Possible Future of AI

Based on ideas in Toby Walsh, “It’s Alive!”, La Trobe University Press 2017

Note: none of this material is examinable

Where have we been

1950 – Turing predicts “thinking machines” by 2000
1956 – Dartmouth Summer Research Project on AI
1965 – Dendral expert system for reasoning about molecular chemistry
1969 – Perceptrons (early neural network) by Minsky and Papert
1972 – Shakey the robot (computer vision, path planning, A* search)
1984 – Cyc project to encode all commonsense knowledge
1986 – Backpropagation for multi-layer neural networks
1997 – Chess grand master Kasparov loses to IBM Deep Blue
2005 – DARPA Grand Challenge for autonomous vehicles won
2011 – IBM Watson wins Jeopardy! game show
2015 – AlphaGo uses deep RL and tree search to beat Go master

DARPA’s third wave of AI

1st wave: handcrafted knowledge, 2nd wave: statistical learning

3rd wave: contextual reasoning

– AI functions “more as colleague than as tool”

New Capabilities: real-time analysis of sophisticated cyber attacks, detection of fraudulent imagery, human language technologies, control of prosthetic limbs

Robust AI: reliable and verifiable operation in complex environments

High Performance AI: 1000x faster, 1000x less power

Next Generation AI: explainable AI, ethical AI, common sense AI

https://www.darpa.mil/work-with-us/ai-next-campaign

An example scenario

How can we tell the difference between real news and fake news online?

Acknowledgements for this work

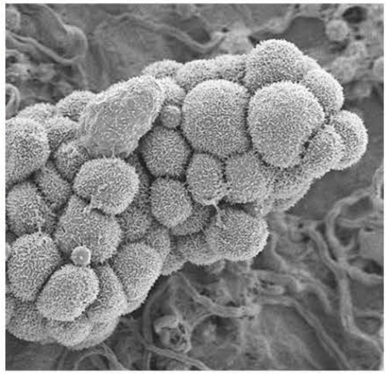
Shanika Karunasekera, Sarah Erfani, James Bailey, Tansu Alpcan, Ben Rubinstein, Yi Han, Vinh Nguyen, Sandamal Weerasinghe, Ling Luo, Amila Silva, Hadi Dolatabadi
University of Melbourne

Olivier de Vel, Paul Montague, Tamas Abraham, David Hubczenko
Defence Science and Technology Group

Deep Learning is Everywhere

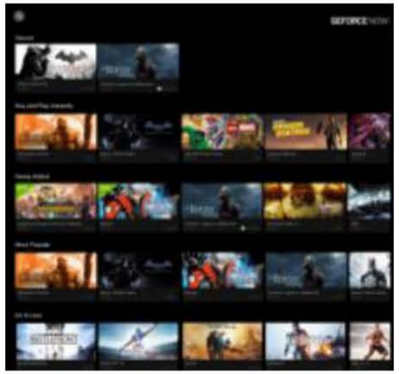
INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

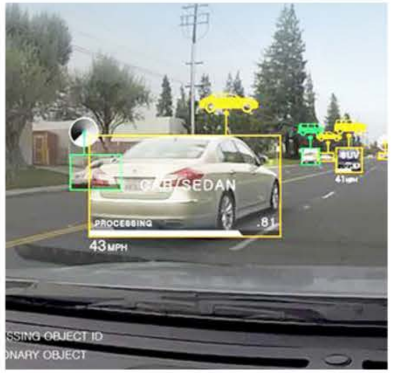


MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery



AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

How can it be used to combat disinformation campaigns?

How can adversaries disrupt defences that use machine learning?

An Example of Disinformation that Fooled Me

Bigfoot

Scenario: Social Botnet for Disinformation

Malicious actors often use a **botnet** of automated accounts on a social media platform such as Twitter to amplify the impact of their influence campaigns, i.e., a force multiplier for trolls

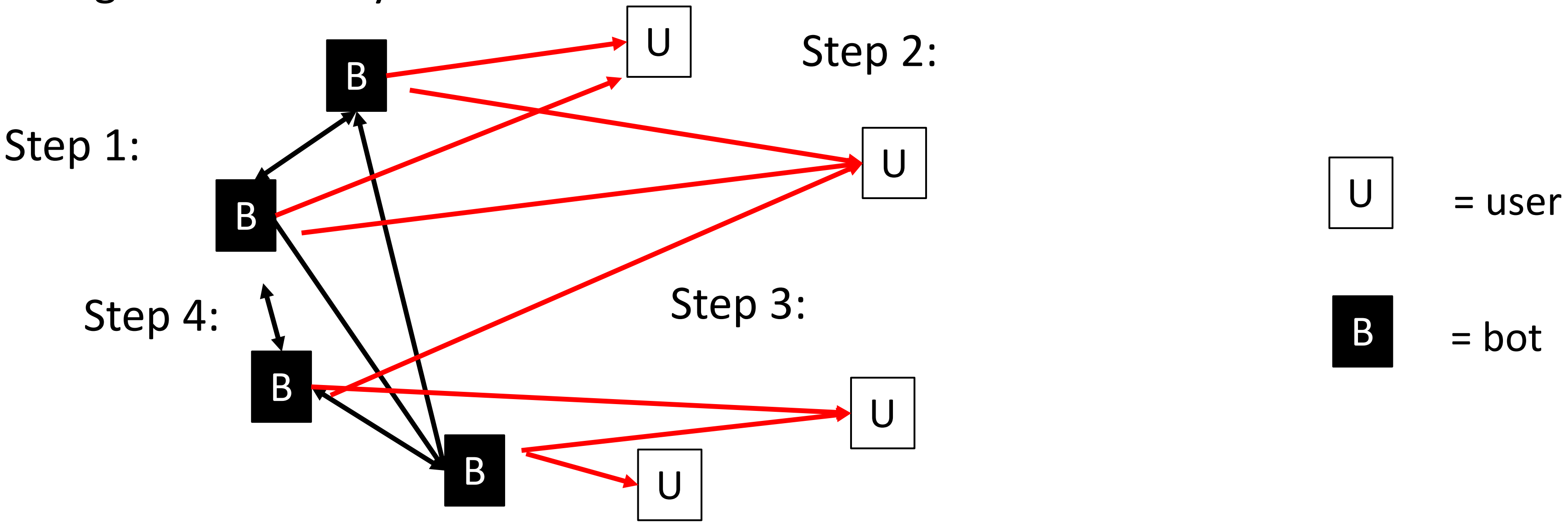
How does a botnet work?

(1) **Infiltrate** target community on social media

(2) Use botnet to **influence** discourse

Step 1: Infiltrate Target Community

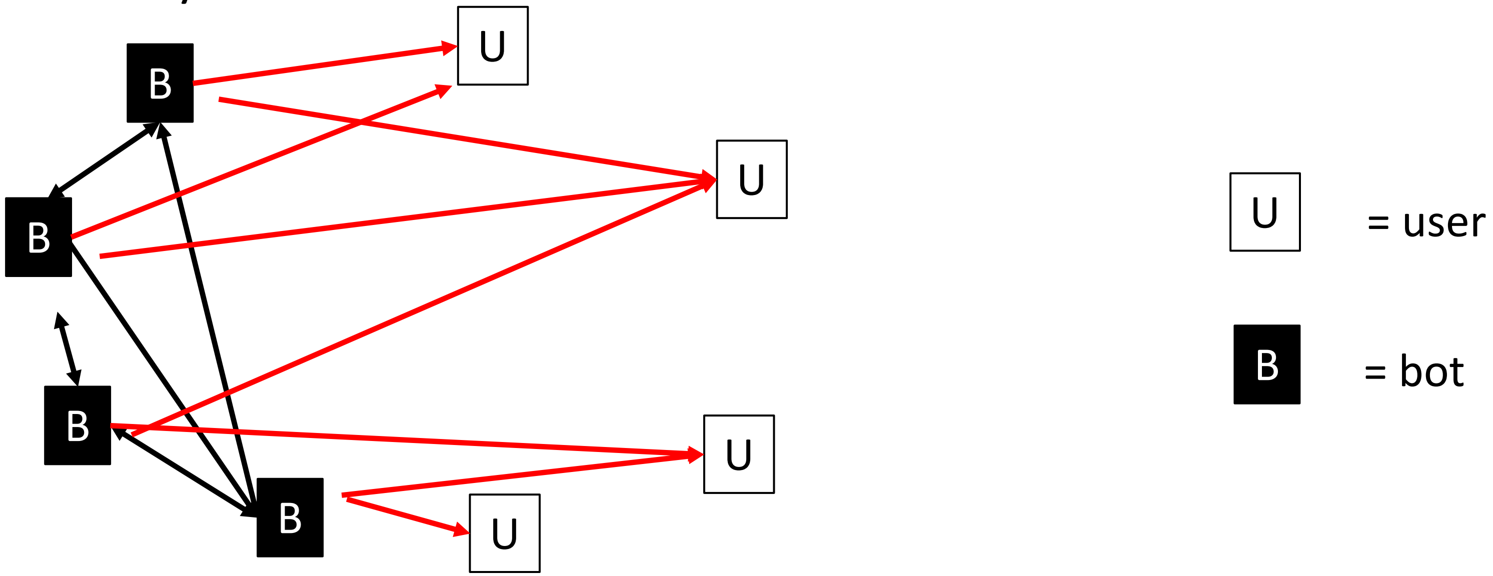
1. Attach social bot software to automate new/repurposed Twitter accounts
2. Pick a community of users to target
3. Each bot follows a subset of popular users in that community (red links)
4. Bots in target community follow each other (black links)
5. Bots make several posts before interacting with real users (manual tweets, plagiarise tweets, synthetic tweets)
6. Repeat steps 3-5 until a sufficient number of bots have been followed by users in the target community



Step 1: Infiltrate Target Community

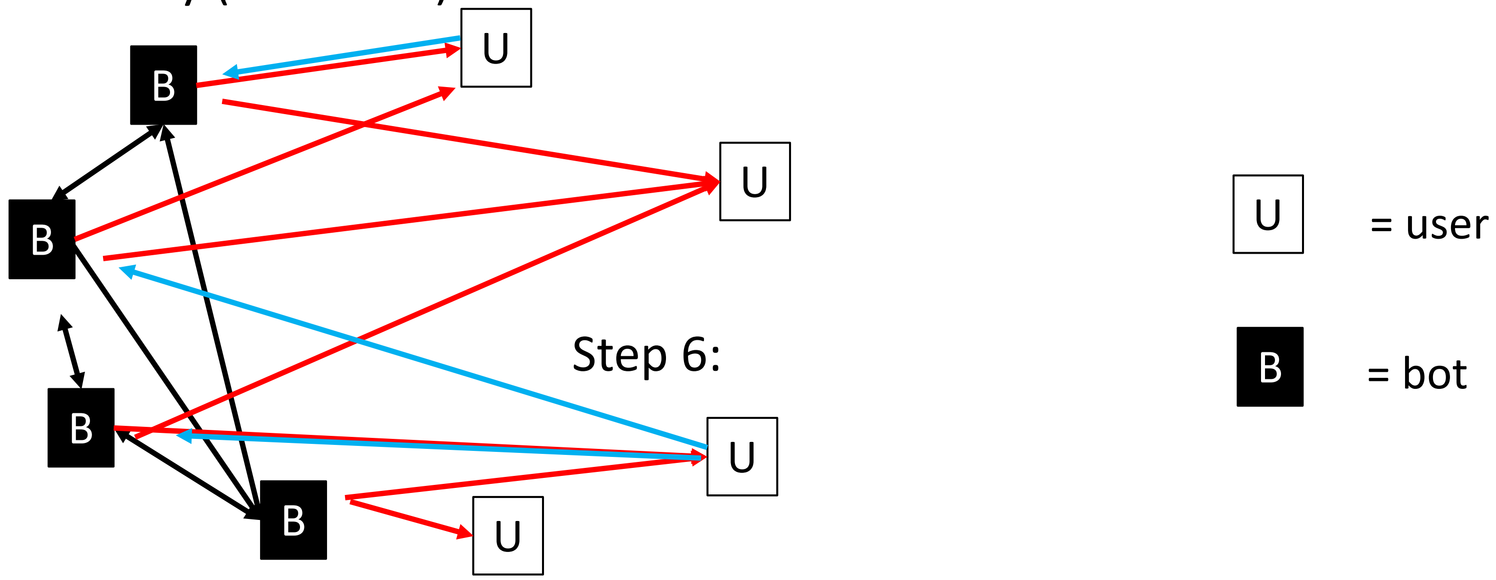
1. Attach social bot software to automate new/repurposed Twitter accounts
2. Pick a community of users to target
3. Each bot follows a subset of popular users in that community
4. Bots in target community follow each other
5. Bots make several posts before interacting with real users (manual tweets, plagiarise tweets, synthetic tweets)
6. Repeat steps 3-5 until a sufficient number of bots have been followed by users in the target community

Step 5:



Step 1: Infiltrate Target Community

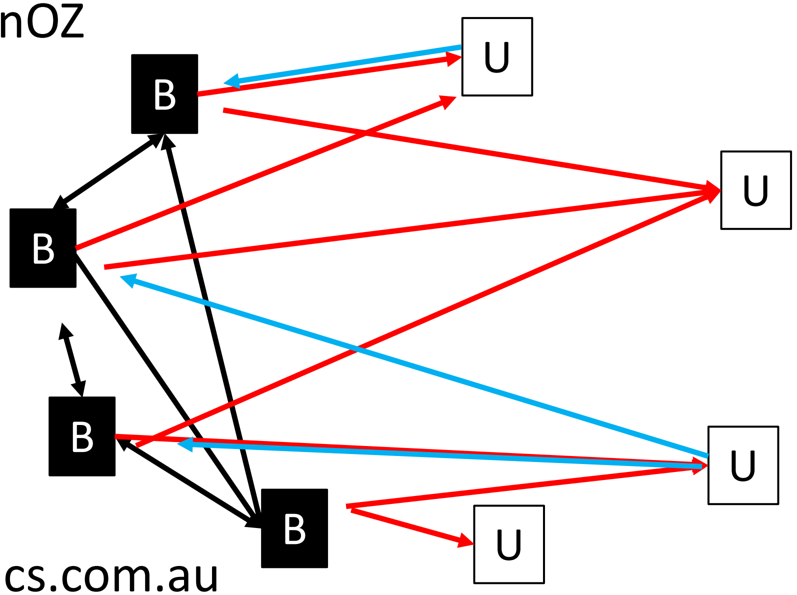
1. Attach social bot software to automate new/repurposed Twitter accounts
2. Pick a community of users to target
3. Each bot follows a subset of popular users in that community
4. Bots in target community follow each other
5. Bots make several posts before interacting with real users (manual tweets, plagiarise tweets, synthetic tweets)
6. Repeat steps 3-5 until a sufficient number of bots have been followed by users in the target community (blue links)



Step 2: Use Botnet to Influence Discourse

1. Spread politically motivated rumours
   (fake grassroots activity to give impression of popular support)
2. Promote disinformation news sites
3. Create sufficient noise to disrupt reasonable discussions

(1) #BigfootInOZ



(2)
http://bigfootpics.com.au

(3)

#Bigfoot4PM

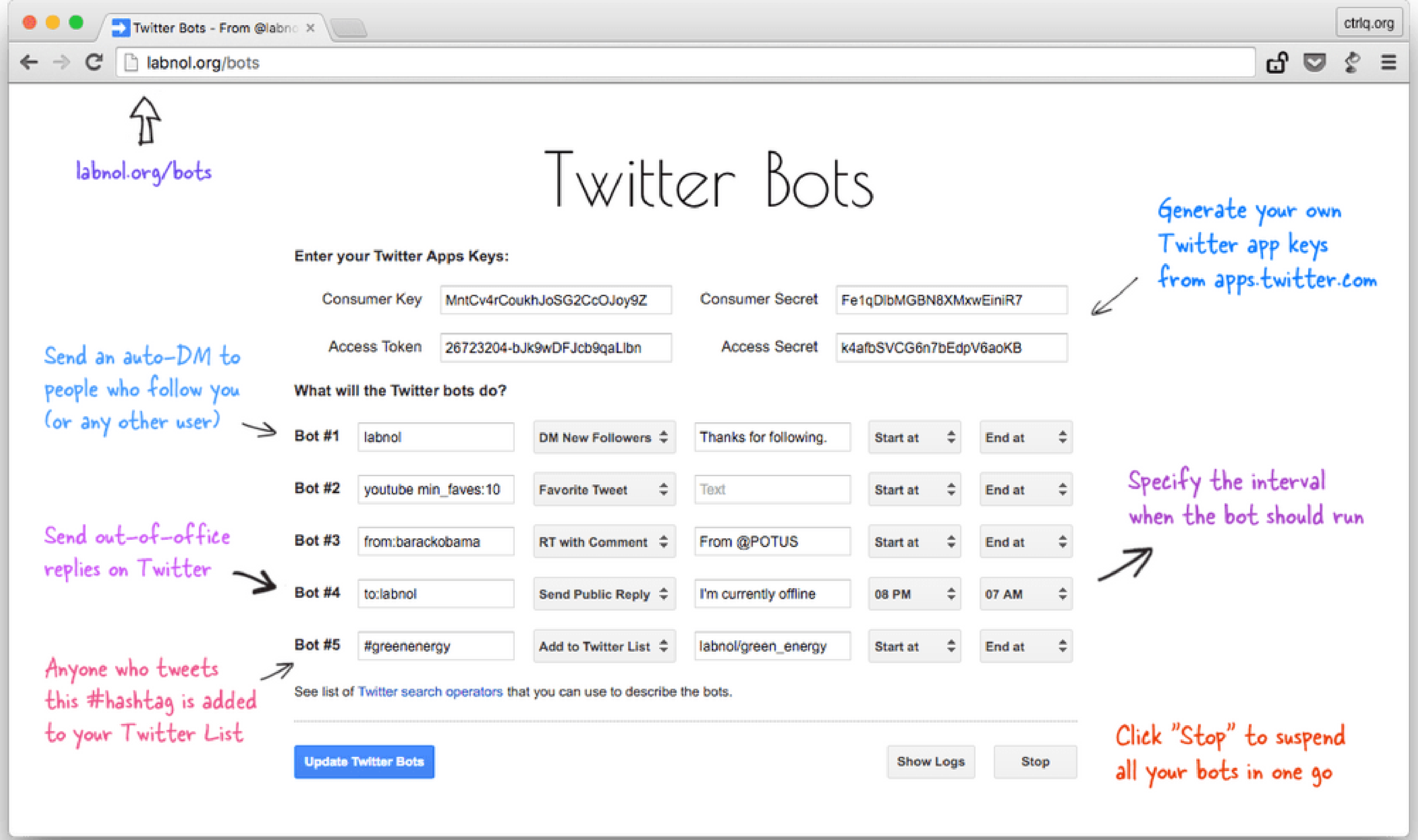
#YetiSux

#DropBears

#BigfootCausedCovid

How to Program a Bot

Bots can be implemented using a scripting language to specify pre-programmed behaviours in response to received tweets and messages



labnol.org/bots

Twitter Bots

Generate your own Twitter app keys from apps.twitter.com

Enter your Twitter Apps Keys:

Consumer Key	MntCv4rCoukhJoSG2CcOJoy9Z	Consumer Secret	Fe1qDlbMGBN8XMxwEiniR7
Access Token	26723204-bJk9wDFJcb9qaLlbn	Access Secret	k4afbSVCG6n7bEdpV6aoKB

What will the Twitter bots do?

Bot #1	labnol	DM New Followers	Thanks for following.	Start at	End at
Bot #2	youtube min_faves:10	Favorite Tweet	Text	Start at	End at
Bot #3	from:barackobama	RT with Comment	From @POTUS	Start at	End at
Bot #4	to:labnol	Send Public Reply	I'm currently offline	08 PM	07 AM
Bot #5	#greenenergy	Add to Twitter List	labnol/green_energy	Start at	End at

See list of Twitter search operators that you can use to describe the bots.

Update Twitter Bots | Show Logs | Stop

Send an auto-DM to people who follow you (or any other user)

Send out-of-office replies on Twitter

Anyone who tweets this #hashtag is added to your Twitter List

Specify the interval when the bot should run

Click "Stop" to suspend all your bots in one go

Source: https://www.labnol.org/internet/write-twitter-bot/27902/

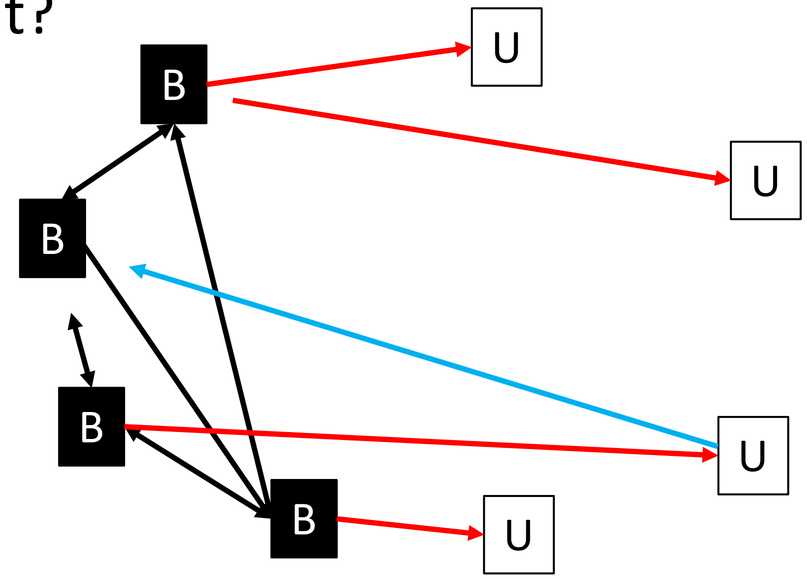
Social Cyber Security

What decisions do analysts need to make to disrupt these botnets?

Q: Is this an automated account?

Q: Is this tweet abnormal for this account?

Q: Are these accounts acting together to form a botnet?



Q: Who/what is the target of the botnet?

Q: What level of influence does the botnet have on the target community?

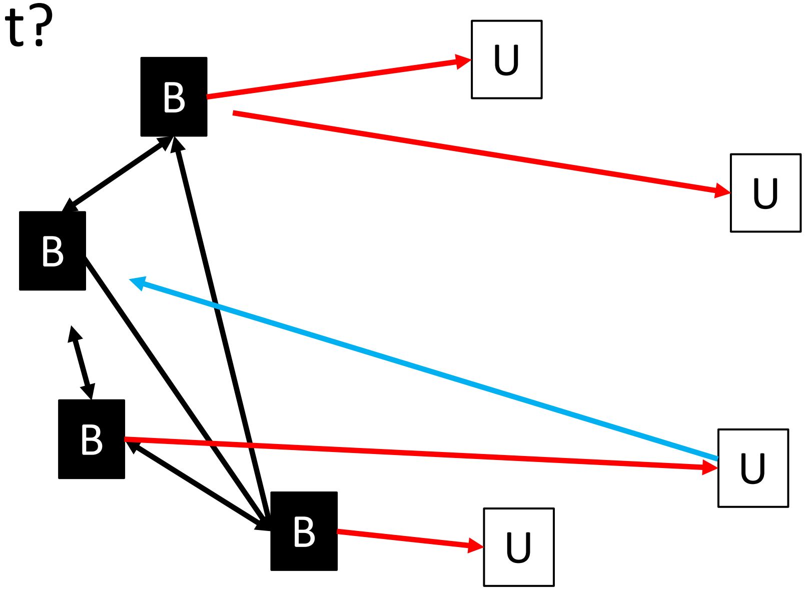
Helping to Automate these Decisions

How can artificial intelligence (AI) and machine learning (ML) be used to help analysts make these decisions?

Q: Is this an automated account?

Q: Is this tweet abnormal for this account?

Q: Are these accounts acting together to form a botnet?



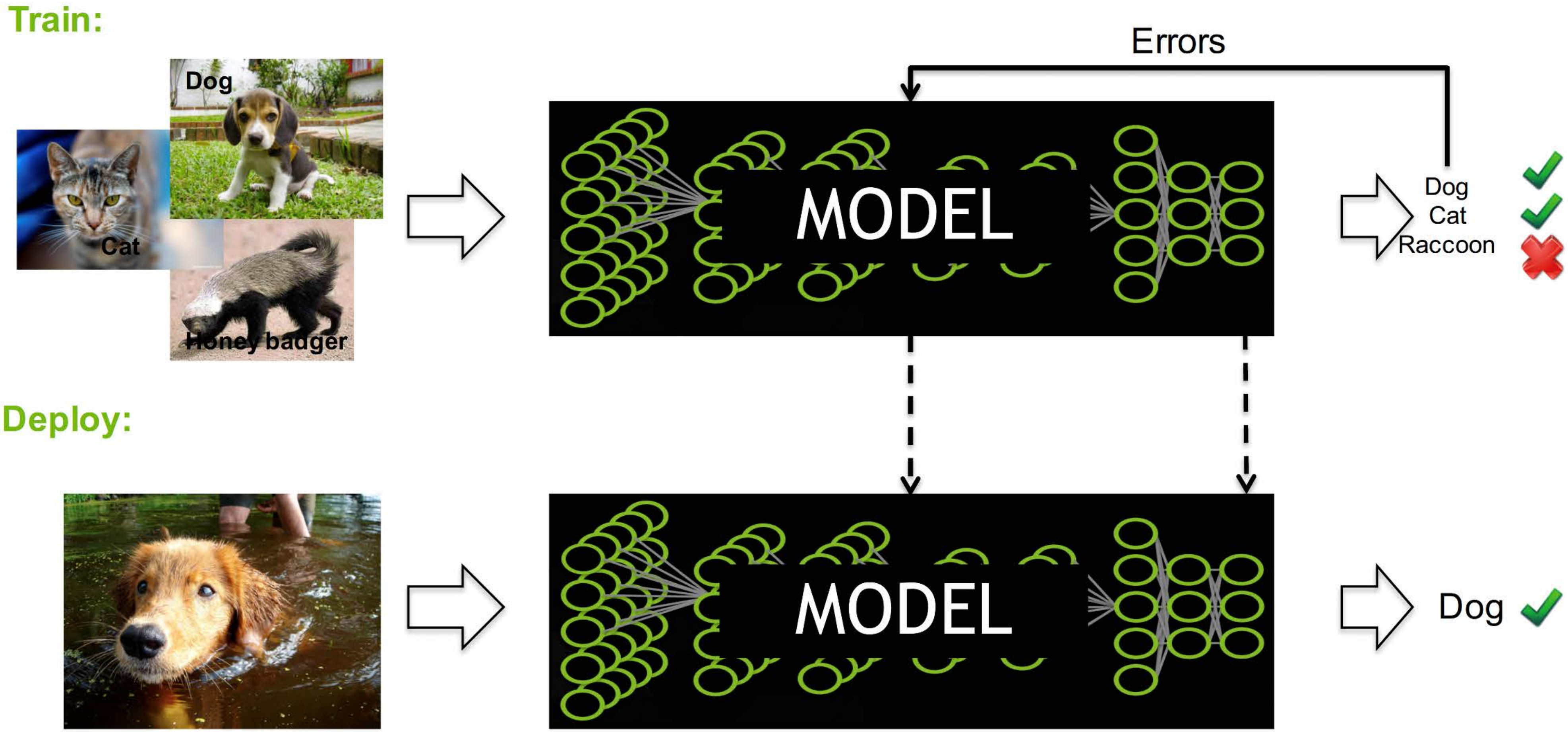
Classification

Anomaly detection

Clustering

Classification

Can we learn a model that predicts the category of a given example?



Train:

Deploy:

Clustering

What are the natural categories in a dataset?

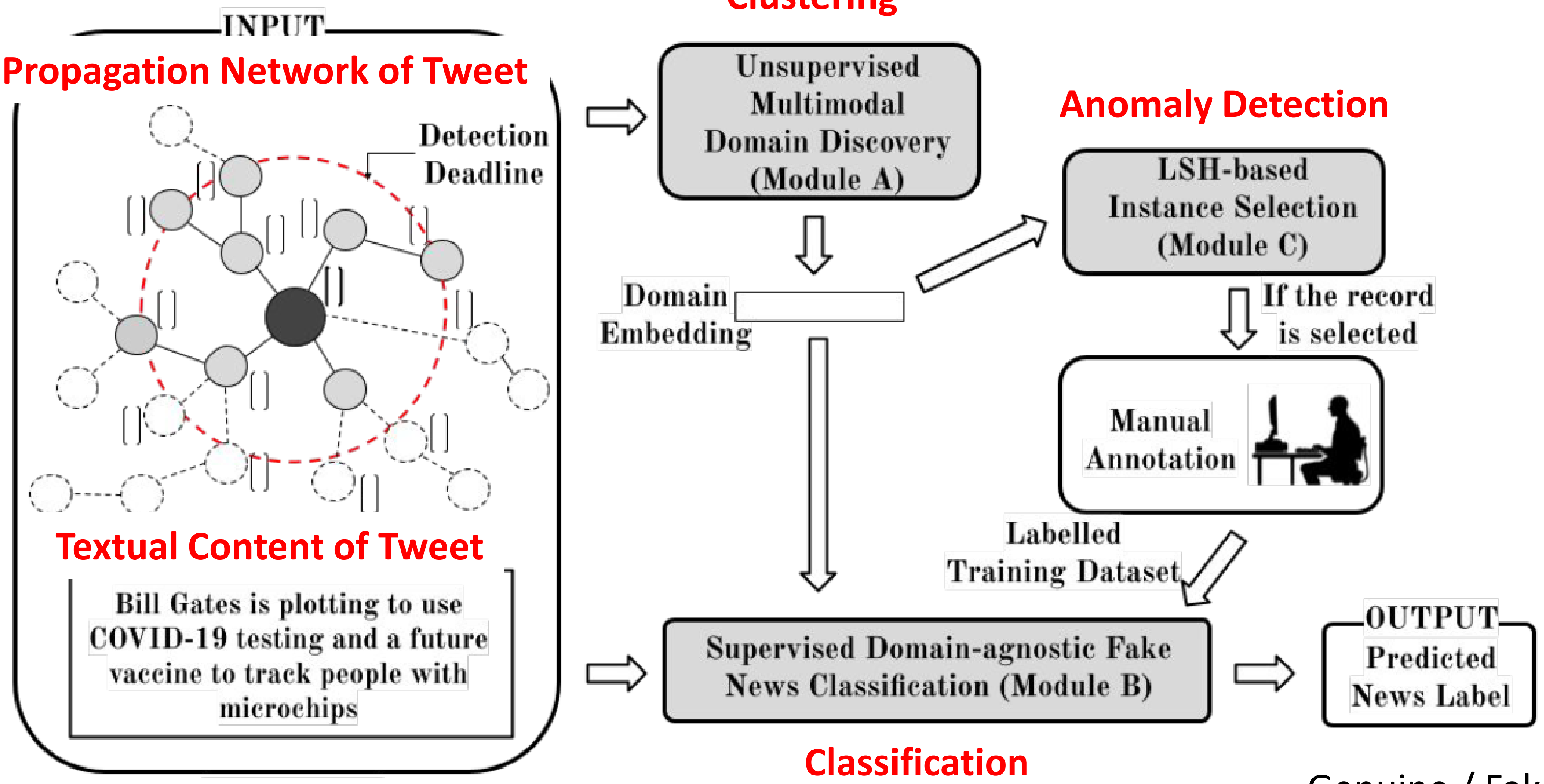
Consider a collection of animals.

How many different types of animals are there here?

Anomaly Detection

Can we learn a model of what is “normal” so that we can spot anomalies?

Deep Learning for Fake News Detection


INPUT
Detection
Deadline
Bill Gates is plotting to use COVID-19 testing and a future vaccine to track people with microchips
Unsupervised Multimodal Domain Discovery (Module A)
Domain Embedding
LSH-based Instance Selection (Module C)
If the record is selected
Manual Annotation
Labelled Training Dataset
Supervised Domain-agnostic Fake News Classification (Module B)
OUTPUT
Predicted News Label

Clustering

Propagation Network of Tweet

Anomaly Detection

Textual Content of Tweet

Classification

Tweet

Genuine / Fake

Silva, Luo, Karunasekera, Leckie (2021) “Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data.” AAAI 2021

So What Could Go Wrong?

Source: Winnetka Animal Hospital

So What Could Go Wrong?

- Intelligent adversaries know they are being monitored by a system based on machine learning
- Adversaries can modify their behaviour to manipulate the machine learning model into making the wrong decision
- Types of adversarial attacks on machine learning:
  - Poisoning the **training** data to bias the learned model of “normal” behaviour
  - Manipulating the **test** data in ways that are imperceptible to humans to fool the learned model – “Adversarial Noise”

What Can We Do About Adversarial Attacks on ML

A major focus of our research is on **AI Assurance:**

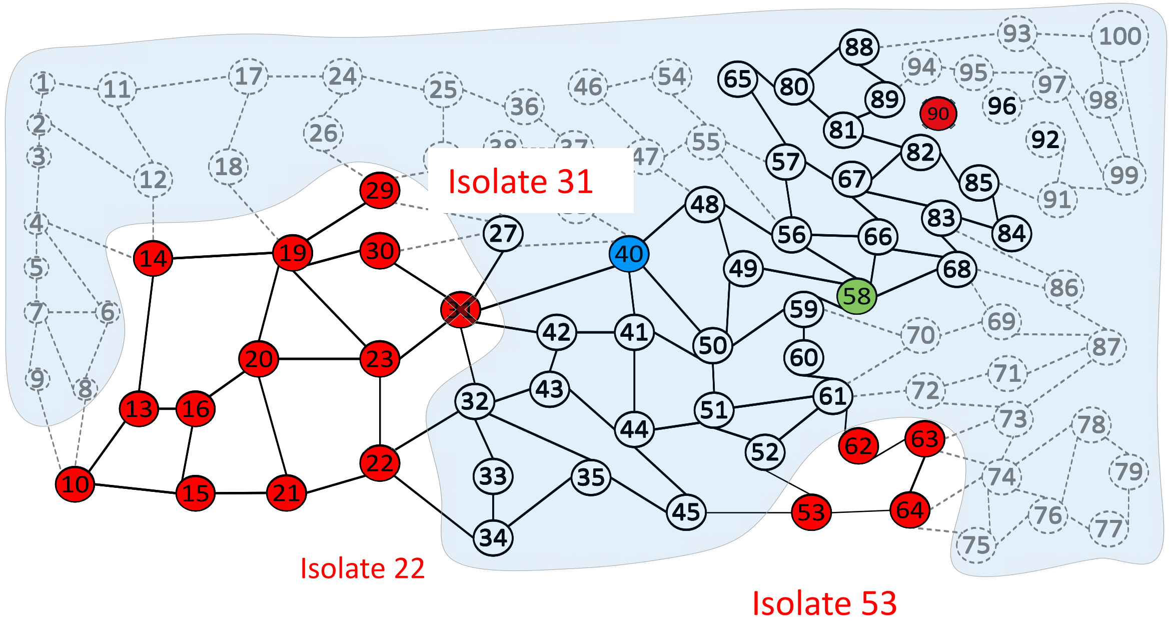
- Robustness in ML against Adversarial attacks,
  Privacy in ML models, Explainability of ML models

Examples of our work on Adversarial ML:

1. Detection and filtering of adversarial examples during training
2. Identifying new types of adversarial attacks
   so that we can devise better defences
3. Developing machine learning models that are resistant to attacks
   during testing / deployment

Our Work: Adversarial Attacks in Cyber Defence

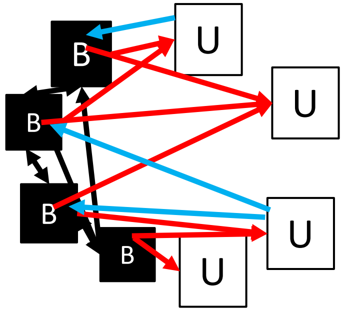
- Problem: How to detect the spread of malicious actors in networks
- Devised a method to detect and isolate the spread of malicious activity, while being resistant to adversarial manipulation
  (collaboration with DST Group)



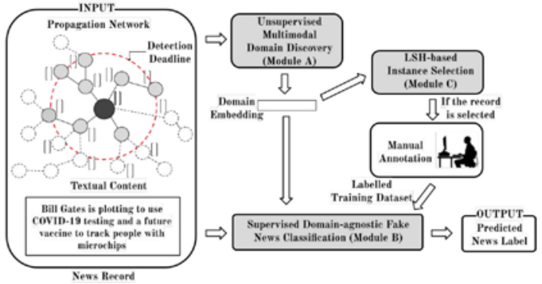
Han, Hubczenko, Montague, de Vel, Abraham, Rubinstein, Leckie, Alpcan, Erfani:
"Adversarial Reinforcement Learning under Partial Observability in Autonomous Computer Network Defence." IJCNN 2020

Conclusion

- Botnets are critical infrastructure for disinformation campaigns



- AI and ML can help automate the detection of these botnets, as a key step in defending against these campaigns


INPUT
Propagation Network
Detection Deadline
Textual Content
News Record
Unsupervised Multimodal Domain Discovery (Module A)
Domain Embedding
LSH-based Instance Selection (Module C)
If the record is selected
Manual Annotation
Labelled Training Dataset
Supervised Domain-agnostic Fake News Classification (Module B)
OUTPUT
Predicted News Label

Bill Gates is plotting to use COVID-19 testing and a future vaccine to track people with microchips

- However, adversarial ML creates a new "attack surface" that can disrupt our AI-based defences

Challenges for the Future

- While we can use AI to help automate defence, attackers can use AI to improve their attacks
- What will an **AI-enabled botnet** look like?
- In this AI arms-race, **does AI favour the attacker or defender**?

Botnet Resources

- *Reverse Engineering Socialbot Infiltration Strategies in Twitter,* Carlos A. Freitas, Fabrício Benevenuto, Saptarshi Ghosh, Adriano Veloso, https://arxiv.org/abs/1405.4927
- *Reverse engineering Russian Internet Research Agency tactics through network analysis*, Charles Kriel, Alexa Pavliuc, https://stratcomcoe.org/download/file/fid/80484
- *Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration*, Philip N Howard, Samuel Woolley, Ryan Calo, https://www.tandfonline.com/doi/full/10.1080/19331681.2018.1448735
- *BotCamp: Bot-driven Interactions in Social Campaigns*,
  Noor Abu-El-Rub and Abdullah Mueen, https://www.cs.unm.edu/~nabuelrub/BotCamp/
- *Social Cybersecurity: An Emerging National Security Requirement*,
  David M. Beskow, Kathleen M. Carley, https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/Mar-Apr-2019/117-Cybersecurity/

Adversarial Learning Resources

- *Explaining and Harnessing Adversarial Examples*, Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, https://arxiv.org/pdf/1412.6572.pdf
- *Robust Physical-World Attacks on Machine Learning Models*, Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song, https://arxiv.org/pdf/1707.08945.pdf
- *Practical Black-Box Attacks against Machine Learning*, Nicolas Papernot, Patrick McDaniel, Somesh Jha, Z. Berkay Celik, Ananthram Swami, https://arxiv.org/pdf/1602.02697.pdf
- *Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality*. Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. Houle, Dawn Song, James Bailey. https://openreview.net/pdf?id=B1gJ1L2aW
- *Adversarial Examples Are Not Bugs, They Are Features,* Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry, NeurIPS 2019, https://arxiv.org/abs/1905.02175
- Kaggle competition: https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack

Deep Learning Resources

Reading List

- http://neuralnetworksanddeeplearning.com/chap1.html
- http://neuralnetworksanddeeplearning.com/chap2.html

Further Resources

- http://www.wired.com/2014/01/geoffrey-hinton-deep-learning
- http://chronicle.com/article/The-Believers/190147/

Courses

- https://class.coursera.org/neuralnets-2012-001
- https://www.coursera.org/course/ml

Book

- http://www.deeplearningbook.org/

So, back to the future

How far can we go with AI?

Predictions of human - AI parity

In 2012, Mueller and Bostrom surveyed AI researchers:

“when is it 50% likely we will build a machine that does most jobs at least as well as an average human?”

Median response: 2040

“when is it 90% likely that high-level machine intelligence is achieved?”

Median response: 2075

What are the ethical limits of AI?

- Trolley car dilemma
- Algorithmic discrimination
- Privacy vs public good
- Humans and machines are indistinguishable
- Killer robots
- Equity: AI winners and AI losers in society

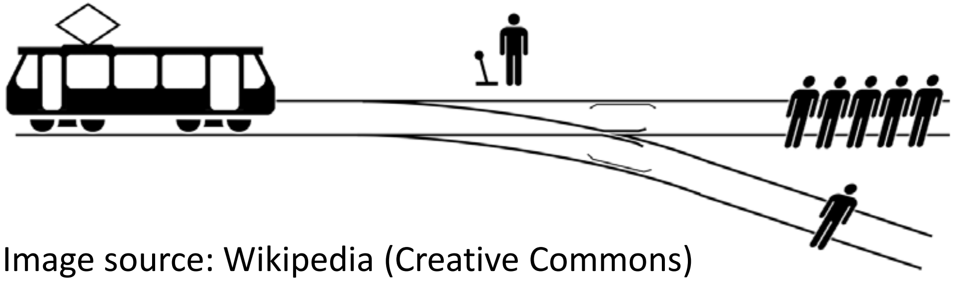


Image source: Wikipedia (Creative Commons)

10 Predictions for 2050

1. You are banned from driving
2. You see the doctor daily
3. Marilyn Monroe is back in the movies
4. A computer hires and fires you
5. You talk to rooms
6. A robot robs a bank
7. Germany loses to a robot soccer team
8. Ghost ships, planes and trains cross the globe
9. TV news is made without humans
10. We live on after death

What are your predictions?

Can you make a prediction of what AI will be able to do by 2050?

Email your ideas to Chris at caleckie@unimelb.edu.au by 9am Thu

We'll report any gems in Thursday's final lecture

We will also give the results of the Project Tournament!