Data Mining Concept

Sunee Pongpinigpinyo

References

- Discovering Knowledge in Data
  - *Daniel T Larose, 2005*
- Data Mining: Concepts and Techniques, 2nd Edition, 2005
  - *Micheline Kamber, Jiawei Han*
- Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, 2005
  - *Ian H. Witten, Eibe Frank*
- Introduction to Data Mining, 2006
  - *Pang-Ning Tan, Michael Steinbach, and Vipin Kumar*

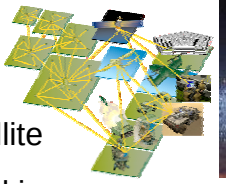
Why Mine Data? Commercial Viewpoint

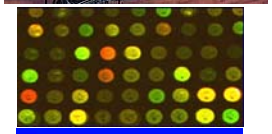
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

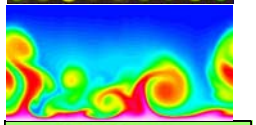
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

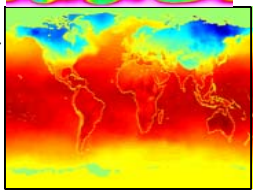
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data







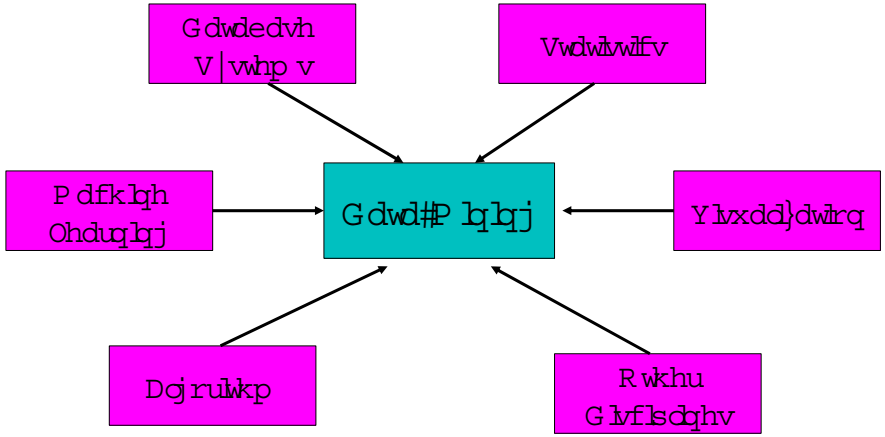


- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

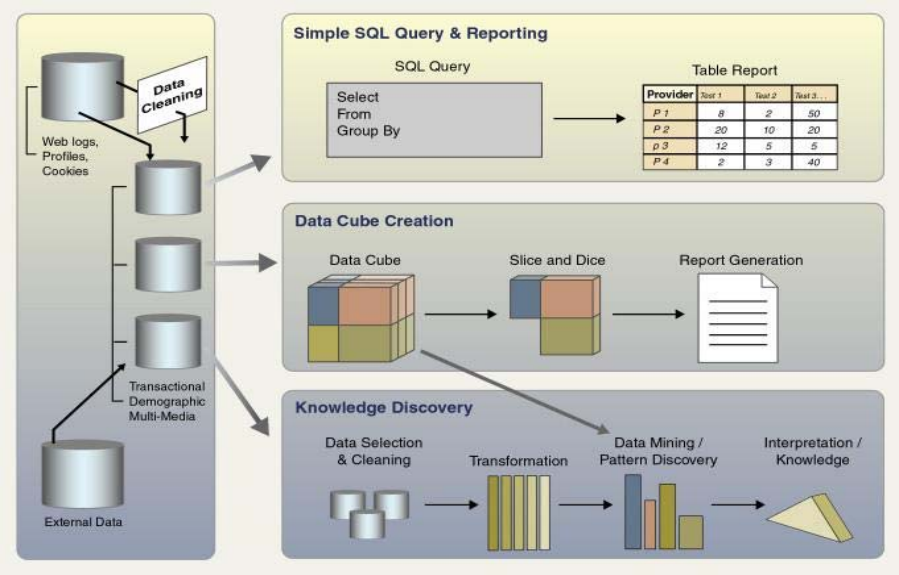
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

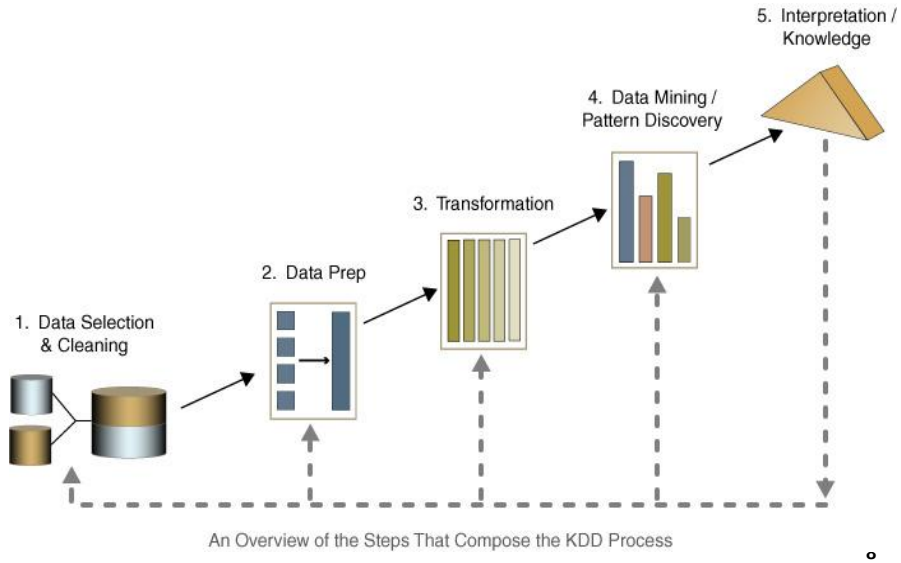
Data Mining: Confluence of Multiple Disciplines



Data Management Environments and Data Mining



Knowledge Discovery In Databases Process

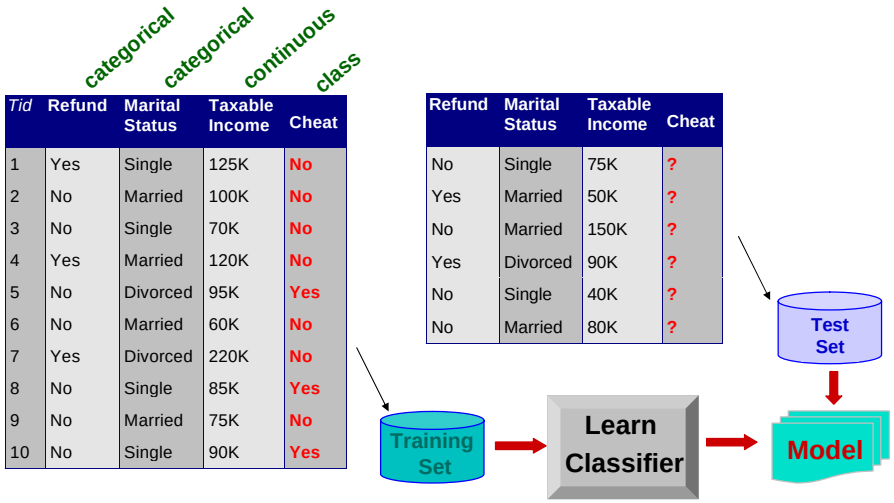


An Overview of the Steps That Compose the KDD Process

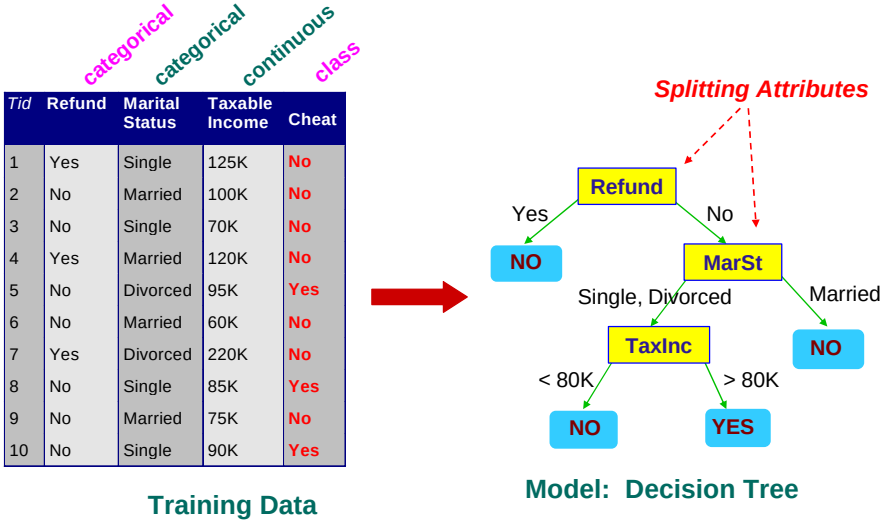
Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
  - Object-relational database
  - Spatial and temporal data
  - Time-series data
  - Multimedia database
  - Heterogeneous and legacy database
  - Text databases & WWW

Classification Example

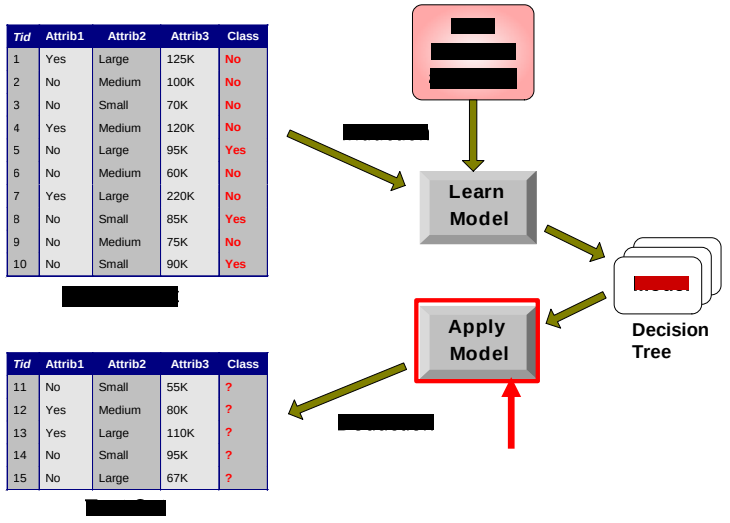


Example of a Decision Tree



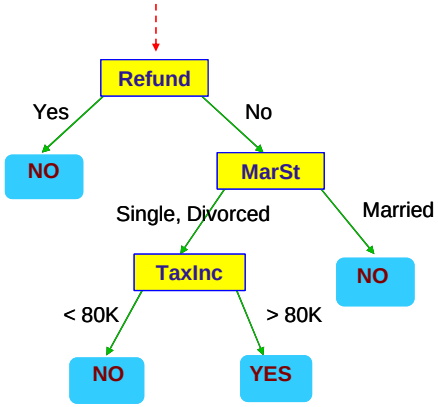
There could be more than one tree that fits the same data!

Decision Tree Classification Task



Apply Model to Test Data

Start from the root of tree.



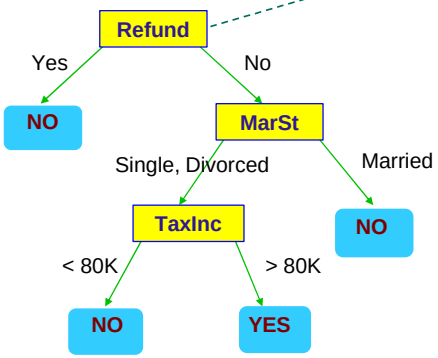
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

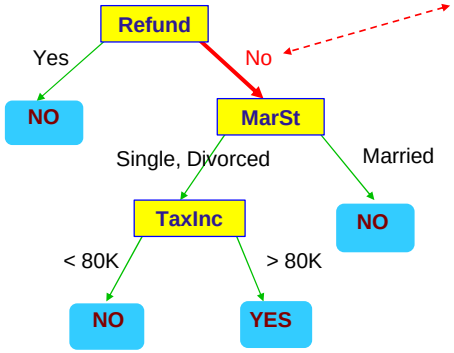
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

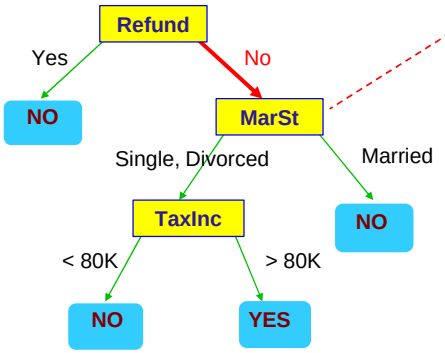
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

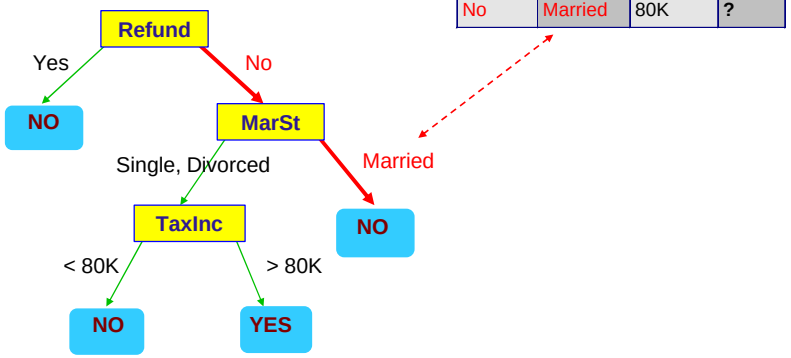
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

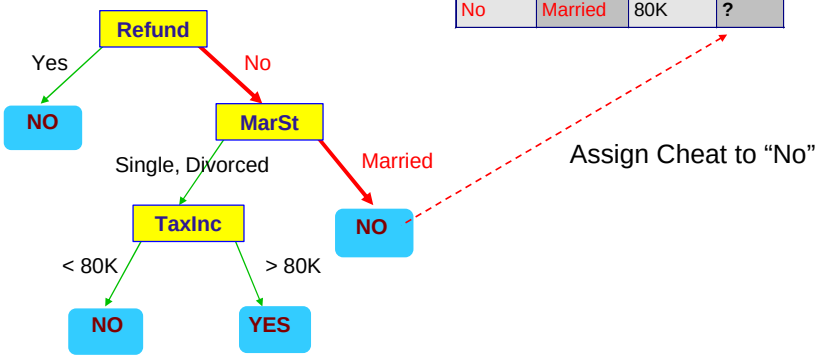
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



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 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

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

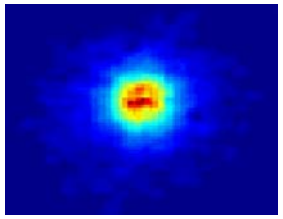
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (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

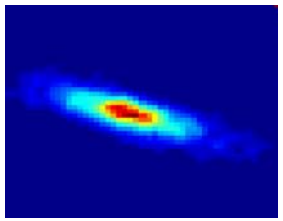
Classifying Galaxies

Courtesy: http://aps.umn.edu

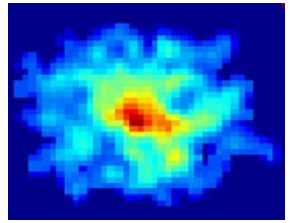
Early



Intermediate



Late



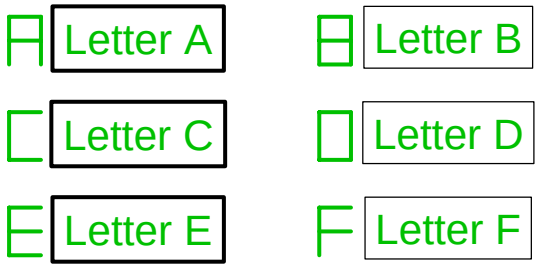
Class:
- **Stages of Formation**

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

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

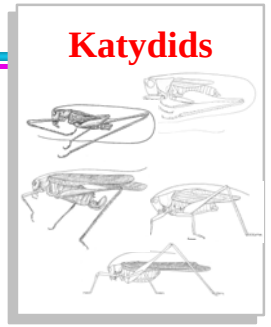
Letter Recognition

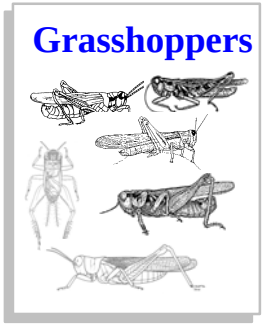
View letters as constructed from 5 components:



x4

Given a collection of annotated data. (in this case 5 instances of **Katydids** and five of **Grasshoppers)**, decide what type of insect the unlabeled example is.



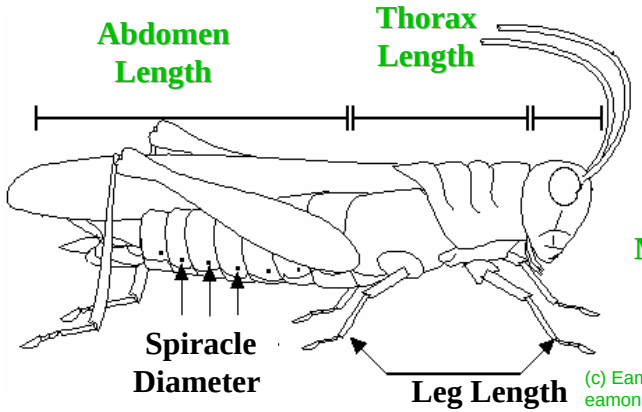


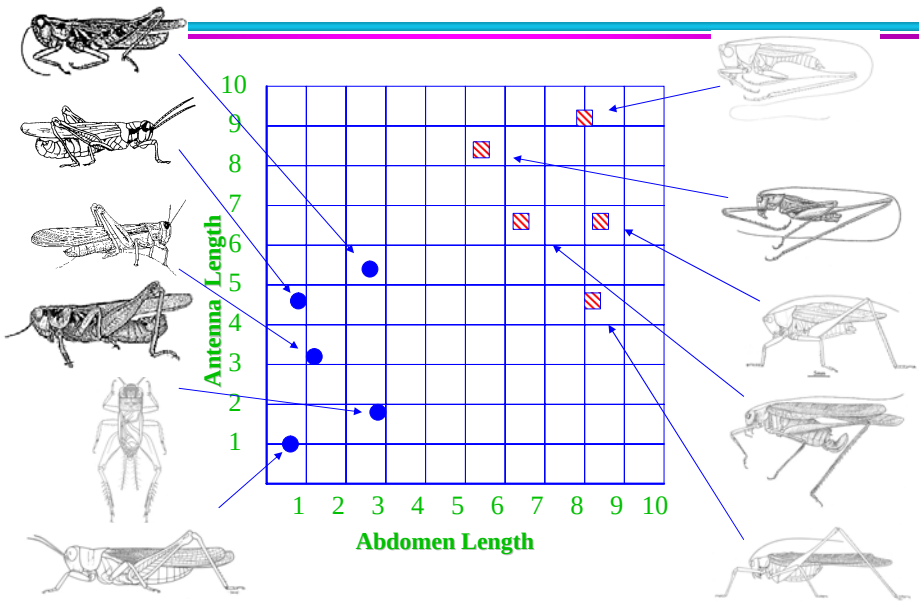
ภาพนิ่ง 26

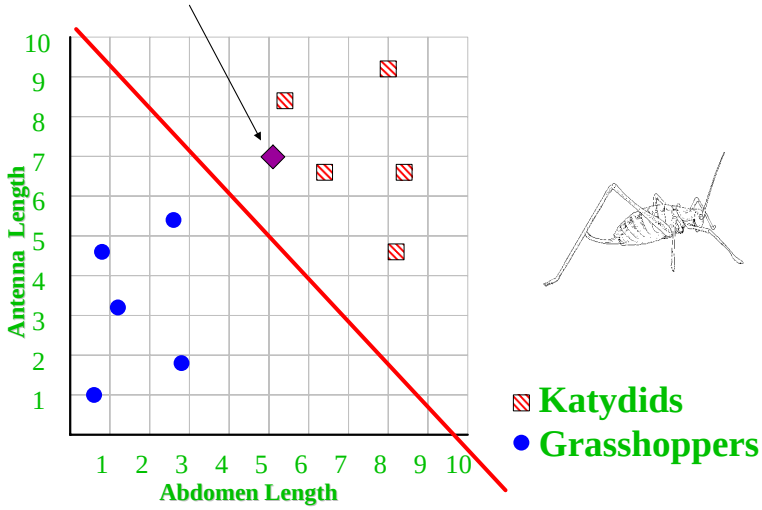
u1

Color {Green, Brown, Gray, Other}

Has Wings?







Clustering Definition

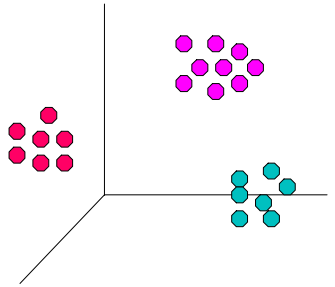
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

☒Hxfdgkdq#Glvwdqfh#Edvhg#Foxvwhulqj#lq#60G#vsdfh1

Intracluster distances are minimized

Intercluster distances are maximized



What is Similarity?

Clustering: Application

- 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.

Hierarchical Clustering Example

Iris Data Set

Setosa

Versicolor

Virginica

- The data originally appeared in Fisher, R. A. (1936). "The Use of Multiple Measurements in Axonomic Problems," Annals of Eugenics 7, 179-188.
- Hierarchical Clustering Explorer Version 3.0, Human-Computer Interaction Lab, University of Maryland, http://www.cs.umd.edu/hcil/multi-

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

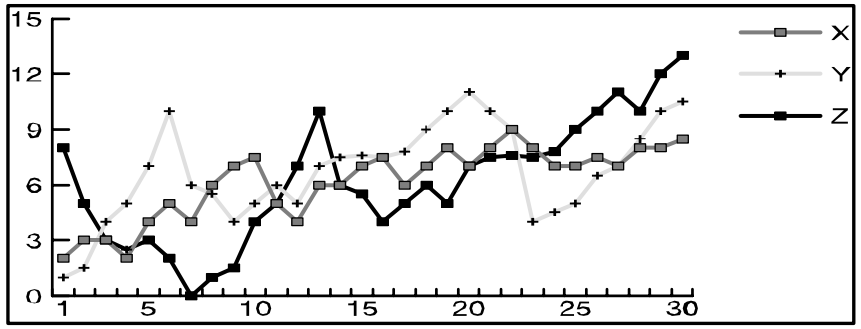
    {doughnut , … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Doughnut in the antecedent => Can be used to see which products would be affected if the store discontinues selling Doughnut .
  - Doughnut in antecedent *and* Potato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

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.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Ex: Time Series Analysis

- Example: Stock Market
- Predict future values
- Determine similar patterns over time
- Classify behavior

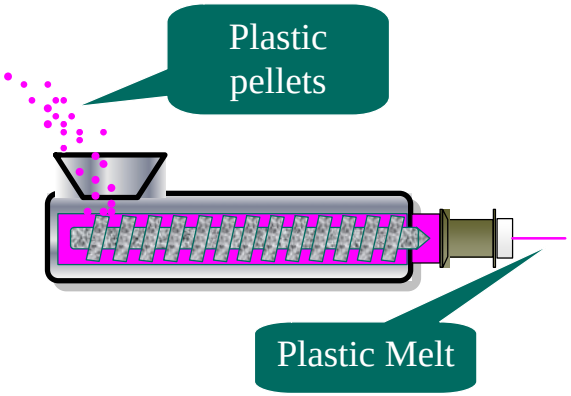


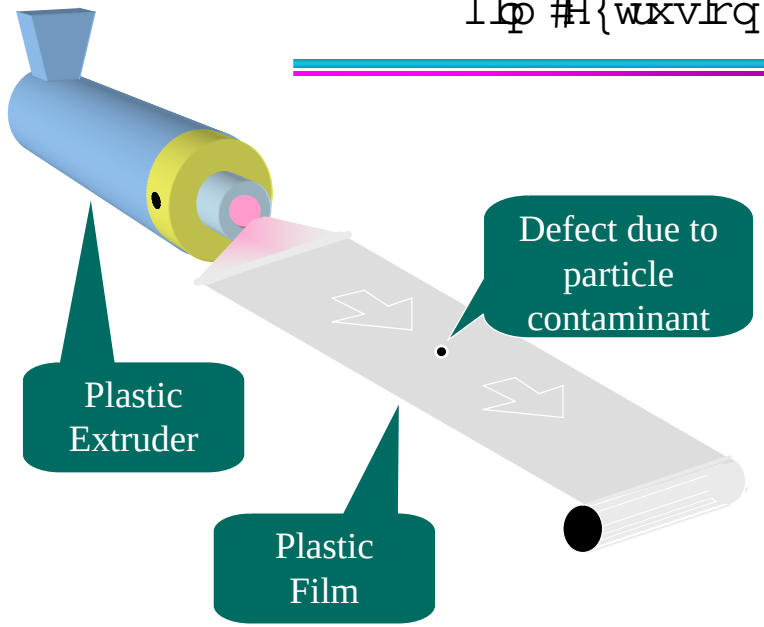
Examples of data mining in science & engineering

1. Data mining in Chemical Engineering

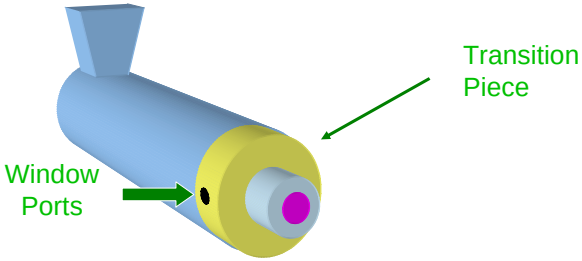
   "Data Mining for In-line Image Monitoring of Extrusion Processing" K.Torabi, L D. Ing, S. Sayad, and S.T. Balke

Plastics Extrusion

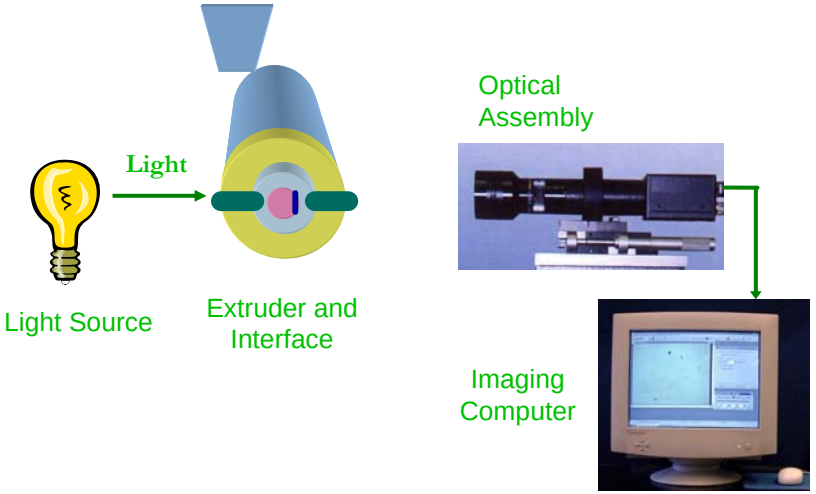




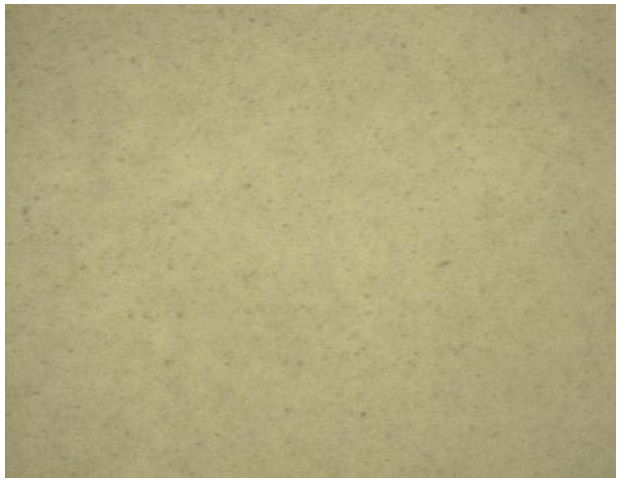
In-Line Monitoring



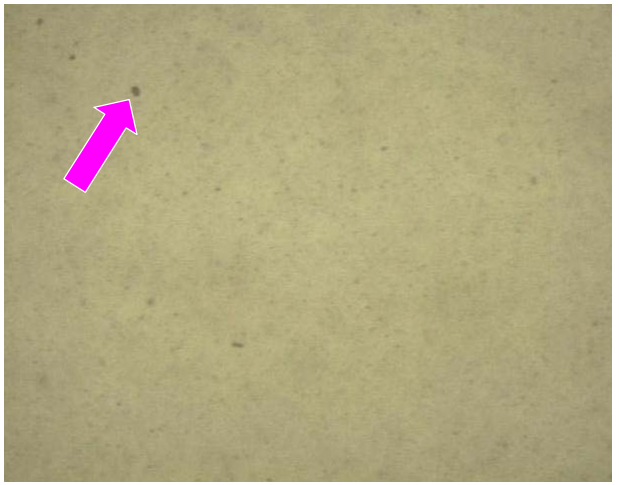
In-Line Monitoring



Melt Without Contaminant Particles (WO)



Melt With Contaminant Particles (WP)



Basic Steps in Data Mining

1. Define the problem
2. Build data mining database
3. Explore data
4. Prepare data for modeling
5. Build model
6. Evaluate model
7. Deploy model

1. Define the problem

Classify images into those with particles (WP) and those without particles (WO).

WO

WP

2. Build a data mining database

- 2000 Images
- 54 Input variables all numeric
- One output variables with two possible values (With Particle and Without Particle)

4. Prepare data for modeling

- Pre-processed images to remove noise
- Dataset 1 with sharp images: 1350 images including 1257 without particles and 91 with particles
- Dataset 2 with sharp and blurry images: 2000 images including 1909 without particles and blurry particles and 91 with particles
- 54 Input variables, all numeric
- One output variable, with two possible values (WP and WO)

5. Build a model

Classification:

- OneR
- Decision Tree
- 3-Nearest Neighbors
- Naïve Bayesian

6. Evaluate Models

10 -fold cross-validation

Dataset	*Attrib.*	*Class*	*One-R*	*C4.5*	*3.N.N*	*Bayes*
Sharp Images	54	2	99.9	99.8	99.8	95.8
Sharp + Blurry Images	54	2	98.5	97.8	97.8	93.3
Sharp + Blurry Images	54	3	87	87	84	79

If pixel_density_max < 142 then WP

7. Deploy model

❖ A Visual Basic program will be developed to implement the model.