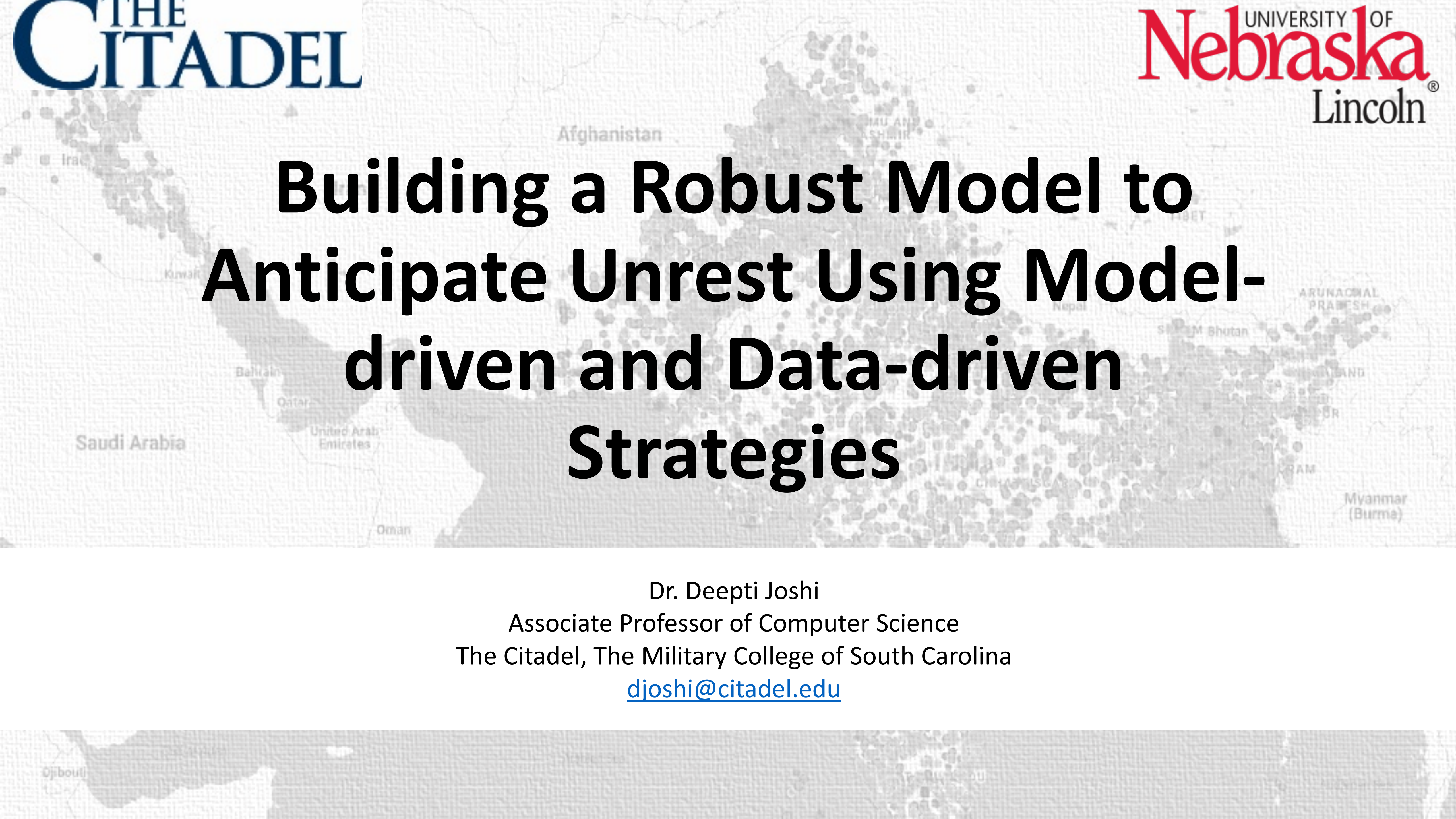


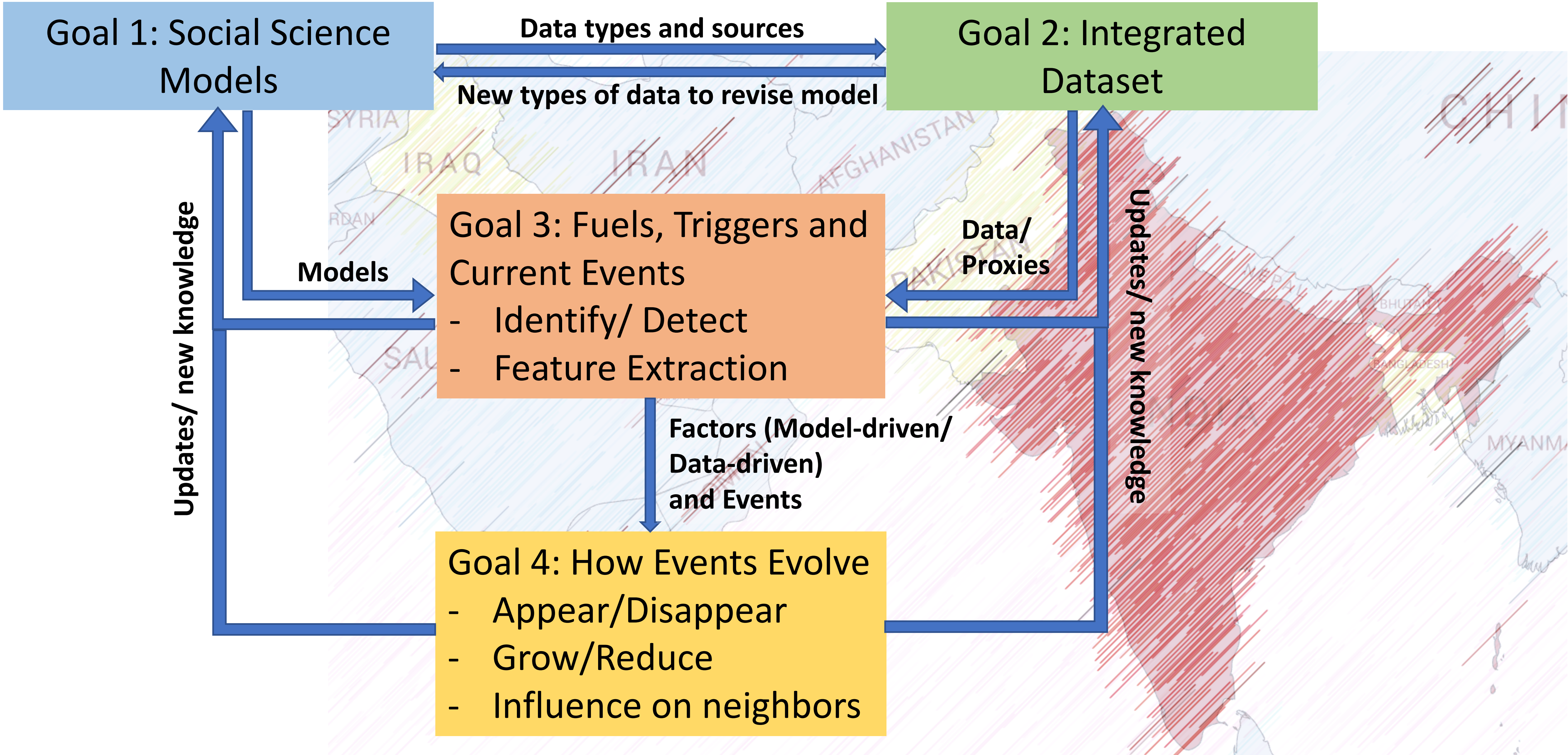
Building a Robust Model to Anticipate Unrest Using Model-driven and Data-driven Strategies

Dr. Deepti Joshi
Associate Professor of Computer Science
The Citadel, The Military College of South Carolina
djoshi@citadel.edu

Over-arching/ Long-term goal

- Develop an integrated theory-based and data-driven framework to *understand and detect the fuels and triggers for social unrest and ultimately anticipate the onset and spread of unrest in a broad range of countries*
- Increase **situational awareness** in the places of interest around the world
- Current focus is on **India**
- Next steps include expanding to **Iraq, Bangladesh and Pakistan**.

Anticipate Unrest Events / Improved Situational Awareness



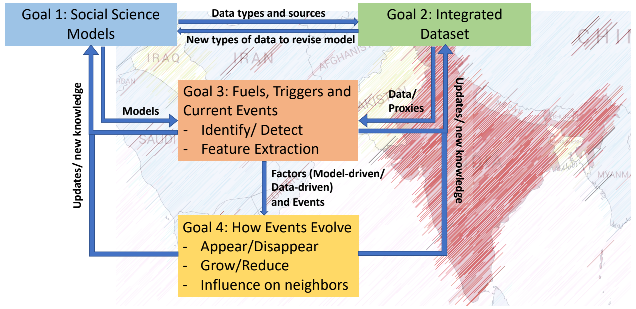
Goal 1: Social Science Models

Goal 2: Integrated Dataset

Goal 3: Fuels, Triggers and Current Events
- Identify/ Detect
- Feature Extraction

Goal 4: How Events Evolve
- Appear/Disappear
- Grow/Reduce
- Influence on neighbors

Anticipate Unrest Events / Improved Situational Awareness



Goal 1: Social Science Models

Data types and sources

New types of data to revise model

Goal 2: Integrated Dataset

Updates/ new knowledge

Models

Goal 3: Fuels, Triggers and Current Events
- Identify/ Detect
- Feature Extraction

Data/ Proxies

Updates/ new knowledge

Factors (Model-driven/ Data-driven) and Events

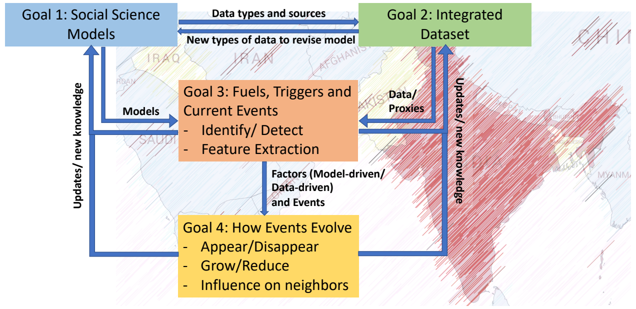
Goal 4: How Events Evolve
- Appear/Disappear
- Grow/Reduce
- Influence on neighbors

Goal 1: Social Science Models

- To leverage social science theories of unrest to identify factors that influence unrest in the long term
  - Current focus on grievance-based approaches

Dependent Variable/Outcome	Independent Variables/ Theoretical Concept	Independent Variables : Operationalization/Measurement	Social Science Data Sources	Proxy in Social Science Models
Conflict	Inequality (broadly defined)	Regionally Concentrated Ethnic Groups; Income Discrepancies (regional GDP); Fiscal Decentralization; Spending on federal grants and shared revenues; Ethnic political representation	Census of India	Religious Fractionalization, Percent Schedule Caste, Percent Schedule Tribe

Anticipate Unrest Events / Improved Situational Awareness



Goal 1: Social Science Models

Data types and sources

New types of data to revise model

Goal 2: Integrated Dataset

Updates/ new knowledge

Models

Goal 3: Fuels, Triggers and Current Events
- Identify/ Detect
- Feature Extraction

Data/ Proxies

Updates/ new knowledge

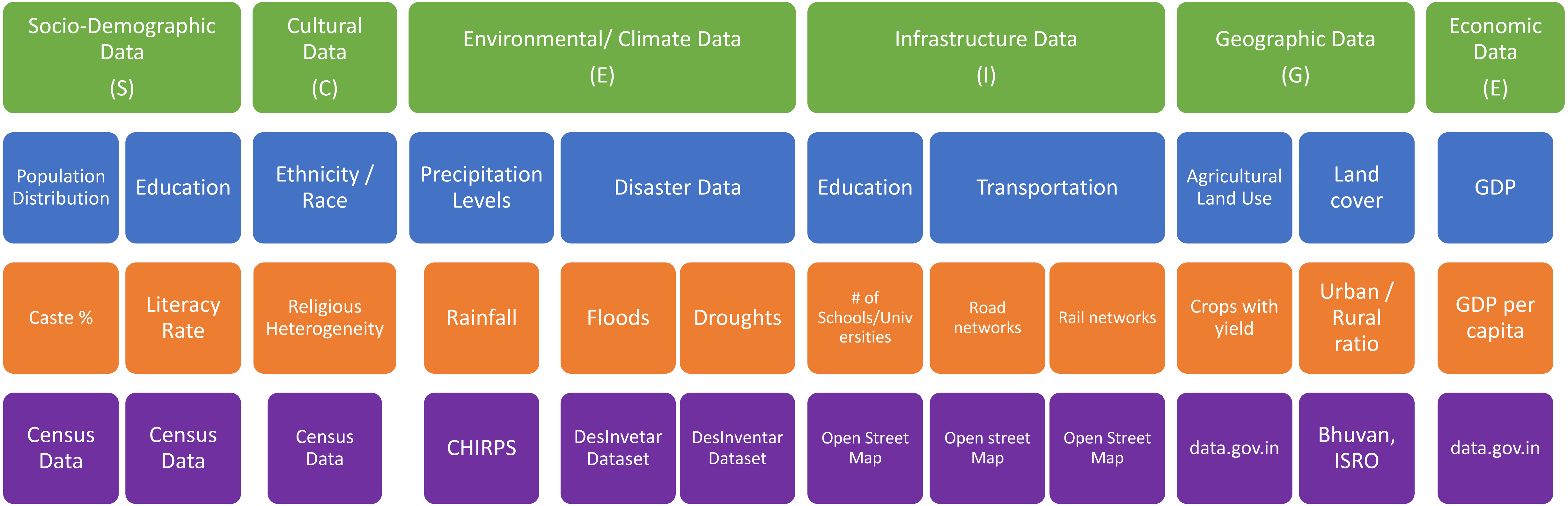
Factors (Model-driven/ Data-driven) and Events

Goal 4: How Events Evolve
- Appear/Disappear
- Grow/Reduce
- Influence on neighbors

Goal 2: Integrated Dataset

- To develop integrated datasets for SCEIGE factors, unrest event data, and situational awareness with raw data
  - **SCEIGE factors**: To scrape, collect, download, and clean data related to socio-demographic (S), cultural (C), environmental (E), infrastructure (I), geographic (G), and economic (E) factors, which are further analyzed as potential sources of fuels and triggers for unrest
  - **Unrest event datasets**: To extract, download, and store reports/counts of unrest events within our region of interest from the existing databases of GDELT, ACLED and ICEWS
    - Additionally, study the advantages and disadvantages of these datasets
  - **Raw data for situational awareness**: To scrape and download original news articles from regional and national sources within our region of interest
    - In addition, collect images that represent unrest within with our region of interest

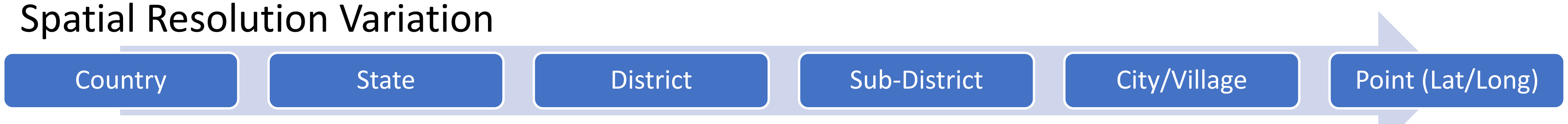
Historical SCEIGE Datasets for Region of Interest



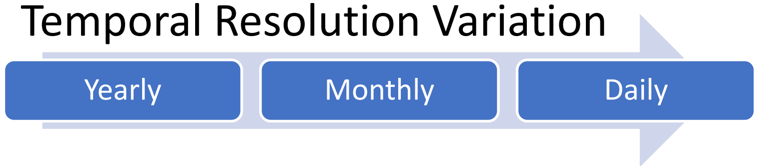
Socio-Demographic Data (S)		Cultural Data (C)	Environmental/ Climate Data (E)			Infrastructure Data (I)			Geographic Data (G)		Economic Data (E)
Population Distribution	Education	Ethnicity / Race	Precipitation Levels	Disaster Data		Education	Transportation		Agricultural Land Use	Land cover	GDP
Caste %	Literacy Rate	Religious Heterogeneity	Rainfall	Floods	Droughts	# of Schools/Universities	Road networks	Rail networks	Crops with yield	Urban / Rural ratio	GDP per capita
Census Data	Census Data	Census Data	CHIRPS	DesInvetar Dataset	DesInventar Dataset	Open Street Map	Open street Map	Open Street Map	data.gov.in	Bhuvan, ISRO	data.gov.in

Challenges:

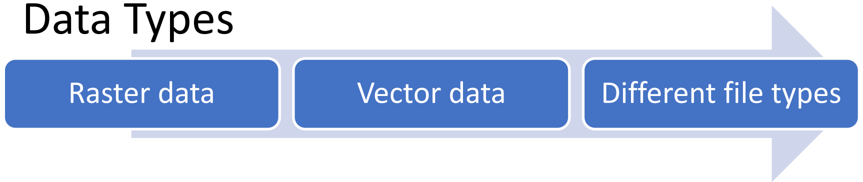
Spatial Resolution Variation



Temporal Resolution Variation



Data Types



Existing Unrest event Databases

GDELT

ACLED

ICEWS

ICEWS database attributes

Event ID	Event Date	Source Name	Source Sectors
Source Country	Event Text	CAMEO Code	Intensity
Target Name	Target Sectors	Target Country	Story ID
Sentence Number	Publisher	City	District
Province	Country	Latitude	Longitude

https://dataverse.harvard.edu/dataverse/icews

ACLED database attributes

data_id	iso	event_id_cnty	event_id_no_cnty
event_date	year	time_precision	event_type
sub_event_type	actor1	assoc_actor_1	inter1
actor2	assoc_actor_2	inter2	interaction
region	country	admin1	admin2
admin3	location	latitude	longitude
geo_precision	source	source_scale	notes
fatalities	timestamp	iso3	

https://acleddata.com/data-export-tool/

GDELT database attributes

GLOBALEVENTID	SQLDATE	MonthYear	Year
FractionDate	Actor1Code	Actor1Name	Actor1CountryCode
Actor1KnownGroupCode	Actor1EthnicCode	Actor1Religion1Code	Actor1Religion2Code
Actor1Type1Code	Actor1Type2Code	Actor1Type3Code	Actor2Code
Actor2Name	Actor2CountryCode	Actor2KnownGroupCode	Actor2EthnicCode
Actor2Religion1Code	Actor2Religion2Code	Actor2Type1Code	Actor2Type2Code
Actor2Type3Code	IsRootEvent	EventCode	EventBaseCode
EventRootCode	QuadClass	GoldsteinScale	NumMentions
NumSources	NumArticles	AvgTone	Actor1Geo_Type
Actor1Geo_FullName	Actor1Geo_CountryCode	Actor1Geo_ADM1Code	Actor1Geo_Lat
Actor1Geo_Long	Actor1Geo_FeatureID	Actor2Geo_Type	Actor2Geo_FullName
Actor2Geo_CountryCode	Actor2Geo_ADM1Code	Actor2Geo_Lat	Actor2Geo_Long
Actor2Geo_FeatureID	ActionGeo_Type	ActionGeo_FullName	ActionGeo_CountryCode
ActionGeo_ADM1Code	ActionGeo_Lat	ActionGeo_Long	ActionGeo_FeatureID
DATEADDED	SOURCEURL		

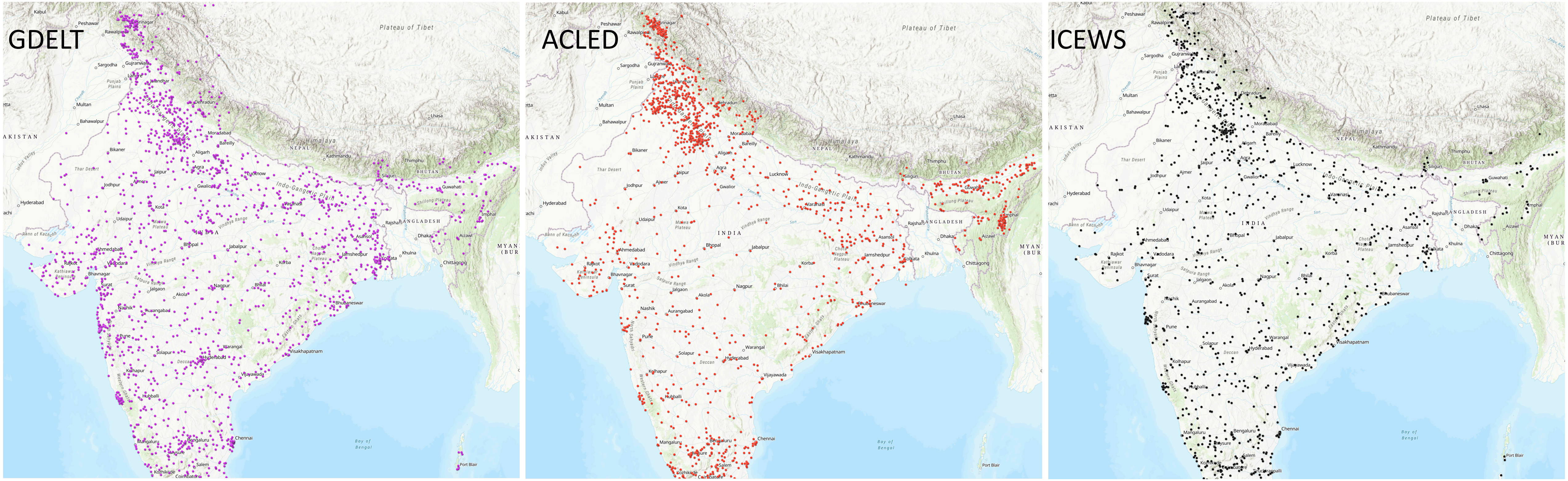
http://data.gdeltproject.org/events/index.html

Unrest Event Databases: A Comparison

	GDELT	ACLED	ICEWS
# of reported protest events in 2016	59,422	9,692	8,491
# of unique locations for protest events in 2016	1,978	1,578	1,081
# of sources in 2016 for India (protest events)	1,908	136	112

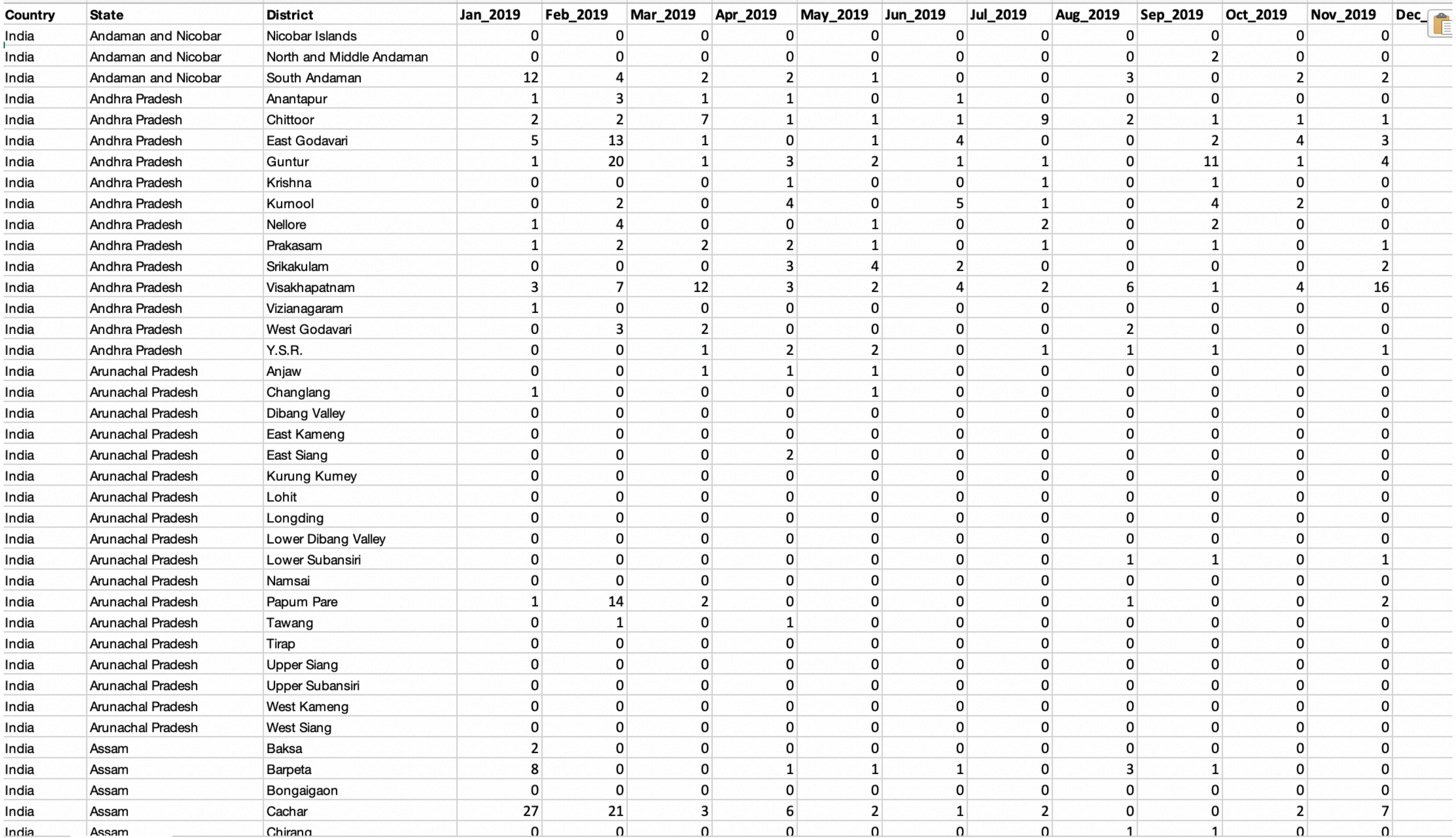
Unrest Event Databases: Challenges

- Duplication
- Extent of Bias
- Accuracy
- Coverage
- Multiple Sources
- Types of sources
- Event coding
- National/ Regional



Protest Event Locations in the Year of 2016 in India as reported in the three databases

De-duplicating GDELT and computing event frequency for each district per month in India



Country	State	District	Jan_2019	Feb_2019	Mar_2019	Apr_2019	May_2019	Jun_2019	Jul_2019	Aug_2019	Sep_2019	Oct_2019	Nov_2019	Dec_
India	Andaman and Nicobar	Nicobar Islands	0	0	0	0	0	0	0	0	0	0	0	
India	Andaman and Nicobar	North and Middle Andaman	0	0	0	0	0	0	0	0	2	0	0	
India	Andaman and Nicobar	South Andaman	12	4	2	2	1	0	0	3	0	2	2	
India	Andhra Pradesh	Anantapur	1	3	1	1	0	1	0	0	0	0	0	
India	Andhra Pradesh	Chittoor	2	2	7	1	1	1	9	2	1	1	1	
India	Andhra Pradesh	East Godavari	5	13	1	0	1	4	0	0	2	4	3	
India	Andhra Pradesh	Guntur	1	20	1	3	2	1	1	0	11	1	4	
India	Andhra Pradesh	Krishna	0	0	0	1	0	0	1	0	1	0	0	
India	Andhra Pradesh	Kurnool	0	2	0	4	0	5	1	0	4	2	0	
India	Andhra Pradesh	Nellore	1	4	0	0	1	0	2	0	2	0	0	
India	Andhra Pradesh	Prakasam	1	2	2	2	1	0	1	0	1	0	1	
India	Andhra Pradesh	Srikakulam	0	0	0	3	4	2	0	0	0	0	2	
India	Andhra Pradesh	Visakhapatnam	3	7	12	3	2	4	2	6	1	4	16	
India	Andhra Pradesh	Vizianagaram	1	0	0	0	0	0	0	0	0	0	0	
India	Andhra Pradesh	West Godavari	0	3	2	0	0	0	0	2	0	0	0	
India	Andhra Pradesh	Y.S.R.	0	0	1	2	2	0	1	1	1	0	1	
India	Arunachal Pradesh	Anjaw	0	0	1	1	1	0	0	0	0	0	0	
India	Arunachal Pradesh	Changlang	1	0	0	0	1	0	0	0	0	0	0	
India	Arunachal Pradesh	Dibang Valley	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	East Kameng	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	East Siang	0	0	0	2	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Kurung Kumey	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Lohit	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Longding	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Lower Dibang Valley	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Lower Subansiri	0	0	0	0	0	0	0	1	1	0	1	
India	Arunachal Pradesh	Namsai	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Papum Pare	1	14	2	0	0	0	0	1	0	0	2	
India	Arunachal Pradesh	Tawang	0	1	0	1	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Tirap	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Upper Siang	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	Upper Subansiri	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	West Kameng	0	0	0	0	0	0	0	0	0	0	0	
India	Arunachal Pradesh	West Siang	0	0	0	0	0	0	0	0	0	0	0	
India	Assam	Baksa	2	0	0	0	0	0	0	0	0	0	0	
India	Assam	Barpeta	8	0	0	1	1	1	0	3	1	0	0	
India	Assam	Bongaigaon	0	0	0	0	0	0	0	0	0	0	0	
India	Assam	Cachar	27	21	3	6	2	1	2	0	0	2	7	
India	Assam	Chirang	0	0	0	0	0	0	0	1	1	0	0	

De-duplication strategy:

If two or more rows have the same date, country, latitude, longitude, event code, avg. tone, Goldstein scale and SourceURL, then drop duplicates, and keep only the first row

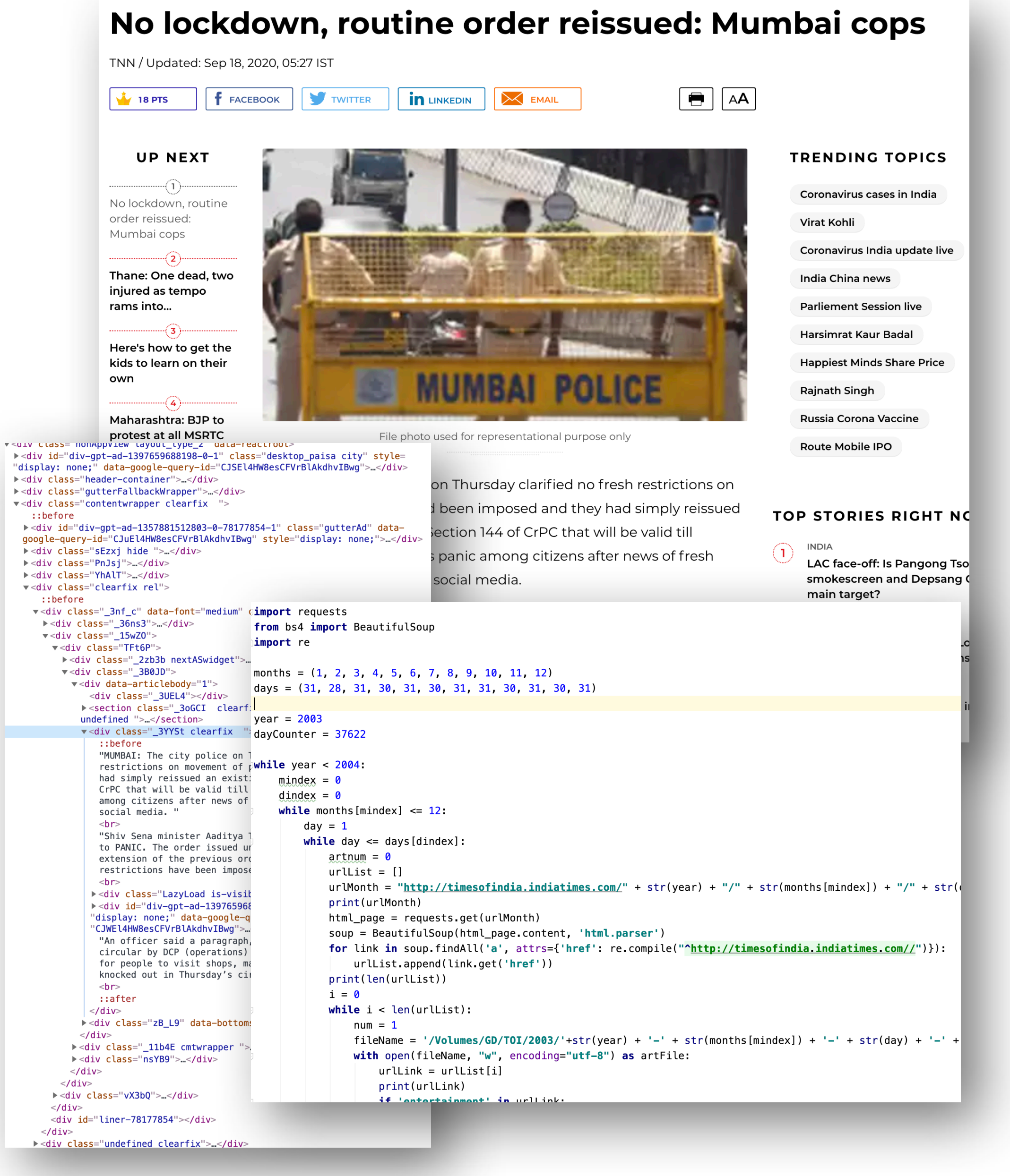
Raw Data (Collected)

News articles

National sources | Regional sources

Indian Newspaper	Category	Years Available	Total# (01/01/2015 – 12/31/2019)
The Times of India	National	01/2001 - Present	1,030,810
The Hindu	National	08/2009 - Present	878,817
The Pioneer	National	01/2011 - Present	278,842
Economic Times	National	01/2001 - Present	680,835
Incredible Orissa	Regional (East India)	02/2014 - Present	85,415
The Assam Tribune	Regional (North East India)	02/2010 – Present	20,680
Kashmir Observer	Regional (North India)	05/2012 – Present	2,857

Process of scraping news articles



No lockdown, routine order reissued: Mumbai cops

TNN / Updated: Sep 18, 2020, 05:27 IST

UP NEXT

1 No lockdown, routine order reissued: Mumbai cops

2 Thane: One dead, two injured as tempo rams into...

3 Here's how to get the kids to learn on their own

4 Maharashtra: BJP to protest at all MSRTC

File photo used for representational purpose only

on Thursday clarified no fresh restrictions on
d been imposed and they had simply reissued
Section 144 of CrPC that will be valid till
s panic among citizens after news of fresh
social media.

TRENDING TOPICS

- Coronavirus cases in India
- Virat Kohli
- Coronavirus India update live
- India China news
- Parliement Session live
- Harsimrat Kaur Badal
- Happiest Minds Share Price
- Rajnath Singh
- Russia Corona Vaccine
- Route Mobile IPO

TOP STORIES RIGHT NO

1 INDIA
LAC face-off: Is Pangong Tso smokescreen and Depsang main target?

```
▼<div class="nonAppView layout_type_2" data-reactroot>
  ▶<div id="div-gpt-ad-1397659688198-0-1" class="desktop_paisa city" style=
  "display: none;" data-google-query-id="CJSEl4HW8esCFVrBlAkdhvIBwg">…</div>
  ▶<div class="header-container">…</div>
  ▶<div class="gutterFallbackWrapper">…</div>
  ▼<div class="contentwrapper clearfix  ">
     ::before
    ▶<div id="div-gpt-ad-1357881512803-0-78177854-1" class="gutterAd" data-
    google-query-id="CJuEl4HW8esCFVrBlAkdhvIBwg" style="display: none;">…</div>
    ▶<div class="sEzxj hide ">…</div>
    ▶<div class="PnJsj">…</div>
    ▶<div class="YhAlT">…</div>
    ▼<div class="clearfix rel">
       ::before
      ▼<div class="_3nf_c" data-font="medium" c
        ▶<div class="_36ns3">…</div>
        ▼<div class="_15wZO">
          ▼<div class="TFt6P">
            ▶<div class="_2zb3b nextASwidget">…
            ▼<div class="_3B0JD">
              ▼<div data-articlebody="1">
                 <div class="_3UEL4"></div>
                ▶<section class="_3oGCI  clearf
                undefined ">…</section>
                ▼<div class="_3YYSt clearfix  ">
                   ::before
                   "MUMBAI: The city police on T
                   restrictions on movement of p
                   had simply reissued an exist
                   CrPC that will be valid till
                   among citizens after news of
                   social media. "
                   <br>
                   "Shiv Sena minister Aaditya T
                   to PANIC. The order issued un
                   extension of the previous ord
                   restrictions have been impose
                   <br>
                  ▶<div class="LazyLoad is-visib
                  ▶<div id="div-gpt-ad-13976596
                  "display: none;" data-google-q
                  "CJWEl4HW8esCFVrBlAkdhvIBwg">…
                   "An officer said a paragraph,
                   circular by DCP (operations)
                   for people to visit shops, ma
                   knocked out in Thursday's cir
                   <br>
                   ::after
                 </div>
                ▶<div class="zB_L9" data-bottoms
               </div>
              ▶<div class="_11b4E cmtwrapper ">
              ▶<div class="nsYB9">…</div>
             </div>
           </div>
          ▶<div class="vX3bQ">…</div>
         </div>
         <div id="liner-78177854"></div>
       </div>
      ▶<div class="undefined clearfix">…</div>
```

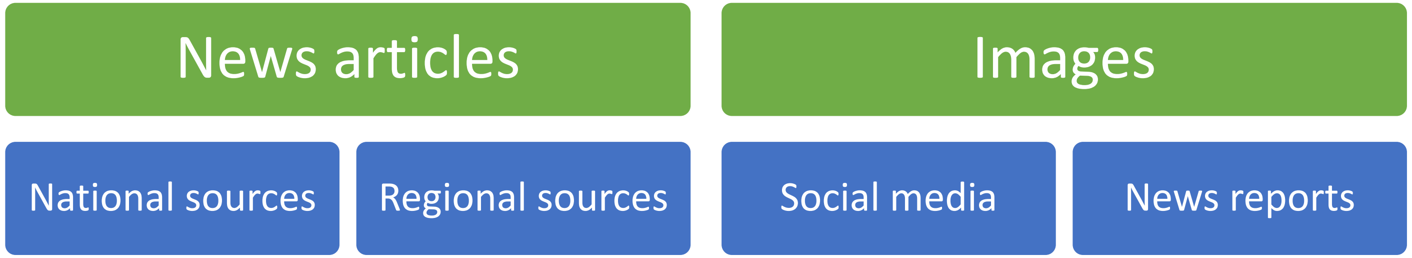
```python
import requests
from bs4 import BeautifulSoup
import re

months = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
days = (31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)

year = 2003
dayCounter = 37622

while year < 2004:
    mindex = 0
    dindex = 0
    while months[mindex] <= 12:
        day = 1
        while day <= days[dindex]:
            artnum = 0
            urlList = []
            urlMonth = "http://timesofindia.indiatimes.com/" + str(year) + "/" + str(months[mindex]) + "/" + str(
            print(urlMonth)
            html_page = requests.get(urlMonth)
            soup = BeautifulSoup(html_page.content, 'html.parser')
            for link in soup.findAll('a', attrs={'href': re.compile("^http://timesofindia.indiatimes.com//")}):
                urlList.append(link.get('href'))
            print(len(urlList))
            i = 0
            while i < len(urlList):
                num = 1
                fileName = '/Volumes/GD/TOI/2003/'+str(year) + '-' + str(months[mindex]) + '-' + str(day) + '-' +
                with open(fileName, "w", encoding="utf-8") as artFile:
                    urlLink = urlList[i]
                    print(urlLink)
                    if 'entertainment' in urlLink:
```

Raw Data (Planned)



News articles

Images

National sources

Regional sources

Social media

News reports

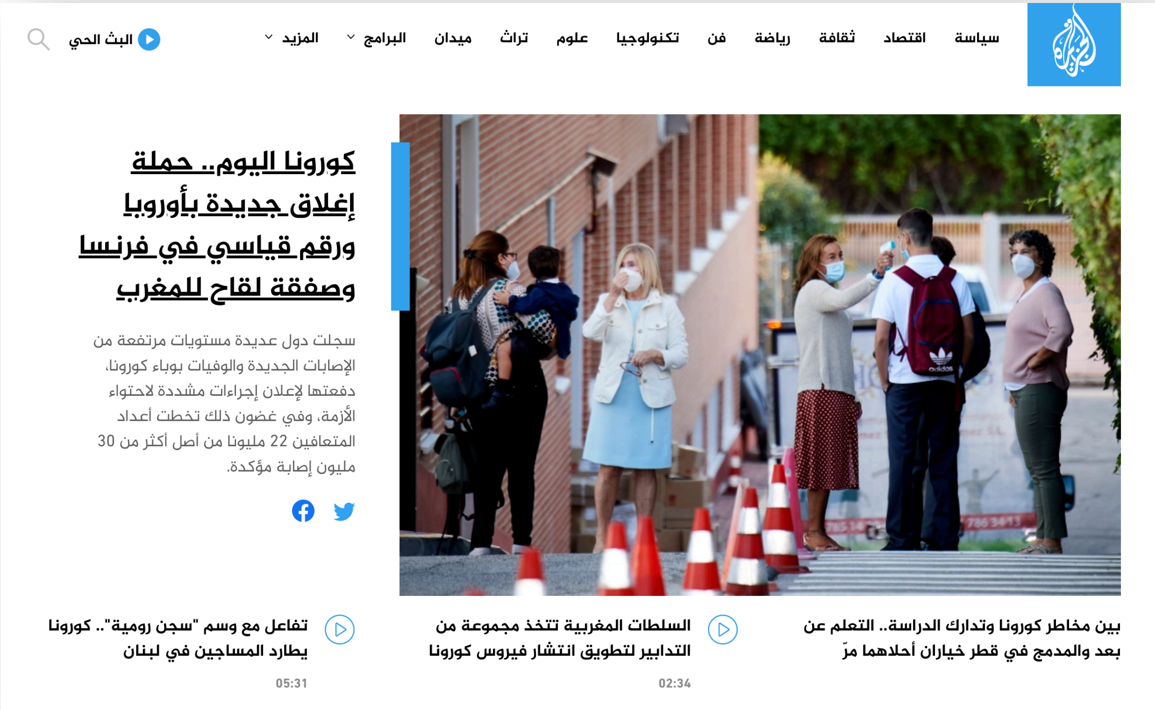
Newspaper	Country, Category	Years Available
Deccan Herald	India, Regional (South India)	06/2009 - Present
Dawn	Pakistan	01/2011 - Present
The Daily Star	Bangladesh	01/2011 - Present
Al Jazeera	Iraq and Arabic Peninsula	-

Image source	Sample Topics/ Category
News reports	All/ Article Topic
Flickr	Protest – Violent and Peaceful
	Fire
	Crowds

Violent Protest

Peaceful Protest

Multilingual News Articles



كورونا اليوم.. حملة إغلاق جديدة بأوروبا ورقم قياسي في فرنسا وصفقة لقاح للمغرب

سجلت دول عديدة مستويات مرتفعة من الإصابات الجديدة والوفيات بوباء كورونا، دفعتها لإعلان إجراءات مشددة لاحتواء الأزمة، وفي غضون ذلك تخطت أعداد المتعافين 22 مليونا من أصل أكثر من 30 مليون إصابة مؤكدة.

بين مخاطر كورونا وتدارك الدراسة.. التعلم عن بعد والمدمج في قطر خياران أحلاهما مرّ

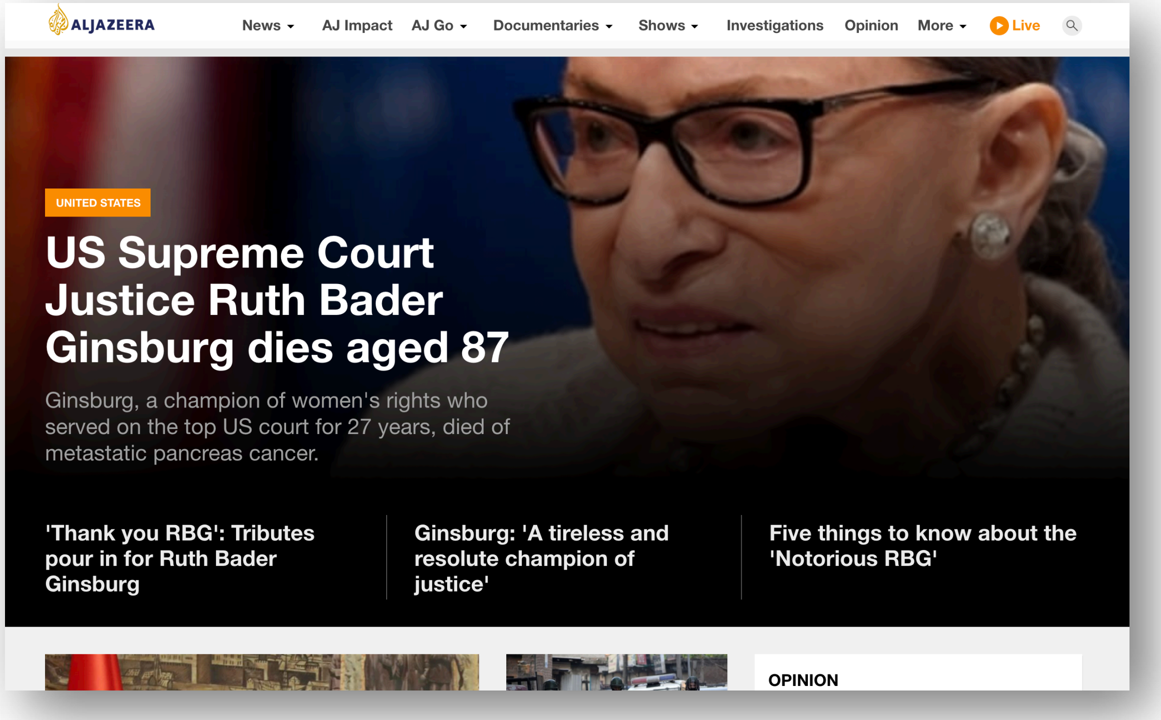
السلطات المغربية تتخذ مجموعة من التدابير لتطويق انتشار فيروس كورونا

02:34

تفاعل مع وسم "سجن رومية".. كورونا يطارد المساجين في لبنان

05:31

https://www.aljazeera.net/ on September 18, 2020 at 11:10 pm



UNITED STATES

US Supreme Court Justice Ruth Bader Ginsburg dies aged 87

Ginsburg, a champion of women's rights who served on the top US court for 27 years, died of metastatic pancreas cancer.

'Thank you RBG': Tributes pour in for Ruth Bader Ginsburg

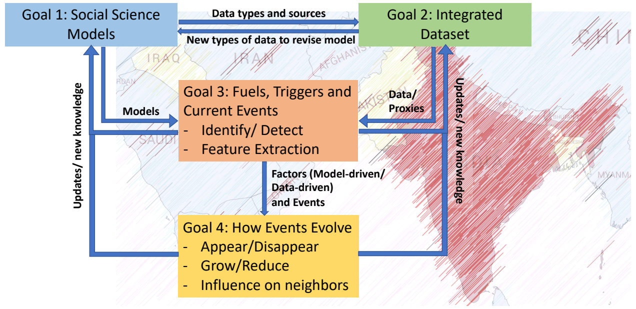
Ginsburg: 'A tireless and resolute champion of justice'

Five things to know about the 'Notorious RBG'

OPINION

https://www.aljazeera.com/ on September 18, 2020 at 11:10pm

Anticipate Unrest Events / Improved Situational Awareness

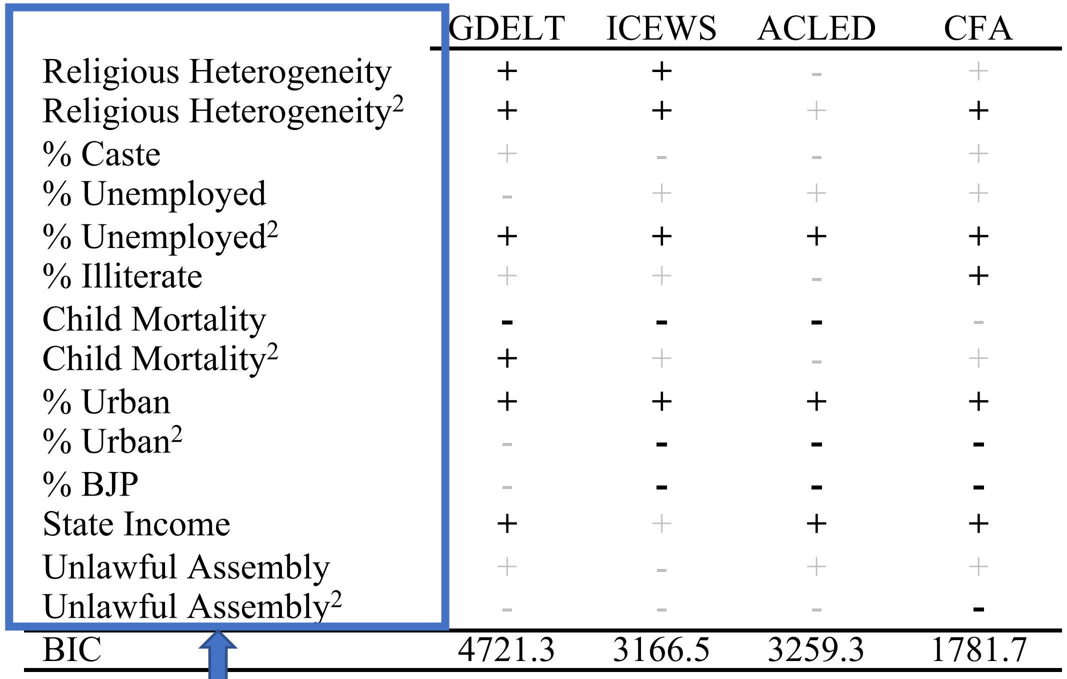


Goal 3: Fuels, Triggers and Current Events

- To identify fuels and triggers for the emergence of social unrest to understand the lifecycle of such dynamic event in complex environments
  - **Fuel Process Identification**: To identify SCEIGE factors fueling or mitigating spread of social unrest
  - **Trigger/Inhibit Process Identification:** To identify government actions that lead to positive and negative sentiments and model their relationships with unrest events
  - **Case Study**: To identify roles of disasters on social unrest
  - **Multi-Agent Modeling**: To investigate the roles of fuels and triggers

Multivariate Analyses of District-Level Protests in India: Comparing Results Across Datasets

- Cross-sectional analyses of 2016 protests at district level (n=524)
  - Zero-Inflated Negative Binomial Regression
  - Using 2011 data as predictors (Census)
- Main findings
  - Directions of results indicate similar trends across all DVs (significance levels vary)
    - Indicates data sources are capturing the underlying patterns in protests
  - Results from CFA seems to smooth out results from three data sources
- Major Takeaways
  - Combining sources mitigates validity concerns
  - Potential solution for robust cross-sectional and longitudinal analyses beyond India



	GDELT	ICEWS	ACLED	CFA
Religious Heterogeneity	+	+	-	+
Religious Heterogeneity^2	+	+	+	+
% Caste	+	-	-	+
% Unemployed	-	+	+	+
% Unemployed^2	+	+	+	+
% Illiterate	+	+	-	+
Child Mortality	-	-	-	-
Child Mortality^2	+	+	-	+
% Urban	+	+	+	+
% Urban^2	-	-	-	-
% BJP	-	-	-	-
State Income	+	+	+	+
Unlawful Assembly	+	-	+	+
Unlawful Assembly^2	-	-	-	-
BIC	4721.3	3166.5	3259.3	1781.7

Long term (fuel) indicators for unrest

Incorporating Trigger/Inhibiter Models

- Government actions can trigger or inhibit social unrest activities
  - Positive actions
  - Negative actions
- Basic Idea
  - Simulating all three models (social unrest spread model, positive action spread model, negative action spread model)
  - Integrating simulated results
    - E.g., using social unrest spread model as anchor, and using the other models to establish confidence bounds
    - E.g., using positive/negative action spread model to modify ”transmission rates” in the social unrest spread model

Trigger/Inhibiter Models

What are Positive/Negative Actions?

- Using GDELT’s existing CAMEO Event Code and Goldstein scale
- Using VADER’s sentiment analysis
- Basic Idea:
    - A CAMEO-coded event (e.g., ID 56: Apologize) (with corresponding Goldstein scale) may trigger unrest or inhibit unrest
    - A CAMEO-coded event with corresponding Goldstein scale with matching VADER sentiment score may be considered as positive or negative *more* confidently
    - Use GDELT to find the trigger/inhibitor events anchored around the ACLED reported protest events

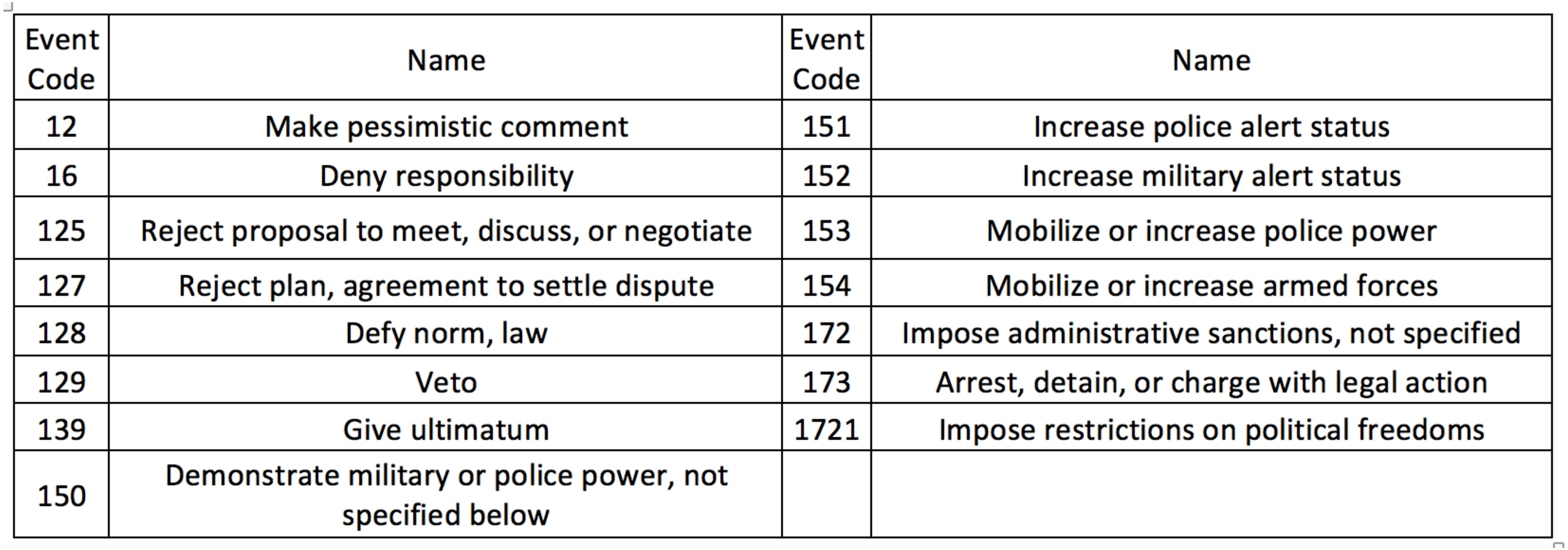
Trigger/ Inhibiter Models

CAMEO Positive Event Codes

Event Code	Name	Event Code	Name
13	Make optimistic comment	53	Rally support or behalf of
14	Consider policy option	54	Grant diplomatic recognition
15	Acknowledge or claim responsibility	55	Apologize
18	Make emphatic comment	56	Forgive
30	Express intent to cooperate, not specified below	57	Sign formal agreement
31	Express intent to engage in material cooperation	81	Ease administrative sanctions, not specified below
32	Express intent to provide diplomatic cooperation such as policy support	82	Ease political dissent
33	Express intent to provide material aid	83	Accede to requests or demands for political reform not specified below
35	Express intent to yield, not specified	84	Return, release, not specified below
36	Express intent to meet or negotiate	311	Express intent to cooperate economically
37	Express intent to settle dispute	312	
38	Express intent to accept mediation	331	Express intent to provide economic aid
40	Consult, not specified below	332	Express intent to provide military aid
41	Discuss by telephone	333	Express intent to provide humanitarian aid
42	Meet a visit	334	Express intent to provide military protection or peacekeeping
43	Host a visit	353	Express intent to release persons or property
44	Meet at third location	356	Express intent to de-escalate military engagement
45	Mediate	811	Ease restrictions on political freedoms
46	Engage in negotiation	831	Accede to demands for change in policy
50	Engage in diplomatic cooperation, not specified	833	Accede to demands for rights
51	Praise or endorse	841	Return, release, persons
52	Defend Verbally		

Trigger/Inhibiter Models

CAMEO Negative Event Codes



Event Code	Name	Event Code	Name
12	Make pessimistic comment	151	Increase police alert status
16	Deny responsibility	152	Increase military alert status
125	Reject proposal to meet, discuss, or negotiate	153	Mobilize or increase police power
127	Reject plan, agreement to settle dispute	154	Mobilize or increase armed forces
128	Defy norm, law	172	Impose administrative sanctions, not specified
129	Veto	173	Arrest, detain, or charge with legal action
139	Give ultimatum	1721	Impose restrictions on political freedoms
150	Demonstrate military or police power, not specified below		

Trigger/Inhibiter Models
Goldstein scale score

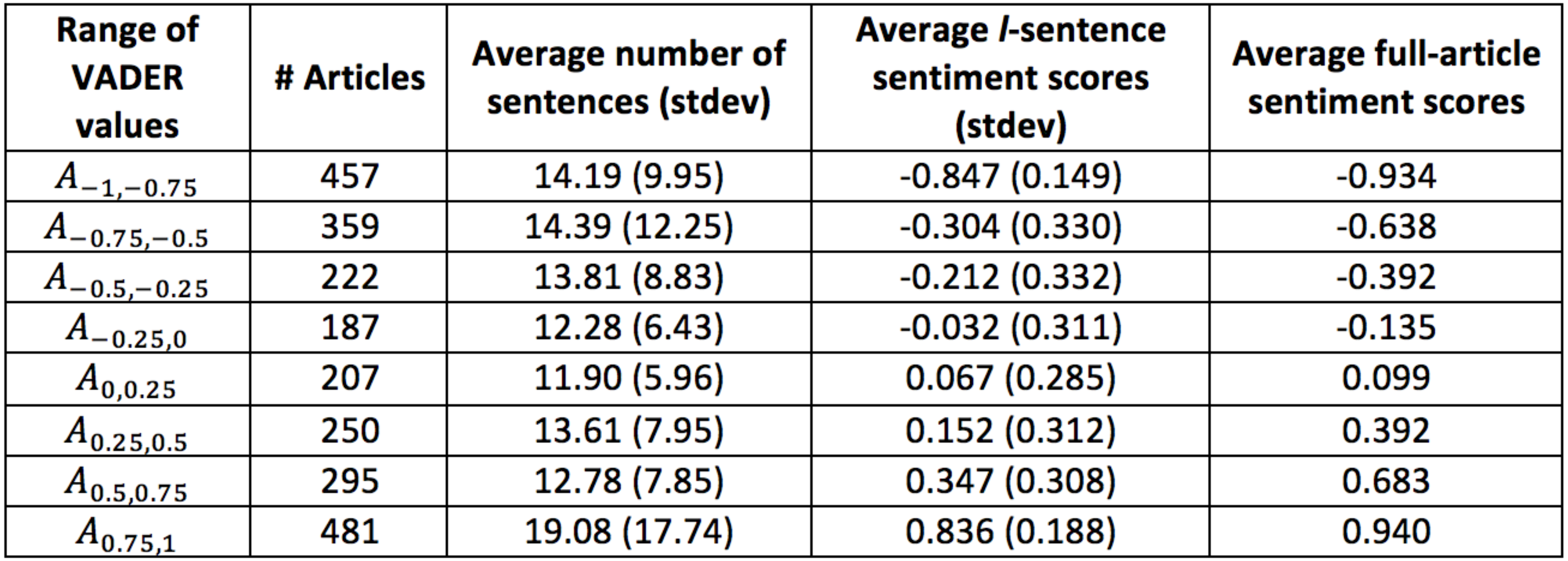
- GDELT provides Goldstein scale score (Goldstein, 1992) for each CAMEO code
  - A conflict-cooperation scale originally created for World Event/Interaction Survey (WEIS)
  - A numeric score from -10 to 10: the theoretical potential impact that type of event will have on the stability of the region
    - Positive Goldstein score implies that the event has positive impact and vice-versa

Trigger/Inhibiter Models
VADER + GDELT

- Use the Valance Aware Dictionary for sentiment Reasoning (VADER) (Gilbert & Hutto, 2014) model to perform the sentiment analysis on each article associated with each recorded GDELT event
  - numeric sentiment score -1 (negative) to 1 (positive)
- **Events with a positive Goldstein value should have a positive VADER sentiment value, and vice versa**
- **However**:
  - Sentiment score of an article varies depending on the number of sentences in the article that are used in the VADER analysis
  - Events with a positive Goldstein value does not always have a positive VADER sentiment, and vice versa

Trigger/Inhibiter Models

Consistency in VADER sentiment scores



Range of VADER values	# Articles	Average number of sentences (stdev)	Average *l*-sentence sentiment scores (stdev)	Average full-article sentiment scores
$A_{-1,-0.75}$	457	14.19 (9.95)	-0.847 (0.149)	-0.934
$A_{-0.75,-0.5}$	359	14.39 (12.25)	-0.304 (0.330)	-0.638
$A_{-0.5,-0.25}$	222	13.81 (8.83)	-0.212 (0.332)	-0.392
$A_{-0.25,0}$	187	12.28 (6.43)	-0.032 (0.311)	-0.135
$A_{0,0.25}$	207	11.90 (5.96)	0.067 (0.285)	0.099
$A_{0.25,0.5}$	250	13.61 (7.95)	0.152 (0.312)	0.392
$A_{0.5,0.75}$	295	12.78 (7.85)	0.347 (0.308)	0.683
$A_{0.75,1}$	481	19.08 (17.74)	0.836 (0.188)	0.940

- The two subsets (very negative, very positive) are most consistent and with average scores within range

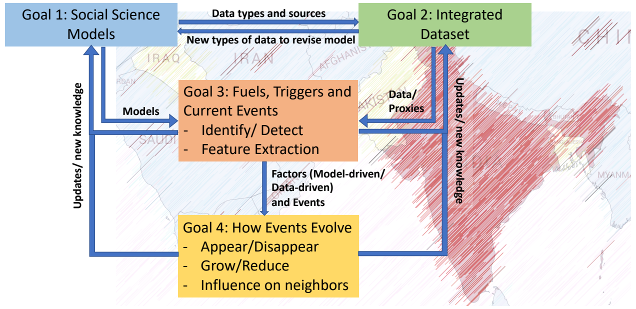
Trigger/Inhibiter Models

Filtering using CAMEO + VADER + Goldstein

- Step 1. Use positive and negative event codes (with the corresponding Goldstein scale) to select events
- Step 2. Compute VADER and retain events with a highly positive or highly negative VADER sentiment score

Goal 4: Event Evolution

Anticipate Unrest Events / Improved Situational Awareness



Goal 1: Social Science Models

Goal 2: Integrated Dataset

Goal 3: Fuels, Triggers and Current Events
- Identify/ Detect
- Feature Extraction

Goal 4: How Events Evolve
- Appear/Disappear
- Grow/Reduce
- Influence on neighbors

- To investigate how social unrest events evolve and interact in a dynamic environment to support analysis and anticipation
  - **Event Understanding**: To extract 5Ws from articles linked to events
  - **Similarity**: To determine distance between events
  - **Clustering**: To identify event clusters and an event's neighborhood
  - **Spatio-Temporal Interaction**: To investigate how interactions between regions impact spread of social unrest
  - **Multi-Agent Modeling**: To simulate event-to-event interactions and evaluate emergent behaviors

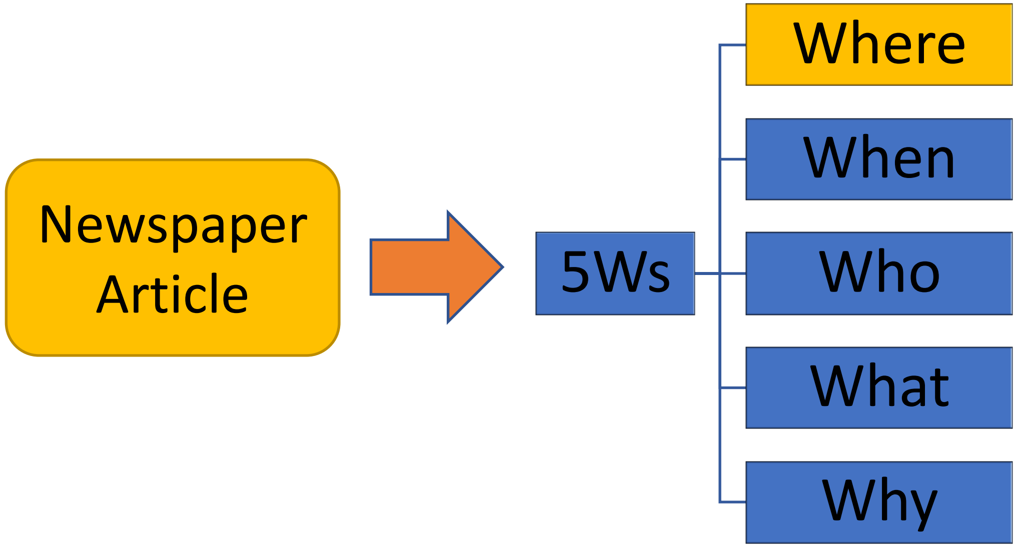
What is 5Ws?

- Journalistic 5Ws
  - Fundamental questions that every story should be able to answer
- “Who,” “What,” “When,” “Where,” and “Why”
- **5W1H**: 5Ws + “How”
- Critical to understanding of events, information gathering, and problem solving in general

Why 5W Analysis?

- Need to accurately understand unrest patterns
  - Are the unrest events random or clustered in space? (Where)
  - How does the unrest spread in space and time? (Where, When)
  - What are the principal drivers of the unrest? (Who, What, Why)
- Challenges with current unrest event databases
  - Duplication: Same event recorded multiple times in the database
  - Inconsistency: Place names, different spatial, and temporal resolutions
  - Absence of information: Unavailability of one or more of the Ws
  - Inaccuracy: Default locations used when the actual location is not available
- Informs the agent-based model for computation of similarity between events

5W Analysis



Here we focus on “Where”

Multi-Source Multi-Document “Where” Identification

```
MSMDW(D) {
  // $D = \{d_1, d_2, \cdots, d_n\}$
  $Locations = Groups = W = \emptyset$
  for each $d_i \in D$ do
     $w_i \leftarrow$ findLocations($d_i$)
     $W \leftarrow W \cup w_i$
  end for
  for each $d_i \in D$ do
     for each $d_j \in D$ do
        S ← similarity ($d_i, d_j$)
     end for
  end for
  $eventGroups \leftarrow$ computeGroups($D, S, W$)
  for each $g \in eventGroups$ do
     $L_g \leftarrow$ groupLocations($g$)
     $RL_g$ = Rank($L_g$)
     for each $d \in g$ do
           $d.placenames = RL_g$
     end for
  end for
}
```

Find all place names in *all* documents

Compute similarity between *all* documents and group them

Rank *all* place names in each group and associate with *all* the articles in the group

Key Problem

Find all place names in a document

Finding Locations in a Document

- Gazetteers provide a mapping from place names to geo-coordinates
- However, many place names are present not in gazetteers
    - E.g. villages India (55%)
- Do not provide a spatial hierarchy
    - Important to know if two events occurred in the same administrative unit
    - Nearby locations in adjacent states may have different unrest dynamics

Improving the Identification of Place Names

- Challenges with NER based place name recognition
  - Trained with a specific corpus
  - Retraining with a new set of place names is not trivial
  - Many place names are not recognized
    - E.g., SCB Medical College and Hospital in Cuttack will be the first … (Source: Incredible Orissa)
    - Auto Rickshaw Drivers Hold Protest in Srinagar. (Source: Kashmir Observer)
- Cues
  - Place names have commonalities in different regions defined by language or other cultural factors.
    - E. g., Place names with 'halli' as a suffix are very common in South India, especially Karnataka (Benniganahalli, Marathahalli, Hosahalli, etc.)
    - Place names with the suffix 'pur' are found abundantly in all regions of India except the state of Kerala. The suffix 'pur' in Malayalam language (the official language of Kerala) is 'puram'. So, place names in Kerala have the suffix 'puram' instead of 'pur'. (e.g., Thiruvanthapuram, Malappuram)
  - Prepositions
    - Place name prepositions
- Goal: Identify possible place names missed by standard NER using spatial cues and verify

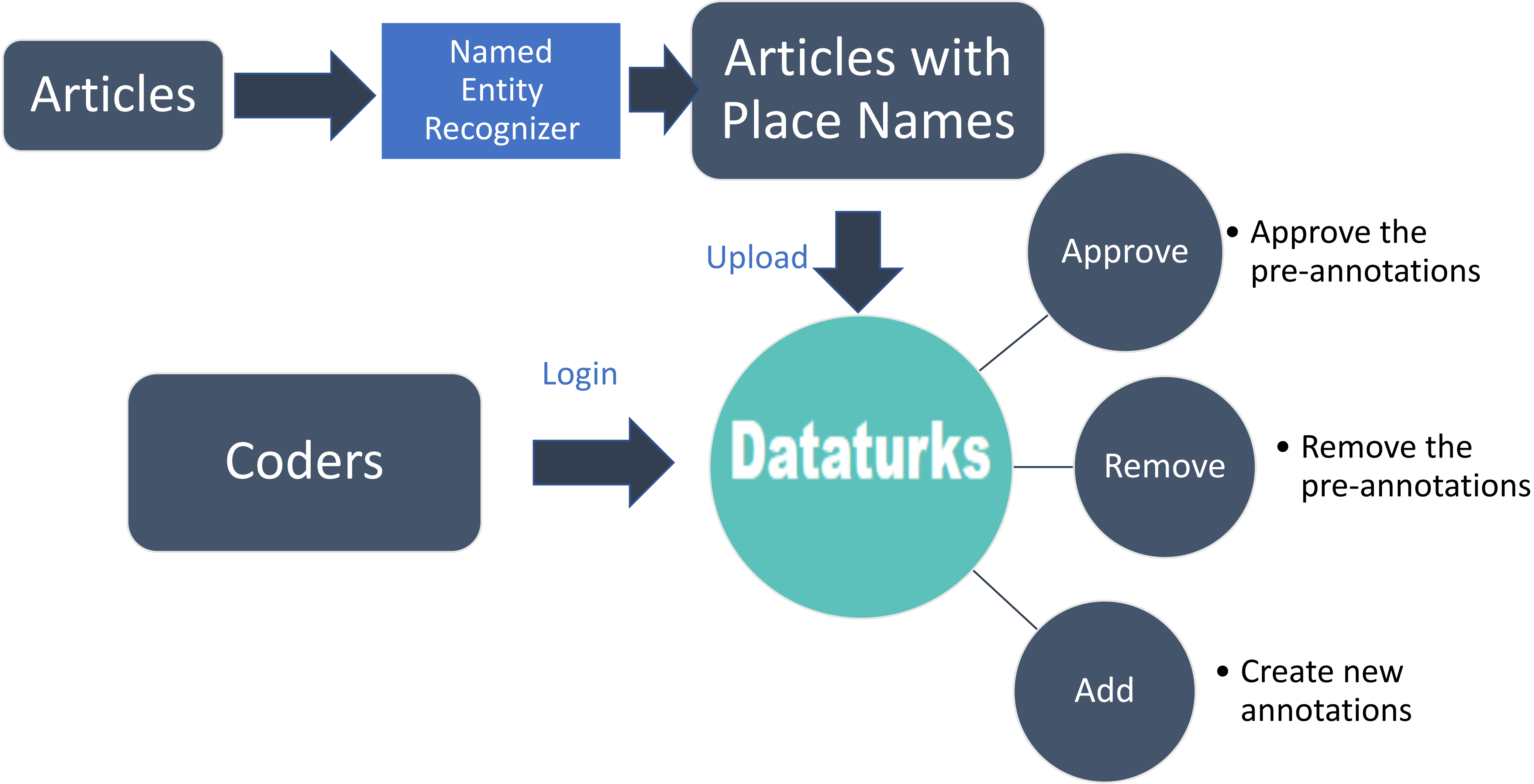
An Enhanced Hierarchical Gazetteer (for India)

- NGA Gazetteer (Number of place names = 659,513)
- Compile ALL state, district, cities, sub-districts, and villages from Indian Census (2010) (Number of place name hierarchies = 668,179)
  - There is a natural hierarchy in the data
  - However, geo-coordinates are NOT available
    - Geocoded from multiple sources (OpenStreetMap, ArcGIS, Google Maps)
  - Other open sources are used for update
- Add hierarchical information to the place names *only* in the NGA Gazetteer
  - Identify NGA place names that don’t appear in the census
  - 358,000 place names (Populated Places, Vegetation, Hypsographic, Hydrographic, etc.)
- **Enhanced Hierarchical Gazetteer** (Number of place names with hierarchies and geocodes = 1,026,179)

Validation of the “Where” detection algorithm

- There is a lack of a standardized dataset with verified ground truth
- Critical for demonstrating the efficacy of the algorithms and comparison of algorithms
- Solution: Develop our own standardized and validated dataset

Developing Ground Truth - 1
“Where” Candidate Annotation Annotation Process



Articles

Named Entity Recognizer

Articles with Place Names

Upload

Coders

Login

Dataturks

Approve

- Approve the pre-annotations

Remove

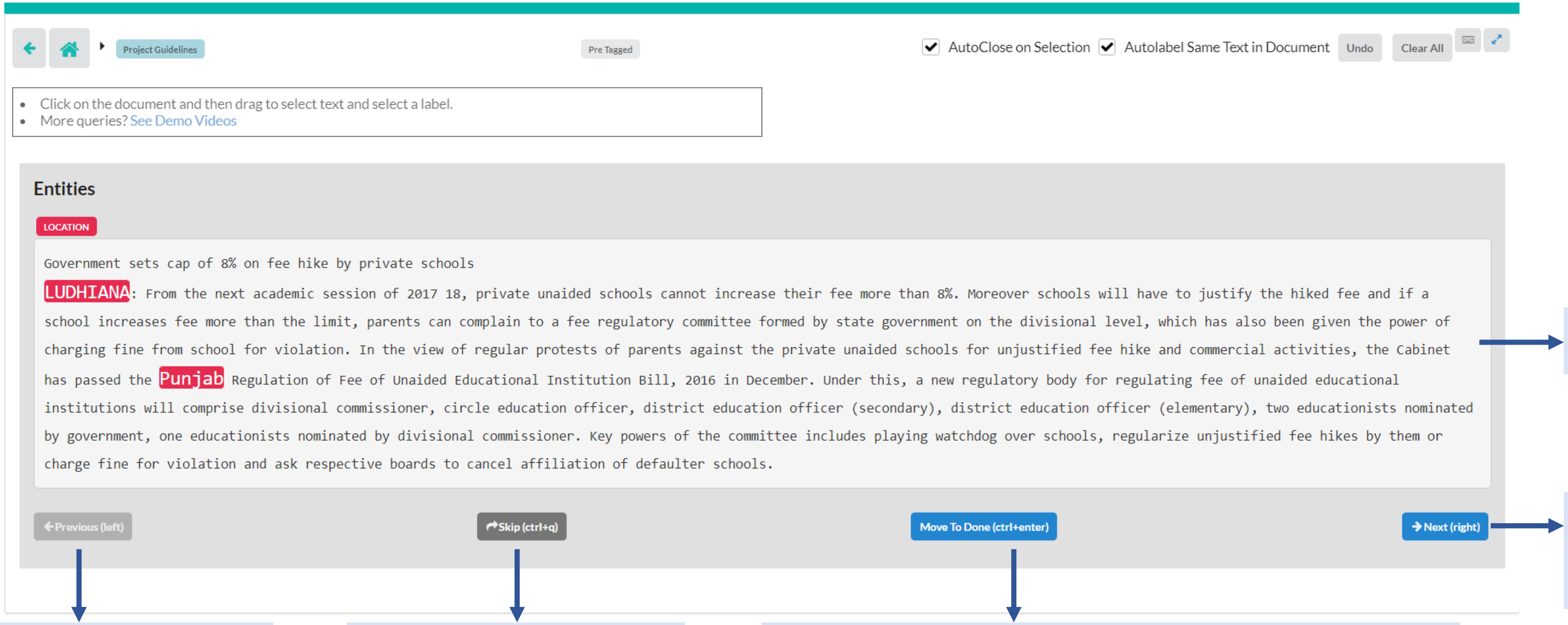
- Remove the pre-annotations

Add

- Create new annotations

Developing Ground Truth - 2
“Where” Candidate Annotation Platform

- **Dataturks**: Web-based platform for text annotation



☑ AutoClose on Selection ☑ Autolabel Same Text in Document

- Click on the document and then drag to select text and select a label.
- More queries? See Demo Videos

```
Government sets cap of 8% on fee hike by private schools
LUDHIANA: From the next academic session of 2017 18, private unaided schools cannot increase their fee more than 8%. Moreover schools will have to justify the hiked fee and if a
school increases fee more than the limit, parents can complain to a fee regulatory committee formed by state government on the divisional level, which has also been given the power of
charging fine from school for violation. In the view of regular protests of parents against the private unaided schools for unjustified fee hike and commercial activities, the Cabinet
has passed the Punjab Regulation of Fee of Unaided Educational Institution Bill, 2016 in December. Under this, a new regulatory body for regulating fee of unaided educational
institutions will comprise divisional commissioner, circle education officer, district education officer (secondary), district education officer (elementary), two educationists nominated
by government, one educationists nominated by divisional commissioner. Key powers of the committee includes playing watchdog over schools, regularize unjustified fee hikes by them or
charge fine for violation and ask respective boards to cancel affiliation of defaulter schools.
```

News article

Navigate to the next news article

Navigate to the previous news article

Skip the news article

Mark news article as done. Articles once marked as done cannot be edited.

Developing Ground Truth - 3
Data Annotation Process

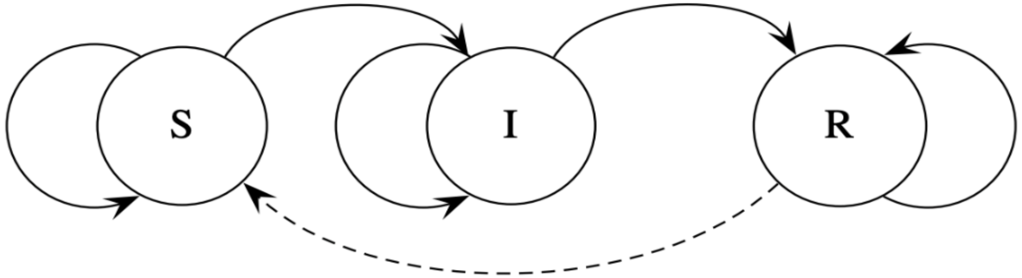
- **Goal -** Create a statistically reliable ground truth dataset
- **Process**
    - Use multiple coders for annotation (place names and “where”)
    - Train the coders with a set of documents and measure consistency
    - Repeat training with a new set of documents until an acceptable level of consistency is achieved
    - Annotate the selected news articles

Agent-Based Modeling of Social Unrest

- Modeling unrest behaviors using three models:
  - (1) spread model for how social unrest activities spread
  - (2) trigger model for how government actions trigger social unrest activities
  - (3) inhibit model for how government actions inhibit social unrest activities
- Integrating
  - Social science models for factors underlying social unrest (fuels)
  - Spatial interaction theories incorporating communication and transportation facilities
  - **Epidemiological model of disease spread**

Agent Design Based on SIR Model

- Each region (e.g., a district) is considered as an agent
- Each agent performs “conceptual” actions, of changing their state based on state transition probabilities



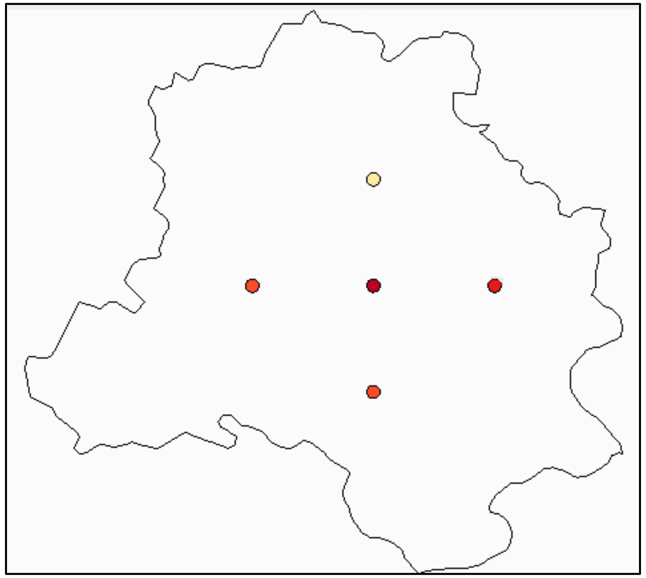


Figure showing possible state transitions among three different states: S = Susceptible, I = Infected, R = Recovered

Key Steps ABM based on SIR disease spread model

- Modeling each region (e.g., a district) as an agent, and an agent’s state (susceptible, infected, recovered) based on the presence of social unrest event, using the SIR model
- Defining a vector of SCEIGE factors to characterize each region, based on available data and guided by models and theories (Chenoweth & Ulfelder, 2017)
  - Pilot, will incorporate more accurate factors in the future
- Defining a distance metric for computing similarity between regions to establish each region’s neighborhood
- Preparing and incorporating ground-truth into simulation and evaluation

Chenoweth, E., & Ulfelder, J. (2017). Can structural conditions explain the onset of nonviolent uprisings? *Journal of Conflict Resolution*, *61*(2), 298–324.

Defining SCEIGE Vector ($\vec{v}_i^n$) (pilot)

- SCEIGE = Socio-Demographic, Cultural, Economic, Infrastructural, Geographical, Environmental

Category	Factor	Variable	Temporal Scale	Temporal Resolution	Extrapolation
Socio-Demographic	Minority group	Scheduled Caste	Yearly	Census 2001 & 2011	Linear regression
Cultural	Ethnic group	Scheduled Tribe	Yearly	Census 2001 & 2011	Linear regression
Economic	Economic	GDP growth rate	Yearly	1999 to 2007	Curve fitting using Fourier function
Infrastructural	Education	Literacy rate	Yearly	Census 2001 & 2011	Linear regression
Geographical	Land Cover	Urban land ratio	Yearly	2011 & 2015	Linear regression
Environmental	Climate	Standard Precipitation Index (SPI3)	Monthly	1989-2018	Curve fitting using Fourier function

Computing Neighborhood Distance Metric

- Distance function currently uses a weighted sum of the geospatial distance and the vector distance between two regions

$$D_{ij}^{n} = 0.5 * d_{ij}^{n} + 0.5 * V_{ij}^{n}$$

- Where d_{ij}^{n} is the geospatial distance between two regions at time n and V_{ij}^{n} is the vector distance between regions i and j at time n
  - Geospatial distance between two region refers to the distance between the centroid of the two regions

Ground Truth Preparation and Simulation

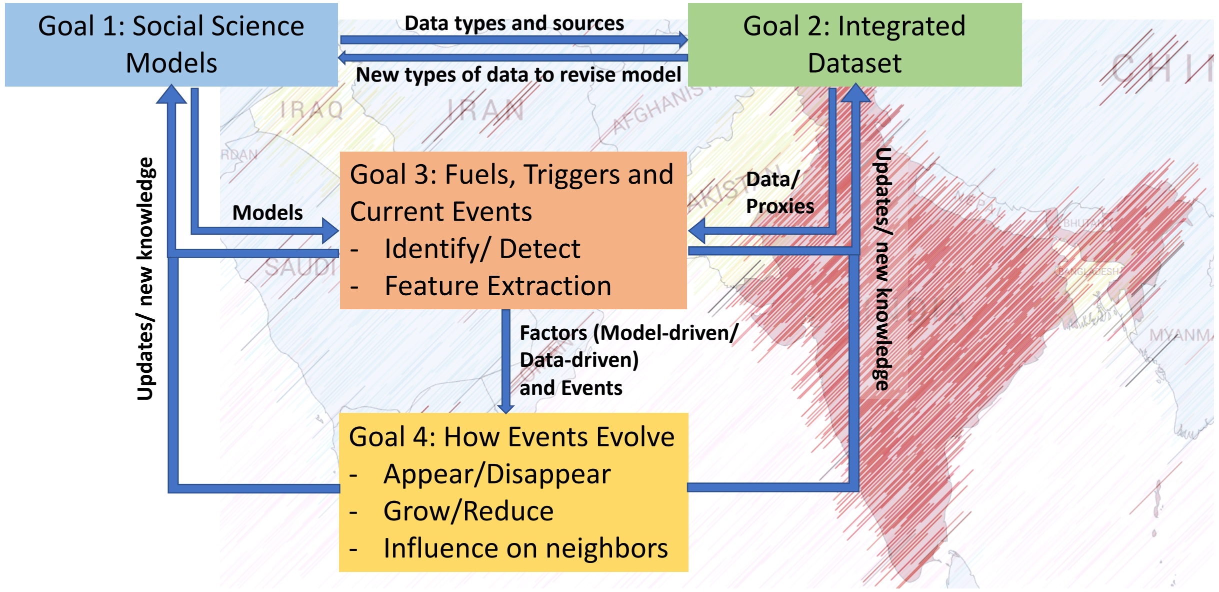
- We use ACLED (The Armed Conflict Location and Event Data Project) (Raleigh et al., 2010) for events that happened in Tamil Nadu from 2016 to 2019 on a monthly scale
- We count the number of events occurring in each district to identify the state of each district
  - A district is Infected (“I”) for each month that it has at least one recorded event
  - If a district has no recorded events: months prior to being infected are considered as susceptible “S” and months after being infected are considered as recovered “R”

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: an armed conflict location and event dataset: special data feature. *Journal of Peace Research*, *47*(5), 651–660.

On-Going Work

- Ongoing data collection and storage
- Develop strategies for incomplete, inconsistent and noisy data
- Continuous monitoring for new sources and source data formats
- Continuous search for SCIEGE based proxies of unrest and datasets to compute them
- Gazetteer extension and enhancement
- Developing gold standard training datasets for 5W analysis
- Integrated multi-agent models and simulations based on the SIR and SIS models, along with the trigger/inhibitor models.

Anticipate Unrest Events / Improved Situational Awareness



Goal 1: Social Science Models

Goal 2: Integrated Dataset

Data types and sources

New types of data to revise model

Goal 3: Fuels, Triggers and Current Events
- Identify/ Detect
- Feature Extraction

Models

Data/ Proxies

Updates/ new knowledge

Updates/ new knowledge

Factors (Model-driven/ Data-driven) and Events

Goal 4: How Events Evolve
- Appear/Disappear
- Grow/Reduce
- Influence on neighbors

Thank you for your attention!

Questions?

djoshi@citadel.edu