



Bài 2

DỮ LIỆU CHO HỌC MÁY

Giảng viên: TS. Nguyễn Ngọc Giang

SĐT: 0862411011

Email: giangnn.cntt@dainam.edu.vn

- Biểu diễn một bài toán học máy [Mitchell, 1997]
 - **Một công việc (nhiệm vụ) T**
 - **Đối với các tiêu chí đánh giá hiệu năng P**
 - **Thông qua (sử dụng) kinh nghiệm E**

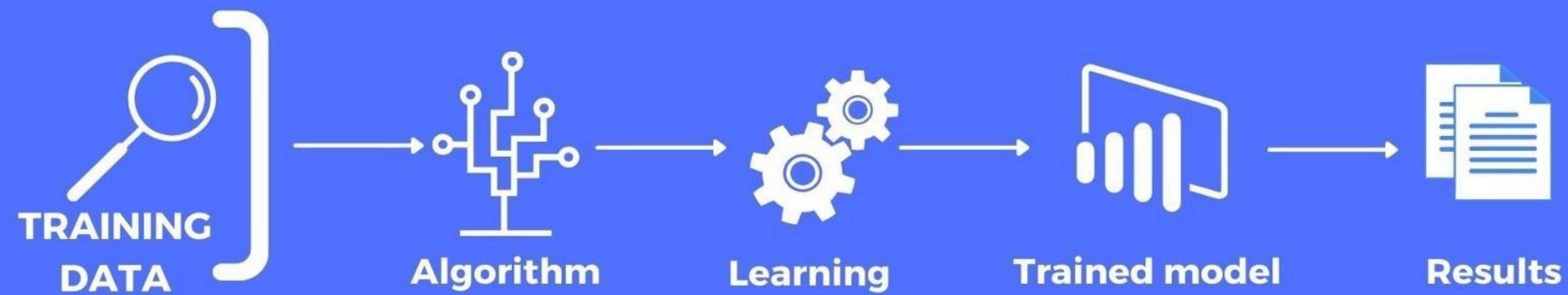
- Lọc thư rác – Email spam filtering
 - *T: Dự đoán (để lọc) những thư điện tử nào là thư rác (spam email)*
 - *P: % of các thư điện tử gửi đến được phân loại chính xác*
 - *E: Một tập các thư điện tử (emails) mẫu, mỗi thư điện tử được biểu diễn bằng một tập thuộc tính (vd: tập từ khóa) và nhãn lớp (thư thường/ thư rác) tương ứng.*

- Phân loại các trang Web
 - T: Phân loại các trang Web theo các chủ đề đã định trước
 - P: Tỷ lệ (%) các trang Web được phân loại chính xác
 - E: Một tập các trang Web, trong đó mỗi trang Web gắn với một chủ đề

- Phân loại hoa Iris
 - **T:** *Dự đoán một bông hoa thuộc loại nào trong 3 loại của hoa Iris*
 - **P:** *% các bông hoa được dự đoán đúng loại*
 - **E:** *Một tập dữ liệu các bông hoa, trong đó mỗi bông hoa bao gồm các thông số pedal_width, pedal_height, sepal_width, sepal_height và nhãn của bông hoa thuộc loại gì trong 3 loại Setosa, Versicolor, Virginica*



Machine Learning Process

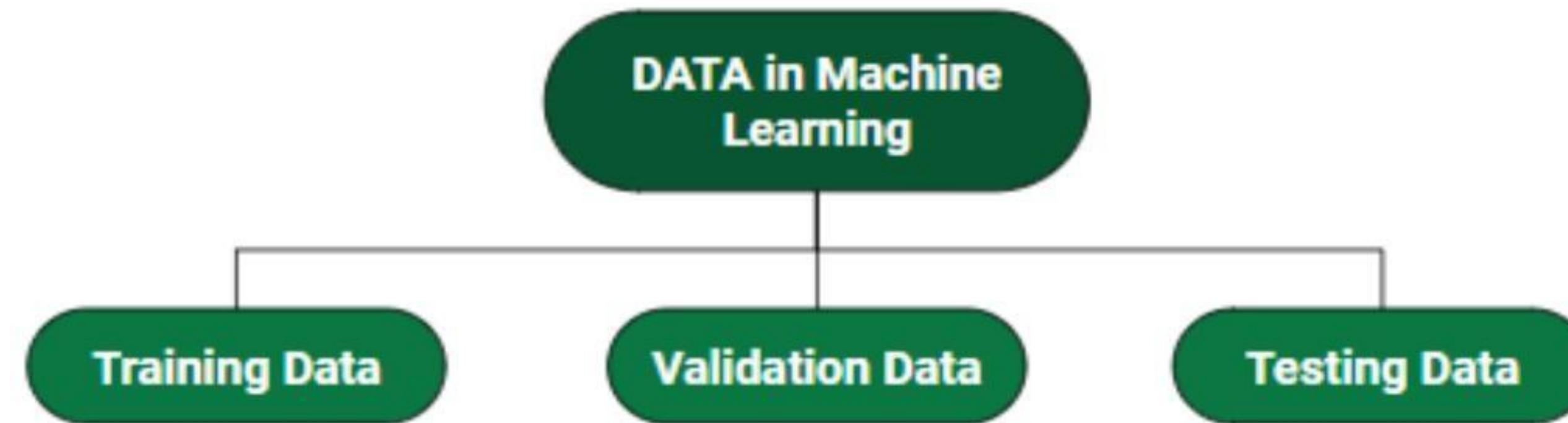


- **Giới thiệu về dữ liệu cho học máy**
- **Xây dựng dữ liệu cho học máy**
- **Kho dữ liệu mẫu cho học máy**
- **Ôn tập về xác suất thống kê**
- **Sử dụng thư viện trong Python**
- **Luyện tập tiếng Anh**

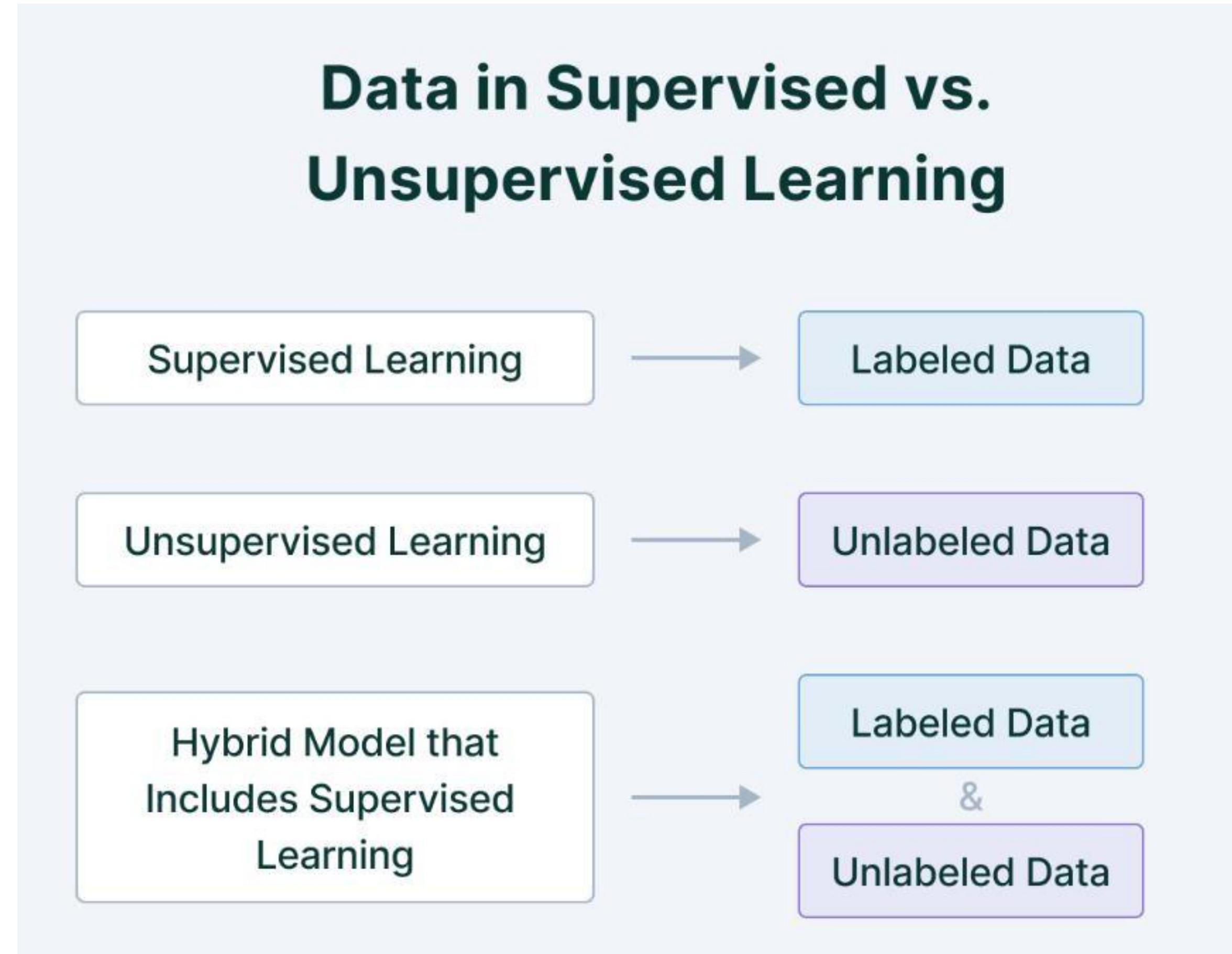


Giới thiệu

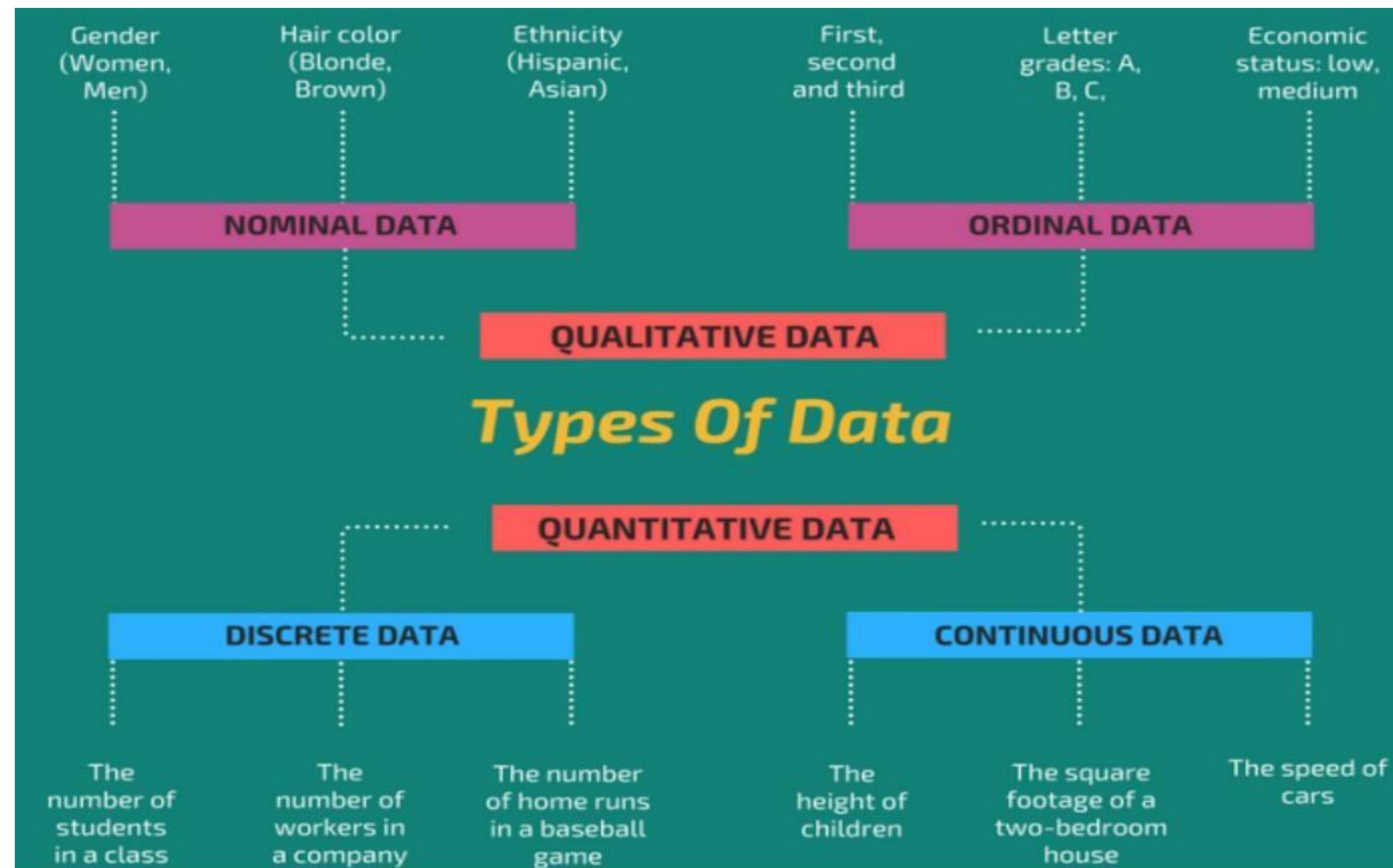
- Tại sao phải cần dữ liệu cho học máy?



- Các loại dữ liệu cho học máy

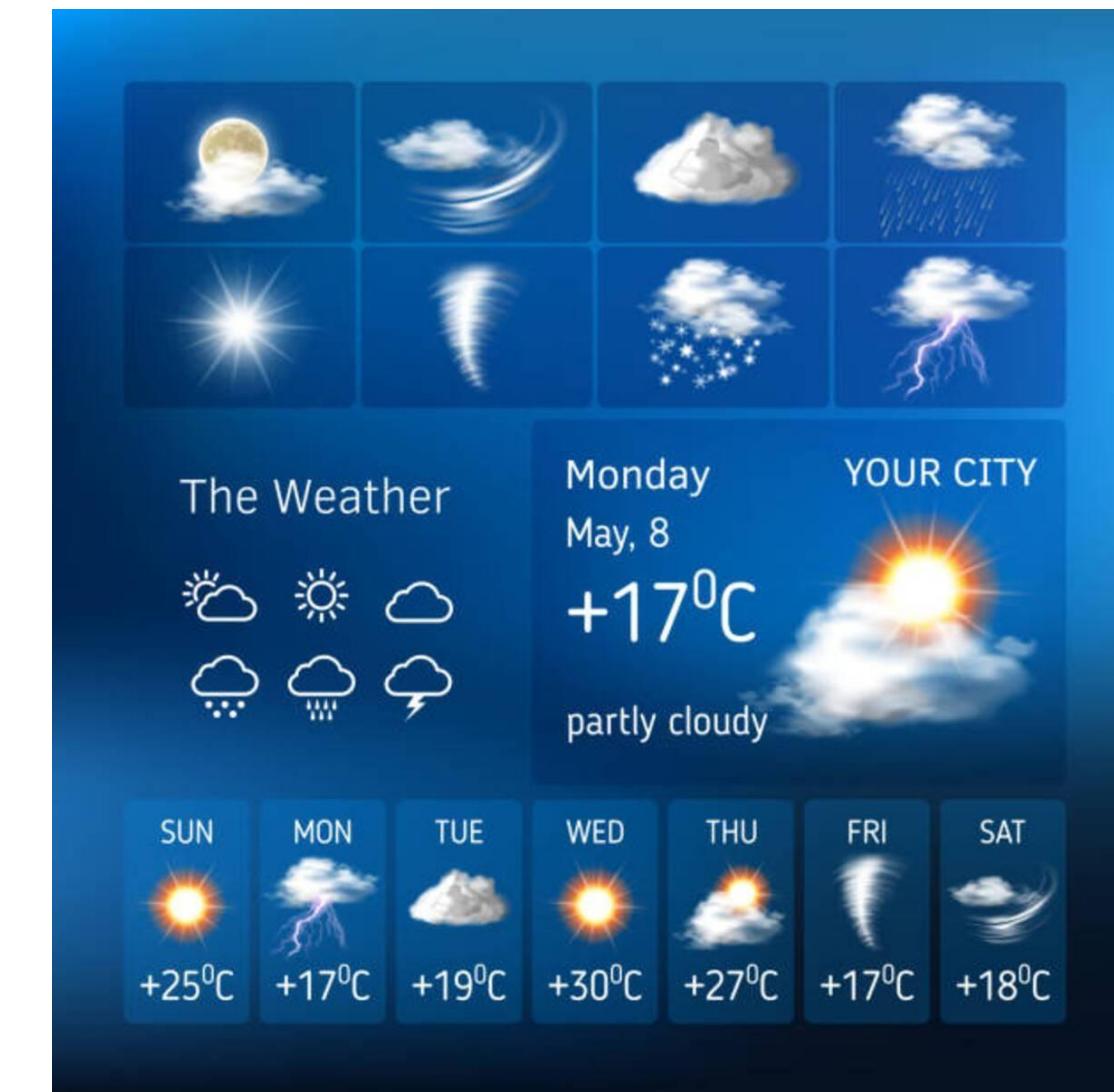


• Các loại dữ liệu cho học máy





outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa



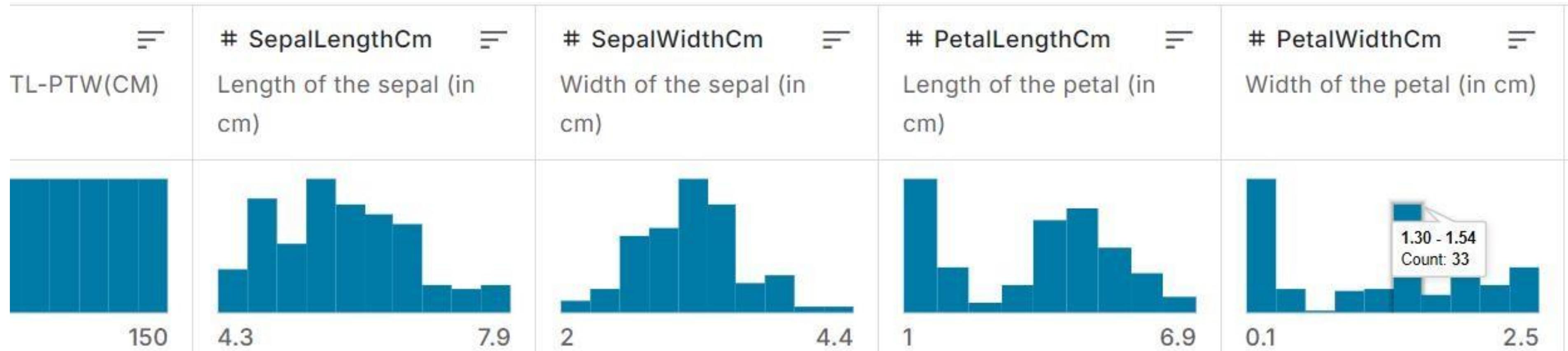
Iris Setosa



Iris Versicolor



Iris Virginica



Iris Setosa

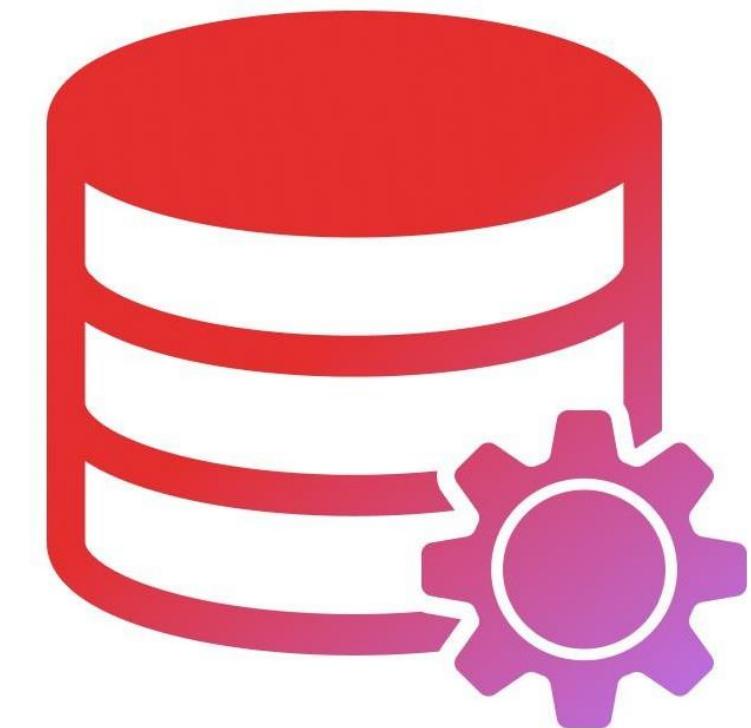


Iris Versicolor

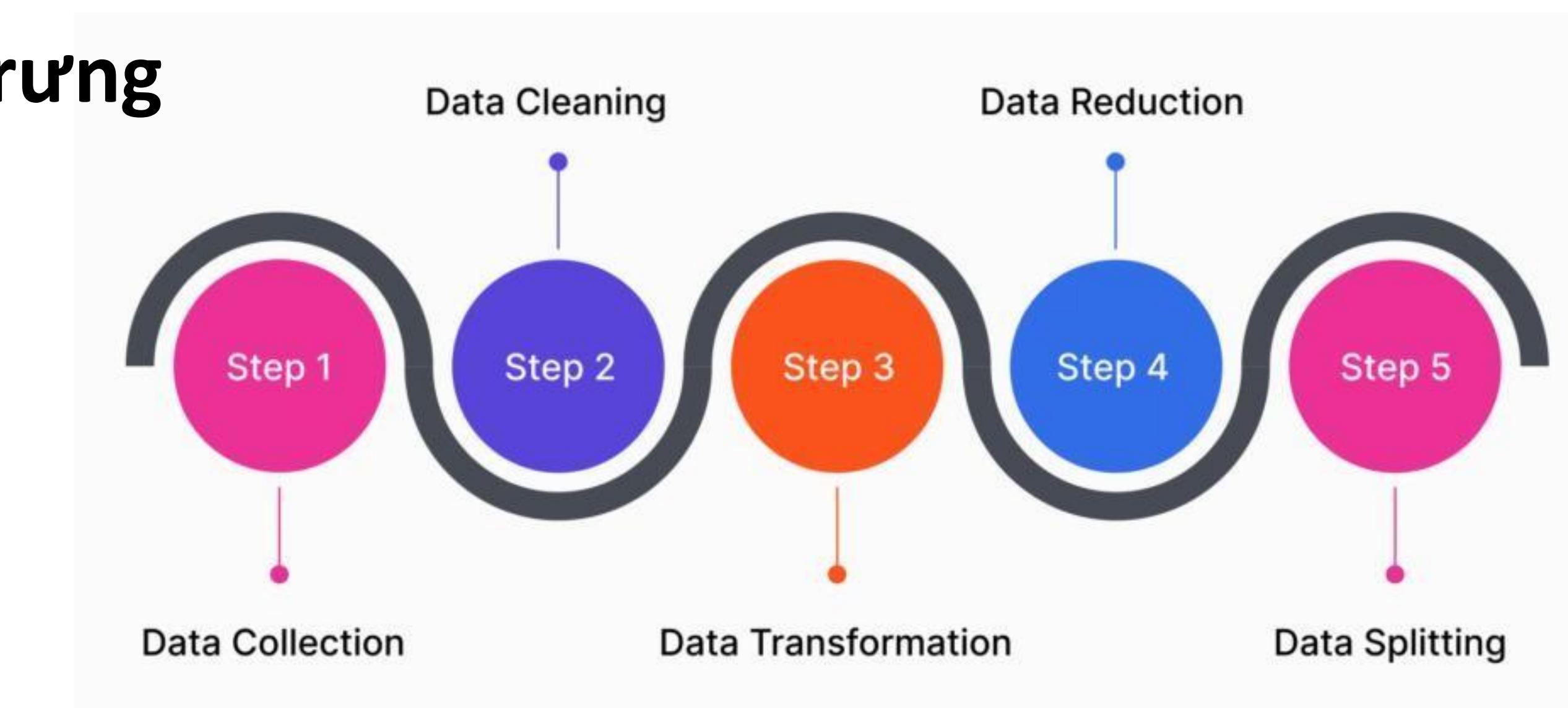


Iris Virginica

- **Một số yêu cầu đối với dữ liệu cho học máy**
 - ✓ Thể hiện được đặc trưng của đối tượng máy cần học
 - ✓ Có cấu trúc rõ ràng
 - ✓ Hạn chế thấp nhất dữ liệu giao nhau
 - ✓ Hạn chế thấp nhất dữ liệu rác
 - ✓ Tỷ lệ cân bằng giữa các mẫu



- Thu thập dữ liệu
- Làm sạch dữ liệu
- Chuẩn hóa và chuyển đổi dữ liệu
- Lựa chọn đặc trưng



<https://www.pecan.ai/blog/data-preparation-for-machine-learning/>

XÂY DỰNG DỮ LIỆU

- Cấu trúc dữ liệu cho học máy

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

- Phân tích kỹ thuật trích chọn đặc trưng



- Phân tích kỹ thuật trích chọn đặc trưng



- Phân tích kỹ thuật trích chọn đặc trưng



- Luyện tập thu thập dữ liệu cho học máy



Kho dữ liệu mẫu

≡ kaggle

+ Create

Home

Competitions

Datasets

Models

Code

Discussions

Learn

More

Search

Sign In

Register

Datasets

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

+ New Dataset

Search datasets

Filters

All datasets

Computer Science

Education

Classification

Computer Vision

NLP

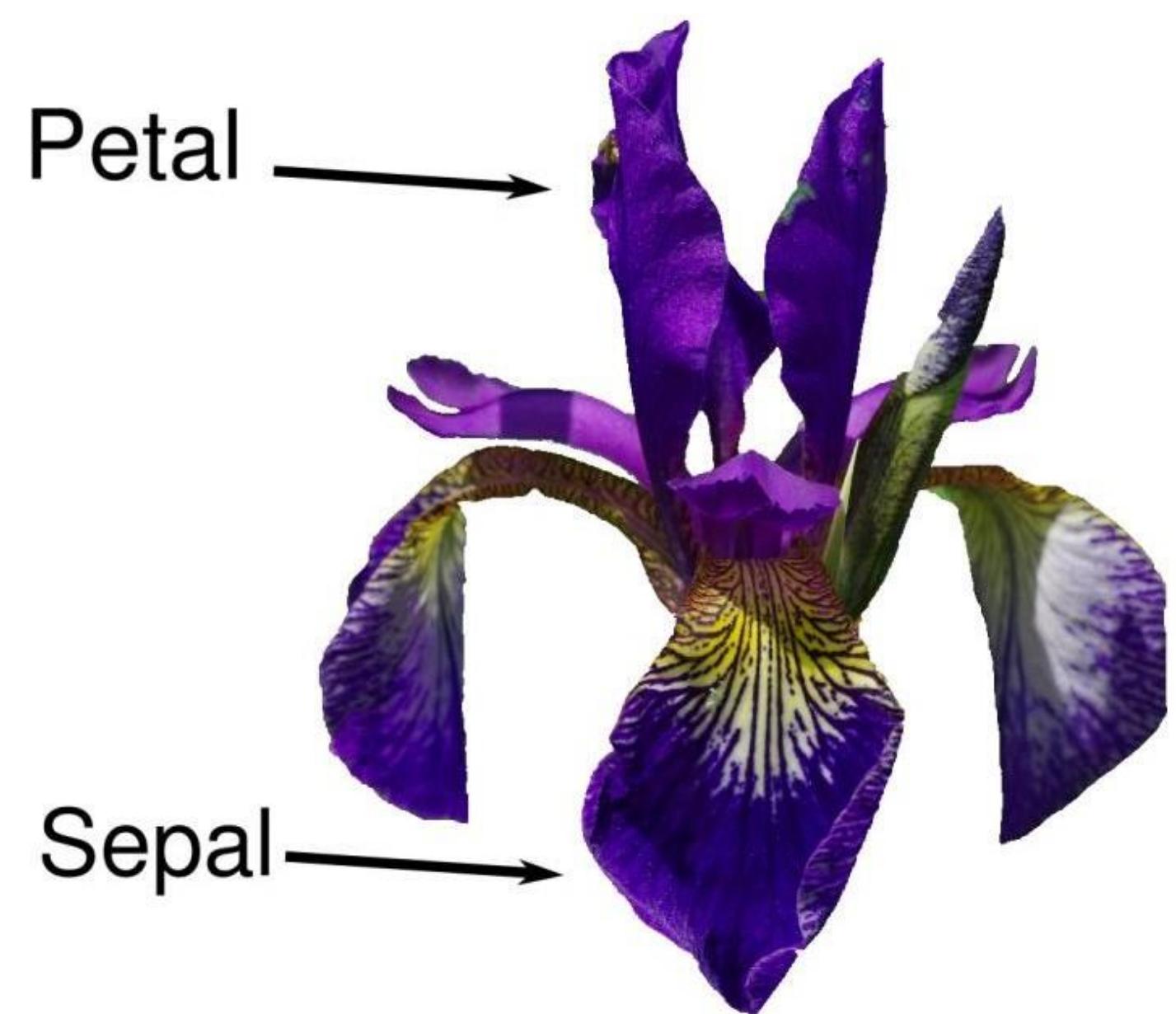
Data Visualization

Pre-Trained Model





Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa
23	4.6	3.6	1	0.2	Iris-setosa
24	5.1	3.3	1.7	0.5	Iris-setosa
25	4.8	3.4	1.9	0.2	Iris-setosa



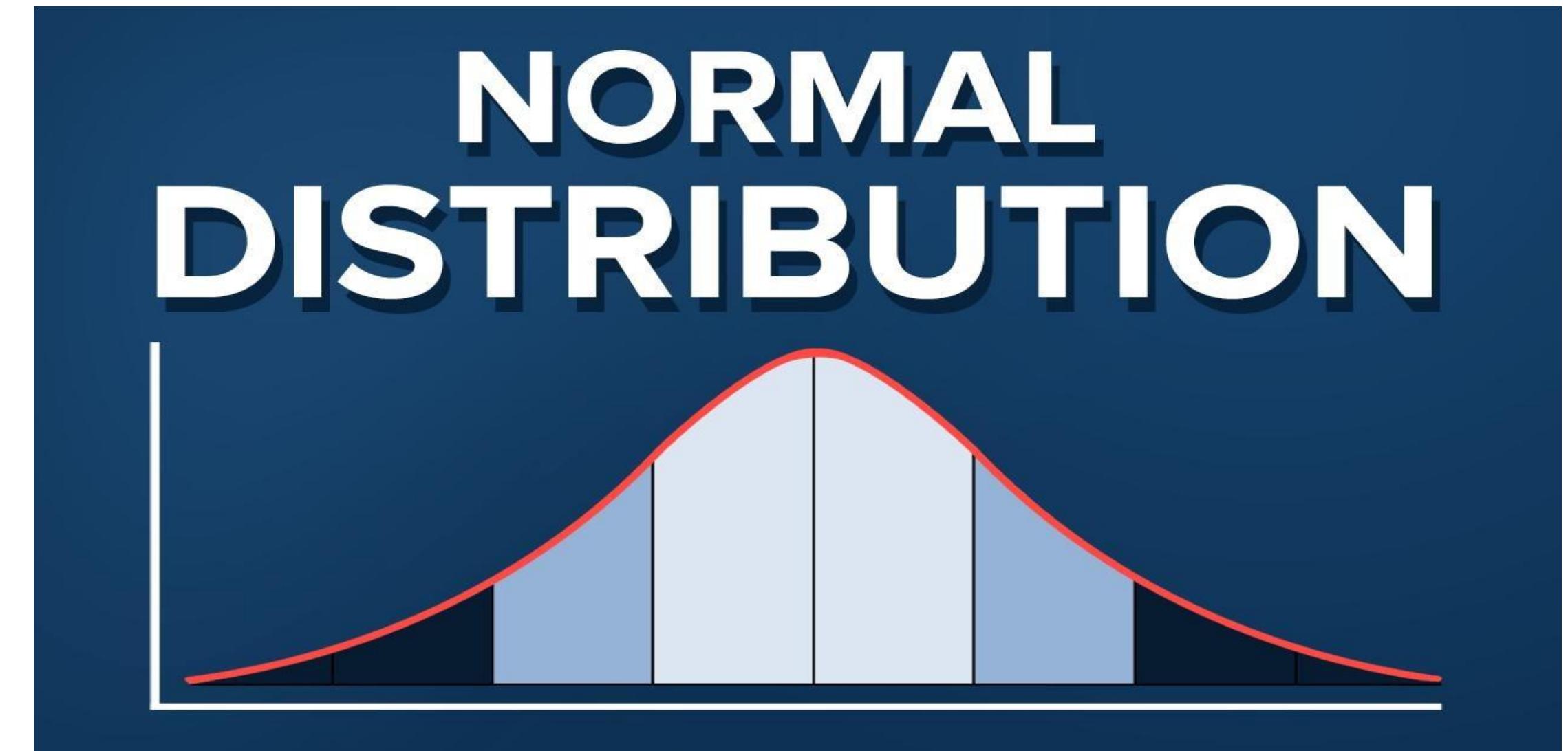


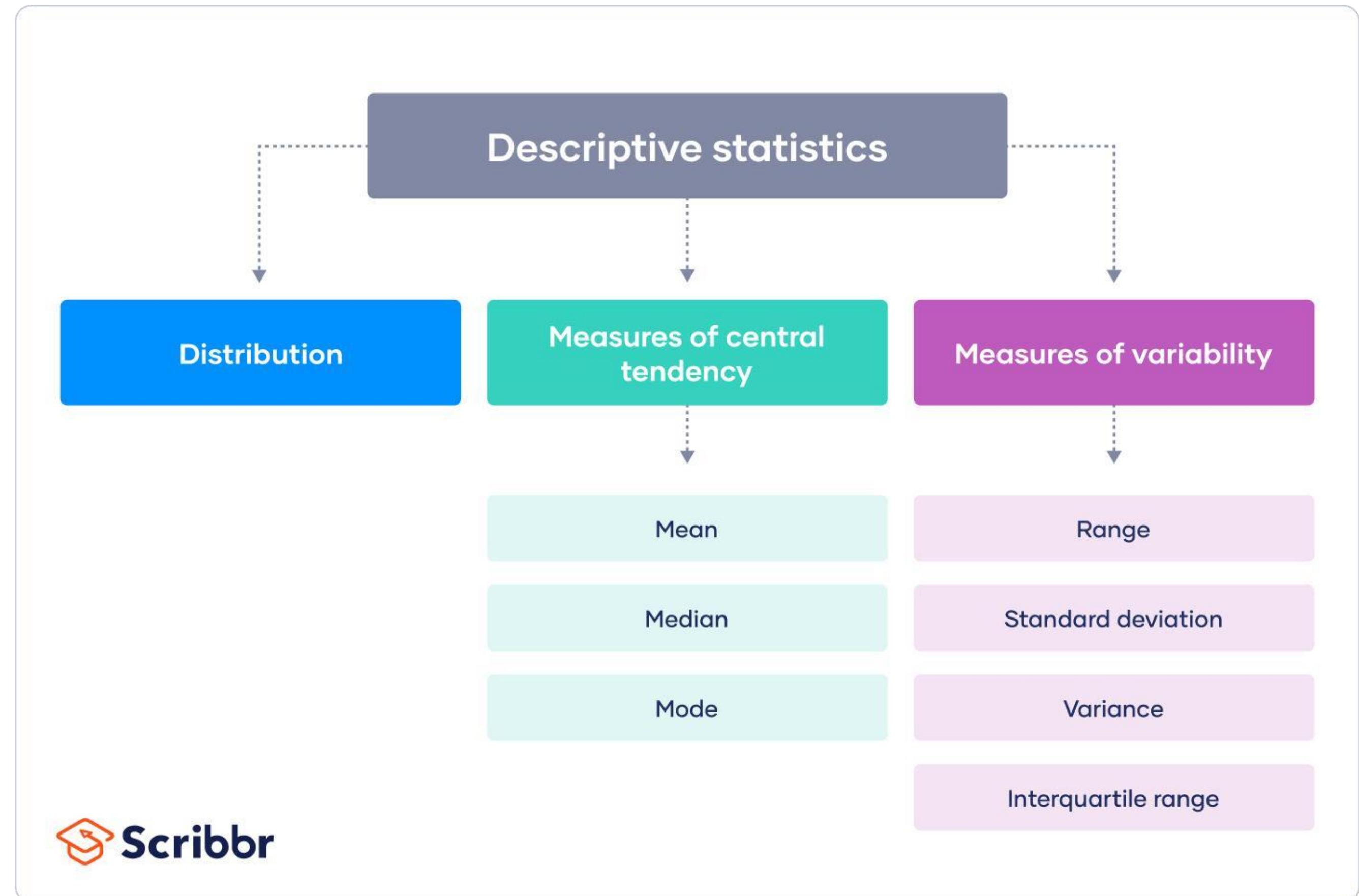
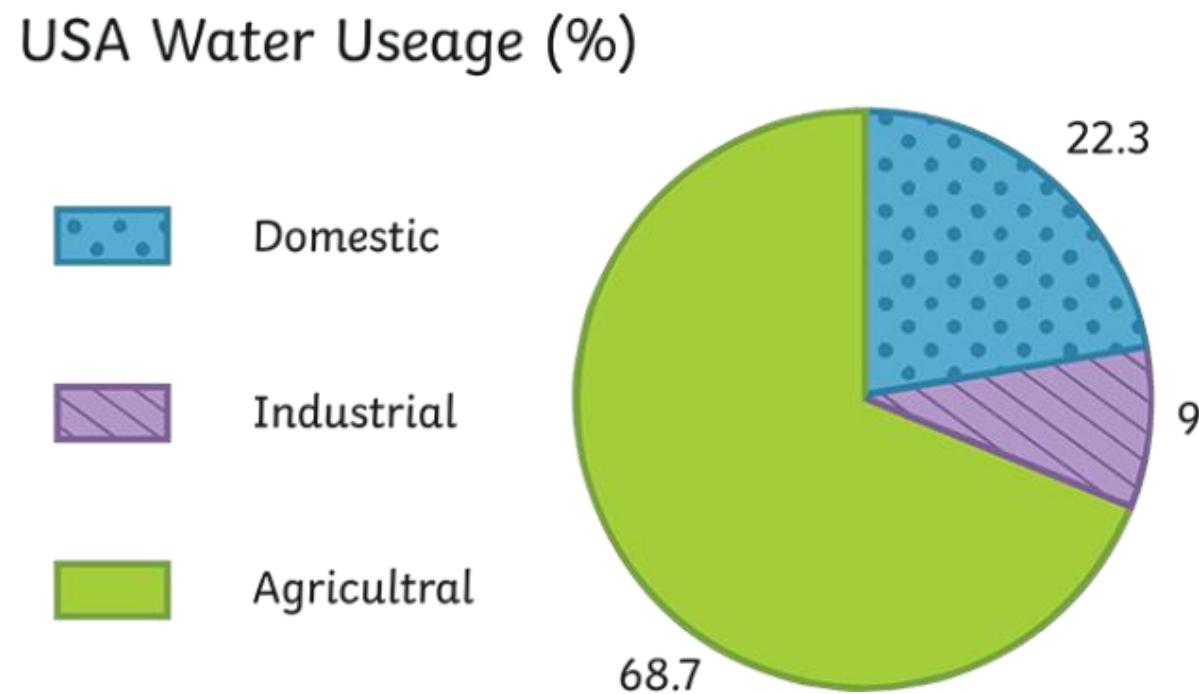
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5



Xác xuất và thống kê

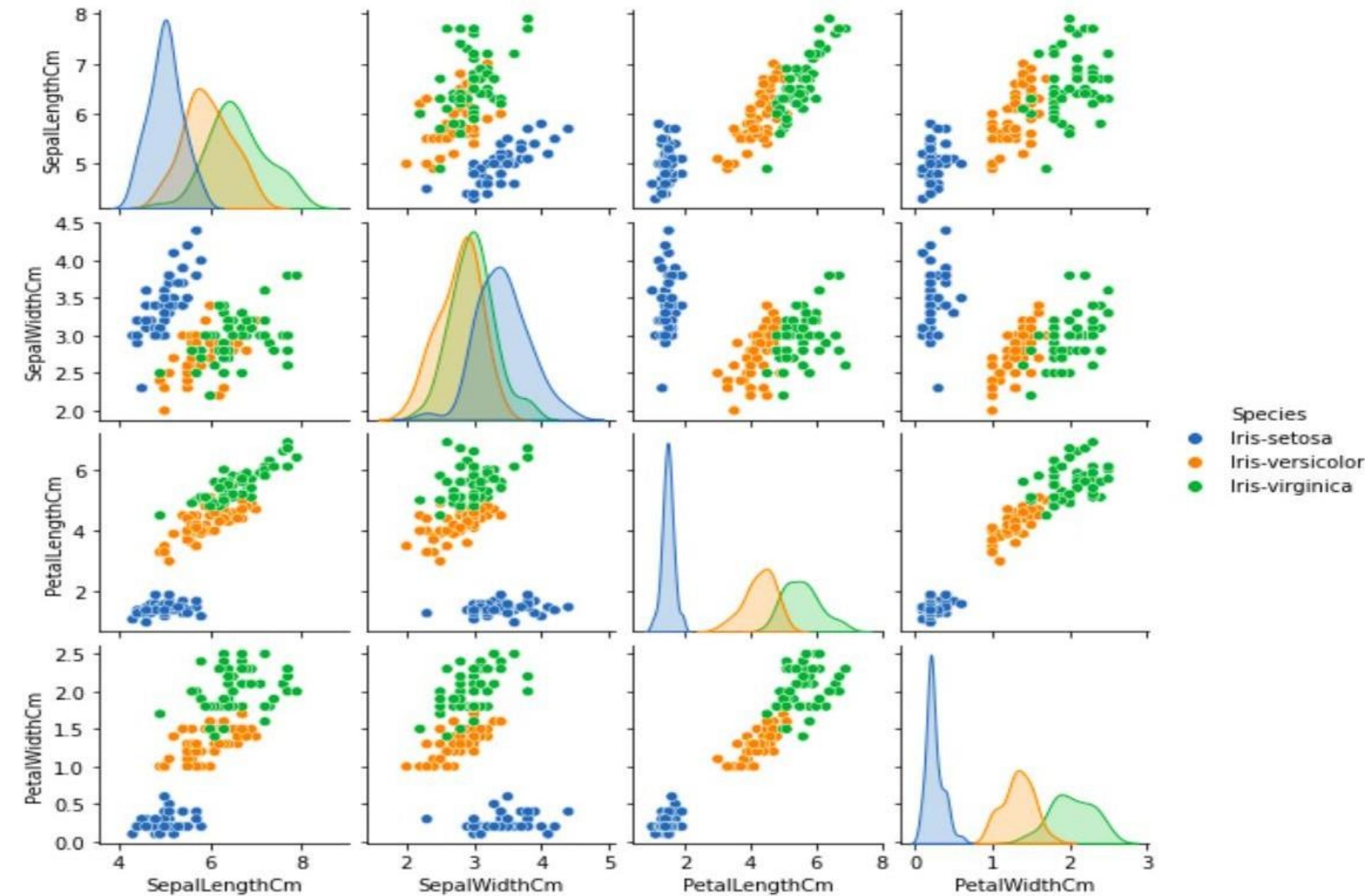
- Số ngẫu nhiên
- Hàm sinh số ngẫu nhiên trong ngôn ngữ lập trình
- Xác suất xuất hiện
- Hàm phân bố







Thống kê đối với Iris Dataset



Sử dụng thư viện trong Python

- Scikit-learn

- NumPy

- SciPy

- Matplotlib

- Pandas

```
In [2]: 1 # Load our necessary libraries
2 import pandas as pd
3 import numpy as np
```

```
In [3]: 1 # Create data into CSV (that we'll import later on)
2 # Let's say the data is for a veterinarian keeping tabs on his clients.
3
4 raw_data = {'pet_name': ['Woof', 'Chester', 'Rex', 'Mystery', 'Pumpkin'],
5             'pet_last_name': ['Smith', 'Kim', "", 'Taylor', ""],
6             'good_pet_score': [96, 34, 89, 92, 79],
7             'type': ['dog', 'cat', 'mini-dinosaur', "unknown", "bird"],
8             'amount_owed': ["5000", "9,000", 570, 622, 190]}
9 df = pd.DataFrame(raw_data, columns = ['pet_name', 'pet_last_name', 'good_pet_score', 'type', 'amount_owed'])
10 df
```

Out[3]:

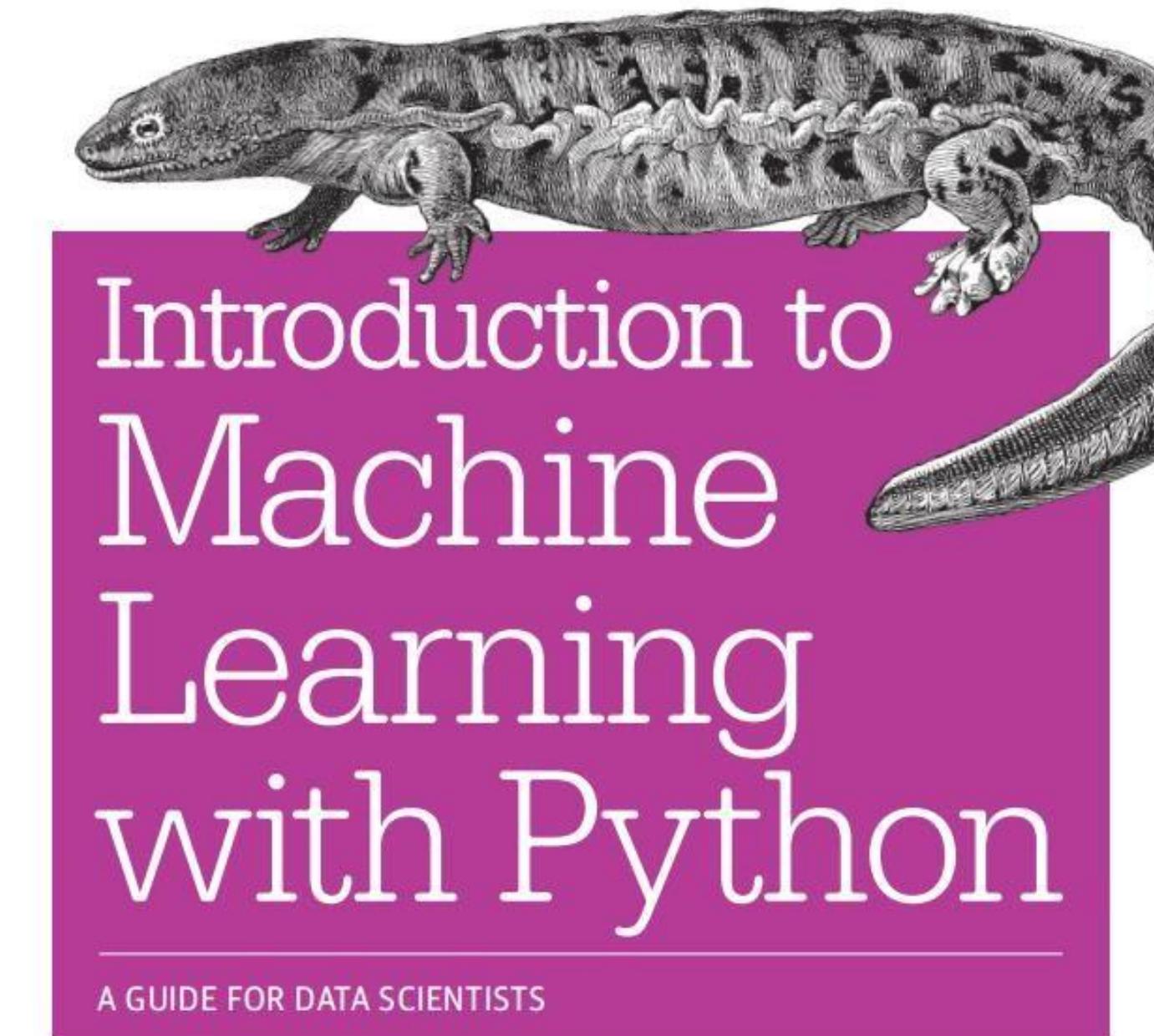
	pet_name	pet_last_name	good_pet_score	type	amount_owed
0	Woof	Smith	96	dog	5000
1	Chester	Kim	34	cat	9,000
2	Rex		89	mini-dinosaur	570
3	Mystery	Taylor	92	unknown	622
4	Pumpkin		79	bird	190



**PRACTICE
ENGLISH
EVERYDAY
30 Minutes**



O'REILLY®



Andreas C. Müller & Sarah Guido

- Ngôn ngữ lập trình Python
- Bài luyện tập tiếng Anh



