VICTORIA UNIVERSITY OF WELLINGTON
*Te Whare Wananga o te Upoko o te Ika a Maui*

School of Engineering and Computer Science

COMP 422 — Week 1

# Introduction to Data Mining and KDD Process

Mengjie Zhang (Meng)

*mengjie@ecs.vuw.ac.nz*

# Introduction to Data Mining

- Why DM/KDD?
- KDD vs DM
- Examples
- Process of KDD
- Relationship to other disciplines
- DM vs data warehousing
- DM vs query tools
- Mining complex types of data: multimedia, time-series, text, WWW
- Challenges/Problems

# Why Data Mining

- *Data comes like water out of a fire hydrant. You can't drink it (Anon).*
- *We are drowning in information but starving for knowledge (John Naisbett).*
- Hardware advances in data collection and storage have far outpaced software advances in data analysis and manipulation.
- Organizations collect more data than they can handle.
- Data that may never be analyzed is still collected out of fear of missing something that might be important.
- As databases grow, decision making directly from their contents is not feasible; knowledge derived from the data is needed.
- Supermarket chains, credit card companies, banks routinely generate daily volumes of 100MB.
- Scientific and remote sensing instruments collect gigabytes of data everyday.

# Is Data Mining Really Applied?

- Is data mining really applied or is it only hype?
- Yes. But only in recent years.
- Why not earlier?
- Applicable machine learning techniques The sudden rise of interest in DM become possible.
- Over the past few years, learning techniques have expanded enormously: Neural networks, genetic algorithms, genetic programming, ...
- KDD/DM conferences: Pacific Asian, International,...

# KDD/DM

- *Knowledge Discovery in Databases (KDD)* is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad].
- The non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data [Adriaans]
- “Golden Nuggets”
- *Data Mining (DM)* is a part of the KDD process relating to methods for extracting patterns from data [Fayyad].
- *Data Mining* is a problem solving methodology that finds a logical or mathematical description, of a complex nature, of patterns and regularities in a set of data [Decker and Focardi].

# KDD and DM (Continued)

- Data Mining is often related to learning/adaptive algorithms and methods.
- In some current usage, KDD = DM.
- Knowledge extraction, information discovery, information harvesting, data archaeology, data dredging, data pattern processing, image classification, object detection/recognition, ...
- KDD/DM is not new techniques but rather a multi-disciplinary field of research: all make a contribution (later)

# Examples of Nuggets

- Fraudulent credit card transactions
- Good/bad loan risks
- New class of stars
- Put beer and disposable nappies together and you may sell more of each
- Put perfume and greeting cards together and you’ll sell more of each
- Inspect credit card transactions, find people who brought scuba gear and lessons and send discount coupons for Carribean cruise

# Examples of Nuggets (Continued)

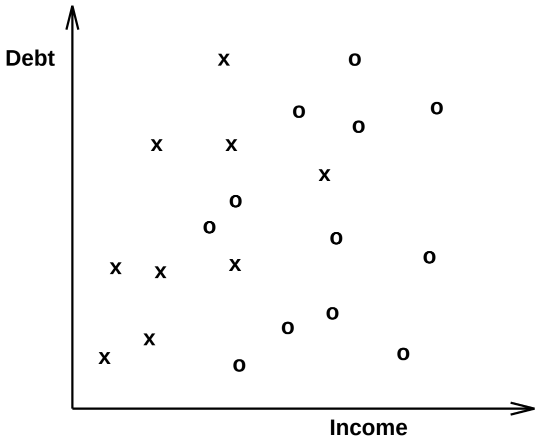
- Recognition of specific market segments that respond to specific characteristics
- Ineffective advertising
- Recognition of a particular face in a database of photographs
- Finding all cyclones in a database of satellite images
- Detection of tumors in a database of X-rays
- Detection of haemorrhages and micro-aneurisms from a set of retina images.

# DM Example

Consider the data from a loan database:

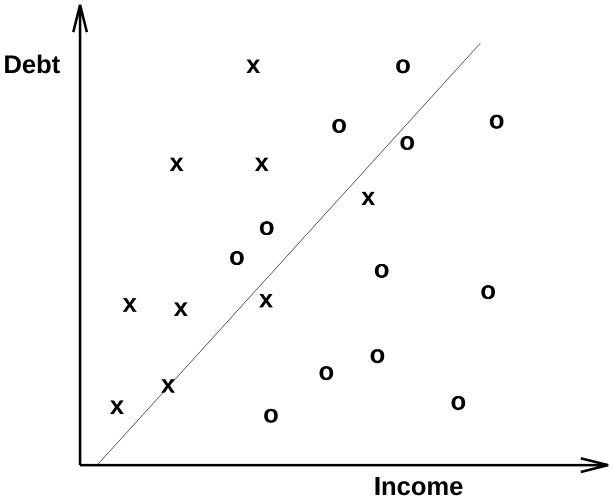
| Income | Debt | Defaulted? |
|---|---|---|
| $20,000 | $1,000 | No |
| $50,000 | $25,000 | Yes |

SCATTER PLOT



What can be discovered? Patterns? Regularities?

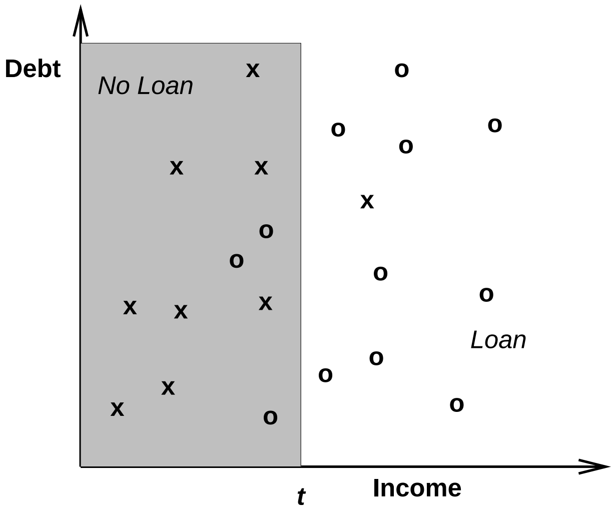
# Regression Line



$y = mx + c$
$Debt = m \cdot Income + c$

- Does the regression line tell us anything?
- The correlation coefficient?

# Threshold

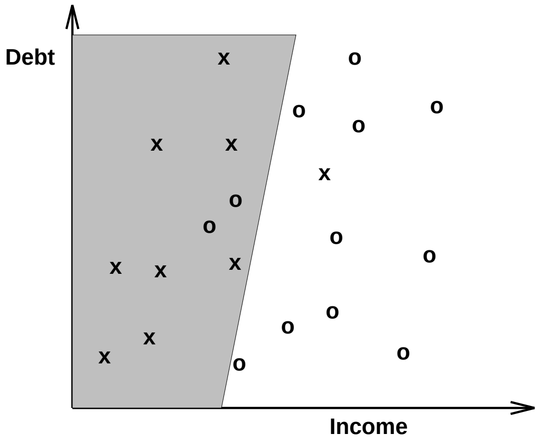


Suppose that we expect the classification rule is:

$$\text{if (Income} > t\text{) then grant loan}$$

- How many errors would we make on this data?
- How many errors on new customers?
- Can we find a better rule?

# Linear Decision Boundary



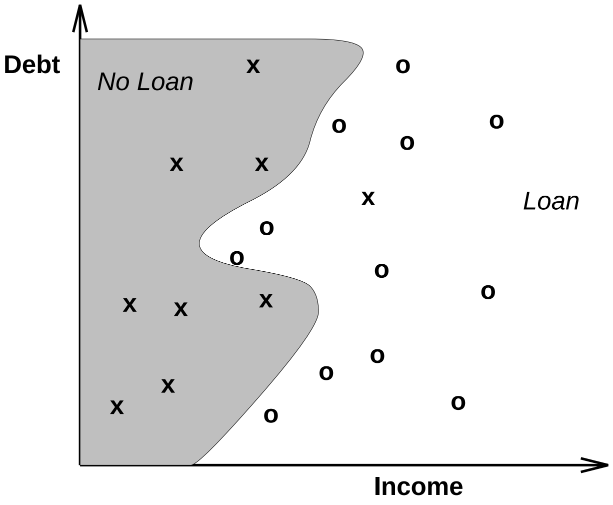
If the equation of the separating line is

$$Debt = \alpha \cdot Income + \beta$$

we can 'extract' the classification rule:
**if** $Debt < \alpha \cdot Income + \beta$ **then** *grant loan*

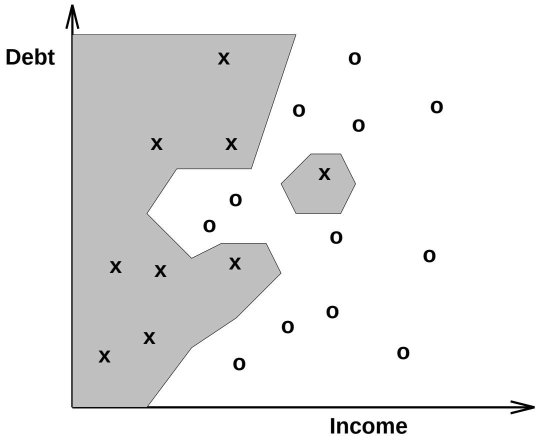
- How many errors would we make on this data?
- How many errors on new customers?
- How can we find the best line?
- Can we find a better rule?

# Non Linear Regions



- How many errors would we make on this data?
- How many errors on new customers?
- How can we find the best curve? What is its equation?
- Neural Networks can give us the regions but not the equation of the separating curve.

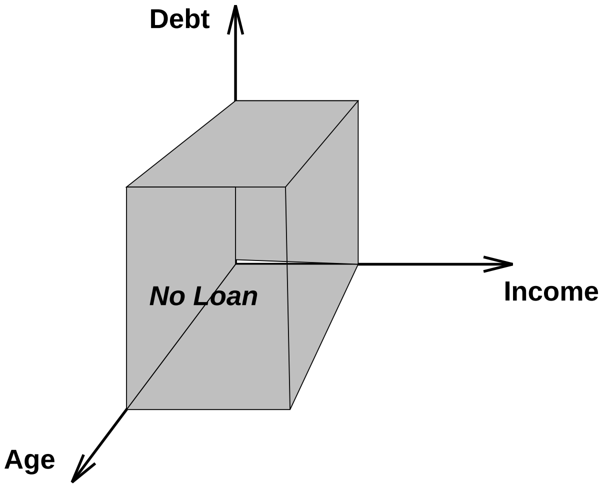
# Nearest Neighbour



- Each unknown point is given the classification of its closest neighbours.
- How many errors on new customers?
- How can we find the best curve? What is its equation?
- Nearest neighbour can give us the regions but not the equation of the separating curve.

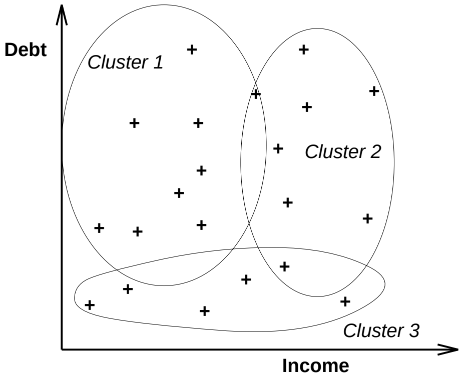
# Higher Dimensions

- Suppose we believe that older people are more likely to pay off loans than younger people. We can include age in the decision.

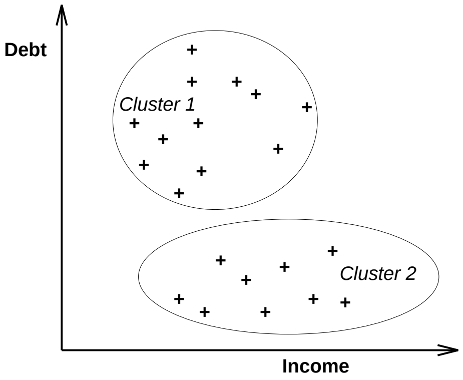


- If we also include bank balance how can we visualize the result?
- How can we include male vs female and other non numeric data?

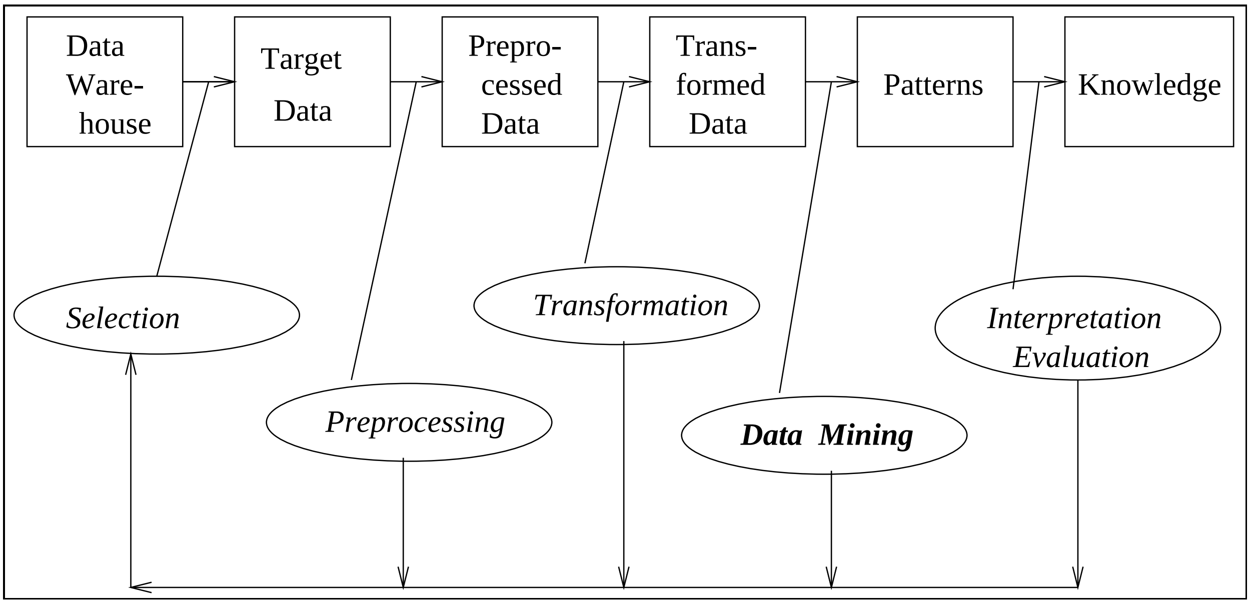
# Clustering



- We seek interesting and useful groupings.
- The clusters above are not much good, those below look more useful.



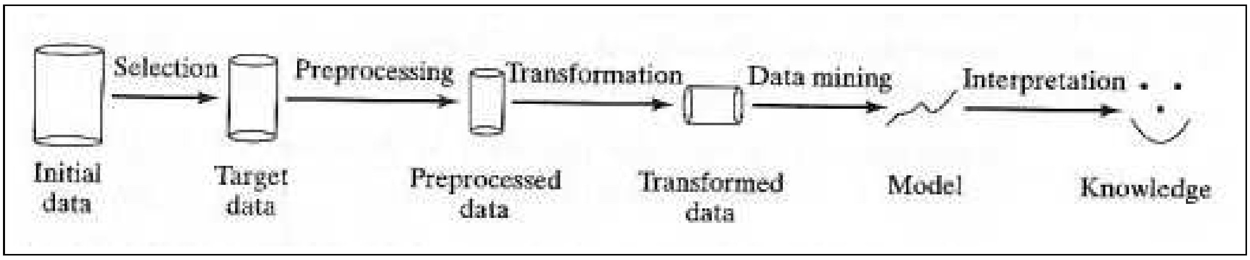
# Process of KDD



# Process of KDD (Continued)

1. Develop an understanding of the application domain: relevant prior knowledge, goals and priorities of the end user.
2. Create target data set: Which variables should be used?
3. Data cleansing and preprocessing: Remove noise, outliers, missing fields, coding of time sequence information, known trends.
4. Data reduction and projection: Determine the most relevant features, derive useful features, dimensionality reduction transformations.
5. Choose data mining tasks: Classification? regression? clustering? trend analysis? model fitting? association discovery?
6. Choose/develop data mining methods.
7. Apply data mining to extract patterns, models, etc.
8. Interpretation and evaluation of patterns.
9. Use the discovered knowledge (to the new data sets).

# Process of KDD



[Dunhan 2003]

# Process of KDD

- Selection: The data needed for the DM/KDD process may be obtained from many different and heterogeneous data sources. The first step obtains the data from various DBs, files, and non-electronic sources.
- Preprocessing: for incorrect, missing data, conflict data (from different sources), ...
- Transformation: data from different sources (with different formats) are converted into a common format. Also consider *data reduction, feature selection and extraction*.
- **Data mining:** Based on the data mining task being performed, this step applies algorithms to the transformed data to generate the desired results.
- Interpretation/evaluation: interpret the results/hidden patterns — symbolic rules, visualisation, etc.

# Applications and Methods

- Predictive Modeling/Classification
  - (Symbolic) Decision Tree induction
  - (Symbolic) Rule Induction
  - Neural Classifier
  - Evolutionary Classifier – genetic algorithms and genetic programming
- Database Segmentation/Summarization
  - Symbolic Clustering
  - Bayesian Clustering
  - Neural Clustering
  - Evolutionary Clustering???
- Link/Association Analysis
- Deviation detection
- Dependency Modeling
- Visualization, ...

# Choice of Data Mining Methods

Main factors which influence choice of data mining methods are:

- Kind of input data
  - Numeric only
  - Symbolic only
  - Mixed Symbolic and numeric
- Supervised vs unsupervised
  - Each input record has a pre-assigned class (supervised)
  - No pre-assigned classes (unsupervised)
- Output of the Method
  - A decision tree
  - A list of rules
  - A mathematical formula
  - A program
  - A black box

# Components of DM Algorithms

- Model Representation: The “language” for decision patterns (equations, decision trees, neural nets,...)
  - Too simple and nothing can be discovered.
  - Too complex and results are hard to interpret and overfitting becomes likely.
- Model Evaluation
- Search Method
  - Parameter Search: Given that we have fixed on a model type, how do we get the best parameters, e.g. if we decide on a linear decision boundary, how do we find the best line?
  - Model search: What model type would be best, e.g. linear or curved boundary, or nearest neighbour model?

# Types of KDD

- Top-down Discovery: The analyst suggests hypotheses and patterns to look for. Results are analyzed for support for a hypothesis.
- Bottom-up Discovery: The system automatically explores the database and suggests patterns supported by the data. The analyst determines whether the patterns are significant or not.
- Mixed: The analyst focuses on an area of search, the system proposes potentially significant patterns, the analyst frames new hypotheses in the light of the patterns......

# Data Mining vs Data Warehousing

- **A data warehouse** is a subject-oriented, integrated, time-variant, and non-volatile collection od data in support of management's decisions. [Inmon92]
- Data mart: smaller, local data warehouses. A data mart is more specialised, more accessible, and lot of smaller than an enterprise-wide data warehouse. It is often used as the first step for many organisations.
- DM is *often* (not always!) discussed as an after-market for data warehouses and/or data marts.
- There are two ways of performing DM techniques:
  - Directly on the existing data warehouses/data marts
  - By extracting the part of the information which is of interest to the end-user from the existing data warehouses/data marts

# Data Mining vs DB Query Tools

- A DB query can be viewed as a simple DM task
- Example: a query from an employement DB — *find out all the people names whose salary is more than $100,000*
- what a DM task?
- Queries in DB applications are usually well defined with precise results. DM applications are often vaguely defined with imprecise results.
- Basic DB queries always output either a subset of the DB or aggregates of the data. A DM tool often outputs a KDD object.
- KDD object: a rule, a decision tree, a neural network, a program,
- KDD objects are **not** part of the DB, does **not** exist before executing the DM algorithm/tool.

# Data Mining vs DB Query Tools

- DM tools and query tools are complementary
- A data mining tool does not replace a query tool, but it does give a lot of additional possibilities
- e.g. a large DB containing millions of records that describe your customers' purchases over the last ten years — there is a wealth of potentially useful knowledge:
  - Who bought which product on what date?
  - what is the average turnover in a certain sales region in July?
  - What is the optimal segmentation of my clients?
  - How do I find the most important different customer profiles?
- Use normal query tools (SQL) or DM methods (NNs GAs)?
- Knowledge/patterns to find are hidden?
- SQLs may take days/months, DM algorithms could find the answer automatically within a short time (minutes/hours)
- If you know exactly what you are looking for, use SQL; if you know only vaguely what you are looking for, turn to DM.

# Summary

- KDD/DM definitions
- KDD process
- Choice of DM methods
- Applications and examples
- Data mining vs data warehousing, data mining vs query tools
- Questions to think:
  - DM related fields, the relationship between DM and these fields
  - DM tasks
  - How to find these solutions?