



Bài giảng môn học:

Kỹ nghệ tri thức và học máy (4080540)

CHƯƠNG 3: HỌC CÓ GIÁM SÁT - 01 (Supervised Learning)

Giảng viên: Đặng Văn Nam

Email: dangvannam@humg.edu.vn

Nội dung chương 3

1. Các bước xây dựng một mô hình học máy
2. Datasets
3. Học có giám sát
4. Phân loại học có giám sát
5. Thuật toán phân loại (KNN, Decision Tree)
6. Thuật toán hồi quy ()
7. Đánh giá độ chính xác của mô hình phân lớp, hồi quy

1. Các bước cơ bản xây dựng một mô hình học máy.

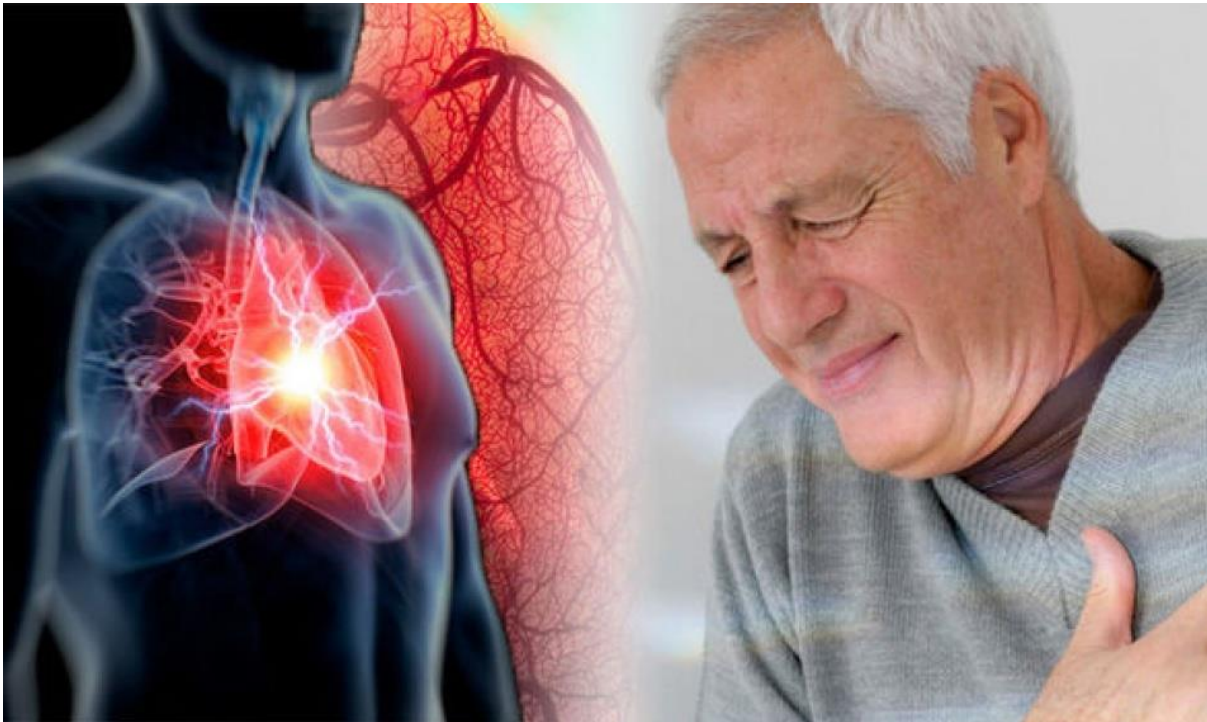
1. Các bước xây dựng một mô hình học máy



- Để xây dựng một mô hình học máy thực hiện qua 7 bước:
 - 1) Thu thập dữ liệu (Data collection)
 - 2) Chuẩn bị dữ liệu (Data preparation)
 - 3) Lựa chọn mô hình phù hợp (Choosing a model)
 - 4) Huấn luyện mô hình (Training)
 - 5) Đánh giá mô hình (Evaluation)
 - 6) Tùy chỉnh tham số của mô hình (Parameter tuning)
 - 7) Dự đoán với mô hình xây dựng được (Prediction)

Ví dụ chi tiết:

- Các bước xây dựng một mô hình học máy cho bài toán dự đoán bệnh nhân có bị bệnh đau tim hay không?



- Tham khảo các bước thực hiện trong file code trên Jupyter Notebook

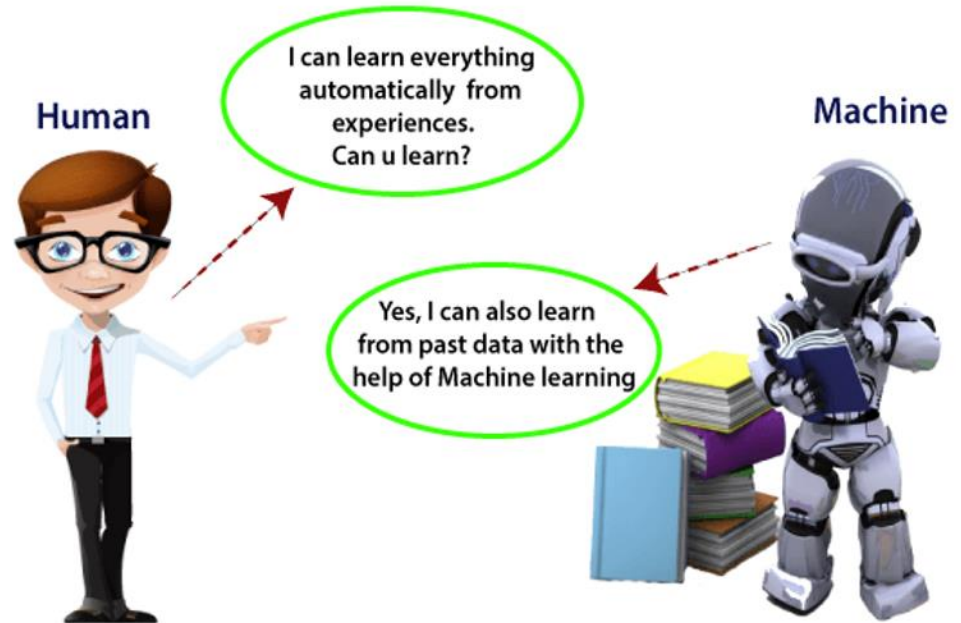
2. Datasets

Datasets

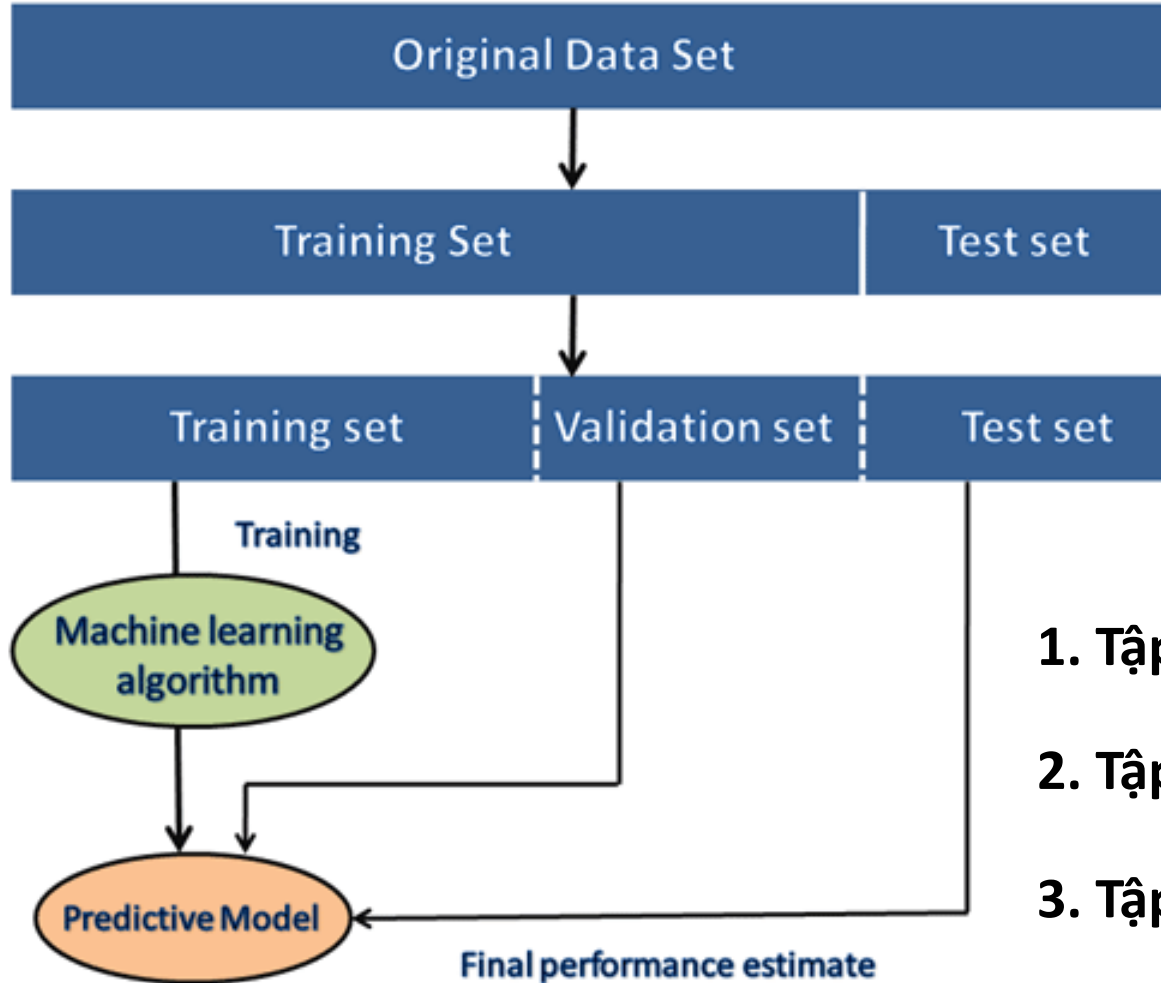
Dữ liệu có vai trò quan trọng trong việc xây dựng ứng dụng học máy cho bất kỳ bài toán nào.

Chất lượng và **khối lượng** dữ liệu ảnh hưởng trực tiếp đến mô hình học máy.

Dữ liệu từ các nguồn sau khi được tổng hợp và xử lý sẽ thu được các tập dữ liệu – **Datasets** phục vụ cho việc xây dựng model.



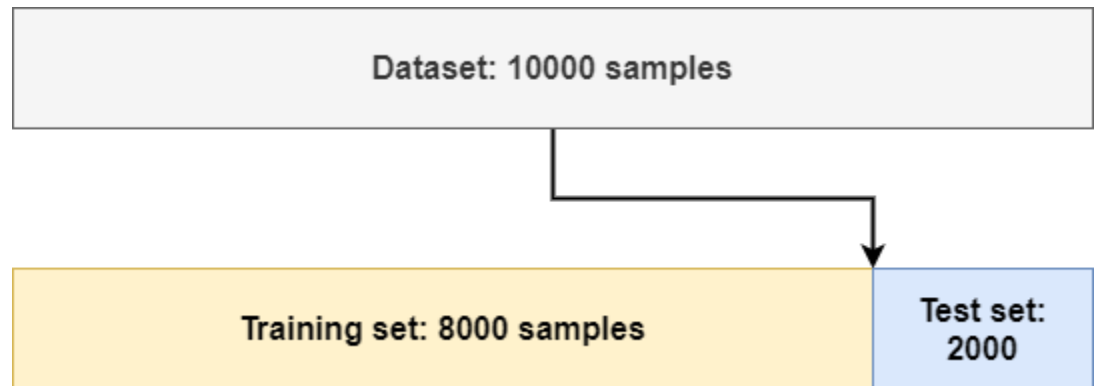
Dataset



1. Tập huấn luyện (Training Set)
2. Tập kiểm tra (Test Set)
3. Tập kiểm chéo (Validation Set)

Dataset

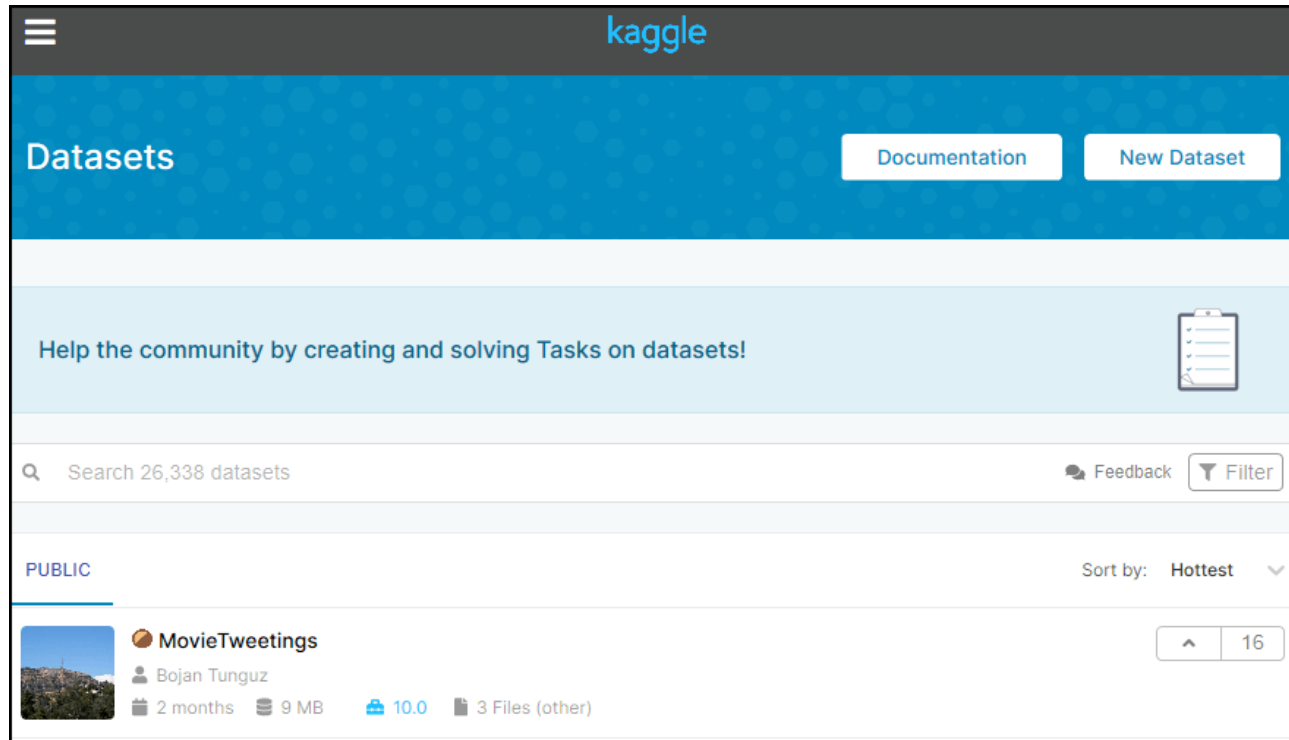
- **Tập huấn luyện (Training Set)** bao gồm các điểm dữ liệu sử dụng trực tiếp trong việc xây dựng mô hình.
- **Tập kiểm tra (Test set)** gồm các dữ liệu được dùng để đánh giá hiệu quả của mô hình. Tập kiểm tra đại diện cho dữ liệu mà mô hình chưa từng thấy, có thể xuất hiện trong quá trình vận hành mô hình trên thực tế.



- Để đảm bảo tính phổ quát, dữ liệu kiểm tra không được sử dụng trong quá trình xây dựng mô hình.
- Điều kiện cần để một mô hình hiệu quả: Kết quả đánh giá trên tập huấn luyện và tập kiểm tra đều cao.

Một số nguồn Dataset cho ML

Kaggle



- Kaggle là một trong những nguồn cung cấp dữ liệu tốt nhất cho các nhà khoa học dữ liệu và những người học về ML.
- <https://www.kaggle.com/datasets>.






Một số nguồn Dataset cho ML

UCI Machine Learning Repository



Browse Through: 488 Data Sets

Table View [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (360) Regression (107) Clustering (90) Other (55)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Attribute Type Categorical (38) Numerical (325) Mixed (55)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Data Type Multivariate (374) Univariate (24) Sequential (51) Time-Series (99) Text (55) Domain-Theory (23) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Area Life Sciences (110) Physical Sciences (52) CS / Engineering (178)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998

- UCI là một nguồn cung cấp Dataset tuyệt vời cho việc xây dựng các model học máy. Ra đời năm 1987, UCI được các sinh viên, giáo sư, nhà nghiên cứu sử dụng rộng rãi
- <https://archive.ics.uci.edu/ml/index.php>.

Một số nguồn Dataset cho ML

Datasets via AWS

Registry of Open Data on AWS



About

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS.](#)

See [all usage examples](#) for datasets listed in this registry.

See datasets from [Facebook Data for Good](#), [NASA Space Act Agreement](#), [NIH STRIDES](#), [NOAA Big Data Program](#), [Space Telescope Science Institute](#), and [Amazon Sustainability Data Initiative](#).

Search datasets (currently 190 matching datasets)

Add to this registry

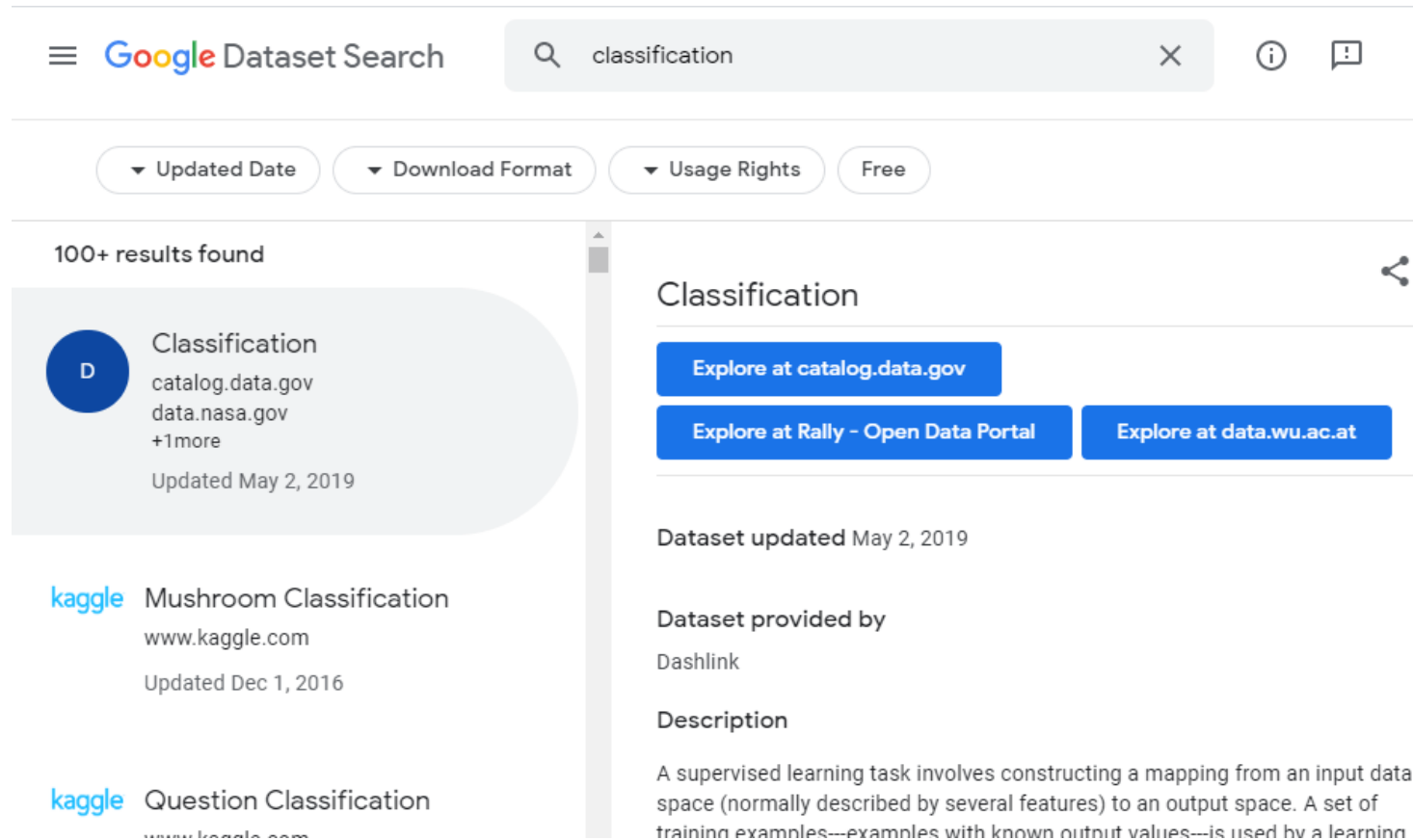
If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub repository](#).

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

- <https://registry.opendata.aws/> .

Một số nguồn Dataset cho ML

Google's Dataset Search Engine

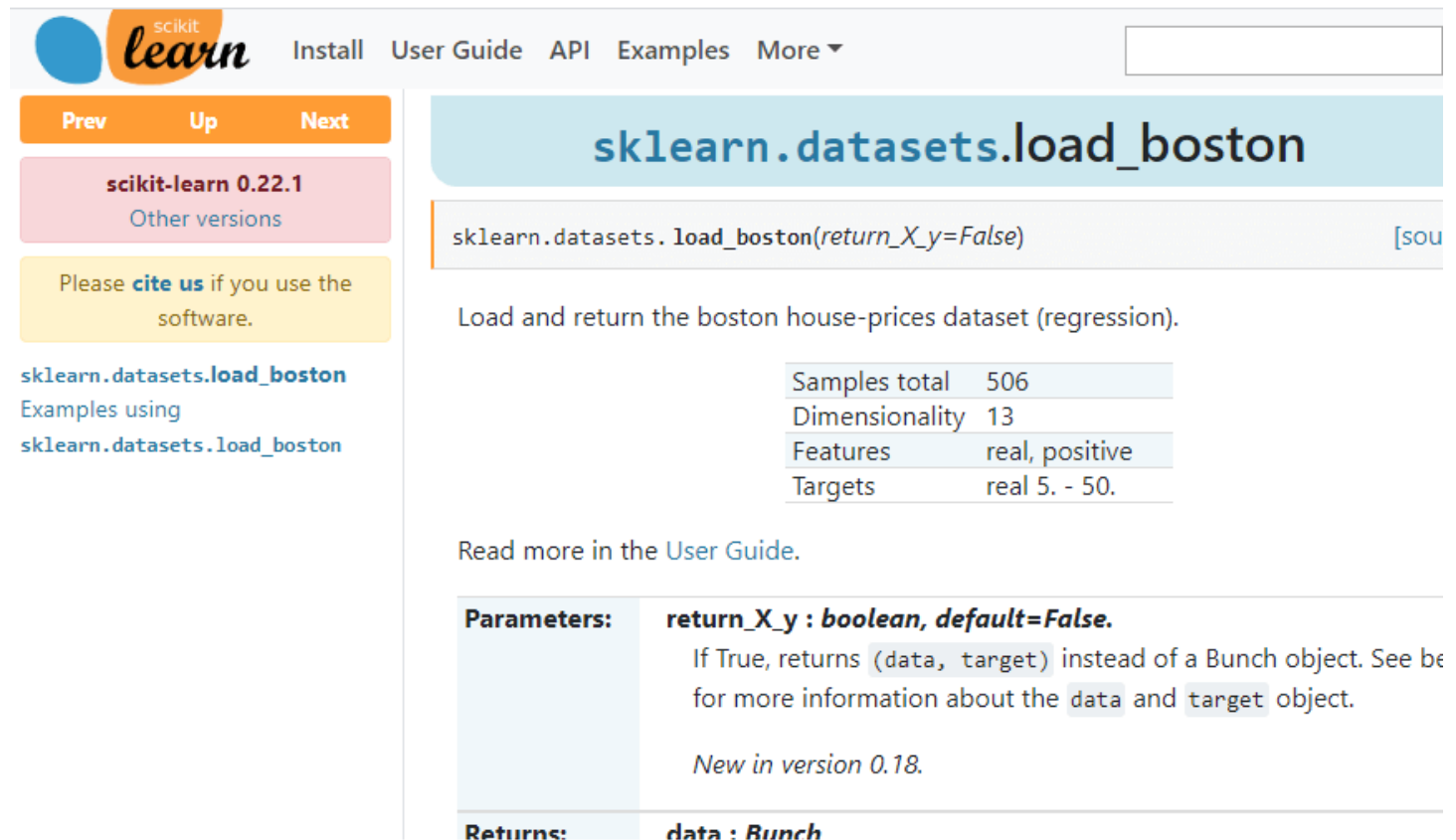


The screenshot shows the Google Dataset Search interface. At the top, the search bar contains the word "classification". Below the search bar, there are filters for "Updated Date", "Download Format", "Usage Rights", and "Free". The results section shows "100+ results found". The first result is a dataset titled "Classification" with sources "catalog.data.gov", "data.nasa.gov", and "+1more", updated on May 2, 2019. Below this, there are two more results from Kaggle: "Mushroom Classification" and "Question Classification". On the right side, there is a detailed view of the "Classification" dataset, including buttons to "Explore at catalog.data.gov", "Explore at Rally - Open Data Portal", and "Explore at data.wu.ac.at". It also shows the dataset was updated on May 2, 2019, and is provided by Dashlink. The description states: "A supervised learning task involves constructing a mapping from an input data space (normally described by several features) to an output space. A set of training examples—examples with known output values—is used by a learning

- <https://toolbox.google.com/datasetsearch>.

Một số nguồn Dataset cho ML

Scikit-learn dataset



The screenshot shows the Scikit-learn website interface. On the left, there's a sidebar with the Scikit-learn logo, navigation links (Install, User Guide, API, Examples, More), and version information (scikit-learn 0.22.1). The main content area displays the function `sklearn.datasets.load_boston`. Below the function name, the code `sklearn.datasets.load_boston(return_X_y=False)` is shown. A description states: "Load and return the boston house-prices dataset (regression)." A table provides dataset statistics: Samples total (506), Dimensionality (13), Features (real, positive), and Targets (real 5. - 50.). The "Parameters" section details the `return_X_y` parameter, and the "Returns" section indicates it returns a `Bunch` object.

Parameter	Value
Samples total	506
Dimensionality	13
Features	real, positive
Targets	real 5. - 50.

- Các ví dụ trong việc xây dựng model trong môn học sẽ chủ yếu lấy từ nguồn này
- <https://scikit-learn.org/stable/datasets/index.html>.

3. Học có giám sát (supervised learning)

Học có giám sát là gì?

Một thuật toán học máy được gọi là học có giám sát (supervised learning) nếu việc xây dựng mô hình dự đoán mối quan hệ giữa đầu vào và đầu ra được thực hiện dựa trên các cặp (đầu vào - input, đầu ra - label) đã biết trong tập huấn luyện. Đây là nhóm thuật toán phổ biến nhất trong các thuật toán machine learning.

Tập dữ liệu học (Training data) bao gồm các quan sát (Examples, Observations), mà mỗi quan sát được gắn kèm với một giá trị đầu ra mong muốn (Label)

Sample
↓
Label



dog

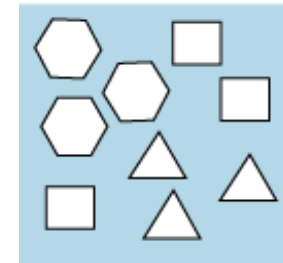


cat



horse

Labeled Data



Labels



Học có giám sát là gì?

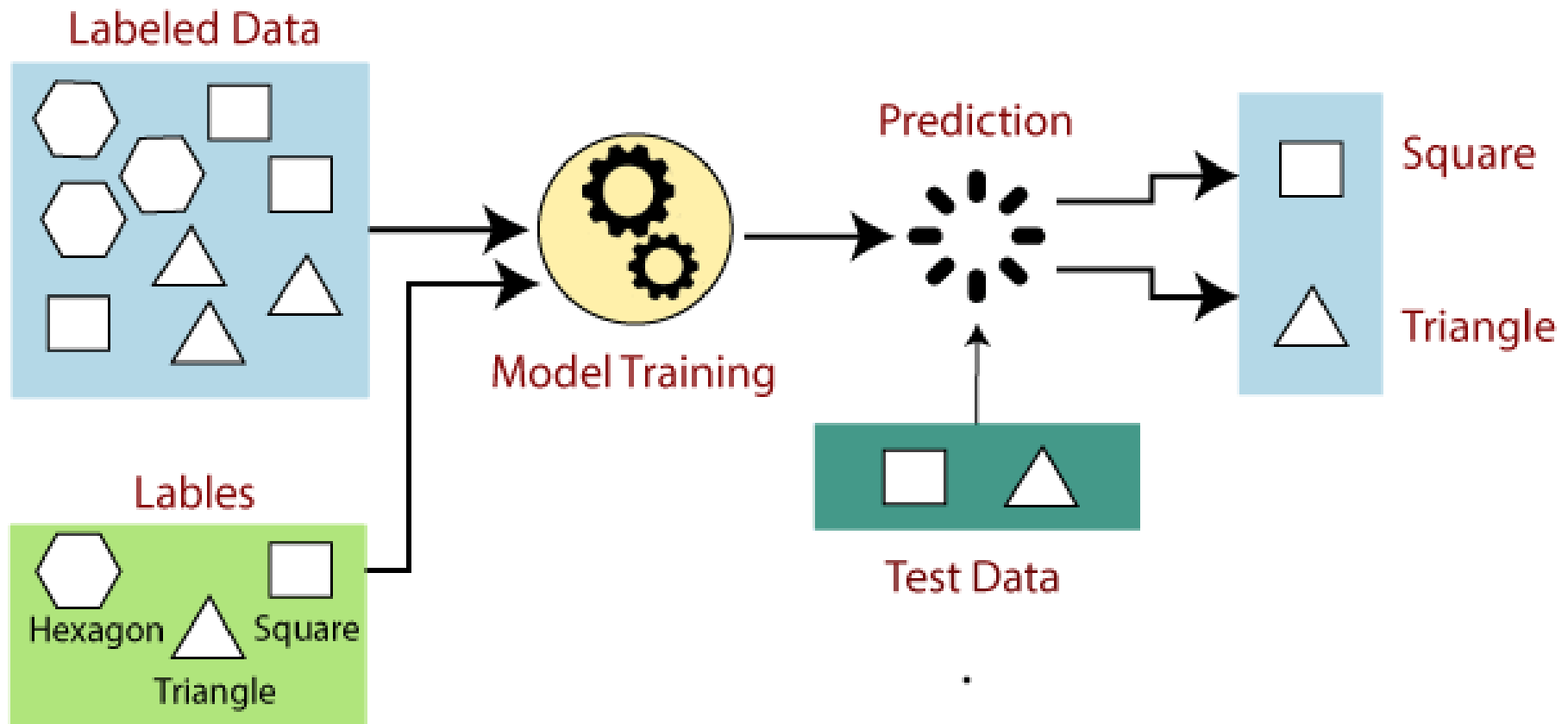
Dataset sẽ bao gồm:

- Các thuộc tính đầu vào (Biến độc lập) – Features (input)
- Thuộc tính mục tiêu (Biến phụ thuộc) – Target (label)

Biến độc lập (Features)								Biến phụ thuộc (Target) - Label
id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
Patient_01	63	Male	Typical angina	145	233	150	6	0
Patient_02	67	Male	Asymptomatic	160	286	108	3	1
Patient_03	67	Male	Asymptomatic	120	229	129	7	1
Patient_04	37	Male	Non-anginal pain	130	250	187	3	0
Patient_05	41	Female	Atypical angina	130	204	172		0
Patient_16	56	Male	Atypical angina	120	236	178	3	0
Patient_07	62	Female	Asymptomatic	140	268	160	3	1
Patient_08	57	Female	Asymptomatic	120	354	163	3	0
Patient_19	63	Male	Asymptomatic	130	254	147	7	1
Patient_10	53	Male	Asymptomatic	140	203	155	7	1
Patient_110	57	Male	Asymptomatic	140	192	148	6	0
Patient_120	56	Female	Atypical angina	140	294	153	3	0
Patient_130	56	Male	Non-anginal pain	130	256	142	6	1
Patient_140	44	Male	Atypical angina	120	263	173	7	0
Patient_150	52	Male	Non-anginal pain	172	199	162	7	0
Patient_160	57	Male	Non-anginal pain	150	168	174	3	0
Patient_170	48	Male	Atypical angina	110	229	168	7	1
Patient_180	54	Male	Asymptomatic	140	239	160	3	0
Patient_190	48	Female	Non-anginal pain	130	275	139	3	0
Patient_200	49	Male	Atypical angina	130	266	171	3	0
Patient_21	64	Male	Typical angina	110	211	144		0

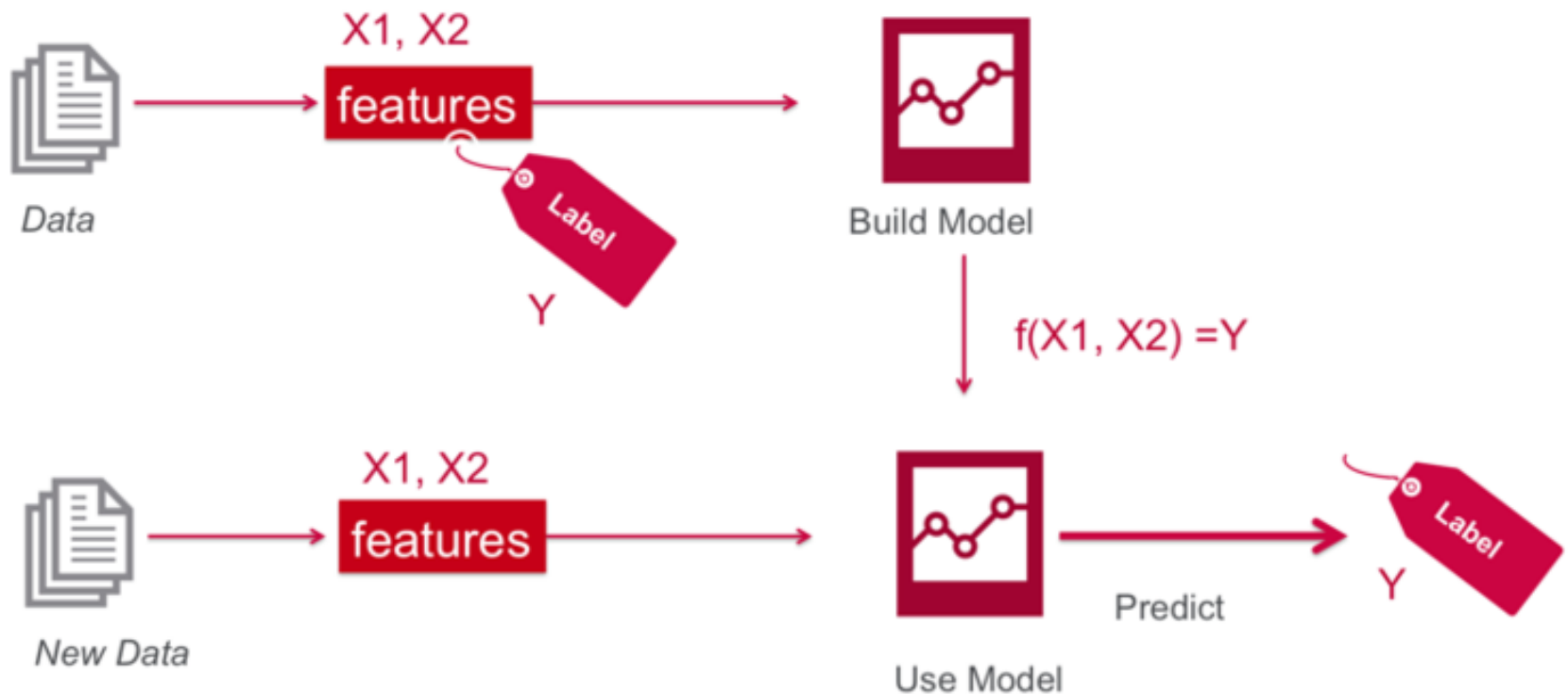
Học có giám sát là gì?

Quá trình huấn luyện và kiểm thử với Mô hình học máy có giám sát



Học có giám sát là gì?

- Bản chất của Supervised learning là học một hàm f phù hợp với tập dữ liệu hiện có và có khả năng tổng quát hóa cao.
- Hàm học được sau đó sẽ dùng để dự đoán cho các quan sát mới.





Thank you!