

Chapter 1: INTRODUCTION

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.” -Arthur Samuel (1959)



Traditional Programming



Machine Learning



Định nghĩa

Definition (Machine learning (ML))

Tom Mitchell - "Máy tính được gọi là học từ kinh nghiệm (dữ liệu) E với tác vụ (dự đoán, phân lớp, gom nhóm) T và được đánh giá bởi độ đo (độ chính xác) P nếu máy tính khiến tác vụ T này cải thiện được độ chính xác P thông qua dữ liệu E cho trước."

Defining the Learning Task

Improve on task T, with respect to
performance metric P, based on experience E

T: Playing checkers

P: Percentage of games won against an arbitrary opponent

E: Playing practice games against itself

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors

P: Average distance traveled before a human-judged error

E: A sequence of images and steering commands recorded while observing a human driver.

T: Categorize email messages as spam or legitimate.

P: Percentage of email messages correctly classified.

E: Database of emails, some with human-given labels

Machine learning vs AI

Ta cần AI để tạo ra các thiết bị thông minh, nhưng để chúng thực sự thông minh và ứng xử như con người, ta cần ML.

Chinh phục AI mặc dù vẫn là mục đích tối thượng của machine learning, nhưng hiện tại machine learning tập trung vào những mục tiêu ngắn hạn hơn như: Làm cho máy tính có những khả năng nhận thức cơ bản của con người như nghe, nhìn, hiểu được ngôn ngữ, giải toán, lập trình, ...

Các ứng dụng như:

- Trợ lý ảo: IBM Watson, Google Now, Cortana, alexa, Siri, Bixby, ...
- Robot: Big Dog robot (chó robot trong quân đội mỹ), asimo (Honda), ...
- Các hệ thống phương tiện thông minh, xe không người lái (Google), xe tự lái trong quân đội, ...

Machine learning vs Big Data

Big Data thực chất không phải là một ngành khoa học chính thống, chỉ là cụm từ được xây dựng bởi truyền thông. Big Data là một hệ quả tất yếu của mạng Internet ngày càng có nhiều kết nối. Như Facebook, Twitter, youtube, ...



Machine learning vs Big Data

Với kho dữ liệu đồ sộ và chứa một khối tri thức khổng lồ. Và từ những dữ liệu này ta có thể hiểu thêm về con người và xã hội. Cụ thể:

- Từ danh sách tìm kiếm của người dùng \Rightarrow sở thích của người dùng và giới thiệu những thứ phù hợp với nhu cầu và sở thích của người dùng.
- Từ mối quan hệ và tương tác của người dùng trên MXH \Rightarrow gom nhóm cộng đồng theo sở thích, công việc, ...
- Từ các tương tác của người sử dụng, có thể phát hiện ra các hành vi sai phạm, ...

Big Data chỉ thực sự bắt đầu từ khi ta hiểu được giá trị của thông tin ẩn chứa trong dữ liệu, và có đủ tài nguyên cũng như công nghệ để có thể khai thác chúng trên quy mô khổng lồ. Machine learning chính là thành phần mấu chốt của công nghệ đó.

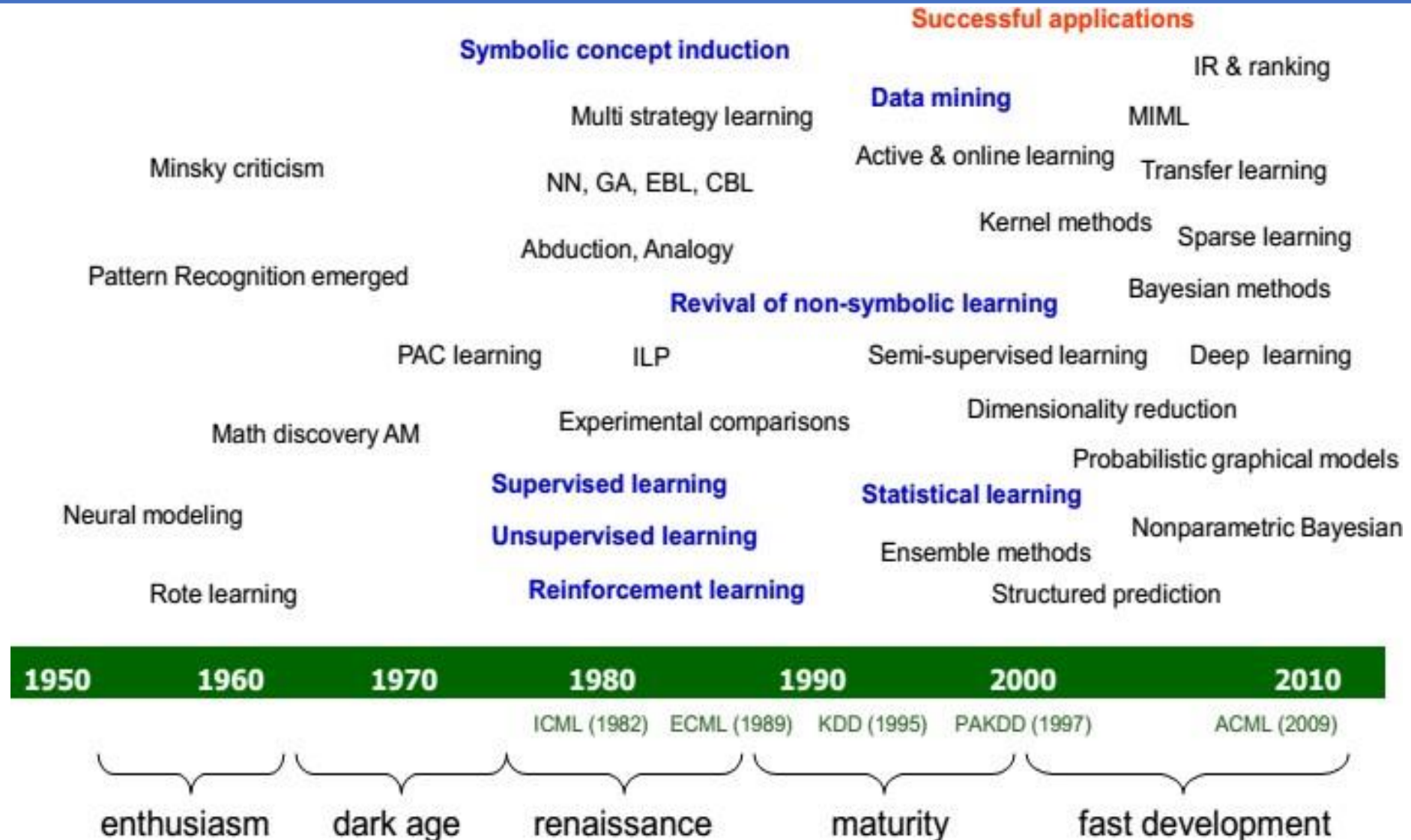
Machine learning vs Phân tích dự báo

ML có mối liên hệ mật thiết đối với thống kê. Tuy nhiên ML, không chỉ đơn thuần sử dụng các mô hình thống kê để ghi nhớ lại sự phân bố dữ liệu, nó có khả năng tổng quát hóa những gì đã được nhìn thấy và dự đoán cho những trường hợp chưa được nhìn thấy. Như vậy ta có thể nói ML có thể dự đoán tương lai, nhưng chỉ khi tương lai có mối liên hệ mật thiết với hiện tại.

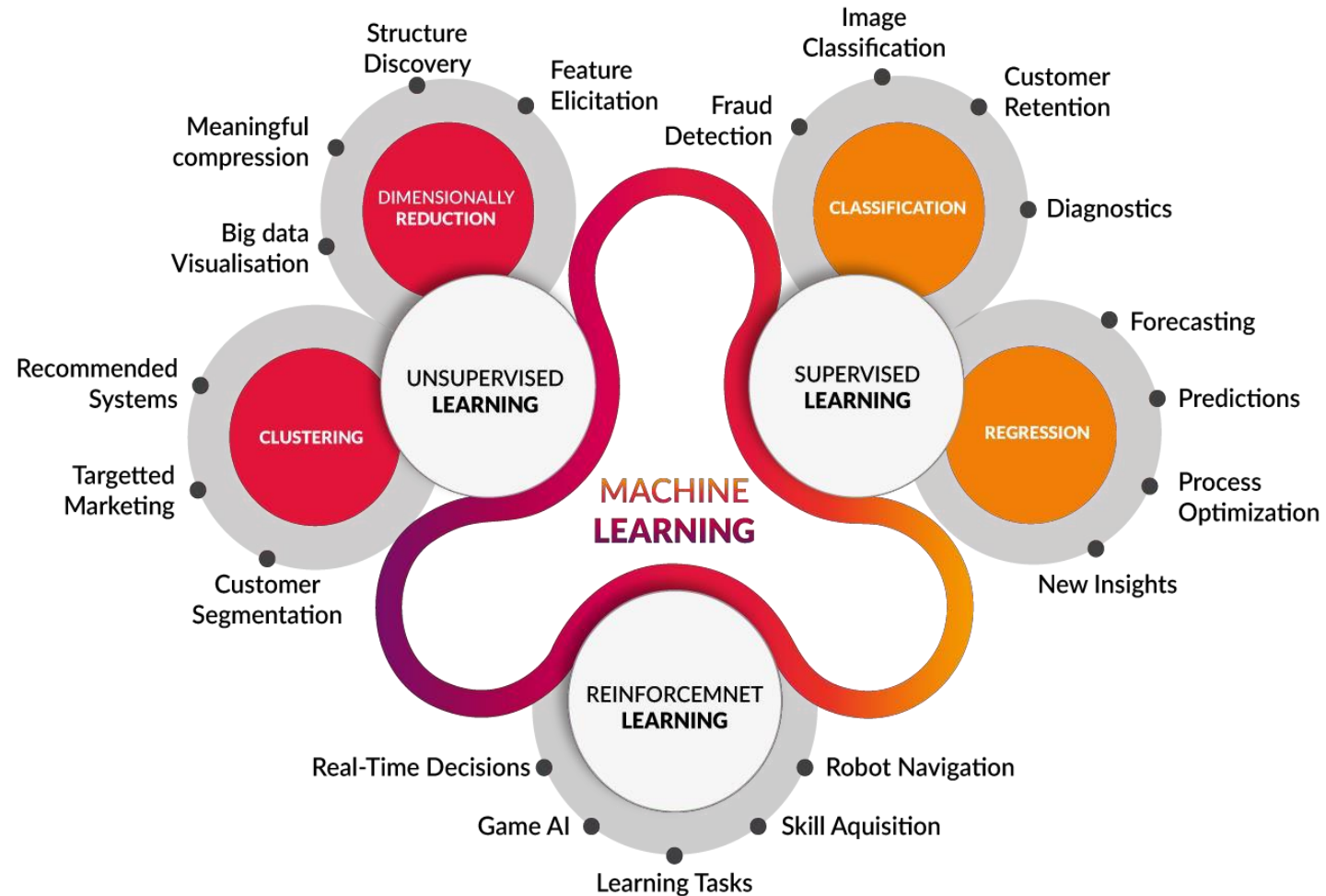
Một số ứng dụng dự đoán như:

- Dự đoán xu hướng thị trường, bất động sản, chính trị, ...
- Dự báo thời tiết, khi hậu, thiên tai, ...
- ...

Quá trình Phát Triển (Pre-Deep Learning)



Các phương pháp học máy



Supervised learning

Definition

Supervised learning là khi ta có tập quan sát $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ và tập nhãn tương ứng $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$, trong đó x_i, y_i là các vector. Các cặp dữ liệu biết trước $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ được gọi là training data. Từ tập training data, ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập \mathcal{X} sang một phần tử tương ứng của tập \mathcal{Y} :

$$y_i \approx f(x_i), \forall i = 1, 2, \dots, N \quad (1)$$

Vì vậy kết quả của dạng toán này phụ thuộc vào tập dữ liệu training set có tính "right answers".

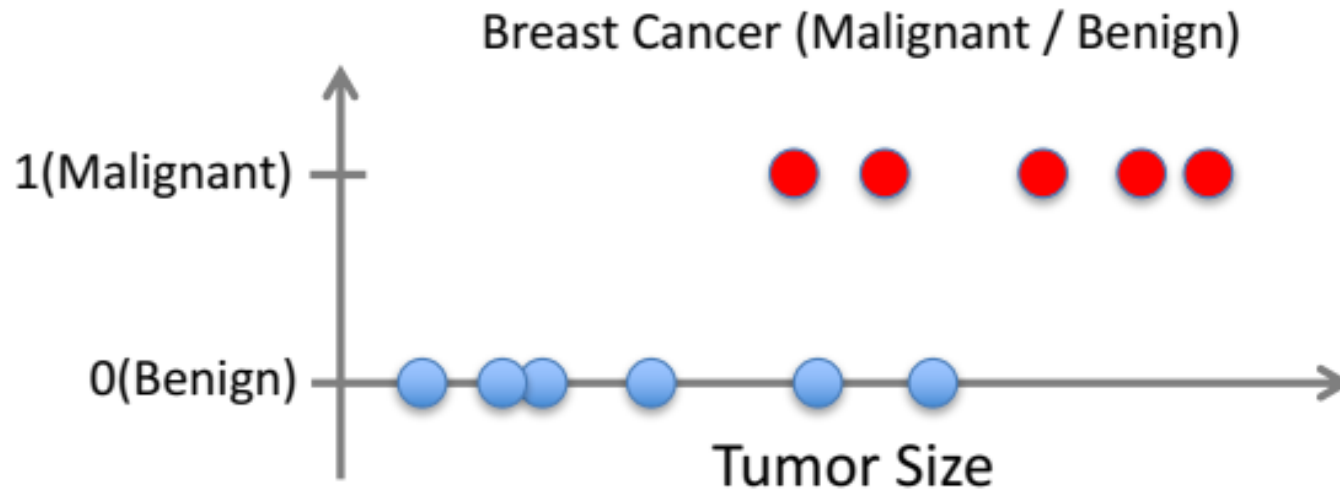
Supervised learning

Thuật toán supervised learning được chia nhỏ thành hai loại chính:

- Classification (Phân loại) nếu các label của input data được chia thành một số hữu hạn nhóm. Ví dụ: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không.
- Regression (Hồi quy) Nếu label không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ: một căn nhà rộng x (m^2), có y phòng ngủ và cách trung tâm thành phố z km sẽ có giá là bao nhiêu?

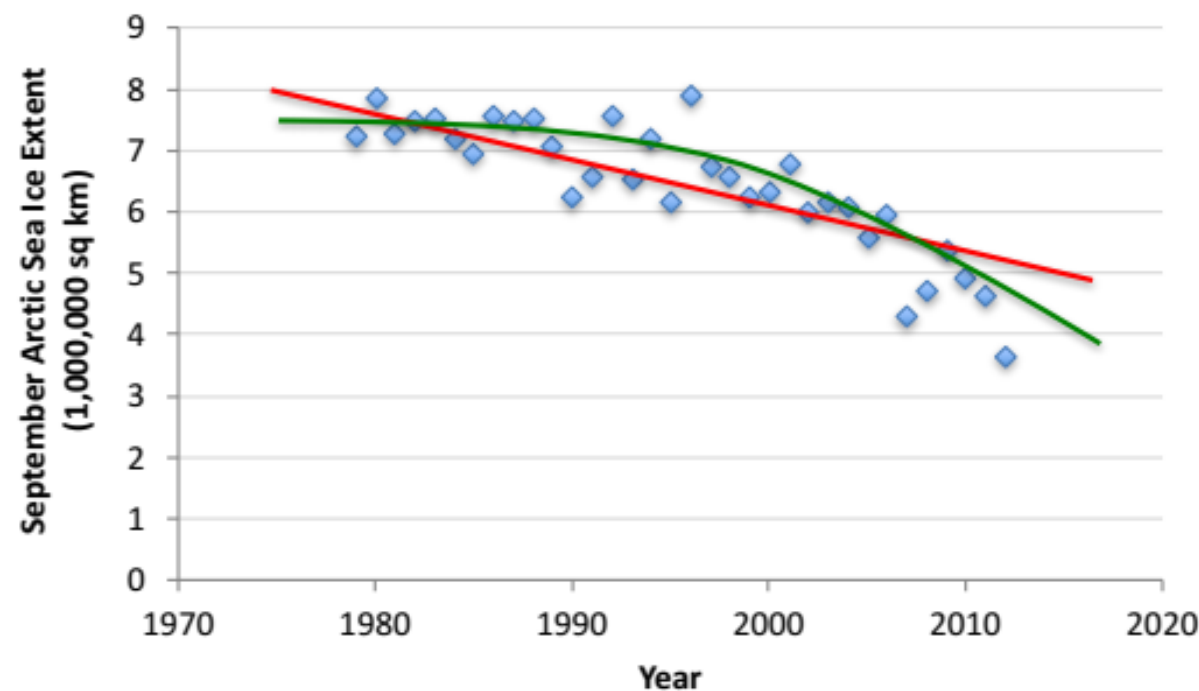
Supervised Learning: Classification

- Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
 - y is categorical == classification



Supervised Learning: Regression

- Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
 - y is real-valued == regression



Định nghĩa

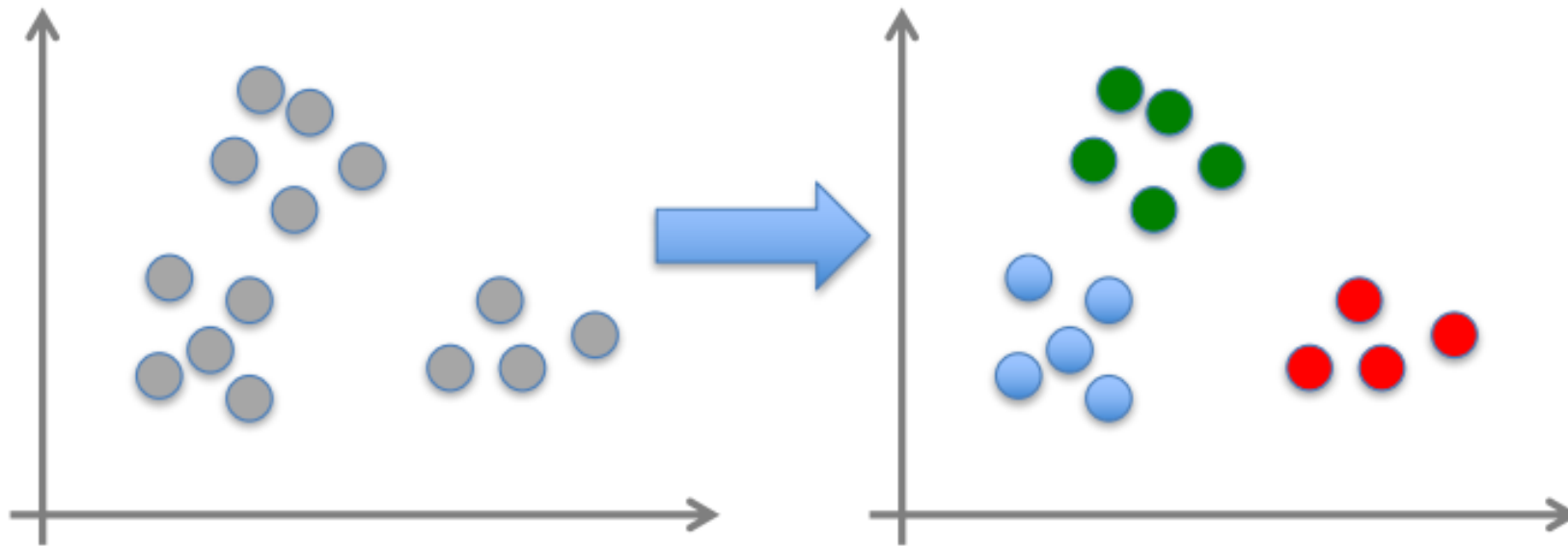
Unsupervised learning là khi chúng ta chỉ có dữ liệu quan sát đầu vào \mathcal{X} mà không biết nhãn \mathcal{Y} tương ứng. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó.

Unsupervised learning được chia nhỏ thành hai loại:

- Clustering: Là bài toán phân nhóm toàn bộ dữ liệu \mathcal{X} thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong nhóm.
VD: Phân nhóm khách hàng dựa trên hành vi mua hàng, phát hiện cộng đồng trong mạng xã hội, ...
- Association: Là bài toán khám phá quy luật dựa trên tập dữ liệu cho trước.

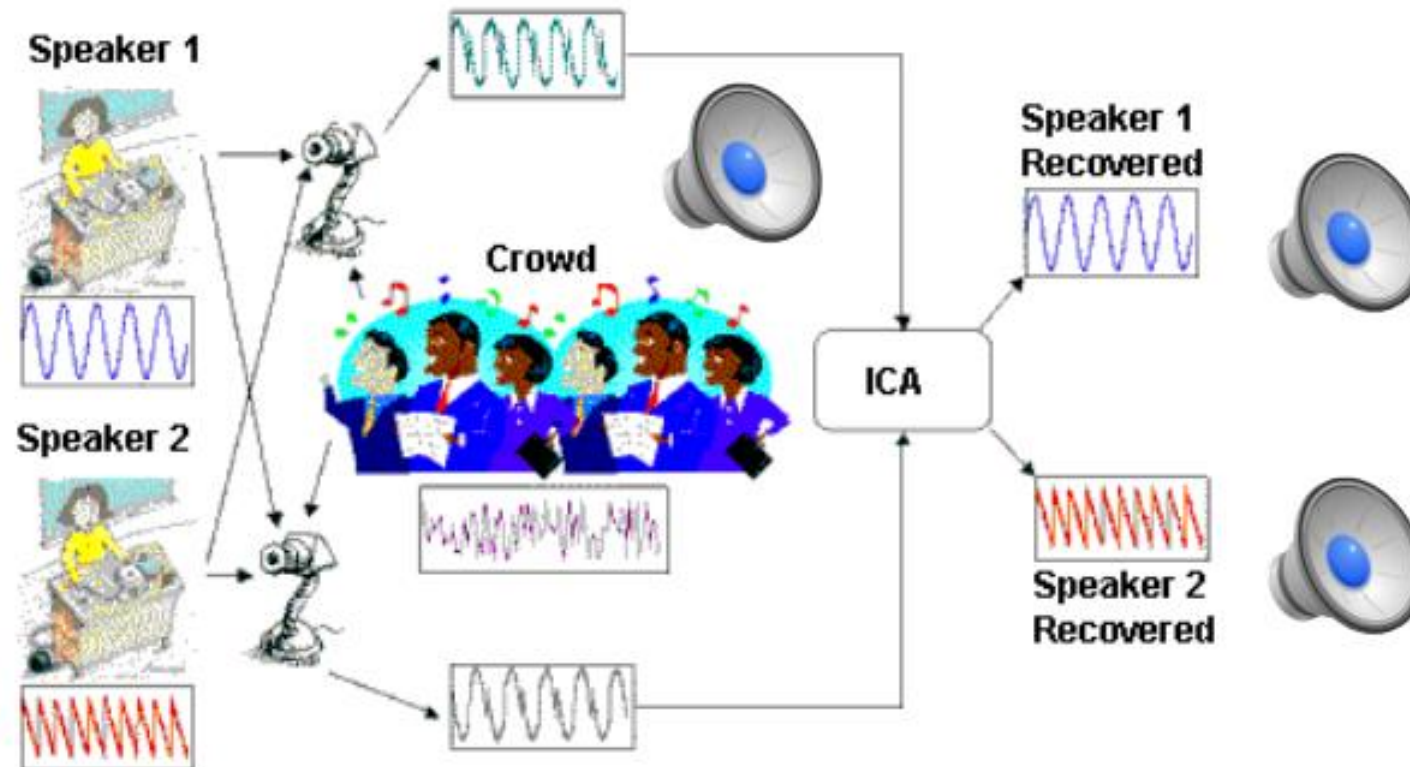
Unsupervised Learning

- Given x_1, x_2, \dots, x_n (without labels)
- Output hidden structure behind the x 's
 - E.g., clustering



Unsupervised Learning

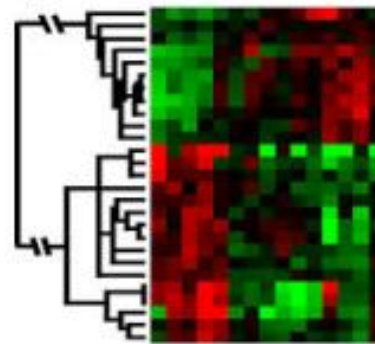
- Independent component analysis – separate a combined signal into its original sources



When Do We Use Machine Learning?

ML is used when:

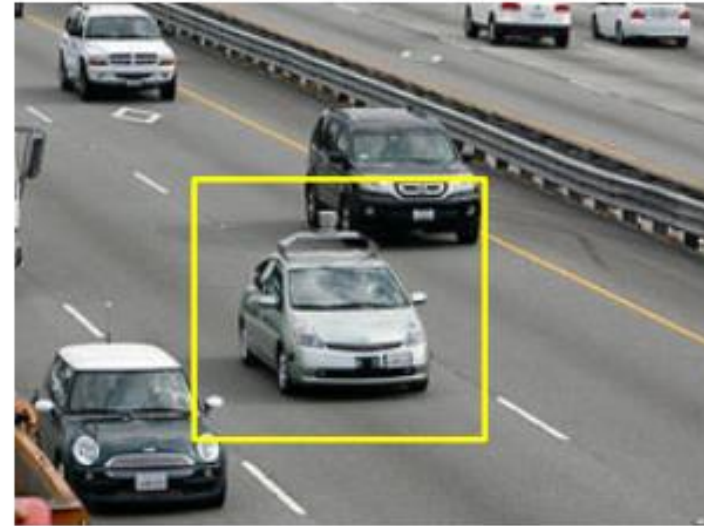
- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)



Learning isn't always useful:

- There is no need to “learn” to calculate payroll

Autonomous Cars



- Nevada made it legal for autonomous cars to drive on roads in June 2011
- As of 2013, four states (Nevada, Florida, California, and Michigan) have legalized autonomous cars

Penn's Autonomous Car →
(Ben Franklin Racing Team)

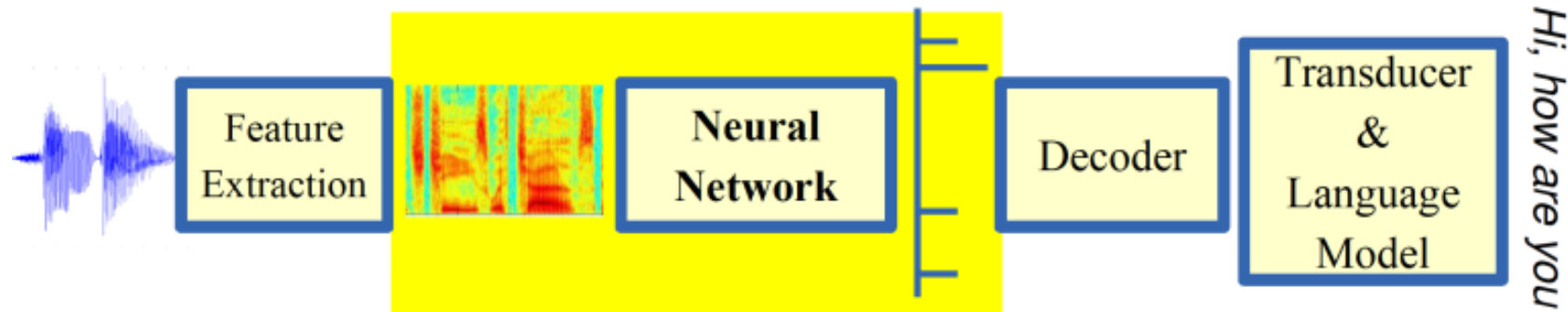


Scene Labeling via Deep Learning

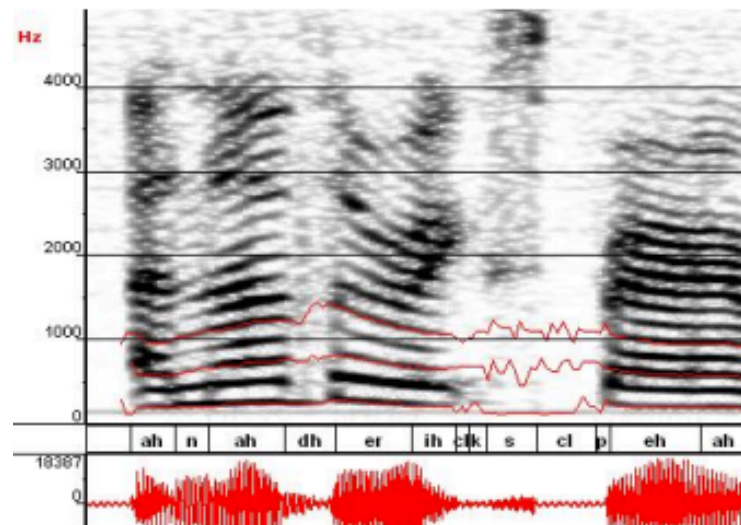


Machine Learning in Automatic Speech Recognition

A Typical Speech Recognition System



ML used to predict of phone states from the sound spectrogram



Deep learning has state-of-the-art results

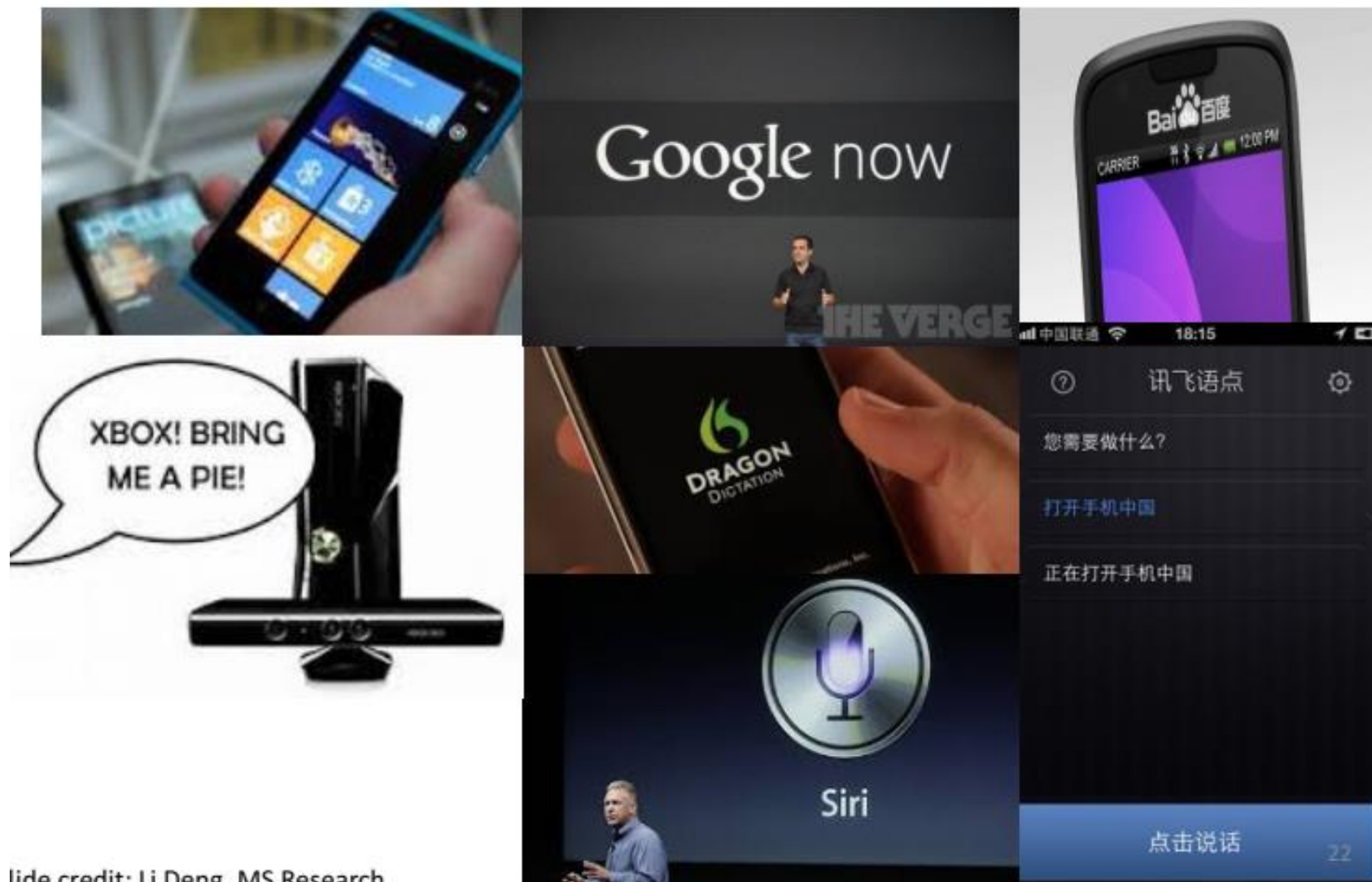
# Hidden Layers	1	2	4	8	10	12
Word Error Rate %	16.0	12.8	11.4	10.9	11.0	11.1

Baseline GMM performance = 15.4%

[Zeiler et al. "On rectified linear units for speech recognition" ICASSP 2013]

21

Impact of Deep Learning in Speech Technology

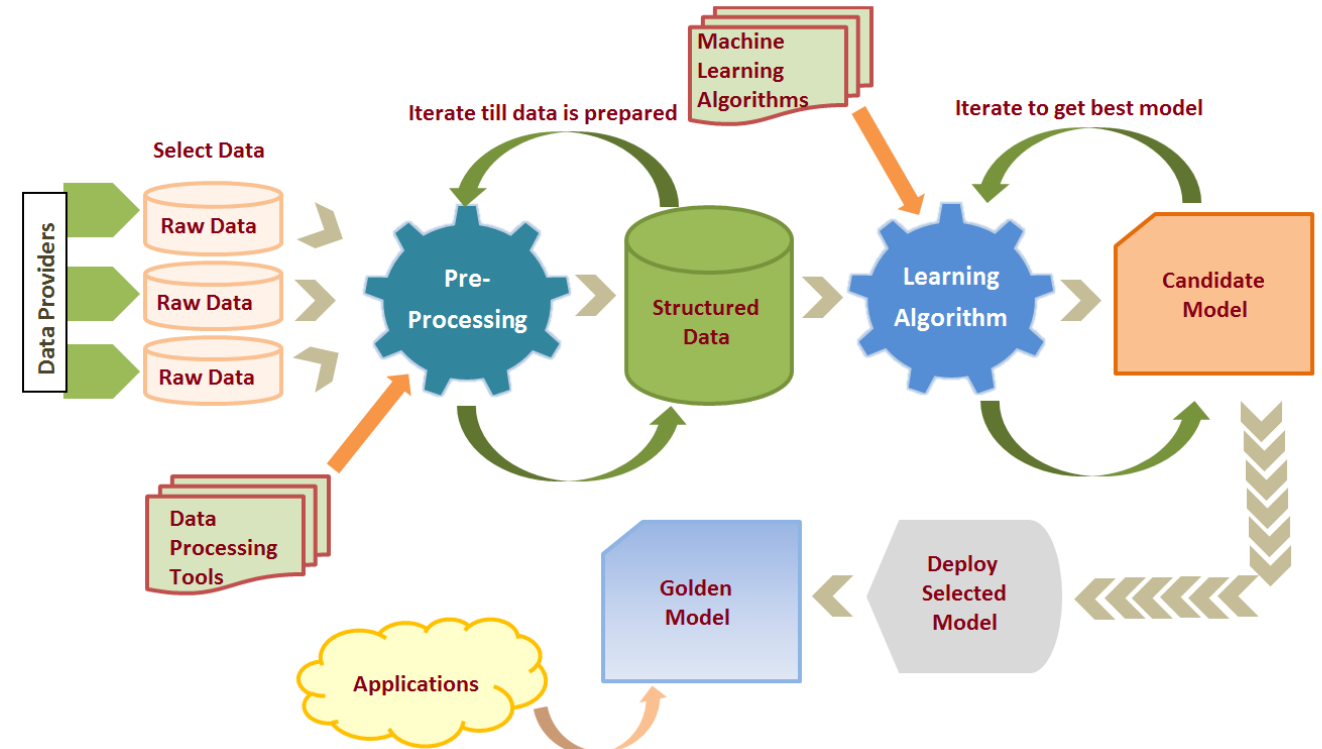


lide credit: Li Deng, MS Research

QUY TRÌNH XÂY DỰNG HỆ THỐNG

Các bước xây dựng mô hình học máy

- Thu thập dữ liệu
- Chuẩn bị dữ liệu
- Lựa chọn mô hình
- Huấn luyện mô hình
- Đánh giá mô hình
- Thay đổi tham số/mô hình
- Áp dụng mô hình



Thu thập dữ liệu (Data Collection)

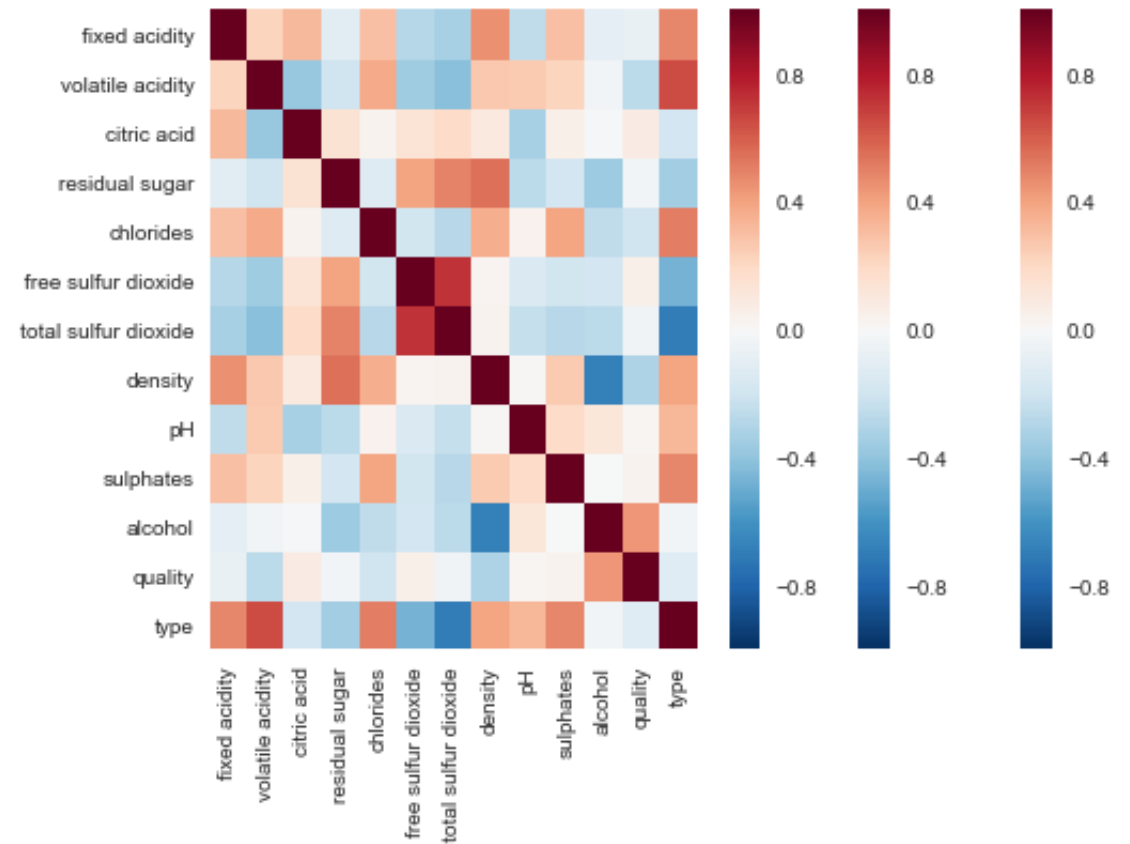
- Chất lượng và khối lượng dữ liệu ảnh hưởng trực tiếp đến mô hình học máy
- Dữ liệu thực tế hay dữ liệu phòng lab
- Dữ liệu thực tế: dữ liệu của bạn hay dữ liệu nguồn khác
- Đánh giá dữ liệu: độ lớn, nguồn, độ phức tạp, độ mất mát ...
- Lưu trữ dữ liệu: Tập trung hay phân tán

Chuẩn bị dữ liệu (Data Preparation)

- Lý do chuẩn bị dữ liệu
 - Phù hợp với thuật toán, công cụ
 - Dữ liệu không sạch: không đầy đủ, nhiễu, không nhất quán
- Các vấn đề trong chuẩn bị dữ liệu
 - Khám phá dữ liệu
 - Làm sạch dữ liệu
 - Tích hợp dữ liệu
 - Biến đổi, rời rạc hóa và chuẩn hóa dữ liệu
 - Cân bằng dữ liệu
 - Rút gọn thuộc tính

Khám phá dữ liệu (Data exploration)

- Thuộc tính ảnh hưởng tới quyết định
- Mối liên hệ giữa các thuộc tính

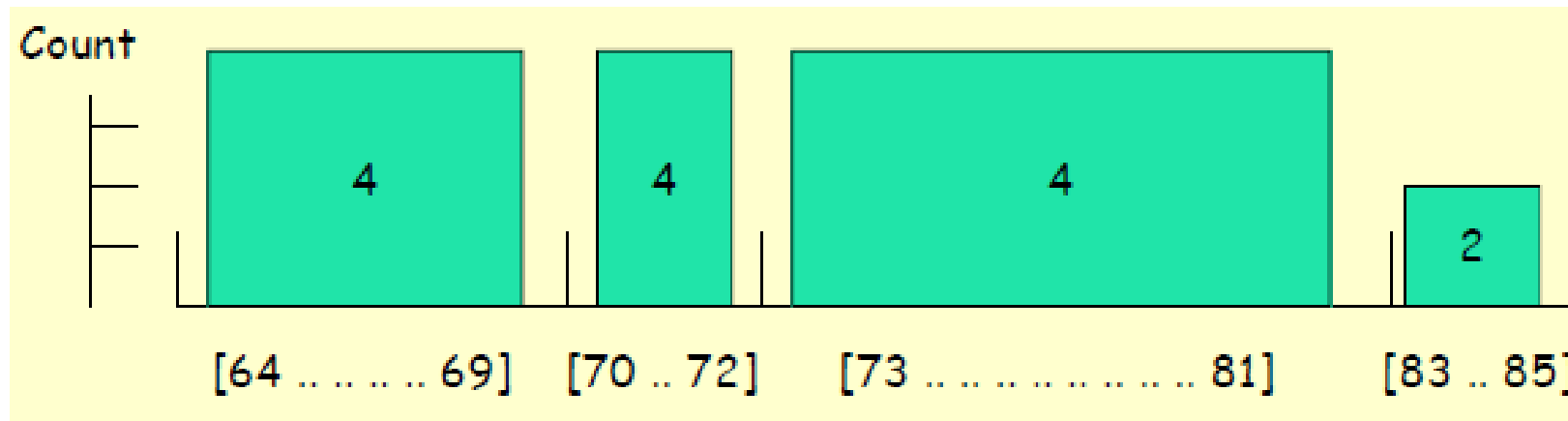


Làm sạch dữ liệu (Data cleansing)

- Dữ liệu mất mát (missing data)
 - Vì một lý do nào đó mà dữ liệu bị mất mát (dữ liệu không được thu thập, lỗi lưu trữ ...)
 - Một vài giải pháp:
 - Bỏ qua bản ghi hoặc thuộc tính chứa thông tin bị mất mát
 - Điền thông tin mới với: ngẫu nhiên hoặc liên quan đến các dữ liệu xung quanh
- Dữ liệu có nhiều: giá trị không phù hợp ...
- Dữ liệu trùng lặp
 - Xảy ra khi tổng hợp nhiều nguồn tin khác nhau.

Biến đổi dữ liệu

- Rời rạc hóa (discretization)
 - Biến đổi dữ liệu từ dạng liên tục (continuous) sang rời rạc (discrete)
 - Nhiều model yêu cầu dữ liệu ở dạng rời rạc: cây phân lớp
 - Cho phép thu gọn dữ liệu



Biến đổi dữ liệu

- Chuẩn hóa dữ liệu (normalization)
 - Với nhiều mô hình dựa trên độ đo khoảng cách (distance-based method), việc chuẩn hóa giúp cho các thuộc tính có sự ảnh hưởng cân bằng với nhau
 - Ví dụ tuổi từ 0-99, lương từ 1 triệu VNĐ tới 1 tỉ VNĐ
 - Chuẩn hóa các thuộc tính về các khoảng tương tự nhau hoặc miền giá trị từ 0 tới 1
 - Các phương pháp:
 - min-max normalization

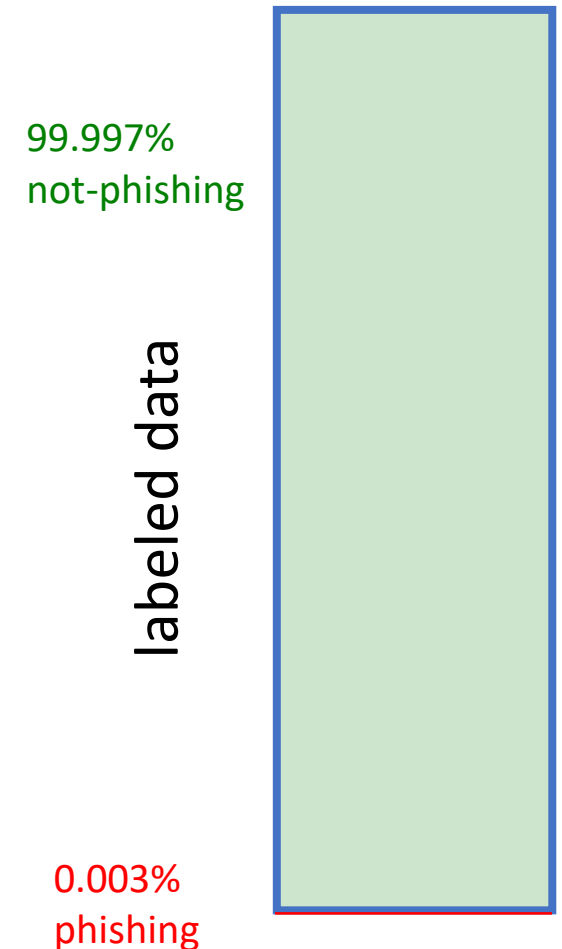
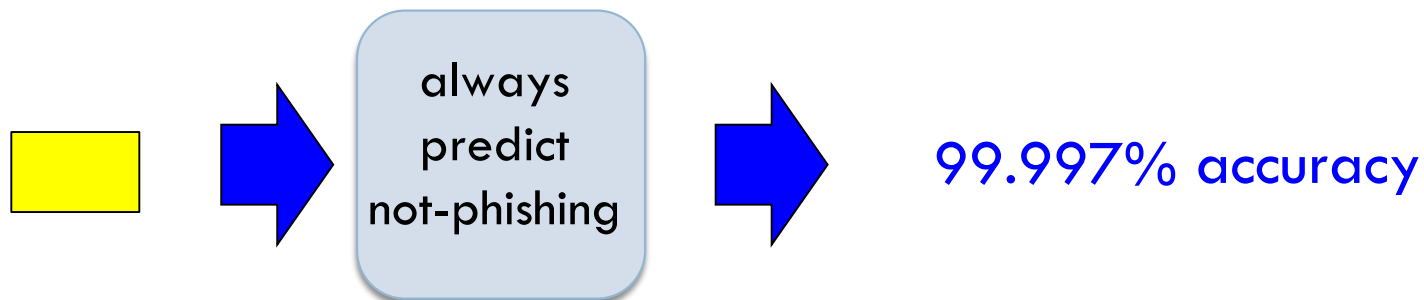
$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- z-score normalization (standardization)

$$z = \frac{x - \mu}{\sigma}$$

Cân bằng dữ liệu

- Dữ liệu bất cân bằng khi một lớp đối tượng có lượng bản ghi lớn hơn hẳn các lớp còn lại
- Bài toán phát hiện phishing
 - Trong 1 triệu email mới có khoảng 30 là phishing



Rút gọn thuộc tính (Feature Selection)

- Rút gọn thuộc tính là quá trình chọn tập con tối ưu các thuộc tính theo một số điều kiện nhất định.
- Tại sao phải rút gọn
 - Tăng hiệu quả của mô hình: tăng tốc độ, độ chính xác và giảm độ phức tạp
 - Trực quan hóa dữ liệu
 - Giảm bớt nhiễu và những ảnh hưởng không cần thiết

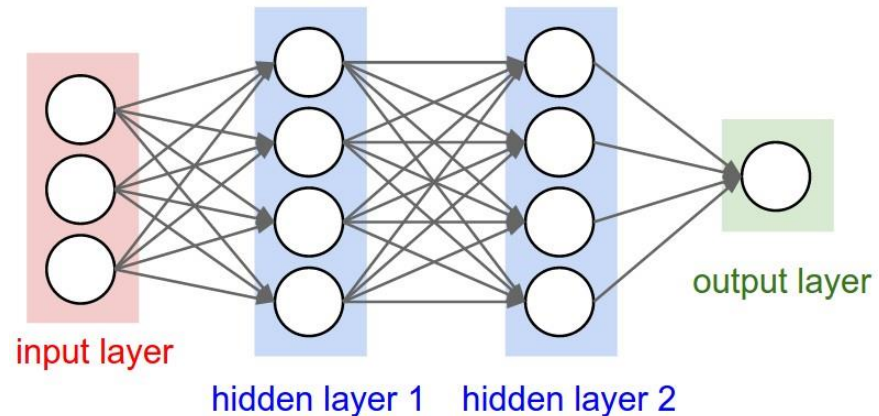
ID	Nhiệt độ	Đau đầu	Nôn mửa	Cúm
1	High	High	Yes	Yes
2	High	Low	No	No
3	Low	High	Yes	Yes
4	Low	Low	No	No

Lựa chọn mô hình (Model selection)

- Mô hình phù hợp với bài toán và dữ liệu
 - Phân loại ảnh, âm thanh hay văn bản
 - Dữ liệu rời rạc hay dữ liệu số
 - Nhiều hay ít thuộc tính
 - Phân loại hay phân cụm

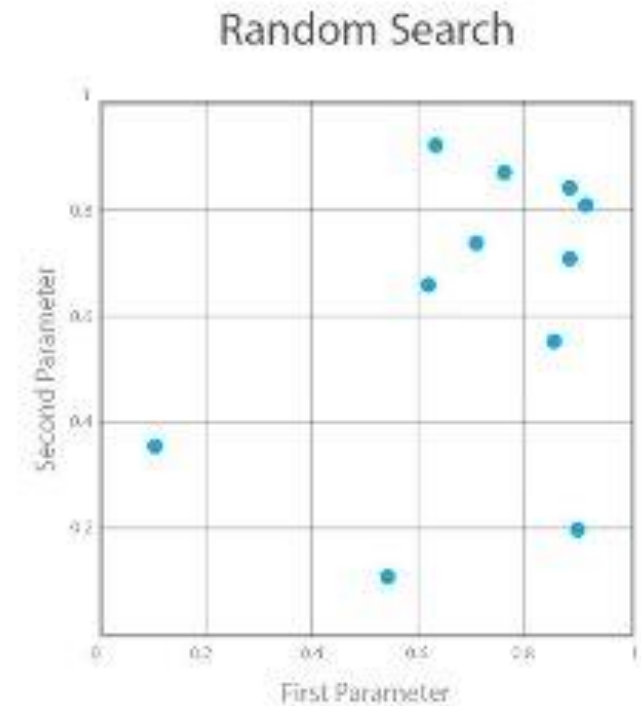
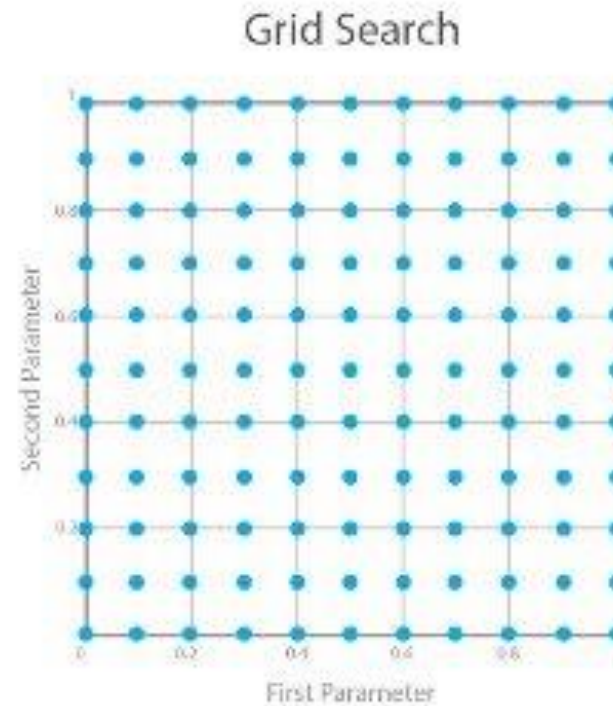
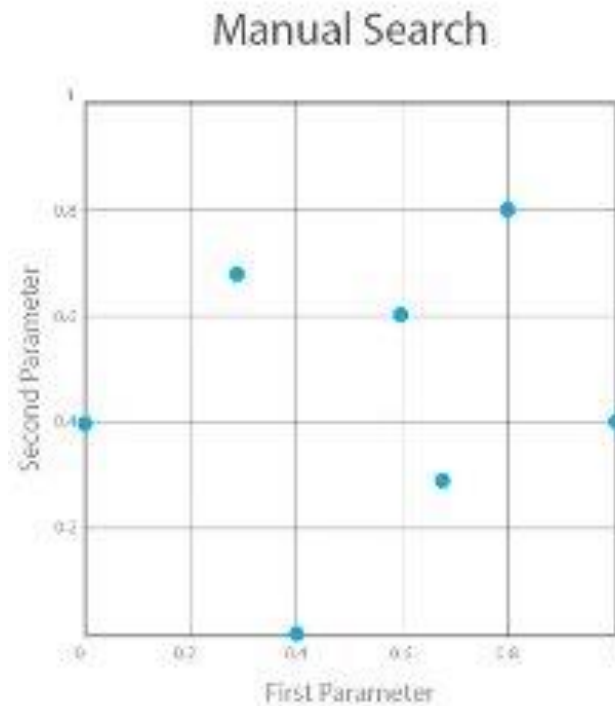
Lựa chọn mô hình

- Lựa chọn siêu tham số (hyper parameter)
 - Decision trees
 - Độ sâu, số lượng lá
 - SVM
 - Kernel trick/feature extraction
 - Boosting
 - Number of rounds
 - Neural network
 - Learning rate
 - Mini-batch size...



Lựa chọn mô hình

- Lựa chọn siêu tham số (hyper parameter)

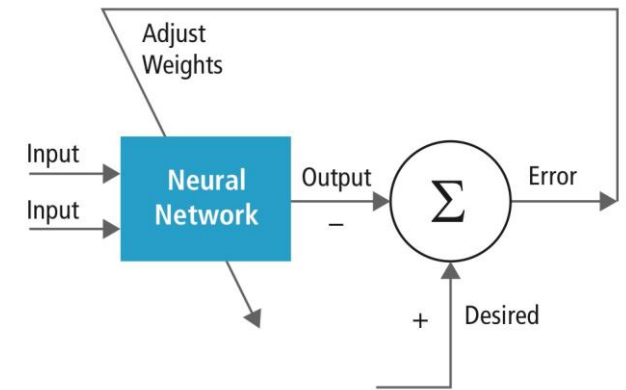


Huấn luyện mô hình (Training)

- Là các bước tìm kiếm giá trị các tham số của mô hình sao cho mô hình xấp xỉ được tốt nhất phân bố của dữ liệu

$$y = f(x)$$

output prediction function Input data



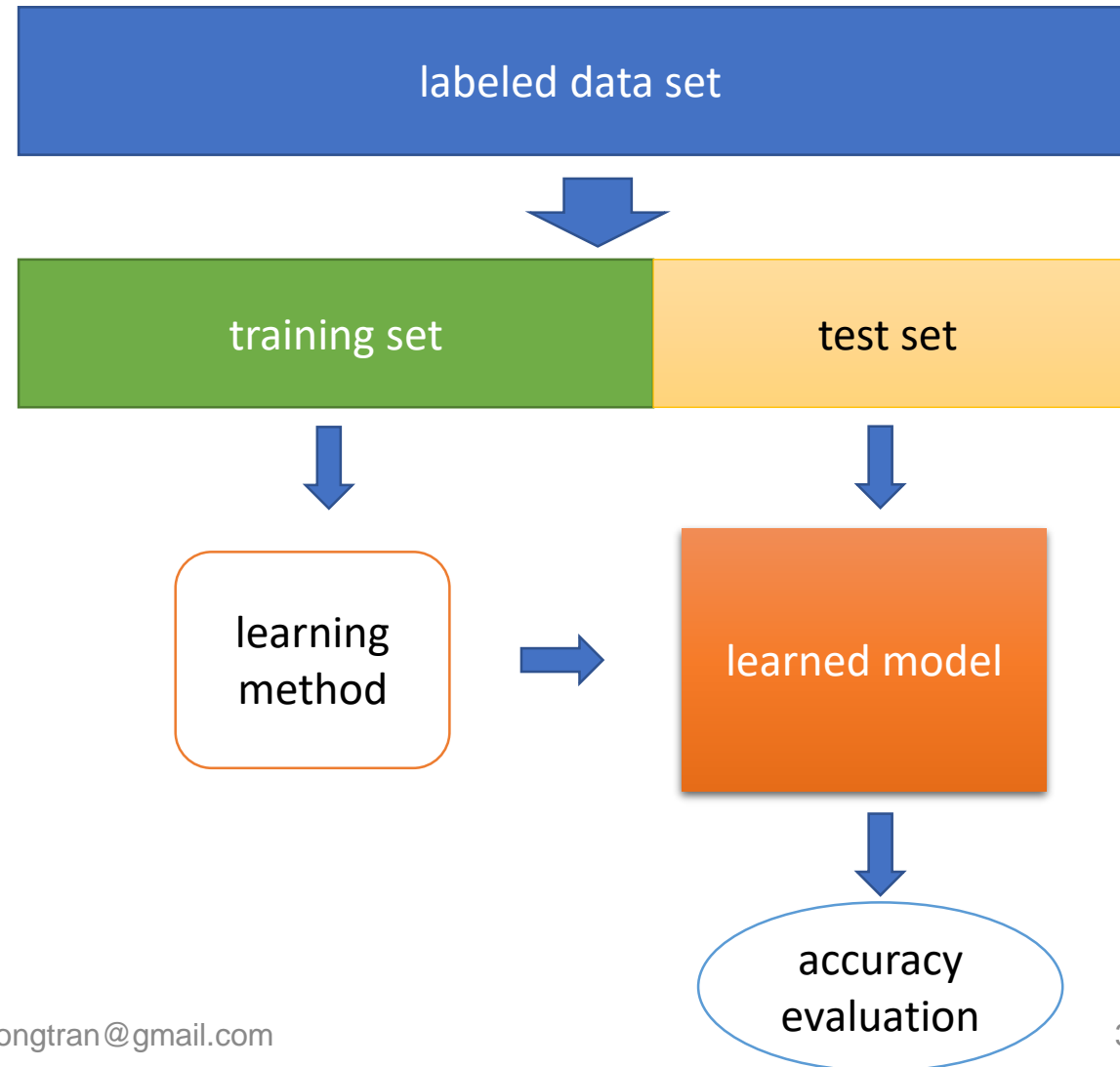
- Huấn luyện:** cho một tập dữ liệu huấn luyện đã được gán nhãn $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, ước lượng hàm dự đoán f bằng cách cực tiểu hóa lỗi dự đoán
- Kiểm tra:** áp dụng f với dữ liệu mới (chưa được huấn luyện) \mathbf{x} để dự đoán output $y = f(\mathbf{x})$

Đánh giá (Evaluation)

- Đánh giá mô hình
 - Sử dụng dữ liệu kiểm tra (validation data, test data)
 - Tách biệt với tập huấn luyện
 - Đảm bảo tính khách quan
 - Ước lượng trước hiệu năng hệ thống khi vận hành thật

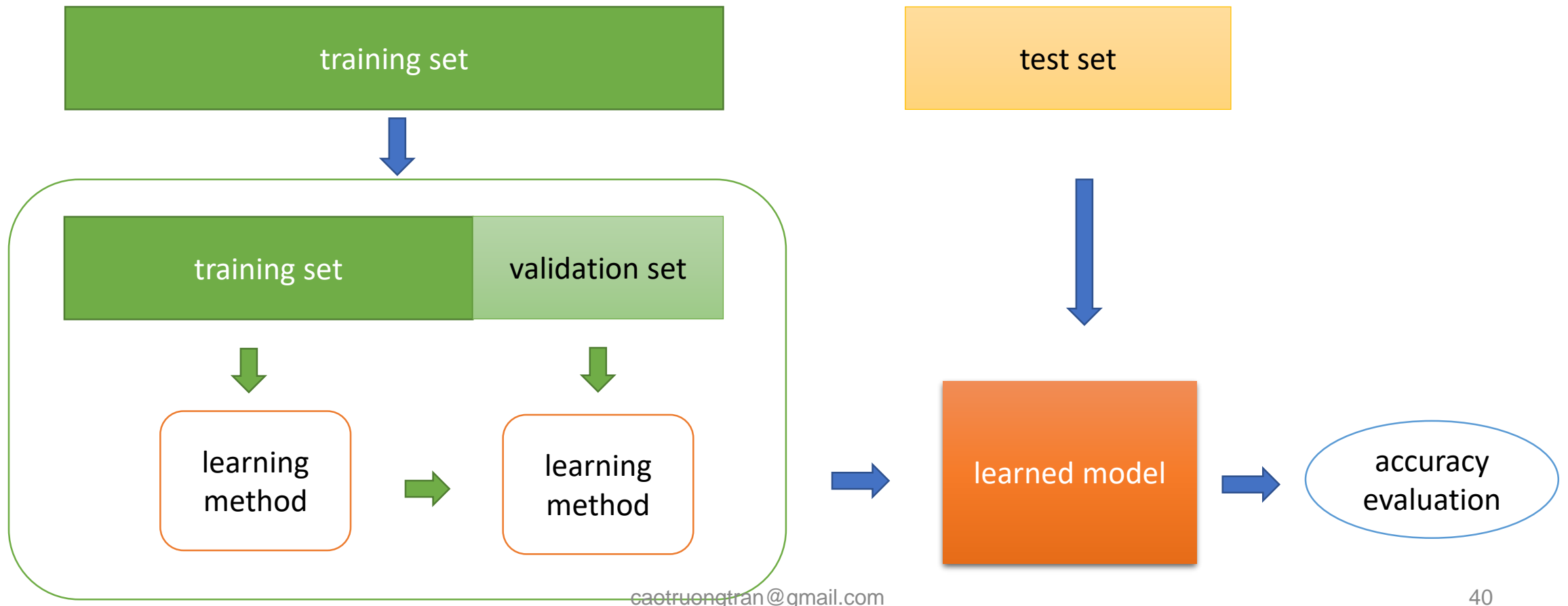
Tập dữ liệu đánh giá mô hình

- Khi huấn luyện mô hình, chưa có dữ liệu kiểm tra
- Sử dụng một phần dữ liệu đã có để đánh giá mô hình trước khi sử dụng

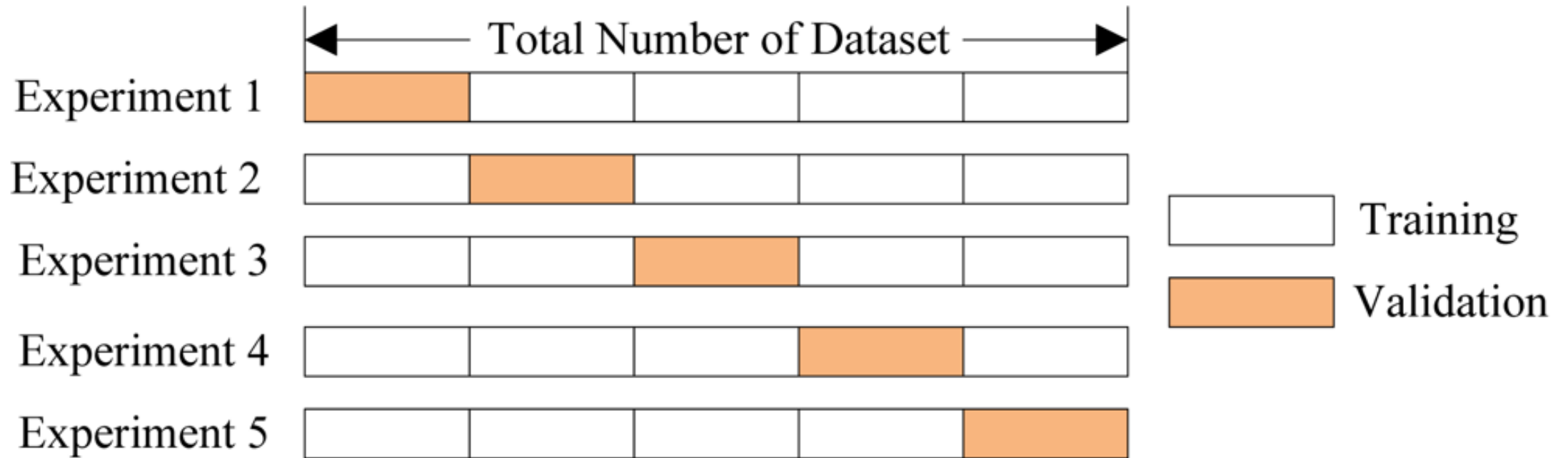


Kiểm định (Validation)

- Kiểm định trong quá trình huấn luyện



Kiểm định chéo (Cross Validation)



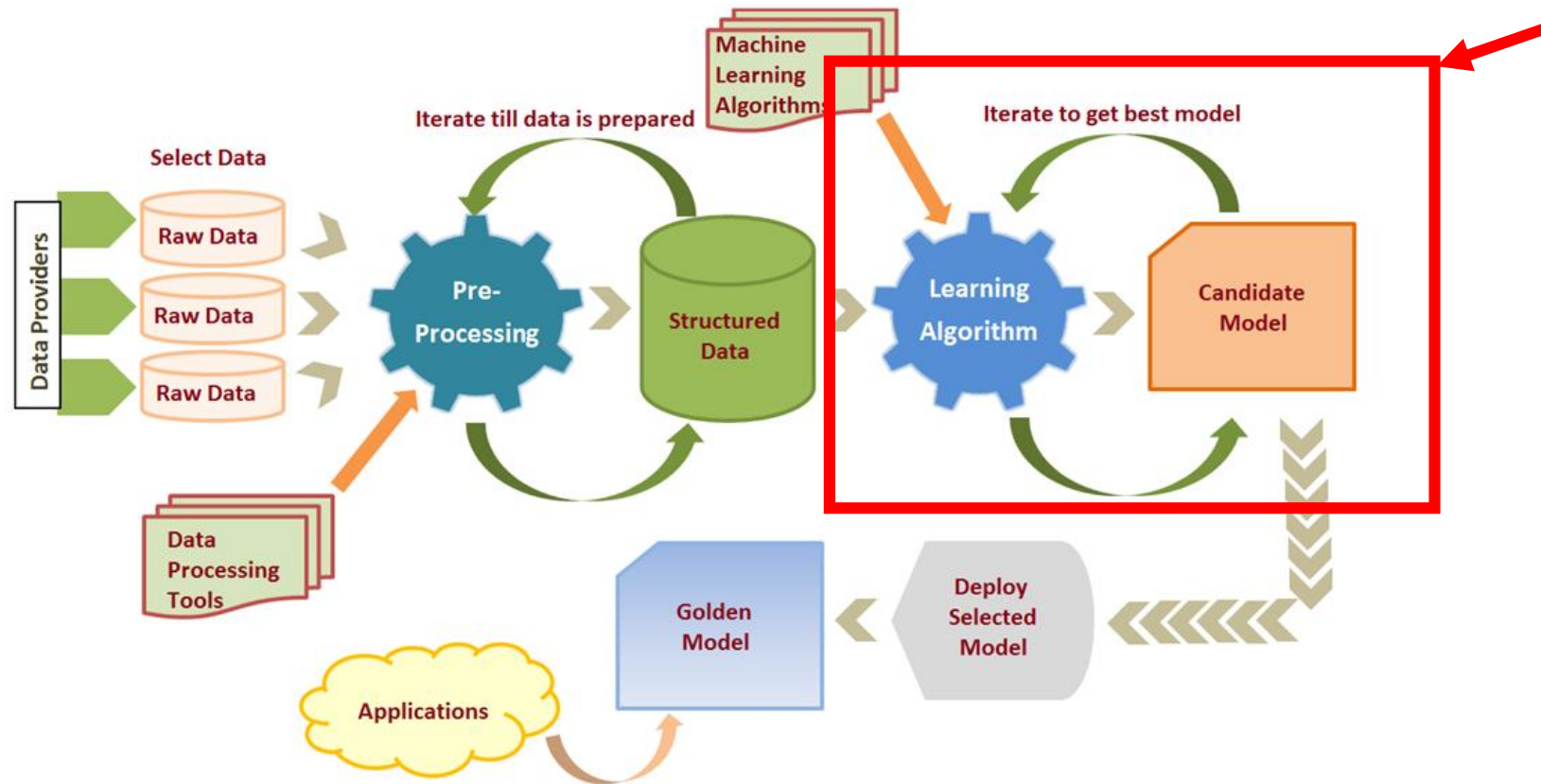
Đánh giá (Evaluation)

		Predicted class		
		yes	no	Total
Actual class	yes	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	P + N

Measure	Formula
accuracy, recognition rate	$\frac{TP + TN}{P + N}$
error rate, misclassification rate	$\frac{FP + FN}{P + N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
precision	$\frac{TP}{TP + FP}$
F , F_1 , F -score, harmonic mean of precision and recall	$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
F_β , where β is a non-negative real number	$\frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

Tổng hợp

Các thuật toán cơ bản



NỀN TẢNG và CÔNG CỤ

Giới thiệu các nền tảng cho học máy và phân tích dữ liệu
Cài đặt các bộ công cụ sử dụng trong khóa học

Python



- Ngôn ngữ thông dịch (Làm việc tương tự Java nhưng độ linh hoạt cao hơn), dễ tiếp cận
- **Python** là ngôn ngữ lập trình rất phổ biến trong cả môi trường học thuật và công nghiệp.
- Được thiết kế gần với ngôn ngữ tự nhiên (tiếng Anh), sử dụng nhóm từ vựng toán học
- Đa nền tảng (Windows, Mac, Linux, etc)
- Ngôn ngữ hỗ trợ lập trình thủ tục, hướng đối tượng, lập trình hàm
- Tồn tại hai phiên bản tương đối độc lập: Python 2 vs Python 3
- Quản lý và cài đặt Packages sử dụng PIP

Python



- Khuyến nghị sử dụng Python 3
- Cài đặt từ gói phân phối chính thức tại python.org/downloads/

Release version	Release date	Click for more	
Python 3.5.6	2018-08-02	Download	Release Notes
Python 3.4.9	2018-08-02	Download	Release Notes
Python 3.7.0	2018-06-27	Download	Release Notes
Python 3.6.6	2018-06-27	Download	Release Notes
Python 2.7.15	2018-05-01	Download	Release Notes
Python 3.6.5	2018-03-28	Download	Release Notes
Python 3.4.8	2018-02-05	Download	Release Notes
Python 3.5.5	2018-02-05	Download	Release Notes

[View older releases](#)

Python



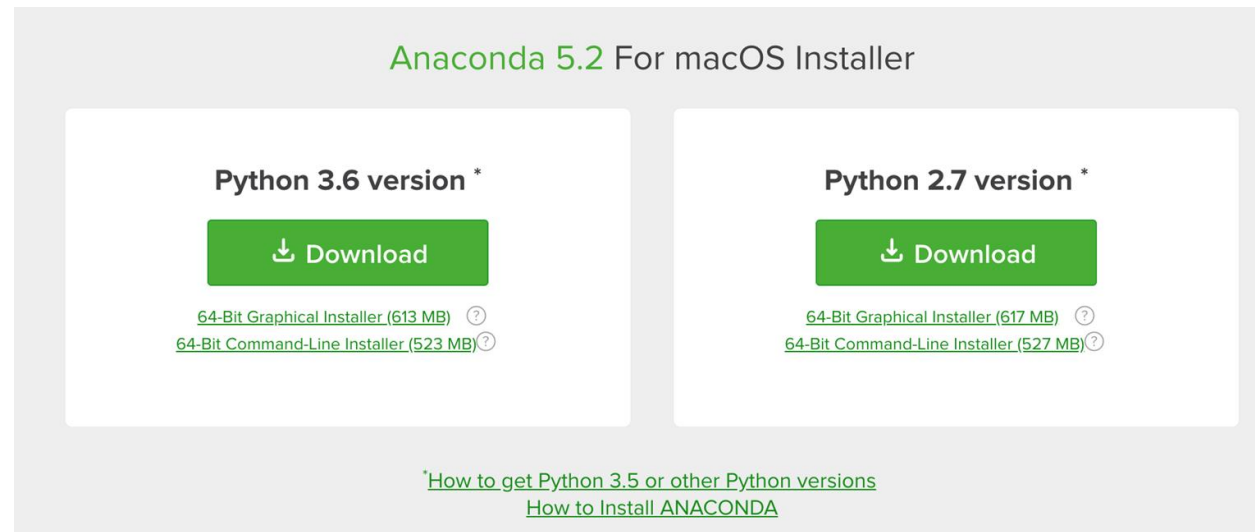
- Chọn “Add Python 3.7 to PATH”



Python



- Cài đặt Python với Anaconda
 - Download Anaconda tại: <https://www.anaconda.com/download/>



NumPy

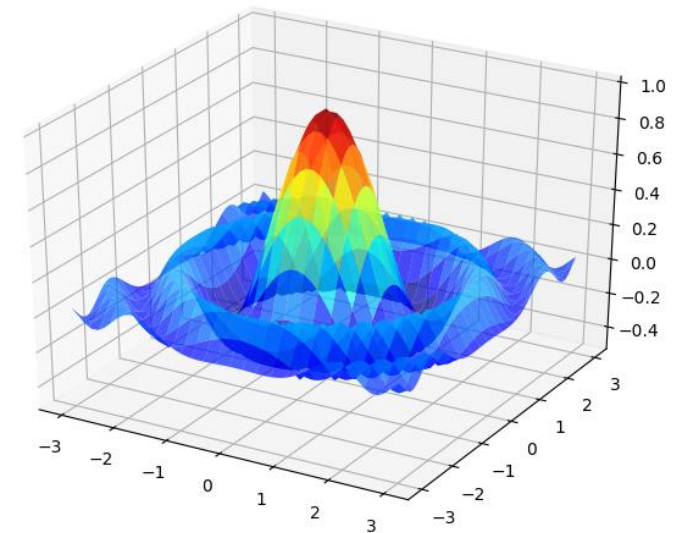
- Là gói thư viện cơ bản dành cho khoa học máy tính trên Python
- Được thiết kế cho việc tính toán hiệu quả trên dữ liệu mảng, ma trận kích thước lớn, nhiều chiều
- NumPy bao gồm:
 - N-dimensional array object
 - Các hàm tính toán phức tạp
 - Linear Algebra, Fourier Transform
 - Bộ sinh số ngẫu nhiên
- Cài đặt thông qua PIP:
> `pip install numpy`

Matplotlib



- Là gói thư viện vẽ đồ thị 2D cho Data visualization
- Tạo plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc.. với chỉ vài dòng code
- Matplotlib được tích hợp nhiều plugins, đặc biệt là mplot3d cho đồ thị 3D
- Cài đặt thông qua PIP:

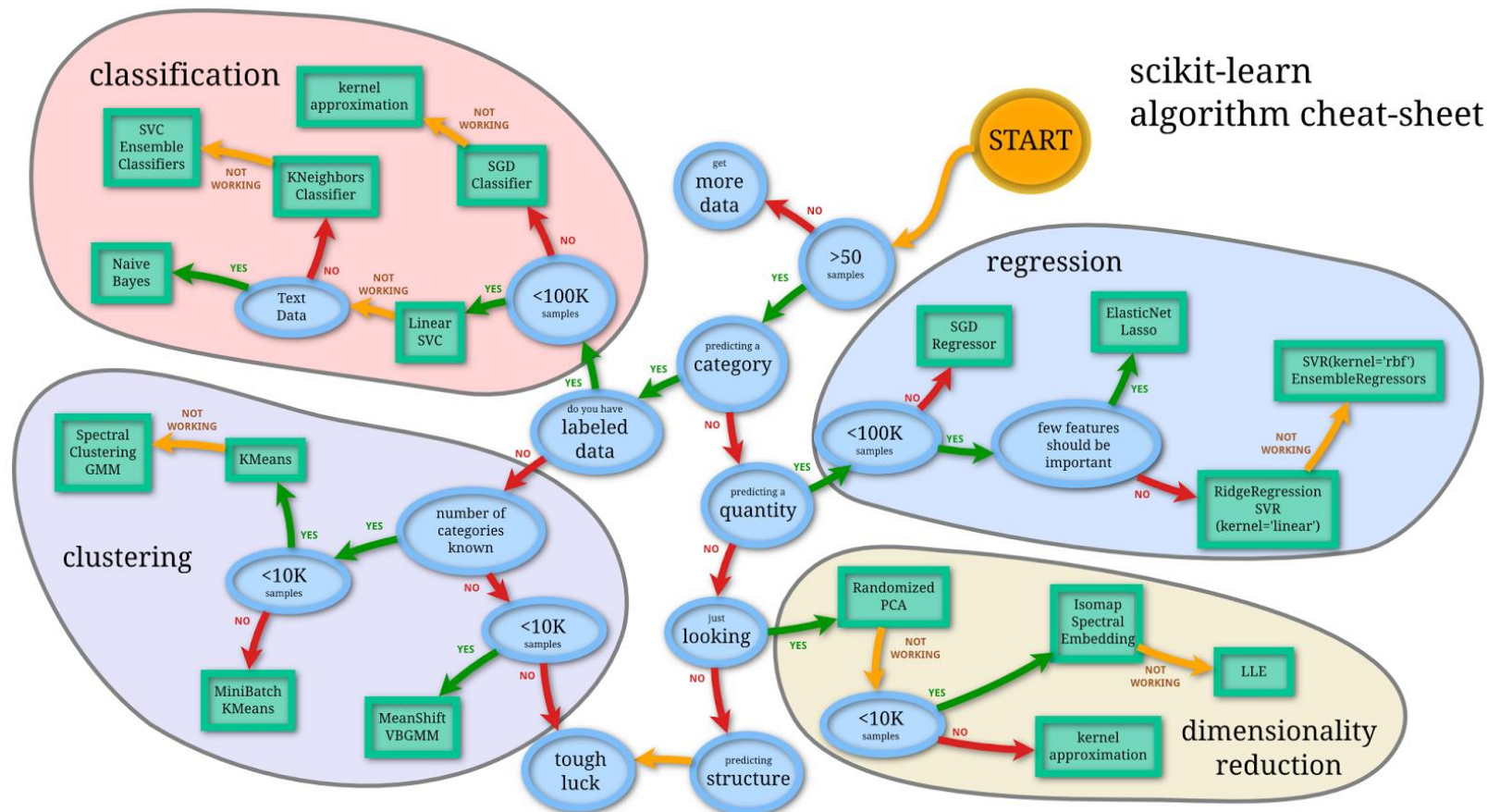
```
> pip install matplotlib
```



Scikit-Learn



- Là thư viện Machine Learning mã nguồn mở nổi tiếng



Scikit-Learn



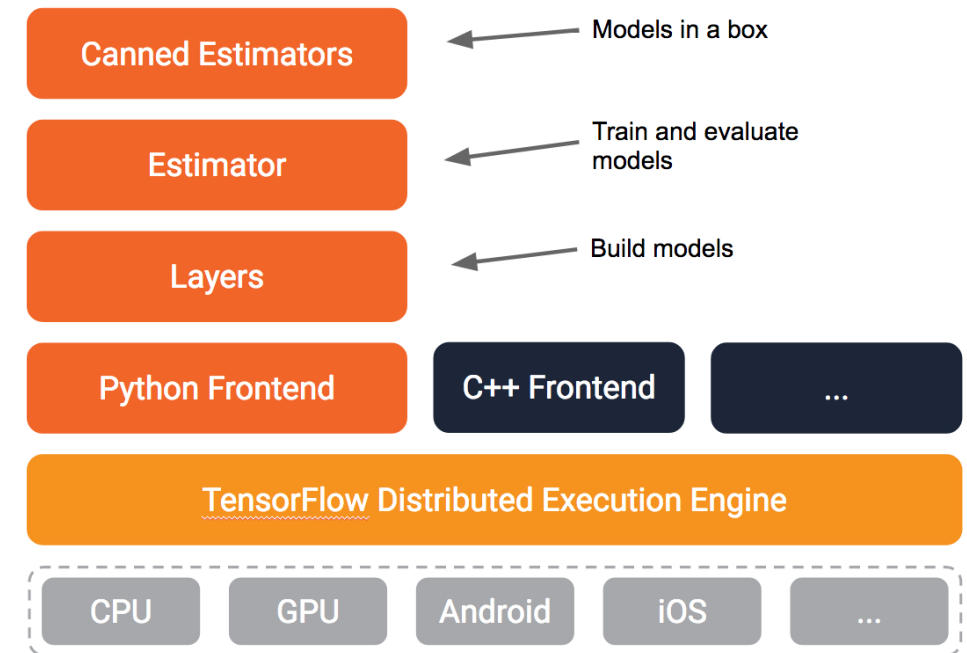
- Mã nguồn mở, được cấp phép cho môi trường doanh nghiệp
- Built trên Numpy, Scipy, và Matplotlib
- Là công cụ hiệu quả cho khai phá dữ liệu và phân tích dữ liệu
- Cài đặt đầy đủ các thuật toán học máy
- Cài đặt thông qua PIP:

```
> pip install scikit-learn
```

Tensorflow



- Là thư viện mã nguồn mở cho Deep Learning, hỗ trợ bởi Google
- Được cài sẵn rất nhiều kỹ thuật tính toán tối ưu, các thuật toán học máy
- Có rất nhiều pre-trained models cho Tensorflow
- Framework tính toán hết sức hiệu quả
- Có thể thực hiện tính toán trên GPU để tăng tốc độ
- Hiện tại đã hỗ trợ gần như đầy đủ các nền tảng, kể cả Mobiles, Raspberry, Browsers (với tensorflow.js)



Tensorflow

- Với máy tính không có GPU của NVIDIA, Cài đặt bản TensorFlow được built cho CPU thông qua PIP:
 - > `pip install tensorflow`
- Cài đặt bản built cho GPU:
 - > `pip install tensorflow-gpu`
- Kiểm tra cài đặt
 - Với tensorflow CPU:
 - > `python`
 - >> `import tensorflow as tf`
 - >> `hello = tf.constant("Hello, Tensorflow")`
 - >> `sess = tf.Session()`
 - >> `print(sess.run(hello))`

Nếu cài đặt thành công, kết quả được in ra:

```
Hello, Tensorflow
```

Tensorflow



- Kiểm tra cài đặt
 - Với tensorflow GPU:
 - > python
 - >> import tensorflow as tf
 - >> sess = tf.Session(config=
tf.ConfigProto(log_device_placement=True)
)

```
C:\Windows\system32\cmd.exe - python
from _conv import register_converters as _register_converters
>>> sess = tf.Session(config=tf.ConfigProto(log_device_placement=True))
2018-08-17 00:09:51.596312: I T:\src\github\tensorflow\tensorflow\core\platform\cpu_feature_guard.cc:141] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2
2018-08-17 00:09:51.848890: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:1392] Found device 0 with properties:
name: GeForce GTX 750 Ti major: 5 minor: 0 memoryClockRate(GHz): 1.1105
pciBusID: 0000:01:00:0
totalMemory: 2.08GiB freeMemory: 1.65GiB
2018-08-17 00:09:51.852873: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:1471] Adding visible gpu devices: 0
2018-08-17 00:09:52.575981: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:952] Device interconnect StreamExecutor with strength 1 edge matrix:
2018-08-17 00:09:52.577968: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:958] 0
2018-08-17 00:09:52.579220: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:971] 0:  N
2018-08-17 00:09:52.588585: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:1884] Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:0 with 1411 MB memory) -> physical GPU (device: 0, name: GeForce GTX 750 Ti, pci bus id: 0000:01:00:0, compute capability: 5.0)
Device mapping:
/job:localhost/replica:0/task:0/device:GPU:0 -> device: 0, name: GeForce GTX 750 Ti, pci bus id: 0000:01:00:0, compute capability: 5.0
2018-08-17 00:09:52.757488: I T:\src\github\tensorflow\tensorflow\core\common_runtime\direct_session.cc:288] Device mapping:
/job:localhost/replica:0/task:0/device:GPU:0 -> device: 0, name: GeForce GTX 750 Ti, pci bus id: 0000:01:00:0, compute capability: 5.0
>>>
```


Các toolkit khác





Q&A

Thank you!