# 금융 통계학 기본(1) – 소개

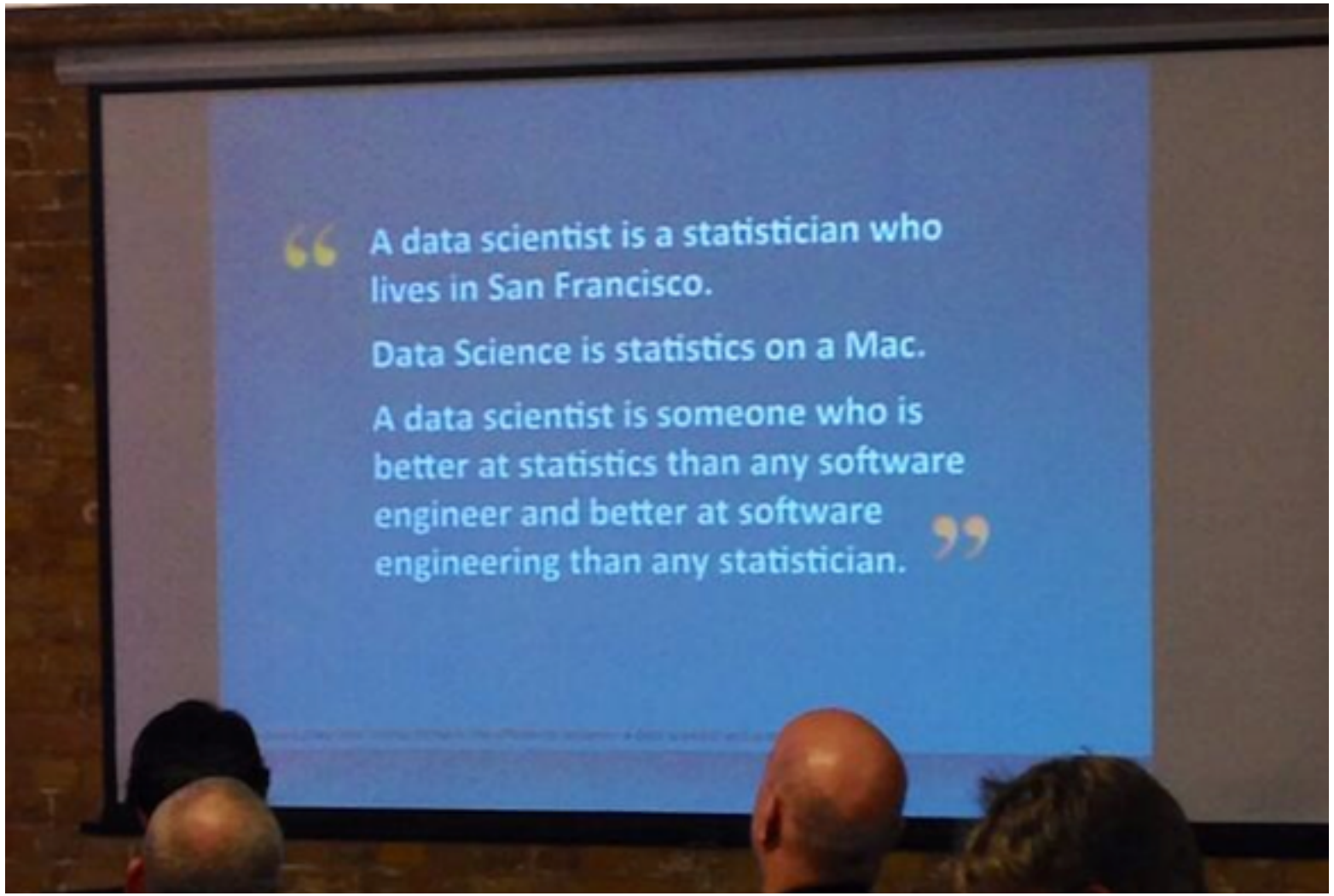
이승준 fb.com/plusjune

# 질문

- 왜 통계가 필요한가?
- 확률과 통계는 왜 같이 따라다니나?
- 데이터 과학과 통계학의 차이?
- 금융 통계학과 일반적인 통계학은 어떻게 다른가?
- 통계학을 알면 수익을 낼 가능성이 높아지는가?
- 빅데이터, 데이터 사이언스, 통계학의 관계?

# Everything Changes, But Nothing Changes

# What is Data Science ?



"Statistics is the **science** of learning from **data**..."

- 미국 통계학회(American Statistical Association) 통계학 정의

# 통계학

- 관심 현상을 수치화하여 측정하고 분석
- 데이터 수집 → 자료 정리 data reduction → 정보제공/의사결정
- 데이터에 기반한 결과를 산출

# 통계학 구성

통계학은 크게 기술 통계descriptive statistics 와 추론 통계inferential statistics로 구성

- 기술 통계: 1)중심경향, 2)산포경향, 3)데이터 모양을  측정, 통계량 statistic 산출
- 추론 통계: 추정 estimation 과 가설검정 testing hypothesis

- 다양한 통계학의 분야들 (수리통계, 데이터 마이닝, 다변량자료 분석, 범주형 자료 분석, 비모수통계학, 생존자료분석, 수리통계학, 시계열자료 분석, 실험계획법…)

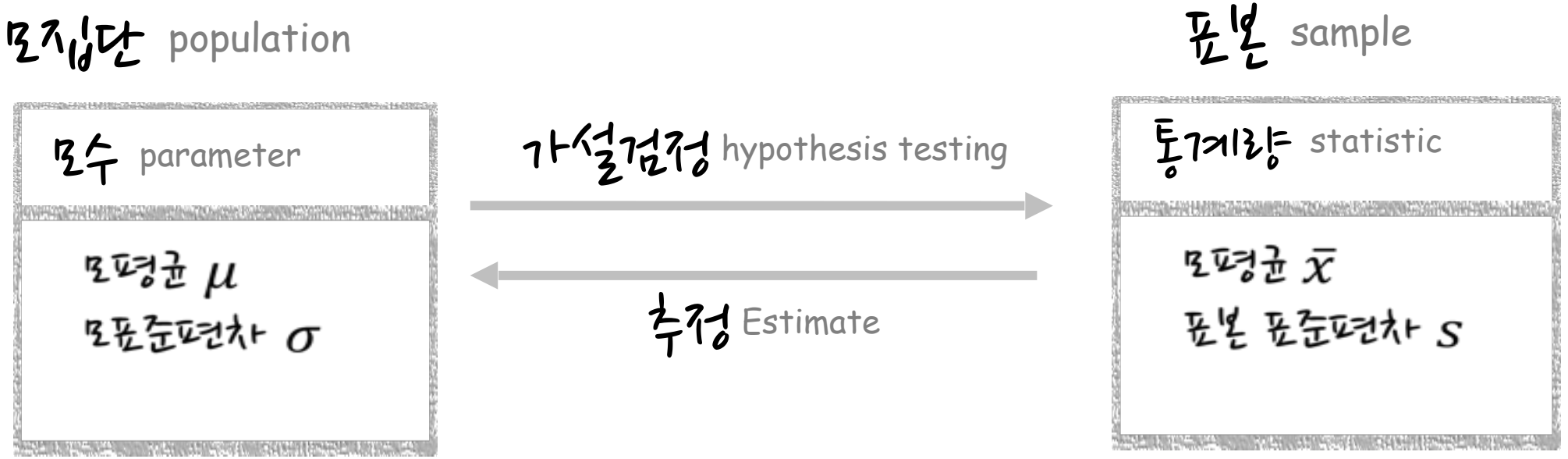
# 통계학 훑어보기

- 조사 대상 전체를 모집단 population, 모집단의 특성을 모수 parameter 라고 하며,
- 모집단에서 요소를 무작위로 random 골라낸다 sampling. 골라낸 것을 표본 sample이라 한다
- 랜덤하게 추출하기 때문에 확률 probability이 함께 쓰인다.
- 표본의 특성 값들을 통계량 statistic 이라고 한다. (통계학 statistics 이란 용어 기원)

결국, 통계학이란 대상의 특성을 추정하는 과정

대상 "전체"population 를 한번에 조사할 수 없을 때 조사 대상의 "부분" sample 을 조사하고,
그 결과를 가지고 조사 대상 전체를 추청 (추측하여 정함)

# 모집단/모수 와 표본/통계량 관계



# 데이터 사이언스

데이터로부터 의미 있는 정보를 추출해내는 과정

~~자료 과학? 데이터 과학? 데이터 사이언스?~~

# 데이터 사이언스의 과정

1) 수집 Collect Data : Question, Collect data

2) 처리 Prepare Data : Organize, Cleanse

3) 분석 Analysis Data : Pattern, Filter, Relationship, Summarize

4) 적용 Apply Data : Visualize, Report, Share, Make decisions

# 빅데이터

- 기존 데이터베이스 관리도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합
- 3V (Volume, Variety, Velocity) + Value(가치)

"데이터가 수단에서 그 자체가 탐구의 대상으로 진화"

# 빅데이터 & 스몰데이터

빅데이터의 약속

① 개인화, 맞춤형 제안

② 비정형적인 데이터에서 상관관계

질문,

- 빅데이터, 우리에게 꼭 필요한 것인가?

# 스몰데이터에서 시작하자

- 이미 가지고 있는 것(정형 데이터), 얼마나 잘 쓰고 있나?
- 모든 데이터는 테이블이다
- 빅데이터를 작게 만들 수 있다
- 도구의 문제나 데이터 크기의 문제 보다 컨텍스트의 문제

# No Silver Bullet

11월, Google은 자사의 머신러닝 소프트웨어인 TensorFlow를 오픈소스로 공개
에릭 클랩튼이 자기 기타를 아무나 연주하게 허락한 것과 같다고 보면 됨.

하지만, 그 기타로 연주한다고 에릭 클랩튼 되는거 아니라는거

# 상관관계 vs. 인과관계

- 애리조나 주는 다른 주보다 폐결핵으로 죽는 사람이 많다: 잘못된 추론
- 인구 10만명당, 경찰 수가 많을수록 범죄 건수가 많다: 역방향 인과 관계
- 맥주 소비량과 영아 사망률의 음의 상관관계가 나타난다: 제 3의 상관관계

데이터 분석에서 밝혀주는 것은 상관관계 Correlation 뿐,
인과관계 Causation 는 설명해주지 않는다.

# 예측하지 말고 측정하라 (DDDM)

- "모든 의사결정은 데이터에 기반한다" – 구글의 규범
- 뱅뱅이론: 세상이 사실은 그렇게 돌아가지 않고 있다
- “우리의 가정은 틀렸다. 그래서 데이터를 봐야 한다” – 하용호 (데이터 사이언티스트)
- “상상하지 말라. 무얼 상상하건 실제와 다르다” – 송길영 (다음소프트 부사장)

## "예측"하지 말고 "측정"하라

절대적인 확신을 갖지 말고 가능성 위주로 탐구하는 확률적 사고가 중요.