CS5228 LECTURE 1: INTRODUCTION

Bryan Hooi

School of Computing

National University of Singapore

OUTLINE

OUTLINE

- Course Logistics
- What is Data Mining?
- Data Mining Approaches
- Data Mining Concepts
- Preprocessing

COURSE INFORMATION

Lectures: Fridays 6.30pm - 8.30pm, LT18

Announcements and course materials will be posted on LumiNUS

If you have questions:

- Ask on LumiNUS forums
- Questions about Assignment 1: email Wang Yiwei (e0409763@u.nus.edu)
- Questions about Assignment 2: email Wang Wenjie (wangwenjie@u.nus.edu)
- Other questions: ask me (bhooi@comp.nus.edu.sg)

COURSE STAFF

Lecturer:

- Bryan Hooi
  Email: bhooi@comp.nus.edu.sg
  Office: COM2-03-15
  Office hours: Thursdays 4pm, or by request

TAs:

- Wang Yiwei
  Email: e0409763@u.nus.edu
- Wang Wenjie
  Email: wangwenjie@u.nus.edu
- Siddharth Bhatia
  Email: siddharth@comp.nus.edu.sg

ASSESSMENT

Assignment 1: worth 25% (due 6 Mar)

Assignment 2: worth 25% (due 27 Mar)

Group Project: worth 50% (due 17 Apr)

ASSIGNMENTS

Assignments will involve programming, as well as theoretical / conceptual questions

Python is the primary programming language

Discussion is allowed, but all code and write-ups must be done **individually**

Submission should be via LumiNUS

One late period can be used for either assignment

- This extends the deadline to the following Monday 11:59pm
- Late submissions beyond this point will incur 10% deduction per day
- No need to send any emails to use it, just submit 1 of your assignments late

PROJECT

Group project (2-3 students per group)

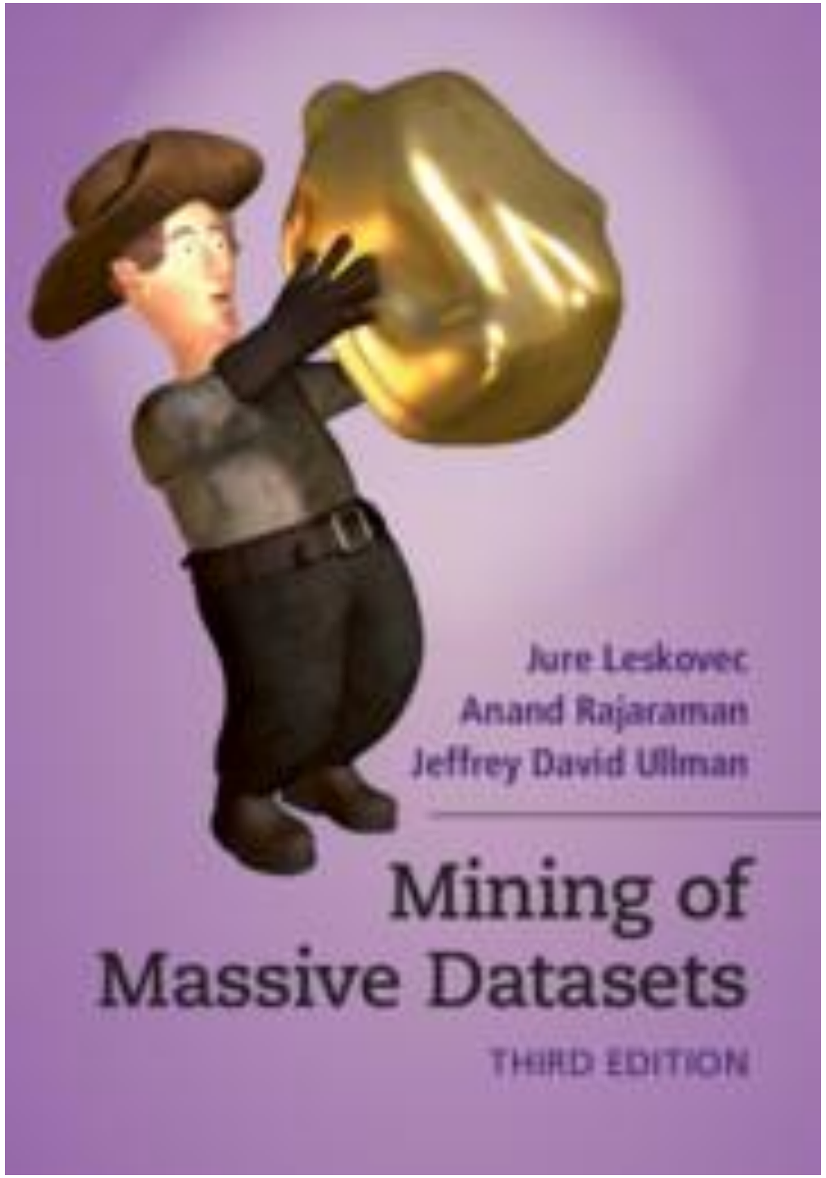
There will be a few topics which you can choose from:

- Kaggle-in-class competition
- Data analysis projects using pre-selected datasets from various domains (e.g. financial, medical, sports, …)
- Self-proposed project
  - If you are interested in this, please come talk to the course staff first to make sure the task is reasonable, does not require excessive data cleaning, etc.

REFERENCE

Textbook (useful but not required):

- Mining of Massive Datasets. Jure Leskovec, Anand Rajaraman, Jeff Ullman
- Freely available online: http://www.mmds.org/



COURSE OBJECTIVES

By the end of the course, you should expect to:

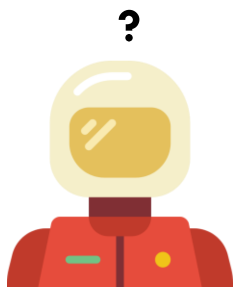
- Have a good knowledge of fundamental **concepts** and **algorithms** of data mining
- Be able to **apply** them to perform data mining tasks for new applications in practice

COURSE OBJECTIVES

By the end of the course, you should expect to:

- Have a good knowledge of fundamental **concepts** and **algorithms** of data mining
- Be able to **apply** them to perform data mining tasks for new applications in practice

UserID	Height (m)	Country	…
1	1.61	SG	…
2	1.50	US	…
3	***NA***	MY	…
…	…	…	…



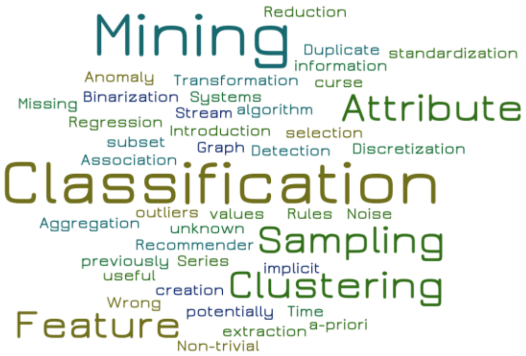
COURSE OBJECTIVES

By the end of the course, you should expect to:

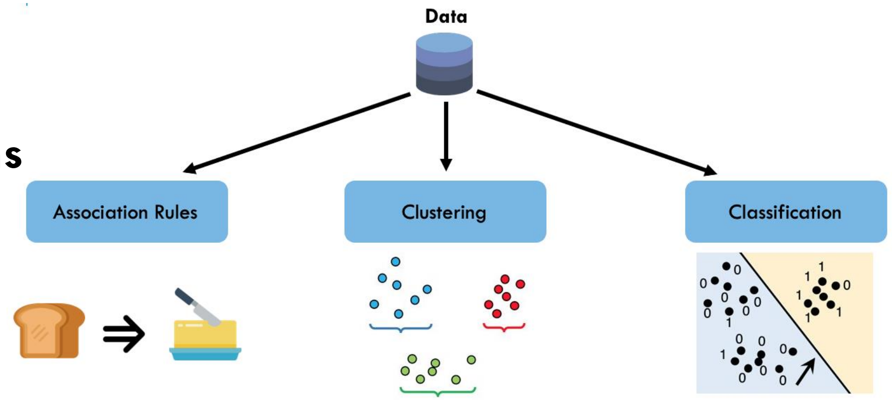
- Have a good knowledge of fundamental **concepts** and **algorithms** of data mining
- Be able to **apply** them to perform data mining tasks for new applications in practice

UserID	Height (m)	Country	…
1	1.61	SG	…
2	1.50	US	…
3	***NA***	MY	…
…	…	…	…

Concepts



Algorithms



LESSON PLAN

Week	Date	Topics	Due Dates
1	17 Jan	Introduction	
2	24 Jan	Association Rules	
3	31 Jan	Clustering 1	
4	*7 Feb*	*No Class (Conference)*	
5	14 Feb	Clustering 2	
6	21 Feb	Classification 1	
Recess	*28 Feb*		
7	6 Mar	Classification 2	Assignment 1 Due
8	13 Mar	Classification 3	
9	20 Mar	Recommender Systems	
10	27 Mar	Graph Mining	Assignment 2 Due
11	3 Apr	Stream Mining	
12	*10 Apr*	*No Class (Good Friday)*	
13	*17 Apr*	*No Class (Conference)*	Project Due

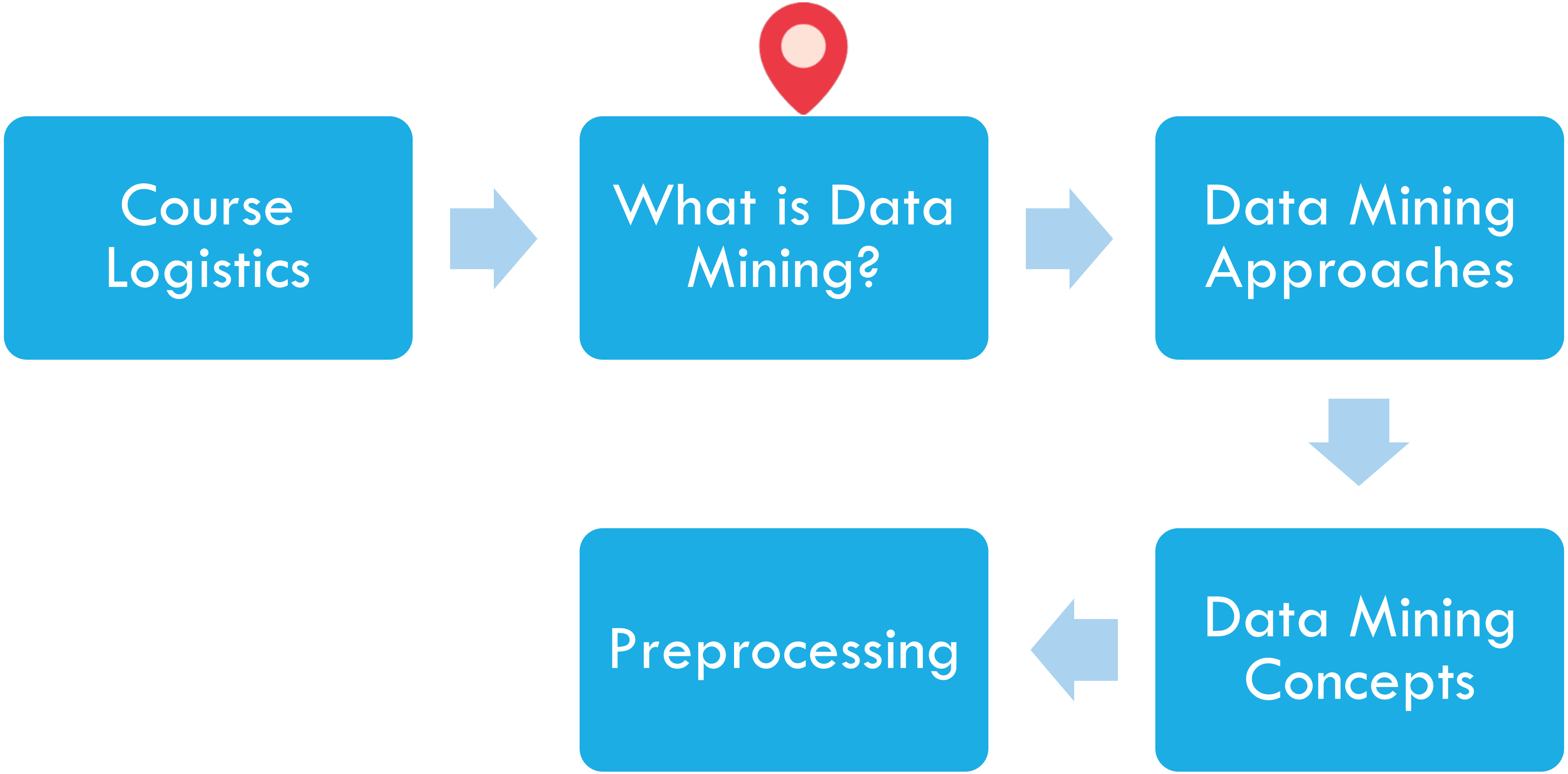
Introducing Activity flows

Treat different states of each activity as a separate slide, so you can engage and instruct using only the keyboard. Check any multiple choice activity then press "Insert as..." below.

Dismiss

ANY QUESTIONS?

OUTLINE


Course Logistics
What is Data Mining?
Data Mining Approaches
Data Mining Concepts
Preprocessing

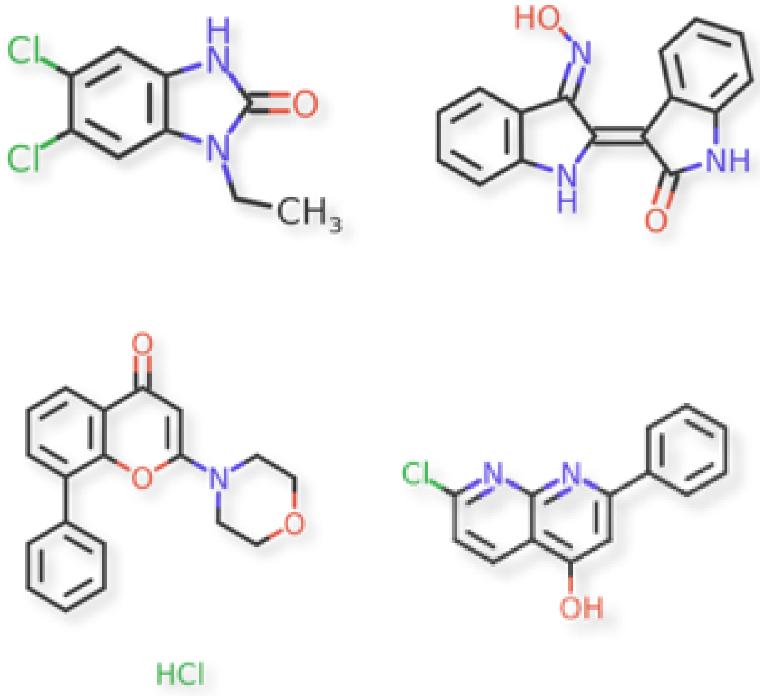
WHAT IS DATA MINING?

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

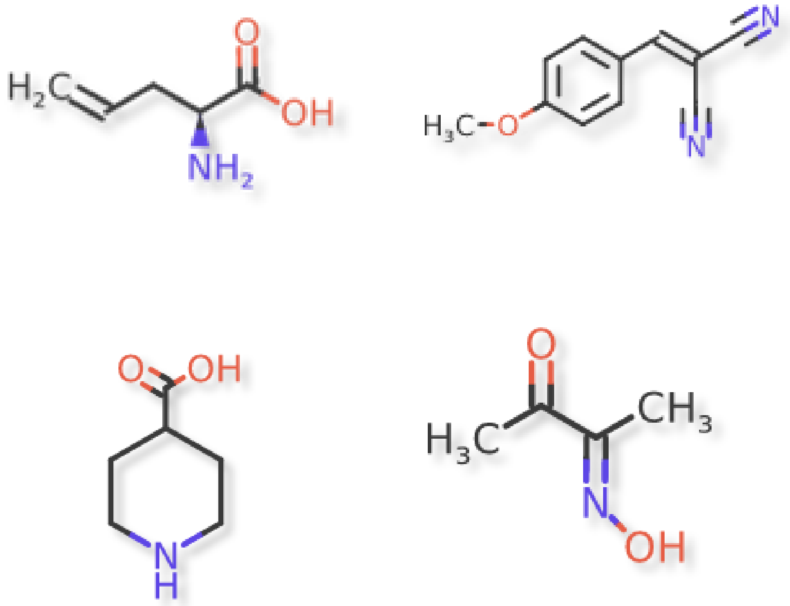
William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus

Q: WHAT RULE CHARACTERIZES TOXIC MOLECULES?

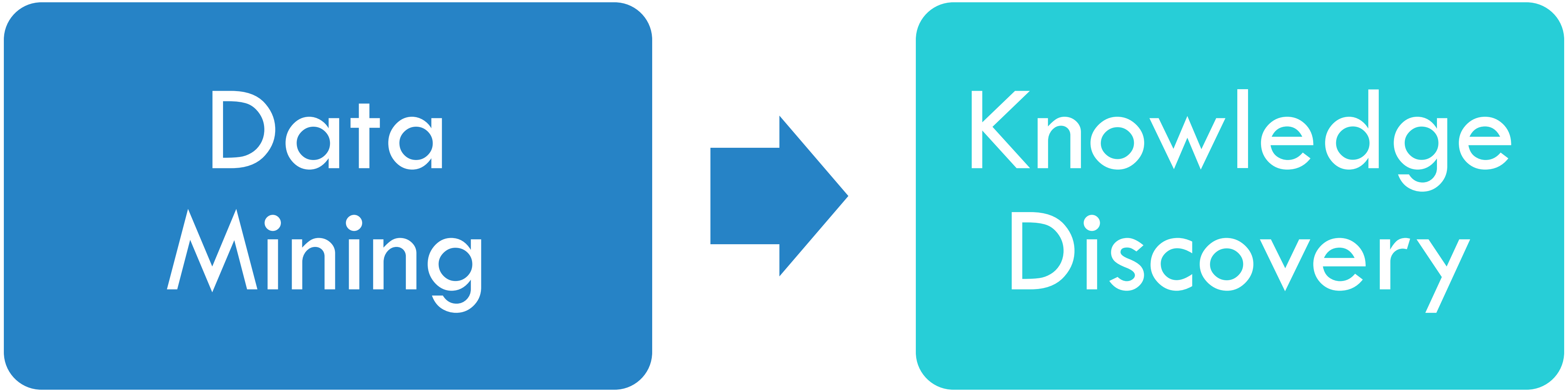
Toxic



Non-toxic



WHAT IS DATA MINING AND KNOWLEDGE DISCOVERY?



(Approach) (Goal)

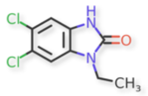
THE DATA MINING PROCESS

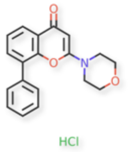
- Feature Selection
- Normalization

- Clustering
- Classification
- Association Rules

- Visualization
- Pattern Interpretation

THE DATA MINING PROCESS





Data

Preprocessing

- Feature Selection
- Normalization

Data Mining

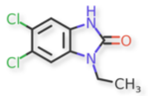
- Clustering
- Classification
- Association Rules

Postprocessing

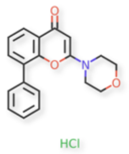
- Visualization
- Pattern Interpretation

Knowledge

THE DATA MINING PROCESS



(2, 0, 0, 1, …)



(4, 1, 0, 1, …)

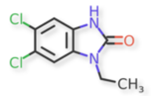
Data → **Preprocessing** → **Data Mining** → **Postprocessing** → Knowledge

Preprocessing
- Feature Selection
- Normalization

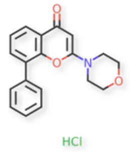
Data Mining
- Clustering
- Classification
- Association Rules

Postprocessing
- Visualization
- Pattern Interpretation

THE DATA MINING PROCESS



(2, 0, 0, 1, …)



(4, 1, 0, 1, …)

if  > 1:
    toxic
else
    nontoxic

Data

Preprocessing

- Feature Selection
- Normalization

Data Mining

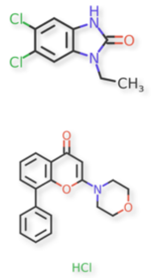
- Clustering
- Classification
- Association Rules

Postprocessing

- Visualization
- Pattern Interpretation

Knowledge

THE DATA MINING PROCESS



(2, 0, 0, 1, …)

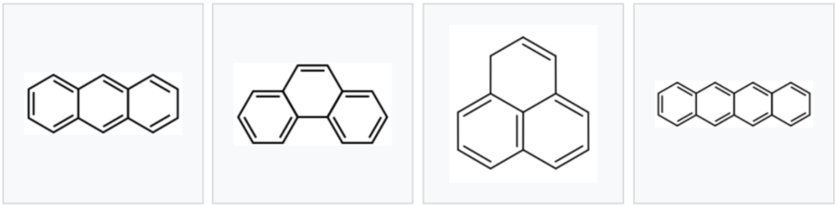
(4, 1, 0, 1, …)

if  > 1:
    toxic
else
    nontoxic

Polycyclic aromatic hydrocarbon

From Wikipedia, the free encyclopedia

Principal PAH Compounds



Data

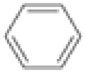
→ Preprocessing → Data Mining → Postprocessing → Knowledge

Preprocessing
- Feature Selection
- Normalization

Data Mining
- Clustering
- Classification
- Association Rules

Postprocessing
- Visualization
- Pattern Interpretation

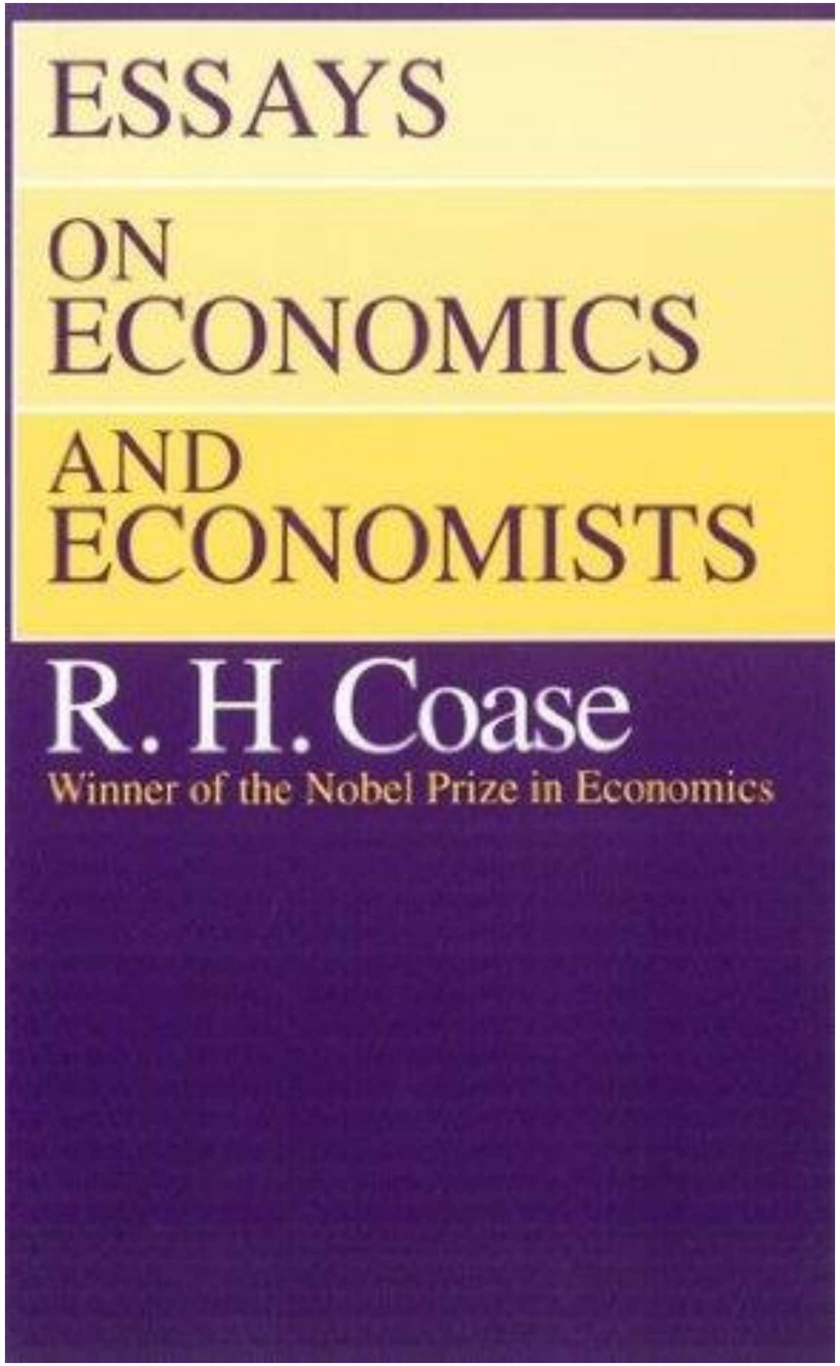
```
if <benzene> > 1:
    toxic
else
    nontoxic
```



HOW DO WE KNOW IF OUR PATTERNS ARE MEANINGFUL?

If you torture the data long enough, it will confess to anything.

R. H. Coase



OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

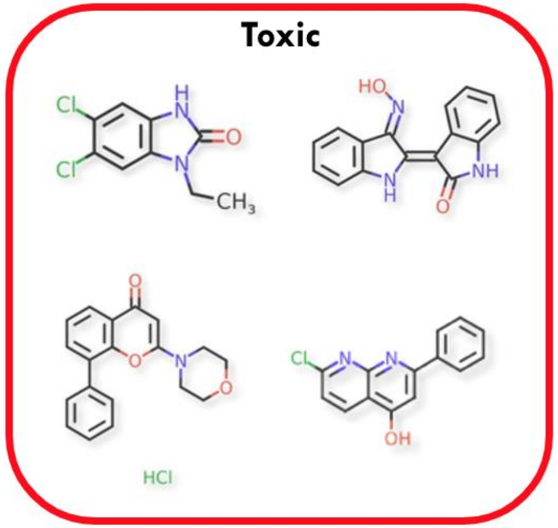
Article | Authors | Metrics | Comments | Related Content

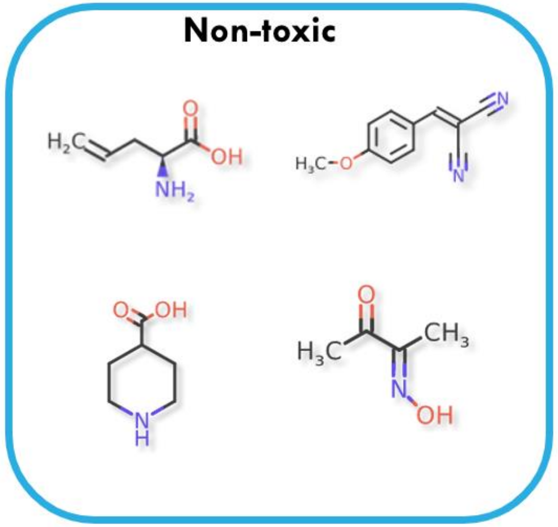
HOW DO WE KNOW IF OUR PATTERNS ARE MEANINGFUL?

Patterns should be **generalizable**: i.e. they should remain accurate on new, unseen data

If the training data is too small or biased, this can lead to lack of generalizability.

if  > 1:
  toxic
else
  nontoxic



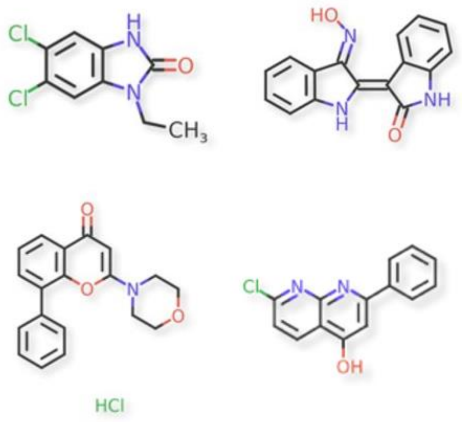


HOW DO WE KNOW IF OUR PATTERNS ARE MEANINGFUL?

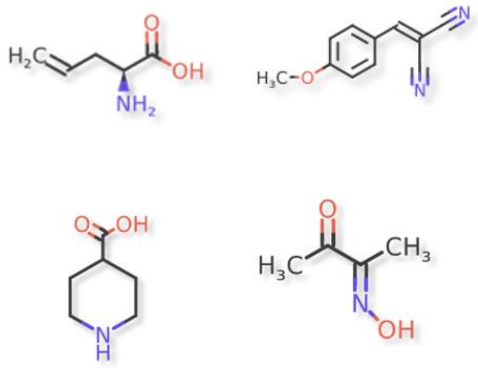
<u>Bonferroni's Principle</u> (roughly): if you look for more patterns than your dataset can support, you are bound to find ***false positives*** (patterns that are not actually present)

if  > 1:
    toxic
else
    nontoxic

Toxic



Non-toxic



OUTLINE

Course Logistics

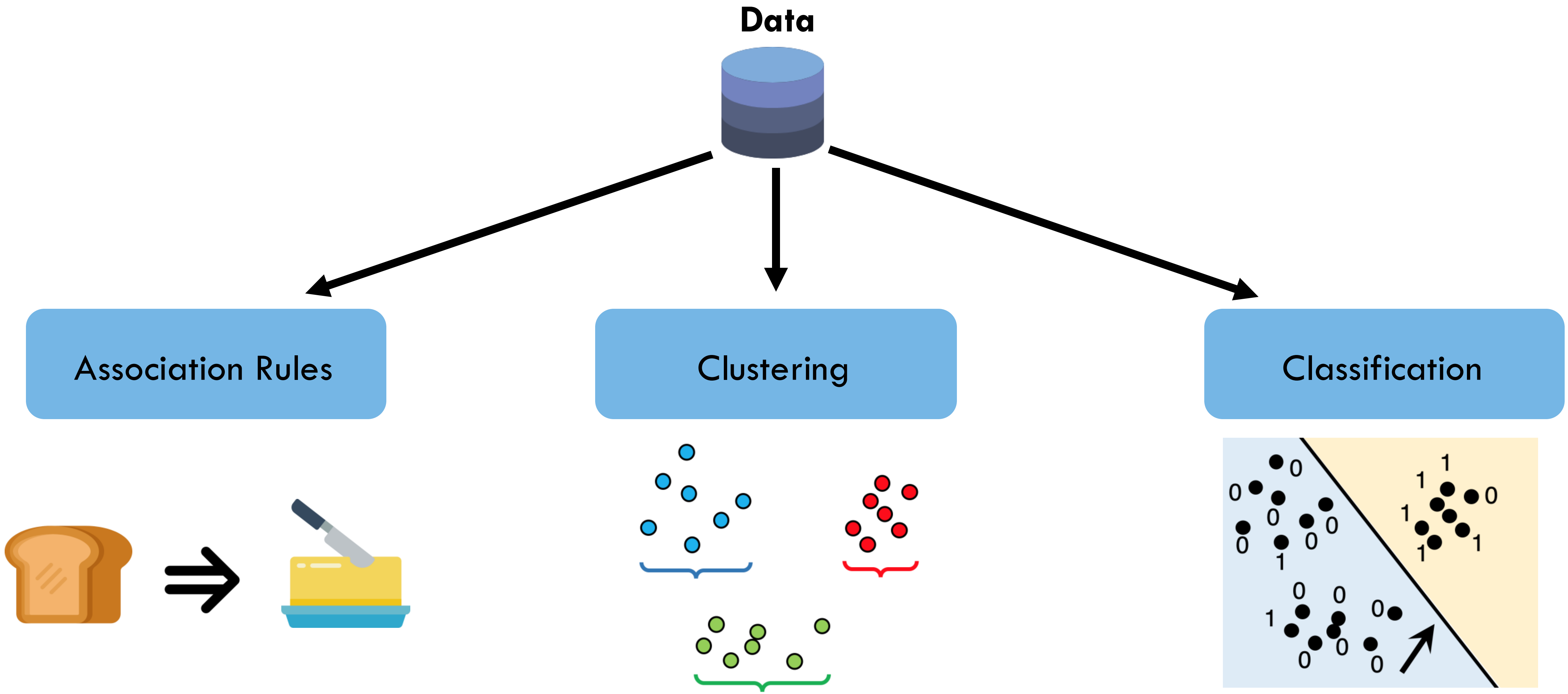
What is Data Mining?

Data Mining Approaches

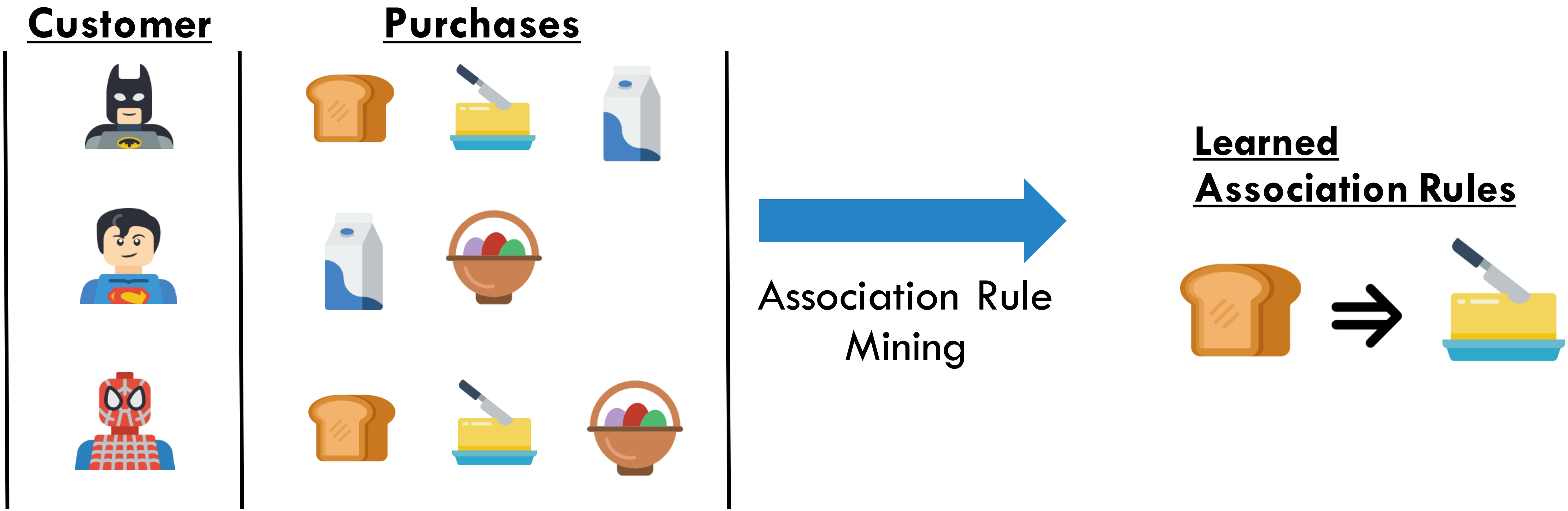
Data Mining Concepts

Preprocessing

DATA MINING APPROACHES



ASSOCIATION RULE MINING



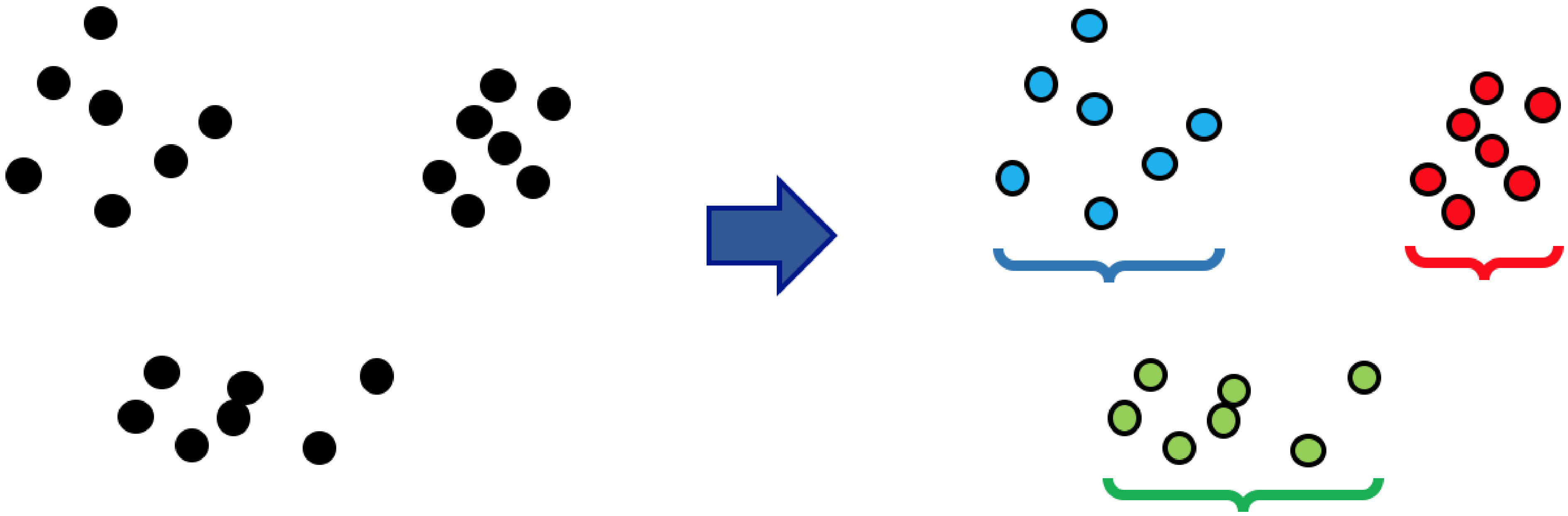
Goal: Given multiple records (e.g. sets of items bought by customers), find **rules** that predict occurrence of an one item based on occurrences of others

ASSOCIATION RULES: EXAMPLE APPLICATIONS

Market basket analysis: e.g. for inventory management

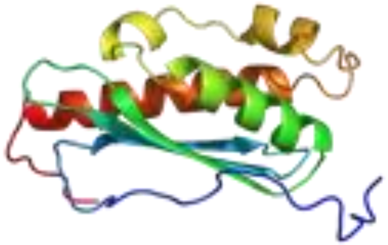
Medical: finding rules relating patient symptoms, test results and diseases

CLUSTERING



Goal: Separate a set of objects into groups of similar points (low **intra-cluster** distances; high **inter-cluster** distances)

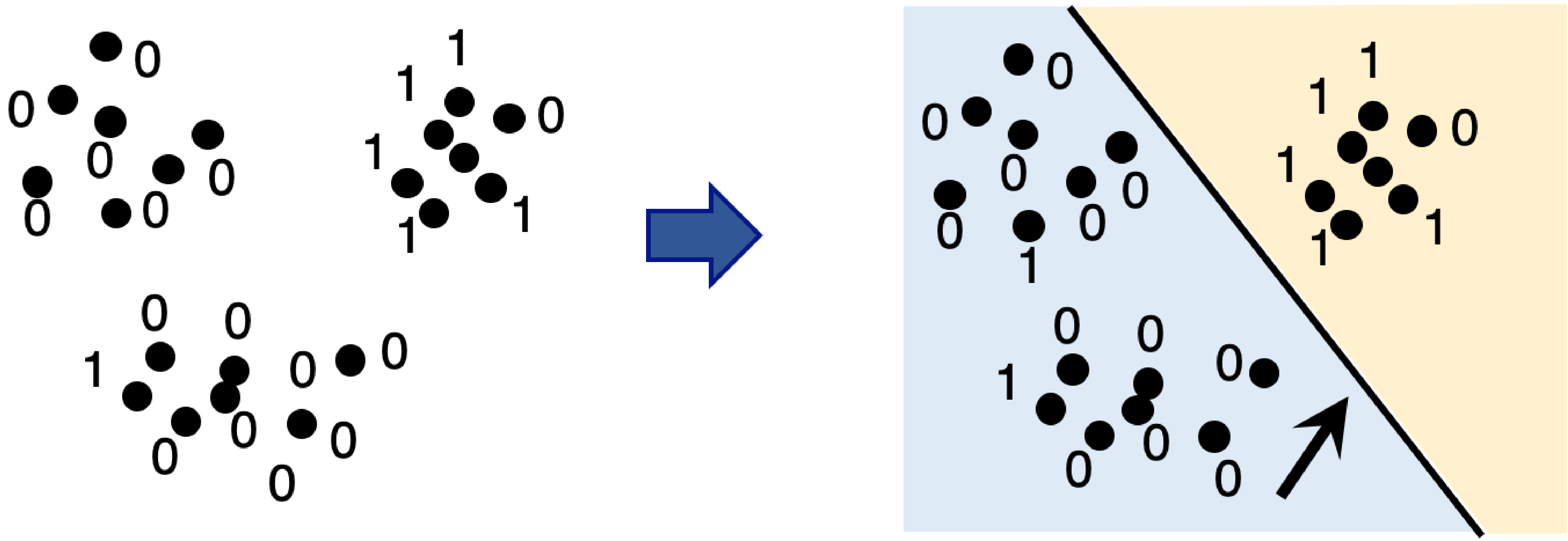
CLUSTERING: EXAMPLE APPLICATIONS



Microbiology: find groups of related genes / proteins

Search & Information Retrieval: grouping similar search (or news) results

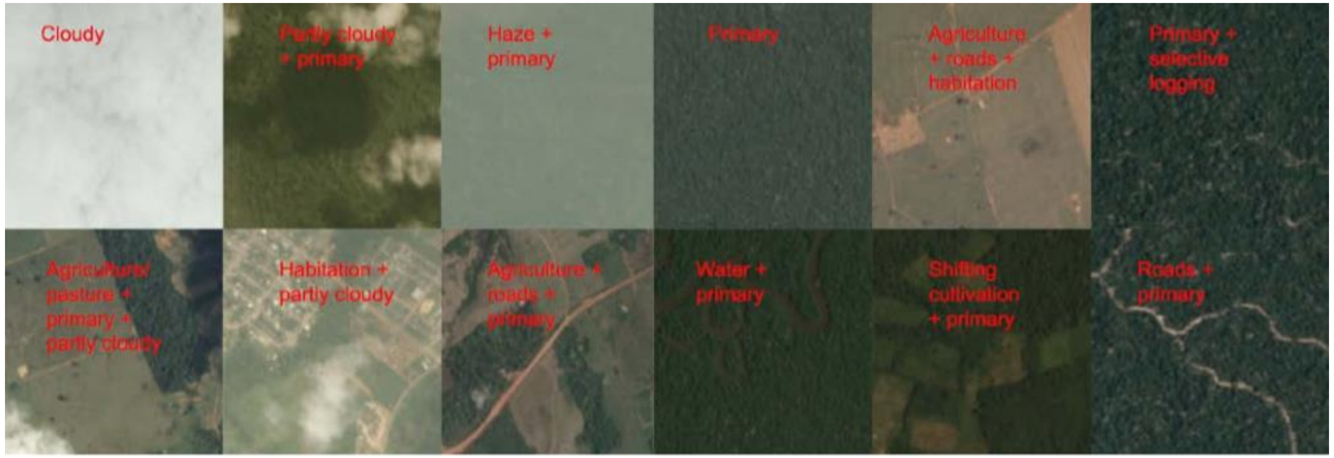
CLASSIFICATION



Goal: Assign data points into categories based on labelled data

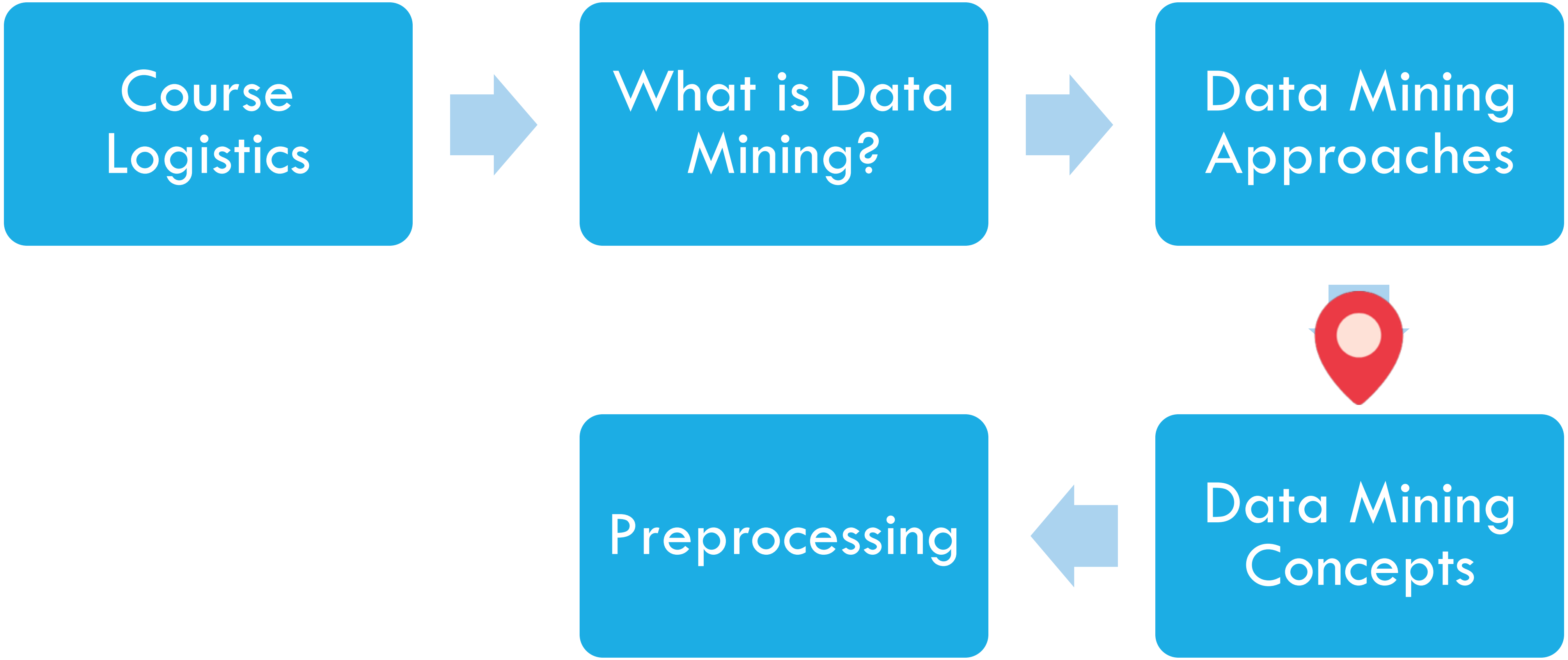
CLASSIFICATION: EXAMPLE APPLICATIONS

Email Spam Detection



Terrain classification: label satellite images by land coverage / use

OUTLINE


Course Logistics
What is Data Mining?
Data Mining Approaches
Data Mining Concepts
Preprocessing

DATASETS: DEFINITIONS

Attributes / Features are properties of each object

Objects / Records

UserID	Country	Height (m)	…
1	SG	1.61	…
2	US	1.50	…
3	MY	1.91	…
…	…	…	…

Attributes / feature values are the numbers or symbols assigned to an attribute for a particular object

TYPES OF ATTRIBUTES

Continuous attributes have real numbers as attribute values

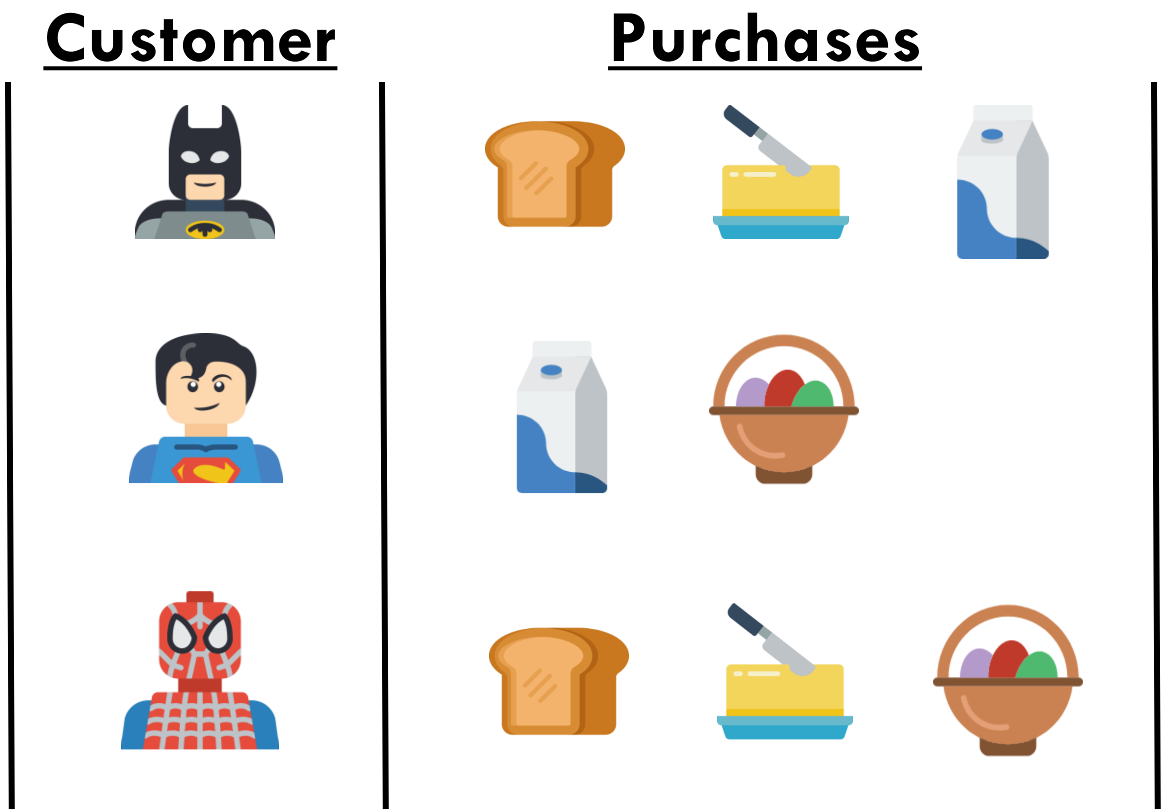
- E.g. height, weight
- Represented in practice using floating point variables

Discrete attributes have categories or integers as attribute values

- E.g. zip codes, counts, words
- Binary attributes are a special case of discrete attributes

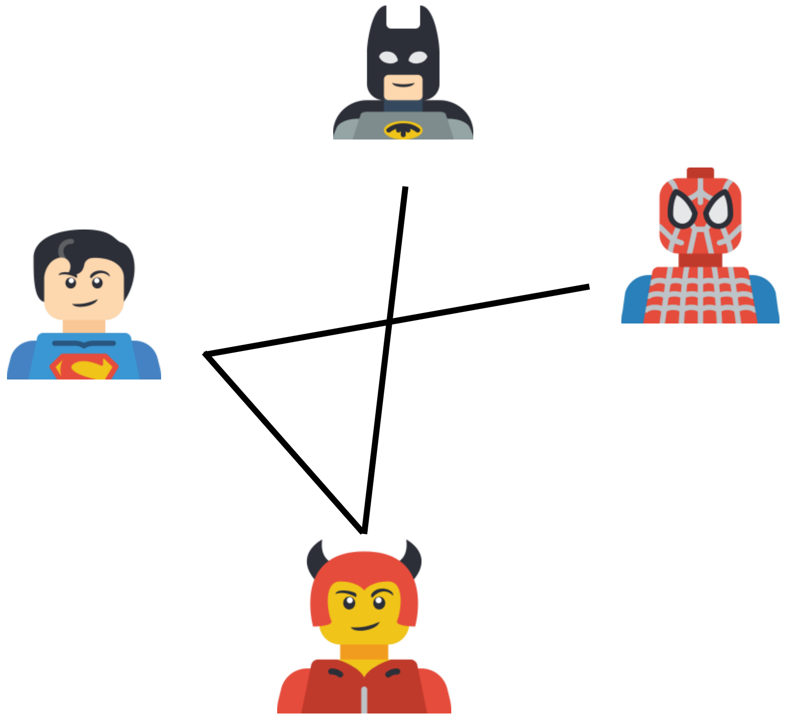
UserID	Height (m)	Country	…
1	1.61	SG	…
2	1.50	US	…
3	1.91	MY	…
…	…	…	…

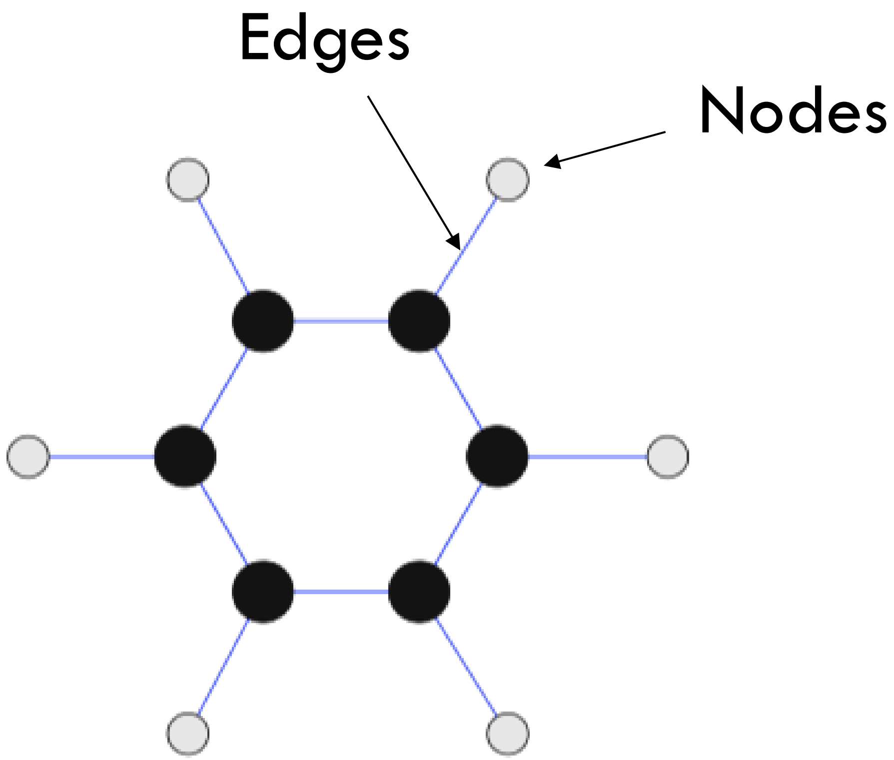
TRANSACTION DATA



Each record (transaction) is a set of items; e.g. products purchased by a customer during a single shopping trip

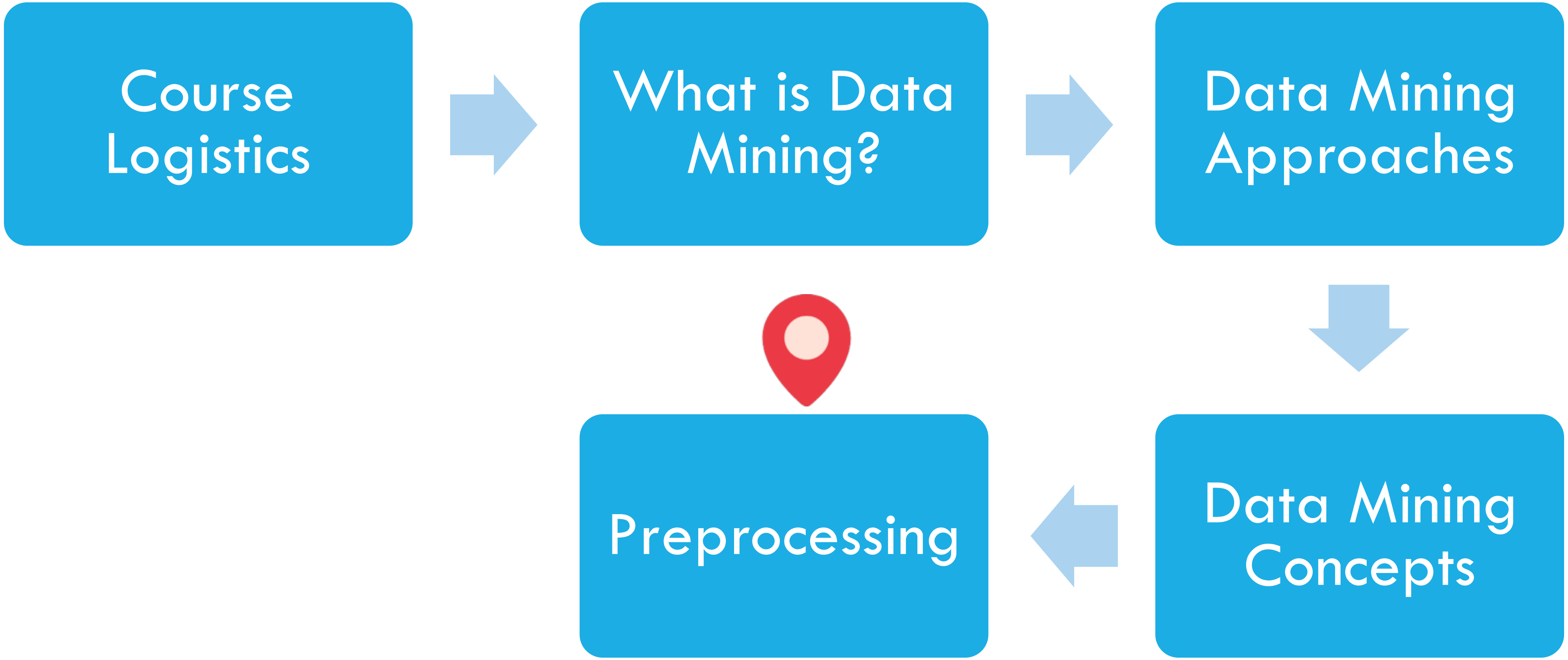
GRAPH DATA





Graph data consists of objects (nodes) connected by a set of links (edges); e.g. nodes can represent webpages, social network users, proteins etc., and edges represent relationships of any kind; e.g. friendships

OUTLINE


Course Logistics
What is Data Mining?
Data Mining Approaches
Data Mining Concepts
Preprocessing

DATA PREPROCESSING

Data Quality ("Cleaning")

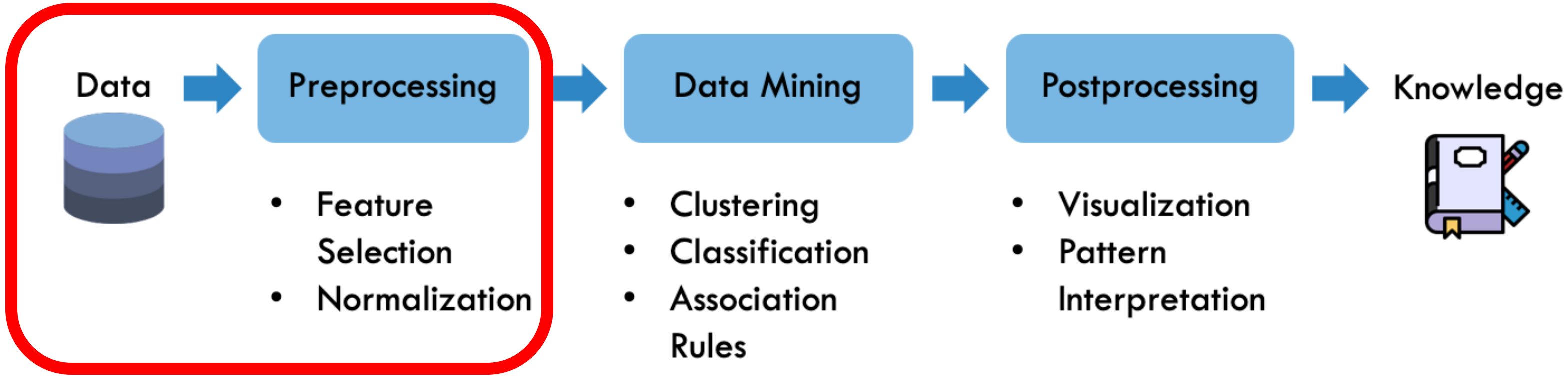
- **Outliers**
- **Missing Values**
- **Duplicates**

Aggregation

Dimensionality Reduction

Feature Creation

Discretization and Binarization



DATA QUALITY

The most important point is that poor data quality is an unfolding disaster.

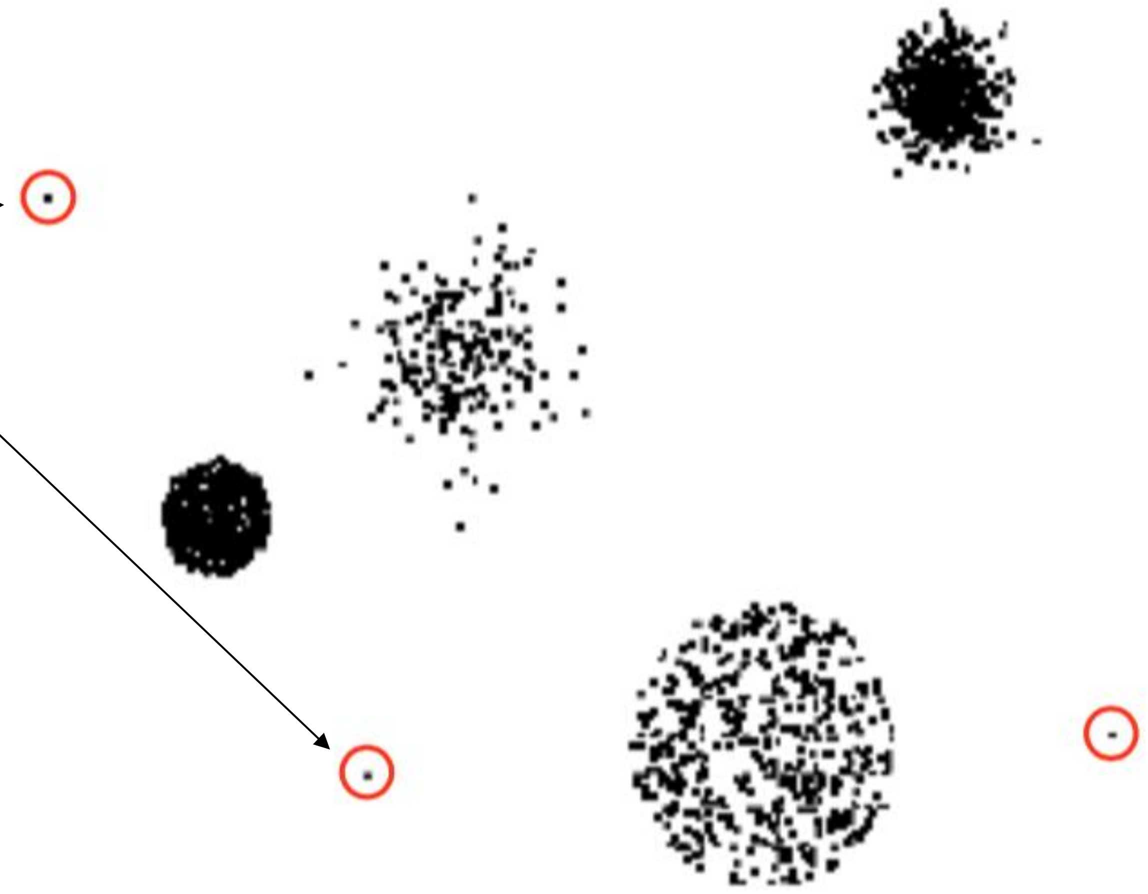
Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.

Thomas C. Redman, DM Review, August 2004

DATA QUALITY: OUTLIERS

Outliers are objects that are considerably different from other objects in the data set

- In some cases, they interfere with data analysis
- In some cases, they are the goal of our analysis: e.g. credit card fraud, network intrusions
- Before eliminating them, it is best to inspect the data to understand why they occur



DATA QUALITY: MISSING VALUES

Why is data missing?

- Information was not collected: e.g. people decline to give weight
- Attributes may not be applicable to all cases

How to handle missing values?

- Eliminate objects with missing values
- Or: fill in the missing values ("imputation")
  - E.g. based on the mean / median of that attribute
  - Or: by fitting a regression model to predict that attribute given other attributes

UserID	Height (m)	Country	...
1	1.61	SG	...
2	1.50	US	...
3	***NA***	MY	...
...	...	...	...

Median Imputation

UserID	Height (m)	Country	...
1	1.61	SG	...
2	1.50	US	...
3	***1.55***	MY	...
...	...	...	...

DATA QUALITY: DUPLICATES

Objects appear multiple times in the dataset, e.g. due to merging data from different sources

- E.g. same person with multiple email addresses

UserID	Height (m)	Country	…
1	1.61	SG	…
2	1.50	US	…
2	1.50	US	…
…	…	…	…

Deduplication

UserID	Height (m)	Country	…
1	1.61	SG	…
2	1.50	US	…
…	…	…	…

AGGREGATION

Combining attribute values:

- E.g. aggregating days into weeks
- Or: aggregating cities into countries

This can help in reducing the number of attribute values

It also makes the data more "stable", e.g. week-frequency data is less variable and easier to predict

UserID	Day	Country	…
1	1	SG	…
2	3	US	…
3	9	MY	…
…	…	…	…

Aggregation

UserID	Week	Country	…
1	1	SG	…
2	1	US	…
3	2	MY	…
…	…	…	…

DIMENSIONALITY REDUCTION

This approximates high-dimensional data using a smaller number of dimensions

Why?

- Efficiency
- Remove irrelevant features
- Can help avoid **"curse of dimensionality"** (next slide)

UserID	Day	Country	...
1	1	SG	...
2	3	US	...
3	9	MY	...
...	...	...	...

Dimensionality Reduction

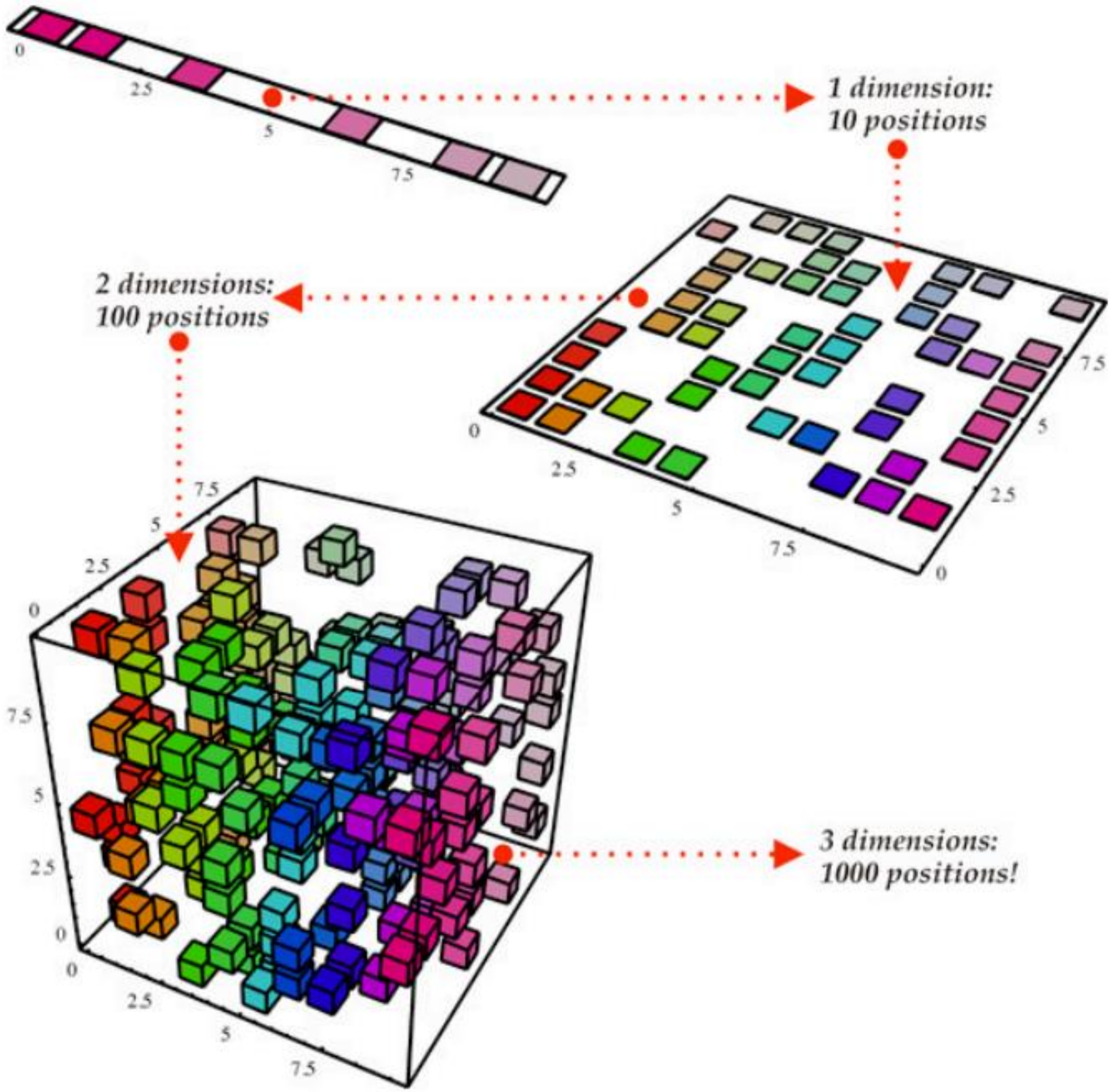
Var1	Var2
1	1
2	1
3	2
...	...

CURSE OF DIMENSIONALITY

As the number of dimensions increases, the amount of space the data has to cover grows exponentially

Hence, the space becomes sparser and sparser

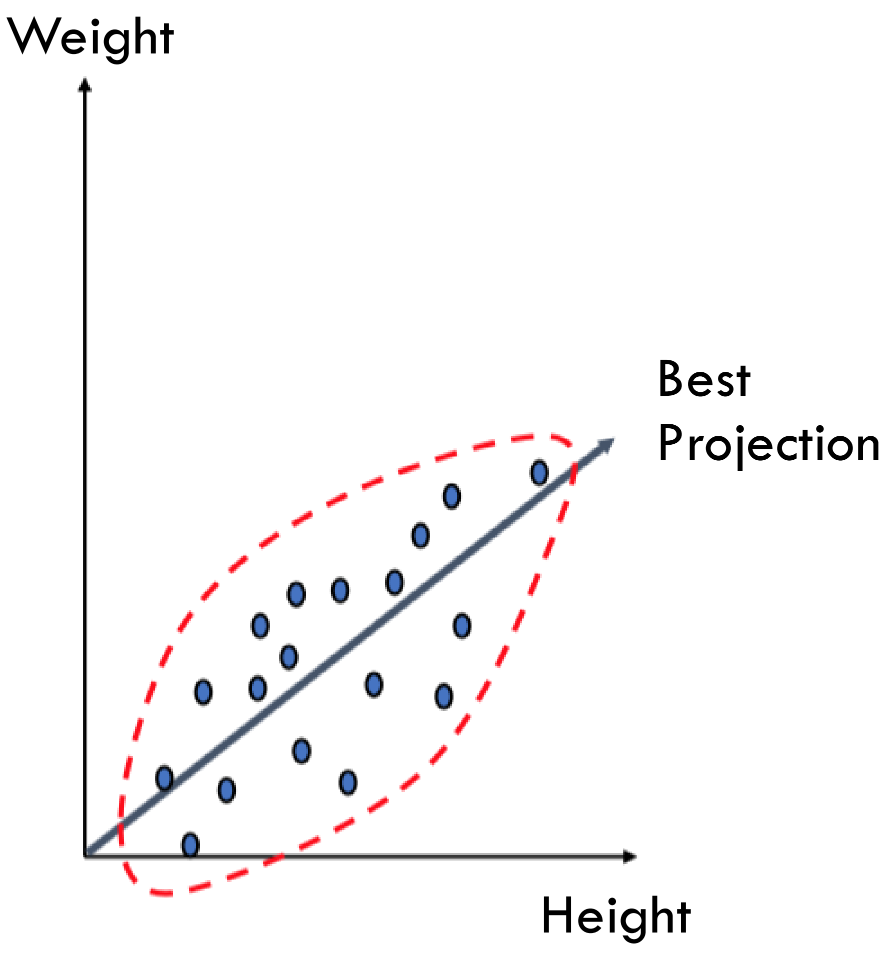
Many algorithms fail to make effective predictions as there are no nearby points to use to predict a given test point



https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

PRINCIPAL COMPONENT ANALYSIS

PCA reduces dimensionality by finding the best projection that captures the largest amount of variation in the data



https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

FEATURE CREATION

Create new features that capture the important information in data better than the original features

2 general approaches:

- Feature extraction
  - E.g. extracting edges from images
- Feature transformation
  - E.g. dividing mass by volume to get density
  - Or: apply simple functions to a feature, e.g. log(x), |x|

https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

DISCRETIZATION

Convert continuous features into discrete features by mapping them into "buckets":

- Ex: round the "height" variable to nearest cm
- Many algorithms (e.g. regression) are more flexible when given discrete input

Var1	Height
1	163.4cm
2	164.5cm
3	193.4cm
…	…

Var1	Var2
1	163cm
2	165cm
3	193cm
…	…

https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

ONE-HOT ENCODING

Convert discrete feature to a series of binary features.

E.g. the first record has group 2, so we set its 2nd binary feature to 1, and all the rest to 0.

This lets us use any method for numerical features (e.g. regression) on discrete features.

Group
2
1
3
…

Group1	Group2	Group3
0	1	0
1	0	0
0	0	1
…	…	…

https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376