HEARTCOUNT

KOREA

LOGIN

Module I
Data Literacy

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

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

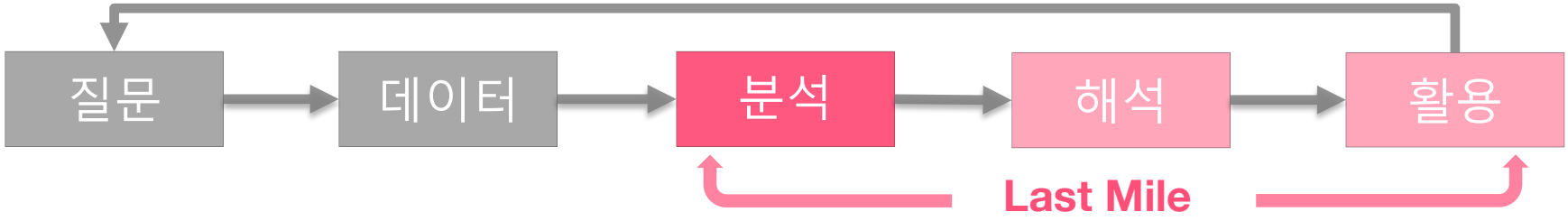
들소보다 물고기를 잡는 게 1.7배 더 생산적

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

Data Literacy: Fixing Last Mile Problem

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

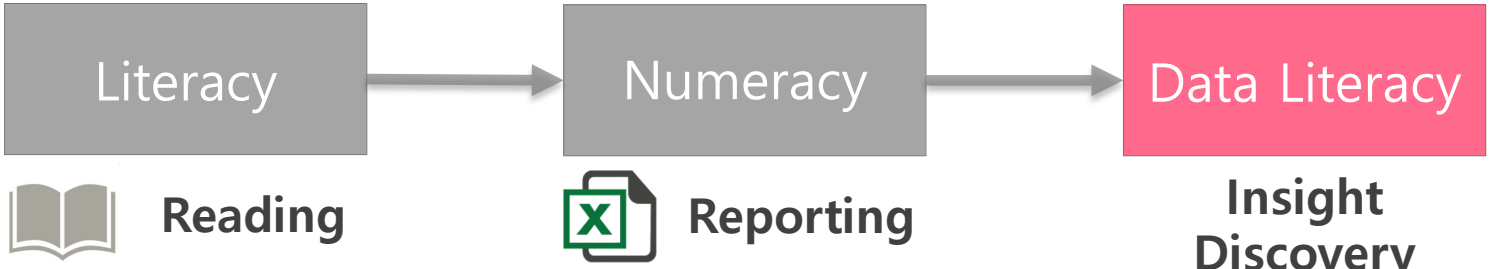
해결책

- 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

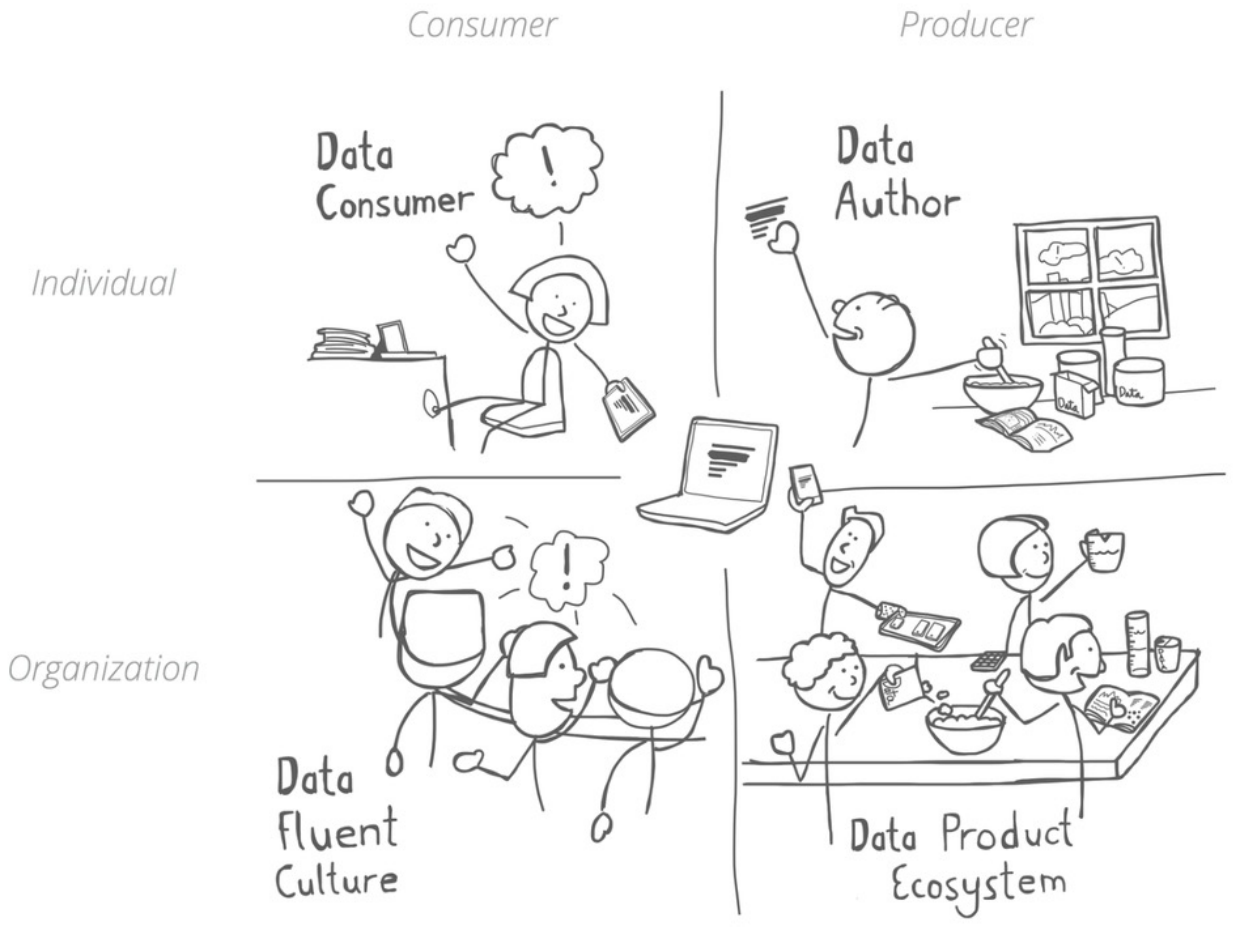


- 텍스트 해석
- 엑셀과 씨름
- 과거집계·시사점X
- 질문·해석·활용에 집중

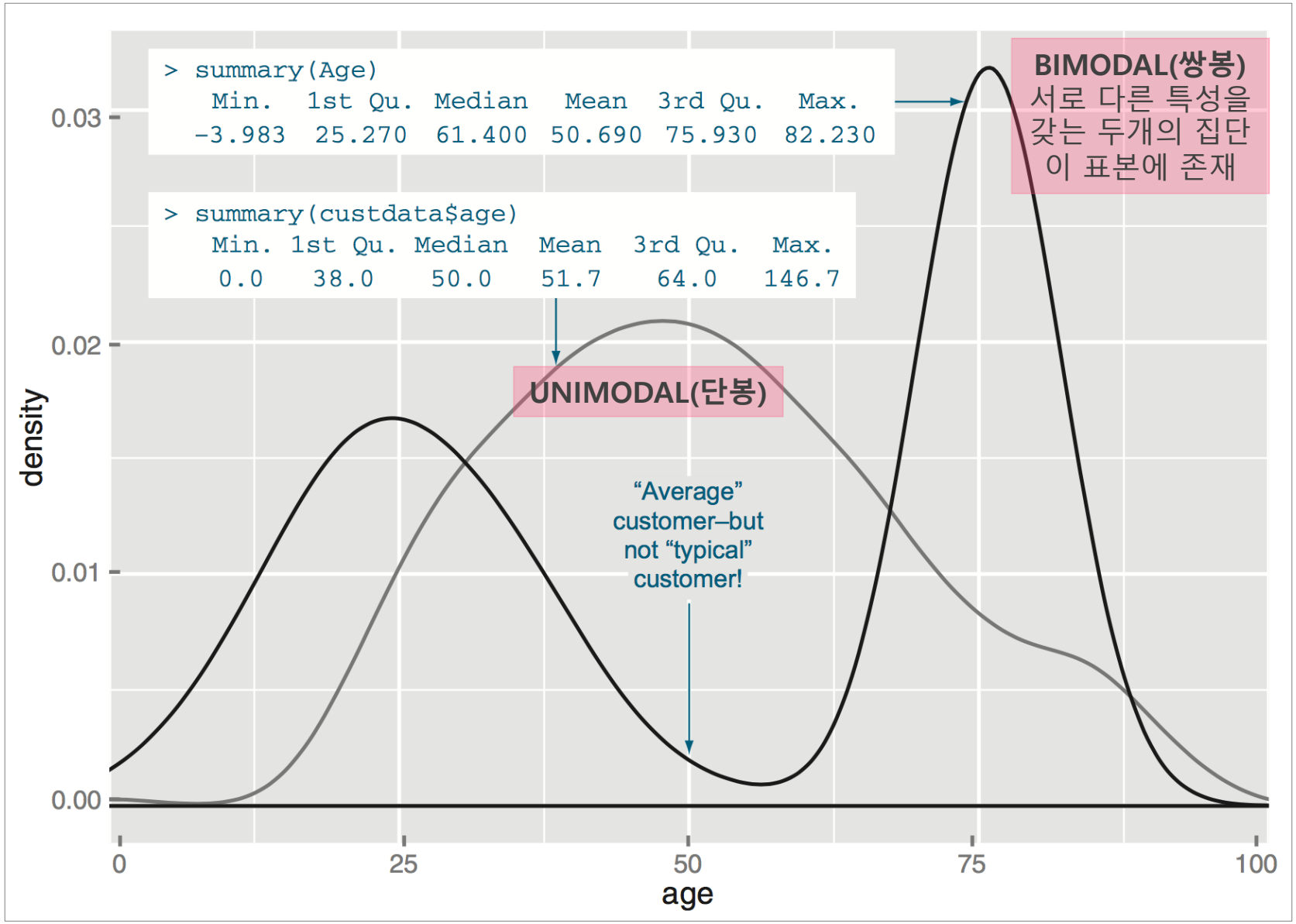
Data Literacy(Fluency) Framework

데이터 분석은 생산자와 소비자 간의 사회적 · 상호적 활동

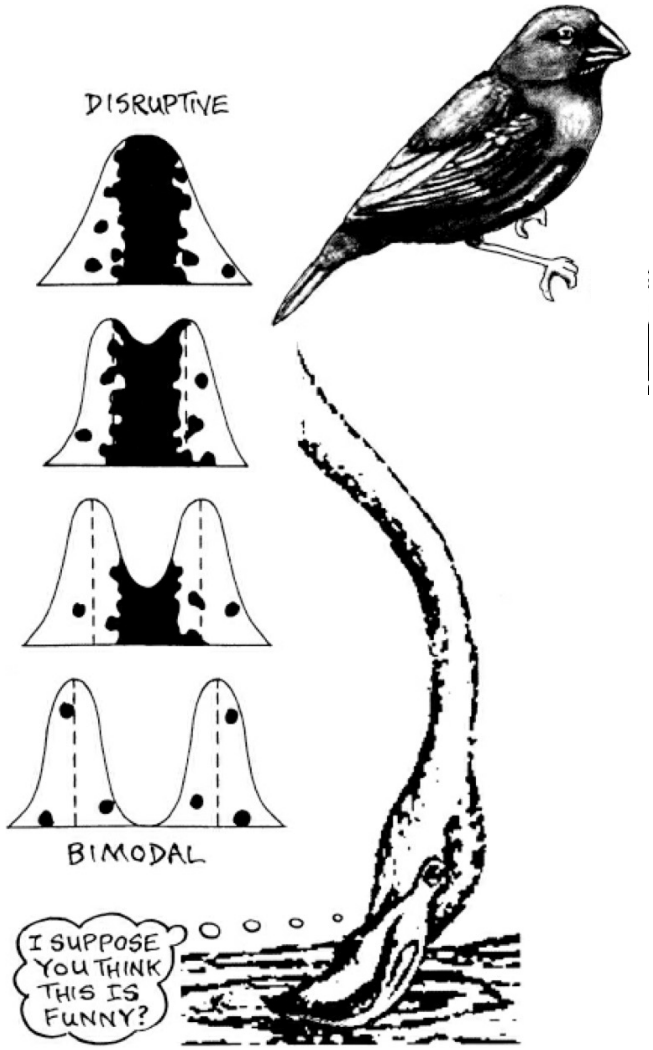
분석 결과를 소비하는 사람이 있어야 분석하는 사람이 존재할 수 있고
좋은 분석을 생산하는 사람이 있어야 또 결과를 소비(활용)하는 사람이 존재



Our First Data Literacy: One Hump vs. Two Humps

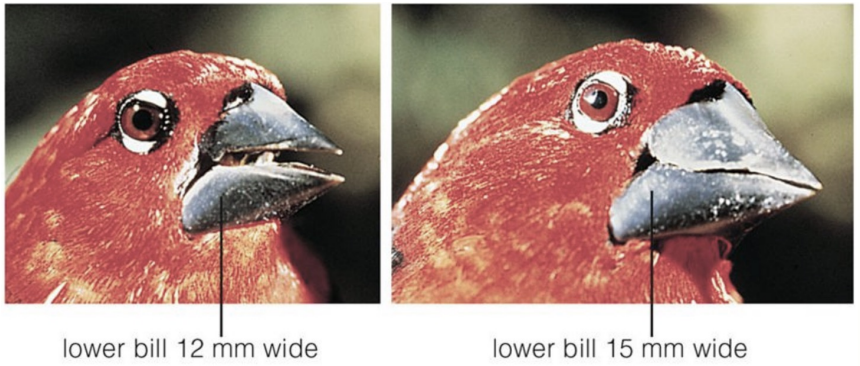


Our First Data Literacy: BIMODAL DISTRIBUTION (쌍봉분포)



African Seedcracker

자연선택에 의해 작고 부드러운 씨앗을
먹는 작은 부리의 새와 크고 단단한
씨앗을 먹는 큰 부리의 새로 나뉨



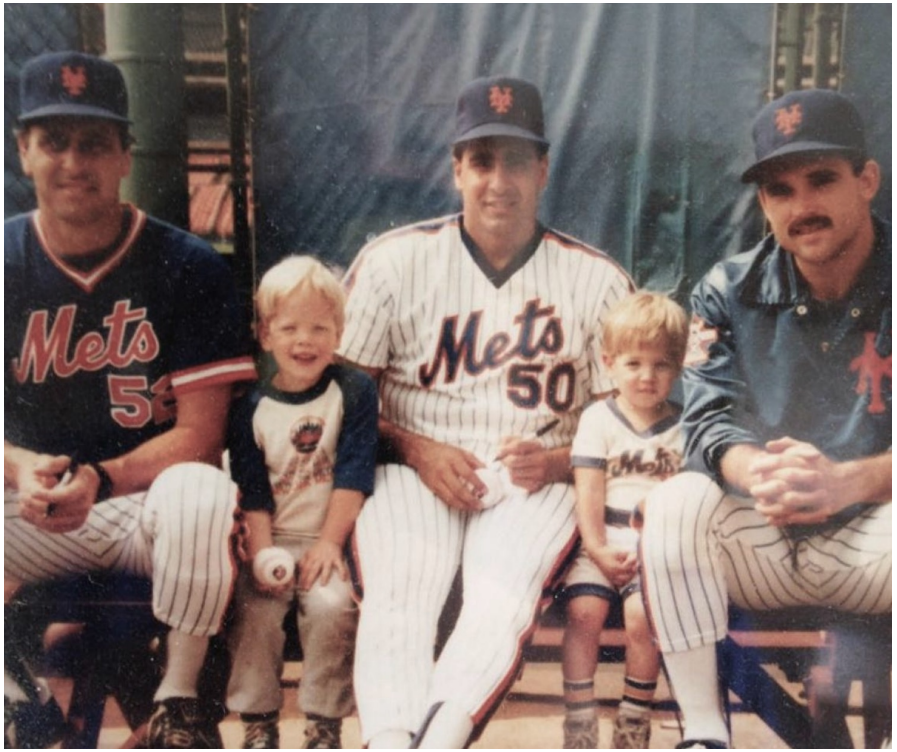
데이터를 읽을 수 있어야 새로운 해석도 가능

- **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?]**

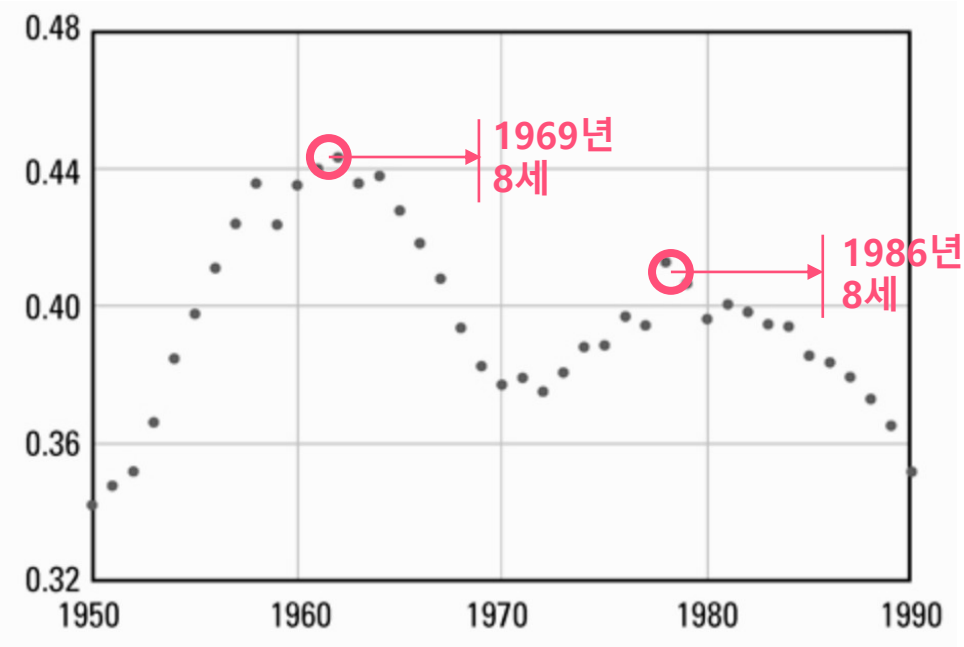
1962년과 1979년에 태어난 남자들은 왜 Mets구단 팬이 되었나?

Hint: 1969년과 1986년에 모두 8세가 되었음



출생년도에 따른 NY Mets 팬들 비율

[대상: 뉴욕 거주하는 남자 야구 팬들]

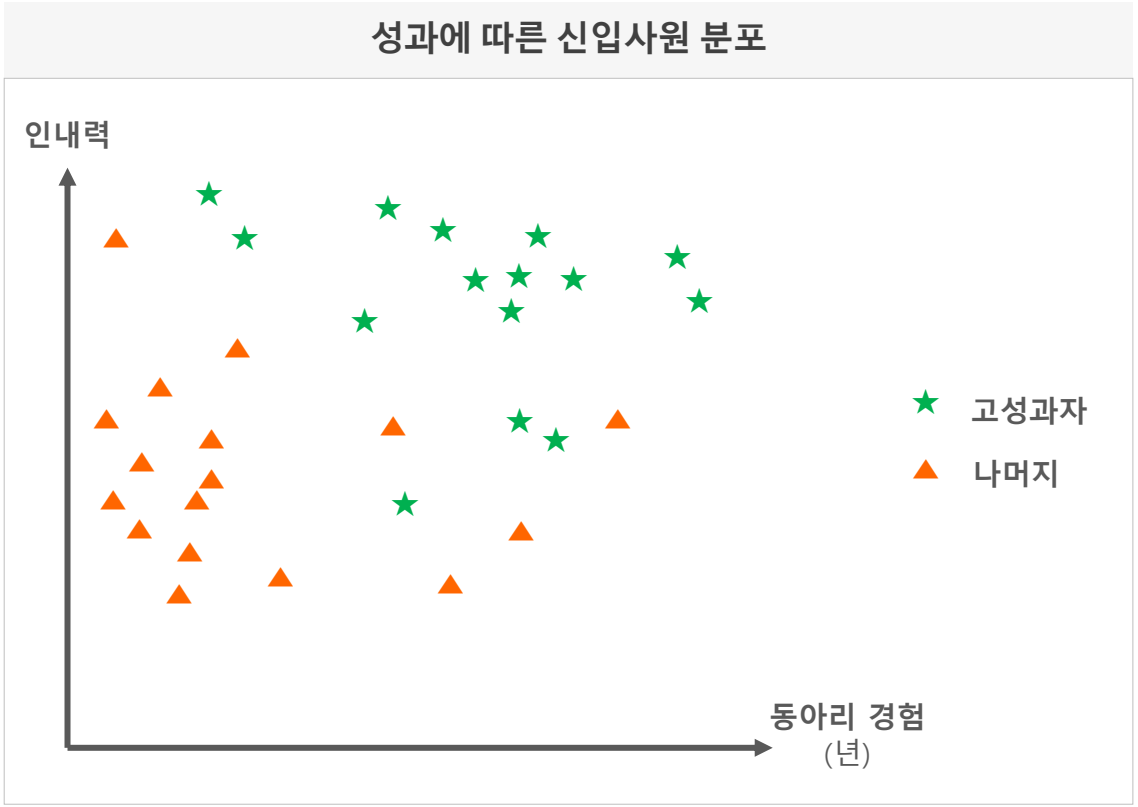


Decision Tree - Minimizing Entropy

Decision Tree Algorithm

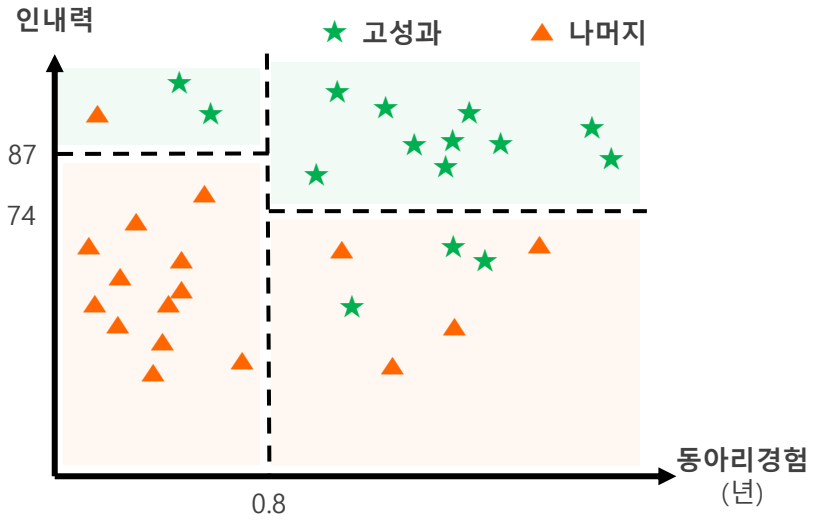
- **Purity:** *엔트로피를 최소화하도록(= 끼리끼리 모이도록) 공간 구획
- **Homogeneity:** 동질적 집단이 밀집한 세그먼트의 논리적 규칙 찾기

***엔트로피(Entropy): Measure of Impurity**

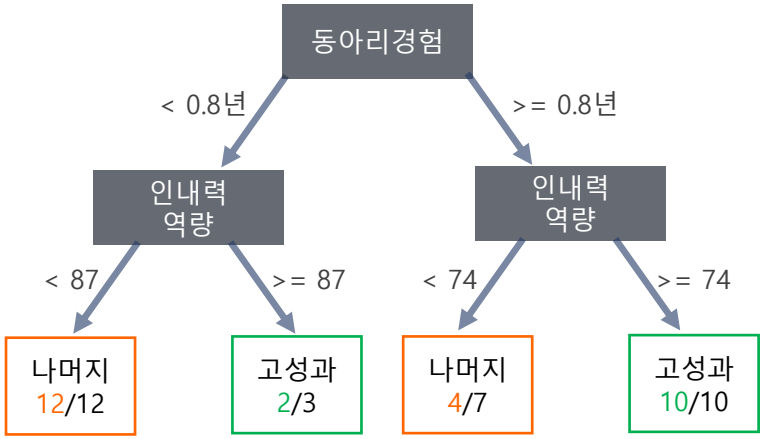


Decision Tree - Minimizing Entropy

성과에 따른 신입사원 분포



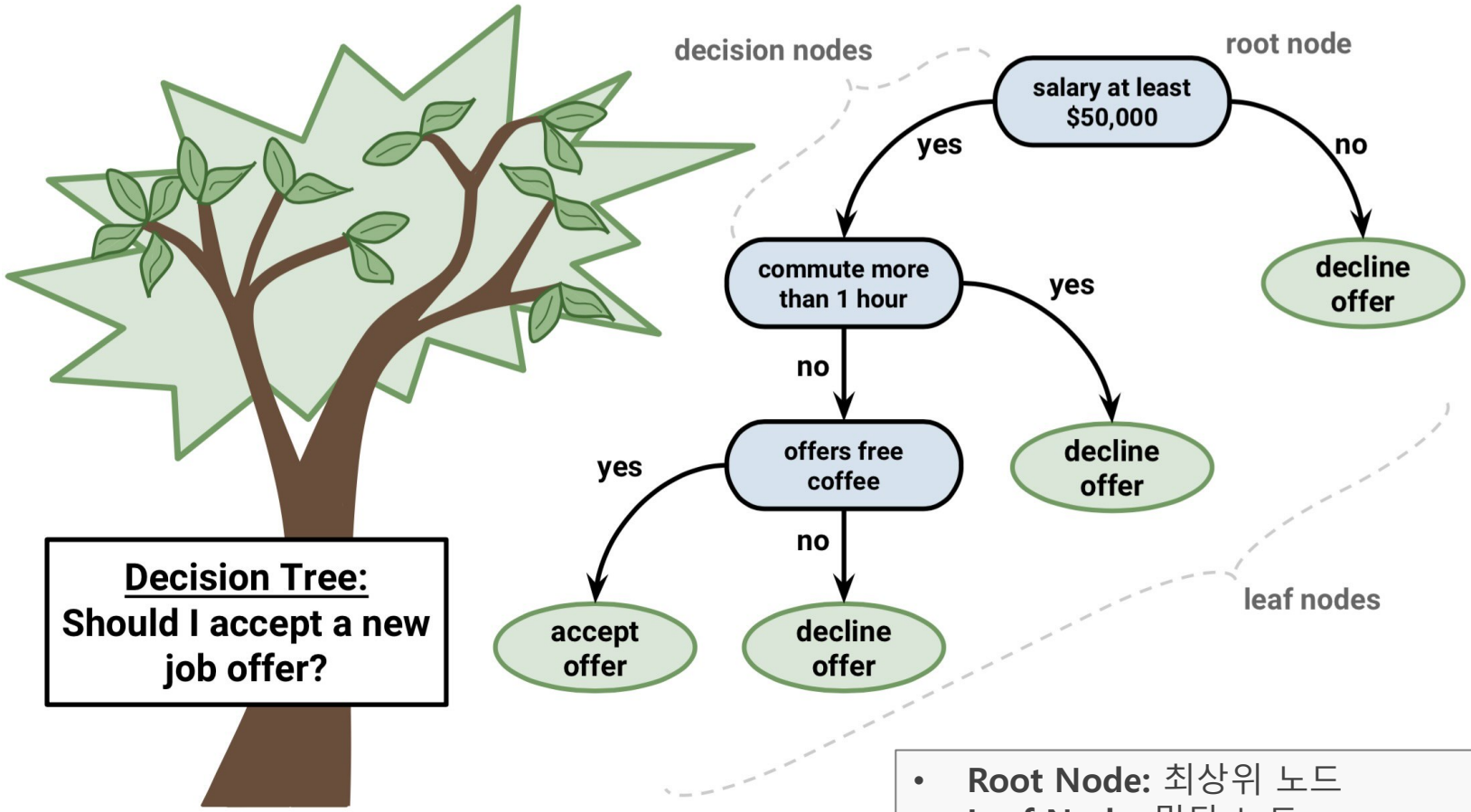
고성과 신입사원 분류모형 (의사결정트리)



분류규칙

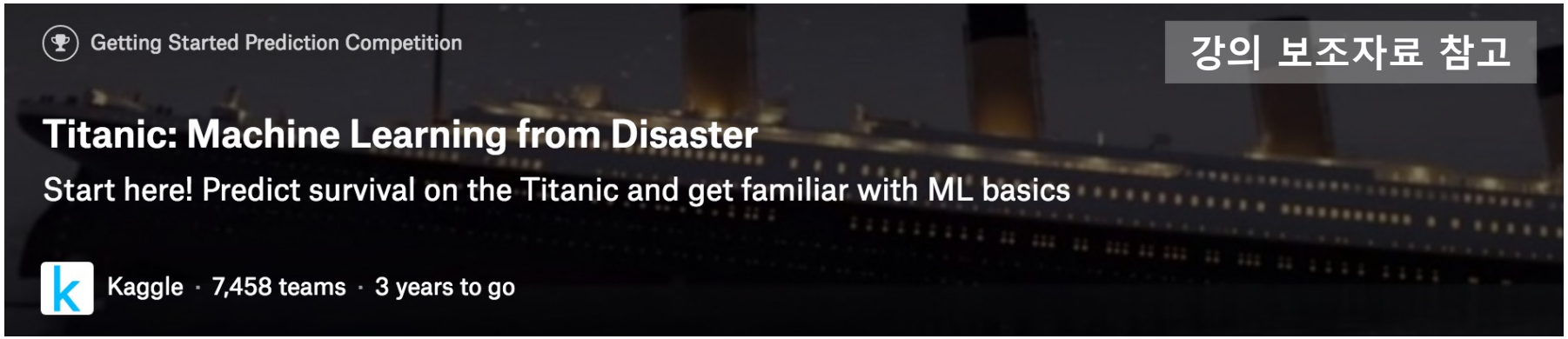
규칙 (Rule Set)	확률 (Probability)
IF (동아리경험 < 0.8년) and (인내력역량 < 87) then Class = 저성과 신입사원	100% (12/12)
IF (동아리경험 < 0.8년) and (인내력역량 >= 87) then Class = 고성과 신입사원	67% (2/3)
IF (동아리경험 >= 0.8년) and (인내력역량 < 74) then Class = 저성과 신입사원	57% (4/7)
IF (동아리경험 >= 0.8년) and (인내력역량 >= 74) then Class = 고성과 신입사원	100%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

Decision Tree – Terminology



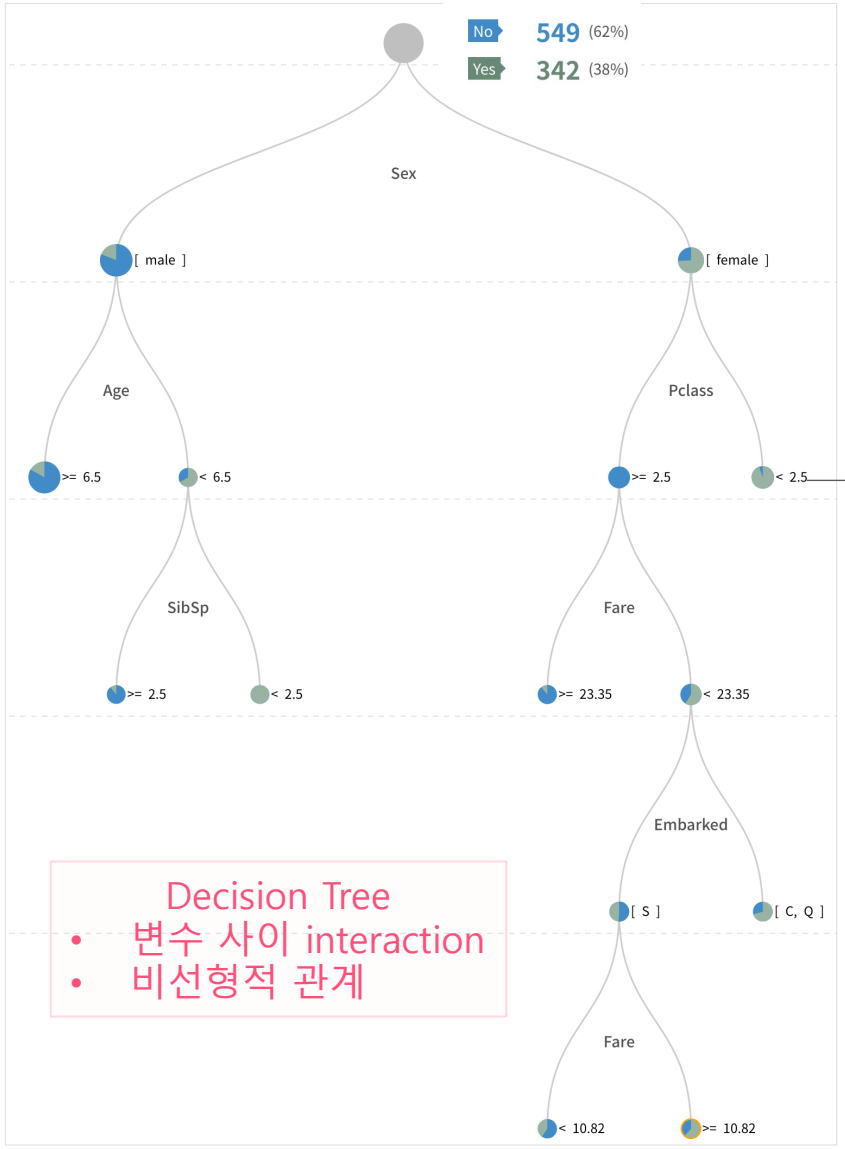
- **Root Node:** 최상위 노드
- **Leaf Node:** 말단 노드
- **Decision Node:** 의사결정 노드
- **Splitting:** 동질적 집단으로 쪼개는 일
- **Pruning:** 트리가 너무 길어지지 않게 하는 일

Decision Tree – Titanic Dataset



Variable	Definition	Key
survival	Survival **(target)**	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



Decision Tree
- 변수 사이 interaction
- 비선형적 관계

세그먼트 특성

클래스	Yes
크기	170 레코드
순도	94.71% (161/170)
타겟 비율	47.08% (161/342)
평균 KPI 차이	+56.32 (94.71 – 38.38)

세그먼트 분류규칙

Sex = [female]

Pclass < 2.5

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

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

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

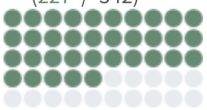
Decision Tree – Confusion Matrix

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 아니다고 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 **89%** = $\frac{227}{227 + 28}$

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