<https://looperxx.github.io/CS224n-2019-19-Safety%2C%20Bias%2C%20and%20Fairness/>

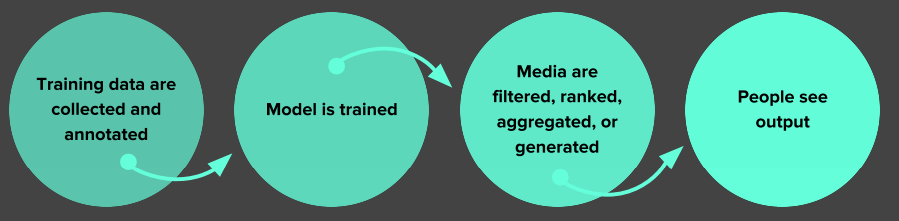
Prototype Theory: object에는 자신을 상징하는 stored central prototypical notion(prototype)이 있다. 이것을 잘 이용해보자..?

테스트의 피실험자들은 의사가 여자라는 또는 다른 가능성이 있을 수 있다는 것을 무시하는 경향이 있다.

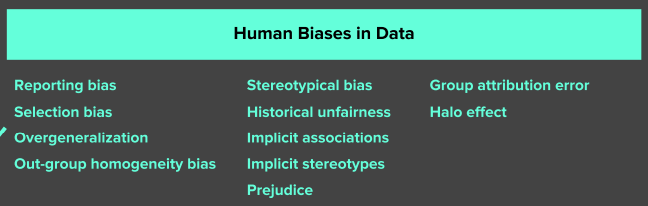
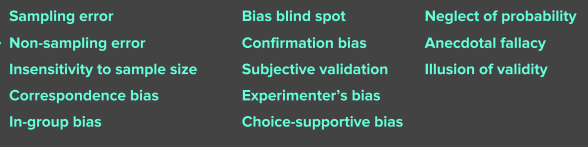
또한 frequency는 실제 세계의 반영과 크게 연관있는 특성이 아니다.

이런 경향들이 있는 실제 세계 vs 현실의 부족? 해야할 일 들이 있다.

또한 4가지 단계로 구성된 pipeline이 있는데



과 같다.

**Training data are collected and annotated**데이터를 모을 때부터 bias가 존재한다.(사람마다 생각이 다르니까 데이터도 다르다.)  
  
또한 annotation & collection 할 때도 bias가 존재한다.  
  
몇 가지 bias들을 소개하겠다.  
reporting bias: 사람들이 보고, 느끼고 공유하는 것은 사실 실제 frequency와는 다르다.  
selection bias: random 하게 샘플하는 것이 아닌 특정한 어느 곳에서 데이터를 고르려는 습성?  
out-group homogeneity bias: 4마리의 검은 고양이가 사실은 달라도 남들이 보기에는 다 똑같은 고양이로 보이는 그런 현상  
따라서 이런 bias 들 때문에 좋은 데이터셋을 구축하기 어려울 수 있다.  
또한 annotations도 그 annotator의 관점을 반영한 것이지 객관적인 worldview를 반영한 것이 아니라서 bias가 생길 수 있다.

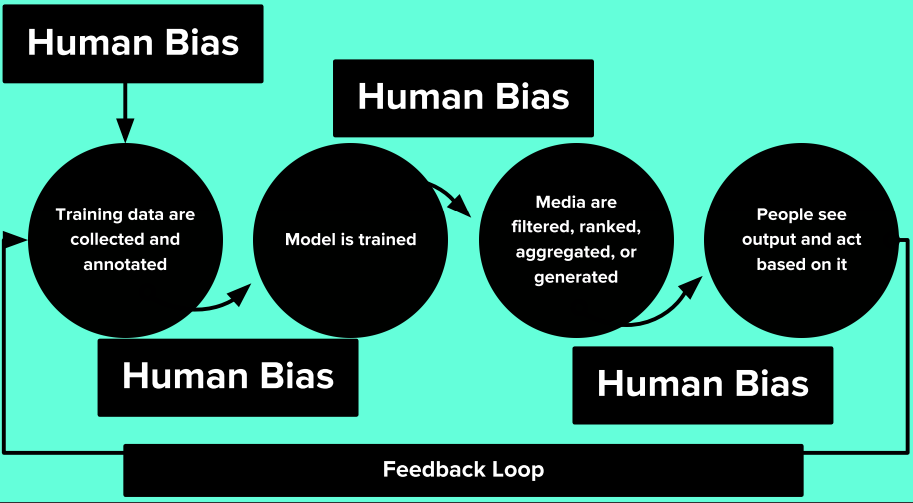
또한 데이터를 해석할 때 bias가 생기는데

Confirmation bias: 이전에 경험한 것, 경험한 사실에 맞게, 해석하고 조사하려는 경향

Overgeneralization: 너무 통상적인 사실만 언급하는 일반화(overfit의 반대일 듯)

Correlation fallacy: 원인과 결과의 관계가 알맞지 않는 경우가 있다.

Automation bias: 자동화된 기계의 의사 결정 시스템의 제안을 좋아하는 경향이 있다.



최종적으로 bias의 데이터 -> bias의 악순환 - > bias network effect

**그렇다면 bias는 always bad 인가?**

Bias can be good / bad / neutral 여러 상황에 따라 달라진다.

하지만 만약 너무 skewed data로 기계가 학습한다면 기계는 amplify injustice 할 것이다.

또한 Predicting policing & sentencing & criminality를 할 수 있다.

Predicting Homosexuality(gay using facial info)

여러 bias에 대한 경우의 예시를 설명했다.

Precision & Recall에 대하여 설명했다.(False Positive vs False Negative) 얘네는 항상 tradeoff가 존재하며 적절한 중재점을 찾아야 한다.

False Positive: 안 가려져도(땡이 레이블) 되는 부분을 blur-가림(정답으로 예측) – 안전빵 느낌

False Negative: 가려져야할(정답이 레이블) 부분이 안가려짐(틀리게 예측) – 도박 느낌

Ex) 첫번째는 precision을 높이려고 쓸데 없는 짓을 안하려는 것 / 두번째는 recall을 높이려고 쓸데 없는 짓을 하는 것(좋은 결과를 가져올 수도 있음.)

하지만 의외로 대개 False Negative를 택하는 것이 좋을 수 있다.

Conversation AI Toxicity