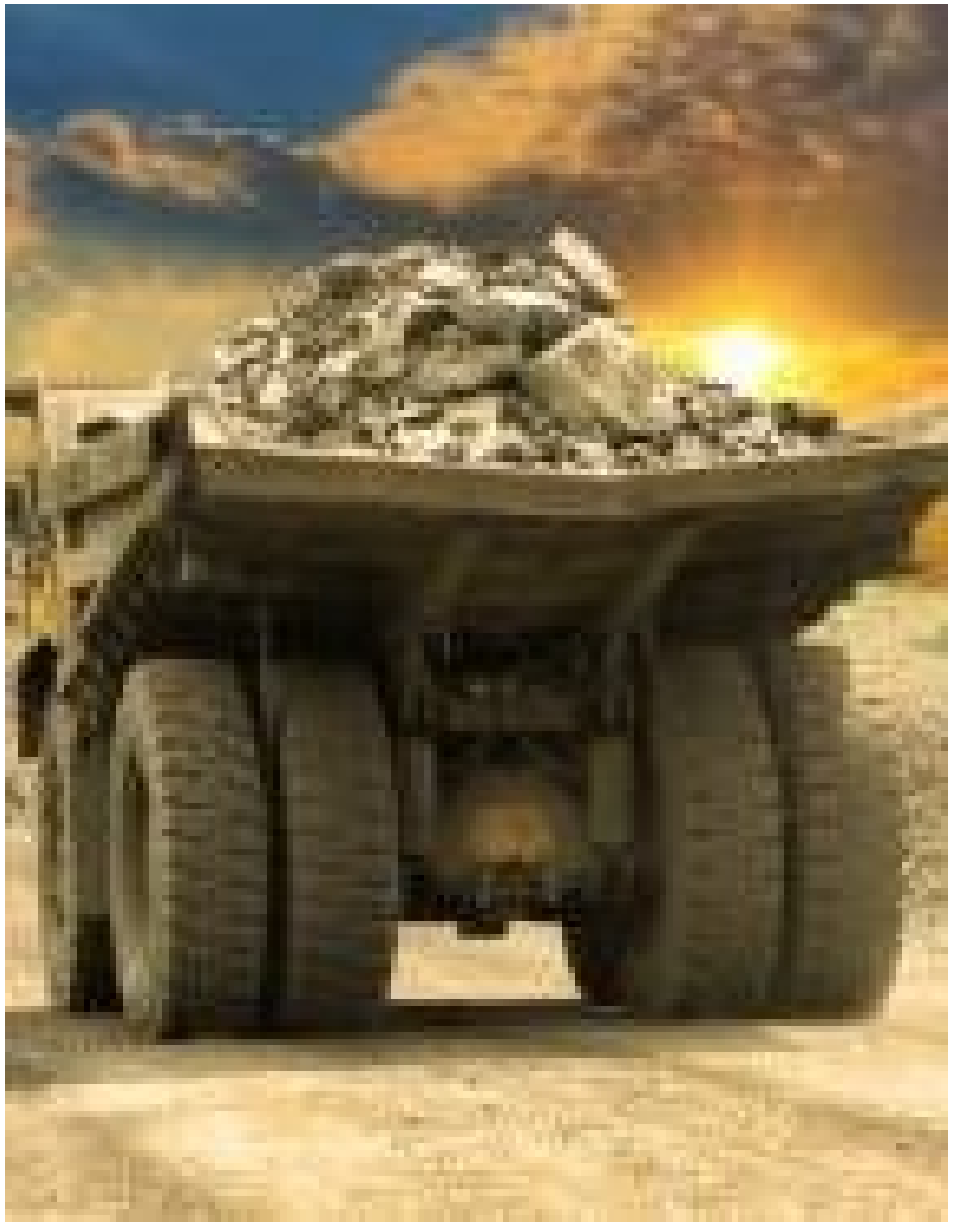
Introduction to Data Mining Methods and Tools

by Michael Hahsler

Agenda

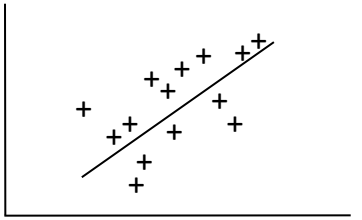
- What is Data Mining?
- **Data Mining Tasks**
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues



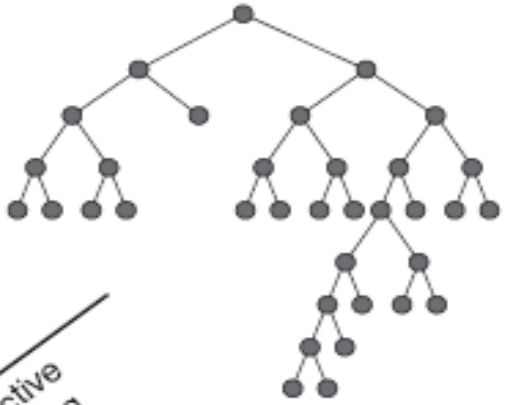
Data Mining Tasks

Descriptive Methods	Find human-interpretable patterns that describe the data.
Predictive Methods	Use some features (variables) to predict unknown or future value of other variable.

Data Mining Tasks



Regression



Classification

Data

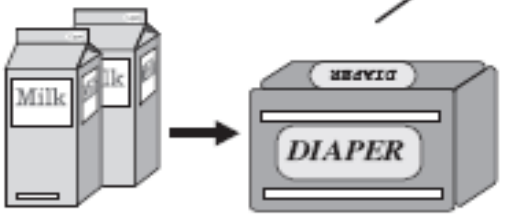
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

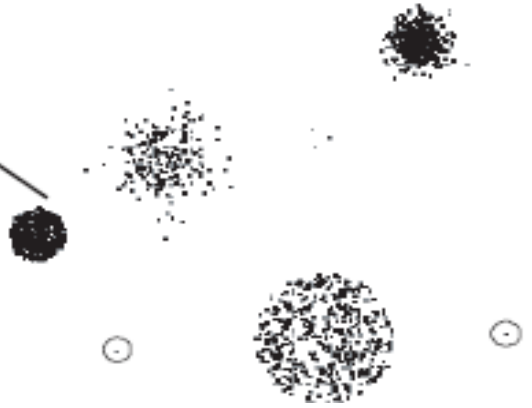
Cluster Analysis

Predictive Modeling

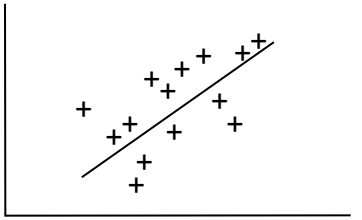
Association Analysis

Anomaly Detection

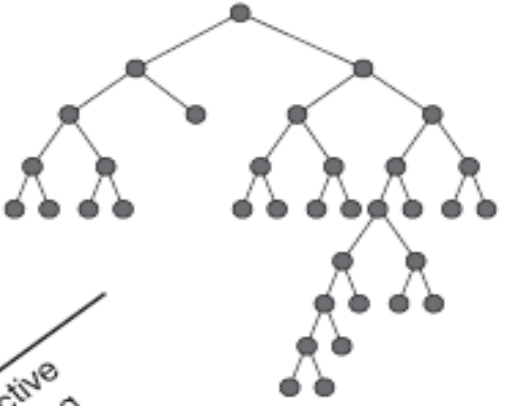




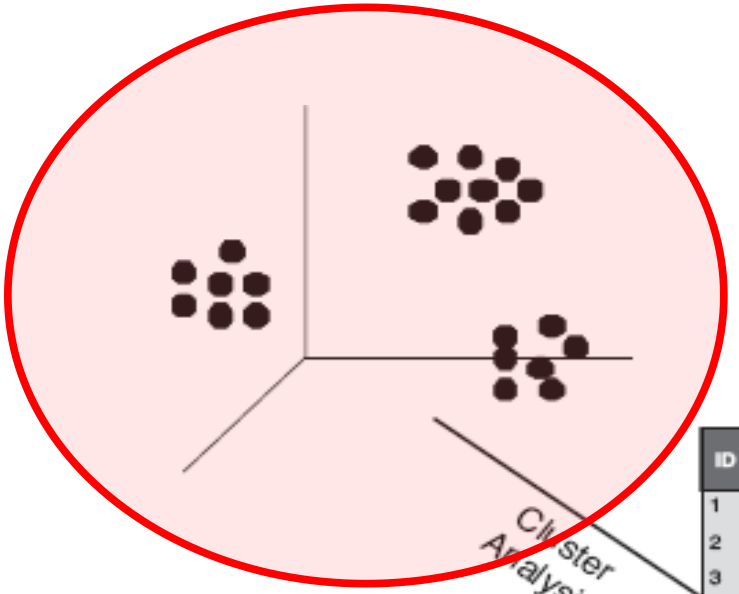
Data Mining Tasks



Regression



Classification



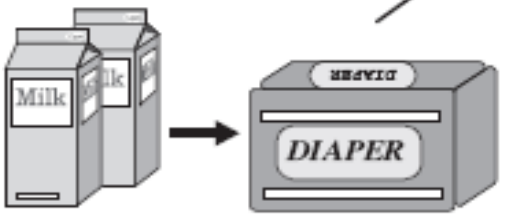
Cluster Analysis

Data

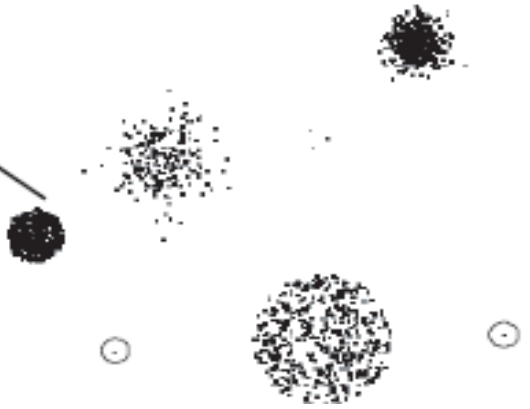
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive Modeling

Association Analysis



Anomaly Detection

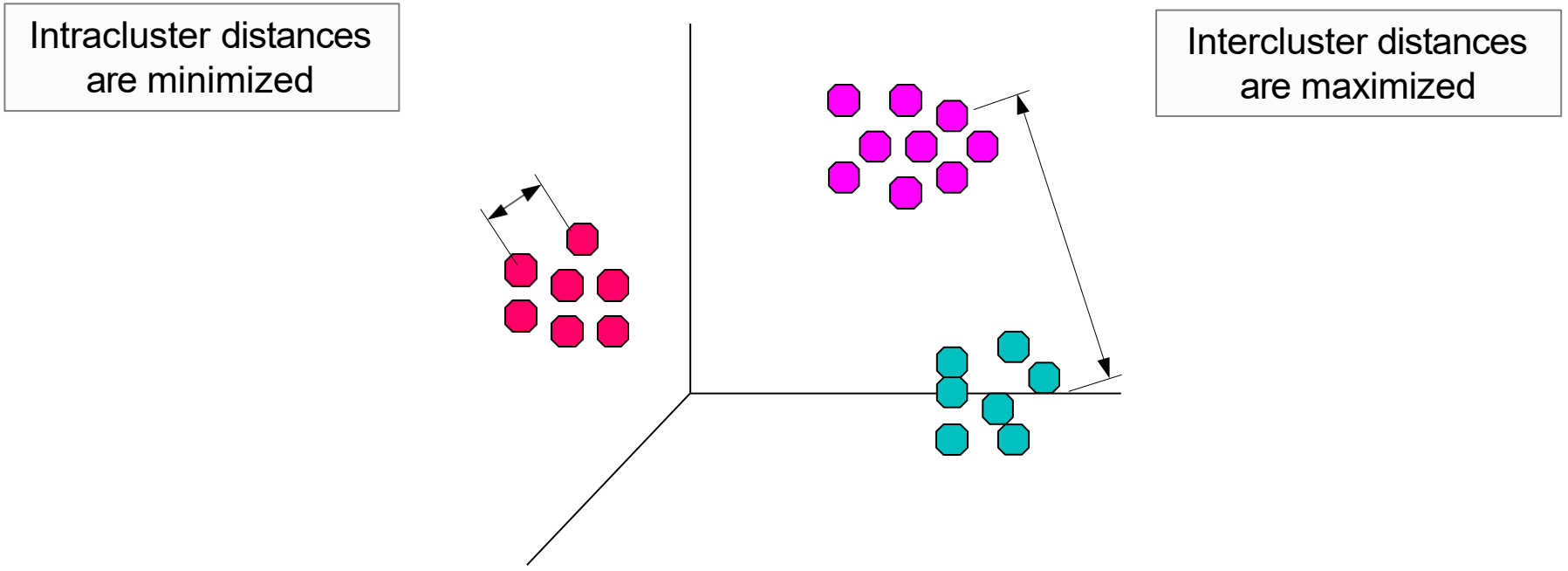


Clustering

Group points such that

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Ideal grouping is not known → Unsupervised Learning



Euclidean distance based clustering in 3-D space.

Clustering: Market Segmentation

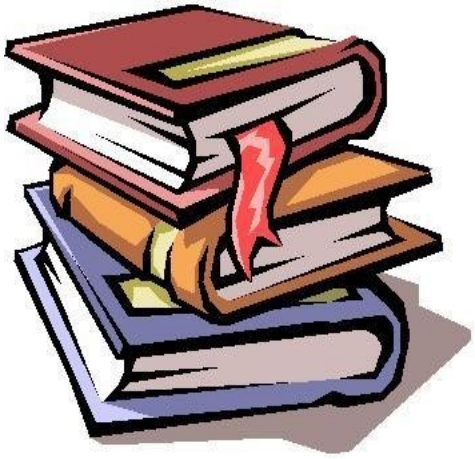
Goal: subdivide a market into distinct subsets of customers. Use a different marketing mix for each segment.

Approach:

1.Collect different attributes of customers based on their geographical and lifestyle related information and observed buying patterns.

2. Find clusters of similar customers.

Clustering Documents



Goal: Find groups of documents that are similar to each.

Approach: Identify frequently occurring terms in each document. Define a similarity measure based on term co-occurrences. Use it to cluster.

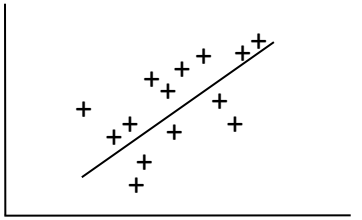
Gain: Can be used to organize documents or to create recommendations.

Clustering: Data Reduction

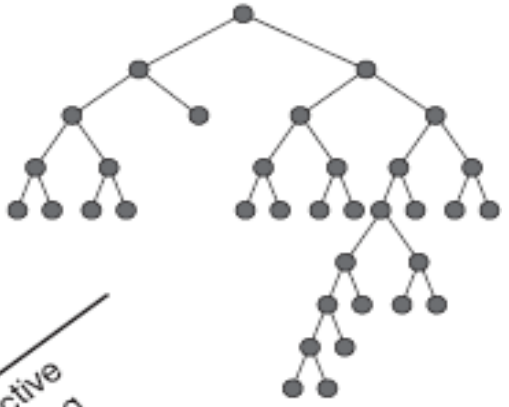
Goal: Reduce the data size for predictive models.

Approach: Group data given a subset of the available information and then use the group label instead of the original data as input for predictive models.

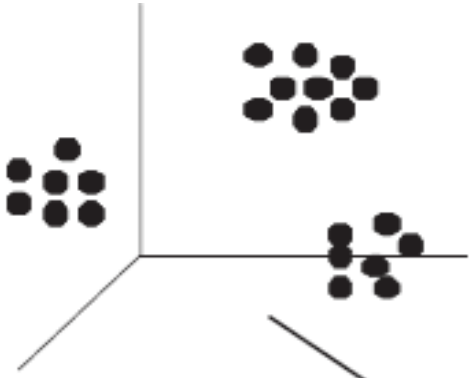
Data Mining Tasks



Regression



Classification



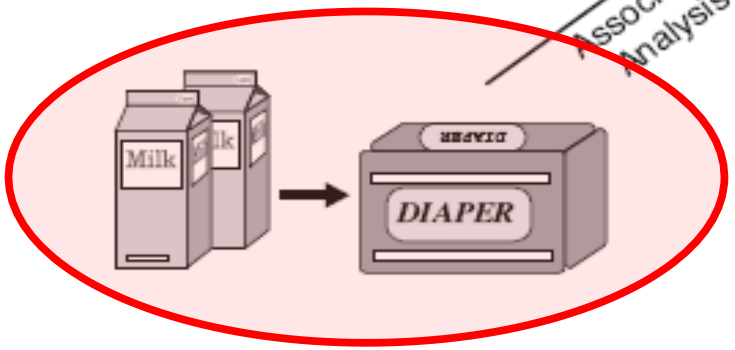
Cluster Analysis

Data

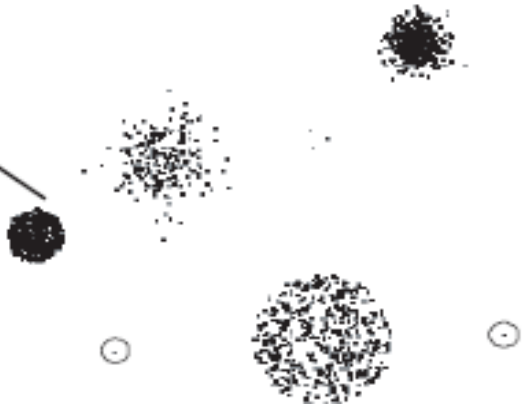
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive Modeling

Association Analysis



Anomaly Detection



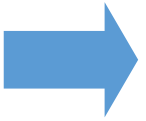
Association Rule Discovery

- Given is a set of transactions. Each contains a number of items.
- Produce dependency rules of the form

  LHS → RHS
- which indicate that if the set of items in the LHS are in a transaction, then the transaction likely will also contain the RHS item.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Transaction data



{Milk} → {Coke}

{Diaper, Milk} → {Beer}

Discovered Rules

Association Rule Discovery Marketing and Sales Promotion

- Let the rule discovered be

  {Potato Chips, … } → {Soft drink}

- **Soft drink as RHS**: What should be done to boost sales? Discount Potato Chips?
- **Potato Chips in LHS**: Shows which products would be affected if the store discontinues selling Potato Chips.
- **Potato Chips in LHS and Soft drink in RHS**: What products should be sold with Potato Chips to promote sales of Soft drinks!

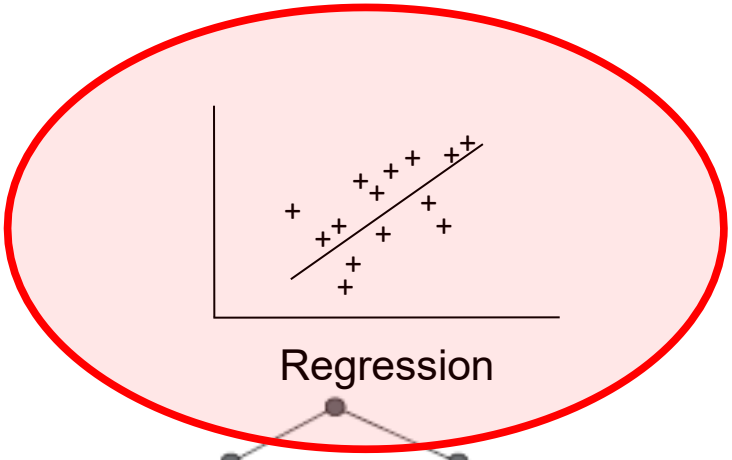
Association Rule Discovery Supermarket shelf management

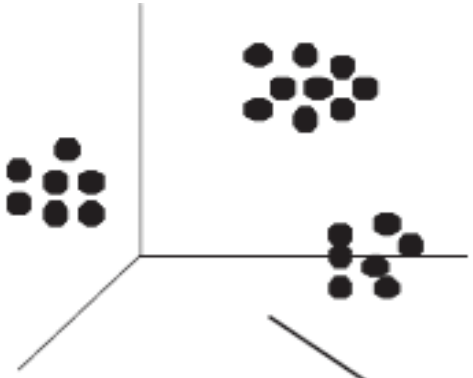
- **Goal**: To identify items that are bought together by sufficiently many customers.
- **Approach**:
  - Process the point-of-sale data to find dependencies among items.
  - Place dependent items
    - close to each other (convenience).
    - far from each other to expose the customer to the maximum number of products in the store.

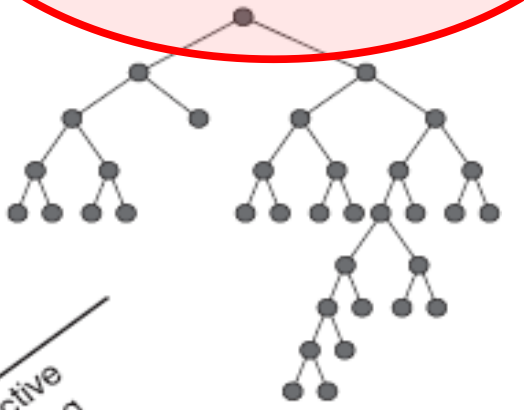
Association Rule Discovery Inventory Management

- **Goal**: Anticipate the nature of repairs to keep the service vehicles equipped with right parts to speed up repair time.
- **Approach**: Process the data on tools and parts required in previous repairs at different consumer locations and discover co-occurrence patterns.

Data Mining Tasks







Classification

Cluster Analysis

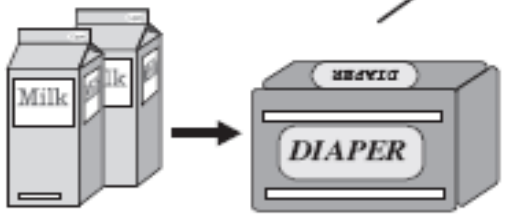
Predictive Modeling

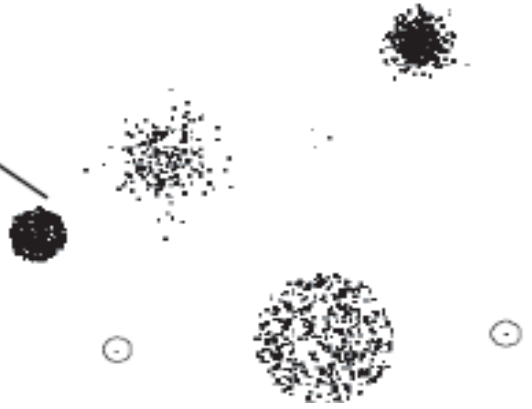
Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Association Analysis

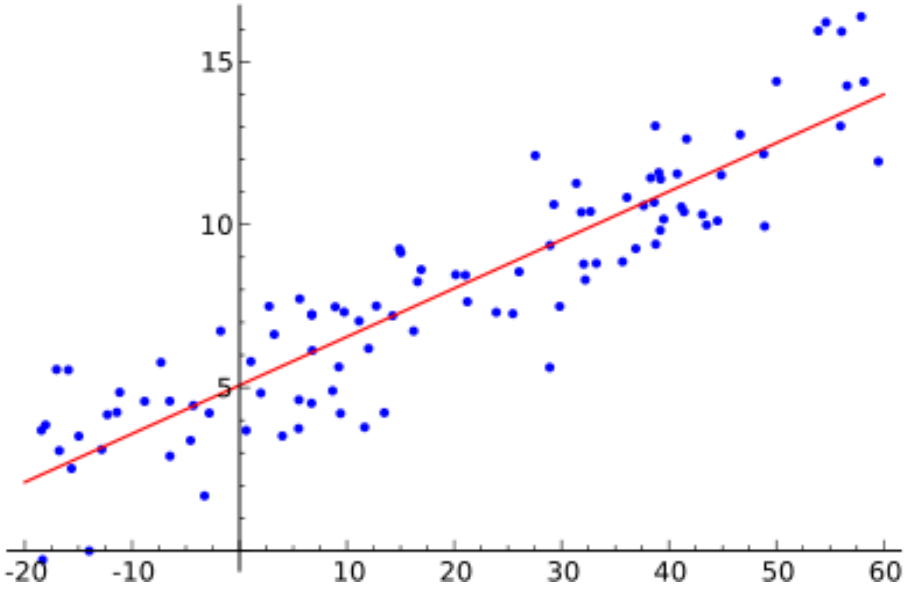
Anomaly Detection





Regression

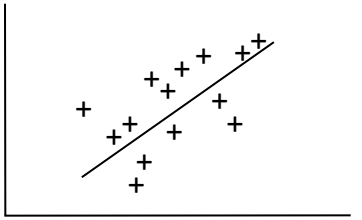
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Studied in statistics and econometrics.



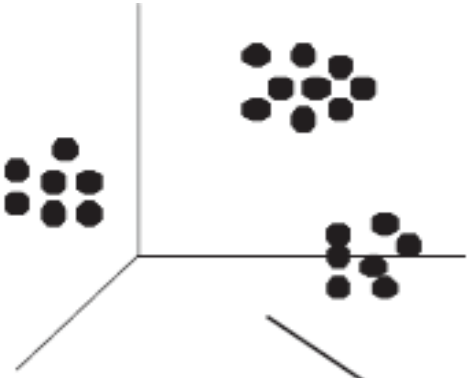
Applications:

- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices (autoregressive models).

Data Mining Tasks



Regression

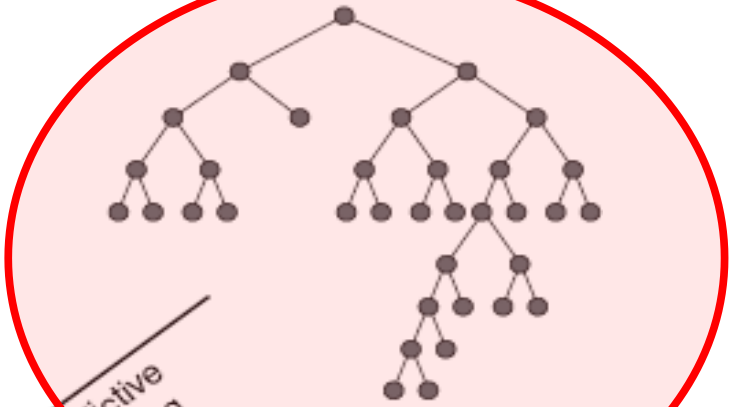


Cluster Analysis

Data

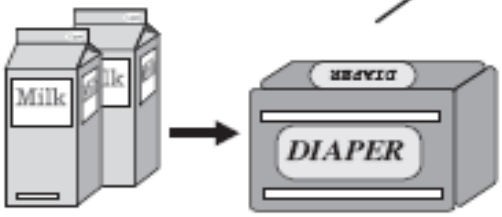
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive Modeling

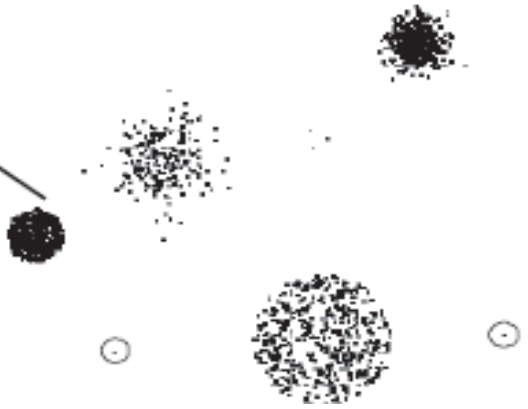


Classification

Association Analysis



Anomaly Detection



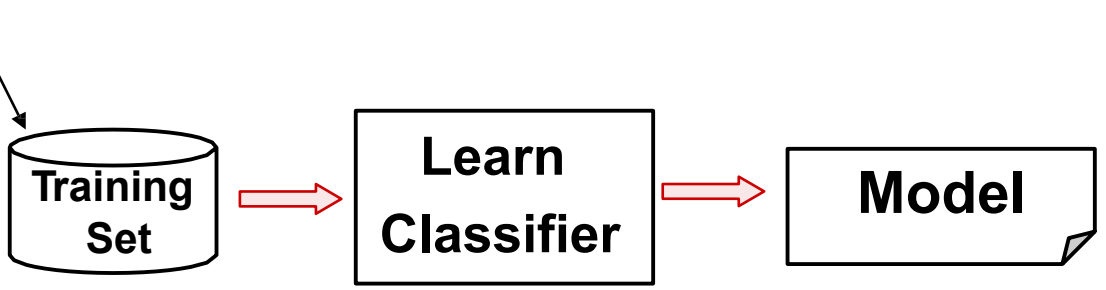
Classification

Find a **model** for the class attribute as a function of the values of other attributes/features.

Class information is available → **Supervised Learning**

class

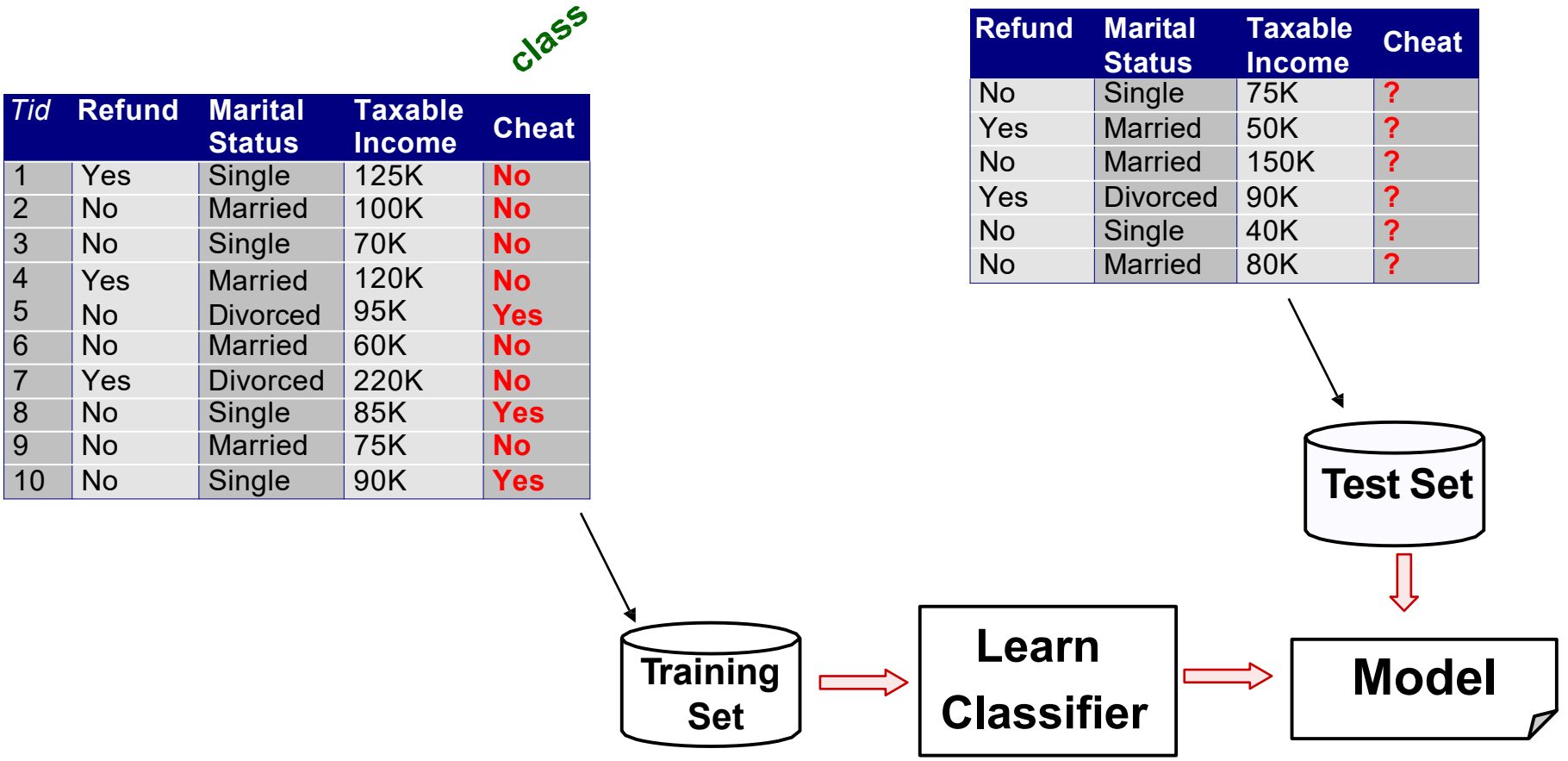
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Classification

Find a **model** for the class attribute as a function of the values of other attributes/features.

Goal: assign new records to a class as accurately as possible.



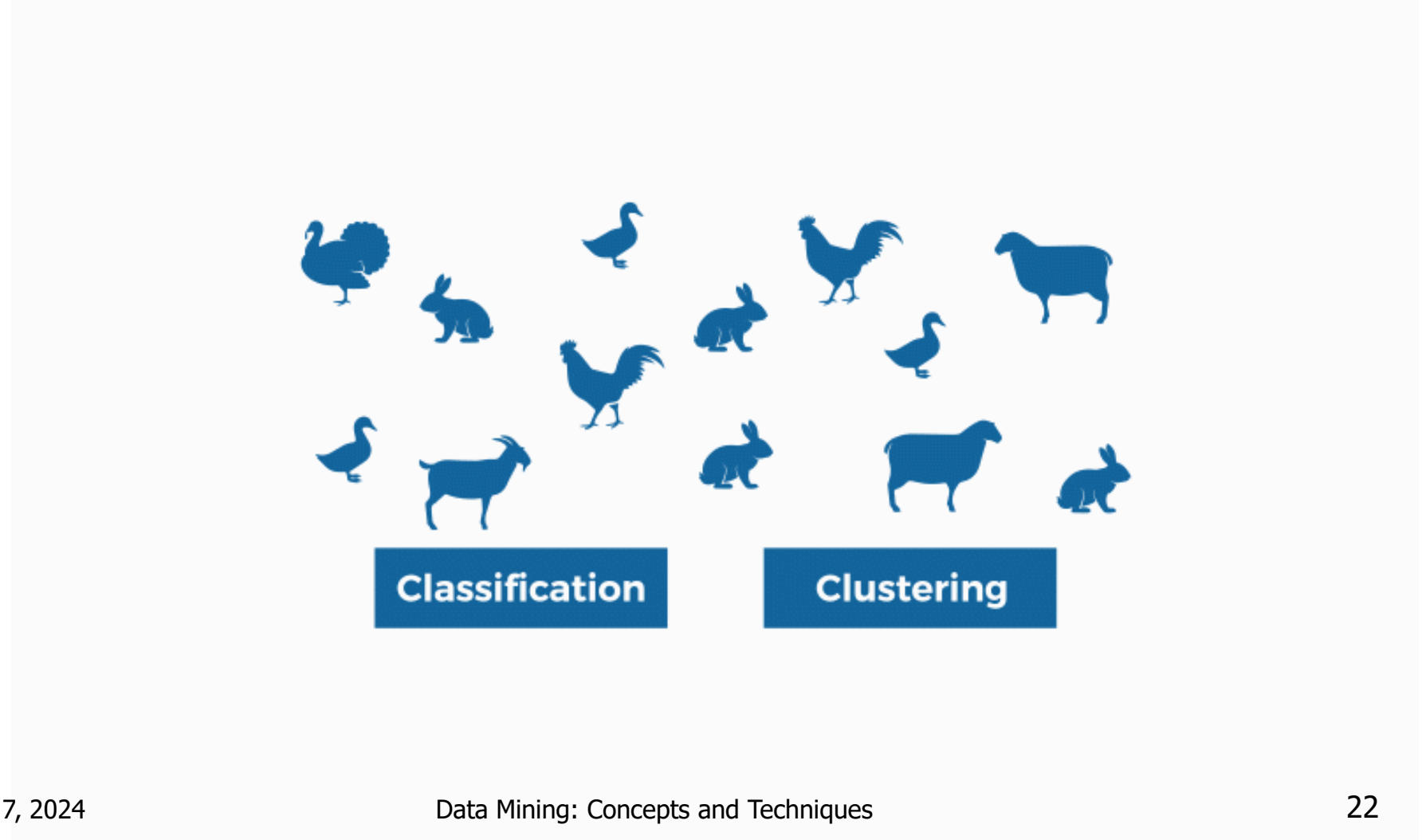
Classification: Direct Marketing

- Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new product.
- Approach:
  - Use the data for a similar product introduced before or from a focus group. We have customer information (e.g., demographics, lifestyle, previous purchases) and know which customers decided to buy and which decided otherwise. This buy/don't buy decision forms the class attribute.
  - Use this information as input attributes to learn a classifier model.
  - Apply the model to new customers to predict if they will buy the product.

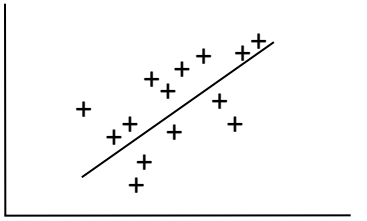
Classification: Customer Attrition/Churn

- 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 (frequency, recency, complaints, demographics, etc.).
  - Label the customers as loyal or disloyal.
  - Find a model for disloyalty.
  - Rank each customer on a loyal/disloyal scale (e.g., churn probability).

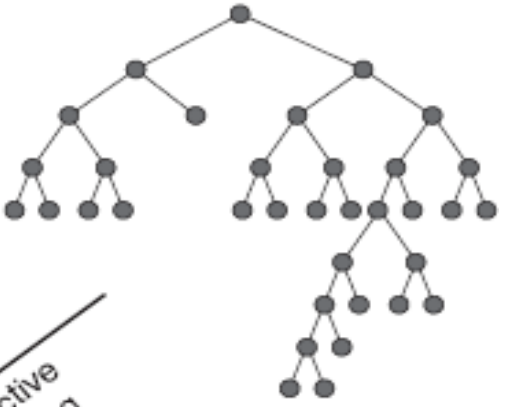
Classification vs



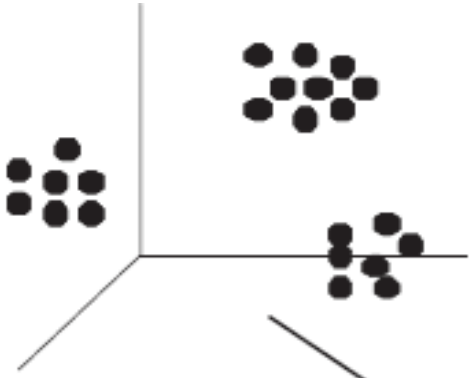
Data Mining Tasks



Regression



Classification



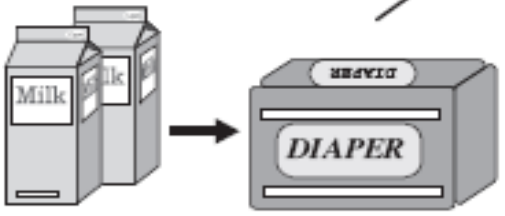
Cluster Analysis

Data

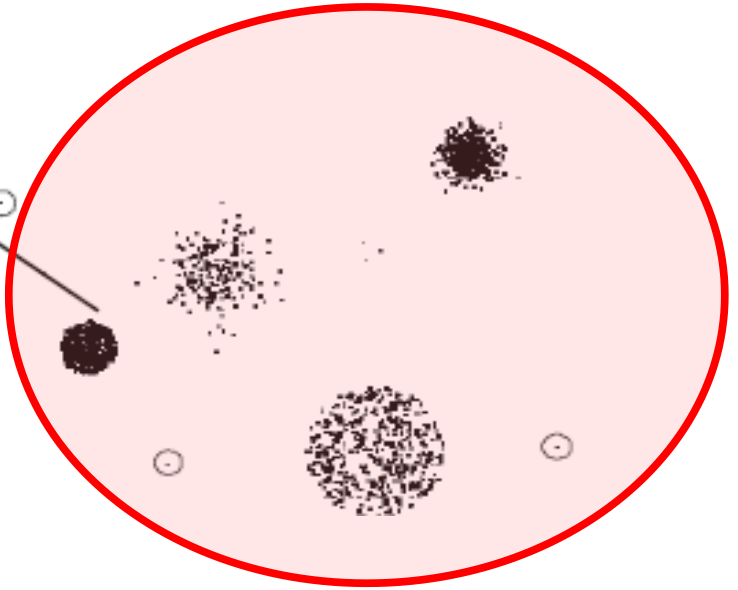
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive Modeling

Association Analysis



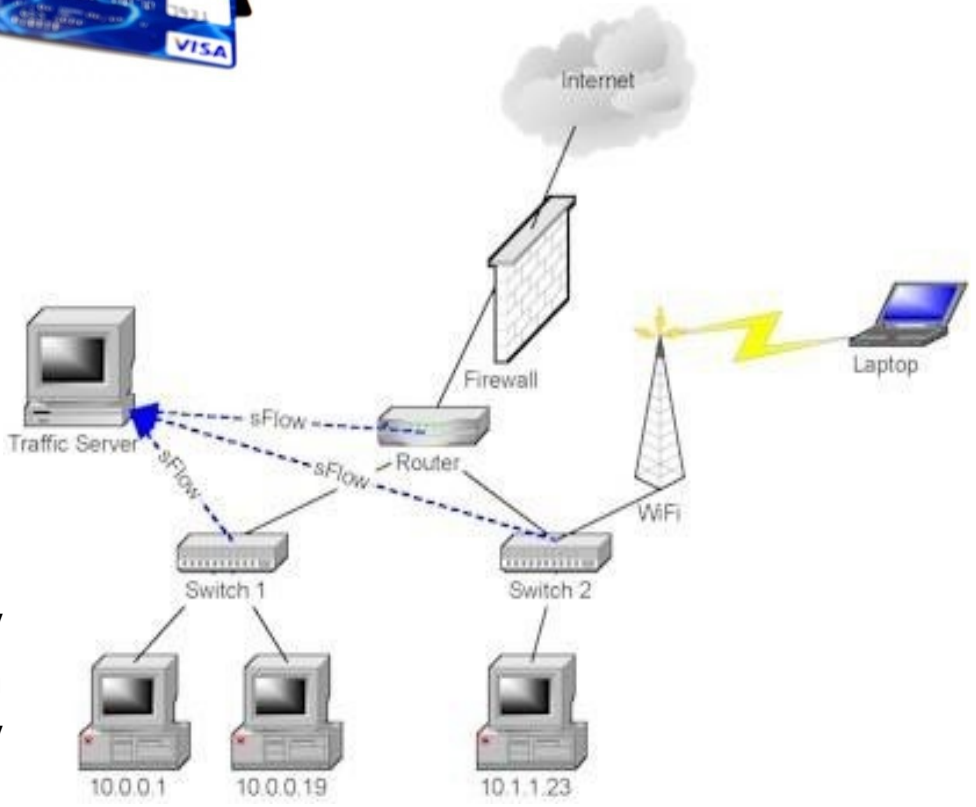
Anomaly Detection



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

Other Data Mining Tasks

Text mining – document clustering, topic models

Graph mining – social networks

Data stream mining/real time data mining

Mining spatiotemporal data (e.g., moving objects)

Visual data mining

Distributed data mining