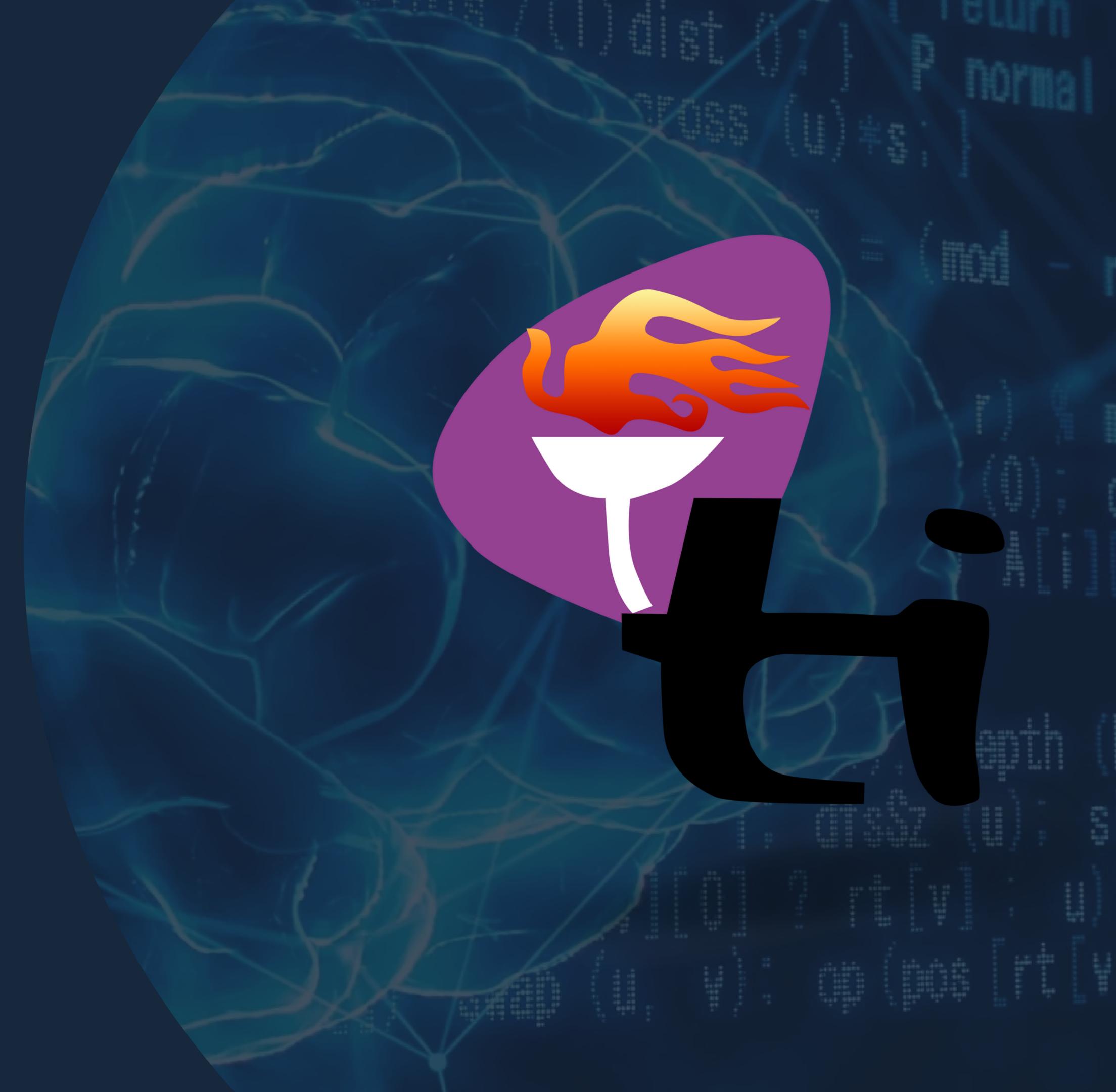


Workshop

Fundamental AI / ML

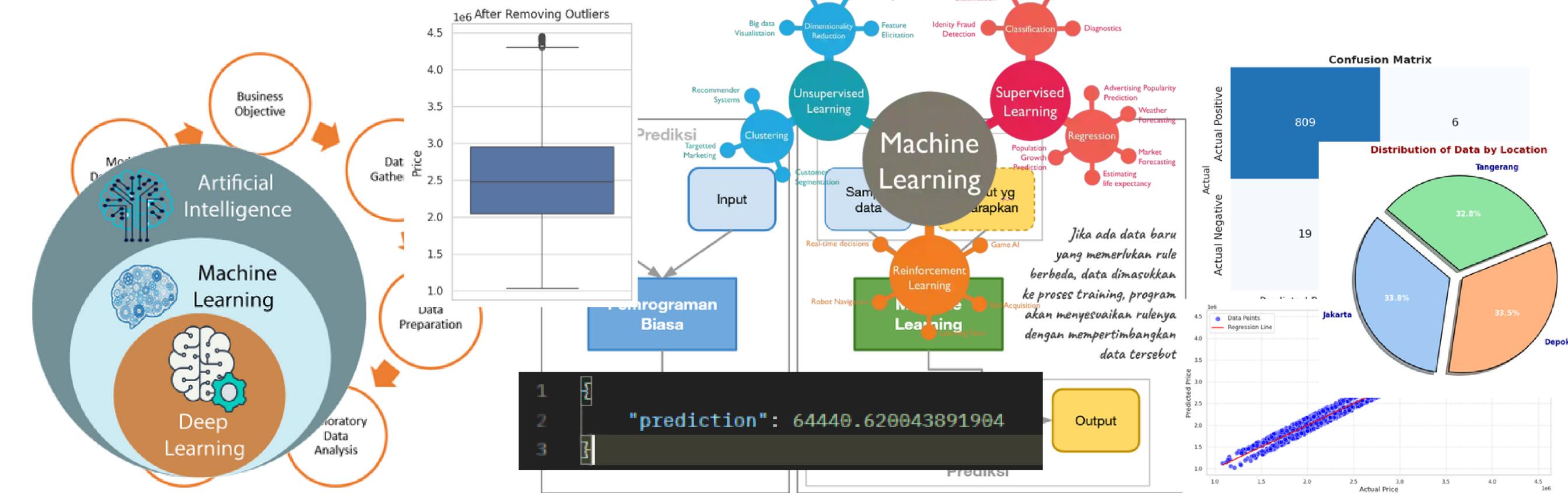
Laboratorium Informatika



Overview

Memahami fundamental - fundamental AI/ML, **mulai dari apa itu data**, apa itu dataset, apa itu AI dan machine learning, bagaimana mesin bisa belajar dari data, apa itu model sampai **implementasi linear regression model** untuk prediksi.

Apaan tuh ?



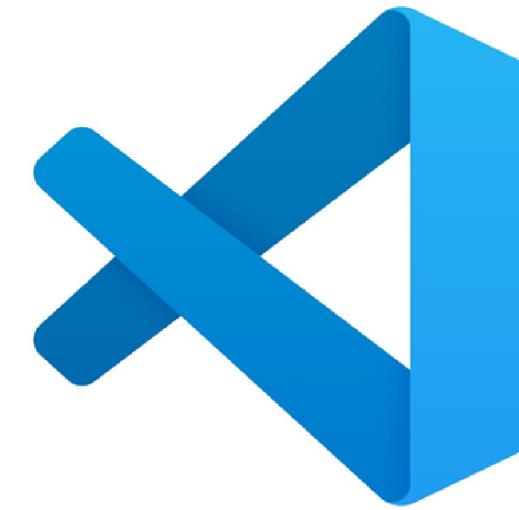
Problem Questions

1. Apa itu Data ?
2. Apa itu Machine Learning / AI ?
3. Apa itu Model Machine Learning ?
4. Apa dan kenapa harus Regresi Linear ?
5. Bagaimana cara implementasinya ?

Tools



1 . Google Colab



2 . VSCode



POSTMAN

3 . Postman



Q1

Apa itu Data ?

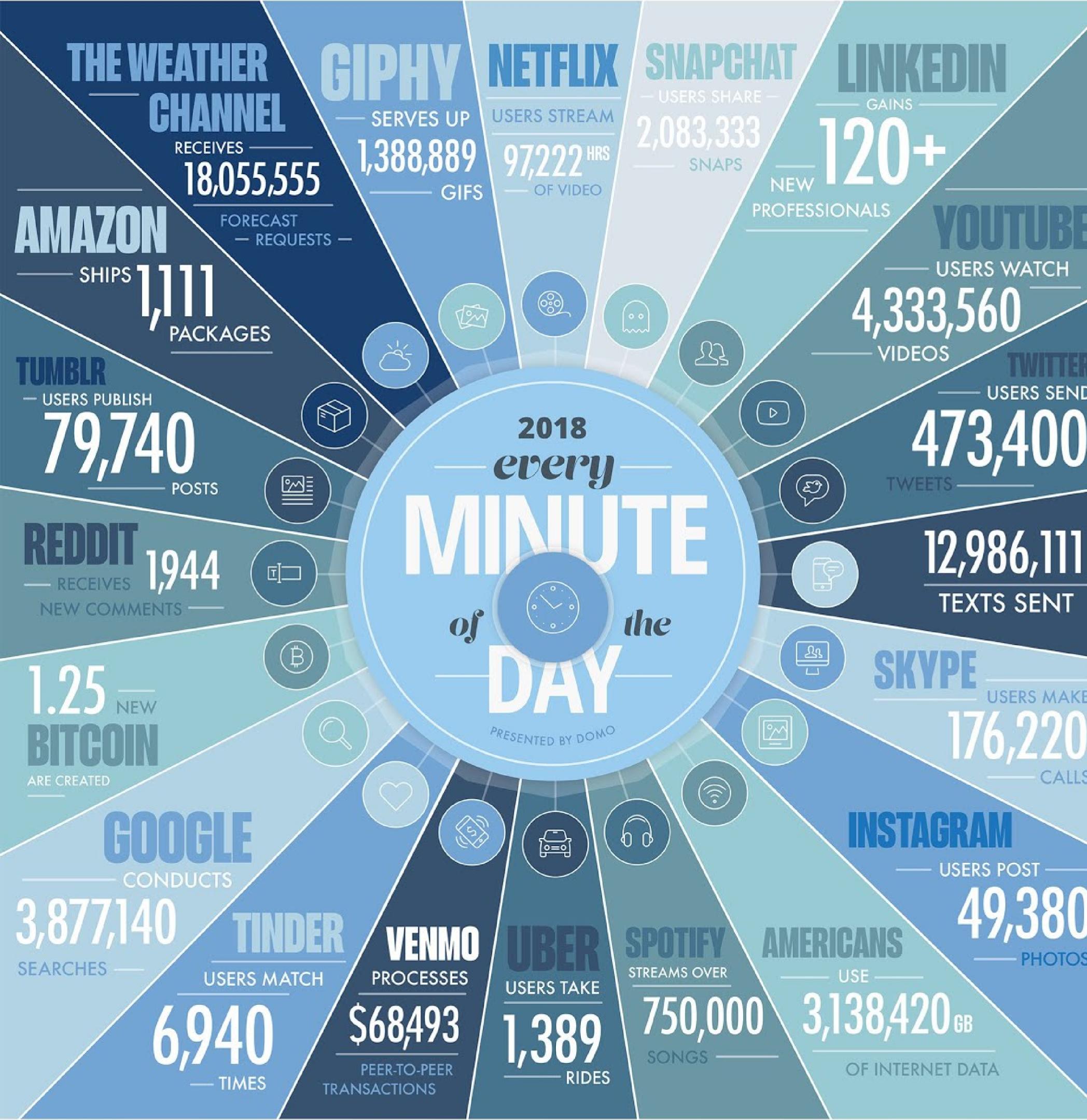
Definisi Data

Data adalah **kumpulan fakta atau informasi yang dikumpulkan** melalui pengamatan, pengukuran, atau eksperimen. **Data dapat berupa angka, teks, gambar, atau suara, dan biasanya disusun dalam format tertentu** yang memungkinkan untuk diproses lebih lanjut, seperti tabel, grafik, atau basis data.



"In the context of information technology, **data is the raw input that is processed** to provide meaningful information." (Sumber: Journal of Information Technology)

"Data refers to the representation of **facts, concepts, or instructions in a formalized manner** suitable for communication, interpretation, or processing by human or **automatic means**." (Sumber: IEEE Standard Glossary of Software Engineering Terminology)



Data is Never Sleep !



Data

Qualitative

Quantitative



Discrete
Data

Continuous
Data

- 5 kids
- 96 workers
- 3 laptops

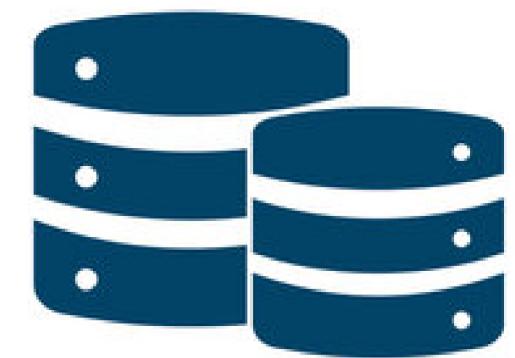
- 3.25 kg
- 1.32 miles
- 7.25 inches

Perbedaan Data Diskrit VS Data Kontinu

Data Diskrit (<i>Discrete Data</i>)	Data Kontinu (<i>Continuous Data</i>)
Berbentuk bilangan bulat	Berbentuk bilangan pecahan atau desimal
Memiliki nilai tetap dan tak bisa dibagi	Rentang nilai luas karena bisa dibagi
Bisa dihitung, namun tak bisa diukur	Bisa diukur, namun tak bisa dihitung
Digambarkan terutama dalam grafik	Direpresentasikan dalam bentuk histogram

Definisi Dataset

Dataset adalah kumpulan data yang disusun dalam format yang terstruktur atau semi-struktur untuk tujuan analisis atau pengolahan lebih lanjut. **Dataset biasanya terdiri dari entitas data yang terorganisir dalam baris (observasi) dan kolom (atribut atau fitur), dimana setiap baris mewakili satu instance atau sampel data, dan setiap kolom mewakili atribut atau variabel yang menggambarkan fitur dari instance tersebut.**



DATASET

Menurut Journal of Data Science: "Dataset adalah kumpulan nilai dari variabel kualitatif atau kuantitatif yang terdiri dari satu set item yang digunakan sebagai dasar untuk penalaran, diskusi, atau perhitungan."

Kaggle Dataset

≡ kaggle

- + Create
- ⌚ Home
- 🏆 Competitions
- 💾 Datasets
- 👤 Models
- Code
- 💬 Discussions
- 🎓 Learn
- ⌄ More

Your Work

VIEWED

- Indonesia President...
- House Prices - Adv...
- Data Analytics Com...
- Seleksi Data Scienc...

View Active Events

Search

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset Your Work

Search datasets

All datasets Computer Science Education Classification Computer Vision NLP Data Visualization Pre-Trained Model

Filters

Trending Datasets

See All

A thumbnail for the "ICC Men's T20 World Cup 2024 Matches" dataset, featuring the tournament logo and a stadium background.

ICC Men's T20 World Cup 2024 Matches

Blueboy · Updated 2 days ago

Usability **10.0** · 2 kB

1 File (CSV)

A thumbnail for the "Employee Attrition data prediction" dataset, showing people working in an office.

Employee Attrition data prediction

MrSimple · Updated 12 days ago

Usability **10.0** · 24 kB

1 File (CSV)

A thumbnail for the "Data Science Jobs & Salaries 2024" dataset, featuring a green board with "DATA SCIENCE" written on it.

Data Science Jobs & Salaries 2024

Ritik Sharma · Updated 20 days ago

Usability **10.0** · 310 kB

1 File (CSV)

A thumbnail for the "Exam Score and Grant Data of Erasmus Applicants" dataset, featuring a collage of European landmarks like Big Ben and the Eiffel Tower.

Exam Score and Grant Data of Erasmus Applicants

Eren Acar · Updated 14 days ago

Usability **10.0** · 6 kB

1 File (CSV)

Contoh Dataset



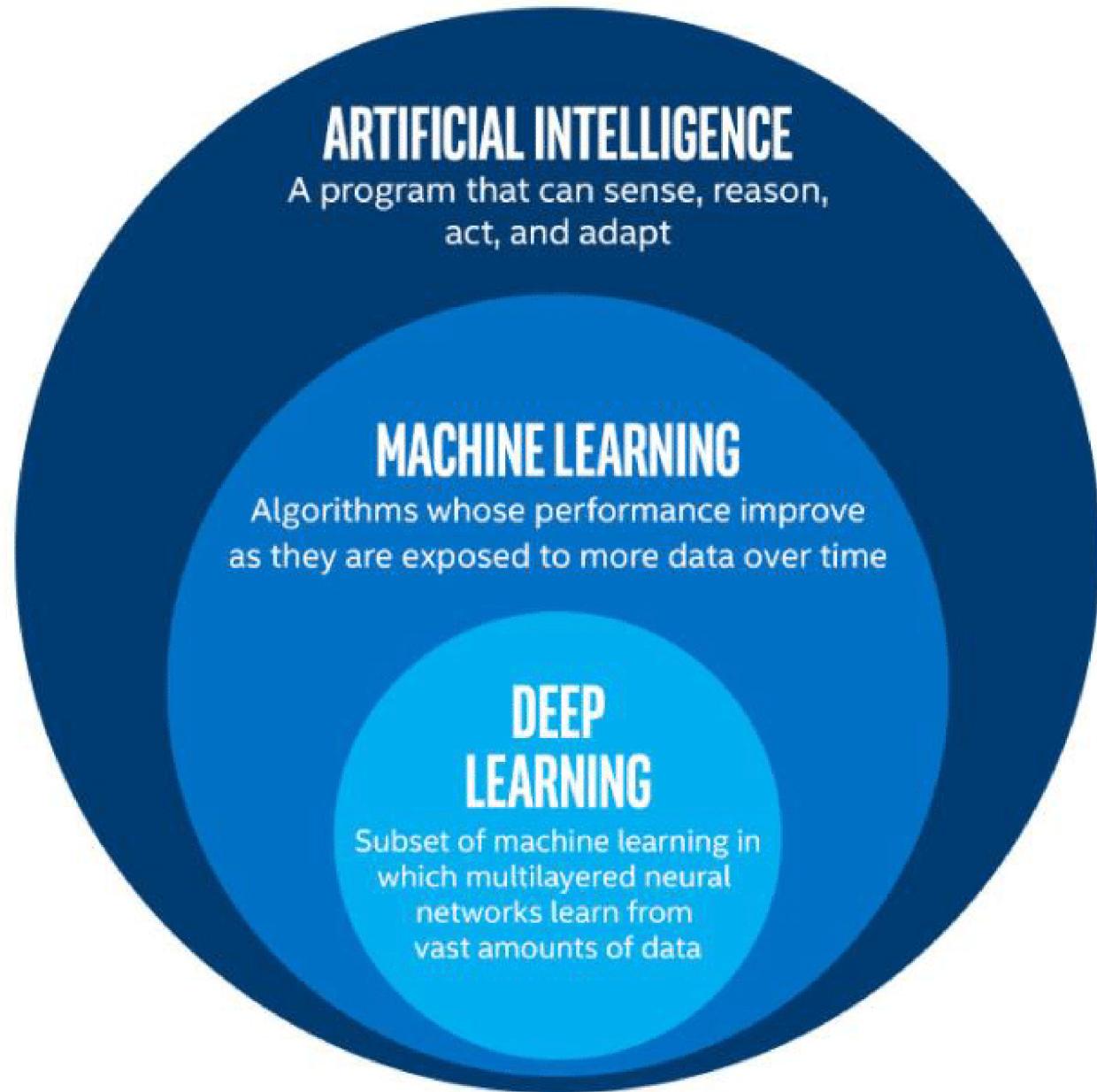
Q2

**Apa itu Machine
Learning / AI ?**

Definisi AI & ML

Artificial Intelligence (AI) bidang ilmu komputer yang bertujuan untuk membuat mesin cerdas yang dapat melakukan tugas yang biasanya memerlukan kecerdasan manusia. Ini termasuk pemrosesan bahasa alami, pengenalan gambar, dan pengambilan keputusan.

Machine Learning (ML) adalah cabang dari AI yang fokus pada pengembangan sistem komputer yang dapat belajar dari data dan meningkatkan kinerjanya dari waktu ke waktu tanpa perlu diprogram ulang secara eksplisit. ML memungkinkan komputer untuk mengidentifikasi pola yang rumit dan membuat prediksi berdasarkan data.

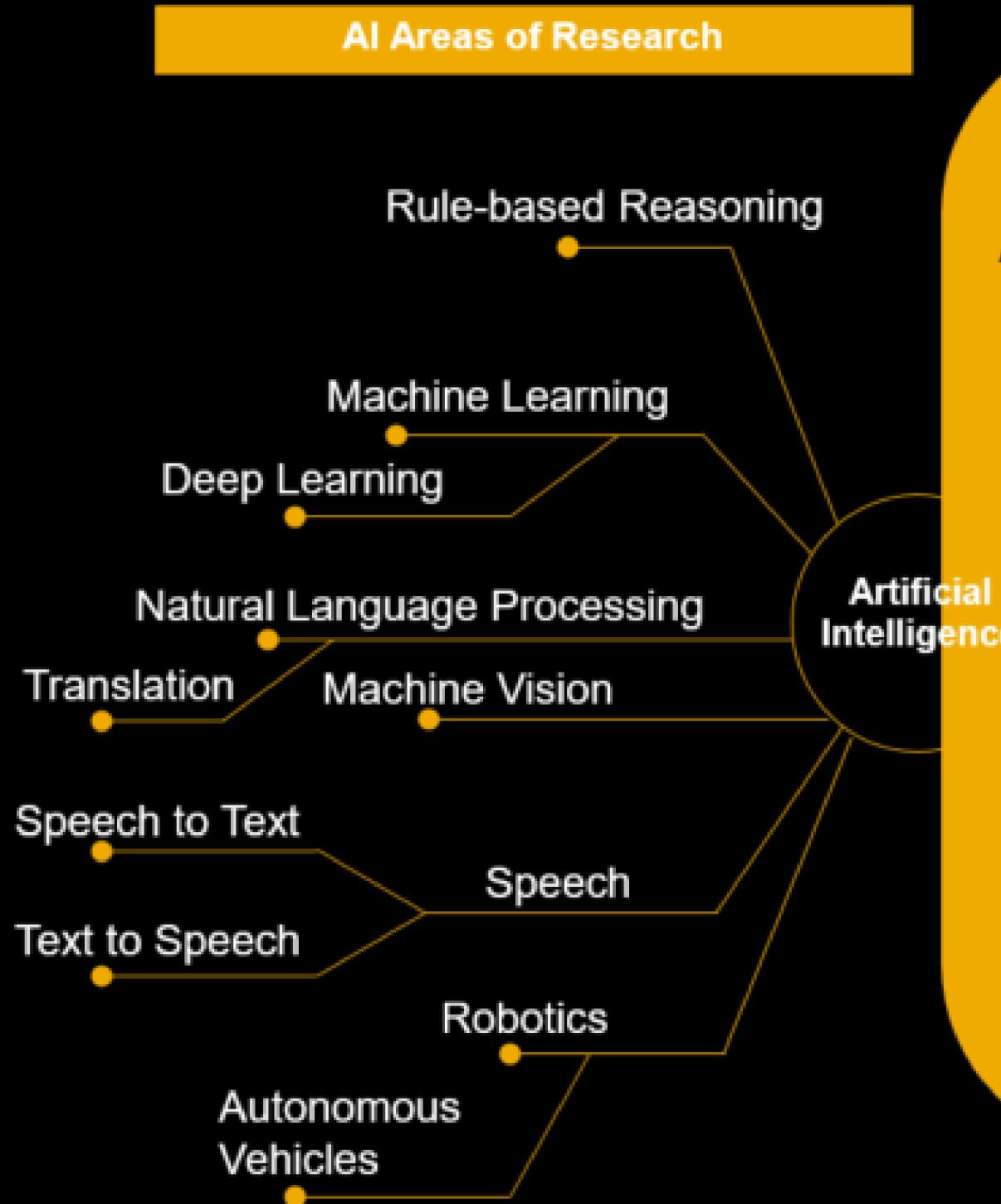


Definisi AI & ML

Perbedaan antara AI dan ML:

- **AI adalah konsep umum yang mencakup pengembangan mesin cerdas yang dapat meniru kecerdasan manusia dalam berbagai konteks.**
- **ML, sebagai sub-bidang AI, khususnya berfokus pada penggunaan data untuk melatih sistem komputer agar dapat memahami pola dan membuat keputusan atau prediksi.**

Evolutions



Artificial Intelligence (AI)

Human Intelligence Exhibited by Machines

Amazon purchase prediction Smart Email Categorization

Machine Learning (ML)

An Approach to Achieve Artificial Intelligence

Google Maps speed of traffic Facebook facial recognition
Netflix video recommendation

Deep Learning (DL)

A Technique for Implementing Machine Learning

Self-Driving Cars
Speech Recognition Robotics

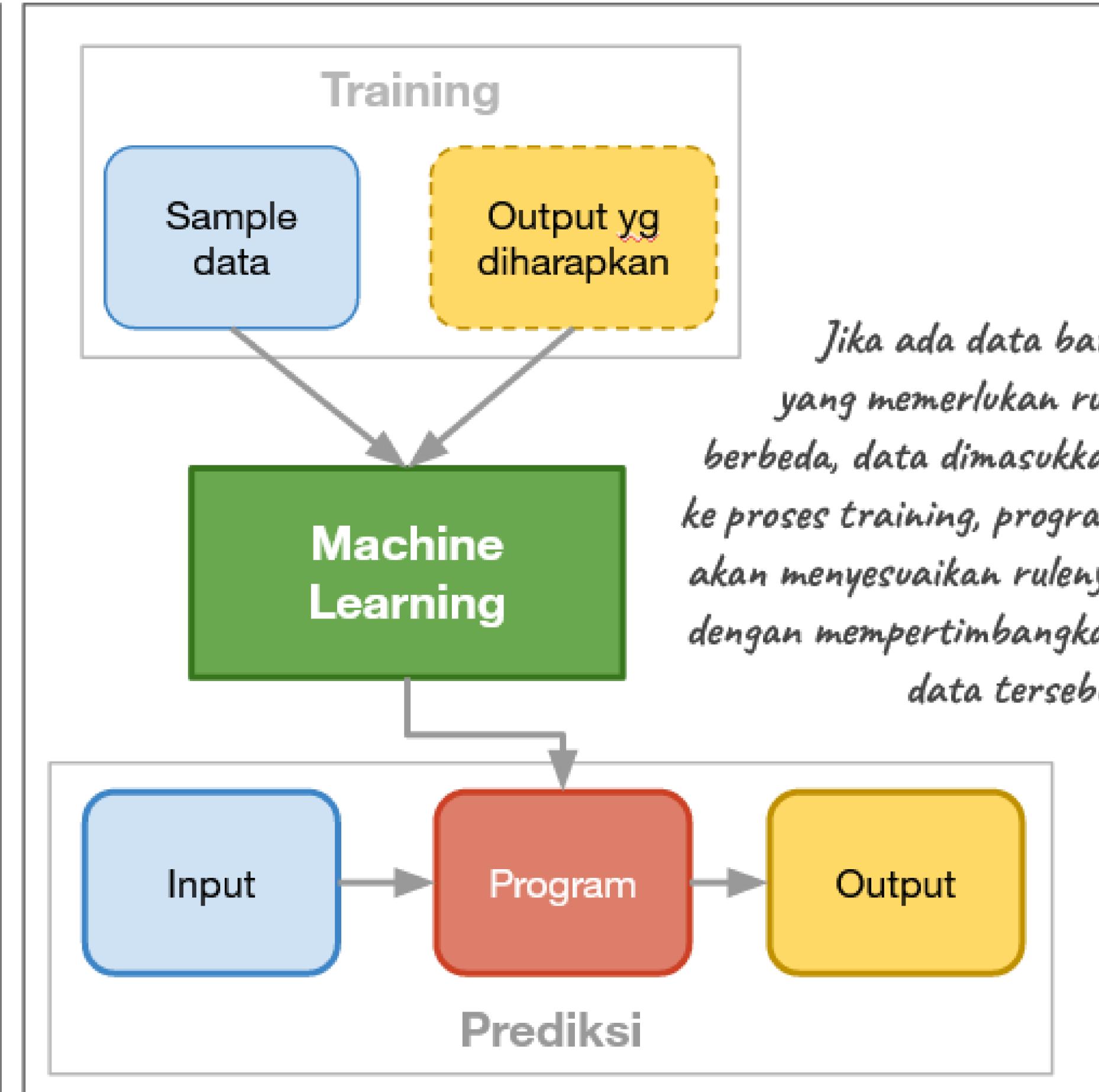
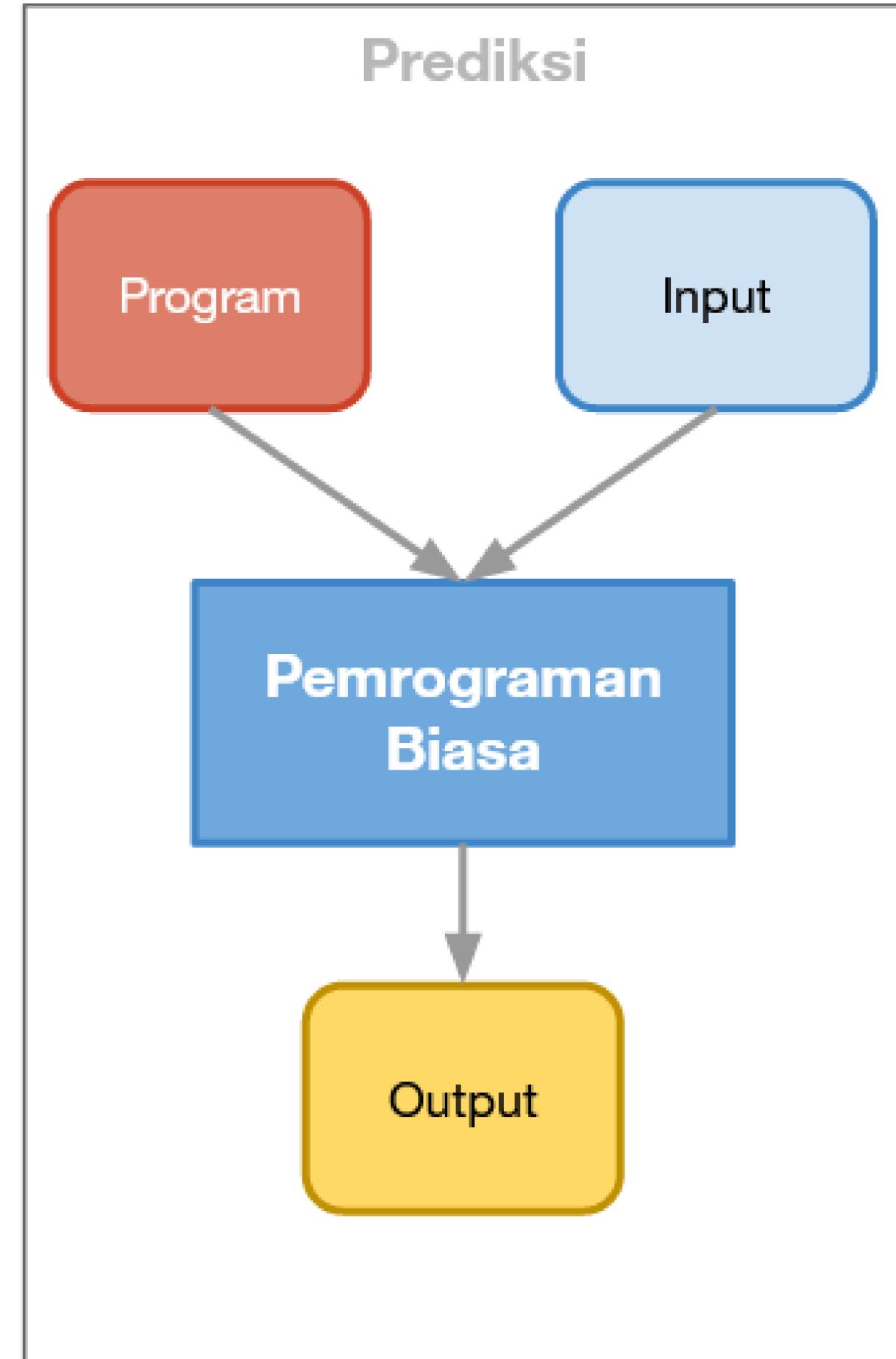
Data Science

Scientific methods, algorithms and systems to extract knowledge or insights from big data

Perbedaan ML dengan Tradisional

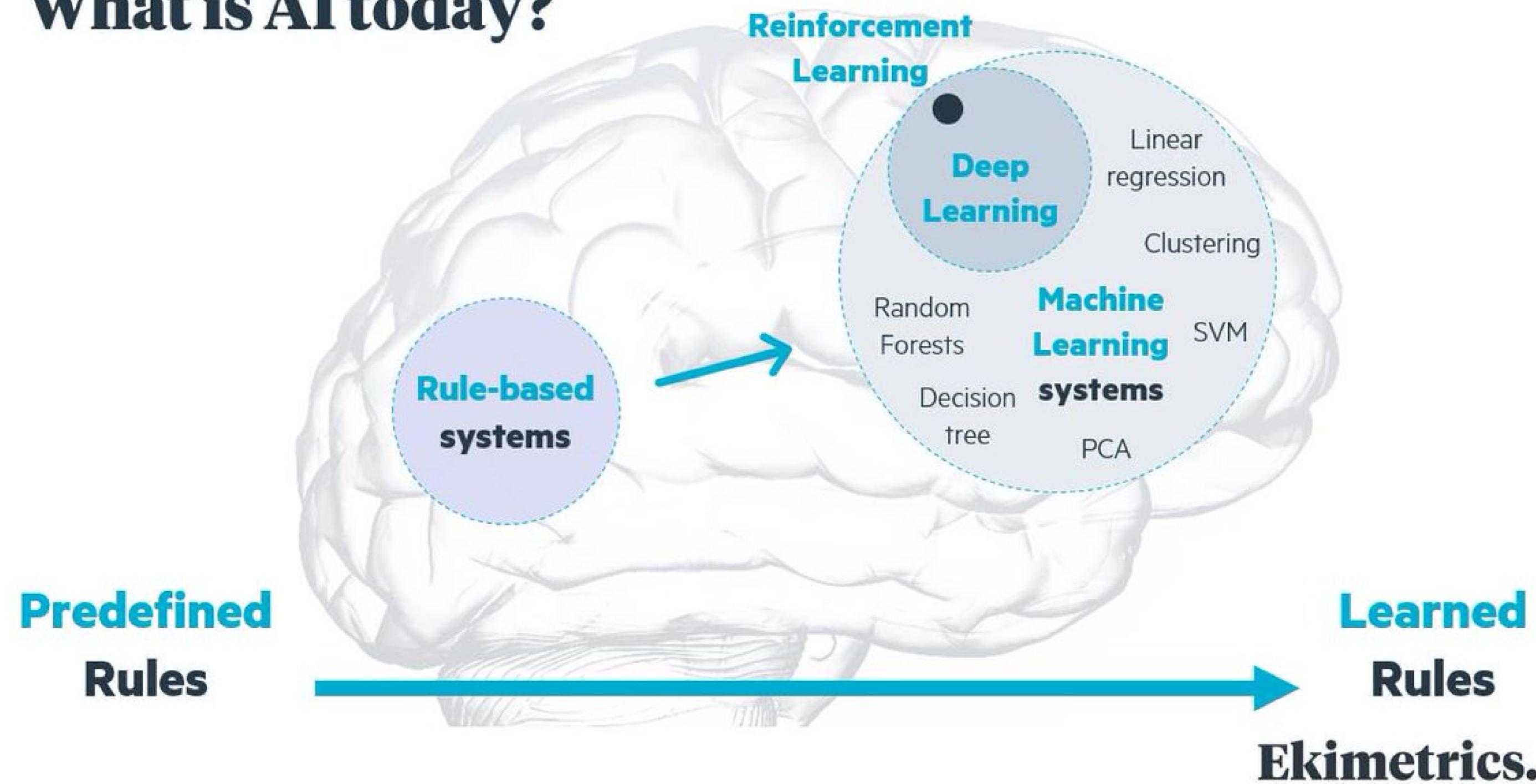


Jika ada data baru yang memerlukan rule berbeda, harus dilakukan perubahan manual pada program.

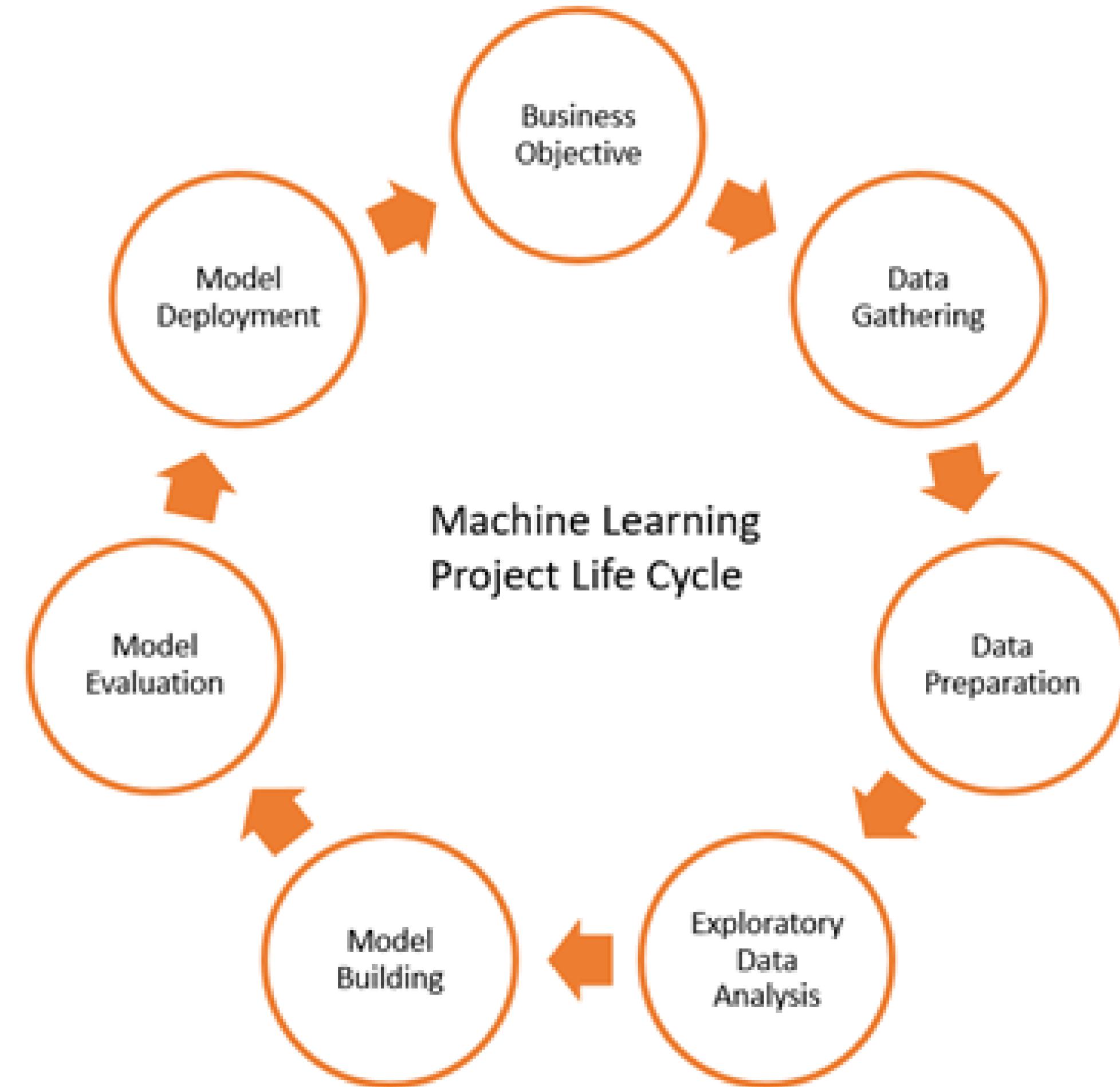


Evolutions

What is AI today?









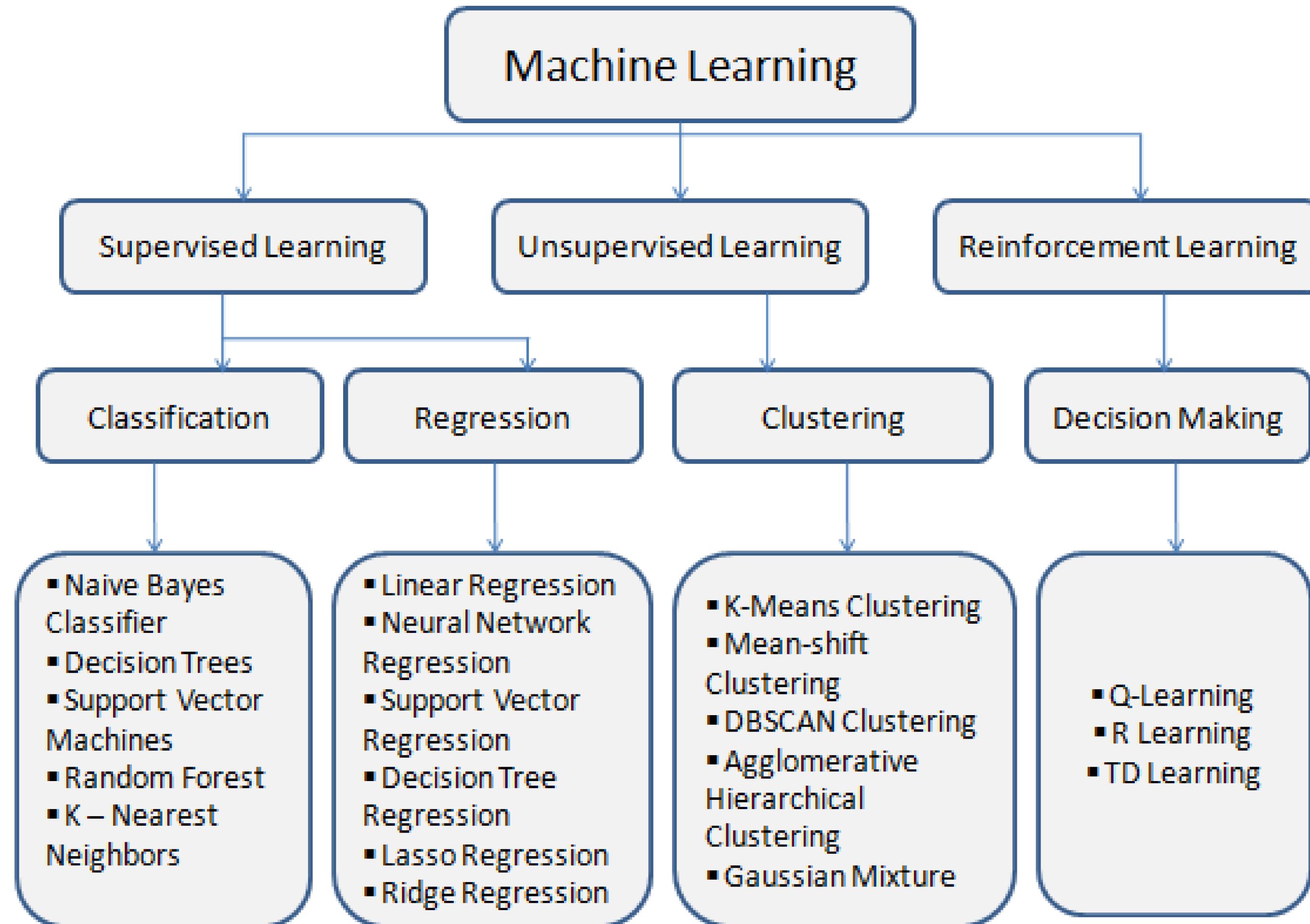
Q3

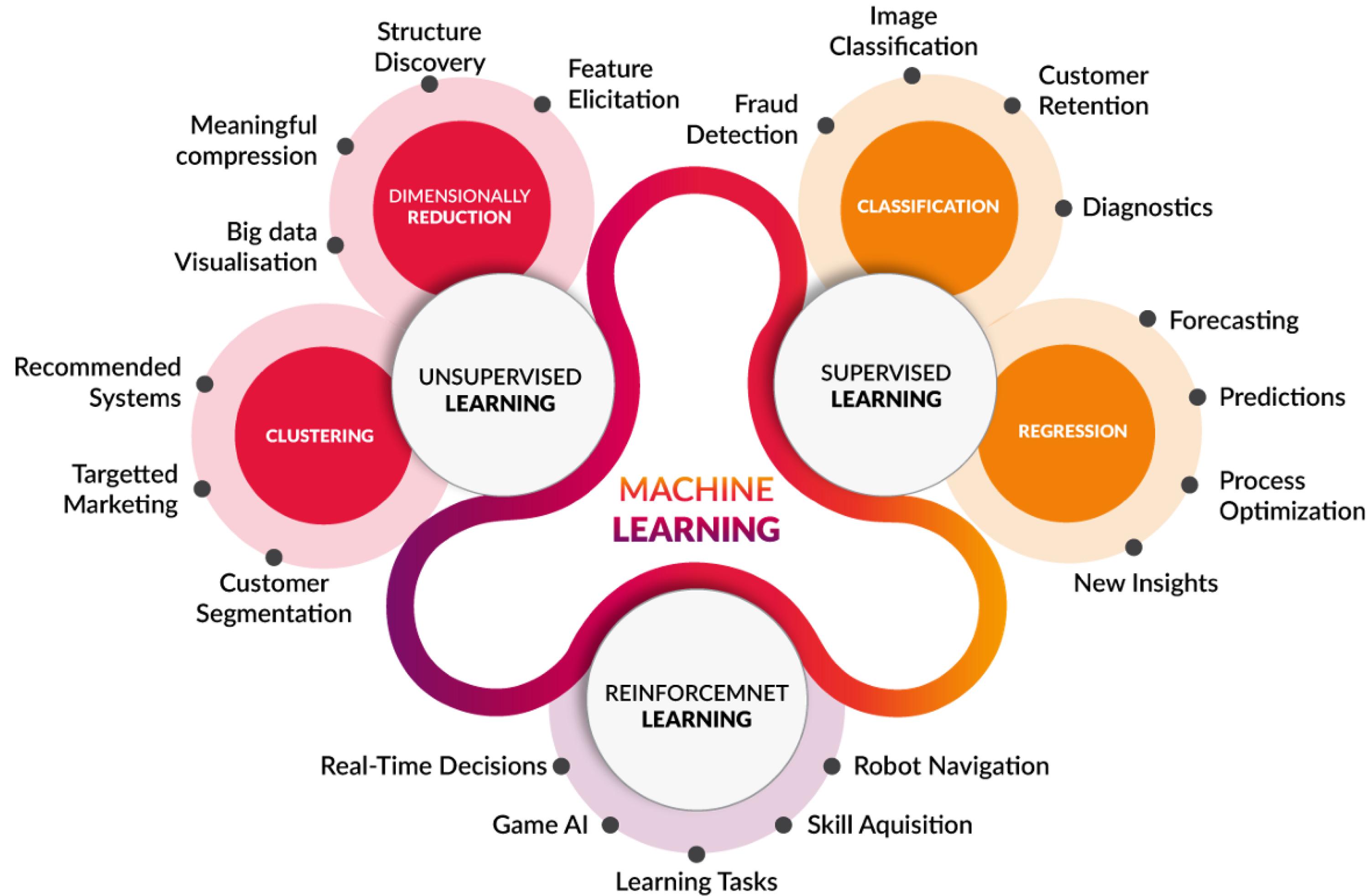
**Apa itu Model
Machine Learning ?**

Model ML

Model Machine Learning adalah **alat yang mempelajari pola dari data untuk membuat prediksi atau keputusan**. Proses melibatkan pelatihan model menggunakan data historis, menguji kinerjanya, dan kemudian menggunakan model **untuk memprediksi atau mengambil keputusan berdasarkan data baru**.

Model ini sangat berguna dalam berbagai aplikasi, seperti pengenalan gambar, prediksi cuaca, deteksi penipuan, dan banyak lagi.



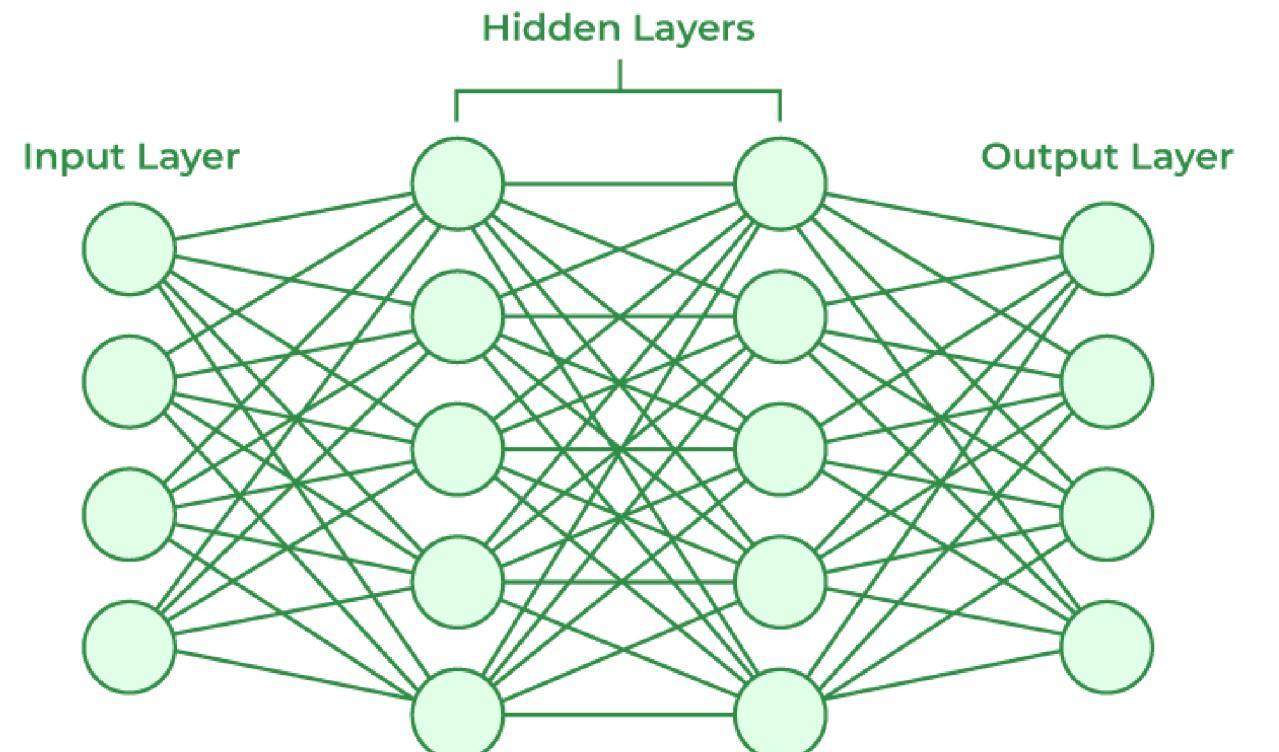


Build in Model

```
from sklearn.linear_model import LinearRegression  
LinearRegression()  
[+] ▾ LinearRegression  
LinearRegression()
```

Auto ML

Neural Network



Cost Function

Cost function (fungsi biaya) adalah **fungsi matematis yang mengukur seberapa buruk atau bagus performa model machine learning** dalam memprediksi nilai target yang sebenarnya dari data latih. Tujuan utama dari cost function adalah untuk memberikan representasi numerik tentang seberapa baik model memetakan data input ke output yang diharapkan. Untuk Regresi Linear, cost function yang umum digunakan adalah **Mean Squared Error (MSE)** atau **Mean Absolute Error (MAE)**.

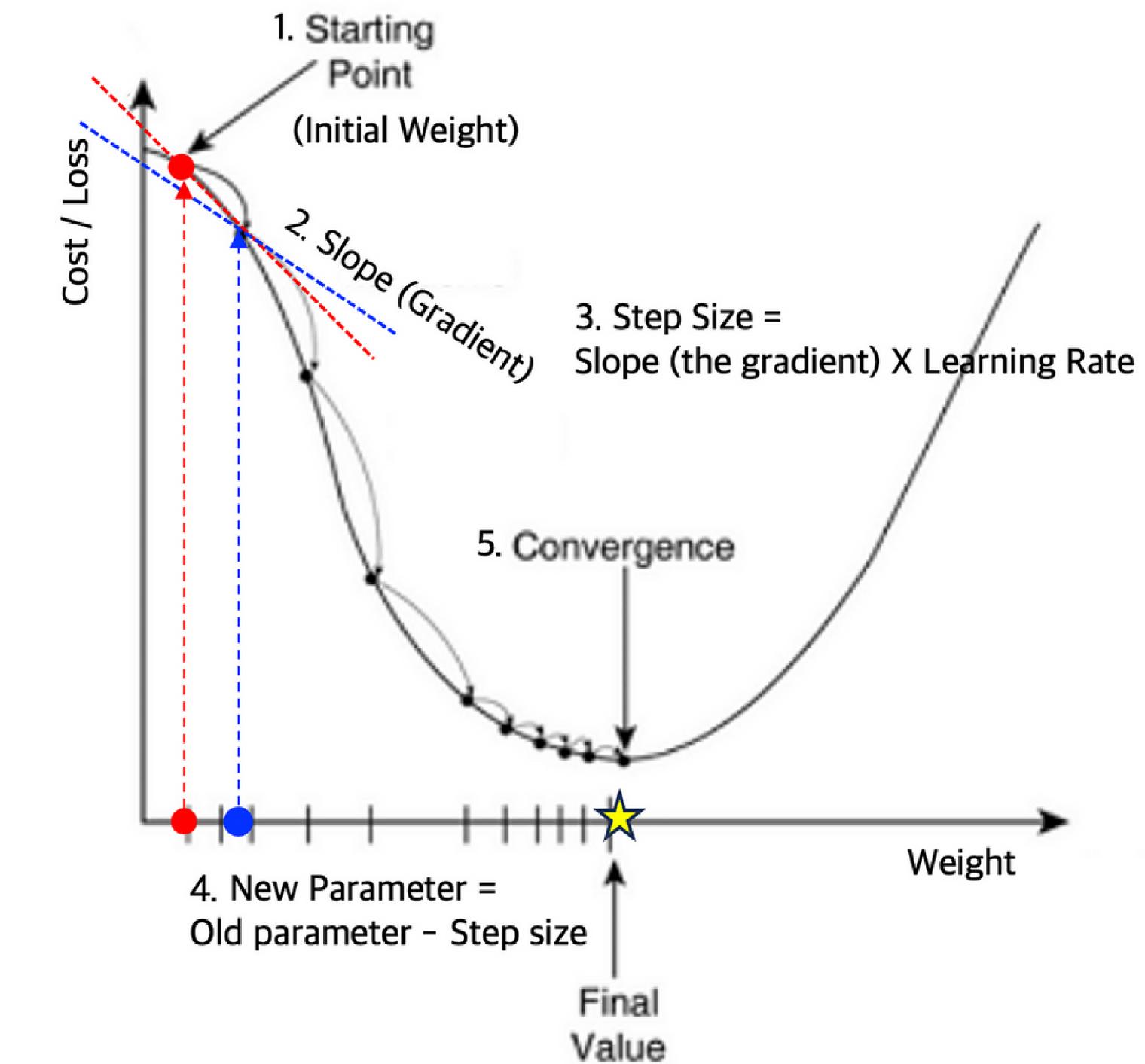
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

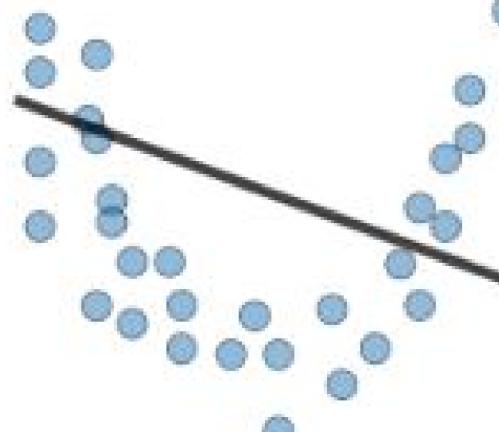
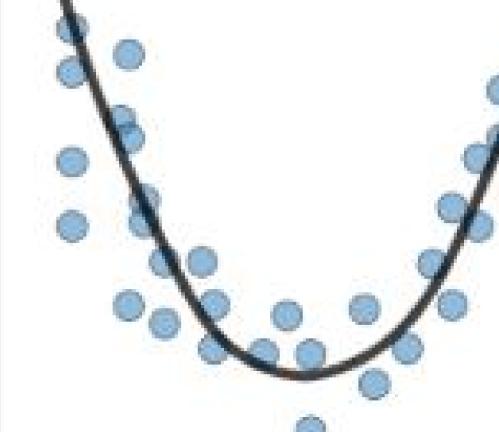
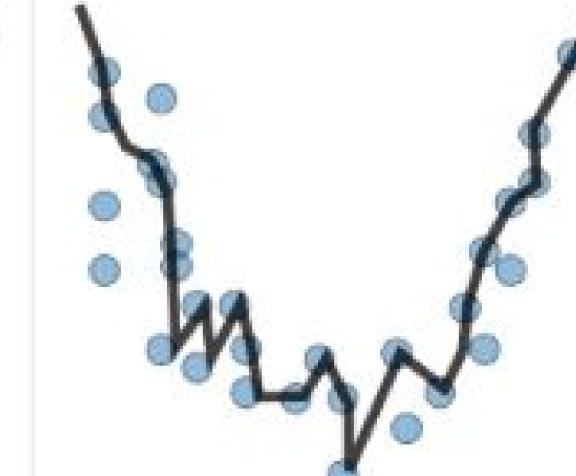
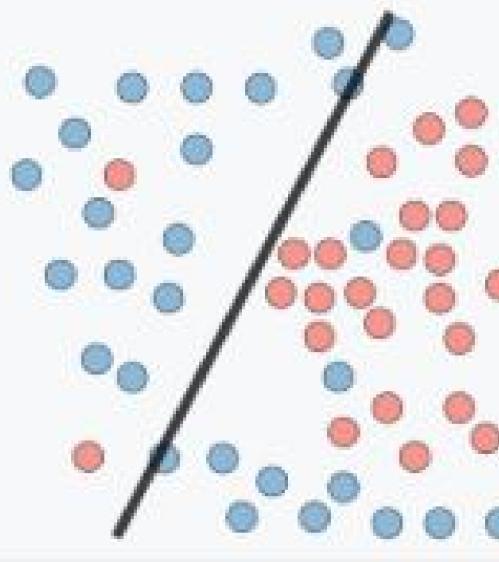
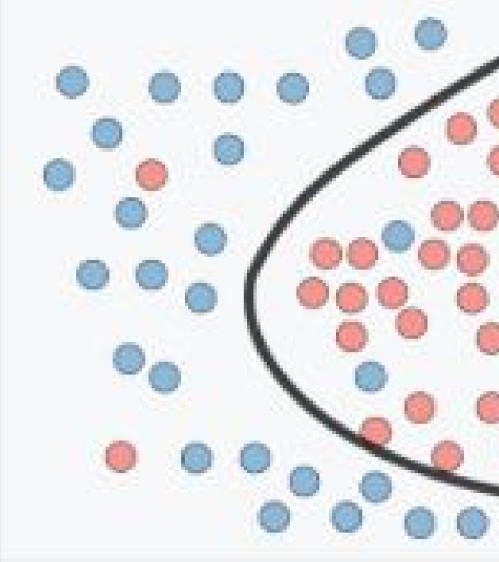
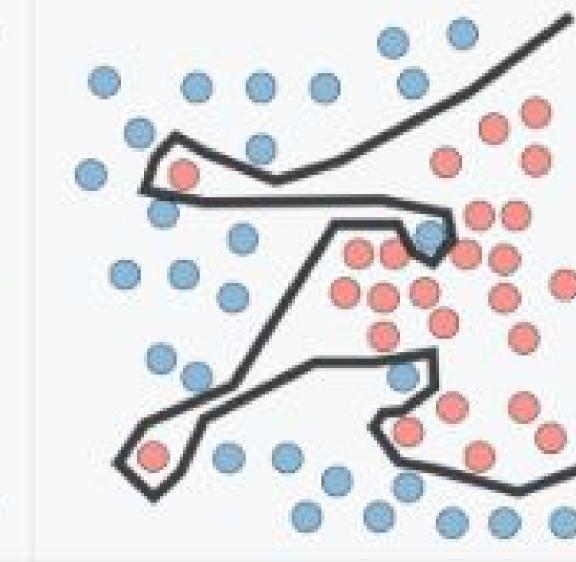
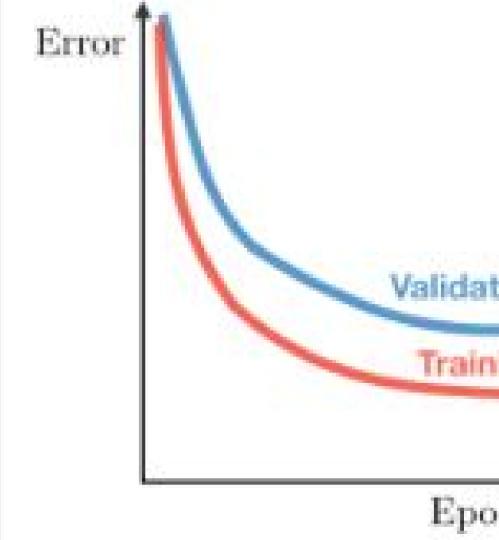
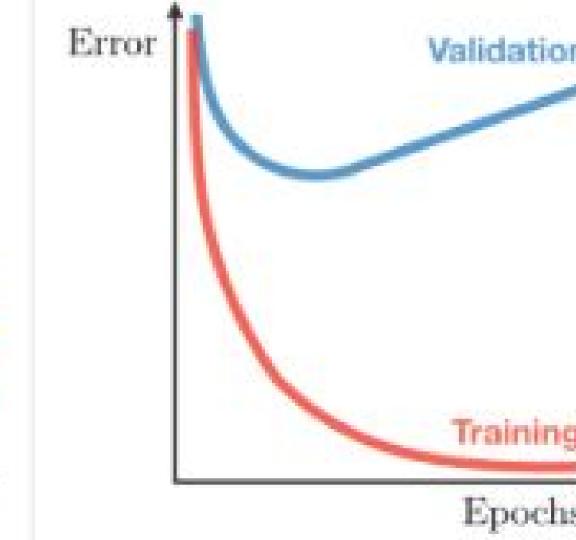
$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

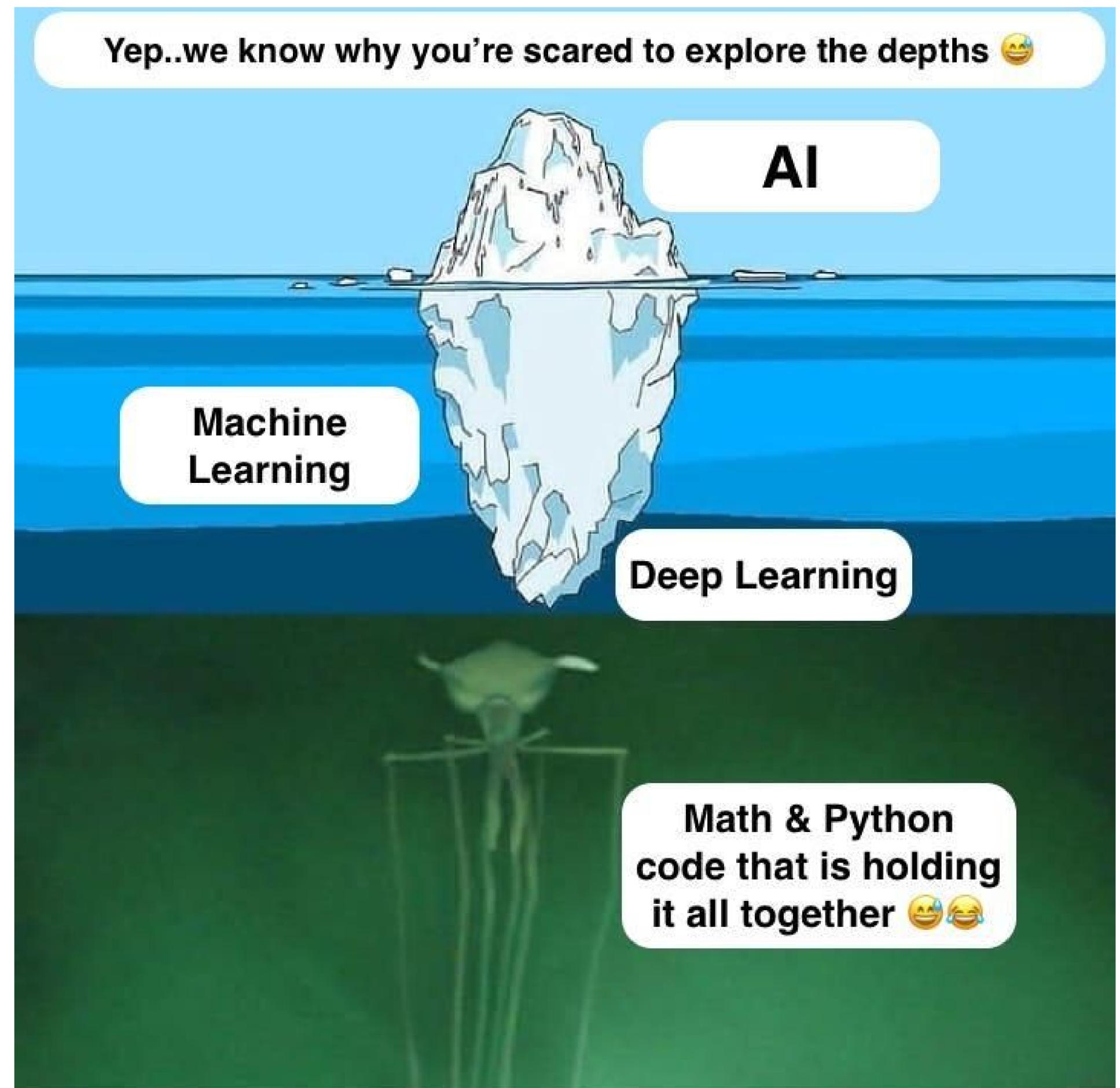
Gradient Descent

Gradient Descent adalah **sebuah algoritma optimisasi** yang digunakan untuk mencari nilai minimum dari sebuah fungsi matematis.

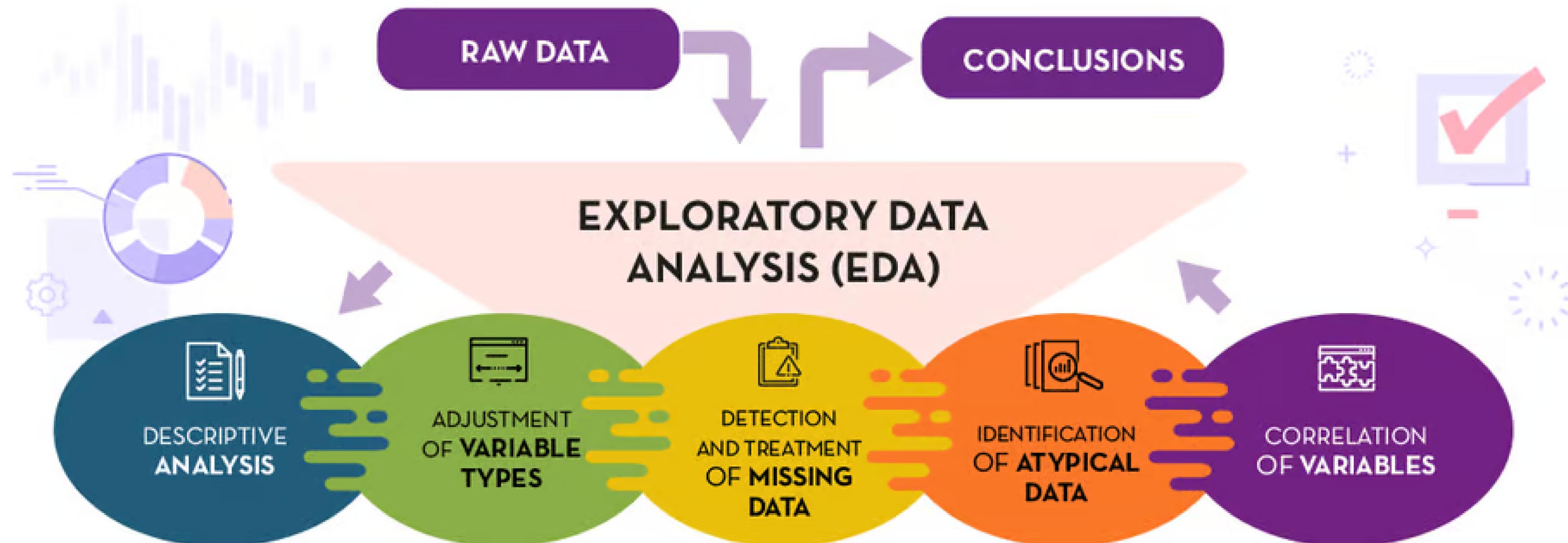
Secara khusus, dalam konteks machine learning, Gradient Descent digunakan **untuk meminimalkan fungsi biaya (cost function)** yang mengukur seberapa baik atau buruk performa model dalam memprediksi nilai target dari data.



	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			



Exploratory Data Analysis



Desc Analysis

```
data.shape  
  
(10000, 5)  
  
data.columns  
  
Index(['Brand', 'Engine Size', 'Horsepower', 'Mileage', 'Price'], dtype='object')  
  
data.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10000 entries, 0 to 9999  
Data columns (total 5 columns):  
 #   Column      Non-Null Count Dtype     
---  --          --          --         
 0   Brand        10000 non-null  object    
 1   Engine Size  10000 non-null  int64    
 2   Horsepower   10000 non-null  int64    
 3   Mileage      10000 non-null  int64    
 4   Price         10000 non-null  float64  
dtypes: float64(1), int64(3), object(1)  
memory usage: 390.8+ KB
```

Exploratory Data Analysis

Data kosong (missing data)

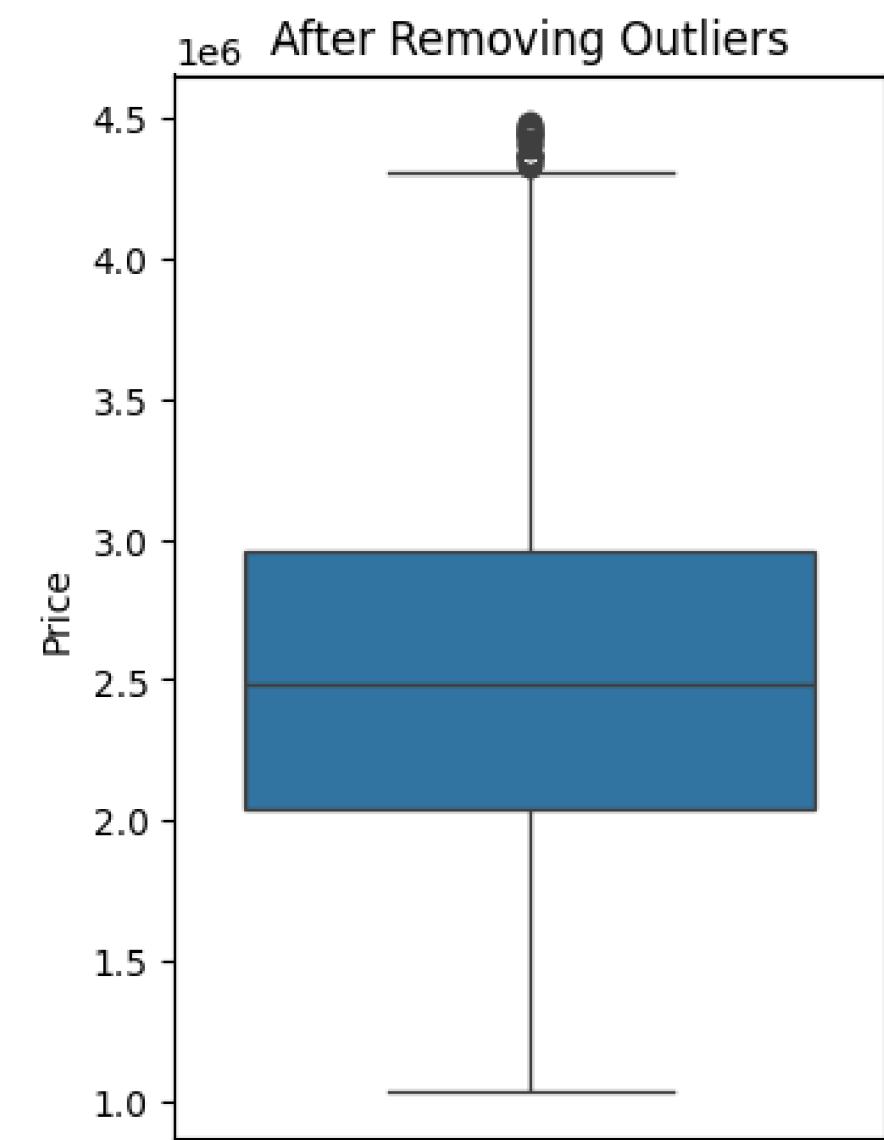
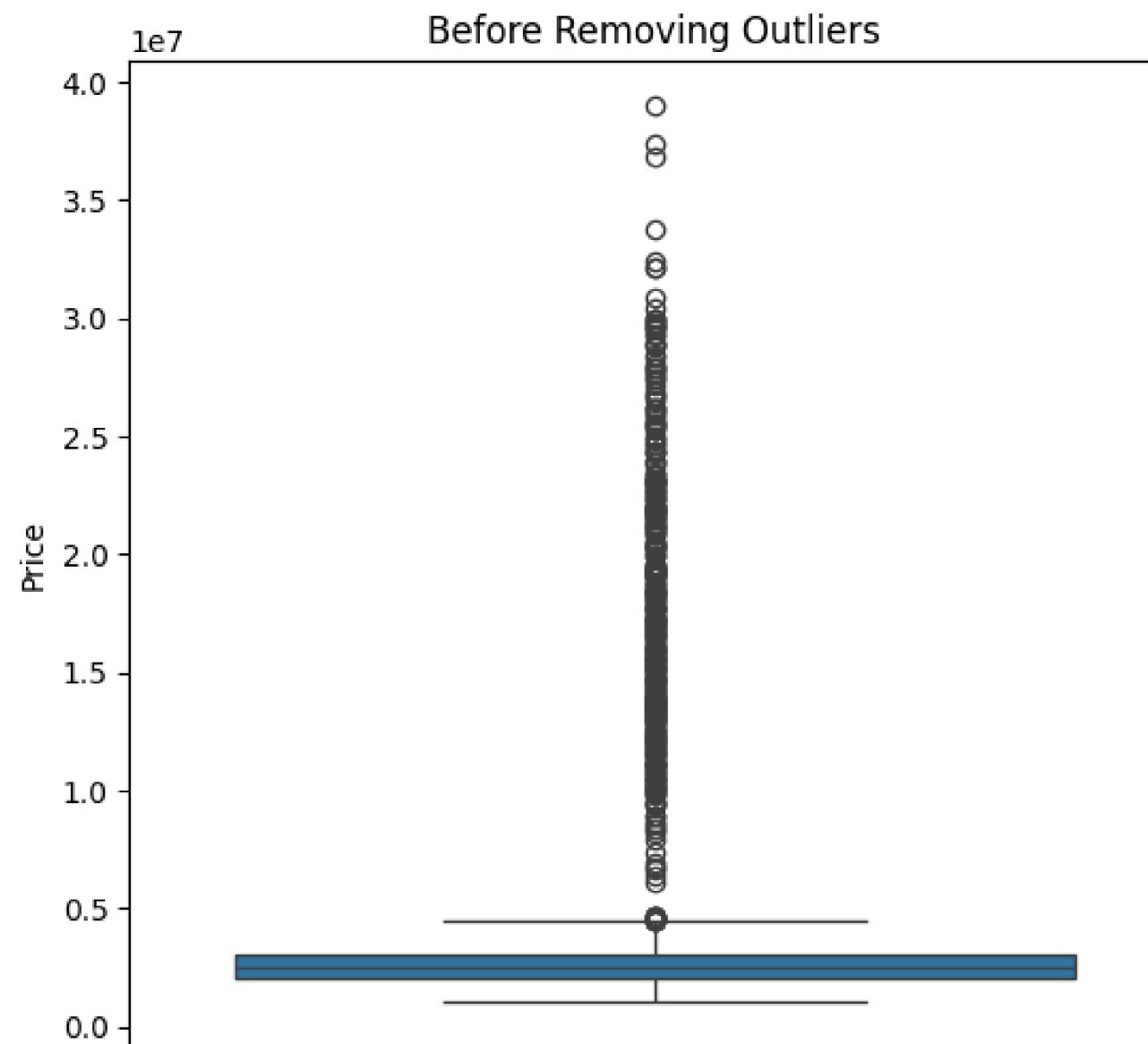
nilai yang tidak ada atau tidak tercatat dalam dataset. Hal ini dapat terjadi karena berbagai alasan, seperti kesalahan dalam pengumpulan data, kegagalan sensor, atau kesalahan dalam proses pengolahan data. Data kosong dapat muncul dalam bentuk nilai yang hilang atau tidak terdefinisi (misalnya, NaN untuk "Not a Number" dalam pengolahan data numerik) atau nilai yang tidak ada dalam dataset (misalnya, nilai yang tidak diisi dalam formulir atau kuesioner).

Honda	5	347	61761	121569.9
BMW		125	11352	
BMW	4		62983	00990.04
Honda	5	326	64353	118159.7
Toyota	5	388	48804	124010.6
BMW	5		100382	132971.76
BMW	1	367		11205.912.000.
Honda	5	315	133157	109037.3
Honda	4		131537	83525.3
Toyota		285	117947	78063.3
BMW	1	270	2610	90674.4
Toyota	2	235	149823	54174.7
BMW	2		86015	
Honda	4	362	174349	104870.1
BMW	1	274	127518	7821622.000.0

Exploratory Data Analysis

Outlier

nilai yang secara signifikan berbeda dari sebagian besar nilai lain dalam dataset. Secara statistik, outlier dapat dianggap sebagai data yang jauh dari nilai-nilai lainnya atau data yang tidak mengikuti pola umum dari sisa data.



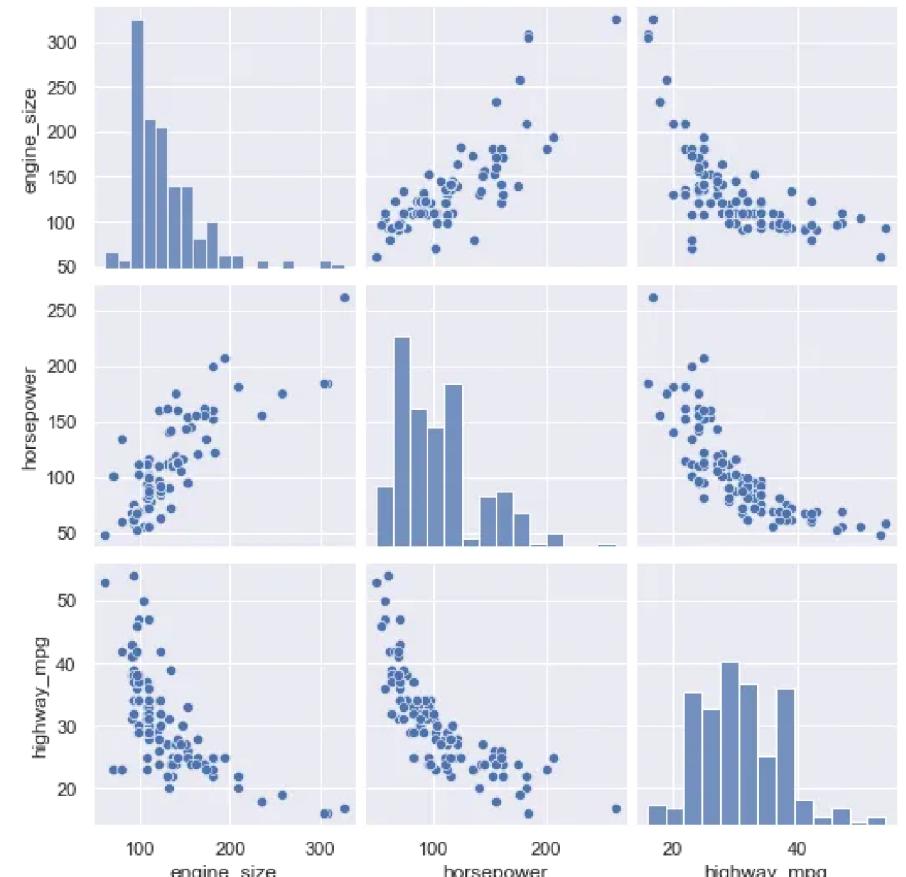
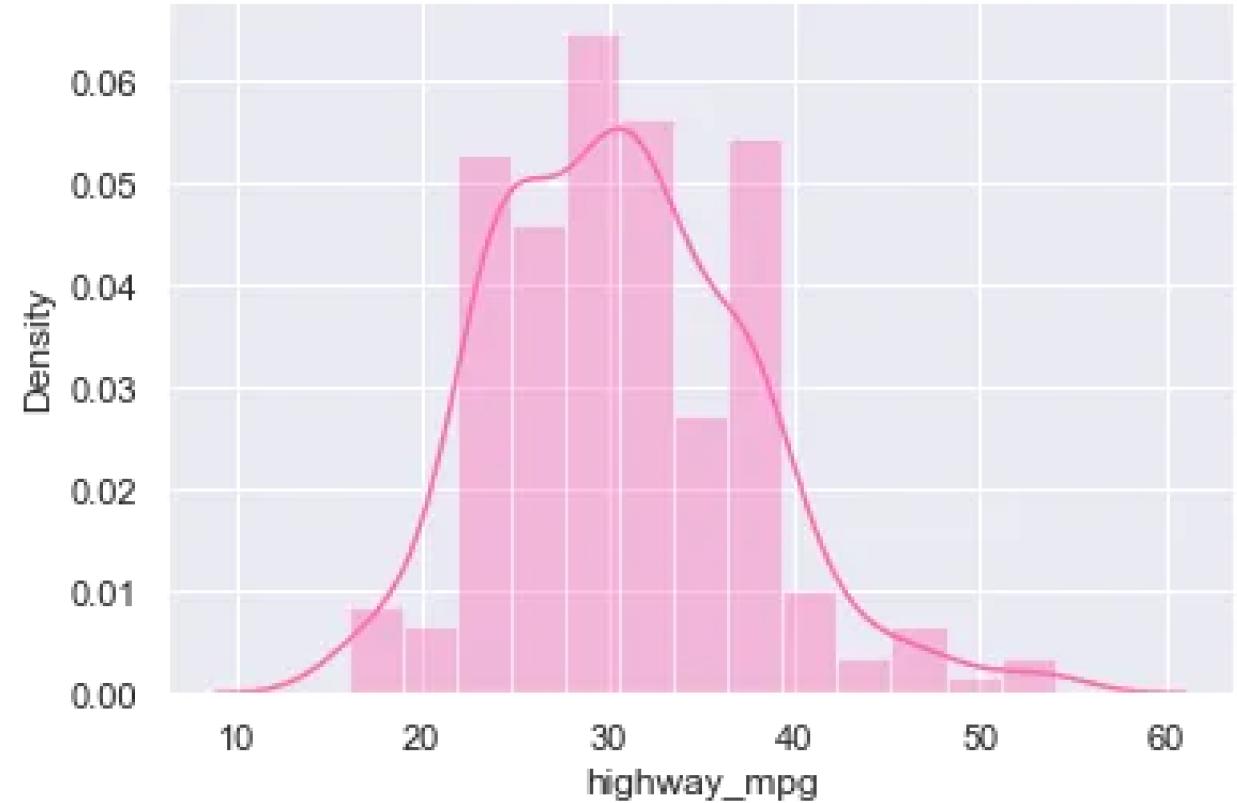
Exploratory Data Analysis

Univariat

berfokus pada satu variabel pada satu waktu, seperti menghitung statistik deskriptif atau menggambarkan distribusi data untuk variabel tunggal. Tujuan utamanya adalah untuk memahami karakteristik atau perilaku variabel tersebut secara terpisah dari variabel lainnya.

Bivariat

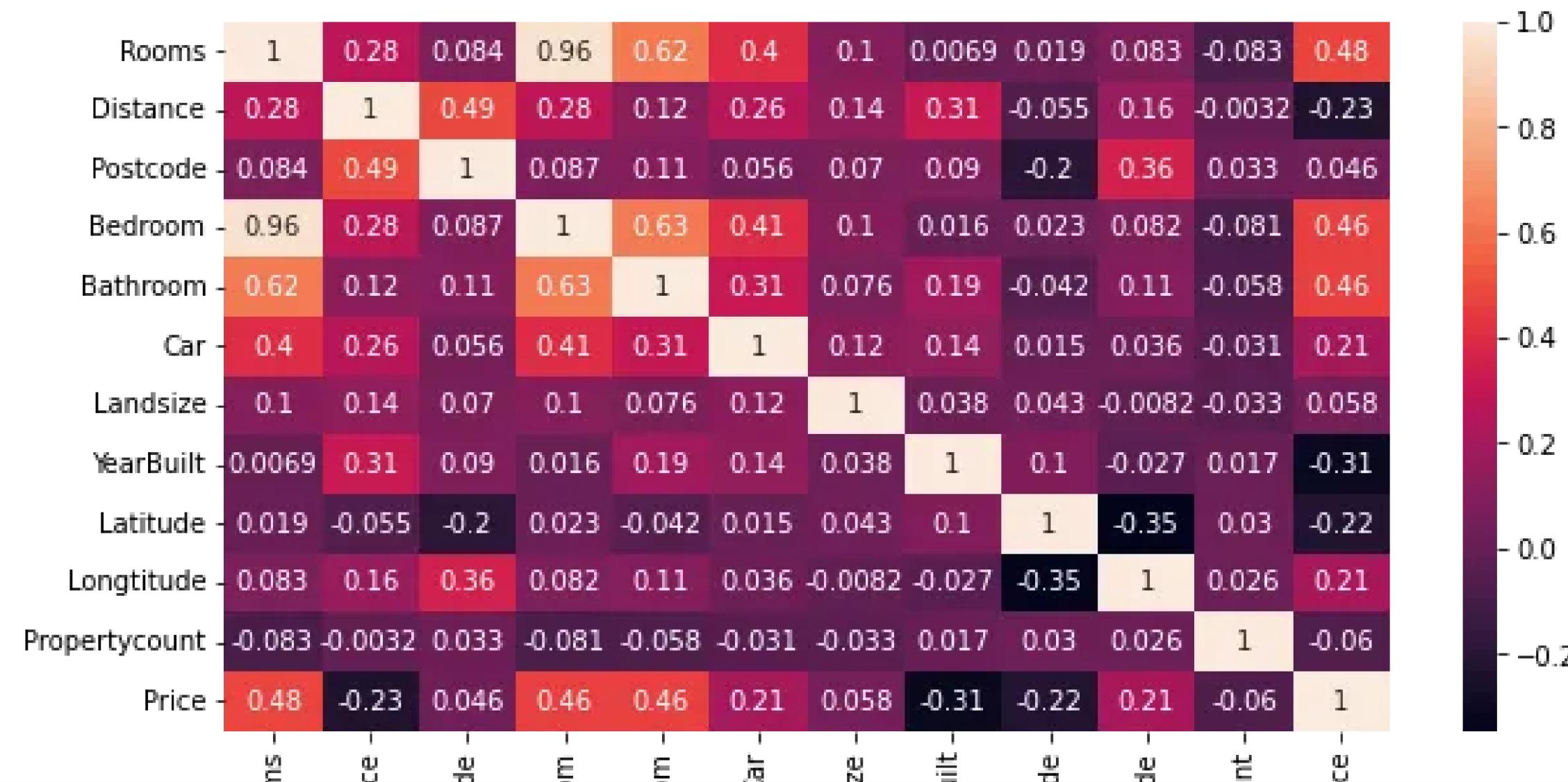
penelitian hubungan antara dua variabel secara bersamaan. Metode ini digunakan untuk mengeksplorasi korelasi, perbedaan, atau interaksi antara dua variabel, seperti mengamati pola dalam scatter plot atau menghitung koefisien korelasi antara dua dataset.



Exploratory Data Analysis

Multivariat

melibatkan lebih dari dua variabel dan bertujuan untuk memahami interaksi yang kompleks di antara variabel-variabel tersebut. Ini sering melibatkan teknik seperti analisis regresi multiple, analisis faktor, atau clustering untuk mengidentifikasi pola atau struktur yang kompleks dalam data yang melibatkan banyak variabel secara bersamaan.





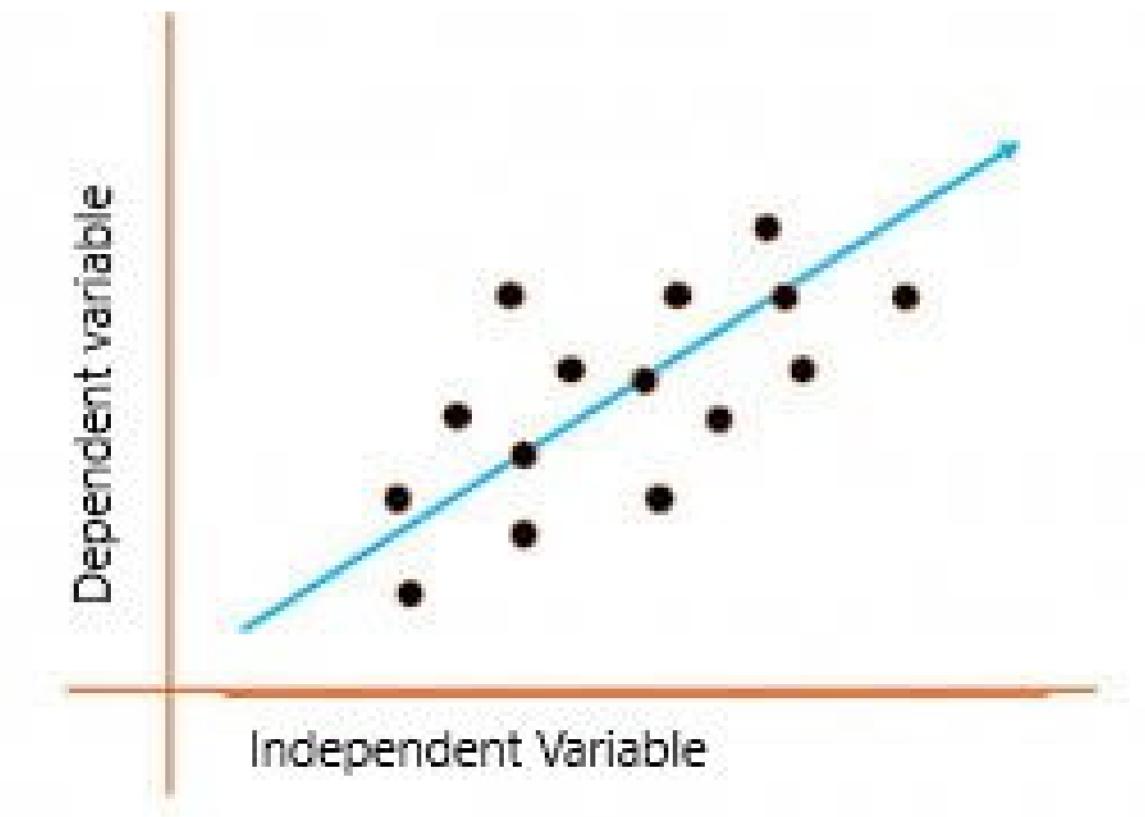
Q4

**Apa dan kenapa
harus Regresi Linear ?**

Linear Regression

Regresi Linear adalah salah satu metode statistik yang digunakan untuk **memodelkan hubungan antara variabel dependen (target) dengan satu atau lebih variabel independen (fitur)**.

Regresi linear sering digunakan untuk menganalisis dan **memodelkan data yang kontinu** dan mengasumsikan bahwa **hubungan antara variabel-variabel tersebut dapat dijelaskan dengan model linier**.



Variable Dependent (target) & Independent (features)

A	B	C	D	E	F
	Brand	Engine Size	Horsepower	Mileage	Price
0	Toyota	5	344	123394	110877.6
1	BMW	1	262	81419	79683.72
2	Honda	4	256	89838	90187.2
3	BMW	4	208	145651	9.628.787.999.9
4	BMW	4	113	172376	70141.68
5	BMW	2	253	31960	92952.0
6	Honda	4	129	149435	58003.5
7	Toyota	3	107	73911	41.075.910.000.
8	BMW	5	102	191401	7.909.427.999.9
9	BMW	1	318	154353	8.446.043.999.9
10	Toyota	1	294	91213	62936.7
11	Toyota	5	128	64895	71422.5
12	Honda	3	302	97455	86953.5
13	Honda	2	296	10502	86277.8
14	Toyota	1	284	183372	51003.8

Note : Lakukan Feature Engineering untuk data Kualitatif

Raw Data

```
0 : {  
    house_info : {  
        num_rooms: 6  
        num_bedrooms: 3  
        street_name: "Shorebird Way"  
        num_basement_rooms: -1  
    }  
    ...  
}
```

Feature Vector

```
[  
    6.0,  
    1.0,  
    0.0,  
    0.0,  
    0.0,  
    9.321,  
    -2.20,  
    1.01,  
    0.0,  
    ...,  
]
```

Feature Engineering

Raw data doesn't come to us as feature vectors.

Process of creating features from raw data is **feature engineering**.

Persamaan Regresi Linear

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Annotations for the components:

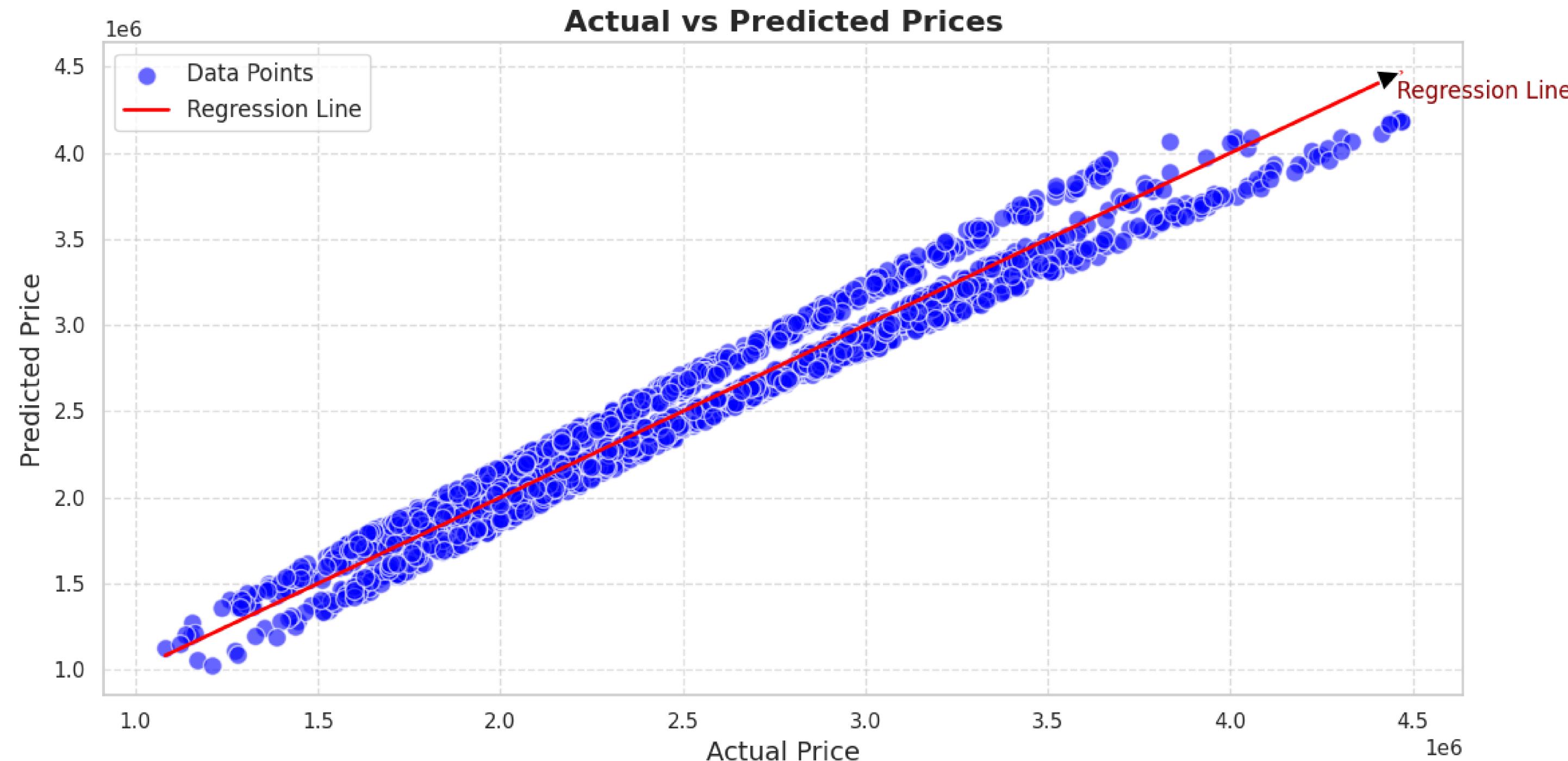
- Dependent Variable: Points to Y_i
- Population Y intercept: Points to β_0
- Population Slope Coefficient: Points to β_1
- Independent Variable: Points to X_i
- Random Error term: Points to ε_i

Brackets indicate the components:

- A blue bracket under $\beta_0 + \beta_1 X_i$ is labeled "Linear component".
- A blue bracket under ε_i is labeled "Random Error component".

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_n \cdot X_n$$

Persamaan Regresi Linear



Confusion Matrix

		Confusion Matrix	
		Predicted Positive	Predicted Negative
Actual	Actual Positive	809	6
	Actual Negative	19	1085
Predicted			

Confusion Matrix

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

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

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

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: 0.9870
Precision: 0.9926
Recall: 0.9771
F1-score: 0.9848

Untuk apa Classification Report ?

Keempat metrik evaluasi ini penting dalam evaluasi kinerja model klasifikasi:

1. **Accuracy** untuk mengukur seberapa akurat model dalam memprediksi seluruh kelas. Ini adalah rasio prediksi benar (baik positif maupun negatif) dibandingkan dengan total jumlah data.
2. **Precision** untuk mengukur seberapa akurat model dalam memprediksi kelas positif. Precision memberikan informasi tentang berapa persen dari prediksi positif yang sebenarnya benar.
3. **Recall** untuk mengukur seberapa baik model dalam menemukan semua instance yang relevan dari kelas positif. Recall memberi tahu kita berapa persen dari total kelas positif yang berhasil dideteksi oleh model.
4. **F1-score** memberikan keseimbangan antara precision dan recall. Nilai F1-score yang tinggi menunjukkan bahwa precision dan recall juga tinggi, sehingga model dapat dianggap baik dalam memprediksi kelas positif.



Q5

**Bagaimana cara
implementasinya ?**

**See You
Tomorrow !**