

# **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)?

## 2021 *This Is What Happens In An Internet Minute*



# AVERAGE TIME SPENT ON MEDIA (CHINA)

JAN  
2021

## DAILY TIME SPENT ON MEDIA

THE AVERAGE DAILY TIME\* THAT INTERNET USERS AGED 16 TO 64 SPEND ON DIFFERENT KINDS OF MEDIA AND DEVICES



CHINA

TIME SPENT USING THE  
INTERNET (ALL DEVICES)



5H 22M

TIME SPENT WATCHING TELEVISION  
(BROADCAST AND STREAMING)



3H 12M

TIME SPENT USING  
SOCIAL MEDIA



2H 04M

TIME SPENT READING PRESS MEDIA  
(ONLINE AND PHYSICAL PRINT)



2H 45M

TIME SPENT LISTENING TO  
MUSIC STREAMING SERVICES



1H 31M

TIME SPENT LISTENING  
TO BROADCAST RADIO



1H 12M

TIME SPENT LISTENING  
TO PODCASTS



1H 15M

TIME SPENT PLAYING VIDEO  
GAMES ON A GAMES CONSOLE



1H 21M

22

**SOURCE:** GWI (Q3 2020). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SURVEY OF INTERNET USERS AGED 16 TO 64. SEE [GLOBALWEBINDEX.COM](https://www.globalwebindex.com) FOR MORE DETAILS.  
**\*NOTES:** CONSUMPTION OF DIFFERENT MEDIA MAY OCCUR CONCURRENTLY. TELEVISION INCLUDES BROADCAST (LINEAR) TELEVISION AND CONTENT DELIVERED VIA STREAMING AND VIDEO-ON-DEMAND SERVICES. PRESS INCLUDES ONLINE AS WELL AS PHYSICAL PRINT MEDIA. BROADCAST RADIO DOES NOT INCLUDE INTERNET RADIO.

we  
are  
social



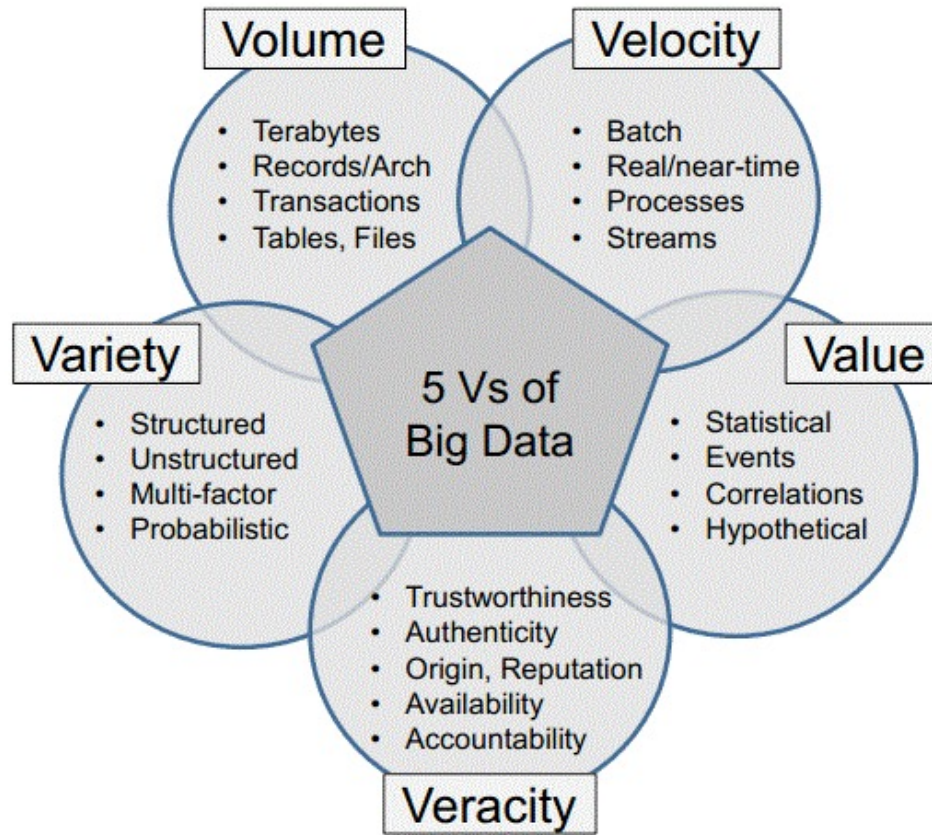
Hootsuite®

# 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?



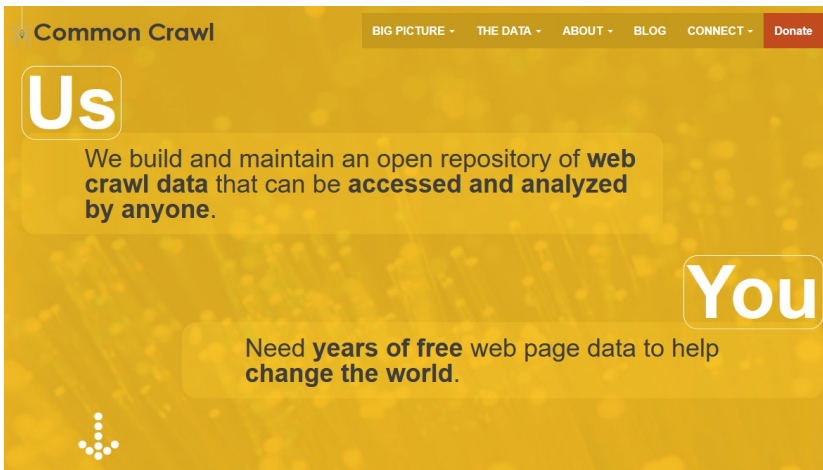
The 5Vs  
of  
Big Data



# WHY “BIG DATA”?

## Many factors that contribute to this concept

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



**Common Crawl**

**Us**

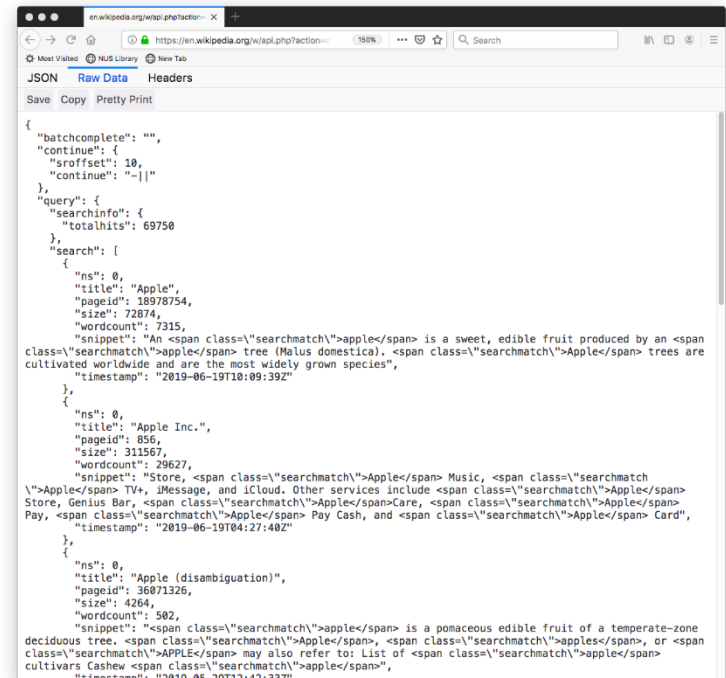
We build and maintain an open repository of **web crawl data** that can be **accessed and analyzed** by anyone.

**You**

Need **years of free** web page data to help **change the world**.



Facebook Graph API



```
{
  "batchcomplete": "",
  "continue": {
    "sroffset": 10,
    "continue": "--||"
  },
  "query": {
    "searchinfo": {
      "totalhits": 69750
    },
    "search": [
      {
        "ns": 0,
        "title": "Apple",
        "pageid": 18978754,
        "size": 72874,
        "wordcount": 7315,
        "snippet": "An <span class="searchmatch" apple</span> is a sweet, edible fruit produced by an <span class="searchmatch" apple</span> tree (Malus domestica). <span class="searchmatch" Apple</span> trees are cultivated worldwide and are the most widely grown species",
        "timestamp": "2019-06-19T10:09:39Z"
      },
      {
        "ns": 0,
        "title": "Apple Inc.",
        "pageid": 856,
        "size": 311567,
        "wordcount": 29627,
        "snippet": "Store, <span class="searchmatch" Apple</span> Music, <span class="searchmatch" Apple</span> TV, <span class="searchmatch" Apple</span> iMessage, and iCloud. Other services include <span class="searchmatch" Apple</span> Store, <span class="searchmatch" Apple</span> Care, <span class="searchmatch" Apple</span> Pay, <span class="searchmatch" Apple</span> Cash, and <span class="searchmatch" Apple</span> Card",
        "timestamp": "2019-06-19T04:27:40Z"
      },
      {
        "ns": 0,
        "title": "Apple (disambiguation)",
        "pageid": 36071326,
        "size": 4264,
        "wordcount": 502,
        "snippet": "<span class="searchmatch" apple</span> is a pomaceous edible fruit of a temperate-zone deciduous tree. <span class="searchmatch" Apple</span>, <span class="searchmatch" apples</span>, or <span class="searchmatch" Apple</span> may also refer to: List of <span class="searchmatch" apple</span> cultivars <span class="searchmatch" Apple</span> Pay Cash, and <span class="searchmatch" Apple</span> Card",
        "timestamp": "2019-05-29T12:42:33Z"
      }
    ]
  }
}
```



# BIG DATA ANALYTICS

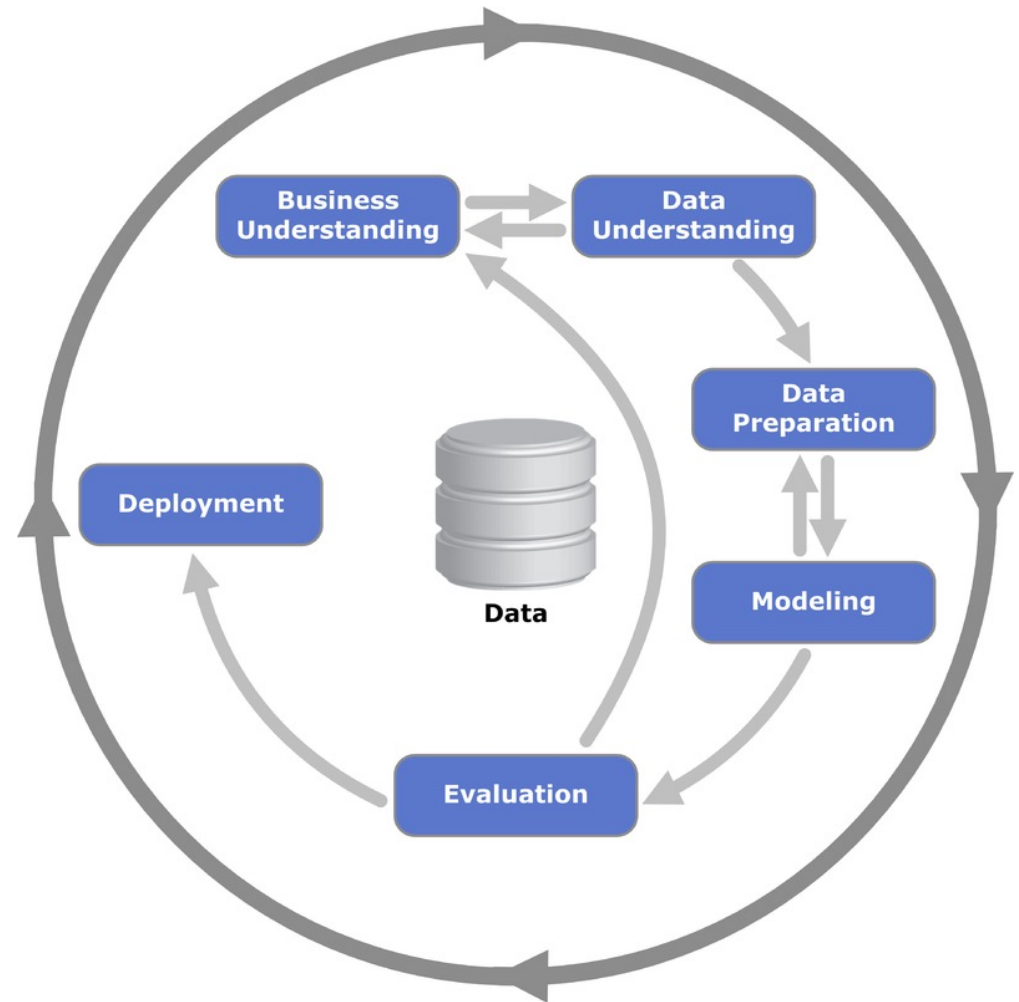
**Big Data by itself, regardless of the size, type, or speed, is worthless**

**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)**

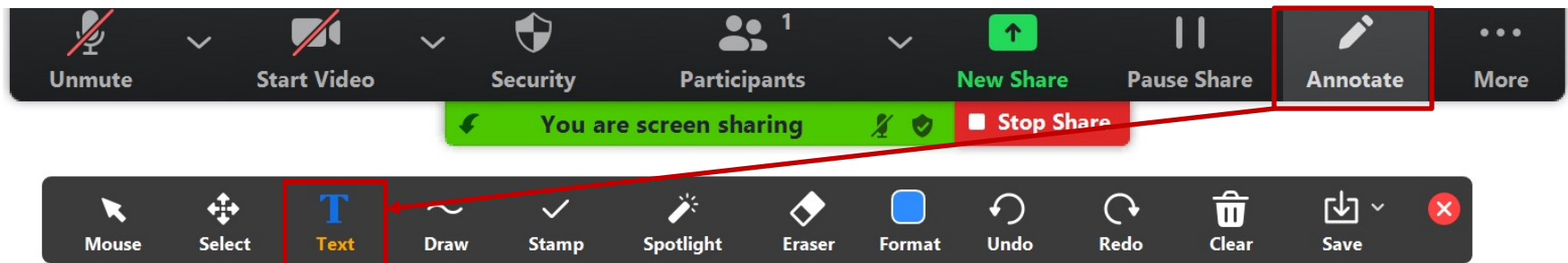
- It does NOT start with the data!

# 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

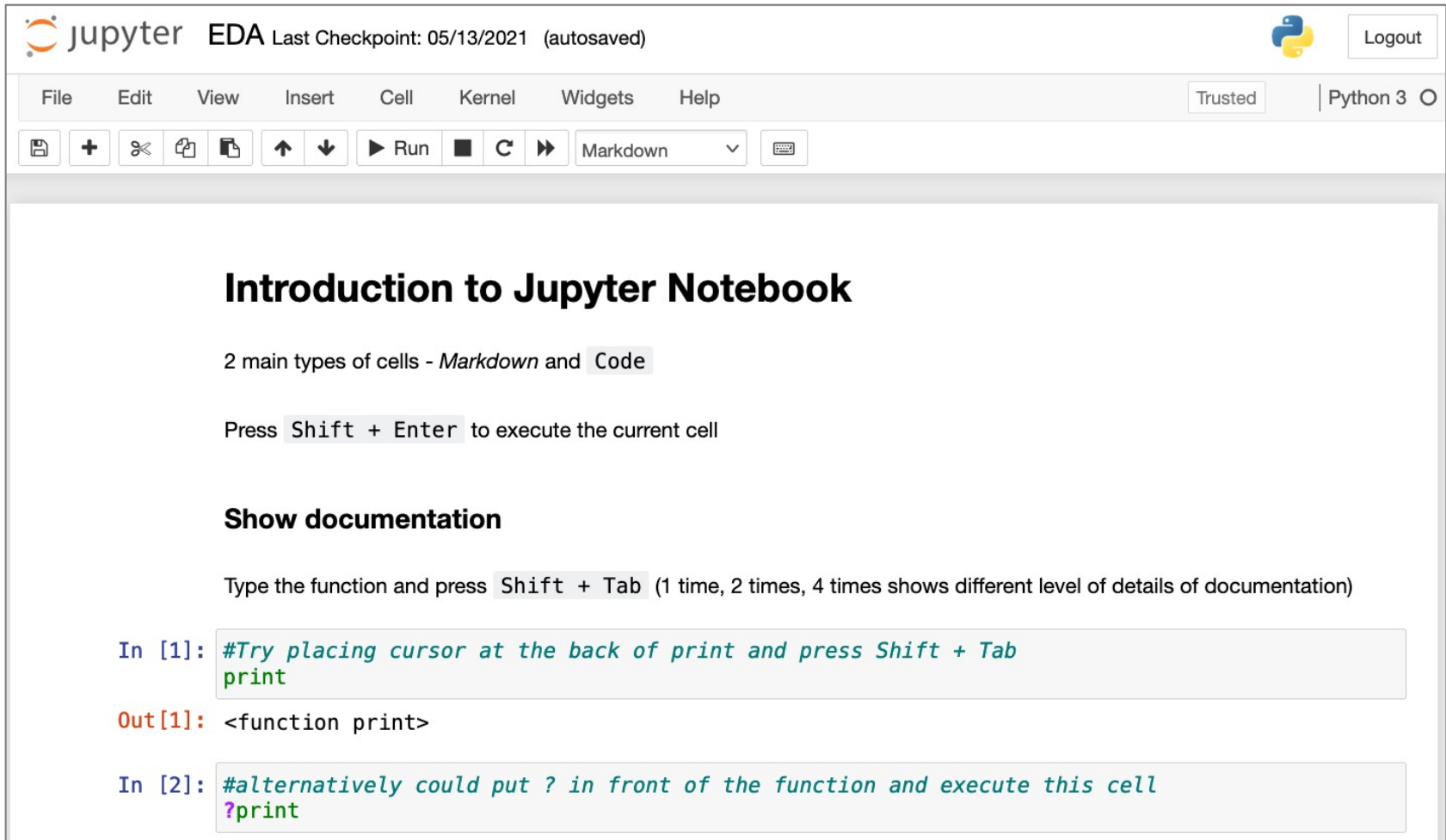


# STEP 2 – DATA UNDERSTANDING

**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)



The screenshot shows a Jupyter Notebook interface with the title "EDA" and a last checkpoint of "05/13/2021 (autosaved)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for saving, adding, deleting, and running cells. The main content area displays the following text:

## Introduction to Jupyter Notebook

2 main types of cells - *Markdown* and *Code*

Press **Shift + Enter** to execute the current cell

### Show documentation

Type the function and press **Shift + Tab** (1 time, 2 times, 4 times shows different level of details of documentation)

```
In [1]: #Try placing cursor at the back of print and press Shift + Tab
print
```

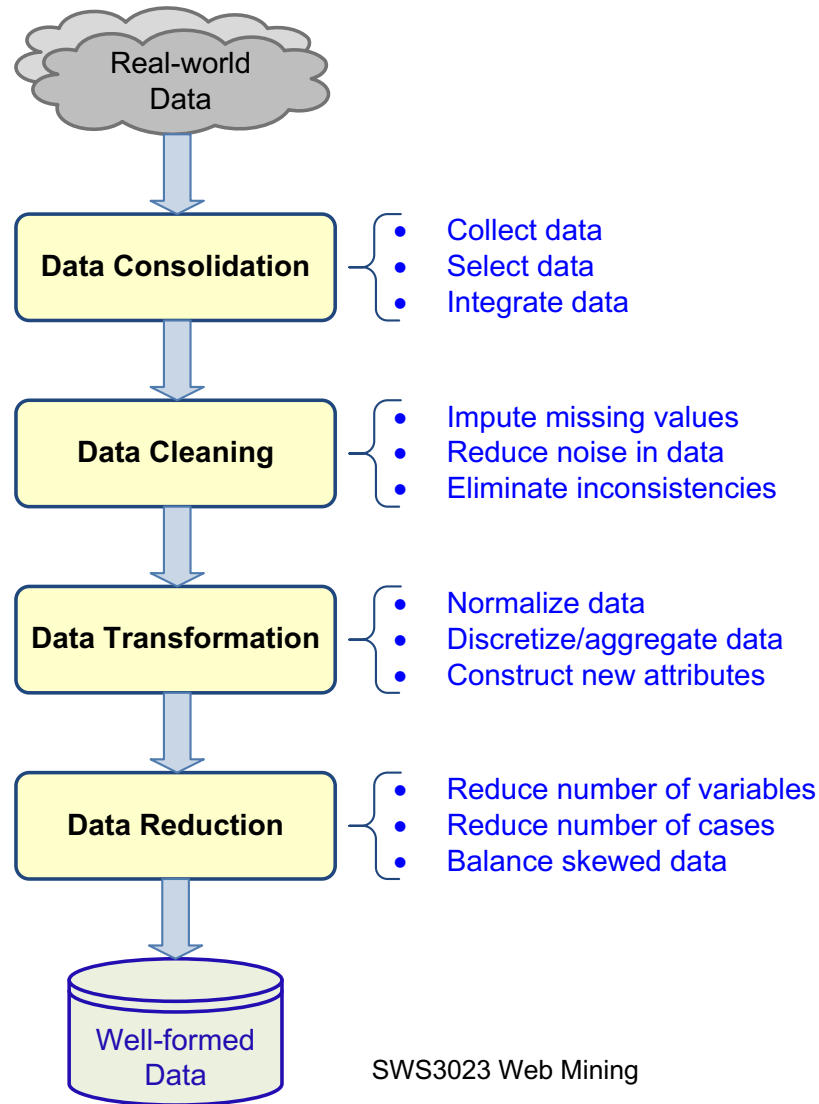
```
Out[1]: <function print>
```

```
In [2]: #alternatively could put ? in front of the function and execute this cell
?print
```

# 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:**

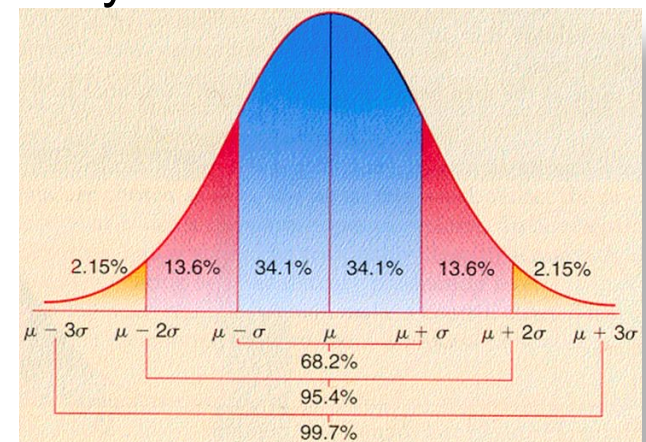
- One way to detect outliers is based on the concept of **statistical dispersion**

# DETECTING OUTLIERS (3-SIGMA)

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

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

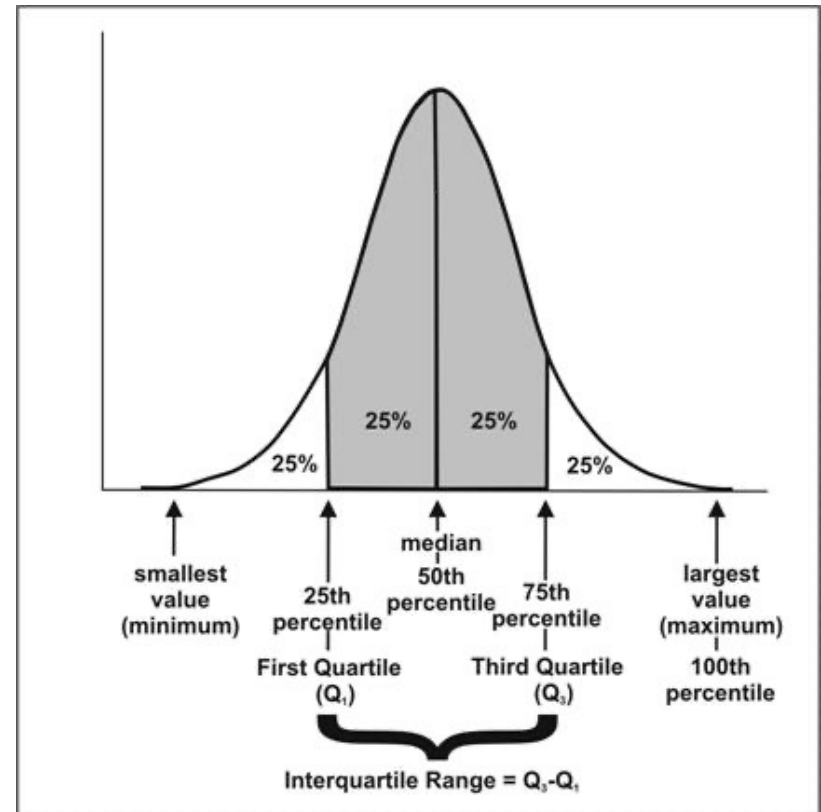
Values that falls outside  $(\bar{\mu} \pm 3\bar{\sigma})$  can be considered as suspicious outliers.



# DETECTING OUTLIERS (BOXPLOT)

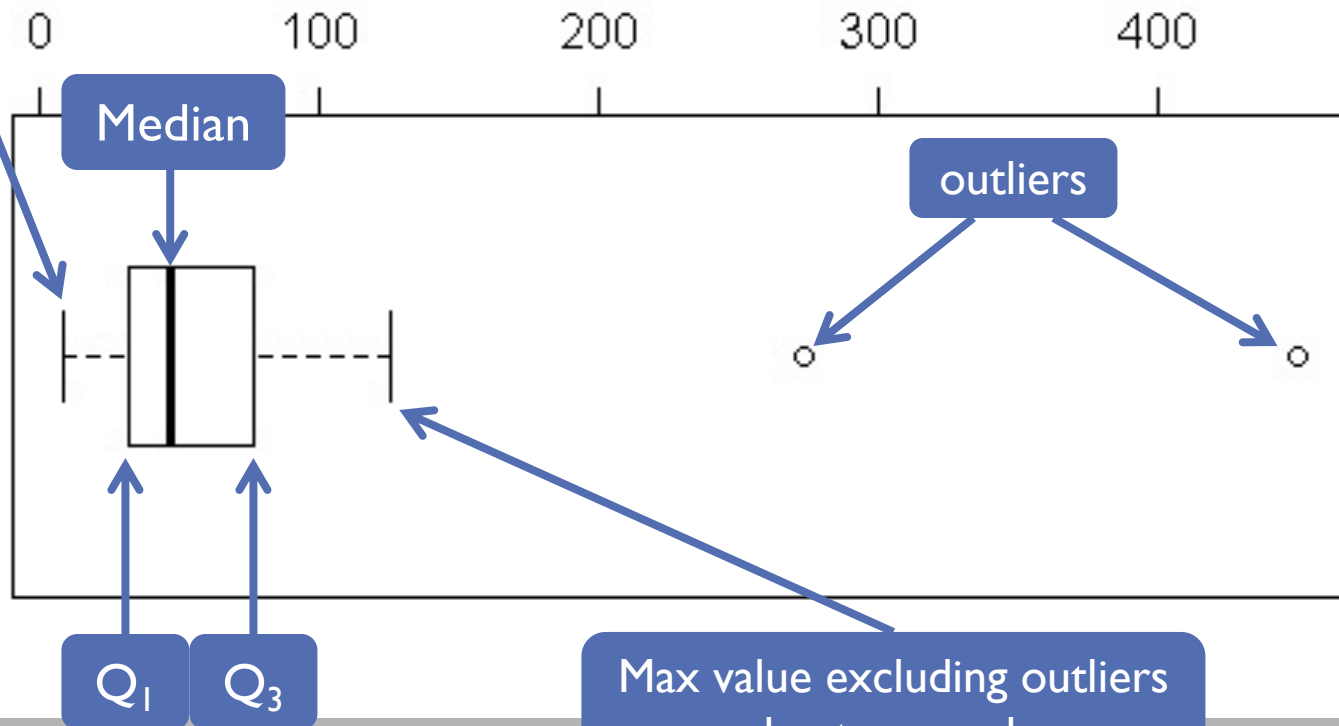
## Interquartile Range (IQR)

- 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)

Min value excluding outliers  
and extreme values



Max value excluding outliers  
and extreme values

# DETECTING OUTLIERS (BOXPLOT)

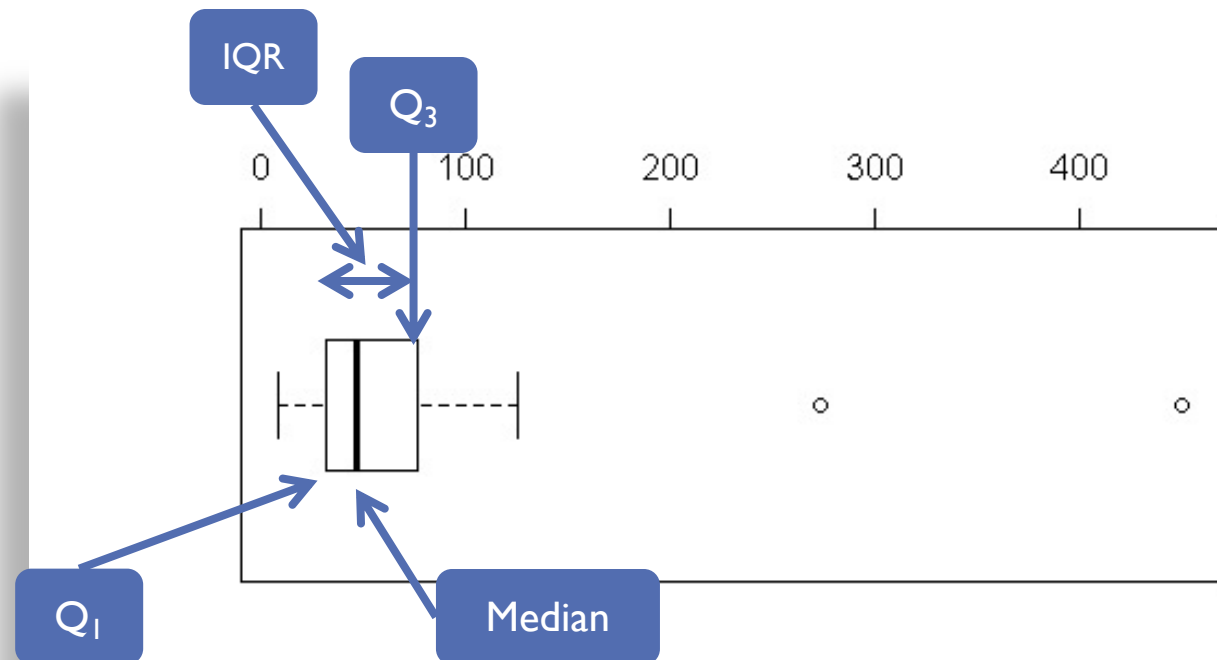
Data: [45, 450, 20, 69, 9, 66, 11, 42, 9, 126, 47, 43, 24, 94, 89, 16, 83, 59, 57, 273, 70, 45, 40]

Median = 47

Q1 = 32

Q3 = 76.5

IQR = 44.5





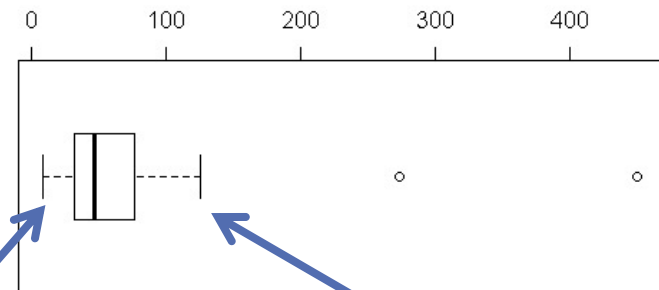
# DETECTING OUTLIERS (BOXPLOT)

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:

- $1.5 \times \text{IQR} = 66.75$
- $QL - 1.5 \times \text{IQR} = -34.75$
- $QU + 1.5 \times \text{IQR} = 143.25$
- Min = 9
- Max = 126



Min value excluding outliers  
and extreme values

Max value excluding outliers  
and extreme values

# DETECTING OUTLIERS (BOXPLOT)

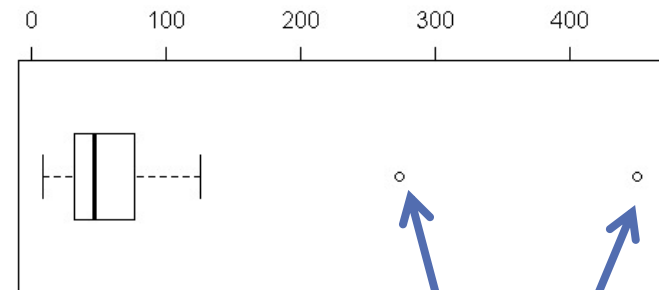
Data: [45, 450, 20, 69, 9, 66, 11, 42, 9, 126, 47, 43, 24, 94, 89, 16, 83, 59, 57, 273, 70, 45, 40]

## Outliers

- $< Q1 - 1.5 \cdot IQR$
- $> Q3 + 1.5 \cdot IQR$

## Extreme values

- $< Q1 - 3 \cdot IQR$
- $> Q3 + 3 \cdot IQR$



Outliers / Extreme values

# 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
<p>12 <b>Predictive Analytics I</b> (10-12pm)</p> <p><b>1<sup>st</sup> consultation</b> (1-6pm)</p>	<p>13 <b>Predictive Analytics II</b> (10-12pm)</p> <p><b>Lab 1 / 1<sup>st</sup> consultation</b> (1-6pm)</p>	<p>14 <b>Mining Web Content II</b> (10-12pm)</p> <p><b>Lab 2 / Ad hoc consultation</b> (1-6pm)</p>	<p>15 <b>Machine Learning</b> (10-12pm)</p> <p><b>Lab 3 / Ad hoc consultation</b> (1-6pm)</p>	<p>16 <b>Machine Learning</b> (10-12pm)</p> <p><b>Ad hoc consultation</b> (1-6pm)</p>	
<p>19 <b>Lab 5 / 3<sup>rd</sup> consultation</b> (10am-6pm)</p>	<p>20 <b>Lab 6 / 3<sup>rd</sup> consultation</b> (10am-6pm)</p>	<p>21 <b>Lab 7 / Ad hoc consultation</b></p> <p><b>Ad hoc help from TA</b> (10am-6pm)</p>	<p>22 <b>4<sup>th</sup> consultation</b></p> <p><b>Ad hoc help from TA</b> (10am-6pm)</p>	<p>23 <b>4<sup>th</sup> Consultation</b></p> <p><b>Ad hoc help from TA</b> (10am-6pm)</p>	<p>24 <b>Ad hoc consultation</b> (10-6pm)</p>
<p>26 <b>Ad hoc consultation</b></p> <p><b>Ad hoc help from TA</b> (10am-6pm)</p>	<p>27 <b>Ad hoc consultation</b></p> <p><b>Ad hoc help from TA</b> (10am-6pm)</p>	<p>28 <b>Project Showcase</b> (12-6pm)</p>	<p>29</p>	<p>26</p>	<p>26</p>

Will learn more about data manipulation in Lab 1 & 2

# MODELING



# 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

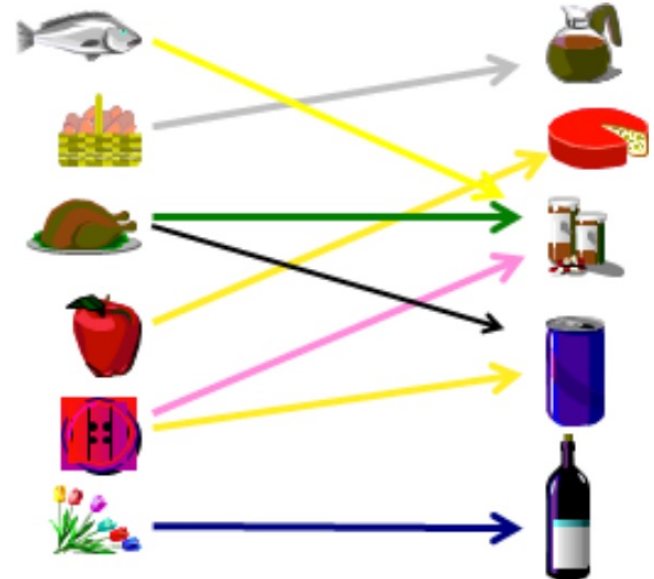


**One of the key problems that a supermarket wants to tackle is how to place the products**

**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



# 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

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

# 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)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

# 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 become obsolete

# WHAT'S NEXT?

## Mining Web Content I