Data Science - Lecture 3

Introduction To Data Science

Dr. Faisal Kamiran

Award winning Data Scientist and Professor

What is today’s agenda?

Today we are going to learn following things :

- Introduction to Data Mining
- Basics of
  - Classification
  - Clustering
  - Association Rule Mining
  - Sequential Pattern Mining

What is (not) Data Mining

What is not Data Mining?

- Look up phone number in phone directory

What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)

Origins of Data Mining

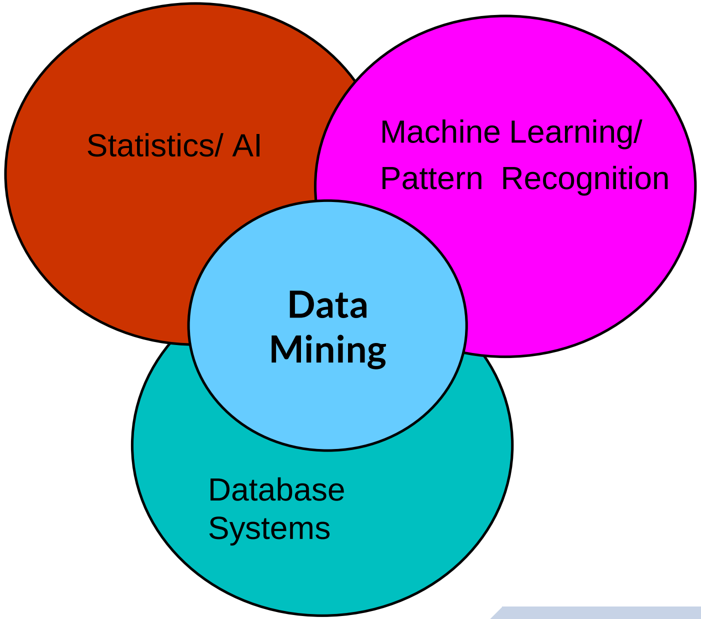
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems.

- Traditional Techniques may be unsuitable due to

  - Enormity of data

  - High dimensionality of data

  - Heterogeneous, distributed nature of data

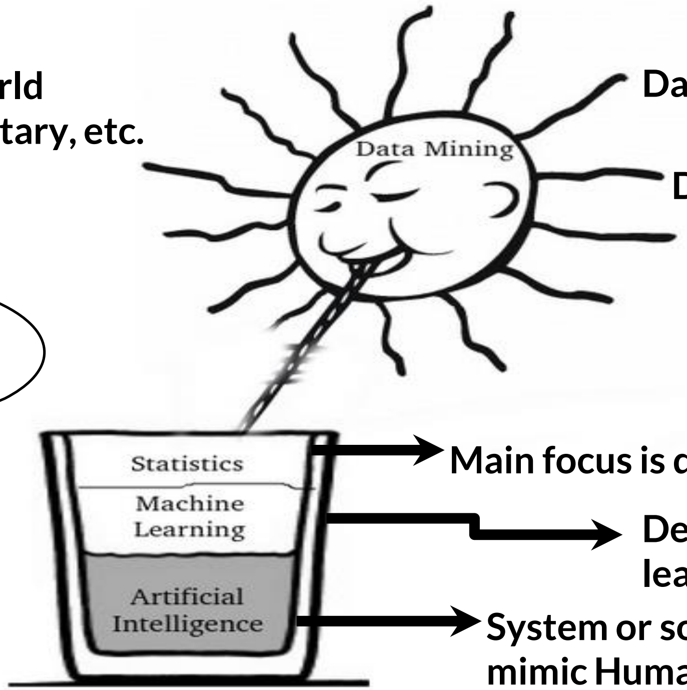


Origins of Data Mining

Solve real world business, Military, etc. problems

Descriptive & Inferential

- NLP
- Computer Vision
- Robotics
- Reasoning



Data preprocessing

Data analysis

Practical application and Development of algorithms

Main focus is data

Development of self learning algorithms

System or software to mimic Human

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

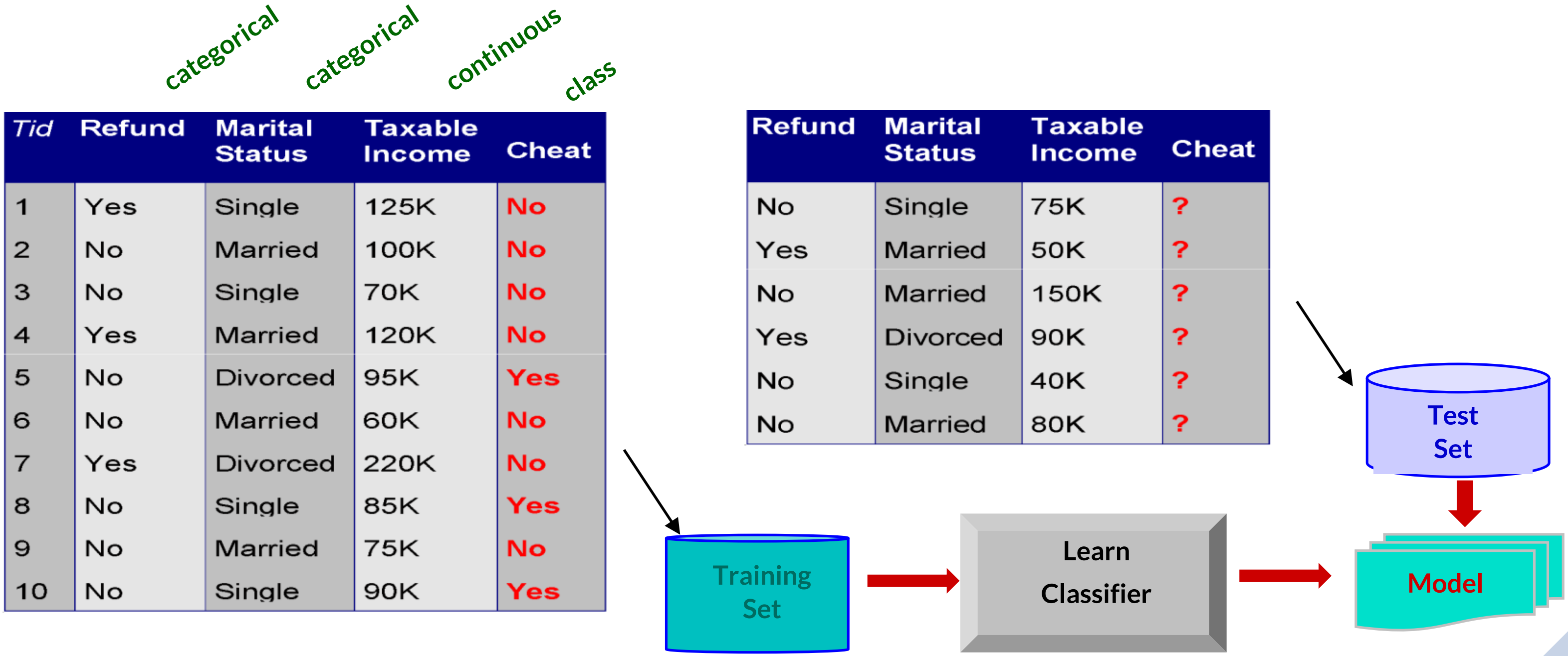
Classification : Definition

- Given a collection of records (*training set*)
  - 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: <u>previously unseen</u> 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.

Classification : Example



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	?

Classification : Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification : Application 2

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

Classification : Application 3

- Customer Attrition/Churn:
    - Goal: To predict whether a customer is likely to be lost to a competitor.
    - Approach:
        - Use detailed record of transactions with each of the past and present customers, to find attributes.
            - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
        - Label the customers as loyal or disloyal.
        - Find a model for loyalty.

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:

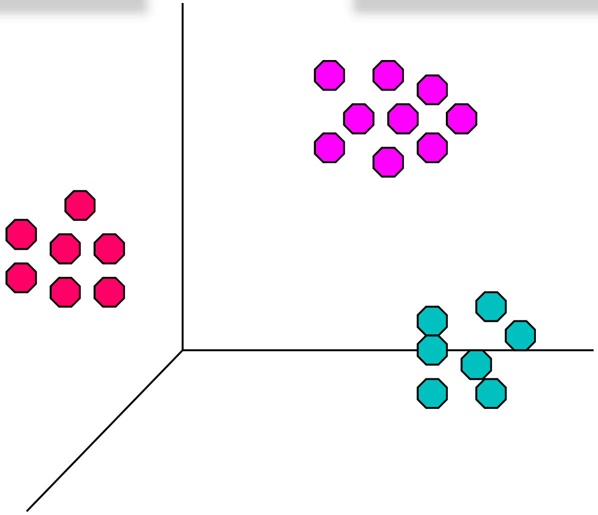
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

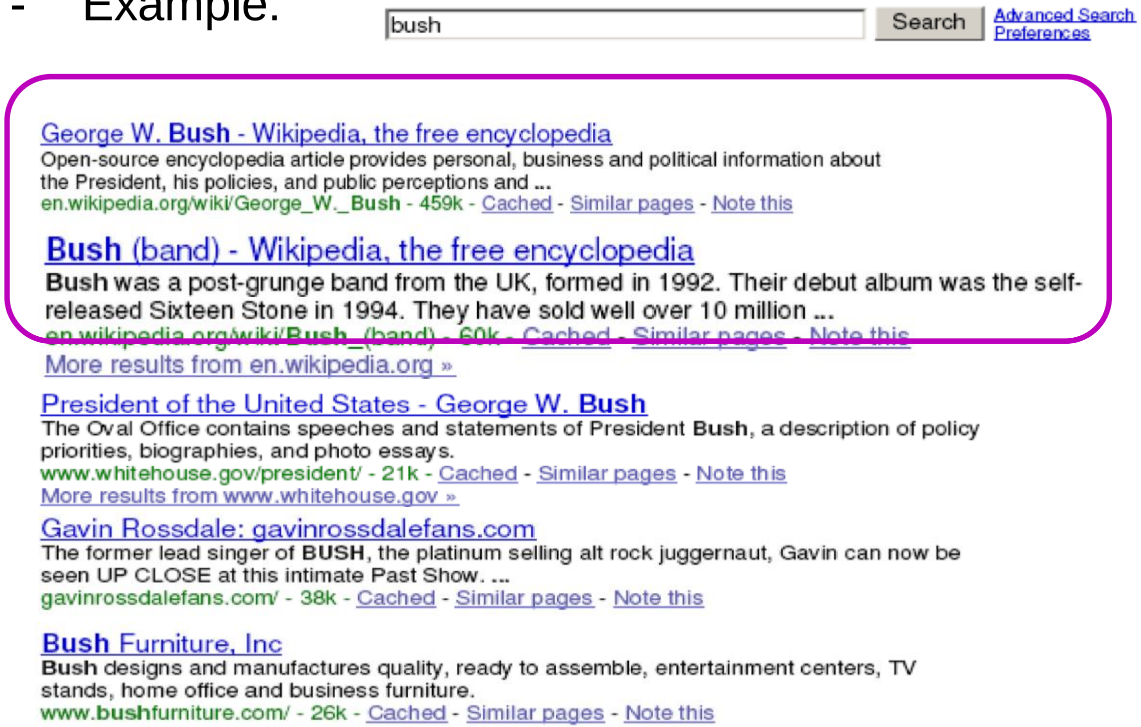
Intracluster distances are minimized

Intercluster distances are maximized



Data Mining Techniques : Clustering

- Example:



bush

George W. **Bush** - Wikipedia, the free encyclopedia
Open-source encyclopedia article provides personal, business and political information about the President, his policies, and public perceptions and ...
en.wikipedia.org/wiki/George_W._**Bush** - 459k - Cached - Similar pages - Note this

Bush (band) - Wikipedia, the free encyclopedia
Bush was a post-grunge band from the UK, formed in 1992. Their debut album was the self-released Sixteen Stone in 1994. They have sold well over 10 million ...
en.wikipedia.org/wiki/**Bush**_(band) - 60k - Cached - Similar pages - Note this
More results from en.wikipedia.org »

President of the United States - George W. **Bush**
The Oval Office contains speeches and statements of President **Bush**, a description of policy priorities, biographies, and photo essays.
www.whitehouse.gov/president/ - 21k - Cached - Similar pages - Note this
More results from www.whitehouse.gov »

Gavin Rossdale: gavinrossdalefans.com
The former lead singer of **BUSH**, the platinum selling alt rock juggernaut, Gavin can now be seen UP CLOSE at this intimate Past Show. ...
gavinrossdalefans.com/ - 38k - Cached - Similar pages - Note this

Bush Furniture, Inc
Bush designs and manufactures quality, ready to assemble, entertainment centers, TV stands, home office and business furniture.
www.**bush**furniture.com/ - 26k - Cached - Similar pages - Note this

Clustering : Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

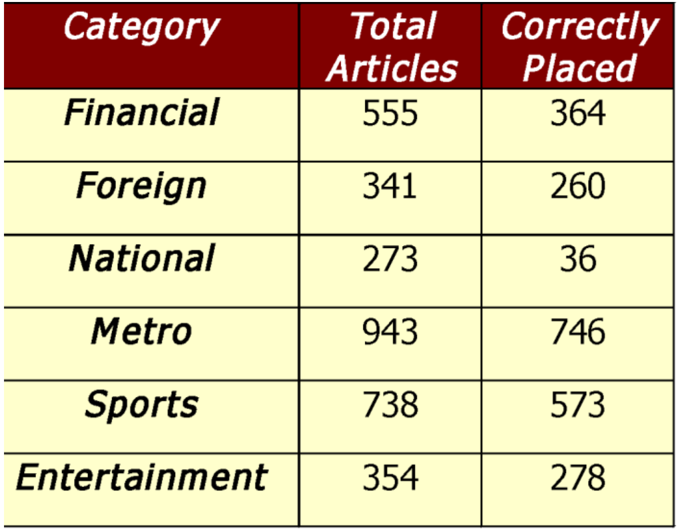
Clustering : Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.

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



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

Classification vs Clustering

Classification

- **Input:** We have a Training set containing data that have been previously categorized

- **Task:** Based on this training set, the algorithms finds the category that the new data points belong to

- Since a Training set exists, we describe this technique as **Supervised learning**

Clustering

- **Input:** We do not know the characteristics of similarity of data in advance

- **Task:** Using statistical concepts, we split the datasets into sub-datasets such that the Sub-datasets have “Similar” data

- Since Training set is not used, we describe this technique as **Unsupervised learning**

Supervised vs Unsupervised Learning

Supervised Learning

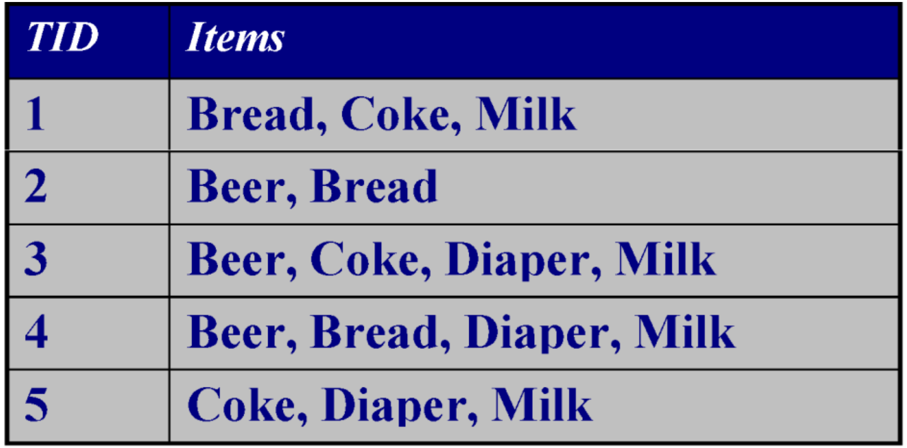
- Correct results/labels during the training are given.
- Resultant models are generalized ones, usually fast and accurate

Unsupervised Learning

- Correct results/labels are **NOT** given in input data
- Usually computationally expensive
- Grouping of input data w.r.t. its statistical properties

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}

Association Rule Discovery : Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Coke, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Coke in the antecedent => Can be used to see which products would be affected if the store discontinues selling coke.
  - Coke in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Coke to promote sale of Potato chips!

Association Rule Discovery : Application 2

- Supermarket shelf management.
  - Goal: To identify items that are purchased together by many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery : Application 3

- Inventory Management:
    - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
    - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery : 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.

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

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

Sequential Pattern Discovery : Example

- In point-of-sale transaction sequences,

  - Computer Bookstore:
    - (Intro_To_Visual_C) (C++_Primer) --> (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    - (Shoes) (Racket, Racketball) --> (Sports_Jacket)

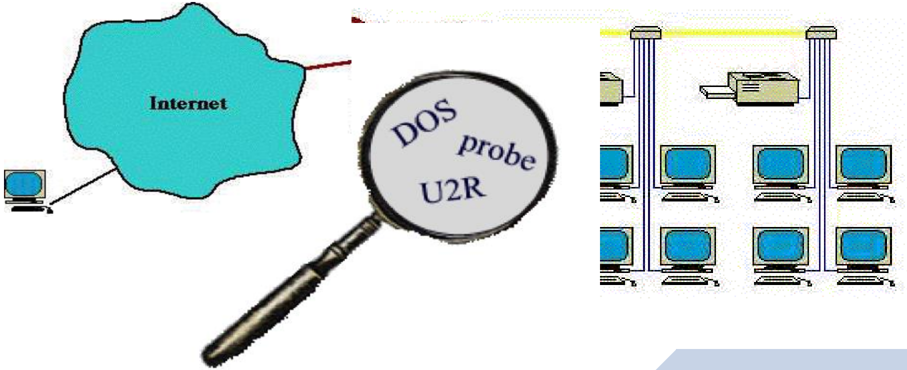
Regression

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

- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Time series prediction of stock market indices.
  - Income prediction on basis of qualifications and other characteristics of individuals

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

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Open Source Data Mining Tools

- Python
- R
- Weka
- Knime
- Rapidminer
- Matlab
- Tableau

Contribution of Data Mining

- Less expenditures
  - Automated systems instead of manual ones
  - Selection of customers to mail new promotions of the company

- Effective decision making
  - Careful expansion of the business
  - Product selection
  - Pricing

Contribution of Data Mining

- Increased sales
  - Shelf management to increase the sale of certain items
  - What types of products can be sold together?
  - How does one retain profitable customers?

Data Mining Real World Success Stories

- Bank of America identified savings of $4.8 million in 2 years by using a credit risk management system, i.e., examination of only borderline applicants.

- BBC’s data mining based program scheduler determines the timing to show programs as good as the best planner but at much less cost.

Data Mining Real World Success Stories

- Bell Atlantic developed telephone technician dispatch system. They must decide what type of technician to dispatch to resolve the reported complain.

- Bell Atlantic save more than 10 million dollars per year by using data mining rule based system because they make fewer erroneous decisions.

Data Mining Real World Success Stories

- Safeway (UK)'s data mining system found that the top - spending 25% customers often purchase a particular cheese product ranked below 200 in sales.

- Normally, without the data - mining results, the product would have been discontinued and would disappoint the best customers.

- Safeway continues to order this cheese, although it is ranked low in sales.

“Questions ?