6 chương :

chương 1: nhập môn học máy

Chương 2: tổng quan về bộ công cụ hỗ trợ phân tích dữ liệu Anaconda

Chương 3: Trực quan hóa dữ liệu

chương 4; tiền xử lý dữ liệu

chương 5: hồi quy tuyến tính

chương 6: hồi quy logistic

chương 7: Mô hình phân lớp k lân cận và Naive Bayes

chương 8: Phương pháp đánh giá mô hình phân loại

tài liệu điện tử:

scikit-learn

Chương 1:

1.1 Giới thiệu về học máy

Machine learning is the field of study that gives computers the ability to learn without being explicily programmed ( Học máy là lĩnh vực nghiên cứu mang lại cho máy tính khả năng học hỏi mà không cần lập trình rõ ràng.)

-Arthur samuel,1959

A computer program is said to learn from experience E with respect to some task T and some performance...

\* Nhiệm vụ T (Task T)

+ Các nhiệm vụ trong ML thường được mô tả thông qua việc một hệ thống ML xử lí một điểm dữ liệu( data point) như thế nào.

+ Mỗi điểm dữ liệu bao gồm nhiều đặc trưng(feature) khác nhau, mỗi feature thường được biểu diễn dưới dạng một con số.

\* Phép đánh giá P (Performance P)

+ Để kiểm tra hiệu năng của một thuật toán ML, chúng ta cần phải thiết kế các phép đánh giá có thể đo đạc được kết quả.

+ Dữ liệu sẽ được hia thành hai phần riêng biệt: tập huấn luyện ( training set) và tập kiểm thử (test set)

Tập huấn luyện sẽ được dùng để tìm các tham số mô hình

Tập kiểm thử được dùng để đánh giá năng lực của mô hình tìm được.

\* Kinh nghiệm E(Experience E)

+ Việc huấn luyện các mô hình ML có thể coi là việc cho chúng trải nghiệm trên các tập dữ liệu (data set) - chính là training set.

+ Các tập dữ liệu khác nhau sẽ cho các mô hình các trải nghiệm khác nhau.

+ Chất lượng của các tập dữ liệu này cũng ảnh hưởng tới hiệu năng của mô hình.

ML is an approach to learn complex patterns from existing data and use these patterns to make predictions on unseen data(ML là một cách tiếp cận để tìm hiểu các mẫu phức tạp từ dữ liệu hiện có và sử dụng các mẫu này để đưa ra dự đoán về dữ liệu chưa được nhìn thấy)

(1) Học(learn): Hệ thống có khả năng tự học hỏi, nó phải có thứ gì đó để học hỏi, hệ thống ML hc từ dữ liệu.

ví dụ: Nếu muốn xây dựng hệ thống ML để học các dự đoán giá thuê phòng trên Airbnb, bạn cần cung cấp tập douwx liệu trong đó mỗi đầu vào là một danh sách có các đặc điểm liên quan( diện tích , số phòng, khu vực lân cận, tiện nghi, xếp hạng của danh sách đó,...) và đầu ra liên quan là giá cho thuê của danh sách đó.

Sau khi học được, hệ thống ML này sẽ có thể dự đoán giá của một danh sách mới dựa trên các đặc điểm của nó.

(2) Mẫu phức tạp( Complex Patterns)

Giải pháp Ml chỉ hữu ích khi có các mẫu để học.

vd: Không sử dụng ML để dự đoán kết quả của việc đổ xúc xắc vì không có khuôn mẫu nào về cách tạo ra những kết quả này. Tuy nhiên, có thể sử dụng ML để dự đoán bệnh,...

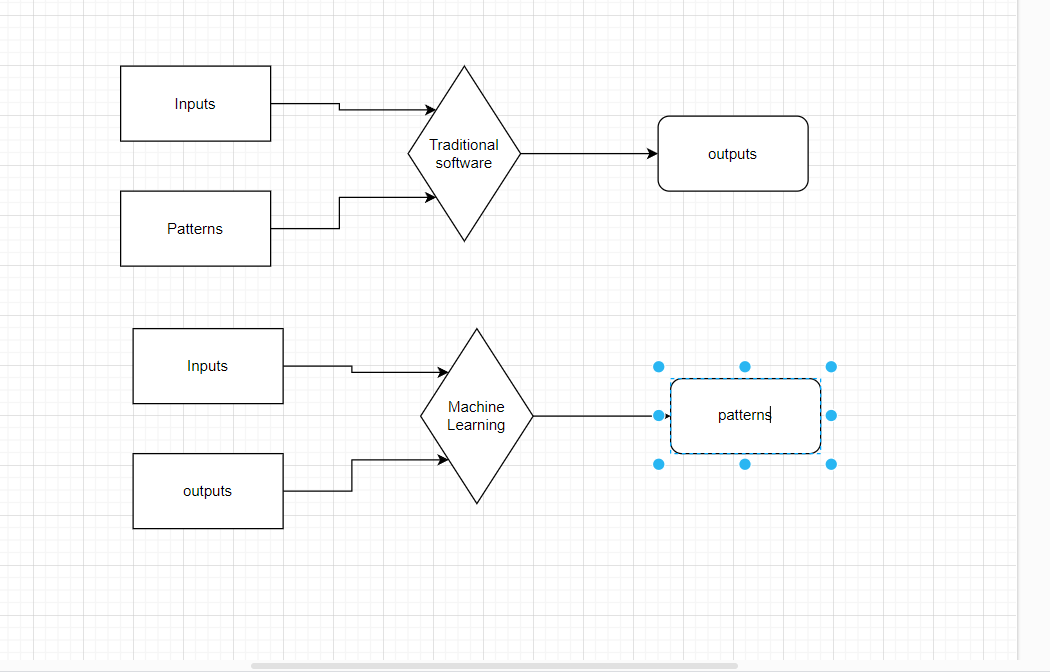
Việc một mẫu có tồn tại thì tập dữ liệu hoặc thuật toán ML có sẵn có thể không đủ tốt

vd: có thể có một xu hướng trong cách các dòng tweet của Elon Musk ảnh hưởng đến giá tiền điện tử.

Hãy xem xét một trang web như Airbnb có rất nhiều danh sách nhà; mỗi danh sách đi kèm với một mã zip.

Nếu muốn sắp xếp danh sách theo các tiểu bang mà chúng nằm ở đó, chúng ta sẽ ko cần hệ thống ML.

Vì mẫu này rất đơn giản -mỗi mã zop tương ứng với một trạng thái đã biết- bạn chỉ cần sử dụng bảng tra cứu.



(3) Dữ liệu có sẵn hoặc có thể thu nhập (Existing data)

(4) Dự đoán giá trị tương lai của 1 vấn đề dựa trên dữ liệu lịch dữ đã được học về vấn đề đó.

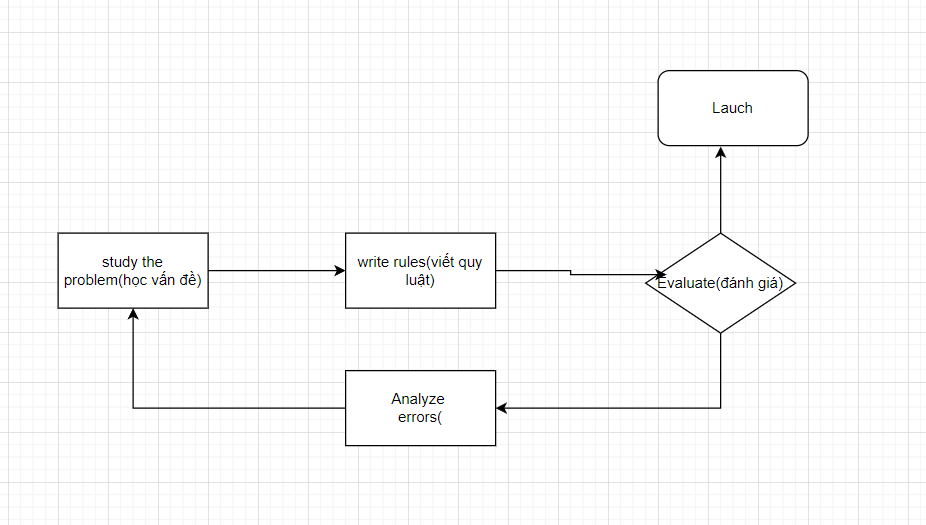
(5) Dữ liệ mới và có thể chưa từng gặp( unseen data)

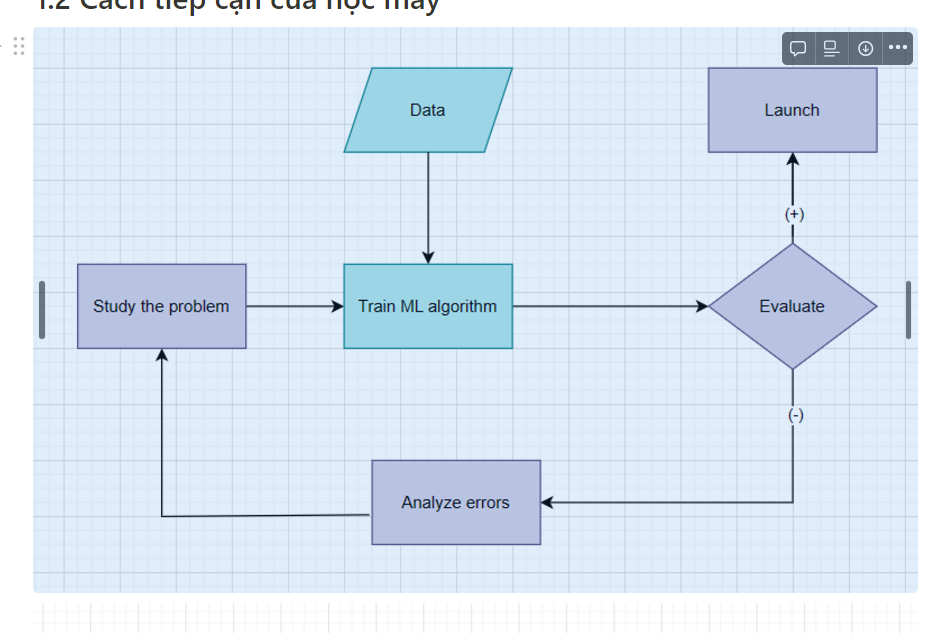
+ Các mẫu mà mô hình được học từ dữ liệu hiện có chỉ hữu ích nếu unseen data cũng có cá mẫu (patterns) gần tương tự như model đã được học.

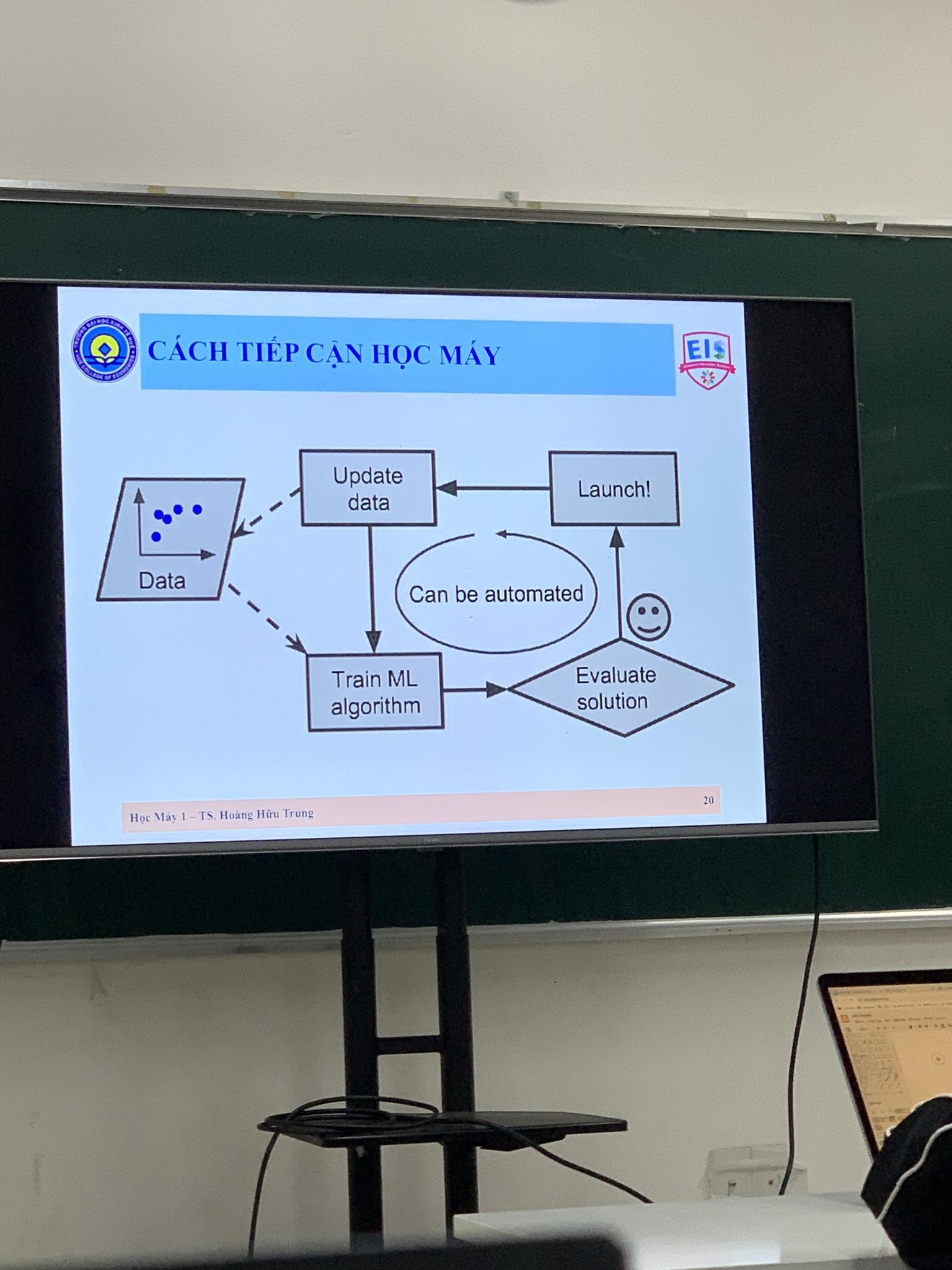
+ Một mô hình tốt là mô hình dự đoán đúng hoặc dự báo được xu hướng trong tương lai của vấn đề được học ( với dữ liệu mới)

1.2 Cách tiếp cận của học máy

1.Cách tiêp cận truyền thống



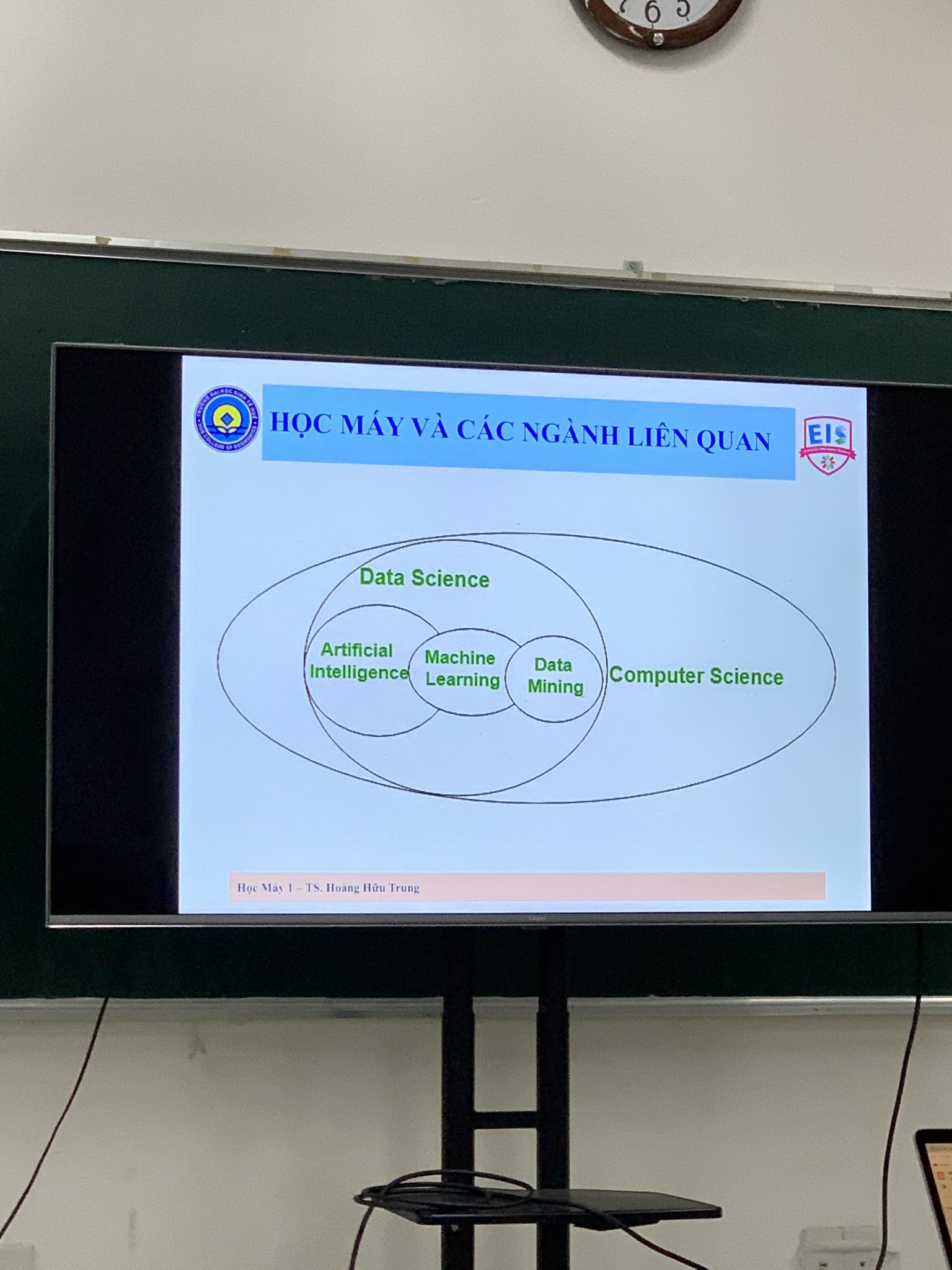


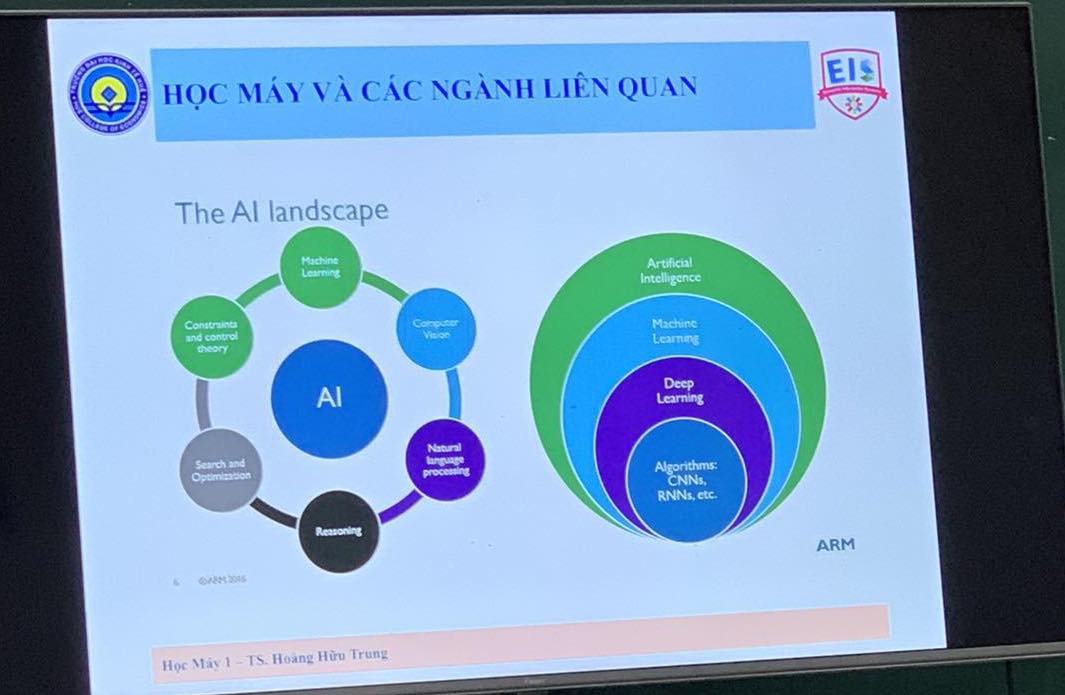


1.3 Ứng dụng của học máy

* Nhận dạng hình ảnh và video( computer vision): sử dụng AI để nhận dạng và phân tích nội dung từ hình ảnh và video, từ việc nhận biết khuôn mặt cho đến phân loại đối tượn
* Xử lí ngôn ngưr tự nhiên (Natural language Processing): sử dụng AI đẻ hiểu và tạo ra văn bản tự nhiên, từ chatbot trả lời câu hỏi đến dịch thuật tự động.
* Tự động hóa công việc: Áp dụng AI để thực hiện tự động các tác vụ lặp đi lặp lại hoặc phức tạp, giúp tiết kiệm thời gian và tăng hiệu suất.
* Dự đoán và phân tích dữ liệu: Sử dụng AI để dự đoán xu hướng, biểu đồ và phân tích dữ liệu lớm để rút ra thông tin hữu ích.
* Y tế và chăm sóc khỏe: Sử dụng AI trong chẩn đoán bệnh, phát hiện tế bào ác tính và tư vấn về điều trị dựa trên dữ liệu y tế
* Thiết kế và nghệ thuật: Sáng tạo nghệ thuật và thiết kế dựa trên AI, từ việc tạo tranh và âm nhạc đến thiết kế sản phẩm.
* Dịch vụ tài chính: sử dụng AI để dự đoán xu hướng thị trường tài chính, quản lí rủi ro và phát hiện gian lận
* Hỗ trợ ra quyết định: cung cấp thông tin và dự đoán để hổ trợ quyết định trong lĩnh vực kinh doanh, chính trị và xã hội
* Học tập trực tuyến: sử dụng AI để cung cấp nội dung học tập tùy chỉnh và đề xuất các khóa học dựa trên sở thích và tiến độ học viên.

1.4 Học máy và các ngành liên quan

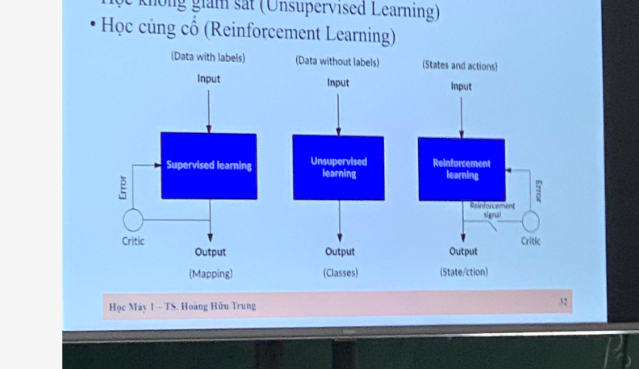




1.5 Phân loại học máy

Học máy được phân thành các loại chính:

* Học có giám sát( supervised learning)
* Học không giám sát( Unsupervised learning)
* Học củng cố ( Reinforcenment learning)



1.5.1 Học có giám sát( supervised learning)

- Học có giám sát ( supervised learning) là một nhóm thuật toán sử dụng dữ liệu được gán nhãn nhằm mô hình hóa mối quan hệ giữa biến đầu vào x và biến đầu ra y.

- Hai nhóm bài toán cơ bản trong học có giám sát là classsification(phân loại) và regression( hồi quy), trong đó biến đầu ra của bài toán phân loại có các giá trị rời rạc trong khi biến đầu ra của bài toán hồi quy có các giá trị liên tục.

- Với supervised learning, bên cạnh xây dựng các mô hình mạnh, việc thu nhập và gán dữ liệu tốt và gán nhãn dữ liệu tốt và hợp lí cũng đóng vai trò then chốt để giải quyết các bài toán trong thực tế.

1.5.2 Học không giám sát ( unsupervised learning)

- Ngược lại supervised learning, unsupervised learning(học không giám sát) là một nhóm thuật toán sử dụng dữ liệu không có nhãn.

- Các thuật toán the cách tiếp cận này hướng đến việc mô hình hóa được được cấu trúc hay thông tin ẩn trong dữ liệu

- hay nói cách khác, sử dụng các phương pháo này thiên về việc mô tả tính chất hay đặc tính của dữ liệu.

- Thông thường, các thuật toán này dựa trên những thông tin sau:

+ Mối quan hệ tương tự ( similarity) giữa các ví dụ( được gọi là instance) trong dữ liệu như trong các thuật toán clustering (phân cụm)

+ xác suất đồng xuất hiện của các đối tượng như trong Association mining.

+ Các phép biến đổi ma trận để trích xuất các đặc trung như PCA, SVD.

- Ứng dụng học không giám sát:

+ Chia phân khúc sản phẩm

+ Chia phân khúc khách hàng

+ Phát hiện sự tương đồng

+ Hệ thống khuyến nghị

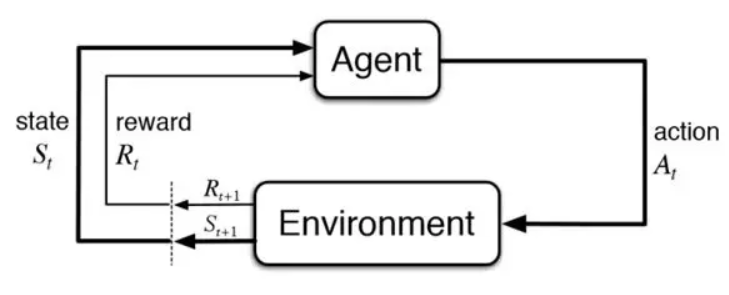
+ Gắn nhãn các tập dữ liệu không được gắn nhãn

1.5.3 Học củng cố ( Reinforcement learning)

- Học củng cố ( Reinforcenment learning) là một lĩnh vực quan trọng trong trí tuệ nhân tạo, nơi máy tính học hành và ra quyết định thông qua việc tương tác với môi trường.

- Trong học tăng cường, một hệ thống hoặc “agent” (người chơi) tương tác với một môi trường, thực hiện các hành động và nhân được phản hồi trong hình thức phần thưởng hoặc hình phạt.

- Mục tiêu của học tăng cường là học cách lựa chọn hàng động để tối ưu hóa một mục tiêu cụ thể theo thời gian.



1.6 Các nguồn dữ liệu miễn phí cho học máy

