Data Mining
(EECS 6412)

Instructor: Aijun An
Department of Electrical Engineering and Computer Science
York University
Email: aan@cse.yorku.ca or aan@eecs.yorku.ca

(http://www.eecs.yorku.ca/course/6412/)

Class Time and Location

- Time:
  - Mon & Wed: 11:30am – 1:00pm
- Location:
  - Mondays: BERG 313
  - Wednesdays: BERG 211

2

Outline of this lecture

- Why data mining?
- What is data mining?
- Course information

3

Why Data Mining?

- Data flood
  - Vast amounts of data are produced every day
- Data are generated by:
  - Bank, telecom, other business transactions ...
  - Scientific fields: astronomy, biology, etc
  - Web (text, images, audio, video, etc.)
- Big Data era
  - The amount of data created every two days was estimated in 2013 to be 5 exabytes (5 billion gigabytes)
  - Big in 5Vs: - Volume, Velocity, Variety, Veracity, Value

4

Big Data Examples

- Library data
  - The U.S. Library of Congress Web Capture team claims that
    - As of April 2011, the Library has collected about 235 terabytes of data
    - And it adds about 5 terabytes per month.
- Climate science
  - In 2010, Germany's Climate Research Centre (DKRZ) generated 10,000 TB of data per year
- Didi Research (August 2017)
  - 70TB+ new data / day
  - 4500TB data processed / day
  - 20 billion+ routing requests / day
  - 14 billion location points /day

5

Big Data Examples

- Web
  - 1998: 26 million pages
  - 2003: Google searches 4+ billion pages, many hundreds TB
  - 2005: Google searches 8+ billion pages
  - 2008: 1+ trillion (1,000,000,000,000) unique URLs.
- Social networks
  - Twitter:
    - 2010: 50 million tweets per day
    - 2011: 200 million tweets per day
    - 2013: 500 million tweets per day (6,000 tweets per second)
  - Facebook:
    - Collect 600 terabytes of data every day (April 2014)

6

Big Data Examples

29 MILLION

NUMBER OF MAILS SENT EVERY SECOND

375 MEGABYTES

DATA CONSUMED BY HOUSEHOLDS EACH DAY

20 HOURS

VIDEO UPLOADED TO YOUTUBE EVERY MINUTE

24 PETABYTES

DATA PER DAY PROCESSED BY GOOGLE

500 MILLION

TWEETS PER DAY

700 BILLION

TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH

13 EXABYTES

DATA SENT AND RECEIVED BY MOBILE INTERNET USERS

398 ITEMS

PRODUCTS ORDERED ON AMAZON PER SECOND

7

Hidden Information in Data

- Very few data will ever be looked at/analyzed by human
- There is often information hidden in the data, which is
  - not readily evident
  - but useful
- For example:
  - What topics are people discussing in social media?
  - What types of credit card transactions are fraud?
  - What items are often bought together?
  - What types of customers likes a particular product?

8

Hidden Information is Valuable

- It can help companies to better understand and serve customers
  - E.g., recommendations made by Amazon or Netflix
- It can allow companies to optimize their processes
  - E.g., Uber is able to predict demand, dynamically price journeys and send the closest driver to the customers
- It can improve our health care
  - E.g., predict flu outbreaks and fast-track drug development
- It can help us to improve security
  - E.g., foil terrorist attacks and detect cyber crime

9

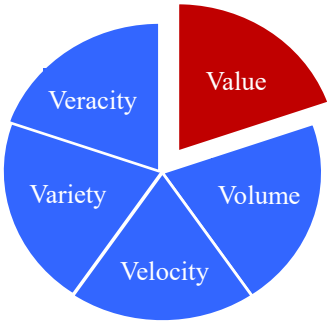
Gap between Data and Knowledge

- There is a huge gap between the stored data and the knowledge that could be constructed from the data
  - Many businesses have not made use of their data
  - Such businesses lose their competitive advantages
- Tools are needed to understand and make use of data
  - Translate the data into useful knowledge
- Data mining is such a tool.

10

Why Data Mining? (Summary)

- Wide availability of huge amounts of data
  - Big data era
- Imminent need for turning such data into useful information and knowledge



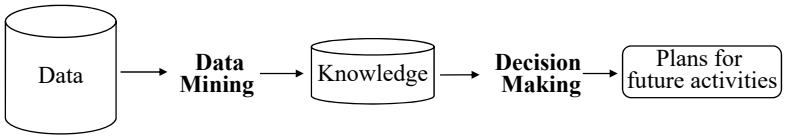
0.5% ever analyzed and used

(MIT Technology review)

11

What is Data Mining?

- Mining knowledge from data
  - Data = *raw* information
  - Knowledge = *patterns* or *models* 'behind' the data



12

What Is Data Mining?

- Data mining [Han, 2001]
  - process of extracting interesting (*non-trivial, implicit, previously unknown* and *potentially useful*) knowledge or patterns from data in *large* databases.
- Objectives of data mining:
  - Discover knowledge that characterizes general properties of data
  - Discover patterns on the previous data in order to make predictions on future data

13

Example 1

- Data about investors

Customer ID	Account Type	Margin Account	Transaction Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30–39	Tennis	40–59K
1013	Custodial	No	Broker	0.5	F	50–59	Skiing	80–99K
1245	Joint	No	Online	3.6	M	20–29	Golf	20–39K
2110	Individual	Yes	Broker	22.3	M	40–49	Fishing	40–59K
1001	Individual	Yes	Online	5.0	M	30–39	Golf	60–79K

- Possible business questions
  - Can we develop a general characterization/profile of different investor types? (characterization)
  - What characteristics distinguish between Online and Broker investors? (classification)
  - Can we develop a model which will predict the average trades/month for a new investor? (regression)

14

Example 2

Hypothetical Training Data for Disease Diagnosis

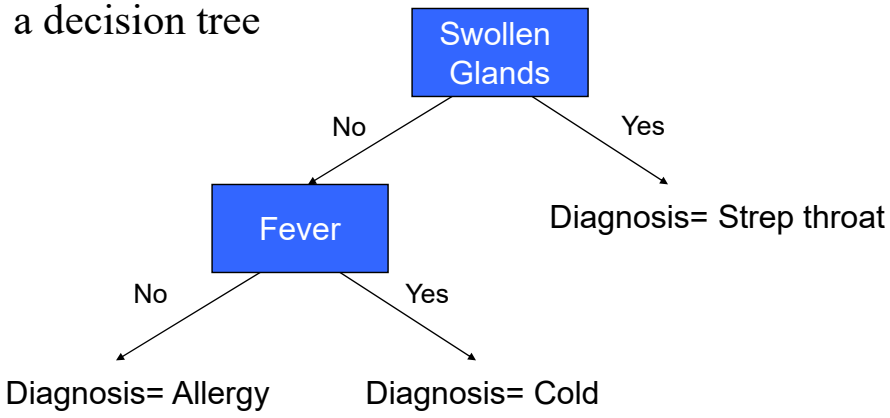
Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

- In this example dataset there are attributes corresponding to Symptoms, and an attribute of Diagnosis.
- The natural question is to predict the Diagnosis (class) [the Output variable] from the symptoms [the input variables].

15

Example 2 (Cont'd)

Pattern discovered:
a decision tree



16

Example 2 (Cont'd)

Use of the Decision Tree for Prediction

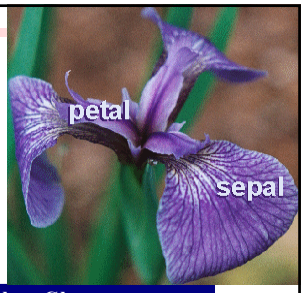
Data Instances with an Unknown Classification

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?

What are the predicted diagnoses?

17

Example 3



Part of training data for Iris plants classification

Sepal Length	Sepal Width	Petal Length	Petal Width	Class
5.1	3.5	1.4	0.2	Iris Setosa
4.9	3	1.4	0.2	Iris Setosa
4.7	3.2	1.3	0.2	Iris Setosa
5.4	3.9	1.7	0.4	Iris Setosa
4.8	3	1.4	0.1	Iris Setosa
7	3.2	4.7	1.4	Iris Versicolour
6.4	3.2	4.5	1.5	Iris Versicolour
5.5	2.3	4	1.3	Iris Versicolour
5	2	3.5	1	Iris Versicolour
5.4	3	4.5	1.5	Iris Versicolour
6	3	4.8	1.8	Iris Virginica
6.9	3.1	5.4	2.1	Iris Virginica
6.5	3	5.2	2	Iris Virginica
6.2	3.4	5.4	2.3	Iris Virginica
5.9	3	5.1	1.8	Iris Virginica

18

Generated Rules for Iris Plants Classification

Rules for Class = Iris Setosa (1 rule)

Rule 1: (petal_length<=1.9) → **(class = Iris Setosa)**

Rules for Class= Iris Versicolour (3 rules)

Rule 1: (1.9<petal_length<=4.9) ∧ (petal_width<=1.6) → **(class = Iris Versicolour)**

Rule 2: (sepal_length>5.8) ∧ (petal_length<=5.1) ∧ (petal_width<=1.7) → **(class = Iris Versicolour)**

Rule 3: (sepal_length>5.8) ∧ (sepal_width>3.0) ∧ (petal_length<=4.8) → **(class = Iris Versicolour)**

Rules for Class = Iris Virginica (4 rules)

Rule 1: (petal_length>4.8) ∧ (petal_width>1.7)→ **(class = Iris Virginica)**

Rule 2: (sepal_length>6.2) ∧(petal_length>5.0)→ **(class = Iris Virginica)**

:

Summary: Total number of rules = 8

Average length of rules = 2.75

Training accuracy = 99.33%

19

Example 4: Assessing Credit Risk

- Situation: Person applies for a loan
- Task: Should a bank approve the loan?
- Need to predict the credit risk of the person
  - people with bad credit are not likely to repay.

20

Example 4: Credit Risk - Results

- Banks develop credit models using a variety of data mining methods.
  - Learn the models from the previous data
  - (Successfully) predict if a person is likely to default on a loan
- Widely deployed in many countries

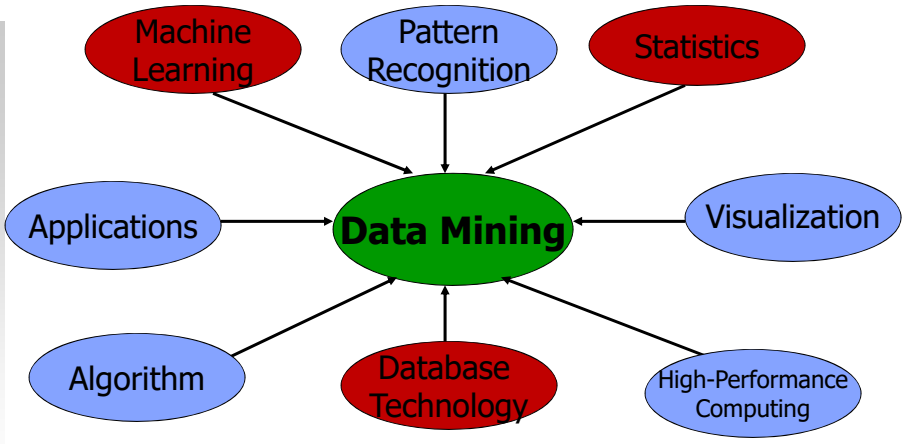
21

Example 5: Successful e-commerce

- A person buys a book (product) at amazon.com.
- Task: Recommend other books (products) this person is likely to buy
- Amazon uses data mining, e.g., frequent pattern mining on transaction histories and may find:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful

22

Data Mining: Confluence of Multiple Disciplines



23

Some Related Fields

- Machine Learning
  - Subfield of AI
- Exploratory data analysis
  - Subfield of Statistics
- Data warehousing and OLAP (Database)
  - Data summarization and aggregation
- Data mining
  - inherits techniques from ML and statistics
  - emphasizes processing very large data sets
  - finds more types of patterns and handle more types of data

24

Data Mining vs DBMS

- Database queries (handled by DBMS)
  - Find all credit card applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 of goods in the last month.
  - Find all customers who have purchased milk
- Data mining queries
  - Identify new credit card applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (clustering)
  - Find all items which are frequently purchased with milk. (association)

Data Mining vs Information Retrieval

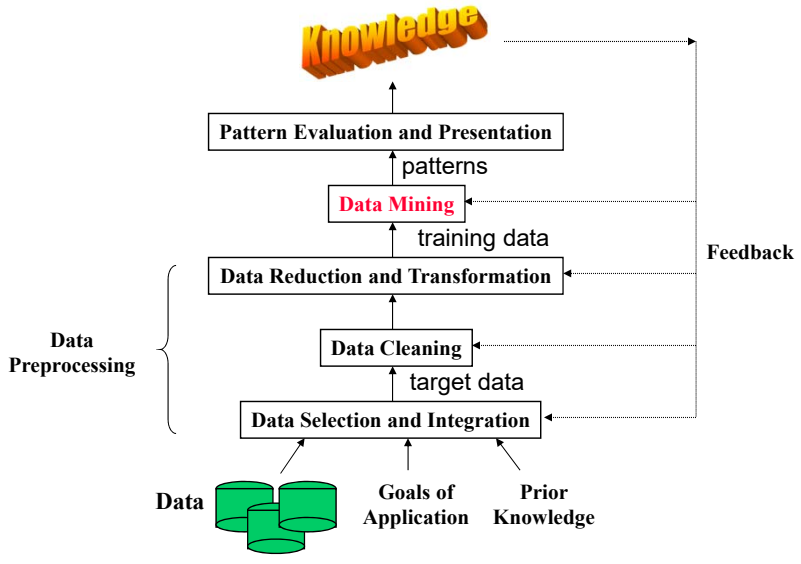
- Information Retrieval (IR)
  - Get back or retrieve information that is relevant to a user query from a collection of documents (or another type of data, such as WWW, images).
  - The information is presented usually in the same way it is stored.
- Data Mining
  - Reveal insight into the data that is not quite obvious.
  - Uncover some hidden patterns by exploring data

Alternative Name: KDD

- Knowledge discovery in databases (KDD)
  - used by AI, Machine Learning Community since 1989
- Data mining
  - Used by DB, business people since 1990
- Two names are now used interchangeably
- KDD is also considered as a process including data mining

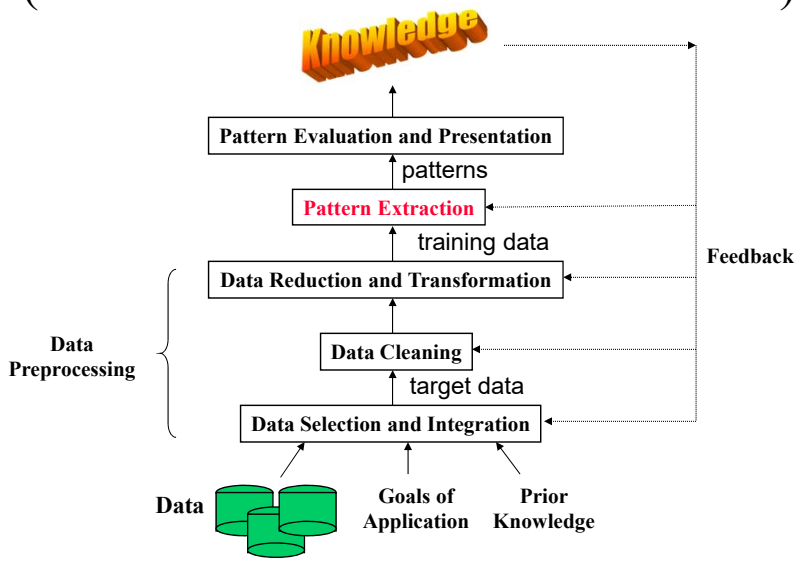
27

Data Mining: Core of a KDD Process



28

Process of Data Mining and KDD (When the two are considered the same)



29

Outline of this class

- *Why data mining?*
- *What is data mining?*
- **Course information**

30

Course Objectives

- To provide an introduction to data mining
- To understand the fundamental concepts, principles, algorithms and applications of data mining.
- To gain hands-on experience by
  - using data mining software to mine patterns from real world data
  - implementing some data mining algorithms.
- To prepare yourself for pursuing research in data mining and for big data jobs

31

Course Prerequisites

- Concepts of probability and its basic rules and theories.
- Knowledge of data structures
- Basic concepts of (relational) database systems

32

Course Content

- Introduction to data mining
- Association rule mining
- Sequential pattern mining
- Mining classification models
  - Data preprocessing
  - Decision tree learning
  - Decision rule learning
  - Bayesian classification
  - Neural networks
  - K-nearest neighbour method
  - Support vector machines (if time allows)
- Clustering analysis
- Applications of data mining, such as text mining
- Other topics if time permits, such as SVM, anomaly detection, sentiment analysis, Web mining, etc.

33

Reference Books and Materials

- Reference books
  - Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 3rd Edition, 2011.
  - Charu C. Aggarwal, Data Mining, The Textbook, Springer, 2015.
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.
  - Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, 2005.
  - The first book and many other books can be downloaded from a link at the course website
- Some conference/journal papers
  - To be posted on the course web site

34

Evaluation and Grading

- Assignments (25%)
  - 2 or 3 assignments
- Paper review and presentation (10%)
- Project (25%)
- Final exam (30%)
- Participation (10%)

35

Paper Review and Presentation

- Each student will select a data mining topic or a research paper from a list.
- Make a 40 minute presentation to the class
  - What are the objectives of the paper/topic? What problems does it solve?
  - A brief literature review for the subject area.
  - What techique(s) or algorithm(s) does the paper propose?
  - In-dept desciption of the technique or algorithm.
  - Experimental results for the technique or algorithm if the paper reports them.
  - What are the advantages of the technique or algorithm? What are the limitations of the technique or algorithm?
  - What are the open (unresolved) issues in the research?

36

Project

- Categories of projects
  - Research-oriented
  - Application-oriented
  - Implementation and evaluation-oriented
- Components:
  - Project proposal (1-2 pages)
  - Design of your solution
  - Implementation and experiments
    - Using Java, Python, C++ or C (or another language with good reason)
  - Written project report (in a research paper format)
  - Class presentation of project (10-15 minutes)
  - System demo if we have time
- Can be done in groups of 2 students
- Detailed requirements will appear at the course web site

37

Major Data Mining Journal and Conferences

- Journals
  - Data Mining and Knowledge Discovery
    - http://www.acm.org/sigmod/dblp/db/journals/datamine/index.html
  - IEEE Transactions on Knowledge and Data Engineering
    - http://www.acm.org/sigmod/dblp/db/journals/tkde/index.html
  - ACM Transactions on Knowledge Discovery from Data
    - http://tkdd.acm.org/
- Data mining conferences
  - ACM SIGKDD, IEEE ICDM, SIAM DM, PKDD/ECML, PAKDD
- Database conferences
  - ACM SIGMOD, VLDB, IEEE ICDE, EDBT, ACM CIKM

38

Data Mining News and Data Sets

- Kdnuggets
  - http://www.kdnuggets.com
- ACM SIGKDD
  - http://www.acm.org/sigkdd/
- ACM SIGMOD
  - http://www.acm.org/sigmod/
- Data Sets (see course web page for the links)
  - UCI machine learning repository
  - KDD Cup
  - Kaggle
  - Frequent itemset mining repository

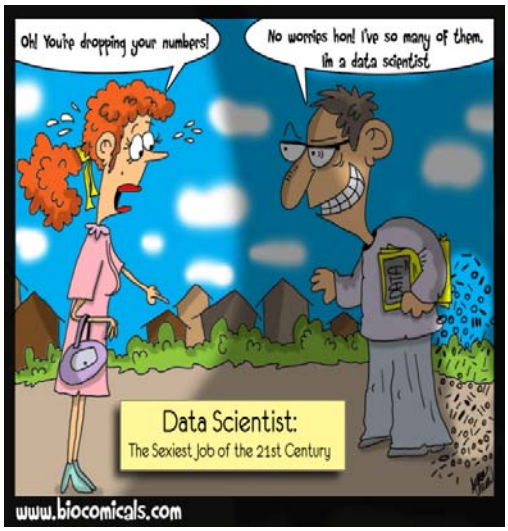
39

Instructor Information

- Office hours
  - Mondays: 1:00pm-2:00pm
- Office location
  - LAS 2048
- Email: aan@cse.yorku.ca or aan@eecs.yorku.ca
- Course web site:
  - http://www.eecs.yorku.ca/course/6412/

40

Data Scientist



Data Scientist:

The Sexiest Job of the 21st Century

Harvard
Business Review
Oct. 2012