Data Mining: Introduction

马锦华

数据科学与计算机学院

中山大学

About The Instructor

- Education
  - Bachelor and master in mathematics, SYSU
  - PhD in computer science, HKBU

- Research Experience
  - Postdoc at Rutgers
  - Postdoc at Johns Hopkins
  - Postdoc at HKBU

- Research interests
  - Machine learning: feature fusion, transfer learning, etc.
  - Computer vision: intelligent video surveillance, etc.
  - Medical data analysis: diagnosis and prediction models, etc.

About This Course

- Instructor's contact
  - Office: 超算中心5楼529C
  - Email: majh8@mail.sysu.edu.cn

- Teaching assistant
  - 王子佳, 2582822457@qq.com
  - 谢国添, 1224617026@qq.com

- Lecture hours and venue
  - Monday 1-2节 (1-9 weeks), D303
  - Wednesday 9-10节, A306

- Prerequisite
  - Linear Algebra, Statistics, Data Structure, Programing

About This Course

- Course Contents
  - Supervised learning
    - Linear regression, logistic regression, SVM, decision tree, ensemble methods, neural networks, overfitting
  - Unsupervised learning
    - PCA, manifold learning, clustering
  - Recommendation system
    - Collaborative Filtering, UV-Decomposition
  - Association analysis
    - Apriori Algorithm, frequent itemset
  - Recent advanced techniques
    - PageRank, hashing, Stochastic gradient descent, CNN
- Slides Website: http://www.xieguotian.cn/

About This Course

- Recommended readings
  - 周志华. 机器学习. 清华大学出版社, 2016.
  - Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman. Mining of Massive Datasets, Second Edition. Cambridge University Press, 2014.
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to data mining. Pearson, 2006.
  - T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed. Springer, 2009.

- Assessment
  - 3~4 Assignments, mid-term project, final open-ended project, attendance

Why Data Mining?

- Large-scale Data is Everywhere!

You Tube 300 hours video uploaded to YouTube every minute
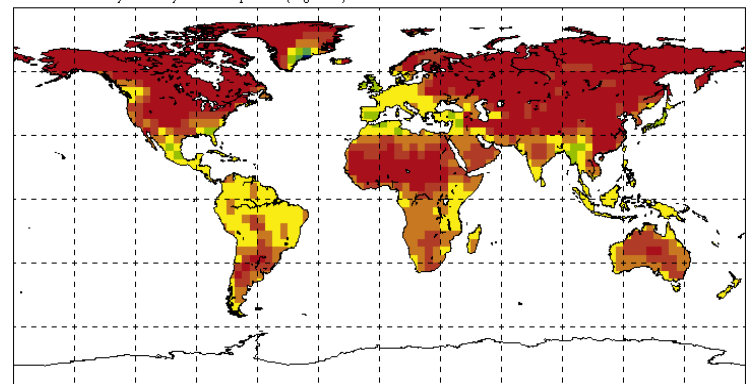
500 million tweets on Twitter per day

30 billion pieces of content shared on Facebook every month

- Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs

CCCma/A2a January to January Mean Temperature (degrees C) 2080s relative to 1961-90

Predicting the impact of climate change

Why Data Mining?

- Commercial Viewpoint

- – Provide better, customized services for an edge, e.g.
  - Finance, Medicine, Manufacturing, Customer Relationship Management, Fraud Detection, etc.
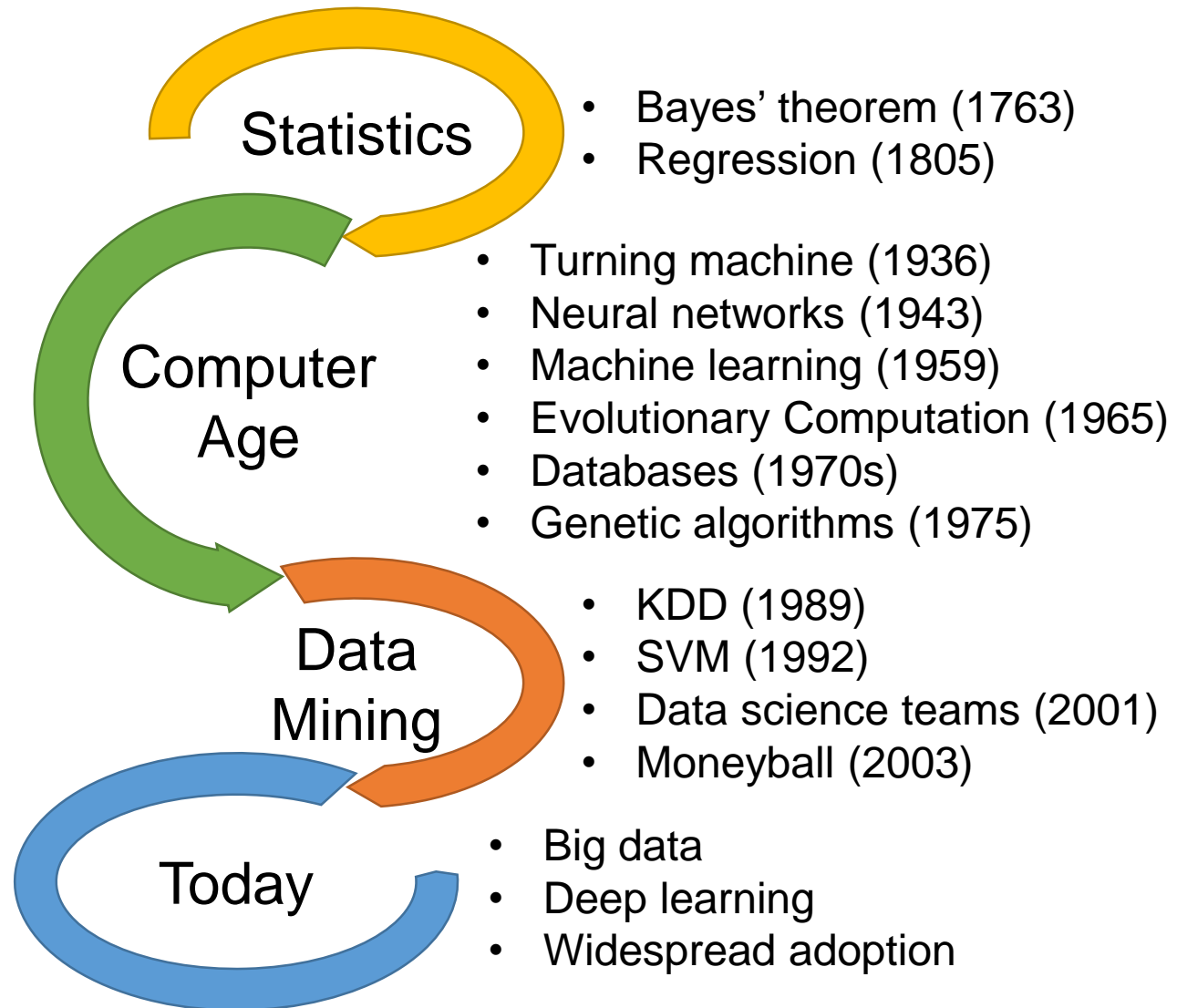- – Conclusion: help to find a **GOOD JOB**

Why Data Mining?

- ## Research Viewpoint
  - Related refereed journals, e.g.,
    - IEEE Trans. on Knowledge and Data Engineering (TKDE)
    - IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)
    - Journal of Machine Learning Research
    - International Journal of Computer Vision
  - Related refereed conferences, e.g.,
    - ACM Knowledge Discovery and Data Mining (KDD)
    - International Joint Conference on Artificial Intelligence (IJCAI)
    - International Conference on Machine Learning (ICML)
    - International Conference on Computer Vision (ICCV)
  - Conclusion: help to pursue **ADVANCED DEGREE**

What is Data Mining?

History of Data Mining

Statistics
- Bayes' theorem (1763)
- Regression (1805)

Computer Age
- Turning machine (1936)
- Neural networks (1943)
- Machine learning (1959)
- Evolutionary Computation (1965)
- Databases (1970s)
- Genetic algorithms (1975)

Data Mining
- KDD (1989)
- SVM (1992)
- Data science teams (2001)
- Moneyball (2003)

Today
- Big data
- Deep learning
- Widespread adoption

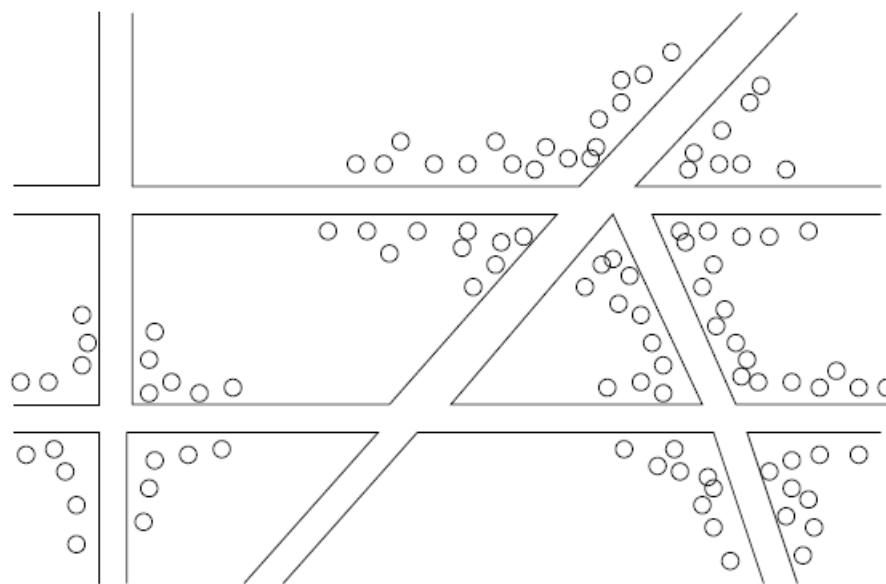# What is Data Mining?

- **Definition**
  - Extract useful patterns from (usually large-scale) data

- **But to extract the knowledge data needs to be**
  - Stored (systems)
  - Managed (databases)
  - And ANALYZED ⬅ this class

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

What is Data Mining?

- Discovery of cholera (霍乱)
  - Instance of clustering to solve the cholera problem in London

Plotting cholera cases on a map of London

  - Knowledge: high chance of cholera in clusters around intersections with contaminated wells that had become

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 "Amazon"

- What is Data Mining?

  – Certain names are more prevalent in certain US locations (O'Brien, O'Rourke, O'Reilly… in Boston area)

  – Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

Data Mining Tasks

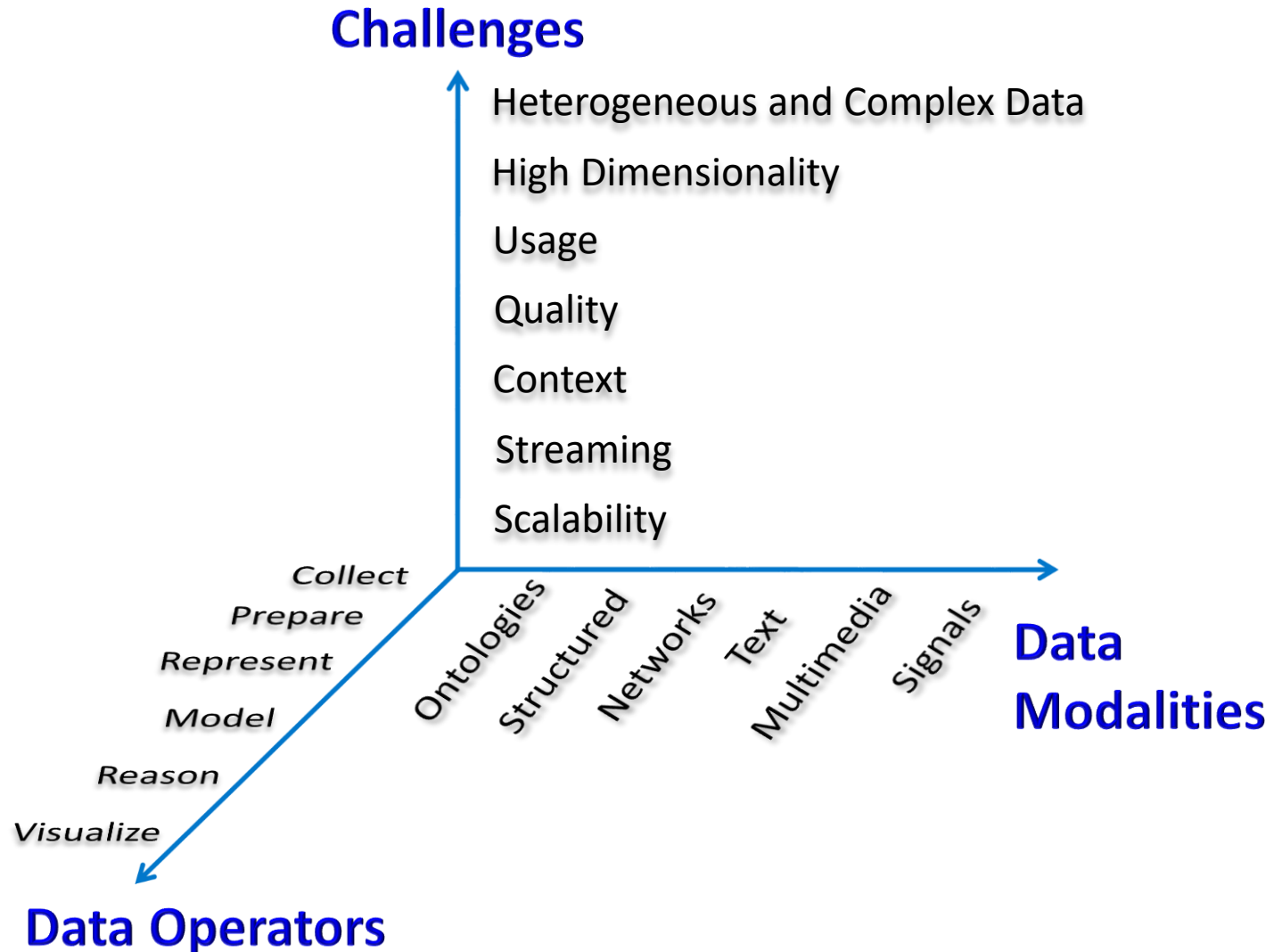- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables, e.g.,
    - Regression: time series prediction of stock market indices
    - Classification: classify credit card transactions as legal or not
    - Recommender systems: predict the someone's rating or preference for a movie

- **Descriptive methods**
  - Find human-interpretable patterns that describe data
    - Clustering: find groups of documents that are similar to each other based on the important terms
    - Association Analysis: market-basket analysis for rule-based sales promotion

What Matters for Data Mining?

Challenges

Heterogeneous and Complex Data

High Dimensionality

Usage

Quality

Context

Streaming

Scalability

Collect
Prepare
Represent
Model
Reason
Visualize

Ontologies Structured Networks Text Multimedia Signals

Data Modalities

Data Operators

Data Mining: Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model
- **In this class we will do both!**

CS Theory

Machine Learning

Data Mining

Database systems
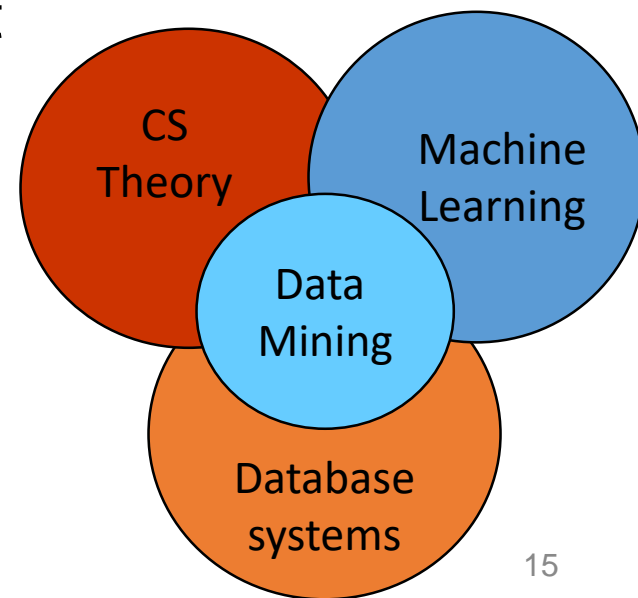
Quiz

- Discuss whether or not each of the following activities is a data mining task
  - Predicting the outcomes of tossing a (fair) pair of dice
    - No. Since the die is fair, this is a probability calculation.
  - Monitoring the heart rate of a patient for abnormalities
    - Yes, known as anomaly detection or classification problem.
  - Extracting the frequencies of a sound wave
    - No. This is signal processing.
  - Dividing customers according to their gender
    - No. This is a simple database query.
  - Predicting the future stock price of a company using historical records
    - Yes, known as predictive modelling, solved by e.g. regression

How do you want that data?

References

- J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org
- P.-N. Tan, M. Steinbach, V. Kumar: Introduction to data mining, Second Edition, https://www-users.cs.umn.edu/~kumar001/dmbook/index.php