Lab 12

Today's plan

01 **Data serialization**

02 **Reading Checkpoint**

03 **Final project work time**

04 **Extras: Huggingface, Zenodo**

Data Serialization

JSON, numpy, Pickle, shelve

Data Serialization: Review

- Serialization: writing data from a computer program to persistent memory on disc as a file
- Deserialization: loading data from a file into a computer program
- Different serialization options for different data types
  - CSV for tabular data
  - JSON for objects
  - Numpy files for arrays
- Why?
  - Allows us to share data and objects
  - Can use with many different programming languages
  - Can store data for future reuse

Lab 12 notebook

Reading Checkpoint

You'll return it until the end of class

Final Project

Project 3) Final Project

The final project involves using a website to present a free-form report-style analysis where you will investigate some questions of interest on datasets from the **Berkeley Police Transparency Hub**.

https://bpd-transparency-initiative-berkeleypd.hub.arcgis.com/search

As you know, BPD shares the following datasets:

- Stop Data
- Calls for Service
- Use of Force
- Collision Data (Traffic Safety)
- Crime Data (Arrests)

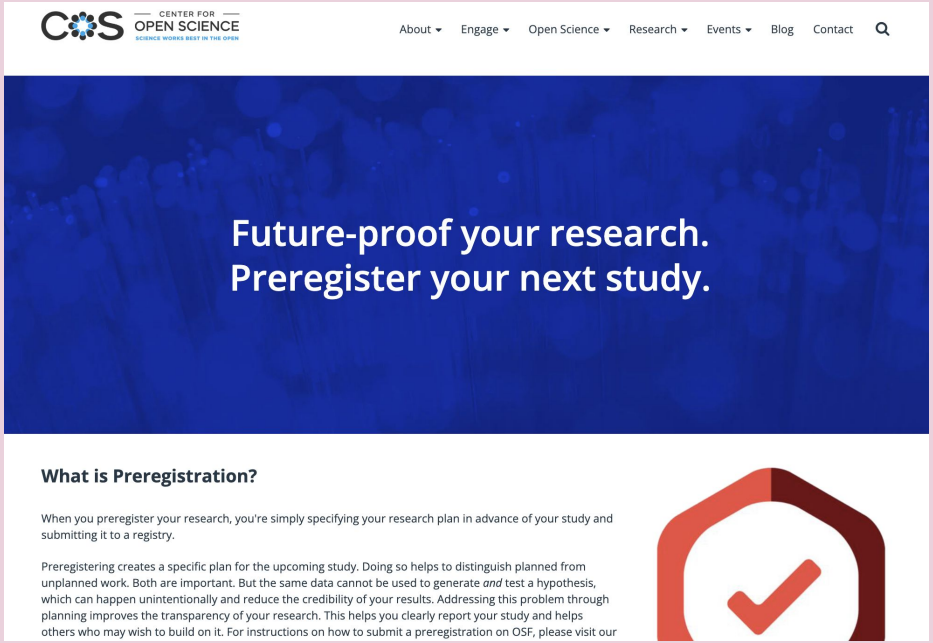
You have to use at least three of the datasets for your project.

> ⓘ **Note**
>
> Implementing sophisticated or novel statistical models is not a requirement of the project (if you'd like to, go for it!), but you should nonetheless conduct an organized, detailed, and thorough anaylsis of your questions.
>
> Most importantly, your project should exemplify the workflows and practices that we have been developing throughout the semester.

Final Project - Preregistration

Goals:

- Specify your research topic and questions *before* performing any analysis or study
- Avoid p-hacking and HARKing (Hypothesizing After Results are Known)
- Defines hypotheses and analysis independent of data
- Increases transparency



Final Project - work time

1. Form groups of 3-4
2. Discuss which datasets you will use
3. Begin brainstorming project ideas
    a. 2 ideas per person ideally!
4. Submit form (1 per team)



Extras

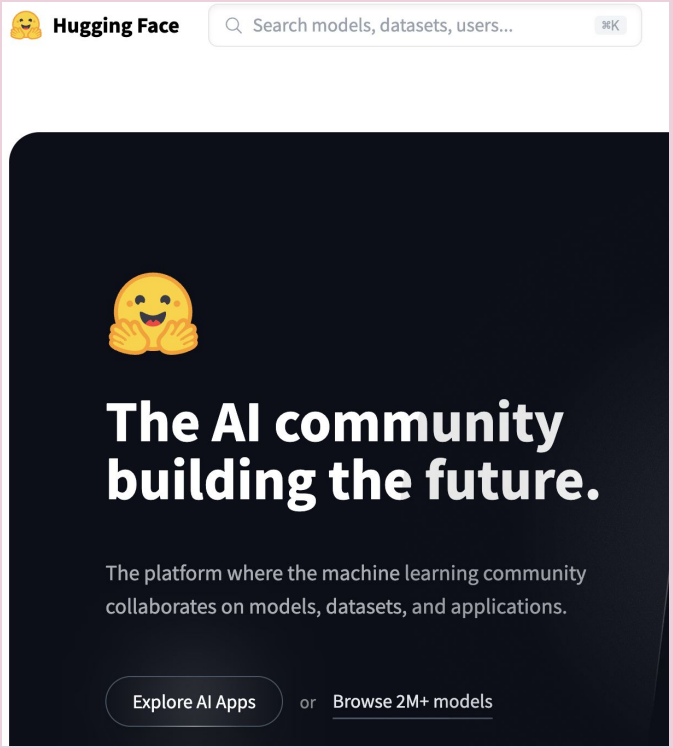
Huggingface, Zenodo

Huggingface

Problem: sometimes we have large data or models that we should not store on github

Solution: Huggingface

- Provides version control via gitlfs
- Can be fully open source or private
- Used by many companies and AI projects



Zenodo

Problem: How do I cite a dataset or code? How do I make my data or code citable?

Solution: Zenodo

- Generates a unique DOI for data or code
- Allows researchers to make their data or code citable
- Free and open source
- Allows for version control and integrates with github

