UTS

Behavioral Data Science

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.

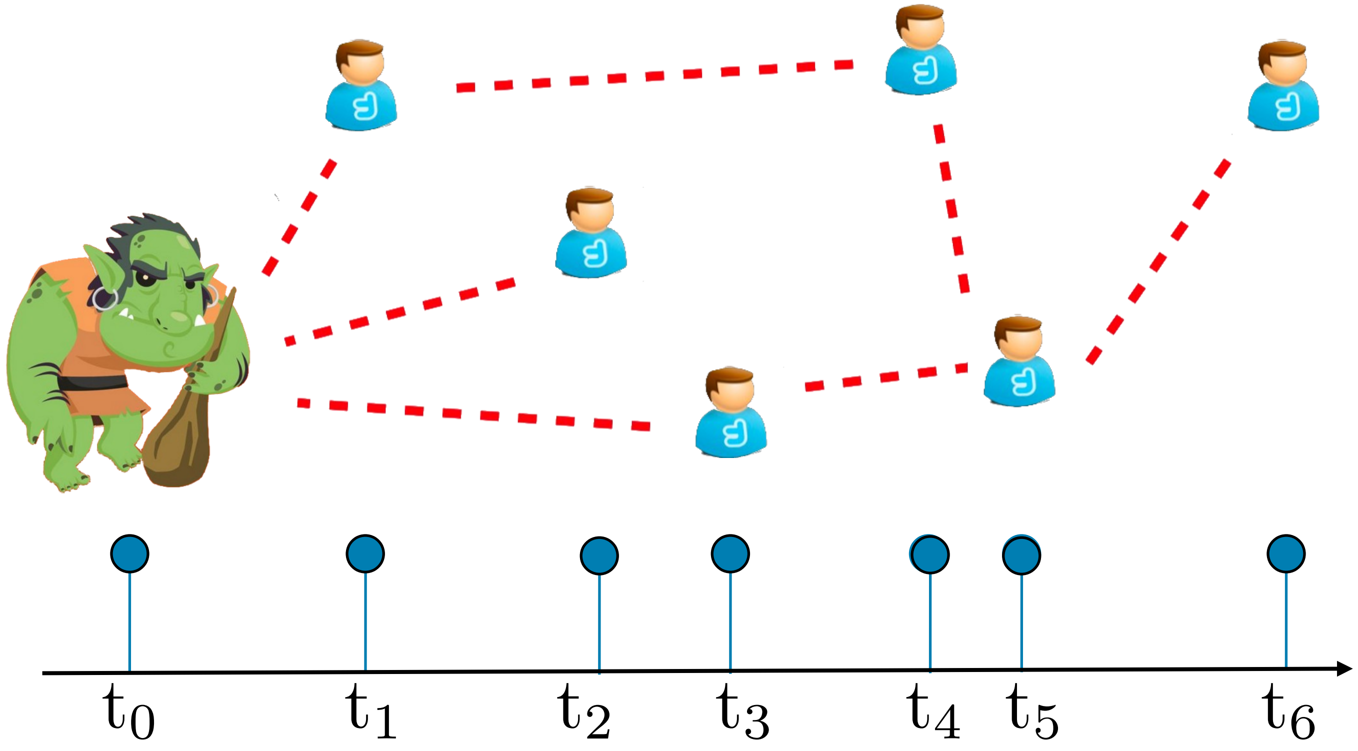
Challenge: beyond content-based detectors

Content- and user-based detection tools:
language nuances, language drift and adversarial attacks

IO are designed to elicit particular reactions from the target audience

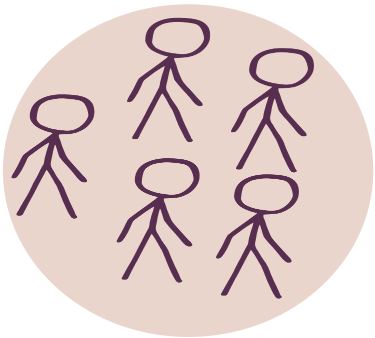
Distinguish users and content types based on on the reaction of online social systems (**no content)**

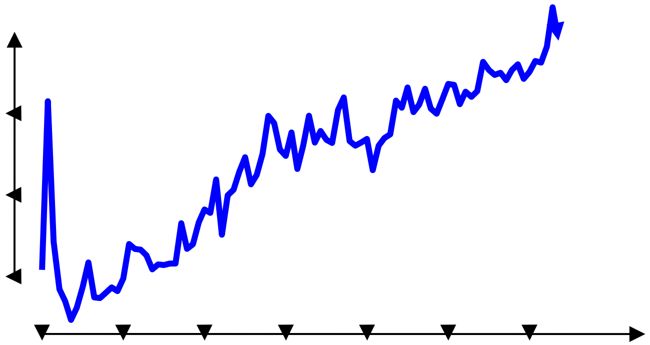
Build early detection systems based on information spread patterns within the user population



The Behavioral Data Science

1.

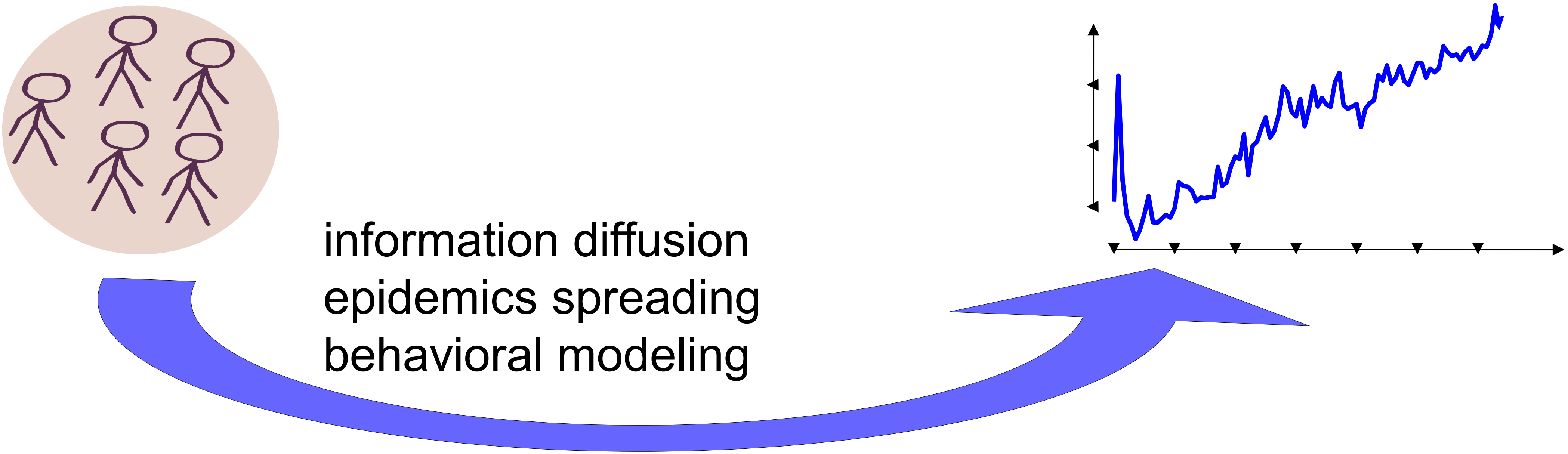




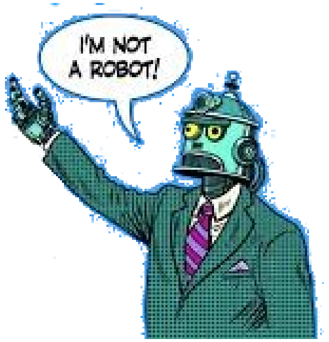
information diffusion
epidemics spreading
behavioral modeling

The Behavioral Data Science

1.



2.

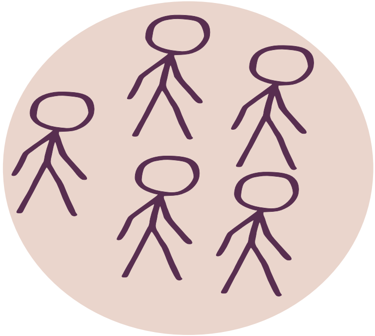


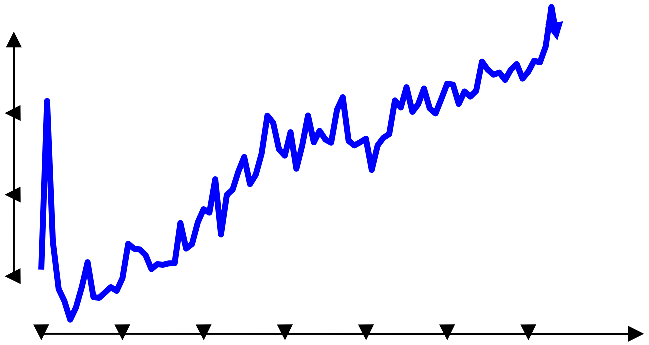
[Rizoiu et al ICWSM'18]

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

The Behavioral Data Science

1.

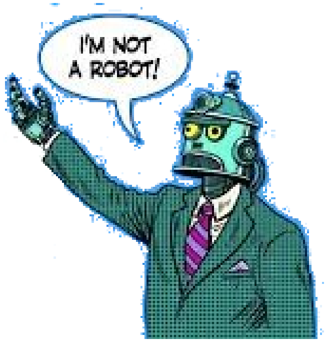




information diffusion
epidemics spreading
behavioral modeling

3.

2.



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

Behavioral DS capabilities in Influence Operations space

Response level	Monitor	Detect	Predict	Mitigate
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.	Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels	Model the impact of networks and influencers on the virality and reach of problematic messages	Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest
	Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.	Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message	Track the spread of problematic messages across and between online platforms and into the real-world	Identify key message inoculation points in social networks based on how information flows and gains velocity

Our founders in the mis-, dis-, IO and IW spaces

Australian Government
Department of Defence
Defence Science and Technology Group

Real-time detection of disinformation campaigns

Australian Government
Department of Home Affairs

Information integrity initiative: fighting misinformation in Australia

DEFENCE INNOVATION NETWORK
To go far, go together

INFORMATION WARFARE DIVISION
AUSTRALIAN DEFENCE FORCE

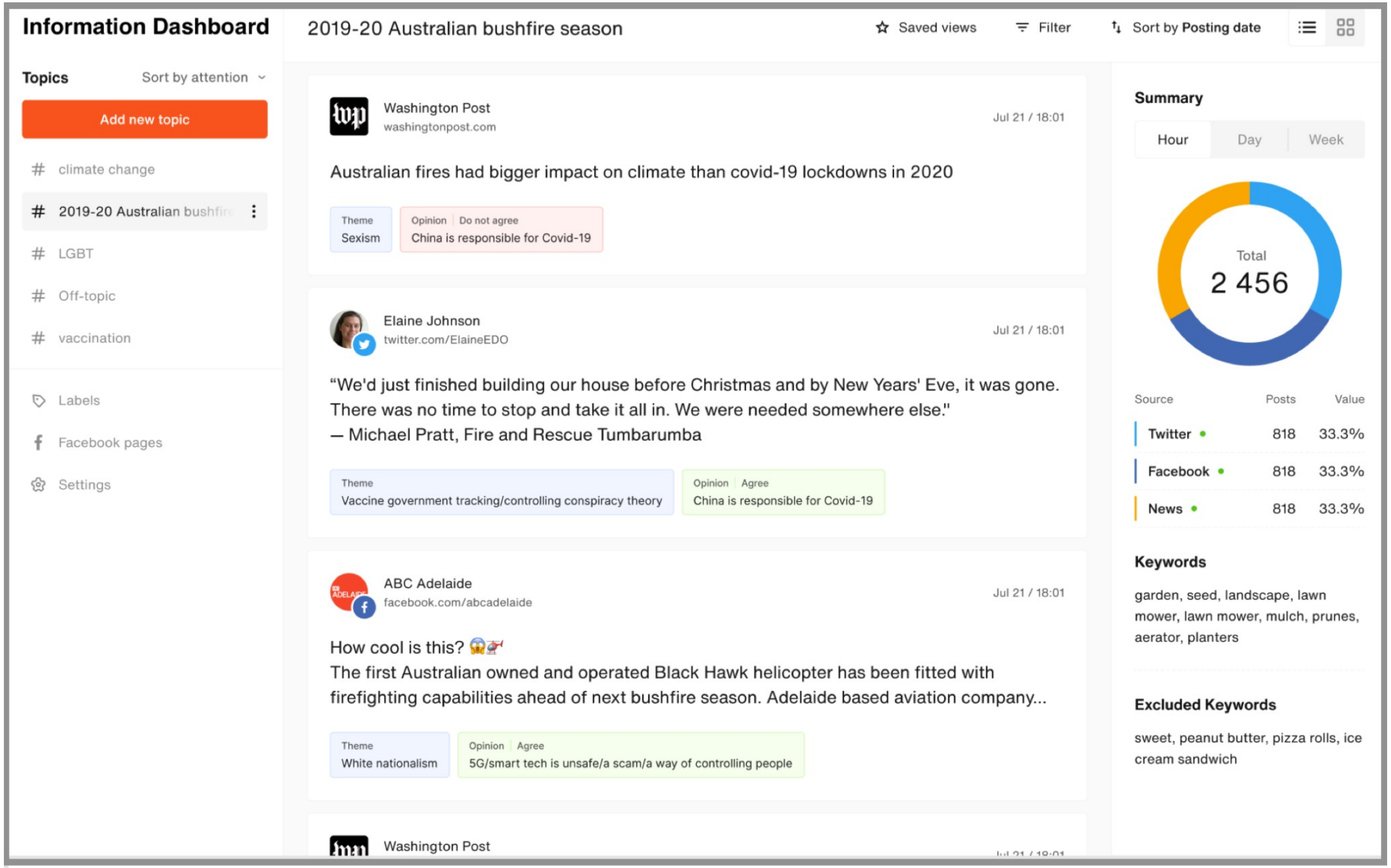
Effectiveness of Information Operations in the Pacific

Australian Government
Department of Defence
Defence Science and Technology Group

Information Warfare STaR Shot “Developing Situational Awareness”

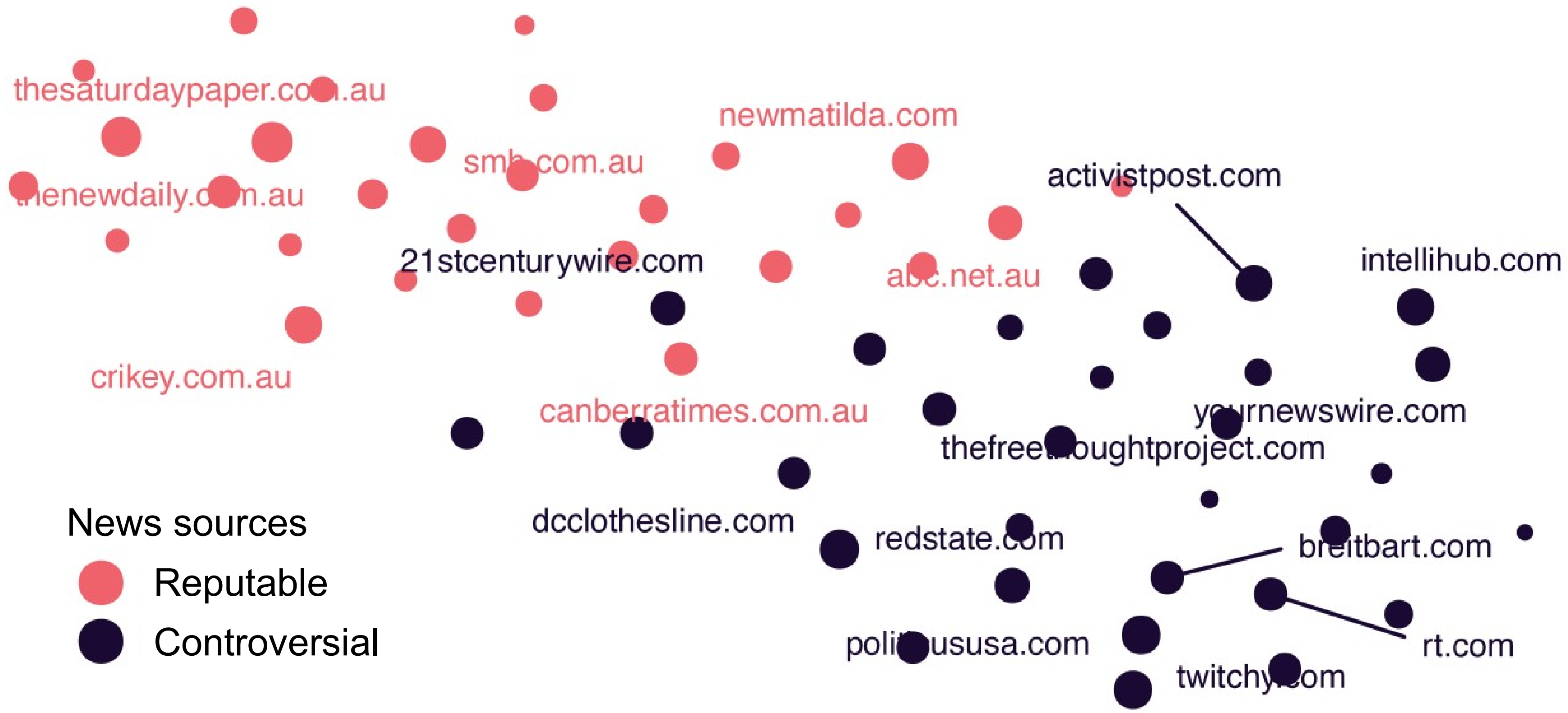
Hate Speech propagation on Social Media

Monitor: Monitoring discussion spaces (TRL: 3)



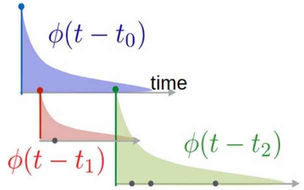
Graphical interface of the Information Dashboard

Detect: separating controversial from reputable



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

Detect controversial news without content analysis



evently

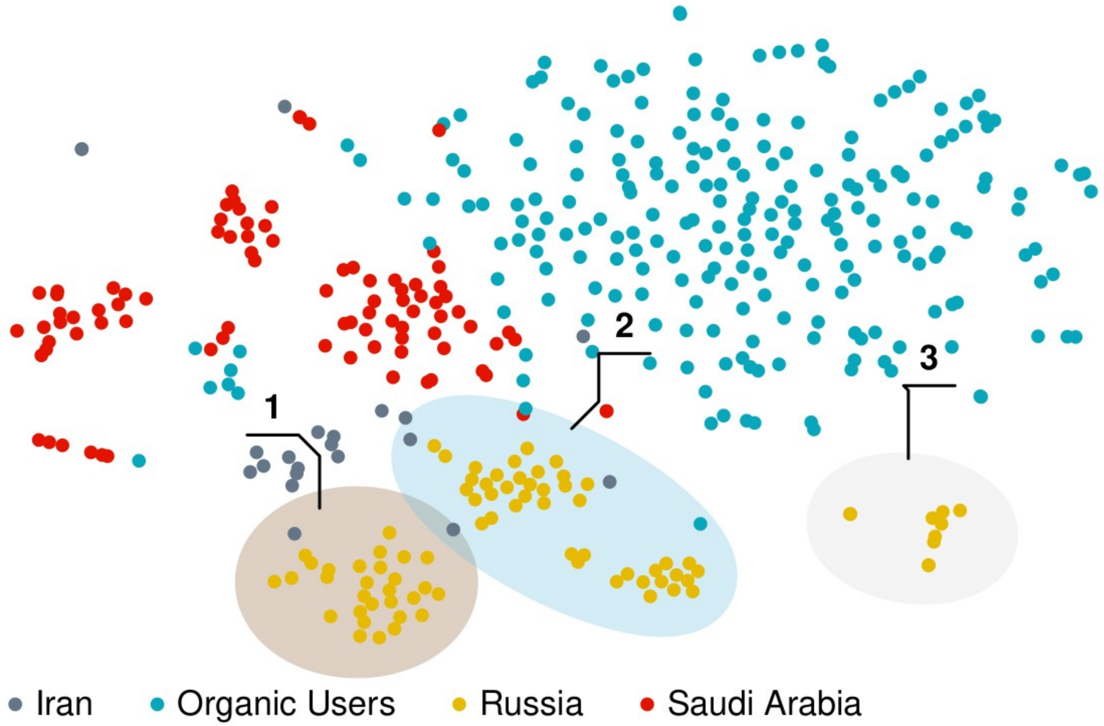
https://www.behavioral-ds.science/theme1_content/evently/

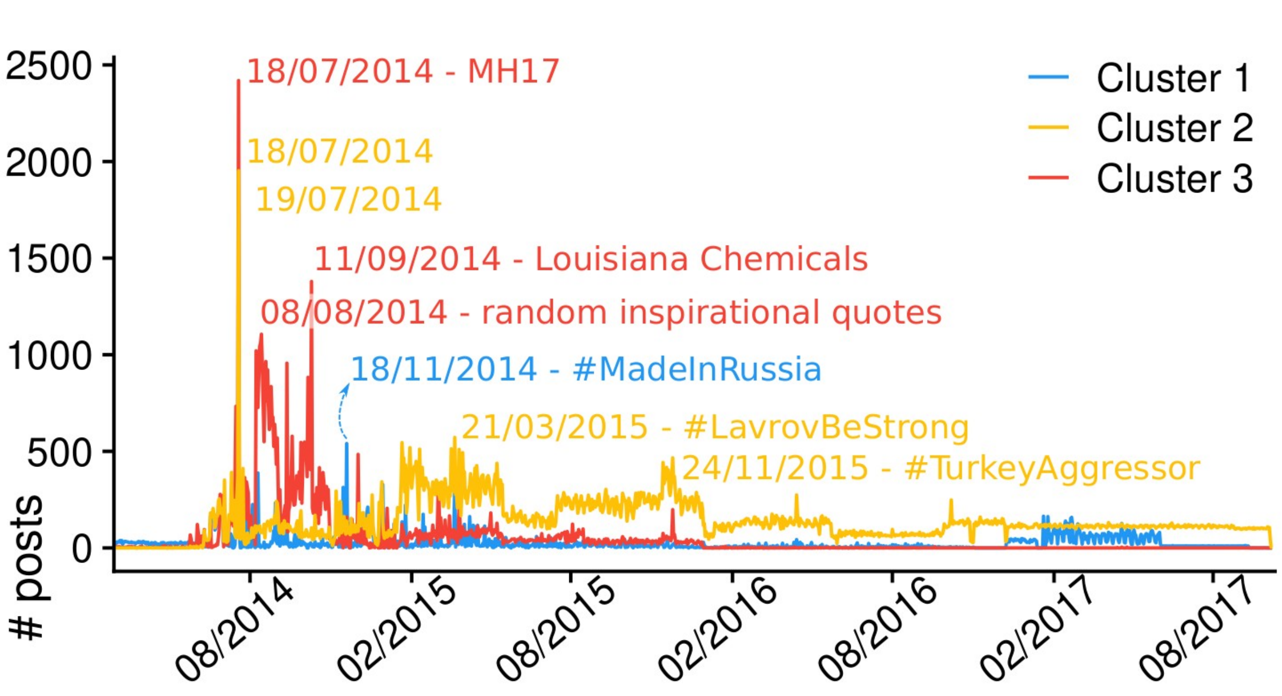
The technical detail:

Mathematical generative modelling; Hawkes processes; joint modelling

Kong, Q., Rizoiu, M. A., & Xie, L. (2020). Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In International Conference on Information and Knowledge Management, Proceedings (pp. 645–654). New York, NY, USA: ACM. https://doi.org/10.1145/3340531.3411861

Detect: identify agent types and coordinated behavior





IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.

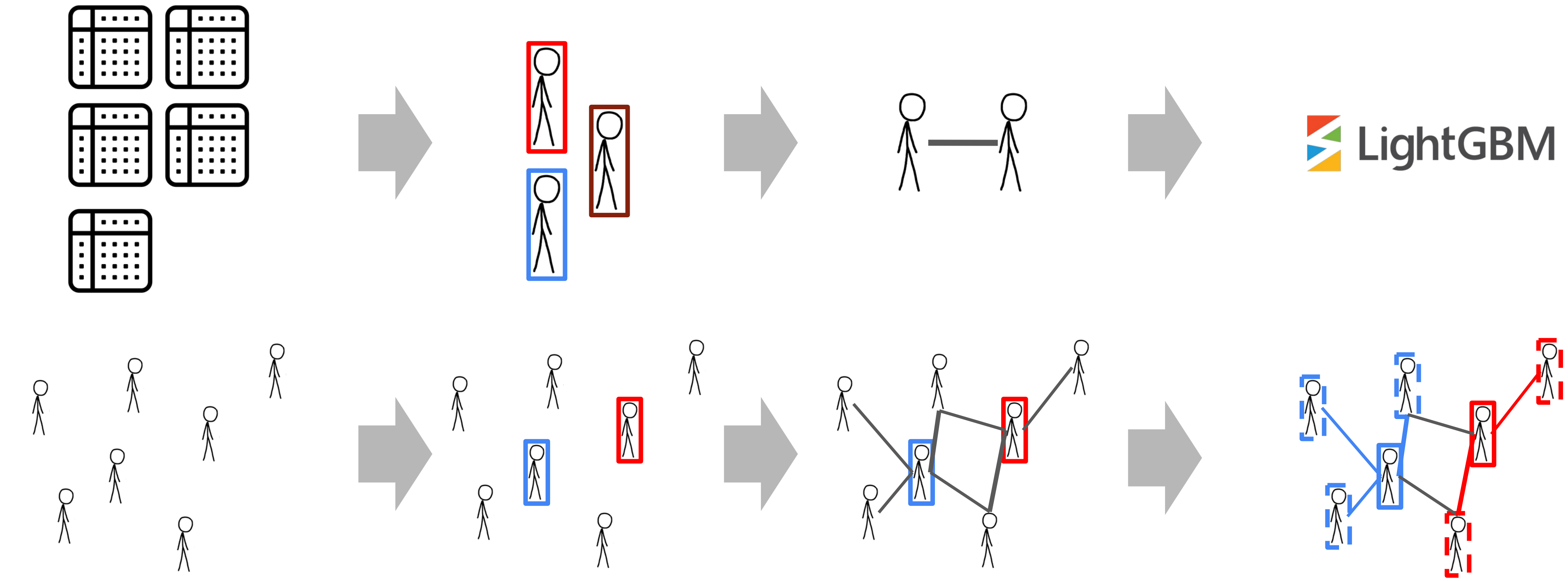
Qualitative investigations uncovers strategies of Russian trolls farms:
C1: Russian news with patriotic framing;
C2: Regional and conservative news;
C3: tweet in English, *#music*, *#usa*, relationship advice

The technical detail:

Interval-censored Transformer Hawkes; Twitter Moderation Research Consortium dataset; partial data loss

Kong, Q., Calderon, P., Ram, R., Boichak, O., & Rizoiu, M.-A. (2023). Interval-censored Transformer Hawkes: Detecting Information Operations using the Reaction of Social Systems. In ACM Web Conference (WWW’23) (pp. 1–9). https://doi.org/10.1145/3543507.3583481

React: far-right and extremist ideology detection



No data labelling required for new datasets

Seed users: use fringe news sharing to label a small subset of users

Text-based homophily quantification – “what you write is who you are”

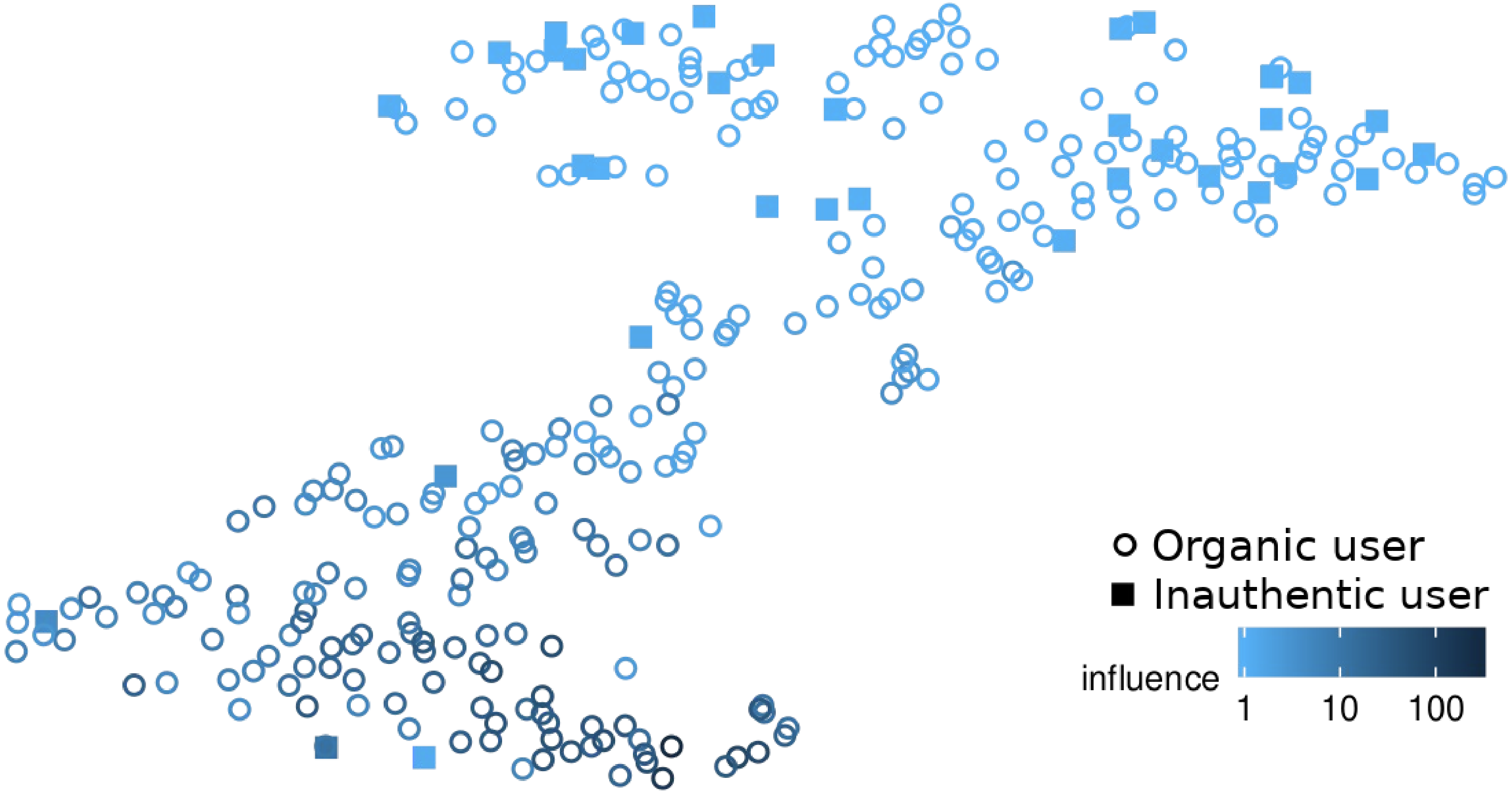
Use advanced Machine Learning to label all users

The technical detail:

Ideology proxies; homophily lenses (text, follower, URLs); automatic user labelling

Ram, R., Thomas, E., Kernot, D., & Rizoiu, M.-A. (2022). Detecting Extreme Ideologies in Shifting Landscapes: an Automatic & Context-Agnostic Approach. Retrieved from http://arxiv.org/abs/2208.04097

React: Identify influential inauthentic users (bots)



Identify users engaged in influence operations

Estimate their impact on the wider community

birdspotter

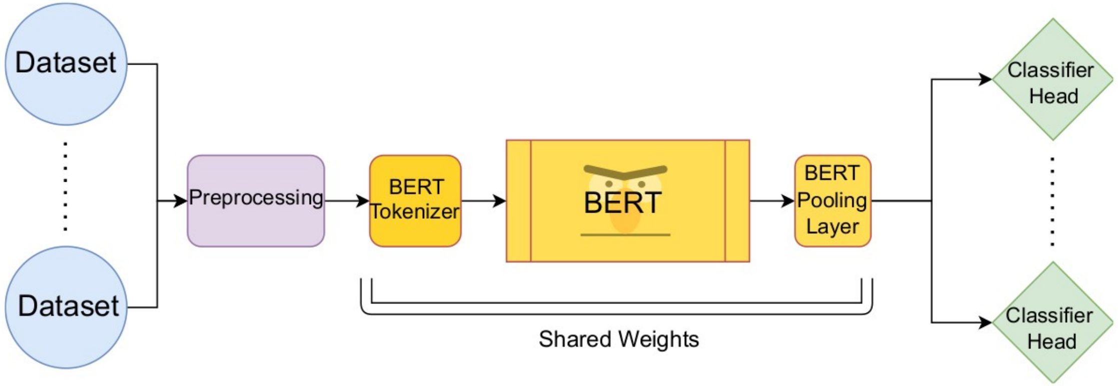
https://www.behavioral-ds.science/theme2_content/birdspotter/

The technical detail:

Influence estimation using stochastic modelling; content-free analysis

Ram, R., & Rizoiu, M.-A. (2021). Conductance and Social Capital: Modeling and Empirically Measuring Online Social Influence. Retrieved from http://arxiv.org/abs/2110.12569

React: Detecting Hate Speech in Unseen Domains



Novel learning paradigm to leverage many disparate datasets to learn a single hate speech representation

Improved performances over the state-of-the-art, generalizable to novel datasets.

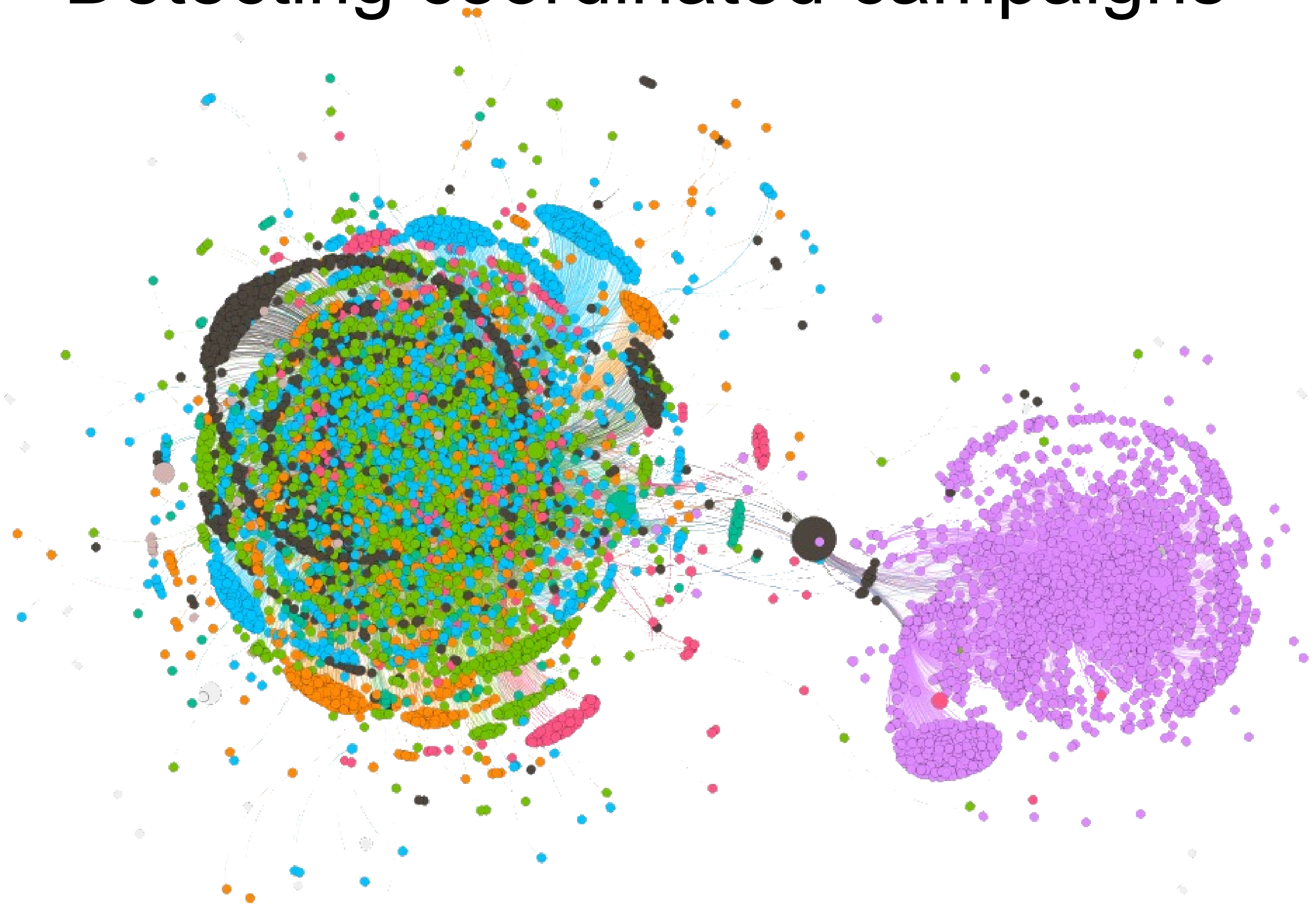
		Testing Dataset									
	Model	Davidson	Waseem	Reddit	Gab	Fox	Storm-front	Mandl	HatEval	PubFigs-L	# Wins
MTL	MTL-NCH	**0.6822**	0.3801	**0.8456**	**0.8738**	**0.6150**	**0.6826**	0.5312	**0.6449**	0.6175	6
MTL	MTL-MV	0.6455	0.4048	0.8263	0.8660	0.6030	0.6771	0.4834	0.6315	**0.6231**	1
BERT baseline trained on:	Davidson		0.5556	0.5914	0.6731	0.4932	0.4597	0.5690	0.5414	0.5469	0
	Waseem	0.6136		0.6000	0.6427	0.5519	0.5356	0.5099	0.5784	0.5611	0
	Reddit	0.6135	0.4957		0.8083	0.5229	0.5559	0.4900	0.5741	0.5402	0
	Gab	0.5720	0.4595	0.8375		0.5075	0.5645	0.4277	0.5664	0.5185	0
	Fox	0.4285	0.4249	0.4234	0.4651		0.3865	0.4159	0.4490	0.3926	0
	Stormfront	0.4533	0.5467	0.5822	0.6487	0.5740		0.5104	0.5664	0.5659	0
	Mandl	0.3336	0.4822	0.4066	0.4582	0.4010	0.3518		0.4546	0.3633	0
	HatEval	0.5849	0.5824	0.5700	0.5796	0.5532	0.5466	0.5348		0.5432	0
	PubFigs-L	0.6351	**0.6048**	0.5970	0.6600	0.5546	0.5249	**0.5963**	0.5858		2

The technical detail:

Transfer learning; language models fine-tuning;

Yuan, L., & Rizoiu, M.-A. (2022). Detect Hate Speech in Unseen Domains using Multi-Task Learning: A Case Study of Political Public Figures. Retrieved from http://arxiv.org/abs/2208.10598

Detecting coordinated campaigns



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

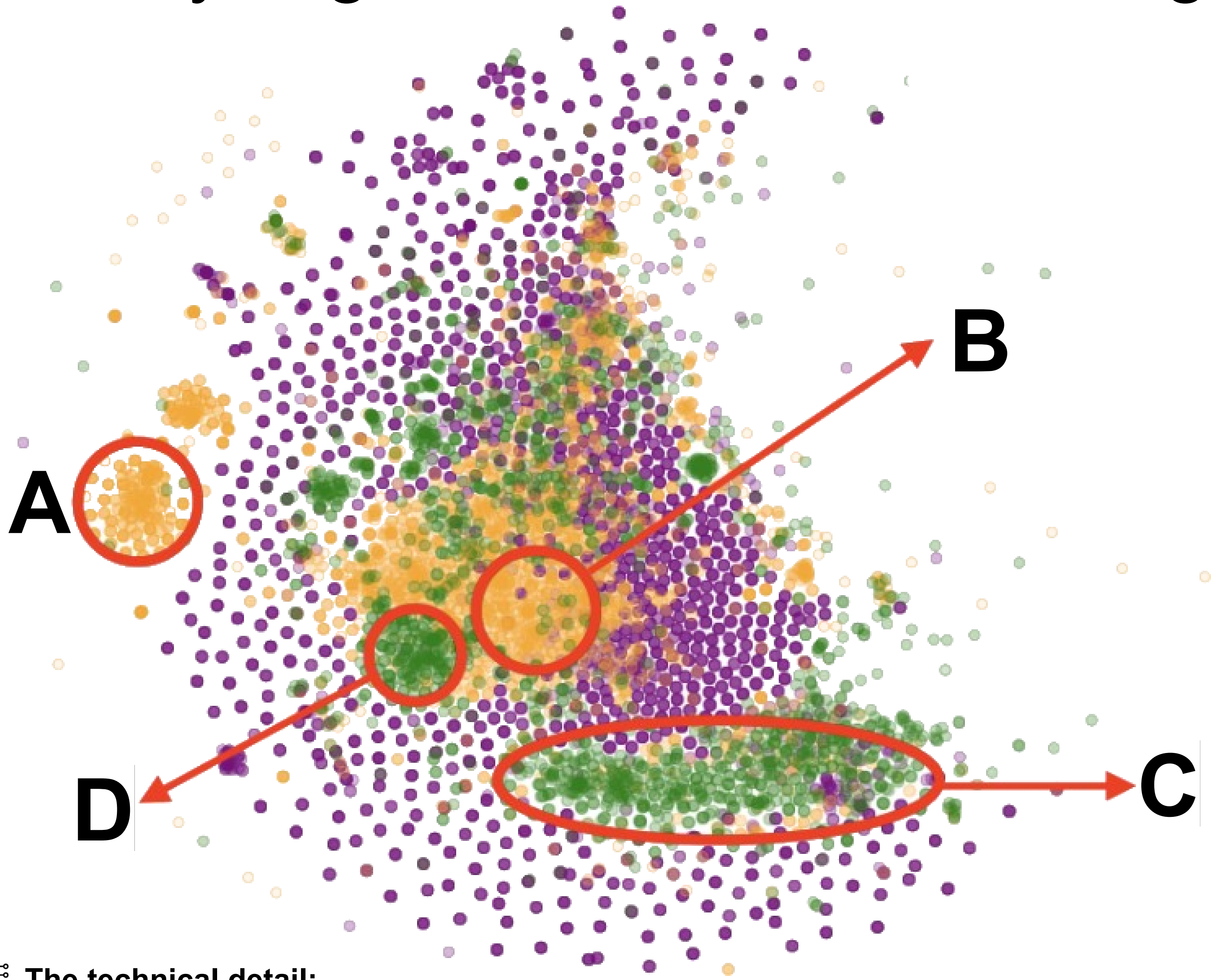
Disinformation cluster: tightly connected, coordinated and timed retweeting

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

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