Introduction

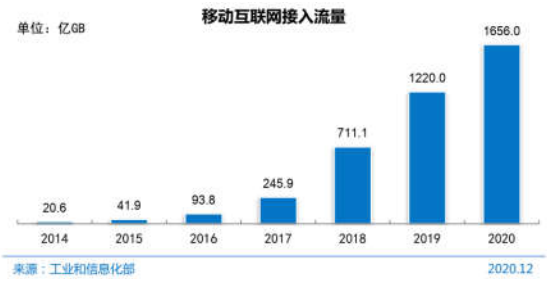
(Data Mining: Method and Application)

徐华

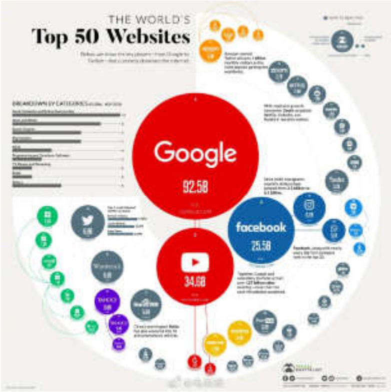
xuhua@tsinghua.edu.cn

Motivation - Background

- The Explosive Growth of Data: from TeraBytes (TB) to PetaBytes (PB): B,KB,MB,GB,TB,PB(Big Data),EB,ZB,YB,DB,NB
  - Data collection and data availability
  - Ex. The changing flow of WebPages in China (2020-12)



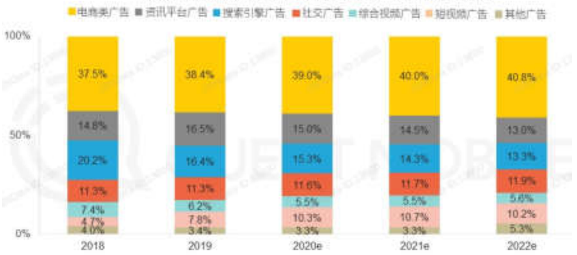
The Webpage Flow in China from CNNIC(2020-12)



Motivation - Background

- About Big Data
  - > 1PB
  - “Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…
  - Features: “5V”
    - Volume (规模大)
    - Variety （种类繁多）
    - Velocity （速度快）
    - Veracity （不确定性）
    - Value （价值）

2018-2022年中国互联网典型媒介类型广告市场份额分布



注：1、广告形式为互联网媒介投放广告，不包括直播、软植、综艺节目冠名、赞助等广告形式；2、互联网媒介渠道分类以QuestMobile TRUTH分类为基础，部分渠道依据广告形式进行了合并，具体为：1）社交广告、综合视频、短视频广告包含APP与QuestMobile TRUTH一致；2）资讯平台广告包括综合资讯行业、垂直资讯行业如汽车、财经、体育等及浏览器平台；3）电商类广告包括电商平台、生活服务平台行业；4）搜索引擎广告包含搜索引擎平台信息流广告；3、参照公开财报数据结合QuestMobile AD INSIGHT广告洞察数据库进行估算。

Source：QuestMobile AD INSIGHT广告洞察数据库，营销研究院 2020年4月

Motivation - Background

- Features of Big Data

	Banking and Securities	Communications, Media and Services	Education	Government	Healthcare Providers	Insurance	Manufacturing and Natural Resources	Retail	Transportation	Utilities	Wholesale Trade
Volume of Data	🟥	🟥	⬛	🟥	🟨	🟨	🟥	🟧	🟧	🟧	🟦
Velocity of Data	🟥	🟧	⬛	🟧	🟥	🟨	🟥	🟧	🟧	🟧	🟦
Variety of Data	🟦	🟥	🟦	🟥	🟧	🟨	🟧	🟨	🟨	🟨	⬛
Underutilized "Dark Data"	🟦	🟨	🟥	🟥	🟨	⬛	🟥	🟦	🟦	🟨	🟧

Potential big data opportunity on each dimension is:

🟥	Very hot (compared with other industries)
🟧	Hot
🟨	Moderate
🟦	Low
⬛	Very low (compared with other industries)

Motivation-Commercial Viewpoints

- Commercial Viewpoints
  - Data Sources: Web data, e-commerce, purchases at department/grocery stores, Bank/Credit Card, transactions
  - Computers have become cheaper and more powerful
  - Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Motivation – Scientific Viewpoints

- Scientific Viewpoints
  - 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

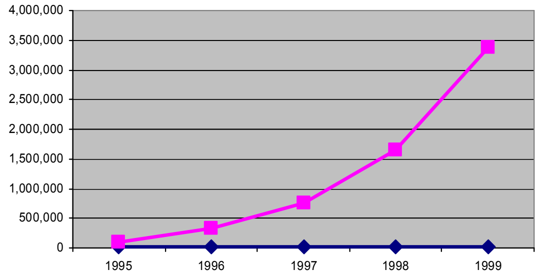
NASA's Perseverance Mars Rover @NASAPersevere · 3h

I love rocks. Look at these right next to my wheel. Are they volcanic or sedimentary? What story do they tell? Can't wait to find out.

#CountdownToMars
mars.nasa.gov/mars2020

Motivation: Why Data mining?

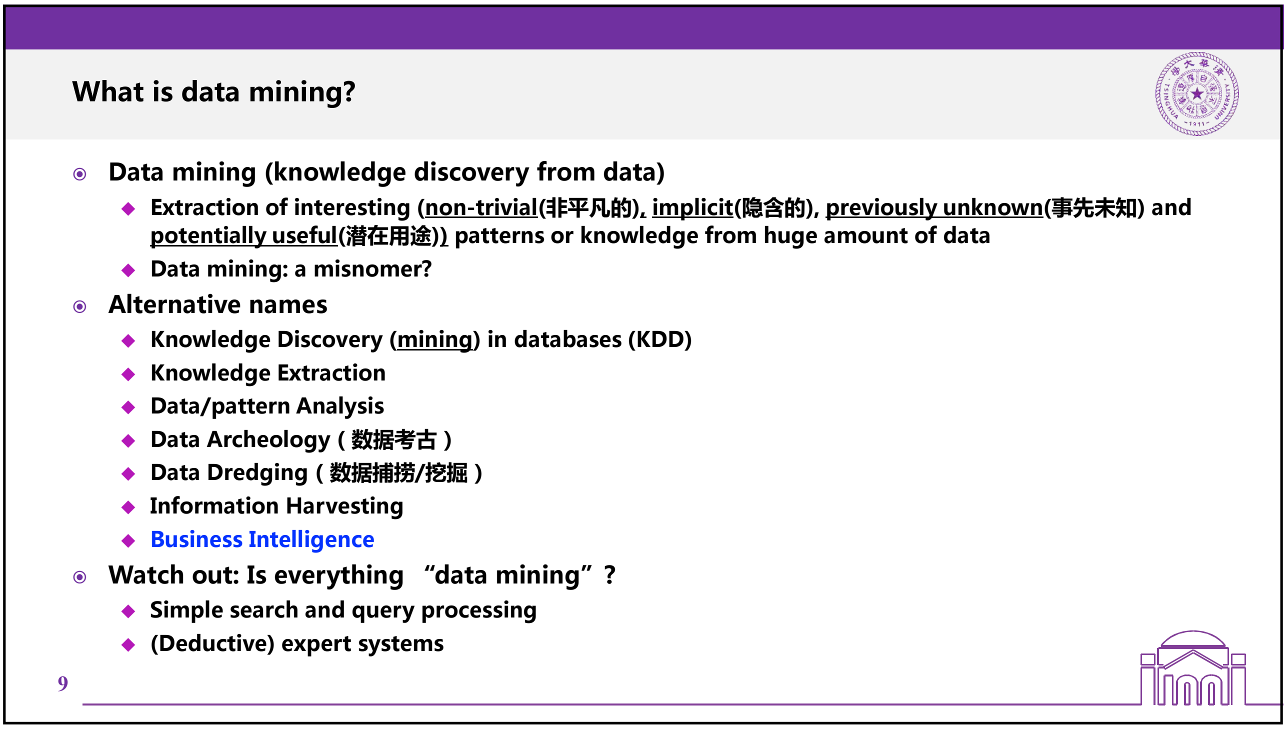
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all. ” We are drowning in data, but starving for knowledge ! ”
- “Necessity is the mother of invention” —Data mining—Automated analysis of massive data sets



From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

Evolution of Database Technology

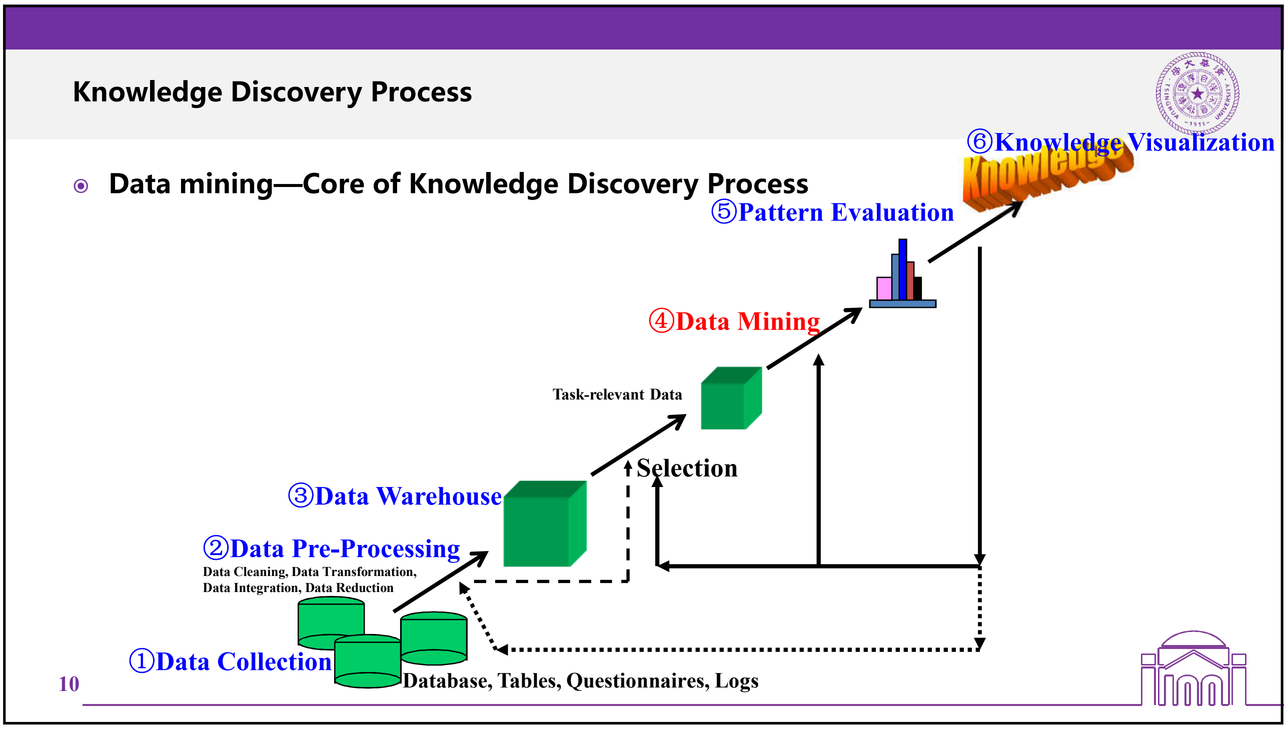
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems



What is data mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial(非平凡的), implicit(隐含的), previously unknown(事先未知) and potentially useful(潜在用途)) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge Discovery (mining) in databases (KDD)
  - Knowledge Extraction
  - Data/pattern Analysis
  - Data Archeology（数据考古）
  - Data Dredging（数据捕捞/挖掘）
  - Information Harvesting
  - Business Intelligence
- Watch out: Is everything “data mining”？
  - Simple search and query processing
  - (Deductive) expert systems

9



Knowledge Discovery Process

- Data mining—Core of Knowledge Discovery Process

⑥Knowledge Visualization

Knowledge

⑤Pattern Evaluation

④Data Mining

Task-relevant Data

Selection

③Data Warehouse

②Data Pre-Processing

Data Cleaning, Data Transformation, Data Integration, Data Reduction

①Data Collection

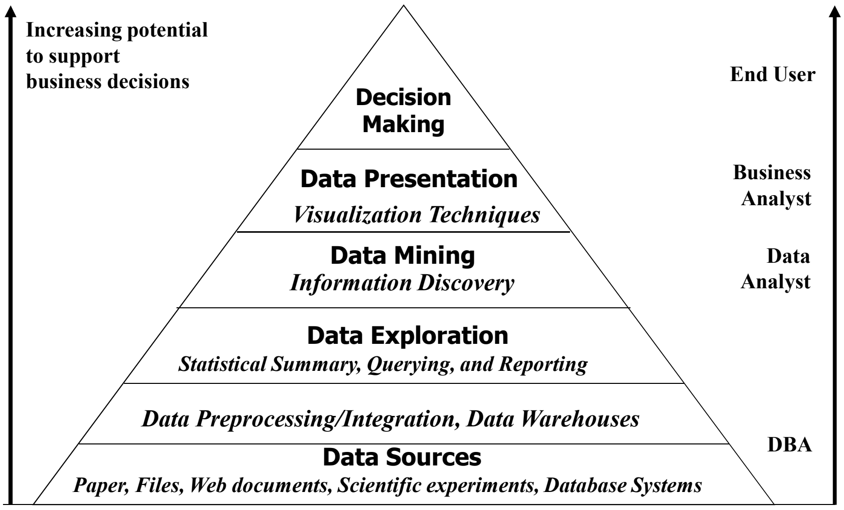
Database, Tables, Questionnaires, Logs

10

Data Mining v.s. KDD

- **Knowledge Discovery in Databases (KDD):** process of finding useful information and patterns in data.
- **Data Mining:** Use of algorithms to extract the information and patterns derived by the KDD process.

Data Mining and Business Intelligence



Increasing potential to support business decisions

Decision Making — End User

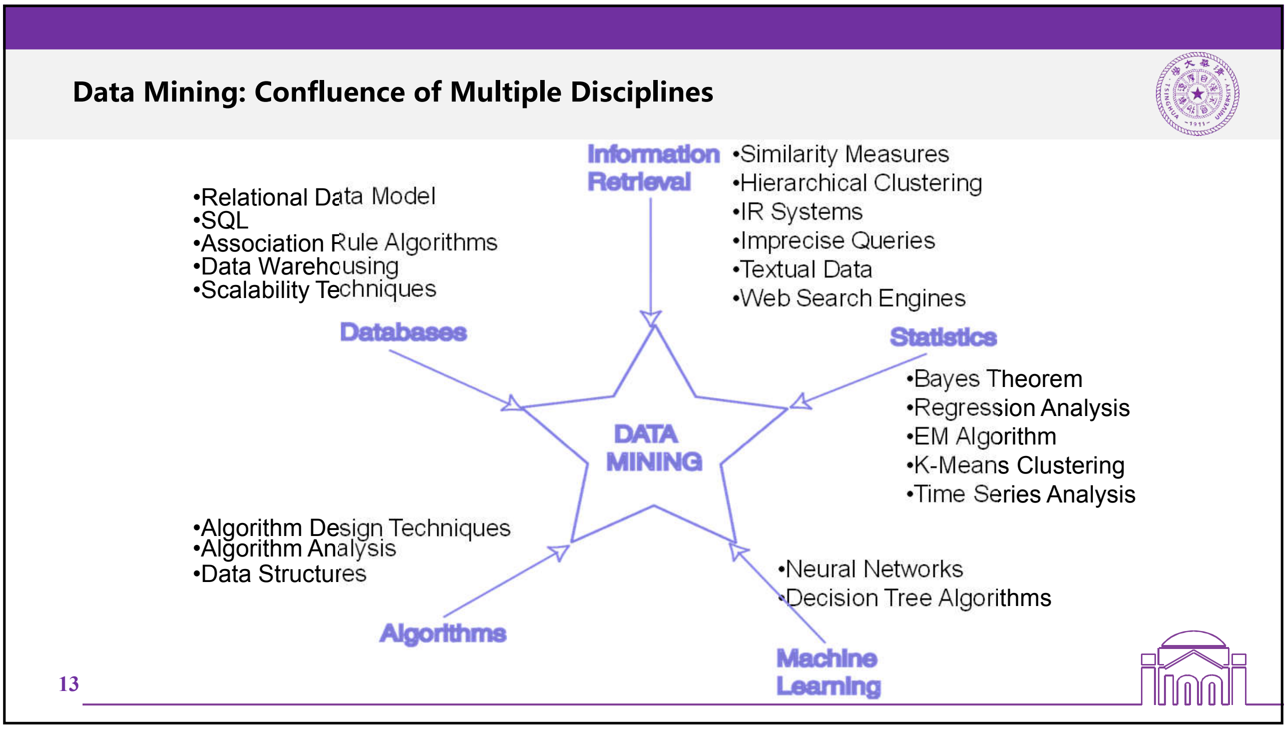
Data Presentation
Visualization Techniques — Business Analyst

Data Mining
Information Discovery — Data Analyst

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses — DBA

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems



Data Mining: Confluence of Multiple Disciplines

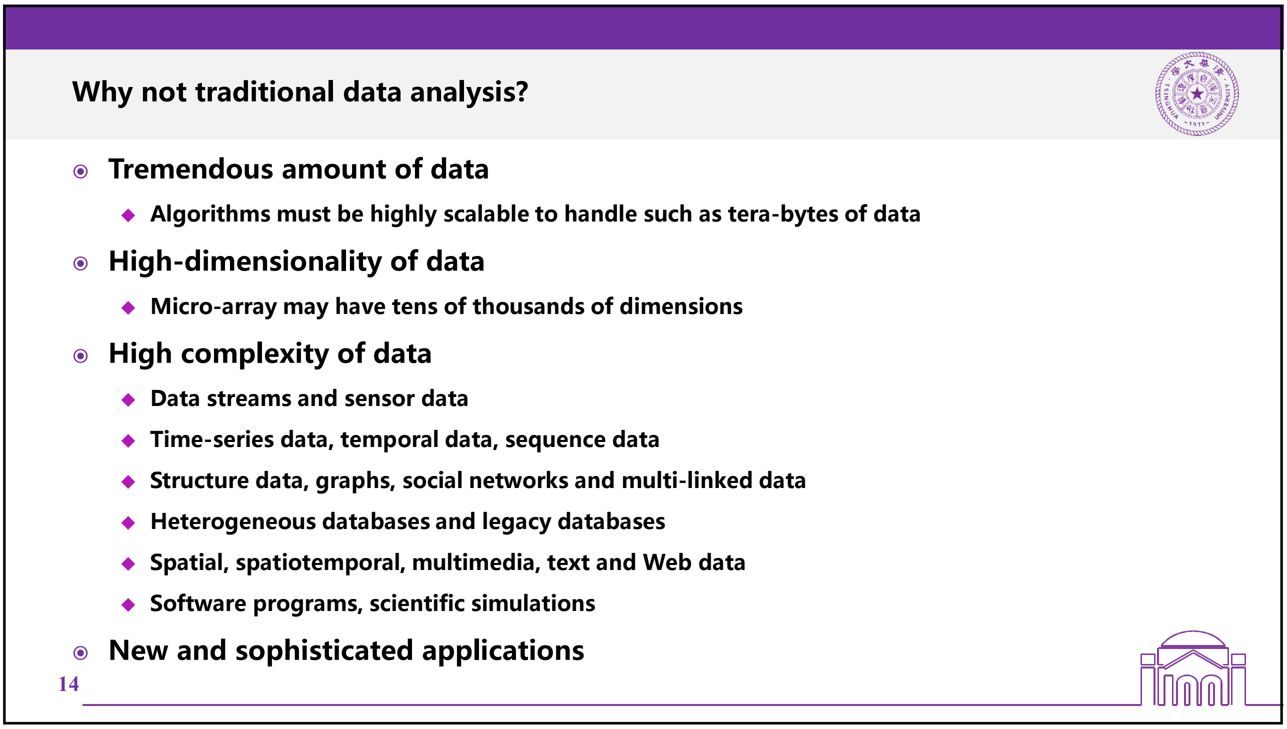
Information Retrieval
- Similarity Measures
- Hierarchical Clustering
- IR Systems
- Imprecise Queries
- Textual Data
- Web Search Engines

Databases
- Relational Data Model
- SQL
- Association Rule Algorithms
- Data Warehousing
- Scalability Techniques

Statistics
- Bayes Theorem
- Regression Analysis
- EM Algorithm
- K-Means Clustering
- Time Series Analysis

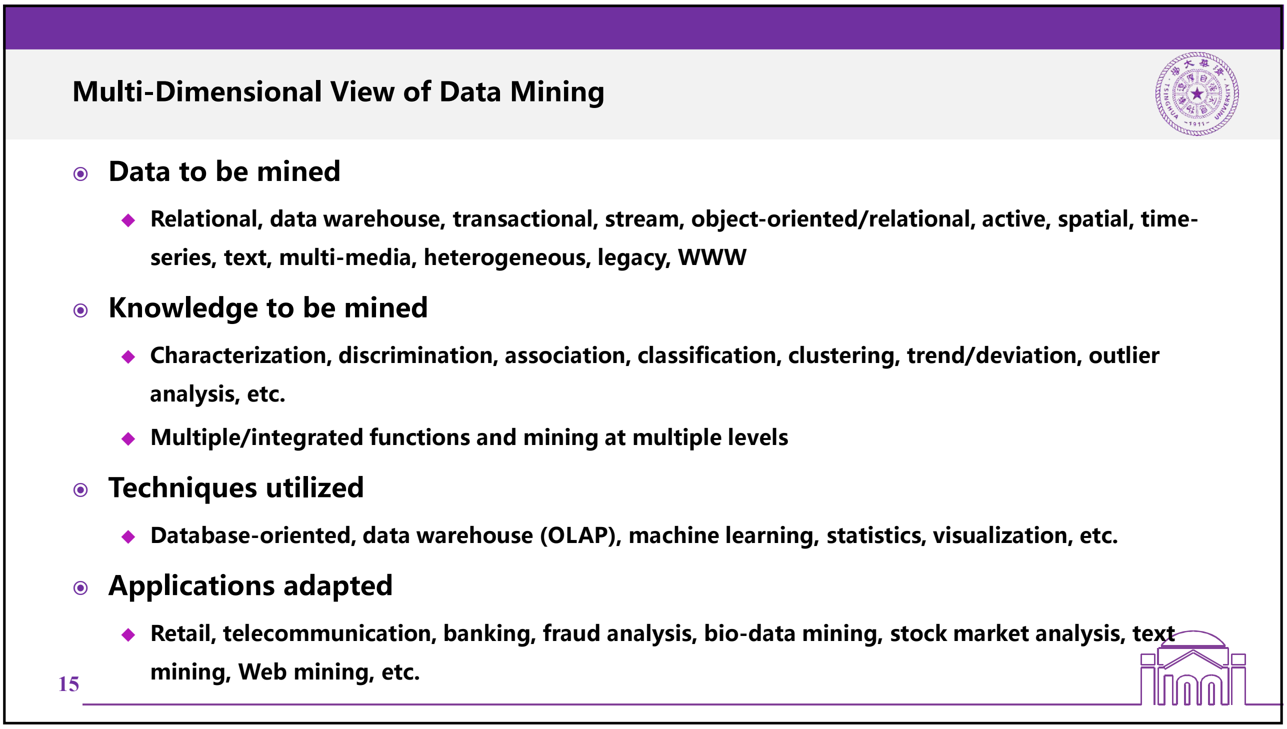
Algorithms
- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

Machine Learning
- Neural Networks
- Decision Tree Algorithms



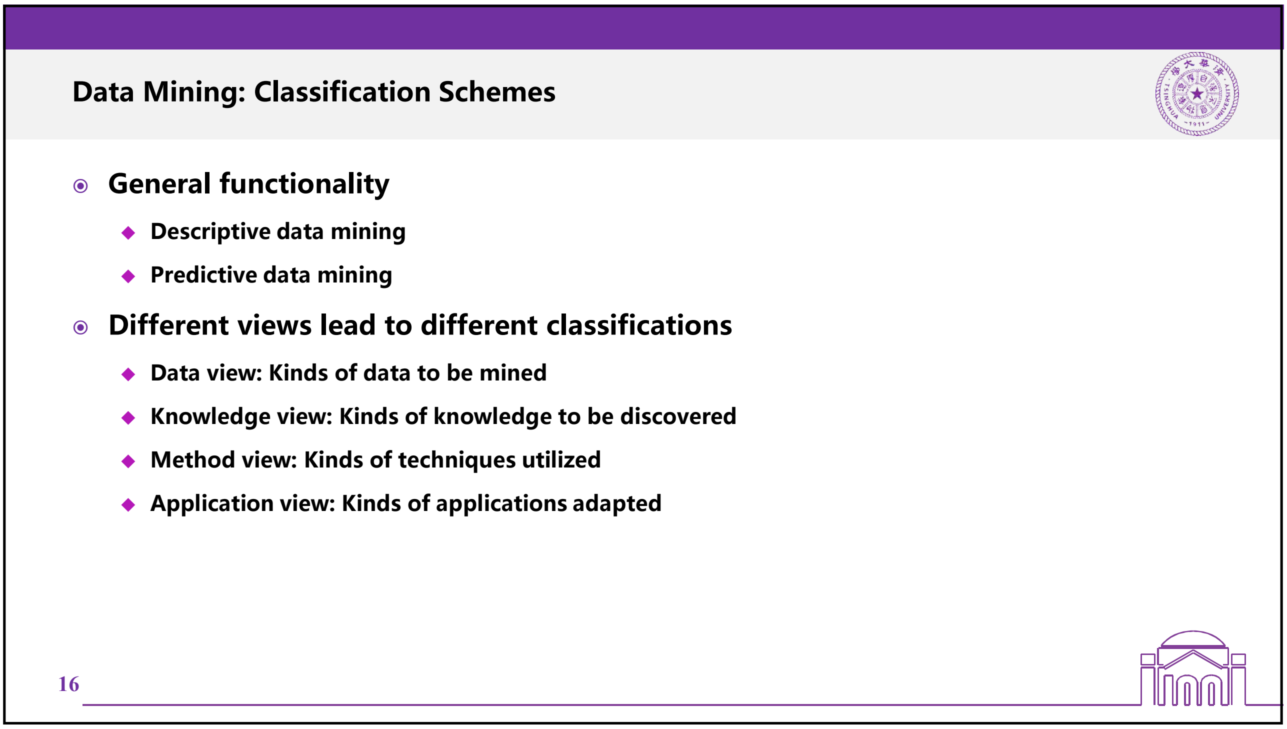
Why not traditional data analysis?

- **Tremendous amount of data**
  - **Algorithms must be highly scalable to handle such as tera-bytes of data**
- **High-dimensionality of data**
  - **Micro-array may have tens of thousands of dimensions**
- **High complexity of data**
  - **Data streams and sensor data**
  - **Time-series data, temporal data, sequence data**
  - **Structure data, graphs, social networks and multi-linked data**
  - **Heterogeneous databases and legacy databases**
  - **Spatial, spatiotemporal, multimedia, text and Web data**
  - **Software programs, scientific simulations**
- **New and sophisticated applications**



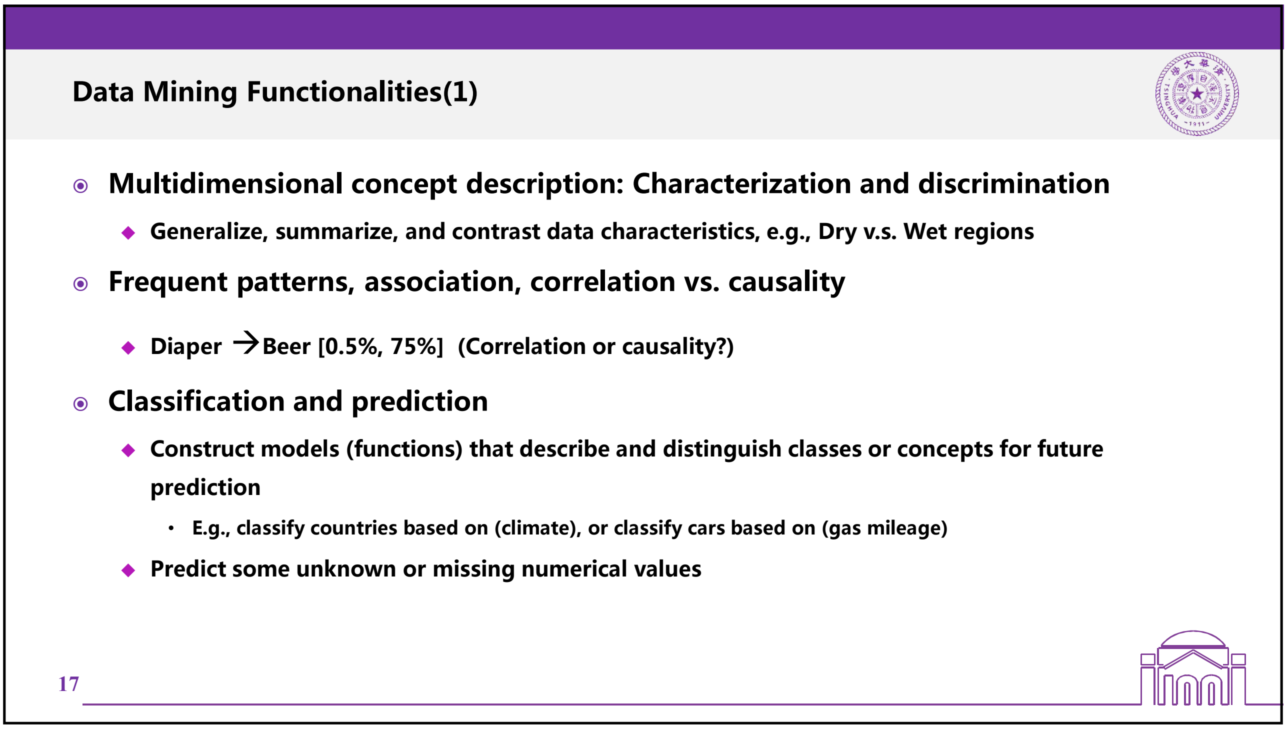
Multi-Dimensional View of Data Mining

- Data to be mined
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- Knowledge to be mined
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- Techniques utilized
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.



Data Mining: Classification Schemes

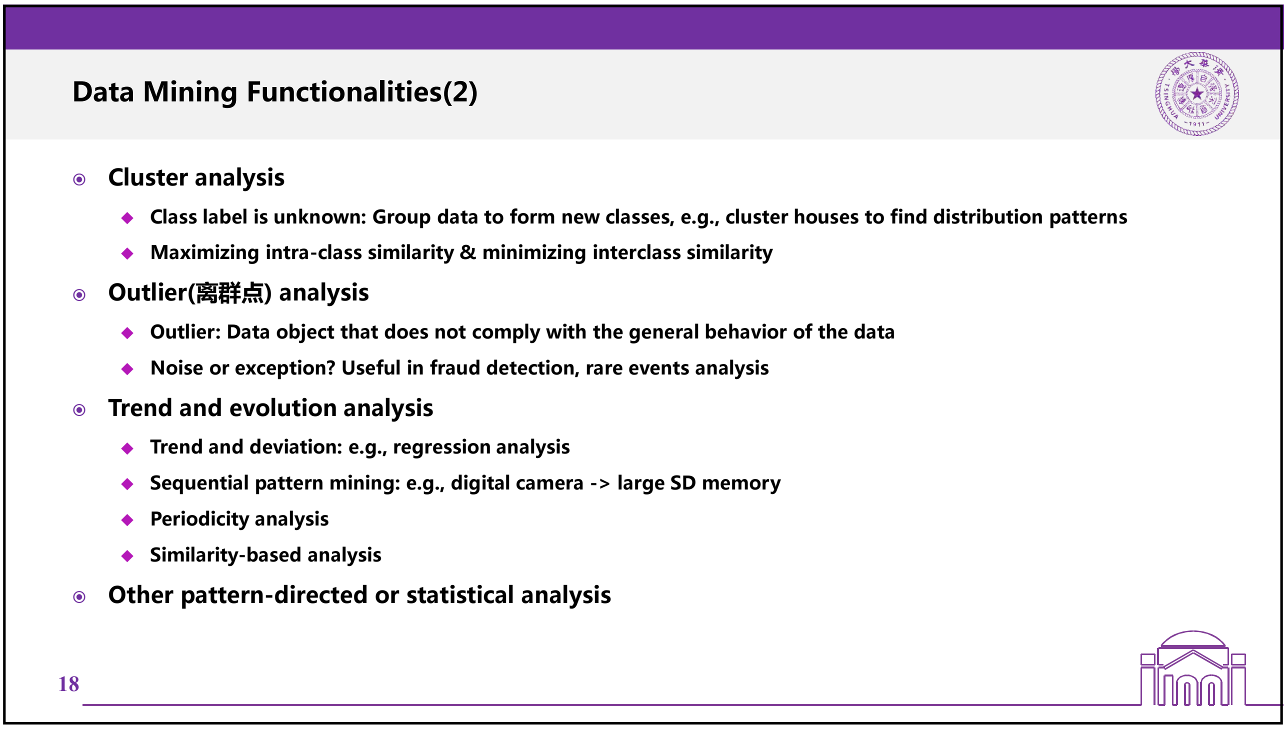
- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted



Data Mining Functionalities(1)

- **Multidimensional concept description: Characterization and discrimination**
  - Generalize, summarize, and contrast data characteristics, e.g., Dry v.s. Wet regions
- **Frequent patterns, association, correlation vs. causality**
  - Diaper →Beer [0.5%, 75%] (Correlation or causality?)
- **Classification and prediction**
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown or missing numerical values

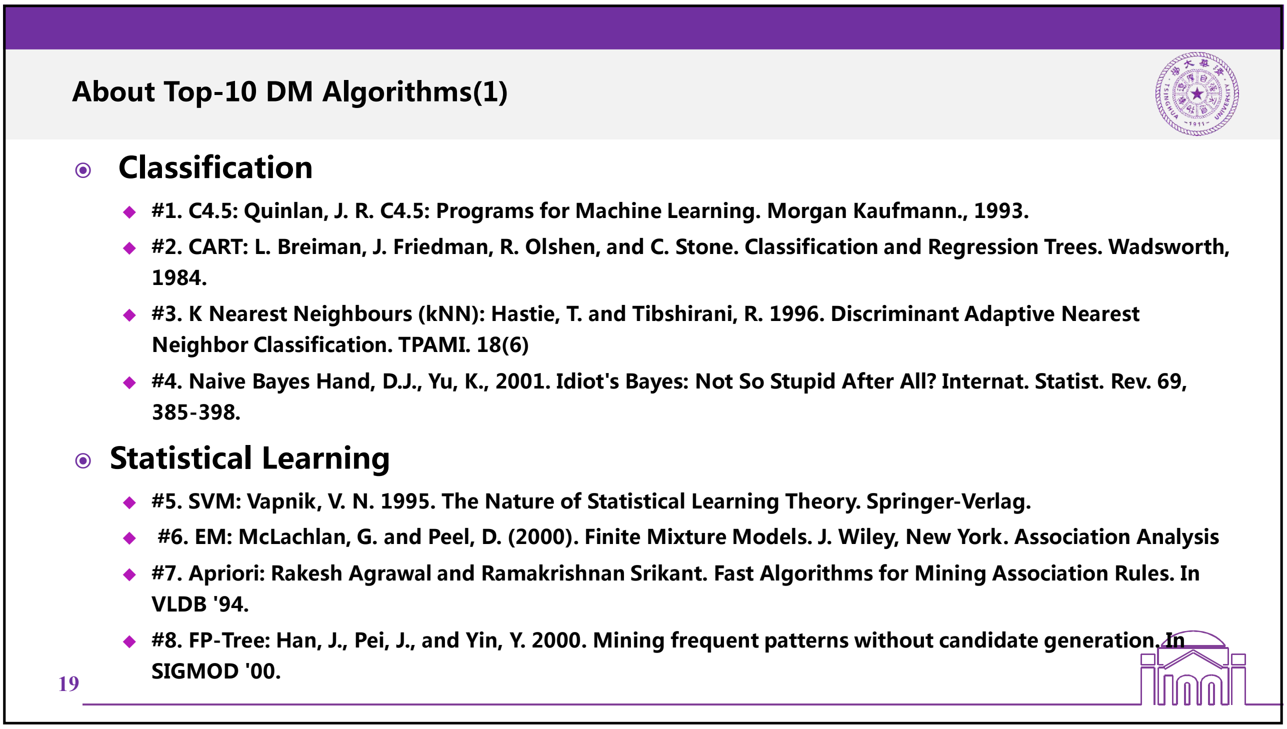
17



Data Mining Functionalities(2)

- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- **Outlier(离群点) analysis**
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- **Trend and evolution analysis**
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera -> large SD memory
  - Periodicity analysis
  - Similarity-based analysis
- **Other pattern-directed or statistical analysis**

18



About Top-10 DM Algorithms(1)

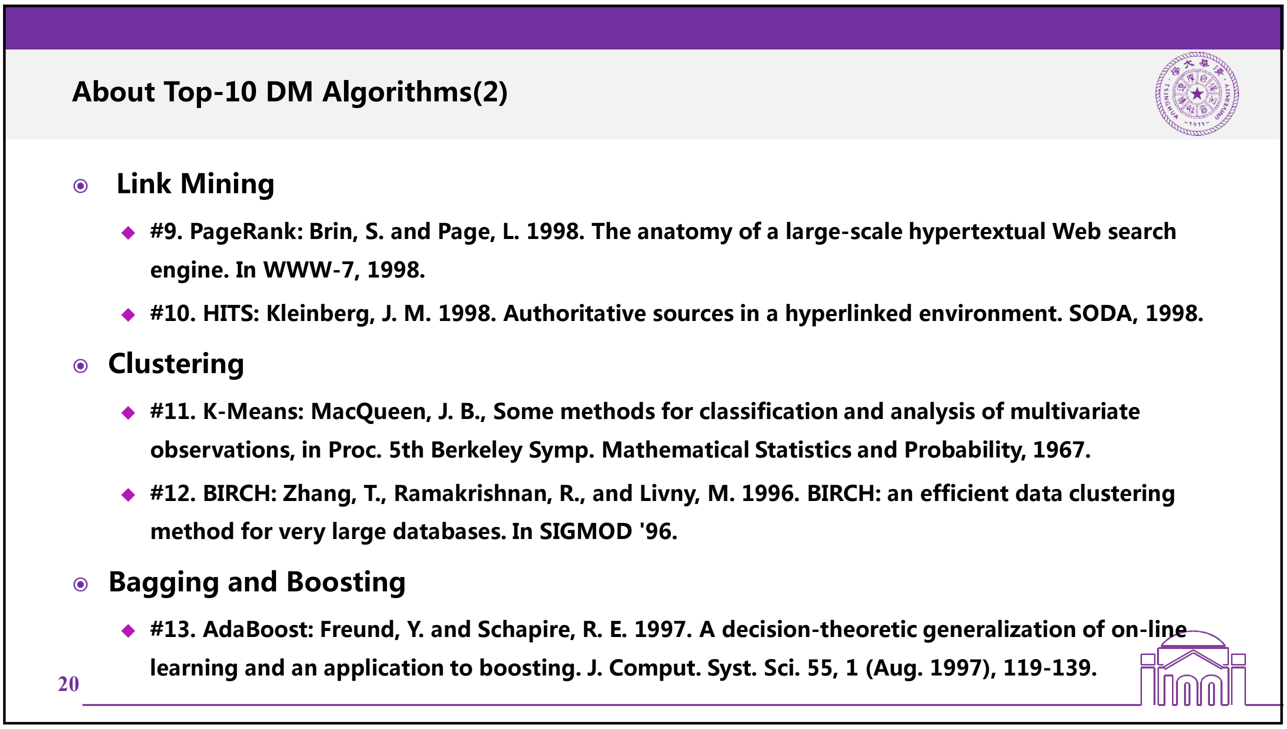
Classification

- #1. C4.5: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.
- #2. CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.
- #3. K Nearest Neighbours (kNN): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. TPAMI. 18(6)
- #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.

Statistical Learning

- #5. SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
- #6. EM: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York. Association Analysis
- #7. Apriori: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
- #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.

19



About Top-10 DM Algorithms(2)

Link Mining

- #9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
- #10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.

Clustering

- #11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
- #12. BIRCH: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.

Bagging and Boosting

- #13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

20

About Top-10 DM Algorithms(3)

- **Sequential Patterns**
  - #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- **Integrated Mining**
  - #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- **Rough Sets**
  - #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- **Graph Mining**
  - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

About Top-10 DM Algorithms(4)

- **Selected at ICDM2007**
  - #1: C4.5 (61 votes)
  - #2: K-Means (60 votes)
  - #3: SVM (58 votes)
  - #4: Apriori (52 votes)
  - #5: EM (48 votes)
  - #6: PageRank (46 votes)
  - #7: AdaBoost (45 votes)
  - #7: kNN (45 votes)
  - #7: Naive Bayes (45 votes)
  - #10: CART (34 votes)

Major Issues in Data Mining

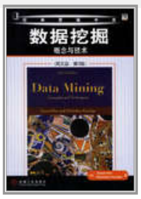
- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

Textbook

Data Mining: Methods and Applications
(Second Edition)
by XU HUA
Tsinghua University Publishers, March 2022.
ISBN: 978-7-302-60144-9
ISBN: 978-7-302-36901-1

Data Mining: Methods and Applications—Application Examples
by XU HUA
Tsinghua University Publishers, Aug 2017.
ISBN: 978-7-302-47211-7

References(1)

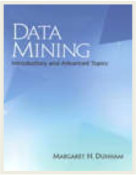


Data Mining: Concepts and Techniques

by Jiawei Han & Micheline Kamber
Morgan Kaufmann Publishers, March 2006.
ISBN: 1-55860-901-6

数据挖掘教程

Margaret H.Dunham 著，郭崇慧，田凤占，靳晓明等译
清华大学出版社, 2005年5月.
ISBN: 7-302-10533-2



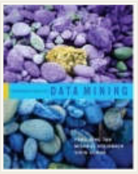
Data Mining: Introductory and Advanced Topics

by Margaret H. Dunham
Prentice Hall; Aug 2002
ISBN-10: 0130888923

References(2)

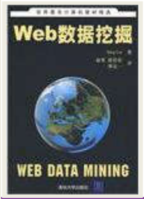
数据挖掘概念与技术（原书第3版）

韩家炜，堪博　著，范明，孟小峰　译
机械工业出版社，2007年3月
ISBN：9787111205388



Introduction to Data Mining

by Pang-Ning Tan, Michael Steinbach, Vipin Kumar
Addison Wesley; May 2005
ISBN-13: 978-0321321367



Web数据挖掘

Bing Liu　著，余勇，薛贵荣，韩定一　译
清华大学出版社，2009年4月
ISBN：978-7-302-19338-8

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

- KDD: ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining
  - KDD08, KDD09 , KDD10 , KDD11 , KDD12, KDD13, KDD14, KDD15
- SDM: SIAM Data Mining Conf.
  - SDM08, SDM09, SDM10, SDM11, SDM12 , SDM13, SDM14, SDM15
- ICDM: IEEE Int. Conf. on Data Mining
  - ICDM08, ICDM09, ICDM10, ICDM11, ICDM12, ICDM13, ICDM14, ICDM15
- PKDD : Conf. on Principles and Practices of Knowledge Discovery and Data Mining
  - PKDD08, PKDD09, PKDD10, PKDD11, PKDD12, PKDD13, PKDD14, PKDD15
- PAKDD: Pacific-Asia Conf. on Knowledge Discovery and Data Mining
  - PAKDD08, PAKDD09, PAKDD10, PAKDD11, PAKDD12, PAKDD13,PAKDD14, PAKDD15

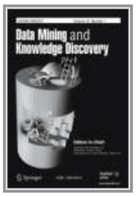
Journals

Data Mining and Knowledge Discovery

Geoffrey I. Webb

http://springer.lib.tsinghua.edu.cn/

清华IP可直接登录



IEEE Trans. On Knowledge and Data Eng.

Xindong Wu

http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=69

清华IP直接登录

SIGKDD Explorations

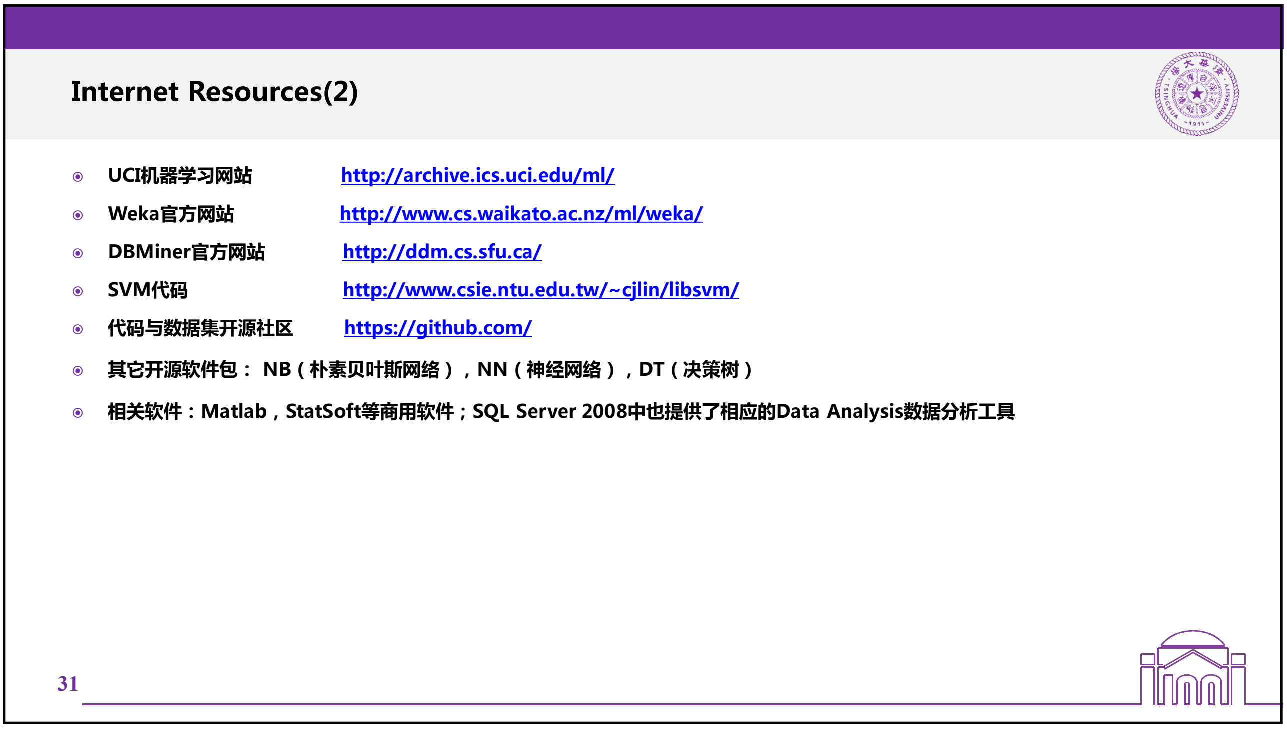
Osmar R. Zaiane

http://www.sigkdd.org/explorations/issue.php?issue=current

清华IP直接登录

Internet Resources(1)

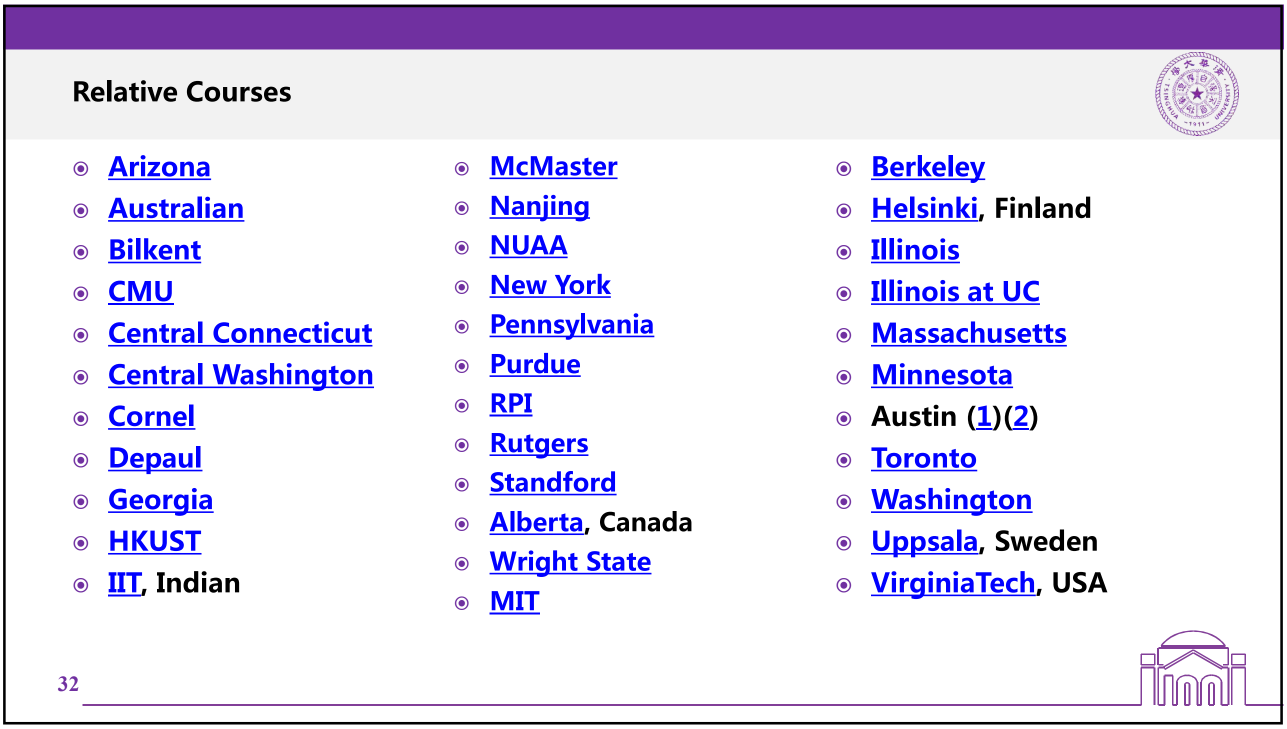
- UCI数据集: http://kdd.ics.uci.edu/
- CMU数据集： http://lib.stat.cmu.edu/datasets/
  http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
- 时序数据集： http://www.stat.wisc.edu/~reinsel/bjr-data/
- 金融数据集： http://lisp.vse.cz/pkdd99/Challenge/chall.htm
- 癌症基因数据集： http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
- 综合数据集：http://www.cs.nyu.edu/~roweis/data.html
- 数据集列表：http://www.kdnuggets.com/datasets/index.html
- 美国政府开放数据：http://data.gov
- 中国地方政府开放数据：北京 http://www.bjdata.gov.cn/
  上海 http://datashanghai.gov.cn



Internet Resources(2)

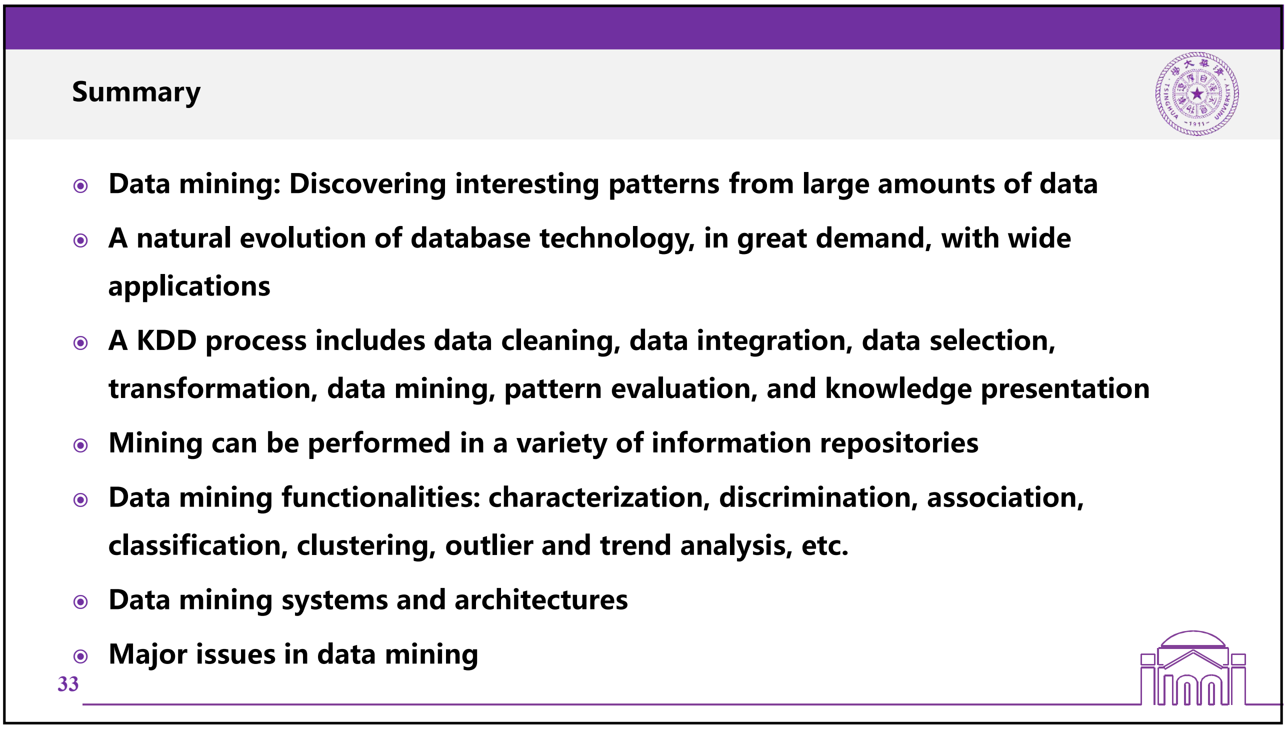
UCI机器学习网站	http://archive.ics.uci.edu/ml/
Weka官方网站	http://www.cs.waikato.ac.nz/ml/weka/
DBMiner官方网站	http://ddm.cs.sfu.ca/
SVM代码	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
代码与数据集开源社区	https://github.com/

- 其它开源软件包： NB（朴素贝叶斯网络），NN（神经网络），DT（决策树）
- 相关软件：Matlab，StatSoft等商用软件；SQL Server 2008中也提供了相应的Data Analysis数据分析工具



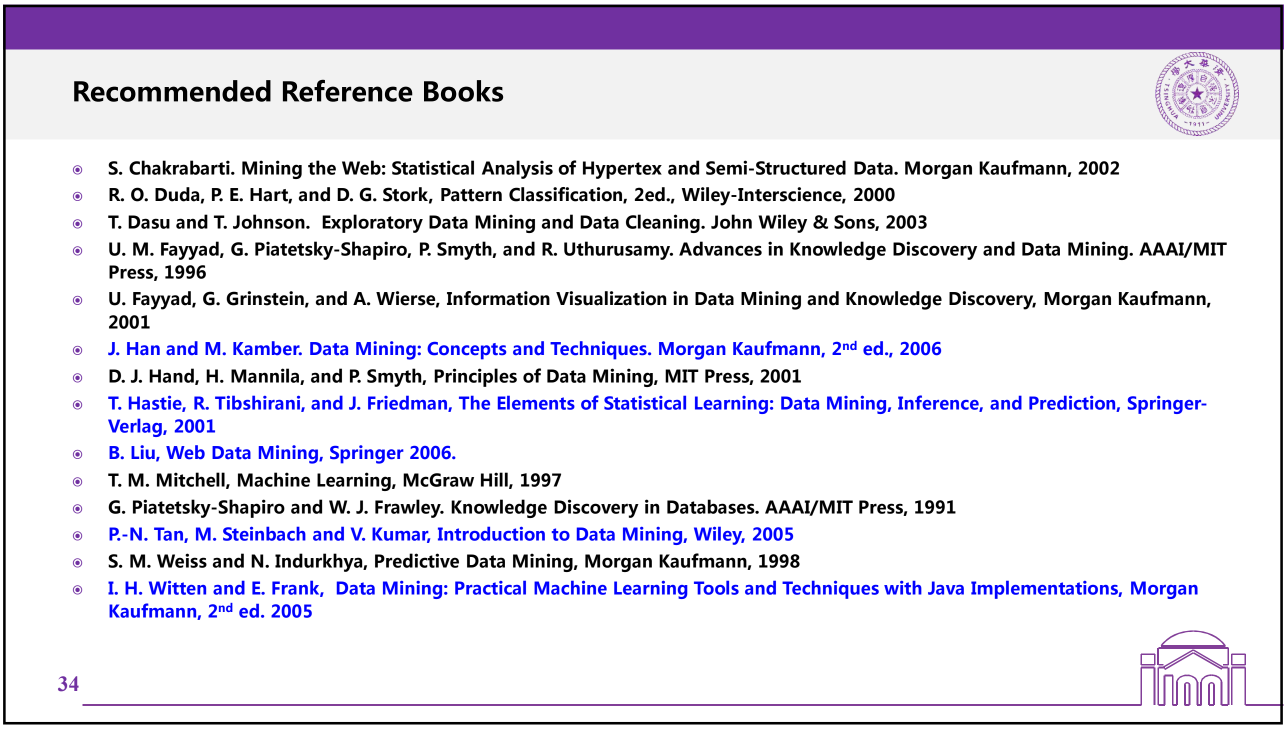
Relative Courses

- Arizona
- Australian
- Bilkent
- CMU
- Central Connecticut
- Central Washington
- Cornel
- Depaul
- Georgia
- HKUST
- IIT, Indian
- McMaster
- Nanjing
- NUAA
- New York
- Pennsylvania
- Purdue
- RPI
- Rutgers
- Standford
- Alberta, Canada
- Wright State
- MIT
- Berkeley
- Helsinki, Finland
- Illinois
- Illinois at UC
- Massachusetts
- Minnesota
- Austin (1)(2)
- Toronto
- Washington
- Uppsala, Sweden
- VirginiaTech, USA



Summary

- Data mining: Discovering interesting patterns from large amounts of data
- A natural evolution of database 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 in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining



Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

概述部分结束！