Data Mining

Chapter 1: Introduction

Yunming Ye, Baoquan Zhang

School of Computer Science

Harbin Institute of Technology, Shenzhen

Agenda

- Overview of this Course
- What is Data Mining
- Data mining: Tasks & Applications
- Structured Formulation of Data

1.1 Overview of this Course

TA Information

- 贾鹏飞
- Email: jiapengfei@stu.hit.edu.cn
- Tel:18463102736

- 于德民
- Email: deminy@stu.hit.edu.cn
- Tel: 15898867086

Course Homepage

- QQ discussion group
- Group number：649742601
- Group name： 2024春-数据挖掘

Textbook

Data Mining: Concepts and Techniques (Third Edition) by J. Han and M. Kamber, Morgan Kaufmann Publishers, 2012

Other Recommended Books

- Ian，Goodfellow, Yoshua，Bengio, Aaron Courville. Deep Learning. The MIT Press, ISBN: 978-0262035613, USA 2016.
- Pang-Ning Tan，Michael Steinbach，Anuj Karpatne ，Vipin Kumar. Introduction to Data Mining（Second Edition, 英文影印版）. 机械工业出版社, ISBN：9787111637882, 2019.

Course Coverage

第 9 周	1	5-6	数据挖掘的基本概念和流程介绍；数据的结构化表示；数据理解的主要问题及方法。	叶允明
第 10 周	1	5-6	数据预处理的主要问题及方法；回归与分类预测概述；线性回归。	叶允明
第 10 周	3	5-6	支持向量机概述；硬间隔线性支持向量机。	叶允明
第 10 周	6	5-6	软间隔线性支持向量机；非线性支持向量机。	叶允明
第 11 周	1	5-6	决策树归纳；集成学习方法。	叶允明
第 11 周	3	5-6	聚类方法概述；经典聚类算法（一）。	叶允明
第 12 周	1	5-6	经典聚类算法（二）；离群点检测概述；经典离群点检测算法。	叶允明
第 12 周	3	5-6	推荐系统概述；经典推荐算法。	叶允明
第 13 周	3	5-6	深度神经网络（一）：CNN 系列。	张保权
第 14 周	1	5-6	深度神经网络（二）：RNN 系列。	张保权
第 14 周	3	5-6	深度神经网络（三）：Transformer 系列。	张保权
第 15 周	3	5-6	深度神经网络（四）：GAN 系列。	张保权
第 15 周	7	5-6	深度神经网络（五）：扩散模型系列。	张保权
第 16 周	1	5-6	关系图学习方法专题。	张保权
第 16 周	3	5-6	小样本学习、大模型技术专题。	张保权

Class Format and Requirements

- Pre-requisites : data structure and algorithms, a good working knowledge of Python or Java (for the project implementation).
- Lecture & discussion
- Final grade will be determined as follows:
  - Homework
  - Final exam

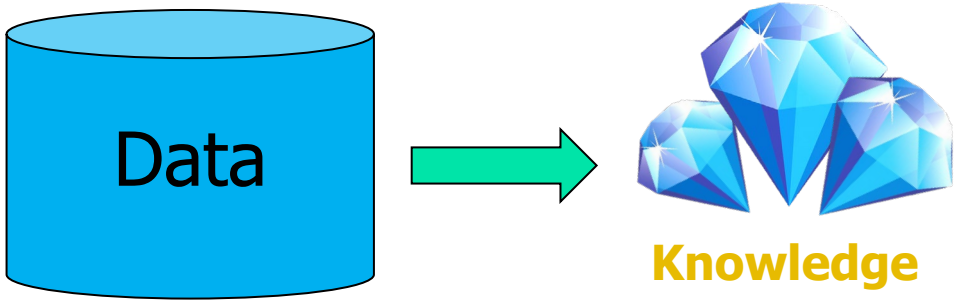
What can you learn from this course?

- How to design and implement a data mining project in real applications
  - Data mining as a process or workflow
  - Data mining software tools
- Classical & advanced data mining algorithms
  - Including classical and new algorithms
- Basic ideas in data mining research

1.2 What is Data Mining?

Necessity Is the Mother of Invention

- Data explosion problem
  - Automated data collection tools, widely used database systems, Internet and WWW applications ……
- We are drowning in data, but starving for knowledge!



客户名	年龄	性别	月入账	月消费	…
张小明	26	男	30k	5k	
李小红	29	女	25k	15k	
……					

“年龄在25-30、没有入账20-30k、消费支出5k-10k，则很可能购买A基金产品”

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
  - Apply extracted knowledge for decision making

How to implement a data mining project?

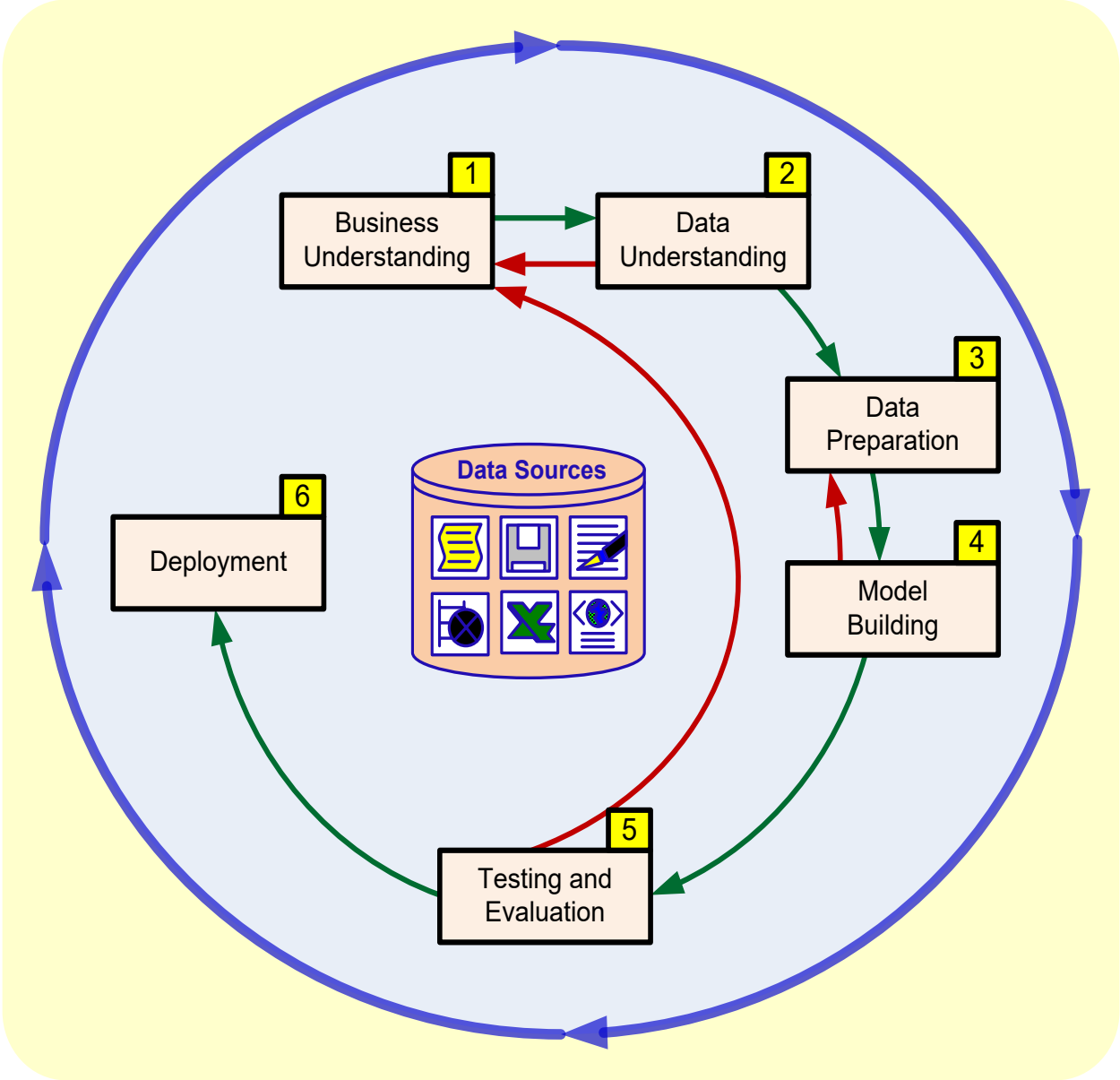
Example: weather prediction

- Predicting the mean temperature of next day
  - daily climate time series data

Date	Mean Temp	Humidity	Wind Speed	Pressure
...	...	...	...	...
2017-01-02	7.40	92.00	2.980	1017.80
2017-01-03	7.17	87.00	4.63	1018.67

How to Implement a Data Mining Project

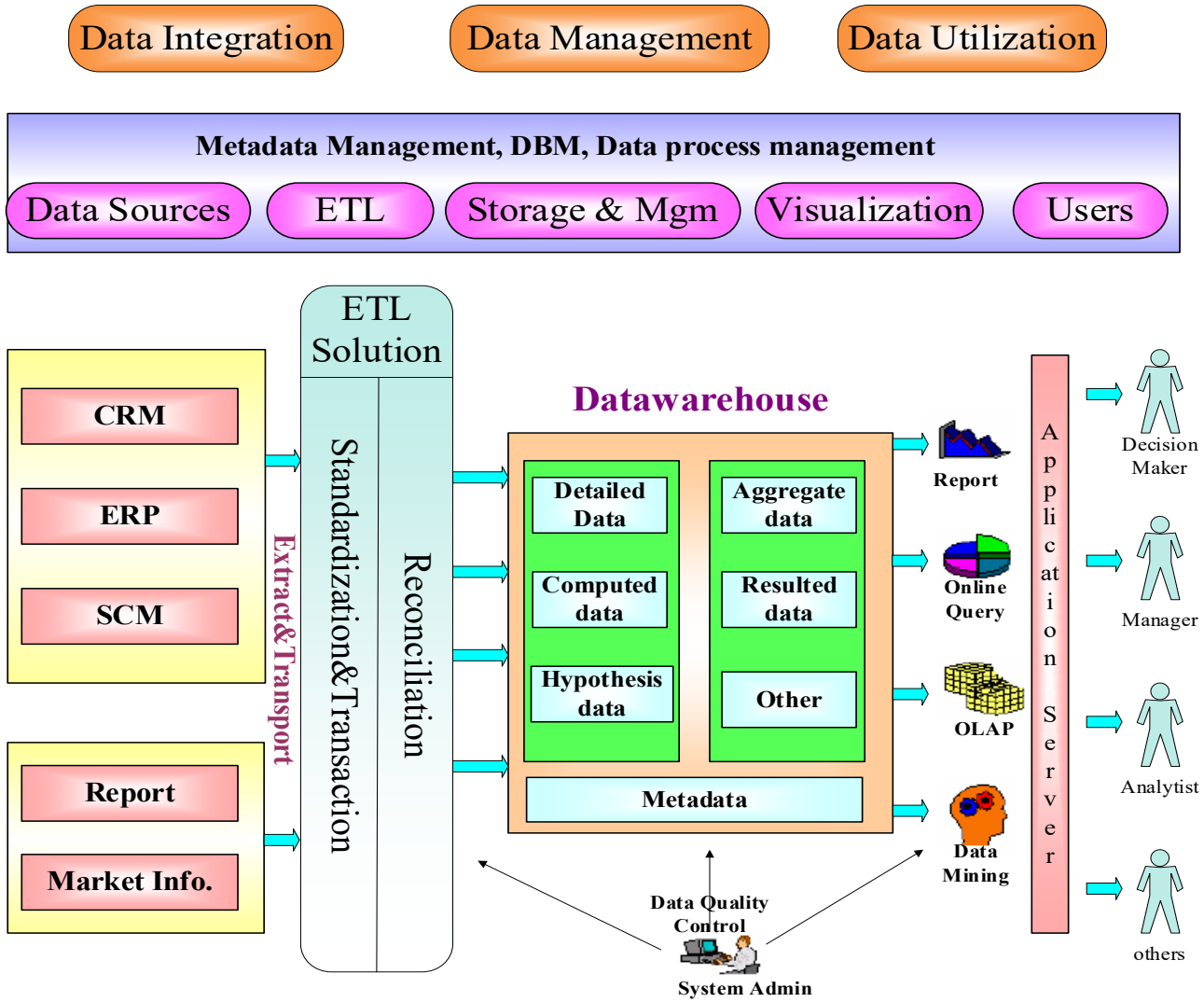
- Process model for practical DM project
  - Workflow
  - Iterative

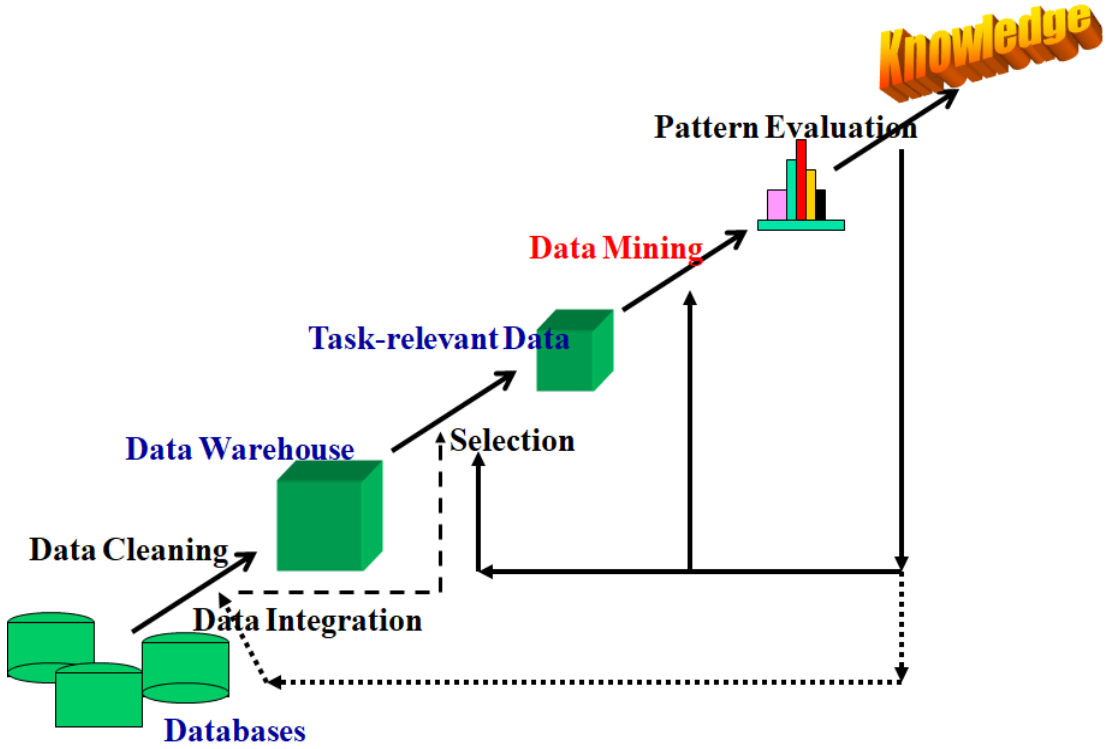


Data Mining: On What Kinds of Data?

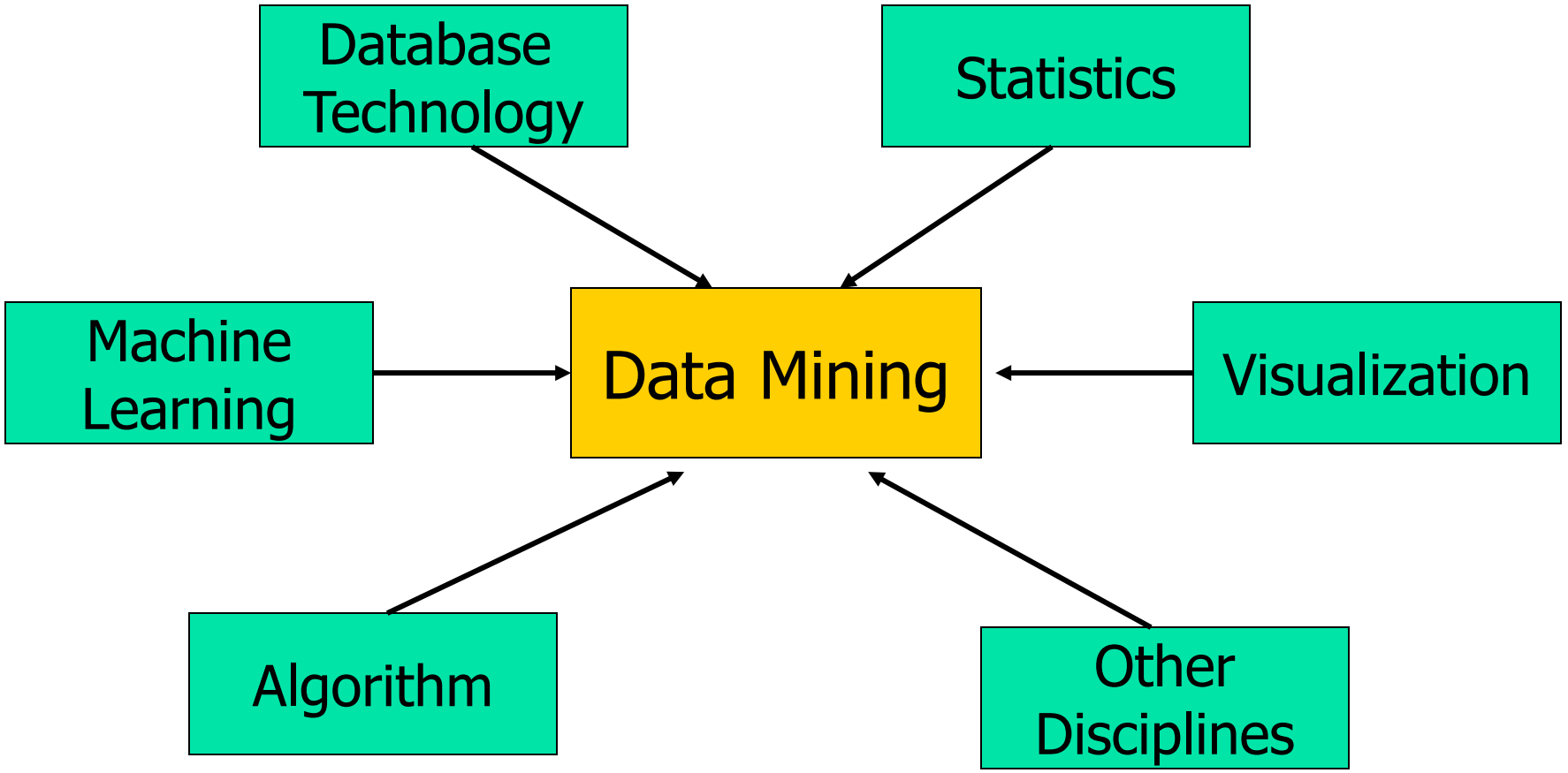
- Traditional database and appllications
  - Relational database, data warehouse, transactional database
- Advanced database and advanced applications
  - Text databases and the World-Wide Web
  - Multimedia data
  - Data streams and time-series data, sequence data (incl. biosequences)
  - Spatial data and spatiotemporal data
  - Graphs, social networks and link databases

The role of Data Mining System in Business Intelligence Platform





Data Mining: Confluence of Multiple Disciplines



Major Research Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Handling noise and incomplete data
  - Incorporation of background knowledge
  - Parallel, distributed and incremental mining methods
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
- Applications and social impacts
  - Domain-specific data mining
  - Protection of data security, integrity, and privacy

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)

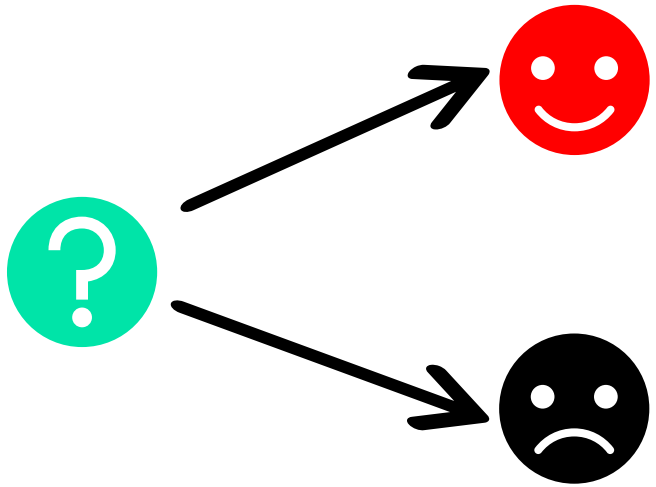
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

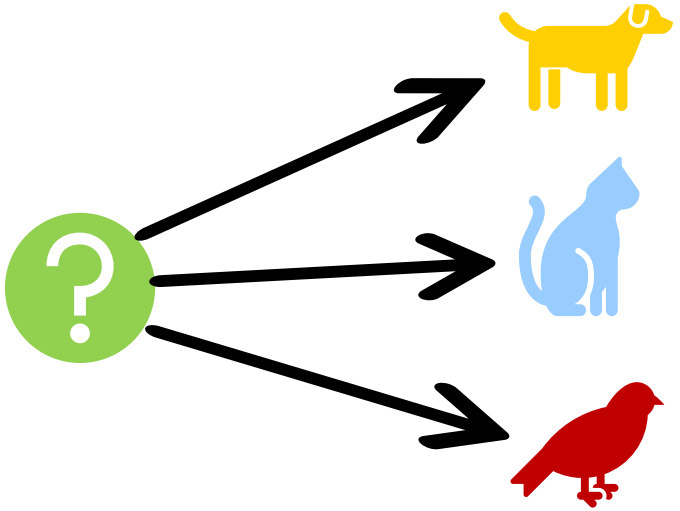
1.3 Data Mining: Tasks & Applications

Classification

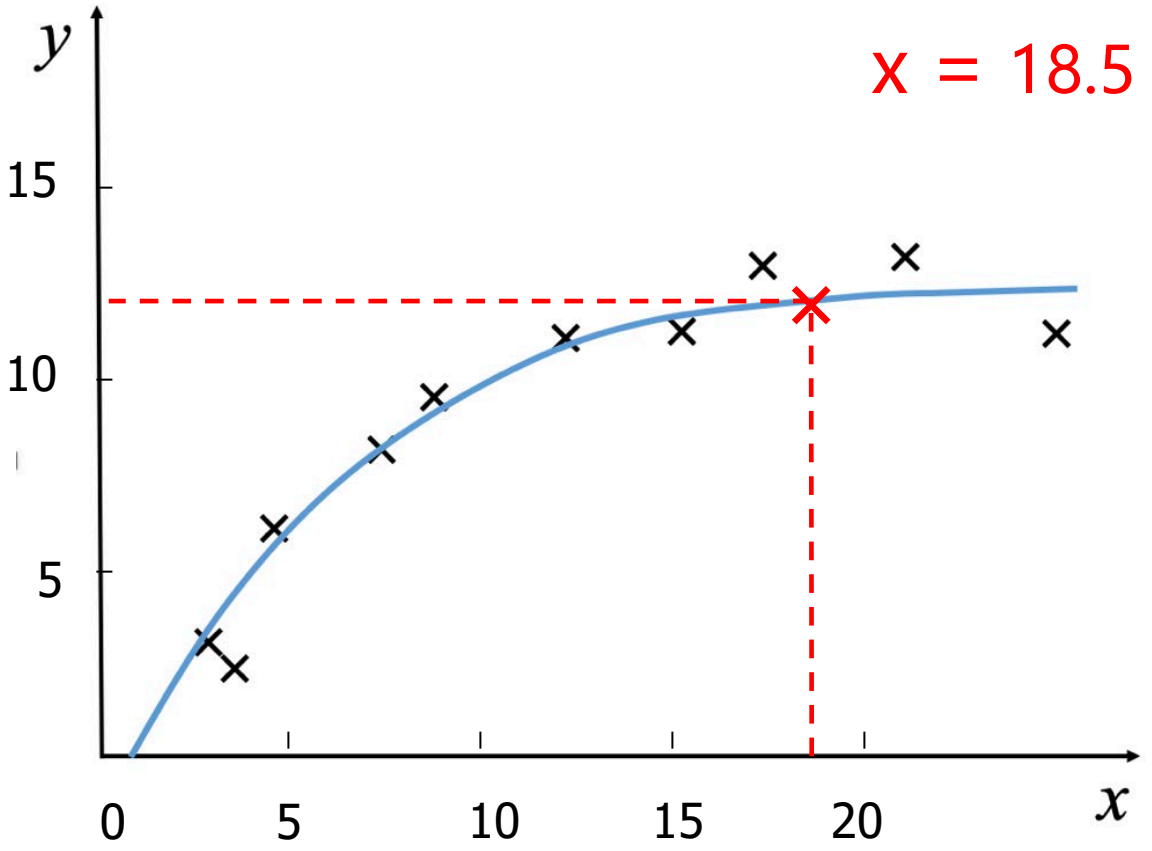
来深圳出差，组织深圳和北京的小伙伴们一起陶陶居了一把，多么美好的记忆啊。

毕竟是深圳，广式点心一类的还是比不上广州，深圳同事考虑再三推荐陶陶居，于是我们就……





Regression



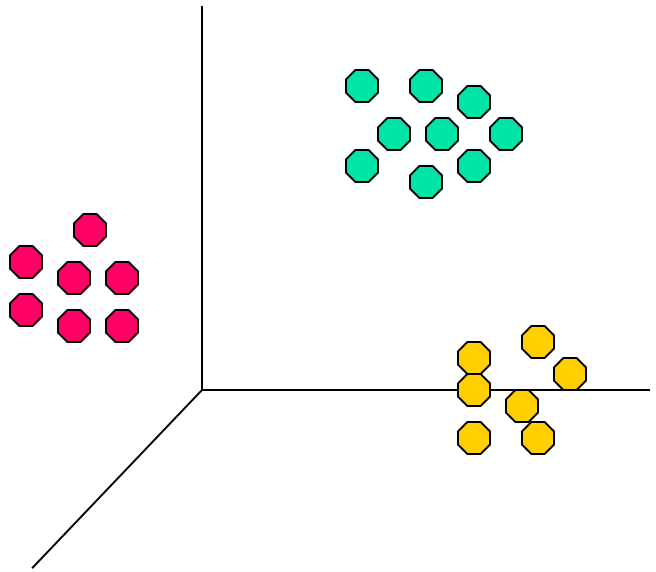
x = 18.5 , y = ?

Clustering Definition

- Given a set of data points, find 'high-quality' clusters

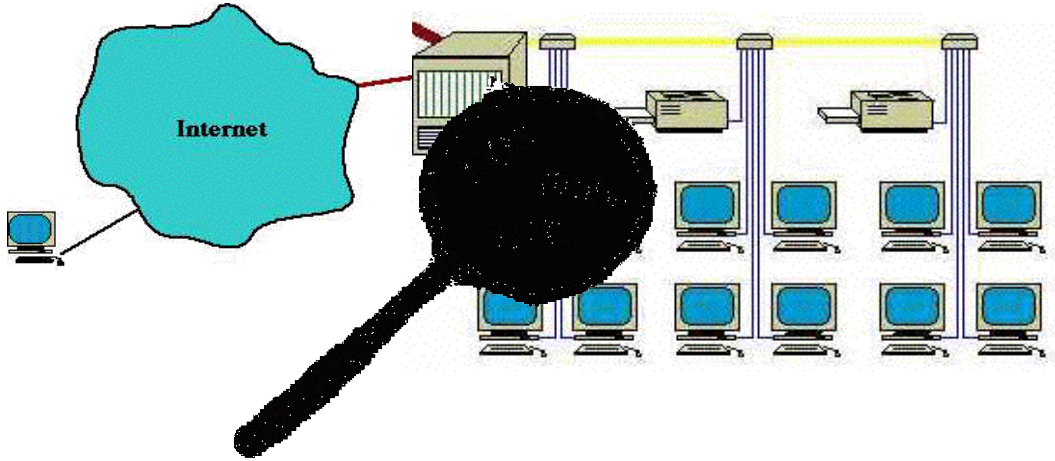
Intracluster distances are minimized

Intercluster distances are maximized



Outlier/Anomaly Detection

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



Association Rule Discovery

- 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}

Recommendation System: Collaborative Filtering

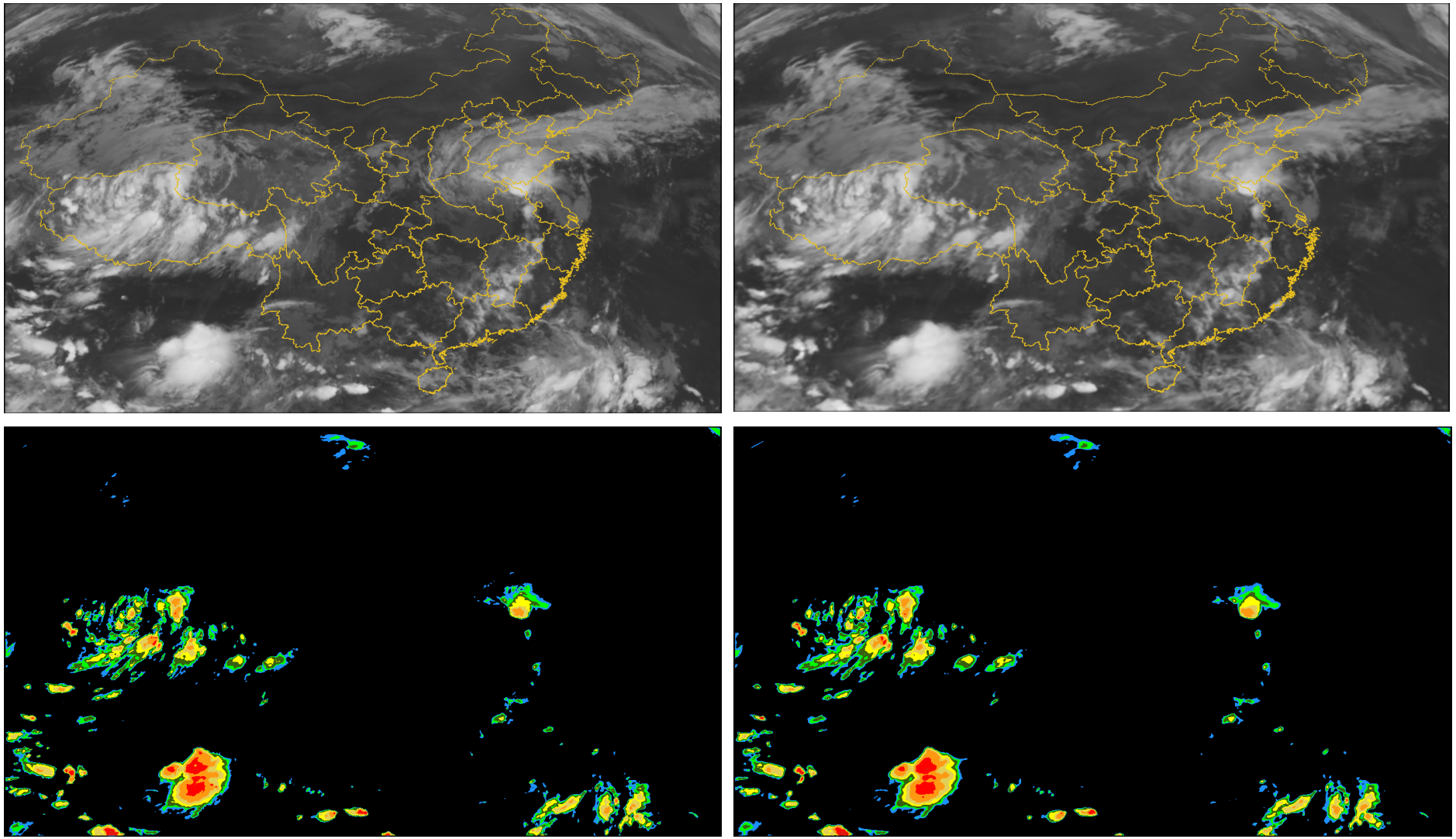
	Book1	Book2	Book3	Book4	Book5	Book6	……
User1							
User2							
User3							
User4							
User5							
User6	?	?		?	**?**	?	?

Other Advanced Tasks

- Structured prediction

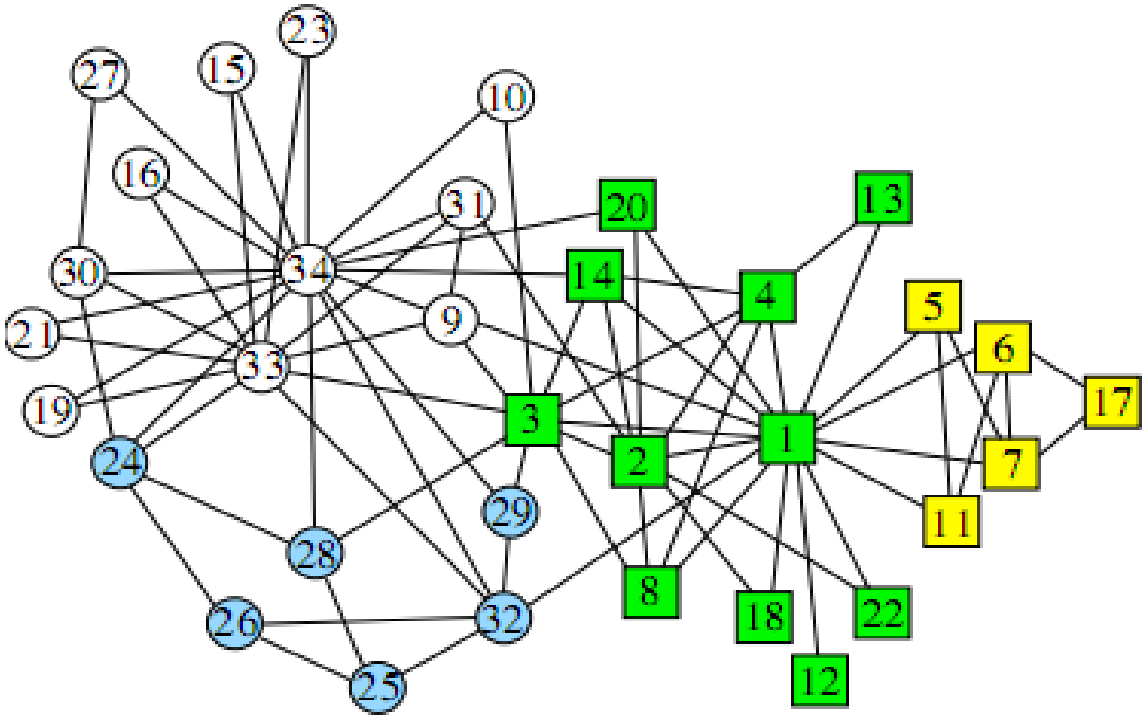
Ground truth

Prediction result



Other Advanced Tasks

- Graph mining tasks



Why 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, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis
  - ……

1.4 Structured Formulation of Data

What is Data?

- Data: observation and measurement of world
- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

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

Structured Data

- Data Structuring:
  - [*Object* – *Attribute* - *Attribute Value*]

- Simplest structured data:

$$\boldsymbol{D} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_i}, \ldots, \boldsymbol{x_m}\},\ (i = 1 \ldots m),$$

$$\text{where } \boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in}\}\ (j = 1 \ldots n),\ x_{ij} \in \boldsymbol{R}$$

Types of Attributes

- Numeric attribute (Continuous)
  - ✓ Has real numbers (floating-point) as attribute values
  - ✓ Sales, temperatures, length, salary
- Categorical attribute (Discrete)
  - ✓ Nominal
    - Examples: ID numbers, eye color, zip codes
    - One-hot encoding
  - ✓ Ordinal
    - Examples: {“A”, “B+”, “C”}
    - represented as (or mapped to) integer variables

Document Data

"Tom is playing ball"

"He is a team coach"

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Color Image Data

(a)

120 121 122 ... 155 156 157 —Channel B

121 187 188 189 ... 217 219 220 —Channel G

121 188 226 227 228 ... 245 248 249 —Channel R

188 227 228 229 ... 246 246 247

... 227 228 230 ... 247 246 246

29 ...

27 33 ...

36 31 131 129 126 ... 126 126 130

41 126 119 118 ... 123 123 123

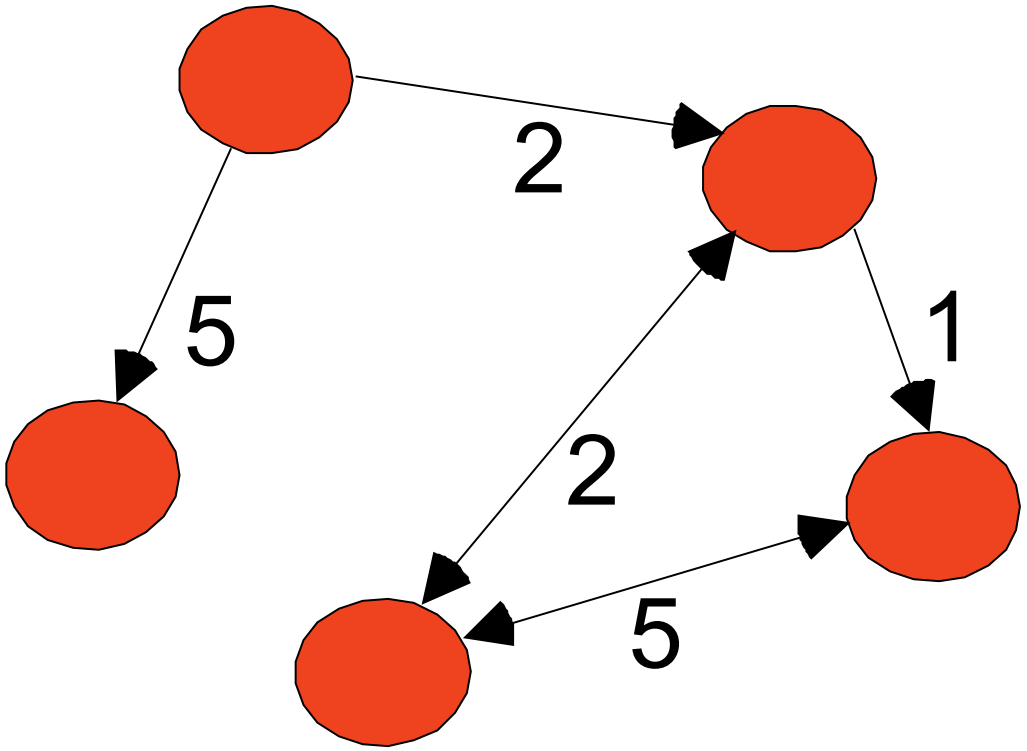
122 119 122 ... 142 145 145

(b)

226, 227, ···, 145	187, 188, ···, 87	120, 121, ···, 82
Channel R	Channel G	Channel B

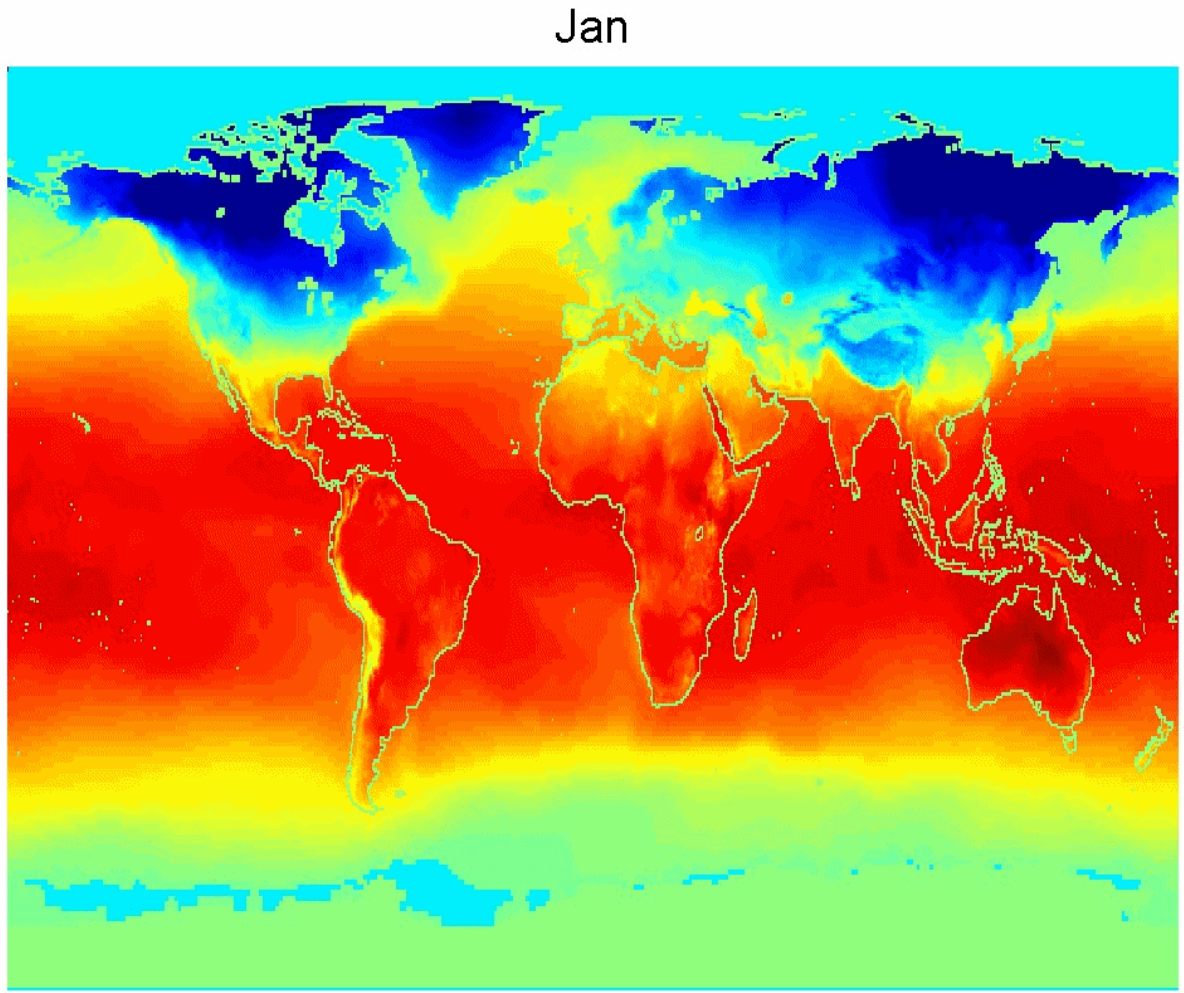
(c)

Graph Data



Spatio-Temporal Data

Average Monthly Temperature of land and ocean



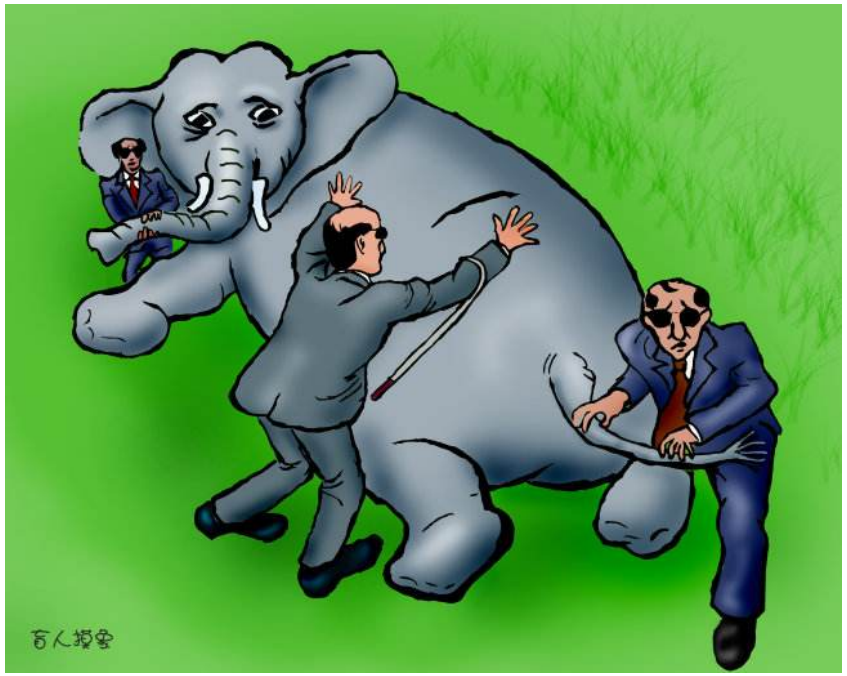
39

Example: Sales Forecasting

- Problem: How to predict product sales effectively?
- What data do you need?
- How to process the data?
- How to modeling?

Core Challenge for Data Mining

- Heterogeneous and uncertain attributes from multiple data sources
- Find **correlation** in big data



图片引自http://blog.sina.com.cn/s/blog_6773d7b90100jnsd.html

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Acknowledgements

- Some text, figures and formulations are from WWW. Thanks for their sharing. If you have copyright claim please contact with me at yym@hit.edu.cn.
- This lecture is distributed for nonprofit purpose.

Thank You for Your Attention

Contact me at: yym@hit.edu.cn

Tel: 26033008, 1376019623

Address: Rm.1402, H# Building