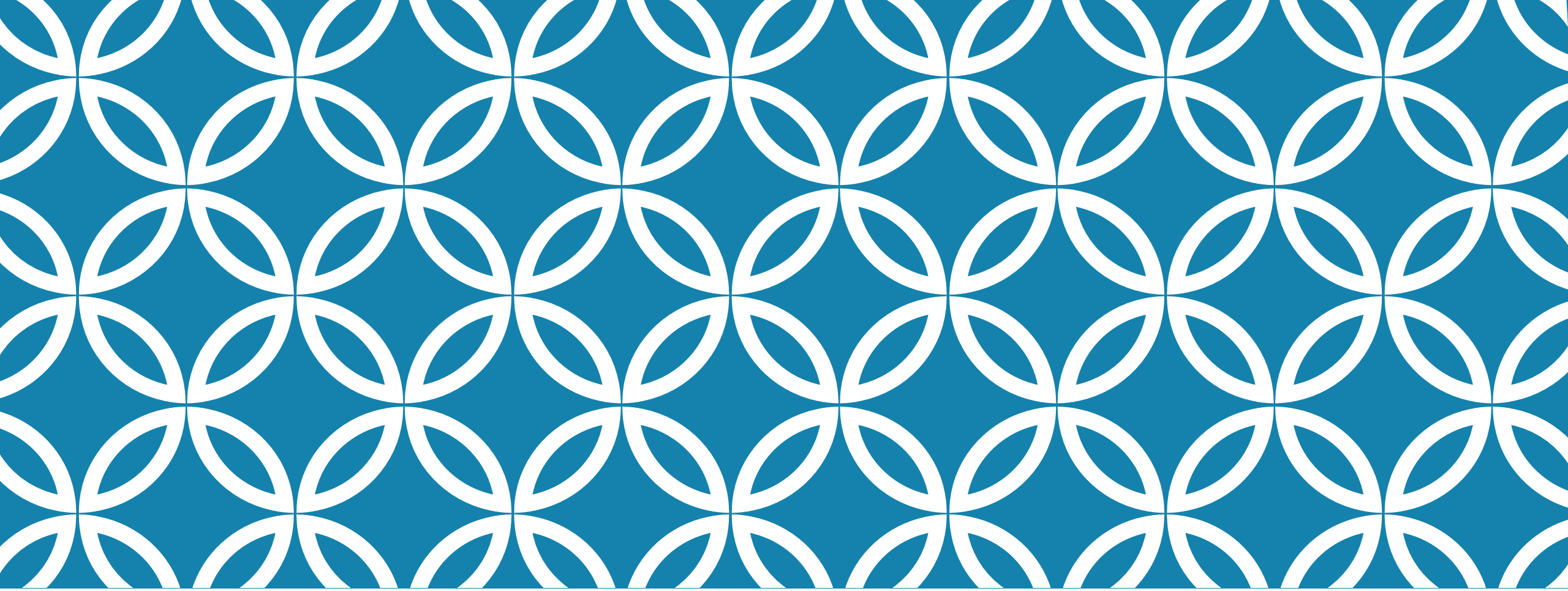


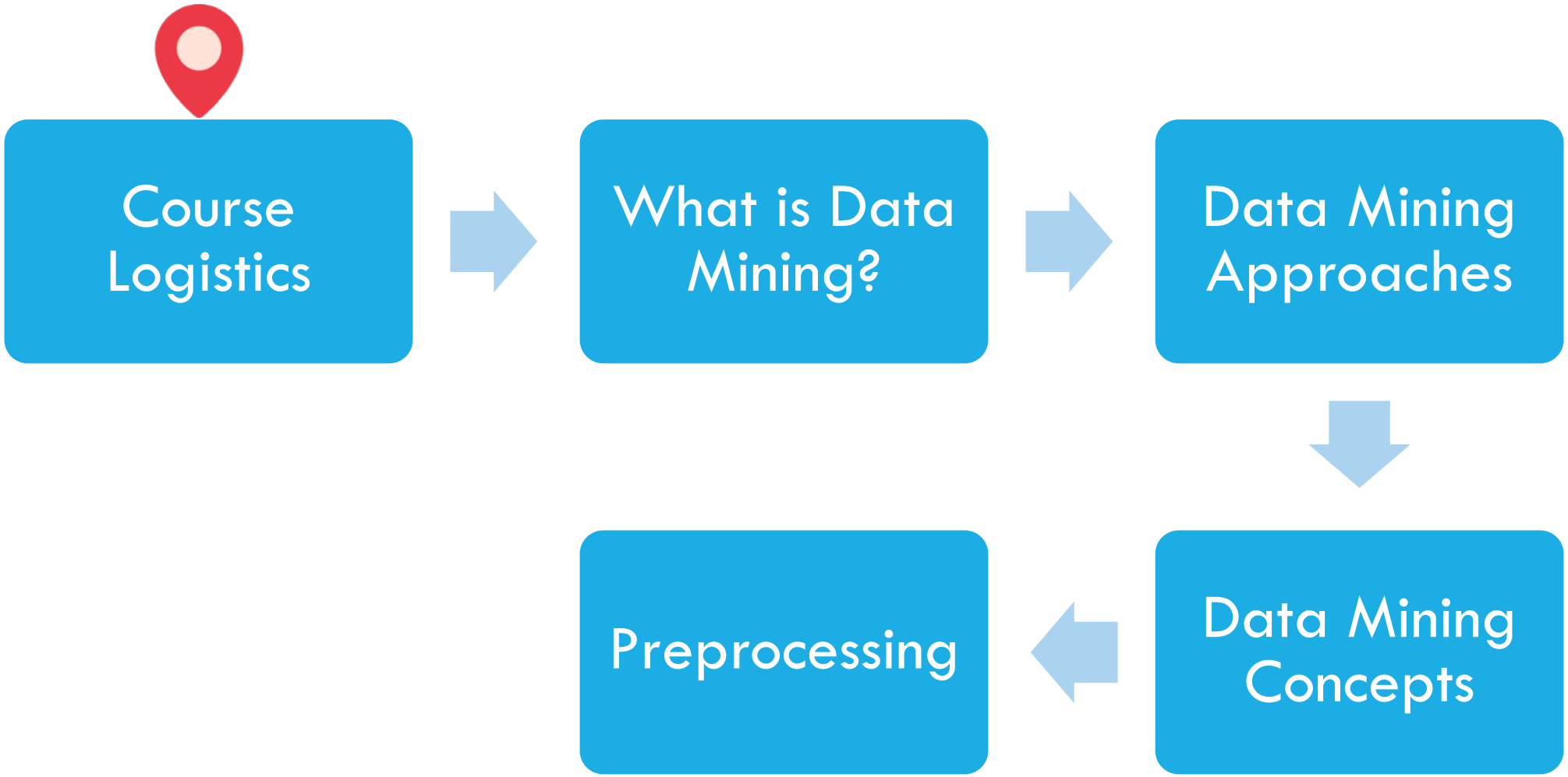
CS5228 LECTURE 1: INTRODUCTION

Bryan Hooi

School of Computing

National University of Singapore

OUTLINE


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

COURSE INFORMATION

- **Lectures & Tutorials**
  - Lectures: Friday 6.30pm – 8.30pm; Tutorials: Fridays 8.30pm – 9.30pm (not every week; see schedule)
  - Physical and Zoom classes (all recorded)
  - Zoom: can access via Canvas "Zoom" link (both lectures and tutorials)
  - Announcements & materials on Canvas

- **Where to ask questions**
  - Canvas discussion (you are also strongly encouraged to answer questions!)
  - Email to teaching team

COURSE STAFF

Lecturer:

- Bryan Hooi (bhooi@comp.nus.edu.sg)
- Office: COM3-02-22; office hours: 1.30 – 2.30pm or by appointment

TAs:

- He Xiaoxin (he.xiaoxin@u.nus.edu)
- Liu Xu (liuxu12@u.nus.edu)
- Jiang Yangfan (yangfan.jiang@u.nus.edu)
- Nguyen Thong Thanh (e0998147@u.nus.edu)

ASSESSMENT

Assignment 1: worth 25%

Assignment 2: worth 25%

Group Project: worth 50%

ASSIGNMENTS

Assignments will involve conceptual questions, and data analysis (with programming)

Python is the primary programming language

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

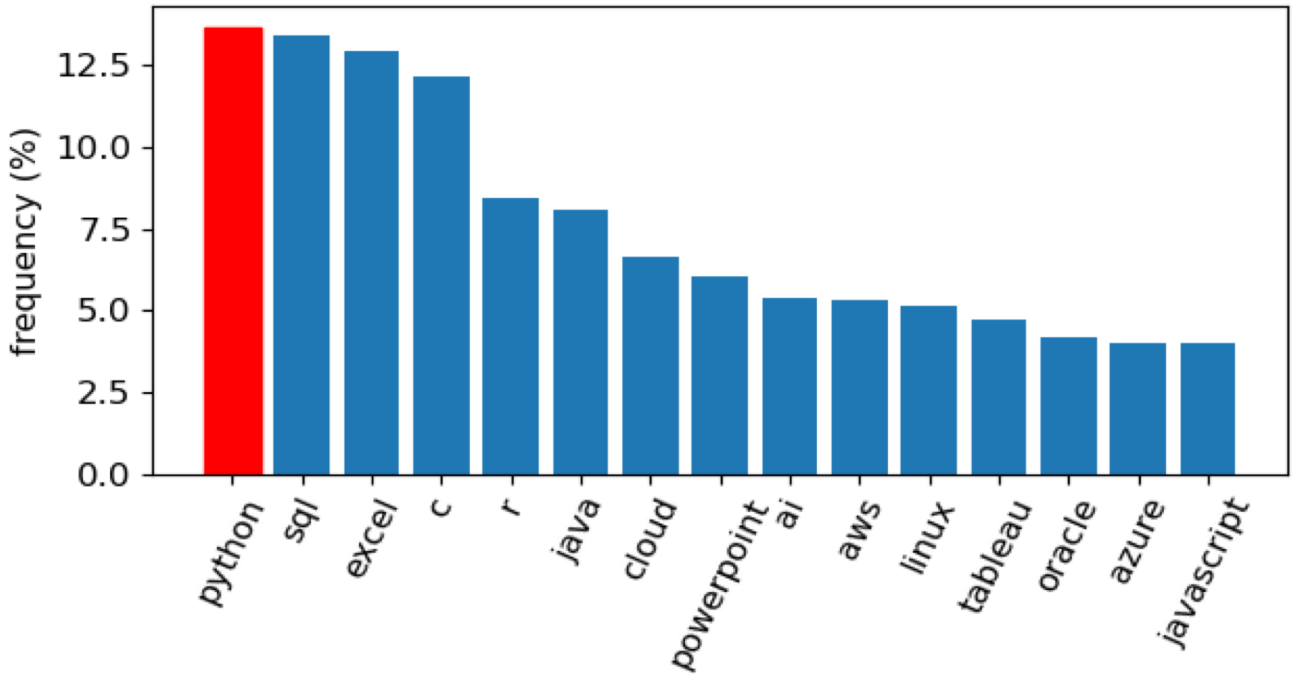
Submission should be via Canvas

4 late days can be used for either assignment

- Each extends the deadline by 24 hours
- No need to send any emails to use it, just submit your assignments late

WHY PYTHON?

- Analysis of job descriptions
  - 15k+ job offers from JobStreet
    (data analyst, data engineer, data scientist)
  - Quick-&-dirty keyword extraction



PROJECT

Group project (3-4 students per group)

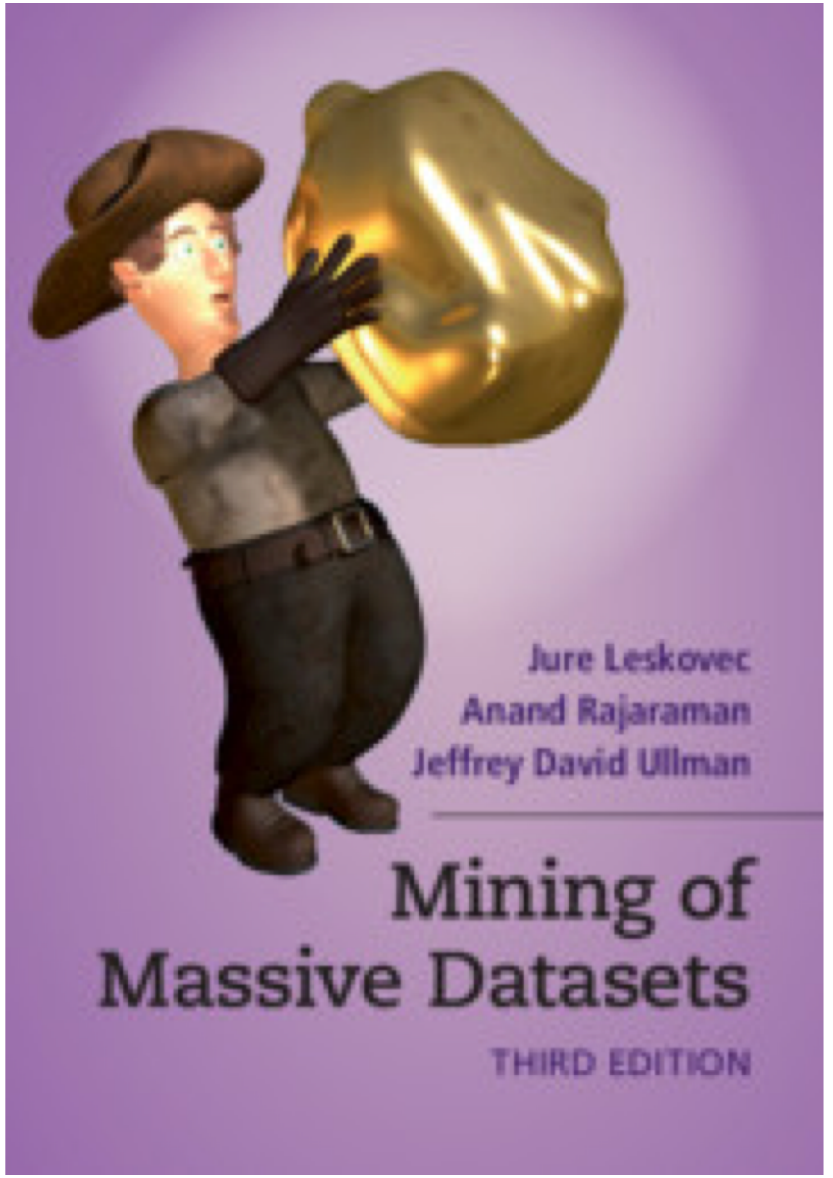
There will be 2 options which you can choose from:

1. Kaggle-in-class competition
2. Self-proposed project

REFERENCE

Textbook (useful but not required):

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



SCHEDULE

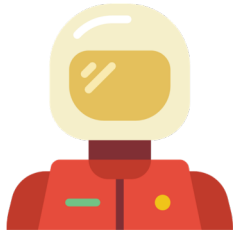
Week	Date	Topics	Tutorials	Important Dates
1	13 Jan	Introduction		
2	20 Jan	No class (public holiday)		
3	27 Jan	Clustering I	Tutorial 1	
4	3 Feb	Clustering II		Release A1
5	10 Feb	Association Rules	Tutorial 2	
6	17 Feb	Regression & Classification I		
Recess		No class		
7	3 Mar	Regression & Classification II	Tutorial 3	A1 due (Sunday 11.59pm), release A2
8	10 Mar	Regression & Classification III		
9	17 Mar	Recommender Systems	Tutorial 4	
10	24 Mar	Graph Mining		
11	31 Mar	Data Stream Mining	Tutorial 5	A2 due (Sunday 11.59pm)
12	7 Apr	No class (public holiday)		
13	14 Apr	Review & Outlook		Project due (Sunday 11.59pm)

COURSE OBJECTIVES

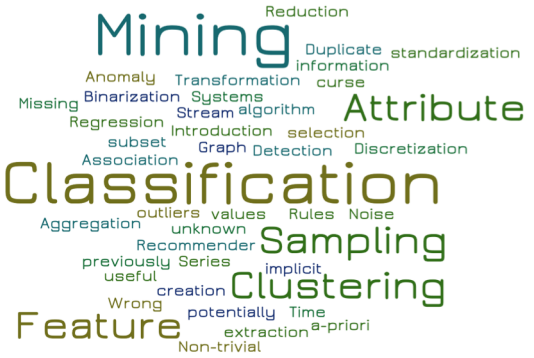
By the end of the course, you should expect to:

- 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

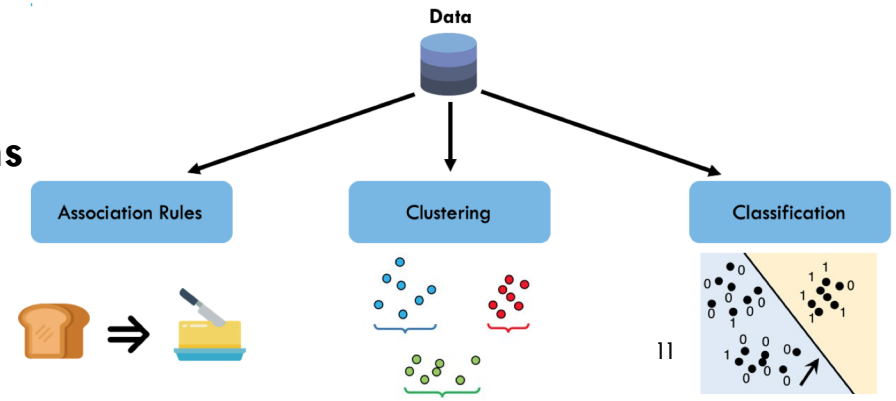
UserID	Height (m)	Country	…
1	1.61	SG	…
2	1.50	US	…
3	***NA***	MY	…
…	…	…	…



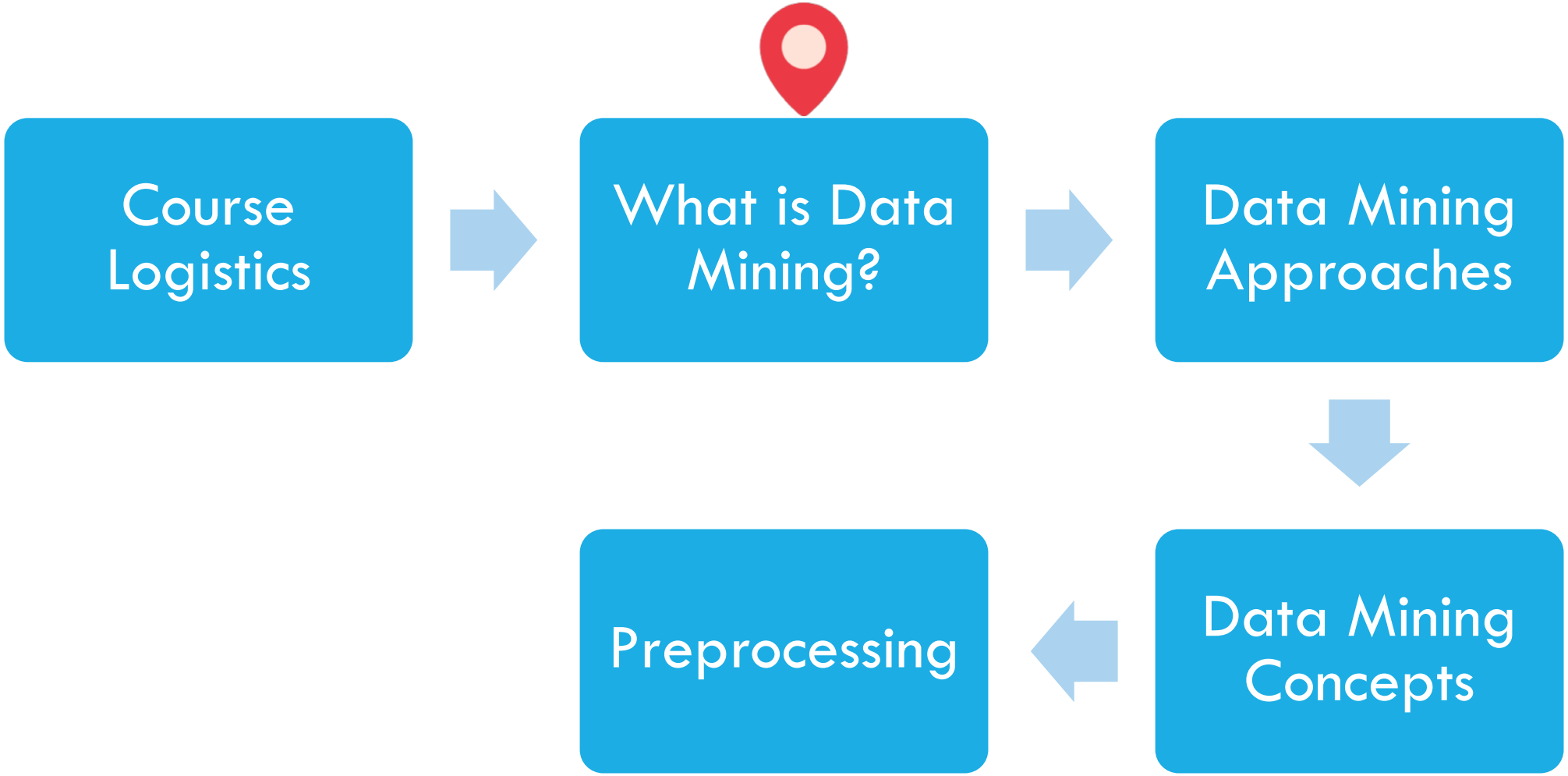
Concepts



Algorithms



OUTLINE


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

WHAT IS DATA MINING?

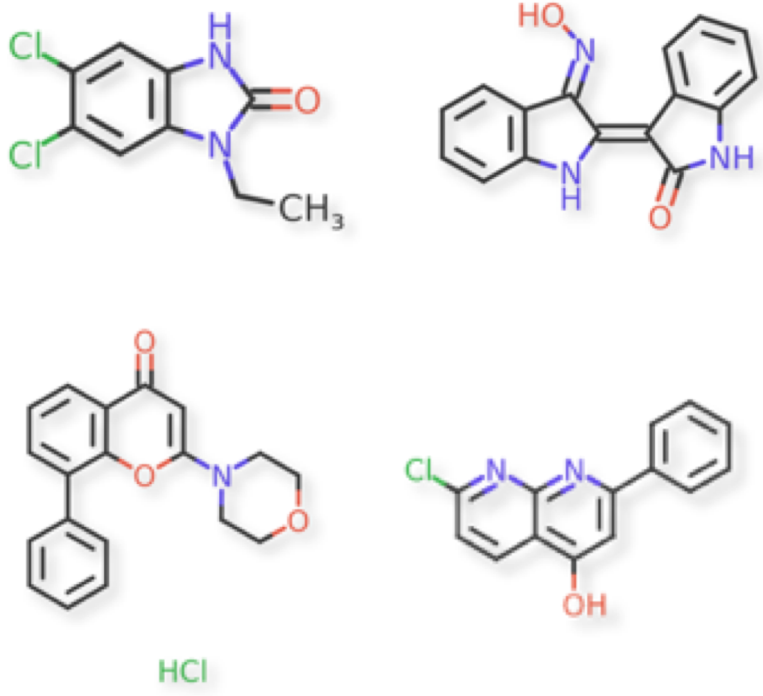
Not just lookup (looking up a phone number from a phone book)

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

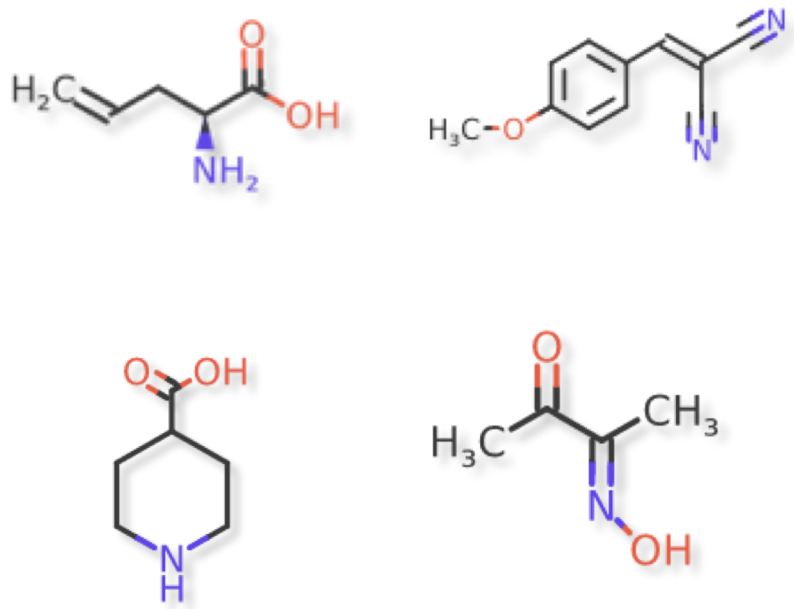
William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus

Q: WHAT RULE CHARACTERIZES TOXIC MOLECULES?

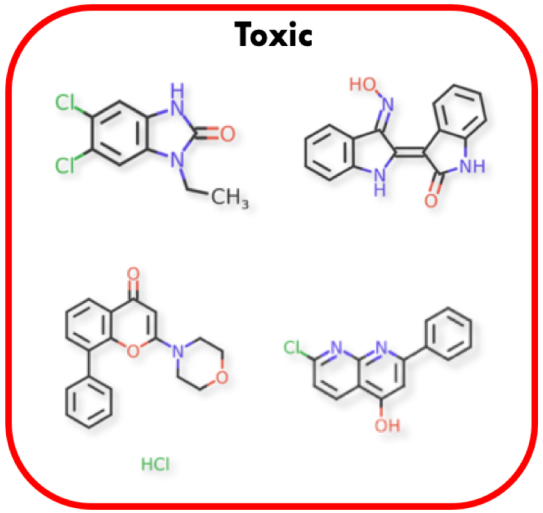
Toxic

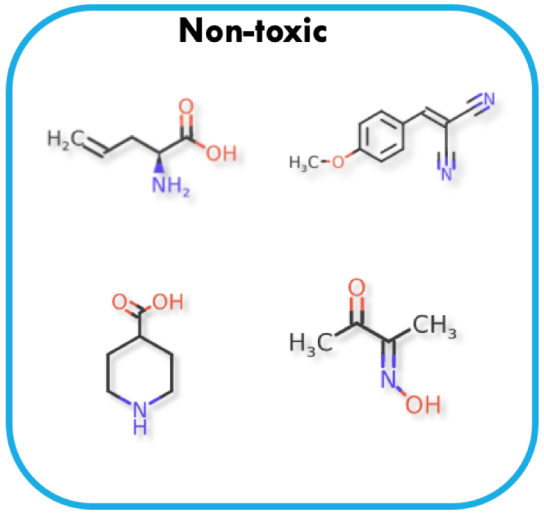


Non-toxic



Q: WHAT RULE CHARACTERIZES TOXIC MOLECULES?





if  > 1:

    toxic

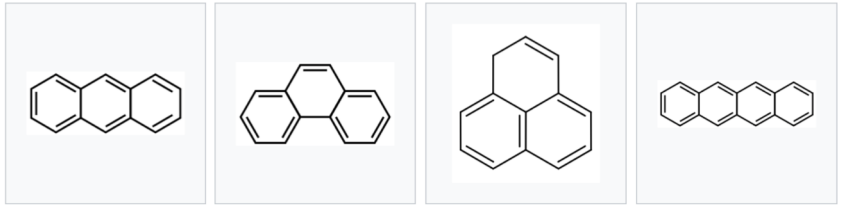
else

    nontoxic

Polycyclic aromatic hydrocarbon

From Wikipedia, the free encyclopedia

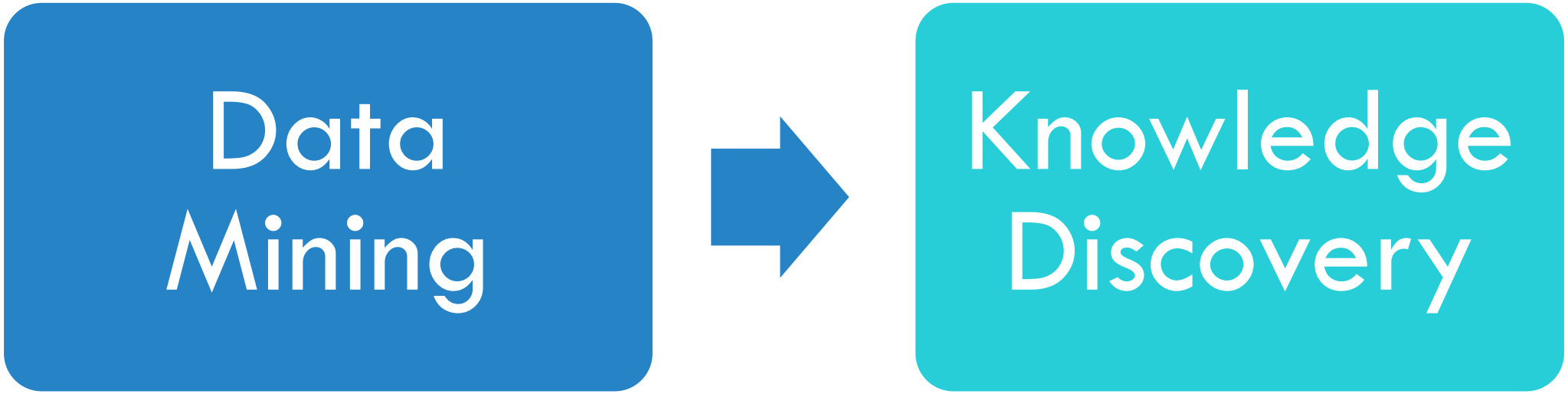
Principal PAH Compounds



Cancer [edit]

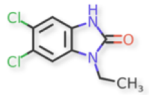
PAHs have been linked to skin, lung, bladder, liver, and stomach cancers in well-established animal model studies.[72] human carcinogens are identified in the section "Regulation and Oversight" below.

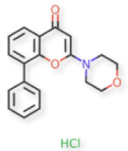
WHAT IS DATA MINING AND KNOWLEDGE DISCOVERY?



(Approach) **(Goal)**

THE DATA MINING PROCESS





(2, 0, 0, 1, …)

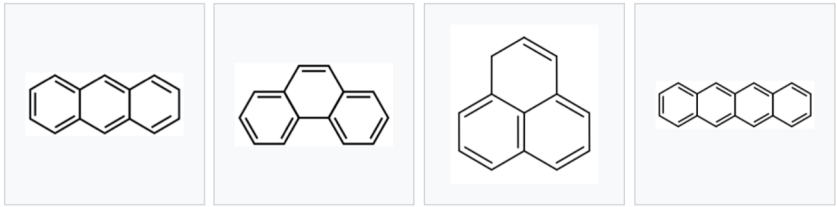
(4, 1, 0, 1, …)

if  > 1:
    toxic
else
    nontoxic

Polycyclic aromatic hydrocarbon

From Wikipedia, the free encyclopedia

Principal PAH Compounds

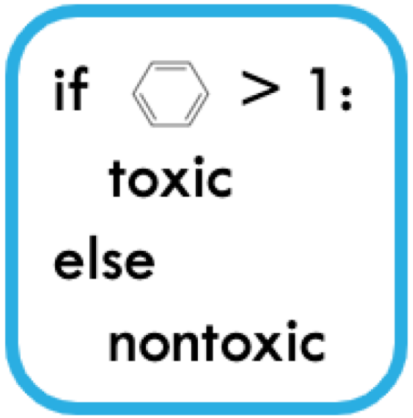


Data → Preprocessing → Data Mining → Postprocessing → Knowledge

- Feature Selection
- Normalization

- Clustering
- Classification
- Association Rules

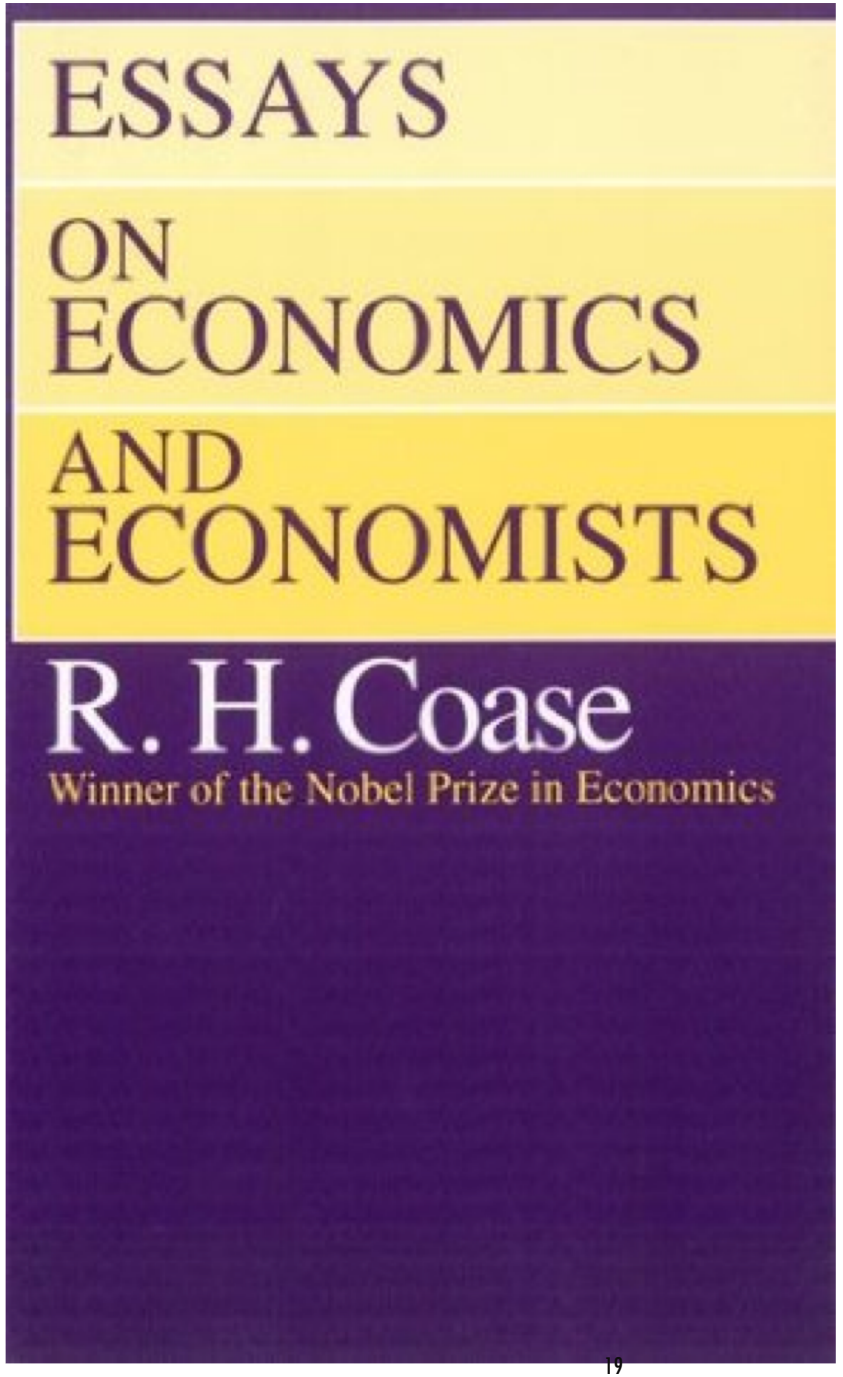
- Visualization
- Pattern Interpretation



HOW DO WE KNOW IF OUR PATTERNS ARE ACTUALLY TRUE?

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

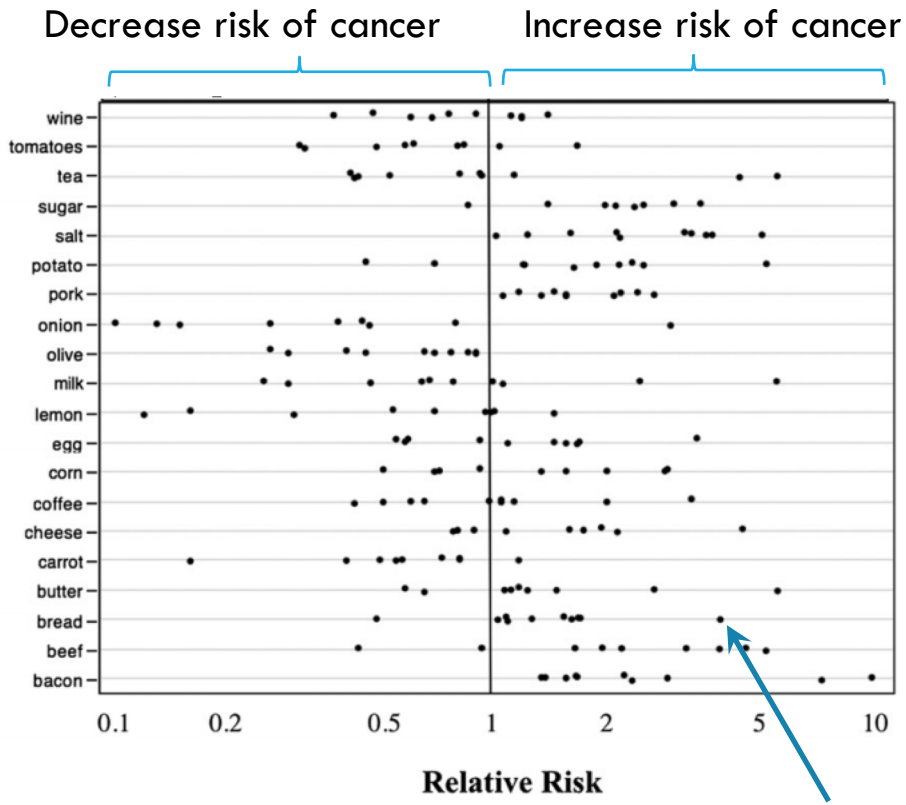
John P. A. Ioannidis

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

Ioannidis, John PA. "Why most published research findings are false." *PLoS medicine* 2.8 (2005): e124.

Optional

ARE MOST PUBLISHED RESEARCH FINDINGS FALSE?



Each point represents 1 study

Reasons:

- Publication bias: huge number of studies with low sample size, and only positive results are published
- Fraud / conflict of interest
- Flexibility in analysis

Conclusion: misuse of statistics leads to false or spurious patterns. An important part of data mining is to ensure that the patterns we discover are real and meaningful.

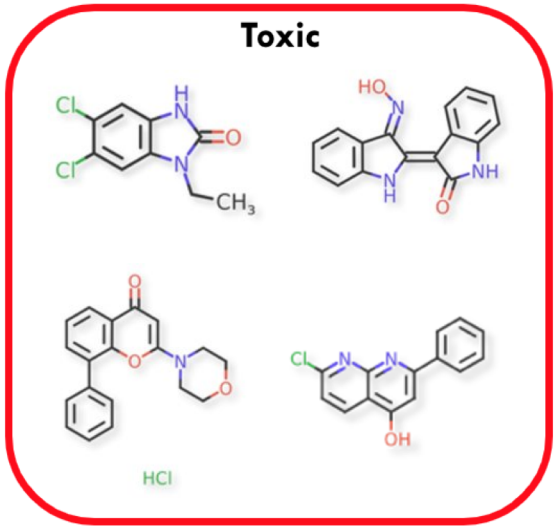
[1] Schoenfeld, Jonathan D. and John P. A. Ioannidis (2013). "Is everything we eat associated with cancer? A systematic cookbook review". In: The American journal of clinical nutrition 97.1, pp. 127–134.

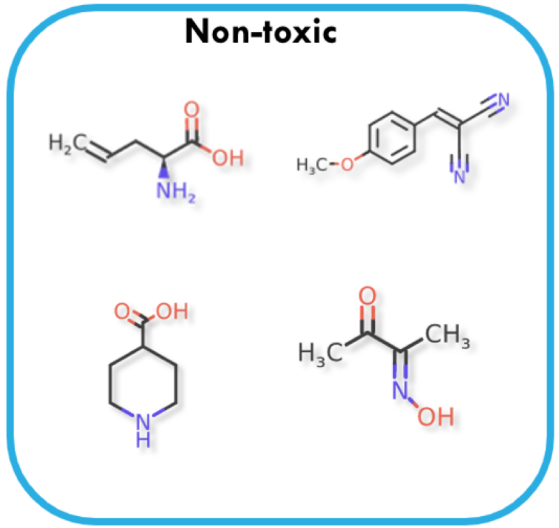
HOW DO WE KNOW IF OUR PATTERNS ARE MEANINGFUL?

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

If the data we are mining from is too small or biased, this can lead to lack of generalizability.

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



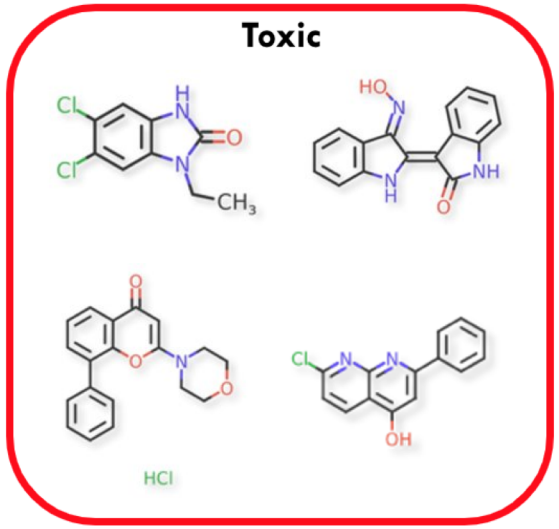


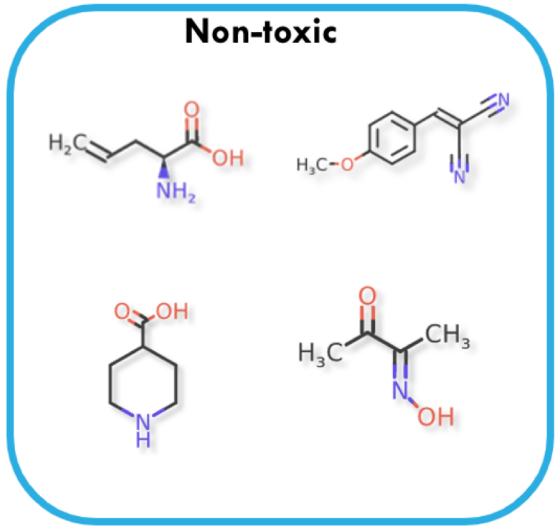
Optional

HOW DO WE KNOW IF OUR PATTERNS ARE MEANINGFUL?

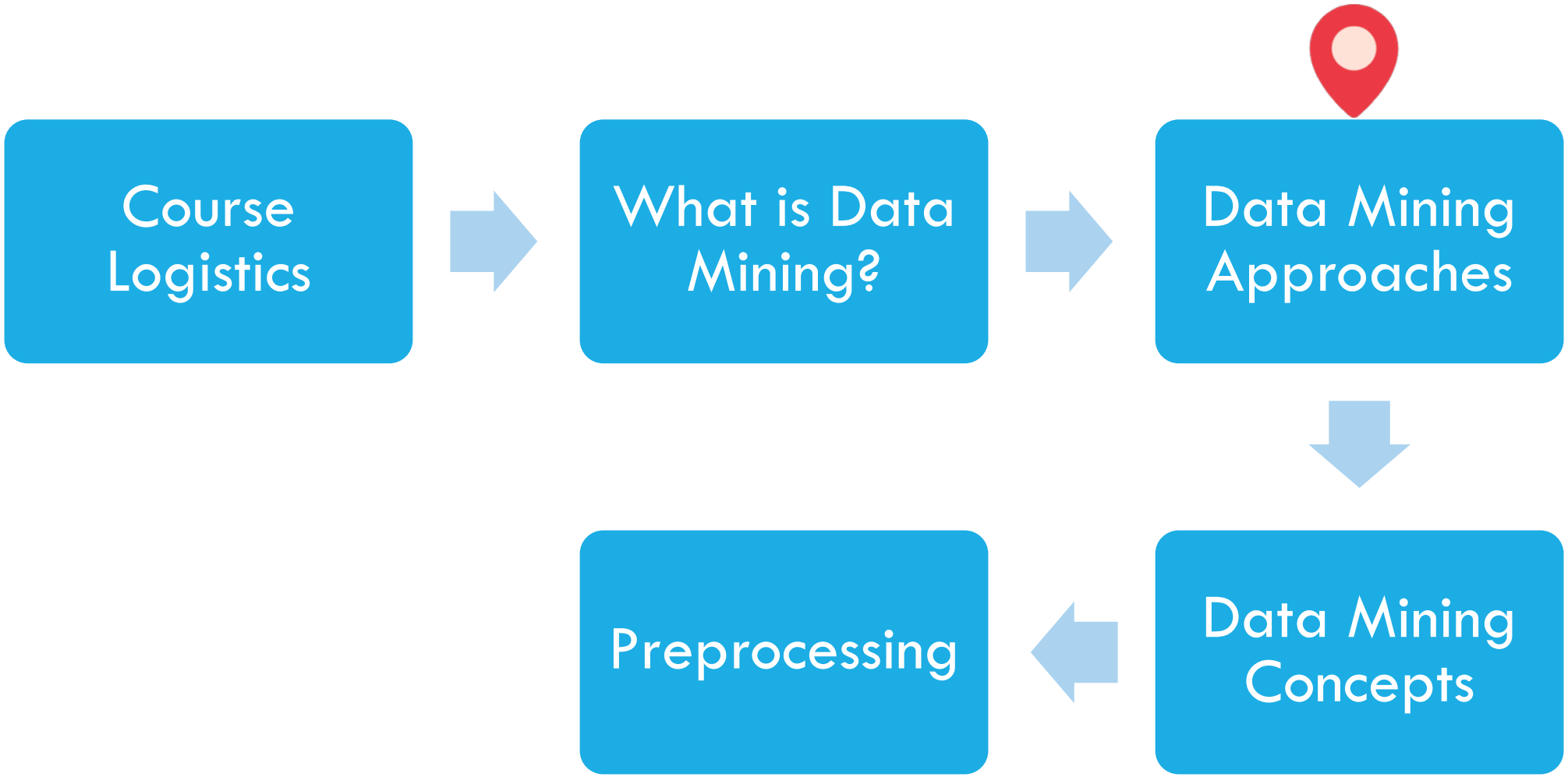
Bonferroni's Principle (roughly): if you look for more patterns than your dataset can support, you are bound to find ***false positives*** (patterns that are not actually present)

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

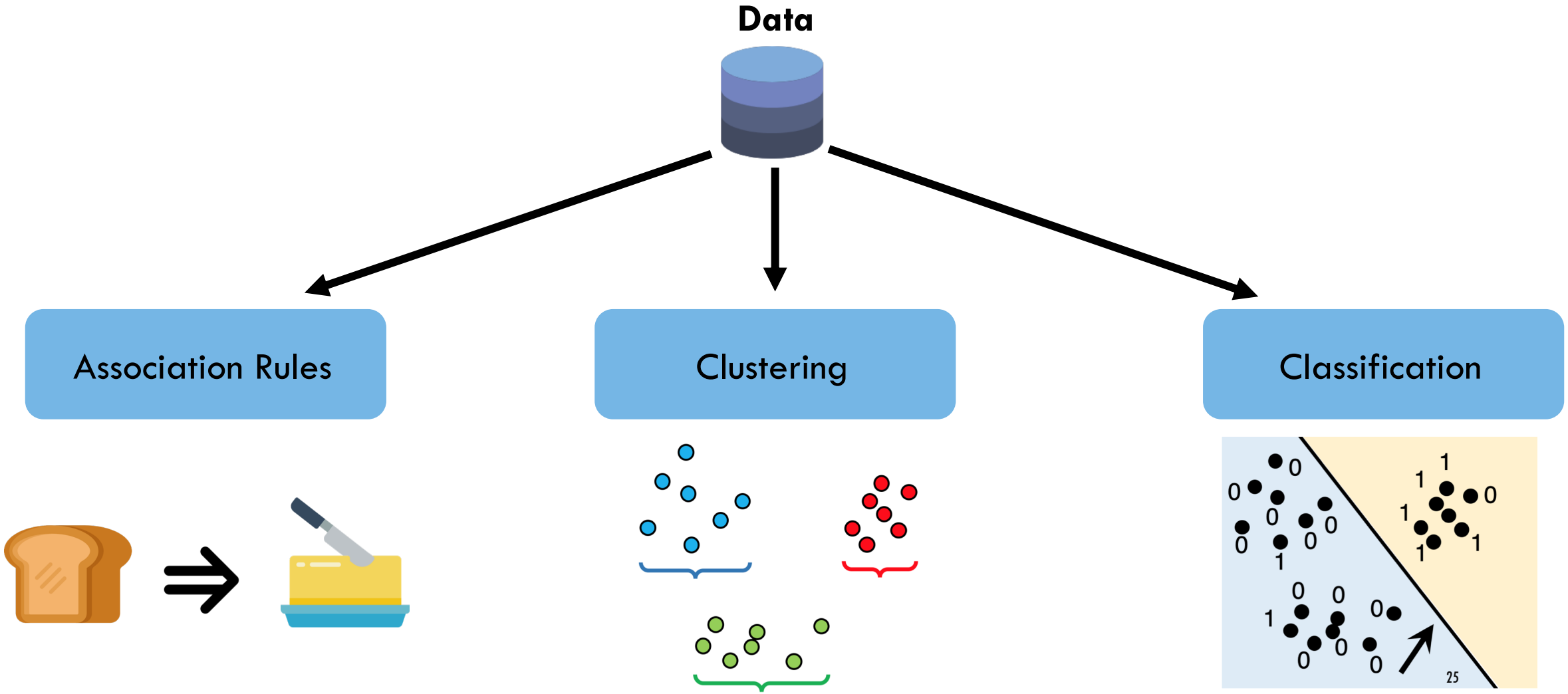




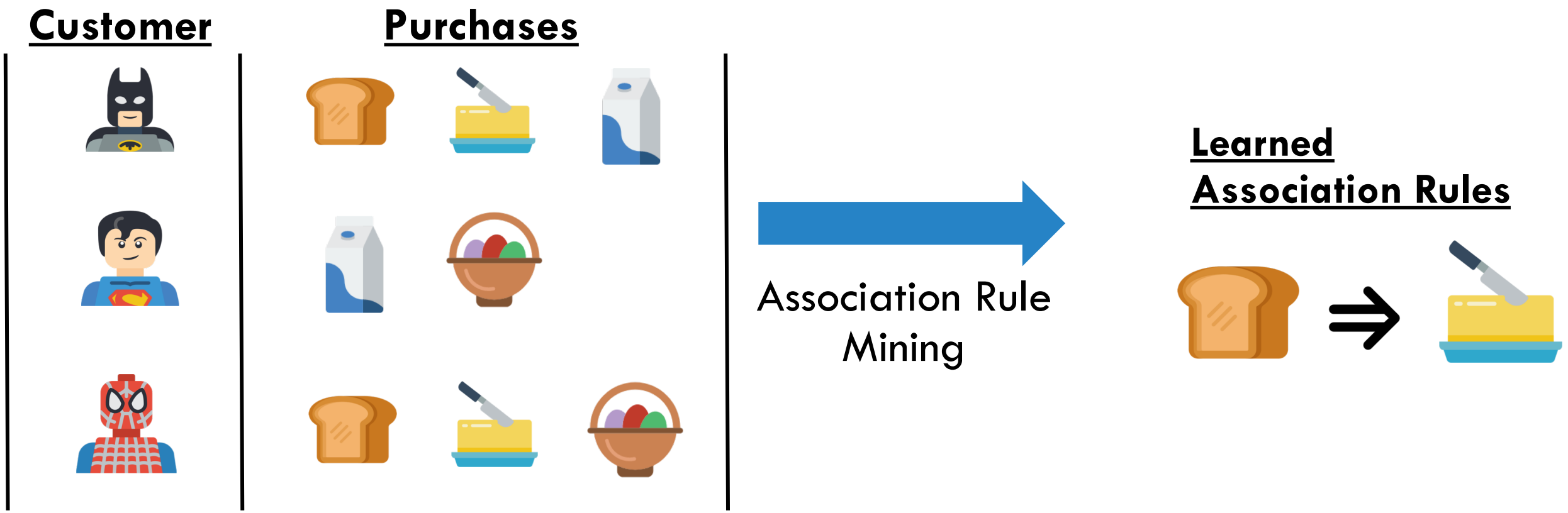
OUTLINE


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

DATA MINING APPROACHES



ASSOCIATION RULE MINING



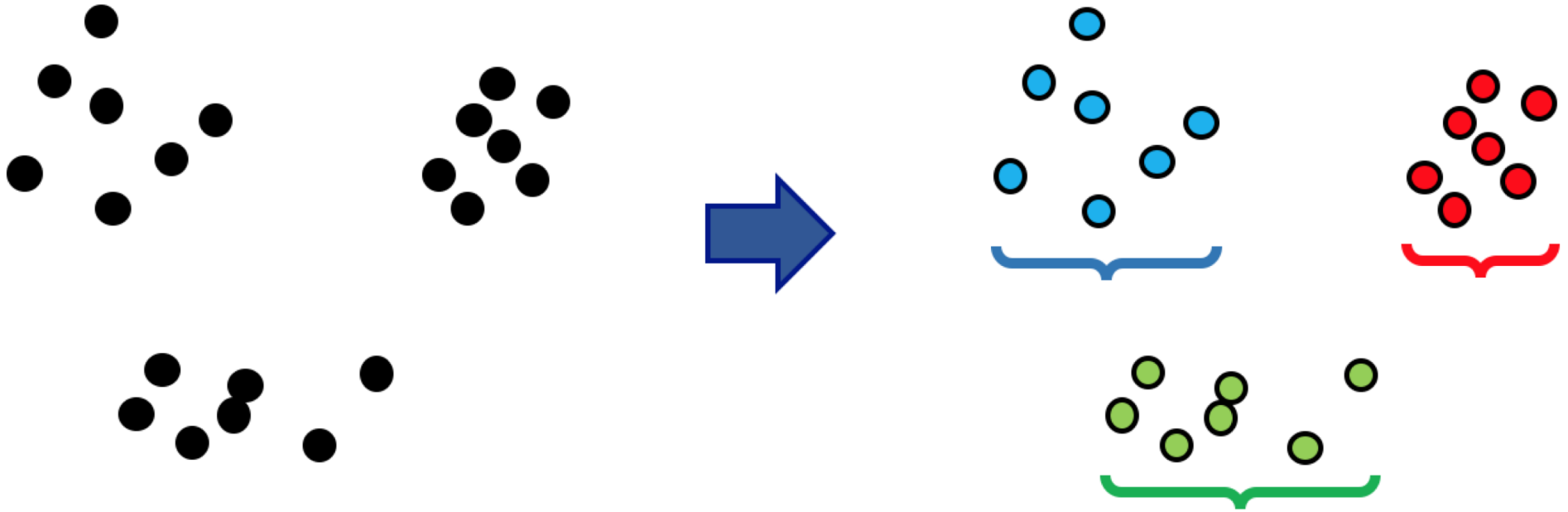
Goal: Given multiple *transactions* (e.g. *sets* of items bought by customers), find **rules** that predict occurrence of an one item based on occurrences of others

ASSOCIATION RULES: EXAMPLE APPLICATIONS

Market basket analysis: e.g. for inventory management

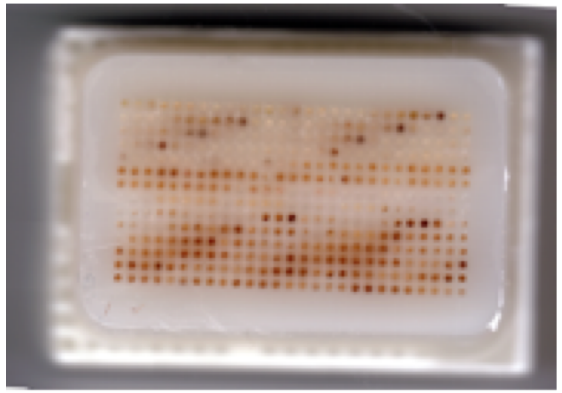
Medical: finding rules relating patient symptoms, test results and diseases

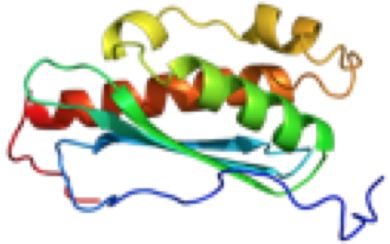
CLUSTERING



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

CLUSTERING: EXAMPLE APPLICATIONS

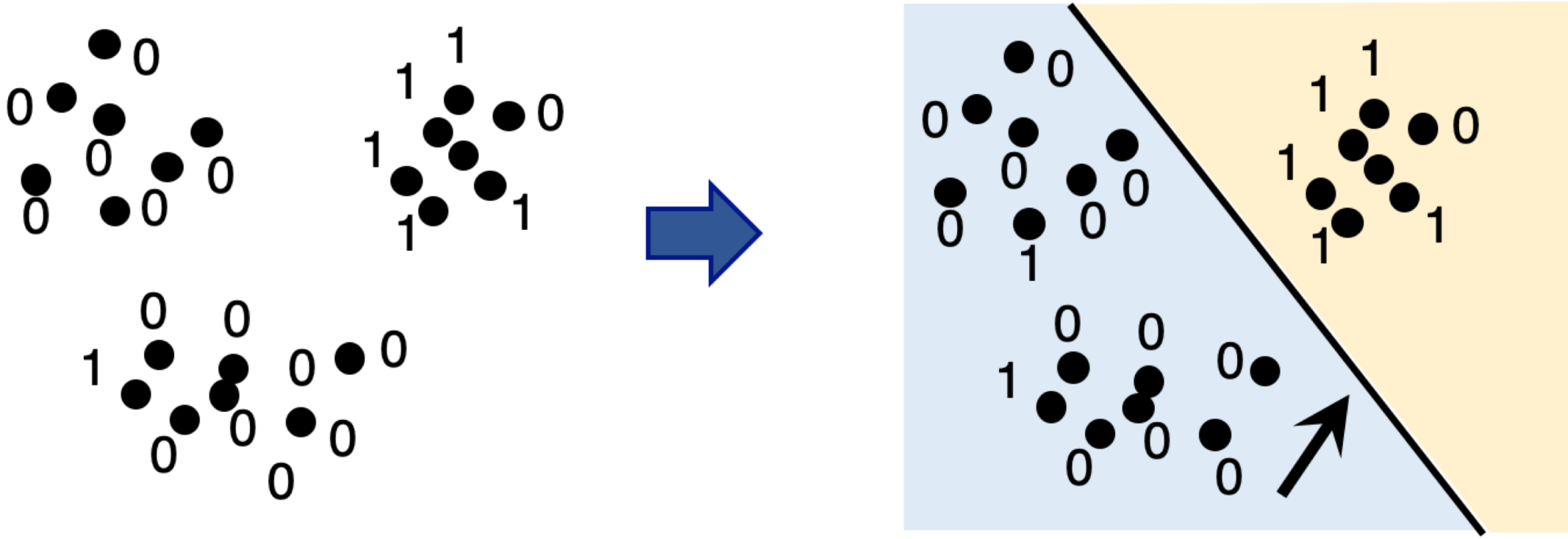




Microbiology: find groups of related genes / proteins

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

CLASSIFICATION & REGRESSION

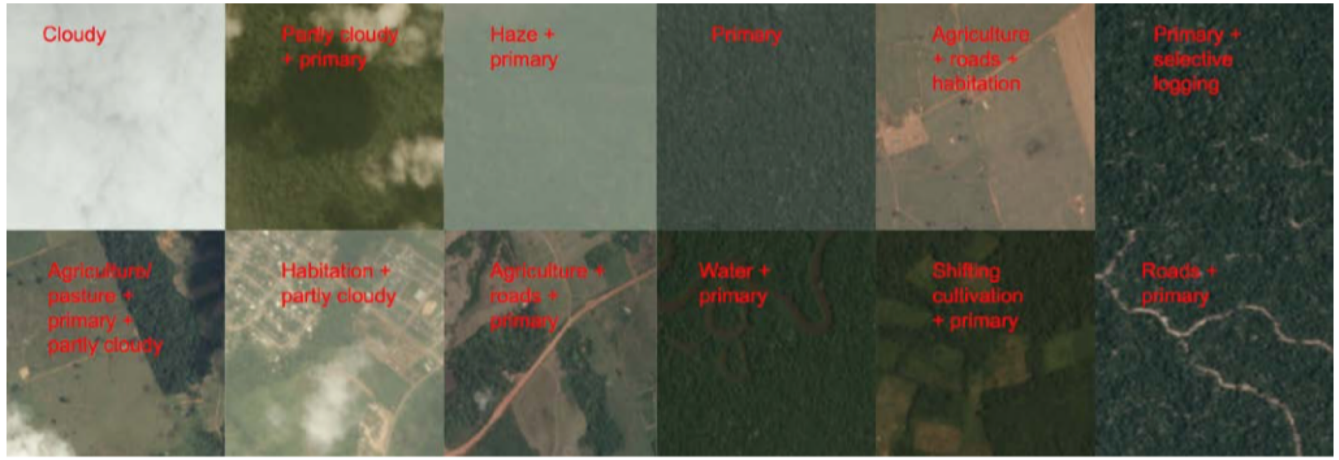


Classification: Given some labelled training data, learn a decision rule to predict a category as a function of other attributes

- If we are instead predicting a numerical attribute, the problem is called Regression

CLASSIFICATION: EXAMPLE APPLICATIONS

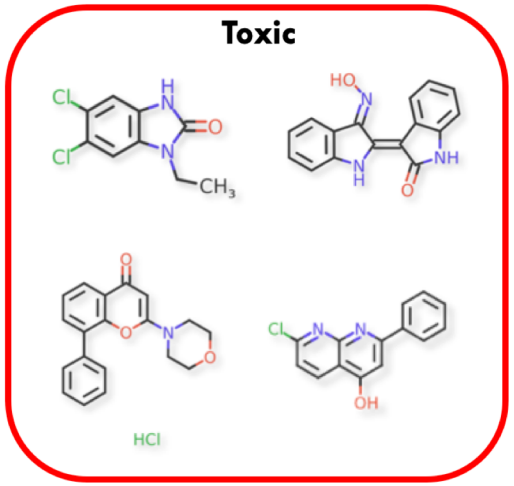
Email Spam Detection

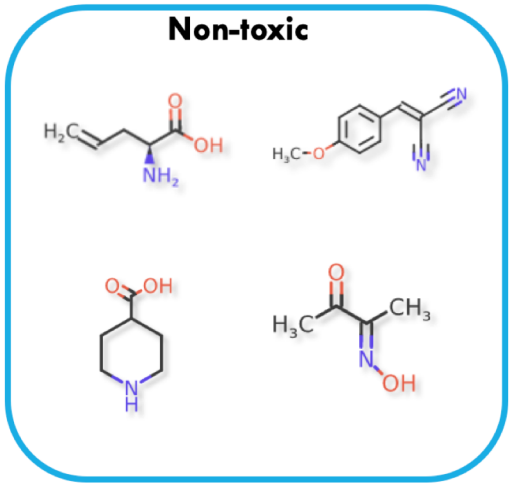


Terrain classification: label satellite images by land coverage / use

Q: THE MOLECULE TOXICITY TASK WAS AN EXAMPLE OF…?

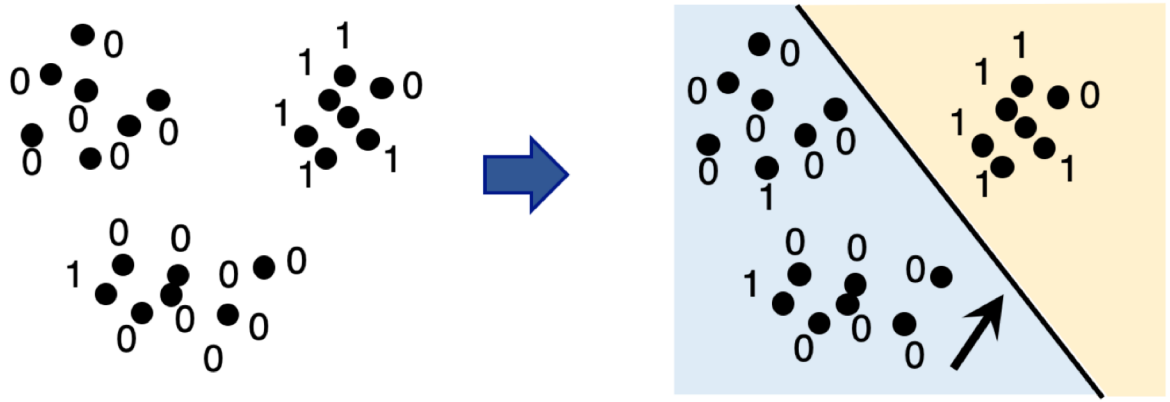
1. Association Rules
2. Clustering
3. Classification
4. Regression

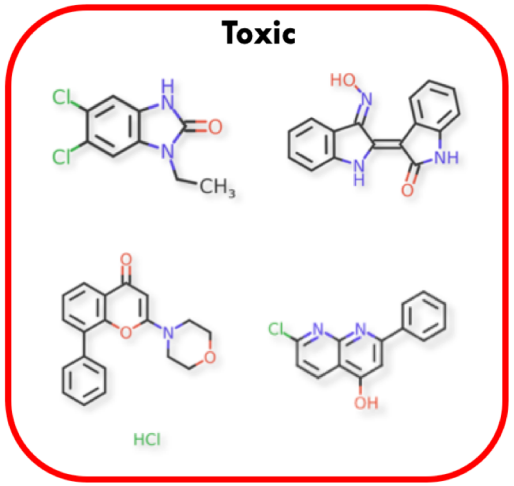


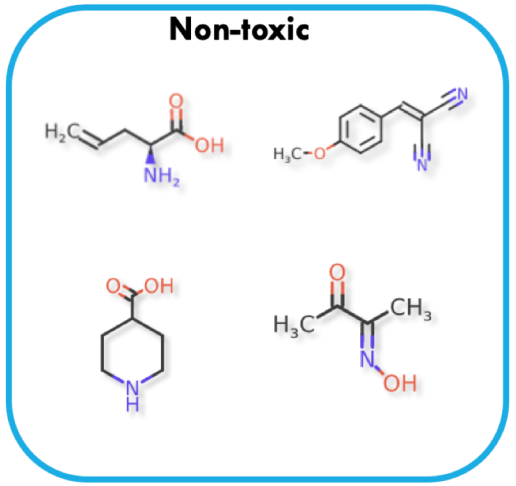


Q: THE MOLECULE TOXICITY TASK WAS AN EXAMPLE OF…?

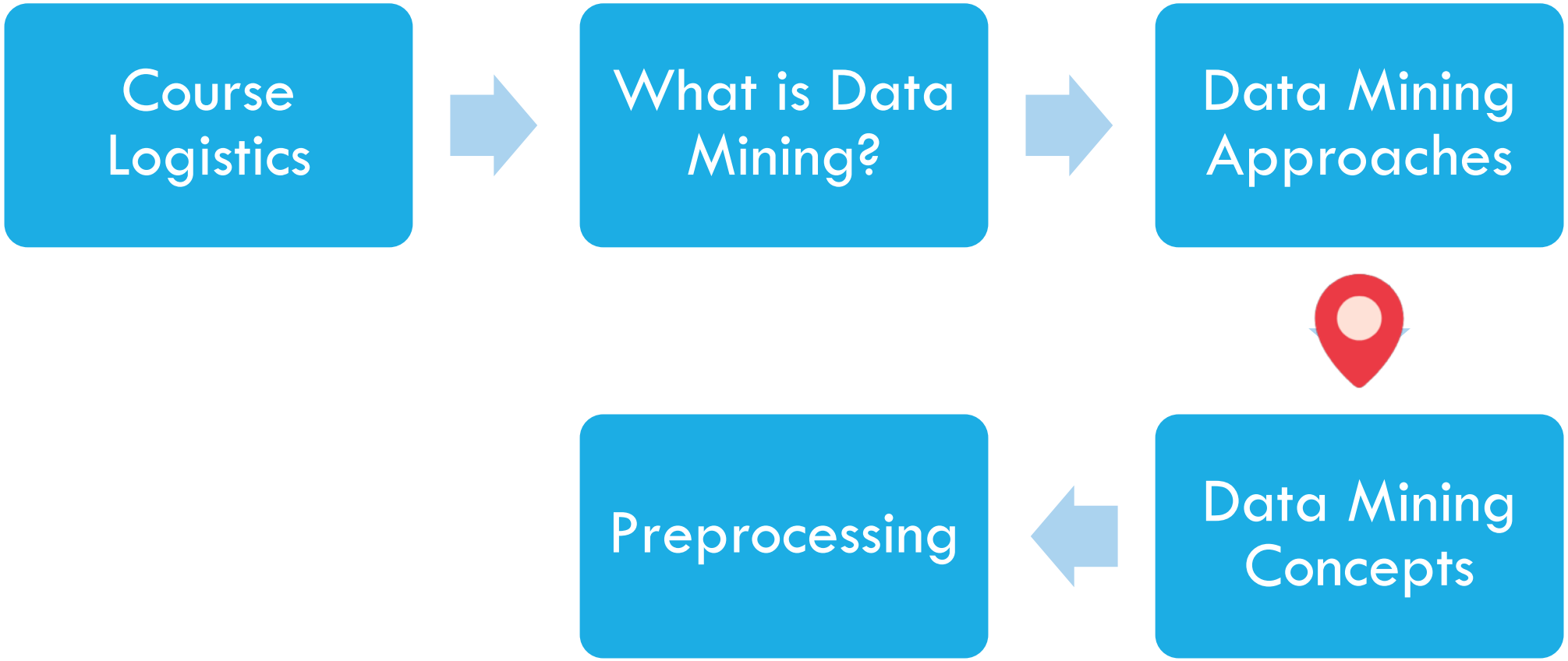
1. Association Rules
2. Clustering
3. Classification







OUTLINE


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

DATASETS: DEFINITIONS

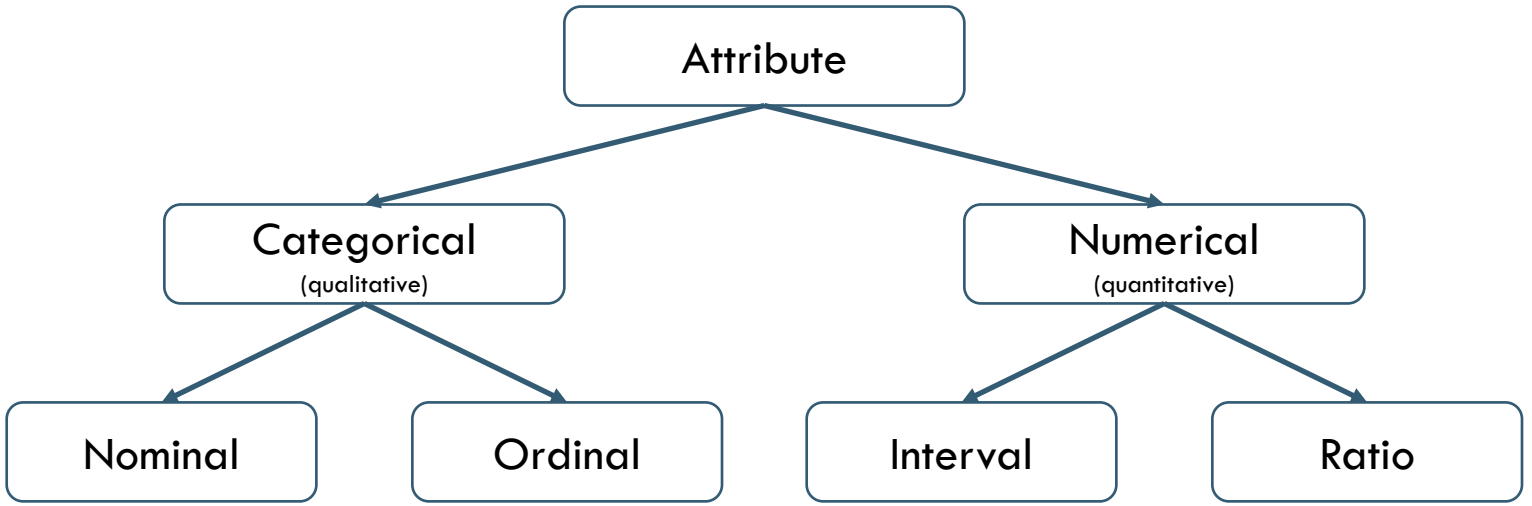
Attributes / **Features** are properties of each object

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



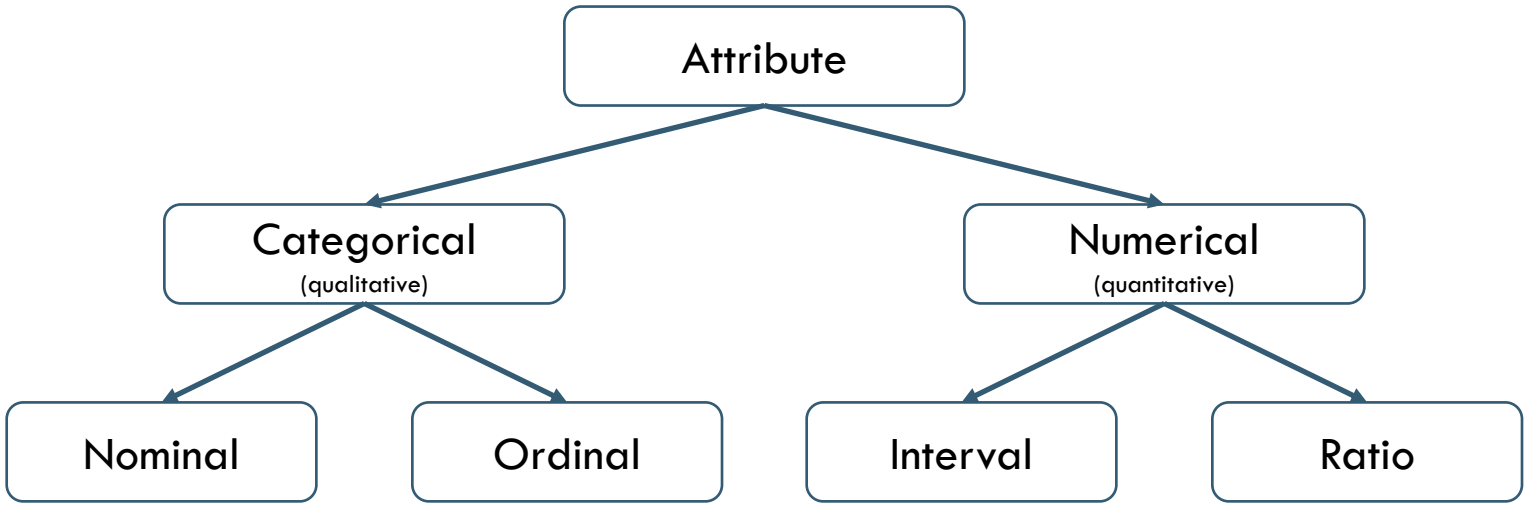
- Values are only labels
- Operations: =, ≠
- E.g.: eye color

- Values are labels with a meaningful order
- Operations: =, ≠, <, >
- E.g.: education level

- Values are measurements with a meaningful distance
- Operations: =, ≠, <, >, +, -
- E.g.: body temperature in ℃

- Values are measurements with a meaningful ratio
- Operations: =, ≠, <, >, +, -, *, /
- E.g.: height, weight

TYPES OF ATTRIBUTES



ID	Age	Education	Birth year
101	23	Masters	1987
102	35	Bachelor	1976
103	26	Masters	1950
104	41	PhD	1999
105	18	Bachelor	2000
…	…	…	…

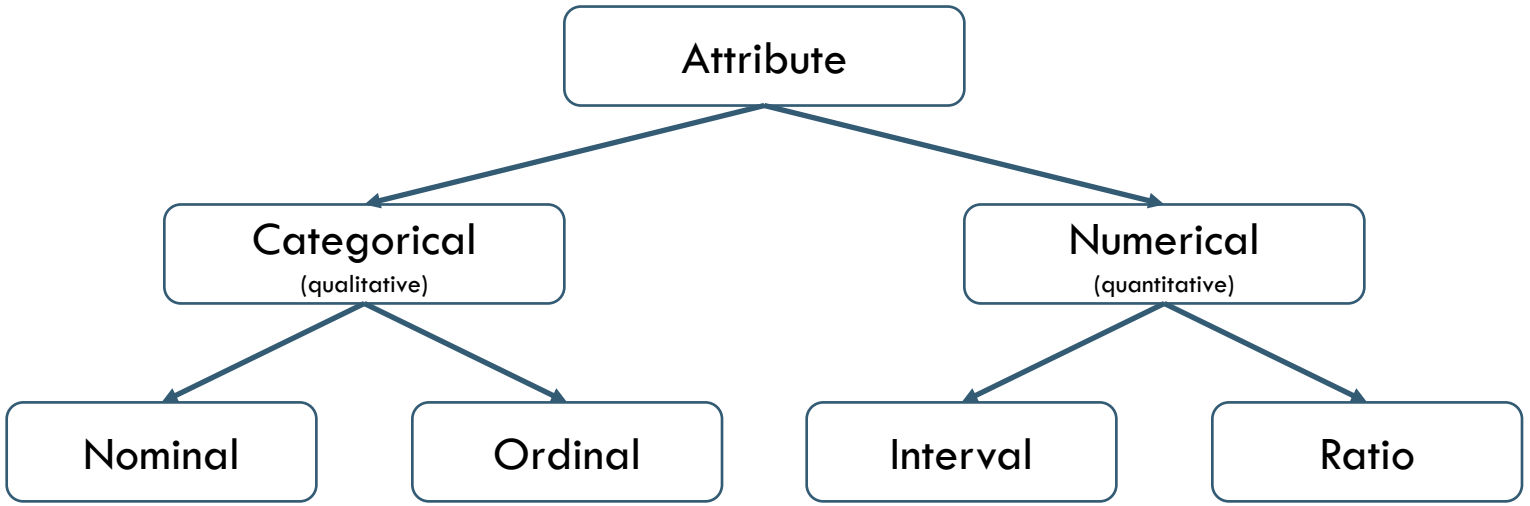
- Values are only labels
- Operations: =, ≠
- E.g.: eye color

- Values are labels with a meaningful order
- Operations: =, ≠, <, >
- E.g.: education level

- Values are measurements with a meaningful distance
- Operations: =, ≠, <, >, +, -
- E.g.: body temperature in °C

- Values are measurements with a meaningful ratio
- Operations: =, ≠, <, >, +, -, *, /
- E.g.: height, weight

TYPES OF ATTRIBUTES



- Values are only labels
- Operations: =, ≠
- E.g.: eye color

- Values are labels with a meaningful order
- Operations: =, ≠, <, >
- E.g.: education level

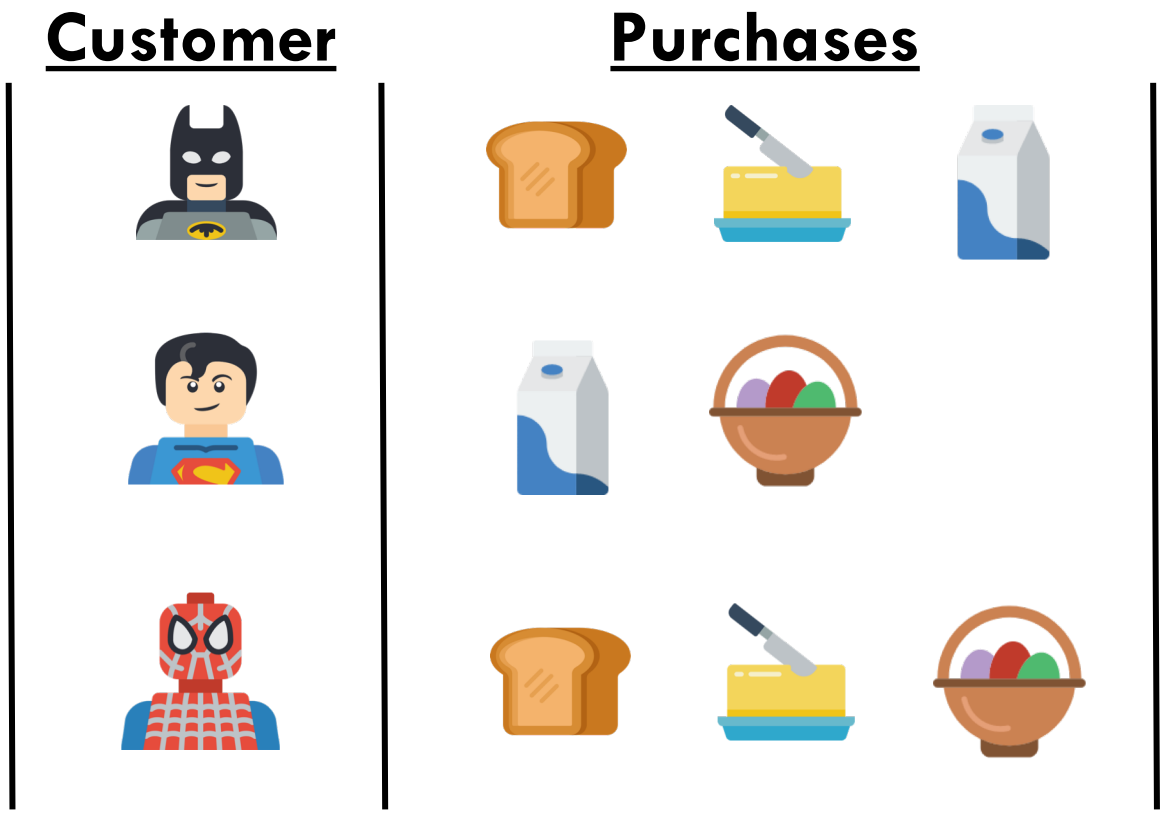
- Values are measurements with a meaningful distance
- Operations: =, ≠, <, >, +, -
- E.g.: body temperature in °C

- Values are measurements with a meaningful ratio
- Operations: =, ≠, <, >, +, -, *, /
- E.g.: height, weight

Nominal Ratio Ordinal Interval

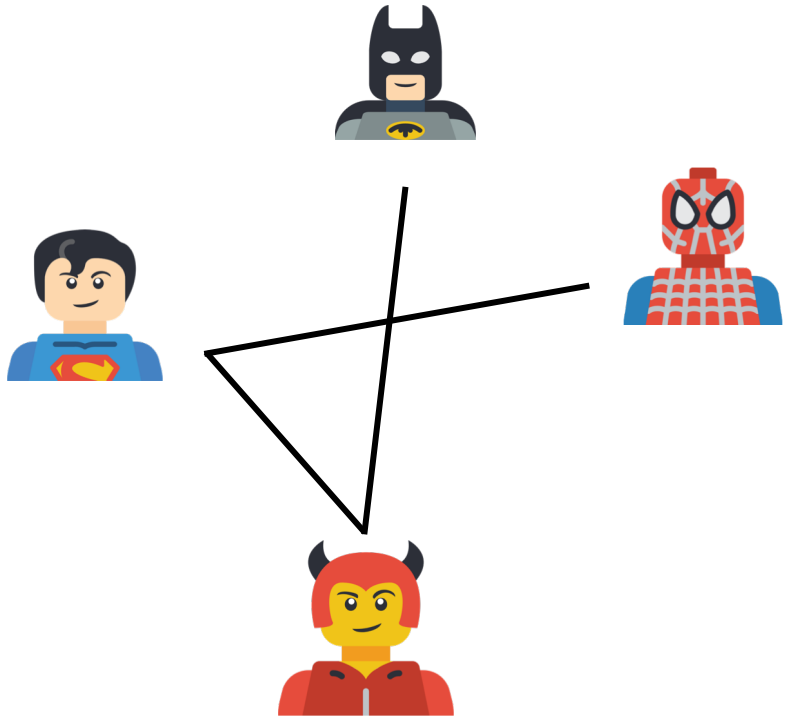
ID	Age	Education	Birth year
101	23	Masters	1987
102	35	Bachelor	1976
103	26	Masters	1950
104	41	PhD	1999
105	18	Bachelor	2000
…	…	…	…

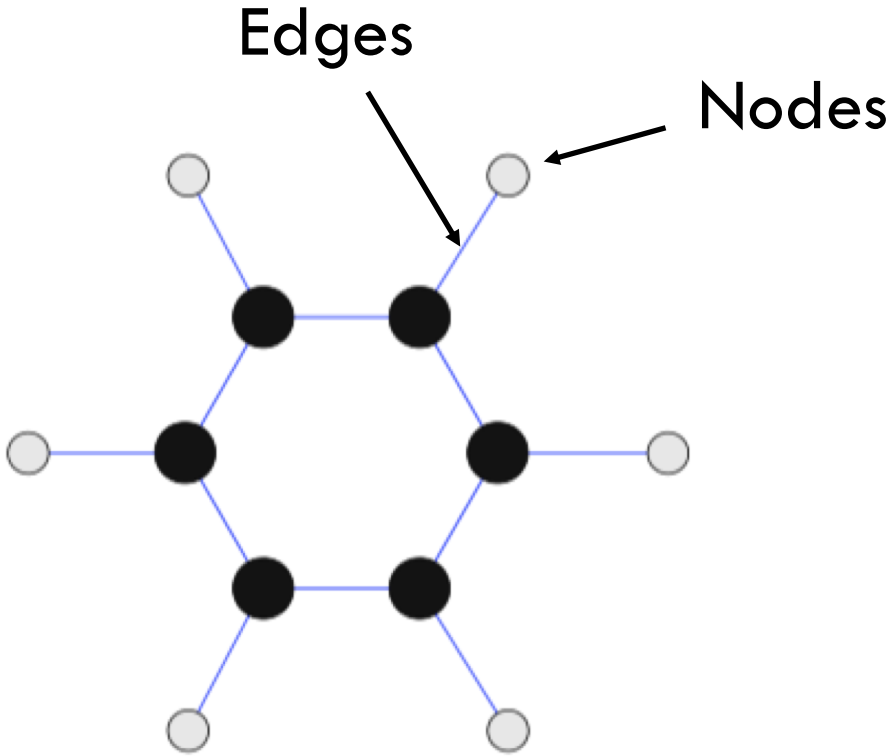
TRANSACTION DATA



Each record (or 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 users, and edges represent relationships of any kind; e.g. friendships

EXPLORATORY DATA ANALYSIS (EDA)

EDA is: simple plots or transformations to understand your data better

Useful for:

- Assessing data quality
- Basic sanity checks
- Initial insights
- Formulate new questions

No formal process with strict rules!

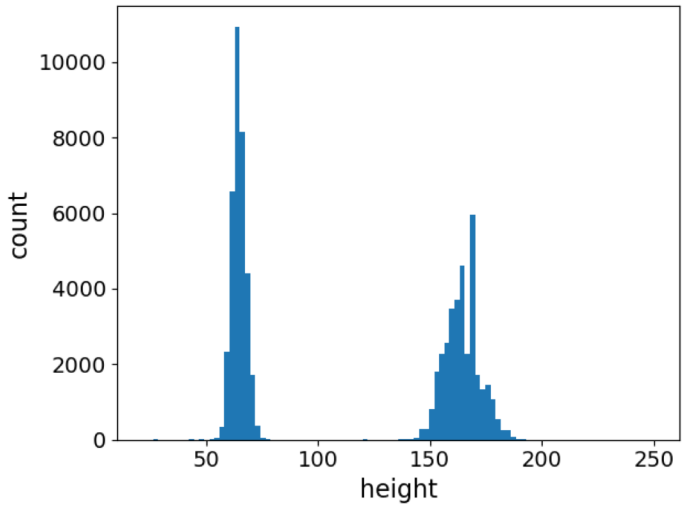
Running example:
Cardiovascular Disease Dataset
(modified to make some points)

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Source: Cardiovascular Disease dataset

EDA - HISTOGRAMS

- Using **histograms** to inspect distribution of data values

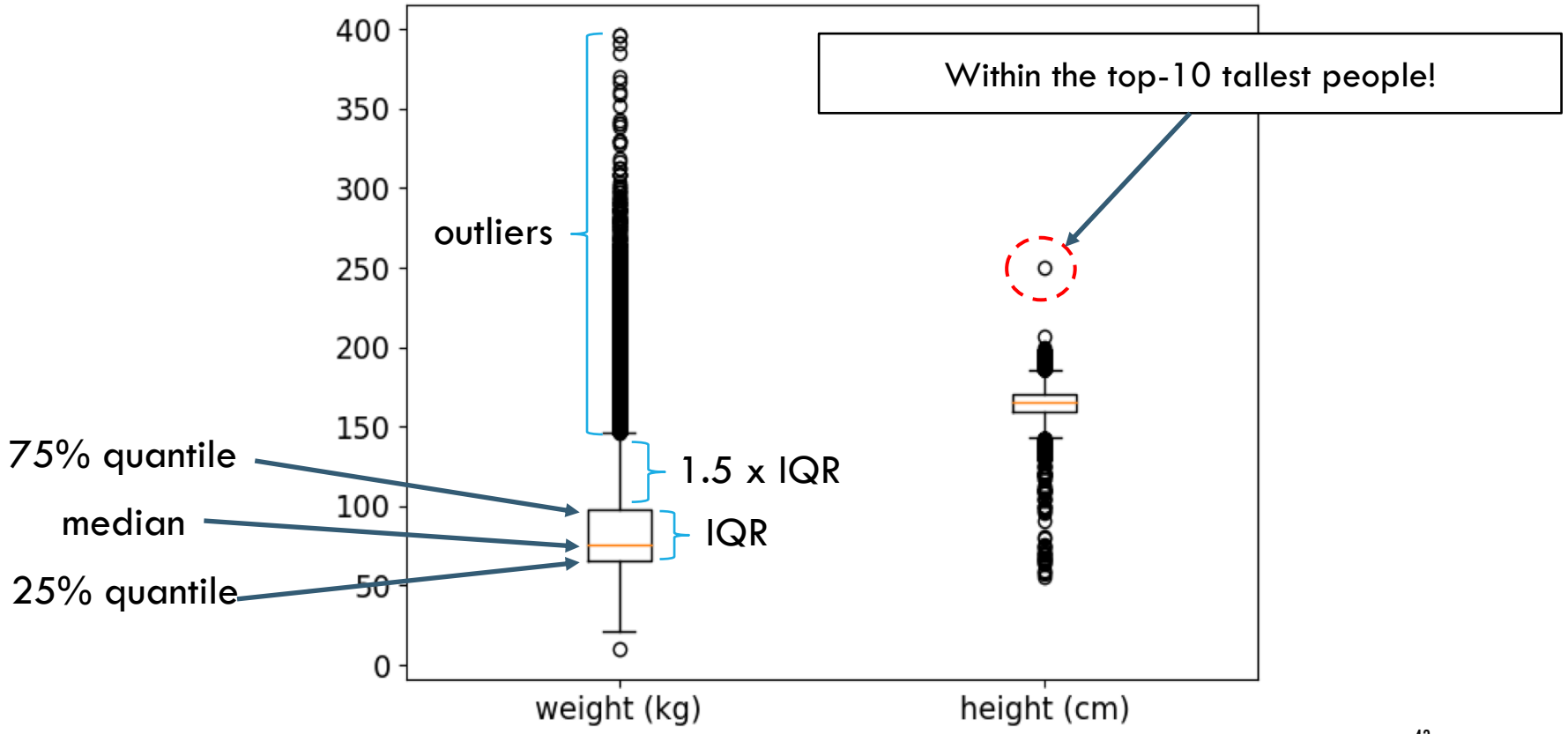


Noise in the height values

- 50% measured in inches
- 50% measured in centimeters

EDA - BOXPLOTS

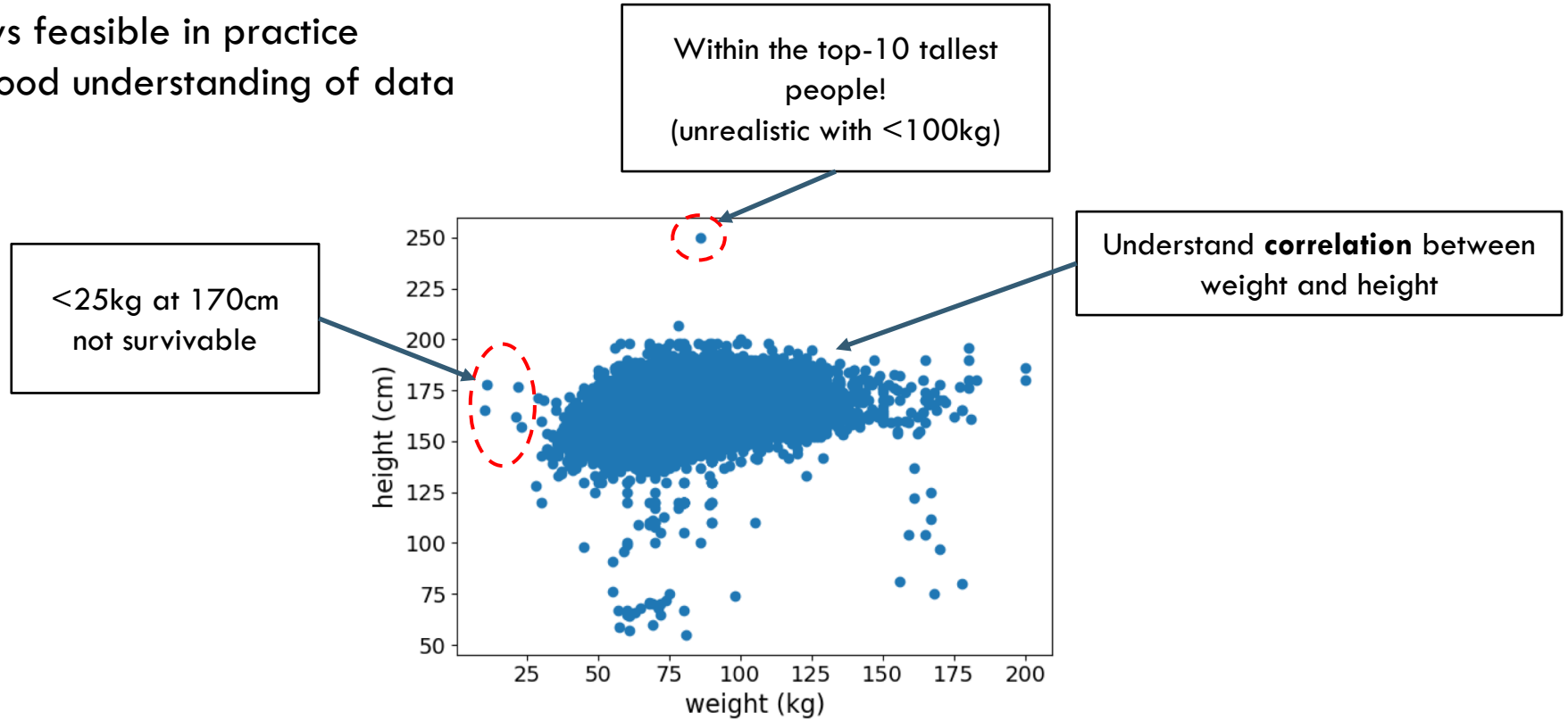
Boxplots show the distribution of a feature in a summarized way:



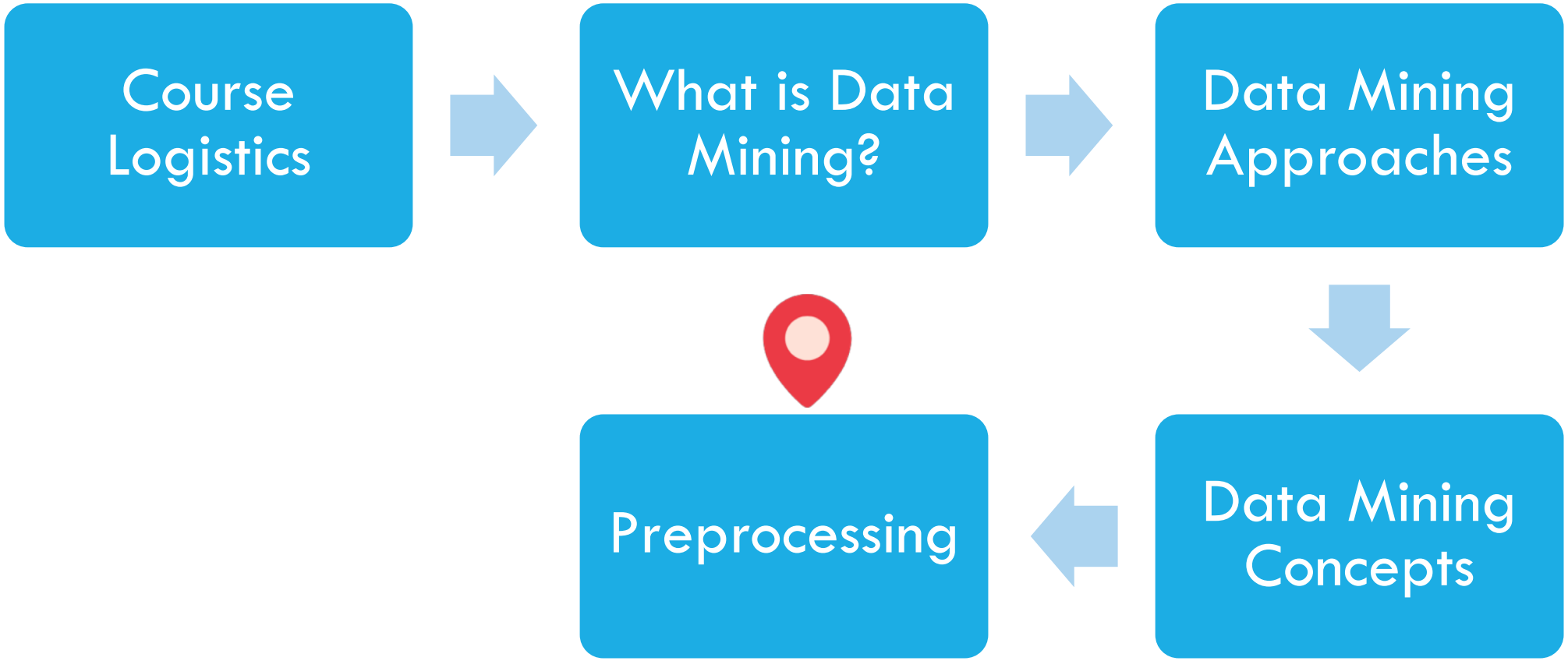
EDA - SCATTERPLOTS

Using scatter plot to inspect correlations

- Not always feasible in practice
- Require good understanding of data



OUTLINE


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

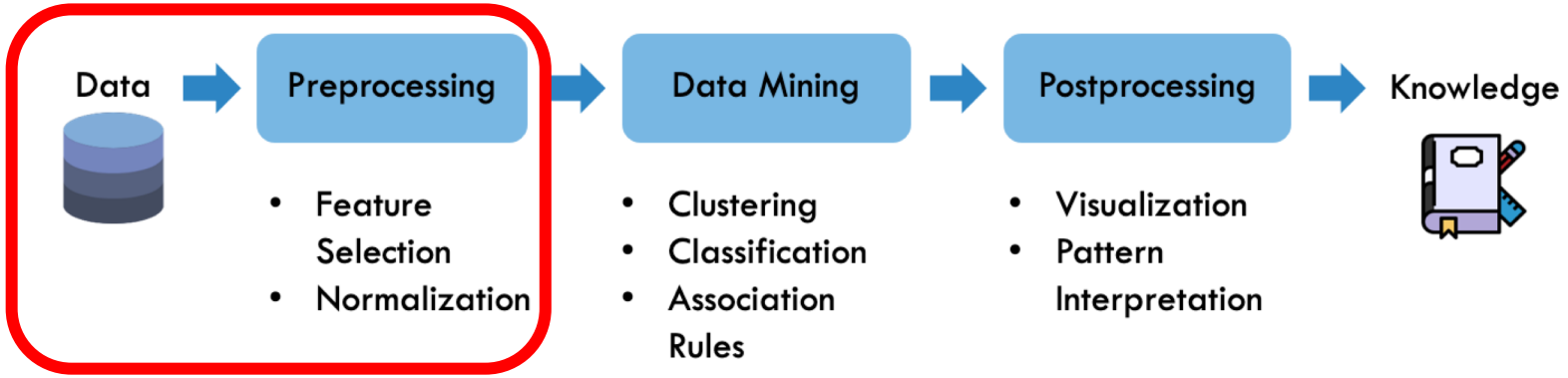
DATA PREPROCESSING

Improving Data Quality ("Cleaning")

- Remove Outliers
- Missing Values
- Duplicates

Transformations

- Normalization
- Aggregation
- Feature Creation
- Discretization
- One-Hot Encoding
- Dimensionality Reduction



DATA QUALITY

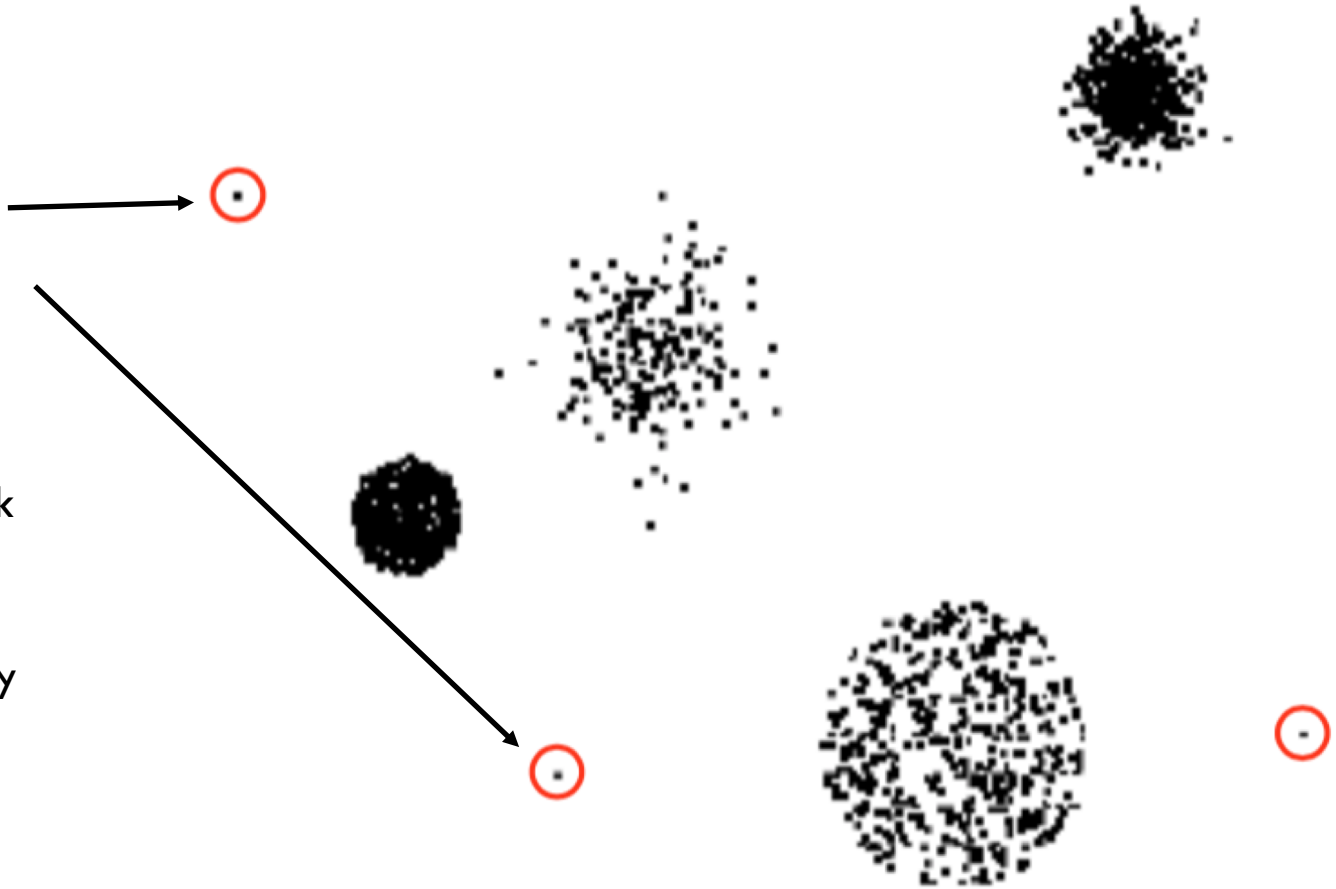
"Garbage in, garbage out"

Purposes: Preprocessing can be used to **fix data quality** issues, or to transform the data into a form to **ease analysis**

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 (rows) 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: 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 (rows) 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
  - **Dummy variables**: optionally insert a column which is 1 if the variable was missing, and 0 otherwise

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

Median Imputation

UserID	Height (m)	Missing?	Country	...
1	1.61	0	SG	...
2	1.50	0	US	...
3	***1.55***	1	MY	...
...	...		...	...

DATA QUALITY: DUPLICATES

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

- E.g. same person with multiple email addresses
- **But:** duplicates could also arise from genuinely different objects! Use domain knowledge before removing duplicates

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	…
…	…	…	…

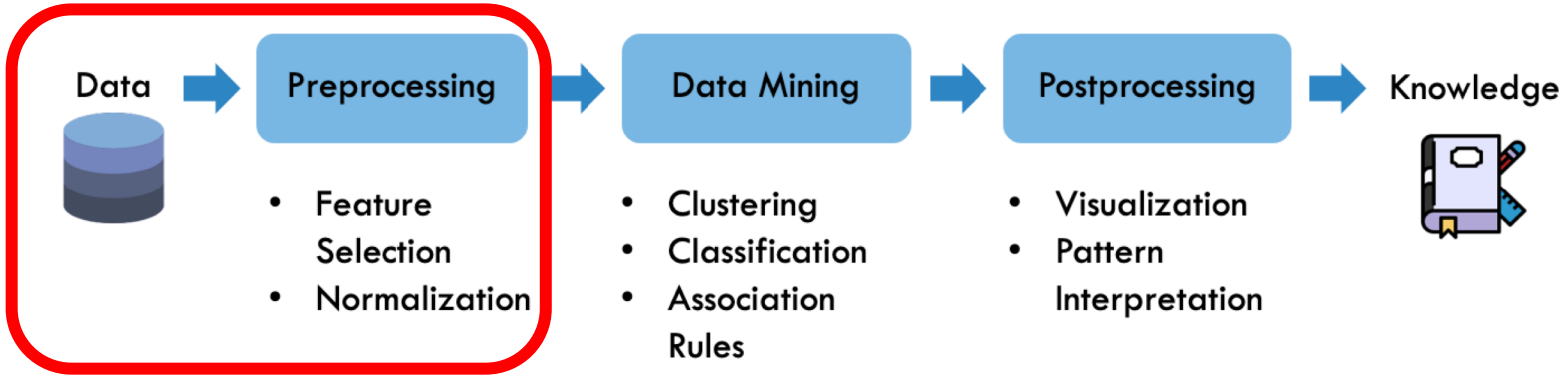
DATA PREPROCESSING

Improving Data Quality ("Cleaning")

- Remove Outliers
- Missing Values
- Duplicates

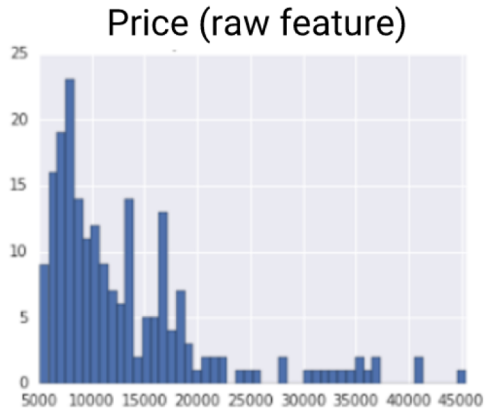
Transformations

- Normalization
- Aggregation
- Feature Creation
- Discretization
- One-Hot Encoding
- Dimensionality Reduction

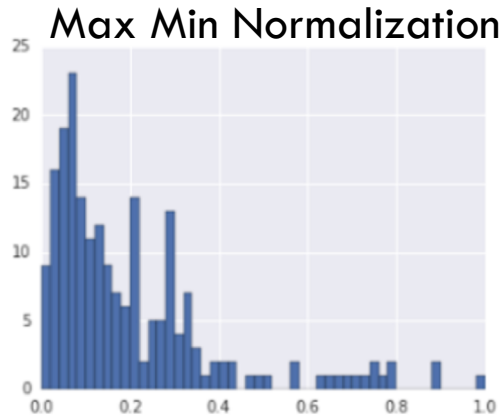


NORMALIZATION

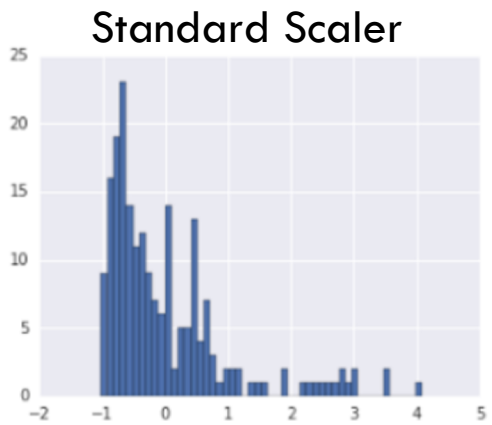
Normalization helps to map the data to a more "suitable" range, e.g. from 0 to 1 (for human analysis, or for subsequent data mining)



$$\frac{x - \min(x)}{\max(x) - \min(x)}$$



$$\frac{x - mean(x)}{std(x)}$$



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

Pro

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

Con

Data is less “granular”, so information may be lost

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

Pros

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

Cons

- Loss of information, particularly if too low dimension size is used

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

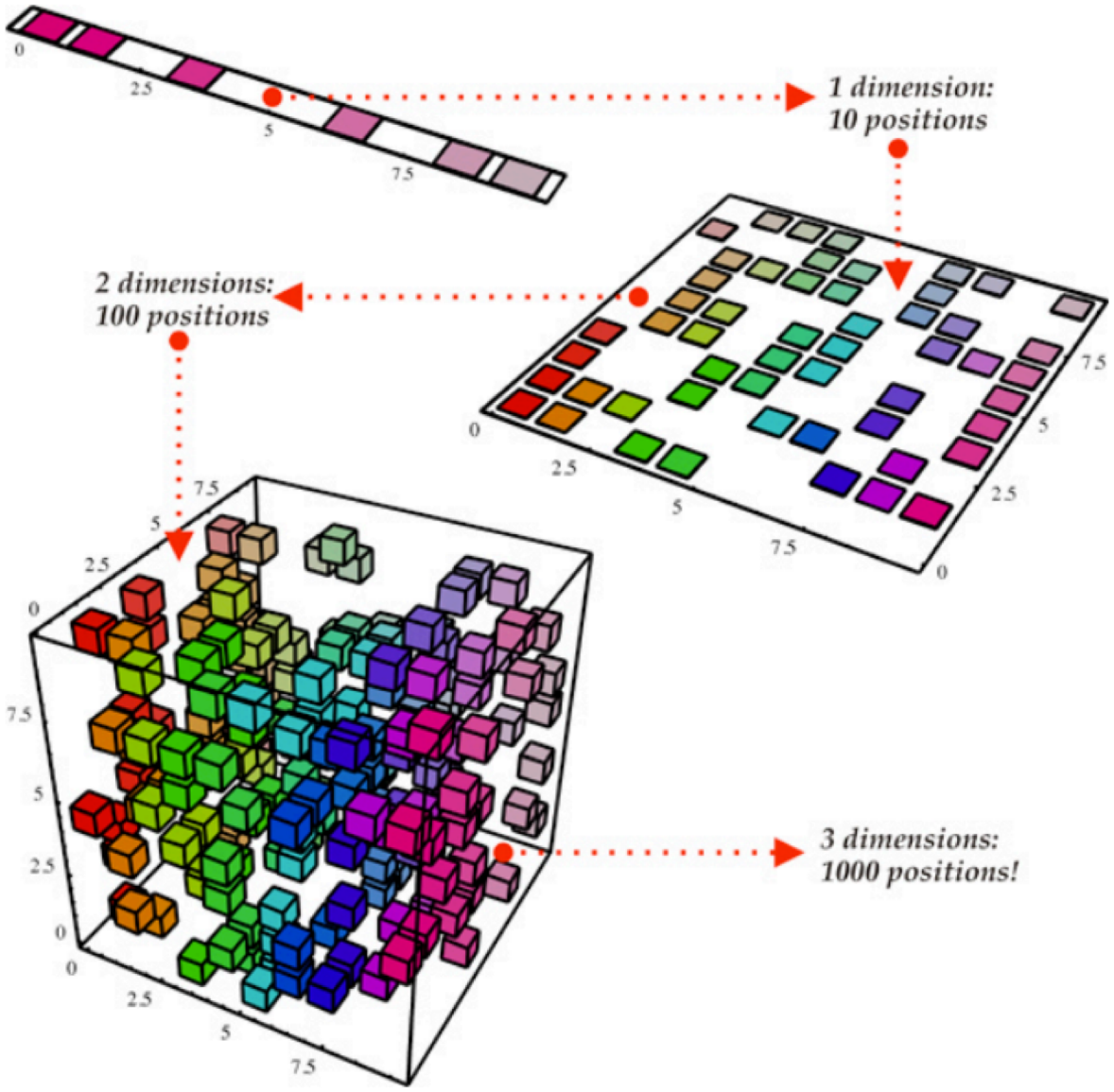
Main observation: many algorithms perform poorly in high-dimensional spaces

- Especially: **distance** and **neighborhood**-based algorithms (we will cover these later on the class)

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

Hence, the space becomes sparser (i.e. emptier)

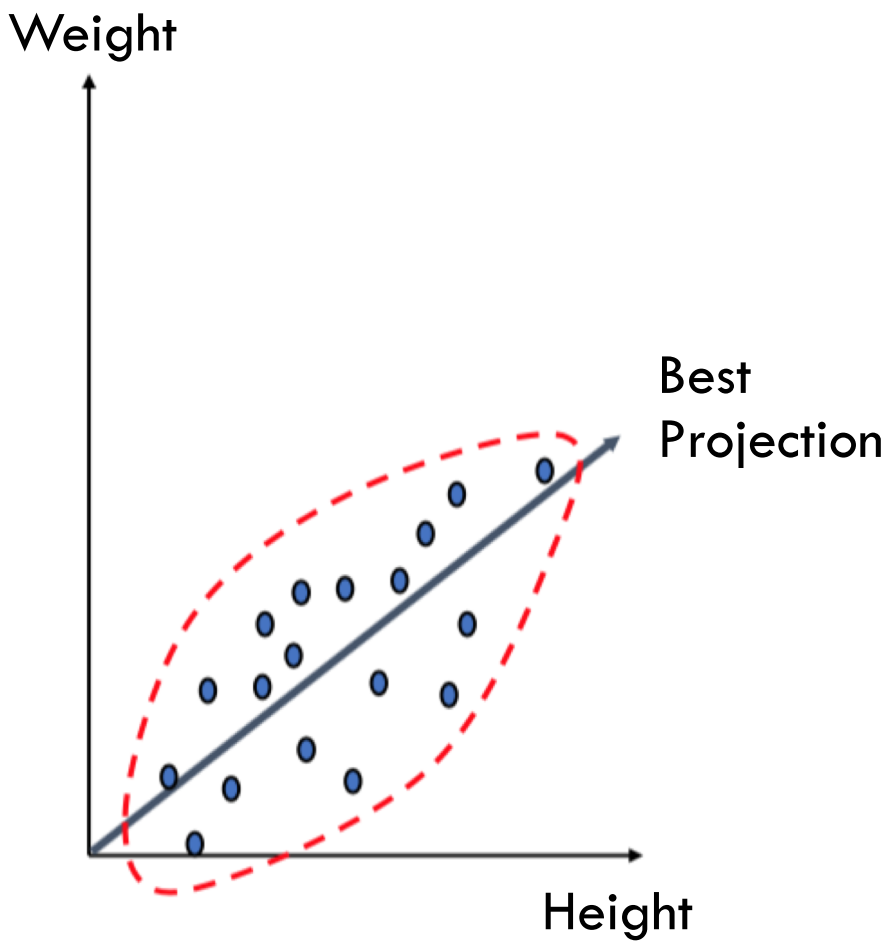
Many algorithms fail to make effective predictions at any given test point, as there are no nearby data to use



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

PRINCIPAL COMPONENT ANALYSIS

PCA reduces dimensionality by finding the best projection line that minimizes the total squared distance between the data points and the line.

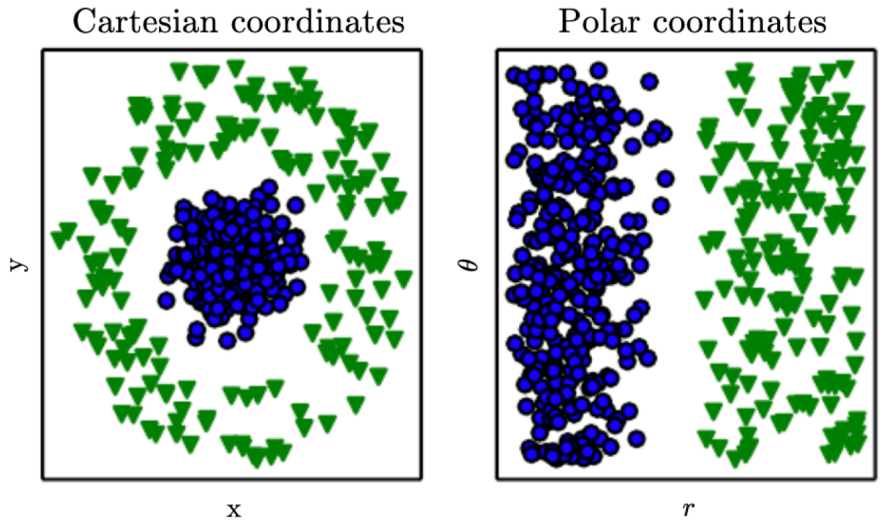


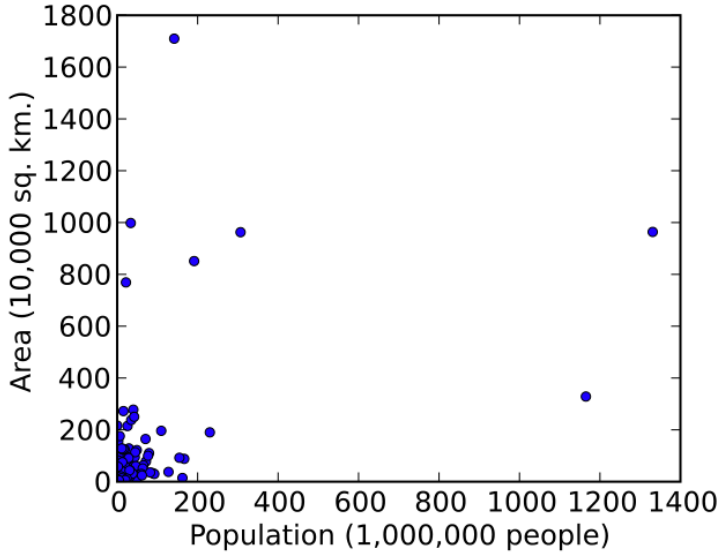
https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

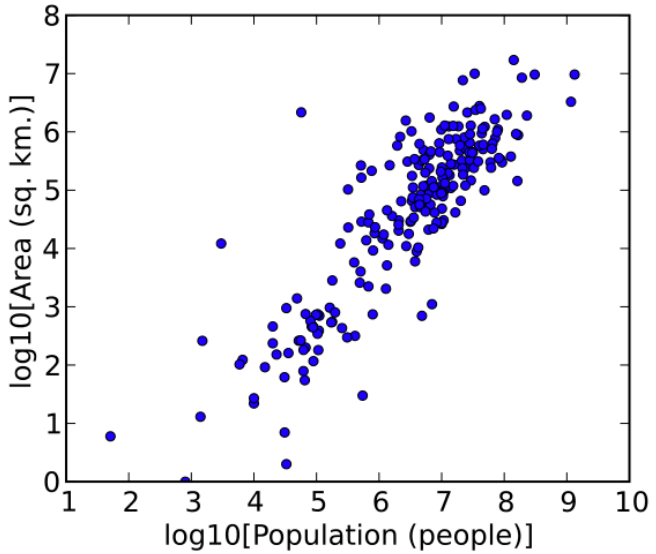
FEATURE ENGINEERING

Transform features, or create new features, that capture the important information in data better than the original features. Examples:

- **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://www.deeplearningbook.org/

DISCRETIZATION (OR BINNING)

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

- Ex: group the "height" variable into buckets of 10cm

Pro

Some algorithms are more flexible when given discrete (grouped) variables

Con

Loss of fine-grained information

Var1	Height
1	143.4
2	144.5
3	193.4
…	…

Var1	Bucket
1	(0, 150]
2	(0, 150]
3	(170, 190]
…	…

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 apply algorithms which can handle binary features (e.g. linear regression)

Group
2
1
3
…

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