R로 배우는 데이터 분석

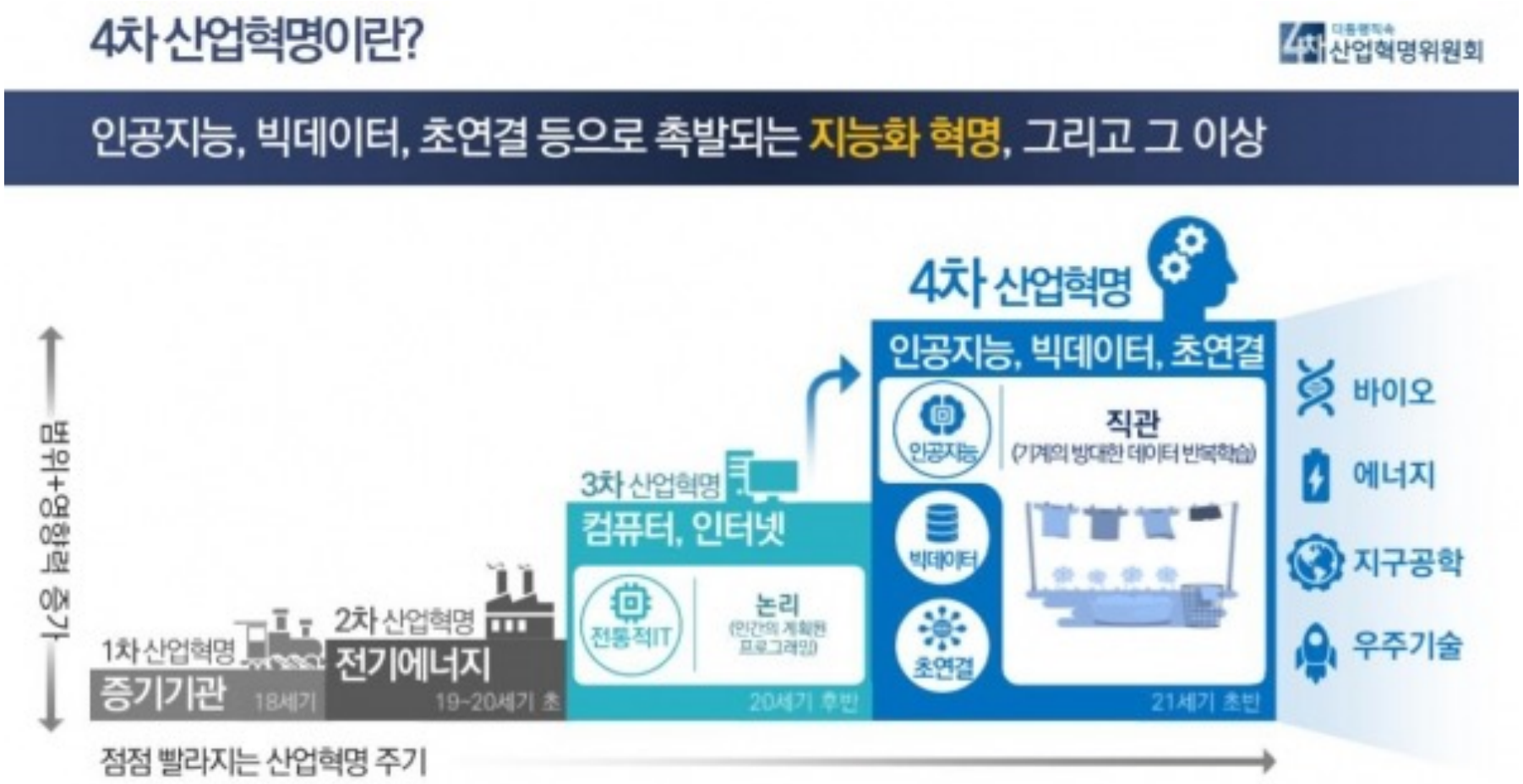
홍 윤 호

서울대학교 의과대학 신경과학 교실

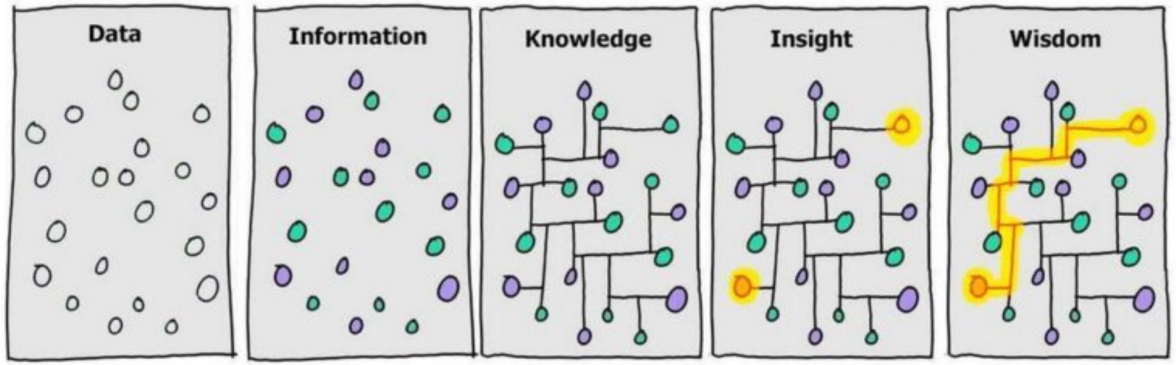
결론부터

- 데이터란 무엇이고, 왜 중요한가?
- 데이터를 분석할 때 빠지기 쉬운 함정과 대책

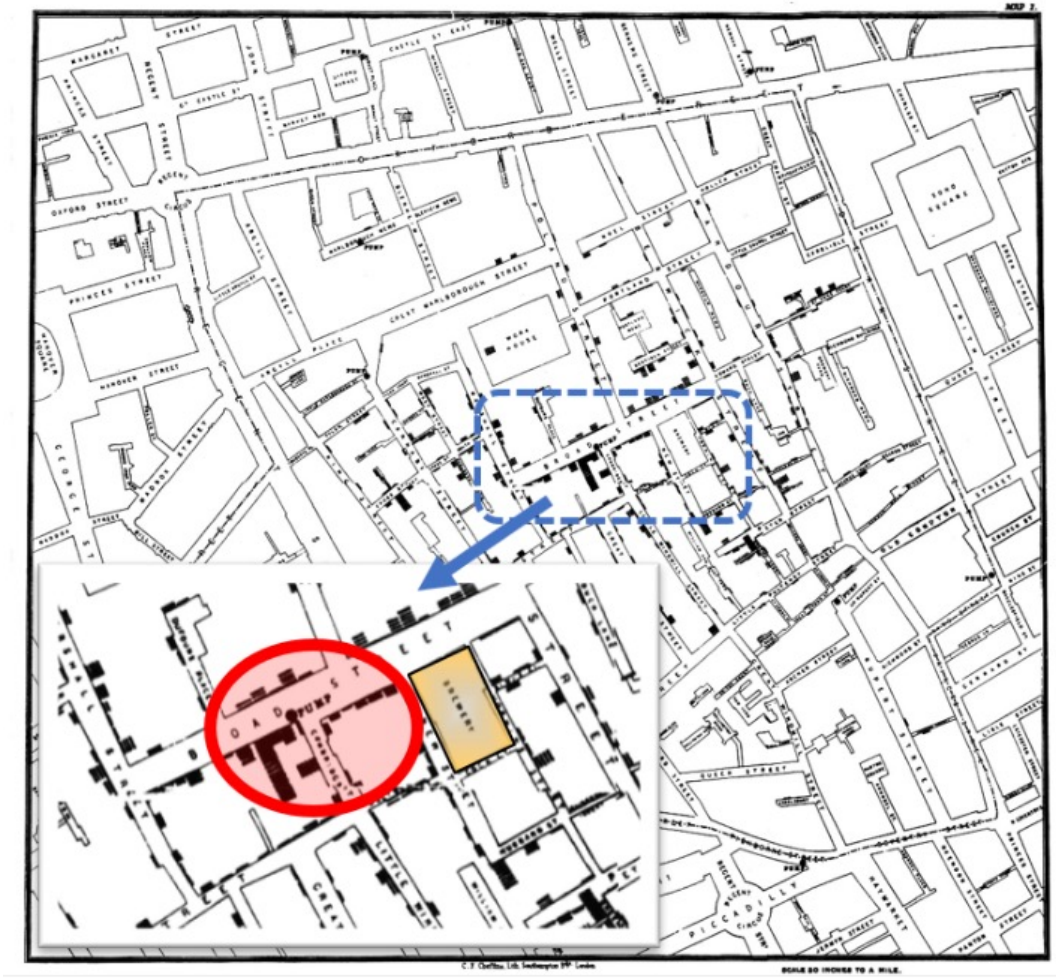
4차 산업혁명의 핵심: 데이터



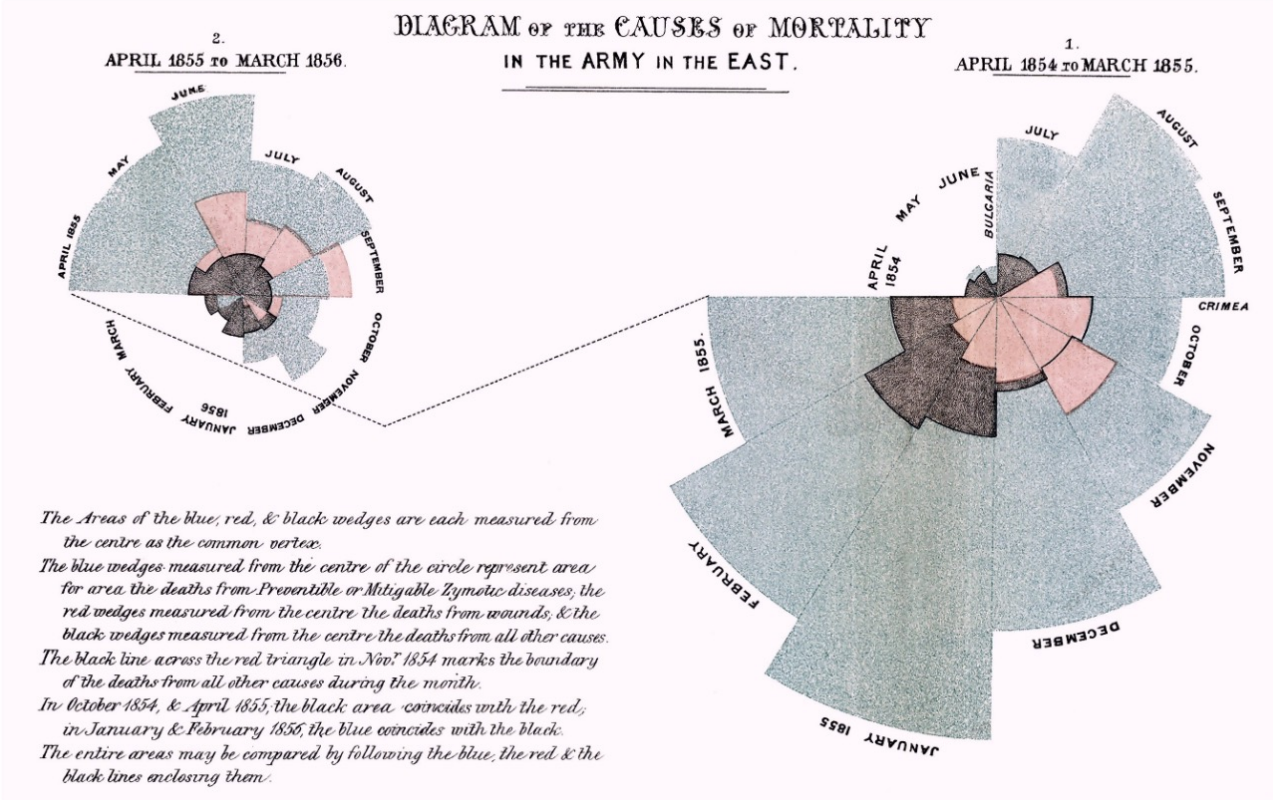
데이터란 무엇이고, 왜 중요한가?



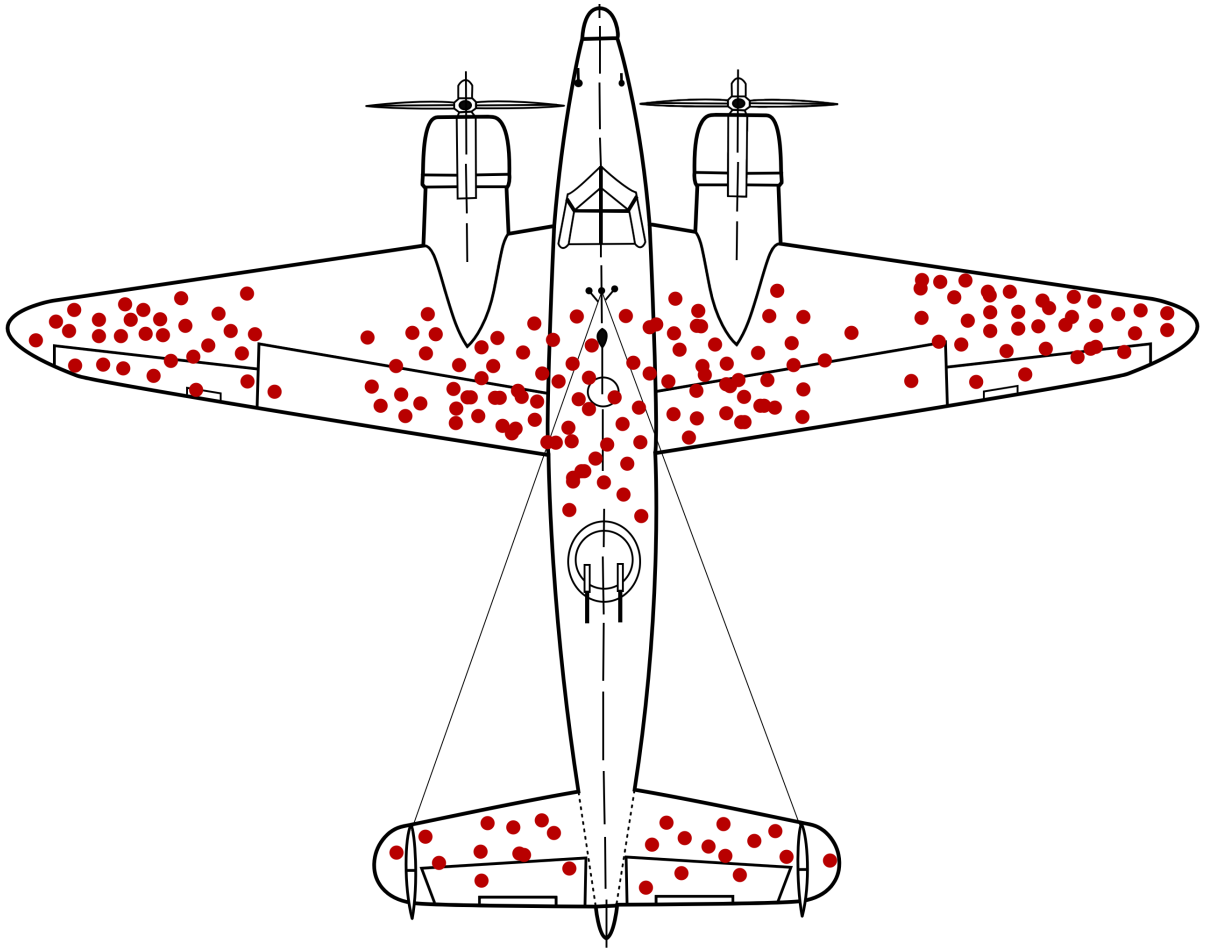
존 스노우
콜레라 지도

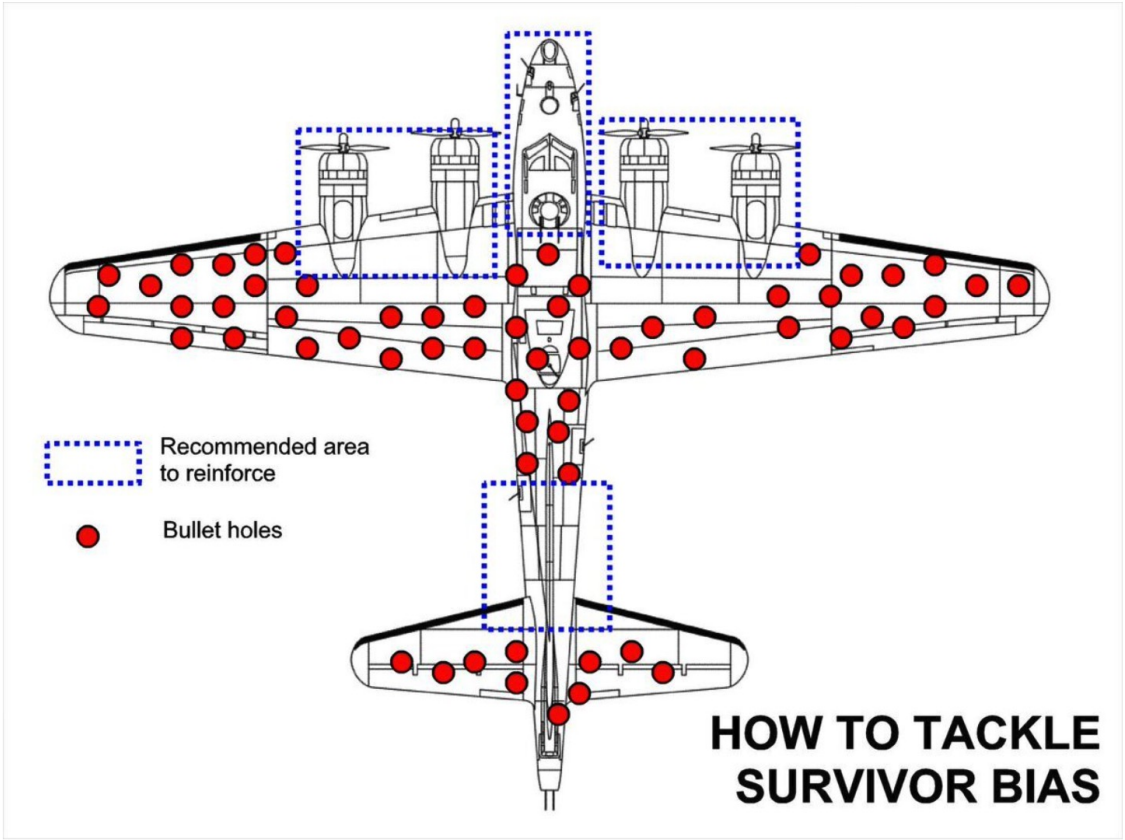


나이팅게일
로즈 다이어그램

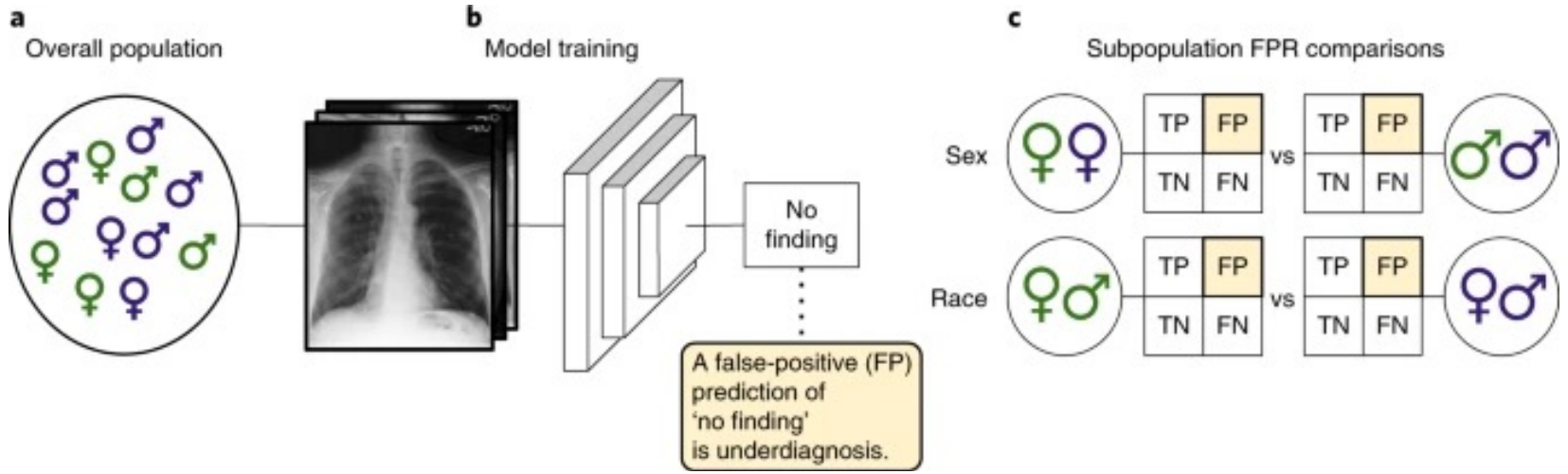


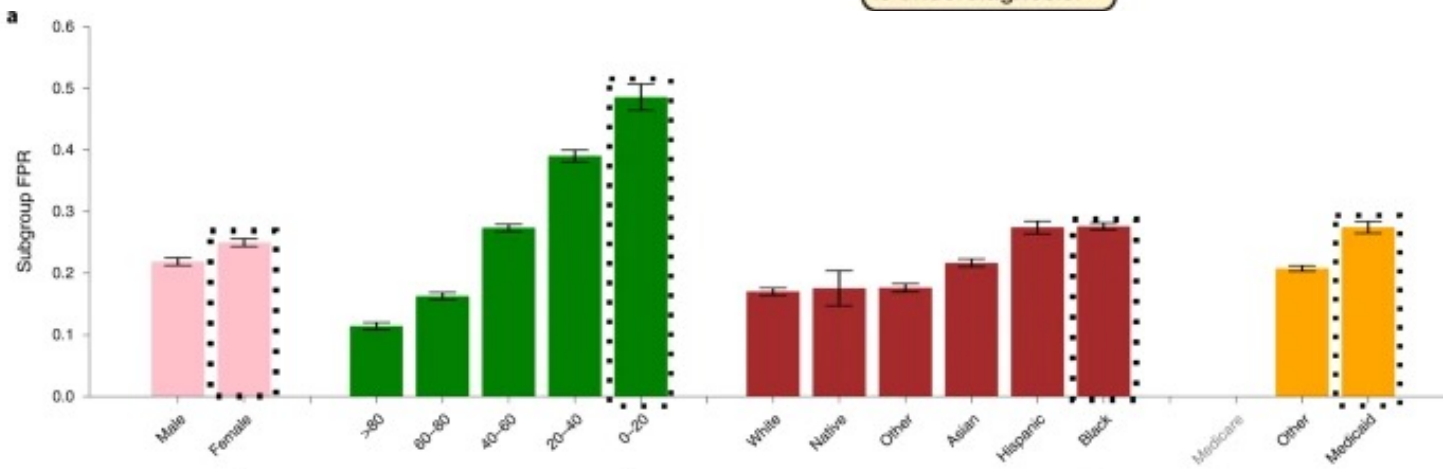
같은 데이터, 전혀 다른 해석





인공지능의 편향: 표본은 적절한가?





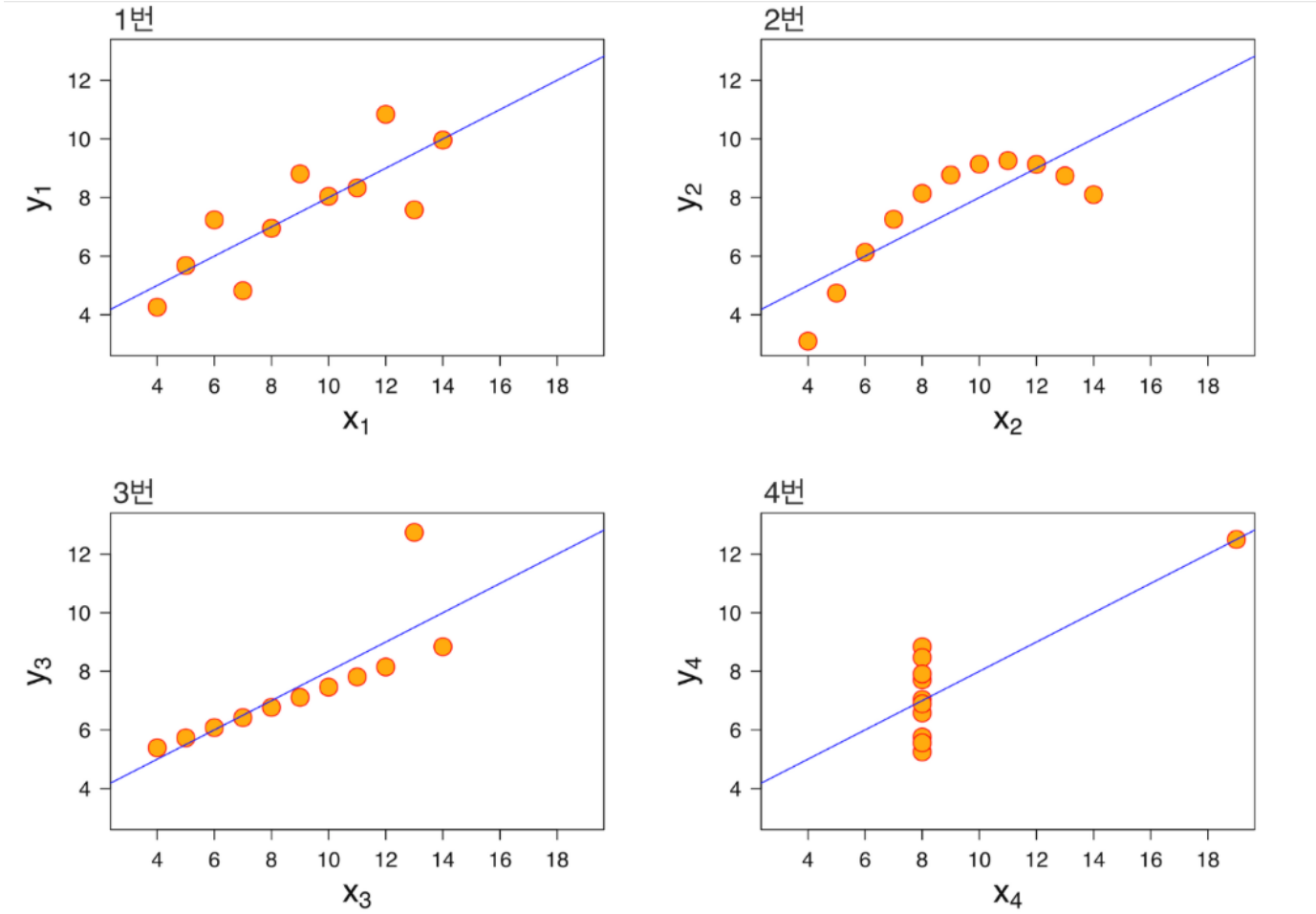
Nature medicine, 2021,
Underdiagnosis bias of AI in under-served populations

결측 데이터(missing)

"

Complete case analysis can lead to incomplete understanding

평균의 함정



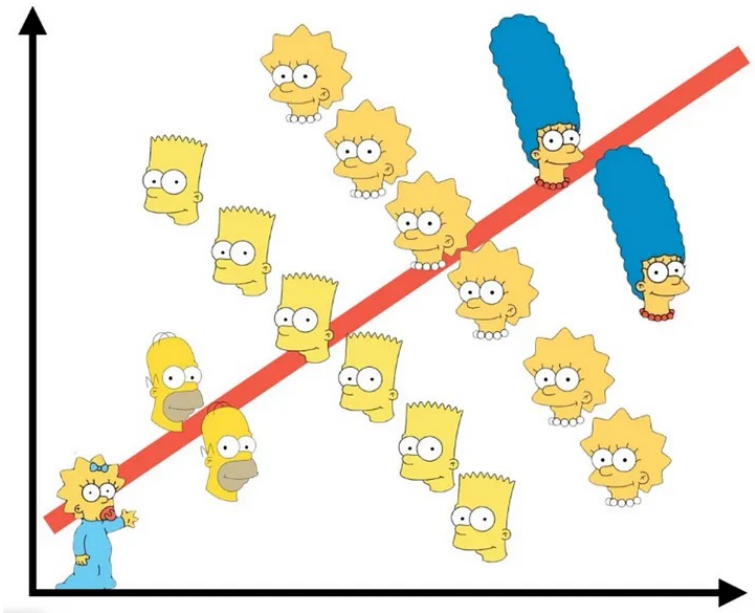
Anscombe's quartet
(앤스컴 쿼르텟)

심슨의 역설: 전체와 부분

전체	지원자	합격자	합격률
남학생	1150	815	71%
여학생	820	185	23%

문학부	지원자	합격자	합격률
남학생	150	15	10%
여학생	700	85	12%

공학부	지원자	합격자	합격률
남학생	1000	800	80%
여학생	120	100	83%

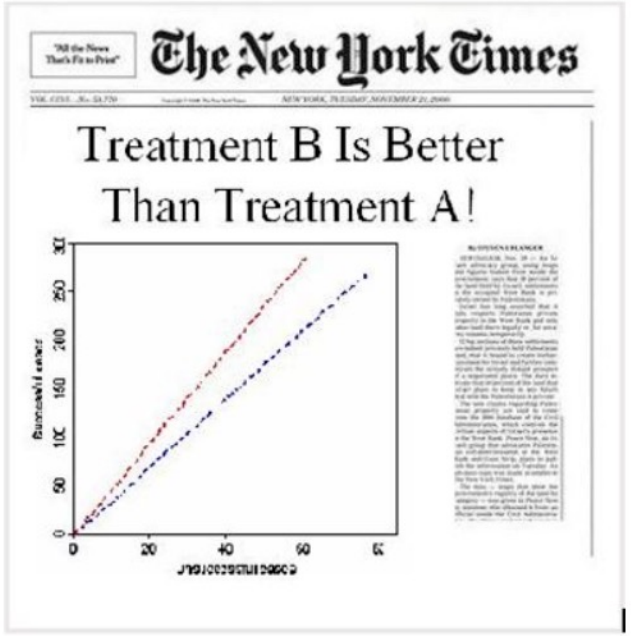


숲도 보고 나무도 보자!

어떤 나무를 보아야 할까?

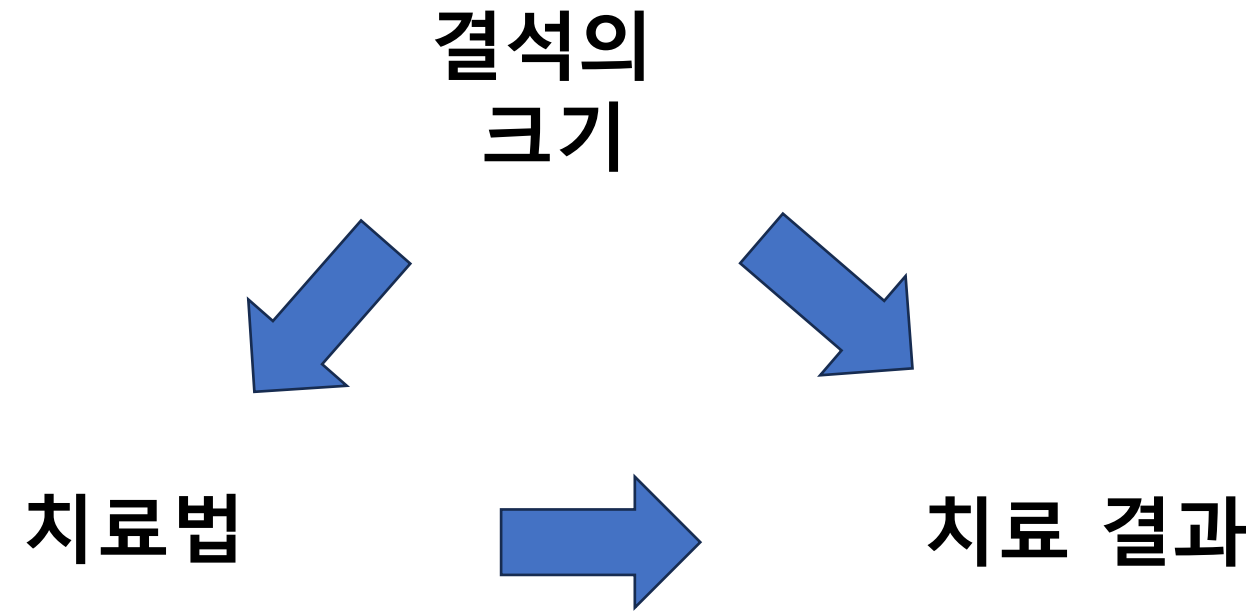
Treatment A	Treatment B
78% (273/350)	83% (289/350)

	Treatment A	Treatment B
Small Stone	93% (81/87)	87% (234/270)
Large Stone	73% (192/263)	69% (55/80)


The New York Times

Treatment B Is Better Than Treatment A!

“교란 변수(confounder)”라는 나무



상관관계와 인과관계: 초콜릿과 노벨상

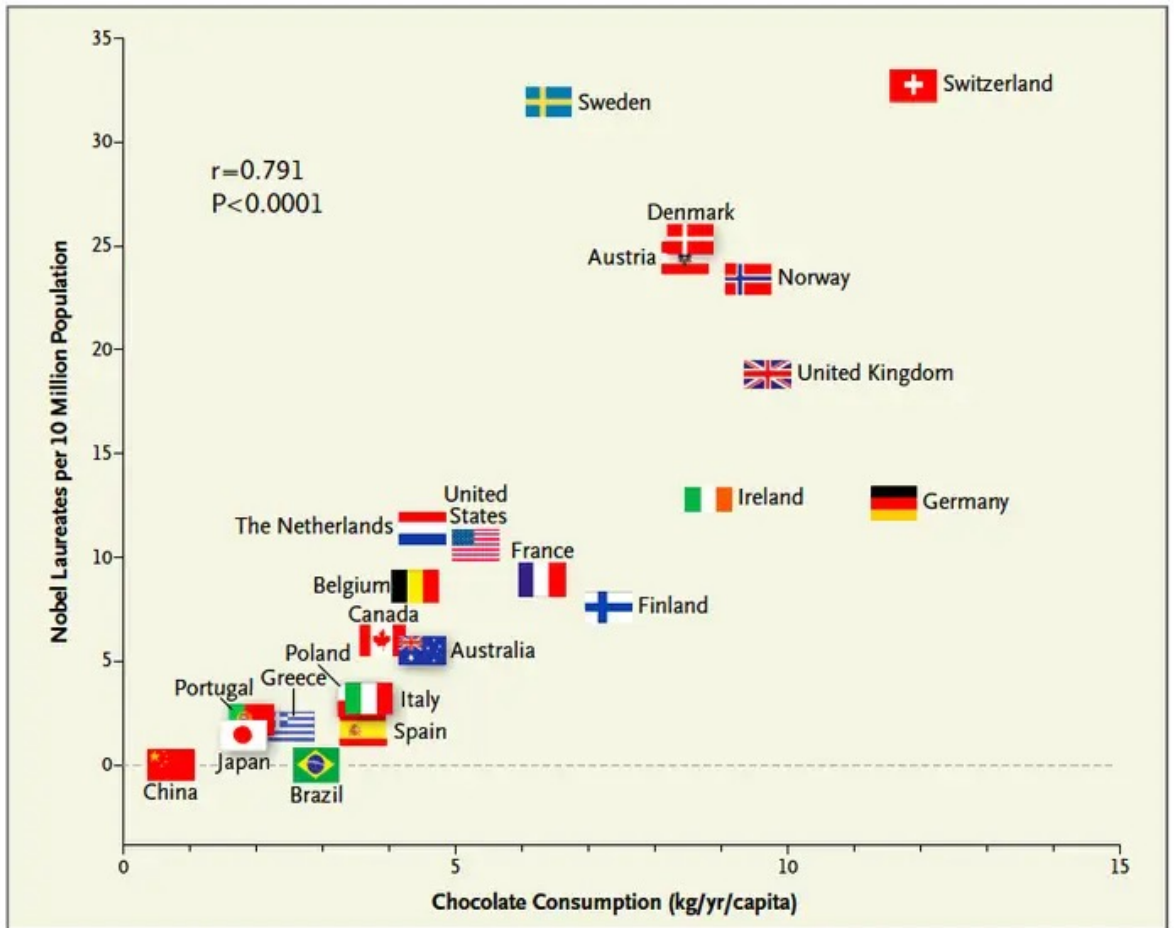


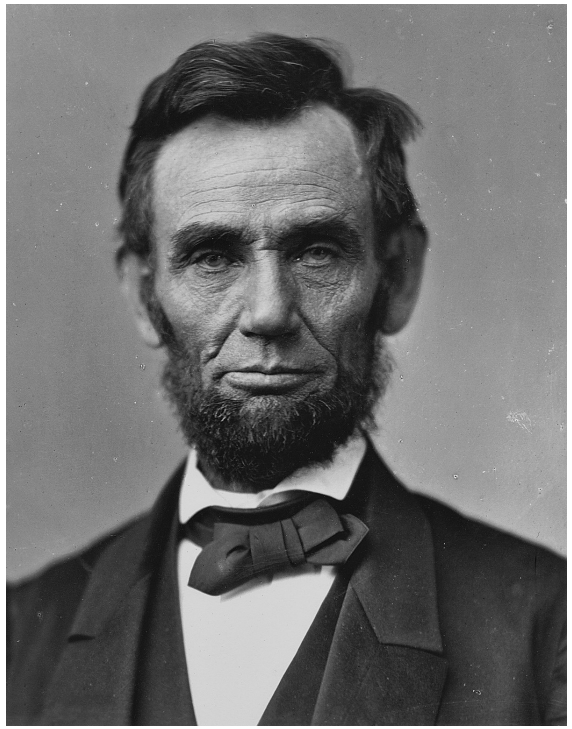
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

인과성 추론의 4가지 요건

1. 원인이 결과보다 시간적으로 먼저 일어나야 한다. (시간적 선후관계)
2. 원인과 결과는 함께 변화해야 한다.(상관관계)
3. 제3의 변수에 의한 영향이 아니어야 한다.(교란변수의 영향 배제)
4. 그럴 듯 해야 한다.(그럴듯함)

다시 결론으로...

- 데이터란 무엇이고, 왜 중요한가?
- 데이터를 분석할 때 빠지기 쉬운 함정과 대책
  - 표본 편향
  - 결측치
  - 전체와 부분
  - 상관관계와 인과관계



"

If I had eight hours to chop down a tree, I'd spend six hours sharpening my axes