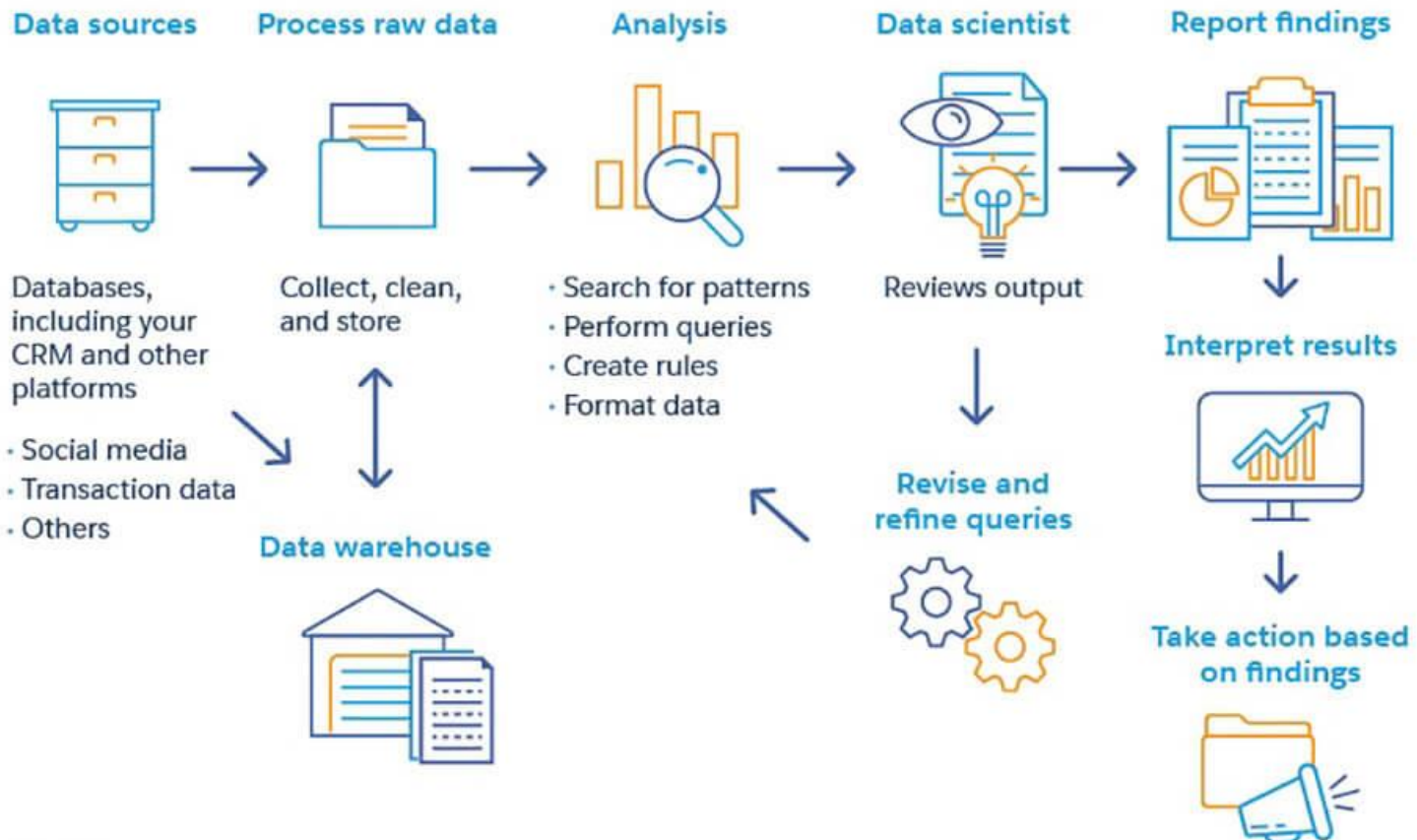


# Data Analysis Process



# How Data Mining Works



SOURCE:

[slideshare.net/PowerPoint-Templates/data-mining-process-powerpoint-presentation-templates](https://slideshare.net/PowerPoint-Templates/data-mining-process-powerpoint-presentation-templates)



# How Predictive Analytics Works

Collect data



Clean data



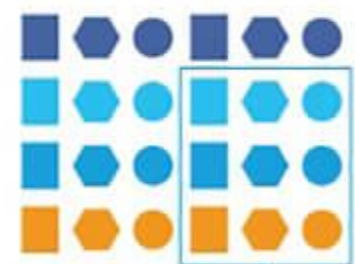
hindsight

Identify patterns



insight

Make predictions

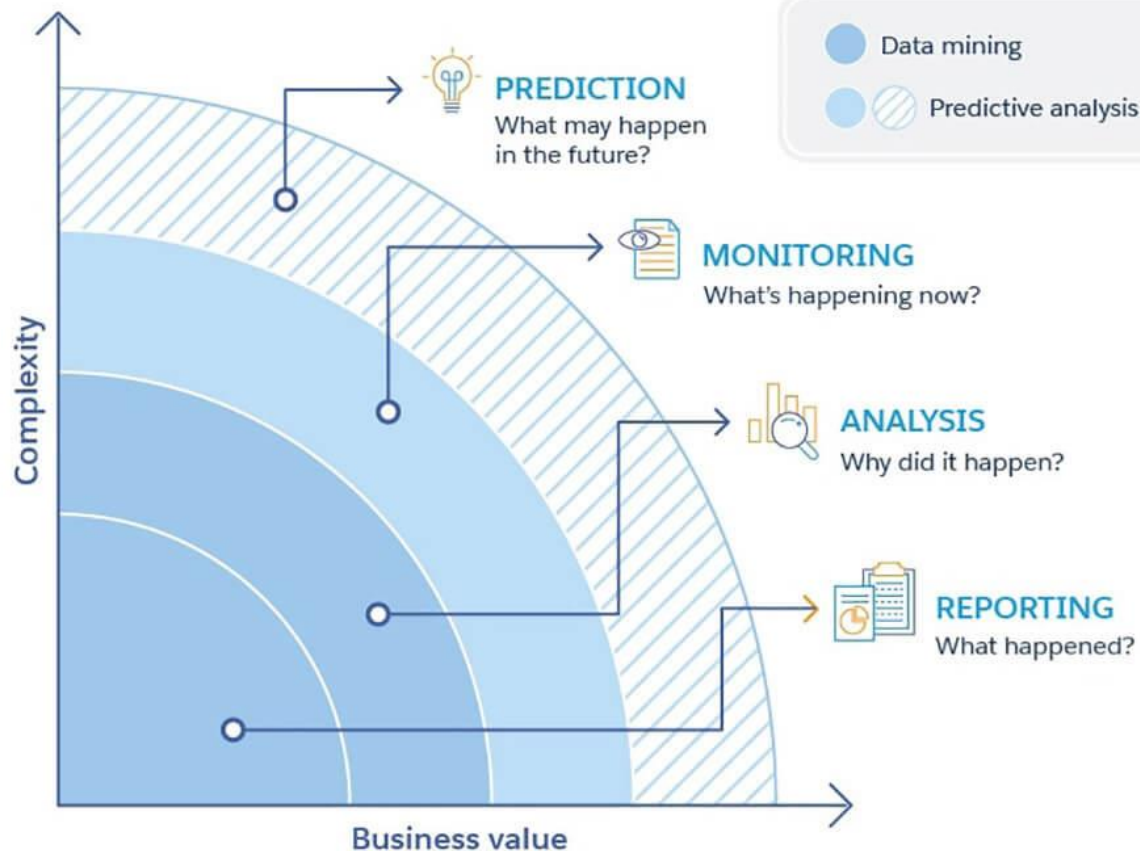


foresight

SOURCE: amadeus.com



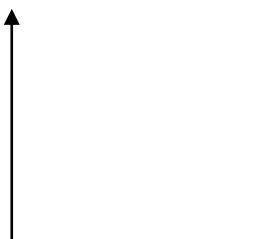
# How Data Mining and Predictive Analytics Work Together



# 데이터 포인트

- 현상을 관측한 단위

- Point (포인트)
- Sample (샘플)
- Instance (인스턴스)
- Record (레코드)
- Observation (관측치)
- Vector (벡터)




id	$X_1$	$X_2$	...	$X_p$	$Y$
1	$x_{11}$	$x_{12}$	...	$x_{1,p}$	$y_1$
2	$x_{21}$	$x_{22}$	...	$x_{2,p}$	$y_2$
...	...	...	...	...	...
$n$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	$y_n$

- 현상들을 설명/표현하는 요소
- Variable, Feature, Attribute, Factor, Field, Column, ...

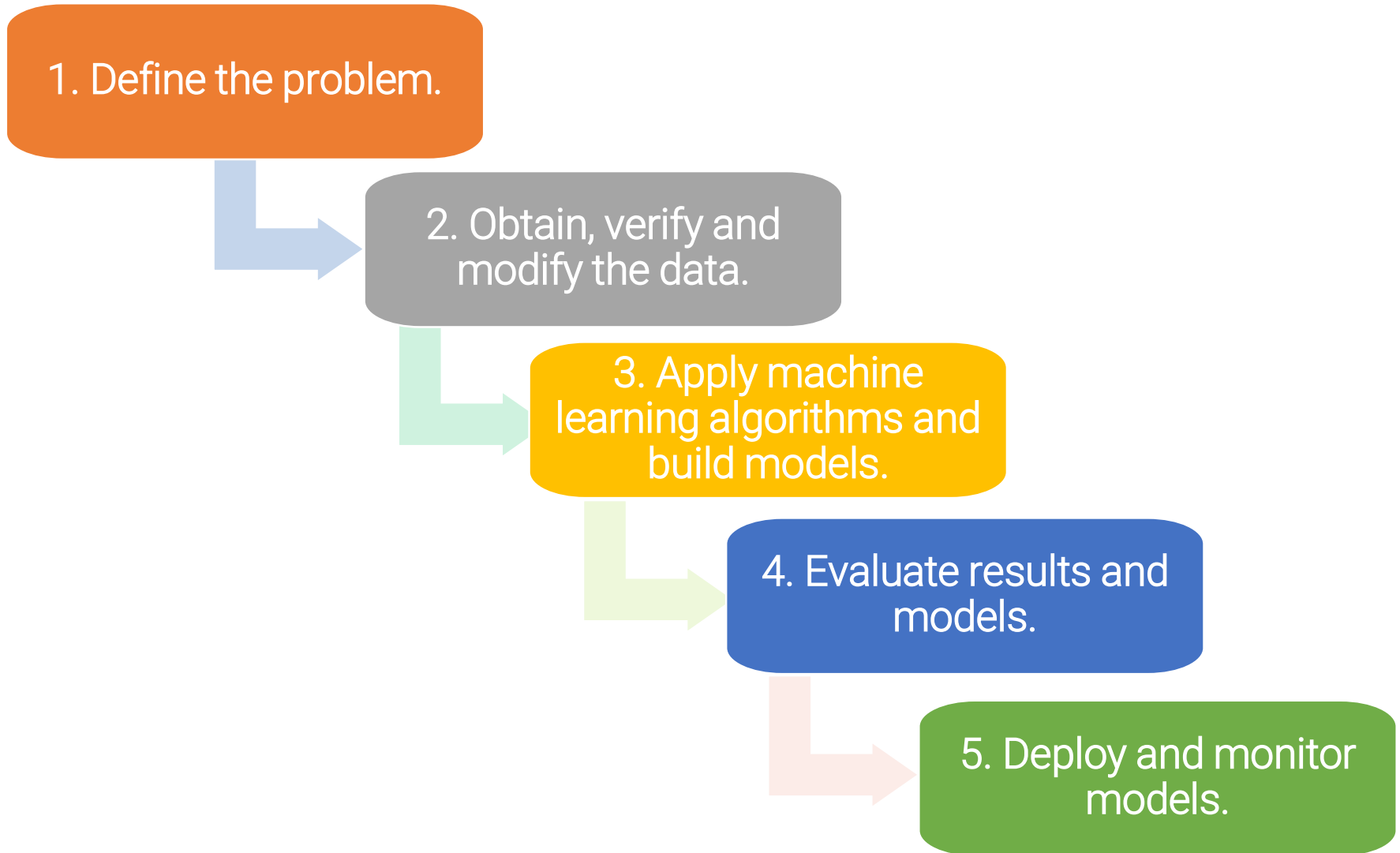
- Predictor variables (예측변수)
- Input variables (입력변수)
- Independent variables (독립변수)

- Target variables (타겟변수)
- Output variables (출력변수)
- Dependent variables (종속변수)



id	$X_1$	$X_2$	...	$X_p$	$Y$
1	$x_{11}$	$x_{12}$	...	$x_{1,p}$	$y_1$
2	$x_{21}$	$x_{22}$	...	$x_{2,p}$	$y_2$
...	...	...	...	...	...
$n$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	$y_n$

# Steps in the machine learning workflow



1. Define the problem.

2. Obtain, verify and  
modify the data.

3. Apply machine  
learning algorithms and  
build models.

4. Evaluate results and  
models.

5. Deploy and monitor  
models.

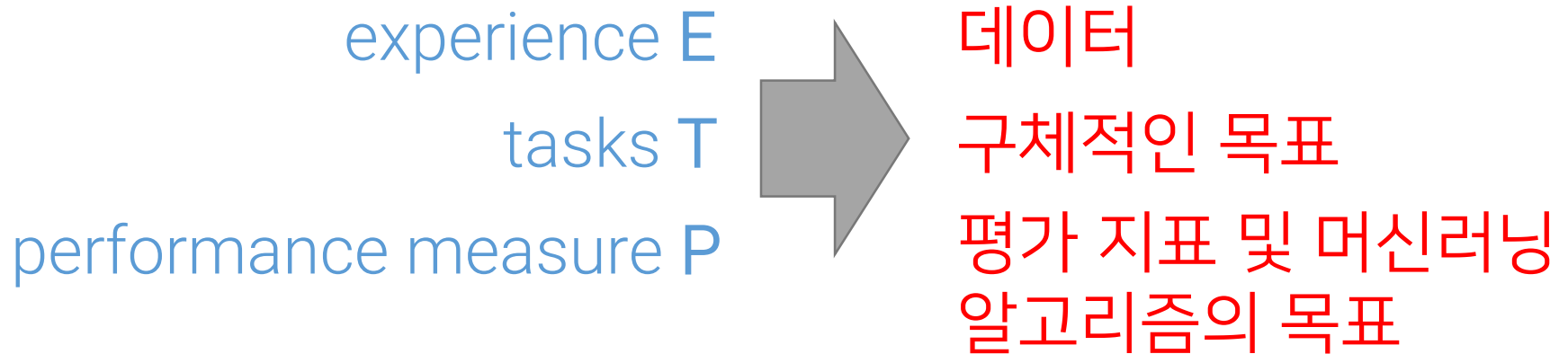


# 1. Define the problem

## : Data, Task, and Performance measure

- What is learning?

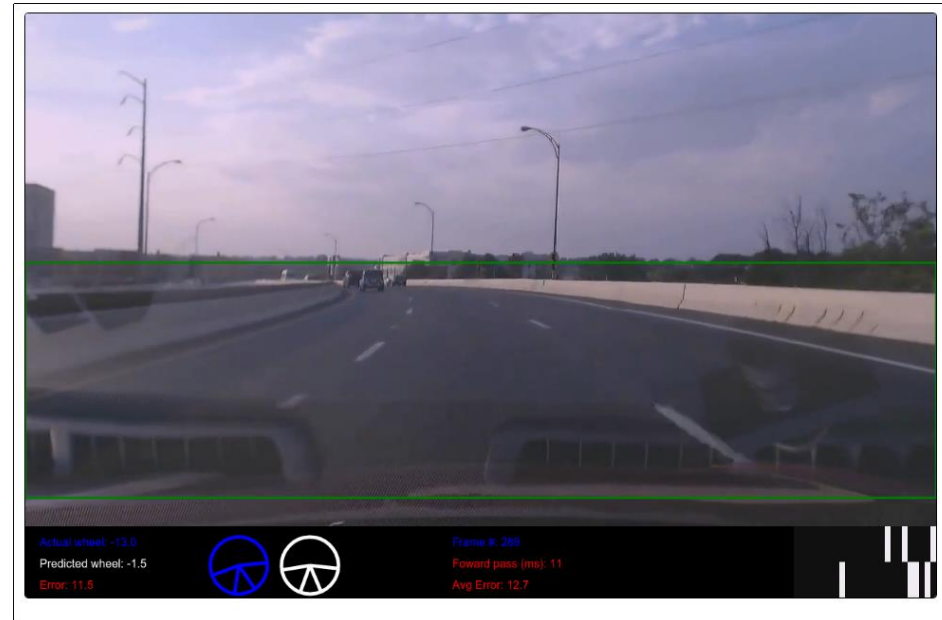
➔ A computer program is said to *learn* from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$  (Mitchell, 1997).



# What is learning?

## Example 1: Self-driving car

- task  $T$ ?
- performance  $P$ ?
- experience  $E$ ?



# What is learning?

## Example 2: AI for chess or go

- task T?
- performance P?
- experience E?

World Chess Champion Garry Kasparov playing IBM AI Deep Blue in 1996. Kasparov won this first match but lost the rematch a year later.



# What is learning?

## Example 3: Predicting Product Quality at the Factory

- task T?
- performance P?
- experience E?



# What is learning?

## Example 4: Churn detection

- task T?
- performance P?
- experience E?

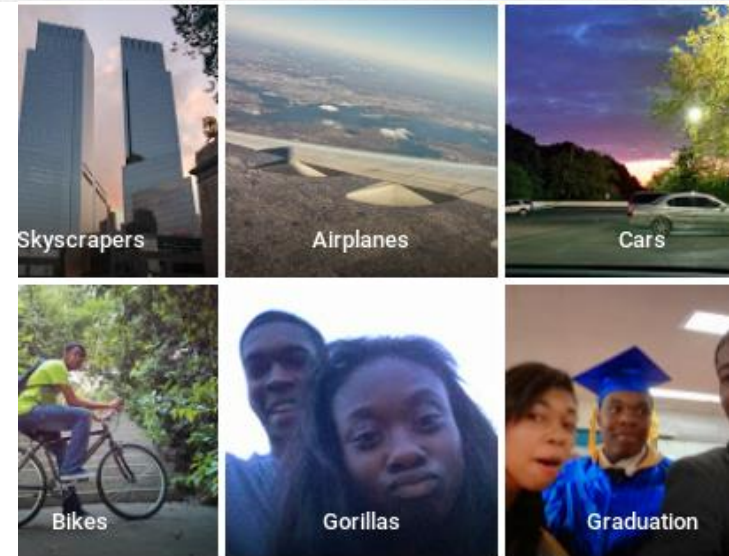
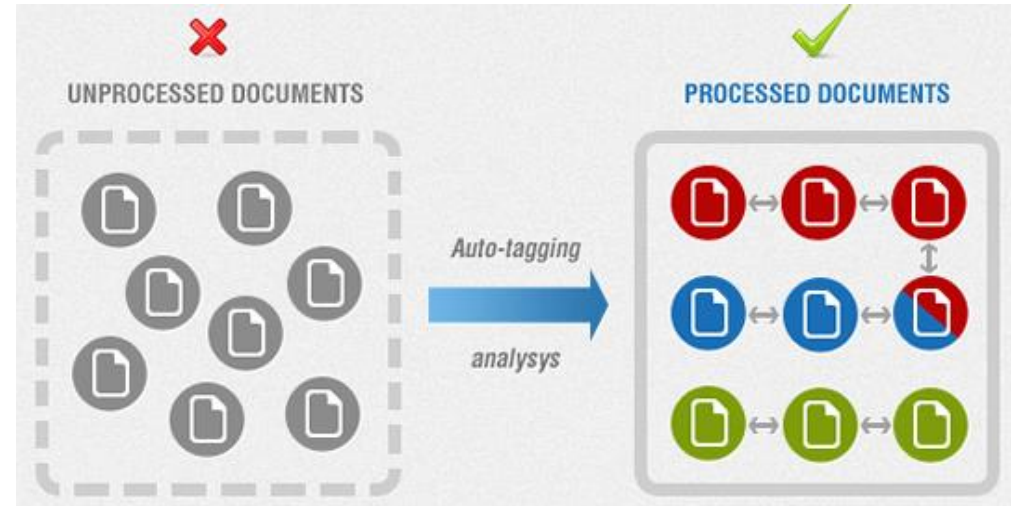




# What is learning?

## Example 5: Auto-tagging documents or photos

- task T?
- performance P?
- experience E?



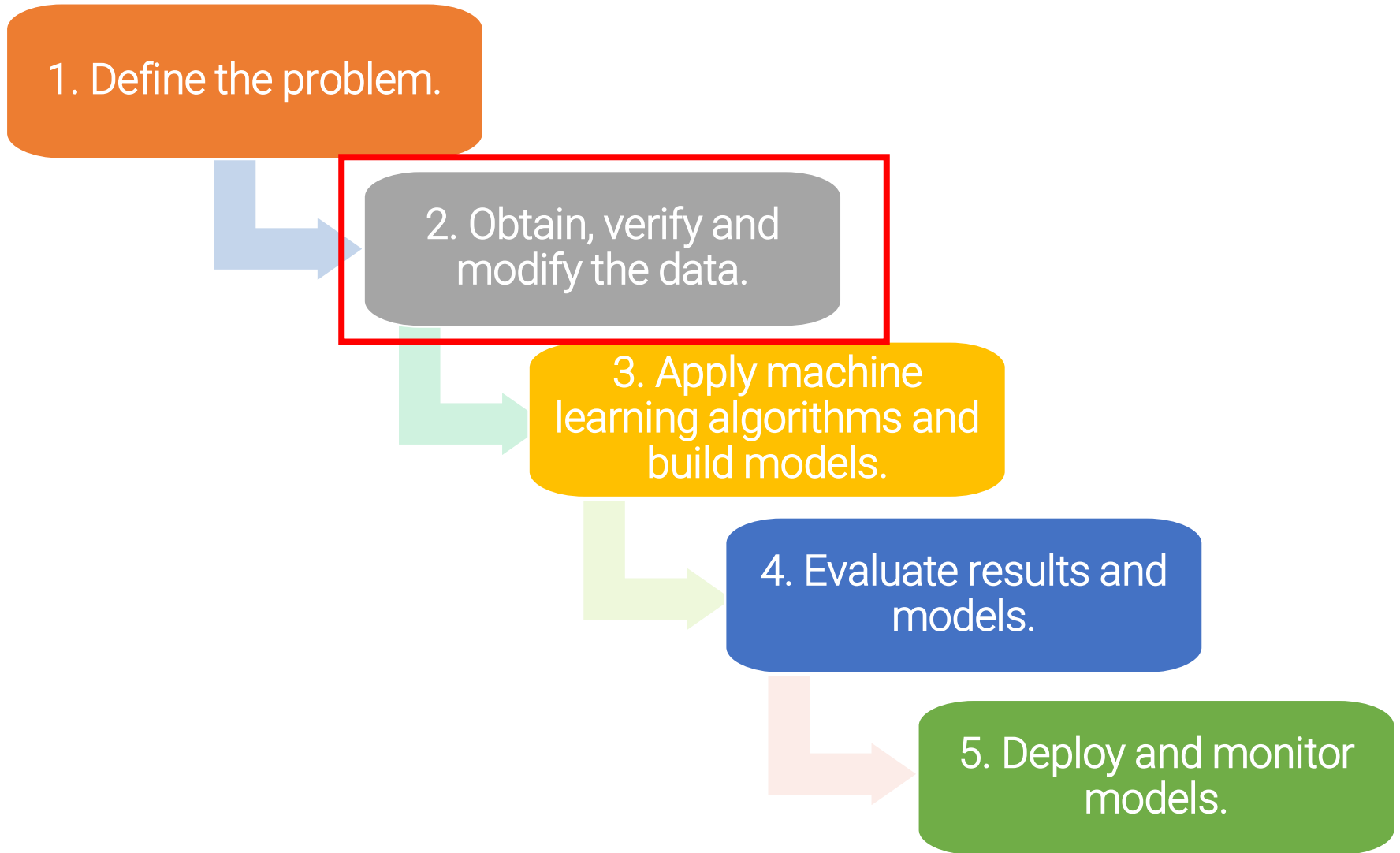
# 1. Define the problem

- 해결하고자 하는 문제를 정확히 정의하자.
- 정의된 문제와 관련이 있는 데이터가 있는지 파악하자.
  - 내부에 존재하는가?
  - 외부에 존재하는가?
- 무엇을 예측하고자 하는지, 아니면 현상을 잘 설명/표현하는 것이 중요한지 파악하자.
  - Machine learning으로 풀어야 하는 문제인가? 인공지능이 필요한가? 아니면 데이터 시각화로 풀 수 있는 문제인가?
  - Supervised learning? ( $Y=f(X)$ ) / Unsupervised learning → ( $p(X)$ )

# Problem Formulation의 중요성

- 데이터 분석은 많이 대중화가 되었고, 이제 여러 분야의 사람들이 자신만의 영역에서 데이터 분석을 수행 중
- 많은 경우,
  - 자신들이 갖고 있는 데이터로 무엇을 할 수 있는지 잘 모름
  - 자신들이 원하는 정보를 추출하기 위해 어떠한 분석 방법을 사용해야 할 지 잘 모름
- 따라서, 해당 필드의 문제를 데이터로 접근하여 풀 수 있도록 문제를 잘 정의하고 구조를 세우는 니즈가 증가할 것임





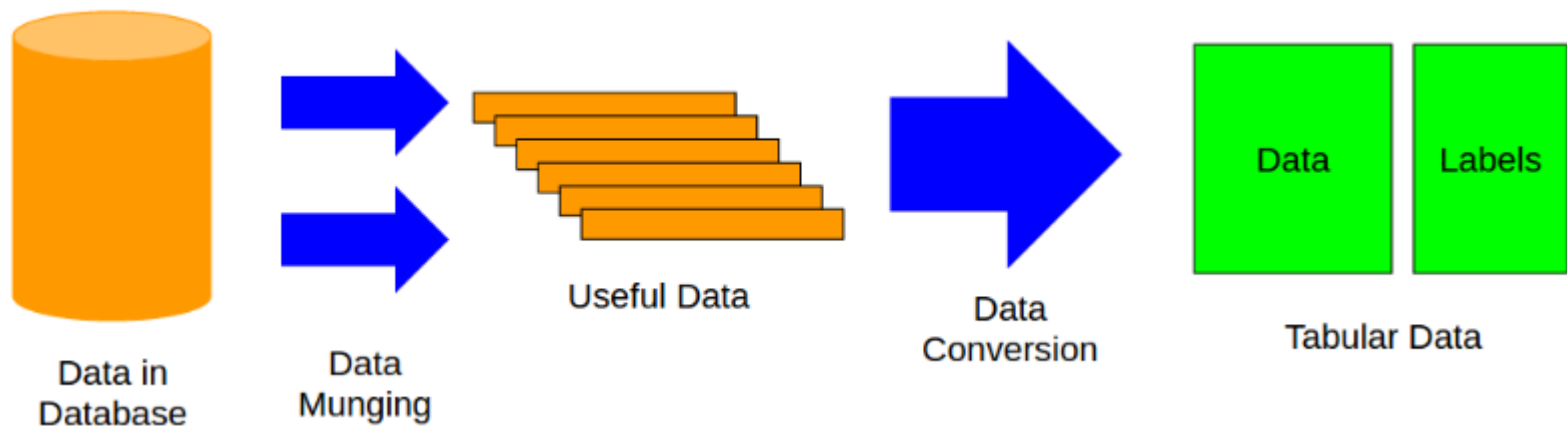
## 2. Obtain and verify the data

- 정의된 문제와 관련이 있는 데이터가 있는지 파악하자.
  - 내부에 존재하는가?
  - 외부에 존재하는가?
- Supervised learning? ( $Y=f(X)$ ) / Unsupervised learning  $\rightarrow$  ( $p(X)$ )

**Y와 X를 찾자!**

# Converting the data to an analyzable form

- Before applying the machine learning models, the data must be converted to a tabular form. This whole process is the most time consuming and difficult process and is depicted in the figure below.
- Tabular data is most common way of representing data in machine learning or data mining.



# Quality of Data

- 데이터의 품질이 좋아야 뭐라도 나온다.

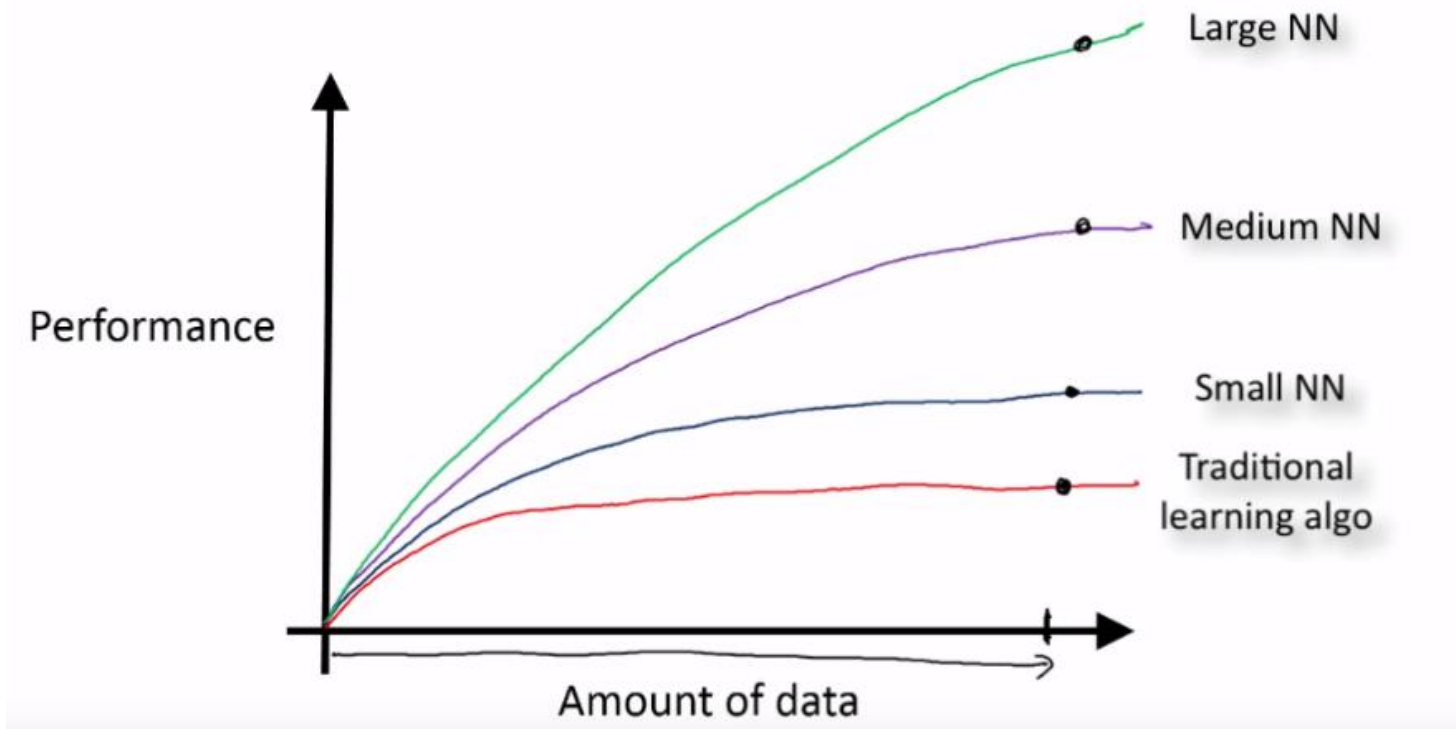


<https://www.linkedin.com/pulse/20140922000317-25059308-garbage-in-garbage-out/>  
<https://kerriknox.liberty.me/anarchy-during-the-gold-rush-from-eyewitnesses/>

# Quantity of Data

- 가능한 한 많은 데이터를 확보하는 것이 좋다.

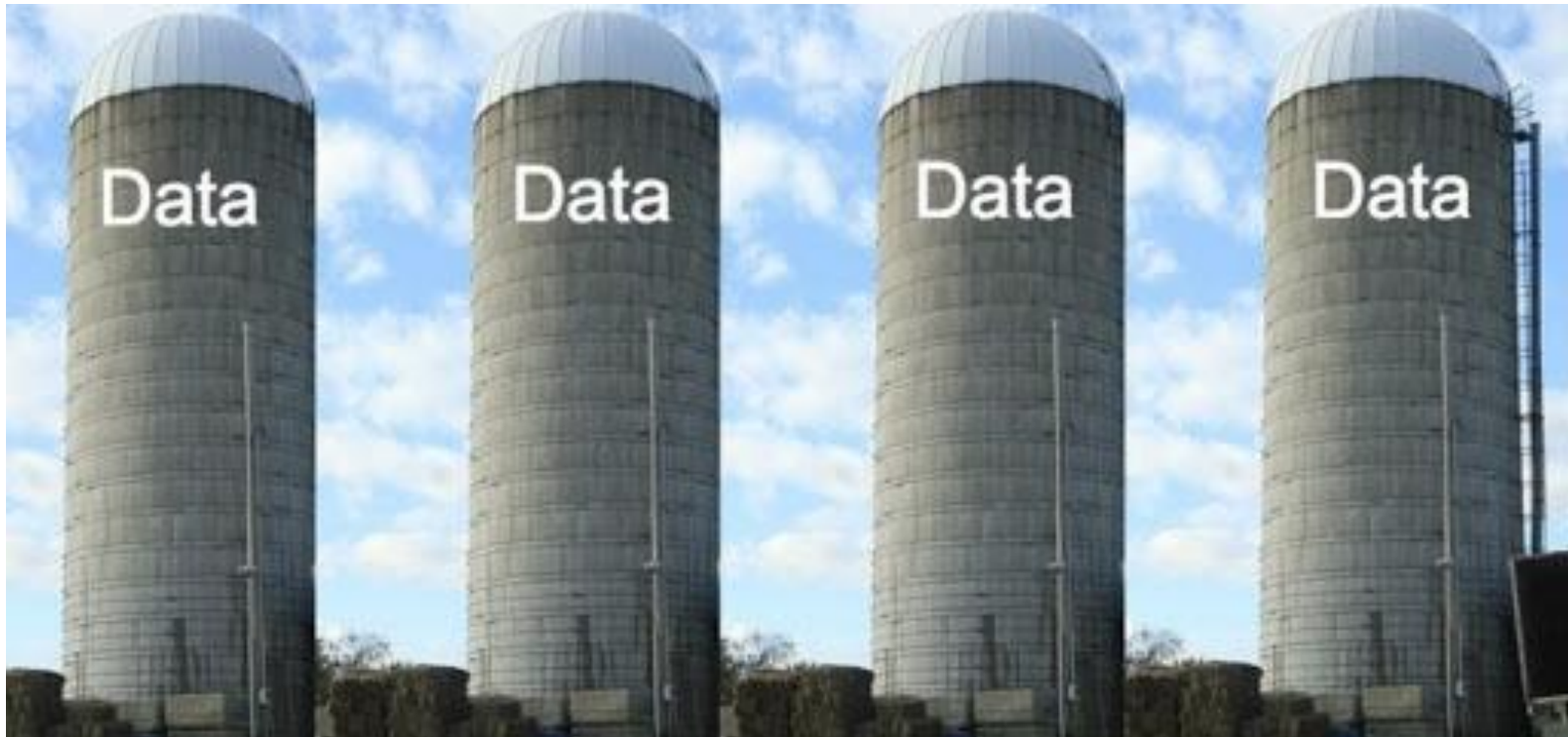
One picture explaining the rise of Deep Learning



a lecture slide made by Andrew Ng

# Breaking data silos

- How to connect and integrate the data



# Modifying the data → Preprocessing

- **Outlier**

- “A value that the variable cannot have” or “An extremely rare value” (ex: age 990, height -150cm, ...)
- There are a number of outliers in a real database due to many reasons.

- **How to deal with outliers?**

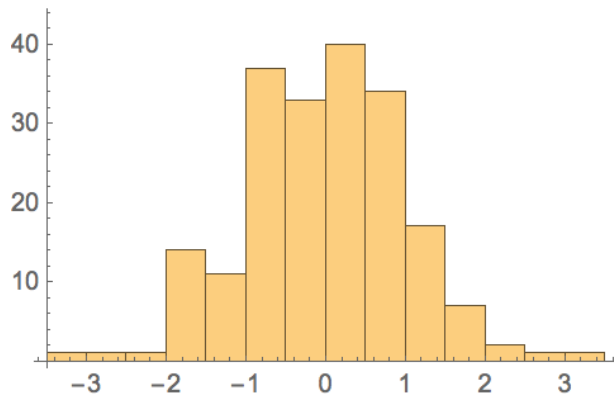
- Ignore the record with outliers if total record is sufficient.
- Replace with another value (mean, median, estimate from a certain pdf, etc) if total records are insufficient.

# Modifying the data → Preprocessing

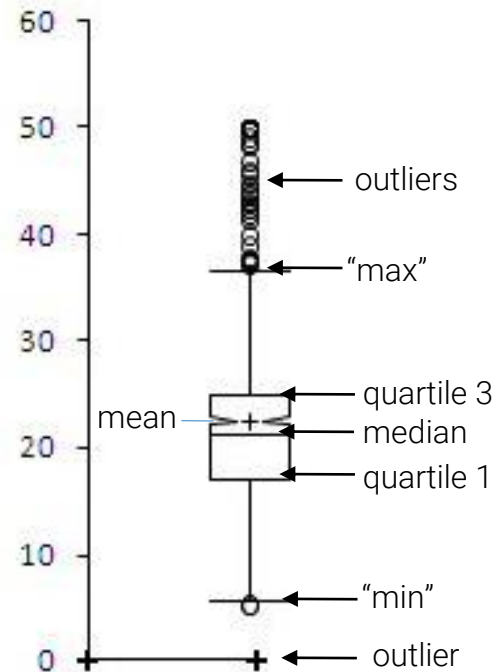
- 데이터 시각화 및 탐색을 통해 outlier를 판단

## Histogram:

- shows the distribution of a single variable.
- possible to check the normality.



## Box plot





# Modifying the data → Preprocessing

- Missing value (결측치)
  - A variable is missing when it has null value in database although it should have a certain real value.
  - Operational errors, human errors.
- How to deal with missing values?
  - Ignore the record with missing values if total record is sufficient.
  - Replace with another value (mean, median, estimate from a certain pdf, etc) if total records are insufficient.

# Modifying the data → Feature engineering

- Feature engineering

- Raw data 혹은 그 이후 단계의 data를 더 나은 표현으로 바꾸기 위한 변수를 생성/추출/변환하는 과정
- 데이터의 표현, 즉 변수들이 어떻게 구성되느냐에 따라 머신러닝 모델 성능에 엄청난 영향을 미치므로, feature engineering은 매우 중요

- Types

- Feature transformation (변수 변환) and generation (생성)
- Feature (subset) selection (변수 선택)
- Feature extraction (변수 추출)

# Modifying the data → Feature engineering

- Type of variables: Quantitative variable

- 많고 적음을 나타내는 수치로 된 자료
- 사칙 연산 가능

## 계수형/이산형 (Count/Discrete)

- 셀 수 있는 정수의 형태
- 형제 수, 보험 가입 건 수 등

## 연속형 (Continuous)

- 셀 수 없는 소수점을 포함
- 키, 무게, 길이 등

## 구간형 (Interval)

- 차이만 의미가 있음
- 온도: 20도는 10도보다 2배 뜨겁다(X)

## 비율형 (Ratio)

- 차이와 비율이 모두 의미가 있음
- 20kg은 10kg보다 2배 무겁다 (O)

# Modifying the data → Feature engineering

- Nominal (명목형)

- 계절 (봄, 여름, 가을, 겨울), 지역 (서울, 대전, 대구, ...)
- 1-of-C coding, one-hot encoding 을 수행한다.  
(1 nominal variable → C binary dummy variables)

Season		d1	d2	d3	d4
spring		1	0	0	0
summer		0	1	0	0
fall		0	0	1	0
winter		0	0	0	1

- Ordinal (순서형)

- 설문항 (매우 나쁘다, 나쁘다, 보통이다, 좋다, 아주 좋다),  
습도 (낮다, 보통, 높다), ...
- 각 수준에 알맞은 숫자를 대입
  - 예1) 매우 나쁘다 → 1, 나쁘다 → 2, 보통이다 → 3, 좋다 → 4, 아주 좋다 → 5
  - 예2) 낮다 → 0, 보통 → 0.5, 높다 → 1

# Modifying the data → Feature engineering

- Normalization (Standardization)
  - Eliminate the effect caused by different measurement scale or unit
  - z-score:  $(\text{value} - \text{mean}) / (\text{standard deviation})$

Original data

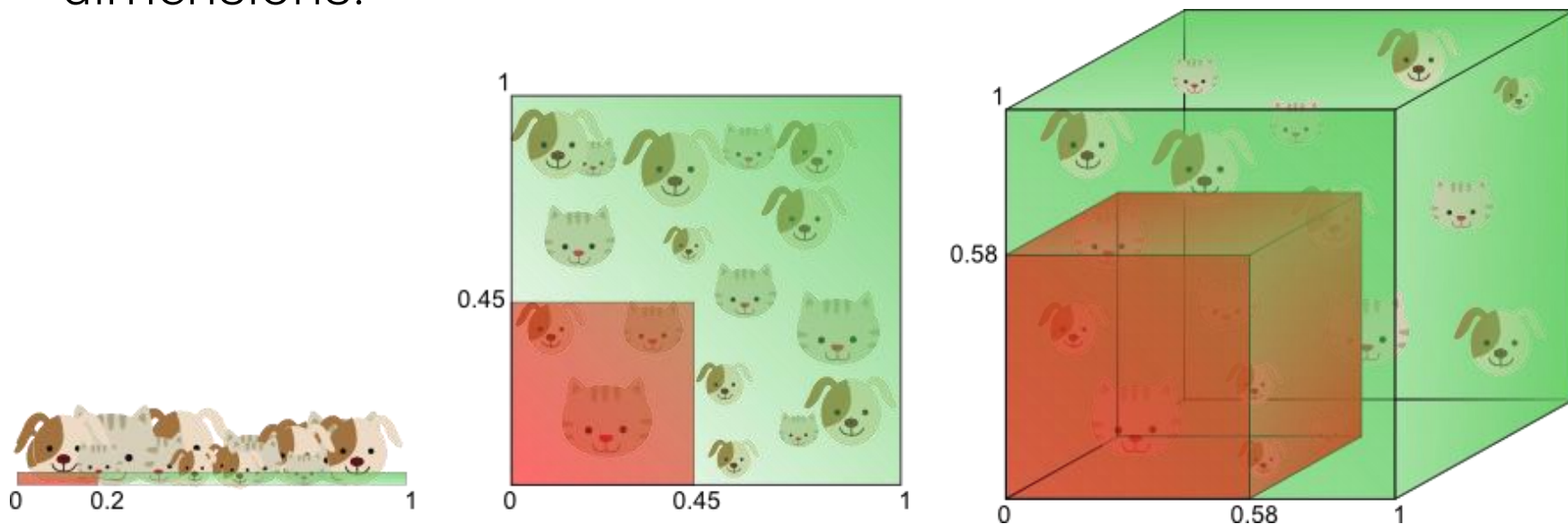
Id	Age	Income
1	25	1,000,000
2	35	2,000,000
3	45	3,000,000
...	...	...
Mean	35	2,000,000
Stdev	5	1,000,000

Normalized data

Id	Age	Income
1	-2	-1
2	0	0
3	2	1
...	...	...
Mean	0	0
Stdev	1	1

# Curse of dimensionality

- The amount of training data needed to cover 20% of the feature range grows exponentially with the number of dimensions.



# Privacy, Security, Regulation

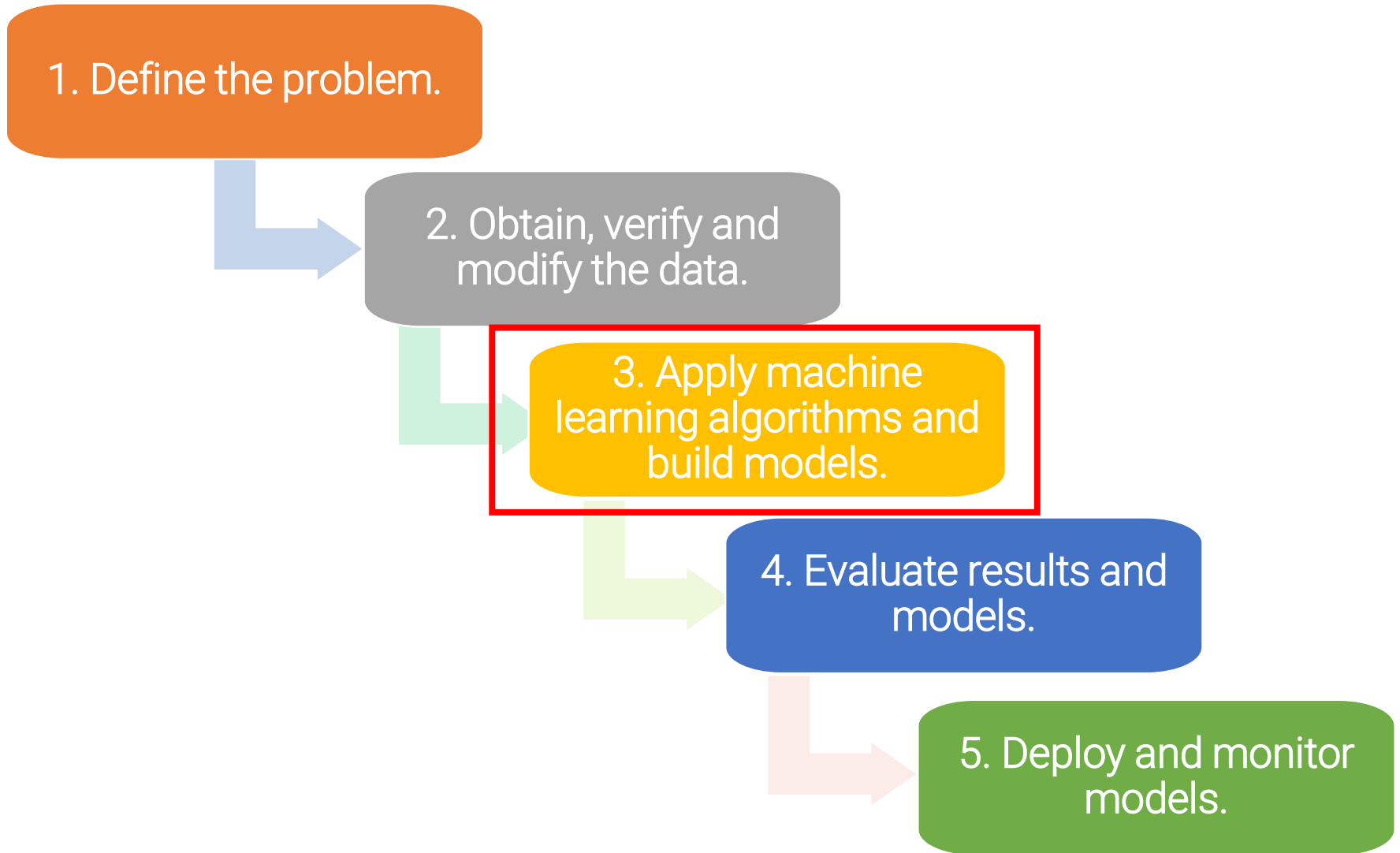


## 2. Obtain, verify and modify the data.

데이터 획득과 검증에 많은 공을 들여야 한다.

- 가능한 한 많은 수의 데이터를 얻어라.
- 만약 예측 모델링을 원한다면 클래스를 같이 확보해라.
- 내가 앞서 설정한 목표 (purpose) 와 구체적인 태스크 (task) 에 데이터가 적합한지 판단해라.
- 데이터 내 분포 (혹은 데이터 생성 배경) 가 일관적인지 판단해라.





### 3. Apply machine learning algorithms and build models.

- Models

- Classification

- Logistic regression, k-nearest neighbor, naïve bayes, classification trees, neural networks, linear discriminant analysis

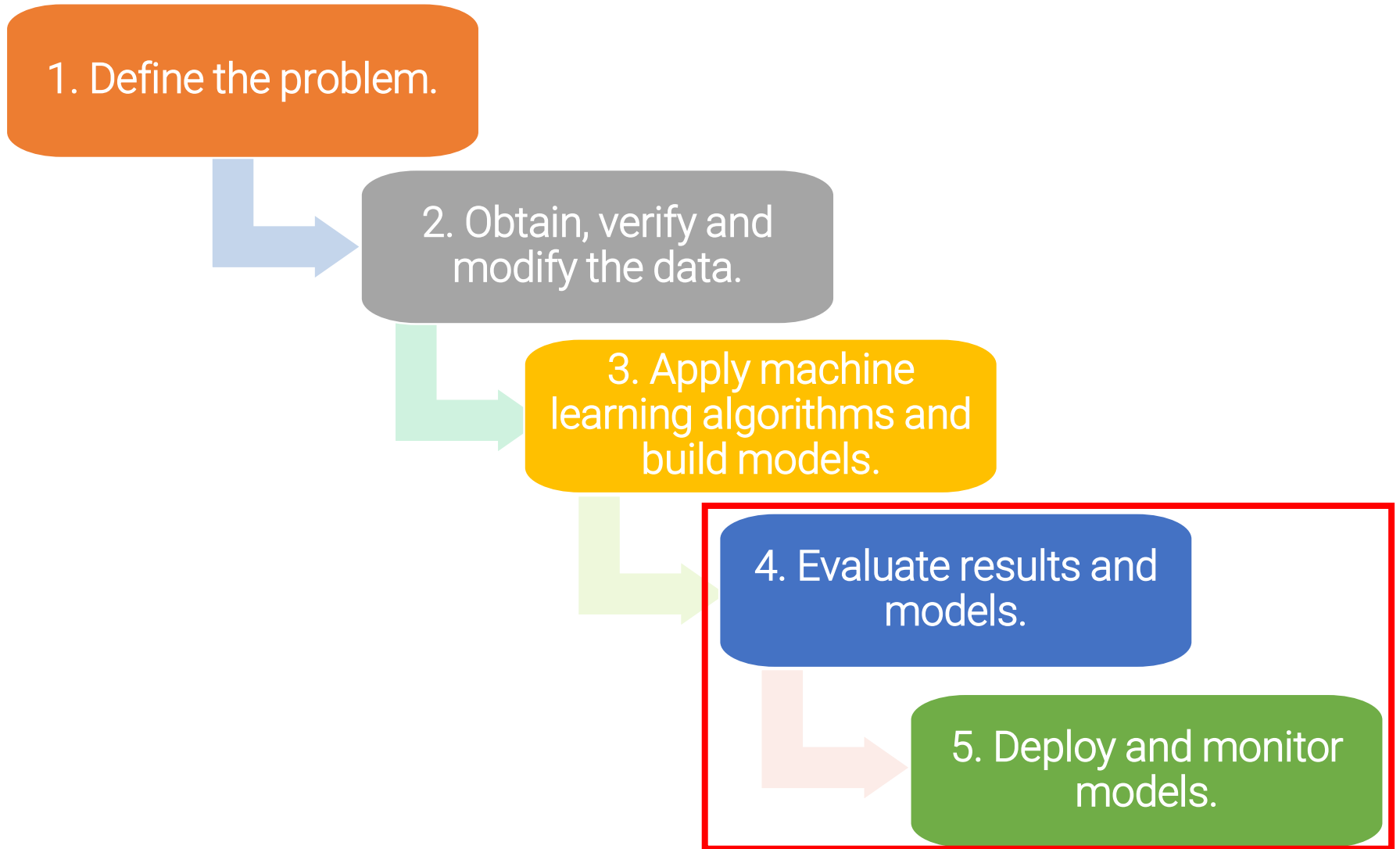
- Regression

- Linear regression, k-nearest neighbor, regression trees, neural networks

- Clustering

- Hierarchical clustering, K-Means clustering

**각 모델과 이에 대한 학습법이 존재한다.**



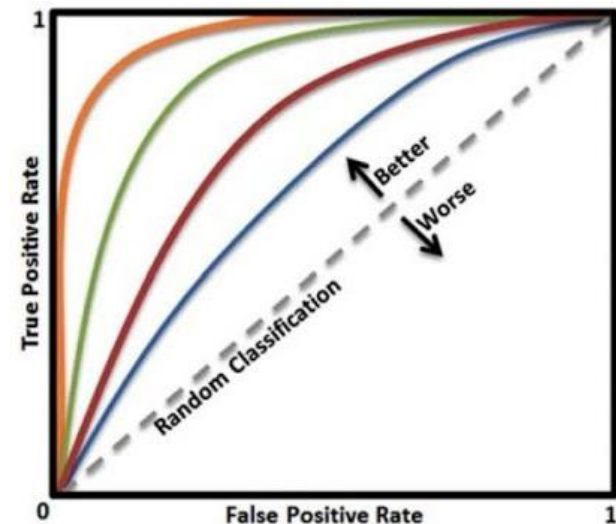
## 4. 결과와 모델 평가

- 모델 평가를 통해, 나의 문제에 가장 적합한 모델이 무엇인지 찾을 수 있음.
- 모델의 평가지표는 곧 머신러닝 알고리즘의 목표 함수가 되기도 한다.

Confusion matrix

		Predicted	
		1(+)	0(-)
Actual	1(+)	True positive, Sensitivity (A)	False negative, Type I error (B)
	0(-)	False positive, Type II error (C)	True negative, Specificity (D)

ROC curve



## 5. 모델 적용 및 모니터링

- **만든 모델을 실제 비즈니스 프로세스에 적용**
  - 현업에서 모델을 실제로 업무에 적용하는 것은 생각보다 많은 이슈를 야기할 수 있으며, 이를 위해 충분한 consensus를 생성하고 토의해야 함.

## 5. 모델 적용 및 모니터링

- 모델은 한 번 만들어지고 끝나는 것이 아님.
  - 데이터는 굉장히 '동적'이다. 외부 요인과 사람들의 인식 변화 등 여러 요인으로 데이터는 계속 살아 움직인다.
  - 내가 만든 모델이 언제까지 계속 쓸 수 있는지 모니터링
    - 기존 모델에 새로운 데이터를 적응시킬 수 있는가?
    - 다시 모델을 학습하여 새로운 모델을 생성할 것인가?