[기업 연계를 위한 AI-IoT 과정]

인공지능

AI, 두 번째 날

2019.08.13.

첫 번째 날

- 인공지능 개론
- 인공지능 사례
- 머신러닝 개론
- Classification 모델 생성

두 번째 날

- Regression 모델 생성
- 머신러닝 알고리즘
- Scikit-learn 모델 생성

세 번째 날

- 딥러닝 이론 소개
- 이미지 분류 방법
- Keras를 활용한 이미지 분류 모델 생성

네 번째 날

다섯 번째 날

- AI 프로젝트

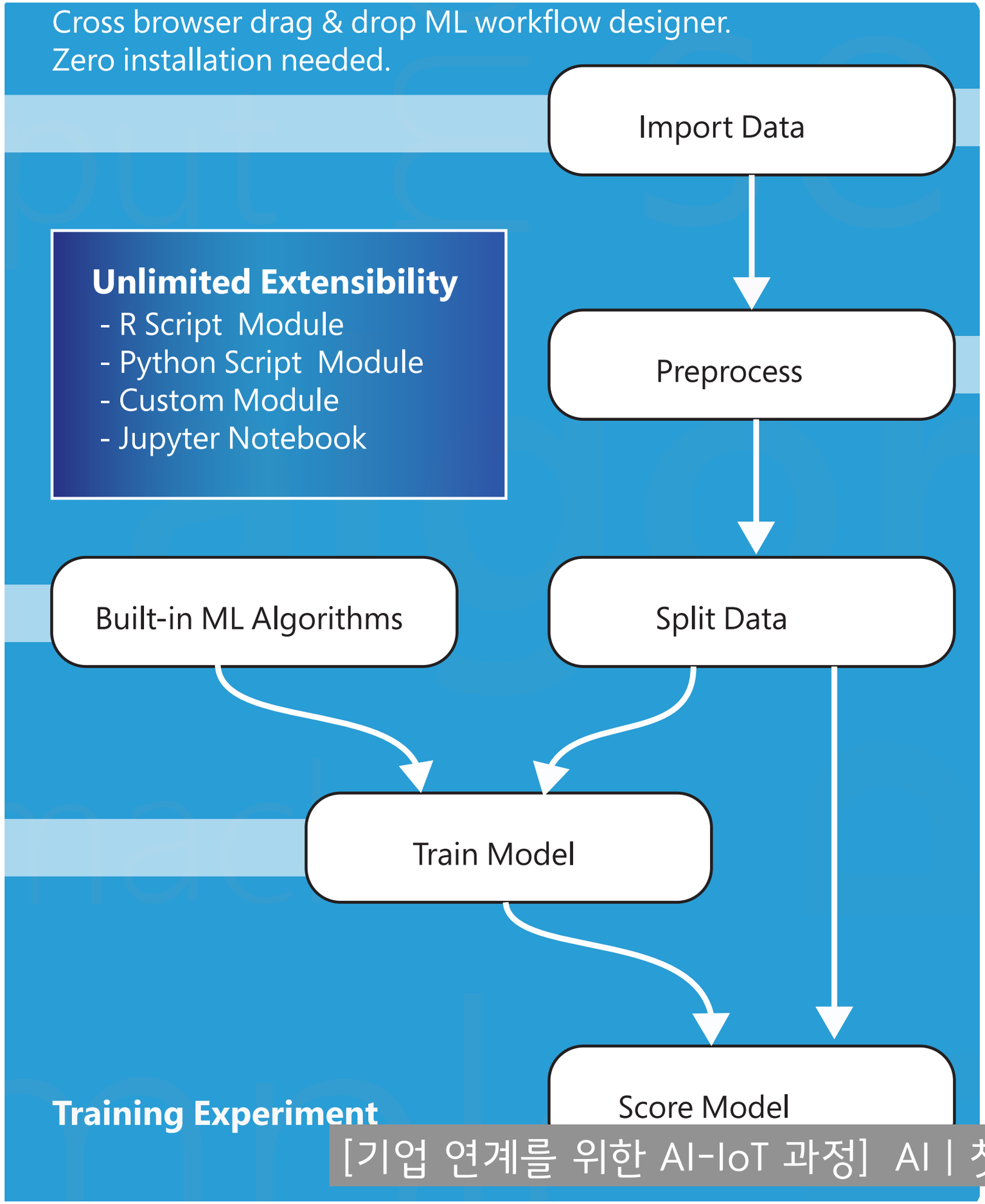
인공지능 기술을 이해하고

생활과 연결할 수 있는 방법을

구상하고 제작할 수 있다

두 번째 날

- Regression 모델 생성
- 머신러닝 알고리즘
- Scikit-learn 모델 생성


Cross browser drag & drop ML workflow designer.
Zero installation needed.
Import Data
Unlimited Extensibility
- R Script Module
- Python Script Module
- Custom Module
- Jupyter Notebook
Preprocess
Built-in ML Algorithms
Split Data
Train Model
Score Model
Training Experiment

문제 정의

데이터 셋 준비

모델 설정

모델 훈련 / 평가

모델 활용

총 111개 모듈

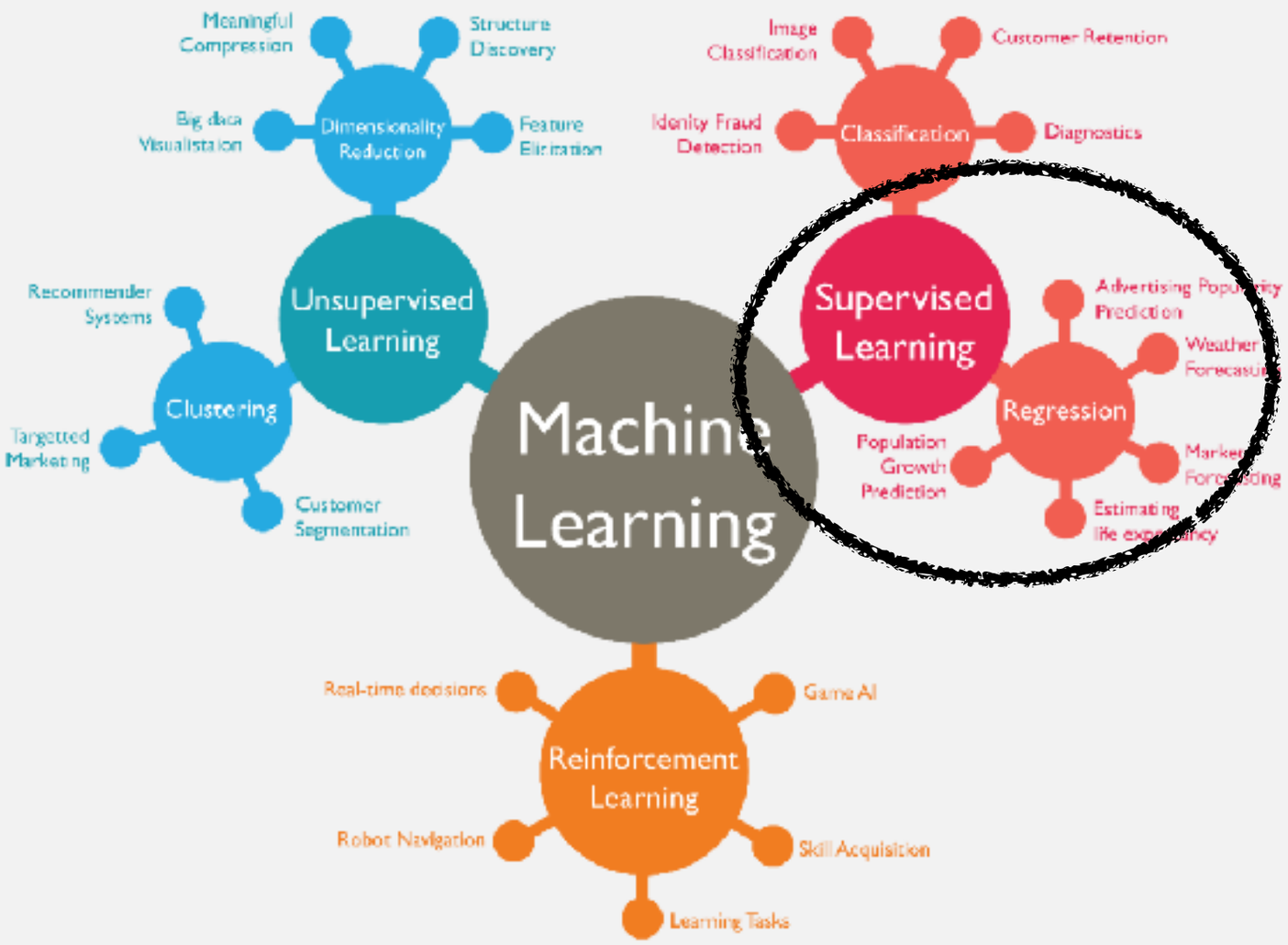
문제 정의

어떤 모델을 만들 것인가?

MACTORES

Data Science Team

Machine Learning Bubble Chart



https://www.mactores.com

미국 45개 월 마트 부서별 주간 판매액 예측 모델

[기업 연계를 위한 AI-IoT 과정] AI | 두번째 날 | 전미정

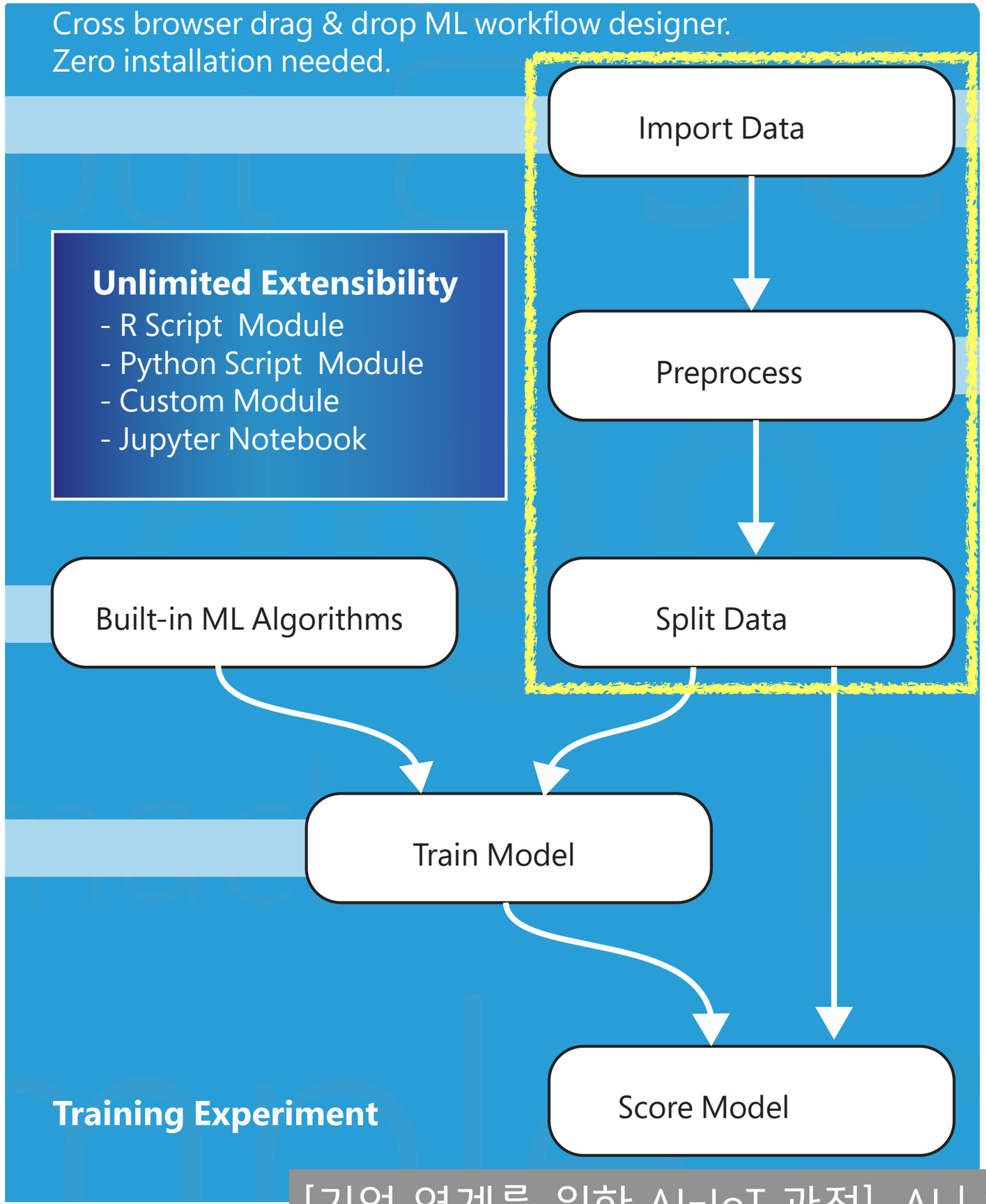
Let’s start HOL 

part1. 데이터 전처리

part2. 모델 학습 / 예측

part1. 데이터 전처리

- 데이터 지원 형식
- 데이터 업로드
- 데이터 합치기
- 데이터 전처리


Cross browser drag & drop ML workflow designer.
Zero installation needed.
Import Data
Preprocess
Split Data
Unlimited Extensibility
- R Script Module
- Python Script Module
- Custom Module
- Jupyter Notebook
Built-in ML Algorithms
Train Model
Score Model
Training Experiment

미국 45개 월 마트 부서별 주간 판매액 예측 모델

[기업 연계를 위한 AI-IoT 과정] AI | 두번째 날 | 전미정

[기업 연계를 위한 AI-IoT 과정] AI | 두번째 날 | 전미정

Two Sigma: Using News to Predict Stock Movements

Use news analytics to predict stock price performance

Featured · Kernels Competition · 2 months to go · news agencies, time series, finance, money

$100,000
2,927 teams

Jigsaw Unintended Bias in Toxicity Classification

Detect toxicity across a diverse range of conversations

Featured · Kernels Competition · a month to go · biases, nlp, text data

$65,000
2,138 teams

LANL Earthquake Prediction

Can you predict upcoming laboratory earthquakes?

Research · 14 days to go · earth sciences, physics, signal processing

$50,000
4,127 teams

Google Landmark Recognition 2019

Label famous (and not-so-famous) landmarks in images

Research · 14 days to go

$25,000
217 teams

Google Landmark Retrieval 2019

Given an image, can you find all of the same landmarks in a dataset?

Research · 14 days to go

$25,000
127 teams

Walmart Recruiting - Store Sales Forecasting

Use historical markdown data to predict store sales

691 teams · 5 years ago

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Join Competition

Overview

- **Description**
- Evaluation
- Prizes
- Timeline

One challenge of modeling retail data is the need to make decisions based on limited history. If Christmas comes but once a year, so does the chance to see how strategic decisions impacted the bottom line.

45개 Walmart 부서별 주간 판매량

2010-02-05 ~ 2012-11-01

12열 X 8,190행

feature.csv

- 지점
- 날짜
- 온도
- 연료비
- 프로모션 * 5
- 소비자 물가 지수
- 실업률
- 휴일 여부

5열 X 421,570행

train.csv

- 지점
- 부서
- 날짜
- 판매량
- 휴일 여부

3열 X 45행

stores.csv

- 지점
- 유형
- 규모

http://bit.ly/walmart_data3

Walmart_dataset

features.csv

train.csv

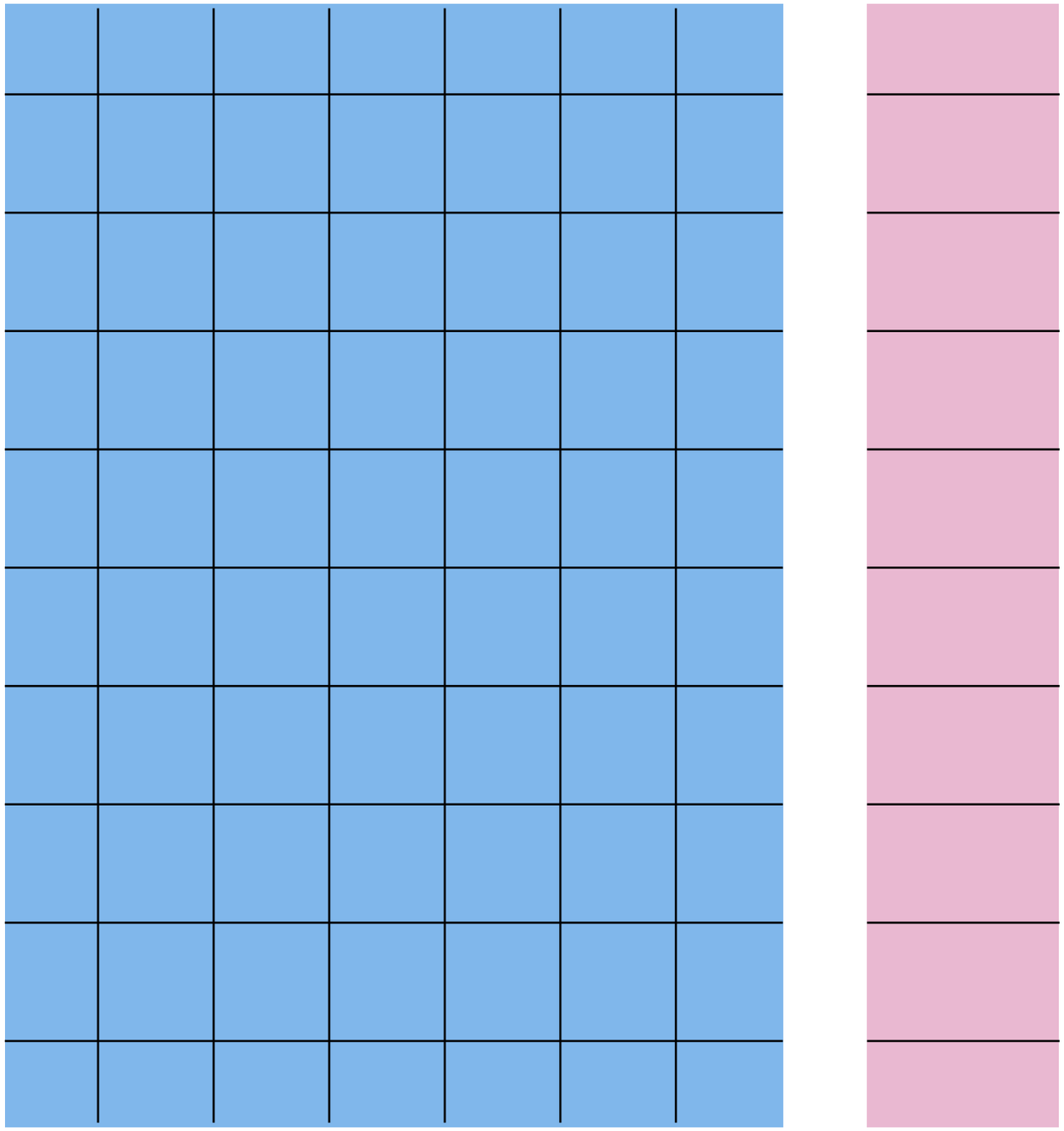
store.csv

login

https://studio.azureml.net

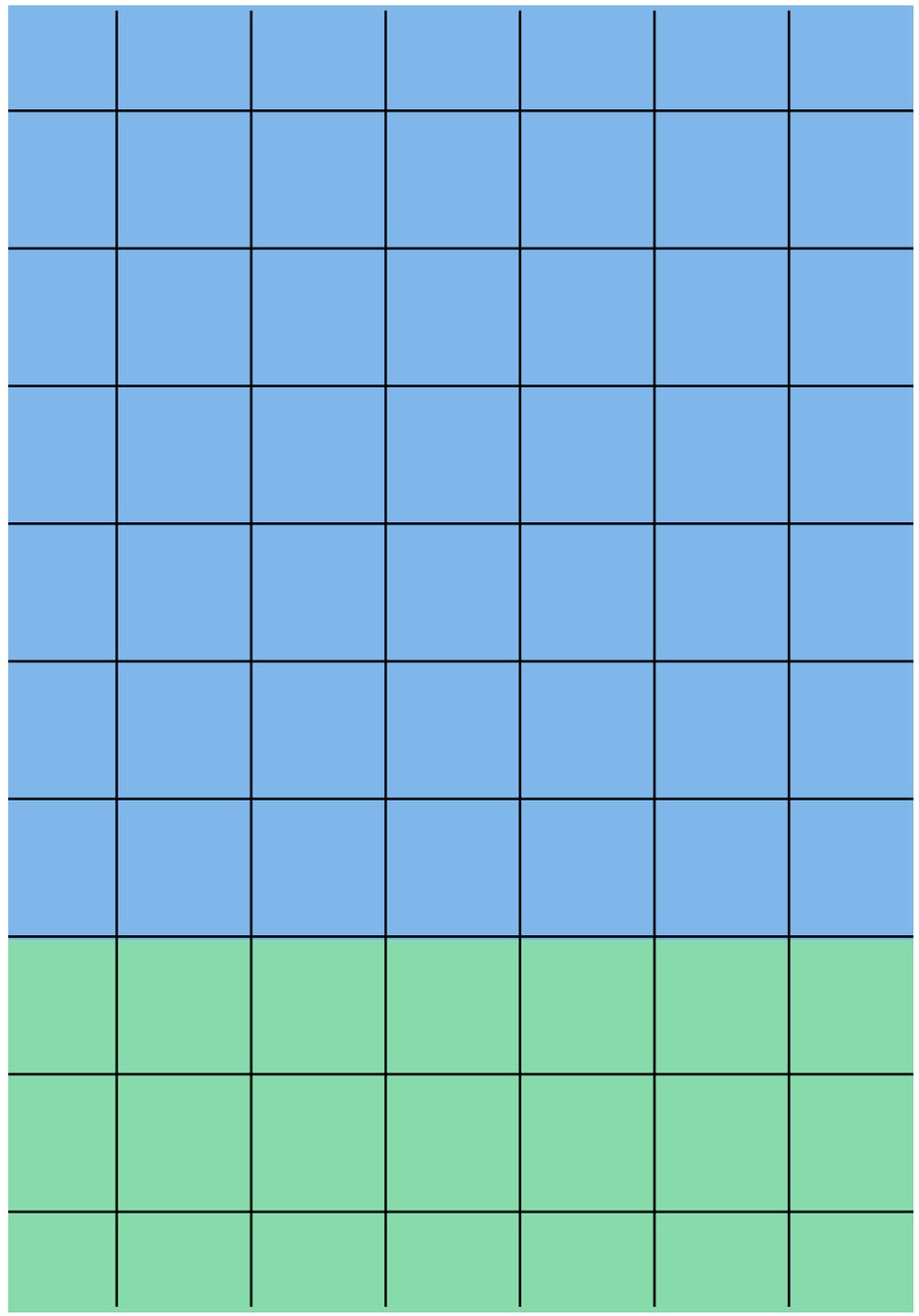
X

Y = *Label*



X

Y = *Label*



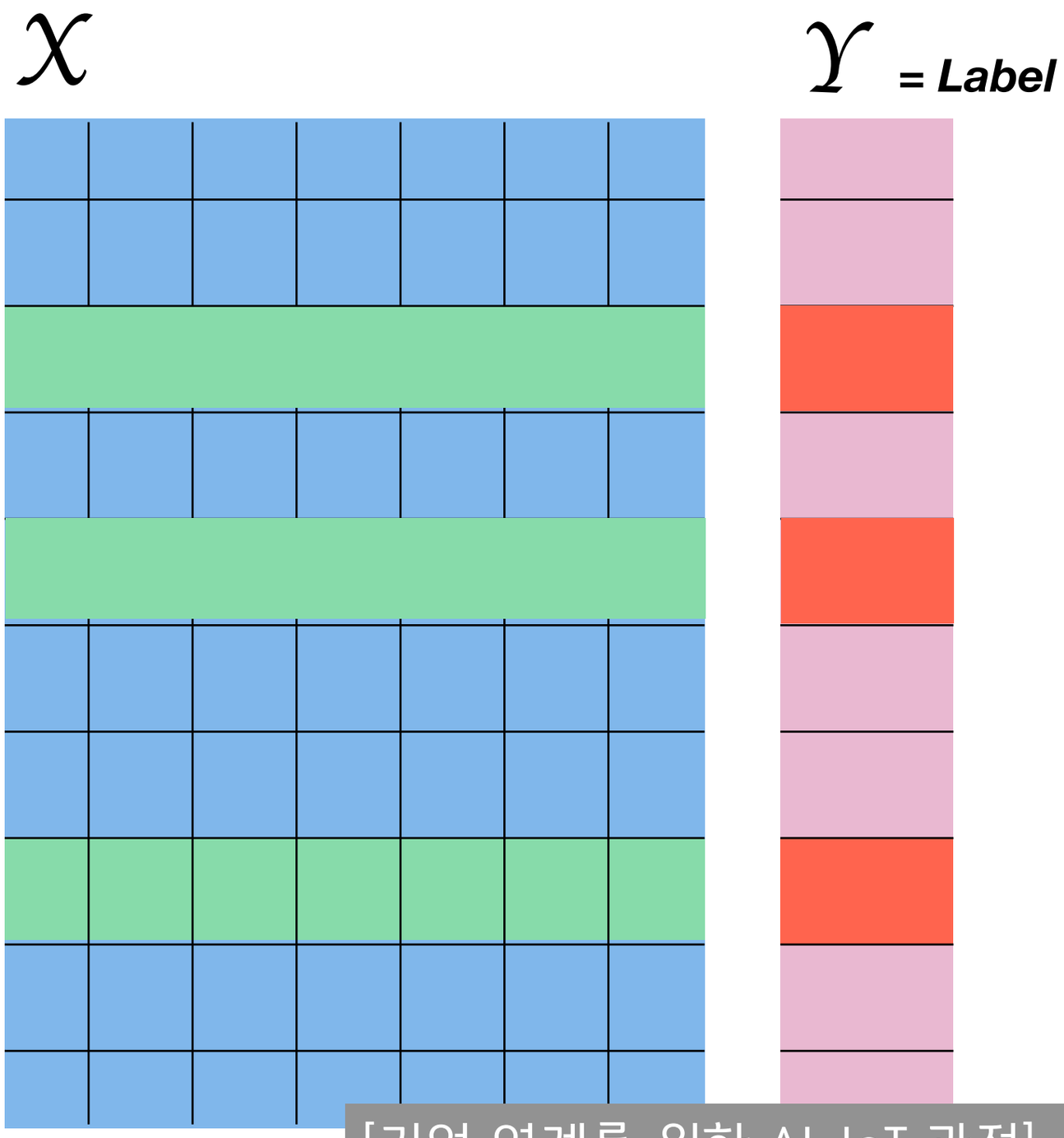
XTrain

YTrain

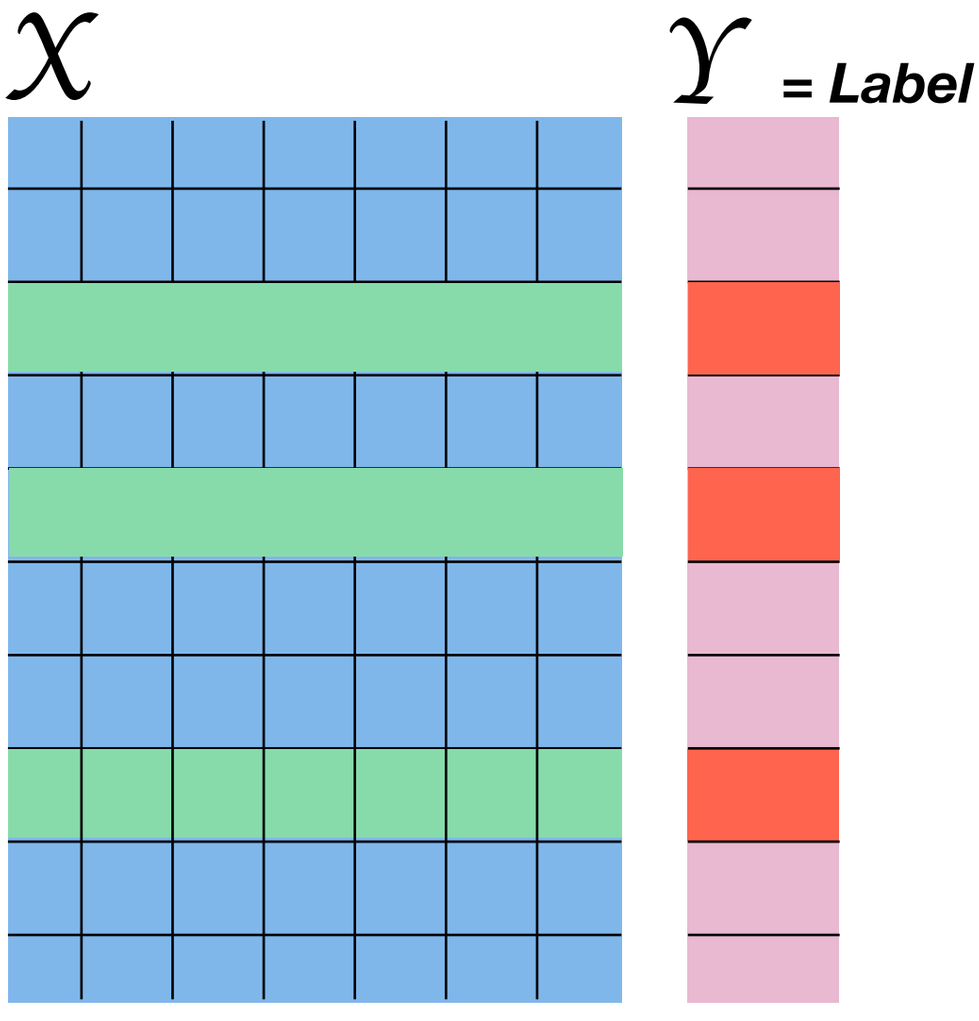
XTest

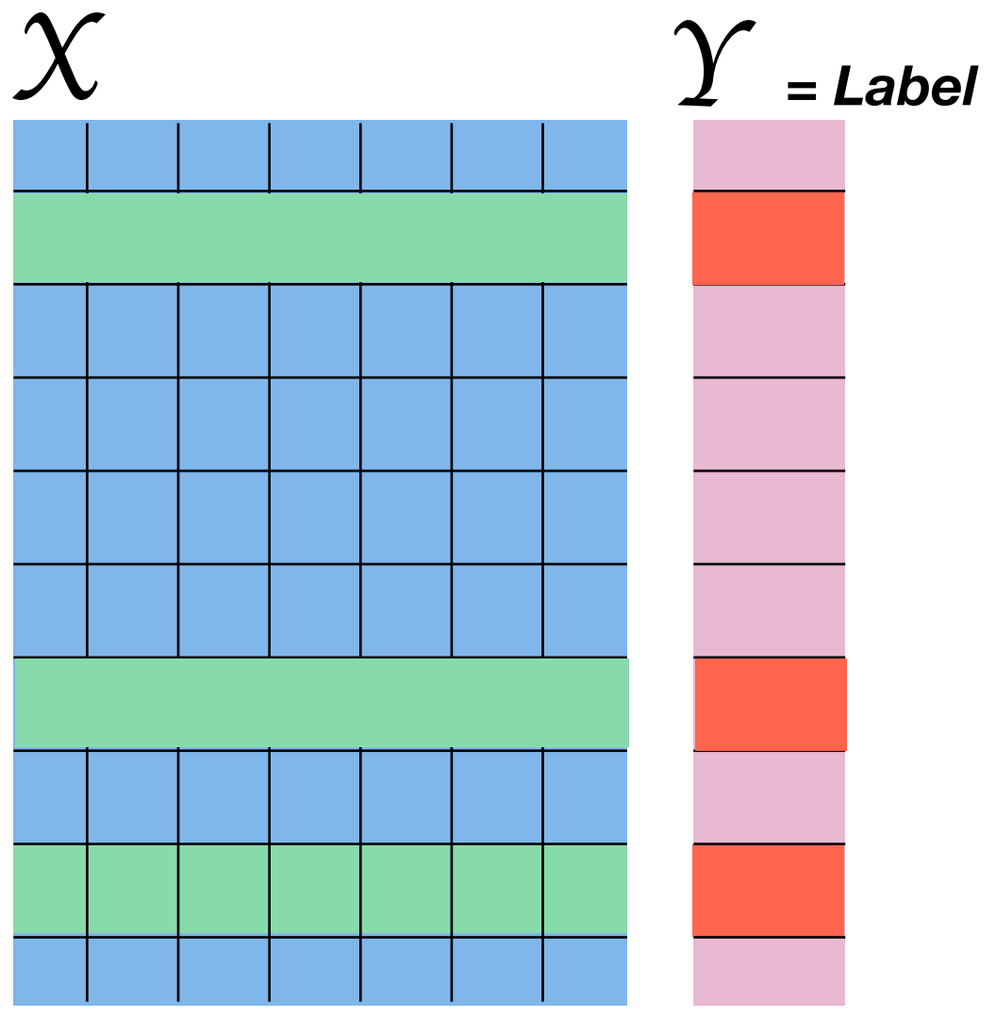
YTest

Randomized split



Random Seed

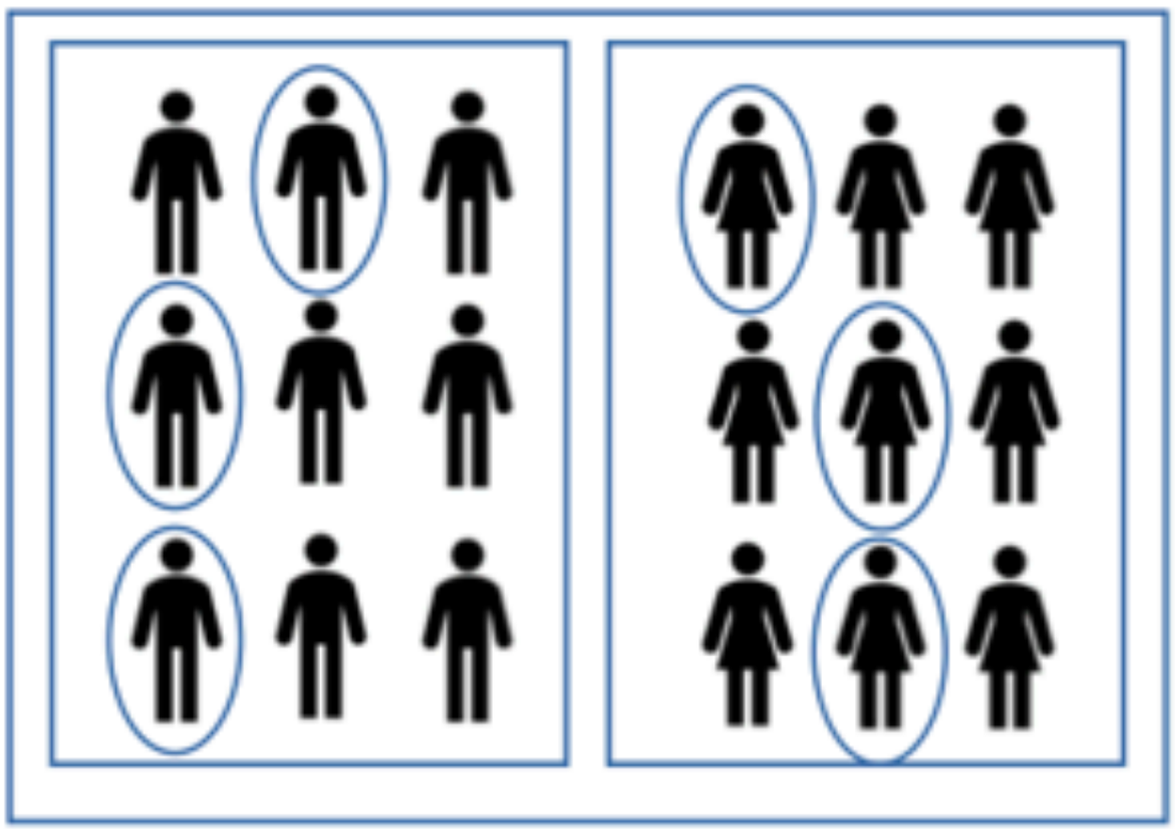




Stratified Split

Normal Sampling 6 of 18

Stratified Sampling 6 of 18



https://dlbjbjzgnk95t.cloudfront.net/1106000/1106385/graphic%201%20-%20srs.png

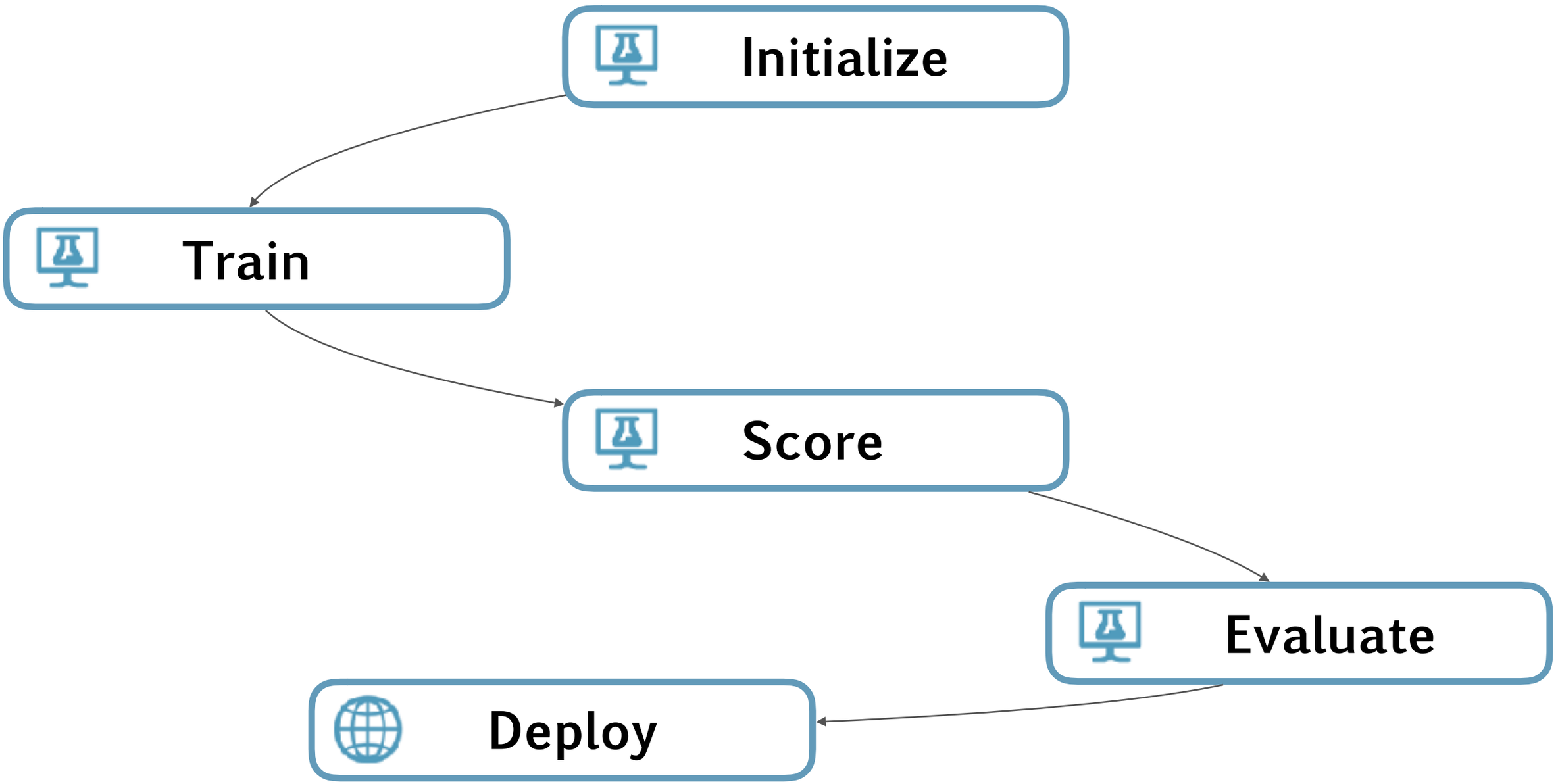
part2. 모델 학습 / 예측

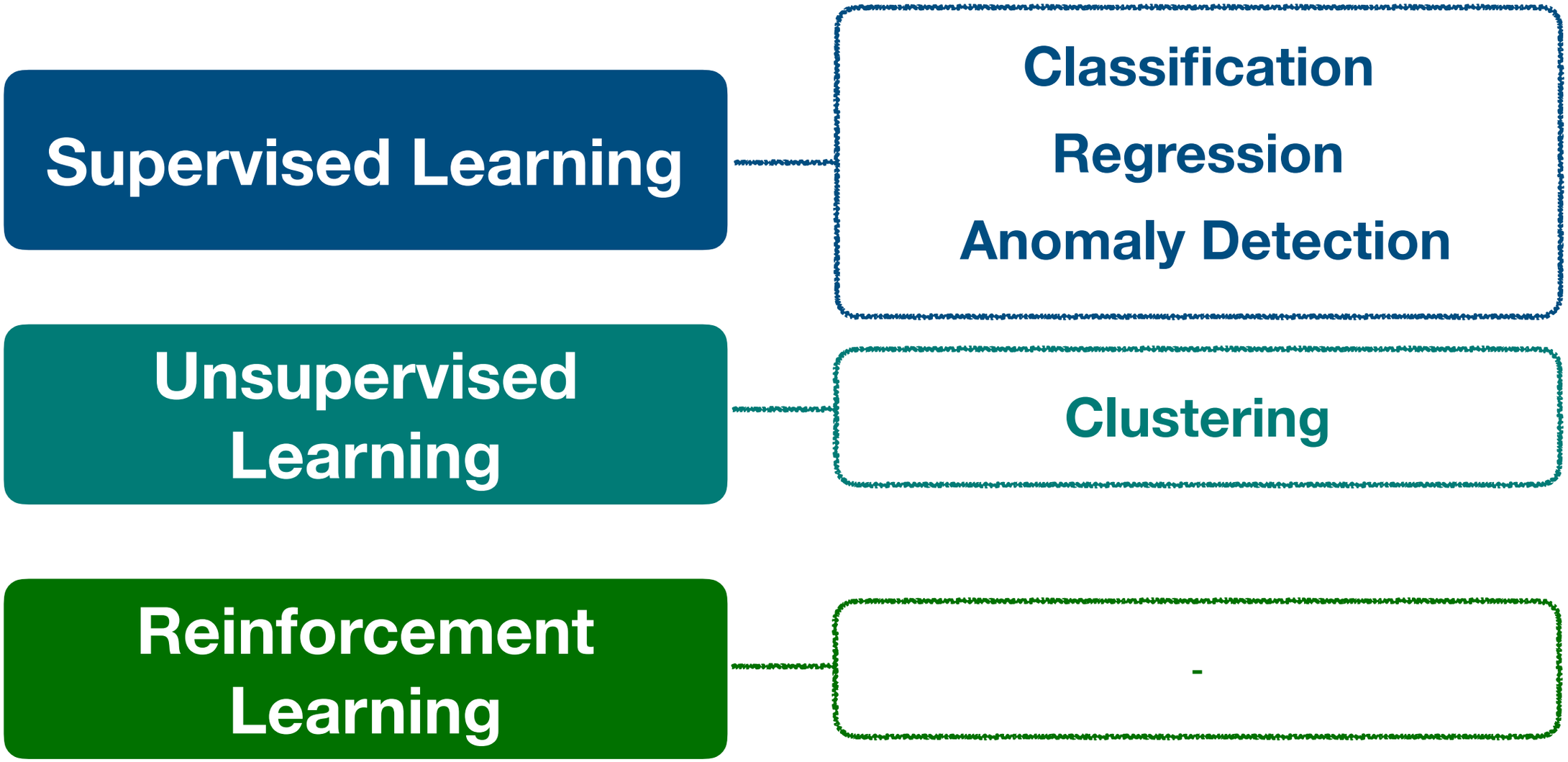
– 다양한 모델 살펴보기

– 올바른 모델 선택

– 모델 학습, 평가

– 모델 배포, 활용


Initialize
Train
Score
Evaluate
Deploy


Supervised Learning
Classification
Regression
Anomaly Detection
Unsupervised Learning
Clustering
Reinforcement Learning
-

Accuracy

Parameter

Training Time

Feature

Linearity

Etc.

https://docs.microsoft.com/ko-kr/azure/machine-learning/studio/algorithm-choice

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points

Discovering structure

Predicting categories

Predicting values

Three or more

Two

ANOMALY DETECTION

- **One-class SVM** — >100 features, aggressive boundary
- **PCA-based anomaly detection** — Fast training

CLUSTERING

- **K-means**

MULTI-CLASS CLASSIFICATION

- Fast training, linear model — **Multiclass logistic regression**
- Accuracy, long training times — **Multiclass neural network**
- Accuracy, fast training — **Multiclass decision forest**
- Accuracy, small memory footprint — **Multiclass decision jungle**
- Depends on the two-class classifier, see notes below — **One-v-all multiclass**

REGRESSION

- **Ordinal regression** — Data in rank ordered categories
- **Poisson regression** — Predicting event counts
- **Fast forest quantile regression** — Predicting a distribution
- **Linear regression** — Fast training, linear model
- **Bayesian linear regression** — Linear model, small data sets
- **Neural network regression** — Accuracy, long training time
- **Decision forest regression** — Accuracy, fast training
- **Boosted decision tree regression** — Accuracy, fast training, large memory footprint

TWO-CLASS CLASSIFICATION

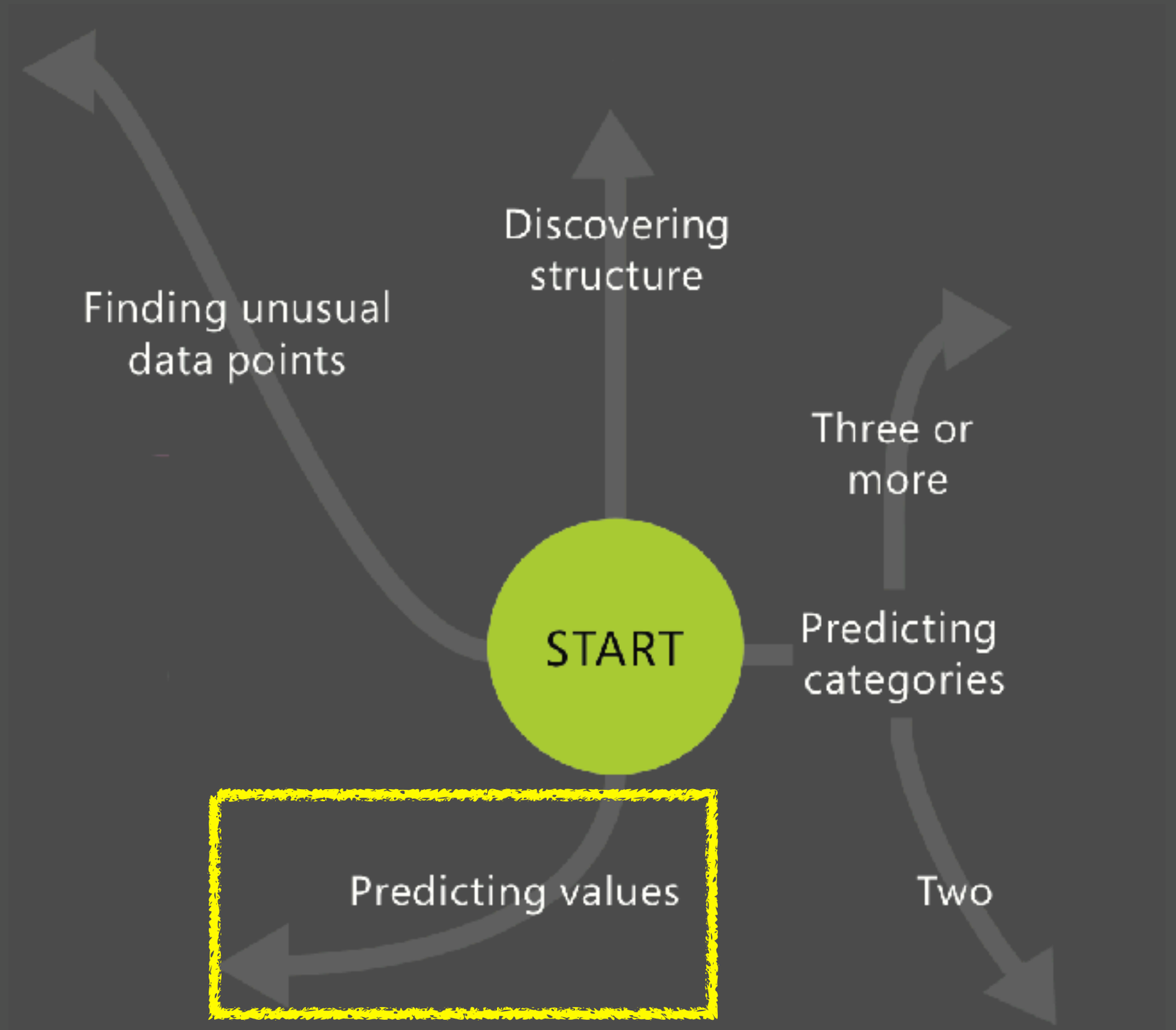
- **Two-class SVM** — >100 features, linear model
- **Two-class averaged perceptron** — Fast training, linear model
- **Two-class logistic regression** — Fast training, linear model
- **Two-class Bayes point machine** — Fast training, linear model

- Accuracy, fast training — **Two-class decision forest**
- Accuracy, fast training, large memory footprint — **Two-class boosted decision tree**
- Accuracy, small memory footprint — **Two-class decision jungle**
- >100 features — **Two-class locally deep SVM**
- Accuracy, long training times — **Two-class neural network**

Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com
Download this poster: http://aka.ms/MLCheatSheet

Microsoft

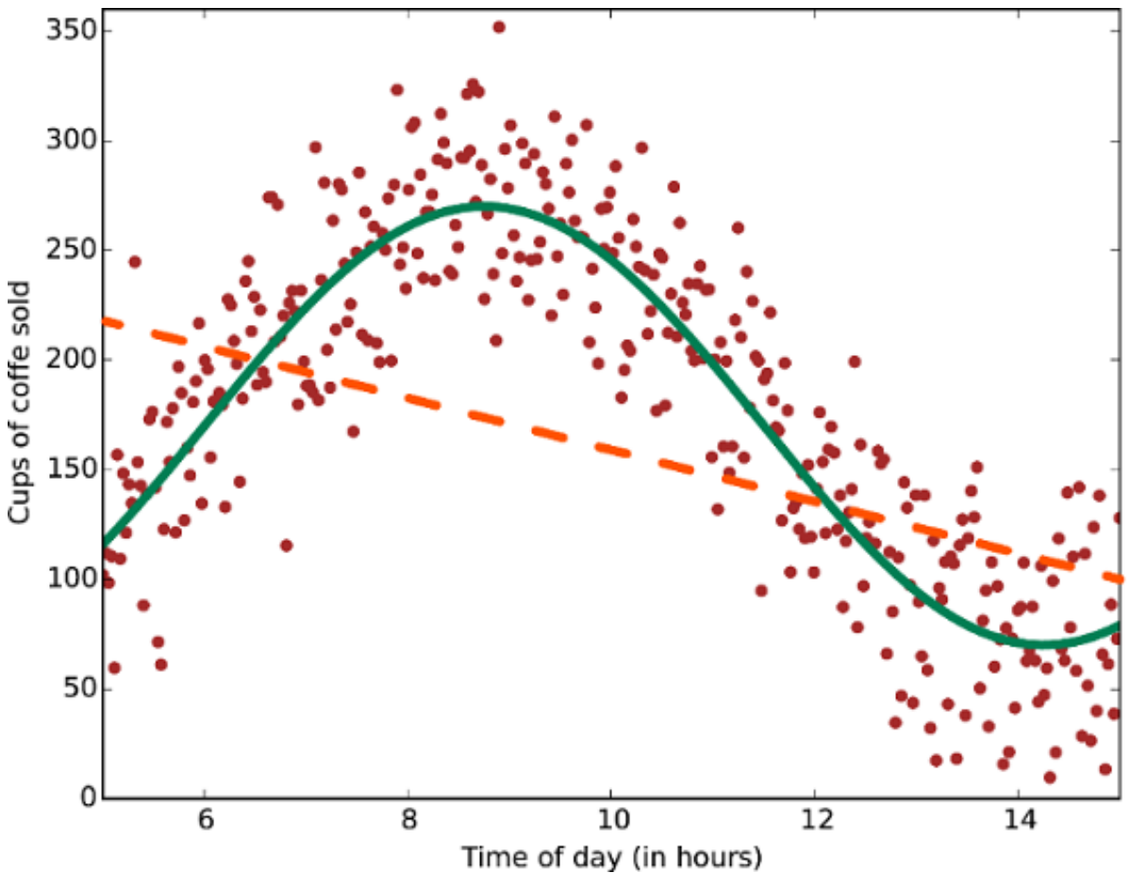
[기업 연계를 위한 AI-IoT 과정] AI | 두번째 날 | 전미정


Finding unusual data points
Discovering structure
Three or more
START
Predicting categories
Predicting values
Two

REGRESSION

Ordinal regression	Data in rank ordered categories
Poisson regression	Predicting event counts
Fast forest quantile regression	Predicting a distribution
Linear regression	Fast training, linear model
Bayesian linear regression	Linear model, small data sets
Neural network regression	Accuracy, long training time
Decision forest regression	Accuracy, fast training
Boosted decision tree regression	Accuracy, fast training, large memory footprint

Regression



- 키에 따른 신발 사이즈
- 시간에 따른 커피 소비량
- 햇빛 노출 시간과 주근깨 개수
- 달 위상에 따른 주요 도시의 범죄 수
- 기온과 인터넷 쇼핑 장바구니 물품 수

Regression

Ordinal Regression	데이터 내 상대적 순서나 랭킹 예측 ex) 강연 참석자의 선호도, URL 즐겨찾기 순서	0
Poisson Regression	어떤 이벤트가 발생할 횟수 예측 이산분포를 따르며 음의 정수값 X ex) 비행기 탑승에 따른 병원 방문 횟수	5
Fast Forest Quantile Regression	값의 분산/분포 예측 ex) 성적 예측를 통한 학생들의 발달 단계 평가	9
Linear Regression	가장 일반적인 선형 회귀 알고리즘	4

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-regression

Regression

Bayesian Linear Regression	Bayesian 접근법을 선형회귀에 적용	2
Neural Network Regression	신경망 회로(DNN), 비선형 문제에 활용 Customizable algorithm	9
Decision Forest Regression	의사 결정 트리, 비선형 문제에 활용 효율적인 메모리 사용 및 계산 (overfitting 주의)	6
Boosted Decision Tree Regression	이전 트리에 종속되어 있어 메모리 사용이 큼 정확도가 높음, 앙상블 모델에 활용	5

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-regression

REGRESSION

Ordinal regression	Data in rank ordered categories
Poisson regression	Predicting event counts
Fast forest quantile regression	Predicting a distribution
Linear regression	Fast training, linear model
Bayesian linear regression	Linear model, small data sets
Neural network regression	Accuracy, long training time
Decision forest regression	Accuracy, fast training
Boosted decision tree regression	Accuracy, fast training, large memory footprint

Metrics

Mean Absolute Error	3197.596091
Root Mean Squared Error	5469.159134
Relative Absolute Error	0.211708
Relative Squared Error	0.058495
Coefficient of Determination	0.941505

Regression

p = predicted value
a = actual value

Mean Absolute Error(MAE)	$MAE = \frac{\sum_{i=1}^{n} \lvert p_i - a_i \rvert}{n}$
Root Mean Squared Error(RMSE)	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}}$
Coefficient of Determination	R^2

Relative Absolute Error(RAE)	$RAE = \frac{\sum_{i=1}^{n} \lvert p_i - a_i \rvert}{\sum_{i=1}^{n} \lvert \overline{a} - a_i \rvert}$
Relative Squared Error(RSE)	$RSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i-1}^{n} (\overline{a} - a_i)^2}$

Index	Error	| Error |	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4

MAE(Mean Absolute Error)

$$\frac{\Sigma|p-a|}{n}$$

$$\frac{2+2+2+2+2}{5}=2.0$$

MSE(Mean Square Error)

$$\sqrt{\frac{\Sigma(p-a)^2}{n}}$$

$$\sqrt{\frac{4+4+4+4+4}{5}}=2.0$$

Index	Error	| Error |	Error^2
1	0	0	0
2	1	1	1
3	3	3	9
4	3	3	9
5	3	3	9

MAE(Mean Absolute Error)

$$\frac{\Sigma|p-a|}{n}$$

$$\frac{0+1+3+3+3}{5}=2.0$$

MSE(Mean Square Error)

$$\sqrt{\frac{\Sigma(p-a)^2}{n}}$$

$$\sqrt{\frac{0+1+9+9+9}{5}}=2.4$$

Index	Error	| Error |	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	10	10	100

MAE(Mean Absolute Error)

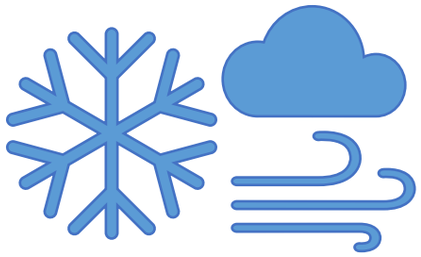
$$\frac{\Sigma|p-a|}{n}$$

$$\frac{0+0+0+0+10}{5}=2.0$$

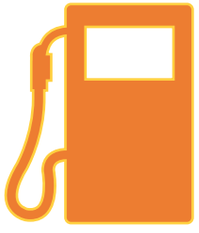
MSE(Mean Square Error)

$$\sqrt{\frac{\Sigma(p-a)^2}{n}}$$

$$\sqrt{\frac{0+0+0+0+100}{5}}=4.5$$



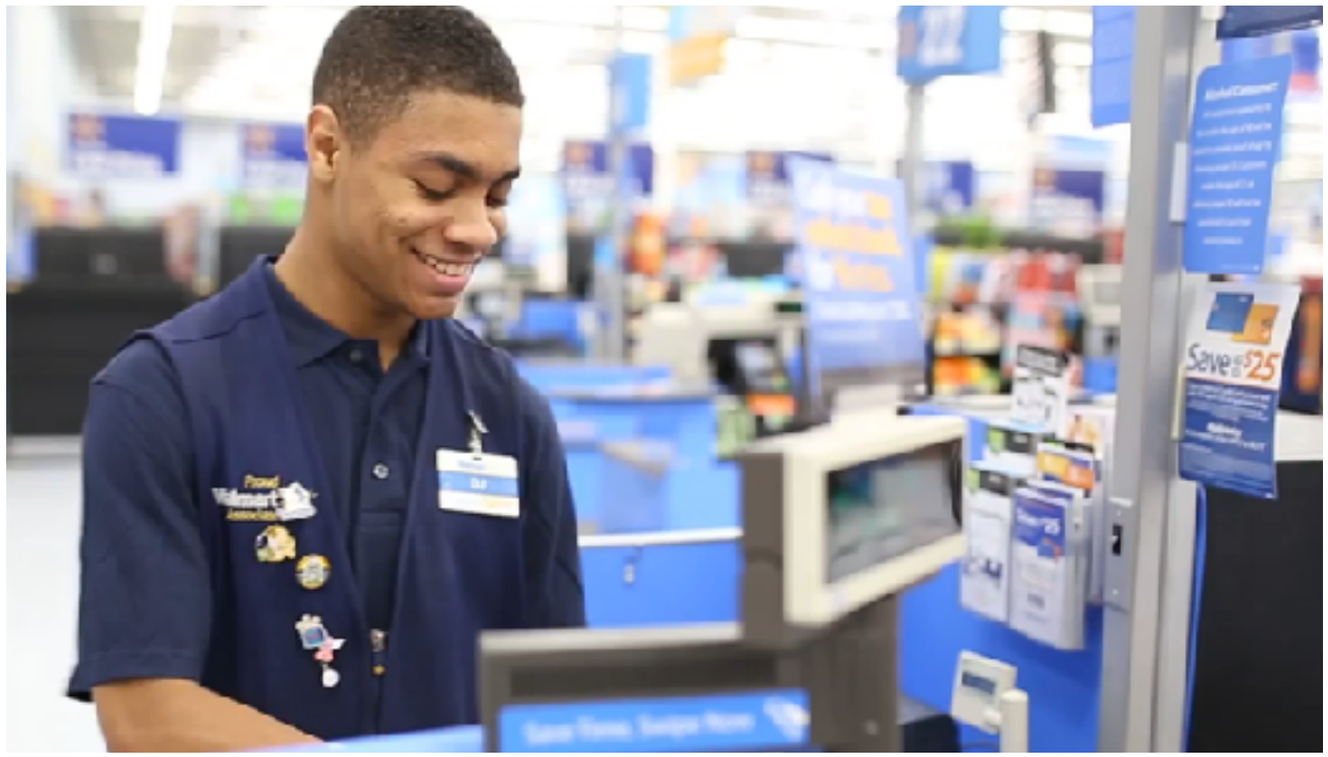
Weather Forecast API

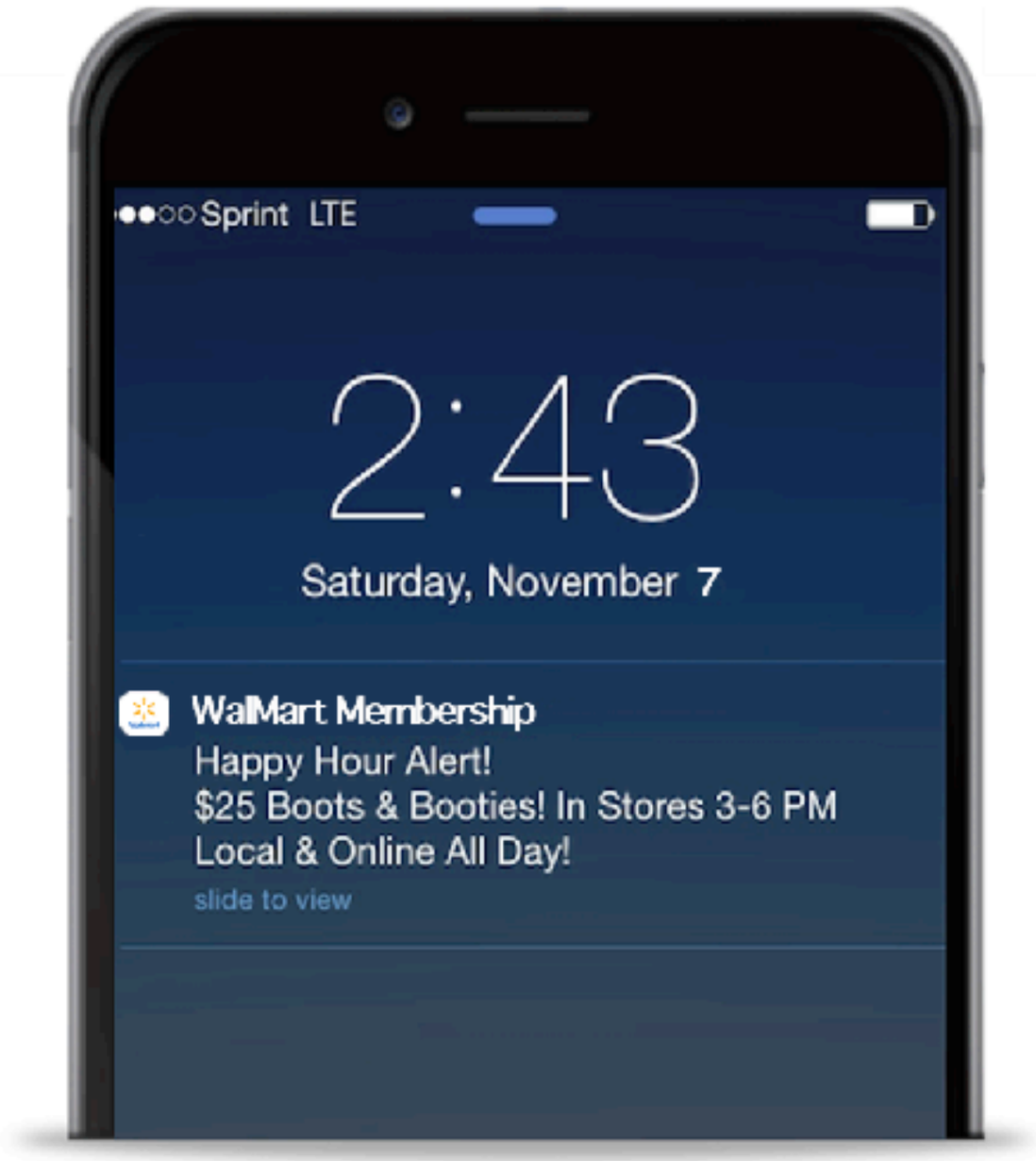


Fuel Price Forecast API

	A	B	C	D	E	F	G	H	I
1	Store	Date	Temperature	Fuel_Price	Dept	Weekly_Sales	IsHoliday	Type	Size
2	1	2010-02	42.31	2.572	1	24924.5	FALSE	A	151315
3	1	2010-02	43.31	2.813	2	50605.27	FALSE	A	151315
4	1	2010-02	42.31	2.572	3	13740.12	FALSE	A	151315
5	1	2010-02	42.31	2.572	4	39954.04	FALSE	A	151315
6	1	2010-02	42.31	2.572	5	32229.38	FALSE	A	151315
7	1	2010-02	43.31	2.813	1	24924.5	FALSE	A	151315
8	1	2010-02	42.31	2.572	2	50605.27	FALSE	A	151315
9	1	2010-02	43.31	2.813	3	13740.12	FALSE	A	151315
10	1	2010-02	42.31	2.572	4	39954.04	FALSE	A	151315
11	1	2010-02	43.31	2.572	5	32229.38	FALSE	A	151315
12	1	2010-02	42.31	2.813	1	24924.5	FALSE	A	151315
13	1	2010-02	43.31	2.572	2	50605.27	FALSE	A	151315
14	1	2010-02	42.31	2.813	3	13740.12	FALSE	A	151315
15	1	2010-02	43.31	2.572	4	39954.04	FALSE	A	151315
16	1	2010-02	43.31	2.813	5	32229.38	FALSE	A	151315

Storage(Azure Blob)

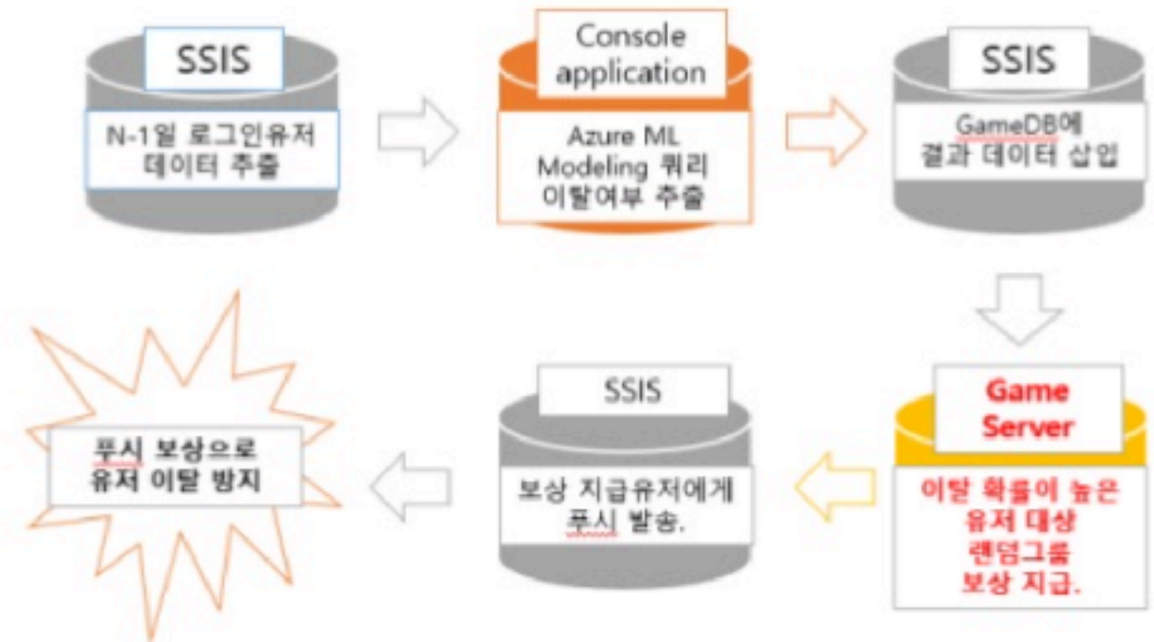



Sprint LTE
2:43
Saturday, November 7
WalMart Membership
Happy Hour Alert!
$25 Boots & Booties! In Stores 3-6 PM
Local & Online All Day!
slide to view

https://www.slideshare.net/devcatpublications/ss-75457149

실제 게임에 적용하기

- 지금까지 Azure ML 과 마비노기 듀얼 접속 데이터를 활용하여,
  유저 재방문 유도 시스템 모델링을 진행하였고, 이를 활용하여 실제 게임에 적용.



https://www.slideshare.net/devcatpublications/ss-75457149

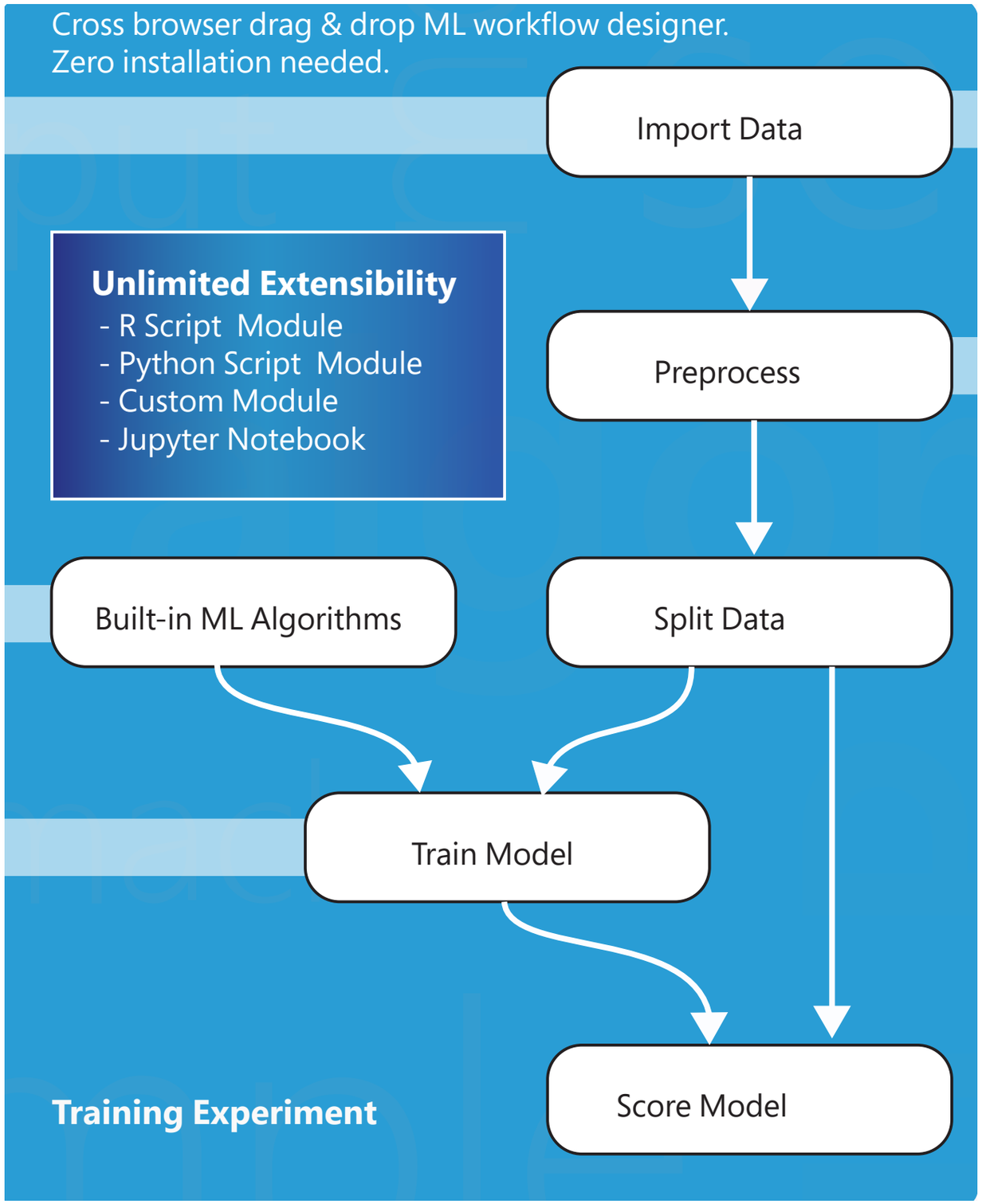
문제 정의

데이터 셋 준비

모델 설정

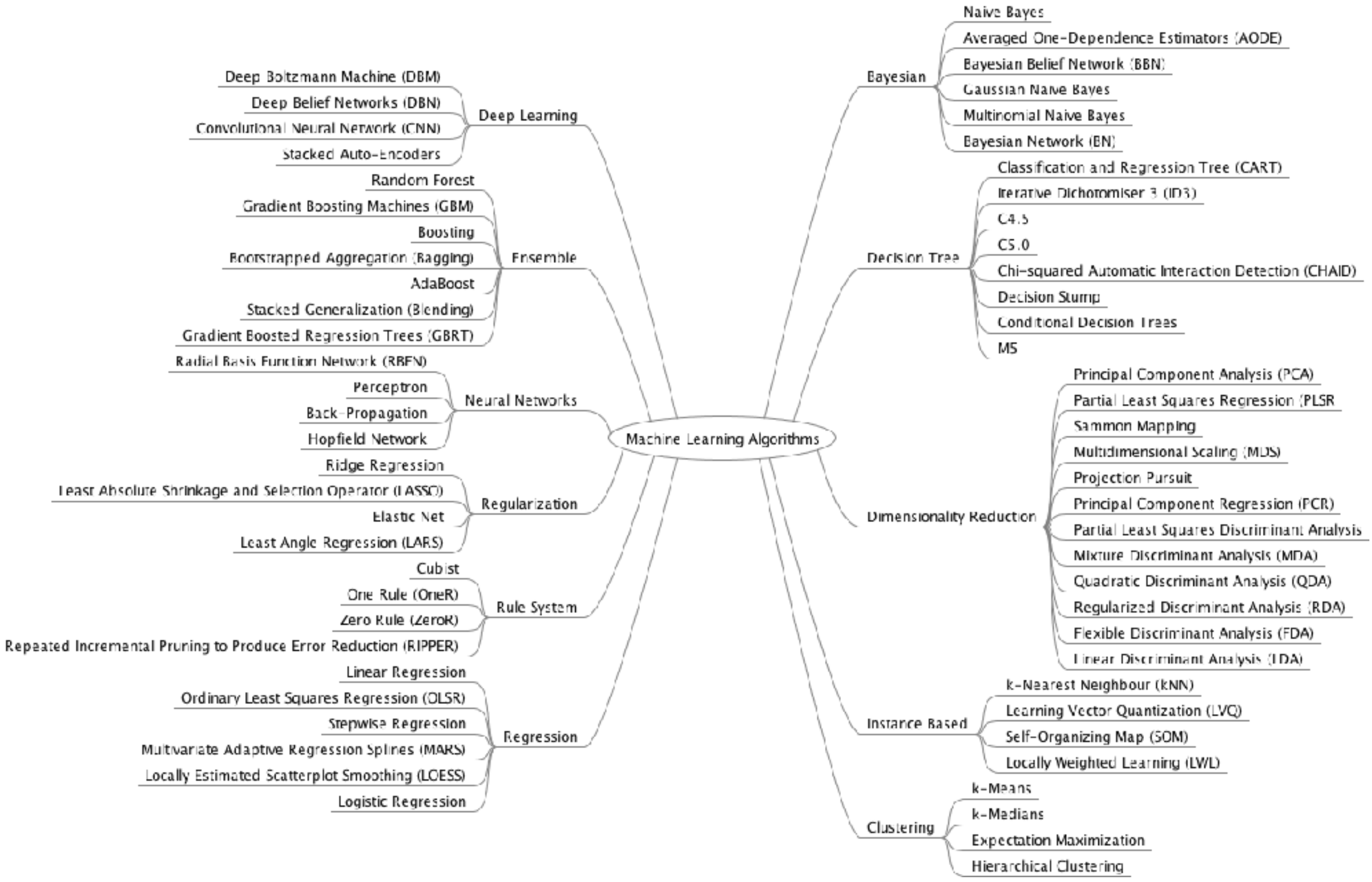
모델 훈련 / 평가

모델 활용



두 번째 날

- Regression 모델 생성
- 머신러닝 알고리즘
- Scikit-learn 모델 생성

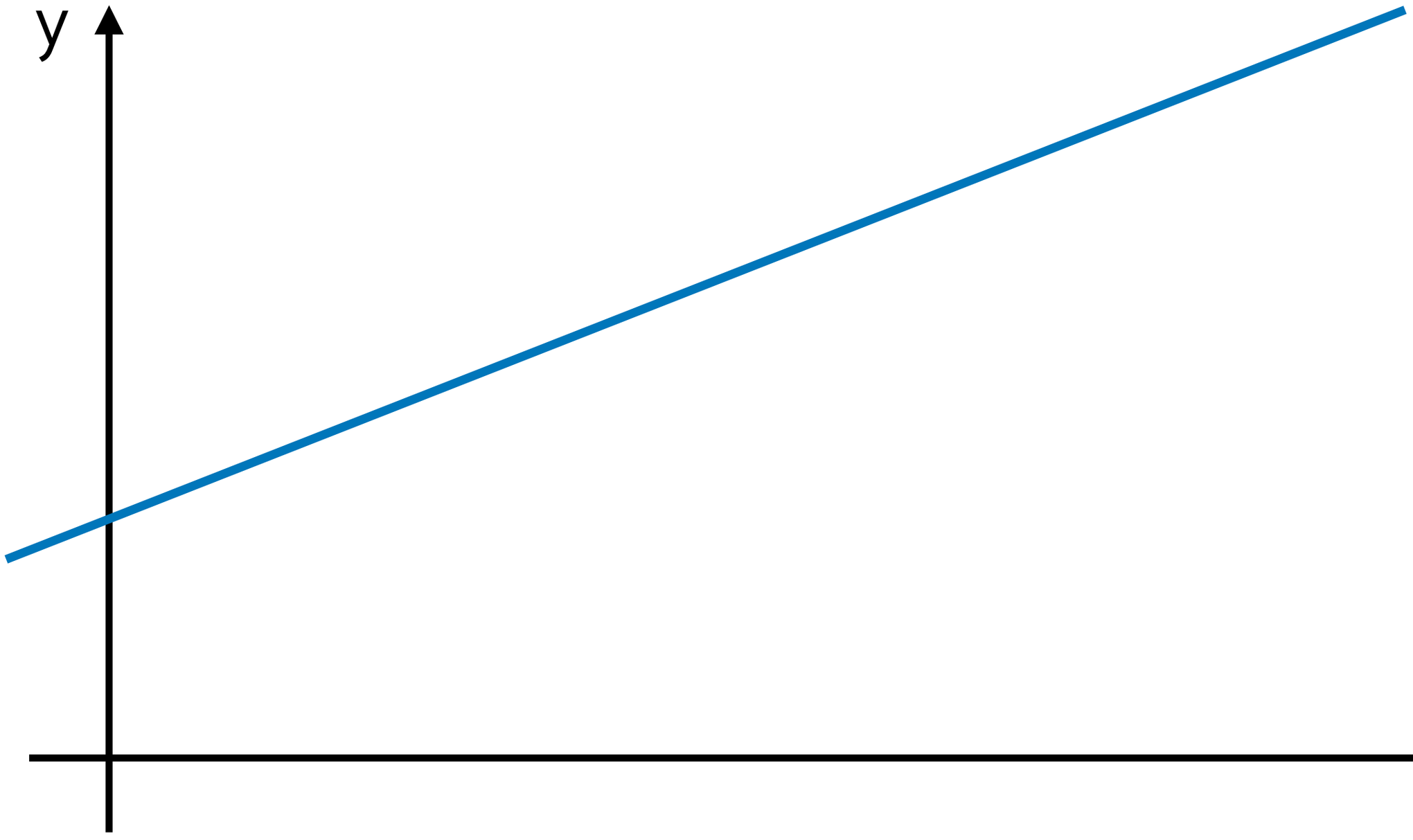


https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Regression

Simple Regression

$$y = ax + b$$



Years	Salary
1.5	3,100
2.5	3,900
4.2	4,300
4.9	4,600
5.1	4,900
6.7	5,400
8.3	6,700
9.5	9,200
13.0	12,900

$$y = ax + b$$

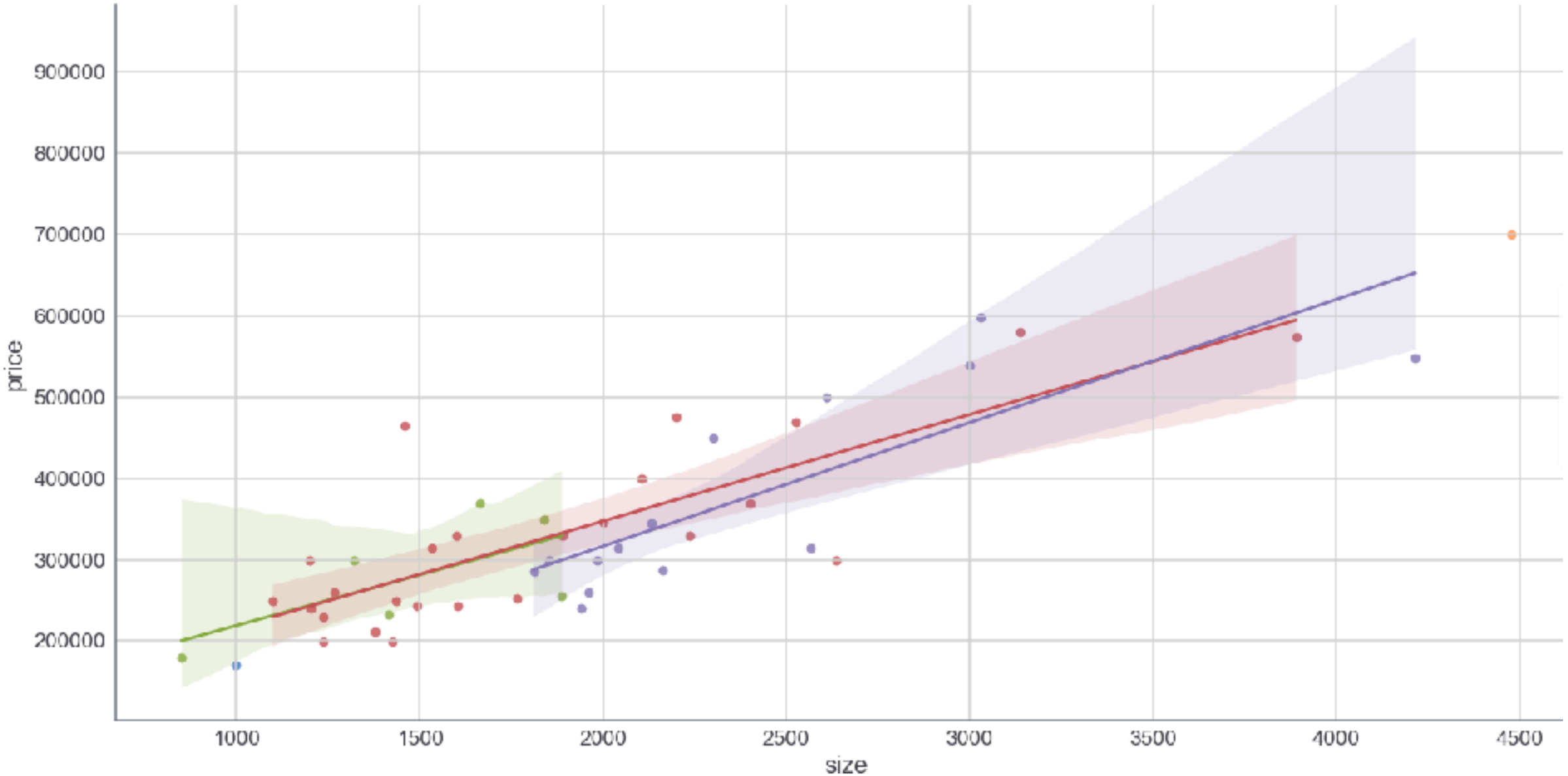
Multiple Linear Regression

Simple Linear Regression

$$y = ax + b$$

Multiple Linear Regression

$$y = k + ax_0 + bx_1 + cx_2 + \ldots + nx_n$$



https://medium.com/we-are-orb/multivariate-linear-regression-in-python-without-scikit-learn-7091b1d45905

Years	Age	Dept.	Salary
1.5	26	영업	3,100
2.5	26	개발	3,900
4.2	28	홍보	4,300
4.9	32	개발	4,600
5.1	31	지원	4,900
6.7	35	인사	5,400
8.3	39	디자인	6,700
9.5	41	홍보	9,200
13.0	46	관리	12,900

$$y = k + ax_0 + bx_1 + cx_2$$

Polynominal Regression

Simple Linear Regression

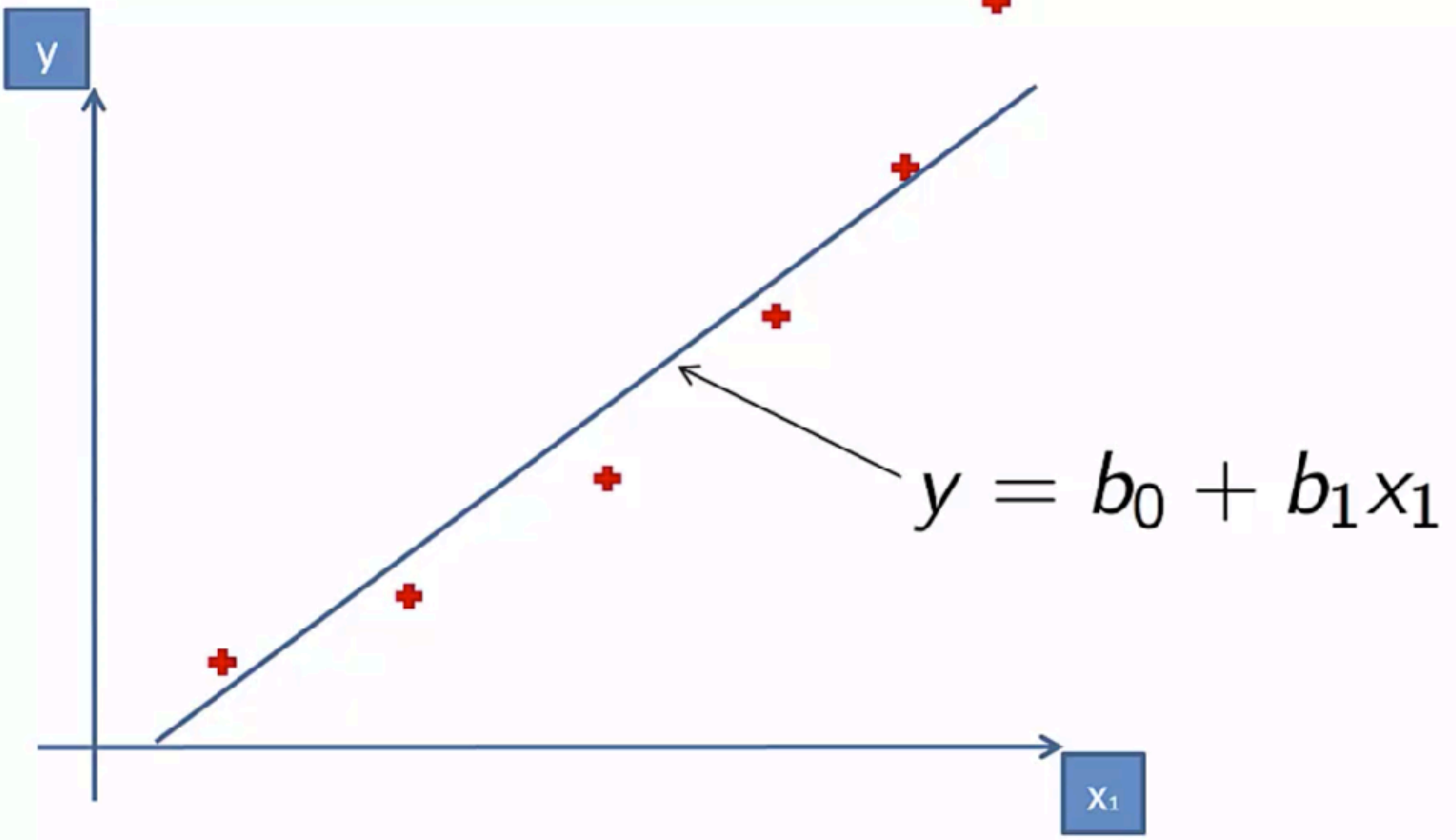
$$y = ax + b$$

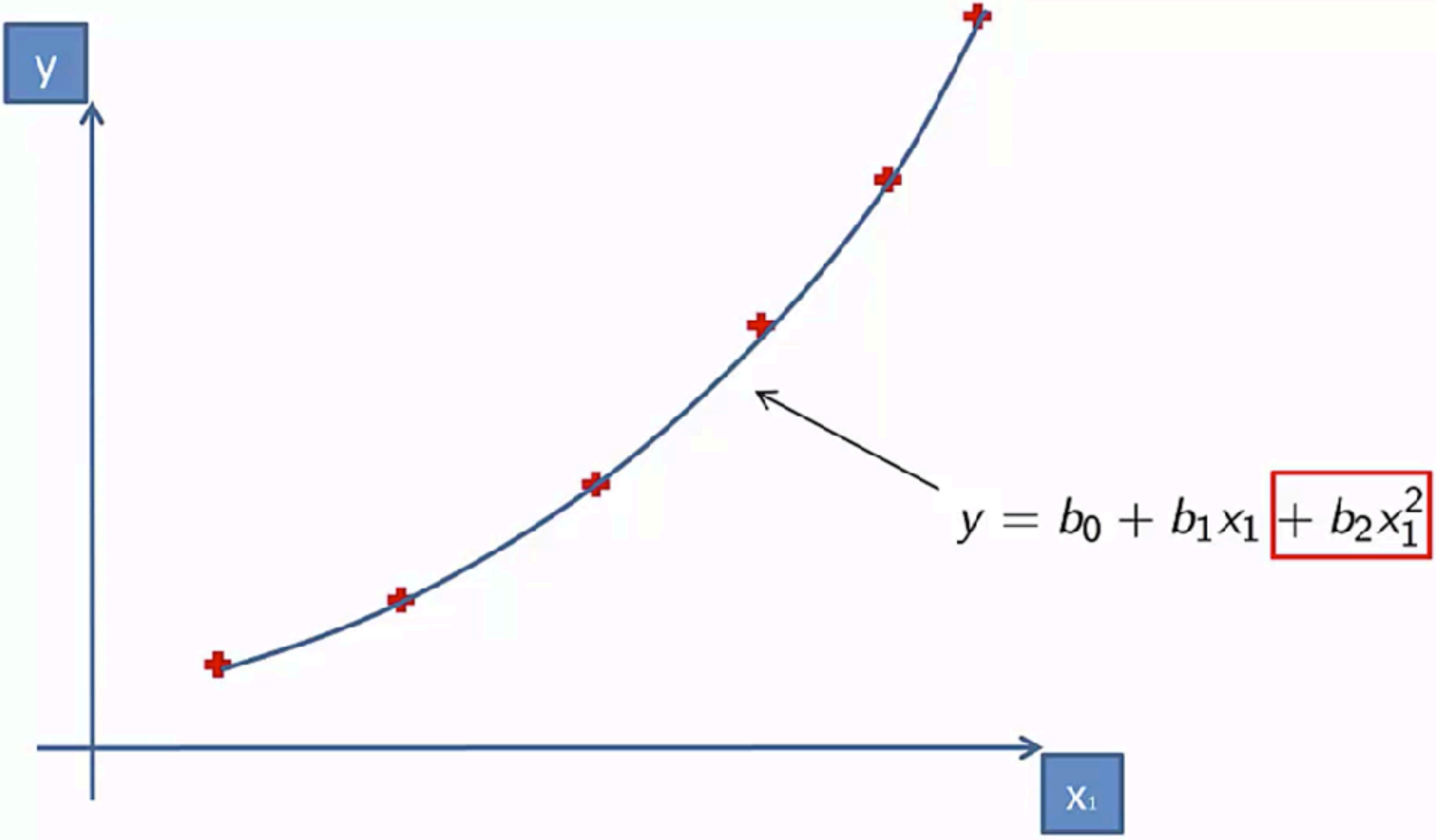
Multiple Linear Regression

$$y = k + ax_0 + bx_1 + cx_2 + \ldots + nx_n$$

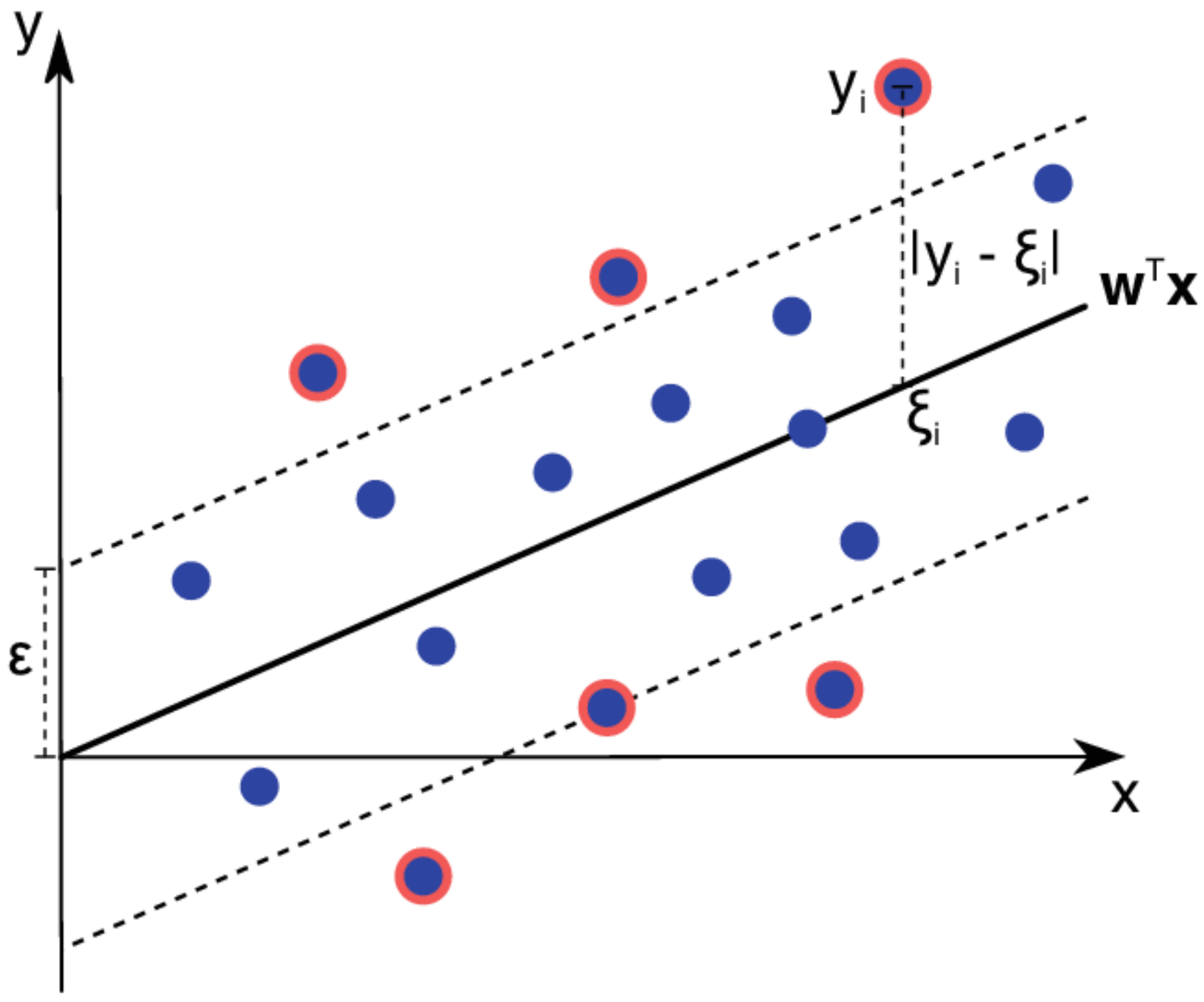
Polygonal Regression

$$y = k + ax + bx^2 + cx^3 + \ldots + nx^n$$


y
$y = b_0 + b_1 x_1$
x_1

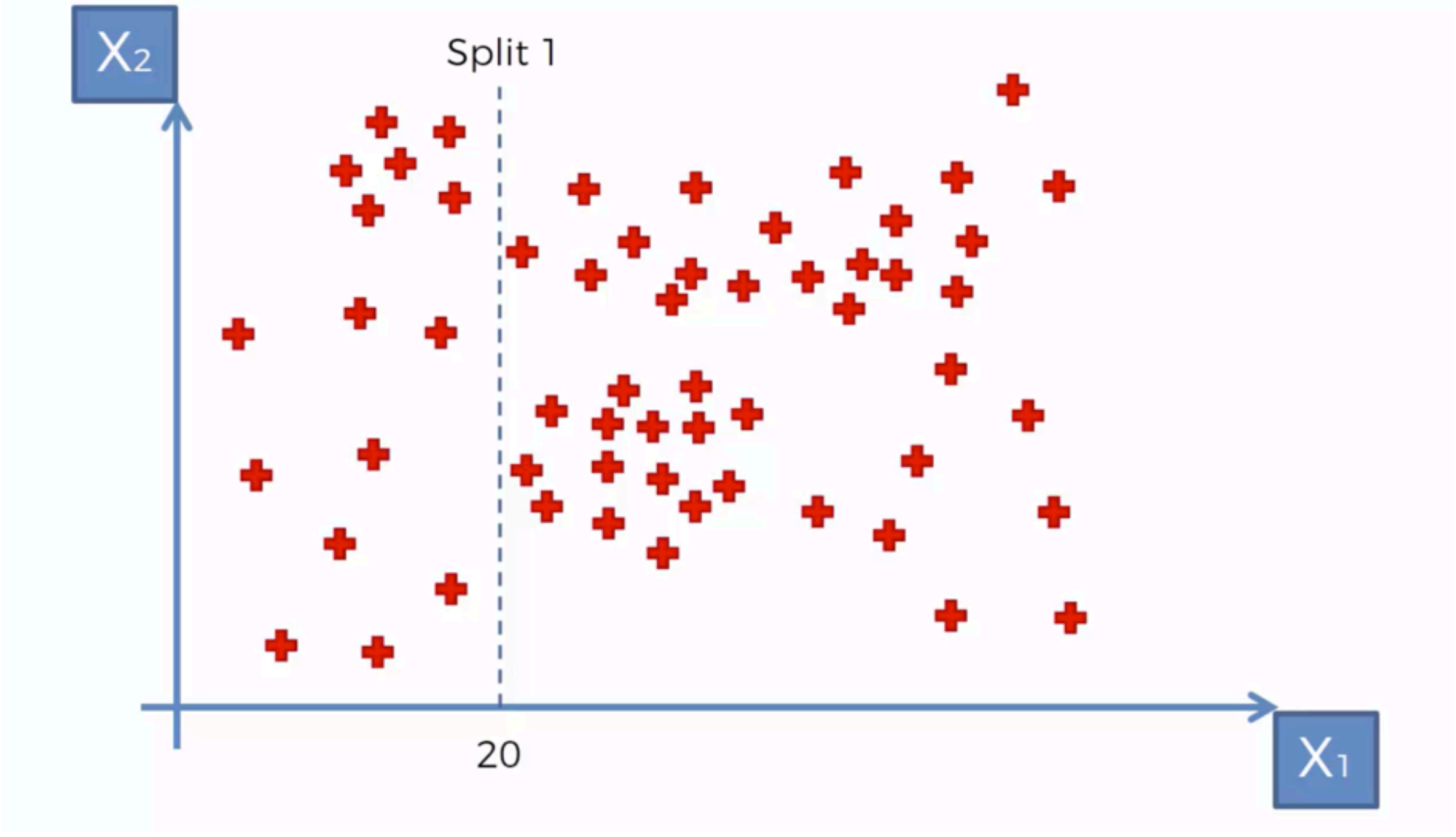

y
$y = b_0 + b_1 x_1 + b_2 x_1^2$
x_1

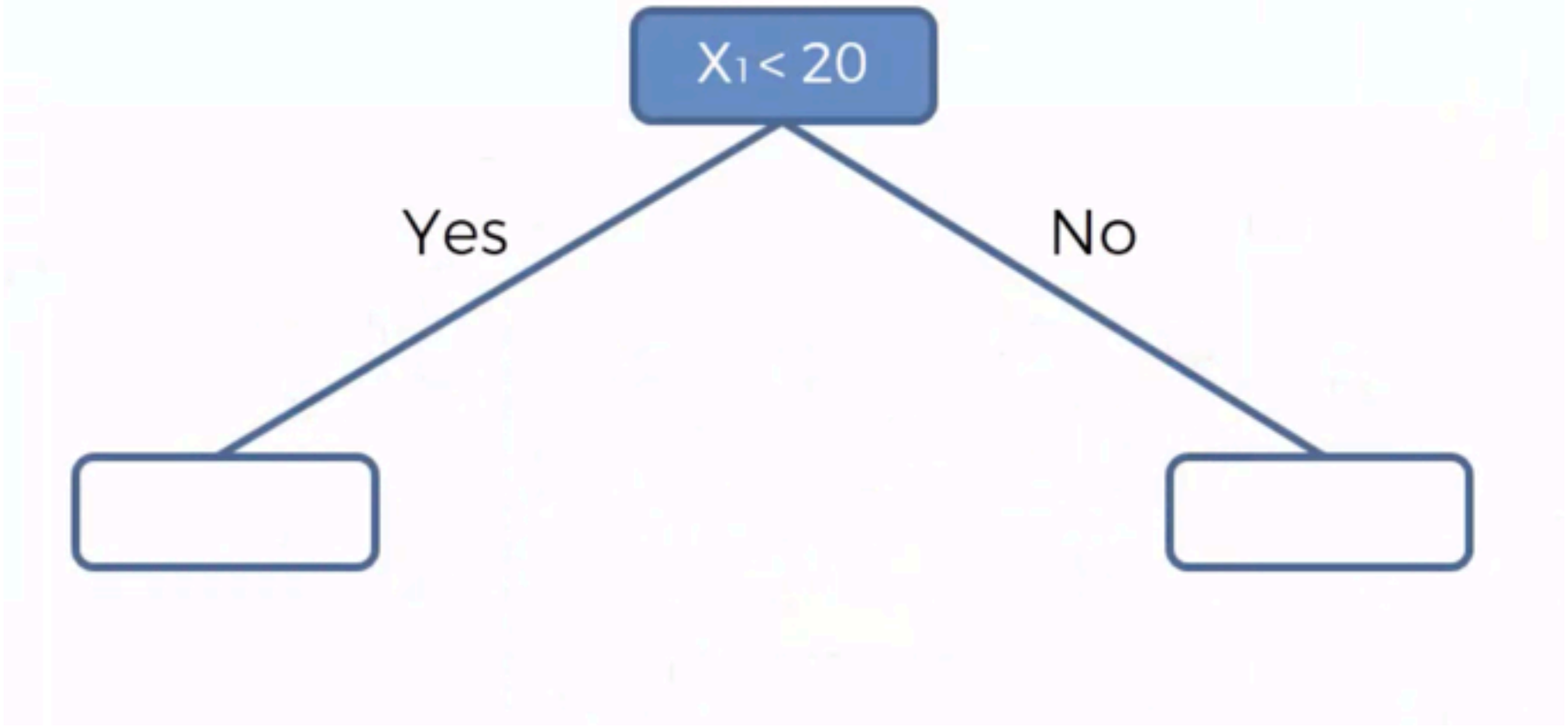
SVR
(Supported Vector Regression)

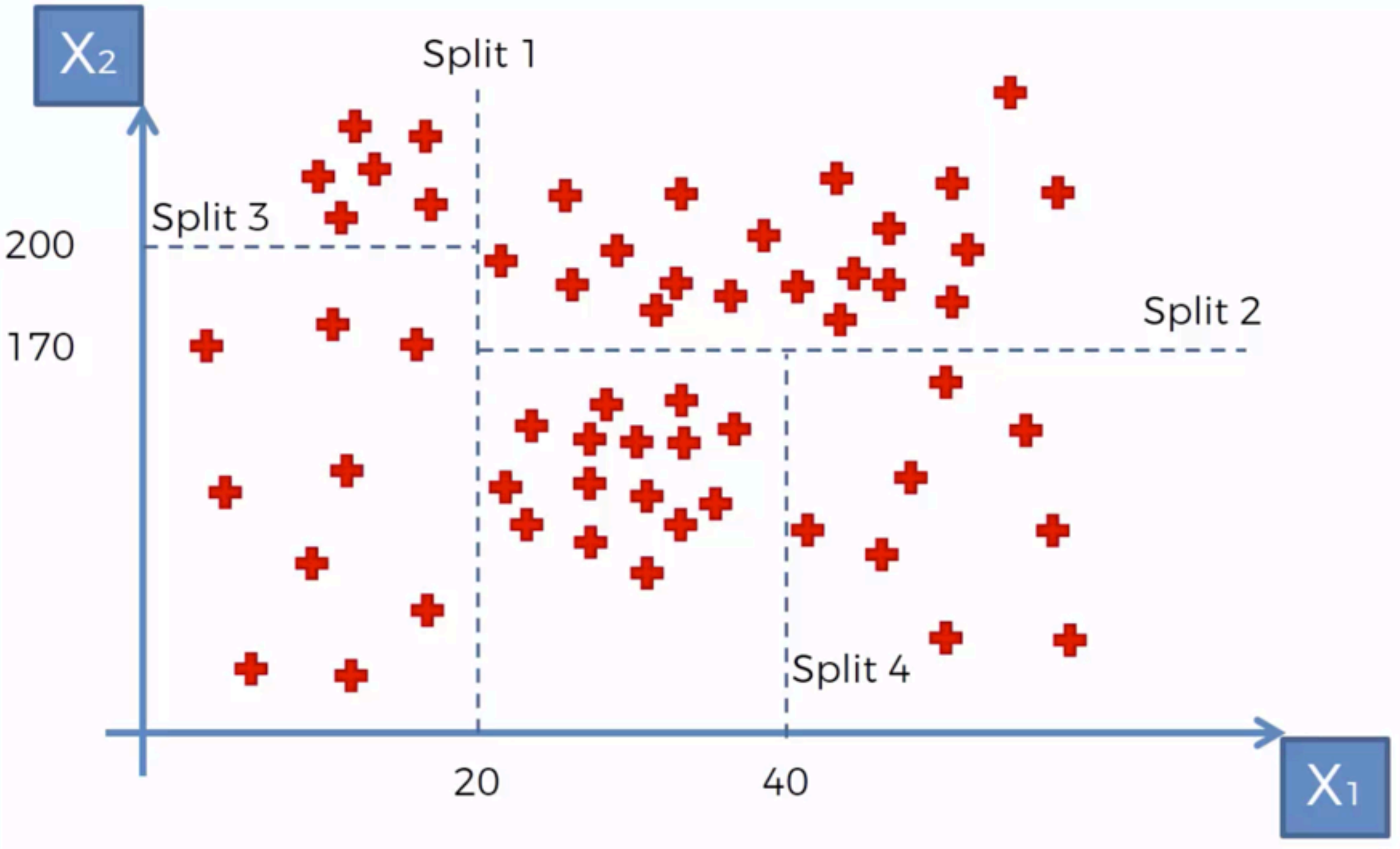


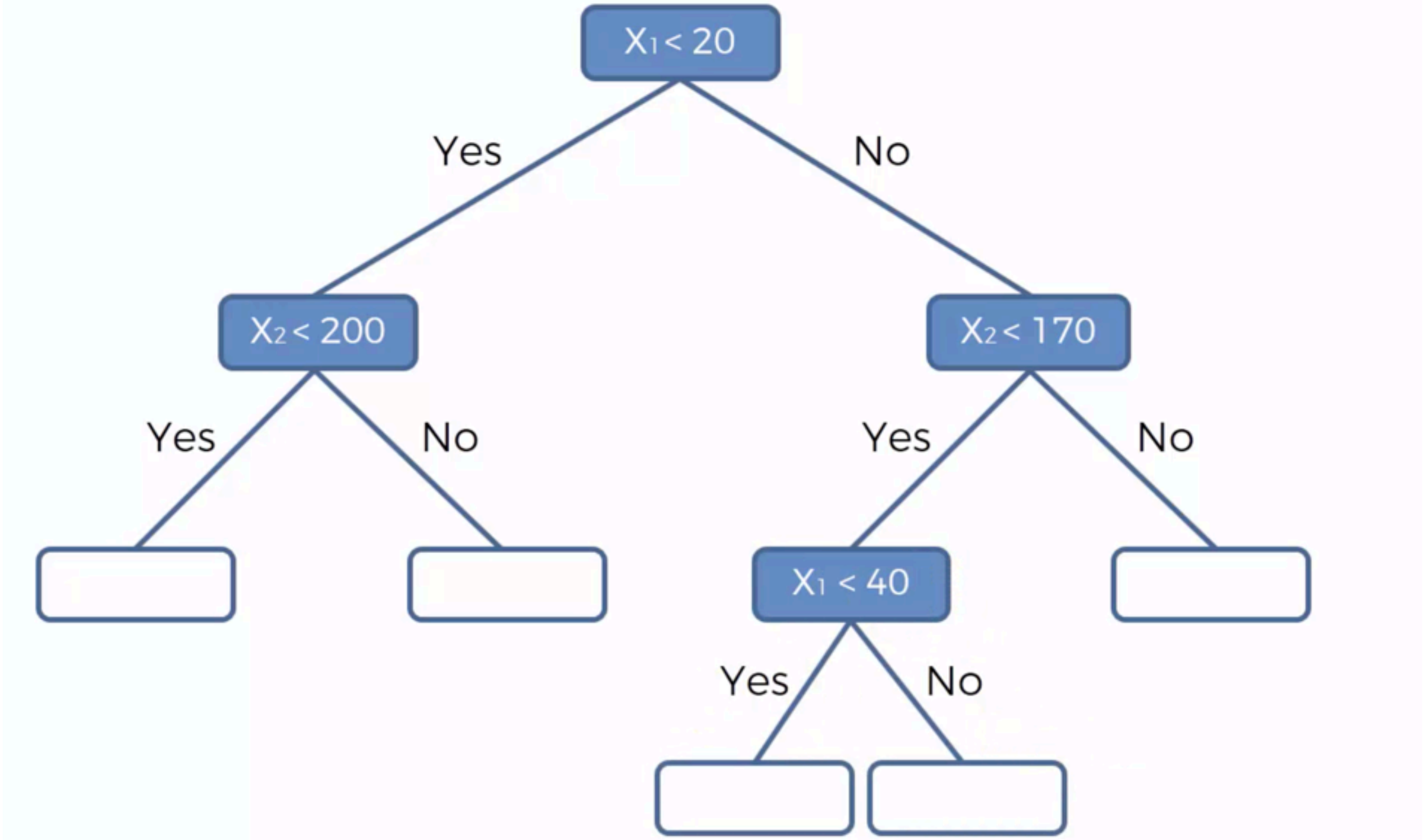
https://www.researchgate.net/profile/Frank_Boeckler/publication

Decision Tree


X2
Split 1
20
X1

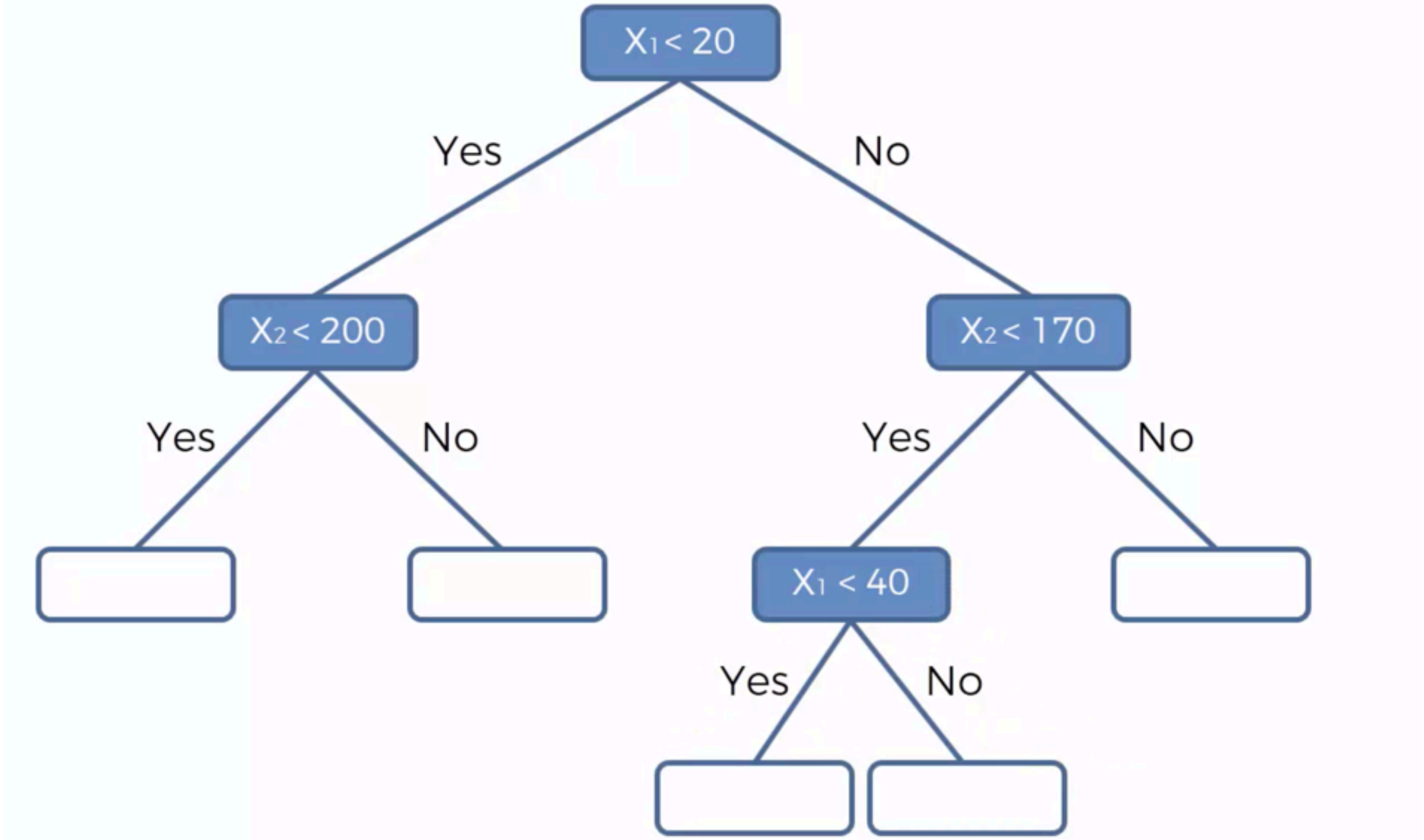

X1 < 20
Yes
No

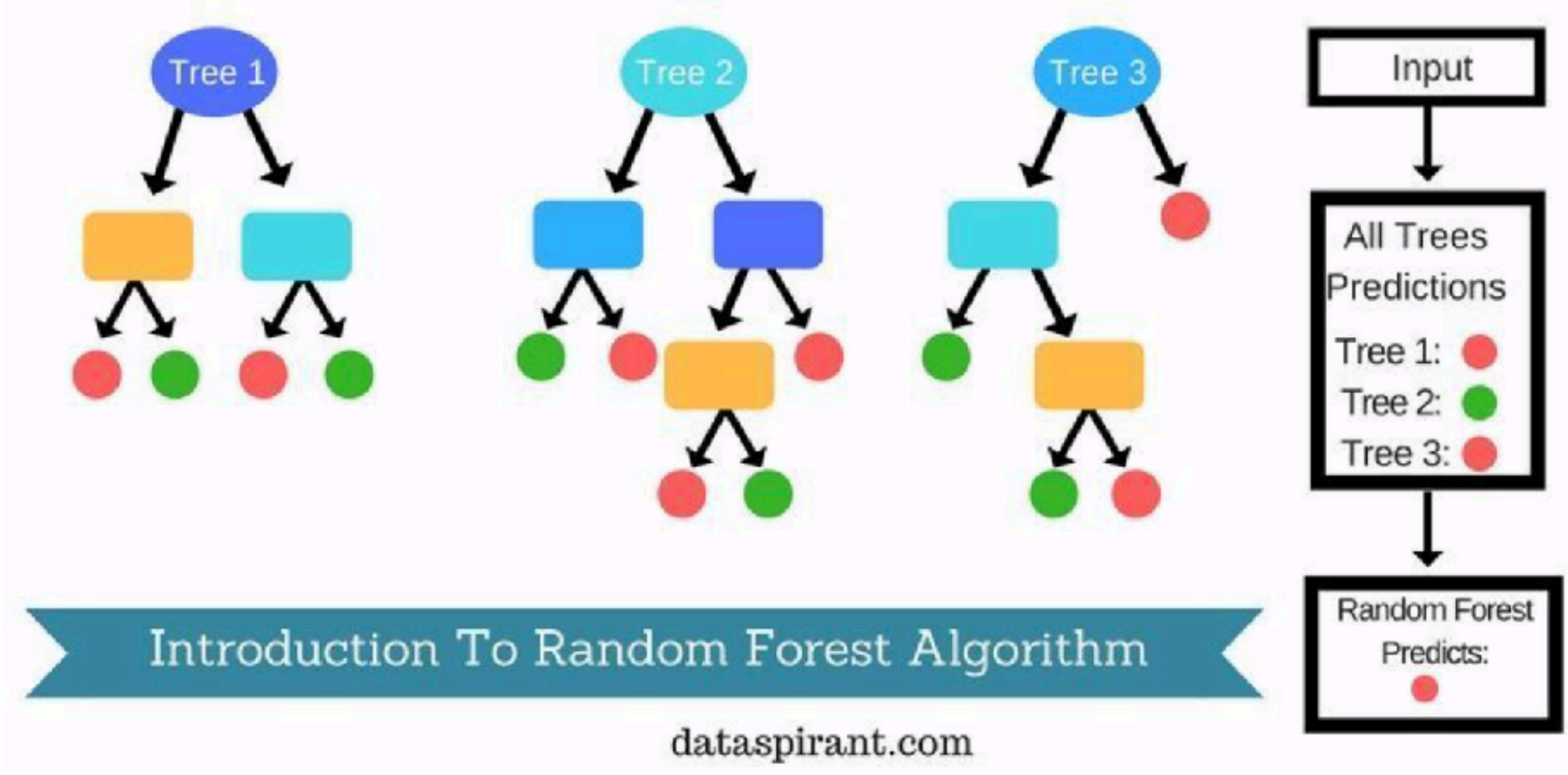

X2
Split 1
Split 3
200
170
Split 2
Split 4
20
40
X1


X1 < 20
Yes
No
X2 < 200
X2 < 170
Yes
No
Yes
No
X1 < 40
Yes
No

Random Forest

[기업 연계를 위한 AI-IoT 과정] AI | 두번째 날 | 전미정


X1 < 20
Yes
No
X2 < 200
X2 < 170
Yes
No
Yes
No
X1 < 40
Yes
No



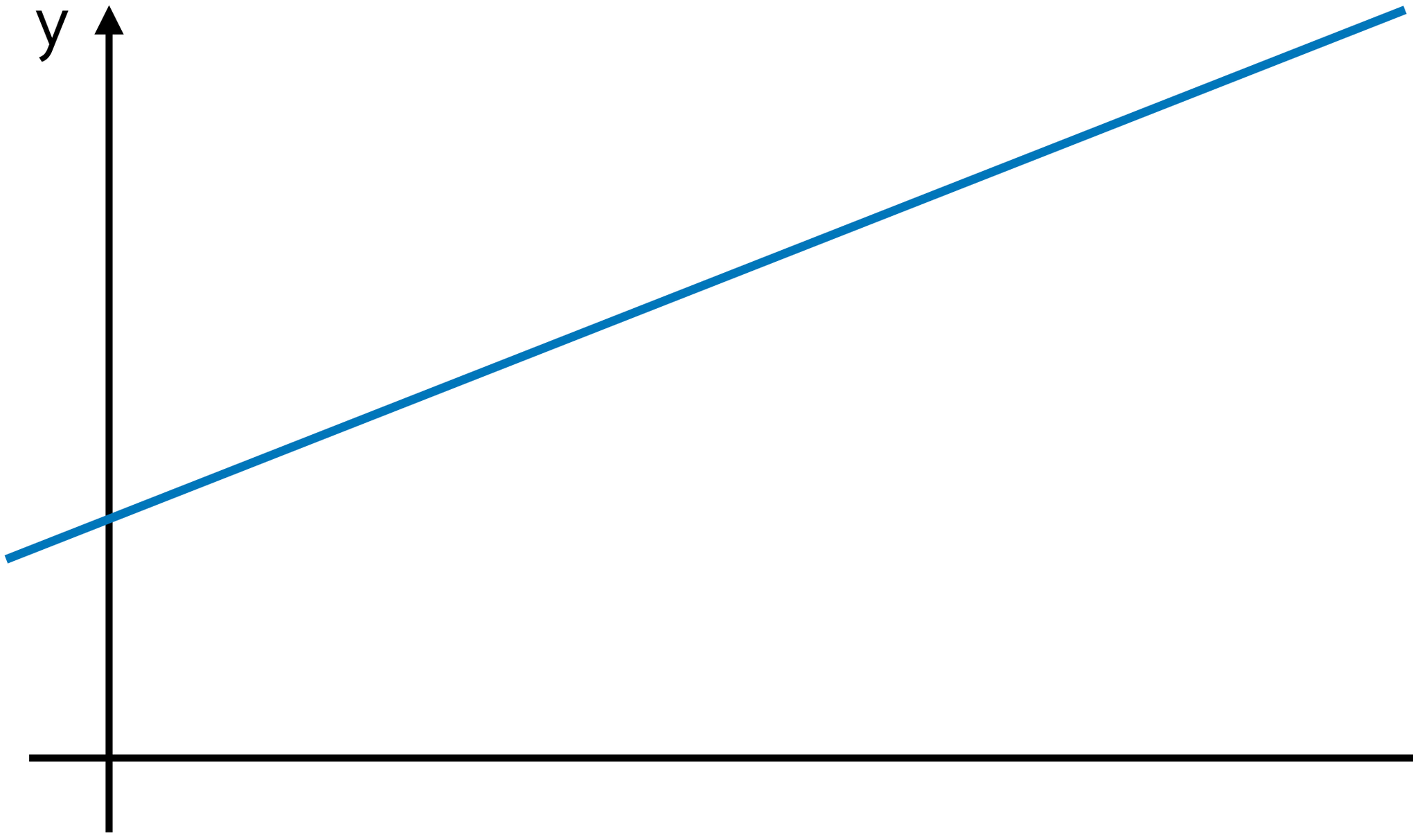
https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

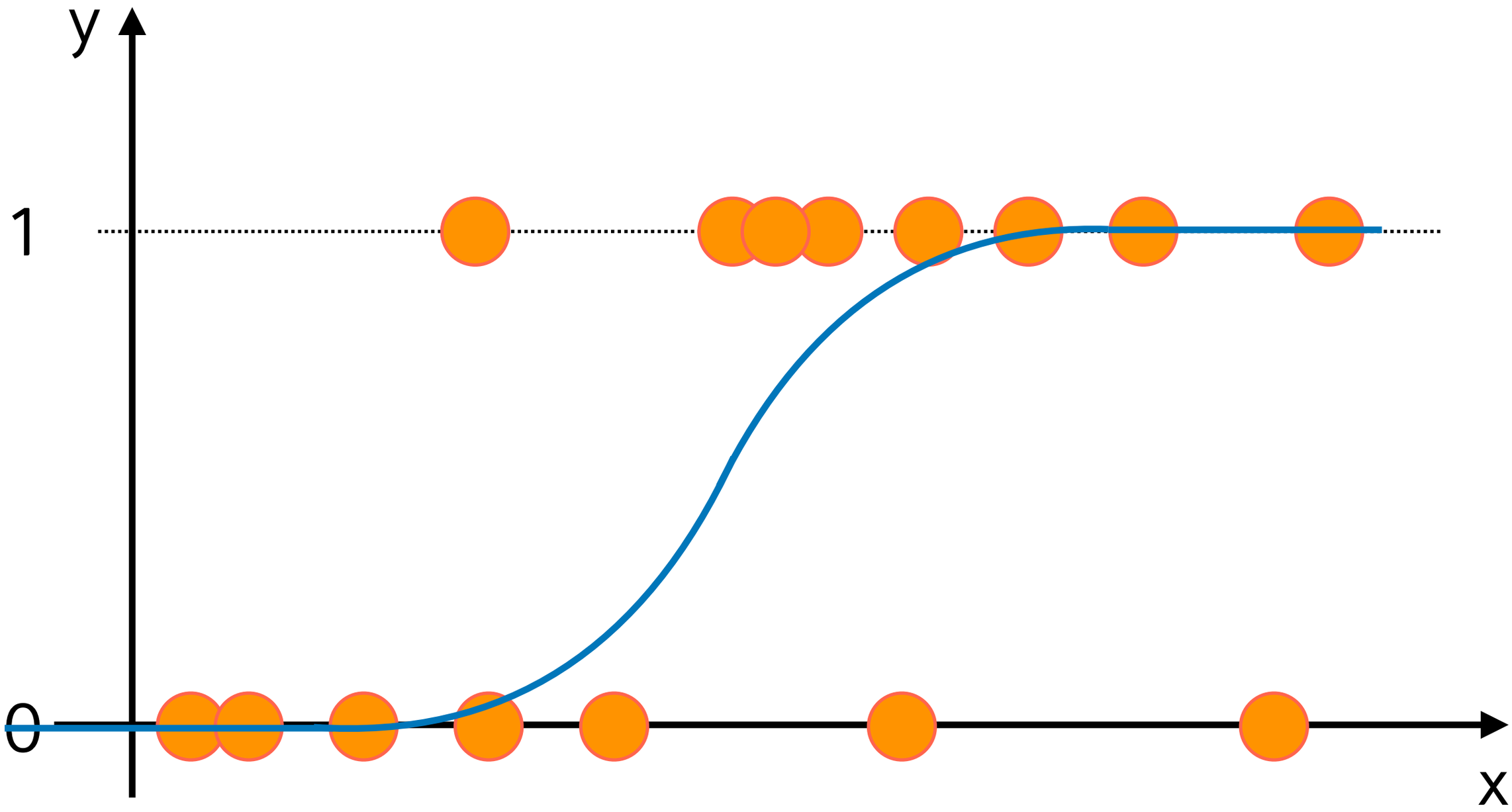
Regression

Classification

Logistic Regression

$$y = ax + b$$

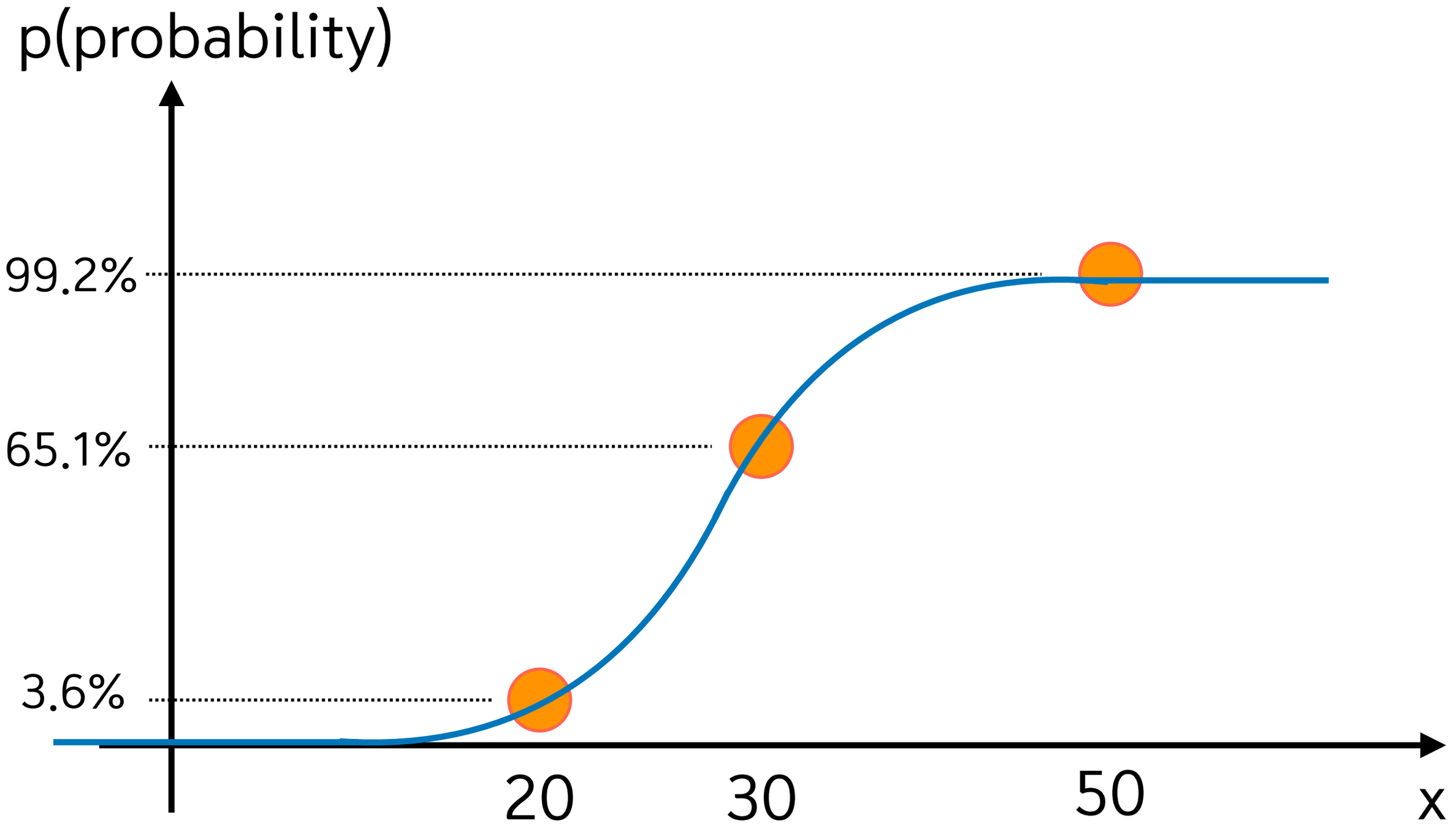


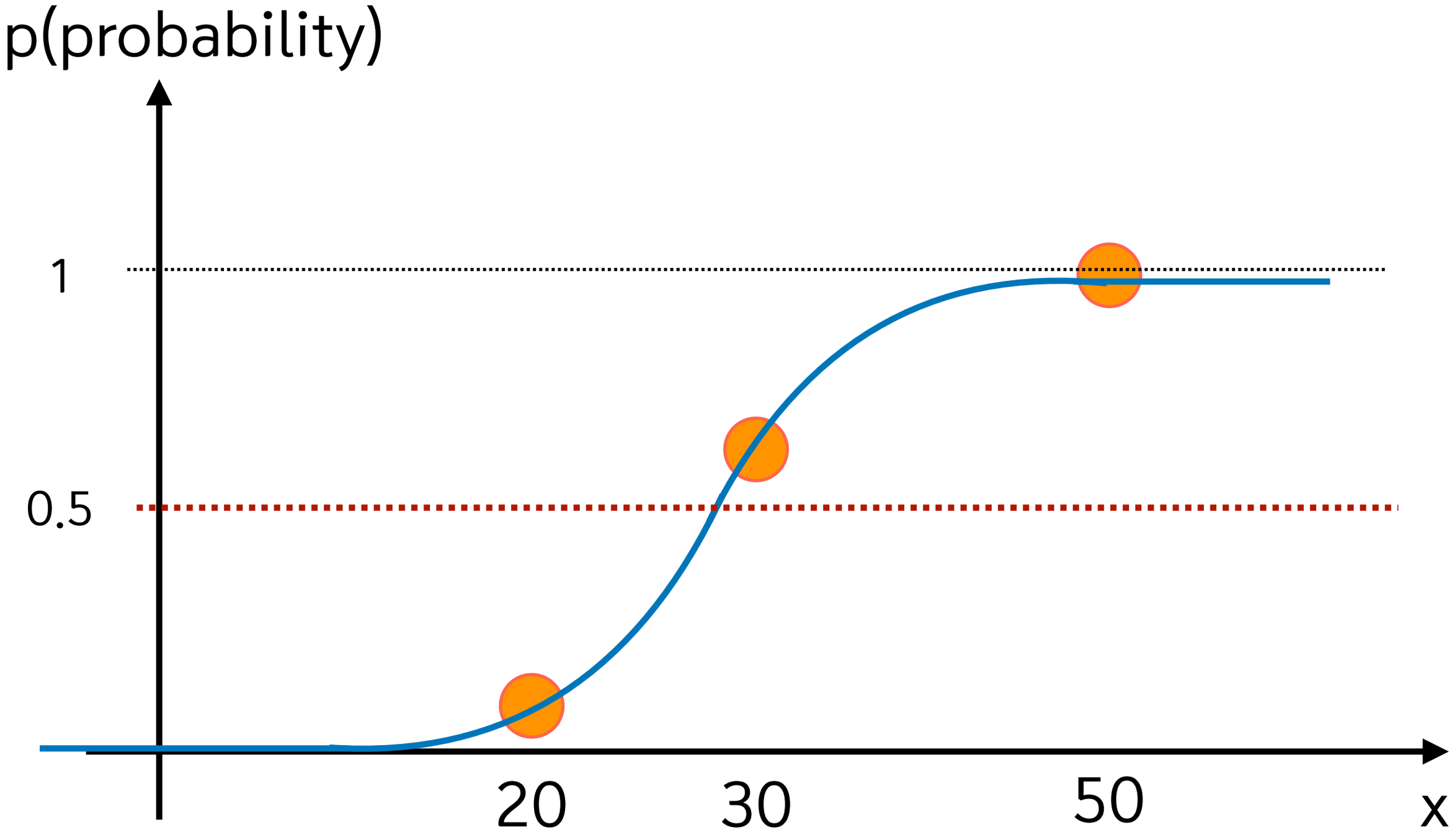

y
1
θ
x

$$y = a + bx$$

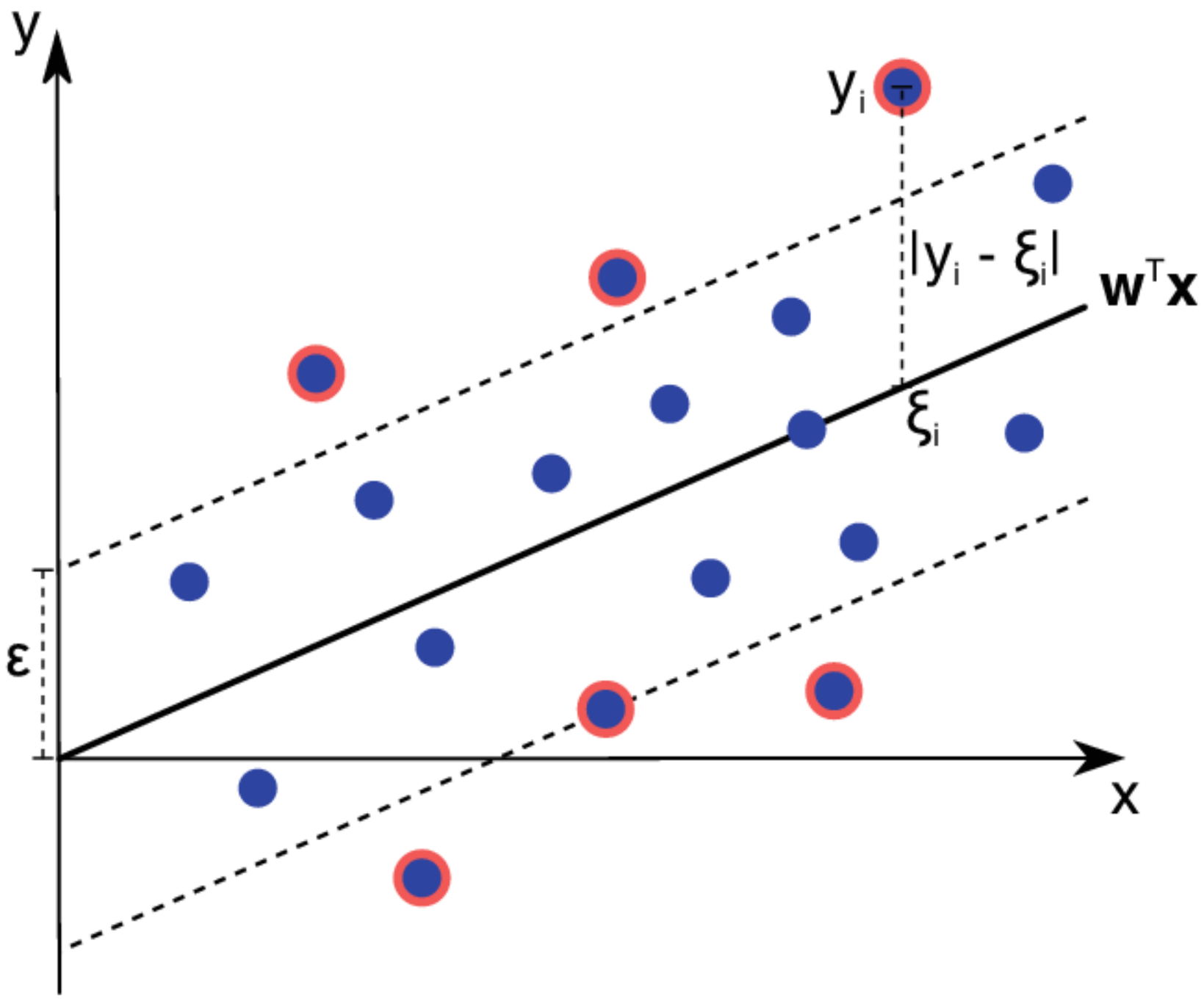
$$p = \frac{1}{(1 + e^{-y})}$$

$$ln(\frac{p}{1-p}) = a + bx$$


p(probability)
99.2%
65.1%
3.6%
20
30
50
x


p(probability)
1
0.5
20
30
50
x

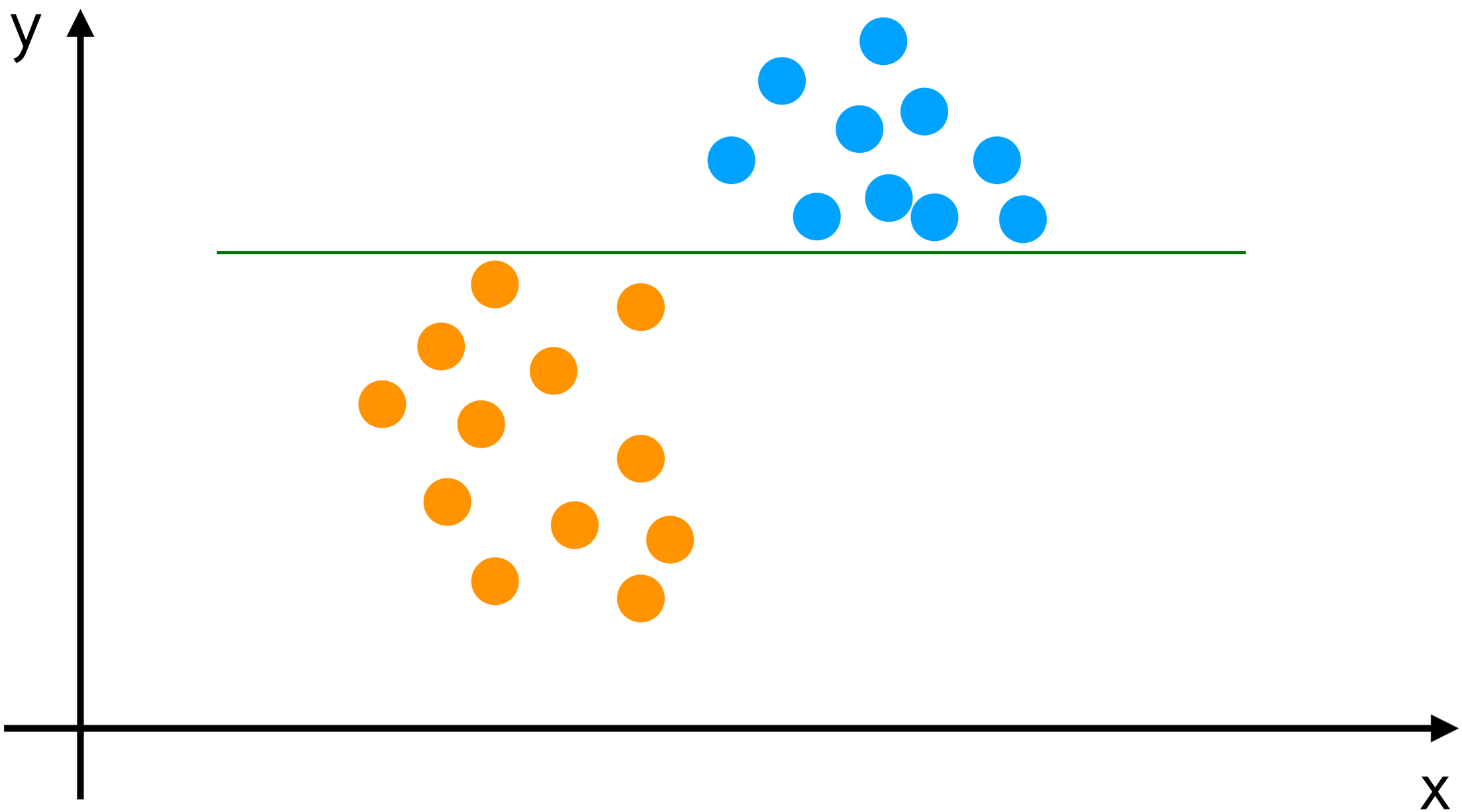
SVR
(Supported Vector Regression)



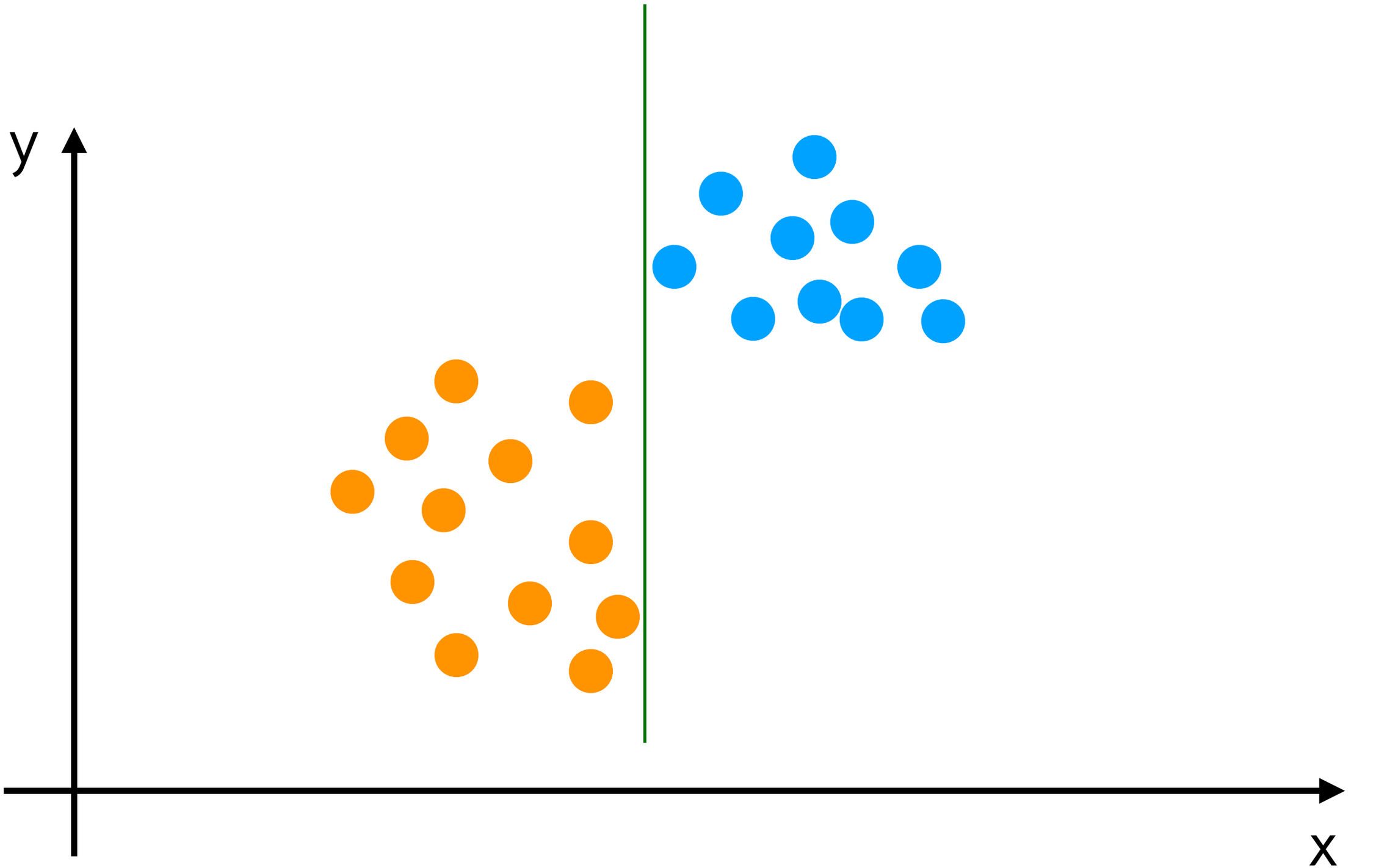
https://www.researchgate.net/profile/Frank_Boeckler/publication

SVM
(Supported Vector Machine)

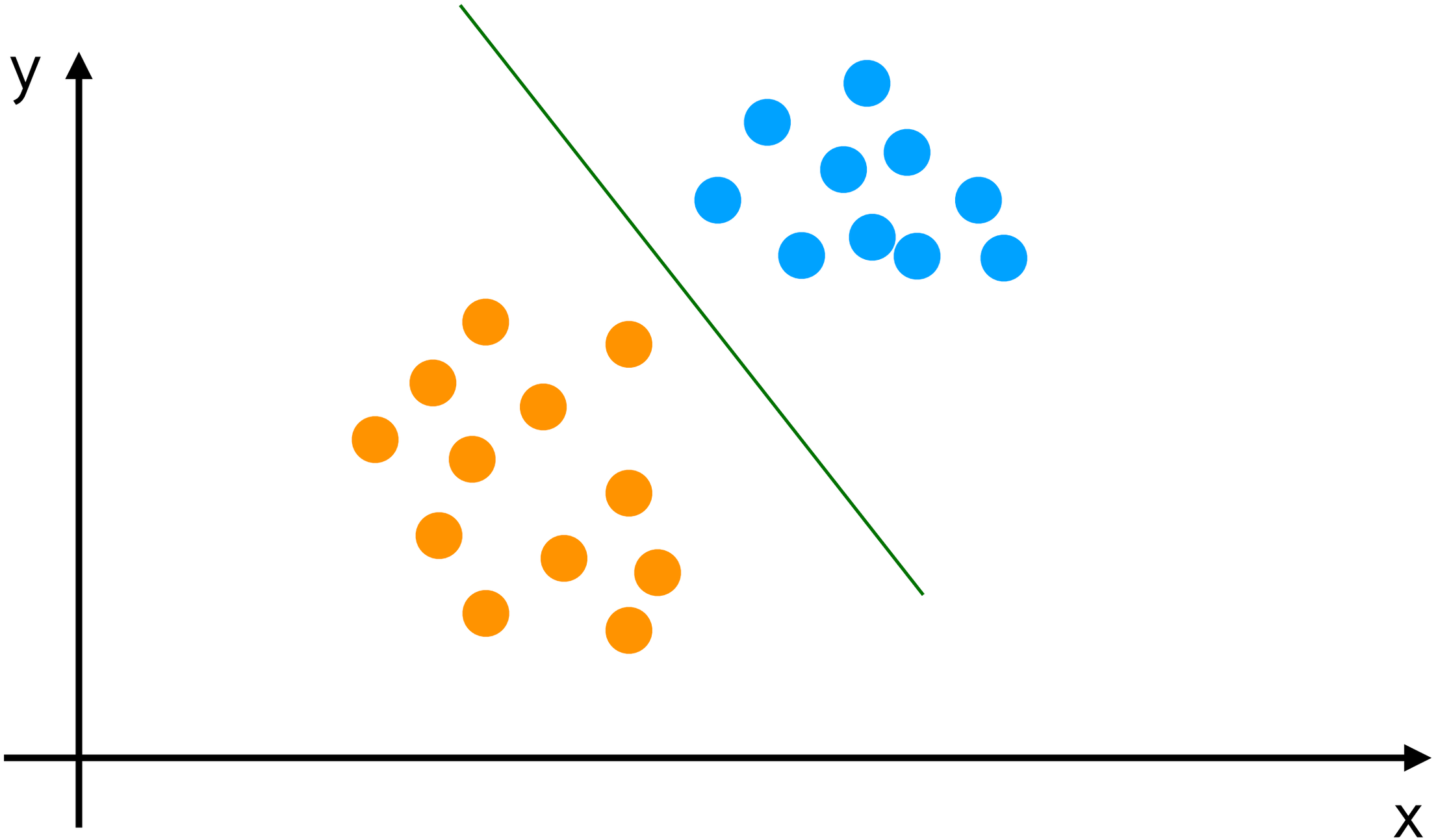
SVM(Supported Vector Machine)



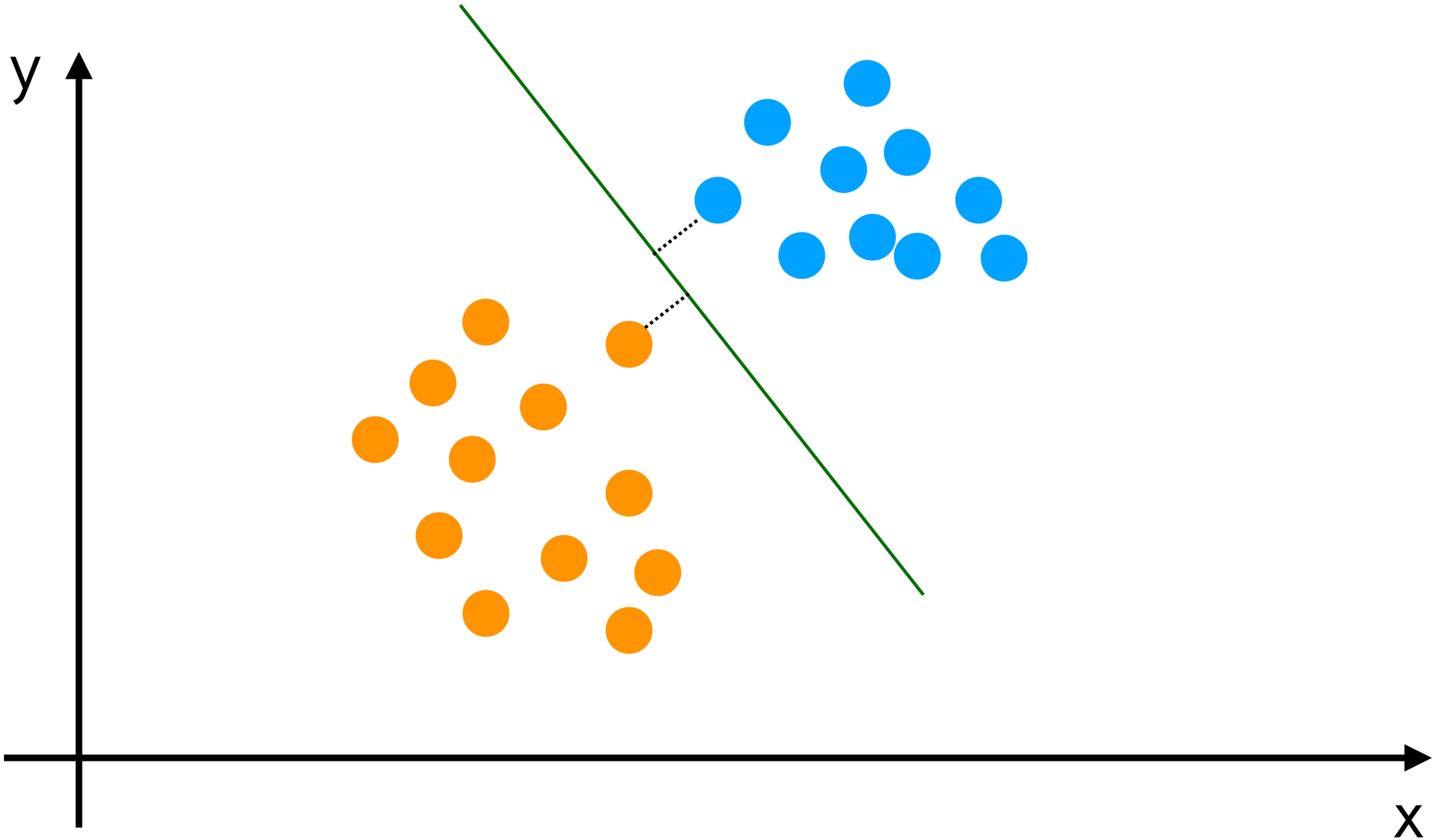
SVM(Supported Vector Machine)



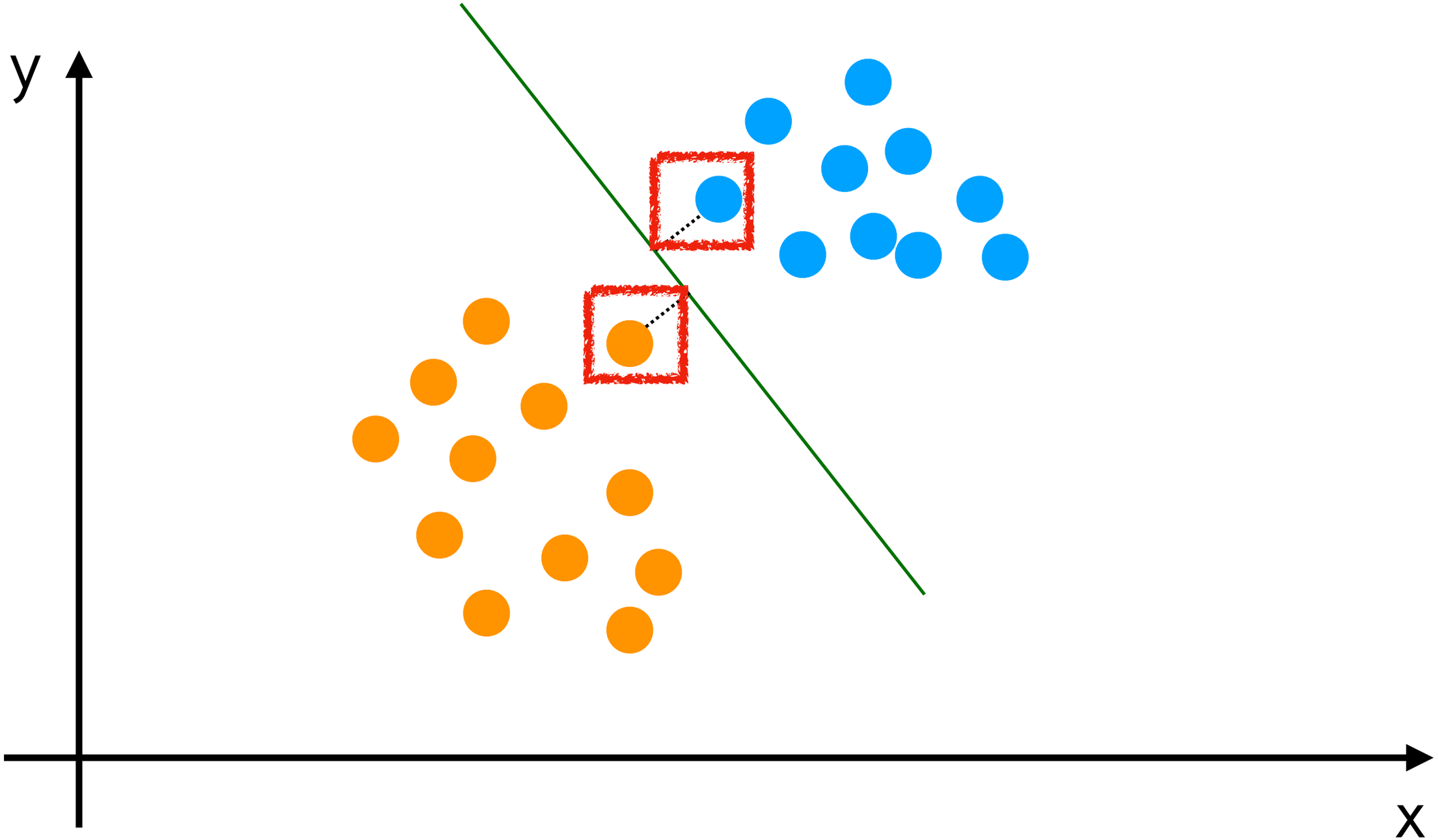
SVM(Supported Vector Machine)



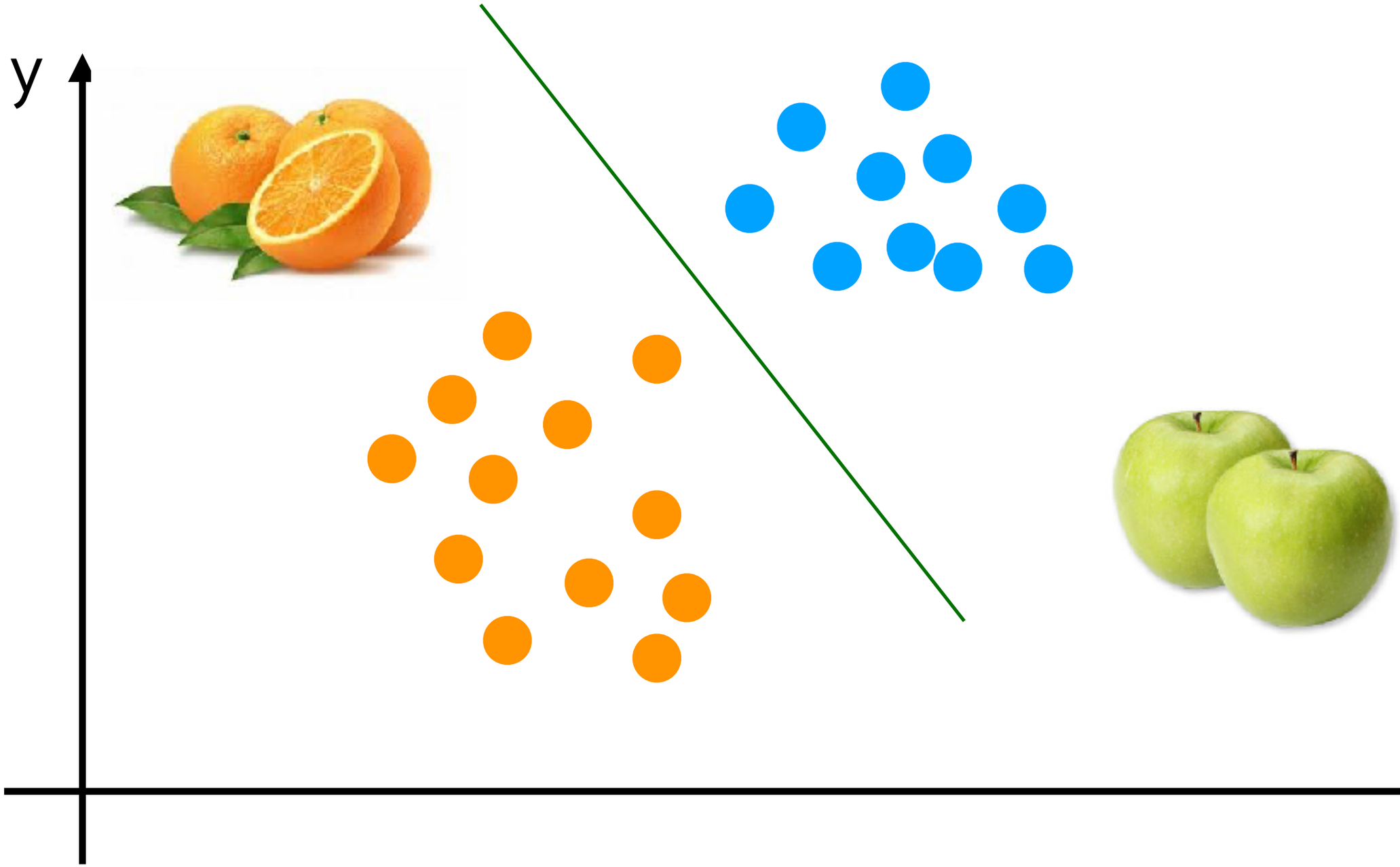
SVM(Supported Vector Machine)



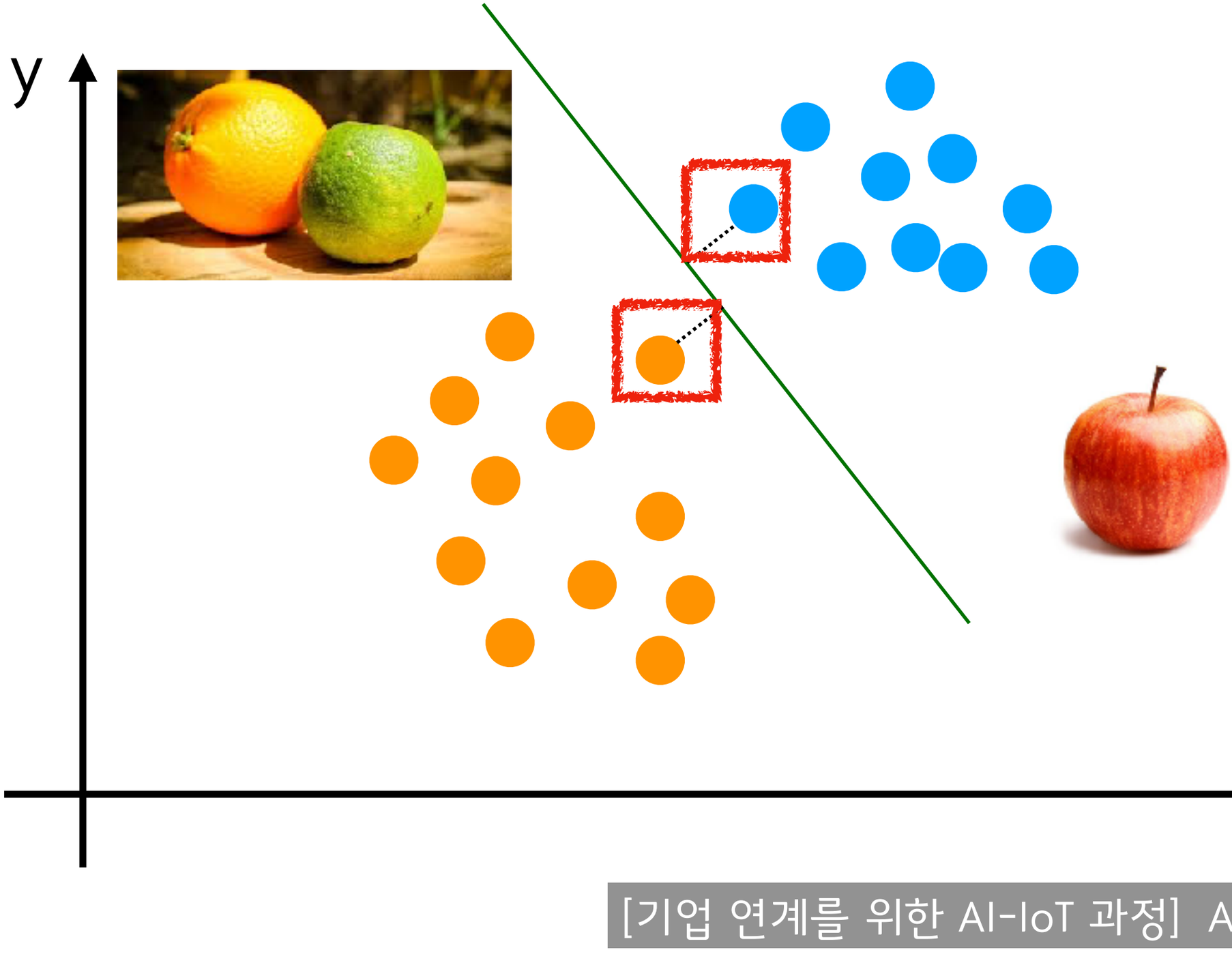
SVM(Supported Vector Machine)



SVM(Supported Vector Machine)



SVM(Supported Vector Machine)

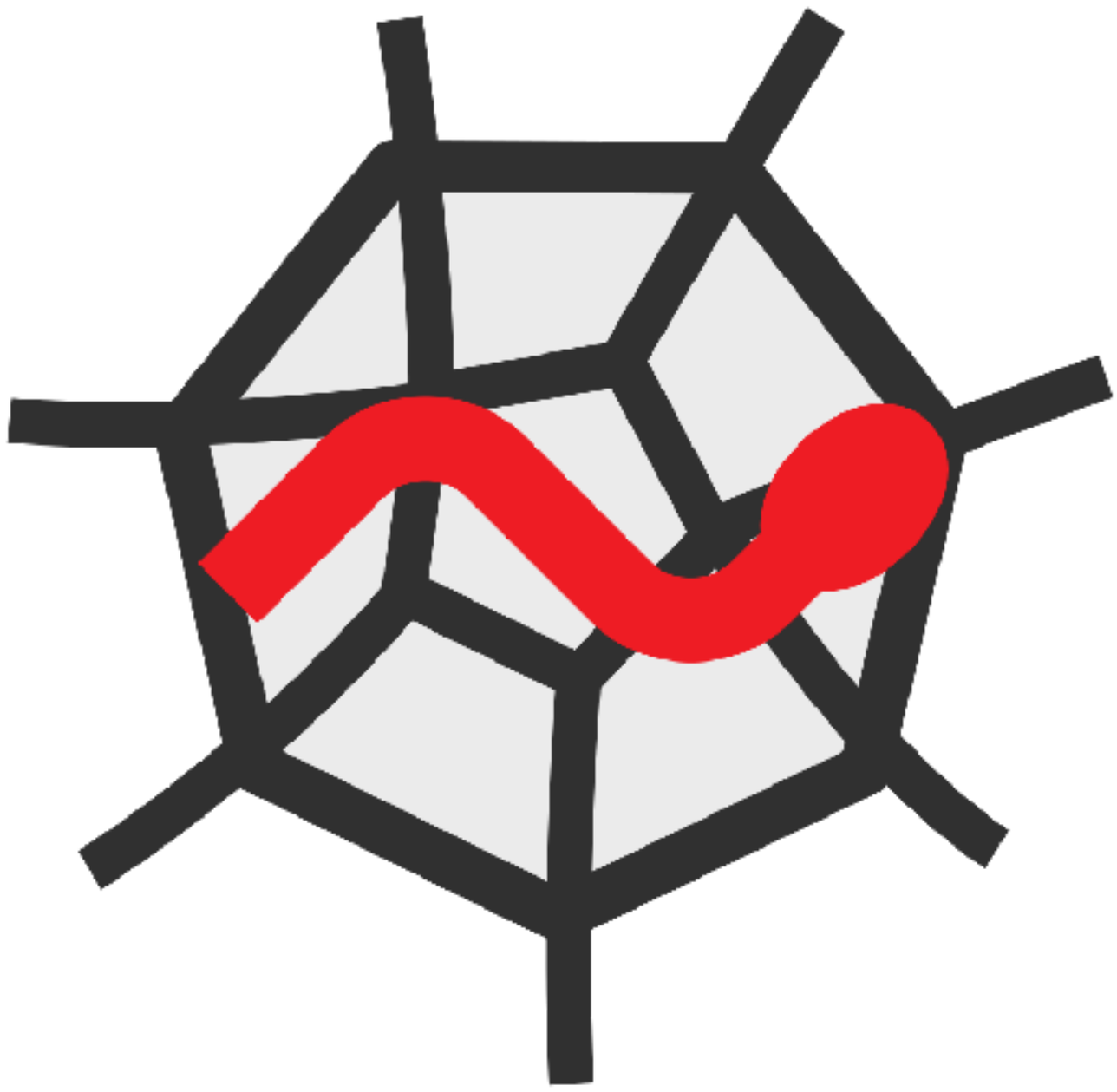


“All models are wrong but some are useful”

– George Box

두 번째 날

- Regression 모델 생성
- 머신러닝 알고리즘
- Scikit-learn 모델 생성



SPYDER


scikit
learn

1. .csv 데이터 가져오기

2. Feature/label 나누기

3. Missing Data

4. Categorical

5. Train/Test set

6. Standard

Label Encoding

name	Label
France	0
German	1
Spain	2

→

name	One Hot Encoding Format
France	[1, 0, 0]
German	[0, 1, 0]
Spain	[0, 0, 1]

세 번째 날

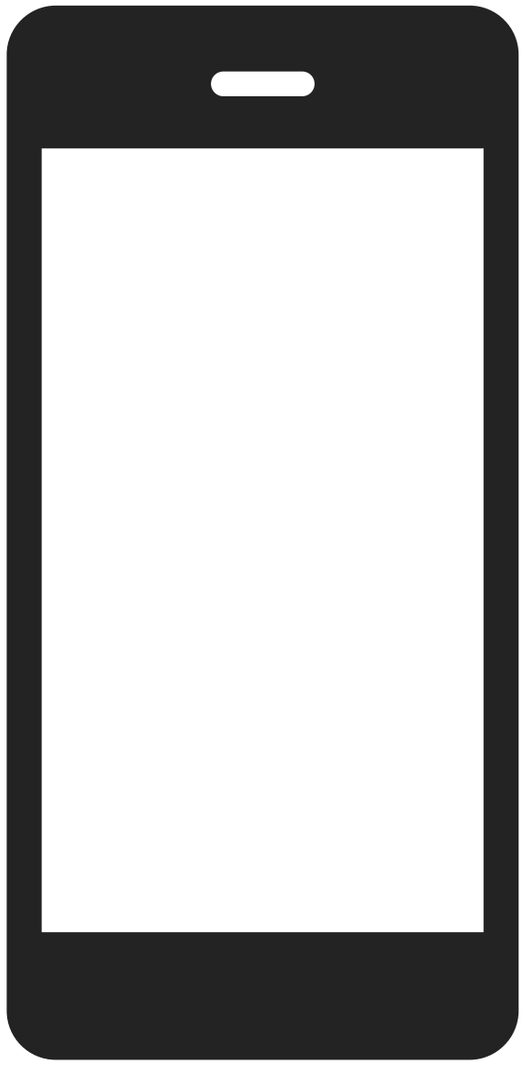
- 딥러닝 이론 소개
- 이미지 분류 방법
- Keras를 활용한 이미지 분류 모델 생성

네 번째 날

다섯 번째 날

AI on Android

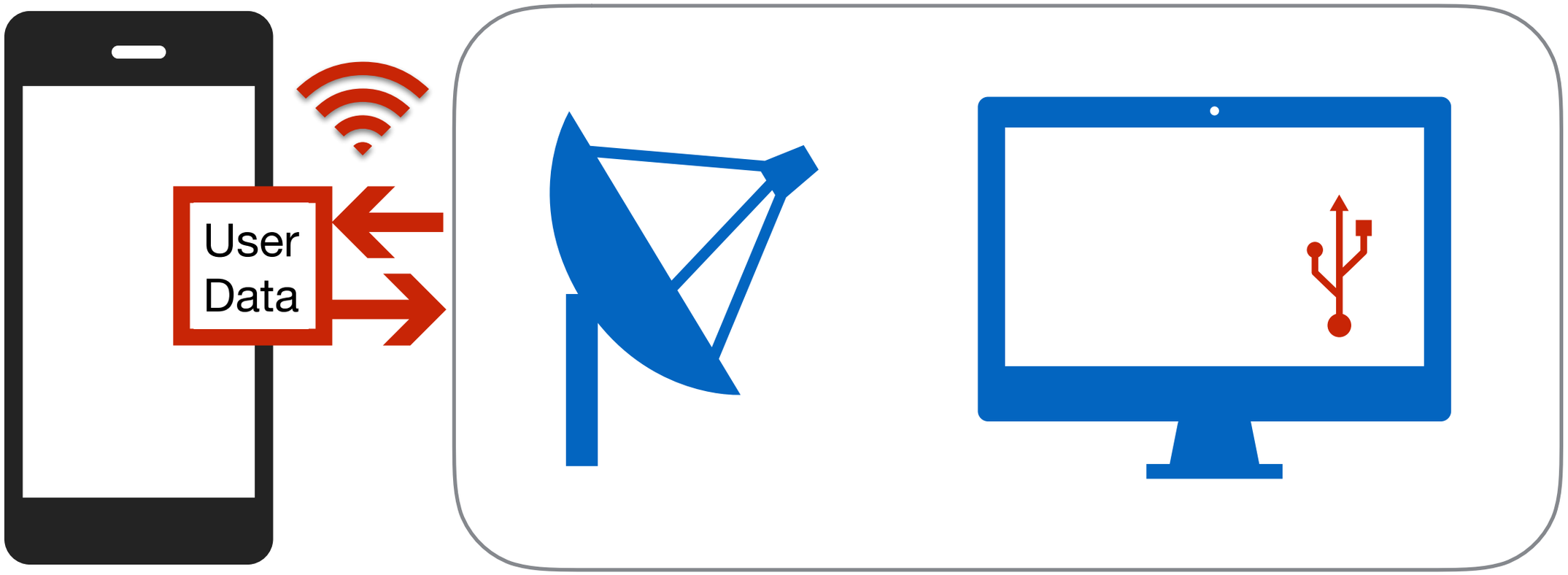
모바일 딥러닝



Mobile Engine

Pre-trained Model

Before 모바일 딥러닝



Mobile Device

Sever

모바일 딥러닝

Mobile Engine

Pre-trained Model

왜?

사용자
정보 보호

사용자
LTE 보호

서버 비용
절약

접근성
향상

https://developer.apple.com/videos/play/wwdc2017/703/

왜 이제서야?

0.1 s 😎

1 s 😯

5 s 😫

Geekbench 4 - cross-platform processor benchmark

MacBook Pro (13-inch Mid 2017)

Single-Core Score	Multi-Core Score
4440	9414

Geekbench 4.2.3 Tryout for Mac OS X x86 (64-bit)

Result Information

Upload Date	June 27 2018 11:20 PM
Views	7

System Information

System Information	
Operating System	macOS 10.13.5 (Build 17F77)
Model	MacBook Pro (13-inch Mid 2017)
Motherboard	Apple Inc. Mac-B4831CEBD52A0C4C MacBookPro
Memory	8192 MB 2133 MHz LPDDR3
Northbridge	
Southbridge	

iPhone X

Single-Core Score	Multi-Core Score
4255	10015

Geekbench 4.2.3 for iOS AArch64

Result Information

Upload Date	June 28 2018 12:21 AM
Views	3

System Information

System Information	
Operating System	iOS 11.4
Model	iPhone X
Motherboard	D221AP
Memory	2785 MB
Processor Information	
Name	Apple A11 Bionic

감정 분석

컨퍼런스 듣길 잘했군! —> 😄

이미지 분류

—> 해변

음악 태깅

—> 클래식

손글씨 분석

—> 7

문장 번역

I’m happy. —> 난 행복해

문자열 예측

집에 갈때... —> 메로나

https://developer.apple.com/videos/play/wwdc2017/703/

Object Detection

Classification

Regression

Decision Tree Classifier

Random Forest Classifier

SVM

Logistic Regression

Random Forest Regression

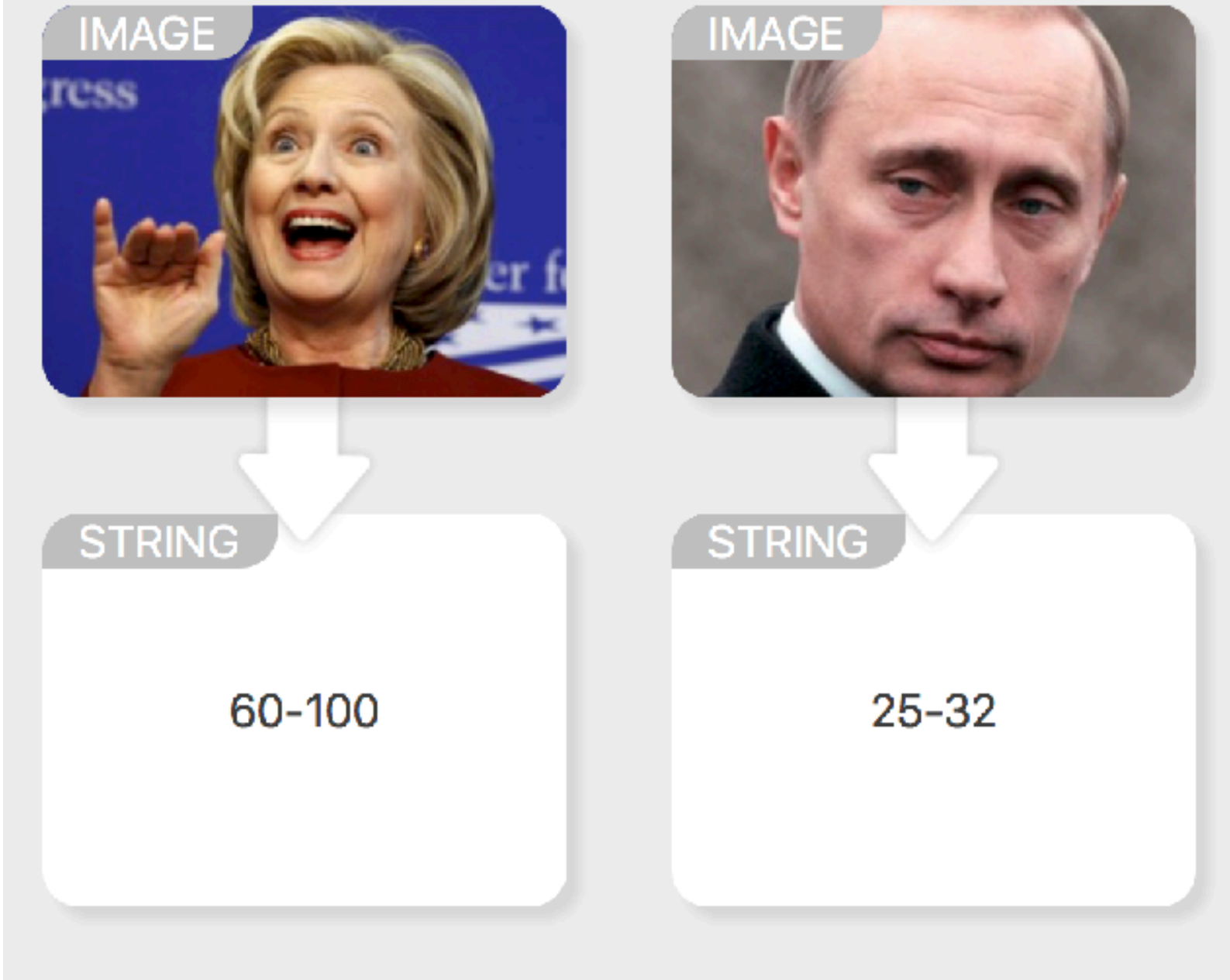
Linear Regression

AgeNet

43.5 MB

Input: Image

Output: String

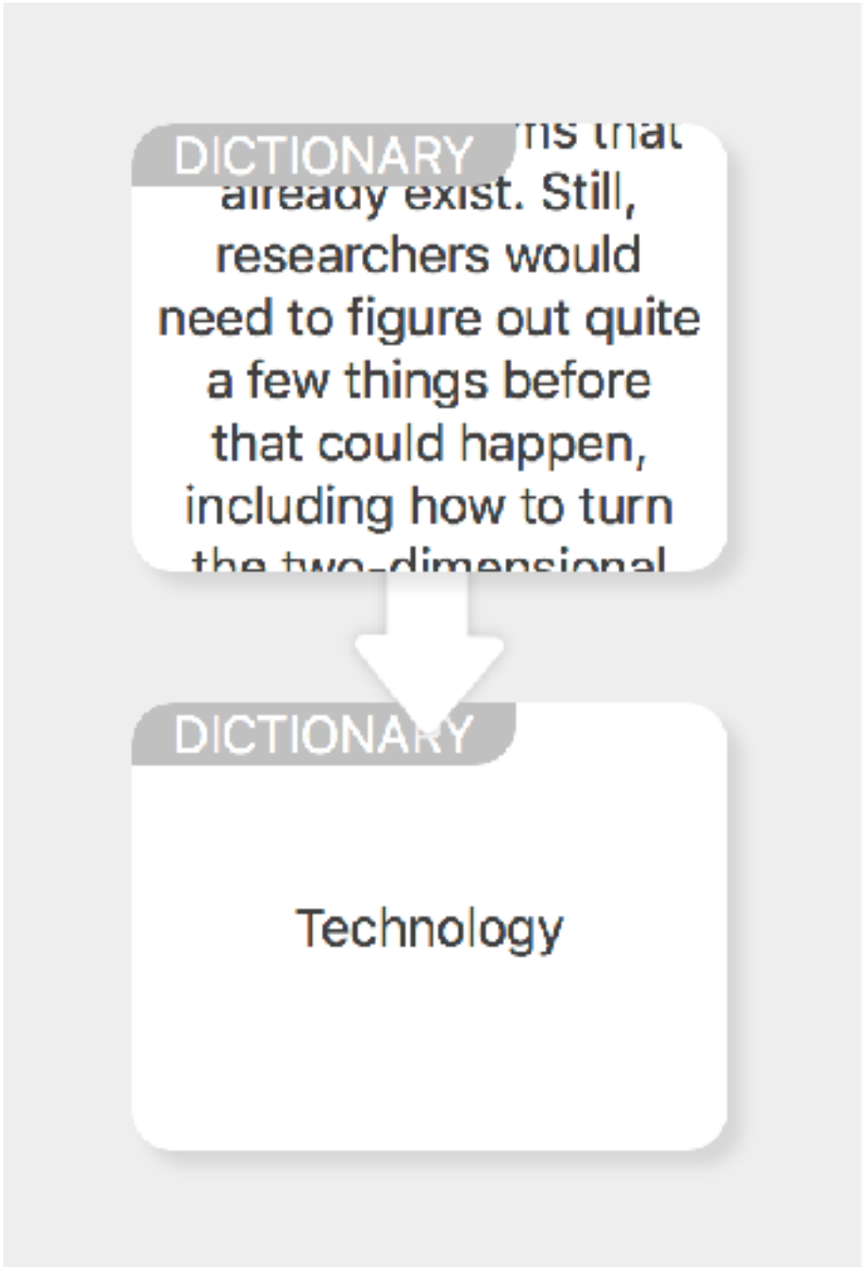


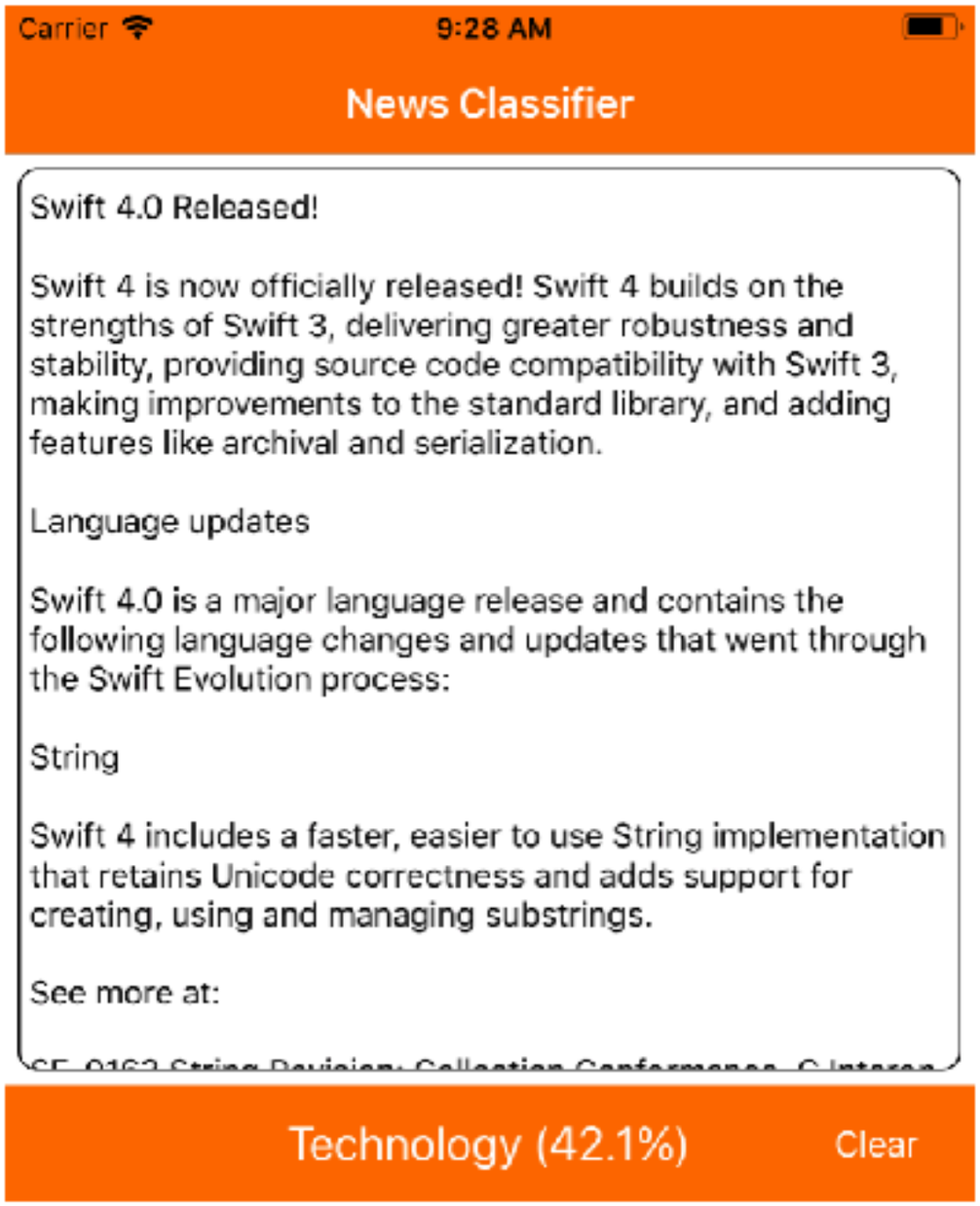
DocumentClassification

1.4 MB

Input: String

Output: String



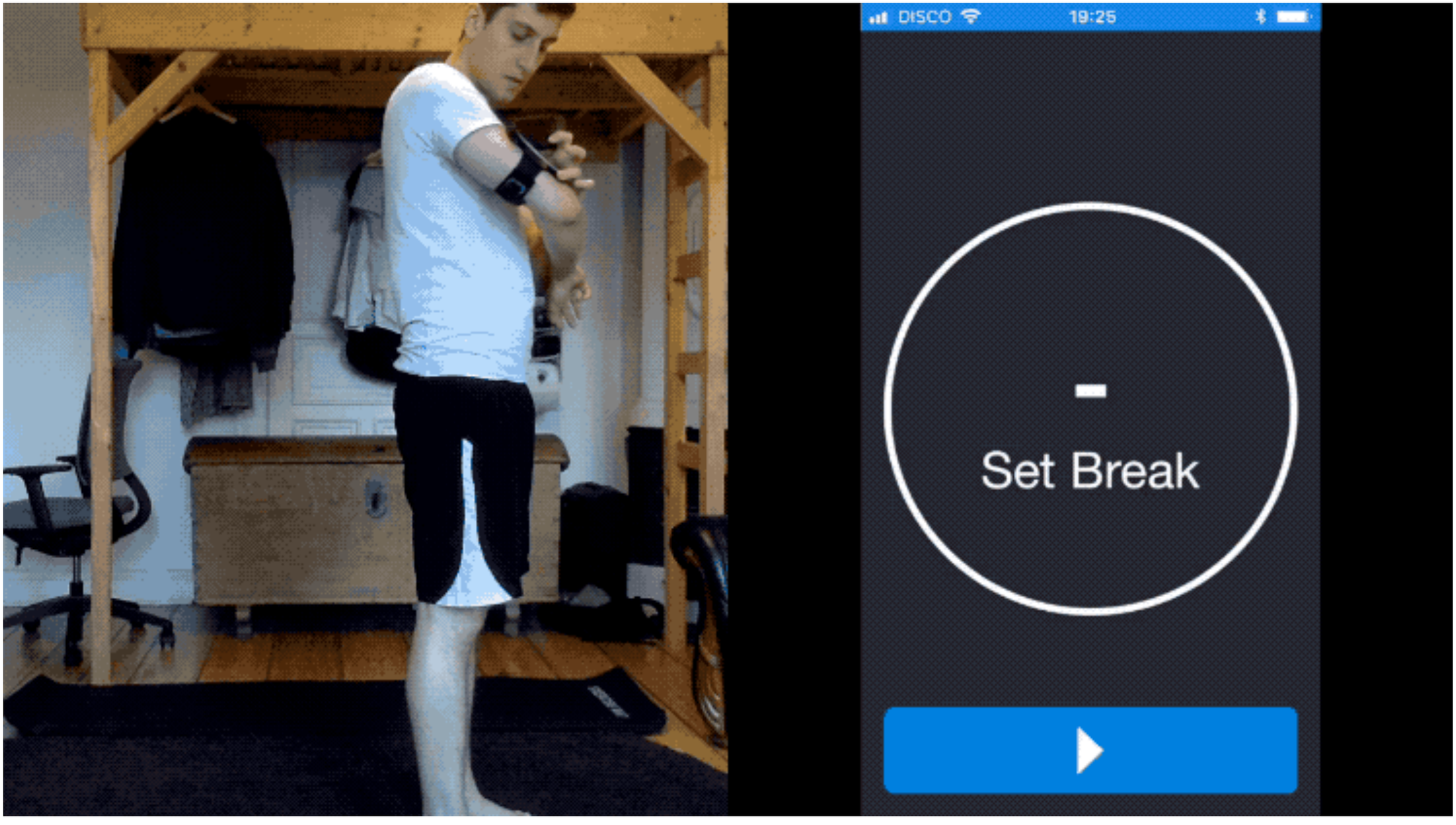


Exermote

80 KB

Input: Accelerations

Output: Array



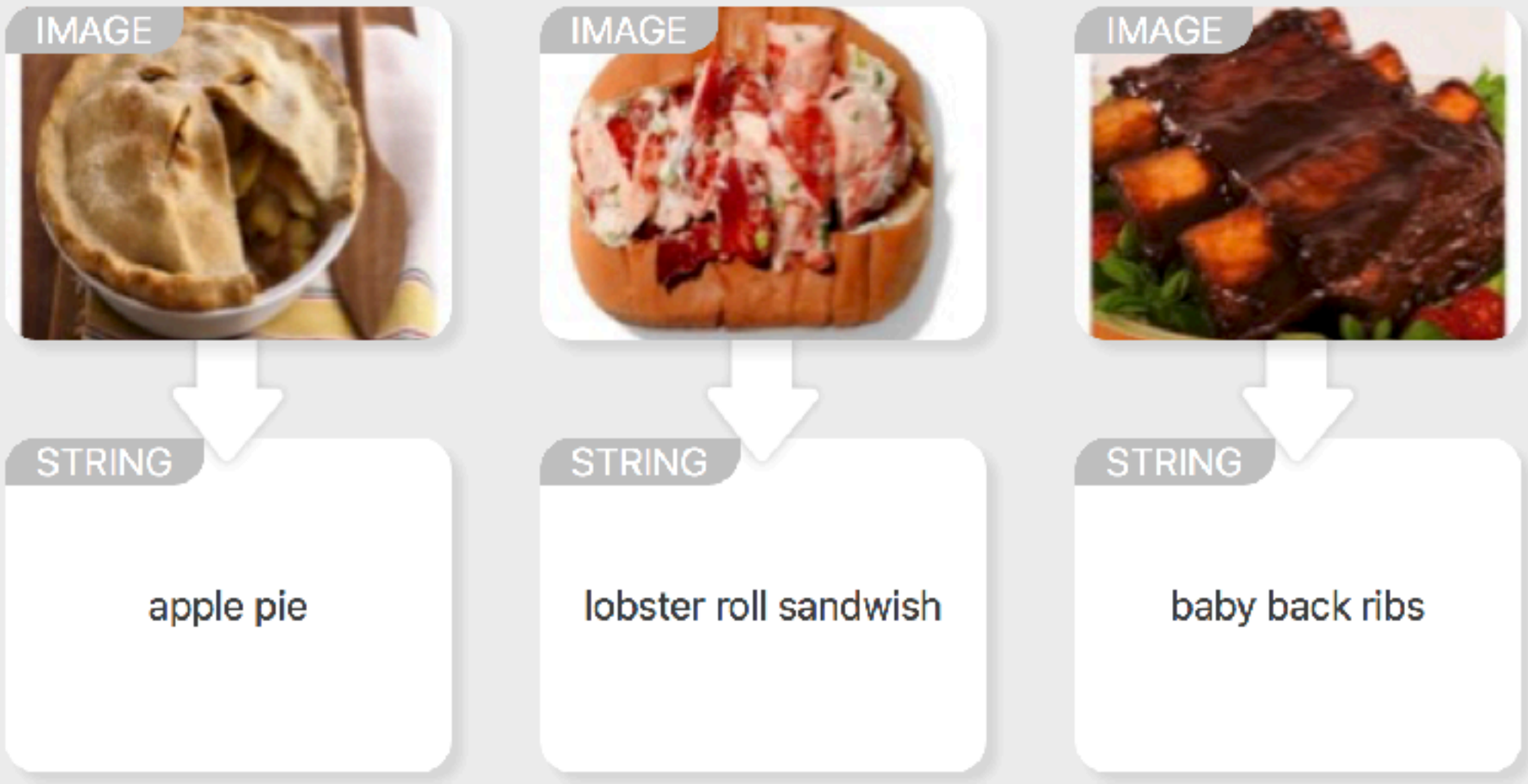
[기업 연계를 위한 AI-IoT 과정] AI | 두번째 날 | 전미정

https://github.com/Laurabert/Exermote/tree/master/ExermoteInference

Food101

83.3 MB

Input: Image



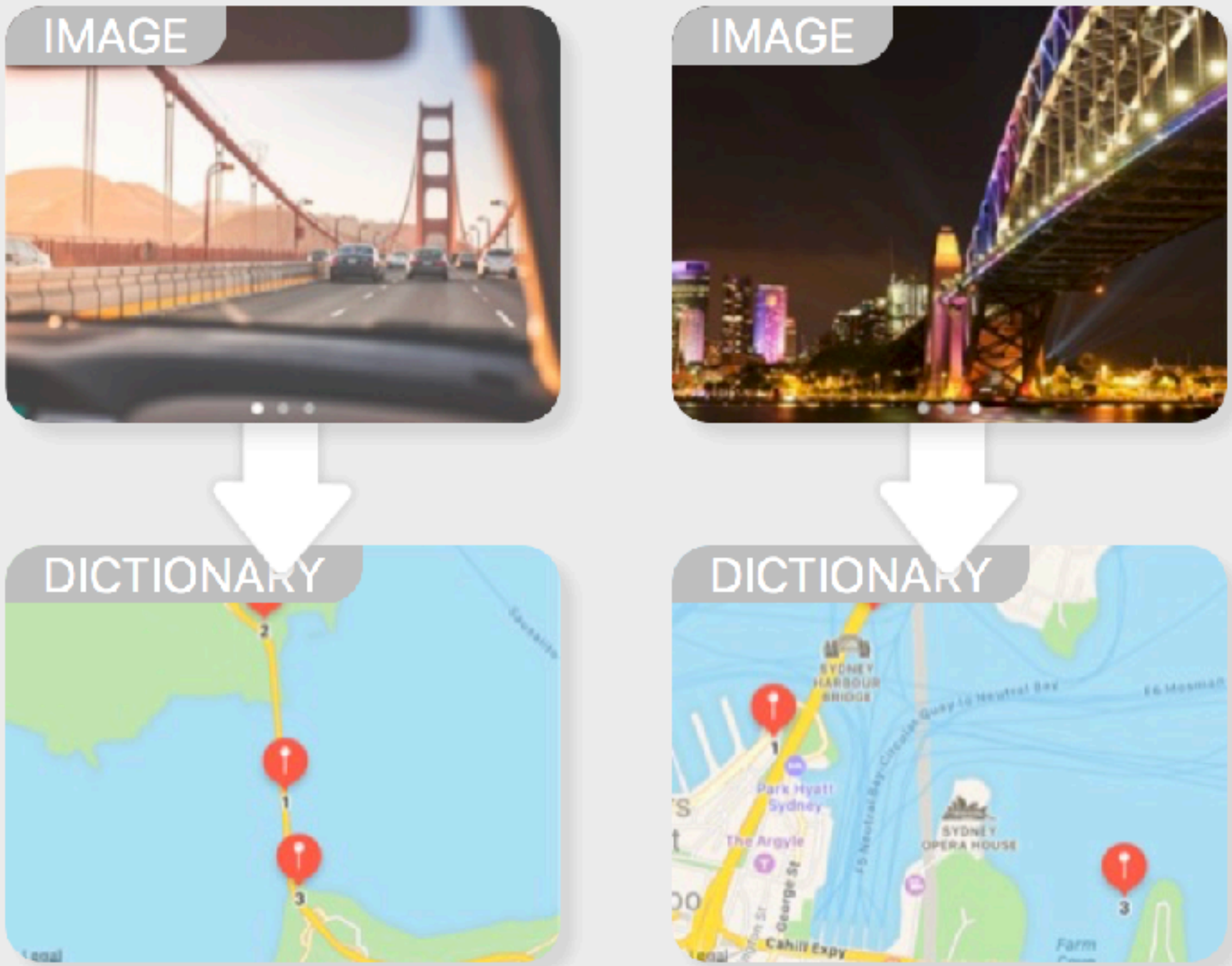
Output: String

RN1015k500

284.4 MB

Input: Image

Output: Dictionary



Webs for more ML

- 캐글 데이터셋
- UCI 데이터셋
- Azure 갤러리
- Azure ML Studio에서 R 사용하기 실습
- Azure Machine Learning Service
- Azure Notebook
- Azure Cognitive Sevice

Books for more ML

입문자가 읽기 좋은 **머신러닝** 책

- 핸즈온 머신러닝_오렐리앙 제롱(한빛미디어)
- 처음 배우는 머신러닝_김승연,정용주(한빛미디어)
- 모두의 딥러닝_조태호(길벗)

그리고 **딥러닝**에 대해 더 공부하고 싶다면

- 블록과 함께하는 파이썬 딥러닝 케라스 이야기_김태영(디지털북스)
- 케라스 창시자에게 배우는 딥러닝_프랑소와 숄레(길벗)
- 밑바닥부터 시작하는 딥러닝_사이토 고키(한빛미디어)

수고하셨습니다 

ninevincentg@gmail.com