Mahidol University
Wisdom of the Land

Mahidol University

Lecture2:
Machine Learning Framework

ITDS251x2 Fundamental Machine Learning

Instructor: Dr. Tipajin Thaipisutikul (Aj. Tip)
Contact: tipajin.tha@mahidol.edu

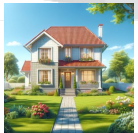
Today Agenda:

- Module1: Machine Learning and Data Science Framework
- Module2: Machine Learning Types
- Module3: 6 steps in ML & DS Framework
- Module4: Matplotlib and Data Visualization
- Module5: Overview of Linear Regression Model

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

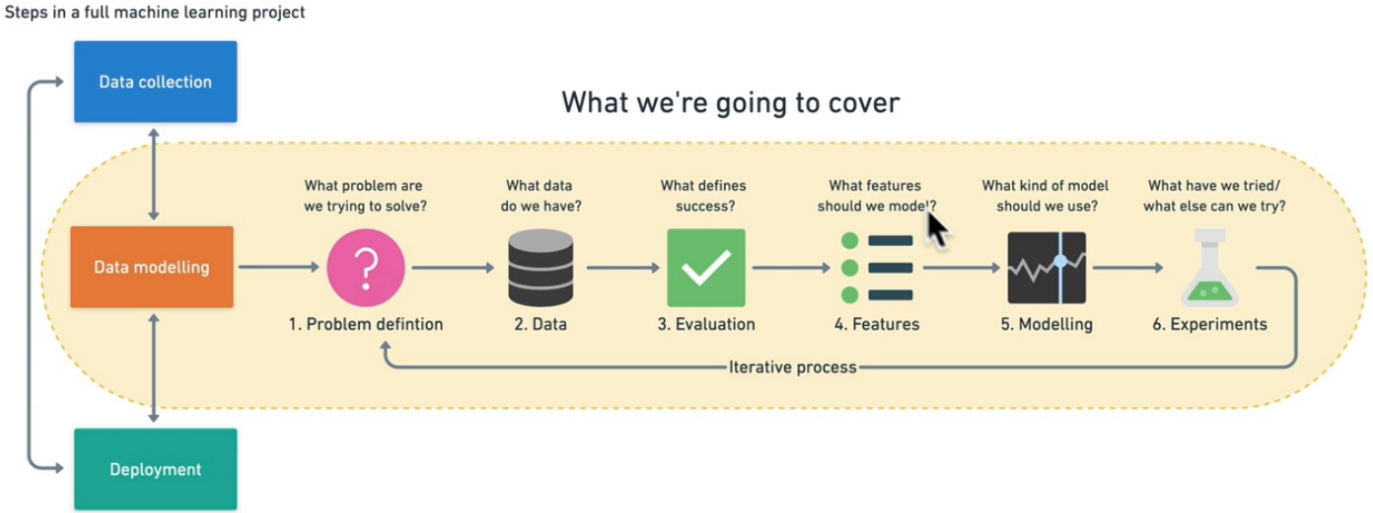
What is Machine Learning?

Welcome! You're our new Machine Learning and Data Science expert!



Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

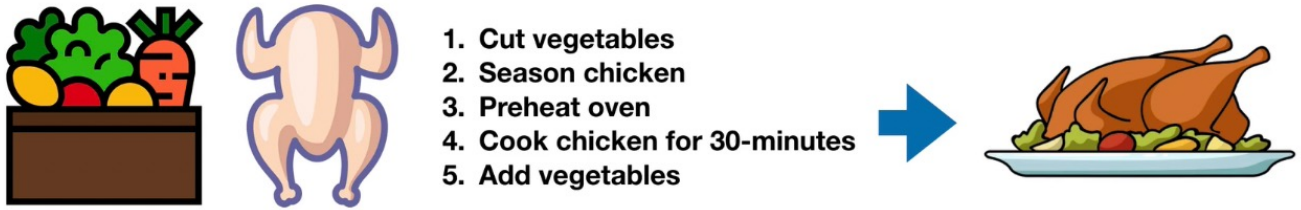
Introduction to Machine Learning and Data Science Framework



Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

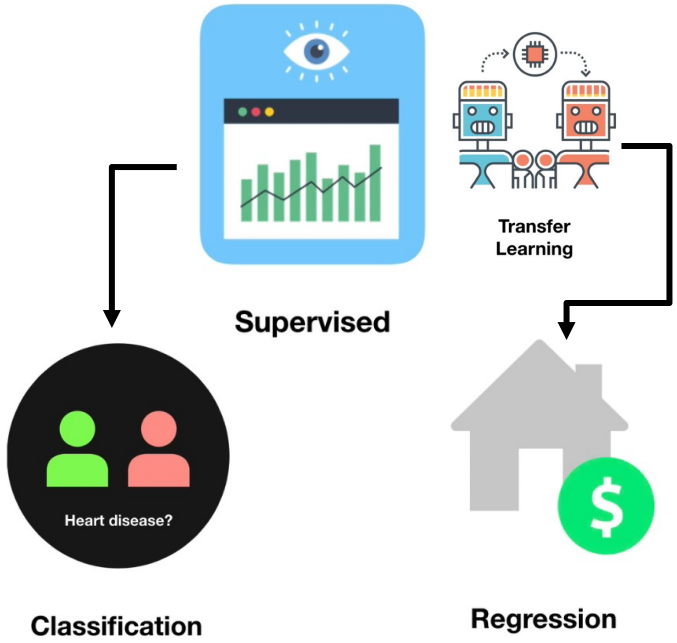
When shouldn't you use ML? เมื่อไรที่เราไม่ควรใช้ ML?

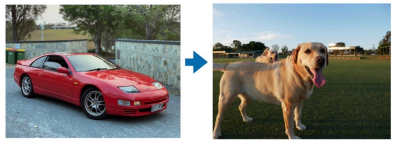
- Will a simple hand-coded instruction based system work?



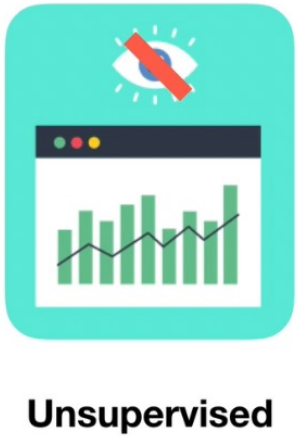
ถ้าเรารู้วิธีการทำอาหารจานนี้อยู่แล้ว เขียนโปรแกรมสั่งไปเลยดีกว่า เป็น AI แต่ไม่ได้เป็น ML

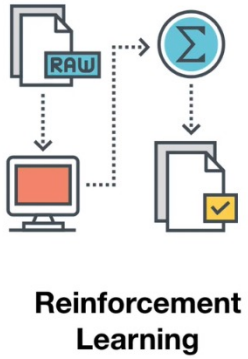
1. Problem Definition





“What problem are we trying to solve?”





☝ ทั้งนี้การแบ่งประเภทมีประโยชน์ต่อการตีโจทย์และการนำเอาเทคนิคที่มีอยู่แล้วไปประยุกต์ใช้

Machine Learning Types

ปัญหาทางด้าน Machine Learning มักจะถูกแบ่งออกไปตามคุณลักษณะของข้อมูลที่นำมาใช้ในการเรียนรู้ โดยเราสามารถกำหนดประเภทคร่าว ๆ ได้ดังต่อไปนี้

1. ⭐ **Supervised Learning**: e.g., Regression, Classification
2. ⭐ **Unsupervised Learning**: e.g., Clustering, Dimensionality reduction, Anomaly detection

Reinforcement Learning: e.g., Q-Learning, Policy Learning

💡 **Supervised Learning**:
Learn a function that maps an input, x, to an output, $\hat{y}$, using training samples of inputs and labels, $D = \{(x_1, y_1), (x_2, y_2), ...\}$

เรียนฟังก์ชันเพื่อแปลงค่า x เป็นค่า $\hat{y}$ โดยเรียนรู้จากข้อมูลที่ประกอบไปด้วยเซตของ x และ y.

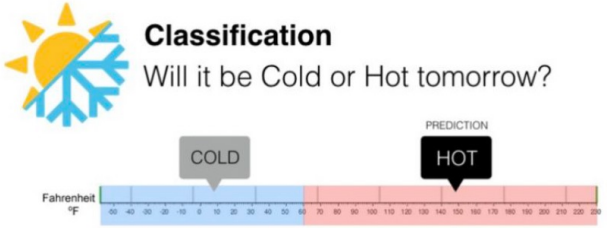
Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Machine Learning Types

Classification and Regression

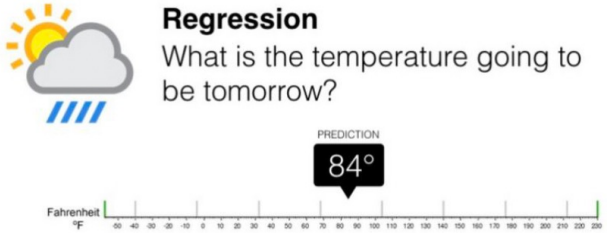
เราสามารถแบ่งประเภทย่อย ๆ ได้อีก 2 ประเภทใหญ่จากคุณสมบัติของ output ($\hat{y}$)

1. **Classification**: $\hat{y}$ เป็นตัวแปรที่เป็นประเภท (discrete)



https://gombru.github.io/2018/05/23/cross_entropy_loss/

2. **Regression**: $\hat{y}$ เป็นตัวแปรที่เป็นประเภทเลขจำนวนจริง (continuous)



https://gombru.github.io/2018/05/23/cross_entropy_loss/

Classification

- "Is this example one thing or another?"
- Binary classification = two options
- Multi-class classification = more than two options

Regression

- "How much will this house sell for?"
- "How many people will buy this app?"

Machine Learning Types

แบบฝึกหัด

- คาดการ์ณเกรดของนักศึกษา — Classification or Regression
- คาดการ์ณคะแนนสอบ — Classification or Regression
- Spam mail filter — Classification or Regression
- PM2.5 forecast — Classification or Regression

Note: Transfer Learning (การเรียนรู้แบบถ่ายโอน) เป็นเทคนิคใน Machine Learning ที่นำความรู้หรือรูปแบบที่ได้จากการฝึกโมเดลใน ชุดข้อมูล หรือ งานหนึ่ง ไปปรับใช้กับงานใหม่ โดยไม่ต้องเริ่มการฝึกจากศูนย์ ซึ่งช่วยประหยัดเวลาและทรัพยากรในการฝึกโมเดล โดยเฉพาะในกรณีที่ชุดข้อมูลใหม่มีขนาดเล็กหรือมีทรัพยากรจำกัด

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

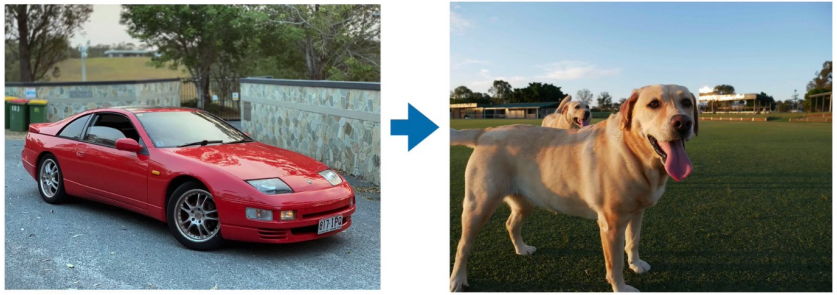
Machine Learning Types

Supervised Learning:
Learn a function that maps an input, x, to an output, $\hat{y}$, using training samples of inputs and labels, $D = \{(x_1, y_1), (x_2, y_2), ...\}$

เรียนฟังก์ชันเพื่อแปลงค่า x เป็นค่า $\hat{y}$ โดยเรียนรู้จากข้อมูลที่ประกอบไปด้วยเซตของ x และ y.

Transfer Learning

Transfer learning



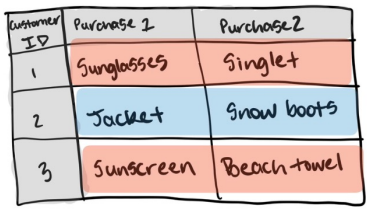
Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

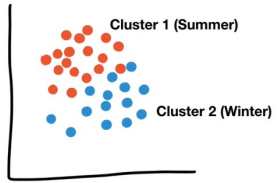
Machine Learning Types

Unsupervised Learning:
Learn a function that maps an input, x, to an output, $\hat{y}$, using training samples of inputs and labels, $D = \{x_1, x_2, ...\}$

เรียนฟังก์ชันเพื่อแปลงค่า x เป็นค่า $\hat{y}$ โดยเรียนรู้จากข้อมูลที่ประกอบไปด้วยเซตของ x

Unsupervised learning แบ่งออกเป็นหลายประเภท





แบบฝึกหัด

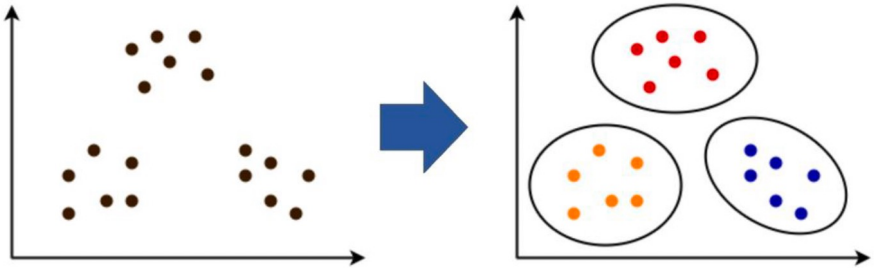
- A. แยก pixels บนรูปภาพเป็นกลุ่มต่าง ๆ
- B. ตรวจจับการทำงานที่ผิดปกติของเครื่องจักร

- I Clustering
- II Deminsionality Reduction
- III Anomaly Detection

Machine Learning Types

1 Clustering

จัดกลุ่มข้อมูลเป็นชุด ๆ $\hat{y}$ เป็นสมาชิกกลุ่มของข้อมูล x



https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

ทั้งนี้อาจจะจัดกลุ่มแบบ hard clusters (แค่หนึ่งกลุ่ม) หรือ soft clusters (เปอร์เซนต์การมีส่วนร่วม)

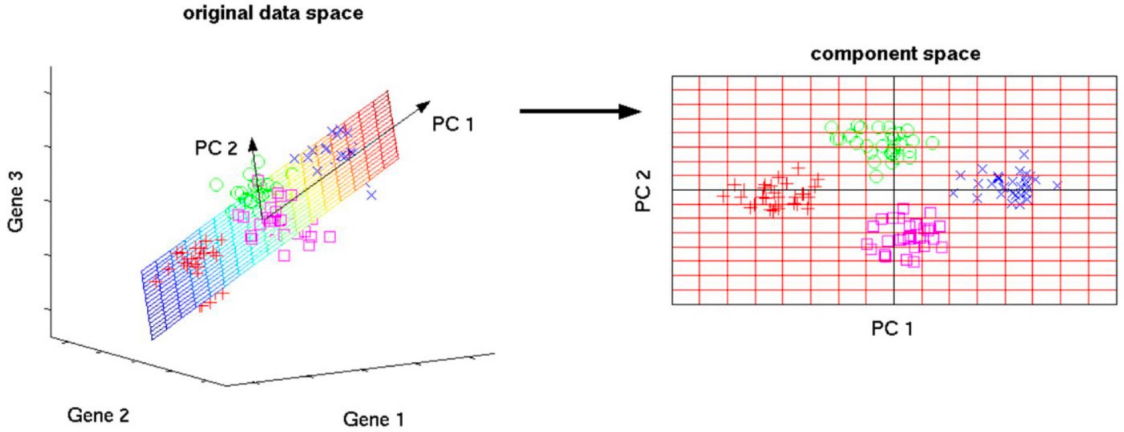
ตัวอย่าง

1. Market segmentation: จัดกลุ่มลูกค้าที่มีพฤติกรรมคล้าย ๆ กัน
2. Document segmentation: จัดกลุ่มเอกสารที่มีเนื้อหาคล้าย ๆ กัน

Machine Learning Types

2 Dimensionality Reduction

ลดมิติของข้อมูล $\hat{y}$ เป็นข้อมูล x ที่มีมิติลดลงและมักจะมีค่าเปลี่ยนแปลงไป



http://www.nlpca.org/pca_principal_component_analysis.html

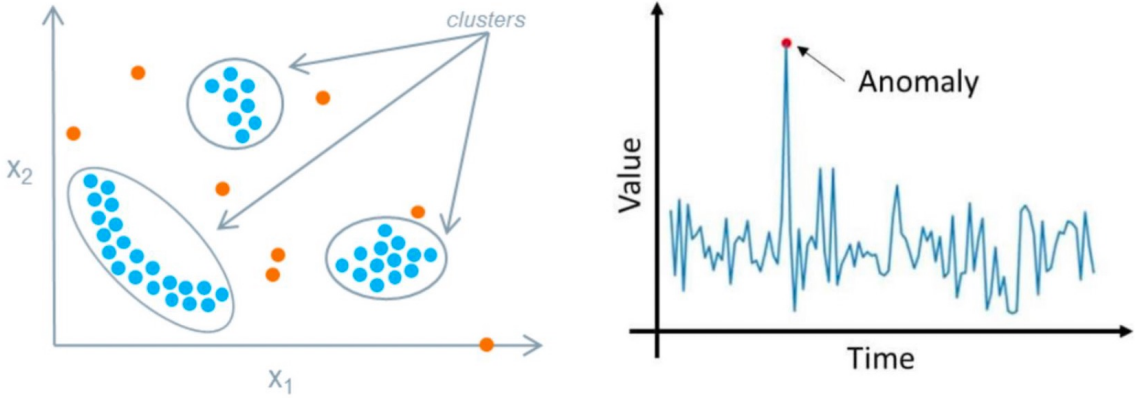
ตัวอย่าง

1. Data visualization: ลดมิติข้อมูลลงเพื่อให้แสดงผลบนหน้าจอได้ เช่น จาก 5 มิติให้เป็น 3 มิติ
2. Data compression: ลดมิติข้อมูลลงเพื่อให้ขนาดข้อมูลลดลง

Machine Learning Types

3 Anomaly Detection

ระบุว่าข้อมูลไหนผิดแปลกไปจากข้อมูลส่วนใหญ่ $\hat{y}$ เป็นค่า True หรือ False (ส่วนใหญ่เป็นค่าความน่าจะเป็น)

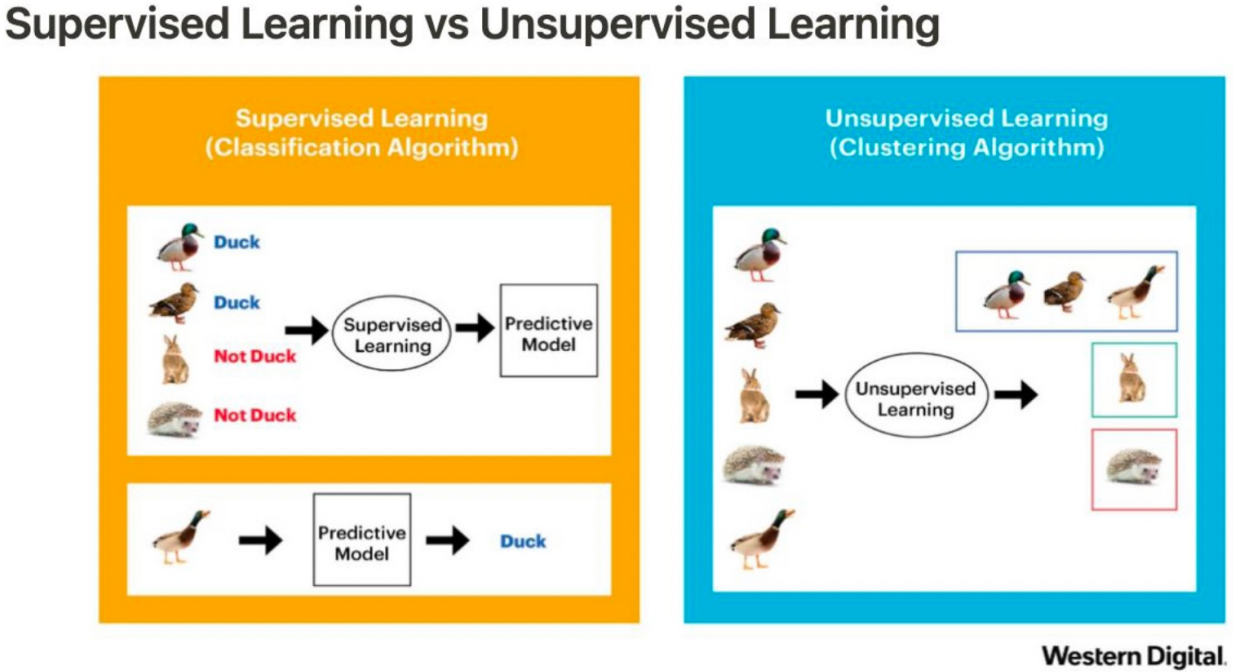


Left: https://developer.mindsphere.io/apis/analytics-anomalydetection/api-anomalydetection-overview.html
Right: https://medium.com/datadriveninvestor/how-machine-learning-can-enable-anomaly-detection-eed9286c5306

ตัวอย่าง

1. Credit card fraud detection: ตรวจจับการซื้อที่อาจจะเป็นการโดนโกงบัตรเครดิต
2. Network intrusion detection: ตรวจจับรูปแบบของ network traffic ที่อาจจะเป็นการโจมที

Machine Learning Types



https://blog.westerndigital.com/machine-learning-pipeline-object-storage/

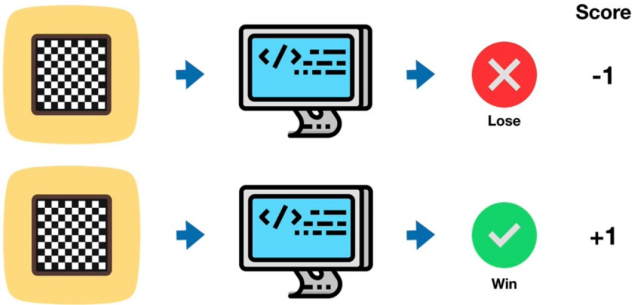
Reinforcement Learning

Reinforcement Learning:
Learn a function that maps an input, x, to an output, $\hat{y}$, using training samples based on
- interaction: observation and action (x, $\hat{y}$) and
- reward (r).

เรียนฟังก์ชันเพื่อแปลงค่า x เป็นค่า $\hat{y}$ โดยเรียนรู้จากข้อมูลที่มาจากการปฏิสัมพันธ์กับสิ่งแวดล้อม โดยประกอบไปด้วยข้อมูลจากการสังเกตุ (x) การกระทำ ($\hat{y}$) และรางวัล (r)

โดยคุณสมบัติที่สำคัญของ Reinforcment learning คือ

1. มี Reward แทนที่จะเป็น Label
2. Reward มักจะไม่ค่อยมี
3. ข้อมูลเกิดจากผลลัพธ์ของแบบจำลองเอง



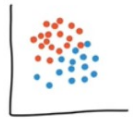
Matching your problem

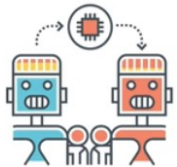
Supervised Learning

"I know my inputs and outputs."

Unsupervised Learning

"I'm not sure of the outputs but I have inputs."





Transfer Learning

"I think my problem may be similar to something else."

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

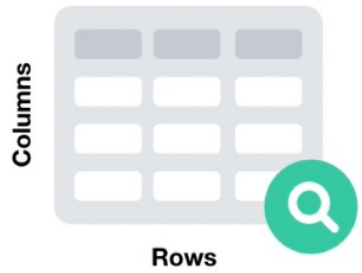
"What kind of data do we have?"

Rows

Columns

ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4328	110kg	M	120/80	4	Yes
5681	64kg	F	130/90	1	NO
7911	81kg	M	130/80	0	NO

Table 1.0: Patient records



Structured

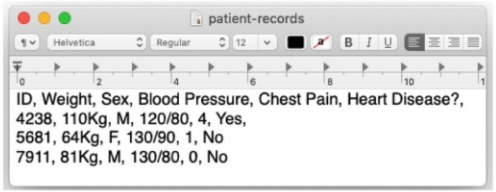
First of all, thank you for being so amazing.
This machine learning course is incredible.
Thank you for keeping it simple!

Unstructured

Different types of data

"The more data the better."

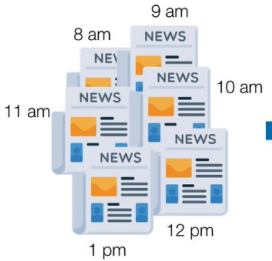
Static



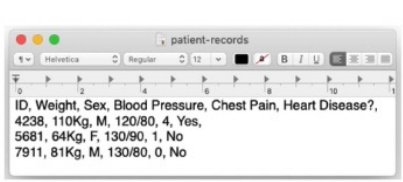
ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4326	110Kg	M	120/80	4	yes
5681	64Kg	F	130/90	1	No
7911	81Kg	M	130/80	0	No

Table 1.0: Patient records

Streaming



Data Science WorkFlow

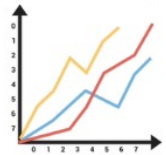


Static data

ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4528	110kg	M	120/80	4	Yes
5681	64kg	F	130/90	1	No
7911	81kg	M	130/80	0	No

Table 1.0: Patient records

Data Analysis



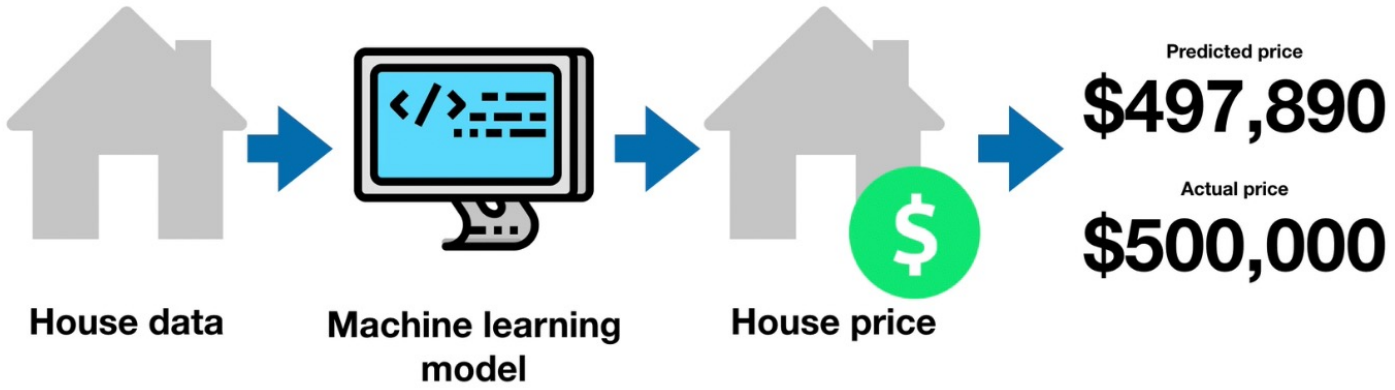
Machine learning model

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

3. Evaluation

“What define success for us?”

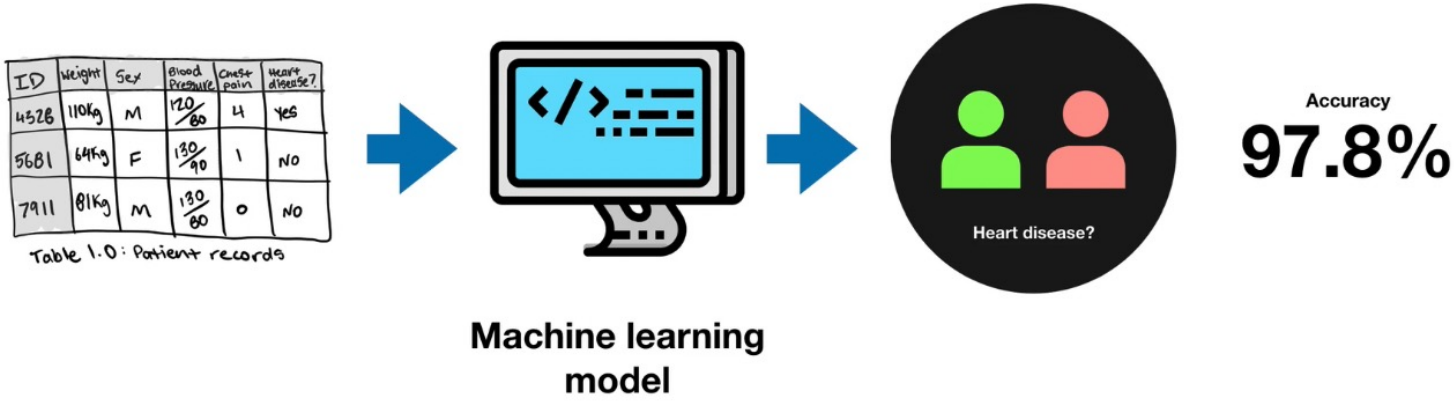
- **Email Spam Detection**
- **Customer Churn Prediction**
- **Stock Market Prediction**



Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

3. Evaluation

"For this project to be worth pursuing further, we need a machine learning model with over 99% accuracy."



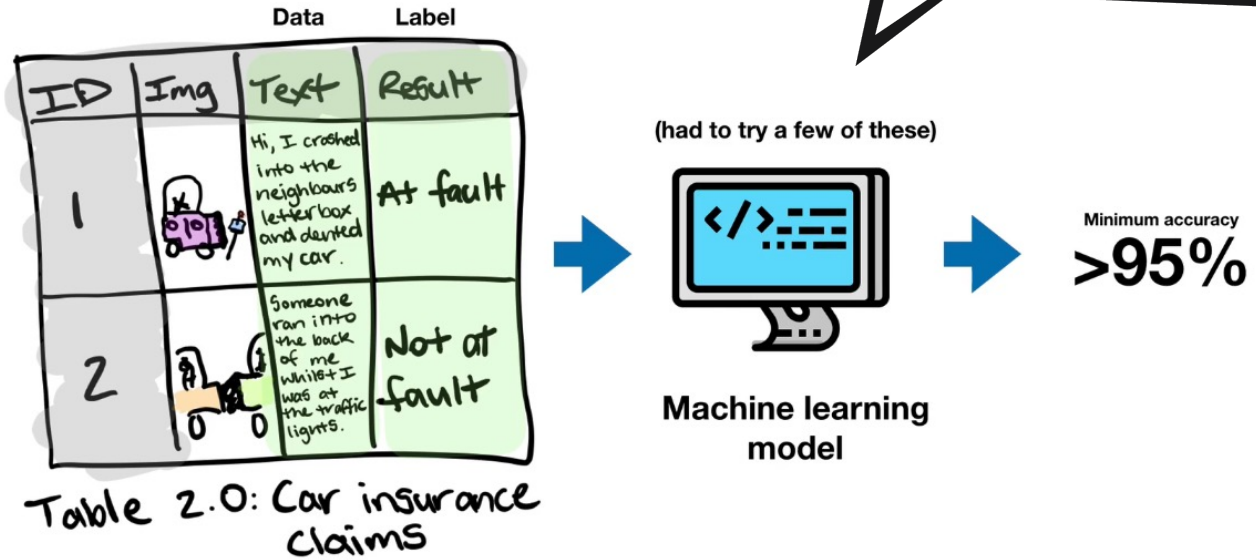
Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

3. Evaluation

Classification	Regression	Recommendation
Accuracy	Mean absolute error (MAE)	Precision at K
Precision	Mean squared error (MSE)	
Recall	Root mean squared error (RMSE)	

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Classifying car insurance claims

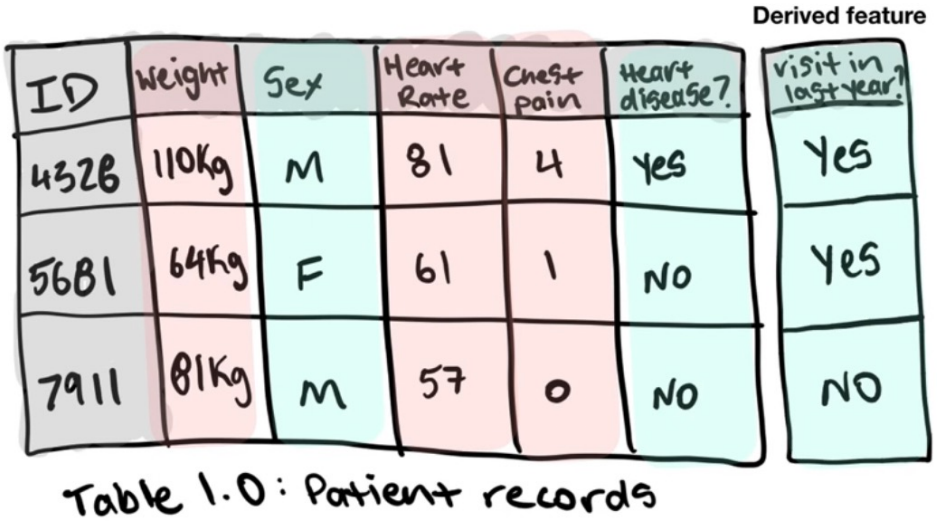


“What do we already know about the data?”

	Feature variables				Target variable
ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110kg	M	81	4	YES
5681	64kg	F	61	1	NO
7911	81kg	M	57	0	NO

Table 1.0: Patient records

“What do we already know about the data?”



ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?	visit in last year?
4326	110kg	M	81	4	yes	Yes
5681	64kg	F	61	1	NO	Yes
7911	81kg	M	57	0	NO	NO

Derived feature

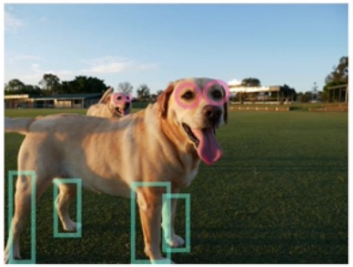
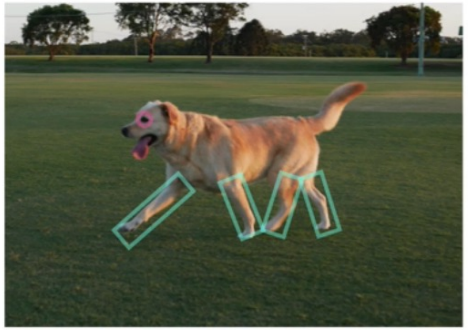
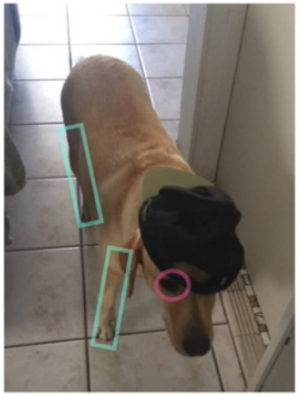
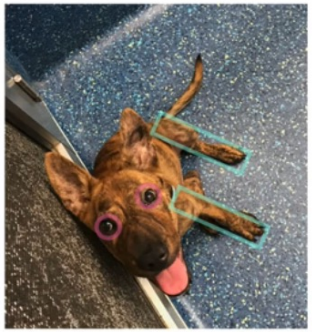
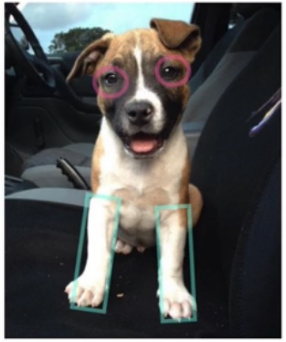
Table 1.0: Patient records

Numerical features

Categorical features

Feature engineering Looking at different features of data and creating new ones/altering existing ones

Different features of data


Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

What features should we use?

What are features of your problems?

ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?	Last eaten food
4326	110kg	M	81	4	yes	fries
5681	64kg	F	61	1	NO	?
7911	81kg	M	57	0	NO	?

Table 1.0: Patient records

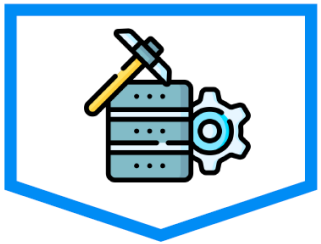
Want > 10% coverage

Feature coverage How many samples have different features? Ideally, every sample has the same features.

5. Modelling

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

A Roadmap for building ML Systems



1. Data Pre-Processing

- Import the data
- Clean the data
- Split into Training & Testing sets
- Feature Scaling

- Feature Extraction and Scaling
- Feature Selection
- Dimentionality Reduction
- Sampling

2. Modelling

- Build the model
- Train the model
- Make predictions

- Model Selection
- Cross-validation
- Performance Metrics
- Hyperparameter Optimization

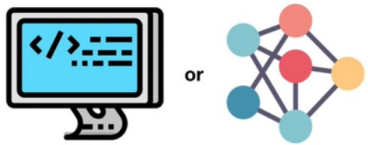
3. Evaluation

- Calculate Performance metrics
- Make a verdict

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

"Based on our problem and data, what model should we use?"

3 parts to modelling

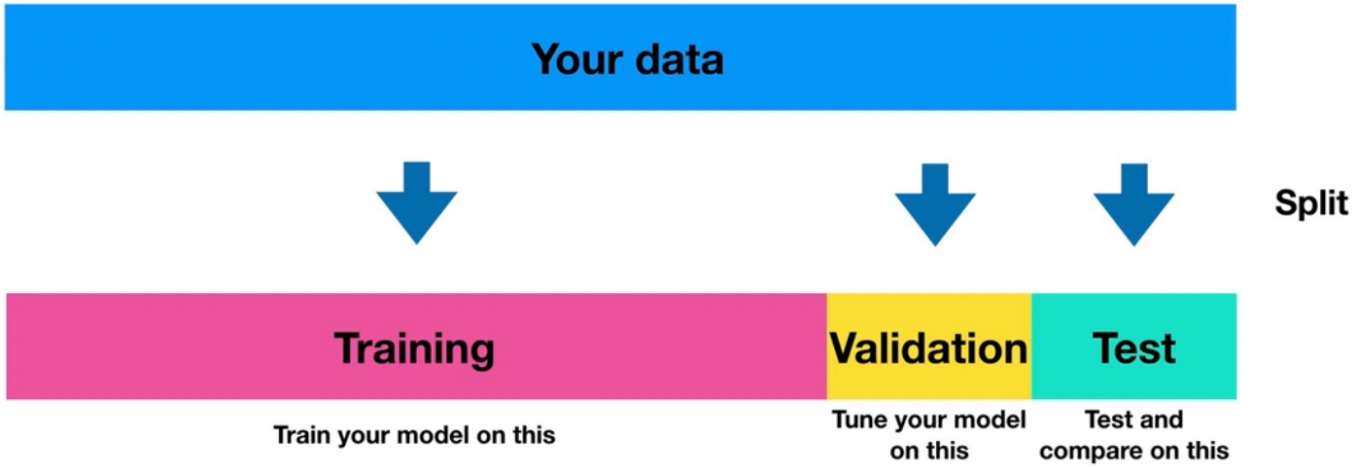
1. Choosing and training a model 

2. Tuning a model 

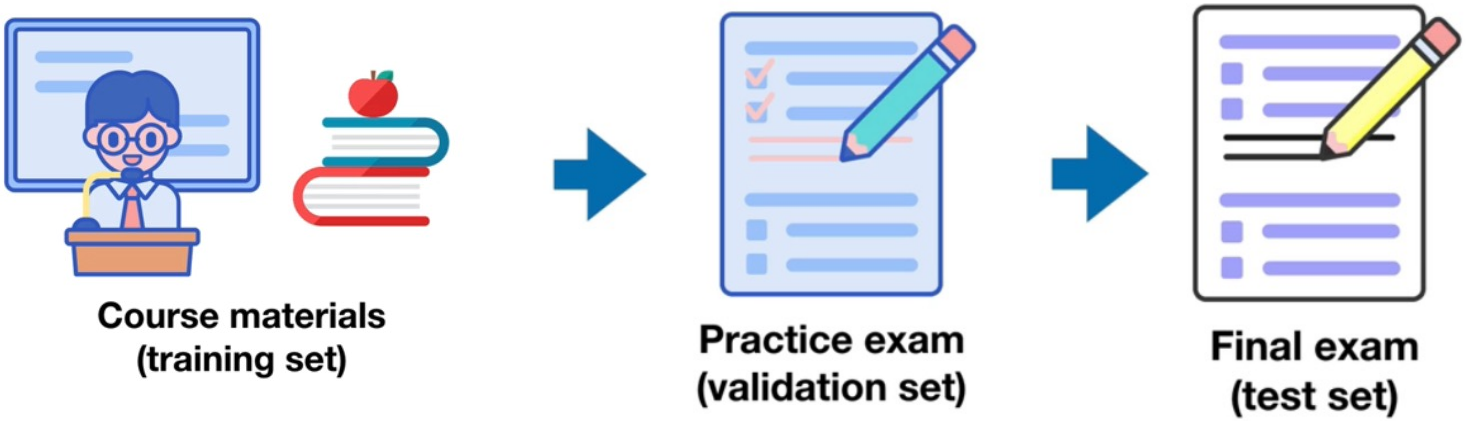
3. Model comparison 

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

The most important concept in machine learning
(the training, validation, and test sets or 3 sets)



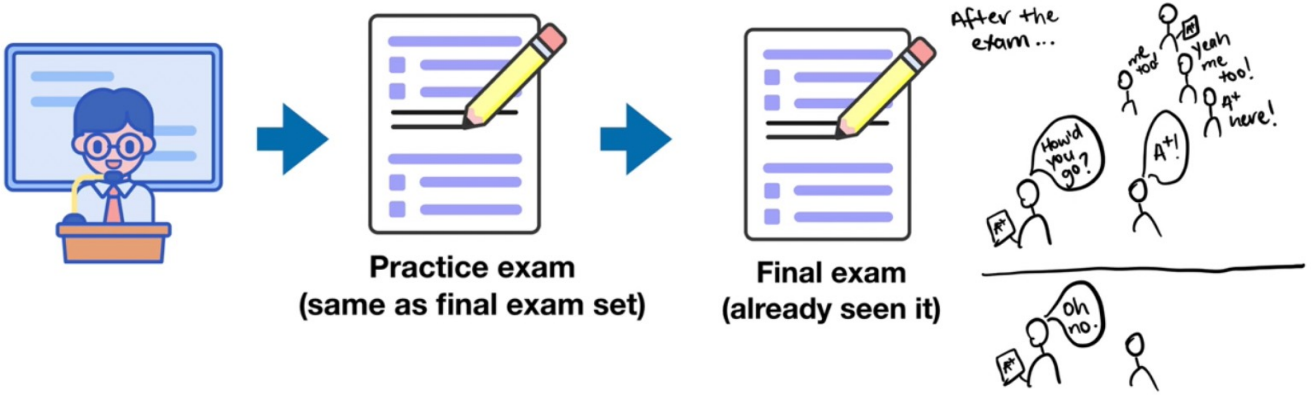
The most important concept in machine learning
(the training, validation, and test sets or 3 sets)



Generalization The ability for a machine learning model to perform well on data it hasn't seen before.

5. Modelling (where might this go wrong?)

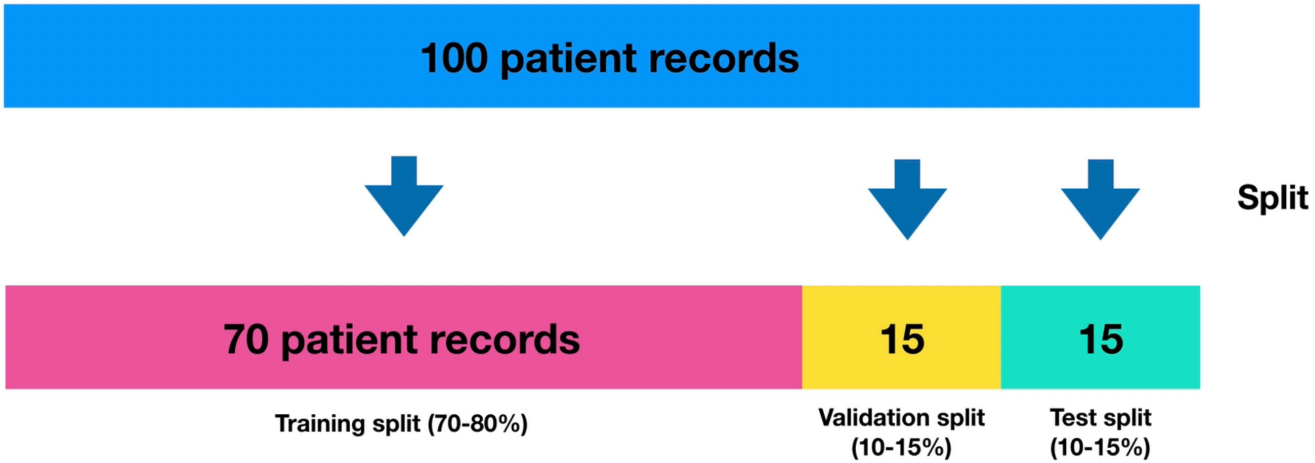
The most important concept in machine learning
(the training, validation, and test sets or 3 sets)



Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

The most important concept in machine learning
(the training, validation, and test sets or 3 sets)

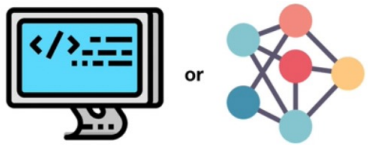
What was the last thing you testing your ability on?



“Based on our problem and data, what model should we use?”

3 parts to modelling

1. Choosing and training a model



Training Data

2. Tuning a model

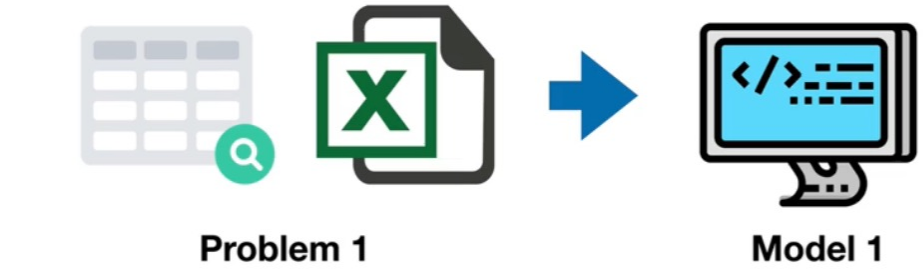
Validation Data

3. Model comparison

Test Data

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Choosing a model



Structured Data

CatBoost

dmlc XGBoost

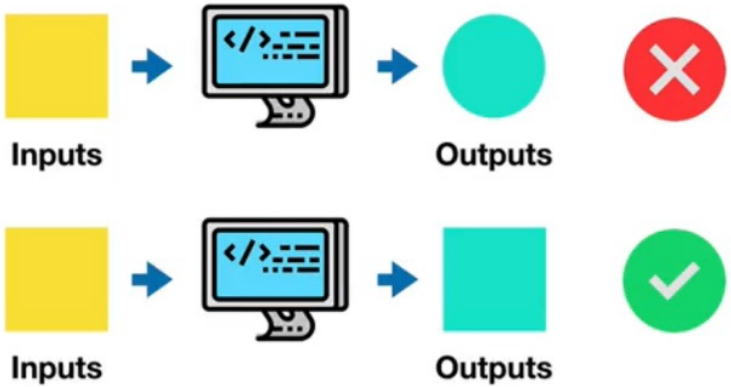
Random Forest

Unstructured Data

Deep Learning

Transfer Learning

5. Modelling



X (data) y (label)

ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110kg	M	81	4	yes
5681	64kg	F	61	1	NO
7911	81kg	M	57	0	NO

Table 1.0: Patient records

Training Data

“Based on our problem and data, what model should we use?”

3 parts to modelling

1. Choosing and training a model

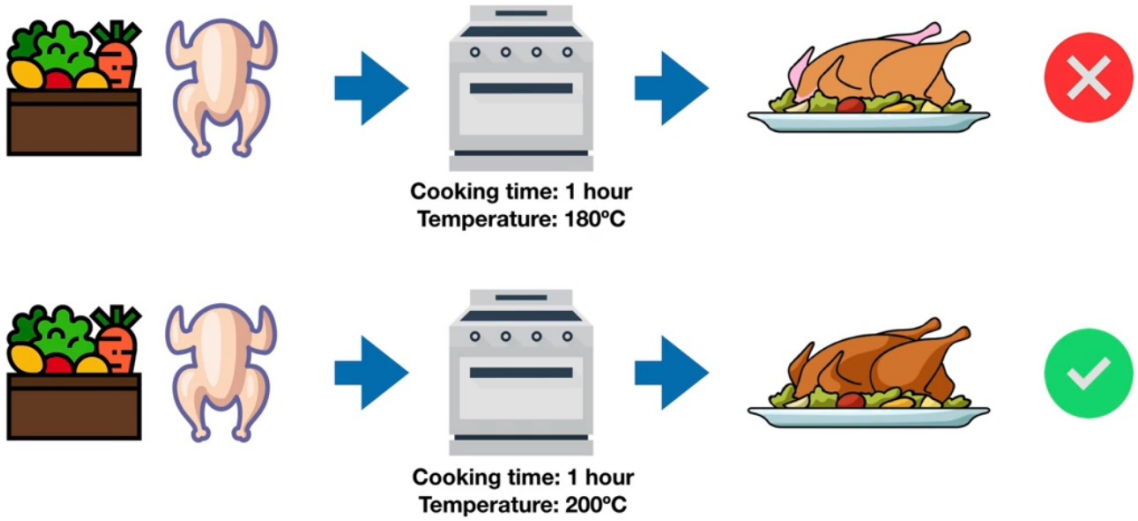
Training Data

2. Tuning a model

Validation Data

3. Model comparison

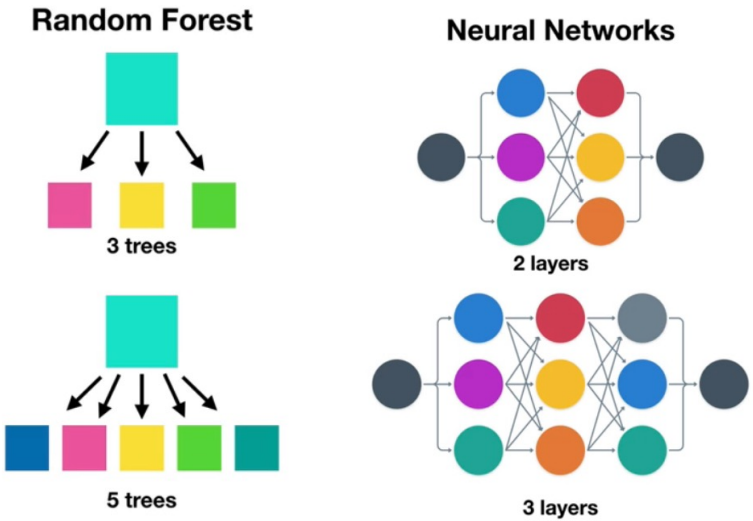
Test Data



Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

2. Tuning a model

Validation Data

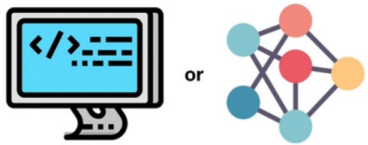


Validation Set ถูกใช้เพื่อประเมินและปรับแต่งโมเดล (เช่น โครงสร้าง) ระหว่างการพัฒนาโมเดล หรือ หลังจากเทรนจบในแต่ละโครงสร้าง แต่ยังไม่ใช่การประเมินขั้นสุดท้าย (ซึ่งควรใช้ Test Set) เพื่อให้ได้โมเดลที่มีความสามารถทั่วไป (generalize) ดีที่สุด

“Based on our problem and data, what model should we use?”

3 parts to modelling

1. Choosing and training a model



Training Data

2. Tuning a model

Validation Data

3. Model comparison

Test Data

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Testing a model

Data Set	Performance
Training	98%
Test	96%

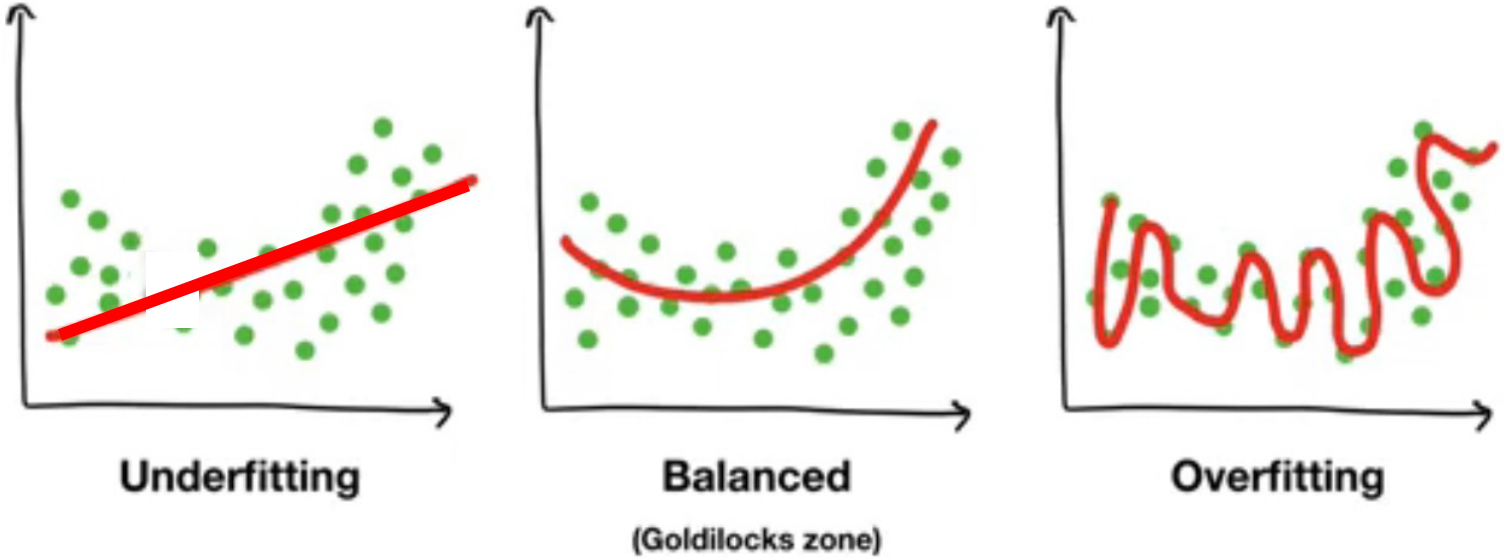
Underfitting (potential)

Data Set	Performance
Training	64%
Test	47%

Overfitting (potential)

Data Set	Performance
Training	93%
Test	99%

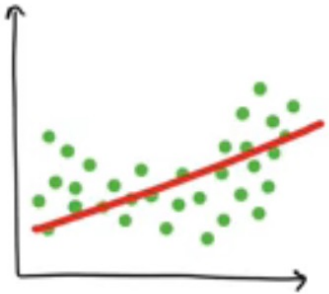
Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]


Underfitting
Balanced
(Goldilocks zone)
Overfitting


Practice exam
Final exam
Training Data
Test Data
Data leakage
Overfitting

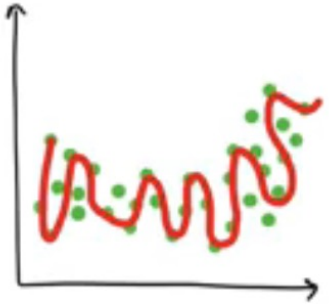

Training Data
Test Data
Data mismatch
Underfitting

Fixes for overfitting and underfitting



Underfitting

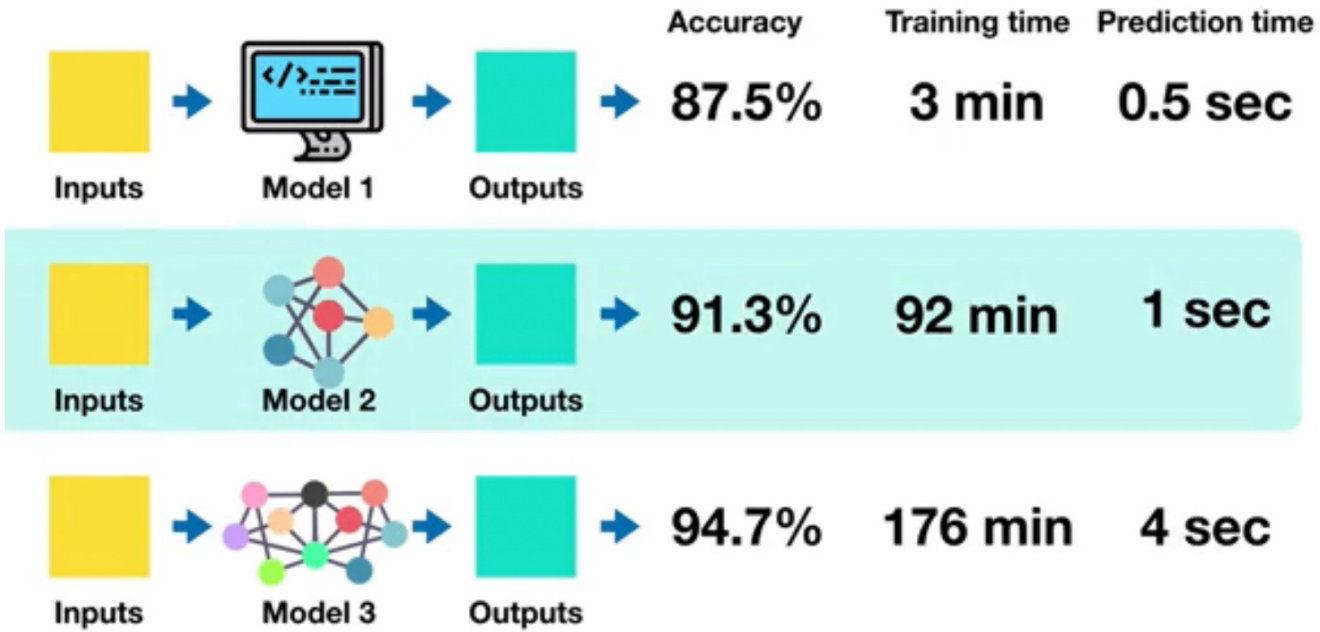
- Try a more advanced model
- Increase model hyperparameters
- Reduce amount of features
- Train longer



Overfitting

- Collect more data
- Try a less advanced model

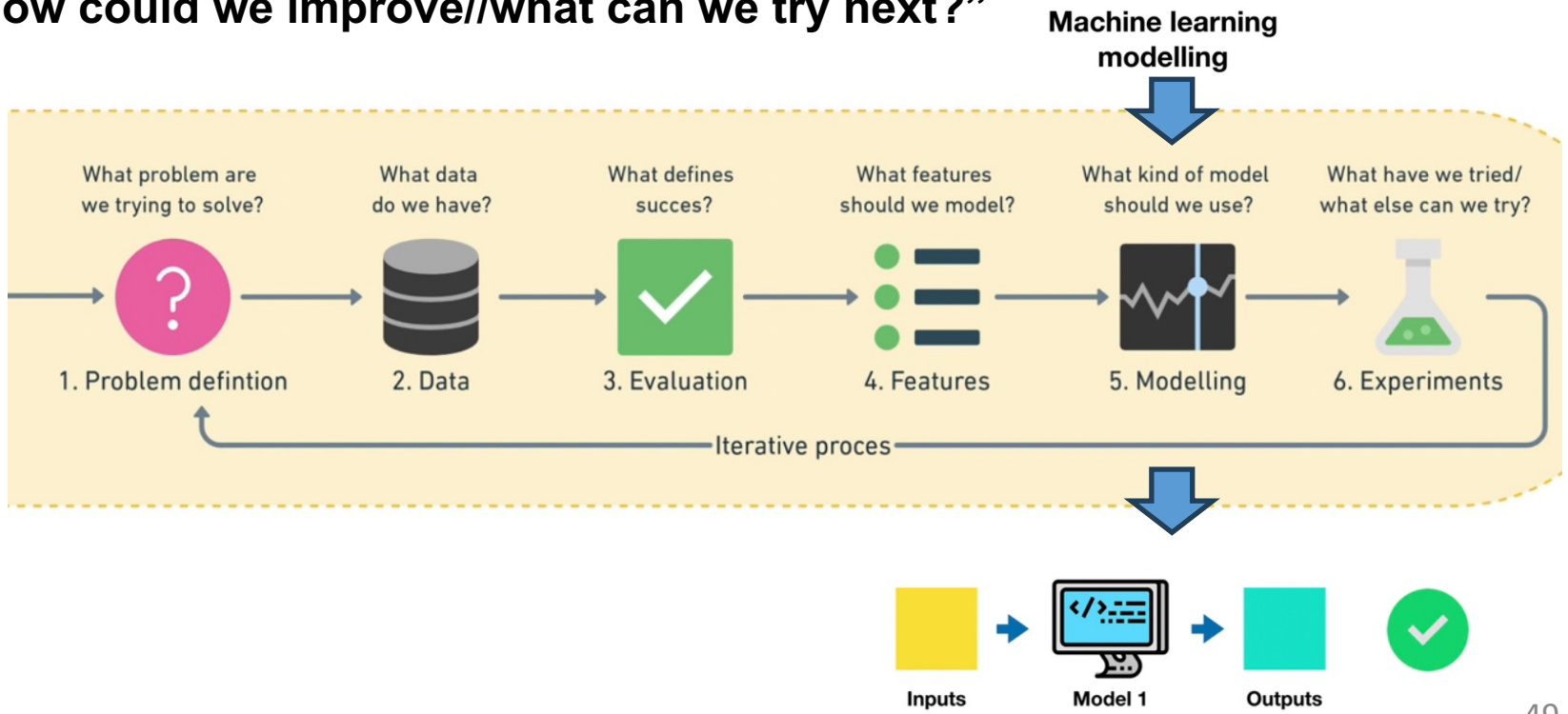
5. Modelling



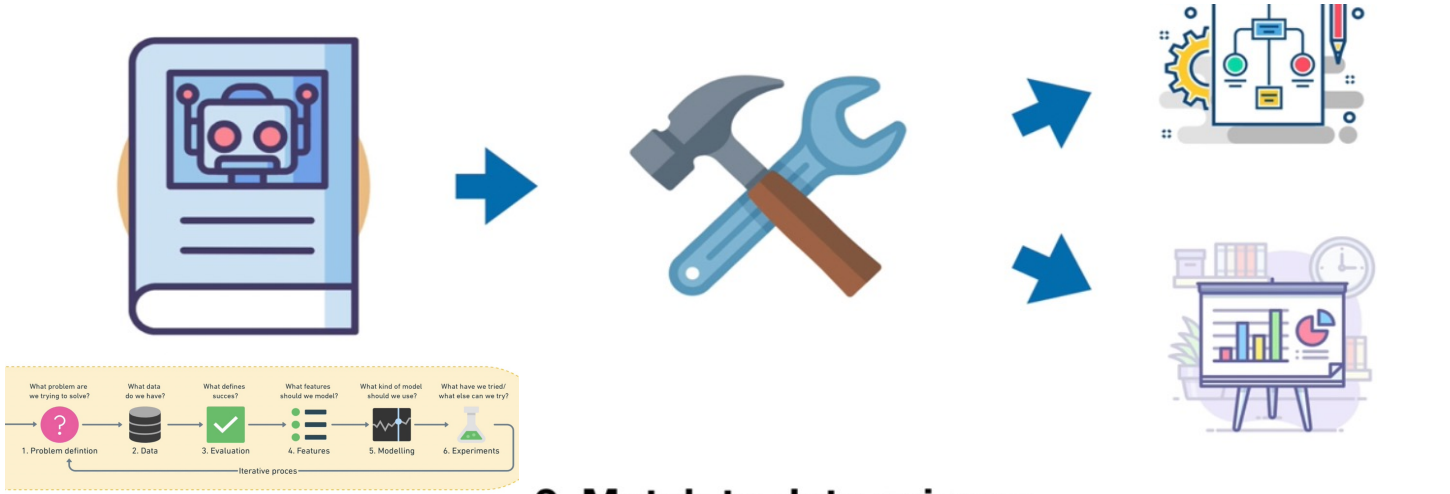
Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

- Want to avoid overfitting and underfitting (head towards generality)
- Keep the test set separate at all costs
- Compare apples to apples
- One best performance metric does not equal best model

“How could we improve//what can we try next?”



What's Next?



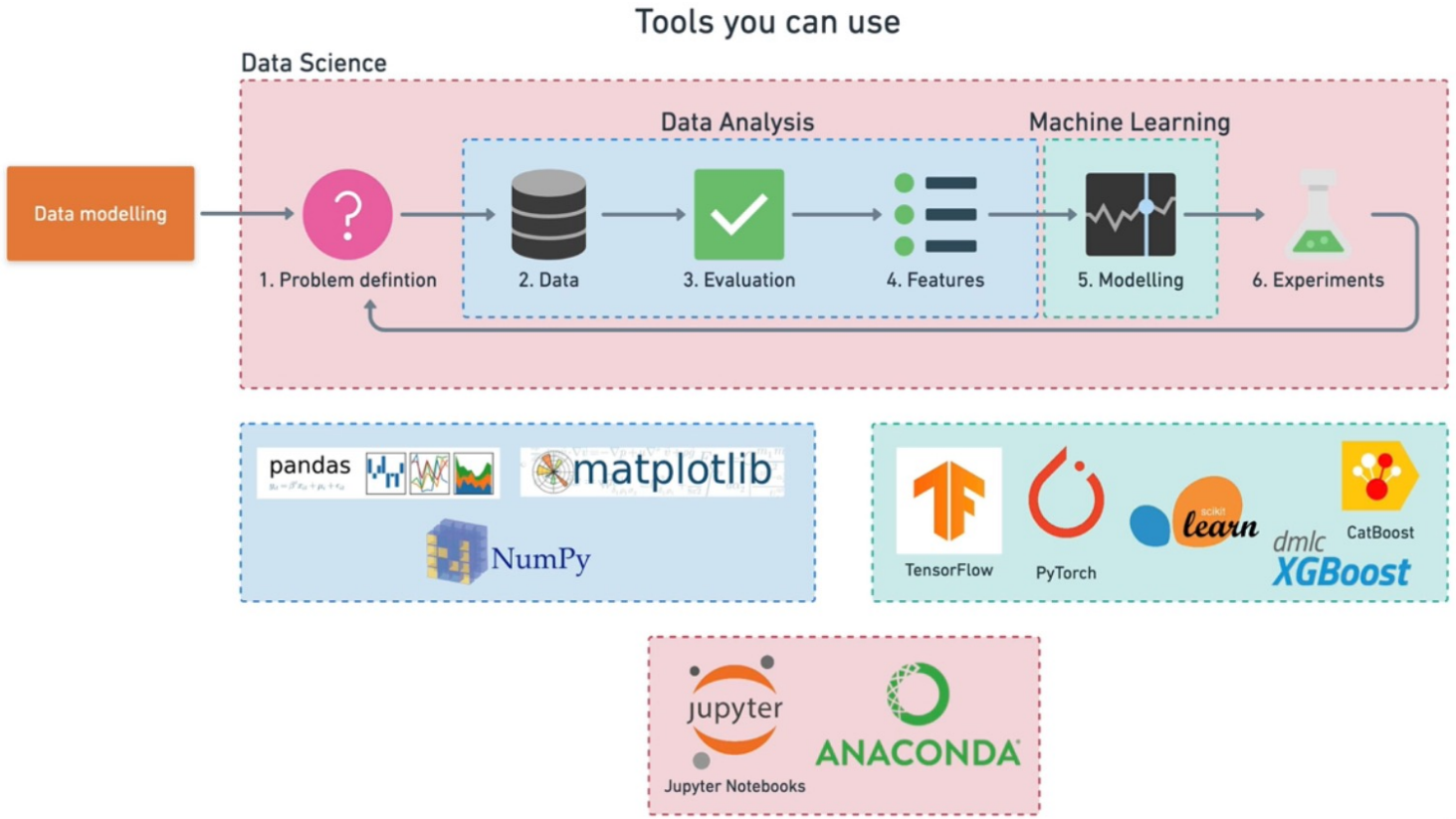
1. Create a framework

2. Match to data science and machine learning tools

3. Learn by doing

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Tool we will use



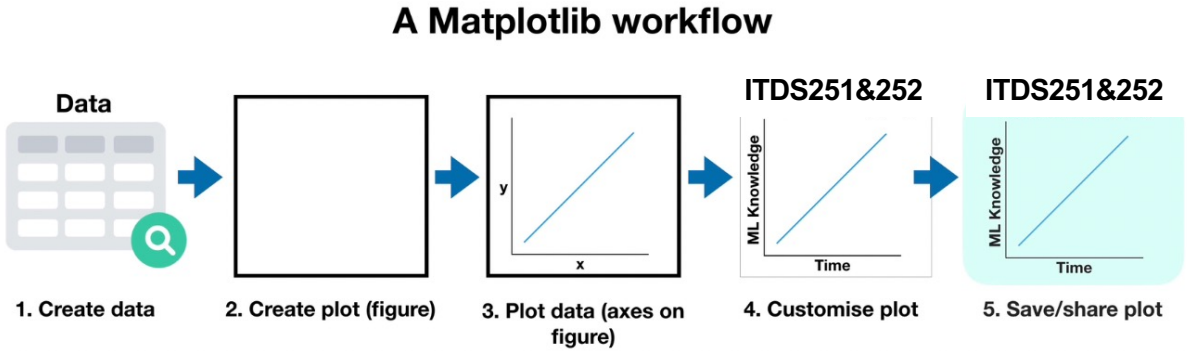
Matplotlib: Plotting and Data Visualization

Hospital needs to visualize some data. People need your help!

DEMO

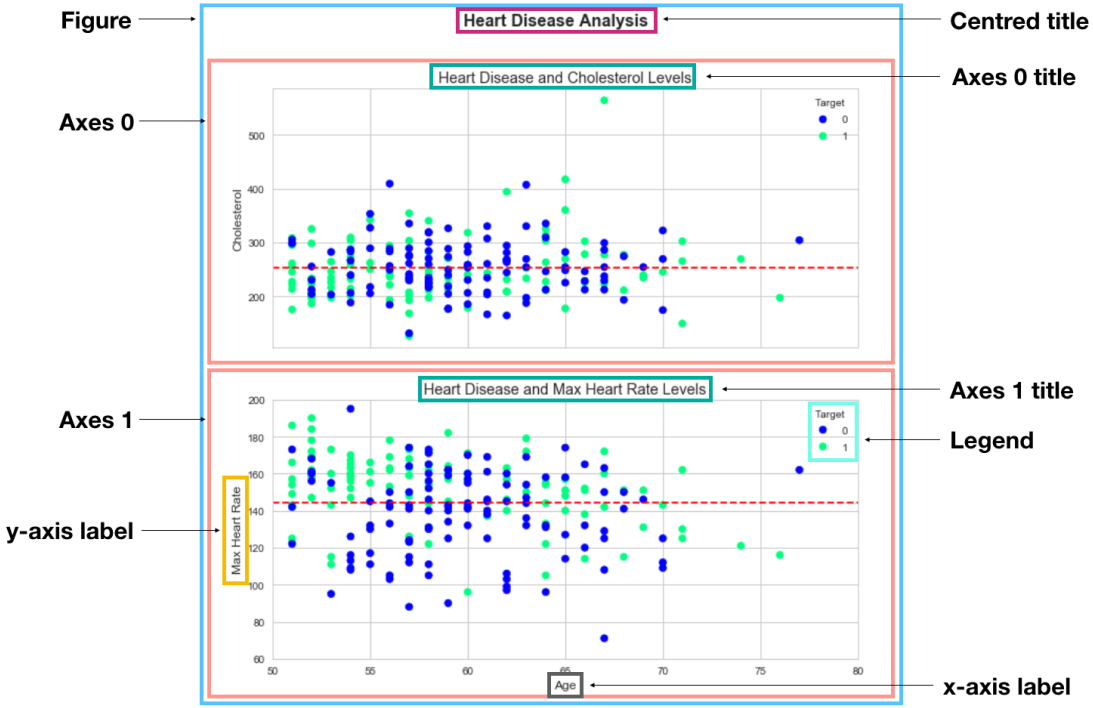
matplotlib

What are we going to cover?

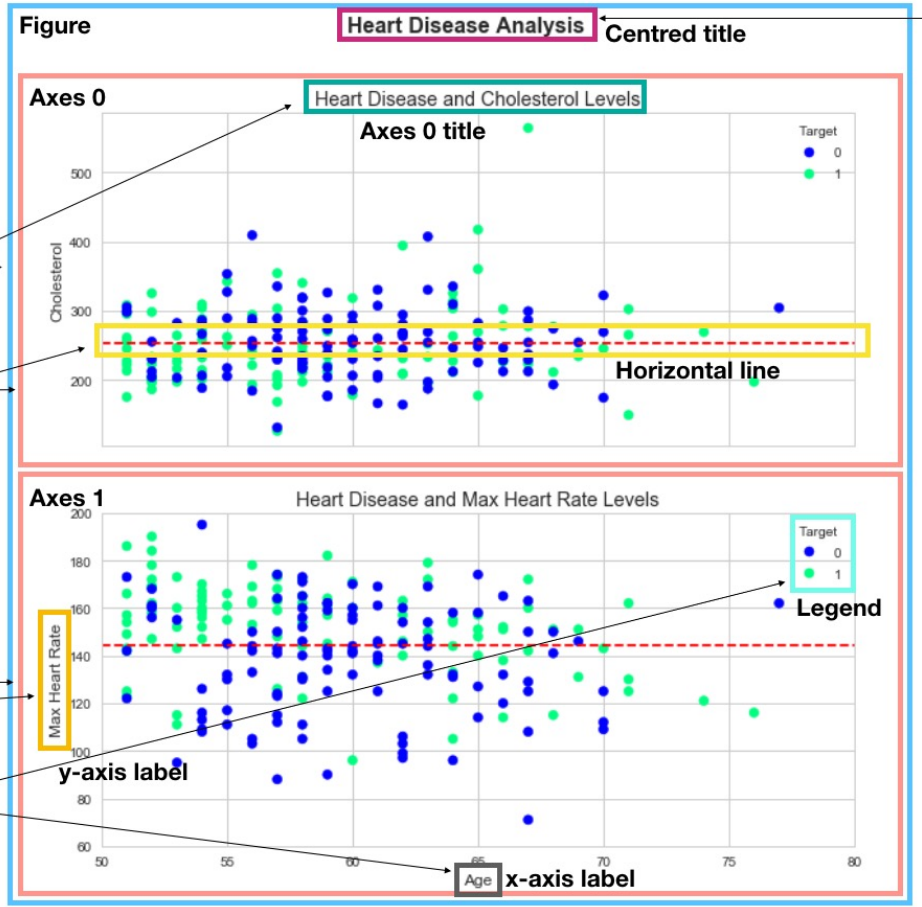


Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Anatomy of a Matplotlib Plot



```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
plt.style.use('seaborn-whitegrid') # set plot style

# Read in and manipulate data
heart_disease = pd.read_csv('../data/heart-disease.csv')
over_50 = heart_disease[heart_disease['age'] > 50]

# Create figure (plot) with 2 axes
fig, (ax0, ax1) = plt.subplots(nrows=2,
                               ncols=1,
                               sharex=True,
                               figsize=(10, 10))

# Add data, titles, meanline (axhline) and legend to axes 0
scatter = ax0.scatter(over_50["age"],
                      over_50["chol"],
                      c=over_50["target"],
                      cmap='winter')
ax0.set(title="Heart Disease and Cholesterol Levels",
        ylabel="Cholesterol",
        xlim=[50, 80])
ax0.axhline(y=over_50["chol"].mean(),
            color='r',
            linestyle='--',
            label="Average");
ax0.legend(*scatter.legend_elements(), title="Target")

# Add data, titles, meanline (axhline) and legend to axes 1
scatter = ax1.scatter(over_50["age"],
                      over_50["thalach"],
                      c=over_50["target"],
                      cmap='winter')
ax1.set(title="Heart Disease and Max Heart Rate Levels",
        xlabel="Age",
        ylabel="Max Heart Rate",
        ylim=[60, 200])
ax1.axhline(y=over_50["thalach"].mean(),
            color='r',
            linestyle='--',
            label="Average");
ax1.legend(*scatter.legend_elements(), title="Target")

# Title the figure
fig.suptitle('Heart Disease Analysis', fontsize=16, fontweight='bold');
```



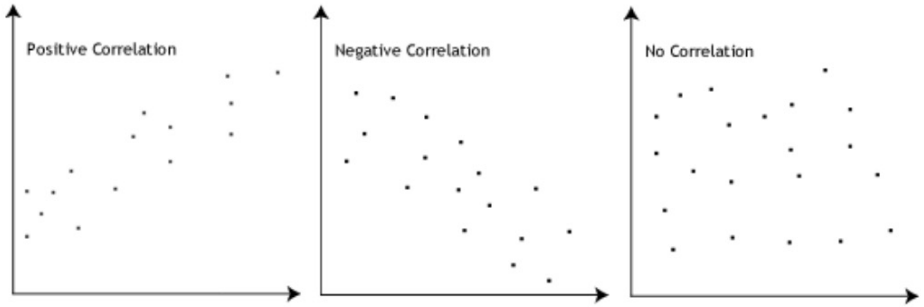
Correlations Between Attributes

WHAT IS CORRELATION?

It's a measure of how well two variables are related to each other. There are positive as well as negative correlation.

1. **Positive Correlation**: It refers to the extent to which the two variables increases or decreases in parallel (think of this as **directly proportional***, one increases other will increase, one decreases other will follow the same).
2. **Negative Correlation**: It refers to the extent to which one of the two variables increases as the other decreases (think of this as **inversely proportional***, one increases other will decrease or if one decreases other will increase).

The most common correlation in statistics is the **Pearson correlation:** the measure of the strength of **linear association** between two variables.



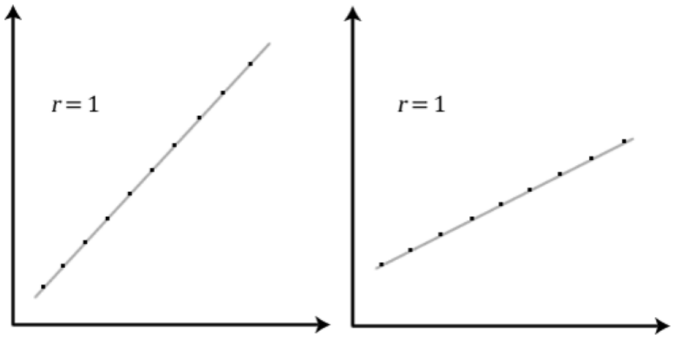
Correlations Between Attributes

But what does it really represents mathematically?

Basically, a **Pearson Product Moment Correlation (PPMC)** attempts to draw a line to best fit through the data of the given two variables, and the Pearson correlation coefficient "**r**" indicates how far away all these data points are from the line of best fit.

The value of "**r**" ranges from +1 to -1 where:

- **r**= +1/-1 represents that al our data points lie on the line of best fit only i.e there is no data point which shows any variation from the line of best fit.

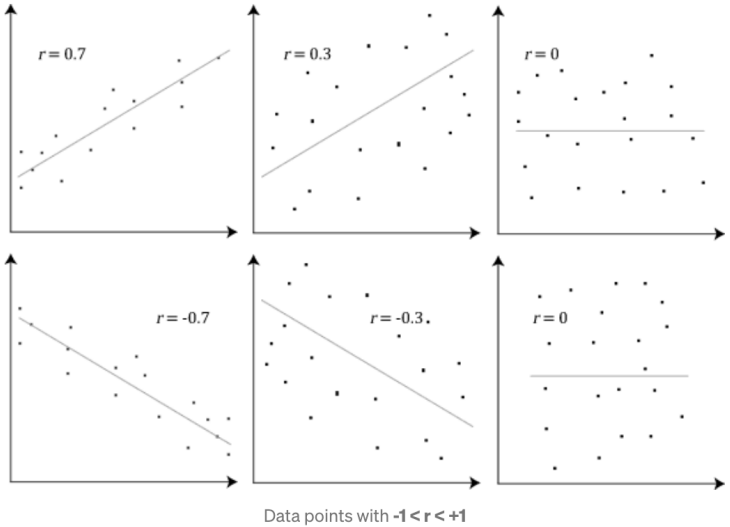


Data points with r=1

Correlations Between Attributes

- Hence, the stronger the association between the two variables, the closer **r** will be to +1/-1.
- **r** = 0 means that there is no correlation between the two variables.
- The values of **r** between +1 and -1 indicate that there is a variation of data around the line.
- The closer the values of **r** to 0, the greater the variation of data points around the line of best fit.

It is also important to realize that the value of Pearson's coefficient, **r**, is not a measure of the slope of the line (i.e the line of best fit). We can see an example in the plot above with **r=1.**



Correlations Between Attributes

Formula of Pearson Correlation coefficient:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

An example with calculating Pearson Coefficient:

Find the value of the correlation coefficient from the following table:

SUBJECT	AGE (X)	GLUCOSE LEVEL (Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Age and Glucose levels of 6 subjects

We'll calculate the value of **r** using the formula mentioned above. For using that formula we need to compute $\Sigma(X*Y)$, $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(X^2)$, $\Sigma(Y^2)$.

The table below shows the computed values of all the summations mentioned above.

SUBJECT	AGE (X)	GLUCOSE LEVEL (Y)	X*Y	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Correlations Between Attributes

Formula of Pearson Correlation coefficient:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

We'll calculate the value of **r** using the formula mentioned above. For using that formula we need to compute **Σ(X*Y), Σ(X), Σ(Y), Σ(X²), Σ(Y²)**.

The table below shows the computed values of all the summations mentioned above.

SUBJECT	AGE (X)	GLUCOSE LEVEL (Y)	X*Y	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

From our table we get:

- **Σ(X) = 247**
- **Σ(Y) = 486**
- **Σ(X*Y) = 20,485**
- **Σ(X²) = 11,409**
- **Σ(Y²) = 40,022**
- **n is the sample size, in our case = 6**

r = 6(20,485) — (247 × 486) / [√[[6(11,409) — (24^{72})] × [6(40,022) — 48^{62}]]]

r = 0.5298.

The range of the correlation coefficient is from **-1** to **+1**. Our result is **0.5298** or **52.98%**, which means the variables have a **moderate positive correlation**.

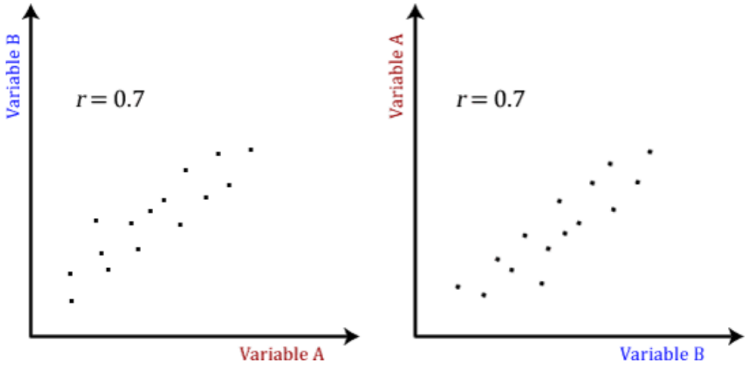
Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]

Problems with Pearson correlation? (Potential)

The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally.

Example-1 If we are trying to find the correlation between a high-calorie diet and diabetes, we might find a high correlation of **.8**. However, we could also get the same result with the variables switched around. In other words, we could say that diabetes causes a high-calorie diet.

Example-2 We might want to find out whether basketball performance is correlated with a person's height. We might, therefore, plot a graph of performance against height and calculate the Pearson correlation coefficient. Let's say, for example, that $r = .67$. That is, as height increases so do basketball performance. This makes sense. However, if we plotted the variables the other way around and wanted to determine whether a person's height was determined by their basketball performance (which makes no sense), we would still get $r = .67$. This is because the Pearson correlation coefficient makes no account of any theory behind why you chose the two variables to compare. This is illustrated below:

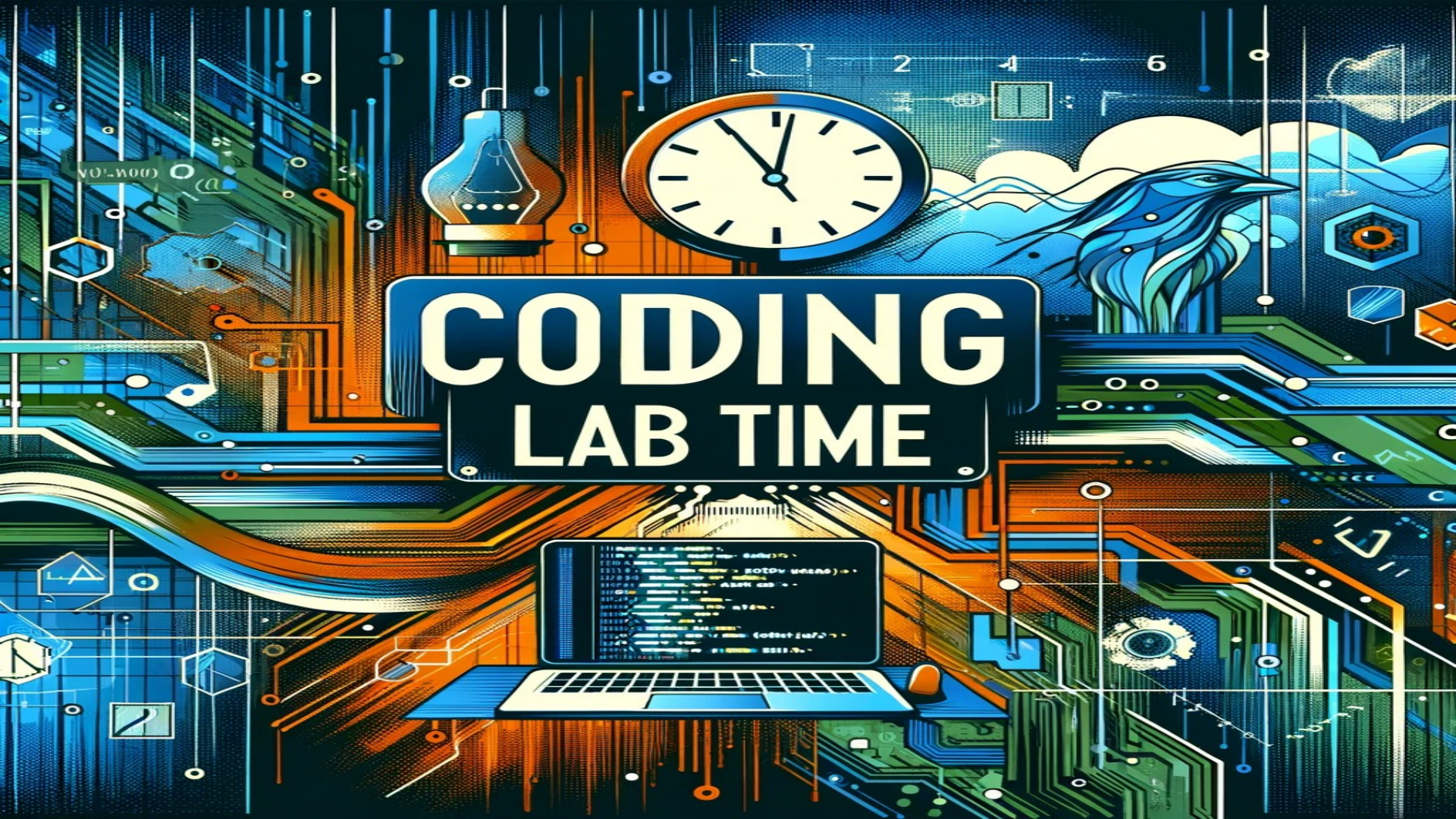


Correlations Between Attributes

Quiz:

1) เมื่อเวลาที่นักเรียนใช้ในการเรียนเพิ่มขึ้น คะแนนสอบเฉลี่ยของเขาก็เพิ่มขึ้นด้วย
2) เมื่อพนักงานได้รับเงินเดือนสูงขึ้น ประสิทธิภาพการทำงานก็เพิ่มขึ้น
3) นักเรียนที่มีการขาดเรียนบ่อย จะมีคะแนนลดลง
4) ยิ่งผู้ชายมีอายุมากขึ้น ผมของเขาก็จะน้อยลง

Contact: tipajin.tha@mahidol.edu. [Credited Tipajin T.]


CODING
LAB TIME