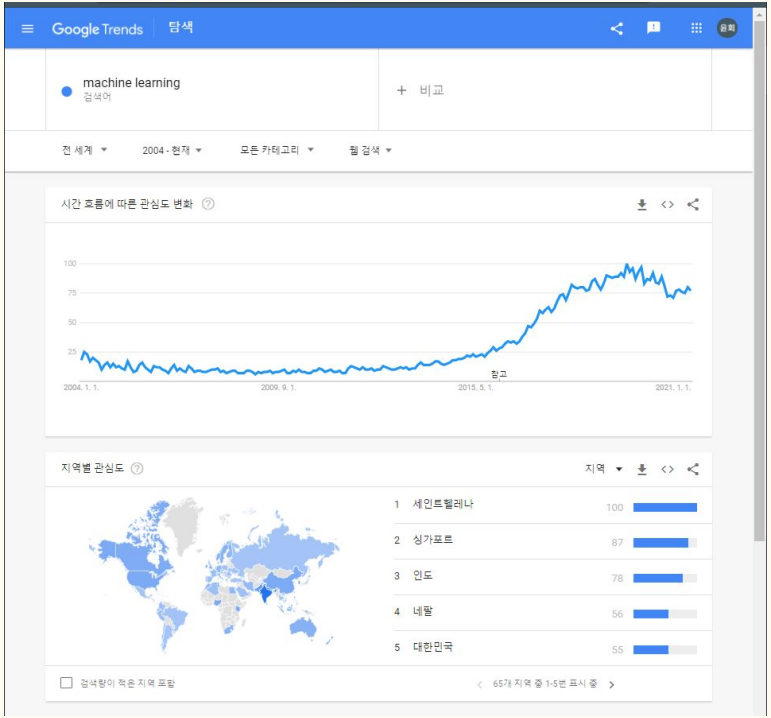
01 Introduction to ML

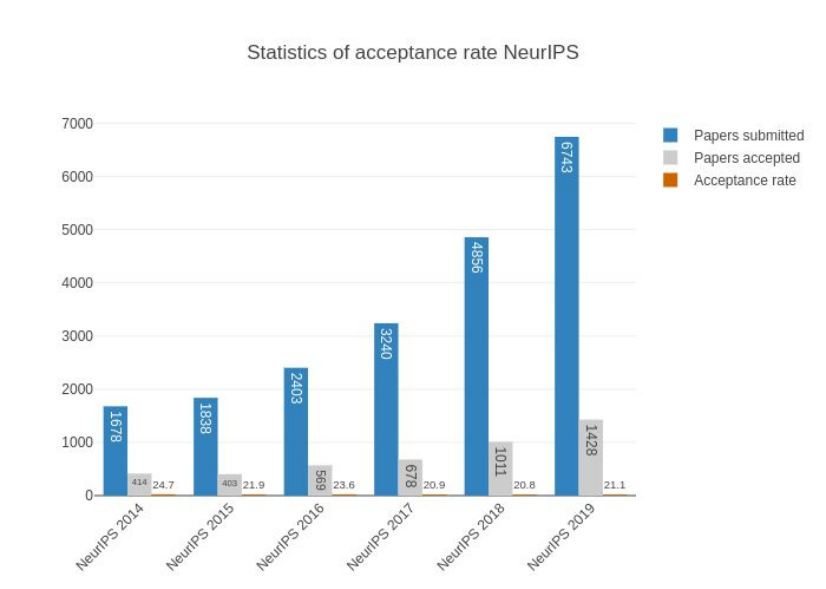
Google Trends

- Google Trends analyzes the popularity of search queries in Google Search
  - Interest over time
  - Interest by region



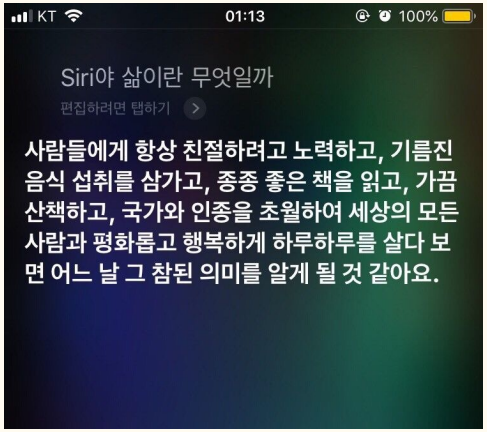
ML Conferences

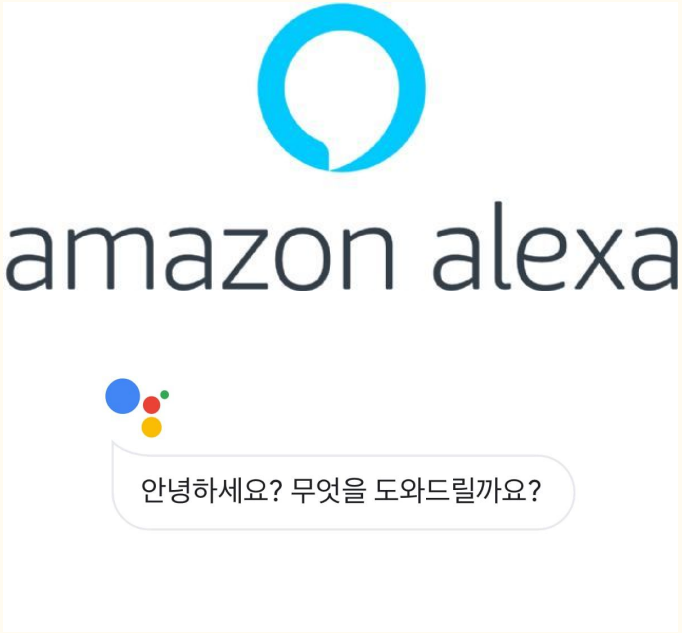
- NIPS (Neural Information Processing Systems



Machine Learning Applications

- Speech recognition
  - 음성인식 기술은 사람의 언어의 의미구조를 이해하기 위해 머신러닝을 사용한다





Machine Learning Applications

- Fraud detection
  - 신용카드 거래사기
  - Fraud detection systems try to recognize fraudulent transactions so that customers are not charged for items that they did not purchase
- Kaggle
  - https://www.kaggle.com/mlg-ulb/creditcardfraud

Machine Learning Applications

- Recommendation
  - Recommendation engines use machine learning to recommend things that you may like
  - 추천 엔진은 기계 학습을 사용하여 사용자가 좋아할 수 있는 항목을 추천합니다.
    - Netflix, Amazon, Spotify

Machine Learning Applications

- 얼굴 인식 : 이미지에서 얼굴을 식별
- 이메일 필터링 : 이메일을 스팸 또는 비스팸으로 분류
- 의학적 진단 : 환자를 질병의 유병자 또는 비고유자로 진단
- 날씨 예측: 예를 들어 내일 비가 올지 여부를 예측

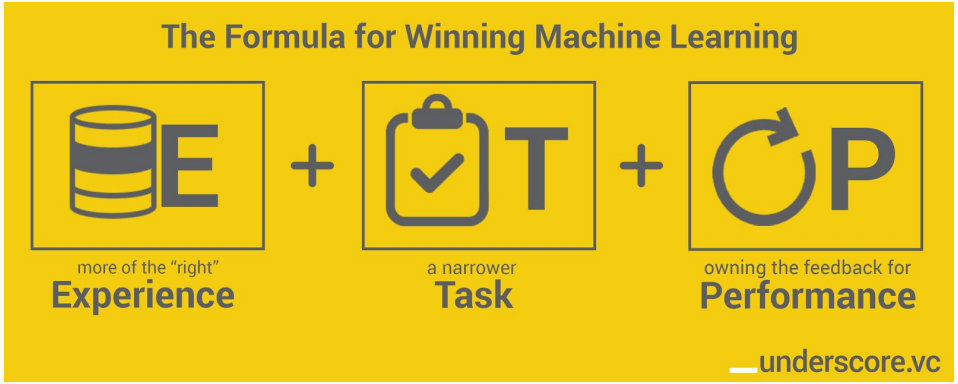
What is Machine Learning?

- The field of study that gives computers the ability to **learn without being explicitly programmed (**Arthur Samuel, 1959)

- 기존 : 사람이 데이터를 통해 학습을 한 뒤 이것을 다시 컴퓨터에 instruct
- ML : 컴퓨터가 데이터를 통해 학습할 수 있는 능력을 부여.

What is Machine Learning?

- A computer program is said to learn from **experience E** with respect to some **task T** and some **performance measure P**, if its performance on T, as measured by P, improves with experience E (Tom Mitchell, 1997)
- Example
  - T : flag spam for new emails
  - E : examples of spam & regular emails
  - P : accuracy(the ratio of correctly classified emails)



E * T = P

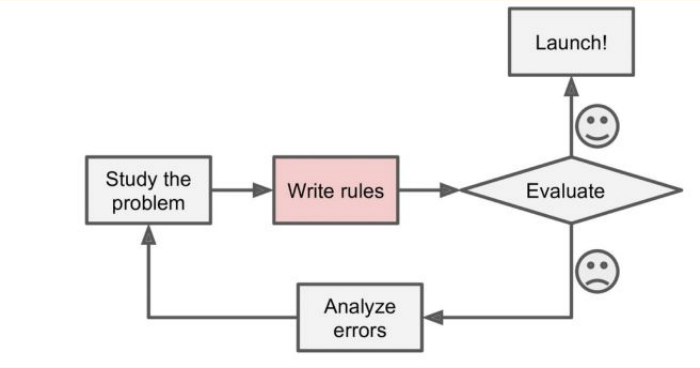
Experience	*	Task	=	Performance
Input Data: • Housing prices • Customer transactions • Clickstream data • Images		**Task:** • Predict prices • Segment customers • Optimize user flows • Categorize images		**Performance:** • Accurate prices • Coherent groupings • KPI lifts • Correctly sorted images

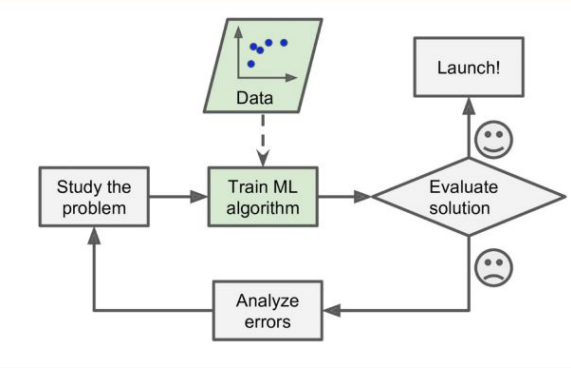
Why use Machine Learning?

- 기존 스팸 필터의 작동 방식
    1. 스팸메일이 어떻게 생겼는지를 본다.
        a. 보내는 사람, 메일 본문의 형태에서 어떠한 패턴을 발견한다.
    2. 각 패턴마다 detection algorithm 을 만든다.
        a. 만약 그러한 패턴이 감지되면 스팸으로 분류한다.
    3. 시스템이 충분히 좋아질 때 까지 1,2를 반복한다.
- 단점
    - 패턴을 직접 발견해야 하는데 만약 패턴이 광범위하다면 다 찾아내기가 힘들다.
    - 룰이 많아질수록 프로그램의 유지보수 측면에서 관리가 힘들어진다.

Why use Machine Learning?

- A spam filter(the machine learning-based approach)
  - 자동으로 스팸에서 빈번하게 쓰이는 단어의 패턴을 감지한다.
- 강점
  - 프로그램이 훨씬 짧아지며, 관리가 쉽고, 더 정확하다
  - 사람의 개입 없이 새로운 패턴을 발견하고, 변화를 적용한다.





Why use Machine Learning?

- Machine learning is ideal for problems that
  - 기존 방법으로는 너무 복잡하다.
  - 알려진 알고리즘이 없다.

- E.g., Speech recognition
  - 시끄러운 환경에서 수십 개의 언어로 수백만 명의 매우 다른 사람들이 말하는 수천 개의 단어를 인식합니다.
  - 가장 좋은 해결책은 각 단어에 대한 많은 예제 녹음이 주어지면 스스로 학습하는 알고리즘을 작성하는 것입니다.

Machine Learning vs. others

Machine Learning vs. Data Science

- Data Science
  - Data Science is the study of the generalizable extraction of knowledge from data.
    - 데이터에서 지식을 일반화 할 수 있는 추출에 대한 연구
  - Data Science is an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information.
    - 데이터 과학은 방대한 정보 수집의 수집, 준비, 분석, 시각화, 관리 및 보존과 관련된 새로운 작업 영역입니다.
  - Data Science is an academic program offered by many universities.
    - 학제 프로그램 이름

• Data Science programs teach **Data mining, Machine Learning, Natural Language Processing, Information Retrieval,** Etc.

Machine Learning vs. Data Mining

- Data Mining
  - Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
  - 데이터 마이닝은 기계 학습, 통계 및 데이터베이스 시스템의 교차점에서 방법을 포함하는 대규모 데이터 세트에서 패턴을 발견하는 프로세스입니다.

- E.g.,What are the characteristics of people using iPhone?
  - Age, education, income, occupation, etc. (statistics)
- E.g., Can we group these people based on the characteristics?
  - Clustering (machine learning)

Machine Learning vs. Artificial Intelligence

Artificial Intelligence

- Artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals.
- Artificial intelligence is used when a machine mimics "cognitive" functions, such as "learning" and "problem solving".
- 기계 지능이라고도 불리는 인공 지능(AI)은 인간과 다른 동물이 보여주는 자연 지능과 달리 기계가 보여주는 지능입니다.
- 인공 지능은 기계가 "학습" 및 "문제 해결"과 같은 "인지" 기능을 모방할 때 사용됩니다.

- **Machine Learning is an early step towards Artificial Intelligence.**
    - 머신러닝은 AI 를 위한 앞단계 라고 보시면 됩니다.
- Artificial Intelligence also includes
    - Knowledge reasoning / Robotics / Etc.,
- People are trying to achieve artificial intelligence in some fields
    - Autonomous driving (self-driving car) / Machine translation/Etc.,

Machine Learning vs. Deep Learning

- Deep Learning
  - Deep learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
  - 딥 러닝은 작업별 알고리즘과 달리 학습 데이터 표현을 기반으로 하는 광범위한 기계 학습 방법 제품군의 일부입니다.
- Machine Learning
  - Decision tree
  - Support Vector Machines
  - Artificial Neural Networks
    - Simple (shallow) neural network
    - Deep neural networks (deep learning)
- Because deep learning addresses many problems that have not been studied in the realm of traditional neural networks, it sometimes considered as a separate field.
- 딥러닝은 기존에 해결하지 못한 부분을 해결하는 등 새로운 문제들을 다루기 때문에 아얘 별도의 분야로 보기도 함.

Types of Machine Learning

Types of Machine Learning

- Amount and type of supervision they get during training.
  - **Supervised learning**
  - **Unsupervised learning**
  - **Semi-supervised learning**
  - **Reinforcement learning**

- Ability to learn incrementally from a stream of incoming data.
  - **Batch learning**
  - **Online learning**

- The way they generalize to unseen data.
  - **Instance-based learning**
  - **Model-based learning**

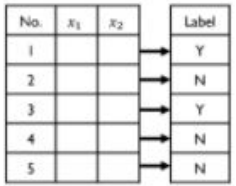
- Amount and type of supervision they get during training.
  - **Supervised learning**
  - **Unsupervised learning**
  - **Semi-supervised learning**
  - **Reinforcement learning**

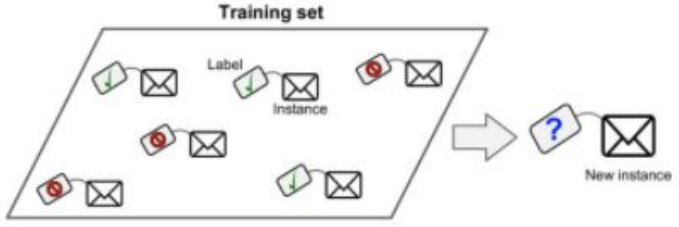
Supervised Learning

- The training data you feed to the algorithm includes the desired solutions, called labels.
- 알고리즘에 제공하는 교육 데이터에는 레이블이라고 하는 원하는 솔루션이 포함됩니다.
- The goal is usually to predict labels for unseen examples.
- 목표는 일반적으로 보이지 않는 예에 대한 레이블을 예측하는 것입니다.
- Humans can label datasets manually to create labeled datasets.
- 인간은 수동으로 데이터 세트에 레이블을 지정하여 레이블이 지정된 데이터 세트를 생성할 수 있습니다.

Supervised Learning – Classification

- Predict the categorical class labels
  - Categorical variable is a variable that can take on one of a limited, and usually fixed number of possible value
  - 범주형 변수는 제한적이고 일반적으로 고정된 수의 가능한 값 중 하나를 취할 수 있는 변수입니다.
    - Blood type
    - race
    - languages



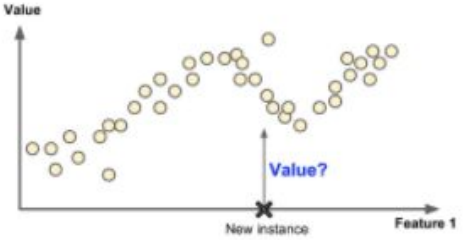


A spam filter is trained with many example emails along with their labels (i.e., spam or not)

Supervised Learning – Regression

- Predict a target numeric value.
    - Blood pressure
    - Stock price
    - Temperature

No.	x_1	x_2	Label
1			30,000
2			40,000
3			28,000
4			15,000
5			10,000

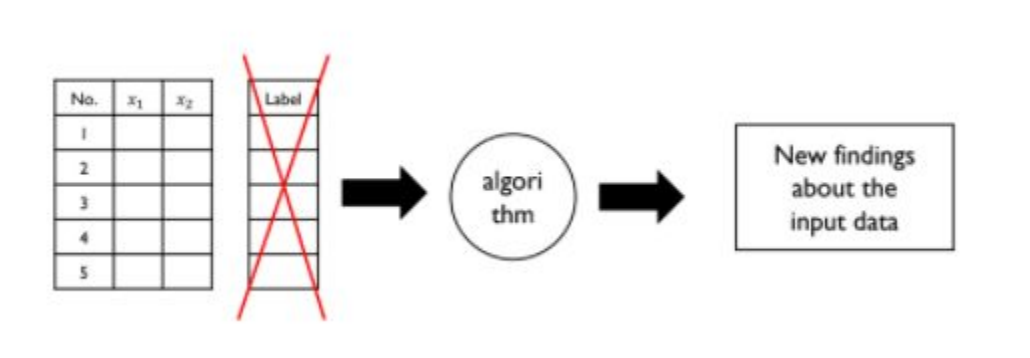


A system predicts the price of a car given a set of features such as mileage, age, brand, etc.

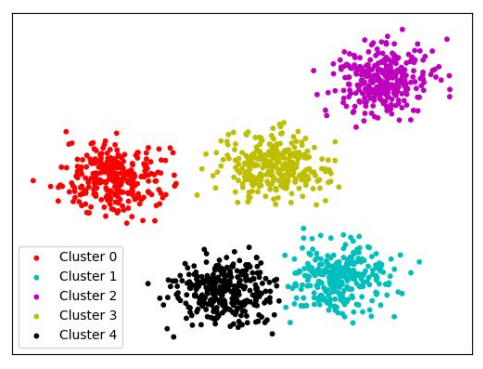
Unsupervised Learning

The training data you feed to the algorithm is **unlabeled.**

- The goal is usually to understand the input data better



Unsupervised Learning – Clustering



Detects groups of similar objects

- The goal is to group a set of objects in such a way that objects in the same group (clust
other than to those in other groups.

- 목표는 동일한 그룹(클러스터)에 있는 개체가 다른 그룹에 있는 개체보다 서로 더 유사하도록 개체 집합을 그룹화하는 것입니다

  - E.g., Cluster website visitors into different groups based on their behaviors.

  .- 예: 웹사이트 방문자를 행동에 따라 다른 그룹으로 묶습니다

  - E.g., Cluster search results into groups of similar web pages

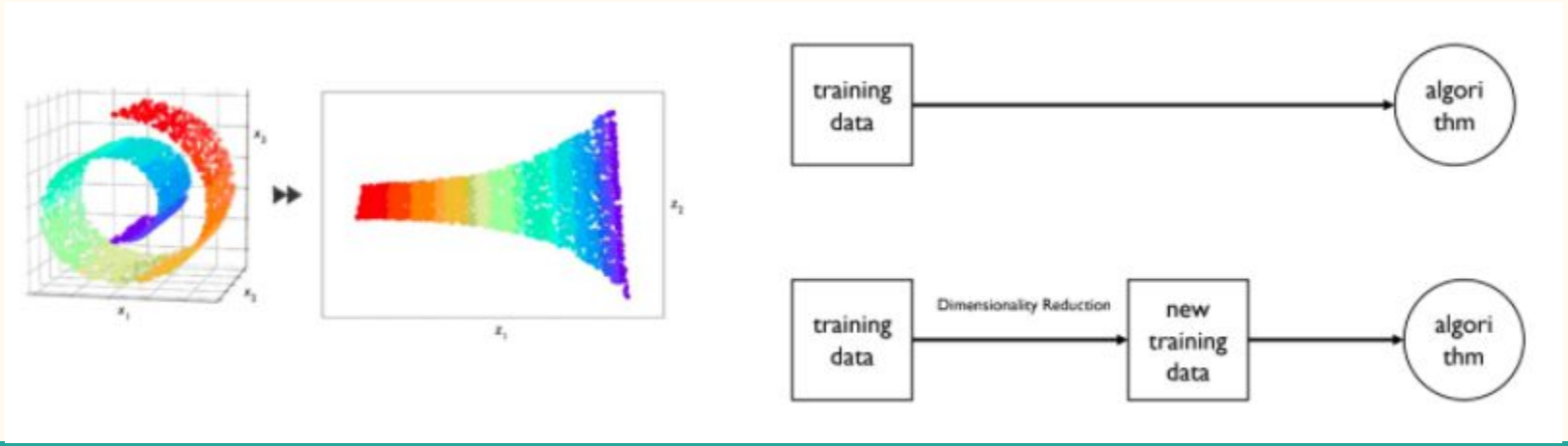
  .- 예: 검색 결과를 유사한 웹 페이지 그룹으로 묶음

- Clustering is usually followed by another exploratory analysis (e.g., explore data within a specific group, labeling, etc.)

주로 데이터 탐색 과정에서 사용합니다.

Unsupervised Learning – Dimensionality Reduction

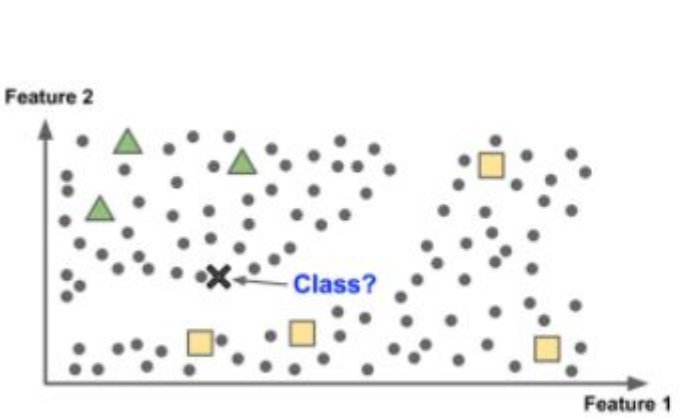
- Reduces the number of dimensions of high-dimensional data.(고차원 데이터의 차원 수를 줄입니다.)
  - It can be used to make the data more interpretable. (데이터를 해석가능하도록 만들때)
  - It can be used as a preprocessing method to generate better input data. (더 좋은 입력데이터 생성을 위한 처리법)
    - Data analysis (e.g., classification, regression, etc.) can be done in the reduced space more accurately than in the original space. (원래의 공간보다 더 축소된 정보가 더 정확할 수 있다.)



Semi-supervised Learning

- The training data you feed to the algorithm is partially labeled, usually a large amount of unlabeled data and a small amount of labeled data.
- 많은 라벨 없는 데이터 + 약간의 라벨 있는 데이터

- Most semi-supervised learning are combination **of unsupervised and supervised algorithms**.



Reinforcement Learning

• In reinforcement learning, the learning system, called an agent, can observe the environment, select and perform actions, and get rewards in return or penalties in the form of negative rewards.
• 강화 학습에서 에이전트라고 하는 학습 시스템은 환경을 관찰하고, 행동을 선택하여 수행하고, 부정적인 보상의 형태로 보상이나 벌칙을 받을 수 있습니다.

• The agent learns by itself what is the best strategy, called a policy, to get the most reward over time.
• 에이전트는 시간이 지남에 따라 가장 많은 보상을 받기 위해 정책이라고 하는 최상의 전략이 무엇인지 스스로 학습합니다.

- A policy defines what action the agent should choose when it is in a given situation
- **정책**은 주어진 **상황**에서 에이전트가 선택해야 하는 **작업**을 정의합니다.
  - E.g., DeepMind's AlphaGo
    - It learned its winning policy by analyzing millions of games.

- Ability to learn incrementally from a stream of incoming data.
  - **Batch learning**
  - **Online learning**

Batch Learning

• The system is incapable of learning incrementally, it must be trained using all the available data.
가능한 모든 데이터를 사용하여 학습.

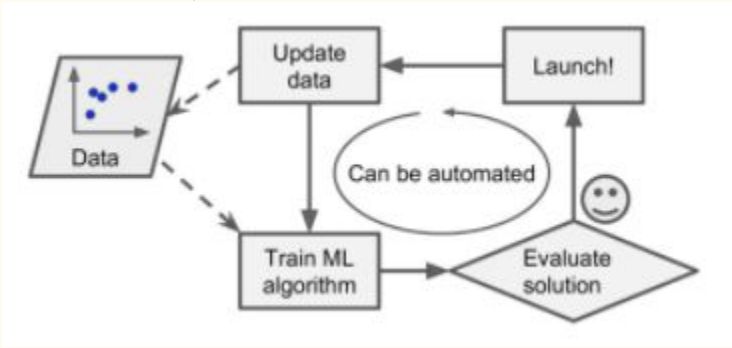
• This generally takes a lot of time and computing resources, so it is typically done offline (offline learning).

시간과 자원이 많이 필요하므로 오프라인 학습.

• If you want the system to know about new data, (새로운 데이터 학습하려면? 처음부터!)

  • Retrain the system from scratch on the full dataset (old data + new data).

  • Stop the old system and replace it with the new one.



Online Learning

• The system can learn incrementally by being fed data instances sequentially, either individually or by small groups called mini-batches.

• The system can learn about new data on the fly, as it arrives.

• Great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously.

• Online learning can also be used to train systems on huge datasets that cannot fit in one machine's main memory.

Online Learning

• The system can learn incrementally by being fed data instances sequentially, either individually or by small groups called mini-batches.
• 시스템은 개별적으로 또는 미니 배치라고 하는 소규모 그룹별로 순차적으로 데이터 인스턴스를 공급받아 점진적으로 학습할 수 있습니다

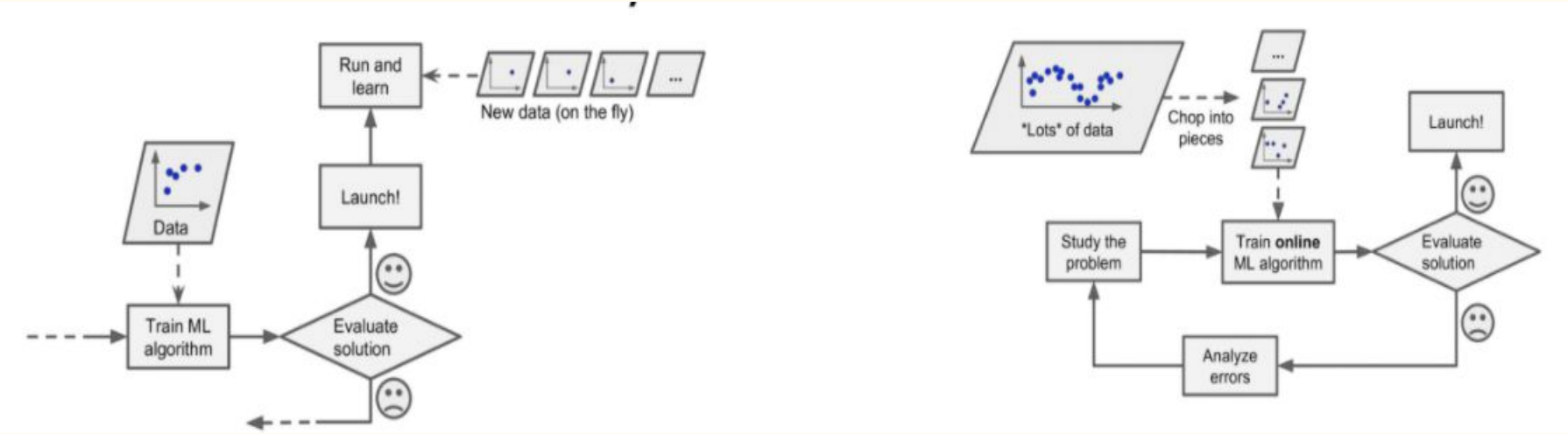
• The system can learn about new data on the fly, as it arrives.
• 시스템은 새로운 데이터가 도착하는 즉시 학습할 수 있습니다

• Great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously. .• 데이터를 연속적인 흐름(예: 주가)으로 수신하고 빠르게 또는 자율적으로 변화에 적응해야 하는 시스템에 적합합니다

• Online learning can also be used to train systems on huge datasets that cannot fit in one machine's main memory.

• 온라인 학습은 한 기계의 주 메모리에 들어갈 수 없는 거대한 데이터 세트로 시스템을 훈련하는 데 사용할 수도 있습니다.

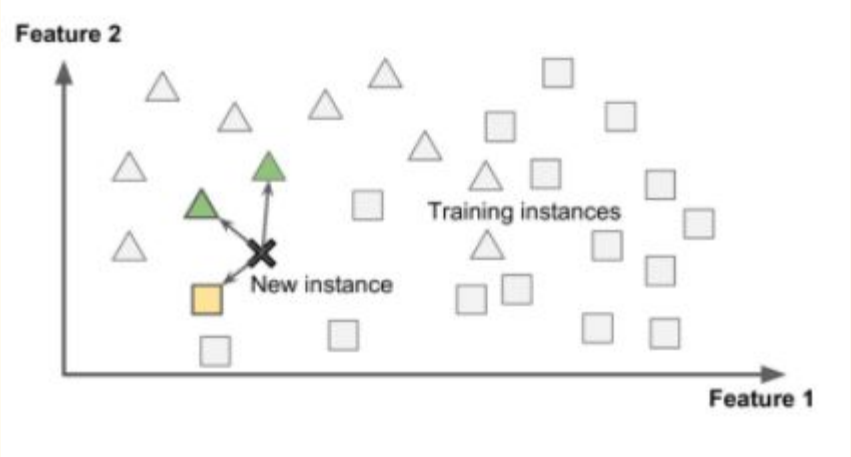
Online Learning



- The way they generalize to unseen data.
  - **Instance-based learning**
  - **Model-based learning**

Instance-based Learning

- Instead of explicit generalization, the system compares new instances with existing instances in the training set based on a measure of similarity.
- 명시적인 일반화 대신에, 유사성 측정을 기반으로 훈련 셋의 기존 인스턴스와 새로운 인스턴스를 비교합니다.
  - E.g., k-nearest neighbors (k-NN)



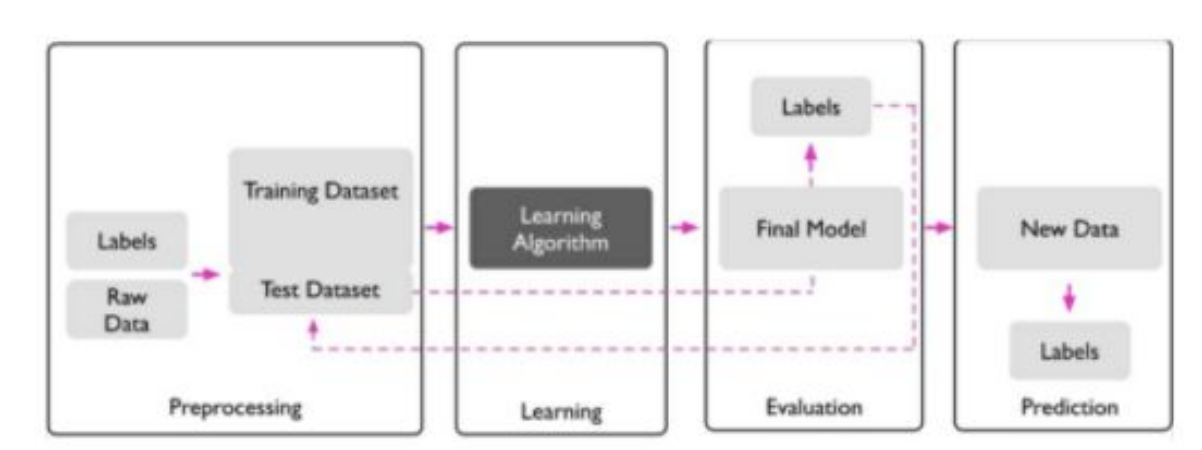
Model-based Learning

- Build a model based on examples, then use that model to make predictions.
먼저 예시(데이터) 로 모델을 만든 후, 그 모델로 예측함

- The approach
    - Study the data
    - Select a model
    - Train it on the training data
    - Apply the model to make predictions on new cases

Machine Learning Workflow

1. Preprocessing: getting data into shape
2. Learning: selecting an algorithm and training models
3. Evaluation: evaluating models
4. Prediction: predicting unseen data instances



Preprocessing: getting data into shape

- Why preprocessing?
  - Garbage in, garbage out
  - Real world data are generally
    - Incomplete: missing values
    - Noisy: containing errors or outliers
    - Inconsistent: containing discrepancies

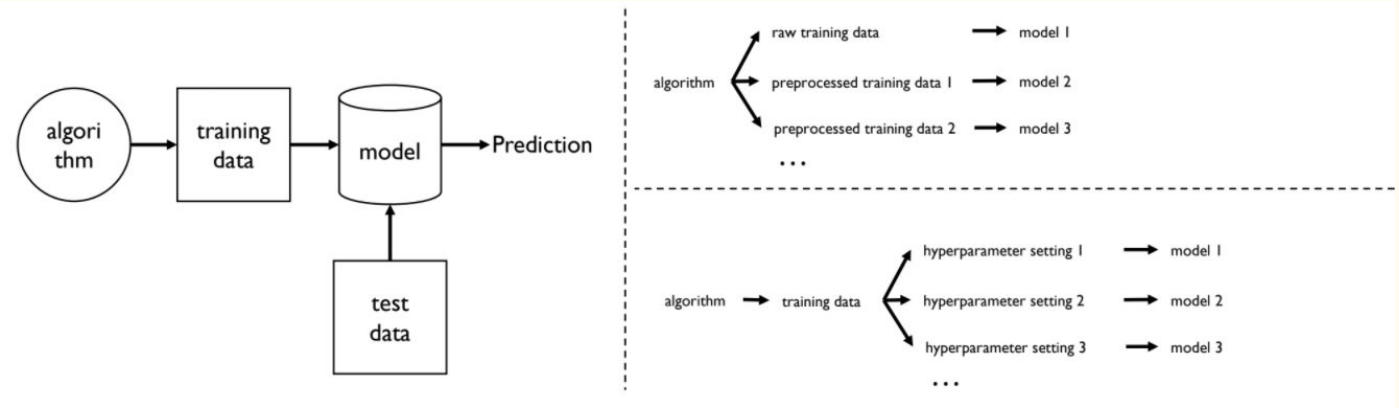
Student	Gender	GPA	Midterm (0-100)	Final
100	M	4.0	90	?
101	F	3.5	800	?
102	Male	NA	70	?
103	F	3.8	85	?
104	Female	3.2	75	?
105	M	3.6	20	?

Learning: selecting an algorithm and training models

- Algorithm : The way to learn models from data
- (데이터로부터 모델을 학습하는 법)

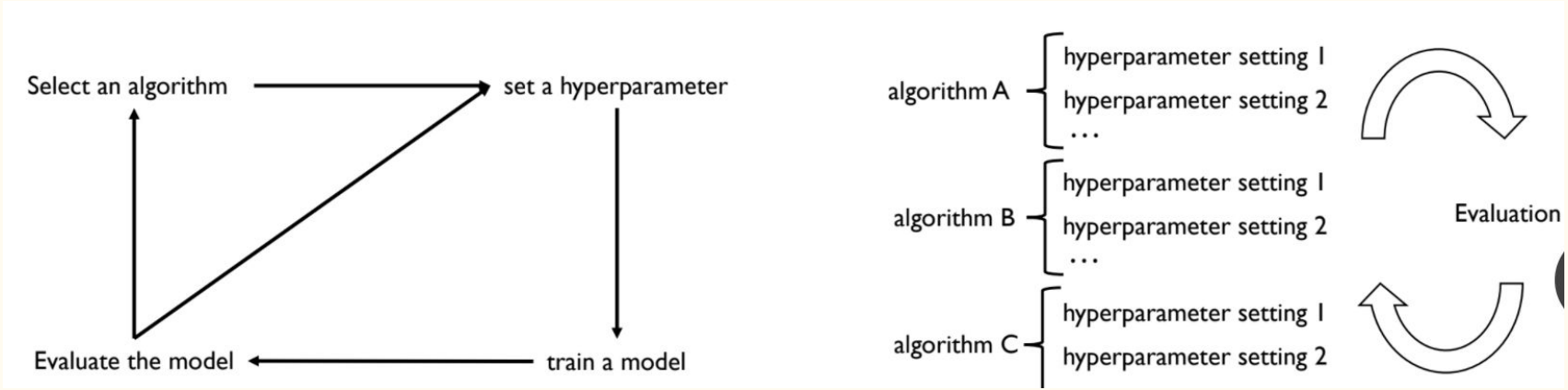
- Model : The product we get by applying an algorithm to data

(데이터에 알고리즘 적용하여 얻은 결과물, 학습한 결과)



Evaluation: evaluating models

- Evaluating & re-learning
  - An iterative process

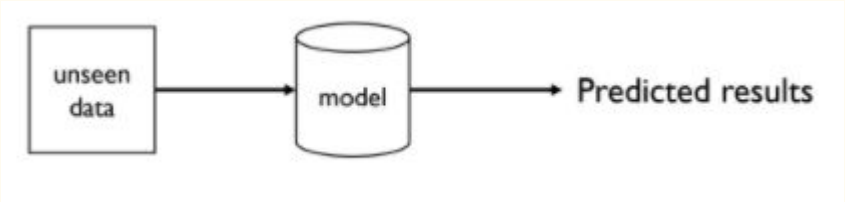


Prediction: predicting unseen data instances

- Apply the trained model to unseen data

처음 보는 데이터를 학습된 모델에 적용하면? 예측된 값이 나온다.

  - The best-performing model



Challenges of Machine Learning (Data)

- Insufficient quantity of training data
  - For simple problems: thousands of examples (데이터 너무 적음)
  - For complex problems: millions of examples (데이터 너무 많음)

  - “We don’t have better algorithms. We just have more data.” —Peter Norvig (Research Director at Google)

  - Better data ≠ more data

Challenges of Machine Learning (Data)

- Non-representative training data (트레이닝 데이터는 대표성을 띄지 않을 때)
    - It is crucial to use a training set that is representative of the cases you want to generalize to. (일반화 하고 싶은 케이스의 대표적인 값이 트레이닝으로 들어가는게 중요하다.
    - Machine Learning is only guaranteed to work for data generated by the same distribution that generated its training data.(ML 이 유일하게 보장하는건, 트레이닝 데이터로부터 동일한 분산으로 생성된 데이터에 한해서이다.)
    - E.g., patients of different country/race/gender
        - physical condition
        - eating habits

Challenges of Machine Learning (Data)

- Poor-quality data (데이터의 질이 떨어짐)
  - Training data may have **errors, outliers, missing values, and noise**
  - **Data preparation** accounts for about 80% of the work of data scientists [1].
    - Data scientists spend 60% of their time on cleaning and organizing data.
    - Collecting data sets comes second at 19% of their time.
    - Mining data for patterns (9%)

Challenges of Machine Learning (Data)

- Irrelevant features (관계 없는 특성)
    - Coming up with a good set of features to train on using feature engineering (feature selection & extraction) is important. (피쳐 엔지니어링을 통해 좋은 학습셋을 만드는 것 중요
    - More features are not always the better. (정보가 많다고 무조건 좋은건 아님)
    - One of the important problems in ML is finding features that play important roles in prediction. (예측에 중요한 역할을 하는 피쳐를 찾는 것이 중요하다.)
        - E.g., hospital readmission
            - A patient who had been discharged from a hospital is admitted again within a specified time interval (usually 30 days).
            - Finding important features (e.g., blood pressure).

Challenges of Machine Learning (Algorithm)

- Overfitting the training data (오버피팅 문제)
  - A model performs well on the training data, but it **does not generalize well.**
  - Happens when the model is too complex relative to the amount of the training data.

- Solutions
  - Simplifying the model by selecting one with fewer parameters, by reducing the number of attributes in the training data, or by constraining the model (i.e., regularization).
  - Gathering more training data.
  - Reducing the noise in the training data (e.g., fix data errors and remove outliers)

Challenges of Machine Learning (Algorithm)

Underfitting the training data (언더피팅 문제)

- Happens when the model is too simple to learn the underlying structure of the data.

- Solutions
- Selecting a more powerful model, with more parameters.
- Feeding better features to the learning algorithm through feature engineering.
- Reducing the constraints on the model (e.g., reducing the regularization hyperparameters)

Transfer learning