TROOPERS

Homophonic Collisions

•••

Hold me closer, Tony Danza

Justin Ibarra, Reagan Short, 2023

Hold me closer ... *tiny dancer*??

Who are we

Reagan Short

- Technical Director at BlueVoyant
- Operations Innovation
- Focused on supporting eradication of evil over millions of endpoints by enhancing detection, processing, and communications

@ReaganShort

Pronounces "regex" as: R EH1 G EH1 K S

Justin Ibarra

- Threat Research and Detection Engineering team lead at Elastic
- Threat research primarily on endpoint, cloud, and web
- Build detection and prevention capabilities for EDR, SIEM, and other features at scale

@br0k3ns0und

Pronounces "regex" as: R EH1 JH EH1 K S

Agenda

- Motivation and origin
- Determining (near) homophones
- Homophonic collisions in the wild
- Defensive opportunities
- What's next?

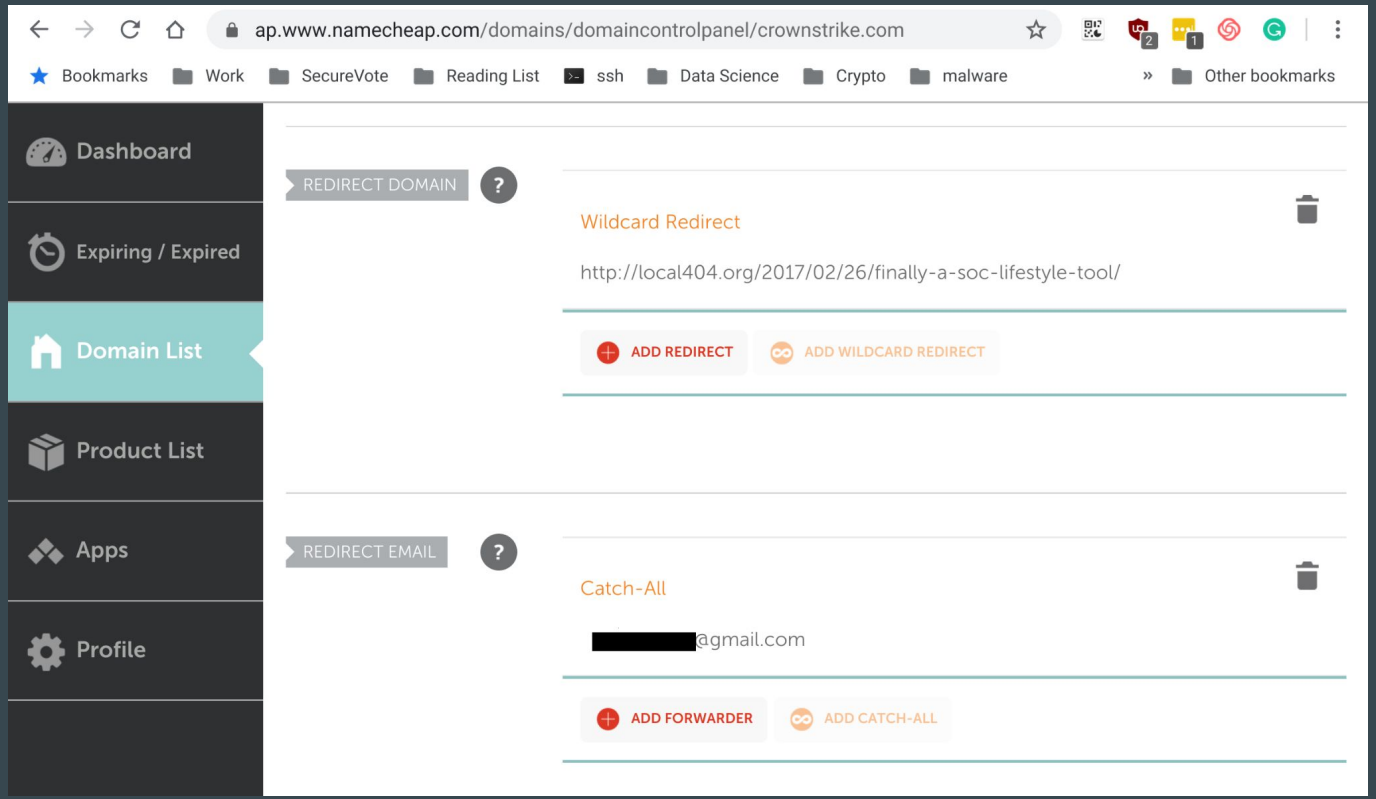
Overview and Motivation

- The continuous integration of language-based “human-machine interaction” increases the impact and urgency of this issue
- Speech-to-text and text-to-speech is resurging with LLMs like ChatGPT
- Highly prevalent risk

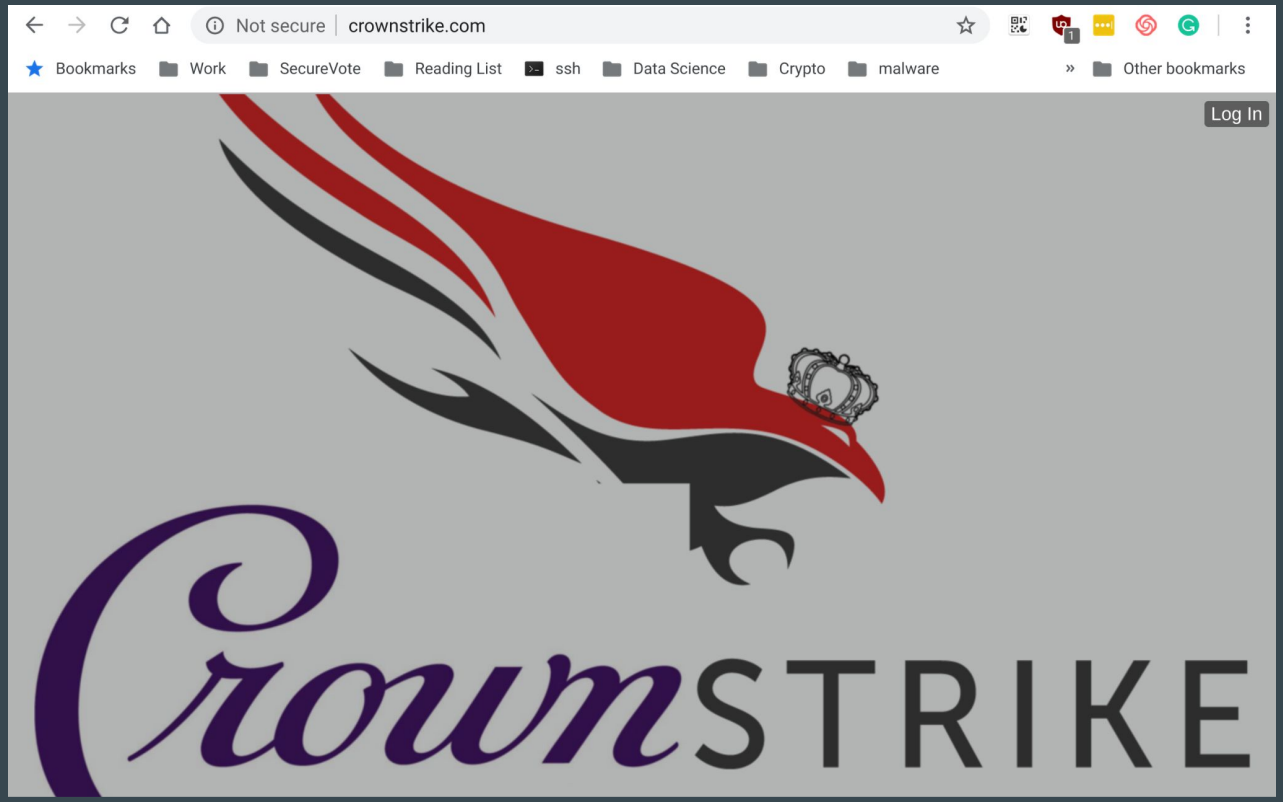
Origin Story

- 2016: Parody domain registered as part of a local SOC training/”crack me” site.
- Catch-all email address created to fwd to personal gmail
- Almost immediately, “oopsies” started flowing in (will showcase some receipts at end)
- Turned into a mutual-interest research project

Origin Story



Origin Story



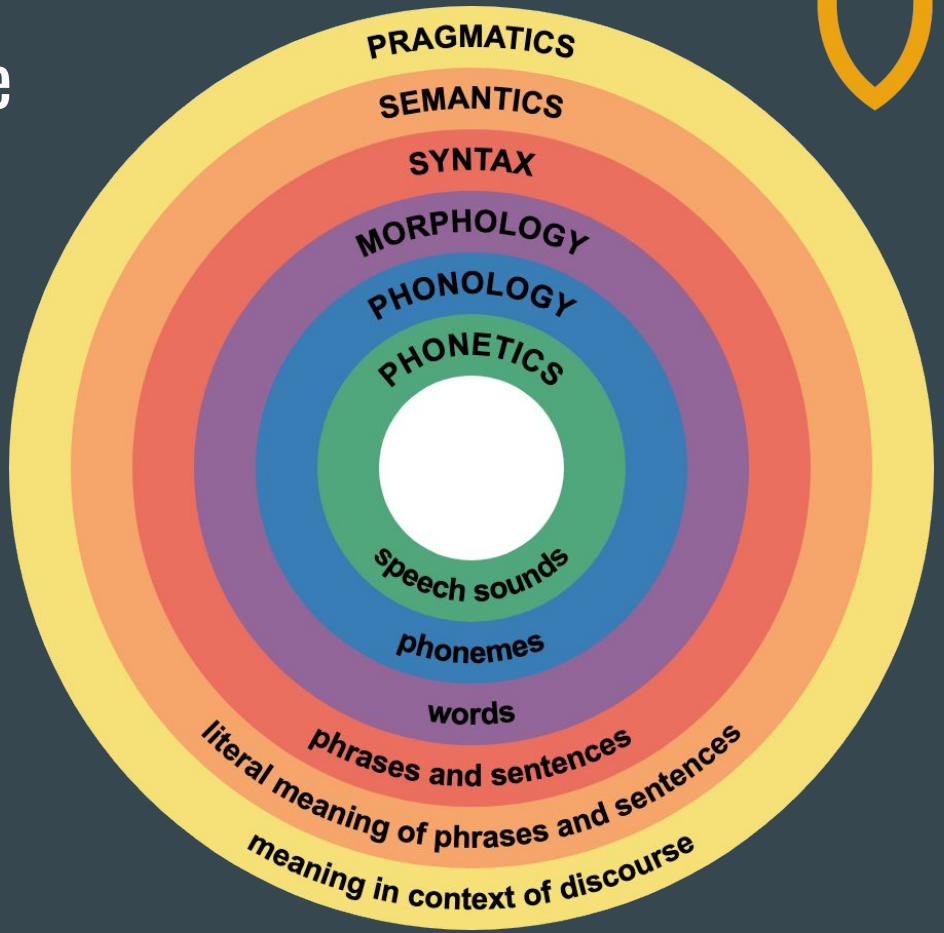
Determining (near) homophones

What's a homophonic collision?

Homophonic collision: a scenario where two or more words or phrases, despite being spelled differently, share the same (or similar) pronunciation, referred to as homophones (or near-homophones).

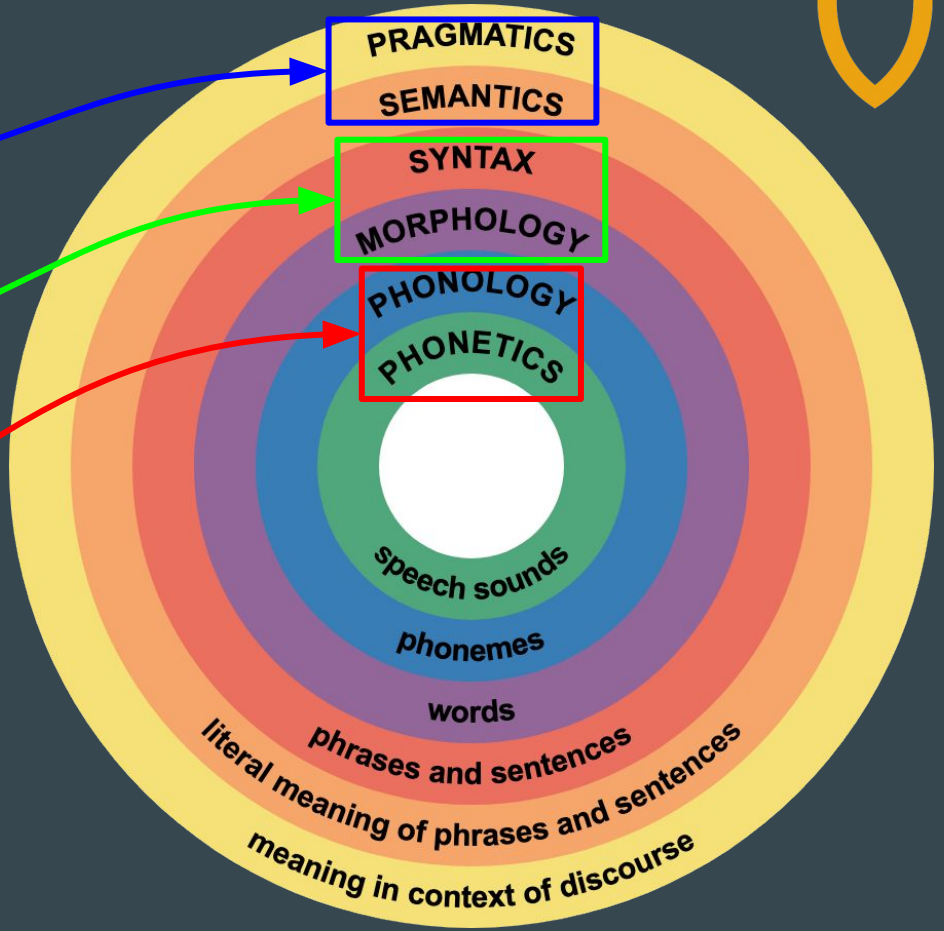
- This can often cause confusion in both spoken and written language if the context doesn't make it clear which word is being referred to.
    - In spoken language, context usually clarifies the intended meaning.
    - In written language, correct spelling is essential to avoid misunderstandings.

Major levels of linguistic structure



Coils of Communication Chaos

- Phishing
  - *Reset bank of america pin to receive funds*
- Homoglyph attacks / domain squatting
  - *Bankofarnerica dot com*
- Homophonic collisions / soundsquatting
  - *Bankofamiracle dot com*



Composition of an audible word

- Phonetic pronunciation comprised of phonemes
- **Phonemes** represent distinct phones (audible sounds or gestures)
  - abstractions of speech sounds that are perceived as equivalent to each other in a given language
- B | IH0 | S | AY1 | D | Z - besides
- B | IY2 | S | AY1 | D | Z - besides, bsides
- B | EH1 | D | S | AY2 | D | Z - bedsides
- T | R | UW1 | P | ER0 | Z - troopers
- T | R | UW1 | P | ER1 | S - true purse

Composition of homophones

- A homophone is two or more words having the same pronunciation but different meanings, origins, or spelling
- D | IH1 | R - deer, dear
- N | UW1 - new, knew
- N | Y | UW1 - new, knew

- Phonemes are identical
- Sometimes across multiple pronunciations (new, knew)

Composition of near-homophones

- A near-homophone is pronounced *almost* the same as another word but has a different spelling and meaning
- P | EY1 | JH - page
- R | EY1 | JH - rage
- D | R | AY1 | V | ER0 - driver
- D | AY1 | V | ER0 - diver

- There is correlation across phonemes, but not completely

Determining near-homophones

- Multiple approaches and algorithms
- Heavily leveraged the Natural Language Toolkit (NLTK)
- CMUdict module - Carnegie Mellon pronunciation dictionary
- Phonemic transcription via the ARPABET

Vowels[2]

ARPABET		IPA	Example(s)
1-letter	2-letter		
a	AA	ɑ~ɒ	b**a**lm, b**o**t (with father–bother merger)
@	AE	æ	b**a**t
A	AH	ʌ	b**u**tt
c	AO	ɔ	c**augh**t, st**o**ry
W	AW	aʊ	b**ou**t
x	AX	ə	comm**a**
–	AXR[3]	ə˞	lett**er**, forw**ar**d
Y	AY	aɪ	b**i**te

Consonants[2]

ARPABET		IPA	Example
1-letter	2-letter		
b	B	b	**b**uy
C	CH	tʃ	**Ch**ina
d	D	d	**d**ie
D	DH	ð	**th**y
F	DX	ɾ	bu**tt**er
L	EL	l̩	bott**le**
M	EM	m̩	rhyth**m**
N	EN	n̩	butt**on**

Determining near-homophones - syllabic partitioning

- AO2 | R | G | AH0 | N | AH0 | Z | EY1 | SH | AH0 | N - organization
- Can leverage phonemic patterns to make assumptions such as syllabic partitioning
  - English' many exceptions make this difficult

AB	Description
0	No stress
1	Primary stress
2	Secondary stress
3...	Tertiary and further stress

- [R | IY0], [T | R | AE1], [K | SH | AH0 | N] - retraction
- [R | IH1], [CH | UW0], [AH0 | L] - ritual
- [EH0], [N | JH | OY1], [M | AH0 | N | T] - enjoyment (should be eN-joy-ment)
- [IH0], [N | F | L | EH1], [K | SH | AH0 | N] - inflection (should be iN-fle-ction)
- [AO2], [R | G | AH0], [N | AH0], [Z | EY1], [SH | AH0 | N] - organization

Determining near-homophones - splitting

- Splitting the word is imperative to isolate discrete phonemic chunks
- IH0 | R | OW1 | ZH | AH0 | N - erosion
  - Sub words
    - ER0 - er
    - AA1 | N, AO1 | N - on
    - AY1 | AH0 | N, AY1 | AA2 | N - ion
  - Syllabic partitioning
    - [IH0], [R | OW1], [ZH | AH0 | N]
    - Reassemble chunks across all combinations
    - Reverse lookup (phoneme to grapheme)
    - [R | OW1] - reaux, rheault, rho, ro, roe, roh, rohe, row, rowe, wroe
      - e-roe-sion, e-rowe-sion, e-row-sion, etc.

Qualifying near-homophones

- Not all near-homophones are equal, so relative value (scoring) is important
- Several algorithms
  - Absolute phoneme correlation
  - Contiguous relative phoneme correlation
  - Phoneme Levenshtein distance (x2)
- Potential Model Approach
  - Using state of the phonemic representation for training models
  - Convert phonemes to numeric representation (also a potential for Levenshtein approach)
- Future Research around Acoustic Approach
  - Audio Samples
  - Fast Fourier Transforms

Challenges and opportunities

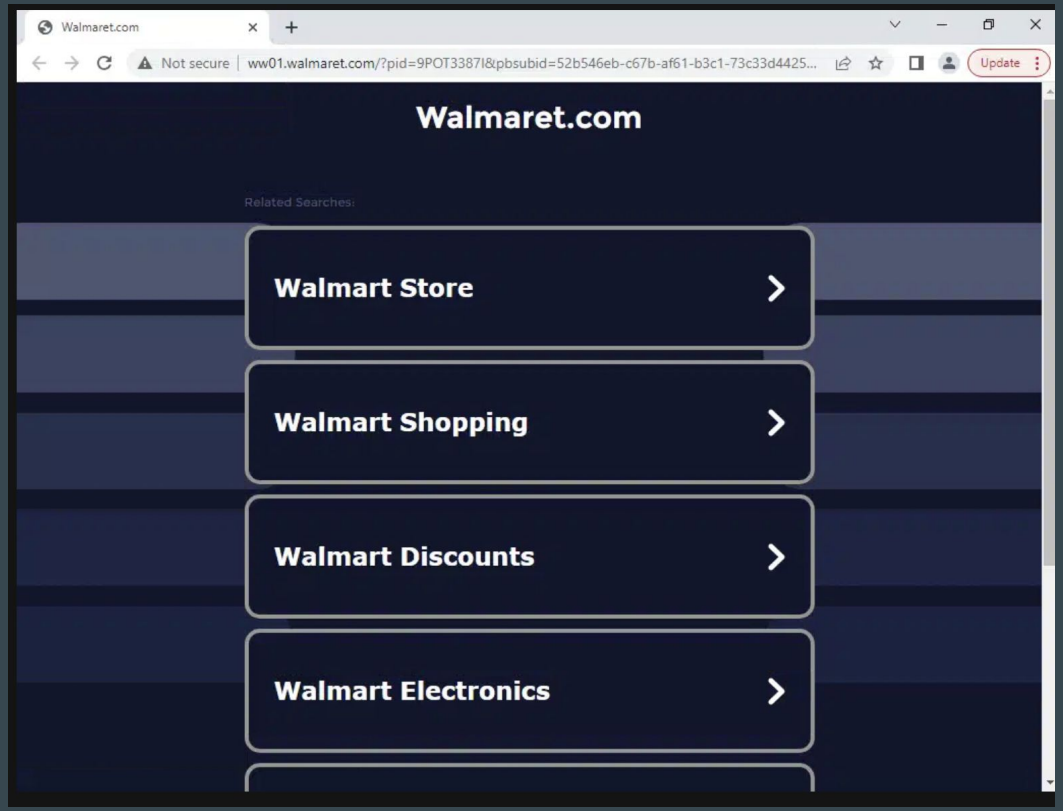
- Tech lexicon
  - Fun words: **colorama**, **asciimatics**,
  - Tech speak and colloquialisms: **bsides** (besides)
  - Abbreviation pronunciation: **io** (eye oh, ee oh)
- Dialects
  - Cultural, regional, familial
- Language-specific exceptions
  - English

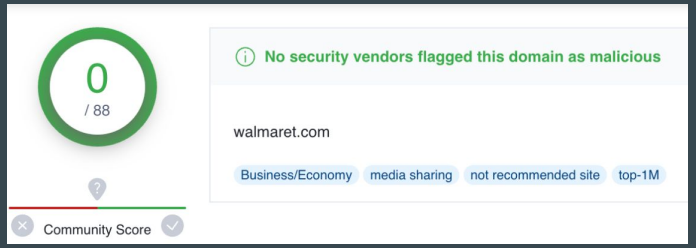
Homophonic collisions in the wild

Finding vulnerable domains

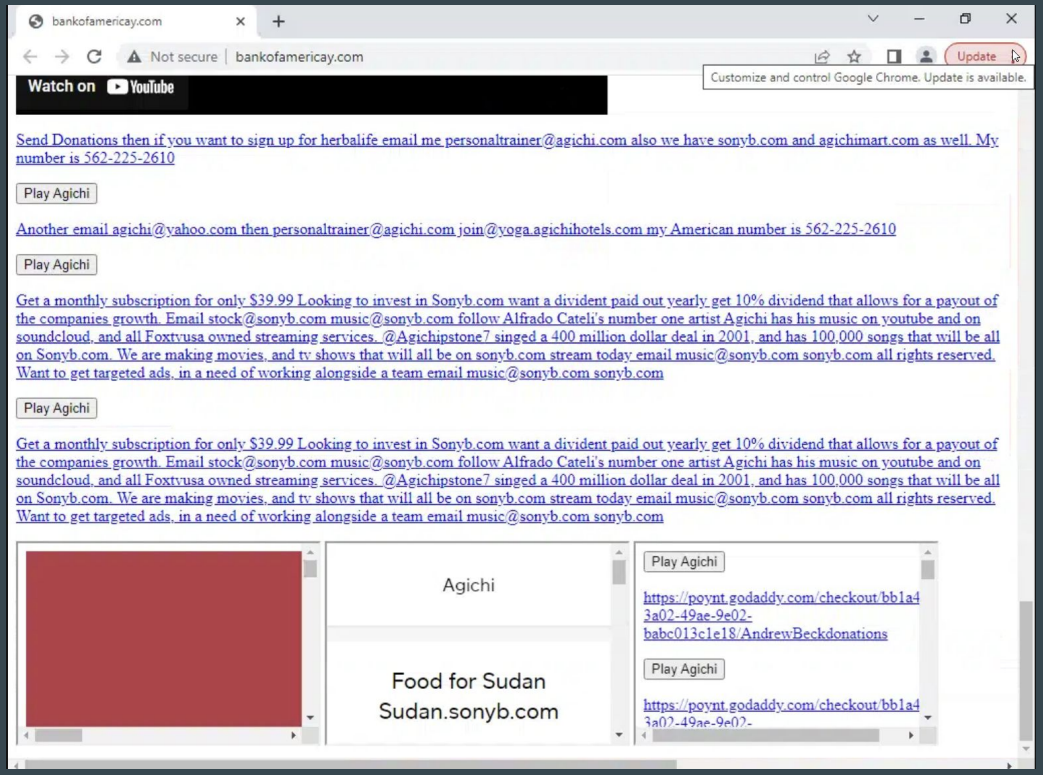
- Filtered Alexa top 1000 to US-based *.com*
- Determined homophones and near-homophones (similar to DNSTwist) and scanned all combinations:
  - Is registered?
  - Is connected?
  - Is suspicious (VT)

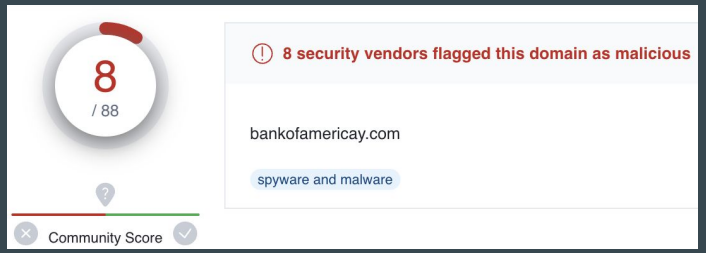
Vulnerable domains: Walmaret .com





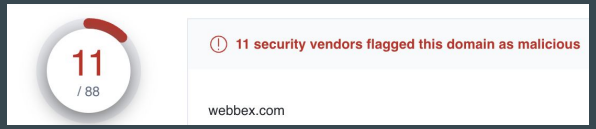
Vulnerable domains: bankofamericay .com



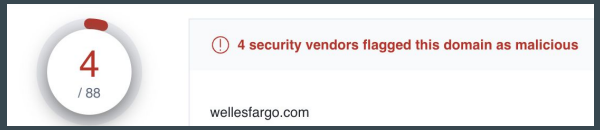


Vulnerable domains

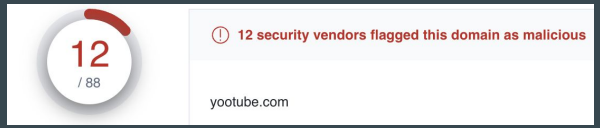
- Webex: webbex



- Wells fargo: wellesfargo



- Youtube: yootube



Finding vulnerable packages

- Took most popular pypi package downloads list
- Determined homophones and near homophones and scanned
- Validated against DataDog GuardDog malicious software data set
- Homophonic collision packages;
  - aiiohttp
  - coloroma
  - colouramas
  - colurama
  - dequests
  - fequests
  - gequests
  - requesfs
  - requeste

Other risks and threat vectors

- Speech2Text / Text2Speech
  - Wrappers around LLMs like ChatGPT
  - Assistants like Siri and Google
- Other “registered namespaces”
  - Social media handles / usernames
  - Email addresses (soundalike@gmail.com)

Super official in-the-wild examples

- Examples of real world sensitive data collected from homophonic collision incidents

Super official in-the-wild examples

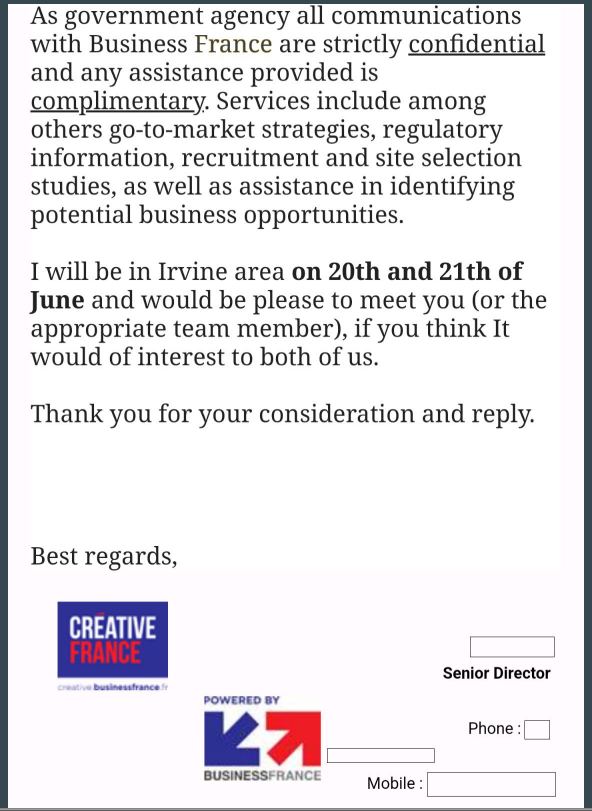
Super official in-the-wild examples

10:56AM	**Flight Information** **Carrier:** DELTA AIR LINES INC **Operated by:** DELTA AIR LINES INC **Airline Booking Ref:** HG7 **Flight:** DL 687 **Class:** BUSINESS **Origin:** Atlanta, GA, Hartsfield Jackson Intl Arpt (ATL) **Departing:** Monday 20 May 2019 at 10:56AM **Departure Terminal:** MAYNARD JACKSON INTL TERM **Destination:** Grand Cayman Island, Owen Roberts Arpt (GCM)

Super official in-the-wild examples

Trip Routing Information:	**PU:** -- : 55 CHURCH STREET NEW YORK 55 Church St New York, NY 10007 (United States of America) **WT:** -- : 340 MacARTHUR BVD MAHWAH 340 MacArthur Blvd Mahwah, NJ 07430 (United States of America) **DO:** -- : 55 CHURCH STREET NEW YORK 55 Church St New York, NY 10007 (United States of America)

Super official in-the-wild examples



As government agency all communications with Business France are strictly confidential and any assistance provided is complimentary. Services include among others go-to-market strategies, regulatory information, recruitment and site selection studies, as well as assistance in identifying potential business opportunities.

I will be in Irvine area **on 20th and 21th of June** and would be please to meet you (or the appropriate team member), if you think It would of interest to both of us.

Thank you for your consideration and reply.

Best regards,

CREATIVE FRANCE
creative.businessfrance.fr

[redacted]
Senior Director

POWERED BY
BUSINESSFRANCE

Phone : [redacted]

[redacted]

Mobile : [redacted]

Super official in-the-wild examples

CREDIT CARD AUTHORIZATION

For your protection and to authorize us to charge your credit card, please complete and sign this form. When completed, e-mail t[redacted] fax [redacted] at [redacted], or mail t[redacted], Irvine, CA 92614.

Card Type (check one) ☐ Visa ☐ Master Card ☐ Discover xAmerican Express

Card Number: __378[redacted]01183_____

Expiration Date __12/20_____ (MM/YY)

Super official in-the-wild examples

INCOME WITHHOLDING FOR SUPPORT

(X) ORIGINAL INCOME WITHHOLDING ORDER/NOTICE FOR SUPPORT (IWO)
( ) AMENDED IWO
( ) ONE-TIME ORDER/NOTICE FOR LUMP SUM PAYMENT
( ) TERMINATION OF IWO

Date: January 30, 20[redacted]

(X) Child Support Enforcement (CSE) Agency ( ) Court ( ) Attorney ( ) Private Individual/Entity (Check One)

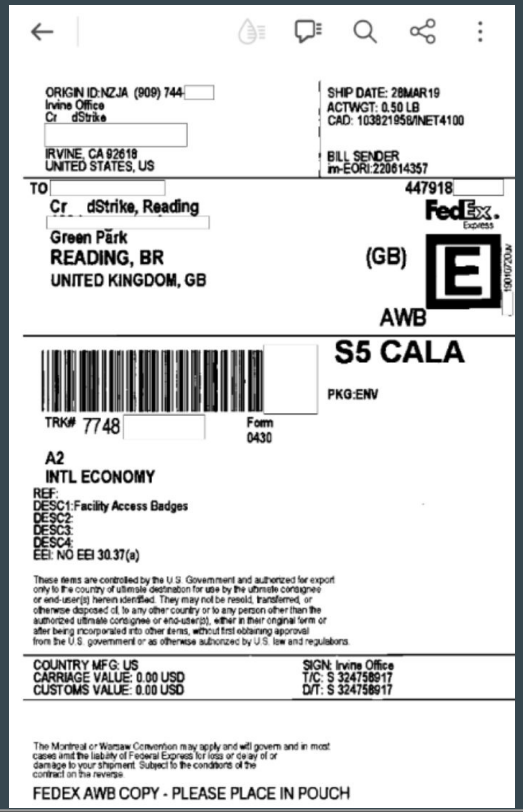
NOTE: This IWO must be regular on its face. Under certain circumstances you must reject this IWO and return it to the sender (see IWO instructions www.acf.hhs.gov/programs/css/resource/income-withholding-for-support-instructions). If you receive this document from someone other than a state or tribal CSE agency or a court, a copy of the underlying order must be attached.

State/Tribe/Territory	Indiana	Remittance ID (include w/payment)	0007[redacted]
City/County/Dist./Tribe	[redacted]	Order ID	02C0[redacted]710
Private Individual/Entity		CSE Agency Case ID	0007[redacted]

Employer/Income Withholder's Name
CRO[redacted]INC
Employer/Income Withholder's Address
[redacted]

RE: Employee/Obligor's Name (First, M, Last)
[redacted]
Employee/Obligor's Social Security Number
[redacted]
Employee/Obligor's Date of Birth

Super official in-the-wild examples



ORIGIN ID:NZJA (909) 744-
Irvine Office
Cr dStrike

IRVINE, CA 92618
UNITED STATES, US

SHIP DATE: 28MAR19
ACTWGT: 0.50 LB
CAD: 103821958/INET4100

BILL SENDER
im-EORI:220614357

TO
Cr dStrike, Reading
Green Park
READING, BR
UNITED KINGDOM, GB

447918

FedEx. Express

(GB)

AWB

S5 CALA

PKG:ENV

TRK# 7748

Form 0430

A2
INTL ECONOMY
REF:
DESC1:Facility Access Badges
DESC2:
DESC3:
DESC4:
EEI: NO EEI 30.37(a)

These items are controlled by the U.S. Government and authorized for export only to the country of ultimate destination for use by the ultimate consignee or end-user(s) herein identified. They may not be resold, transferred, or otherwise disposed of, to any other country or to any person other than the authorized ultimate consignee or end-user(s), either in their original form or after being incorporated into other items, without first obtaining approval from the U.S. government or as otherwise authorized by U.S. law and regulations.

COUNTRY MFG: US
CARRIAGE VALUE: 0.00 USD
CUSTOMS VALUE: 0.00 USD

SIGN: Irvine Office
T/C: S 324758917
D/T: S 324758917

The Montreal or Warsaw Convention may apply and will govern and in most cases limit the liability of Federal Express for loss or delay of or damage to your shipment. Subject to the conditions of the contract on the reverse.

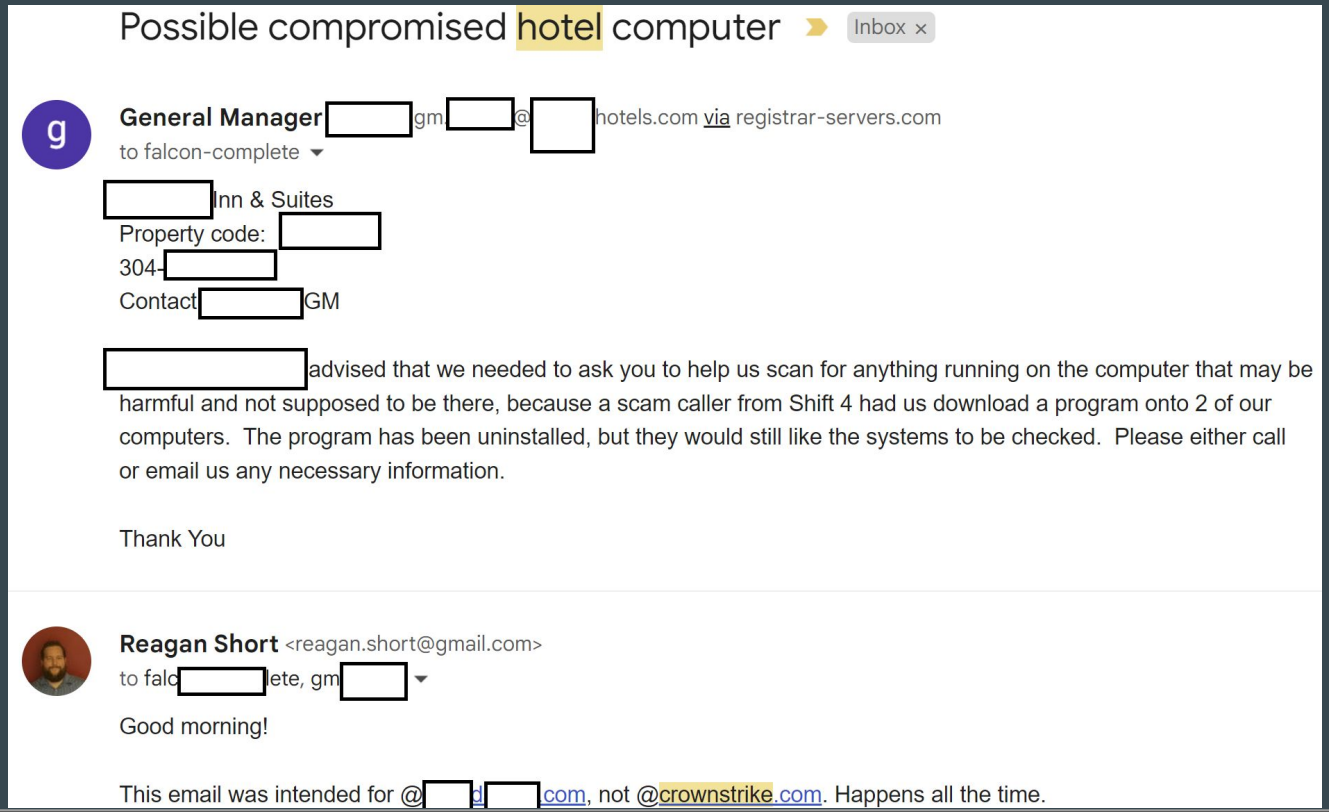
FEDEX AWB COPY - PLEASE PLACE IN POUCH

A2

INTL ECONOMY

REF:
DESC1:Facility Access Badges
DESC2:
DESC3:
DESC4:
EEI: NO EEI 30.37(a)

Super official in-the-wild examples



Possible compromised hotel computer Inbox ×

General Manager [redacted]gm[redacted]@[redacted]hotels.com via registrar-servers.com
to falcon-complete

[redacted] Inn & Suites
Property code: [redacted]
304-[redacted]
Contact [redacted] GM

[redacted] advised that we needed to ask you to help us scan for anything running on the computer that may be harmful and not supposed to be there, because a scam caller from Shift 4 had us download a program onto 2 of our computers. The program has been uninstalled, but they would still like the systems to be checked. Please either call or email us any necessary information.

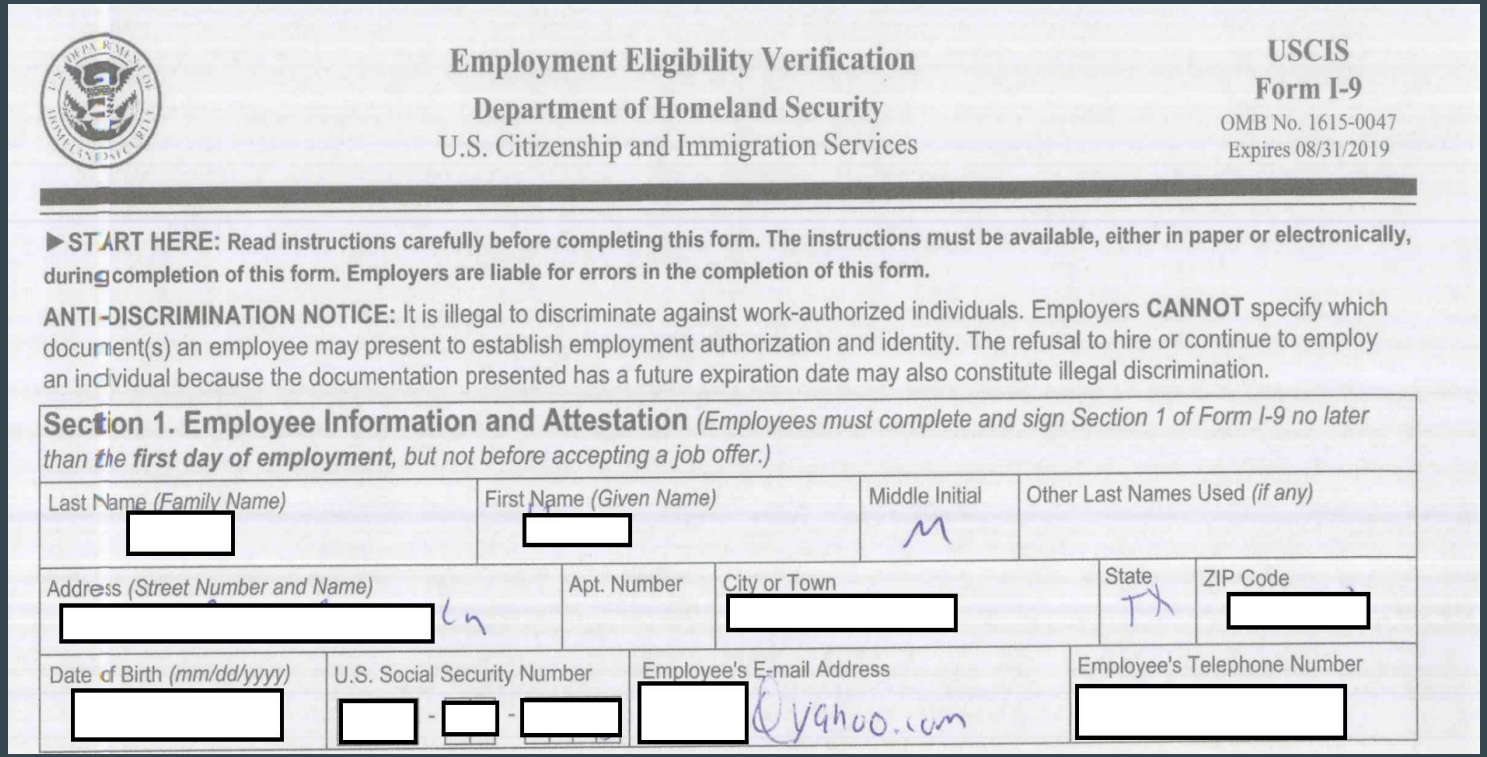
Thank You

Reagan Short <reagan.short@gmail.com>
to falc[redacted]lete, gm[redacted]

Good morning!

This email was intended for @[redacted]d[redacted]com, not @crownstrike.com. Happens all the time.

Super official in-the-wild examples



Employment Eligibility Verification
Department of Homeland Security
U.S. Citizenship and Immigration Services

USCIS
Form I-9
OMB No. 1615-0047
Expires 08/31/2019

▶START HERE: **Read instructions carefully before completing this form. The instructions must be available, either in paper or electronically, during completion of this form. Employers are liable for errors in the completion of this form.**

ANTI-DISCRIMINATION NOTICE: It is illegal to discriminate against work-authorized individuals. Employers **CANNOT** specify which document(s) an employee may present to establish employment authorization and identity. The refusal to hire or continue to employ an individual because the documentation presented has a future expiration date may also constitute illegal discrimination.

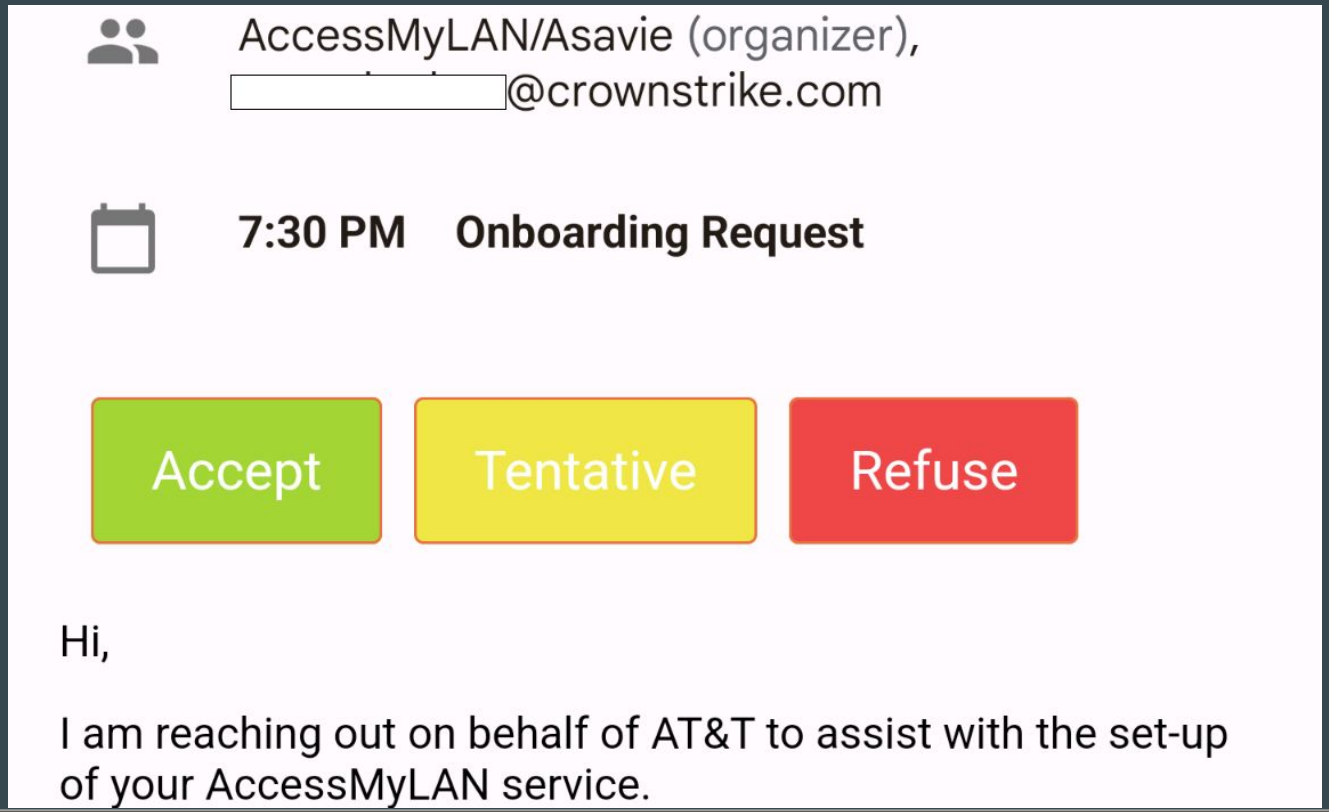
Section 1. Employee Information and Attestation *(Employees must complete and sign Section 1 of Form I-9 no later than the **first day of employment**, but not before accepting a job offer.)*

Last Name *(Family Name)*	First Name *(Given Name)*	Middle Initial	Other Last Names Used *(if any)*
[redacted]	[redacted]	M	

Address *(Street Number and Name)*	Apt. Number	City or Town	State	ZIP Code
[redacted]		[redacted]	TX	[redacted]

Date of Birth *(mm/dd/yyyy)*	U.S. Social Security Number	Employee's E-mail Address	Employee's Telephone Number
[redacted]	[redacted]	[redacted]@yahoo.com	[redacted]

Super official in-the-wild examples



Super official in-the-wild examples

From: **Tuscany Suites Casino** <noreply@canarytechnologies.com>
Date: Fri, Mar 31, 2023 at 9:04 AM
Subject: Tuscany Suites & Casino - Complete pre-arrival check-in
To: [redacted]@besideslv.org>

TUSCANY
SUITES & CASINO

Contactless Check-In

We look forward to your upcoming stay with us at The Tuscany Suites and Casino! Below are your reservation details:

Name: [redacted]
Your Stay: Mar 31st, 2023 - Apr 03rd, 2023
Confirmation: 537727

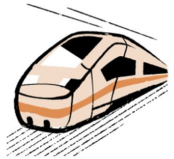
Super official in-the-wild examples

From: **BILL** <account-services@hq.bill.com>
Date: Wed, May 17, 2023 at 1:32 AM
Subject: Security BSides Las Vegas, Inc, a payment from [redacted]s on the way!
To: Security BSides Las Vegas, Inc <sponsors@besideslv.org>

bill

You've got money on the way from

[redacted]



$12,000.00
Total amount

Invoice #
1045

Defensive opportunities

Detection methodologies and opportunities

- API lookup for real-time lookup or pre-emptive preparation
  - Real-time/event-driven processing
  - Preprocessing known namespaces of interest
- Detection rule looking for all homophones from explicitly defined words
- Subscribe to domain registration and SSL certificate registration
- Expose resources to pypi and other package ecosystems to detect

Detection methodologies and opportunities

- Generate homophones and near homophones list
- Explicitly alert on domain collisions connection attempts

```
network where dns.question.name : ("truepurse.de",
                                   "trueparse.de",
                                   "troophearse.de",
                                   "troophurts.de")
```

Detection methodologies and opportunities

- Generate homophones and near homophones list
- Explicitly alert on domain collisions successful connection

```
network where url.domain : ("truepurse.de",
                            "trueparse.de",
                            "troophearse.de",
                            "troophurts.de") and
  and http.response.status_code >= 200 and http.response.status_code <= 200
```

Detection methodologies and opportunities

- Generate homophones and near homophones list
- Explicitly alert on domain collisions successful connection downloading a file

```
sequence by process.parent.entity_id with maxspan=5m
  [network where url.domain : ("truepurse.de",
                               "trueparse.de",
                               "troophearse.de",
                               "troophurts.de") and
    and http.response.status_code >= 200 and http.response.status_code <= 200
  [file where event.type == "creation"]
```

What's next?

phonemenal

- Phonemenal.io (coming soon)
  - API, near-homophone lookup, etc.
- Python package
  - `pip install phonemenal` (coming soon)
  - A phenomenal toolkit for having fun with phonemes and homophonic collisions

Thank you

@br0k3ns0und | @ReaganShort