BERLIN DATA we decided for:

Newspaper Data :

1. ACLED got an account(Paula) only allowed to get data from past 3 years :

newspaper data:

> event type: riot or protest; subtype: violent or non-violent; organization names; newspaper: names, national or city news, time reported; location: quarters, latitude, longitude, population of quarters, …; crowd size;

<https://docs.google.com/spreadsheets/d/1nj3jeYxOHhshLIZ3mbw4kQncFdy3UgL43krZ7qkPrLk/edit?usp=drivesdk> - exel sheet here you can download as csv yourself

* we have started changing columns so please only work on this link ! so we don't mix up sheets !!!

Our GENERAL FOLDER for everything FindMyProtest related is:

* <https://drive.google.com/drive/folders/1Hrvnl6baYQ2TZEKw4ujPBzbtJYIEExWc>

Each one has their own Colab\_Notebook even for our Mentor Session we have one!

* <https://drive.google.com/drive/folders/1KBfC9gAlAPAS5_Z4PCOgT5GuFAyyGfiS>

Our Main inspo for what problems we could solve with that data is:

<https://interaktiv.tagesspiegel.de/lab/demo-atlas-berlin-wofuer-wird-in-meiner-nachbarschaft-demonstriert/>

<https://protestdata.eu/bundesweite-analyse>

Problem models:

1. ecological protests sort them into different organizations find out if they are right wing or left

multi-class problem for Data Science

or even

NER(named-entity-recognition)

**Problem:**

* 1. **Predict from what is happening right now on social media that will happen in future.** We have information from Social Media (TWITTER/X), and there are #s, topics, catchy phrases, videos
  + Assumption: Social media gives signal about probable protests
* 2. We have social media information (posts and invites for protests)
* 3. We have social media information (abouts protests that happened) 1990 - 2020 and how did ~~they develop~~

We pre-select #s. Then we filter by the #s

Protest-type -> [‘Protection of the forest’, ‘war’,...]

Protection of the forest: [#saveamazonia, #savetherainforest, …] -> “Protection of the forest

(1.000) data points

Flu: [#covid, #flu, …] -> Flu

(1.000) data points

5.000 data points randomly from twitter (Taylor Swift, US election, Soccer, Dota2, …) → [“

**Problem v2 - proposal:**

Given some texts from social media (twitter/x, facebook, tiktok) infer the protests and the “\*power” that they will may happen.

*\*Power is the strength that the protest may have, we could measure it through related posts*

by Tugba

By using the [**Versammlungen im Land Berlin**](https://daten.berlin.de/datensaetze/versammlungen-im-land-berlin) data we can have regression and classification tasks. Regression is a type of supervised learning task where the goal is to predict a continuous target variable.

In regression, the output variable takes continuous values. For example duration of protest in the future.

We can evaluate the regression model with some metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc.

In classification, the output variable takes categorical values. For example, next protest area can predict depends on the zip and street.

We can evaluate the classification model with some metrics like accuracy, precision, recall, F1-score, etc.