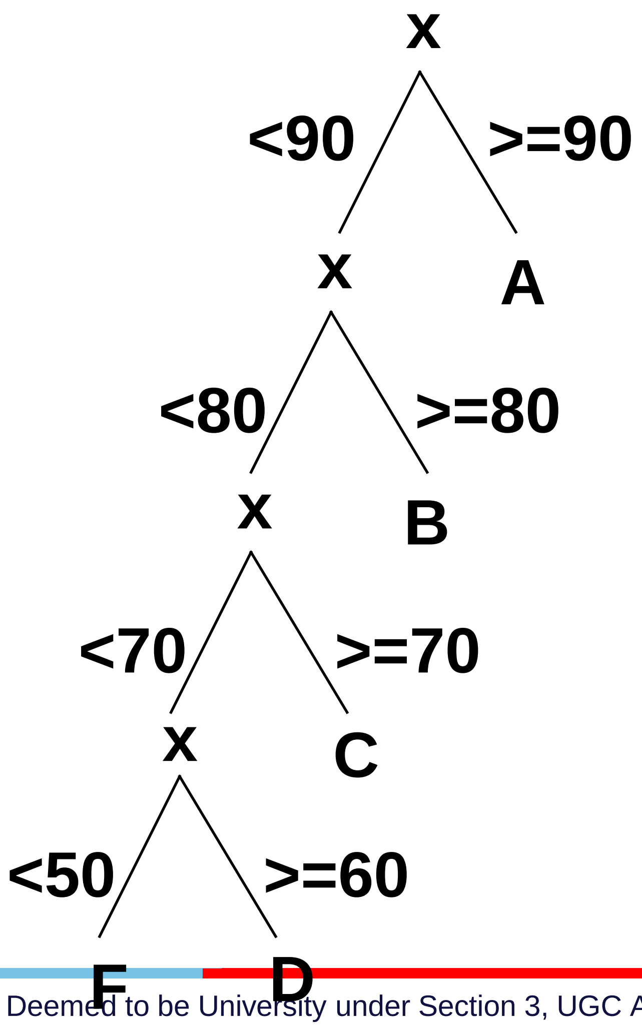
- Credit approval
- Target marketing
- Medical diagnosis
- Treatment effectiveness analysis
- Image recognition

Some More Applications

- Teachers classify students’ grades as A, B, C, D, or E.
- Identify mushrooms as poisonous or edible.
- Predict when a river will flood.
- Identify individuals with credit risks.
- Speech recognition
- Pattern recognition

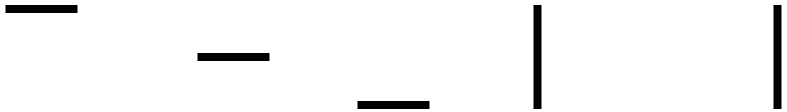
Grading: A Simple Example

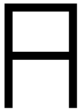
- If x >= 90 then grade =A.
- If 80<=x<90 then grade =B.
- If 70<=x<80 then grade =C.
- If 60<=x<70 then grade =D.
- If x<50 then grade =E.



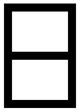
Classification Example: Letter Recognition

View letters as constructed from 5 components:

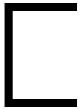




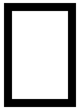
Letter A



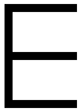
Letter **B**



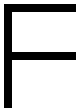
Letter C



Letter **D**



Letter E



Letter F

Supervised vs. Unsupervised Learning

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

Classification: A Two-Step Process

- **Model construction: describing a set of predetermined classes**
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the *class label attribute*
  - The set of tuples used for model construction: *training set*
  - The model is represented as classification rules, decision trees, or mathematical formulae

Performance Evaluation

- Accuracy of Classification
- Classification is a fuzzy problem, the correct answer may depend on user
- %age of tuples places in correct class
- Cost of incorrect assignment

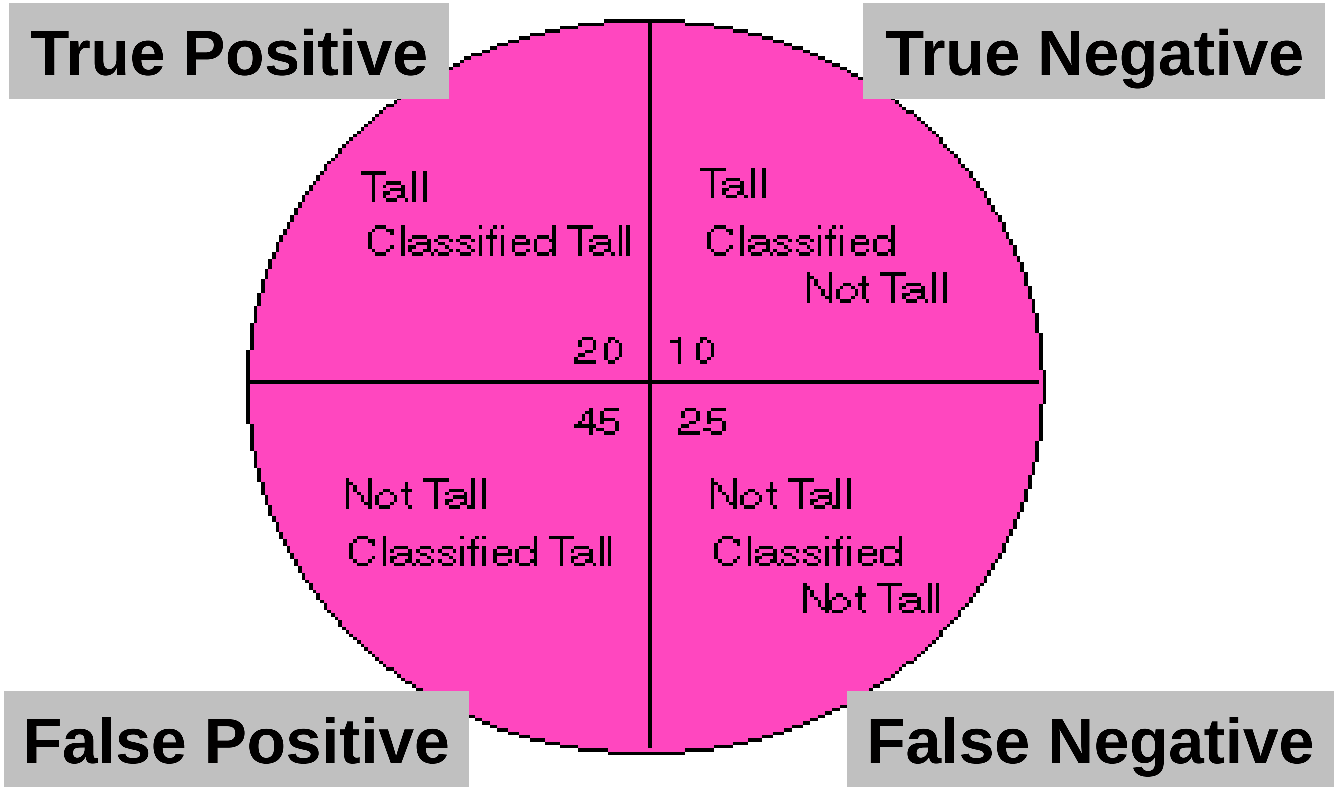
Classifier Accuracy

- **Partition: Training-and-testing (holdout)**
  - **use two independent data sets, e.g., training set (2/3), test set(1/3)**
  - **used for data set with large number of samples**
  - **Variation: random subsampling (repeated k times)**
- **K-fold Cross-validation**
  - **divide the data set into k subsamples**
  - **use k-1 subsamples as training data and one sub-sample as test data**
  - **training and testing is performed k times**
  - **for data set with moderate size**

Height Example Data

Name	Gender	Height	Output1	Output2
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Wo rth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

Classification Performance



True Positive

True Negative

Tall

Classified Tall

20

Tall

Classified Not Tall

10

45

Not Tall

Classified Tall

25

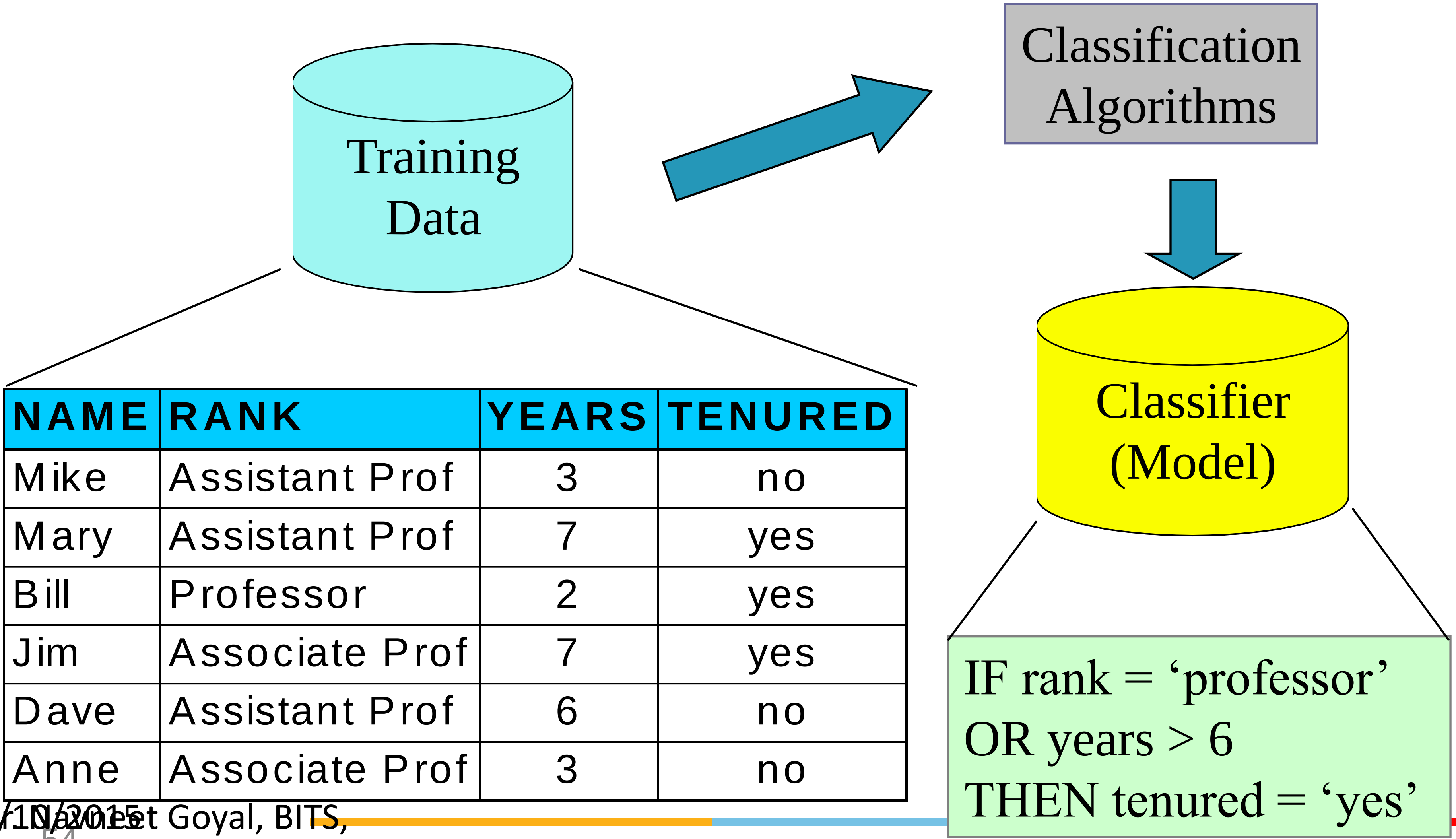
Not Tall

Classified Not Tall

False Positive

False Negative

Classification: Model Construction



Training Data

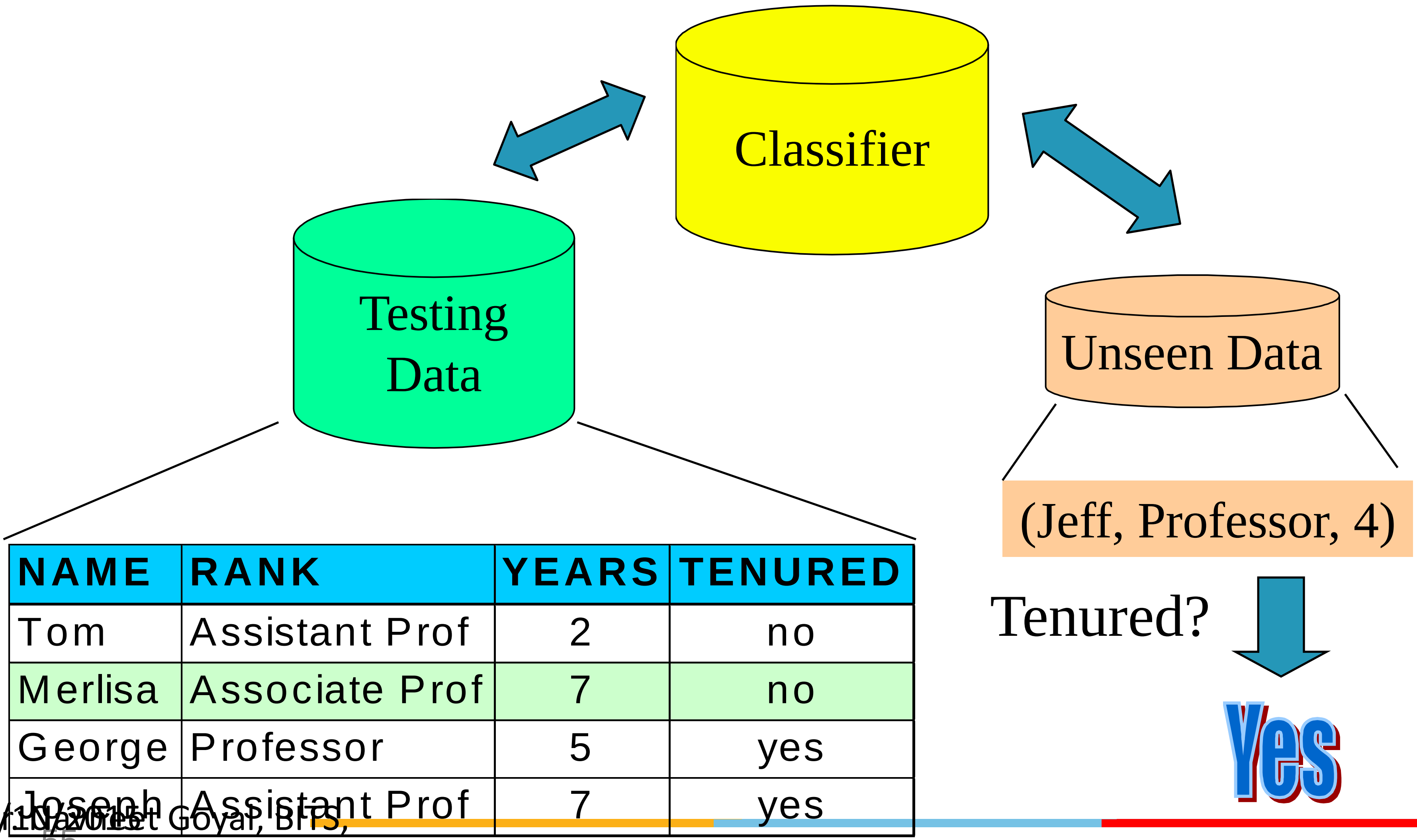
Classification Algorithms

Classifier (Model)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Classification: Use the Model



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Classification: Algorithms

- Classification by Decision Tree Induction
- Bayesian Classification
- Classification by Back Propagation

Clustering

Clustering of data is a method by which large sets of data is grouped into clusters of smaller sets of similar data.

Objects in one cluster have high similarity to each other and are dissimilar to objects in other clusters.

It is an example of unsupervised learning.

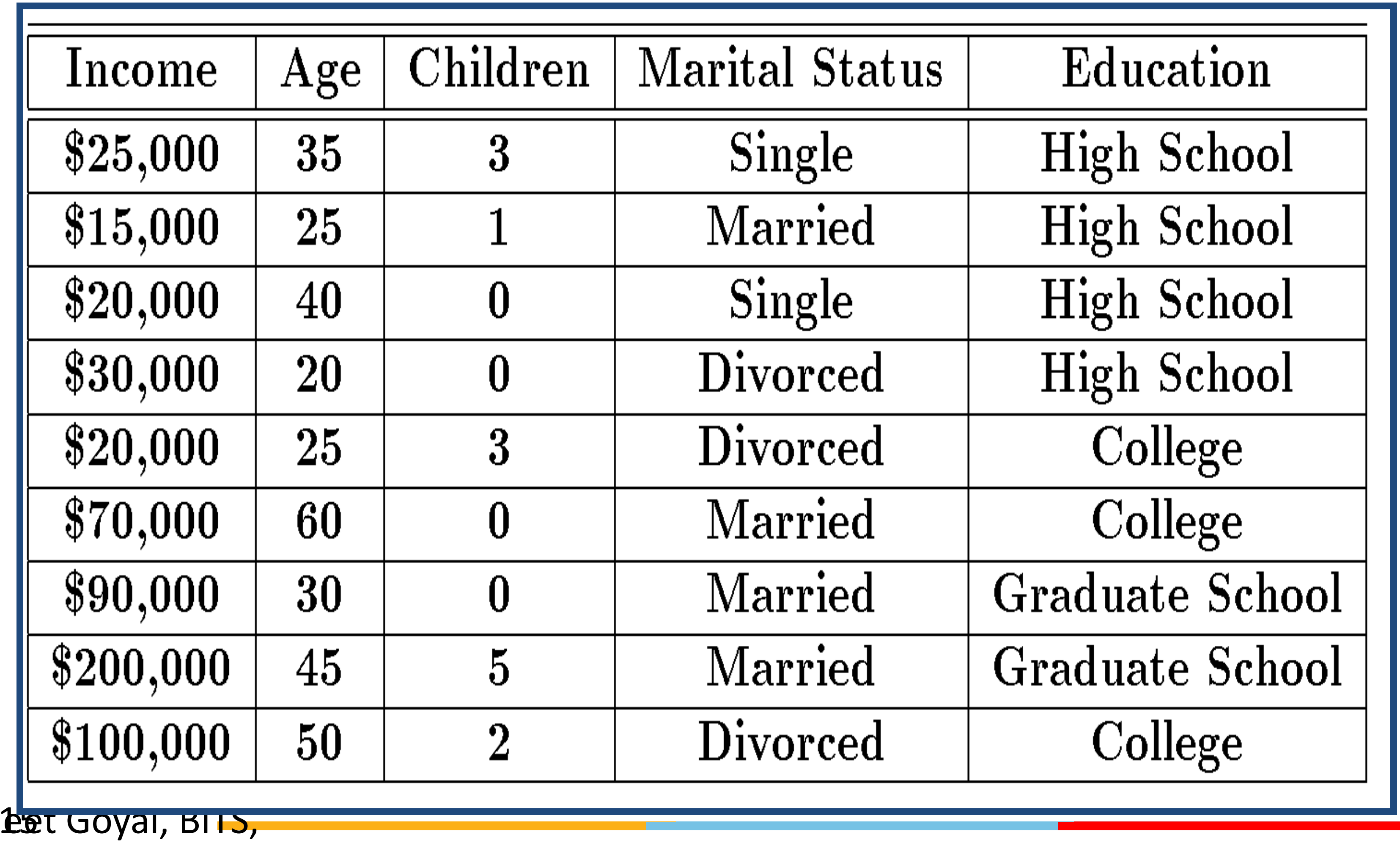
Clustering Applications

- *Segment* customer database based on similar buying patterns.
- Group houses in a town into neighborhoods based on similar features.
- Identify new plant species
- Identify similar Web usage patterns

Clustering Applications

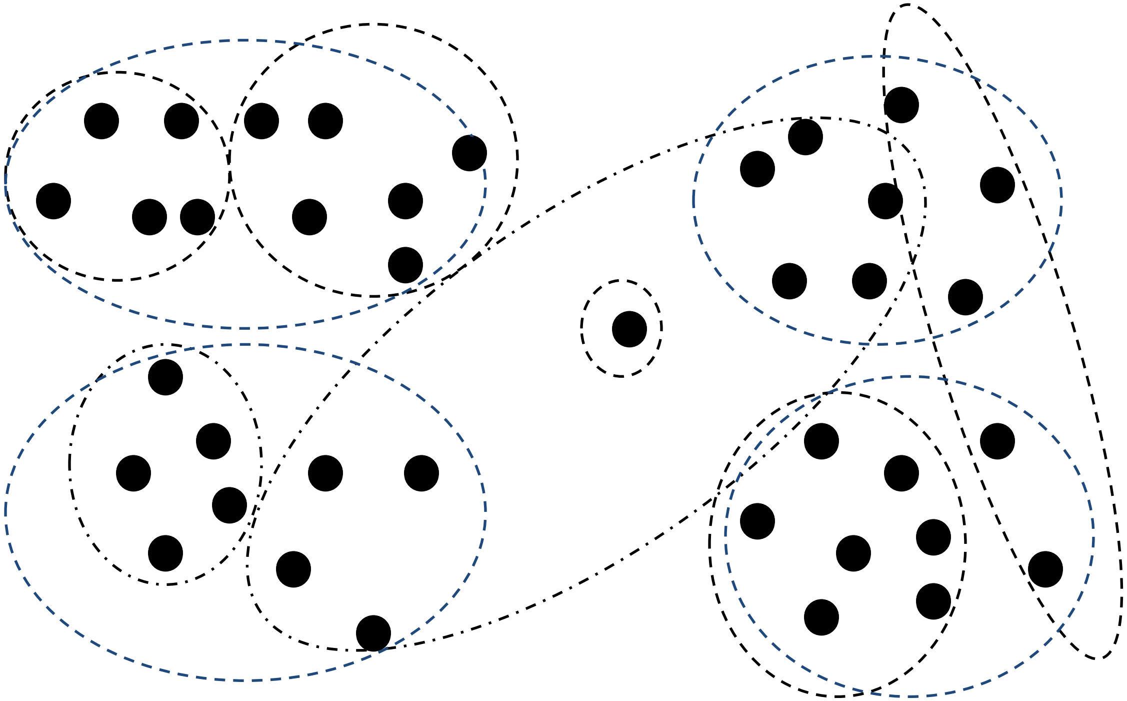
- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use:** Identification of areas of similar land use in an earth observation database
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults

Clustering Example



Income	Age	Children	Marital Status	Education
$25,000	35	3	Single	High School
$15,000	25	1	Married	High School
$20,000	40	0	Single	High School
$30,000	20	0	Divorced	High School
$20,000	25	3	Divorced	College
$70,000	60	0	Married	College
$90,000	30	0	Married	Graduate School
$200,000	45	5	Married	Graduate School
$100,000	50	2	Divorced	College

Clustering Houses



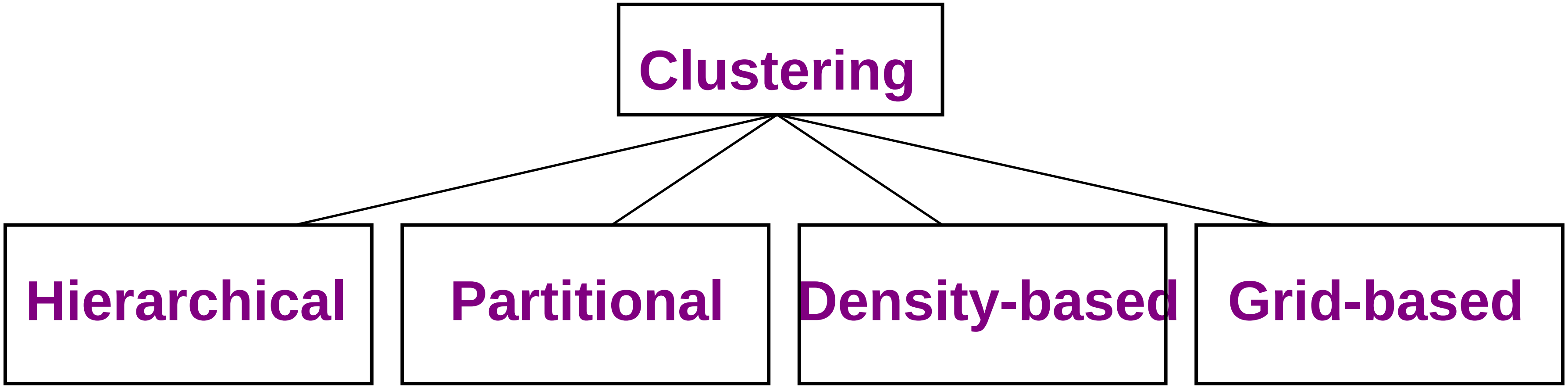
Geographic Distance Based

Size Based

Clustering vs. Classification

- No prior knowledge
  - Number of clusters
  - Meaning of clusters
- Unsupervised learning

Clustering Approaches



Clustering

Hierarchical

Partitional

Density-based

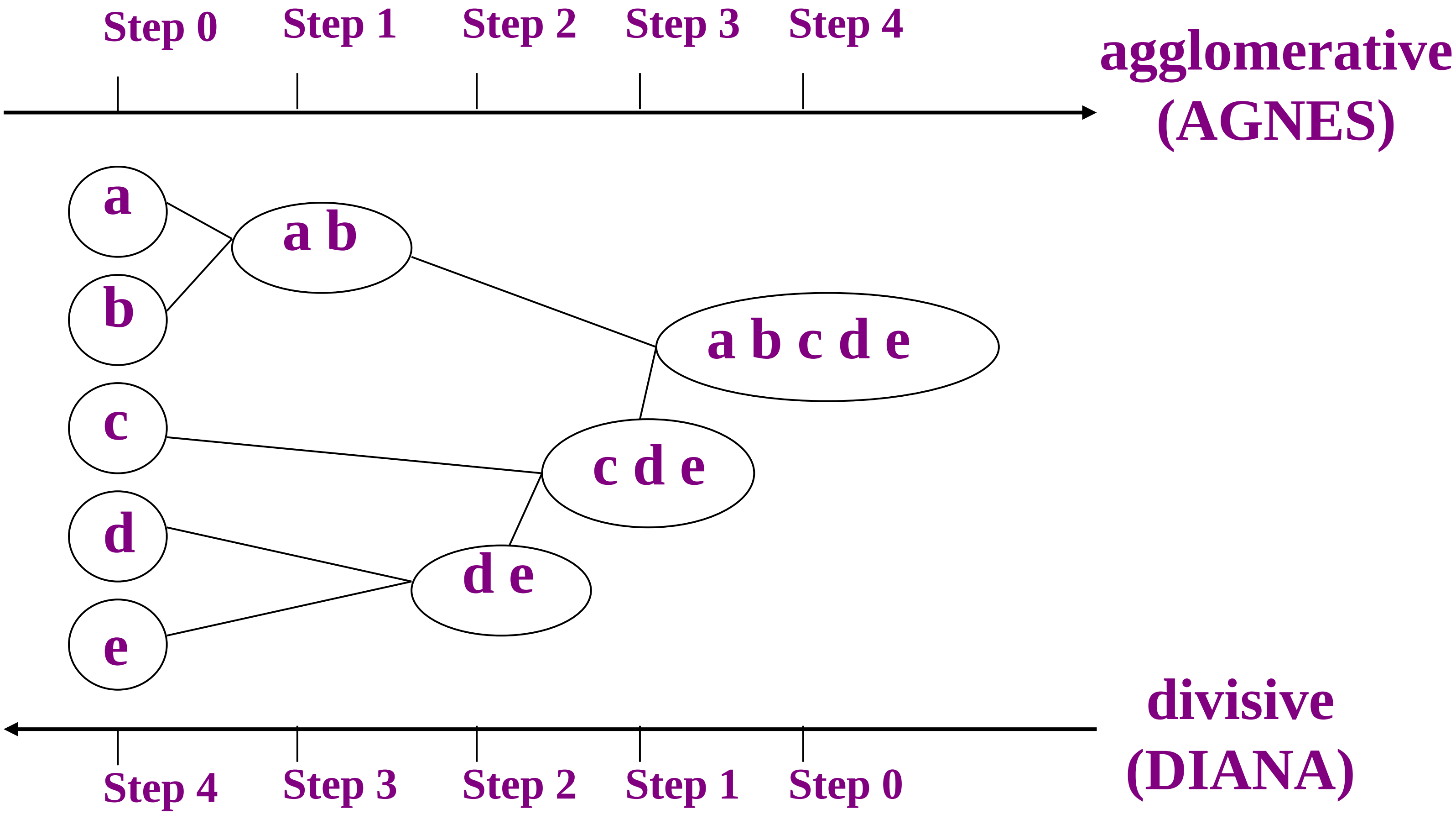
Grid-based

Hierarchical Methods

Creates a hierarchical decomposition of a given set of data objects

- Agglomerative
  - Initially each item in its own cluster
  - Clusters are merged iteratively
  - Bottom up
- Divisive
  - Initially all items in one cluster
  - Large clusters are divided successively
  - Top down

Hierarchical Clustering



Step 0 Step 1 Step 2 Step 3 Step 4

agglomerative (AGNES)

a

b

c

d

e

a b

d e

c d e

a b c d e

Step 4 Step 3 Step 2 Step 1 Step 0

divisive (DIANA)

Partitioning Methods

Given a DB of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k<=n$ such that

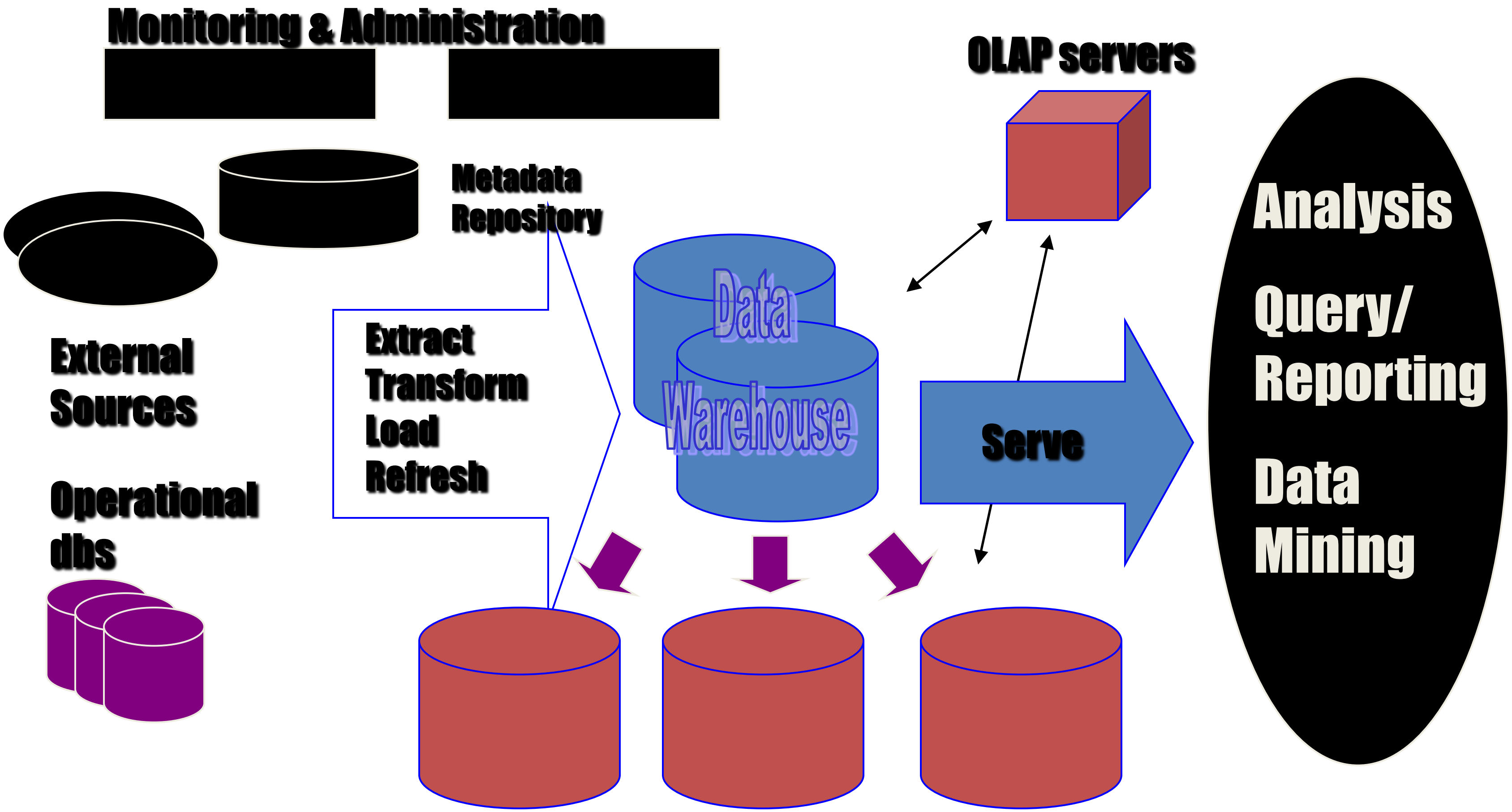
1. **Each group must contain atleast one object, and**
2. **Each object must belong to exactly one group**

Density-based Methods

- Most partitioning-based methods cluster objects based on distances between them
- Can find only spherical-shaped clusters
- Density-based clustering
- Continue growing a given cluster as long as the density in the 'neighborhood' exceeds some threshold.

Hierarchical Algorithms

- **Single Link**
- **MST Single Link**
- **Complete Link**
- **Average Link**



Monitoring & Administration

Metadata Repository

External Sources

Operational dbs

Extract Transform Load Refresh

Data Warehouse

Serve

OLAP servers

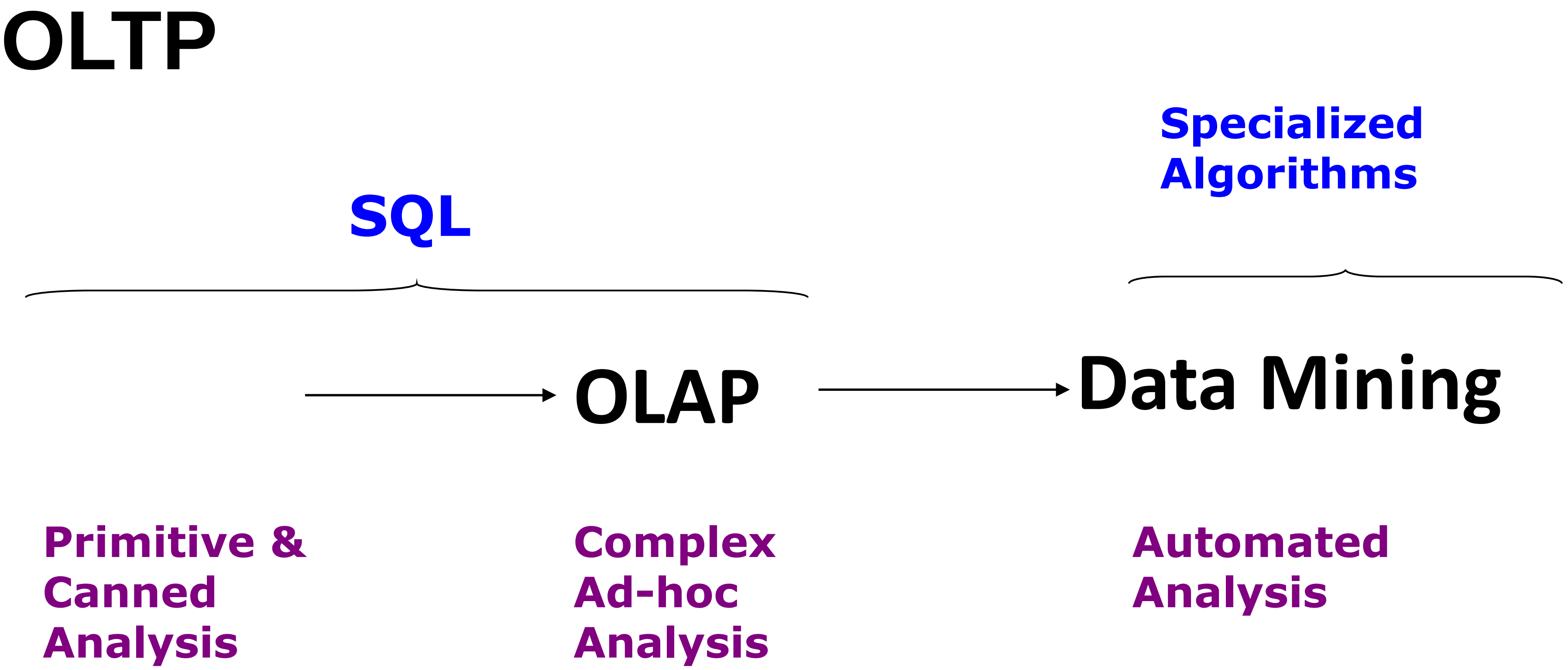
Analysis

Query/ Reporting

Data Mining

Data Marts

Continuum of Analysis



OLTP

SQL

Specialized Algorithms

OLAP

Data Mining

Primitive & Canned Analysis

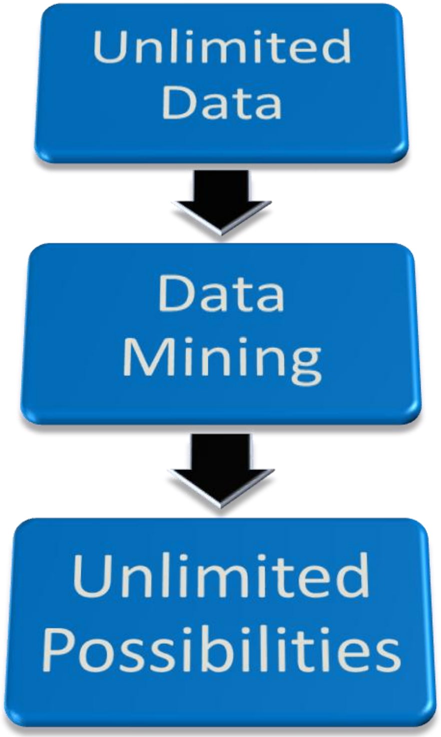
Complex Ad-hoc Analysis

Automated Analysis

Data Mining

My definition of Data Mining
"Data Mining is a family of techniques that transforms raw data into actionable information/knowledge"

Data Mining



Thank You