CHAPTER 01

Basics of machine learning

01 What is machine learning?

02 Machine Learning Processes and Types

03 Building a machine learning environment

01
What is machine learning?

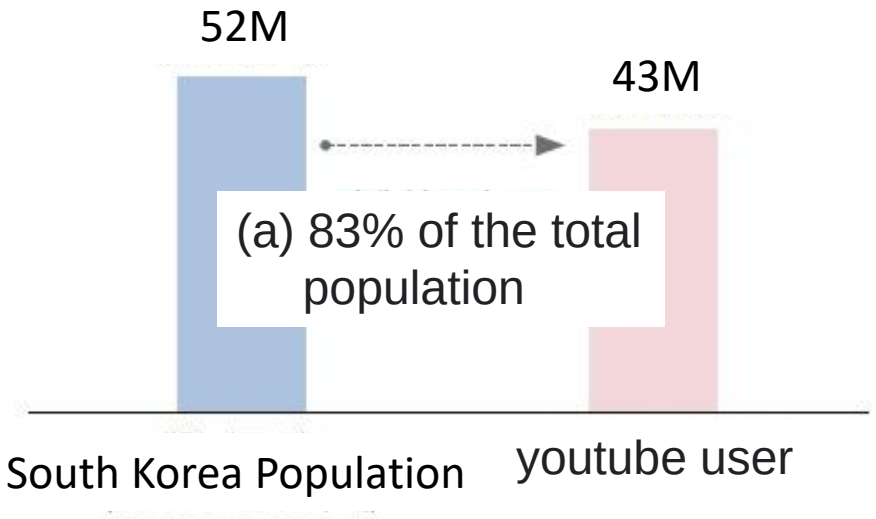
01 What is machine learning?

1. The age of machine learning

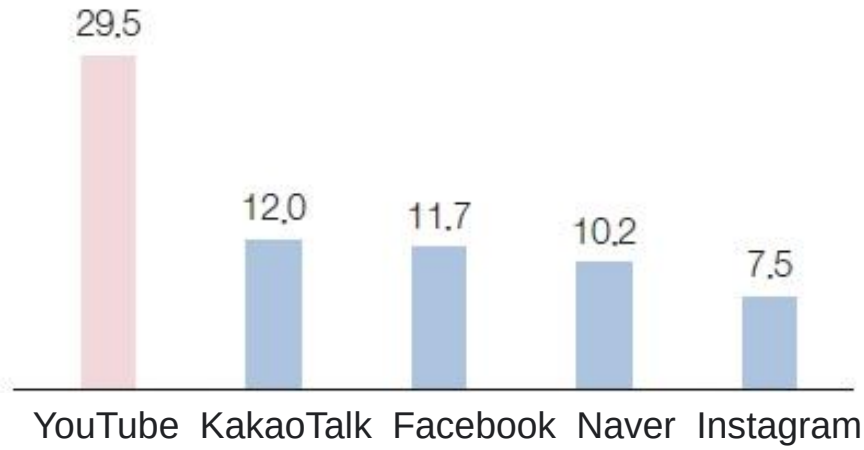
- People are drawn to a certain 'logic' made by the machine to watch the video.
- ‘google profile’ → ‘Data and privacy protection’ → ‘Ad personalization’

01 What is machine learning?

- YouTube applied a deep learning-based recommendation algorithm in 2016
- YouTube is the #1 most popular smartphone app



(a) YouTube app users



(b) Average monthly usage time for major apps

01 What is machine learning?

- Algorithms: A set of procedures or methods to solve a problem.
- Machine learning: Algorithms in which machines learn patterns and automate them

- YouTube makes a program (machine learning) that learns about the pattern of individuals watching YouTube videos, and then continues to recommend the next video according to the pattern (algorithm).

2. Practical use cases for machine learning

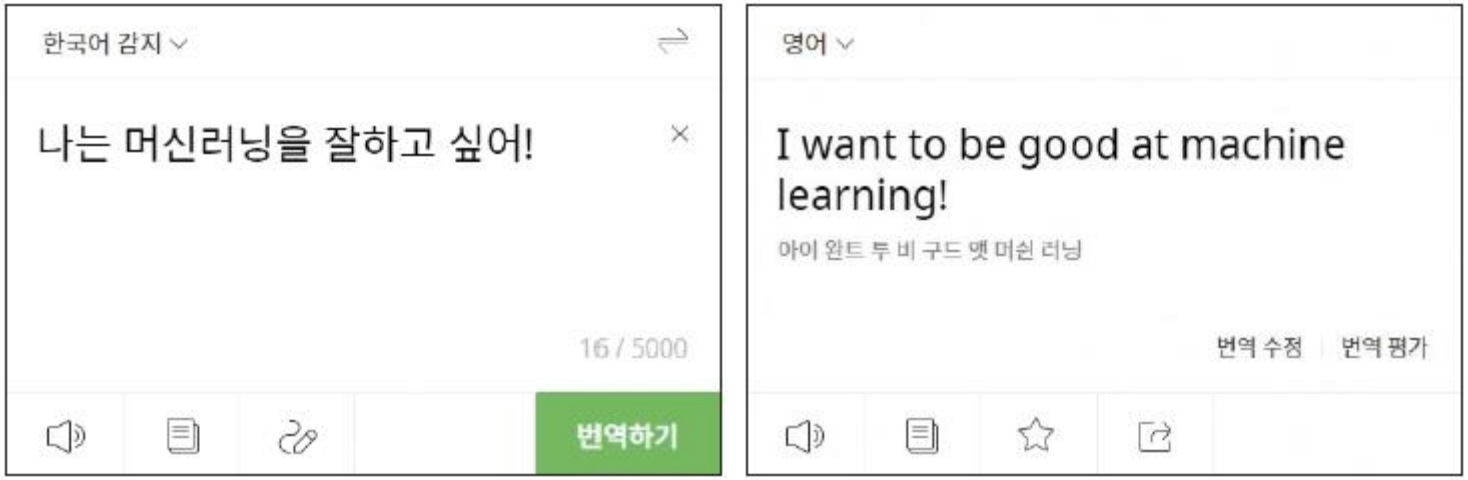
2.1 Recommendation system

- Recommend purchasing products similar to the products added to the shopping cart in the online shopping mall

01 What is machine learning?

2.2 translation

- Train machine learning to translate the translation of a professional translator to translate new sentences
- No linguistic experts on the Google Translate team



2.3 self-driving car

- By using image processing technology, one of the machine learning technologies, various images on the road are learned, and the computer in the vehicle judges and operates it by itself.

01 What is machine learning?

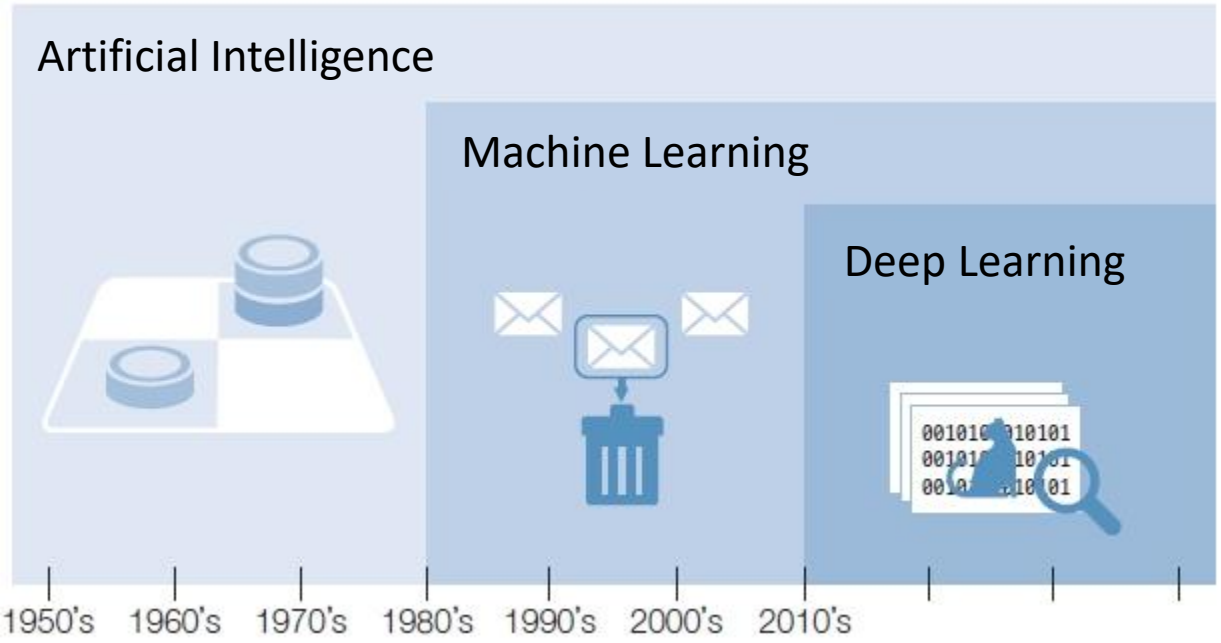
2.4 Chatbot

- Supports user-computer conversations based on machine learning
- Due to the COVID-19 outbreak, artificial intelligence call centers that respond by mimicking human voices appear
- Machine learning learns human conversation patterns and

3. Machine learning keywords

3.1 Artificial Intelligence, Machine Learning, Deep Learning

- Artificial Intelligence ⊃ Machine Learning ⊃ Deep Learning



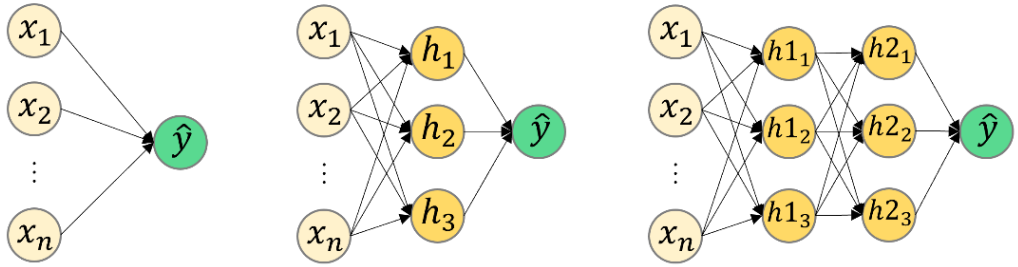
01 What is machine learning?

- Artificial Intelligence (AI): A technology that enables computers to learn, think, and make their own decisions.
  - After the term 'artificial intelligence' was first used at the Dartmouth Conference in 1956, it began to develop into a full-fledged academic field.

Dartmouth Conference in 1956

01 What is machine learning?

- Machine Learning: A technology that allows a computer to learn patterns and rules by learning data from a computer.
  - Previously, people directly databased knowledge and then programmed it to be processed by computers.
  - Machine learning programs a mathematical model to classify data, so by inputting data, an already created mathematical model is applied as a rule to solve many problems.
- Deep learning: A technology used to cluster or classify objects or data based on a neural network among machine learning techniques.

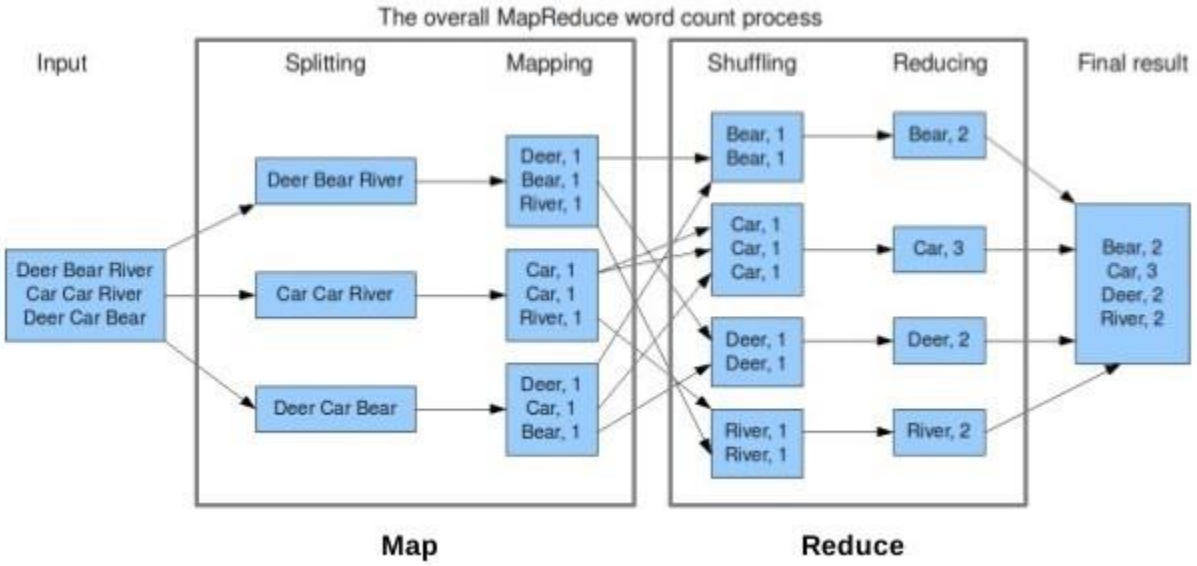


3.2 Statistical Machine Learning and Deep Learning

- As deep learning develops rapidly, traditional machine learning and recently used deep learning are often viewed as a parallel relationship.
- Deep learning algorithms are often created by computer scientists.
  - Deep learning and error backpropagation were proposed by Professor Jeffrey Hinton.
- Existing machine learning has been developed mainly in the field of statistics.
  - Algorithms such as regression analysis or decision trees

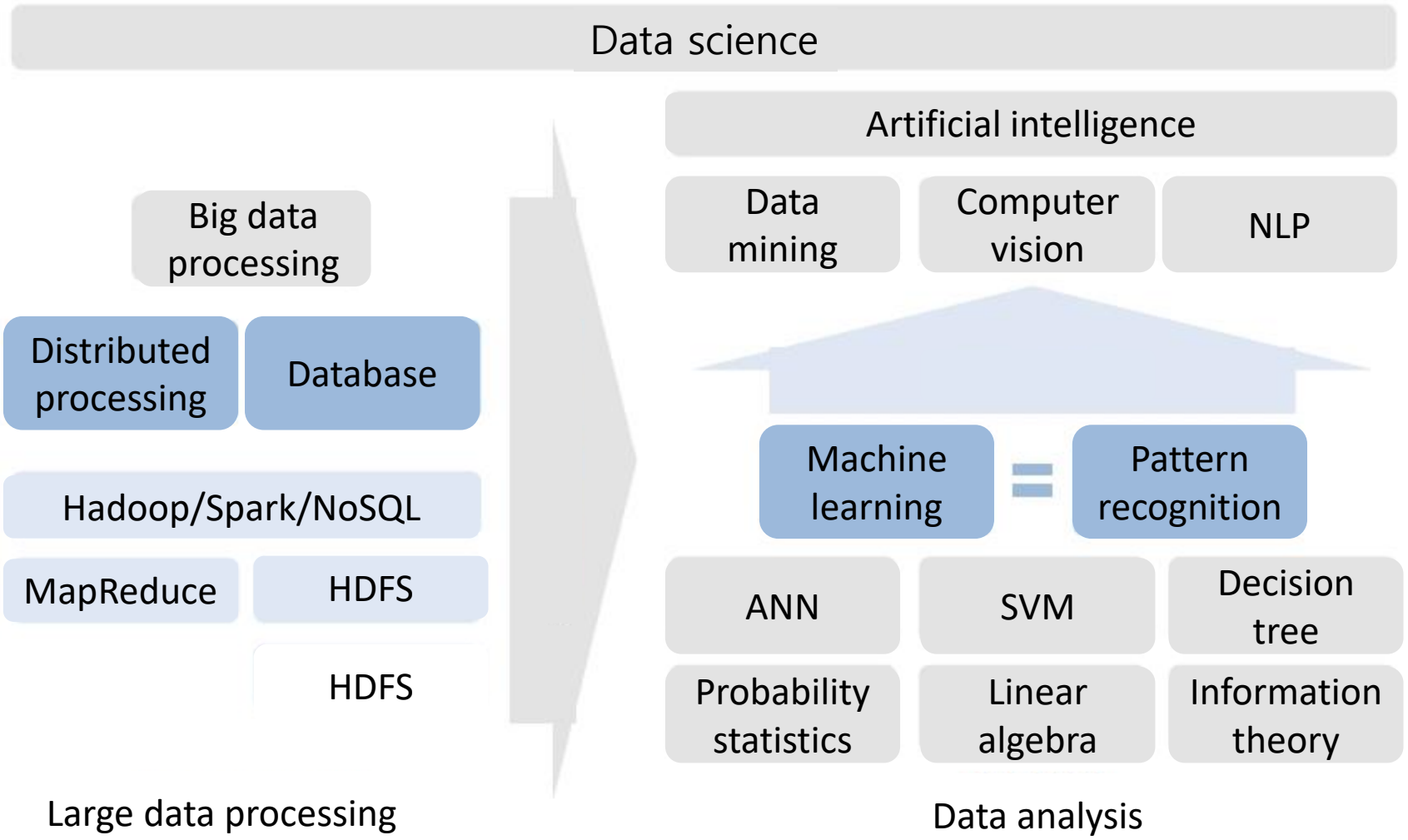
01 What is machine learning?

- GFS (Google File System): A system designed to store and manage Google's huge search system by distributing data across multiple computers.
- Map Reduce: A system that tells how to process distributed and stored data in cooperation with multiple computers at the same time
- Hadoop: A system developed by expanding the concept of Google's distributed processing system. Handling large amounts of data today



01 What is machine learning?

Artificial intelligence ecosystem



02
Machine Learning Processes and Types

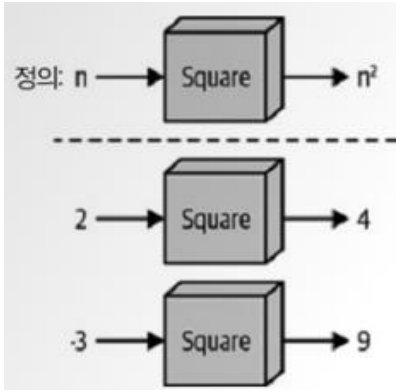
1. Learning process in machine learning

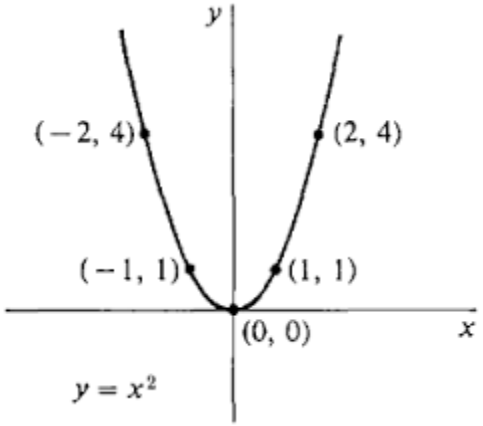
- Basic model of machine learning

$$\hat{y} = f(x)$$

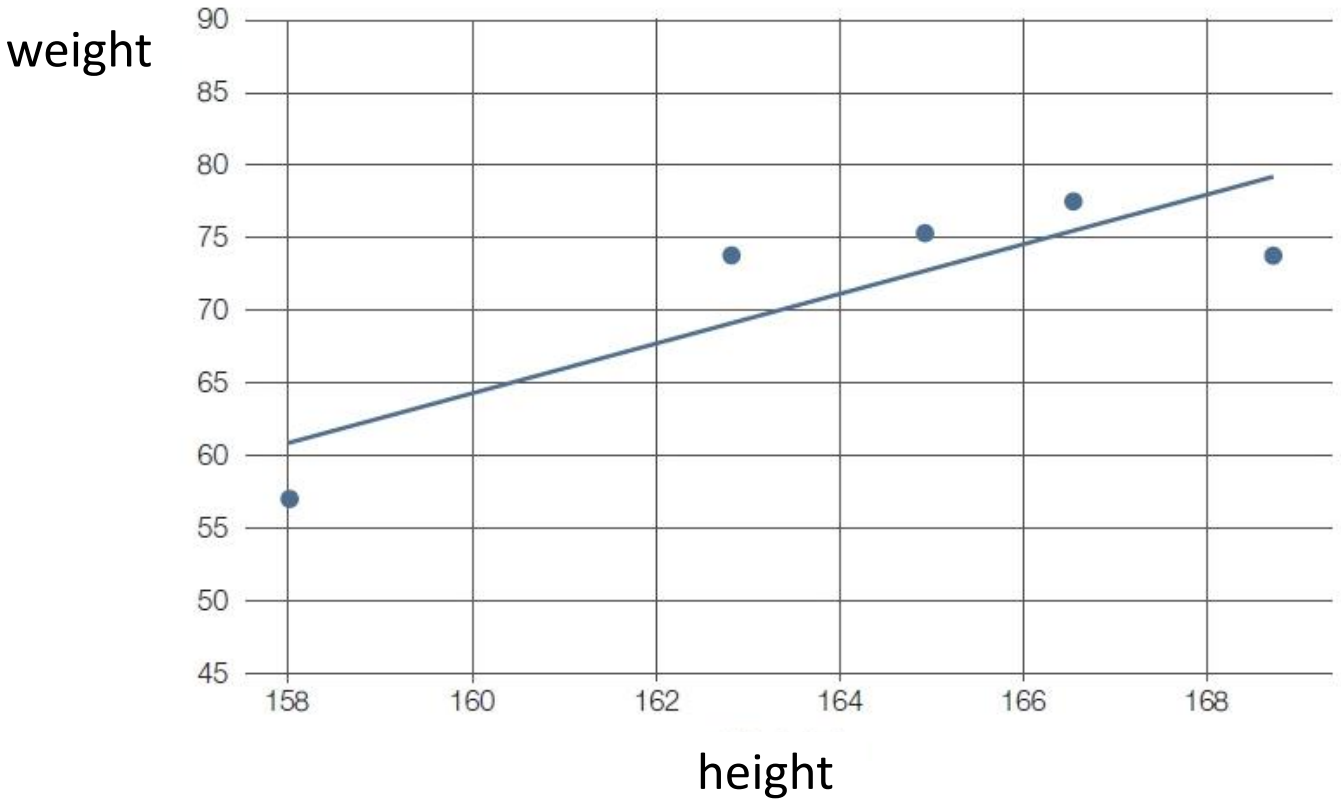
- When the data (x) is put into the function $f(\,)$ generated by the machine learning model trained, the result is a prediction $\hat{y}$.
- Machine learning gives us the $f(\,)$ we are looking for.

$$f_1(x) = x^2$$





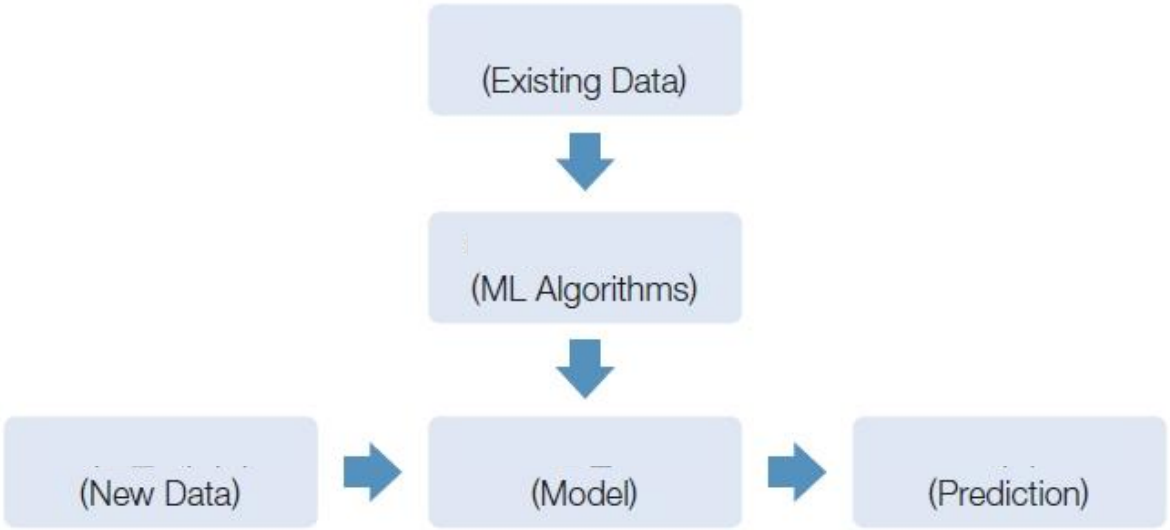
height	158	168.7	162.8	166.5	164.9
weight	57.1	74.1	74.1	77.9	75.5



- Model: Expression of correlation as an expression

$$y = \alpha x + \beta$$

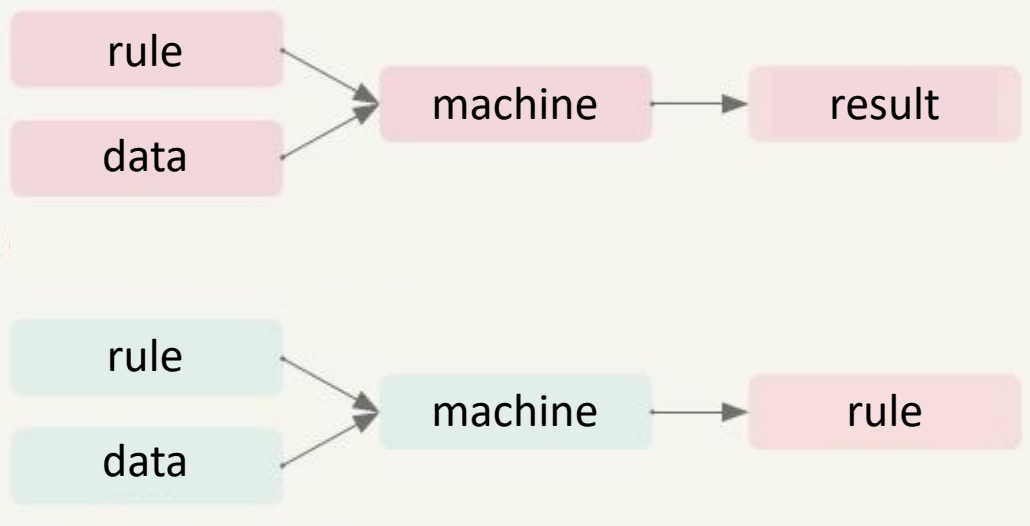
- In a situation where you do not know which model will best fit the relationship between height and weight, you can find the model you are looking for if you complete a linear equation like this
  - Algorithm to find suitable α and β to predict weight for new height



- In general, a model can be expressed by various algorithms.
- Sometimes an 'algorithm' can be expressed as a 'formula'

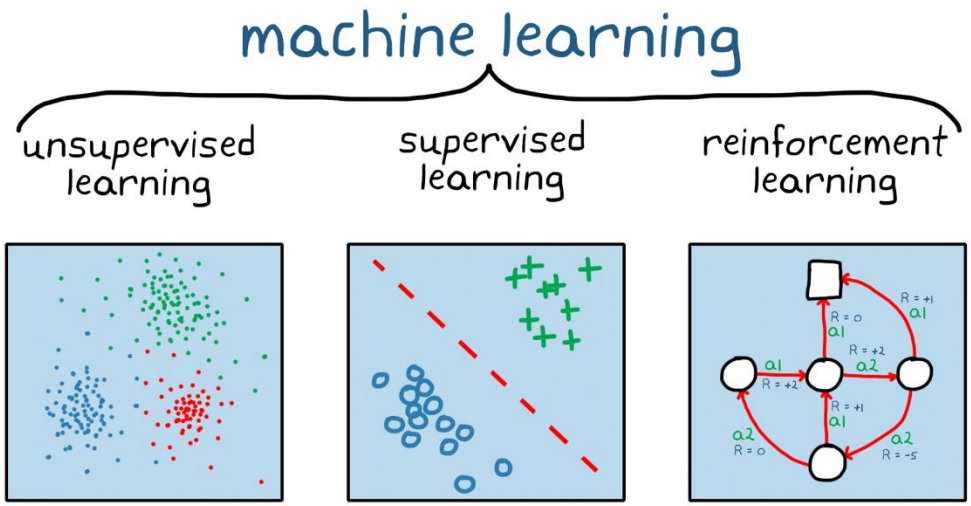
Programming before machine learning

- Existing programming techniques have a structure in which a person specifies programming rules and then inputs data to obtain results.
- In machine learning-based programming, a structure that extracts rules from data and results when results and data are input



2. Types of Machine Learning

- Supervised learning: learning problems and answers together
- Unsupervised learning: A computer learns itself without the help of an assistant. The computer uses the training data to find regularities between the data.
- Distinguish based on the existence of ground truth “y”



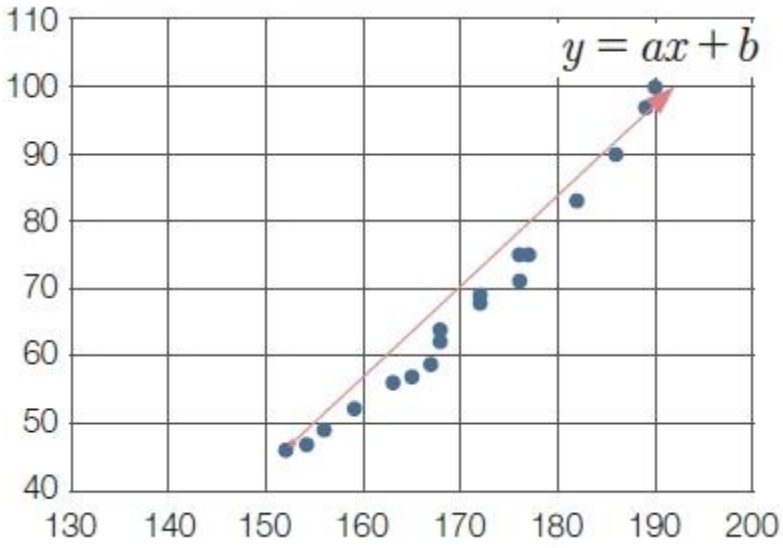
2. Types of Machine Learning

- Reinforcement learning: A computer learns rules like a game while simulating the rules that exist in the world.
- Generation: A model that learns various rules that exist in the world creates something new that does not exist in the world.
  - computer generated human face
  - A computer communicates with a human in the form of a chatbot

2.1 regression

- Regression: Describe the relationship between the independent variable x and the dependent variable y with a functional formula
- Techniques for creating mathematical models that represent trend lines

height	weight	height	weight
152	46	172	69
154	47	172	68
156	49	176	71
159	58	176	75
163	56	177	75
165	57	182	83
167	59	186	90
168	64	189	97
168	62	190	100



02 Machine Learning Processes and Types

유전인자, 방사선 노출, 식이
X_1 X_2 X_3

음주, 흡연, 스트레스
X_4 X_5 X_6

$H(x)=w_1x_1+w_2x_2+w_3x_3+ w_4x_4+w_5x_5+w_6x_6+ b$

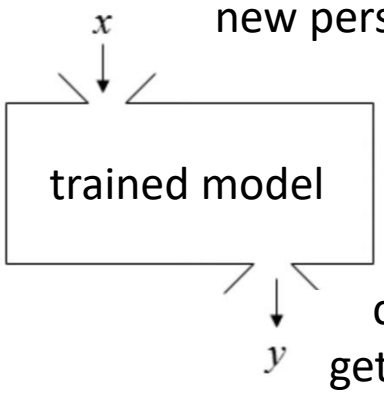
W는 가중치, X는 각 요인

$$H(x) = w_1x_1+w_2x_2+w_3x_3+\ldots+b$$

암에 걸릴 확률 = 유전인자의 중요도(W_1)* 유전인자 수치(X_1)+
방사선노출의 중요도(W_2)*방사선 노출량(X_2)+...

(W = 각 X에 대한 중요도를 구별하기 위해 도입)

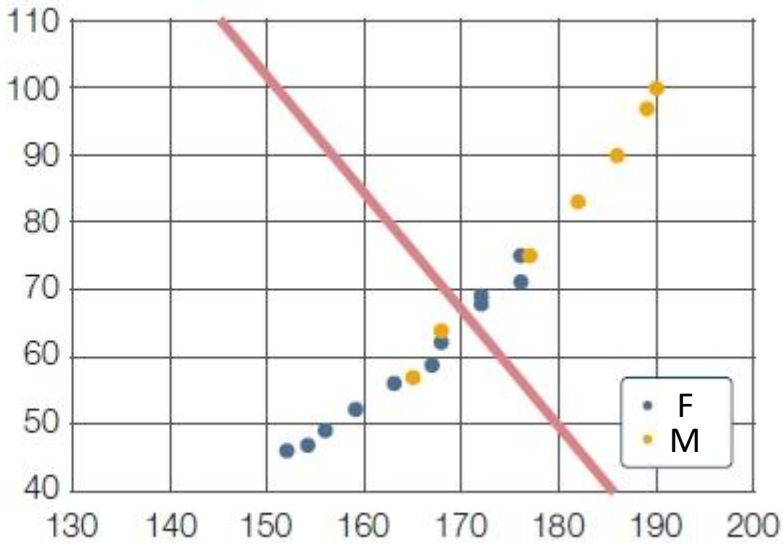
결론(암걸렸는지 유무) = 유전인자의 영향도(W_1)* 개별 유전인자(X_1)+
방사선노출의 영향도(W_2)* 개별 방사선 노출량(X_2)+
식습관 영향도(W_3)* 개별 식습관(X_3)+ ... +b



2.2 Classification

- Classification: to divide data according to some criteria
- Binary classification: classify one out of two values
- Multi-class classification: run 3 or more classifications

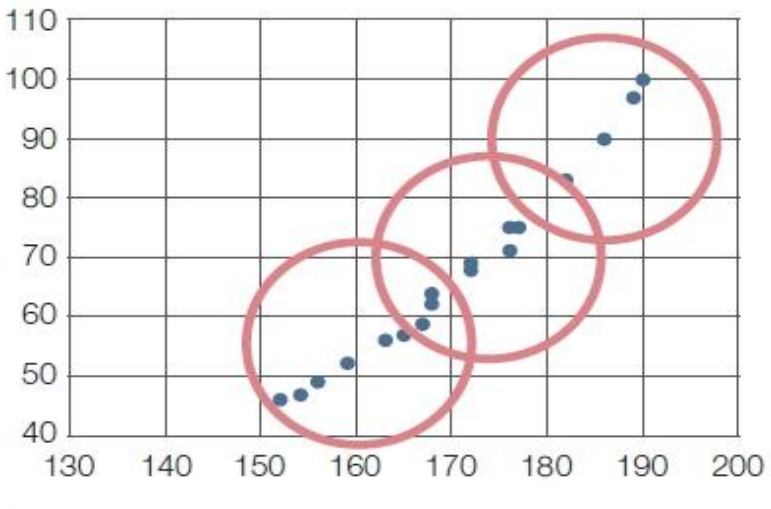
sex	height	weight	sex	height	weight
F	152	46	M	172	69
F	154	47	F	172	68
F	156	49	F	176	71
F	159	58	F	176	75
F	163	56	M	177	75
M	165	57	M	182	83
M	167	59	M	186	90
M	168	64	M	189	97
F	168	62	M	190	100



2.3 Clustering

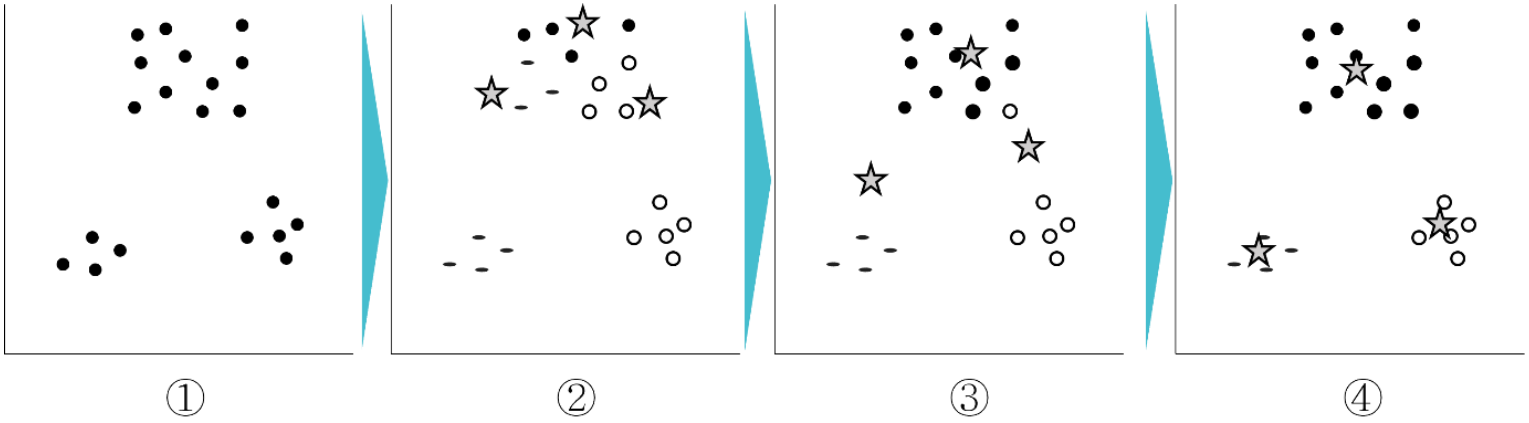
- Clustering: A technique in which a model finds a classification criterion on its own and collects groups without providing a separate classification criterion for previously collected data.
- Create 3 basketball teams of similar level

height	weight	height	weight
152	46	172	69
154	47	172	68
156	49	176	71
159	58	176	75
163	56	177	75
165	57	182	83
167	59	186	90
168	64	189	97
168	62	190	100



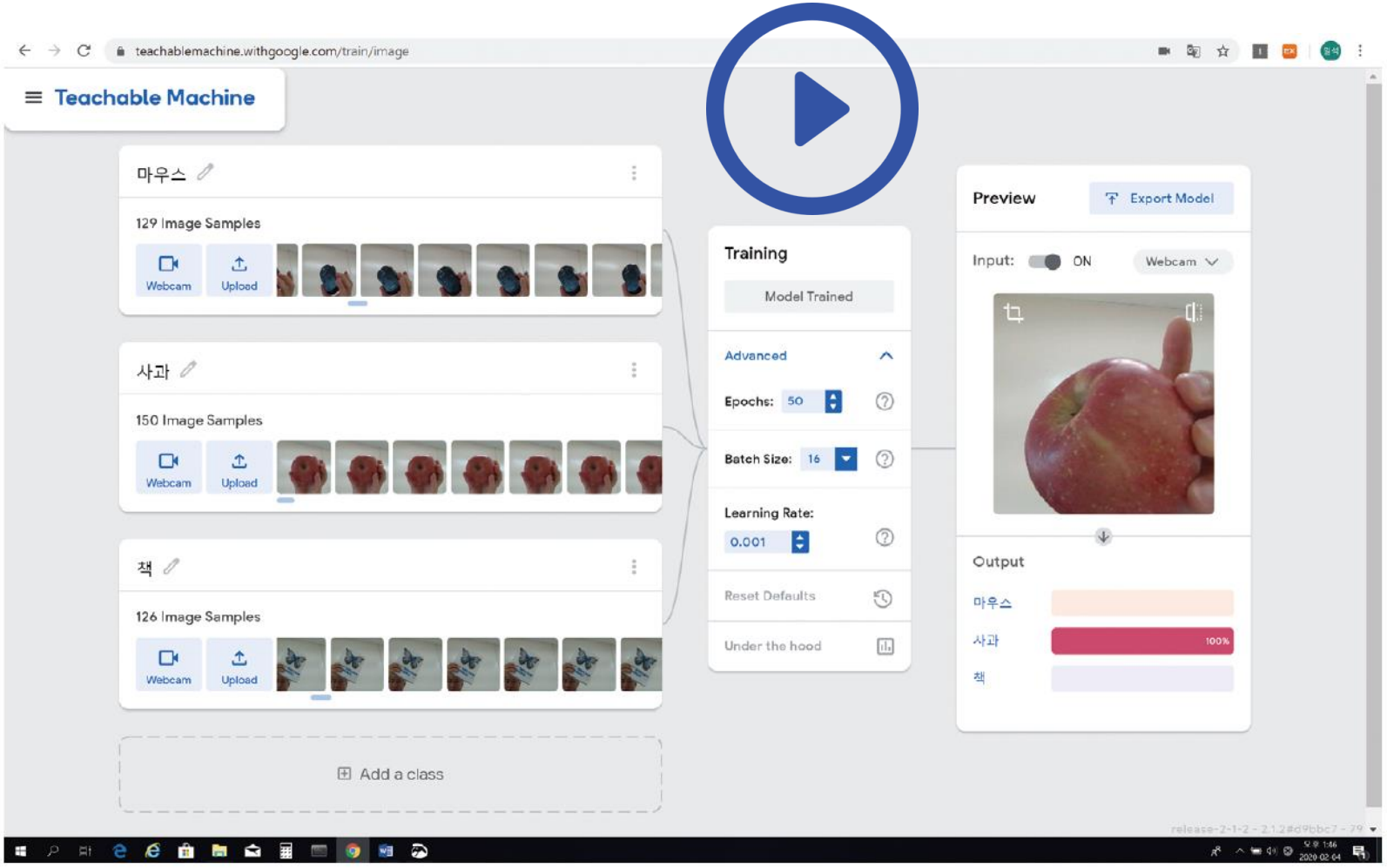
2.3 Clustering

- k-means clustering
  - Algorithms that group observations with similar characteristics



- Step 1: Select k centroids in an arbitrary data space
- Step 2: Calculate the Euclidean distance between each centroid and the observations
- Step 3: Assign observations close to each centroid to the cluster
- Step 4: Calculate the Euclidean distance between the observations of the assigned cluster and the corresponding centroid.
- Step 5: Move the centroid to the center of the cluster (minimum point of distance between observations in the cluster)
- Step 6: Repeat steps 2-5 until the center point no longer moves

2.4 teachable machine



03
Building a machine learning environment

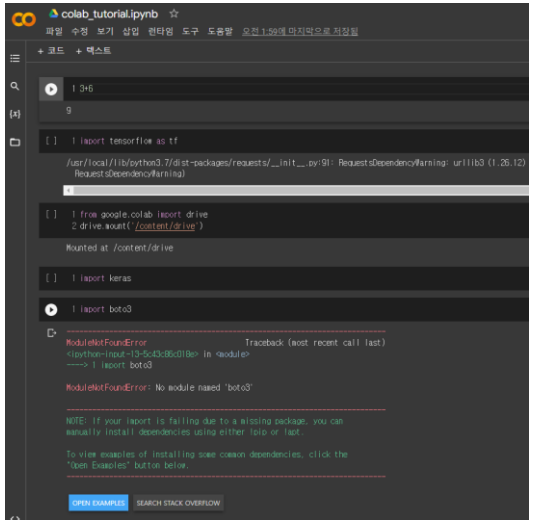
3.1 Cloud and Standalone

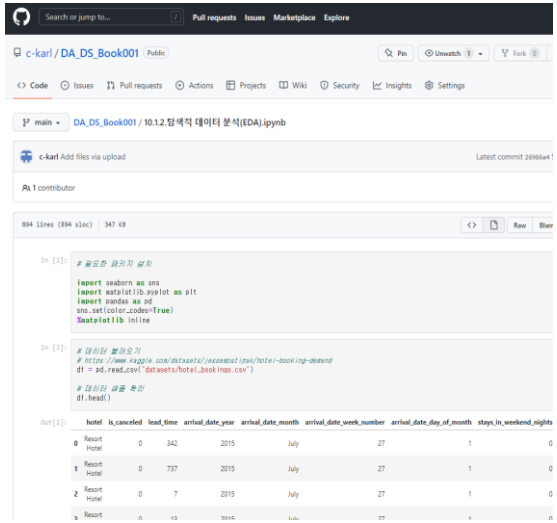
- Cloud
  - Programs and data are stored and managed on the server
    - The server has most of the environment, so you can program right away when you log in
    - You can develop anywhere as long as you have an internet connection. Collaboration possible.
  - Google's Colab, Amazon's SageMaker, Microsoft's Azure
  - There is a limit to not having an optimal environment for my project

- Standalone
  - It is possible to build an environment that is optimal for you. Programs and data stored on your computer
  - You have to install the software and build the environment yourself

3.2 Introduce the Colab and demonstration with Github

- Jupyter laptops powered by Google Cloud
- GPU available for free
- Github friendly environment
- Easy cloud-based collaboration
- Easy to use data science library





3.3 practice

- Upload the dataset to a new Colab notebook to check it out.
- Check each statistic in the dataset.
- Save the note as a copy on Github.
- Connect to Github, check the saved notes, and access the Colab link.

- 새로운 Colab 노트에 데이터셋을 업로드하여 확인한다.
- 데이터셋의 각 통계치를 확인한다.
- 해당 노트를 Github에 사본으로 저장한다.
- Github에 접속하여 저장된 노트를 확인 및 Colab 링크로 접속해본다.

3.4 Demonstration of Logistic regression

```
[ ]
# RobustScaler 적용

# 숫자형 변수 분리
df_num = df[['BMI','PhysicalHealth','MentalHealth','SleepTime']]
df_nom = df2.drop(['BMI','PhysicalHealth','MentalHealth','SleepTime'],axis=1)

# 숫자형 변수 RobustScaler 적용
RobustScaler = RobustScaler()
df_robust = RobustScaler.fit_transform(df_num)

# 컬럼명 결합
df_num2 = pd.DataFrame(data=df_robust, columns=df_num.columns)

# 숫자형 데이블과 더미화 문자형 데이블 결합
df3 = pd.concat([df_num2,df_nom],axis=1)

df3.head()
```

	BMI	PhysicalHealth	MentalHealth	SleepTime	HeartDisease_Yes	Smoking_Yes	AlcoholDrinking_Yes	Stroke_Yes	DiffWalking_Yes	Sex_Male	...	Diabetic_Yes	Diabetic_Yes (during pregnancy)	Physic
0	-1.453315	1.5	10.0	-1.0	0	1	0	0	0	0	...	1	0	
1	-0.947226	0.0	0.0	0.0	0	0	0	1	0	0	...	0	0	
2	-0.102842	10.0	10.0	0.5	0	1	0	0	0	1	...	1	0	
3	-0.423545	0.0	0.0	-0.5	0	0	0	0	0	0	...	0	0	
4	-0.491204	14.0	0.0	0.5	0	0	0	0	1	0	...	0	0	

5 rows × 38 columns