CS653 – Data Mining

Instructor: Xiaobai Liu

SDSU Machine Vision and Perception Lab

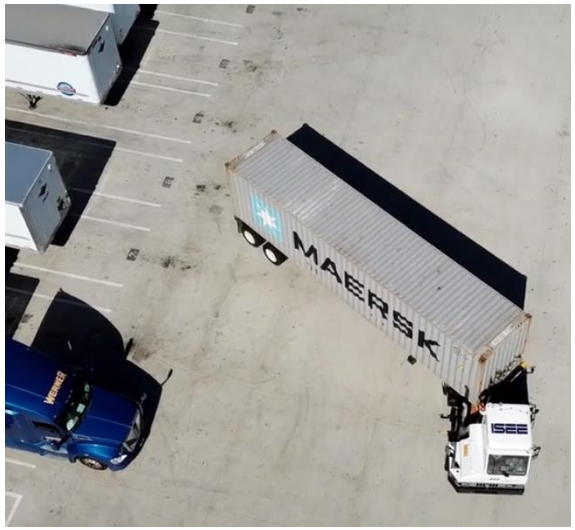
SDSU MVP Lab:

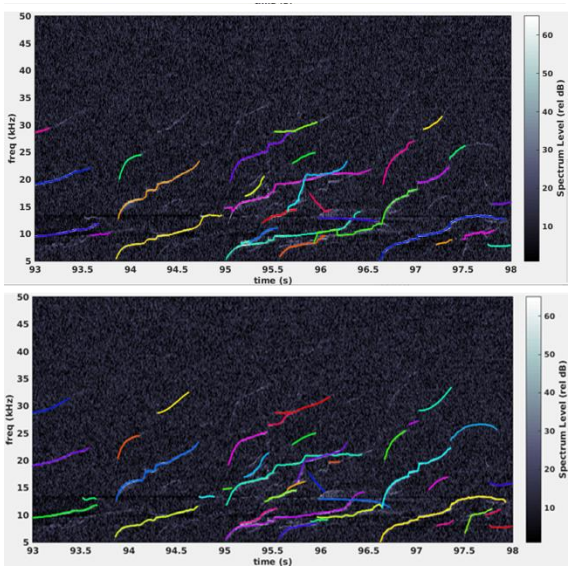
PI: Dr. Xiao-Bai Liu

Three doctoral students,
Five Master Students

Research Areas:

AI, Robotics, Autonomous-Driving and Bioacoustics





General Information

- Email: Xiaobai.liu@sdsu.edu
- Office Hours: 3:30-4:15 AM, Tuesday/Thursday
- Office: GMCS #547

Course structure

- The course has four parts:
  - Lectures - Introduction to the main topics
  - Homework Assignments
  - Exams
  - Course Project
- Lecture slides are available on SDSU Canvas

Grading

- Homework Assignments (HAs, 10%)
  - HA1
  - HA2
  - HA3
  - HA4
  - HA5
- In-class Quiz (10%)
- Mid-term exam (25%)
- Final Exam (45%)
- Course project (10%)

Prerequisites

- CS 210 Data Structure
- Data mining involves many Inter-discipline topics
  - Machine learning
  - Pattern recognition
  - Statistics
  - Probability theory

Teaching materials

- **Recommended textbook**
  - Web Data Mining: Exploring Hyperlinks, Contents and Usage data. By Bing Liu, Second Edition, Springer, ISBN 978-3-642-19459-7. (Online Version, Free Download)
  - Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Pearson/Addison Wesley, ISBN 0-321-32136-7.
- References:
  - Data mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber, Morgan Kaufmann, ISBN 1-55860-489-8.
  - Principles of Data Mining, by David Hand, Heikki Mannila, Padhraic Smyth, The MIT Press, ISBN 0-262-08290-X.
  - Machine Learning, by Tom M. Mitchell, McGraw-Hill, ISBN 0-07-042807-7

Syllabus

Week 1	**Introduction**
Week 2	Data Pre-processing
Week 3	Association Rules
Week 4	
Week 5	Supervised Learning
Week 6	
Week 7	Unsupervised Learning
Week 8	
Week 9	Social network analysis
Week 10	
Week 11	Opinion Mining and sentiment analysis
Week 12	
Week 13	Recommender systems and collaborative filtering
Week 14	
Week 15	Final Project

Programing Language

- Python 3
  - Not a programming course
  - Self-study materials will be provided
  - Homework assignment 1

Important Dates

- Mid-term exam, Oct 16, 2025
- Final exam, Dec 16, 2025, 1-3pm
- Course Project Due, December 11,2025 (last day of classes)

Feedback and suggestions

- Your feedback and suggestions are most welcome!
  - I need it to adapt the course to your needs.
  - Let me know if you find any errors in the class material.
- Share your questions and concerns with the class – very likely others may have the same.

A brief Introduction of Data Mining

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Regression [Predictive]
- Abnormal Detection [Predictive]

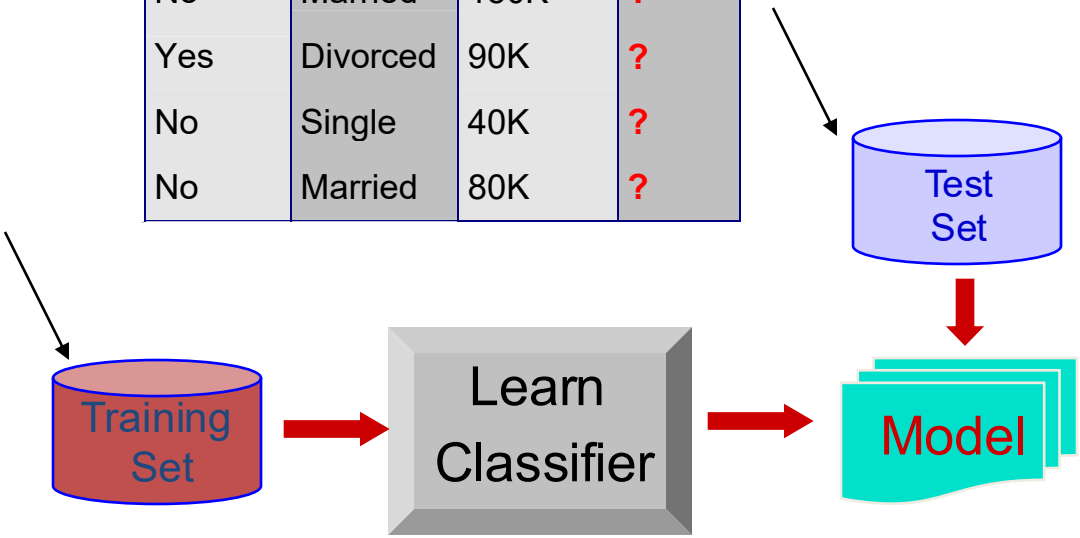
Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

	categorical	categorical	continuous	class
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**

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

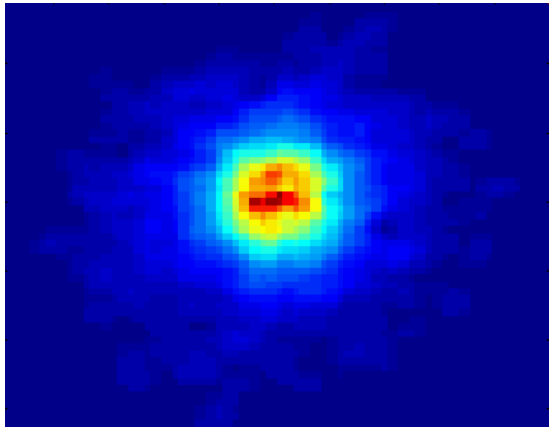
Classification: Application 3

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

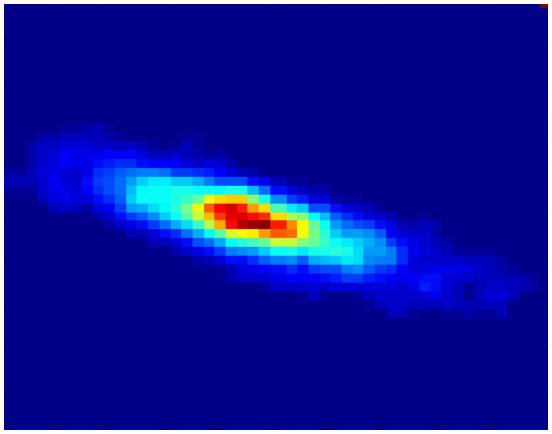
Early



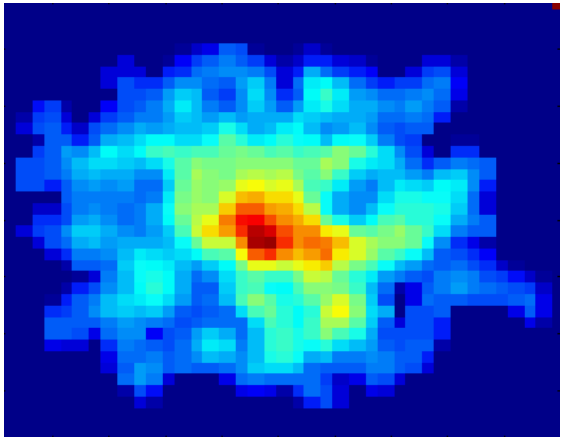
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Data Mining Tasks...

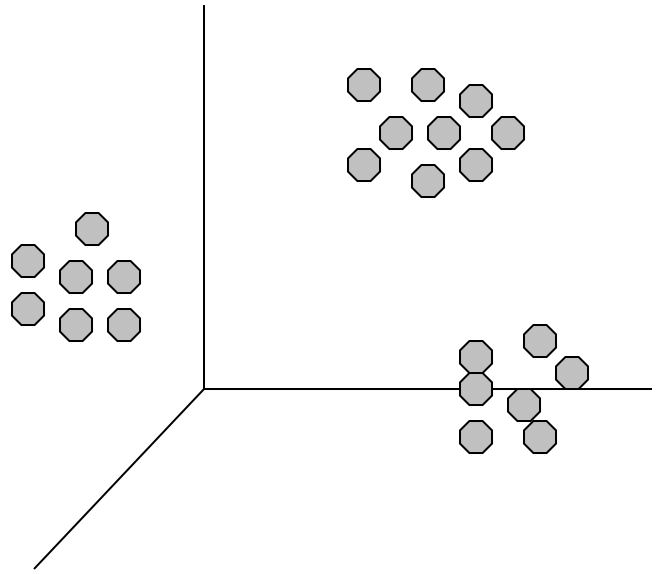
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Regression [Predictive]
- Abnormal Detection [Predictive]

Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

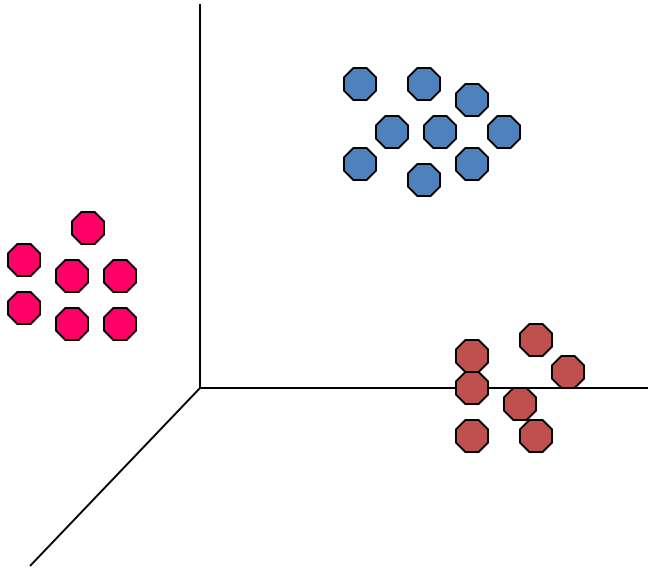


Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	*Total Articles*	*Correctly Placed*
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Regression [Predictive]
- Abnormal Detection [Predictive]

Association Rule Discovery: Definition

- Given a set of records each of which contains 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}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper, then he is very likely to buy milk.

Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Mining

- Sequential pattern mining: A sequential rule: A$\rightarrow$ B, says that event A will be immediately followed by event B with a certain confidence

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Regression [Predictive]
- Abnormal Detection [Predictive]

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

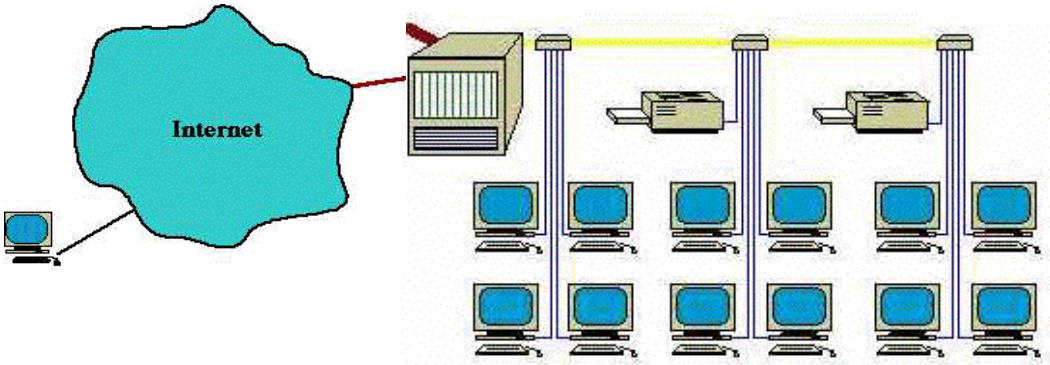
Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Abnormal Detection [Predictive]

Deviation/Anomaly Detection

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

  - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Resources

- ACM SIGKDD
- Data mining related conferences
  - Data mining: KDD, ICDM, SDM, …
  - Databases: SIGMOD, VLDB, ICDE, …
  - AI: AAAI, IJCAI, ICML, ACL, …
  - Web: WWW, WSDM, …
  - Information retrieval: SIGIR, CIKM, …
- Kdnuggets: http://www.kdnuggets.com/
  - News and resources. You can sign-up!

Review: what is data mining?

- Data mining is also called *knowledge discovery and data mining* (KDD)
- Data mining is
  - extraction of useful patterns from data sources, e.g., databases, texts, web, images, etc.
- Patterns must be:
  - valid, novel, potentially useful, understandable

Review: Classic data mining tasks

- Classification:

  mining patterns that can classify future (new) data into known classes.

- Association rule mining

  mining any rule of the form $X \rightarrow Y$, where X and Y are sets of data items. E.g.,

  Cheese, Milk$\rightarrow$ Bread [sup =5%, confid=80%]

- Clustering

  identifying a set of similarity groups in the data

Classic data mining tasks (contd)

- Deviation detection:

  discovering the most significant changes in data

- Data visualization: using graphical methods to show patterns in data.

Review: why is KDD important?

- Computerization of businesses produce huge amount of data
  - How to make best use of data?
  - Knowledge discovered from data can be used for competitive advantage.
- Online e-businesses are generating even larger data sets
  - Online retailers (e.g., amazon.com) are largely driving by data mining.
  - Web search engines are information retrieval (text mining) and data mining companies

Review: why is data mining necessary?

- Make use of your data assets
- There is a big gap from stored data to knowledge; and the transition won't occur automatically.
- Many interesting things that one wants to find cannot be found using database queries

  "find people likely to buy my products"

  "Who are likely to respond to my promotion"

  "Which movies should be recommended to each customer?"

Review: why data mining?

- The data is abundant.
- The computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every company is doing (or has to do) it

Review: related fields

- Data mining is an multi-disciplinary field:
  - Machine learning
  - Statistics
  - Databases
  - Information retrieval
  - Visualization
  - Natural language processing
  - etc.

Review: data mining applications

- Marketing, customer profiling and retention, identifying potential customers, market segmentation.
- Engineering: identify causes of problems in products.
- Scientific data analysis, e.g., bioinformatics
- Fraud detection: identifying credit card fraud, intrusion detection.
- Text and web: a huge number of applications …
- Any application that involves a large amount of data …

Pre-review: Data mining (KDD) process

- Understand the application domain
- Identify data sources and select target data
- Pre-processing: cleaning, attribute selection, etc
- Data mining to extract patterns or models
- Post-processing: identifying interesting or useful patterns/knowledge
- Incorporate patterns/knowledge in real world tasks

A popular domain: text mining

- Data mining on text
  - Due to online texts on the Web and other sources
  - Text contains a huge amount of information of almost any imaginable type!
  - A major direction and tremendous opportunity!
- Main topics
  - Text classification and clustering
  - Information retrieval
  - Information extraction
  - Opinion mining