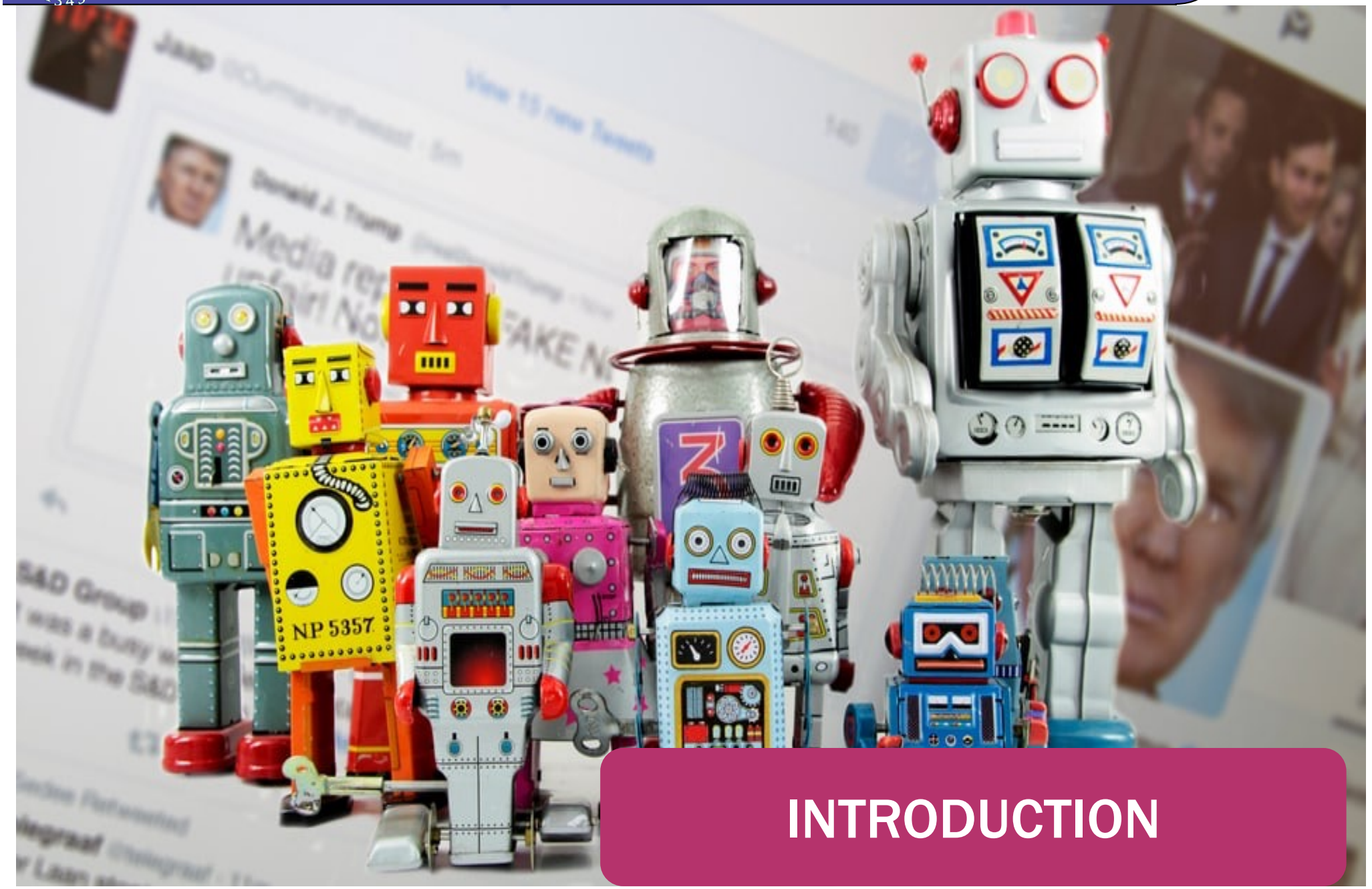
Social Bot Detection

Lorenzo Cima

lorenzo.cima@phd.unipi.it; lorenzo.cima@iit.cnr.it



INTRODUCTION

Malicious Accounts

Social media are fertile ground for the proliferation of malicious accounts

Why?

- Open platforms
- Anonymity
- Programmatic access (APIs)

For what purposes?

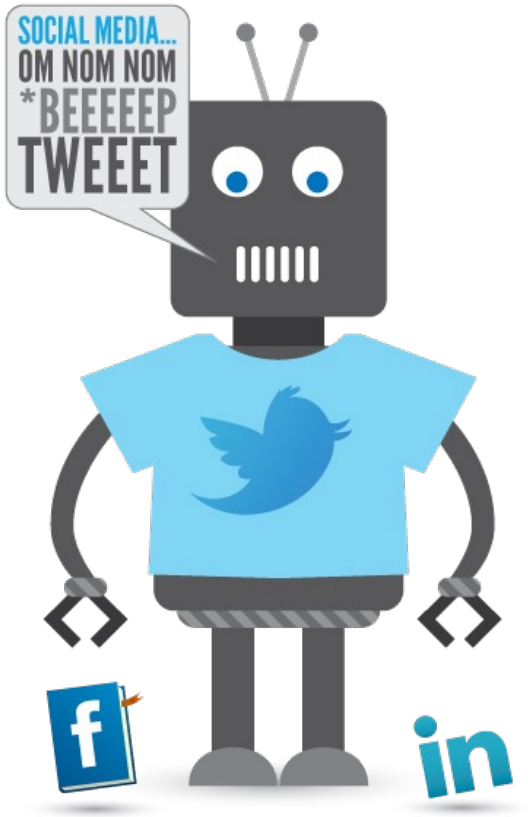
- Influence public opinions
- Economic purposes
- For fun

What Is a Social Bot?

A social media **account controlled by software**,
in a (more or less) **automatic** way

A social bot can automatically:

- "Read" what others say
- Create and post messages
- Like and reshare content
- Establish social relationships
- Contact other users



Malicious and Benign Bots

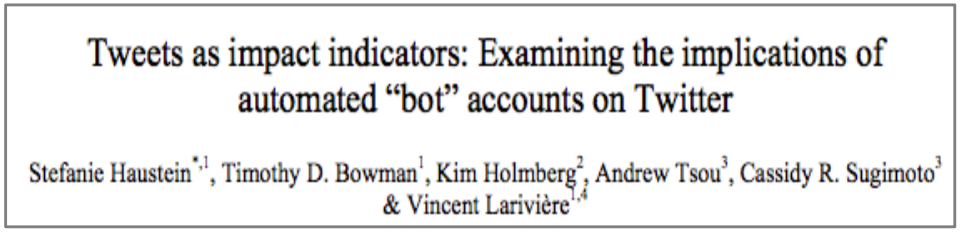
Not all social bots are **malicious**, some are **neutral** or even **benign:**

- News aggregation and dissemination


NEWS BOTS
Automating news and information dissemination on Twitter

Tetyana Lokot and Nicholas Diakopoulos

news dissemination


Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter

Stefanie Haustein[*,1], Timothy D. Bowman[1], Kim Holmberg[2], Andrew Tsou[3], Cassidy R. Sugimoto[3] & Vincent Larivière[1,4]

scientific impact

Malicious and Benign Bots

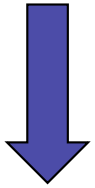
Not all social bots are **malicious**, some are **neutral** or even **benign:**

- Support and coordination in emergency situations

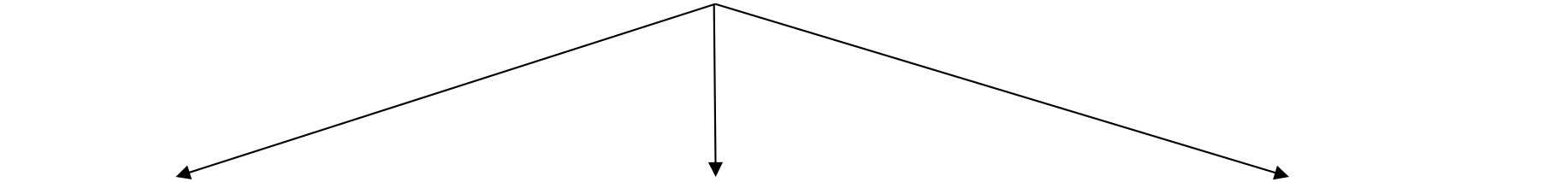
Malicious and Benign Bots

Not all social bots are malicious, some are **neutral** or even **benign:**

- News aggregation and dissemination
- Support and coordination in emergency situations
- and more…



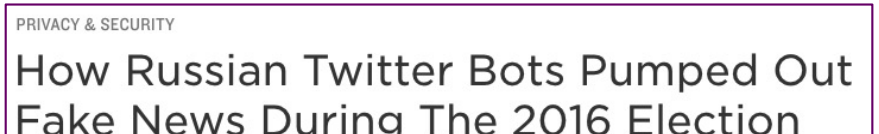
The majority of social bots however is **malicious**!



Malicious Bots

The majority of social bots however is **malicious**!

- Mis- and disinformation
- Opinion manipulation
- Spam
- Polarization
- Hate speech
- Viruses, malware
- Phishing, scams


PRIVACY & SECURITY

How Russian Twitter Bots Pumped Out Fake News During The 2016 Election


Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter

STEFANO CRESCI, Institute of Informatics and Telematics, IIT-CNR, Italy
FABRIZIO LILLO, Department of Mathematics, University of Bologna, Italy and Scuola Normale Superiore of Pisa, Italy
DANIELE REGOLI, Azimut Analytics srl, Milano, Italy and Scuola Normale Superiore of Pisa, Italy
SERENA TARDELLI and MAURIZIO TESCONI, Institute of Informatics and Telematics, IIT-CNR, Italy

Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi. 2019. Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter. *ACM Trans. Web* 13, 2, Article 11 (April 2019), 27 pages.
https://doi.org/10.1145/3313184

11


BBC NEWS
Technology

Massive networks of fake accounts found on Twitter


BY EMILIO FERRARA, ONUR VAROL, CLAYTON DAVIS, FILIPPO MENCZER, AND ALESSANDRO FLAMMINI

The Rise of Social Bots


The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race

Stefano Cresci s.cresci@iit.cnr.it
Roberto Di Pietro roberto.di_pietro@nokia-bell-labs.com
Marinella Petrocchi m.petrocchi@iit.cnr.it
Angelo Spognardi angsp@dtu.dk
Maurizio Tesconi m.tesconi@iit.cnr.it


FINANZA & MERCATI | Così i tweet dei robot insidiano le new

INCHIESTA

Così i tweet dei robot insidiano le news dei listini di Borsa

di Vittorio Carlini | 23 marzo 2018


The Brexit Botnet and User-Generated Hyperpartisan News

2019, Vol. 37(1) 38-54
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439317734157
journals.sagepub.com/home/ssc
SAGE

Marco T. Bastos[I] and Dan Mercea[I]


SOCIAL MEDIA

Facebook and Twitter Bots Are Starting to Influence Our Politics, a New Study Warns

Malicious Bots

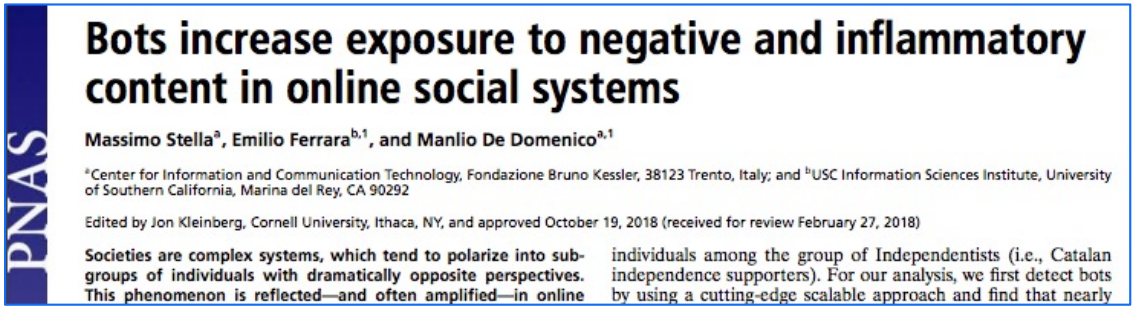
The majority of social bots however is **malicious**!

Shao et al. "The spread of low-credibility content by social bots." *Nature communications* 9.1 (2018)

- Social bots play a disproportionate role in spreading articles from low-credibility sources
- Humans reshare content posted by bots
- Successful low-credibility sources are heavily supported by bots

Malicious Bots

The majority of social bots however is **malicious**!



PNAS

Bots increase exposure to negative and inflammatory content in online social systems

Massimo Stella[a], Emilio Ferrara[b,1], and Manlio De Domenico[a,1]

[a]Center for Information and Communication Technology, Fondazione Bruno Kessler, 38123 Trento, Italy; and [b]USC Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292

Edited by Jon Kleinberg, Cornell University, Ithaca, NY, and approved October 19, 2018 (received for review February 27, 2018)

Societies are complex systems, which tend to polarize into subgroups of individuals with dramatically opposite perspectives. This phenomenon is reflected—and often amplified—in online individuals among the group of Independentists (i.e., Catalan independence supporters). For our analysis, we first detect bots by using a cutting-edge scalable approach and find that nearly

Stella et al. "Bots increase exposure to negative and inflammatory content in online social systems." *Proceedings of the National Academy of Sciences* 115.49 (2018): 12435-12440.

- Bots act from peripheral areas of the social system to target influential humans
- They flood legitimate users with violent content, increasing their exposure to negative and inflammatory narratives
- They exacerbate social conflict online

Malicious Bots

The majority of social bots however is **malicious**!

SOCIAL SCIENCE

The spread of true and false news online

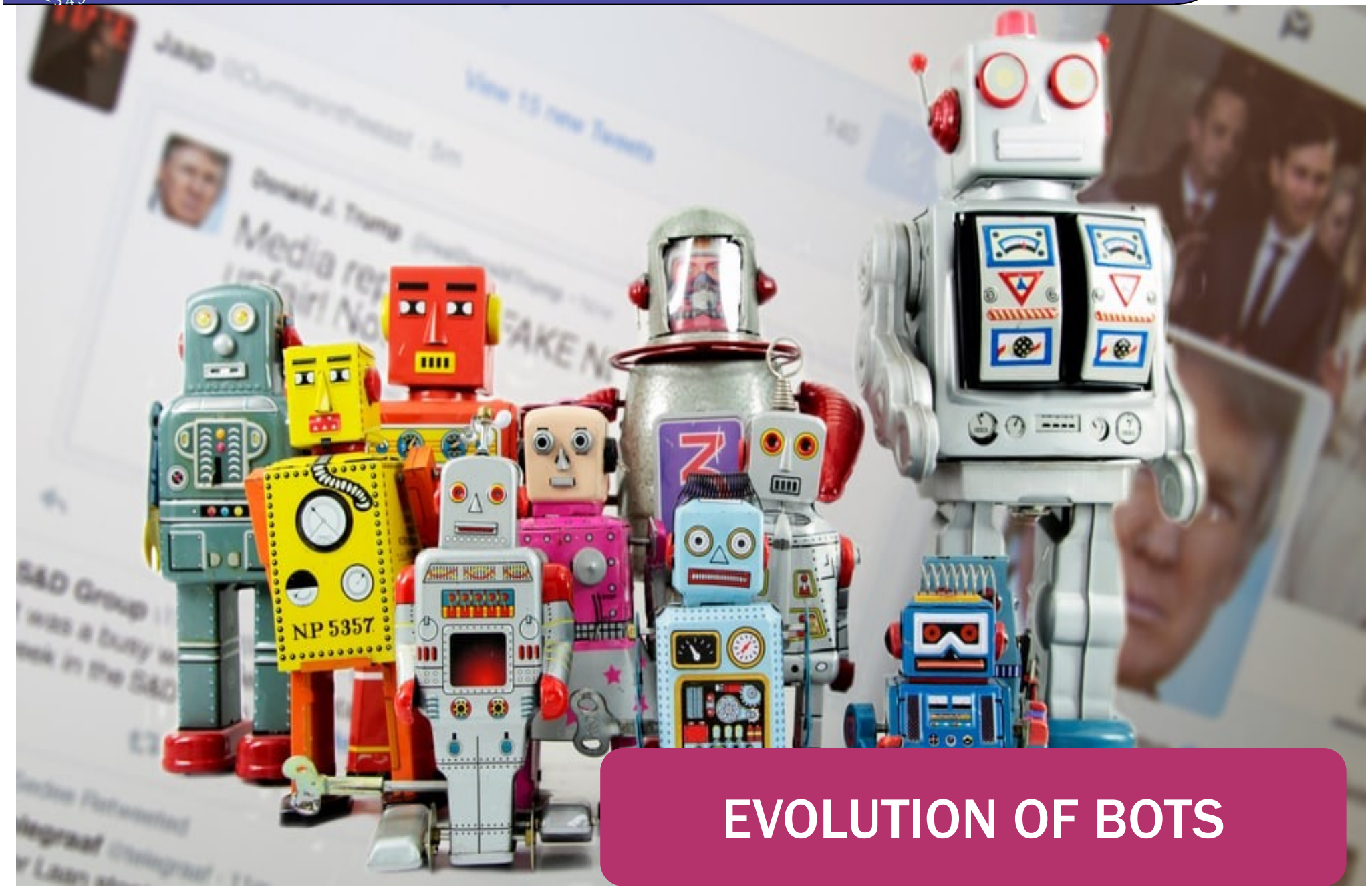
Soroush Vosoughi,[1] Deb Roy,[1] Sinan Aral[2]*

Vosoughi et al. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

- Contrary to conventional wisdom, **robots accelerated the spread of true and false news at the same rate**
- False news spreads more than the truth because humans, not robots, are more likely to spread it

We need some countermeasures

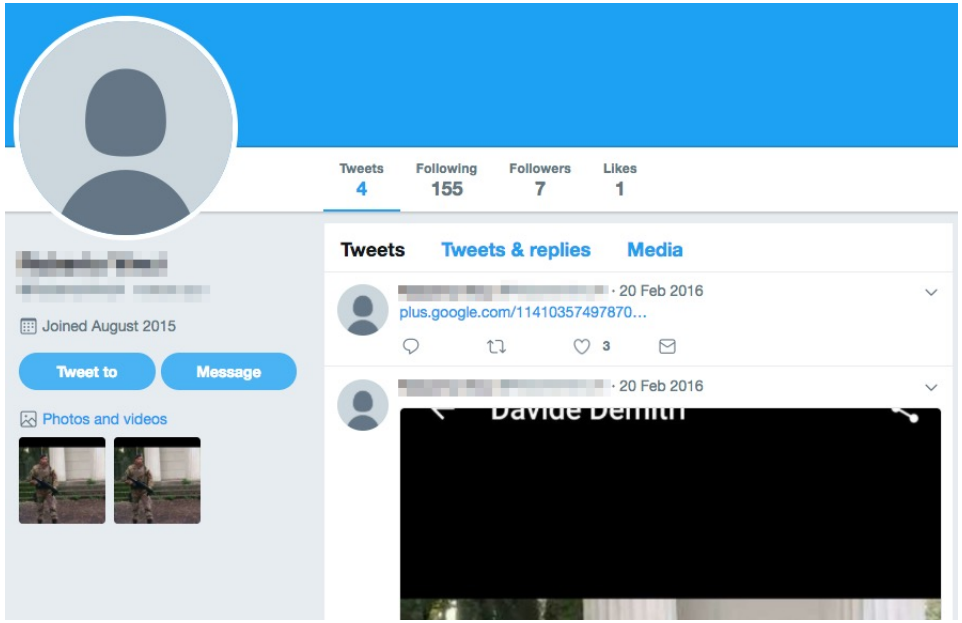
to unmask this automated social manipulation

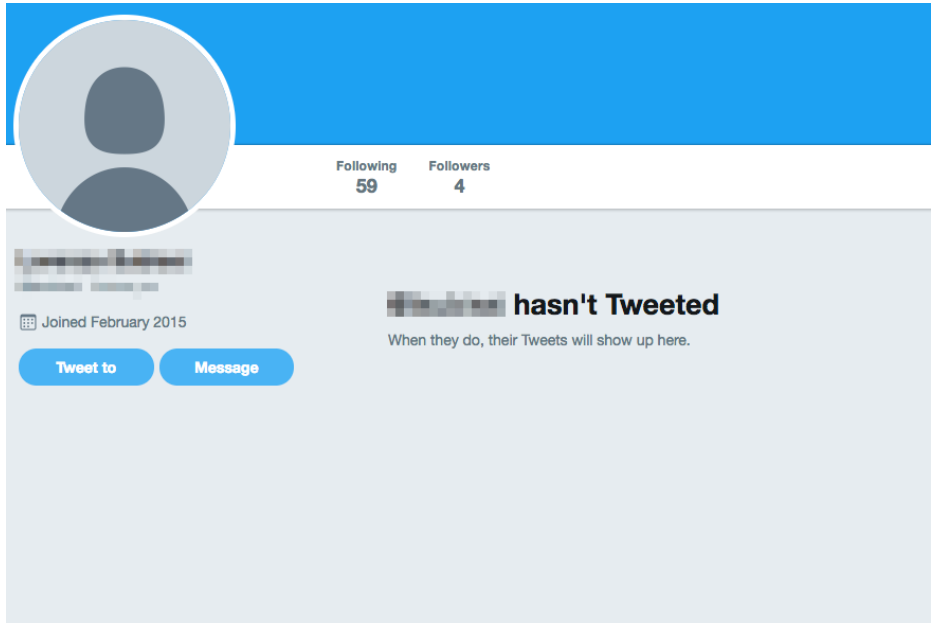


EVOLUTION OF BOTS

Evolution of Bots

Much time ago…

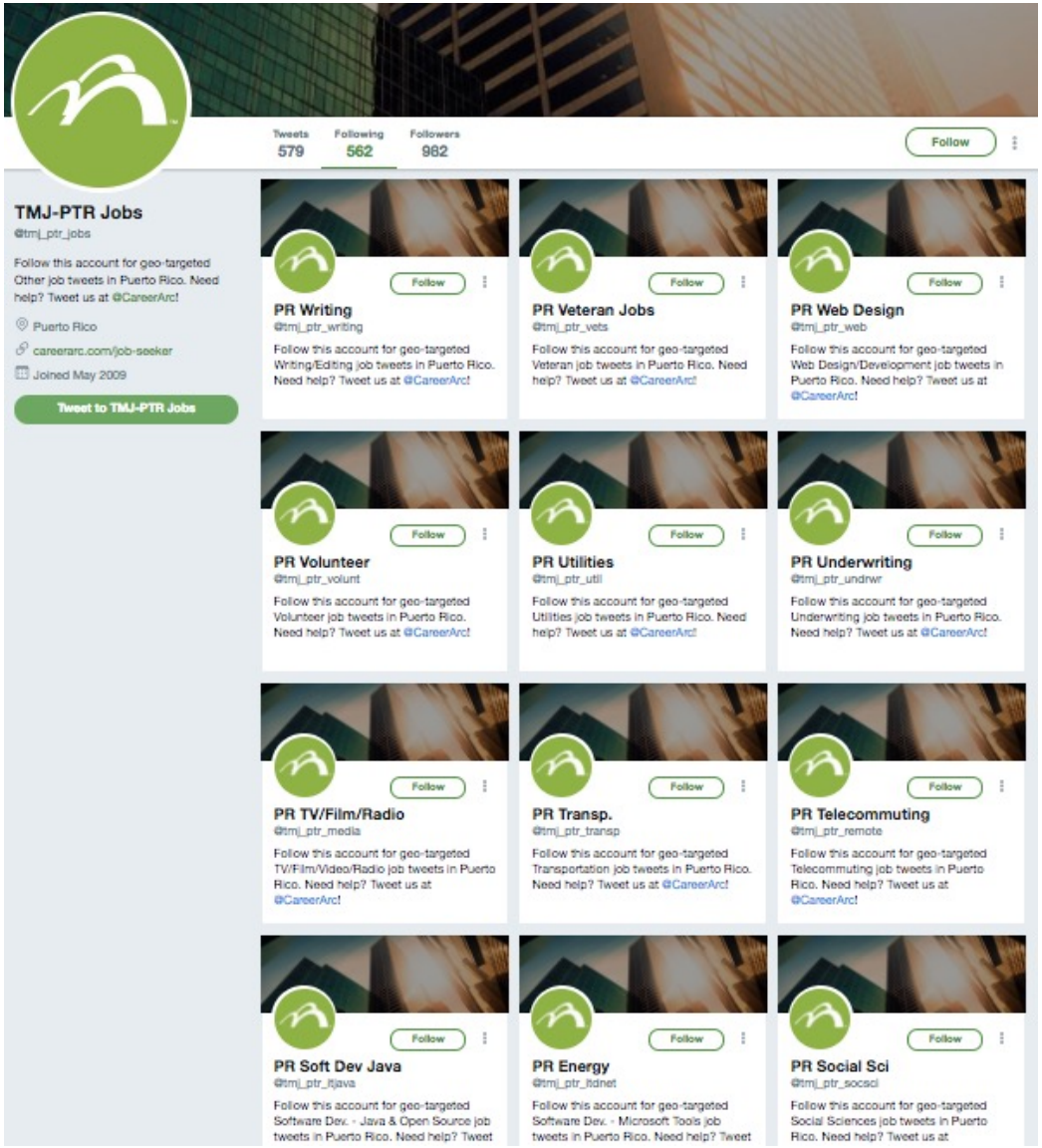




Evolution of Bots

Some time ago…

Botnets

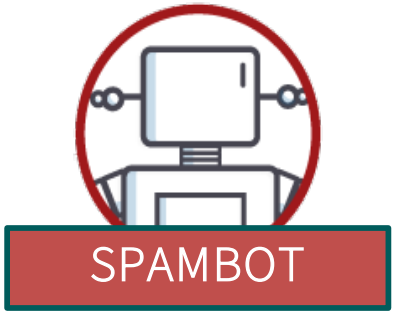

Tweets 579
Following 562
Followers 982
Follow
TMJ-PTR Jobs
@tmj_ptr_jobs
Follow this account for geo-targeted Other job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
Puerto Rico
careerarc.com/job-seeker
Joined May 2009
Tweet to TMJ-PTR Jobs
PR Writing
@tmj_ptr_writing
Follow this account for geo-targeted Writing/Editing job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Veteran Jobs
@tmj_ptr_vets
Follow this account for geo-targeted Veteran job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Web Design
@tmj_ptr_web
Follow this account for geo-targeted Web Design/Development job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Volunteer
@tmj_ptr_volunt
Follow this account for geo-targeted Volunteer job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Utilities
@tmj_ptr_util
Follow this account for geo-targeted Utilities job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Underwriting
@tmj_ptr_undrwr
Follow this account for geo-targeted Underwriting job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR TV/Film/Radio
@tmj_ptr_media
Follow this account for geo-targeted TV/Film/Video/Radio job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Transp.
@tmj_ptr_transp
Follow this account for geo-targeted Transportation job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Telecommuting
@tmj_ptr_remote
Follow this account for geo-targeted Telecommuting job tweets in Puerto Rico. Need help? Tweet us at @CareerArc!
PR Soft Dev Java
@tmj_ptr_itjava
Follow this account for geo-targeted Software Dev. - Java & Open Source job tweets in Puerto Rico. Need help? Tweet
PR Energy
@tmj_ptr_itdnet
Follow this account for geo-targeted Software Dev. - Microsoft Tools job tweets in Puerto Rico. Need help? Tweet
PR Social Sci
@tmj_ptr_socsci
Follow this account for geo-targeted Social Sciences job tweets in Puerto Rico. Need help? Tweet us at

Novel Social Bots

- Social bots are organized **in groups**, with specific purposes (botnets)
- They have a **detailed** profile, with:
  - a profile picture (*stolen…*);
  - a biography (*fake…*);
  - many followers and friends (*real!*)
- They post famous quotes, videos, memes, …
- They are almost **indistinguishable** from legitimate accounts!

Novel Social Bots

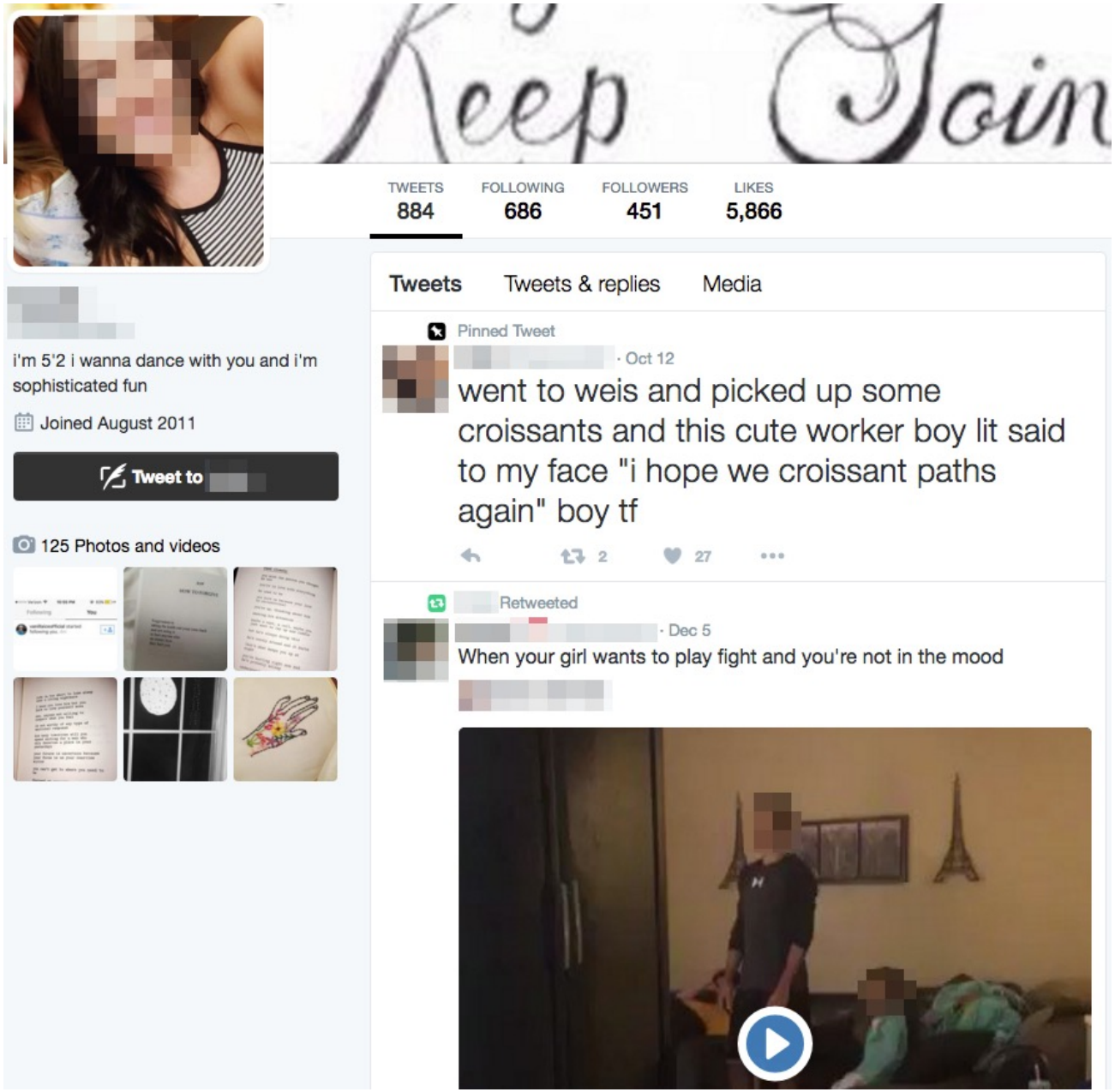
bot or real?



Novel Social Bots

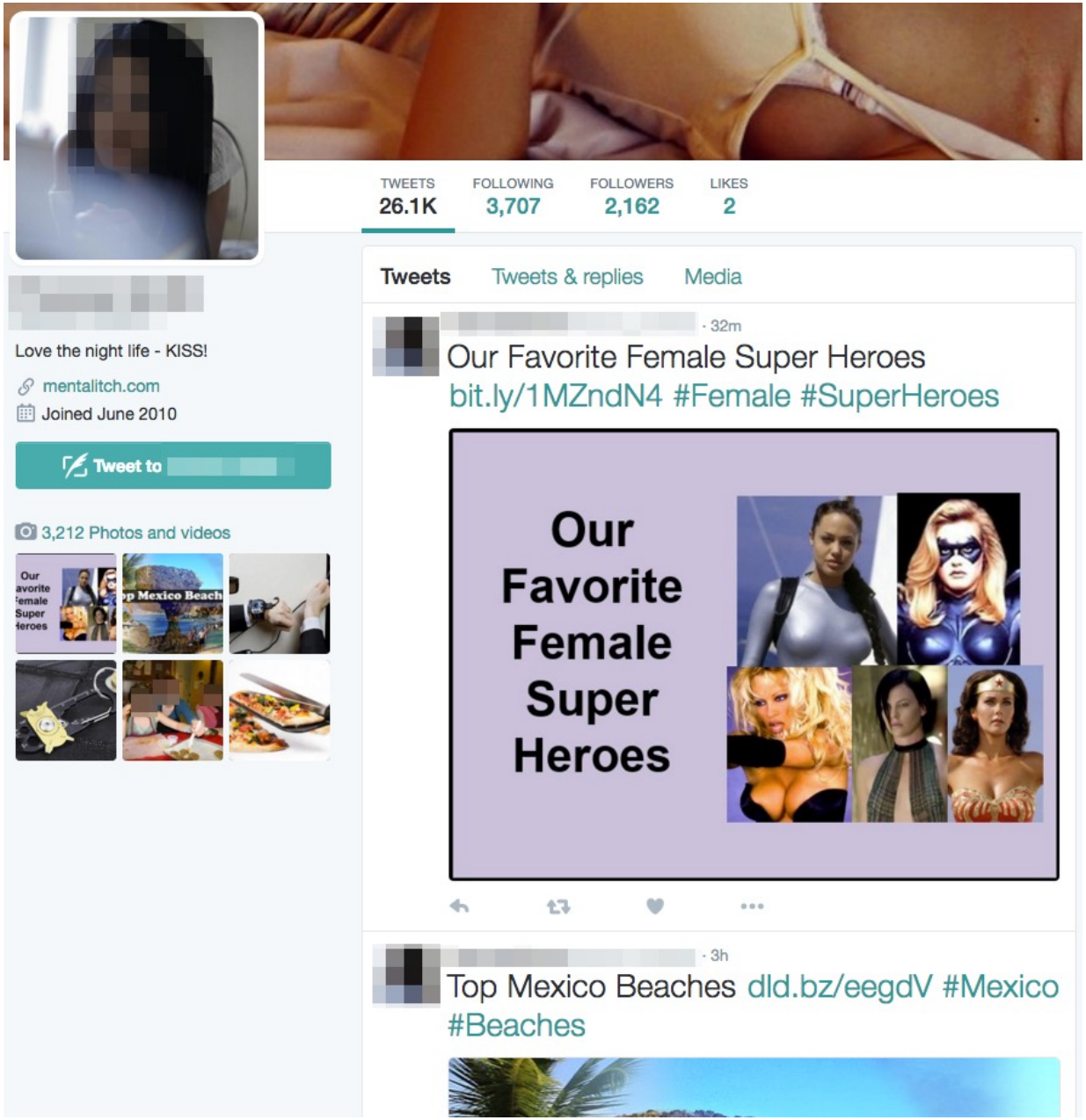
bot or real?

Novel Social Bots

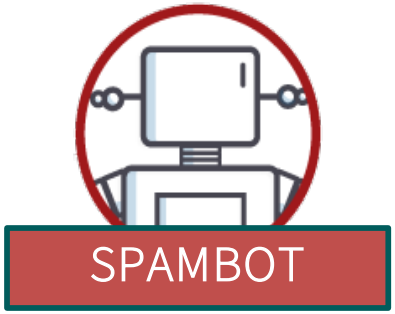


bot or real?

Novel Social Bots



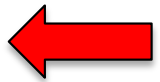
bot or real?



Crowdsourcing experiment

- Are users capable of **detecting social bots**?
- 3 social bots out of 4 **go undetected!**

Detection accuracy (acc)

- old bots **91%** 
- social bots **24%**  
- real accounts **92%** 

Cresci, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." *Proceedings of the 26th International Conference on World Wide Web Companion*. ACM, 2017.

Crowdsourcing experiment

- Are users capable of **distinguishing old bots, novel social bots, and legitimate accounts**?
- Humans were more in agreement when **wrongly** classifying social bots (acc. = 0.2, k = 0.2) than when **correctly** classifying old bots (acc. = 0.9, k ~ 0)

Human classification agreement (k)

- old bots **0.007 (very low)**
- social bots **0.186 (low)**
- real accounts **0.410 (moderate)**

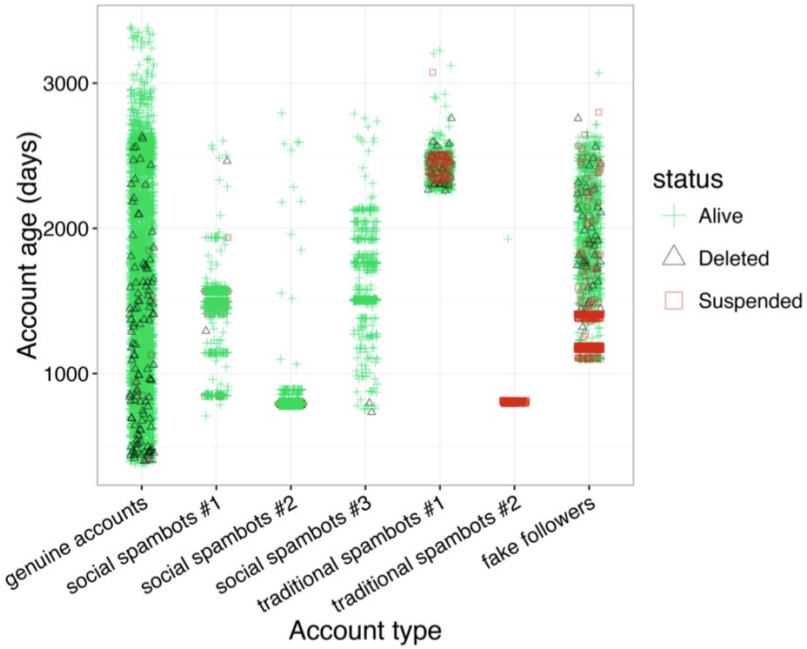
Cresci, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." ***Proceedings of the 26th International Conference on World Wide Web Companion*****. ACM, 2017.**

Twitter Defenses

- Twitter is not capable of effectively removing novel social bots

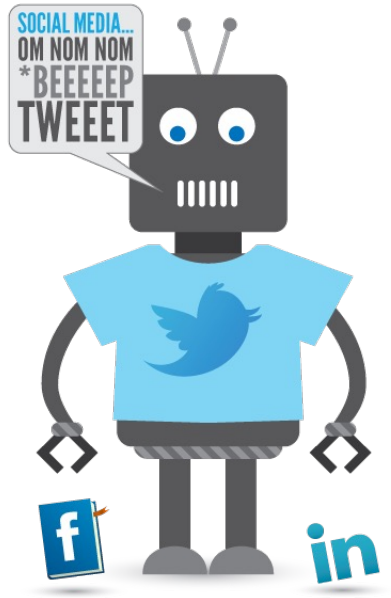
Twitter removal rate

- old bots **60%**
- social bots **4%**



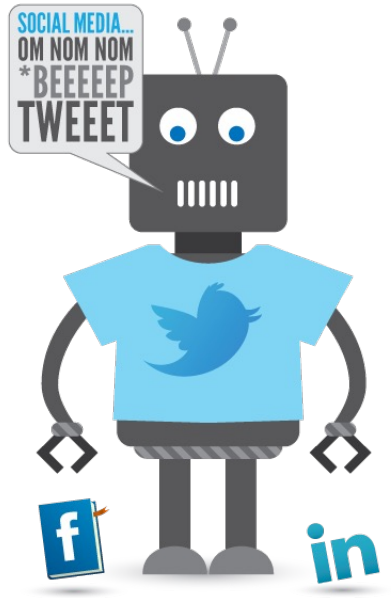
Detection Techniques

- **Naive detection**: all the accounts that interact in the same timestamp (second) are bots
- Helps to easily discover self-declared bots (not deceiving bots)
- A large number of novel bots remain undetected



Detection Techniques

- The majority of detection techniques are **based on machine learning,** often **supervised** (*e.g., classification algorithms)*
- Each account is analyzed **singularly**
- Usage of **general-purpose** ML algorithms (*e.g., decision trees, random forest, SVM)*
- The focus is on **features** rather than algorithms (*i.e., which account characteristics can be exploited by the detector?)*



Botometer

Botometer®

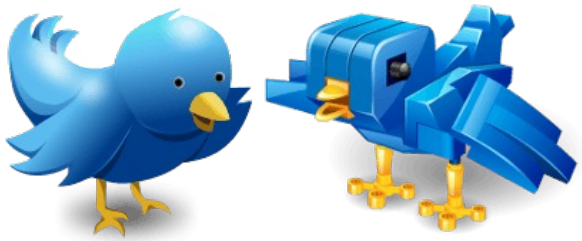
An OSoMe project (bot•o•meter)

- Most widely used social bot classification system
  - Web application: *https://botometer.iuni.iu.edu/#!/*
  - REST API: *https://botometer.iuni.iu.edu/#!/api*

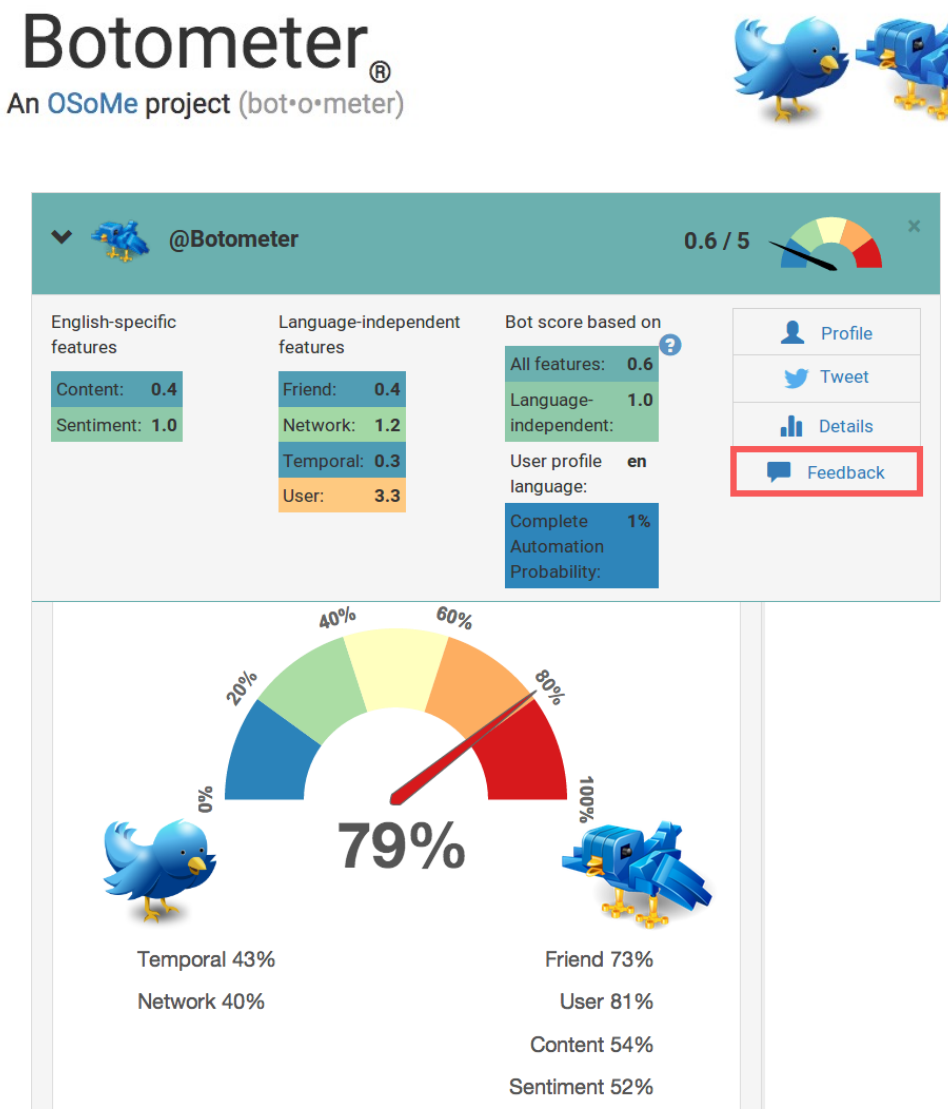
- Ensemble of **supervised Random Forest** classifiers

Botometer

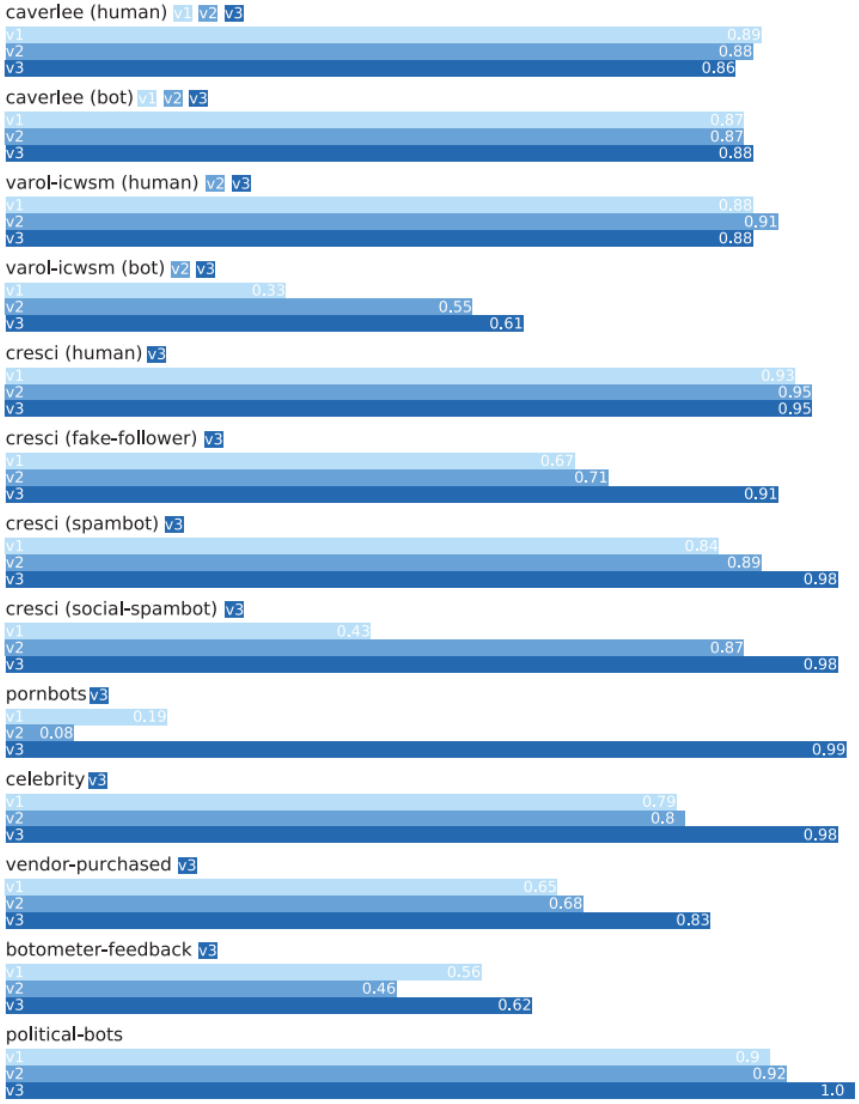
- More than **1,200 features** divided in 6 groups:
  1. **Network**: built from retweets, mentions, and hashtag co-occurrences
  2. **User**: language, geographic locations, account creation time, …
  3. **Friends**: statistical properties of friends (number of followers and posts, …)
  4. **Temporal**: timing patterns of posts (tweet rate and inter-tweet times, …)
  5. **Content**: typical natural language processing (NLP) features
  6. **Sentiment**: emoticon scores, overall happiness, ...



Botometer (results)



interface



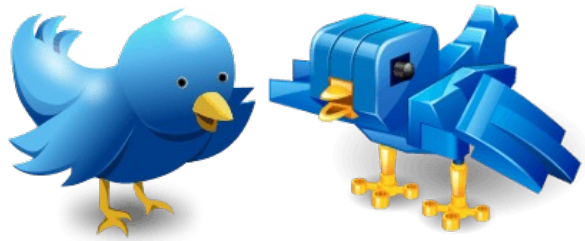
system evaluation (accuracy)

Detection with Novel Bots

- Botometer with novel social bots achieve moderate unsatisfactory results
- However, it is the best state-of-the-art technique

Performance against novel bots (acc.)

- Botometer **53%**
- classification techniques **39%**
- clustering techniques **40%**



Novel social bots could not be analyzed account-by-account
We need to analyze the collective behaviours of groups of users

GROUP DETECTION

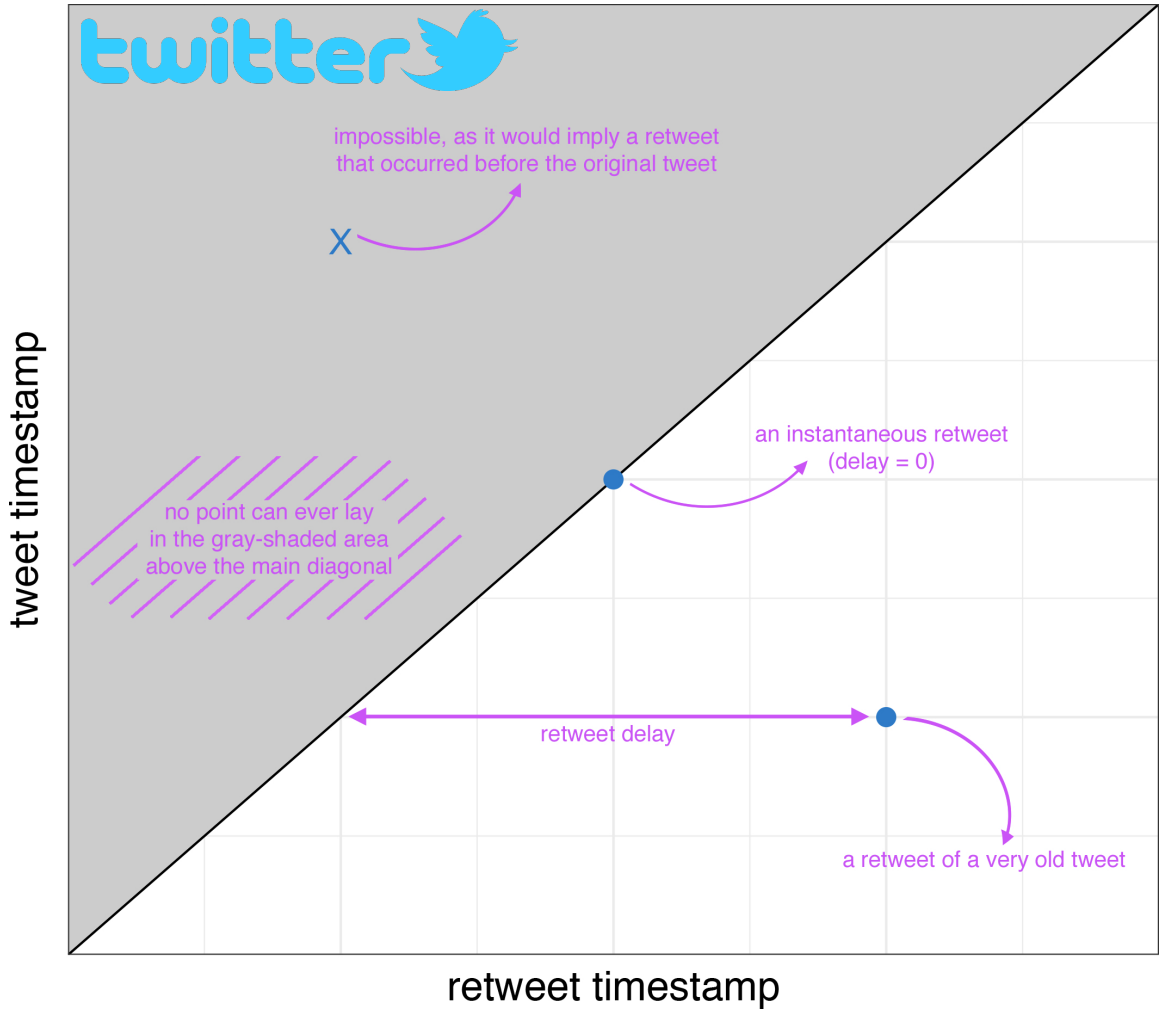
Group Analysis

- A large enough group of bots **will still leave traces of automation**
- Group of bots share a common purpose
  - f.e. increasing someone's popularity
- Groups are subjected to **synchronized or coordinated** behaviours

Temporal Behaviours

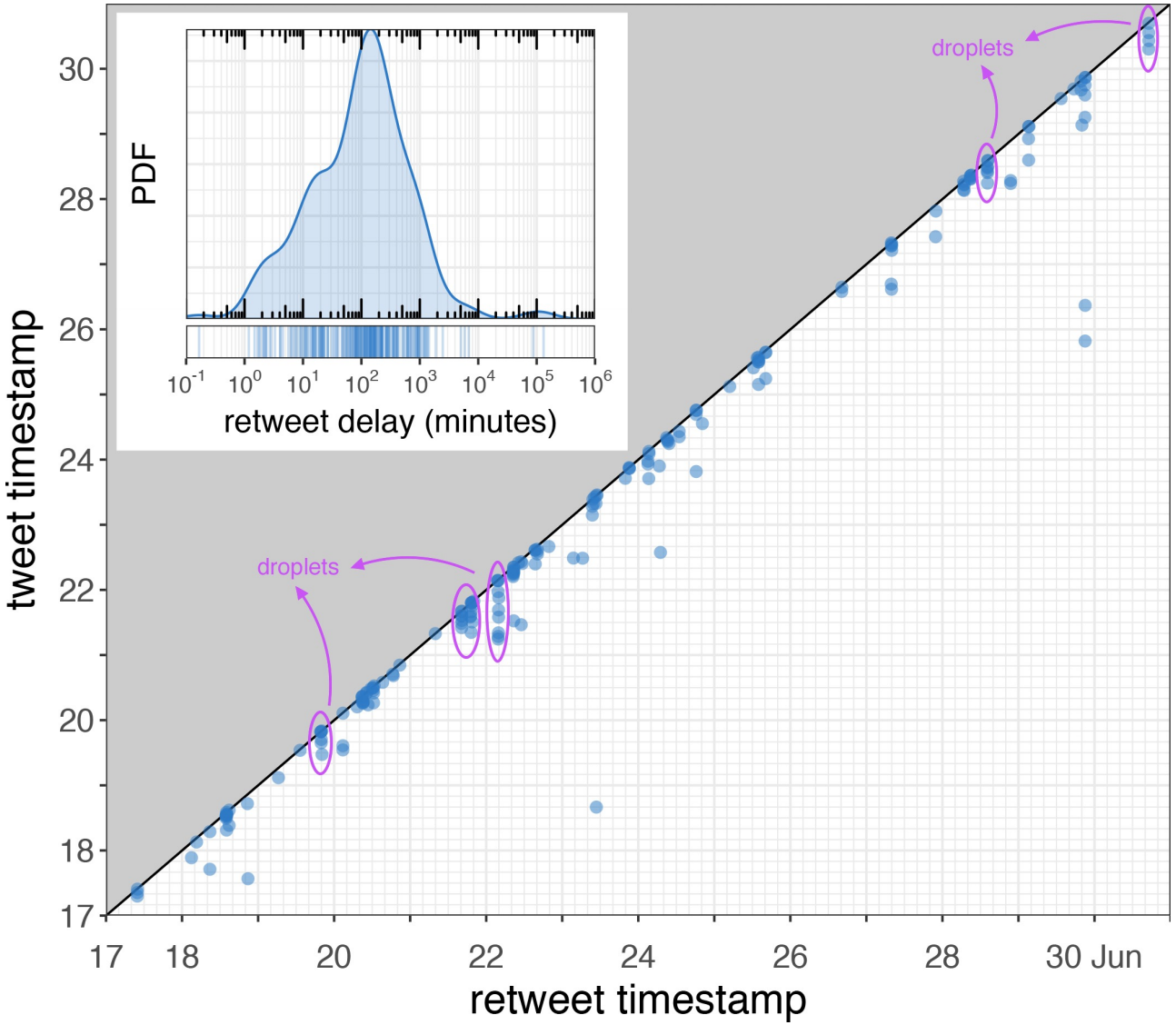
How can we recognize bots?

- *Intuition*: many bots **coordinatedly** retweet the same content



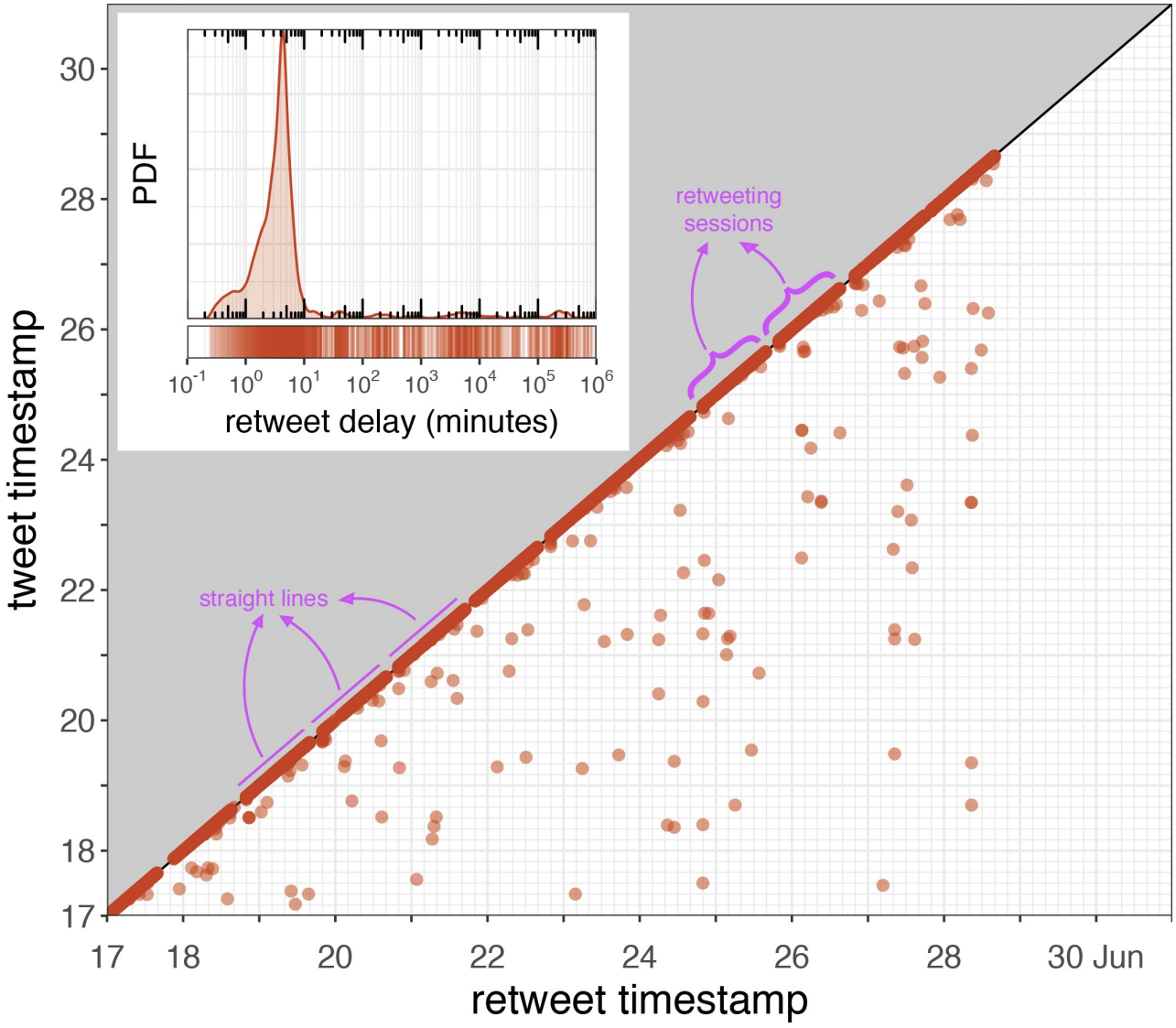
Temporal Behaviours

Normal Behaviour



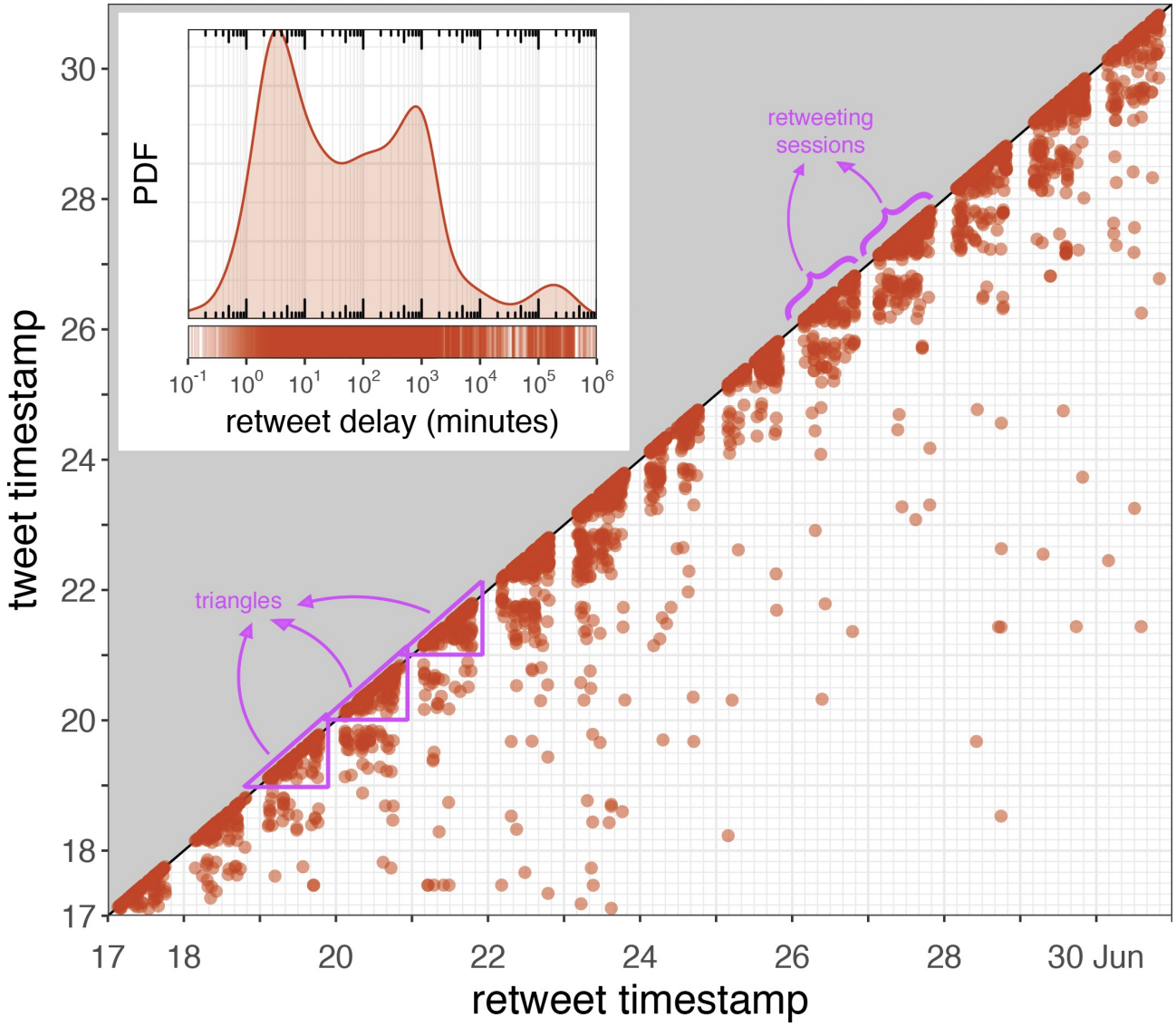
Temporal Behaviours

Suspicious Behaviour (straight line)



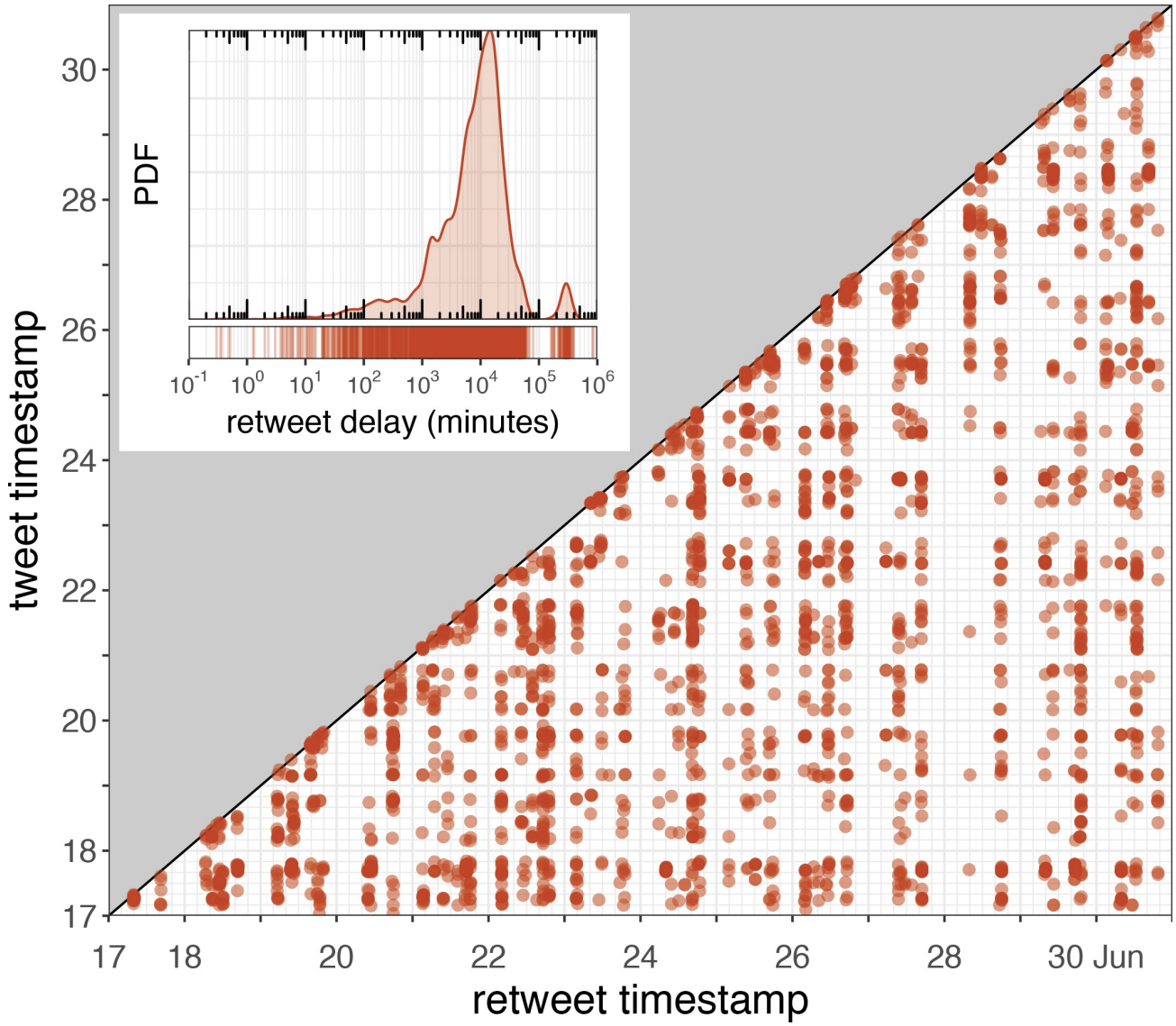
Temporal Behaviours

Suspicious Behaviour (triangle)

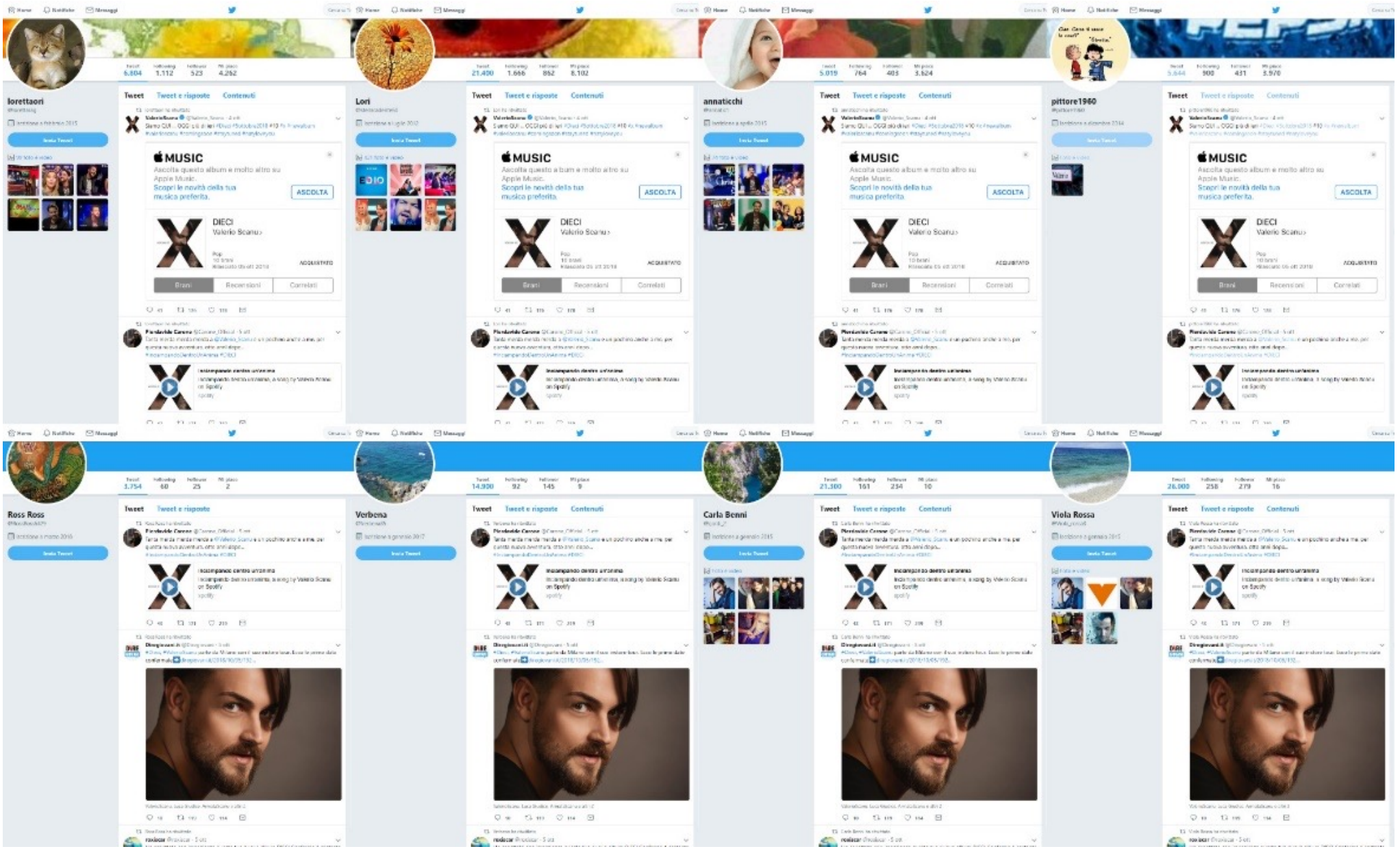


Temporal Behaviours

Suspicious Behaviour (waterfall)



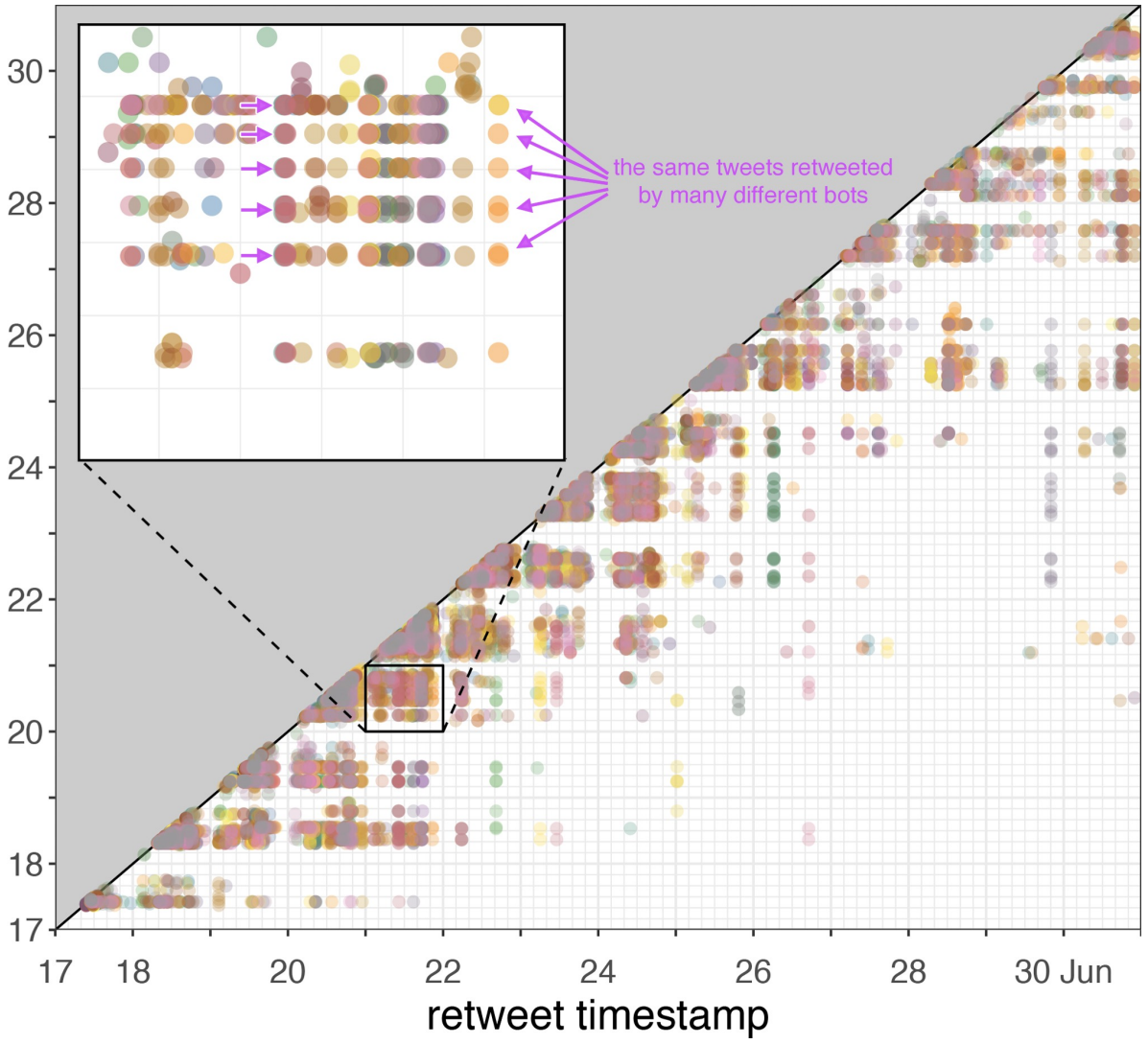
Real Examples



Real Examples

Singer botnet (Twitter)

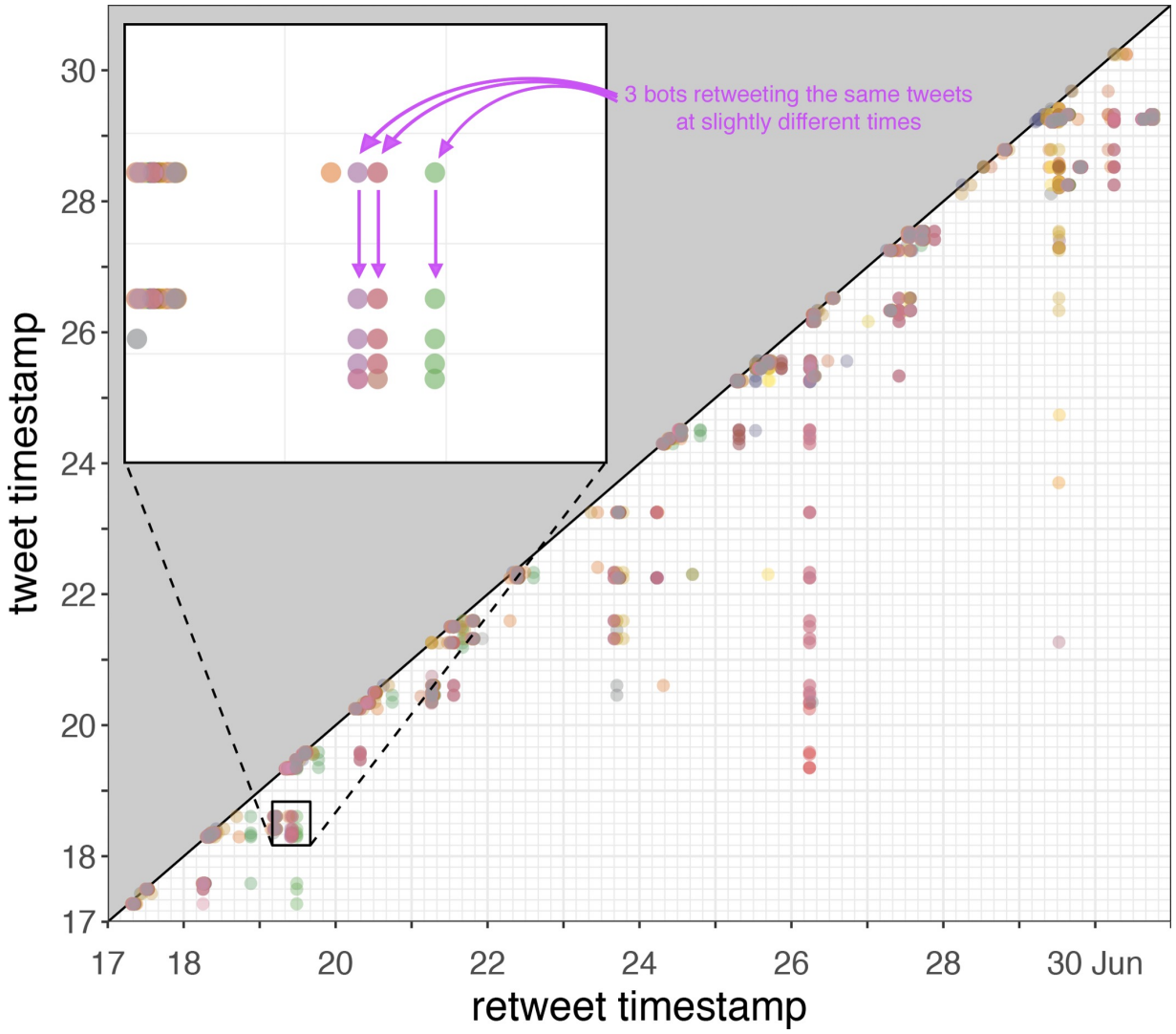
- almost 300 bots
- they only retweeted 2 accounts related to an italian pop singer:
  - @Valerio_Scanu
  - @ArmataScanu



Real Examples

Cars botnet (Twitter)

- group of 44 bots
- they only retweeted 3 accounts:
  - @citroenitalia
  - @peugeotitalia
  - @motorionline



Action-based Behaviours

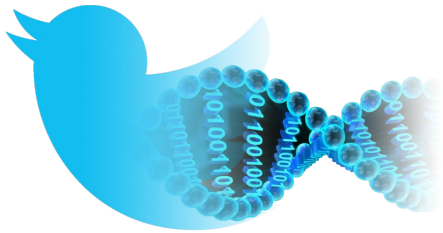
DIGITAL DNA

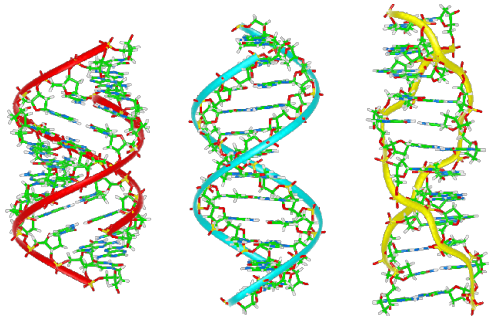
T ← tweet,
R ← retweet,
P ← reply

A ← adenine,
G ← guanine,
T ← thymine,
C ← cytosine

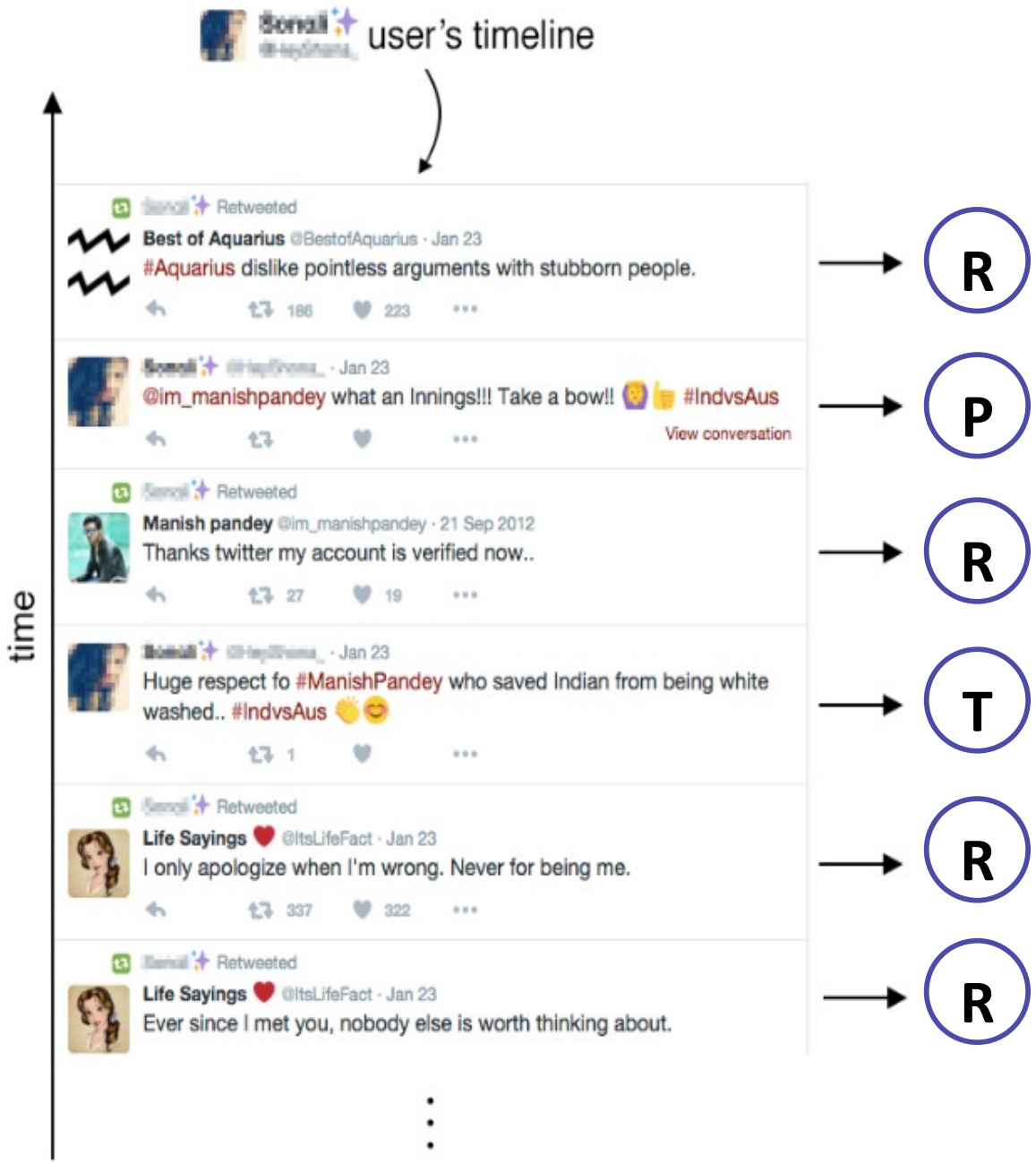
…RRTRPRTPRRPRTPRPTPRRTRPR
…RPRTPTTRPTRPTPRRRRTPPRPP
…TTTRRRPPTPRPTPRTRPTRRRTP
…PRTRPRTPPPPRTPRRPRTPPRRT
…TRTRPRTPRRPRTPRPTPTPPRTT
…TRPPRTPPTRPPTPRRTTTPPRPR

…AGTCTCCATTTTCAGGTCGTA
…GTTTAAGATCGCCTCATCACC
…AGGCAATTCGCCTGAACTGG
…AGTCTCGATCCTTTCCTCGTT
…AAAATCGAACGCCTTGTCGG
…ATTCTCCATCGCCTAAACAAC





Digital DNA



T ← tweet,
R ← retweet,
P ← reply

DIGITAL DNA:
…RRTRPR…

Digital DNA

How can we recognize bots?

- *Intuition*: automated accounts have similar sequences
- Computation of the Longest Common Substring (LCS)

…TRRRPRRTRRPRTPRPTPRRTRPR
…RPRTPTTRRRPRRTPRRRRTPPRP
…TTTRRRPRRRPRRTRTRPTRRRTP
…PRTRPRTPPPPRTPRRRRRPRRTR

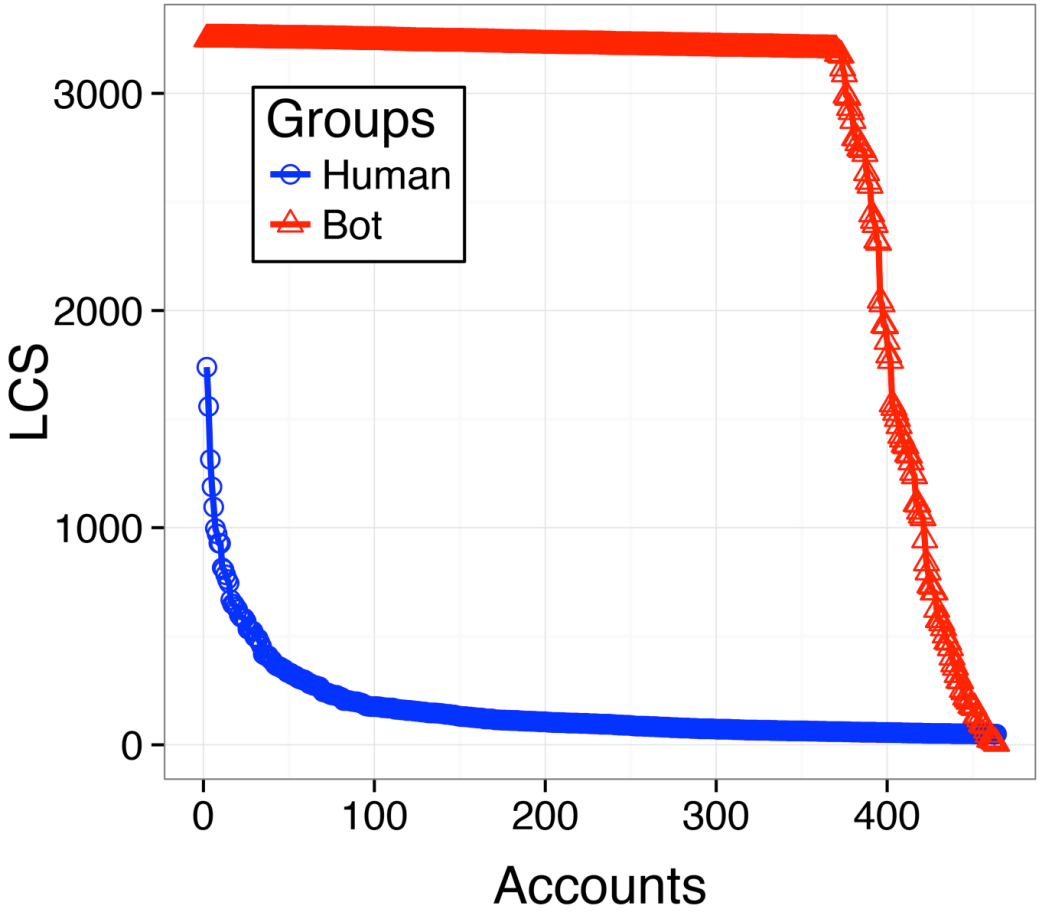
RRRPRRT

LCS: 7

Digital DNA

How can we recognize bots?

- *Intuition*: automated accounts have similar sequences

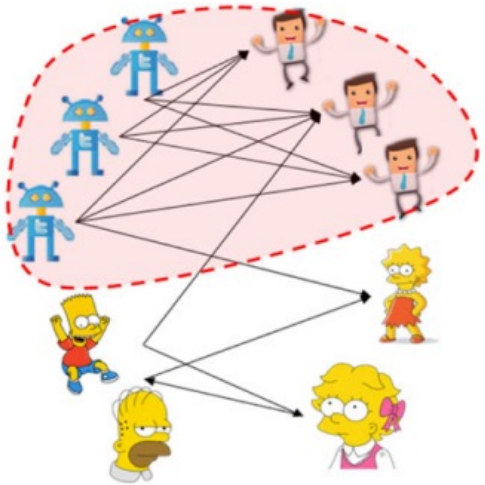


Network-based Behaviours

How can we recognize bots?

- Possible networks:
  - retweeters of a set of tweets;
  - followers of a set of accounts;
  - likers of a set of pages on Facebook;
  - users that shared a set of URLs;
  - …

Twitter-Style Network



Undirected/directed Graph

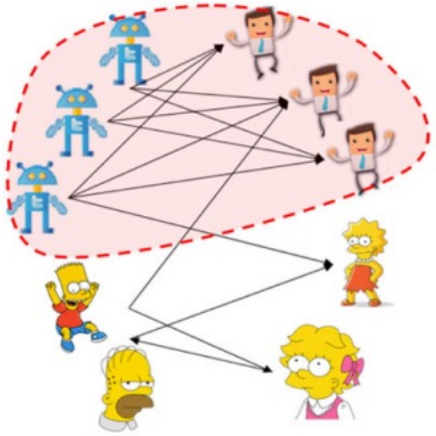
	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0

Network-based Behaviours

Intuition: automated accounts have patterns in the **adjacency matrix**

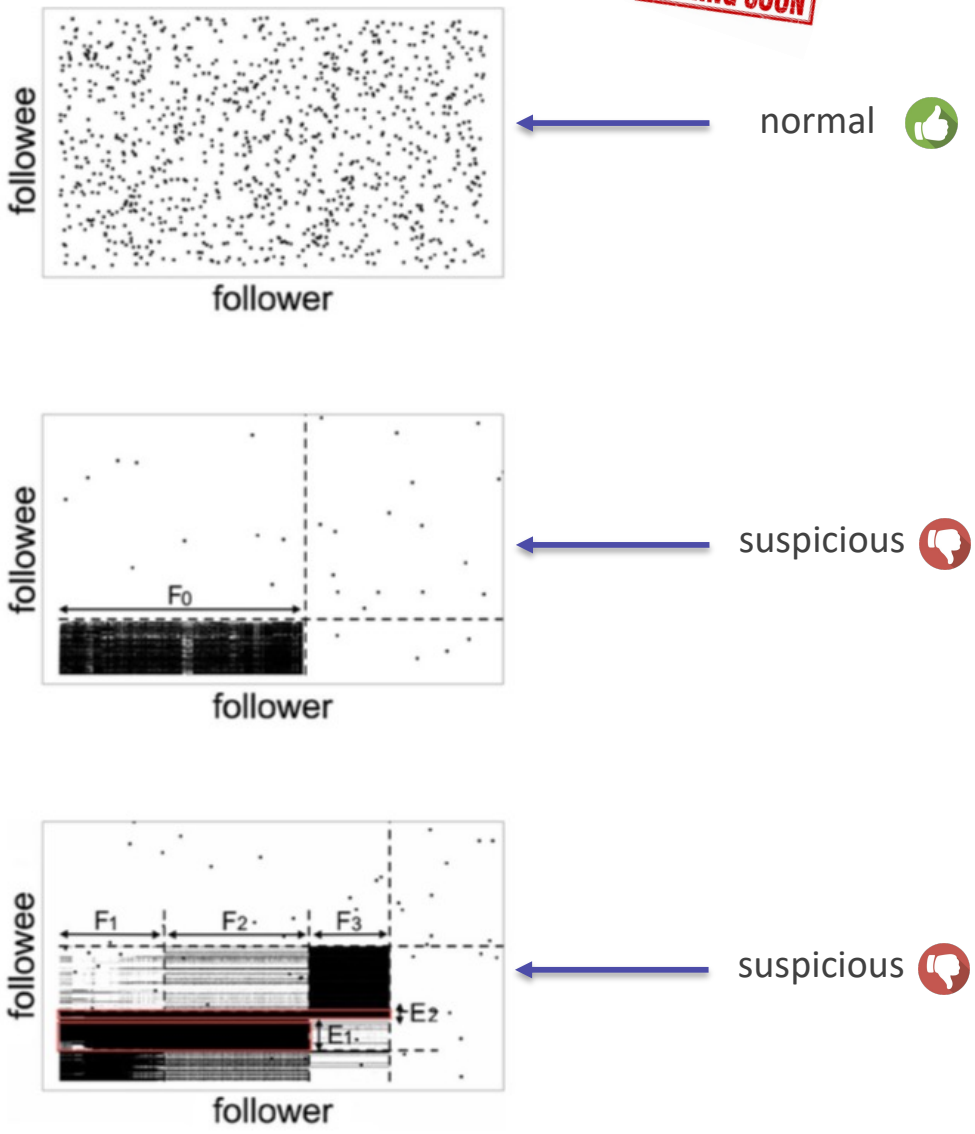
COMING SOON

Twitter-Style Network



Undirected/directed Graph

	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0



And Now?

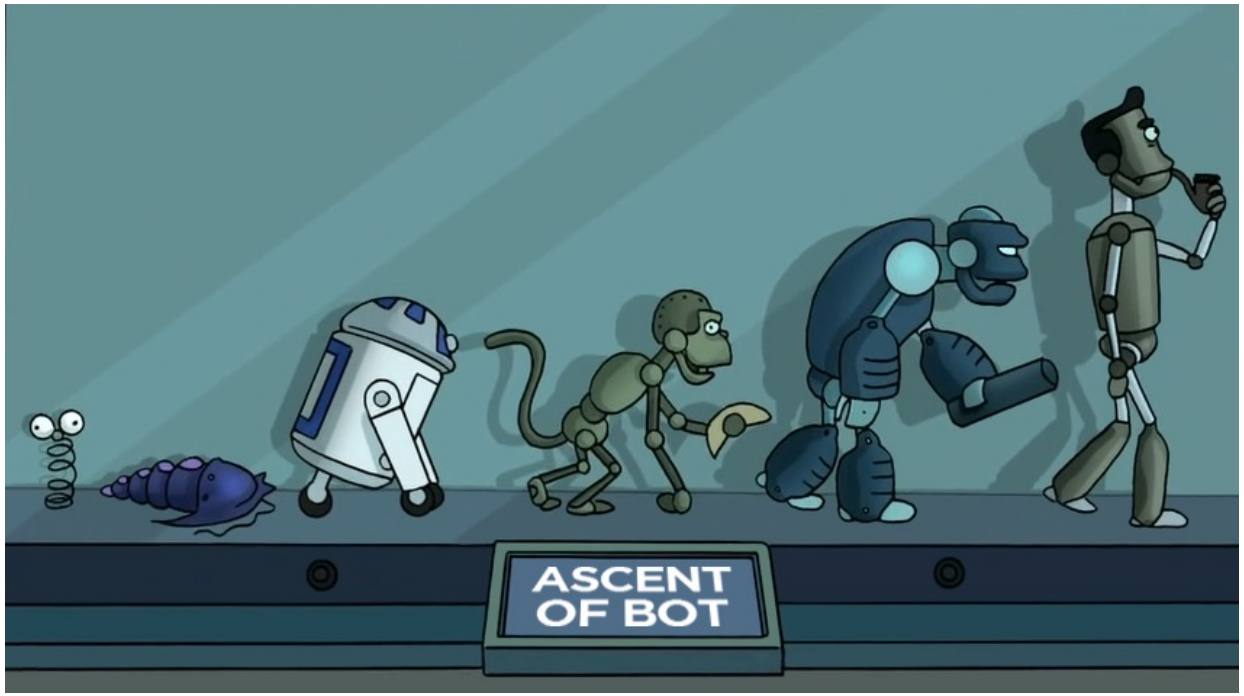
Is the problem solved now? **NO**

- No systematic comparison of **novel** vs **traditional** approaches
- Novel approaches produce **great improvements** in detection

And Now?

Is the problem solved now? **NO**

- However, malicious accounts evolve, in order to evade established detection techniques
- New accounts are much more sophisticated than previous ones



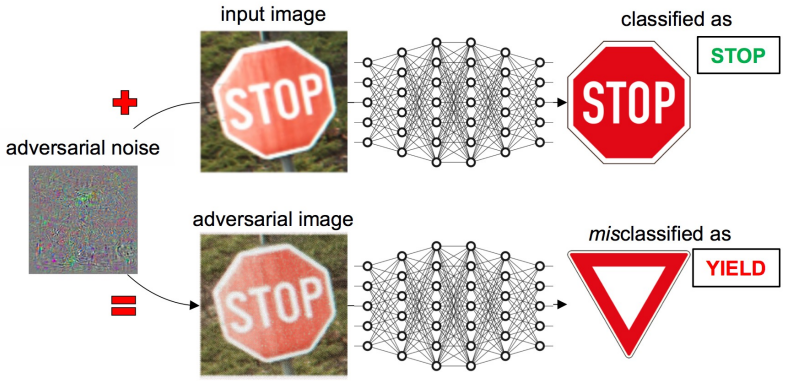
And Now?

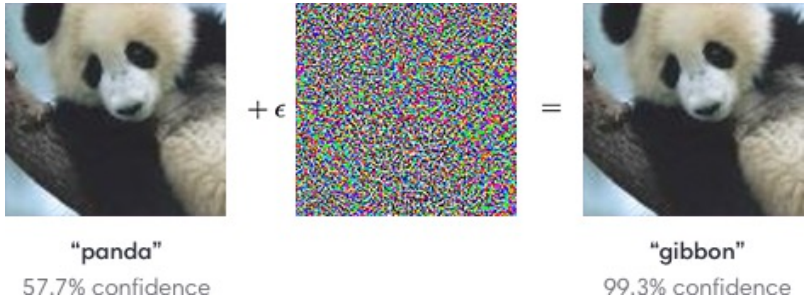
Is the problem solved now? **NO**

- Anyway, the bar is raised for bot developers

Advanced Modern Approaches

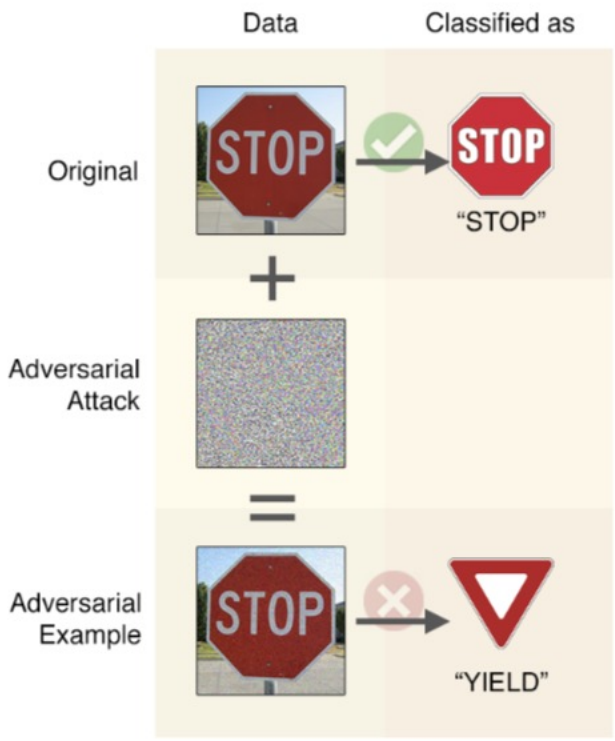
- Many machine learning techniques were originally designed for **stationary** and **benign** environments
- When these assumptions are violated techniques start making big mistakes (*e.g., wrong classifications)*
- Bot detection is neither stationary (**bots evolve**) nor benign (bots try to **evade detection**)!



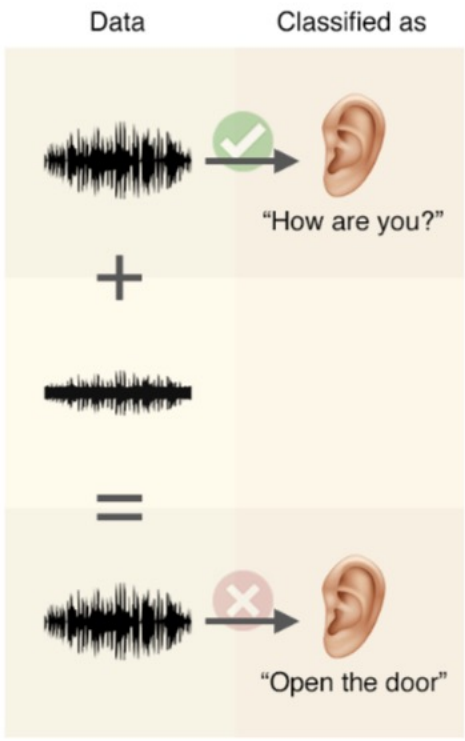


Advanced Modern Approaches

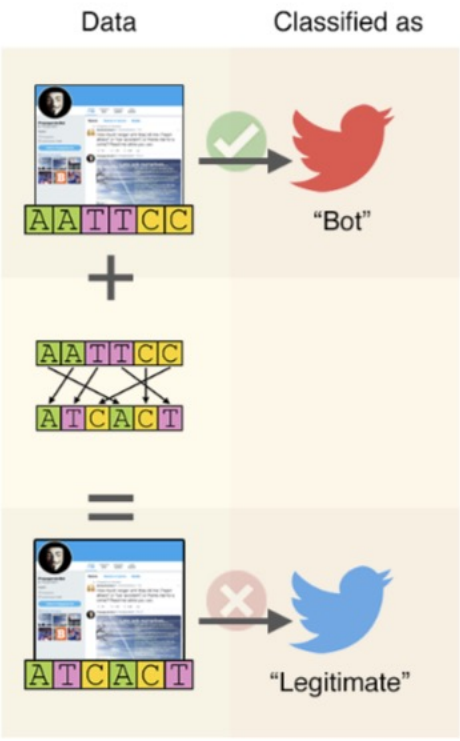
bot evolutions = adversarial attack to bot detectors



(a) Computer vision. Images can be modified by adding adversarial noise so as to fool image classification systems (e.g., those used by autonomous vehicles).



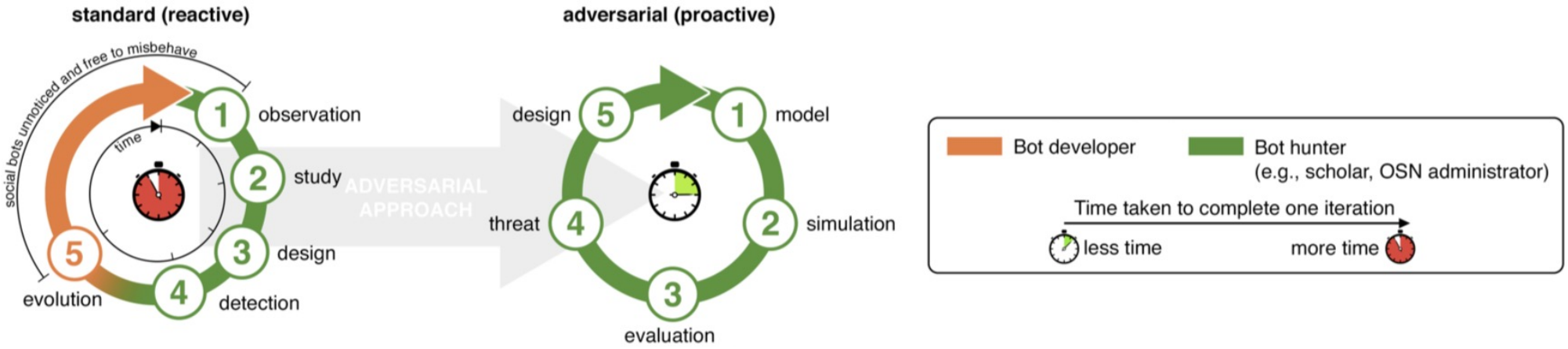
(b) Automatic speech recognition. Adding adversarial noise to a speech waveform may result in wrong textual translations.

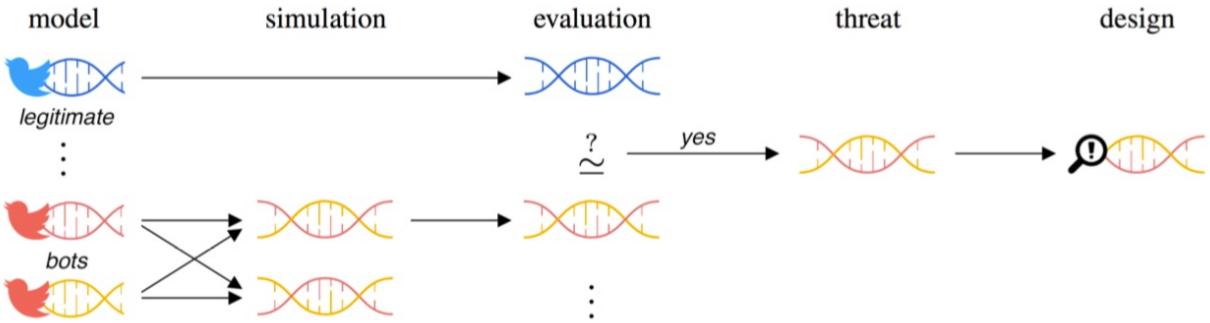


(c) Social bot detection. Similarly to computer vision and automatic speech recognition, adversarial attacks can alter the features of social bots, without impacting their activity, thus allowing them to evade detection.

Adversarial Bot Detection

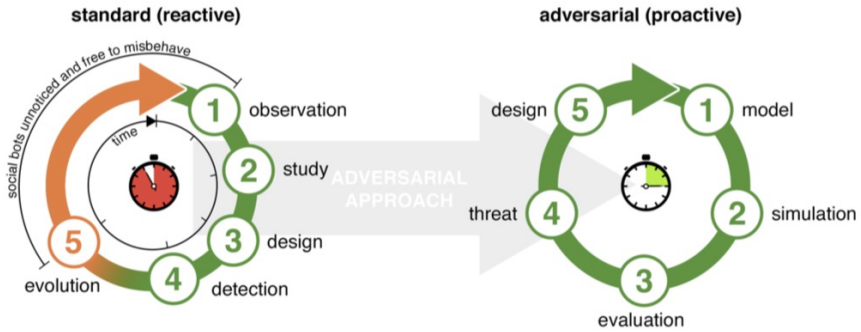
The purpose now is to develop a **proactive approach** reinforcing novel techniques (temporal, action-based, network-based…)





Adversarial Bot Detection

The purpose now is to develop a **proactive approach** reinforcing novel techniques (temporal, action-based, network-based…)

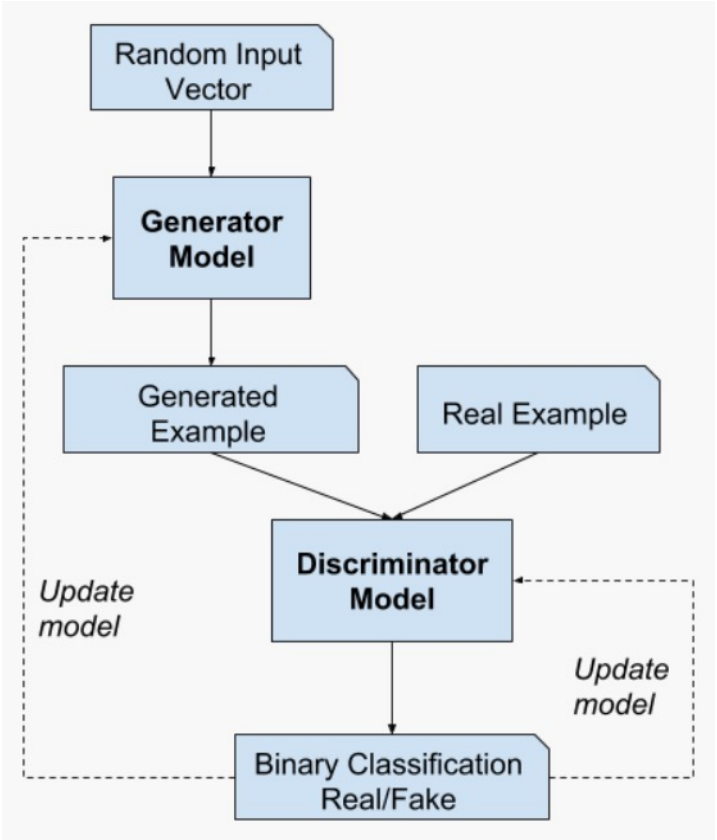


- *model* → digital DNA
- *simulation* → genetic algorithms ("evolved" bots)
- *evaluation* → bot detection technique (e.g., social fingerprinting)
- *threat* → "evolved" bot that evades detection
- *design* → new or modified technique that detects the "evolved" bot

Adversarial Bot Detection

The proactive approach is based on the usage of **Generative Adversarial Nets (GANs)**

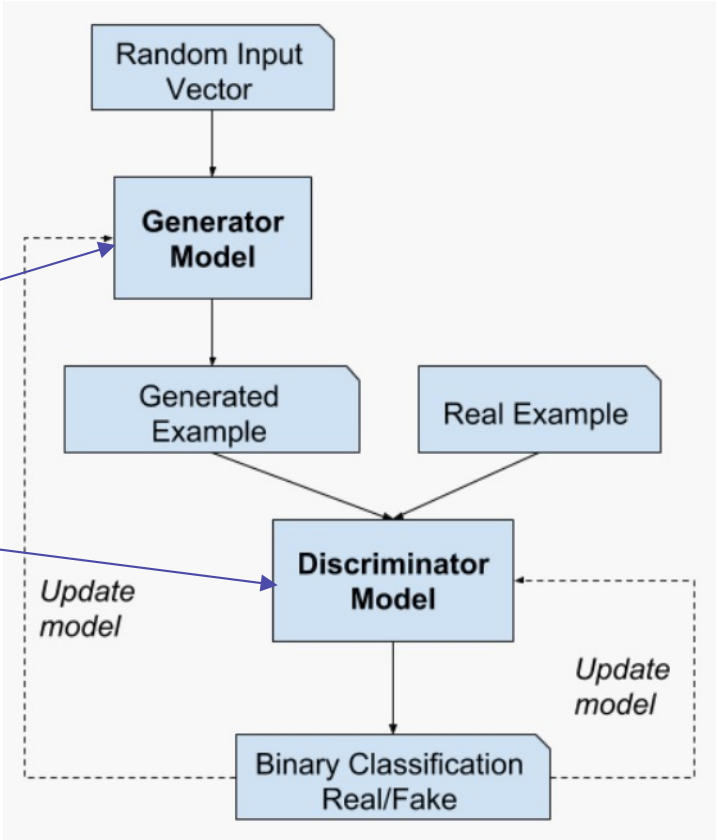
- GANs are based on a game theoretic scenario where **the generator competes against the discriminator**
- (typical) goal: **train the generative model**
- training stops when the discriminator **cannot tell apart real samples from fake** ones



Adversarial Bot Detection

The proactive approach is based on the usage of **Generative Adversarial Nets (GANs)**

- a bot creator competes against a bot detector



- goal: **train a better** (more robust) **bot detector**

Adversarial Bot Detection

Anyway, a continue evolution in the techniques is required!