Bias in Data Sets

No Easy Solution

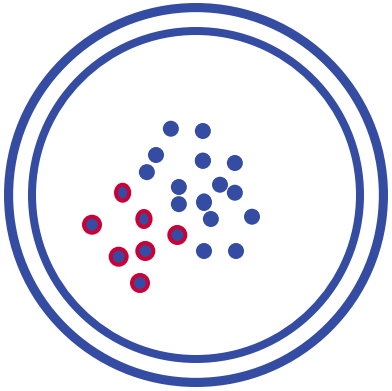
Overview

Bias isn't simply a cultural issue

- Mathematical formulations
  - Systematic failure of clf on cluster of data points
  - Hidden variable
- Examples
  - Race encoded in zip code
  - Gender bias hidden in language
- What is fair?
  - How can you even recognize errors?

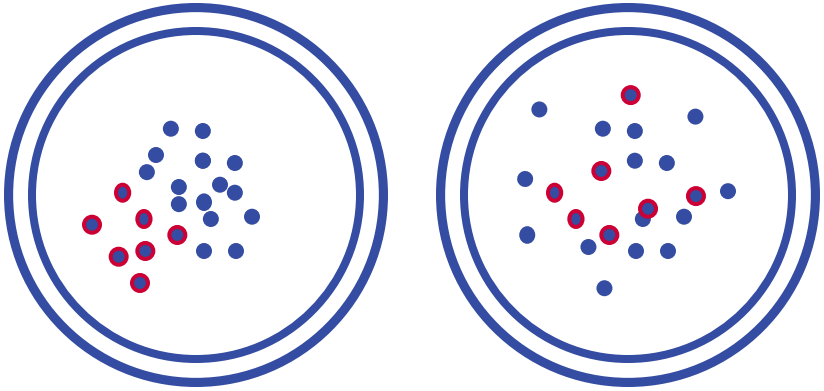
Not Just Culture

- Bias culturally is often spoken of as related to intent rather then effect.
- Lacking intent, trained models can only be evaluated based on how they preform.
- Fortunately, bias in algorithms is clearly defined. It means the algorithms consistently deliver results clustered around a result other than the optimal.
- In this case, we are thinking of when algorithms are biased on only subpopulation rather then the entire data set (see red below).



Mathematical Formulations

- Systematic failure of clf on cluster of data points.
- Currently, the most accepted way to manage bias is to acknowledge that classifiers should look to mitigate variance in accuracy (or other valuation metric) among subpopulations, even if this means decreasing overall accuracy for the entire population.



Hidden Variable

This definition highlights a problem in the naive approach, which is simply to exclude membership of the cluster. For instance, people have said their algorithm can't be racist/sexist because they didn't include race/sex as a variable.

This has two fundamental problems:

1. The hidden variable may be encoded in other seen variables.
2. It can make it harder to identify if the approach is problematic.

DataScience@SMU

Race Bias in Housing

Race Encoded in Zip Codes

- It's important to remember that many examples of biased data sets are a reflection of an active policy of discrimination.
- Racial geographic distribution of people in Chicago is a direct result of an overt racist policy for racial segregation.
- This means that geographical analysis serves as a particularly problematic way of gaining an unbiased view of the city.

DataScience@SMU

Gender Bias in Language Use

Sexism Encoded in Text

- Word2Vec does a very good job at mapping language usage and encoding the relationships it discovers.
- Vector for king, remove vector for man, replace it with woman and you get queen.
- Vector for doctor, remove man, replace it with woman and you get nurse.
- While there are not texts (that I know of) attempting to skew the use of the term doctor toward men and nurse toward women, there are very likely societal factors that do just that, and moreover, our use of language likely consciously and unconsciously supports these problems.
- Should we then continue to just attempt to have our algorithms fit to reality or fit to what we perceive reality should be?

DataScience@SMU

One Formulation of Fair

What Is Fair?

The complexity of mathematically deciding on an appropriate way to both fit reality and exclude bias is a daunting task.

Fortunately, we know there is a fantastic step that can successfully mitigate bias in data sets.

1. Direct involvement of under represented populations

University of Washington found a fantastic solution to under-representation of women in programming. They involved more women. Unsurprisingly, when stakeholders of historically disenfranchised communities are involved in the process of algorithm generation, they can identify and deconstruct problems that those who are outside their communities may not be able to do.

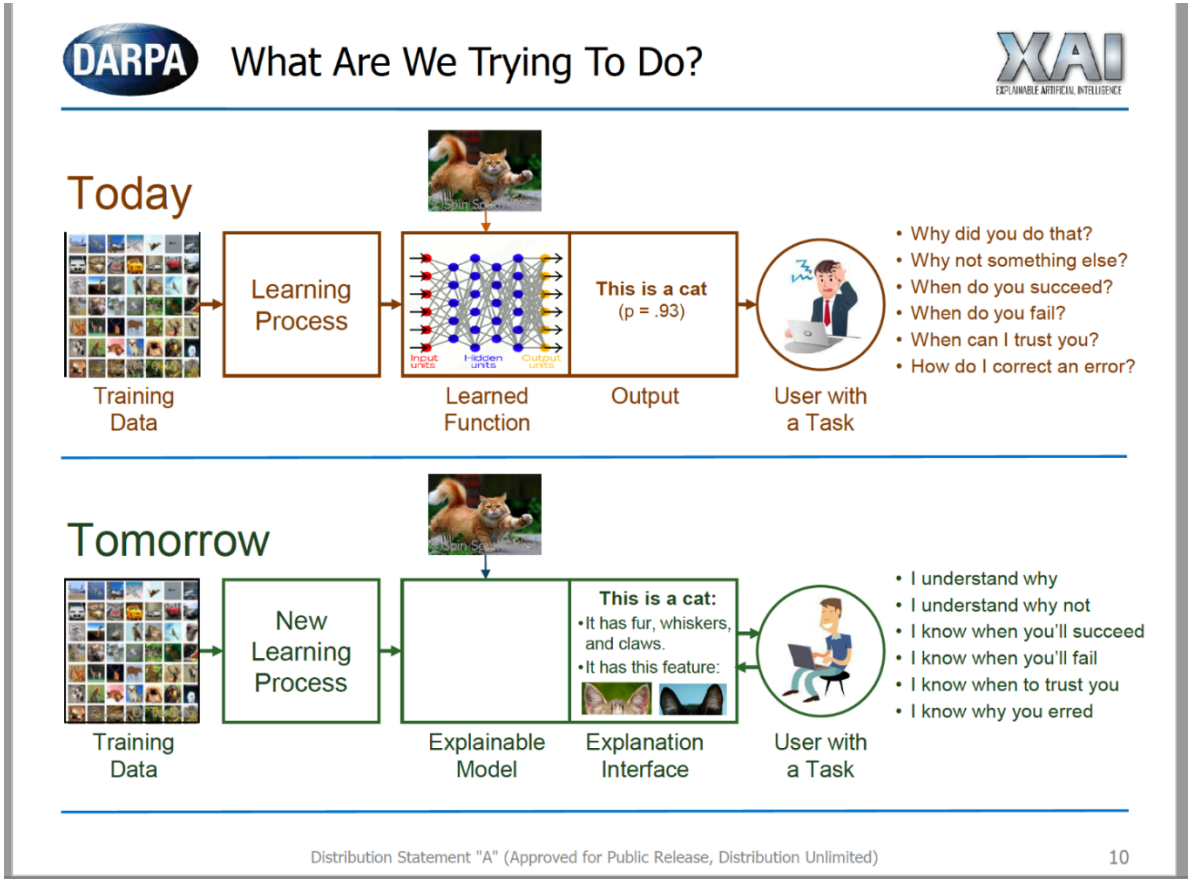
Interpretability

One way of detecting (and countering bias) is developing interpretable models.

What is an interpretable/explainable model?

- There isn't a unified technical definition of interpretability. Several related questions about *black box* models usually arise: Do we understand how the model works? Do we understand the role of each parameter? Is the model complexity such that it can to be examined by a human?
- A nice reference on the topic: *The Mythos of Model Interpretability*, Z. Lipton (https://arxiv.org/pdf/1606.03490.pdf)

DARPA's XAI Program



A slide from DARPA's Explainable AI (XAI) program:
https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_184754.pdf

DataScience@SMU