

Covid Rumours in Historical Context

Data and digital methods

Marty Steer (and the research team)

<https://historyandrumour.blogs.sas.ac.uk>

DHRH Seminar, 2022-03-21

CONSPIRACY

Outline

- The Project
- The (contemporary) Data
- The Analyses
- The Pipeline
- The Dashboard
- The End

1916 - emergence of Radio Waves

1918 - Spanish Flu **outbreak**

2003 - 3G introduced to the world

2003 - SARS **outbreak**

2009 - 4G introduced to the world

2009 - Swine flu **outbreak**

2019/20 - 5G introduced to the world

2019/20 - Coronavirus **outbreak** #WWG1WGA #COVID-19 #5G
#5GRollout

PATTERN: 120, 120, 120, 120

The Project

- AHRC funded project, ran for a year between 2020-2022
- Jo Fox, David Coast, James Smith, Jacob Forward, Kunika Kono, Rich Williams and Myself
- Aimed to conduct a longitudinal study
 - Contemporary data – diachronic linguistic analysis of COVID-19 tweets
 - Historical data – documents and archival evidence of past pandemics
- Policy paper in collaboration with History & Policy
- Built Classifiers, Taxonomies and a Semantic field explorer

The (contemporary) Data

- Corpus size
 - 35.5 million tweets (18.2m English)
 - Longitudinal scale
 - April 2020 to January 2022
 - 576 days
 - 12.3 million twitter accounts
 - 27GB text compressed
 - 234GB uncompressed

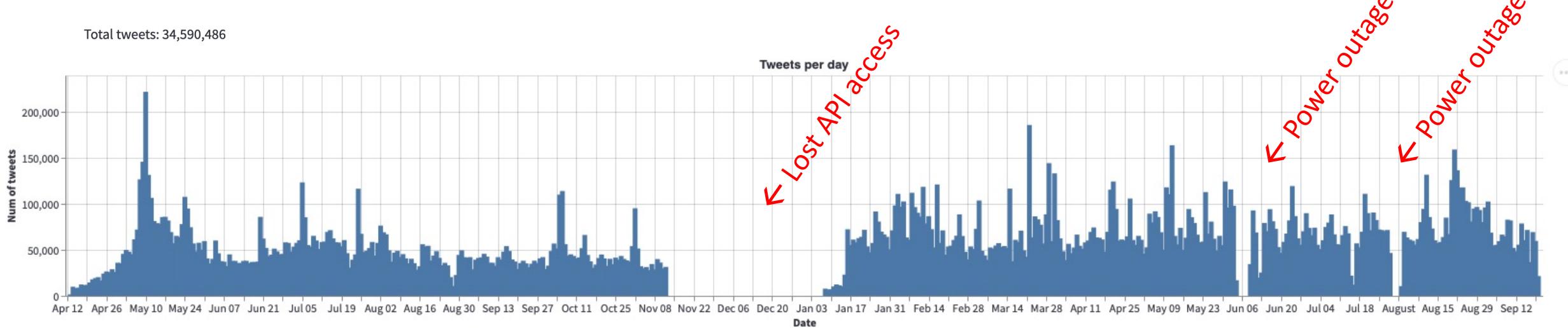
The Data

Twitter statistics

A statistical overview of tweets collected between April 2020 and January 2022 and related activities preceding and during the data collection period.

Tweets collected

Total tweets: 34,590,486



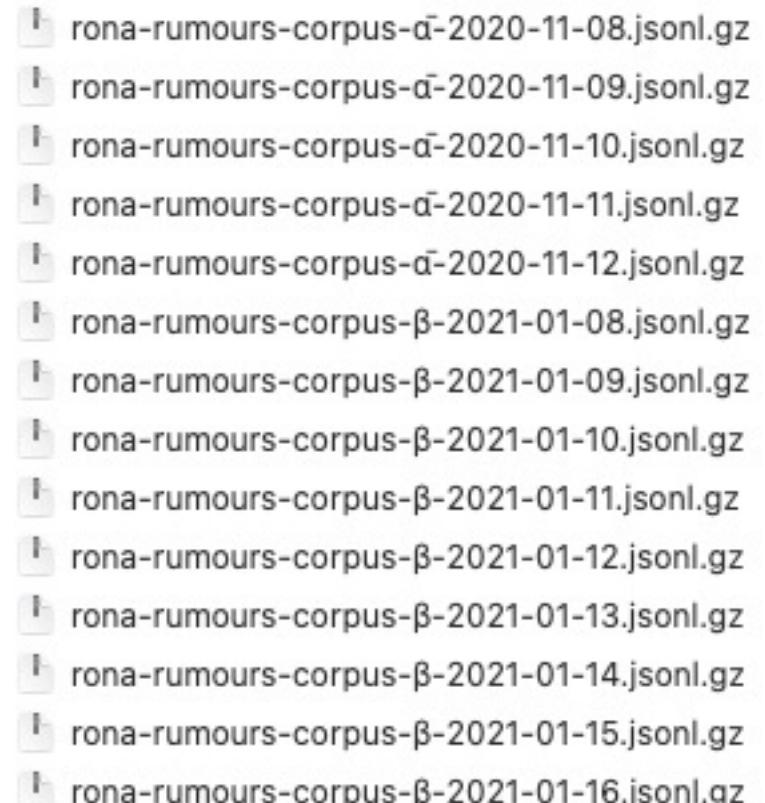
The Data

- Collection method
 - Twarc – Documenting the now
 - Crontab nightly rotation
 - JSONL text files
 - Rsync across local network
 - Raspberry Pi 4



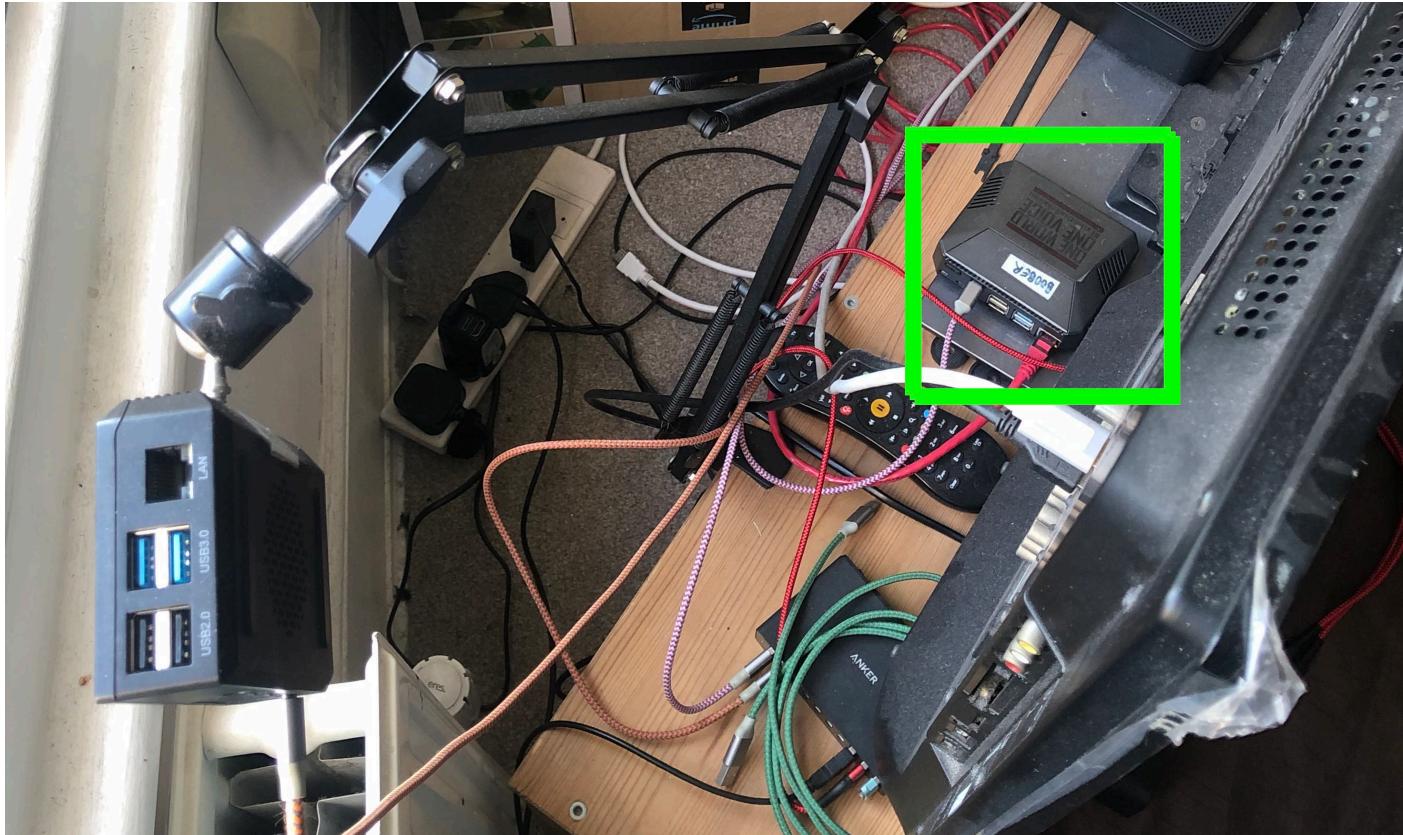
The Data

- Collection method
 - Twarc – Documenting the now
 - Crontab nightly rotation
 - JSONL text files
 - Rsync across local network
 - Raspberry Pi 4



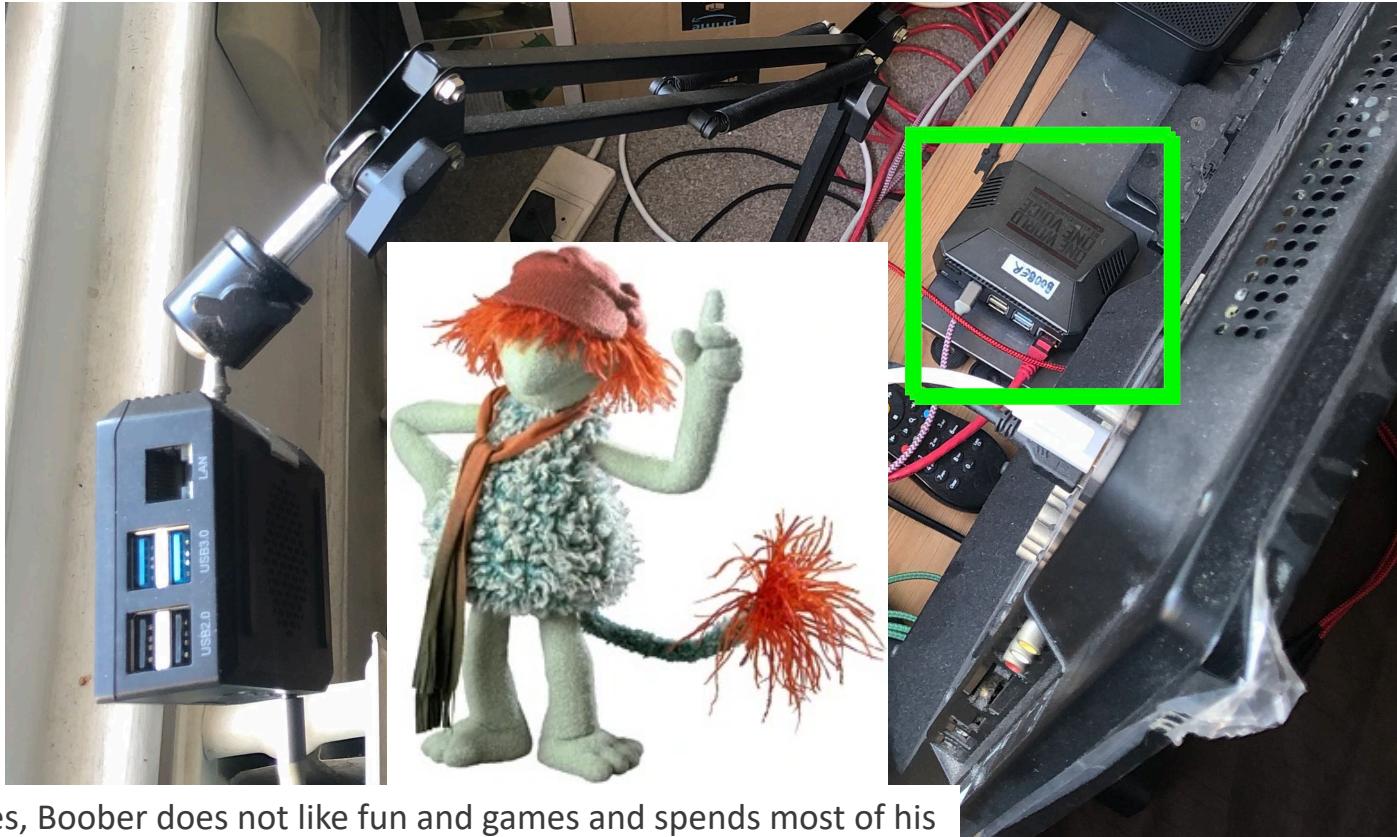
The Data

- Collection method
 - Twarc – Documenting the now
 - Crontab nightly rotation
 - JSONL text files
 - Rsync across local network
 - Raspberry Pi 4 (called boober)



The Data

- Collection method
 - Twarc – Documenting the now
 - Crontab nightly rotation
 - JSONL text files
- Rsync across local network
- Raspberry Pi 4 (called boober)
- Low resource compute
- Tiny team



Boober

Unlike other Fraggles, Boober does not like fun and games and spends most of his time worrying about doom and disease. When he's not worrying about himself, he's busy warning others. He is easily frightened and suffers from a variety of phobias... Because of his fears, he is very knowledgeable about health and superstition. He has a lucky charm for just about anything -- often multiples.

https://muppet.fandom.com/wiki/Boober_Fraggle

The Data

- Data management
 - Azure blob storage
 - Secure enclave
 - OneDrive – Confidential sharing
 - DropBox – General documents
 - Backups
 - 2TB USB-C M.2 SSD – Twistor
 - 6TB LAN NAS – Marjory

The Data

- Data management
 - Azure blob storage
 - Secure enclave
 - OneDrive – Confidential sharing
 - DropBox – General documents
 - Backups
 - 2TB USB-C M.2 SSD – Twistor
 - 6TB LAN NAS – Marjory

This is a shared folder.

My files > Research > COVID-tweets

file:///Users/martinsteer/Data/rona-rumours-corpus-alpha/wall.html

FB Moving Past the Finger Pointing - About Facebook Majority of Covid misinformation came from 12 peo... faculty.washington.edu/kstarbi/maddock_starb...

Title Here
created on the command line with [tware](#)

AscensionAgent
ascensionagent

RT @LotusOak2: #5G is the real silent killer! A live 5G mast is chucking out insane radiation. Birds are dying massively.
<https://t.co/kQD...>

843 retweets, 0 likes
Mon Apr 13 23:59:57 +0000 2020

New Type
NewType36999768

RT @CyrusAParsa1: Cheetah Robots Can Hunt You, Quarantine You With Swarm Tech on #5G From Big Tech-Socialist Governments to Tag You With Bi...

300 retweets, 0 likes
Mon Apr 13 23:59:55 +0000 2020

Andrew Morris
andrewmorrisuk

RT @HaroldSinnott: #5G will aid #SelfDrivingCars and #AutonomousVehicles @geoworldmedia via @MikeQuindazzi #AI #IoT #MachineLearning #BigDa...

46 retweets, 0 likes
Mon Apr 13 23:59:37 +0000 2020

Stephani M
s_macca02

RT @CyrusAParsa1: 5G WEAKENS IMMUNITY TO CHINA MADE CORONA VIRUS, OTHER ILLNESS' FOR HUMANS WATCH VIDEO, SUBSCRIBE FOR DAILY UPDATES ON U...

180 retweets, 0 likes
Mon Apr 13 23:59:23 +0000 2020

David Lowdell
DALowdell

@veteranstoday This is a complete bullshit story. #BBC is being used for psyop here. Nobody thinks 5G causes #Coronavirus. The link is between #5G and the chips that Bill Gates is proposing to insert into all of us through his vaccine program.

0 retweets, 2 likes
Mon Apr 13 23:59:20 +0000 2020

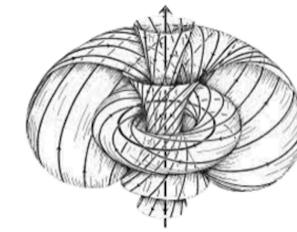
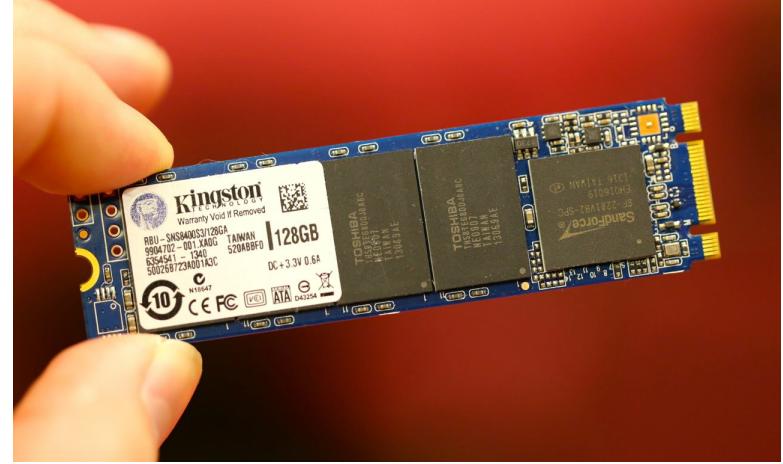
TheBlade
The_BladeRaker

RT @CyrusAParsa1: Cheetah Robots Can Hunt You, Quarantine You With Swarm Tech on #5G From Big Tech-Socialist Governments to Tag You With Bi...

300 retweets, 0 likes
Mon Apr 13 23:59:14 +0000 2020

The Data

- Data management
 - Azure blob storage
 - Secure enclave
 - OneDrive – Confidential sharing
 - DropBox – General documents
- Backups
 - 2TB USB-C M.2 SSD – Twistor
 - 6TB LAN NAS – Marjory



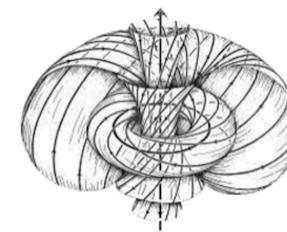
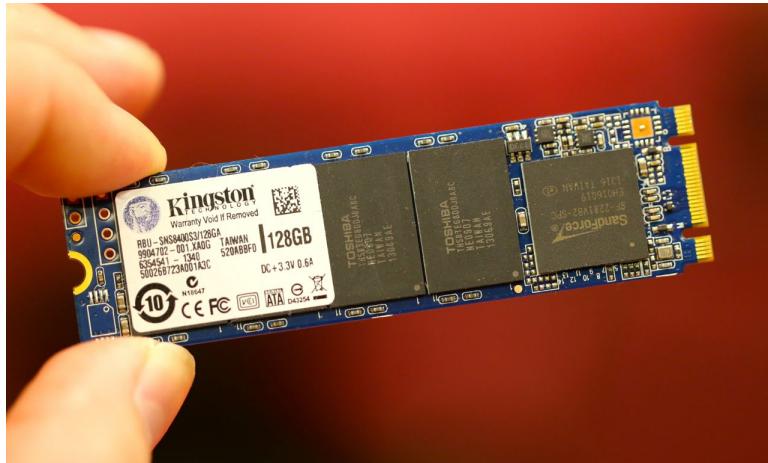
Twistor

The Data

- Data management
 - Azure blob storage
 - Secure enclave
 - OneDrive – Confidential sharing
 - DropBox – General documents
 - Backups
 - 2TB USB-C M.2 SSD – Twistor
 - 6TB LAN NAS – Marjory



Marjory



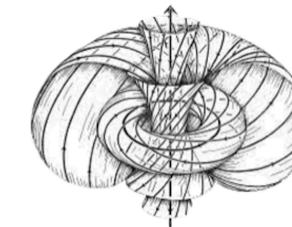
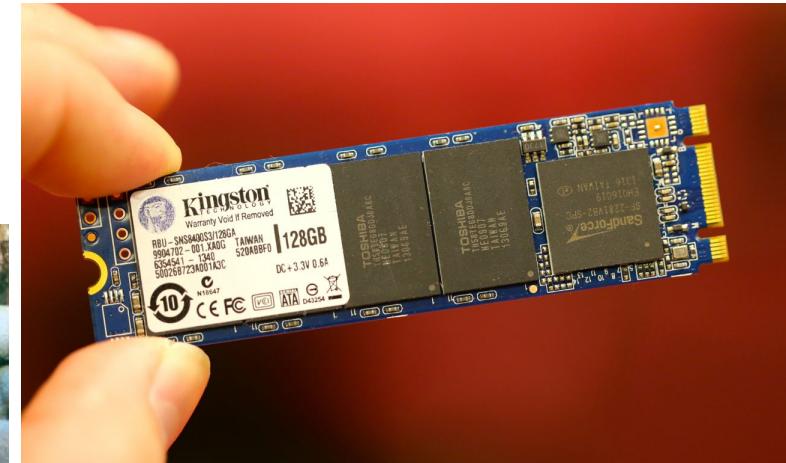
Twistor

The Data

- Data management
 - Azure blob storage
 - Secure enclave
 - OneDrive – Confidential sharing
 - DropBox – General documents
- Backups
 - 2TB USB-C M.2 SSD – Twistor
 - 6TB LAN NAS – Marjory



Marjory



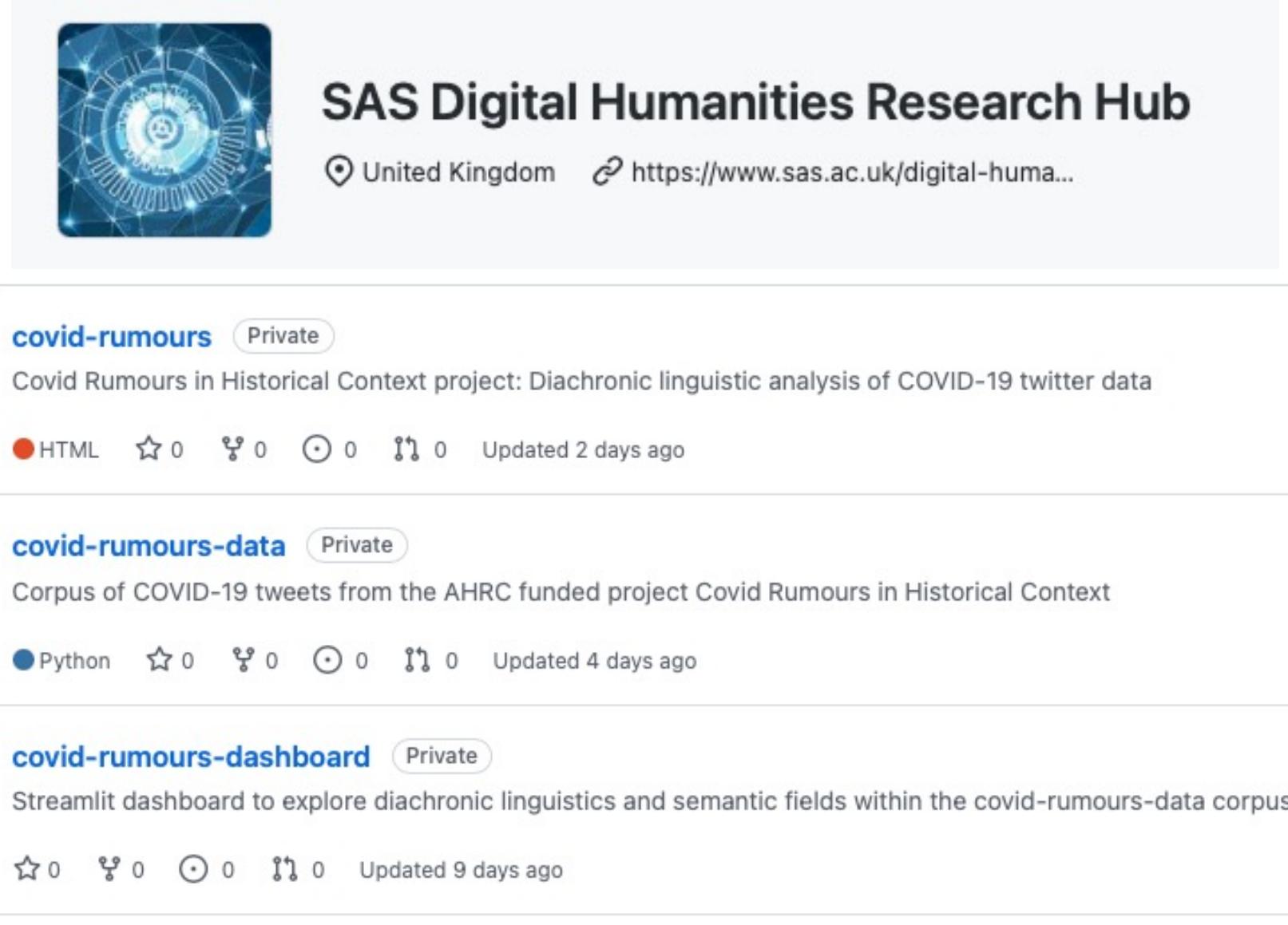
Twistor

Marjory the Trash Heap is a large, matronly, sentient compost heap from [Fraggle Rock](#). ... the Trash Heap knows all and sees all. In fact, Marjory *is* all: "I'm orange peels, I'm coffee grounds, I'm wisdom!"

https://muppet.fandom.com/wiki/Marjory_the_Trash_Heap

The Data

- Version control
 - Git
 - DVC
 - Makefile



SAS Digital Humanities Research Hub
United Kingdom <https://www.sas.ac.uk/digital-huma...>

covid-rumours Private
Covid Rumours in Historical Context project: Diachronic linguistic analysis of COVID-19 twitter data
HTML 0 stars 0 forks 0 issues 0 Updated 2 days ago

covid-rumours-data Private
Corpus of COVID-19 tweets from the AHRC funded project Covid Rumours in Historical Context
Python 0 stars 0 forks 0 issues 0 Updated 4 days ago

covid-rumours-dashboard Private
Streamlit dashboard to explore diachronic linguistics and semantic fields within the covid-rumours-data corpus
0 stars 0 forks 0 issues 0 Updated 9 days ago

The Data



Data Version Control

- Version control
 - Git
 - DVC
 - Makefile



The Data

- Version control

- Git
- DVC
- Makefile

Managing
Projects with



GNU Make

```
# ---
# tweets
# useful metadata fields from all tweets
T_CHUNKS_DIR = tweets
$(BUILD_DIR)/$(T_CHUNKS_DIR)/%.csv: $(DATA_DIR)/*.jsonl.gz
    mkdir -p $($D)
    gzip -d $^ | jq -r '[.id_str, .created_at, .user.name,
        .user.id_str, .user.created_at, .lang,
        .possibly_sensitive, .quote_count, .reply_count,
        .retweet_count, .favorite_count] | @csv' > $@
```

The Data

- Security and Privacy
 - Twitter's T&C
 - Share only Tweet ID's and limited metadata
 - Sensitive content
 - Sharing 'fake news' inappropriate
 - Secure enclave
 - No sensitive data stored in code repository
 - Derive Pseudonymous data

The Data

- Security and Privacy
 - Twitter's T&C
 - Share only Tweet ID's and limited metadata
 - Sensitive content
 - Sharing 'fake news' inappropriate
 - Secure enclave
 - No sensitive data stored in code repository
 - Derive Pseudonymous data

'An Interactive Anti-Coronavirus Toolkit'

Science Fiction tells of a future where suppression of human nature leads to eugenics, psychosis and murder – *That future is now.*

Author:

Independent Research Consultant
[REDACTED] Ltd
London, England, Great Britain
[REDACTED]@btinternet.com

127pg PDF URL found in our corpus

The Data

- Security and Privacy
 - Twitter's T&C
 - Tweet ID's and metadata
 - Sensitive content
 - Sharing 'fake news' inappropriate
 - Secure enclave
 - No sensitive data stored in code repository
 - Derive Pseudonymous data

master covid-rumours / .dvc / config



martysteer git submodule add git@github.com:SAS-DH/

1 contributor

8 lines (8 sloc) | 212 Bytes

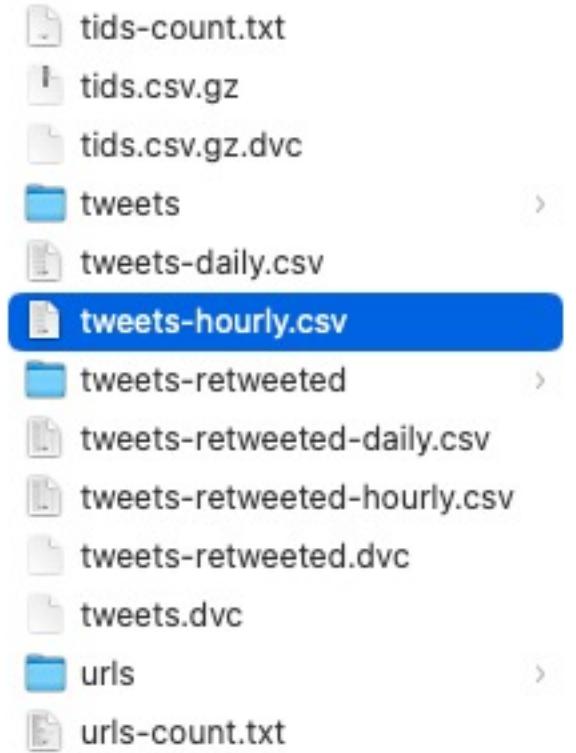
```
1 [core]
2     remote = azureprod
3     analytics = false
4     ['remote "marjory"']
5     url = ssh://marjory/home/dvcremotes
6     ['remote "azureprod"']
7     url = azure://twml/covid-rumours
8     account_name = sasprodtwittermlsa1
```

The Data

- Security and Privacy
 - Twitter's T&C
 - Tweet ID's and metadata
 - Sensitive content
 - Sharing 'fake news' inappropriate
 - Secure enclave
 - No sensitive data stored in code
 - Derive Pseudonymous data

created_at	tweet_count
2020-04-13 19:00:00	137
2020-04-13 20:00:00	482
2020-04-13 21:00:00	395
2020-04-13 22:00:00	321
2020-04-13 23:00:00	282
2020-04-14 00:00:00	238
2020-04-14 01:00:00	265
2020-04-14 02:00:00	244
2020-04-14 03:00:00	246
2020-04-14 04:00:00	213
2020-04-14 05:00:00	234
2020-04-14 06:00:00	308
2020-04-14 07:00:00	415
2020-04-14 08:00:00	471
2020-04-14 09:00:00	
2020-04-14 10:00:00	418

CSV



The Data

- What about the Content itself?

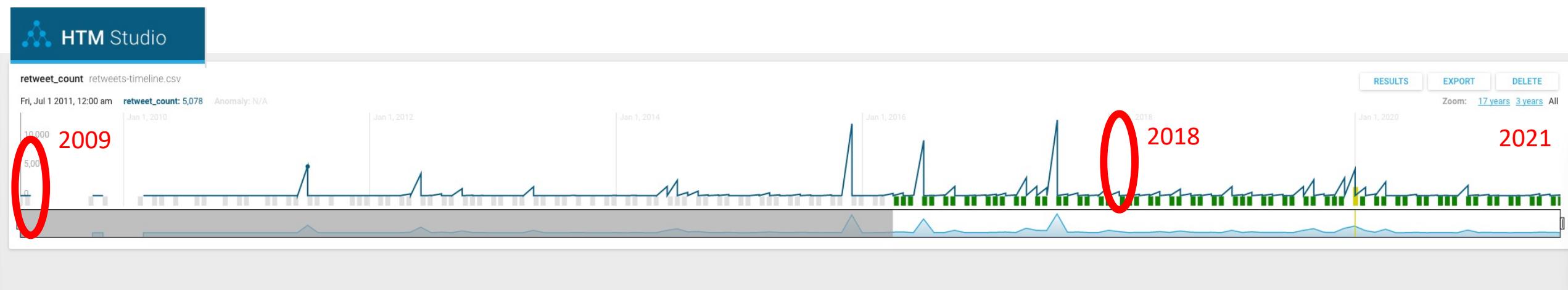
The Data

Content – Retweets, timeseries by original tweet date

- 2009-10,1,IMMUNOCONTRACEPTIVE HIDDEN IN THE FLU VACCINE
<http://bit.ly/XXXXXv> #nwo #h1n1 #swineflu #depopulation nssm200
- 2018-07,62,"#5G Call To Action Community Summit –
Premiere "There are 1000s of studies that prove that 5G cellular
is doing harm to our bodies, biology.

← Vaccine sterilizations
(2009)

← 5G doing harm
(2018)



The Data

Content – 5G's rapid semantic shift to full-on conspiracy

CONSPIRACY

@victoriav68 Brainwashing and programming, added Low frequency wave warfare, Ammonium sulfate and Toxic metals being sprayed into the atmosphere also coating your lungs and brain tissue. #5G activates the metal particles

They have 10 years to kill us. #2030

← Anxiety and a fresh new timeline

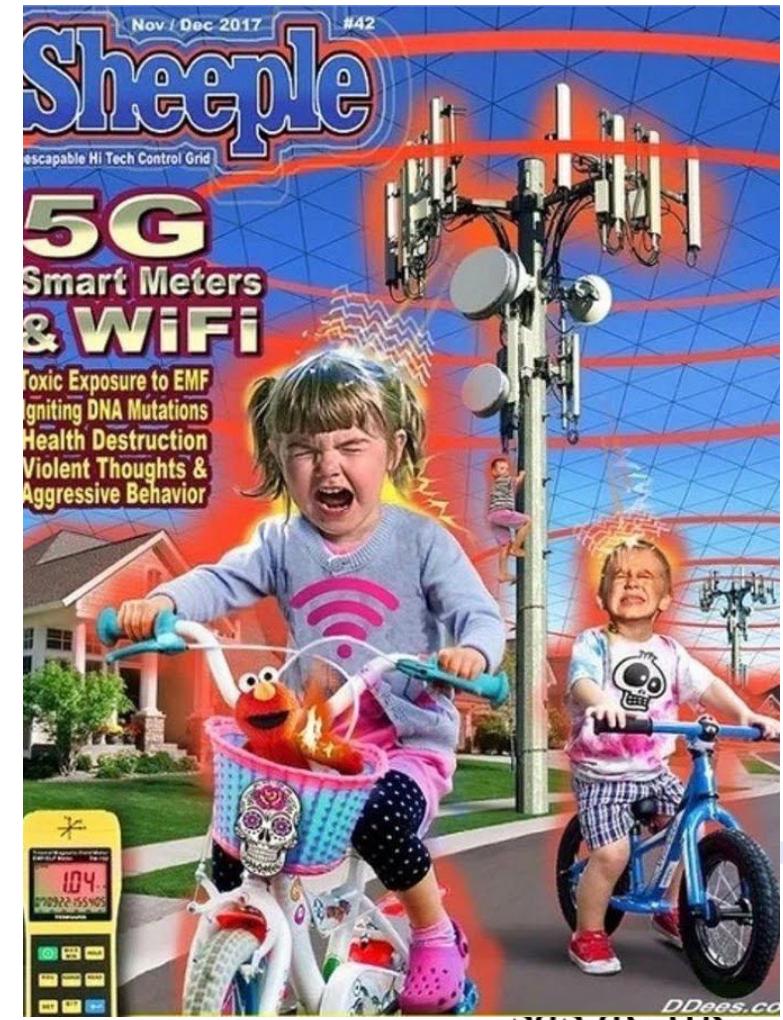
SCORE: 0.48

CONSPIRACY

It's a historical fact that industries have used lobbying power to suppress scientific data questioning safety. tobacco, pesticides, oil & gas to name a few. I'm amazed that telecoms is above suspicion when regulators dismiss peer reviewed science questioning safety. #5g

← Tyrannical Global industries are back

PATTERN: 151



The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

./jq

```
# ---
# tweets
# useful metadata fields from all tweets
T_CHUNKS_DIR = tweets
$(BUILD_DIR)}/${T_CHUNKS_DIR}/%.csv: ${DATA_DIR}/%.jsonl.gz
    mkdir -p ${@D}
    gzip -d ${^} | jq -r '[.id_str, .created_at, .user.name,
        .user.id_str, .user.created_at, .lang,
        .possibly_sensitive, .quote_count, .reply_count,
        .retweet_count, .favorite_count] | @csv' > ${@}
```

The Analyses

```
$ (MODEL_DIR)/grabbag/walls/%.html: $(MODEL_DIR)/grabbag/sample-data/%.jsonl.gz  
[[ -d $($@D) ]] || mkdir -p $($@D)  
gzcat $^ | $(SCRIPT_DIR)/twarc-wall.py --title $($@F) > $($@)
```

- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

A screenshot of a web browser window displaying a tweet wall. The title bar says "file:///Users/martinsteer/Data/rona-rumours-corpus-alpha/wall.html". The main content area is titled "Title Here" and contains several tweets. One tweet from "AscensionAgent" discusses 5G radiation and bird deaths. Another tweet from "New Type" links to a story about Cheetah Robots. A third tweet from "Andrew Morris" links to a story about self-driving cars. The browser interface includes a toolbar with various icons and a status bar at the bottom.

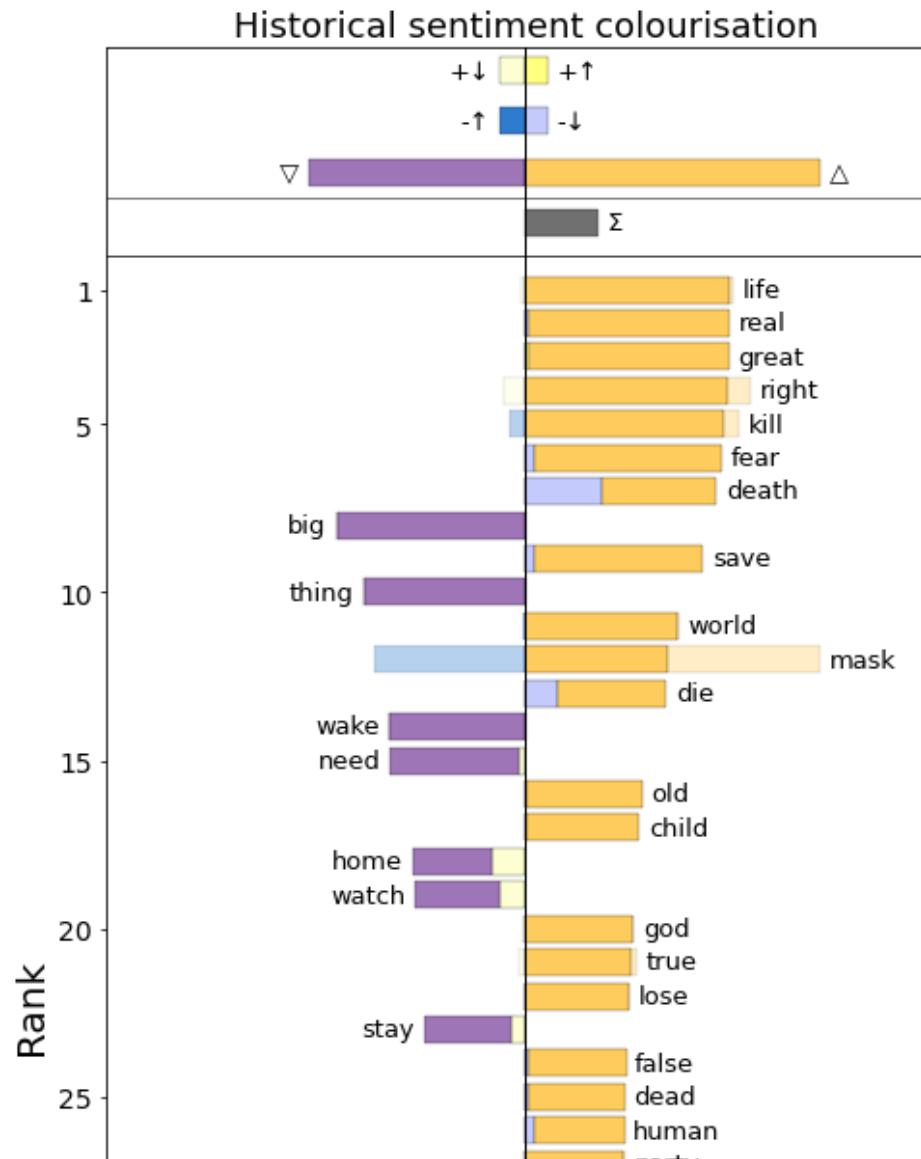
Title Here

created on the command line with [twarc](#)

User	Tweet Content	Retweets	Created At
AscensionAgent	RT @LotusOak2: #5G is the real silent killer! A live 5G mast is chucking out insane radiation. Birds are dying massively. https://t.co/kQD...	843	Mon Apr 13 23:59:57 +0000 2020
New Type	RT @CyrusAParsa1: Cheetah Robots Can Hunt You, Quarantine You With Swarm Tech on #5G From Big Tech-Socialist Governments to Tag You With Bi...	300	Mon Apr 13 23:59:55 +0000 2020
Andrew Morris	RT @HaroldSinnott: #5G will aid #SelfDrivingCars and #AutonomousVehicles @geoworldmedia via @MikeQuindazzi #AI #IoT #MachineLearning #BigDa...	46	Mon Apr 13 23:59:37 +0000 2020
Stephani M	RT @CyrusAParsa1: 5G WEAKENS IMMUNITY TO CHINA MADE CORONA VIRUS, OTHER ILLNESS' FOR HUMANS WATCH VIDEO, SUBSCRIBE FOR DAILY UPDATES		
David Lowdell	@veteranstoday This is a complete bullshit story. #BBC is being used for psyop here. Nobody thinks 5G causes #Coronavirus. The link is between #5G and the chips that Bill Gates has implanted in us all		
TheBlade	RT @CyrusAParsa1: Cheetah Robots Can Hunt You, Quarantine You With Swarm Tech on #5G From Big Tech-Socialist Governments to Tag You With Bi...		

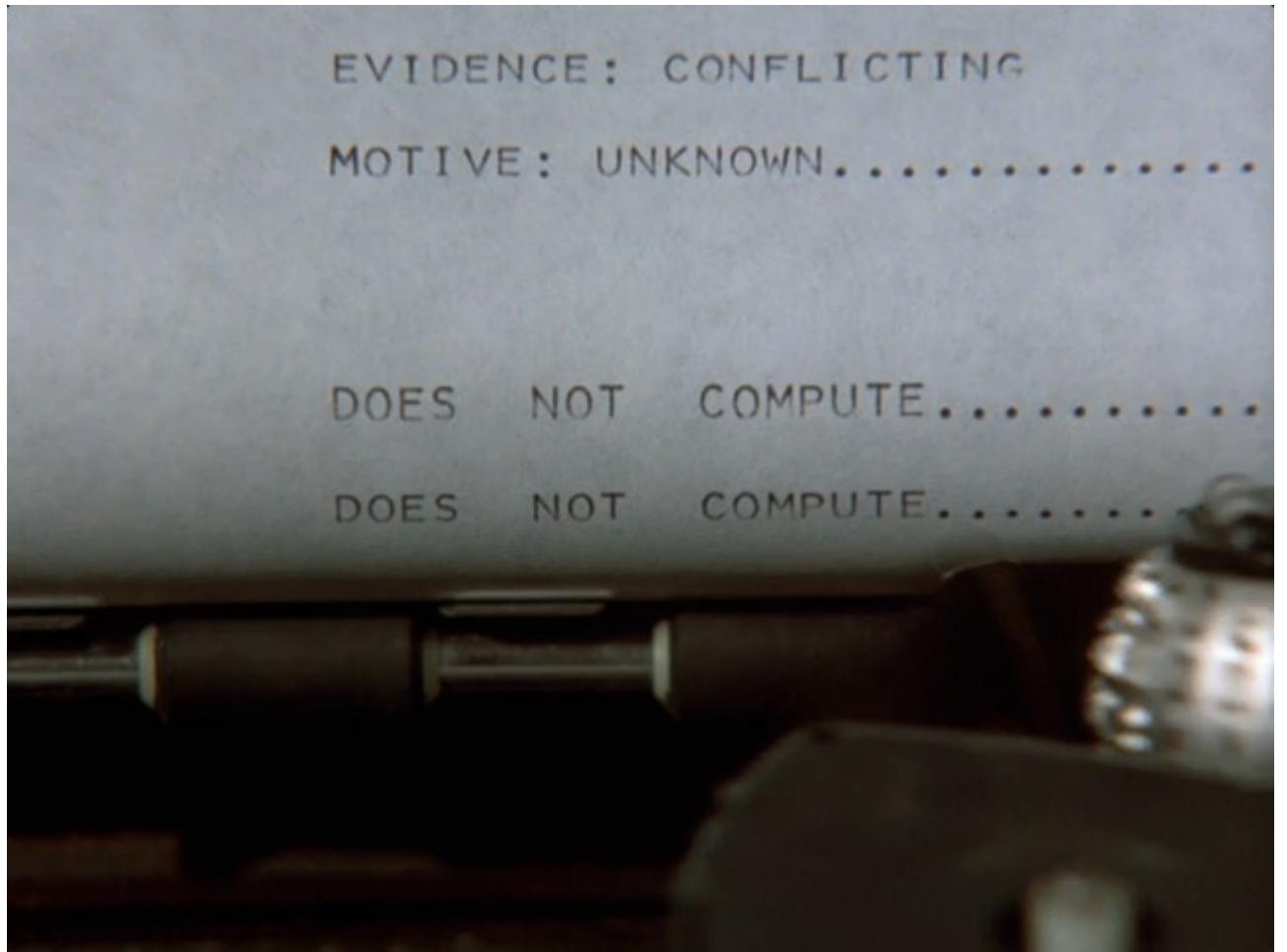
The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data



The Analyses

- Words, users, URLs, hashtags, languages
- ~~Sentiment analysis~~
- ~~User network analyses~~
- ~~Hashtag/Term cluster analyses~~
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data



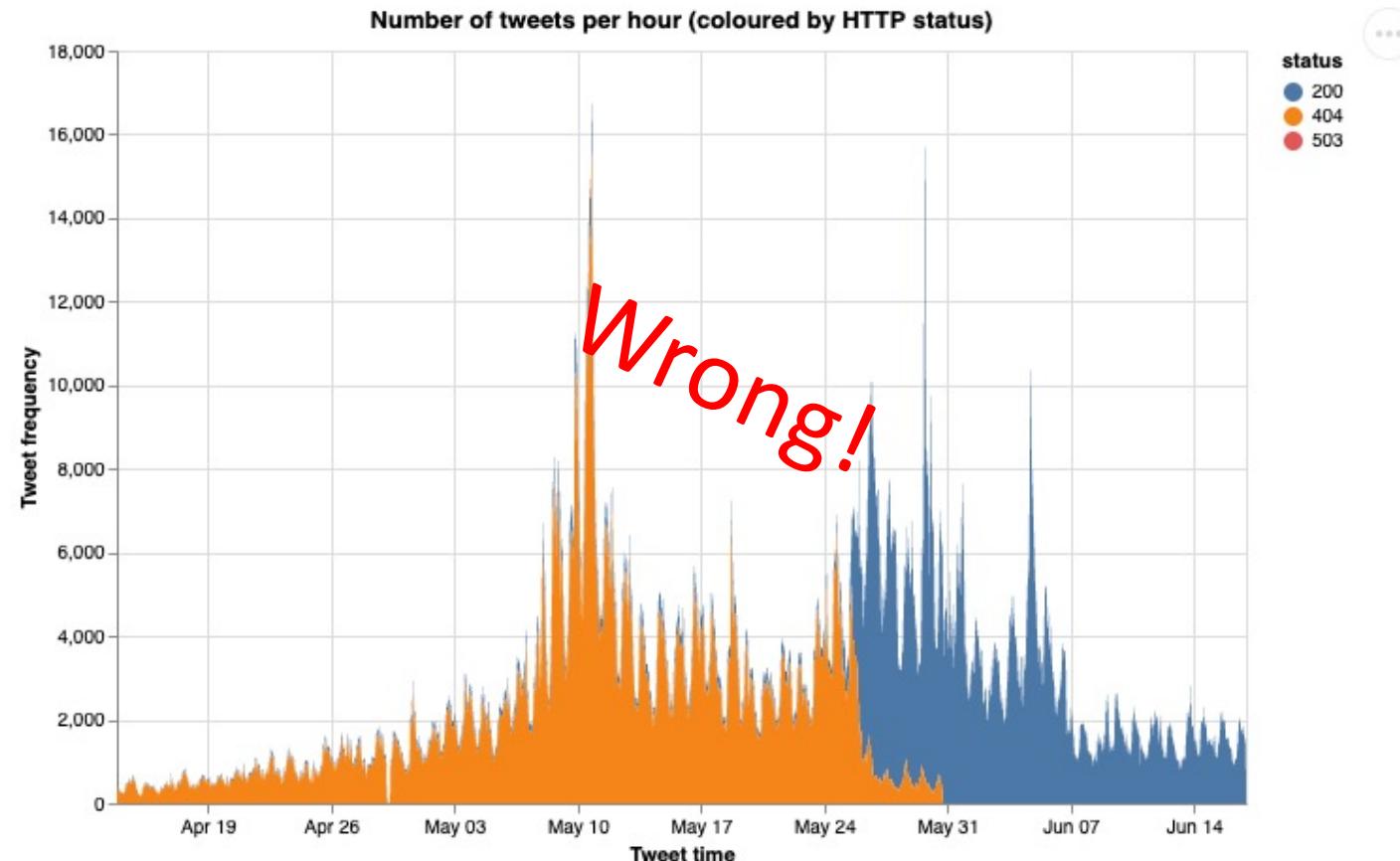
The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

Took 3 months for script to collect data and was entirely wrong!

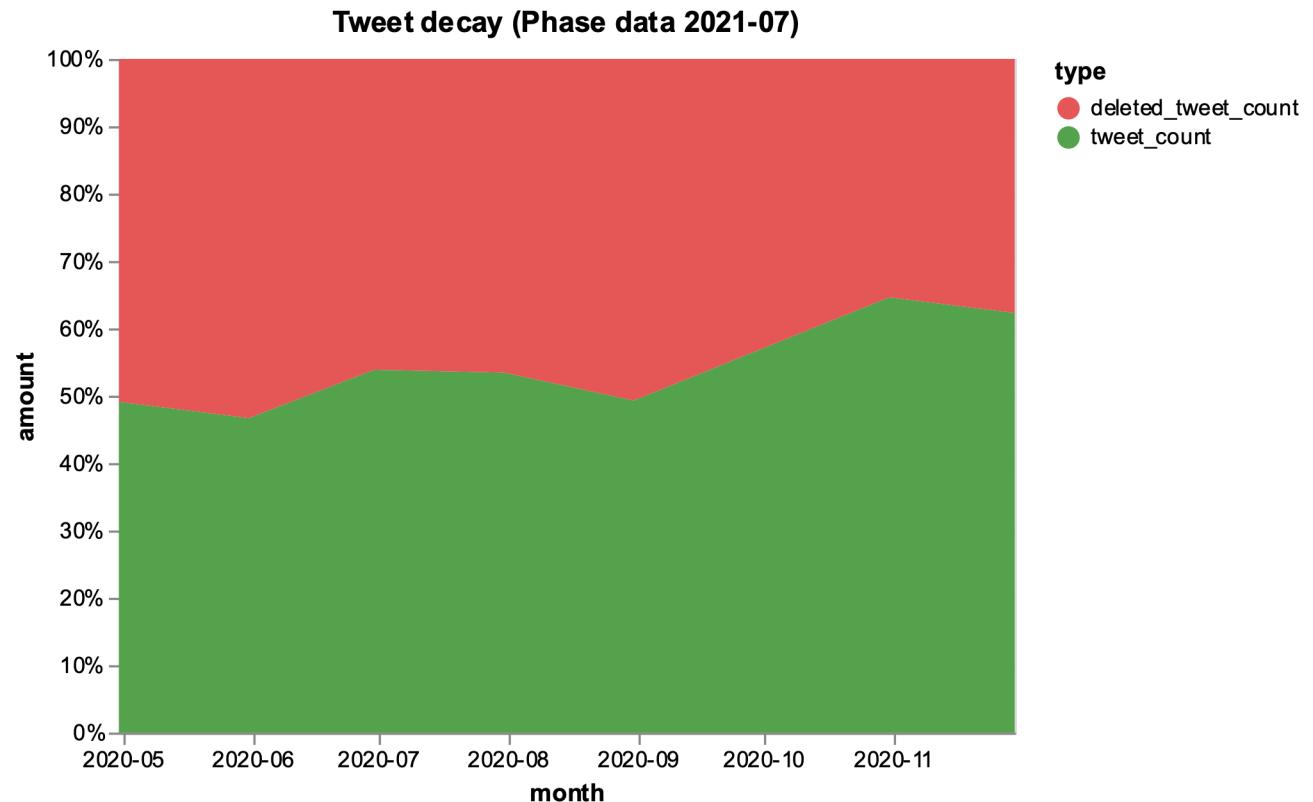
Problem: 404 errors due to jq's LONG INT truncating tweet id's to end in 00
e.g

1249849839975686145 → 1249849839975686100
1249849831100538880 → 1249849831100538800



The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data



Added tweet attrition feature for us

The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

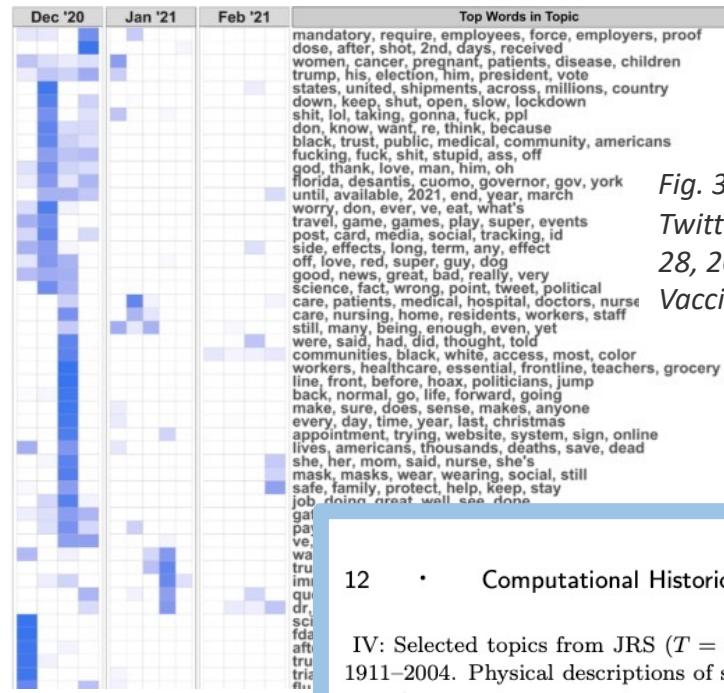
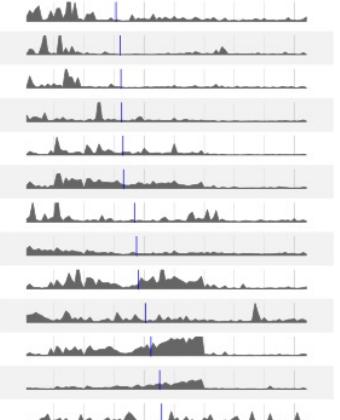


Fig. 3. Weekly variation of data-driven COVID-19 Twitter topics from December 1, 2020 to February 28, 2021. Guntuku, Buttenheim, Sherman, Merchant, Vaccine, Volume 39, Issue 30, 2021

12 • Computational Historiography

IV: Selected topics from JRS ($T = 150$), in order by mean publication year (blue vertical line), 1911–2004. Physical descriptions of sites dominate the earlier 20th century, while recent decades have focused on social and economic history.

fig tomb walls feet level wall tombs side plan room above house building
blocks two floor small large remains
province asia roman antioch galatia minor cilicia name colonia governor capadocia pamphylia strabo pisidia war probably studies part city
inscription stone inscriptions letters name published inscribed two above
high dedication ramsay monument broken below plate monuments block
bronze museum now collection two pl glass fig british objects found head
silver shape long pieces similar plate ashmolean
wall fort britain hadrian antonine roman forts stone scotland occupation
work north vallum turf hill found building richmond milecastle
found roman pottery site ml coins near ware samian road small date occu-
pation fragments hill objects iron gravel well
road river via miles valley bridge route through near modern ancient along
roads map course plain line point bank
found part place two now great light still another remains being discovered
while good under little among
wall ft rampart ditch stone gate walls wide clay tower foundation defences
built bank section roman inner cut building
left right figure hand head relief front figures two side behind back shield
panel arm over round above standing
ft building century house street buildings floor two wide room fourth second
stone rooms bath jrs timber side site
found reads part fragment roman graffiti cut above base site altar name two
tile building now drag face stamp
coins coinage coin silver types gold bronze mint struck reverse series issues
denarii roman denarius value issue type hoards



The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data



Works on long text
No so good on short text
Not diachronic
(unless you structure your corpus that way)

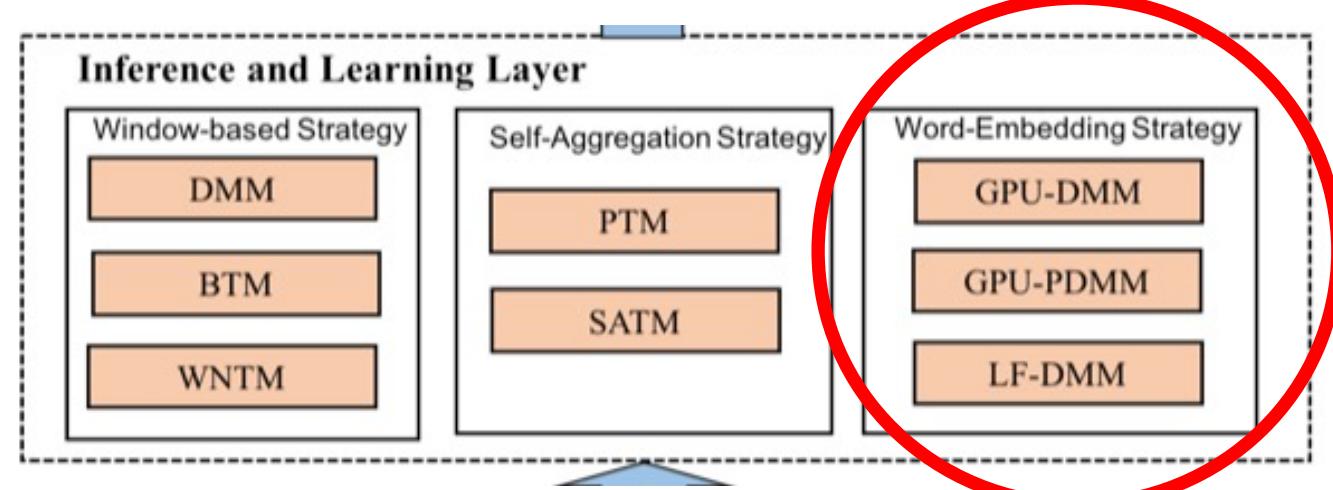
The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

STTM: A Library of Short Text Topic Modeling

This is a Java (Version=1.8) based open-source library for short text topic modeling algorithms. The library is designed to facilitate the development of short text topic modeling algorithms and make comparisons between the new models and existing ones available. STTM is open-sourced at [Here](#).

STTM is maintained by [Jipeng Qiang](#) (Yangzhou, China).



Word embeddings (diachronic and tweets)

- [1] Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, Barbara McGillivray, 2019. "Mining the UK Web Archive for Semantic Change Detection". In *RANLP*.
- **GloVe: Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 200d vectors, 1.42 GB download)**

The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

Top topic words

LFDM (10.18 hours)	GPU-PDMM (50.25 hours)
868 covid	2197 5g
583 depopulation	1416 new
569 new	1304 covid
548 vaccine	1160 scandemic
544 agenda21	1120 covid19
531 5g	965 covidiot
528 time	919 like
524 ve	902 world
521 world	894 pandemic
520 good	874 people

Used topics to get a macro view of the English corpus
Wasn't able to use historical word embeddings successfully
Tweets hashtag themselves as a form of topicality
so this analysis wasn't much more useful than word frequencies

The Analyses

- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data



prodigy

Radically efficient machine teaching.
An annotation tool powered
by active learning.

FROM THE MAKERS OF SPACY

```
- category: "VACCINES"
- vocabulary: ...
- category: "Motives"
- vocabulary:
  - money
  - money-making
  - scheme
  - big pharma
  - Bill Gates
  - manufactured
  - deliberate
  - pretext
  - introduce
  - vaccines
```

Vectors and Terminology

terms.teach BINARY

- Interface: text
- Saves: accepted and rejected terms to the database
- Updates: target vector used for similarity comparison
- Use case: building terminology lists and pre-processing candidates for NER training

```
- category: "Misc"
- category: "Vaccine is unnecessary"
- category: "Vaccines are deliberate"
- category: "Tweets debunking vaccines"
```

The Analyses

- Words, users, URLs, hashtags, languages
 - Sentiment analysis
 - User network analyses
 - Hashtag/Term cluster analyses
 - Short Text Topic modelling
 - Diachronic Topic visualisation
 - Tweet walls
 - Rumour vocabularies
 - Classifiers
 - Phase/Attrition data
- CONSPIRACY
 - ORIGNS
 - CURES
 - VACCINES

```
- category: "CONSPIRACY"
- vocabulary: " "
- category: "COVID IS EXAGGERATE"
- vocabulary: " "
- category: "DEATH OR CASE MIS"
- vocabulary: " "
- category: "GLOBAL CONSPIRACY"
- vocabulary: " "
- category: "TYRANNY"
- vocabulary: " "
- category: "VACCINES"
- category: "DEPOPULATION"
- vocabulary: " "
- category: "MICROCHIPS AND"
- vocabulary: " "
- category: "BIG PHARMA"
- vocabulary: " "
- category: "PLOT BY DEMOCRATIC"
- vocabulary: " "
- category: "SATAN"
- vocabulary: " "
- category: "ZOMBIES"
- vocabulary: " "
- category: "ROBOTS"
- vocabulary: " "
- category: "BLACK LIVES MATTER (BLM)"
- vocabulary: " "
- category: "ANTI-CONSPIRACY"
- vocabulary: " "
```

```
- category: "VACCINES"
- vocabulary: " "
- category: "Motives"
- vocabulary:
  - money
  - money-making
  - scheme
  - big pharma
  - Bill Gates
  - manufactured
  - deliberate
  - pretext
  - introduce
  - vaccines
- category: "Effects"
  - category: "Vaccines are harmful to heal"
  - category: "Vaccines contain microchips"
    - vocabulary:
      - vaccines
      - microchips
      - track
      - population
  - category: "Vaccines won't work / have no"
    - vocabulary:
      - vaccines
      - won't work
      - have never worked
- category: "Misc"
  - category: "Vaccine is unnecessary because"
  - category: "Vaccines are deliberately be"
  - category: "Tweets debunking vaccine rumo
```

The Analyses

- Words, users, URLs, hashtags, languages
 - Sentiment analysis
 - User network analyses
 - Hashtag/Term cluster analyses
 - Short Text Topic modelling
 - Diachronic Topic visualisation
 - Tweet walls
 - Rumour vocabularies
 - Classifiers
 - Phase/Attrition data
- Facet ngrams
 - Initiate field search

Tweets Explorer

Tweets: All

Date range: 2020-04-13 - 2022-01-10

Words to explore

truth X facts X lies X scam X hoax X



Update

Advanced options

Suggest words only from taxonomy category:

None

Origins and causes

Origins and causes > Wuhan lab

Origins and causes > Wuhan lab > Bioweapon

Origins and causes > Wuhan lab > Accidental release

Origins and causes > Historiographical

Origins and causes > Zoonotic

Origins and causes > 5g

The Analyses

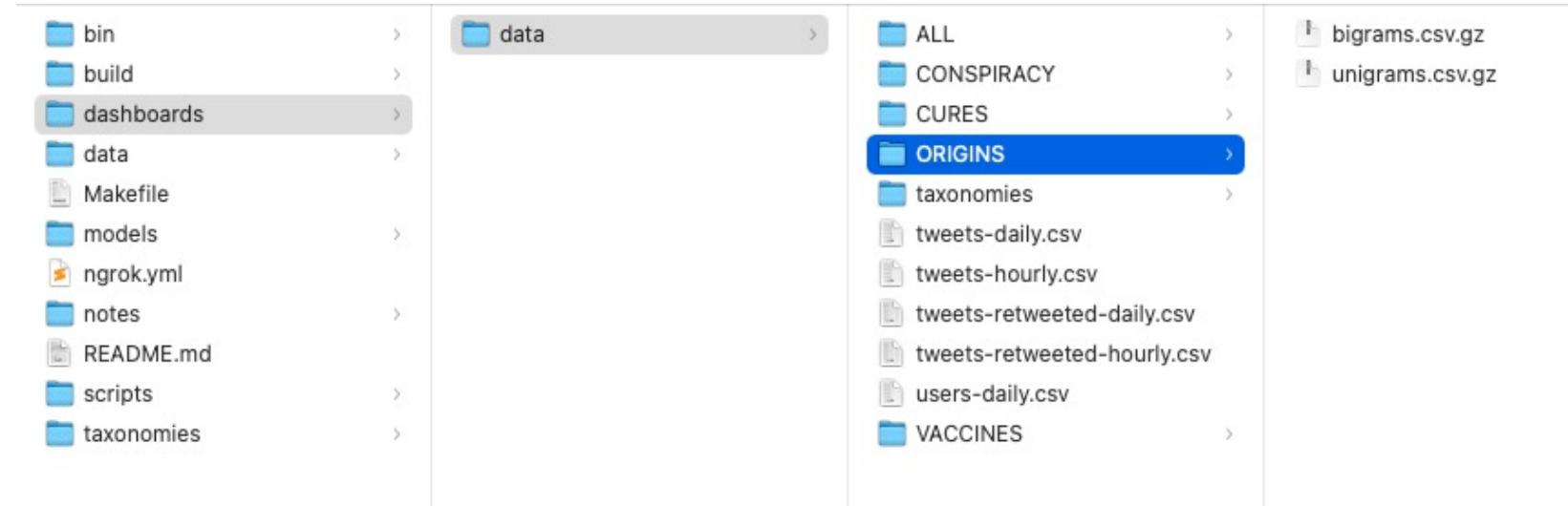
- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

```
===== Training pipeline =====
<0x1b>[38;5;4m Pipeline: ['tok2vec', 'tagger', 'parser', 'attribute', 'lemmatizer', 'ner', 'textcat_multilabel']<0x1b>[0m
<0x1b>[38;5;4m Frozen components: ['tagger', 'parser', 'attribute', 'lemmatizer', 'ner']<0x1b>[0m
<0x1b>[38;5;4m Initial learn rate: 0.001<0x1b>[0m
E   #      LOSS_TOK2VEC  LOSS_TEXTC...  CATS_SCORE  SCORE
---  ---  -----  -----  -----  -----
0     0       0.00        0.25      47.25    0.47
3    1000     0.00        42.83     73.40    0.73
13   2000     0.00        0.86      74.57    0.75
34   3000     0.00        0.05      74.46    0.74
56   4000     0.00        0.02      74.14    0.74
79   5000     0.00        0.01      73.83    0.74
101  6000     0.00        0.01      73.79    0.74
123  7000     0.00        0.00      73.76    0.74
<0x1b>[38;5;2m✓ Saved pipeline to output directory<0x1b>[0m
.../models/conspiracy-classifier/model-last
```

- Historians annotated data using Prodigy for all 4 top-level categories
- Reviewed label agreements and discarded disagreements
- Trained models using
 - Prodigy and then Spacy
 - On CPU and GPU
 - 80/20 cross-validation
 - Not super accurate (avg. 70%) but good enough for prototype

The Pipeline

1. Harvesting
2. Cleaning
3. Preprocessing
4. Sampling
5. Annotating
6. Classifying
7. Visualising
8. Exploring
9. Interpreting



Classified complete English corpus into subsets
84MB of classified temporal word colocations in total
Very small data footprint for the visualization tool
Just the 'signals' we want have been extracted from the corpus

The Pipeline

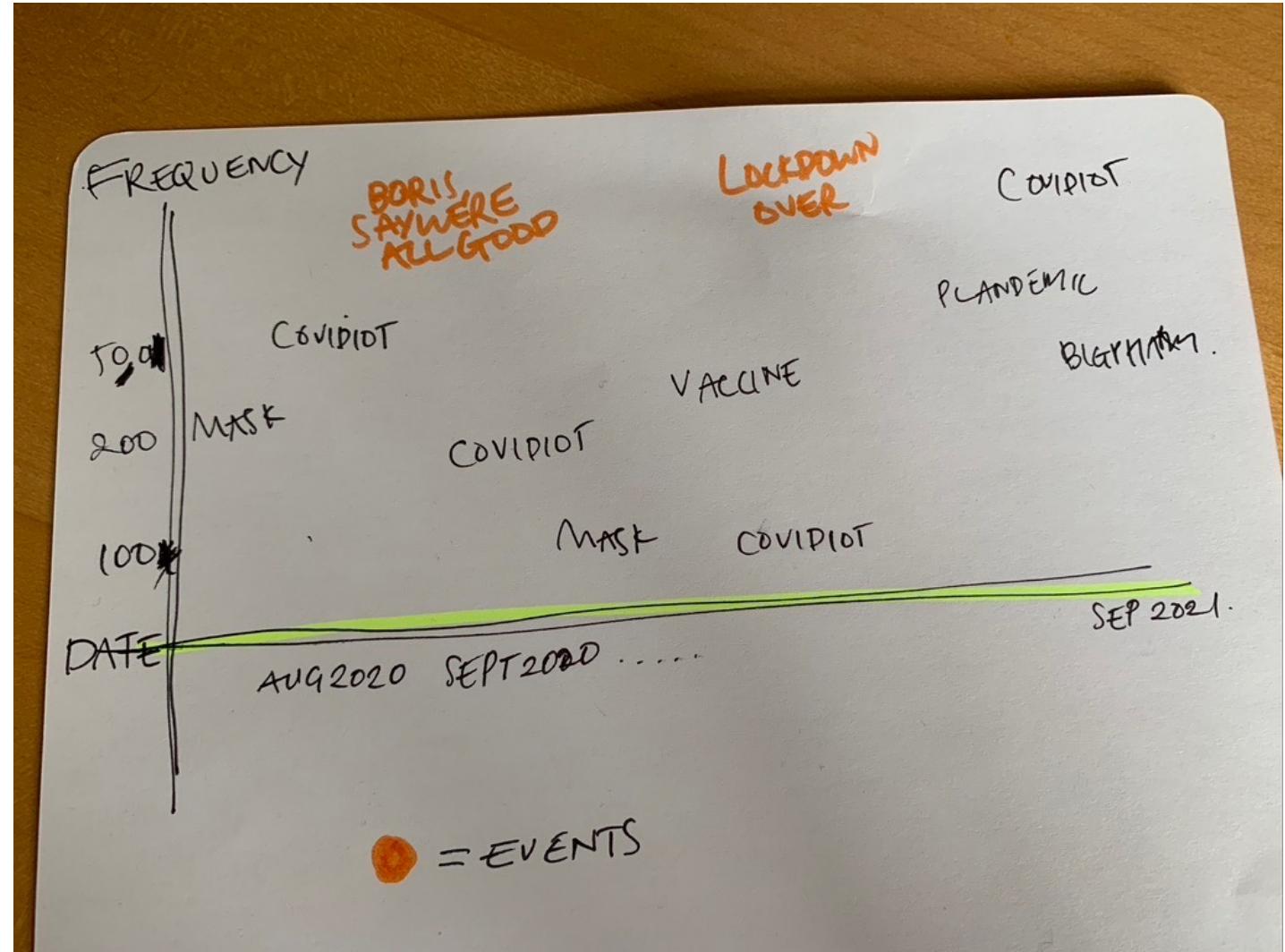
1. Harvesting
2. Cleaning
3. Preprocessing
4. Sampling
5. Annotating
6. Classifying
7. Visualising
8. Exploring
9. Interpreting

The Pipeline

1. Harvesting 234 GB JSONL source data
2. Cleaning Remove stopwords, hash characters, usernames, etc. – a ‘hybrid’ clean
3. Preprocessing Extract useful field frequencies (text, retweets, usernames, etc.)
4. Sampling 4000 random tweets from first 4 months of corpus alpha
5. Annotating Prodi.gy annotation tool for terminology lists and labelling data
6. Classifying Spacy for training the 4 classifier models and classifying the corpus
7. Visualising 84MB “signal processed” data for the dashboard
8. Exploring Compare diachronic ngrams and subsets of collocated semantic fields
9. Interpreting ??

The Dashboard

- Diachronic comparison
- Categorical comparison
- Word frequencies
- Word colocations

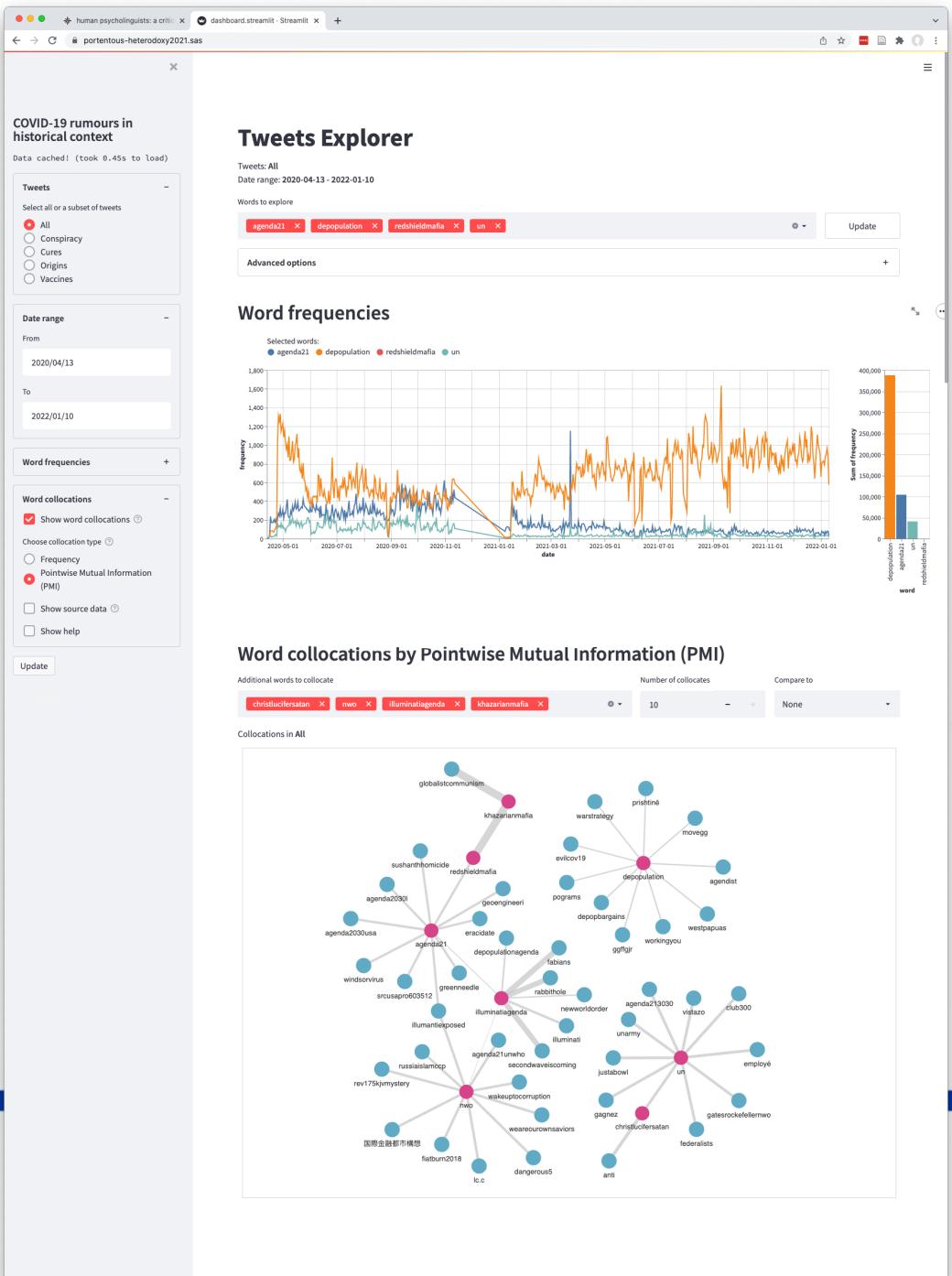


Kunika's prototype sketch

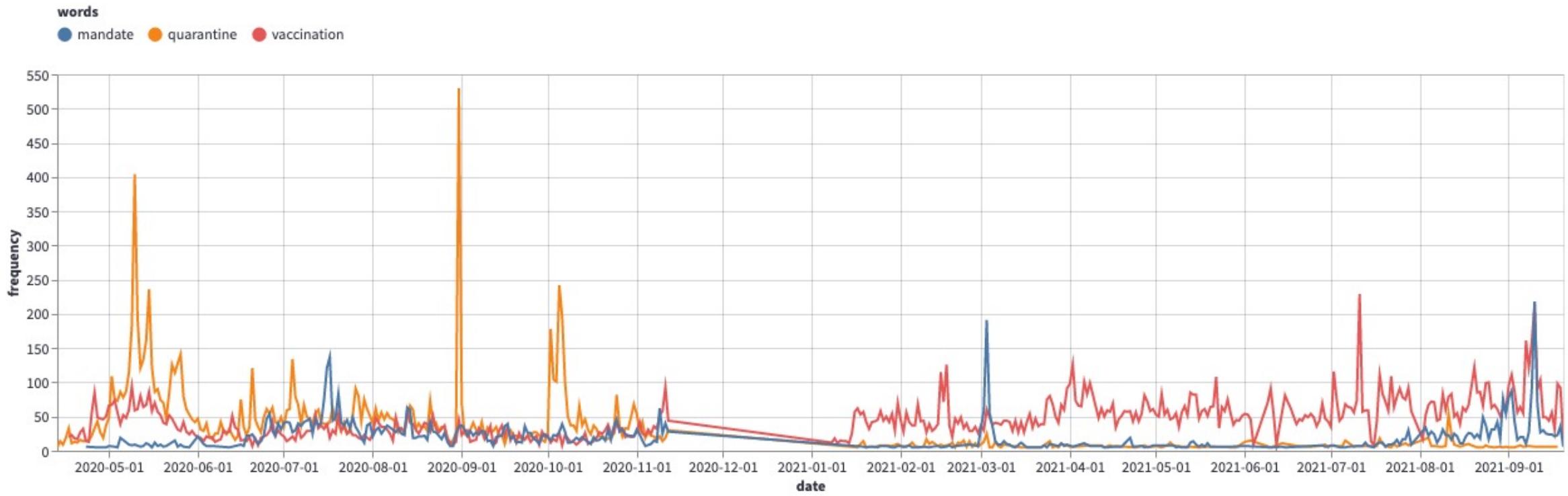
The Dashboard

- Diachronic comparison
- Categorical comparison
- Word frequencies
- Word colocations

Hard to compare many dimensions at once and meaningfully – but we did it!



Word frequencies



Trials of two anti-malarial to	Major breakthrough in...	UK to try the first 'challenge'	The UK announces cha...	British PM, Boris Johnson, ...	Half of the UK population has	Sajid Javid appointed Health
UK human COVID-19 vac...	WHO urges for a rapid increase	UK bans gatherings of m...	UK announces plans that can li...	15 million Britons have no...	7 deaths reported amon...	England removes the vast of
British Prime Minister Boris	US agrees to pay Pfizer \$2bn	AstraZeneca resumes vaccin...	The number of people in the UK	UK sets new record of 873,784	All over 40s in the UK will be	Half of under-30s in th...
The UK begins human testing in	Volunteers in the UK, Brazil	Oxford University resu...	UK secures another 2 million	UK announces that all UK will	UK Prime Minister Boris t...	UK hits 60million vaccine
UK Prime Minister Boris ...	Twitter restricts Donald...	UK announces 4 week national	AstraZeneca states that it its	UK lights candles in of th...	Data shows a third of UK are	European and UK drugs identify
UK Prime Minister Boris c...	Lockdown reimposed on a...	UK announces a new three-tier	BioNTech/Pfizer and Moderna EU	One in three adults in the UK	UK advises against giving the	Britain records no new deaths
UK makes masks compulsory on	UK imposes local lockdown ...	Pfizer announces that...	Oxford-AstraZe...	UK's emergency Covid extended	Prime Minister Boris Johnson	European Medicines Age...
US president Donald Trump is	UK rejects offer to join the	Johnson and Johnson pause...	European Medicines Age...	Queen Elizabeth II urges those	WHO declares that the risks	UK reports its highest number
UK Christ...	UK revo...	UK Health...	England enters	NHS England...	Health experts...	1,200

Event context

The Dashboard – Worldwide medical tyranny

Word collocations by Frequency

Additional words to collocate

mandate X mandates X

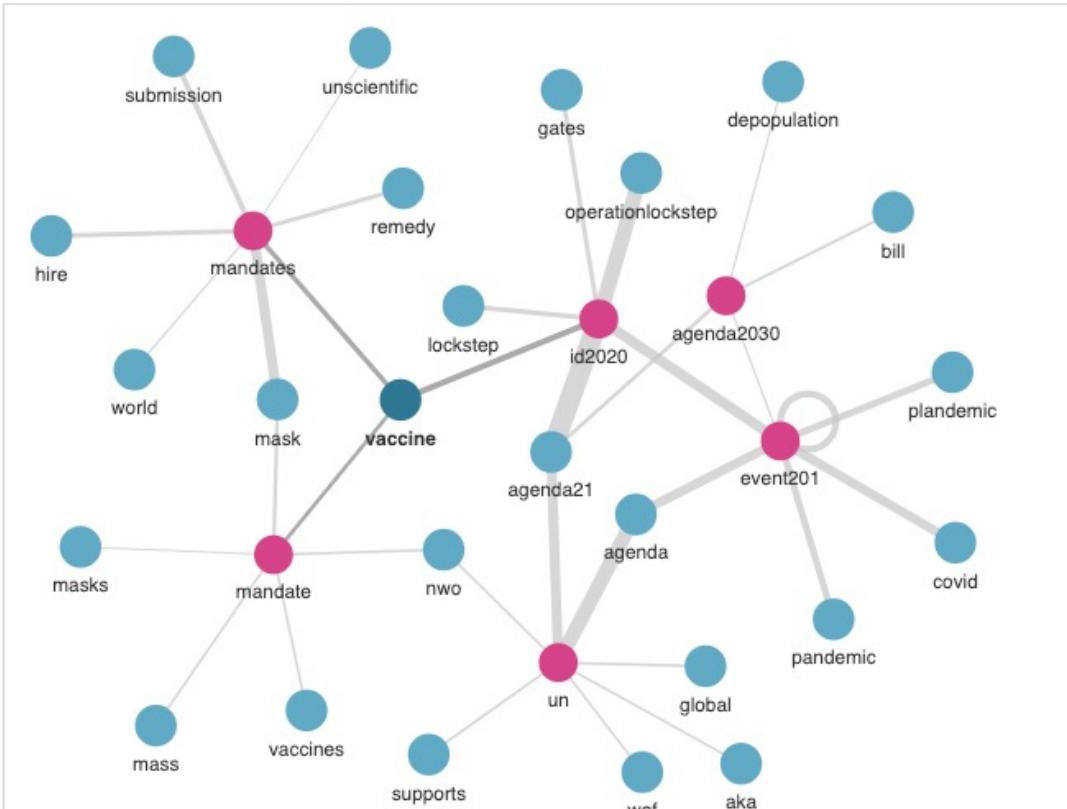
Number of collocates

7

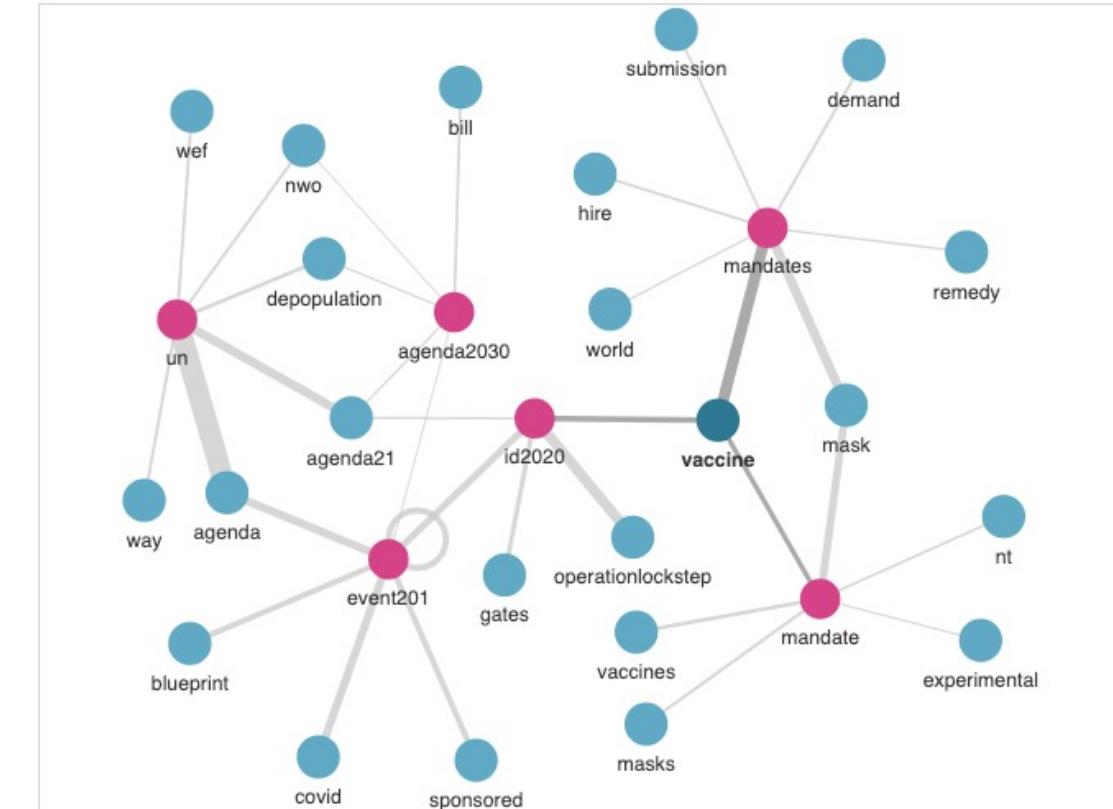
Compare to

Vaccines

Collocations in Conspiracy



Collocations in Vaccines



The Dashboard – Worldwide medical tyranny

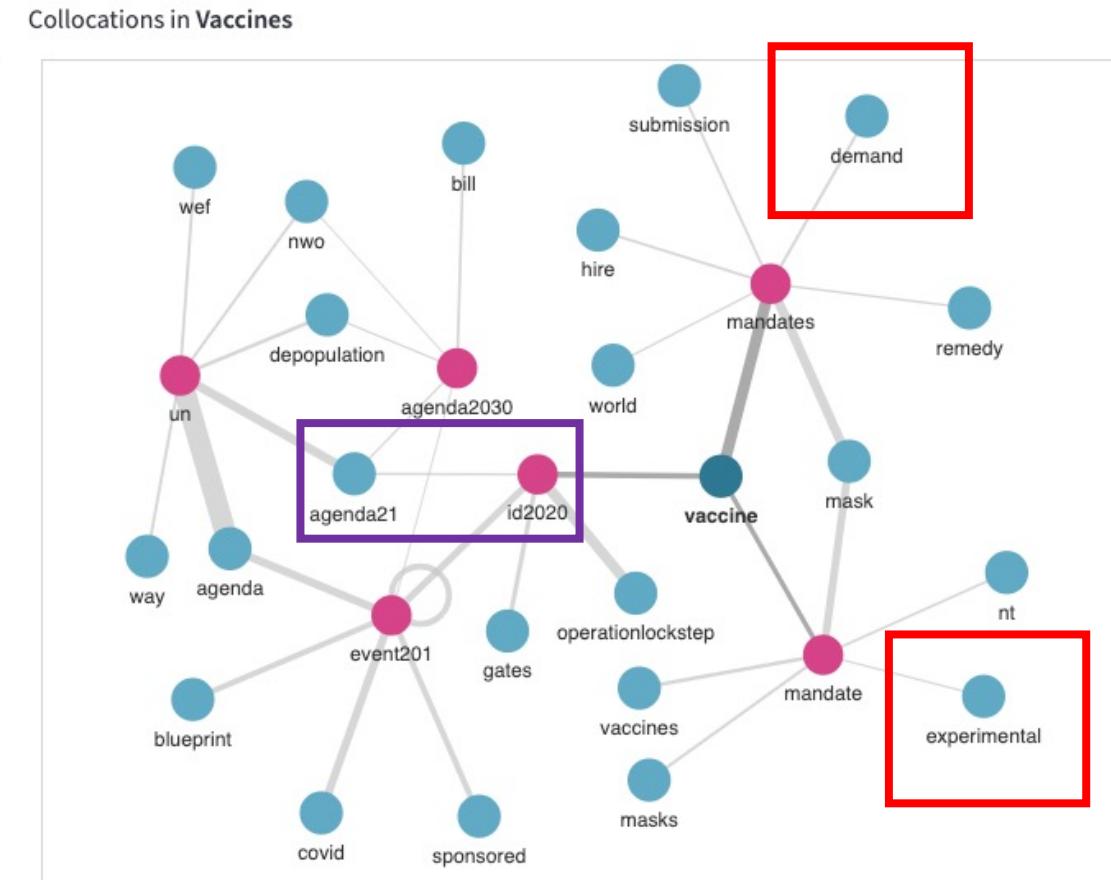
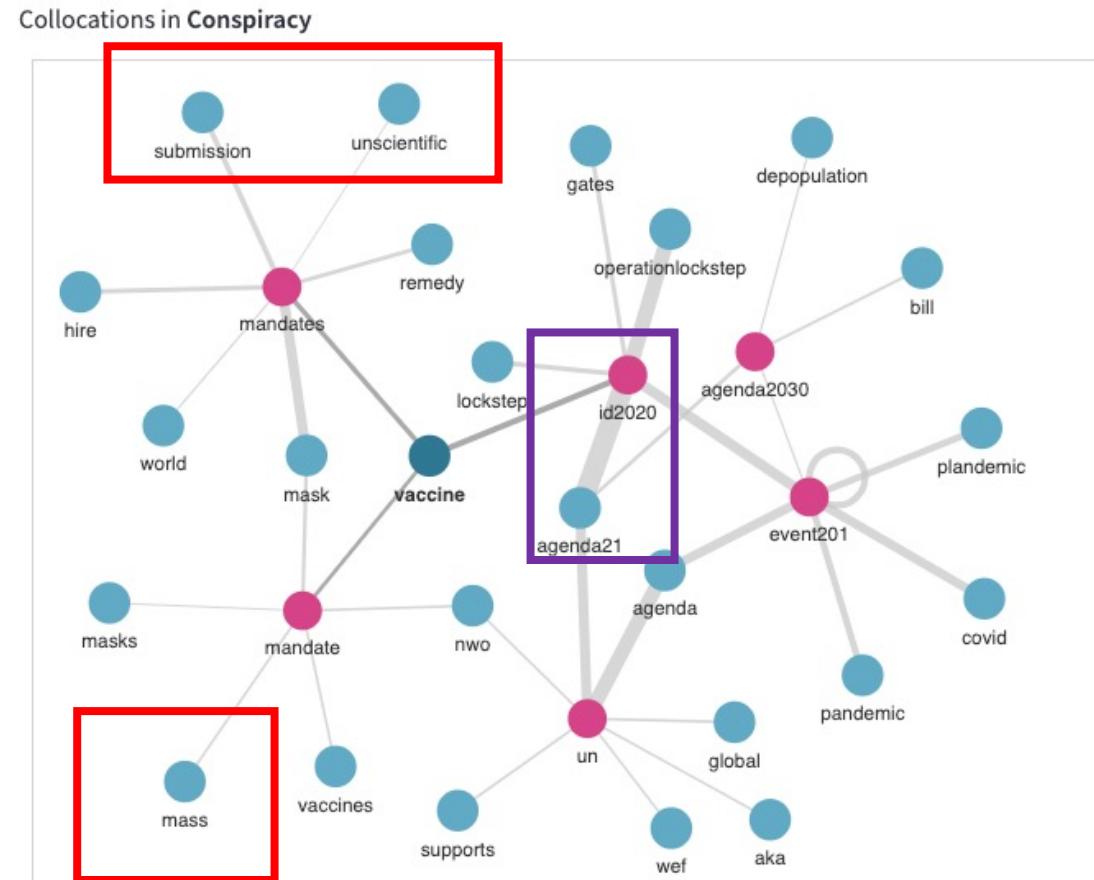
Word collocations by Frequency

Additional words to collocate

mandate **mandates**

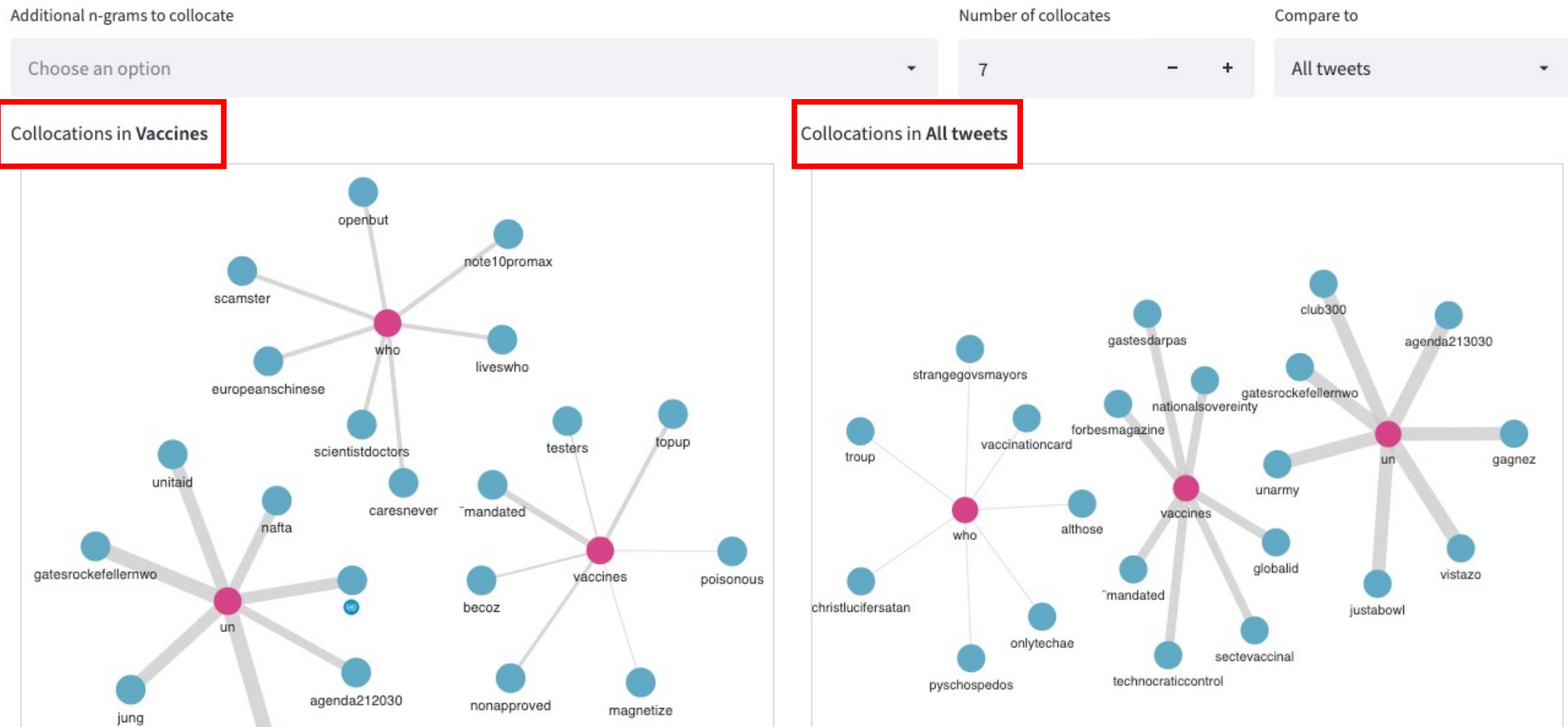
Number of collocates

Compare to



The Dashboard – Frequencies are not enough

Word collocations by Pointwise Mutual Information (PMI)



The Dashboard – Frequencies are not enough

Word collocations by Pointwise Mutual Information (PMI)

Additional n-grams to collocate

Number of collocates

Compare to

Choose an option

7

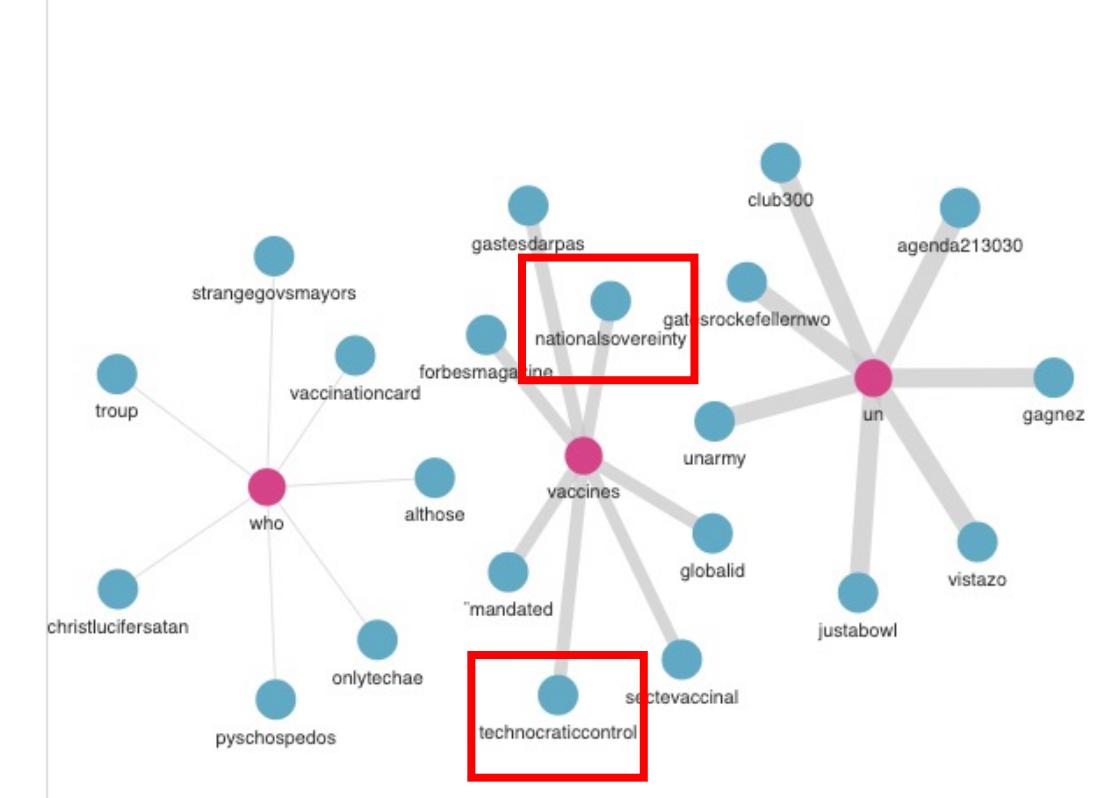
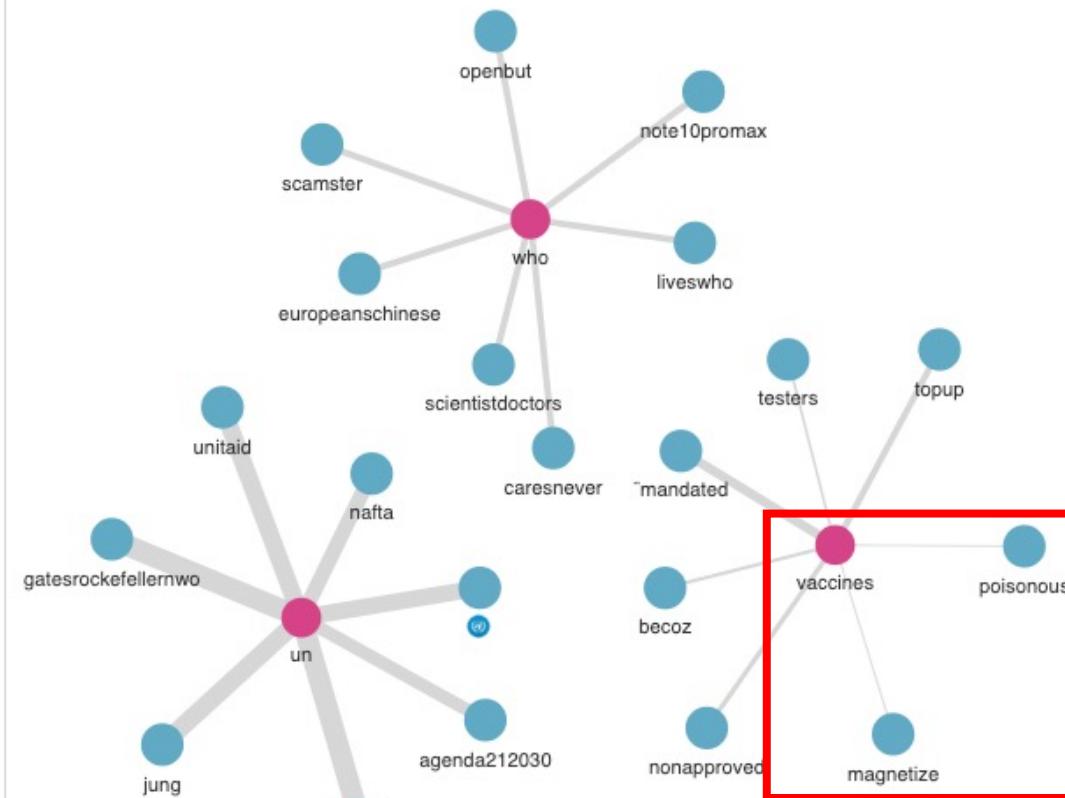
-

+

All tweets

Collocations in Vaccines

Collocations in All tweets



The end

The end

- Challenges with historical data
- Challenges with contemporary data
- Challenges with missing data
- Challenges with interpreting data
- Opportunities

The end – Challenges with historical data

- Qualitative rather than quantitative evidence
- Gaps, omissions and bias
- Representativeness and reliability
- Danger of anachronism



Jenner performing his first vaccination on [James Phipps](#), a boy of age 8. 14 May 1796,
[https://en.wikipedia.org/wiki/Edward_Jenner#/media/File:Jenner_phipps_01_\(cropped\).jpg](https://en.wikipedia.org/wiki/Edward_Jenner#/media/File:Jenner_phipps_01_(cropped).jpg)

The end – Challenges with contemporary data

- Quantitative – 35.5m tweets, 18.2m English, low resource/tiny team
- Longitudinal scale – 17+ months, 576 days – with gaps
- Biases – initial hashtags, single data source, sample tweet API

Hashtags used:

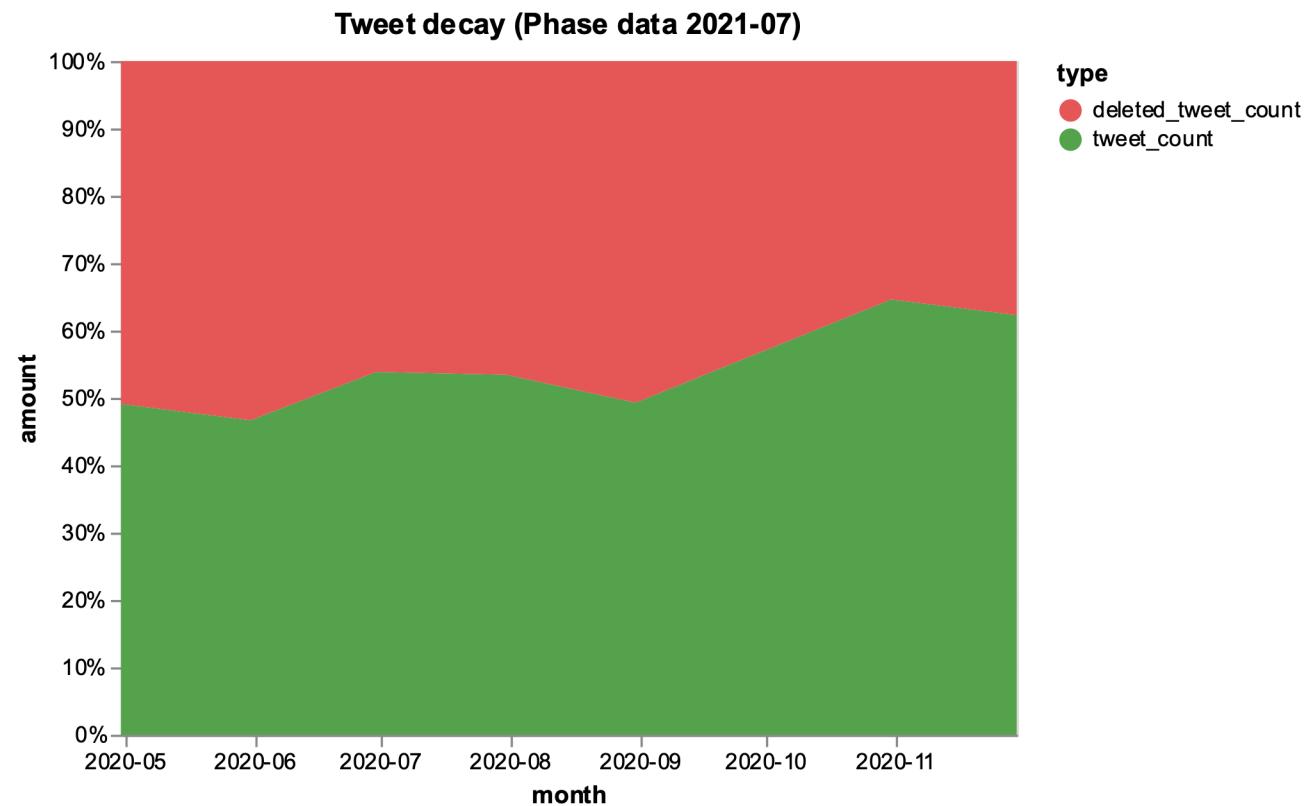
5g, AdvancedMedicine, agenda2020, agenda21, antivaccine, bioweapon, boycottqantas, coronaviruscoverup, coronabollocks, coronahoax, coronavirusliar, coronavirustruth, covidiot, covididiots, depopulation, depopulationagenda, DrRashidAButtar, endlockdown, endthelockdownuk, event201, fakepandemic, filmyourhospital, foodshortages, forcedvaccines, idonotconsent, medicalfreedom, mybodymychoice, newworldordervirus, OR vaccineskill, plandemic, plandemicdocumentary, PlandemicDocumentary, plannedemic, reopenuk, scamdemic, stayathomecowards, stop5g, stopmandatoryvaccination, syringeslaughter, TOF, tof, TruthOrFacts, vaccineagenda, vaccineinjury, vaccineskill, WeChangeTheWorld, wewillnotcomply

Hashtag omissions:

AdvancedMedicine, bioweapon, covididiots, DrRashidAButtar, plandemic, plandemicdocumentary, PlandemicDocumentary, TOF, tof, TruthOrFacts, WeChangeTheWorld, antivaccine, boycottqantas, idonotconsent, medicalf\$reedom, mybodymychoice, stop5g, stopmandatoryvaccination, syringeslaughter, vaccineinjury, vaccineskill

The end – Challenges with missing data

- Reproducibility
 - 40-50% tweets in 2020 are already inaccessible



The end – Challenges with interpreting data

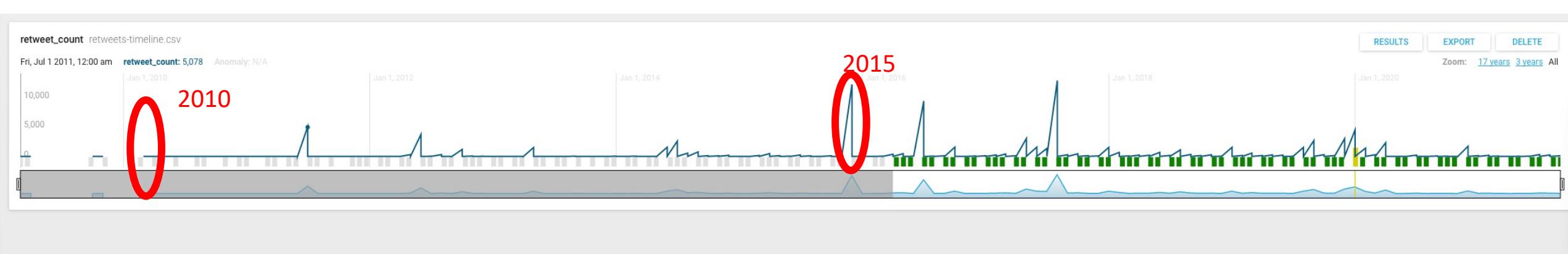
- What is a meaningful semantic shift for our research questions?

2010-03, 1 , 39% OFF \$15.05 Griffin iClear Sketch for iPod Nano 5G (Camo Black)

<http://www.haggleblast.com/home/deal/XXXXXX>

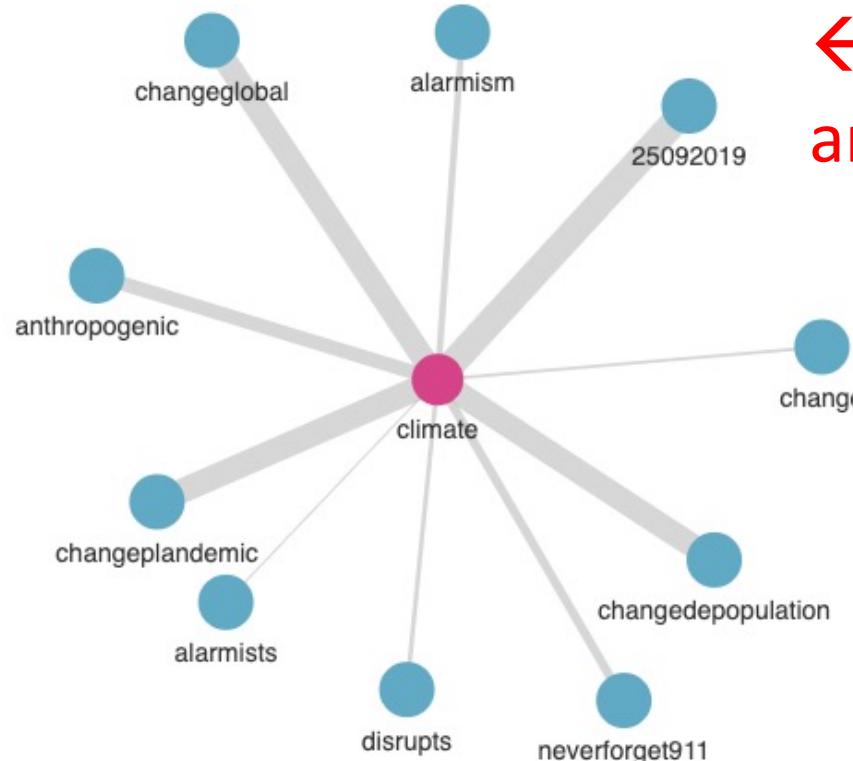
2015-12, 12336, "琥珀糖のつくりかた 水200ccに粉寒天5g入れて沸かして溶けたら砂糖300グラム入れて糸が引くくらいまで中火で煮てリキュールや食紅で作った色水を入れた型に入れて混ぜて冷蔵庫1時間後くらいに切り分けて4日間陰干しで完成！"

- "How to make amber sugar. Put 5g of powdered agar in 200cc of water, boil it and melt it, then add 300g of sugar and simmer on medium heat until the thread is pulled. Cut into pieces and dry in the shade for 4 days to complete!"



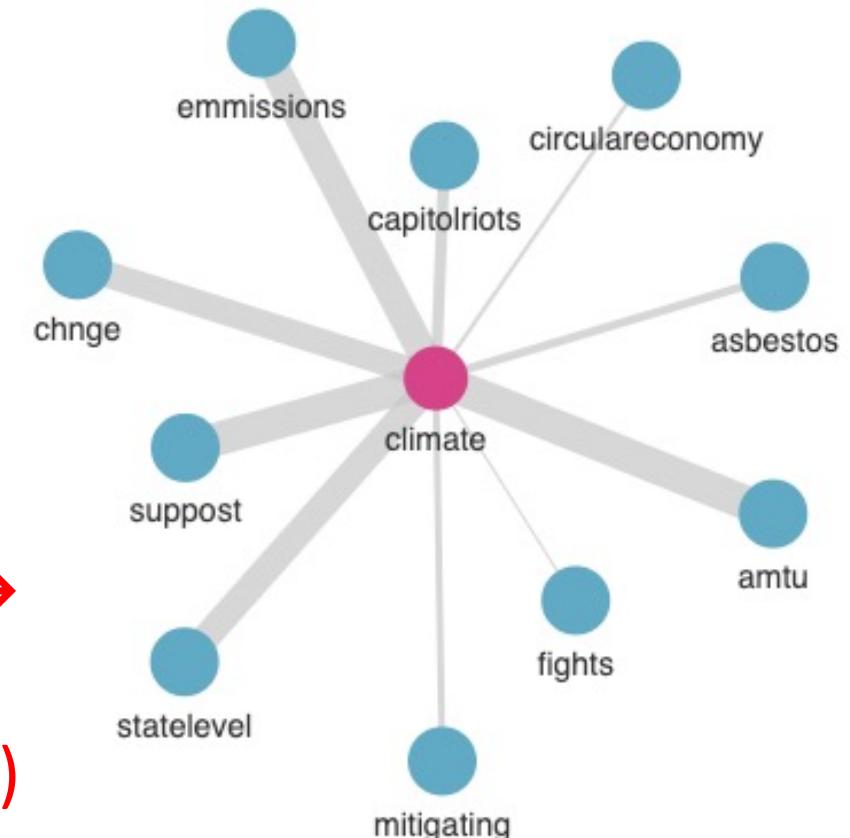
The end – Opportunities

'Climate' change – Anthropogenic environmental phenomenon or Diachronic psycholinguistic signal?



← 2020, alarmism
and disruption

2021, riots, fights →
(and the Asbestos
coverup/conspiracy)



@MichaelSutcli15 @birdonthewire3 @stop1984 @KailashChandOBE
@ChristineJameis @BorisJohnson Thank God that the tories are in power and can crack on with global depopulation to cure the global warming issue, that's the science that is now eventually being listened to.