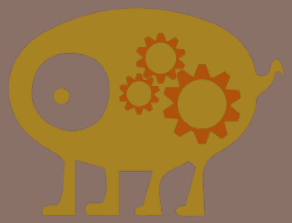


Data Mining
資料探勘

Introduction

Hung-Yu Kao, Fall 2017

An example

What kind of Data?

Which is useful in your Data?

Which is useful in your Data under some problem definition?

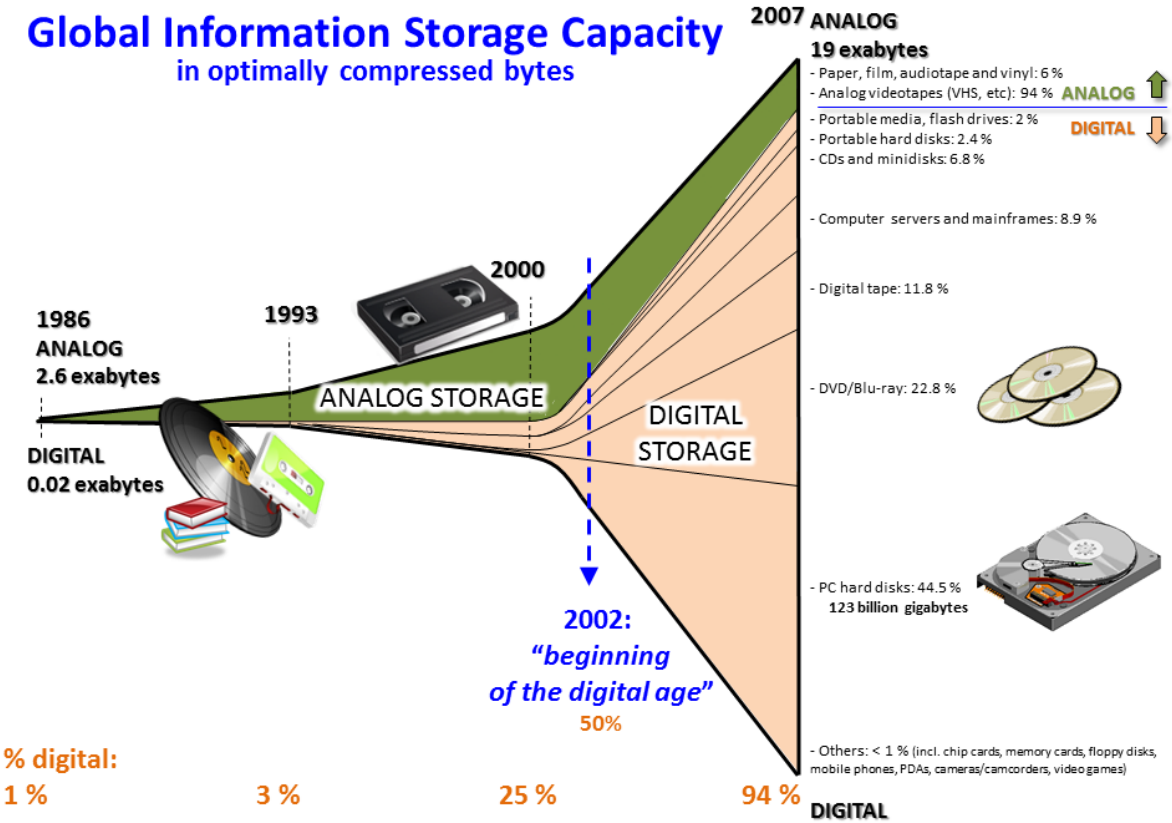
Which is useful in your Data under some problem definition and some evaluation criteria?

How about “Data only”?

- Time to take
- Cities to visit?
- Trace pattern?
- Travel purpose?
- Traffic suggestion?
- Pokemon hunting path?

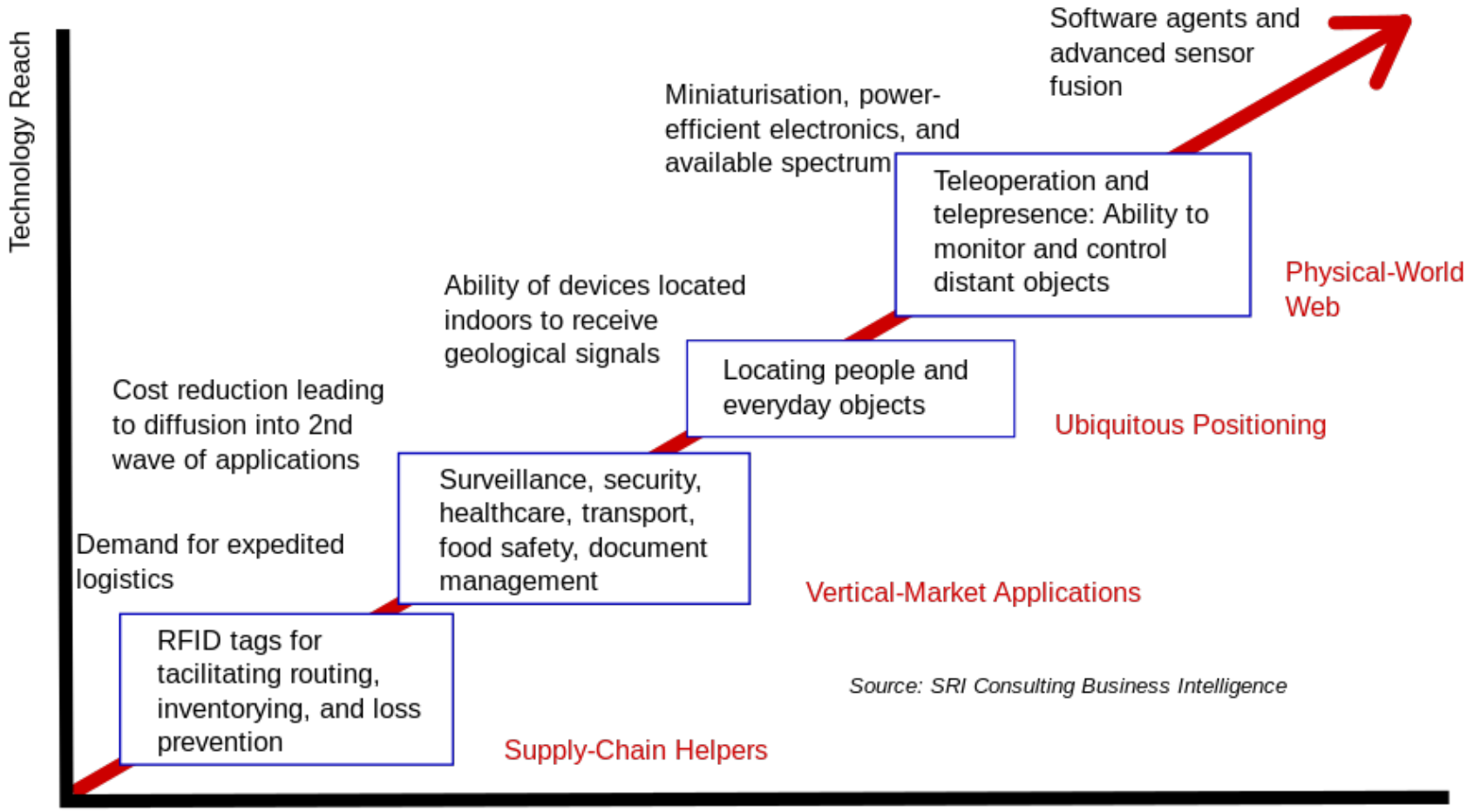
We are data rich

Driving force – Digital Storage



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

More data in the coming years
Everything is on-line

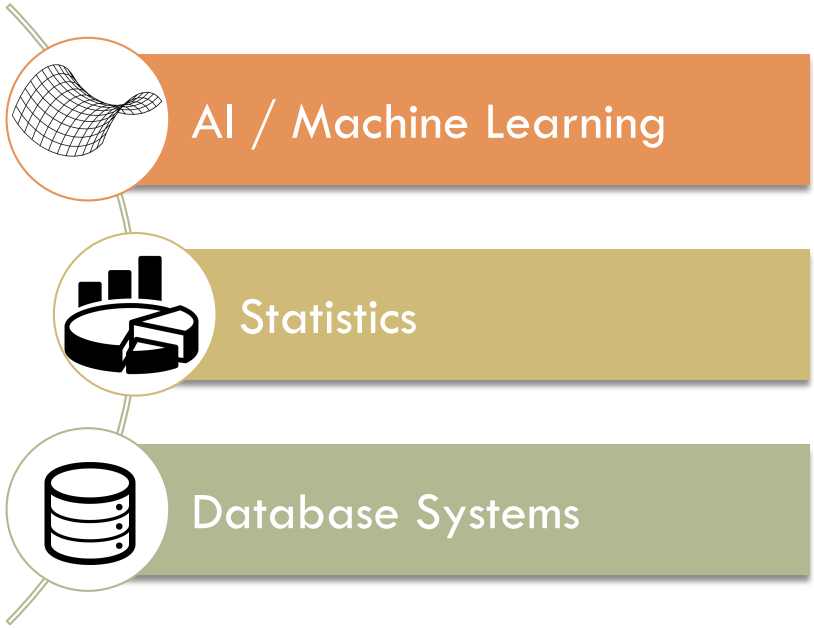


Data Mining

- Interdisciplinary subfield of Computer Science
- the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining).
- Data mining: Knowledge discovery in databases

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



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 "NCKU"

- What is Data Mining?
  - Certain names are more prevalent/popular in certain US locations (O' Brien, O' Rurke, O' Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. NCKU, 鄭成功)

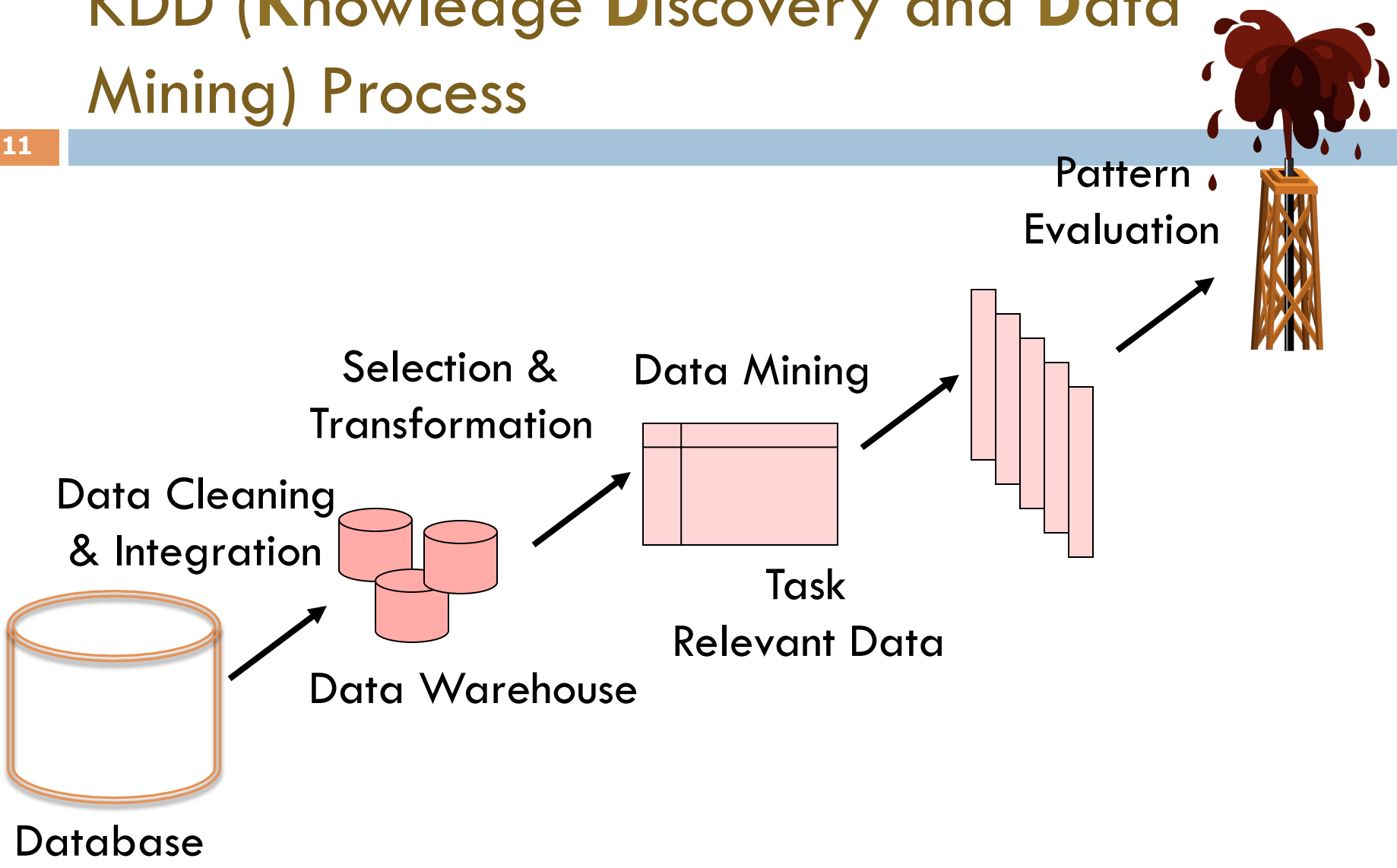
Data Mining Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
    - "Whoscall"

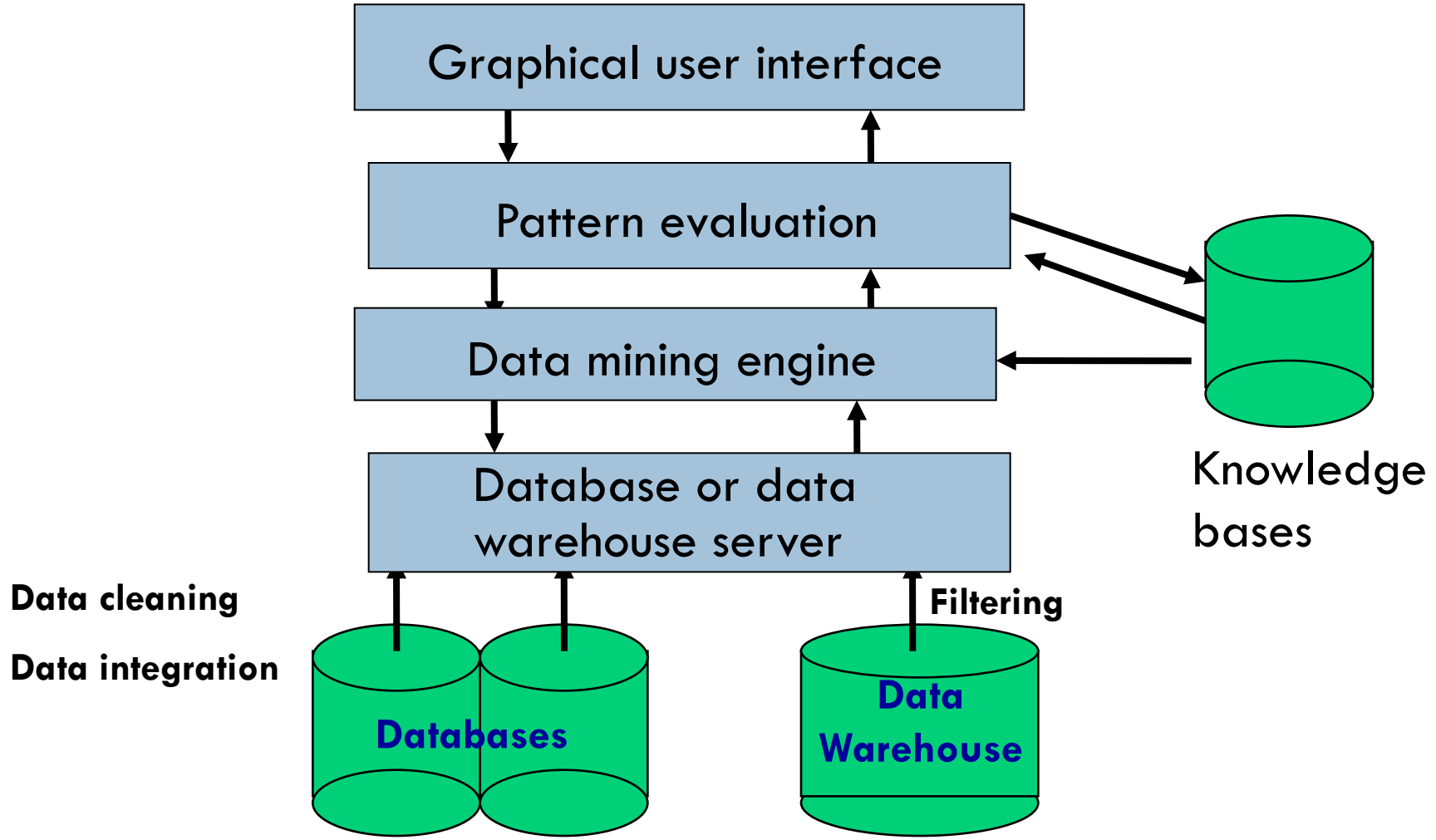
Knowledge Discovery from Databases

- Nontrivial process of extraction of
  - valid (with some degree of certainty)
  - novel (implicit, previously unknown)
  - potential useful
  - ultimately understandable
  - patterns from large collection of data
- Pattern
  - expression in languages describing subset of data
  - model (structure) applicable to subset of data

KDD (Knowledge Discovery and Data Mining) Process



Architecture: Typical Data Mining System



Classification of Data Mining Techniques

- What kinds of databases to work on
- What kind of knowledge to be mined
- What kind of techniques to be utilized

Databases to Work on

- Relational
- Transactional
- Spatial
- Time series data
- Multimedia
- Unstructured text
- Graph

Knowledge to Be Mined

- Association rules
  - Buy(bread) ^ Buy(milk) => Buy(butter)
  - Age(20~29) ^ Income(20~30k) => Buy(CD player)
- Classification
- Clustering
- Time series data analysis
- Semantics

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.

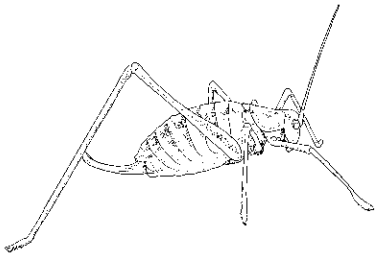
From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

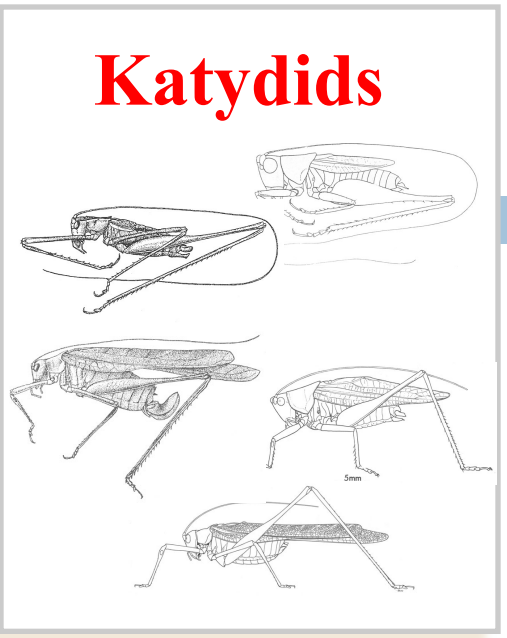
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification example

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



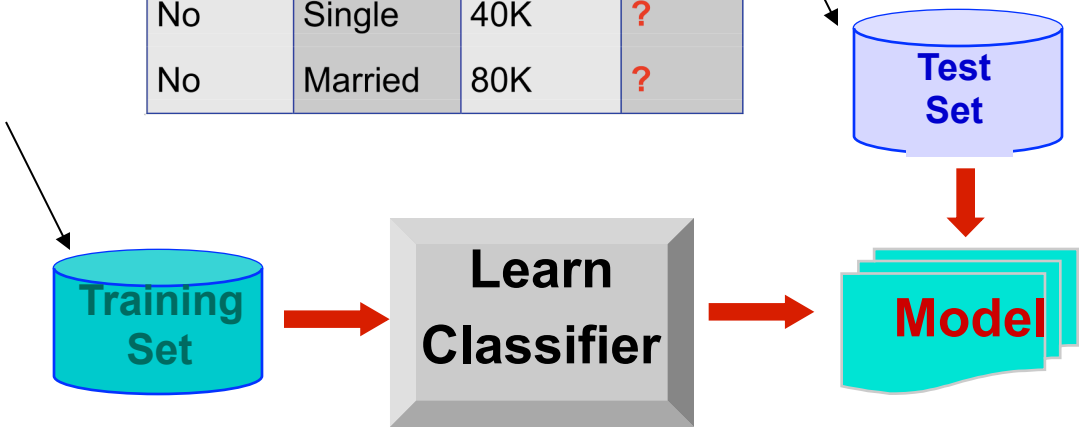
Katydid or **Grasshopper**?



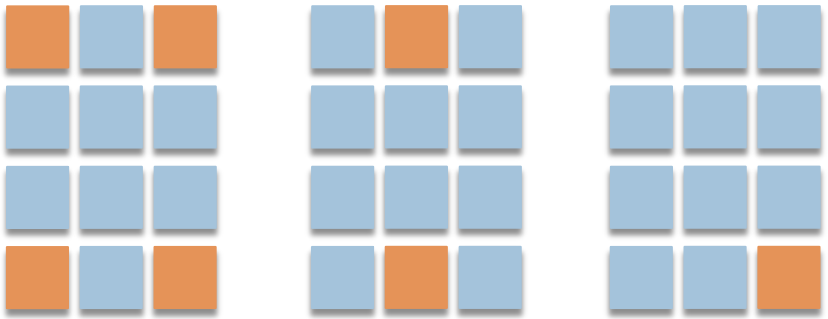
Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (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

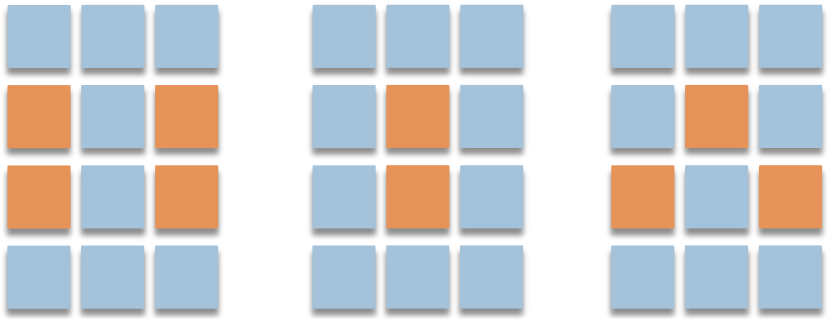
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



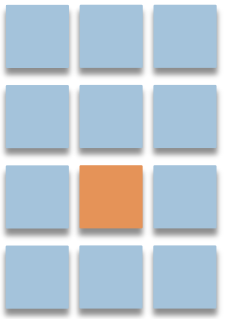
Classification Example



Class A



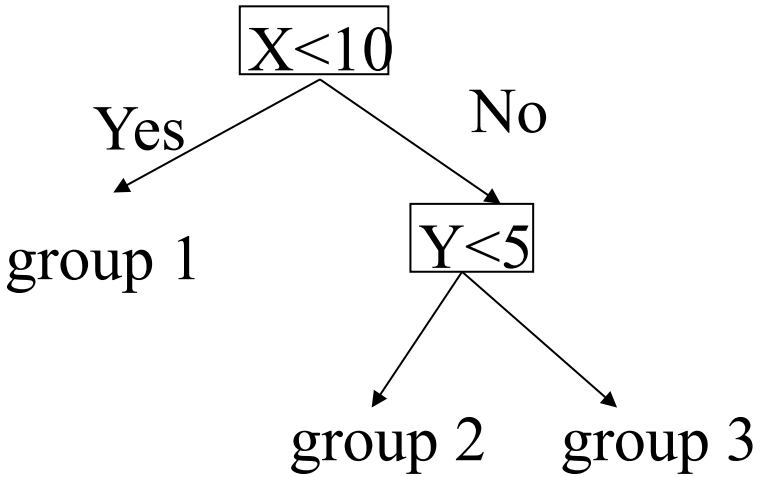
Class B



How about this?

Classification

- Supervised classification
- Organizes data into given classes based on attribute values
- Machine learning



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: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

WSDM Cup 2018 https://wsdm-cup-2018.kkbox.events/

Classification: Application 4

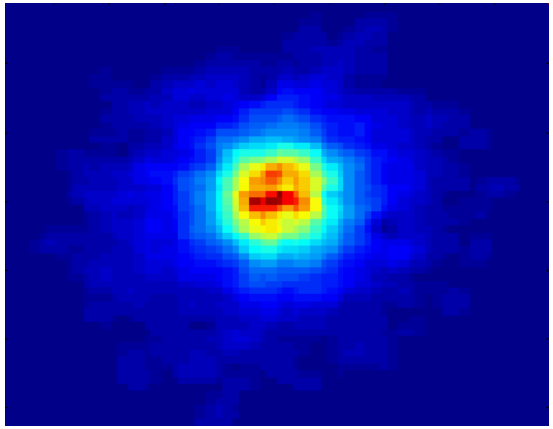
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes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

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

Classifying Galaxies

Courtesy: http://aps.umn.edu

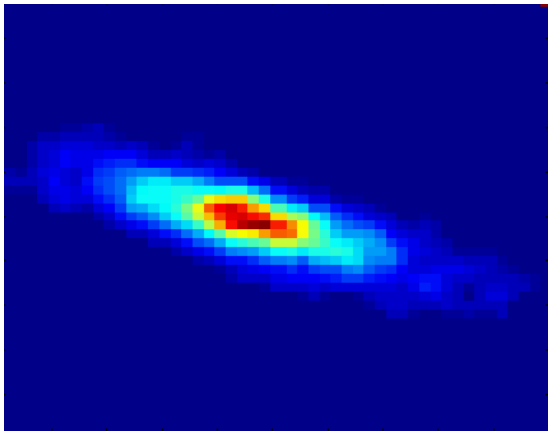
Early



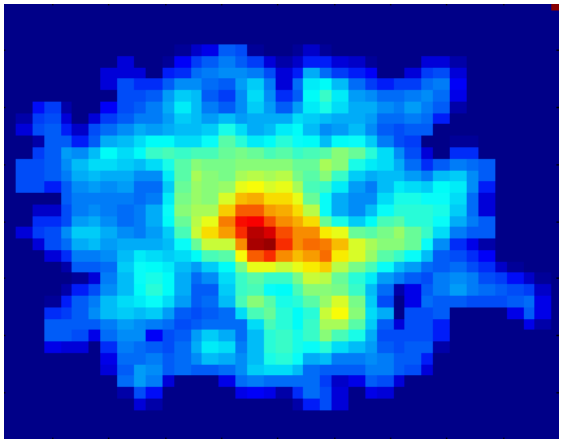
Class:
- Stages of Formation

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

Intermediate



Late



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

Clustering Definition

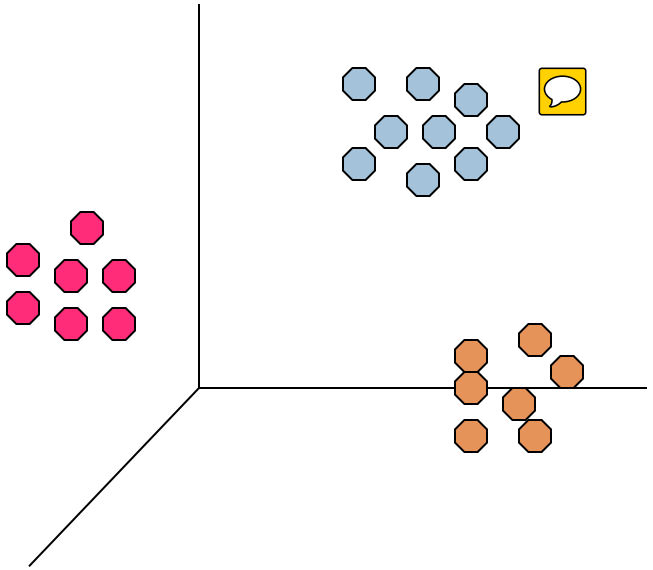
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance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

- Unsupervised classification

Illustrating Clustering

Intracluster distances are minimized

Intercluster distances are maximized



- Euclidean Distance Based Clustering in 3-D space.

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).

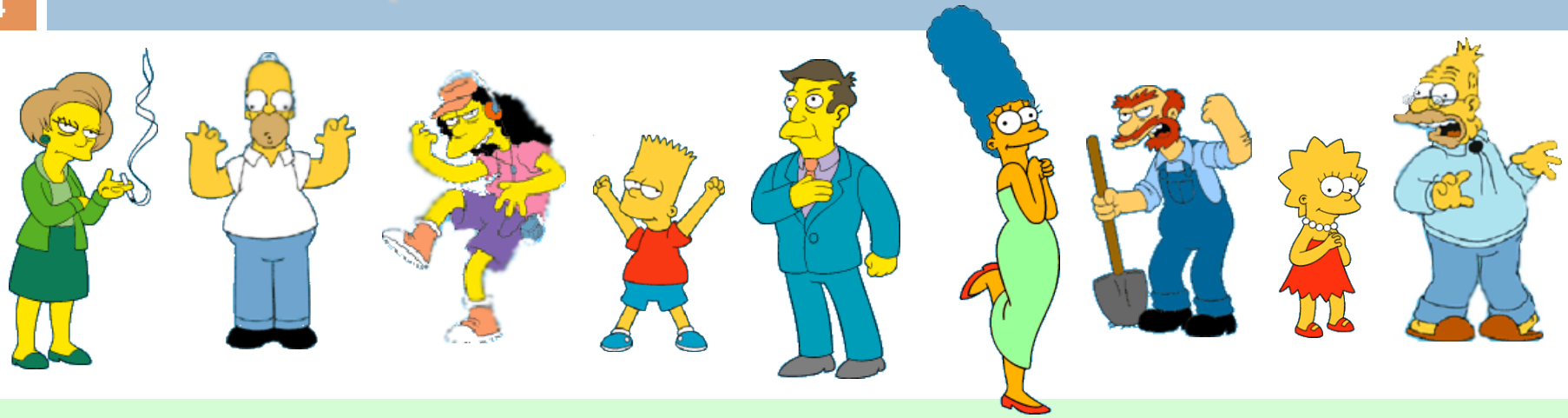
Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

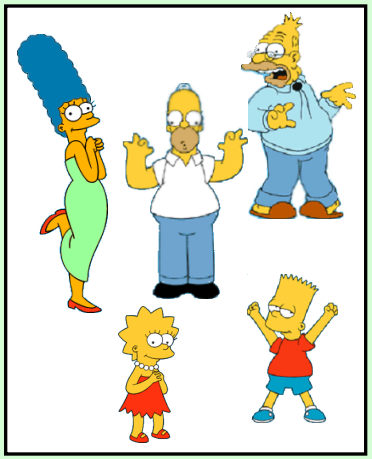
- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	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

What is a natural grouping among these objects?

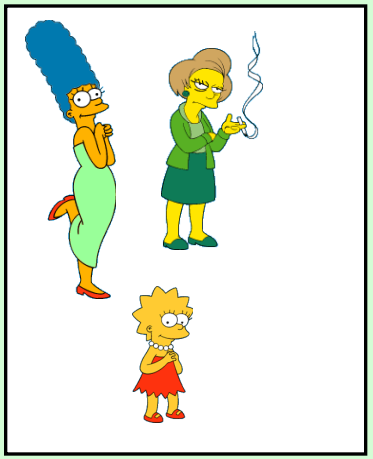


Clustering is subjective



Simpson's Family

School Employees



Females

Males

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

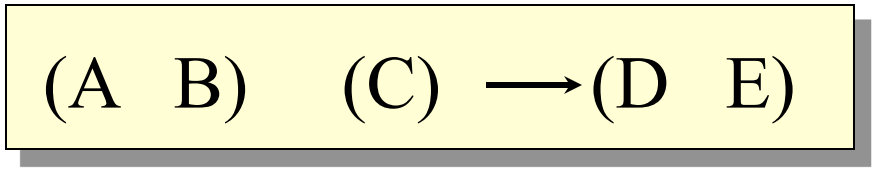
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is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

Association Rule Discovery: Application 3

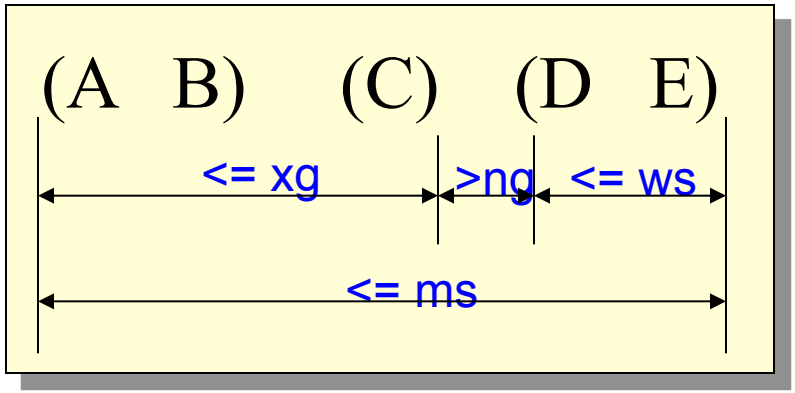
- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.
  - PAKDDCUP 2014

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: 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)

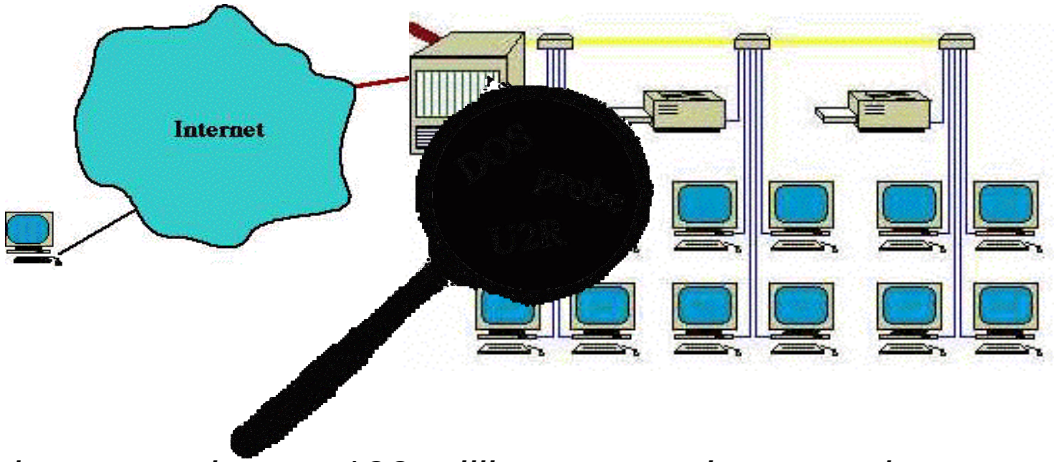
Regression

- 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.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



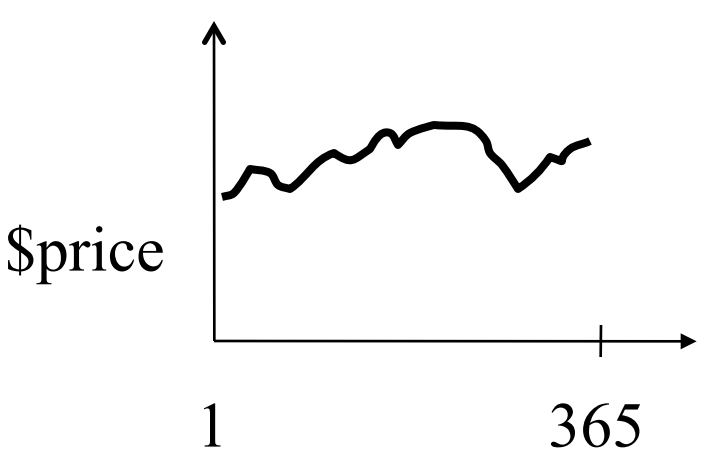
Typical network traffic at University level may reach over 100 million connections per day

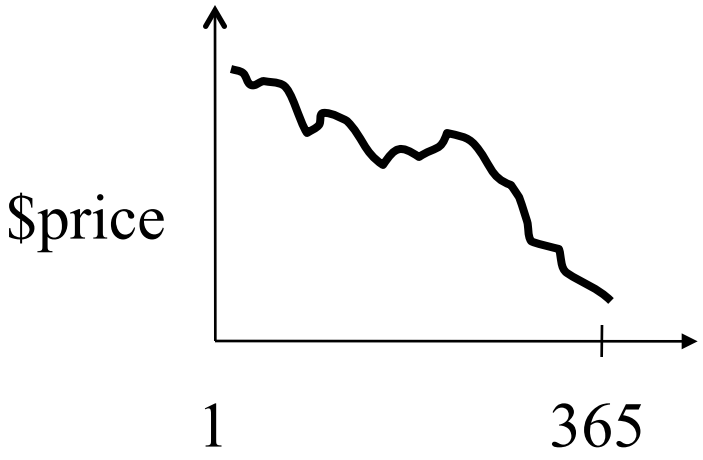
Time Series Analysis

- Trends analysis
- Regression
- Sequential patterns
- Similar sequences

Time Series Database

- Time series
  - Financial, marketing & production: stock price, sales number
  - Scientific: weather data, geological, astrophysics
- Time series DB
  - Databases with many time series of real numbers



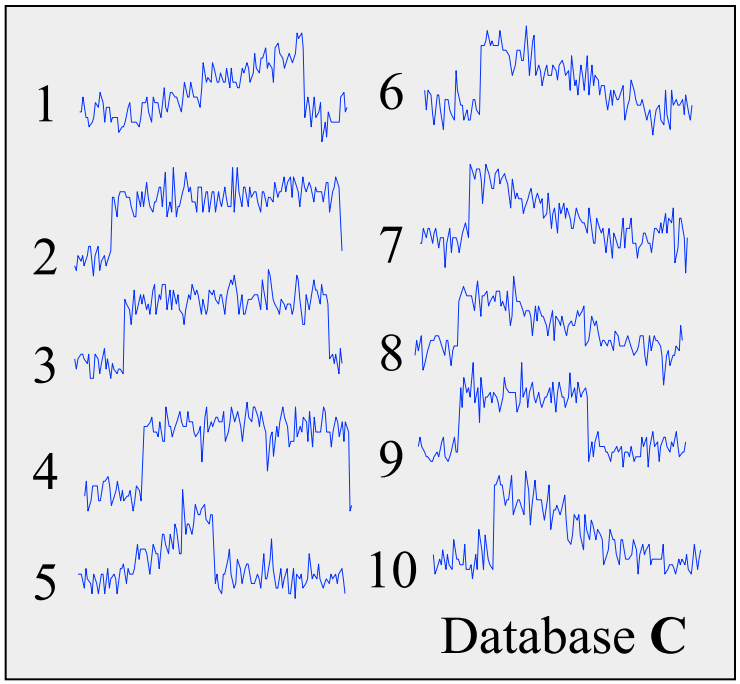


Time Series Database (cont'd)

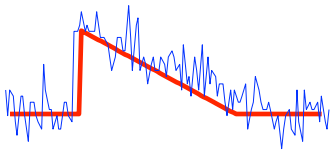
- Query in time series DB
  - Searching for similar patterns
  - Whole matching
  - Subsequence matching
  - Examples
    - Identify companies with similar pattern of growth
    - Determine products with similar selling patterns
    - Discover stocks with similar movement in stock prices
    - Find if a musical score is similar to one of the copyrighted scores

The similarity matching problem can come in two flavors I

Query Q
(template)



1: Whole Matching

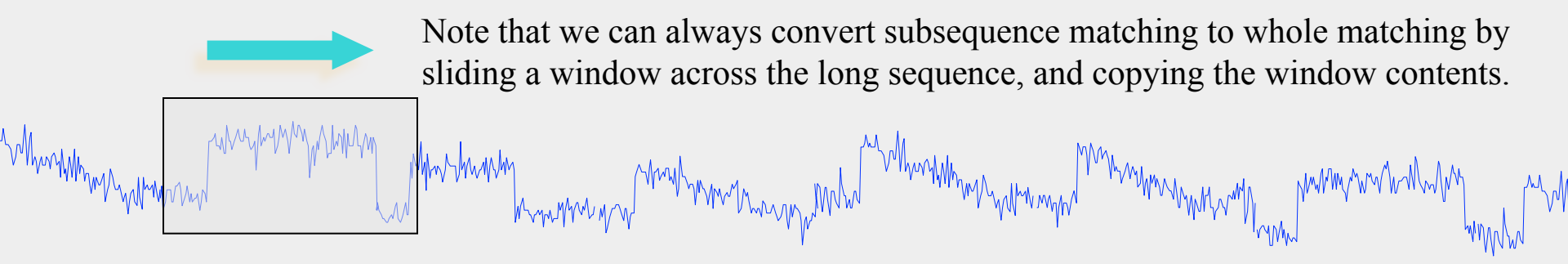


C_6 is the best match.

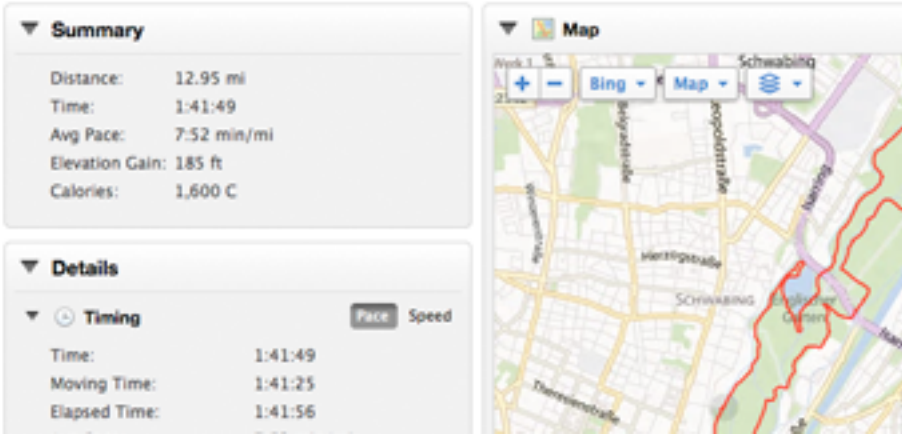
Given a Query Q, a reference database **C** and a distance measure, find the C_i that best matches Q.

The similarity matching problem can come in two flavors II

Given a Query Q, a reference database **C** and a distance measure, find the location that best matches Q.

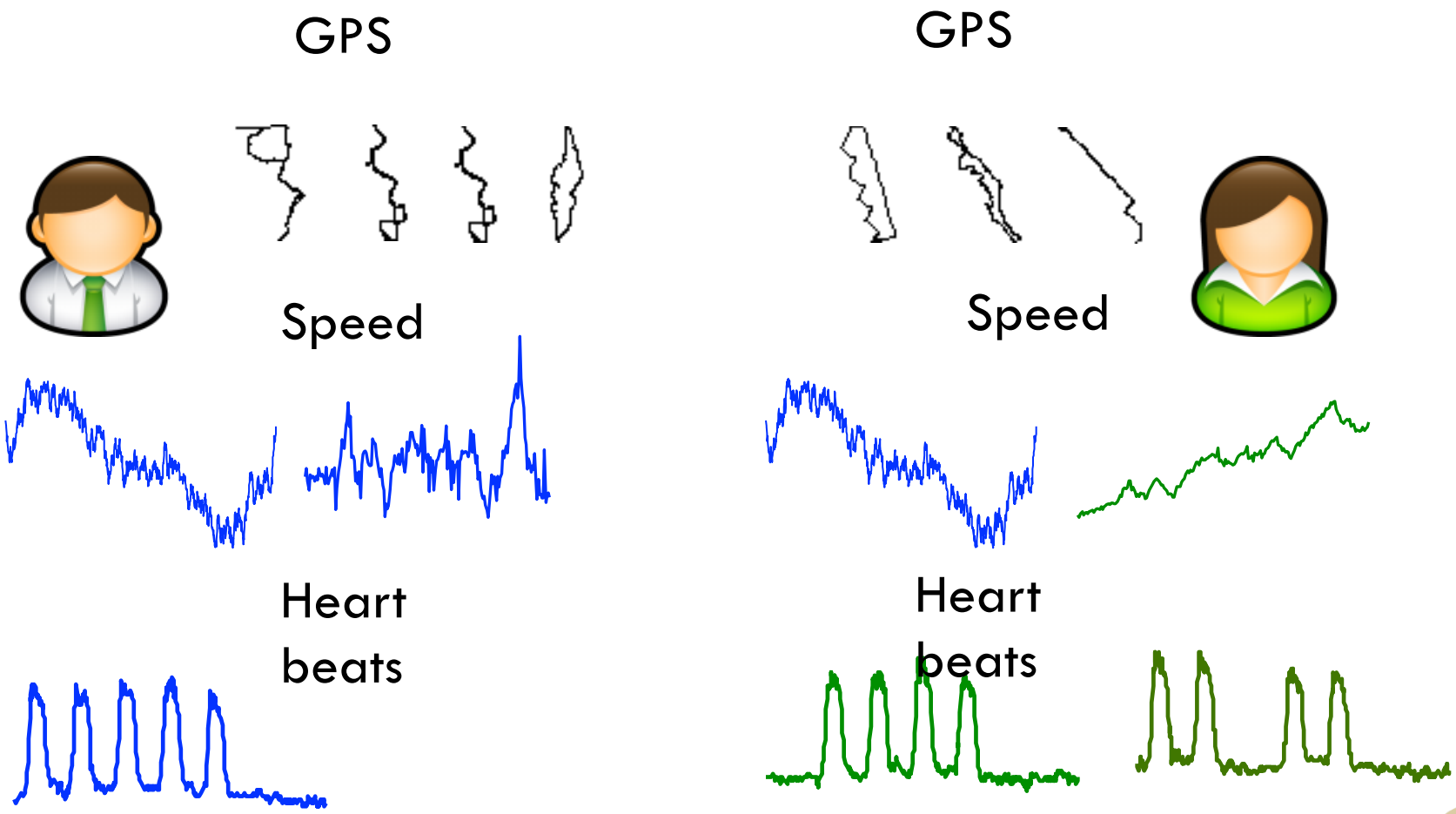


Time series data in IoT Devices

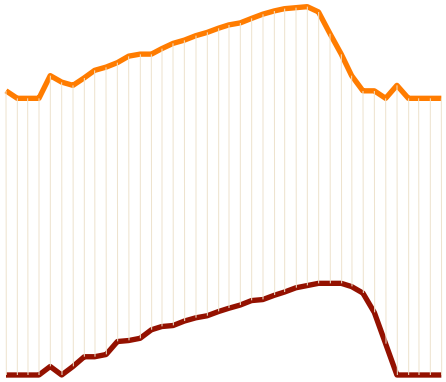


在 Garmin Connect 上追蹤、分析、分享與互相鼓勵 (from Garmin Connect Web)

Measuring similarity among users

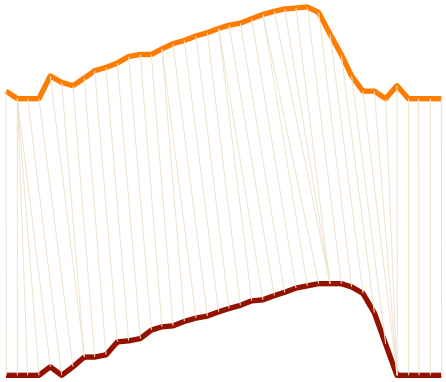


Key issue: Similarity of time series data



Fixed Time Axis

Sequences are aligned "one to one".



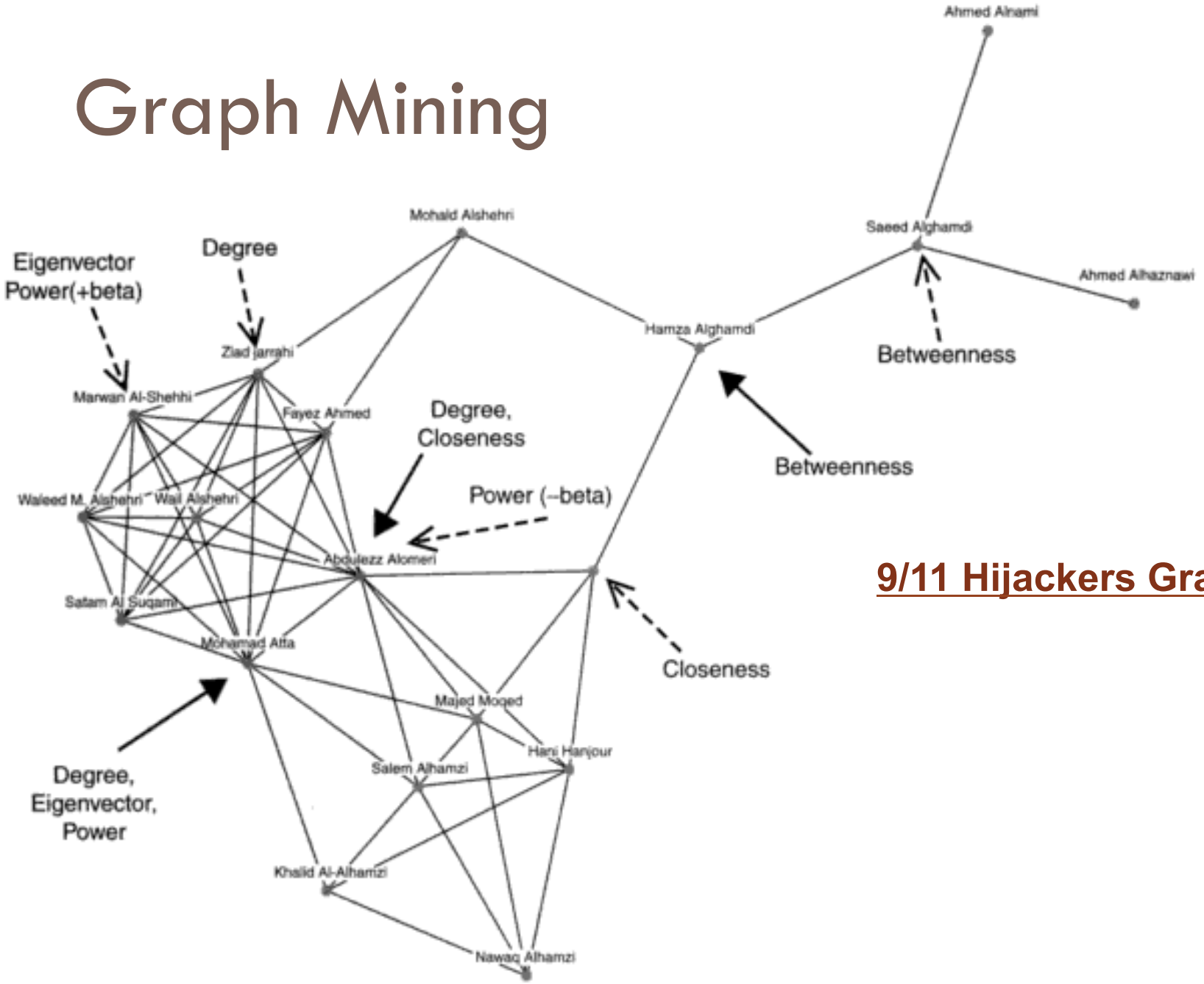
"Warped" Time Axis

Nonlinear alignments are possible.

Existing methods: DTW, LCSS, EDR

Time complexity is huge by computing all similarity between users' GPS traces

Graph Mining



9/11 Hijackers Graph

Reference from "The Text Mining Handbook", Ronen Feldman, James Sanger, P257.

Performance Measurement

- Efficiency
  - Time
  - Accuracy, Precision, Recall, Purity, P@k, ROC, …
- Effectiveness (interestingness)
  - Objective measures; based on statistics & structures of patterns
    - e.g. support, confidence
  - Subjective: based on user's beliefs in data
    - e.g. unexpectedness, novelty

Interestingness

- A pattern is interesting if it is
  - Easily understood by humans
  - Valid on new or test data with some degree of certainty
  - Potentially useful
  - Validates some hypothesis that a user seeks to confirm

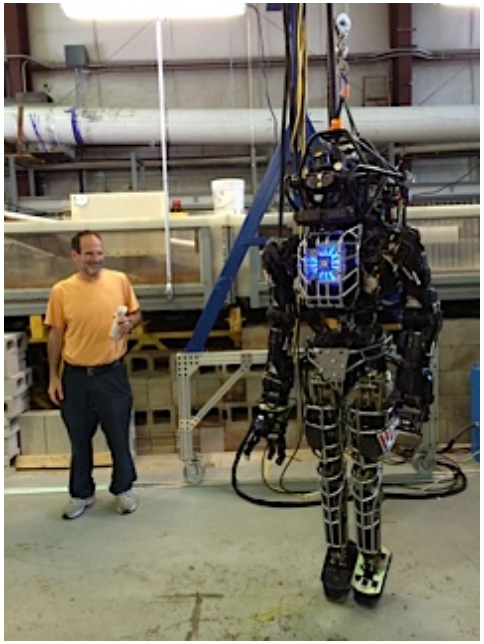
Challenges of Data Mining

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

Features & Challenges of KDD

- Big data
- Feature engineering
  - Deep learning

A chatbot developed by Microsoft (Tay.ai) has gone rogue on Twitter, swearing and making racist remarks and inflammatory political statements.



Dream the future world:

More data about human behavior and knowledge of human

Google Now, Siri and Cortana show one example of mining and understanding your behavior

Ultimate goal: We are building brain for robots