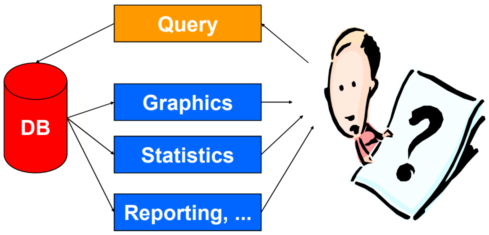
Data Mining
Basics

Instructor: Sharma Chakravarthy
sharmac@cse.uta.edu
The University of Texas at Arlington

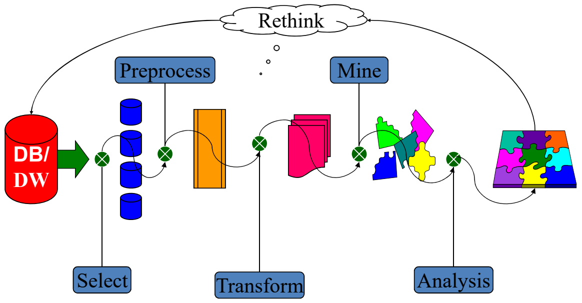
Contrasting with Traditional Data Analysis

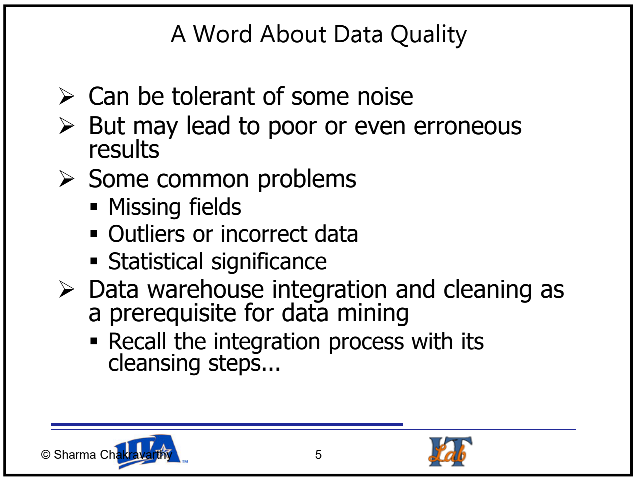


Data Mining Process

- Collect, Assess, and transform (DW)
- Select: reduces cost, increases speed
- Explore: summarize, Segment, visualize
- Modify: data filtering, variable selection
- Model: regression, neural nets, decision trees, associations, sequences
- Interpret results (BI or business intelligence)

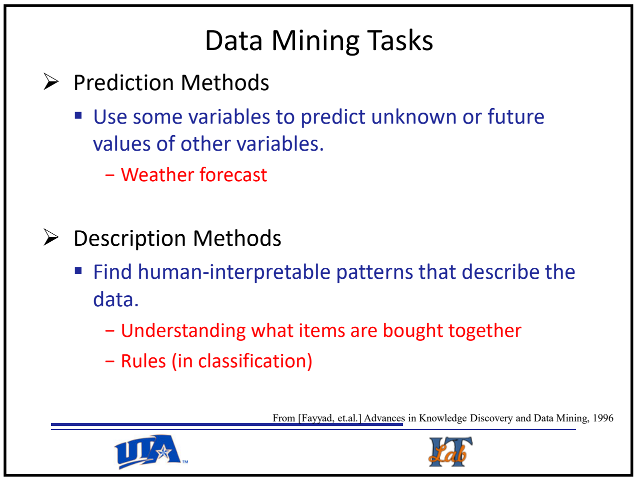
Data Mining Cycle





A Word About Data Quality

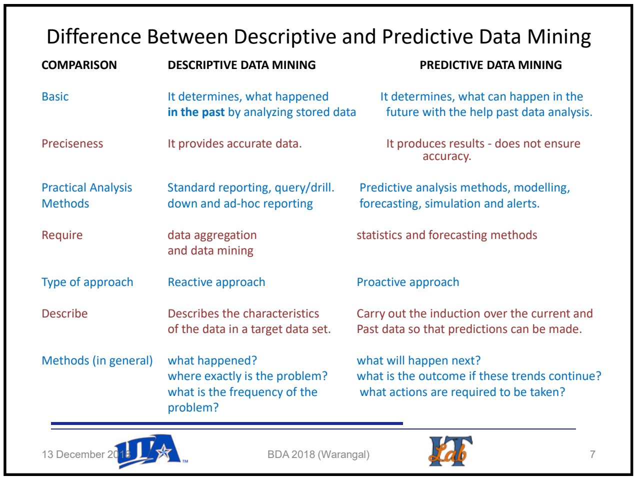
- Can be tolerant of some noise
- But may lead to poor or even erroneous results
- Some common problems
  - Missing fields
  - Outliers or incorrect data
  - Statistical significance
- Data warehouse integration and cleaning as a prerequisite for data mining
  - Recall the integration process with its cleansing steps...



Data Mining Tasks

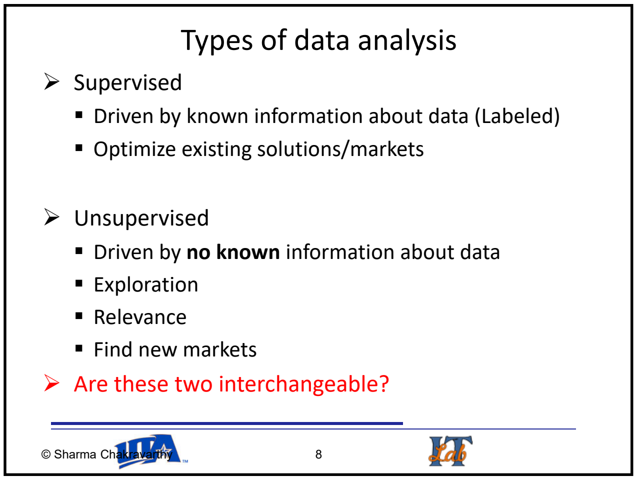
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
    - Weather forecast
- Description Methods
  - Find human-interpretable patterns that describe the data.
    - Understanding what items are bought together
    - Rules (in classification)

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



Difference Between Descriptive and Predictive Data Mining

COMPARISON	DESCRIPTIVE DATA MINING	PREDICTIVE DATA MINING
Basic	It determines, what happened **in the past** by analyzing stored data	It determines, what can happen in the future with the help past data analysis.
Preciseness	It provides accurate data.	It produces results - does not ensure accuracy.
Practical Analysis Methods	Standard reporting, query/drill. down and ad-hoc reporting	Predictive analysis methods, modelling, forecasting, simulation and alerts.
Require	data aggregation and data mining	statistics and forecasting methods
Type of approach	Reactive approach	Proactive approach
Describe	Describes the characteristics of the data in a target data set.	Carry out the induction over the current and Past data so that predictions can be made.
Methods (in general)	what happened? where exactly is the problem? what is the frequency of the problem?	what will happen next? what is the outcome if these trends continue? what actions are required to be taken?



Types of data analysis

- Supervised
  - Driven by known information about data (Labeled)
  - Optimize existing solutions/markets
- Unsupervised
  - Driven by **no known** information about data
  - Exploration
  - Relevance
  - Find new markets
- Are these two interchangeable?

DM Approaches

- Classification [predictive]
- Clustering [descriptive]
- Association rules [descriptive]
- Text classification [descriptive]
- Anomaly detection [predictive, descriptive]
- Graph Mining [predictive, descriptive]
- …

Classification: Definition

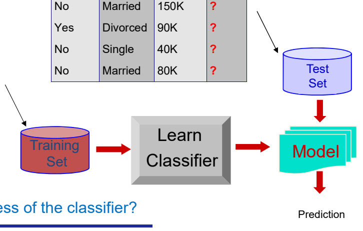
- Input: A collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class or label* (labeled data set)
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: <u>previously unseen</u> records should be assigned 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.
  - Termed cross-validation!

Classification Example

categorical categorical continuous class

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

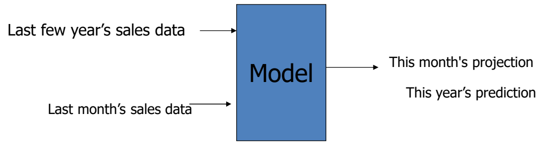
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



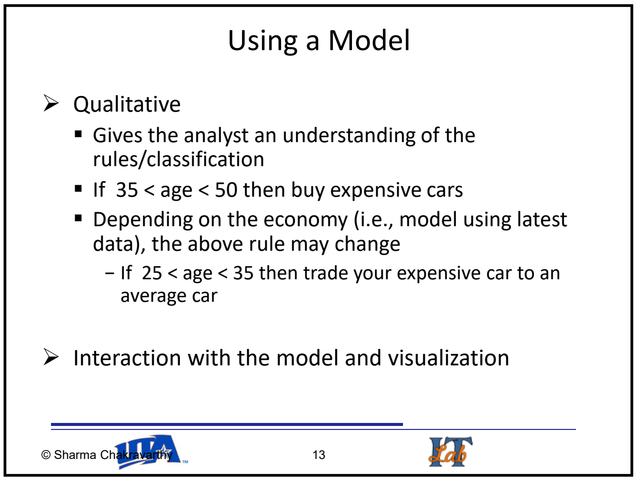
How do we verify the correctness of the classifier?

Predictive Modeling

- A “black box” that makes predictions about the future data based on information from past and present

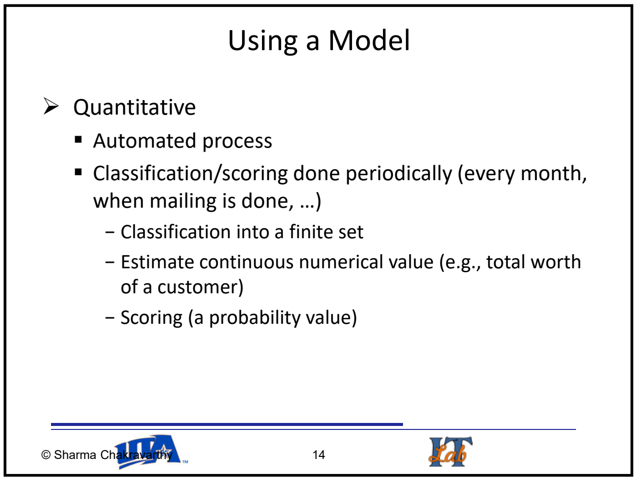


Usually Large number of inputs available



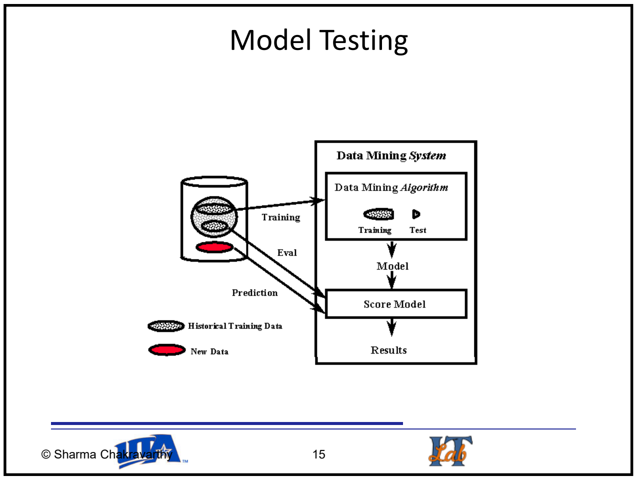
Using a Model

- Qualitative
  - Gives the analyst an understanding of the rules/classification
  - If 35 < age < 50 then buy expensive cars
  - Depending on the economy (i.e., model using latest data), the above rule may change
    - If 25 < age < 35 then trade your expensive car to an average car
- Interaction with the model and visualization

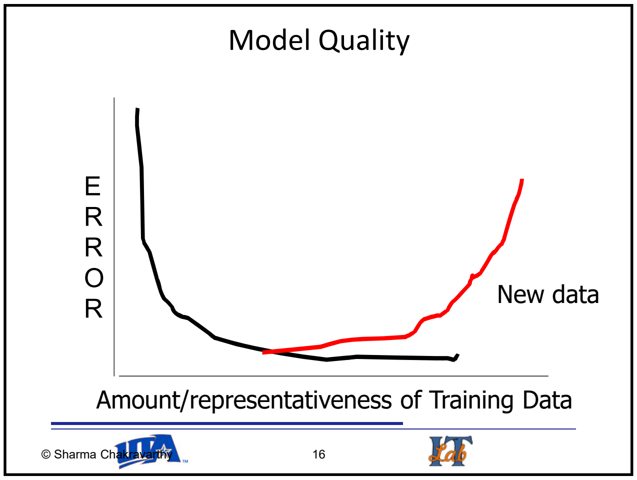


Using a Model

- Quantitative
  - Automated process
  - Classification/scoring done periodically (every month, when mailing is done, …)
    - Classification into a finite set
    - Estimate continuous numerical value (e.g., total worth of a customer)
    - Scoring (a probability value)



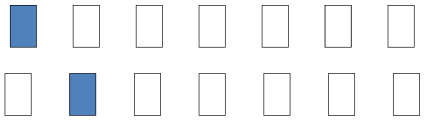
Model Testing



Model Quality

Cross-Validation
(k-fold cross-validation)

- Randomly partition the data into k sets (of equal size)
- Use set i for validating and build the model using the rest (sets 1, 2, …, i-1, i+1, …, k)
- Repeat the above process for i from 1 through k



Why do Cross-Validation?

- Does it improve accuracy of the model?
- No!

- Then why is it done?
  - It measures the predictive performance of the model
  - Averaged to give an estimate of the model's predictive performance
  - If the accuracy varies, you may want to generate a different model with varying training and test cases!

- It is a model validation technique

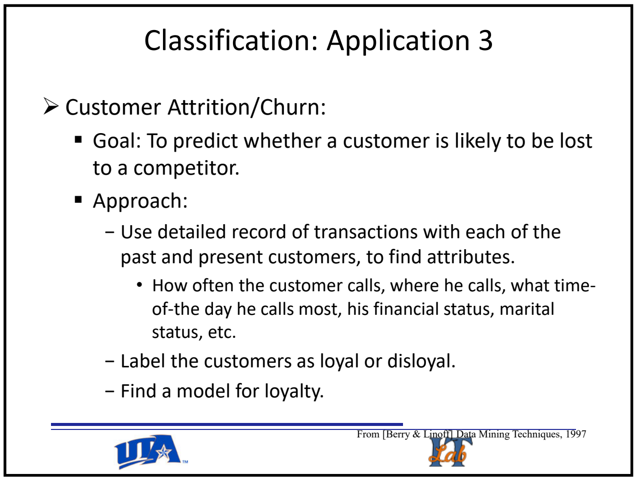
Classification: Application 1

- 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.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

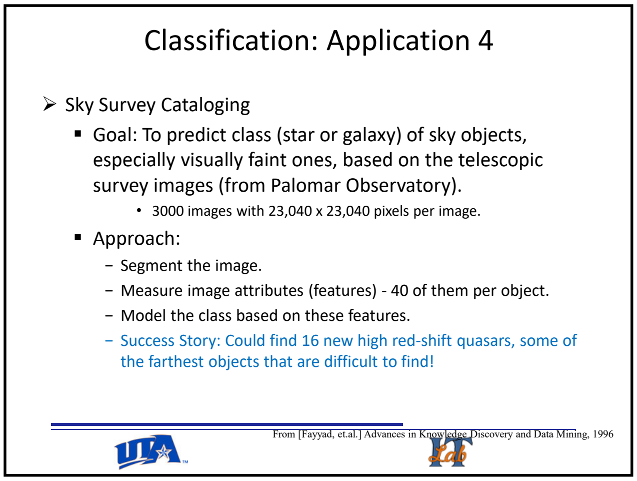
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions 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.

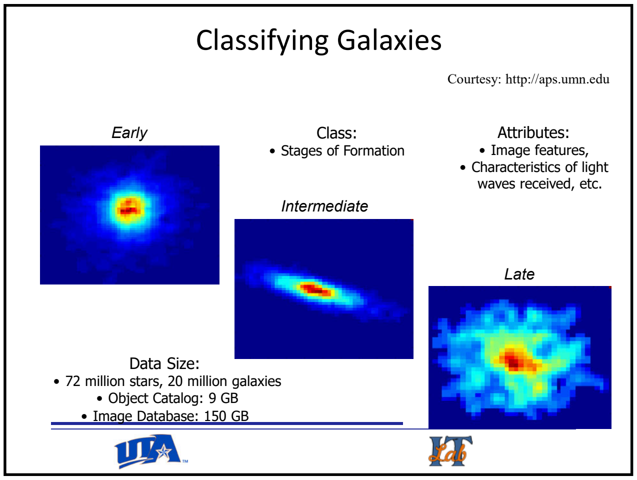
From [Berry & Linoff] Data Mining Techniques, 1997



Classification: Application 4

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



Classifying Galaxies

Courtesy: http://aps.umn.edu

Early

Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate

Late

Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Thank You !!!

For more information visit:
http://itlab.uta.edu

13 December 20 24 BDA 2018 (Warangal)