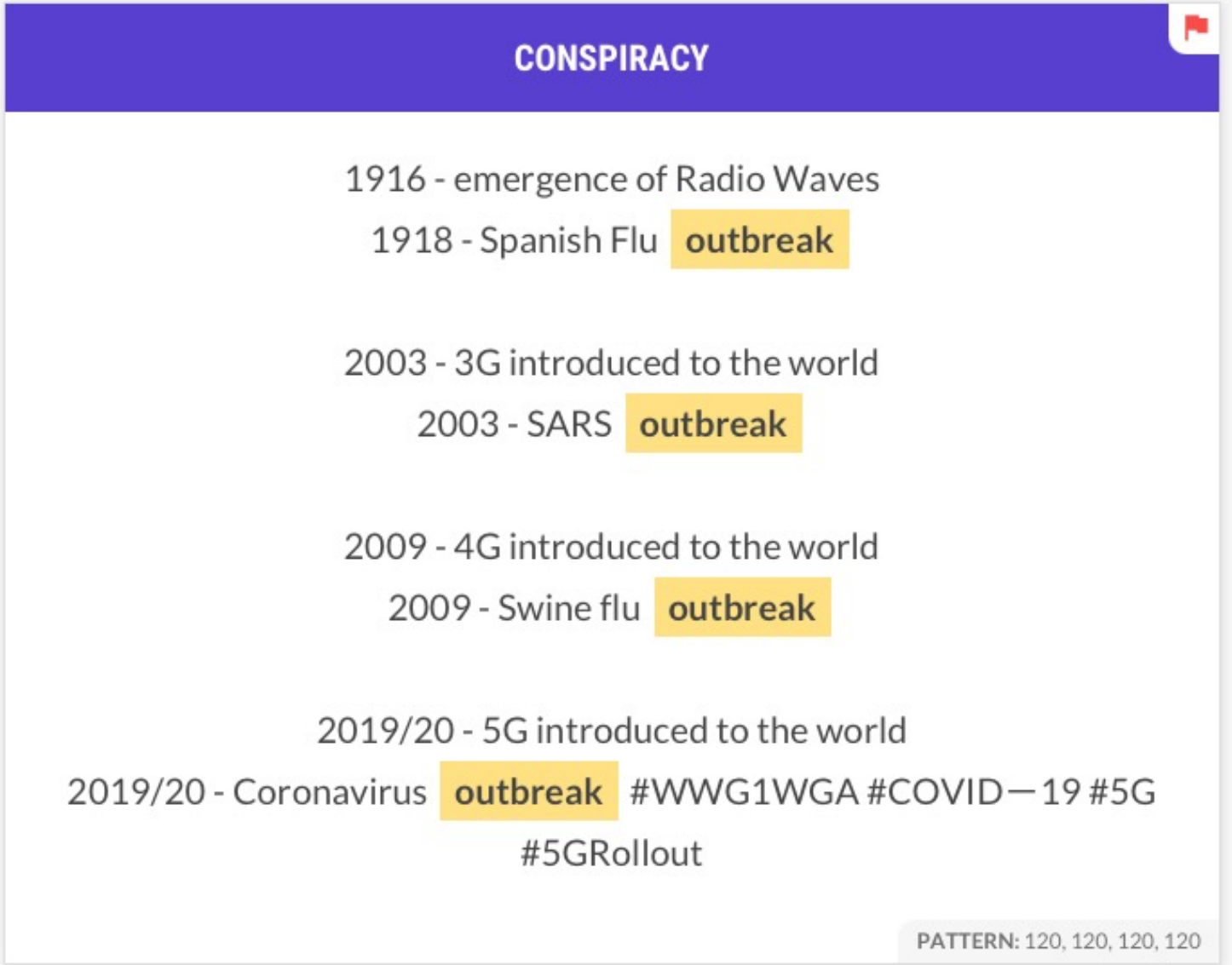
Covid Rumours in Historical Context
Data and digital methods

Marty Steer (and the research team)
https://historyandrumour.blogs.sas.ac.uk

DHRH Seminar, 2022-03-21

Outline

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



Example tweet implying historical colocation is causation

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

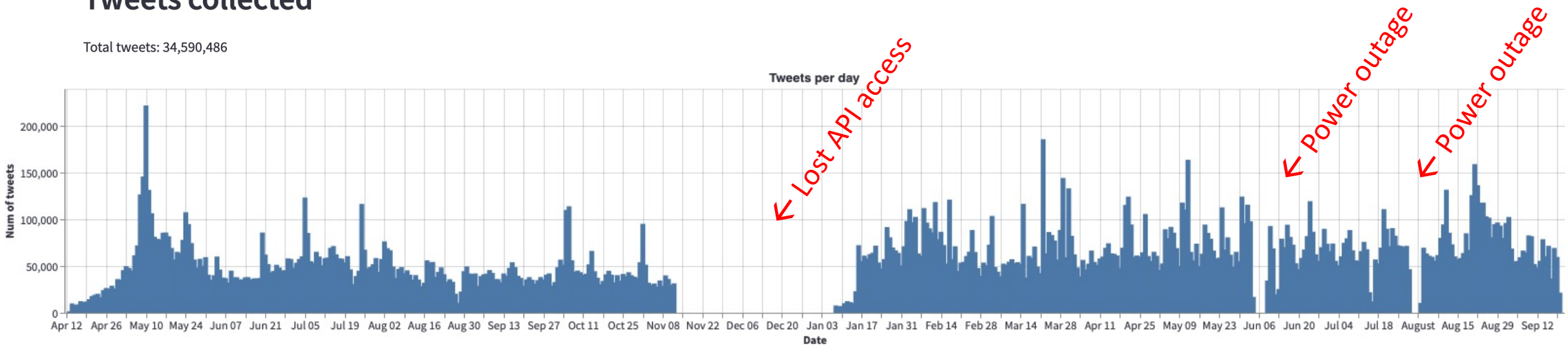
The Data

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

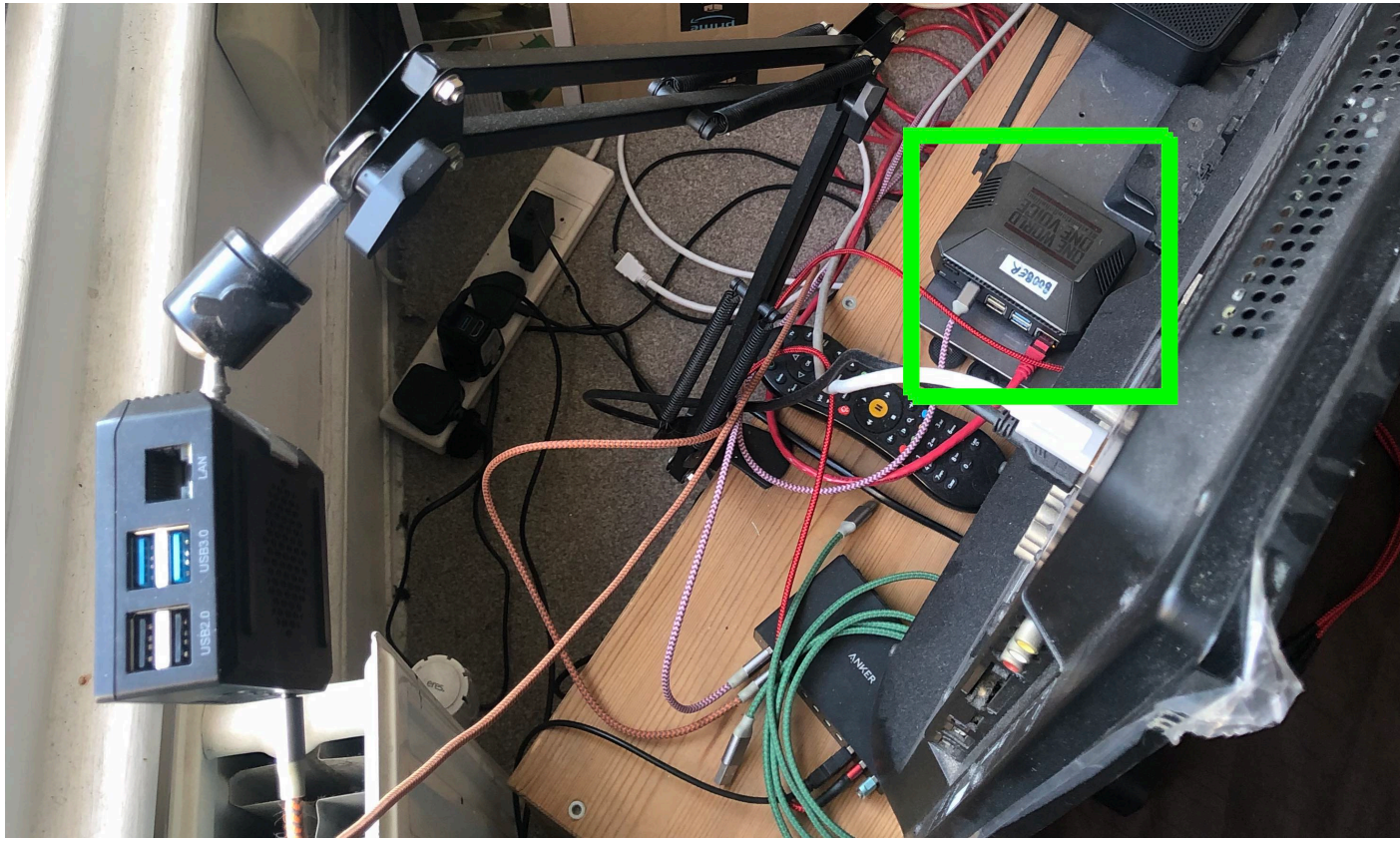
- Collection method
  - Twarc – Documenting the now
  - Crontab nightly rotation
  - JSONL text files
  - Rsync across local network
  - Raspberry Pi 4

rona-rumours-corpus-α-2020-11-08.jsonl.gz
rona-rumours-corpus-α-2020-11-09.jsonl.gz
rona-rumours-corpus-α-2020-11-10.jsonl.gz
rona-rumours-corpus-α-2020-11-11.jsonl.gz
rona-rumours-corpus-α-2020-11-12.jsonl.gz
rona-rumours-corpus-β-2021-01-08.jsonl.gz
rona-rumours-corpus-β-2021-01-09.jsonl.gz
rona-rumours-corpus-β-2021-01-10.jsonl.gz
rona-rumours-corpus-β-2021-01-11.jsonl.gz
rona-rumours-corpus-β-2021-01-12.jsonl.gz
rona-rumours-corpus-β-2021-01-13.jsonl.gz
rona-rumours-corpus-β-2021-01-14.jsonl.gz
rona-rumours-corpus-β-2021-01-15.jsonl.gz
rona-rumours-corpus-β-2021-01-16.jsonl.gz

The Data

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

Boober

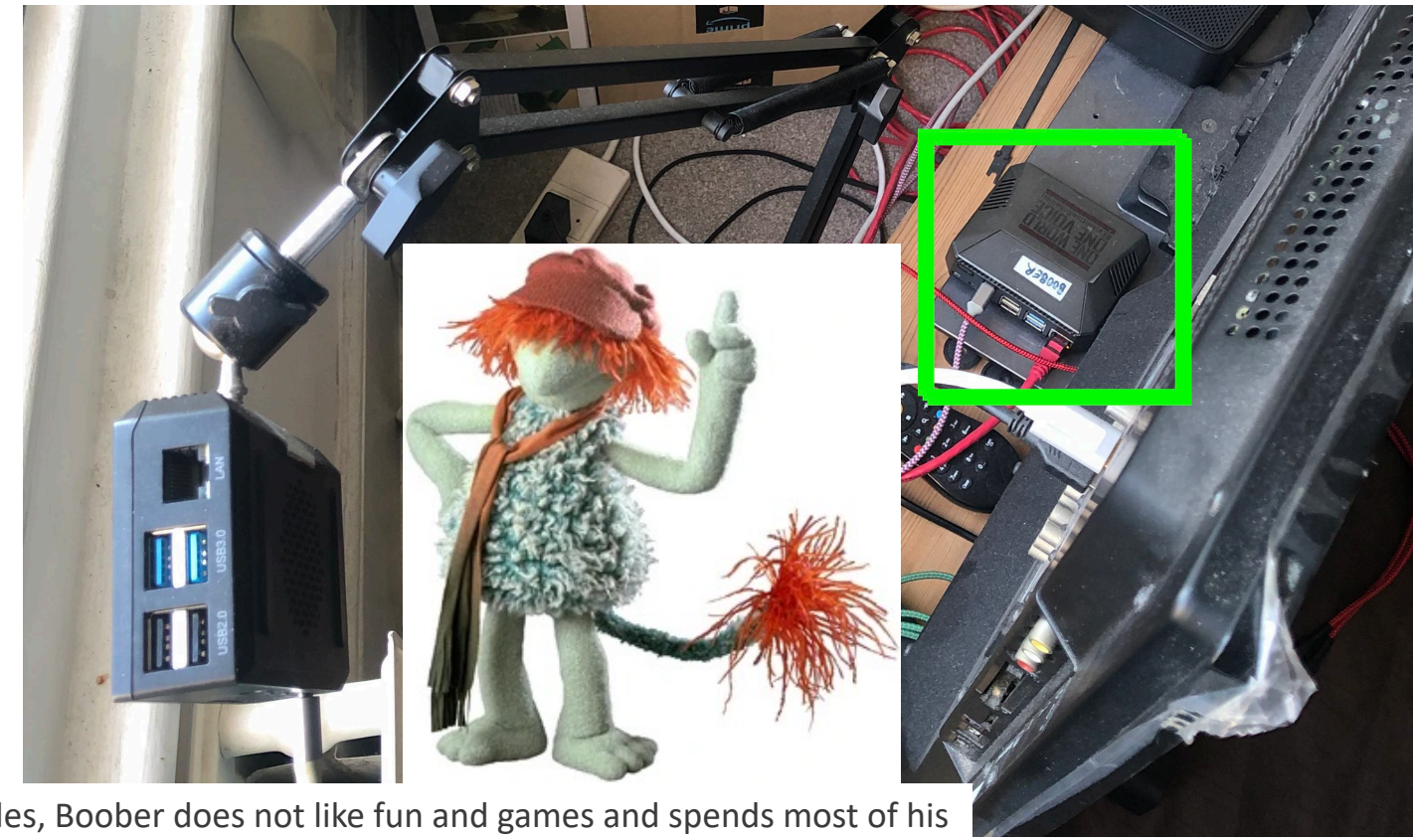


The Data

- 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



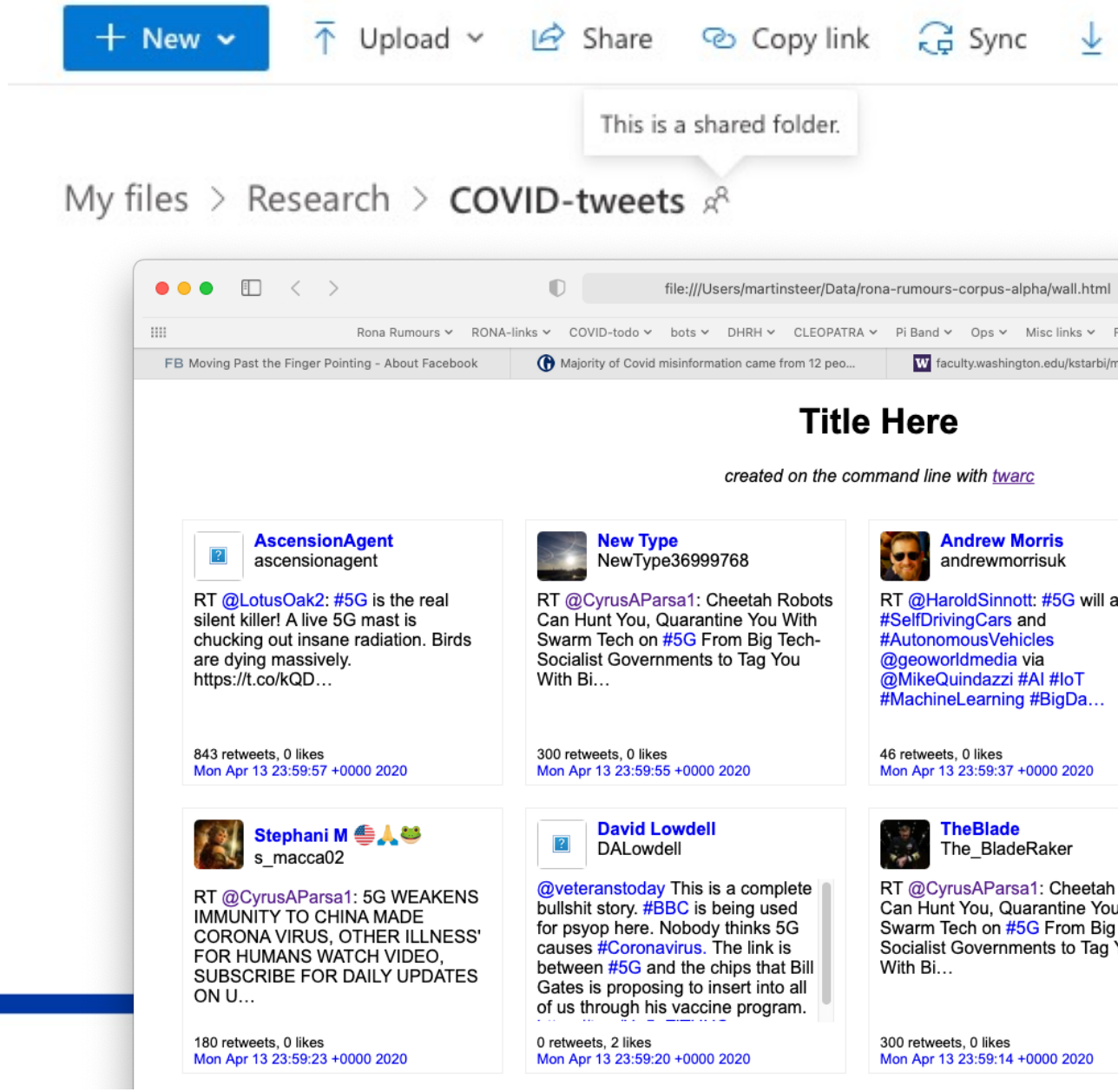
Unlike other Fraggles, Boober does not like fun and games and spends most of his time worrying about doom and disease. When he's not worrying about himself, he's busy warning others. He is easily frightened and suffers from a variety of phobias... Because of his fears, he is very knowledgeable about 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

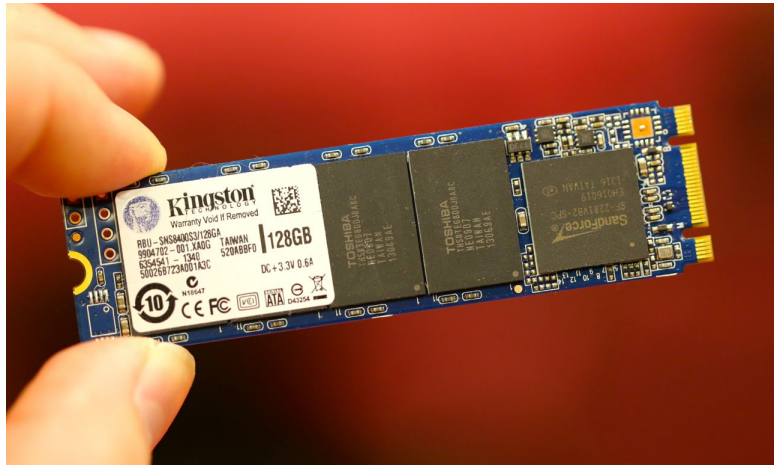
The Data

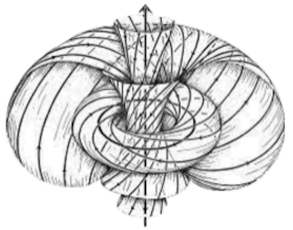
- Data management
  - Azure blob storage
    - Secure enclave
  - OneDrive – Confidential sharing
  - DropBox – General documents
  - Backups
    - 2TB USB-C M.2 SSD – Twistor
    - 6TB LAN NAS – Marjory



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



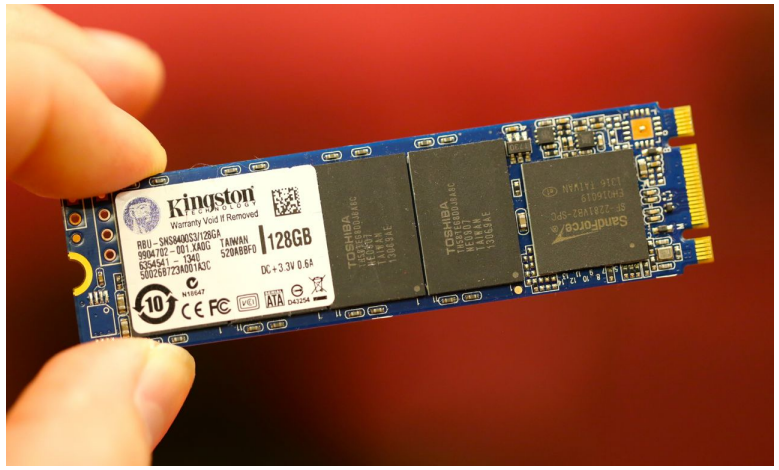


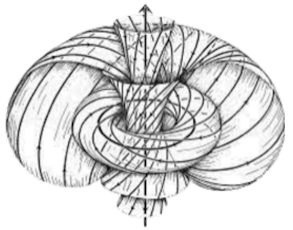
Twistor

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





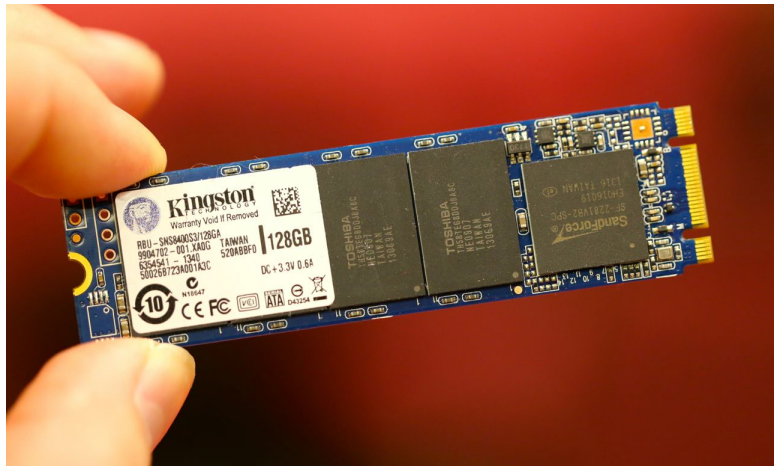
Twistor

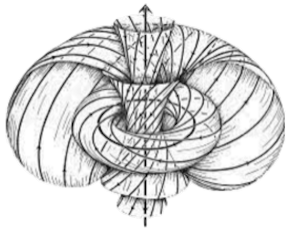
```
(base) ➜  covid-rumours git:(master) dvc pull --remote marjory
  7% Transferring|                                |39/577 [00:43<15:08,  1.69s/fil
 57%|                |1023aa643eb87a47fc92c0ecc83.84M/6.80M [00:12<00:04,     662kB/
 99%|                |42ce021abefd9affcc2b4d37676.56M/6.63M [00:12<00:00,     582kB/
 87%|                |de268fea2ce3f4ea3bce81813a4.31M/4.93M [00:12<00:00,     666kB/
100%|                |4d254a47ea74d9e7e3c4407ca94.15M/4.15M [00:30<00:00,     669kB/
  0%|                |a6df1ba7099526c1b1e82e5194ffc6     0.00/?  [00:00<?,       ?B/
  0%|                |29c815c4b4e71b68911f482693ac0b     0.00/?  [00:00<?,       ?B/
  0%|                |b5be857b0596640e11ed5df663e87a     0.00/?  [00:00<?,       ?B/
 90%|                |581c796de2070fc78ec0b4d49d6.09M/6.75M [00:10<00:01,     668kB/
 85%|                |072ef7ecb2190a9036c721a7bf5.67M/6.69M [00:09<00:01,     667kB/
```

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 0 0 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 0 0 0 Updated 4 days ago

covid-rumours-dashboard Private

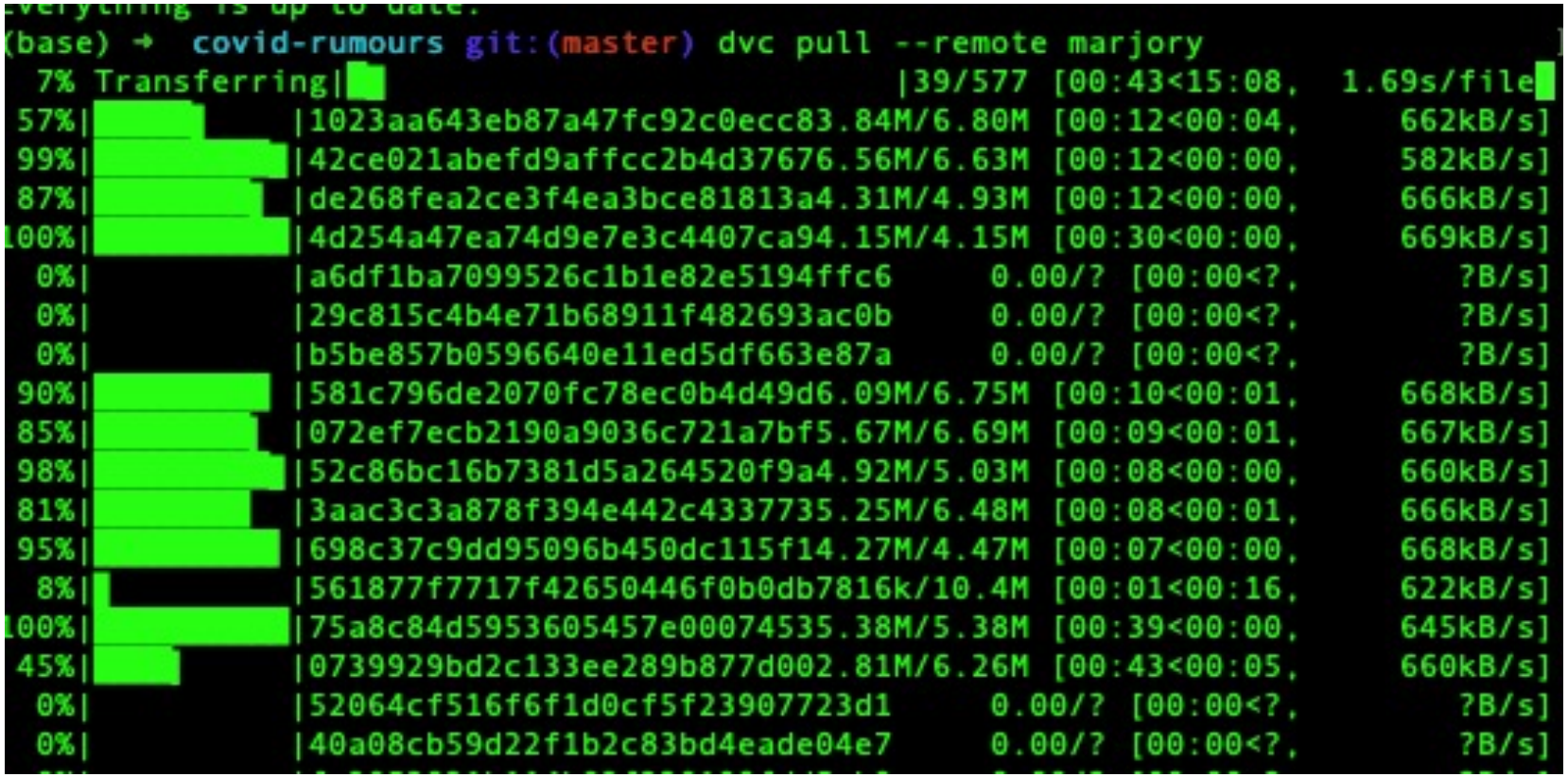
Streamlit dashboard to explore diachronic linguistics and semantic fields within the covid-rumours-data corpus

☆ 0 0 0 0 Updated 9 days ago

The Data

Data Version Control

- Version control
  - Git
  - DVC
  - Makefile



The Data

- Version control
  - Git
  - DVC
  - Makefile


```
# ---
# tweets
# useful metadata fields from all tweets
T_CHUNKS_DIR = tweets
$(BUILD_DIR)/$(T_CHUNKS_DIR)/%.csv: $(DATA_DIR)/%.jsonl.gz
	mkdir -p $(@D)
	gzcat $^ | 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:
[redacted]
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

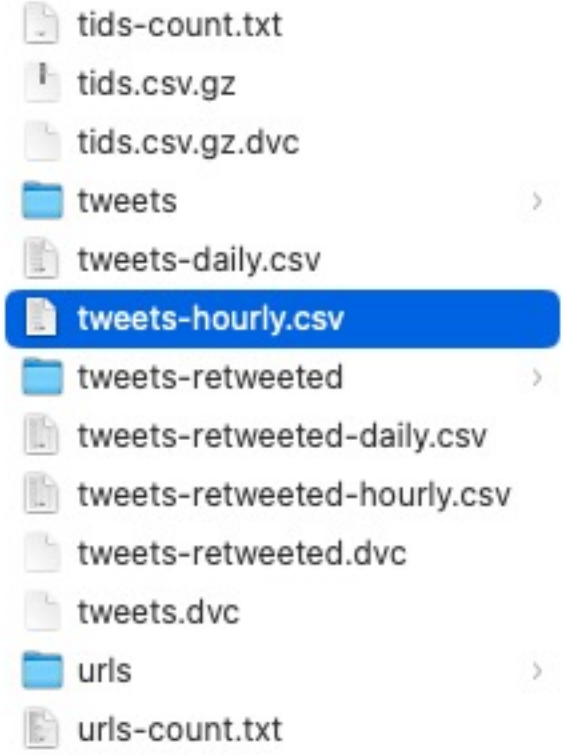
1 contributor

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' inappropriat
  - Secure enclave
    - No sensitive data stored in code
  - Derive Pseudonymous data



The Data

- What about the Content itself?

The Data

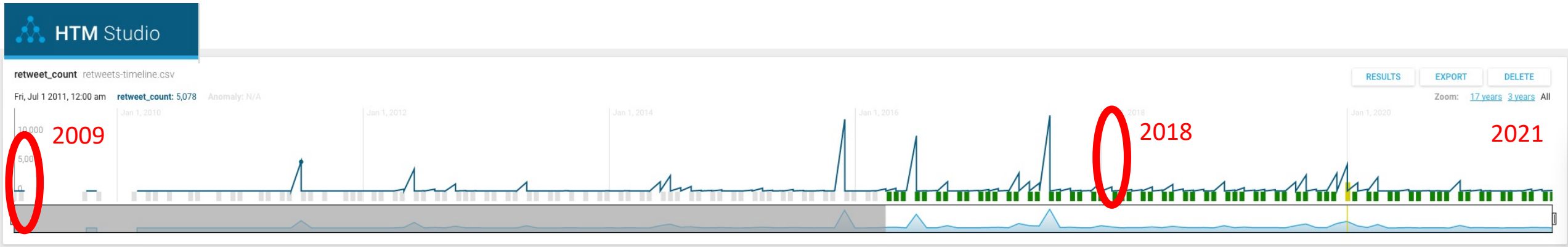
Content – Retweets, timeseries by original tweet date

- 2009-10,1,IMMUNOCONTRACEPTIVE HIDDEN IN THE FLU VACCINE http://bit.ly/XXXXXv #nwo #h1n1 #swineflu #depopulation nssm200

← Vaccine sterilizations (2009)

- 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.

← 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

./jq

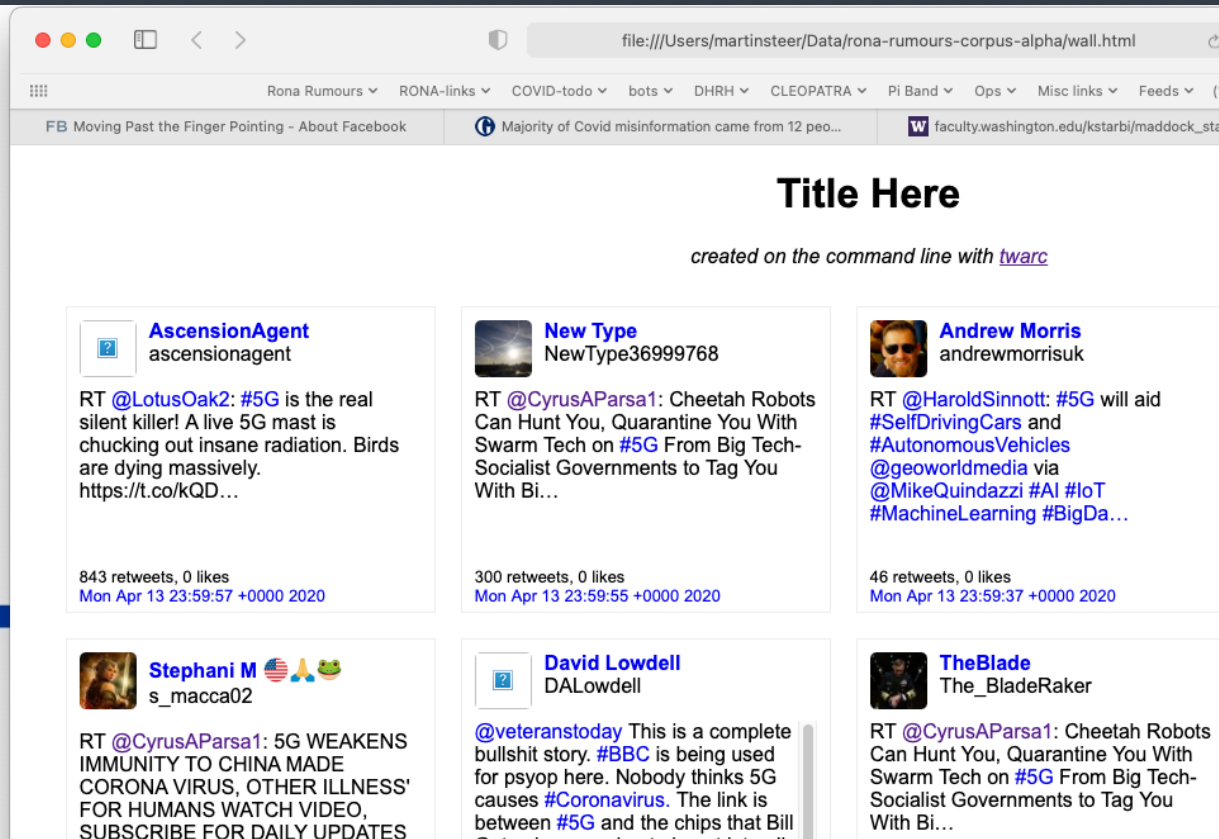
- Words, users, URLs, hashtags, languages
- Sentiment analysis
- User network analyses
- Hashtag/Term cluster analyses
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

```
# ---
# tweets
# useful metadata fields from all tweets
T_CHUNKS_DIR = tweets
$(BUILD_DIR)/$(T_CHUNKS_DIR)/%.csv: $(DATA_DIR)/%.jsonl.gz
	mkdir -p $(@D)
	gzcat $^ | 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

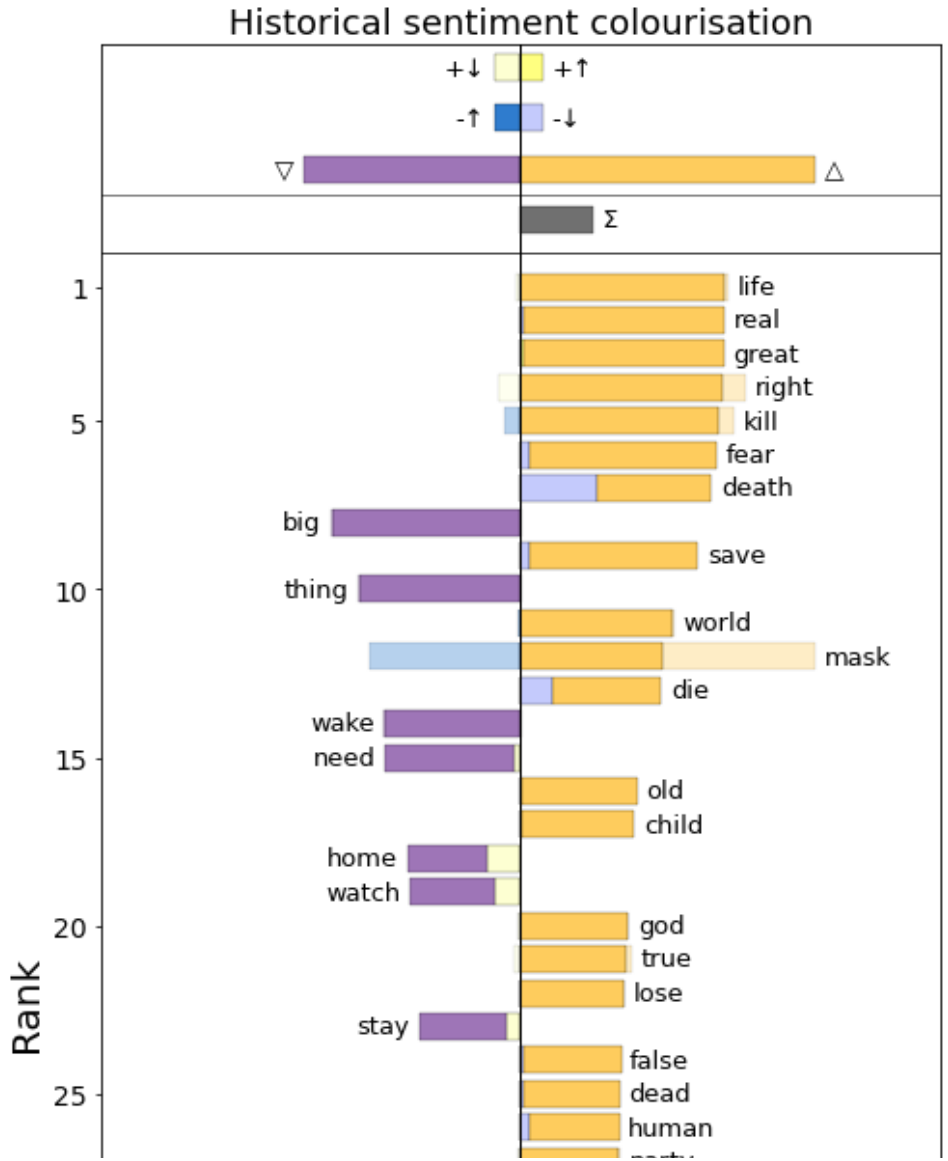
```
$(MODEL_DIR)/grabbag/walls/%.html: $(MODEL_DIR)/grabbag/sample-data/%.jsonl.gz
    [[ -d $(@D) ]] || mkdir -p $(@D)
    gzcat $^ | $(SCRIPT_DIR)/twarc-wall.py --title $(@F) > $@
```

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



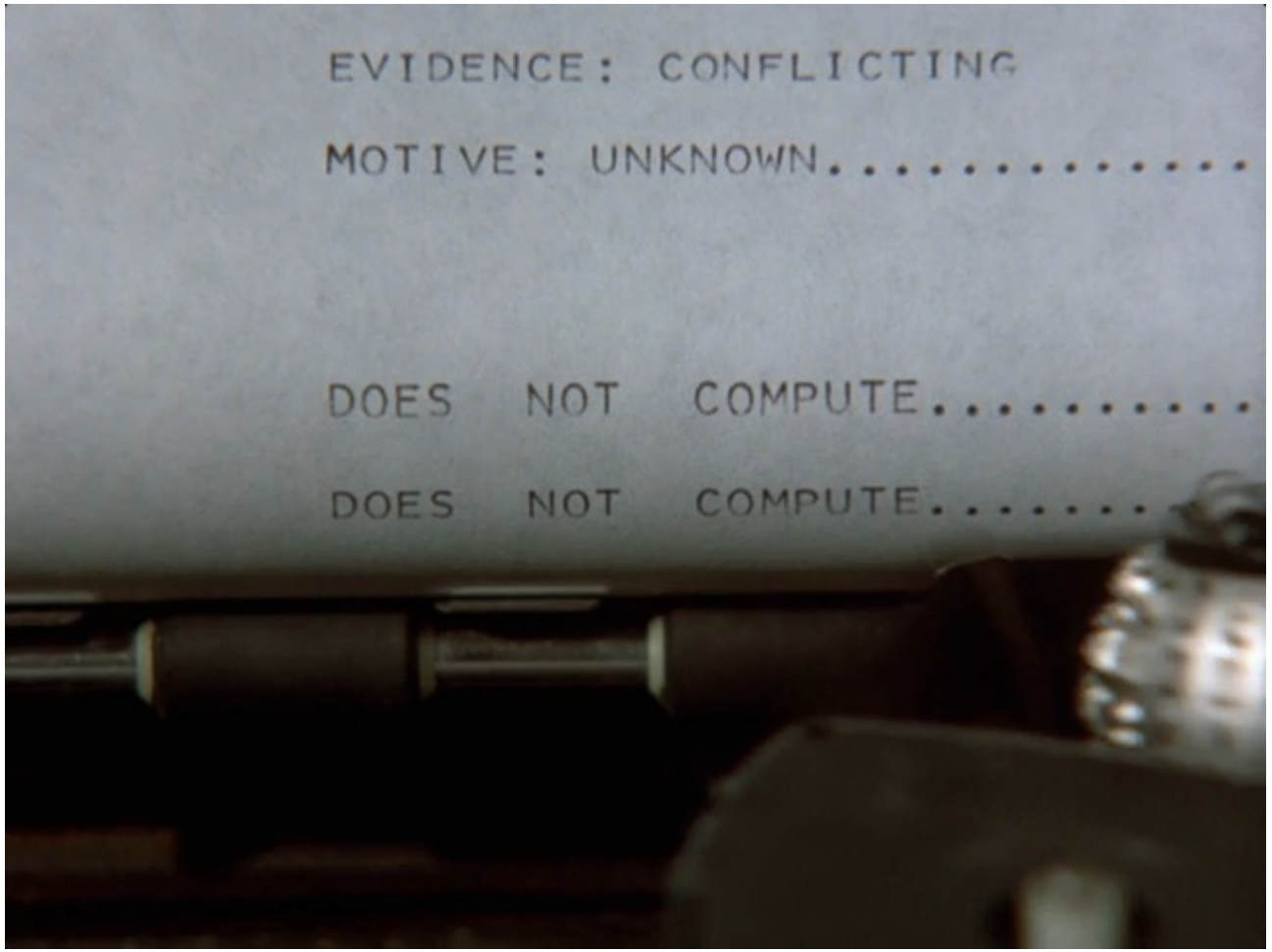
The Analyses

- 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



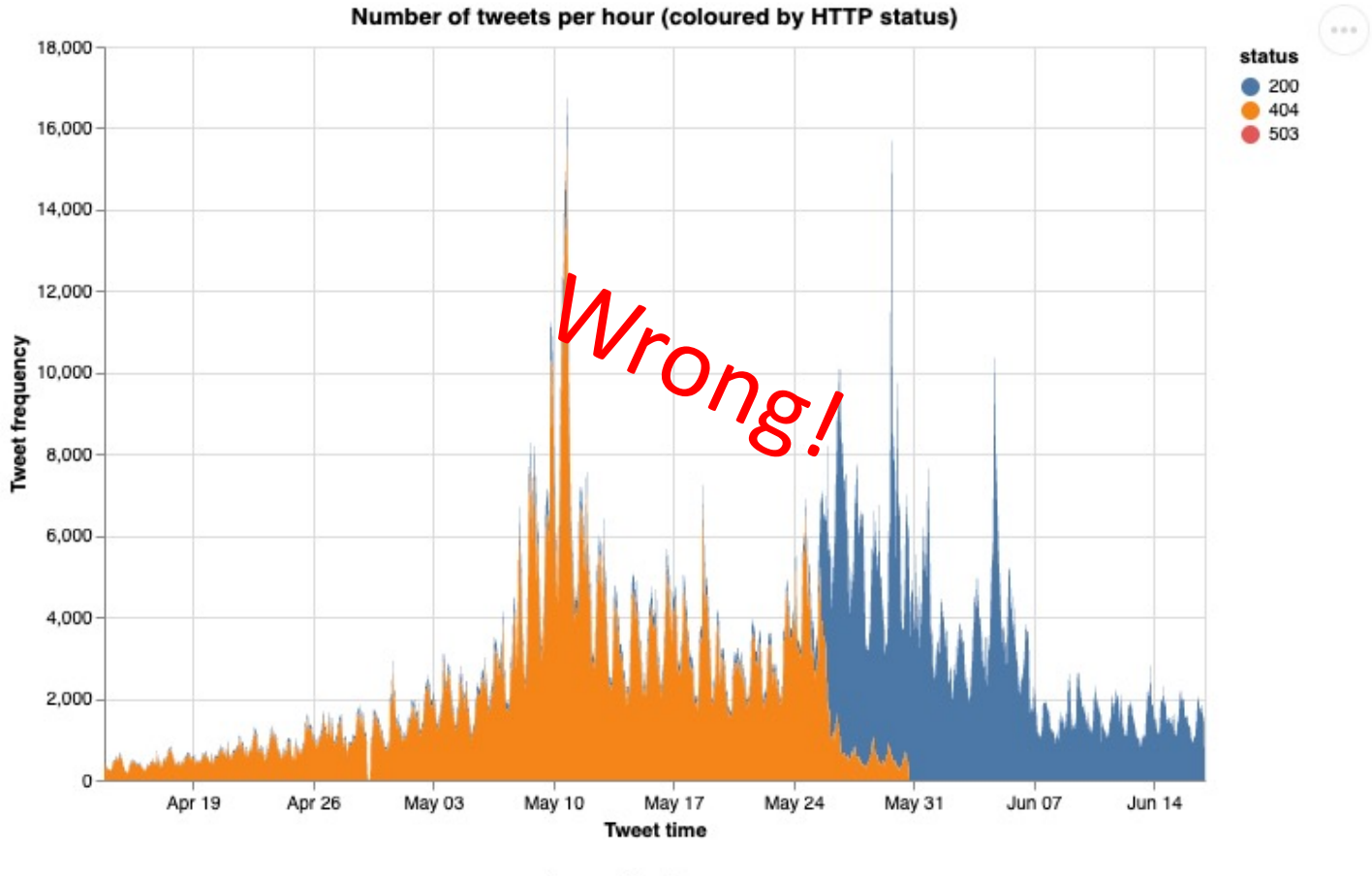
The Analyses

- 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!

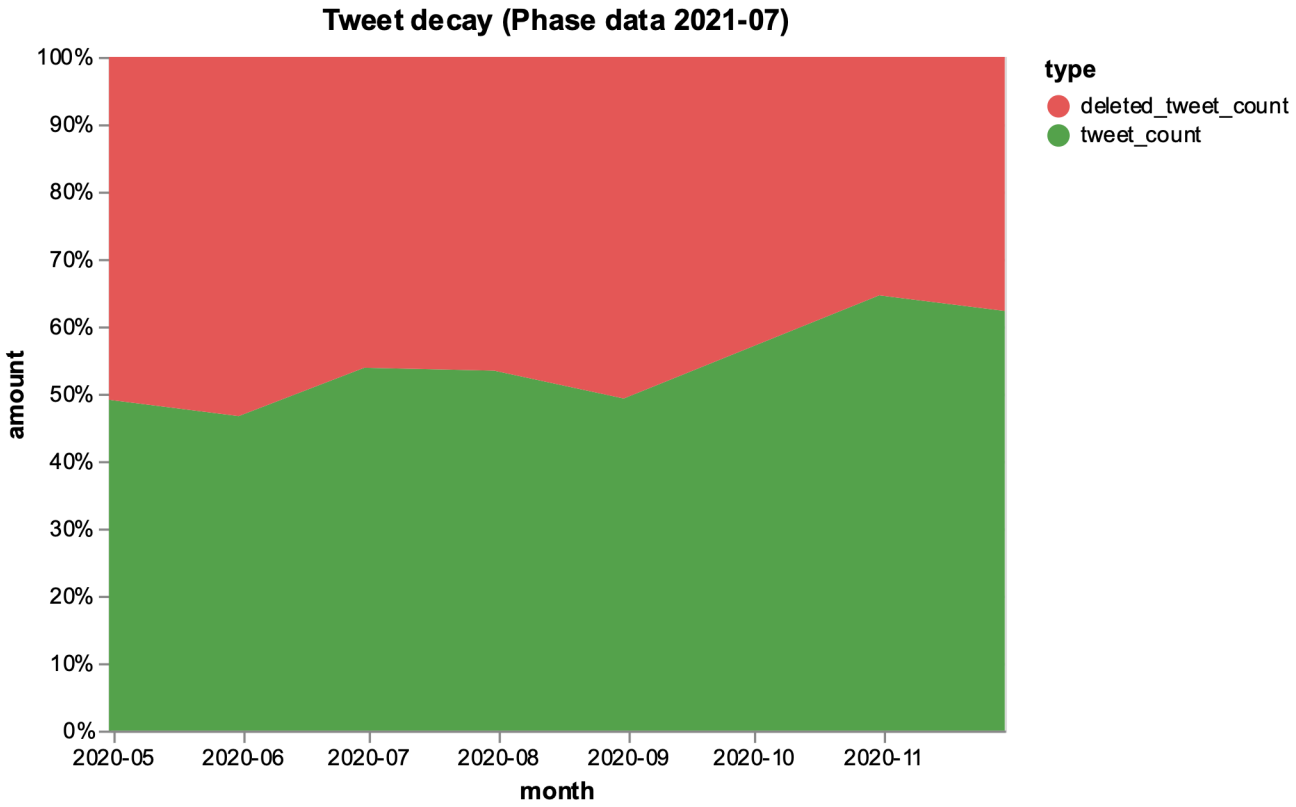
Problem: 404 errors due to jq's LONG INT truncating tweet id's to end in 00 e.g

1249849839975686145	→	1249849839975686**00**
1249849831100538880		1249849831100538**00**



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



MINET Added tweet attrition feature for us

https://github.com/medialab/minet

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

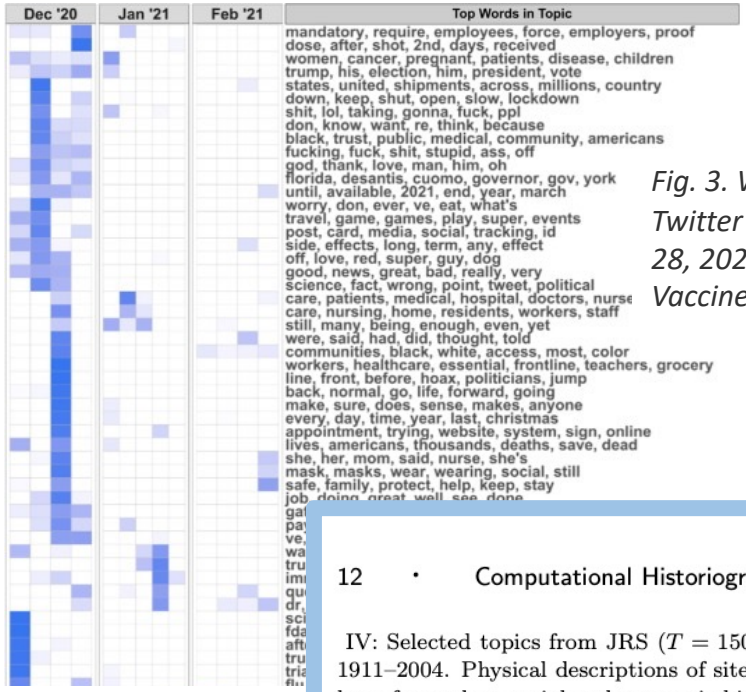
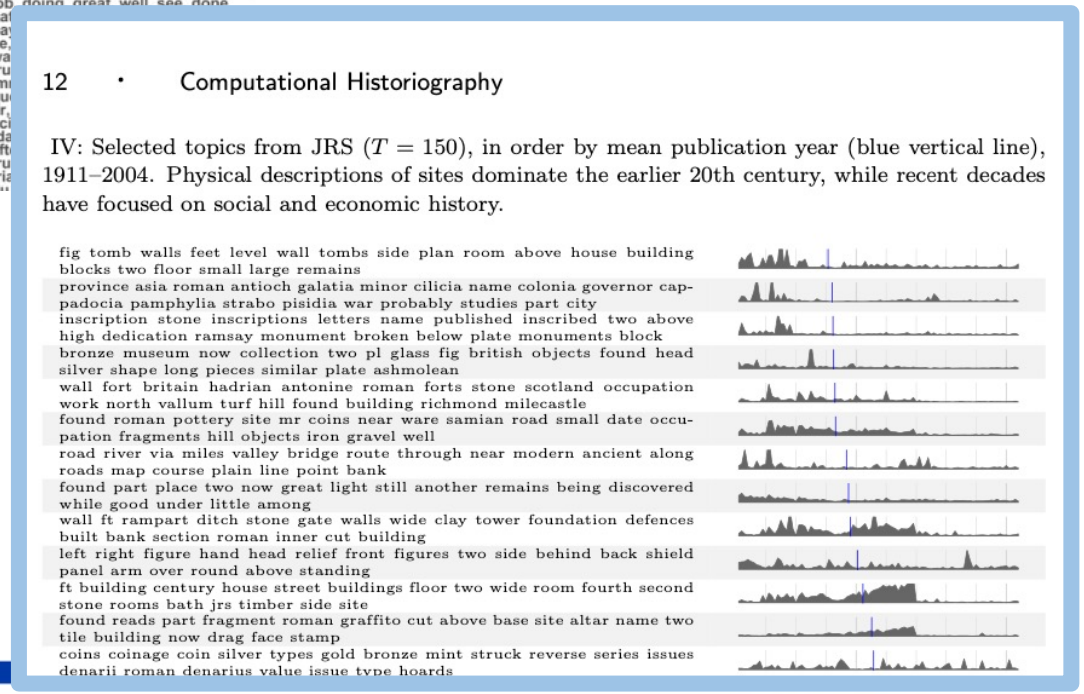


Fig. 3. Weekly variation of data-driven COVID-19 Twitter topics from December 1, 2020 to February 28, 2021. Guntuku, Buttenheim, Sherman, Merchant, Vaccine, Volume 39, Issue 30, 2021



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 cappadocia 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 mr coins near ware samian road small date occupation 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 graffito 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

David Mimno. 2012. Computational historiography, https://doi.org/10.1145/2160165.2160168, http://www.perseus.tufts.edu/publications/02-jocch-mimno.pdf

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)

https://mimno.github.io/Mallet/index

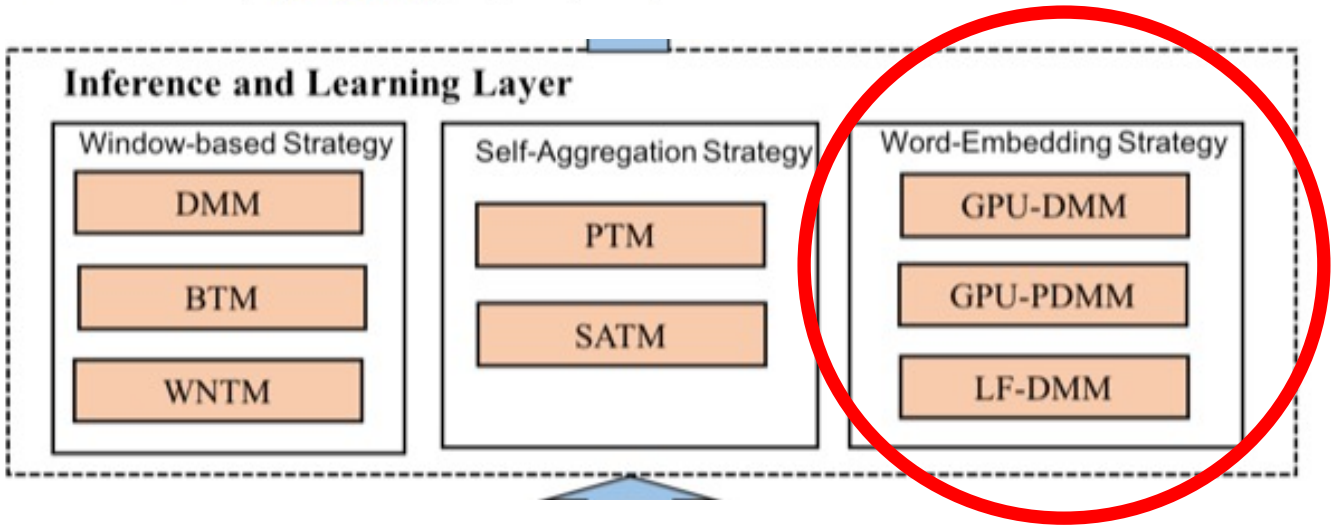
The Analyses

- 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

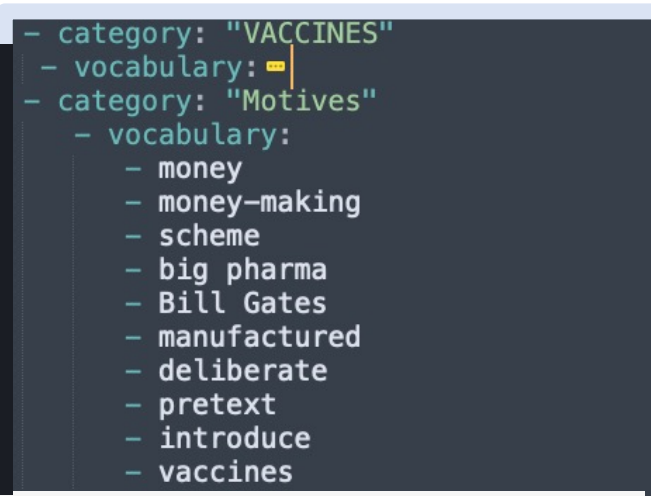
Top topic words

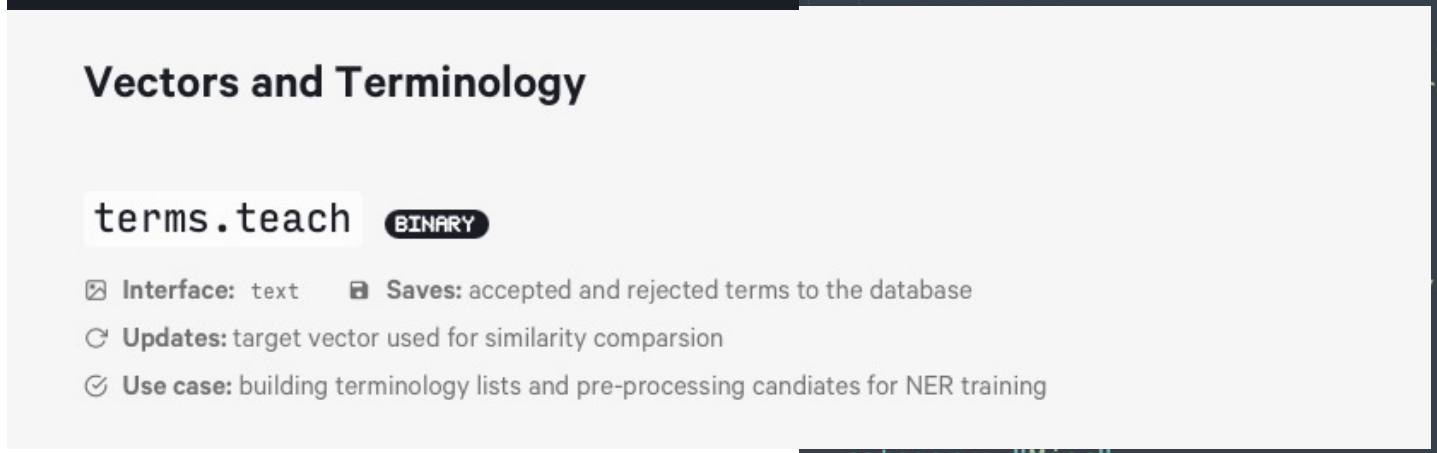
LFDMM (10.18 hours)	GPU-PDMM (50.25 hours)
868 covid	2197 5g
583 depopulation	1416 new
569 new	1304 covid
548 vaccine	1160 scamdemic
544 agenda21	1120 covid19
531 5g	965 covidiot
528 time	919 like
524 ve	902 world
521 world	894 plandemic
520 good	874 people

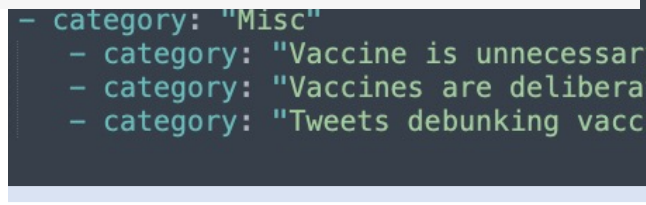
Used topics to get a macro view of the English corpus
Wasn't able to use historical word embeddings successfully
Tweets hashtag themselves as a form of topicality
so this analysis wasn't much more useful than word frequencies

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





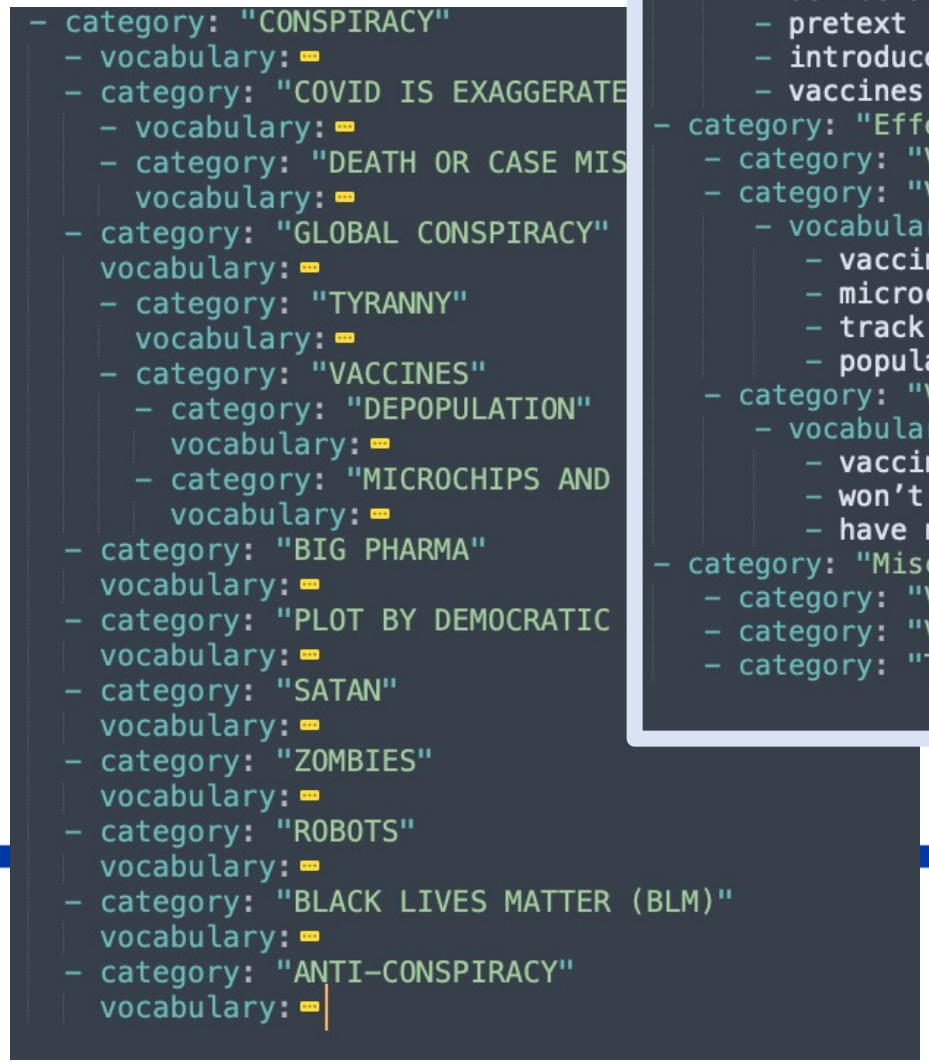


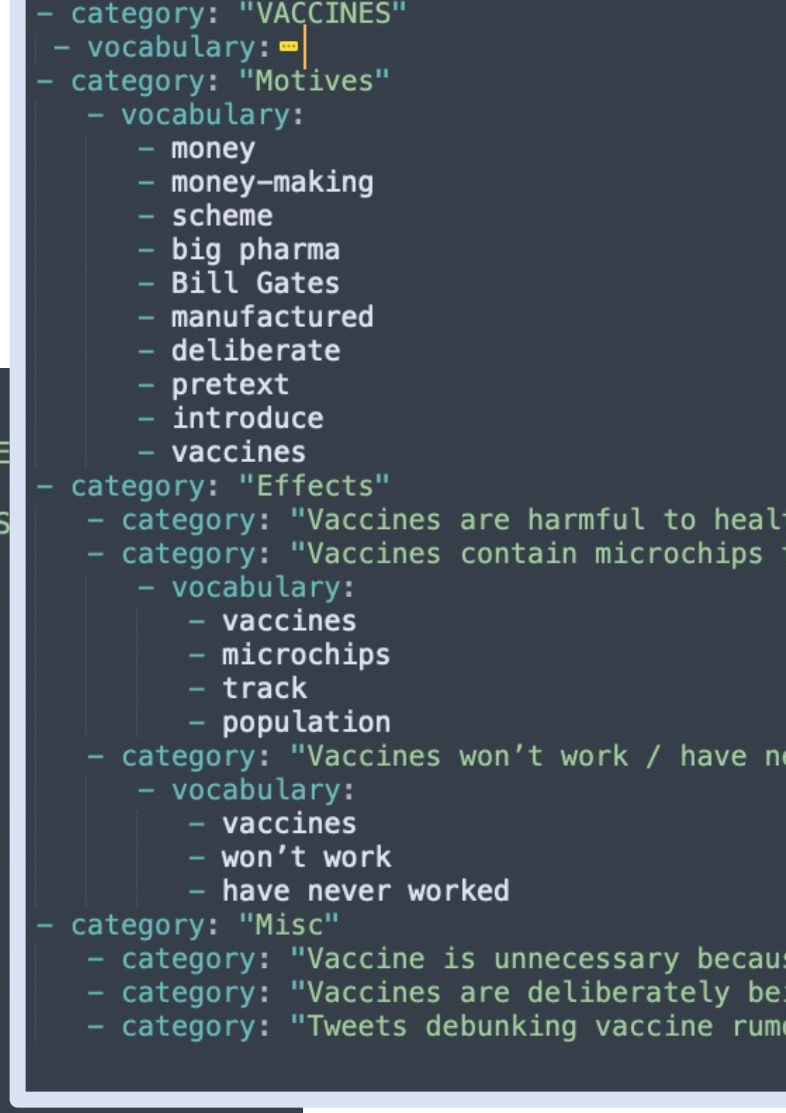
- Build terminology lists
- Refine conceptual boundaries of the categories
- Used to mine corpus for biased (relevant) training data

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

- CONSPIRACY
- ORIGNS
- CURES
- VACCINES

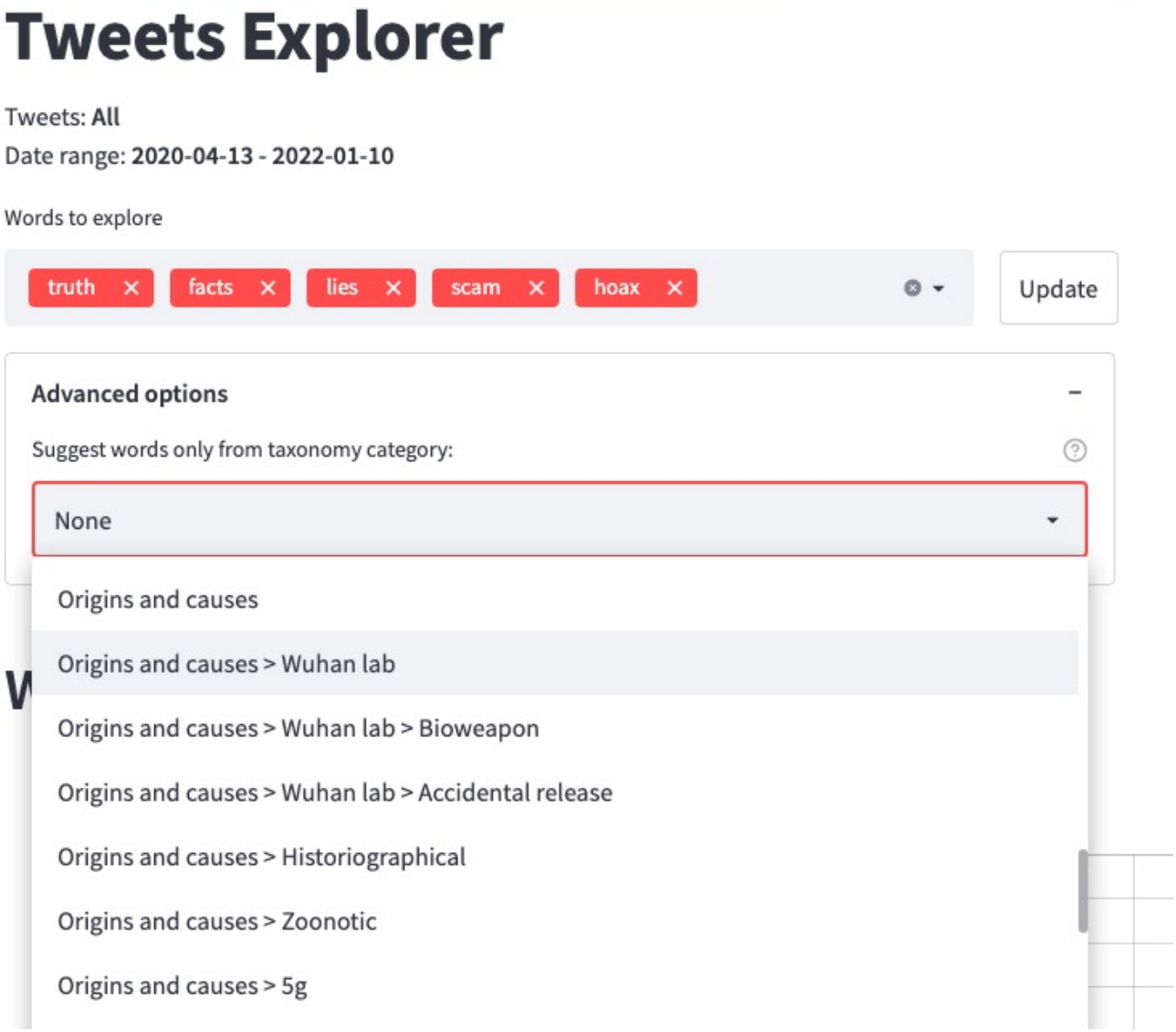




The Analyses

- 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

- Facet ngrams
- Initiate field search



The Analyses

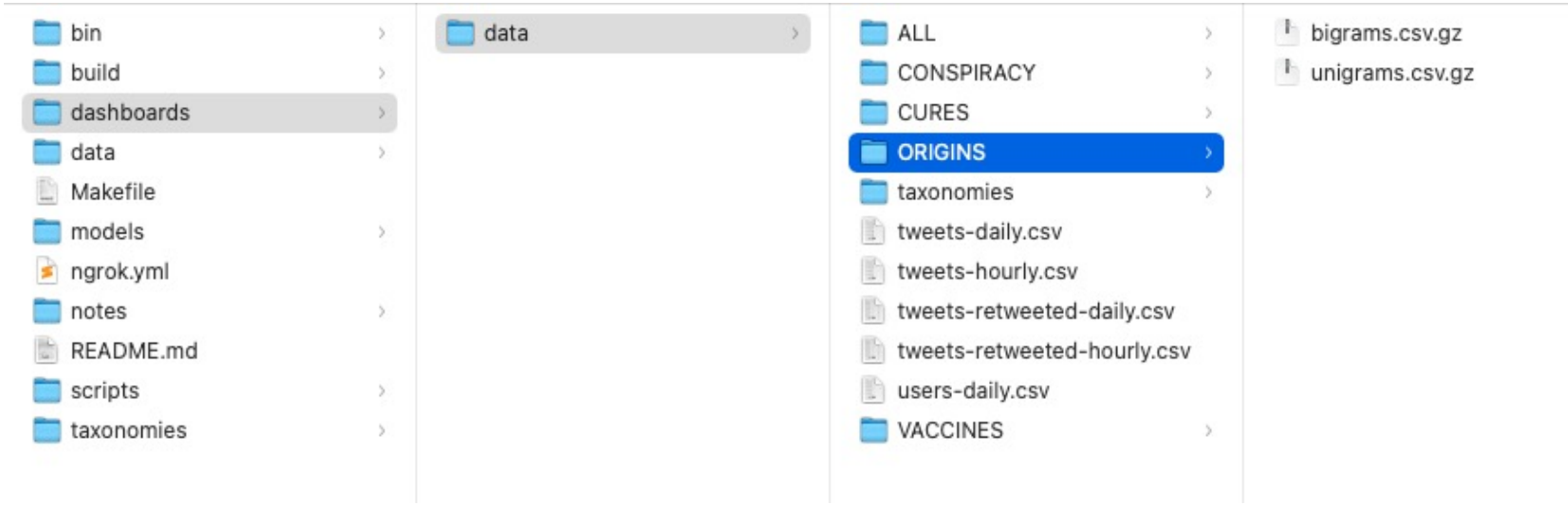
- Words, users, URLs, hashtags, languages
- ~~Sentiment analysis~~
- ~~User network analyses~~
- ~~Hashtag/Term cluster analyses~~
- Short Text Topic modelling
- Diachronic Topic visualisation
- Tweet walls
- Rumour vocabularies
- Classifiers
- Phase/Attrition data

```
============================= Training pipeline =============
<0x1b>[38;5;4mℹ Pipeline: ['tok2vec', 'tagger', 'parser', 'att
'lemmatizer', 'ner', 'textcat_multilabel']<0x1b>[0m
<0x1b>[38;5;4mℹ Frozen components: ['tagger', 'parser', 'attri
'lemmatizer', 'ner']<0x1b>[0m
<0x1b>[38;5;4mℹ 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

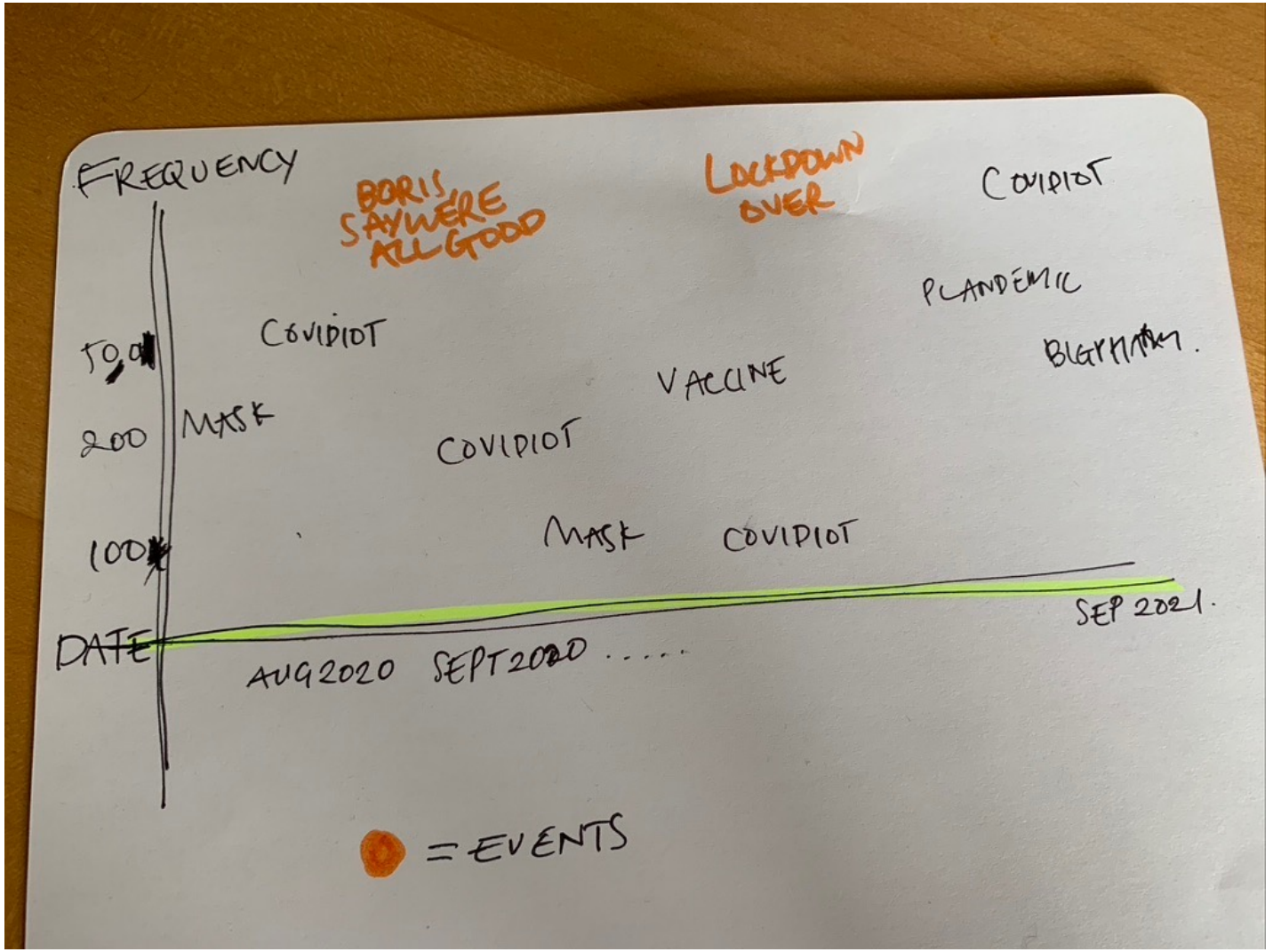
1. Harvesting
2. Cleaning
3. Preprocessing
4. Sampling
5. Annotating
6. Classifying
7. Visualising
8. Exploring
9. Interpreting

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

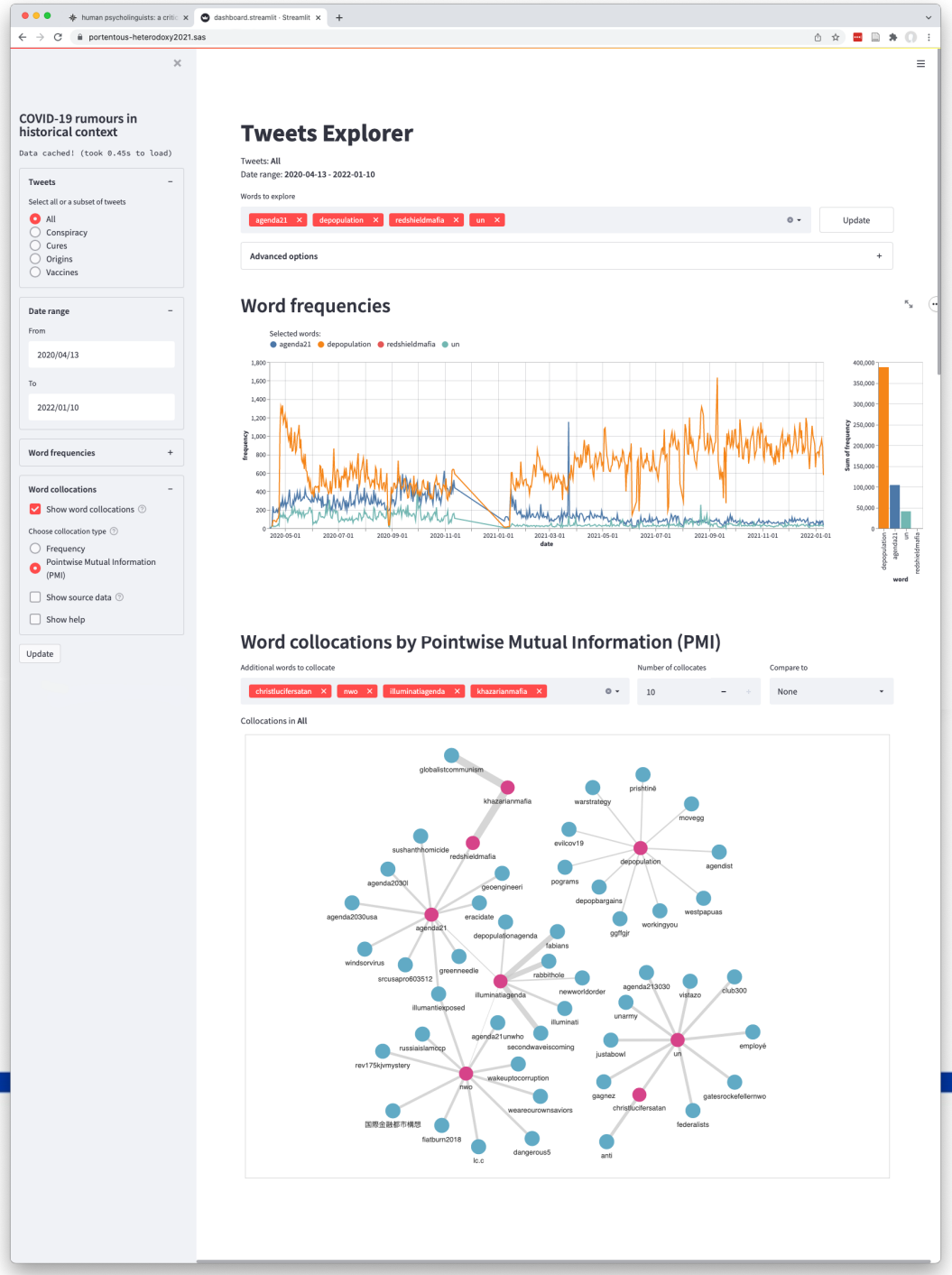


Kunika's prototype sketch

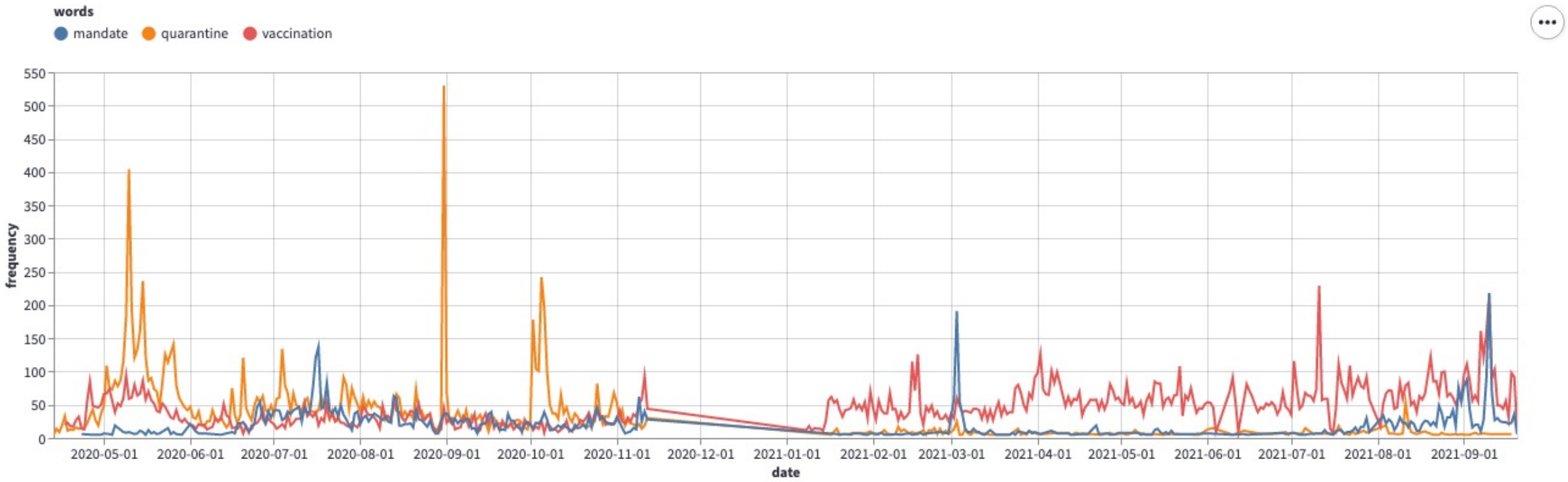
The Dashboard

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

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



Word frequencies



The Dashboard – Worldwide medical tyranny

Word collocations by Frequency

Additional words to collocate

mandate ✕ mandates ✕

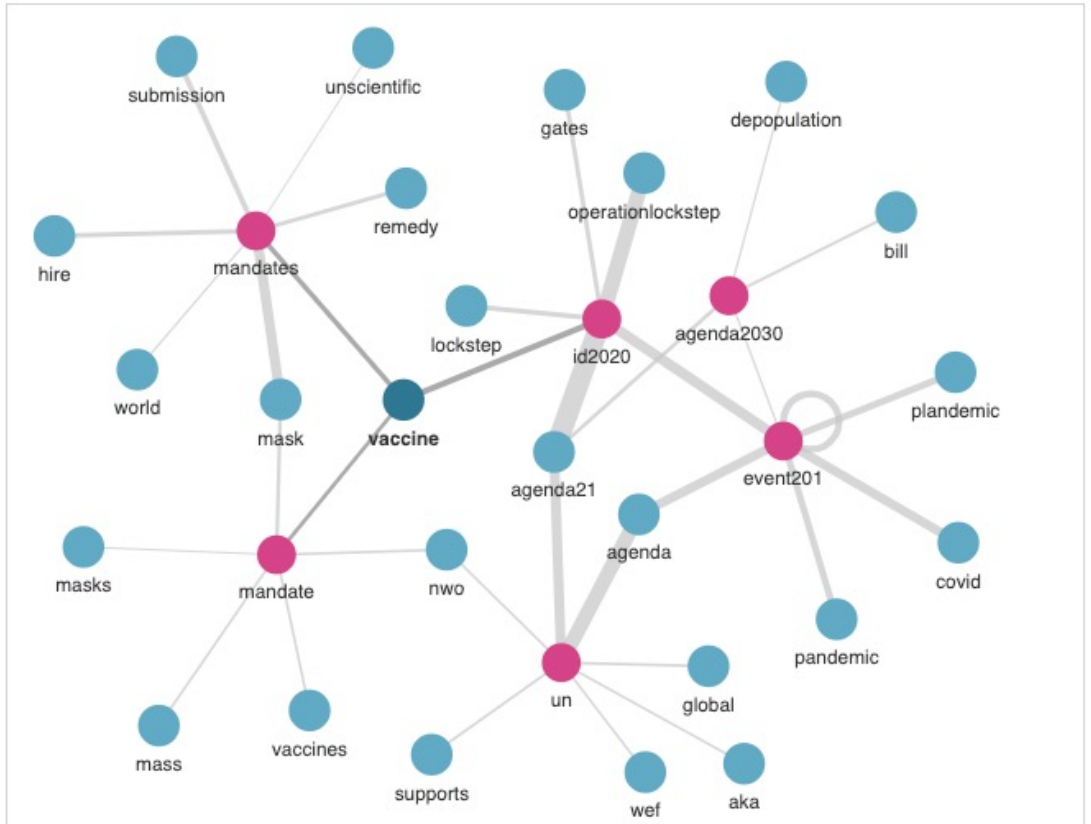
Number of collocates

7 − +

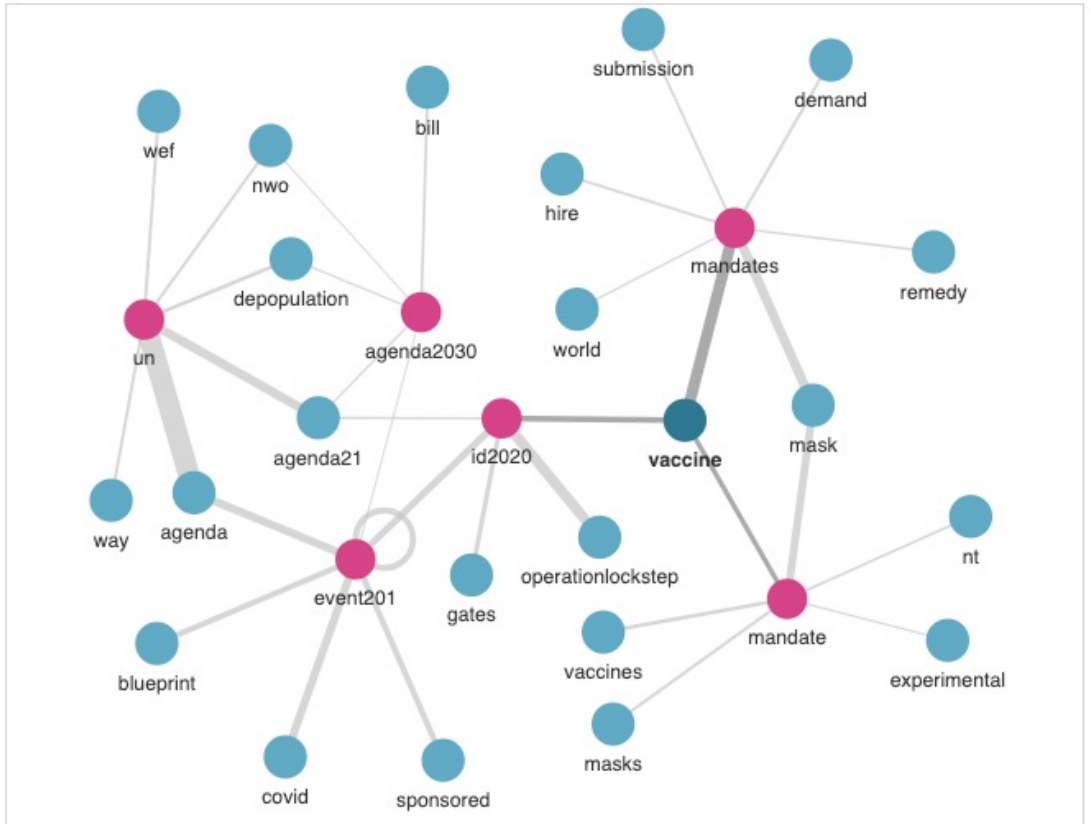
Compare to

Vaccines

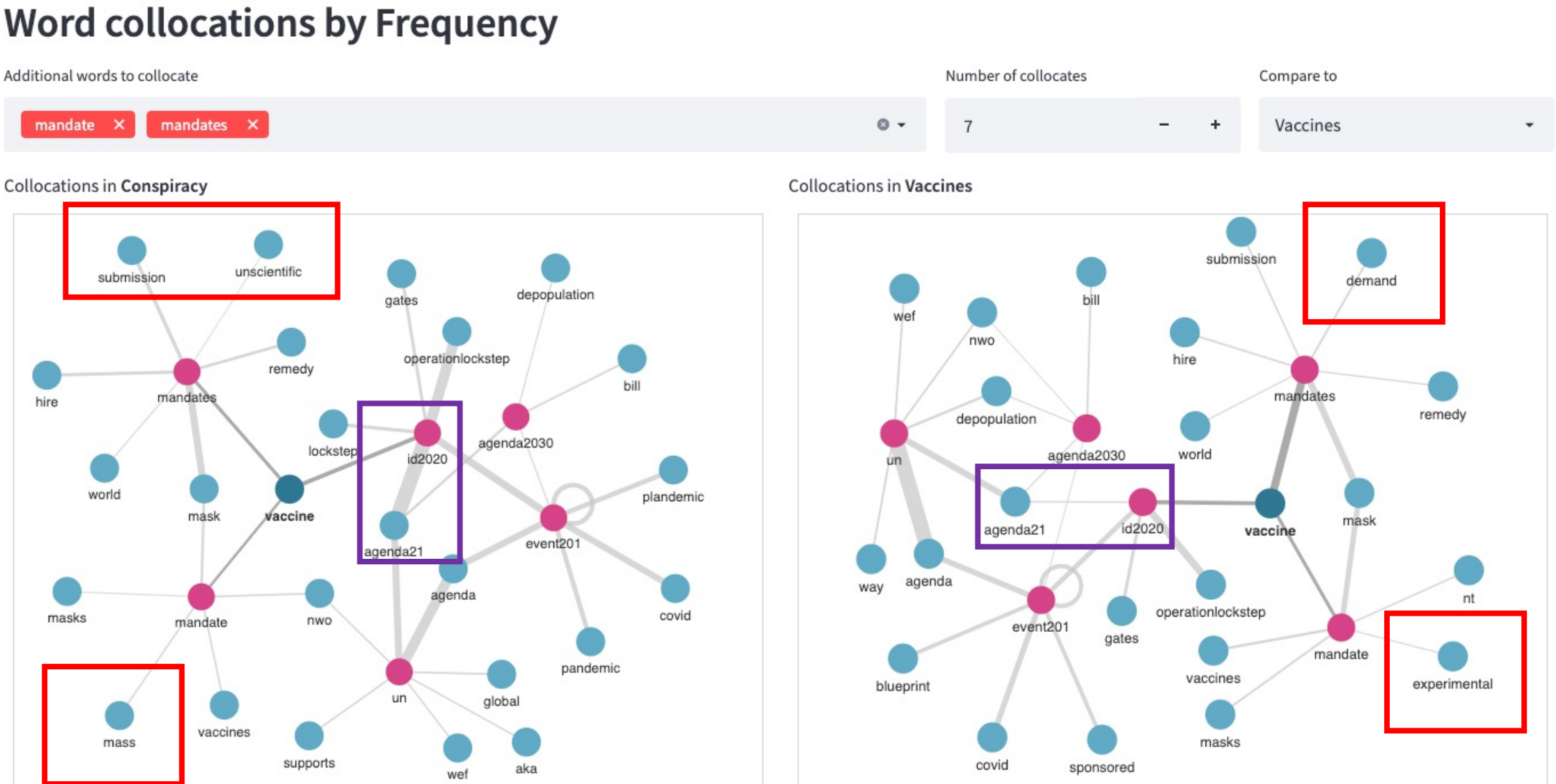
Collocations in **Conspiracy**



Collocations in **Vaccines**



The Dashboard – Worldwide medical tyranny



The Dashboard – Frequencies are not enough

Word collocations by Pointwise Mutual Information (PMI)

Additional n-grams to collocate

Choose an option

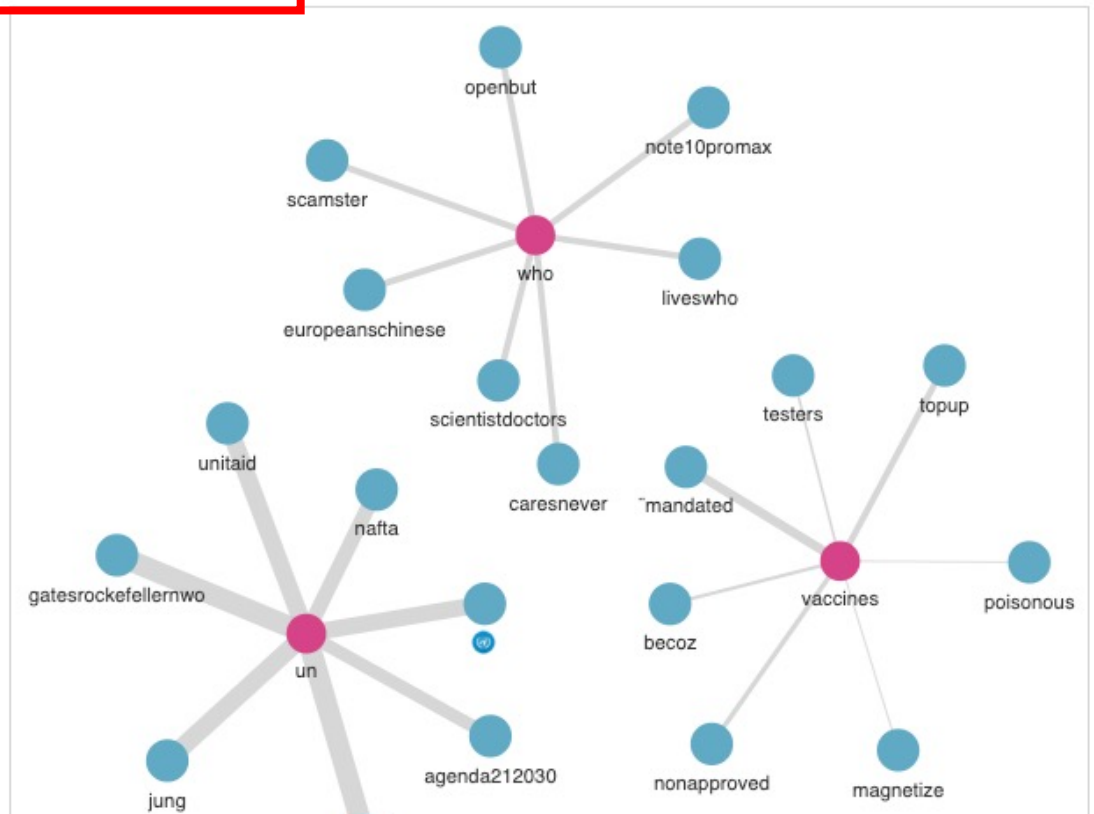
Number of collocates

7 − +

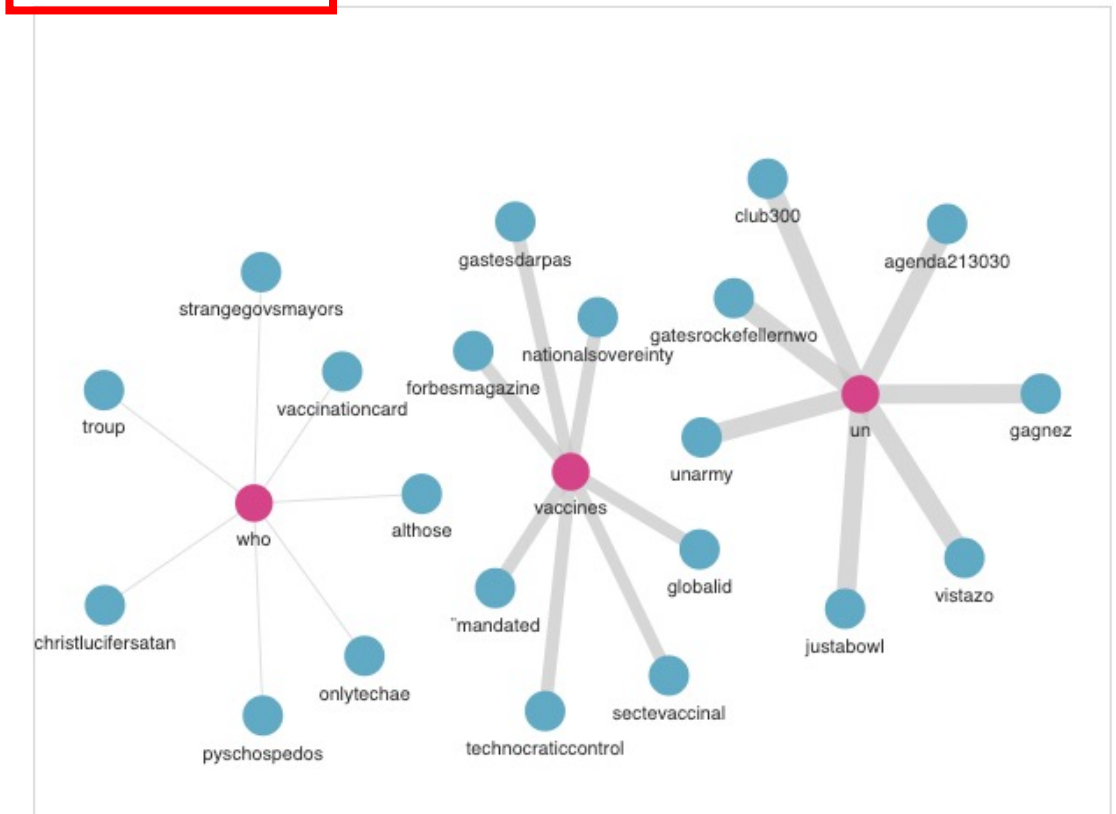
Compare to

All tweets

Collocations in **Vaccines**



Collocations in **All tweets**



The Dashboard – Frequencies are not enough

Word collocations by Pointwise Mutual Information (PMI)

Additional n-grams to collocate

Choose an option

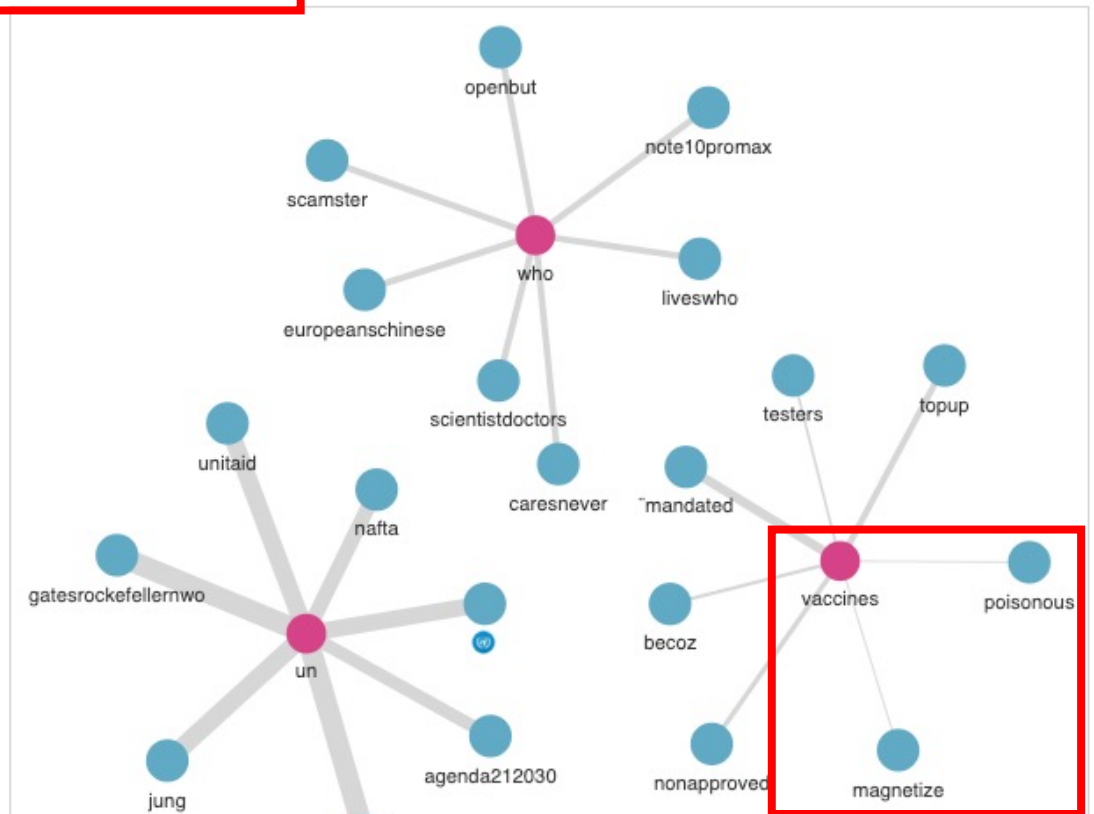
Number of collocates

7 − +

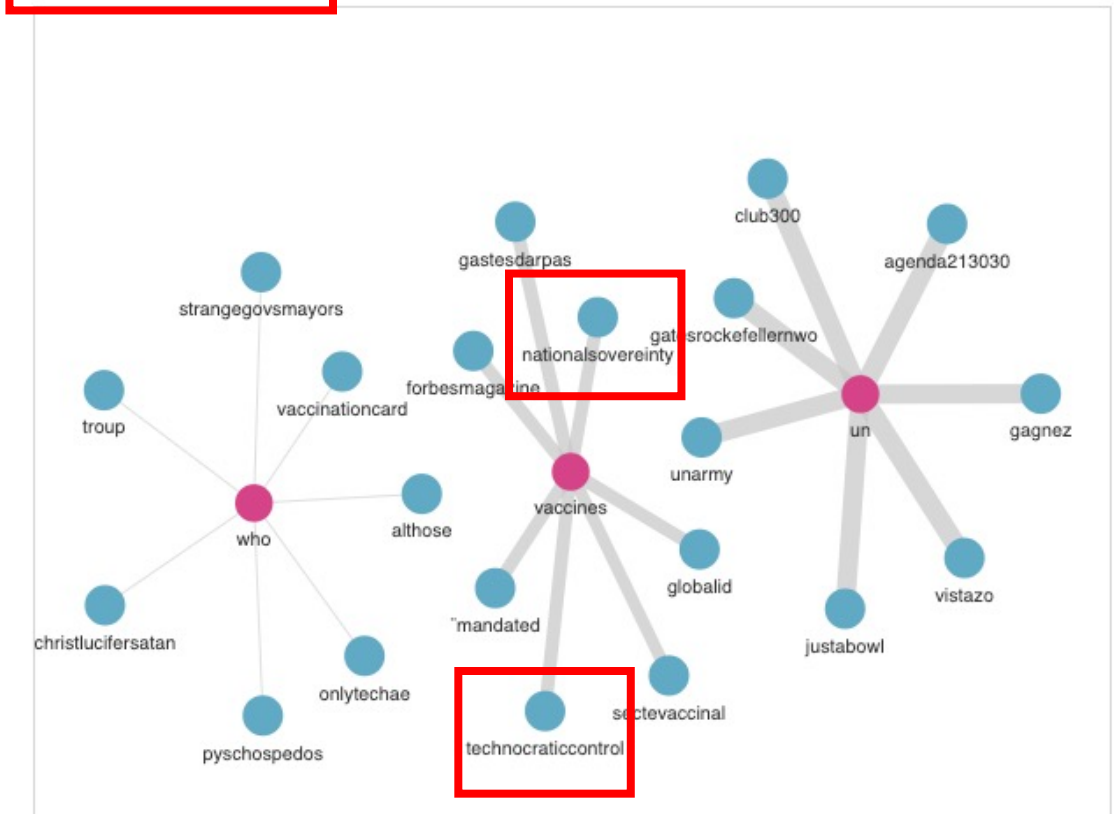
Compare to

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

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:

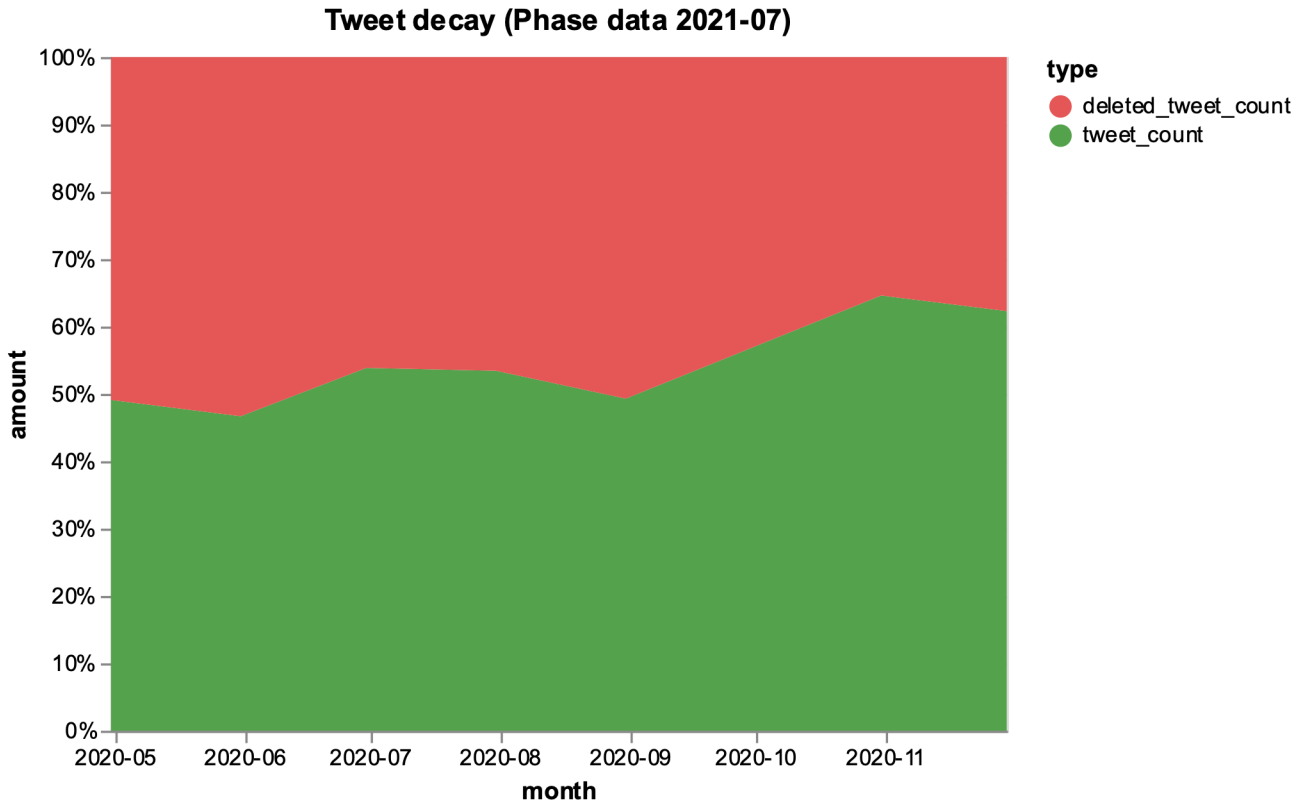
5g, AdvancedMedicine, agenda2020, agenda21, antivaccine, bioweapon, boycottqantas, cornaviruscoverup, coronabollocks, coronahoax, coronavirusliar, coronavirustruth, covidiot, covidiots, depopulation, depopulationagenda, DrRashidAButtar, endlockdown, endthelockdownuk, event201, fakepandemic, filmyourhospital, 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, covidiots, 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



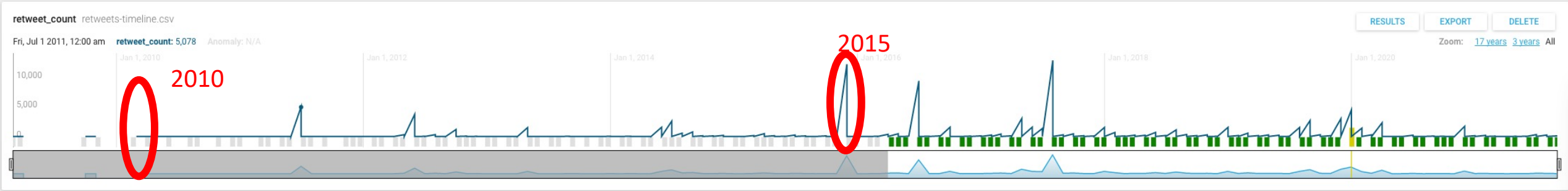
The end – Challenges with interpreting data

- 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!"

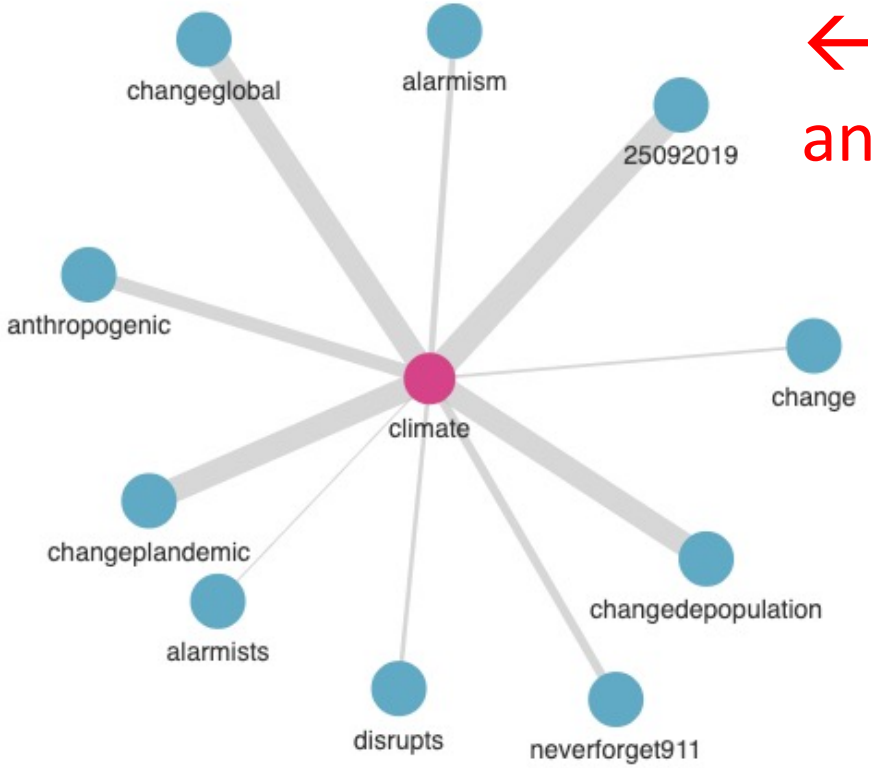


The end – Opportunities

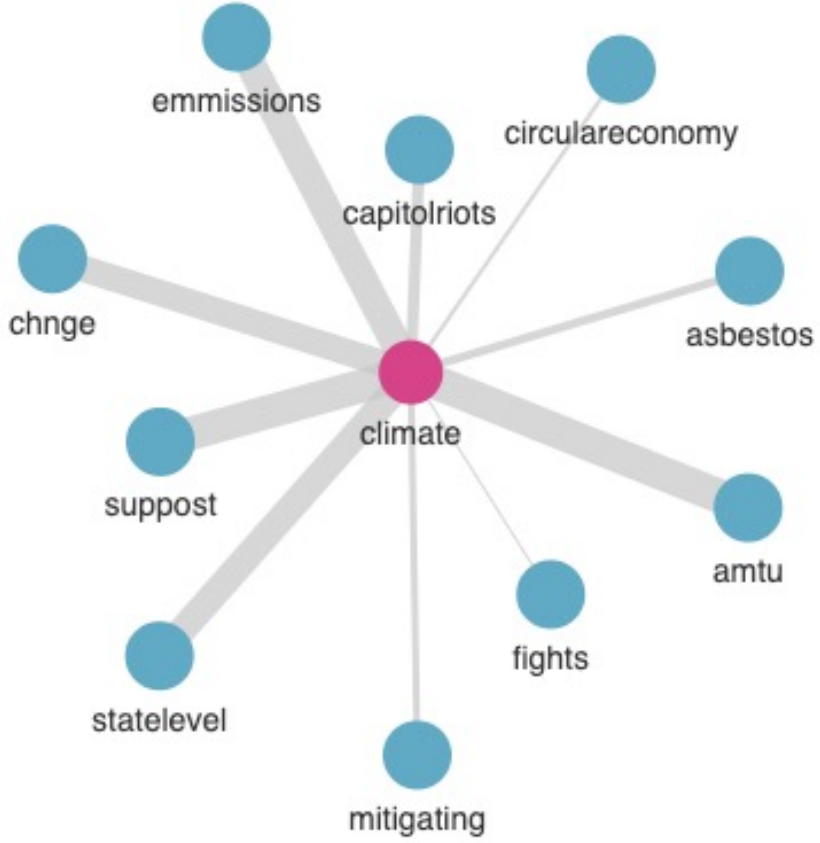
CURES

@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.

‘Climate’ change – Anthropogenic environmental phenomenon or Diachronic psycholinguistic signal?



← 2020, alarmism and disruption



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