Using Deep Learning to Uncover Darkweb Malicious Actors and Their Close Circle

By @rodsoto @josephzadeh

$Whoami

Rod Soto
Director of Security Research at JASK
Co-Founder of Hackmiami/Hack The Valley %27
President Pacific Hackers Conference

Joseph Zadeh
Director of Data Science at JASK
Co-Founder of Hack The Valley %27

Dark web

Dark Web VS Deep Web

Deep Web

- Business intranets, web archives, password-protected websites. Many times larger than surface web.

Dark Web

- Websites are not indexed by normal search engines. Accessible mainly via TOR .

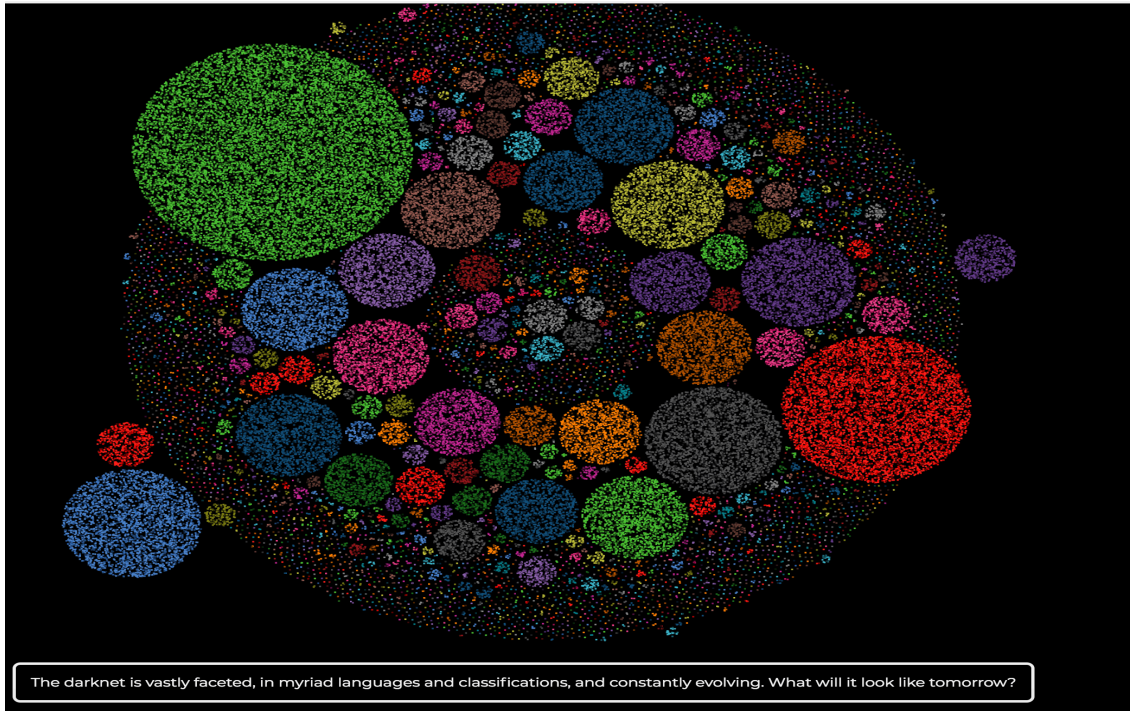
*darknets: The darknets which constitute the dark web include small, friend-to-friend peer-to-peer networks, as well as large, popular networks like Tor, Freenet, I2P and Riffle operated by public organizations and individuals. *

The Dark Web facts

- Part of Deep Web NOT all of it.
- Accessible mainly via TOR (Onionland)
- Not indexed by standard search engines
- Provides a level of anonymity acceptable for criminal operators
- Heavily monitored, regulated by LE actions
- Low latency
- Driven by cryptocurrencies (BTC,XMR, ETH, LTC, etc)

DarkOwl: What Does the Darknet Look Like?

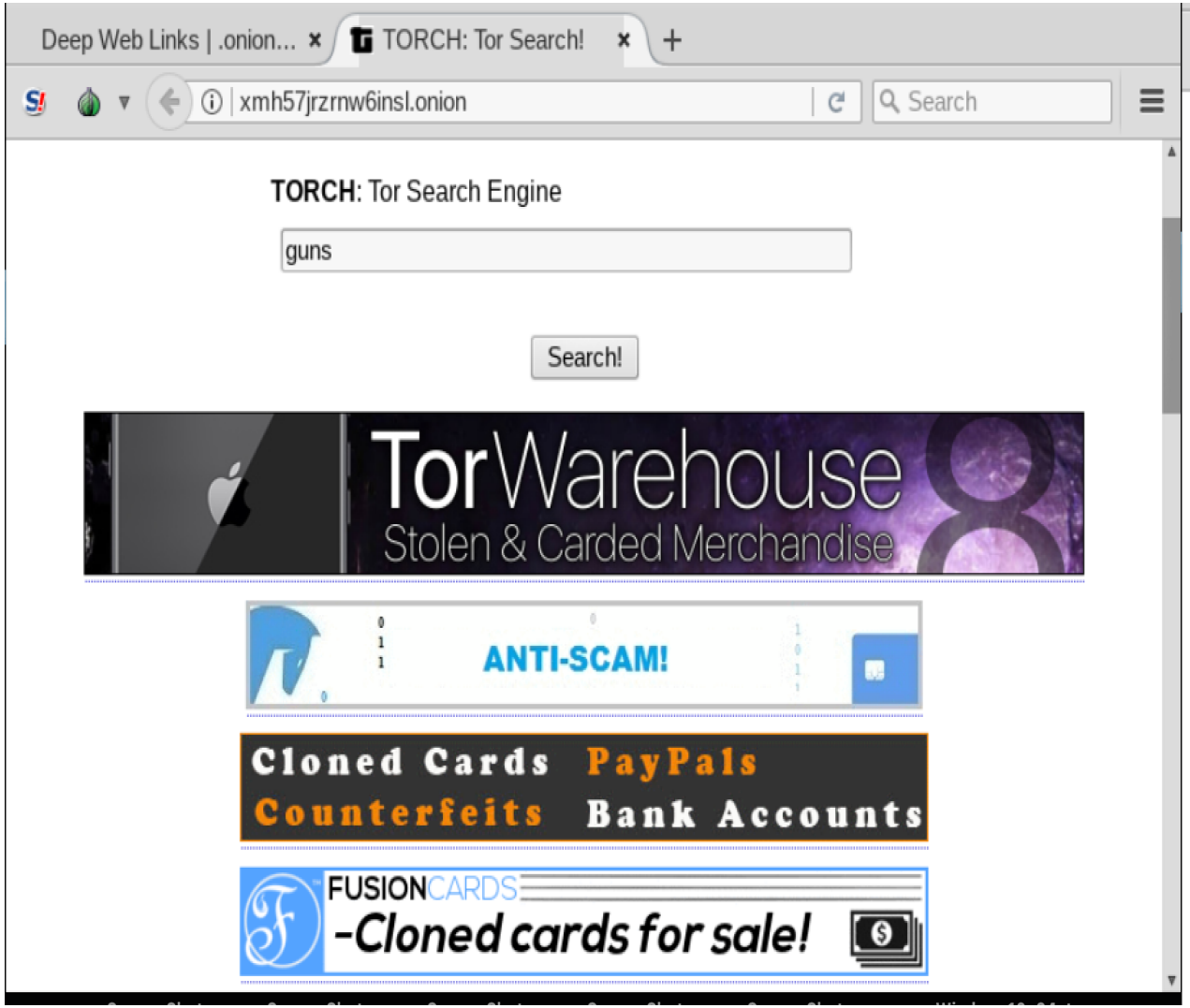
- Darkowl has an index of over 500,000,000 sites across the entire DarkWeb.
- One of the key partners we owe a big thanks to contributing time and access to data is DarkOwl!! Viz from here: https://mapthedark.com/

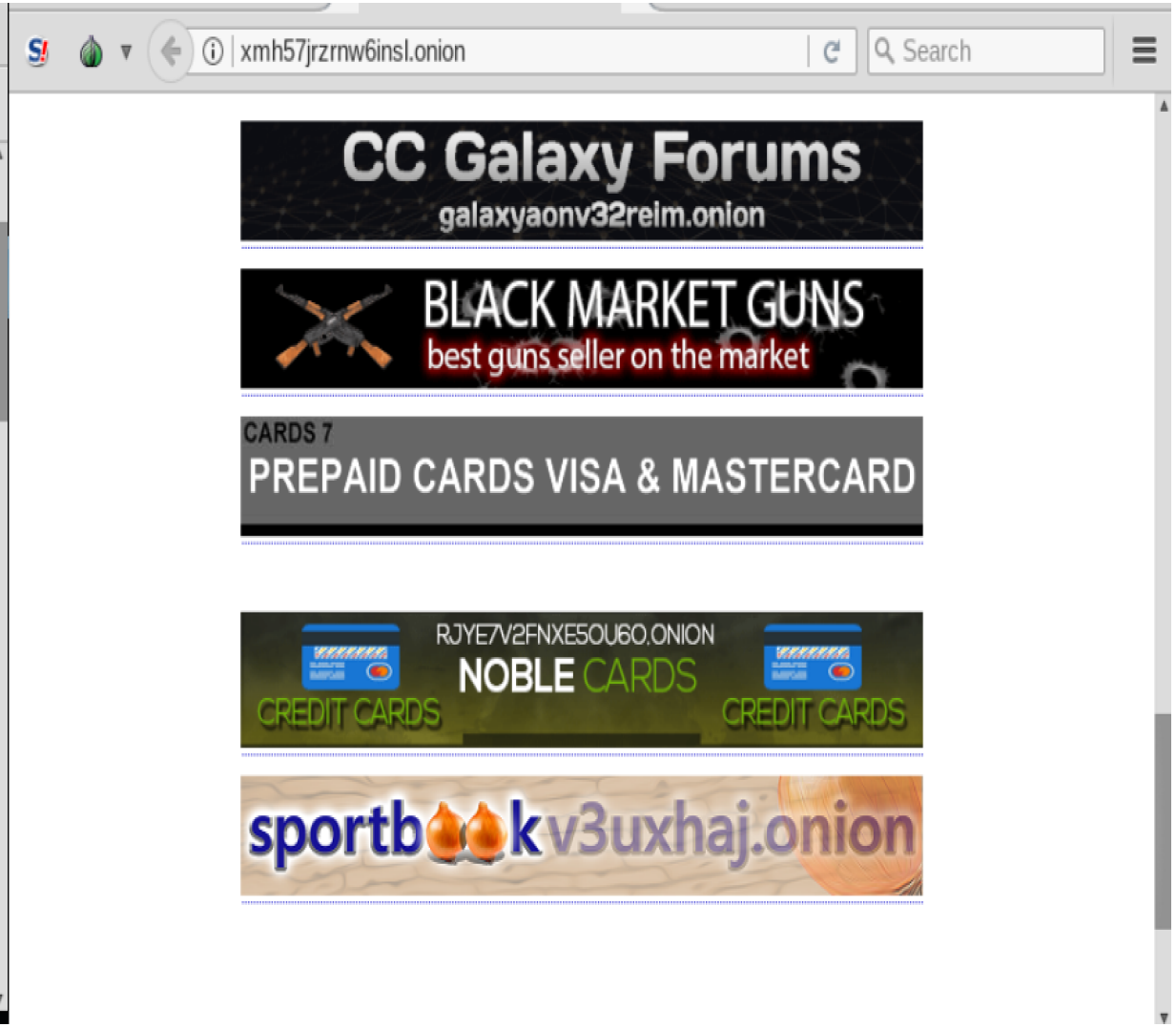


DARKOWL™

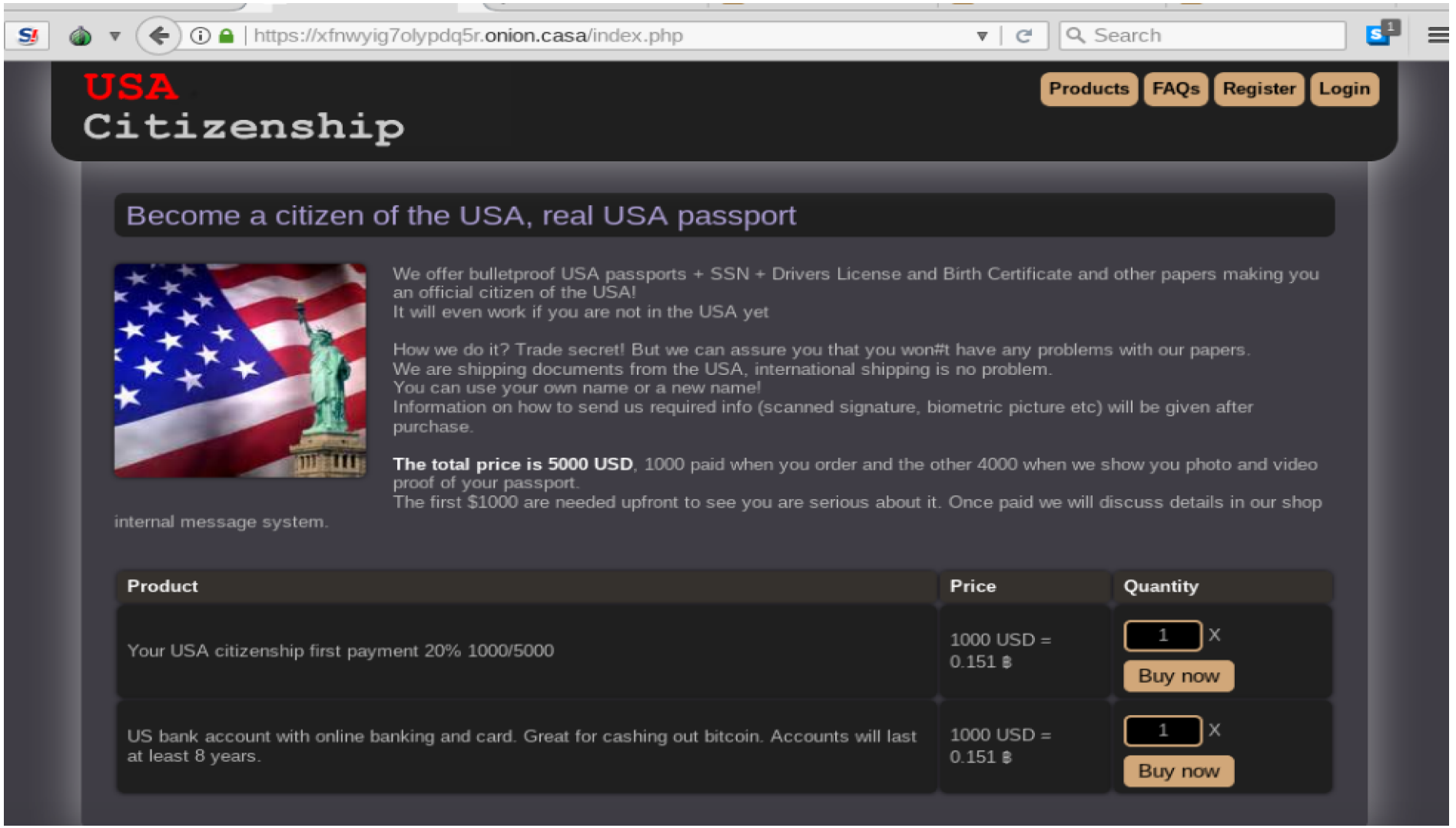
JASK

Looks a lot like the 90s internet





You can buy anything...



JASK

Some of the ‘noticeable’ items in the dark web

- Buy guns
- Buy stolen or fake identities (Including citizens
- Drugs like nowhere
- Piracy (All kinds of legal and illegal content)
- Assassination/Terrorism
- Sex offenders haven (trafficking, CP, etc)

Web based Hidden Services in February 2016[25][26]

Category	Percentage
Violence	0.3
Arms	0.8
Social	1.2
Hacking	1.8
Illegitimate pornography	2.3
Nexus	2.3
Extremism	2.7
Unknown	3.0
Other illicit	3.8
Finance	6.3
Drugs	8.1
Other	19.6
None	47.7

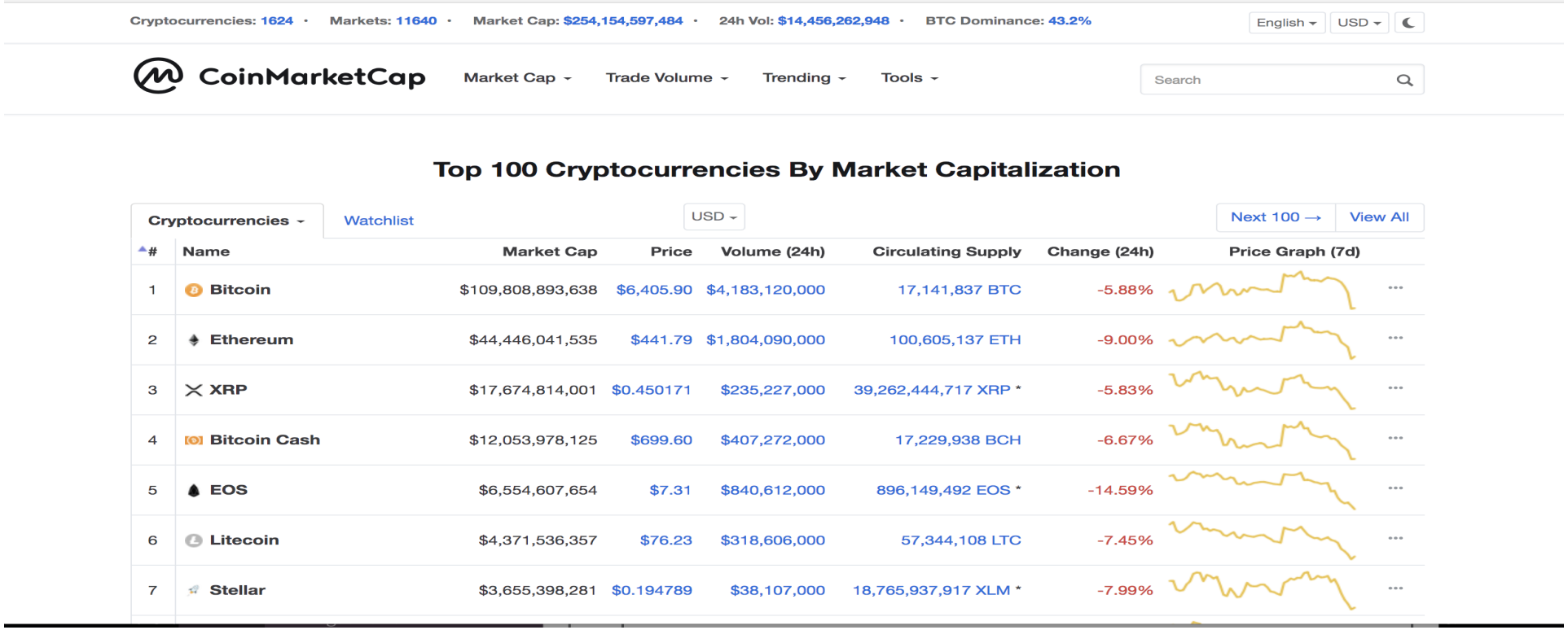
JASK

Cryptocurrency the big enabler

Most of the trading activities in the dark web are done using Cryptocurrency

“a digital currency in which encryption techniques are used to regulate the generation of units of currency and verify the transfer of funds, operating independently of a central bank.”

Cryptocurrency the big enabler



Cryptocurrencies: 1624 · Markets: 11640 · Market Cap: $254,154,597,484 · 24h Vol: $14,456,262,948 · BTC Dominance: 43.2%

English · USD

CoinMarketCap Market Cap · Trade Volume · Trending · Tools Search

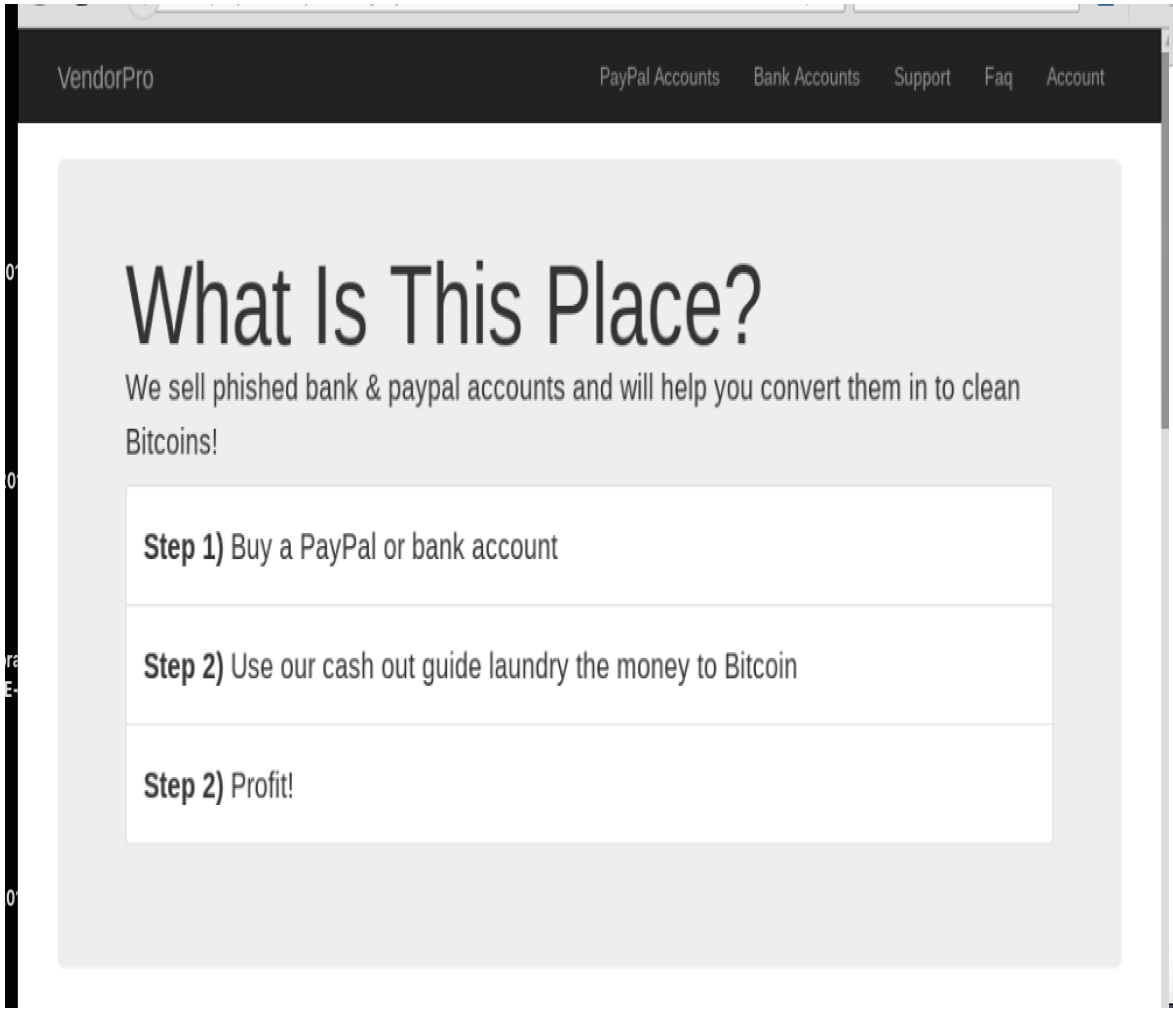
Top 100 Cryptocurrencies By Market Capitalization

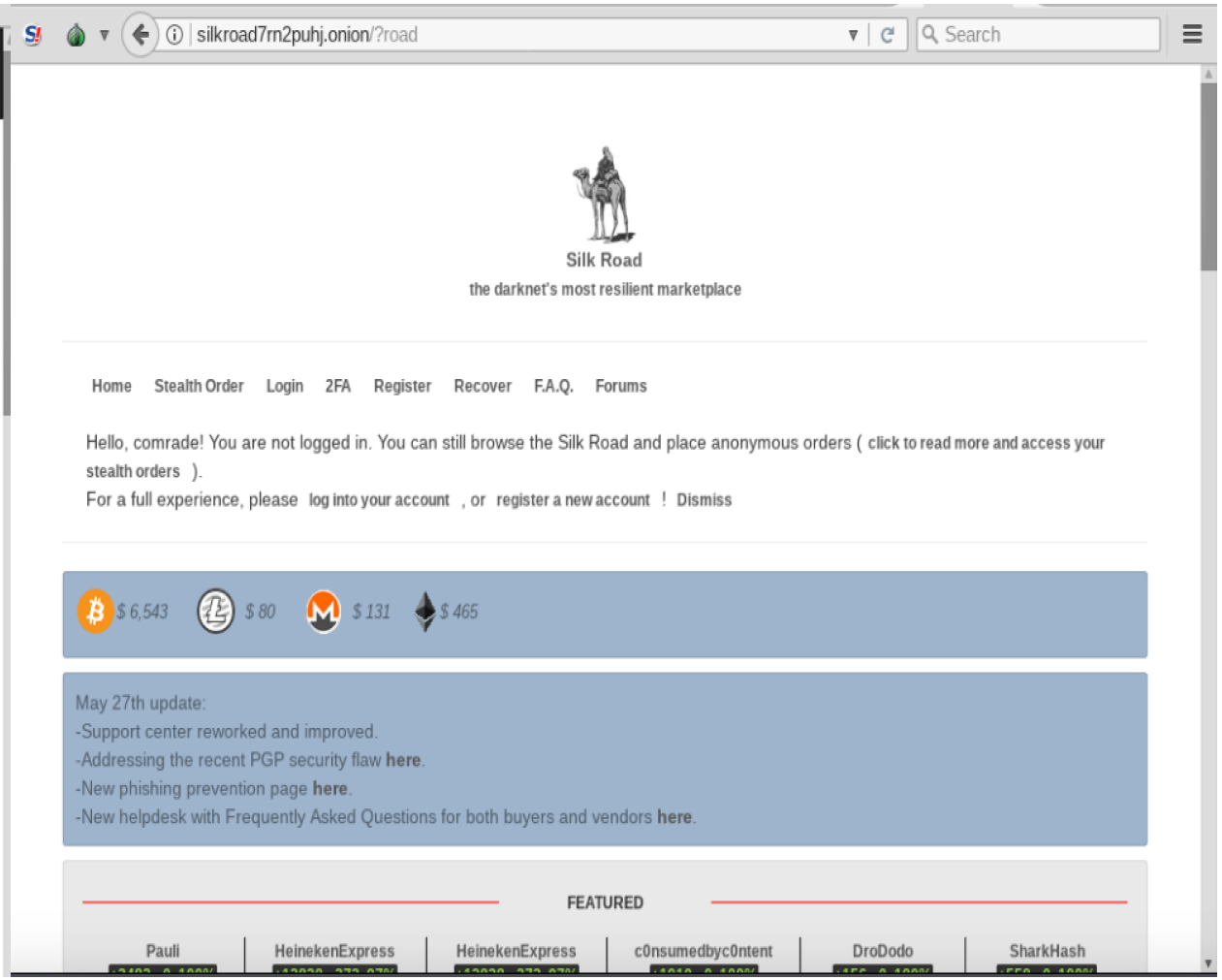
Cryptocurrencies · Watchlist · USD · Next 100 → · View All

#	Name	Market Cap	Price	Volume (24h)	Circulating Supply	Change (24h)	Price Graph (7d)
1	Bitcoin	$109,808,893,638	$6,405.90	$4,183,120,000	17,141,837 BTC	-5.88%	
2	Ethereum	$44,446,041,535	$441.79	$1,804,090,000	100,605,137 ETH	-9.00%	
3	XRP	$17,674,814,001	$0.450171	$235,227,000	39,262,444,717 XRP *	-5.83%	
4	Bitcoin Cash	$12,053,978,125	$699.60	$407,272,000	17,229,938 BCH	-6.67%	
5	EOS	$6,554,607,654	$7.31	$840,612,000	896,149,492 EOS *	-14.59%	
6	Litecoin	$4,371,536,357	$76.23	$318,606,000	57,344,108 LTC	-7.45%	
7	Stellar	$3,655,398,281	$0.194789	$38,107,000	18,765,937,917 XLM *	-7.99%	

JASK

Cryptocurrency the big enabler





JASK

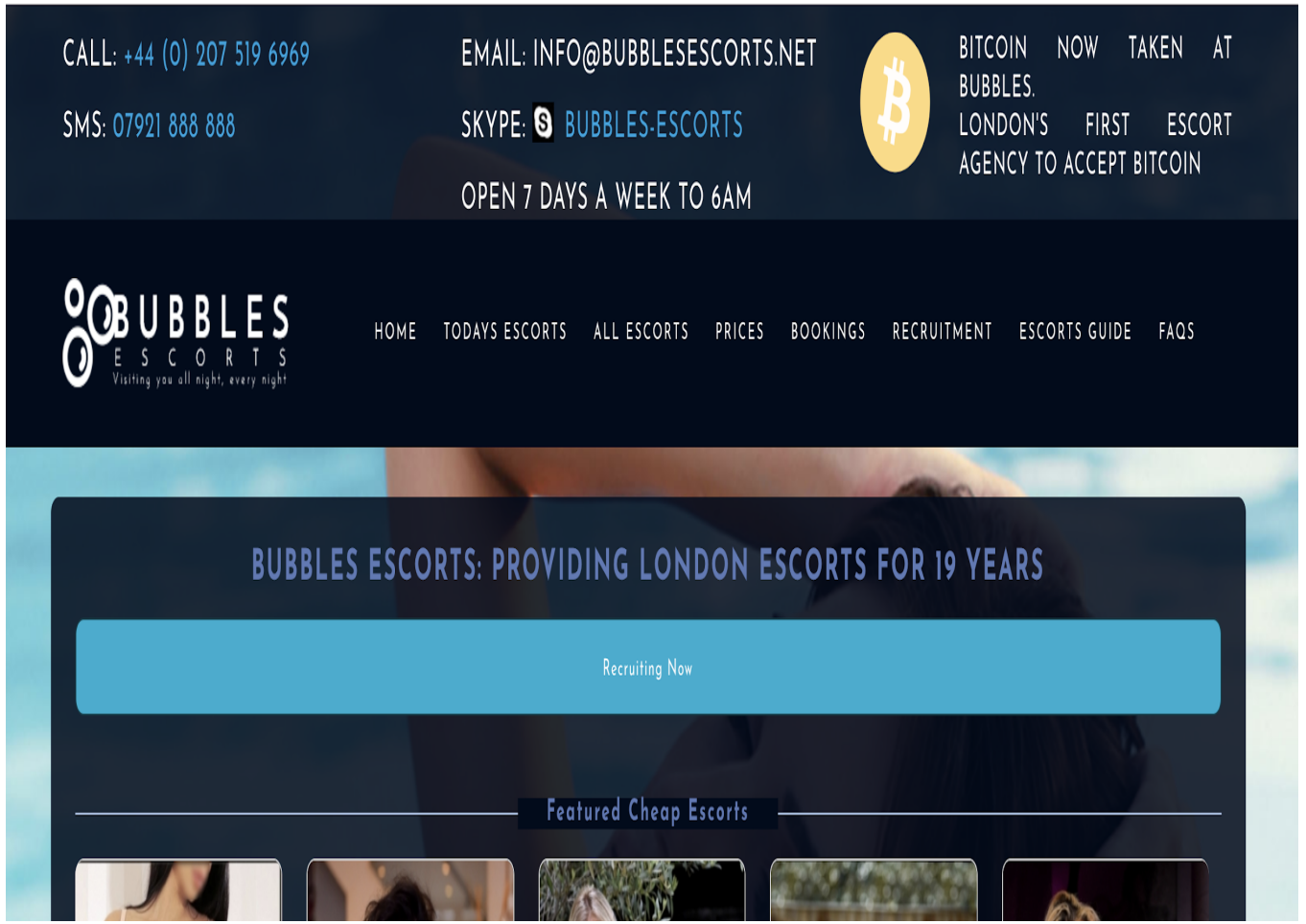
Cryptocurrency the big enabler

Why is cryptocurrency preferred?

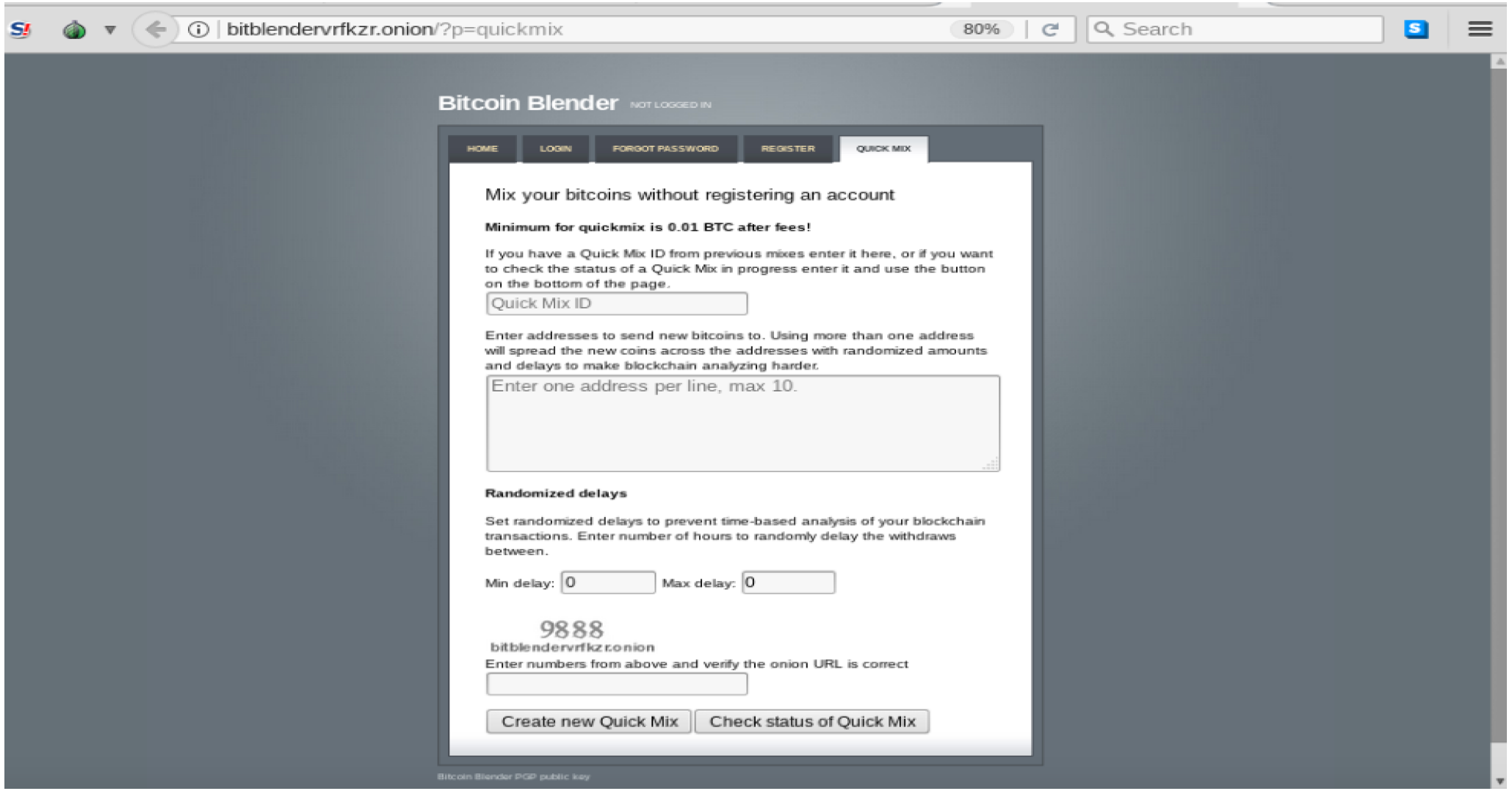
- Decentralized -- Not dependent on banks
- Anonymous -- A level of anonymity comfortable for criminals
- Not regulated -- No limits, formal tracking
- Accepted worldwide -- no borders, no oversight from gov or int org

Cryptocurrencies & Human trafficking

- Human traffickers using cryptocurrency to avoid prosecution
- Clearnet/Surface Web now showing crypto as payment form



Cryptocurrency big enabler



JASK

Human trafficking facts

- Estimated of 150 BILLION market per year *
- At any given time in 2016, an estimated 40.3 million people are in modern slavery, including 24.9 million in forced labor and 15.4 million in forced marriage.*
- Commercial sexual exploitation accounting for $99 billion of that total.*
- Traffickers are driven by the level of anonymity provided by dark web/darknets and cryptocurrency.
- What you see in Clearnet (Internet as we all know it) is tip of the iceberg, transactions and logistics mainly performed using dark web/crypto

2016 UNODC Global Report on Trafficking in Persons: Trafficking Crime Is Evolving

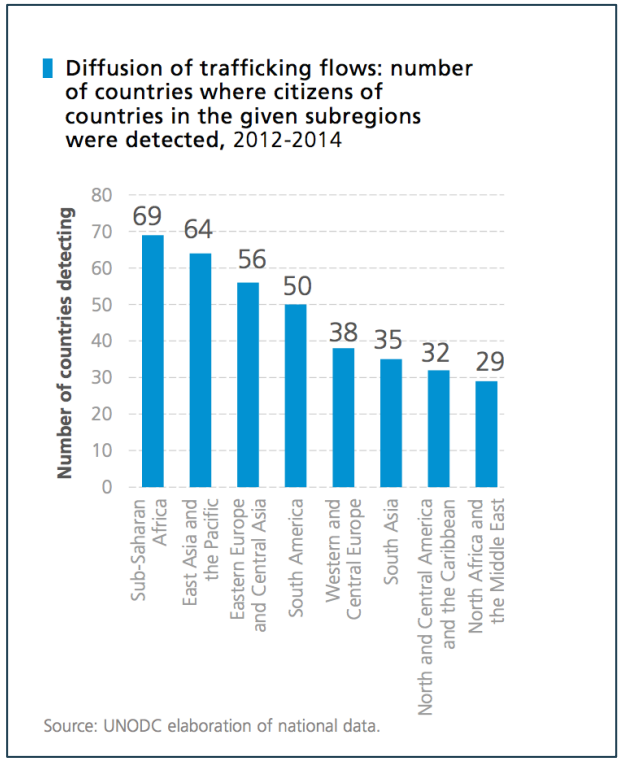
2) HOW HAS TRAFFICKING IN PERSONS CHANGED IN RECENT YEARS?

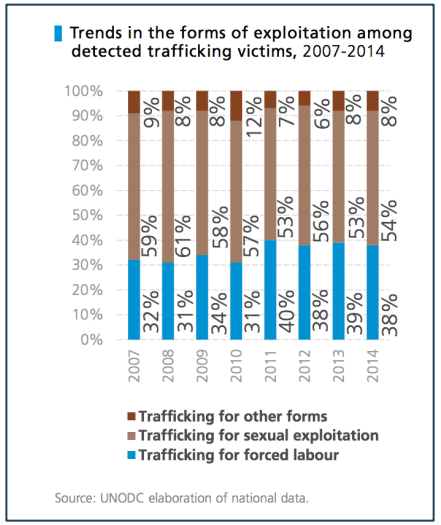
These shifts indicate that the common understanding of the trafficking crime has evolved. A decade ago, trafficking was thought to mainly involve women trafficked from afar into an affluent country for sexual exploitation. Today, criminal justice practitioners are more aware of the diversity among offenders, victims, forms of exploitation and flows of trafficking in persons, and the statistics may reflect this increased awareness.

7) TRAFFICKING THE MOST VULNERABLE: CHILDREN

The share of detected child victims has returned to levels last seen in 2009, after seven years of increases. Despite this trend, still more than a quarter of the detected trafficking victims in 2014 were children.

In Sub-Saharan Africa and Central America and the Caribbean, a majority of the detected victims are children. There are several reasons, such as demographics, socio-economic factors, legislative differences and countries' institutional frameworks and priorities. There seems to be a relation between a country's level of development and the age of detected trafficking victims. In the least developed countries, children often comprise large shares of the detected victims.



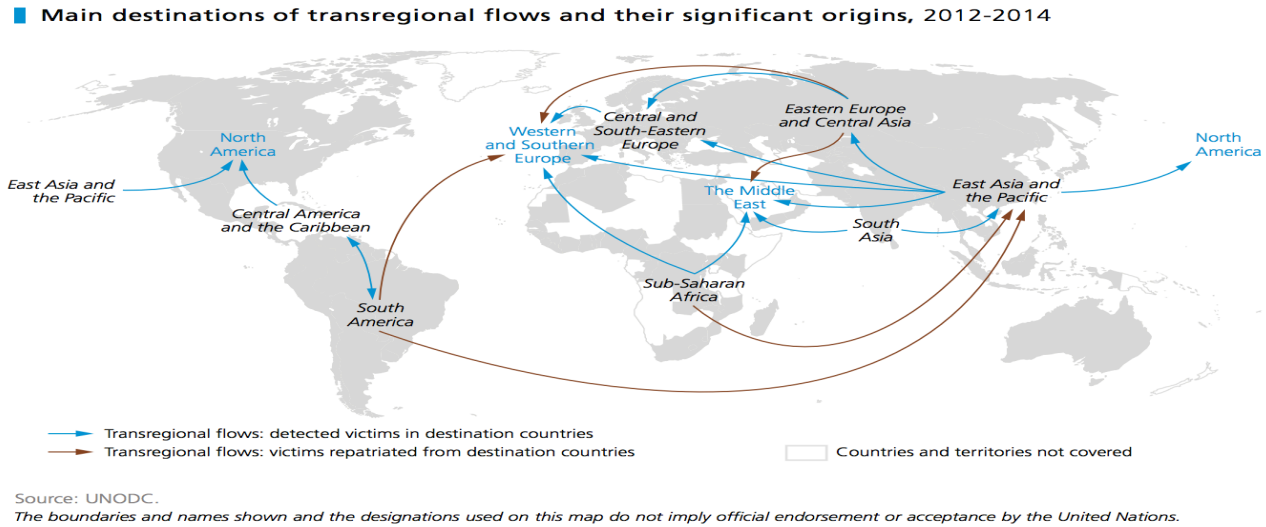


JASK

Tracking Cyber Crime Groups

As dark web provides an effective mean of coverage, the use of new techniques applied to research can provide advantages:

- Automated scraping, monitoring, indexing, & labeling of targeted data (Details next)
- Cryptocurrency tracking (I.E chain analysis)
- Use of machine learning and data driven methods to uncover patterns, establish entities and identify specific types of anonymized behavior



JASK

Data Driven Security Analytics on the Darkweb: Hackers Helping Save Lives

Facebook and Microsoft battle child porn

May 21, 2011

The Facebook website is displayed on a laptop computer on May 9, in San Anselmo, California. Facebook and Microsoft on Friday formally unveiled an alliance to ferret out child porn and those that share such images at the world's leading online social network.

Microsoft donates image-matching tools to fight child porn trafficking

Hacker group battles child porn

Ethical Hackers Against Pedophilia are using their high-tech cracking talents to identify people they say post child porn on the Net.

/ FEBRUARY 2, 1998 1:10 PM PST

Twitter to introduce PhotoDNA system to block child abuse images

Microsoft-developed system may be introduced this year once complication of handling pictures posted alongside billions of tweets can be overcome

Anti-human trafficking group uses data to track criminals

by Selena Larson @selenalarson

August 17, 2017: 3:33 PM ET

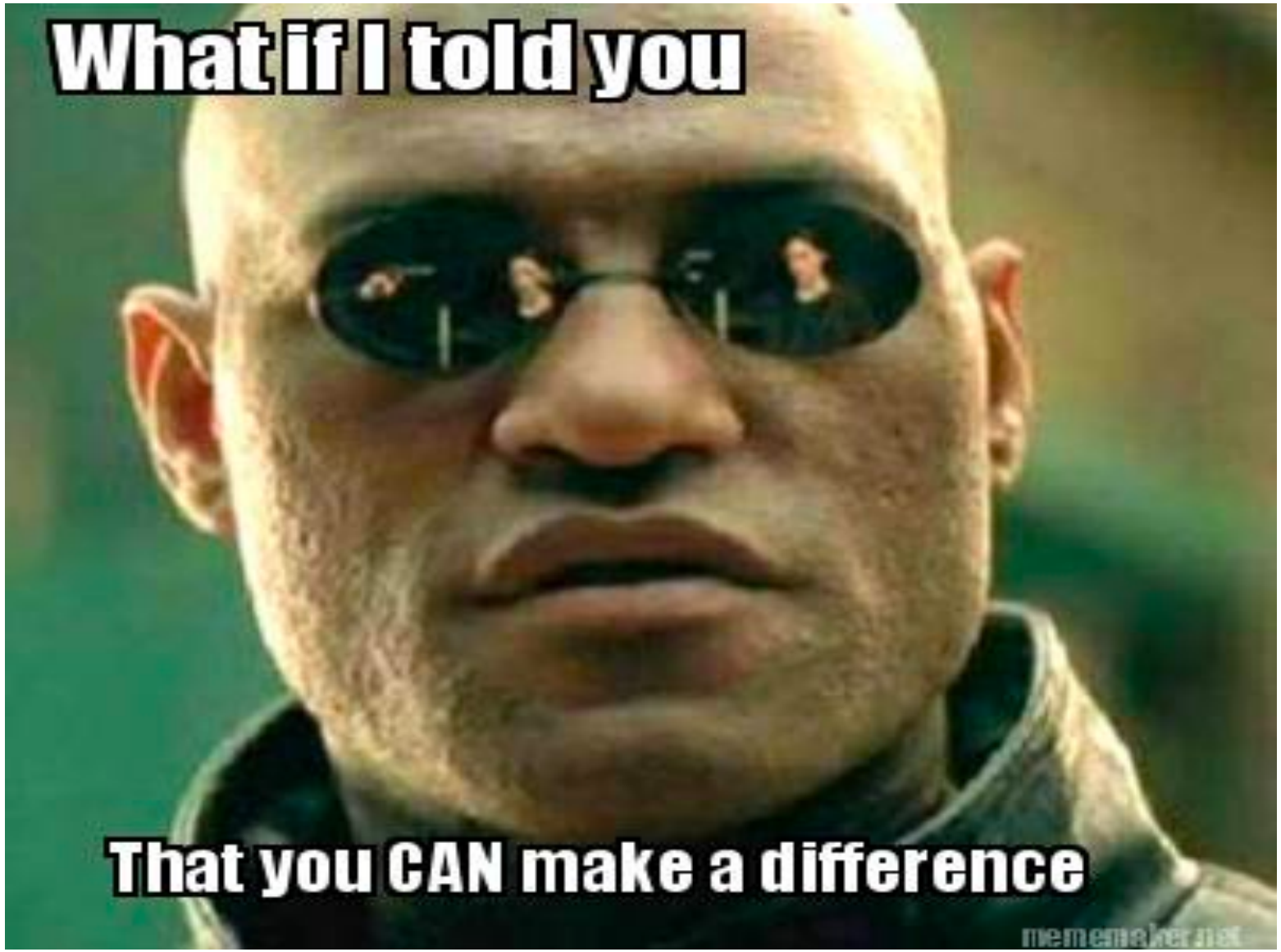
Recommend 0

How did we get involved in this type of research?

- Private/public sector collaboration at grass roots level
- Silicon valley partnerships and “data science for good” programs
- Child safety hackathons at Facebook every year and similar volunteer type engineering/research opportunities

Call to arms (In the spirit of BSides “I am The Calvary”)

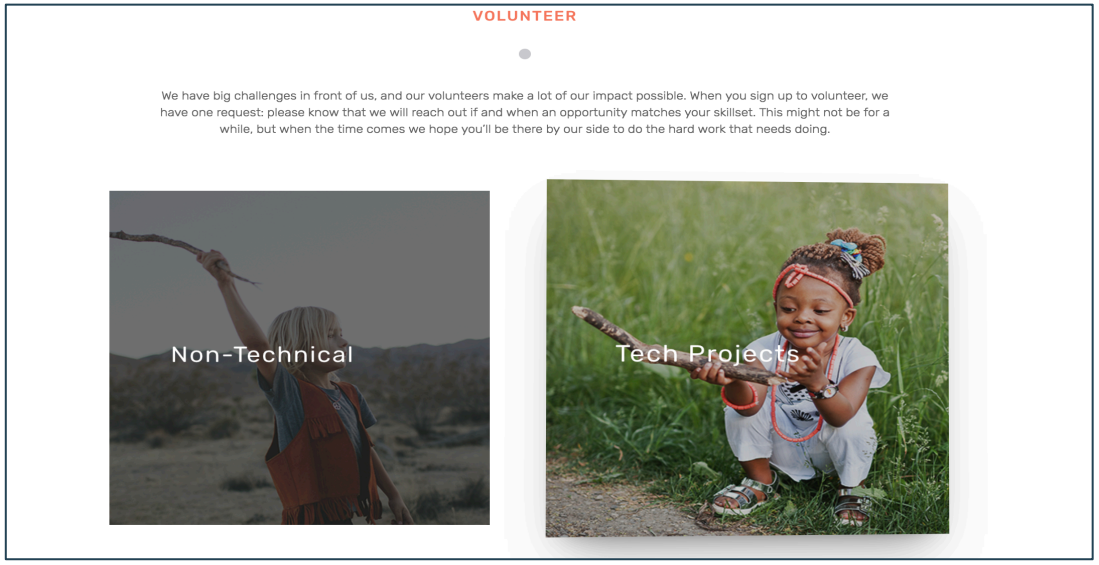
- Quote at this years BSides panel: “There are certain problems that industry /private sector wont work on and there are problems that the public sector cannot do” – there is a grey area where technical experts are needed to help hack and contribute to a global movement
- It starts at the grass roots level with volunteering and collaboration with NGOs, Legislative processes and industry partnerships
- What I like about the set of related problems: very hard computer science problems, security and encryption reverse engineering behavioral profiling and ***digital forensics***



JASK

We need hackers like you: THORN and Global Emancipation examples

- https://www.wearethorn.org/join-us/
- https://www.globalemancipation.ngo/engage-with-us/



Want to help in the fight against global human trafficking? Want to build world-class data analytics to solve one of the most important human rights problems?

The Global Emancipation Network is always looking for opportunities to connect with those who want to help combat human trafficking. Key areas in which our organization is always investing include:

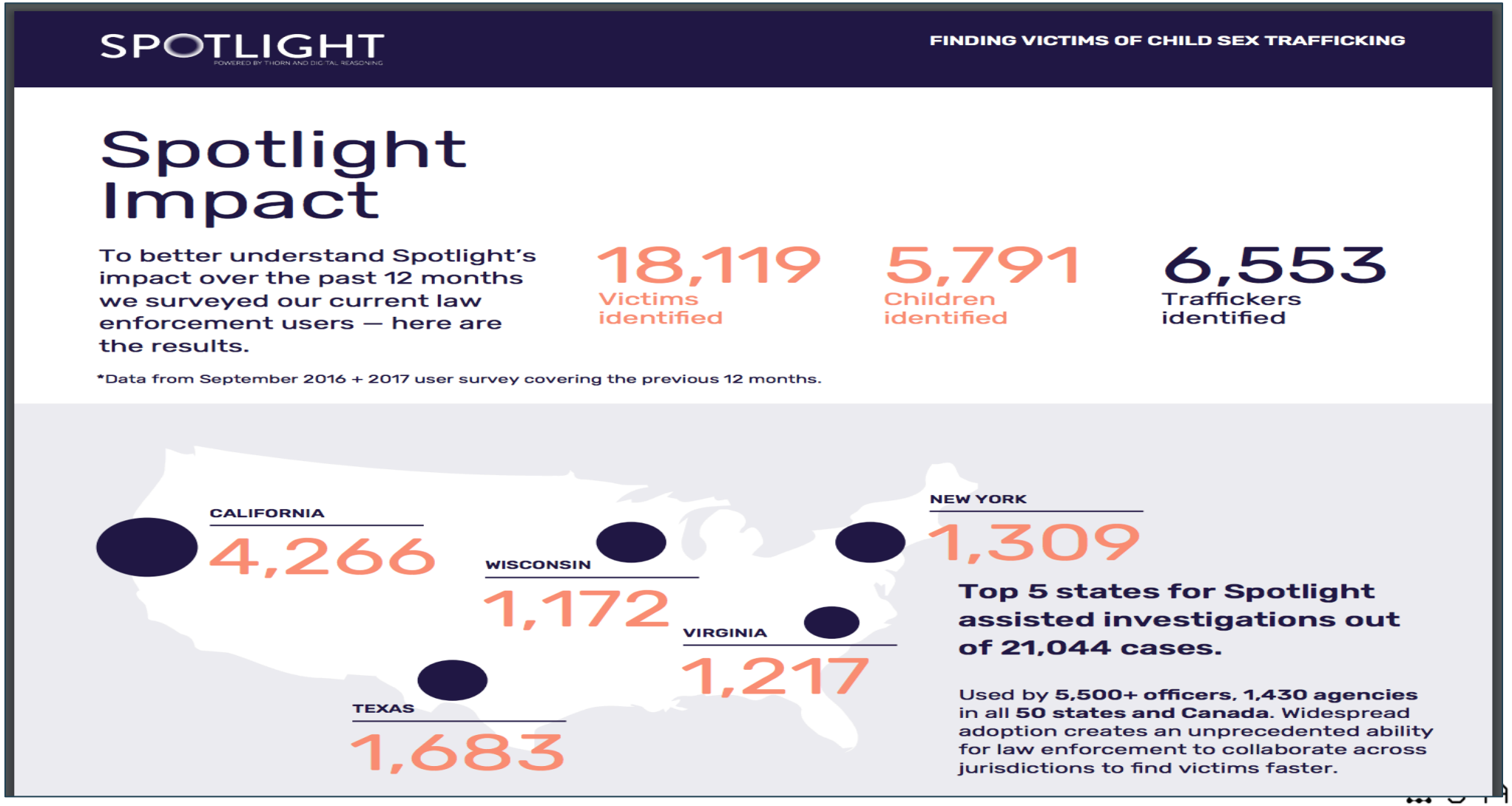
Partner organizations involved in the human trafficking problem such as policy analysts, legal services, rescue operations, shelters, medical aid units, law enforcement, military task forces, and more

Data scientists wishing to volunteer to construct analytics and work with big-data sets to identify traffickers, victims, and trafficking pipelines

Researchers wishing to volunteer on a variety of tasks involved in the analysis of trafficking

Technology providers interested in applying their products and services to the fight against human trafficking

Impact: Thorn and Global Emancipation Network have been involved in helping rescue multiple children In slave labor

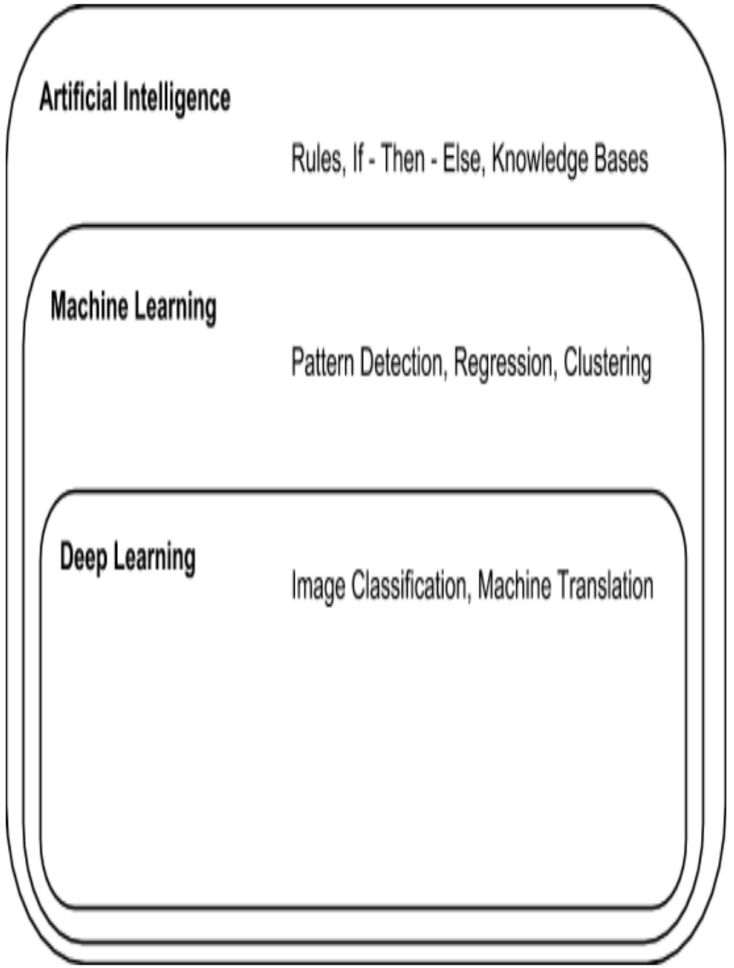


Challenging Research Problems

- There are so many open problems and ways hackers can help drive research some great examples are:
  - Global Emancipation Research: http://www.kdd.org/kdd2017/papers/view/backpage-and-bitcoin-uncovering-human-traffickers
  - https://github.com/EricSchles/traffickingGrab: A tool used to scrape popular websites for instances of human trafficking. Presently looks for sex trafficking.
  - https://code.google.com/archive/p/nudetech/: this tool can be integrated with various file and media sharing sites such as YouTube, Flickr, Rapidshare, Megaupload etc to prevent images/videos containing nudity or pornography from being uploaded to their system.
  - https://traffickcam.com/: Help fight trafficking by uploading photos of your hotel room. These photos will be used to determine where perpetrators of sex trafficking are committing their crimes.
- Other examples of a couple models we will talk about in next couple slides
  - Modeling populations of content producers vs content consumers in CP forums
  - Image correlation and similarity matching

Deep Neural Nets (DNN): Elevator Pitch

- Deep learning in some sense is synonymous with the state of the art in techniques for a variety of statistical learning problems (both supervised and unsupervised)
- One of the biggest ways deep learning has impacted my work in last couple years is DNN can be applied to raw data and the model itself will learn which features(stats/columns) are most important (Feature Learning)
    - https://github.com/jzadeh/aktaion: OSS Project we have three versions of V1 we built 100 features by hand to predict when a bro http.log has a payload in a sliding window. V2 we built another 50 or so features to add to the ones we have. V3 with a Convolution network or LSTM requires zero features
- Downside: Building DNN models are computationally expensive. Breakthroughs in training algorithms and the commoditization of compute has helped accelerate the ability to make the time to learn feasible for practical applications
- Interpreting the model is also challenging with neural nets



Key Ingredient for security use cases: operator feedback (labeling)

Domain Name	Total Count	Risk Factor	Exploit	Flow Pattern	UA Stats	URI Patterns
Yyfaimjmocdu.com	144	6.05	0	1	0	0
Jjeyd2u37an30.com	6192	5.05	0	1	0	0
Cdn4s.steelhousemedia.com	107	3	0	0	0	0
Log.tagecade.com	111	2	0	1	0	0
Go.vidprocess.com	170	2	0	0	0	0
Statse.webtrndslive.com	310	2	0	1	0	0
Cdn4s.steelhousemedia.com	107	1	0	0	0	0
Log.tagcade.com	111	1	0	1	0	0

Machine Learning: Value through operator feedback base case

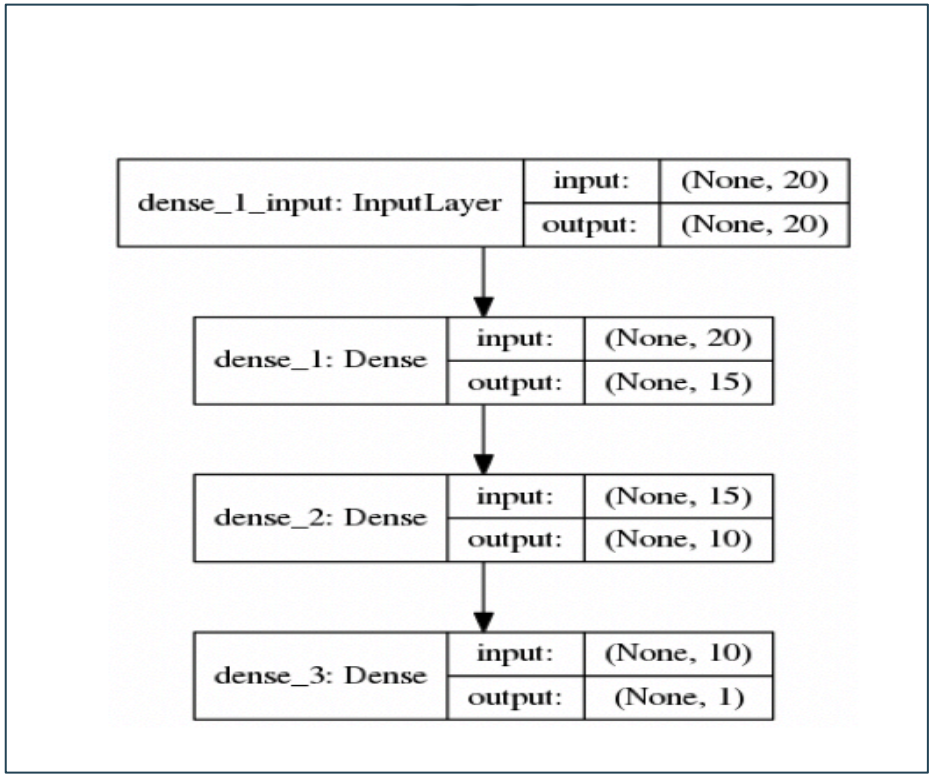
Domain Name	Total Count	Risk Factor	Exploit	Flow Pattern	UA Stats	URI Patterns	Outcome
Yyfaimjmocdu.com	144	6.05	0	1	0	0	Malicious
Jjeyd2u37an30.com	6192	5.05	0	1	0	0	Malicious
Cdn4s.steelhousemedia.com	107	3	0	0	0	0	Benign
Log.tagecade.com	111	2	0	1	0	0	Benign
Go.vidprocess.com	170	2	0	0	0	0	Benign
Statse.webtrndslive.com	310	2	0	1	0	0	Benign
Cdn4s.steelhousemedia.com	107	1	0	0	0	0	Benign
Log.tagcade.com	111	1	0	1	0	0	Benign

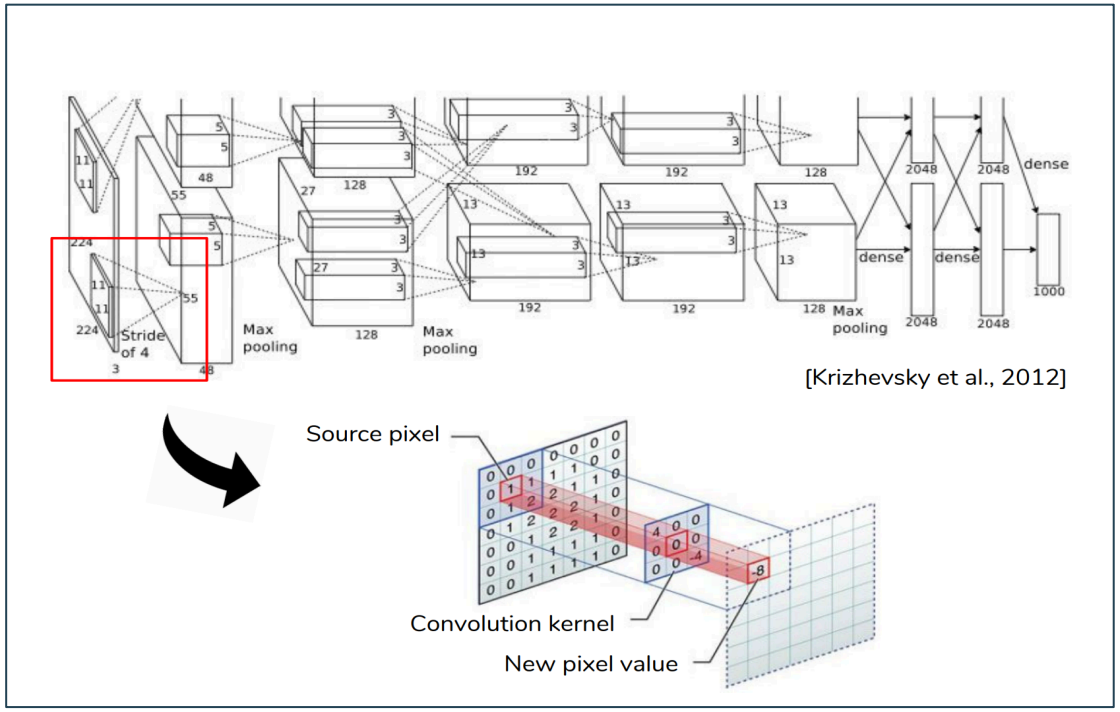
Human Expertise is manually encoded into a format computer understand: Sometimes this process is called Labeling or “Truth-ing” the data.

Manual Feature Engineering

```
package com.aktaion.ml.behaviors

import scala.util.matching.Regex

class ExploitationUriBehaviors extends MicroBehaviorSet {
  val uriMaxPathDepth = MicroBehaviorData("MaxPathDepth", "Maximum path length of all URI's observed")
  val uriMinPathDepth = MicroBehaviorData("MinPathDepth", "Minimum path length of all URI's observed")
  val uriMaxLength = MicroBehaviorData("MaxUriLength", "Maximum length of all URI's observed")
  val uriMinLength = MicroBehaviorData("MinUriLength", "Minimum length of all URI's observed")
  val uriDistinct = MicroBehaviorData("UniqNumberOfUri", "Unique count of URI's in window")
  val uriMaxEntropy = MicroBehaviorData("UriMaxEntropy", "Maximum entropy of all URI's observed")
  val uriMinEntropy = MicroBehaviorData("UriMinEntropy", "Minimum entropy of all URI's observed")
  val base64Match = MicroBehaviorData("UriBase64", "URI's obseved contain large number of base 64 strings")
  val percentEncondingMatch = MicroBehaviorData("UriBase64", "URI's obseved contain large number of percent encoded strings")

  val behaviorVector = List(uriMaxPathDepth,
    uriMinPathDepth,uriMaxLength,
    uriMinLength,uriDistinct,
    uriMaxEntropy,uriMinEntropy)

  val encodingBase64 = new Regex( """^(?:[A-Za-z0-9+/]{4})*(?:[A-Za-z0-9+/]{2}==|[A-Za-z0-9+/]{3}=)?$""")
  val encodingBase64modified = new Regex( """^([A-Za-z0-9+/]{4})*([A-Za-z0-9+/]{4}|[A-Za-z0-9+/]{3}=|[A-Za-z0-9+/]{2}==)$""")
  val encodingPercentSimple = new Regex( """%[A-Fa-f0-9]{2}""")
}

class ExploitationTimingBehaviors extends MicroBehaviorSet {
  val maxTimeIntervalA = MicroBehaviorData("MaxTimeIntervalA", "Difference in timestamp between event 1 and event 2")
  val maxTimeIntervalB = MicroBehaviorData("MaxTimeIntervalB", "Difference in timestamp between event 2 and event 3")
  val maxTimeIntervalC = MicroBehaviorData("MaxTimeIntervalC", "Difference in timestamp between event 3 and event 4")
  val maxTimeIntervalD = MicroBehaviorData("MaxTimeIntervalD", "Difference in timestamp between event 4 and event 5")
```

Modeling in the era of deep learning: the magic is in the architecture and the algorithms that update the weights in each part of the network





Good references for the topic at hand mixed with the practical advice on ML for security

- Dean Teffer, JASK Principal Scientists blog on the difference between AI, ML and Deep Learning: https://jask.com/artificial-intelligence-vs-machine-learning/
- https://jask.com/applied-machine-learning-in-security-part-2/
- https://clarifai.com/blog/what-convolutional-neural-networks-see-at-when-they-see-nudity/
- Great deck with lots of applications of deep learning relevant to forensics: http://www.ic.unicamp.br/~ariadne/MC039/1s2017/Sandra-Avila.pdf
- https://www.wearethorn.org/blog/eliminating-child-sexual-abuse-material-hash-values/
- https://www.wearethorn.org/blog/hashing-detect-child-sex-abuse-imagery/

Stories from the trenches: Modeling Darkweb “forum” interactions

Data set: DarkOwl for querying Darknet scrapes (500,000,000 darkweb pages indexed) and Thorn collaboration with law enforcement

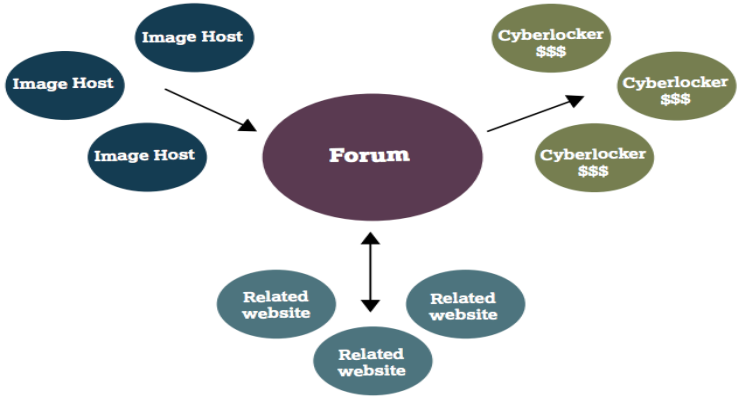
Key statistical assumption: Forum interactions /child abuse are tightly linked to specific subgroups of the population. There is a very small number of active users who are creating/posting the majority of the content

Modeling method core concept: Latent Dirichlet Allocation (NLP) + PageRank (Graph Algorithms) + Object Detection (DNN)

Graph of the population is made based on @mentions and similar type tweets on darknet sites suspect of sharing CP and correlated with metadata and telemetry (e.g darkowl data set)

Distribution Methods

Of the 1,765 images being distributed via Image Hosts in this Study, 1,526 (86%) had been individually embedded into 16 forums dedicated to distribution of captures of live-streamed child sexual abuse. As the following diagram shows, these forums are at the centre of distribution networks for captures of live-streamed child sexual abuse.



Source: 2018 Trends in Online Child Sexual Exploitation: Examining the Distribution of Captures of Live-streamed Child Sexual Abuse

JASK

Stories from the trenches: Modeling Darkweb “forum” interactions

Algorithm:

Step 1: Build the graph. Process all data into sets of users (nodes) and forum posts. If another user is mentioned in a post draw an edge between those two users.

Step 2: NLP Tools. Topic model. Aggregate all posts for a single user and call this a “document.” (lots of preprocessing tricks to reduce the cardinality of the vocabulary here as well as build NLP features computed for upstream correlations). For each document we run LDA to cluster users into similar groups (“Topics”)

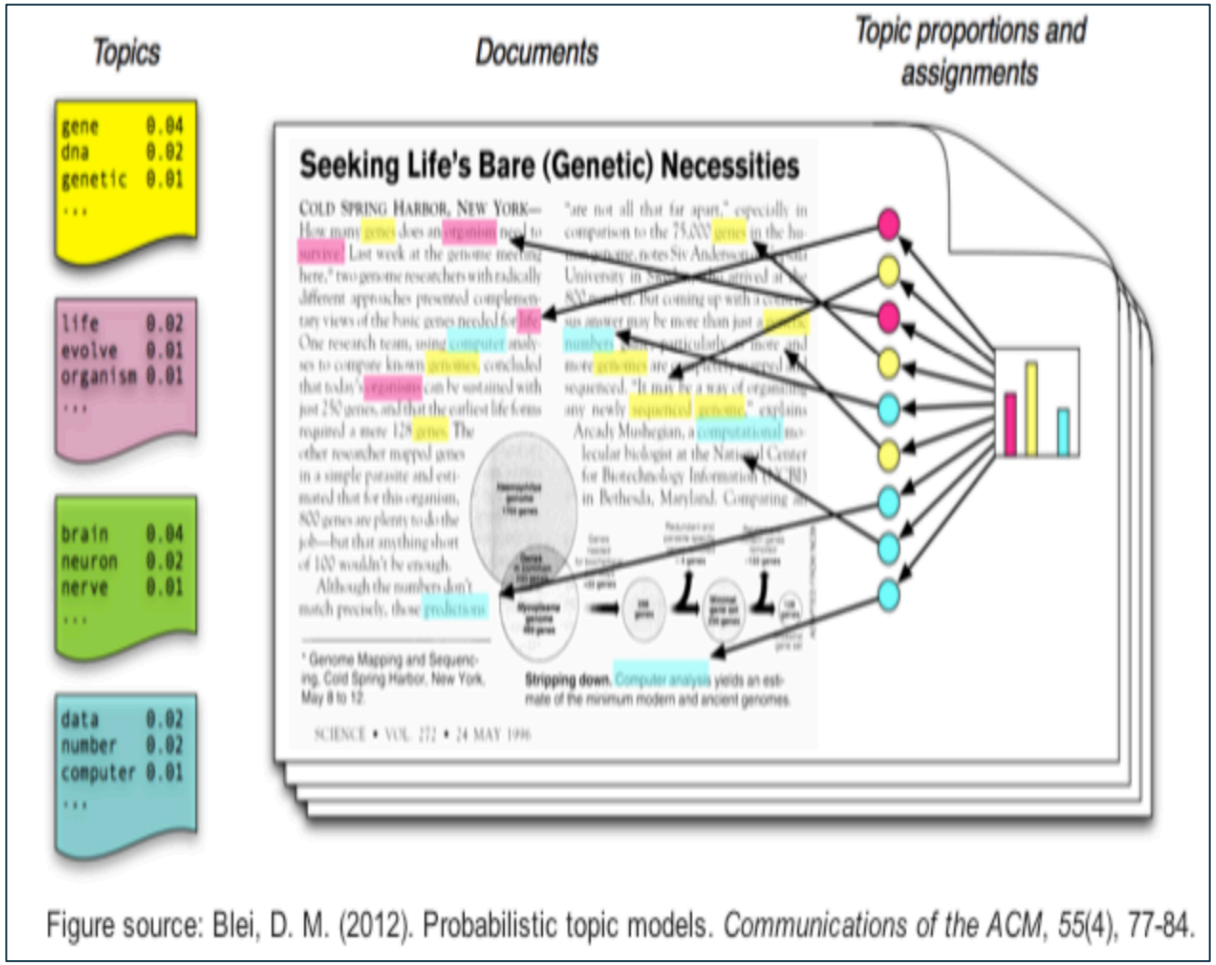


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

JASK

Stories from the trenches: Modeling Darkweb “forum” interactions

Algorithm:

Step 3: Use PageRank to assign popularity scores to more influential users and correlate the

Step 4: Data reduction and correlation. Assign labels /rank to clusters of users that upload most content and are highly active/ interconnected

Step 5: Time to hunt for OpSec fails in coms (NLP) and image object recognition (DNN) to help triangulate with local law possible human trafficking victims in posted content

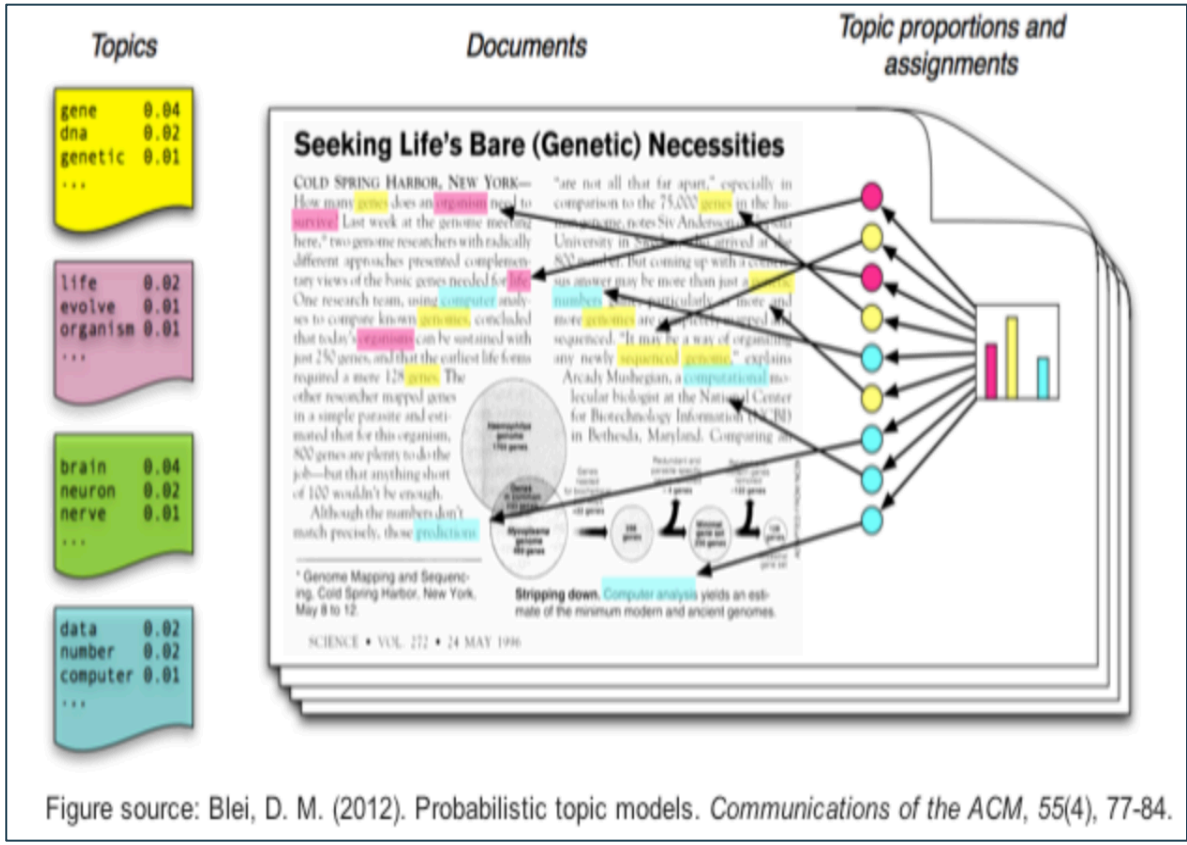
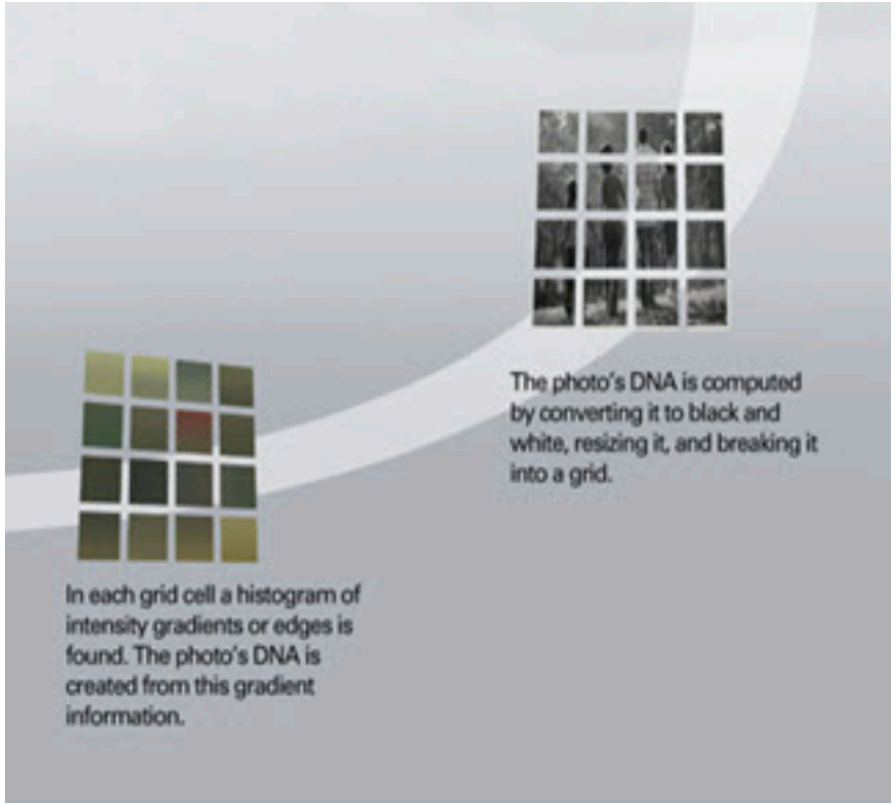


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

JASK

Stories from the trenches: Correlating images to missing persons reports

- Of the nearly 25,000 runaways reported to the National Center for Missing and Exploited Children (NCMEC) in 2017, one in seven were likely victims of child sex trafficking. And that's just in the US. Around the world, even with under reporting, the numbers are equally staggering.
- POC originally built by the MS team at Facebook hackathon tested the idea on 5000 missing persons photos in California law unsolved cases: The pictures were correlated to scrapes of escort sites (huge asymptotic complexity here)
- During the 24 hour hackathon a missing woman was identified as correlating to an escort post despite photos being taken a few years apart and the proper evidence was given to law enforcement for cross border collaboration

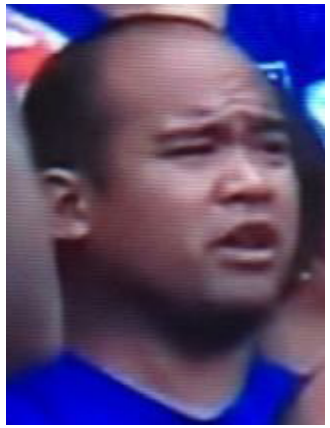


PhotoDNA

PhotoDNA provides a way to create a unique signature - similar to a finger print - from a photo that will remain consistent even after it is edited or manipulated.







JASK

Stories from the trenches: Correlating Images/Text To Content Producers And OpSec Fails

- Major algorithmic challenge in image similarity for forensics workflows is searching billions of pictures to match one: its an interesting problem that intersects classic search bounds with how lightweight our hashing method is
- Relevant ideas/literature
  - Classic Reference: Fleck M.M., Forsyth D.A., Bregler C. (1996) Finding naked people.
  - PhotoDNA: https://www.microsoft.com/en-us/photodna
  - H. Farid. Reining in Online Abuses. Technology and Innovation, 2018.
  - "Small Codes and Large Image Databases for Recognition" A. Torralba, R. Fergus and Y. Weiss
  - Hashing by Deep Learning Wei Liu IBM T. J. Watson Research Center Yorktown Heights, New York 10598, USA
  - A. Torralba, R. Fergus and W. T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition,"
  - A. Torralba, R. Fergus and Y. Weiss, "Small codes and large image databases for recognition,"

Table 1: The characteristics of eight recently proposed deep learning based hashing methods.

Deep learning based hashing methods	Data domain	Unsupervised or supervised?	Learning features?	Hierarchy of deep neural networks
Semantic Hashing [12]	text	unsupervised	no	4
Restricted Boltzmann Machine [13]	text and image	supervised	no	4 and 5
Tailored Feed-Forward Neural Network [10]	text and image	supervised	no	6
Deep Hashing [8]	image	unsupervised	no	3
Supervised Deep Hashing [8]	image	supervised	no	3
Convolutional Neural Network Hashing [14]	image	supervised	yes	5
Deep Semantic Ranking Hashing [15]	image	supervised	yes	8
Deep Neural Network Hashing [6]	image	supervised	yes	10

Conclusions

- Human trafficking is a global problem and can use all the help from smart people volunteering and working on projects
- The Darkweb is one place where statistical techniques help de-obfuscate and model targets of interests behavioral profile. The darkweb is but a small part of the bigger picture though (plenty of use cases exists just scrapping the regular internet for instance)
- There are fascinating intersections of research that touch security, algorithms, geo politics and forensics (some of the problems are very hard)
- ML/AI are buzzwords that are useful sometimes in tactical applications when we know the right tool for the job. DNN represents (often times) the state of the art in terms of prediction power

Q&A

- Nostalgic shout-outs to a throwback era: Genocide2600 & EHAP (Ethical Hackers Against Pedophilia)
  - 1998: “Hacker group battles child porn” https://www.cnet.com/news/hacker-group-battles-child-porn/
  - https://en.wikipedia.org/wiki/Genocide2600