Discrimination

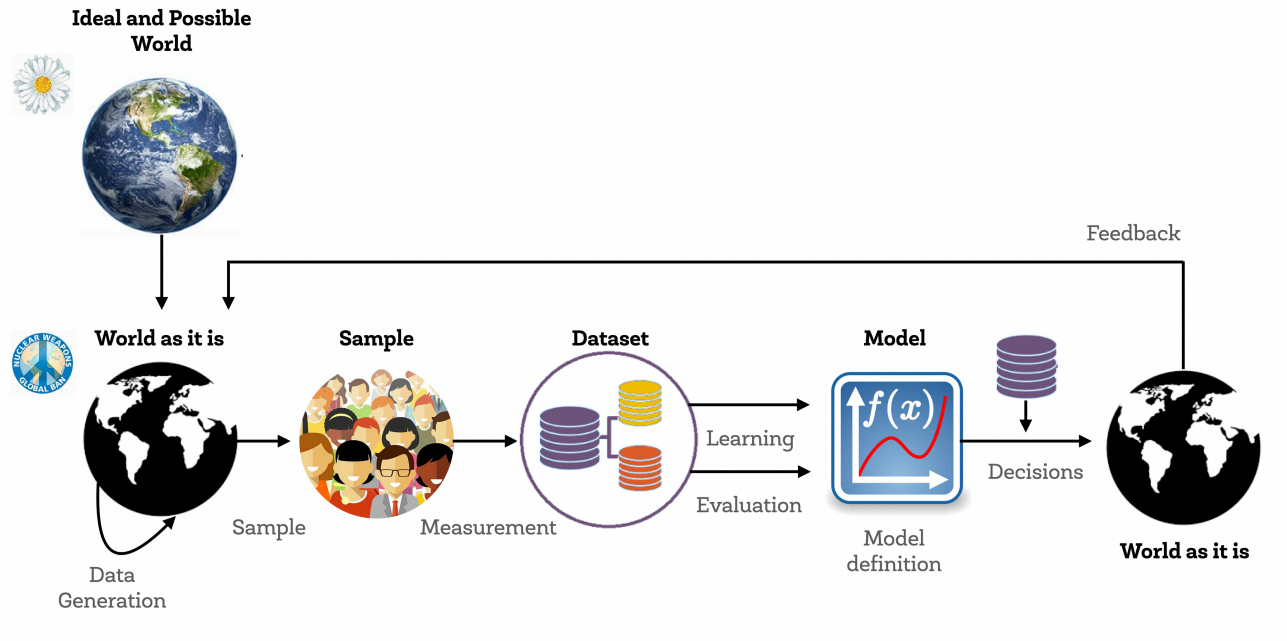
Exercise 1: ML Loop

A 2016 investigation found a racial disparity in Amazon's same-day delivery service: in many cities, Black residents were about half as likely as White residents to live in a ZIP code where the service was offered.

Consider a hypothetical machine learning system that uses data available to Amazon to determine which ZIP codes would be profitable targets for same-day delivery service. Since the rollout happens gradually, data from already-serviced neighborhoods helps inform the decision of where to expand.

Give three distinct reasons why racial disparities might arise in the predictions of such a system. Place each of these reasons in one (or more) of the stages of the machine learning loop: real world, measurement, learning, action, and feedback.

Feel free to read about the investigation here, but the question asks you to imagine a hypothetical machine learning system and so the actual details of the investigation are not relevant. The reasons you list must be plausible but don't have to match what actually happened.



Exercise 2: The naive statistician

Google's image generation algorithm failure which resulted in the inability to create images of Caucasians is an example of an **unintended consequence caused by how the algorithm was tuned**[1].

This is the content of a threat we could read in X, written by an statistician, about how to solve this issue (in an apparently obvious way):

"We basically want our AI image generator to roughly match the real world proportion of people of various races after factoring in the constraints of the user prompt.

This is *slightly* challenging because the information about what the racial composition of various populations and subpopulations is just isn't going to be in the image training data (nor should it be).

That's why we need to build a separate module that estimates the proportion of people of each race in whatever subgroup is mentioned in the prompt.

To do this, we need to get accurate global and historical demographic data and put it in a separate database. This task shouldn't be that hard if we have the resources of a billion dollar company. We can next use the data to make our estimates.

Let's say we get a prompt for an image of a US doctor. Our racial composition module estimates that US doctors are maybe 10% black. Then with 10% probability, we add the term "African American" to the user prompt.

The user will then receive an image of a black doctor with 10% probability. That's it. I think it's really that easy.

We don't have to do this perfectly, just well enough that the proportions of each race don't look wildly off to the average user."

This approach shows a lot of "statistical common sense" but there are significant shortcomings. Could you identify them? The right thing to do is a very complex question, both philosophically and practically...

Write short report (max 1 page) defining the problems you identify and a possible solution.

Hints:

- Which features should we consider? Is this selection generalizable to all countries?
- Demographics can be measured at different geographical levels. Which is the relationship between demographic measures and the content of the query?
- What about asking for a historical picture?
- How do we treat fact versus fiction (user intent, expected use of the image, etc.)? If someone wants to have a Black queen in Bridgerton, why not? If that same image is going to illustrate a news article, maybe not.
- Etc.