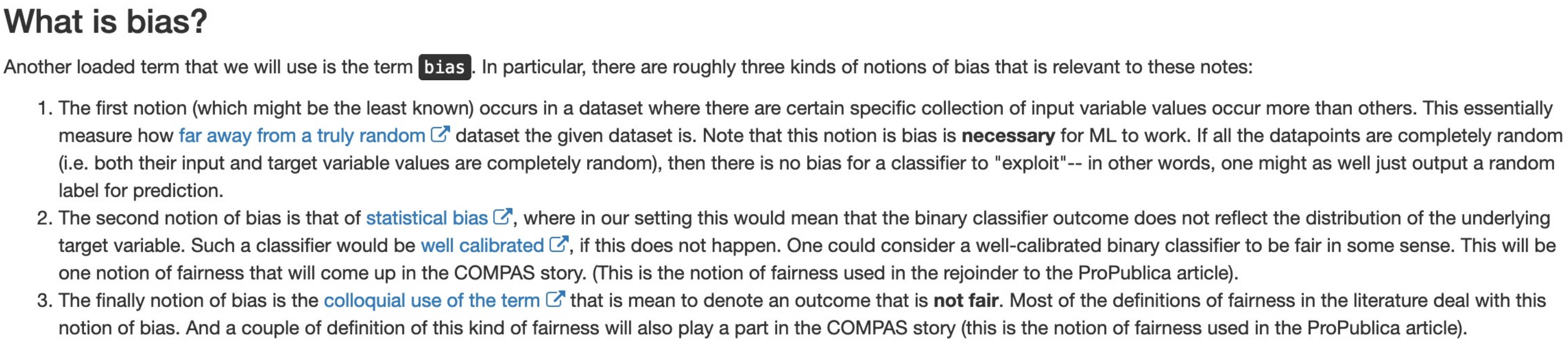
ML and Society

Apr 19, 2022

What is bias?

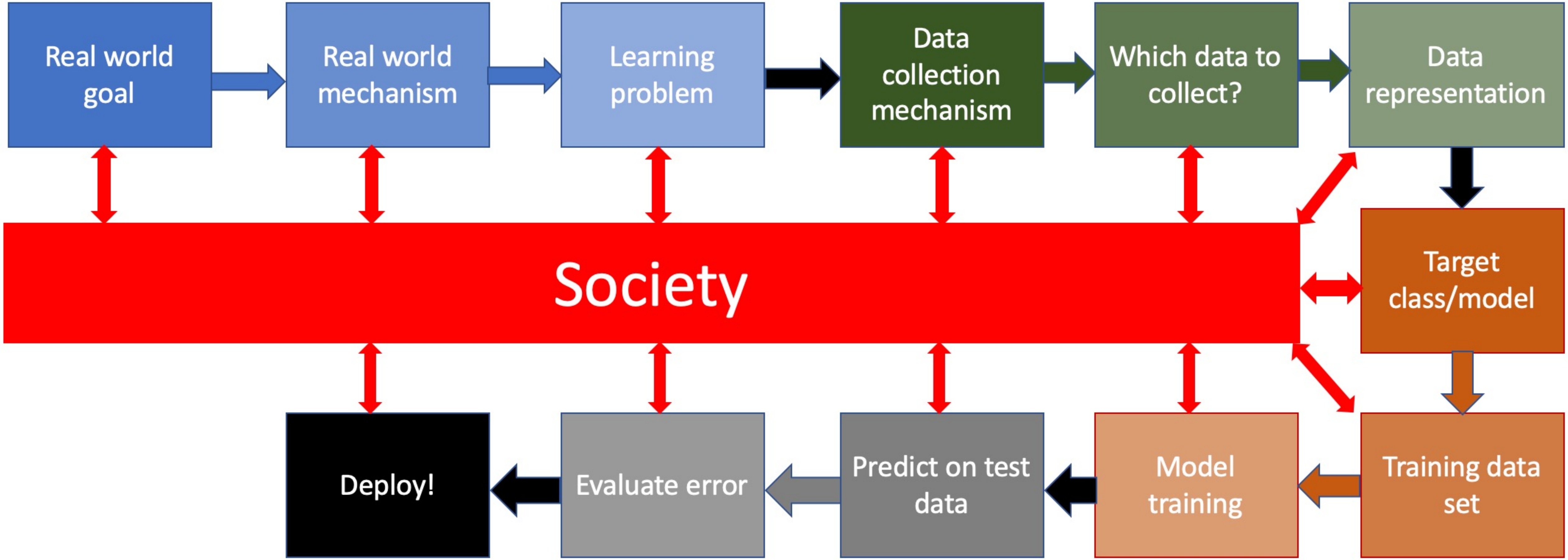


What is bias?

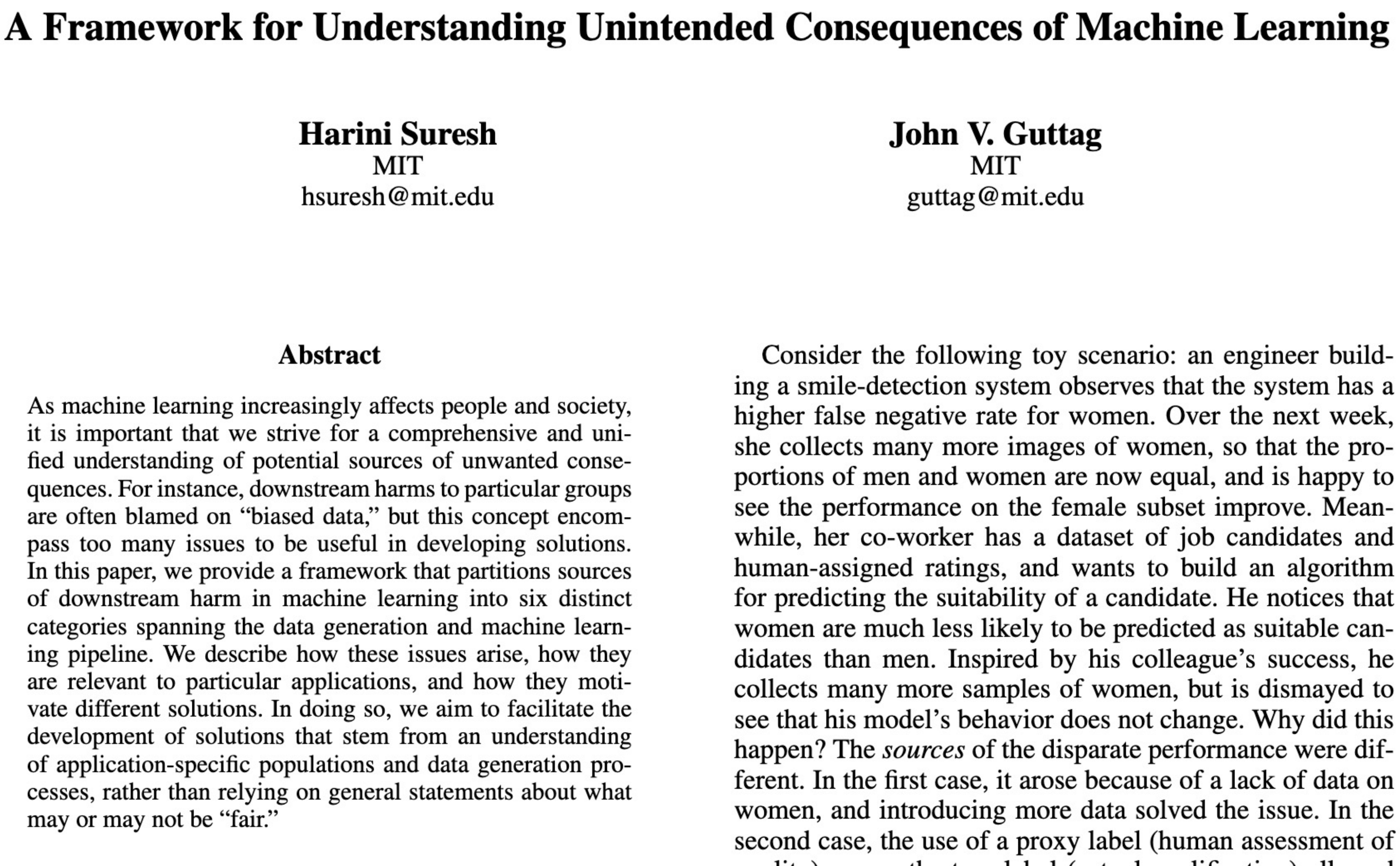
Another loaded term that we will use is the term `bias`. In particular, there are roughly three kinds of notions of bias that is relevant to these notes:

1. The first notion (which might be the least known) occurs in a dataset where there are certain specific collection of input variable values occur more than others. This essentially measure how far away from a truly random dataset the given dataset is. Note that this notion is bias is **necessary** for ML to work. If all the datapoints are completely random (i.e. both their input and target variable values are completely random), then there is no bias for a classifier to "exploit"-- in other words, one might as well just output a random label for prediction.
2. The second notion of bias is that of statistical bias, where in our setting this would mean that the binary classifier outcome does not reflect the distribution of the underlying target variable. Such a classifier would be well calibrated, if this does not happen. One could consider a well-calibrated binary classifier to be fair in some sense. This will be one notion of fairness that will come up in the COMPAS story. (This is the notion of fairness used in the rejoinder to the ProPublica article).
3. The finally notion of bias is the colloquial use of the term that is mean to denote an outcome that is **not fair**. Most of the definitions of fairness in the literature deal with this notion of bias. And a couple of definition of this kind of fairness will also play a part in the COMPAS story (this is the notion of fairness used in the ProPublica article).

Bringing society back into the picture



Six categories of bias of the 3rd kind



A Framework for Understanding Unintended Consequences of Machine Learning

Harini Suresh
MIT
hsuresh@mit.edu

John V. Guttag
MIT
guttag@mit.edu

Abstract

As machine learning increasingly affects people and society, it is important that we strive for a comprehensive and unified understanding of potential sources of unwanted consequences. For instance, downstream harms to particular groups are often blamed on "biased data," but this concept encompass too many issues to be useful in developing solutions. In this paper, we provide a framework that partitions sources of downstream harm in machine learning into six distinct categories spanning the data generation and machine learning pipeline. We describe how these issues arise, how they are relevant to particular applications, and how they motivate different solutions. In doing so, we aim to facilitate the development of solutions that stem from an understanding of application-specific populations and data generation processes, rather than relying on general statements about what may or may not be "fair."

Consider the following toy scenario: an engineer building a smile-detection system observes that the system has a higher false negative rate for women. Over the next week, she collects many more images of women, so that the proportions of men and women are now equal, and is happy to see the performance on the female subset improve. Meanwhile, her co-worker has a dataset of job candidates and human-assigned ratings, and wants to build an algorithm for predicting the suitability of a candidate. He notices that women are much less likely to be predicted as suitable candidates than men. Inspired by his colleague's success, he collects many more samples of women, but is dismayed to see that his model's behavior does not change. Why did this happen? The *sources* of the disparate performance were different. In the first case, it arose because of a lack of data on women, and introducing more data solved the issue. In the second case, the use of a proxy label (human assessment of

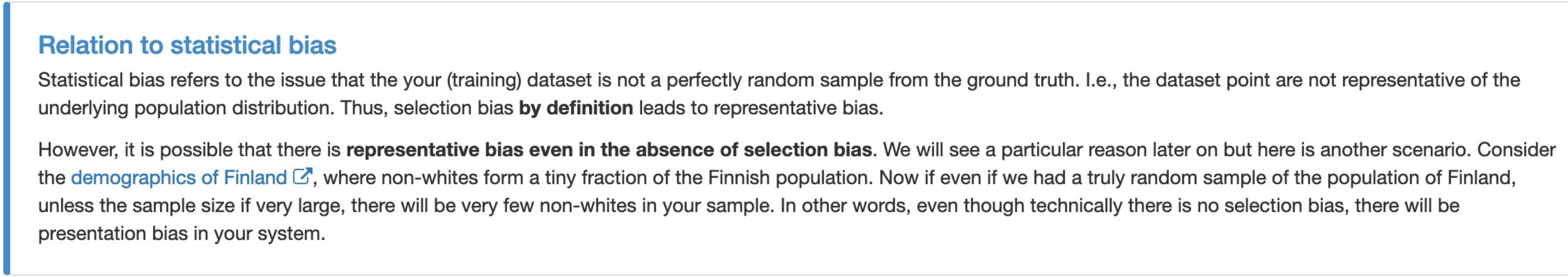
Historical Bias

Bias ingrained in society

Cannot be avoided even with **perfect sampling** of the population

Representation bias

Certain section(s) of population excluded in your dataset



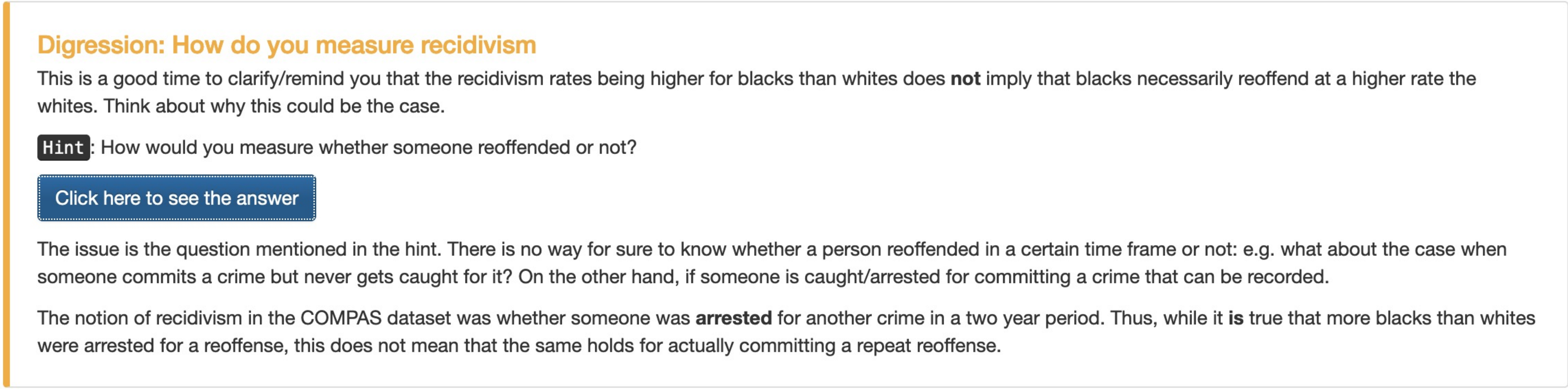
Relation to statistical bias

Statistical bias refers to the issue that the your (training) dataset is not a perfectly random sample from the ground truth. I.e., the dataset point are not representative of the underlying population distribution. Thus, selection bias **by definition** leads to representative bias.

However, it is possible that there is **representative bias even in the absence of selection bias**. We will see a particular reason later on but here is another scenario. Consider the demographics of Finland, where non-whites form a tiny fraction of the Finnish population. Now if even if we had a truly random sample of the population of Finland, unless the sample size if very large, there will be very few non-whites in your sample. In other words, even though technically there is no selection bias, there will be presentation bias in your system.

Measurement bias

Using a proxy variable instead of the “real” variable



Digression: How do you measure recidivism

This is a good time to clarify/remind you that the recidivism rates being higher for blacks than whites does **not** imply that blacks necessarily reoffend at a higher rate the whites. Think about why this could be the case.

`Hint`: How would you measure whether someone reoffended or not?

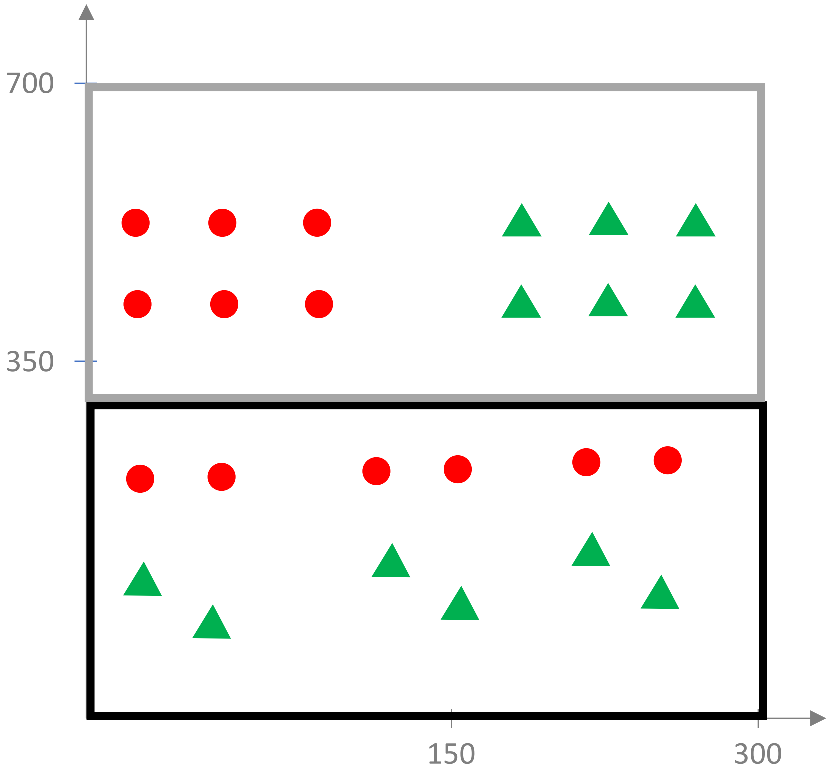
The issue is the question mentioned in the hint. There is no way for sure to know whether a person reoffended in a certain time frame or not: e.g. what about the case when someone commits a crime but never gets caught for it? On the other hand, if someone is caught/arrested for committing a crime that can be recorded.

The notion of recidivism in the COMPAS dataset was whether someone was **arrested** for another crime in a two year period. Thus, while it **is** true that more blacks than whites were arrested for a reoffense, this does not mean that the same holds for actually committing a repeat reoffense.

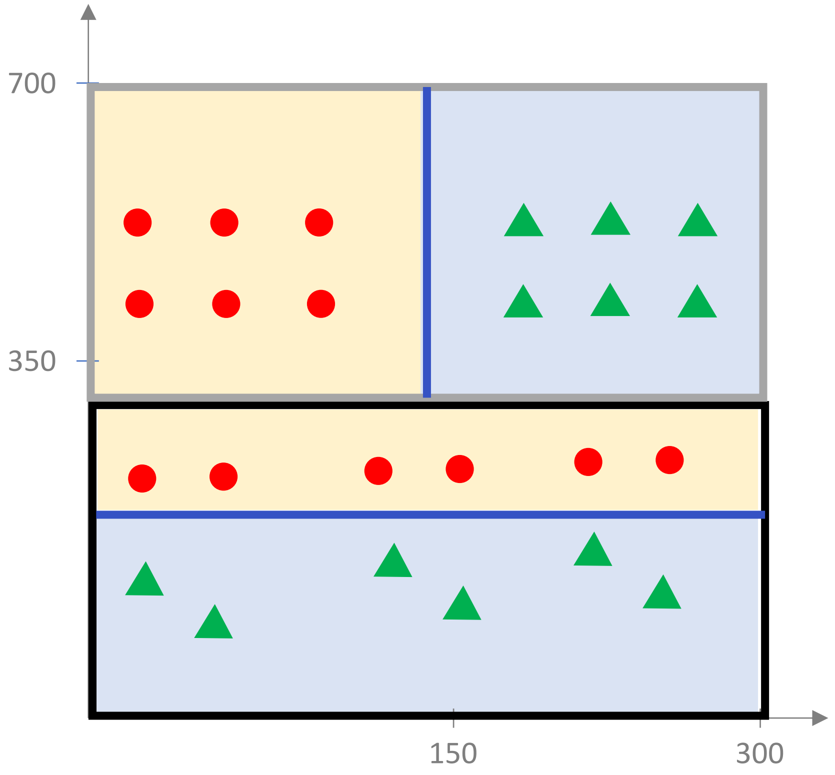
Aggregation bias

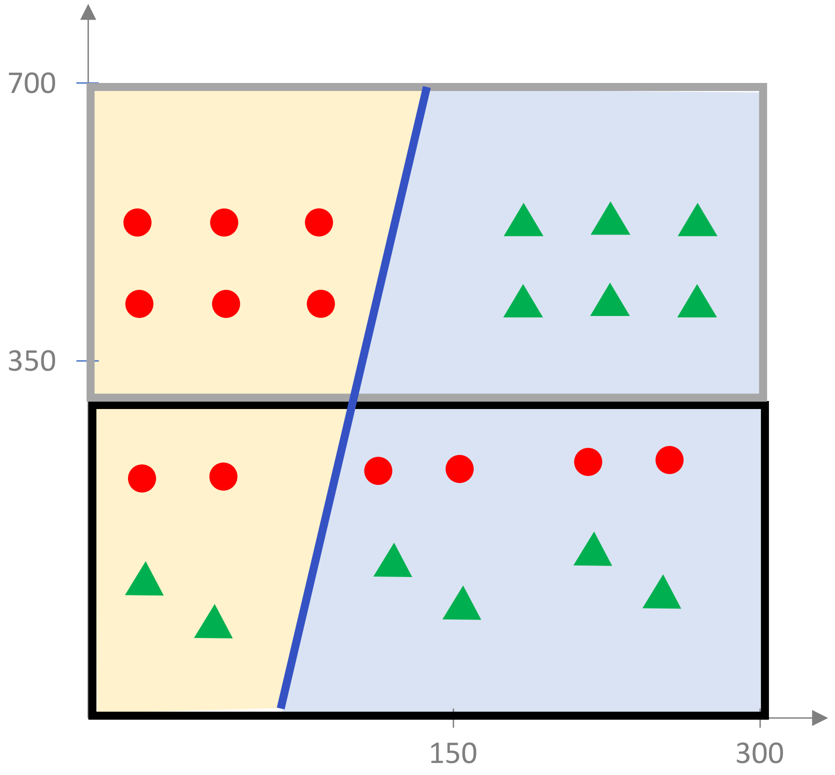
Use a single model where prevalence changes based on groups

Consider an example

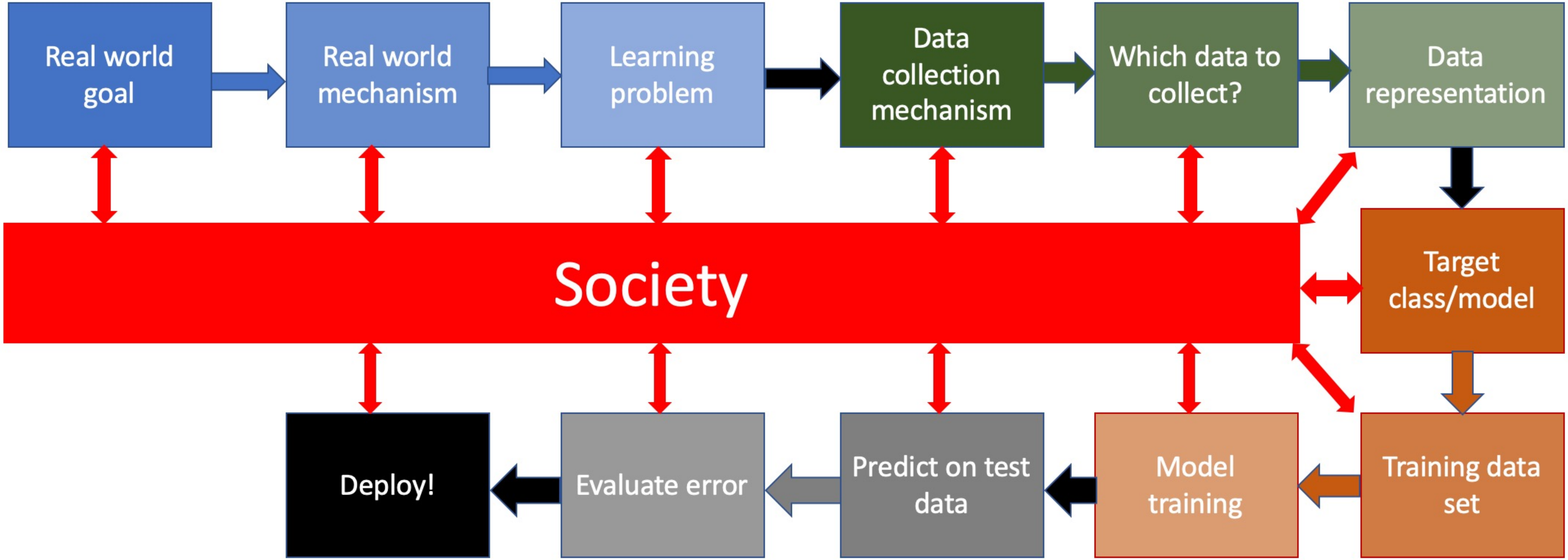


Two linear models vs. one linear model

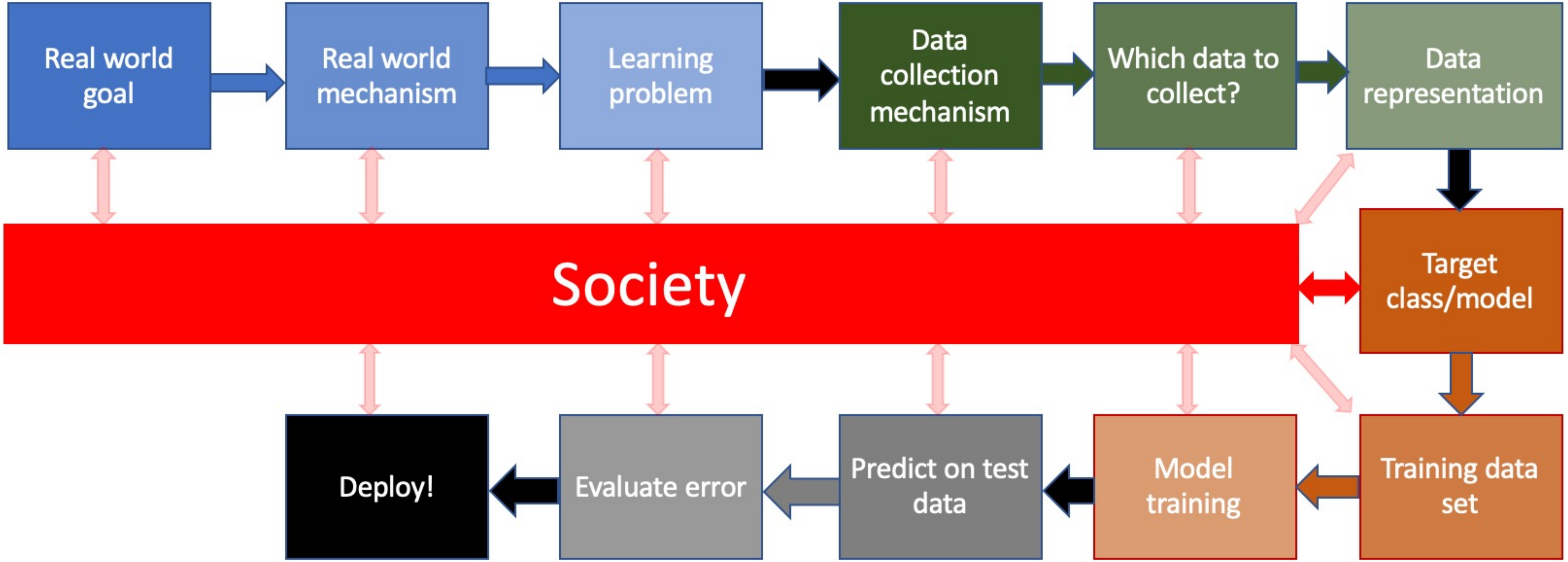




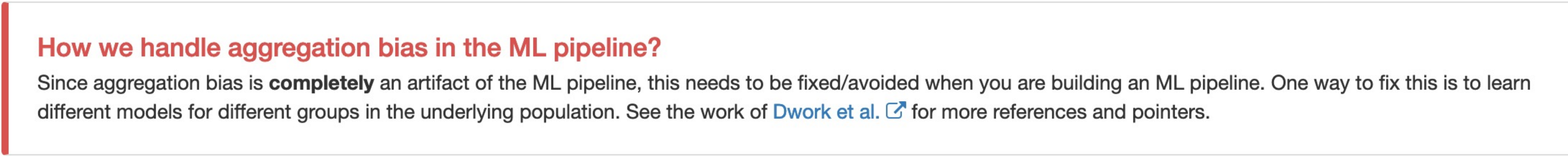
Back to the ML pipeline



What are the relevant interactions?



How do we handle aggregation bias?

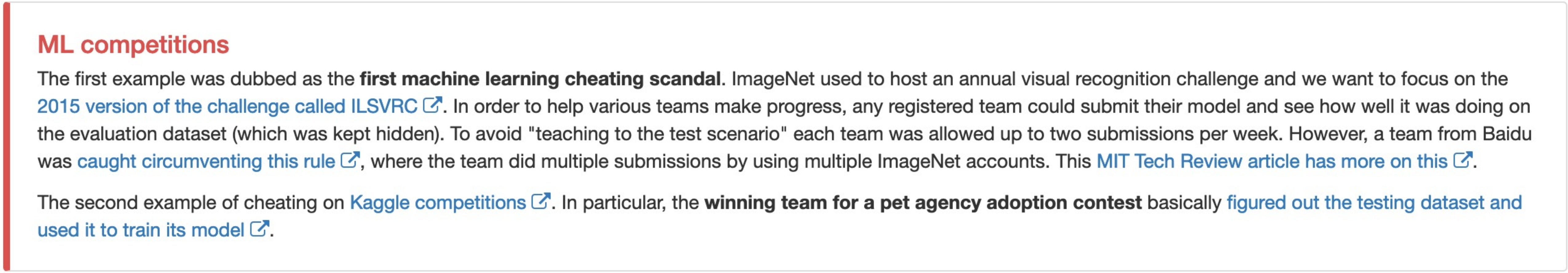


How we handle aggregation bias in the ML pipeline?

Since aggregation bias is **completely** an artifact of the ML pipeline, this needs to be fixed/avoided when you are building an ML pipeline. One way to fix this is to learn different models for different groups in the underlying population. See the work of Dwork et al. for more references and pointers.

Evaluation bias

Reason #1: Training to the benchmark

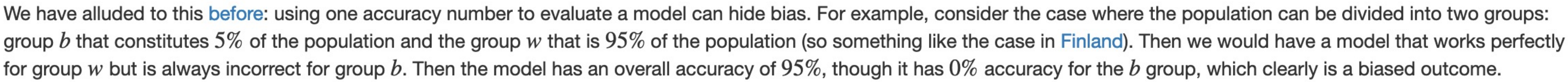


ML competitions

The first example was dubbed as the **first machine learning cheating scandal**. ImageNet used to host an annual visual recognition challenge and we want to focus on the 2015 version of the challenge called ILSVRC. In order to help various teams make progress, any registered team could submit their model and see how well it was doing on the evaluation dataset (which was kept hidden). To avoid "teaching to the test scenario" each team was allowed up to two submissions per week. However, a team from Baidu was caught circumventing this rule, where the team did multiple submissions by using multiple ImageNet accounts. This MIT Tech Review article has more on this.

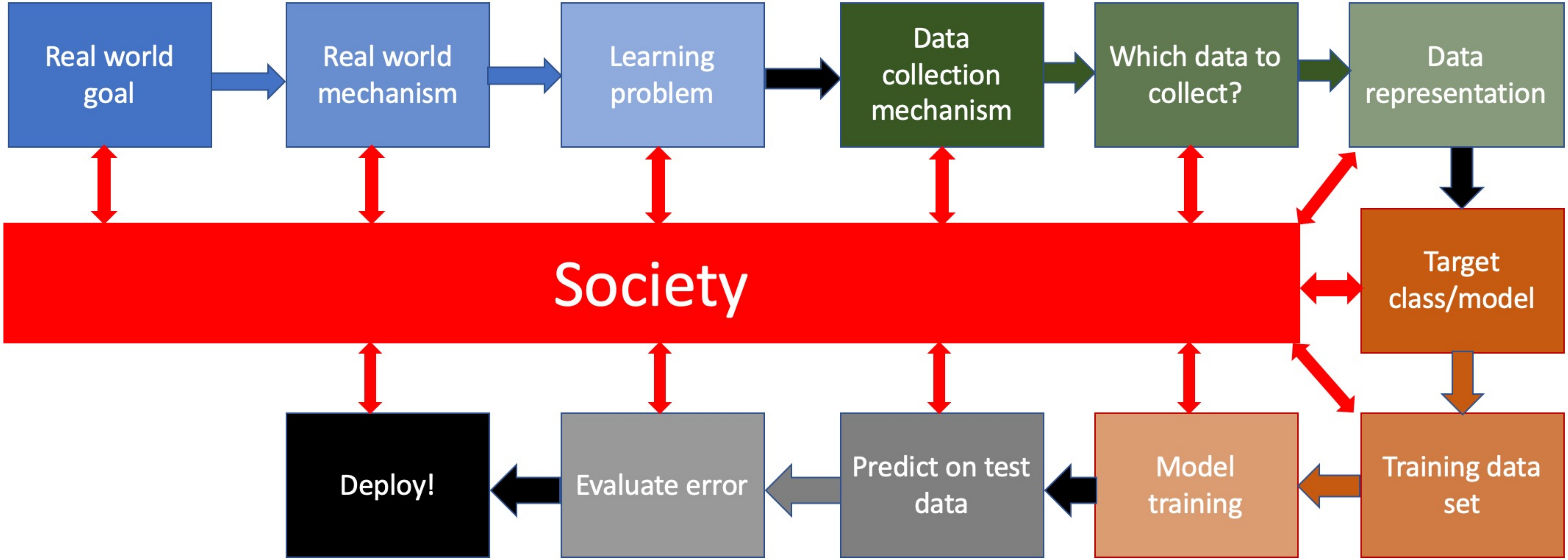
The second example of cheating on Kaggle competitions. In particular, the **winning team for a pet agency adoption contest** basically figured out the testing dataset and used it to train its model.

Reason #2: Using a single accuracy number

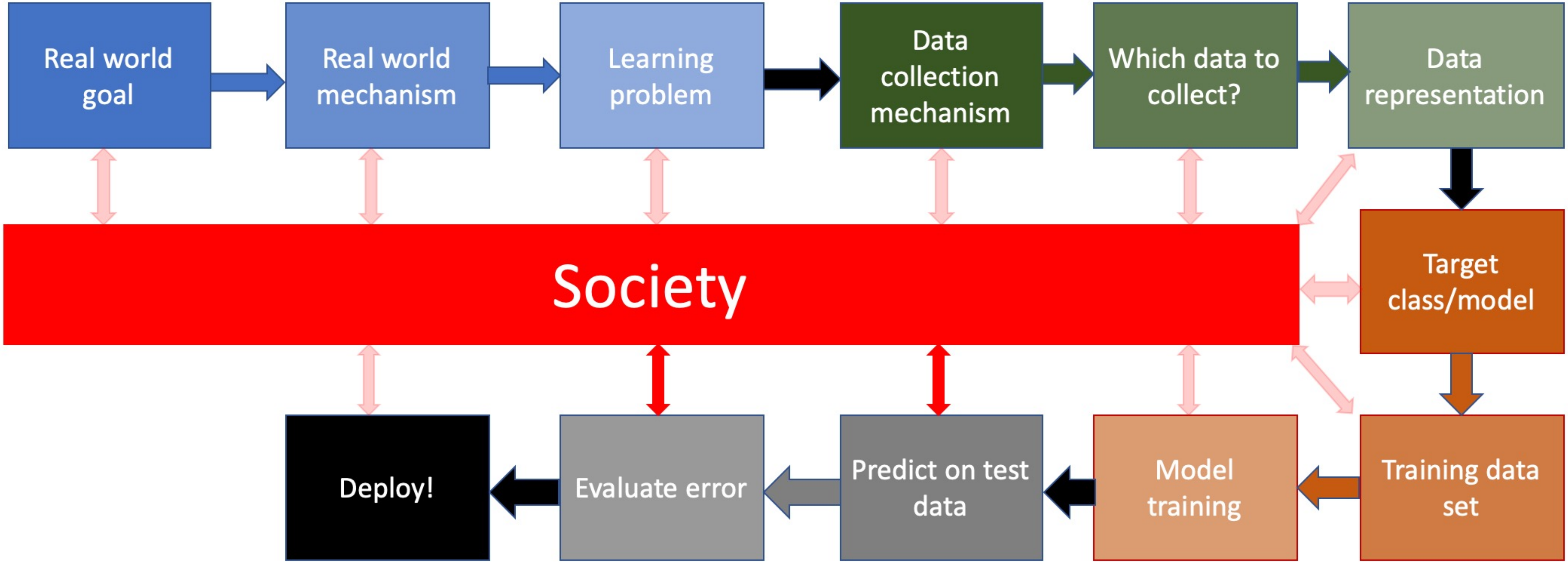


We have alluded to this before: using one accuracy number to evaluate a model can hide bias. For example, consider the case where the population can be divided into two groups: group b that constitutes 5% of the population and the group w that is 95% of the population (so something like the case in Finland). Then we would have a model that works perfectly for group w but is always incorrect for group b. Then the model has an overall accuracy of 95%, though it has 0% accuracy for the b group, which clearly is a biased outcome.

Back to the ML pipeline



What are the relevant interactions?



Real world goal

Real world mechanism

Learning problem

Data collection mechanism

Which data to collect?

Data representation

Society

Target class/model

Training data set

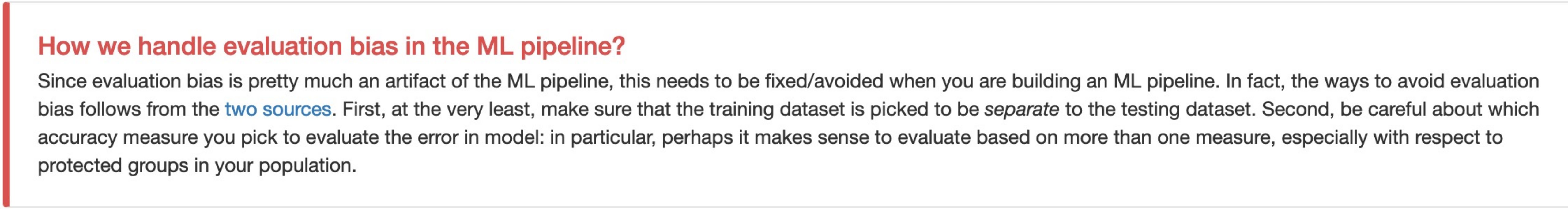
Model training

Predict on test data

Evaluate error

Deploy!

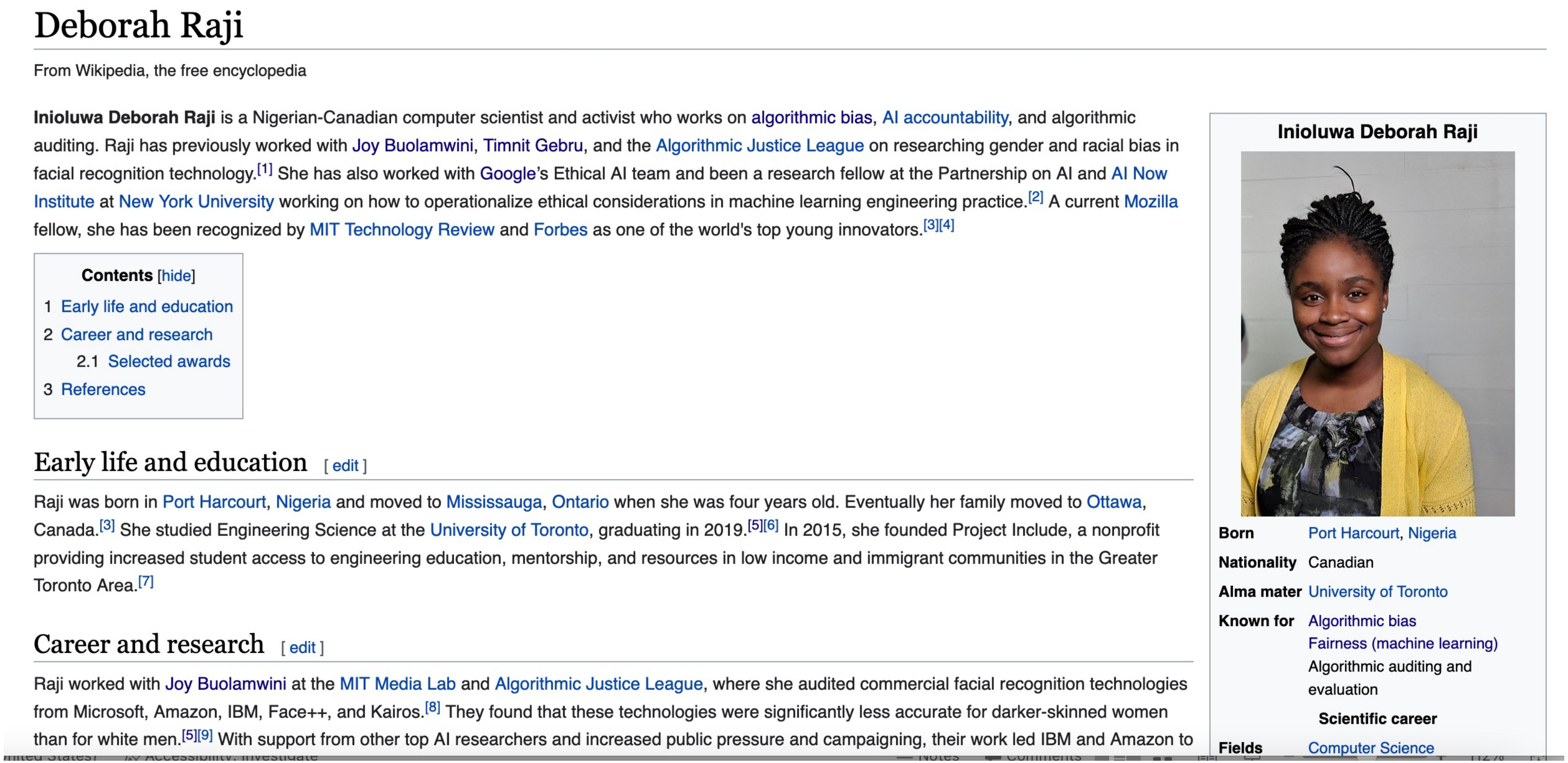
How do we handle evaluation bias?



How we handle evaluation bias in the ML pipeline?

Since evaluation bias is pretty much an artifact of the ML pipeline, this needs to be fixed/avoided when you are building an ML pipeline. In fact, the ways to avoid evaluation bias follows from the two sources. First, at the very least, make sure that the training dataset is picked to be *separate* to the testing dataset. Second, be careful about which accuracy measure you pick to evaluate the error in model: in particular, perhaps it makes sense to evaluate based on more than one measure, especially with respect to protected groups in your population.

Passphrase for today: Deborah Raji



Deborah Raji

From Wikipedia, the free encyclopedia

Inioluwa Deborah Raji is a Nigerian-Canadian computer scientist and activist who works on algorithmic bias, AI accountability, and algorithmic auditing. Raji has previously worked with Joy Buolamwini, Timnit Gebru, and the Algorithmic Justice League on researching gender and racial bias in facial recognition technology.[1] She has also worked with Google's Ethical AI team and been a research fellow at the Partnership on AI and AI Now Institute at New York University working on how to operationalize ethical considerations in machine learning engineering practice.[2] A current Mozilla fellow, she has been recognized by MIT Technology Review and Forbes as one of the world's top young innovators.[3][4]

Contents [hide]

Early life and education [edit]

Raji was born in Port Harcourt, Nigeria and moved to Mississauga, Ontario when she was four years old. Eventually her family moved to Ottawa, Canada.[3] She studied Engineering Science at the University of Toronto, graduating in 2019.[5][6] In 2015, she founded Project Include, a nonprofit providing increased student access to engineering education, mentorship, and resources in low income and immigrant communities in the Greater Toronto Area.[7]

Career and research [edit]

Raji worked with Joy Buolamwini at the MIT Media Lab and Algorithmic Justice League, where she audited commercial facial recognition technologies from Microsoft, Amazon, IBM, Face++, and Kairos.[8] They found that these technologies were significantly less accurate for darker-skinned women than for white men.[5][9] With support from other top AI researchers and increased public pressure and campaigning, their work led IBM and Amazon to

Inioluwa Deborah Raji

Born	Port Harcourt, Nigeria
Nationality	Canadian
Alma mater	University of Toronto
Known for	Algorithmic bias Fairness (machine learning) Algorithmic auditing and evaluation

Scientific career

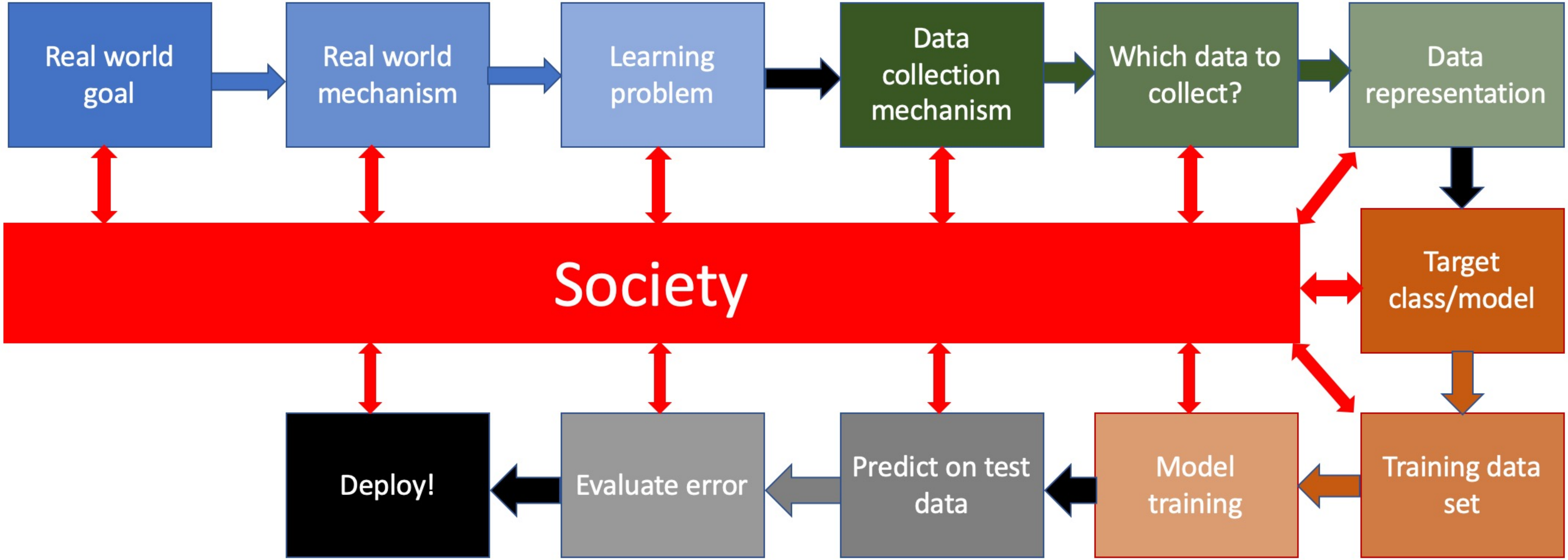
Fields	Computer Science

Deployment bias

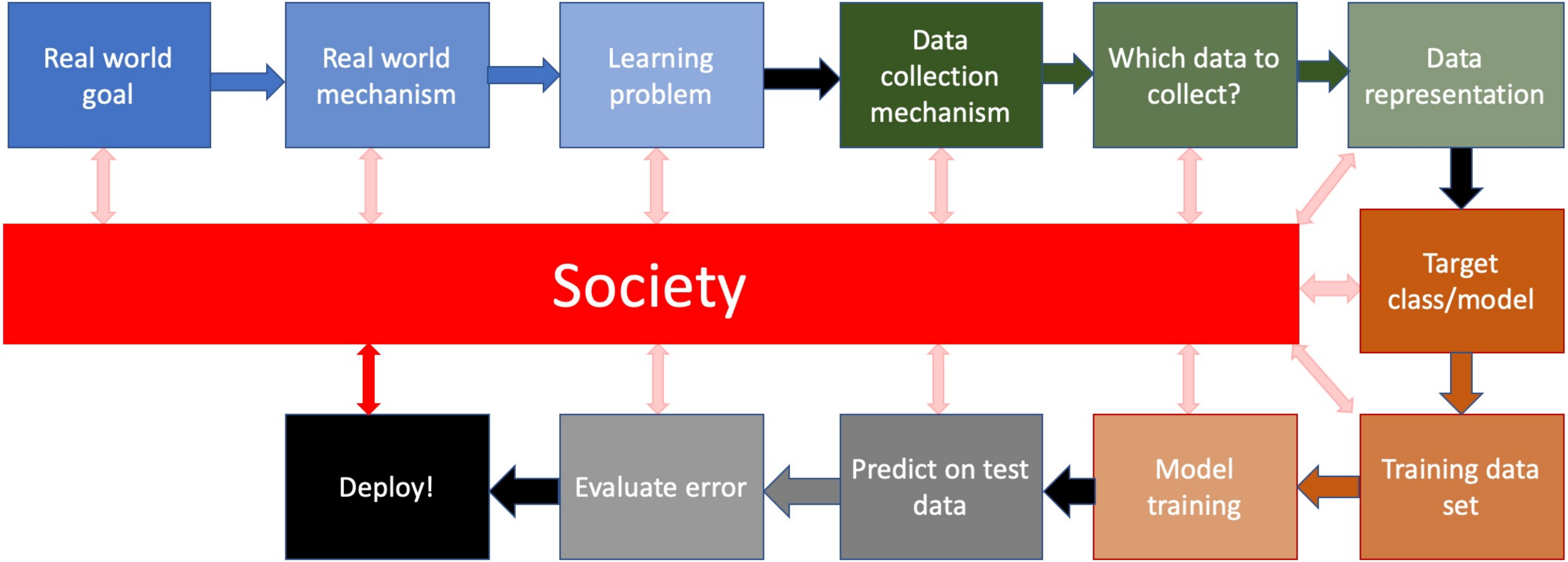
ML pipeline used in a way it was not designed for

More examples

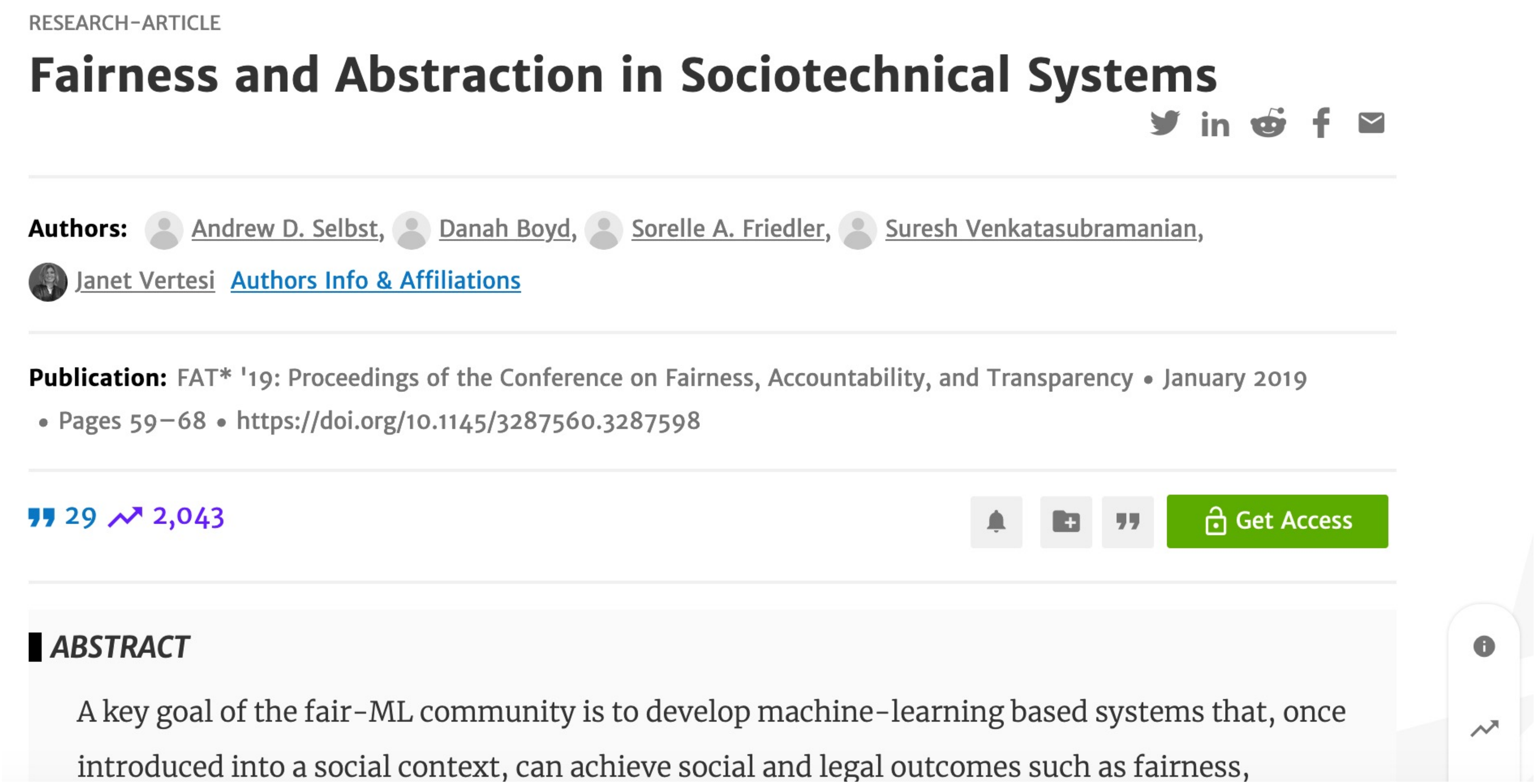
Back to the ML pipeline



What are the relevant interactions?



How do we handle deployment bias?



RESEARCH-ARTICLE

Fairness and Abstraction in Sociotechnical Systems

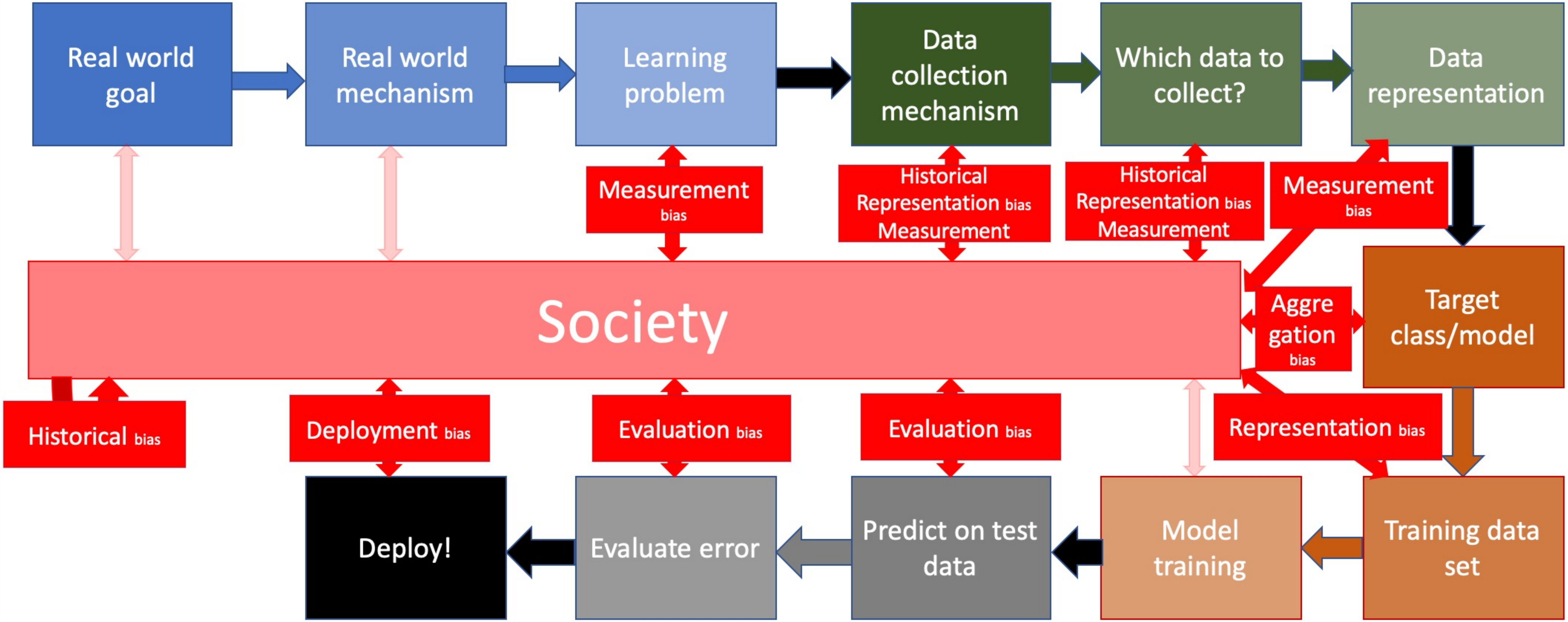
Authors: Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, Janet Vertesi Authors Info & Affiliations

Publication: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency • January 2019 • Pages 59–68 • https://doi.org/10.1145/3287560.3287598

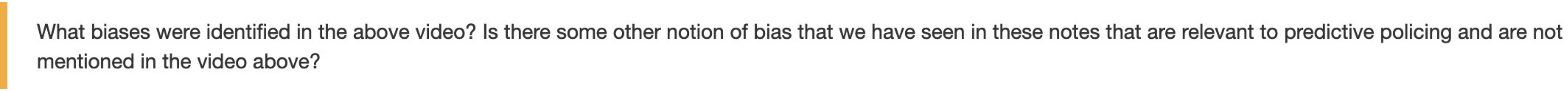
ABSTRACT

A key goal of the fair-ML community is to develop machine-learning based systems that, once introduced into a social context, can achieve social and legal outcomes such as fairness,

All biases in one place…



Exercise



> What biases were identified in the above video? Is there some other notion of bias that we have seen in these notes that are relevant to predictive policing and are not mentioned in the video above?