Open-Source Intelligence

A Gentle Introduction

Anton Sobolev

UT Dallas

Data Science for Social Inquiry, June 2023

Social Inquiry Challenges: Causality & Data

How do we know: X ⇝ Y, not Y ⇝ X ?

□ **Causality**

  □ Ice-Cream ⇝ Shark Attacks?

  □ Economic Growth ⇝ Democracy?

□ **Huge Progress**

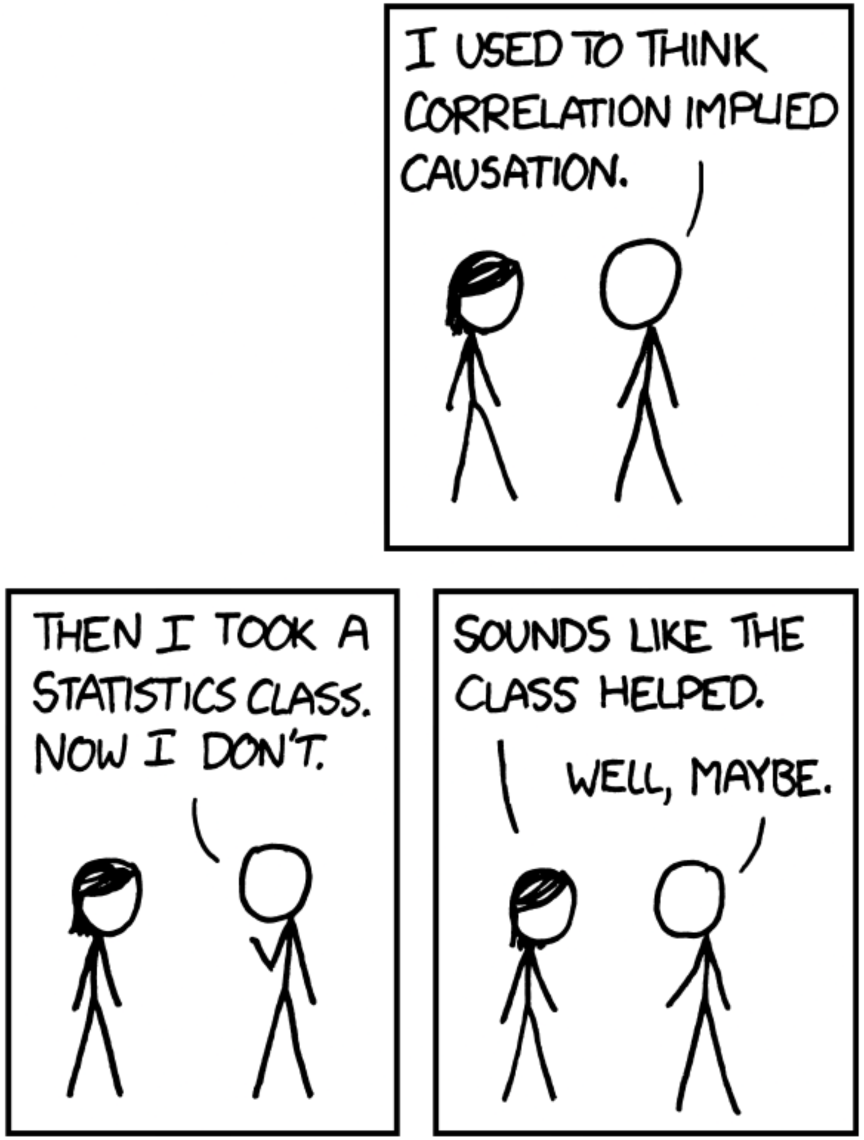
  □ Quasi-Experiments

  □ Causal Graphs:
  Account for Alternatives

  □ Causality as Missing Data

□ **Main Factor of CI success**

  □ Medical Studies!


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

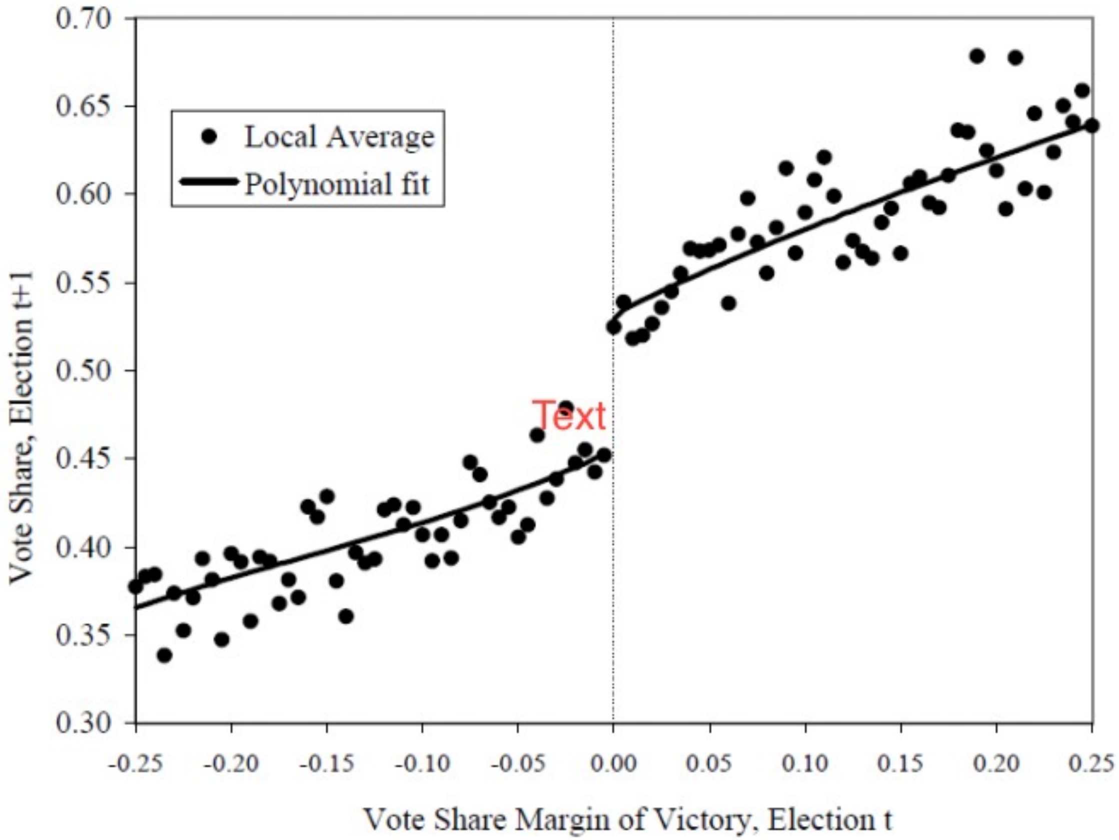
Social Inquiry Challenges: Causality & Data

How do we know: $X \rightsquigarrow Y$, not $Y \rightsquigarrow X$?

- □ **Causality**
  - □ Ice-Cream ⇝ Shark Attacks?
  - □ Economic Growth ⇝ Democracy?

- □ **Huge Progress**
  - □ Quasi-Experiments
  - □ Causal Graphs:
    Account for Alternatives
  - □ Causality as Missing Data

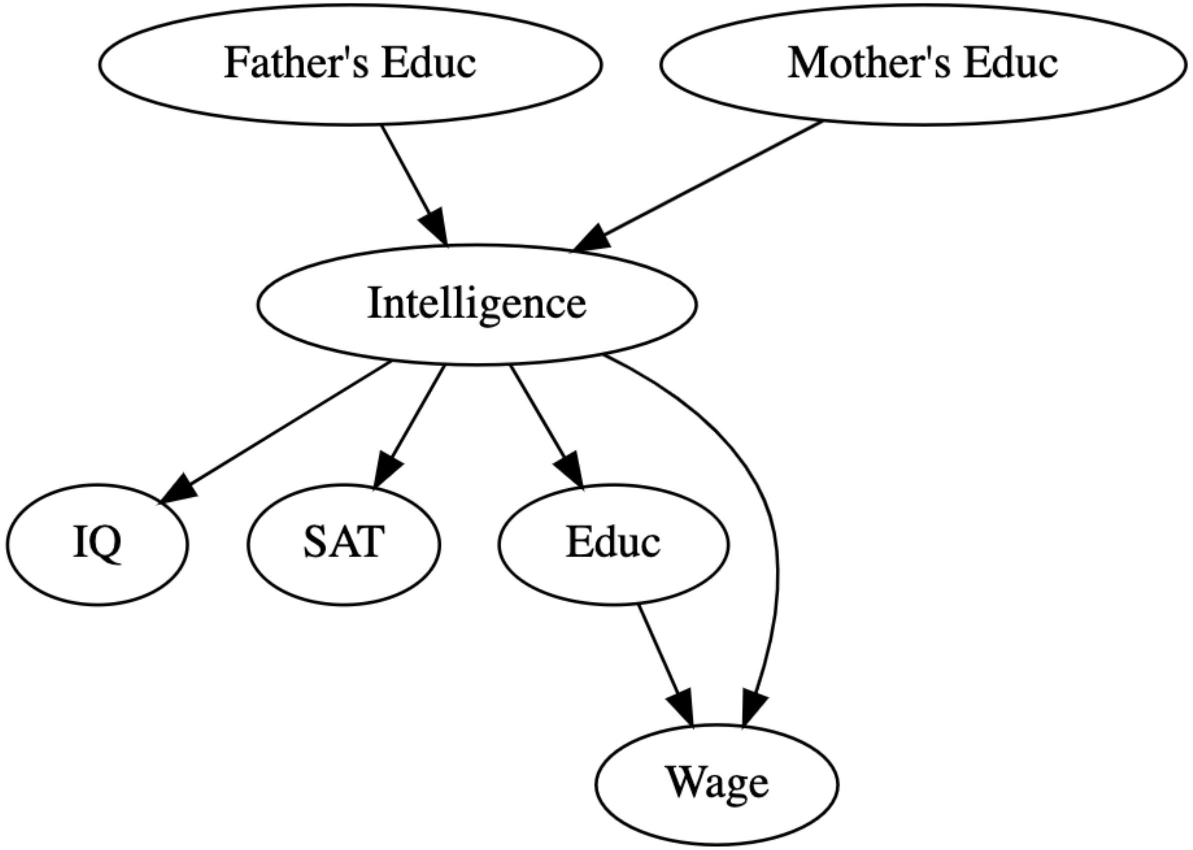
- □ **Main Factor of CI success**
  - □ Medical Studies!



Social Inquiry Challenges: Causality & Data

How do we know: X ⇝ Y, not Y ⇝ X ?

- □ **Causality**
  - □ Ice-Cream ⇝ Shark Attacks?
  - □ Economic Growth ⇝ Democracy?
- □ **Huge Progress**
  - □ Quasi-Experiments
  - □ Causal Graphs: Account for Alternatives
  - □ Causality as Missing Data
- □ **Main Factor of CI success**
  - □ Medical Studies!



Social Inquiry Challenges: Causality & Data

How do we know: $X \rightsquigarrow Y$, not $Y \rightsquigarrow X$?

- □ **Causality**
  - □ Ice-Cream ⇝ Shark Attacks?
  - □ Economic Growth ⇝ Democracy?
- □ **Huge Progress**
  - □ Quasi-Experiments
  - □ Causal Graphs: Account for Alternatives
  - □ Causality as Missing Data
- □ **Main Factor of CI success**
  - □ Medical Studies!

Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	0.3
P2	0	0.8	0.6
P3	1	0.3	0.2
P4	0	0.3	0.1
P5	1	0.5	0.5
P6	0	0.6	0.5
P7	0	0.3	0.1

Social Inquiry Challenges: Causality & Data

How ~~the heck~~ do we get data?

□ **Issue #1: Measurement**

- □ Political Ideology
- □ Racism & Job Discrimination
- □ Media Bias
- □ Kid's Trauma
- □ Anti-Dictator Attitudes

New School is a part of of OSINT

□ **Issue #2: Get Data**

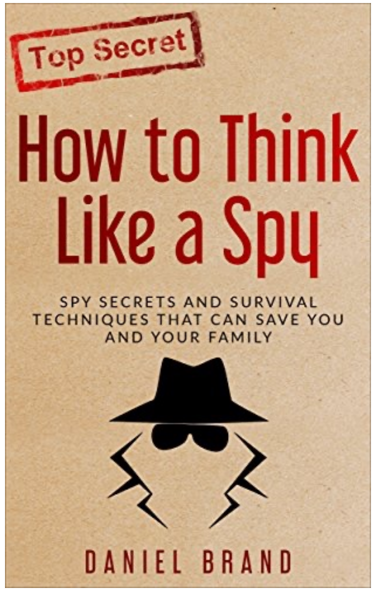
Old School

- □ **Collected by someone:** Public Statistics
- □ **Created by yourself:** Surveys

New School

- □ **Auto-Generated:** Social Media,
  CCTV, Cellphones

OSINT: Very-Very New Phenomena ~~[well,almost]~~

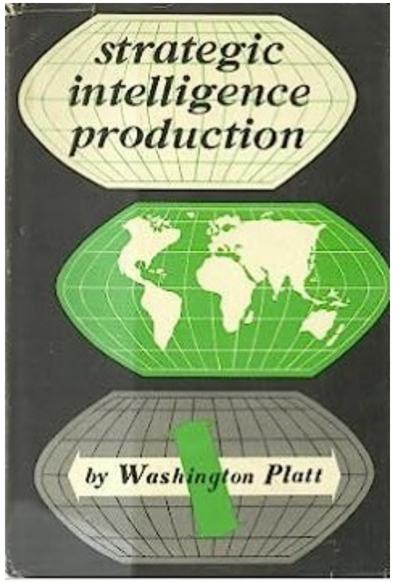


Top Secret

How to Think Like a Spy

SPY SECRETS AND SURVIVAL TECHNIQUES THAT CAN SAVE YOU AND YOUR FAMILY

DANIEL BRAND

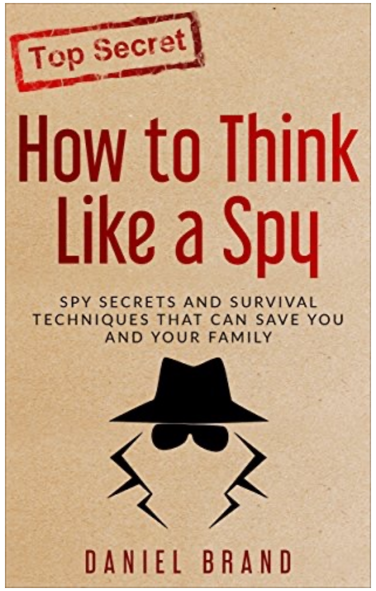


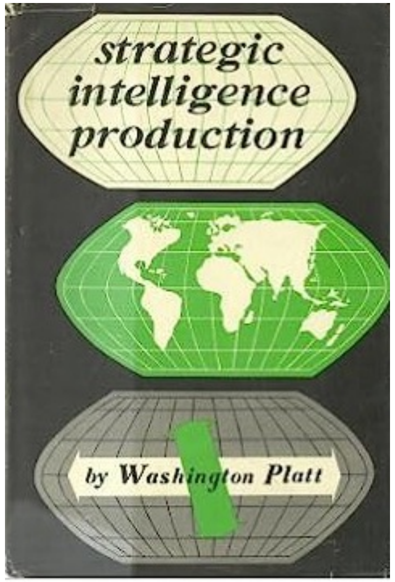
strategic intelligence production

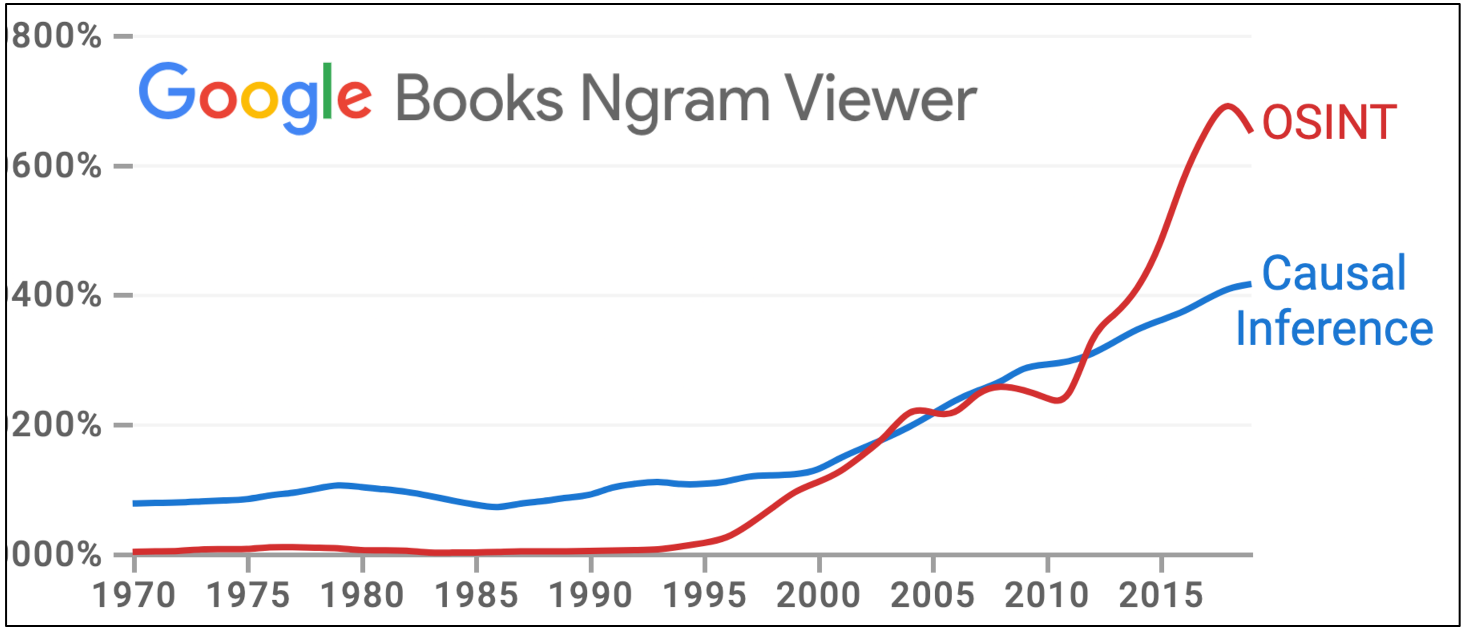
by Washington Platt

- □ **Kremlinology**
  - □ Soviet Statistics Lies
    (China today?)
  - □ Total Control: Spying is hard
    [in contrast to soviet spies in the US]

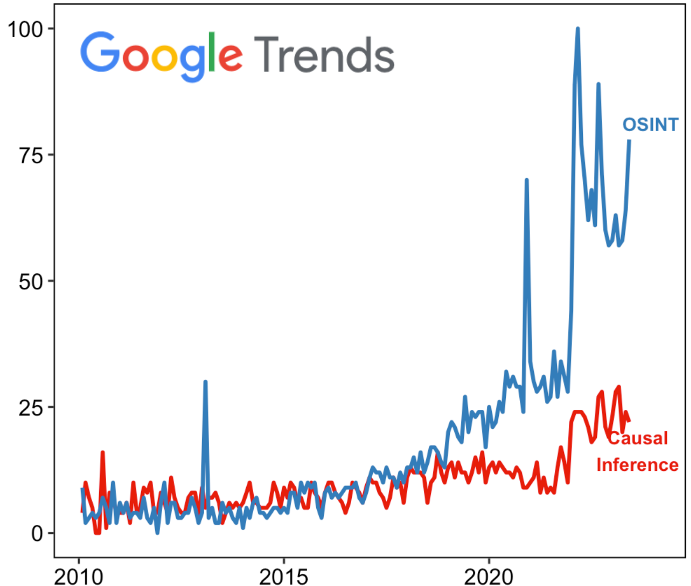
OSINT: Very-Very New Phenomena ~~[well,almost]~~


Top Secret
How to Think Like a Spy
SPY SECRETS AND SURVIVAL TECHNIQUES THAT CAN SAVE YOU AND YOUR FAMILY
DANIEL BRAND





Google Queries Related to ...



OSINT Applications

- □ **International Relations**
  - □ Open-Dat **“**Espionage”
  - □ Military operations
  - □ Anti-terrorism
- □ **Corporate Sector:**
  - □ Competitive Intelligence (Uber)
  - □ Market strategy
  - □ Military operations
- □ **Public Policy:**
  - □ Sensetive Issues (teenage pregnancy / racism / bullying)
- □ **Criminal Investigations**
  - □ Crypto-investigations

    good guys: “money laundering”
    bad guys: “Repress donors of political opposition”

OSINT Applications

- □ **International Relations**
  - □ Open-Dat “Espionage”
  - □ Military operations
  - □ Anti-terrorism
- □ **Corporate Sector:**
  - □ Competitive Intelligence (Uber)
  - □ Market strategy
  - □ Military operations
- □ **Public Policy:**
  - □ Sensetive Issues (teenage pregnancy / racism / bullying)
- □ **Criminal Investigations**
  - □ Crypto-investigations
    good guys: “money laundering”
    bad guys: “Repress donors of political opposition”

And Social Sciences!

Get Data Exmple: Protest Behavior

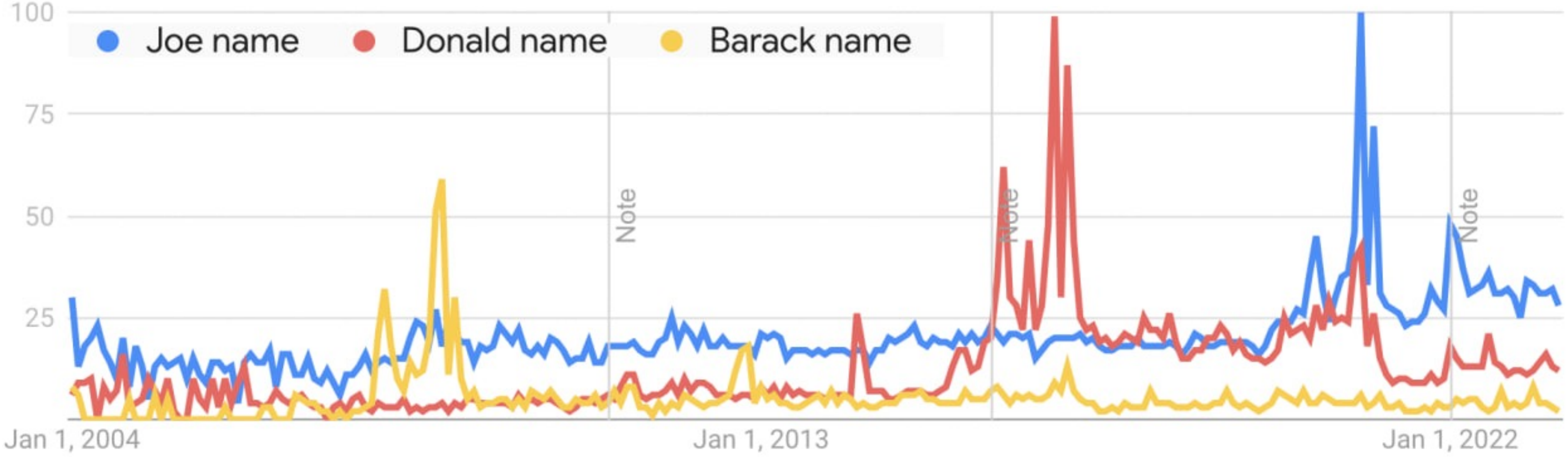
- □ **Old School**
  - □ Resoucres

- □ **New school**
  - □ Understanding the case
  - □ Data Generated around

Get Data Exmple: Protest Behavior

- □ **Old School**
  - □ Resoucres

- □ **New school**
  - □ Understanding the case
  - □ Data Generated around


● Joe name
● Donald name
● Barack name
100
75
50
25
Note
Note
Note
Jan 1, 2004
Jan 1, 2013
Jan 1, 2022

How the heck we get the data?

□ In location x_i, individuals who look for protest campaign information *also* search:

- "revolution"
- "anti-corruption reform"

□ **Proposed Approach**

- Identify the largest cluster *[robust to outliers]*
- Calculate cluster's centroid *[n-dimensional space]*
- **Fragmentation Score**: average distance to the centroid *[Manhattan distance]*

Old School

New School

How the heck we get the data?

□ In location x_i, individuals who look for protest campaign information *also* search:
- “revolution”
- “anti-corruption reform”

□ **Proposed Approach**
- Identify the largest cluster *[robust to outliers]*
- Calculate cluster’s centroid *[n-dimensional space]*
- **Fragmentation Score**: average distance to the centroid *[Manhattan distance]*

Theory

□ In location x_i, individuals who look for protest campaign information *also* search:

- "revolution"
- "anti-corruption reform"

Useful Tools by BellingCat

This Paper

□ Initial Research Questions Not this paper!

- Does lack of unified agenda among protesters reduce chances of campaign's success? *[Protest Fragmentation Hypothesis]*
- Do scholars mistakenly categorize de-facto separate campaigns as a single entity? *["Under The Same Flag" Hypothesis]*

□ Current Goal: Method to estimate campaign fragmentation*

□ Desired Properties

- **Behavior-based measure:** ~~media reports~~, ~~surveys~~, ~~expert opinions~~
- **Explicit interpretation:** ~~Likert scale~~, ~~composite measures~~ *[Polity IV]*
- **Comparability:** cross-country / cross-campaign comparison

* **Campaign Fragmentation** – variation in the goals of a protest campaign among protesters

Focus

□ Sub-national differences in the demand for information related to a protest campaign

└ **Assumption:** High variation ⇝ High protest fragmentation

□ Correlated behaviors

└ **Key idea:** Individuals who look for the same information related to a protest campaign share similar views regarding the goals of this campaign

□ Implementation

├ **Input data:** Search queries (Google Trends)

└ **Key feature:** Ability to identify *other* search queries individuals conduct when they seek for protest-campaign information

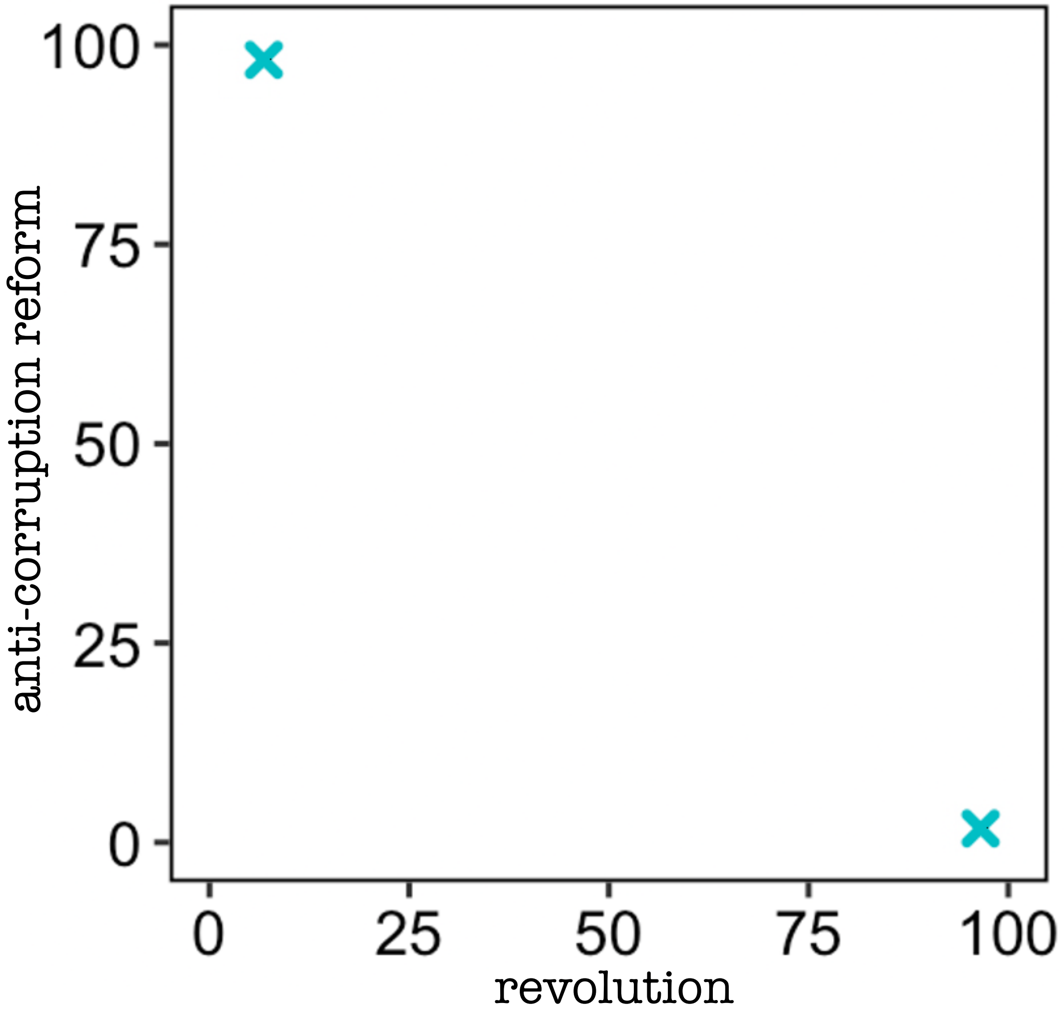
Theory

□ In location x_i, individuals who look for protest campaign information *also* search:

- "revolution"
- "anti-corruption reform"

Dissimilar interest in topics
⇝ High protest campaign fragmentation

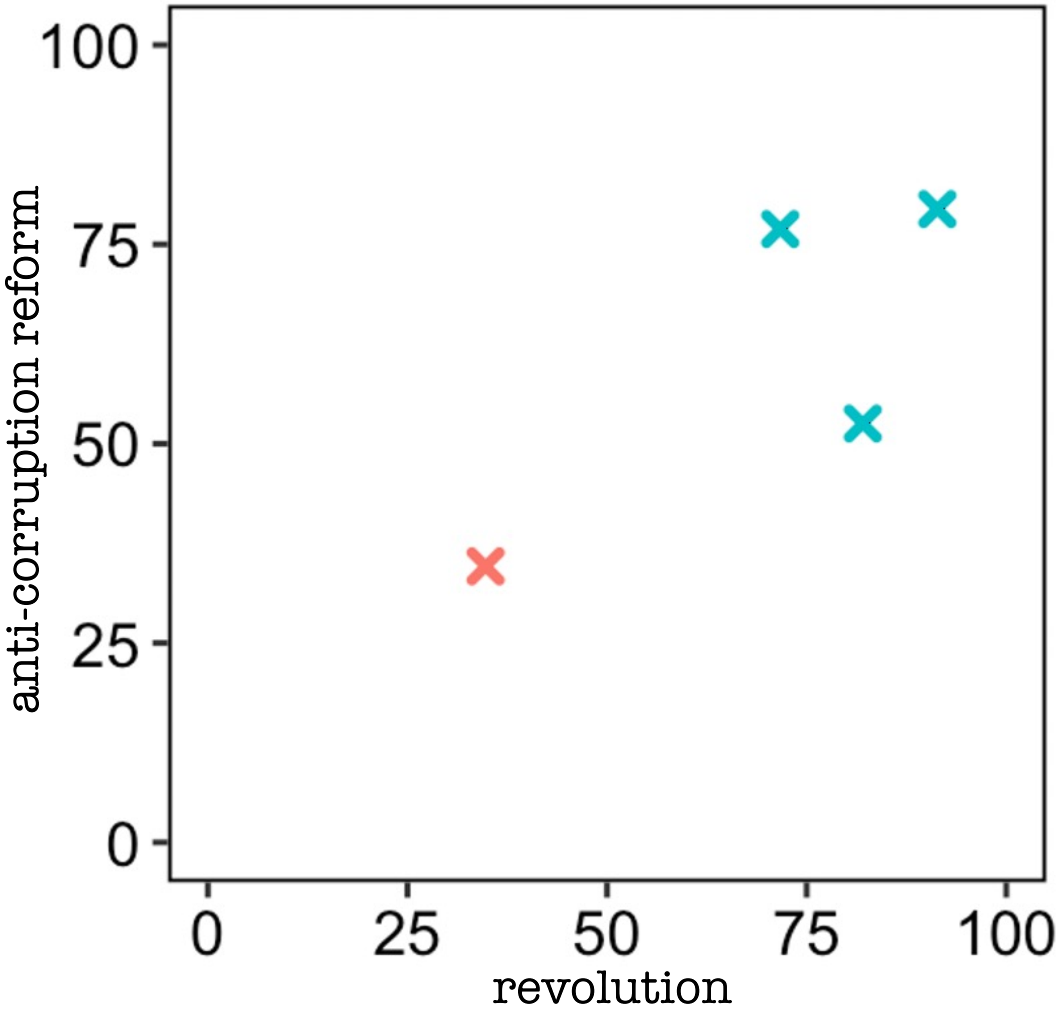
Case 2: High Fragmentation



Theory

□ In location x_i, individuals who look for protest campaign information *also* search:
- “revolution”
- “anti-corruption reform”

Case 3: Multiple Locations

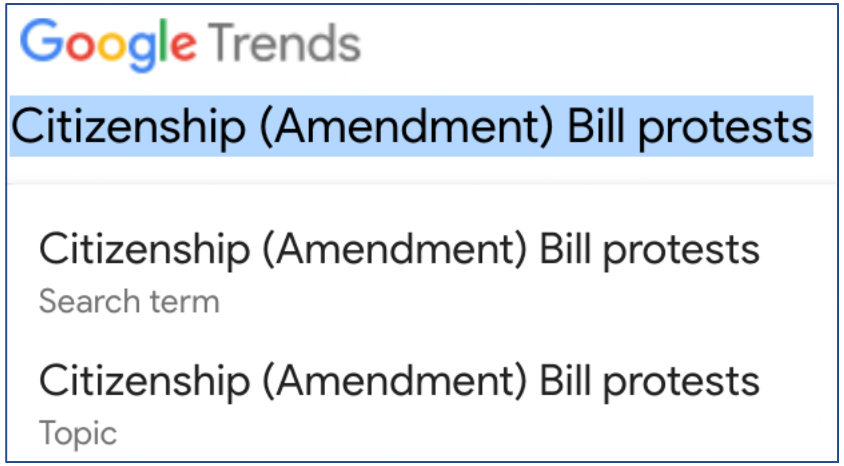


Implementation with G-Trends

Example: Citizenship Amendment Act protests (India, 2019)

□ **Google trends**

- Score [0-100] based on the volume of search queries
- Provides data for separate queries and queries aggregated into topics
- Identifies queries / topics correlated with the initial query / topic
- Subnational level data

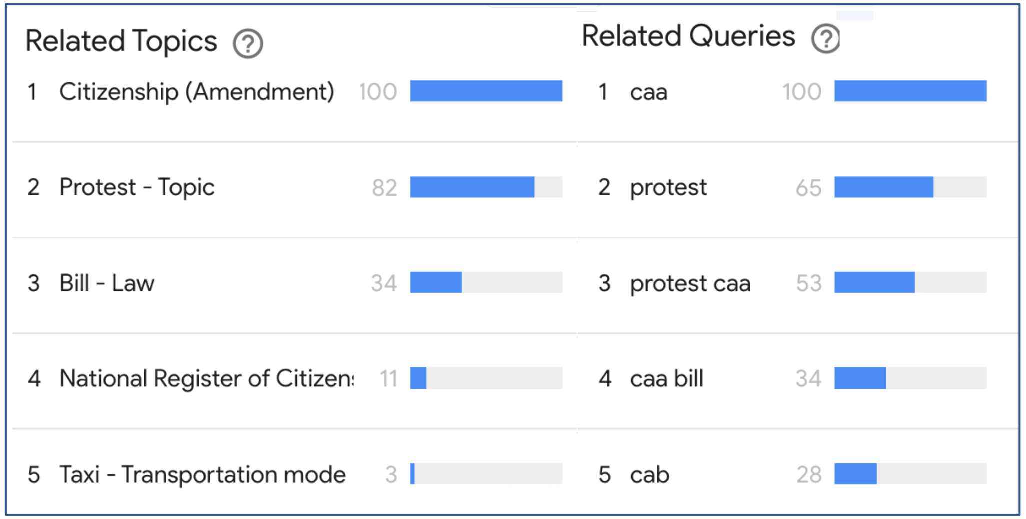


Google Trends

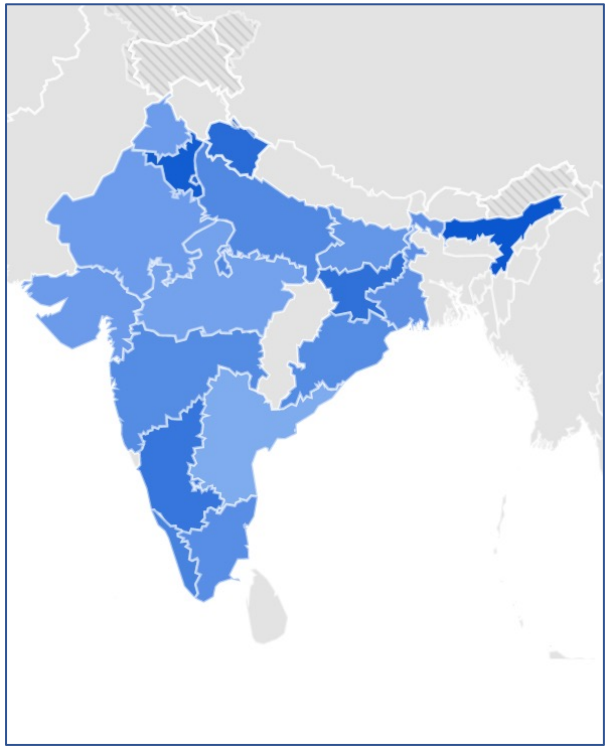
Citizenship (Amendment) Bill protests

Citizenship (Amendment) Bill protests
Search term

Citizenship (Amendment) Bill protests
Topic



Related Topics ⓘ		Related Queries ⓘ	
1 Citizenship (Amendment)	100	1 caa	100
2 Protest - Topic	82	2 protest	65
3 Bill - Law	34	3 protest caa	53
4 National Register of Citizen	11	4 caa bill	34
5 Taxi - Transportation mode	3	5 cab	28

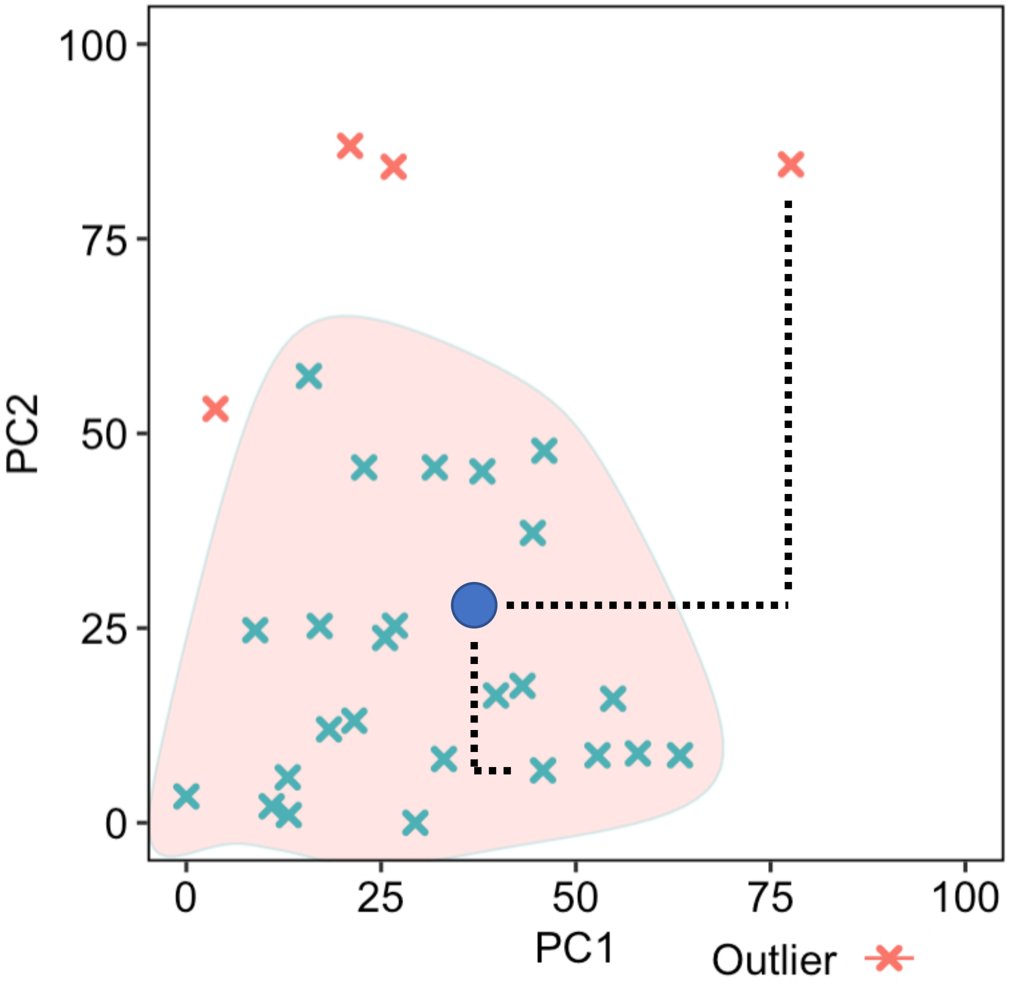


Implementation with G-Trends

Example: Citizenship Amendment Act protests (India, 2019)

□ **Citizenship Amendment Act protests (India, 2019)**

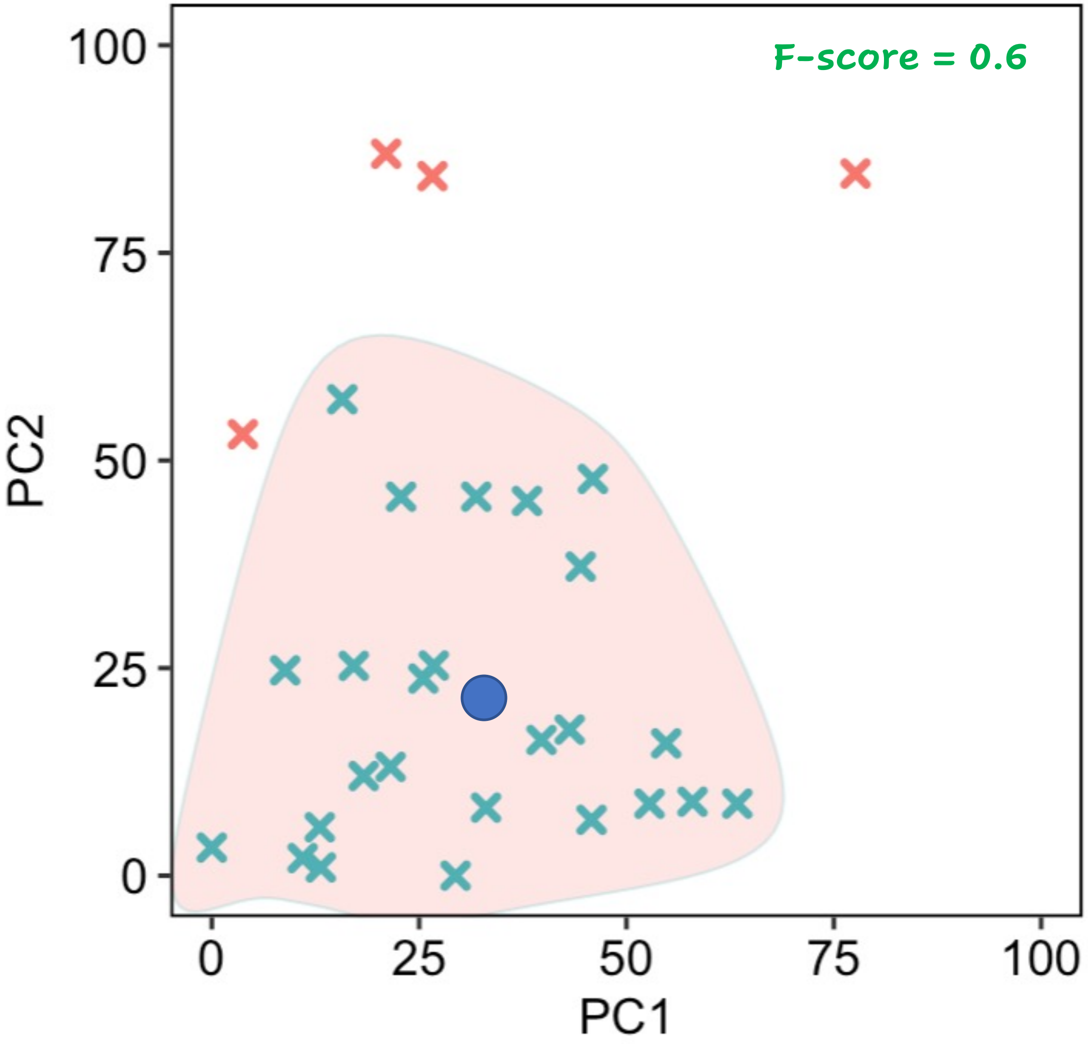
- Identify the protest movement topic
- Identify first 10 correlated topics
- Identify largest cluster *[via DBSCAN]*
- Calculate centroid
- Calculate mean distance D *[Manhattan]*
- Adjust $(100 - \frac{D}{2})/100$



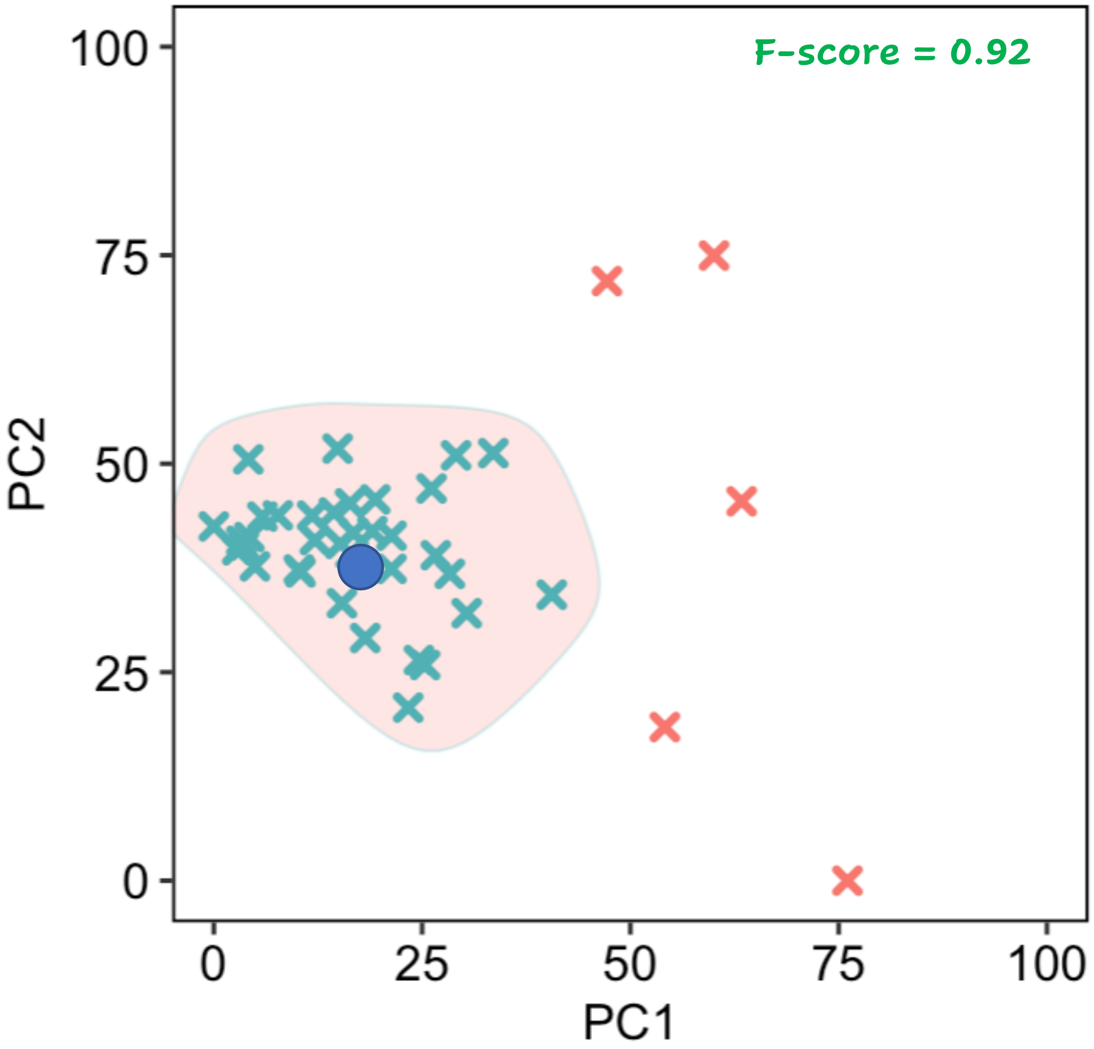
Fragmentation Score = 0.6

* Principal Components are used for illustration purposes

Comparison India-2019 *VS* US-2017

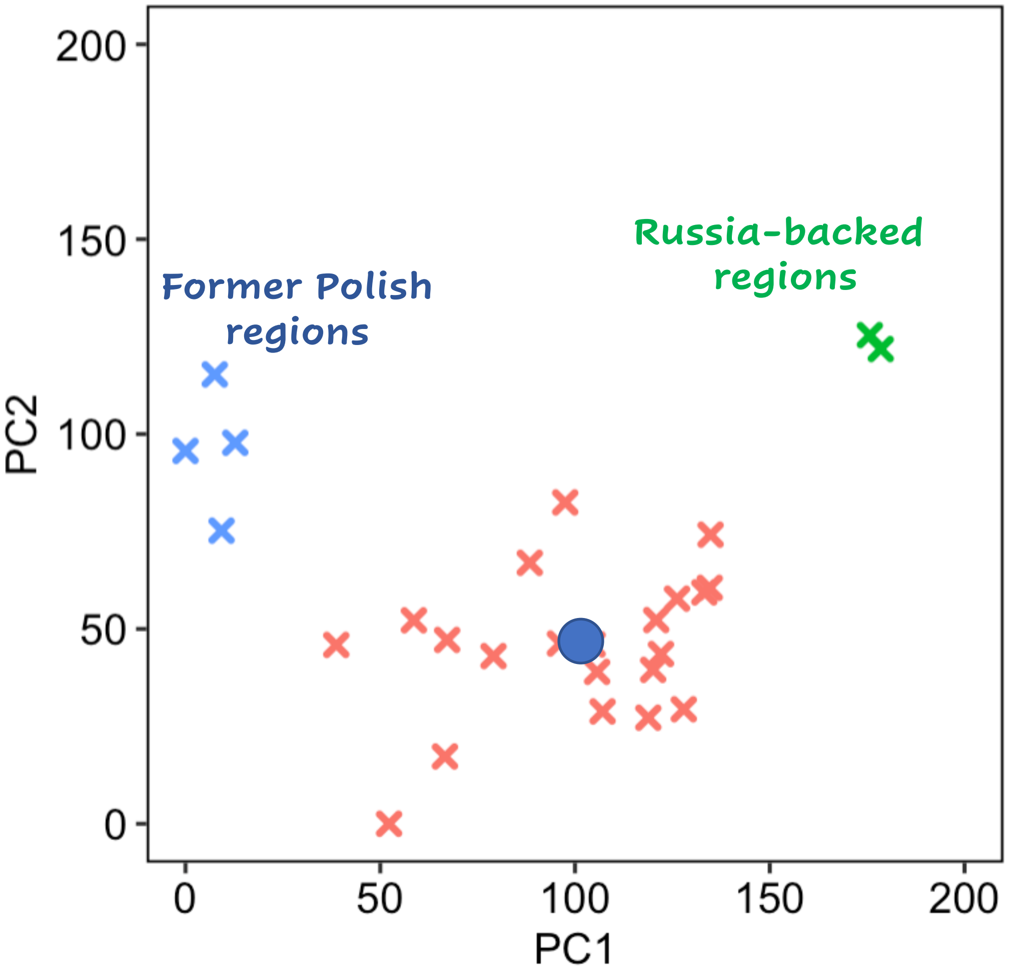


Citizenship Amendment Act protests (India, 2019)

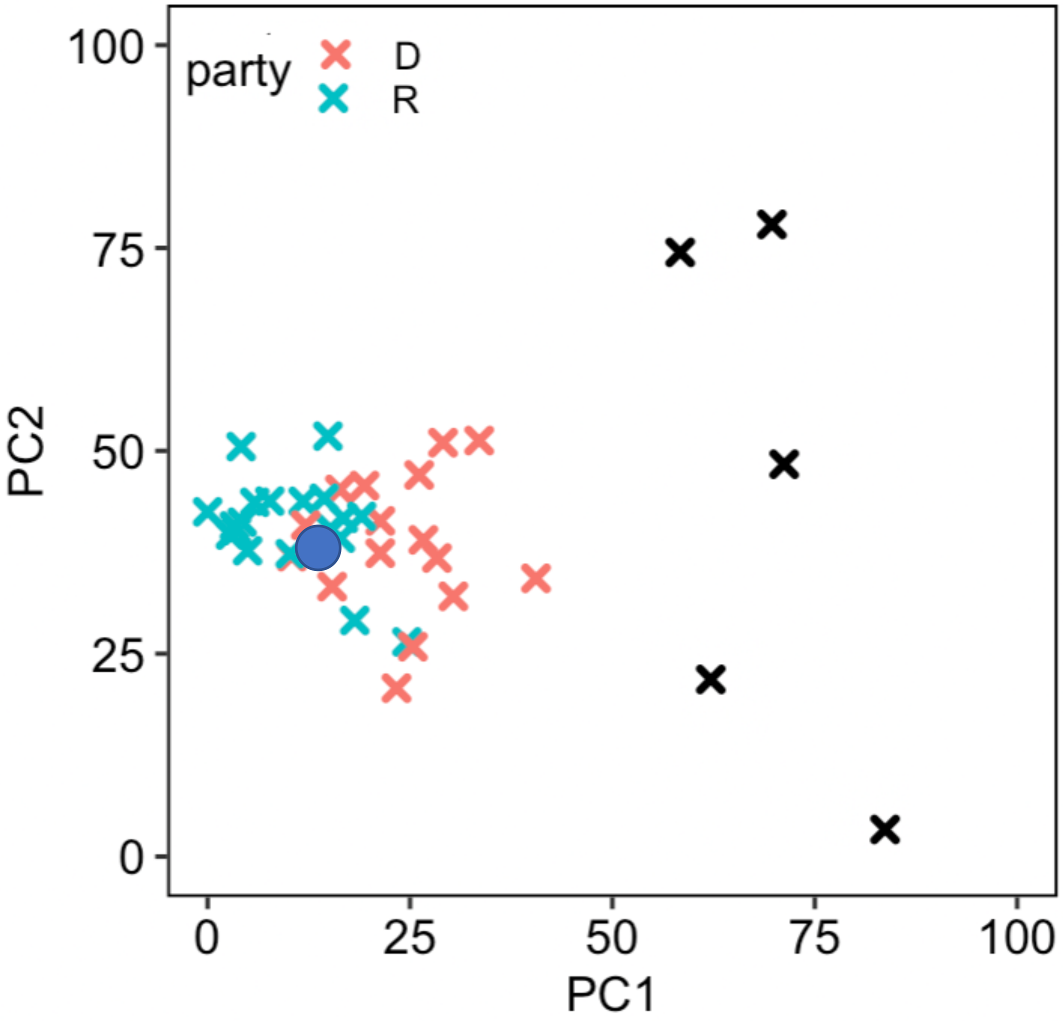


Women's March (US, 2017)

Validity Check Potentially Important Variables



Ukraine EuroMaidan (Ukraine, 2013-14)



Women's March (US, 2017)

Feedback, please 🙏

Thank you!

□ Next steps

- **Robustness:** □ How does the F-score change depending on the number of included correlated queries?
  □ Alternative clustering
- Correlation with closely-related measures: *NAVCO* [# of organizations, vertical/horizontal communication]

□ Does it make sense to …

- **Adjustment:** Adjust for the region population / internet users?
- **Multiple clusters:** should we calculate F-score separately for each?