Introduction to Data Mining

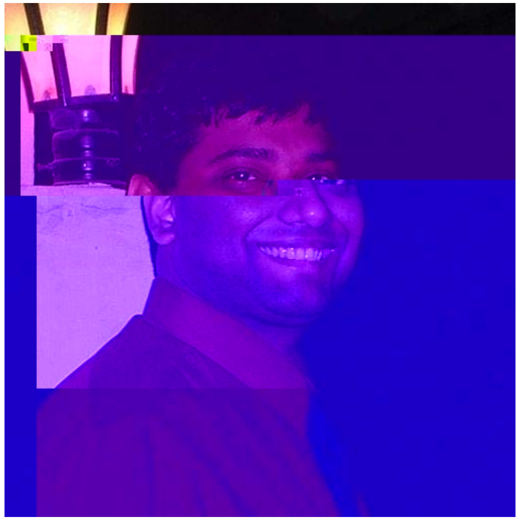
Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Outline

- **Overview**
- Introduction
- The Data Mining Process
- The Basic Data Types
- The Major Building Blocks
- Scalability and Streaming
- Application Scenarios
- Summary
- Mathematical Background

Textbook

- Charu C. Aggarwal. Data Mining: The Textbook. Springer, May 2015.
  - http://www.charuaggarwal.net/Data-Mining.htm



Distinguished Research Staff Member
IBM T. J. Watson Research Center

Textbook

- Charu C. Aggarwal. Data Mining: The Textbook. Springer, May 2015.
  - http://www.charuaggarwal.net/Data-Mining.htm
- Reference
  - David Hand, Heikki Mannila and Padhraic Smyth. Principles of Data Mining. The MIT Press, 2001.
  - Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3 edition, 2011.
  - Ian H. Witten Eibe Frank and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 3 edition, 2011.
  - Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2007.

Course Plan (1)

- Introduction to Data Mining
- Data Preparation
- Similarity and Distances
- Association Pattern Mining
- Cluster Analysis

Course Plan (2)

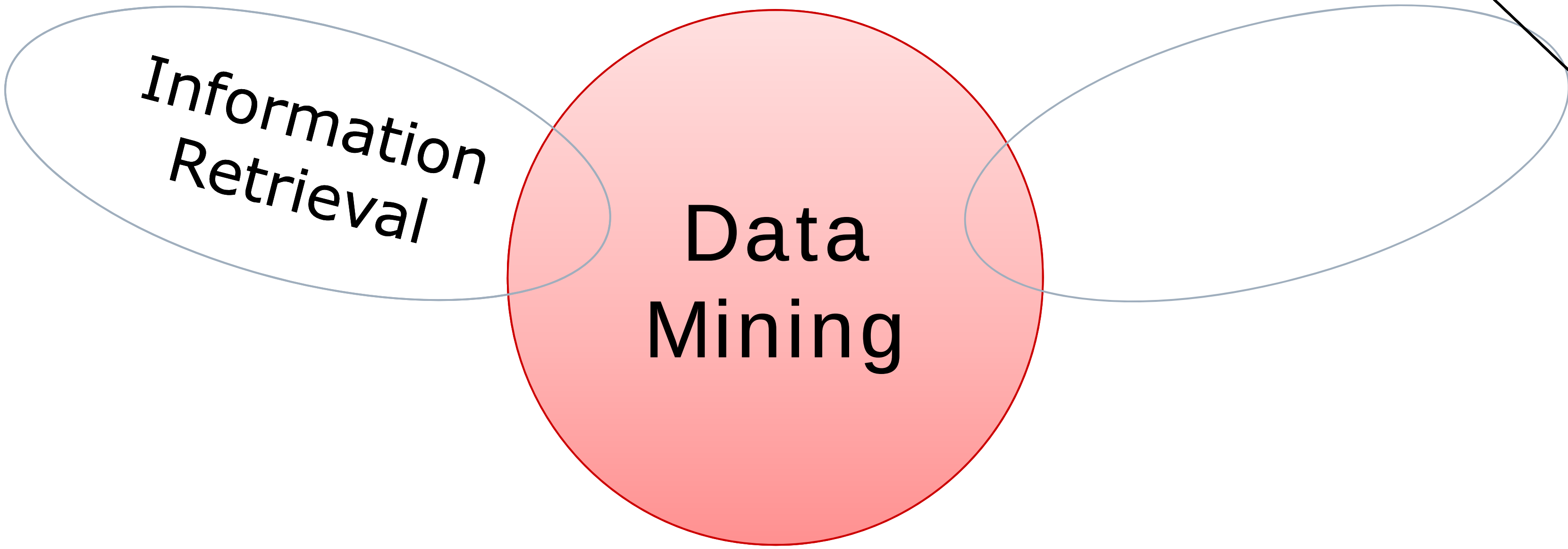
- Data Classification
  - Decision Trees, Naïve Bayes
  - SVM, Ensemble Methods
- Linear Methods for Regression
  - Least Square, Ridge Regression, Lasso
- Mining Text Data
  - LSA, PLSA, Co-clustering
- Mining Web Data
  - Ranking, Recommender Systems
- Big Data Mining
  - Online, Randomized, Distributed

Grading

- Homework (70)
  - Document Processing
  - Association Pattern Mining
  - Classification
  - Ensemble
  - Clustering
  - Competition

  http://lamda.nju.edu.cn/yehj/DM16/dm16.html
- Final Exam (30)
  - Hong Qian (qianh@lamda.nju.edu.cn)
  - Han-Jia Ye (yehj@lamda.nju.edu.cn)
  - Jia-Lve Chen (chenjl@lamda.nju.edu.cn)

Interdiscipline



Resources

- WWW
  - Google, Wikipedia
- Top Conferences
  - **SIGKDD**, WWW, SIGIR, ACM MM
  - ICML, NIPS, VLDB, SIGMOD
  - AAAI, IJCAI, CVPR, ICCV
- Top Journals
  - **TKDE**, **TKDD**, TPAMI, TMM
  - JMLR, ML, PR, TODS, TIP

Outline

- Overview
- **Introduction**
- The Data Mining Process
- The Basic Data Types
- The Major Building Blocks
- Scalability and Streaming
- Application Scenarios
- Summary
- Mathematical Background

Introduction

- What is data mining?

The study of collecting, cleaning, processing, analyzing, and gaining useful insights from data

Introduction

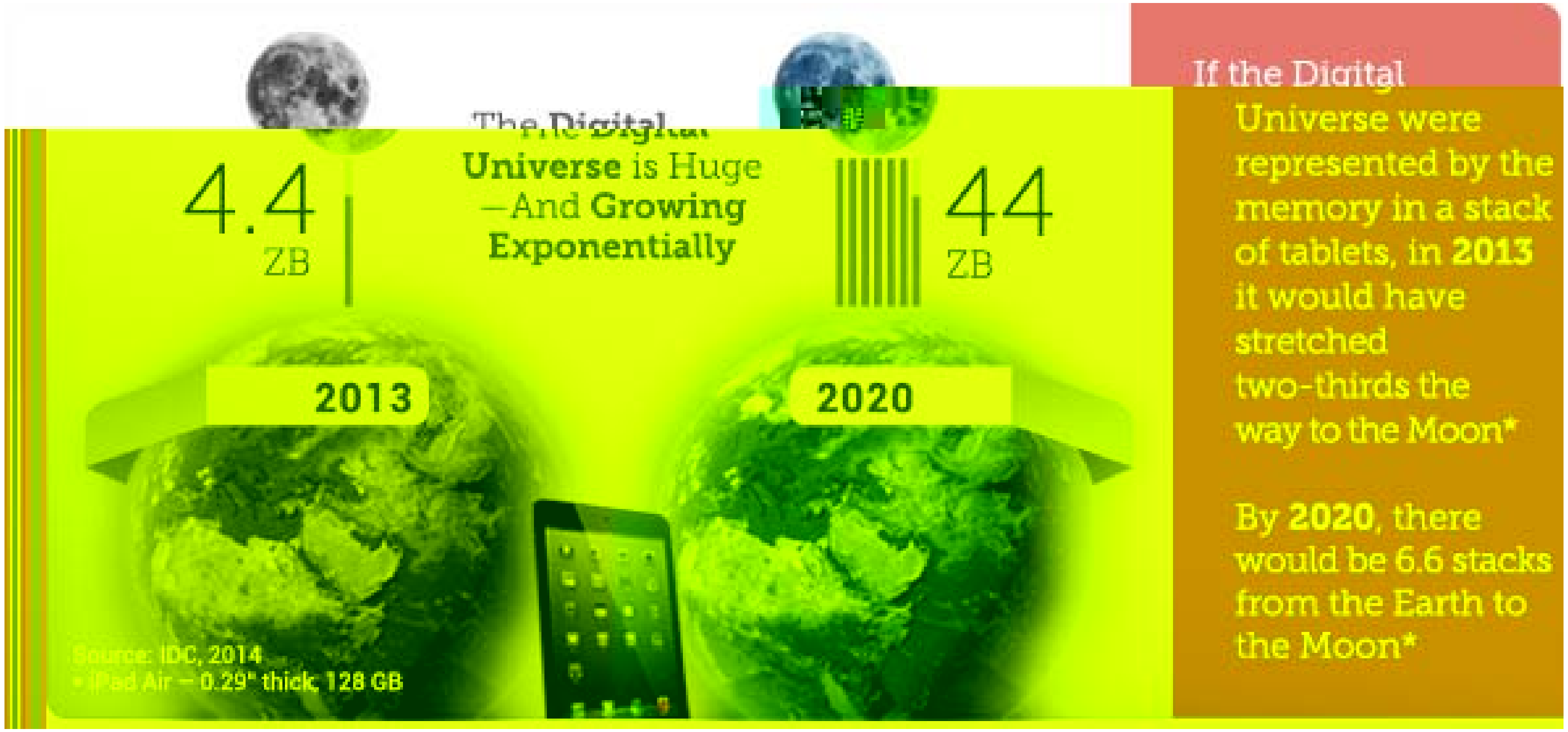
What is data mining?

> The study of collecting, cleaning, processing, analyzing, and gaining useful insights from data

Why do we need?

- Data is the new oil
- We have entered the Era of Big Data

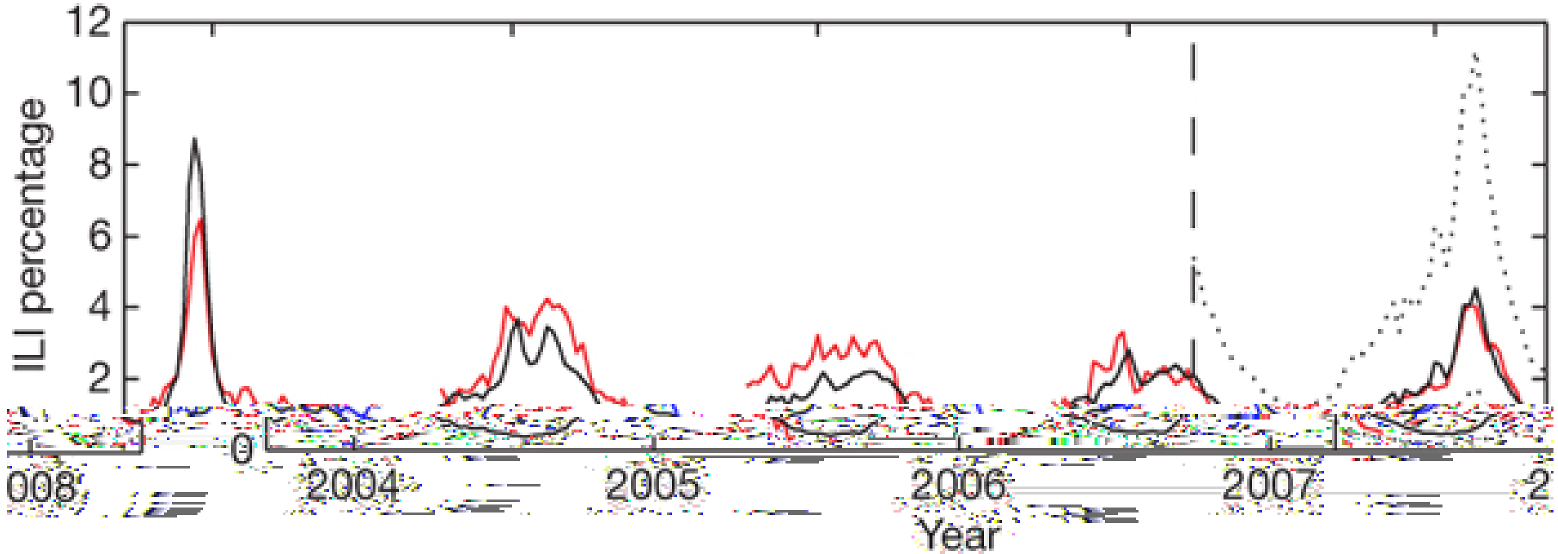
Big Data



1 Zb 1000 EB 1000, 000 PB 1000, 000, 000 TB

http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Google Flu Trends



- Google's prediction: a reporting lag of about one day
- Traditional surveillance systems: a 1–2-week reporting lag

Ginsberg et al. Detecting influenza epidemics using search engine query data. Nature 457, 1012-1014, 2009.

Outline

- Overview
- Introduction
- **The Data Mining Process**
- The Basic Data Types
- The Major Building Blocks
- Scalability and Streaming
- Application Scenarios
- Summary
- Mathematical Background

The Data Mining Process (1)

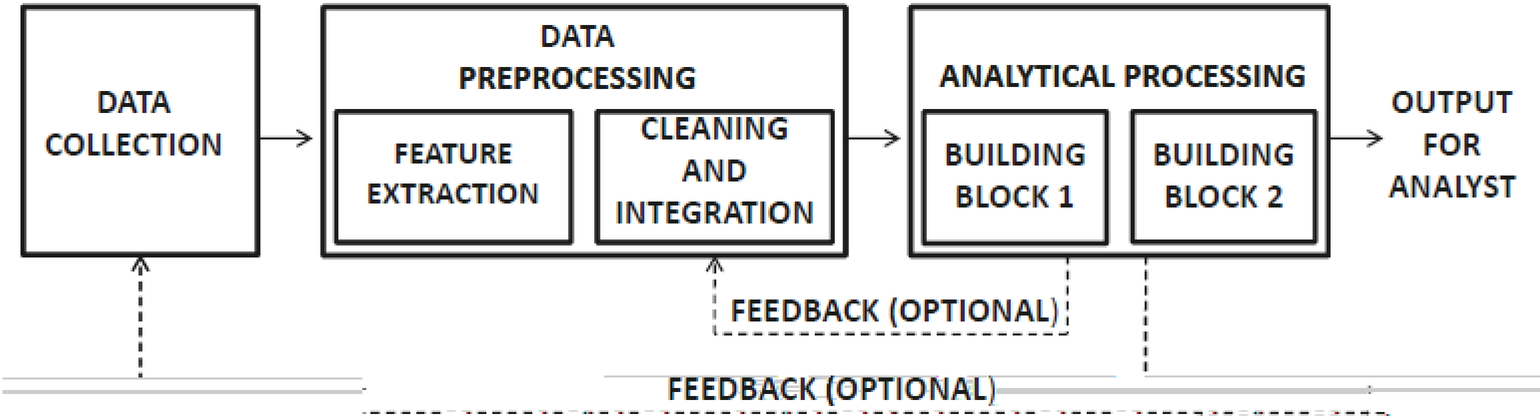


Figure 1.1: The data processing pipeline

- Data Collection
  - Hardware, Software, Human
- Feature Extraction
  - Multidimensional, Time series

The Data Mining Process (2)

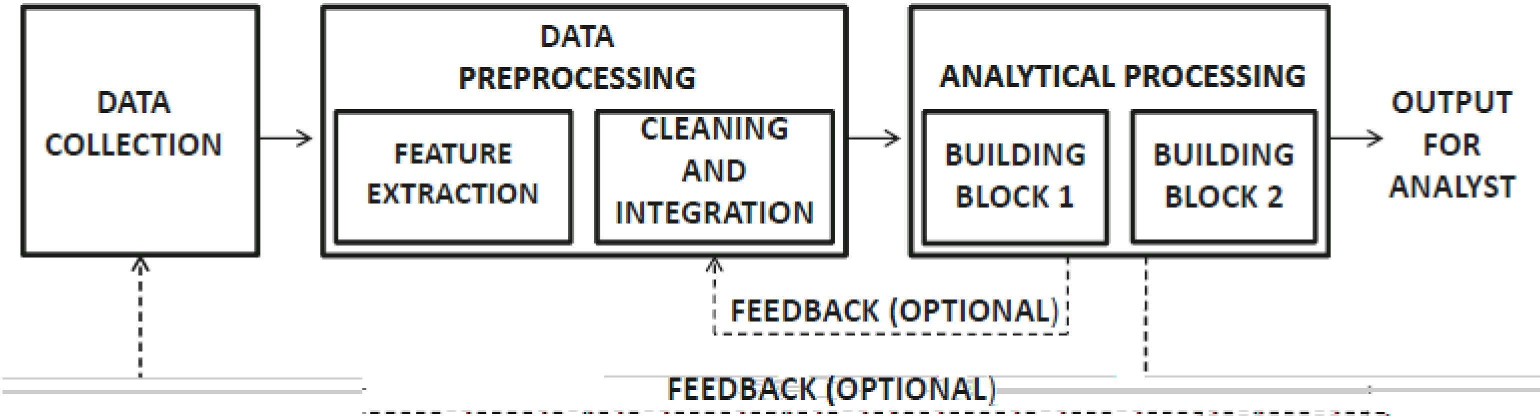


Figure 1.1: The data processing pipeline

- Data Cleaning and Integration
  - Handle missing and erroneous values
  - Integrate data from multiple sources
- Analytical Processing and Algorithms

A Recommendation Scenario (1)

Example 1.2.1 *Consider a scenario in which a retailer has Web logs corresponding to customer accesses to Web pages at his or her site. Each of these Web pages corresponds to a product, and therefore a customer access to a page may often be indicative of interest* [illegible] *ing customers. The retailer wants to make targeted product recommendations to customers us the customer demographics and buying behavior.*

1. Data Collection

- Web logs at the site

```
98.206.207.157 - - [31/Jul/2013:18:09:38 -0700] "GET /productA.htm
HTTP/1.1" 200 328177 "-" "Mozilla/5.0 (Mac OS X) AppleWebKit/536.26
(KHTML, like Gecko) Version/6.0 Mobile/10B329 Safari/8536.25"
"retailer.net"
```

- Demographic information within the retailer database

A Recommendation Scenario (2)

Example 1.2.1 *Consider a scenario in which a retailer has Web logs corresponding to customer accesses to Web pages at his or her site. Each of these Web pages corresponds to a product, and therefore a customer access to a page may often be indicative of interest* [illegible] *ing customers. The retailer wants to make targeted product recommendations to customers us the customer demographics and buying behavior.*

2. Feature Extraction

- A specific choice of features extracted from the Web page accesses

3. Data Cleaning

- Estimate, Remove, Normalization

4. Data Integration

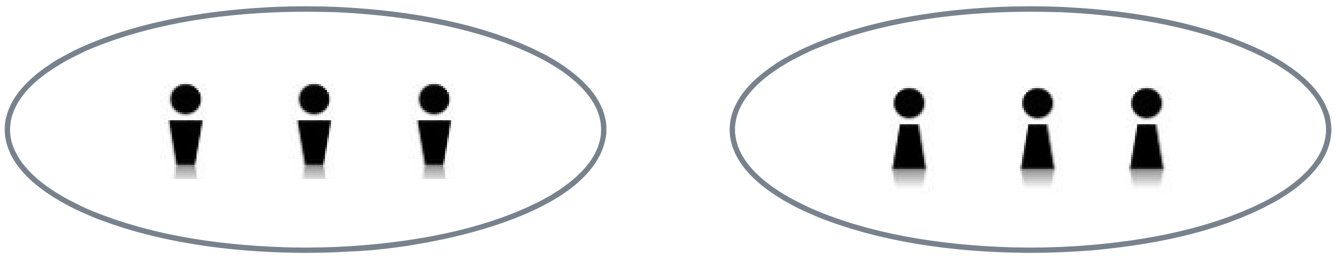
- Add demographics information

A Recommendation Scenario (3)

Example 1.2.1 *Consider a scenario in which a retailer has Web logs corresponding to customer accesses to Web pages at his or her site. Each of these Web pages corresponds to a product, and therefore a customer access to a page may often be indicative of interest* [illegible] *ing customers. The retailer wants to make targeted product recommendations to customers us the customer demographics and buying behavior.*

5. Making Recommendation

- Partition customers by clustering



- Recommend based on behaviors of customers in the same group

The Data Preprocessing Phase

- Rarely explored to the extent that it deserves
  -

1. Feature Extraction
   - HTML, System logs
2. Data Cleaning
   - Erroneous, Missing, Inconsistent
3. Feature Selection and Transformation
   - High-dimensionality, Heterogeneous

Outline

- Overview
- Introduction
- The Data Mining Process
- **The Basic Data Types**
- The Major Building Blocks
- Scalability and Streaming
- Application Scenarios
- Summary
- Mathematical Background

The Basic Data Types

- Nondependency-oriented Data
  - Data records do not have any specified dependencies between either the data items or the attributes

Table 1.1: An example of a multidimensional data set

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	[illegible]	Native American	10598

- Dependency-oriented Data
  - Implicit or explicit relationships may exist between data items
  - Social Network, Time Series

Nondependency-Oriented Data (1)

□ Multidimensional Data (Vectors)

Definition 1.3.1 (Multidimensional Data) *A multidimensional data set $\mathcal{D}$ is a set of n records, $\overline{X_1} \ldots \overline{X_n}$, such that each record $\overline{X_i}$ contains a set of d features denoted by $(x_i^1 \ldots x_i^d)$.*

- Record, data point, instance, example, transaction, entity, tuple, object, feature-vector
- Fields, attributes, dimensions, features.

Nondependency-Oriented Data (2)

□ Quantitative Multidimensional Data

Table 1.1: An example of a multidimensional data set

code	Name	Age	Gender	Race	ZIP
5139	John S.	45	M	African American	05
10598	Manyona L.	31	F	Native America	

- Numerical in the sense that they have a natural ordering
- Continuous, numeric, or quantitative
- Convenient for analytical processing
  - ✓ Mean, Variance, +-*/

Nondependency-Oriented Data (3)

- Categorical Data

Table 1.1: An example of a multidimensional data set

... code	Name	Age	Gender	Race	ZIP
...139	John S...	45	M	African American	0...
10598	[illegible]	31	F	Native American	[illegible]

  - Take on **discrete unordered** values
  - Unordered discrete-valued Data
- Mixed Attribute Bata
  - A combination of categorical and numeric attributes

Nondependency-Oriented Data (4)

- **Binary Data**
  - A special case of multidimensional categorical data
    - ✓ Each categorical attribute may take on one of at most two discrete values
  - A special case of multidimensional quantitative data
    - ✓ An ordering exists between the two values
- **Setwise Data**
  - A set element indicator $\mathbb{I}(\dot\omega) \begin{cases} 1, \text{if } \dot\omega \quad \aleph \\ 0, \text{otherwise} \end{cases}$

Nondependency-Oriented Data (5)

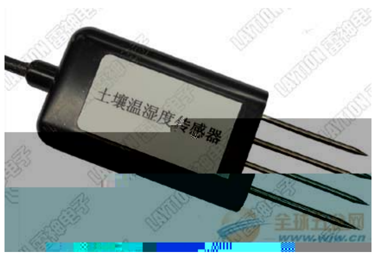
- Text Data
  - A string—a dependency-oriented data
    - Natural Language Processing
  - Document-term matrix— a multidimensional quantitative data

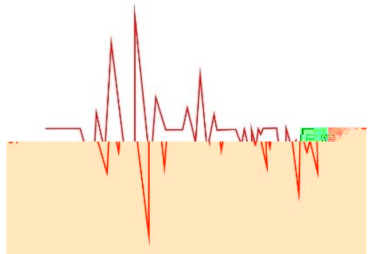
    - Nontrivial

Dependency-Oriented Data (1)

Implicit Dependencies

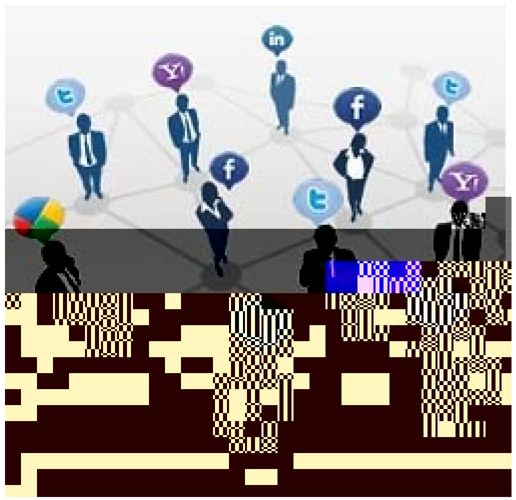
- Dependencies are known to “typically” exist





Explicit dependencies

- Graph or network data where edges are used to specify relationships



Dependency-Oriented Data (2)

- Time-Series Data
  - Contextual attributes
    - ✓ Define the context on the basis of which the implicit dependencies occur in the data
  - Behavioral attributes
    - ✓ Represent the values that are measured in a particular context

Definition 1.3.2 (Multivariate Time-Series Data) *A time series of length n and dimensionality d contains d numeric features at each of n time stamps $t_1 \ldots t_n$. Each time-stamp contains a component for each of the d series. Therefore, the set of values received at time stamp t_i is $\overline{Y_i} = (y_i^1 \ldots y_i^d)$. The value of the jth series at time stamp t_i is y_i^j.*

Dependency-Oriented Data (3)

- Discrete Sequences and Strings
  - The categorical analog of time-series data
    - ✓ Event logs: a sequence of user actions

      ```
      Login Password Login Password Login Password ....
      ```

    - ✓ Biological data: strings of nucleotides
      - Contextual attribute is position

Definition 1.3.3 (Multivariate Discrete Sequence Data) *A discrete sequence of length*
stamps n and dimensionality d contains d discrete feature values at each of n different time-
$\ldots y_i^d$), $t_1 \ldots t_n$. Each of the n components Y_i contains d discrete behavioral attributes (y_i^1
collected at the ith time-stamp.

  - Strings, when $d = 1$

Dependency-Oriented Data (4)

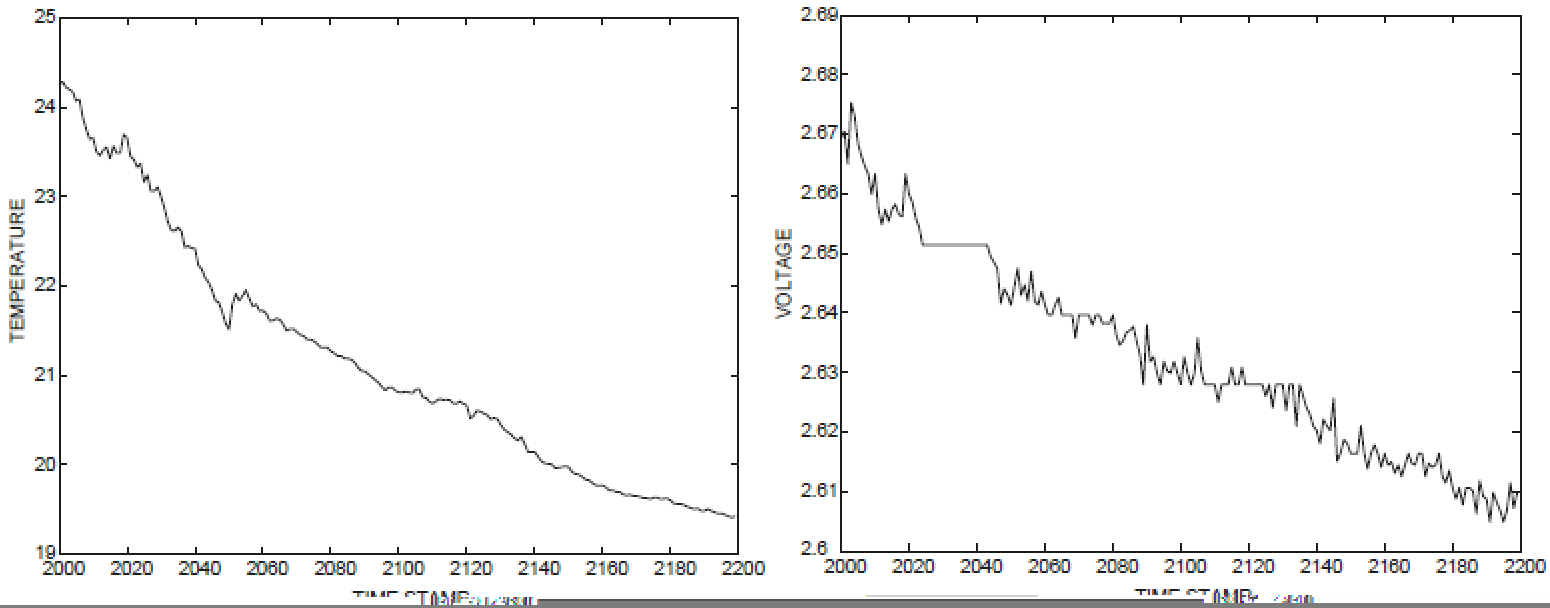
- Spatial Data

Definition 1.3.4 (Spatial Data) *A d-dimensional spatial data record contains d behavioral attributes and one or more contextual attributes containing the spatial location. Therefore, a d-dimensional spatial data set is a set of d dimensional records* $\overline{X_1} \ldots \overline{X_n}$*, together with a set of n locations* $L_1 \ldots L_n$*, such that the record* $\overline{X_i}$ *is associated with the location* L_i*.*

- Spatiotemporal Data
  - Both spatial and temporal attributes are contextual
  - The temporal attribute is contextual, whereas the spatial attributes are behavioral
    - ✓ Trajectory analysis

Dependency-Oriented Data (5)

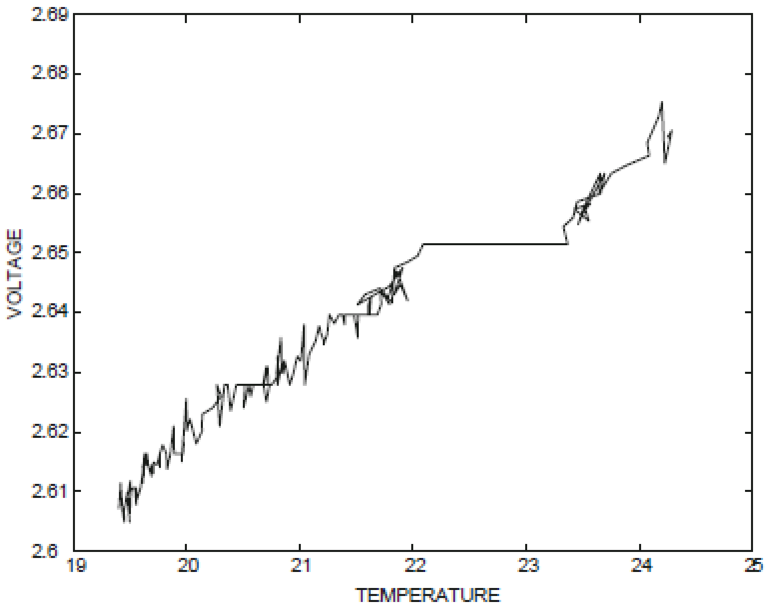
- 2- or 3-dimensional time-series data
  - Can be mapped onto trajectories



(a) Temperature (b) Voltage

Dependency-Oriented Data (6)

- 2- or 3-dimensional time-series data
  - Can be mapped onto trajectories



(c) Temperature-voltage trajectory

Dependency-Oriented Data (7)

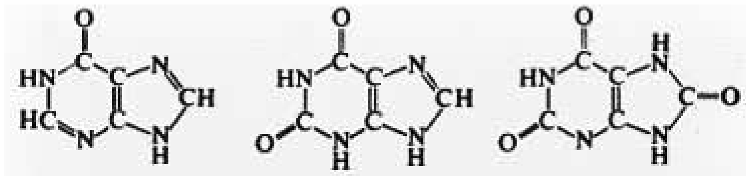
Network and Graph Data

- A Single Network

Definition 1.3.5 (Network Data) *A network $G = (N, A)$ contains a set of nodes N and a set of edges A, where the edges in A represent the relationships between the nodes. In some cases, an attribute set $\overline{X_i}$ may be associated with node i, or an attribute set $\overline{Y_{ij}}$ may be associated with edge (i, j).*

  - ✓ Web graph with directed edges corresponding to directions of hyperlinks
  - ✓ Facebook social network with undirected edges corresponding to friendships

- A database containing many small graphs
  - ✓ Chemical compound databases



次黄嘌呤 黄嘌呤 尿酸

Outline

- Overview
- Introduction
- The Data Mining Process
- The Basic Data Types
- **The Major Building Blocks**
- Scalability and Streaming
- Application Scenarios
- Summary
- Mathematical Background

The Major Building Blocks (1)

- Consider a multidimensional database $\mathcal{D}$ with n records and d attributes

$$\begin{array}{cc} & \text{Ώ columns/attributes} \\ \text{ὲ rows/records} & \begin{bmatrix} 1 & 0 & 2 \\ 0 & 3 & 5 \\ 3 & 0 & 8 \\ 7 & 0 & 1 \\ & & \\ 0 & 2 & 1 \end{bmatrix} \end{array}$$

Data Matrix

Я ص ز

- Thinking in linear algebra

The Major Building Blocks (2)

- Consider a multidimensional database $\mathcal{D}$ with n records and d attributes
  - Relationships between columns
    - ✓ Positive or negative association pattern mining problem (e.g., Synonym)
    - ✓ Data classification (i.e., Prediction)
  - Relationships between rows
    - ✓ Clustering
    - ✓ Outlier analysis

Association Pattern Mining (1)

- Sparse binary databases

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad 0,1$$ ن ۀ

[redacted]

Minimum Support	Frequent Patterns	Support
2/5	{2,3}	3/5
	{1,4}	2/5

Association Pattern Mining (2)

Association Rule Mining

- Rule

$$A \Rightarrow B$$

  - ✓ If ȍ appears, then ǒ also ap

- The confidence of the rul

ȕ(ǒ|ȍ) $\frac{\text{ȕ ȍ} \cap \text{ǒ}}{\text{ȕ ȍ}}$

$$\frac{\text{support(ȍ ǒ)}}{\text{support(ȍ)}}$$

Association Pattern Mining (2)

- Association Rule Mining
  - Rule

$$A \Rightarrow B$$

    - ✓ If ő appears, then ǒ also appears
  - The confidence of the rule

$$\frac{\text{support}(ő \quad ǒ)}{\text{support}(ő)}$$

Definition 4.2 (Association Rules) Let A and B be two sets of items. The rule $A \Rightarrow B$ is said to be valid at support level s and confidence level c, if the following two conditions are satisfied:

1. *The support of the item set A is at least s.*
2. *The confidence of $A \Rightarrow B$ is at least c.*

Association Pattern Mining (2)

Association Rule Mining

- Rule

$$A \Rightarrow B$$

  - ✓ If ŏ appears, then ǒ also appears

- The confidence of the rule

support(ŏ ǒ)

{Bread} {Milk} with ω̃ 0.8
{Beer, Cigarette} {Male} with ω̃ 0.9

Definition

B is said to be ... ditions are satisfied:

Data Clustering (1)

Definition 1.4.3 (Data Clustering) *Given a data matrix D (database $\mathcal{D}$), partition its rows (records) into sets $C_1, \ldots, C_K$, such that the rows (records) in each cluster are "similar" to one another.*

- An Informal Definition
  - How to measure the similarity?
    - Human (Nationality, Gender, Age)
  - What is the number of sets?
  - Do sets overlap with each other?
  - How to measure the quality of a partition?

Data Clustering (2)

Definition 1.4.3 (Data Clustering) *Given a data matrix D (database $\mathcal{D}$), partition its rows (records) into sets $\mathcal{C}_1 \ldots \mathcal{C}_k$, such that the rows (records) in each cluster are "similar" to one another.*

- Relevant Applications
  - Customer segmentation
    - ✓
  - Data summarization
    - ✓ Identifying representative points
  - Application to other data mining problems
    - ✓ Outlier analysis

Outlier Detection

Definition 1.4.4 (Outlier Detection) *Given a data matrix D, determine the rows of the data matrix that are very different from the remaining rows in the matrix.*

- Abnormalities, Discordants, Deviants, or Anomalies

  Outlier $\neq$ G

- Applications
  - Intrusion-detection s stems
  - Credit card fraud
  - Interesting sensor events, Earth science
  - Medical diagnosis, Law enforcement

What will happen if you lose your campus card?

Data Classification (1)

- Class Label
  - A particular feature in the data
- The Goal
  - Learn the relationships of the remaining features in the data with respect to this special feature
- Training data
  - The class label is known
- Testing Data
  - The class label is missing

Data Classification (2)

Definition 1.4.5 (Data Classification) *Given an $n \times d$ training data matrix $\mathcal{D}$ (database $\mathcal{D}$), and a class label value in $\{1 \ldots k\}$ associated with each of the n rows in $\mathcal{D}$ (records in $\mathcal{D}$), create a training model $\mathcal{M}$, which can be used to predict the class label of a d-dimensional record $\overline{Y} \notin \mathcal{D}$.*

- Relation to Clustering
  - Supervised vs Unsupervised
- Relation to Association Pattern Mining
  - Classification based on association rules
- Relation to Outlier Detection
  - Supervised outlier detection can be modeled as a classification problem

Data Classification (3)

Definition 1.4.5 (Data Classification) *Given an $n \times d$ training data matrix D (database $\mathcal{D}$), and a class label value in $\{1, \ldots, k\}$ associated with each of the n rows in D (records in $\mathcal{D}$), create a training model $\mathcal{M}$, which can be used to predict the class label of a d-dimensional record $\overline{Y} \notin \mathcal{D}$.*

- Applications
  - Target marketing
    - ✓ Predict buying behaviors
  - Intrusion detection
    - ✓ Predict the possibility of intrusions
  - Supervised anomaly detection
    - ✓ Identify records belonging to rare class

Impact of Complex Data Types on Problem Definitions (1)

Problem	Time series	Spatial	Sequence	Networks
Patterns	Motif-mining Periodic pattern	Colocation patterns	Sequential patterns Periodic Sequence	Structural patterns
	Trajectory patterns			
Clustering	Shape clusters	Spatial clusters	Sequence clusters	Community detection
	Trajectory clusters			
Outliers	Position outlier Shape outlier	Position outlier Shape outlier	Position outlier Combination outlier	[illegible] Community outliers
	Trajectory outliers			
Classification	Position classification Shape classification	Position classification Shape classification	Position classification Sequence classification	Collective classification Graph classification
	Trajectory classification			

Impact of Complex Data Types on Problem Definitions (2)

- Pattern Mining with Complex Data Types
  - Be temporally contiguous, as in time-series Motifs
  - Be periodic, as in periodic patterns
  - Be frequent subgraphs, in networks
- Clustering with Complex Data Types
  - The similarity function is significantly affected by the data type
  - Community detection in networks

Impact of Complex Data Types on Problem Definitions (3)

- Outlier Detection with Complex Data Types
  - A sudden jump in the value of a time series will result in a position outlier
- Classification with Complex Data Types
  - Class labels are attached to a specific position
  - Class labels are attached to individual nodes in a very large network
  - Class labels are attached small graphs

Outline

- Overview
- Introduction
- The Data Mining Process
- The Basic Data Types
- The Major Building Blocks
- **Scalability and Streaming**
- Application Scenarios
- Summary
- Mathematical Background

Scalability Issues and the Streaming Scenario

- The data are stored on one or more machines, but it is too large to process efficiently
  - Distributed Learning
- The data are generated continuously over time in high volume, and it is not practical to store it entirely
  - Online Learning
  - One-pass constraint
  - Concept drift (e.g., popular clothes)

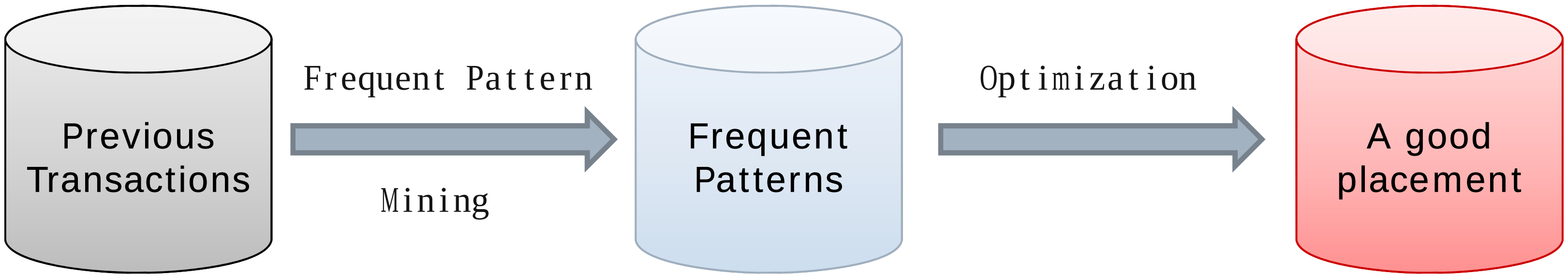
Outline

- Overview
- Introduction
- The Data Mining Process
- The Basic Data Types
- The Major Building Blocks
- Scalability and Streaming
- **Application Scenarios**
- Summary
- Mathematical Background

Store Product Placement

- The famous "beer and diapers" story

Application 1.6.1 (Store Product Placement) A merchant has a set of d products
ns bought together with previous transactions from the customers containing baskets of iter
n the shelves to increase together. The merchant would like to know how to place the product o



- For each placement, define a score based on frequent patterns

Customer Recommendations (1)

Application 1.6.2 (Product Recommendations) *A merchant has an $n \times d$ binary matrix D representing the buying behavior of n customers across d items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.*

- A simple solution based on association rule mining
  - Find associate rules at particular levels of support and confidence

$$A \Rightarrow B$$

  - If a customer have bought items in A, then it is likely he/she will buy items in B.

Customer Recommendations (2)

Application 1.6.2 (Product Recommendations) *A merchant has an $n \times d$ binary matrix D representing the buying behavior of n customers across d items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.*

- A second solution based on clustering
  - For a customer, find the most similar customers
  - Recommendation based on items bought by customers similar to him/her

Customer Recommendations (3)

Application 1.6.2 (Product Recommendations) *A merchant has an $n \times d$ binary matrix D representing the buying behavior of n customers across d items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.*

- A hybrid approach
  - Apply clustering to partitioning customers to similar groups
  - In each group, use association pattern mining to make recommendations

Outline

- Overview
- Introduction
- The Data Mining Process
- The Basic Data Types
- The Major Building Blocks
- Scalability and Streaming
- Application Scenarios
- **Summary**
- Mathematical Background

Summary

- The Data Mining Process
  - Collection → Proprocessing → Analytical
- The Basic Data Types
  - Nondependency-Oriented Data
  - Dependency-Oriented Data
- The Major Building Blocks
  - Association Pattern Mining
  - Data Clustering
  - Outlier Detection
  - Data Classification

Outline

- Overview
- Introduction
- The Data Mining Process
- The Basic Data Types
- The Major Building Blocks
- Scalability and Streaming
- Application Scenarios
- Summary
- **Mathematical Background**

Mathematical Background

- Linear algebra
- Analysis
- Probability and Statistics
- Convex Optimization