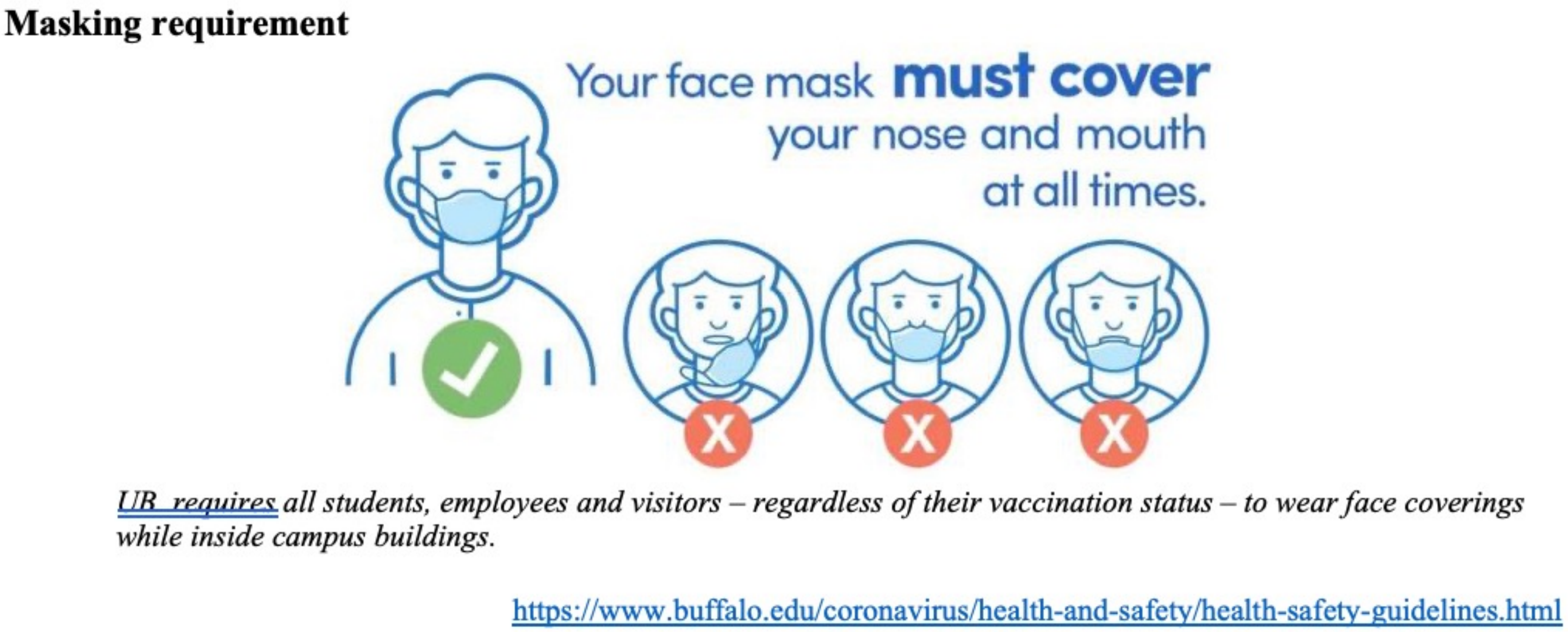
ML and Society

Feb 8, 2022

Please have a face mask on



Masking requirement

Your face mask **must cover** your nose and mouth at all times.

UB requires all students, employees and visitors – regardless of their vaccination status – to wear face coverings while inside campus buildings.

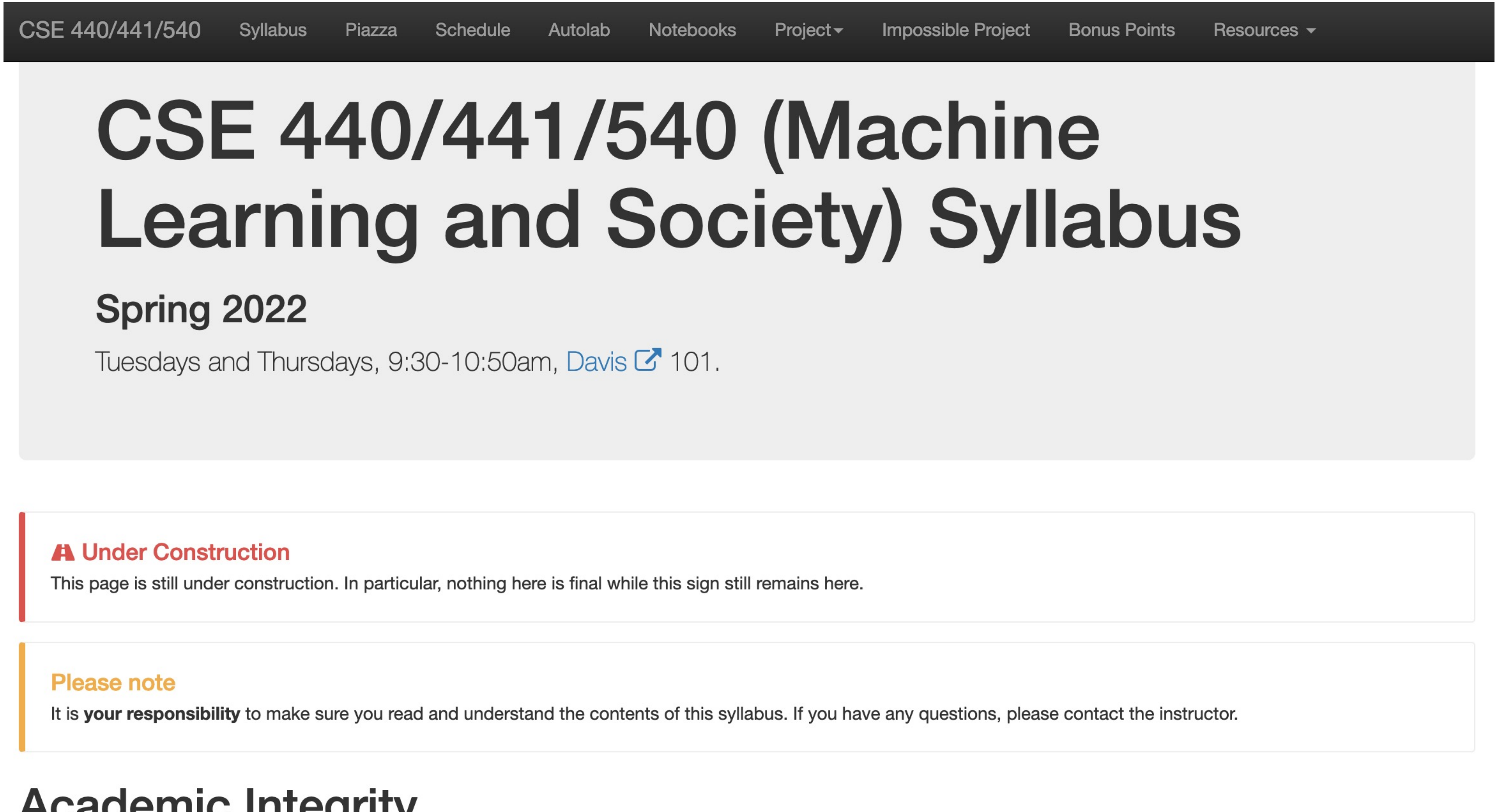
https://www.buffalo.edu/coronavirus/health-and-safety/health-safety-guidelines.html

Who does my machine learning model serve?

How do I know?

What can I do about it?

Read the syllabus carefully!



CSE 440/441/540 (Machine Learning and Society) Syllabus

Spring 2022

Tuesdays and Thursdays, 9:30-10:50am, Davis 101.

Under Construction

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

Please note

It is **your responsibility** to make sure you read and understand the contents of this syllabus. If you have any questions, please contact the instructor.

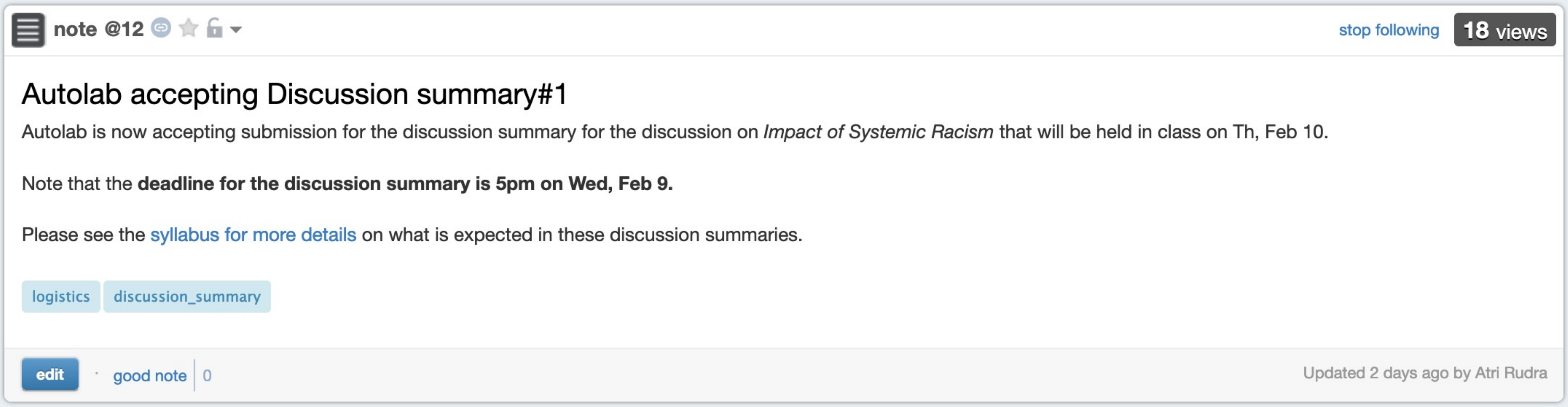
Academic Integrity

Due in the next couple of days

Wed 5pm: Discussion summary on Autolab

Th 5pm: Project choices due (via Google form)

Autolab accepting discussion summaries



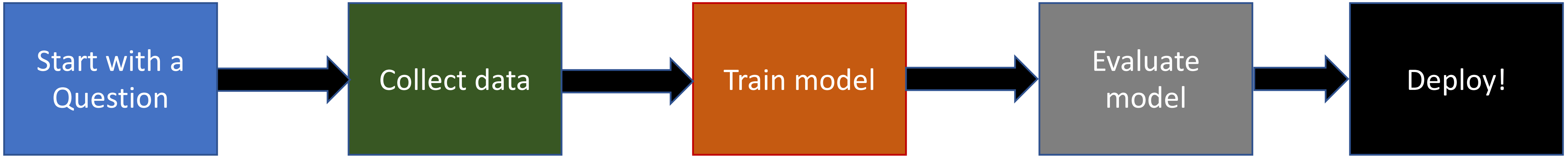
Autolab accepting Discussion summary#1

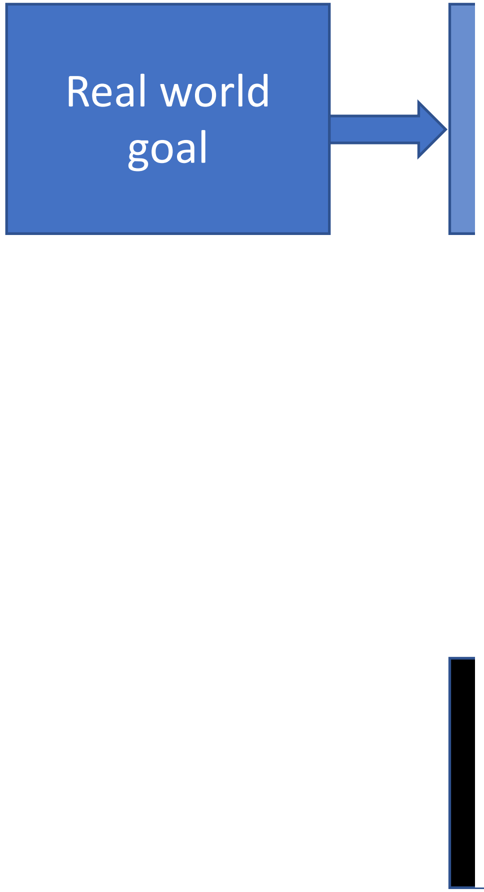
Autolab is now accepting submission for the discussion summary for the discussion on *Impact of Systemic Racism* that will be held in class on Th, Feb 10.

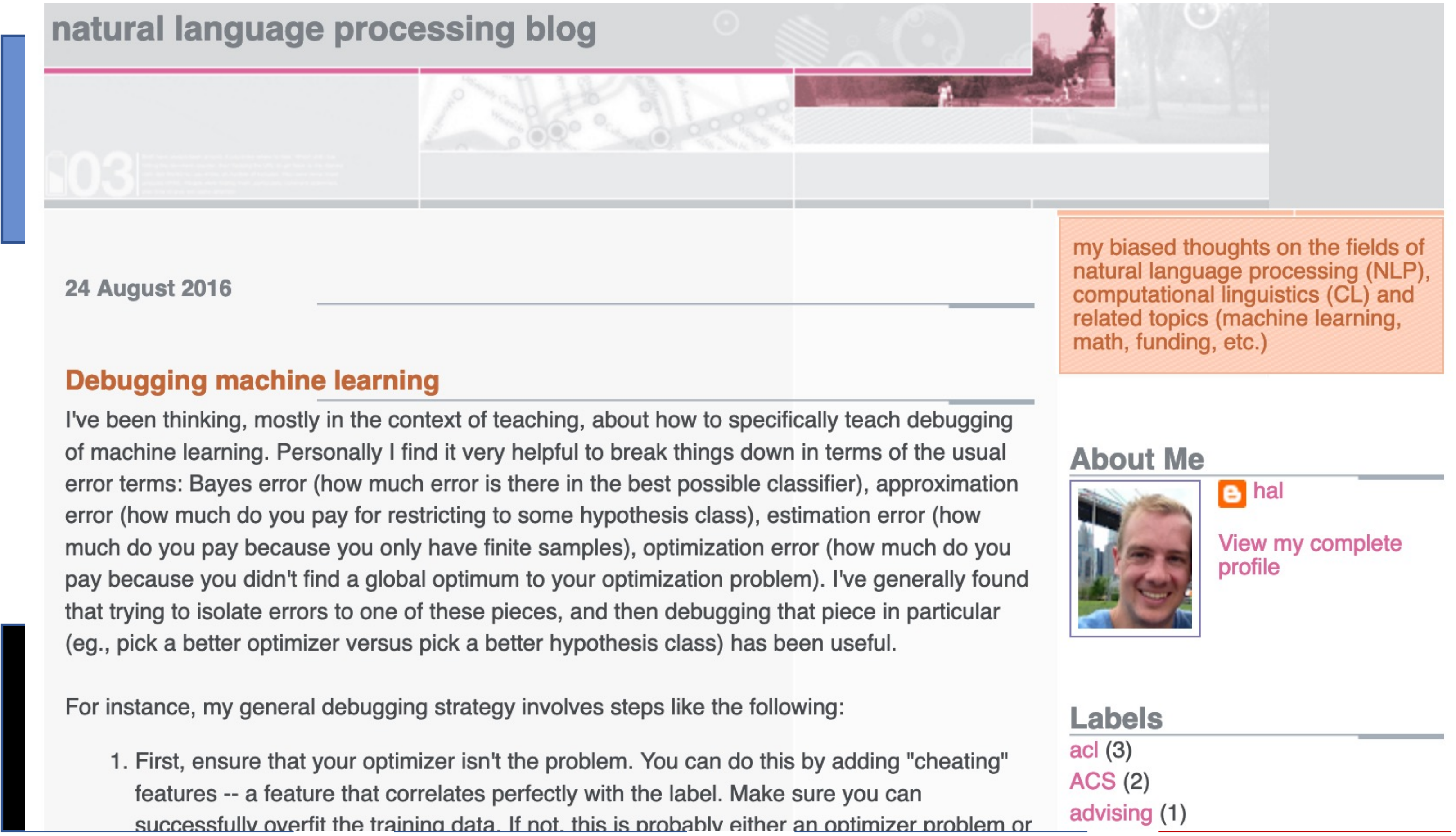
Note that the **deadline for the discussion summary is 5pm on Wed, Feb 9.**

Please see the syllabus for more details on what is expected in these discussion summaries.

Updated 2 days ago by Atri Rudra







natural language processing blog

my biased thoughts on the fields of natural language processing (NLP), computational linguistics (CL) and related topics (machine learning, math, funding, etc.)

24 August 2016

Debugging machine learning

I've been thinking, mostly in the context of teaching, about how to specifically teach debugging of machine learning. Personally I find it very helpful to break things down in terms of the usual error terms: Bayes error (how much error is there in the best possible classifier), approximation error (how much do you pay for restricting to some hypothesis class), estimation error (how much do you pay because you only have finite samples), optimization error (how much do you pay because you didn't find a global optimum to your optimization problem). I've generally found that trying to isolate errors to one of these pieces, and then debugging that piece in particular (eg., pick a better optimizer versus pick a better hypothesis class) has been useful.

For instance, my general debugging strategy involves steps like the following:

1. First, ensure that your optimizer isn't the problem. You can do this by adding "cheating" features -- a feature that correlates perfectly with the label. Make sure you can successfully overfit the training data. If not, this is probably either an optimizer problem or

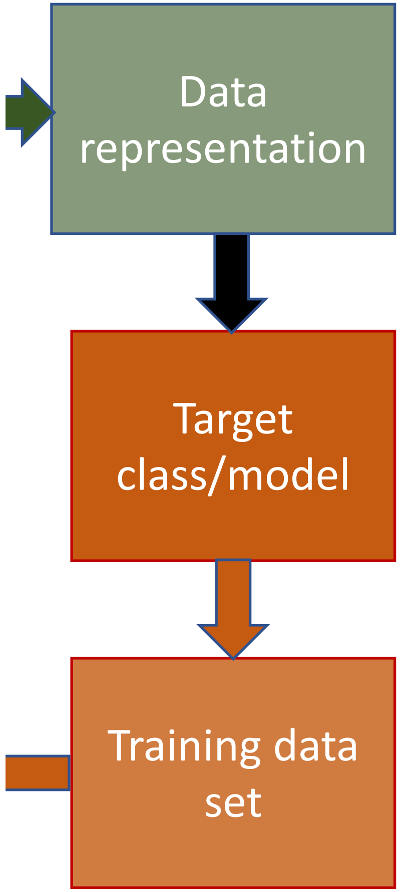
About Me

hal

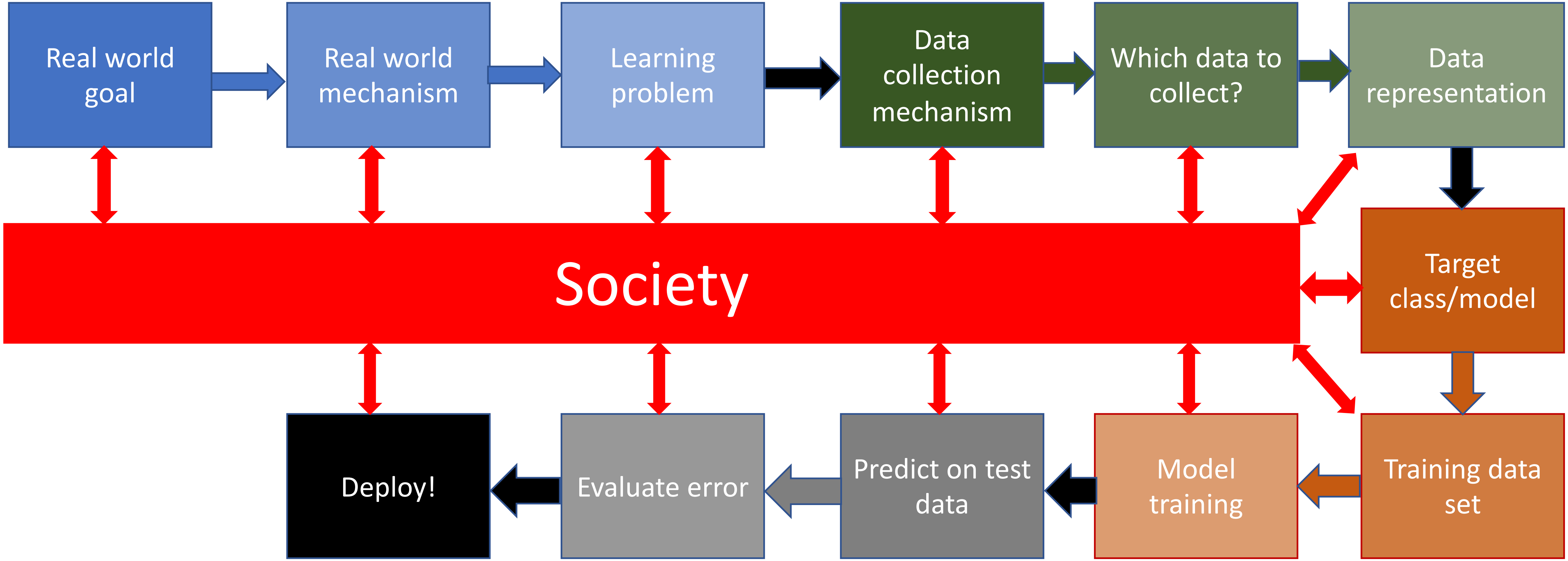
View my complete profile

Labels

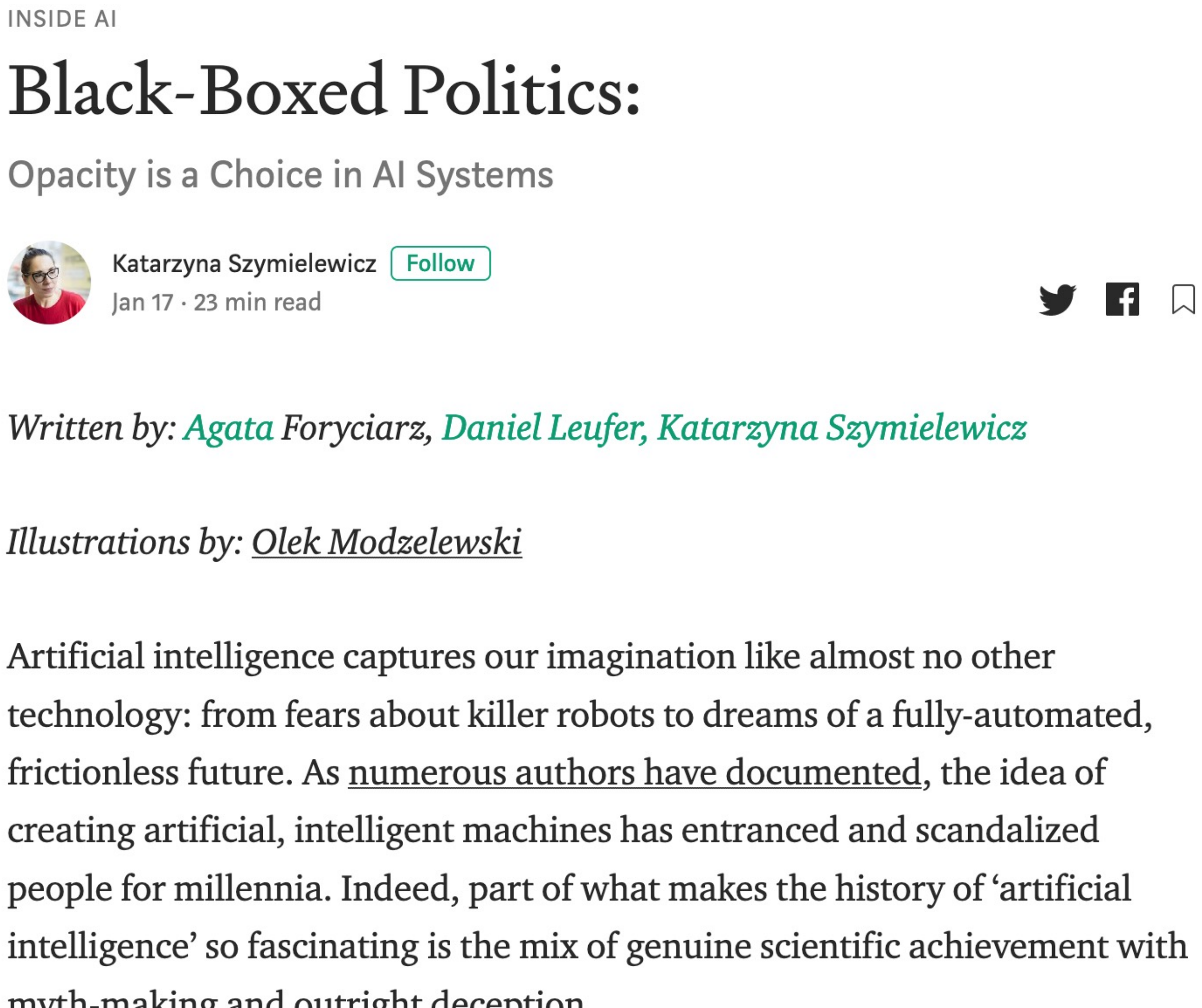
acl (3)
ACS (2)
advising (1)



Putting society back in



Not the only ML+society pipeline in town



INSIDE AI

Black-Boxed Politics:

Opacity is a Choice in AI Systems

Katarzyna Szymielewicz Follow

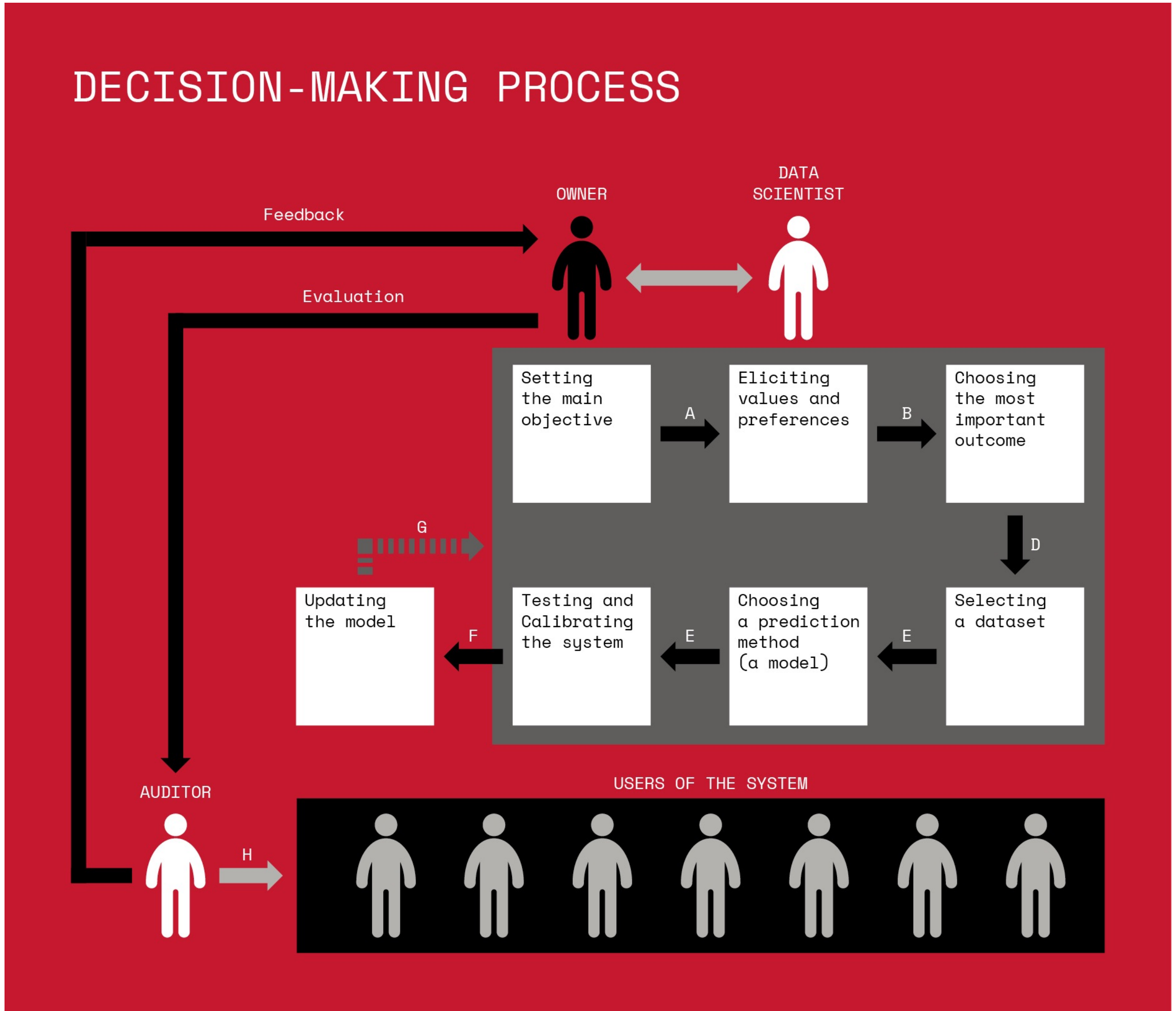
Jan 17 · 23 min read

Written by: Agata Foryciarz, Daniel Leufer, Katarzyna Szymielewicz

Illustrations by: Olek Modzelewski

Artificial intelligence captures our imagination like almost no other technology: from fears about killer robots to dreams of a fully-automated, frictionless future. As numerous authors have documented, the idea of creating artificial, intelligent machines has entranced and scandalized people for millennia. Indeed, part of what makes the history of 'artificial intelligence' so fascinating is the mix of genuine scientific achievement with myth-making and outright deception

Not the only ML+society pipeline in town



DECISION-MAKING PROCESS

OWNER

DATA SCIENTIST

Feedback

Evaluation

Setting the main objective

Eliciting values and preferences

Choosing the most important outcome

Selecting a dataset

Choosing a prediction method (a model)

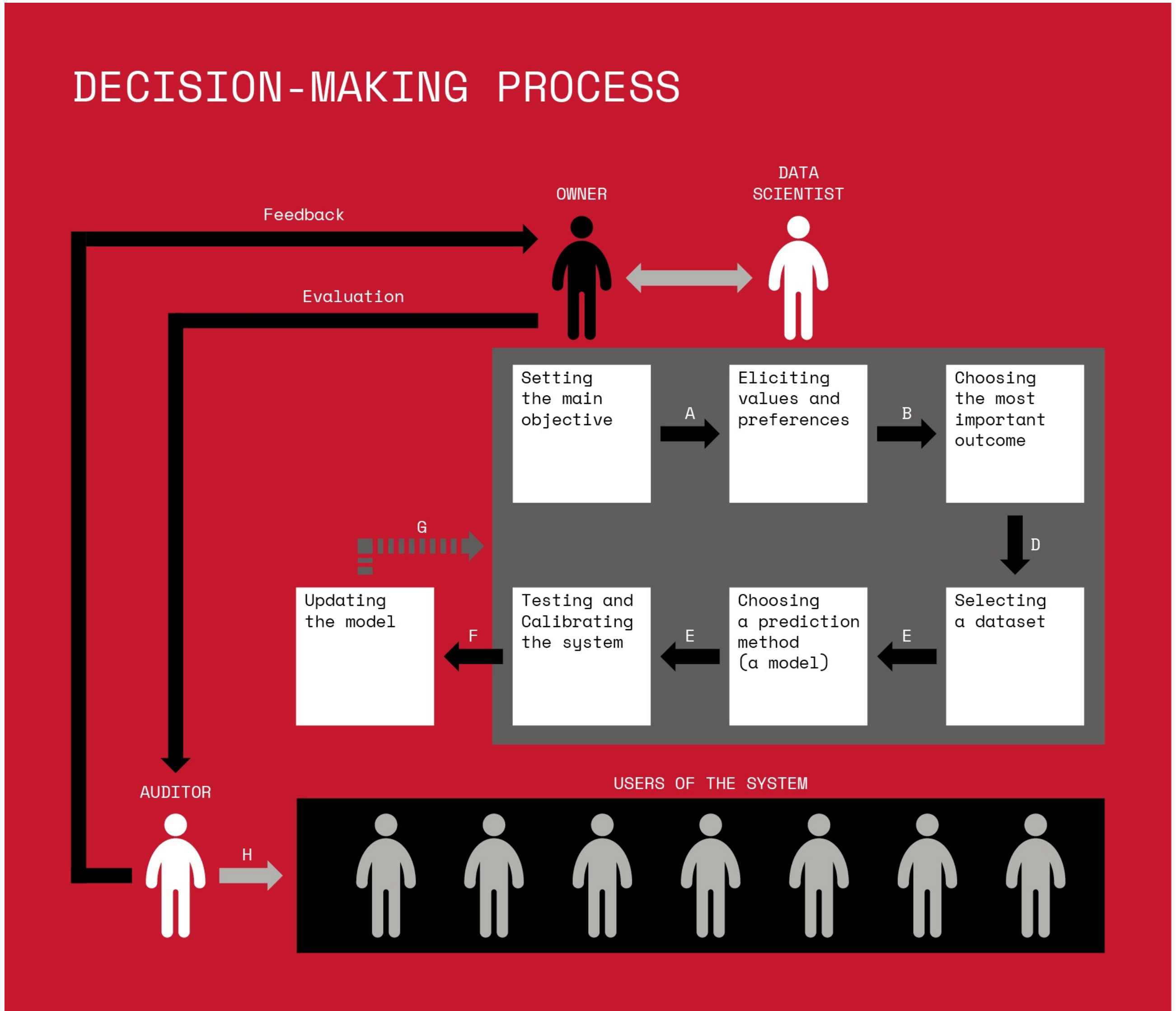
Testing and Calibrating the system

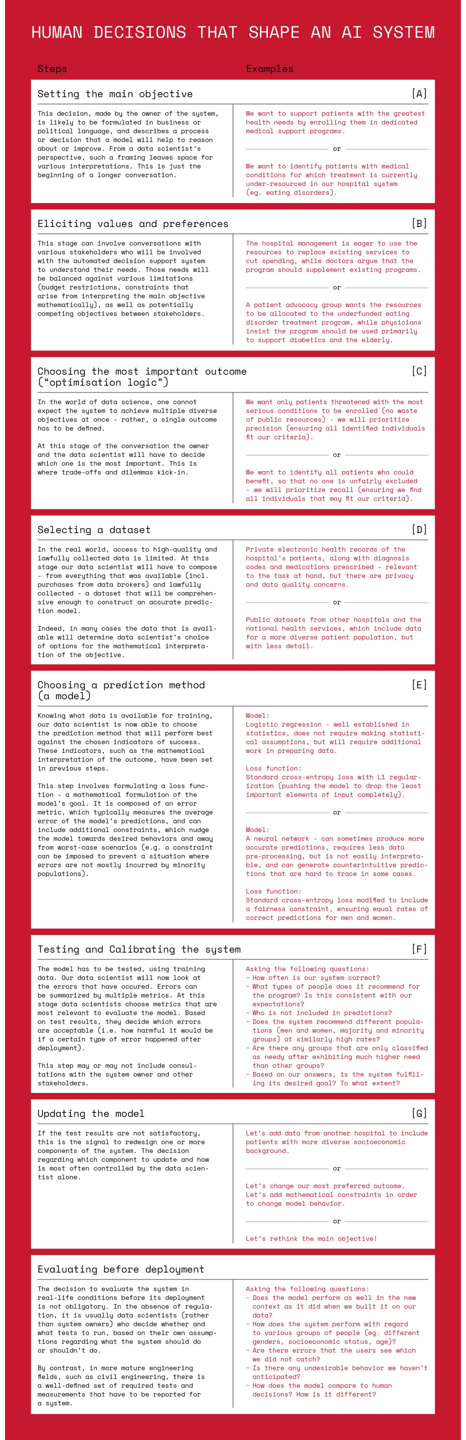
Updating the model

AUDITOR

USERS OF THE SYSTEM

A walkthrough


DECISION-MAKING PROCESS
OWNER
DATA SCIENTIST
Feedback
Evaluation
Setting the main objective
A
Eliciting values and preferences
B
Choosing the most important outcome
D
Selecting a dataset
E
Choosing a prediction method (a model)
E
Testing and Calibrating the system
F
Updating the model
G
AUDITOR
H
USERS OF THE SYSTEM



HUMAN DECISIONS THAT SHAPE AN AI SYSTEM

Steps

Examples

Setting the main objective [A]

This decision, made by the owner of the system, is likely to be formulated in business or political language, and describes a process or decision that a model will help to reason about or improve. From a data scientist's perspective, such a framing leaves space for various interpretations. This is just the beginning of a longer conversation.

We want to support patients with the greatest health needs by enrolling them in dedicated medical support programs.

or

We want to identify patients with medical conditions for which treatment is currently under-resourced in our hospital system (eg. eating disorders).

Eliciting values and preferences [B]

This stage can involve conversations with various stakeholders who will be involved with the automated decision support system to understand their needs. Those needs will be balanced against various limitations (budget restrictions, constraints that arise from interpreting the main objective mathematically), as well as potentially competing objectives between stakeholders.

The hospital management is eager to use the resources to replace existing services to cut spending, while doctors argue that the program should supplement existing programs.

or

A patient advocacy group wants the resources to be allocated to the underfunded eating disorder treatment program, while physicians insist the program should be used primarily to support diabetics and the elderly.

Choosing the most important outcome ("optimisation logic") [C]

In the world of data science, one cannot expect the system to achieve multiple diverse objectives at once - rather, a single outcome has to be defined.

At this stage of the conversation the owner and the data scientist will have to decide which one is the most important. This is where trade-offs and dilemmas kick-in.

We want only patients threatened with the most serious conditions to be enrolled (no waste of public resources) - we will prioritize precision (ensuring all identified individuals fit our criteria).

or

We want to identify all patients who could benefit, so that no one is unfairly excluded - we will prioritize recall (ensuring we find all individuals that may fit our criteria).

Selecting a dataset [D]

In the real world, access to high-quality and lawfully collected data is limited. At this stage our data scientist will have to compose - from everything that was available (incl. purchases from data brokers) and lawfully collected - a dataset that will be comprehensive enough to construct an accurate prediction model.

Indeed, in many cases the data that is available will determine data scientist's choice of options for the mathematical interpretation of the objective.

Private electronic health records of the hospital's patients, along with diagnosis codes and medications prescribed - relevant to the task at hand, but there are privacy and data quality concerns.

or

Public datasets from other hospitals and the national health services, which include data for a more diverse patient population, but with less detail.

Choosing a prediction method (a model) [E]

Knowing what data is available for training, our data scientist is now able to choose the prediction method that will perform best against the chosen indicators of success. These indicators, such as the mathematical interpretation of the outcome, have been set in previous steps.

This step involves formulating a loss function - a mathematical formulation of the model's goal. It is composed of an error metric, which typically measures the average error of the model's predictions, and can include additional constraints, which nudge the model towards desired behaviors and away from worst-case scenarios (e.g. a constraint can be imposed to prevent a situation where errors are not mostly incurred by minority populations).

Model:
Logistic regression - well established in statistics, does not require making statistical assumptions, but will require additional work in preparing data.

Loss function:
Standard cross-entropy loss with L1 regularization (pushing the model to drop the least important elements of input completely).

or

Model:
A neural network - can sometimes produce more accurate predictions, requires less data pre-processing, but is not easily interpretable, and can generate counterintuitive predictions that are hard to trace in some cases.

Loss function:
Standard cross-entropy loss modified to include a fairness constraint, ensuring equal rates of correct predictions for men and women.

Testing and Calibrating the system [F]

The model has to be tested, using training data. Our data scientist will now look at the errors that have occured. Errors can be summarized by multiple metrics. At this stage data scientists choose metrics that are most relevant to evaluate the model. Based on test results, they decide which errors are acceptable (i.e. how harmful it would be if a certain type of error happened after deployment).

This step may or may not include consultations with the system owner and other stakeholders.

Asking the following questions:
- How often is our system correct?
- What types of people does it recommend for the program? Is this consistent with our expectations?
- Who is not included in predictions?
- Does the system recommend different populations (men and women, majority and minority groups) at similarly high rates?
- Are there any groups that are only classified as needy after exhibiting much higher need than other groups?
- Based on our answers, is the system fulfilling its desired goal? To what extent?

Updating the model [G]

If the test results are not satisfactory, this is the signal to redesign one or more components of the system. The decision regarding which component to update and how is most often controlled by the data scientist alone.

Let's add data from another hospital to include patients with more diverse socioeconomic background.

or

Let's change our most preferred outcome.
Let's add mathematical constraints in order to change model behavior.

or

Let's rethink the main objective!

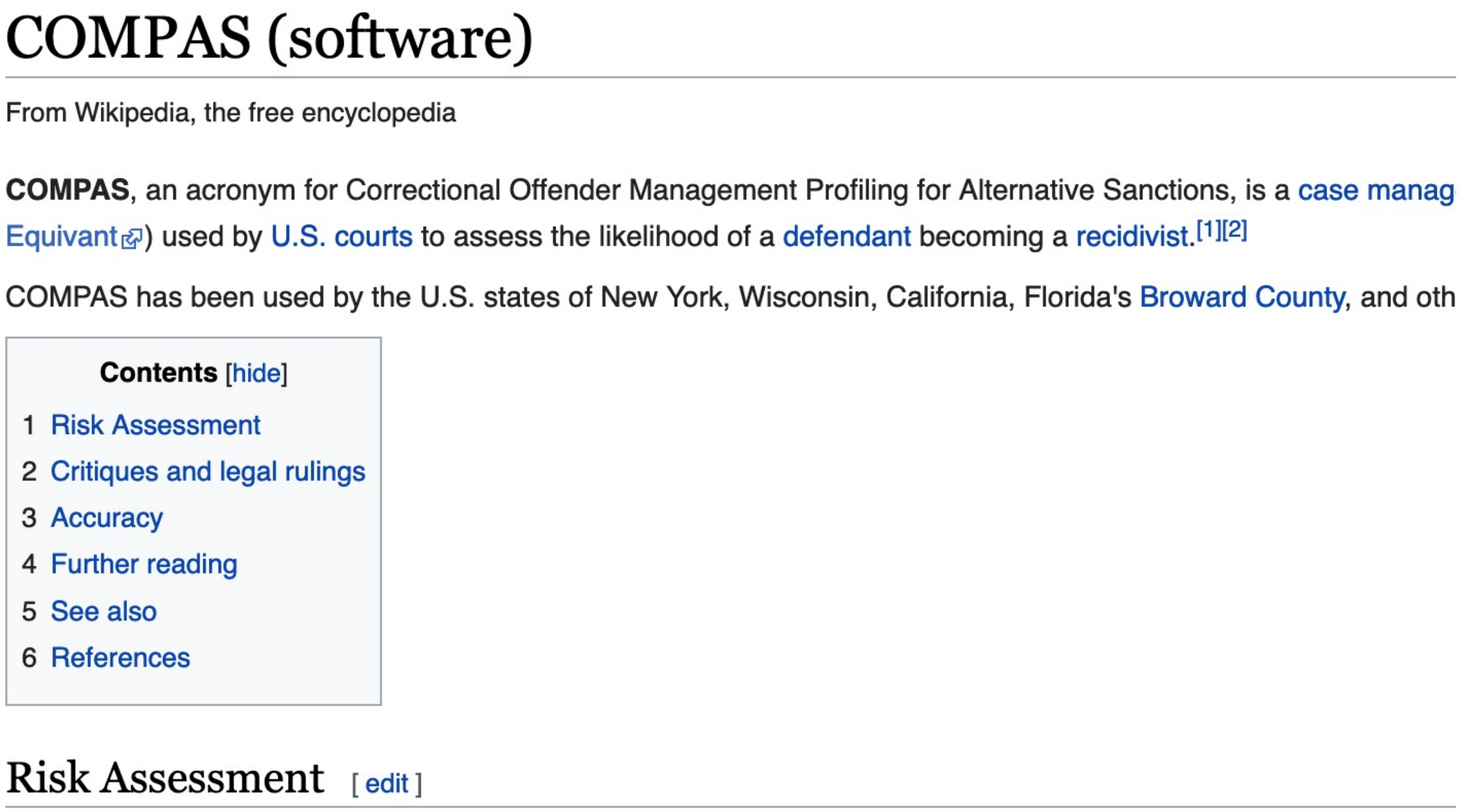
Evaluating before deployment

The decision to evaluate the system in real-life conditions before its deployment is not obligatory. In the absence of regulation, it is usually data scientists (rather than system owners) who decide whether and what tests to run, based on their own assumptions regarding what the system should do or shouldn't do.

By contrast, in more mature engineering fields, such as civil engineering, there is a well-defined set of required tests and measurements that have to be reported for a system.

Asking the following questions:
- Does the model perform as well in the new context as it did when we built it on our data?
- How does the system perform with regard to various groups of people (eg. different genders, socioeconomic status, age)?
- Are there errors that the users see which we did not catch?
- Is there any undesirable behavior we haven't anticipated?
- How does the model compare to human decisions? How is it different?

Have you heard of COMPAS?



COMPAS (software)

From Wikipedia, the free encyclopedia

COMPAS, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case manag

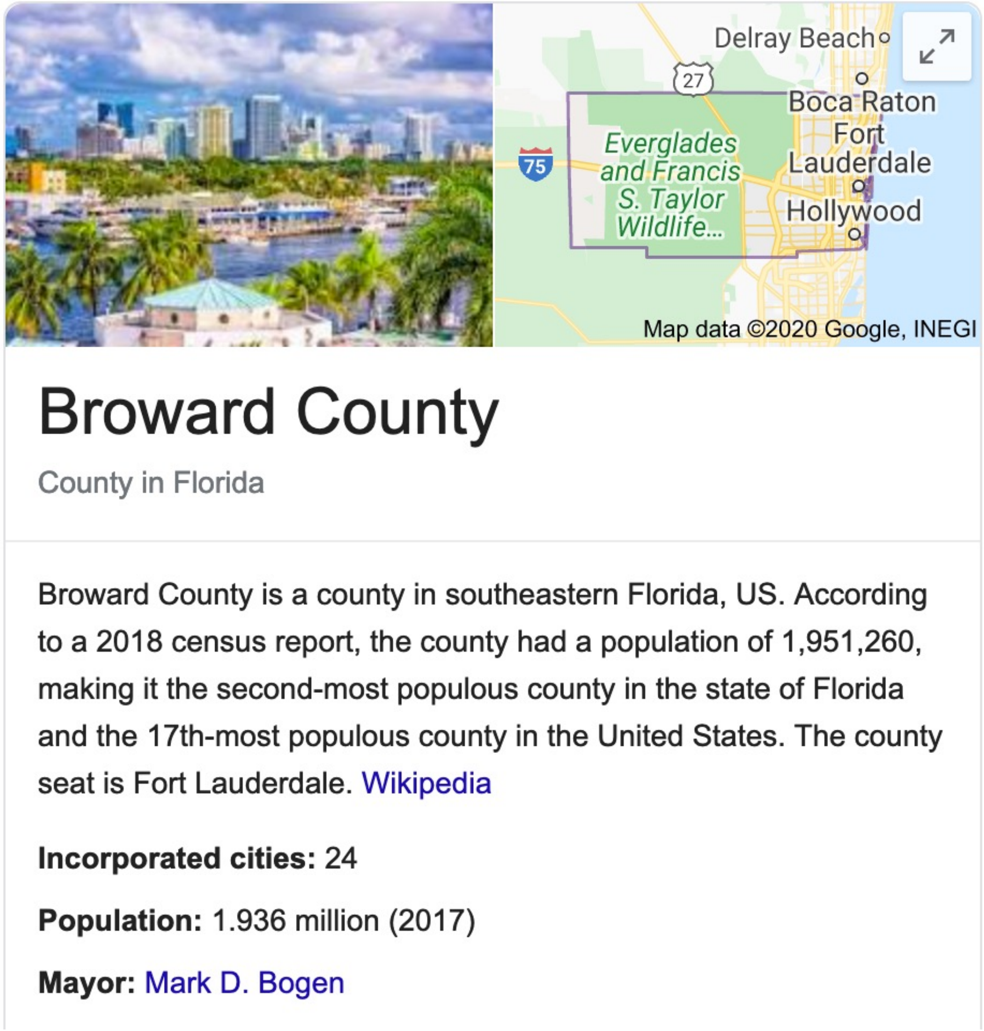
Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and oth

Contents [hide]

1 Risk Assessment
2 Critiques and legal rulings
3 Accuracy
4 Further reading
5 See also
6 References

Risk Assessment [edit]



Delray Beach
27
Boca Raton
75
Everglades and Francis S. Taylor Wildlife...
Fort Lauderdale
Hollywood
Map data ©2020 Google, INEGI

Broward County

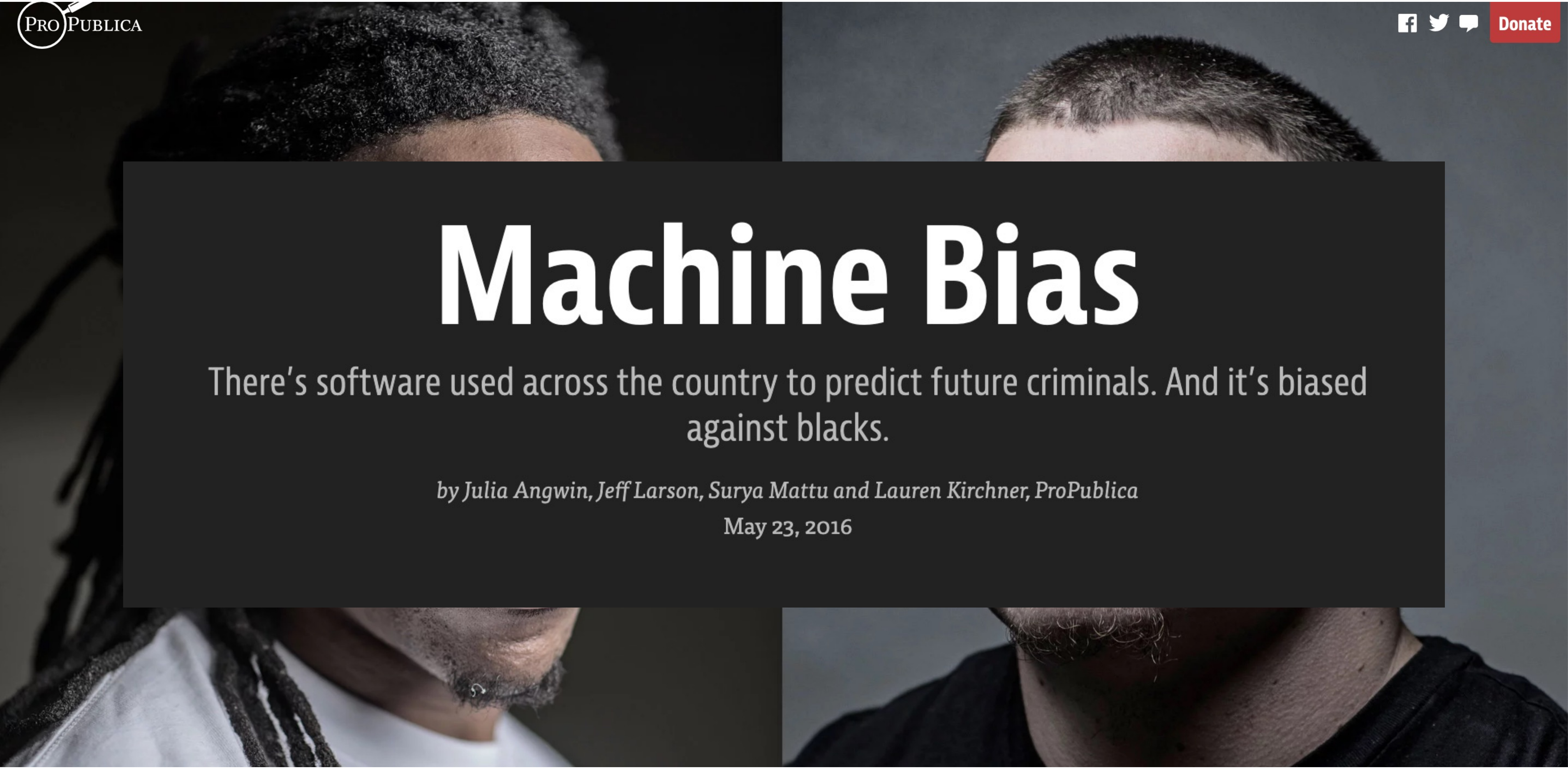
County in Florida

Broward County is a county in southeastern Florida, US. According to a 2018 census report, the county had a population of 1,951,260, making it the second-most populous county in the state of Florida and the 17th-most populous county in the United States. The county seat is Fort Lauderdale. Wikipedia

Incorporated cities: 24

Population: 1.936 million (2017)

Mayor: Mark D. Bogen



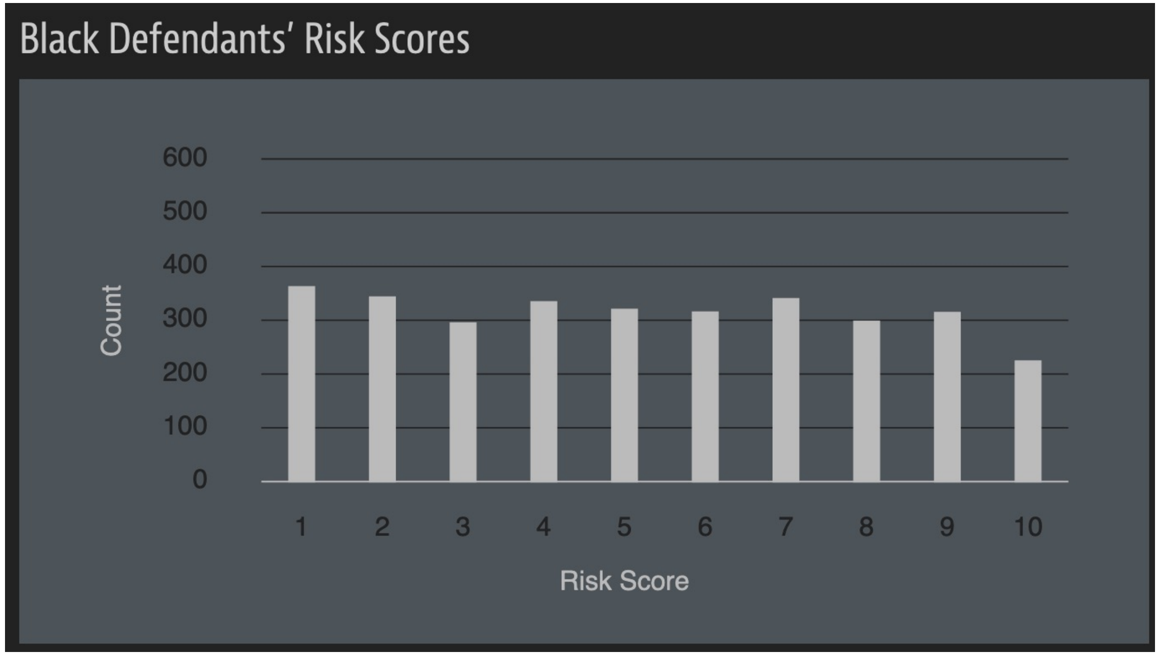
Machine Bias

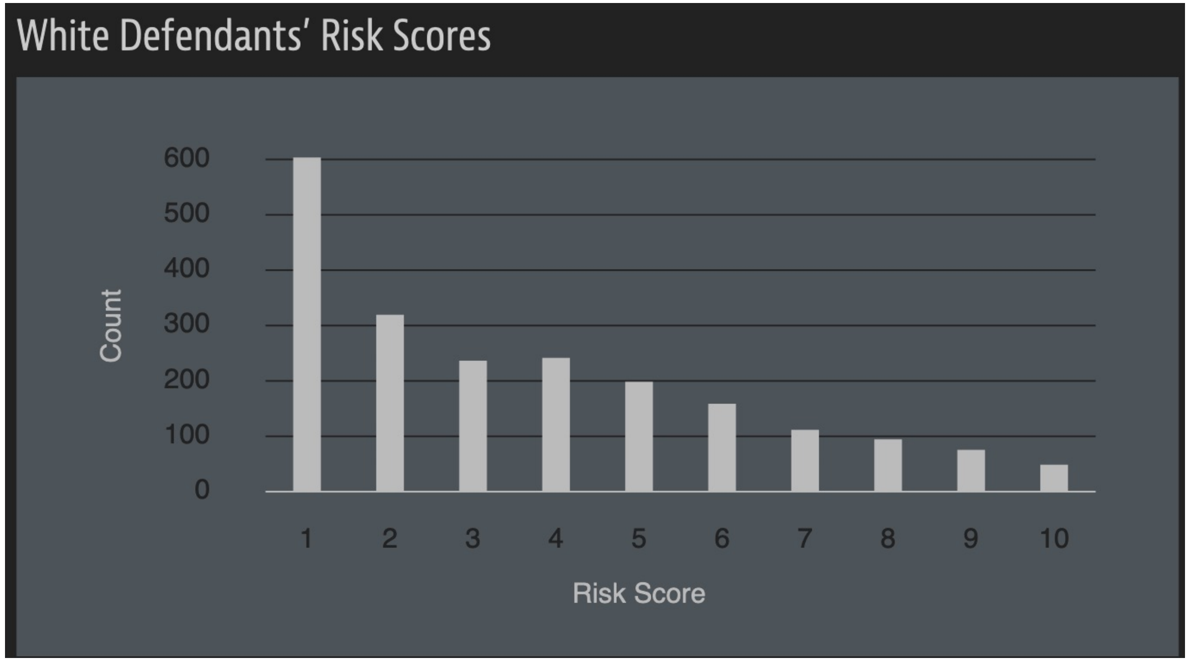
There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

A sample of their result

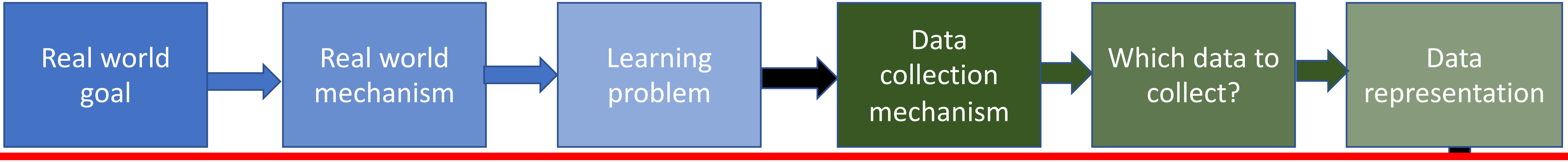




False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."

Anthony W. Flores
California State University, Bakersfield
Kristin Bechtel
Crime and Justice Institute at CRJ
Christopher T. Lowenkamp
Administrative Office of the United States Courts
Probation and Pretrial Services Office

A walkthrough



Real world goal

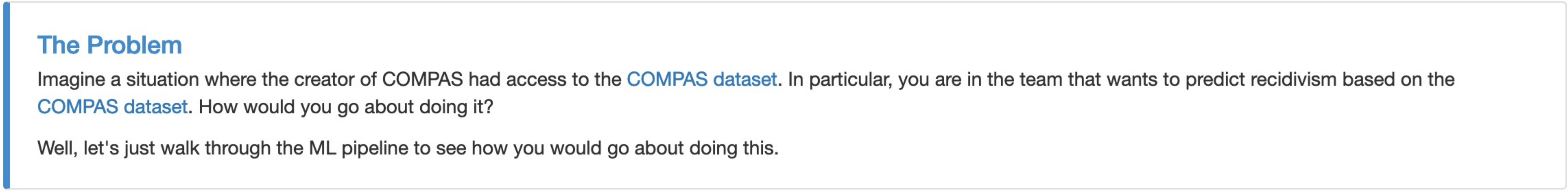
Real world mechanism

Learning problem

Data collection mechanism

Which data to collect?

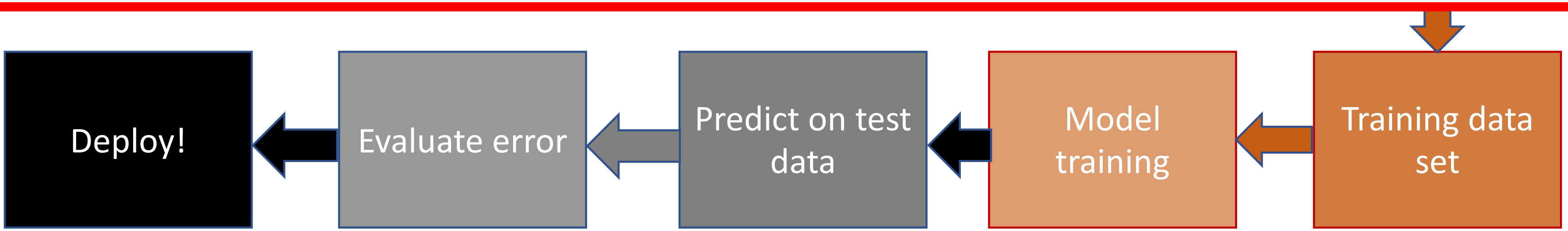
Data representation



The Problem

Imagine a situation where the creator of COMPAS had access to the COMPAS dataset. In particular, you are in the team that wants to predict recidivism based on the COMPAS dataset. How would you go about doing it?

Well, let's just walk through the ML pipeline to see how you would go about doing this.



Training data set

Model training

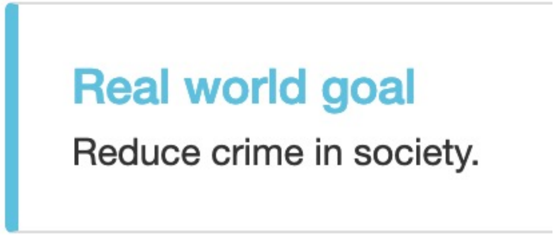
Predict on test data

Evaluate error

Deploy!

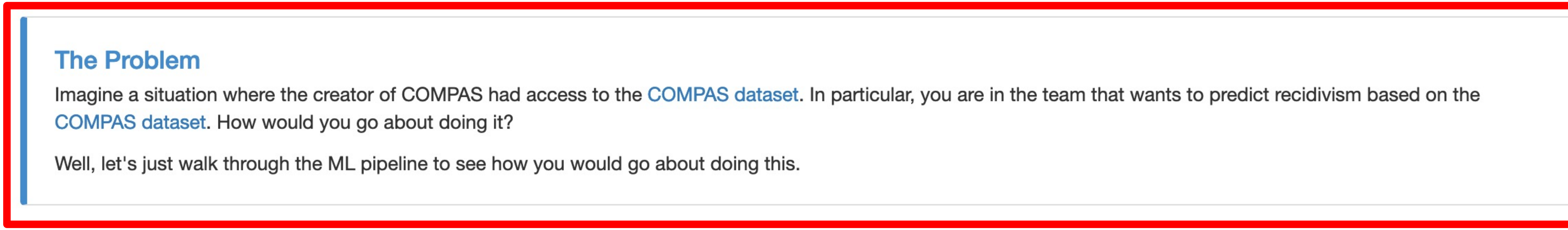
Real world goal

Real world goal



Real world goal

Reduce crime in society.

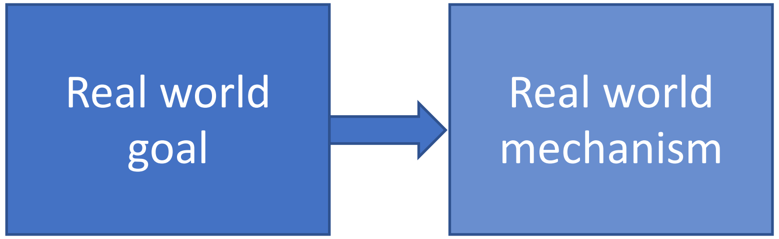


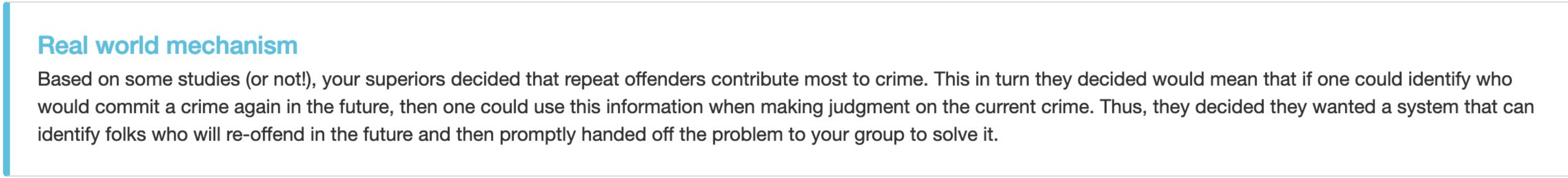
The Problem

Imagine a situation where the creator of COMPAS had access to the COMPAS dataset. In particular, you are in the team that wants to predict recidivism based on the COMPAS dataset. How would you go about doing it?

Well, let's just walk through the ML pipeline to see how you would go about doing this.

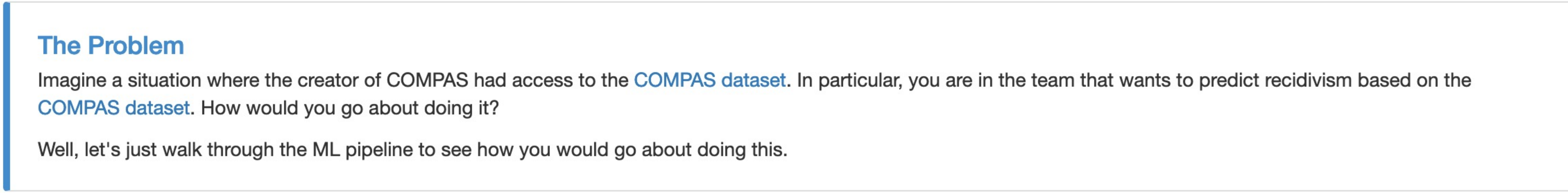
Real world mechanism





Real world mechanism

Based on some studies (or not!), your superiors decided that repeat offenders contribute most to crime. This in turn they decided would mean that if one could identify who would commit a crime again in the future, then one could use this information when making judgment on the current crime. Thus, they decided they wanted a system that can identify folks who will re-offend in the future and then promptly handed off the problem to your group to solve it.

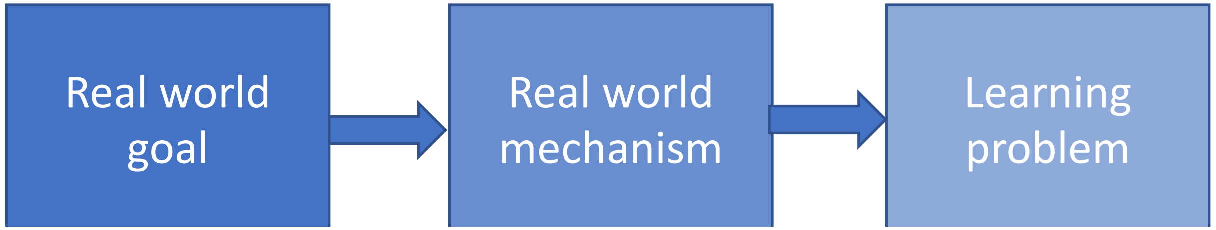


The Problem

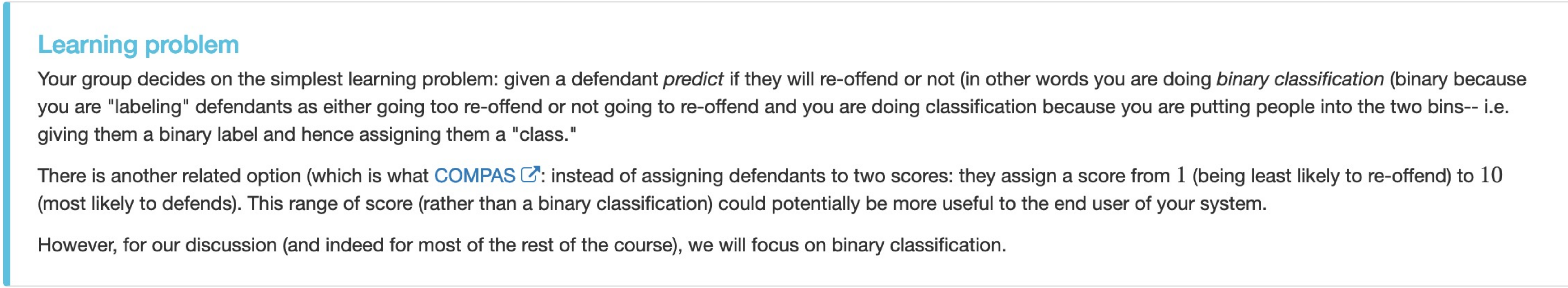
Imagine a situation where the creator of COMPAS had access to the COMPAS dataset. In particular, you are in the team that wants to predict recidivism based on the COMPAS dataset. How would you go about doing it?

Well, let's just walk through the ML pipeline to see how you would go about doing this.

Learning problem



Target variable might not be obvious in other scenarios!

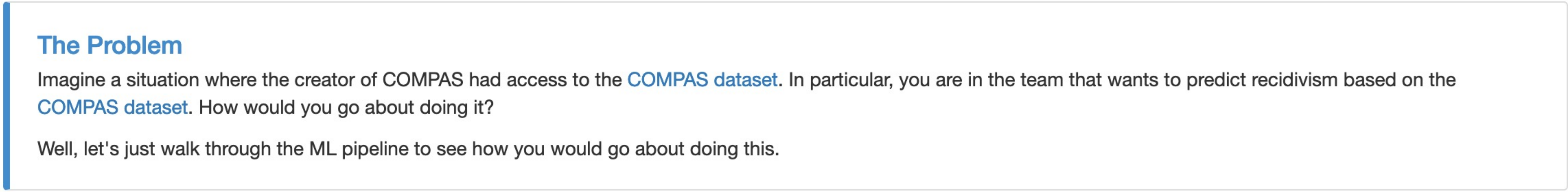


Learning problem

Your group decides on the simplest learning problem: given a defendant *predict* if they will re-offend or not (in other words you are doing *binary classification* (binary because you are "labeling" defendants as either going too re-offend or not going to re-offend and you are doing classification because you are putting people into the two bins-- i.e. giving them a binary label and hence assigning them a "class."

There is another related option (which is what COMPAS: instead of assigning defendants to two scores: they assign a score from 1 (being least likely to re-offend) to 10 (most likely to defends). This range of score (rather than a binary classification) could potentially be more useful to the end user of your system.

However, for our discussion (and indeed for most of the rest of the course), we will focus on binary classification.

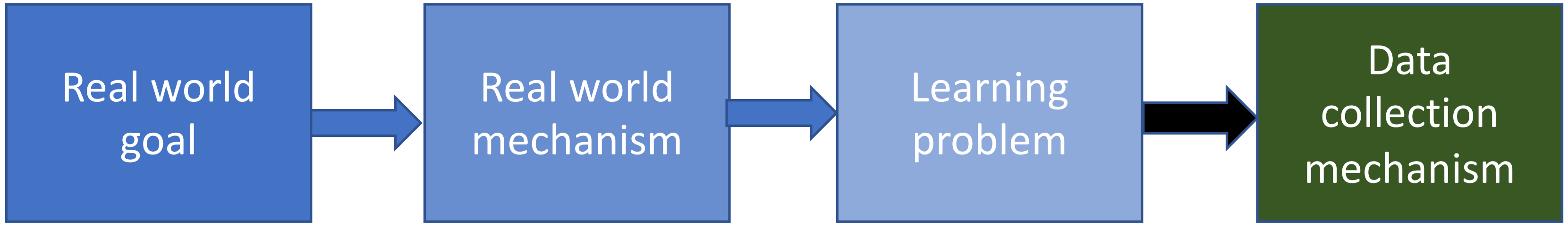


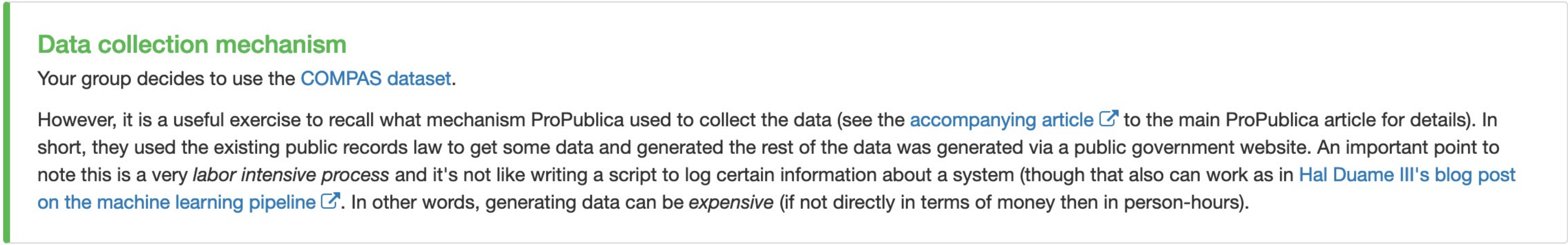
The Problem

Imagine a situation where the creator of COMPAS had access to the COMPAS dataset. In particular, you are in the team that wants to predict recidivism based on the COMPAS dataset. How would you go about doing it?

Well, let's just walk through the ML pipeline to see how you would go about doing this.

Data collection mechanism

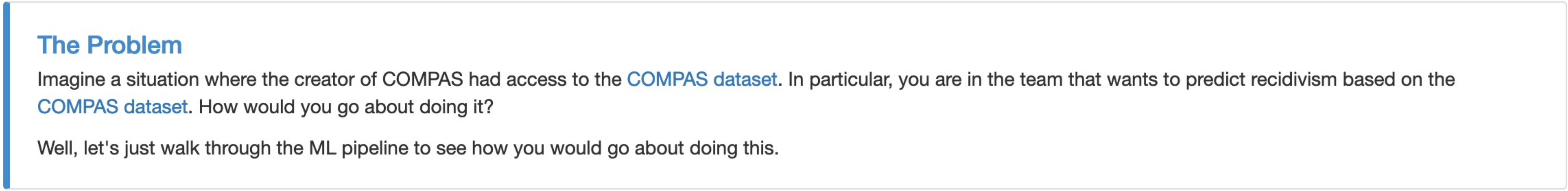




Data collection mechanism

Your group decides to use the COMPAS dataset.

However, it is a useful exercise to recall what mechanism ProPublica used to collect the data (see the accompanying article to the main ProPublica article for details). In short, they used the existing public records law to get some data and generated the rest of the data was generated via a public government website. An important point to note this is a very *labor intensive process* and it's not like writing a script to log certain information about a system (though that also can work as in Hal Duame III's blog post on the machine learning pipeline). In other words, generating data can be *expensive* (if not directly in terms of money then in person-hours).

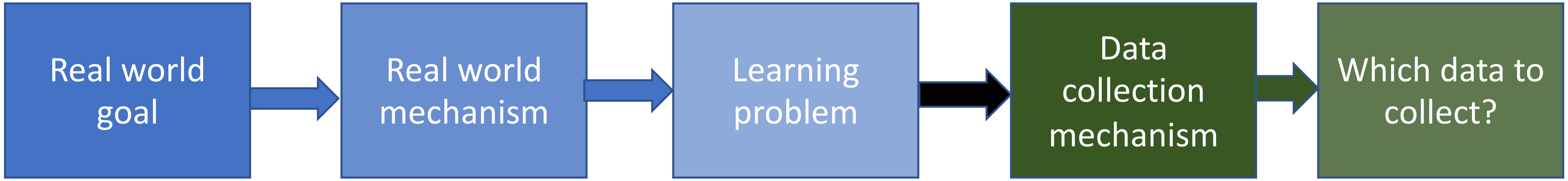


The Problem

Imagine a situation where the creator of COMPAS had access to the COMPAS dataset. In particular, you are in the team that wants to predict recidivism based on the COMPAS dataset. How would you go about doing it?

Well, let's just walk through the ML pipeline to see how you would go about doing this.

Which data to collect?



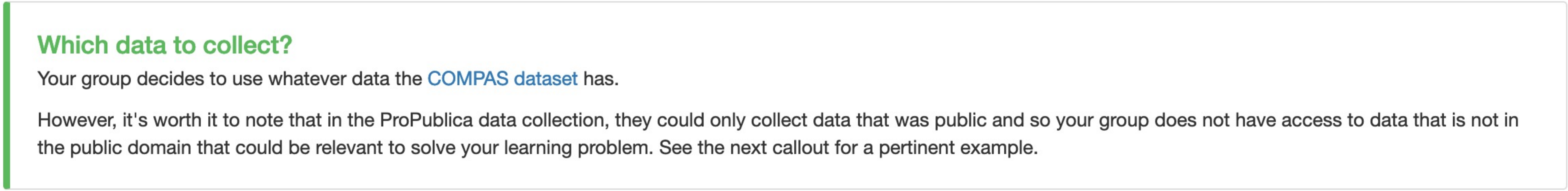
Real world goal

Real world mechanism

Learning problem

Data collection mechanism

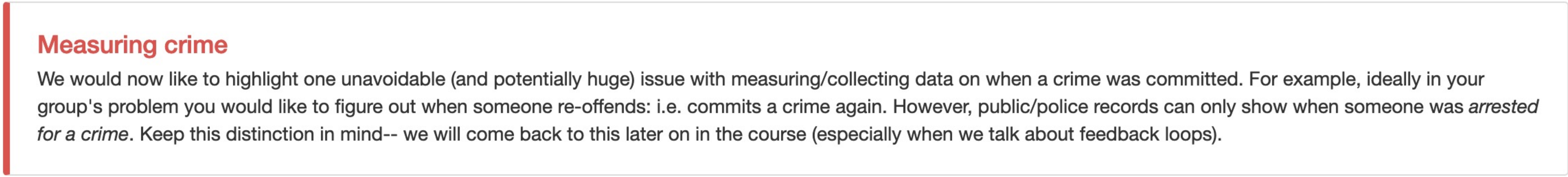
Which data to collect?



Which data to collect?

Your group decides to use whatever data the COMPAS dataset has.

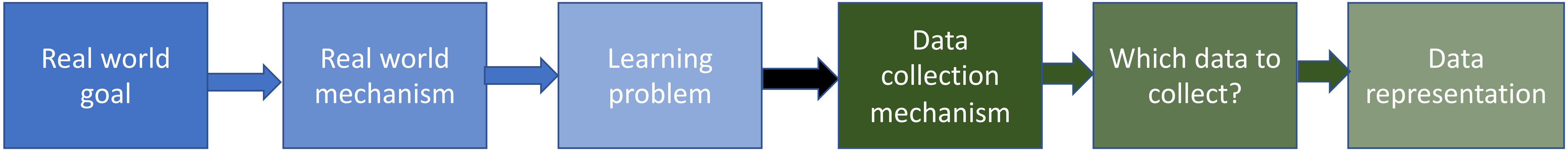
However, it's worth it to note that in the ProPublica data collection, they could only collect data that was public and so your group does not have access to data that is not in the public domain that could be relevant to solve your learning problem. See the next callout for a pertinent example.

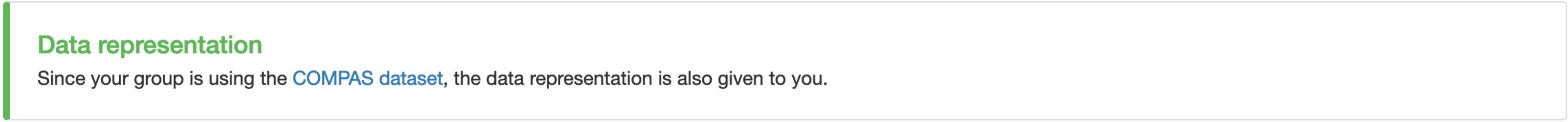


Measuring crime

We would now like to highlight one unavoidable (and potentially huge) issue with measuring/collecting data on when a crime was committed. For example, ideally in your group's problem you would like to figure out when someone re-offends: i.e. commits a crime again. However, public/police records can only show when someone was *arrested for a crime*. Keep this distinction in mind-- we will come back to this later on in the course (especially when we talk about feedback loops).

Data representation

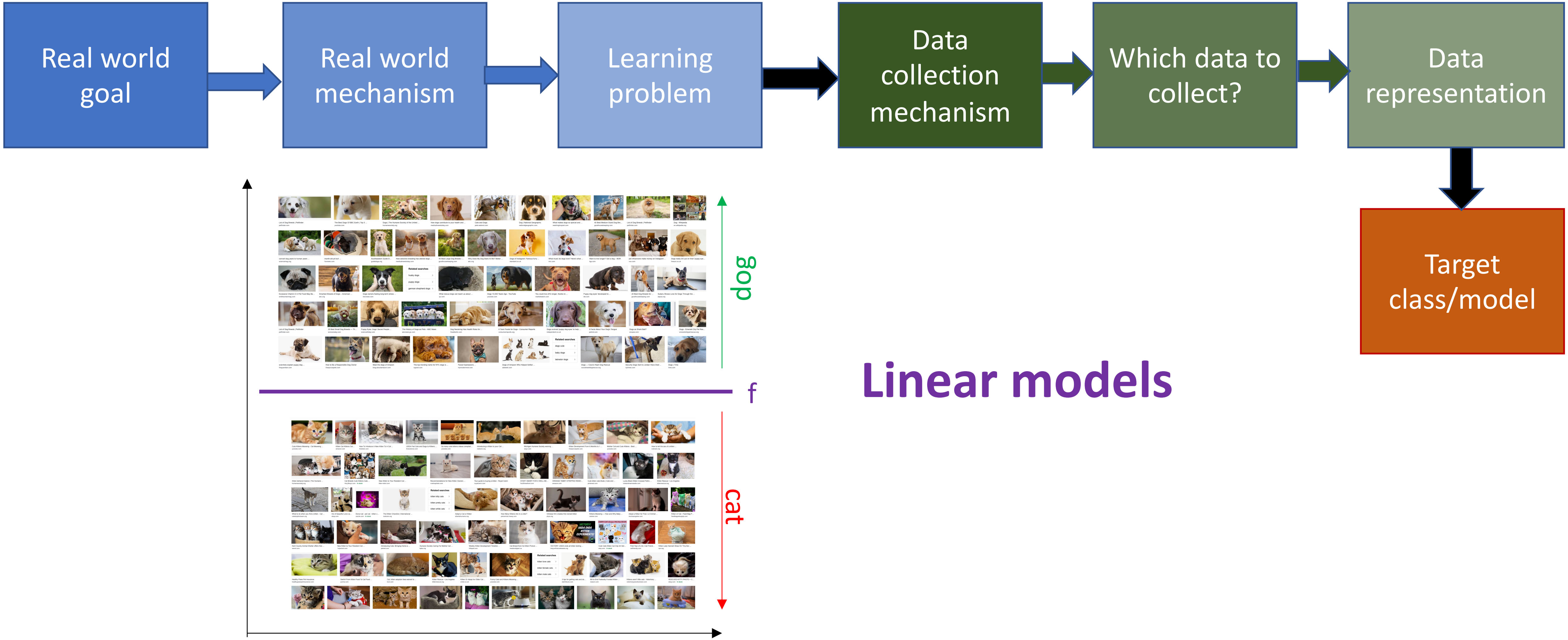




Data representation

Since your group is using the COMPAS dataset, the data representation is also given to you.

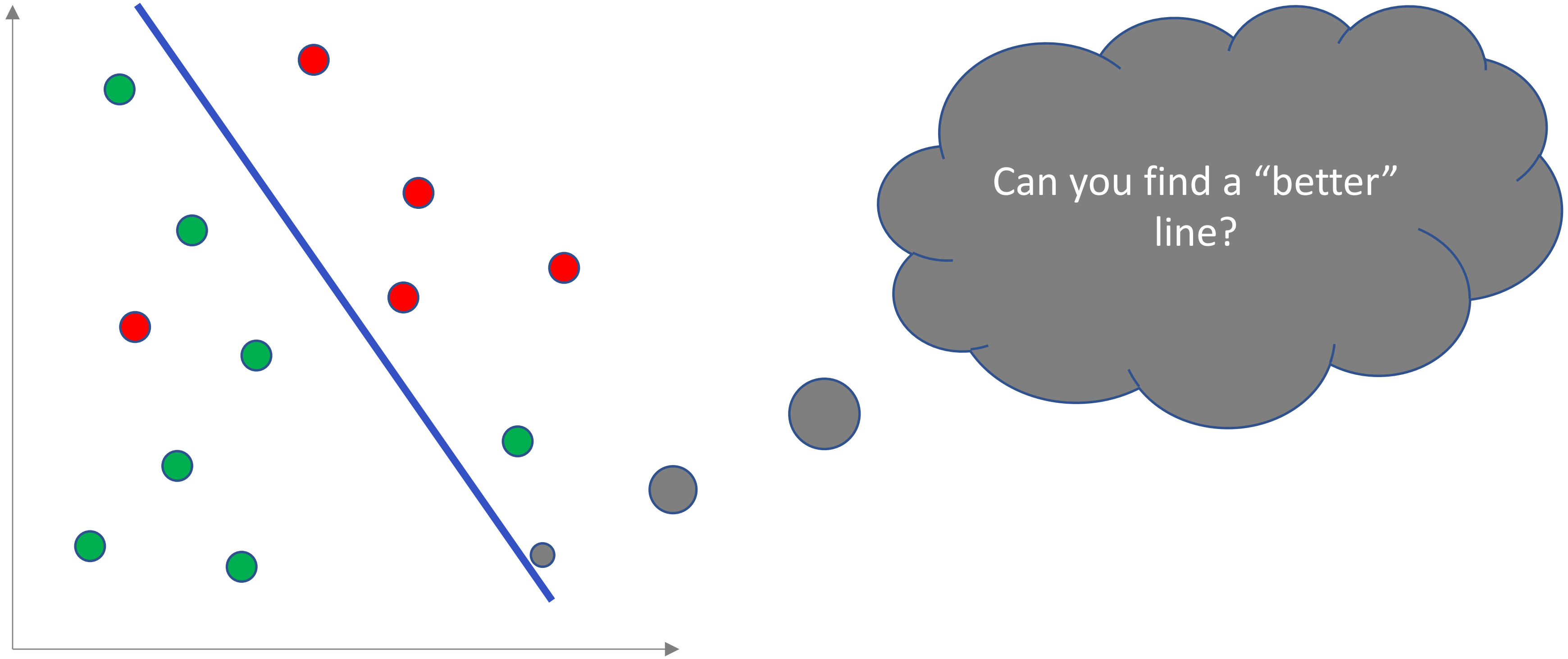
Target class/model



Real world goal → Real world mechanism → Learning problem → Data collection mechanism → Which data to collect? → Data representation → Target class/model

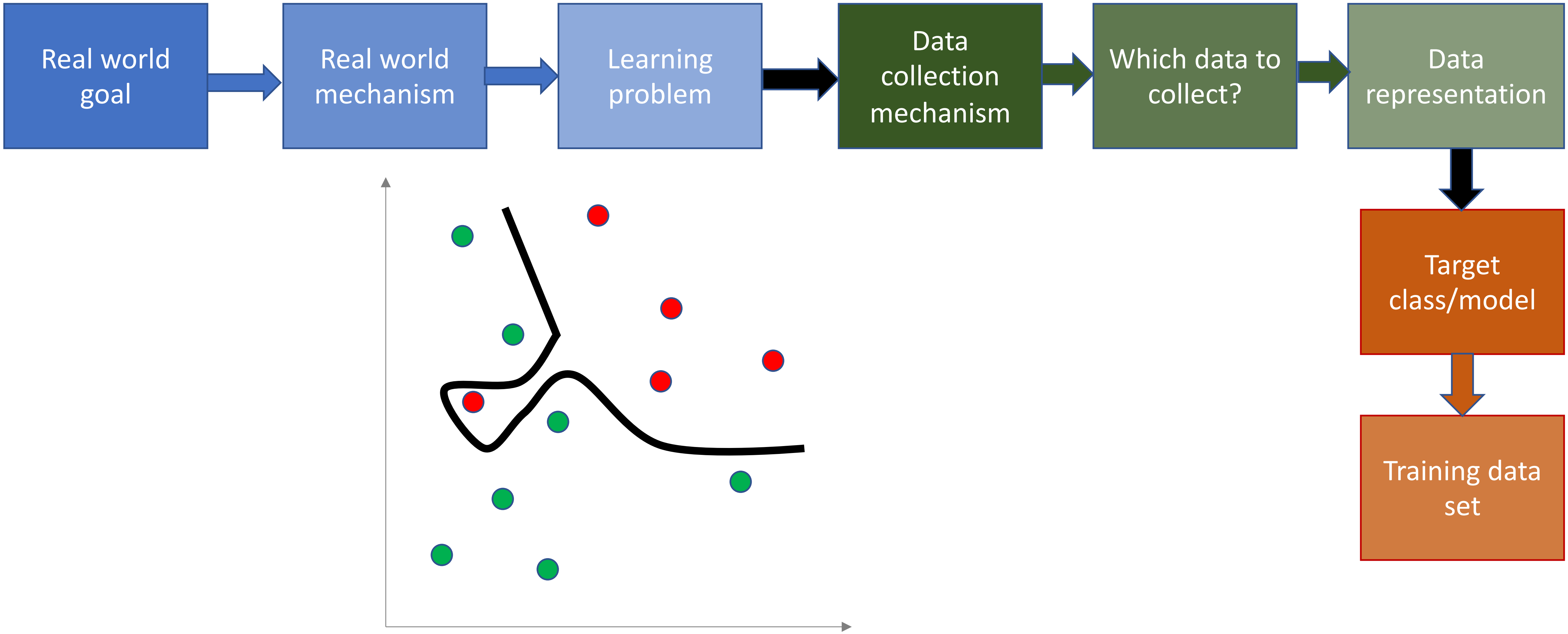
Linear models

Cartoon of how linear model works



Can you find a “better” line?

Training data set



Real world goal

Real world mechanism

Learning problem

Data collection mechanism

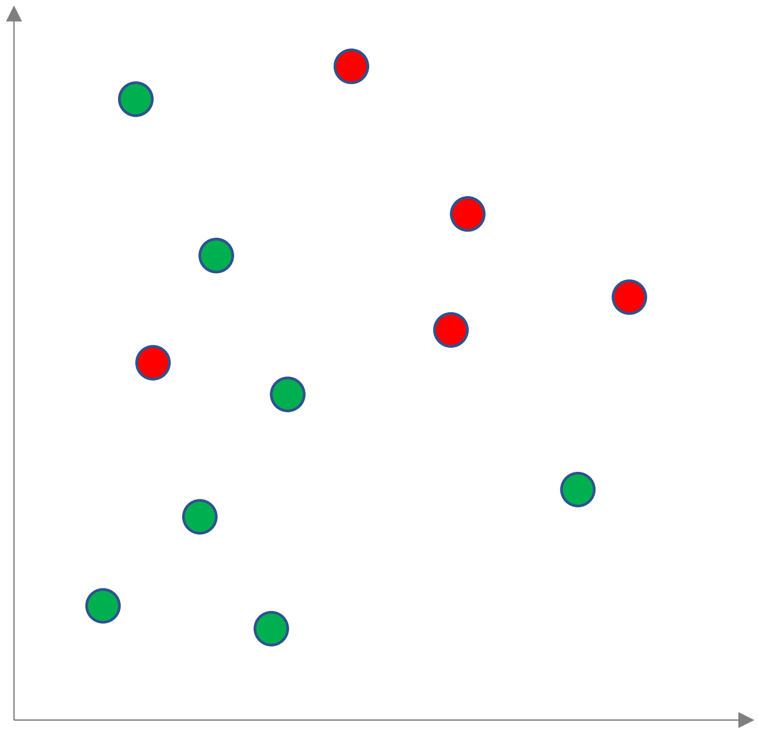
Which data to collect?

Data representation

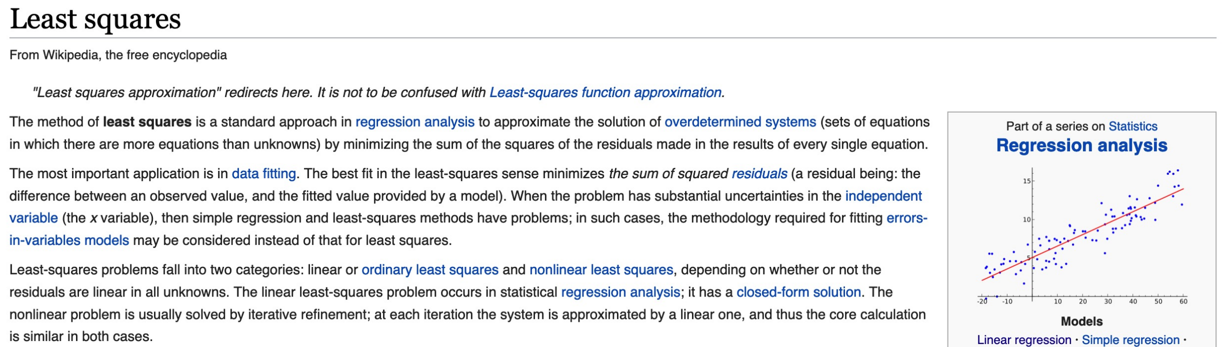
Target class/model

Training data set

Pick (random) half of the dataset



Model training



Least squares

From Wikipedia, the free encyclopedia

"Least squares approximation" redirects here. It is not to be confused with Least-squares function approximation.

The method of **least squares** is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimizing the sum of the squares of the residuals made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes *the sum of squared residuals* (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the *x* variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

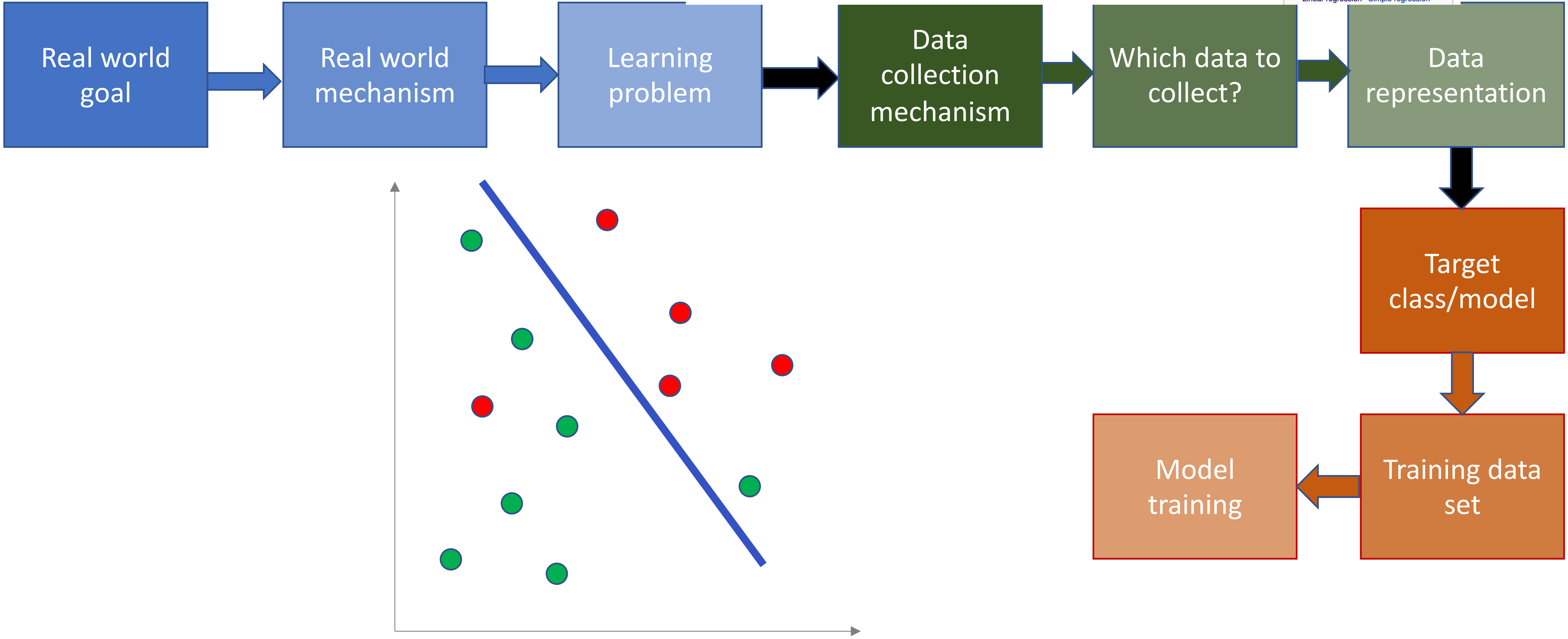
Least-squares problems fall into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The nonlinear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Part of a series on Statistics

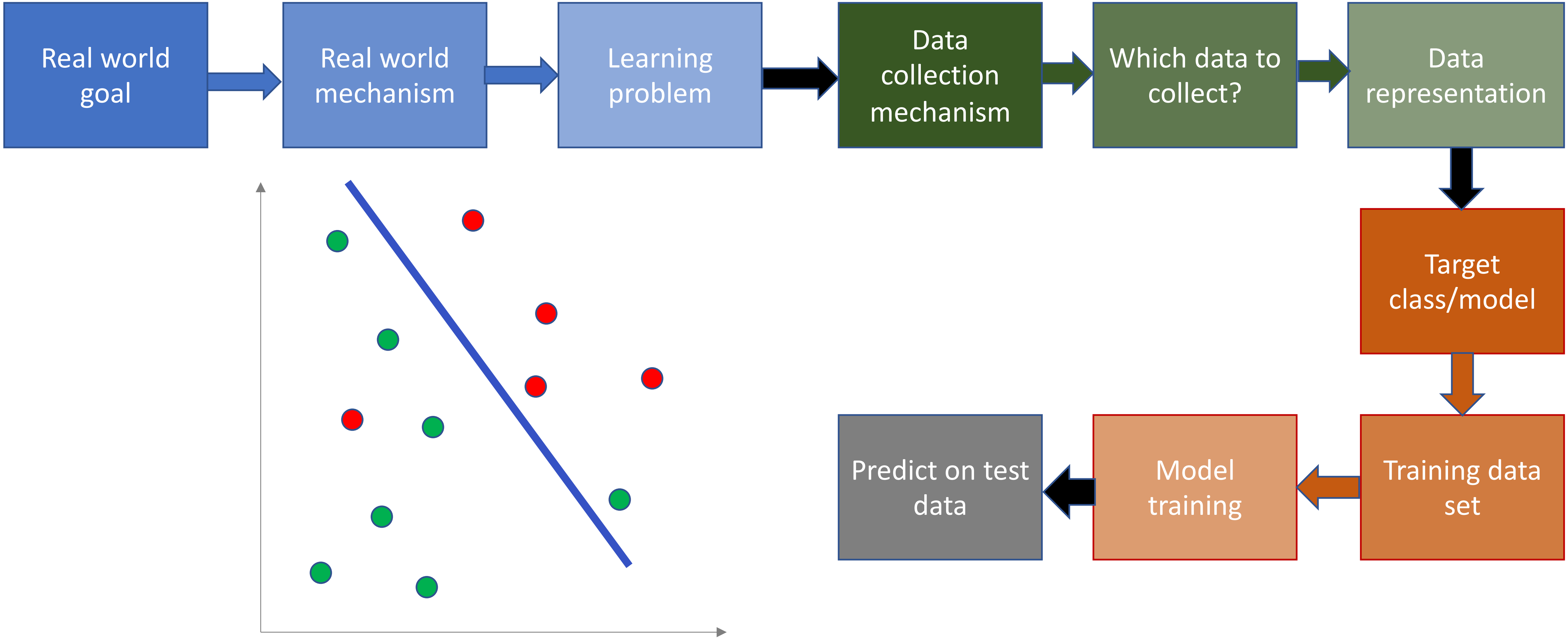
Regression analysis

Models

Linear regression · Simple regression ·



Predict on test data



Real world goal

Real world mechanism

Learning problem

Data collection mechanism

Which data to collect?

Data representation

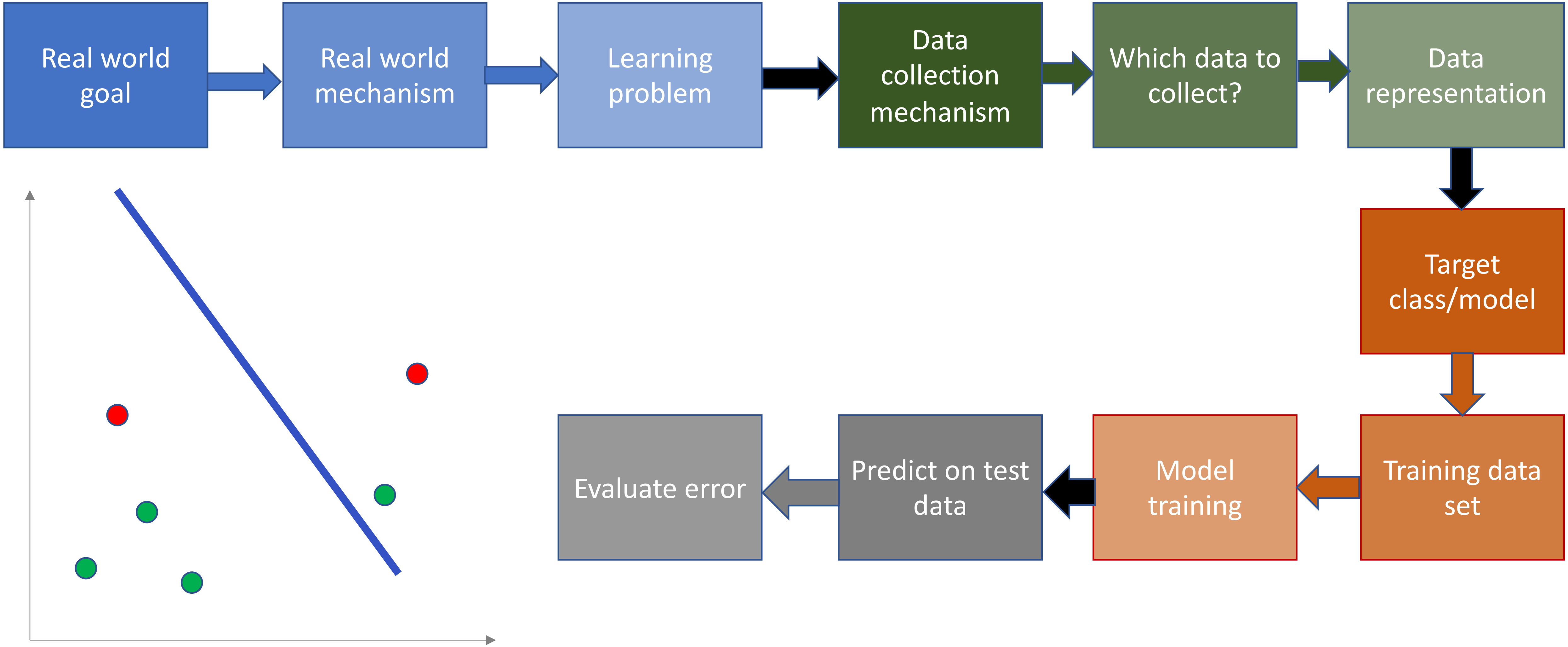
Target class/model

Training data set

Model training

Predict on test data

Evaluate error



Real world goal

Real world mechanism

Learning problem

Data collection mechanism

Which data to collect?

Data representation

Target class/model

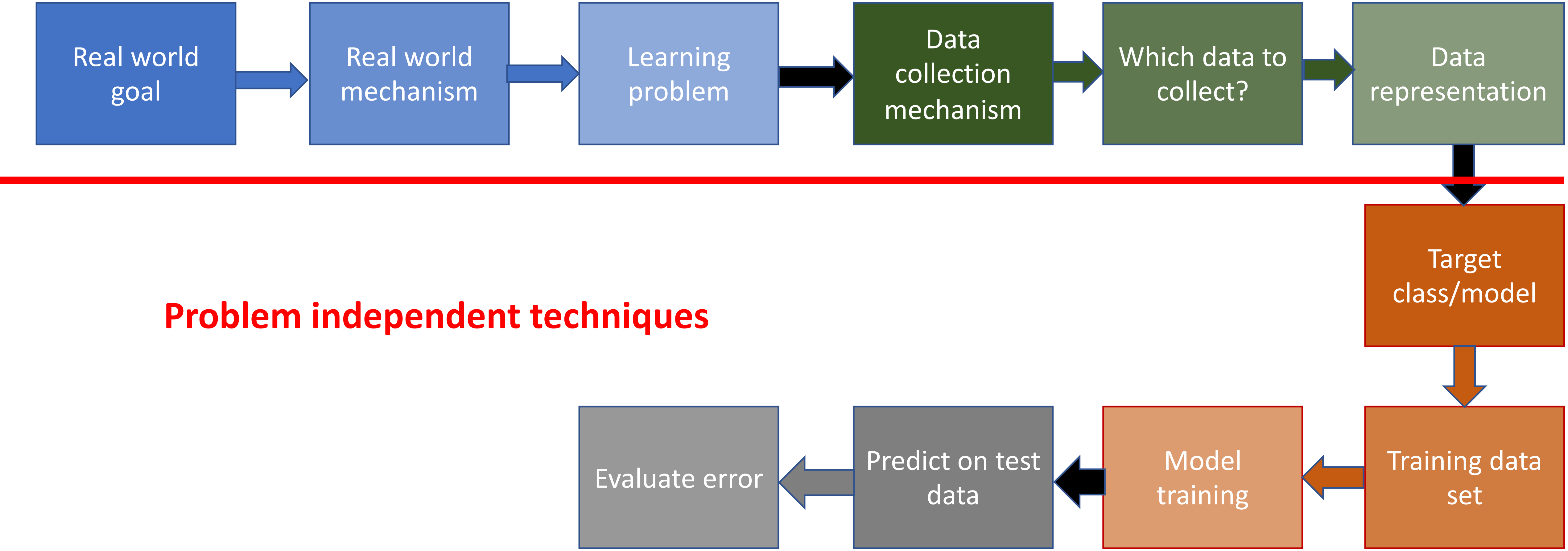
Training data set

Model training

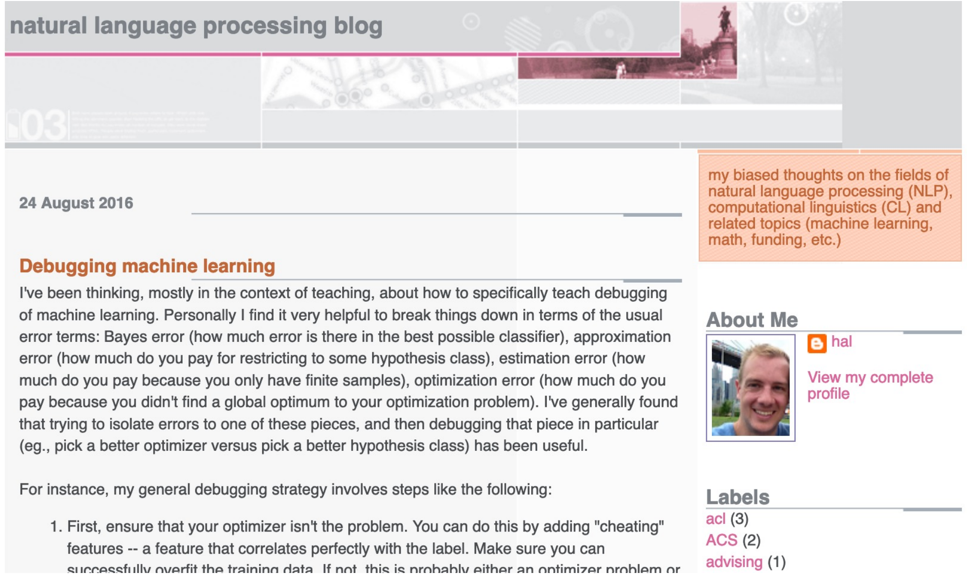
Predict on test data

Evaluate error

Relation to problem statement



Three problems to consider



natural language processing blog

my biased thoughts on the fields of natural language processing (NLP), computational linguistics (CL) and related topics (machine learning, math, funding, etc.)

24 August 2016

Debugging machine learning

I've been thinking, mostly in the context of teaching, about how to specifically teach debugging of machine learning. Personally I find it very helpful to break things down in terms of the usual error terms: Bayes error (how much error is there in the best possible classifier), approximation error (how much do you pay for restricting to some hypothesis class), estimation error (how much do you pay because you only have finite samples), optimization error (how much do you pay because you didn't find a global optimum to your optimization problem). I've generally found that trying to isolate errors to one of these pieces, and then debugging that piece in particular (eg., pick a better optimizer versus pick a better hypothesis class) has been useful.

For instance, my general debugging strategy involves steps like the following:

1. First, ensure that your optimizer isn't the problem. You can do this by adding "cheating" features -- a feature that correlates perfectly with the label. Make sure you can successfully overfit the training data. If not, this is probably either an optimizer problem or

About Me

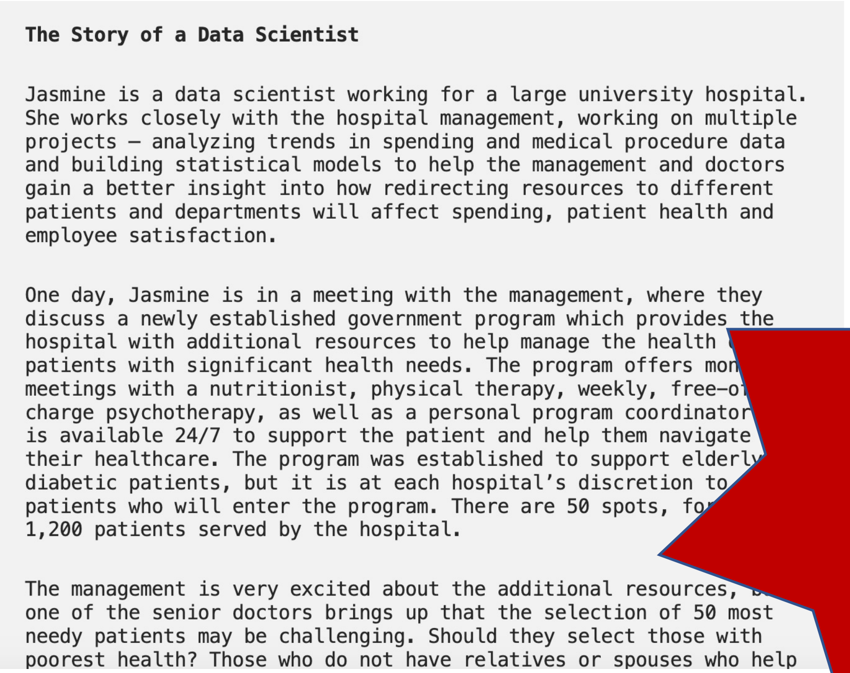
hal

View my complete profile

Labels

acl (3)
ACS (2)
advising (1)

Ad display example



The Story of a Data Scientist

Jasmine is a data scientist working for a large university hospital. She works closely with the hospital management, working on multiple projects – analyzing trends in spending and medical procedure data and building statistical models to help the management and doctors gain a better insight into how redirecting resources to different patients and departments will affect spending, patient health and employee satisfaction.

One day, Jasmine is in a meeting with the management, where they discuss a newly established government program which provides the hospital with additional resources to help manage the health of patients with significant health needs. The program offers mon meetings with a nutritionist, physical therapy, weekly, free-of charge psychotherapy, as well as a personal program coordinator is available 24/7 to support the patient and help them navigate their healthcare. The program was established to support elderly diabetic patients, but it is at each hospital's discretion to patients who will enter the program. There are 50 spots, for 1,200 patients served by the hospital.

The management is very excited about the additional resources, b one of the senior doctors brings up that the selection of 50 most needy patients may be challenging. Should they select those with poorest health? Those who do not have relatives or spouses who help

After the break

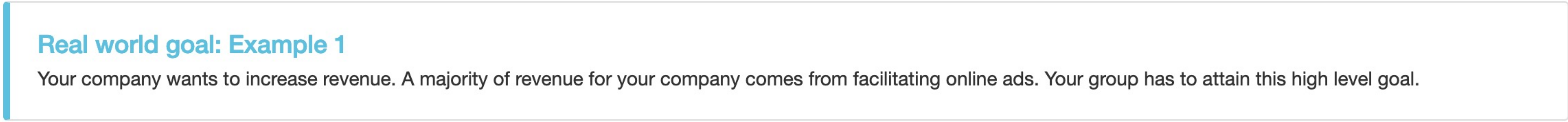
Hospital trying to utilize new govt program

Third example: **YOU** decide!

Pass phrase for today: Kate Crawford

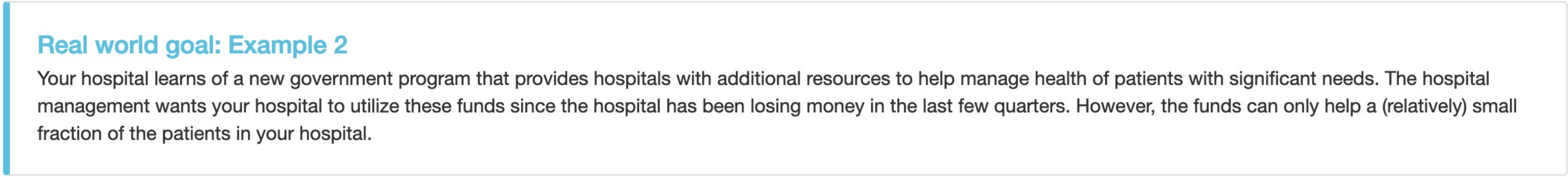
Real world goal

Real world goal



Real world goal: Example 1

Your company wants to increase revenue. A majority of revenue for your company comes from facilitating online ads. Your group has to attain this high level goal.



Real world goal: Example 2

Your hospital learns of a new government program that provides hospitals with additional resources to help manage health of patients with significant needs. The hospital management wants your hospital to utilize these funds since the hospital has been losing money in the last few quarters. However, the funds can only help a (relatively) small fraction of the patients in your hospital.

Real world goal: Your choice

Real world goal

Group 1: How do you pick limited number of students that'll benefit from a course

Group 2: Matching students to colleges where they have the best chance to succeed

Group 3: Given a budget how should a company best advertise to get max benefit

Real world goal: Choices from Spring 2020

Real world goal

Improve user experience for Stampede buses: 1

Figure out which laptop to buy?: 3

Figure out where to live as a student: 3

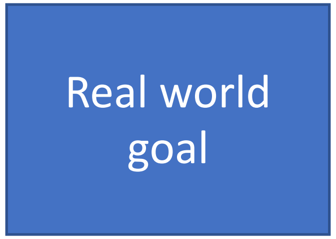
Figure out how to improve student’s class experience: 2

Improve students’ success in getting scholarships: 4

Improve on-campus parking: 3

Figure out how to improve instructor’s class experience: 2

Real world goal: General thoughts

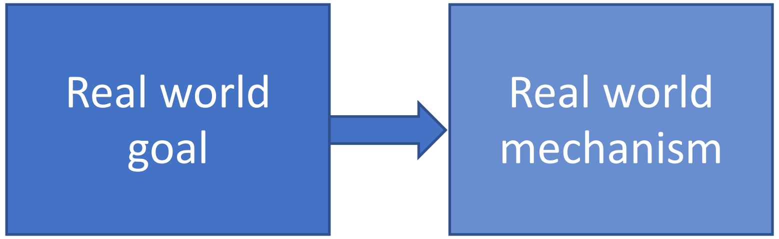


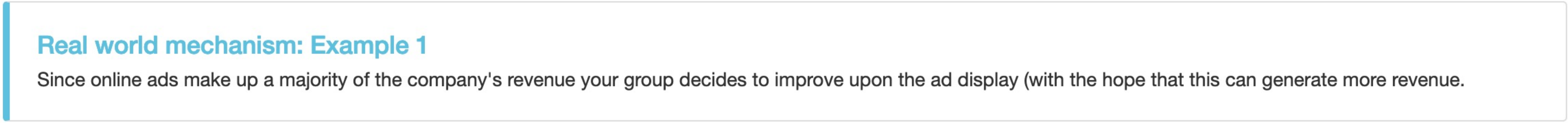
Real world goal

This step generally done at higher management level

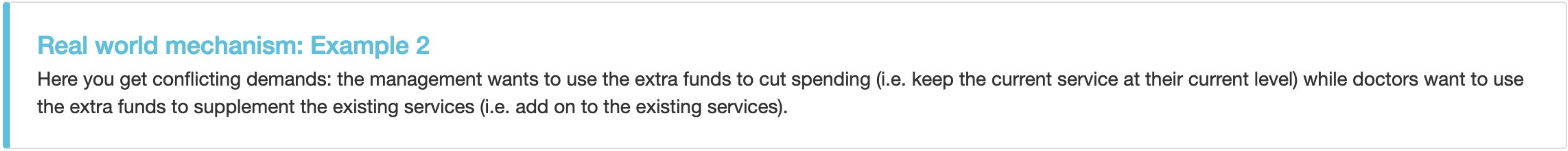
Translating this into something concrete needs remaining steps

Real world mechanism



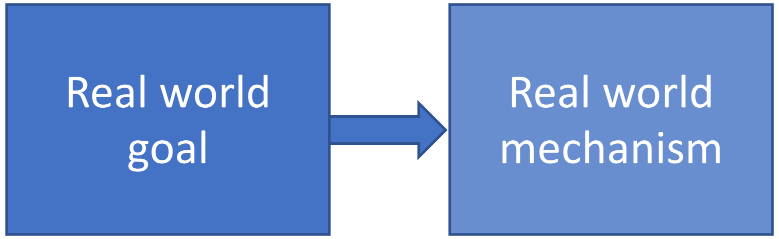

Real world mechanism: Example 1

Since online ads make up a majority of the company's revenue your group decides to improve upon the ad display (with the hope that this can generate more revenue.


Real world mechanism: Example 2

Here you get conflicting demands: the management wants to use the extra funds to cut spending (i.e. keep the current service at their current level) while doctors want to use the extra funds to supplement the existing services (i.e. add on to the existing services).

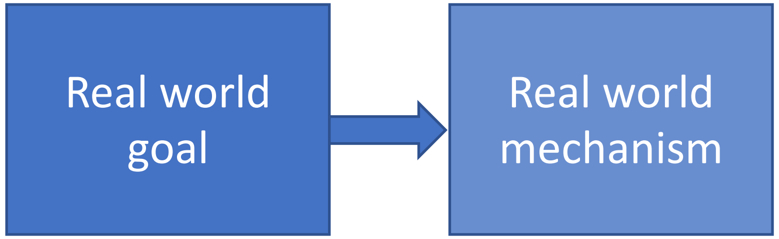
Real world mechanism: Your choice



Choice 1: Benefit = number of degree requirements satisfied

Choice 2: Benefit = Number of options opened for future prospects

Real world mechanism: Spring 2020 choice



Real world goal

Real world mechanism

Improve students' success in getting external scholarships

Improve student access to mentor (to help with the application process): 4

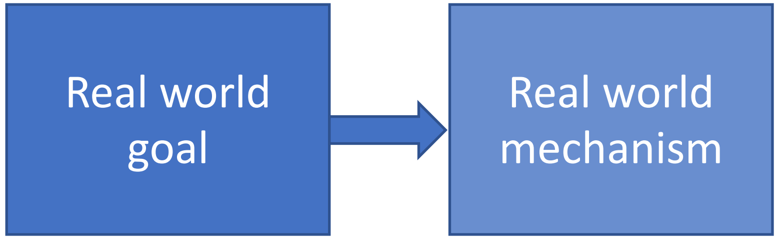
Improve student awareness of existing scholarships: 4

Identify students who are likely to win a scholarship

Motivate students to actually apply: 2

Make it easier to get scholarships: 4

Real world mechanism: General thoughts


Real world goal

Real world mechanism

Is improving ad display the best option?

ALWAYS question if the mechanism captures well the real life goal

There can be competing/incompatible mechanisms

Doctors vs. management in Example 2

CONVENIENCE trap!