Catch me, Yes we can! - Pwning Social Engineers using Natural Language Processing Techniques in Real-Time

Myeongsoo Kim[1], Changheon Song[1], Hyeji Kim[1], Deahyun Park[1], Yeeji Kwon[2], Eun Namkung[1], Ian G. Harris[3], Marcel Carlsson[4]

1. Kookmin University, 2. Seoul Women's University, 3. University of California Irvine, 4. Lootcore

Black Hat USA 2018

Who we are

- Ian G. Harris
  - Professor of Computer Science at the University of California Irvine
  - Research in HW Verification and Security
  - Applies Natural Language Processing techniques

Who we are

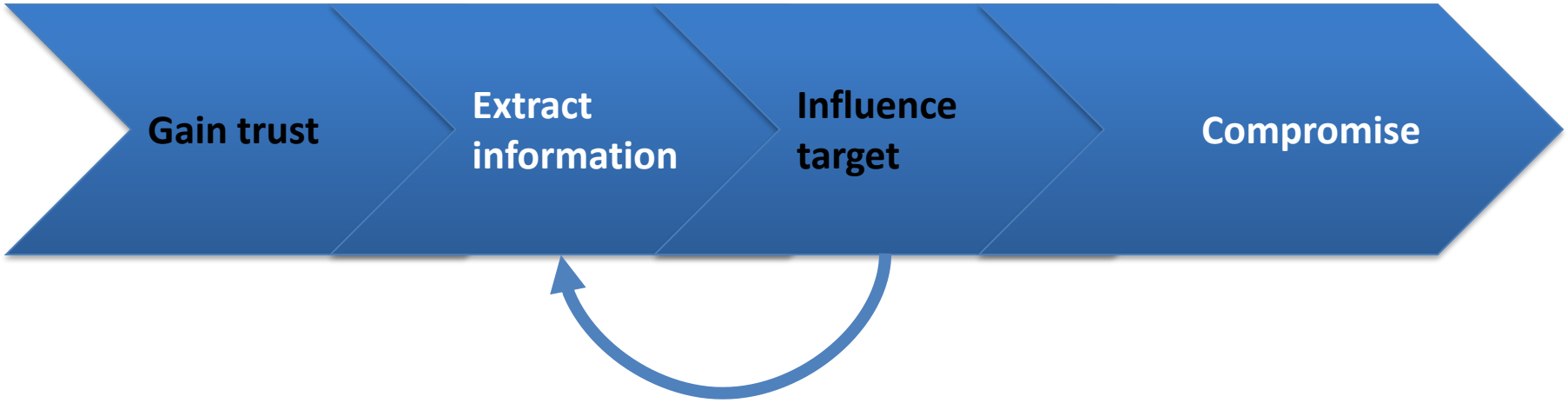
- Marcel Carlsson
  - Principal consultant at Lootcore
  - Red teaming, consulting and security research
  - Holistic scope across human/technology/ process/physical domains incl Social Engineering (meatware pwnage)

Social Engineering (SE) 101

"Any act that influences a person to take an action that may or may not be in their best interest"

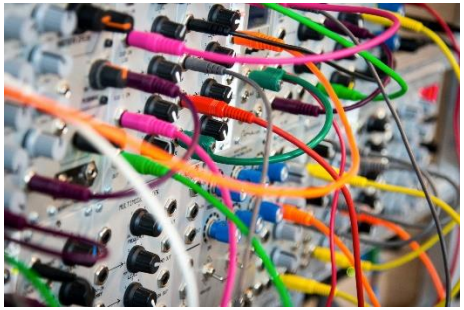
– social-engineer.com

Social Engineering (SE) 101



Social Engineering (SE) 101

Open Source Intelligence Gathering (OSINT)



Persons identified by facial recognition

Extracted meta data after indexing

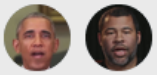
Video input file

fake_obama

Edit

Insights Transcript Search... English

People



Time line for person appearance

Barack Obama

44th President of the United States

Show biography Find in Bing

Appears for 96.34% of the video's duration.

Keywords

time trust

Labels

person man suit indoor necktie

Speech sentiment

Negative (1.87%)
Neutral (79.76%)
Positive (18.36%)

Extracted key words from audio transcript

Spoken word sentiment analysis

Video input file

Extracted meta data after indexing

Translation possible

Video OCR

fake_obama

Edit

Insights | Transcript | Search... | English

Edit Off | Autoscroll On | OCR On

We're entering an era in which our enemies can make it look like anyone is saying anything at any point in time even if they would never say those things so.

For instance they could have me say things like I'm know.

Killmonger was right. Bro Ben Carson is

in the sunken place or about this simply president trump is a total and complete d******.

Now. You see I will never say these things.

At least not in a

Speaker #1

Keyframe 2 Image

public address but someone else would someone.

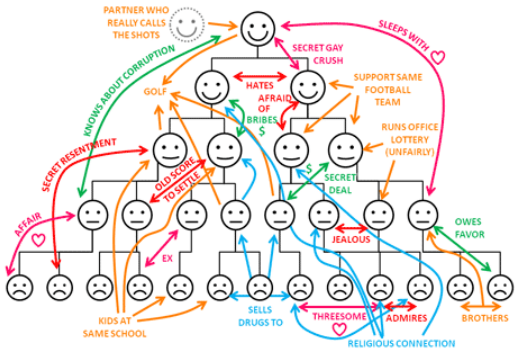
Like Jordan peele. This is a dangerous time.

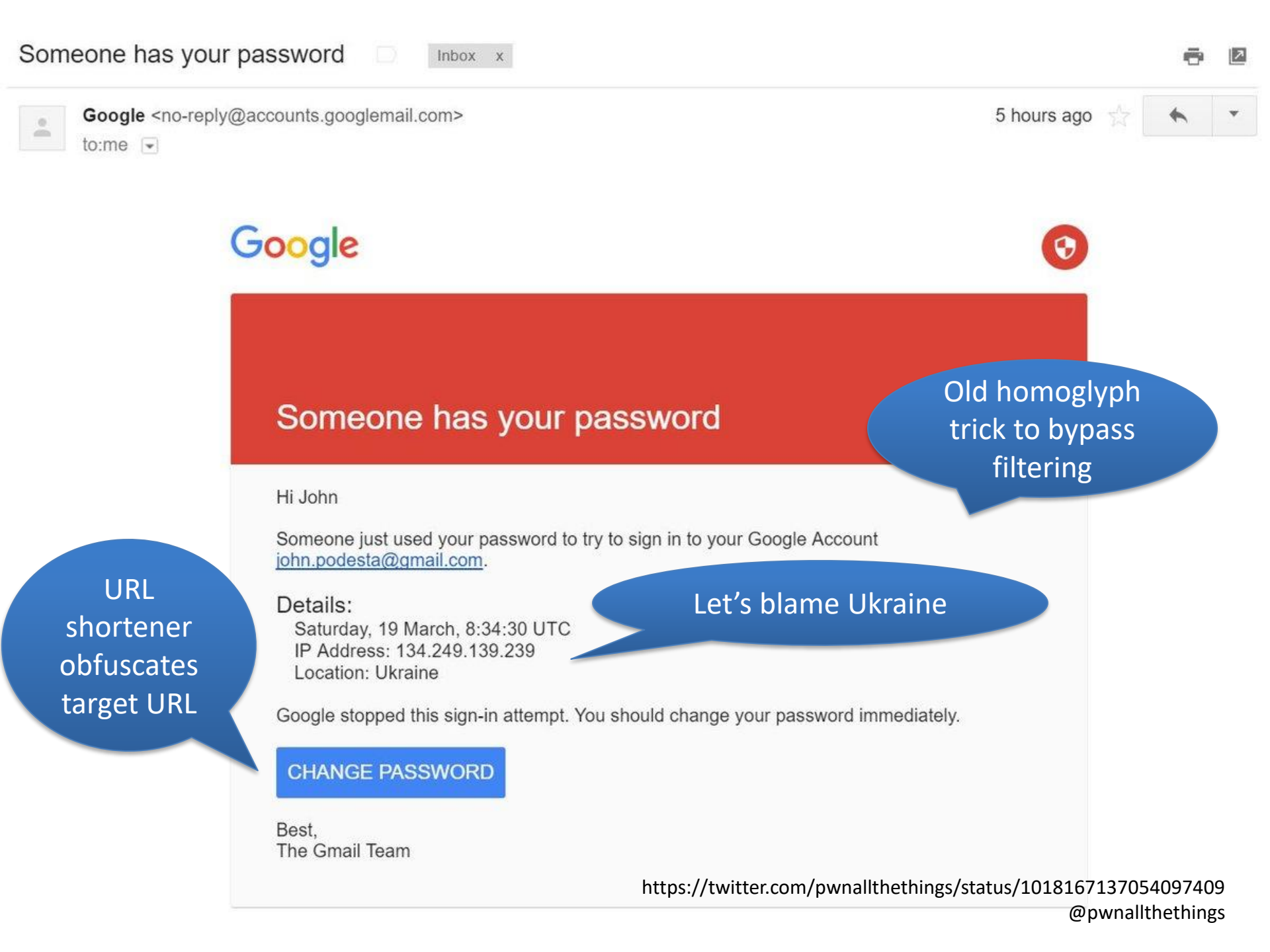
Moving forward we need to be more vigilant with what we trust from the Internet.

It's a time when we need to rely on trusted news sources

Transcripted audio

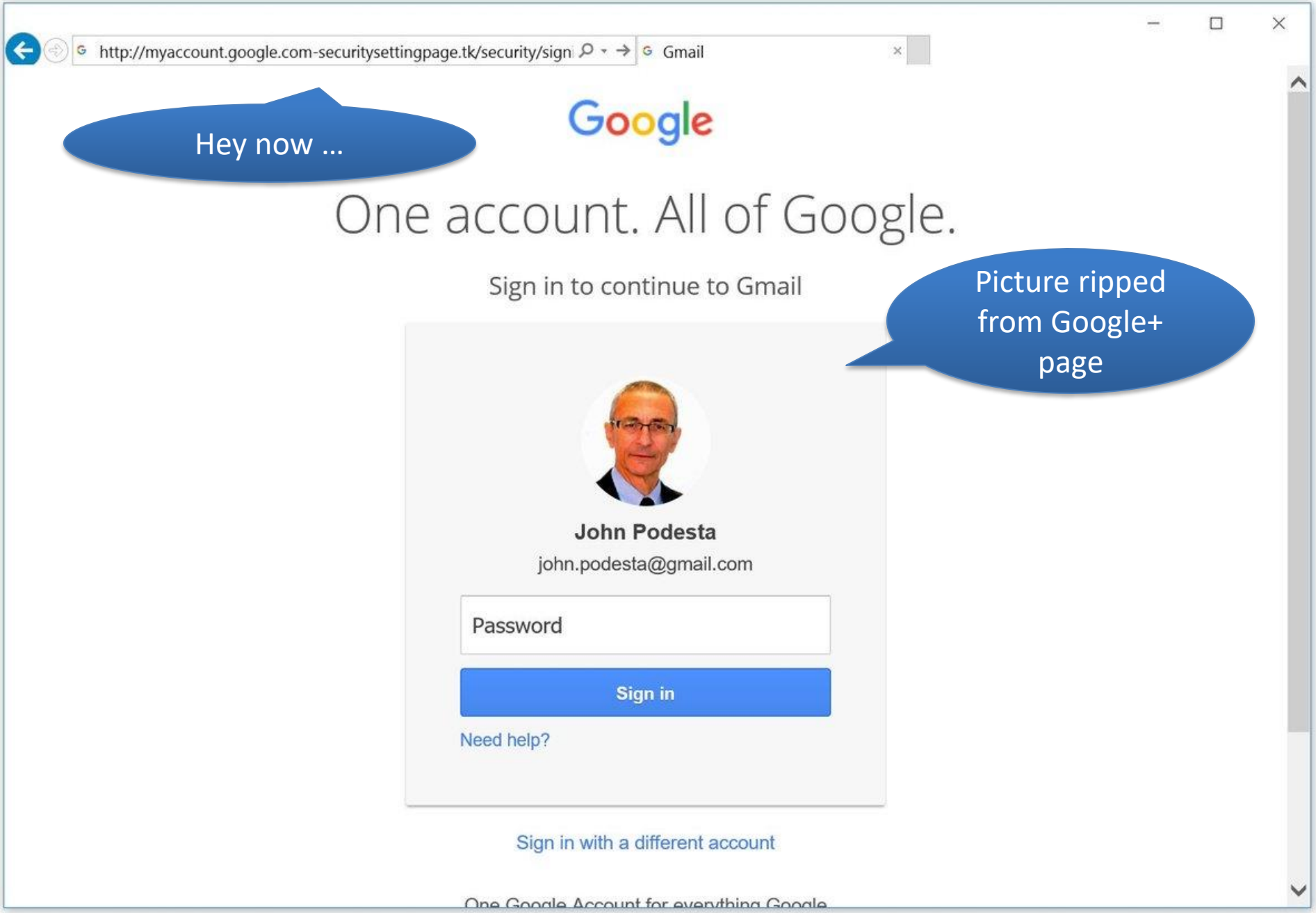
Social Engineering (SE) 101





https://twitter.com/pwnallthethings/status/1018167137054097409

@pwnallthethings



https://twitter.com/pwnallthethings/status/1018167137054097409

@pwnallthethings

Social Engineering (SE) 101

Blended attacks (human/technology/physical/process)

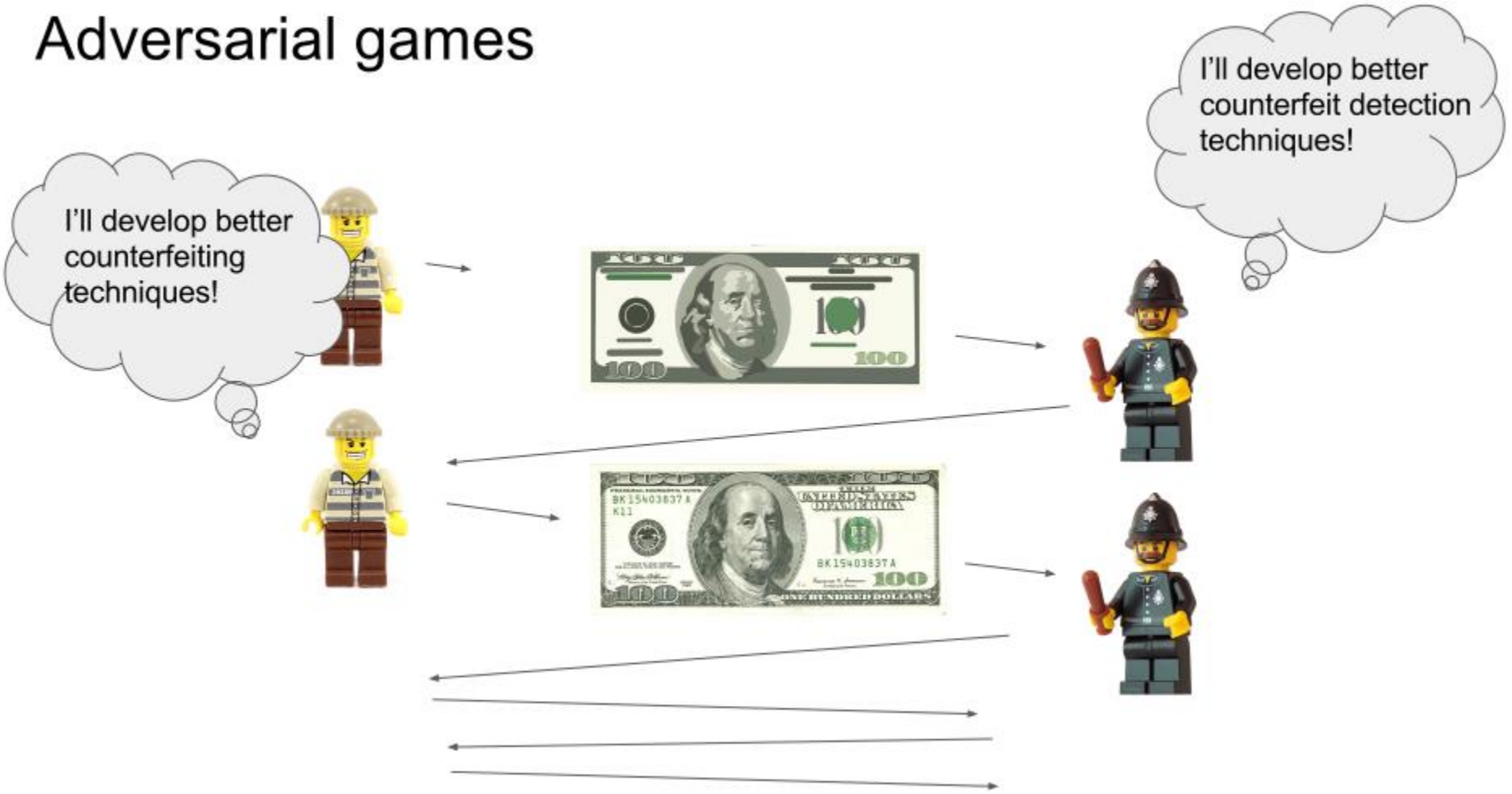
Remote

Email
Messaging
SMS
Voice
etc

Local


COMING SOON

Adversarial games

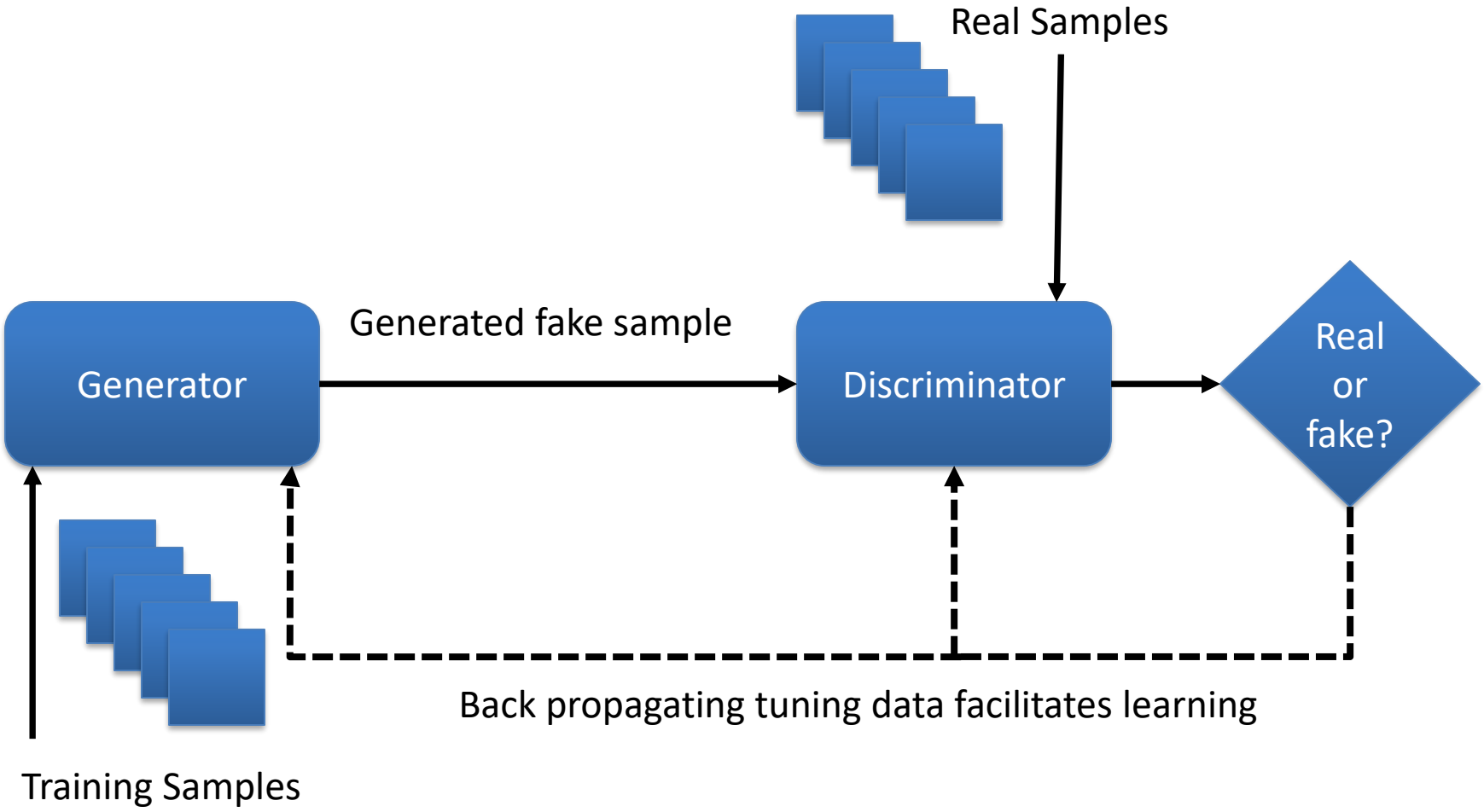


https://github.com/goodfeli/adversarial

"Generative Adversarial Networks." Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. ArXiv 2014.

http://geoffreylitt.com/2017/06/04/enhance-upscaling-images-with-generative-adversarial-neural-networks.html

Generative Adversarial Network (GAN)

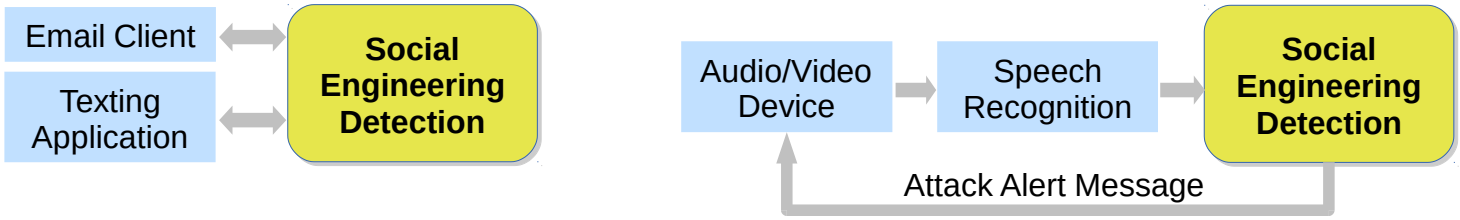


https://github.com/goodfeli/adversarial

"Generative Adversarial Networks." Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. ArXiv 2014.

Photo: Marvel Entertainment; John Byrne and Glynis Wein

Detection of Dialog-Based Attacks

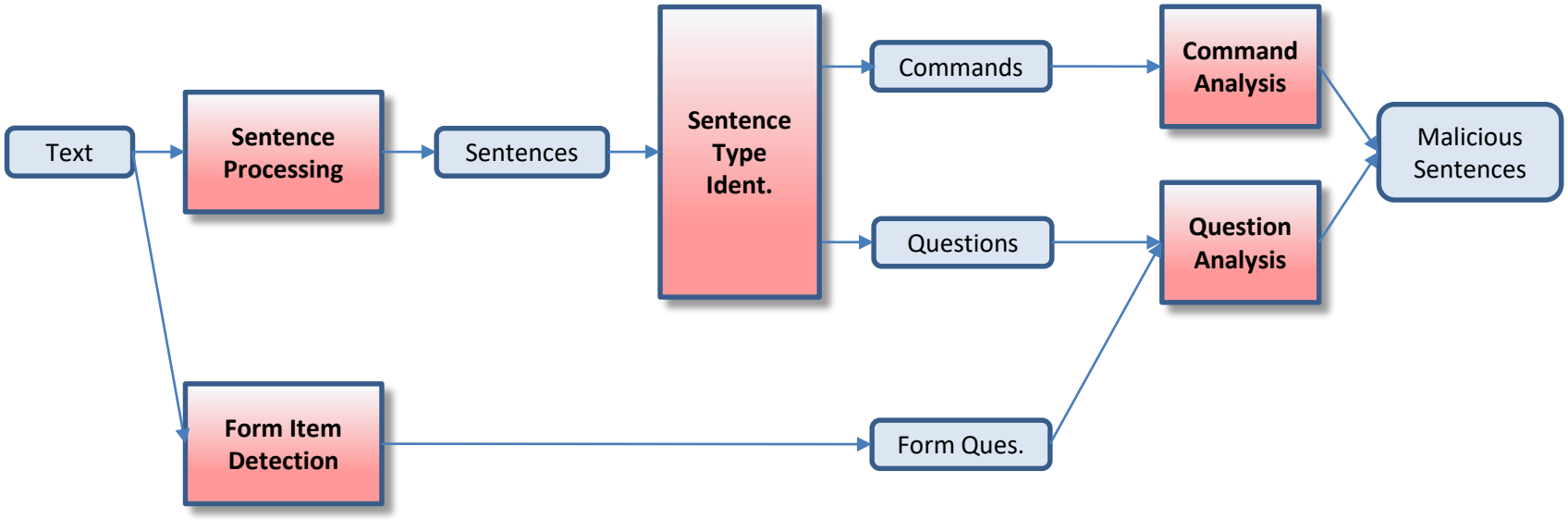


- Difficult because evidence is only in the text of the dialog
- Cannot rely on vector-specific cues
  - images on a phishing website
  - links in a phishing email
- Need to perform some **semantic analysis**
  - consider the meaning of the dialog

Common Features of SE Attacks

- In a social engineering dialog, the attacker must perform one of the following dialog acts:
  1. **Ask an inappropriate question**
     - "What is your social security number?"
  2. **Issue an inappropriate command**
     - "Please click on this link."
- The **topic** of the question/command is forbidden
  - At least for discussion with an untrusted individual

System Structure



- **Question Analysis** and **Command Analysis** are the main steps

Detecting Questions/Commands

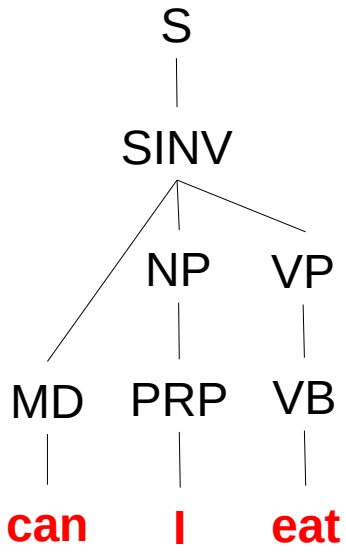
- Parse each sentence using a **syntactic parser**
  - **Stanford Parser**, https://nlp.stanford.edu/software/lex-parser.shtml
- Resulting **parse tree** reveals syntactic structure
  - Parts of speech, phrase decomposition
- Syntactic features are used to identify questions/commands

Question Detection

- Yes/No questions include subject/auxiliary inversion
- The auxiliary verb appears before the subject
  - Auxiliary verbs are “helper” verbs which add meaning
  - “will”, “may”, “can”, etc.
- “I can eat.” vs. “Can I eat?”

Recognition of Yes/No Questions

- SQ or SINV tag



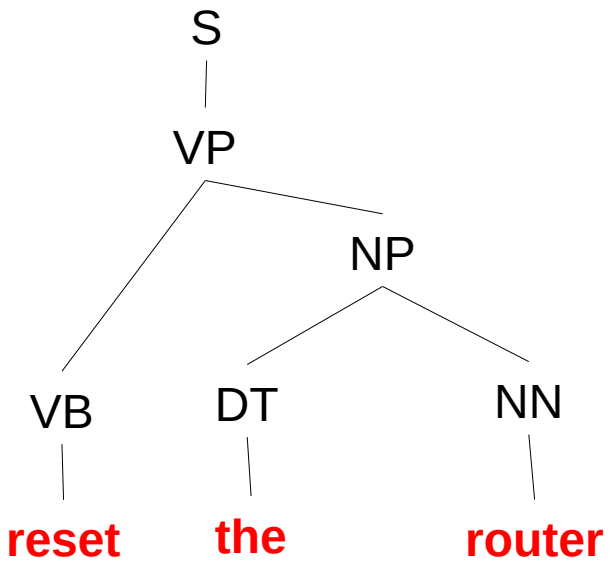
Command Detection

- Imperative clauses express commands
- A direct imperative is a clause which does not contain a subject

"Go home", "Stop right there"

Recognition of Direct Imperatives

- A verb with no preceding noun phrase



Question Analysis

- Our goal is to determine if the answer to a question is private or not
- Sound an alarm if the answer is private data

1. "Where is the bathroom?", answer is not private
2. "What is your social security number?", private, alarm

Question Answer Systems

- User enters a question in natural language
- System provides an answer to the question

Q: “What is the tallest building in South Korea?”

A: Lotte Word Tower

- Search a structured database
  - DBPedia – structured data from wikipedia

Structure of a QA System

- “How old is President Obama?”
- SELECT age FROM agetable WHERE name=“President Obama”

age	name
55	President Obama
47	Ian Harris
35	Rain

QA System for Social Engineering

- We do not need the answer to a question
- We only need to know if the answer is private or not

age	name	private
55	President Obama	No
47	Ian Harris	No
35	Rain	Yes

SELECT private FROM agetable WHERE name="President Obama"

Paralex QA System

"Paraphrase-Driven Learning for Open Question Answering", Anthony Fader and Luke Zettlemoyer and Oren Etzioni, ACL, 2013

rel	arg1	arg2
be_official_language.r	Cantonese	Hong Kong
be_plural_for.r	Bacterium	Bacteria
be_highest_mount.r	Ararat	Turkey

- Searches SQLite database
- Each entry is a triple, (relation, arg1, arg2)

Paralex QA Queries

Natural language:

"What is the nickname of Kansas?"

Query:

SELECT arg2 FROM tuples WHERE rel= "be-nickname.r"
AND arg1= "kansas.e"

Answer:

sunflower-state.e, Private = No

Multiple Queries

- Many SQL queries are generated from each question
- Top ranked SQL query is chosen

"What year was apple founded?"

1. SELECT arg1 FROM tuples WHERE rel= "found.r" AND arg2= "apple.e"
   - Answer is **steve-jobs.e**
2. SELECT arg2 FROM tuples WHERE rel= "be_found_on.r" AND arg1= "apple-computer.e"
   - Answer is **april-1-1976.e**

- If top answer is wrong, private information is wrong

Modification to Database

rel	arg1	arg2
social_security_num.r	<user>	----------
password.r	<user>	----------
location.r	router	----------

- Only keep private triples
- If triple is found in the database, the data is private
- Do not keep actual private data

Privacy from Queries

- Assume that the correct answer is somewhere among the top 15 answers
- A question is private if any of the top 15 answers are private

- Increases the rate of true positives
- May create false positives

Command Analysis

- Determine if the answer to a command is forbidden or not
- Sound an alarm if the command is a forbidden action

1. “Take a left at the next corner.”, command is OK
2. “Please tell me your social security number.”, forbidden, alarm

Command Summarization

- Represent command with **verb-direct object** pair

1. "Take a left at the next corner"
   ("take", "left")
2. "Please give me your password."
   ("give", "password")

- Lookup pair in **topic blacklist**

Verb and Direct Object

- Use **Stanford Typed Dependency Parser** to find the verb and its direct object
- Determines semantic relationships between words

“Please give me your password”

dobj(tell-2, password-5)

- **dobj** relates verb to its direct object

Topic Blacklist

Verb	Direct Object
give	password
send	money

- Pairs can be compiled manually
- We found most relevant pairs in phishing emails
- Used **term-frequency inverse document frequency (TF-IDF)** metric
  - TF-IDF ranking is high if pair is in phishing emails but not in non-phishing emails
  - 100,000 phishing emails and non-phishing emails examined

Experimental Datasets

- Evaluated phishing emails
  - Non-email attacks not available
- Trained with 100,000
  - private answers
  - verb-object blacklist

Database	URL	Size
Scamdex	http://www.scamdex.com	56555
Scamwarners	http://www.scamwarners.com	43241
Scamalot	http://scamalot.com	18149
Antifraudintl	http://antifraudintl.com	69209
Total		187154

- Non-phishing emails taken from the Enron Email Dataset
  - https://www.cs.cmu.edu/~enron/

Experiment Results

	Phishing	Enron
Detected	56616 (True Positive)	14168 (False Positive)
Not-Detected	30432 (False Negative)	72880 (True Negative)

- Precision (TP/(TP+FP)) = 0.80
- Recall (TP/(TP+FN)) = 0.65

- Why so many False Negatives and False Positives?

False Negatives

- 35% of phishing emails were not detected

- Our approach only detects the **punchline** of the attack
  - Malicious question/command
- We cannot detect pretexting or elicitation
- Phishing attacks often involve a **sequence of emails**
- Only the final email may contain the punchline

Analysis of False Negatives

- Manually checked 100 False Negative emails
- 79% were early in the sequence, before the punchline

MY NAME IS MR TERRY ARUMAH FROM GHANA WEST AFRICA . I AM A MARKETING MANGER …
IF YOU ARE INTERESTED **PLEASE YOU CAN CALL US** HERE +2335403977 OR REPLY US HERE OKAY.

- All pretext, invitation to continue the conversation
- Punchline would occur in a later email

False Positives

- 16% of Enron emails which were detected as phishing

- Manually checked 100 False Positive emails
- 97% had malicious commands whose verb-object pairs were not in the blacklist
- i.e. **(“pay”, <number>)** is not in the blacklist

Addressing False Positives

- **Change the blacklist**
  - Add entries manually
  - Decrease TF-IDF threshold
  - May increase False Negatives
- **May need more info than verb-object**
  - Speaker verification may be needed
  - “Pay me $100” is OK if it is your mom talking

Bypassing/Obfuscation

- How might an attacker bypass detection?

1. **Don't ask a question or issue a command.**

2. **State your question/command in a way that uses non-traditional English grammar/words.**

3. **State question/command in a way that our system was not trained for.**

Basically, evade the damn black list e.g. remember the old homoglyph trick

No Question/Command

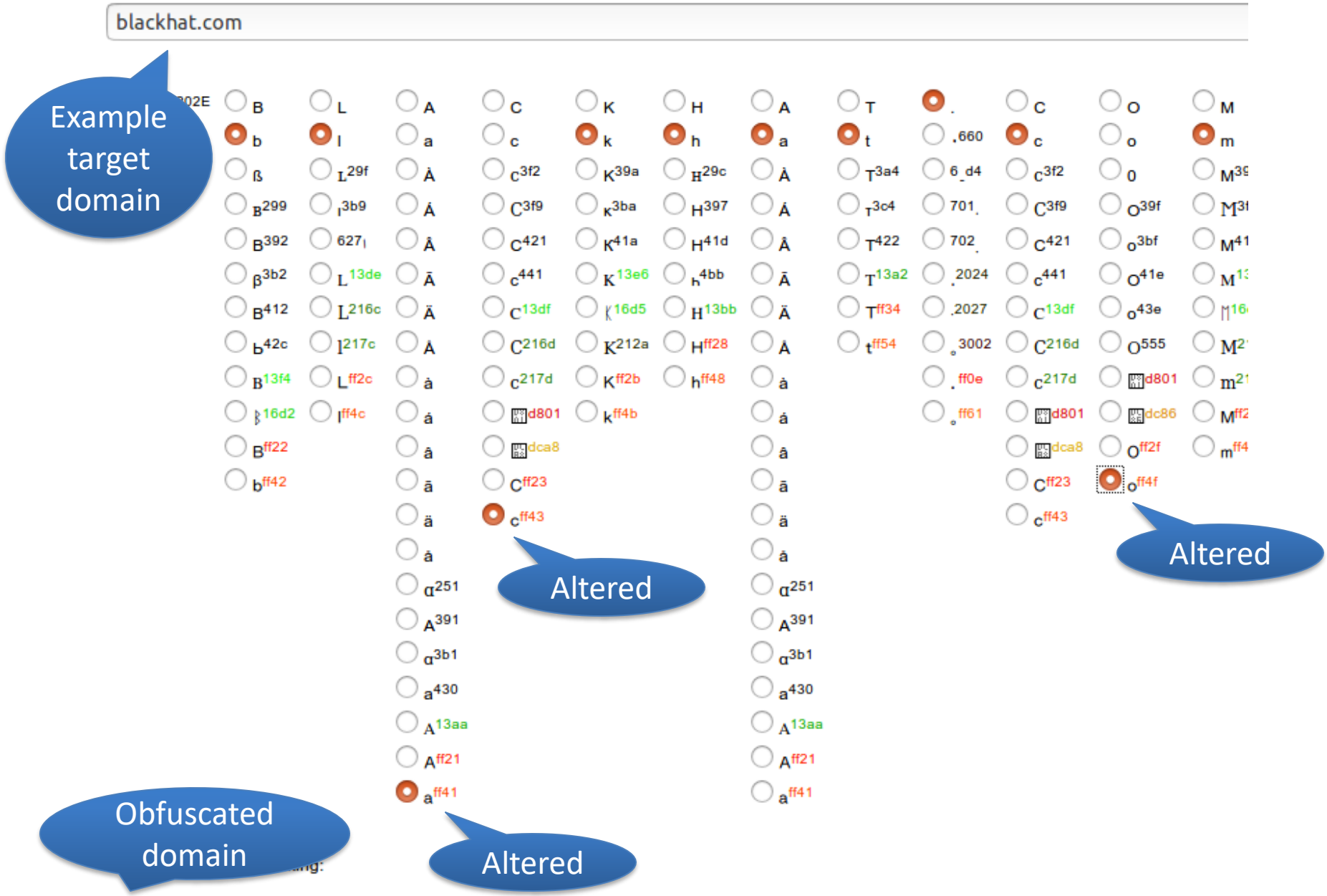
- **Suggest the question/command** without stating it explicitly

"What is your password?" – explicit question
"I can reset your account but I'll need the password first." – implicit question

"Reset the router." – explicit command
"I can fix the problem but the router needs to be reset and I can't do it from here." - implicit command

Non-Traditional English

- Parser (Stanford Parser) has been trained with traditional English sentences
  - Mostly correct grammar
- Non-traditional grammar can cause the parse to be incorrect
  - Slang, patois, Swenglish, fat fingering
  - Leet speak – “G!v3 m3 +h3 m0n3y”


blackhat.com
Example target domain
Altered
Altered
Altered
Obfuscated domain
blackhat.com
This one is so you can copy & paste:
blackhat.com
https://www.irongeek.com/homoglyph-attack-generator.php
@irongeek

Incomplete Training

- Question patterns learned from **WikiAnswers**
  - http://wiki.answers.com
- If a question pattern is used which is not present in WikiAnswers, it would not be recognized

- Topic blacklist learned from:
  - Various phishing email datasets
  - Enron emails (non-phishing)
- Verb-object must be in phishing emails and NOT in Enron emails

Future Research

- Perform scenario-based SE attack experiment with human targets to generate raw data set for further analysis
- Explore GANs within SE audio spoof attack and defense context

Thank You