UE 141 Discovery Seminar on Data Mining

Spring 2013

Jing Gao
SUNY Buffalo

Data Mining

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data

Why Data Mining?

Data

Goal

Knowledge ⟶ Action/Decision

Why Data Mining?

Data

Goal

Data Mining

Knowledge → Action/Decision

Pattern Mining

- ## Data

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- ## Goal
  - Have customers purchase more products while they are in the supermarket!
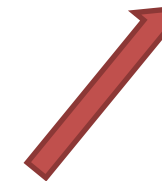
Pattern Mining: From Data to Knowledge to Decision

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke
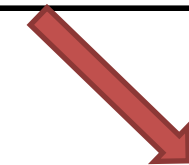
{Bread, Milk}

{Diaper, Beer}

......

Put these items close to each other!

Buy more items!!

Classification

- ## Data

user	age	gender	education	Ad?
	27	Female	Bachelor	Yes
	30	Male	PhD	Yes
	55	Male	Bachelor	No

- ## Goal
  - Increase profit while maintaining advertising cost!
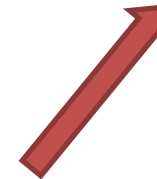
Classification: From Data to Knowledge to Decision

user	age	gender	education	Ad?
	27	Female	Bachelor	Yes
	30	Male	PhD	Yes
	55	Male	Bachelor	No

If a customer is younger than 30, then he/she is likely to buy the product.

Send advertisement only to these customers!

Increase profit!!

- **Data** **Clustering**

user	items
	orange, banana, apple, water
	orange, apple, water
	rice, bread, milk, eggs
	bread, milk, eggs, water
	yogurt, milk, eggs

- **Goal**
  - Increase profit while maintaining advertising cost!

Clustering: From Data to Knowledge to Decision

user	items
	orange, banana, apple, water
	orange, apple, water
	rice, bread, milk, eggs
	bread, milk, eggs, water
	yogurt, milk, eggs

Group 1

Group 2

Group 1: Bob and Alice bought lots of fruits

Group 2: Mary, Mike and Joe bought bread, eggs, milk

Target marketing!
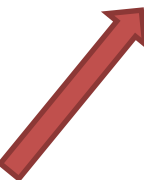
Increase profit!!

Other Examples

- ## Applications
  - Business, Biomedical, Engineering, ...

- ## Examples
  - Netflix collects user ratings of movies (data) => What types of movies you will like (knowledge) => Recommend new movies to you (action) => Users stay with Netflix (goal)
  - Gene sequences of cancer patients (data) => Which genes lead to cancer? (knowledge) => Appropriate treatment (action) => Save life (goal)
  - Road traffic (data) => Which road is likely to be congested? (knowledge) => Suggest better routes to drivers (action) => Save time and energy (goal)

Data Mining Software

Data		Data Mining Software		Knowledge

Input   effectiveness   Output

efficiency

Data Miner

Formulate data mining problems

↓

Use existing software to solve data mining problems

↓

Understand existing data mining techniques, design and implement data mining software

↓

Develop new techniques for data mining tasks
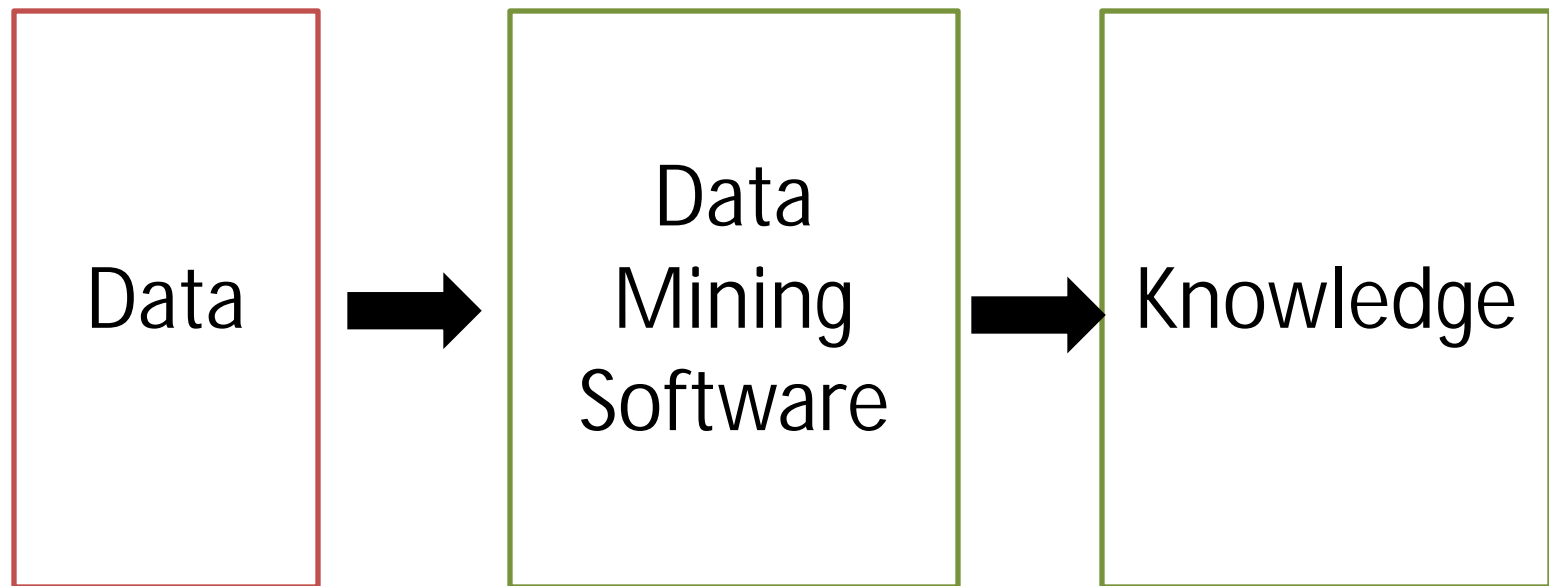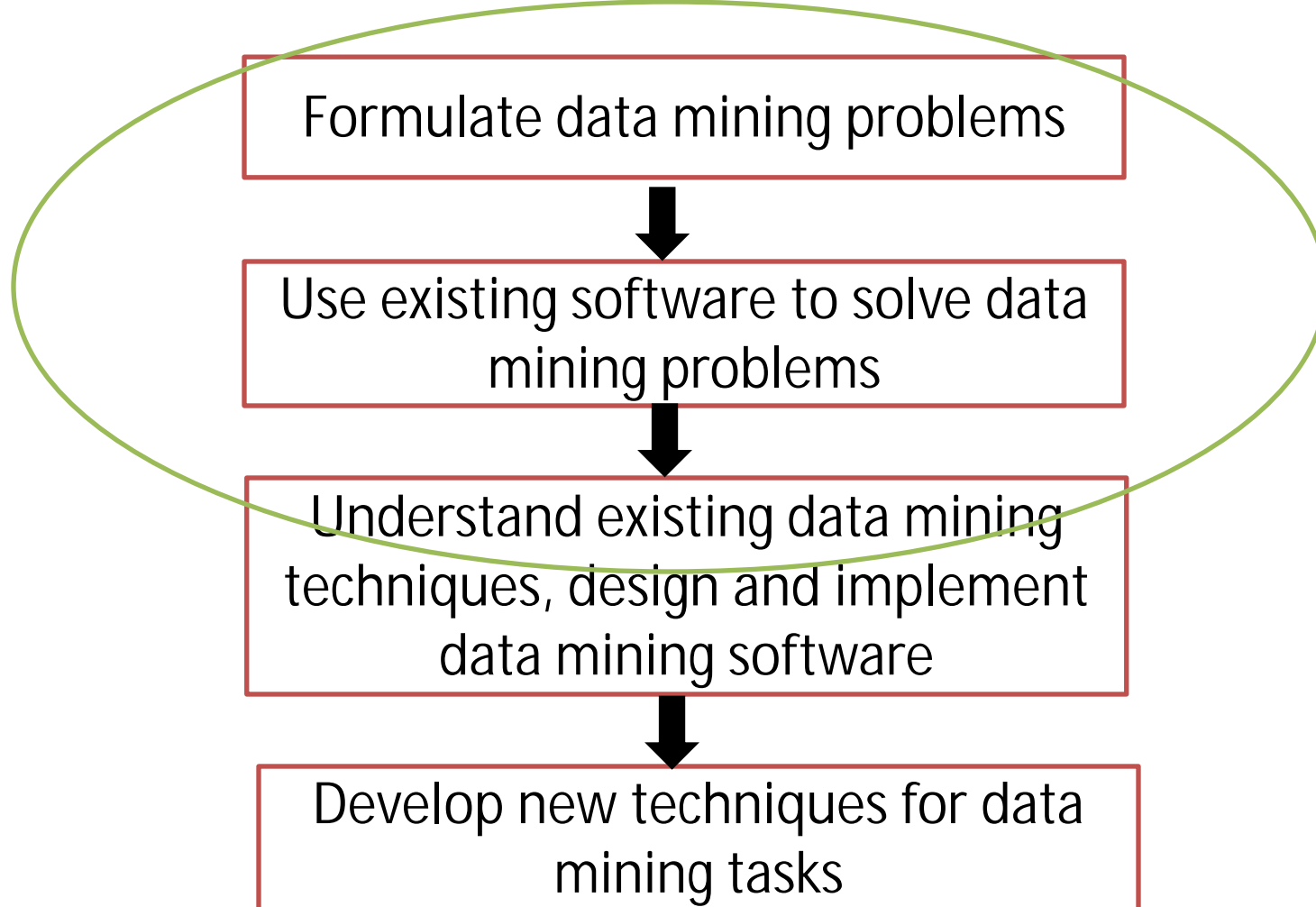
Terms & Courses

- ## Terms
  - Data mining techniques, methods, approaches: The steps to be taken to extract knowledge from data
  - Data mining program: Instructions for computers to conduct a particular data mining task
  - Data mining software: a collection of programs

- ## Suggested courses
  - This course: Formulate data mining problems and use existing software
  - Data mining techniques: Data Mining, Machine Learning, Pattern Recognition, Statistical Learning
  - Data mining programs: Programming, Data Structure, Algorithms
  - Data mining research: Statistics, Optimization, Linear Algebra, Matrix, Cloud computing
  - General data science relevant courses: Database, Information Retrieval