Introduction to Data Mining

Data Mining Outline

- ***Introduction***
- ***Related Concepts***
- ***Data Mining Techniques***

Outline

- Define data mining
- Data mining vs. databases
- Basic data mining tasks
- Data mining development
- Data mining issues

Why Data Mining

- Credit ratings/targeted marketing:
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
  - Identify likely responders to sales promotions
- Fraud detection
  - Which types of transactions are likely to be fraudulent, given the transactional history of a particular customer?
- Customer relationship management:
  - Which of my customers are likely to be the most loyal?

Data Mining helps extract such information

Introduction

- Data is growing at a phenomenal rate
- Users expect more sophisticated (refined) information
- How?

UNCOVER HIDDEN INFORMATION

DATA MINING

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, information harvesting, business intelligence, etc.

- Finding hidden information in a database
- Fit data to a model

Data Mining Algorithm

- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Preference – Technique to choose the best model
- Search – Technique to search the data
  - “Query”

Database Processing vs. Data Mining Processing

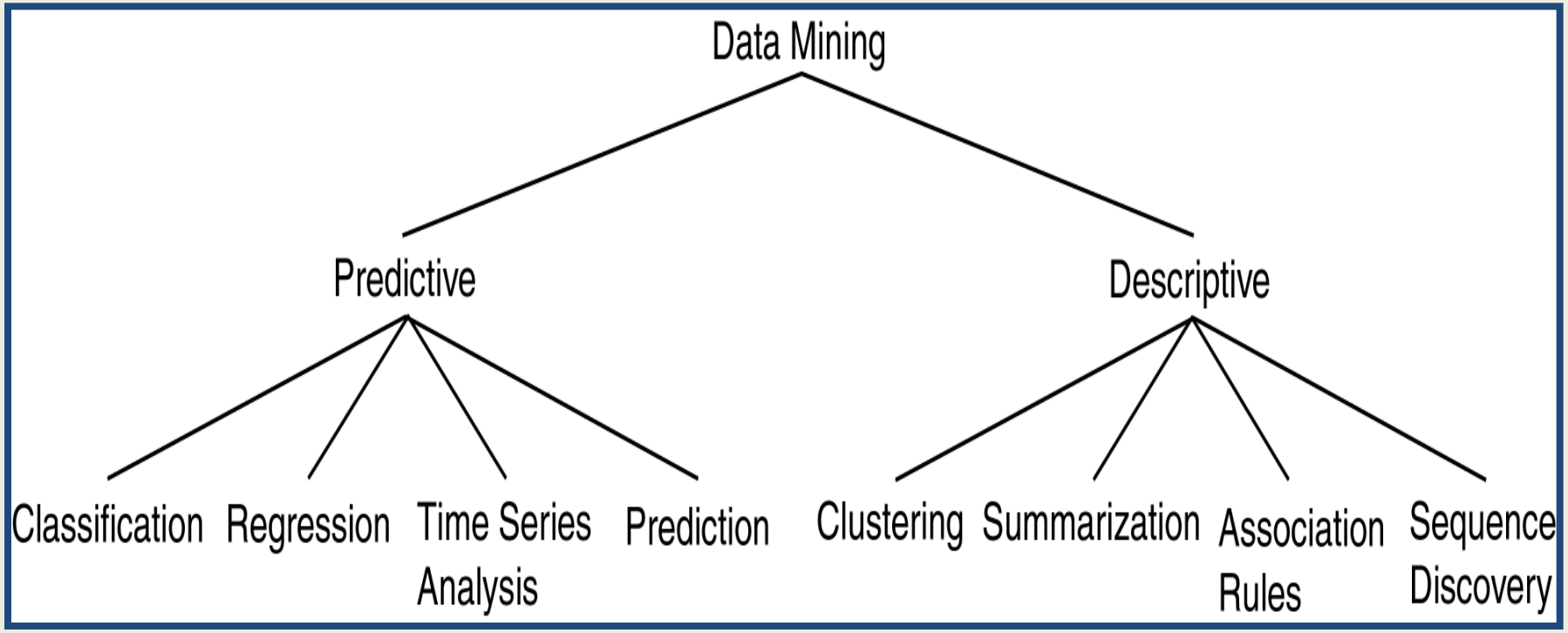
- Query
  - Well defined
  - SQL
- Data
  - Operational data
- Output
  - Precise

- Query
  - Poorly defined
  - No precise query language
- Data
  - Not operational data
- Output
  - Fuzzy

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Models and Tasks



Basic Data Mining Tasks

- ***Classification*** maps data into predefined groups or classes
  - Supervised learning
  - Pattern recognition
  - Prediction
- ***Regression*** is used a model to predict continuous value for a given input.
- ***Clustering*** groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning

Basic Data Mining Tasks (cont'd)

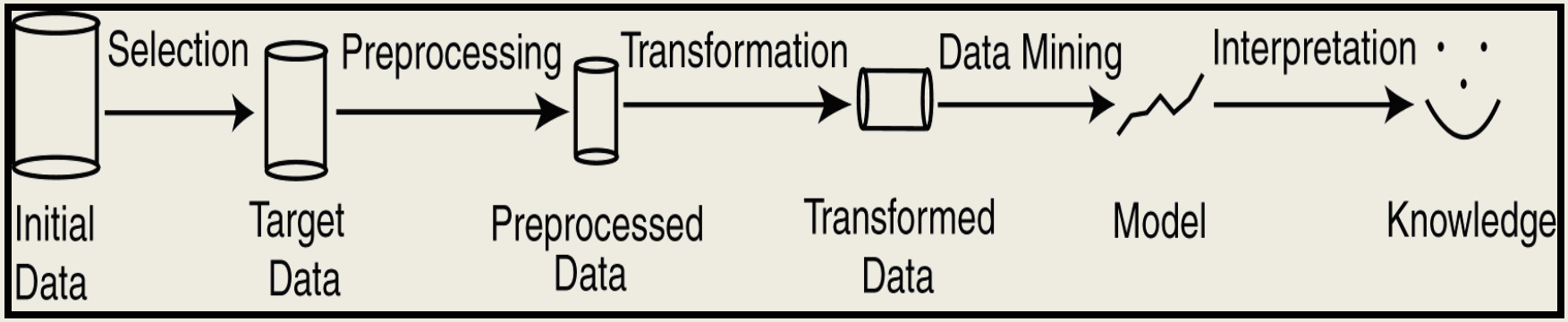
Link Analysis uncovers relationships among data.

- Affinity (similarity) Analysis
- Association Rules
- Sequential Analysis determines sequential patterns.

Data Mining vs. KDD

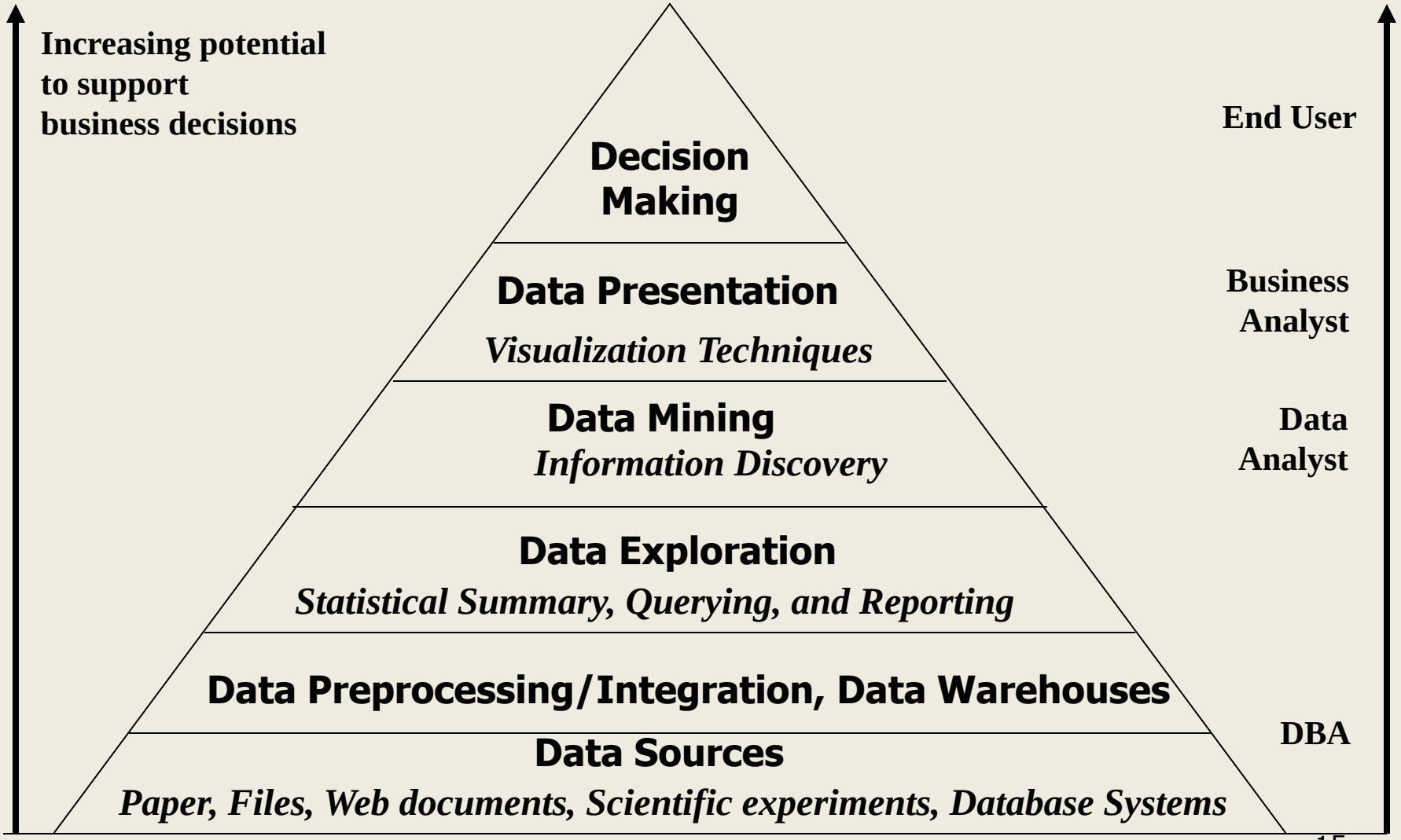
- ***Knowledge Discovery in Databases (KDD):*** process of finding useful information and patterns in data.
- ***Data Mining:*** Use of algorithms to extract the information and patterns derived by the KDD process.

KDD Process

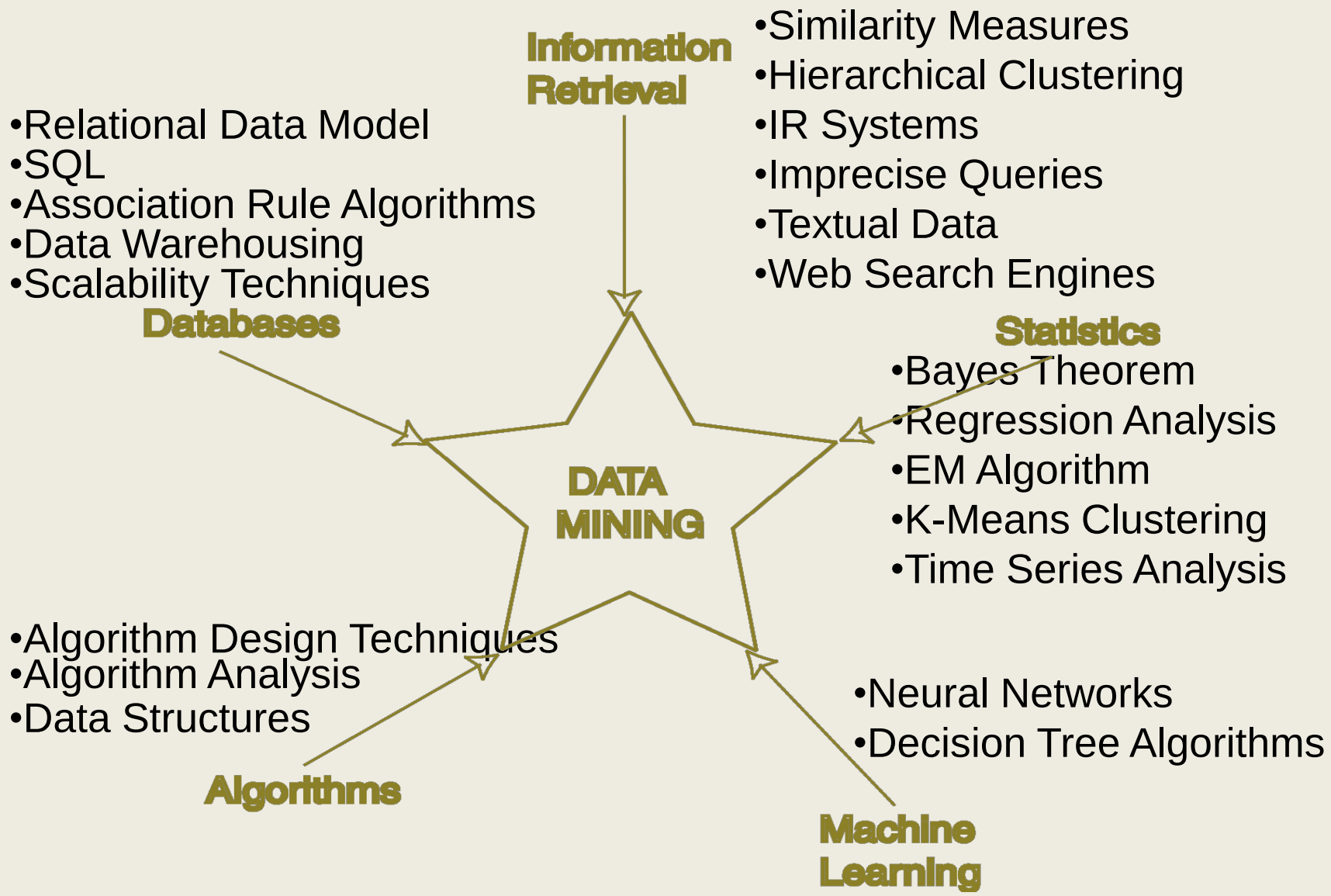


- ***Selection:*** Obtain data from various sources.
- ***Preprocessing:*** Cleanse data.
- ***Transformation:*** Convert to common format. Transform to new format.
- ***Data Mining:*** Obtain desired results.
- ***Interpretation/Evaluation:*** Present results to user in meaningful manner.

Data Mining and Business Intelligence



Data Mining Development: Multiple Disciplines



Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data
  - Multimedia database
  - Text databases
  - The World-Wide Web

KDD Issues

- **Human Interaction**
- **Overfitting**
- **Outliers**
- **Interpretation**
- **Visualization**
- **Large Datasets**
- **High Dimensionality**

KDD issues (cont'd)

- **Multimedia Data**
- **Missing Data**
- **Irrelevant Data**
- **Noisy Data**
- **Changing Data**
- **Integration**
- **Application**

Database Perspective on Data Mining

- Scalability
- Real World Data
- Updates
- Ease of Use

Goal: Examine some areas which are related to data mining.

Related Concepts Outline

- Database/OLTP Systems
- Fuzzy Sets and Logic
- Information Retrieval(Web Search Engines)
- Dimensional Modeling
- Data Warehousing
- OLAP/DSS
- Statistics
- Machine Learning
- Pattern Matching

DB & OLTP Systems

- Schema
  - (ID,Name,Address,Salary,JobNo)
- Data Model
  - ER
  - Relational
- Transaction
- Query:

```
SELECT Name
FROM T
WHERE Salary > 100000
```

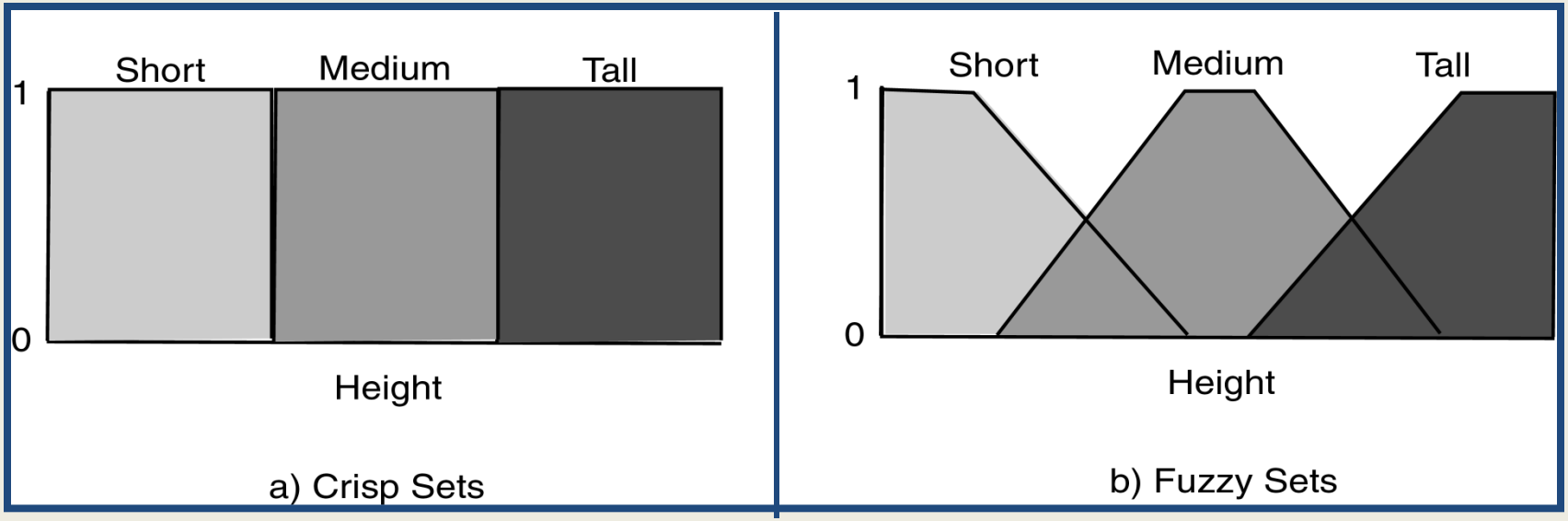
DM: Only imprecise queries

Fuzzy Sets and Logic

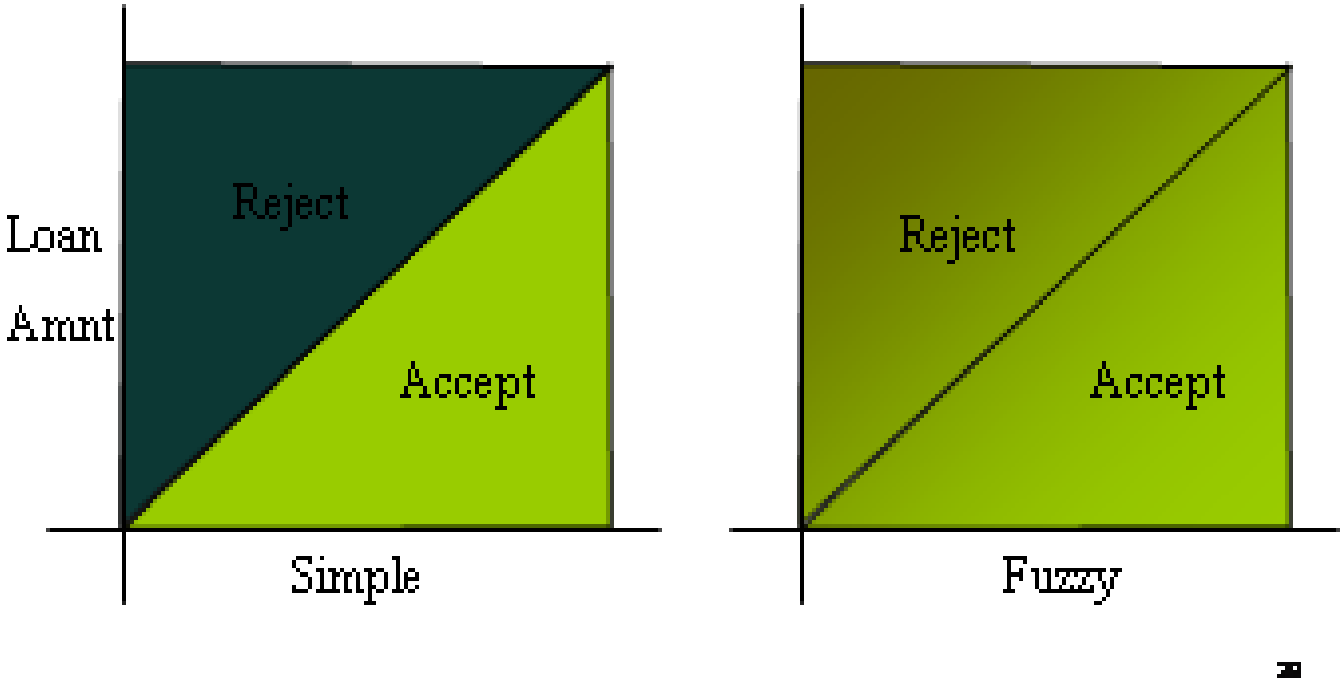
- ***Fuzzy Set:*** Set membership function is a real valued function with output in the range [0,1].
- f(x): Probability x is in F.
- 1-f(x): Probability x is not in F.
- EX:
  - T = {x | x is a person and x is tall}
  - Let f(x) be the probability that x is tall
  - Here f is the membership function

DM: ***Prediction and classification are fuzzy.***

Fuzzy Sets


Short
Medium
Tall
1
0
Height
a) Crisp Sets
Short
Medium
Tall
1
0
Height
b) Fuzzy Sets

Classification/Prediction is Fuzzy



Information Retrieval

- ***Information Retrieval (IR):*** retrieving desired information from textual data.

- Digital Libraries
- Web Search Engines
- Traditionally keyword based
- Sample query:

  Find all documents about “data mining”.

DM: Similarity measures;
Mine text/Web data.

Dimensional Modeling

- View data in a hierarchical manner more as business executives might
- Useful in decision support systems and mining
- ***Dimension:*** collection of logically related attributes; axis for modeling data.
- ***Facts:*** data stored
- Ex: Dimensions – products, locations, date

  Facts – quantity, unit price

DM: May view data as dimensional.

Relational View of Data

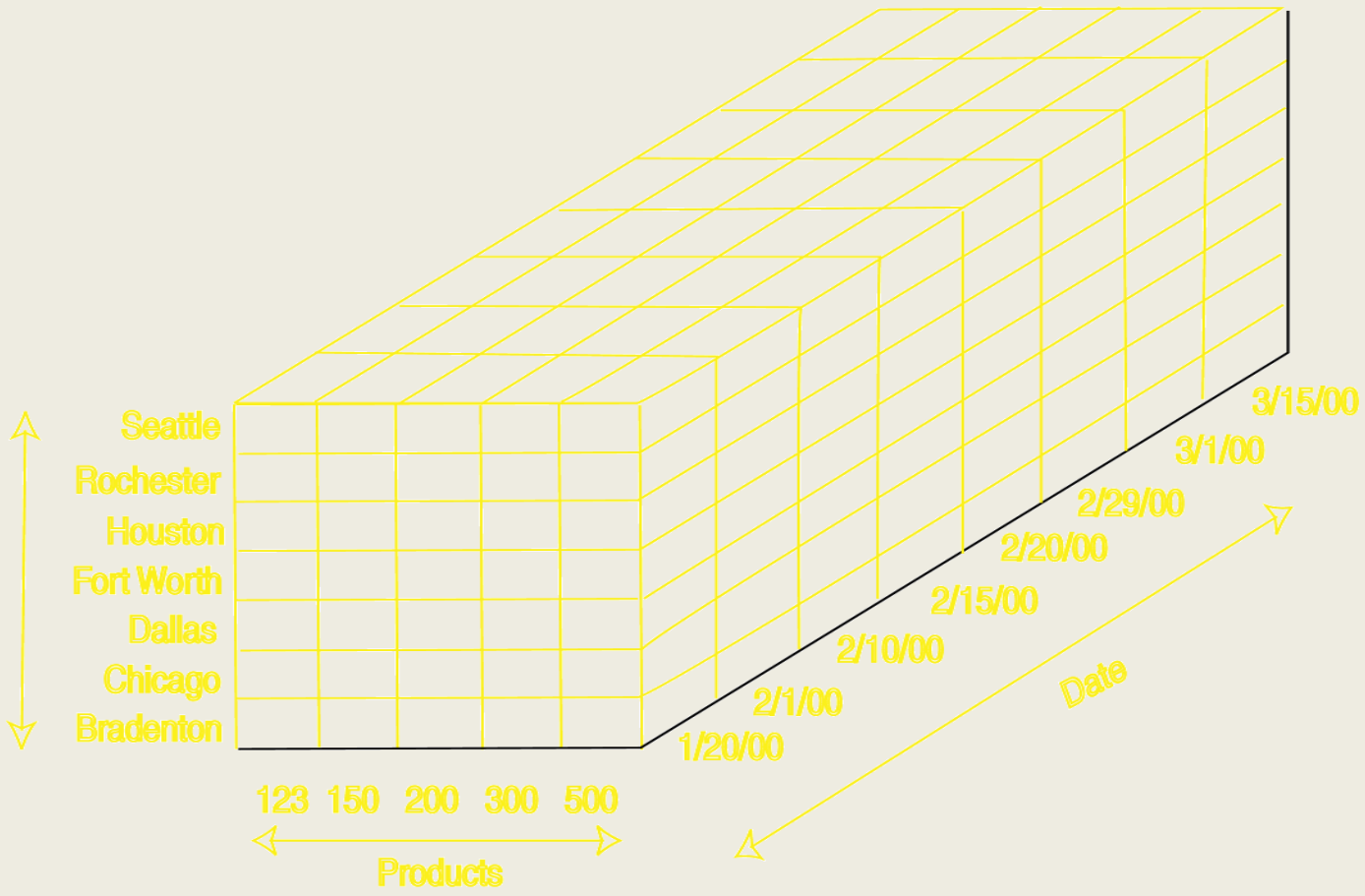
ProdID	LocID	Date	Quantity	UnitPrice
123	Dallas	022900	5	25
123	Houston	020100	10	20
150	Dallas	031500	1	100
150	Dallas	031500	5	95
150	Fort Worth	021000	5	80
150	Chicago	012000	20	75
200	Seattle	030100	5	50
300	Rochester	021500	200	5
500	Bradenton	022000	15	20
500	Chicago	012000	10	25

1

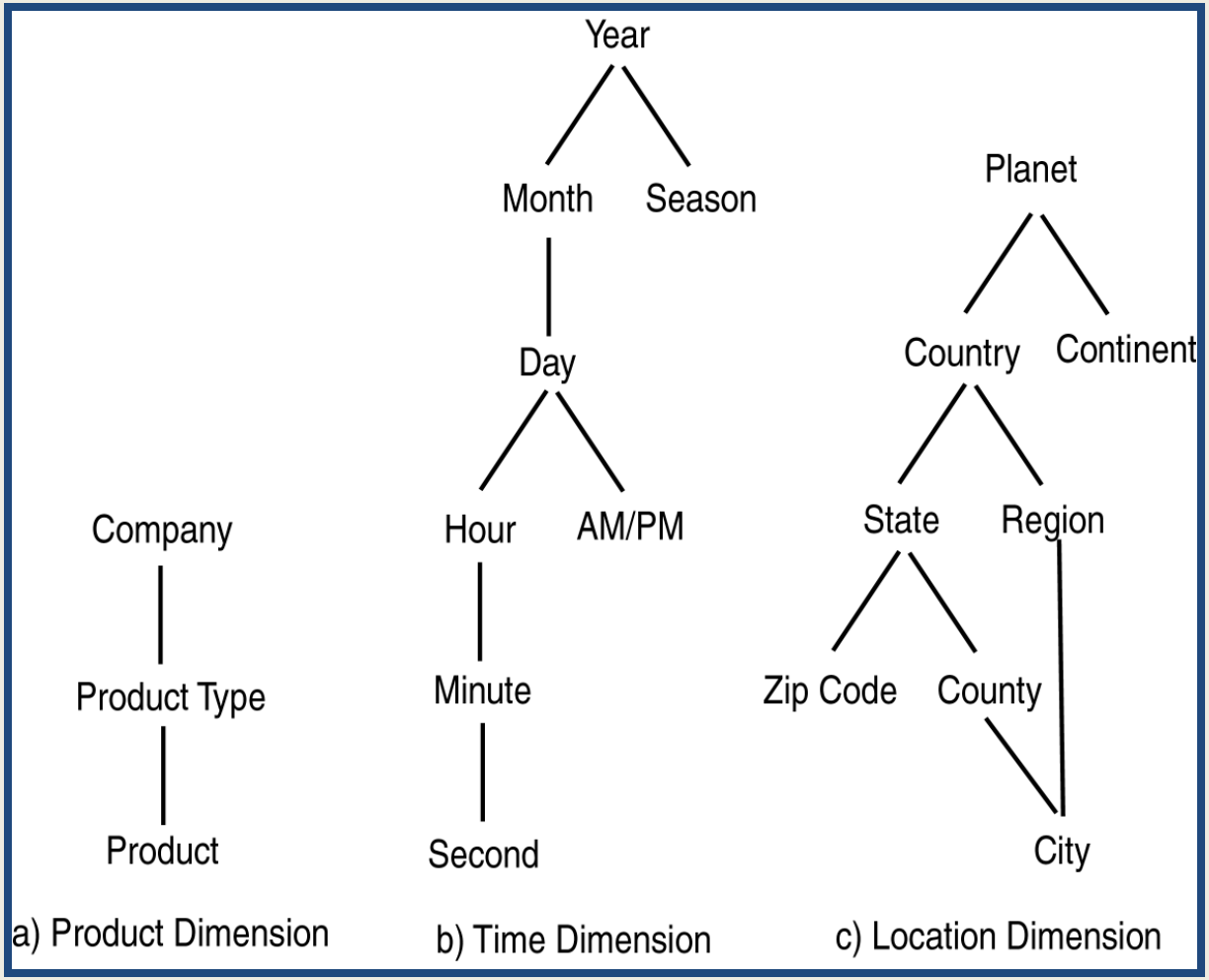
Dimensional Modeling Queries

- ***Roll Up:*** more general dimension
- ***Drill Down:*** more specific dimension
- Dimension (Aggregation) Hierarchy
- SQL uses aggregation
- ***Decision Support Systems (DSS):*** Computer systems and tools to assist managers in making decisions and solving problems.

Cube view of Data



Aggregation Hierarchies



Data Warehouse

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - 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*

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

Operational vs. Informational

	Operational Data	Data Warehouse
Application	OLTP	OLAP
Use	Precise Queries	Ad Hoc
Modification	Dynamic	Static
Orientation	Application	Business
Data	Operational Values	Integrated
Size	Gigabits	Terabits
Level	Detailed	Summarized
Access	Often	Less Often
Response	Few Seconds	Minutes

Statistics

- Simple descriptive models
- ***Statistical inference:*** generalizing a model created from a sample of the data to the entire dataset.
- Data mining targeted to business user

DM: Many data mining methods come from statistical techniques.

Machine Learning

- ***Machine Learning:*** area of AI that examines how to write programs that can learn.
- Often used in classification and prediction
- ***Supervised Learning:*** learns by example.
- ***Unsupervised Learning:*** learns without knowledge of correct answers.
- Machine learning often deals with small static datasets.

DM: Uses many machine learning techniques.

Pattern Matching (Recognition)

- ***Pattern Matching:*** finds occurrences of a predefined pattern in the data.
- Applications include speech recognition, information retrieval, time series analysis.

DM: Type of classification.

DM vs. Related Topics

Area	Query	Data	Results	Output
DB/OLTP	Precise	Database	Precise	DB Objects or Aggregation
IR	Precise	Documents	Vague	Documents
OLAP	Analysis	Multidimensional	Precise	DB Objects or Aggregation
DM	Vague	Preprocessed	Vague	KDD Objects

Data Mining Techniques Outline

Goal: Provide an overview of basic data mining techniques

- Statistical
  - Point Estimation
  - Bayes Theorem
  - Hypothesis Testing
  - Regression and Correlation
- Similarity Measures
- Decision Trees
- Neural Networks
  - Activation Functions
- Genetic Algorithms

Point Estimation

- ***Point Estimate:*** estimate a population parameter.
- May be made by calculating the parameter for a sample.
- May be used to predict value for missing data.
- Ex:
  - R contains 100 employees
  - 99 have salary information
  - Mean salary of these is $50,000
  - Use $50,000 as value of remaining employee's salary.

Is this a good idea?

Bayes Theorem

- ***Posterior Probability:*** P(H|X)
- ***Prior Probability:*** P(H)
- ***Bayes Theorem:***
- Assign probabilities of hypotheses given a data value.

Bayes Theorem : Basics

- Let **X** be a data sample : class label is unknown
- Let H be a *hypothesis* that X belongs to a specified class C
- For classification problems, we want to determine P(H|**X**), the probability that the hypothesis holds given the observed data sample **X**
- P(H) (*prior probability*), the initial probability
  - E.g., **X** will buy computer, regardless of age, income or any other information, for that matter.
- P(H|X) (*posteriori probability*), the probability of observing the sample **X**, given that the hypothesis holds
  - For example, suppose our world of data tuples is confined to customers described by the attributes *age* and *income*, respectively,
  - and that ***X*** is a 35-year-old customer with an income of $40,000.
  - Suppose that *H* is the hypothesis that our customer will buy a computer.
  - Then P(H|X) reflects the probability that customer ***X*** will buy a computer given that we know the customer's age and income.

Naïve Bayesian Classifier: Training Dataset

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data sample
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Bayesian Classifier: An Example

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

- Compute $P(X|C_i)$ for each class

P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019
$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

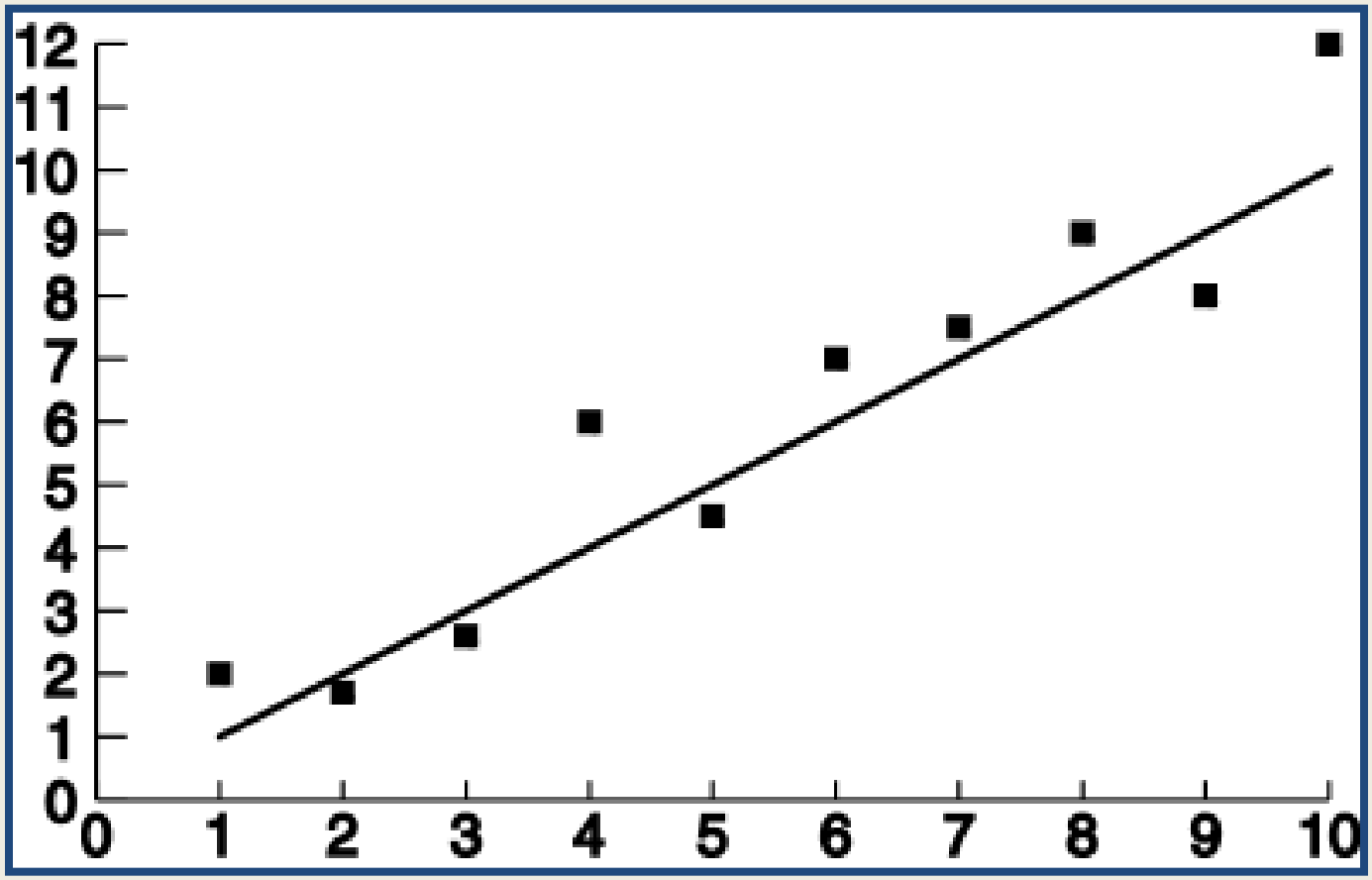
Regression

- Predict future values based on past values
- ***Linear Regression*** assumes linear relationship exists.

$$y = c_0 + c_1 x_1 + \ldots + c_n x_n$$

- Find values to best fit the data

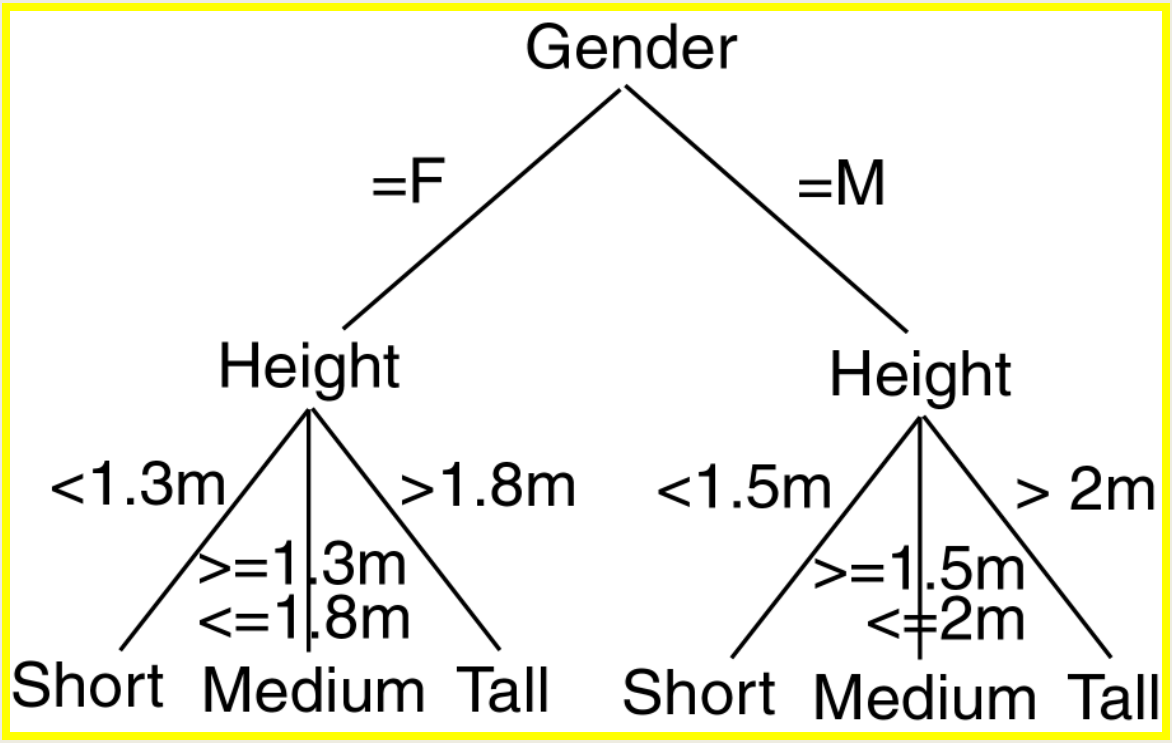
Linear Regression



Decision Trees

- ***Decision Tree (DT):***
  - Tree where the root and each internal node is labeled with a question.
  - The arcs represent each possible answer to the associated question.
  - Each leaf node represents a prediction of a solution to the problem.
- Popular technique for classification; Leaf node indicates class to which the corresponding tuple belongs.

Decision Tree Example



Decision Trees

- A ***Decision Tree Model*** is a computational model consisting of three parts:
  - Decision Tree
  - Algorithm to create the tree
  - Algorithm that applies the tree to data
- Creation of the tree is the most difficult part.
- Processing is basically a search similar to that in a binary search tree (although DT may not be binary).

DT Advantages/Disadvantages

- Advantages:
  - Easy to understand.
  - Easy to generate rules
- Disadvantages:
  - May suffer from overfitting.
  - Does not easily handle nonnumeric data.
  - Can be quite large – pruning is necessary.

Neural Networks

- Based on observed functioning of human brain.
- ***(Artificial Neural Networks (ANN)***
- Used in pattern recognition, speech recognition, computer vision, and classification.

Neural Networks

- **Introduction to Artificial Neural Network-**
- ANN in general is a highly interconnected network of a large no. of processing called 'Neurons' in an architecture inspired by the brain.
- The brain is a highly complex, nonlinear and parallel computer (Information Processing System).

- **ANN Vs Brain:-**

In brain a neuron has three principal components:

1. **Dendrites:-** that carry electrical signals into the cell body.
2. **Cell Body:-** effectively sums and thresholds these incoming signals.
3. **Axon:-** is a single long fiber that carries the signal from the cell body out to other neurons.
4. The point of contact between an axon of one cell and a dendrite of another cell is called a '**synapse'**

ANN	Brain
It is simple (few neuron in connection)	It is complex (10^{11} Neurons and 10^{15} connections)
It is dedicated for specific purpose	It is generalized for all purpose
Response time is fast (it may be in Nanosecond)	Response time is slow (it may be in millisecond)
Design is regular	Design is arbitrary
Activities are synchronous	Activities are asynchronous

Classification by Backpropagation

- Backpropagation: A **neural network** learning algorithm
- Started by psychologists and neurobiologists to develop and test computational analogues of neurons.
- A neural network: A set of connected input/output units where each connection has a **weight** associated with it.
- During the learning phase, the **network learns by adjusting the weights** so as to be able to predict the correct class label of the input tuples.
- Also referred to as **connectionist learning** due to the connections between units

Neural Network as a Classifier

- Weakness
  - Long training time
  - Require a number of parameters typically best determined empirically, e.g., the network topology or ``structure."
  - Poor interpretability: Difficult to interpret the symbolic meaning behind the learned weights and of ``hidden units" in the network

- Strength
  - High tolerance to noisy data as well as their ability to classify patterns on which they have not been trained.
  - They are well-suited for continuous-valued inputs *and outputs*, unlike most decision tree algorithms.
  - They have been successful on a wide array of real-world data, including handwritten character recognition, pathology and laboratory medicine, and training a computer to pronounce English text.
  - Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process.

These above factors contribute toward the usefulness of neural networks for classification and prediction in data mining.

- **Applications of ANN:-**

1. Speech Recognition
2. Pattern Recognition
3. Signature Recognition
4. Financial accounting
5. Digital image processing
6. Weather forecasting

NN Advantages

- Learning
- Can continue learning even after training set has been applied.
- Easy parallelization
- Solves many problems

NN Disadvantages

- Difficult to understand
- May suffer from overfitting
- Structure of graph must be determined a priori.
- Input values must be numeric.
- Verification difficult.

Genetic Algorithms (GA)

- Genetic Algorithm: based on an analogy to biological evolution.
- In general, genetic learning starts as follows. An initial **population** is created consisting of randomly generated rules
- Based on the notion of survival of the **fittest**, a new population is formed to consist of the fittest rules and their offsprings.
- The fitness of a rule is represented by its *classification accuracy* on a set of training examples Offsprings are generated by *crossover* and *mutation.*
- The process continues until a population P evolves *when each rule in P satisfies a prespecified* fitness *threshold .*
- Slow but easily parallelizable and have been used for classification as well as other optimization problems.

GA Advantages/Disadvantages

- Advantages
  - Easily parallelized
- Disadvantages
  - Difficult to understand and explain to end users.
  - Abstraction of the problem and method to represent individuals is quite difficult.
  - Determining fitness function is difficult.
  - Determining how to perform crossover and mutation is difficult.

Some success stories:

- **Data mining applications**
- **Text Mining**
- **Video Mining ---------- Multimedia Mining**

- **Privacy Preserving in Association Rule Mining**

- **Intrusion Detection- Database Intrusion Detection**
  - Network Intrusion Detection