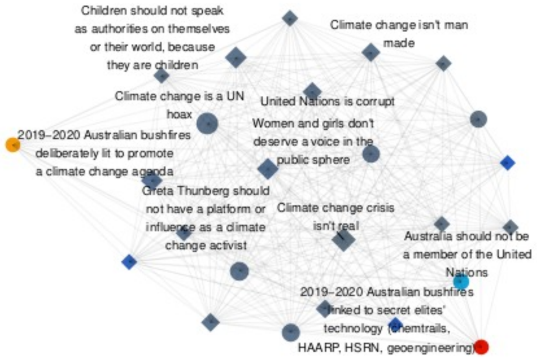
Behavioral Data Science

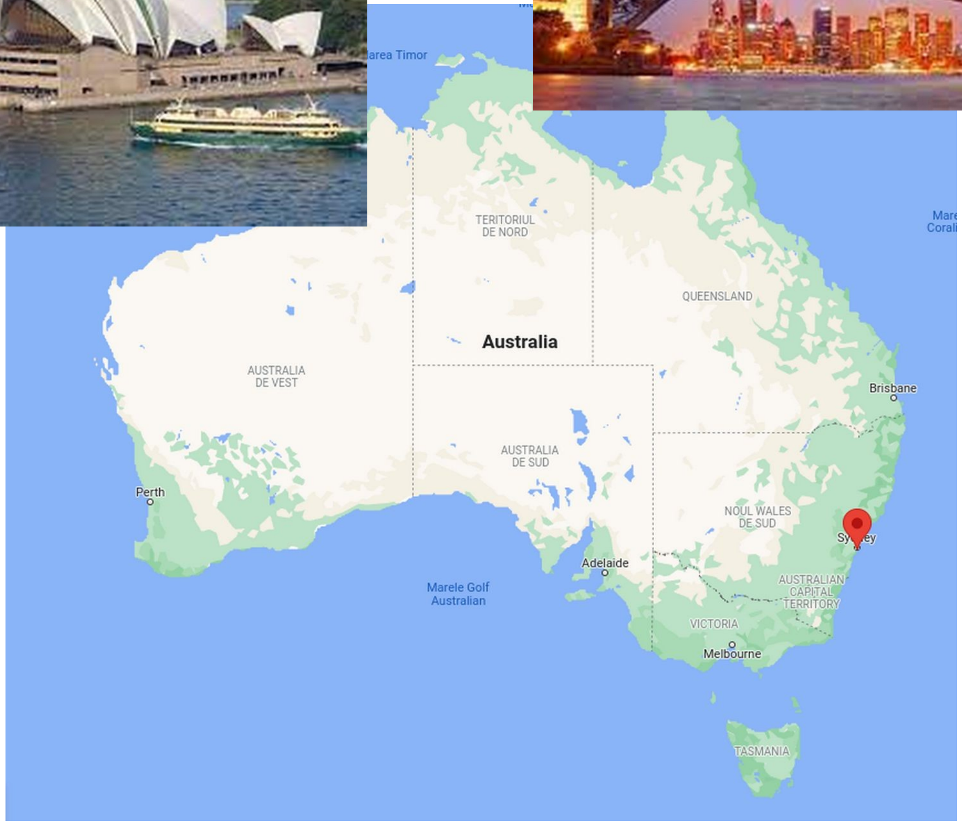
UTS Data Science

UTS UNIVERSITY OF TECHNOLOGY SYDNEY



Mapping Online Problematic Content: Mixing Qualitative Approaches with State-of-the-art Machine Learning

Marian-Andrei Rizoiu | Behavioural Data Science
Marian-Andrei.Rizoiu@uts.edu.au
https://www.behavioral-ds.science



Located in Sydney, Australia

A city campus, iconic brutalist style blended with modern buildings

The research group

Behavioral Data Science

1 PostDoc, 6 PhD, 3 Masters, 1 assistant prof.





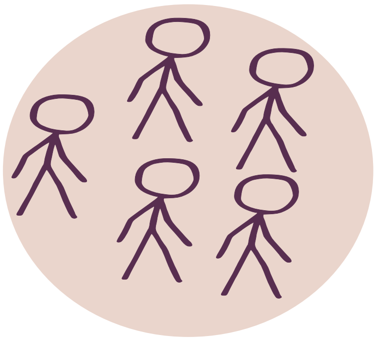




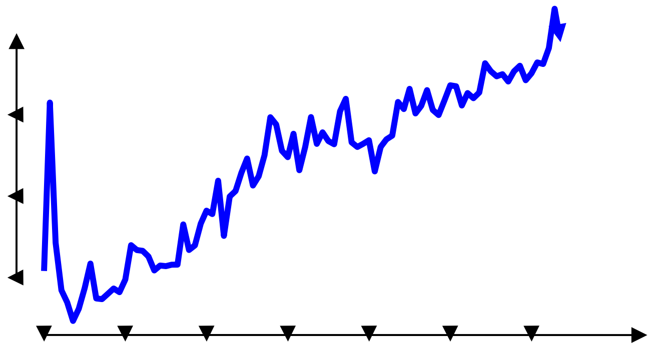


The Behavioral Data Science

1.

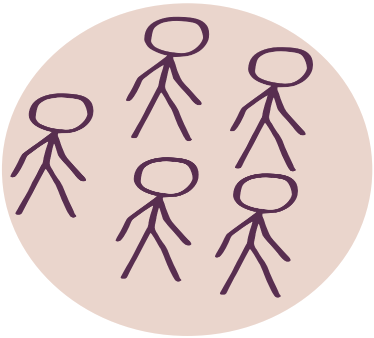


information diffusion
epidemics spreading
behavioral modeling

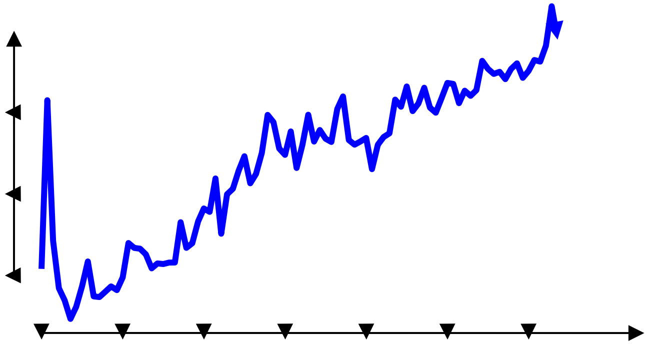


The Behavioral Data Science

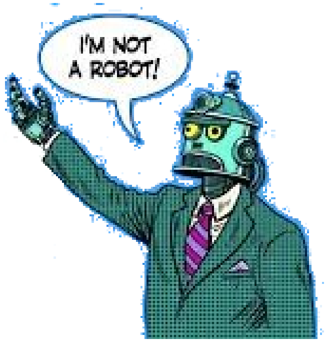
1.



information diffusion
epidemics spreading
behavioral modeling



2.

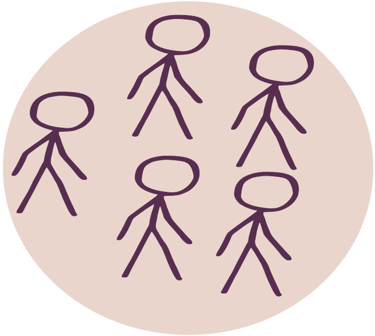


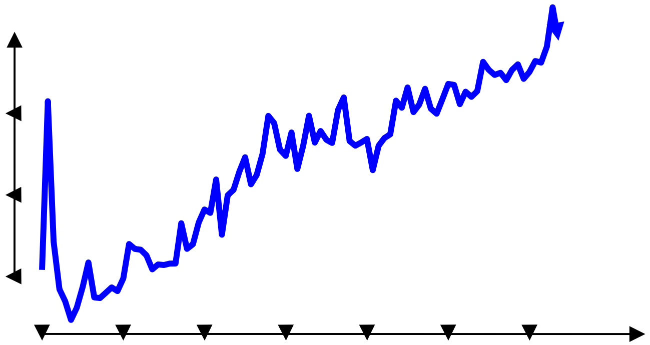
[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

The Behavioral Data Science

1.

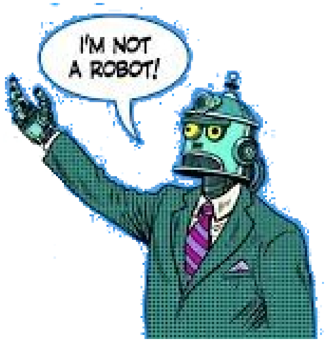




information diffusion
epidemics spreading
behavioral modeling

3.

2.



[Rizoiu et al ICWSM'18] [Kim et al Journ.Comp.SocSci'19]

Presentation plan

The interdisciplinary team:
Lessons learned – doing inter-disciplinary research

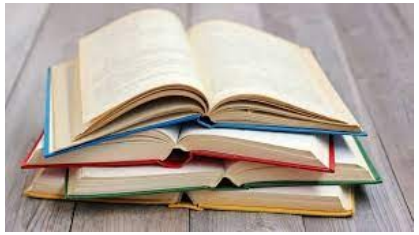
The tools:
Crash course: how to analyse social media data? what is data, what is an API, what is a classifier, and how do you measure performance?

The research:
Slipping to the extremes: combining qualitative research and computer science to fight problematic speech

Lessons learned:
Inter-disciplinary research work

Interdisciplinary approach and team

Communication science



Literature

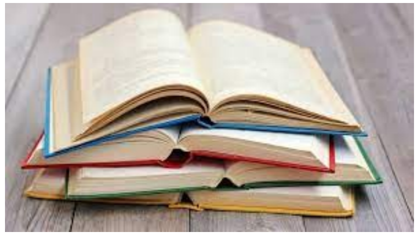
Computer science

Interdisciplinary approach and team

Communication science

Francesco Bailo

Amelia Johns



Literature

Emily Booth

Computer science

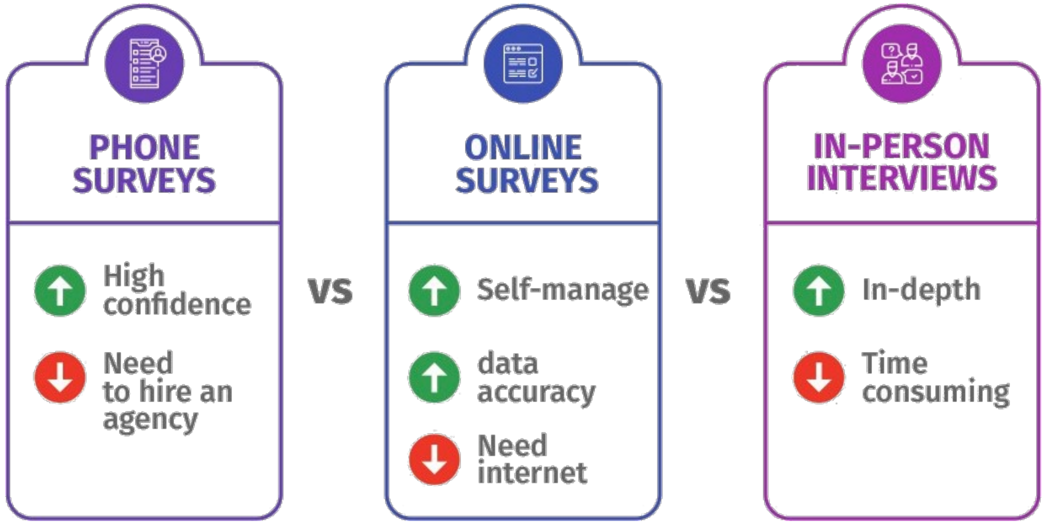
Marian-Andrei Rizoiu

Quyu Kong

Language and methodological differences

- Same **terms** might mean different things to different disciplines: *paper, quantitative, data, database, etc.*
- Approach to collecting data

Computer science approach



Social science approach

Language and methodological differences

- Same **terms** might mean different things to different disciplines: *paper, quantitative, data, database, etc.*
- Approach to collecting data
- Each field has its limits:
  - *Computer science* – simplifying assumptions about how the world works (e.g., we assume that every user is uniformly connected to every other user)
  - *Social science* – naive assumptions (wishful thinking) about what is technically possible (e.g., we will collect all Facebook posts and comments, and extract opinions and sentiment)

The whole is larger than the sum of parts

- Slow start, but cross-polination is worth the effort
- It opens the eyes to a wealth of new problems and approaches – **builds a completely new way of approaching research**
- Complementarity serves individuals better than duplication of skills

- You grow individually – **diversity of ideas leads to better ideas**

Tips & Tricks

- Listen and don't judge
- Try to understand their vocabulary and methods – if an entire field converged on a set of methods, there must be something right in them
- If sometimes happens that your field does it wrong / suboptimal
- Meet regularly – any progress is better than no progress
- Explain your approach and methods – as simple as possible, and repeat if necessary
- Keep engaging and build trust – **inter-disciplinary relations pay off in the long term**

Presentation plan

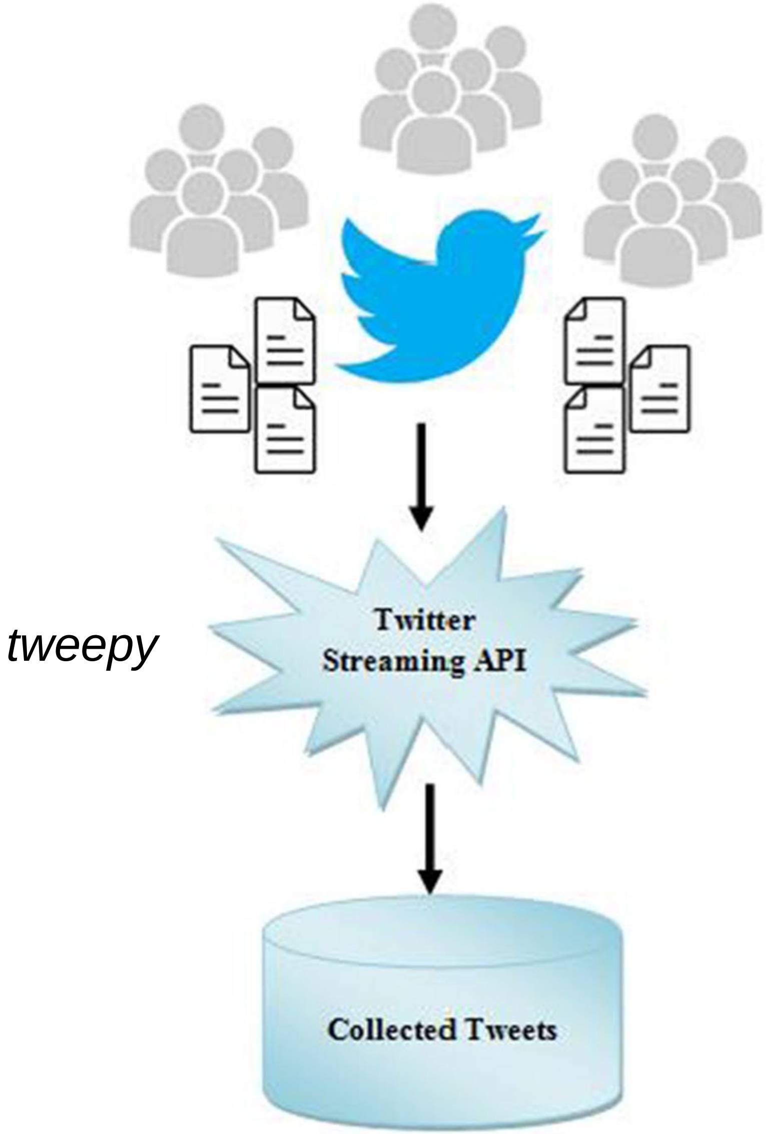
The interdisciplinary team:
Lessons learned – doing inter-disciplinary research

The tools:
Crash course: how to analyse social media data? what is data, what is an API, what is a classifier, and how do you measure performance?

The research:
Slipping to the extremes: combining qualitative research and computer science to fight problematic speech

Quantitative VS qualitative research

Collecting Twitter data (1)



Tuesday June 21, 2022 - APIs and Digital Trace Data

09:30-10:00 Coffee/tea

10:00-11:00 Collecting digital trace data
Tristram Alexander

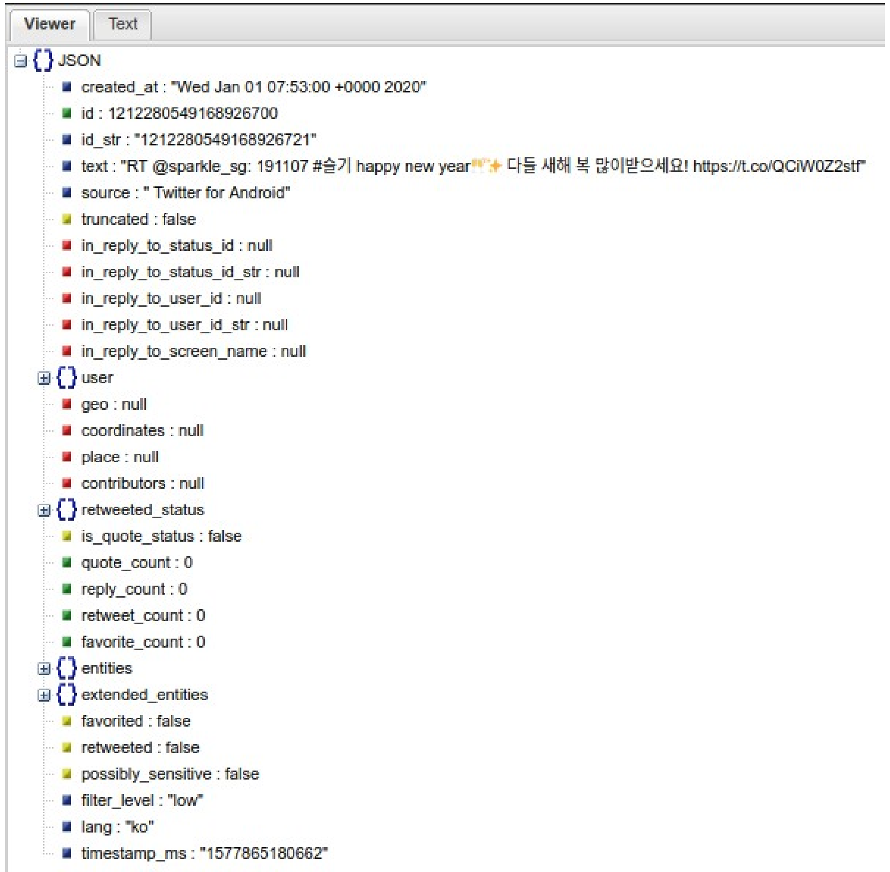
11:00-12:30 Academic research using the Twitter API (workshop)
Suhem Parack
This workshop is open to the public, register here
Click here to request academic access to Twitter API

12:30-13:00 Lunch break

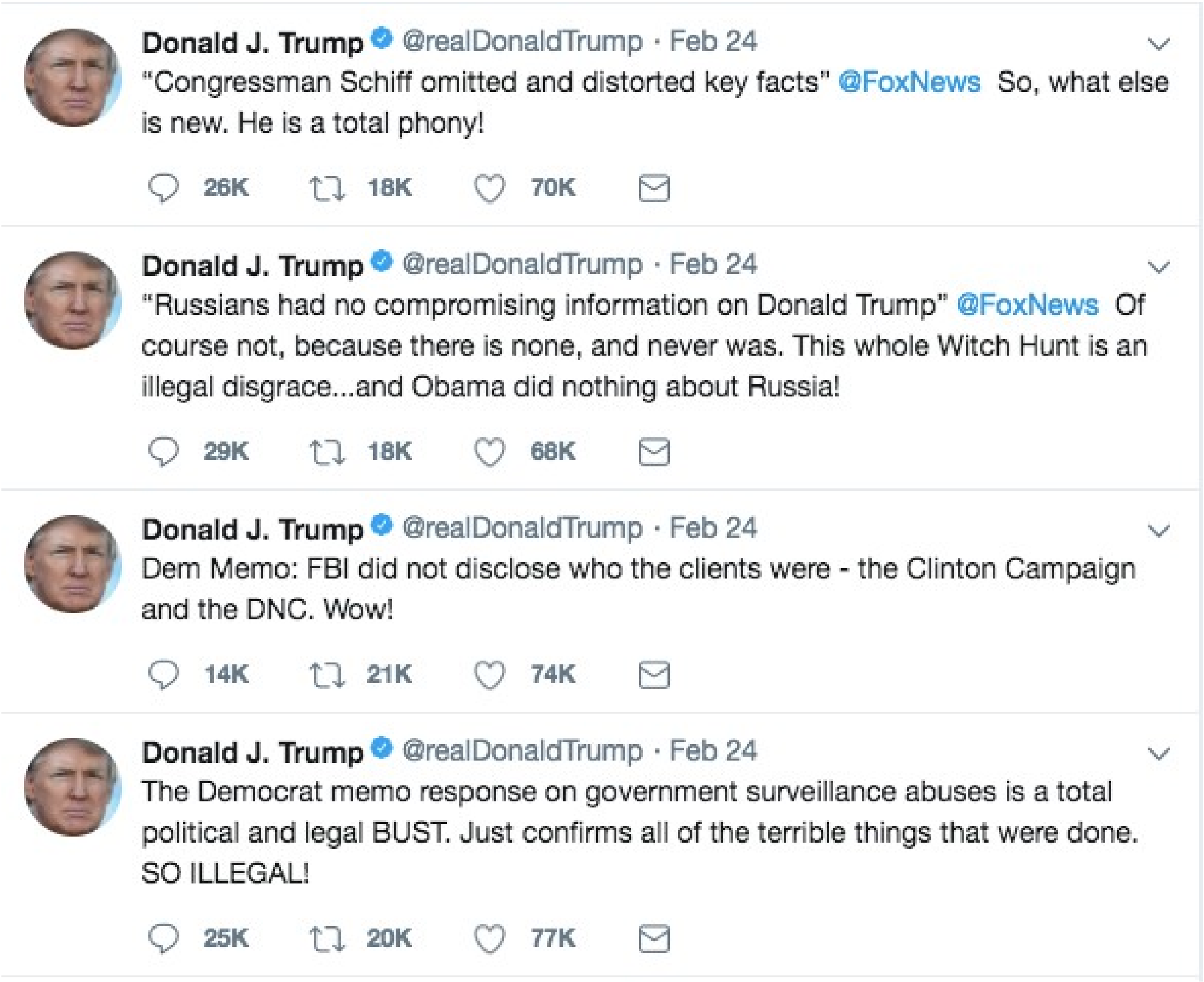
13:00-14:00 Research data management
Simon Musgrave, Ben Foley, Sam Hames
This workshop is open to the public, register here

14:00-16:00 Collecting digital trace data (group exercise)
Tristram Alexander

Collecting Twitter data (2)



Tweet textual analysis (1)

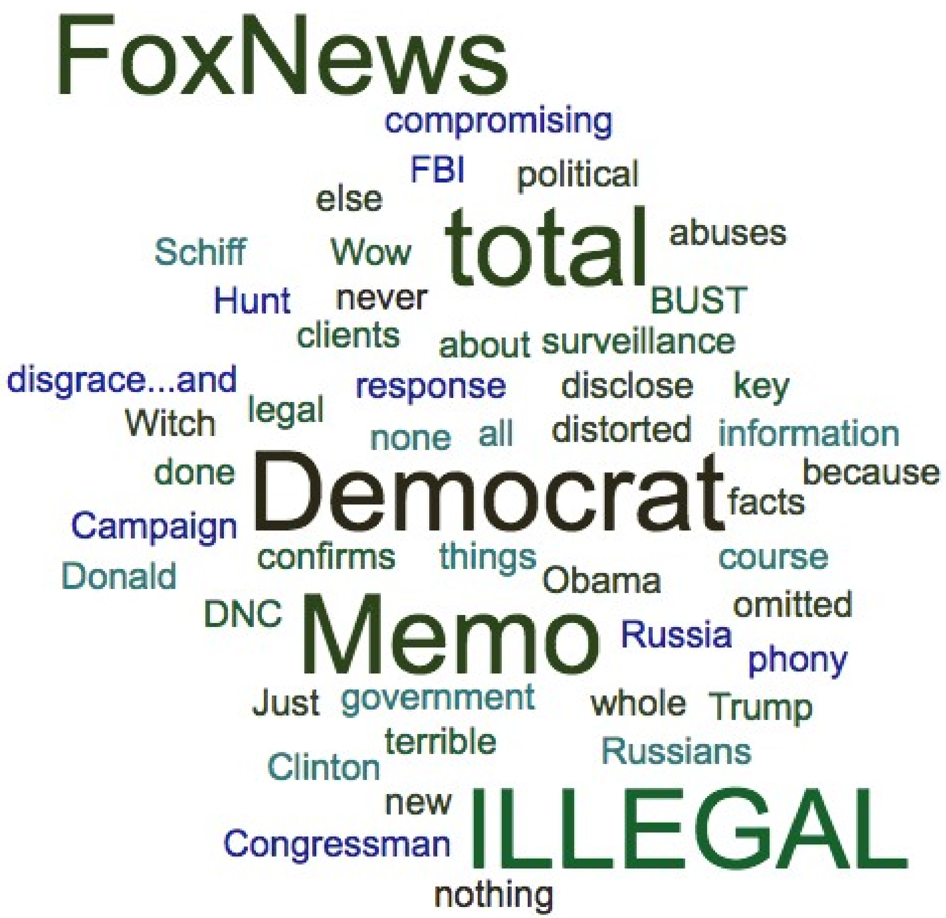


	1	2	3	4
and	1	2	1	1
congressman	1	0	0	0
distorted	1	0	0	0
else	1	0	0	0
facts"	1	0	0	0
foxnews	1	1	0	0
key	1	0	0	0
new	1	0	0	0
omitted	1	0	0	0
phony	1	0	0	0
schiff	1	0	0	0
total	1	0	0	1
what	1	0	0	0
about	0	1	0	0

… truncated for space.

Tweet textual analysis (2)

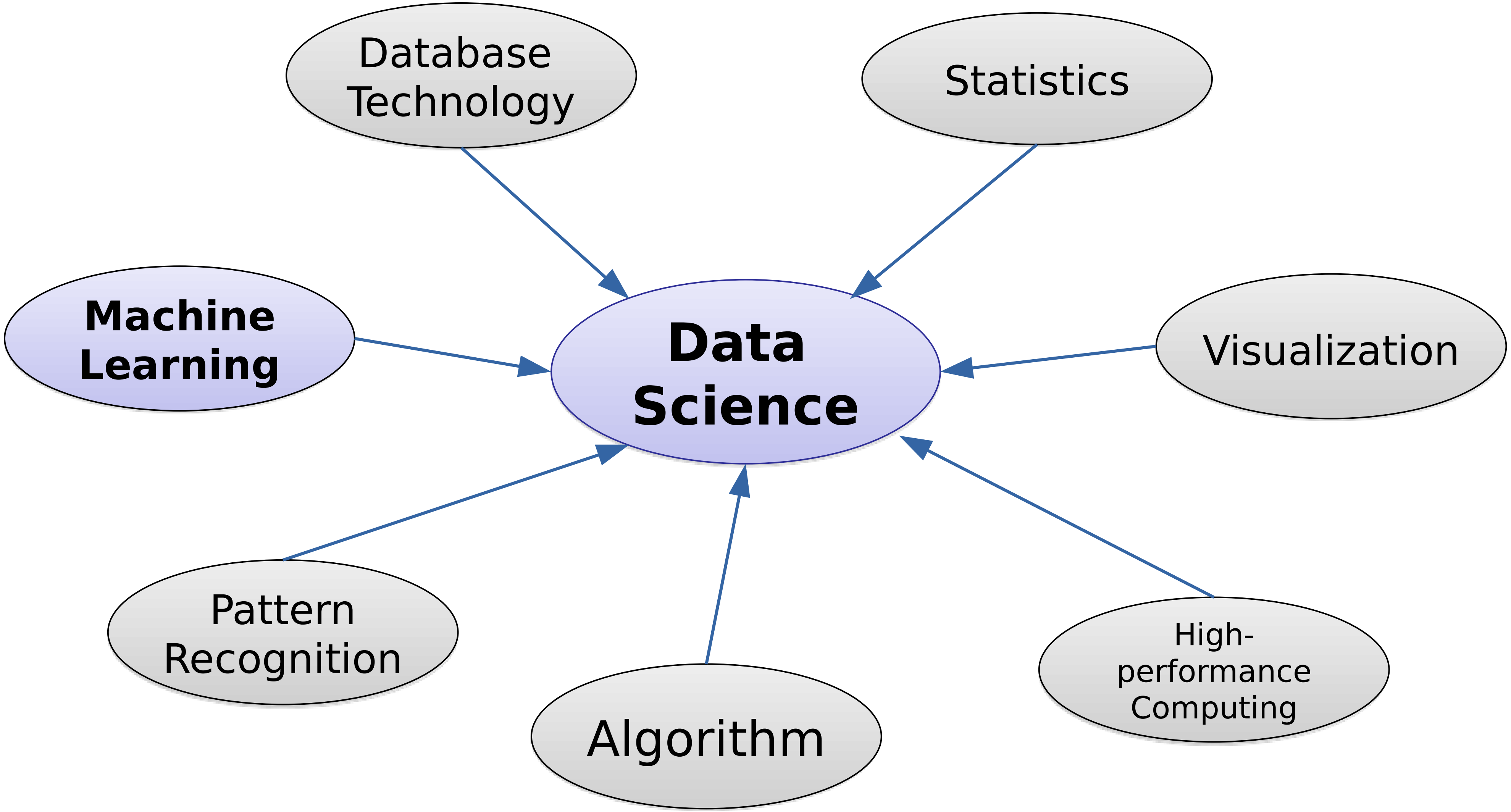
	1	2	3	4
and	1	2	1	1
congressman	1	0	0	0
distorted	1	0	0	0
else	1	0	0	0
facts"	1	0	0	0
foxnews	1	1	0	0
key	1	0	0	0
new	1	0	0	0
omitted	1	0	0	0
phony	1	0	0	0
schiff	1	0	0	0
total	1	0	0	1
what	1	0	0	0
about	0	1	0	0



Quantitative research aims

- Numerically counting things
- Determining the relationship between one thing [an **independent variable**] and another [a **dependent or outcome variable**] within a population
- Prediction, generalisability, causality
- Statistical analysis

Data Science: Confluence of Multiple Disciplines



What is **Machine Learning**?

Machine learning is about prediction

Examples/features	$x_1, \ldots, x_n \sim \mathcal{X}$
Labels/annotations	$y_1, \ldots, y_n \sim \mathcal{Y}$
Predictor	$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Estimate best predictor = training

Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

- No mechanistic model of the phenomenon
- There is relatively large amounts of data (examples, x usually $\mathbb{R}^d$)
- The outcomes (labels, y usually binary) are well defined

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Or the society we live in?

What is **Data**?

- Data=a table (dataset, database, sample)

Variables (attributes, features) = measurements made on objects

	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5	VAR 6	VAR 7	VAR 8	VAR 9	VAR 10	VAR 11	...
Object 1	0	1	2	0	1	1	2	1	0	2	0	...
Object 2	2	1	2	0	1	1	0	2	1	0	2	...
Object 3	0	0	1	0	1	1	2	0	2	1	2	...
Object 4	1	1	2	2	0	0	0	1	2	1	1	...
Object 5	0	1	0	2	1	0	2	1	1	0	1	...
Object 6	0	1	2	1	1	1	1	1	1	1	1	...
Object 7	2	1	0	1	1	2	2	2	1	1	1	...
Object 8	2	2	1	0	0	0	1	1	1	1	2	...
Object 9	1	1	0	1	0	0	0	0	1	2	1	...
Object 10	1	2	2	0	1	0	1	2	1	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Objects (samples, observations, individuals, examples, patterns)

Dimension=number of variables
Size=number of objects

- Objects: samples, patients, documents, images...
- Variables: genes, proteins, words, pixels...

What is **Supervised learning**?

Inputs | Output

A1	A2	A3	A4	Y
-0.69	-0.72	Y	0.47	Healthy
-2.3	-1.2	N	0.15	Disease
0.32	-0.9	N	-0.76	Healthy
0.37	-1	Y	-0.59	Disease
-0.67	-0.53	N	0.33	Healthy
0.51	-0.09	Y	-0.05	Disease

Learning sample

Supervised learning →

$$\hat{Y} = h(A_1, A_2, A_3, A_4)$$

model, hypothesis

- Goal: from the database (learning sample), find a function *h* of the inputs that approximates ***at best** the output*
- Discrete output ⇒ *classification* problem
- Numerical output ⇒ *regression* problem

Two main goals

- **Predictive:**

  Make predictions for a ***new*** sample described by its attributes

A1	A2	A3	A4	Y
0.83	-0.54	T	0.68	Healthy
-2.3	-1.2	F	-0.83	Disease
0.08	0.63	F	0.76	Healthy
0.06	-0.29	T	-0.57	Disease
-0.98	-0.18	F	-0.38	Healthy
-0.68	0.82	T	-0.95	Disease
0.92	-0.33	F	-0.48	?

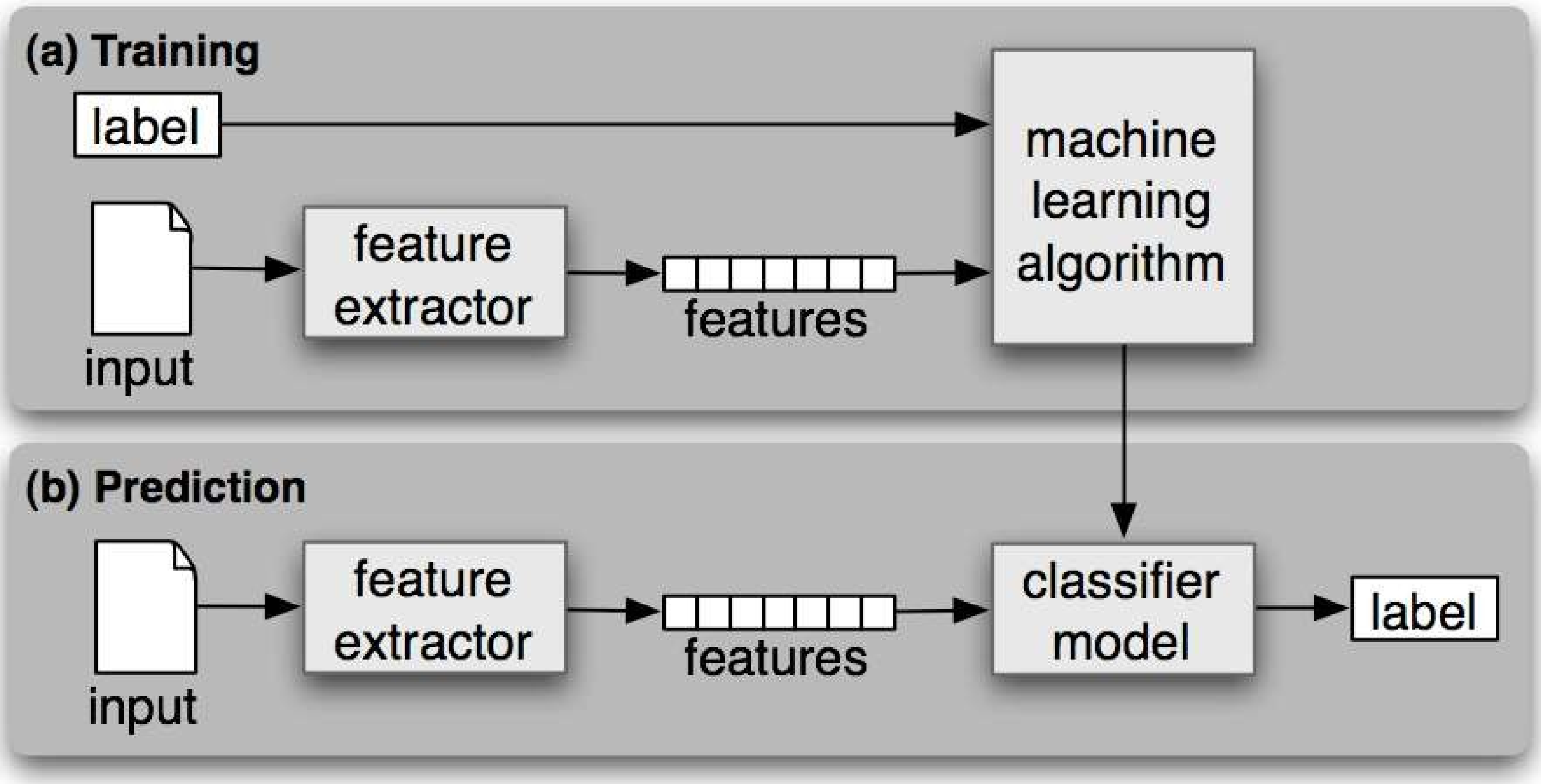
- **Informative:**

  Help to understand the relationship between the inputs and the output

$$Y\text{=disease if } A3\text{=F and } A2<0.3$$

  Find the most relevant inputs

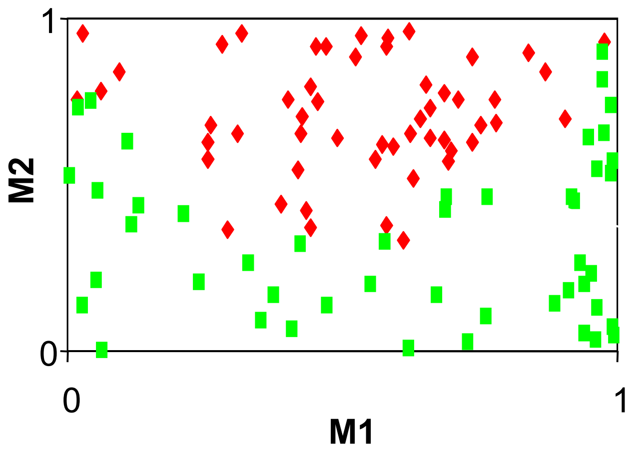
Workflow


(a) Training
label
machine learning algorithm
feature extractor
features
input
(b) Prediction
feature extractor
features
classifier model
label
input

Illustrative problem

- Medical diagnosis from two measurements (eg., weights and temperature)

M1	M2	Y
0.52	0.18	Healthy
0.44	0.29	Disease
0.89	0.88	Healthy
0.99	0.37	Disease
...	...	...
0.95	0.47	Disease
0.29	0.09	Healthy

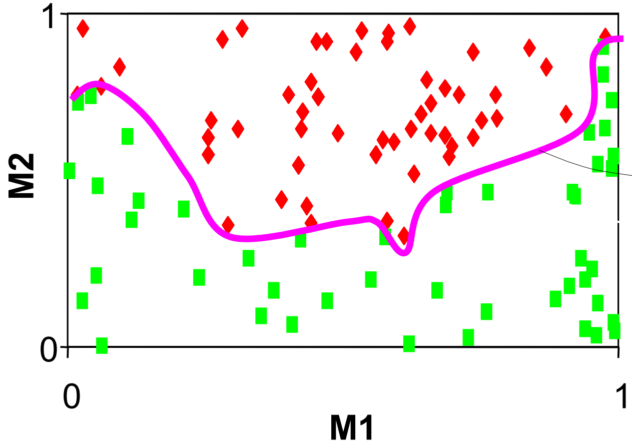


- **Goal:** find a model that classifies at best new cases for which M1 and M2 are known

Illustrative problem

- Medical diagnosis from two measurements (eg., weights and temperature)

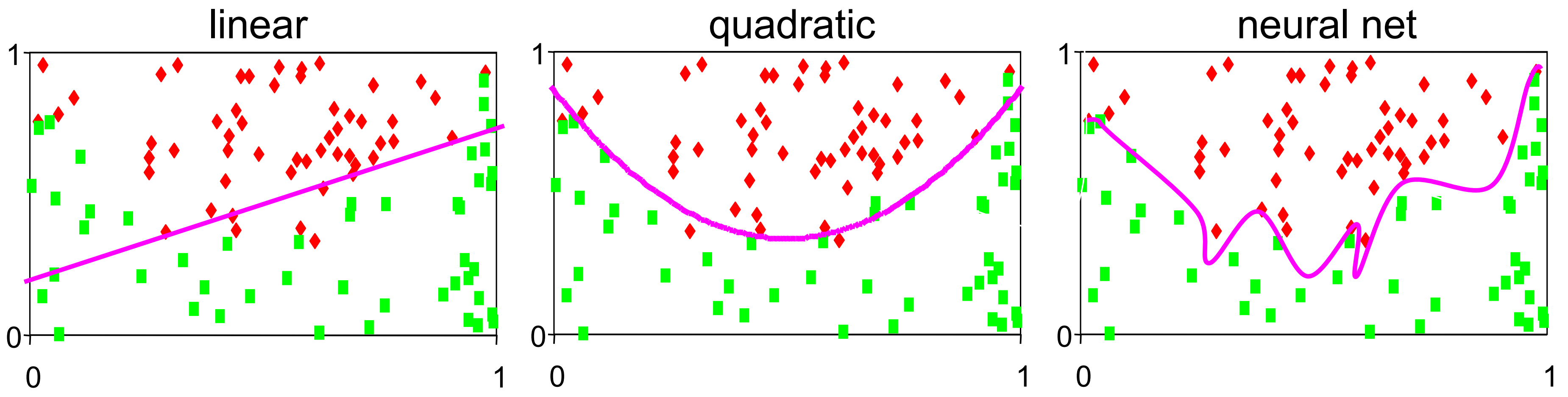
M1	M2	Y
0.52	0.18	Healthy
0.44	0.29	Disease
0.89	0.88	Healthy
0.99	0.37	Disease
...	...	...
0.95	0.47	Disease
0.29	0.09	Healthy



a model obtained by supervised learning

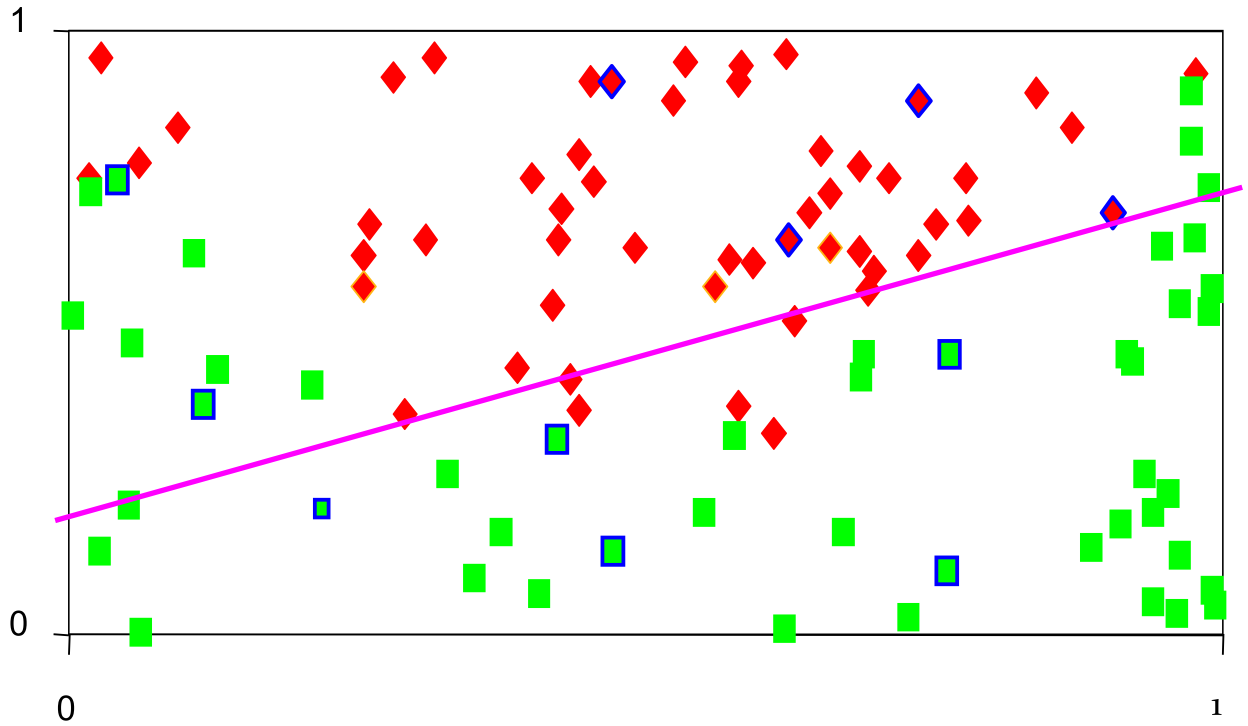
- **Goal:** find a model that classifies at best new cases for which M1 and M2 are known

But which model? So many choices…



How well do we predict future similar data? (*generalisation error*)

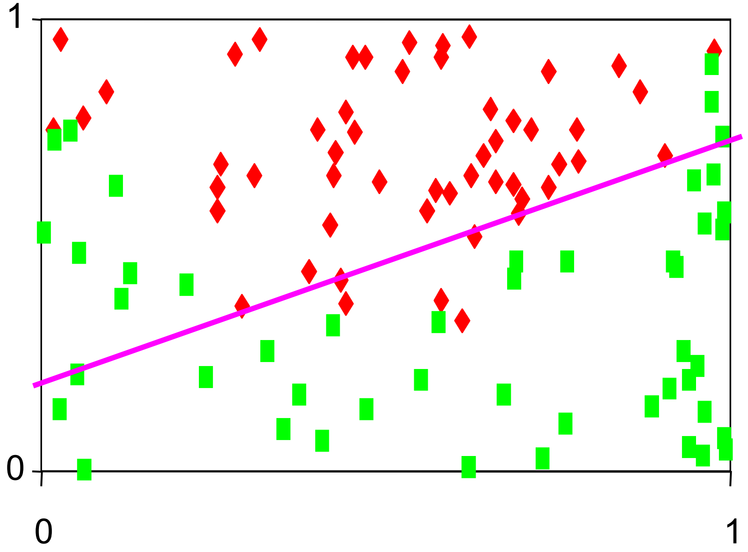
The test set method



1. Randomly choose 30% of the data to be in a test sample
2. The remainder is a learning sample
3. Learn the model from the learning sample
4. Estimate its future performance on the test sample

Which model is the best?

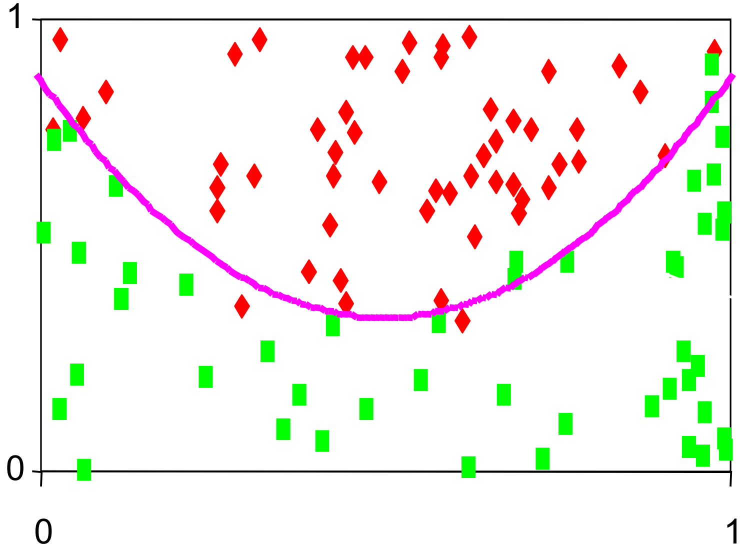
linear



LS error= 3.4%
TS error= 3.5%

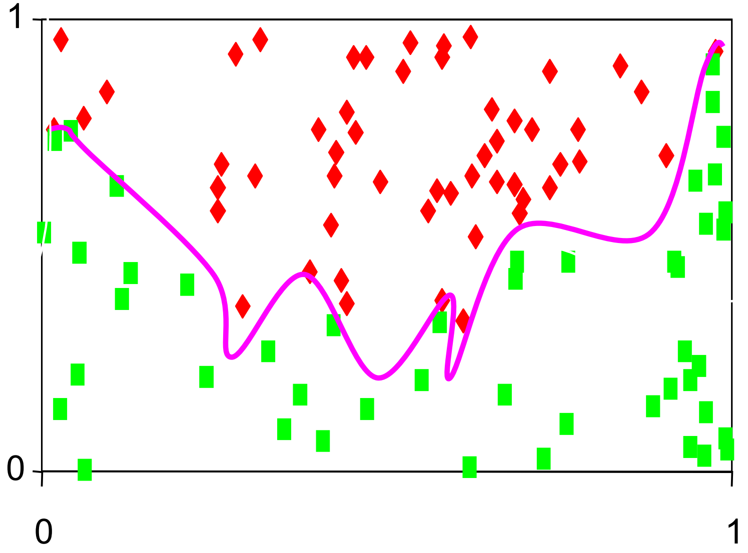
under-fitting

quadratic



LS error= 1.0%
TS error= 1.5%

neural net

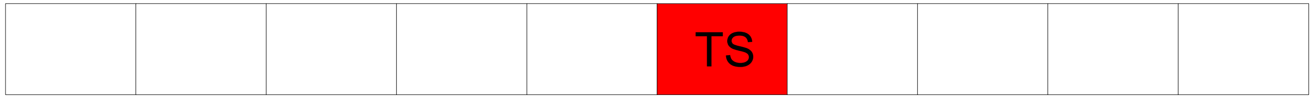


LS error= 0%
TS error= 3.5%

over-fitting

k-fold Cross Validation

- Randomly partition the dataset into *k* subsets (for example 10)

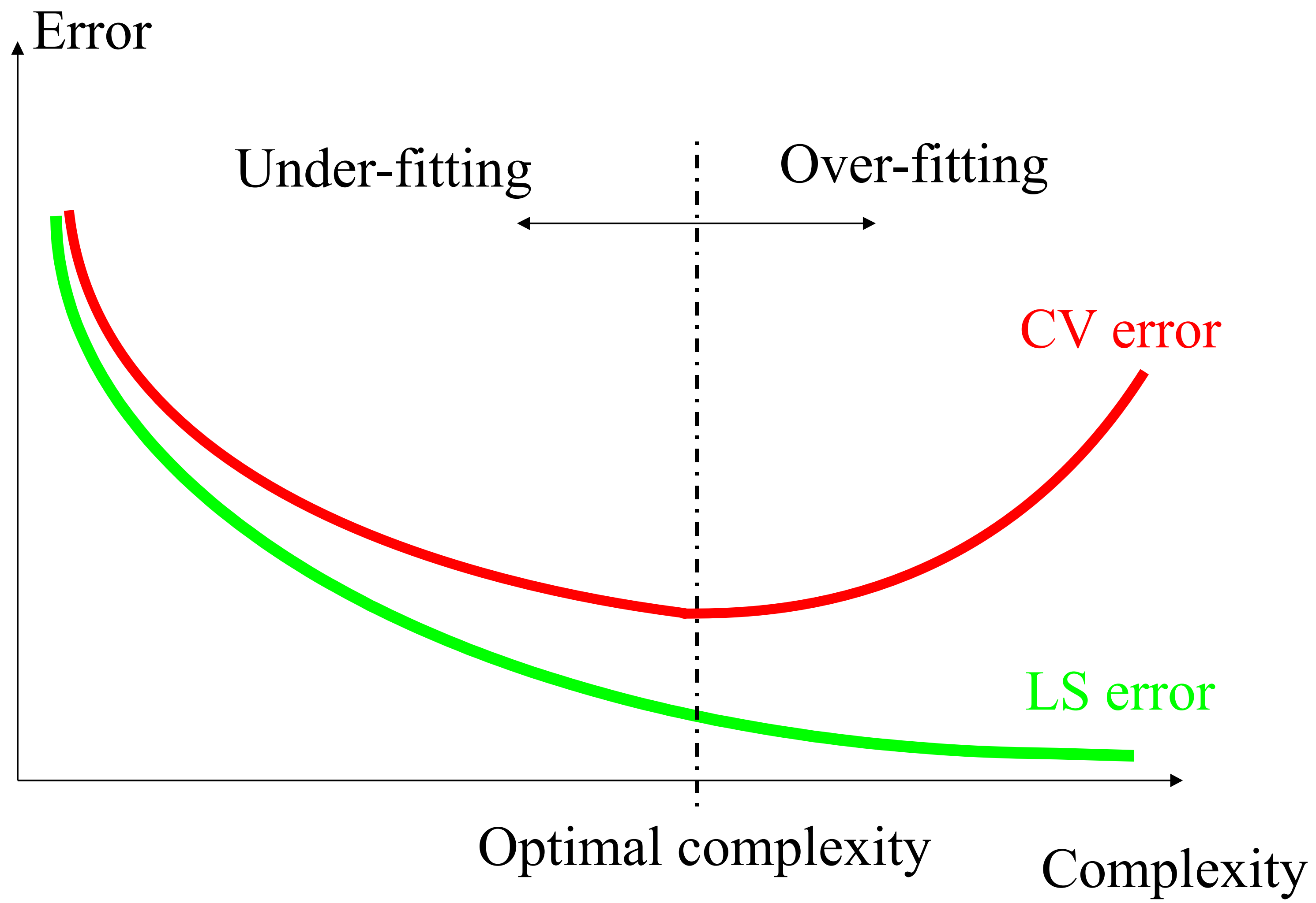


- For each subset:

  learn the model on the objects that are not in the subset

  compute the error rate on the points in the subset
- Report the mean error rate over the *k* subsets

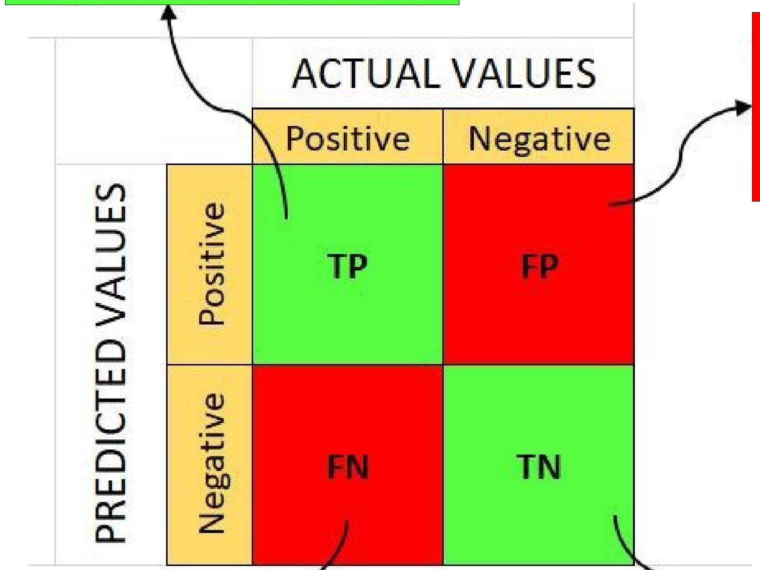
Complexity control



Performance measures

In binary classification, results can be summarized in a contingency table (aka confusion matrix)

True positives:
Both the predicted and true values are positive



False positives:
The predicted value is positive, but true value is negative

False negatives:
The predicted value is negative, but true value is positive

True negatives:
Both the predicted and true values are negative

Various criteria

Error rate = (*FP*+*FN*)/(*N*+*P*)

Accuracy = (*TP*+*TN*)/(*N*+*P*) = 1- Error rate

Recall = TP/P

Specificity = *TN*/(*TN*+*FP*)

Precision = *TP*/(*TP*+*FP*)

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Presentation plan

The interdisciplinary team:
Lessons learned – doing inter-disciplinary research

The tools:
Crash course: how to analyse social media data? what is data, what is an API, what is a classifier, and how do you measure performance?

The research:
Slipping to the extremes: combining qualitative research and computer science to fight problematic speech

Slipping to the Extreme: A Mixed Method to Explain How Extreme Opinions Infiltrate Online Discussions

Quyu Kong,[1,2] Emily Booth,[2] Francesco Bailo,[2] Amelia Johns,[2] Marian-Andrei Rizoiu[1,2]

[1] Australian National University
[2] University of Technology Sydney
quyu.kong@anu.edu.au, emily.booth@uts.edu.au, francesco.bailo@uts.edu.au, amelia.johns@uts.edu.au,
marian-andrei.rizoiu@uts.edu.au

Abstract

Qualitative research provides methodological guidelines for observing and studying communities and cultures on online social media platforms. However, such methods demand considerable manual effort from researchers and may be overly focused and narrowed to certain online groups. In this work, we propose a complete solution to accelerate qualitative analysis of problematic online speech — with a specific focus on opinions emerging from online communities — by leveraging machine learning algorithms. First, we employ qualitative methods of deep observation for understanding problematic online speech. This initial qualitative study constructs an ontology of problematic speech, which contains social media postings annotated with their underlying opinions. The qualitative study also dynamically constructs the set of opinions, simultaneous with labeling the postings. Next, we collect a large dataset from three online social media platforms (Facebook, Twitter and Youtube) using keywords. Finally, we introduce an iterative data exploration procedure to augment the dataset. It alternates between a data sampler, which balances exploration and exploitation of unlabeled data, the automatic labeling of the sampled data, the manual inspection by the qualitative mapping team and, finally, the retraining of the automatic opinion classifier. We present both qualitative and quantitative results. First, we present detailed case studies of the dynamics of problematic speech in a far-right Facebook group, exemplifying its mutation from conservative to extreme. Next, we show that our method succesfully learns from

and Vraga 2018) being recorded in the literature. To date, there exist three primary types of methods for addressing problematic information. The first type concentrated on large-scale monitoring of social media datasets to detect inauthentic accounts (bots and trolls) (Ram, Kong, and Rizoiu 2021) and coordinated disinformation campaigns (Rizoiu et al. 2018). The second group aims to understand which platforms, users, and networks contribute to the "infodemic" (Smith and Graham 2019; Bruns, Harrington, and Hurcombe 2020; Colley and Moore 2020). The third group uses computational modeling to predict future pathways and how the information will spread (Molina et al. 2019). These studies provide valuable insights into understanding how problematic information spreads and detecting which sources are reshared frequently and by which accounts. Though the first and third research approaches offer breadth of knowledge and understanding, there are limitations — they often have less to say about why certain opinions and views gain traction with vulnerable groups and online communities.

Qualitative research methods are well placed to address this gap. They provide rich, contextual insights into the social beliefs, values, and practices of online communities, which shape how information is shared and how opinions are formed (Glaeser and Sunstein 2009; Boyd 2010; Baym 2015; Johns 2020). This is also fundamental to un-

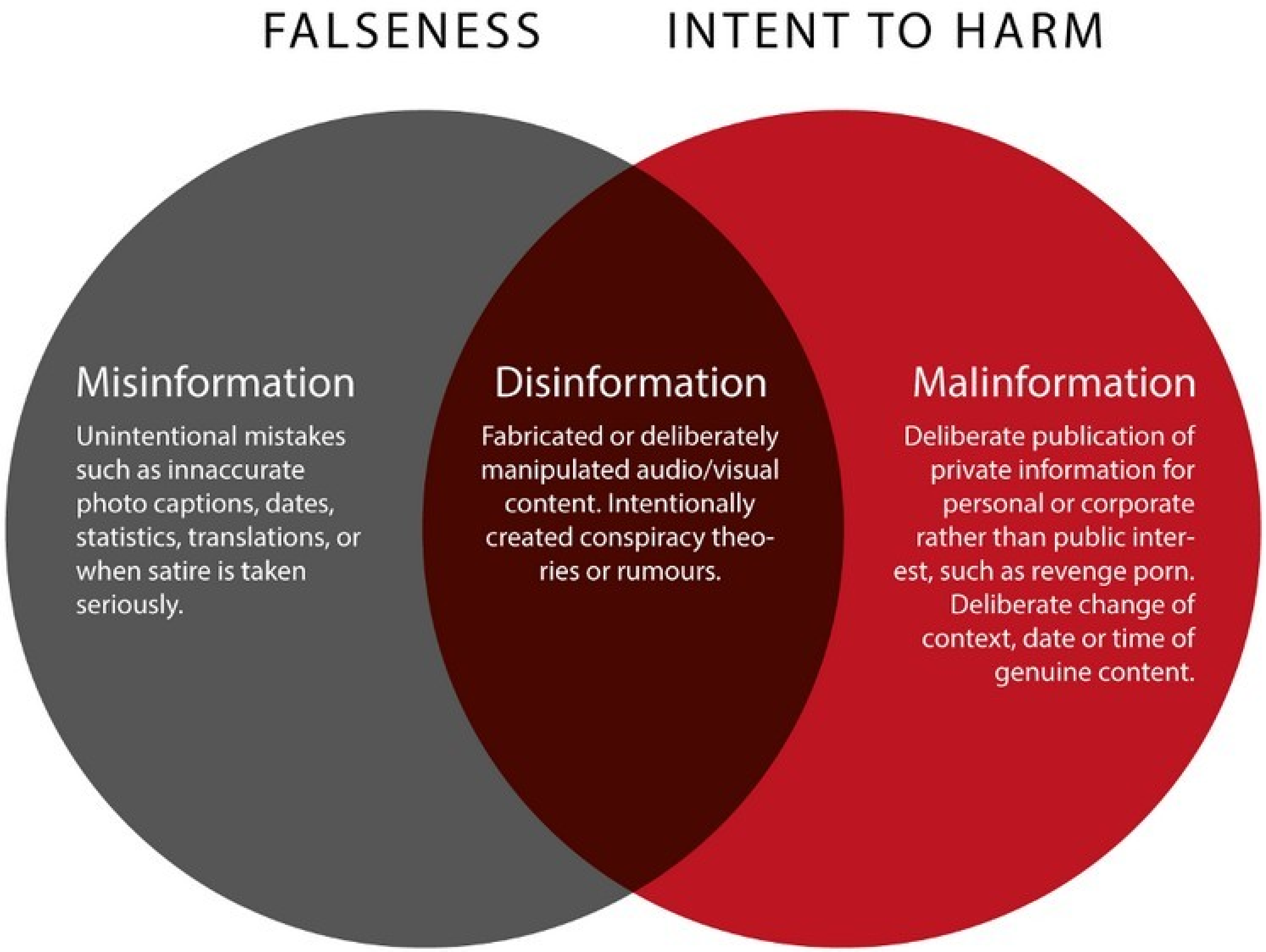
Kong, Q., Booth, E., Bailo, F., Johns, A., & Rizoiu, M.-A. (2022). Slipping to the Extreme: A Mixed Method to Explain How Extreme Opinions Infiltrate Online Discussions. In AAAI International Conference on Web and Social Media. Retrieved from http://arxiv.org/abs/2109.00302

Motivation

Problematic speech is online interactions, speech, and artefacts that are inaccurate, misleading, inappropriately attributed, or altogether fabricated (Jack 2017).

- misinformation
- disinformation
- hate speech

Types of information disorder



The gap of methods

Computational and quantitative

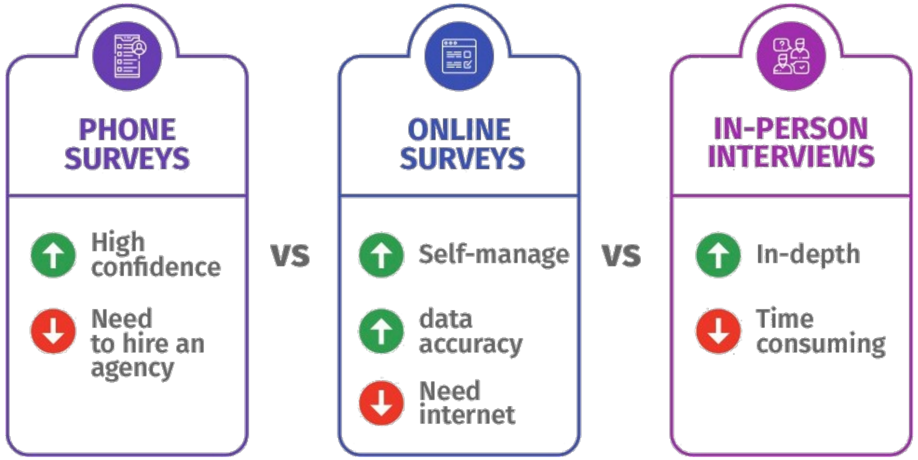
Large-scale monitoring of social media datasets [Kong et al, CIKM'20] [Ram et al, WSDM'21]

Identify platforms, users, and networks that contribute to the "infodemic" [Smith and Graham 2019] [Bruns et al 2020]

Future information spread [Molina et al. 2019]

The gap of methods

Computational and quantitative



Qualitative and ethnographic

Large-scale monitoring of social media datasets [Kong et al, CIKM'20] [Ram et al, WSDM'21]

Identify platforms, users, and networks that contribute to the "infodemic" [Smith and Graham 2019] [Bruns et al 2020]

Future information spread [Molina et al. 2019]

How information is shared and how opinions are formed [Boyd 2010] [Baym 2015]

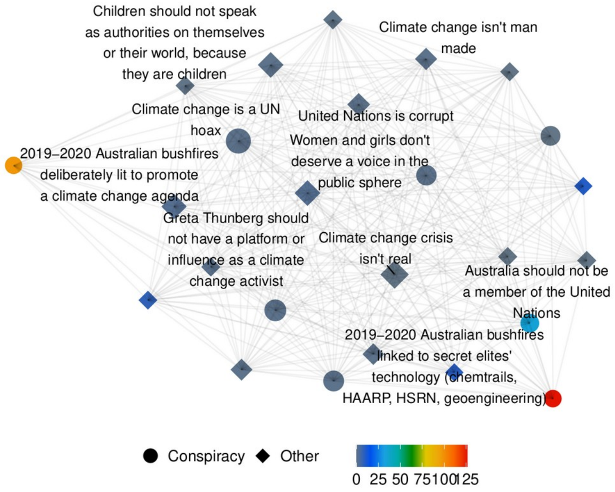
Why opinions and information sources scale to encompass large segments of the online society [Bailo 2020] [Bruns et al 2020]

Research questions

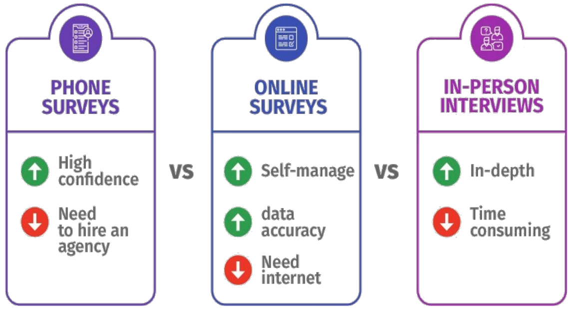
Can we leverage both qualitative and quantitative analysis for studying problematic online speech?

Can we accelerate qualitative research and observations of online behavior with machine learning algorithms?

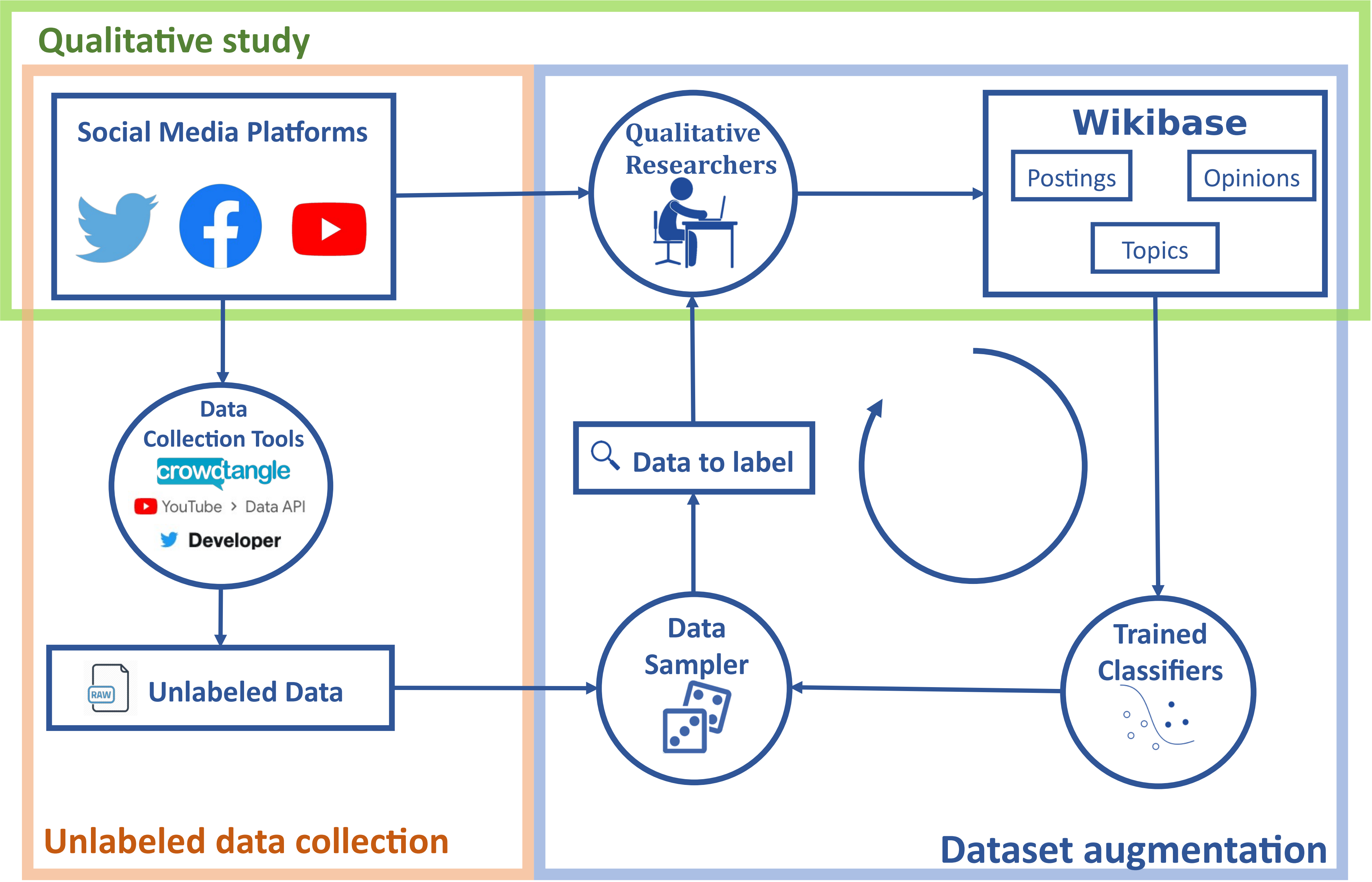
Can we track the dynamics of problematic opinions from online discussions using unlabeled data?



+



Presentation Plan / Overall Approach



1. Qualitative Study

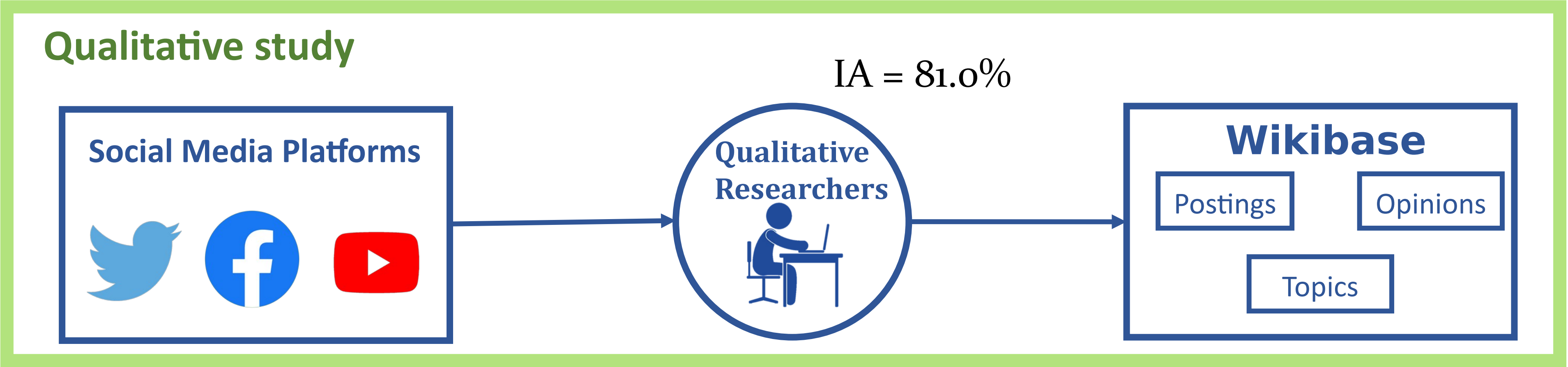
Four topics:

- 2019-20 Australian bushfire season
- Climate change
- COVID-19
- Vaccination

Dec 2019 – Jan 2021

Internet places:

- News stories
- Facebook page monitoring
- Cross-page link tracking
- Platform recommender systems



614 postings and 65 opinions:

- Climate change crisis isn't real
- United Nations is corrupt
- Climate change is a UN hoax
- United Nations want to be the global ruling government
- Experts manipulate data for private or corporate agendas
- Vaccines cause Autism
- The World Health Organization is corrupt
- Men are being chemically emasculated by the government/science/elites
- Covid-19 is the Chinese government's bioweapon

Take Australia Back - Public Facebook group, 11.2K members.

Sample post and comments 1: Jan 10 2020

January 10 ·

Apparently climate change is real 🤷‍♂️
Apparently half of this group are smart enough to disprove the fact tho 🤦‍♂️

42 · 220 Comments 1 Share

[redacted]
#reaserchgeoengineering

Like · 35w

[redacted]
I don't understand any of it but I would like to understand Glaciers and glacial valleys and why in 200 years the sea level in Sydney is exactly the same ie goat island

Like · 35w

7 Replies

[redacted]
Why im a climate change sceptic.

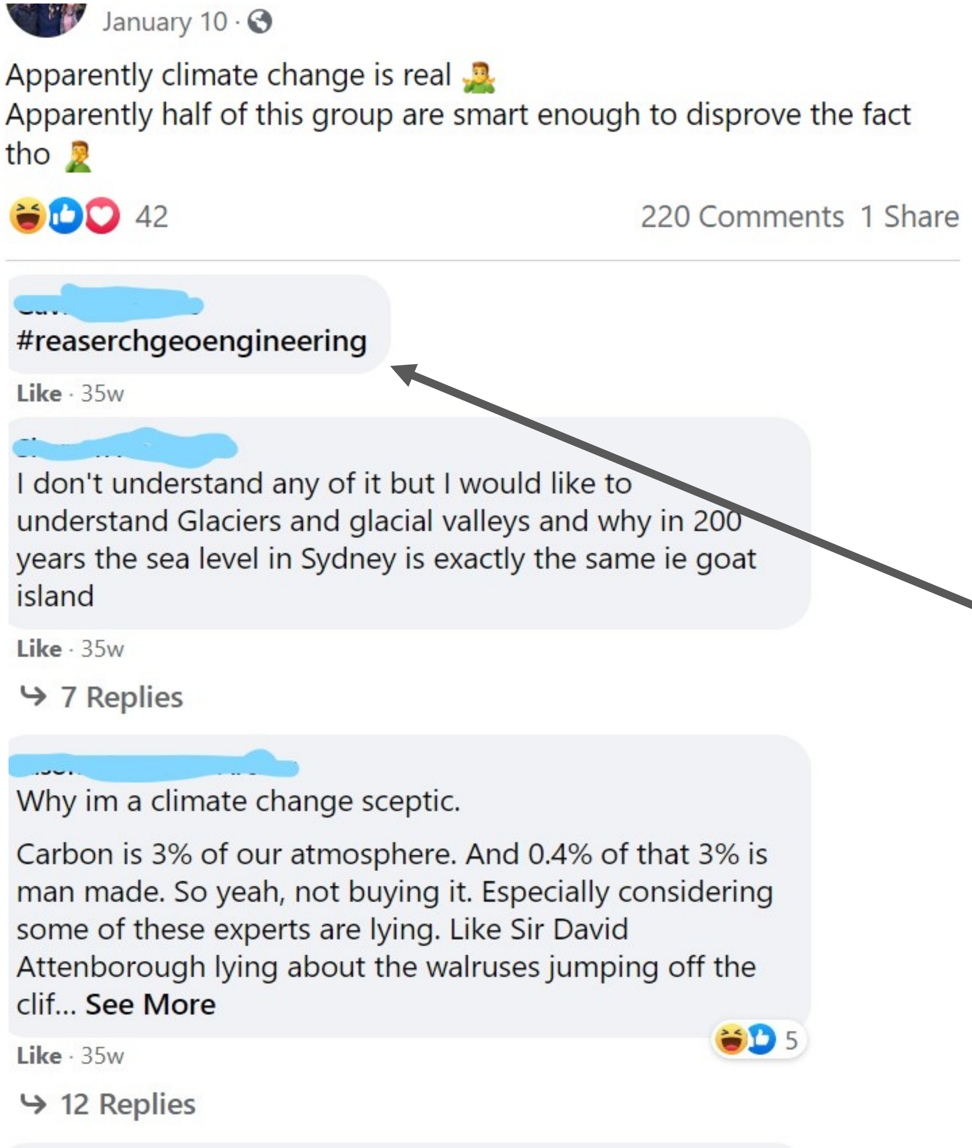
Carbon is 3% of our atmosphere. And 0.4% of that 3% is man made. So yeah, not buying it. Especially considering some of these experts are lying. Like Sir David Attenborough lying about the walruses jumping off the clif... **See More**

Like · 35w · 5

12 Replies

Take Australia Back - Public Facebook group, 11.2K members.

Sample post and comments 1: Jan 10 2020



- 50/50 climate change denial and support
- Some respectful debate but mainly polarising contest and troll-like social practices
- Use of misogynistic and ableist abuse to inflame/polarise/derail opposing opinion
- Small number of conspiracy theories (e.g. chemtrails)
- 40-60+ user group
- Text based comments, few links out, more comments than shares

Sample post and comments 2: 16 September 2020

84 · 9 Comments 58 Shares

Like · Share

1

Like · 5h

Yeah, you can just remove the cloth masks anytime you want. And also they don't silence you as much as muffle your voice.

Like · 5h · Edited

But it covers so much of your emotion and power. There is a reason men do this to women in Islam.

Like · 5h

Like · 1h

how true

I joined this group when we were fighting against scomo cause he's a dick head, now this group is full of dickheads

3

Like · 1h · Edited

Not too mention most of those iron masks had fukn locks on them

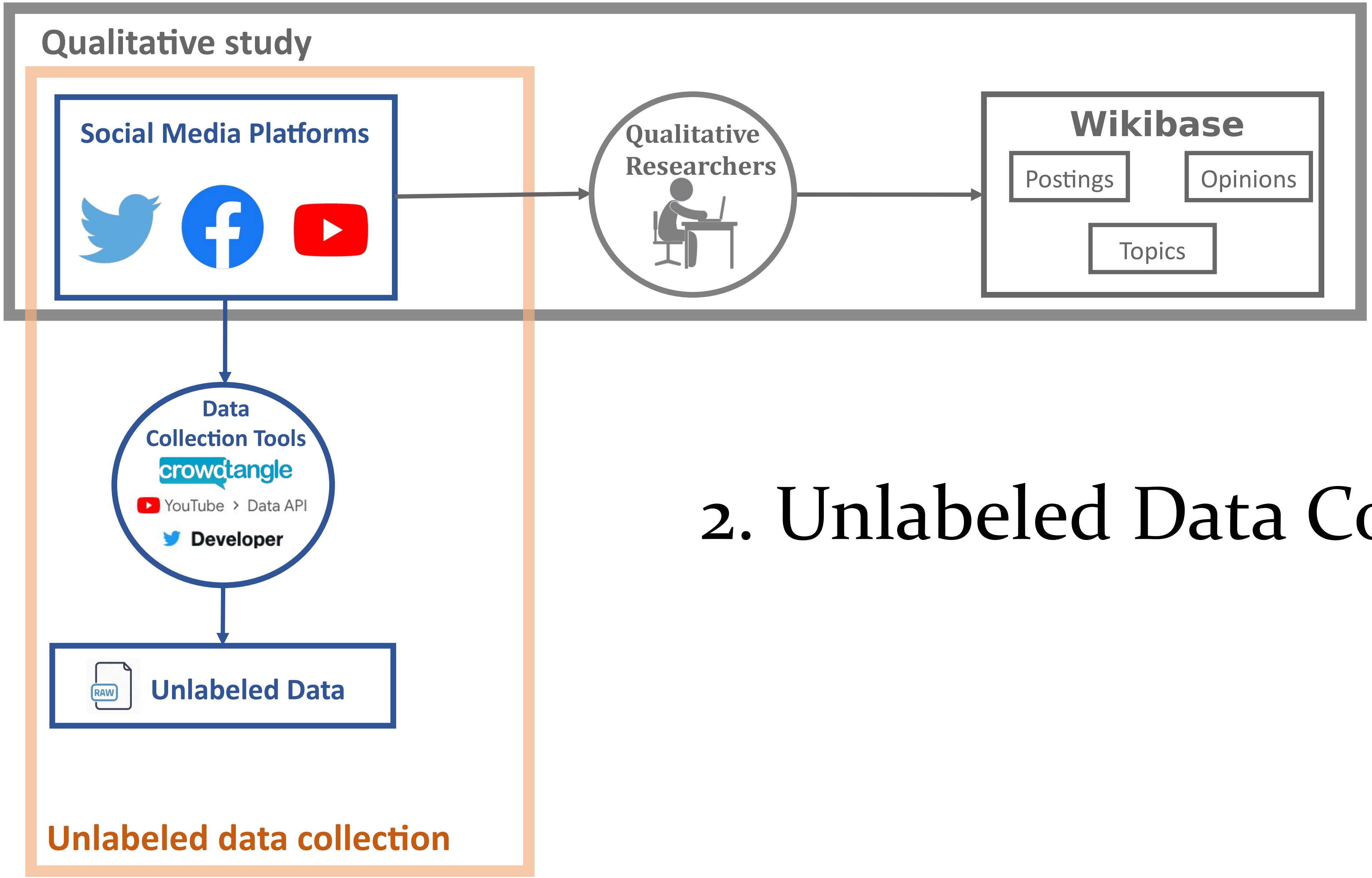
3

Like · 1h

its one of the first thing you learn on your road of indoctrination you mean. 😉

Like · 1h

realise who controls the puppets first.



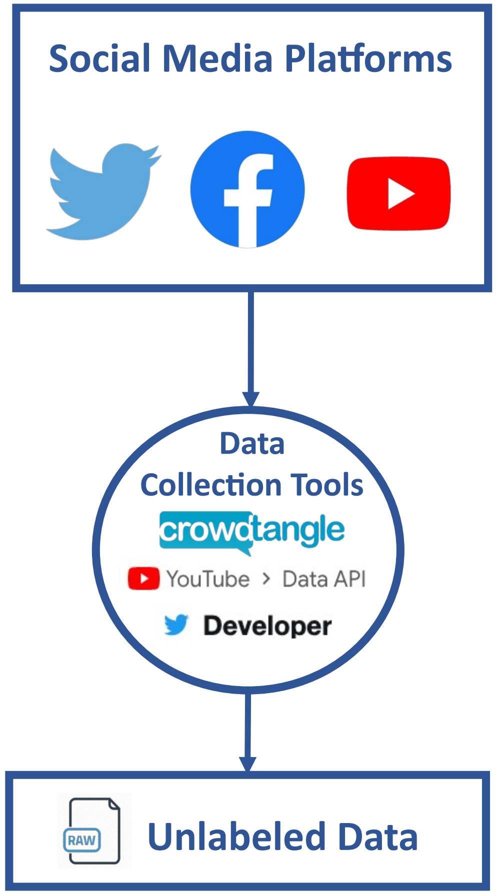
2. Unlabeled Data Collection

2. Unlabeled Data Collection

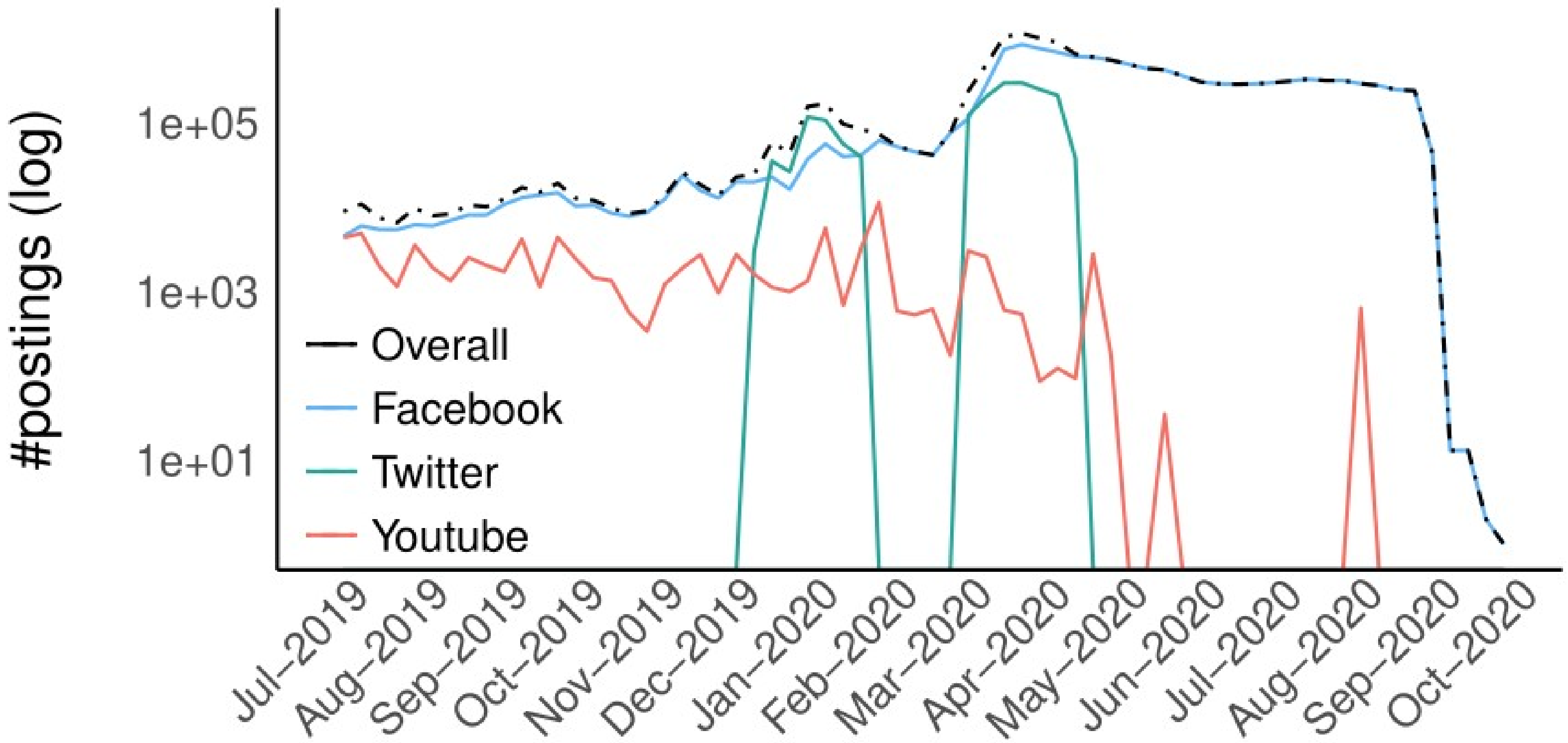
Topics	Selected keywords
2019-20 Australian bushfire season, Climate change	bushfire, australian fires, arson, scottyfrommarketing, liarfromtheshiar, australiaburns, australiaburning, itsthegreensfault, backburning, back burning, climate change, climate mergency, climate hoax, climate crisis, climate action now
Covid-19, Vaccination	covid, coronavirus, covid-19, pandemic, world health organization, vaccine, social distancing, quarantine, plandemic, chinavirus, wuhan, stayhome, MadeinChina, ChinaLiedPeopleDied, 5G, chinacentric

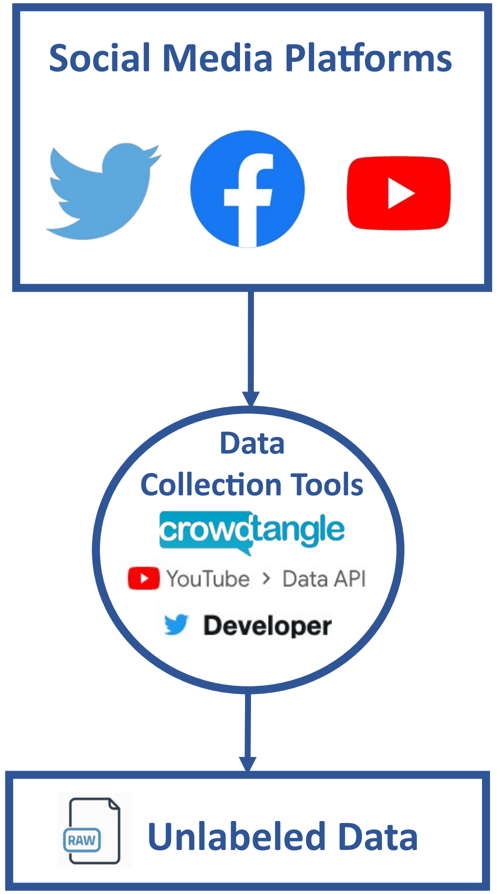
13.3M postings:

- 11.4M Facebook
- 1.8M Twitter
- 91K YouTube comments

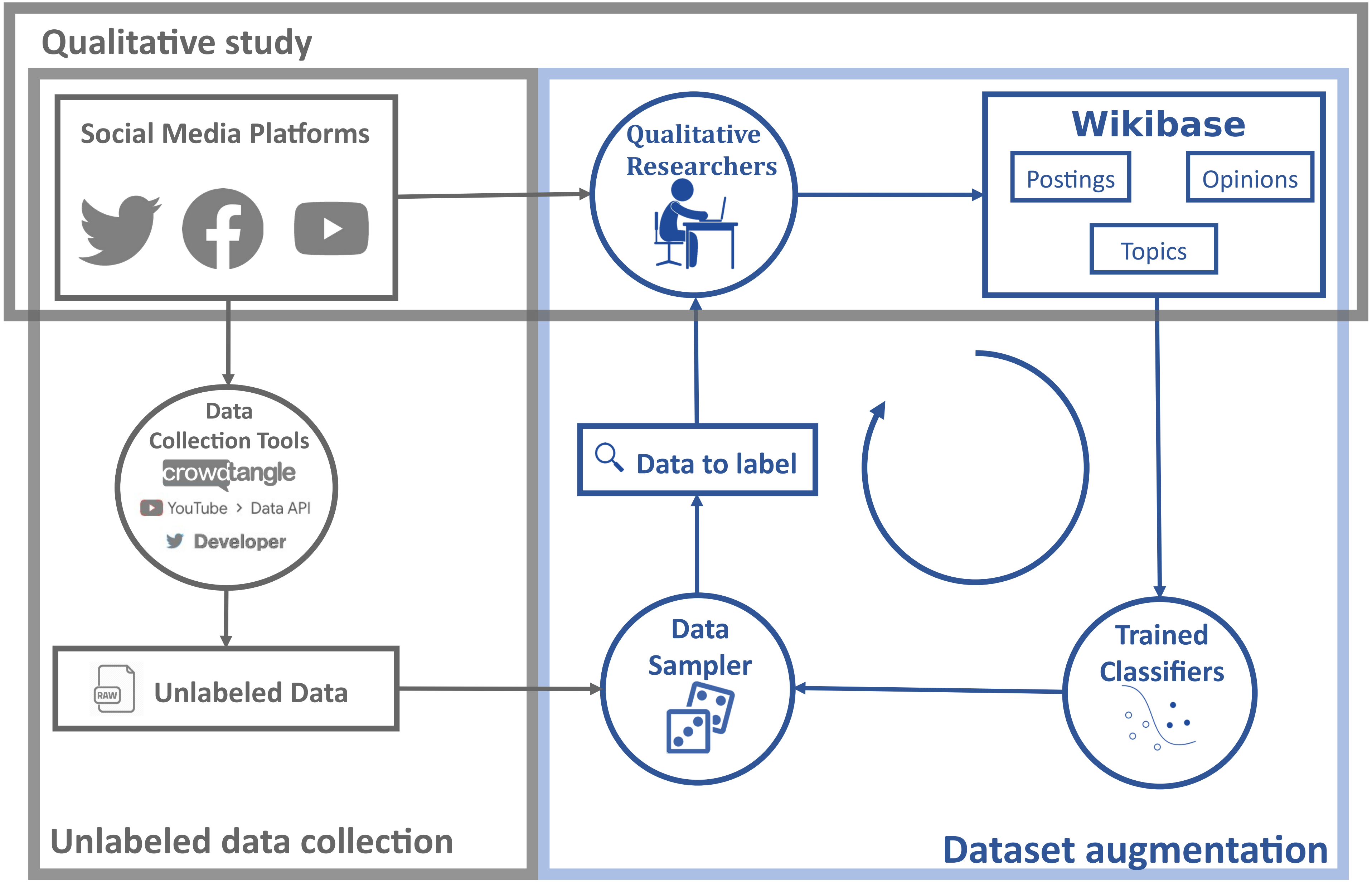


2. Unlabeled Data Collection



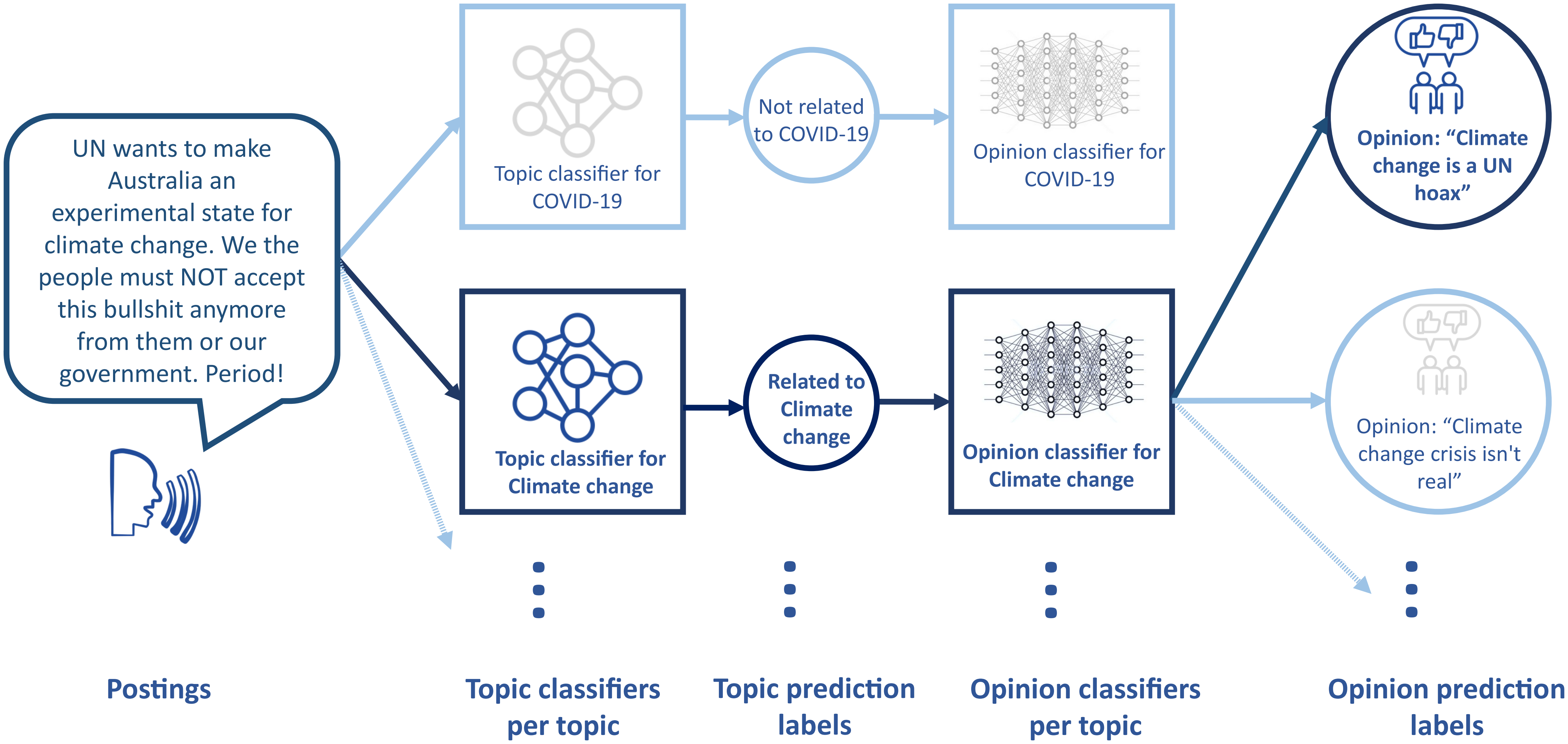


3. Dataset Augmentation



Two levels of classifiers

	RF	SVM	XGBoost	RoBERTa
Macro Accuracy	0.791	0.775	0.779	**0.800**
Macro F1	0.782	0.768	0.768	**0.800**

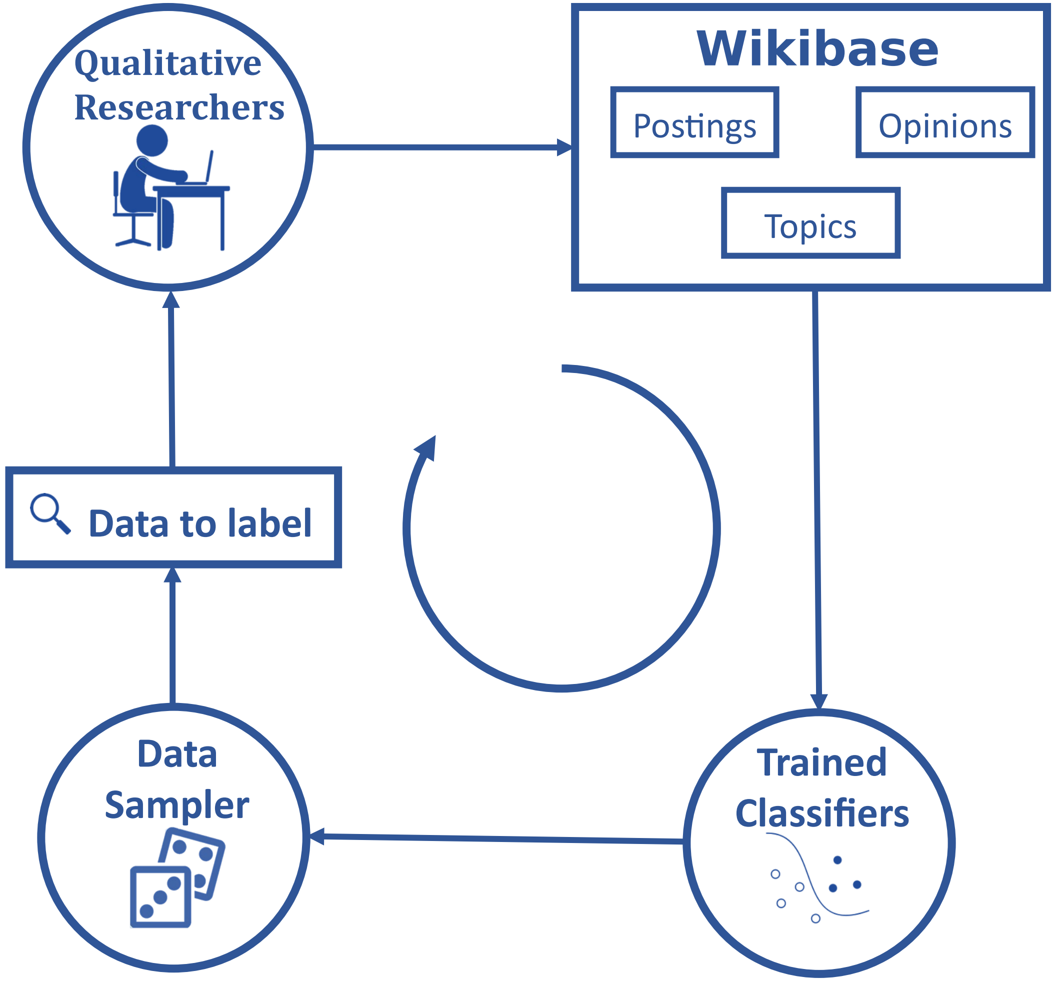


3. Dataset Augmentation

- Human-in-the-loop Machine Learning

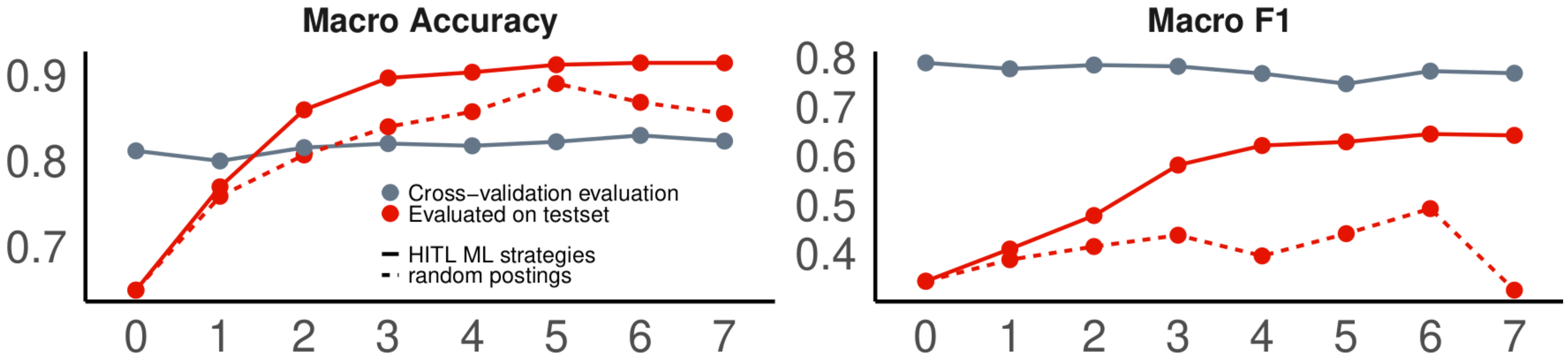
- Three strategies for data sampling:
  - Active learning
    10 posts / iteration / topic
    $u(\boldsymbol{x}) = 1 - p(\hat{y} \mid \boldsymbol{x}; f_{t,i})$
  - Top confidence
    10 posts / iteration / topic
  - Random sampling
    5 posts / iteration / topic

- Iterated until convergence
  - cross-validation error VS test set error
  - gain on test set between two iterations



Results

Human-in-the-loop performance



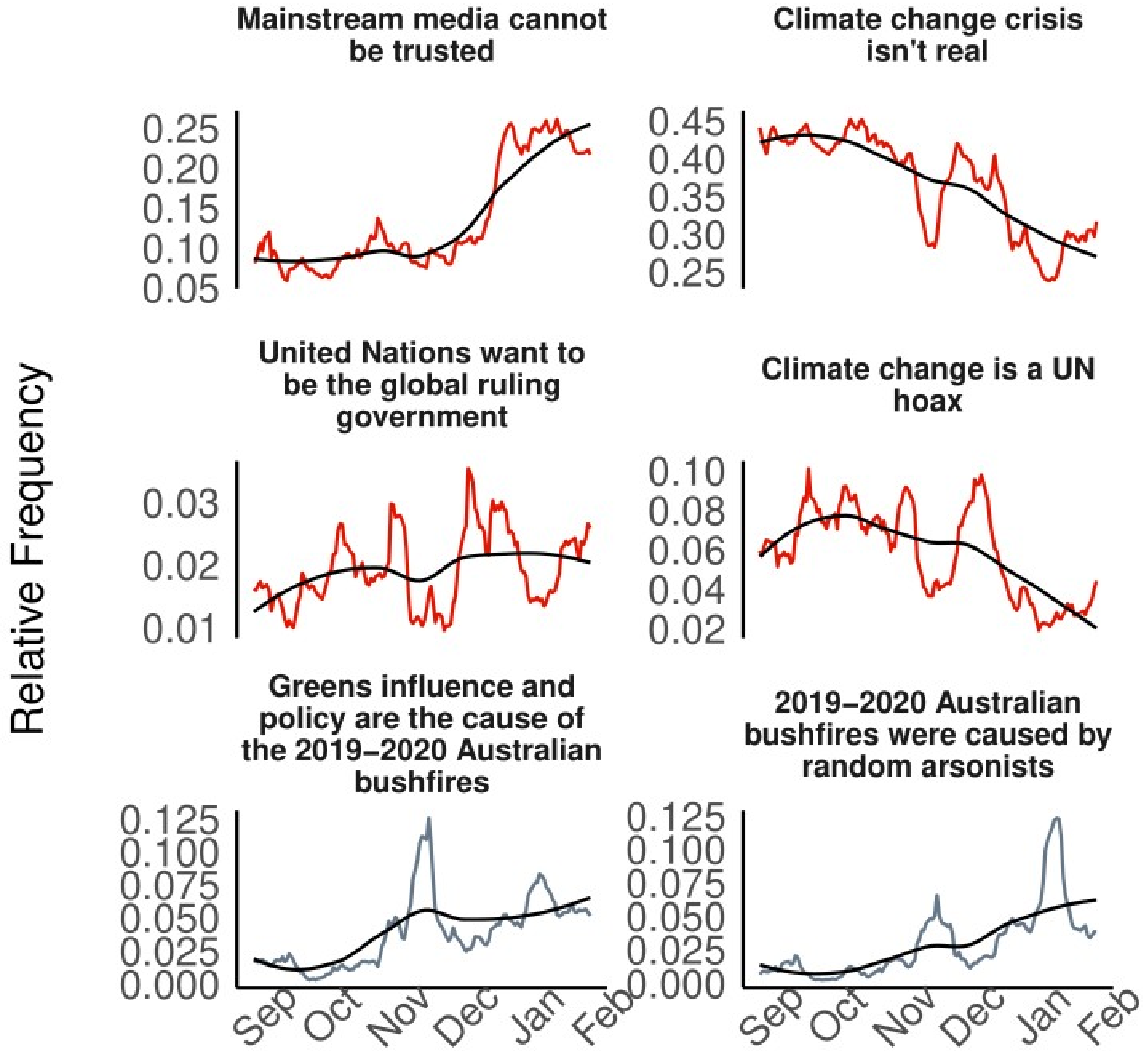
- Performances improve as more batches are performed
- Gap between generalization and test set error reduces
- Improvement plateaus as the process converges
- Human-in-the-Loop outperforms static random selection of samples

	L0	L7
#posts	614	1381
#opinions	65	71

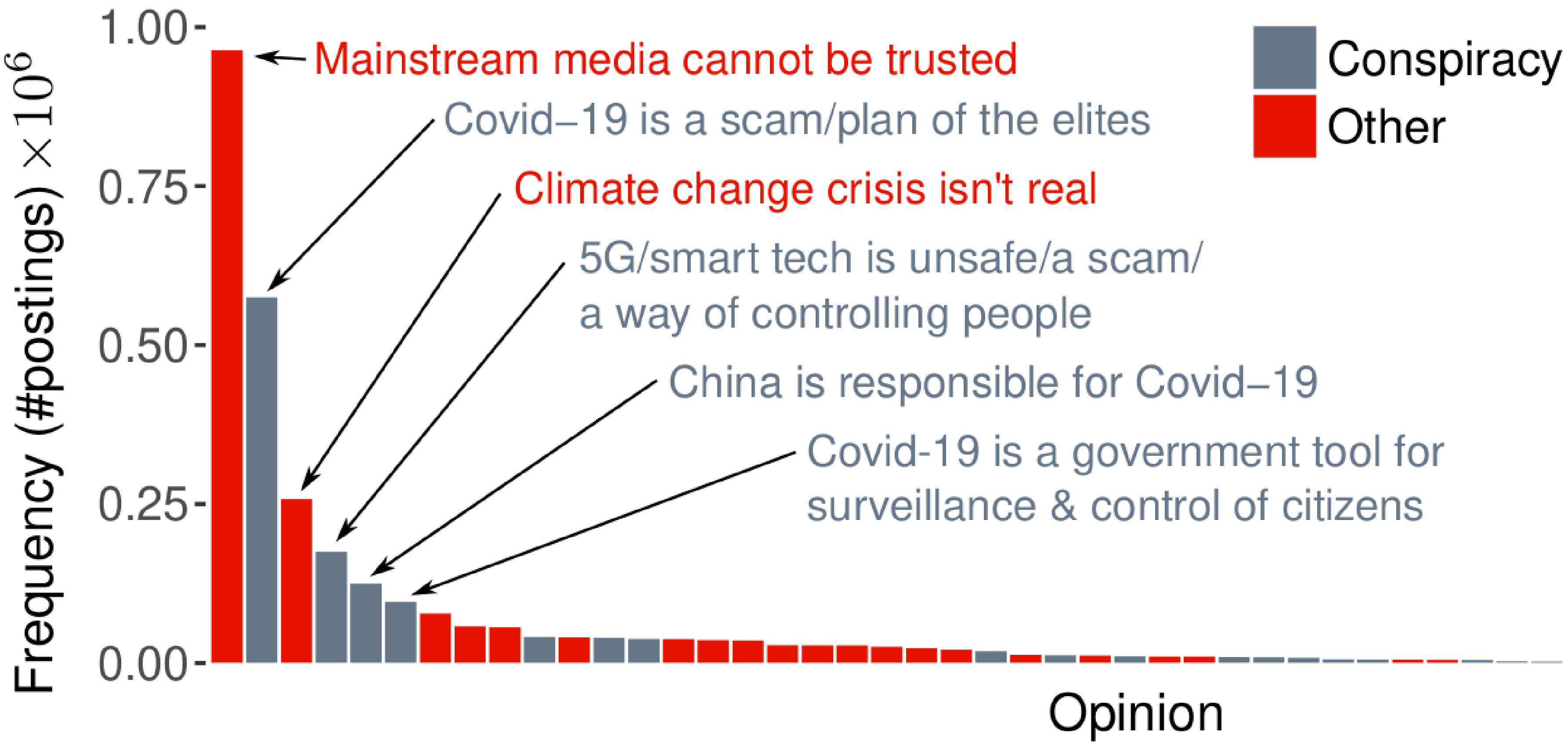
Opinion analysis at scale

Fully labeled dataset stats

- 1.7M postings with at least one opinion
- 314K postings with 2 or more opinions
- 21.26M off-topic postings
- **Total**: 22.96M postings

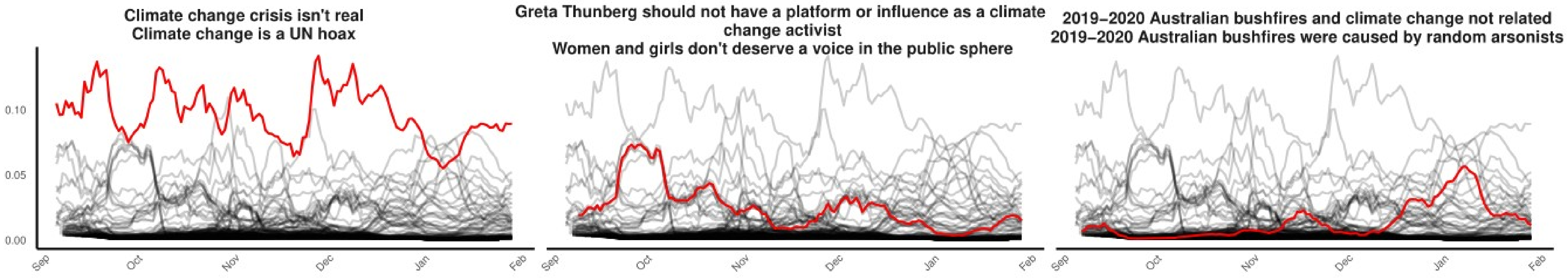


Opinion analysis at scale



- Opinion usage frequency is longtail distributed
- Four of the top six opinions endorse conspiracy theories

Opinion co-occurrence network



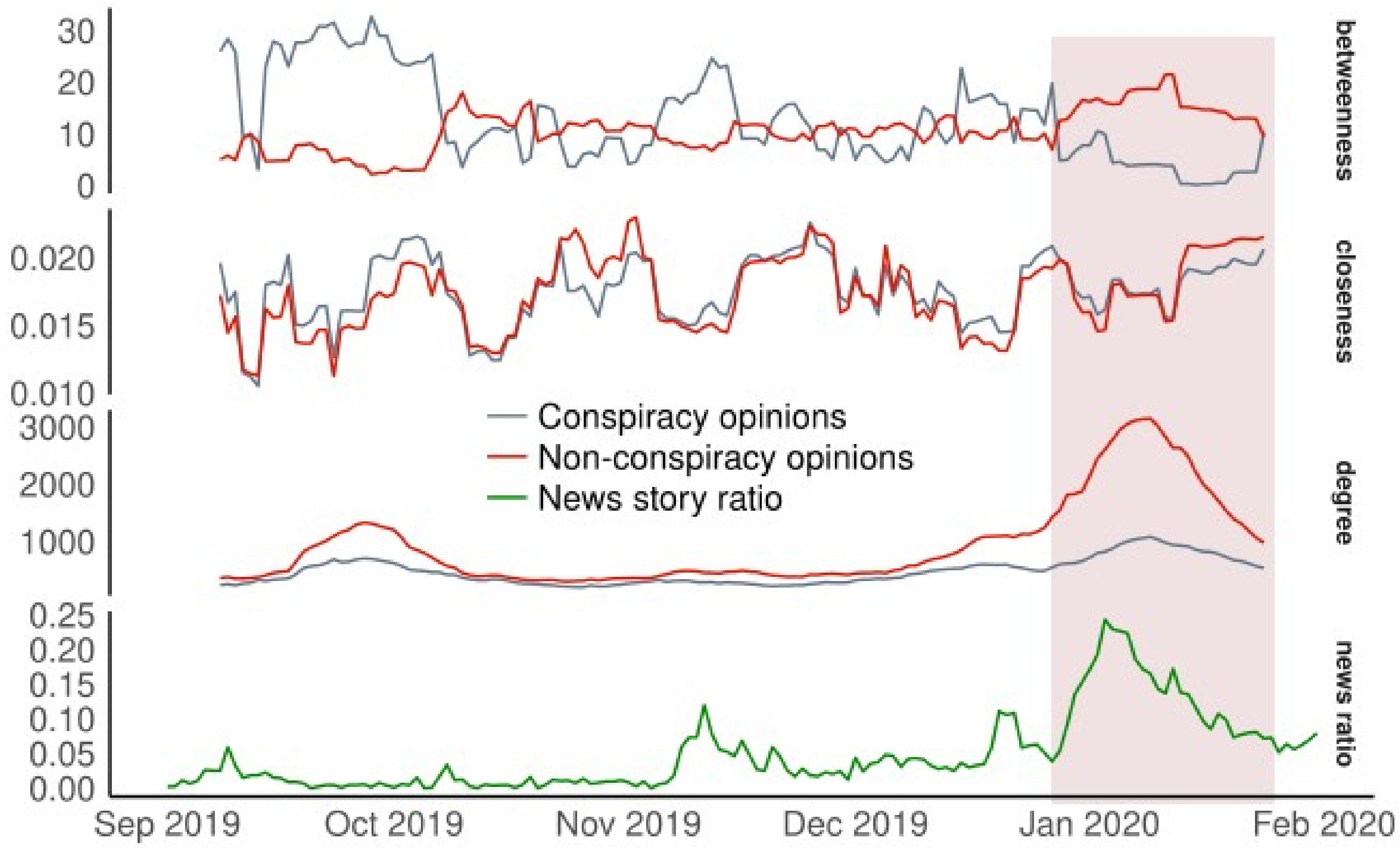
A continuous and relatively strong association between prevalent opinions

Associations with declining relative frequencies

Rising associations – early warnings for their adoption (and possibly normalization) by participants

Centrality of conspiracy opinions and news ratio

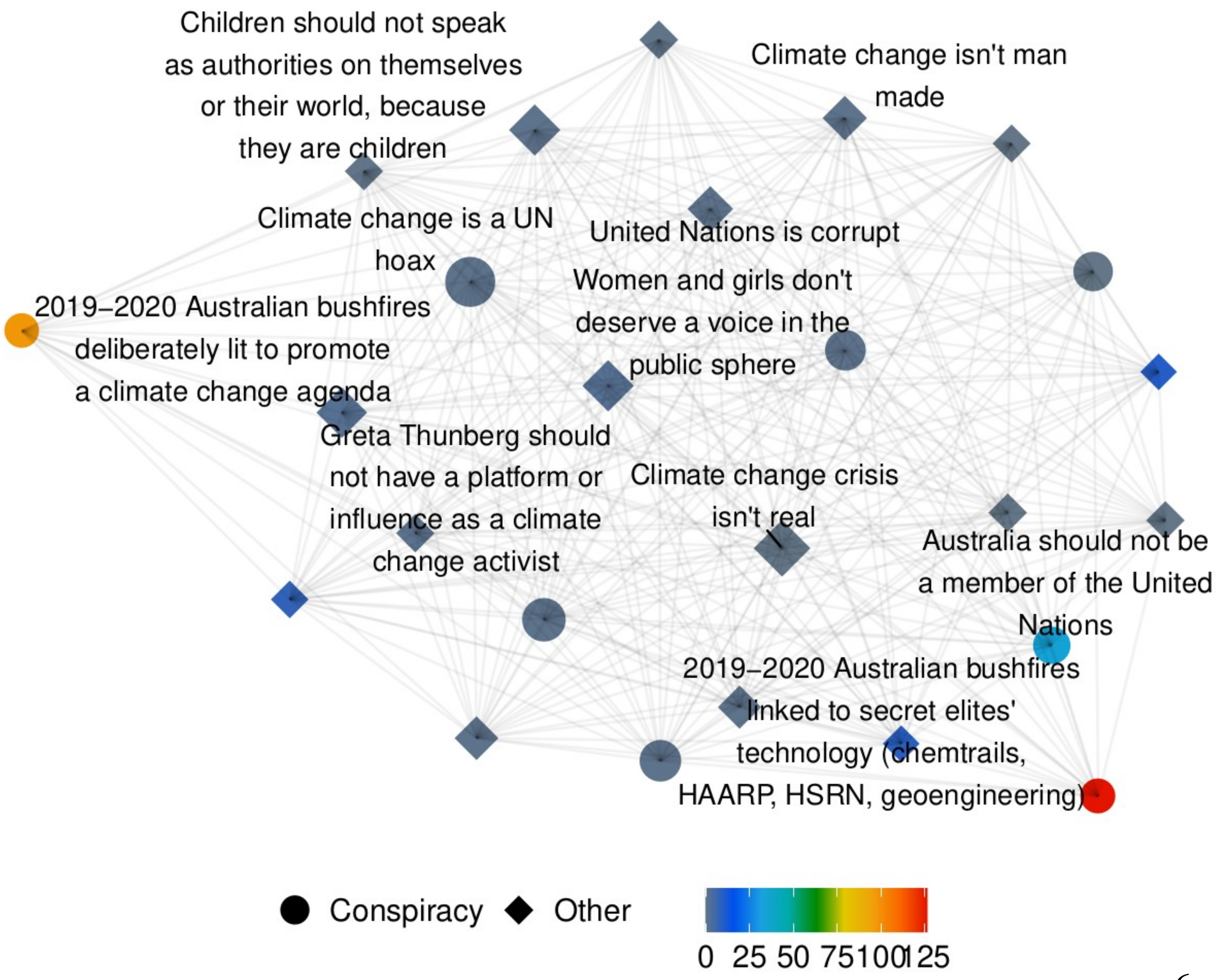
Higher coverage from news media reduces centrality of conspiracy opinions.



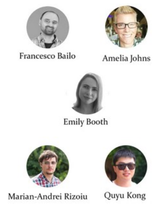
coverage ratios from Media Cloud
(Roberts et al. 2021)

Opinion co-occurrence network

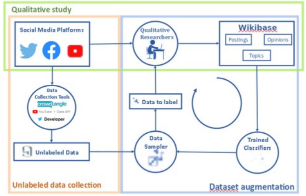
- High betweenness centrality of conspiracy opinions → selectively used in conjunction with many other opinions
- 14 days in late September 2019 – peak betweeness
- Conspiracy opinions are used together with mainstream opinions – rationalize and popularize them



Summary

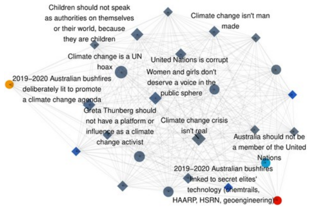


An inter-disciplinary team and methods to solve a difficult task: detecting and mapping the impact of online problematic content



A mixed qualitative and human-in-the-loop Machine Learning approach for detecting problematic content

A representative annotated dataset of online problematic content, qualitative and quantitative analyses



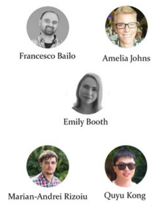
A hypothesis of how fringe opinions infiltrate mainstream discourse via co-occurrence with established opinions

Behavioral Data Science

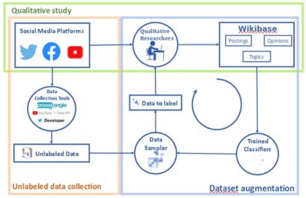
Thank you!

UTS Data Science

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

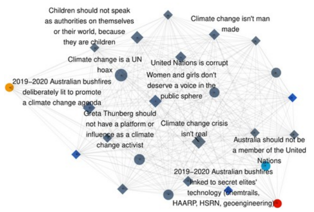


An inter-disciplinary team and methods to solve a difficult task: detecting and mapping the impact of online problematic content



A mixed qualitative and human-in-the-loop Machine Learning approach for detecting problematic content

A representative annotated dataset of online problematic content, qualitative and quantitative analyses



A hypothesis of how fringe opinions infiltrate mainstream discourse via co-occurrence with established opinions