Introduction

- INFO411/911 -

"Introduction to Data Mining and Knowledge Discovery"

Presented by

A/Prof Wanqing Li

Outline

- Def. Knowledge discovery and data mining
- Data mining tasks
- Business/Analytics applications
- Medicine/Science/Engineering applications
- Input data
- Classification of attributes
- Types of data sets
- Data quality
- Data preparation
- Data integration

Introduction

- Data is produced at a phenomenal rate.
- Our ability to collect and store data has grown.
- Users expect more sophisticated information.
- But how?

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

- Example: Search for flight MH370.
  - Lots of data collected (sonar, satellite, etc.)
    - Is information on the whereabouts of MH370 captured by the data?
    - Where is MH370?

Introduction

- Data Mining Objective: Fit data to a model
- Expected Results: Higher-level meta information that may not be obvious when looking at raw data.
- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning

➢ Data Mining is an integral part of knowledge discovery in database (KDD).

- What is the difference of knowledge discovery and data mining?

What is knowledge discovery ?

Knowledge discovery is a process of converting raw data into useful information

Knowledge discovery consists of the following steps:

(1) *Data cleaning* to remove noise and inconsistent data

(2) *Data integration* to combine multiple data sources

(3) *Data selection* to retrieve data from databases

(4) *Data transformation* to get data into forms appropriate for data mining

(5) *Data mining* to extract data patterns

(6) *Pattern evaluation* to identify interesting patterns

(7) *Knowledge presentation* to present mined knowledge to the users

What is data mining ?

Data Mining is the science of developing, identifying, and using suitable machine assisted methods for extracting or uncovering useful information from data.

Other definitions:

- *Data mining* is the process of automatically discovering useful information in large data repositories.
- Data mining means extracting or "mining" knowledge from large amounts of data
- Data mining is an application to sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases
- Data mining is the process of semi-automatically analyzing large databases to find patterns that are: valid, novel, useful, and understandable.

What is data mining ?

Consequences of these definitions:

- “Data in large preexisting data repositories”:
  - Data has already been collected and has been stored.
  - Dataset is fixed.
- “Large amounts of data”:
  - Too much to process manually.
  - It does not mean that the dataset needs to be of a certain size.
- “Develop and identify suitable methods”:
  - There are many methods available. Task is to chose the right one for a given task and, if no suitable method can be found, then develop a suitable method.
  - Requires knowledge and understanding of data, task, and available methods.
- ”novel, useful, understandable results”:
  - Results are non-obvious and it should be possible to act on the result.
  - Humans should be able to interpret patterns and results.

Data mining ?

Common Pitfalls:

- Data needs to be available in a computer readable form (digital data).
  - Digital (binary) system is inherently discrete in nature but most information in the real world is continuous.
- Results need to be presented to users (humans).
  - A machine runs Data Mining methods to extract knowledge. Results are in digital form that can be understood by a machine.
  - It is the users' responsibility to "read" and "understand" the results.

Data scientist & analytics professionals?

Data scientist:

- Is an expert with knowledge of methods in data mining.
- Ability to acquire domain knowledge.
- Ability to understand the (complexity of the) data mining task.
- Ability to identify the most suitable method(s) for a given task.
- Ability to adapt methods to new data mining tasks.
- Ability to develop new methods for a given task.
- Ability to read, interpret, and understand the results.

Data scientist – Job trends 2020-2029

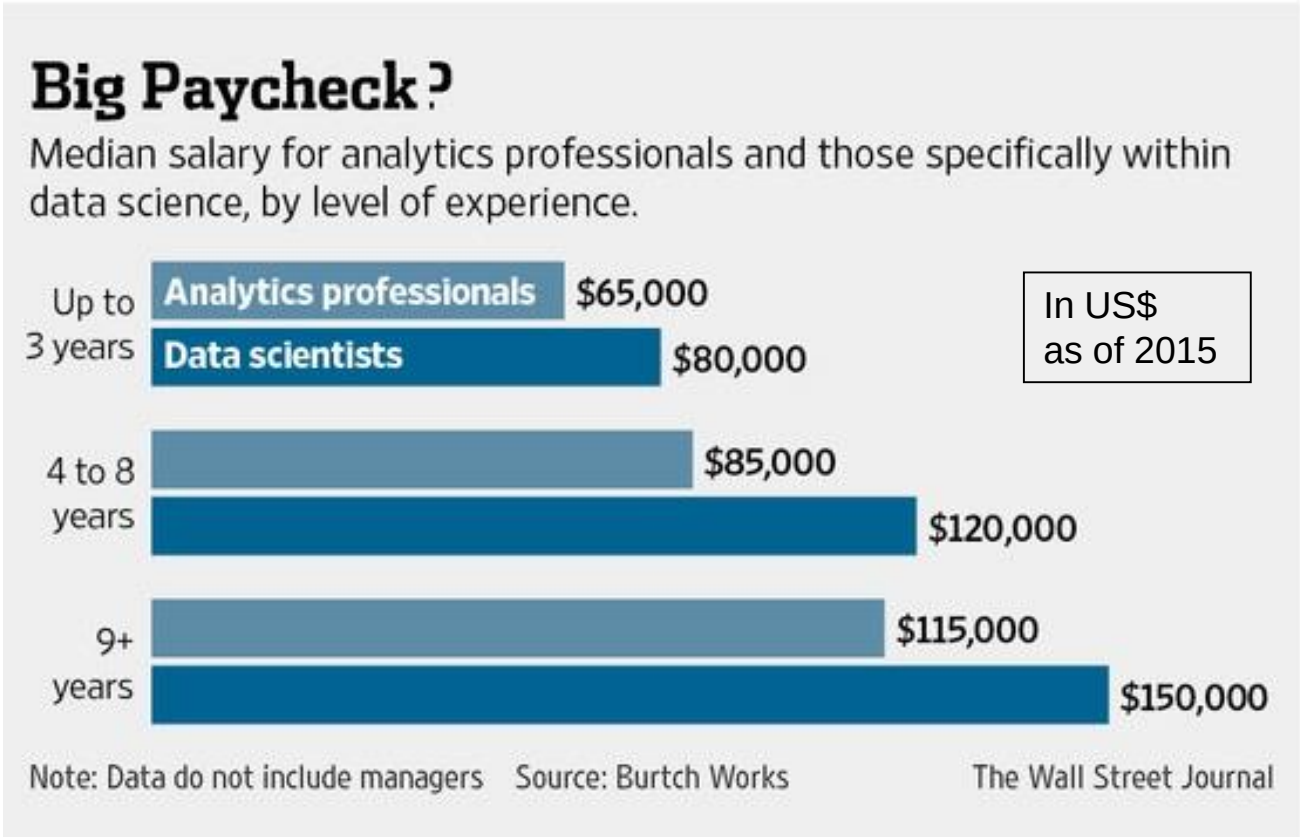
Data science has progressed from a relatively unknown term, to one of the most sought after disciplines in the job market

- **Data Scientist – The sexiest job of the 21st century**
- There will be a sharp increase in demand for data scientists by 2020. According to IBM, an increment by **364,000 to 2,720,000** openings will be generated in the year 2020. This demand will only grow further to an astonishing **700,000 openings**.
- Data Science is predicted to grow over the next decade. It is a staggering fact that over **90% of the data in the world was generated in just 2 years**. It is unimaginable to realize the amount of data that will be generated in the next decade. The demand for data scientists will rise by **28% by 2020 alone**. More and more industries are becoming data hungry and they need data to hold specialized data scientists who can craft products for the customers. About **11.5 Million jobs will be created by 2026 according to U.S. Bureau of Labor Statistics**.

https://data-flair.training/blogs/data-science-job-trends/

Data scientist

The demand for data scientists has created an attractive job market. The wages paid to data scientists reflects the value that employers acknowledge these analysts generate.



Data mining tasks

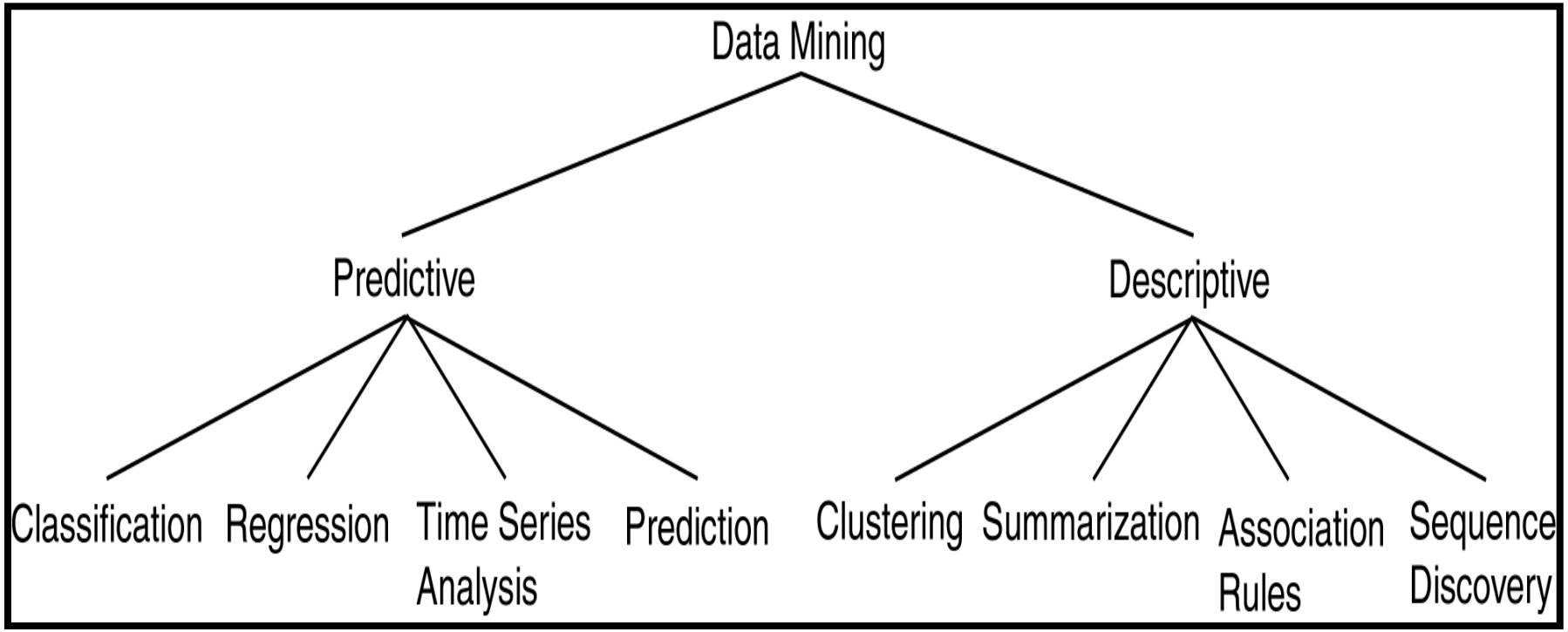
The tasks of data mining can be categorized into:

- ***Predictive tasks:***
  Objective of these tasks is to predict the value of a particular attribute based on the value of other attributes

- ***Descriptive tasks:***
  Objective of these tasks is to derive patterns such as correlations, trends, clusters, trajectories, and anomalies that summarize the underlying relationships in data

- ***Or both***

Descriptive and predictive models and tasks



© Prentice Hall

Data mining tasks

- ***Classification:*** Map data into predefined groups or classes.
  - Supervised learning
  - Pattern recognition
  - Prediction
- ***Regression:*** Map a data item to a real valued prediction variable.
  - Function approximation.
- ***Clustering:*** Group similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning

Data mining tasks (cont'd)

- ***Summarization:*** Map data into subsets with associated simple descriptions.
  - Characterization
  - Generalization
- ***Link Analysis:*** Uncover relationships among data.
  - Affinity Analysis
  - Association Rules
  - Sequential Analysis determines sequential patterns

Examples of Data mining tasks

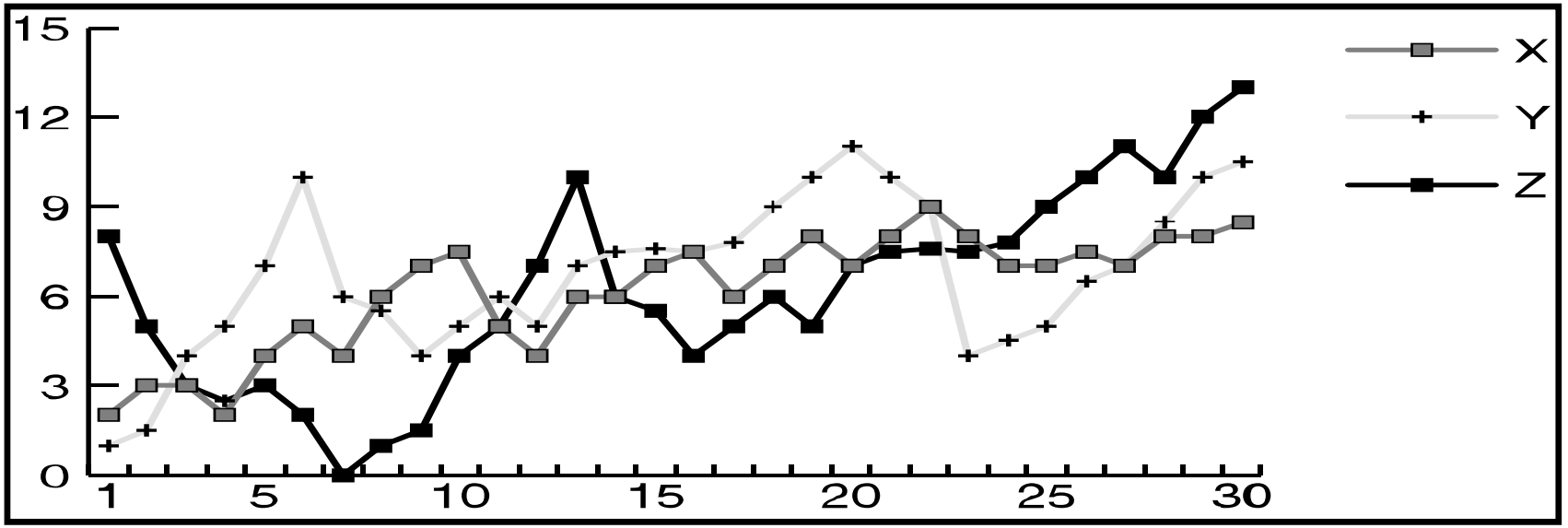
Some examples to*:*

- *Predictive tasks*
- *Association analysis*
- *Cluster analysis*
- *Anomaly detection*

Predictive modeling

Predictive modeling refers to the task of building a model for the target variable as a function of explanatory variables from past observations.

Examples: Predict future values in stock market, electricity consumption, …



Association analysis

Association analysis is used to discover patterns that describe strongly associated features in the data

For example finding the items that are frequently bought together by the customers

Based on the analysis of customer baskets the following rules may have been derived:

- {bread} $\rightarrow$ {butter} that suggests that customers who buy bread also tend to buy butter
- {dollar} $\rightarrow$ {industry} that suggests that newspaper articles that use word dollar also use word industry
- {dangling reference} $\rightarrow$ {memory leak} that suggests that software that has dangling reference error also has memory leaks

Cluster analysis

Cluster analysis finds the groups of closely related observations such that observations that belong to the same cluster are more similar to each other that belong to other cluster

For example grouping the newspaper articles based on their respective topics

Based on the analysis of word frequency pairs `(w,c)` where `w` is a word and `c` is the number of times a word appears in an article it is possible to identify the clusters of articles corresponding to economy and cluster of articles corresponding to health care.

Anomaly detection

Anomaly detection is a task of identifying observations whose characteristics are significantly different from the rest of the data

For example credit card fraud detection

Based on the analysis of legitimate credit card transactions a profile of legitimate transaction is built and when a new transaction arrives it is compared against the profile; if its characteristics are very different from the earlier created profile the transaction is flagged as suspected

Examples of application classes

- *Business/Analytics applications*
- *Medicine applications*
- *Science and Engineering applications*

Business/Analytics applications

- Customer relationship management
  - identify those who are likely to leave for a competitor.
- Targeted marketing.
  - identify likely responders to promotions
- Discovery of segments or groups within a customer data set
  - Identify characteristics of the most successful employees
- Market basket analysis
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production
  - automatically adjust knobs when process parameter changes

Medicine/Science/Engineering applications

- Understand the mapping relationship between the inter-individual variation in human DNA sequences
- Condition monitoring of high voltage electrical equipment
- Dissolved gas analysis on power transformers
- Analysis of factors leading students to choose to engage in behaviors which reduce their learning
- Discovering patterns associating drug prescriptions to medical diagnoses.
- Disease outcome, effectiveness of treatments. Analyze patient disease history: find relationship between diseases

…just to mention a few examples.

Data Mining methods

Some common data mining methods are:

- K-means
- DBscan
- Nearest neighbor classifier
- Neural Networks
- Support Vector Machines
- Decision Tree classifiers
- Bayes classifier
- Association rule mining
- …many more

Pros and cons

- **Nearest Neighbor Classifier:**

  - **Pros**
    + Fast training
    + Algorithm is simple and easy to understand

  - **Cons**
    – Slow during application.
    – No feature selection.
    – Notion of proximity vague
    – Sensitive to noise and outliers
    – Sensitive to imbalanced data

Pros and cons

- **Decision Trees:**

- **Pros**

+Reasonable training time

+ Fast application

+ Easy to interpret

+ Easy to implement

+ Can handle large number of features

- **Cons**
  - Cannot handle complicated relationship between features
  - Simple decision boundaries
  - Sensitive to missing data.
  - Sensitive to noise and outliers

Pros and cons

- **Neural Networks:**

- **Pros**

+Can learn complicated class boundaries

+ Fast application

+ Can handle large number of features

+No feature selection.

+Insensitive to noise and outliers.

+Exceptionally scalable on parallel computers.

- **Cons**
  - Slow training time
  - Hard to interpret
  - Hard to implement
  - Requires setting of parameters via trial and error.
  - Some NNs are sensitive to imbalanced data.
  - Can be sensitive to sparse data.

Pros and cons

- **Support Vector Machines:**

- **Pros**

+Can learn complicated class boundaries

+ Fast application

+ Can handle large number of features

+Simpler than neural networks

+Insensitive to imbalanced data.

- **Cons**
  - Reasonable training time but not scalable to very large data.
  - Hard to interpret
  - Can be sensitive to sparse data.

Data Mining methods

From these examples we find:

- There is no one-fits-all approach to data mining.
- Each method has its own list of pros and cons.
  - It is important to understand the strength and weaknesses of each method, and
  - It is important to understand the property of the data and to understand the task at hand.
- *This subject introduces to common data mining methods. Students will learn to identify suitable methods via exposure to data mining problems.*
- Experience will be the best teacher in this subject: Learn by doing.

The Data

The **first step** in data mining:

- Understand the property of the data!
- Know the quality of the data!
- Familiarize with the domain
  - Where does the data come from, how was the data collected, how is the data stored and how can it be accessed, is the data complete, …?

Input data

Input data can be described by:

- *Data sets*

  Data set is a collection of data objects (records, points, vectors, graphs, observations, etc)

- *Attributes and measurements*

  An attribute is a property or characteristics of an object that may vary either from one object to another or from one time to another

- *Attribute type*

  Attribute type is determined by the properties of its values that correspond to underlying properties of the attribute

Example:

Record

Attribute name

Attribute value

TABLE 9.3 Animal Names and Their Attributes

Animal		Dove	Hen	Duck	Goose	Owl	Hawk	Eagle	Fox	Dog	Wolf	Cat	Tiger	Lion	Horse	Zebra	Cow
is	small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
likes to	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Attribute types

- *Nominal*
  The values of nominal attribute are just different names (i.e. red, coke, Joe Smith)
- *Ordinal*
  The values of an ordinal attribute allow to order objects (i.e. 1st of Feb, 3rd prize)
- *Interval*
  The difference between interval attribute are meaningful (i.e. Age range: 10-30years)
- *Ratio*
  Both differences and ratios are meaningful for ratio attributes (i.e. 1/3)

Other classification of attributes

- *Discrete attribute*

  A discrete attribute has a finite or countably infinite set of values, e.g. zip codes, ID numbers, dates, colours, standard sizes, etc

  A binary attribute is a special case of discrete attribute

- *Continuous attribute*

  A continuous attribute is one whose values are real numbers, e.g. temperature, weight, height, speed, etc

- *Asymmetric attribute*

  An asymmetric attribute is one whose existence is regarded as important

Characteristics of data sets

- *Dimensionality*

  The dimensionality of a data set it the number of attributes that the objects in the data set posses.

- *Sparsity*

  The sparsity of a data set means frequency of attribute appearances in the descriptions of the objects.

- *Resolution*

  The resolution of a data set means an average "distance" between the measurement of the attributes of the data objects.

Types of data sets:

- *Record data*

  No explicit relationship among record or data fields, every record has the same set of attributes

- *Transaction data*

  A set of records where each record involves a set of items

- *Data matrix*

  All records have fixed set of numeric attributes, data objects can be considered as "points" in a multidimensional space where each dimension represents a distinct attribute describing the object

- *Sparse data matrix*

  A data matrix with missing or unavailable elements

Types of data sets: Data ordering

- *Sequential data (temporal data)*

  Extension of record data where each record has a time moment associated with it

- *Sequence data*

  Data set that is a sequence of individual entities, such as a sequence of words or letters

- *Time series data*

  Special type of sequential data in which each record is a time series, i.e., a series series of measurements taken over time

- *Spatial data*

  Record data that have spatial attributes such as positions or areas and other types of attributes

Types of data sets: Graph-based data

- *Data with relationships among objects*

  Relationships among the objects convey important information, the data is represented as a graph

- *Data with objects that are graphs*

  If objects have internal structure then the objects contain sub-objects that have relationships among them

- *Data with objects that are graphs and have relationships amongst objects*

  Graph-of-Graphs

Data quality

- *Measurement and data collection errors*

  Measurement error happens when a value recorded differs from the true value.

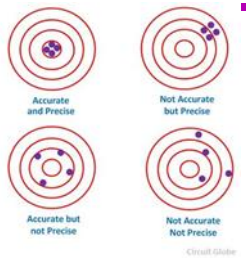
  Data collection error refers to omitting data objects or attributes, or inappropriately including a data object.

- *Noise and artifacts*

  Noise is a random component of a measurement error, it distorts a value or it adds spurious objects

  Artifact is a deterministic distortion of data

Data quality



- *Precision, bias, and accuracy*

  Precision means the closeness of repeated measurements to one another

  Bias means a systematic variation of measurements from the quantity being measured

  Accuracy means the closeness of measurements of the true value of the quantity being measured.

- *Outliers*

  Outliers are either data objects that have characteristics different from the most of other data objects in the data set or …

  … or values of attributes that are unusual with respect to the typical values for that attribute

Data quality

- *Missing values*

  Missing values means that one or more attribute values are not available in a date objects

  Values can be missing because information was not collected, some attributes are not applicable, its presence depends on presence of other values etc

- *Inconsistent values*

  Inconsistent values are values that violate given consistency constraints

- *Duplicate data*

  Duplicate data are data objects that are duplicates or almost duplicates of each other

Data Preprocessing

Step 2 in data mining:

- Data preprocessing and
- Data transformation

Data preprocesing

- *Aggregation*

  Aggregation combines two or more objects into a single object

- *Sampling*

  Sampling selects a subset of the data objects to be analyzed

- *Dimensionality reduction*

  Dimensionality reduction reduces the total number of attributes describing an object creating new that are combinations of the old ones

- *Feature subset selection*

  Feature subset selection reduce the total number of attributes by eliminating nonimportant attributes

Data preprocesing

- *Feature creation*

  Feature creation means creating a new set of attributes from the original one

- *Discretization and binarization*

  Discretization is a transformation of a continuous attribute into a categorical attribute, binarization is a transformation of both continuous and discrete atributes into binary attributes

- *Variable transformation*

  Variable transformation refers to a transformation that is applied to all the values of a variable (attribute)

Data integration

- Data integration means merging data from multiple data sources into a coherent data store

- *Schema integration and object matching*

  Schema integration means matching real world entities in to a common schema

  Object matching means matching identical real world objects that have a bit different descriptions

- *Elimination of redundancies*

  Elimination of redundancies means finding the attributes whose values can be derived from other attributes,e.g. through correlation analysis

Data integration

- *Detection and resolution of data value conflicts*

  Detection and resolution of data value conflicts means identification and elimination of all cases when for the same real world entity the values of the same attributes from different sources may differ

Data Mining

Step 3 in data mining:

- Mining!

But what methods to use?

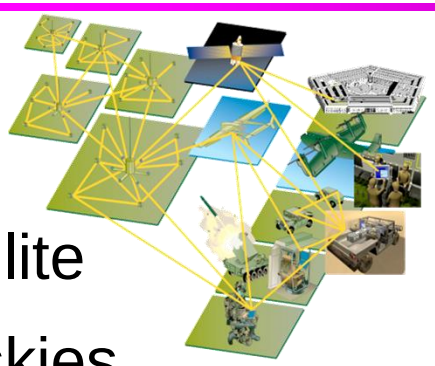
- Depends on the data (covered in step 1 and 2)
- Depends on the task, domain, and problem.
- Lets have a look at some examples.

Why Mine Data? Commercial Viewpoint

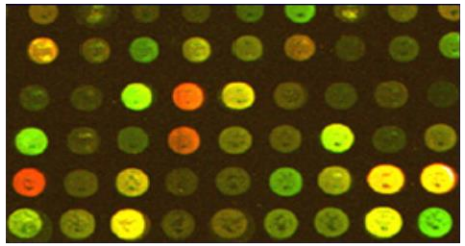
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

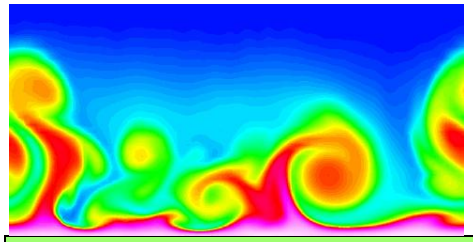
- Computers and data storage have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

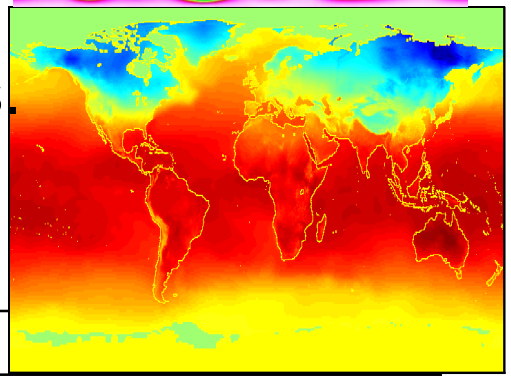
Why Mine Data? Scientific Viewpoint



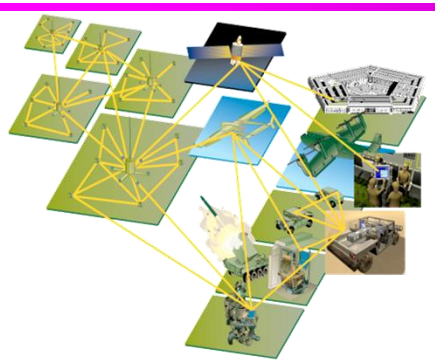
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
  - Video surveillance, traffic control systems.
  - Data generated on social networking sites.
  - …and many more.



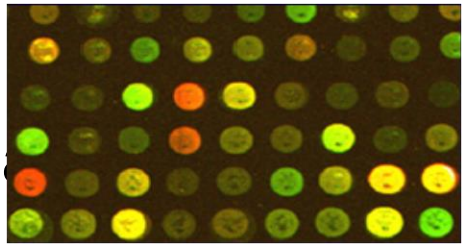


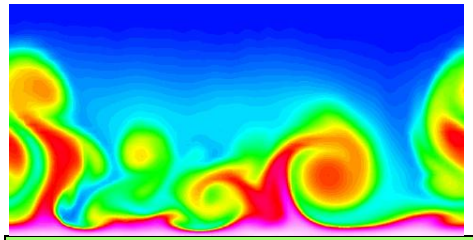


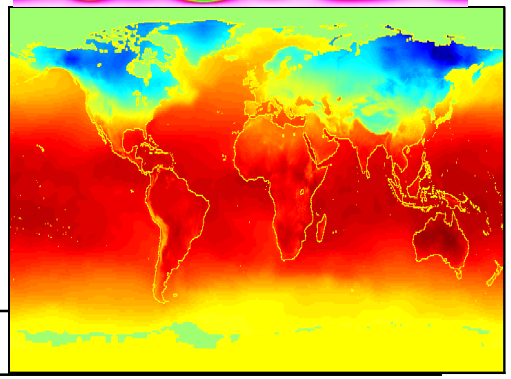
Why Mine Data? Scientific Viewpoint



- Traditional data processing techniques are infeasible for large amounts of raw data.
- Processing in real-time is often necessary.
- Scalability is a main concern in the design of a data mining algorithm.
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation.
  - in finding new insights.





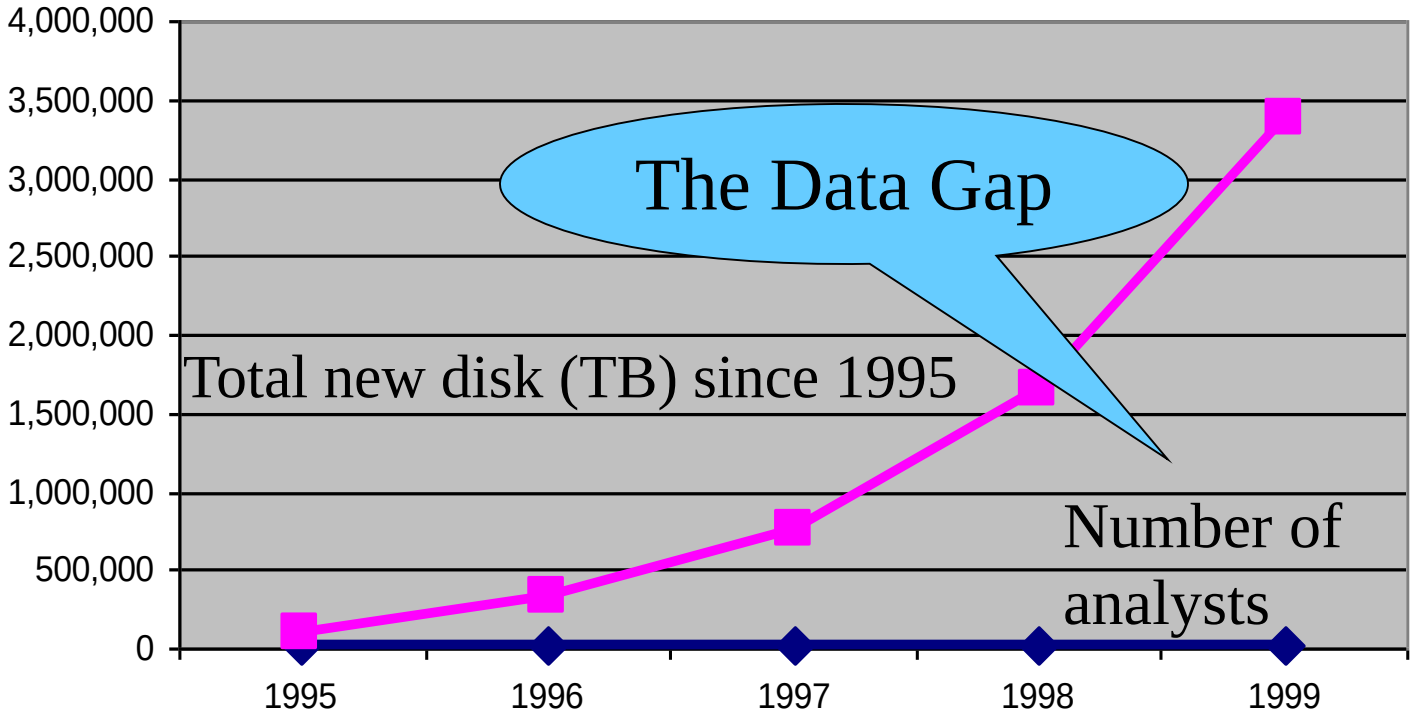


Examples:

- Data collections (Terrabyte TB, Petabytes PT)
  - Google map (http://maps.google.com/)
  - Google sky (http://www.google.com/sky/)
  - National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/)
  - National Aeronautics and Space Administration (NASA) (http://www.nasa.gov/)
  - CANTATA video surveillance (http://www.hitech-projects.com/euprojects/cantata/datasets_cantata/dataset.html)
  - Wikipedia,INEX,…and many others
- 1024byte = 1K; 1024K = 1M; 1024M = 1G; 1024G = 1T

Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident.
- Human analysts may take weeks to discover useful information.
- Much of the data is never analyzed at all.



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

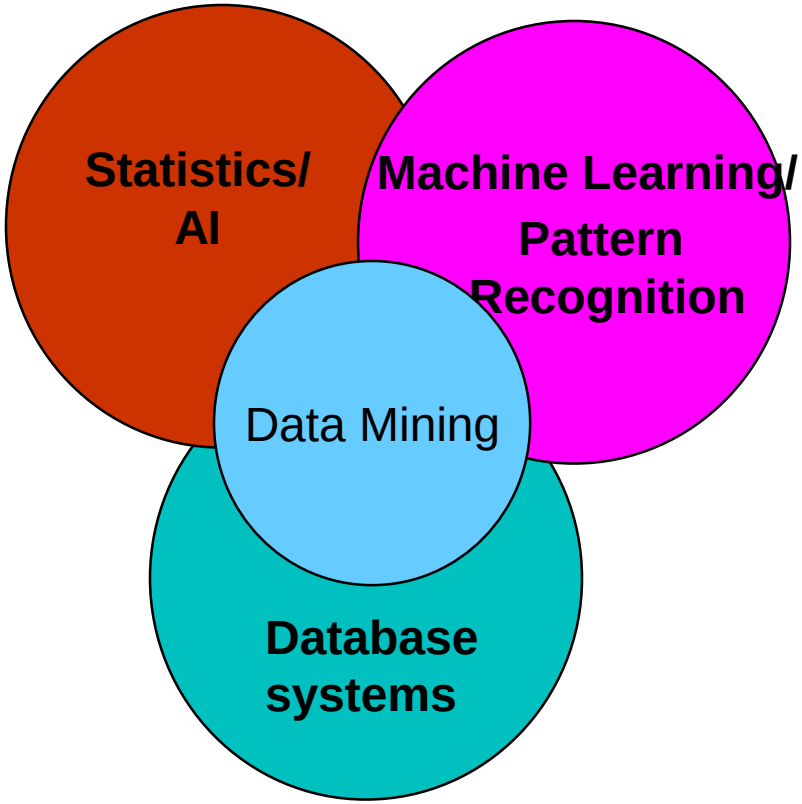
What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon" (Information Retrieval)

- What is Data Mining?
  - Recommend a book to the potential users based on their relatives' and friends' preference. (Collaborative filtering)
  - Group together similar documents returned by search engine according to their context (e.g. amazon.com) (Document clustering)

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



As a summary, what is data mining?

- Algorithms
  - They are soft strategies for processing data.
- Large volume of data
  - Data mining does not focus on small datasets.
- Knowledge discovery
  - Through data mining, we expect to find new knowledge to support or denyl our hypothesis.
  - We expect to find “hidden” relations under these data.

Current Data Mining

- Conferences:
  - KDD: http://www.kdd.org/kdd2017/
  - ICDM: http://icdm2017.rutgers.edu/
  - SDM: http://www.siam.org/meetings/sdm17/
- Journals:
  - IEEE Transactions on Knowledge and Data Engineering (TKDE): http://www2.computer.org/portal/web/tkde/about;jsessionid=E19962782EA80EA8661DFFB342EE98E4
  - Data mining and knowledge discovery: http://www.springerlink.com/content/100254/

Data Mining Tasks

- Classification
- Clustering
- Sequential Pattern Discovery
- Regression
- Deviation Detection

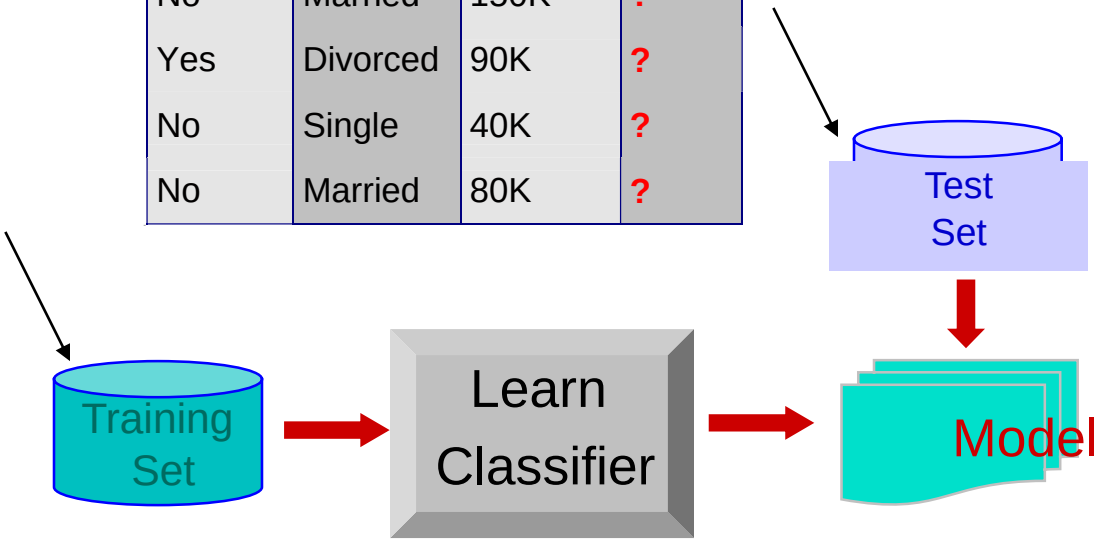
Classification: Definition

- 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: previously unseen 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
		categorical	categorical	continuous	class
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 smart-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

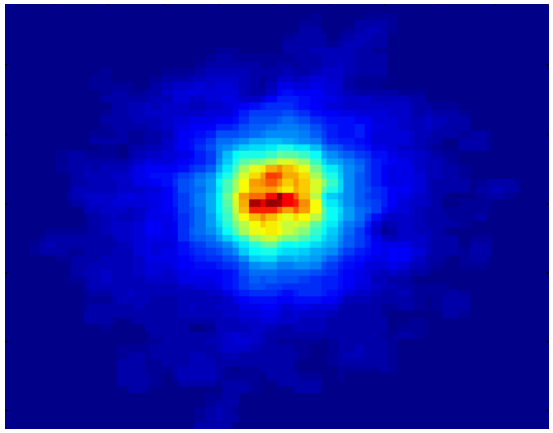
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes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

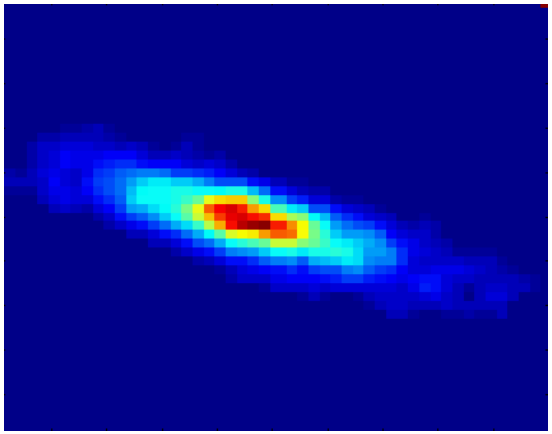
Early



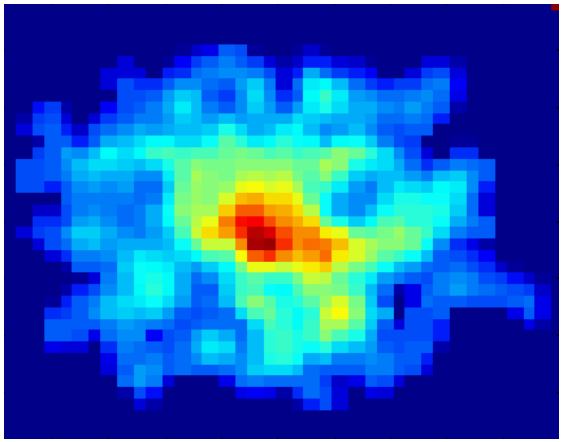
Class:
- Stages of Formation

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

Clustering Definition

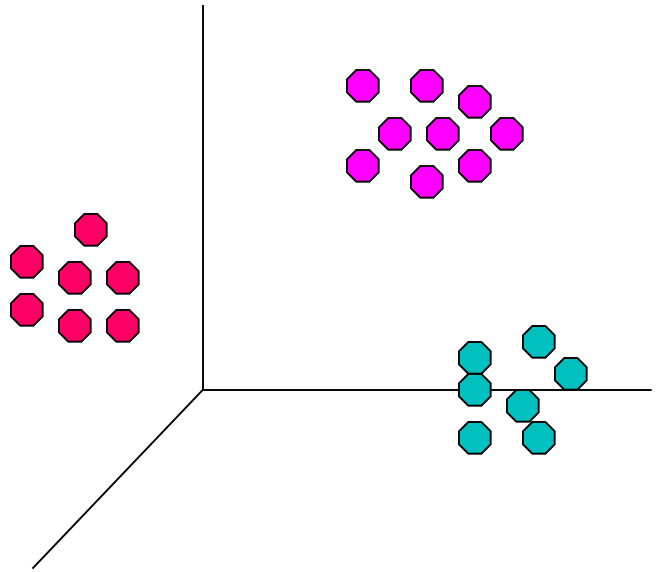
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance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



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

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

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

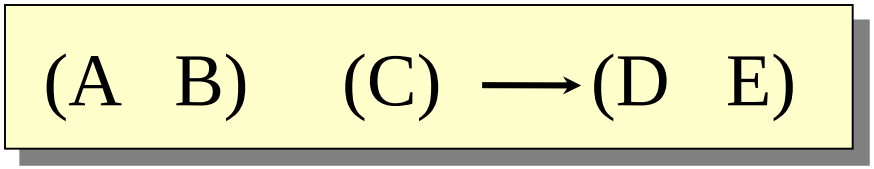
    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips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

Association Rule Discovery: Application 2

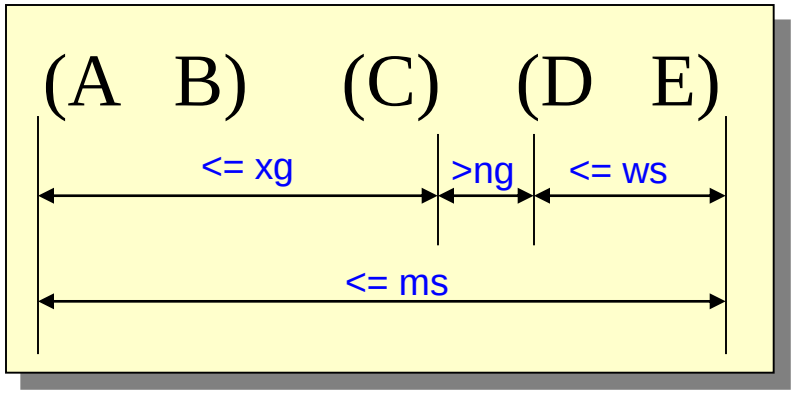
- Supermarket shelf management.
  - **Goal:** To identify items that are bought together by sufficiently many customers.
  - **Approach:** Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then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

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.**



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



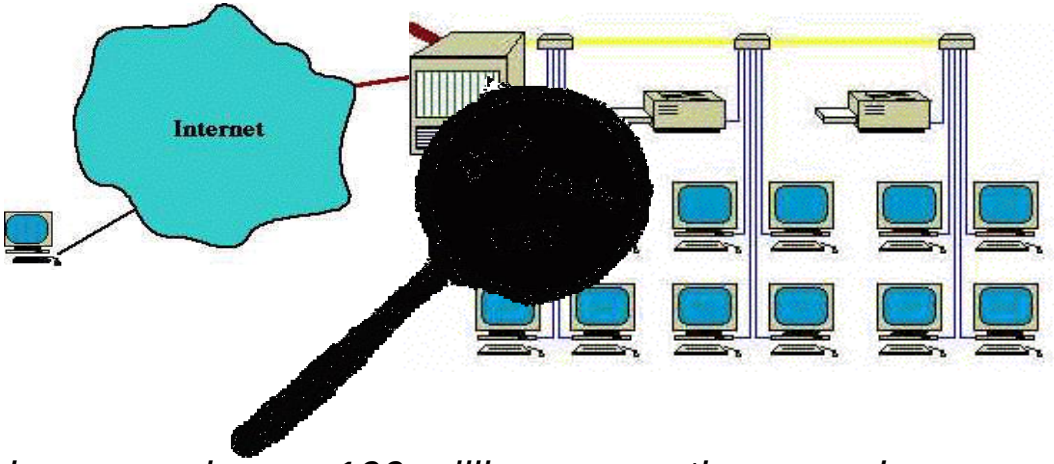
Regression

- 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.

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
- Domain Knowledge
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

References

Pang-Ning Tan and Micheal Steinbach and Vipin Kumar, *Introduction to Data Mining*, Pearson, Addison Wesley, 2006, ISBN 0-321-32136-7
chapters 1, 2.1, 2.2., 2.3

Jiawei Han and Micheline Kamber, *Data Mining* Concepts and 1-55860-901-6
chapters 1.1, 1.2, 1.4, 2.4