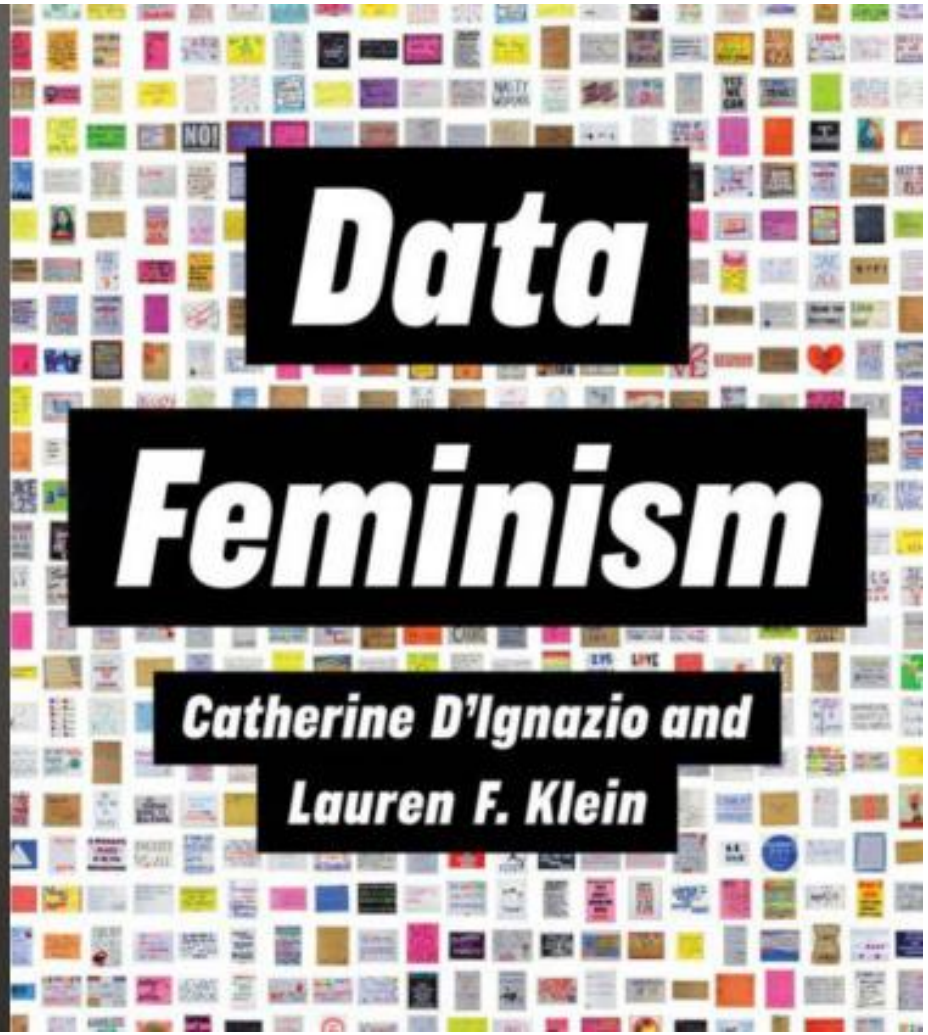
Feedback Loops

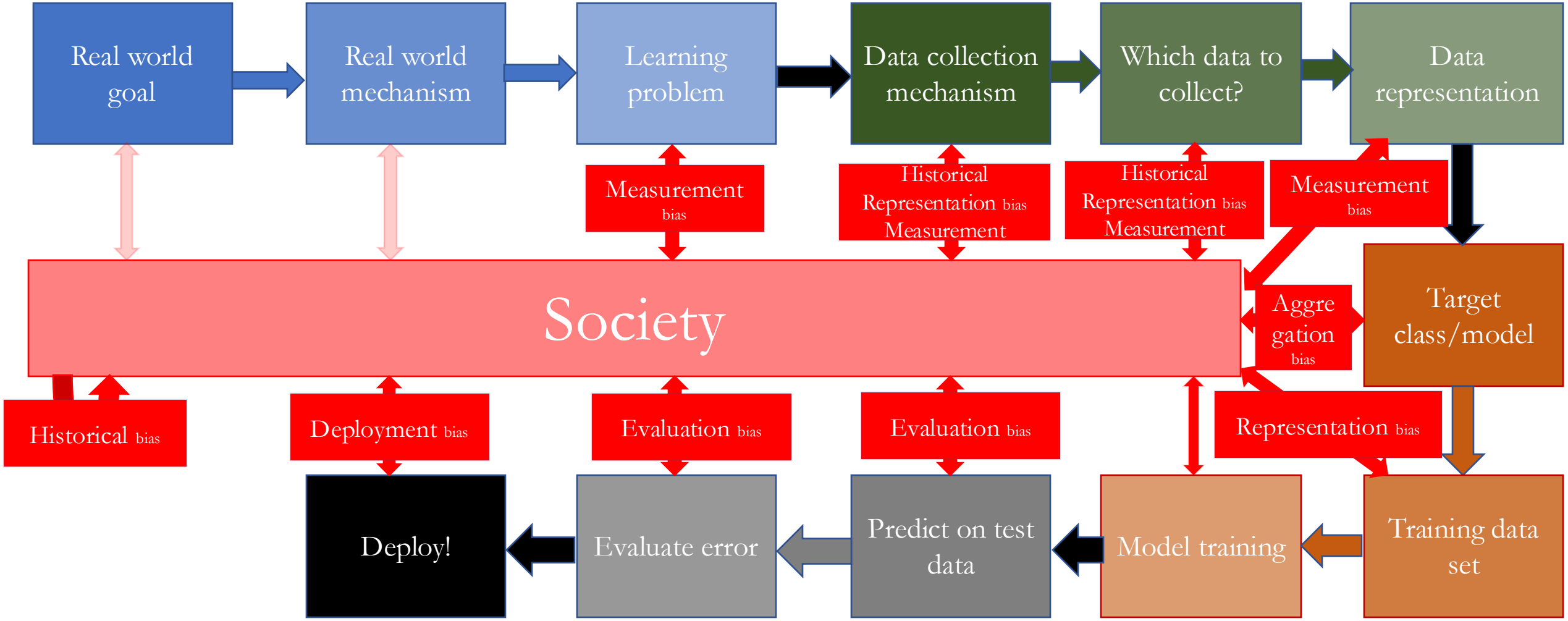
Lecture 2

Spring 2025

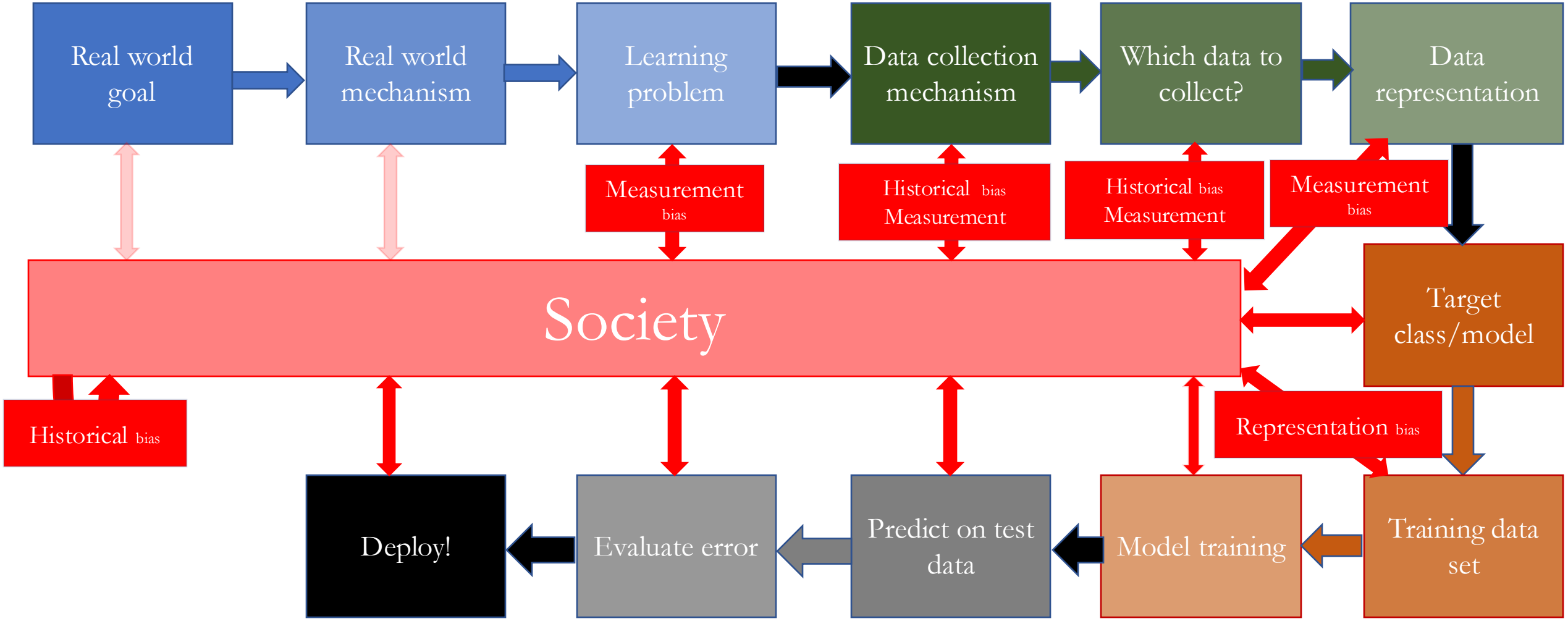
Pass pharse : **Lauren Klein**



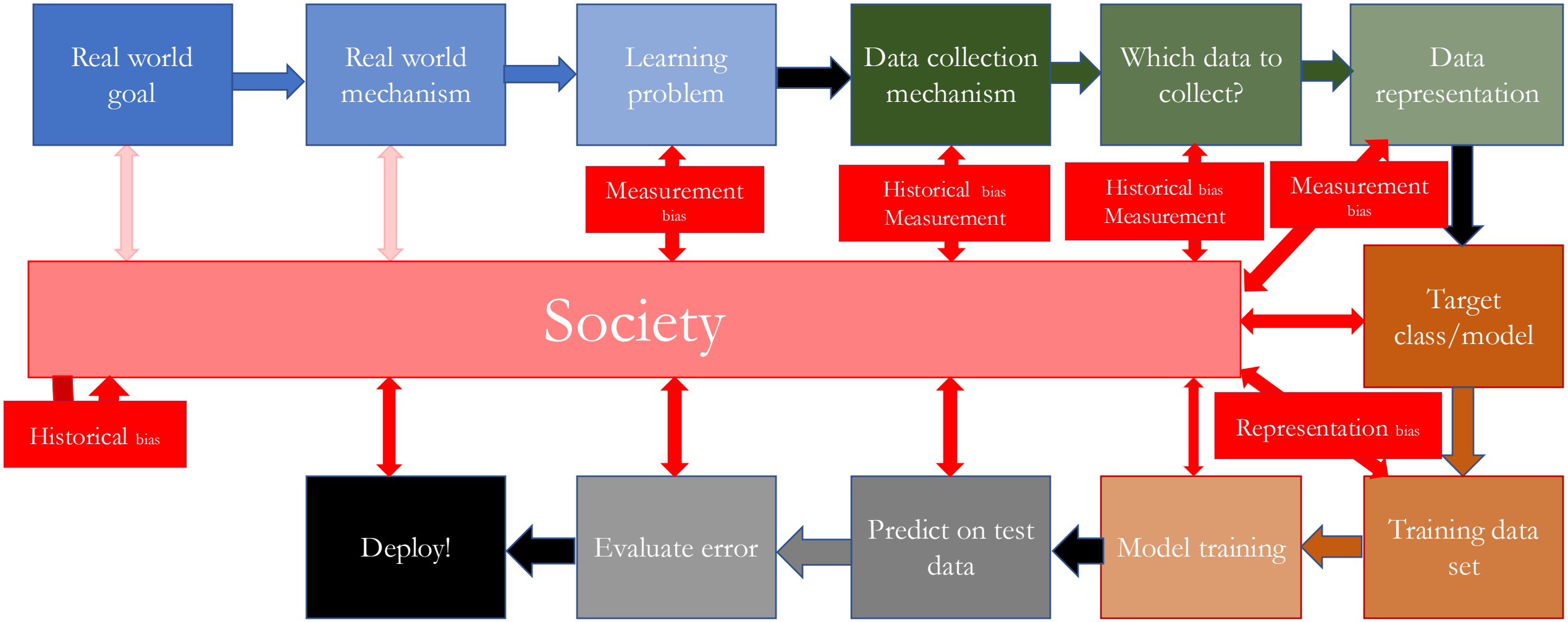
Different biases



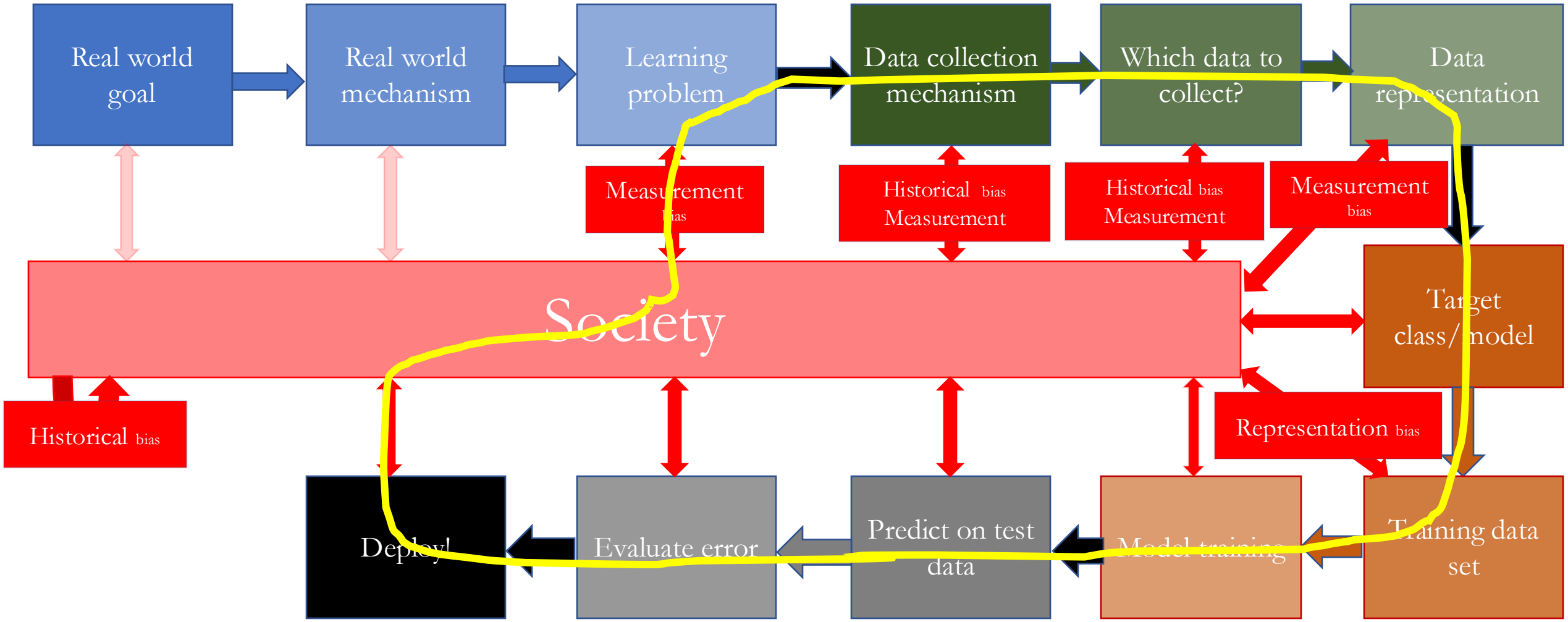
Let us assume…



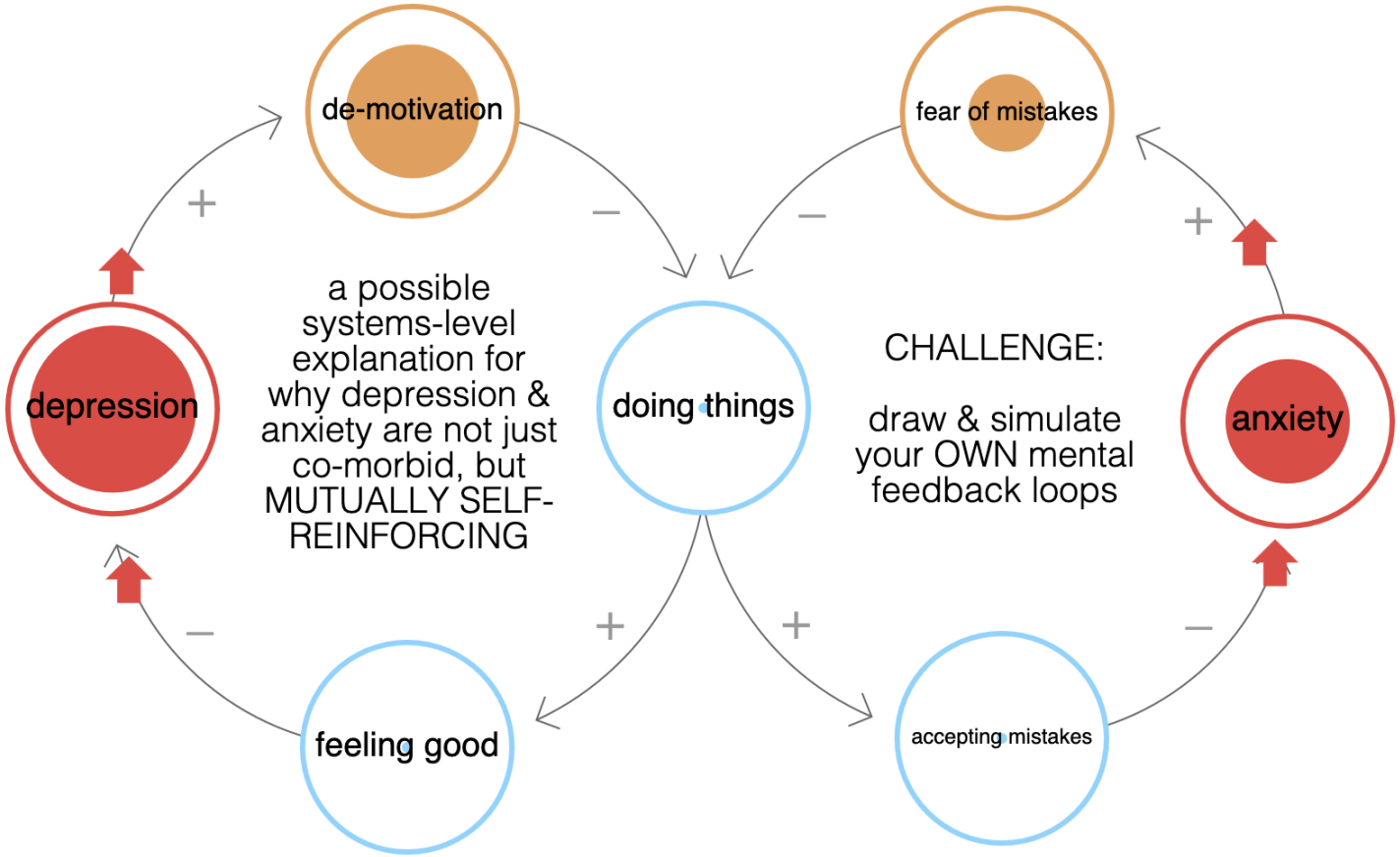
What is a feedback loop



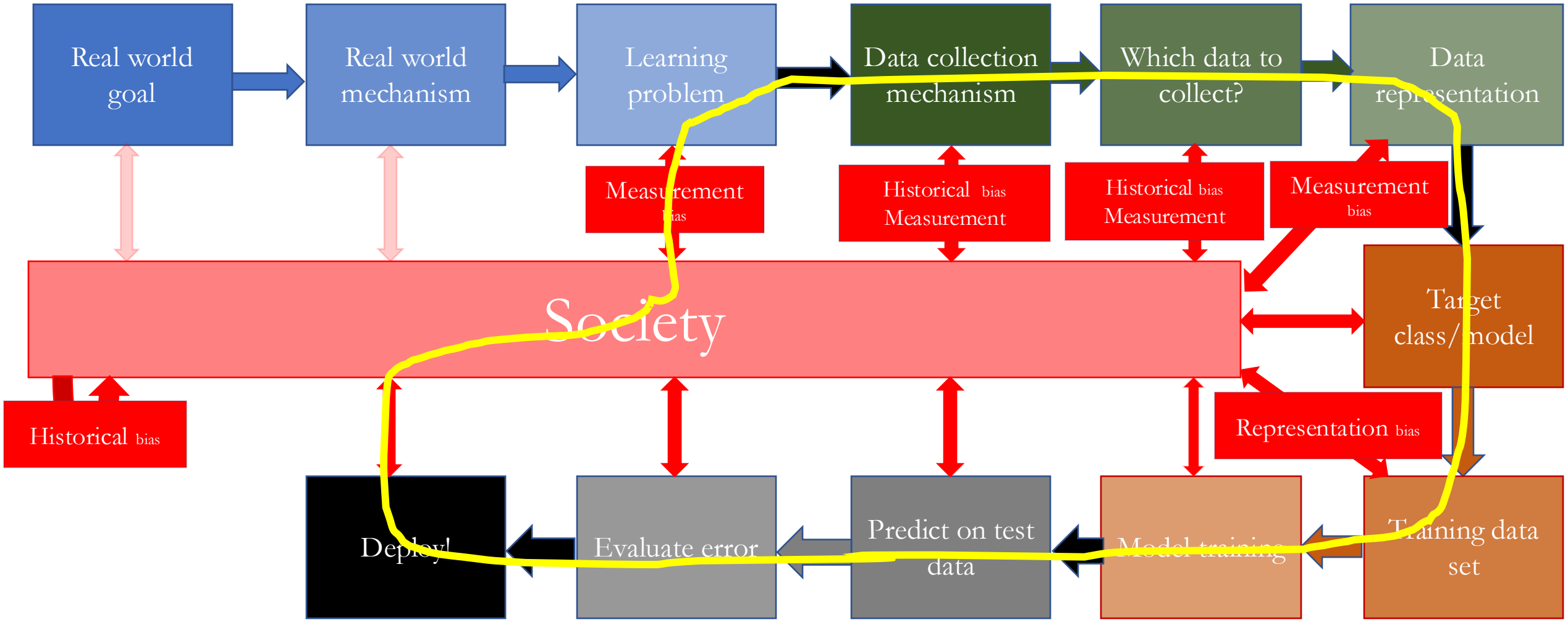
The "loop" in feedback loop



Loopy feedback loop example



The "loop" in feedback loop



Predictive policing

Potential biases in predictive policing

Exercise

Exercise

Figure out how predictive policing can lead to a feedback loop.

Community organization can get results!

Can we formalize this intuition?

How do we “prove” that feedback loops can exist in predictive policing?

Simulation results

Theoretical modeling results

A simulation result

IN DETAIL

To predict and serve?

Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data? **Kristian Lum** and **William Isaac** consider the evidence – and the social consequences

Theoretical modeling: Ensign et al.

CSE 440/441/540 Resources

Feedback Loop and ML

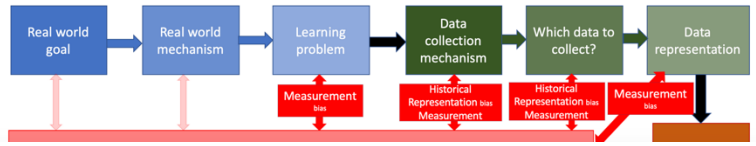
This page talks about how the ML pipeline when deployed in society can lead to a feedback loop.

Under Construction
This page is still under construction. In particular, nothing here is final while this sign still remains here.

A Request
I know I am biased in favor of references that appear in the computer science literature. If you think I am missing a relevant reference (outside or even within CS), please email it to me.

An overview

Recall that we have considered various notions of bias that can creep in when the ML pipeline interacts with society:



Proceedings of Machine Learning Research 81:1–12, 2018 Conference on Fairness, Accountability, and Transparency

Runaway Feedback Loops in Predictive Policing*

Danielle Ensign DANIPHYE@GMAIL.COM
University of Utah

Sorelle A. Friedler SORELLE@CS.HAVERFORD.EDU
Haverford College

Scott Neville DROP.SCOTT.N@GMAIL.COM
University of Utah

Carlos Scheidegger CSCHEID@CSCHEID.NET
University of Arizona

Suresh Venkatasubramanian† SURESH@CS.UTAH.EDU
University of Utah

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Predictive policing systems algorithm) interact: used to determine how to allo its can atten- across a city in order to best pre way feedback, they Discovered crime data (e.g. thout the in- are used to help update the model, a t. the process is repeated. Such systems k loops, predictive been empirically shown to be susceptible to poli line lea runaway feedback loops, where police are repeatedly sent back to the same neighborhoods regardless of the true crime rate.

A LOT of simplifications!

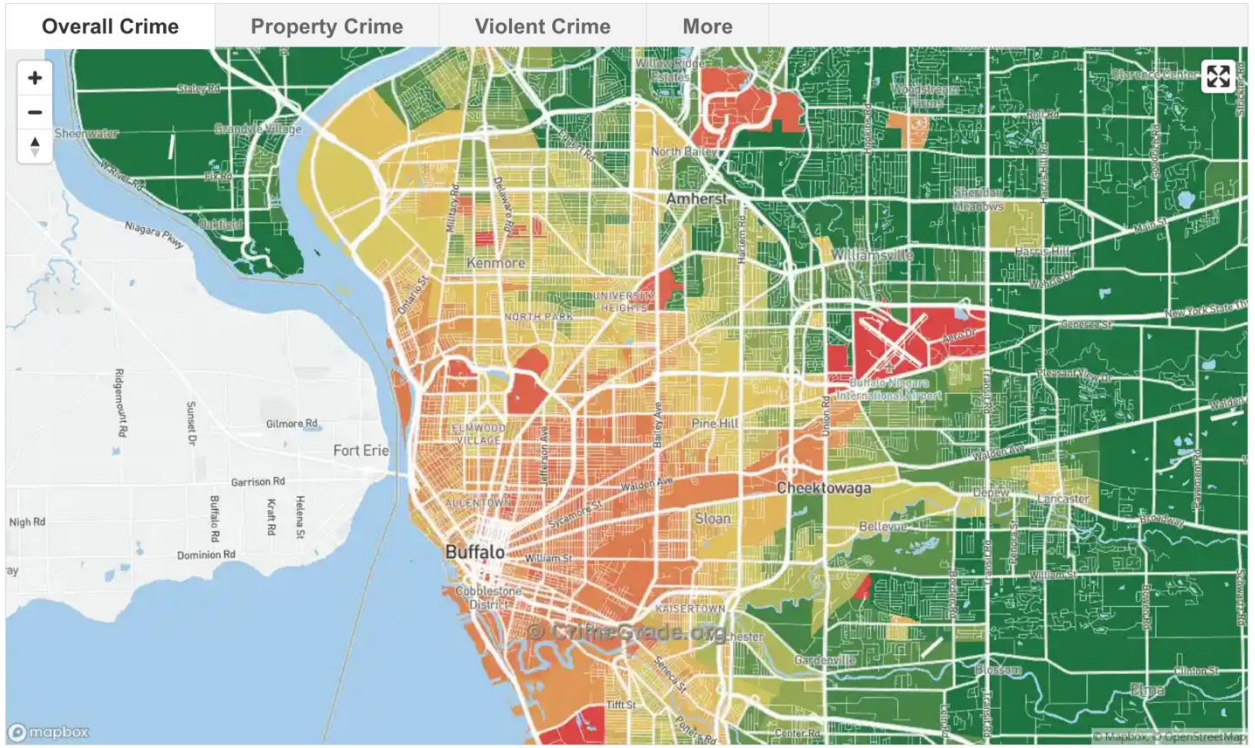
1. Introduction

Basic Setup

Only ONE cop patrol E and W

Crime per Capita in Buffalo

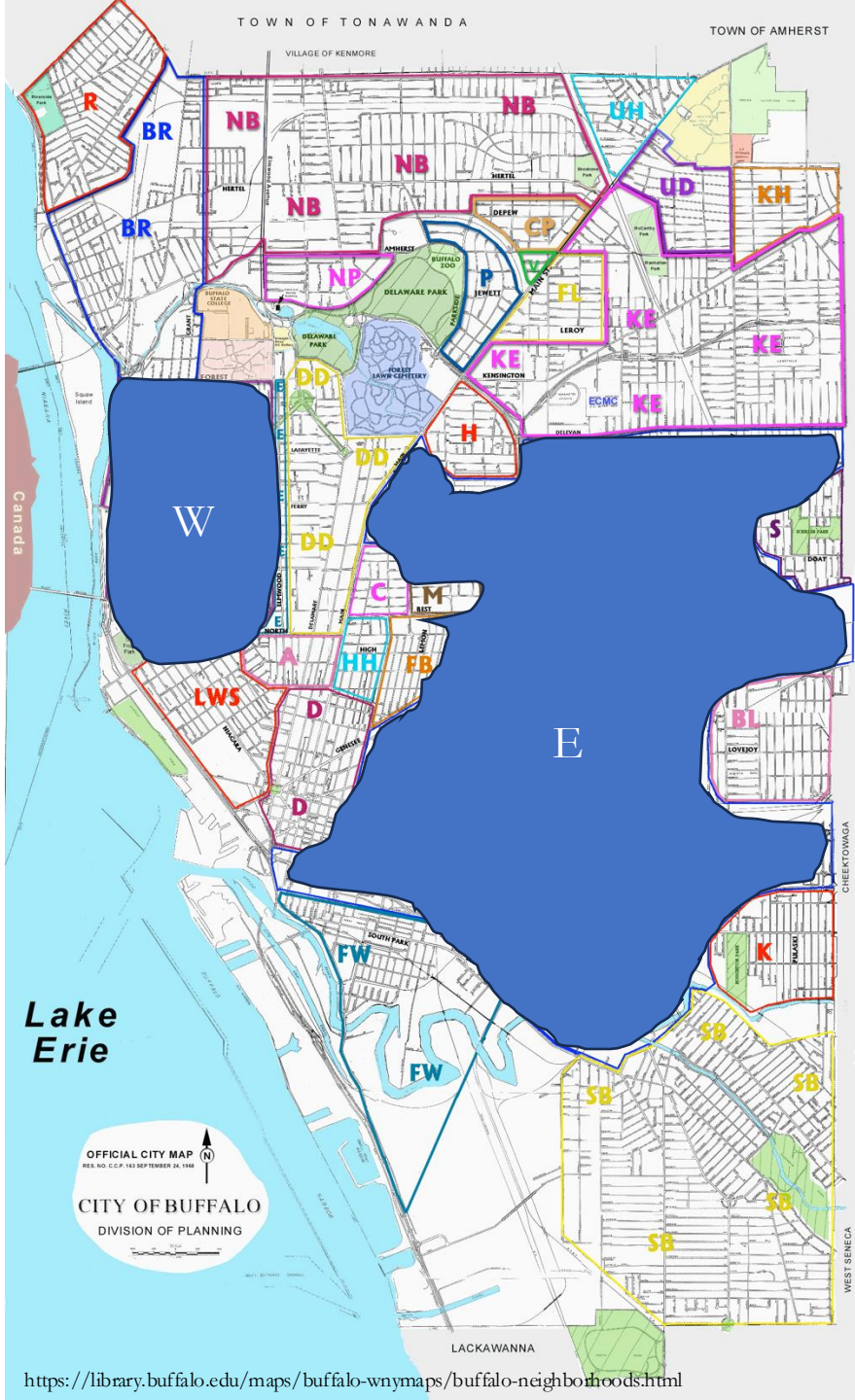
The map below shows crime per 1,000 Buffalo residents.



Crime Grades

https://crimegrade.org/safest-places-in-buffalo-ny/

A+ (dark green) areas are safest



https://library.buffalo.edu/maps/buffalo-wnymaps/buffalo-neighborhoods.html

Assumption 1: One region/day

Cop can only go to one of E or W region per day

Cop will go to E or W with probability proportional to the number of crimes reported in that region

> **Notation alert**
> For any given day t, we will use $n_E^{(t)}$ and $n_W^{(t)}$ to denote the number of observed crimes in E and W respectively from day 0 to day t.

Cop will visit E on t+1 with probability $\dfrac{n_E^{(t)}}{n_E^{(t)} + n_W^{(t)}}$

Cop will visit W on t+1 with probability $\dfrac{n_W^{(t)}}{n_E^{(t)} + n_W^{(t)}}$.

Assumption 2: Unequal crime rates

E and W unequal crime rates (which are known)

> **Notation alert**
>
> The crime rate for E is λ_E and the crime rate for W is λ_W.

Mathematically: $\lambda_E \neq \lambda_W$.

Do NOT need λ_E and λ_W to be far apart $\lambda_E = 10.5\%$ and $\lambda_W = 11\%$

Assumption 3: Observed = actual crime rate

Cop discovers crime at *exactly* the same rate as the actual crime rate in either region

Is this a reasonable assumption?

If cop goes to E:

Discovers one crime with probability λ_E and no crime with probability $1-\lambda_E$

If cop goes to W:

Discovers one crime with probability λ_W and no crime with probability $1-\lambda_W$

An exercise

If cop goes to E:

> Discovers one crime with probability λ_E and no crime with probability 1- λ_E

If cop goes to W:

> Discovers one crime with probability λ_W and no crime with probability 1- λ_W

Notation alert

For any given day t, we will use $n_E^{(t)}$ and $n_W^{(t)}$ to denote the number of observed crimes in E and W respectively from day 0 to day t.

Exercise

Given the above what are the relationships of $n_E^{(t+1)}$ with $n_E^{(t)}$ (and similarly the relationship of $n_W^{(t+1)}$ with $n_W^{(t)}$)?

Solution to exercise

Exercise

Given the above what are the relationships of $n_E^{(t+1)}$ with $n_E^{(t)}$ (and similarly the relationship of $n_W^{(t+1)}$ with $n_W^{(t)}$)?

If the cop visits E with probability λ_E the cop will discover/report one crime and not crime otherwise. In other words,

$$n_E^{(t+1)} = \begin{cases} n_E^{(t)} + 1 & \text{with probability } \lambda_E \\ n_E^{(t)} & \text{with probability } 1 - \lambda_E \end{cases}.$$

And we have a similar result **If the cop visits** W:

$$n_W^{(t+1)} = \begin{cases} n_W^{(t)} + 1 & \text{with probability } \lambda_W \\ n_W^{(t)} & \text{with probability } 1 - \lambda_W \end{cases}.$$

Finally, we have our model…

The evolution of the number of observed crimes

- The process starts with initial values $n_E^{(0)}$ and $n_W^{(0)}$.
- For $t = 1, 2, \ldots$

```
//The process repeats "forever"
```

  - With probability $\frac{n_E^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ do:

```
//Cop visits E
```

    - With probability λ_E set $n_E^{(t+1)} = n_E^{(t)} + 1$
    - Else with probability $1 - \lambda_E$ set $n_E^{(t+1)} = n_E^{(t)}$
  - Otherwise with probability $\frac{n_W^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ do:

```
//Cop visits W
```

    - With probability λ_W set $n_W^{(t+1)} = n_W^{(t)} + 1$
    - Else with probability $1 - \lambda_W$ set $n_W^{(t+1)} = n_W^{(t)}$

Next exercise…

Exercise

To make things concrete assume that $n_E^{(0)} = n_W^{(0)} = 100$ and $\lambda_E = 10.5\%$ and $\lambda_W = 11\%$. What would you consider to be a manifestation of feedback loop as the process above runs?

`Hint`: Think about how the ratios $\frac{n_E^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ and $\frac{n_W^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ evolve as t grows larger. (Side question: why are these ratios something worth monitoring?)

Cop will visit E with probability $\dfrac{n_E^{(t)}}{n_E^{(t)} + n_W^{(t)}}$

Cop will visit W with probability $\dfrac{n_W^{(t)}}{n_E^{(t)} + n_W^{(t)}}$.

Solution to exercise

Exercise

To make things concrete assume that $n_E^{(0)} = n_W^{(0)} = 100$ and $\lambda_E = 10.5\%$ and $\lambda_W = 11\%$. What would you consider to be a manifestation of feedback loop as the process above runs?

`Hint`: Think about how the ratios $\frac{n_E^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ and $\frac{n_W^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ evolve as t grows larger. (Side question: why are these ratios something worth monitoring?)

Cop will visit E with probability $\frac{n_E^{(t)}}{n_E^{(t)} + n_W^{(t)}}$ $\approx \lambda_E$ $\ll \lambda_E$ $\gg \lambda_E$

Cop will visit W with probability $\frac{n_W^{(t)}}{n_E^{(t)} + n_W^{(t)}}$. $\approx \lambda_W$ $\gg \lambda_W$ $\ll \lambda_W$

Does this model have a feedback loop?

The evolution of the number of observed crimes

- The process starts with initial values $n_E^{(0)}$ and $n_W^{(0)}$.
- For $t = 1, 2, \ldots$

```
//The process repeats "forever"
```

  - With probability $\frac{n_E^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ do:

```
//Cop visits E
```

    - With probability λ_E set $n_E^{(t+1)} = n_E^{(t)} + 1$
    - Else with probability $1 - \lambda_E$ set $n_E^{(t+1)} = n_E^{(t)}$
  - Otherwise with probability $\frac{n_W^{(t)}}{n_E^{(t)}+n_W^{(t)}}$ do:

```
//Cop visits W
```

    - With probability λ_W set $n_W^{(t+1)} = n_W^{(t)} + 1$
    - Else with probability $1 - \lambda_W$ set $n_W^{(t+1)} = n_W^{(t)}$

$n_E^{(0)} = n_W^{(0)} = 100$ and $\lambda_E = 10.5\%$ and $\lambda_W = 11\%$.

What do you think will happen?

Answer is yes and in the most extreme sense.

		$\lambda_E > \lambda_W$	$\lambda_E < \lambda_W$
Cop will visit E with probability	$\frac{n_E^{(t)}}{n_E^{(t)} + n_W^{(t)}} \approx \lambda_E$	$= 1$	$= 0$
Cop will visit W with probability	$\frac{n_W^{(t)}}{n_E^{(t)} + n_W^{(t)}}. \approx \lambda_W$	$= 0$	$= 1$

How the heck do you prove such a thing?

Pólya urn model

6 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

In statistics, a **Pólya urn model** (also known as a **Pólya urn scheme** or simply as **Pólya's urn**), named after George Pólya, is a family of urn models that can be used to interpret many commonly used statistical models.

The model represents objects of interest (such as atoms, people, cars, etc.) as colored balls in an urn. In the basic Pólya urn model, the experimenter puts *x* white and *y* black balls into an urn. At each step, one ball is drawn uniformly at random from the urn, and its color observed; it is then returned in the urn, and an additional ball of the same color is added to the urn.

If by random chance, more black balls are drawn than white balls in the initial few draws, it would make it more likely for more black balls to be drawn later. Similarly for the white balls. Thus the urn has a self-reinforcing property ("the rich get richer"). It is the opposite of sampling without replacement, where every time a particular value is observed, it is less likely to be observed again, whereas in a Pólya urn model, an observed value is *more* likely to be observed again. In a Pólya urn model, successive acts of measurement over time have less and less effect on future measurements, whereas in sampling without replacement, the opposite is true: After a certain number of measurements of a particular value, that value will never be seen again.

It is also different from sampling with replacement, where the ball is returned to the urn but without adding new balls. In this case, there is neither self-reinforcing nor anti-self-reinforcing.

Lemma 3 (Renlund (2010)) *Suppose we are given a Pólya urn with replacement matrix of the form*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with a positive number of balls of each kind to start with. Assume that $a, b, c, d \geq 0$ and at least one entry is strictly positive. Then the limit of the fraction of balls of each type exists almost surely. The fraction p of A-colored balls can be characterized as follows:

- *If $a = d, c = b = 0$, then p tends towards a beta distribution.*
- *If not, then p tends towards a single point distribution x^*, where $x^* \in [0, 1]$ is a root of the quadratic polynomial*

$$(c + d - a - b)x^2 + (a - 2c - d)x + c.$$

If two such roots exist, then it is the one such that $f'(x^*) < 0$.

Now let's prove this "lemma 3"!

Just kidding ☺

Now let's prove this "lemma 3"!

Just kidding ☺

Is there a (mathematical) “fix”?

A potential fix

In the above model, Ensign, Friedler, Neville, Scheidegger and Venkatasubramanian suggest the following fix (which they can mathematically prove that it works) is based roughly on the following idea. If the cop visits a specific region most of the time, then it should not be a surprise if they discover a crime in the region and in such a case they should "discount" the crime discovery by not recording such discovery most of the time. On the other hand, if the cop visits region infrequently and they discover a crime, they have learned something "new" and hence would record such crime discoveries most of the time.