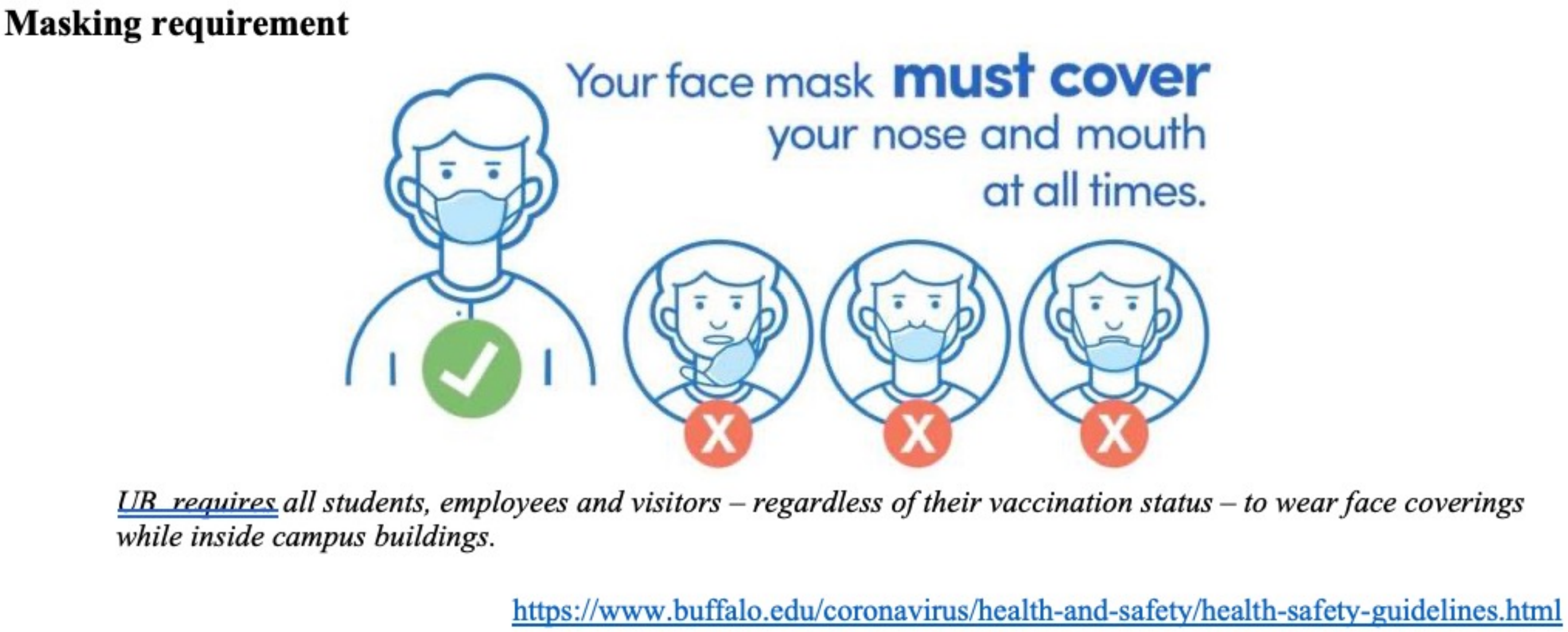
ML and Society

Feb 15, 2022

Please have a face mask on



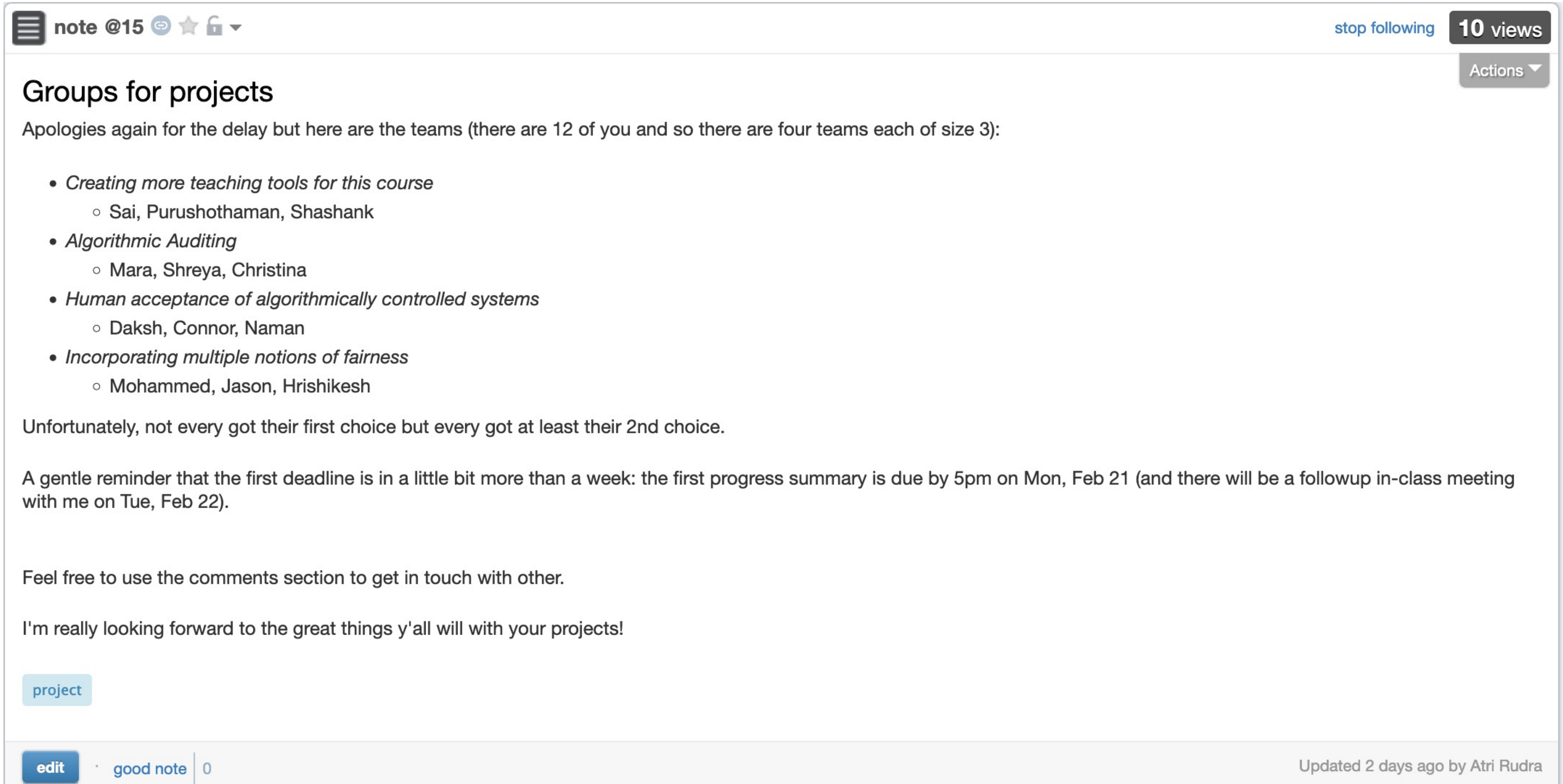
Masking requirement

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

Project groups formed



Groups for projects

Apologies again for the delay but here are the teams (there are 12 of you and so there are four teams each of size 3):

- *Creating more teaching tools for this course*
    - Sai, Purushothaman, Shashank
- *Algorithmic Auditing*
    - Mara, Shreya, Christina
- *Human acceptance of algorithmically controlled systems*
    - Daksh, Connor, Naman
- *Incorporating multiple notions of fairness*
    - Mohammed, Jason, Hrishikesh

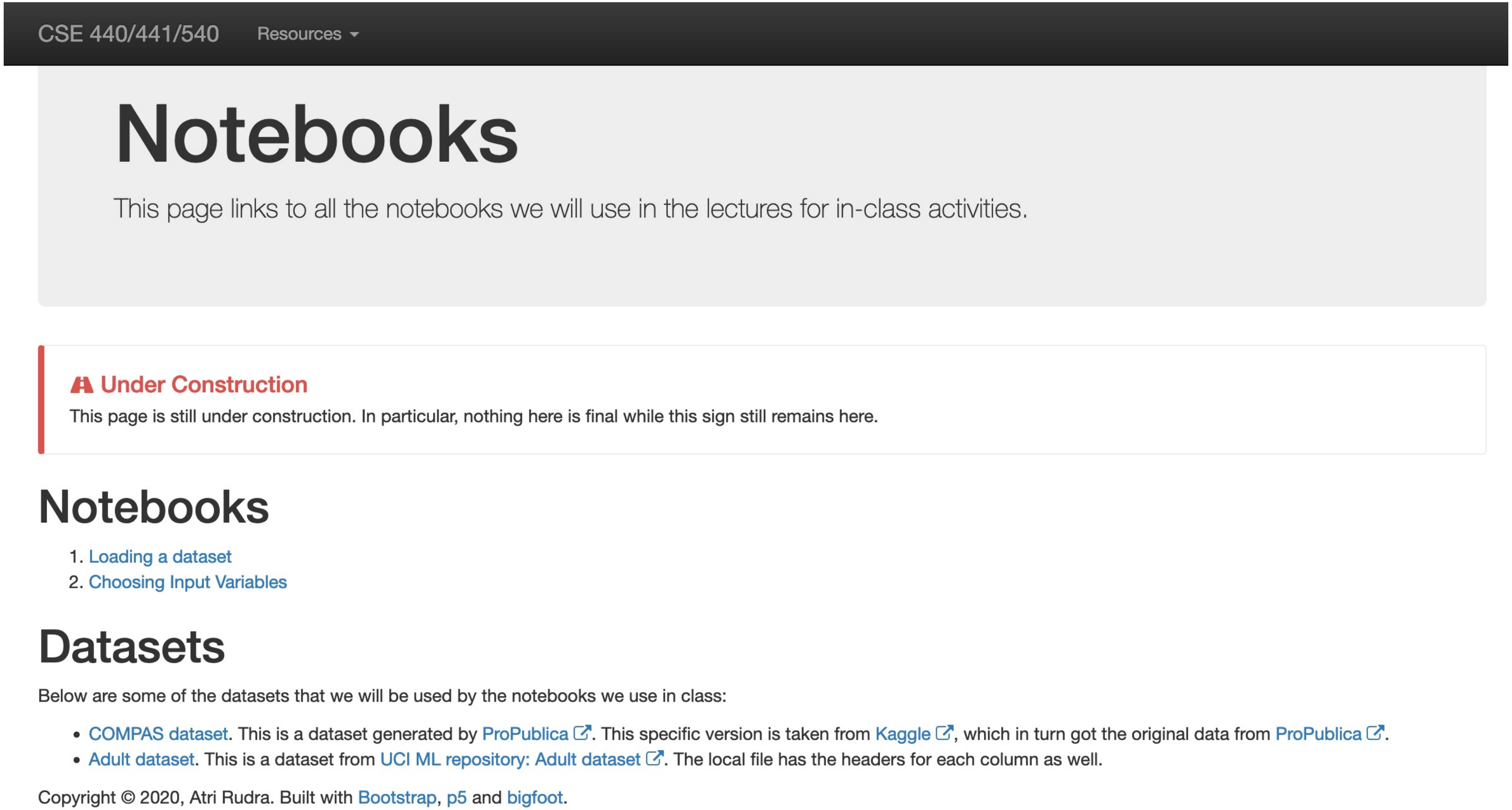
Unfortunately, not every got their first choice but every got at least their 2nd choice.

A gentle reminder that the first deadline is in a little bit more than a week: the first progress summary is due by 5pm on Mon, Feb 21 (and there will be a followup in-class meeting with me on Tue, Feb 22).

Feel free to use the comments section to get in touch with other.

I'm really looking forward to the great things y'all will with your projects!

Some Jupyter notebooks have been added



Notebooks

This page links to all the notebooks we will use in the lectures for in-class activities.

Under Construction

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

Notebooks

1. Loading a dataset
2. Choosing Input Variables

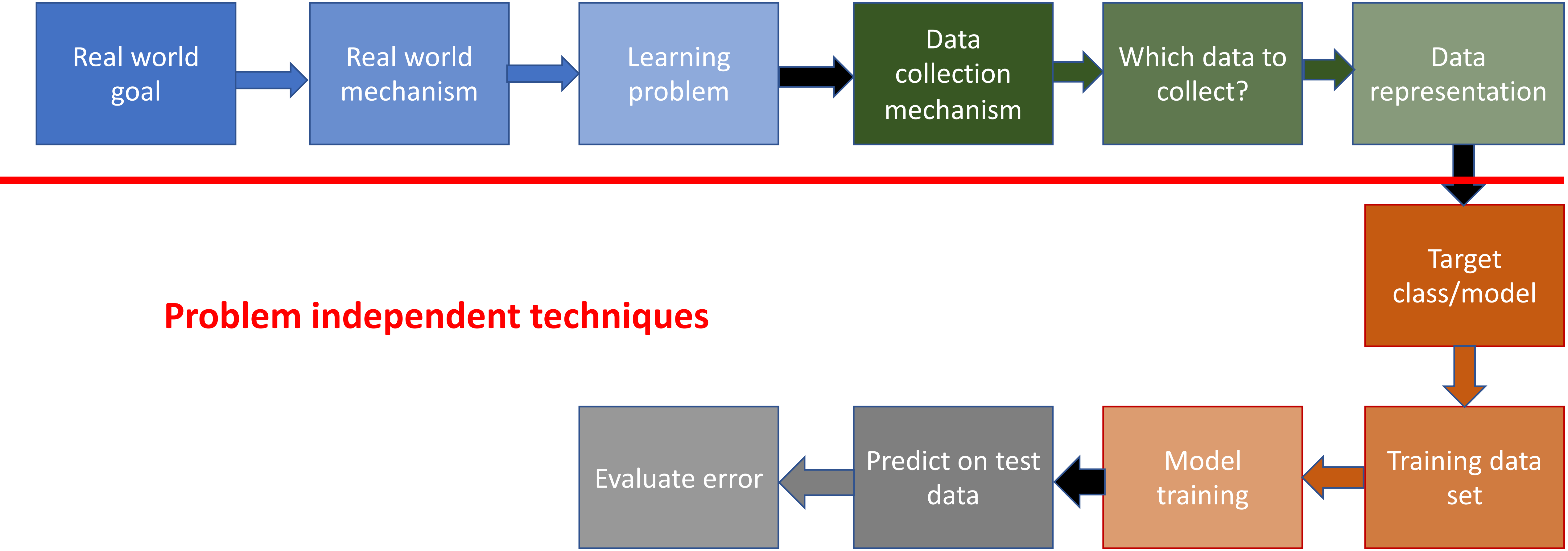
Datasets

Below are some of the datasets that we will be used by the notebooks we use in class:

- COMPAS dataset. This is a dataset generated by ProPublica. This specific version is taken from Kaggle, which in turn got the original data from ProPublica.
- Adult dataset. This is a dataset from UCI ML repository: Adult dataset. The local file has the headers for each column as well.

Copyright © 2020, Atri Rudra. Built with Bootstrap, p5 and bigfoot.

Relation to problem statement



Real world goal

Real world mechanism

Learning problem

Data collection mechanism

Which data to collect?

Data representation

Target class/model

Training data set

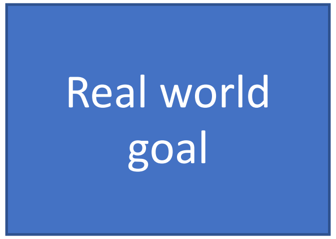
Model training

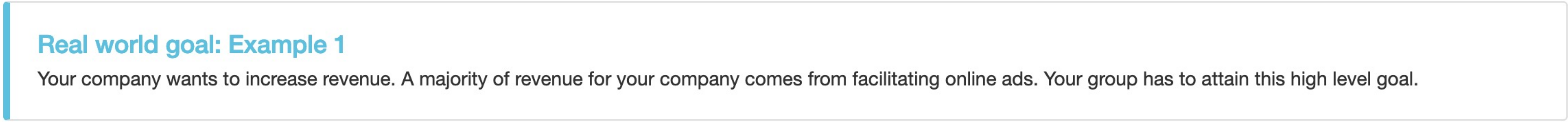
Predict on test data

Evaluate error

Problem independent techniques

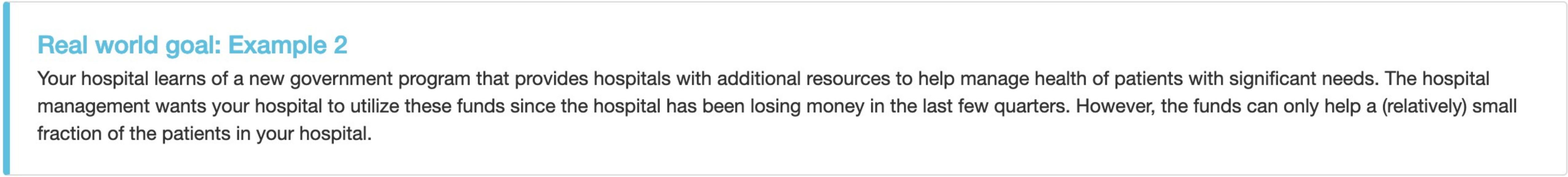
Real world goal





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

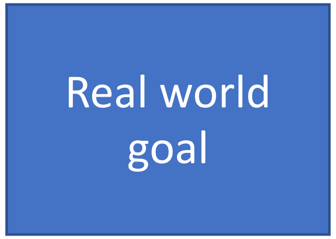
Real world goal

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: General thoughts

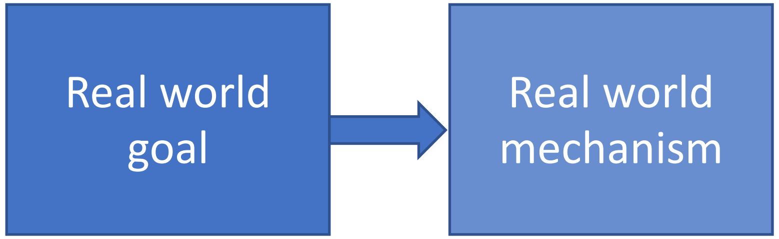


Real world goal

This step generally done at higher management level

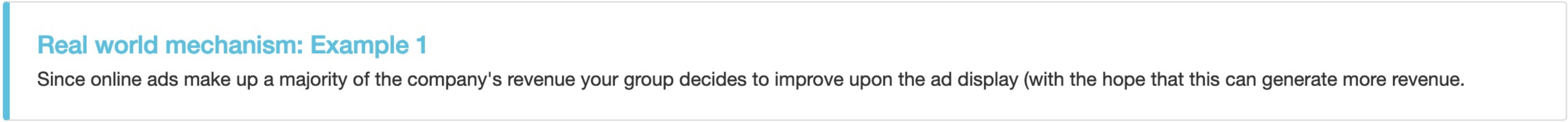
Translating this into something concrete needs remaining steps

Real world mechanism



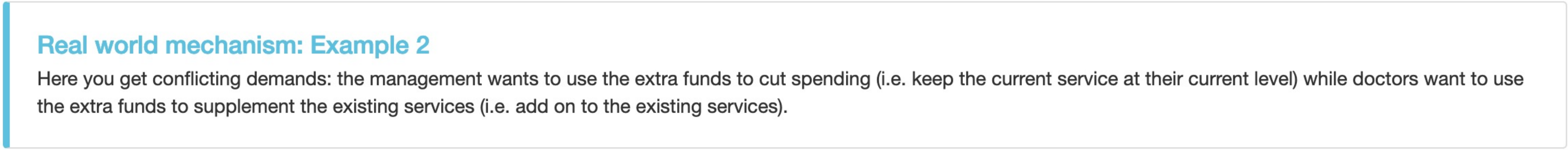
Real world goal

Real world mechanism



Real world mechanism: Example 1

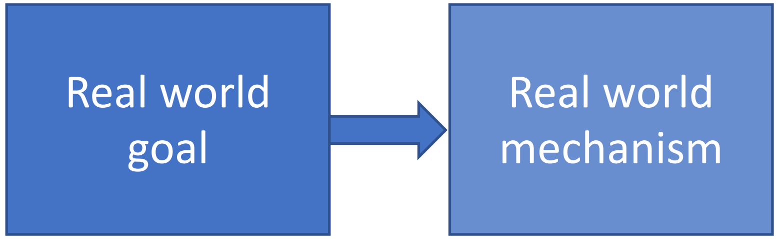
Since online ads make up a majority of the company's revenue your group decides to improve upon the ad display (with the hope that this can generate more revenue.



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).

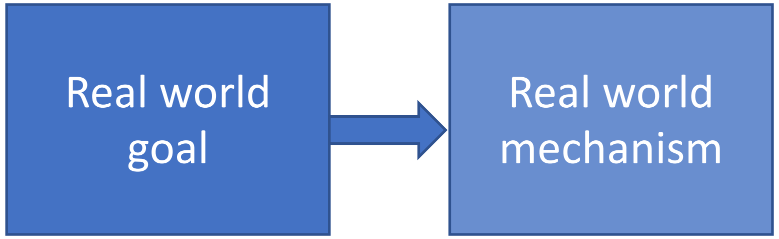
Real world mechanism: Your choice



Choice 1: Benefit = number of degree requirements satisfied

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

Real world mechanism: General thoughts



Real world goal

Real world mechanism

Is improving ad display the best option?

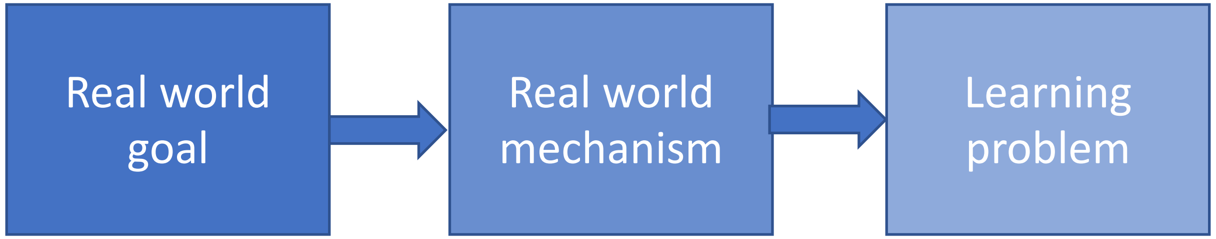
ALWAYS question if the mechanism captures well the real life goal

There can be competing/incompatible mechanisms

Doctors vs. management in Example 2

CONVENIENCE trap!

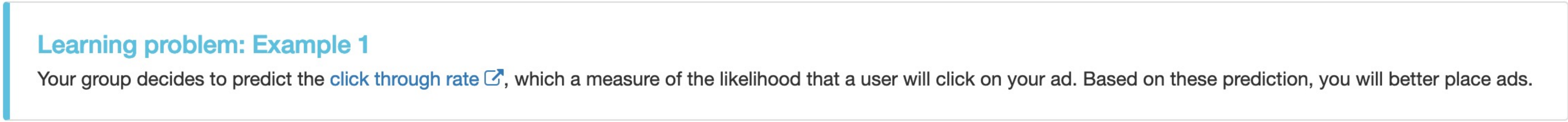
Learning problem



Real world goal

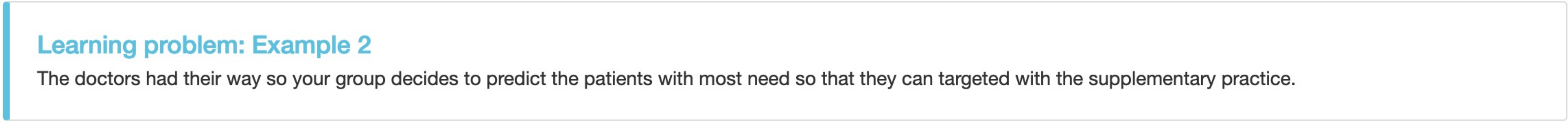
Real world mechanism

Learning problem



Learning problem: Example 1

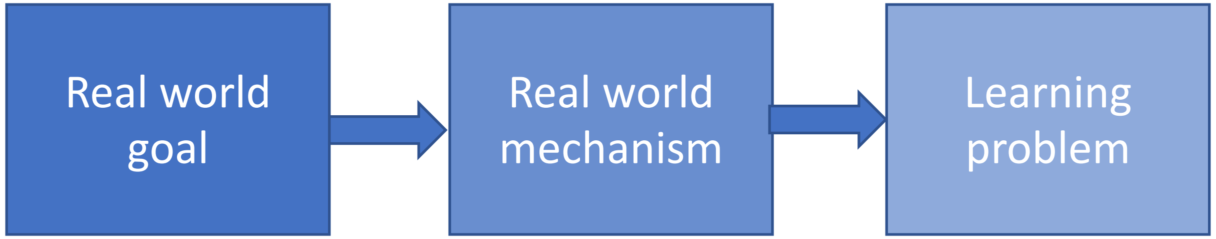
Your group decides to predict the click through rate, which a measure of the likelihood that a user will click on your ad. Based on these prediction, you will better place ads.



Learning problem: Example 2

The doctors had their way so your group decides to predict the patients with most need so that they can targeted with the supplementary practice.

Learning problem: General thoughts



Real world goal

Real world mechanism

Learning problem

Decide NOT to use learning

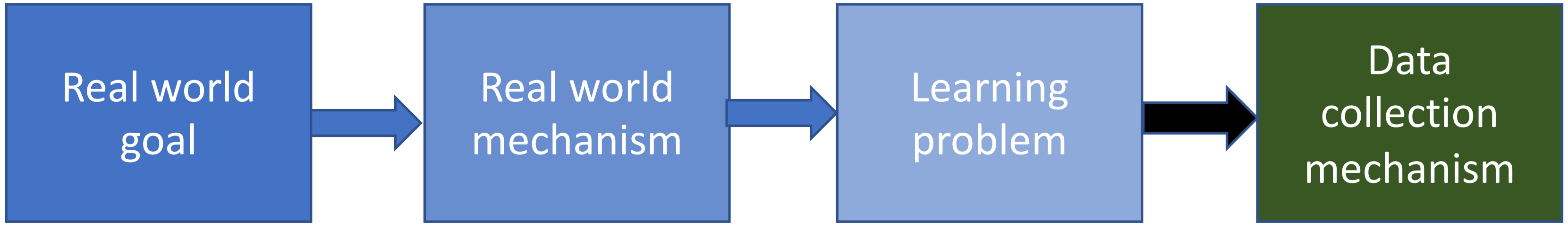
Use of proxies for the real target variable

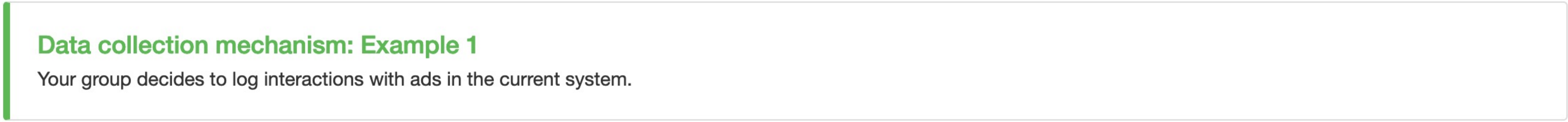
Some has to decide between competing target variables

Choosing the learning problem can have big consequence!

Convenience trap

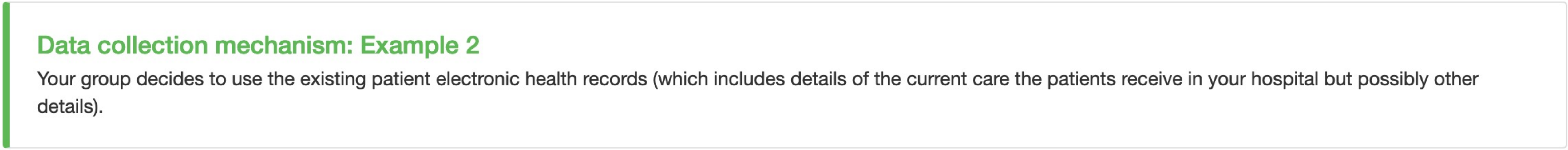
Data collection mechanism





Data collection mechanism: Example 1

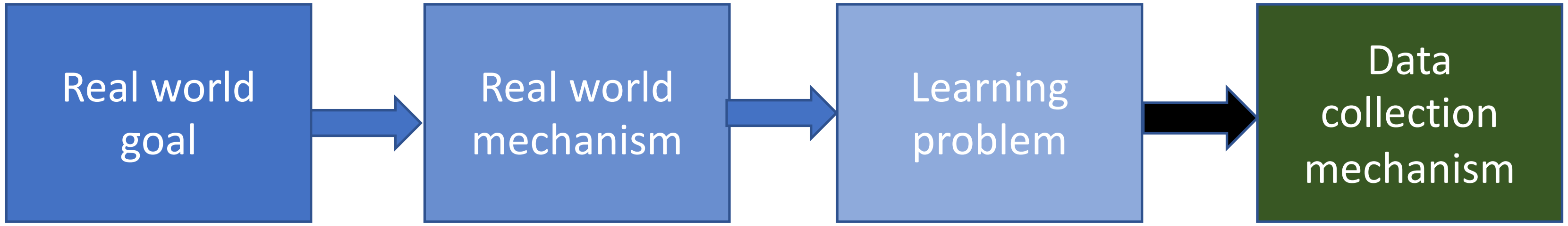
Your group decides to log interactions with ads in the current system.



Data collection mechanism: Example 2

Your group decides to use the existing patient electronic health records (which includes details of the current care the patients receive in your hospital but possibly other details).

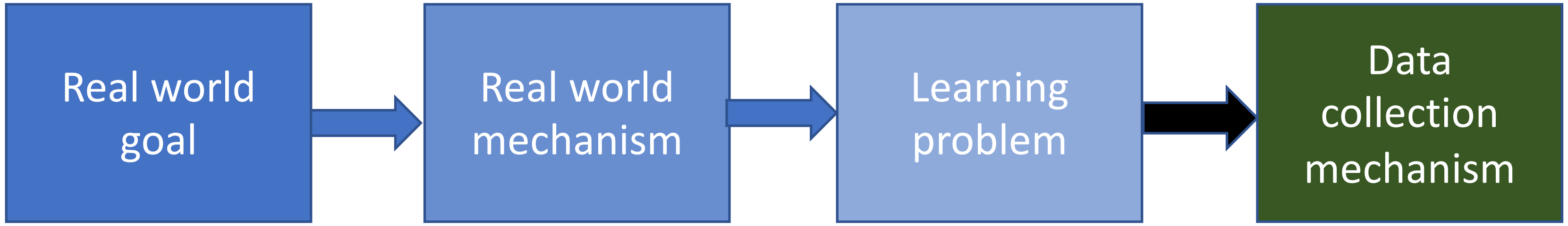
Data collection mechanism: general thoughts



Concept/distribution drift

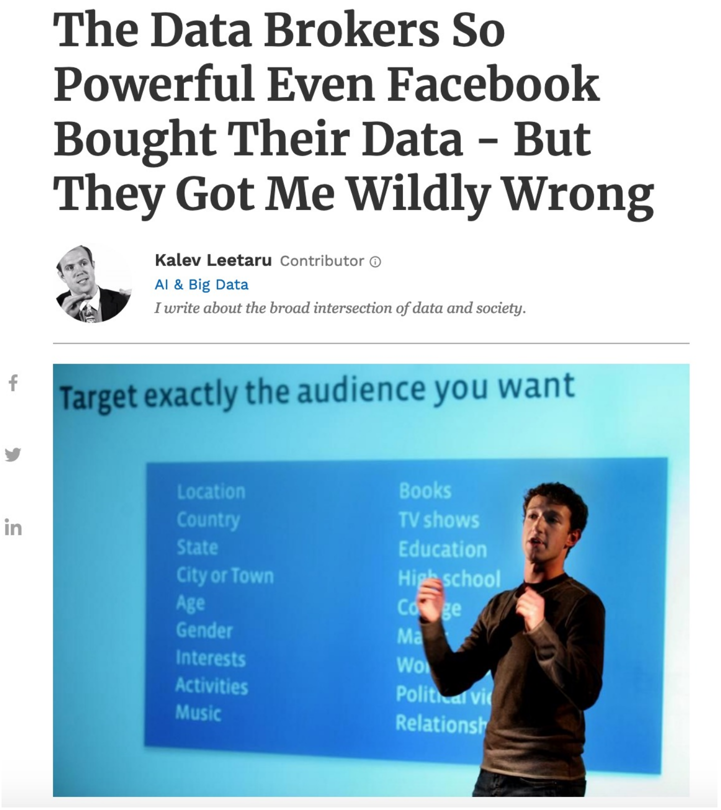
Privacy can be a concern

Data collection mechanism: Data doesn't exist


Real world goal → Real world mechanism → Learning problem → Data collection mechanism

Potential issues?

Use 3rd party data brokers

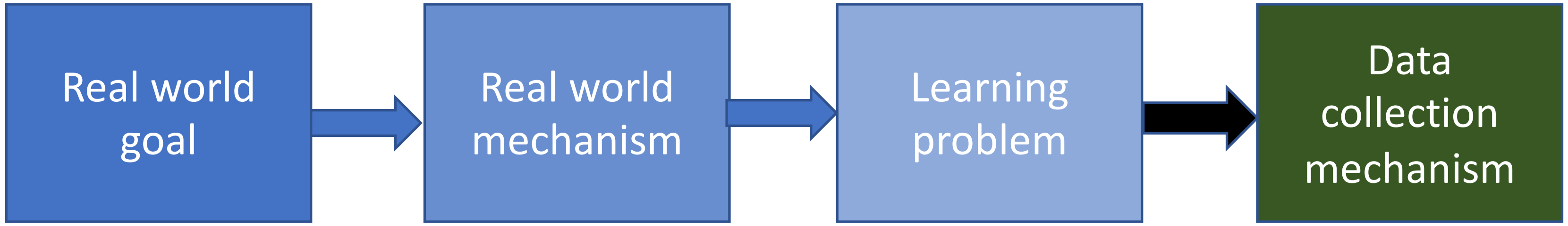

The Data Brokers So Powerful Even Facebook Bought Their Data – But They Got Me Wildly Wrong

Kalev Leetaru Contributor

AI & Big Data

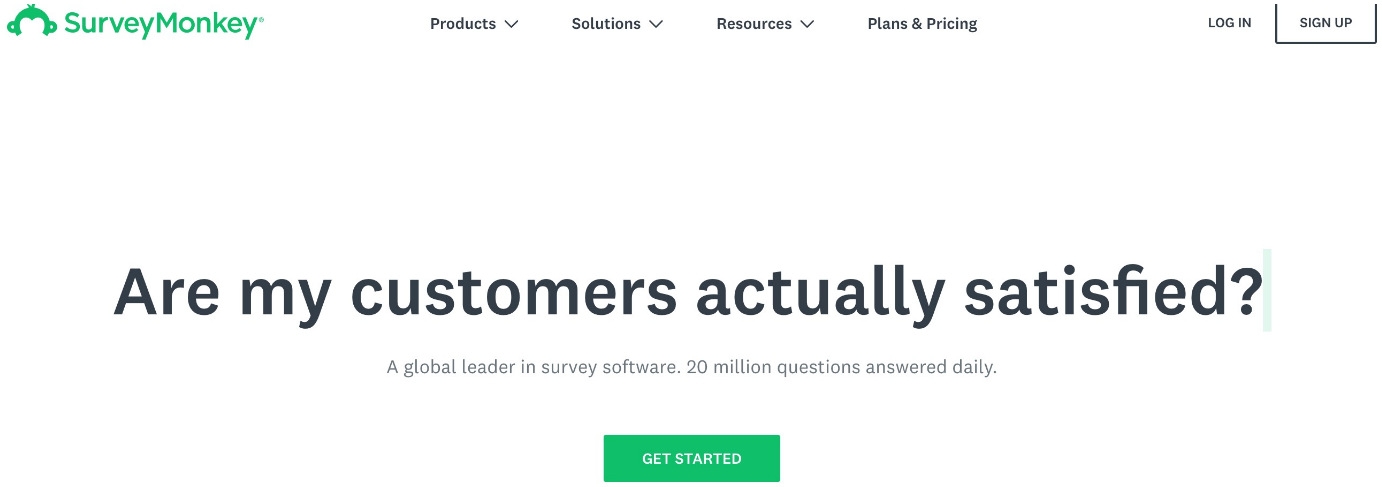
I write about the broad intersection of data and society.

Data collection mechanism: Data doesn't exist



Real world goal → Real world mechanism → Learning problem → Data collection mechanism

Run surveys

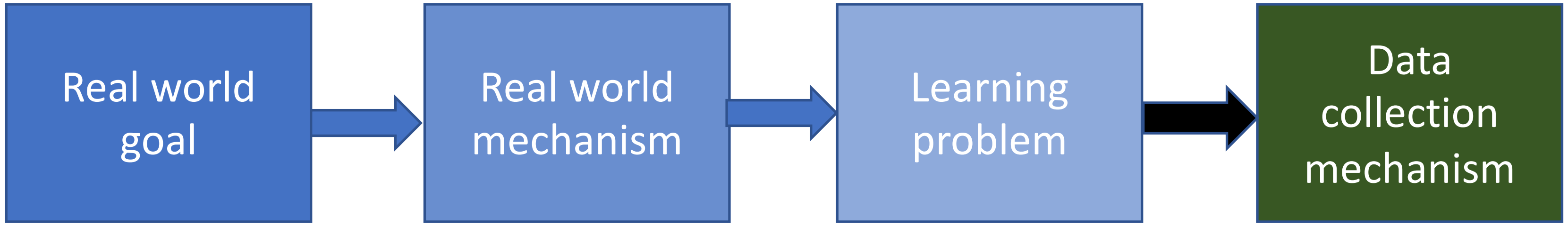


Are my customers actually satisfied?

A global leader in survey software. 20 million questions answered daily.

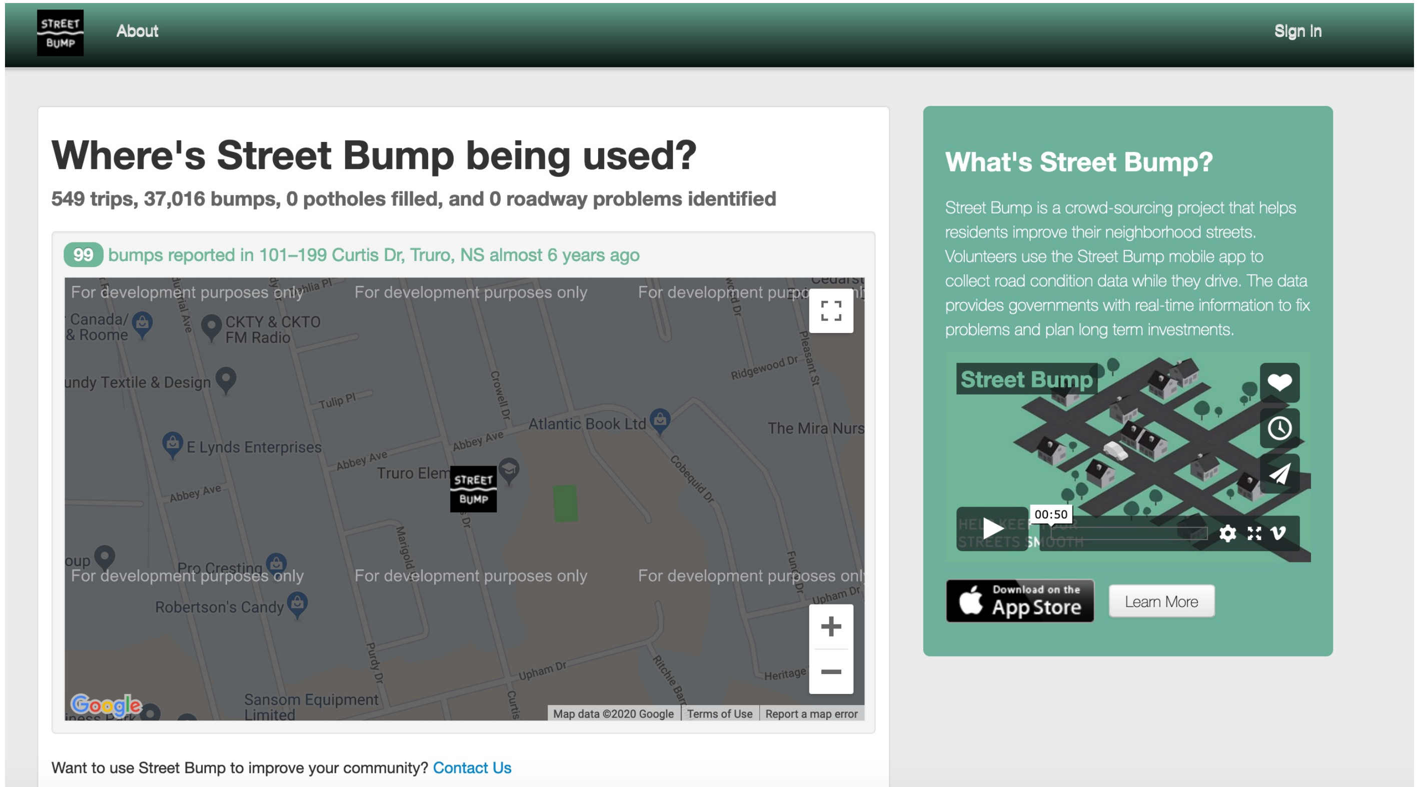
Potential issues?

Data collection mechanism: Data doesn’t exist



Collect data from smartphone

Potential issues?



Where's Street Bump being used?

549 trips, 37,016 bumps, 0 potholes filled, and 0 roadway problems identified

99 bumps reported in 101–199 Curtis Dr, Truro, NS almost 6 years ago

Want to use Street Bump to improve your community? Contact Us

What's Street Bump?

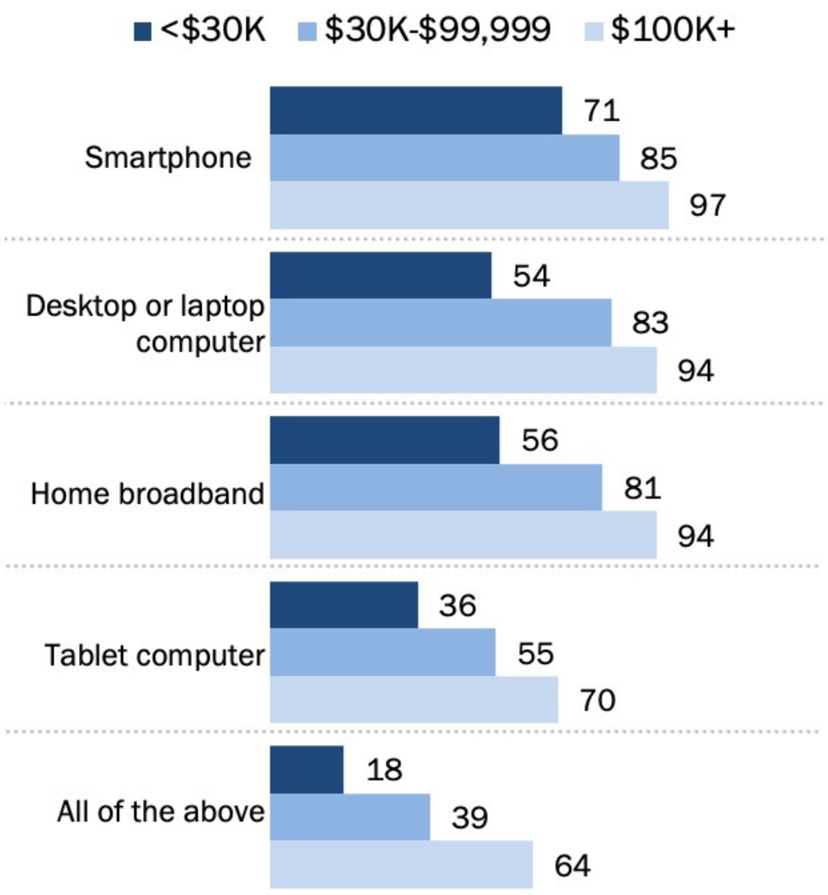
Street Bump is a crowd-sourcing project that helps residents improve their neighborhood streets. Volunteers use the Street Bump mobile app to collect road condition data while they drive. The data provides governments with real-time information to fix problems and plan long term investments.

The smartphone blind-spot

Many of us in CSE assumes that ”everyone” has smartphones

Lower-income Americans have lower levels of technology adoption

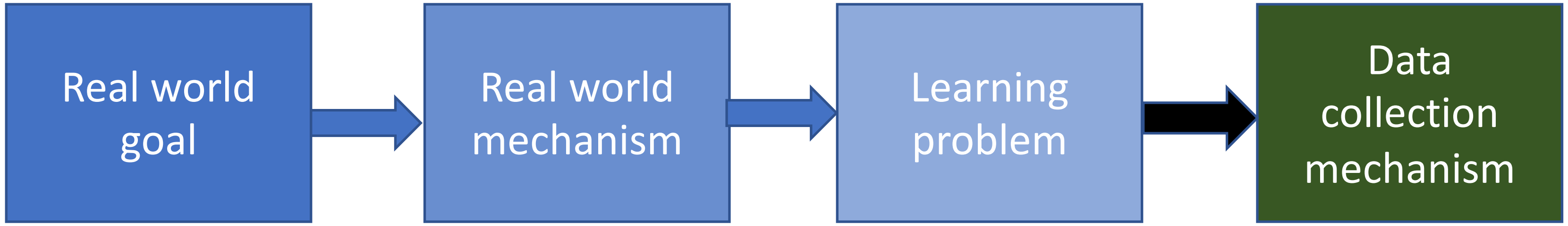
% of U.S. adults who say they have the following …



Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 8-Feb. 7, 2019.

PEW RESEARCH CENTER

Data collection mechanism: Data doesn’t exist

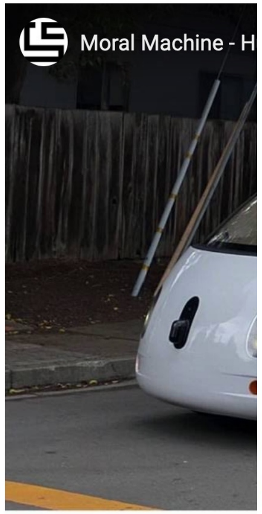


Online video games

Potential issues?

MORAL MACHINE

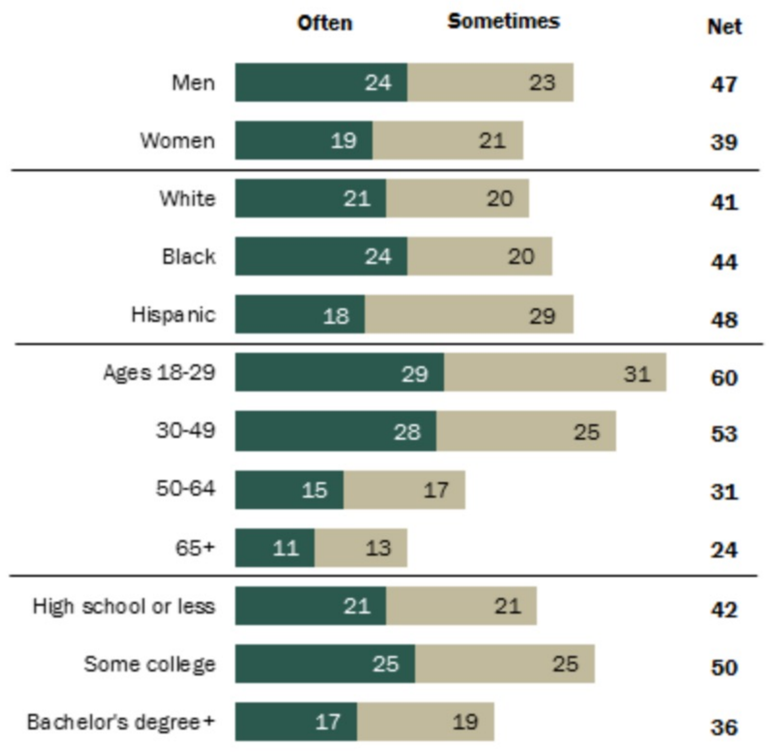
Try our emotional AI
(opens new tab)



By Grendelkhan (Own work) [CC BY-SA 4.0 (http://cre

Younger Americans and men are among the most likely to play video games

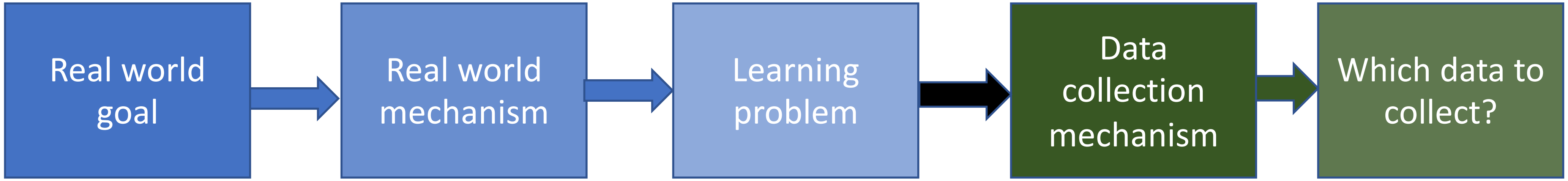
% of adults saying they often/sometimes play video games on a computer, TV, game console, or portable device like a cellphone



Note: Figures may not add to subtotals due to rounding. White and blacks include only non-Hispanics. Hispanics are of any race.
Source: Survey of U.S. adults conducted March 13-27 and April 4-18, 2017.

PEW RESEARCH CENTER

Which data to collect?



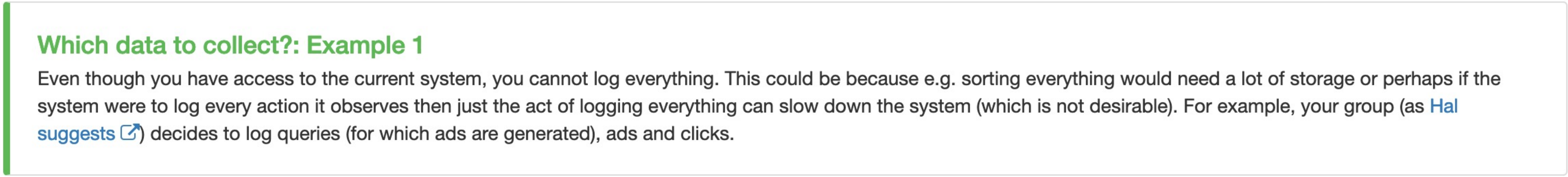
Real world goal

Real world mechanism

Learning problem

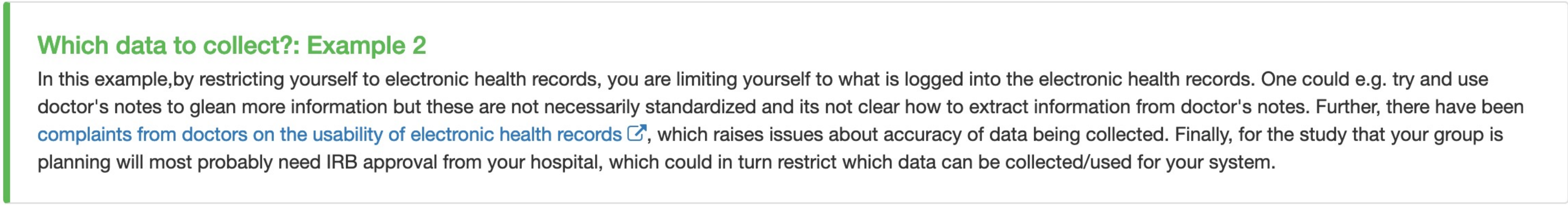
Data collection mechanism

Which data to collect?



Which data to collect?: Example 1

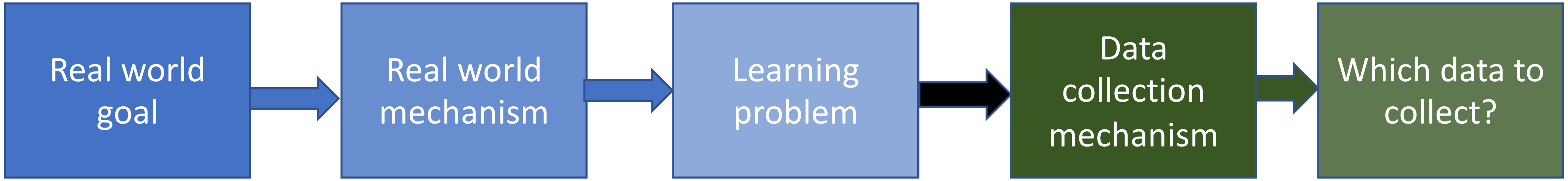
Even though you have access to the current system, you cannot log everything. This could be because e.g. sorting everything would need a lot of storage or perhaps if the system were to log every action it observes then just the act of logging everything can slow down the system (which is not desirable). For example, your group (as Hal suggests) decides to log queries (for which ads are generated), ads and clicks.



Which data to collect?: Example 2

In this example,by restricting yourself to electronic health records, you are limiting yourself to what is logged into the electronic health records. One could e.g. try and use doctor's notes to glean more information but these are not necessarily standardized and its not clear how to extract information from doctor's notes. Further, there have been complaints from doctors on the usability of electronic health records, which raises issues about accuracy of data being collected. Finally, for the study that your group is planning will most probably need IRB approval from your hospital, which could in turn restrict which data can be collected/used for your system.

Which data to collect?: General thoughts

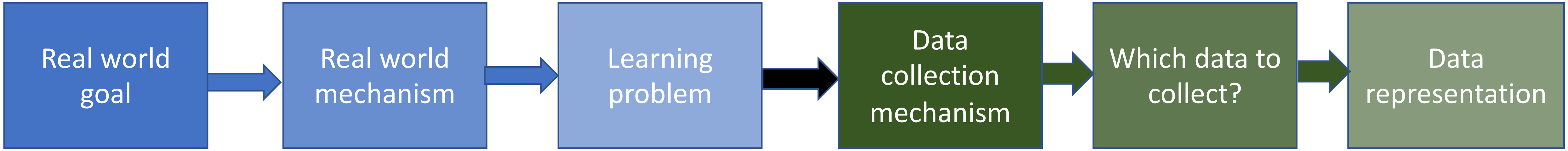


Expense might determine what gets collected

Time to finish a survey also has implications

Other restrictions, e.g. from an IRB

Data representation



Real world goal

Real world mechanism

Learning problem

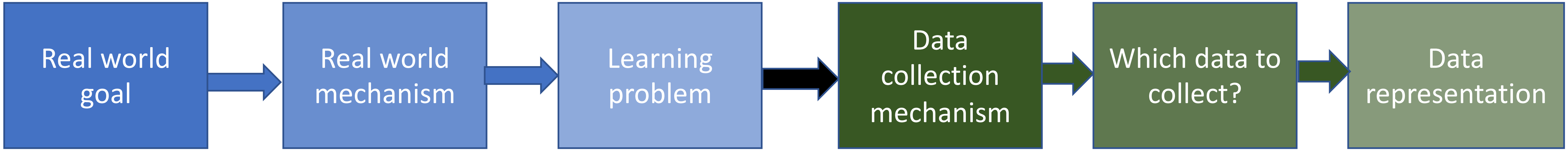
Data collection mechanism

Which data to collect?

Data representation

https://www.history101.com/april-14-2003-the-human-genome-project-completed/

Data representation



Real world goal

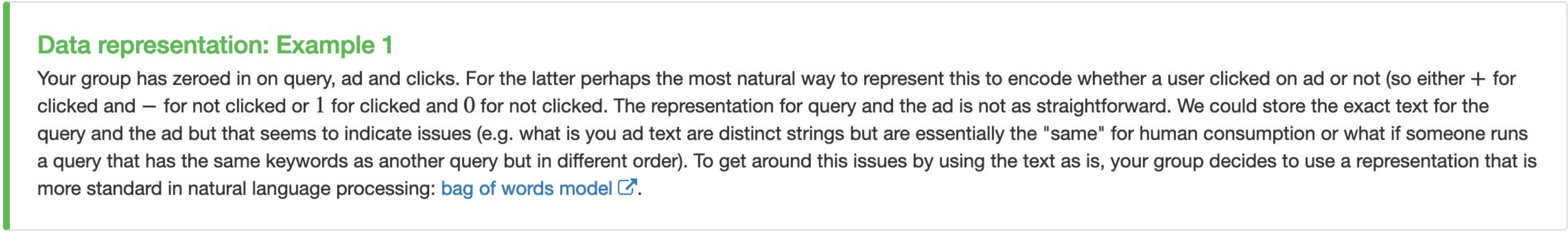
Real world mechanism

Learning problem

Data collection mechanism

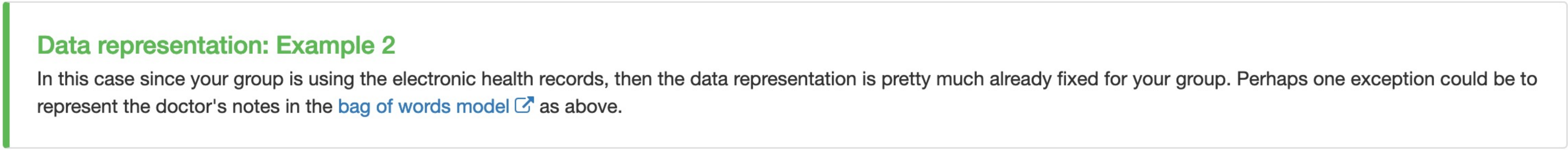
Which data to collect?

Data representation



Data representation: Example 1

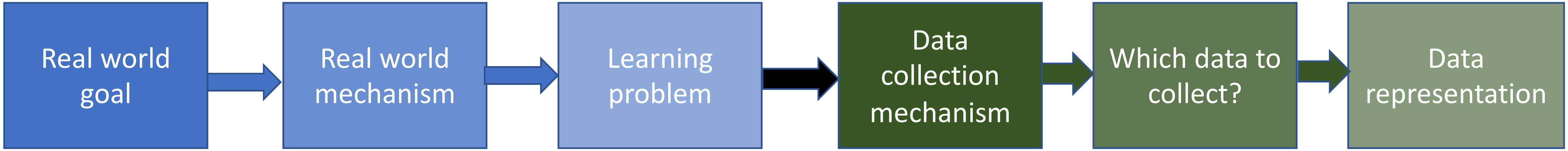
Your group has zeroed in on query, ad and clicks. For the latter perhaps the most natural way to represent this to encode whether a user clicked on ad or not (so either $+$ for clicked and $-$ for not clicked or 1 for clicked and 0 for not clicked. The representation for query and the ad is not as straightforward. We could store the exact text for the query and the ad but that seems to indicate issues (e.g. what is you ad text are distinct strings but are essentially the "same" for human consumption or what if someone runs a query that has the same keywords as another query but in different order). To get around this issues by using the text as is, your group decides to use a representation that is more standard in natural language processing: bag of words model.

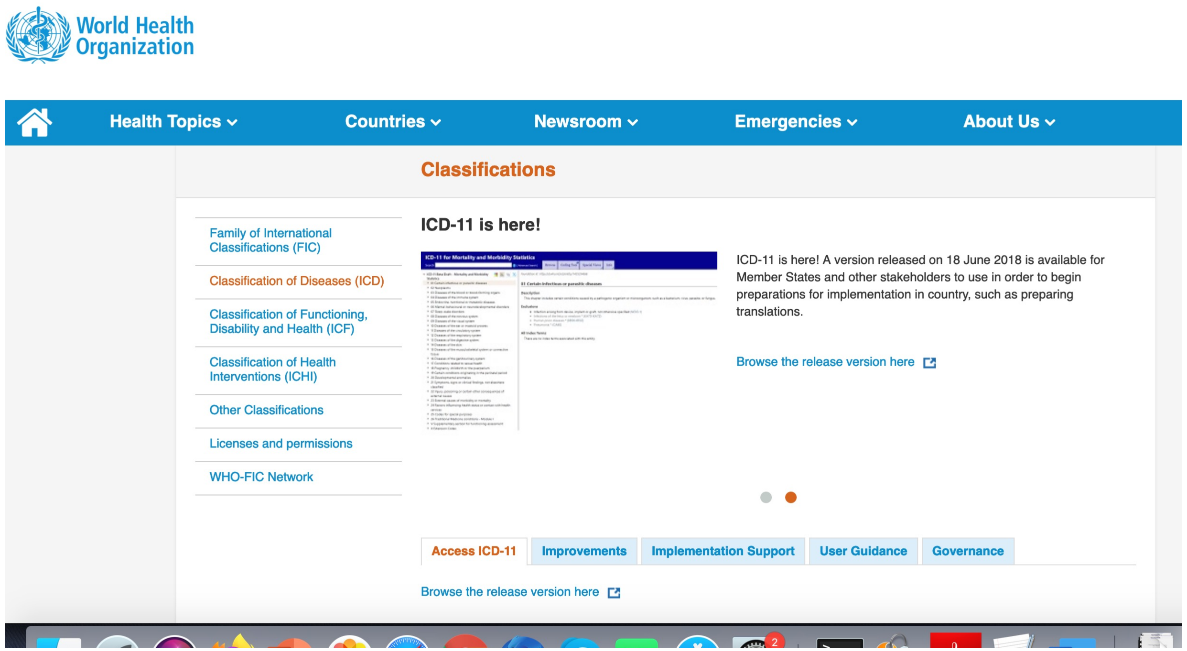


Data representation: Example 2

In this case since your group is using the electronic health records, then the data representation is pretty much already fixed for your group. Perhaps one exception could be to represent the doctor's notes in the bag of words model as above.

Data representation: General thoughts





Classifications

- Family of International Classifications (FIC)
- Classification of Diseases (ICD)
- Classification of Functioning, Disability and Health (ICF)
- Classification of Health Interventions (ICHI)
- Other Classifications
- Licenses and permissions
- WHO-FIC Network

ICD-11 is here!

ICD-11 is here! A version released on 18 June 2018 is available for Member States and other stakeholders to use in order to begin preparations for implementation in country, such as preparing translations.

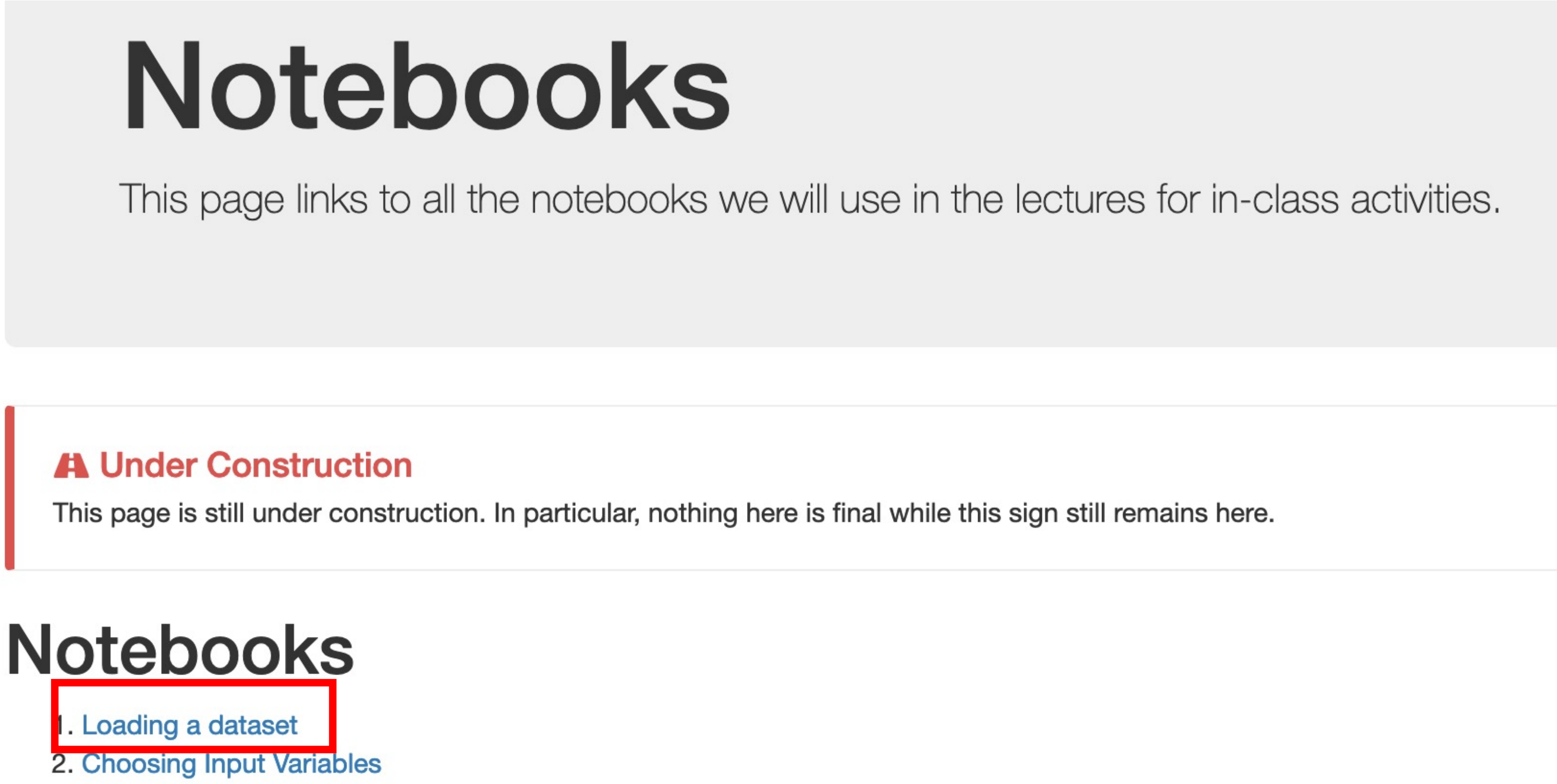
Browse the release version here

Browse the release version here

Categorical data

Jupyter Notebook Exercise

https://colab.research.google.com/



Notebooks

This page links to all the notebooks we will use in the lectures for in-class activities.

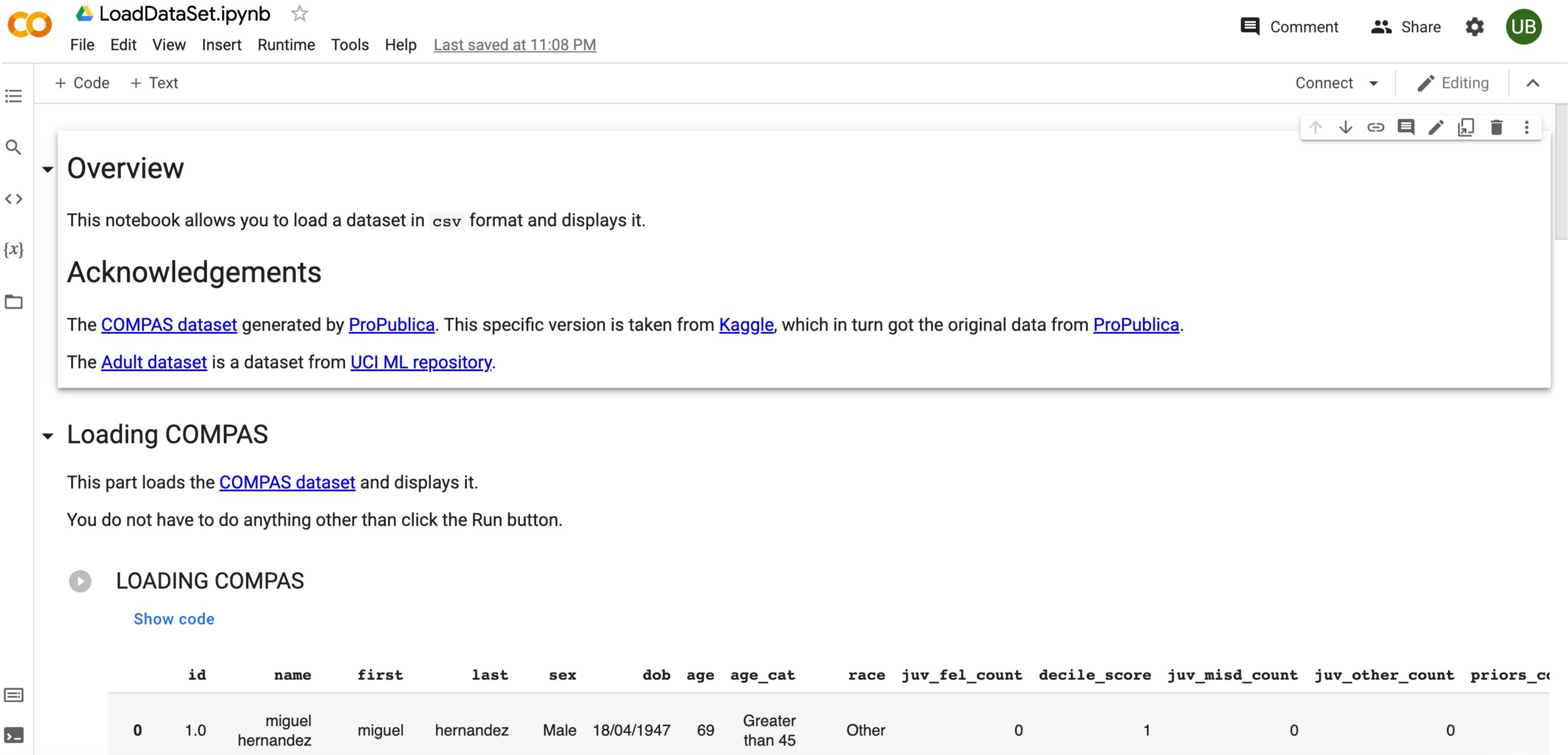
Under Construction

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

Notebooks

1. Loading a dataset
2. Choosing Input Variables

Familiarize yourself with it (and do the Ex.)



CO
LoadDataSet.ipynb
File Edit View Insert Runtime Tools Help Last saved at 11:08 PM
Comment
Share
UB
+ Code + Text
Connect
Editing

Overview

This notebook allows you to load a dataset in `csv` format and displays it.

Acknowledgements

The COMPAS dataset generated by ProPublica. This specific version is taken from Kaggle, which in turn got the original data from ProPublica.

The Adult dataset is a dataset from UCI ML repository.

Loading COMPAS

This part loads the COMPAS dataset and displays it.

You do not have to do anything other than click the Run button.

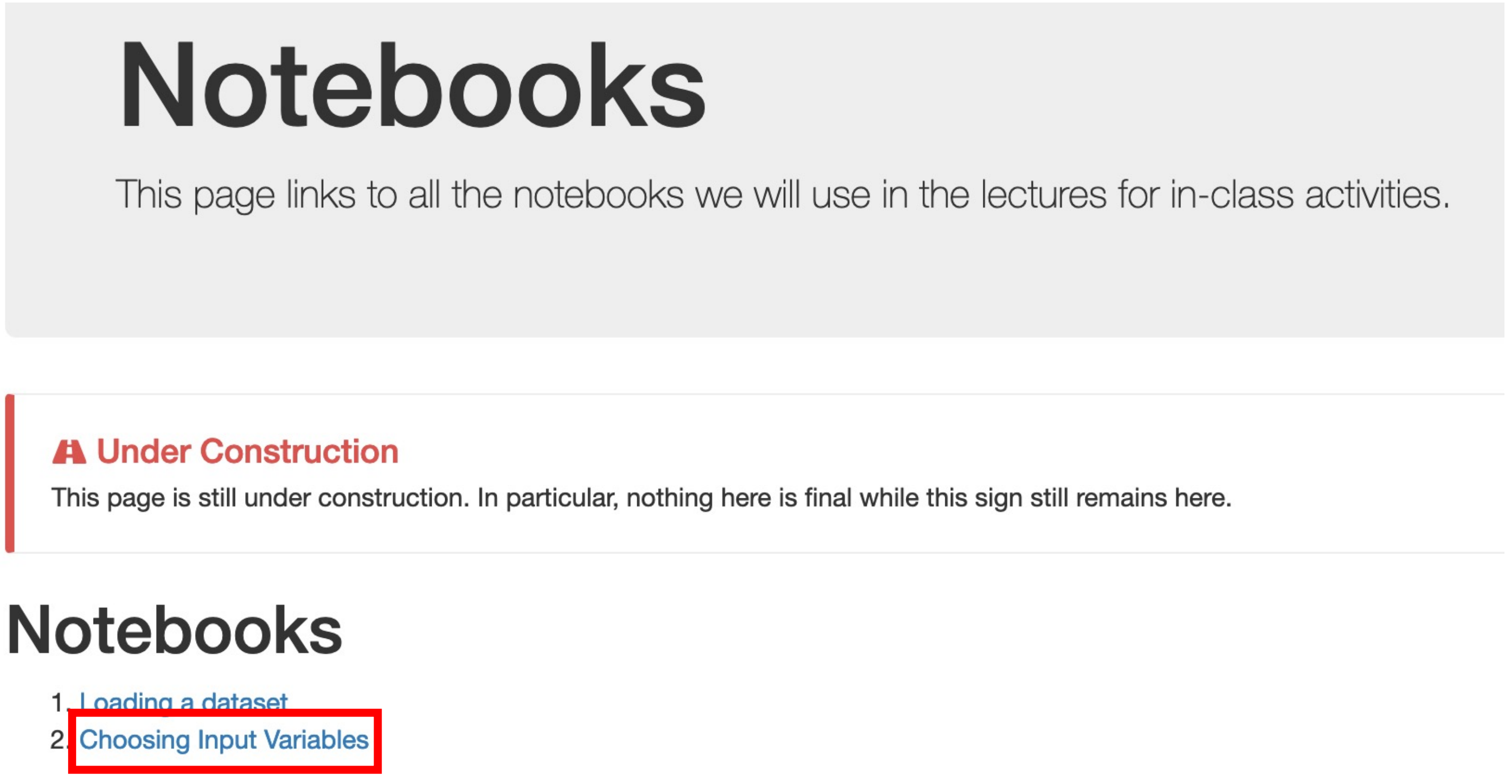
LOADING COMPAS

Show code

	id	name	first	last	sex	dob	age	age_cat	race	juv_fel_count	decile_score	juv_misd_count	juv_other_count	priors_c
0	1.0	miguel hernandez	miguel	hernandez	Male	18/04/1947	69	Greater than 45	Other	0	1	0	0	

Another Jupyter exercise

https://colab.research.google.com/



Notebooks

This page links to all the notebooks we will use in the lectures for in-class activities.

Under Construction

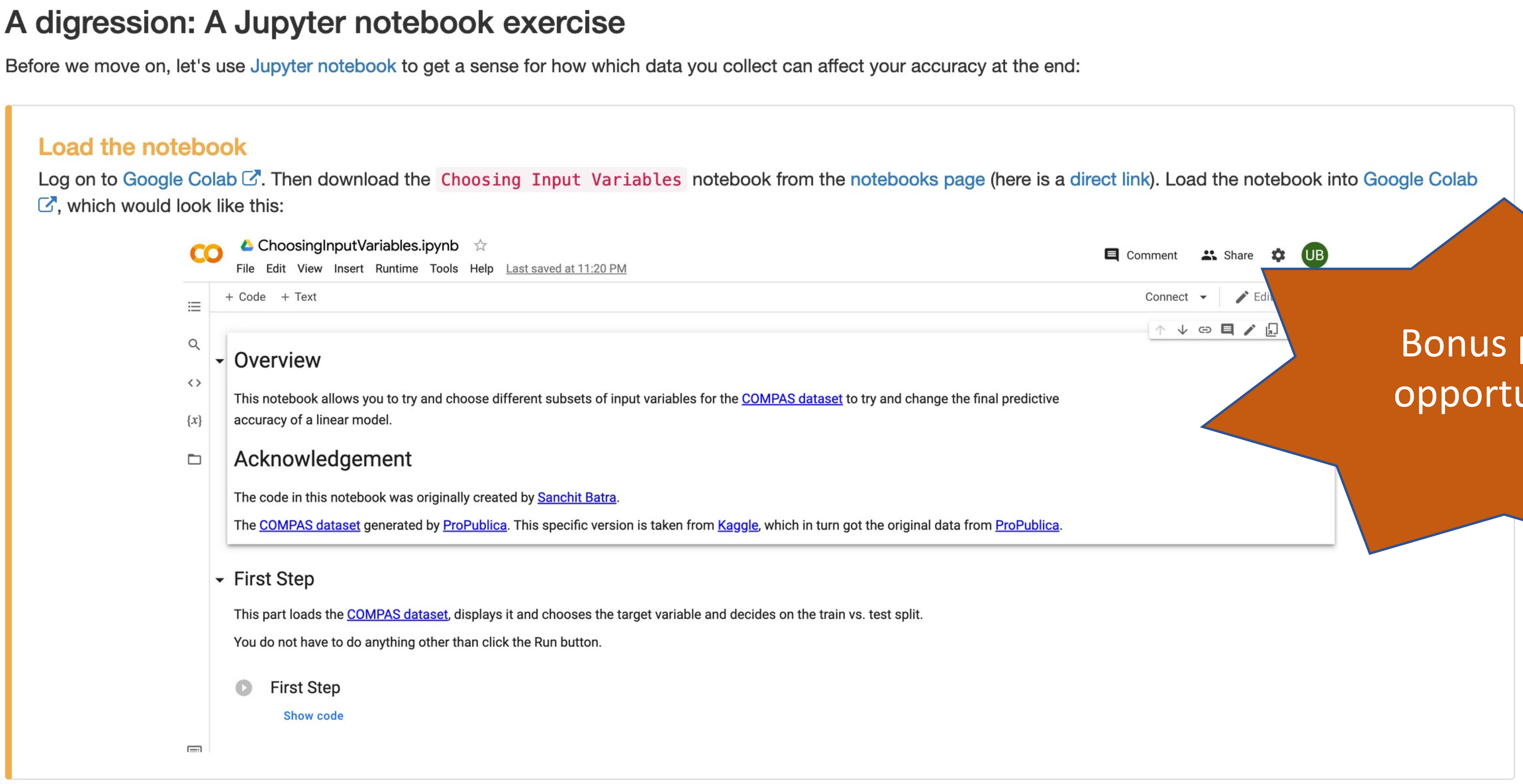
This page is still under construction. In particular, nothing here is final while this sign still remains here.

Notebooks

1. Loading a dataset
2. Choosing Input Variables

Another Jupyter exercise

https://colab.research.google.com/



A digression: A Jupyter notebook exercise

Before we move on, let's use Jupyter notebook to get a sense for how which data you collect can affect your accuracy at the end:

Load the notebook

Log on to Google Colab. Then download the `Choosing Input Variables` notebook from the notebooks page (here is a direct link). Load the notebook into Google Colab, which would look like this:

Overview

This notebook allows you to try and choose different subsets of input variables for the COMPAS dataset to try and change the final predictive accuracy of a linear model.

Acknowledgement

The code in this notebook was originally created by Sanchit Batra.

The COMPAS dataset generated by ProPublica. This specific version is taken from Kaggle, which in turn got the original data from ProPublica.

First Step

This part loads the COMPAS dataset, displays it and chooses the target variable and decides on the train vs. test split.

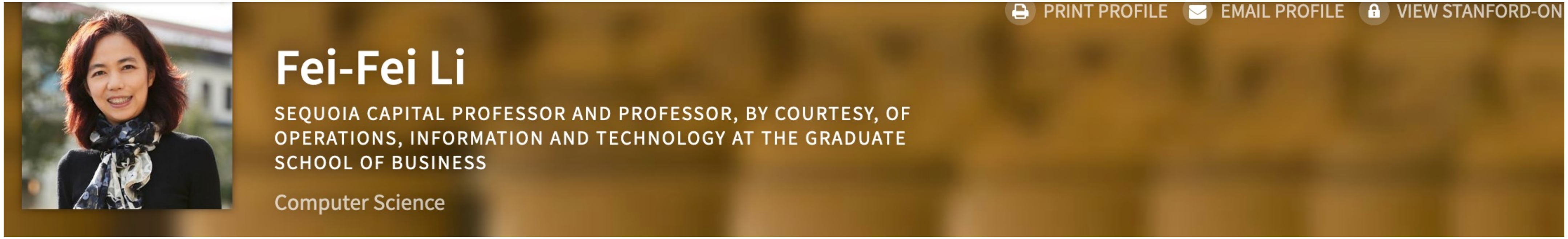
You do not have to do anything other than click the Run button.

First Step

Show code

Bonus point opportunity!

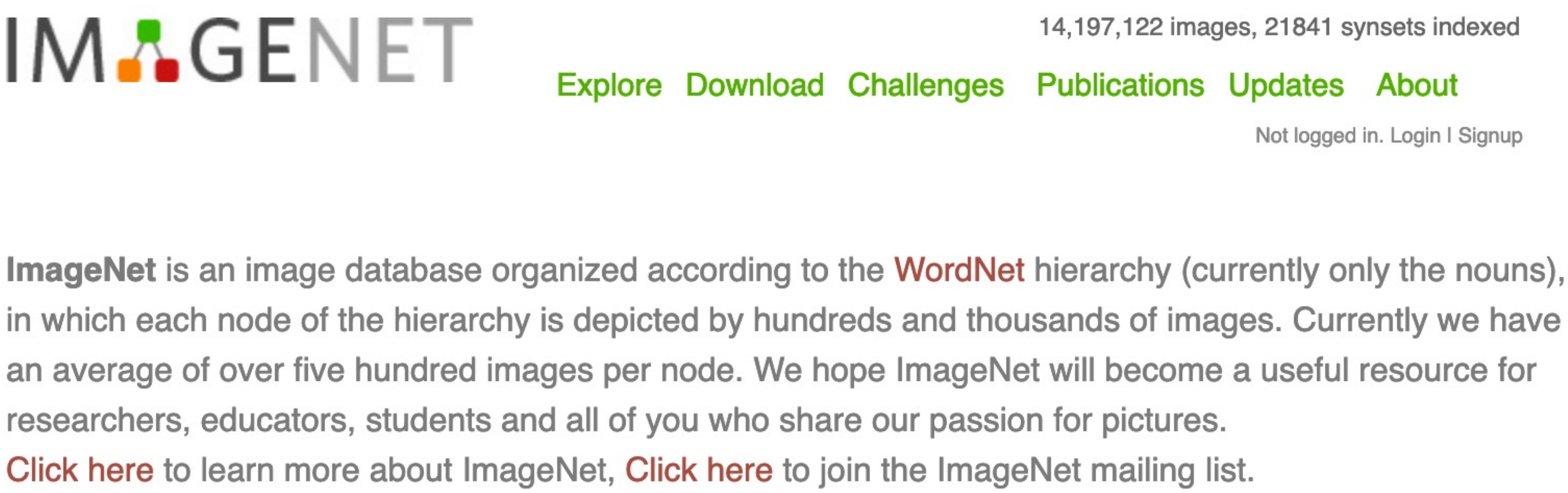
Pass phrase for today: Fei-Fei Li



Fei-Fei Li

SEQUOIA CAPITAL PROFESSOR AND PROFESSOR, BY COURTESY, OF OPERATIONS, INFORMATION AND TECHNOLOGY AT THE GRADUATE SCHOOL OF BUSINESS

Computer Science



IMAGENET
14,197,122 images, 21841 synsets indexed
Explore
Download
Challenges
Publications
Updates
About
Not logged in. Login I Signup

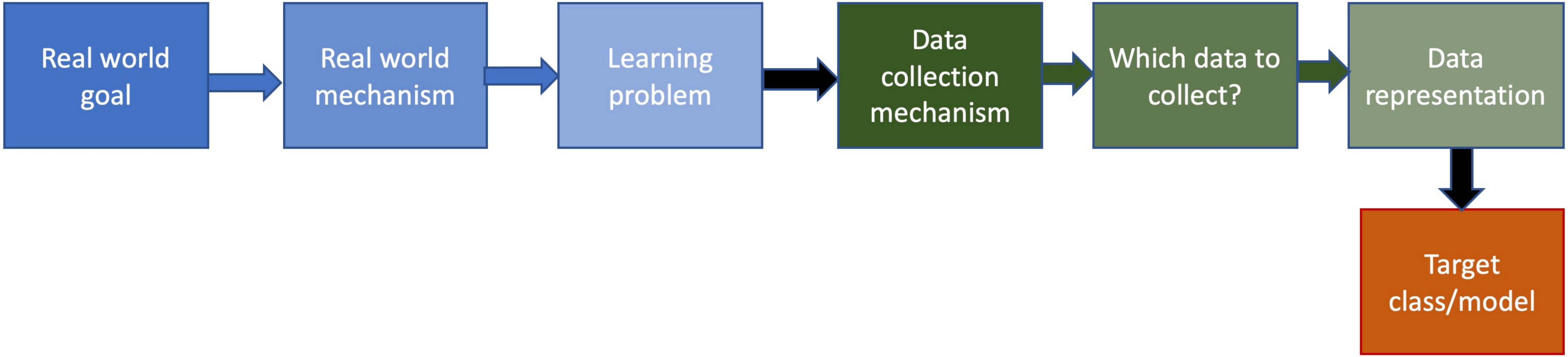
ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

neuroscience. She has published more than 200 scientific articles in top-tier journals and conferences, including Nature, PNAS, Journal of Neuroscience, CVPR, ICCV, NIPS, ECCV, ICRA, IROS, RSS, IJCV, IEEE-PAMI, New England Journal of Medicine, etc. Dr. Li

650-497-5875 (office)

ML model classes



Real world goal

Real world mechanism

Learning problem

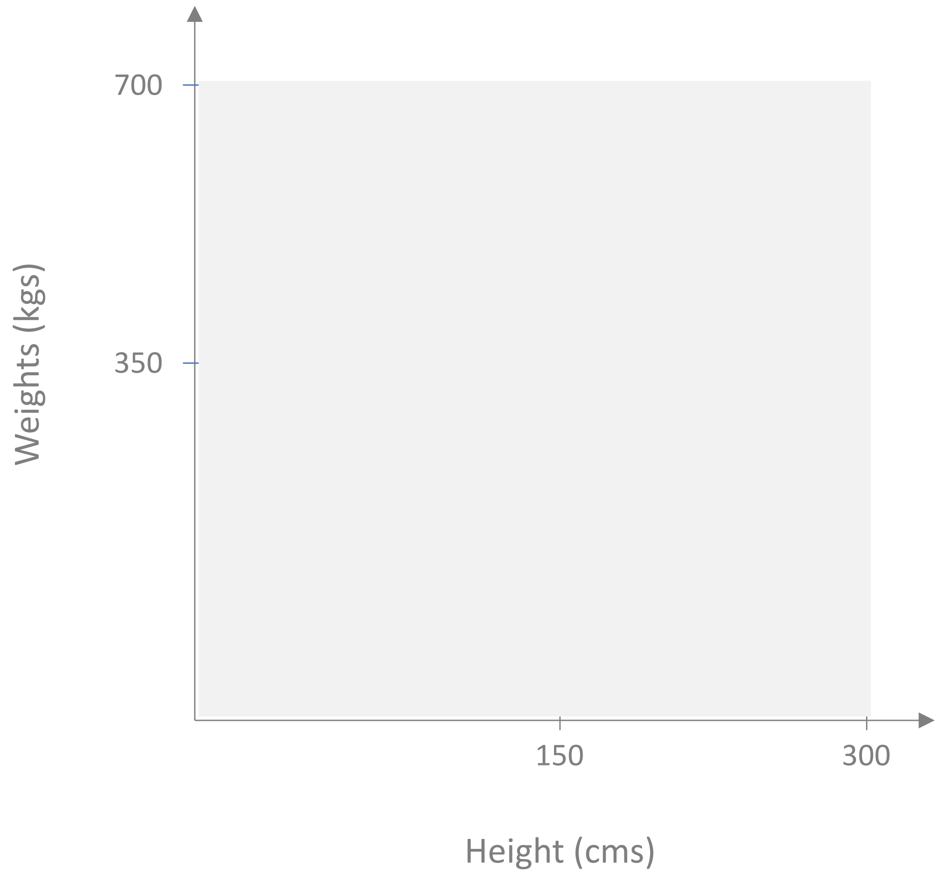
Data collection mechanism

Which data to collect?

Data representation

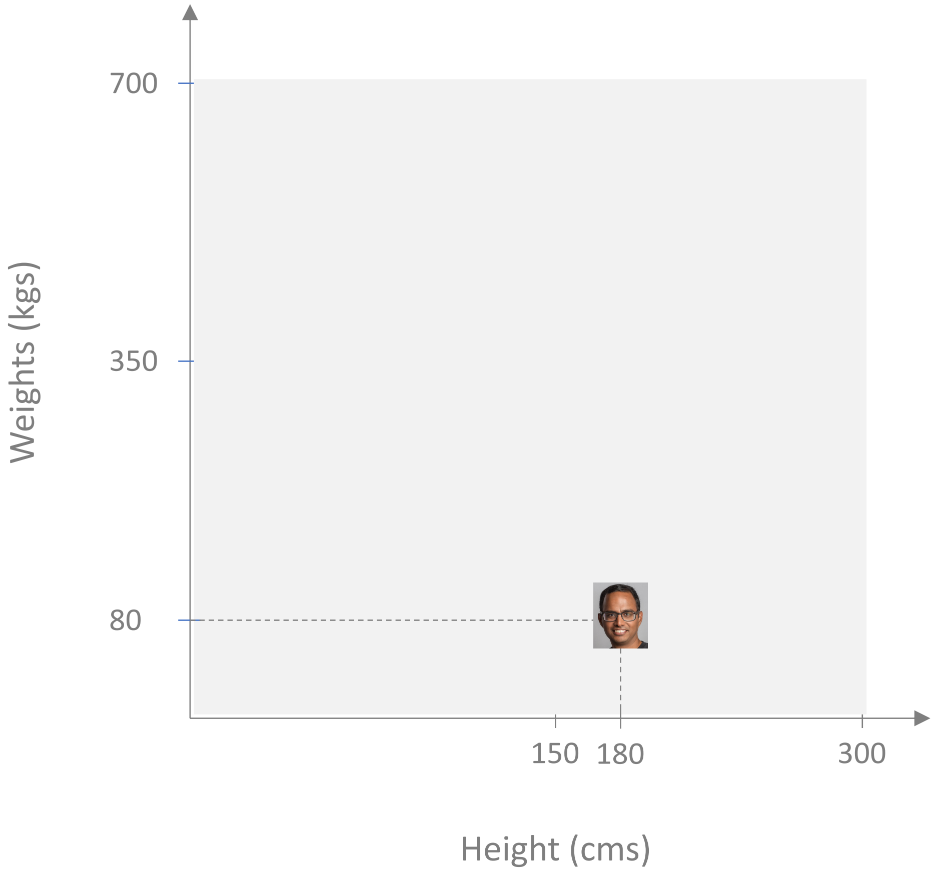
Target class/model

Restrict to two input variables



Predict risk of heart disease

For example…


Weights (kgs)
Height (cms)