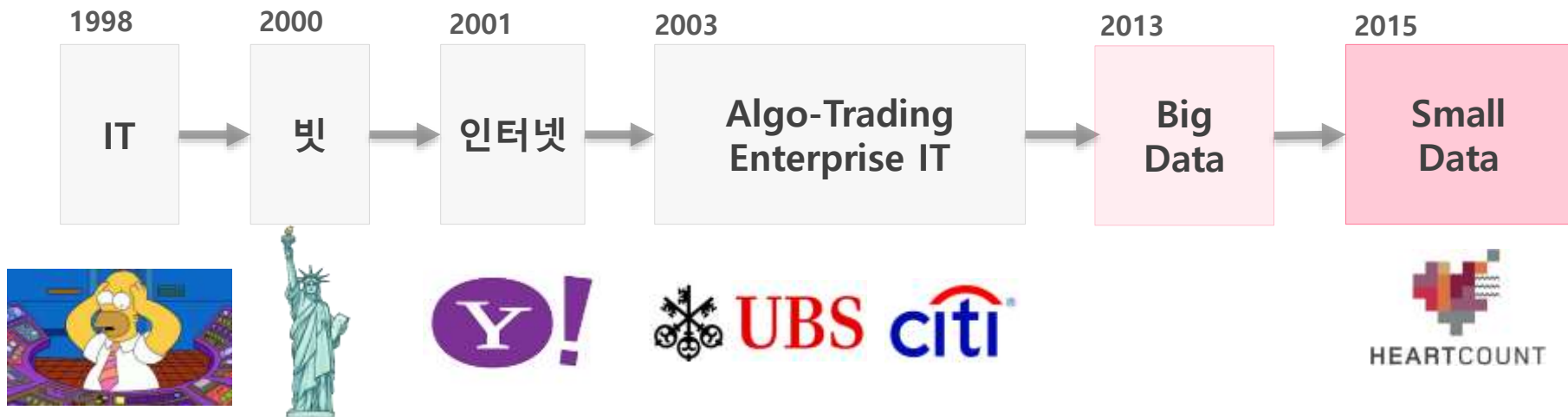


# From Data Literacy To Data Fluency

# 강사 소개

- 양승준: 아이디케이스퀘어드(IDK<sup>2</sup> - I Don't Know What I Don't Know) 대표
- 서비스: **HEARTCOUNT**, Augmented Analytics for Enterprise (SaaS)
- 경력



# 강의에서 다룰 내용

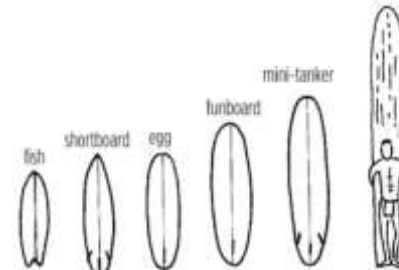
## MIND



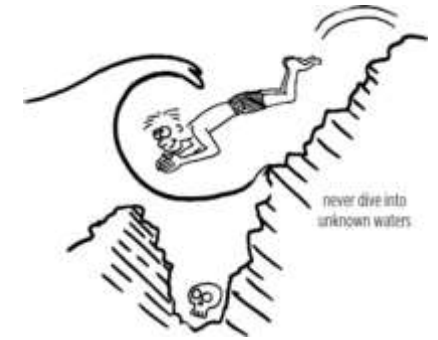
## DATA



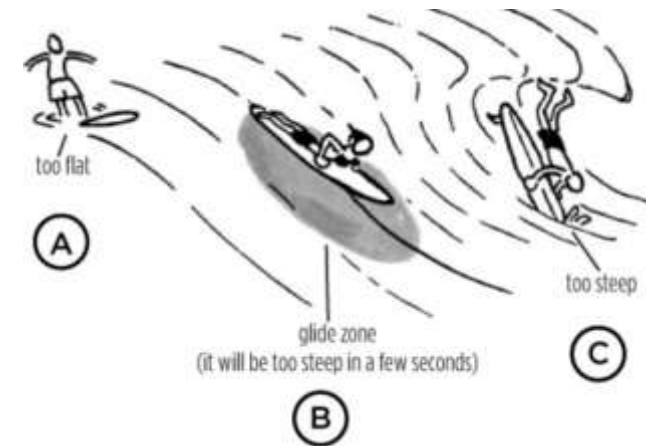
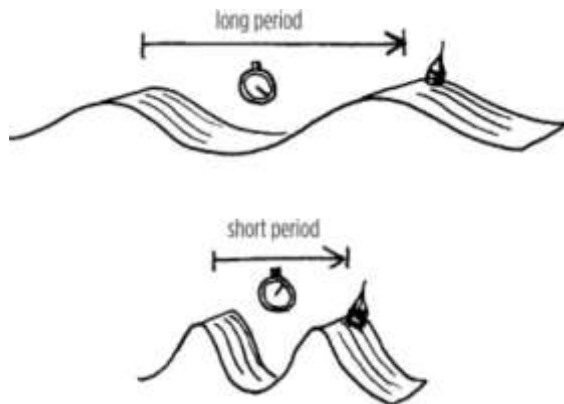
## TOOL



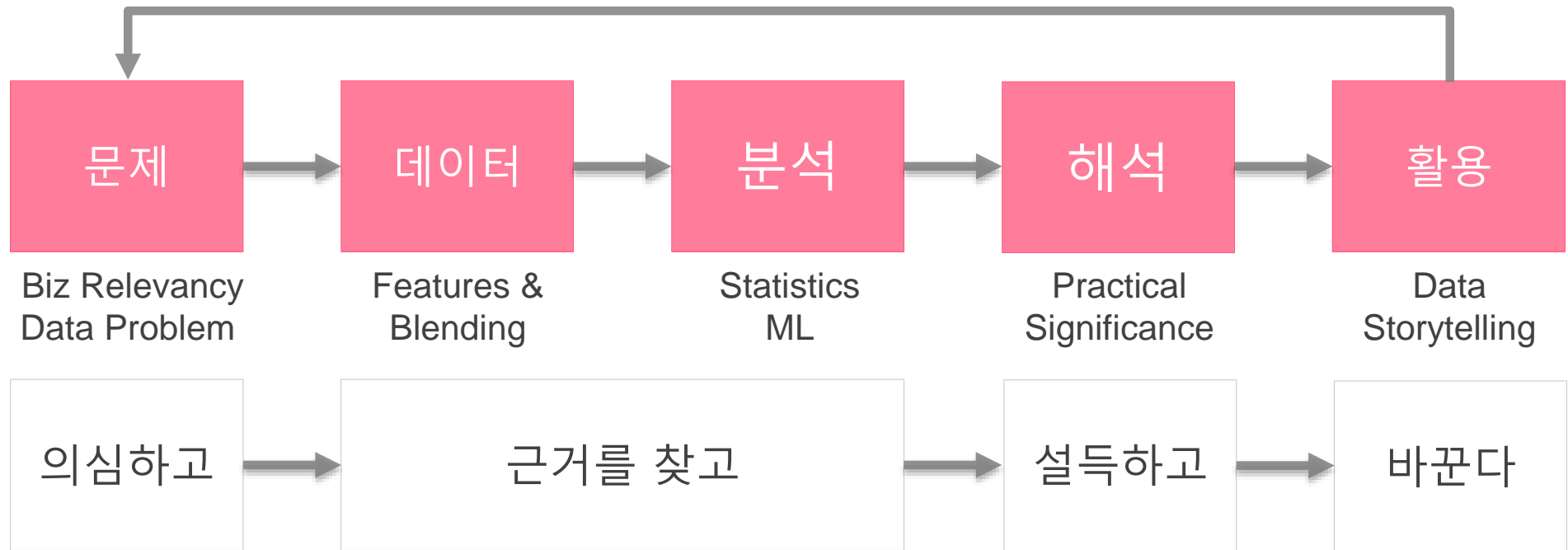
## Context



## BASIC SKILL

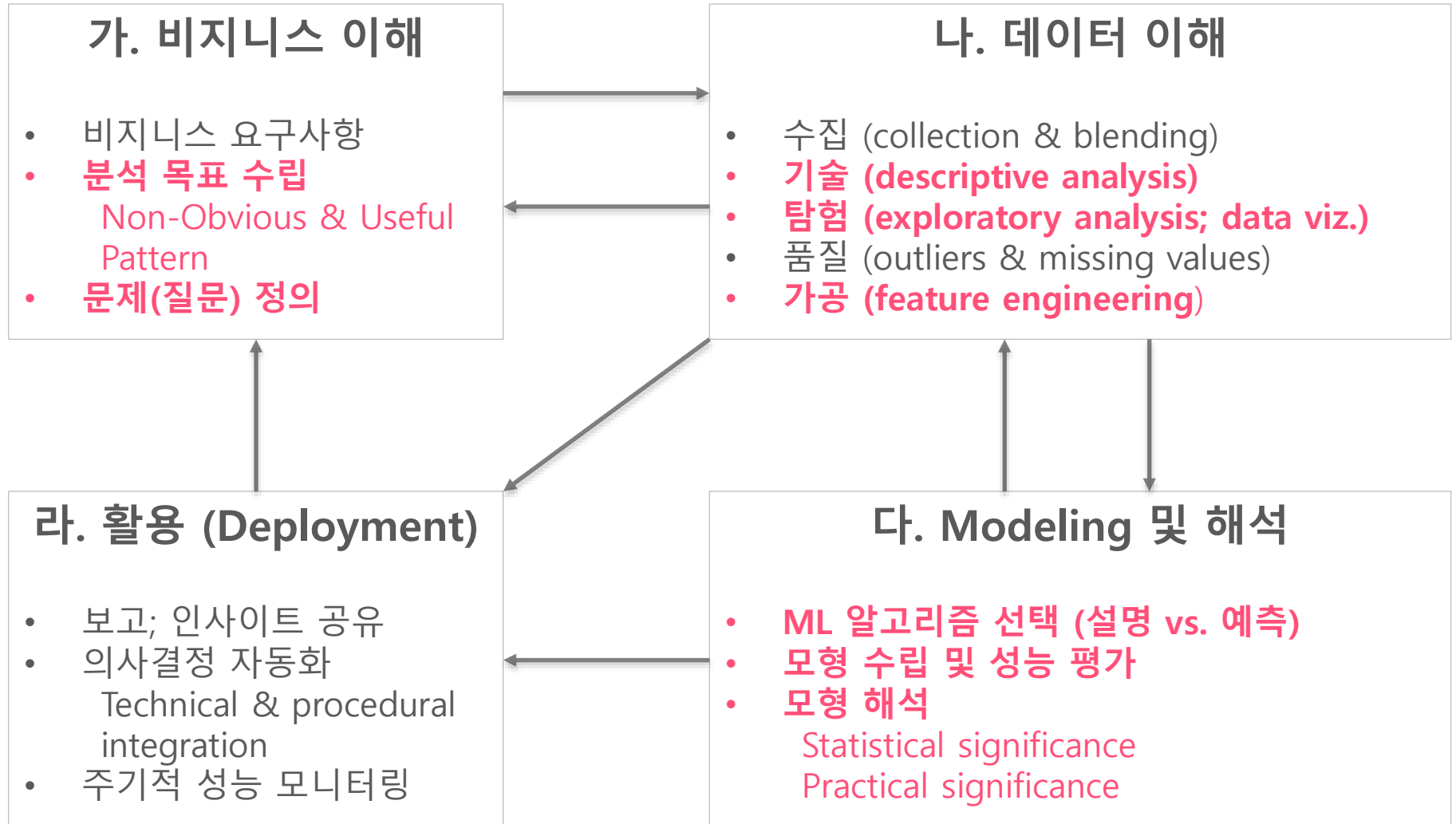


# 데이터 분석 과정



# 데이터 분석 과정 (\*CRISP-DM)

## \*강의 범위



데이터가 답할 수 있는 **질문**: 적절한 데이터와 분석방법  
Frame Real-World Problem into Data Problem



뻔하지 않은 쓸모있는 **패턴**  
Non-Obvious & Useful Pattern



피보고자가 분석결과 **수용(활용)**  
Audience Accepts the Results

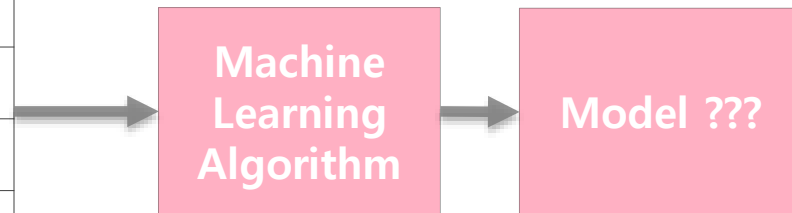
### Machine Learning is not a Magic, but a Math

주어진 데이터(번지수, 지붕 색상)가 문제(집값 예측)를 해결하는데  
사람에게 소용없다면 기계에게도 마찬가지  
사람이 풀 수도 있는 문제를  
기계가 더 빠르고 다양한 관점으로 냉정하게 풀게 하자.

---

Training Data Set

번지수 끝자리	지붕 색상	가격(억)
3	빨강	9.5
7	주황	3.5
2	노랑	3.2
1	빨강	3.5
4	파랑	4.7
.	.	.



## 폭력적 게임을 하면 더 폭력적이 되나?

- ① 개념(Concept)을 정량화하는 일의 주관성
  - 폭력적 게임을 정의할 수 있나?
  - 현실에서의 폭력성을 어떻게 계량화하나?
- ② 인과성을 증명할 수 있나?
  - 공격적 성향의 아이가 폭력적 게임에 더 끌림?

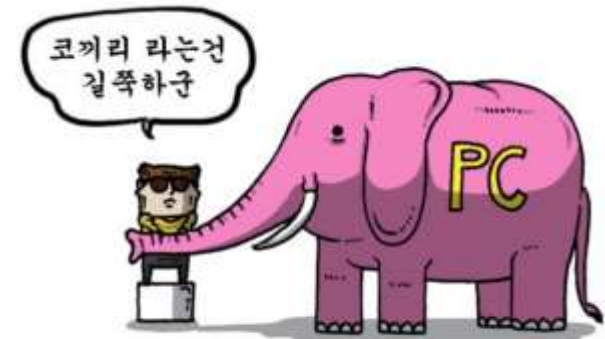
폭력적 게임



폭력적인 아이



- 통계적 상관관계로 인과성에 대해 주장할 수 없음
- 숫자로 환원된 분석 결과는 복잡한 진실에 대한 단면적 요약







### 정보량이 큰(뻔하지 않은) 분석결과

- A. 작년 여자 영업사원의 평균 매출이 남자 영업사원 평균 매출의 80% 정도였다.
- B. 작년에 신규 출시한 탈모 치료제의 경우 연구소 출신 여자 영업사원의 매출이 평균 매출의 350%에 달했다.

- A. 평균 직원 나이가 상위 10%에 속하는 매장들의 평균 인건비 지출이 전체 매장 평균 인건비 지출액보다 20% 높았다.
- B. 55세 이상인 직원이 한 명 이상 근무하는 매장은 그렇지 않은 매장과 비교해 매출은 15%, 고객 만족도는 25% 높았다.

- A. 내일 아침에 동쪽에서 해가 뜰 것이다. = 정보량 뺄임

### 정보 = 특정 질문에 대한 답변

- 뻔한 질문(예, 남녀 영업직원 간 매출 차이가 나는가?)에 대한 답변은 정보량도 낮음
- 엔트로피(불확실성)가 높은 불확실성이 큰 사안에 대한 질문(예, 탈모 치료제를 많이 판매한 직원들의 공통된 특성은 무엇인가?)에 대한 답변은 정보량이 큼

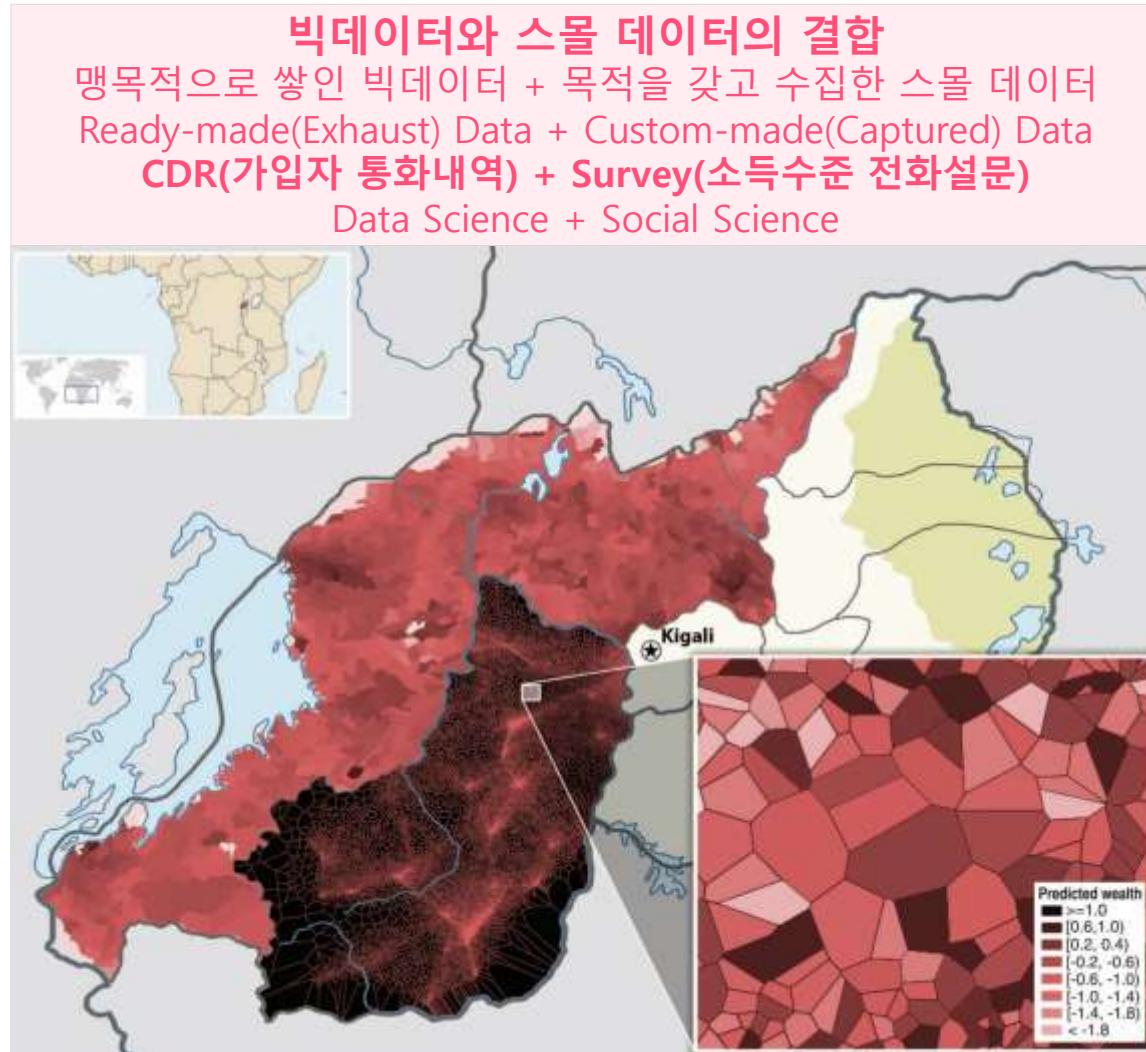


### Good Data Problem

- 작고 구체적인, 덜 뻔한(엔트로피가 큰) 질문
- 이미 확보하고 있거나 쉽게 수집 가능한 데이터에서 답을 찾을 수 있는 문제
  - 신뢰할 수 있고 익숙한 데이터에서 시작
  - 확보 가능한 데이터에 대한 이해 없이 문제부터 정하면 필요한 데이터를 추가 수집, 준비하느라 프로젝트 일정이 지연되거나 나쁜 분석 결과가 나오기 쉬움
- 바람직한 답변이 없고 너무 민감하지는 않은 문제
  - 바람직한 답변이 이미 마음 속이나 조직 내에 정해져 있는 경우
  - 특정한 분석 결과가 조직 내에서 너무 큰 반향과 혼돈을 불러올 수 있다면 분석 및 결과 해석 과정에서 객관성이 흔들릴 수 있음
- 본인 문제(내 호기심) 말고 비즈니스 문제
  - Business Top-Line(매출, 이익, 고객수, NPS, 생산성 등) Alignment

# 데이터에서 유용한 답을 찾을 수 있는 문제

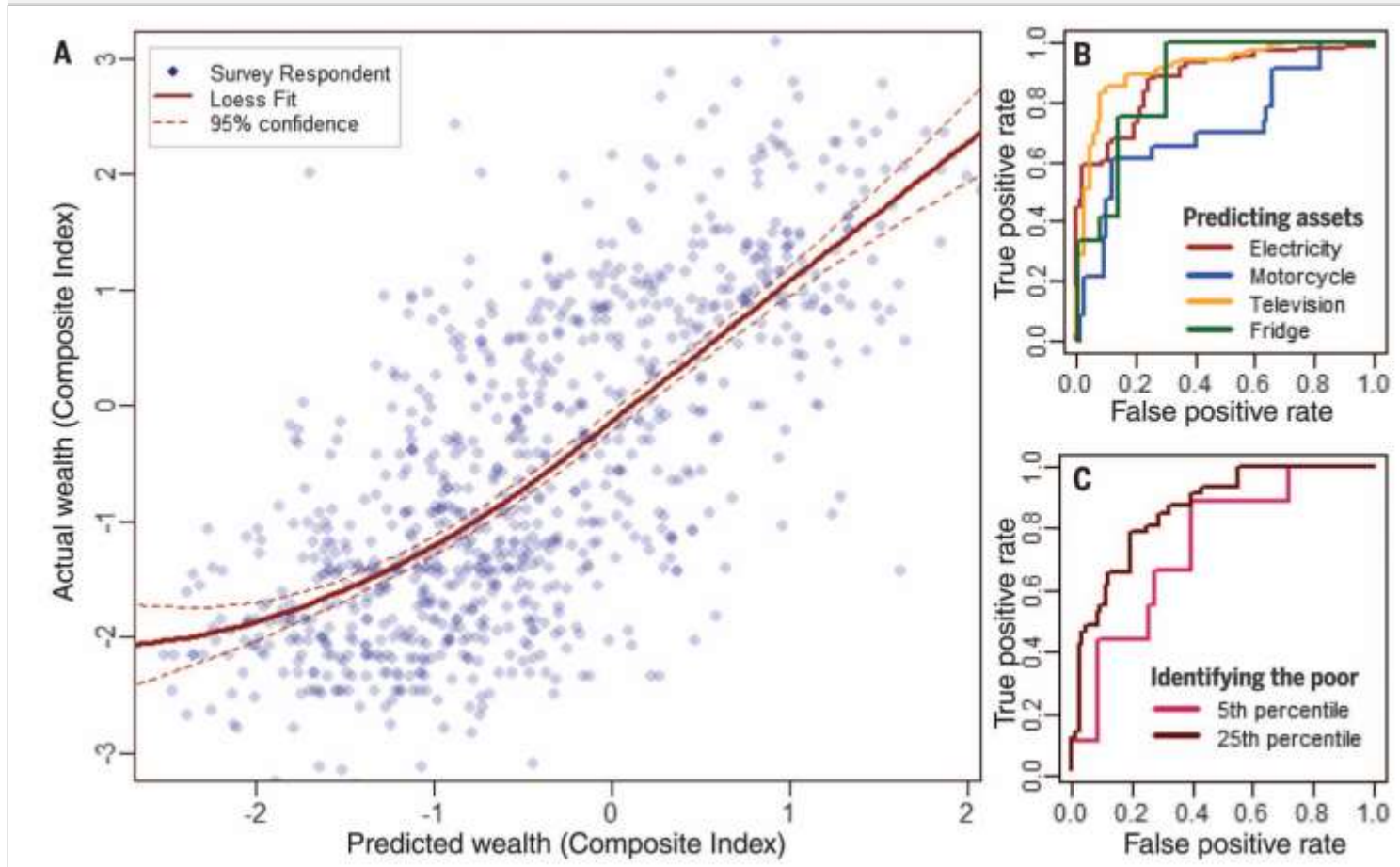
## 856명에게 전화걸어 르완다 전체 지역별 소득수준 분포 확인하기



## 데이터에서 유용한 답을 찾을 수 있는 문제 – cont'd

CDR(백오십만명 통화 내역) + Survey(856명 설문으로 소득수준 조사)  
→ 통화내역(X)만으로 소득수준(Y)을 예측하는 모형 생성

(서베이로 확인한) 실제 소득수준과 (통화 내역만으로) 예측된 소득수준



# 기업 내 데이터 분석 목표

본인 부서 Impact → Business Impact

문제 정의

패턴 발견

패턴 활용

1단계 현실을 객관적으로 이해	<ul style="list-style-type: none"><li>정량적으로 검증이 필요한 문제 정의</li></ul>	<ul style="list-style-type: none"><li>기존 믿음 검증</li><li>새로운 사실 확인</li></ul>	<ul style="list-style-type: none"><li>그릇된 통념 파괴</li><li>새로운 통찰의 공유</li></ul>
2단계 비즈니스 문제를 해결	<ul style="list-style-type: none"><li>본인 부서 말고 비즈니스 문제 정의</li></ul>	<ul style="list-style-type: none"><li>현실 적용이 가능한 패턴(Lever) 찾기</li></ul>	<ul style="list-style-type: none"><li>현실 적용</li><li>Success Metrics 지속적 모니터링</li></ul>

가치의 발견

가치의 완성

# 강의 목표 & 목차

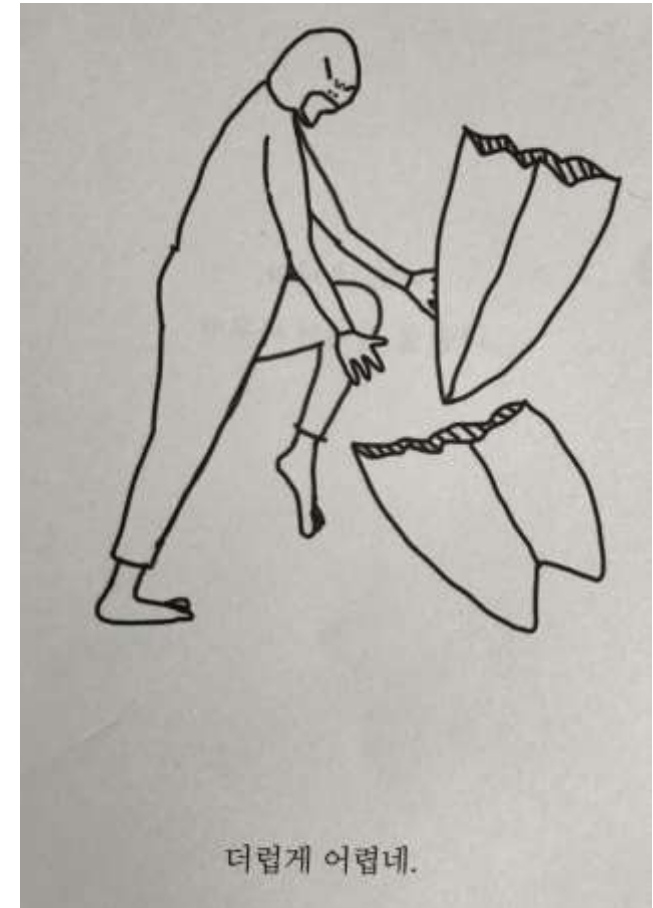
목표: 데이터 분석, 어렵지 않다. 유용하다.

## From Data Literacy to Data Fluency

- **Module I** Data Literacy
- **Module II** Data Understanding
- **Module III** ML and Decision-Making
- **Module IV** Linear Regression Analysis
- **Module V** Decision Tree Algorithm
- **My Two Cents** 참고자료



\*혼자서 공부할 내용



# Module I

## Data Literacy

아이디케이스퀘어드 양승준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)  
<https://www.heartcount.io>

# The Linda Problem

---

린다는 올해 31살로 아직 싱글이며 매우 솔직하고 총명한 여성이다. 철학을 전공했으며 학창시절 소수자 차별과 사회정의 문제에 깊은 관심을 가졌으며 비핵화 운동에도 활발하게 참여하였다.

**Q. 아래 두가지 문장 중 개연성이 더 높은 것은?**

- 1) 린다는 은행원이다.
- 2) 린다는 은행원이며 페미니스트 활동가로 활약하고 있다.



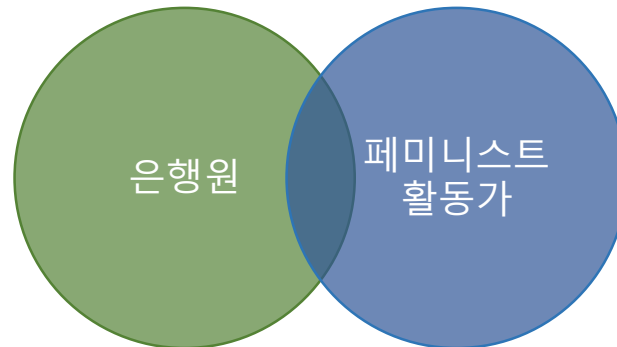
# The Linda Problem

## Plausibility vs. Probability

린다는 올해 31살로 아직 싱글이며 매우 솔직하고 총명한 여성이다. 철학을 전공했으며 학창시절 소수자 차별과 사회정의 문제에 깊은 관심을 가졌으며 비핵화 운동에도.

Q. 아래 두가지 문장 중 개연성이 더 높은 것은?

- 1) 린다는 은행원이다.
- 2) 린다는 은행원이며 페미니스트 활동가로 활약하고 있다.



## 합리성(Rationality)

효용(Utility)을 고려 최선의 방법을 선택  
완전한 정보 + 모든 가능성 고려



의사결정 잘 하고 있나요?

제한된 합리성  
최적의 결정 vs. 만족스러운 결정

# From Literacy to Data Literacy

## Data Literacy: 추상에서 구체로의 이행

- 관습적 믿음/직관(Literacy)에 대한 회의에서 출발
- 데이터를 통해 세상을 보는 안목: 날것의 기록에서 패턴을 찾아 세상에 대한 더 좋은(실용적인) 설명을 찾는 일

현실

직접 본 것 · 한 것



Literacy



추상 · 개념 · 관념의 탄생  
"사냥해서 짐승을 잡았다.  
사냥의 꽃은 들소 잡기"

현실의 기록



들소: 0마리  
물고기: 35마리  
들소: 0마리  
물고기: 65마리  
들소: 0마리  
물고기: 71마리  
물고기: 15마리  
들소: 1마리  
...

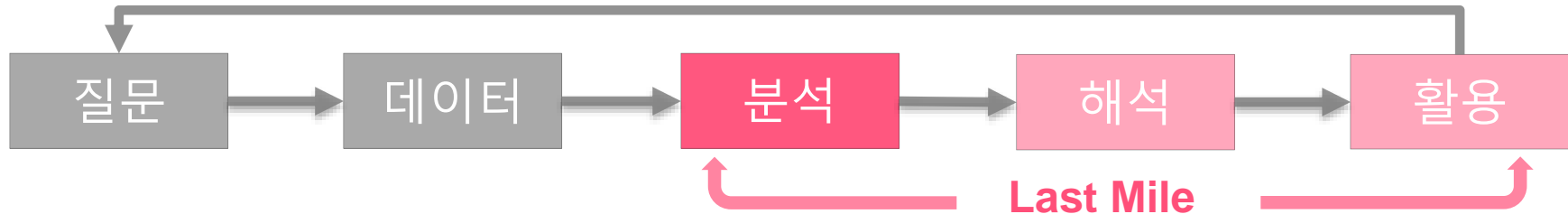
Data Literacy

들소보다 물고기를 잡는 게 1.7배 더 생산적

	사냥횟수	마릿수	kg	kg/사냥
들소	25	2	300	12kg
물고기	35	900	700	20kg

## Last Mile Problem

기업이 데이터에서 쓸모있는 패턴을 발견하여  
더 좋은 의사결정에 활용하지 못하는 문제



### 원인

- **분석 부재:** 엑셀보고; 대쉬보드
- **분석 분리:** 현업과 분석가의 분리;  
[질문→분석→활용] 선순환 X

### 해결책

- 현업 스스로 데이터에 질문, 패턴 발견·해석·활용
- **Data Literacy + Right Tool**

## 현업이 똑똑한 데이터 소비자가 되려면

### Data Literacy

#### 데이터 안목

도메인 지식 활용 데이터에 질문,  
분석결과를 실용적으로 활용

### Right Tool

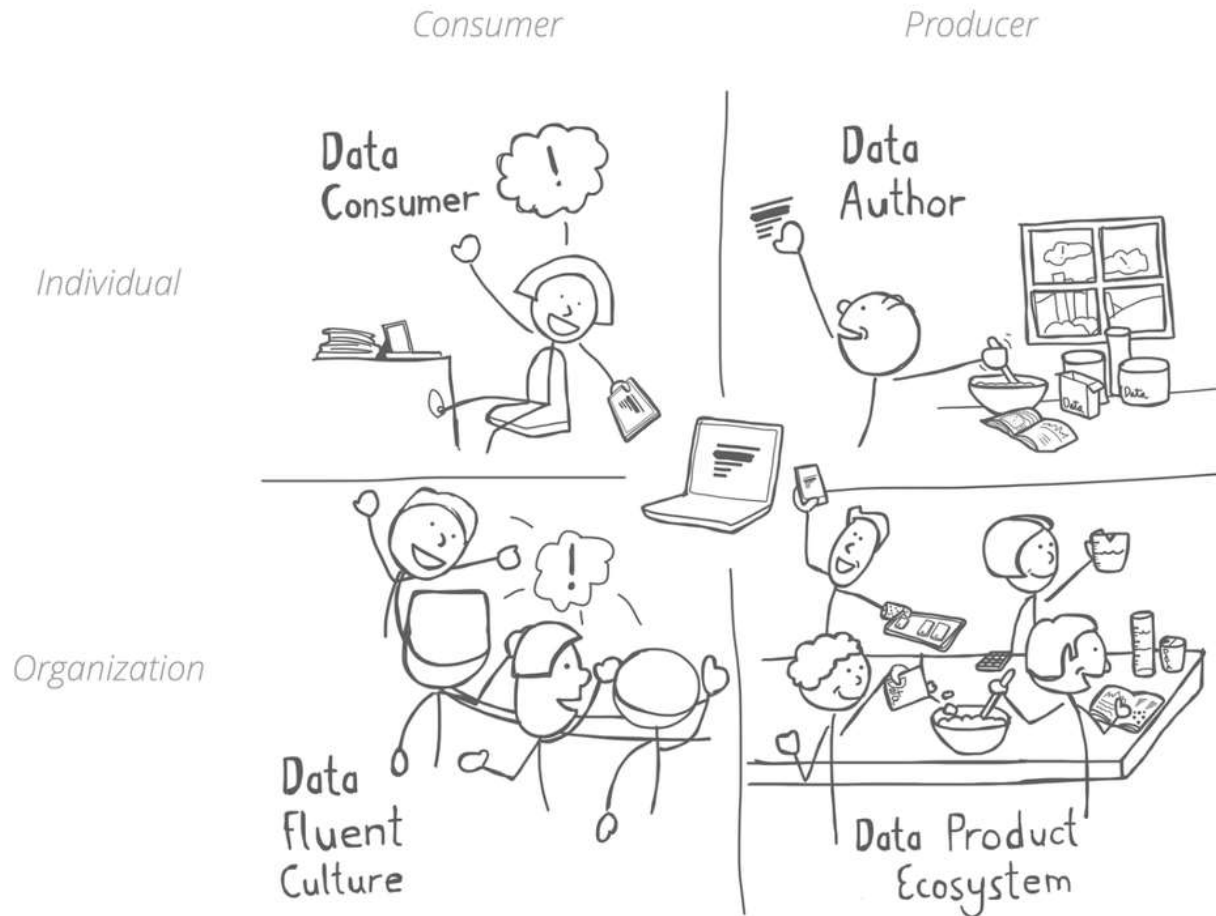
#### 닭잡는 칼

데이터의 특성 분석  
역량·목적에 맞는 도구

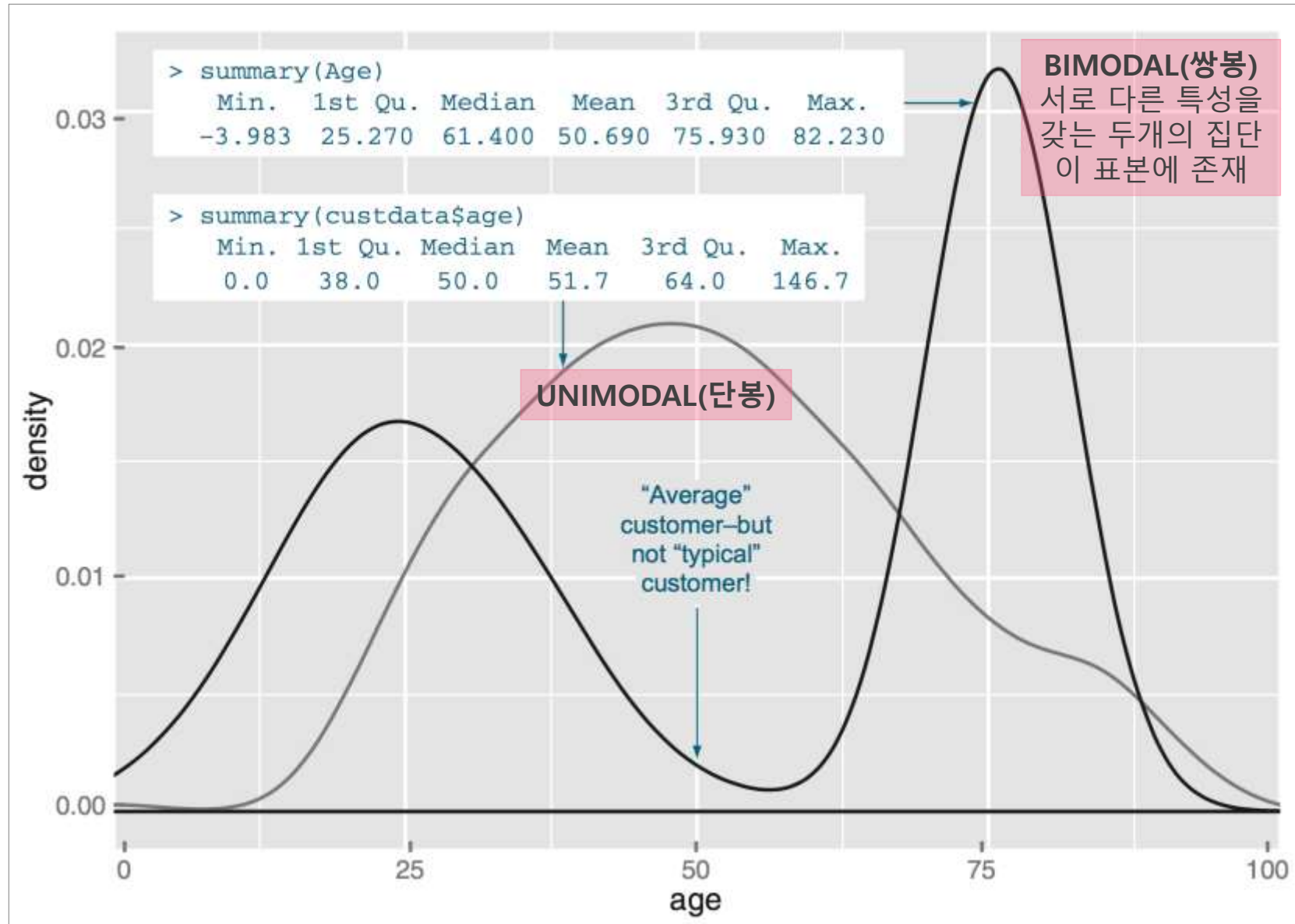


# Data Literacy(Fluency) Framework

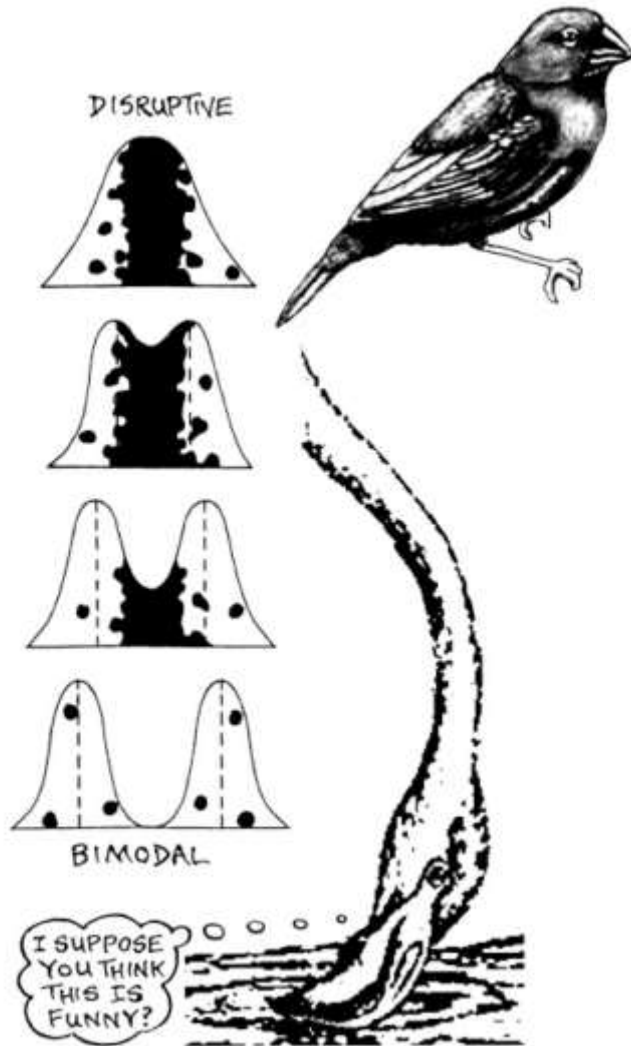
**데이터 분석은 생산자와 소비자 간의 사회적 · 상호적 활동**  
분석 결과를 소비하는 사람이 있어야 분석하는 사람이 존재할 수 있고  
좋은 분석을 생산하는 사람이 있어야 또 결과를 소비(활용)하는 사람이 존재



# Our First Data Literacy: One Hump vs. Two Humps



# Our First Data Literacy: BIMODAL DISTRIBUTION (쌍봉분포)



## African Seedcracker

자연선택에 의해 작고 부드러운 씨앗을 먹는 작은 부리의 새와 크고 단단한 씨앗을 먹는 큰 부리의 새가 나뉘어



lower bill 12 mm wide

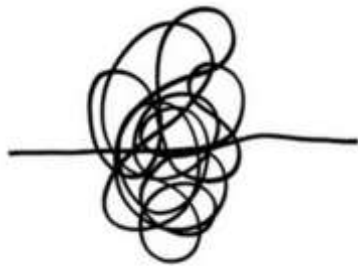


lower bill 15 mm wide

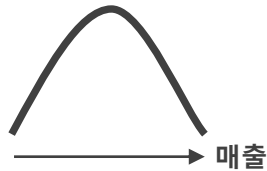


## 데이터를 읽을 수 있어야 새로운 해석도 가능

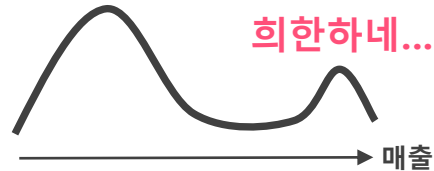
- **Reality:** 복잡계; 실제 작동방식 100% 알 수 없음
- **Belief:** 세상의 작동방식에 대한 최선의(만족스러운) 설명
- **Data:** 세상의 작동방식에 대한 기록; 세상의 샘플링
- **Insight:** 세상에 대한 더 나은 설명, 새로운 해석



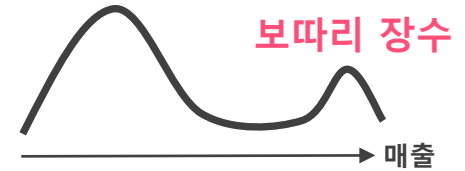
Reality



Belief



Data



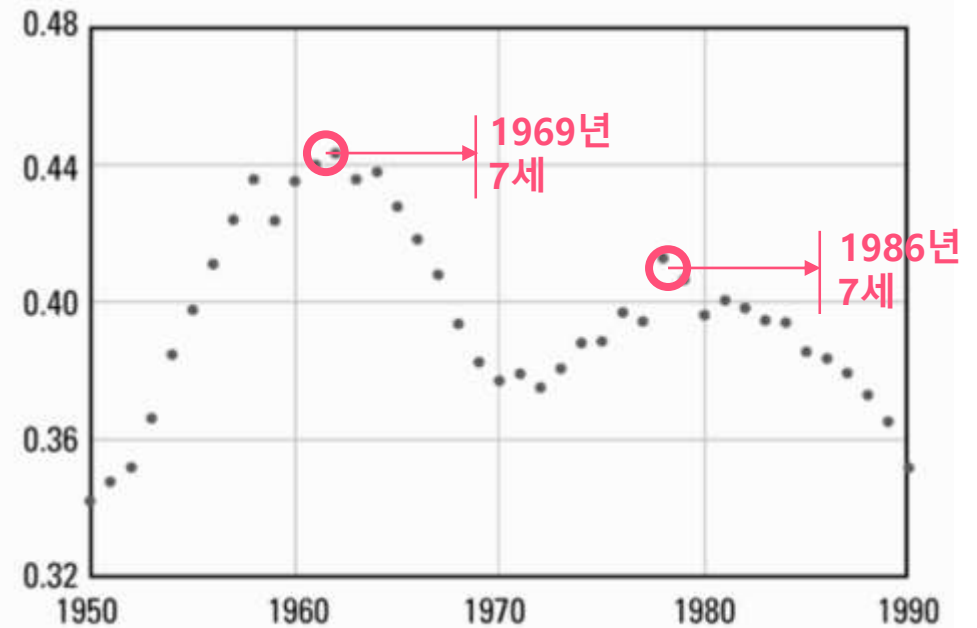
Insight  
[or Inspiration?]

1962년과 1979년에 태어난 남자들은 왜 Mets구단 팬이 되었나?

Hint: 1969년과 1986년에 모두 8세가 되었음



출생년도에 따른 NY Mets 팬들 비율  
[대상: 뉴욕 거주하는 남자 야구 팬들]



# Module II-a

## Data Understanding

아이디케이스퀘어드 양승준 / sidney.yang@idk2.co.kr  
<https://www.heartcount.io>

## 데이터 분석 방법 (X와 Y)

### 분석하는 이유

- 궁금한 것(Y)을 데이터(X)로 더 잘 설명(예측)
- X를 바꾸어서 Y를 개선하기 위해서

### 엑셀 (대쉬보드)

- 성과지표(Y)를 익숙한 관점(범주; X)으로 요약
- 과거에 대한 집계

### 데이터 시각화

- X와 Y를 점, 선, 크기, 색상으로 표현 (탐험분석)
- X와 Y 사이의 패턴(관계) 시각적 발견; 가설 수립

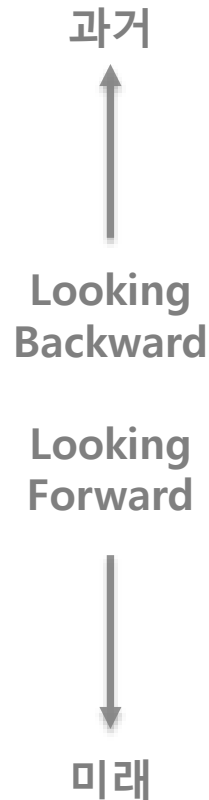
### 통계

- 데이터의 특성과 모양 요약 (기술 분석)
- 독립변수(통제가능; X)와 종속변수(Y) 간 가설 검증

### 기계학습

- 데이터 학습, Feature(X)로 Target(Y)을 예측·설명
- 의사결정 자동화 vs. 더 좋은 의사결정

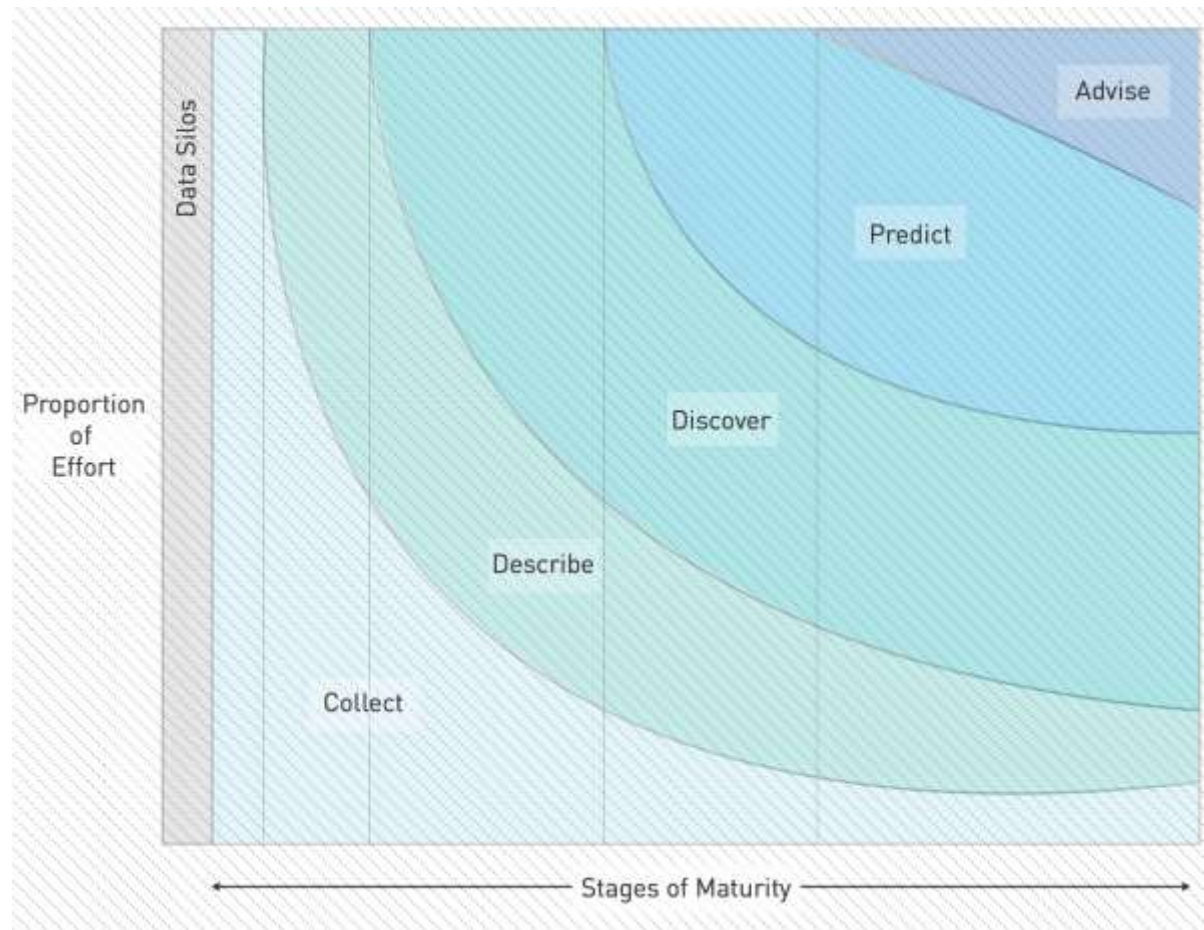
## 데이터 분석 주요기술



- **DESCRIBE (기술 분석) - 엑셀**
  - 데이터 특성과 모양을 (수치적으로) 요약
- **EXPLORE (탐험적 분석) - 데이터 시각화 도구**
  - 가설수립·데이터 감 잡기 위해 패턴 탐험
- **PREDICT/INFER (예측·추론 분석) - 통계/ML**
  - 패턴(모형)을 통해 주어진 문제를 예측·설명

# Data Analysis Maturity Model

데이터 수집 → 데이터 기술(묘사) → 패턴 발견 → 예측 → 활용  
우측으로 갈수록 성숙해진다기보다는 자기에게 필요한 단계를 잘 하면 됨



Source: Booz Allen Hamilton

**EDA(Exploratory Data Analysis) =**  
DESCRIBE (기술 분석) + EXPLORE (탐험 분석)

---

**EDA, 데이터와 함께 떠나는 창의적 여행 (생고생)**



- inspect data structure
- data quality
- summarize
- visualize data
- hypothesis generation
- != modeling

Source: Booz Allen Hamilton

## 데이터에 대해 사실적으로 묘사하는 법

### Description 요약

변수의 대표값과  
모양이 어떻게?

개별 변수(Y)의  
통계값과 분포 확인

### Comparison 비교

변수값의 차이가  
어디서 얼마나 나나?

서로 다른 범주(X) 간  
Y의 특성·모양 비교

### Relationship 관계

변수(Y)의 변화와 관계를  
갖는 다른 변수(X)는?

X와 Y 사이의  
상관관계 파악



# Analysis-Ready Dataset

## 분석하기 좋은 데이터셋

- 국가별로 1999/2000년에 결핵으로 사망한 환자수(Cases)와 전체인구(Population)를 정리한 데이터셋들
- 국가별 연도별 인구 10,000명당 결핵 사망률을 계산하기 가장 좋은 데이터는?

NOT SO GREAT

1

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

2

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

3

country	year	population
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 172006362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

GREAT

4

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

rectangular data  
data frame  
data table  
tidy dataset

# Rectangular Dataset and Key Terms

## 분석하기 좋은 데이터셋

- **Dataset:** 값(Values)들의 집합으로 숫자 또는 범주로 구성
- **Values:** 변수(Variable)와 관측점(Observation)으로 구성
- **Variable:** 동일한 속성(나이, 매출)에 대한 측정값들로 행(Column)을 구성
- **Observation:** 동일한 대상(사람, 매장)에 대한 측정값들로 열(Row)를 구성

country	year	cases	population
Afghanistan	1999	1815	19987071
Afghanistan	2000	2566	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	211258	1272015272
China	2000	211706	128042583

variables

country	year	cases	population
Afghanistan	1999	1815	19987071
Afghanistan	2000	2566	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	211258	1272015272
China	2000	211706	128042583

observations

country	year	cases	population
Afghanistan	1999	1815	19987071
Afghanistan	2000	2566	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	211258	1272015272
China	2000	211706	128042583

values

X  
features  
independent variables  
input (variables)  
predictor  
attribute

Y  
target  
dependent variables  
output (variable)  
response

record  
sample  
Instance  
case

# Raw vs. Aggregated Dataset

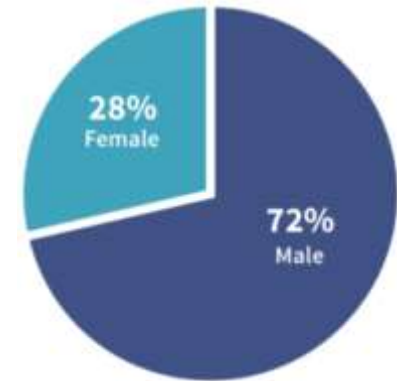
Raw Data

Name	Gender	Coffee
Bob Smith	M	Regular
Jane Doe	F	Regular
Dale Cooper	M	Mocha
Mary Brewer	F	Decaf
Betty Kona	F	Regular
John Java	M	Regular
Bill Bean	M	Regular
Jake Beatnik	M	Mocha
Bob Smith	M	Regular
Jane Doe	F	Regular
Dale Cooper	M	Mocha
Mary Brewer	F	Regular
John Java	M	Decaf
Bill Bean	M	Regular

Aggregated Data

Year	2000	2001	2002
Total sales	19,795	23,005	31,711
Male	12,534	16,452	19,362
Female	7,261	6,553	12,349
Regular	9,929	14,021	17,364
Decaf	6,744	6,833	10,201
Mocha	3,122	2,151	4,146

Q. 2001년 남녀 구매 비율은?



추가 질문

Q. 2001년 Regular Coffee 구매한 여자 고객수?

Q. 남자 고객이 선호하는 커피 종류는?

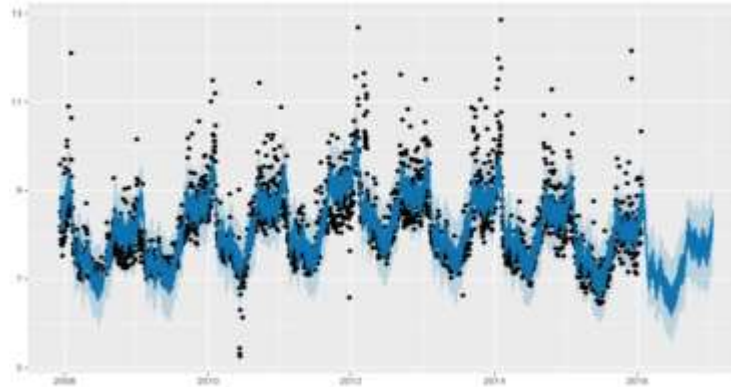
**Raw Data: Zoom-in(새로운 질문) 가능**

# Non-Rectangular Data Structures

Rectangular 구조가 아닌 데이터도 있음; 각 구조에 맞는 별도의 처리 및 분석 기법이 존재

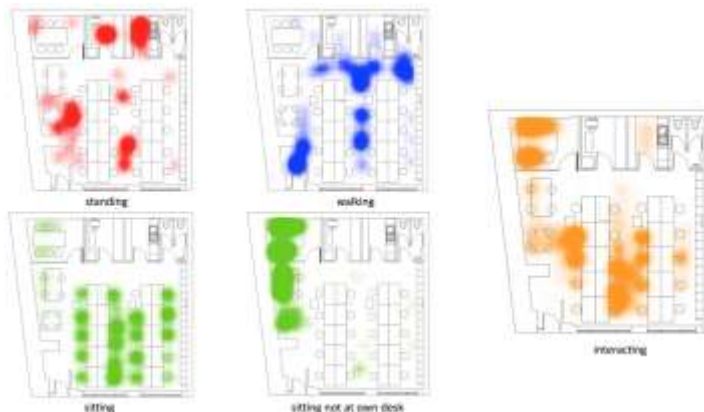
## TIME-SERIES

- 동일 변수 연속적 기록
- Seasonality; Event



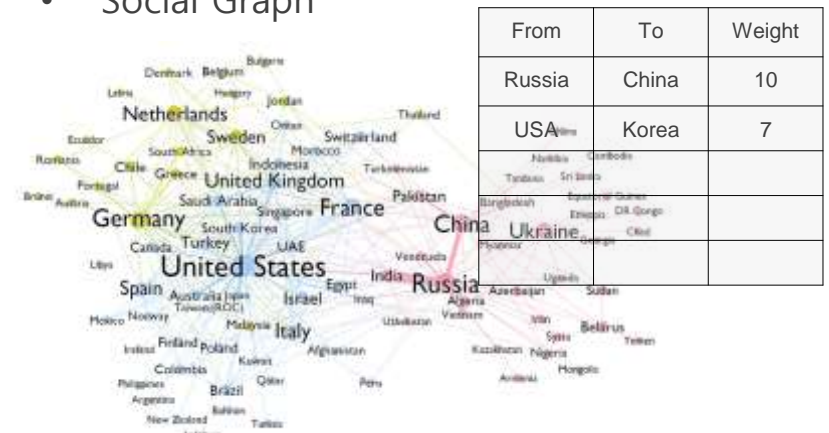
## SPATIAL

- Object에 대한 위치좌표
- Location Analytics; Geo-Statistics



## GRAPH

- Physical, Social, Abstract 관계
- Social Graph

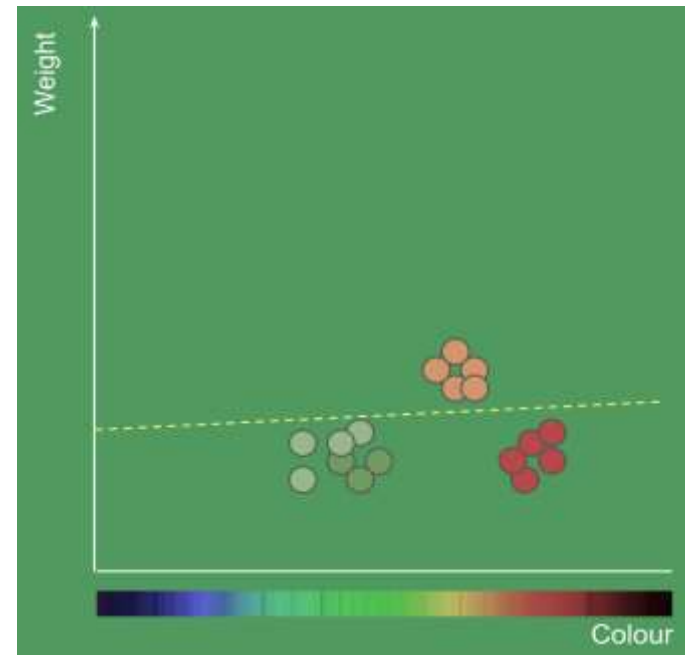
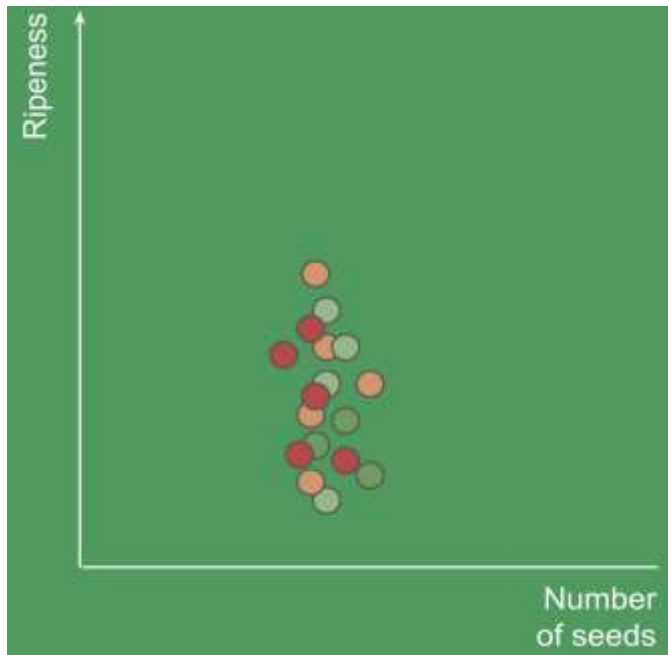


## Features / Attributes / X

**Feature:** Y(Output/Target)를 설명하거나 분류(예측)하는데 사용되는 속성  
좋은 Feature를 발굴하는 것이 참 중요함.



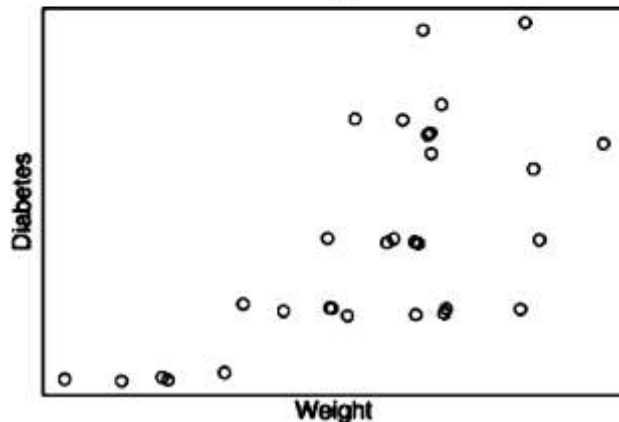
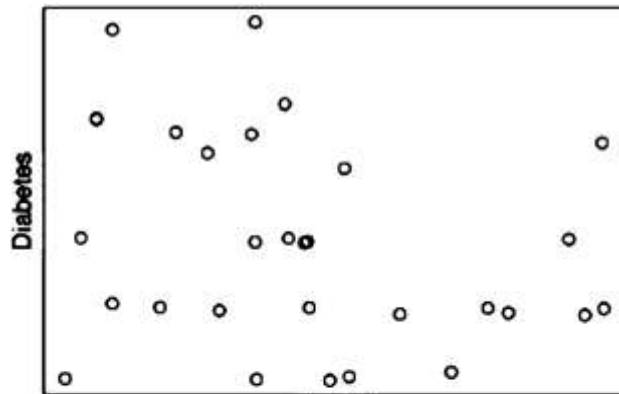
	Ripeness	# of Seeds	Weight (g)	Color	Fruit
	0.56	5	320	Orange	Orange
	0.61	6	280	Red	Apple



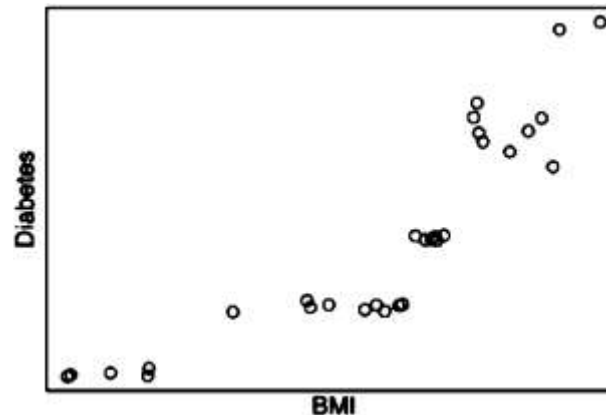
# Features Engineering

## Feature Engineering: From Raw Variable to Derived Variable

Y를 더 잘 설명하거나 분류(예측)할 수 있도록  
기존 변수를 창의적으로 가공하여 새로운 변수를 만드는 일



- 당뇨병 위험도와 상관관계가 높은 변수**
- 같은 몸무게라도 비만도는 키에 좌우됨
  - 비만도를 더 잘 반영할 수 있는(키와 몸무게의 상호작용을 잡아낼 수 있는) 새로운 변수 가공
  - \*BMI(Body Mass Index) =  $\text{kg/m}^2$



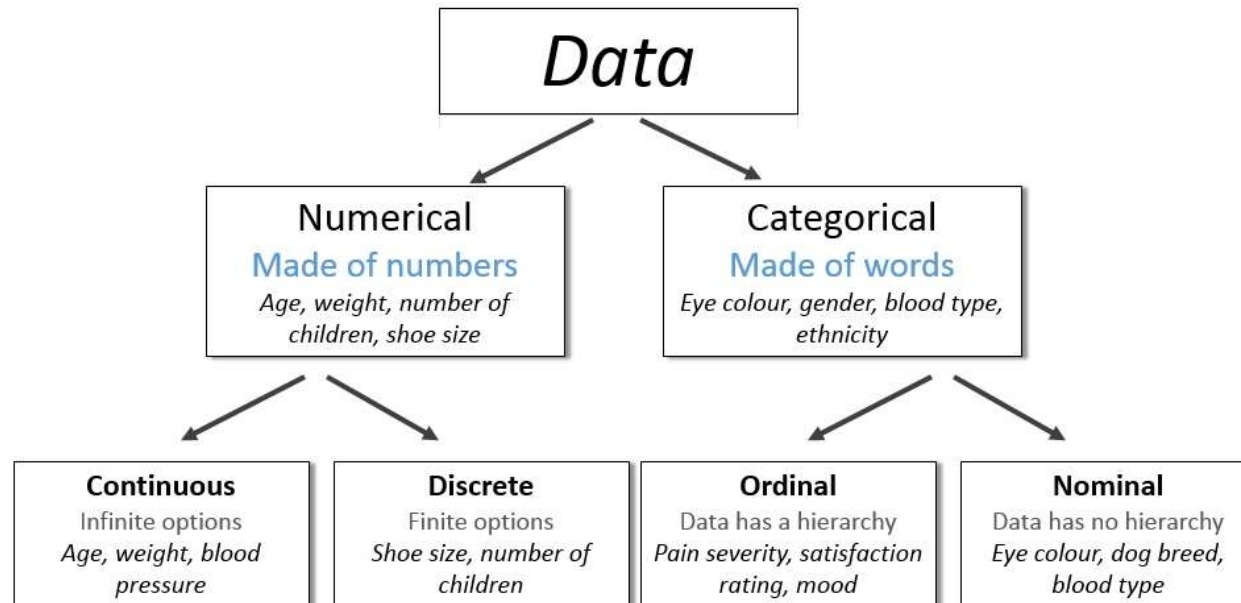
\*발명한 사람의 이름을 따서  
Quetelet Index라고도 함





## 숫자형(Quantitative)과 범주형(Qualitative)

분석: 숫자와 숫자 사이의 연관성, 숫자의 차이를 가져오는  
범주를 발견하는 것

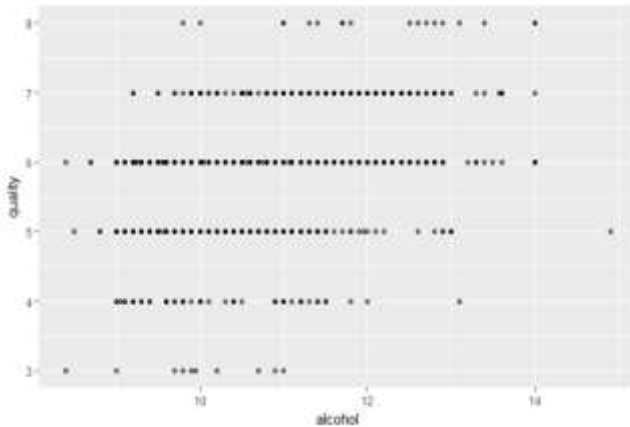


- 숫자형 자료는 이산형(discrete)이나 연속형(continuous)으로 나뉨
- 범주형 자료는 명목형(nominal)이나 순서형(ordinal)으로 나뉨

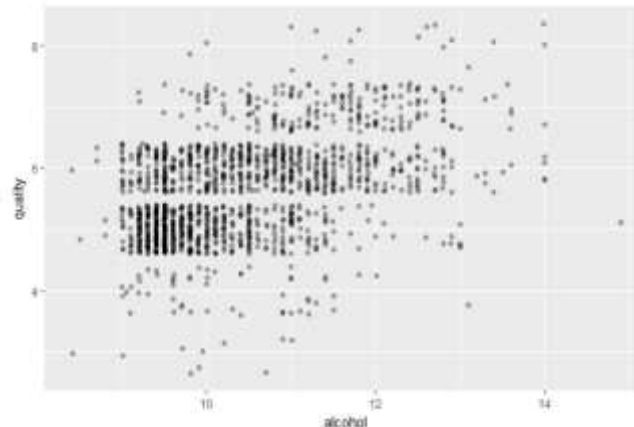
## 변수 유형에 따라 분석 방법과 효과적 시각화 방법이 달라짐

Alcohol(%): 와인 알코올 함량, Quality: 소비자가 매긴 점수

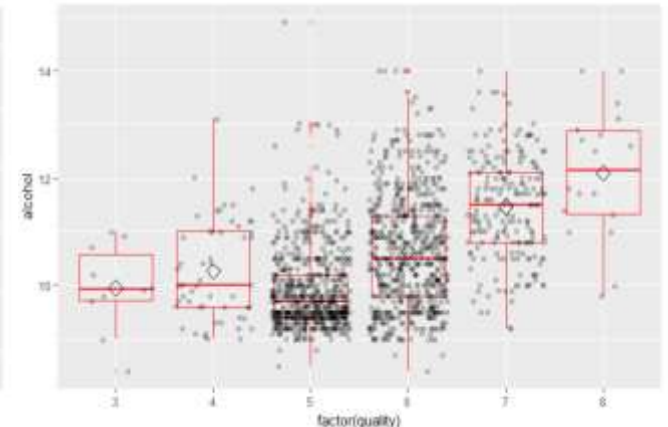
숫자 x 숫자 = Scatterplot  
Overplotting(점이 겹침)!



Jittering 기법으로 Noise 추가  
Jittering(인위적으로 퍼지게)!



Quality를 범주로 처리  
Boxplotting(분포 시각화)!



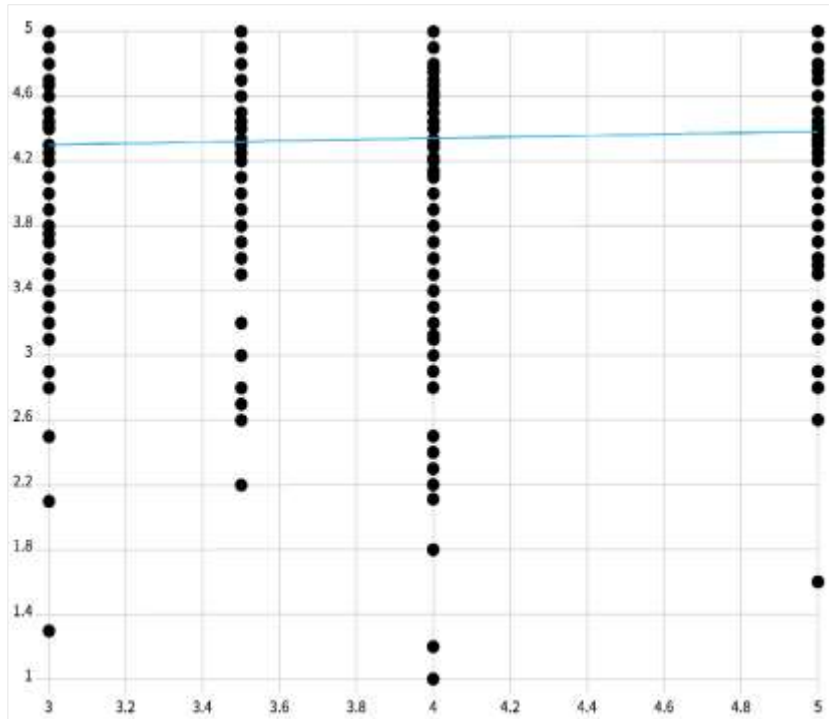


# Data Type에 따른 시각화 방법 – 그 때 알았더라면

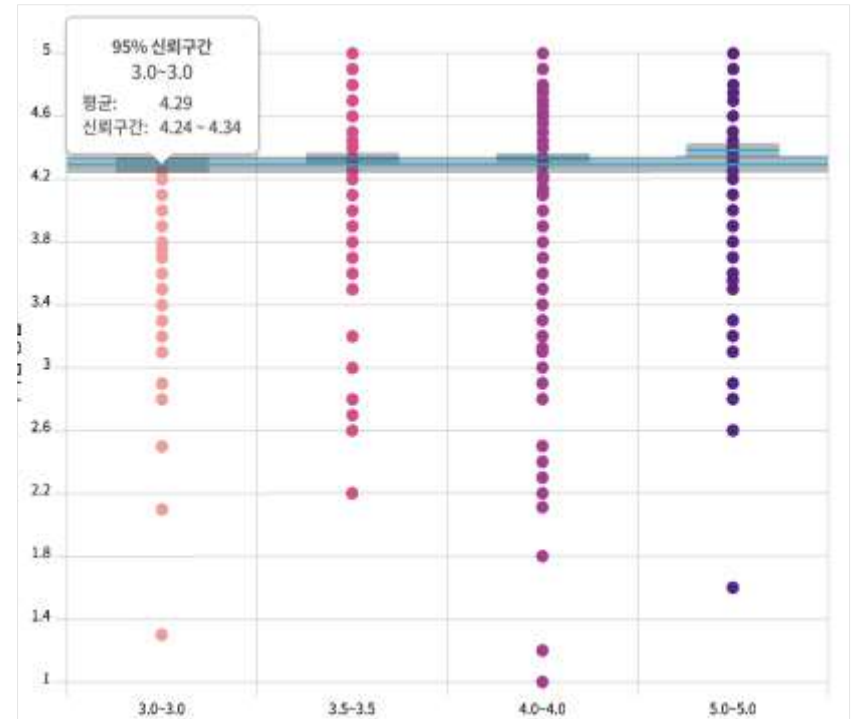
순서형(Ordinal) 변수는  
범주(Category)로 다루는 게 좋다.

Y: 리더십 점수, X: 평가 점수(등급)

X를 숫자로 처리  
X와 Y 간 상관관계(0.06)가 매우 약함



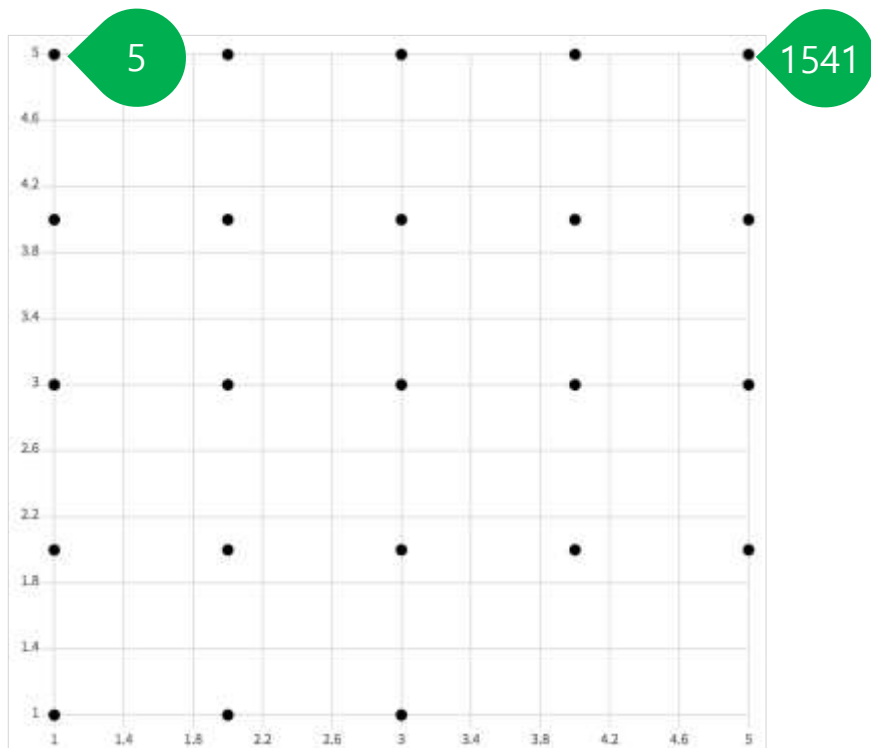
X를 범주로 처리  
서로 다른 범주(점수)간 Y값 차이가 존재함



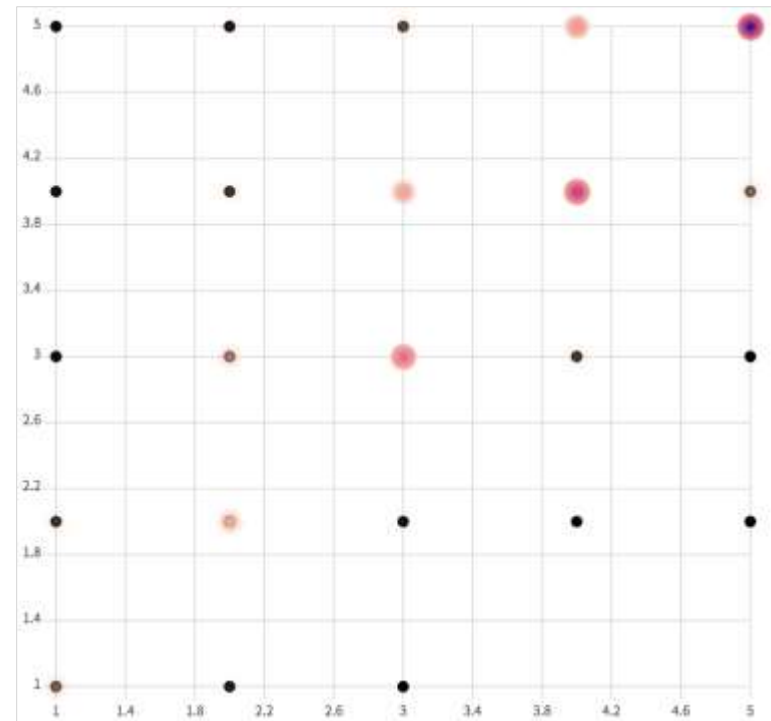
## 순서형x 순서형 변수 간 관계 시각화 서베이(설문) 데이터

개별 레코드의 Density(밀도)를 표현해서 Overplotting 문제를 해결

Overplotting이 심각함



밀도가 높을수록 진하고 크게 표현



# 숫자형 변수를 나누는 또 다른 기준: Interval vs. Ratio



절대적 원점(True Zero)이 있으면 Ratio, 없으면 Interval

시간 = 00:00 : 시간이 없다(빵시)?

나이 = 0살 : 나이가 없다(빵살)

Q. 나누거나(Ratio) 곱해도 말이 되는 것은?  
온도 vs. **몸무게**

Interval (구간 자료)  
 $10\text{도} + 10\text{도} = 20\text{도}$   
 $20\text{도} / 10\text{도} = 2\text{배?}$

Ratio (비율 자료)  
 $50\text{kg} + 50\text{kg} = 100\text{kg}$   
 $100\text{kg} / 50\text{kg} = 2\text{배?}$



# Data Description (Data Profiling)

## 데이터의 특성과 모양을 요약하여 기술하는 방법

### Central Tendency 중심 경향

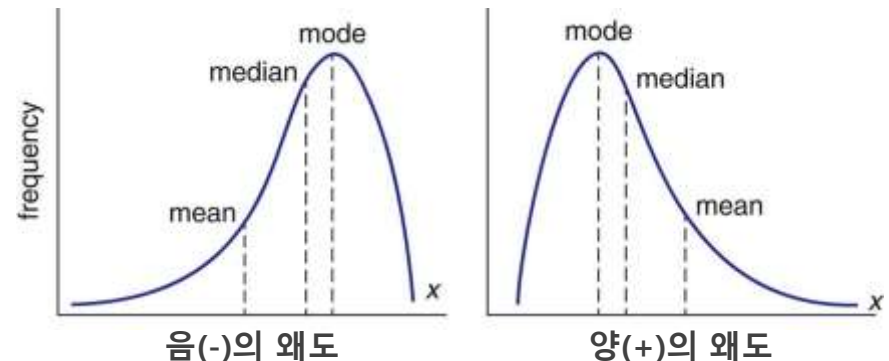
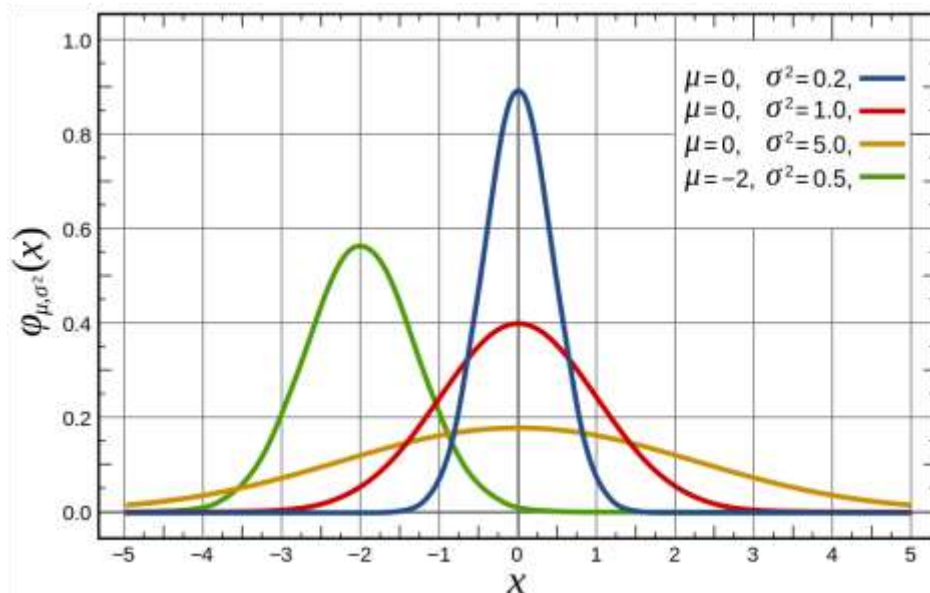
- 평균(Mean)
- 중앙값(Median)
- 최빈값(Mode)

### Dispersion 퍼진 정도

- 범위(Range)
- 분산(Variance)
- 표준편차(SD)
- Percentile

### Shape of Distribution 퍼진 모양(대칭)

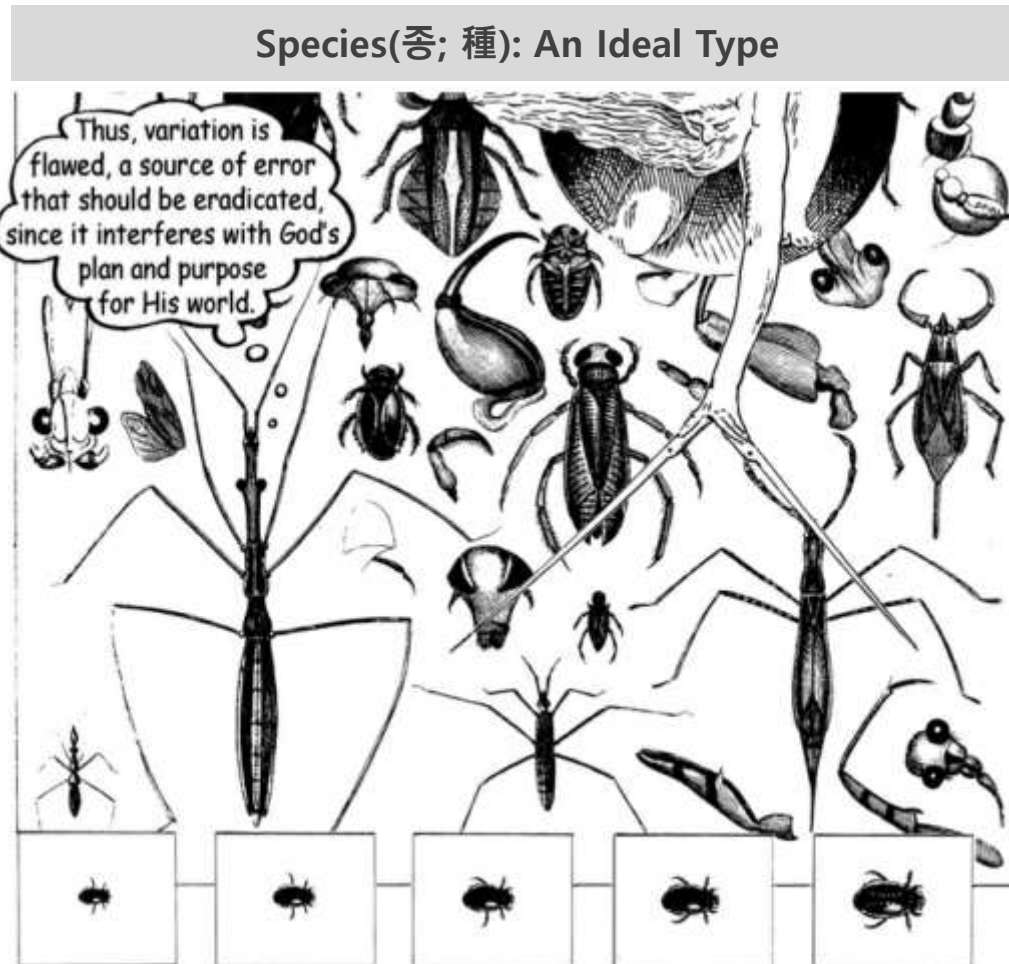
- 왜도 (Skewness)



# The Philosophy of Statistics [19<sup>th</sup> Century]

## 초기의 통계학 - 결정론적 세계관에 바탕을 둔 이데아/본질의 추구

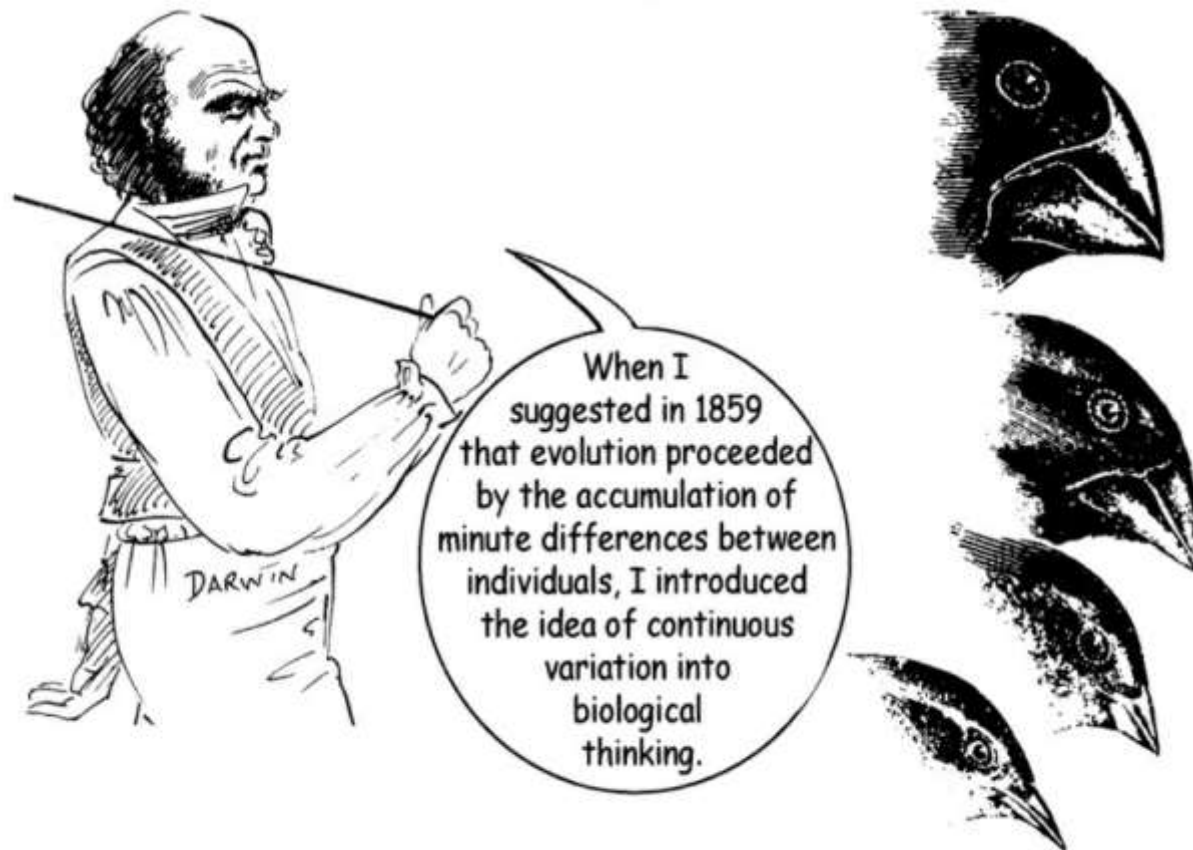
평균값이 대상이 보유한 이상적인 속성이고(Idealized Mean)  
Variation(차이)은 제거해야 할 오류라는 생각이 지배적이었음



# Darwin and Statistical Population [Late 19<sup>th</sup>~Early 20<sup>th</sup> Century]

**다윈의 등장: Type/Essence(본질) → Variation(차이)**

차이(변이)의 점진적 누적에 의해 진화가 이루어진다는 발견  
개별 개체에 존재하는 의미있는 차이(변이)에 관심을 갖기 시작



# Vital Statistics vs. Mathematical Statistics

Average: 집단을 요약 → Variation: 개인(개체)들에 존재하는 차이에 관심

인구통계  
평균, 비율



수리통계  
분산, 관계, 추론  
확률, 유의성

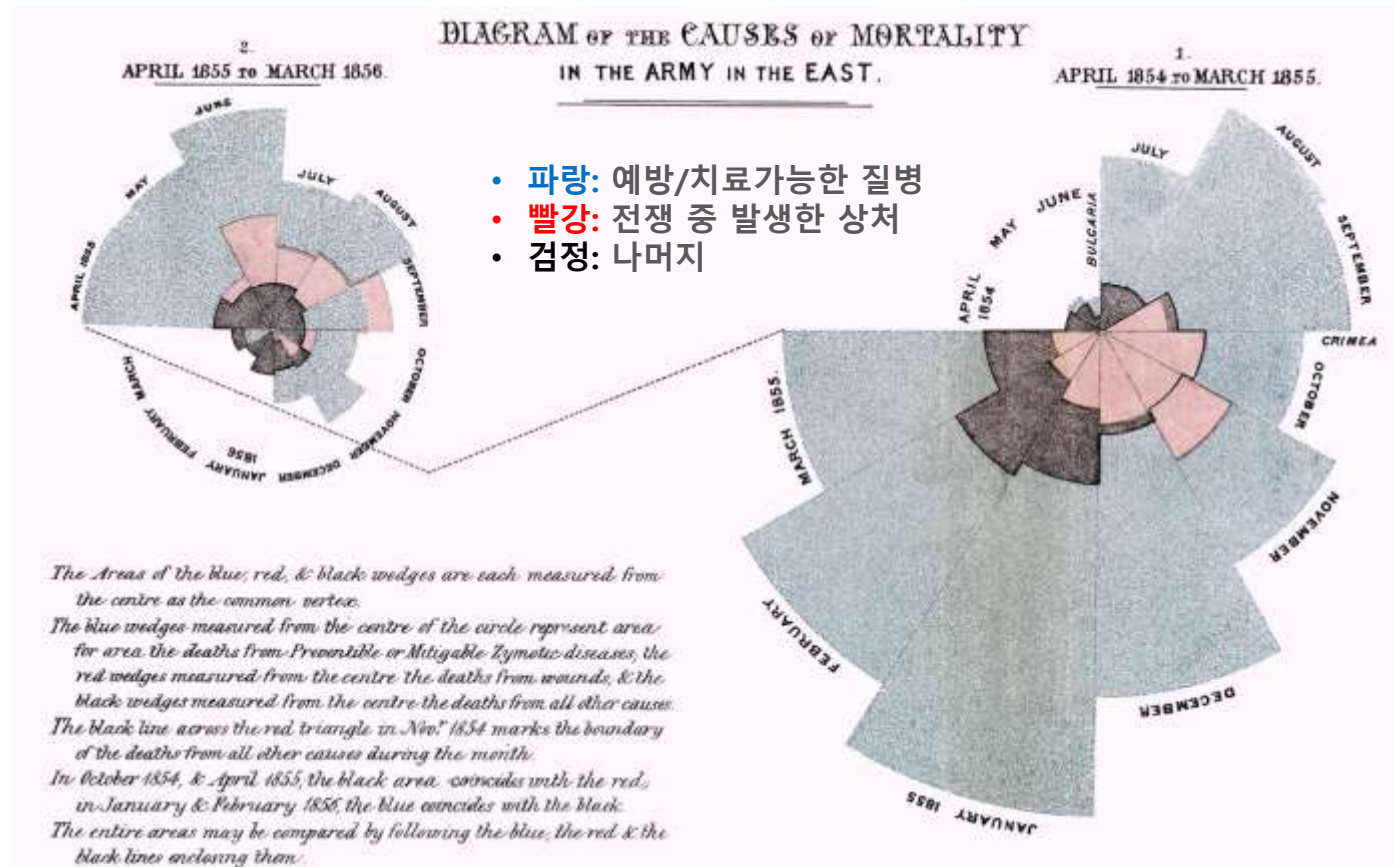


# Florence Nightingale, the Lady with the Lamp

“신의 생각을  
이해하기 위해 우리는  
통계를 공부해야 해요.  
통계를 통해 신이  
목적하신 바를 측정할  
수 있다구요.”

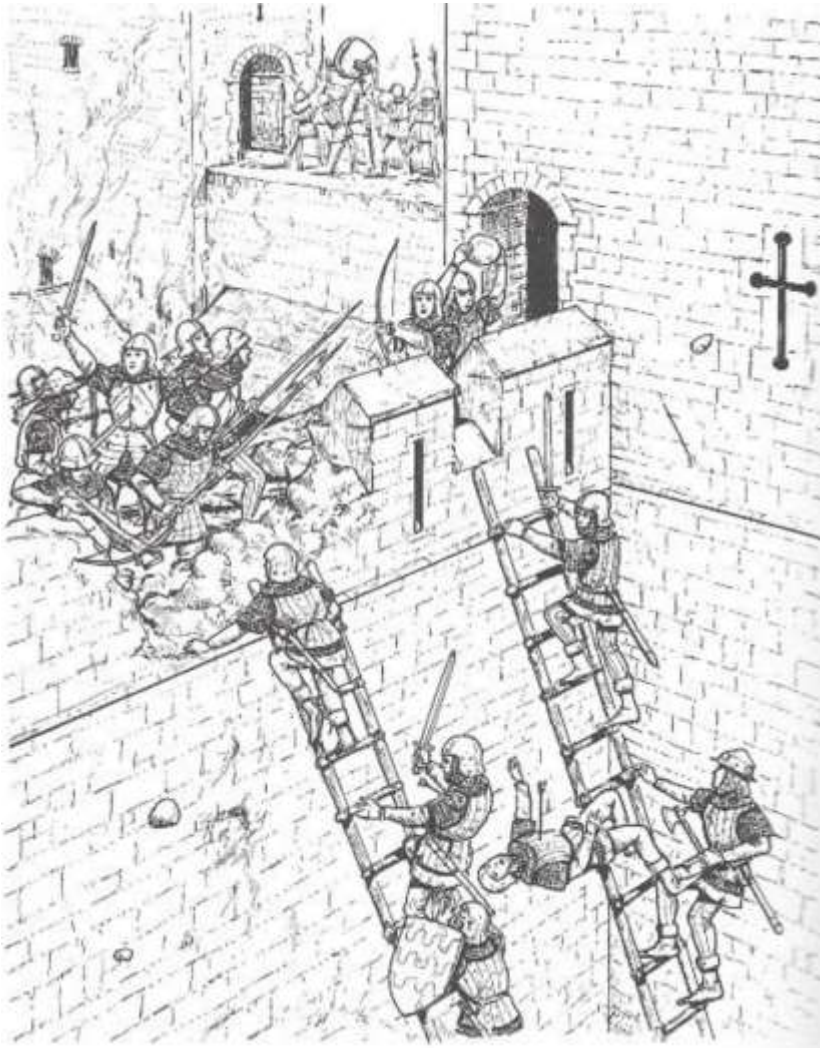


- 크림 전쟁에 여성 최초로 영국군 보건위생장교로 참여
- 표준화되지 않은 질병 분류체계; 주먹구구 사망 데이터 관리
- 전사자 데이터 정리 후, 보고 방식 고민 (테이블? 파이차트?)
- Data Storytelling: 파이차트를 변형하여 데이터를 시각화





# 3 Types of Average: Mean, Median, Mode



성벽의 벽돌 갯수를 병사들이 측정한 값들

병-1	병-2	병-3	병-4	병-5	병-6	병-7	병-8	병-9
13	18	13	15	13	16	14	21	13

Q. 어떤 값을 대표값으로 선택할까?

A. 평균

$$(13+18+13+14+13+16+14+21+13) \div 9 = 15$$

B. 중앙값

13, 13, 13, 13, **14**, 14, 16, 18, 21

C. 최빈값

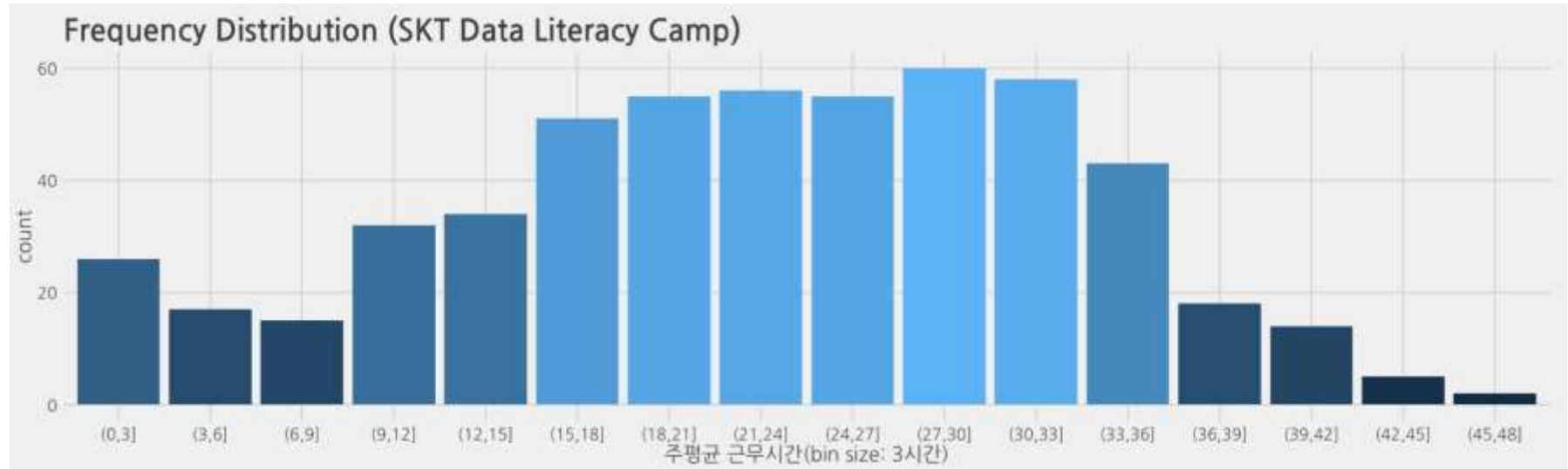
**13** (3번 측정; 다른 값들은 1번씩만 측정됨)

D. 선호값

**16** (병-6)

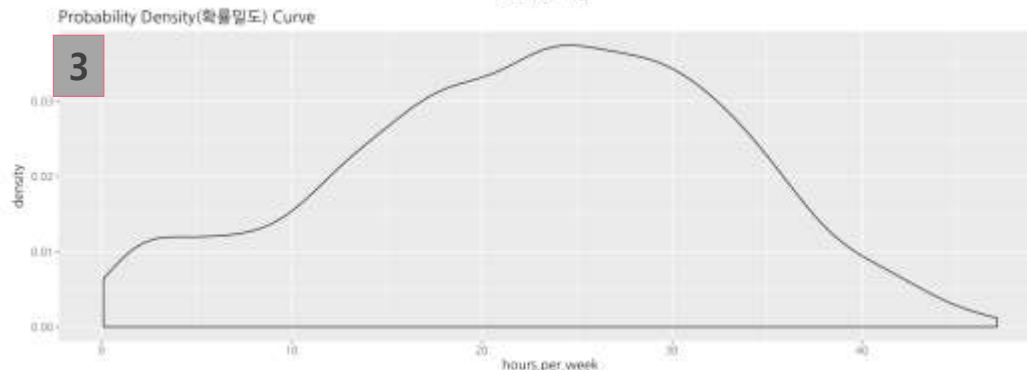
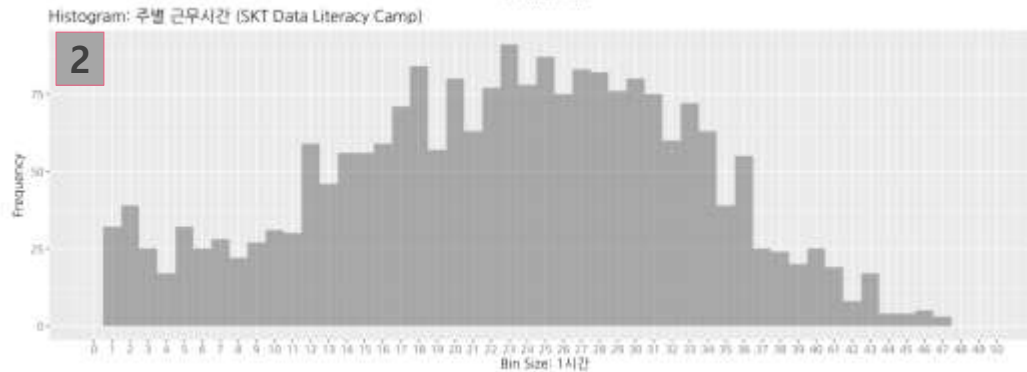
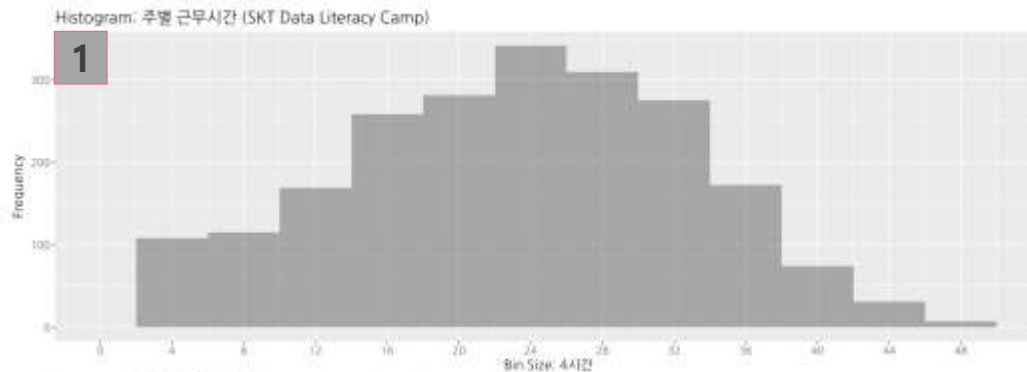
# Histogram vs. Frequency Distribution Table

## 히스토그램과 도수분포표



계급	(0,3]	(3,6]	(6,9]	(9,12]	(12,15]	(15,18]	(18,21]	(21,24]	(24,27]	(27,30]	(30,33]	(33,36]	(36,39]	(39,42]	(42,45]	(45,48]
빈도	26.0	17.0	15.0	32.0	34.0	51.0	55.0	56.0	55.0	60.0	58.0	43.0	18.0	14.0	5.0	2.0
누적 빈도	26.0	43.0	58.0	90.0	124.0	175.0	230.0	286.0	341.0	401.0	459.0	502.0	520.0	534.0	539.0	541.0
비율	4.8	3.1	2.8	5.9	6.3	9.4	10.2	10.4	10.2	11.1	10.7	7.9	3.3	2.6	0.9	0.4
누적 비율	4.8	7.9	10.7	16.6	22.9	32.3	42.5	52.9	63.1	74.2	84.9	92.8	96.1	98.7	99.6	100.0

# Histogram vs. Density Plot



## 1 Histogram – Bin Size: 4시간

- 히스토그램: 도수(빈도)의 분포[도수분포표]를 차트로 표현한 것
- 계급: X축에 표현된 변수의 구간[4시간]

## 2 Histogram – Bin Size: 1시간

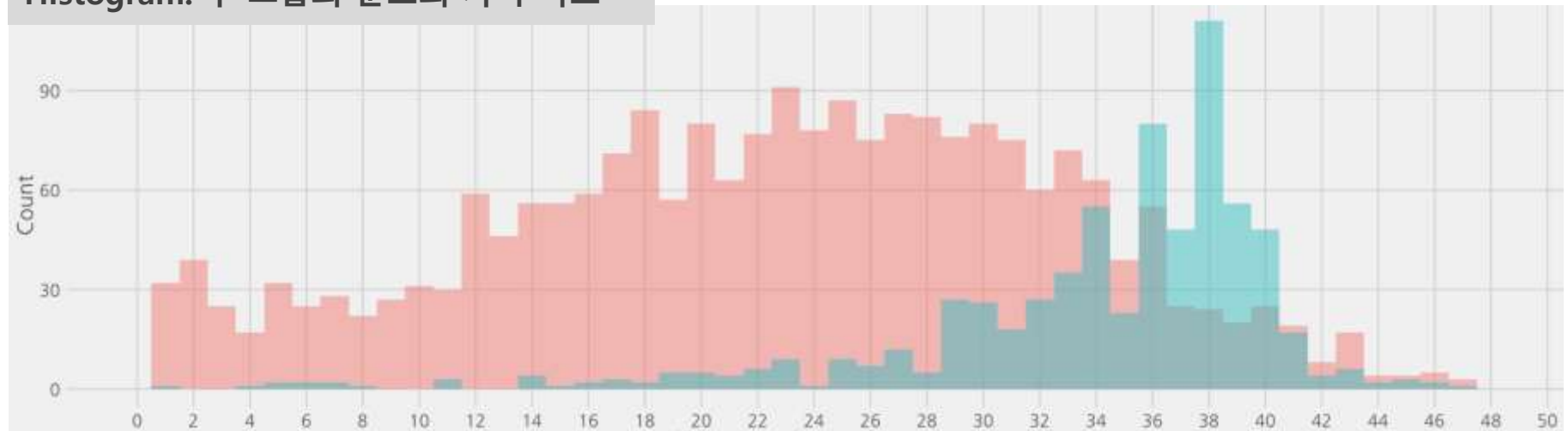
- X축 변수 구간의 크기(Bin Size)를 4시간에서 1시간으로 조정하였음

## 3 Probability Density Curve

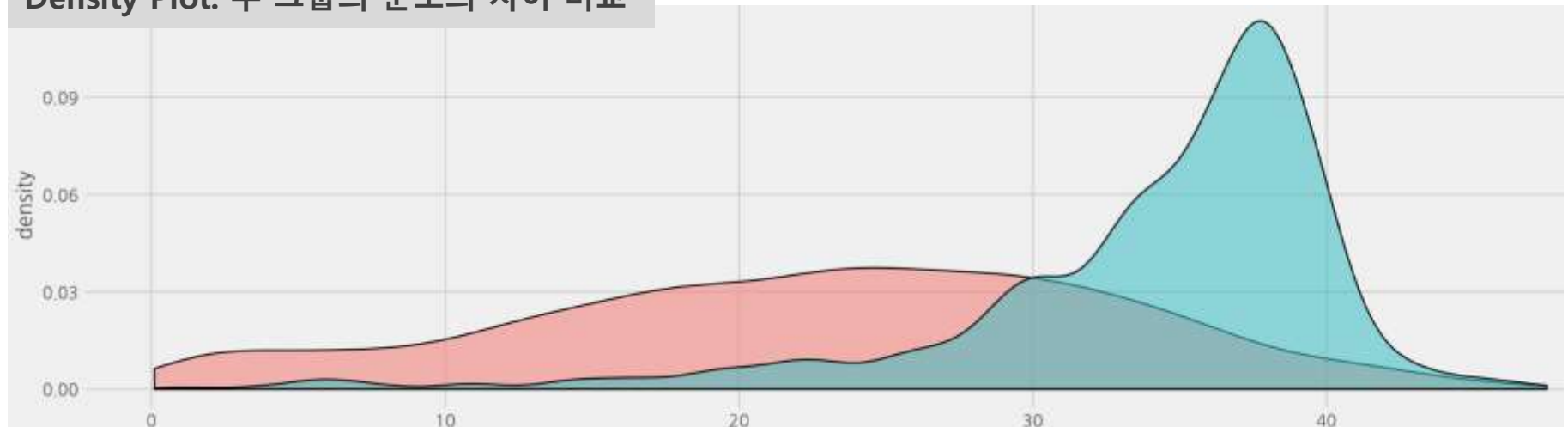
- 확률밀도: X가 연속형 변수일 경우 X값과 이에 대응하는 확률을 나타낸 그래프
- 좌측에서 X가 10~20시간 사이의 값을 가질 확률은 해당 구간의 면적과 동일함

# Histogram vs. Density Plot: 서로 다른 두 집단의 분포를 비교

Histogram: 두 그룹의 분포의 차이 비교

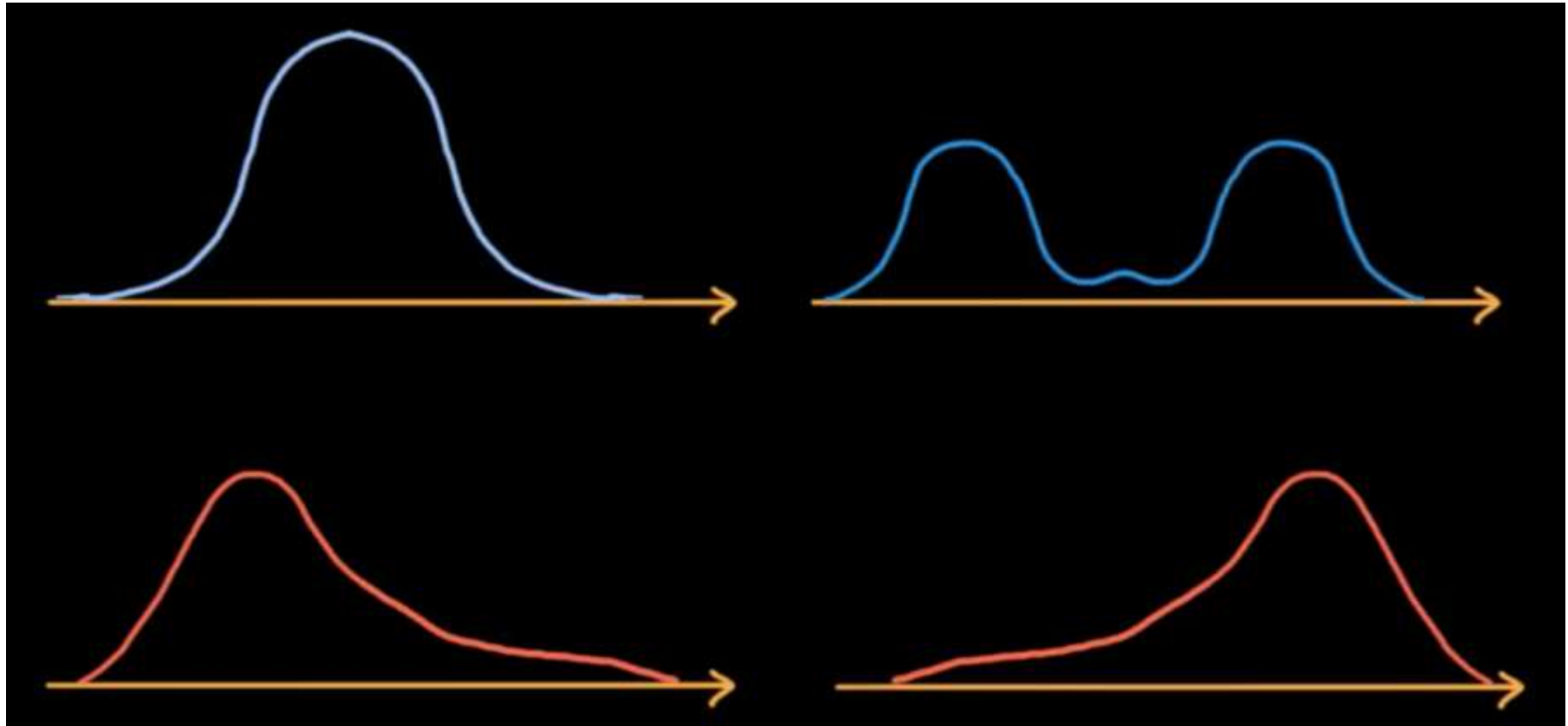


Density Plot: 두 그룹의 분포의 차이 비교





Let's Find Out Mean and Median on Density Curve  
밀도함수에서 평균과 중앙값 찾아봅시다.



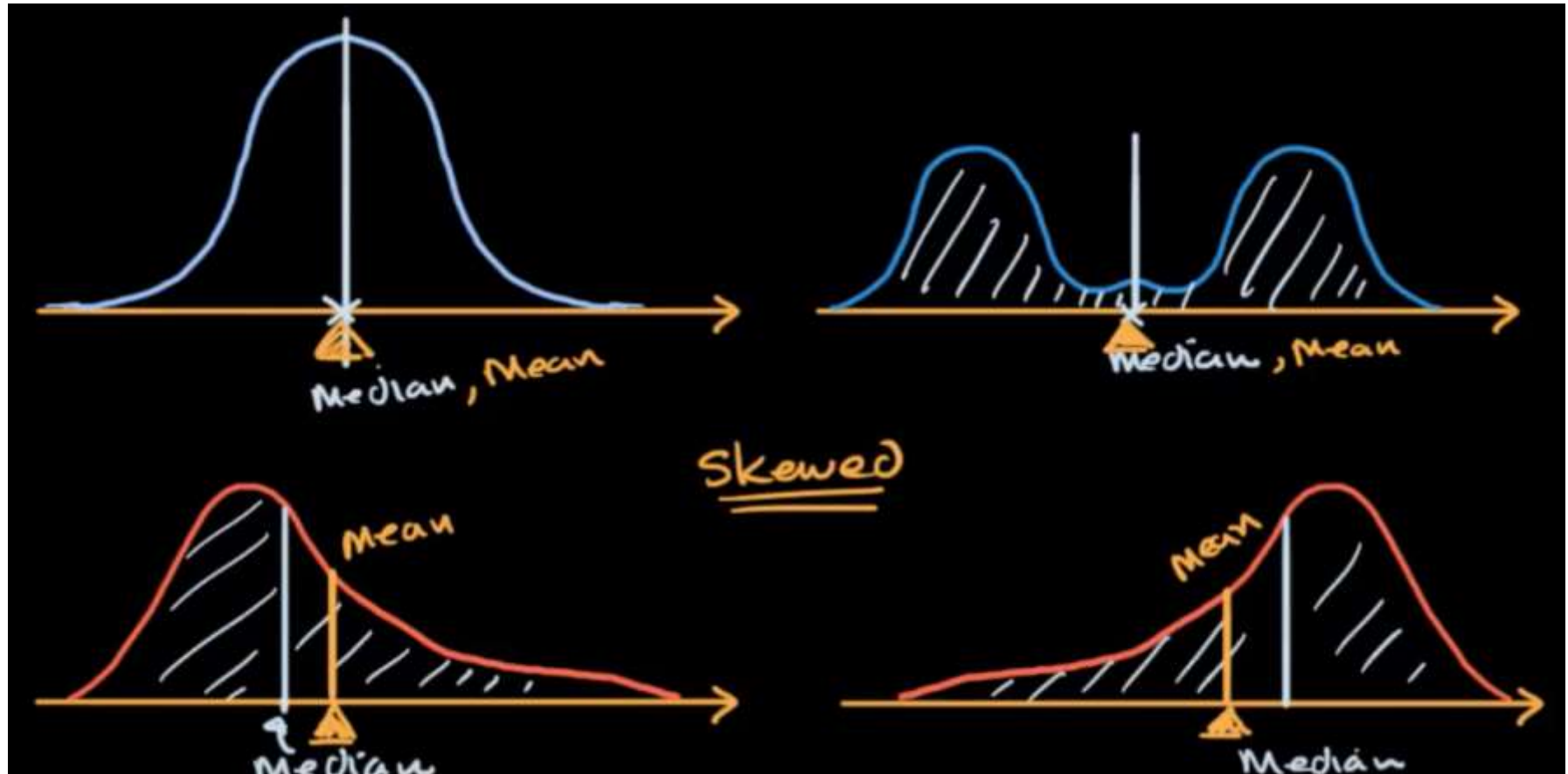
Source: <https://www.khanacademy.org/math/statistics-probability>



Let's Find Out Mean and Median on Density Curve

중앙값: 면적을 분할

평균: 무게 중심



Source: <https://www.khanacademy.org/math/statistics-probability>

# Larger Variation, Greater Sampling Error



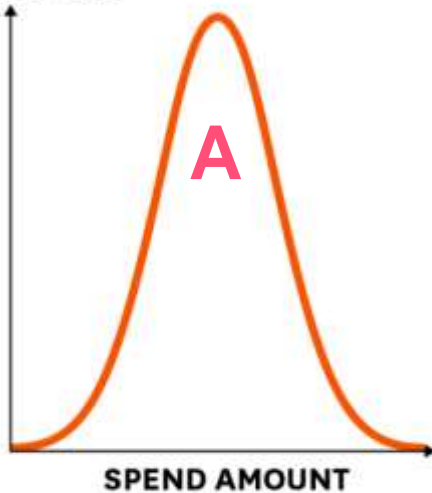
- 신규 캠페인에 노출된 고객: 1인당 **10,000원** 소비
- 기존 캠페인에 노출된 고객: 1인당 **8,000원** 소비

Q. 기존 고객들의 소비액 분포가 A와 B 두가지가 있다고 가정했을 때, 둘 중 신규 캠페인이 더 효과적이라고 주장(일반화)하기에 좋은 분포는?

변화의 폭 ↗    샘플 데이터 신뢰도 ↘    평균값에 대한 확신 ↘

**Lesser Variation**

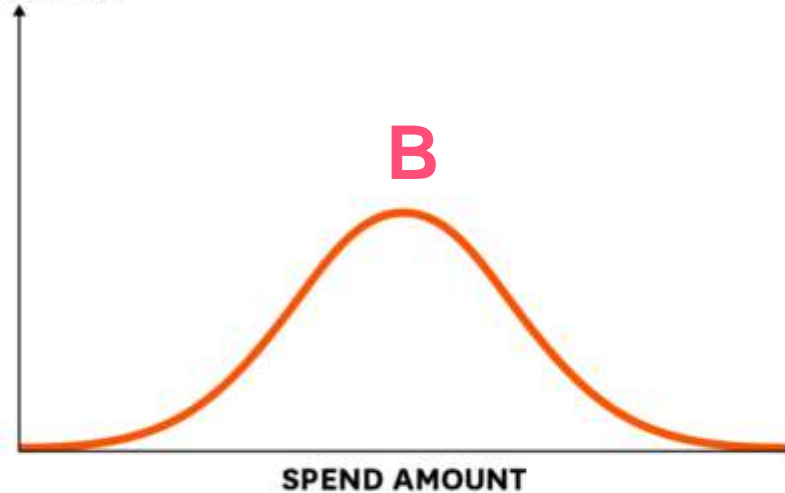
NUMBER OF  
CUSTOMERS



SOURCE THOMAS C. REDMAN

**Greater Variation**

NUMBER OF  
CUSTOMERS



© HBR.ORG

# Percentile

다른 관측(측정)값들을 크기의 분포를 고려했을 때  
특정값의 상대적 위치

Original Value	Ordered Value	Ranking	Percentile	Quartile (사분위)
32	29	1	8%	1 <sup>st</sup> Qu. [~25 <sup>th</sup> ] (최하위 25%)
54	32	2	17%	
74	<b>38</b>	<b>3</b>	<b>25%</b>	
99	41	4	33%	2 <sup>nd</sup> Qu. [~50 <sup>th</sup> ] (차하위 25%)
38	53	5	42%	
55	<b>54</b>	<b>6</b>	<b>50%</b>	
29	55	7	58%	3 <sup>rd</sup> Qu. [~75 <sup>th</sup> ] (차상위 25%)
41	74	8	67%	
134	<b>93</b>	<b>9</b>	<b>75%</b>	
53	99	10	83%	4 <sup>th</sup> Qu. [~100 <sup>th</sup> ] (최상위 25%)
209	134	11	92%	
93	<b>209</b>	<b>12</b>	<b>100%</b>	



## Binning (Feature Engineering)

### Binning: 연속된 숫자형 변수를 범주형 변수로 변환하는 것

- **\_bin**: 변수(나이)값의 범위(20~59)를 기준으로 최대한 균등하게 10개 구간을 생성
- **\_percentile**: 레코드 갯수가 최대한 균등하게 안분되도록 5개 구간을 생성

Ranking	나이	나이_bin	나이_percentile
1	20	20~23	~20 <sup>th</sup> (하위 20%) 3개의 레코드
2	24	24~27	
3	25	24~27	
4	29	28~31	~40 <sup>th</sup> 3개의 레코드
5	33	31~34	
6	33	31~34	
7	39	38~41	~60 <sup>th</sup> 3개의 레코드
8	40	38~41	
9	41	38~41	
10	42	42~45	~80 <sup>th</sup> 3개의 레코드
11	43	42~45	
12	43	42~45	
13	44	42~45	~100 <sup>th</sup> (상위 20%) 3개의 레코드
14	51	50~53	
15	60	58~60	

## Percentile 활용하여 주성분(Principal Component) 찾기

## 비타민, 지방, 섬유질 변수로 채소와 고기 분류하기



- **비타민 빼기 지방:** Percentile 값으로 바꾸면 해당 변수를 정규화하는 효과가 있어서 서로 다른 단위를 갖는 변수들 간 연산이 가능해짐
- **PCA:** 데이터 분류를 용이하게 하는 (=데이터가 최대한 퍼지게 하는) 주성분(Principal Component; 이 경우 Vitamin C + Fiber - Fat)을 찾는 일.



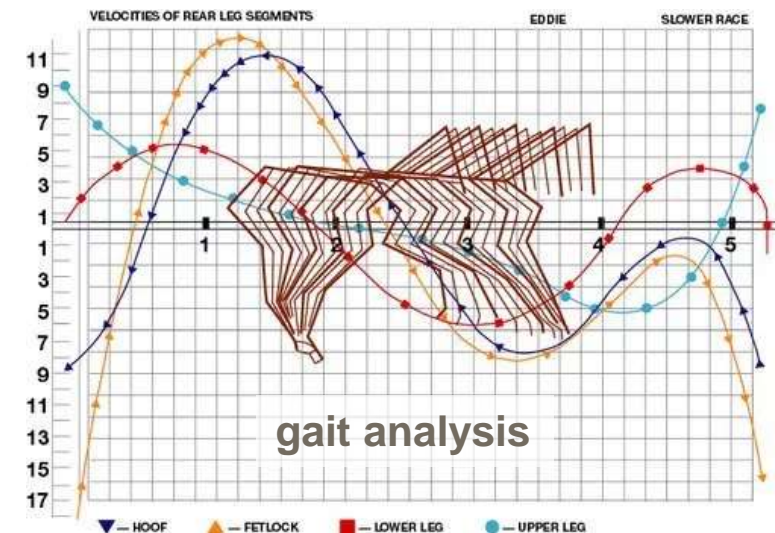
# Finding Secret Feature

## Percentile과 남들은 모르는 좋은 Feature로 돈 번 사례

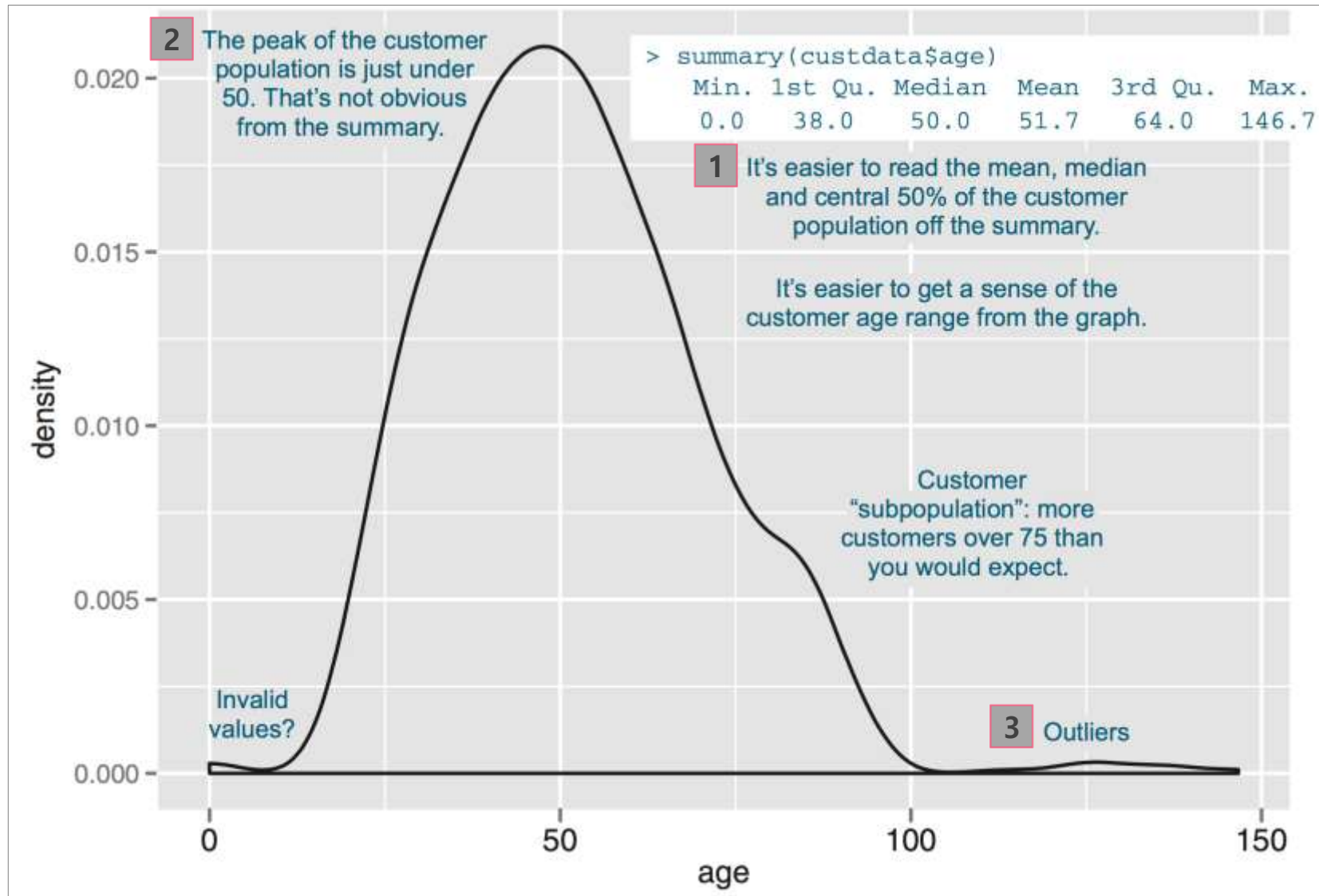
**American Pharoah:** 2015년도에 37년 만에 Triple Crown 달성 (삼관마)  
Jeff Seder: **“Sell your house. But, do NOT sell this horse.”**



Variable	Percentile
Height	56%
Weight	61%
Pedigree	70%
<b>Left Ventricle (좌심실)</b>	<b>99.61%</b>

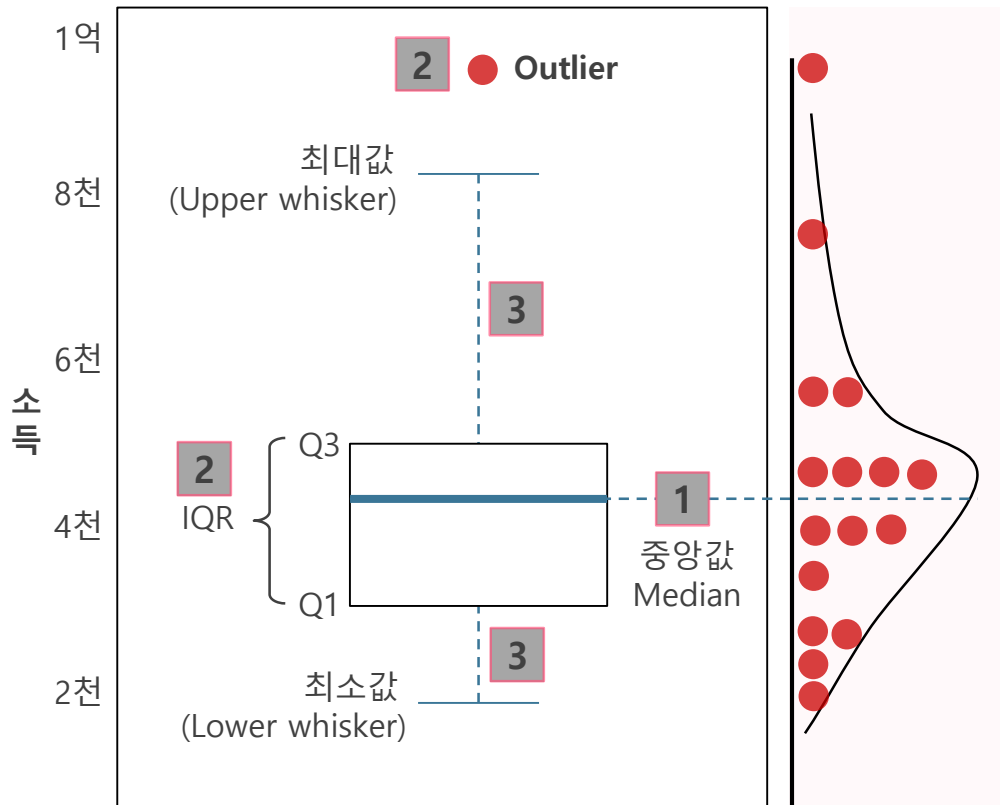


# Summary Statistics and Visualization



# Distribution (boxplot)

## Box-and-whiskers plot (예시)



## Box-and-whiskers plot 해석

### 1 중심 경향

- 중양값(median) 파악

### 2 특이값(Outlier)

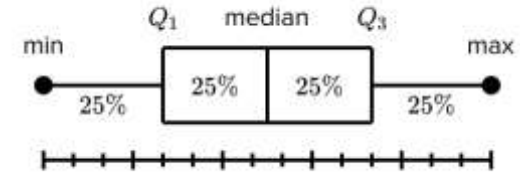
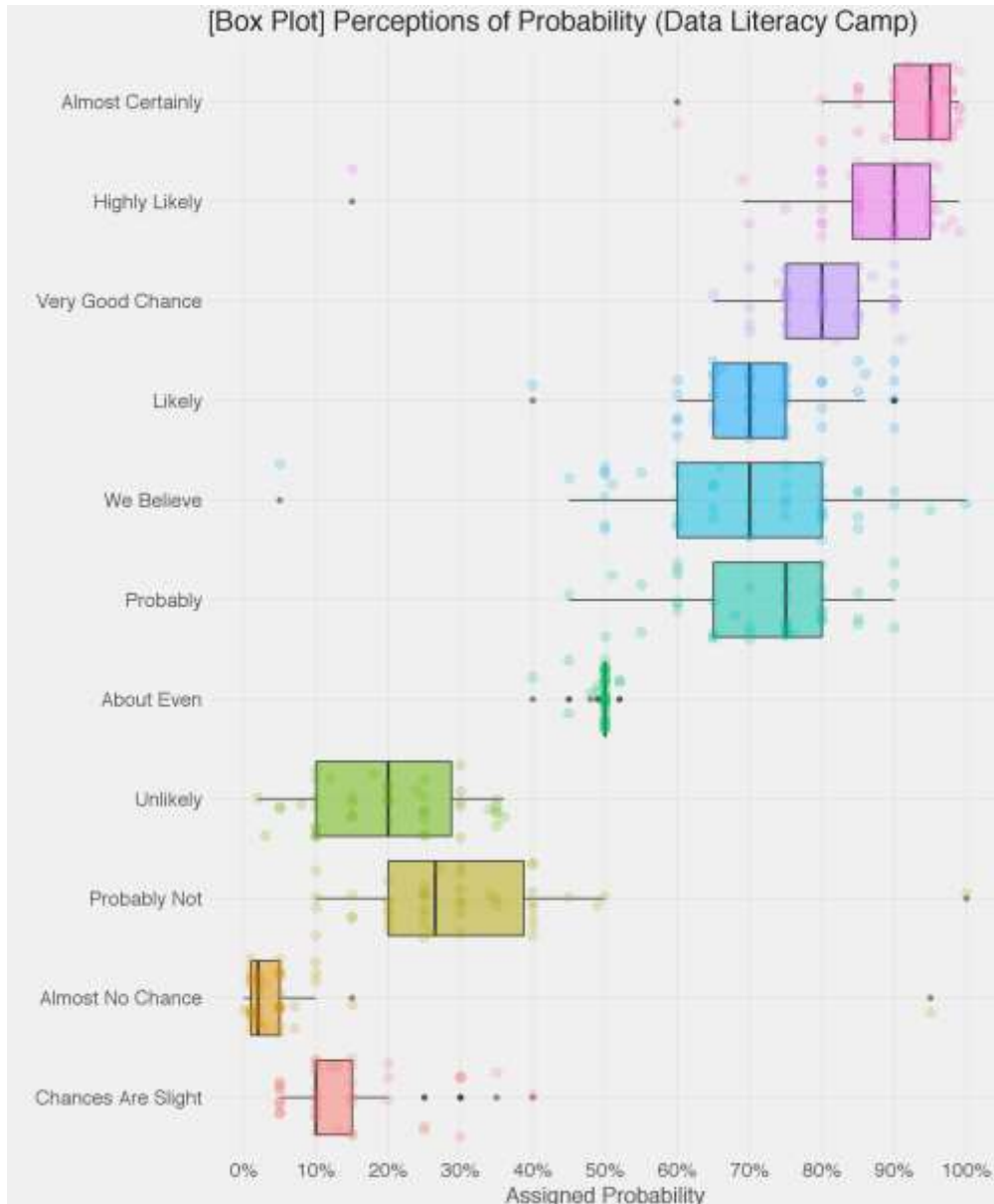
- 특이값(outlier): 중양값에서 편차가 큰 값
- IQR(Inter Quartile Range)의 1.5배 이상 벗어나 있는 값들을 Outlier로 정의

### 3 대칭성 및 분포

- 대칭성(symmetry): 최대값과 최소값까지의 수염 길이 비교

SK 내부 강의자료 참고

# Distribution (boxplot)



- Q. 아마 아닐꺼야(Probably Not)에 대해**
- Range를 구해보세요.
  - 중앙값을 구해보세요.
  - 평균을 가늠해보세요.
  - 1Q(하위 25%)의 답변의 범위를 구해보세요.
  - 응답자의 75%는 최소 몇%(X축) 이상으로 답변했나요?

**Q. 분산(퍼진 정도)이 가장 큰(작은) Phrase는 무엇가요?**

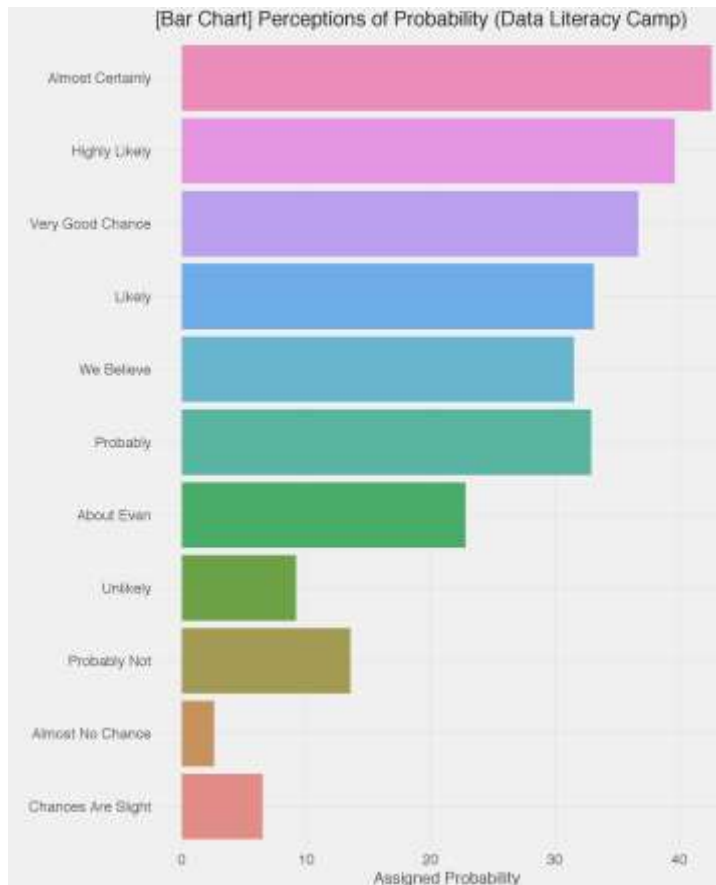
**Q. 음(-)의 왜도(비대칭성)가 가장 큰 Phrase는 무엇인가요?**



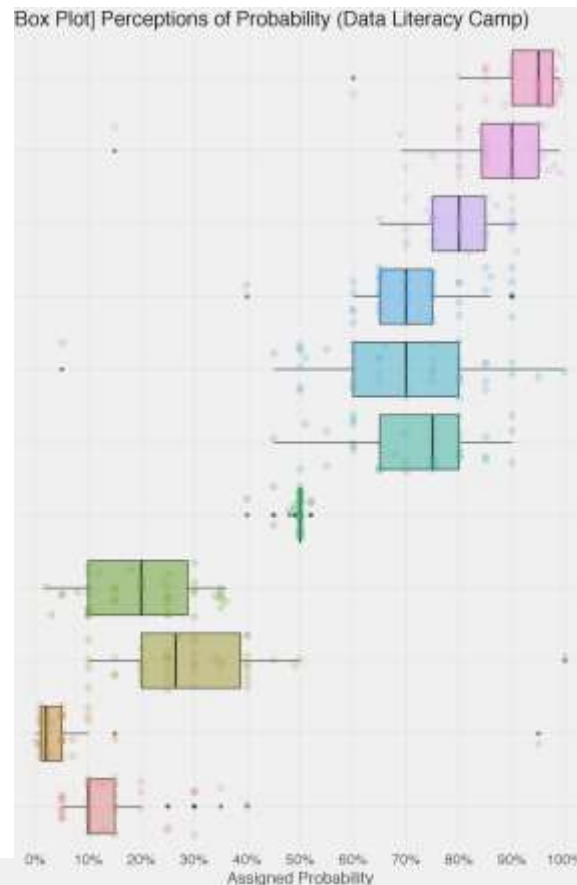


## 데이터의 특성과 모양을 시각화하는 방법

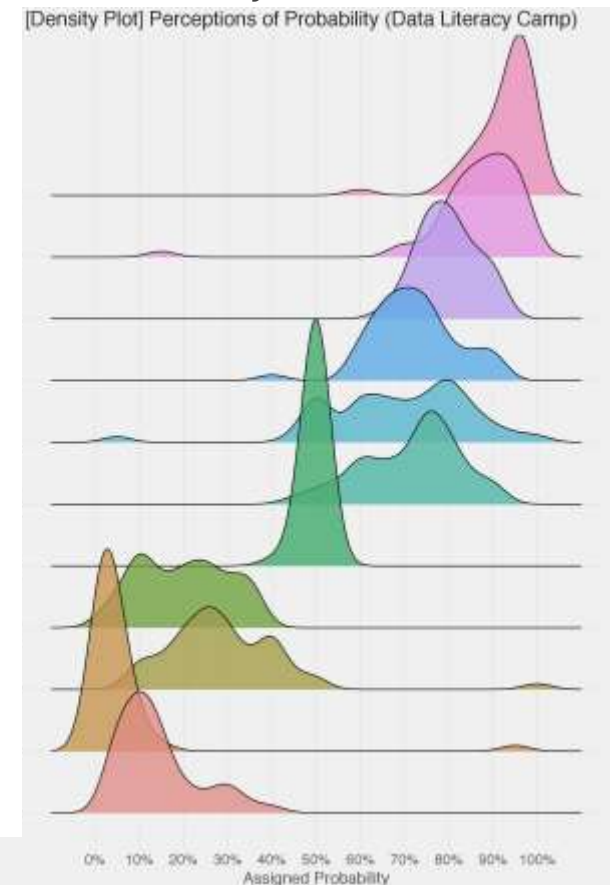
### Bar Chart (평균)



### Box Plot (분포)



### Density Plot (분포)



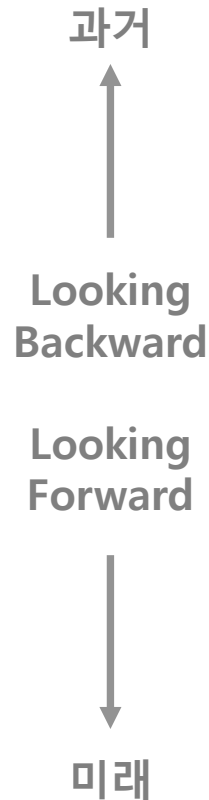
# Module II-a

## Data Understanding

아이디케이스퀘어드 양승준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)  
<https://www.heartcount.io>



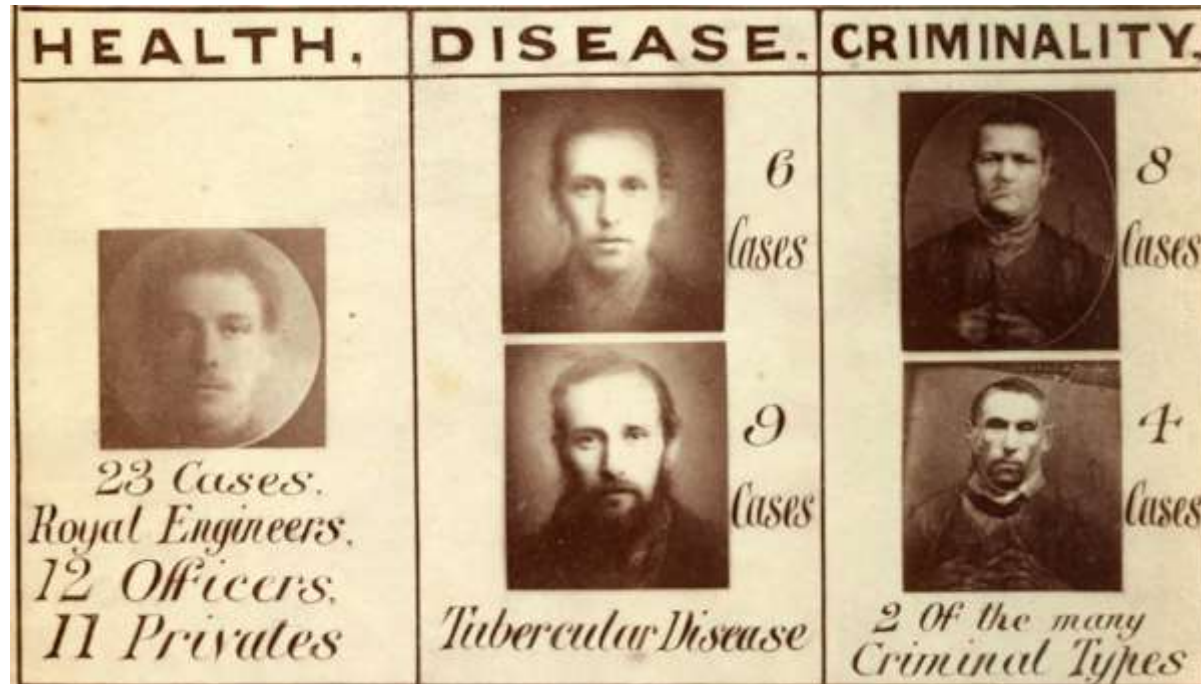
## 데이터 분석 주요기술



- **DESCRIBE (기술 분석) - 엑셀, 대쉬보드**
  - 데이터 특성과 모양을 요약
- **EXPLORE (탐험적 분석) - 데이터 시각화 도구**
  - 가설수립·데이터 감 잡기 위해 패턴 탐험
- **PREDICT/INFER (예측·추론 분석) - 통계/ML**
  - 패턴(모형)을 통해 주어진 문제를 예측·설명

# 평균의 문제 - Data Aggregation

## Average Man – Galton's Composite Portraits



## The Problems with Average: Not Robust!



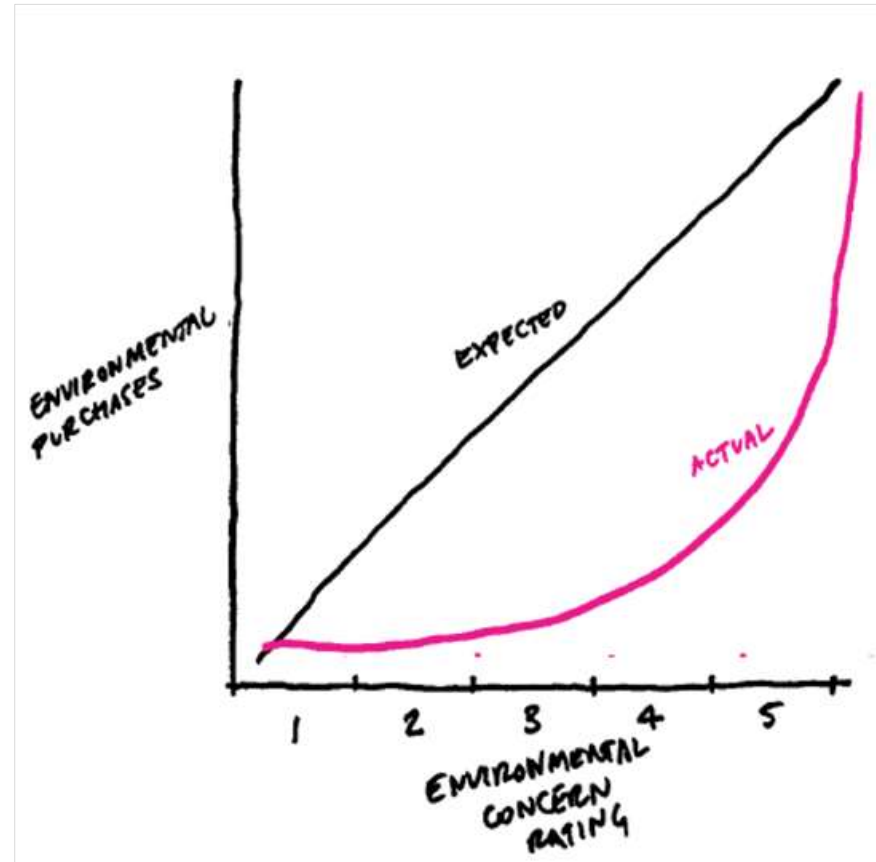
## 평균의 문제 - Linear Thinking vs. Non-Linear Relationship

Q. 친환경 제품을 출시하려 한다.  
어떤 세그먼트에 프로모션해야 하나?

고객 세그먼트	평균 점수
A	4
B	3

A = [4, 4, 4, 4, 4, 4, 4, 4, 4]

B = [1, 1, 1, 1, 5, 5, 5, 5, 5]



## 제한된·익숙한 관점 - Simpson's Paradox

### 심슨의 역설

뭉뚱그린 수치는 현실을 왜곡할 수 있음  
쪼개보는 일(Segmentation; Drill-Down; Dimensions)의 중요성

남녀 지원자 합격률

	지원자 수	합격자 수	합격률
여자	1,000	150	15%
남자	1,000	250	25%

문과대 합격률

	지원자 수	합격자 수	합격률
여자	800	80	10%
남자	200	10	5%

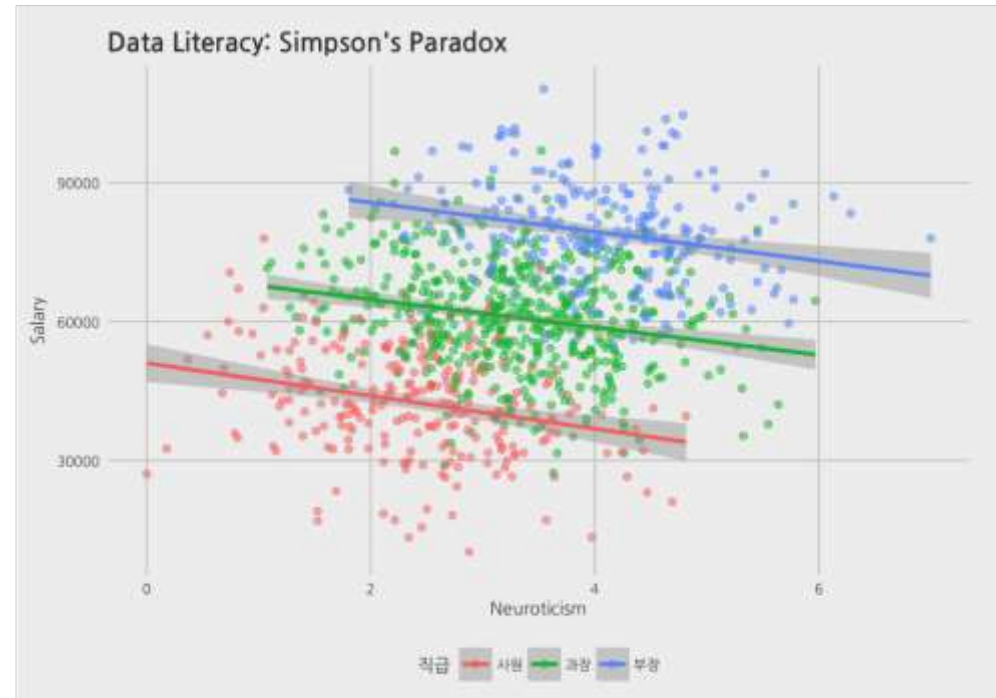
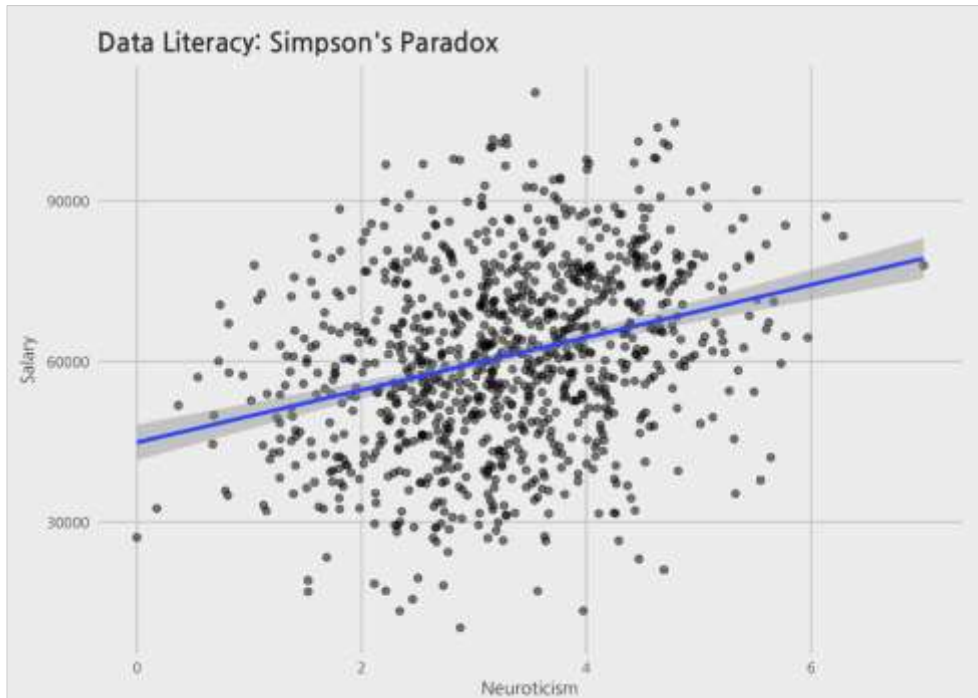
이공대 합격률

	지원자 수	합격자 수	합격률
여자	200	70	35%
남자	800	240	30%

## 제한된·익숙한 관점 - Simpson's Paradox

### 새로운 관점(Dimension)

연봉과 까칠함과의 관계 → 직급별 연봉과 까칠함과의 관계





### Story, not Data, should dictate our choice

Phone 전환률이 중요한 경우  
유료 전환 사용자수가 중요한 경우

Mobile App Conversion Rate: 5.00%

	iOS	Android
Devices	5000	10000
Conversinos	200	550
Conversion Rate	<u>4.00%%</u>	<u>5.50%</u>

	iOS		Android	
	Tablet	Phone	Tablet	Phone
Devices	1500	3500	8000	2000
Conversinos	100	100	500	50
Conversion Rate	<u>6.67%</u>	<u>2.86%</u>	<u>6.25%</u>	<u>2.50%</u>

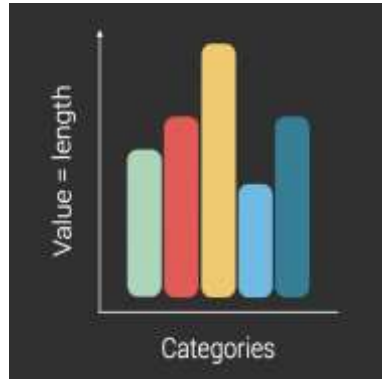
# The Curse of Dimensionality

## 차원의 저주

“If you test enough things, just by random chance, one of them will be statistically significant.”

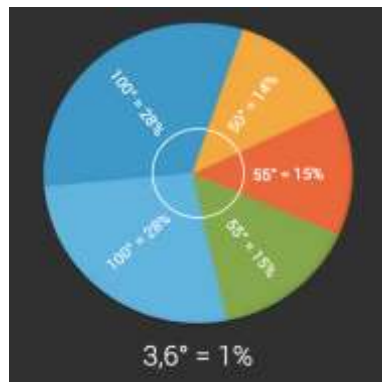
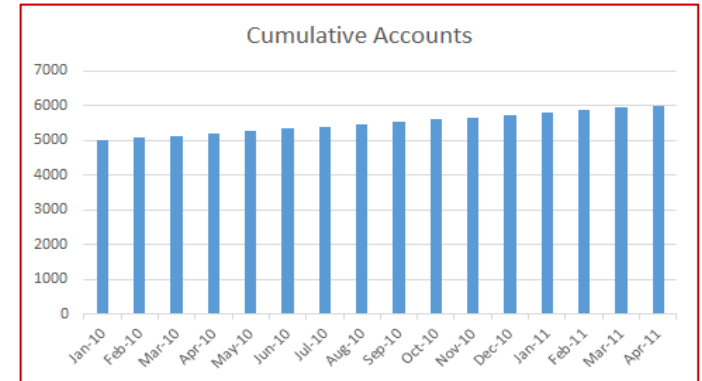
	S&P 지수	Coin 1	Coin 2	Coin 3	Coin 4	Coin 5	.....	.....	Coin 1217	.....	.....	Coin 1999	Coin 2000
1	상승	앞	뒤	앞	앞	뒤			앞			앞	뒤
2	하락	뒤	뒤	앞	앞	뒤			앞			뒤	뒤
3	상승	뒤	앞	뒤	앞	뒤			앞			뒤	앞
4	상승	앞	뒤	앞	뒤	뒤			앞			앞	뒤
5	하락	뒤	앞	뒤	앞	앞			뒤			뒤	뒤
...													
248	상승	앞	뒤	뒤	뒤	앞			앞			뒤	뒤
249	하락	뒤	앞	뒤	뒤	뒤			뒤			뒤	앞
250	하락	앞	뒤	앞	앞	앞			뒤			뒤	뒤

## 1 Dimension



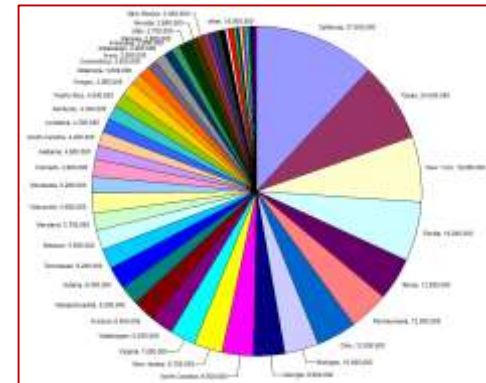
### Bar Chart

- 서로 다른 범주(사업부)간 평균값(매출 평균)의 차이를 비교하는데 효과적
- 시계열에 따른 변화를 표현하기에는 부적절



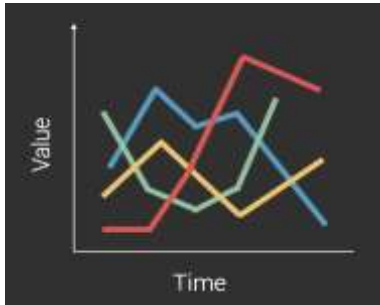
### Pie Chart

- 서로 다른 범주가 전체(매출 총합)에서 차지하는 비율을 대비하는데 효과적
- 범주의 갯수가 많거나 비율이 비슷한 경우 부적절



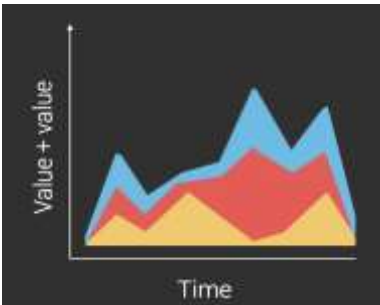
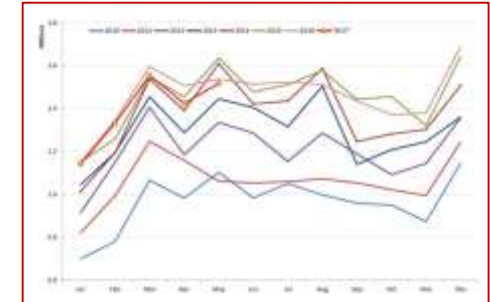


## 2 Dimensions



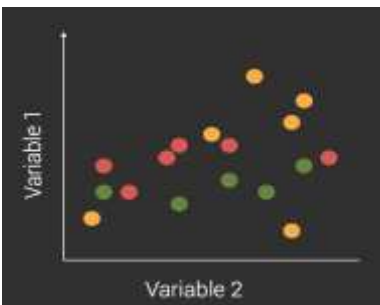
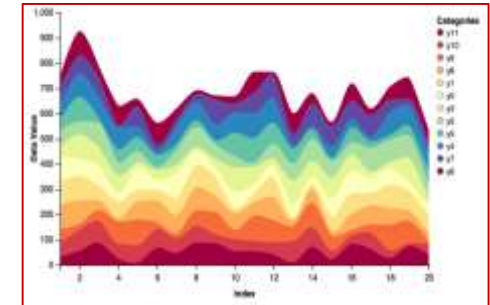
### Line Chart

- 시계열에 따른 추세를 보는데 효과적
- 너무 많은 범주(4~5개 이상)가 함께 표현되는 경우 헷갈림



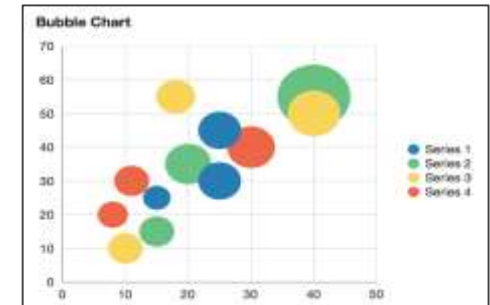
### Stacked Area Chart

- 시간의 흐름에 따른 개별 범주의 전체 크기 내에서의 상대적 크기 변화를 표현하는데 효과적
- 범주가 너무 많으면 역시 헷갈림



### Scatter Plot

- 두개의 숫자형 변수 간 관계를 파악하는데 효과적
- Outlier를 발견하는데도 효과적
- 원 크기/색상으로 4 Dimensions 표현

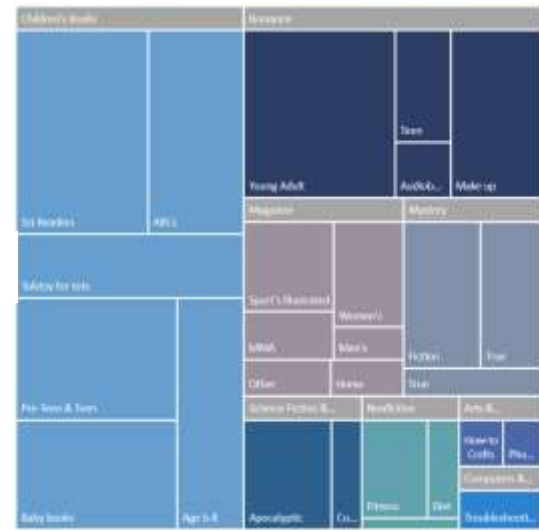


# Data Visualization 101

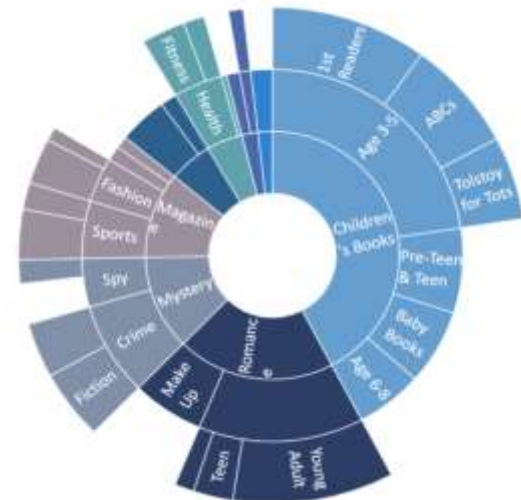
## Hierarchical Data

GENRE	SUB-GENRE	TOPIC	REVENUE
ARTS & PHOTOGRAPHY	How-to Crafts	How-to Crafts	\$ 2,711
ARTS & PHOTOGRAPHY	Coffee-table	Photography	\$ 2,309
CHILDREN'S BOOKS	Baby Books	Baby Books	\$ 16,092
CHILDREN'S BOOKS	Age 3-5	1st Readers	\$ 24,514
CHILDREN'S BOOKS	Age 3-5	ABCs	\$ 17,771
CHILDREN'S BOOKS	Age 3-5	Tolstoy for Tots	\$ 13,295
CHILDREN'S BOOKS	Age 6-8	Age 6-8	\$ 14,046
CHILDREN'S BOOKS	Pre-Teen & Teen	Pre-Teen & Teen	\$ 18,046
COMPUTERS & INTERNET	Troubleshooting	Troubleshooting	\$ 4,527
MYSTERY	Crime	Fiction	\$ 11,186
MYSTERY	Crime	True Crime	\$ 8,790
MYSTERY	Spy	Spy	\$ 6,516
MYSTERY	Spy	True Spy	\$ 3,809
NONFICTION	Health	Diet	\$ 3,293
NONFICTION	Health	Fitness	\$ 6,891
NONFICTION	History	History	\$ 1,131
MAGAZINE	Fashion	Women's	\$ 7,315
MAGAZINE	Fashion	Men's	\$ 2,222
MAGAZINE	Home	Home	\$ 2,612
MAGAZINE	Other	Other	\$ 3,140
MAGAZINE	Sports	Sport's Illustrated	\$ 8,009
MAGAZINE	Sports	MMA	\$ 4,257
ROMANCE	Break up	Teen	\$ 6,205
ROMANCE	Break up	Young Adult	\$ 25,193
ROMANCE	Break up	Audiobooks	\$ 3,045
ROMANCE	Make Up	Make Up	\$ 15,050
SCIENCE FICTION & FANTASY	Apocalyptic	Apocalyptic	\$ 10,200
SCIENCE FICTION & FANTASY	Comics	Comic	\$ 3,456

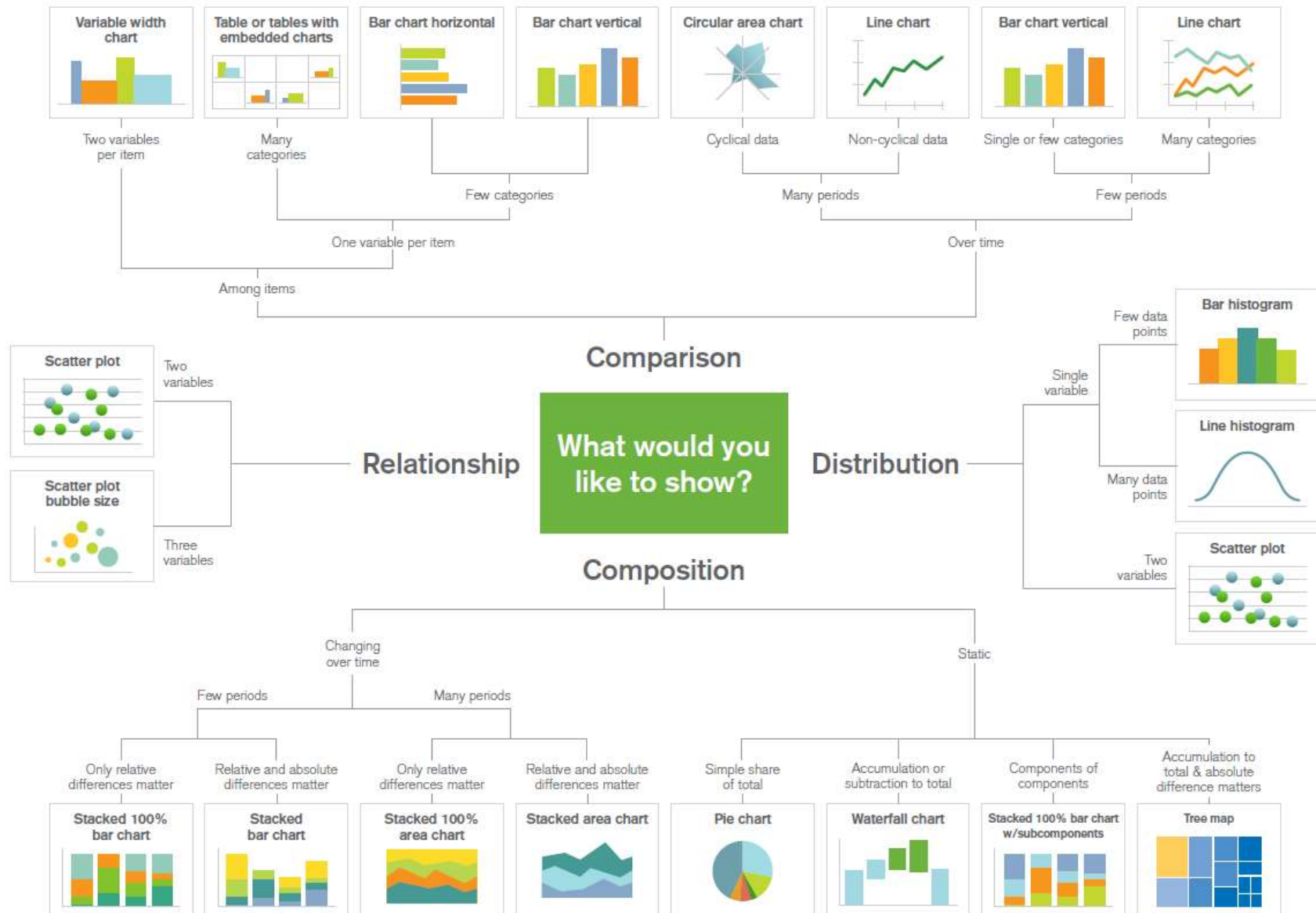
Treemap



Sunburst

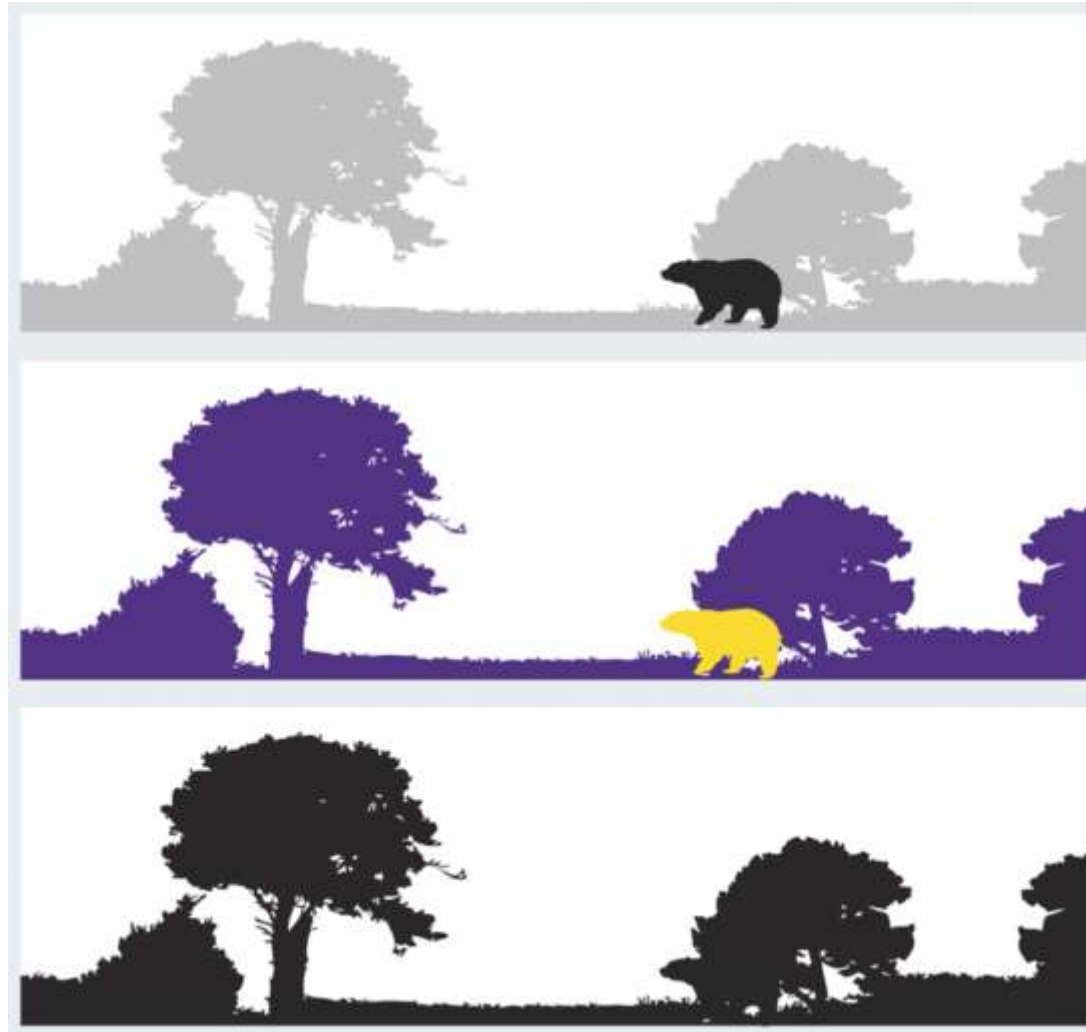


# Data Visualization 101 – 기본 문법 속지



# Data Storytelling – Pre-attentive Processing

전주의 처리 (Pre-attentive Processing)  
주의를 기울이지 않고도 핵심 정보를 인지

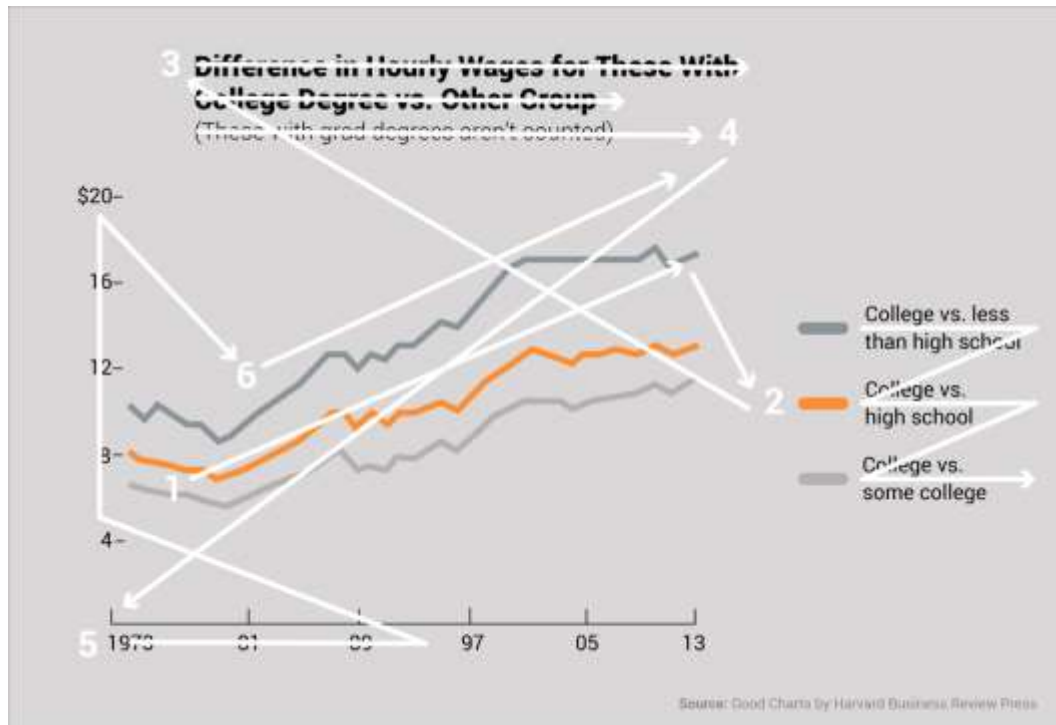


# Data Storytelling – Pre-attentive Processing

## 전주의 처리 (Pre-attentive Processing)

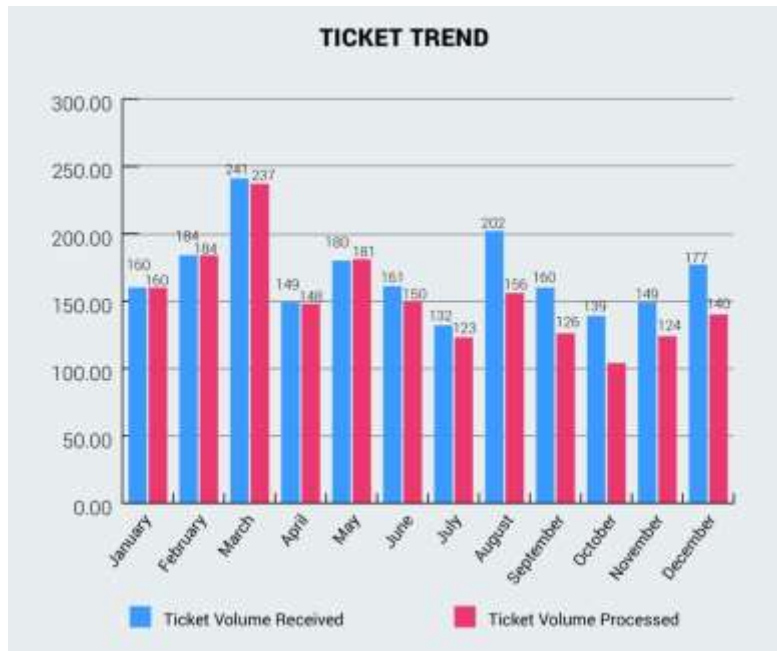
주의를 기울이지 않고도 핵심 정보를 인지하도록 적절히 강조

차트를 볼 때 사람의 눈동자가  
어떤 순서로 어디로 향할지 알 수 없음



# Data Storytelling – Before and After

## Before

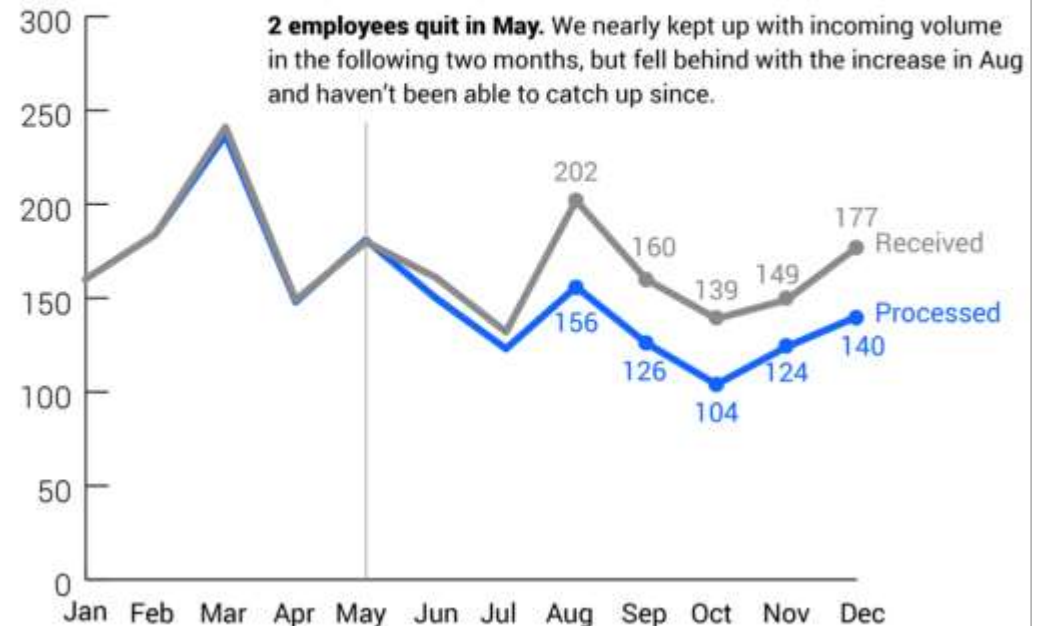


## After: Call to Action

### Please approve the hire of 2 FTEs

to backfill those who quit in the past year

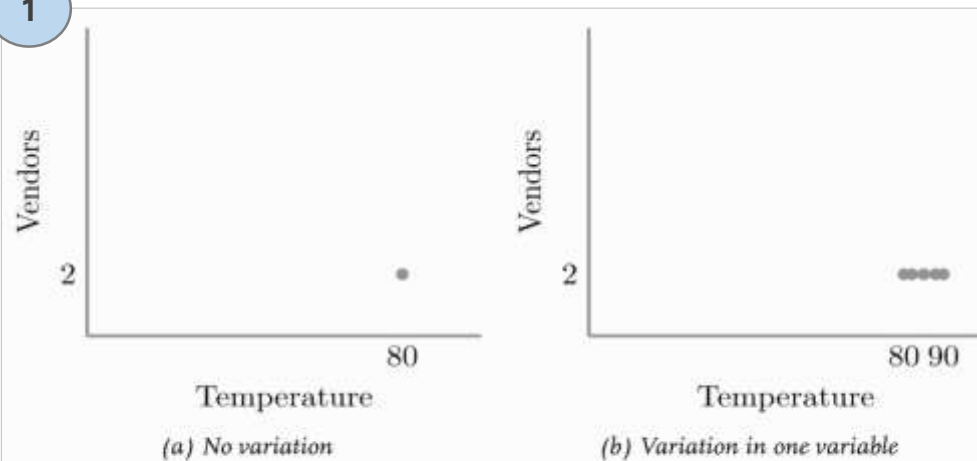
Ticket volume over time





# Correlation and Scatter Plot (source: WHY)

1



**Figure 3-2. Without variation in both variables, we cannot find a correlation.**

3

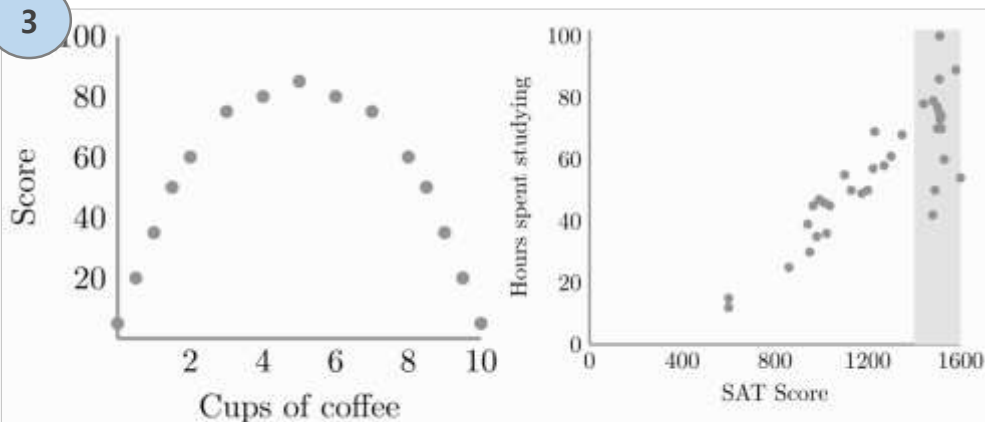
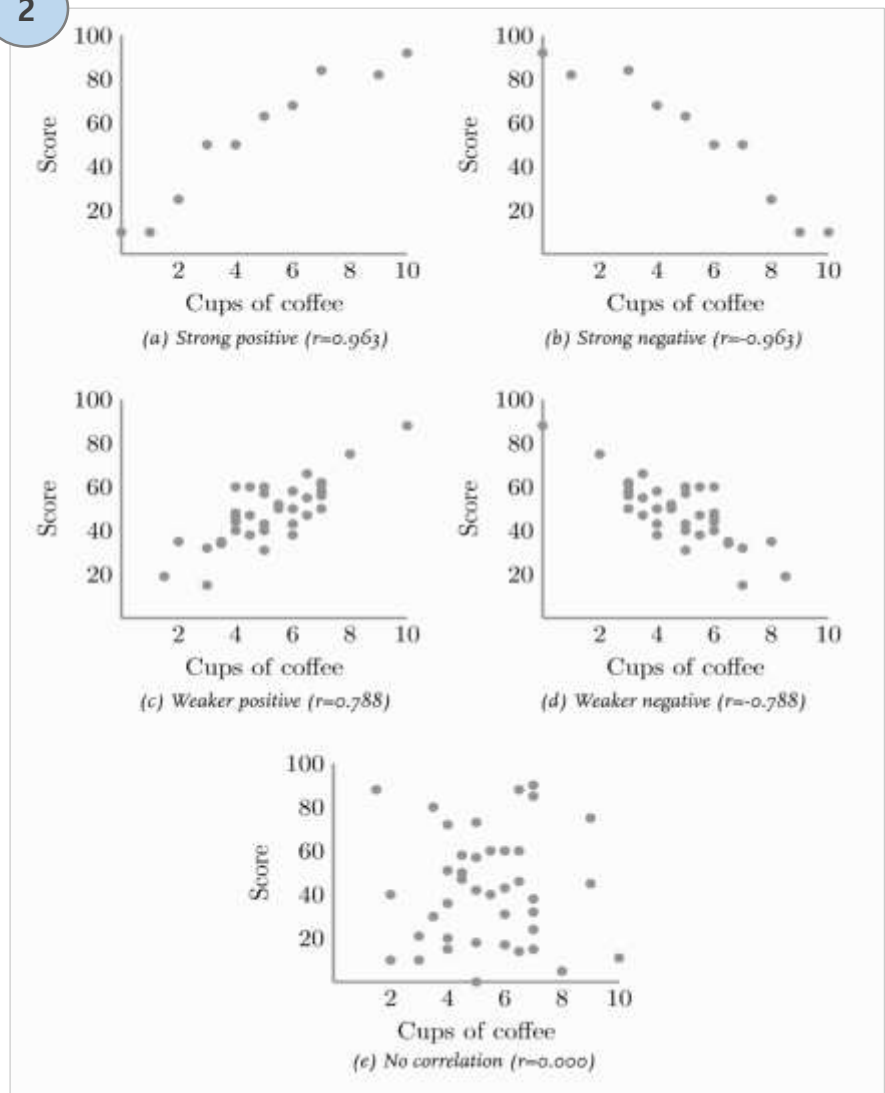


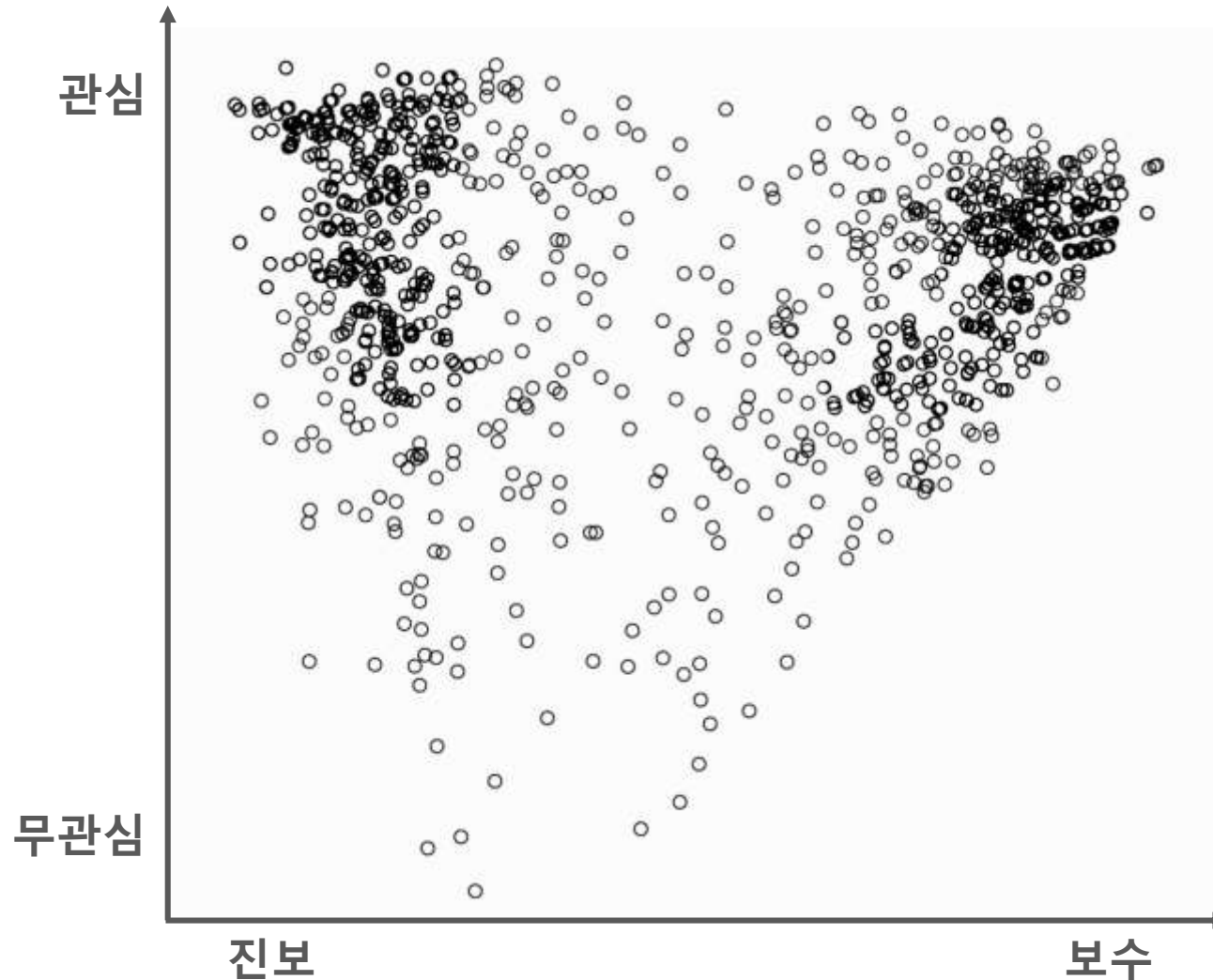
Figure 3-4. Nonlinear relationship ( $r=0.000$ ).

Figure 3-5. Data from only the shaded area represents a restricted range.

2



### “정치적 관심도”와 “정치적 성향(보수-진보)” 간 관계 해석





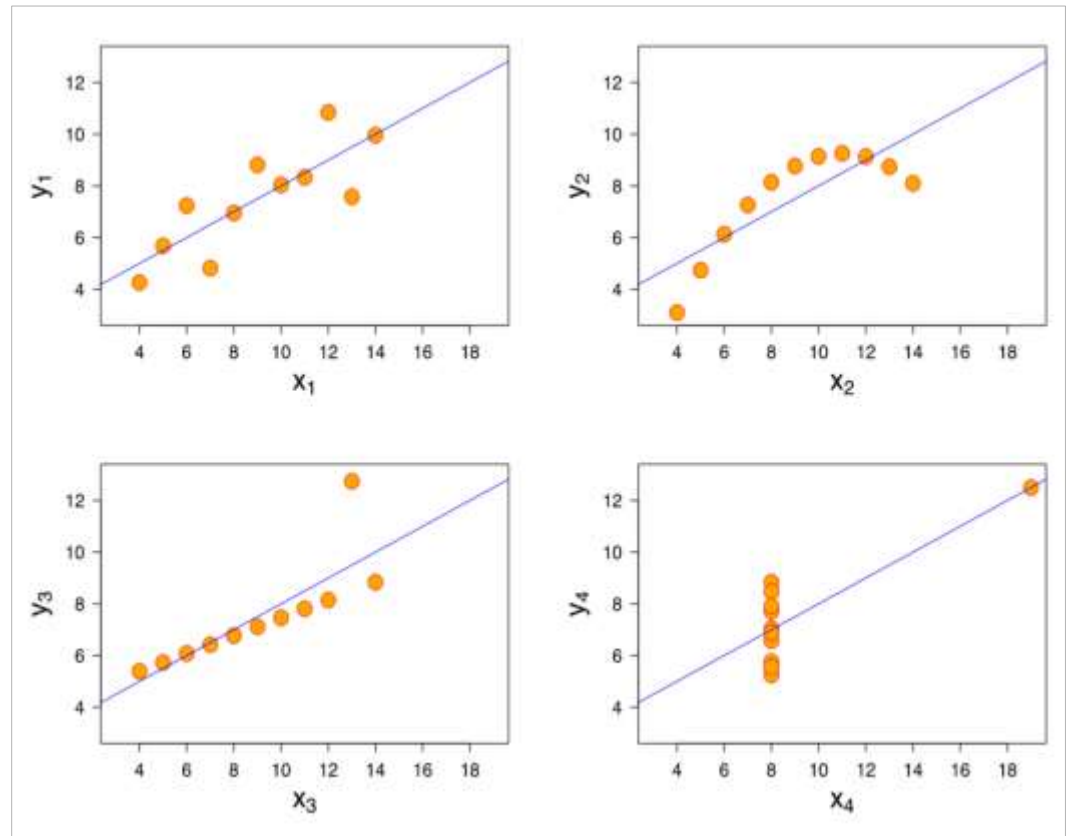
# 데이터 요약 & 시각화

## 동일한 평균, 분산, 상관계수

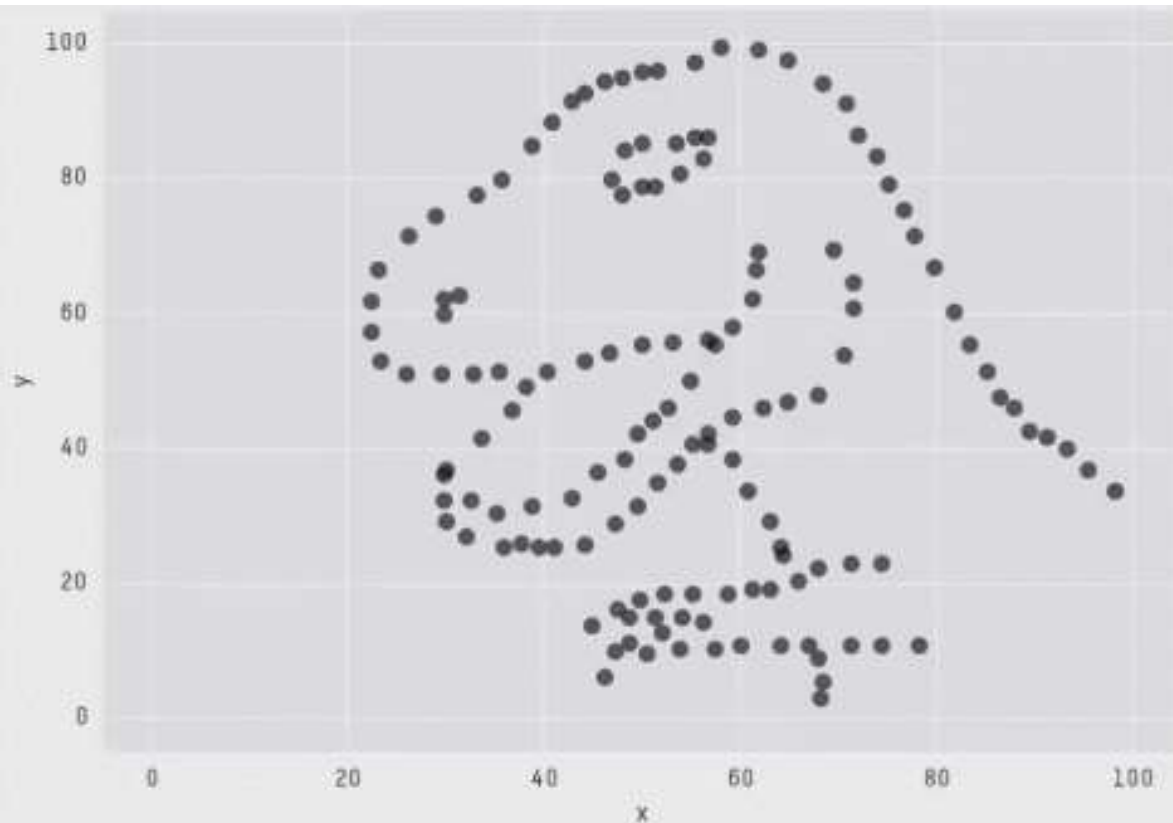
	I		II		III		IV	
	X	Y	X	Y	X	Y	X	Y
평균	9	7.5	9	7.5	9	7.5	9	7.5
분산	11	4.1	11	4.1	11	4.1	11	4.1
상관계수	0.82		0.82		0.82		0.82	

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

## 시각화를 통해 현실의 복잡성이 드러남



## 통계치와 시각화 결과를 함께 확인

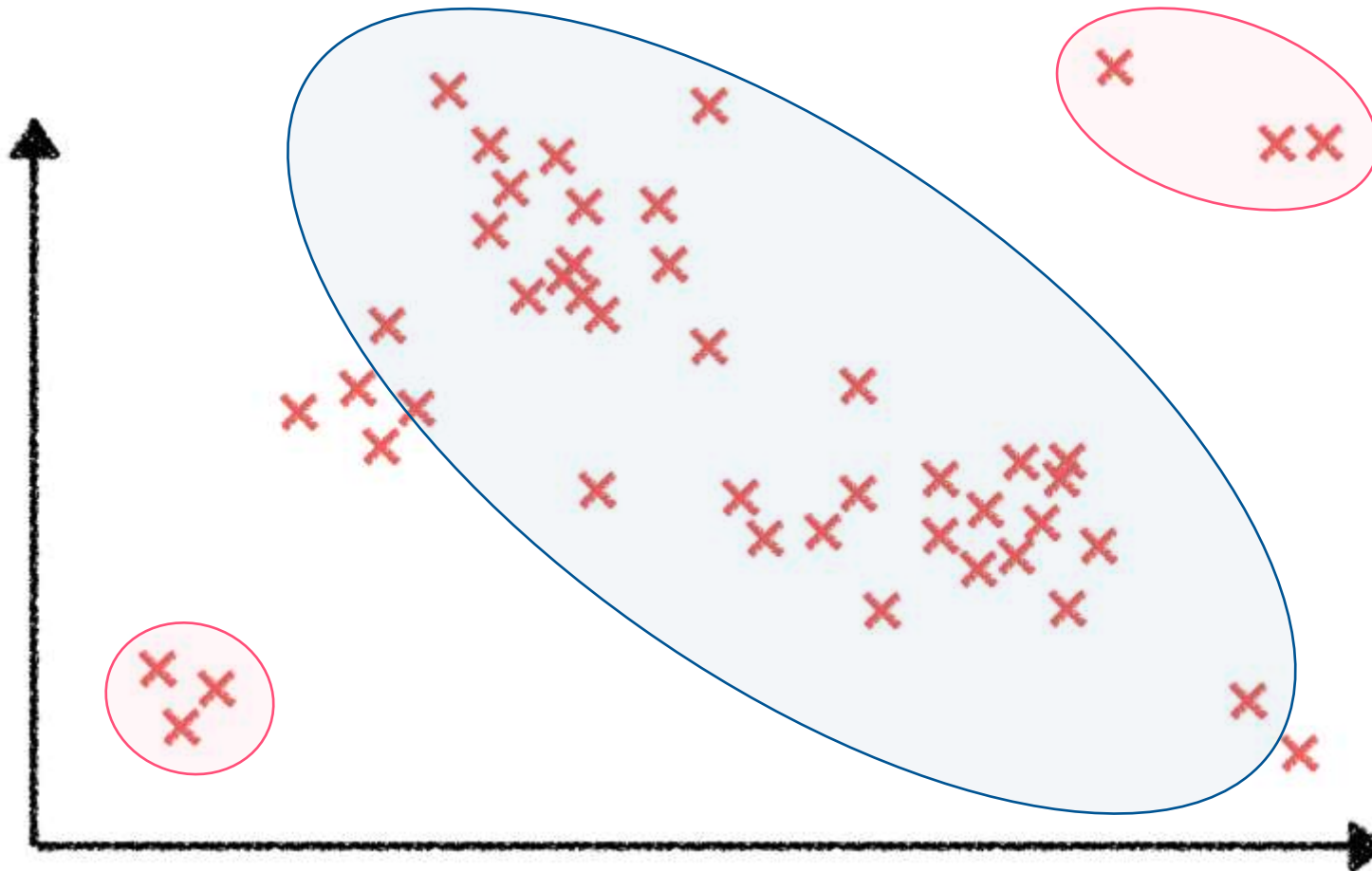


X Mean: 54.2659224  
Y Mean: 47.8313999  
X SD : 16.7649829  
Y SD : 26.9342120  
Corr. : -0.0642526

# Data Viz: Understand and Explain Data

**Signal**  
패턴, 일반적 경향

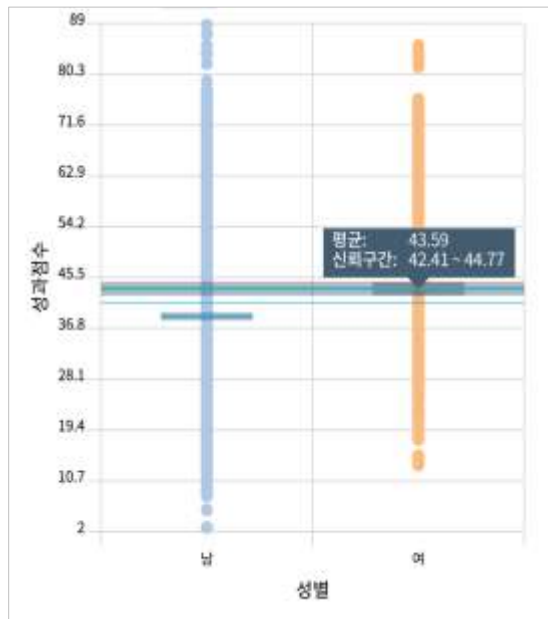
**Noise?**  
이상한 애? vs. 특별한 애?



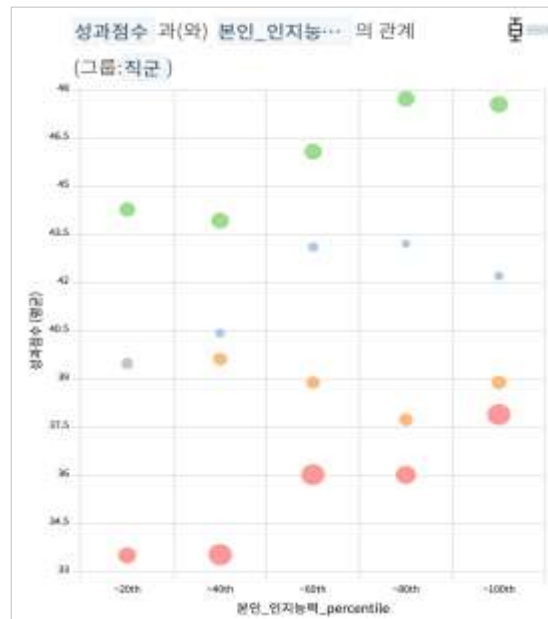
## Data: A New Language of Business Data Viz: A Medium of Data Communication

### 목적에 맞는 데이터 시각화 방법 선택

#### Visual Confirmation 가설을 시각적으로 검증



#### Visual Exploration 뭐라도 하나 걸렸으면



#### Visual Affirmation 검증된 사실을 주장/보고



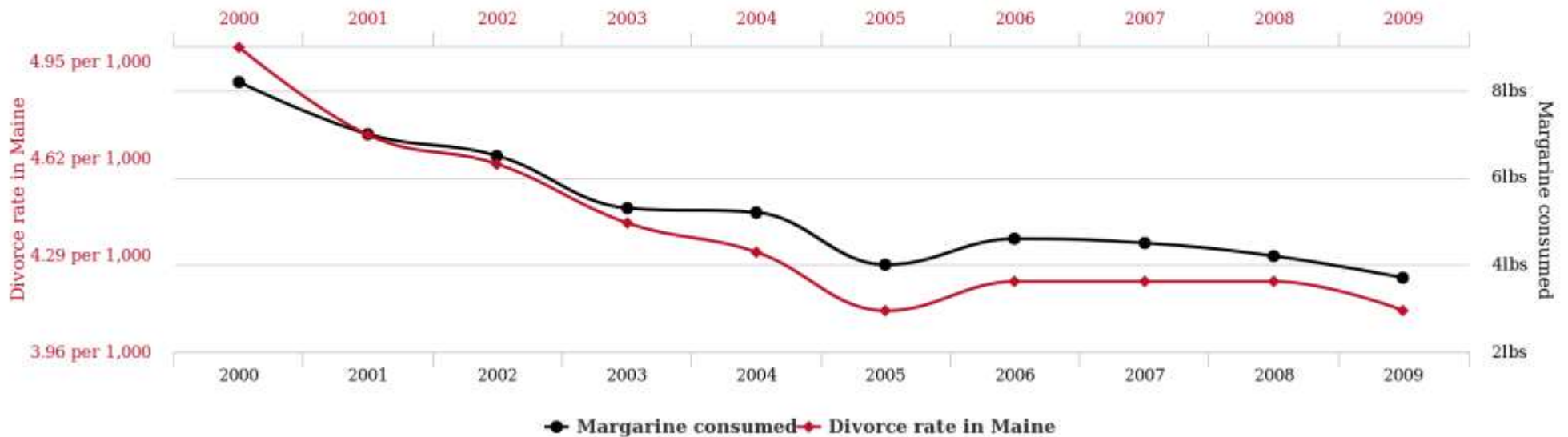


## Spurious Correlation (허위상관)

마요네즈 덜 먹으면 이혼을 덜 할까?

참고) <http://www.tylervigen.com/spurious-correlations>

### Divorce rate in Maine correlates with Per capita consumption of margarine



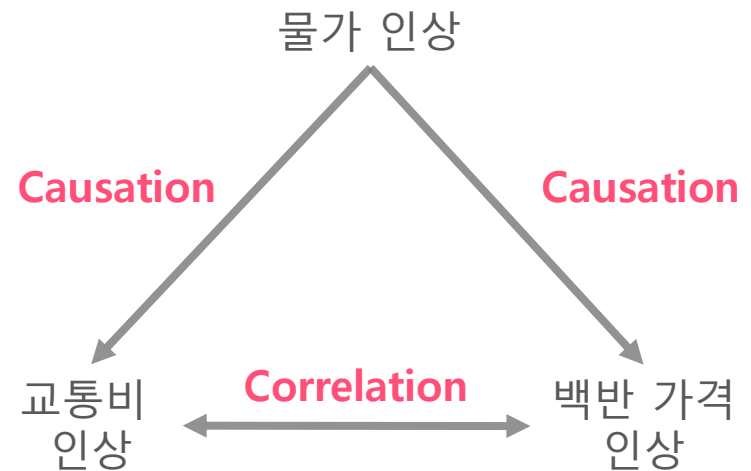
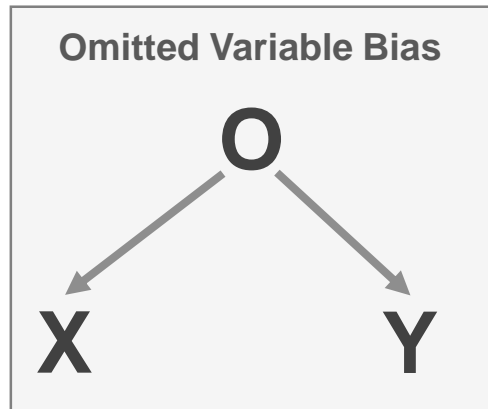
tylervigen.com

# Correlation vs. Causation

**Correlation helps you predict the future;  
Causality lets you change the future.**

- Correlation: X의 변화로 Y의 변화 예측 가능
- Causation: X에 개입해서 Y를 바꿀 수 있음

대표적인 오류

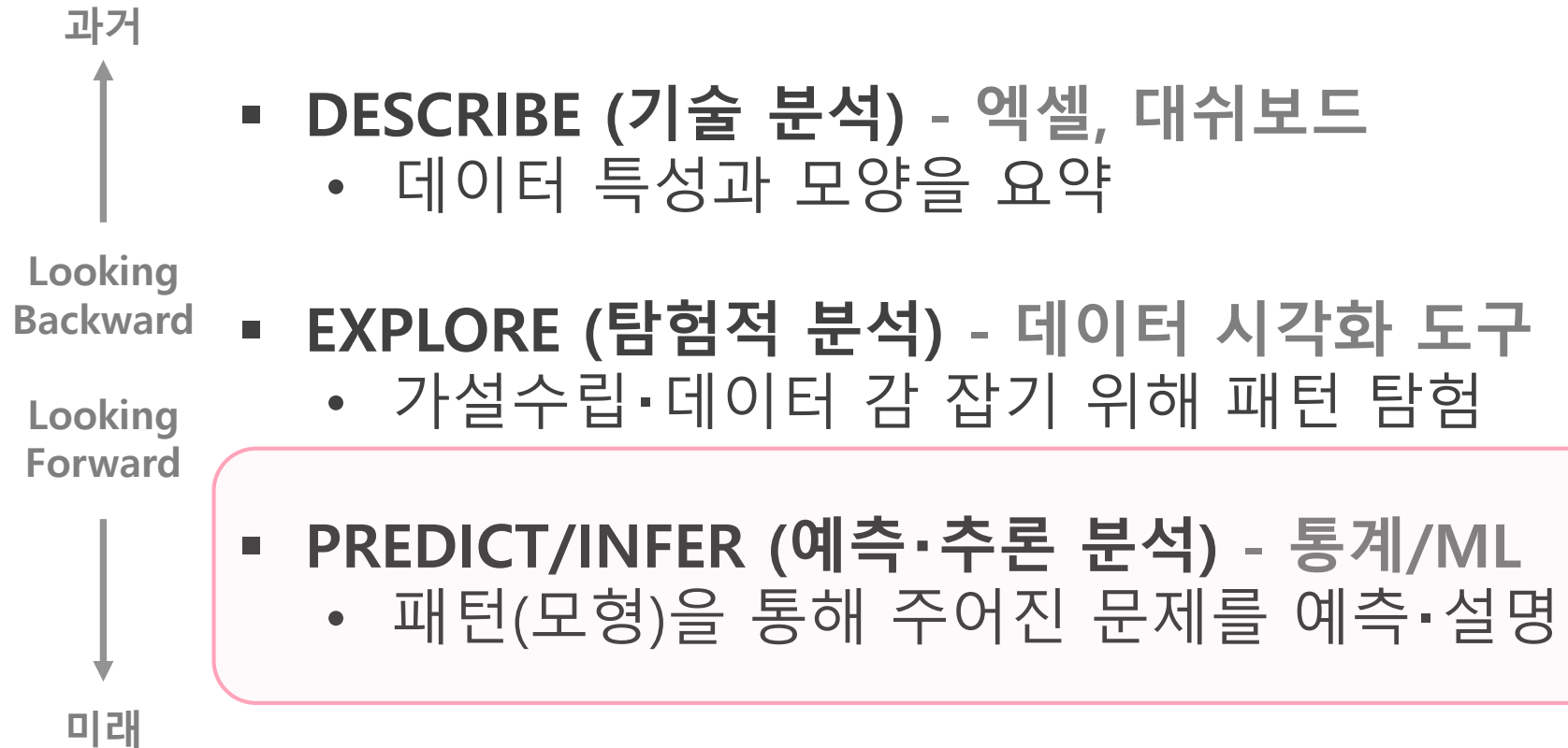


# Module III

## Machine Learning and Enterprise Decision-Making

아이디케이스퀘어드 양승준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)  
<https://www.heartcount.io>

## 데이터 분석 주요기술

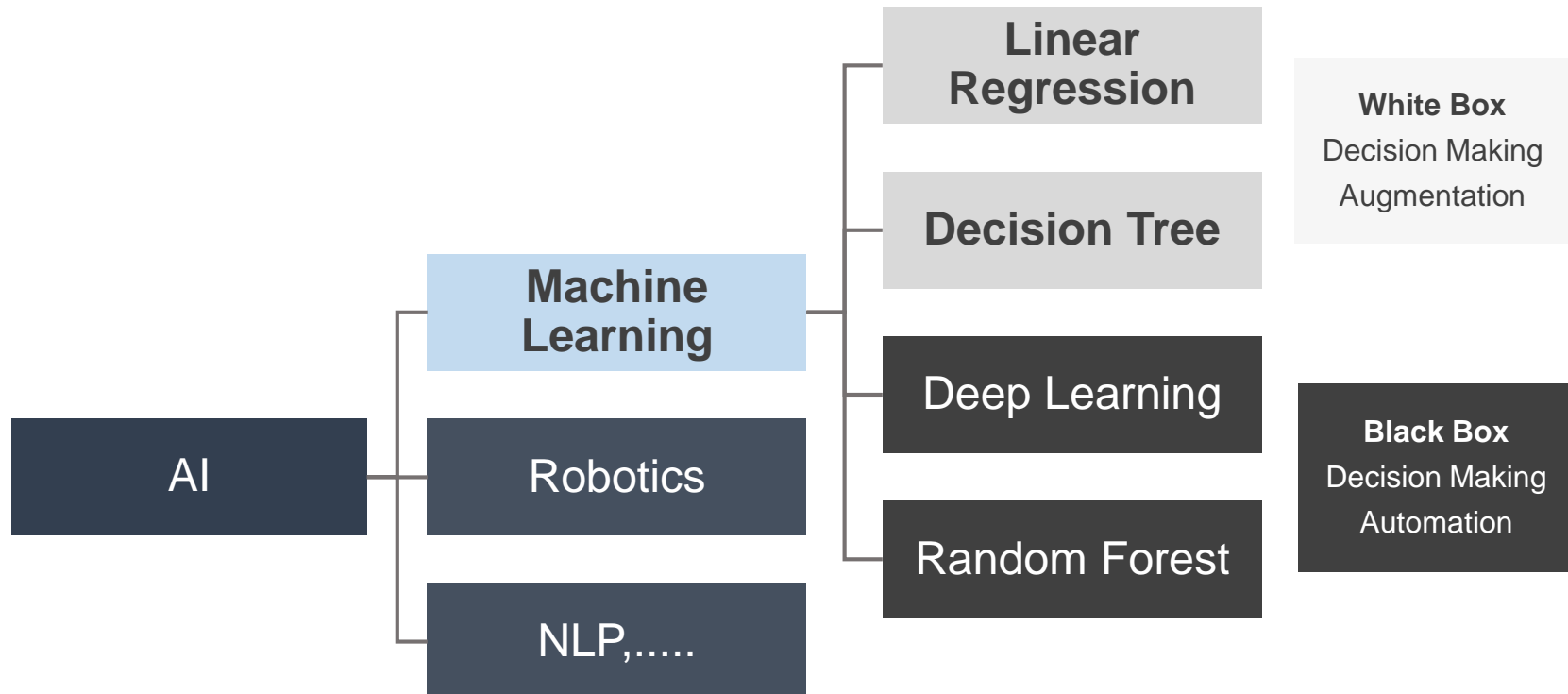




# Machine Learning and Decision Making

## Machine Learning: Decision Making Technology

- 인공지능의 한 분야로 사람으로부터 아주 **제한된** **지시**만 받고 데이터를 **학습**하여 패턴을 추출, 의사결정에 활용하는 기술
- 기계에게 언어로 기술(**codify**)하기 힘든 것을 정답(Label)을 알고 있는 과거 데이터를 제공하여 학습하도록 함
- **폴라니의 역설**: 언어의 해상도가 인식의 해상도보다 낮다. (**할 줄은 아는데 어떻게 하는지 말 못함**)



# Machine Learning and Enterprise Decision Making

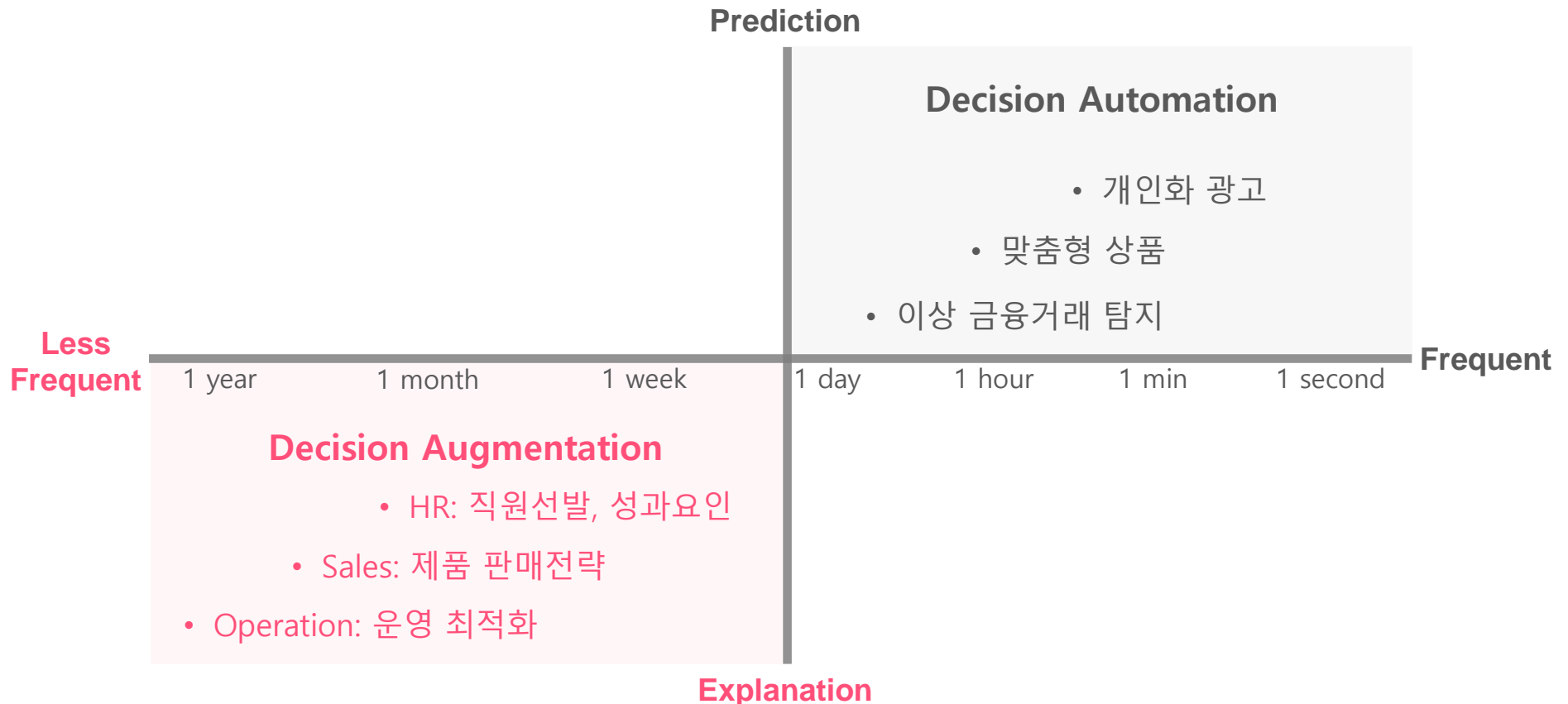
## Are You in Prediction Business or **Explanation Business**?

빈번한 덜 중요한 결정

빅 데이터 → 기계학습 → 예측 통한 의사결정 자동화

**덜 빈번한 중요한 결정**

**스몰 데이터 → 기계학습 → 사실·증거 기반 더 좋은 결정**



## Supervised Machine Learning (지도 학습)

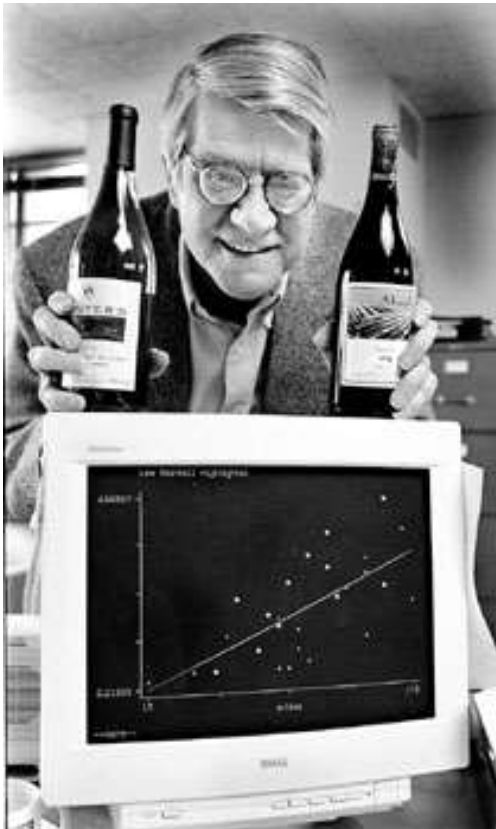
설명/예측하고 싶은 문제에 관련하여  
정답을 알고 있는 데이터가 충분히 있는 경우  
X로부터 Y를 유추하는 패턴을 찾는 것

X (Input)	Y (Output)	APPLICATION
Voice recording	Transcript	Speech recognition
Transaction records	Fraudulent (yes/no)	Fraud detection
Emails	Spam (yes/no)	Spam filtering
Ad + User Profile	Click (yes/no)	Personalized AD
Faces	Names	Face recognition
Korean	English	Language Translation

# Supervised Machine Learning

## Supervised Machine Learning (지도 학습)

모형이 투명한 경우(White Box), 설명·예측 둘 다 활용 가능



### 선형회귀 알고리즘

$$Y = a + b X_1 + c X_2 + d X_3$$

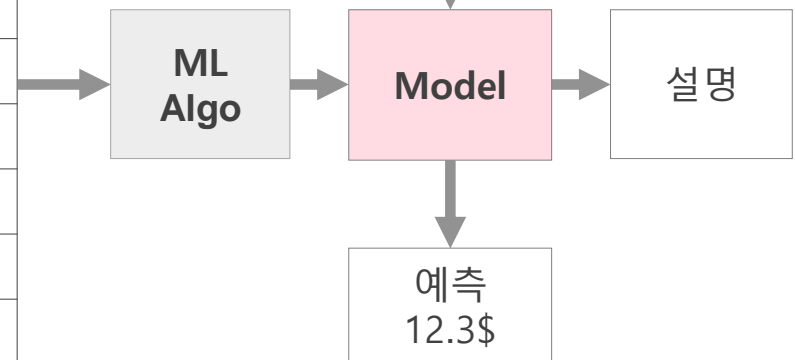
### Ashenfelter's Wine Formula

$$\text{Price} = 12 + 0.001 \text{ 겨울강수량} + 0.06 \text{ 평균온도} - 0.004 \text{ 수확철강수량}$$

### Training Data Set

$X_1$	$X_2$	$X_3$	$Y$
겨울강수량	수확철강수량	평균온도	와인가격
13	35	35	9.5
22	25	25	3.5
25	21	21	3.2
11	18	18	3.5
47	45	45	4.7
.	.	.	.

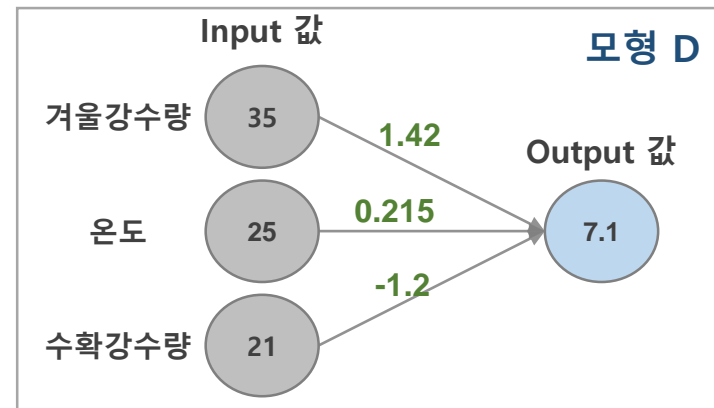
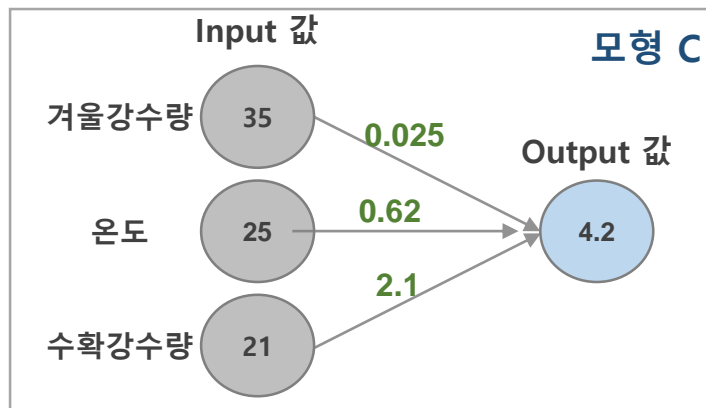
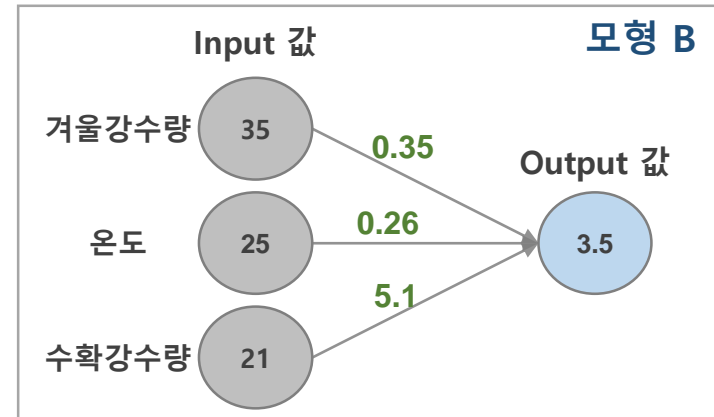
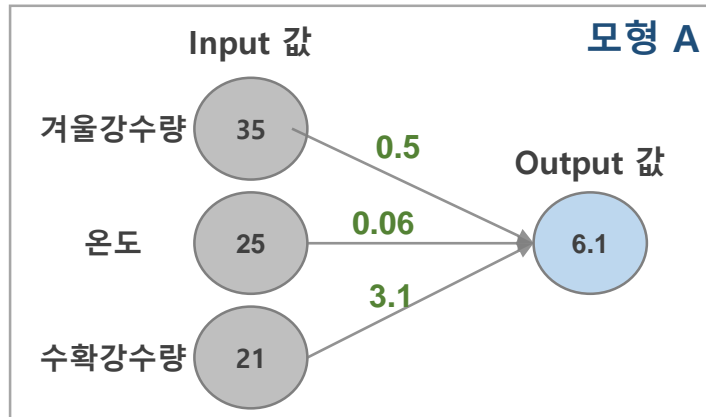
겨울강수량	수확철강수량	평균온도	와인가격
35	25	21	?



# White Box vs. Black Box

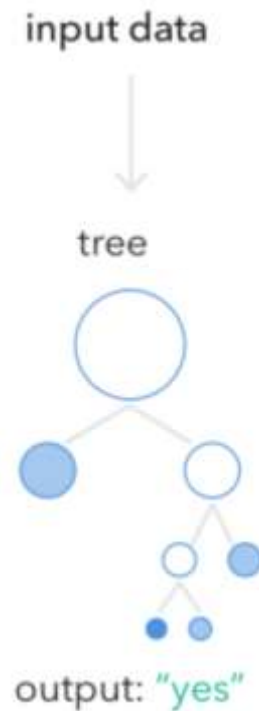
**Black Box:** 더 정확할지언정 인간이 이해하기 힘들다.

$$\text{가격} = 15.4\$ = 0.6 \times \text{A } 6.1 + 0.1 \times \text{B } 3.5 + 0.25 \times \text{C } 4.2 + 0.05 \times \text{D } 7.1$$

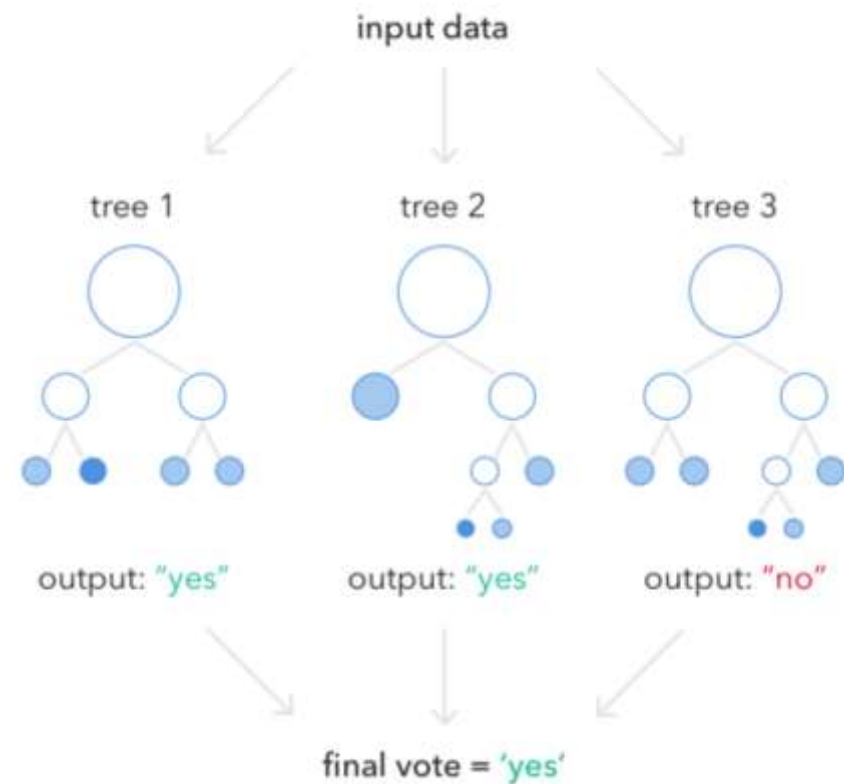


# Trade Off Between Interpretability and Accuracy

# Decision Tree



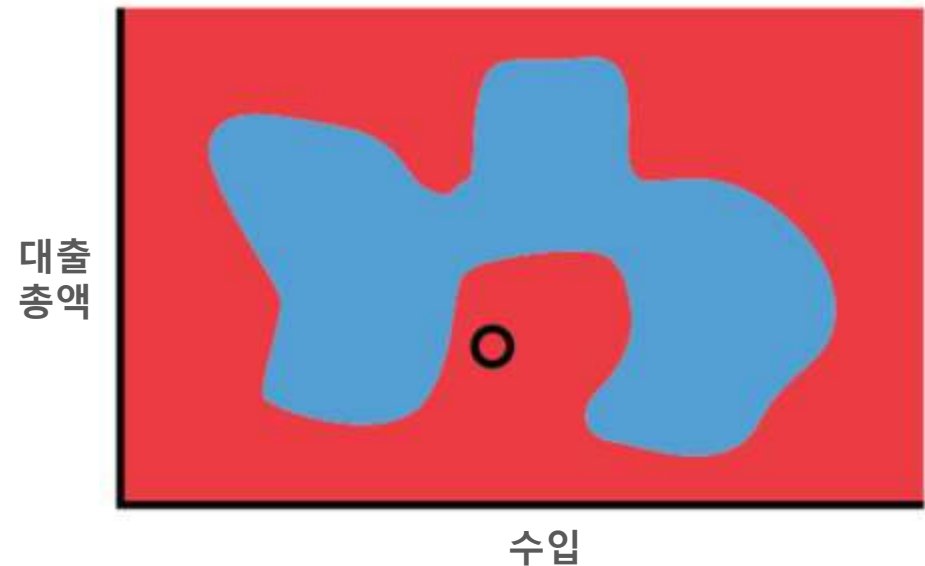
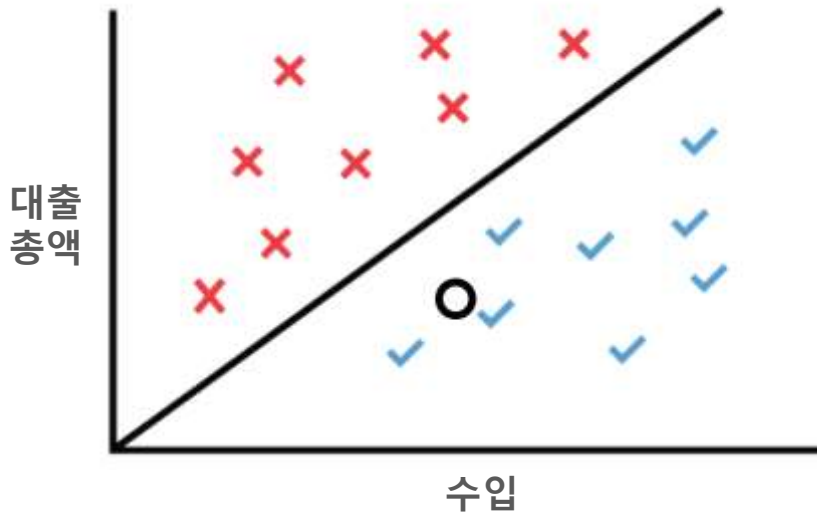
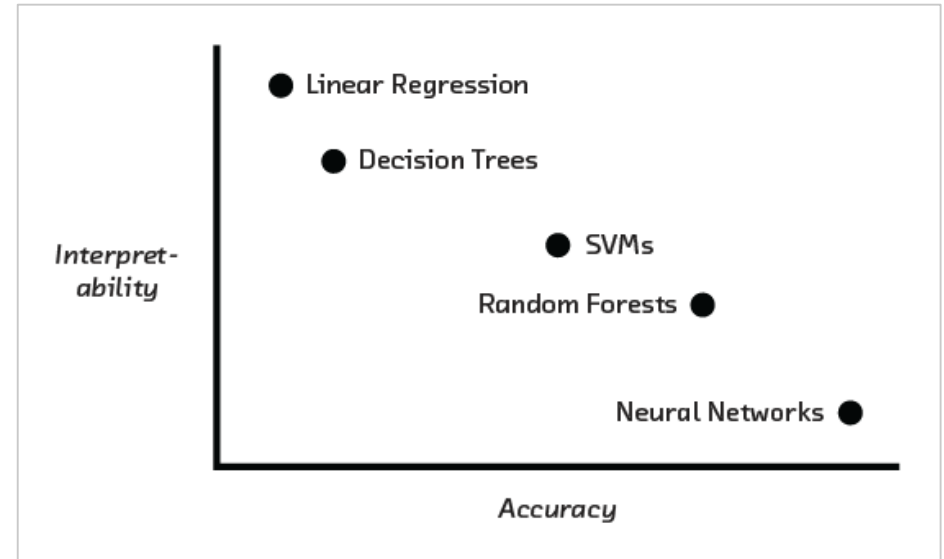
# Random Forest



# Model Interpretability

## 정확함 vs. 올바름

Model should be Accurate  
for the Right Reason



비즈니스 의사결정에  
예측모형을 활용하는 것이  
힘들고 조심스러운 이유

## Model Validity

데이터에 담기 어려운  
현실세계의 우발성, 복잡성으로  
모델의 정확도가 낮음

## Model Diversity

데이터 속에 내재된 편견이  
모델에 반영될 경우  
현실이 오히려 강화/공고화됨

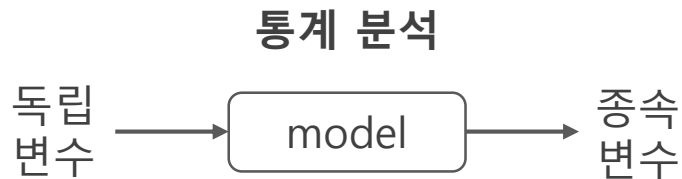
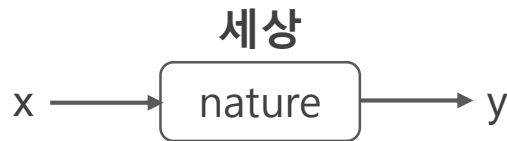




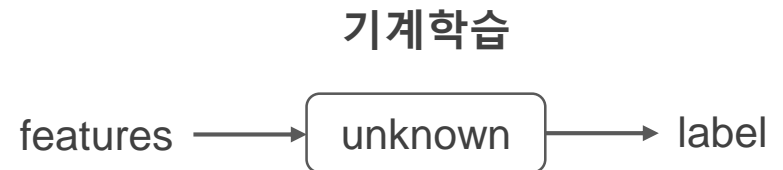
# Statistics vs. Machine Learning

*“Machine Learning is essentially a form of **applied statistics**”*

*“Machine Learning is statistics scaled up to **big data**”*



- **연역적 추론**(Deductive Reasoning): 가설수립 → 가설검증
- **모형의 타당성**: 가설 일반화를 위한 모형의 타당성·재현성이 중요



- **귀납적 추론**(Inductive Reasoning): 데이터 → 패턴
- **모형의 유용성**: 의사결정에 활용하기 위한 모형의 유용성이 중요

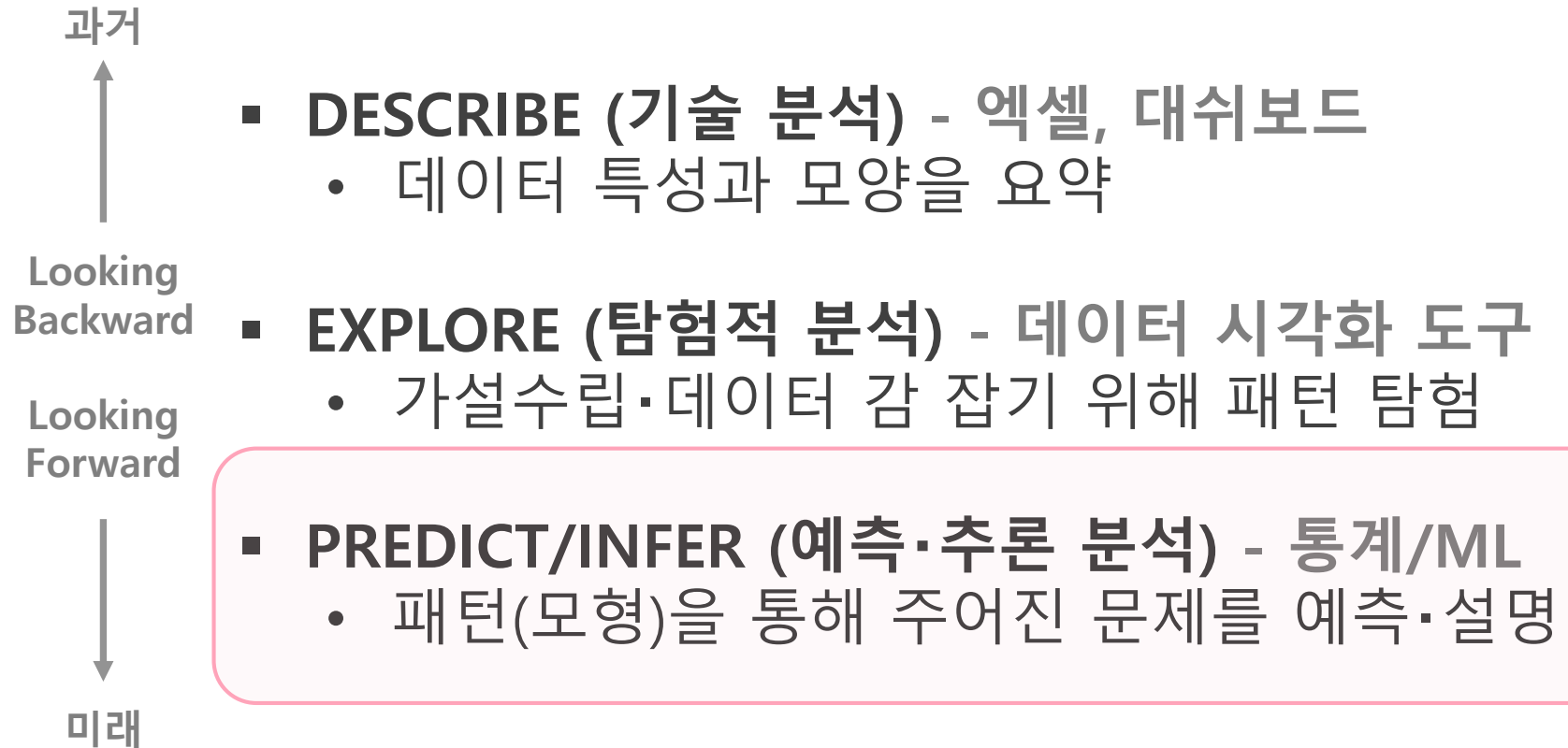
**결국 둘다 데이터를 통해 문제를 해결하는 데 사용됨  
기술과 방법의 차이라기 보다는 분석 목표의 차이**

# Module IV

## Linear Regression Analysis

아이디케이스퀘어드 양승준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)  
<https://www.heartcount.io>

## 데이터 분석 주요기술



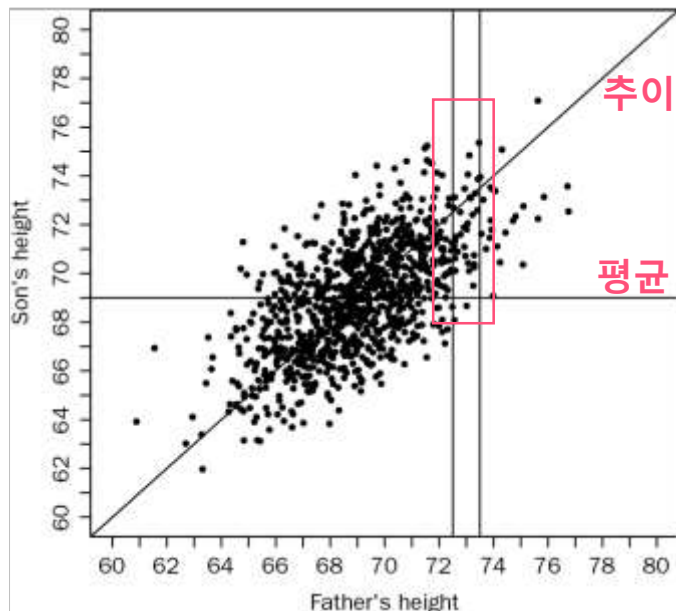
# Regression toward the Mean



## 우연(Chance)의 영향으로 평균(평범함)으로의 회귀

- 왕겨(Bran) 먹은 후 배변에 걸리는 시간(Oral-Anal Transit Time)이
  - 평균(48시간)보다 오래걸렸던 사람은 더 빨라졌고
  - 평균(48시간)보다 짧았던 사람은 더 길어졌고
  - 평균이었던 사람은 큰 변화가 없었다.

## REGRESSION *towards* MEDIOCRITY in HEREDITARY STATURE. By FRANCIS GALTON, F.R.S., &c.



NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.  
(All Female heights have been multiplied by 1.08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	4	5	..
72.5 ..	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72.2
71.5 ..	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5 ..	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5 ..	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5 ..	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	40	68.2
67.5 ..	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67.6
66.5 ..	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67.2
65.5 ..	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66.7
64.5 ..	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65.8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the readings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

# 선형회귀분석 (Linear Regression Analysis)

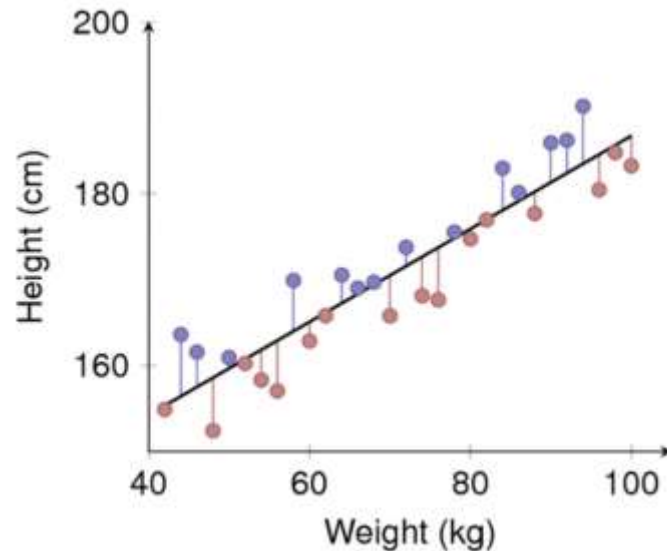
Supervised  
Machine-Learning

Regression Model: Y가 숫자형 변수(매출)인 경우

Classification Model: Y가 범주형 변수(성별)인 경우

## Linear Regression

- 가장 오래되고, 널리 쓰이고, 결과를 이해하기 쉬운 지도학습 알고리즘
- 독립변수(X)를 가지고 숫자형 종속변수(Y)를 가장 잘 설명·예측(**Best Fit**)하는 선형 관계(Linear Relationship)를 찾는 방법 중 하나
- X가 범주형 변수(성별)인 경우, 집단(남·녀) 간 Y값의 차이를 분석



## 계산방법 (Least Squares)

X와 Y 사이에 선형적 관계가 있다는 가정 하에 실제 Y값과 예측한 Y값의 차이를 최소화하는 방정식을 계산

$$Y = b_0 + b_1X + \text{error}$$

- $b_0$ : Y축 절편(Intercept); 예측변수가 0일 때 기대 점수를 나타냄
- $b_1$ : 기울기로 X가 한 단위 증가했을 때의 Y의 평균적 변화값을 나타냄

\*참고: <http://students.brown.edu/seeing-theory/regression/index.html>

# 선형회귀분석 (Linear Regression Analysis)

## P-Value (Probability-Values)

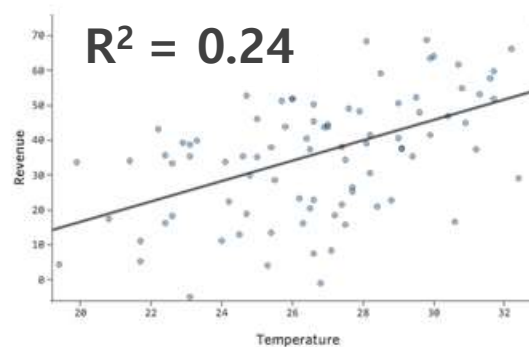
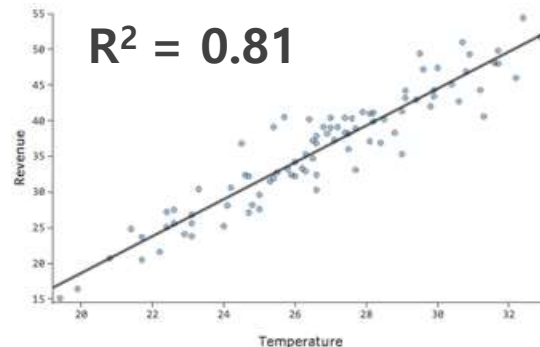
Q. X와 Y 사이에 통계적으로 유의미한 관계가 있나?

- Statistical Significance (통계적 유의성)
- 데이터를 통해 확인한 관계가 우연히 나왔을 확률
- P값이 0.03: 데이터에서 발견한 관계가 운일 확률 3%
- 관계의 세기(Size of an Effect)를 나타내는 것은 아님

## R<sup>2</sup> (R-SQUARED; 결정계수)

Q. X가 Y를 얼마나 잘 설명/예측하는가?

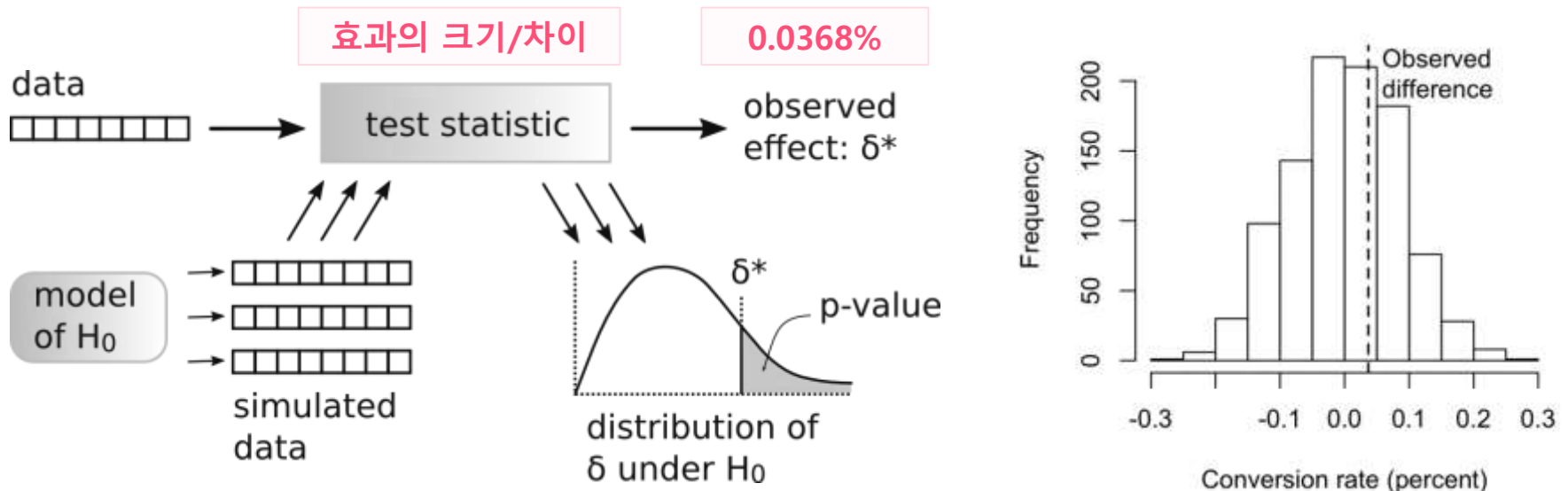
- Goodness of Fit: X로 설명할 수 있는 Y 변화량의 크기



# Statistical Significance and P-Value(Probability-Values)

Outcome	Campaign A	Campaign B
Conversion	200	182
No Conversion	23539	22406
Conversion Rate	0.8425%	0.8057%

1. 캠페인 A의 전환률이 0.0368% 높음; 차이가 의미가 있나?
2. 통계적 유의성 검증이 꼭 필요한가? (이 정도면 작은 샘플을 사용하여 일반화하는 일을 걱정할 필요없는 빅데이터 아닌가?)
3. 두 캠페인 사이의 전환율 차이가 우연은 아닌가?
  - 둘 간 전환율 차이가 없다고 가정( $H_0$ )하고 두 캠페인 결과를 하나로 섞는다
  - 섞은 데이터에서 23739, 22588개를 Resampling하여 두 캠페인 간 전환율 차이를 기록 (1,000번 반복)
  - 차이(test statistic; 검정통계량)가 > 0.0368%보다 큰 경우의 확률을 계산: 30.8% (P-Value: 0.308)

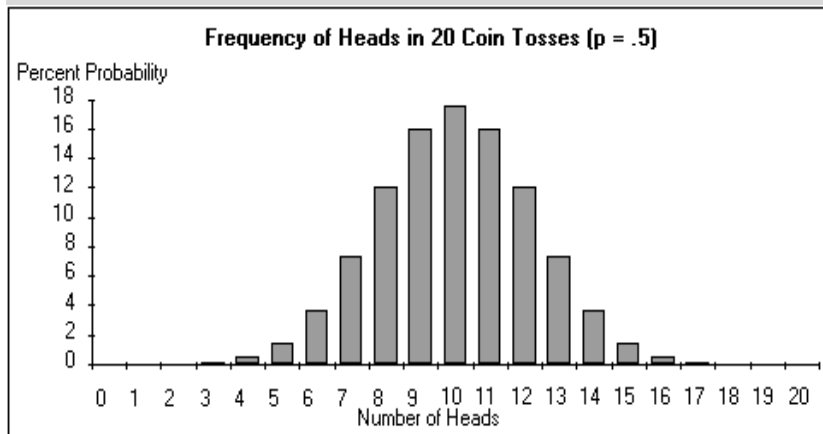




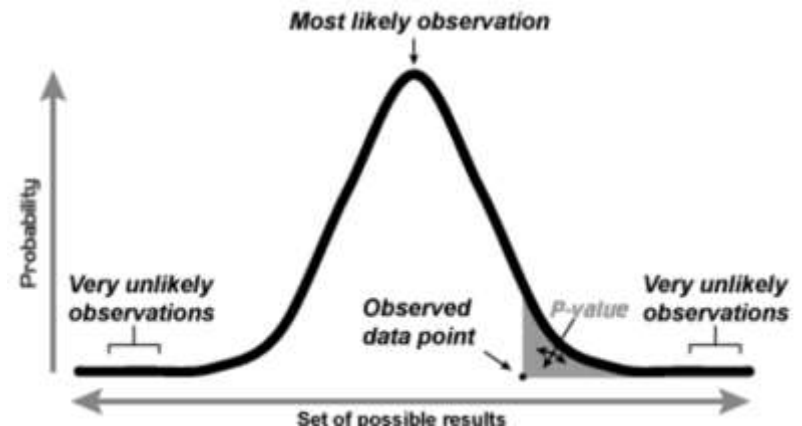
**P값: (선형적) 관계가 우연히 나왔을 확률**  
**작은 P값: 데이터에서 발견한 관계가 우연이 아니다 (=통계적으로 유의미)**

- **귀무 가설( $H_0$ ; Null Hypothesis):** X와 Y 사이에 관계가 없다고 가정  
 **$P = 0.03$ :** 관계가 없단 가정 하에 데이터에서 발견한 관계 혹은 더 극단적인 관계가 관측될 확률 = 3%

정상적인(앞뒷면 확률이 동일한) 동전을 20번 던져 앞면이 나온 횟수에 대한 확률 분포



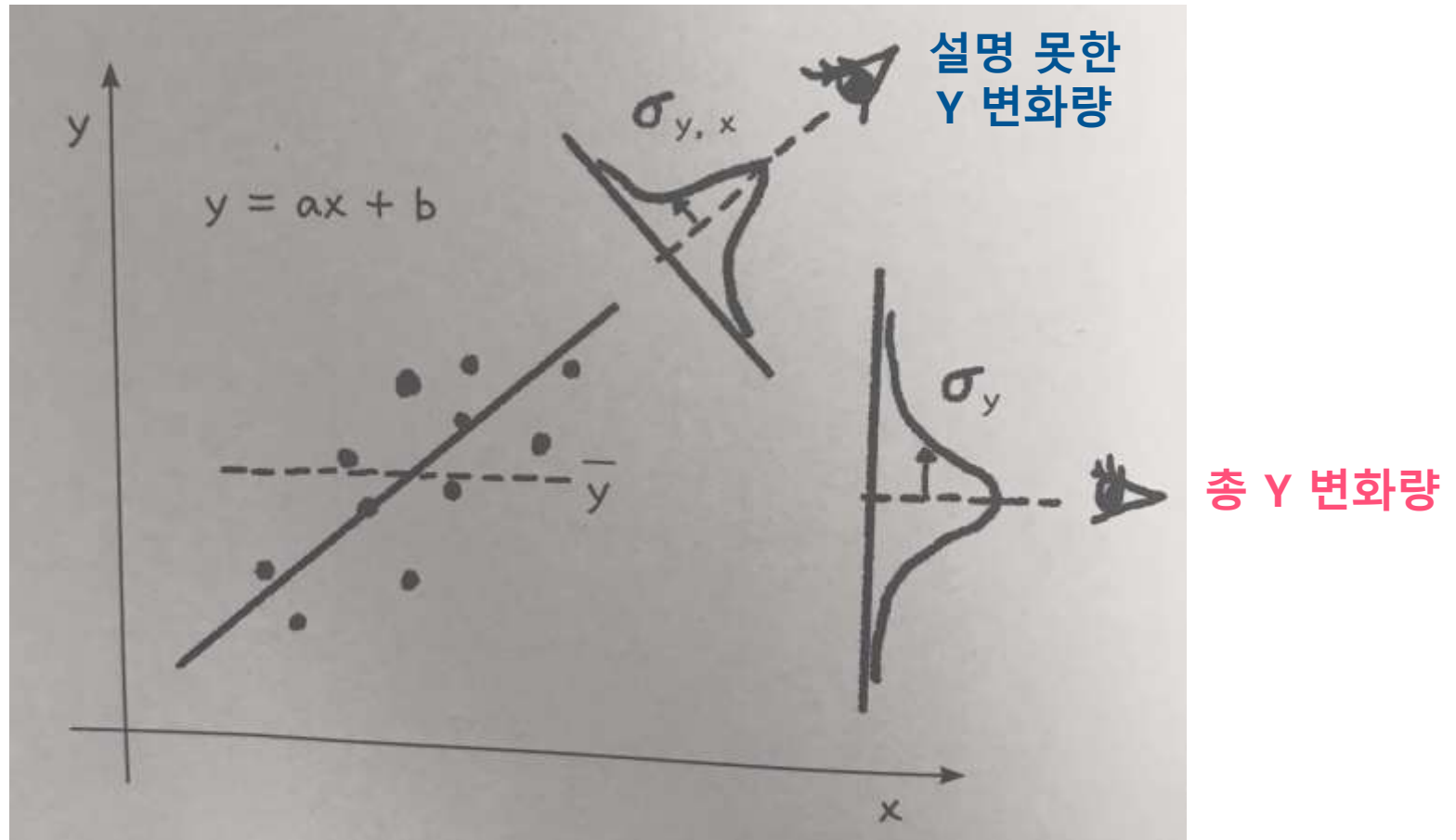
동전이 정상이라면 앞면이 연속 18번 이상 나올 확률이 < 5%. 동전 정상이 아니라고 결론





# 선형회귀분석 (Linear Regression Analysis)

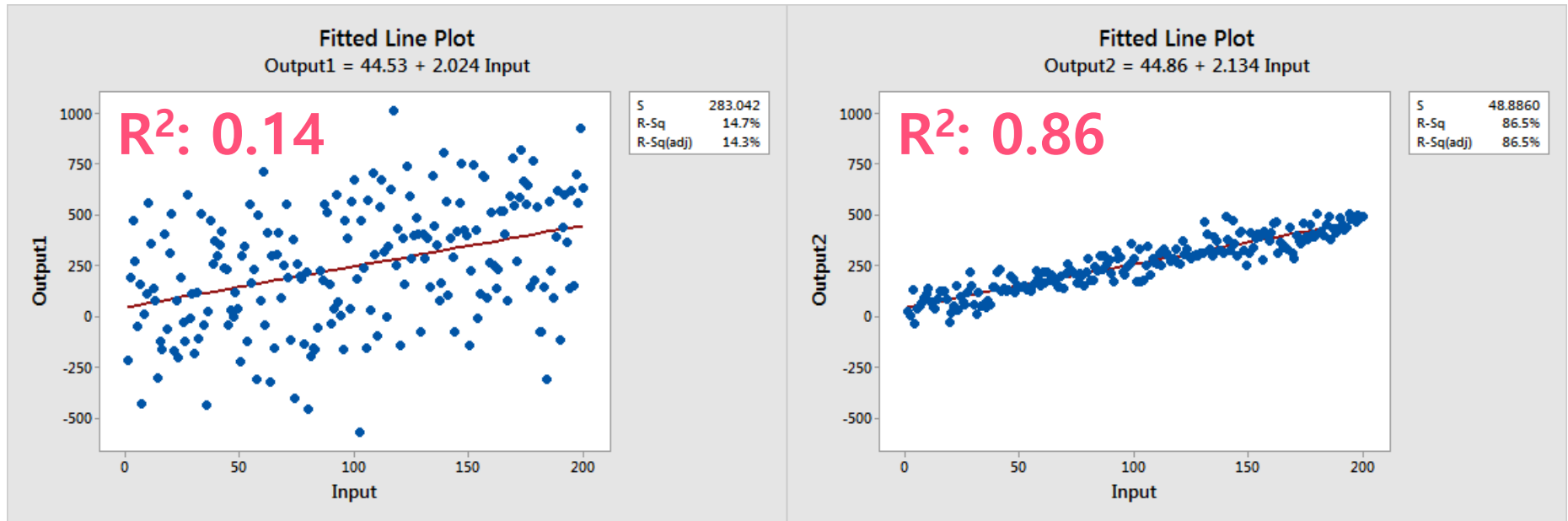
$$\begin{aligned} R^2 &= \text{설명한 Y 변화량} / \text{총 Y 변화량} \\ &= (\text{총 Y 변화량} - \text{설명 못 한 Y 변화량}) / \text{총 Y 변화량} \\ &= 1 - (\text{설명 못 한 Y 변화량} / \text{총 Y 변화량}) \end{aligned}$$



# 선형회귀분석: 결정계수( $R^2$ : R-SQUARED)

낮은 결정계수가 반드시 나쁜 (Inherently Bad) 것은 아님

- 동일한 회귀방정식:  $Y = 44 + 2 \cdot X$ ;  $P < 0.001$
- 우측 모형이 좌측 모형보다 예측 정확도( $R^2$ )는 매우 높음
- 변수 간 경향성은 동일: X 1단위 증가  $\rightarrow$  Y 2단위 증가



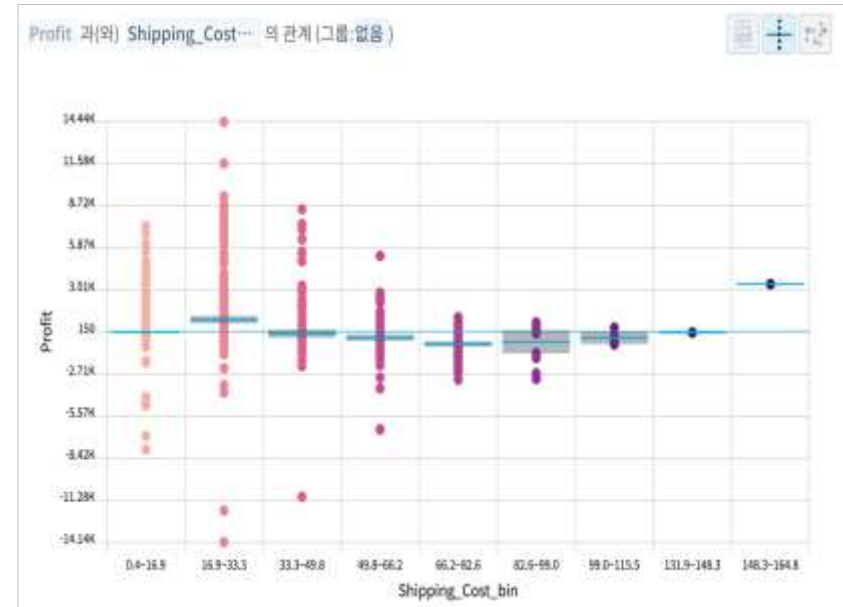
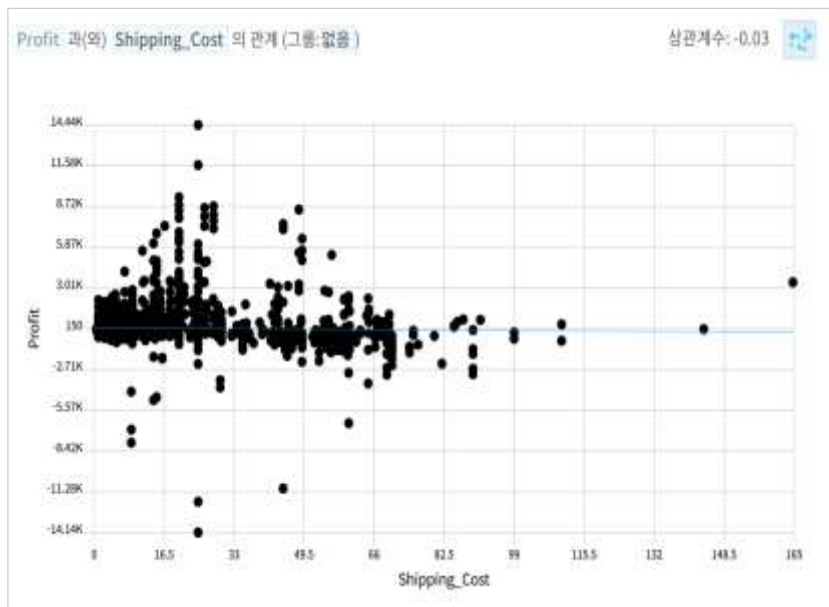
# Binning: 선형회귀분석으로 비선형적 관계 찾기



## Binning: 숫자형 변수를 범주형 변수로 변형

- 이익(Y)과 운송비용(X; 숫자)간에는 선형적 관계 없음
- 운송비용을 범주형 변수로 변형하면 비선형적 관계 발견
- 해석, 서로 다른 운송비용 구간별로 이익의 차이가 존재

No.	변수명	R <sup>2</sup> ⓘ	Adjusted R <sup>2</sup> ⓘ	P-Value ⓘ	레코드 갯수 ⓘ
3	Shipping_Cost_bin	0.078	0.075	0.00000 (< 0.001 ***)	3,319
4	Product_Sub_Category	0.064	0.059	0.00000 (< 0.001 ***)	3,319



# Simple Linear Regression Analysis – Advertisement

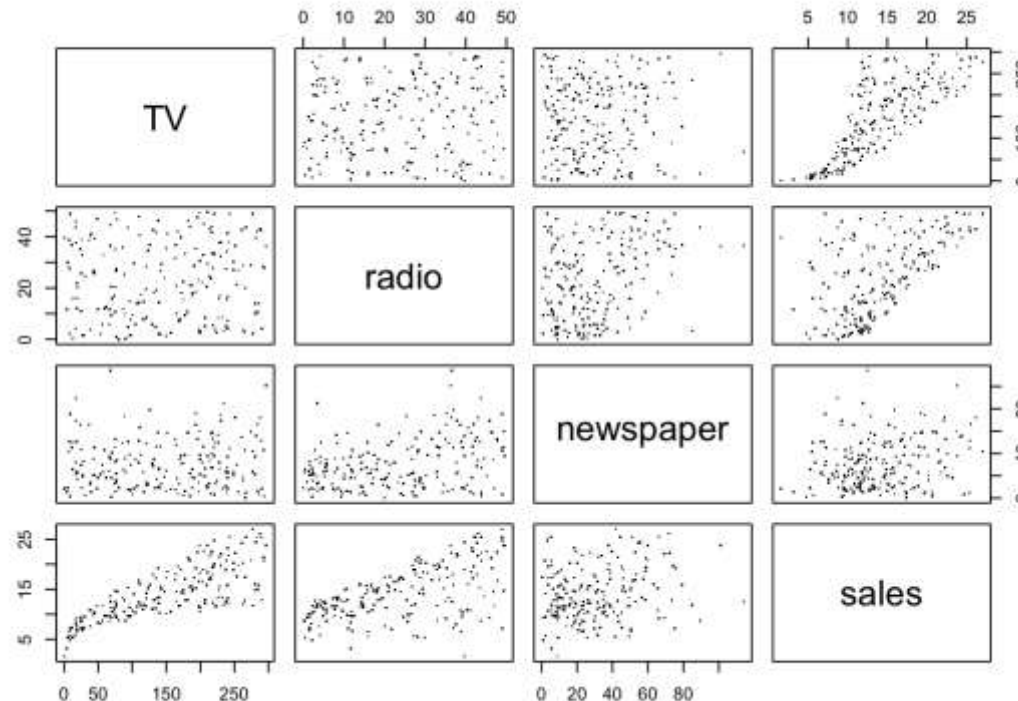


매출에 미치는 매체 영향에 상호작용이 없다고 가정하였을 때

1. TV, Radio, 신문 중 Sales를 가장 정확하게 예측하는 매체는?
2. TV, Radio, 신문 중 Sales 증가에 가장 큰 효과가 있는 매체는?



Advertising.xlsx

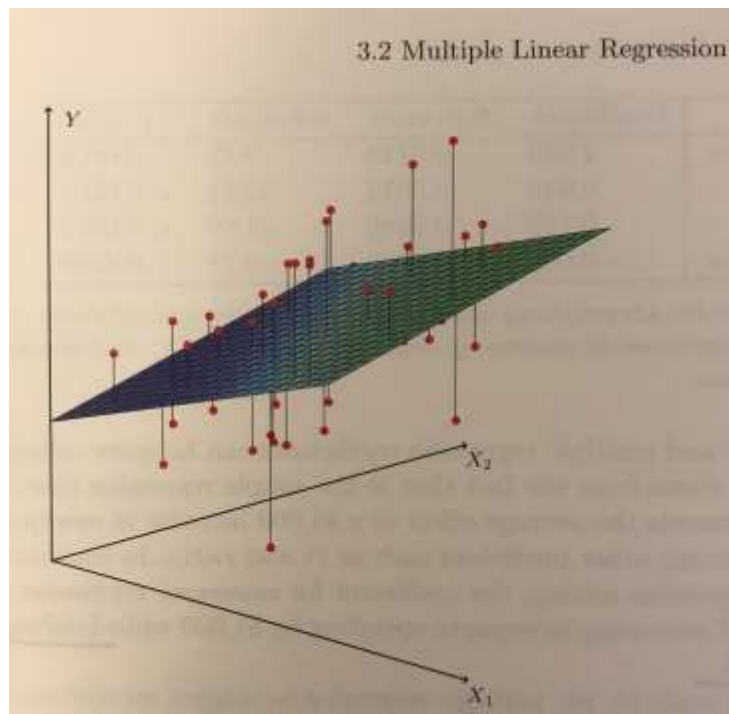


# Multiple Linear Regression Analysis – Advertisement



이번에는 변수 2개[TV, Radio]를 사용하여 Sales와의 관계를 설명·예측하는 회귀모형을 만들어 봅시다.

$$Y = b_0 + b_1X_1 + b_2X_2$$
$$\text{Sales} = 2.9 + 0.045 \times \text{TV} + 0.187 \times \text{Radio}$$



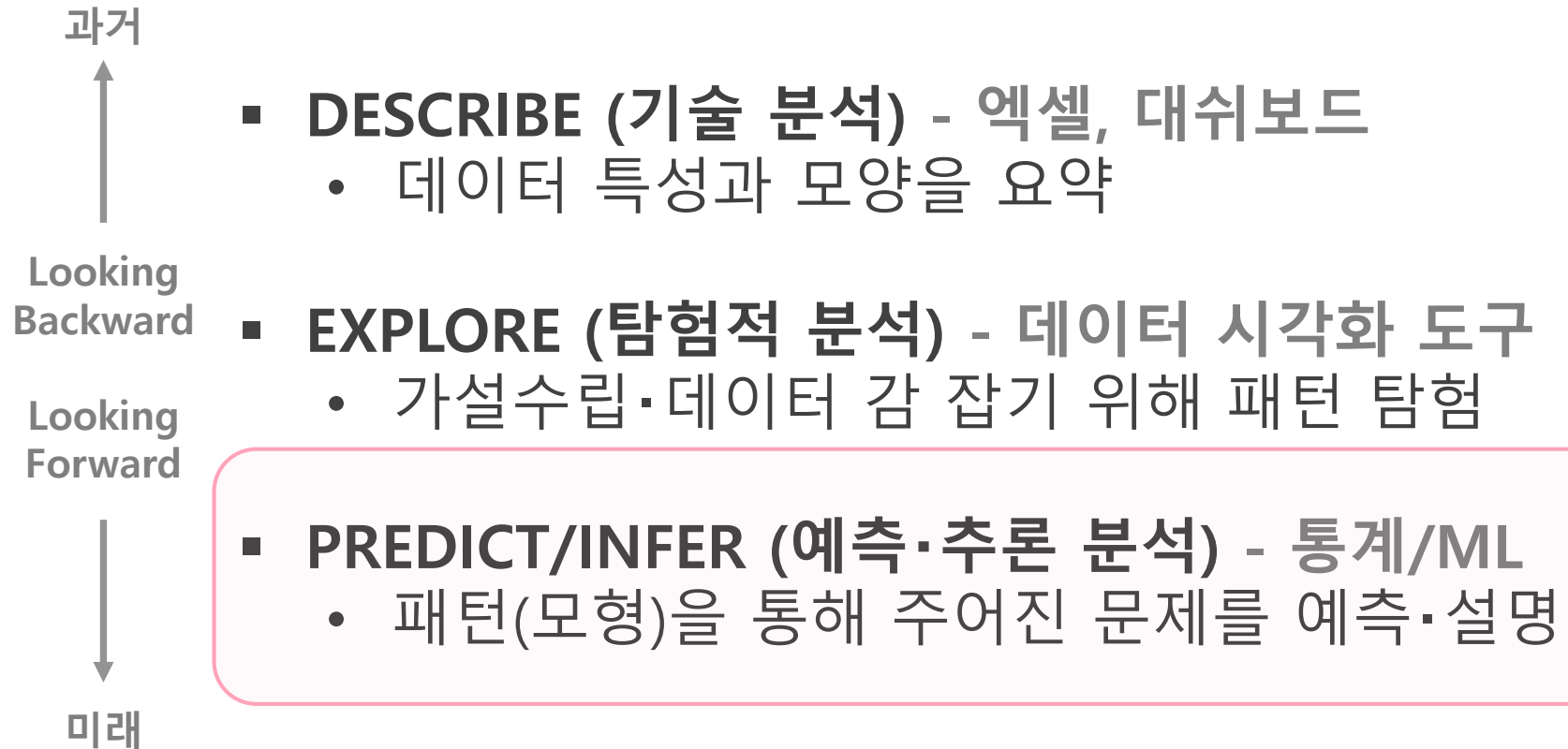
SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.94720339			
R Square	0.897194261			
Adjusted R Square	0.896150548			
Standard Error	1.681360913			
Observations	200			
ANOVA				
	df	SS	MS	F
Regression	2	4860.2348	2430.1174	859.6177183
Residual	197	556.91398	2.8269745	
Total	199	5417.1488		
	Coefficients	Standard Error	t Stat	P-value
Intercept	2.921099912	0.2944897	9.9191929	4.56556E-19
X Variable 1	0.045754815	0.0013904	32.908708	5.43698E-82
X Variable 2	0.187994227	0.00804	23.382446	9.77697E-59

# Module V

## Decision Tree Algorithm

아이디케이스퀘어드 양승준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)  
<https://www.heartcount.io>

## 데이터 분석 주요기술



# Decision Tree and Classification

**Supervised  
Machine-Learning**

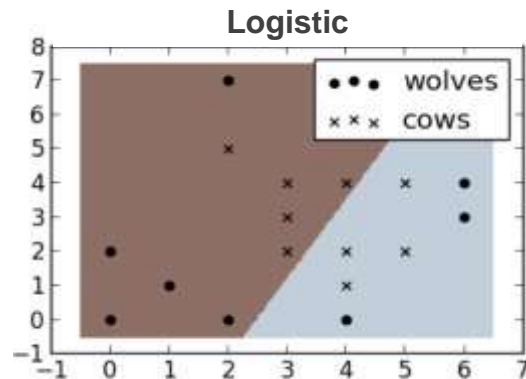
**Regression Model:** Y가 숫자형 변수(매출)인 경우

**Classification Model:** Y가 범주형 변수(성별)인 경우

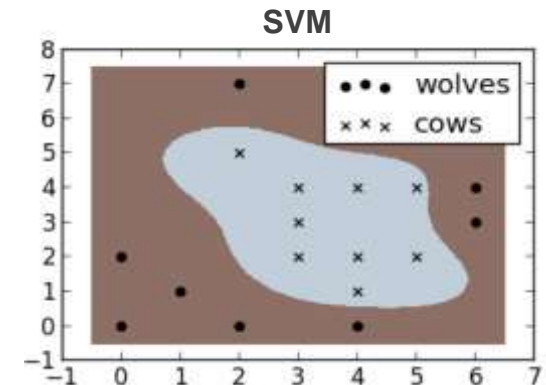
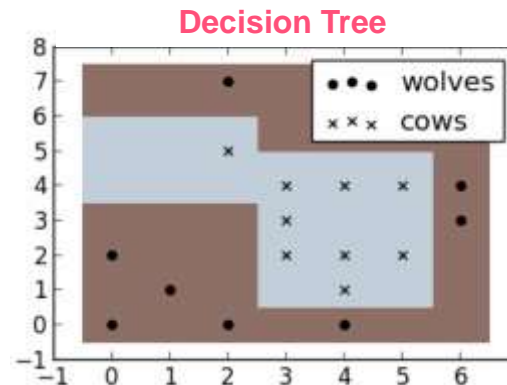
- **Decision Tree:** 의사결정트리; 대표적 Classification Model
- **Classification:** 서로 다른 집단을 구분하는 규칙(경계) 찾기

## Data-Driven Farmer

**Linear Classifier**



**Non-Linear Classifier**

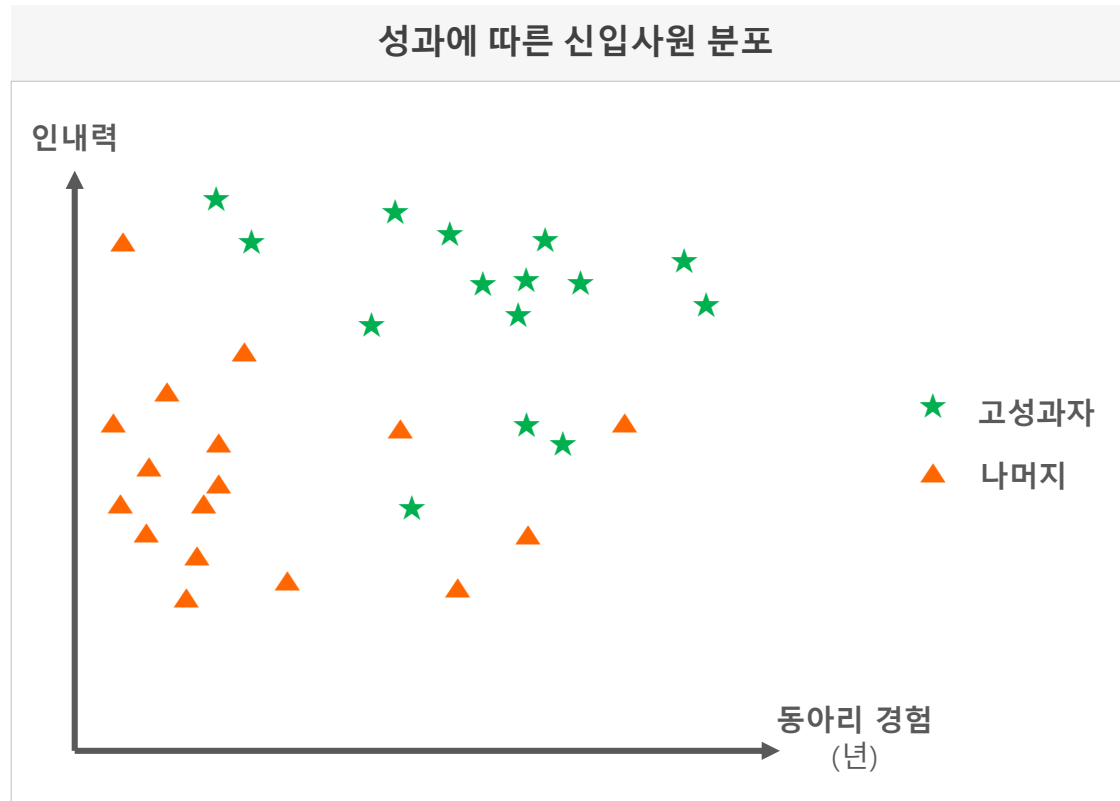




# Decision Tree - Minimizing Entropy

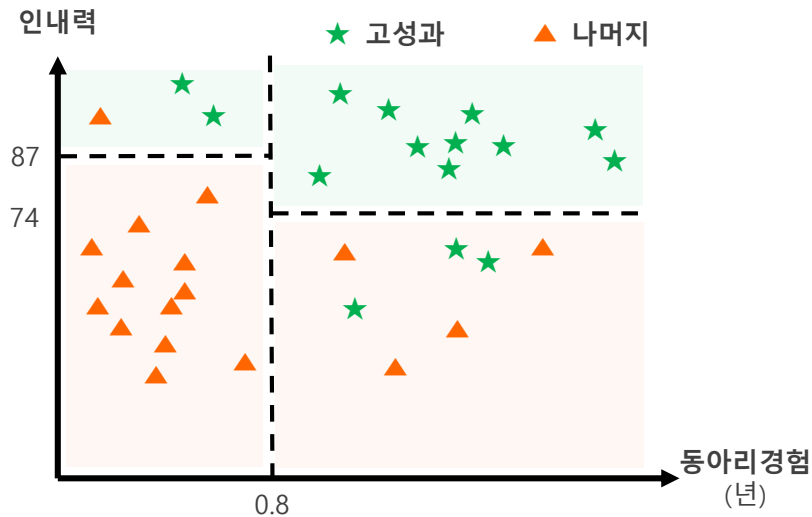
## 엔트로피(Entropy): Measure of Impurity

- Purity: 엔트로피를 최소화하도록(= 끼리끼리 모이도록) 공간을 구획하는 문제
- Homogeneity: 특정 집단이 밀집한 세그먼트의 논리적 규칙을 찾는 문제

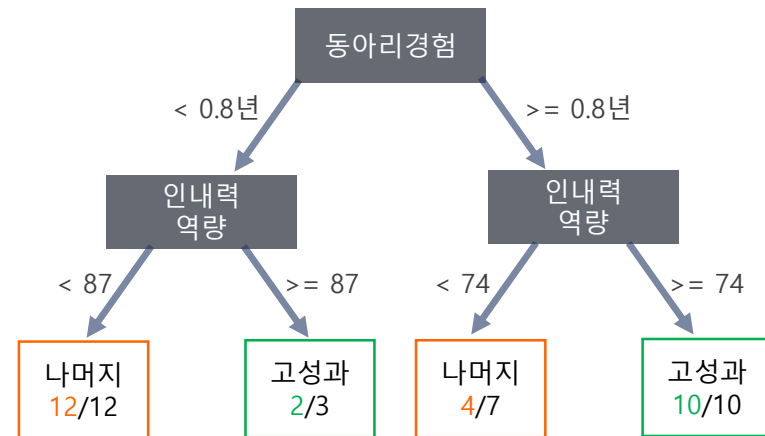


# Decision Tree - Minimizing Entropy

성과에 따른  
신입사원 분포



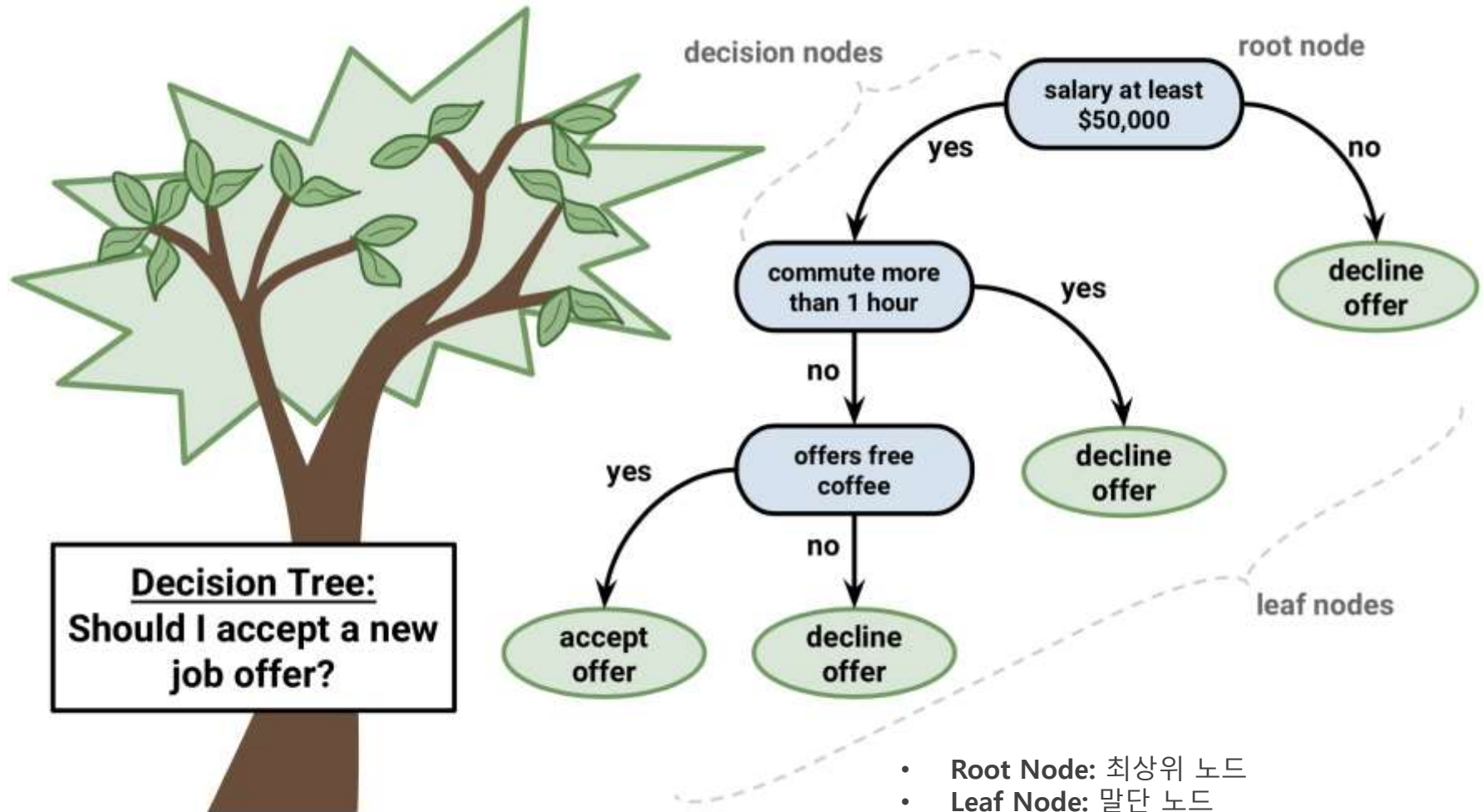
고성과 신입사원  
분류모형 (의사결정트리)



## 분류규칙

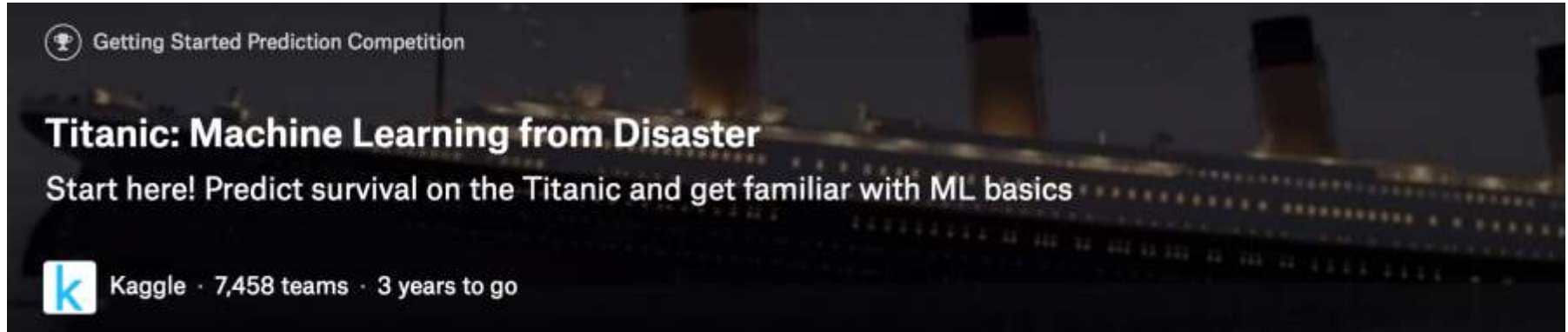
규칙 (Rule Set)	확률 (Probability)
IF (동아리경험 < 0.8년) and (인내력역량 < 87) then Class = 저성과 신입사원	100% (12/12)
IF (동아리경험 < 0.8년) and (인내력역량 >= 87) then Class = 고성과 신입사원	67% (2/3)
IF (동아리경험 >= 0.8년) and (인내력역량 < 74) then Class = 저성과 신입사원	57% (4/7)
IF (동아리경험 >= 0.8년) and (인내력역량 >= 74) then Class = 고성과 신입사원	100% (10/10)

# Decision Tree – Terminology



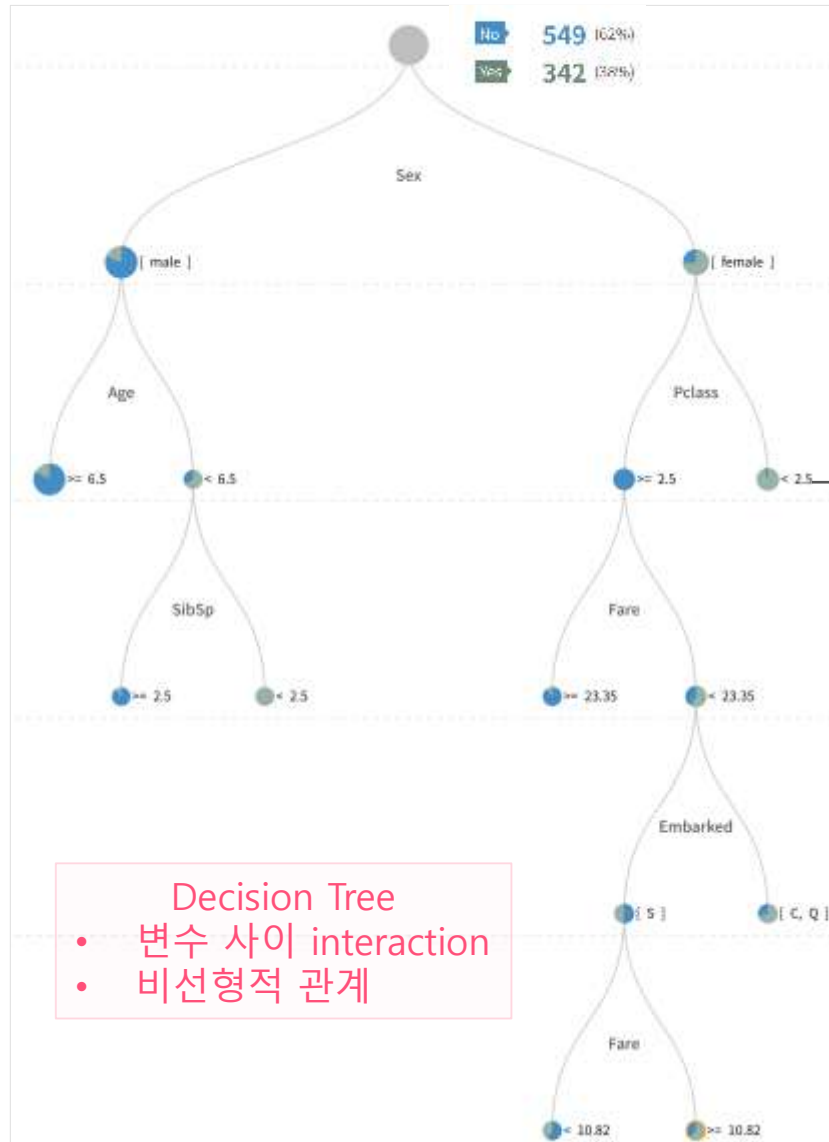
- **Root Node:** 최상위 노드
- **Leaf Node:** 말단 노드
- **Decision Node:** 의사결정 노드
- **Splitting:** 동질적 집단으로 쪼개는 일
- **Pruning:** 트리가 너무 길어지지 않게 하는 일

# Decision Tree – Titanic Dataset

A banner for the Titanic dataset competition. It features a dark background with a night view of the Titanic ship. The text is white and yellow. It includes a 'Getting Started Prediction Competition' icon, the title 'Titanic: Machine Learning from Disaster', a subtitle 'Start here! Predict survival on the Titanic and get familiar with ML basics', and the Kaggle logo with text 'Kaggle · 7,458 teams · 3 years to go'.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)
parch	# of parents / children aboard the Titanic	Parent = mother, father Child = daughter, son, stepdaughter, stepson
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

# Decision Tree – Titanic Dataset



confusion matrix	Yes (Predicted)	No (Predicted)
Yes (Actual)	227 True Positive	115 False Negative
No (Actual)	28 False Positive	521 True Negative

$$\text{Precision(Y)} = \frac{227}{227 + 28} = 89\%$$

$$\text{Recall(Y)} = \frac{227}{227 + 115} = 66\%$$

# Decision Tree – Titanic Dataset

## Confusion Matrix: 분류 모형의 성능을 평가하는 방법

이해하기는 쉬운데 용어가 어려움


- **True Positive:** 맞는 걸 맞다고 하는 것
- **True Negative:** 아닌 걸 아니라고 하는 것
- **False Positive** (I형 오류) : 아닌데 맞다고 하는 것 (거짓을 믿는 것)
- **False Negative** (II형 오류) : 긴데 아니라고 하는 것 (참을 거부하는 것)

Confusion Matrix	Yes (Predicted)	No (Predicted)
Yes (Actual)	227 True Positive	115 False Negative TYPE II Error
No (Actual)	28 False Positive TYPE I Error	521 True Negative

Recall(Y)  
재현률  
66%

$$= \frac{227}{227 + 115}$$

66.4%  
(227 / 342)



Precision(Y)  
정밀도 =  $\frac{227}{227 + 28}$   
89%

Accuracy  
모형 정확도 =  $\frac{227 + 521}{227 + 115 + 28 + 521}$   
84%

## My Two Cents

아이디케이스퀘어드 양승준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)  
<https://www.heartcount.io>

### 할리우드

(캐릭터 + 욕망) / 방해물 = 이야기

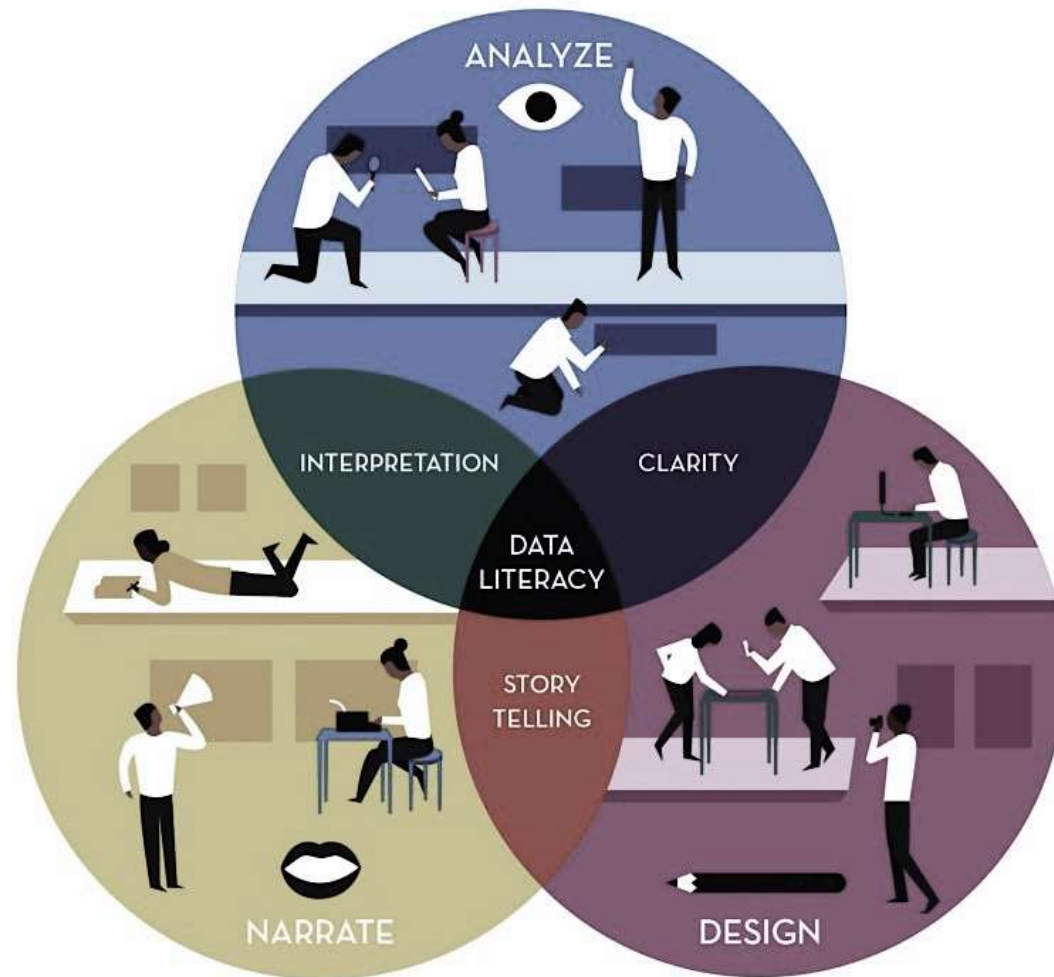
### 데이터 분석

(Data + Data Literacy) / 방해물 = 유용한 패턴



# Data Literacy

데이터에서 발견한 **유용한 사실**을 **공감**할 수 있는 형식으로 표현, **이롭게 활용**되도록 하는 것

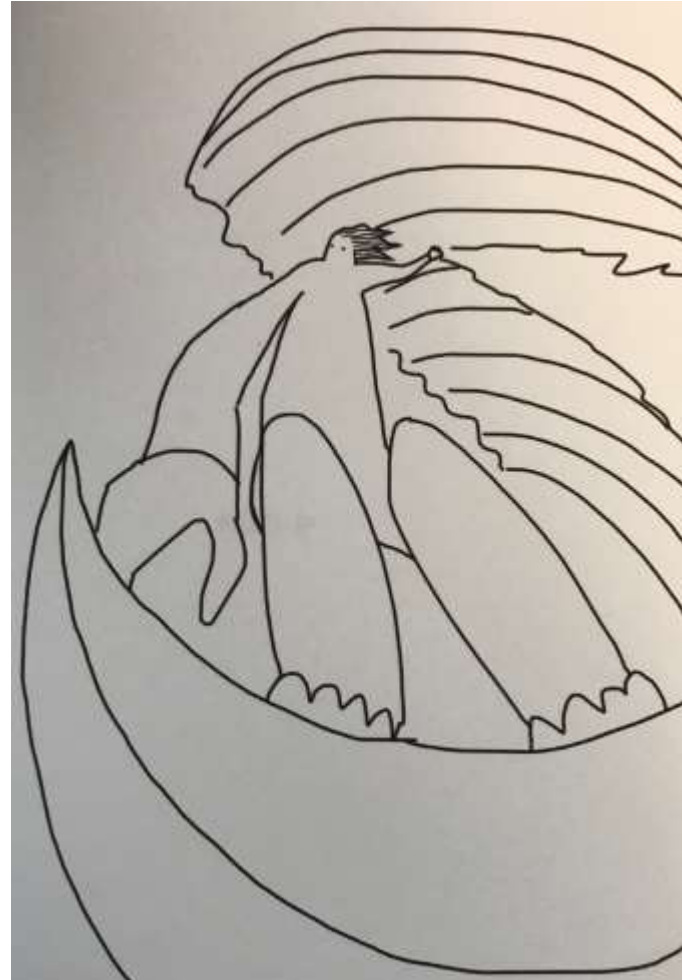


# Data Literacy: from Data Phobia to Data Fluency

Data Phobia



Data Fluency



# The Size of Data is Overrated

개별 레코드에 담긴 패턴(효과/시그널)이 클수록  
패턴 발견을 위해 적은 데이터가 필요

**Bigger Data**  
[Liability]



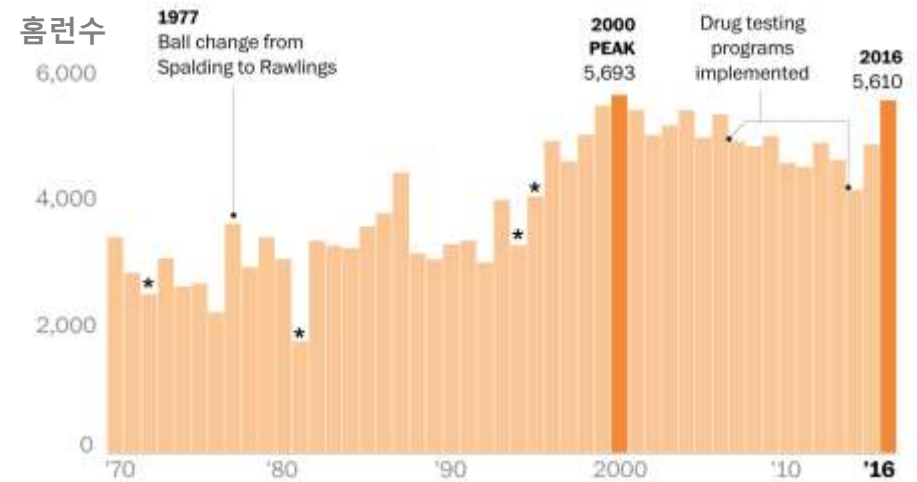
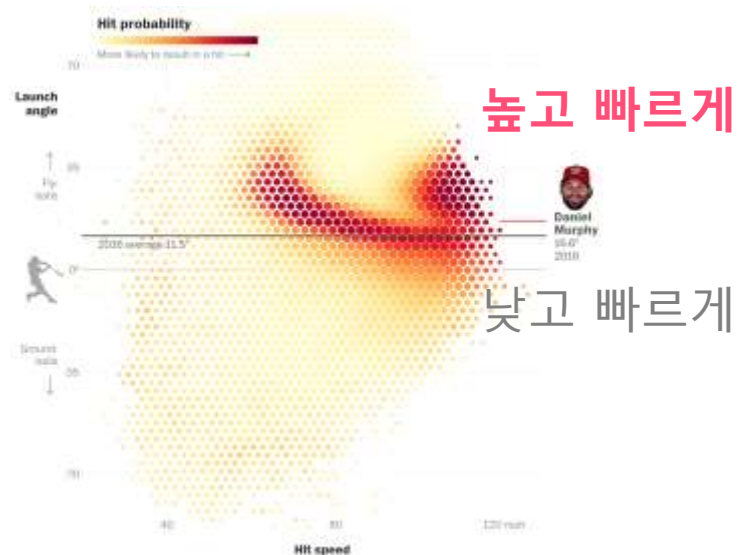
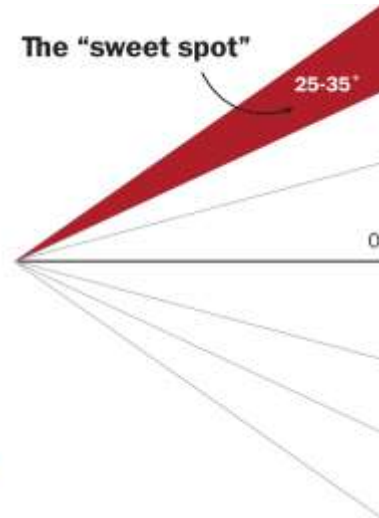
**Smaller Data**  
[Asset; 독점]



# Finding the Right Angle: 땅볼(Ground Ball) vs. 뜀볼(Fly Ball)



Baseball.Dataset.xlsx



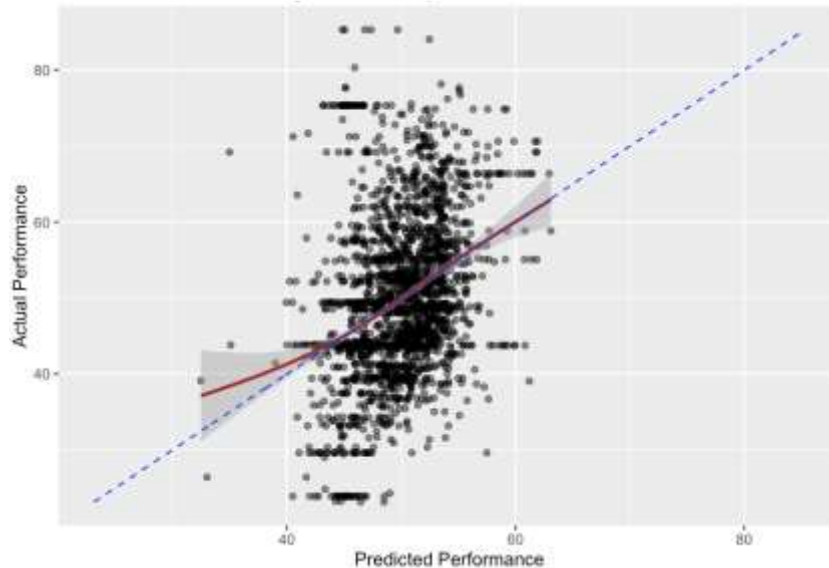


## Model Transparency

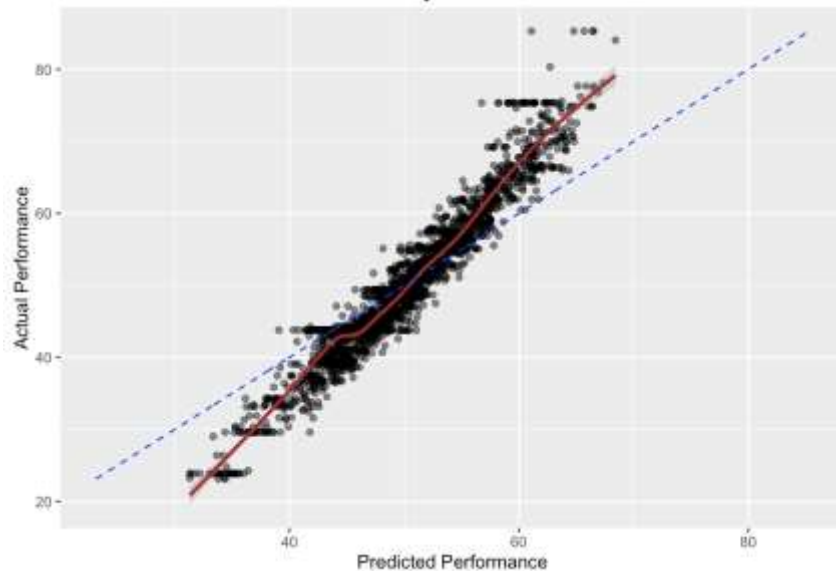
기계의 예측을 맹목적으로 따를 것이 아니라 현실에 직접 개입하려 한다면 모형의 투명성(설명력)이 예측력보다 중요

**\*Right to Explanation (EU GDPR)**

White Box 모형 - Linear Regression  
낮은 예측정확도, 높은 설명력



Black Box 모형 - Random Forest  
높은 예측정확도, 낮은 설명력





당신 회사에 다음과 같이 두가지 모델의 차량이 있다.  
모든 자동차는 일년에 10,000 마일을 주행한다.

**Fleet A**  
10 MPG SUV



**Fleet B**  
20 MPG Sedan

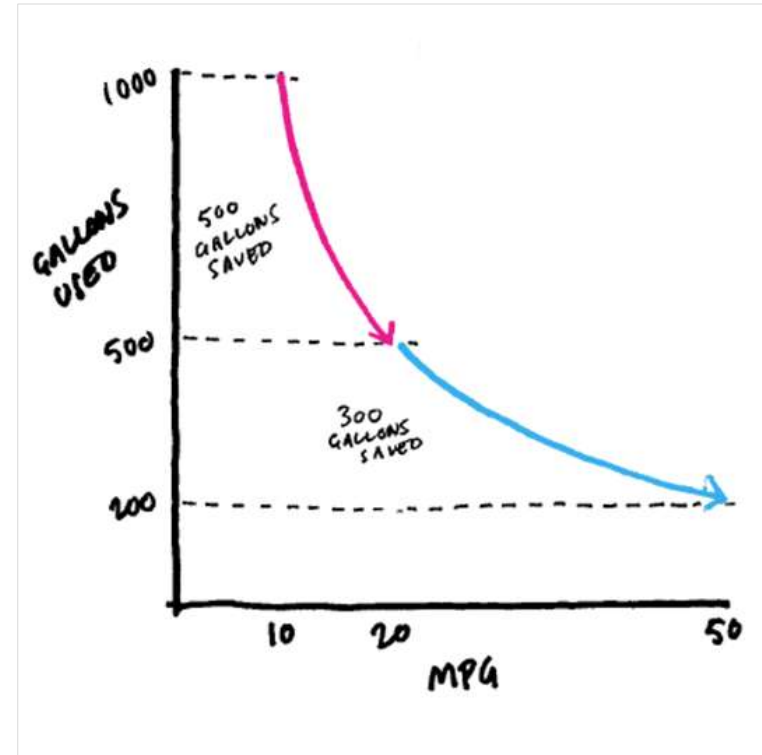
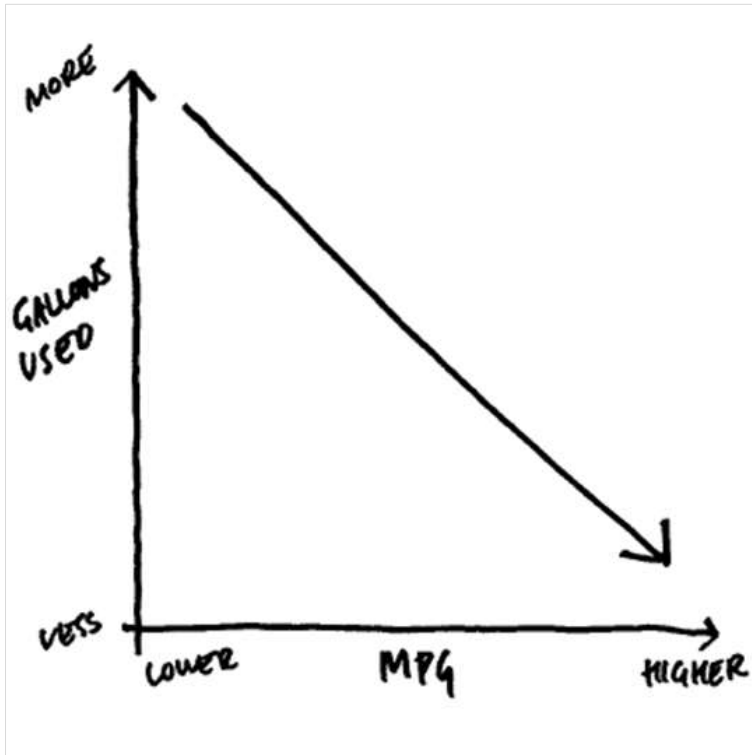


---

**Q. OPEX(기름값)을 줄이기 위한 당신의 선택은?**

- A. 10 MPG 차량을 20 MPG 차량으로 교체
- B. 20 MPG 차량을 50 MPG 차량으로 교체

# Linear Thinking vs. Non-Linear Relationship (source: HBR 2017.05월호)



10,000 마일당 사용한 기름 (Gallons)

	현재	차량 업그레이드 후	절감분
A.	1,000 (@10 MPG)	500 (@20 MPG)	500
B.	500 (@20 MPG)	200 (@50 MPG)	300



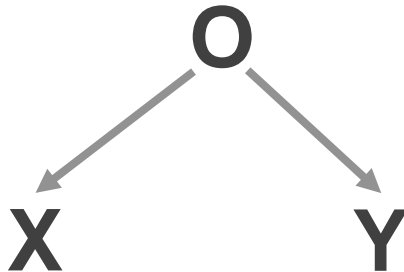
# Correlation vs. Causation



바람이 많이 불면  
나무통 가게가 돈을 번다.

## Two Types of Causality Problems

### Omitted Variable Bias



- 교육 받아서 성과가 좋아 졌나?
- (경기 탓으로) 성과가 안 좋아서 교육 받았나?

### Reverse Causality



- 한 업무에 오래 있어서 성과가 낮나?
- 성과가 낮아서 한자리에 오래 있었나?



## Correlation을 언제 의사결정에 활용해야 하나?

데이터를 통해 [소고기와 우유]를 구매한 고객들의 자동차 사고 발생률이 [라면과 소주]를 구매한 고객보다 높은 걸 확인했다. 보험회사가 취할 행동은?

가. 구매패턴에 따른  
보험료 차등 적용

나. 저위험군 고객들  
타겟 마케팅



## Correlation을 언제 의사결정에 활용해야 하나?

데이터를 통해 [소고기와 우유]를 구매한 고객들의 자동차 사고 발생률이 [라면과 소주]를 구매한 고객보다 높은 걸 확인했다. 보험회사가 취할 행동은?

가. 구매패턴에 따른  
보험료 차등 적용

나. 저위험군 고객들  
타겟 마케팅

관계에 대한  
높은 확신  
(인과관계)

가. 구매패턴에 따른  
보험료 차등 적용

**Don't Act**

**Act**

나. 저위험군 고객들  
타겟 마케팅

관계에 대한  
낮은 확신

실 > 득

실 < 득

# Why Model Transparency Matters!

## 결핵 치사율 예측모형

결핵환자 중 천식을 앓고 있는 사람은 집으로 돌려 보내세요.



Carnegie Mellon University  
Research Showcase @ CMU

Department of Philosophy

Dietrich College of Humanities and Social Sciences

1997

An Evaluation of Machine-Learning Methods for  
Predicting Pneumonia Mortality

투명한 모형과 사람의 판단이  
환자를 살렸습니다.

# Insight Matrix

## X(입력변수)와 Y(목표변수) 간의 관계

○ 기존 믿음(가설)을 정량적으로 검증

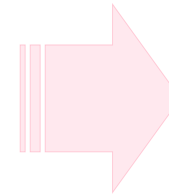
● 몰랐던 사실 발견 (Unknown Unknown)

△ 관계 없음

▲ 관계 있는 줄 알았는데 없음 (Myth Busting)

example) 임직원 성과

X (Input Variables; Features)		
Hiring	성격	▲
	학력/인지능력	○
	채용 경로	●
Culture / Management	리더쉽	○
	보상	▲
	육성	○
Behavior	근태	△
	협업	○
	만족도	▲



Y: 성과

## Change: Targeting vs. Optimization

### Targeting

특정 대상 선정 후 개입

### Optimization

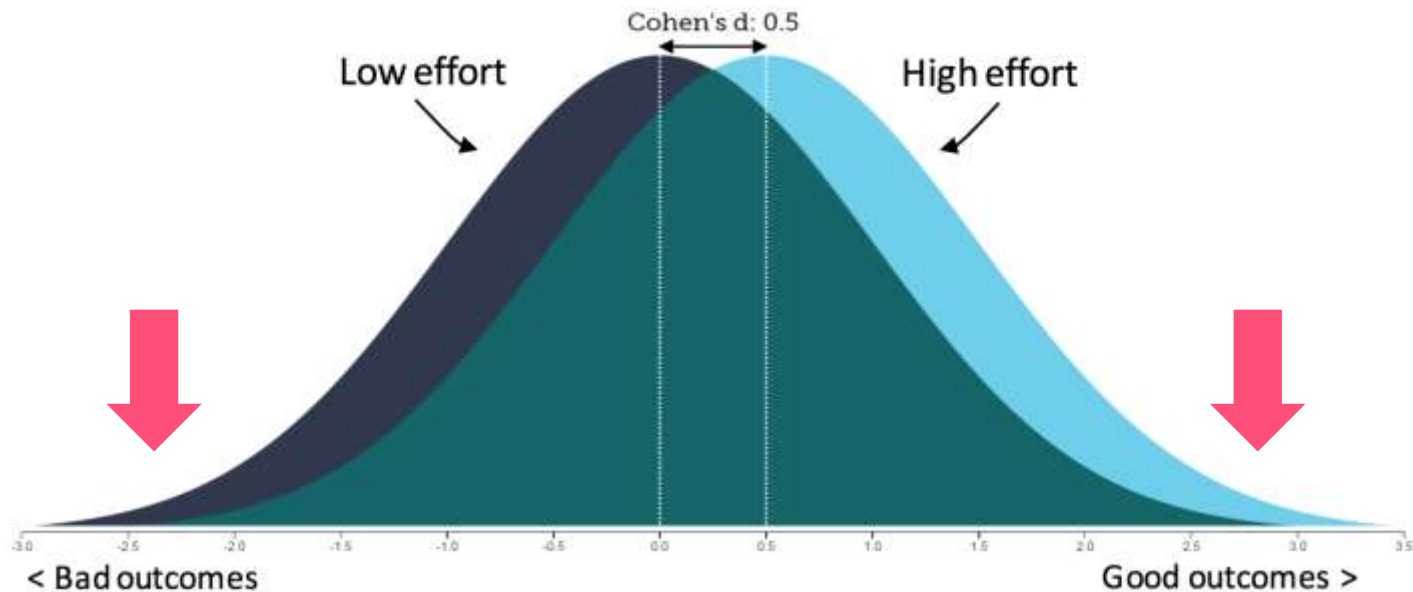
X를 바꾸어서 Y를 개선

Y(퇴사?)	X1	부서	지역	X4	회의 참여	X5	.....
No		재무	서울		32		
No		R&D	부산		14		
No		총무	서울		9		
No		인사	부산		26		
No		R&D	서울		43		
Yes		마케팅	서울		78		
Yes		인사	일본		63		
Yes		인사	일본		51		
Yes		인사	일본		103		

# Noise와 Signal을 어떻게 구분할까?

기량의 역설 (Paradox of Skill)  
기량 ↗ 기량의 변량 ↘ 운(Chance) ↗

## Compare The Extremes



# Data → Insight → **Belief**

① 믿음은 어디에서 오는가?    ② 새로운 사실은 믿음을 바꾸는가?



① Evolution      Experience      Culture      Data



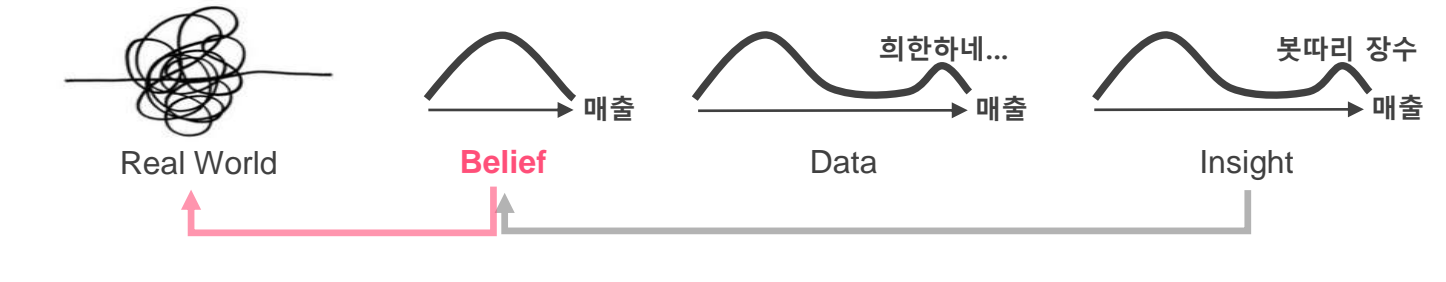
**Belief (Knowledge)**

②

- 우리의 믿음은 문장으로 구성되어 있음
  - 믿음: 적당한 운동은 건강에 좋다.
- 설명: 느끼고 보고 배운 걸로 믿음을 설명
  - 설명: 수영을 하니 몸이 깨운하다.
- 믿음이 바뀌려면 패턴이 쉽고 명확해야 함

# Data → Insight → Belief → Change

## 믿음을 바꿔 세상을 바꾸기



[1990 베트남, 빈곤 아동들의 영양실조 문제]  
"6개월 안에 변화를 만들지 못하면 떠나시오!"

### 구조적 문제들

위생설비, 깨끗한 물, 무지함

**TBU: True But Useless**

### 당장 바꿀 수 있는 것을



작고 부드러운 목소리로