엔지니어를 위한 데이터 문해력

정 성 규
서울대학교 통계학과

데이터, 통계와 의사결정

- 모더나 백신 임상시험: 30000명 대상으로 시험

VS

79억명의 지구인

3만명의 미국인

3만명의 결과가 어떻게 모든 사람에게 적용되는가?

정성규 ©

Photo: Shutterstock, https://www.swedishnomad.com/how-many-people-in-the-world/
https://www.mightyrecruiter.com/

데이터의 불확실성

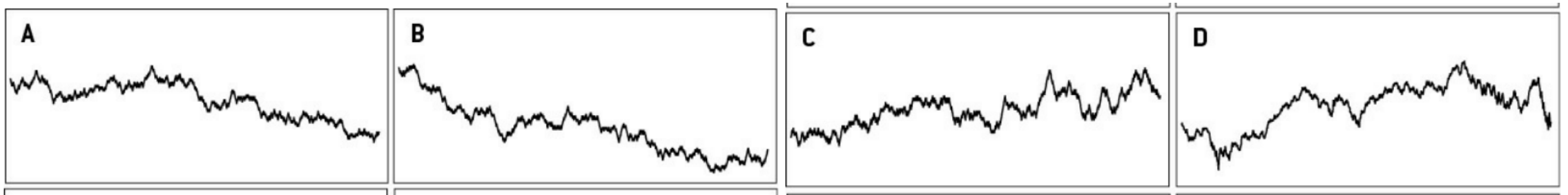
데이터의 불확실성 1: 전체가 아닌 부분으로부터 결론을 내릴 때의 불확실성

어떤 "부분"이 뽑히느냐에 따라 결론이 달라짐

데이터의 불확실성 2: "오늘을 다시 산다면?" 다른 데이터 관측

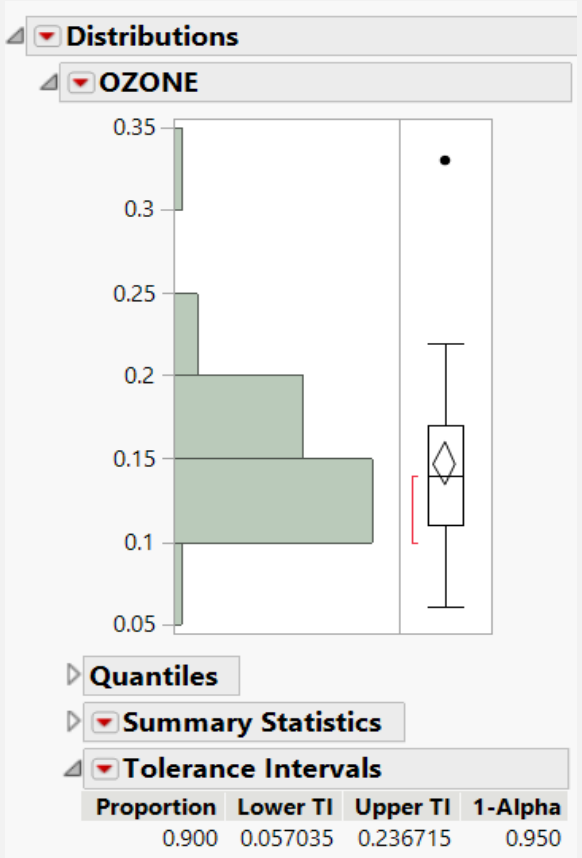
데이터 = 내재된 패턴 + 가늠할 수 없는 노이즈

또는 "신호 + 소음"

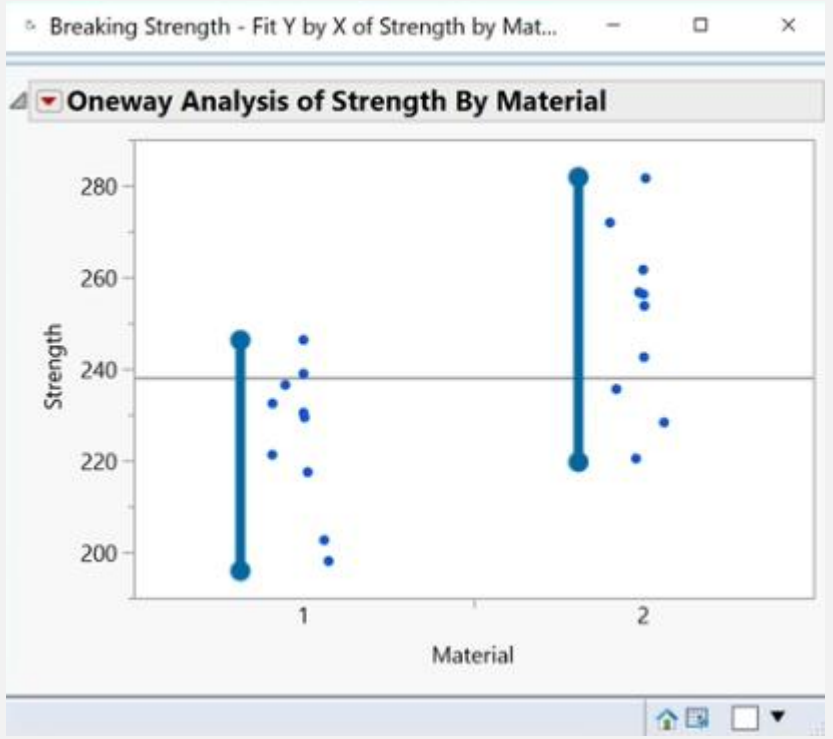


내재적 불확실성

Ozone Concentration in the Atmosphere



Breaking Strength by Material Type



정성규 ©

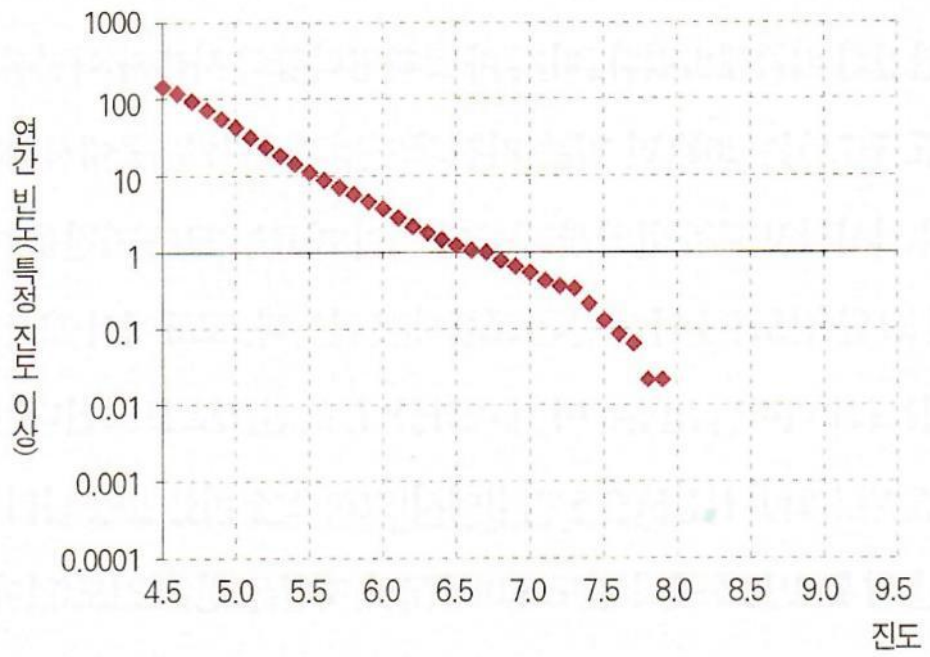
후쿠시마 핵 원자로는 왜 폭발했을까?

2011년 3월 11일 오후 2시 46분 발생한 규모 9.0 동일본대지진 …
사상 최악의 재난

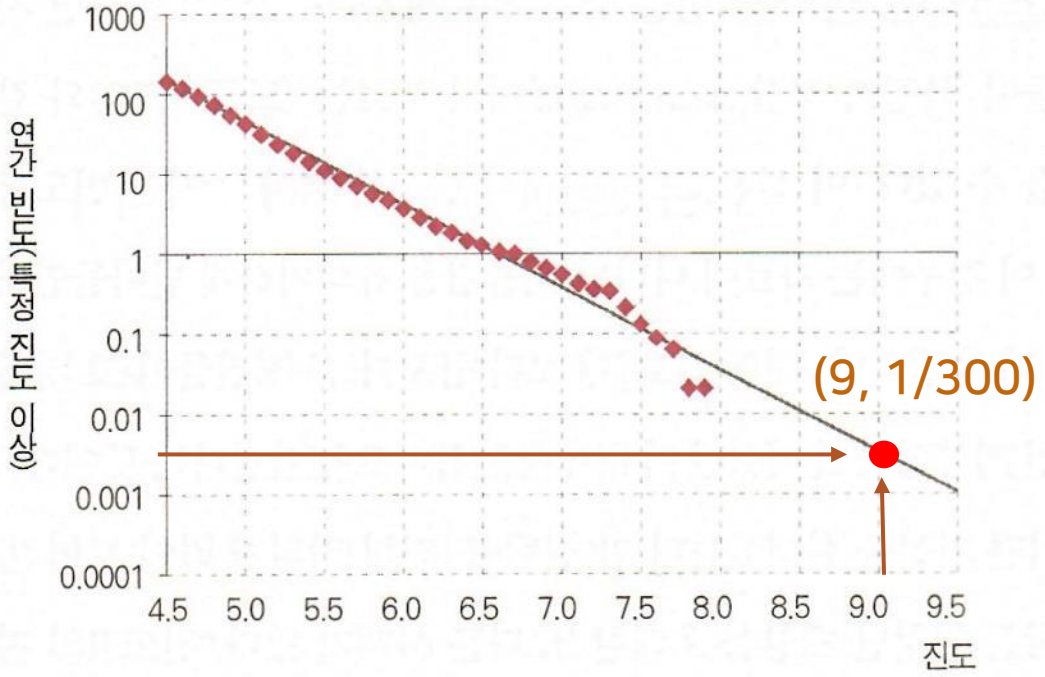
후쿠시마 핵 원자로는 왜 폭발했을까?

구텐베르크-리히터 법칙

| 5-7A | 일본 도호쿠의 지진 빈도(1964년 1월 1일~2011년 3월 10일)



| 5-7B | 일본 도호쿠의 지진 빈도(구텐베르크-리히터 적합)



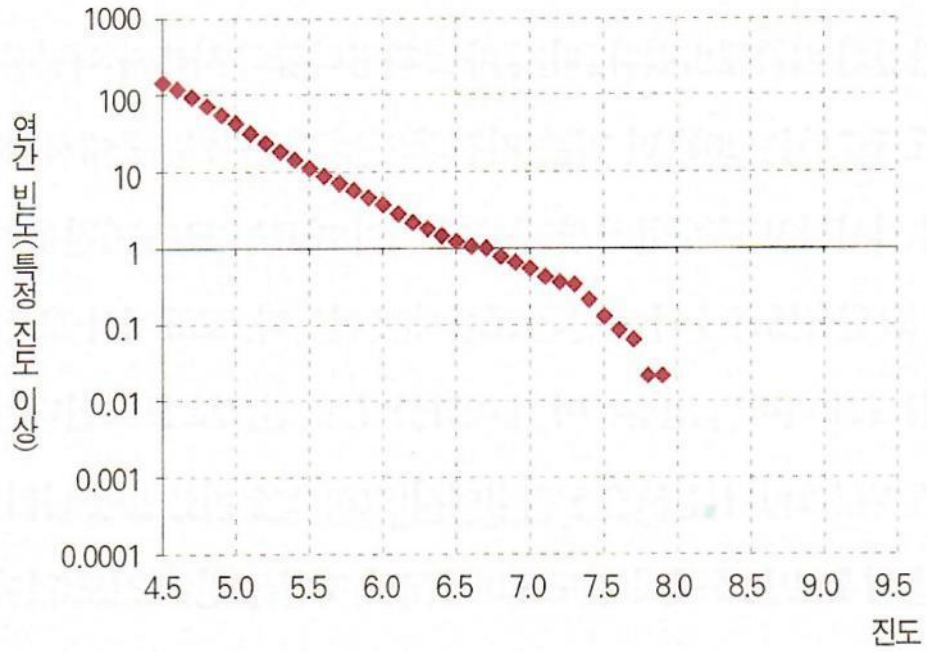
진도 9.0 이상의 지진이 300년 만에 한 번씩 나타난다!

신호와 소음, 네이트 실버, 2012

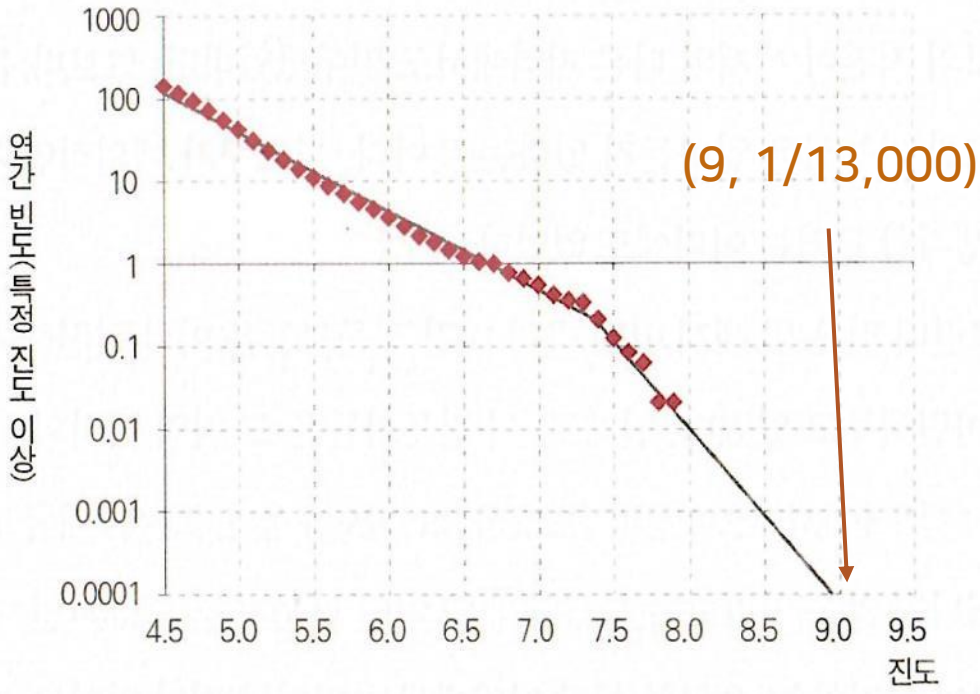
후쿠시마 핵 원자로는 왜 폭발했을까?

추세선 (또는 국소회귀)

| 5-7A | 일본 도호쿠의 지진 빈도(1964년 1월 1일~2011년 3월 10일)



| 5-7C | 일본 도호쿠의 지진 빈도(특징적 적합)



진도 9.0 이상의 지진이 1만 3,000년 만에 한 번씩 나타난다! → 과적합

신호와 소음, 네이트 실버, 2012

후쿠시마 핵 원자로는 왜 폭발했을까?

- 지진의 규모와 횟수에 대한 구텐베르크-리히터 법칙에 따르면 진도 9.1의 지진은 충분히 일어날 수 있다.

- 후쿠시마 지역의 규모와 횟수의 자료에 추세선을 그리면 진도 9.1이 일어날 가능성이 매우 작음

"과적합은 **소음**까지 계산에 넣어
추가점수를 받았을 뿐이다."

네이트실버

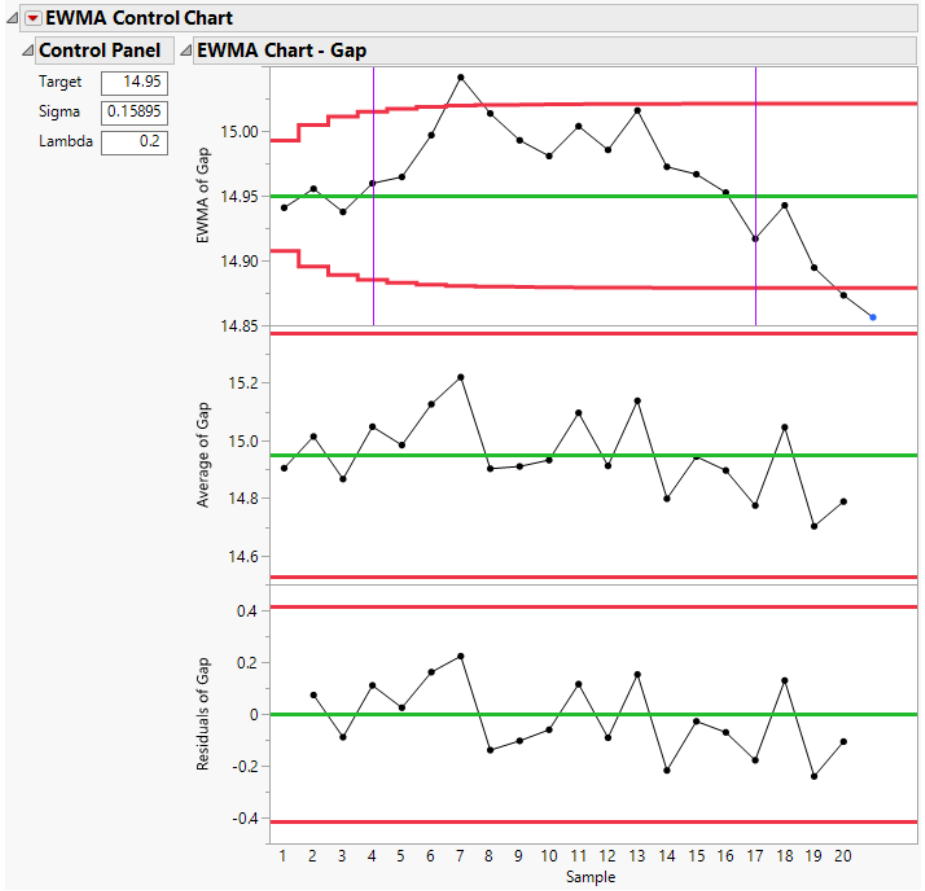
정성규 ©

데이터 문해력: 데이터 분석의 이면

어떻게 달라질 수 있었을까?

라고 질문하라

소음의 크기와 의사결정

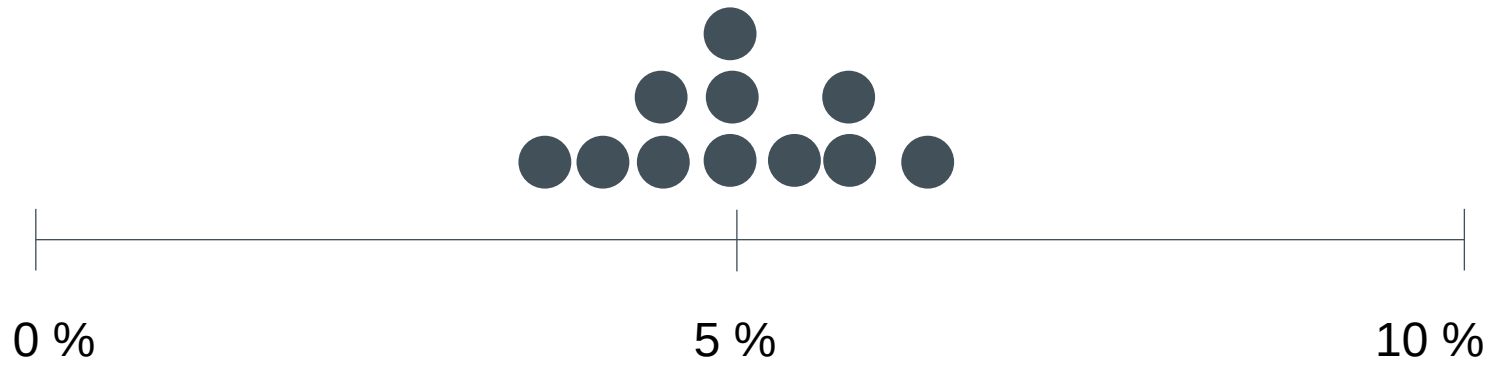


정성규 ©

오늘을 다시 산다면?

- SQC에서의 데이터 변동성: 전체가 아닌 부분만을 보기 때문

- 오늘의 조사 결과 100개 중 4개가 기준 충족
- 오늘을 다시 산다면? 100개 중 5개가 충족
- 오늘을 반복해서 다시 산다면?

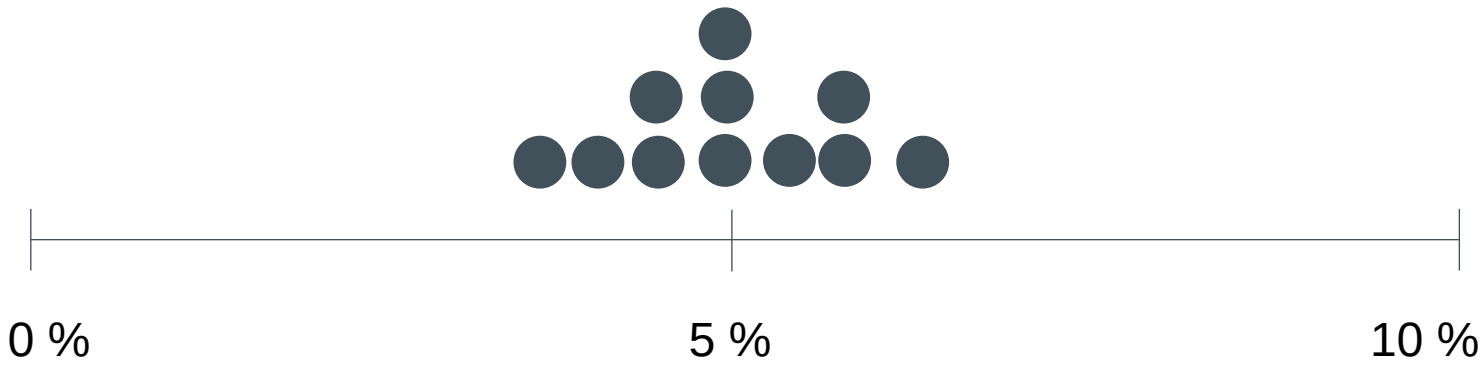


정성규 ©

오늘을 다시 산다면?

- 품질 검사를 반복하여도 “평균” 5%라는 “신호”는 그대로,
  $\pm 2\%$라는 변동 (소음) 이 있음

- 오늘을 한번만 사는 우리는 관측된 4%의 조사결과로부터 “신호”를 추정



정성규 ©

오늘을 다시 산다면?

- 100개가 아닌 1000개를 조사한다면,
  “평균” 5%라는 “신호”는 그대로,
  변동의 양이 $\pm 2\%$에서 $\pm 0.7\%$ 로 줄어듬

큰 수의 법칙

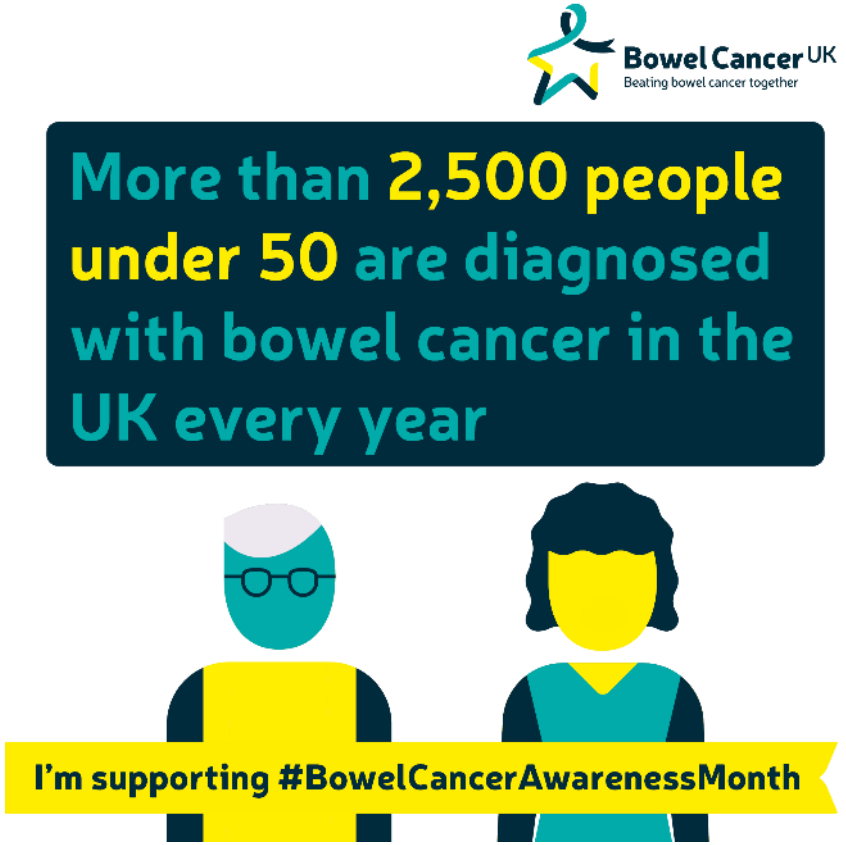
데이터 양이 늘어날 수록 진실에 가까워진다

정성규 ©

Clackmannanshire의 높은 대장암 발병률

영국은 대장암 발병률이 높기로 유명
영국의 Clackmannanshire에서는
대장암 발병률이 750pm에 달했다

What's wrong with Clackmannanshire?



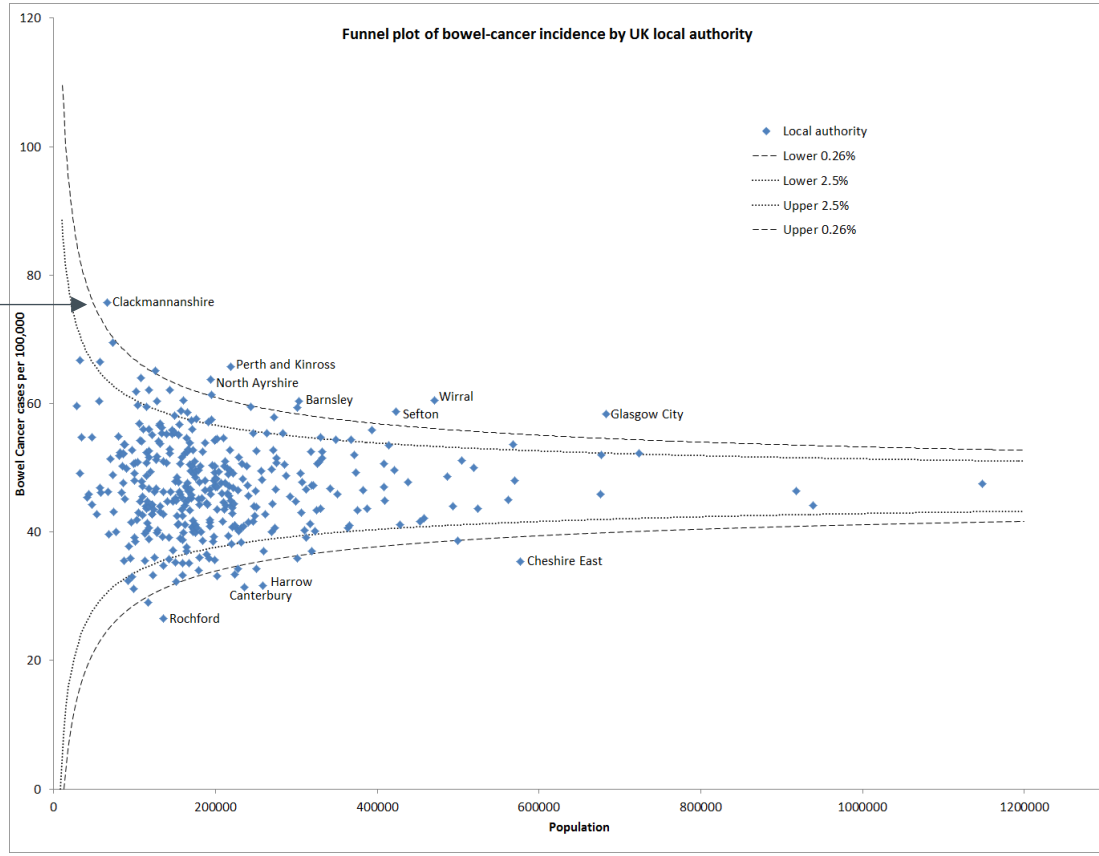
https://healthwatchlancashire.co.uk/
https://en.wikipedia.org/wiki/Clackmannanshire

정성규 ©

Clackmannanshire의 높은 대장암 발병률

- 큰 수의 법칙(의 반대):
  데이터 수가 작을수록 변동이 크다!!

Clackmannanshire 인구수: 5만명



Nothing's wrong with Clackmannanshire!

https://pb204.blogspot.com/2011/10/funnel-plot-of-uk-bowel-cancer.html

데이터 문해력: 데이터 분석의 이면

어떤 데이터를 보지 못하는가?

라고 질문하라

정성규 ©

한국의 위암 환자의
99%는 이 음식을
먹었다.
이 음식은 무엇일까?

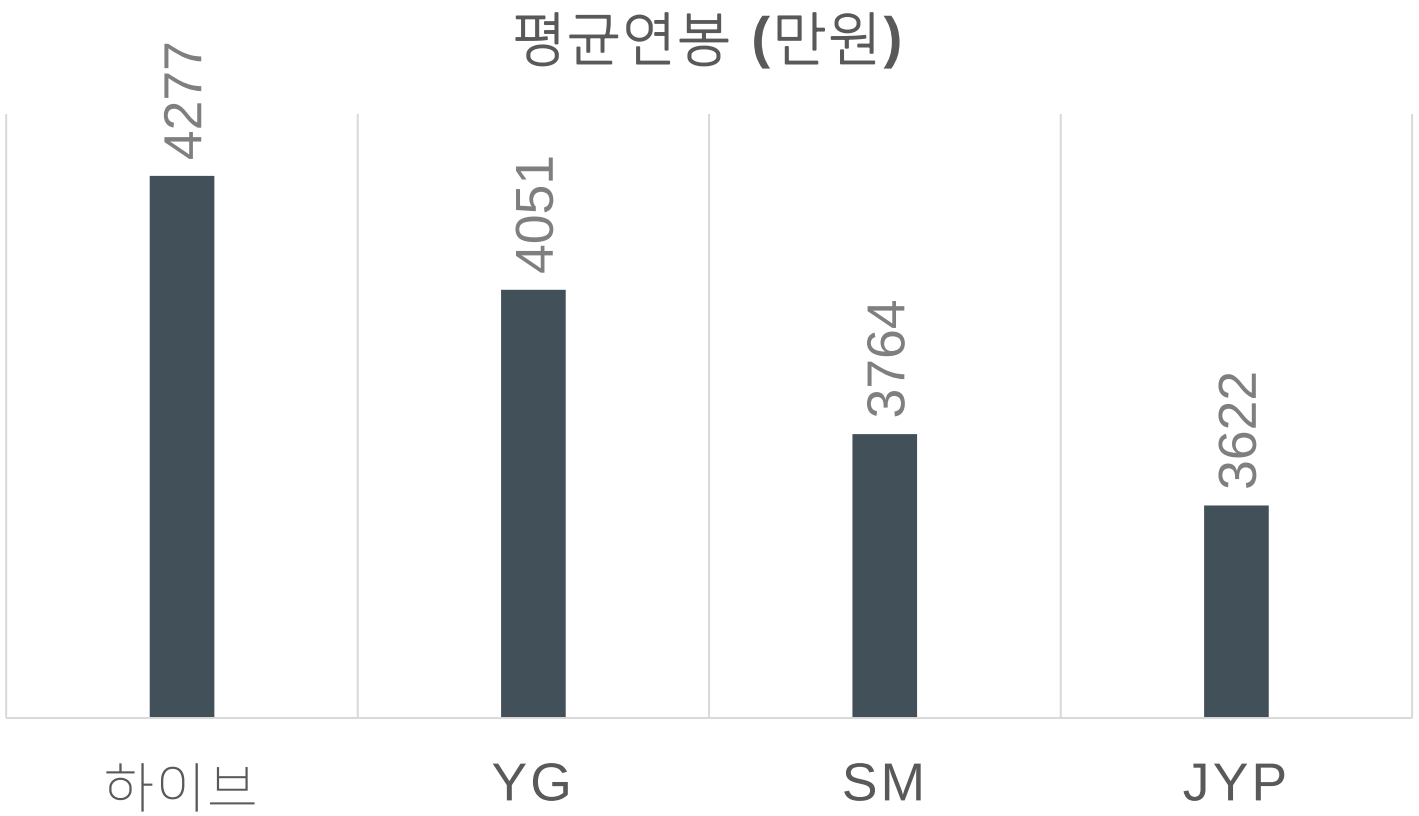
한국의 위암 환자의 99%는 이 음식을 먹었다.
이 음식은 무엇일까?

핵심 질문

- 보지 못한 데이터는 무엇일까?
- 위암을 앓지 않는 한국인의 몇 퍼센트가 이 음식을 먹었을까?

정성규 ©

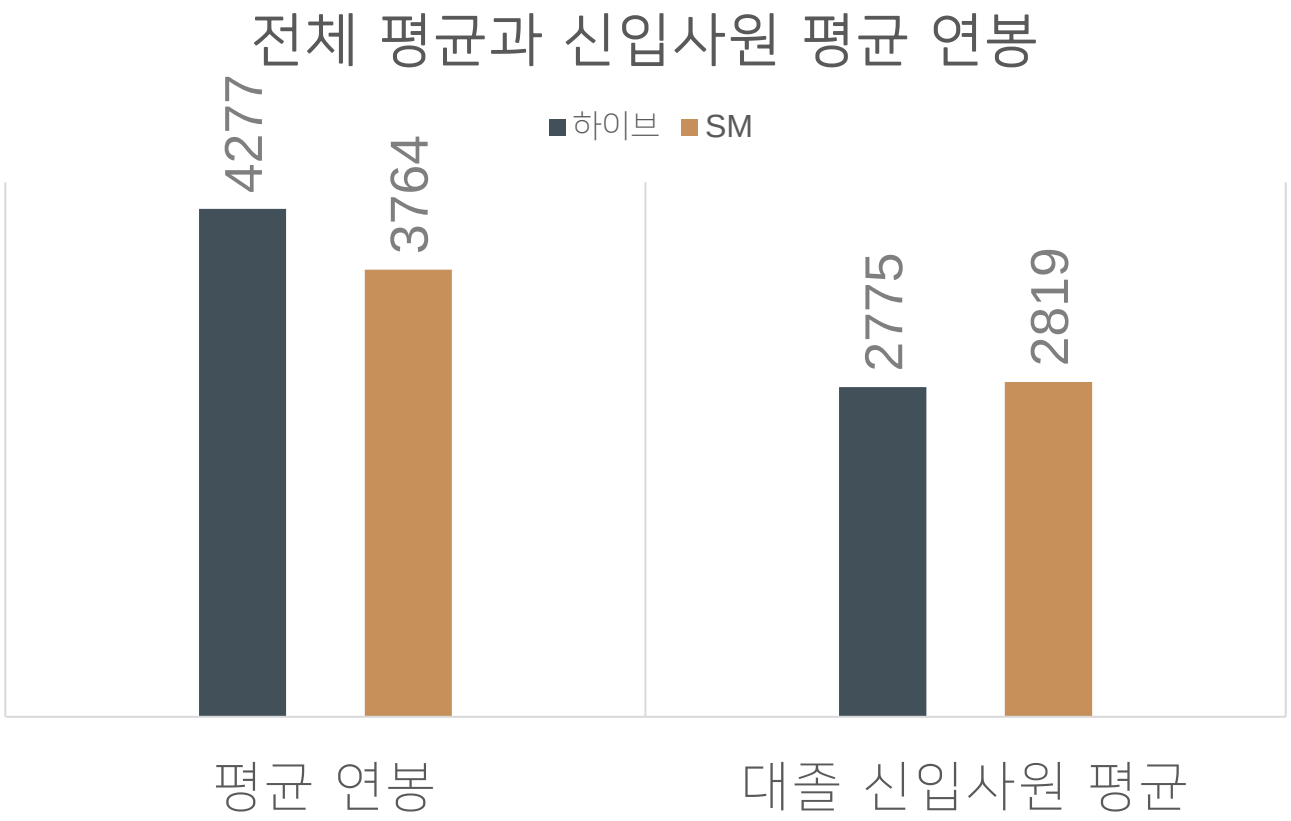
어느 직장에 취직해야 할까?



정성규 ©

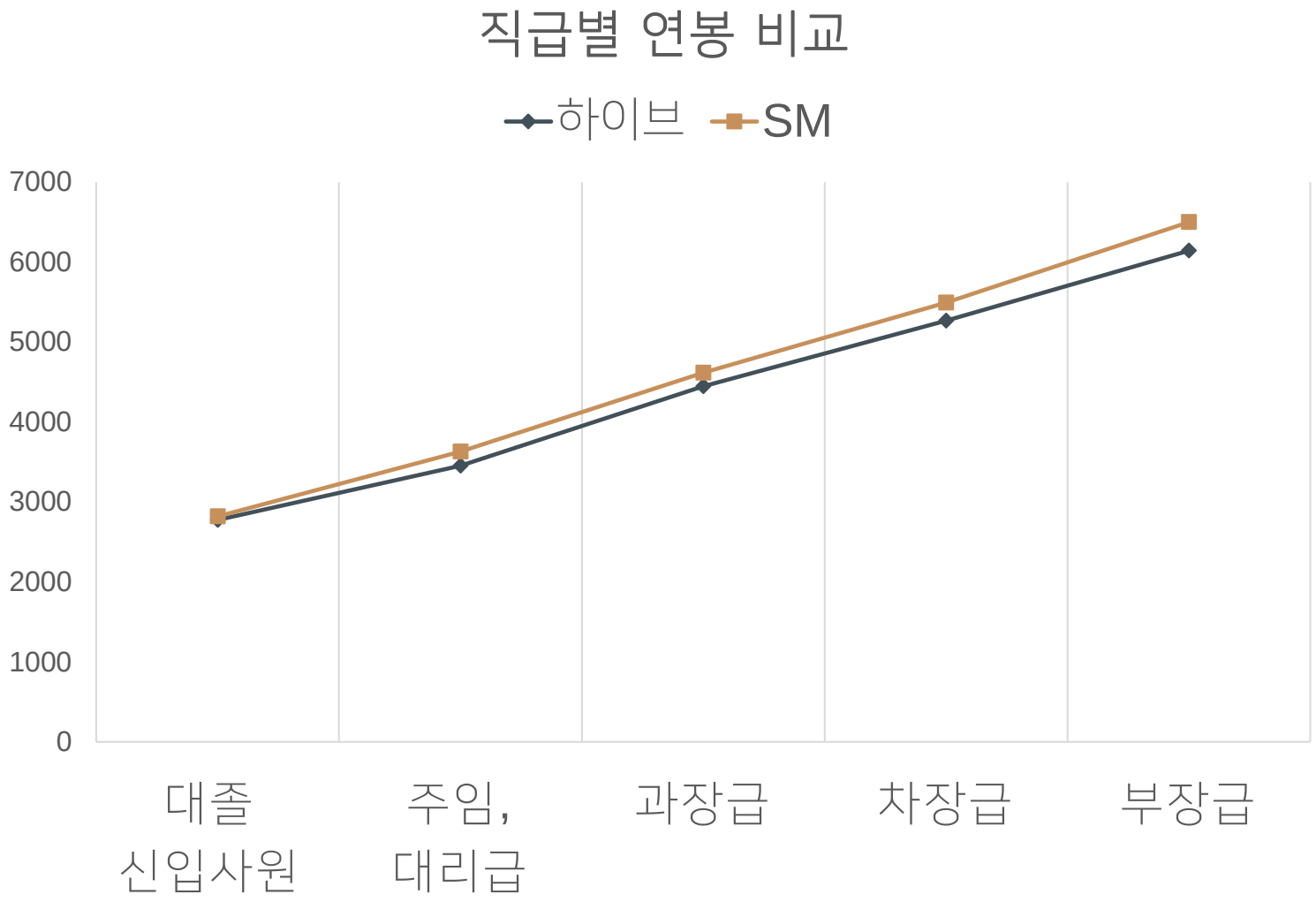
조선일보 김성민 “연예계 빅4 분석…연봉 1위는 ‘방시혁 회사’, 퇴사율 1위는?” 2019-01-18.

어느 직장에 취직해야 할까?



정성규 ©

어느 직장에 취직해야 할까?



정성규 ©

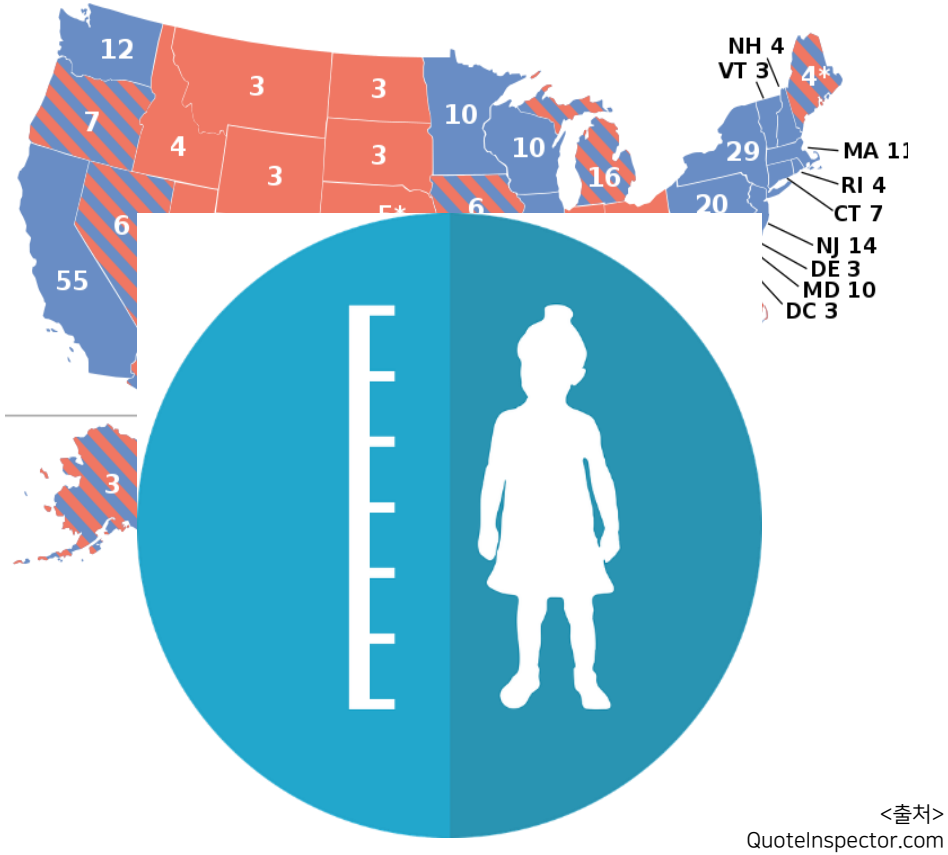
신뢰할 수 있는 예측

성공적인 예측을 위한 세 가지 원칙

정성규 ©

미래 예측

- 다음 주 복권 번호 예측
- 미래의 주가, 주가 지수 예측
- 다음 대통령 선거 결과의 예측
- 내 아이의 키는 얼마나 클 수 있을까?



<출처>
QuoteInspector.com
https://commons.wikimedia.org/wiki/General_election_polls_2016_Clinton_v_Trump.svg

정성규 ©

불가능한 예측과 가능한 예측

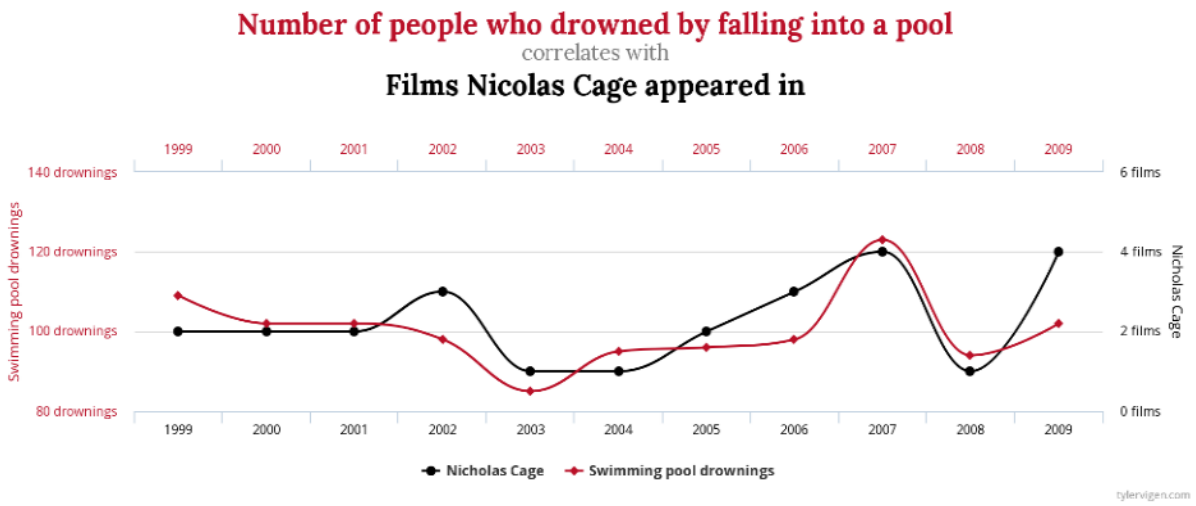
- 원칙 1: 예측의 기반이 되는 데이터가 있어야 한다

데이터가 아예 없는 예측 (불가능)

점쟁이? ~~예측~~ 예언

불가능한 예측과 가능한 예측

- 원칙 1: 예측의 기반이 되는 데이터가 있어야 한다



<출처> www.tylervigen.com

명백히 관련이 없는 데이터에 의한 예측 (불가능)

니콜라스 케이지 배우의 출연작 수로 수영장 익사 사고 예측?

불가능한 예측과 가능한 예측

- 원칙 1: 예측의 기반이 되는 데이터가 있어야 한다

기존의 복권 당첨 번호들이 데이터?

복권 당첨 결과들이 서로 “독립”.

다음 주 복권 당첨 번호는 무작위 (랜덤)
어떤 조합도 모두 같은 확률.

예측 불가능

불가능한 예측과 가능한 예측

- 원칙 2: 다른 사례로부터 배울 수 있어야 한다

부모의 키로 아이의 키를 예측 (175, 165) → ??

데이터: 다른 가족들의 사례

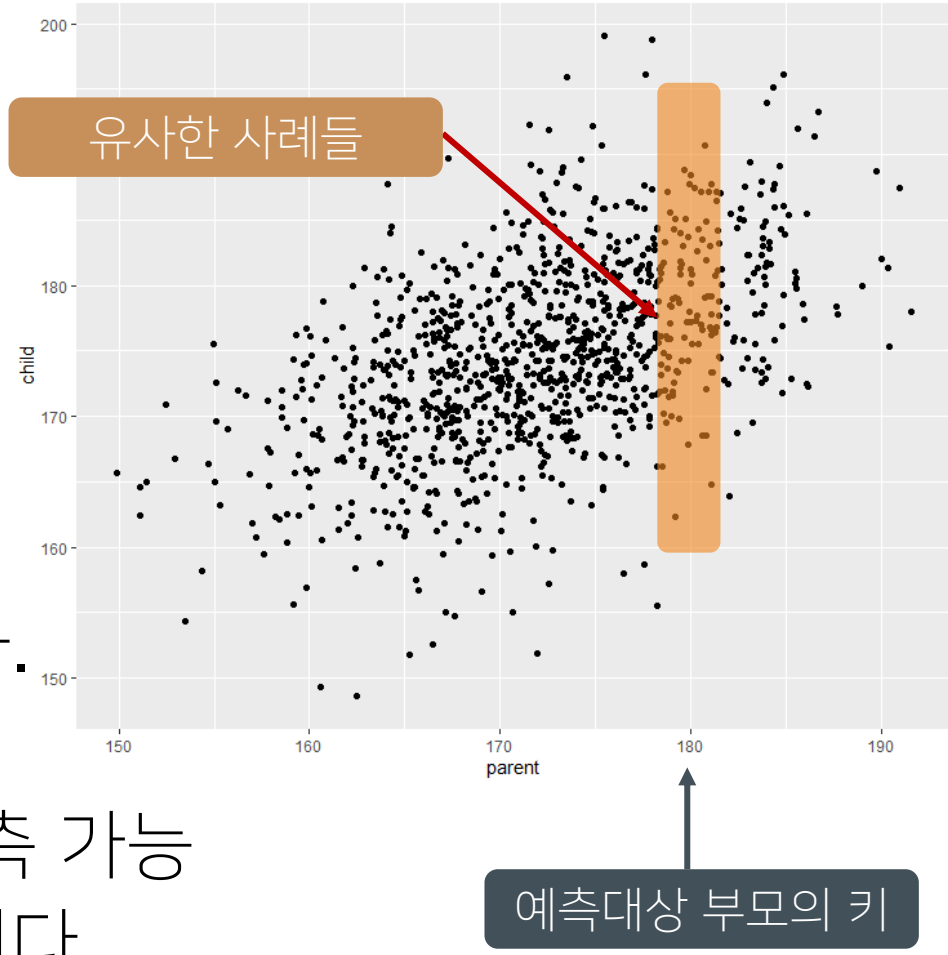
(170, 160) → 172
(177, 155) → 174
(165, 168) → 161
….

유사성의 가정 필요:
예측할 대상이 다른 사례들과 특별히 다르지 않다

정성규 ©

유사성

- 예측 대상 (내 가족)과 사례들 (데이터)이 모두 **같은 패턴**을 따르는 집단에서 **우연히** 뽑힌 한 값이다.
- 예측 대상, 사례들 모두 **같은 분포**를 따른다.
- 데이터로 패턴 (즉, 분포) 파악 가능
- 예측 대상과 유사한 사례들 존재하므로 예측 가능
- 예측의 두 번째 원칙은 곧, 유사성의 원칙이다.



- 한국 사람에 대한 예측을 할 때, 미국 사람 데이터에 기반?
- 한국의 미래에 대한 예측을 할 때, 1960년대의 사례에 기반?

정성규 ©

불가능한 예측과 가능한 예측

- 원칙 2: 다른 사례로부터 배울 수 있어야 한다

<출처> naver.com

주가 지수 예측 (과거부터 오늘까지의 주가) → 미래주가

데이터: 주가의 과거값들

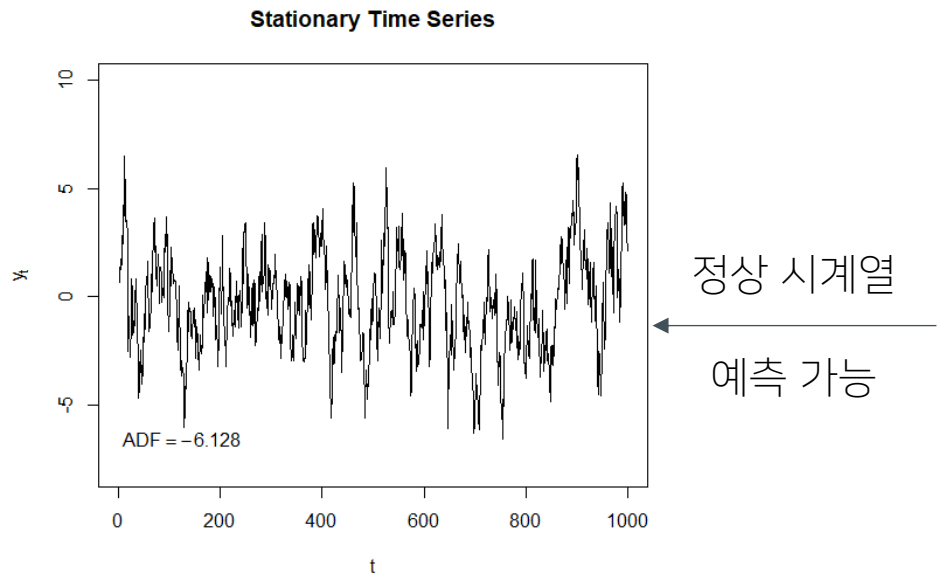
(어제까지의 주가) → 오늘주가
(이틀 전까지의 주가) → 어제주가
(삼일 전까지의 주가) → 이틀 전 주가
…

과연 유사성의 가정이 만족될까?

정성규 ©

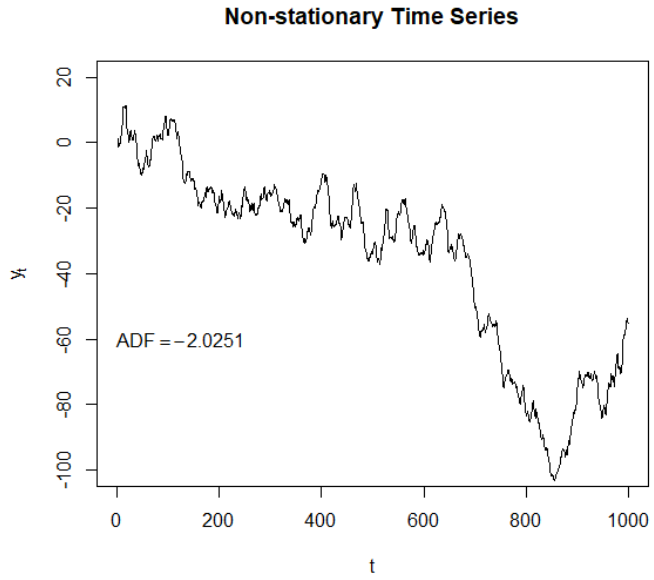
정상성 (stationarity)

- 시간의 흐름에 따라 관측된 데이터 (시계열)에서,
- 과거 값들의 분포와 미래 값들의 분포가 같다.
- 즉, 정상성이 만족되는 시계열 데이터는 사례들과 예측대상이 유사.



정상 시계열

예측 가능



비정상 시계열

예측 불가능

정상(正常): 특별한 변동이나 탈이 없이 제대로인 상태

정성규 ©

사회 · 경제 현상에 대한 예측

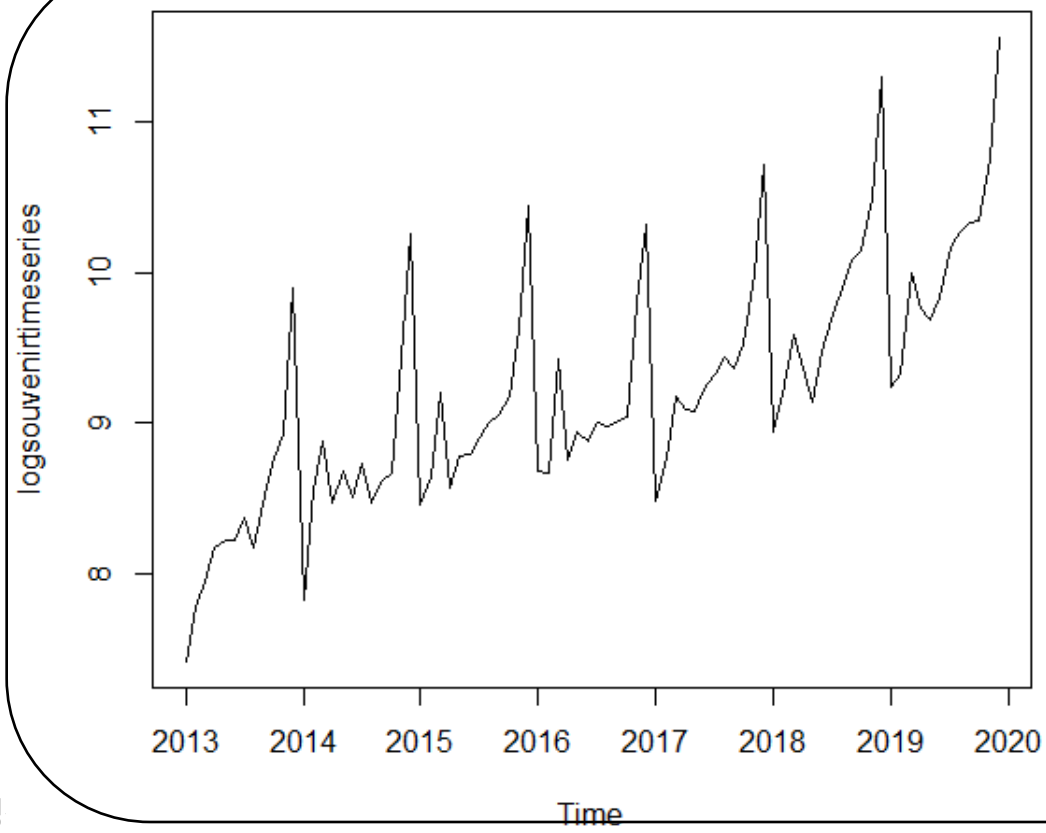
- 주가, 경제지표, 전염병 확산 등
- 대부분 정상성의 가정을 만족하지 않음
- 국가, 기관, 기업의 개입으로 “모형”이 바뀜

<출처> https://corona-live.com/

정성규 ©

불가능한 예측과 가능한 예측

- 원칙 3: 신호와 소음을 분리할 수 있어야 한다



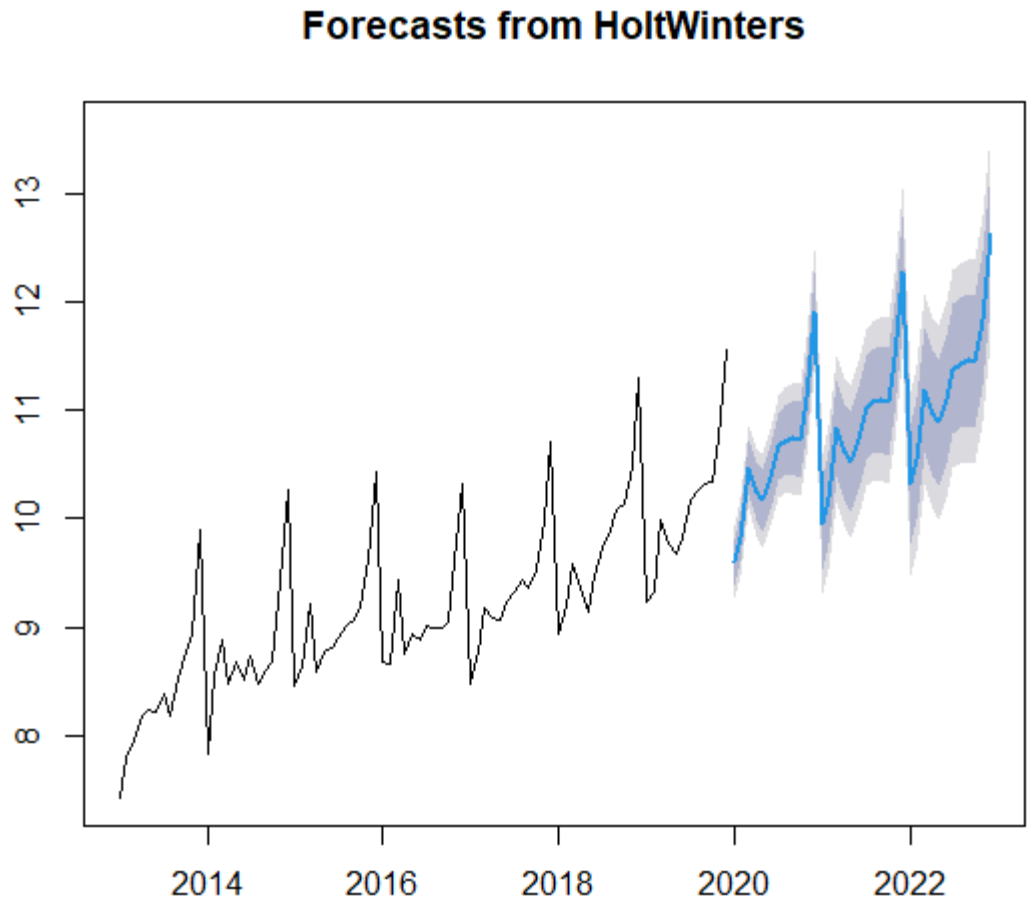
미래 판매량 예측

데이터: 판매량의 과거값들

비정상 시계열이지만, **결정적 추세** 존재

시계열 = 결정적 추세 (신호) + 소음

예측: 결정적 추세 + 소음으로 인한 오차



$$\text{판매량}(t) = \text{증가추세}_t + \text{계절추세}_t + \text{정상소음}_t$$

정성규 ©

신호와 소음

- 예측하려는 대상 y, 조건 x.
- 어떤 예측함수 f 에 대해, y = f(x) + 소음

=

신호

부모의 키
조부모의 키
운동량
….

+

소음

특정할 수 없는
모든 요소들

=

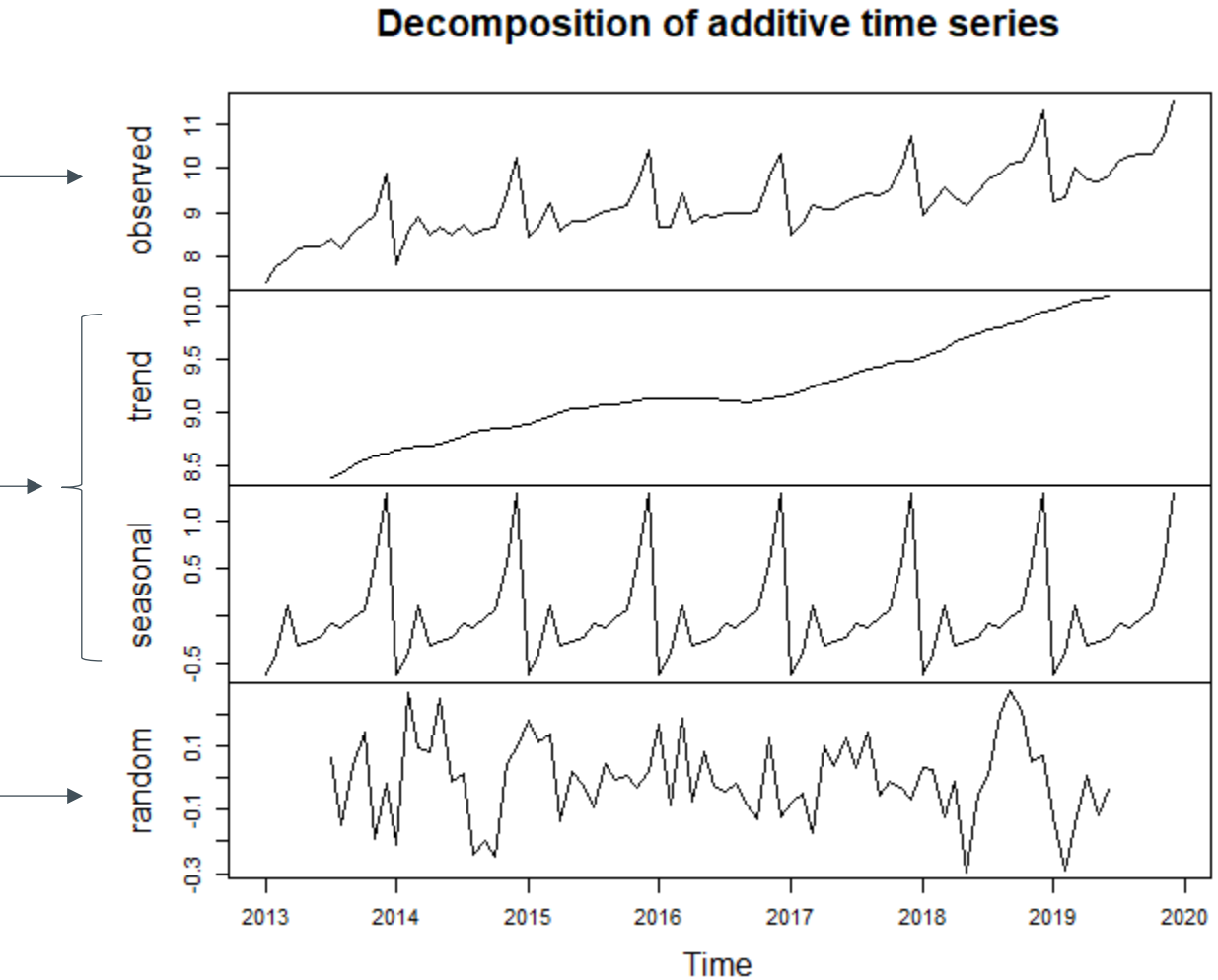
100% 소음

신호와 소음

관측값 (예측 대상)

신호

소음



정성규 ©

왜 예측이 어려운가?

- 원칙 1: 예측하고자 하는 대상과 관련 있는 데이터를 찾기 어렵기 때문

- 원칙 2: 특히 미래의 예측은 시간이 흐름에 따라 시스템 자체가 변하기 때문에 정상성의 가정을 만족하지 않을 가능성이 매우 크다. 즉, "다른 사례"들이 예측 대상과 밀접하게 관계되어 있지 않다.

- 원칙 3: 두 원칙이 지켜지더라도, "소음"의 비중이 크다면 정확한 예측은 어렵다.

정성규 ©

좋은 예측을 위한 제언

원칙 1. 데이터 기반 예측을 하라

- 관찰하지 못한 현상에 대한 예측 (forecast)은 어렵다
- (반복적으로) 관찰 가능한 현상 또는 근미래에 대한 예측 (prediction)은 가능
- 내일에 대한 예측은 오늘의 데이터를 이용

"날마다 새로운 예측을 하라"

정성규 ©

좋은 예측을 위한 제언

원칙 1. 데이터 기반 예측을 하라

원칙 2. 알고리즘은 예측이 아니다

- 알고리즘은 예측모형 구축을 위한 여러 선택지 중 하나일 뿐
- 상황, 문제, 데이터마다 적당한 예측모형이 다르다
- 좋은 알고리즘을 고르기 위해서는 데이터가 대변하는 “모집단”에 대한 통계적 이해 필요
  - 어떤 알고리즘이 당장 좋은 예측을 하는 것처럼 보여도 그 이유를 설명할 수 없으면 결국 좋지 않은 예측으로 판명나는 경우가 많다
  - Garbage in, garbage out: 편향된 데이터를 이용하면 편향된 예측

정성규 ©

좋은 예측을 위한 제언

원칙 1. 데이터 기반 예측을 하라

원칙 2. 알고리즘은 예측이 아니다

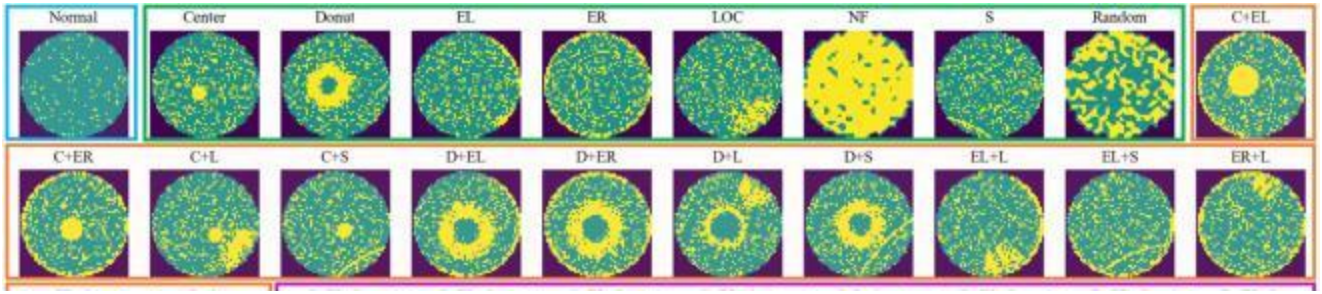
원칙 3. 확률적으로 생각하라

- 알고리즘의 결과 (예측값)은 언제나 불확실성을 내포한다
- 예측값 = “상승” 보다는 P(Y = “상승”) = 0.7
- 데이터의 “신호”와 “소음”을 파악하면 더욱 정확한 예측

정성규 ©

Can AI do better in prediction?

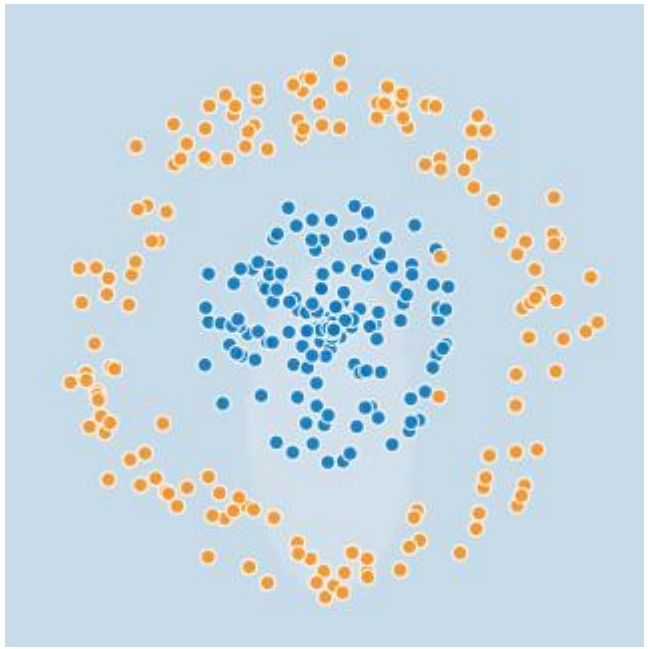
- What is AI?
- AI (or DNN-based learning) is successful with
  - massive dataset,
  - very flexible models,
  - for tasks with inherently very high signal-to-noise ratio.
- With DNN, tedious feature engineering is replaced by deeper networks (or by ingenious network designs)

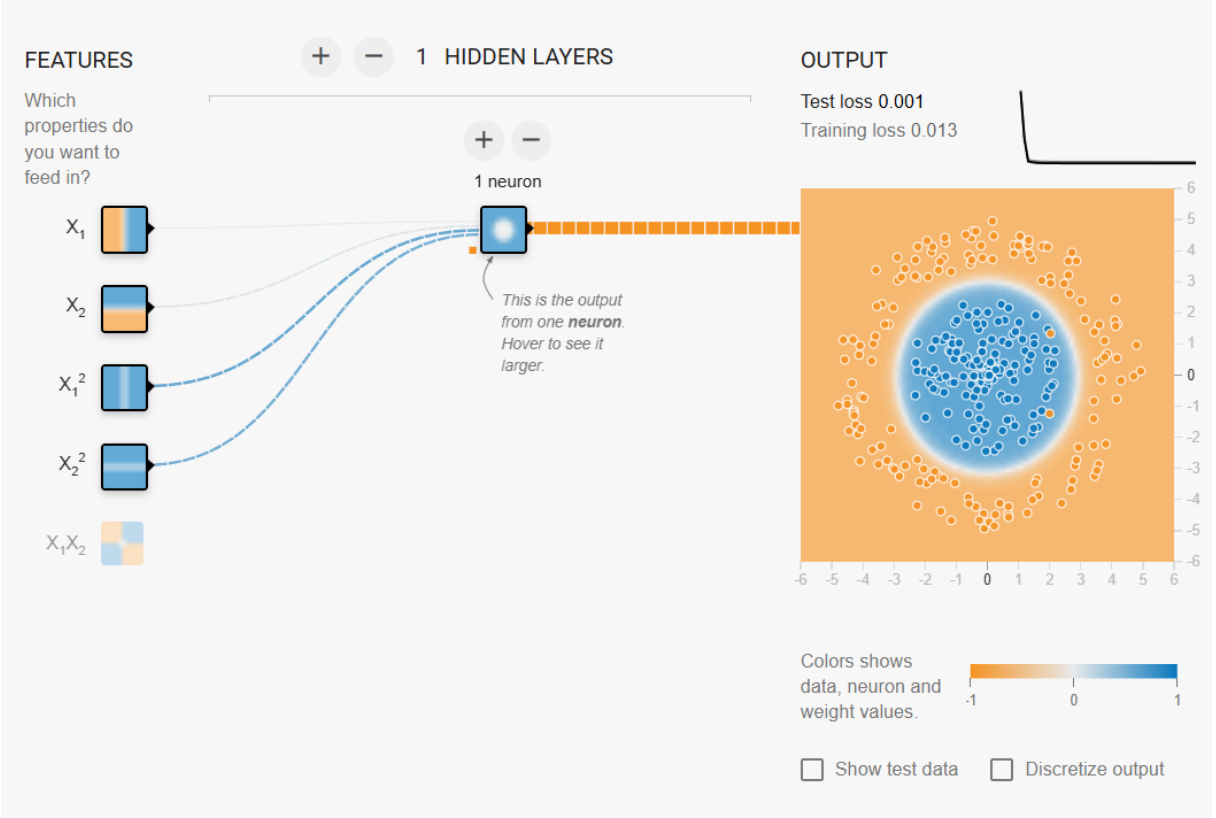


정성규 ©

Data with high signal-to-noise ratio

- A shallow network will do the job
- Other methods do the same (if one can find the definitive feature)

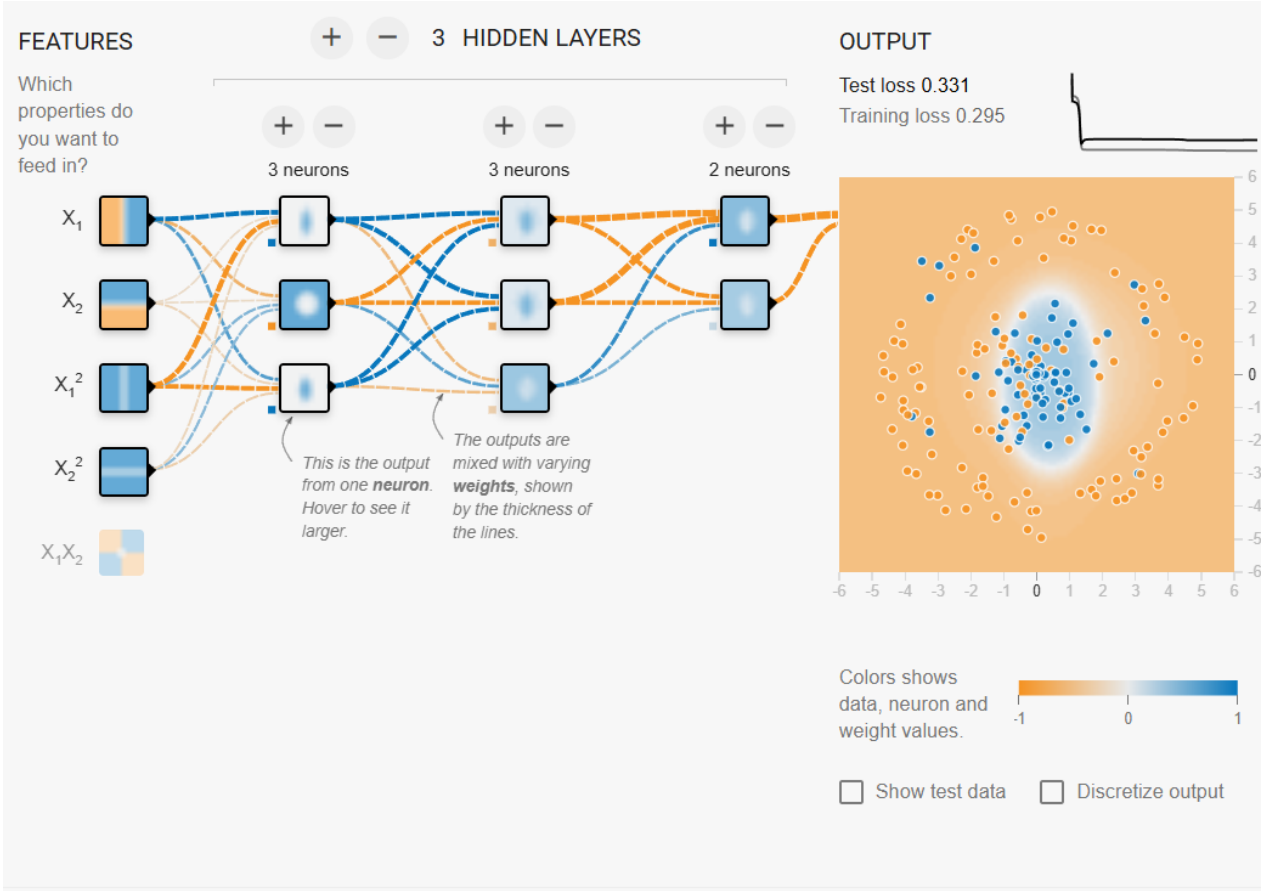




정성규 ©

Data with low signal-to-noise ratio

- DNN tends to fail with
  - Inherent large noise,
  - And lack of sample size.



정성규 ©

나아가며

- 데이터 = 신호 + 소음

  데이터 분석: 유의미한 신호를 추출하고 소음을 이해하여,

  그 속에서 패턴을 파악

- 데이터 문해력

  관측한 데이터를 해석하는 것을 넘어,

  보이지 않는 부분까지 인지하여, 분석의 한계를 이해하며

  전략적 통찰을 이끌어내는 능력

정성규 ©

End of Slide
감사합니다

정성규 ©