"DATA AND WEB MINING"

Introduction

Salvatore Orlando

The slides of this course were partly taken up by tutorials and courses available on the web. More specifically:

Vipin Kumar, Data mining course at University of Minnesota

Jiawei Han, slides of the book *Data mining: concepts and techniques*

Li Yang, Data mining course at Western Michigan University

Giannotti/Pedreschi, PhD course on Data mining at University of Pisa

Goal of the course

- **This course provides the motivation and the fundamentals of data mining (DM)**
- **Analyze with in some detail the main techniques of DM**
- **Use the Web as a case study, and the opportunity to extract useful knowledge from the mining analysis of the hyperlink structure of the Web, as well as from content and usage logs.**

General information about the couse

- **Web site:**
  - Register in moodle.unive.it (**Data and Web Mining [CM0226]**)
  - Website: http://www.dsi.unive.it/~dm
- **The exam is subdivided into two parts:**
  - Written exam (60%)
  - Public presentation of a scholar paper (40%)
- **Text, Reading list, Didactic Material**
  - Slides
  - **P.-N. Tan, M. Steinbach, V. Kumar. *Introduction to Data Mining*. Pearson Addison-Wesley.**
  - **J. Han, M. Kamber. *Data mining: concepts and techniques.* Morgan Kaufmann.**
  - **M. H. Dunham. *Data Mining: Introductory and Advanced Topics.* Prentice Hall.**
  - **Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag, 2006.**
  - **C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.**

Course contents

- Introduction to data mining

- Main techniques
  - Associative Rules and temporal sequences
  - Classification and prediction
  - Clustering
  - ….

- Web Mining
  - Web Content Mining (Web Search)
  - Web Structure Mining (Link Analysys)
  - Web Usage Mining (Recommender Systems)

Wealth of data/information
Lack of knowledge

- **The databases are more and more large**
  - **Terrorbytes!**
  - **A deluge of data, containing a lot of hidden information**
    ⇒ new knowledge

- **What are the technological motivations?**
  - **Technologies to collect data**
    - **Bar code readers, scanners, cameras, etc..**
  - **Technologies to store data**
    - **Databases, data warehouses, other repositories**
  - **Network (Web) as computing and storage platform**

- **An example of data deluge:**
  - ***the WEB and SOCIAL MEDIA !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!***

From the dawn of civilization until 2003, humankind generated five exabytes (1000^6) of data. Now we produce five exabytes every two days…and the pace is accelerating.

Eric Schmidt,
Executive Chairman, Google

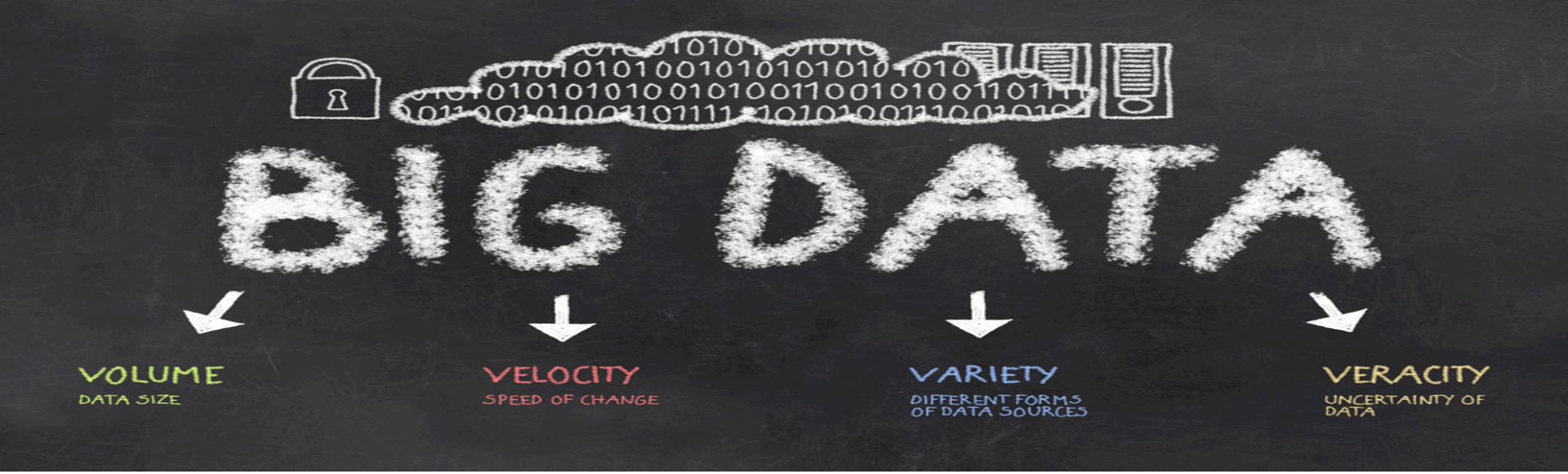
Big Data is fuelled by two things:

1. The increasing 'datafication' of the world, which means we generate new data at frightening rates.
2. Our increasing ability to harness and analyse large and complex sets of data

- Activity Data: Simple activities like listening to music or reading a book are now generating data. Digital music players and eBooks collect data on our activities. Your smart phone collects data on how you use it and your web browser collects information on what you are searching for. Your credit card company collects data on where you shop and your shop collects data on what you buy. It is hard to imagine any activity that does not generate data.

- Conversation Data: Our conversations are now digitally recorded. It all started with emails but nowadays most of our conversations leave a digital trail. Just think of all the conversations we have on social media sites like Facebook or Twitter. Even many of our phone conversations are now digitally recorded.

- Photo and Video Image Data: Just think about all the pictures we take on our smart phones or digital cameras. We upload and share 100s of thousands of them on social media sites every second. The increasing amounts of CCTV cameras take video images and every minute we up-load hundreds of hours of video images to YouTube and other sites.
- Sensor Data: We are increasingly surrounded by sensors that collect and share data. Take your smart phone, it contains a global positioning sensor to track exactly where you are every second of the day, it includes an accelometer to track the speed and direction at which you are travelling. We now have sensors in many devices and products.
- The Internet of Things Data: We now have smart TVs that are able to collect and process data, we have smart watches, smart fridges, and smart alarms. The Internet of Things, or Internet of Everything connects these devices so that the traffic sensors on the road send data to your alarm clock which will wake you up earlier than planned because the blocked road means you have to leave earlier to make your 9am meeting...



With the datafication comes big data, which is often described using the four V's: Volume, Velocity, Variety and Veracity

Volume refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. New big data tools use distributed systems so that we can store and analyse data across databases that are dotted around anywhere in the world.

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Variety refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.) With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

Veracity refers to the messiness or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

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

- Competitive Pressure is Strong
  - Use Data Mining to provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

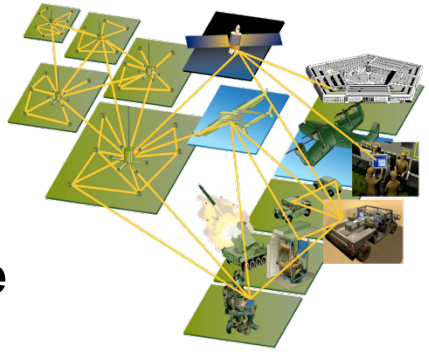
Competing on Analytics

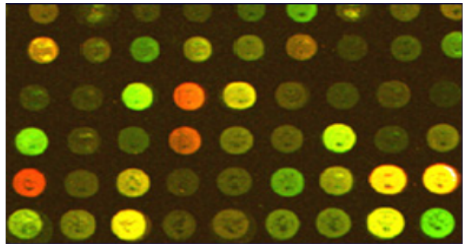
- *"Some companies have built their very businesses on their ability to collect, analyze, and act on data. Every company can learn from what these firms do."*
  Competing on Analytics, by Thomas H. Davenport
- *"Although numerous organizations are embracing analytics, only a handful have achieved this level of proficiency. But analytics competitors are the leaders in their varied fields—consumer products, finance, retail, and travel and entertainment among them."* *Competing on Analytics, by Thomas H. Davenport*
- Business intelligence cited as the top technology priority for CIOs in 2006 (surpassing security) - Gartner 2006
- Sizable market - 11.5% growth in 2005 for a market size of $5.7 billion in worldwide software revenue - IDC 2006
- "Organizations are moving beyond query and reporting" - IDC 2006

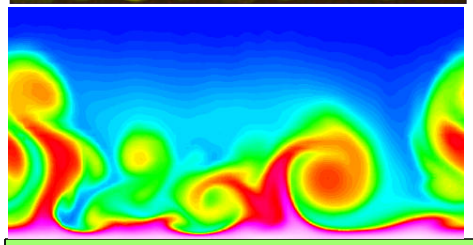
ORACLE

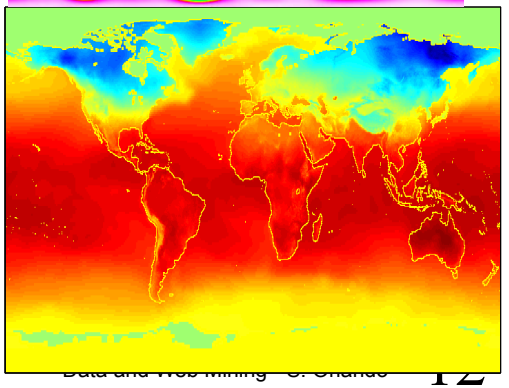
Why Mine Data? Scientific Viewpoint

- 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 (Many Definitions)**
  - **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**

- **Data mining: a misnomer?**
  - **It should be pattern mining in analogy to gold mining**

- **Alternative names:**
  - **Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.**

What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for page containing the term "Amazon"
- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine

Business Intelligence & Analytics

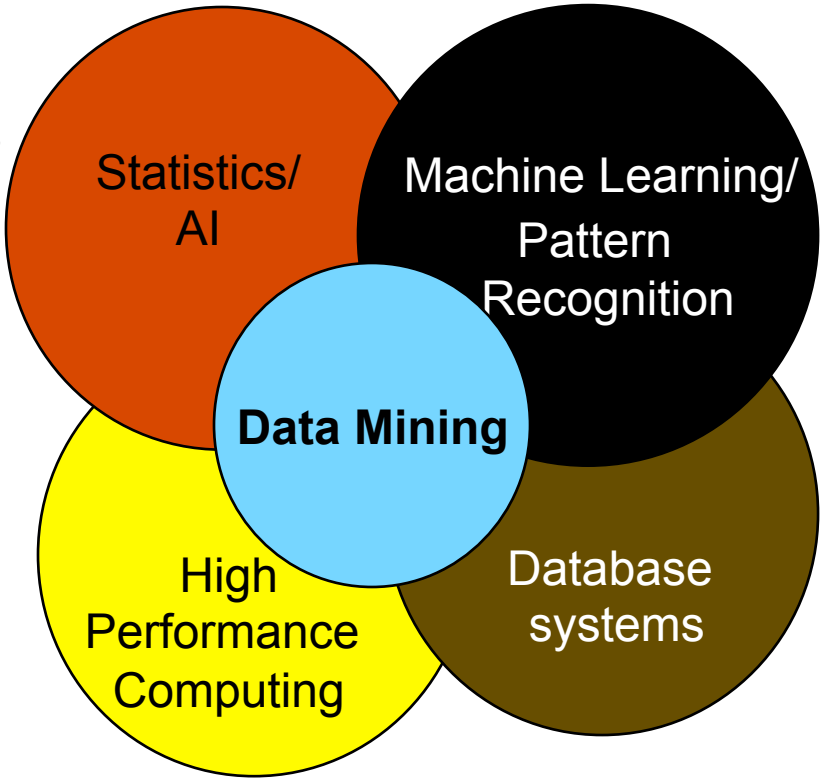
ORACLE DATABASE 11g

Query and Reporting	OLAP	Data Mining
Extraction of detailed and roll up data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns
"Information"	*"Analysis"*	*"Insight & Prediction"*
Who purchased mutual funds in the last 3 years?	**What is the** average **income** of mutual fund buyers, by region, by year?	**Who will buy** a mutual fund in the next 6 months and why?

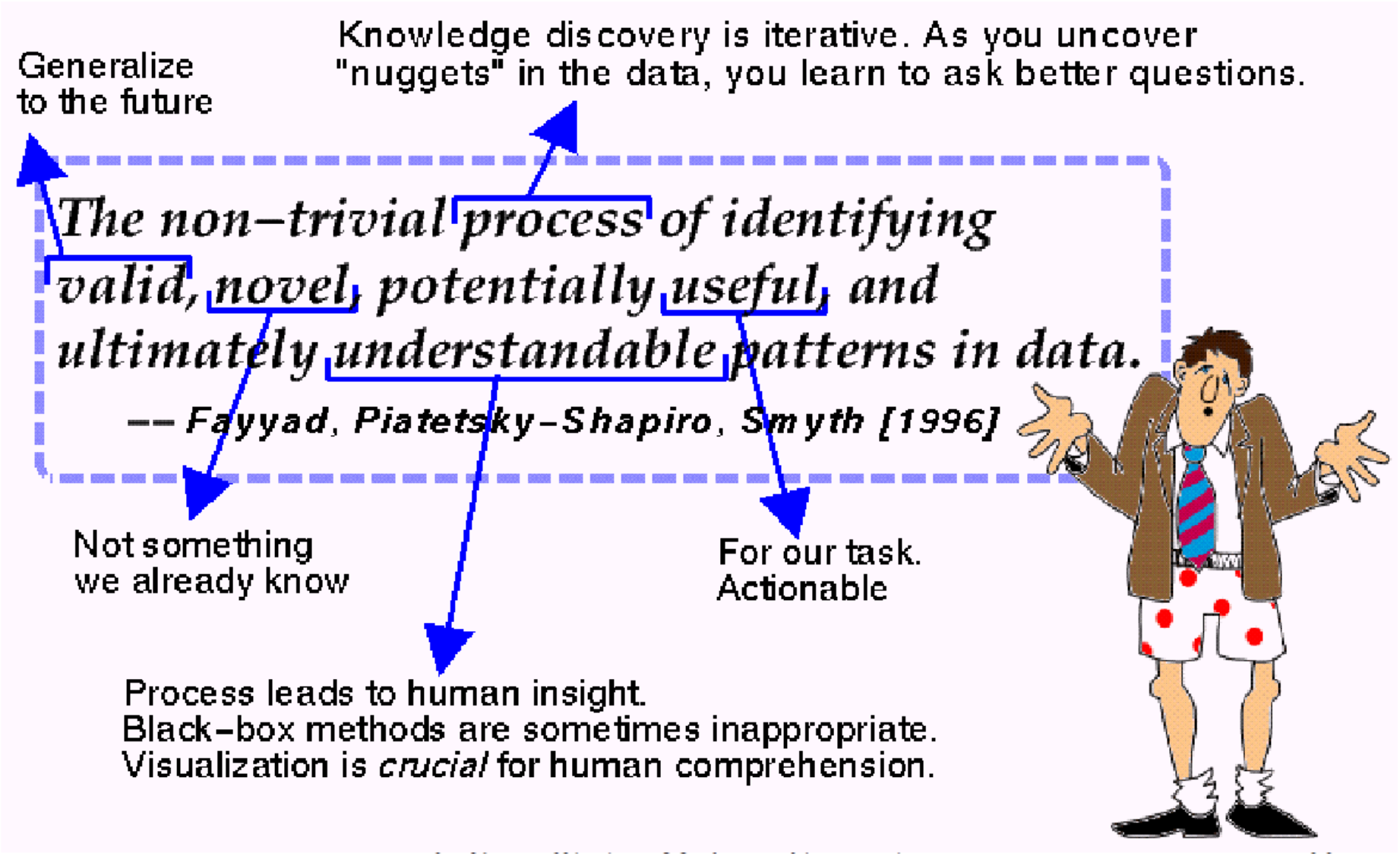
ORACLE

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, database systems, HPC**
- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**

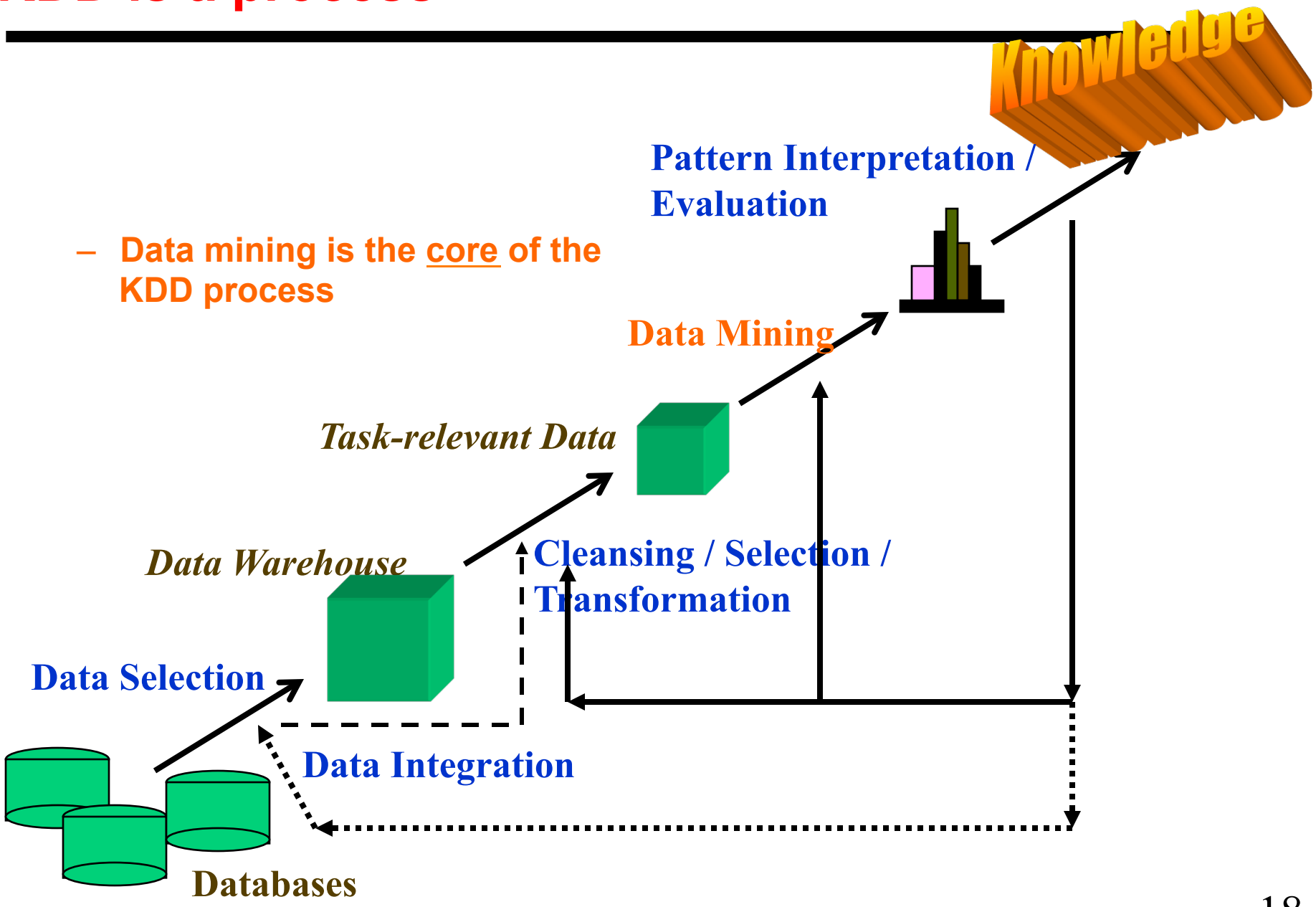


Knowledge Discovery in Database (KDD)



KDD is a process

– Data mining is the core of the KDD process



Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

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.**

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

Data Mining Tasks...

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

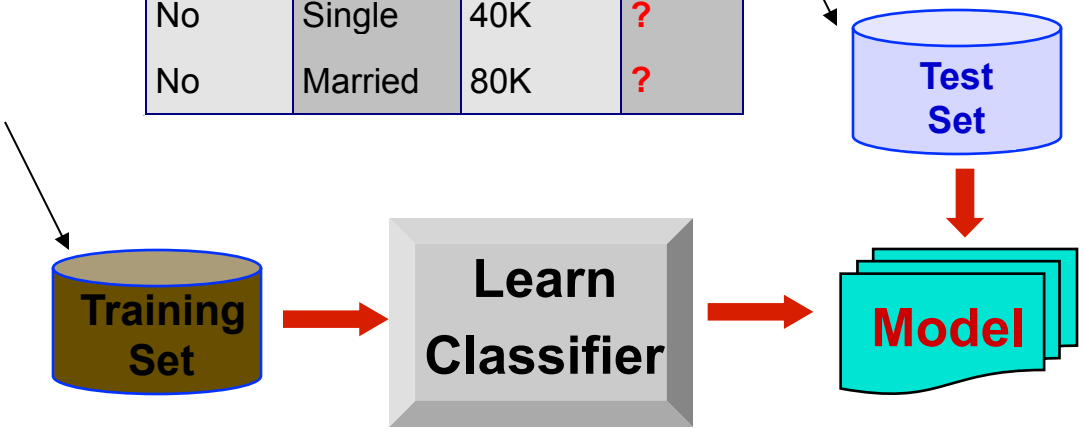
Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*
  - One of the (discrete) attributes is the *class*.
- Find a *model* for the 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

	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	?



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 the customers who decided to buy, and the ones who 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 of a user 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

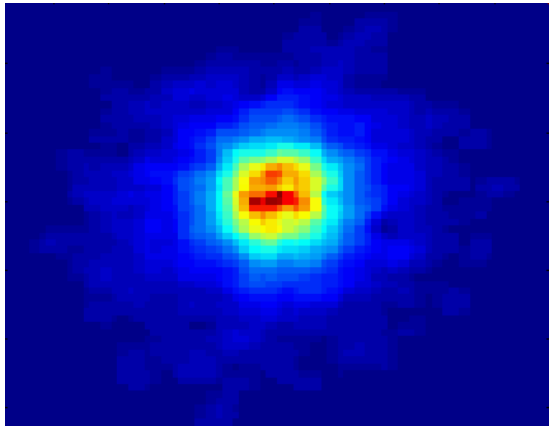
- **Sky Survey Cataloging**
  - **Goal:**
    - **To predict class (star or galaxy) of sky objects, especially visually perceptible, based on the telescopic survey images (from Palomar Observatory).**
      - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach:**
    - **Segment the image.**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

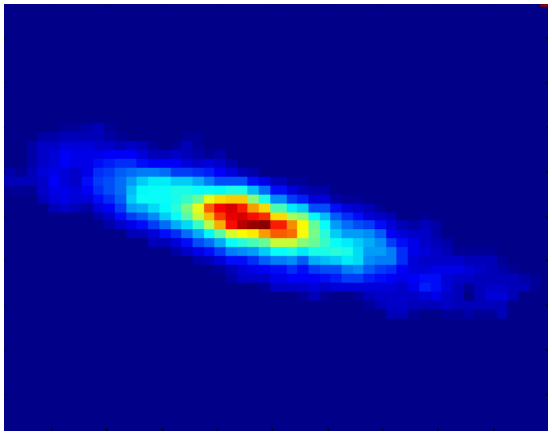
Early



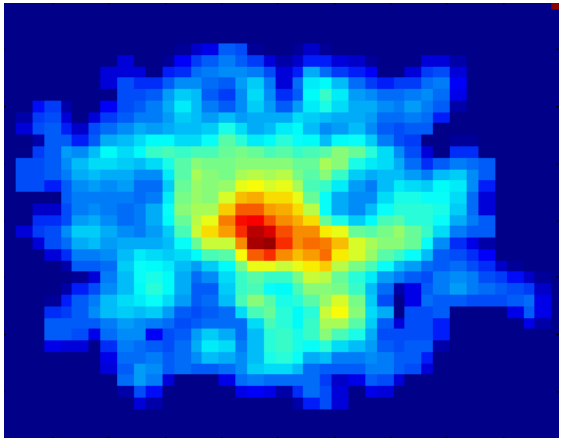
Class:
- **Stages of Formation**

Attributes/Features:
- **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

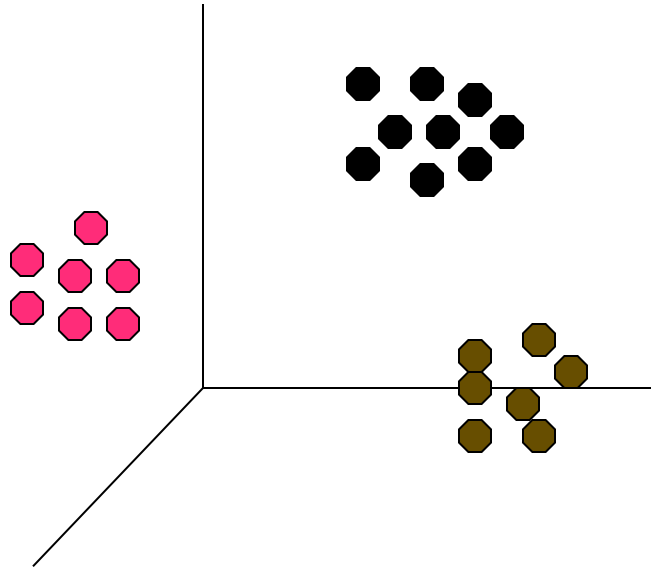
- **Given a set of "data points", each having a set of attributes, and a similarity measure among them, find clusters such that**
  - **Intracluster: Data points in one cluster are more similar to one another.**
  - **Intercluster: Data points in separate clusters are less similar to one another.**
- **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 could 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 wrt these attributes.**
    - **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.**
  - **A possible approach:**
    - **Identify frequently occurring terms in each document (to filter out some terms/dimensions).**
      Form a similarity measure based on the frequencies of these terms.
      Use it to cluster.
  - **Gain:**
    - **Information Retrieval can relate a new document or search term to clusters of similar documents.**

Illustrating Document Clustering

- **Clustering Points: 3204 Articles of Los Angeles Times.**
- **Similarity Measure: Consider 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} associated with a list of days
Similarity Measure: Two points are similar if the events described by them frequently happen together on the same day.
(frequent sets are used 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

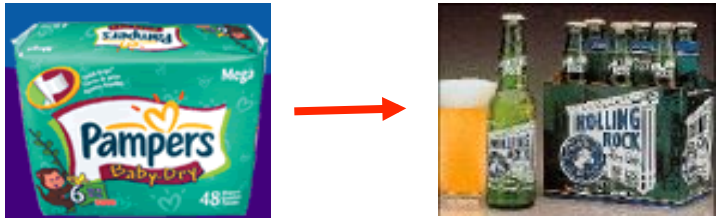
- Given a set of records, each of which contains some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item (or more items) based on occurrences of other items.

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 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! ☺**



Association Rule Discovery: Application 1

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

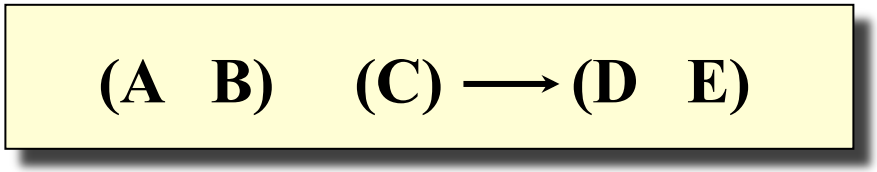
    {Bagels, … } --> {Potato Chips}
  - **Potato Chips as consequent => Can be used to determine what should be done to boost its sales.**
  - **Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.**
  - **Bagels in antecedent *and* Potato 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 3

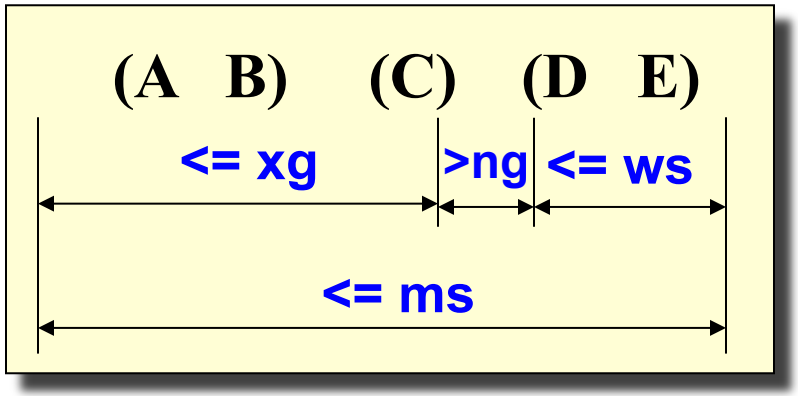
- **Inventory Management:**
  - **Goal:**
    - **A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products, and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.**
  - **Approach:**
    - **Process the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

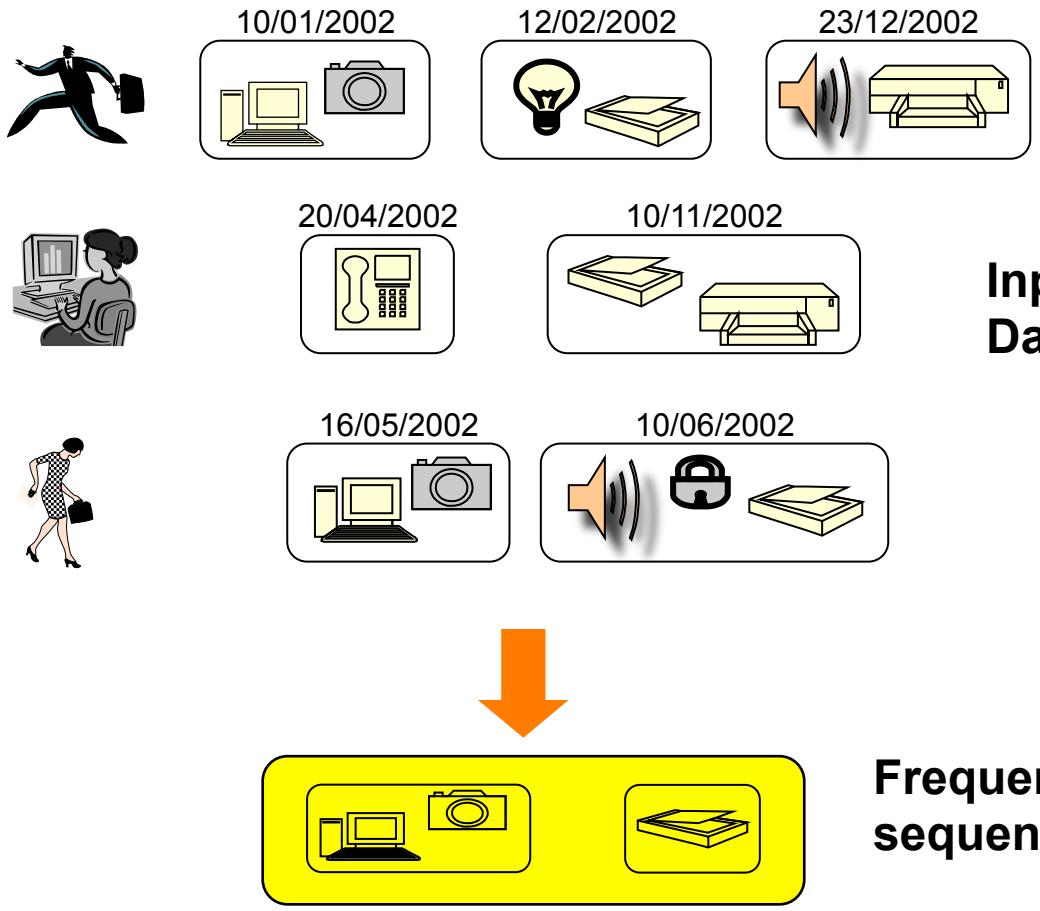


- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential Pattern Discovery: Example

- **Purchases**



Sequential Pattern Discovery: Examples

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) ----> (Fire_Alarm)

- 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

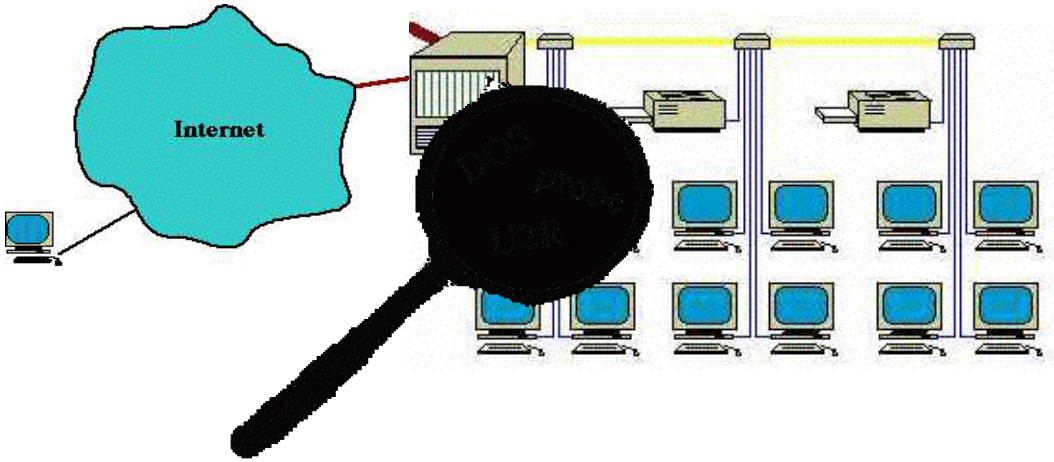
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.

- Examples:
  - 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.

Outlier/Anomaly Detection

- Detect significant deviations from normal behavior

- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection

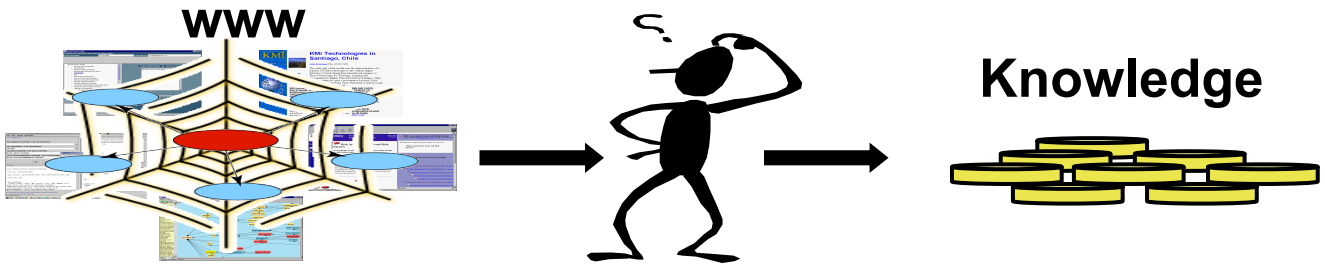


Challenges of Data Mining

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

Web mining

- **Web Mining applies DM to WWW**



- **Data Mining**
  - **Often applied to structured database**
- **Web mining**
  - **Applied to less structured data, dynamic, of huge size**
  - **Not only Web content, but also hyperlinks and access log**

Web mining taxonomy