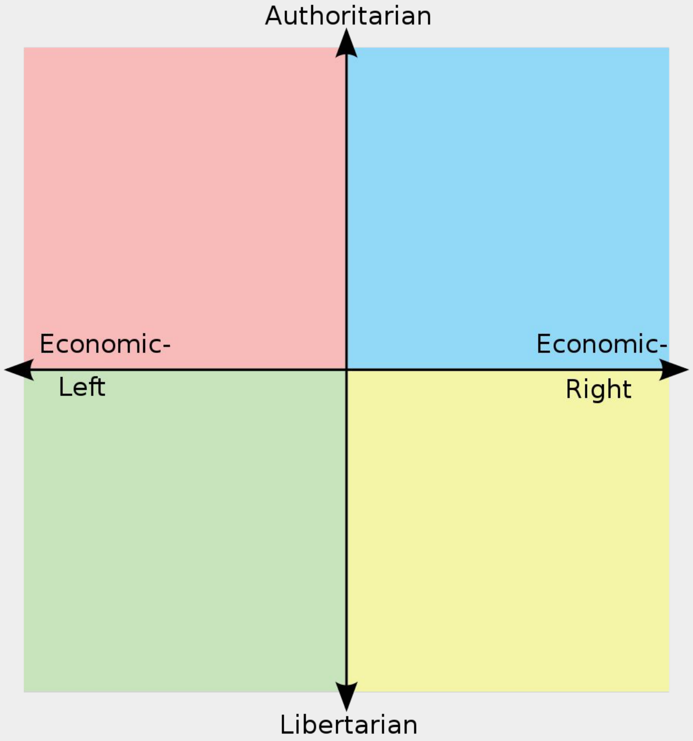
Utilizing X Data to Predict Political Alignment

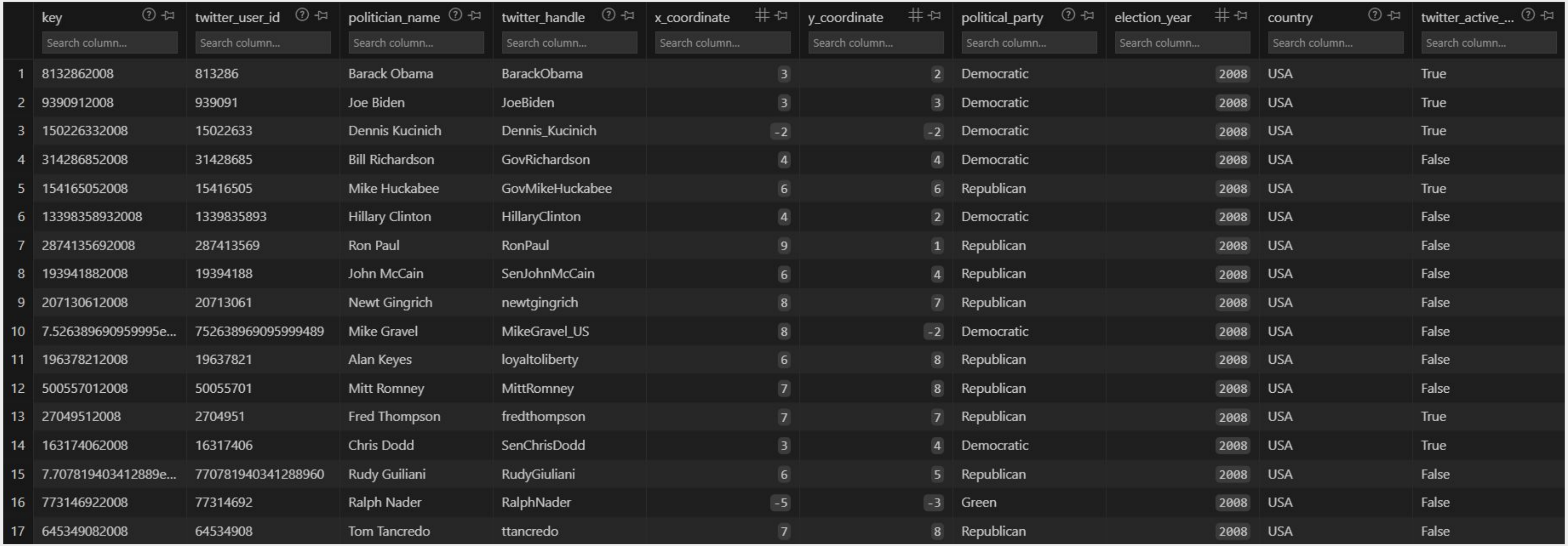
Eloragh Espie and Rylan Vachon

What is the Political Compass?

- Modern political spectrum model
- Attempts to provide a global scale for measuring political values
- Four quadrant graph
  - X axis - economic values
  - Y axis - social values

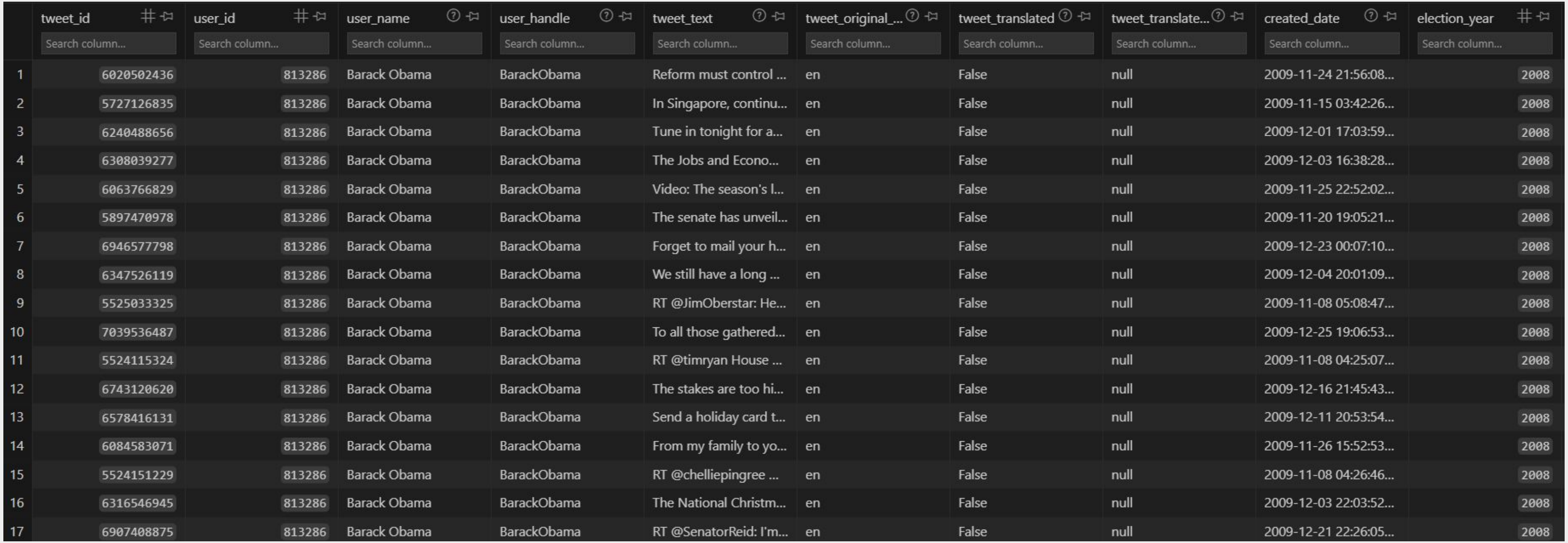


Manual Data Curation



	key	twitter_user_id	politician_name	twitter_handle	x_coordinate	y_coordinate	political_party	election_year	country	twitter_active_...
	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...
1	8132862008	813286	Barack Obama	BarackObama	3	2	Democratic	2008	USA	True
2	9390912008	939091	Joe Biden	JoeBiden	3	3	Democratic	2008	USA	True
3	150226332008	15022633	Dennis Kucinich	Dennis_Kucinich	-2	-2	Democratic	2008	USA	True
4	314286852008	31428685	Bill Richardson	GovRichardson	4	4	Democratic	2008	USA	False
5	154165052008	15416505	Mike Huckabee	GovMikeHuckabee	6	6	Republican	2008	USA	True
6	13398358932008	1339835893	Hillary Clinton	HillaryClinton	4	2	Democratic	2008	USA	False
7	2874135692008	287413569	Ron Paul	RonPaul	9	1	Republican	2008	USA	False
8	193941882008	19394188	John McCain	SenJohnMcCain	6	4	Republican	2008	USA	False
9	207130612008	20713061	Newt Gingrich	newtgingrich	8	7	Republican	2008	USA	False
10	7.526389690959995e...	752638969095999489	Mike Gravel	MikeGravel_US	8	-2	Democratic	2008	USA	False
11	196378212008	19637821	Alan Keyes	loyaltoliberty	6	8	Republican	2008	USA	False
12	500557012008	50055701	Mitt Romney	MittRomney	7	8	Republican	2008	USA	False
13	27049512008	2704951	Fred Thompson	fredthompson	7	7	Republican	2008	USA	True
14	163174062008	16317406	Chris Dodd	SenChrisDodd	3	4	Democratic	2008	USA	True
15	7.707819403412889e...	770781940341288960	Rudy Guiliani	RudyGiuliani	6	5	Republican	2008	USA	False
16	773146922008	77314692	Ralph Nader	RalphNader	-5	-3	Green	2008	USA	False
17	645349082008	64534908	Tom Tancredo	ttancredo	7	8	Republican	2008	USA	False

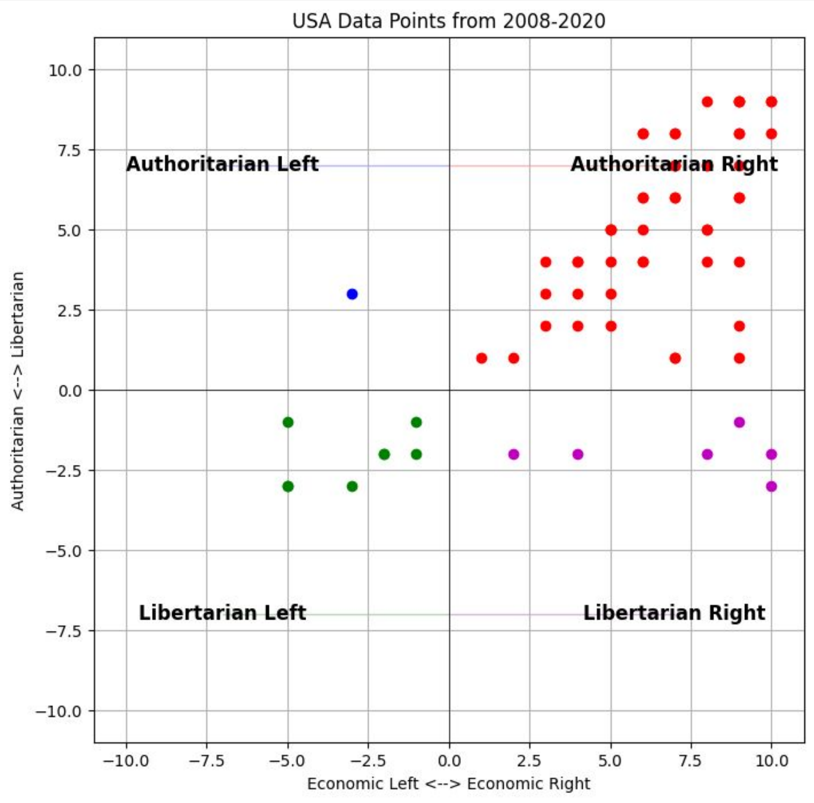
Scraped Data

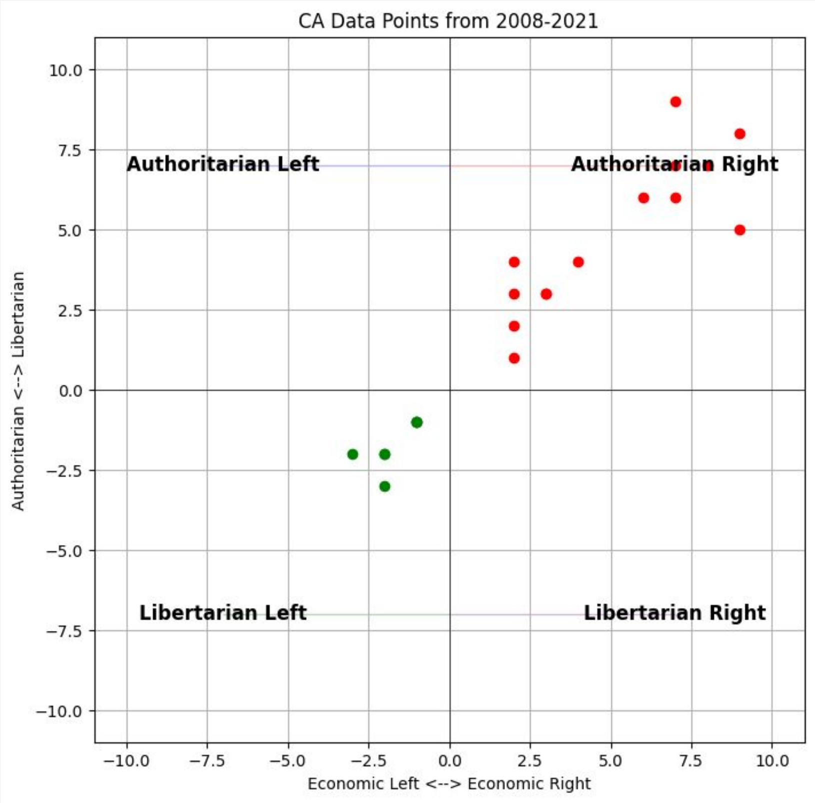


	tweet_id	user_id	user_name	user_handle	tweet_text	tweet_original_...	tweet_translated	tweet_translate...	created_date	election_year
	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...
1	6020502436	813286	Barack Obama	BarackObama	Reform must control ...	en	False	null	2009-11-24 21:56:08...	2008
2	5727126835	813286	Barack Obama	BarackObama	In Singapore, continu...	en	False	null	2009-11-15 03:42:26...	2008
3	6240488656	813286	Barack Obama	BarackObama	Tune in tonight for a...	en	False	null	2009-12-01 17:03:59...	2008
4	6308039277	813286	Barack Obama	BarackObama	The Jobs and Econo...	en	False	null	2009-12-03 16:38:28...	2008
5	6063766829	813286	Barack Obama	BarackObama	Video: The season's l...	en	False	null	2009-11-25 22:52:02...	2008
6	5897470978	813286	Barack Obama	BarackObama	The senate has unveil...	en	False	null	2009-11-20 19:05:21...	2008
7	6946577798	813286	Barack Obama	BarackObama	Forget to mail your h...	en	False	null	2009-12-23 00:07:10...	2008
8	6347526119	813286	Barack Obama	BarackObama	We still have a long ...	en	False	null	2009-12-04 20:01:09...	2008
9	5525033325	813286	Barack Obama	BarackObama	RT @JimOberstar: He...	en	False	null	2009-11-08 05:08:47...	2008
10	7039536487	813286	Barack Obama	BarackObama	To all those gathered...	en	False	null	2009-12-25 19:06:53...	2008
11	5524115324	813286	Barack Obama	BarackObama	RT @timryan House ...	en	False	null	2009-11-08 04:25:07...	2008
12	6743120620	813286	Barack Obama	BarackObama	The stakes are too hi...	en	False	null	2009-12-16 21:45:43...	2008
13	6578416131	813286	Barack Obama	BarackObama	Send a holiday card t...	en	False	null	2009-12-11 20:53:54...	2008
14	6084583071	813286	Barack Obama	BarackObama	From my family to yo...	en	False	null	2009-11-26 15:52:53...	2008
15	5524151229	813286	Barack Obama	BarackObama	RT @chelliepingree ...	en	False	null	2009-11-08 04:26:46...	2008
16	6316546945	813286	Barack Obama	BarackObama	The National Christm...	en	False	null	2009-12-03 22:03:52...	2008
17	6907408875	813286	Barack Obama	BarackObama	RT @SenatorReid: I'm...	en	False	null	2009-12-21 22:26:05...	2008

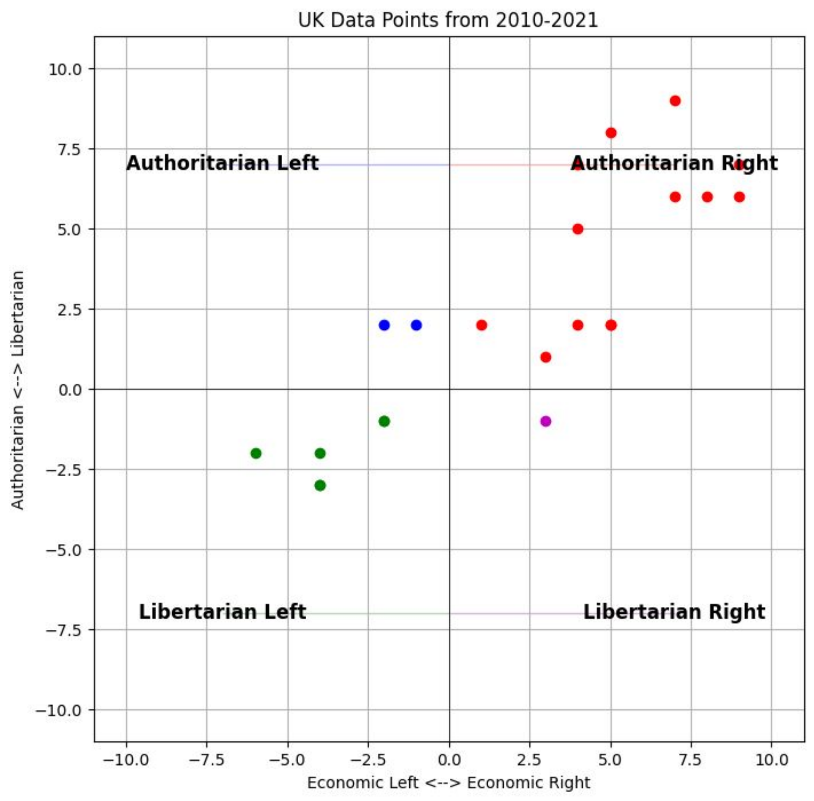
USA

Canada

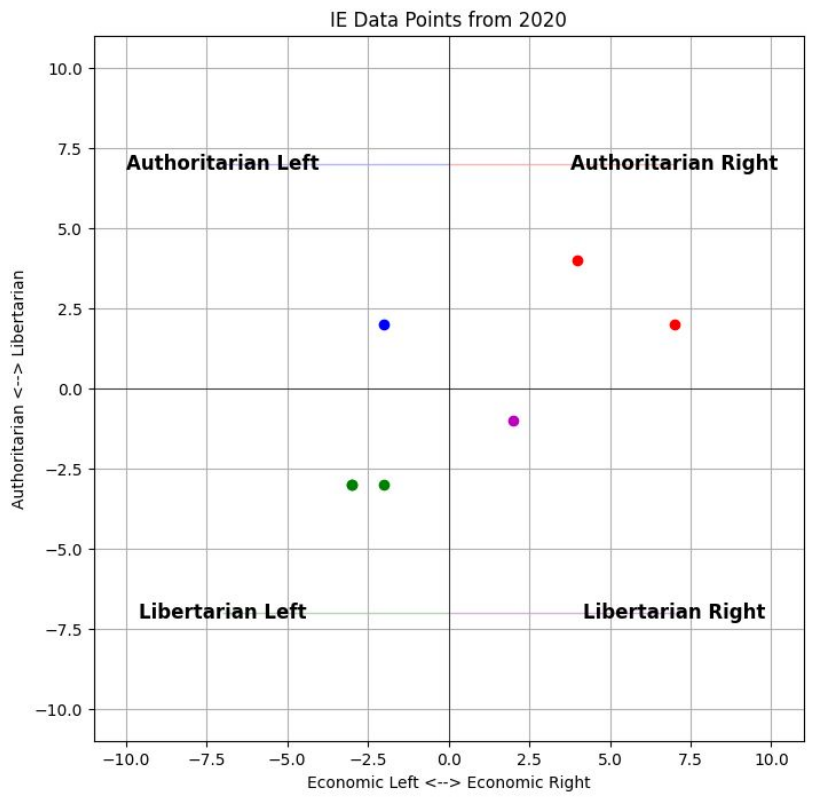

USA Data Points from 2008-2020
Authoritarian Left
Authoritarian Right
Libertarian Left
Libertarian Right
Authoritarian <--> Libertarian
Economic Left <--> Economic Right
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0


CA Data Points from 2008-2021
Authoritarian Left
Authoritarian Right
Libertarian Left
Libertarian Right
Authoritarian <--> Libertarian
Economic Left <--> Economic Right
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0

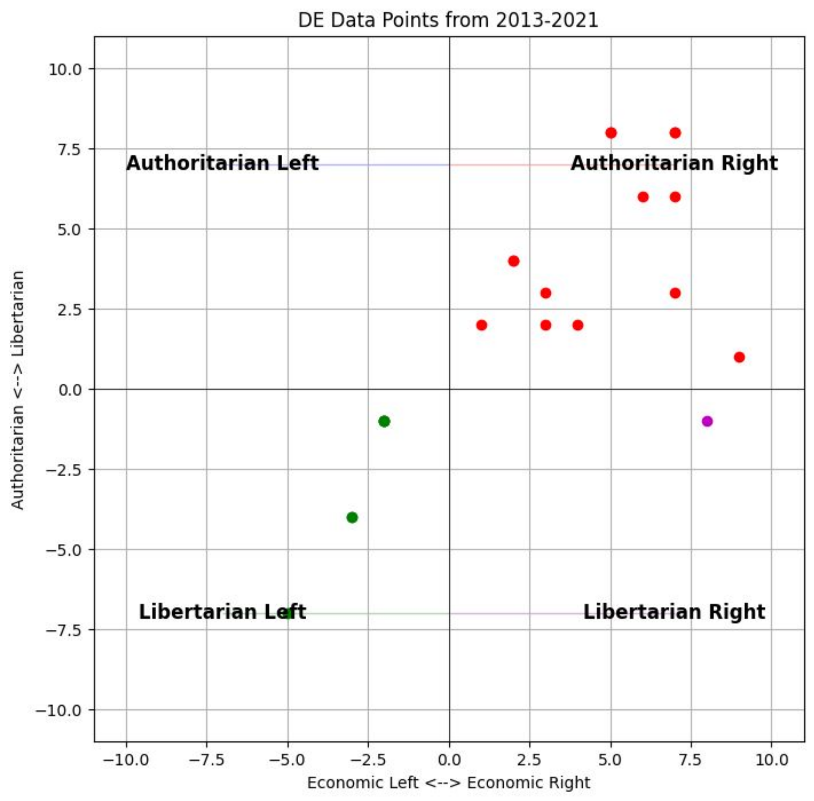
United Kingdom



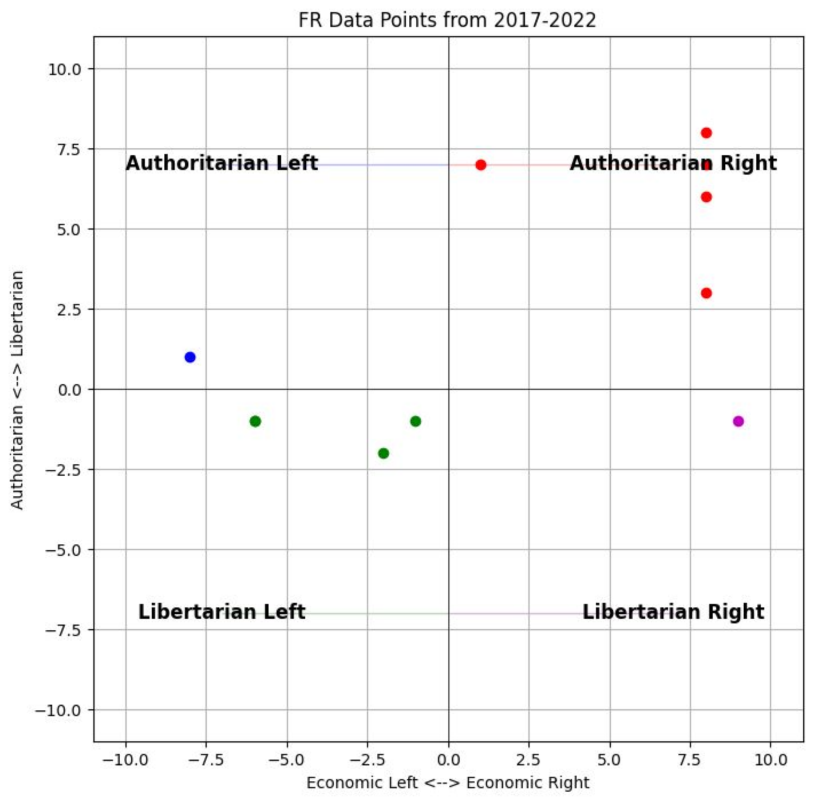
Ireland



Germany

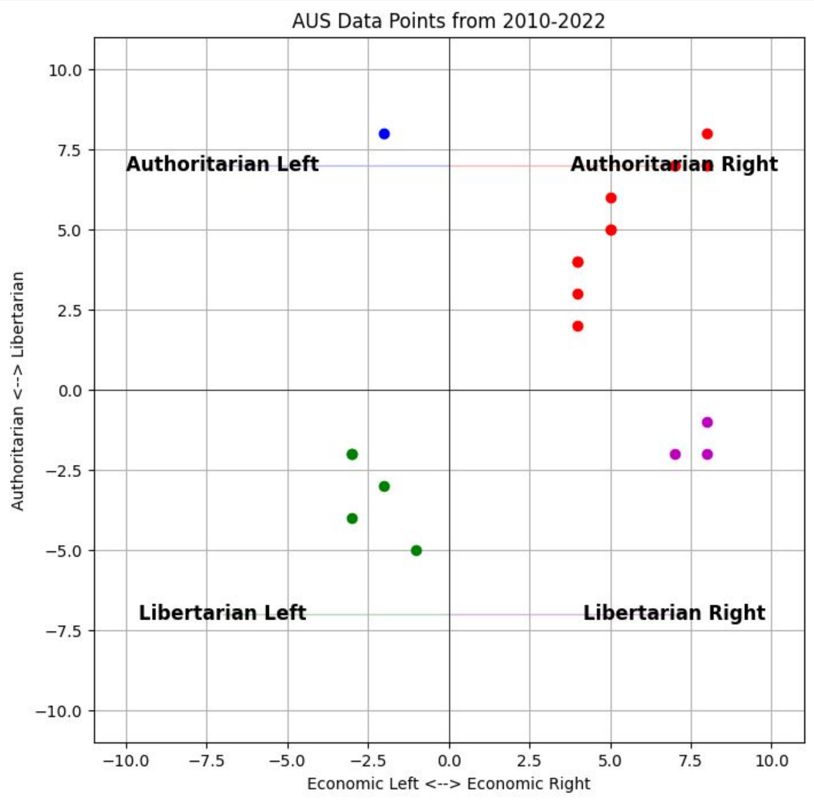


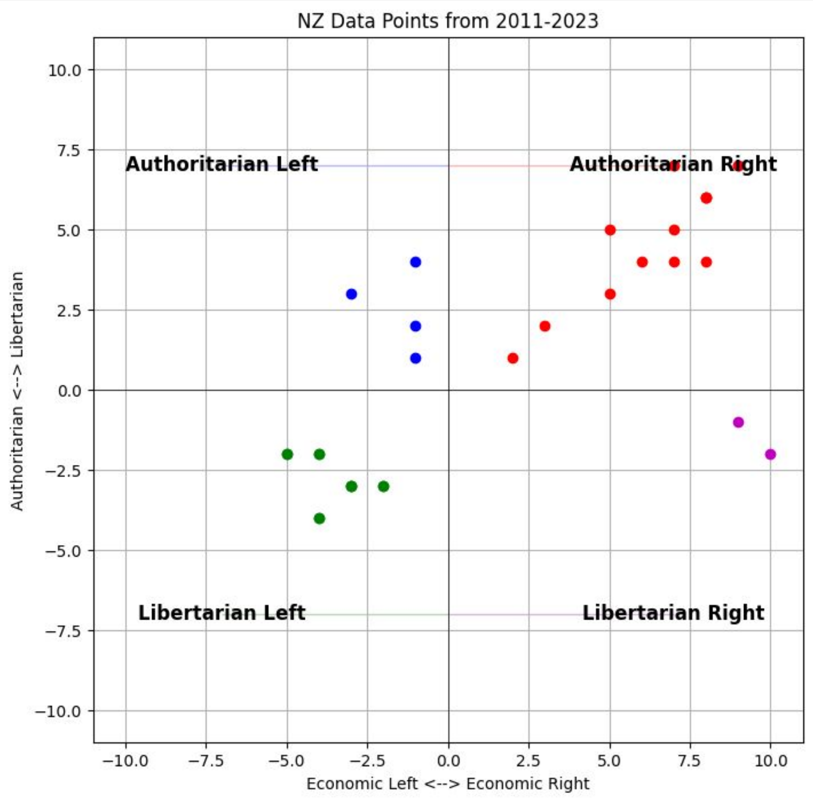
France

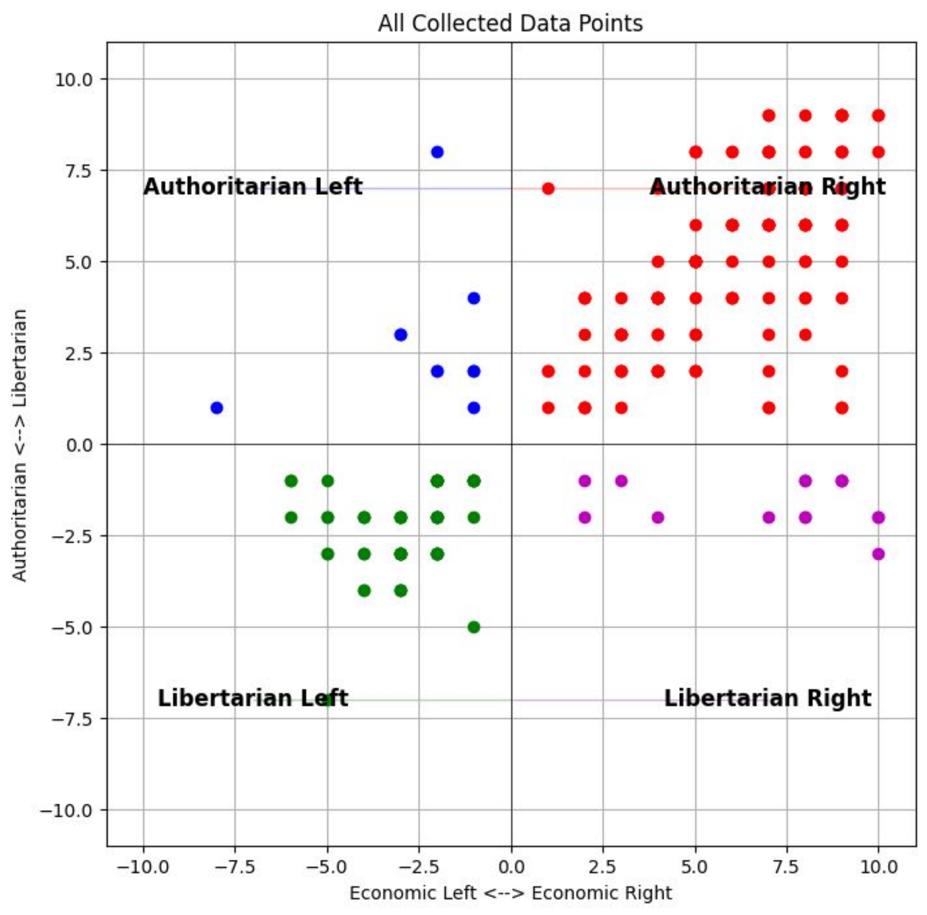


Australia

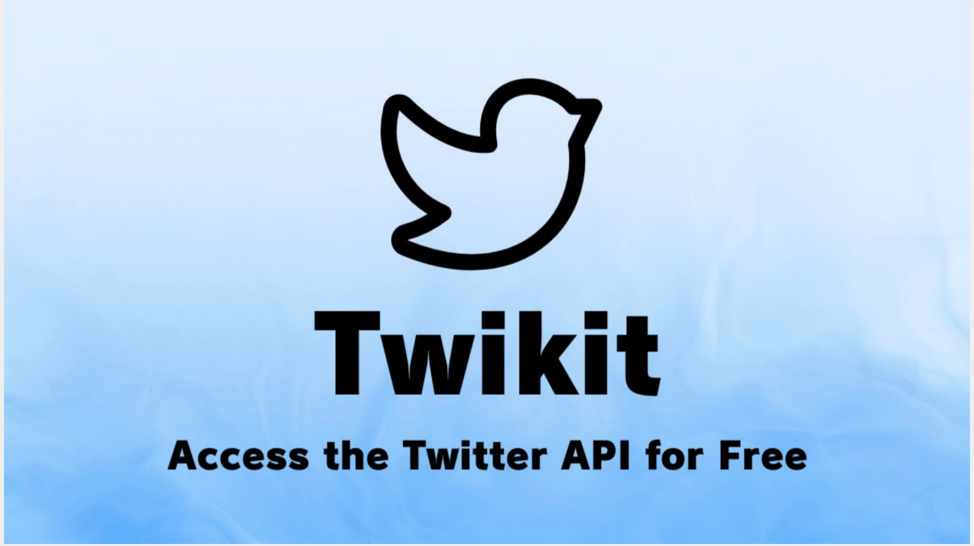
New Zealand





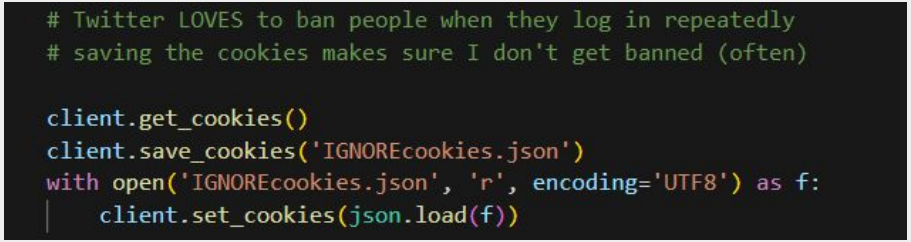

All Collected Data Points
Authoritarian Left
Authoritarian Right
Libertarian Left
Libertarian Right
Authoritarian <--> Libertarian
Economic Left <--> Economic Right
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Subverting Paywalls using API Wrappers



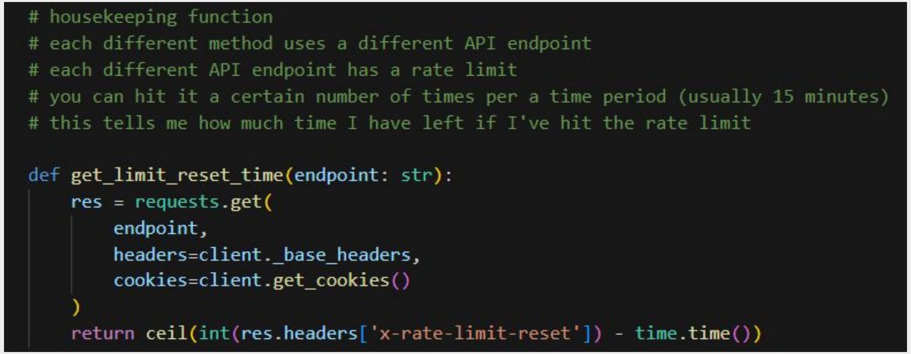
- Twitter API has been paywalled since 2023
- Twitter API wrappers like Twikit can access endpoints for free
- Has pros and cons

API Rate Limits and Authentication Issues



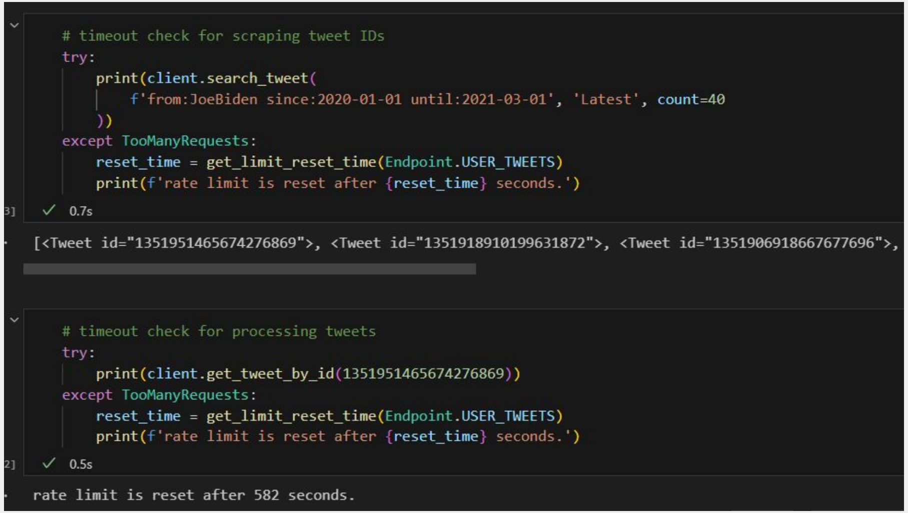
```
# Twitter LOVES to ban people when they log in repeatedly
# saving the cookies makes sure I don't get banned (often)

client.get_cookies()
client.save_cookies('IGNOREcookies.json')
with open('IGNOREcookies.json', 'r', encoding='UTF8') as f:
    client.set_cookies(json.load(f))
```



```
# housekeeping function
# each different method uses a different API endpoint
# each different API endpoint has a rate limit
# you can hit it a certain number of times per a time period (usually 15 minutes)
# this tells me how much time I have left if I've hit the rate limit

def get_limit_reset_time(endpoint: str):
    res = requests.get(
        endpoint,
        headers=client._base_headers,
        cookies=client.get_cookies()
    )
    return ceil(int(res.headers['x-rate-limit-reset']) - time.time())
```



```
# timeout check for scraping tweet IDs
try:
    print(client.search_tweet(
        f'from:JoeBiden since:2020-01-01 until:2021-03-01', 'Latest', count=40
    ))
except TooManyRequests:
    reset_time = get_limit_reset_time(Endpoint.USER_TWEETS)
    print(f'rate limit is reset after {reset_time} seconds.')
```

3] ✓ 0.7s

```
[<Tweet id="1351951465674276869">, <Tweet id="1351918910199631872">, <Tweet id="1351906918667677696">,
```

```
# timeout check for processing tweets
try:
    print(client.get_tweet_by_id(1351951465674276869))
except TooManyRequests:
    reset_time = get_limit_reset_time(Endpoint.USER_TWEETS)
    print(f'rate limit is reset after {reset_time} seconds.')
```

2] ✓ 0.5s

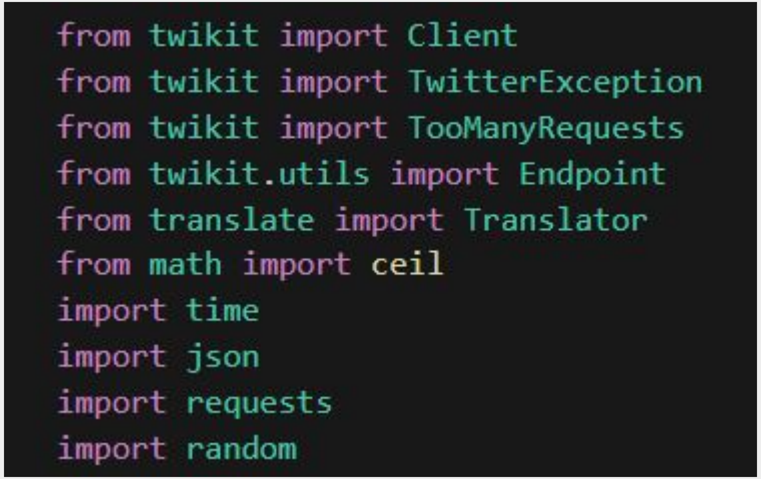
```
rate limit is reset after 582 seconds.
```

Other tools we used

Python has a built in SQLite3 library.

SQL databases are a great way to store large amounts of structured data.

We used several different python libraries and modules to make our code run faster or to get more information about rate limits.



```
from twikit import Client
from twikit import TwitterException
from twikit import TooManyRequests
from twikit.utils import Endpoint
from translate import Translator
from math import ceil
import time
import json
import requests
import random
```

Cleaning and preprocessing the data

Iteration 1 - CountVectorizer

- Simplest option
- Just counting occurrences of words

*Iteration 2 - GloVe Embeddings

- More complex option
- Capture more information about data

*Iteration 3 - BERT

- Most complex option
- Interesting to see how much context matters in short form text data

* potential iterations

Linear Regression to Predict X and Y coordinates

Two linear regression models

- Both will be trained and tested on the same dataset
- One will predict the X (economic) coordinate, the other will predict the Y (social) coordinate

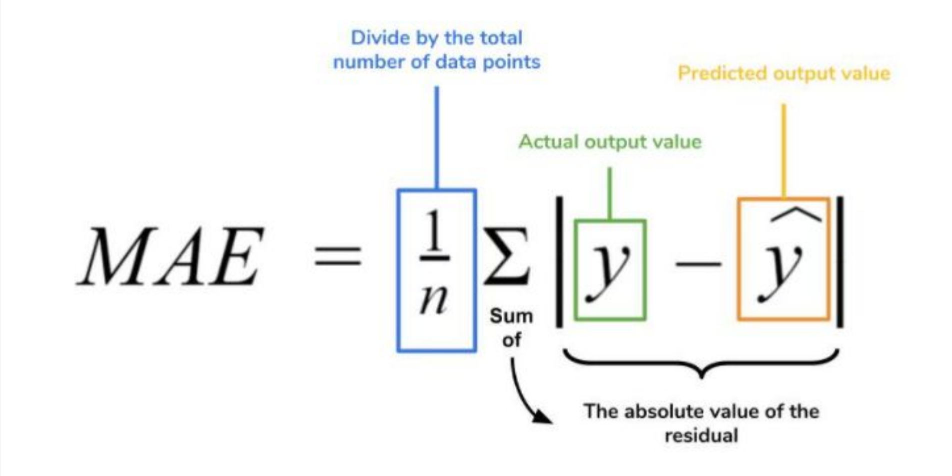
Why linear regression?

- We need a continuous output. Each axis goes from -10 to 10, so there are 21 potential outputs.
- Simplicity in implementation and interpretation.

How are we evaluating the output?

Mean Absolute Error

- Measures the average size of the mistakes in a collection of predictions.
- With the scale of -10 to 10, we hope to keep the MAE within 2 points as an acceptable range for error.


Divide by the total number of data points
Predicted output value
Actual output value
$MAE = \frac{1}{n}\sum\left|y - \hat{y}\right|$
Sum of
The absolute value of the residual

Questions?