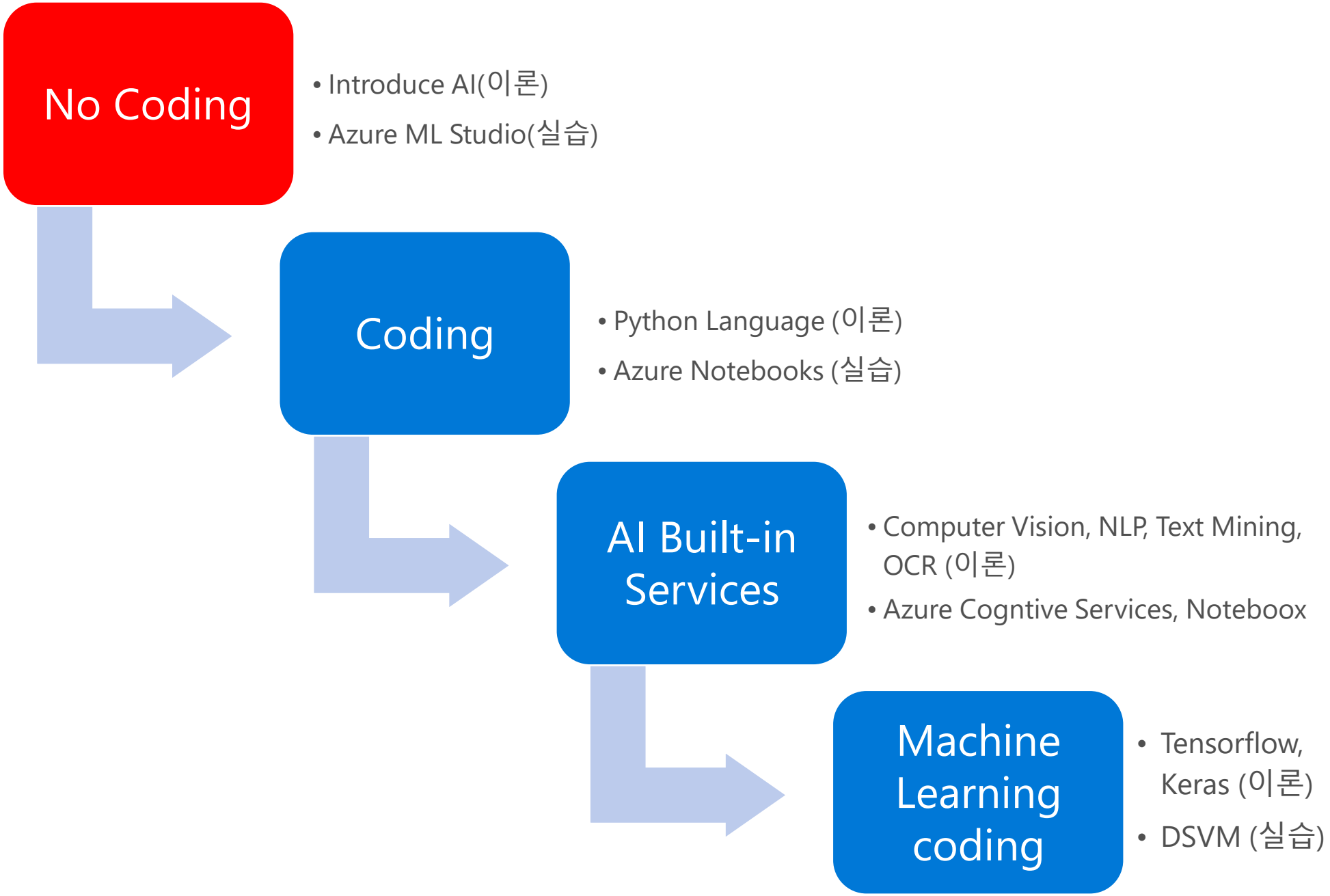
머글들을 위한
Machine Learning 2

김영욱
부장/ PS / Microsoft

youngwook@outlook.com
Blog: Youngwook.com


No Coding
• Introduce AI(이론)
• Azure ML Studio(실습)
Coding
• Python Language (이론)
• Azure Notebooks (실습)
AI Built-in Services
• Computer Vision, NLP, Text Mining, OCR (이론)
• Azure Cogntive Services, Noteboox
Machine Learning coding
• Tensorflow, Keras (이론)
• DSVM (실습)

Steps

First-Understand the Business Domain

Second-Understand the Business Problem

Third- What is the Right Data, Right Column and Right Algorithm

Last-Combine Knowledge With Machine Learning

문제 정의

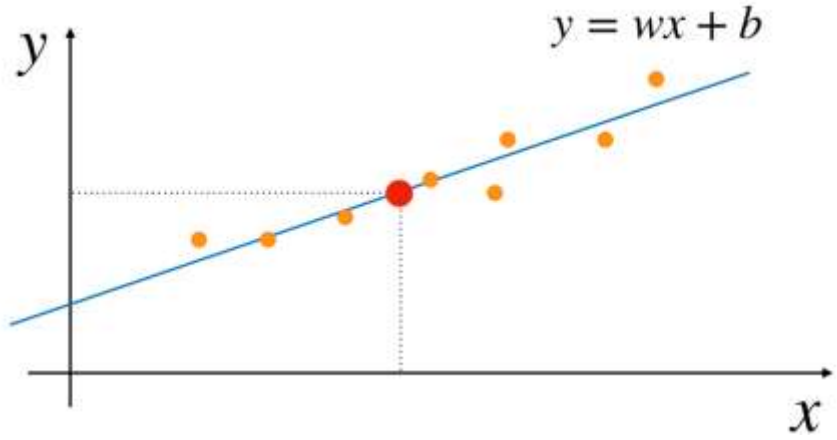
데이터 셋 준비

모델 설정

모델 훈련 / 평가

모델 활용

근속연수	연봉
1.5	3,100
2.5	3,900
4.2	4,300
5.1	4,900
6.7	5,400
8.3	6,700
9.5	9,200
13	12,900



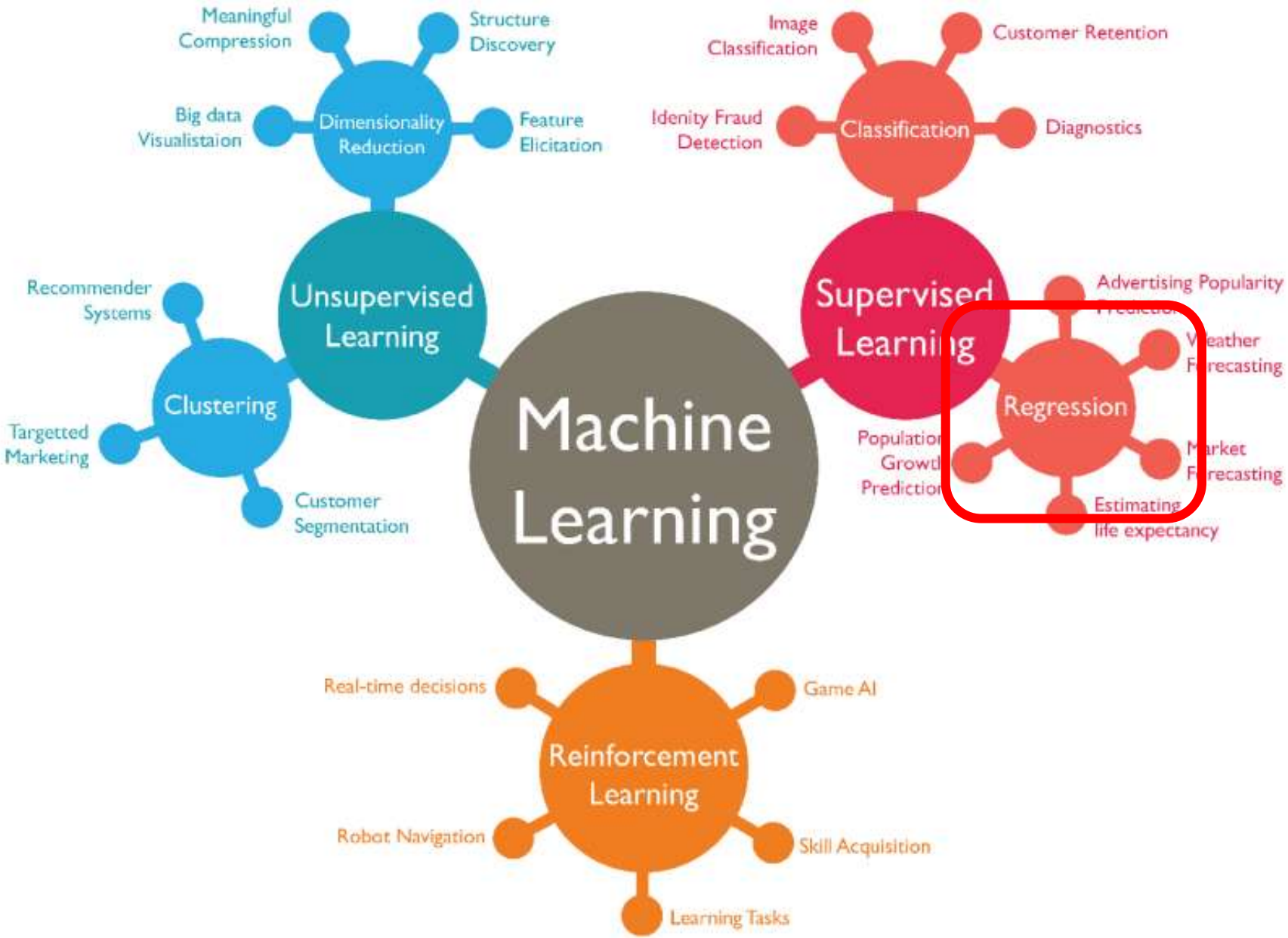
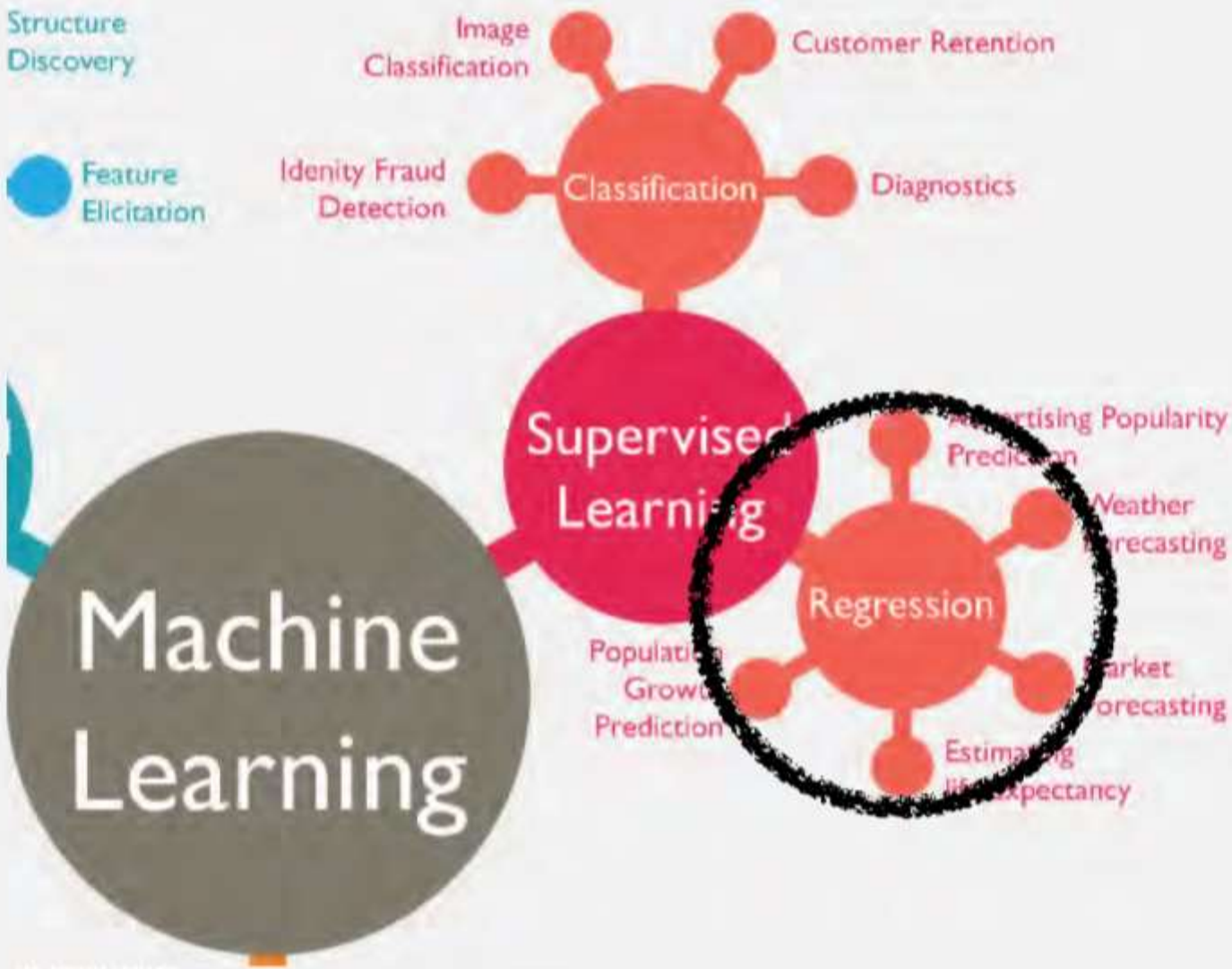


Image credits: https://www.slideshare.net/awahid/big-data-and-machine-learning-for-businesses


Structure Discovery
Feature Elicitation
Image Classification
Customer Retention
Idenity Fraud Detection
Classification
Diagnostics
Supervised Learning
Regression
Machine Learning

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning 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

Three or more

Predicting categories

Predicting values

Two

ANOMALY DETECTION

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

CLUSTERING

- K-means

MULTICLASS 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

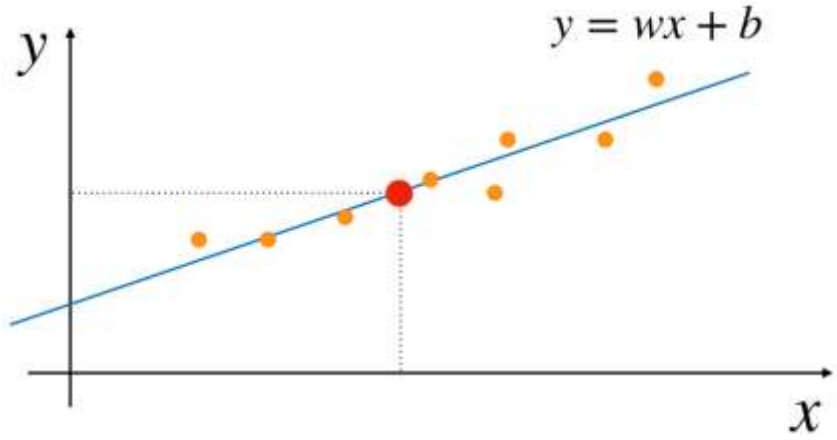
TWO-CLASS CLASSIFICATION

- Two-class SVM — >100 features, linear model
- Two-class averaged perceptron — Fast training, linear model
- Two-class logistic regression — Fast training, linear model
- Accuracy, fast training — Two-class decision forest
- Accuracy, fast training — Two-class boosted decision tree
- Accuracy, small memory footprint — Two-class decision jungle
- >100 features — Two-class locally deep SVM

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

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

근속연수	연봉
1.5	3,100
2.5	3,900
4.2	4,300
5.1	4,900
6.7	5,400
8.3	6,700
9.5	9,200
13	12,900




Data

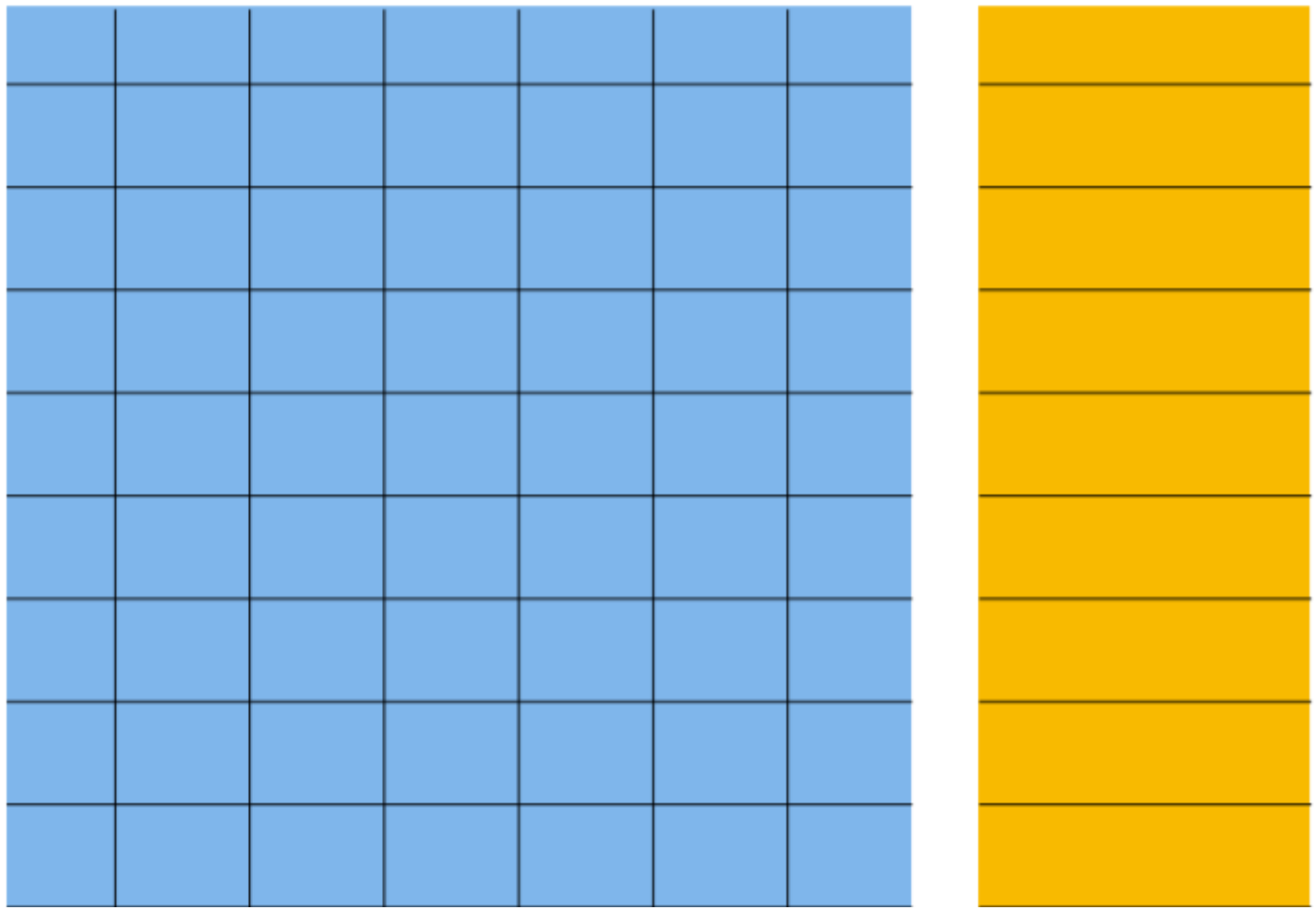
Feature

Label

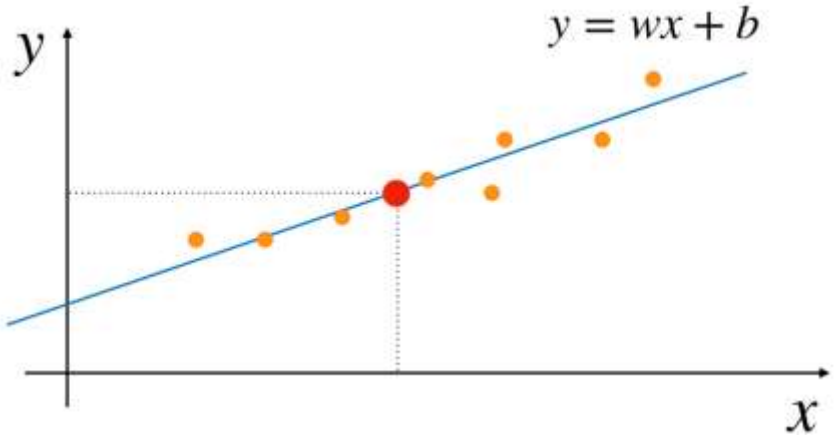
이름	나이	성별	선실 등급	티켓넘버	티켓 요금	부모 자식	형제 자매	키	출항지	생존 여부
Rose	25	여	A	EA-1039	300	2명	1명	167	런던	Y
Jack	20	남	C	GE-3059	29	-	-	178	도버	N
Mark	57	남	B	BA-2031	89	4명	3명	167	뉴포트	N
Andy	48	남	B	NN-3928	102	5명	7명	182	런던	Y

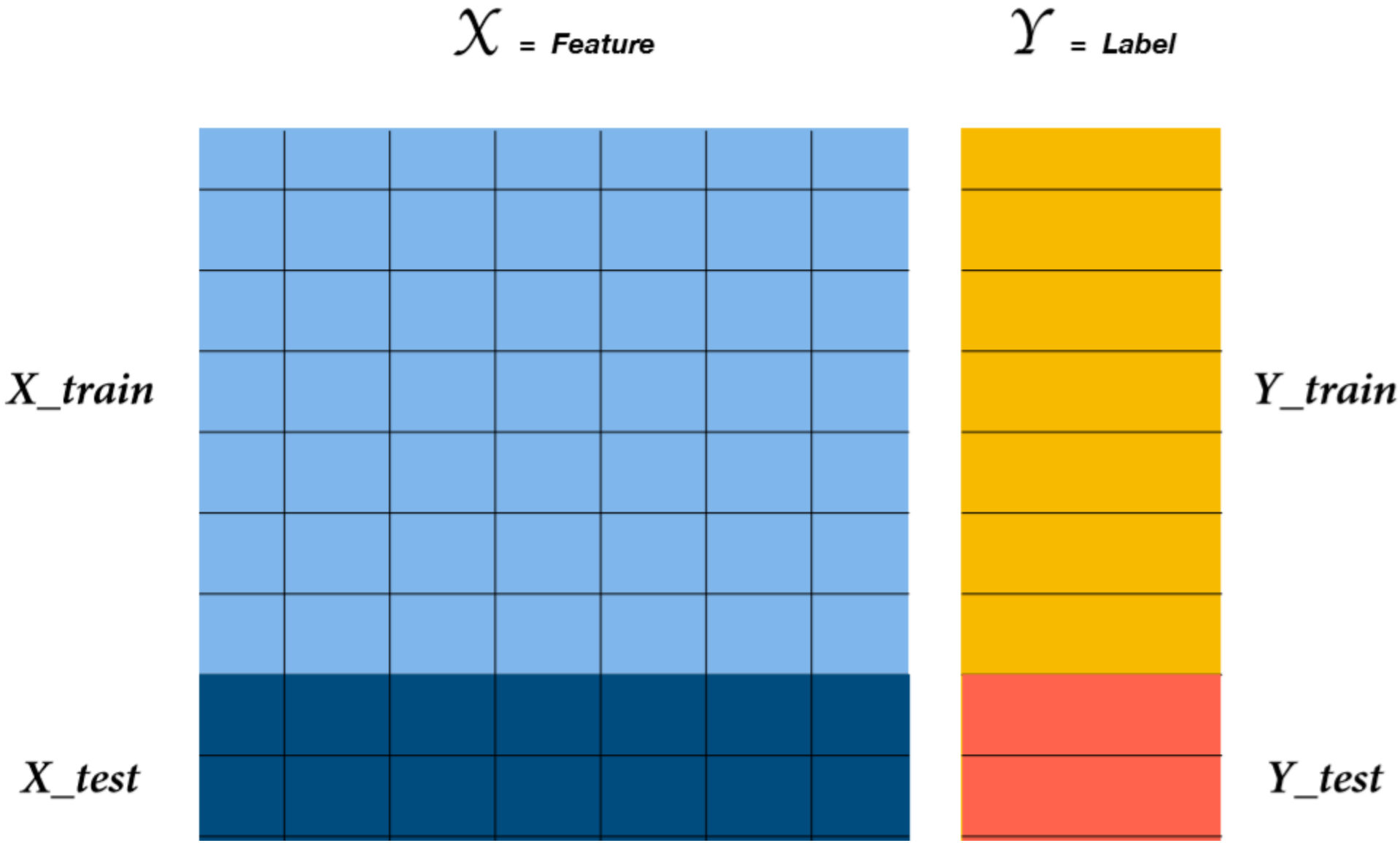
X = *Feature*

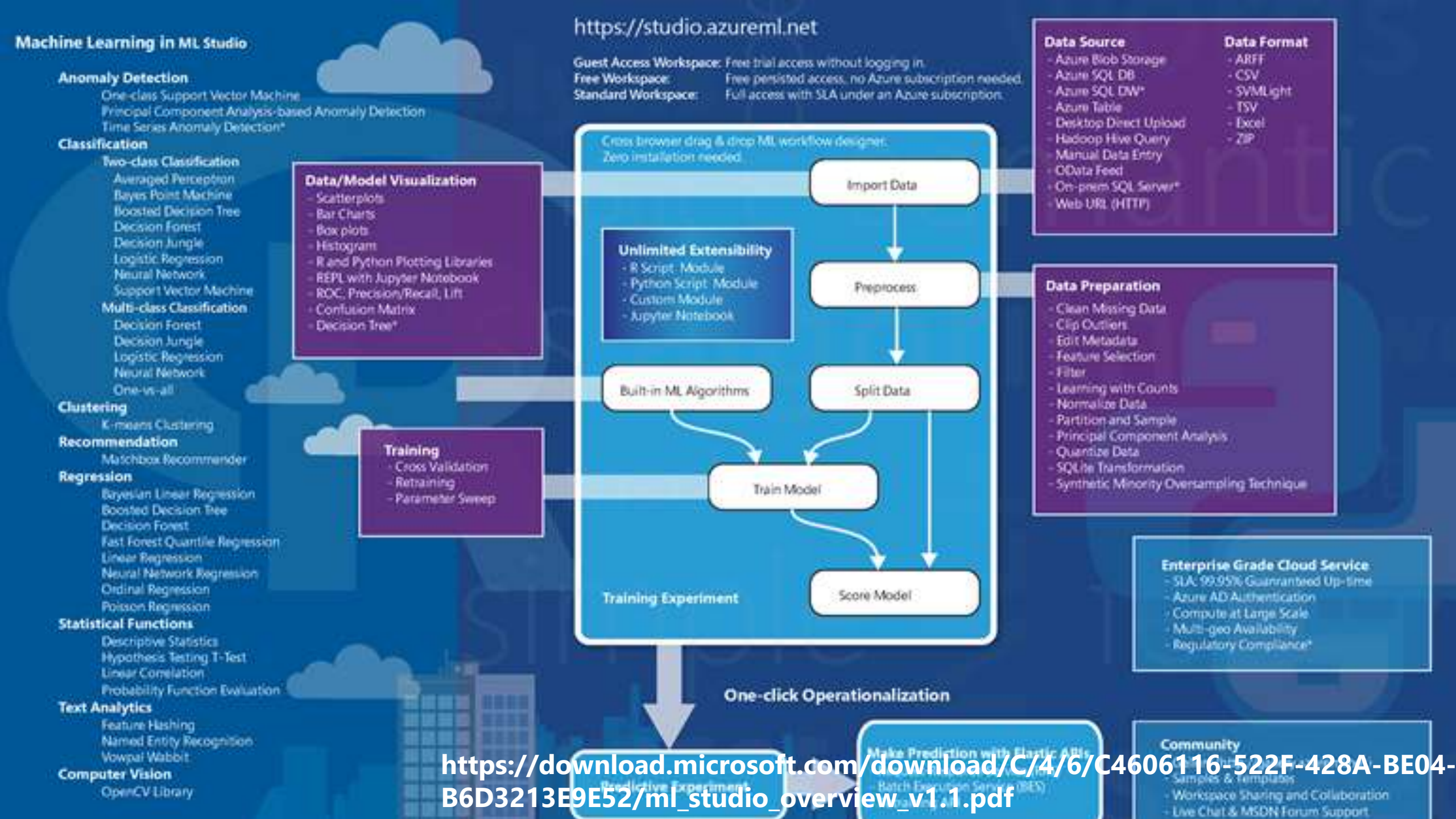
Y = *Label*



근속연수(x)	연봉(y)
1.5	3,100
2.5	3,900
4.2	4,300
5.1	4,900
6.7	5,400
8.3	6,700
9.5	9,200
13	12,900




X = Feature
Y = Label
X_train
X_test
Y_train
Y_test



실습1:

Linear Regression

https://azure.microsoft.com/ko-kr/pricing/details/machine-learning-studio/

스튜디오 가격

Machine Learning Studio는 무료 및 표준 두 계층으로 제공됩니다.

아래의 표는 계층별로 기능을 비교하여 보여 줍니다.

	무료	STANDARD
가격	무료	매월 ML Studio 작업 영역당 ₩11,235.254 ₩1,124.65/스튜디오 실험 시간
Azure 구독	필요 없음	필수
실험당 최대 모듈 수	100	제한 없음
최대 실험 기간	실험당 1시간	실험당 최대 7일, 모듈당 최대 24시간
최대 저장 공간	10GB	제한 없음 - BYO
온-프레미스 SQL에서 데이터 읽기 미리보기	아닙니다.	예
실행/성능	단일 노드	다중 노드
프로덕션 웹 API	아닙니다.	예
SLA	아닙니다.	예

시간당 요금은 실제 서비스 사용에만 적용됩니다. 동시에 적용된 여러 미터가 표시됩니다.


문제 정의
데이터 셋 준비
모델 설정
모델 훈련 / 평가
모델 활용
총 111개 모듈


ThinkPad

감사합니다

Thank you~!