Breaking free of the arms race

Monitor, detect, assess and react to influence operations

Data Science Institute

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

Red Queen effect

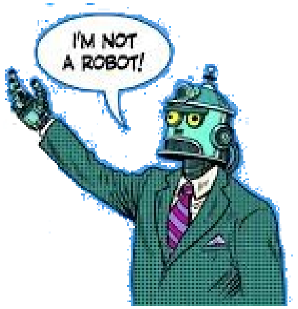
Content-based detectors are sensitive to adversarial training attacks – simply use the detector to train the attacker.

Our detection approach in a nutshell

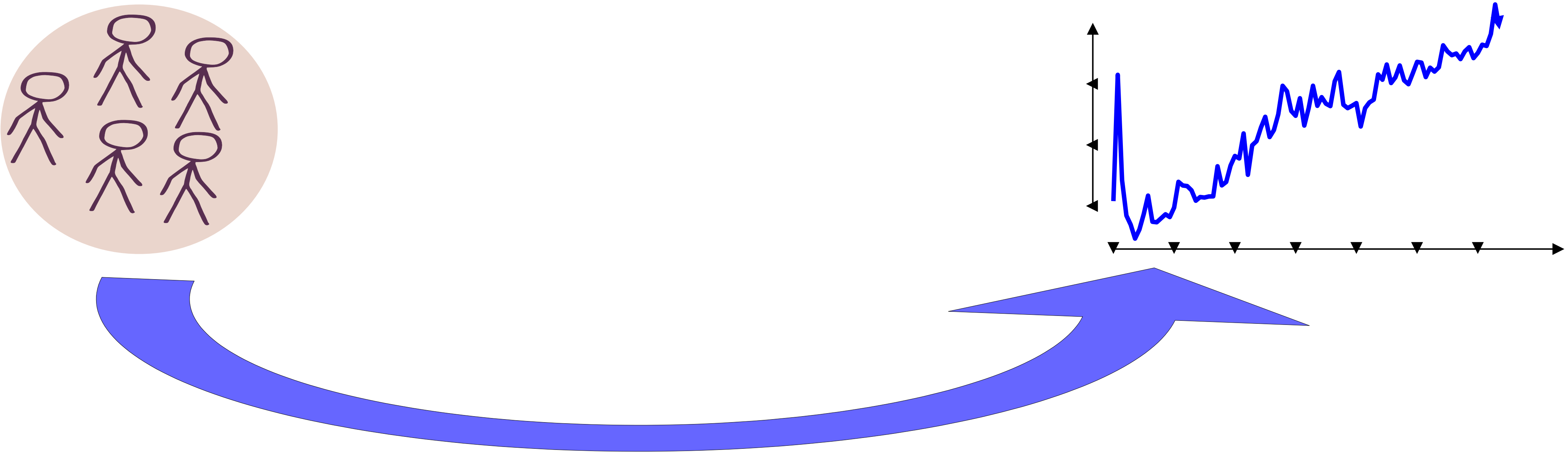
Build social sensors – the reaction of the social system cannot be faked

Early detection systems – information spread patterns within the user population

Distinguish content types and user actions – how online social systems react to them.

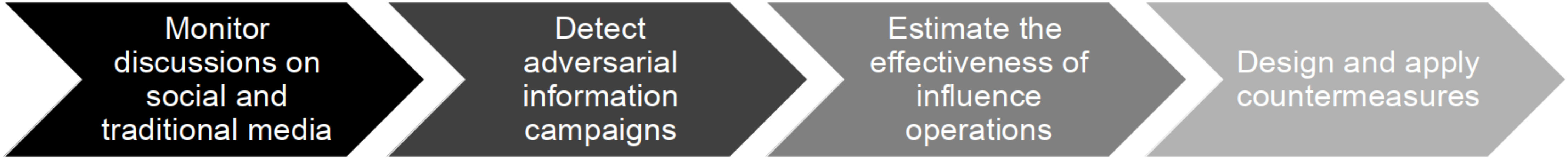


Our detection approach in a nutshell



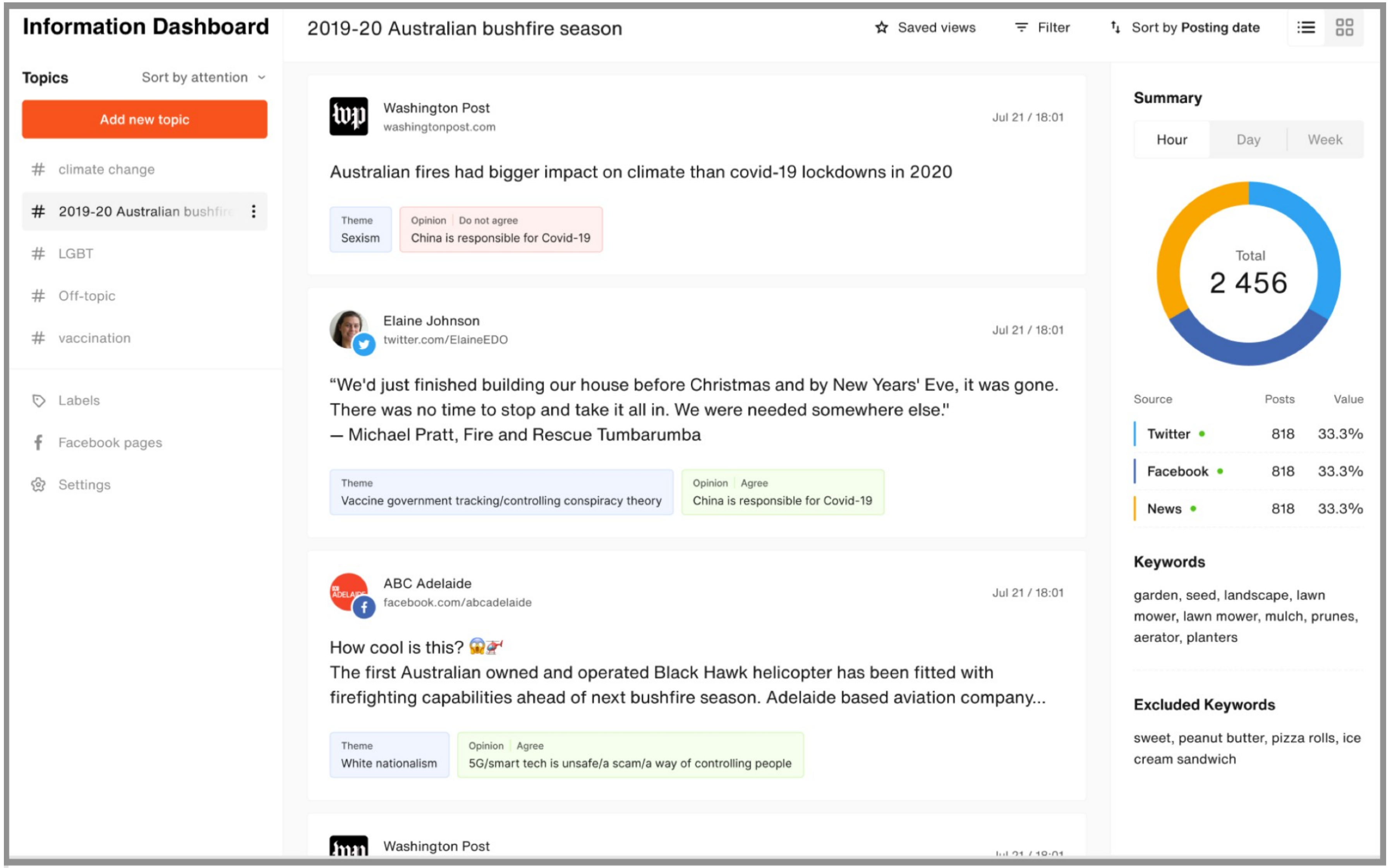
information diffusion
misinformation spread
influence operations

UTS capabilities in the Influence Operations space



Response level	Monitor	Detect	Counter	
Objective	How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?	How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?	What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?	What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?
	Monitor discussions on social and traditional media	Detect adversarial information campaigns	Estimate the effectiveness of influence operations	Design and apply countermeasures
Approach	Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content. Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.	Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message	Model the impact of networks and influencers on the virality and reach of problematic messages Track the spread of problematic messages across and between online platforms and into the real-world	Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest Identify key message inoculation points in social networks based on how information flows and gains velocity

Monitor: Monitoring discussion spaces (TRL: 3)



Washington Post
washingtonpost.com
Jul 21 / 18:01

Australian fires had bigger impact on climate than covid-19 lockdowns in 2020

Theme: Sexism
Opinion | Do not agree: China is responsible for Covid-19

Elaine Johnson
twitter.com/ElaineEDO
Jul 21 / 18:01

“We'd just finished building our house before Christmas and by New Years' Eve, it was gone. There was no time to stop and take it all in. We were needed somewhere else."
— Michael Pratt, Fire and Rescue Tumbarumba

Theme: Vaccine government tracking/controlling conspiracy theory
Opinion | Agree: China is responsible for Covid-19

ABC Adelaide
facebook.com/abcadelaide
Jul 21 / 18:01

How cool is this? 😱🚁
The first Australian owned and operated Black Hawk helicopter has been fitted with firefighting capabilities ahead of next bushfire season. Adelaide based aviation company...

Theme: White nationalism
Opinion | Agree: 5G/smart tech is unsafe/a scam/a way of controlling people

Washington Post
Jul 21 / 18:01

Source	Posts	Value
Twitter ●	818	33.3%
Facebook ●	818	33.3%
News ●	818	33.3%

Keywords

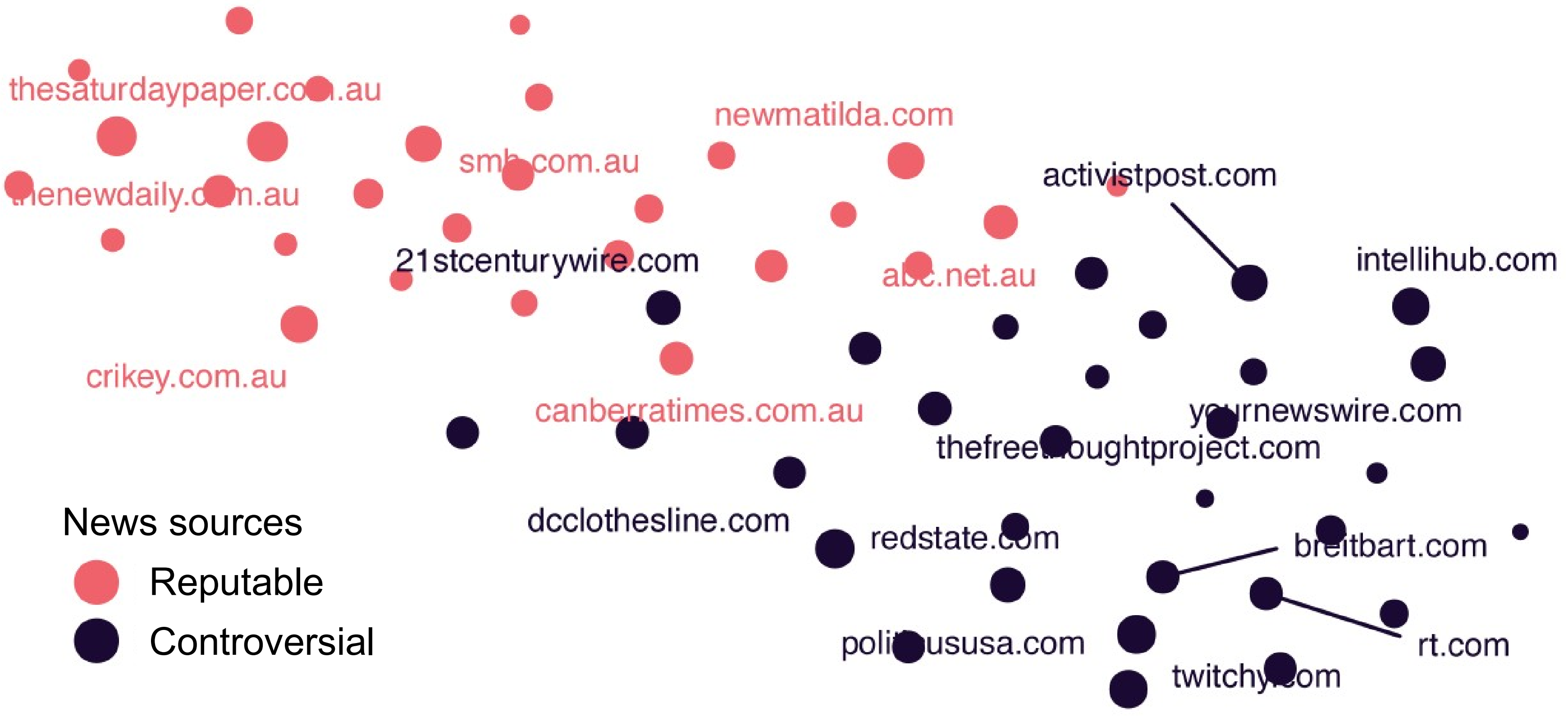
garden, seed, landscape, lawn mower, lawn mower, mulch, prunes, aerator, planters

Excluded Keywords

sweet, peanut butter, pizza rolls, ice cream sandwich

Graphical interface of the Information Dashboard

Detect: separating controversial from reputable (TRL: 5)

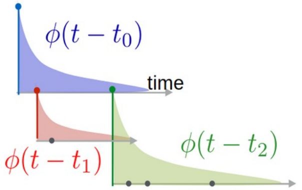


Reputable and controversial sources are separable based solely on how their information spreads

Detect controversial news without content analysis

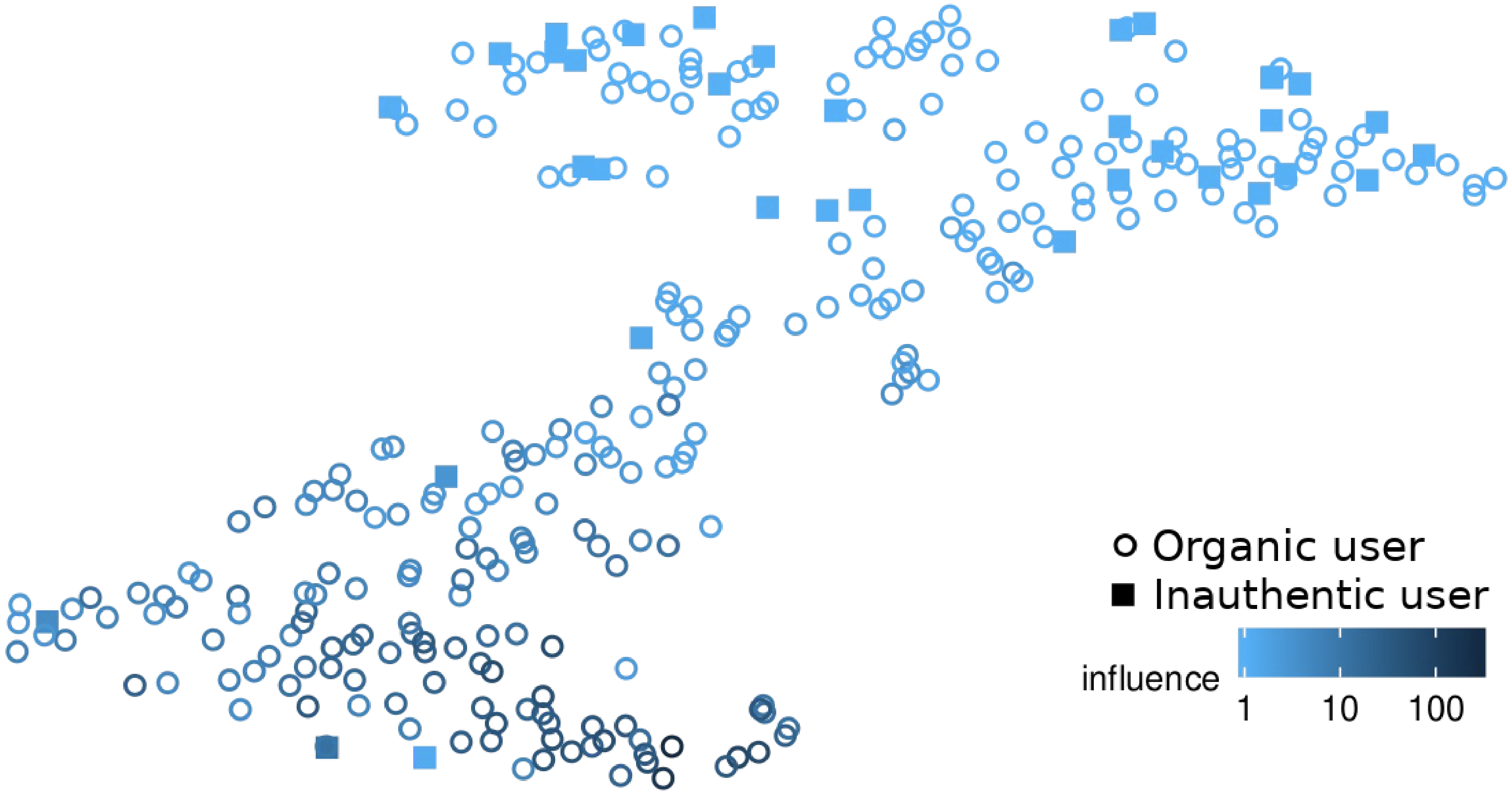
The technical detail:

Mathematical generative modelling; Hawkes processes; joint modelling



evently

React: Identify influential inauthentic users (TRL: 5)



Identify users engaged in influence operations

Estimate their impact on the wider community

The technical detail:

Influence estimation using stochastic modelling; content-free analysis

birdspotter

Prior expertise with Information Disorder

Australian Government
Department of Defence
Defence Science and
Technology Group

Real-time detection of
disinformation campaigns

Hate Speech propagation
on Social Media

FIRST
DRAFT

Detection and debunking
for online misinformation

NSW
GOVERNMENT

Expert roundtable for
Defamation law reform

Australian
National
University

CRAWFORD SCHOOL
OF PUBLIC POLICY

Tracking Disinformation
Campaigns across terrain

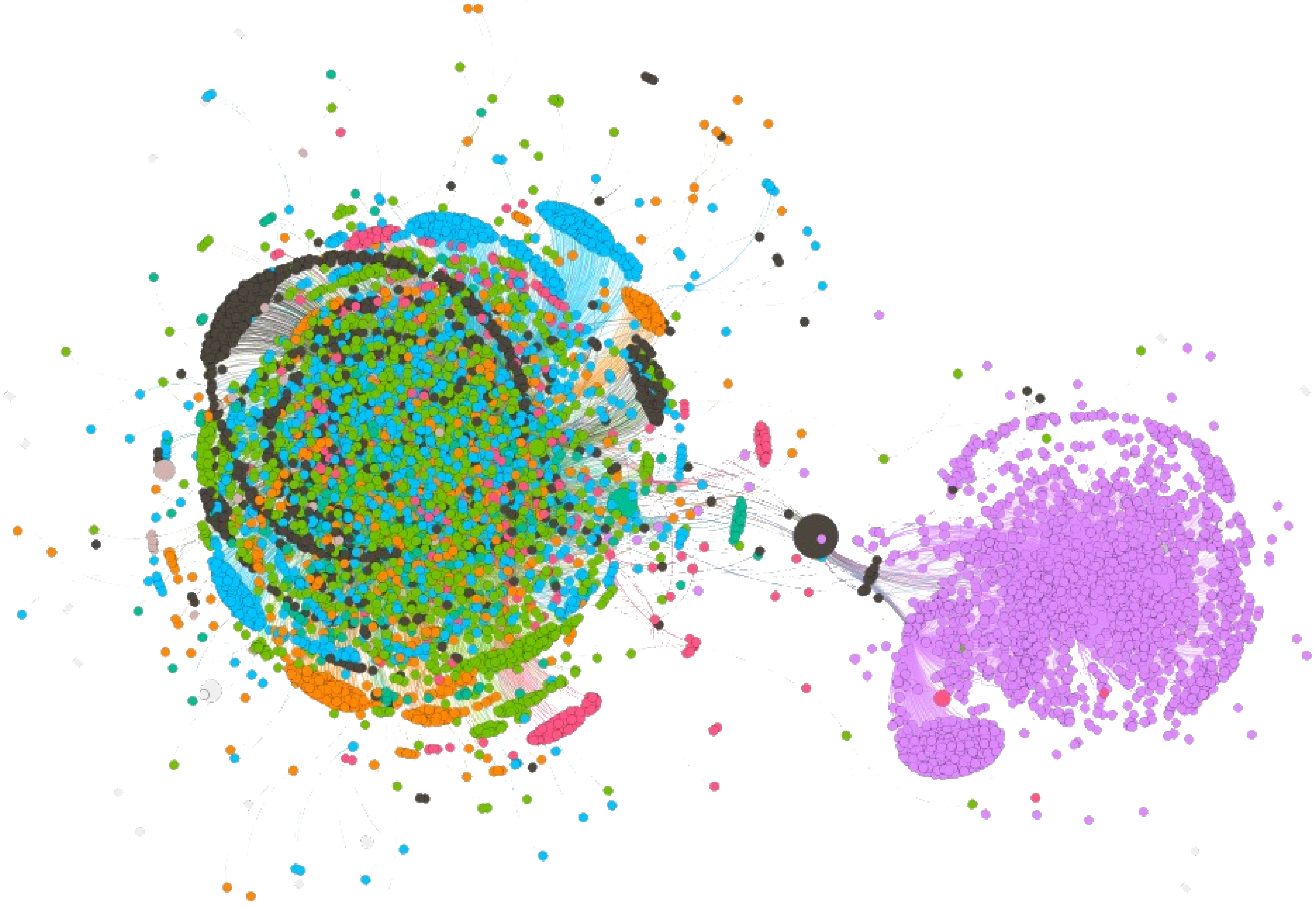
WIKIMEDIA
FOUNDATION

WIKIPEDIA
The Free Encyclopedia

Detecting and quantifying privacy
loss over time

Other examples of expertise

Expertise: Detecting coordinated campaigns



Clear structure with two clusters: disinformation (right) and debunking (left)

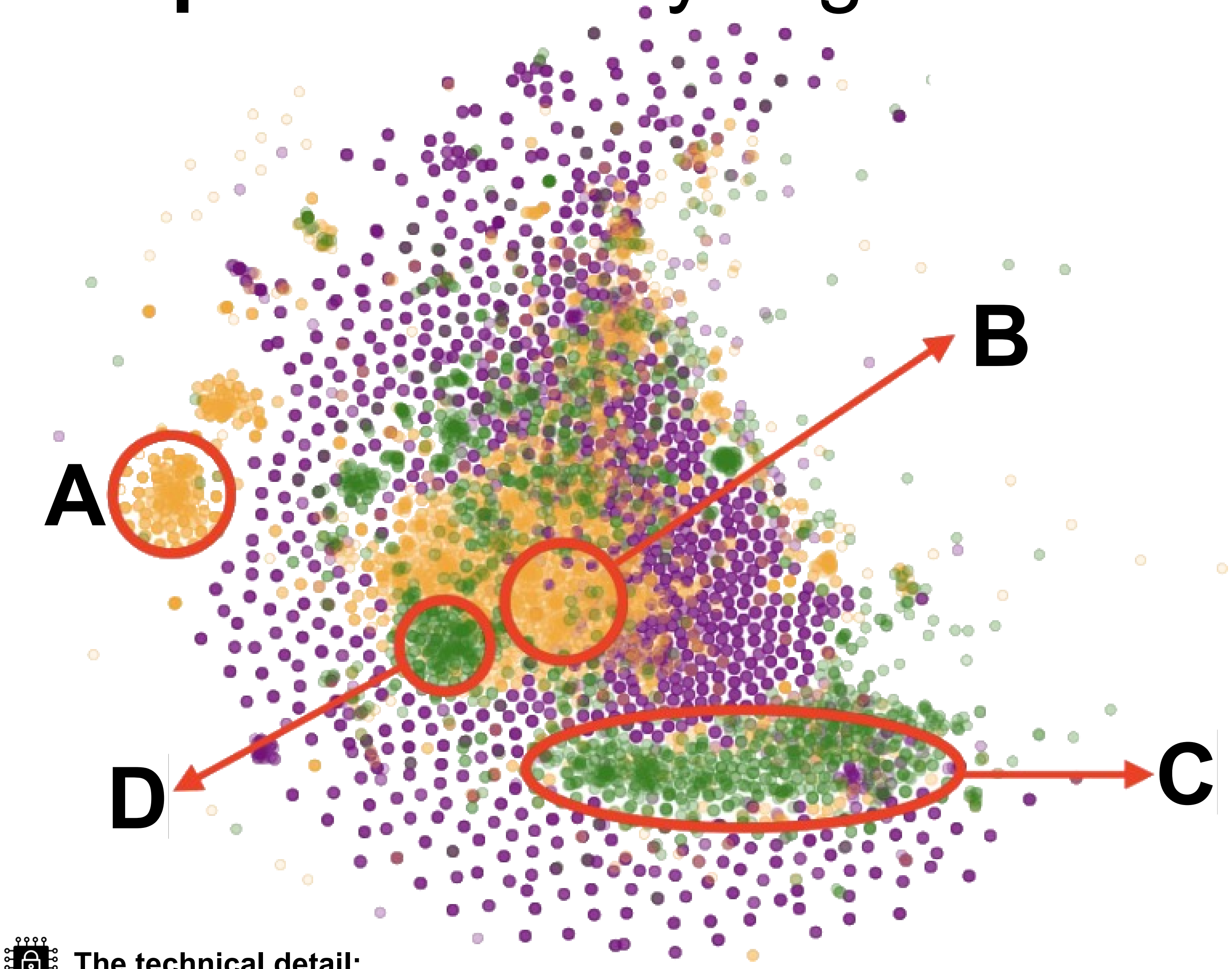
Disinformation cluster: tightly connected, coordinated and timed retweeting

Debunking cluster: organic retweeting, reactionary, loosely connected, multiple communities

The technical detail:

Map information networks from social media; content, interactions, structure and diffusions analyse; social network analysis

Expertise: Analysing coordinated troll strategies



(yellow) right trolls: focused MAGA
(magenta) left trolls: surround discussion
(green) news trolls: selective highlighting

A – *(right trolls)* Hillary cannot be trusted *#ThingsMoreTrustedThanHillary*

B – *(right trolls)* Mimic black Trump supporters *#Blacks4Trump*

C – (news trolls) News about violence and civil unrest *#news*

D – (news trolls) Federal politics, policy and regulation *#politics*

The technical detail:

Semantic edit distance; dimensionality reduction; Twitter trolls

UTS Data Science Institute Capability

Team of 35 full-time data scientists and data engineers with a strong focus on delivering meaningful industry impact.

Deep expertise in cutting-edge social network message diffusion, virality and disinformation, ratified through high-profile publications

UTS is the top ranked university in computer science and engineering in Australia and top 15 in the world and is the top-ranked Australian university in scientific impact and collaboration

A collaborative network of researchers operating in the disinformation space and data science spaces, from PhD student through to senior researcher

Long history of industry project delivery to diverse partners from Australian government, global water utilities, regulatory agencies, and energy, water, transport and education sectors

Data Science Institute members have won industry awards, the Eureka Data Science Prize, the CSIRO Collaboration medal for their work across applied data science initiatives

Experience in the management, leadership and delivery of large-scale collaborative research initiatives and long-term partnerships (including management of a $20m initiative with the federal government)

A voice that stretches beyond academia, with meaningful media engagement record and experience across print, digital, television and radio platforms