

Social Bot Detection

Lorenzo Cima

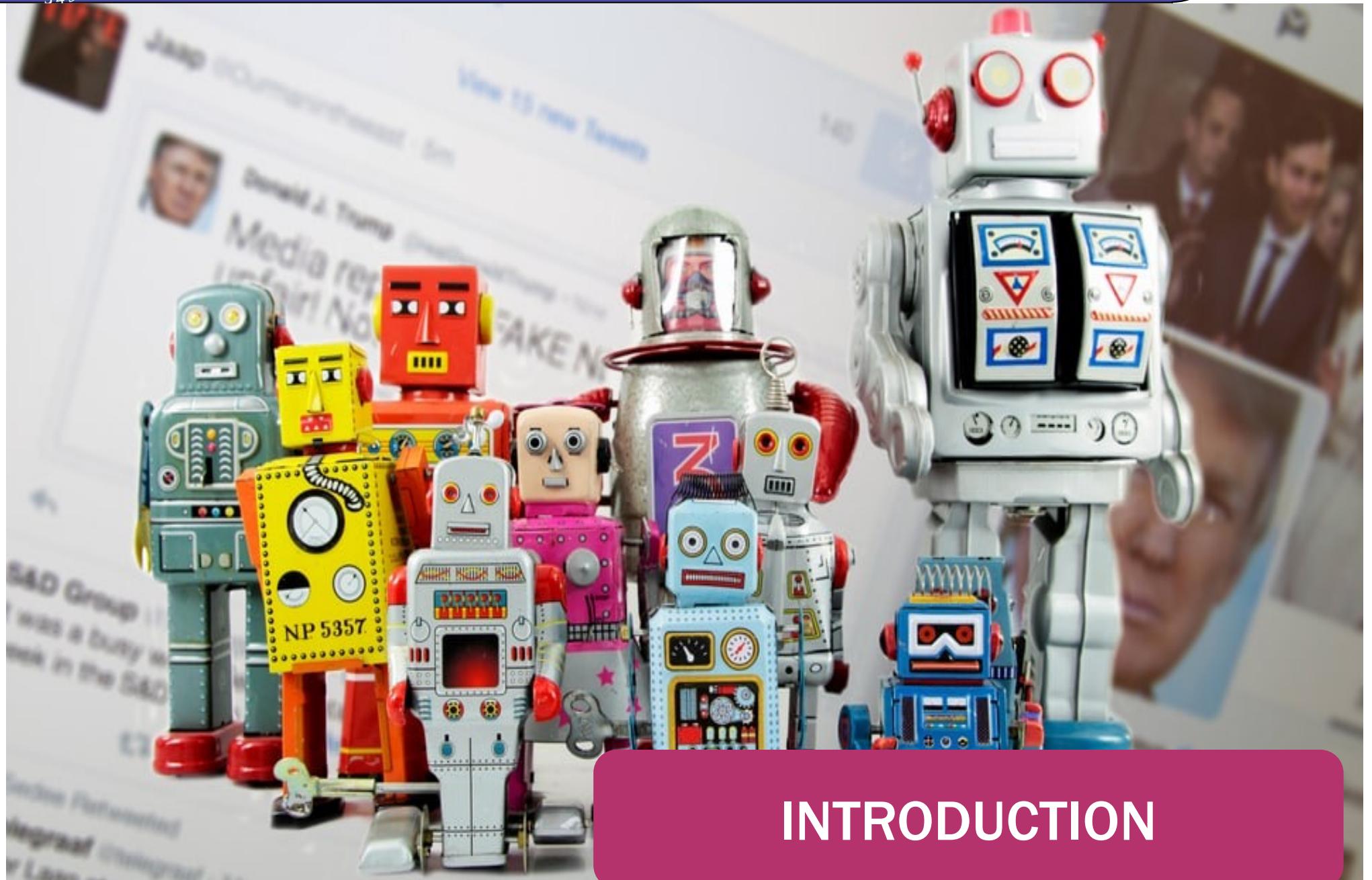


UNIVERSITÀ DI PISA



ISTITUTO
DI INFORMATICA
E TELEMATICA

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



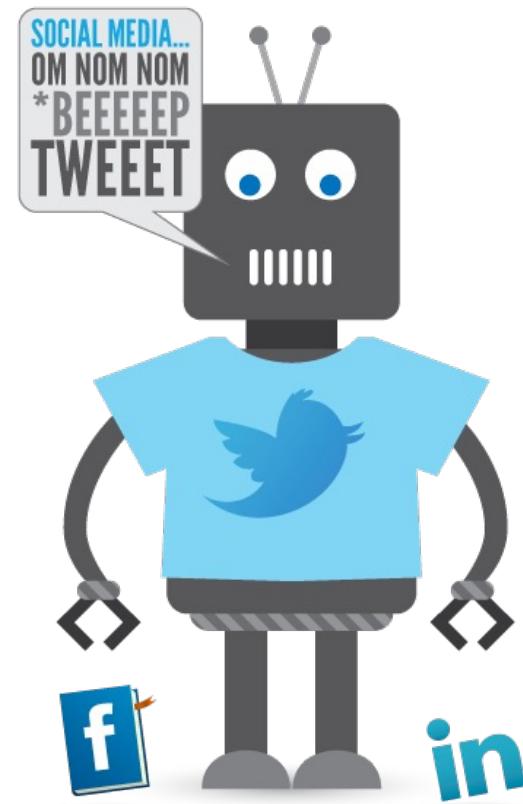


What Is a Social Bot?

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. Bowman¹, Kim Holmberg², Andrew Tsou³, Cassidy R. Sugimoto³
& Vincent Larivière⁴

scientific impact



Malicious and Benign Bots

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

- Support and coordination in emergency situations



The screenshot shows the Twitter profile of USGSted (@USGSted). The profile picture is a circular logo for "USGS TED" featuring a globe and the letters "USGS" and "TED". The bio reads: "Official U.S. Geological Survey earthquake alerts. For other official accounts, and to engage with us on other channels see [usgs.gov/socialmedia](#)". The stats are: Tweets 3,082, Following 1, Followers 87.6K, Likes 5. The timeline shows two tweets:

- USGSted @USGSted · 6h Prelim M5.6 Earthquake southern Mid-Atlantic Ridge Jun-13 03:00 UTC, updates [go.usa.gov/xmJSF](#)
- USGSted @USGSted · Jun 10 Prelim M5.8 Earthquake Mariana Islands region Jun-10 17:14 UTC, updates [go.usa.gov/xmuqW](#)



The screenshot shows the Twitter profile of Social Sensing (@sos_sensing). The profile picture is a logo for "SOS Social Sensing" with a globe and signal waves. The bio reads: "This bot belongs to Social Sensing, an IIT CNR scientific project. It leverages the crowd-sensing paradigm to improve situational awareness during emergencies." The stats are: Tweets 5,577, Following 284, Followers 78, Likes 228. The timeline shows two tweets:

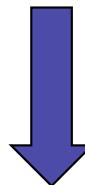
- Social Sensing @sos_sensing · 9h Replying to @ahmedalhassaniy Hi @ahmedalhassaniy, a M4.1 earthquake occurred at 2019-06-13 01:21UTC (21m:56s ago) near Paveh, Iran. Did you feel it? Please answer with one of these Yes
No
- Social Sensing @sos_sensing · Jun 12 Replying to @9o8m3p Hi @9o8m3p, we are aware that Batgram, Pakistan was struck by a Magnitude 5.2 earthquake at 2019-06-12 05:10 UTC (10m:27s ago). Did you sense it? Please reply with "Yes" or "No".



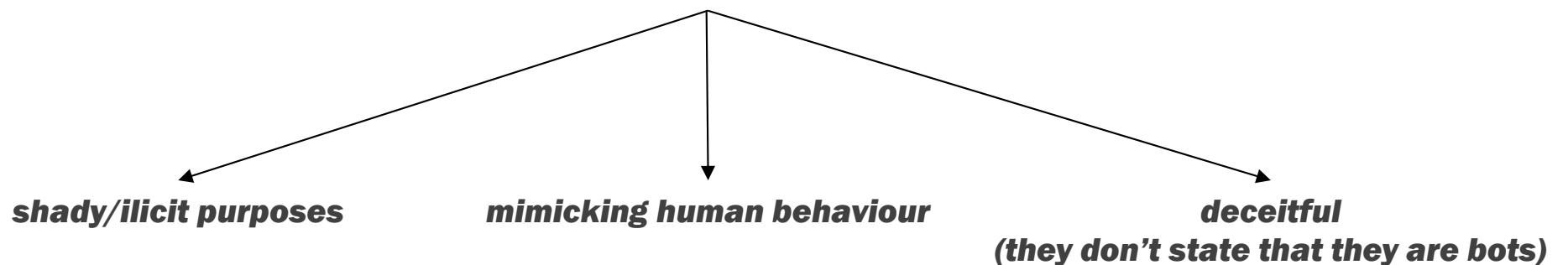
Malicious and Benign Bots

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

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

The Rise of Social Bots

review

2019, Vol. 37(1) 38-54
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/089439317734157
journals.sagepub.com/home/ssc

SAGE

The Brexit Botnet and User-Generated Hyperpartisan News

Marco T. Bastos¹ and Dan Mercea¹

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

Massive networks of fake accounts found on Twitter

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

24 DIREZIONE FINANZA & MERCATI | Così i tweet dei robot insidiano le news

INCHIESTA

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

—di Vittorio Carlini — 23 marzo 2018

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!**

ARTICLE
DOI: 10.1038/s41467-018-06930-7 OPEN

The spread of low-credibility content by social bots

Chengcheng Shao^{1,2}, Giovanni Luca Ciampaglia³, Onur Varol¹, Kai-Cheng Yang¹, Alessandro Flammini^{1,3} & Filippo Menczer^{1,3}

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!**

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}

^aCenter for Information and Communication Technology, Fondazione Bruno Kessler, 38123 Trento, Italy; and ^bUSC 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 sub-groups 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

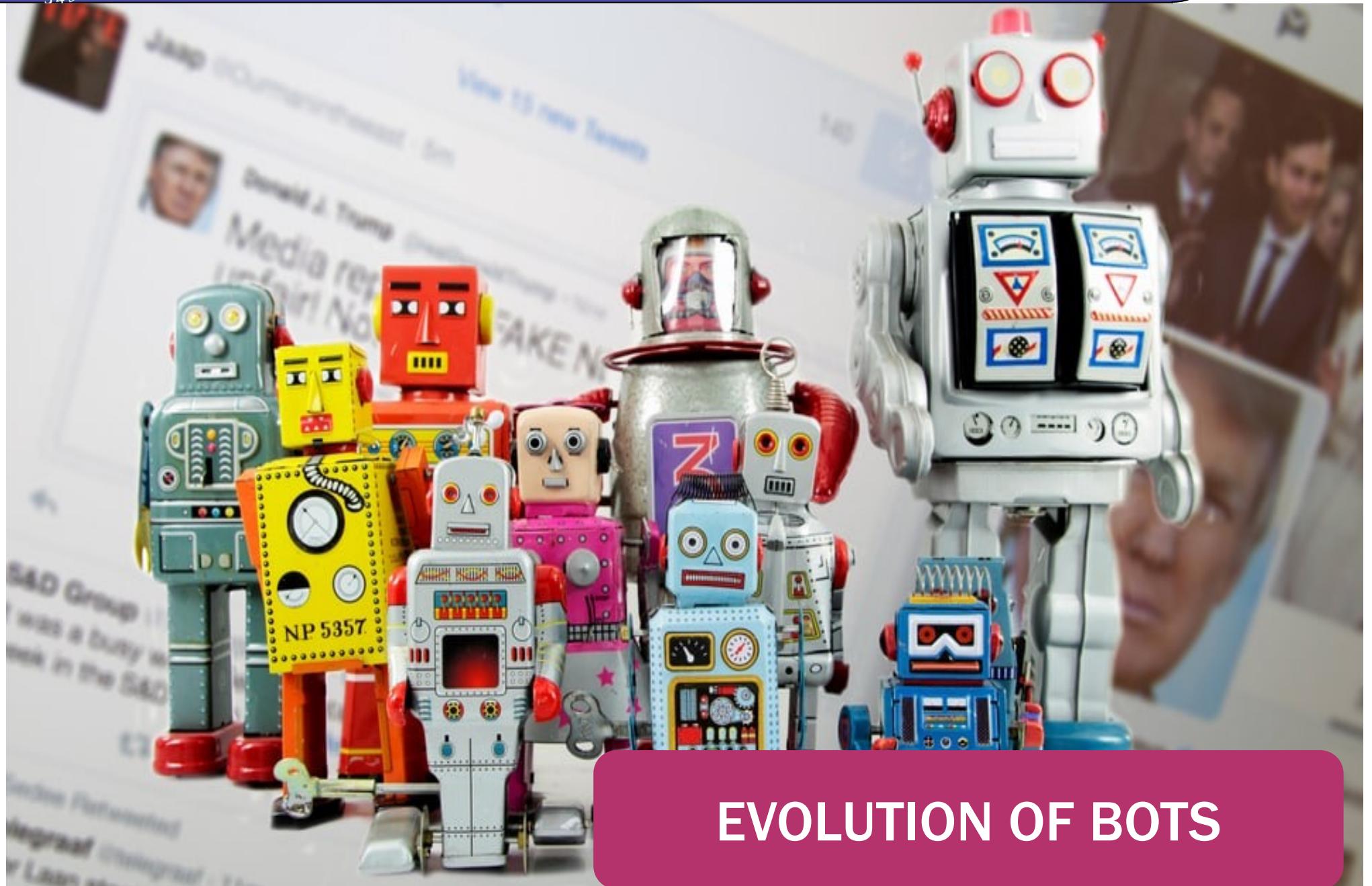
The spread of true and false news online

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...

A screenshot of a Twitter profile page. The user has a large, dark gray placeholder profile picture. At the top, it shows the user has 4 tweets, 155 accounts they follow, 7 followers, and 1 like. Below this, the 'Tweets' section is active, showing two tweets from February 2016. The first tweet links to a Google Plus post. The second tweet is from 'Davide Demitri'. On the left sidebar, it says the user joined in August 2015, has 2 photos and videos, and buttons for 'Tweet to' and 'Message'.

A screenshot of a Twitter profile page. The user has a large, dark gray placeholder profile picture. At the top, it shows the user follows 59 accounts and has 4 followers. Below this, the 'Tweets' section is inactive, displaying the message 'hasn't Tweeted'. On the left sidebar, it says the user joined in February 2015, has 2 photos and videos, and buttons for 'Tweet to' and 'Message'.



Evolution of Bots

Some time ago...

TMJ-CHN Jobs
@tmj_chn_jobs
Follow this account for geo-targeted Other job tweets in China. Need help? Tweet us at @CareerArc!
④ China
🔗 careerarc.com/job-seeker
📅 Joined May 2009

Tweets Tweets & replies

TMJ-CHN Jobs @tmj_chn_jobs · Feb 22 Interested in a #job in #Guangzhou, Guangdong? This could be a great fit: bit.ly/2qNU4Rh #DellJobs #Sales #Hiring #CareerArc

TMJ-CHN Jobs @tmj_chn_jobs · Feb 22 If you're looking for work in #China, check out this #job: bit.ly/2FtCIGz #DellJobs #Sales #Hiring #CareerArc

TMJ-CHN Jobs @tmj_chn_jobs · Feb 22 See our latest #Guangzhou, Guangdong #job and click to apply: Account Executive - bit.ly/2FkZS8g #DellJobs #BusinessMgmt #Hiring #CareerArc

Tweet to TMJ-CHN Jobs

TMJ - SFO Util Jobs
@tmj_sfo_util
Follow this account for geo-targeted Utilities job tweets in San Francisco, CA. Need help? Tweet us at @CareerArc!
④ San Francisco, CA
🔗 careerarc.com/job-seeker
📅 Joined February 2009

Tweets Tweets & replies

TMJ - SFO Util Jobs @tmj_sfo_util · 13h This #job might be a great fit for you: Maintenance Technician II - 2nd shift - bit.ly/2orHIA3 #cintasjobs #Utilities #Pittsburg, CA #Hiring #CareerArc

TMJ - SFO Util Jobs @tmj_sfo_util · 14h We're #hiring! Read about our latest #job opening here: Service Sales Representative - First Aid and Safety - bit.ly/2CeyhHn #cintasjobs #Utilities #Napa, CA #CareerArc

TMJ - SFO Util Jobs @tmj_sfo_util · 15h Join the Cintas Corporation team! See our latest #job opening here: bit.ly/2otFyQ #cintasjobs #Utilities #SouthSanFrancisco, CA #Hiring #CareerArc



Botnets

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

TWEETS FOLLOWING FOLLOWERS LIKES

2,311 9,254 8,572 117

Tweets Tweets & replies Photos & videos

Oggi e per sempre, #metticilafaccia. NO alla mafia

Retweeted

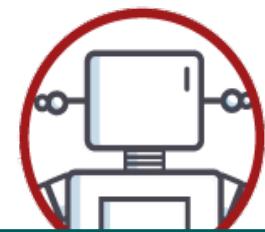
Rai1 @RaiUno · 3 Nov 2015

Aspettando l'ultima puntata di #SottoCopertura, #metticilafaccia su Twitter. → ow.ly/TRHa8 #Rai1 #Rai

Rai 1

Contro la criminalità organizzata,
twittate ora il vostro selfie
con l'hashtag
#METTICILAFACCIA.

bot or real?





Novel Social Bots



TWEETS 33.3K FOLLOWING 3,371 FOLLOWERS 4,039 LIKES 17K LISTS 4

Logopedista. Amo il mare ma sono ipertiroidea.. ma chi se ne frega. Ci vado ugualmente.

Reggio Emilia
Joined May 2013

Tweet to [redacted]

3 Followers you know



1,912 Photos and videos



Tweets [Tweets & replies](#) [Media](#)

· 14m
Italiani in fuga all'estero: nel 2015 espatriati in 107mila. Molti giovani [quotidiano.net/cronaca/italia...](#) via @quotidianonet



Italiani in fuga all'estero: nel 2015 espatriati in 107mila. Molti giovani ...
La meta preferita è la Germania. Lombardia e Veneto le principali regioni di emigrazione. Mattarella: "Cercare soluzioni"
[quotidiano.net](#)

Retweeted · 2h
Il virus del "prendo una immagine erotica/poetica/suggestiva dal web e ci appiccico un aforisma copiato" l'avete contratto su Facebook?

bot or real?



REAL



Novel Social Bots

TWEETS 884 FOLLOWING 686 FOLLOWERS 451 LIKES 5,866

Tweets **Tweets & replies** **Media**

Pinned Tweet · Oct 12
went to weis and picked up some croissants and this cute worker boy lit said to my face "i hope we croissant paths again" boy tf

2 27 ...

Retweeted · Dec 5
When your girl wants to play fight and you're not in the mood

bot or real?



REAL



Novel Social Bots

The profile picture is blurred. The bio reads: "Love the night life - KISS!". It includes links to mentalitch.com and was joined in June 2010. The stats are: TWEETS 26.1K, FOLLOWING 3,707, FOLLOWERS 2,162, LIKES 2. The tweets section shows two tweets:

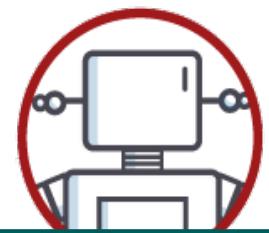
Tweets Tweets & replies Media

Our Favorite Female Super Heroes
bit.ly/1MZndN4 #Female #SuperHeroes

Our Favorite Female Super Heroes

Top Mexico Beach
dld.bz/eegdV #Mexico #Beaches

bot or real?

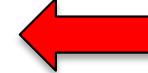


SPAMBOT

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, $k \sim 0$)

Human classification agreement (k)

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

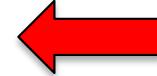
Cresci, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." *Proceedings of the 26th International Conference on World Wide Web Companion*. ACM, 2017.

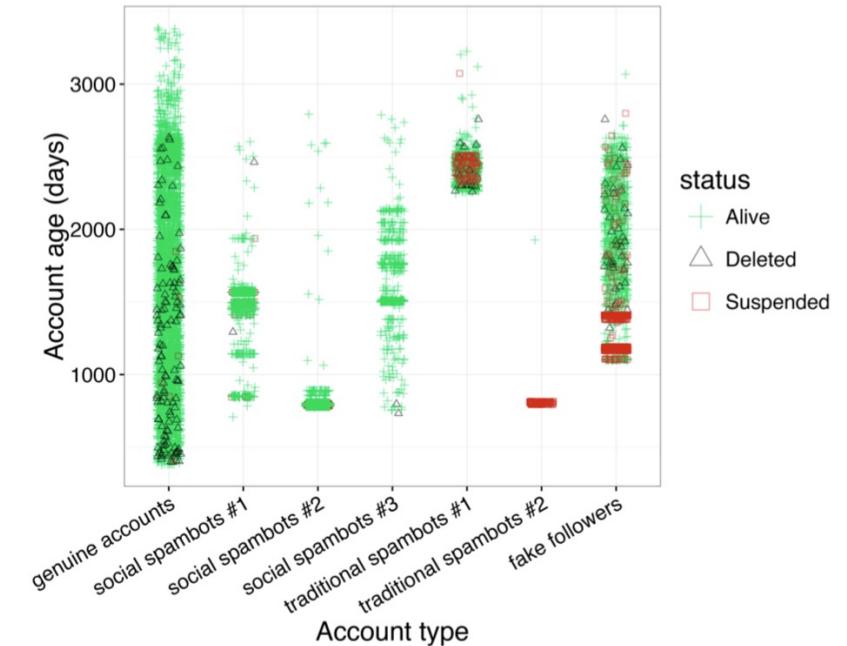


Twitter Defenses

- Twitter is not capable of effectively removing novel social bots

Twitter removal rate

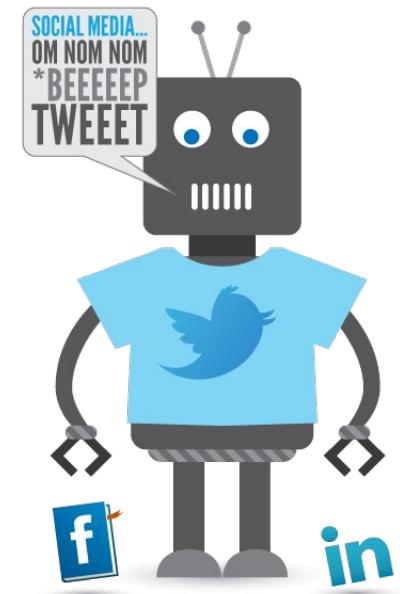
- old bots **60%**
- social bots **4%**  





Detection Techniques

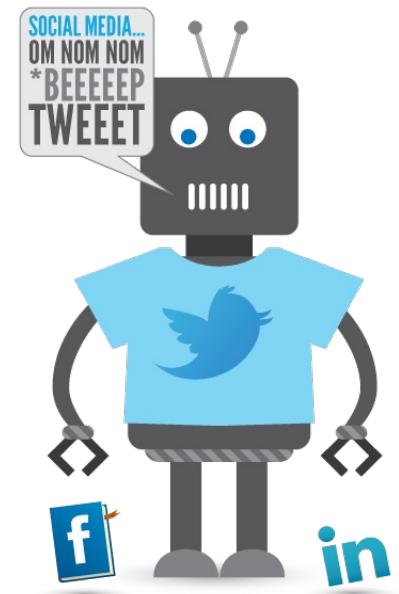
- **Naive detection:** all the accounts that interact in the same timestamp (second) are bots
- Helps to easily discover self-declared bots (not deceiving bots)
- A large number of novel bots remain undetected





Detection Techniques

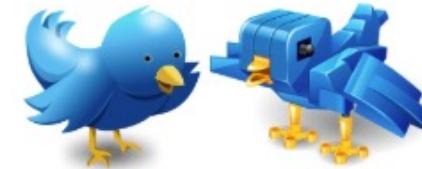
- The majority of detection techniques are **based on machine learning**, often **supervised** (e.g., *classification algorithms*)
- Each account is analyzed **singularly**
- Usage of **general-purpose ML algorithms** (e.g., *decision trees, random forest, SVM*)
- The focus is on **features** rather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





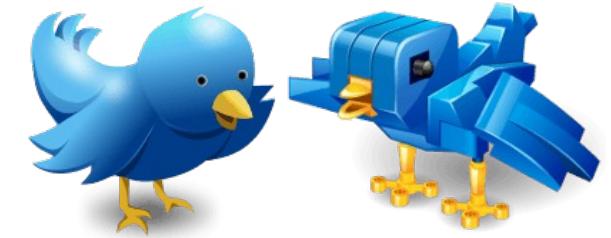
Botometer

Botometer[®]
An OSoMe project (bot•o•meter)



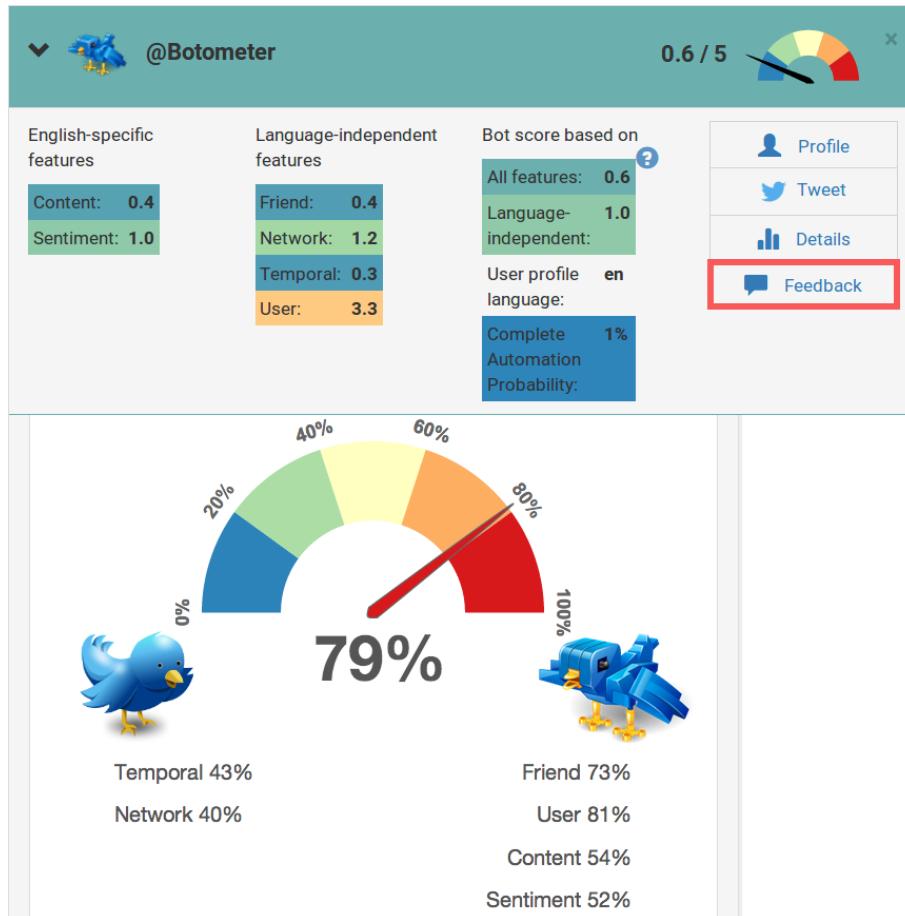
- Most widely used social bot classification system
 - Web application: https://botometer.iuni.iu.edu/#/!
 - REST API: <https://botometer.iuni.iu.edu/#/api>
- Ensemble of **supervised Random Forest** classifiers

- 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, ...

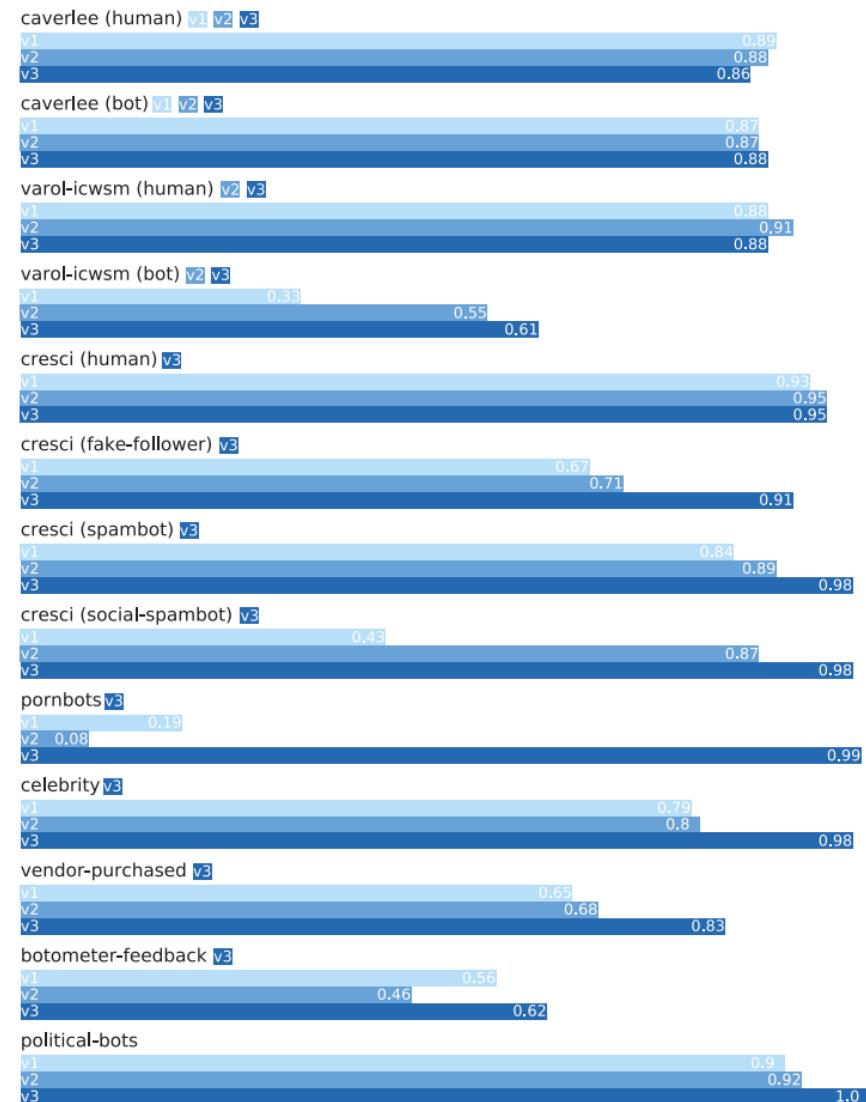


Botometer (results)

Botometer® An OSoMe project (bot•o•meter)



interface



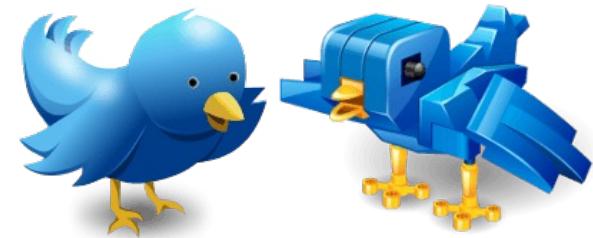
system evaluation (accuracy)

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%**



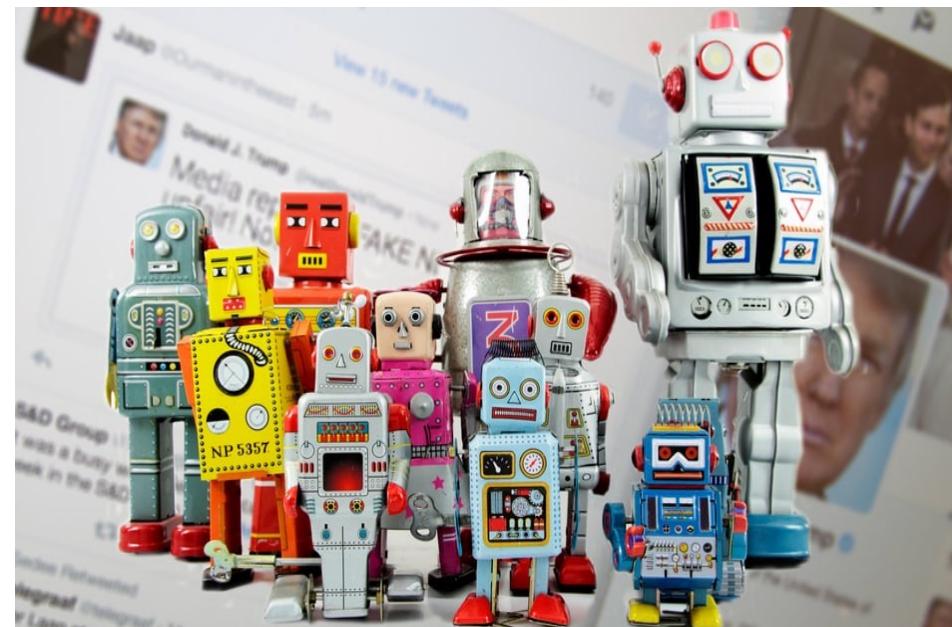
**Novel social bots could not be analyzed account-by-account
We need to analyze the collective behaviours of groups of users**



GROUP DETECTION

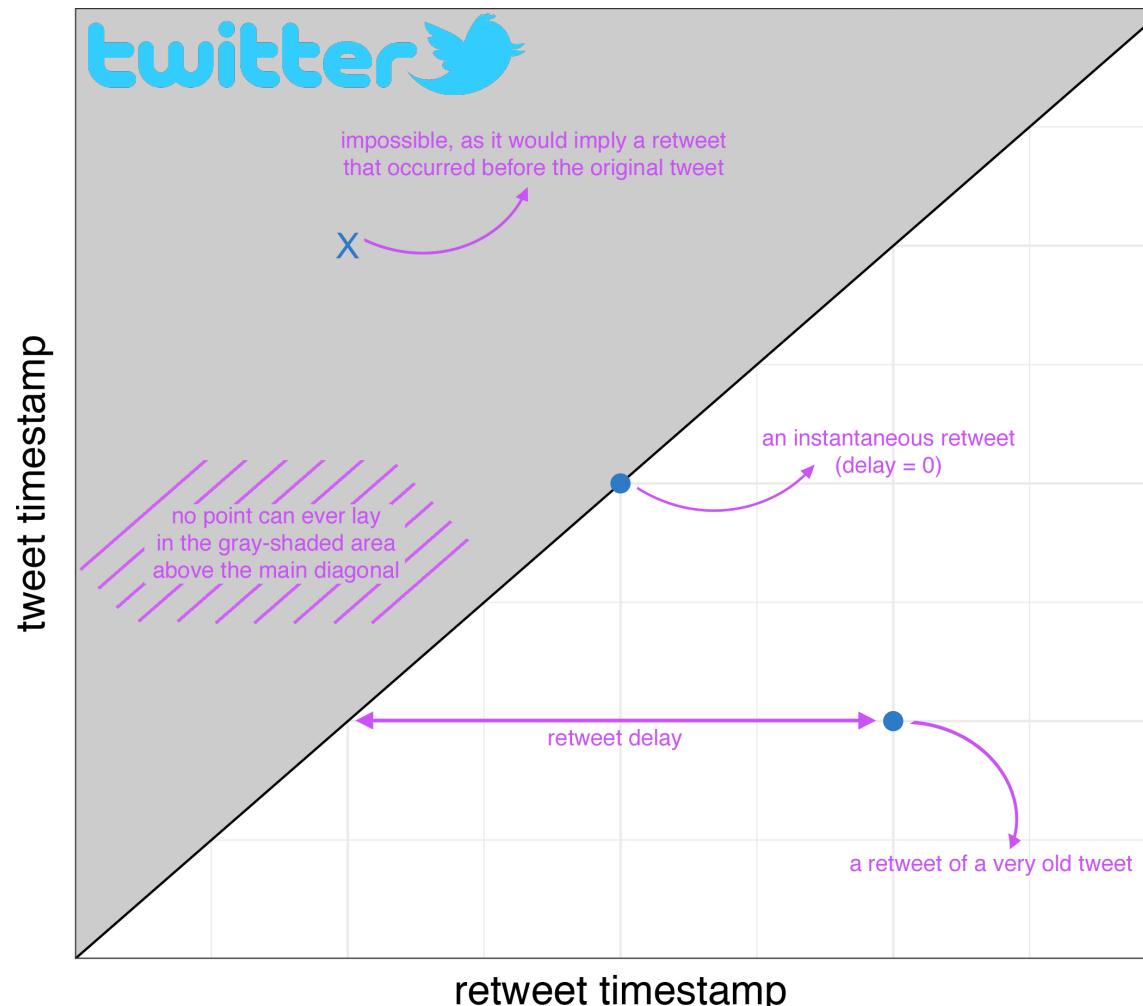
Group Analysis

- 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**



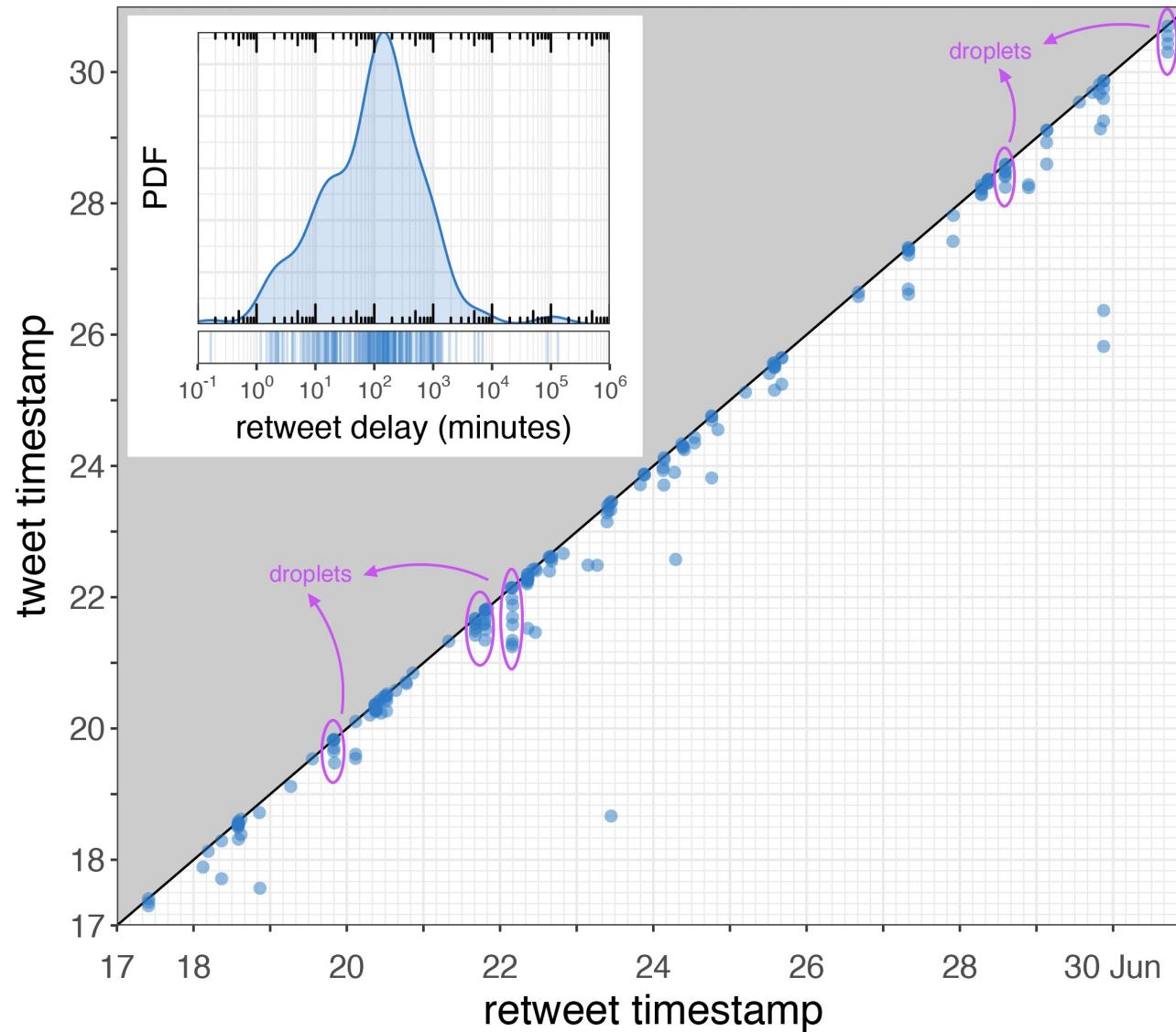
How can we recognize bots?

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



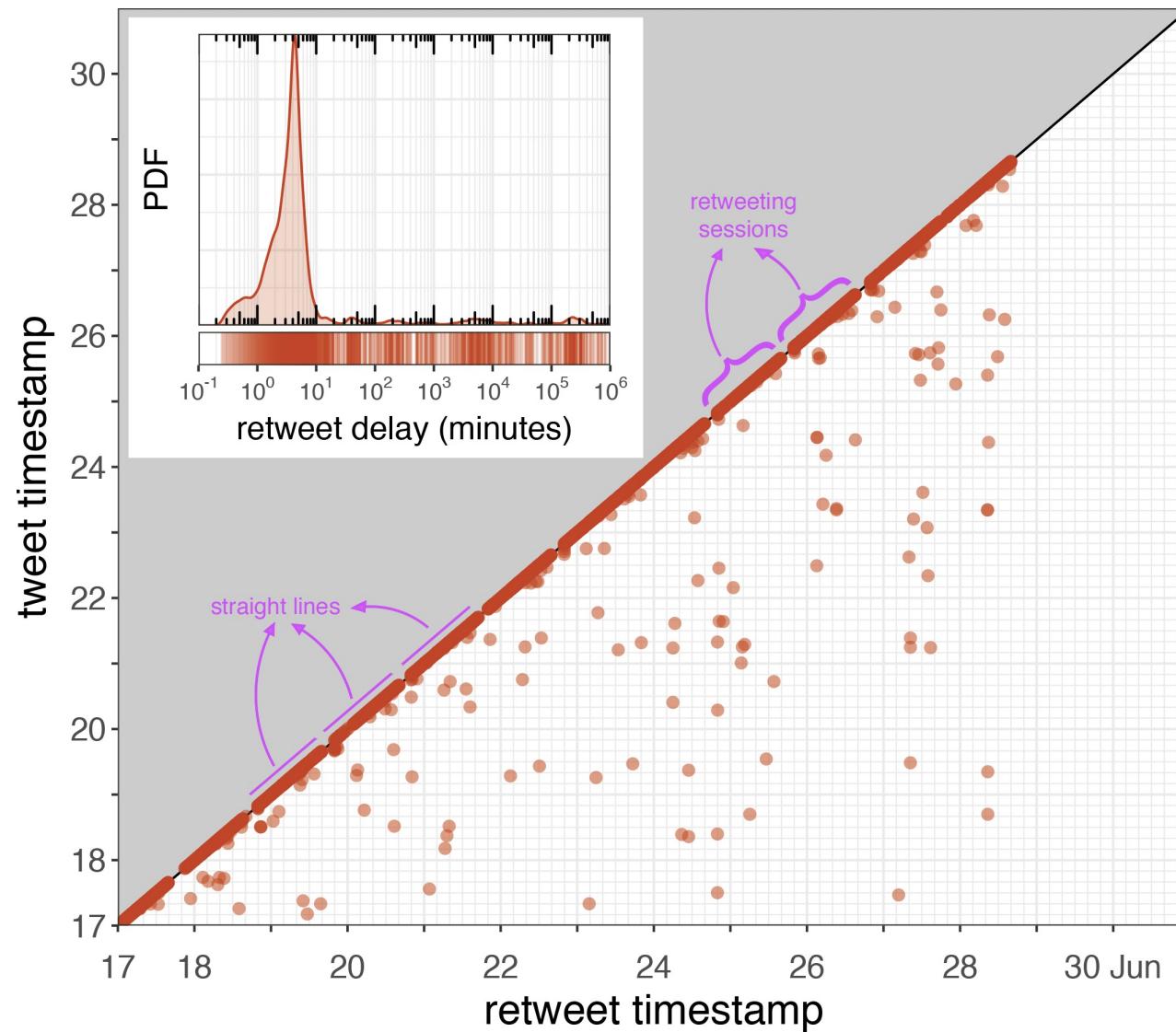
Temporal Behaviours

Normal Behaviour

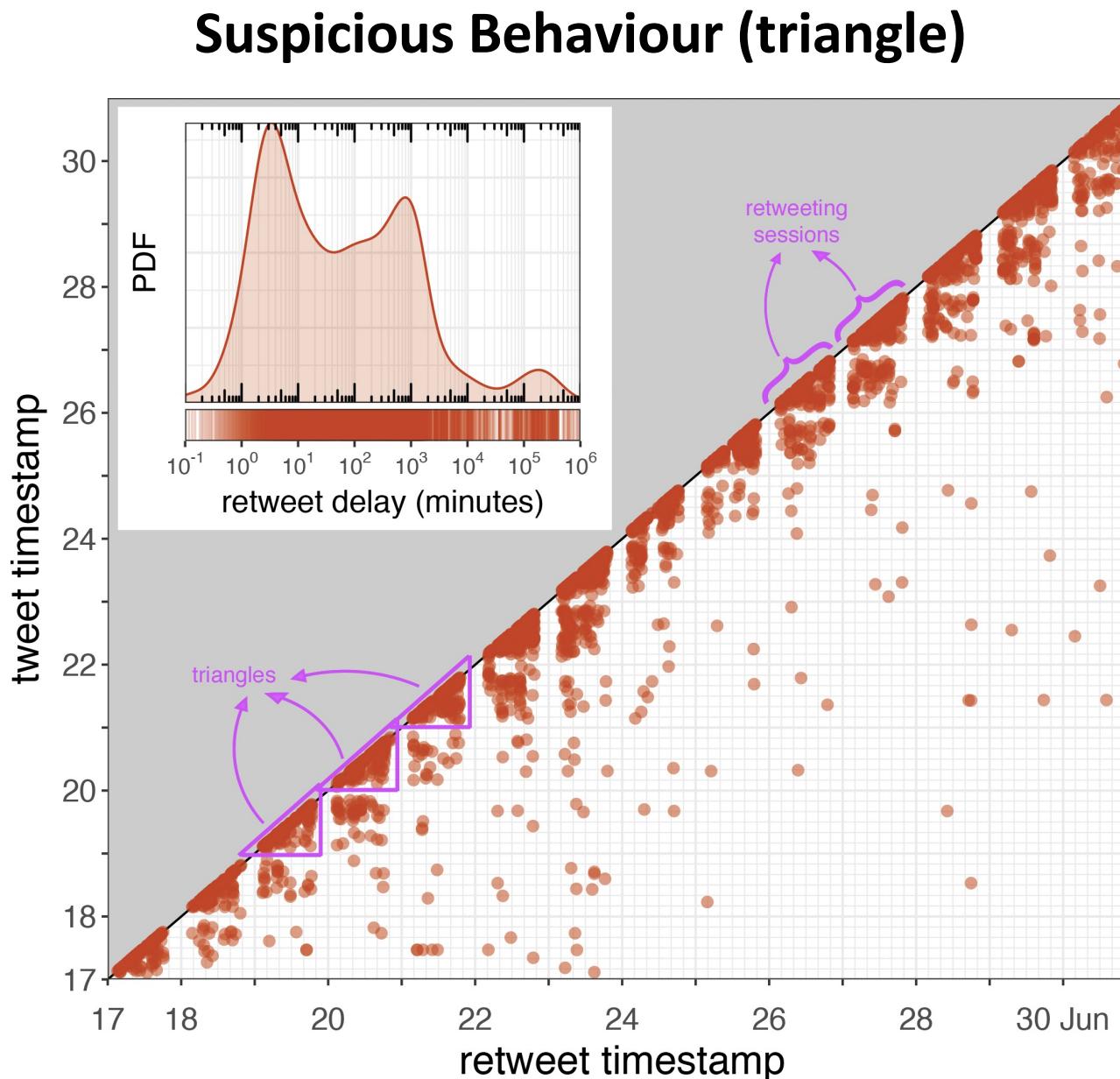


Temporal Behaviours

Suspicious Behaviour (straight line)

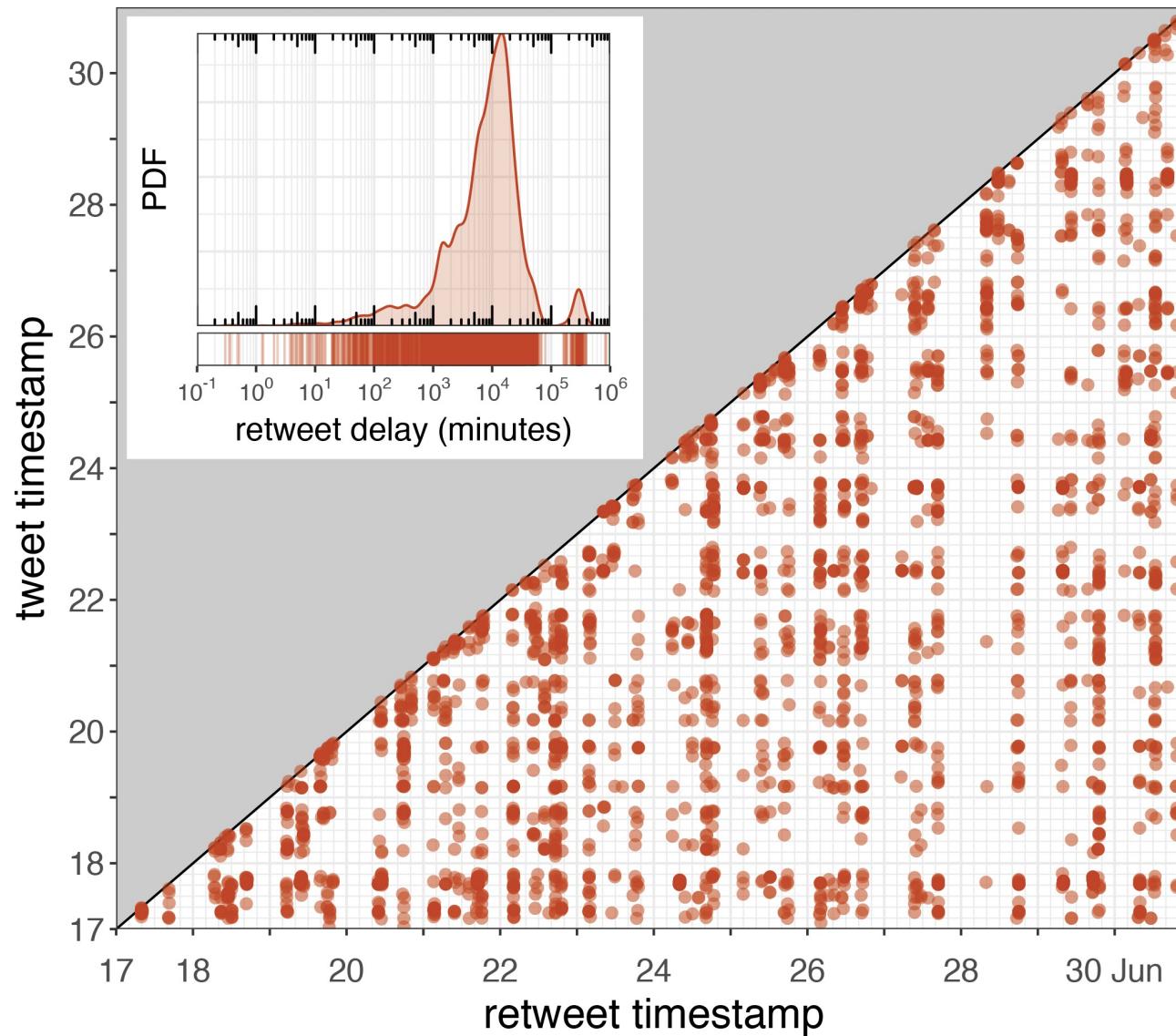


Temporal Behaviours



Temporal Behaviours

Suspicious Behaviour (waterfall)



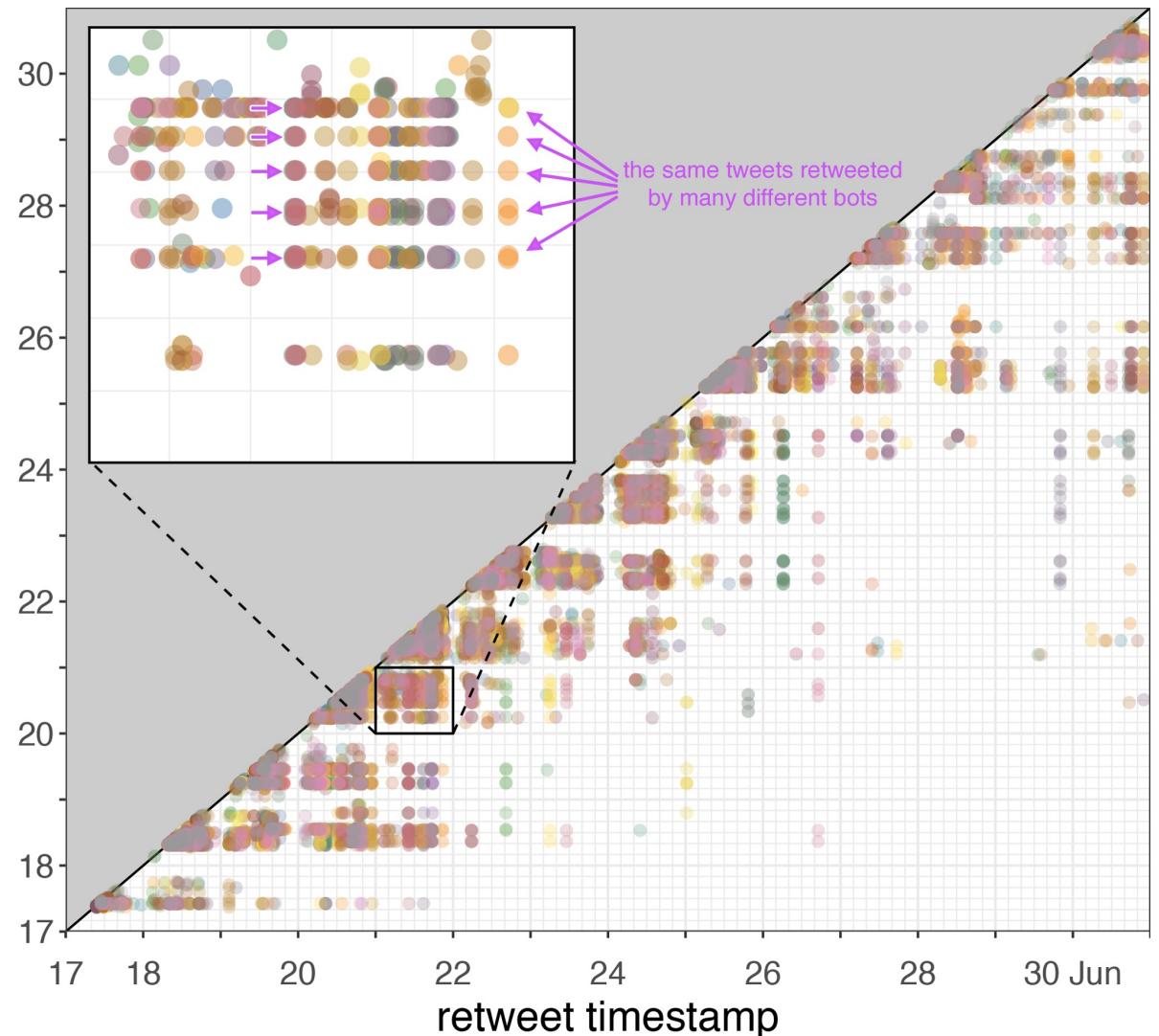


Real Examples

Real Examples

Singer botnet (Twitter)

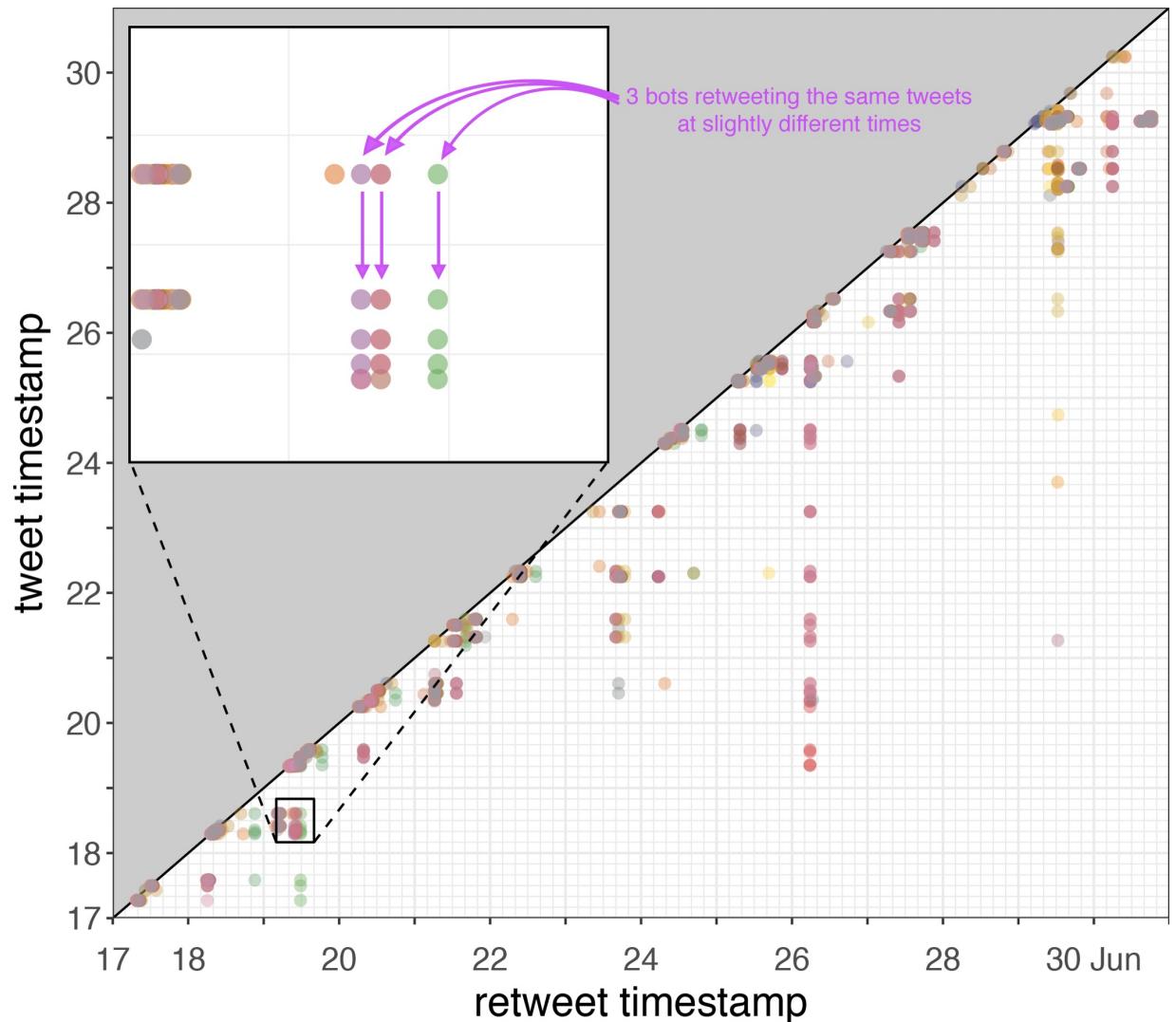
- almost 300 bots
- they only retweeted 2 accounts related to an italian pop singer:
 - @Valerio_Scanu
 - @ArmataScanu



Real Examples

Cars botnet (Twitter)

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





Action-based Behaviours

DIGITAL DNA

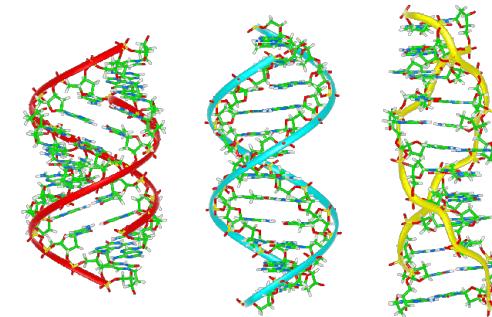
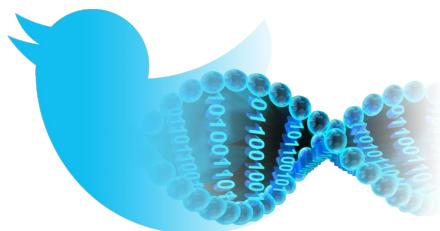
T ← tweet,
R ← retweet,
P ← reply

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



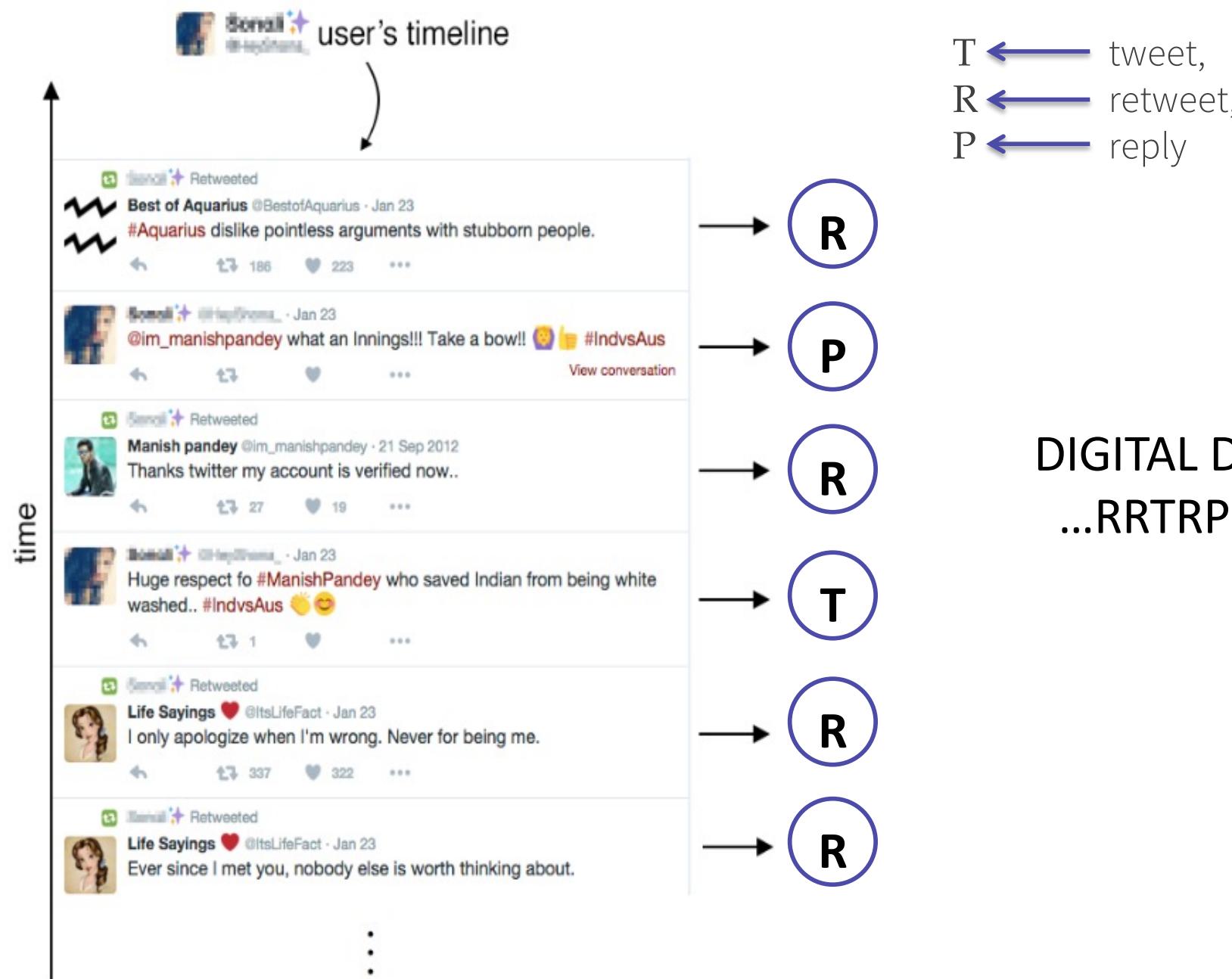
...RRTRPRTPRRPTPRPTPRRTRPR
...RPRTPTRPTRPRTPRRRRTPPRPP
...TTTRRRPPTPRPTPRTRPRTRRRTP
...PRTRPRTPPPRTPRRPTPPRRT
...TRTRPRTPRRPTPRPTPTPPRTT
...TRPPRTPPTRPPTPRRTTTPRPR

...AGTCTCCATTTCAGGTCTGA
...GTTAAAGATCGCCTCATCACC
...AGGCAATTGCCTGAACCTGG
...AGTCTCGATCCTTCCTCGTT
...AAAATCGAACGCCCTGTCGG
...ATTCTCCATCGCCTAAACAAC





Digital DNA





Digital DNA

How can we recognize bots?

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



...T**RRRPRRT**RRPRTPRPTPRRTRPR
...RPRTPTT**RRRPRRT**PRRRRRTPPRP
...TTTRRRP**RRRPRRT**RTRPTRRRTP
...PRTRPRTPPPRTPRR**RRRPRRT**R



RRRPRRT

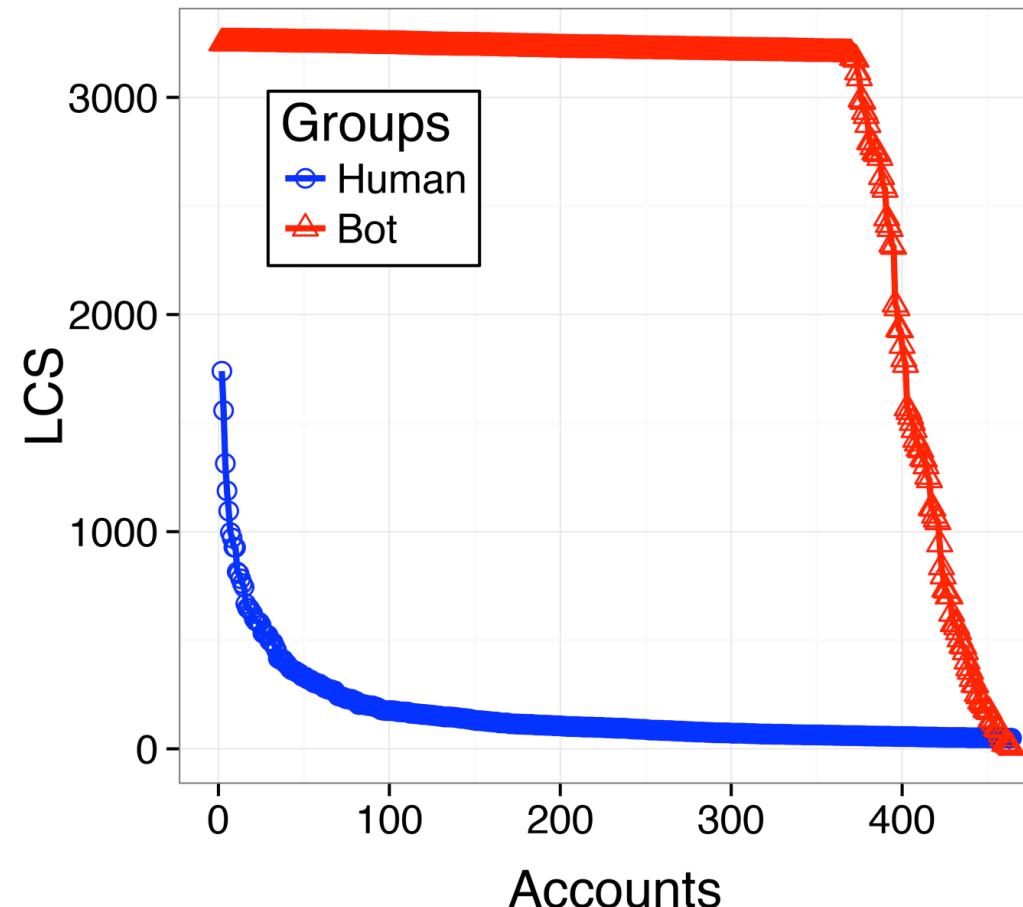
LCS: 7



Digital DNA

How can we recognize bots?

- *Intuition:* automated accounts have similar sequences

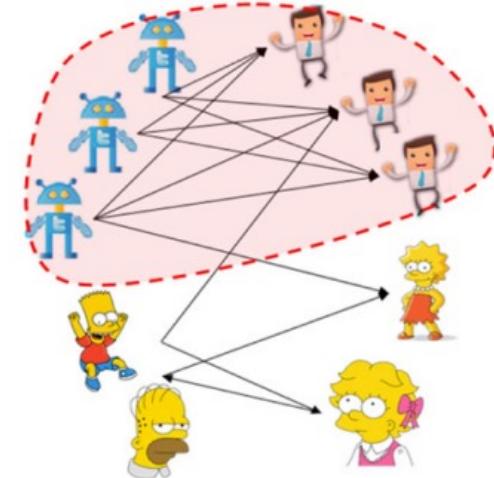


Network-based Behaviours

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

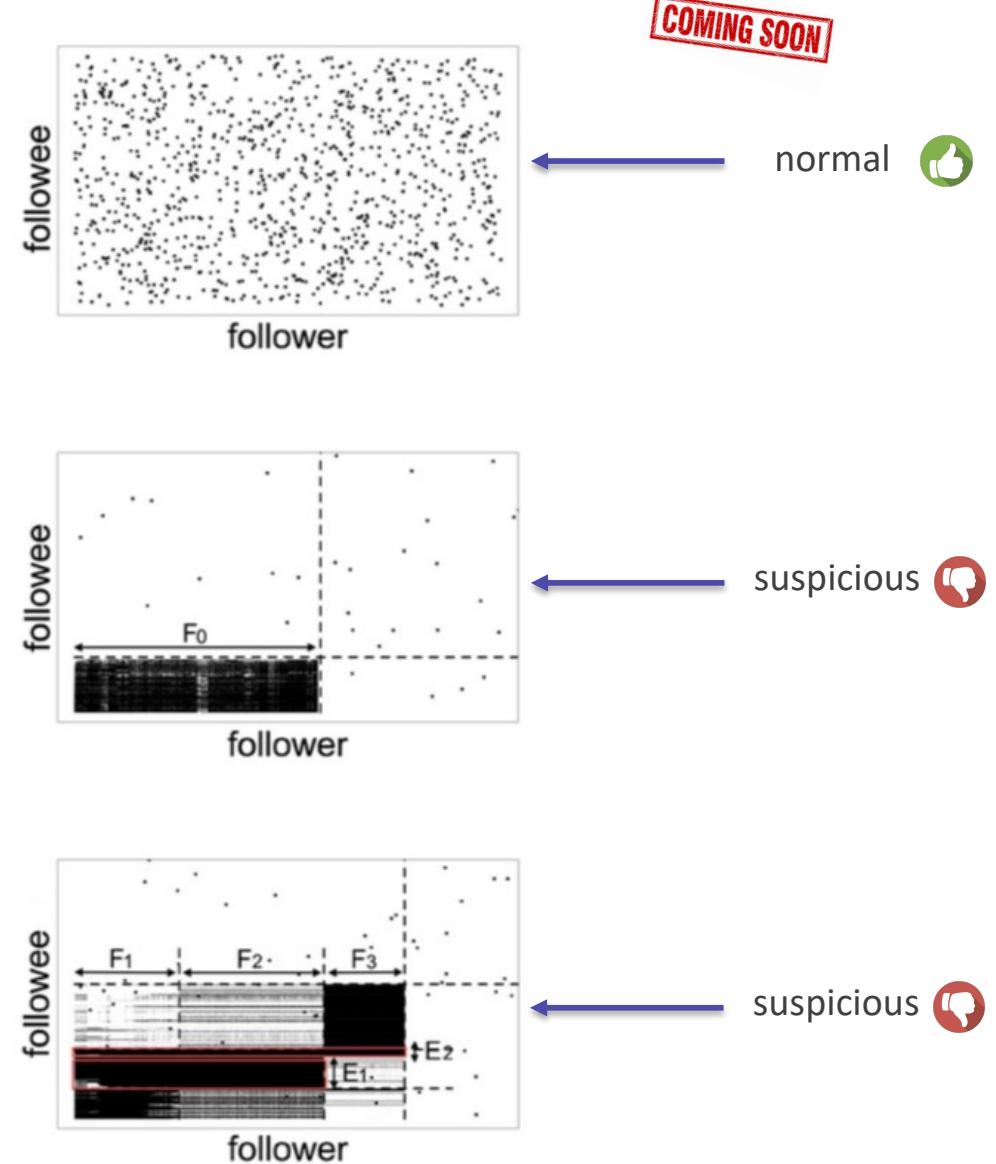
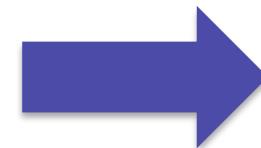
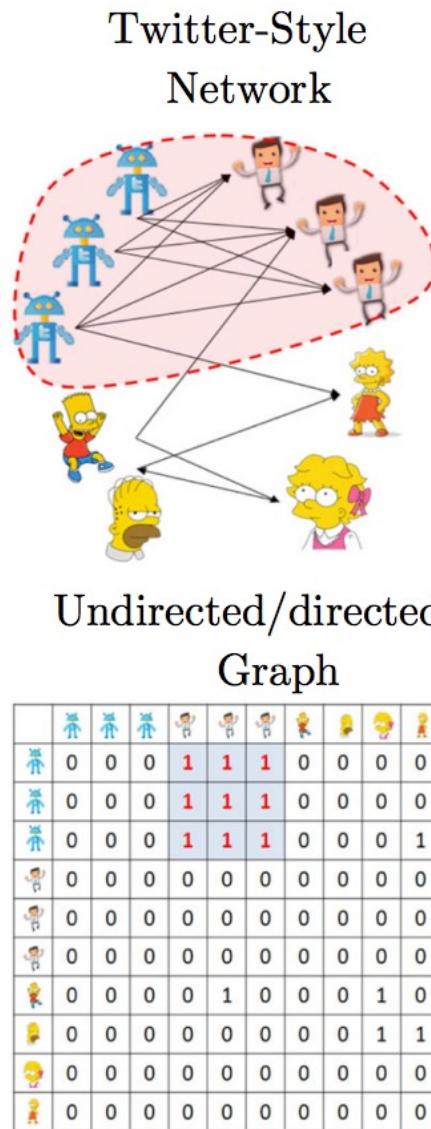


Undirected/directed Graph

	Robot 1	Robot 2	Robot 3	User 1	User 2	User 3	Simpson 1	Simpson 2	Simpson 3	Simpson 4
Robot 1	0	0	0	1	1	1	0	0	0	0
Robot 2	0	0	0	1	1	1	0	0	0	0
Robot 3	0	0	0	1	1	1	0	0	0	1
User 1	0	0	0	0	0	0	0	0	0	0
User 2	0	0	0	0	0	0	0	0	0	0
User 3	0	0	0	0	0	0	0	0	0	0
Simpson 1	0	0	0	0	1	0	0	0	1	0
Simpson 2	0	0	0	0	0	0	0	0	1	1
Simpson 3	0	0	0	0	0	0	0	0	0	0
Simpson 4	0	0	0	0	0	0	0	0	0	0

Network-based Behaviours

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

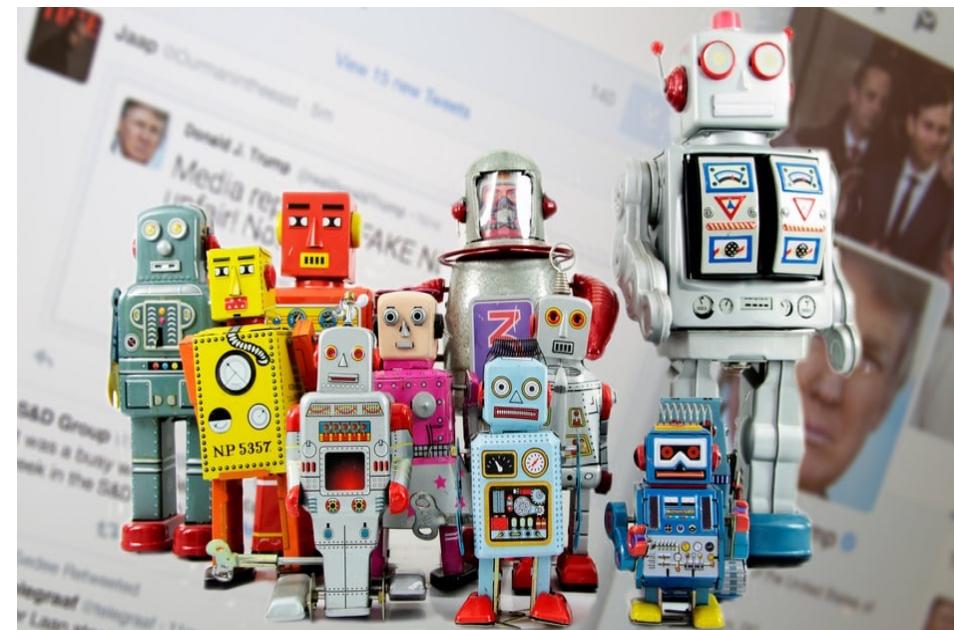




And Now?

Is the problem solved now? NO

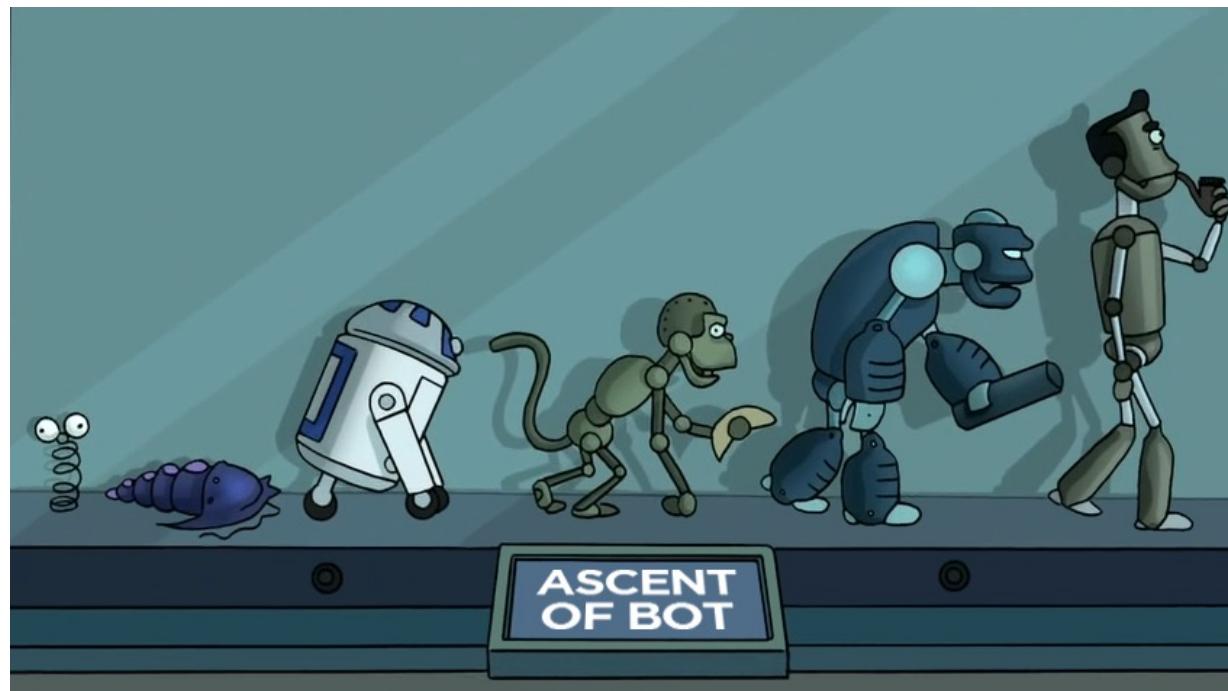
- No systematic comparison of **novel** vs **traditional** approaches
- Novel approaches produce **great improvements** in detection



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

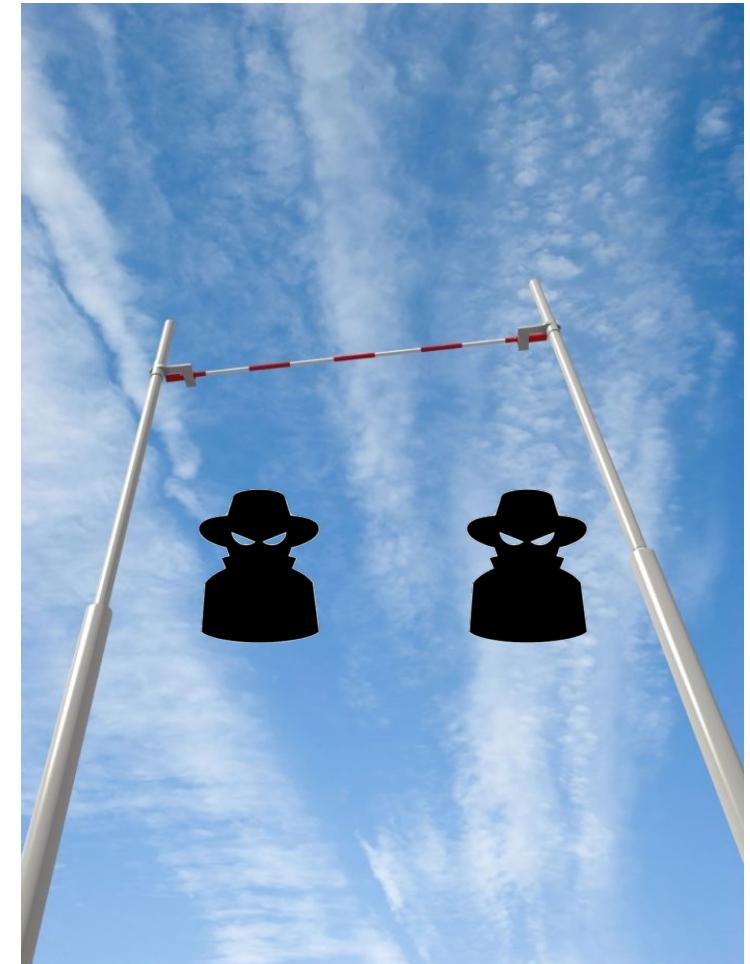




And Now?

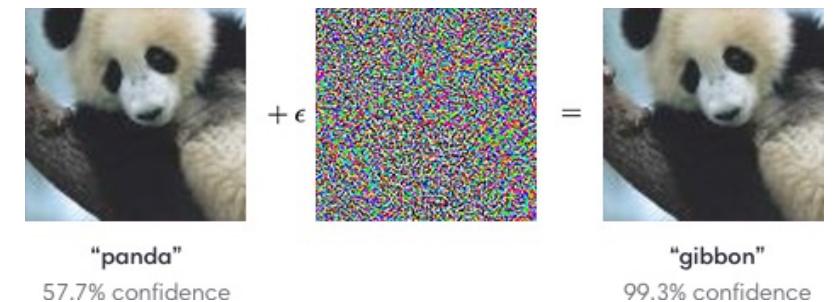
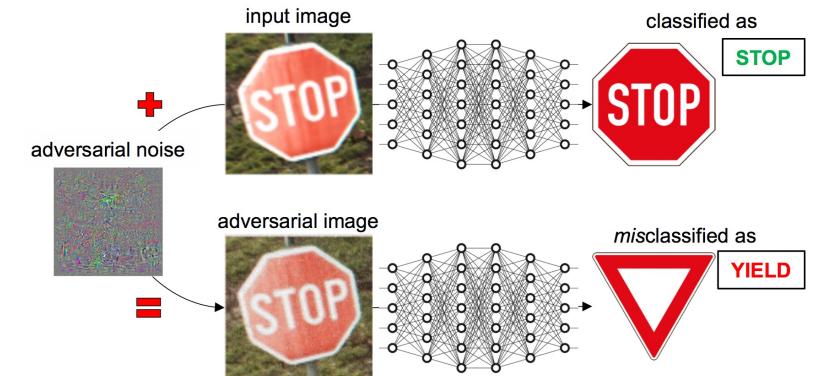
Is the problem solved now? NO

- Anyway, the bar is raised for bot developers



Advanced Modern Approaches

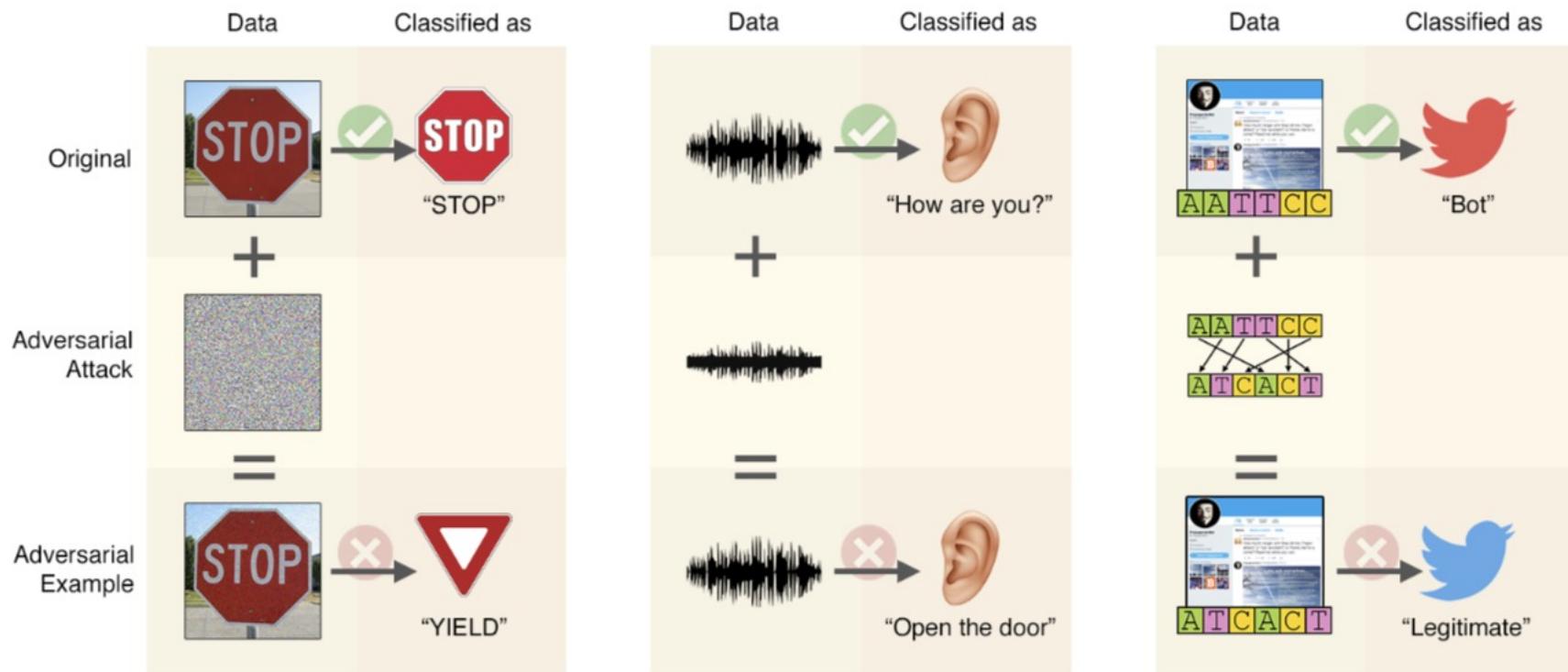
- Many machine learning techniques were originally designed for **stationary** and **benign** environments
- When these assumptions are violated techniques start making big mistakes (*e.g.*, *wrong classifications*)
- Bot detection is neither stationary (**bots evolve**) nor benign (bots try to **evade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





Advanced Modern Approaches

bot evolutions = adversarial attack to bot detectors



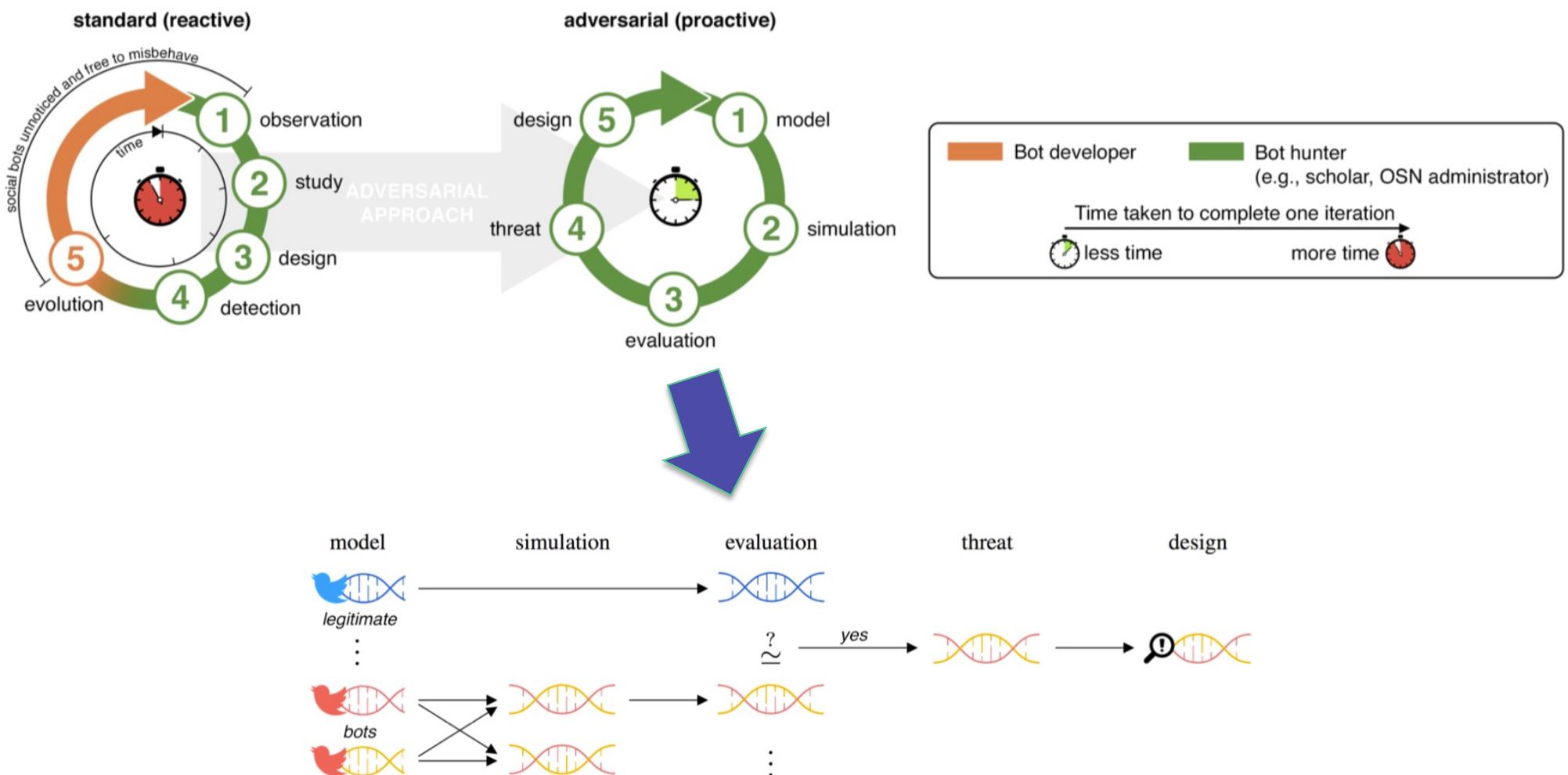
(a) Computer vision. Images can be modified by adding adversarial noise so as to fool image classification systems (e.g., those used by autonomous vehicles).

(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.

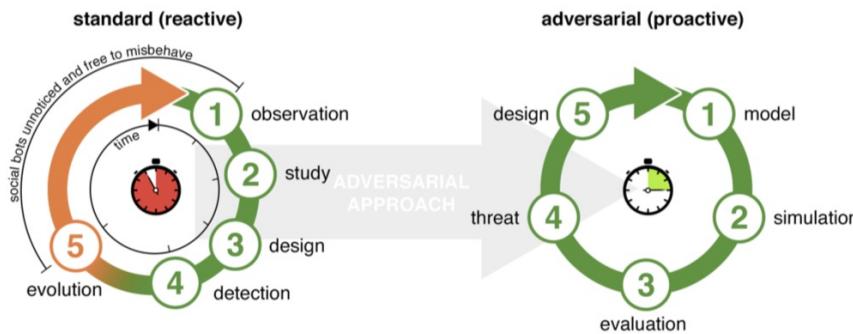
Adversarial Bot Detection

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...)

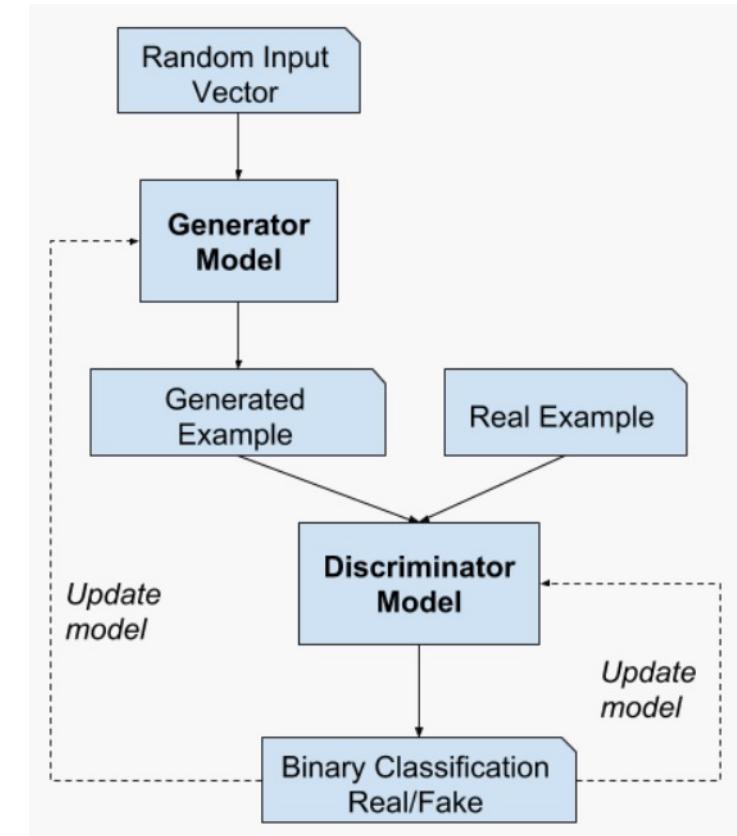


- *model* → digital DNA
- *simulation* → genetic algorithms (“evolved” bots)
- *evaluation* → bot detection technique (e.g., social fingerprinting)
- *threat* → “evolved” bot that evades detection
- *design* → new or modified technique that detects the “evolved” bot

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

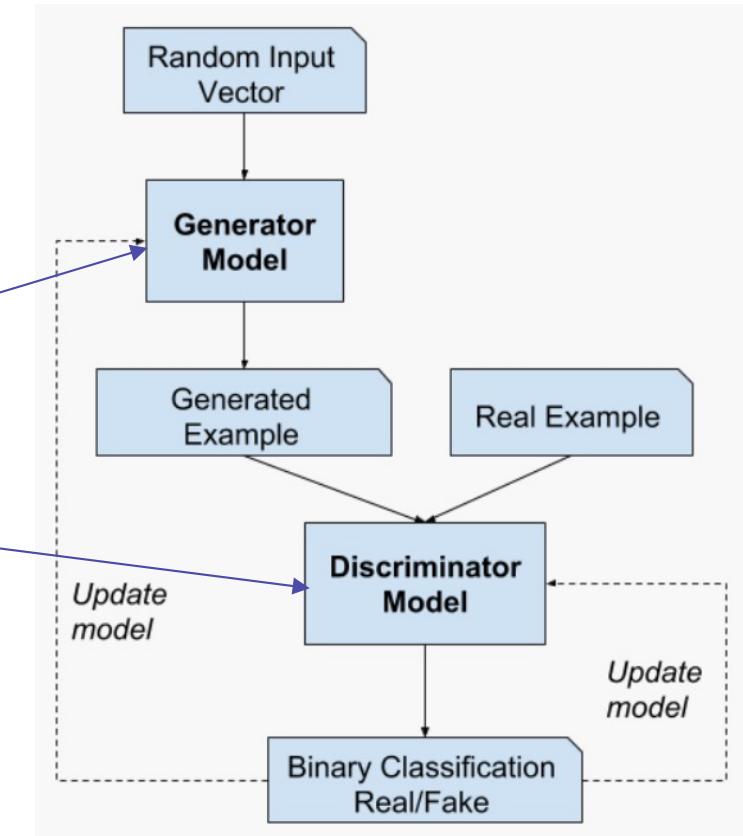
The proactive approach is based on the usage of **Generative Adversarial Nets (GANs)**

- a bot creator competes against a bot detector

*generates plausible
(yet fake) samples*

*distinguishes between
real and fake samples*

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



Adversarial Bot Detection

Anyway, a continue evolution in the techniques is required!

