CSC4008: Techniques for Data Mining

Course Overview

Jan. 14, 2021

Chenye Wu

wuchenye@cuhk.edu.cn

Points of Contact

- Instructor:

  Prof. Chenye Wu, wuchenye@cuhk.edu.cn

- Instructor Office hours:
- Tuesday&Thursday, 3:00pm – 3:30pm
  TD102

Teaching Assistants

Kai Li (Lead TA)	**219019032@link.cuhk.edu.cn**
Siyi Wang	218019023@link.cuhk.edu.cn
Qingyan Meng	219019044@link.cuhk.edu.cn
Yuncheng Jiang	220019054@link.cuhk.edu.cn
Chi Li	220019044@link.cuhk.edu.cn
Zhiwei Tang	220019070@link.cuhk.edu.cn

Course Coverage

- High Dimensional Data Analysis (The Theory Part)

- Recap for Probability Theory

- Sparse Signal Models
- Convex Methods for Sparse Signal Recovery
- Convex Methods for Low-Rank Matrix Recovery
- Decomposing Low-Rank and Sparse Matrices

Course Coverage

- Classical Data Mining (The Practice Part)

- Data Preprocessing and Descriptive Statistical Analysis
- Association and Correlation Analysis, and Pattern Discovery
- Classification and Prediction
- Cluster Analysis and Outlier Detection

Reference Books

- Theory Part

- High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications
- John Wright and Yi Ma, Cambridge University Press

- Practice Part

- Data Mining: Concepts and Techniques, 3rd ed.
- Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems

Tests and Grading Criteria

- Assignments (25%, each 5%):
  - 5 Assignments.
  - No late submission allowed. Please be on time!
  - Solutions will be posted on Blackboard after the submission ddl.

- Midterm Exam (35%):
  - Questions similar to assignment questions
  - Application of what was dicussed in lecture

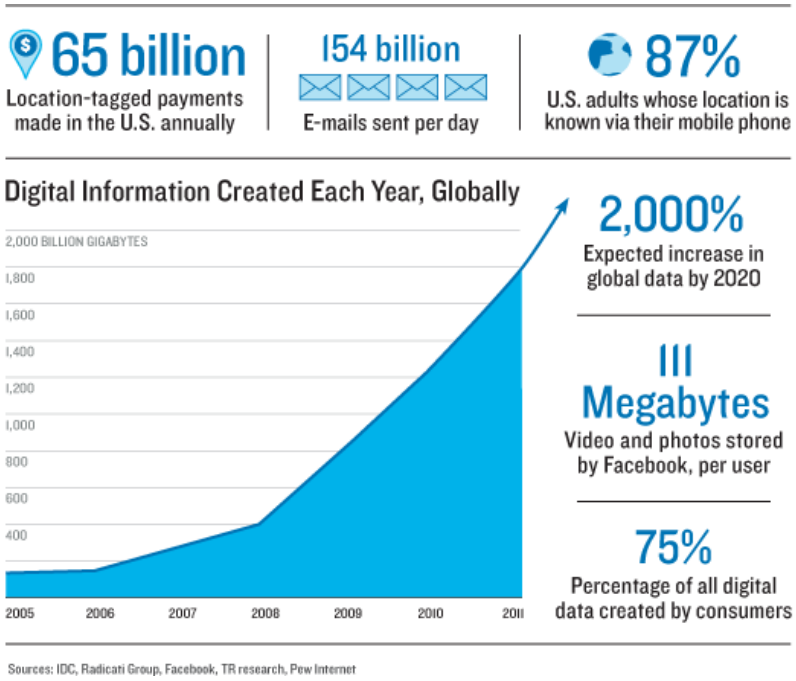
- Project (40%):
  - Will be introduced in Week 3, after the Add/Drop period

Today's Lecture

- What is Data Mining?
- What Makes Data Mining useful?
- How to enable Data Mining?

WHAT IS DATA MINING?

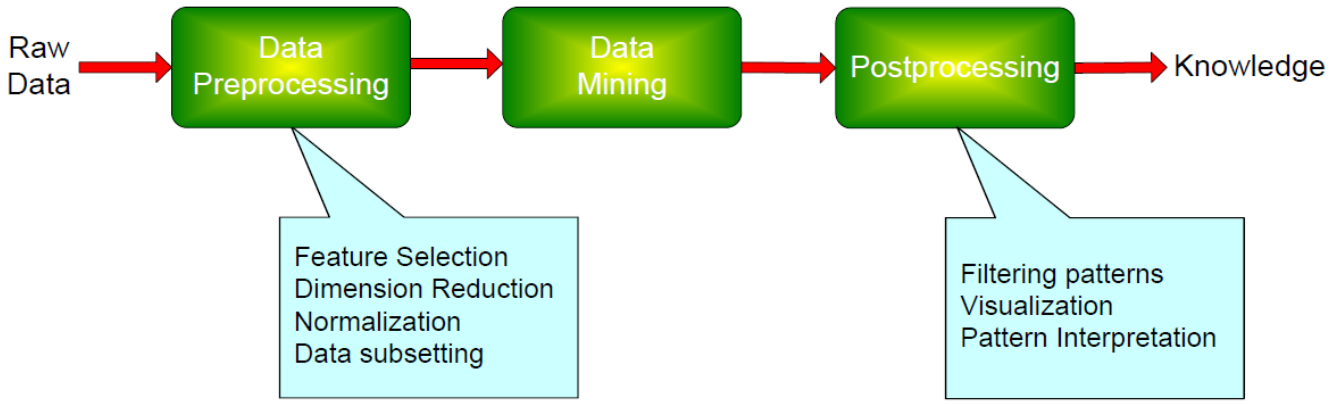
Do ALL Data Contain Value?



How Do Data Convey Information?

Definition of Data Mining

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from (huge amount of) data

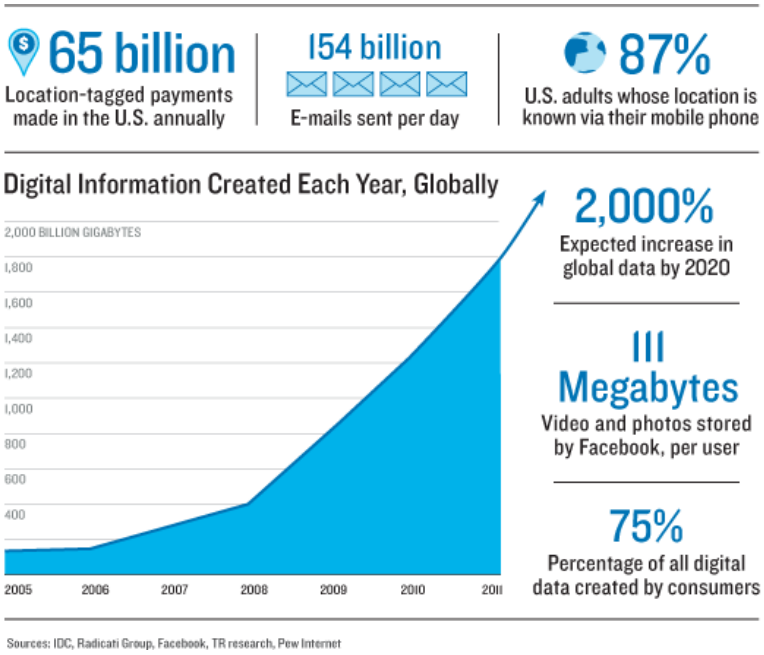


Measurement of Interesting Patterns

- Novelty
  - Not previously know, surprising (used to remove redundant rules)
- Utility
  - Potential usefulness, e.g., support association, noise threshold, etc.
- Simplicity
  - Rule length, decision tree size, etc.
- Certainty
  - Confidence level, classification reliability or accuracy, rule strength, etc.

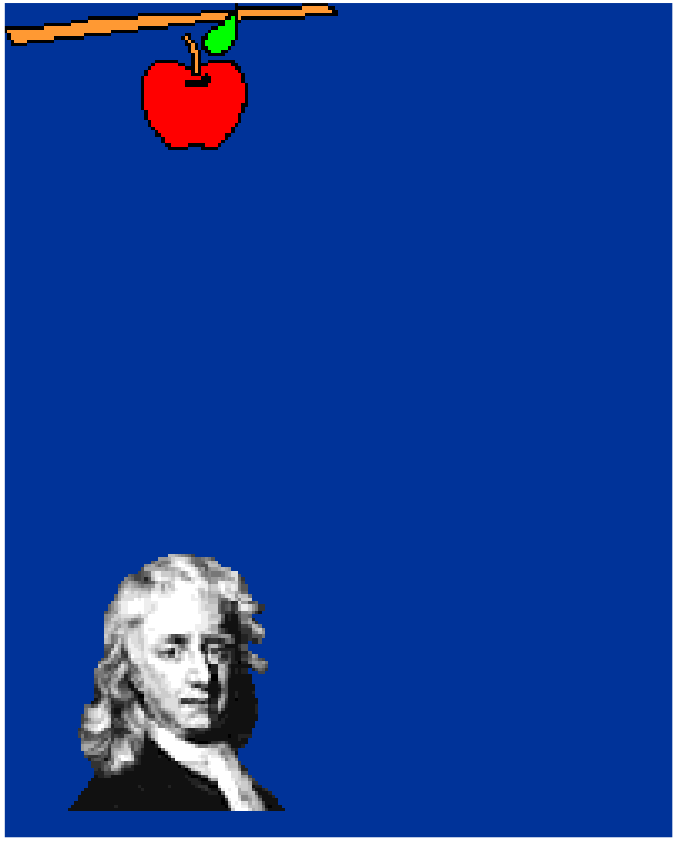
WHAT MAKES DATA MINING USEFUL?

Motivation of Data Mining



- We are drowning in data (i.e., the big data era)
  - Data explosion: Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and to be analyzed.
  - Great data diversity: Stream data, web data, bio-data, semi-structured data, etc.
- Data mining can help generate new hypothesis or help analysts make sense out of the data

A More Scientific Viewpoint

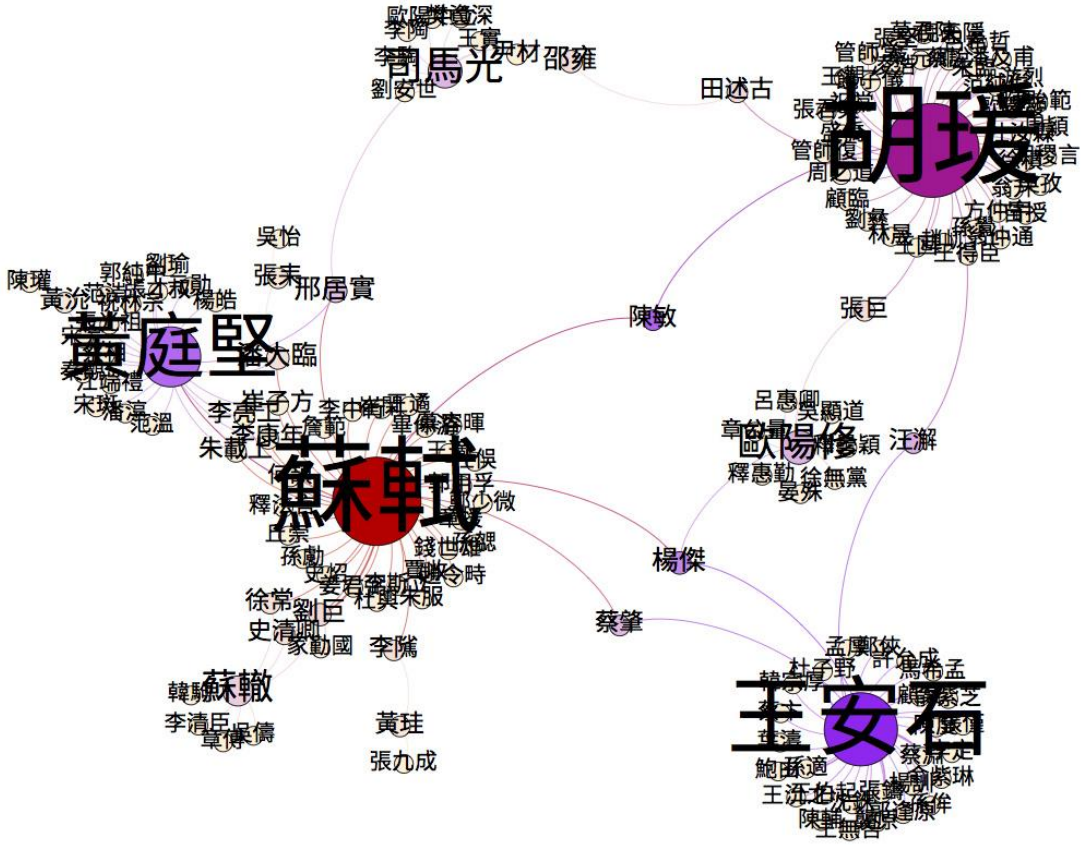


Knowledge Discovery in the old days

- Observe phenomena
- Formulate theory
- Validate theory via experimentation

Discovery of the Universal Law of Gravitation
From: http://csep10.phys.utk.edu/astr161/lect/history/newtongrav.html

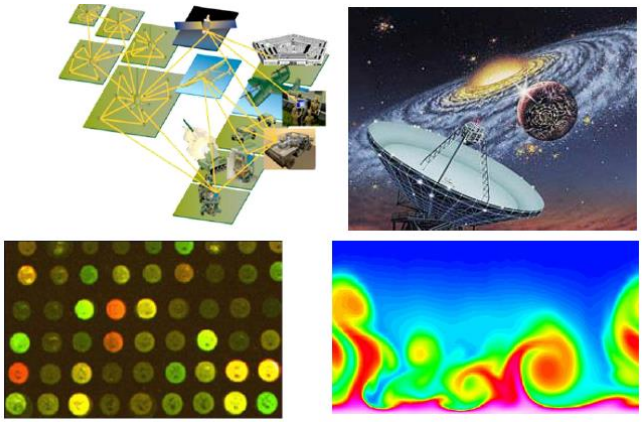
A More Scientific Viewpoint



微博@JasonGKwok：

http://photo.weibo.com/1656966370/wbphotos/large/mid/3836469267566154/pid/62c34ce2gw1erktcgv2q7j20sg0sg7dn

A More Scientific Viewpoint



SVM-derived decision tree

- Science example:
- •P. Raccuglia, et al. Machine-learning-assisted materials discovery using failed experiments, Nature, May 5, 2016.
- machine-learning model outperformed traditional human strategies

Data Mining Also Changes Our Lifestyles

- Clothing
  - E.g., fashion recommendations based on sales figures and client surveys
  - Abu-Mostafa. Machines that think for themselves. *Scientific American*, 289(7):78-81, July 2012.
- Food
  - E.g., restaurant recommendation
  - Wei-Ta Chu, et al. A hybrid recommendation system considering visual information for predicting favorite restaurants. WWW J. 2017.
- Housing
  - E.g., predict home appliance usage
  - Kaustav Basu, et al. A prediction system for home appliance usage. Energy and Buildings, 2013.
- Transportation
  - E.g., recommend new POIs the user has not visited before
  - Wei Zhang, Jianyong Wang. A Location and Time Aware Social Collaborative Retrieval Approach for New Successive Point-of-Interest Recommendation. CIKM 2015.

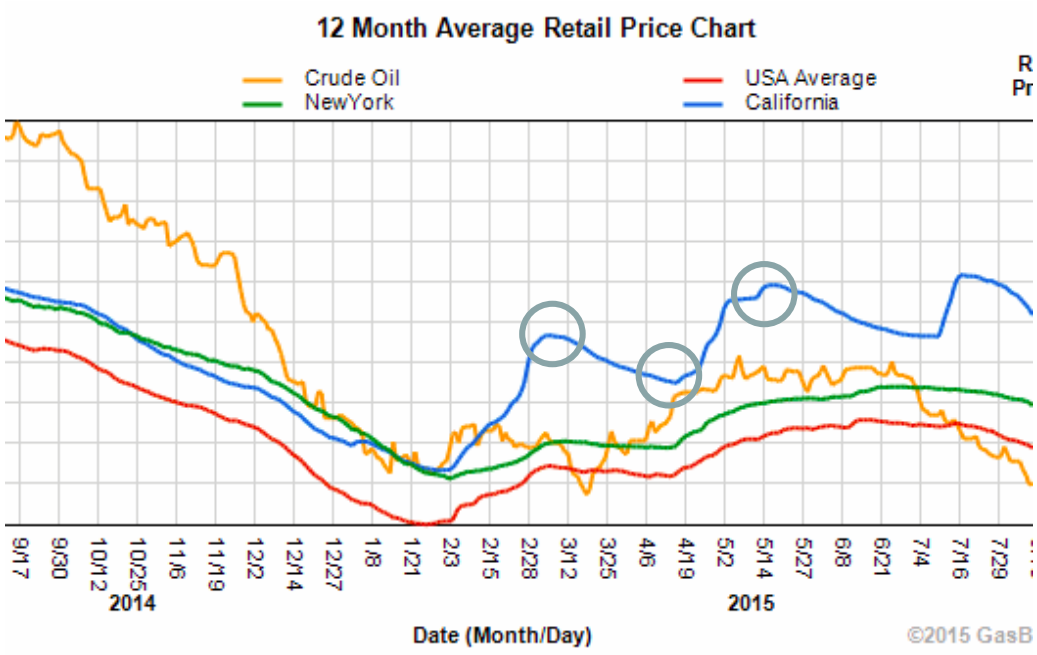
Data Mining has a much broader impact

- Data analysis and decision support
  - Market analysis and management
  - Customer relationship management (CRM), cross selling, market segmentation & target marketing
  - Risk analysis and management
  - Customer retention, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)

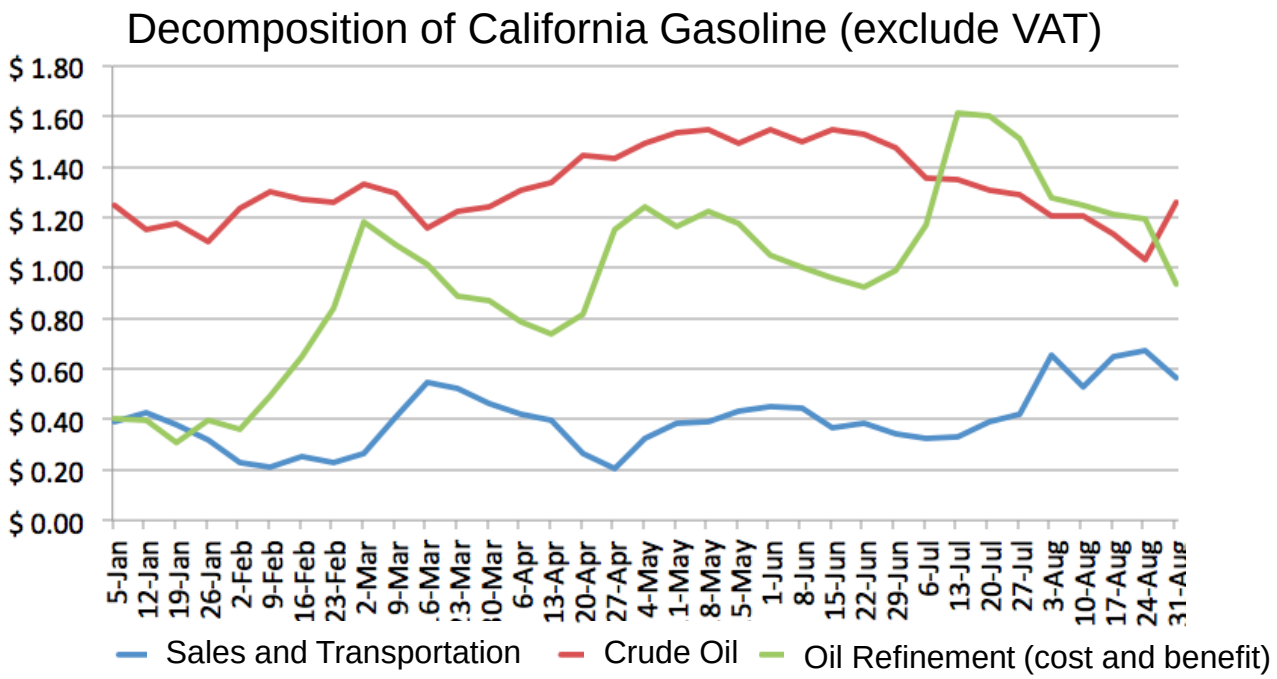
- Other Applications
  - Text mining (news group, email, weibo) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis
  - Social network analysis

HOW TO ENABLE DATA MINING?

The Old Day Tricks: Outlier Dection

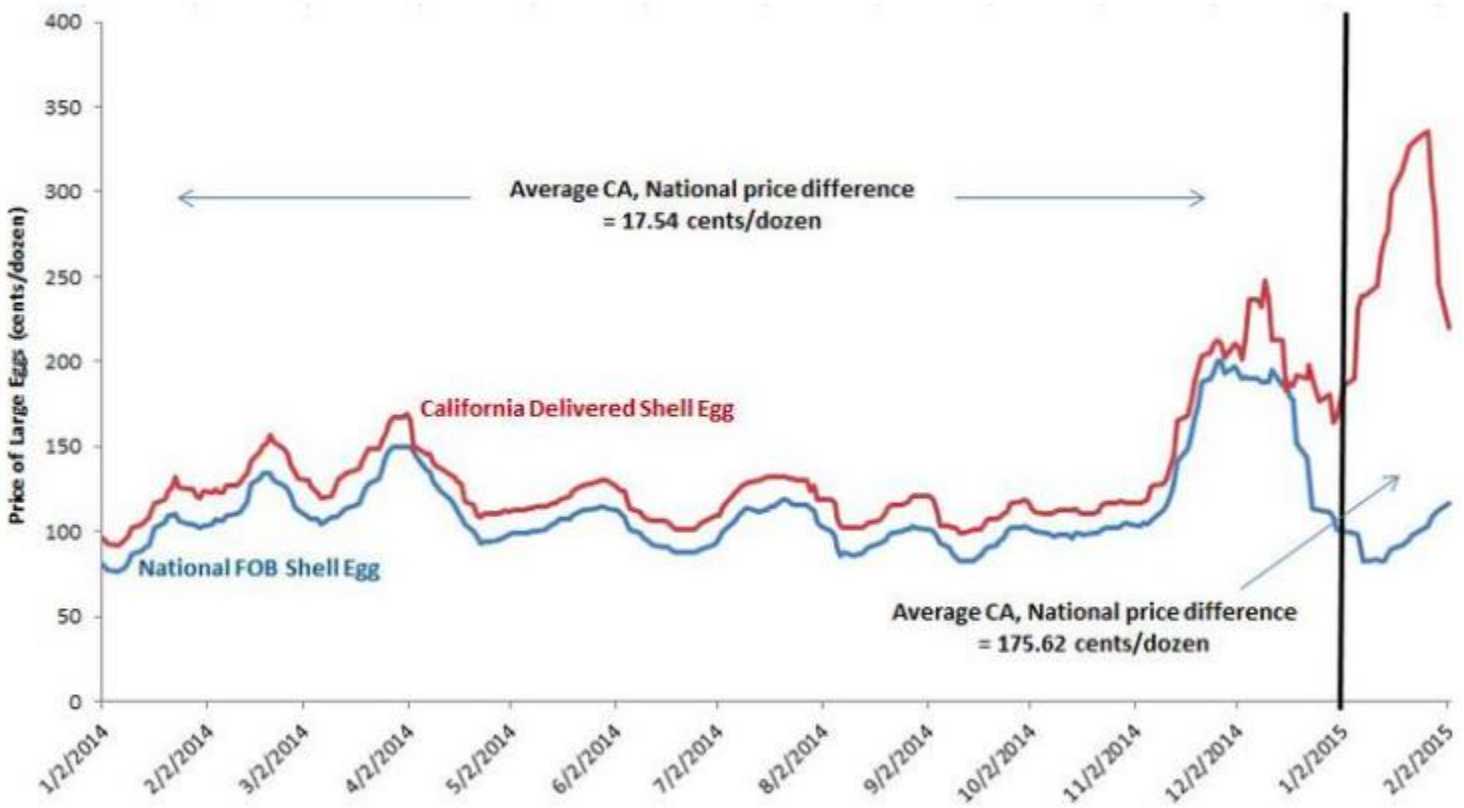


Dive into Details: Difference in Difference



The truth: union and strike

One more example



What happened?

- The Prevention of Farm **Animal Cruelty Act** passed via state ballot initiative by a 64% majority of **California** voters in 2008. The **law**, which came into effect in January 2015, requires that egg-laying **hens** have the ability to fully spread their wings without touching another bird or the side of an enclosure.

- To put it simple, each egg-laying hen has the right to stay in a place of the size similar to a letter page.

Data Mining Functionalities

- Predictive functionalities:
- Use some variables to predict unknown values of other variables

- Descriptive functionalities:
- Find human-interpretable patterns that describe the data

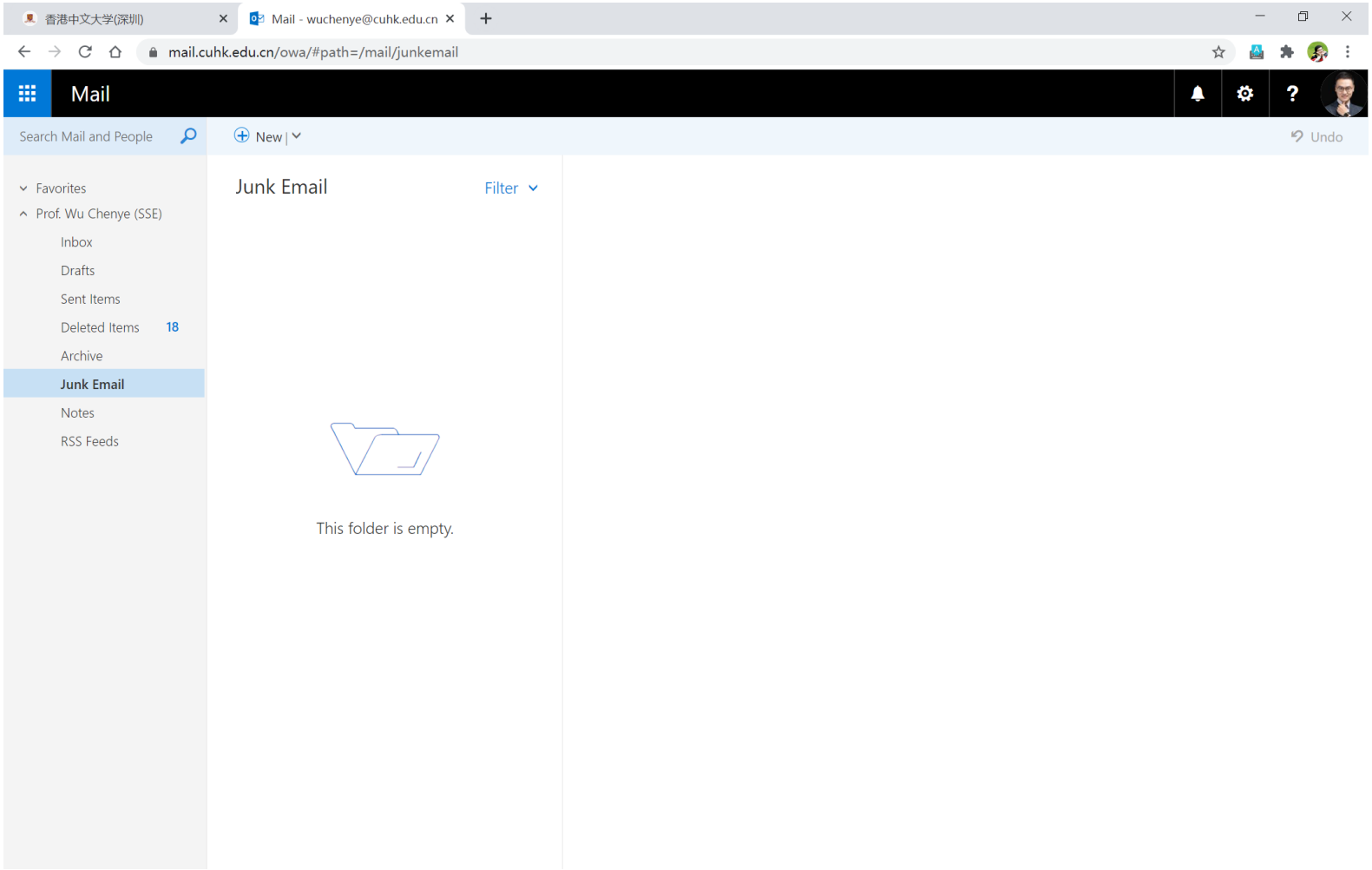
Data Mining Techniques

- Pattern discovery, association, correlation, and causality analysis (descriptive)
- Classification and prediction (predictive)
- Cluster analysis (descriptive)
- Outlier detection (descriptive)
- Trend and deviation: regression analysis (predictive)
- Sequential pattern mining, periodicity analysis (descriptive)

Classification: Definition

- Given a collection of instances (*training set*)
  - Each instance contains a set of *attributes*, one of the attributes is the *class*.
- Task: 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 1



Classification: Example 2

- Opinion mining / Sentiment Analysis

- Goal: Given a set of product reviews on some commercial Web sites (e.g., Amazon.com), classify them into either positive or negative and then summarize each category of reviews.

- Approach:
- Use the labeled reviews to build a classification model
- Classify the unlabeled reviews using the model, and summarize each category of reviews using pattern discovery approaches.

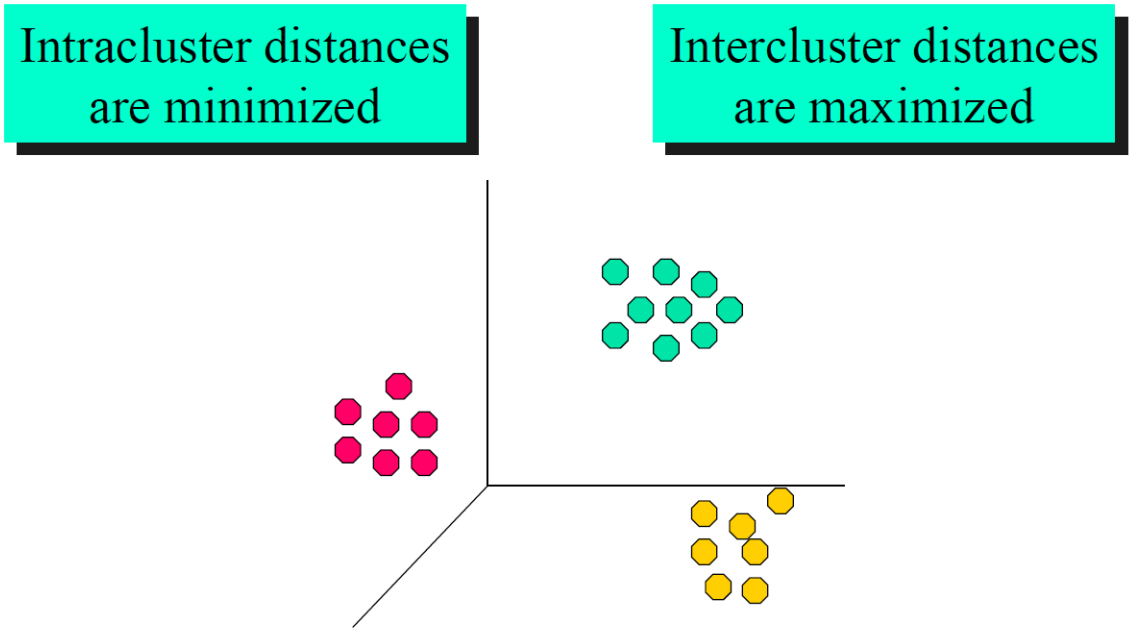
Clustering Definition

- 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 3D space



Common Belief

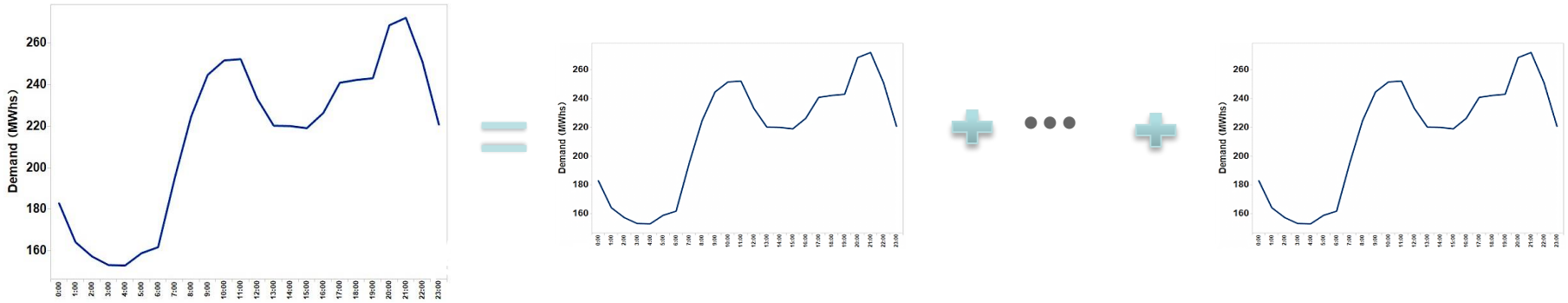
- In power system operation, the critical peak during the year drives the operational cost high.
- Hence, suppose the State Grid is asking you to distinguish bad consumers from good consumers, how would you suggest?

Common Belief

- People tend to believe those large consumers (consume large quantity of energy) are the bad consumers.
- What are the inherent assumptions behind this belief?

Underlying Assumption

- We thought that the energy consumption patterns are similar.



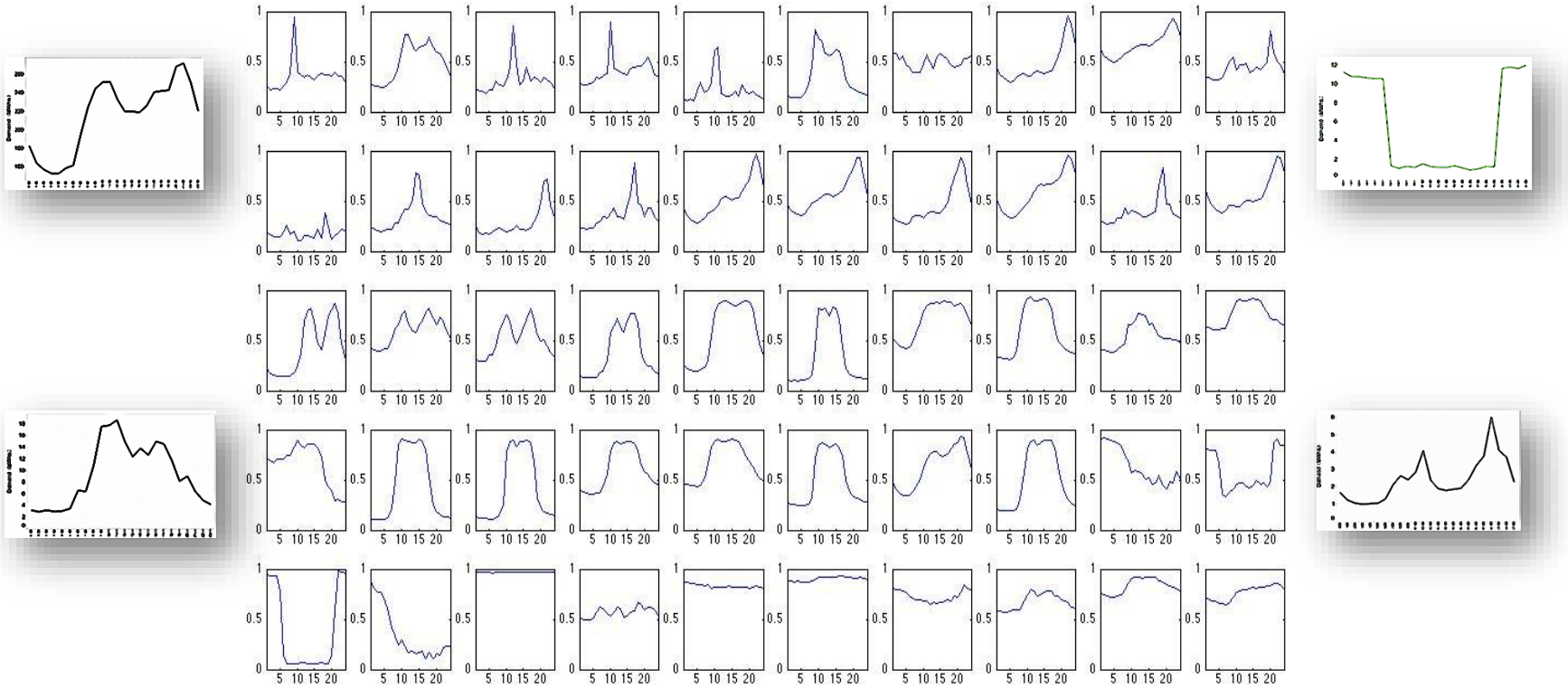
- Based on this assumption, of course, large consumers are bad consumers.

- Hence, we should charge them a higher electricity rate.

Is this assumption true?

How to validate your conclusion?

Ask the Data!

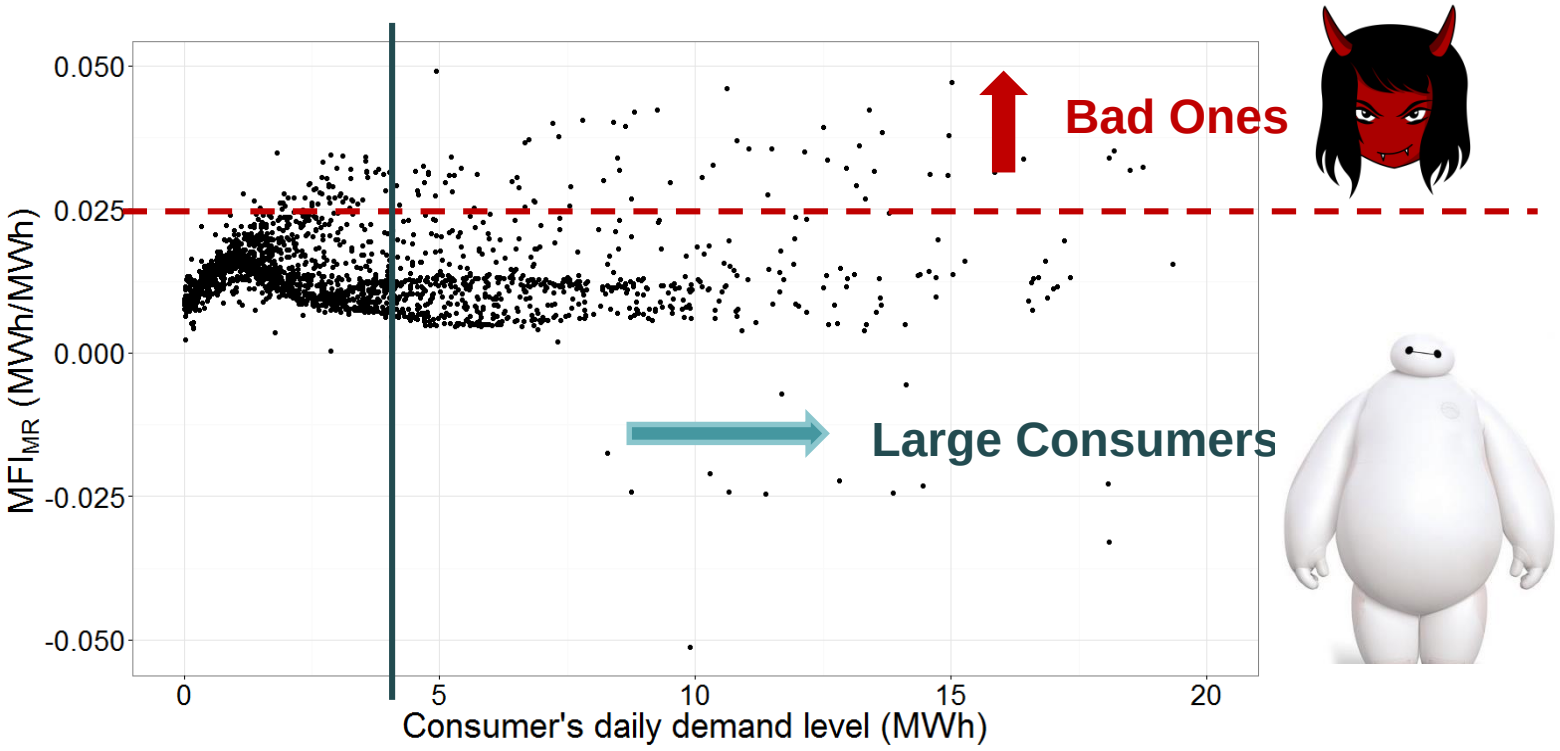


Good consumers v.s. Bad consumers

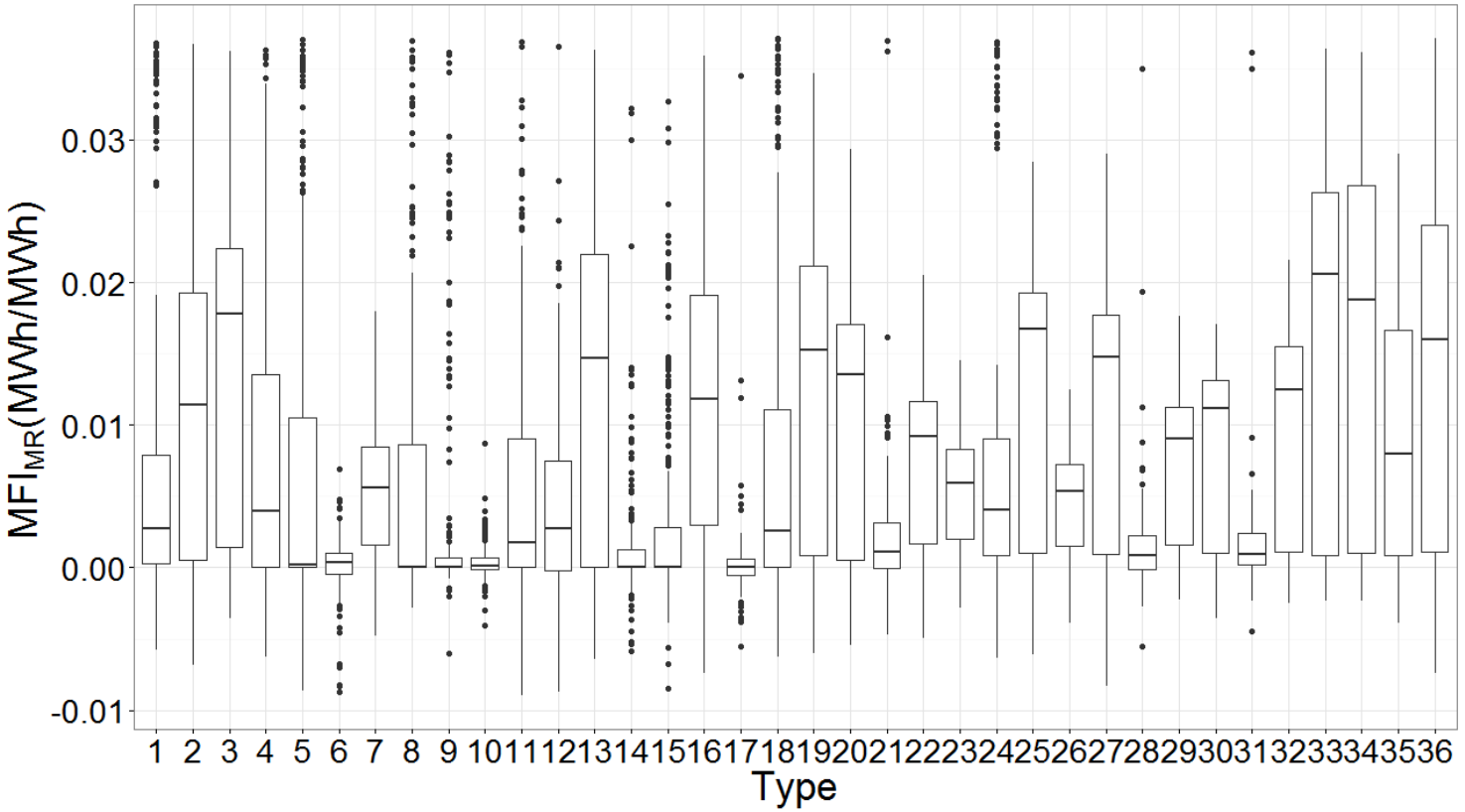
- To determine if one consumer is good or bad, we need to specify the system operator's marginal cost to serve the consumer.
- This marginal cost is uniquely determined by the consumer's load shape!
- Hence, instead of evaluation the consumer's total energy consumption, the system operator needs to evaluate its load profile!
- Yu et al. have developed such an index to use the load profile to determine the marginal serving cost in [1].

[1] Yu et al, Good consumer or bad consumer: Economic information revealed from demand profiles, IEEE Trans. On Smart Grid.

Good consumers v.s. Bad consumers



Good consumers v.s. Bad consumers



Acknowledgement: Most of the course materials are shared by Prof. Gabriela Hug from ETH.

Regression

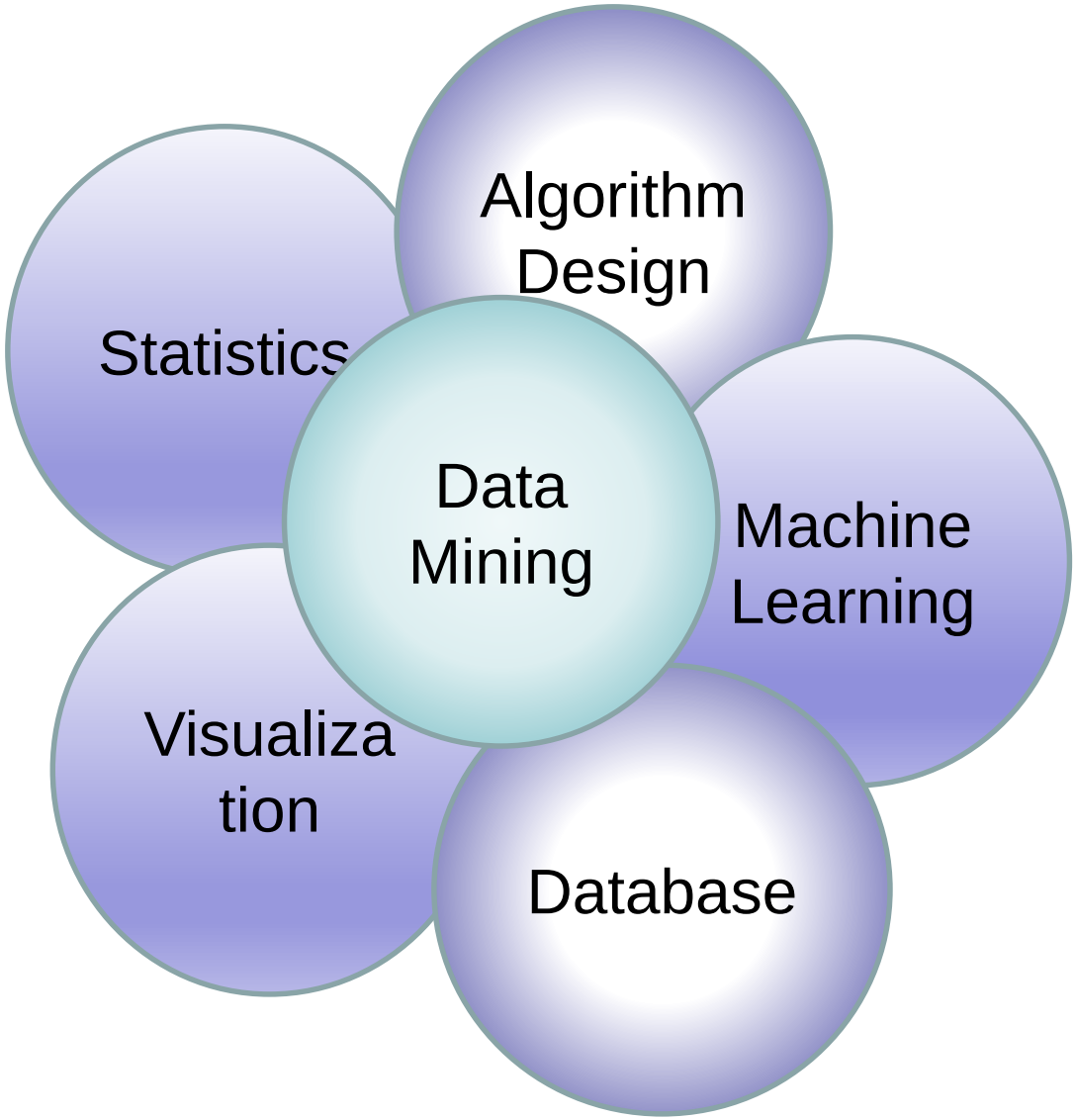
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.

- 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.

Data Mining Processes

- 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, …
- Choosing functions of data mining
- Summarization, classification, regression, association, clustering, …
- Choosing/designing the mining algorithm(s)
- Pattern evaluation and knowledge presentation
- Visualization, transformation, removing redundant patterns, …
- Use of discovered knowledge

An Interdisciplinary Area



Challenges in Data Mining

- Mining methodology
  - Performance: efficiency, effectiveness, and scalability
    - Most data mining problems are NP-Hard (Eg, frequent pattern, decision tree, …)
    - Heuristic based search space pruning
    - Ensembles of local optimal solutions
    - Parallel, distributed, and incremental mining
  - Mining different kinds of knowledge from diverse data types
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- Applications and social impacts
  - Protection of data security, integrity, and privacy, social network analysis, recommender systems, information extraction, …
- User interaction
  - Expression and visualization of data mining results, interactive mining

Where to Find References

- Data Mining and Knowledge Discovery from Data
  - Conferences: **ACM-SIGKDD**, **IEEE-ICDM**, **SIAM-DM**, PKDD, PAKDD, ...
  - Journals: **ACM TKDD**, **DMKD**, KAIS, SAM, …
- AI & Machine Learning
  - Conferences: **ICML**, **AAAI/IJCAI**, **NIPS**（**NeurIPS**）, **COLT**,...
  - Journals: **J. Machine Learning Research**, **Machine Learning**, **Artificial Intelligence**, **J. Artificial Intelligence**, ...
- Web, IR, NLP
  - Conferences: **SIGIR**, **WWW**, **ACL**, **CIKM**, **WSDM**, APWeb, ...
  - Journals: **ACM TOIS**, **ACM TWEB**, **ACM TOIT**, WWW J., …
- Bioinformatics
  - Journal: **Bioinformatics**, ...
- Statistics
  - Conferences: **Joint Stat. Meeting**, ...
  - Journals: **Annals of statistics**, ...

Summary

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A knowledge discovery process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Data mining functionalities: association, classification, clustering, outlier and trend analysis, etc.

Takeaway Message

- Data mining has shown promise but needs much more further research.
- We stand on the brink of great new answers, but even more, of great new questions. – *Matt Ridley*

Copyright Claim

- Some slides/materials used in this class are borrowed from some publicly available lectures on data mining/machine learning/statistics, and the following are some examples:
  - Lectures taught by Prof. Jiawei Han from UIUC, Prof. Pang-Ning Tan from Michigan State University, Prof. Andrew Ng from Stanford University, Prof. Geoffrey Hinton from University of Toronto, Prof. Ani Adhikari from UC Berkeley, Dr. Prof. Hsuan-Tien Lin at National Taiwan University, and Prof. Tom Mitchell from CMU, Prof. Jianyong Wang from Tsinghua University
- The copyright of these borrowed/revised slides still belongs to the original authors, and here we thank them for making their slides publicly available!