https://data

<https://data.wprdc.org/dataset/?q=landslide&sort=score+desc%2C+metadata_modified+desc>

<https://data.wprdc.org/dataset/polling-places-for-allegheny-county-elections-current>

<https://data.wprdc.org/dataset/sidewalk-to-street-walkability-ratio>

Ideas:

* Voter turnout
  + Which factors most contribute to voter turnout (e.g., sidewalks, transit)
* Changes in voter preferences
* Predicting future landslide locations
* 311 Data option (domanin 2)
* Donor project (Domain 1)

**Deadlines**

* Tuesday Apr 1, 2025: Topic proposals and Decision
* Apr 8, 2025 Proposal Due

| Task tracker | | | |
| --- | --- | --- | --- |
| People Assignee | No type Title | Dates Date | Dropdowns Status |
| Person |  | Date | Not started |
| Person |  | Date | Not started |
| Person |  | Date | Not started |

| **Predicting Landslides in Pittsburgh** | | |
| --- | --- | --- |
| Datasets | [**Global Landslides**](https://data.wprdc.org/dataset/landslides/resource/dde1f413-c849-413c-b791-0f861bf219ce) | Database of Landslide events from NASA's Cooperative Open Online Landslide Repository (COOLR). |
| * Landslide-prone areas | Geodatabase of 37 landslide prone areas in Pittsburgh (last updated in 2019) – Does not include data on features of the landslide areas |
| * [Landslide prone areas with feature](https://landslide-portal-alcogis.opendata.arcgis.com/pages/map-tools)s | ArcGIS map that includes different features of landslide prone areas and historic landslides (e.g., steepness, soil grade) |
| * Allegheny County Landslide events | 3 datasets: The Landslides (County Public Works) dataset is maintained by Allegheny County Public Works, showing landslides that have impacted county owned roads. The Landslides (Public Assistance 911) and Landslides (Individual Assistance 911) datasets are recieved by Allegheny County Emergency Services through PEMA's Knowledge Center™ (PEMA-KC). |
| * [Soil Grade](https://websoilsurvey.sc.egov.usda.gov/app/WebSoilSurvey.aspx) USDA | Soil grade and slope for PA |
|  | * [Allegheny County Land use](https://data.wprdc.org/dataset/allegheny-county-land-use-areas/resource/da2190b1-88d0-40c7-9af7-0dc73d8c775c) | Dataset with parcels of land and the land use in Allegheny County. |
| Contribution: | * Predict where the next landslides will most likely occur (e.g., given last months rainfall, predicted rainfall, construction projects, soil, etc…) to mitigate harm and proactively intervene (e.g., make visits to high risk areas to ensure proper drainage). Pittsburgh’s geological factors make it very prone to landslides. | |
| Open questions | * What is the frequency at which the data in the alert system will be updated (i.e., will we have to use an API to get real time weather data) | |
| Notes | * “ Pittsburgh’s notorious “red bed” soil makeup is particularly prone to sliding and is a main cause of many of the slides that have happened, according to the website. Another major cause is water and incorrect drainage.” (source: [local news TRIBLiv](https://triblive.com/local/pittsburgh-allegheny/allegheny-county-releases-interactive-map-locating-landslides-across-the-county/)) * Used ChatGPT to see what datasources might be of use | |

**Voter turnout**

Data:

* Voter turnout ([Link](https://results.enr.clarityelections.com/PA/Allegheny/115752/web.307039/#/detail/0004)): Allegheny county provides precinct-level voter turnout rates as well as candidates votes counts (1323 precinct in the 2022 senate election)
* Polling station location ([Link](https://data.wprdc.org/dataset/polling-places-for-allegheny-county-elections-current)): There are 1327 polling stations in Allegheny county

Difficulty

* The policy goal should be to identify which precincts have lowest turnout, and how to improve it.
  + We may focus on the accessibility such as the [PRT Monthly Average Ridership by Route](https://data.wprdc.org/dataset/prt-scheduled-trip-counts) (2017-2023), [Bus Stop Usage](https://data.wprdc.org/dataset/prt-transit-stop-usage/resource/3f40b94b-4ac4-48f1-8c61-8439d2d2f420) (2021), [Sidewalk Street Ratia](https://data.wprdc.org/dataset/sidewalk-to-street-walkability-ratio/resource/b90ccee1-c0aa-43b9-93e2-8a25e690c393) (2021), Bike Lanes Data ([2019](https://data.wprdc.org/dataset/shape-files-for-bikepgh-s-pittsburgh-bike-map), [2019](https://data.wprdc.org/dataset/data-protected-bike-lanes-oct-2019)). [POGOH trip data](https://data.wprdc.org/dataset/pogoh-trip-data) (2022-2025)
  + The problem is to merge these data will be really time consuming, and how to do that is another big issue.
* If we can conquer these problems, our policy suggestion should focus on how to improve the accessibility of polling stations, just relocates or public transportation

Clustering

**Share Bike usage and expansion**

Data:

* [POGOH trip data](https://data.wprdc.org/dataset/pogoh-trip-data) (2022-2025): start station, end station, start date, end date
* [POGOH station location](https://data.wprdc.org/dataset/station-locations) (2022-2025): name, id, latitude, longitude
* [Bike rack location Downtown Pittsburgh](https://data.wprdc.org/dataset/bike-rack-locations-downtown-pittsburgh) (2016/2017, not updated)
* Supplement with Bus ridership data to identify the potential new location to improve the bike usage
  + [PRT Monthly Average Ridership by Route](https://data.wprdc.org/dataset/prt-monthly-average-ridership-by-route/resource/12bb84ed-397e-435c-8d1b-8ce543108698) (2017-2023)

Policy goal: to identify the potential location for expanding POGOH system to encourage bicycle usage to reduce traffic volume, relieve overcrowding on buses during rush hours, and encourage a healthier daily commute.

Early thoughts/ notes on Predicting

Predicting changes in winning party margin // changes in party win margin

* Description: Identifying which party will win based on changes in socioeconomic factors such as employment,
* Winning party margin - Democrat vs. Republican, challenger vs. incumbent
* Granularity level: Either precinct or county → issue with precinct is ensuring we have demographic data at that level
* Relevant datasets:
  + [Allegheny County Election Results](https://data.wprdc.org/dataset/election-results)
  + [Current Voter Registration Statistics](https://www.pa.gov/agencies/dos/resources/voting-and-elections-resources/voting-and-election-statistics.html)
  + [Commonwealth of Pennsylvania Election Results](https://www.electionreturns.pa.gov)
* Questions for Raja:
  + **Is the topic okay** —> given the expansion out of western pa to state and possibly multiple states? → Given the “policy implications / contributions” → political targeting and the policies of incumbent politicians
    - Make sure we have enough data
    - Can potentially expand to one or two more states
    - How can parties affect the election results? (money, influence, etc.)
  + **Which ML tool would be best** → **clustering** to identify precincts that are most similar (e.g., high margin for incumbent, highly swayed by economic factors vs. less swayed by economic factors and demographic changes, vs. highly swayed by economic factors but not demographic changes) or **predicting** the amount the margin will change
    - Classification: will the incumbent win or not?
    - Prediction: low vs. high margin
* Current approaches:
  + [The Hill and DDHQ - Election Forecasting](https://elections2024.thehill.com/forecast/2024/president/pennsylvania/) → by overall state, doesn’t seem to display information at more granular levels

house, senate

Pattern

Political consultant

Events, what's happening outside

Vote changes (incumbent): features

## **Thursday Research and Discussion Goals**

* **We all research political predictions for mayoral and state elections, and federal elections** 
  + What has been done?
  + What has been found to be impactful?
  + What models might be best to use?
  + What datasets are out there?

## **Proposal Timeline**

* Thursday: Touch base on the research and datasets
* Friday: Writing the research and beginning the analysis
* Saturday: Finalize the research and contributions (e.g., datasets); start preemptive data analysis and cleaning
* Sunday: Finalize the analysis and scope
* Monday: 4/6: finalizing
* Tuesday 4/7:

Proposal Outline

## **Proposal Outline**

#### **How this problem is solved today/previously**

* *What interventions exist/will exist*
* *What data do you have and what additional data will you need?*
* *Important*: You should do data exploration and provide descriptive stats to show that you have enough relevant data to solve this problem

#### **What analysis are you proposing to do?**

* *What is the ML problem? What are some possible outcome variables (labels) that you might use?*
* How will you validate it in the class project? What metrics will you use? Why will those metrics achieve the goal you described above?
* What additional validation will need to be done later?

#### **Caveats and Concerns**

* What are some ethical considerations here around privacy, equity, transparency, and accountability? How do you plan on dealing with them?
* Caveats (due to data limitations, analysis limitations, time limitations, etc.)

##### **Policy recommendations:**

* What kind of recommendations do you hope to give to policymakers based on this analysis/project?
* How will you validate whether what you are proposing will have the desired impact?

Open questions:

* Granularity city v precinct // multi-city
* What is the ML model prediction (e.g., vote share or just party)

Meeting notes

Apr 3, 2025 3:00 PM

Proposal Discussion

# Agenda

| Topic | Time | File | Team member |
| --- | --- | --- | --- |
| Topic 1 |  |  |  |
|  |  |  |  |
| Next steps | 5 |  |  |

# Action items and next steps

# Details

* **We all research political predictions for mayoral and state elections, and federal elections** 
  + What has been done?
  + What has been found to be impactful?
  + What models might be best to use?
  + What datasets are out there?

### All notes recap

* What has been done?
  + Lots of research but generally focused on nation wide elections
  + Most are using social media text data to analyze the results
  + Some research tries to do cross-ntnl prediction with same model and few focus on the local
  + Our contribution is the building of local elections
  + Previous models use Bayesian models to predict the probability of either party winning
  + Mostly focused on 2 party voting // dems and republicans
* What is our contribution
* What has been found to be impactful
  + Polling data
  + Twitter data
  + Campaign finance data
  + Polls
  + Party Bias
* Best models to use
  + Clustering
  + Regression
* Data
  + Simon found
  + County level data may be more accessible
* Wh

### Open questions

* Are we focusing on mayors or state level election?

### Ideas for later

O

Contribution

* Incumbency: policy implication, if incumbency always wins, then they have lower incentives to be better :)

**Data:** <https://sfethics.org/disclosures/campaign-finance-disclosure/campaign-finance-dashboards>

<https://www.sfelections.org/tools/election_data/>

<https://sfethics.org/ethics/2023/12/campaign-finance-dashboards-november-5-2024.html>

<https://www.sfelections.org/tools/election_data/registration_by_party.php#>

**Next step**

1. Case selection (one or two more cities, campaign finance data, dominates by one )
   1. Check the [link](https://ballotpedia.org/Party_affiliation_of_the_mayors_of_the_100_largest_cities)
   2. [Oklahoma city](https://oklahoma.gov/elections/elections-results/election-results.html) (red city): they have electoral results database, but i cannot find the results for mayor idk why
   3. Dellas (largest city that has Republican mayor now)
2. Sentiment analysis (wanna do or not)
   1. CrowdTangle API? (facebook)
   2. Twitter API
3. Decided what models

Zoe's notes

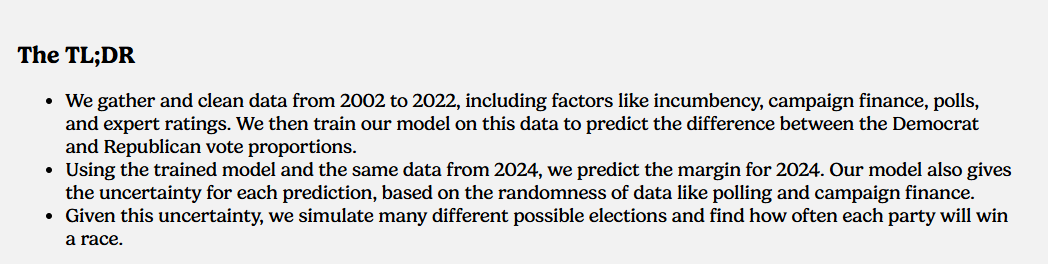
BRainstorm NOtes

* If we do local polling is likely less robust so the model may have other features that rise as the best
* Could not find quick data on the political party of Raleigh, NC mayors
* We would use areal interpolation which uses spatial relationships and densities to distribute values from one polygon layer to another – this can be done w the Tobler library ([tutorial](https://dges.carleton.ca/CUOSGwiki/index.php/Areal_Interpolation_in_Python_Using_Tobler#Areal_Interpolation_in_Python_Using_Tobler_-_Tutorial))
* Variation potential: how do the factors that contribute most heavily to one party leaning or another relate to certain factors?

**Dallas Mayoral Results (1981-1983)**



[Brown students create ML model](https://www.browndailyherald.com/article/2024/09/get-to-know-the-election-result-prediction-model-developed-by-brown-students) (2024)

* Created a [website](http://24cast.org) that made predictions for each state and updated multiple times a day
* Their [methodology](https://24cast.org/methodology) :
* 
* Their [data sources](https://github.com/BPR-Data-Team/ElectionModel2024/blob/main/DataCitations.md)
* Their model was unique because typical election forecasting manually assigns weights to features (e.g., polls) instead of just letting a model determine the most important features
* Expanded on the features typically used. [Key features](https://24cast.org/methodology) included:
  + Campaign financing
  + Polls
  + Incumbent differential
* Prioritized transparency and included “Shapely values” to give users clear insight into how the predictions were made

[How To Buy An Election: A Machine-Learning Approach to Predicting Federal Elections](https://github.com/AndrewBrodsky/election_predictions?tab=readme-ov-file#how-to-buy-an-election--a-machine-learning-approach-to-predicting-federal-elections) (2018)

* Polls are typically the best predictor of election results
* “This project uses machine learning to predict the relationship between campaign contributions and candidates' vote totals for U.S. House of Representatives elections, taking into account district and candidate characteristics.” (Brodsky)
* Previous ML models used to predict election outcomes were Random Forest and SVM and Kmeans clustering
* This study tested random forest model, a gradient boosted regressor, and a simple linear regression. Gradient boosted regressor had the best outcome with R^2 of .82
* “A model using Random Forest and Support Vector Machines predicted election results based on age, gender, and race of individual voters and achieved results that were accurate within 1%.”
* Kaggle includes further examples of some machine learning algorithms to predict 2016 election results, including K-means clustering.” (Brodsky)

[A Machine Learning Based Strategy for Election Prediction](https://par.nsf.gov/servlets/purl/10209597) (2019)

* Used Twitter posts to predict outcome of midterm elections. Looked at tweets related to brett k’s supreme court nomination → if the tweets were negative they forecasted more dem advantage and vice versa
* States that previous models using Twitter to predict the election outcome would use tweets that contained the candidates name, tag them as pos, neg, or pos and neg, and would then use naive bayes sentiment analysis and SVM to predict the outcome
* This study’s contribution is that they analyze Tweets for “high-impact” events in the lead up to the election to understand broader sentiment and the prediction is made based on the “polarity-analysis” of the event

Look to see if incumbents win at such high rates that they’re not incentivised to make changes

Ideas for methods

* K-means clustering
* Random Forest
* Regression
* SVM

Ideas for features

* Campaign finance
* Census bureau demographic data
* Incumbent candidate
* Polls

Ideas for unique contribution

* Local / mayoral level which are potentially more “volatile”..?
* May have the same primary predictors (e.g., campaign finance)
* Also key features may be more city specific..?
* **Develop model to predict mayoral incumbent win using campaign finance, polls, demographics and if time, twitter sentiment analysis**

Data sources:

* [SF campaign finance:](https://sfethics.org/disclosures/campaign-finance-disclosure/campaign-finance-dashboards)
* [New York campaign finance](https://publicreporting.elections.ny.gov/ContributionsByRecipient/ContributionsByRecipient)
* [San Francisco Party Affiliation by District](https://www.sfelections.org/tools/election_data/registration_by_party.php#)