LECTURE 2
CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

LEK HSIANG HUI

OUTLINE

Introduction to Data Mining

CRISP-DM

Introduction to Web Mining

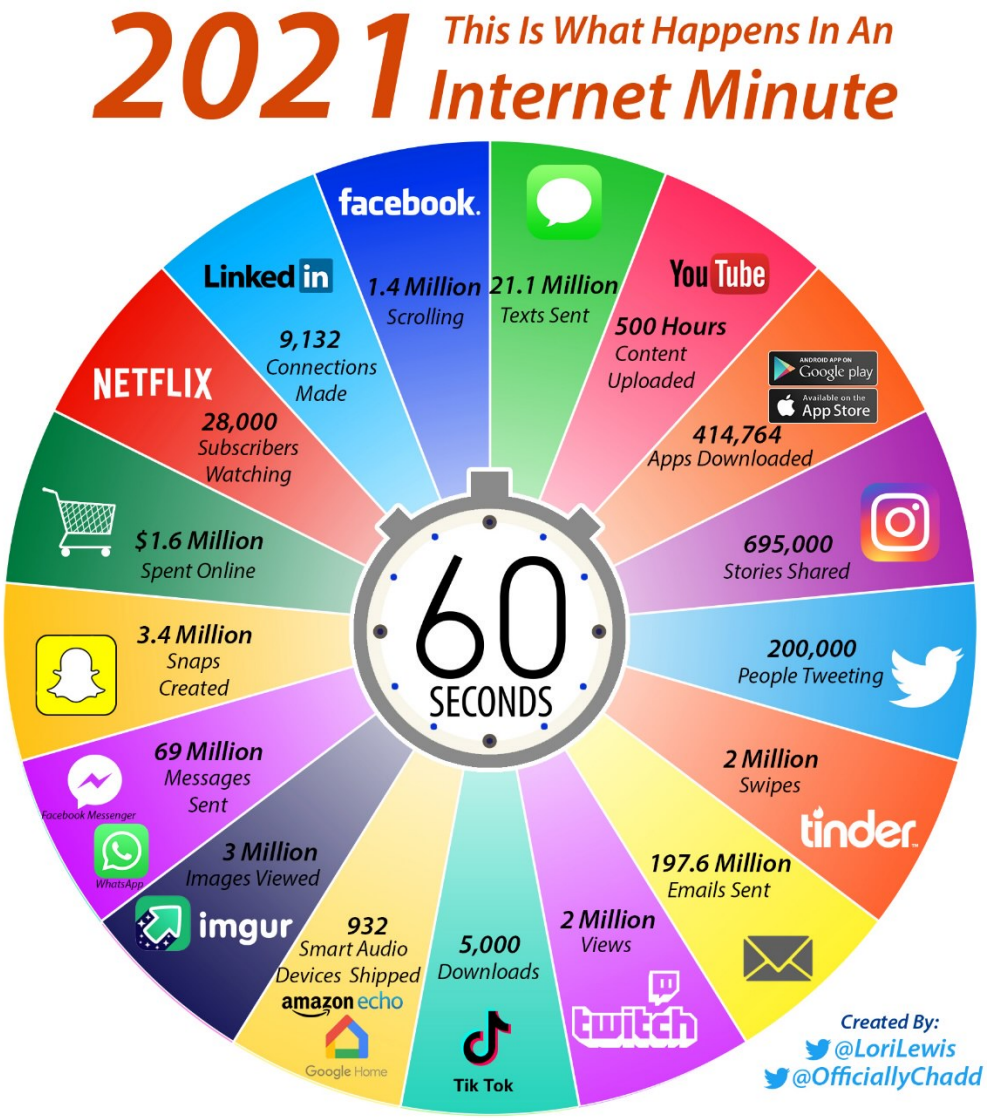
DATA MINING

Since everything is computerized nowadays, data is now stored in digital form (e.g. databases)

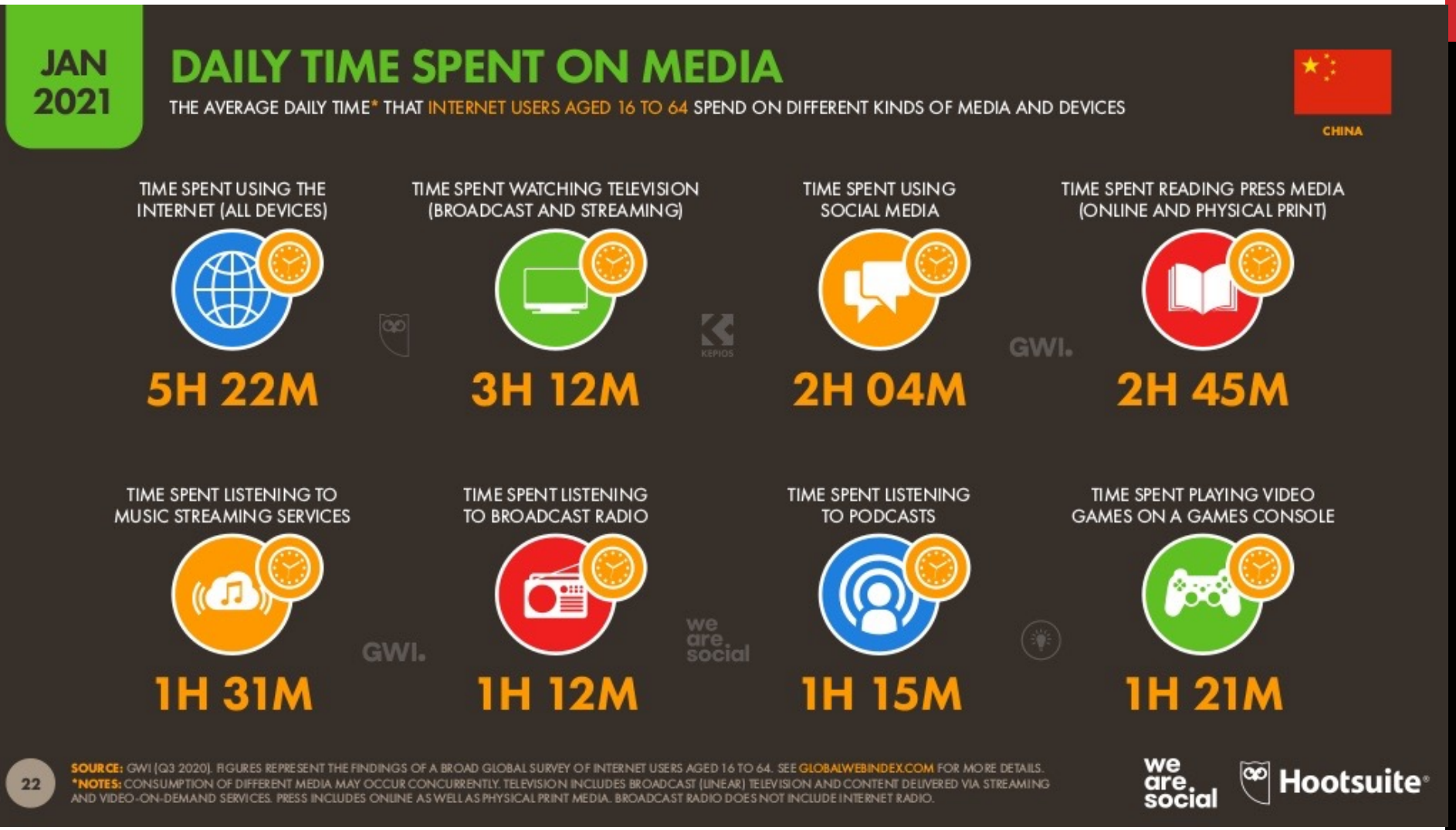
From these databases, the purpose of data mining is to look for patterns so as to discover more insights to the raw data

- Raw Data ➡ Patterns ➡ Knowledge

WHAT HAPPENS IN 1 MINUTE (2021)?



AVERAGE TIME SPENT ON MEDIA (CHINA)

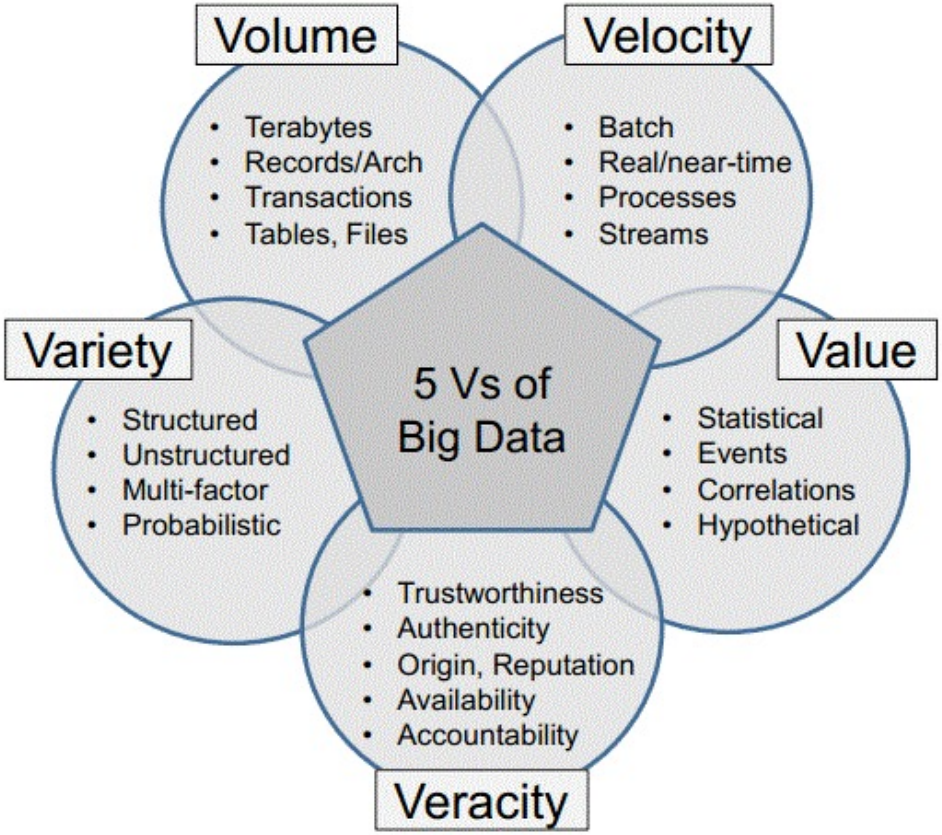


WHAT IS BIG DATA?

Traditionally, "Big Data" = large amount of data

- Transactional data, web pages, text documents, images, call records, social networking data, GPS location, sensor data, call records, medical records, etc

WHAT IS BIG DATA?

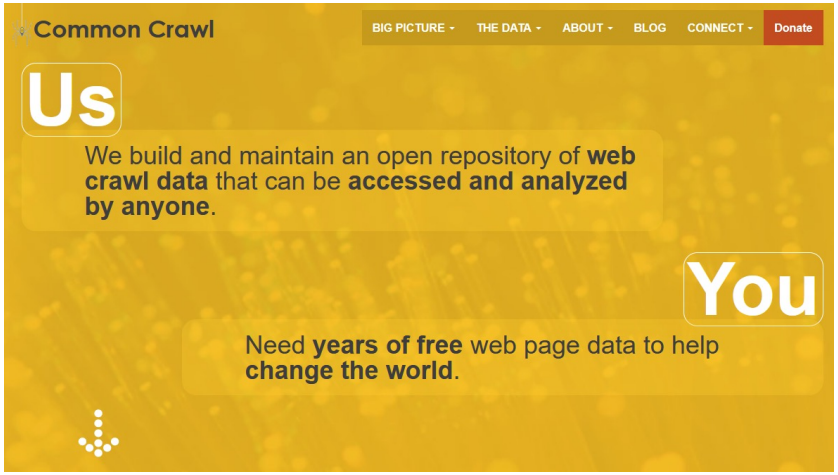


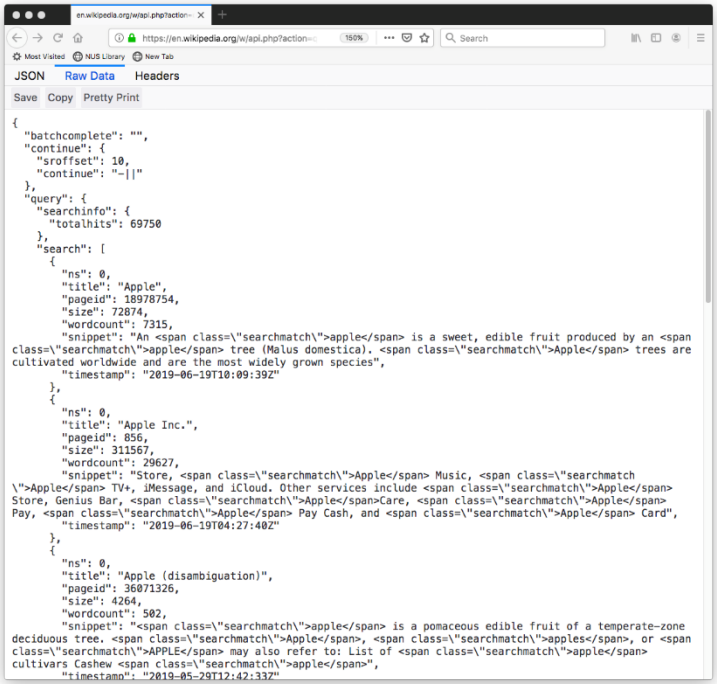
The 5Vs
of
Big Data

WHY "BIG DATA"?

Many factors that contribute to this concept

- Possible to collect data automatically now (Publicly available datasets, APIs, web crawler)





BIG DATA ANALYTICS

Big Data by itself, regardless of the size, type, or speed, is worthless

Big Data + "Big" analytics = value

CRISP-DM

Cross-industry standard process for data mining (CRISP-DM) breaks the process of data mining into 6 major phases

BUSINESS UNDERSTANDING

STEP 1 – BUSINESS UNDERSTANDING

Understand the purpose of the data mining study

- Project objectives
- Requirements of the business
- Rough idea of potential data to use for analysis
- Preliminary plan

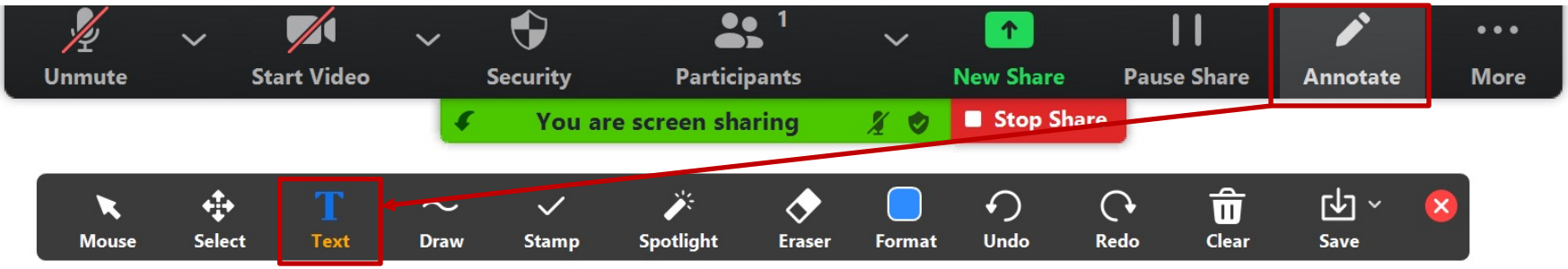
Notice that the process starts with the business understanding (i.e. problem)

- It does NOT start with the data!

DISCUSSION:
STEP 1 – BUSINESS UNDERSTANDING

Suppose you are a data analyst who has been hired by Taobao/Amazon

- Propose some analytics initiatives where the company that will benefit from
  - Use the annotate feature on zoom to propose the ideas



Propose some analytics initiatives where Taobao/Amazon will benefit from

Suppose the initiative is [TO BE INSERTED], what is the modeling problem?

Suppose the modeling problem is [TO BE INSERT], what are the potential data that can be used?

DATA UNDERSTANDING

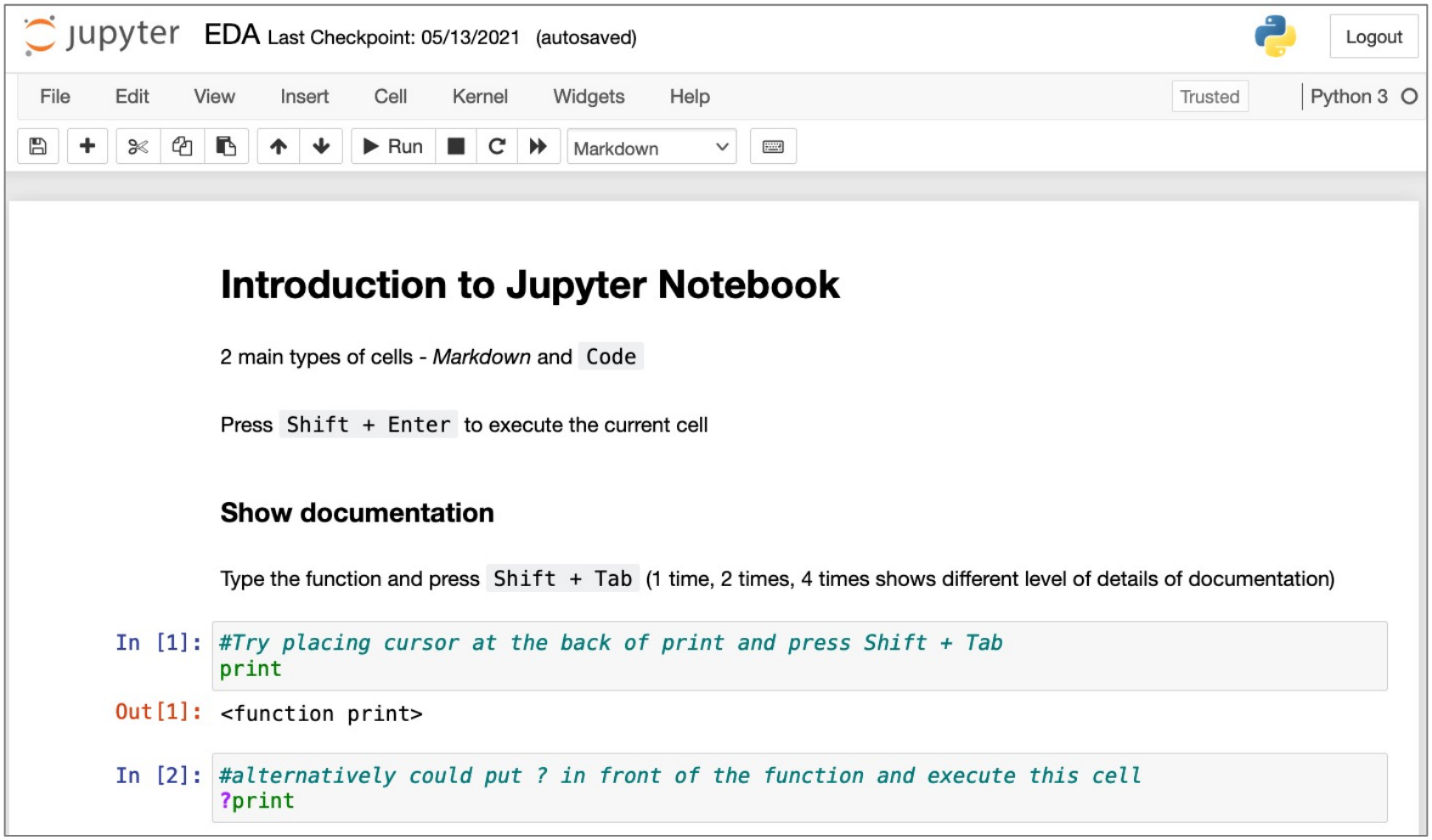
STEP 2 – DATA UNDERSTANDING

Identify the relevant data from the many sources

- Normally: download and use datasets off internet
- Now: learn how to mine the datasets yourself
- Then, perform **Exploratory Data Analysis**
  - Perform statistical analysis
  - Perform various types of visualizations

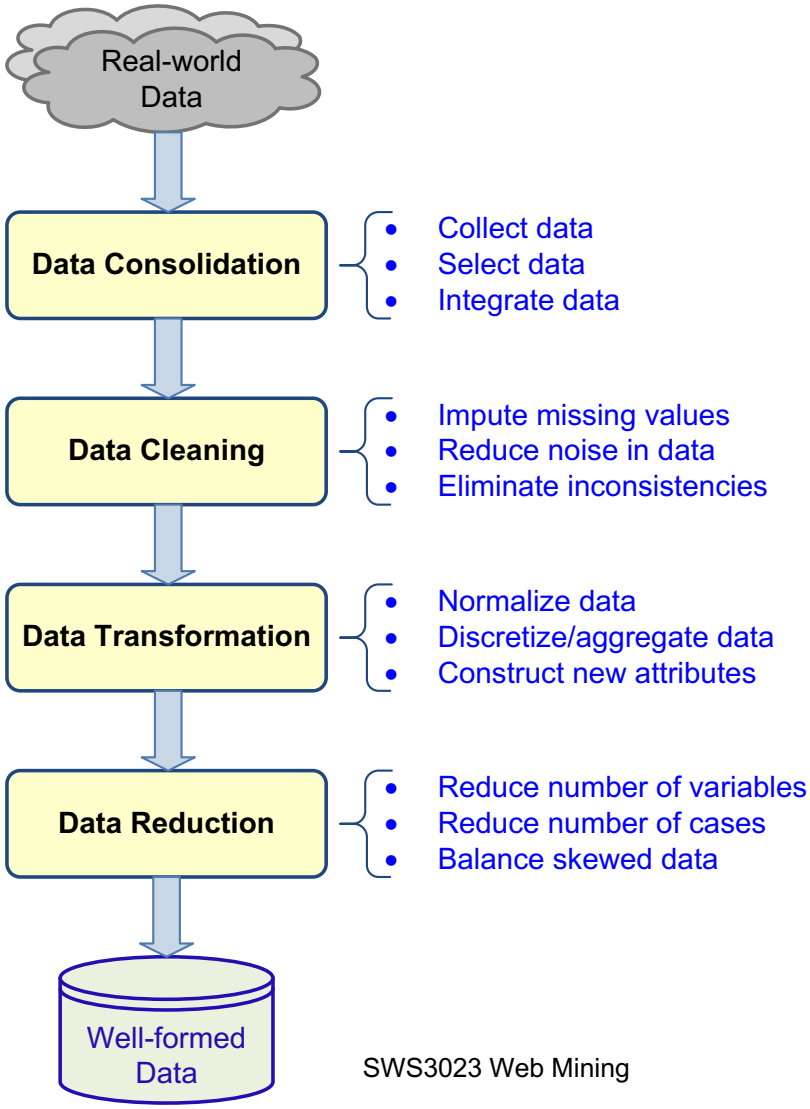
HANDS-ON: EXPLORATORY DATA ANALYSIS (EDA)

Download and access:
EDA.ipynb



DATA PREPARATION

STEP 3 – DATA PREPARATION



DATA CLEANING

The mathematical models for decision making can only achieve accurate and effective results when the data is highly reliable

The data can have several anomalies which affects the quality of the data:

- Incompleteness
- Noise
- Inconsistency

It is important to validate your data before analysis in order to arrive at an accurate conclusion

HANDLING INCOMPLETE DATA

Several techniques for handling incomplete data:

Elimination:

- Discard all records with one or more missing attributes
- Remove the attribute with missing values
- However, elimination will lead to smaller dataset which may affect the model performance

HANDLING INCOMPLETE DATA

Substitution:

- Automatic replacement of missing values
- Missing values of a numerical variable, replace with the mean calculated from the remaining observations
  - Or just considering the nearby observations (time-series data)
  - etc

DETECTING AND HANDLING NOISY DATA

Outliers in dataset need to be identified and either:

- Corrected and regularized
- Entire records containing noise are eliminated

Dispersion:

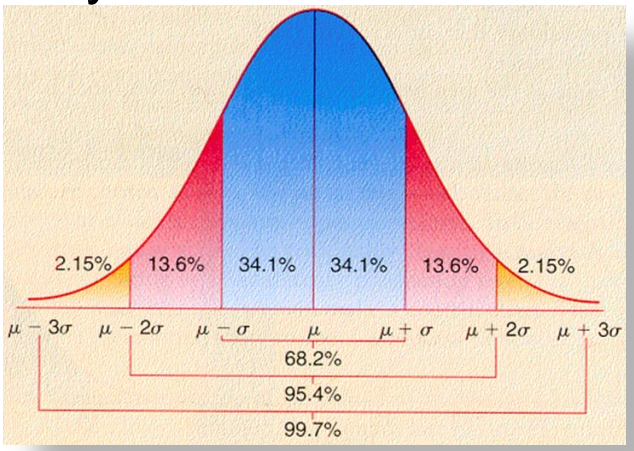
- One way to detect outliers is based on the concept of **statistical dispersion**

DETECTING OUTLIERS (3-SIGMA)

If the distribution of an attribute is normal or approximately normal:

- The interval $(\overline{\mu} \pm \overline{\sigma})$ contains approximately 68% of the observed values.
- The interval $(\overline{\mu} \pm 2\overline{\sigma})$ contains approximately 95% of the observed values.
- The interval $(\overline{\mu} \pm 3\overline{\sigma})$ contains approximately 100% of the observed values.

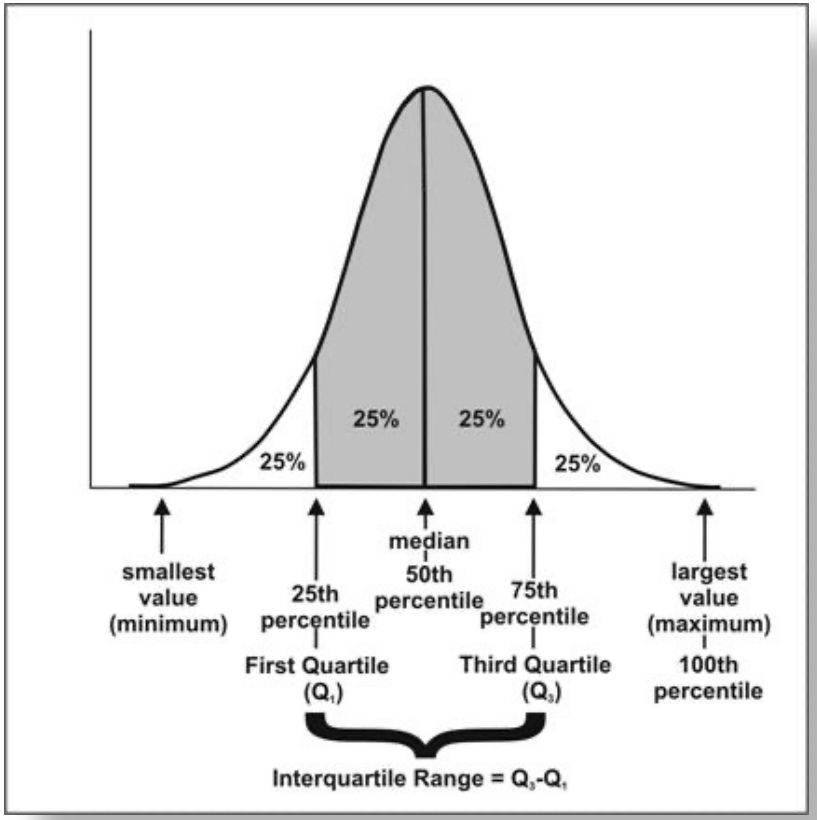
Values that falls outside $(\overline{\mu} \pm 3\overline{\sigma})$ can be considered as suspicious outliers.



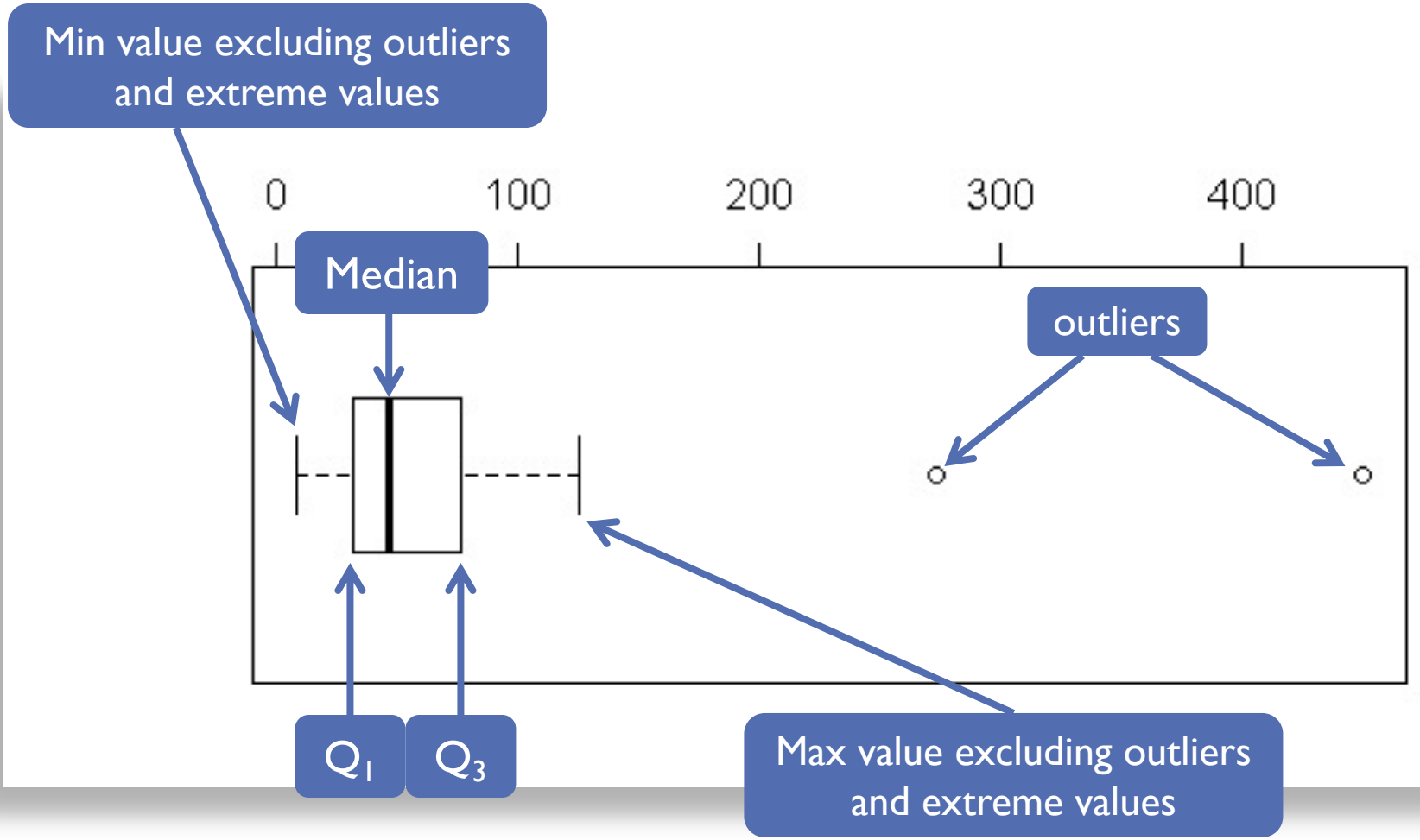
DETECTING OUTLIERS (BOXPLOT)

Interquartile Range (IQR)

- Difference between the 3rd and 1st quartiles
- The interquartile range is useful to identify outliers
- This measure is used in **box-plot**



DETECTING OUTLIERS (BOXPLOT)



DETECTING OUTLIERS (BOXPLOT)

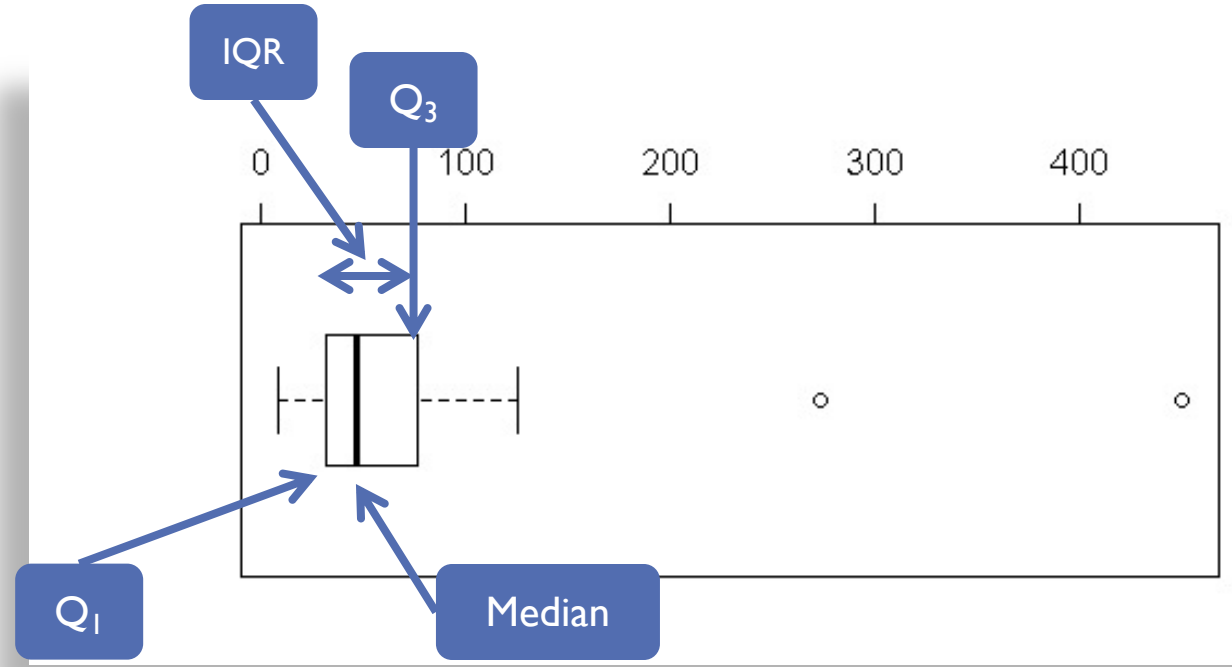
Data: [45, 450, 20, 69, 9, 66, 11, 42, 9, 126, 47, 43, 24, 94, 89, 16, 83, 59, 57, 273, 70, 45, 40]

Median = 47

Q1 = 32

Q3 = 76.5

IQR = 44.5



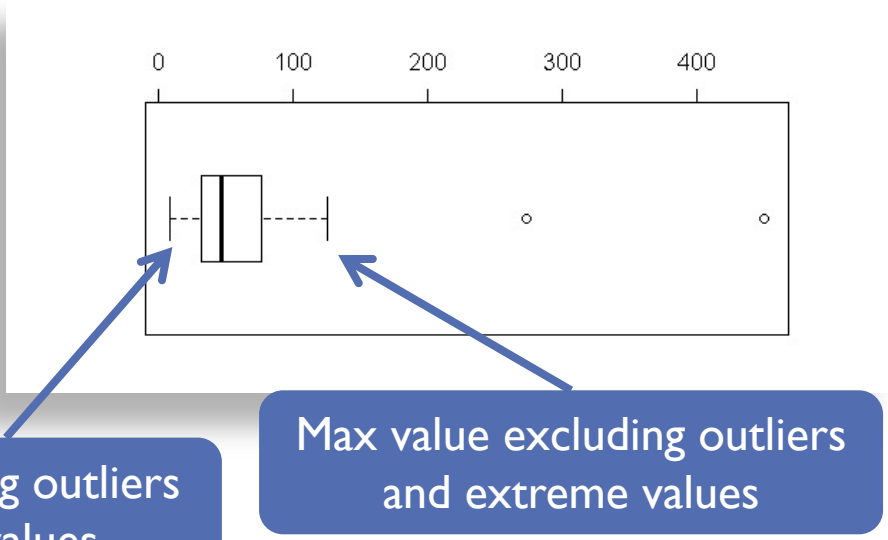
DETECTING OUTLIERS (BOXPLOT)

Data: [45, 450, 20, 69, 9, 66, 11, 42, 9, 126, 47, 43, 24, 94, 89, 16, 83, 59, 57, 273, 70, 45, 40]

Median = 47, Q1 = 32, Q3 = 76.5, IQR = 44.5

To determine whether a point is an outlier or extreme value, we make use of 4 thresholds:

- 1.5*IQR = 66.75
- QL -1.5*IQR = -34.75
- QU +1.5*IQR = 143.25
- Min = 9
- Max = 126



DETECTING OUTLIERS (BOXPLOT)

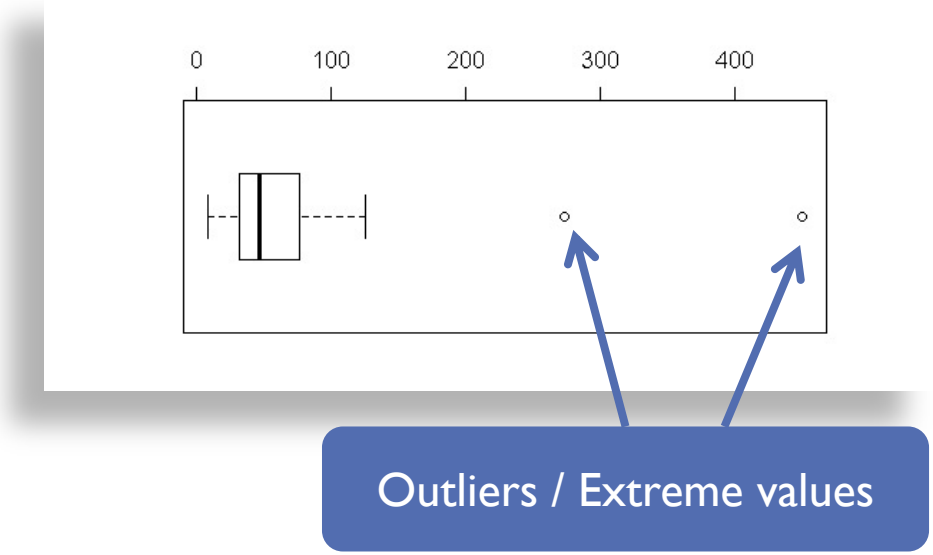
Data: [45, 450, 20, 69, 9, 66, 11, 42, 9, 126, 47, 43, 24, 94, 89, 16, 83, 59, 57, 273, 70, 45, 40]

Outliers

- <Q1 – 1.5*IQR
- >Q3 + 1.5*IQR

Extreme values

- <Q1 – 3*IQR
- >Q3 + 3*IQR



DATA INTEGRATION

Data Wrangling is the process of performing data transformation

- Basic data manipulation operations:
- Selecting of rows/columns
- Filtering by conditions
- Merge datasets together
- Grouping columns
- etc

TEACHING/CONSULTATION SCHEDULE

July 2021					
Mon	Tues	Wed	Thurs	Fri	Sat
12 **Predictive Analytics I** (10-12pm) **1st consultation** (1-6pm)	13 **Predictive Analytics II** (10-12pm) **Lab 1 / 1st consultation** (1-6pm)	14 **Mining Web Content II** (10-12pm) **Lab 2 / Ad hoc consultation** (1-6pm)	15 **Mi** **Co** (10 **La** **co** (1-6pm)	(1-6pm)	
19 **Lab 5 / 3rd consultation** (10am-6pm)	20 **Lab 6 / 3rd consultation** (10am-6pm)	21 **Lab 7 / Ad hoc consultation** **Ad hoc help from TA** (10am-6pm)	22 **4th consultation** **Ad hoc help from TA** (10am-6pm)	23 **4th Consultation** **Ad hoc help from TA** (10am-6pm)	24 **Ad hoc consultation** (10-6pm)
26 **Ad hoc consultation** **Ad hoc help from TA** (10am-6pm)	27 **Ad hoc consultation** **Ad hoc help from TA** (10am-6pm)	28 **Project Showcase** (12-6pm)	29	26	26

Will learn more about data manipulation in Lab1 & 2

MODELING


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

STEP 4 – MODEL BUILDING

Apply and compare various data mining techniques

- Some techniques have specific requirements on the form of data (e.g. need to be numeric)
- Most techniques can only be applied to one type of problem (e.g. classification) while others can be applied for both regression and classification

DATA MINING TECHNIQUES

Data mining techniques:

- Regression
- Classification
- Clustering
- Association Rules
- Neural Networks
- Deep Learning
- etc

Already discussed briefly earlier

ASSOCIATION RULES

Association Rules

- Quite different from Regression and Classification
- Still using data but there is no clear cut prediction that we can derive
- Also different from Clustering in that we are clear in our objectives
- We are interested in mining Association Rules which look like:
  - $X \rightarrow Y$ (X implies Y)
  - E.g. {diaper, milk} $\rightarrow$ {beer}

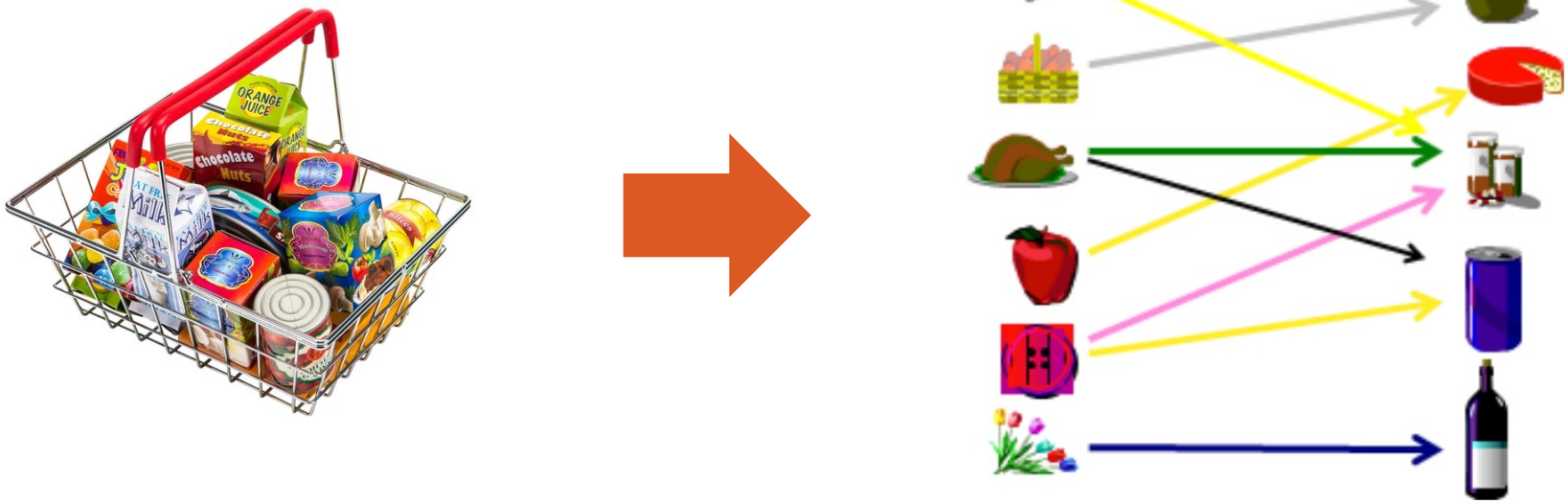
SUPERMARKET PROBLEM

One of the key problems that a supermarket wants to tackle is how to place the products

Place similar category products together for better searching?

MARKET BASKET ANALYSIS

Market Basket Analysis (MBA) aims to find association between groups of items based on the transactions



98% of people who purchased items A and B also purchased C

SUPERMARKET PROBLEM

Probably better to place products that users are likely to buy together in the same location

- Likely to generate more sales that way

EVALUATION


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

STEP 5 – TESTING AND EVALUATION

Evaluate the models developed in step 4 (depending on the problem)

- Regression – how far is the prediction from the actual values
- Classification – classification error rates
- Could also have other evaluation methods for other tasks

We usually divide the labeled data into training and testing data and perform K-Fold Cross Validation

EVALUATION

Regression:

- Mean Square Error (MSE) is typically used

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Classification:

- Accuracy (% correct)

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$

EVALUATION

Classification:

- Other than accuracy/error rate, there are other ways to measure performance of the learning method:
- So far, most of the classification problem we have seen always produce a single class as the classification result.
- However, it is possible in some domains to have multiple values as the classification result

EVALUATION

Classification:

- …
- Information Retrieval (Search Engine context), the classification problem is to retrieve relevant documents (out of a set of documents) based on a given query
  - Classification result should have zero or more documents
  - Since each classification task does not only have one and only one result, using the “accuracy” measure is no longer useful

Documents:
[A, B, C, D, E, F, G]

Given query q_1,
Relevant documents:
[A, B, C]

Given query q_1,
System produce:
[A, B, E, G]

EVALUATION

Classification:

- Information Retrieval: **precision** and **recall** is used instead

$$\text{Precision} = \frac{|\text{relevant document} \cap \text{document retrieved by system}|}{|\text{document retrieved by system}|}$$

$$\text{Recall} = \frac{|\text{relevant document} \cap \text{document retrieved by system}|}{|\text{relevant document}|}$$

Documents:
[A, B, C, D, E, F, G]

Given query q_1,
Relevant documents:
[A, B, C]

Given query q_1,
System produce:
[A, B, E, G]

Precision = 2 / 4
Recall = 2 / 3

EVALUATION

Classification:

- Information Retrieval: **precision** and **recall** is used instead

$$\text{Precision} = \frac{|\text{relevant document} \cap \text{document retrieved by system}|}{|\text{document retrieved by system}|}$$

$$\text{Recall} = \frac{|\text{relevant document} \cap \text{document retrieved by system}|}{|\text{relevant document}|}$$

- **F-measure** is used to factor into both component

$$\text{F} - \text{measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

EVALUATION

Confusion Matrix

		Condition (Expected Result)	
		Positive	Negative
Test outcome (System)	Positive	True Positive (TP)	False Positive (FP) (Type I error)
	Negative	False Negative (FN) (Type II error)	True Negative (TN)

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

DEPLOYMENT


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

STEP 6 – DEPLOYMENT

Development and assessment of model is usually not the end of the project

Depending on the requirements, the deployment phase can be:

- As simple as generating a report
- Or as complex as implementing a system that uses the model for daily operations

Monitoring and maintenance of models

- Over time, the models built may be become obsolete

WHAT'S NEXT?

Mining Web Content I