CS423: Data Mining

Introduction

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

Quote of the day

"Never memorize something that you can look up"

- *Albert Einstein* -

Outline

- Why data mining ?
- What is data mining ?
- What kinds of data can be mined ?
- What kinds of applications are targeted ?
- Challenges in data mining
- Summary

Why data mining?

- Size of data grows from terabytes to exabytes (10^{18})
- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, ...
  - Science: Remote sensing, bioinformatics, scientific simulation, ...
  - Society and everyone: news, digital cameras, YouTube
- We have technologies supporting the collection of data.
  - Automated data collection tools
  - Distributed database system

Why data mining?

- We are drowning in data, but starving for knowledge!
- Top 5 data centres worldwide (as of Jan 2017)
  1. – Digital Reality – San Francisco 
  2. – Global Switch – Singapore 
  3. – DuPont Fabros Technology – Virginia 
  4. – CyrusOne – Pheonix 
  5. – China Telecom – Inner Mongolia 

A closer example

- Jumboplus usage logger produces around 2 millions records per day

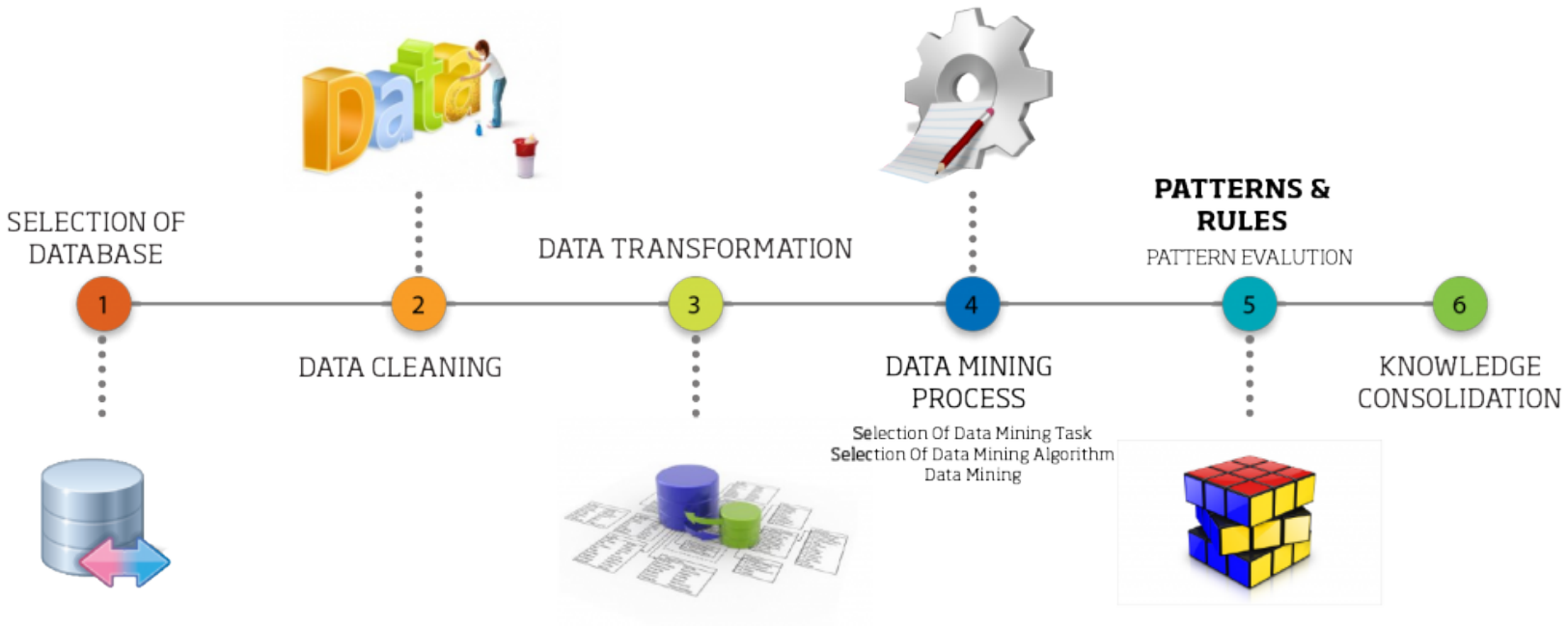
```
"time","mac","user","ip","ap","ssid","apName"
"2017-06-19 00:00:00","00:08:22:04:3C:FC",[illegible]"172.16.161.35","00:00:00:00:00:00",".@ AIS SUPER WiFi",""
"2017-06-19 00:00:00","00:08:22:E6:CF:1D",[illegible]"172.23.59.152","E0:D1:73:1A:BB:C0","@TOT_Wi-Fi","TOT_MED_MED_AP21487"
"2017-06-19 00:00:00","00:08:CA:39:A7:46",[illegible]cmu.ac.th","10.80.137.121","34:BD:C8:55:D0:50","@JumboPlus","Lady8_AP518"
"2017-06-19 00:00:00","00:08:CA:39:FA:02",[illegible]th","10.80.30.91","E0:D1:73:1A:62:F0","@JumboPlus","TOT_MED_MED_AP21568"
"2017-06-19 00:00:00","00:08:CA:3C:5B:1E",[illegible]cmu.ac.th","10.80.185.136","2C:D0:2D:B1:81:F0","@JumboPlus","Suandok_Dorm_AP982"
"2017-06-19 00:00:00","00:08:CA:3C:77:BA",[illegible]c.th","10.80.130.218","50:67:AE:C2:3B:F0","@JumboPlus","TOT_MED_NUR_AP20293"
"2017-06-19 00:00:00","00:08:CA:6A:DB:59",[illegible]mu.ac.th","10.80.166.185","84:80:2D:AA:F6:80","@JumboPlus","TOT_MAE_MHD_AP22090"
"2017-06-19 00:00:00","00:08:CA:85:1E:5A",[illegible]th","10.80.172.106","E0:D1:73:1A:BD:E0","@JumboPlus","TOT_MED_MED_AP21563"
"2017-06-19 00:00:00","00:08:CA:B1:36:4C",[illegible].th","10.80.44.39","E0:D1:73:1A:7D:20","@JumboPlus","TOT_MED_MED_AP21569"
"2017-06-19 00:00:00","00:08:CA:F2:94:8E",[illegible]c.th","10.80.55.135","A0:E0:AF:DC:B0:10","@JumboPlus","Suandok_Dorm_AP975"
"2017-06-19 00:00:00","00:11:7F:39:26:3E",[illegible].th","10.80.182.2","18:9C:5D:96:06:10","@JumboPlus","Flat7_AP746"
"2017-06-19 00:00:00","00:15:AF:CE:6E:1D",[illegible].ac.th","10.80.192.194","0C:F5:A4:41:6B:80","@JumboPlus","TOT_COM_SCI_AP20475"
"2017-06-19 00:00:00","00:16:D4:EF:5F:79",[illegible]ek@cmu.ac.th","10.80.172.124","50:67:AE:C2:66:40","@JumboPlus","TOT_MED_DENT_AP21763"
"2017-06-19 00:00:00","00:17:C4:88:90:FA",[illegible].th","10.80.55.172","E0:D1:73:1A:A7:40","@JumboPlus","TOT_MAE_MHD_AP22085"
"2017-06-19 00:00:00","00:17:CD:2C:20:B3",[illegible]u.ac.th","10.73.24.61","00:00:00:00:00:00","@JumboPlus",""
"2017-06-19 00:00:00","00:19:7E:73:AB:38",[illegible].th","10.80.106.154","18:9C:5D:4A:64:20","@JumboPlus","dorm_40y_AP639"
"2017-06-19 00:00:00","00:1B:9E:2C:E2:A0",[illegible]"0.0.0.0","1C:DE:A7:CF:9B:00","@JumboPlus","TOT_MED_NUR_AP21966"
"2017-06-19 00:00:00","00:1B:B1:A9:75:54",[illegible]n@cmu.ac.th","10.80.114.81","34:62:88:0E:1E:30","@JumboPlus","TOT_MED_NUR_AP20294"
"2017-06-19 00:00:00","00:1E:65:4D:95:10",[illegible].th","10.80.98.224","E0:D1:73:1A:49:00","@JumboPlus","TOT_MAE_MHD_AP22088"
"2017-06-19 00:00:00","00:1F:3A:B9:B7:5D",[illegible].th","10.80.85.183","E0:D1:73:1A:9E:A0","@JumboPlus","TOT_MED_MED_AP21570"
"2017-06-19 00:00:00","00:1F:3B:9E:25:89",[illegible].th","10.80.53.171","18:9C:5D:96:06:10","@JumboPlus","Flat7_AP746"
"2017-06-19 00:00:00","00:1F:3C:5F:A7:B7",[illegible],"10.80.45.82","34:62:88:0E:0F:30","@JumboPlus","TOT_AGR_FIN_AP21028"
"2017-06-19 00:00:00","00:21:00:E5:63:34",[illegible].th","10.80.50.128","18:9C:5D:96:48:40","@JumboPlus","dorm_40y_AP623"
"2017-06-19 00:00:00","00:21:5D:56:15:F4",[illegible]u.ac.th","10.80.168.76","B4:14:89:1A:08:60","@JumboPlus","Gent6_AP267"
"2017-06-19 00:00:00","00:22:FA:18:56:80",[illegible]u.ac.th","10.80.85.157","0C:F5:A4:41:34:30","@JumboPlus","TOT_COM_SCI_AP20562"
"2017-06-19 00:00:00","00:22:FA:25:8F:D6",[illegible]ac.th","10.80.114.66","34:62:88:0E:1E:F0","@JumboPlus","TOT_MED_NUR_AP20288"
"2017-06-19 00:00:00","00:22:FA:2B:6B:FA",[illegible].th","10.80.200.95","28:34:A2:7E:AD:60","@JumboPlus","Suandok_Dorm_AP906"
"2017-06-19 00:00:00","00:22:FA:52:0D:C4",[illegible]ac.th","10.80.43.105","34:62:88:0E:10:00","@JumboPlus","TOT_MED_MED_AP20238"
```

What is data mining?

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

A bigger picture

- Data mining is one phase in the Knowledge Discovery in Databases (KDD) process.



- In practice, performing data mining usually involves a little bit of all steps.

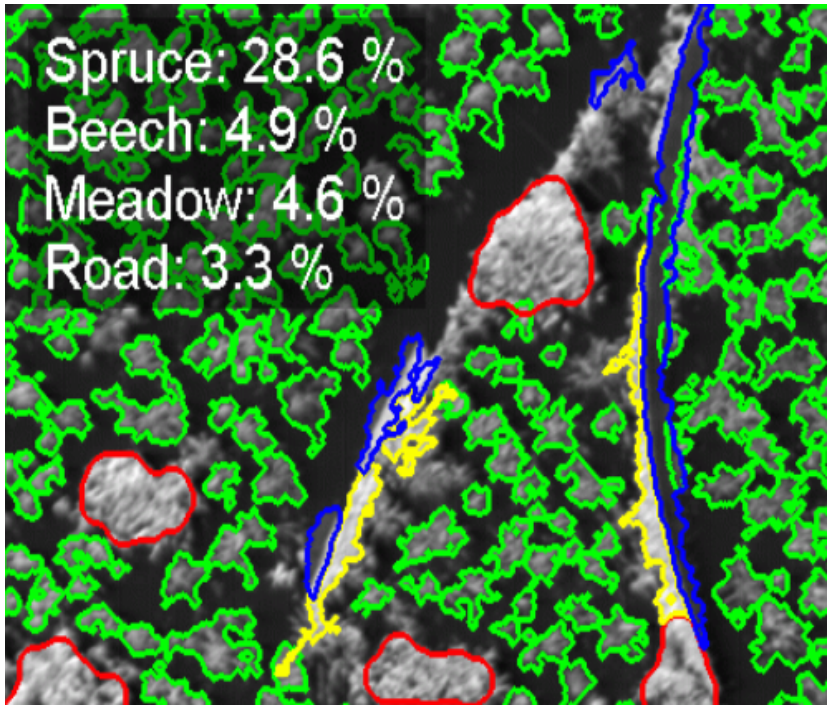
Overview of Focuses in Data Mining

- Data to be mined: transactional data, stream, time-series, text, multi-media, graphs, etc.
- Knowledge to be mined: association, classification, clustering, trend, outlier analysis, etc.
- Techniques utilised: Online analytical processing (OLAP), machine learning, statistics, visualisation,
- Application adapted: Retail, banking, bio-medical, stock market, text and web mining, etc.

Data to be mined ?

- Data streams and sensor data
- Time-series data, temporal data, sequence data (incl. bio-sequences)
- Graphs, social networks and information networks
- Spatial data and spatiotemporal data
- Multimedia database
- Text databases
- The World-Wide Web

[Data to be mined] Data stream



[LEFT] A music recognition application. [RIGHT] Classification of remotely sensed land usage data.

[Data to be mined] Sensor readings

(a) Walking and Jogging

(b) Squatting

(c) Sit-up

(d) Push-up

Predicting sport activities from accelerometer data. (will be presenting at IDEAL 2017, China)

[Data to be mined] Graph

Trying to isolate accounts that the same person has been used to communicate.

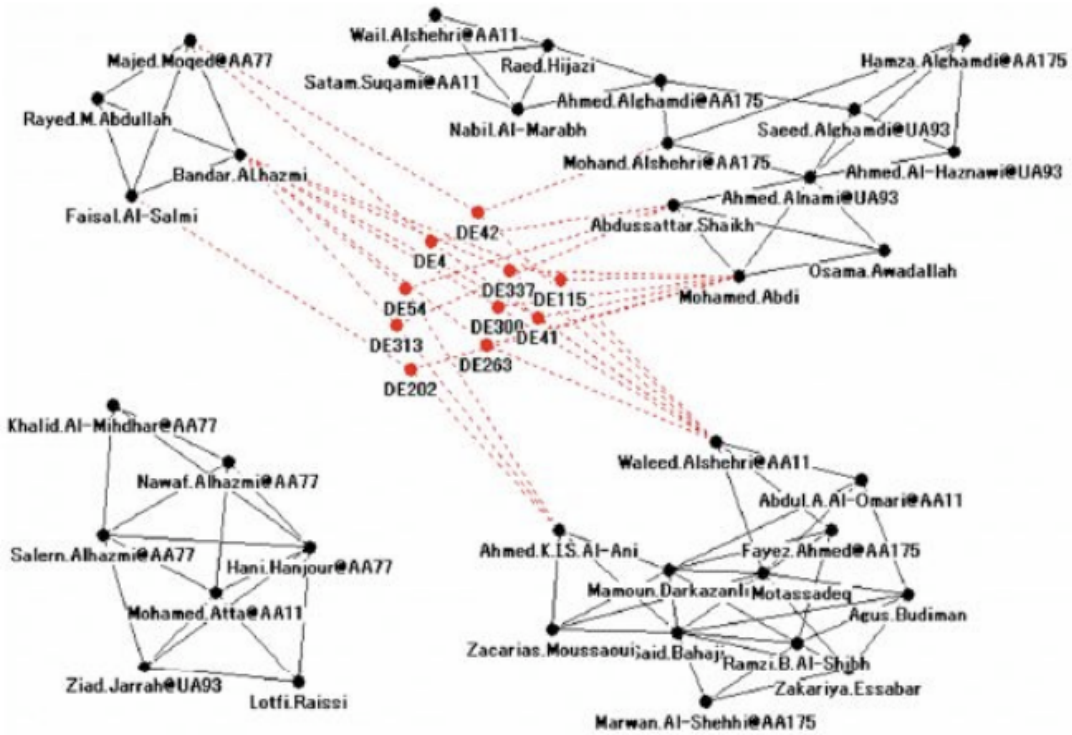


Figure 10 Four clusters and ten of the highly ranked red nodes corresponding to Mustafa A. Al-Hisawi hidden in the suspicious records. Waleed Alshehri and Mohand Alshehri are retrieved as neighbor persons of the red nodes.

[Data to be mined] Image

Generating Image descriptions

"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."

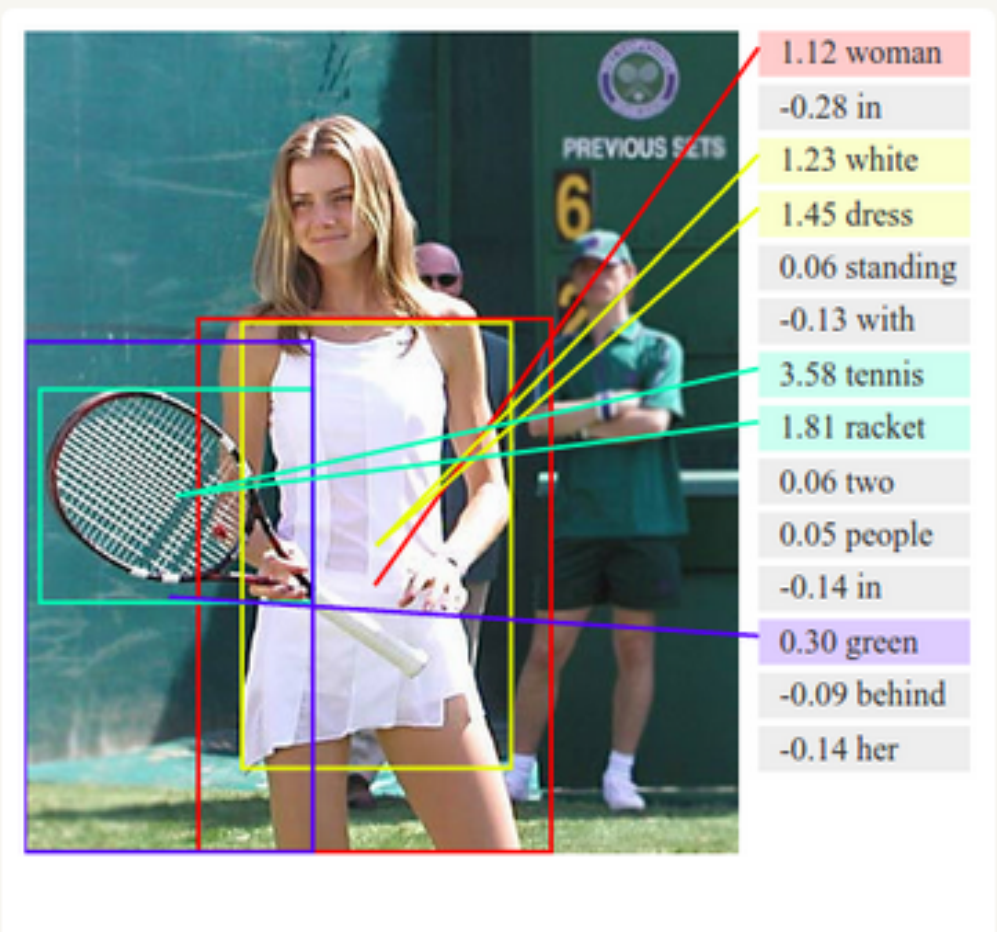
"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

Image Analysis

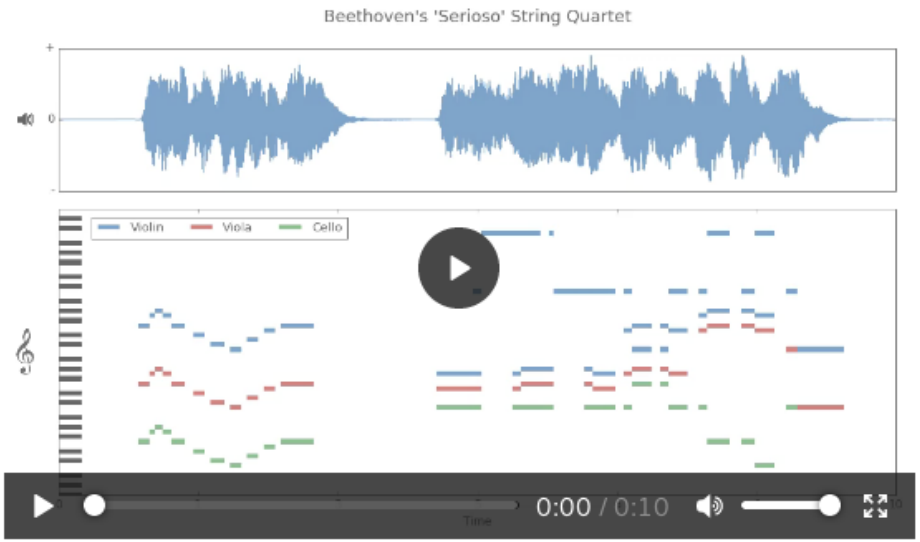
How did they do it ?



[REF] Andrej Karpathy, Li Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions. CVPR (2015)

[Data to be mined] Music

MusicNet: A collection of labeled classical music.



What can you do ?

- Classify the instruments that perform in a recording.
- Classify the composer of a recording.
- Predict the next note in a recording, conditioned on history.

[Knowledge to be mined] Generalisation

- Concept/Class description
  - What are characteristics of good customers/bad customers ?
- Data summarisation
  - Central Tendency Measure – Mean, Mode, Median
  - Dispersion Measure – Standard deviation, Variance
- Meaning of generalisation $\rightarrow$ Must be applicable to unseen data

[Knowledge to be mined] Association/ Correlation

- Frequent patterns
  - What items are frequently purchased together at 7-11 ?
- Association, correlation vs. causality
  - Diaper $\rightarrow$ Beer
  - Fried chicken $\rightarrow$ Sticky rice
  - Coke $\rightarrow$ Dimsum
- How to mine such patterns and rules efficiently in large datasets?
- How to efficiently make use to the rules ?

[Knowledge to be mined] Classification

- Basic idea
  - Construct models (functions) based on some training examples
    - ⋆ Training examples is of the form (input,label)
  - Goal: Predicting correct label of future unseen input.
  - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, nearest neighbours, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, classifying stars, diseases, web pages, ...

[Knowledge to be mined] Cluster analysis

- Basic idea
  - Construct models (functions) based on some training examples
    - ★ Training examples is of the form (input) ; no label is available
  - Goal: Group similar inputs to form new categories.
  - E.g., clustering online customers, clustering second hand houses.
- Typical methods
  - K-means, Gaussian Mixture Model.

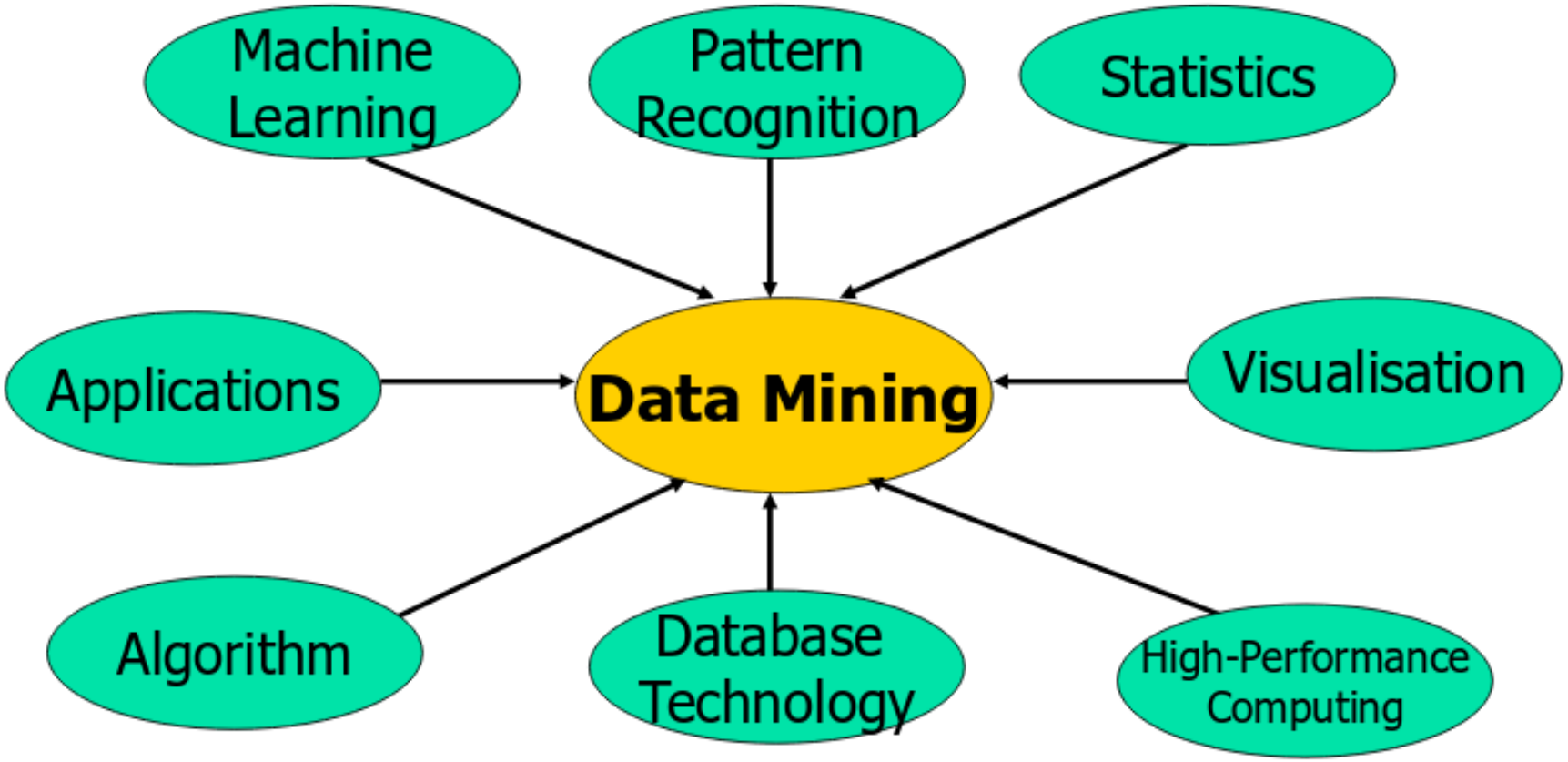
[Knowledge to be mined] Outlier analysis

- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception?
  - One person's garbage could be another person's treasure
- Methods: by product of clustering or regression analysis, ...
- Useful in fraud detection, rare events analysis

Evaluation of knowledge

- Generalisation & Association rule
  - Evaluation is based on coverage and generality.
- Classification
  - Accuracy of the predictive model.
- Clustering
  - Maximising intra-class similarity & minimising interclass similarity
- Outlier analysis
  - Identified items is in agreement with expert's view.
- Other common properties of good model.
  - Timeliness

What kind of techniques are used ?



Applications of data mining

- Web page analysis: from web page classification, clustering to PageRank
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering
- Text analysis, sentiment analysis,
- And many more.

Example

Customers Who Viewed This Item Also Viewed

ESQ Movado Unisex 07301436 ESQ ONE Round Stainless Steel Watch
★★★★½ 14
$150.00 Prime

Movado Men's 0606504 "Museum" Stainless Steel Watch
★★★★½ 8
$695.00 Prime

Movado Men's 0606610 "Museum" Stainless Steel, Black Leather, and Blue Dial Watch
★★★★½ 3
$495.00 Prime

Movado Women's 0606503 "Museum" Stainless Steel and Leather Strap Watch
★★★★½ 4
$495.00 Prime

Movado Men's 606307 Stainless Steel Watch
★★★★½ 3
$1,995.00 Prime

Customers Who Bought This Item Also Bought

Movado Women's 0606503 "Museum" Stainless Steel and Leather Strap Watch
★★★★½ 4
$495.00 Prime

ESQ Movado Unisex 07301436 ESQ ONE Round Stainless Steel Watch
★★★★½ 14
$150.00 Prime

Kenneth Cole Reaction Men's Hematite Tie Clip
★★★★½ 30
$19.53 - $23.00

MICHAEL Michael Kors Mk Logo Crossbody Bag
★★★★½ 38
$97.40 - $229.99

Move Free Advanced Glucosamine Chondroitin Joint Supplement with Hyaluronic Acid, MSM...
★★★★☆ 179
$14.99 Prime

Nuby Hot Safe Spoons 4 Pack BPA FREE
★★★★☆ 65
$2.98

Issues in data mining [1]

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern
  - Constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
- Presentation and visualisation of data mining results

Issues in data mining [2]

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed on a variety of data
- Data mining functionalities: characterisation, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications

References

- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997