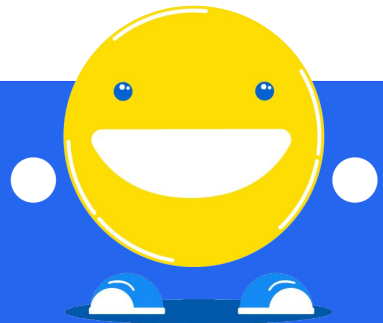
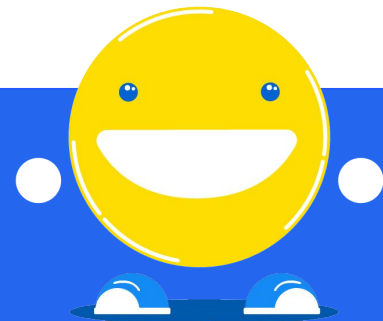


Machine Learning Workshop



2019



5

Introduction to Data Science (2 hours)

5

Data Preprocessing and Cleansing (4 hours)

6

Supervised and Unsupervised Model (4 hours)

6

Project Preparation and Explanation (2 hours)

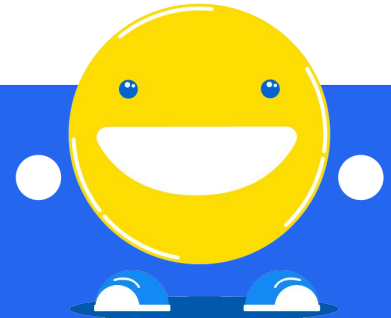
9

Project Presentation (8 hours)

Introduction to Data Science



2019

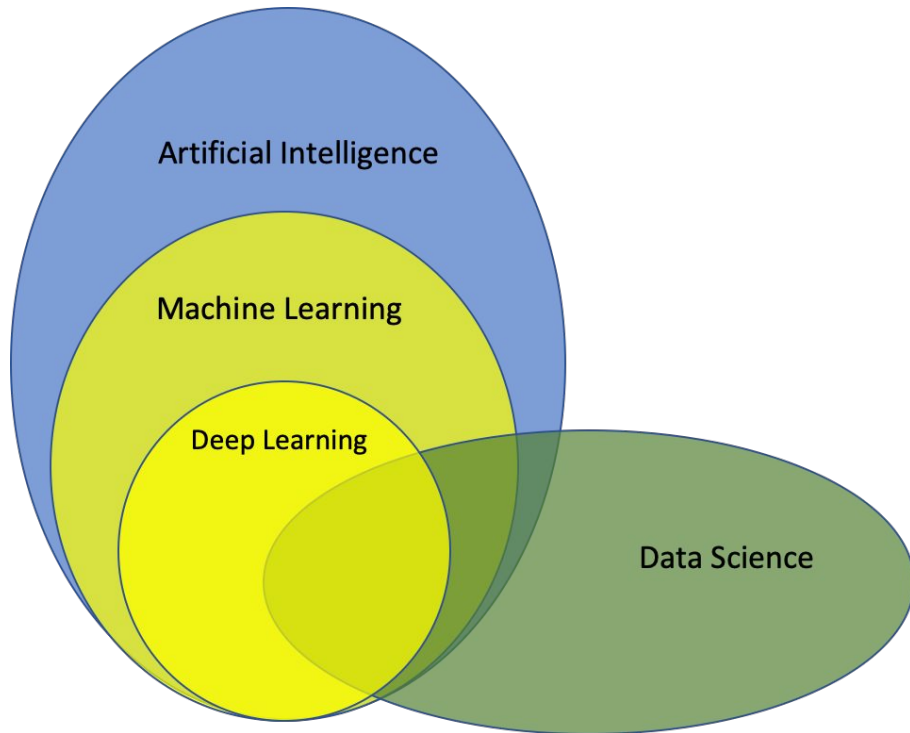


What is data science

Data science is a multi-disciplinary field that uses **scientific methods**, processes, algorithms and systems to **extract knowledge and insights from structured and unstructured data**.

Data Science = statistics + computational technology advancement

Artificial Intelligence? Machine Learning? Data Science? Deep Learning? What do those buzzwords mean?



Artificial Intelligence is the **simulation of human intelligence** processes by machines, especially computer systems.

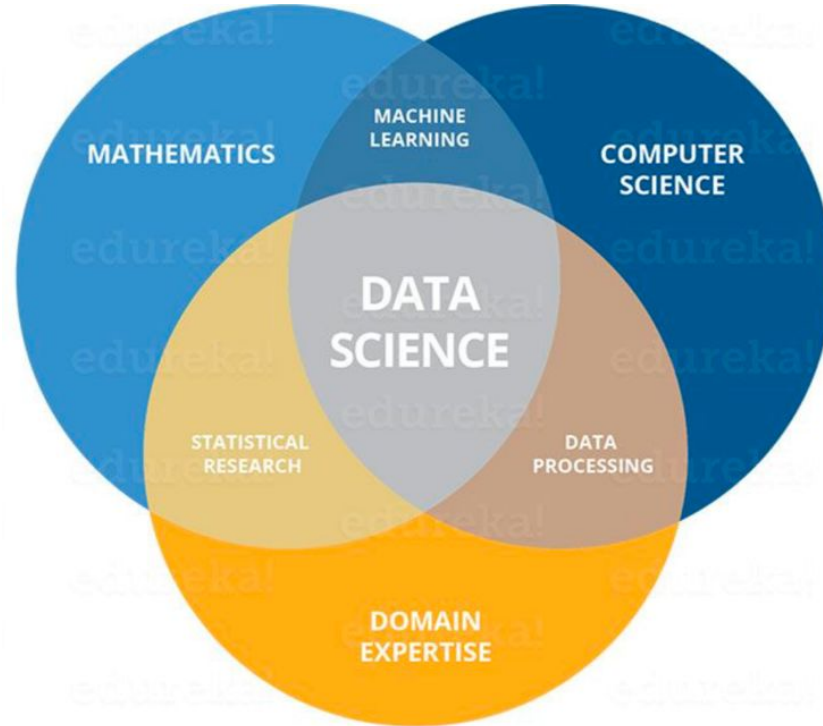
Example: calculator and automated gate

Machine Learning is an application of **artificial intelligence** (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

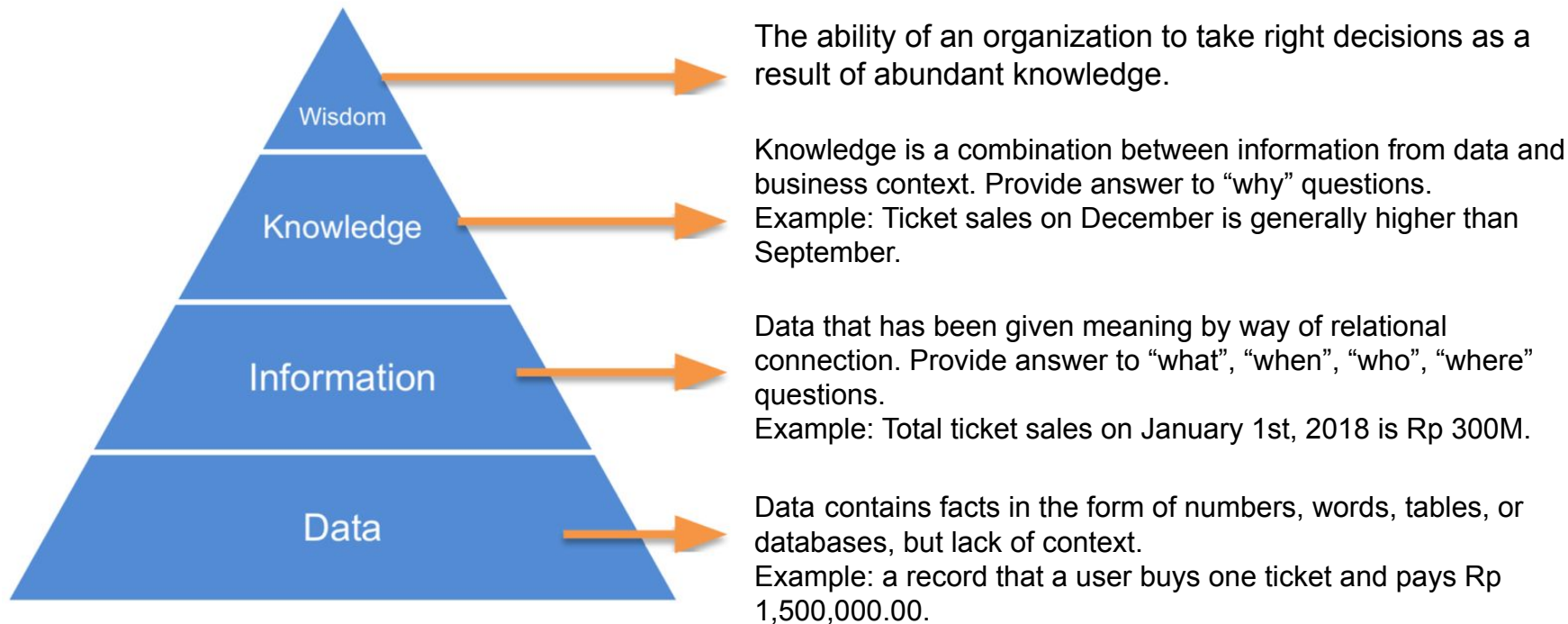
Deep Learning is a set of **machine learning** methods that are based on artificial **neural networks**.

Data Science combines methods and algorithms that are part of artificial intelligence, machine learning, and even deep learning, with other subjects such as statistics and business knowledge.

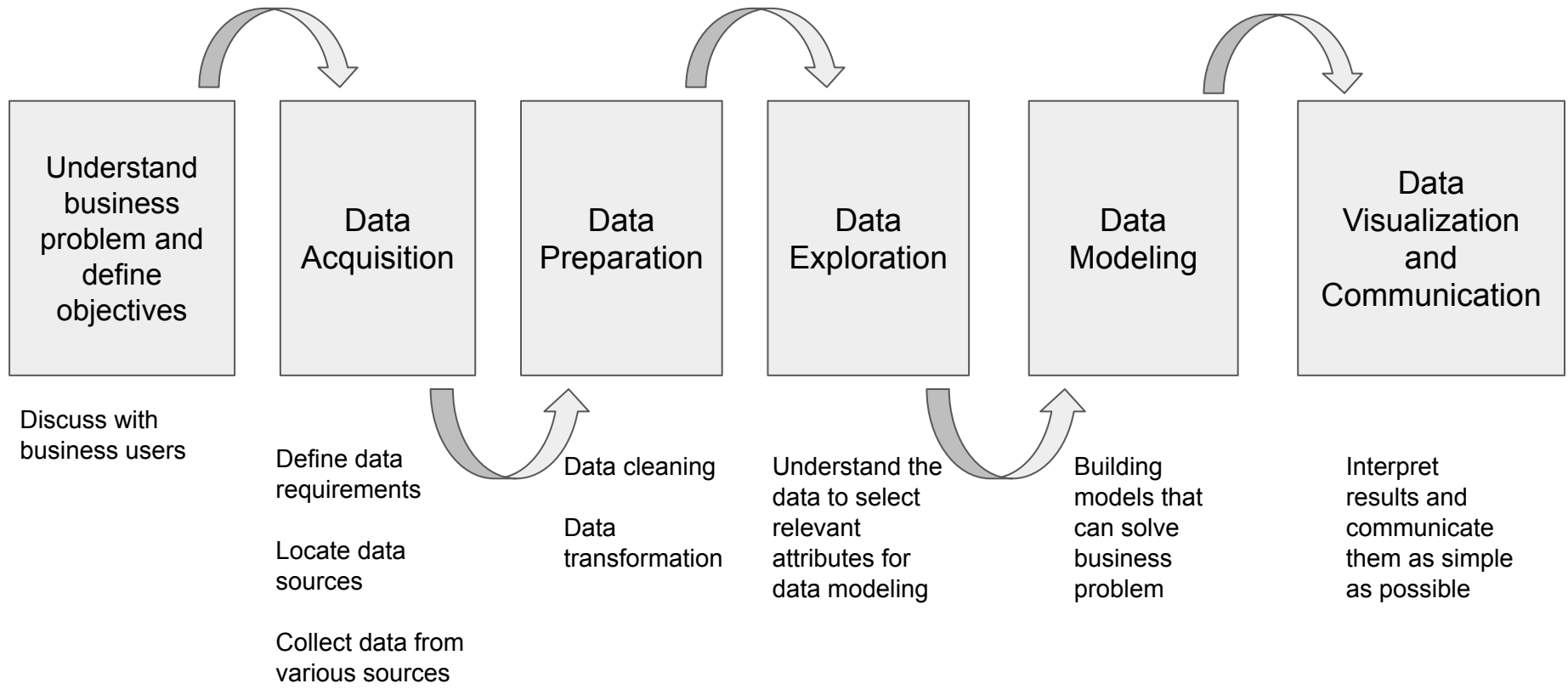
Data science is a combination of concepts from computer science, mathematics, and domain expertise.



Data scientist's job is to convert data into actionable insights that brings values for their organization.



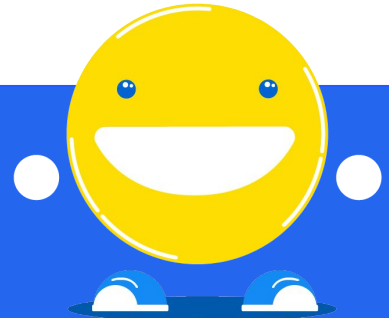
Data Science Lifecycle



Basic Statistics



2019



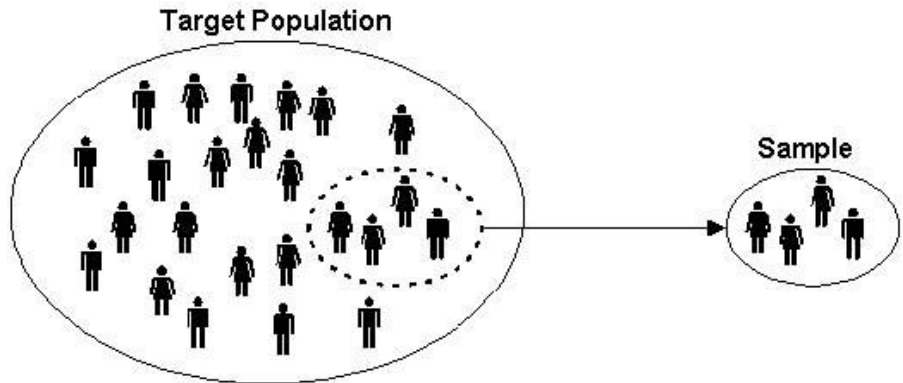
Sample vs Population

Population contains all members of a specified group (the entire list of possible data values) that we want to investigate

example: all Indonesian citizens.

Sample dataset contains a part or a subset of a population.

Example: people in the room, tiket employees, Binus students



Population

Population size depends on the scope of analysis that we want to do.

example:

- National census is against all Indonesian citizens
- Employee happiness measurement is only against all employees of a company.
- All product transaction in tiket.com

It is usually very difficult, if not possible, to collect data of all members in a population. Therefore, we use samples to draw conclusion about the population.

Sample

Samples must be **RANDOM** and **REPRESENTATIVE**.

RANDOM:

A random sample is collected when each member is chosen by chance. Each member of the population is equally likely to be chosen.

REPRESENTATIVE:

The sample accurately reflects the member of the entire population. No group of members are left out.

We have to make sure that our samples are **NOT bias** toward certain attributes.

Descriptive and Inferential Statistics

Descriptive statistics are summary statistics that quantitatively describe or summarizes features of the samples that have been collected.

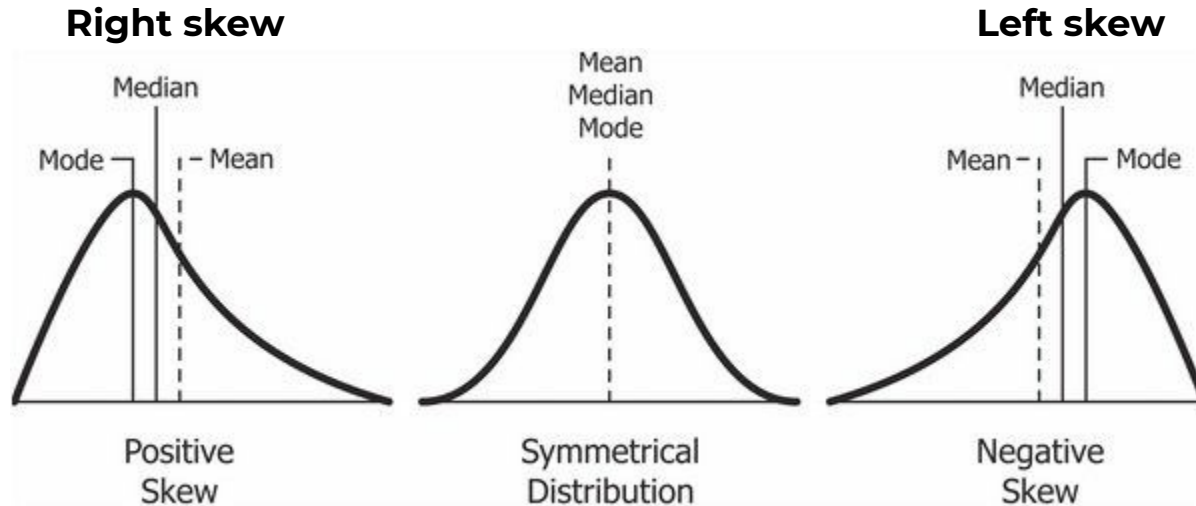
Inferential statistics are statistical methods to make inferences and predictions about a population based on a sample of data taken from the population in question.

Central Tendency

A measure of central tendency is a value that describes a set of data by identifying the central position within the set of data.

1. Mean (or average) = sum of all values over the number of values.
2. Median = middle value of an ordered sample of numerical values.
3. Mode = Value that occurs most frequently.

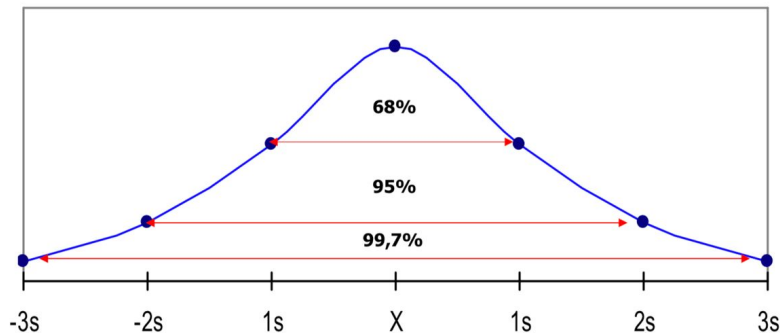
Skewness



Positive skew is when the tail on the right side of the distribution is longer or fatter.

Negative skew is when the tail on the left side of the distribution is longer or fatter.

Normal Distribution



An attribute has a normal distribution when its mean is the same as its median and mode.

When an attribute is normally distributed:

- 68% of its values are between 1 standard deviation from the mean.
- 95% of its values are between 2 times standard deviation from the mean.
- 99.7% of its values are between 3 times standard deviation from the mean.

Normal distribution is the most common distribution in nature. A random event is most likely to follow this distribution. See video.

Quartiles

When we put the values of an attribute in ascending order and divide them into 4 groups with equal number of values, we will get quartiles. Each quartile contains 25% of the data.

Suppose we collect age from 16 students as follows:

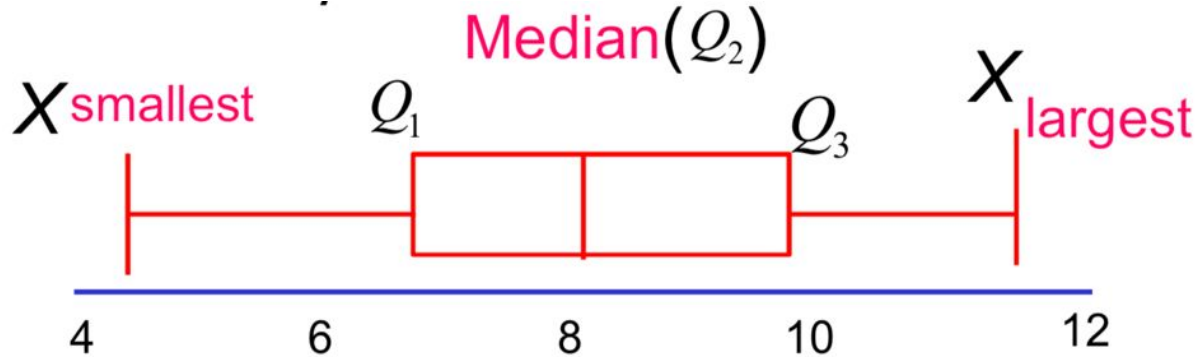
18, 19, 20, **21**, 21, 21, 22, **22**, 22, 22, 22, **23**, 23, 24, 24, **24**

$Q1 = 21$; $Q2 = 22$, $Q3 = 23$, $Q4 = 24$

In the example above, the first quartile is 21 years old. This means that 25% of the students have age less than 21 years old. Second quartile is 22 years old, meaning that 50% of the students have age less than 23 years old. Second quartile is the same as median.

Interquartile Range (IQR)

Quartiles can be visualized using a boxplot as shown below.

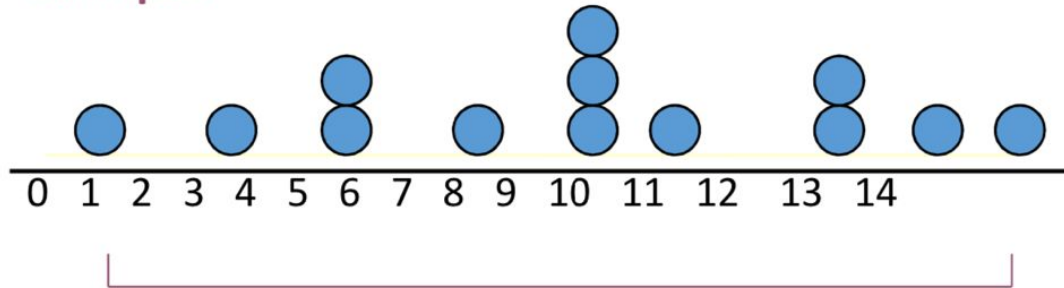


$$\text{IQR} = Q_3 - Q_1$$

Range

Range is the simplest data variance measure. It is the difference between maximum value and minimum value of an attribute.

Example:



$$\text{Range} = 14 - 1 = 13$$

Variance and Standard Deviation

Variance is the average of squared differences from the mean.

Formula of variance =

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

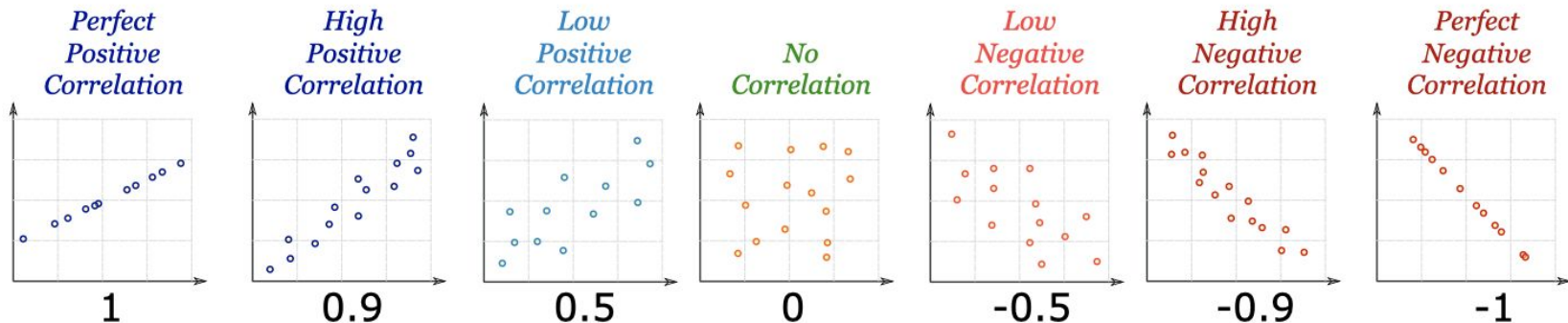
Standard deviation is square root of the variance. It is the most commonly used measure of spread.

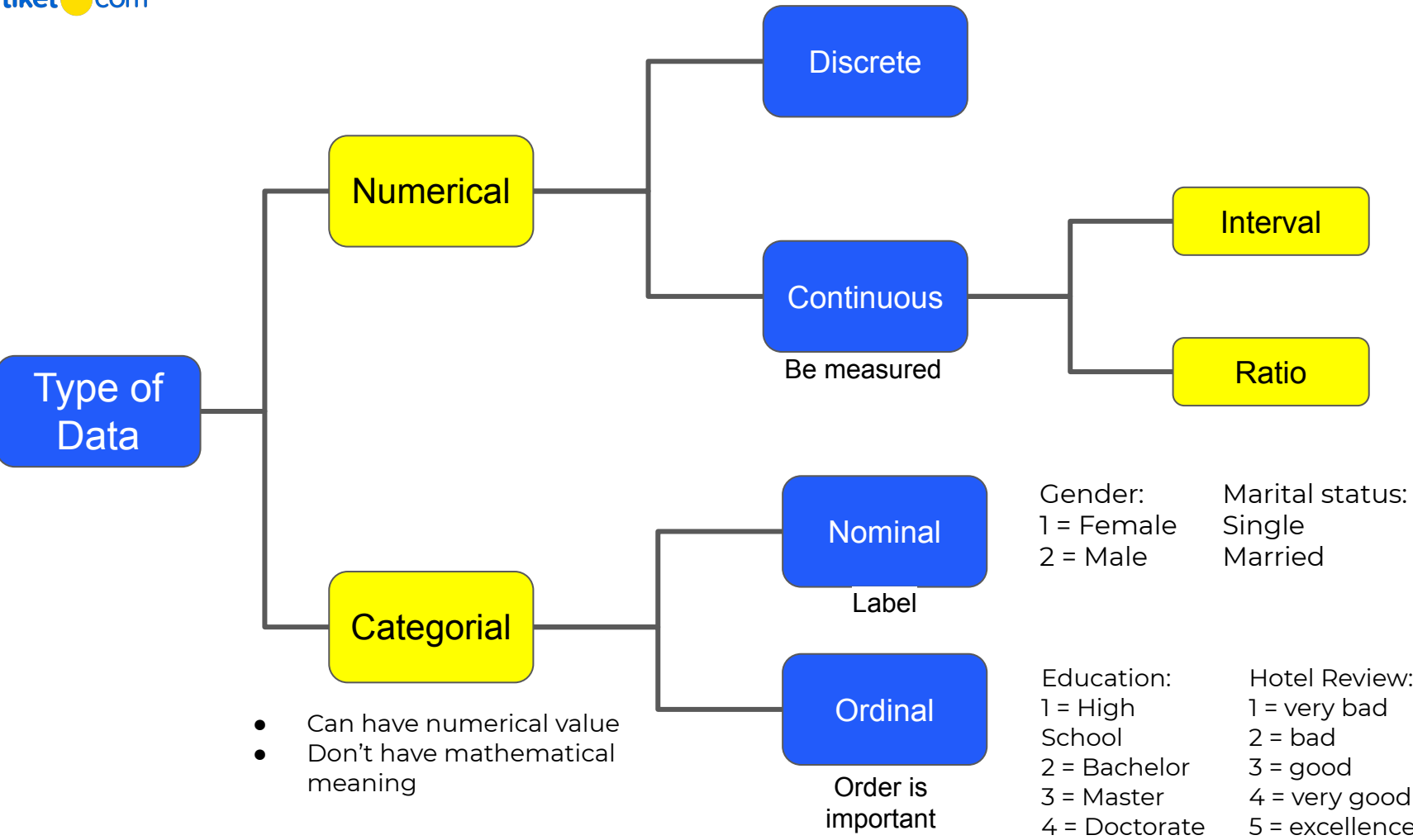
Formula of variance =

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Correlation

Correlation measures the strength of association between two variables and the direction of the relationship. The value of a correlation coefficient varies between +1 and -1.

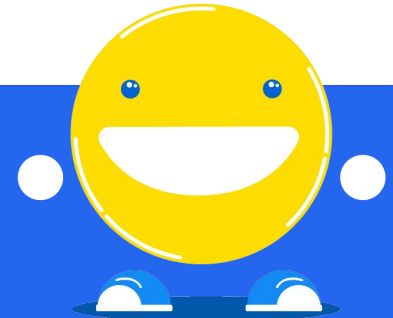




Data Pre-processing and Cleaning



2019



Do you trust your data?

If your business is a house, then data is its foundation.

Before
Data Cleansing
Data quality is 40%



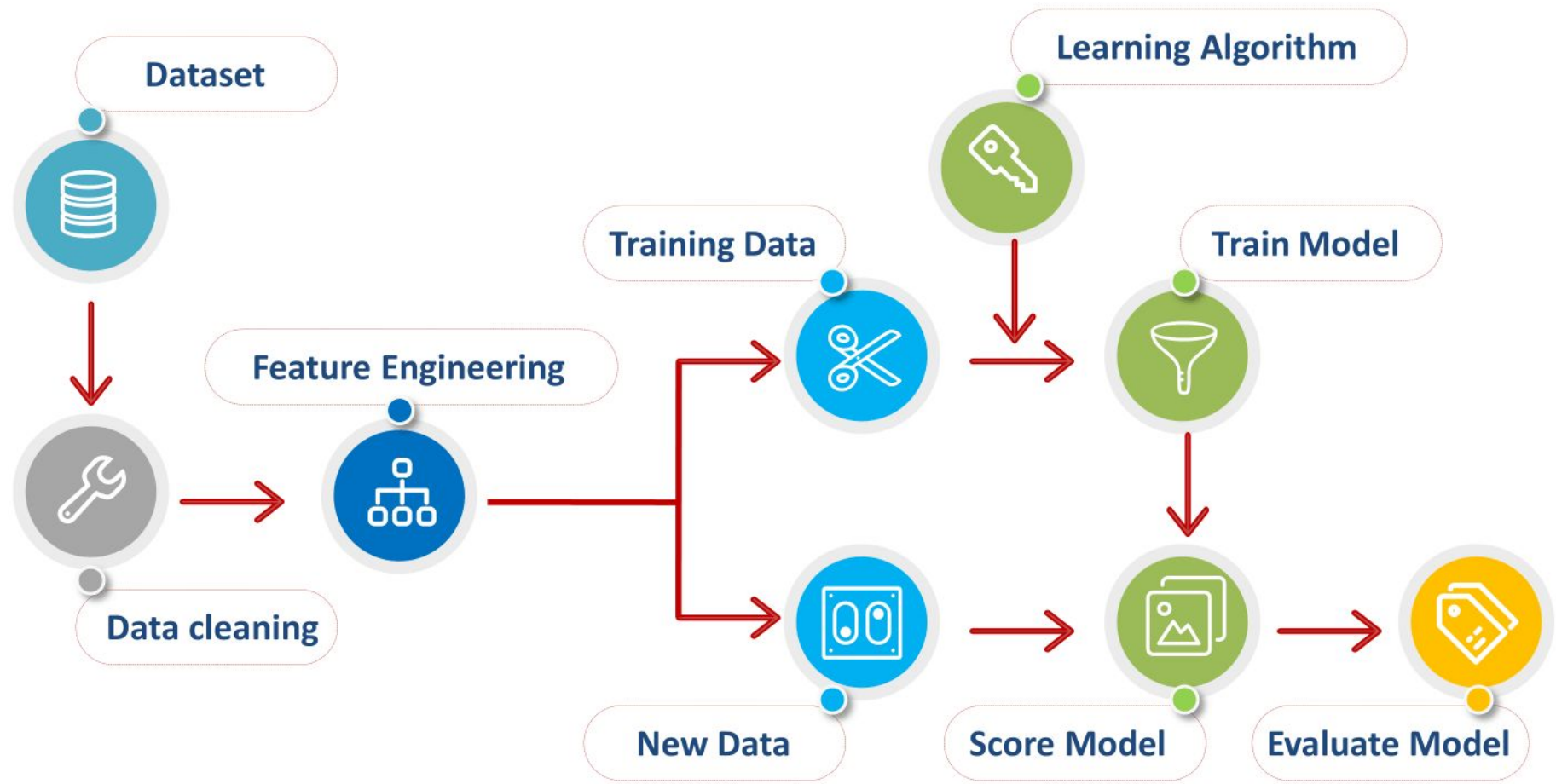
After
Data Cleansing
Data quality is 90%



Why We Need To Pre Processing and Clean The Data?

Real world data is **Dirty and Messy** and these might lead into false conclusion if we used it straightly. These are some of the dirts you can find in the data

- Incomplete Data : Lacks certain attributes values, some missing data. For Example, not all the values on occupation are filled
- Noisy : Contains noise, errors or outliers. For Example, age = -1
- Inconsistent : containing discrepancies. For Example, rating have values 10 and A
- Intentional : disguised missing data. For Example 1 Jan 1970 as birthdate



1

Missing Data

- Ignore
- Fill Manually
- Fill Computed Value

2

Noisy Data

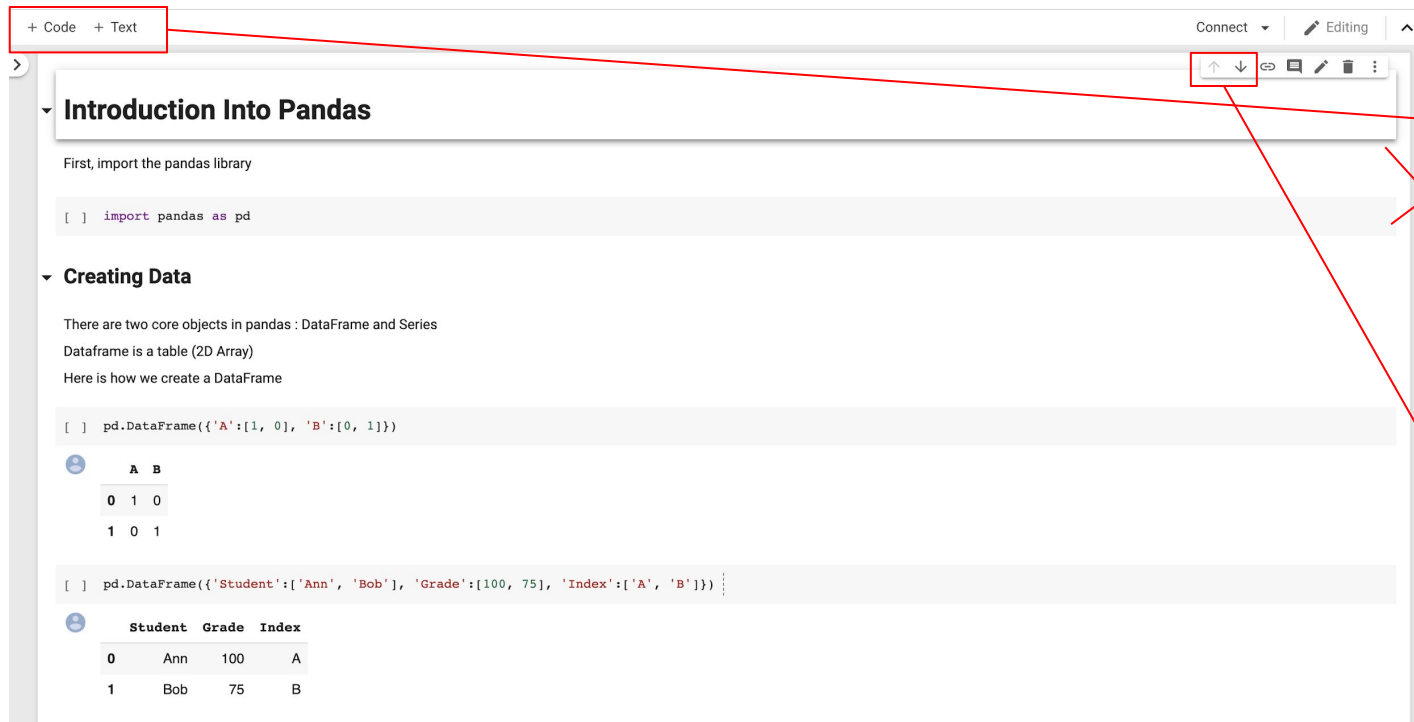
- Binning
- Clustering
- Machine Learning Algorithm
- Remove Manually

3

Inconsistent Data

- External References
- Knowledge Engineering Tools

Intro Google colabs



The screenshot shows the Google Colab interface. At the top left, a red box highlights the '+ Code' and '+ Text' buttons. A red arrow points from this box to the top right of the code editor, where another red box highlights the 'Move up' and 'Move down' icons. A red arrow points from the 'Move up' icon to the text 'Code : for adding code section'. Another red arrow points from the 'Move down' icon to the text 'Text : for adding Markdown section'. A third red arrow points from the 'Move up' icon to the text 'Move current section to go up or down from current position'.

Code Section:

```
[ ] import pandas as pd
```

Text Section:

First, import the pandas library

Creating Data

There are two core objects in pandas : DataFrame and Series
 Dataframe is a table (2D Array)
 Here is how we create a DataFrame

```
[ ] pd.DataFrame({'A':[1, 0], 'B':[0, 1]})
```

| | A | B |
|---|---|---|
| 0 | 1 | 0 |
| 1 | 0 | 1 |

```
[ ] pd.DataFrame({'Student':['Ann', 'Bob'], 'Grade':[100, 75], 'Index':['A', 'B']})
```

| | Student | Grade | Index |
|---|---------|-------|-------|
| 0 | Ann | 100 | A |
| 1 | Bob | 75 | B |

Code : for adding code section

Text : for adding Markdown section

Move current section to go up or down from current position

Intro Google colabs

There are two core objects in pandas : DataFrame and Series

Dataframe is a table (2D Array)

Here is how we create a DataFrame

```
pd.DataFrame({'A':[1, 0], 'B':[0, 1]})
```

| | A | B |
|---|---|---|
| 0 | 1 | 0 |
| 1 | 0 | 1 |

Run : execute the current selected Code

```
[ ] pd.DataFrame({'Student':['Ann', 'Bob'], 'Grade':[100, 75], 'Index':['A', 'B']})
```

| | Student | Grade | Index |
|---|---------|-------|-------|
| 0 | Ann | 100 | A |
| 1 | Bob | 75 | B |

Show the result for the current code that been execute

We use pd.DataFrame to generate these DataFrame objects.

The list of row labels used in a DataFrame is known as Index. For the above example, the index are 0, 1. We can assign values to index by using index parameter when we construct a list

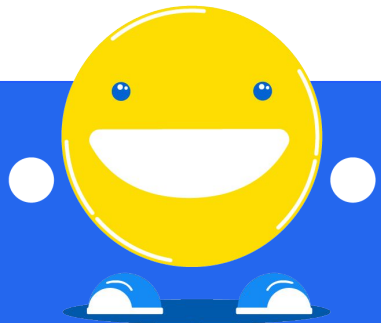
```
[ ] pd.DataFrame({'Student':['Ann', 'Bob'],
                  'Grade':[100, 75], 'Index':['A', 'B'],
                  index= ['Class 1-A', 'Class 1-B']
                  })
```

| | Student | Grade | Index |
|-----------|---------|-------|-------|
| Class 1-A | Ann | 100 | A |

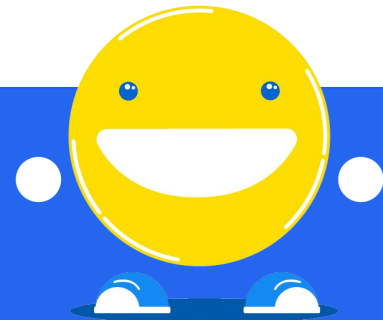
Delete the current selected section

Let's Hands-on

Intro into Machine Learning



2019



What is Learning?

Learning is a process by which the learner **improves** its **performance** on a **task** or a **set of tasks** as a result of **experience** within some **environment**

Learning = Inference + Memorization

Inference = Deduction, Induction, Abduction

What is Machine Learning?

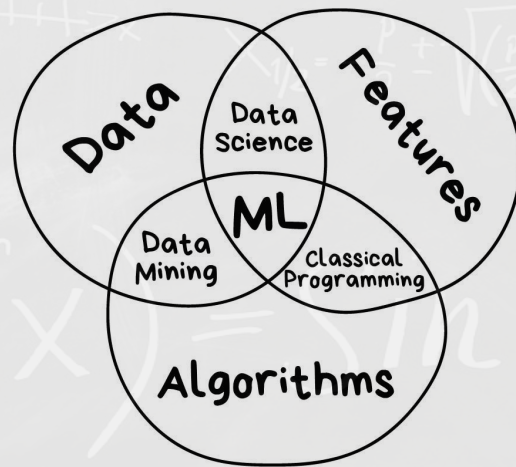
A computer program **M** is said to learn from experience **E** with respect to some class of tasks **T** and performance **P**, if its performance as measured by **P** on tasks in **T** in an environment **Z** improves with experience **E**.

Example:

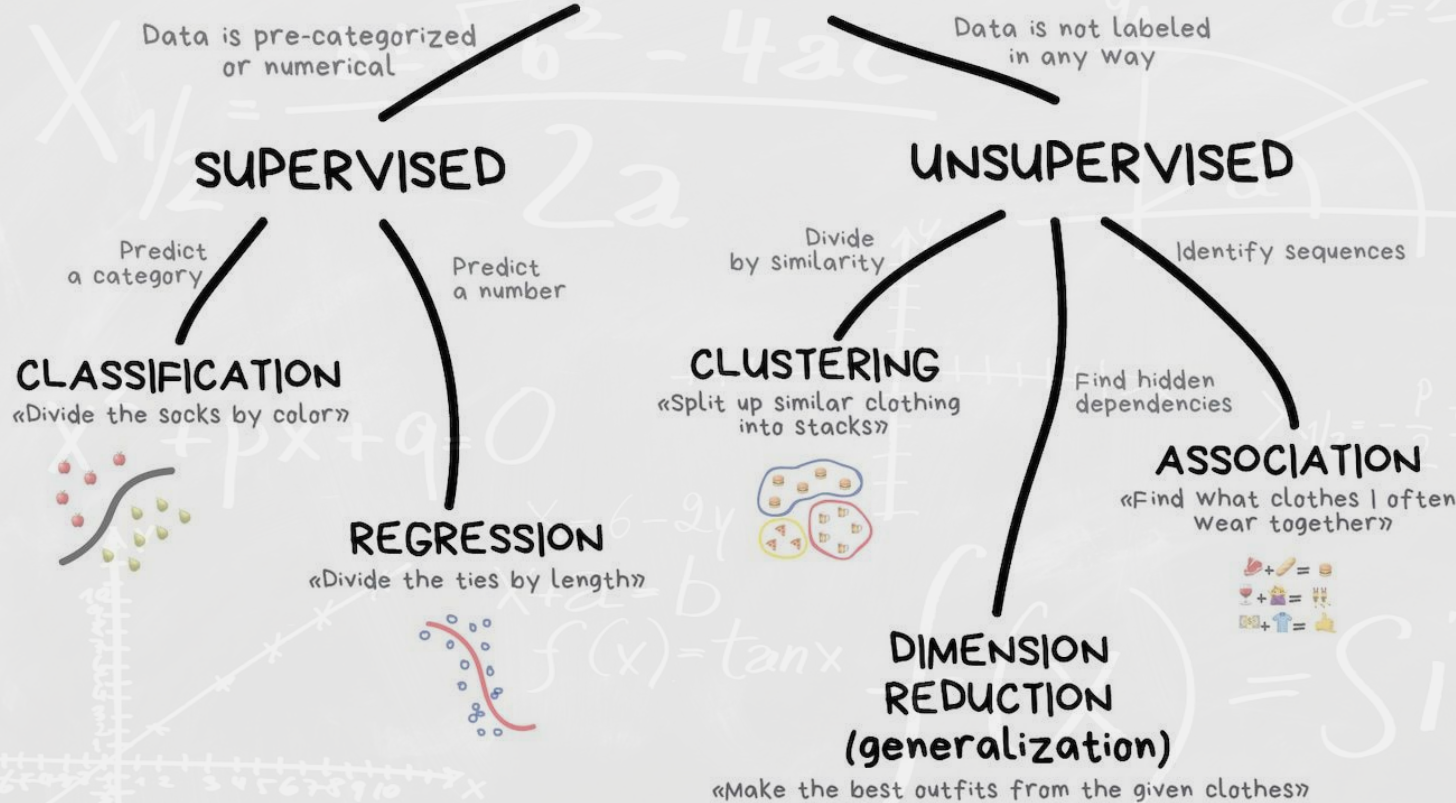
- **T**: Cancer diagnosis
- **E**: A set of diagnosed cases
- **P**: Accuracy of diagnosis on new cases
- **Z**: Noisy measurements, occasionally misdiagnosed training cases
- **M**: A program that runs on a general purpose computer; the **learner**

What is Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
 - Solve the optimization problem
 - Representing and evaluating the model for inference
- Role of Mathematics: Linear algebra and calculus to
 - Solve regression problem
 - Optimization functions

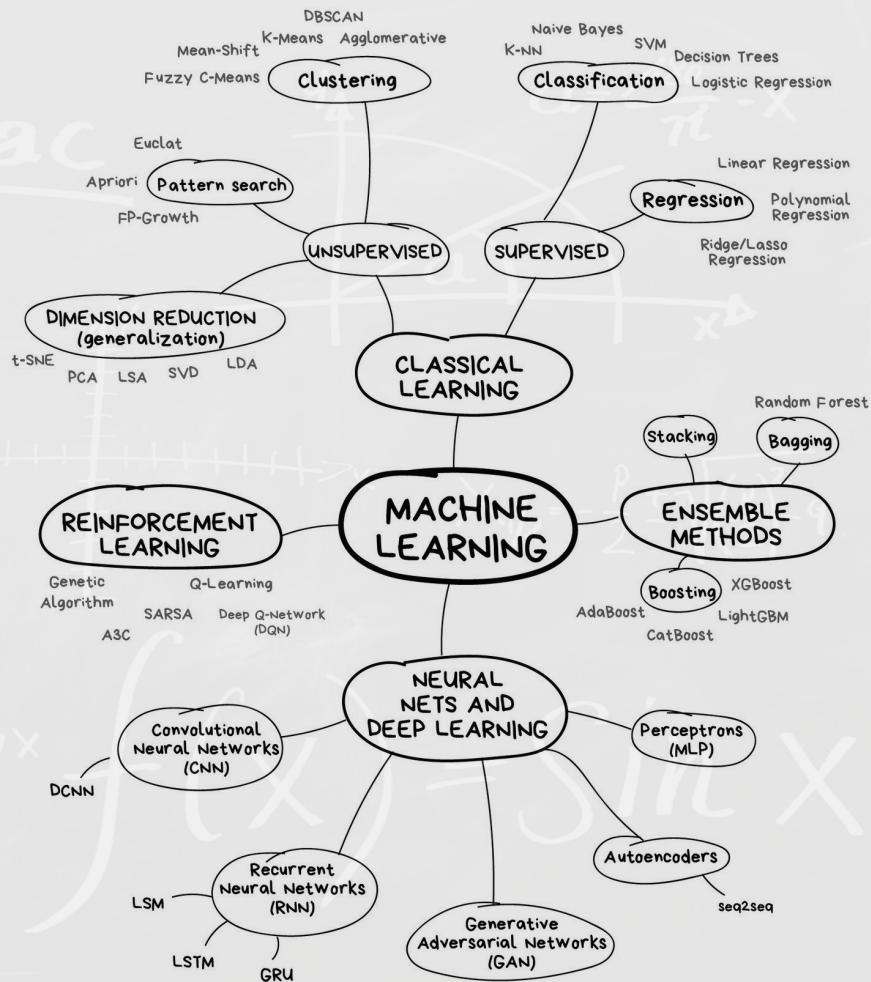


CLASSICAL MACHINE LEARNING



What is Machine Learning?

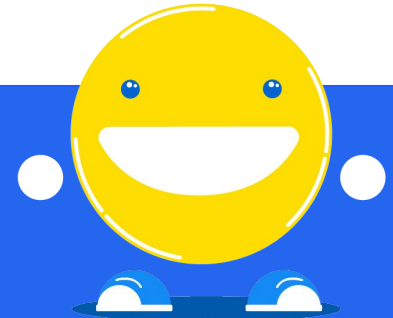
Too lazy to know it all? →

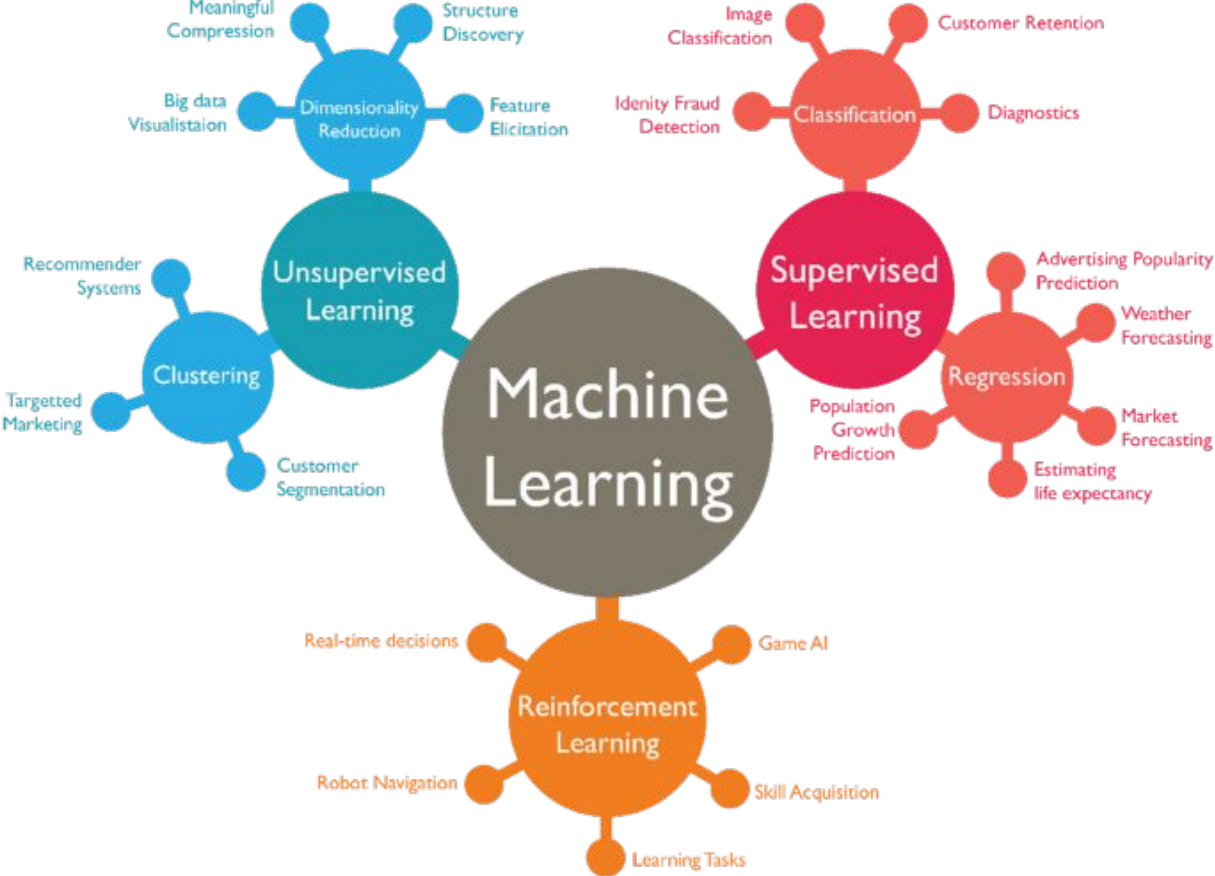


Supervised Learning



2019





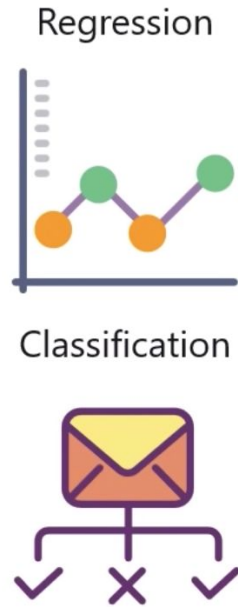
What Is Supervised Learning

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

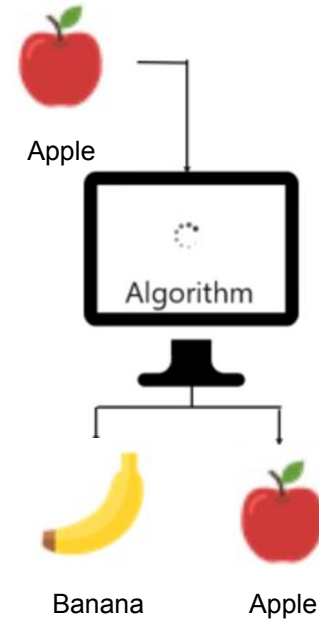
$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

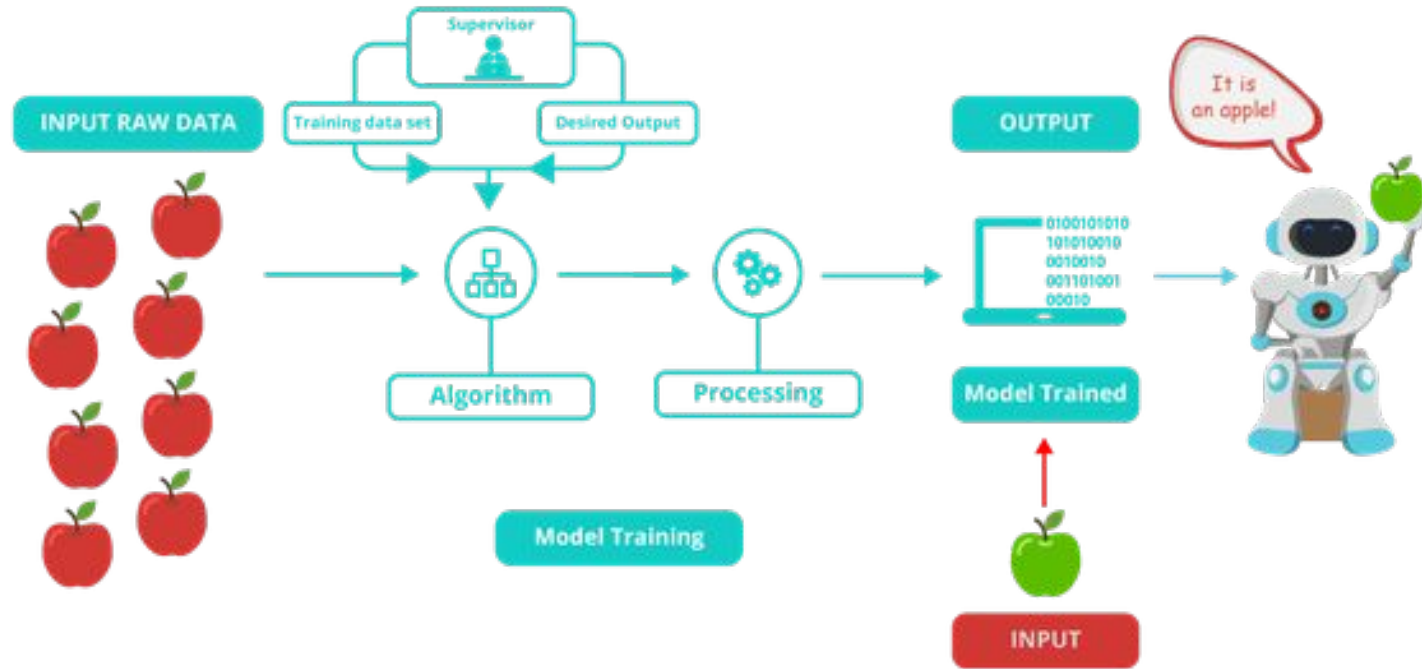
Supervised Problem



Supervised Data



How Supervised Learning Work



Why it is called Supervised

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher **supervising** the learning process.

We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher.

Learning stops when the algorithm achieves an acceptable level of performance.

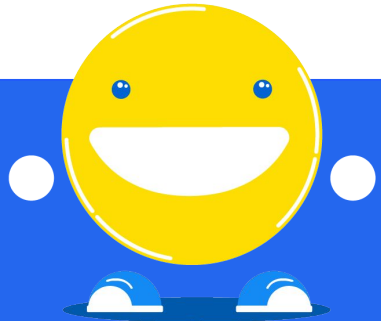
What is inside Supervised Learning ?

Supervised learning problems can be further grouped into regression and classification problems.

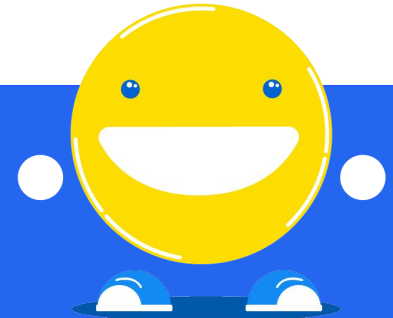
- **Classification:** A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- **Regression:** A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Let's Hands-on

Unsupervised Learning



2019

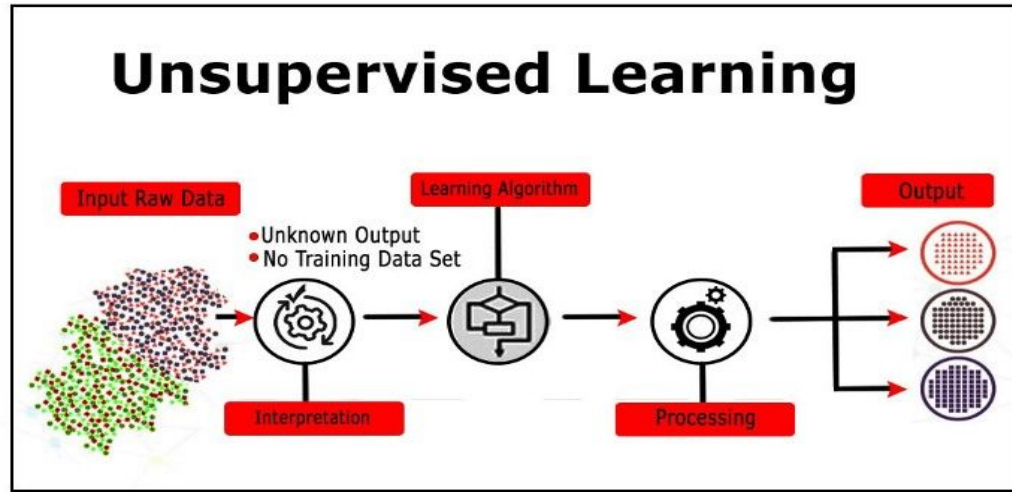


What is Unsupervised Learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

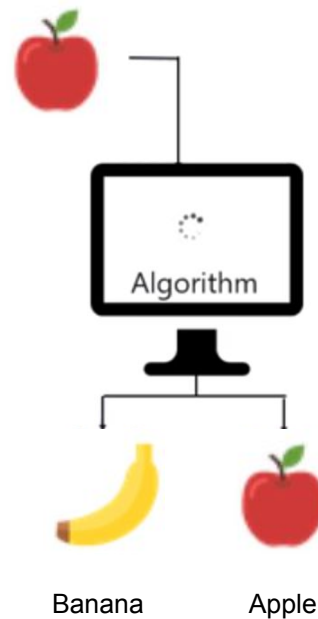
How Unsupervised Learning Work



Unsupervised Problem



Unsupervised Data



Why it is called Unsupervised

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devices to discover and present the interesting structure in the data.

What is inside Unsupervised Learning ?

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

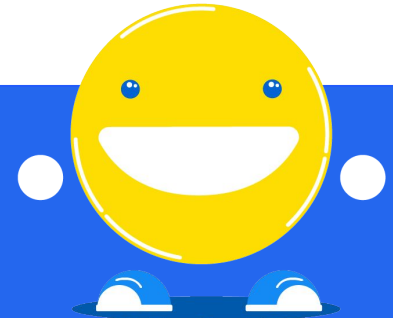
- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

Let's Hands-on

The Project



2019



Project List

1. Predicting Flight Delay
2. Predicting demand for Uber Taxis

Flight Delay - Problem Overview

Delay-free Ticket

In 2015, Ticket.com released delay-free ticket, a kind of ticket that grants a customer a free meal when his flight is departing or arriving late more than 15 minutes. This ticket applies to US top 5 budget airlines, such as Virgin Airlines (VX), JetBlue Airways (B6), Southwest (WN), Frontier (F9), Spirit (NK) for all flights departing to and arriving from selected airports. The list of airports can be found on the next slide.

Since the average number of passengers flying to/from those airports is 5 people, Ticket made a deal with the airlines to buy 5 meals at a reduced price for each flight connecting the selected airports. Ticket paid \$50 for a meal to the airlines. If the flight was not delayed, the paid meals could not be refunded. If the flight was delayed, the airline would give the free meal for Ticket's passengers during flight.

Flight Delay - Problem Overview

A New Strategy

The strategy was too costly so the management asked Data Science team to find a better way. If Tiket could predict delay before it happened, Tiket could bought meals at the same price just for the predicted flights. However, when a flight was delayed but Tiket did not provide free meals, Tiket would break its promise, hence Tiket would suffer a big loss to the brand. To compensate this, Tiket would have to give each passenger a discount voucher with value \$1,000 and also bear a loss to the brand which was estimated to be around \$500 per passenger.

Before suggesting this new strategy, the data science team must be able to find answer to the following questions. Was the new strategy better for Tiket? Could applying predictive modeling to predict flight delay make Tiket be able to save money?

Flight Delay - Problem Overview

List of airports

1. Hartsfield-Jackson, Atlanta (ATL)
2. Los Angeles International Airport (LAX)
3. Chicago O'Hare International Airport (ORD)
4. Dallas/Fort Worth (DFW)
5. Denver International Airport (DEN)
6. John F. Kennedy Airport (JFK)
7. San Francisco International Airport (SFO)
8. Charlotte Douglas Airport (CLT)
9. McCarran International Airport (LAS)
10. Miami International Airport (MIA)

Flight Delay - Problem Overview

Delay-free Ticket Scheme:

Current:

Buy 5 meals @\$50 for every flight.

New Scheme:

Buy 5 meals @\$50 only for predicted flights

Suffered loss for every prediction mistake:

1. \$1,000 per passenger to give discount voucher
2. \$500 per passenger due to loss of trust to the brand

Flight Delay - Datasets

The dataset for this analysis contains all flight records that happened in 2015. The dataset can be downloaded from Google Drive. The main dataset is titled “flights.csv”. Other spreadsheets contains list of airports, list of airlines, and descriptions of the columns.

Flight Delay - Data Exploration Guidance

Help Tiket to answer the following questions:

1. Is this a supervised or unsupervised problem? Is this classification, regression, or clustering problem?
2. Which airline has the highest number of delay?
3. Which airport has the highest number of departing/arriving delay?
4. What is the most popular airport/airline?
5. How long is the longest departing/arriving delay?
6. How much does Tiket have to pay in the old strategy?
7. What cause delay? What attributes have a high correlation with delay?

Flight Delay - Data Preprocessing

1. Load all columns as string.
2. Convert all dates from string into date.
3. Convert scheduled departure, departure_time, scheduled_arrival, and arrival_time into timestamp.
4. Filter airlines and airports
5. Make a label of delay. Departing and Arriving more than 15 minutes late will be classified as delayed.
6. Encode Airlines and Airports into columns using one-hot encoder.
7. Construct more attributes, for example:
 - a. How many flights departing around the same time from the airport?
 - b. How many flights arriving around the same time to the airport?
 - c. How many runways does the airport have?
 - d. What is the status of the previous flight from the same airport?
 - e. etc
8. Be Creative!
9. Calculate correlation between each of your constructed attributes and the delay. Are they highly correlated?

Flight Delay - Presentation Structure

1. Start from problem overview. Explain current condition. What is Tiket's initial cost?
2. Explain how you explore the data
3. Explain what attributes do you construct and the hypothesis behind each of them.
4. Predictive model that you build
5. Results and conclusion
6. Present the results using great visualizations.

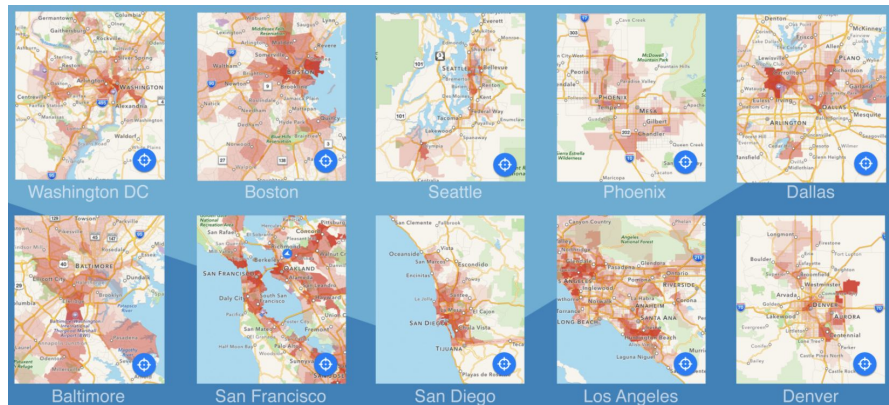
Uber Demand - Problem Overview

Passenger Hotspots

Uber is a US-based ride-hailing service like Gojek and Grab. To better assign drivers and adjust prices, Uber has to track hotspots (areas with high number of bookings) in its service area throughout the days. As shown in the picture, areas with dark red color are hotspots. When these hotspots occur, Uber will notify drivers so many of them can go around the area to find passengers.

A team of data scientists at Uber always monitor booking records and construct a heatmap. They will cluster bookings that happen around the same area based on their coordinate.

They then answer the following questions.
How many clusters are there in every hourly interval? What is the demand in an area at a specified hour on a specified day?



Uber Demand - Dataset

Trip Records

The dataset contains all uber trips happened in Chicago from January 1st, 2019. The following is the list of important columns.

1. Trip id
2. Trip start timestamp
3. Trip end timestamp
4. Pickup centroid latitude
5. Pickup centroid longitude

Uber Demand - Analysis

Number of Hotspots

Can you find out how many clusters there are in every interval hour on a specific day? What is the highest demand of each time and day? It would be better if you can visualize the clusters in a scatter plot.

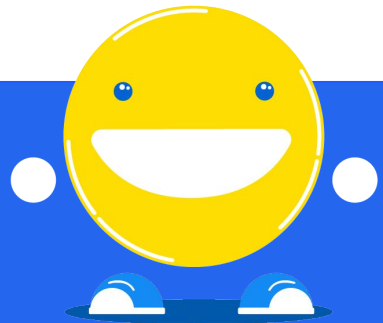
Demand Prediction

Define small squares of geographical area based on longitude and latitude. Predict demand around each areas that will happened at a specific time and on a specific day. The simplest demand prediction is the average demand on a specific time and day. Visualise the heatmaps if possible.

Uber Demand - Presentation Structure

1. Start from problem overview.
2. Explain how you explore the data
3. Explain data preprocessing process that you do.
4. Predictive model that you build
5. Present them in great visualizations
6. Results and conclusion

Machine Learning Workshop



2019

