Biting into the Jawbreaker: Pushing the Boundaries of Threat Hunting Automation (#hamm)

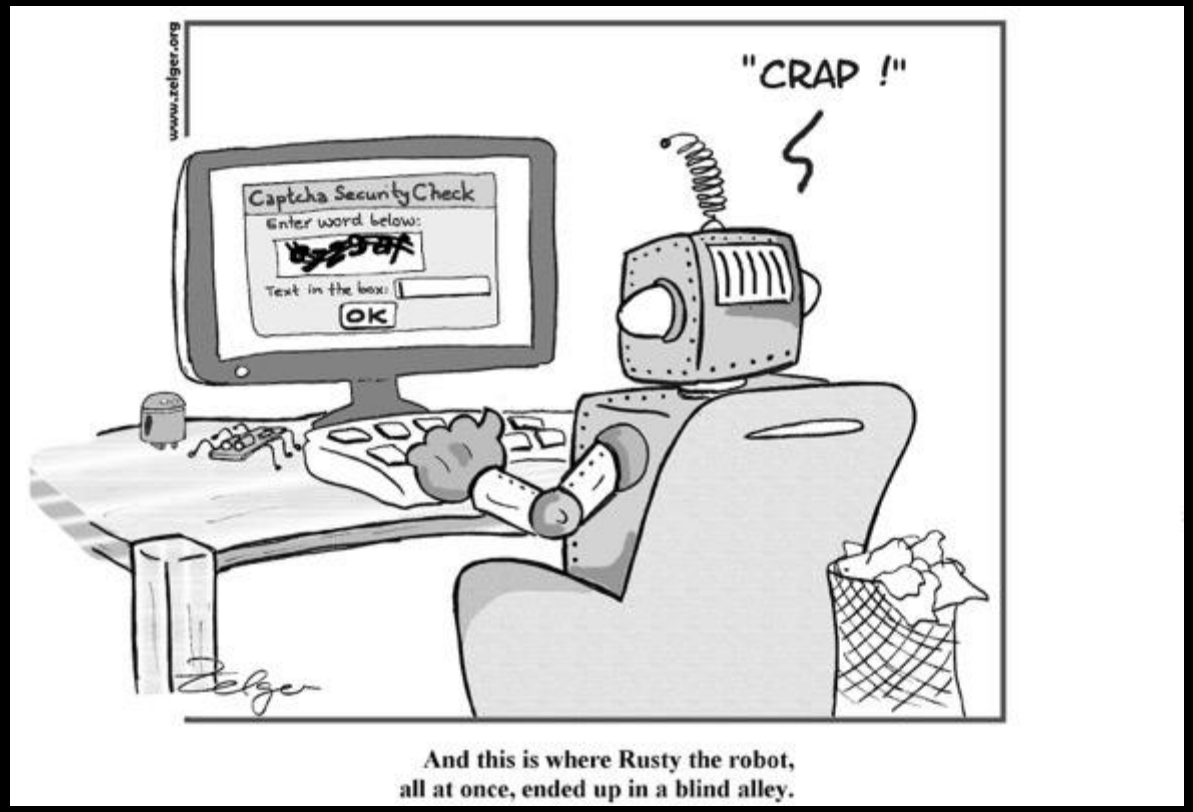
Alex Pinto - Chief Data Scientist – Niddel
@alexcpsec
@NiddelCorp

Agenda

- Who am I?
- Why does this talk exist?

- The Automation Barrier
- The Context Barrier
- The Experience Barrier
- The Creativity Barrier

- Hunting Automation Maturity Model

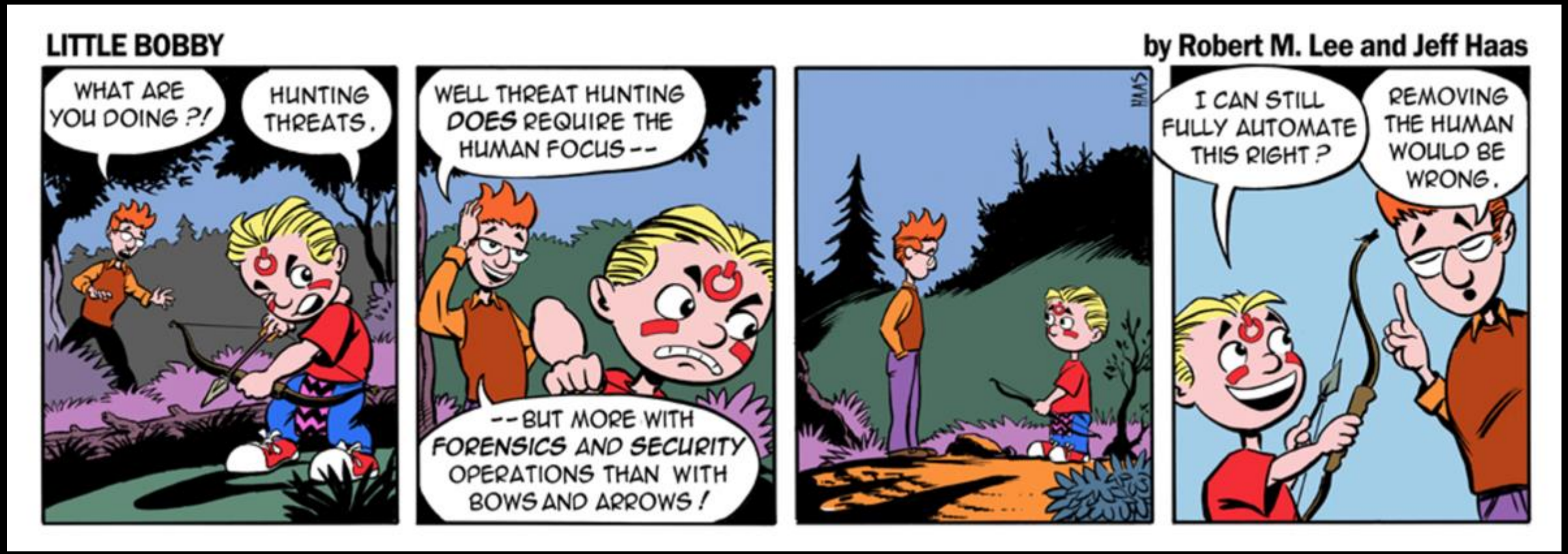


Who am I?

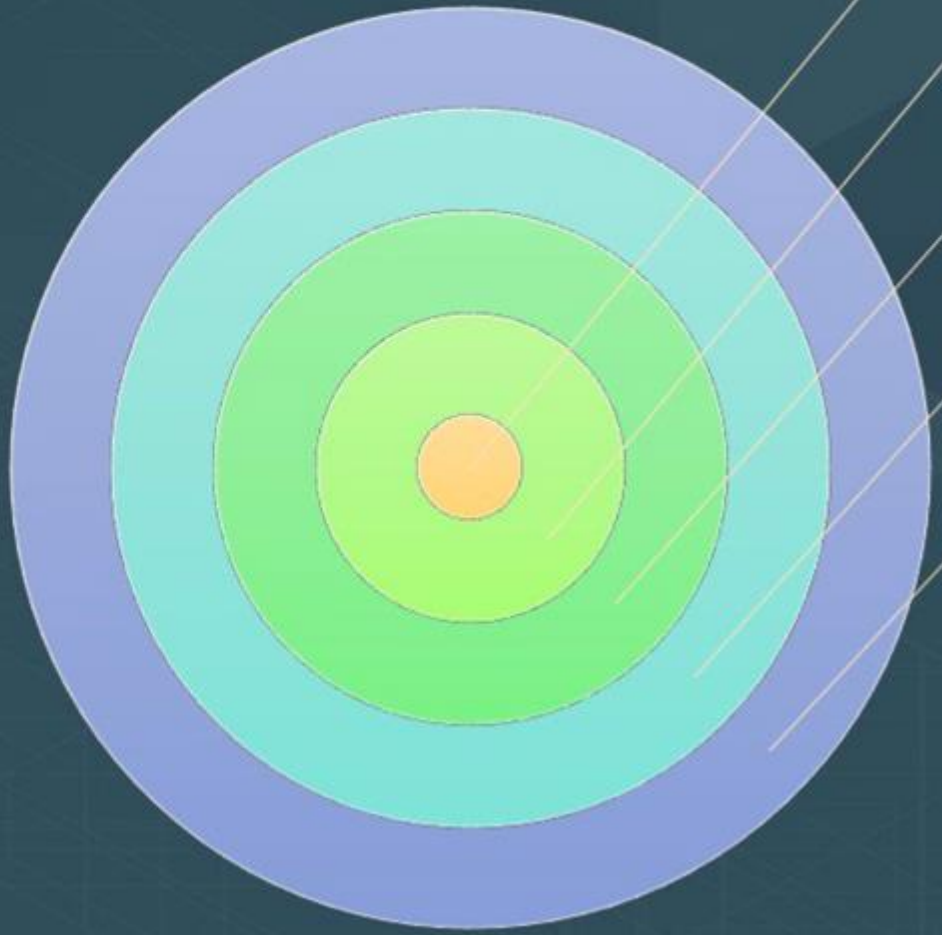
- Brazilian Immigrant
- Security Data Scientist
- Capybara Enthusiast
- Co-Founder at Niddel (@NiddelCorp)
- Founder of MLSec Project (@MLSecProject)
- What is **MLSec Project**? - Community of like-minded infosec professionals working to improve data science and machine learning application in security.
- What is **Niddel**? – Niddel is a security vendor that provides a SaaS-based Autonomous Threat Hunting System

Why does this talk exist?

Like any good story, it all started with a discussion on the Internet



The Simple Truths of Threat Hunting
"Threat Hunting Jawbreaker"



Threat Hunting requires the focus to be on the people

Your job focuses on human adversaries but you may not encounter adversaries

You need the open-mindedness of a new person but hunting is not for new folks

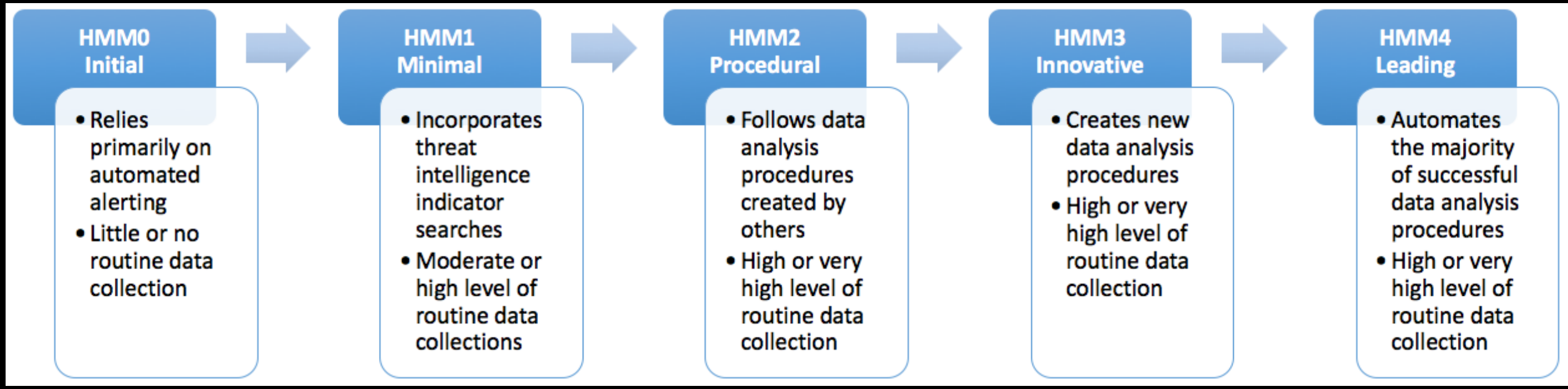
Product vendors will pitch hunting but it's not about the product

You need to rely on automation but you can't fully automate hunting

Slide shamelessly lifted from one of @RobertMLee 's presentations. Accept no imitations.

David Bianco to the Rescue!

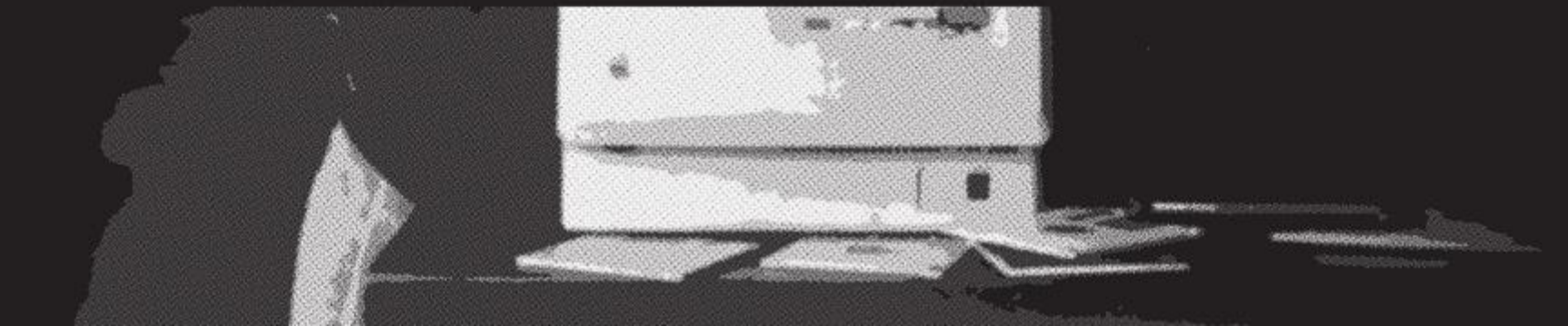
Why not describe hunting automation as a maturity model?



[This is my first presentation without citing the PoP in 3 years]

"Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom."

– Cliff Stoll



The Automation Barrier

Breaking the Automation Barrier

- When 9 of 10 of you think of automation, you think of this.
  - File hashes, YARA Rules, IP addresses, domain names

- Lowest possible bar for a vendor to claim they automate threat hunting
- Batch analysis / ”Retro-hunting”

Choosing Indicators – RIG EK

Active actor registering domains - NOT Domain Shadowing

```
> aa_rig_domain3 <- aa_rig_domain[whois.registrant.email == "a.miroshichenko@yandex.ru"]
> aa_rig_domain3[, list(domain, authority, SOA.host, whois.registrar, whois.registrant, whois.registrant.email)]
                 domain            authority       SOA.host                          whois.registrar whois.registrant    whois.registrant.email
 1:     b594e.s6h3eq.top          s6h3eq.top dns4.regway.com                                PDR LTD             N/A a.miroshichenko@yandex.ru
 2:  downtoughrodidn.xyz  downtoughrodidn.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 3: hedthindownhepro.xyz hedthindownhepro.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 4:     lettanarloro.xyz     lettanarloro.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 5: nothadheeventreb.xyz nothadheeventreb.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 6:    onemaharranse.xyz    onemaharranse.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 7:      othecledewi.xyz      othecledewi.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 8:     perbetrededu.xyz     perbetrededu.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
 9: peugheckbutaning.xyz peugheckbutaning.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
10:   r8406p7.nnm8s2.top          nnm8s2.top dns4.regway.com                                PDR LTD             N/A a.miroshichenko@yandex.ru
11: rinterthersparci.xyz rinterthersparci.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
12:   rolundilittrat.xyz   rolundilittrat.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
13:    ue5ow.otcrhoa.top          otcrhoa.top dns4.regway.com                                PDR LTD             N/A a.miroshichenko@yandex.ru
14:  useressitharrec.xyz  useressitharrec.xyz dns4.regway.com PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM             N/A a.miroshichenko@yandex.ru
15:    x134bi.eowjl2.top          eowjl2.top dns4.regway.com                                PDR LTD             N/A a.miroshichenko@yandex.ru
```

Email a.miroshichenko@yandex.ru is associated with ~234 domains ← Yay! Let's go block this!!

Choosing Indicators – RIG EK

```
> bb_rig_ns <- bb_rig[order(bb_rig$num.hits, decreasing = TRUE)][bucket.type == "whois.registrant.email"][1:10]
> bb_rig_ns
               bucket.type                bucket.value entity.category num.hits
 1: whois.registrant.email       stiviemalone@gmail.com             rig     1353
 2: whois.registrant.email      stivie.malone@gmail.com             rig      734
 3: whois.registrant.email        sizilsanksi@yahoo.com             rig       62
 4: whois.registrant.email   steelehendershot@gmail.com             rig       38
 5: whois.registrant.email               me@robcross.me             rig       33
 6: whois.registrant.email        ilan6741042@gmail.com             rig       22
 7: whois.registrant.email    a.miroshichenko@yandex.ru             rig       15
 8: whois.registrant.email bluestonewebdesigns@gmail.com             rig       14
 9: whois.registrant.email   roger@ticktockholdings.com             rig       14
10: whois.registrant.email      spiros@wsigodigital.com             rig       11
```

```
> bb_asn_rig_ip <-  bb_rig_ip[order(bb_rig_ip$num.hits, decreasing = TRUE)][bucket.type == "asnumber"][1:10]
> bb_asn_rig_ip
    bucket.type bucket.value entity.category num.hits
 1:    asnumber        16276             rig     1669
 2:    asnumber        14576             rig      513
 3:    asnumber        48096             rig      103
 4:    asnumber        48347             rig       91
 5:    asnumber        35415             rig       86
 6:    asnumber       394279             rig       58
 7:    asnumber        60117             rig       28
 8:    asnumber        21100             rig       21
 9:    asnumber        19318             rig       20
10:    asnumber        46606             rig       13
```

AS16276 – OVH SAS ☹ (maybe block?)

AS14576 – Hosting Solution Ltd (actually king-servers.com)

AS48096 – ITGRAD (any Russian offices?)

Choosing Indicators – Context Matters

```
> bb_tld <-  bb_rig[order(bb_rig$num.hits, decreasing = TRUE)][bucket.type == "domain_0"][1:10]
> bb_tld
    bucket.type bucket.value entity.category num.hits
 1:    domain_0          com             rig     1919
 2:    domain_0         info             rig      352
 3:    domain_0          org             rig      181
 4:    domain_0          top             rig       49
 5:    domain_0          net             rig       31
 6:    domain_0           pw             rig       15
 7:    domain_0          xyz             rig       12
 8:    domain_0           co             rig       10
 9:    domain_0         mobi             rig        8
10:    domain_0           cf             rig        4
```

← Can't block this one, lol

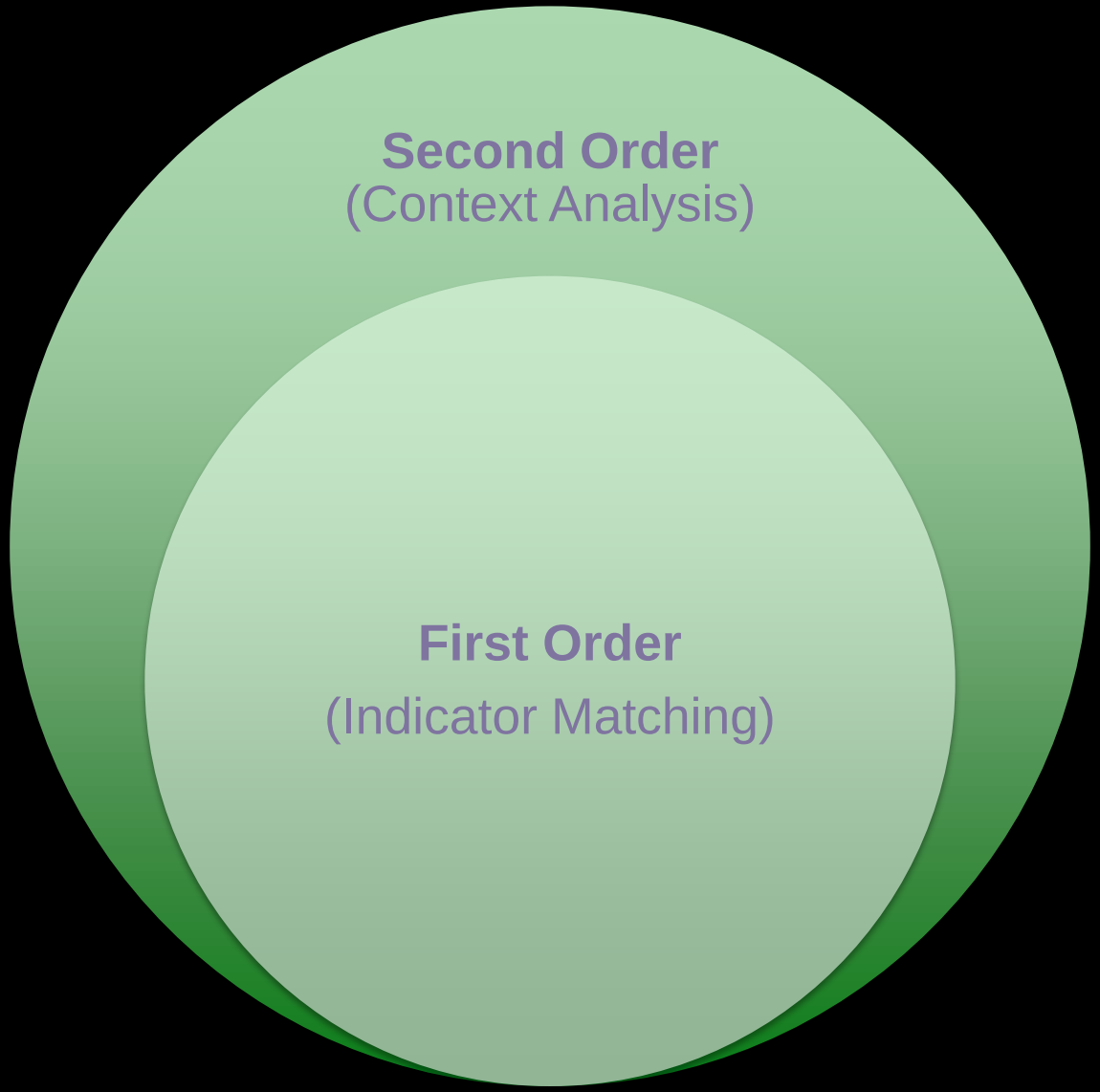
← Or this one either

← Would not touch this one

Without context that ".com" and ".org" are usually ok, automation fails

The Context Barrier

Breaking the Context Barrier



- Using internal and external enrichments to improve decision making
- Internal:
  - Statistical analysis internal data (a.k.a all of the UEBA stuff, PCR, ”stacking”)
  - Knowledge from internal incidents
- External:
  - Pivoting / Visual Aids
  - Statistical analysis from enrichment data (pDNS / WHOIS)

Example - Maliciousness Ratio

Let’s build aggregation metrics for ”good places” and ”bad places” in traffic

We propose a ratio that compares the cardinality of the node connectedness:

- **Bpp** – count of ”bad entities” connected to a specific pivoting point
- **Gpp** – count of ”good entities” connected to a specific pivoting point

$$MR_{pp} = \frac{B_{pp}}{G_{pp} + B_{pp}}$$

Example - Maliciousness Ratio

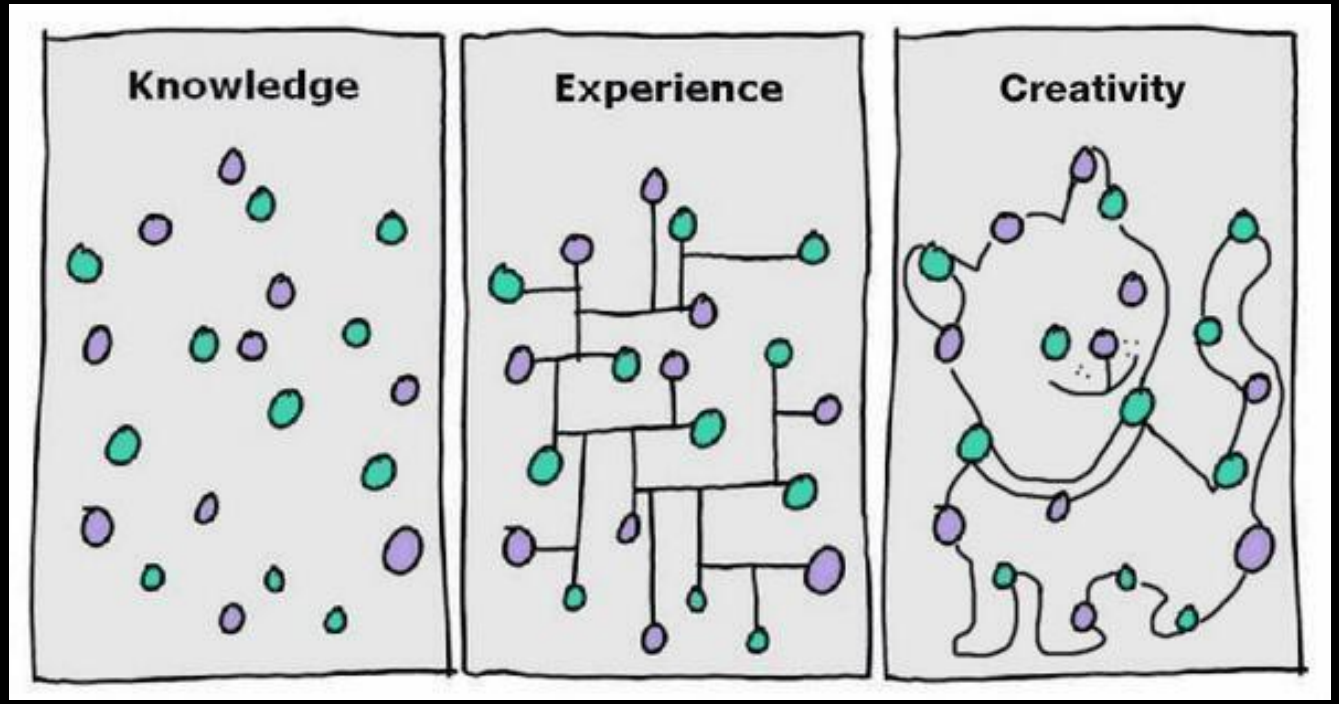
- Looking at the base rate:
  - ASN Base Rate 0.6%
  - Country Base Rate 0.58%
  - TLD Base Rate 1.9%

AS Number	ITGRAD , RU (48096)
IP	5.200.53.81
Hostname	we.soulomissions.org
Port	80 / TCP
Service	HTTP
Country	Russian Federation (RU)

- Telemetry from an pool of Niddel customers:
  - AS48096 – ITGRAD 87.5% => 145.9x more likely
  - Country RU 5.2% => 8.96x more likely
  - .org TLD 2.9% => 1.52x more likely

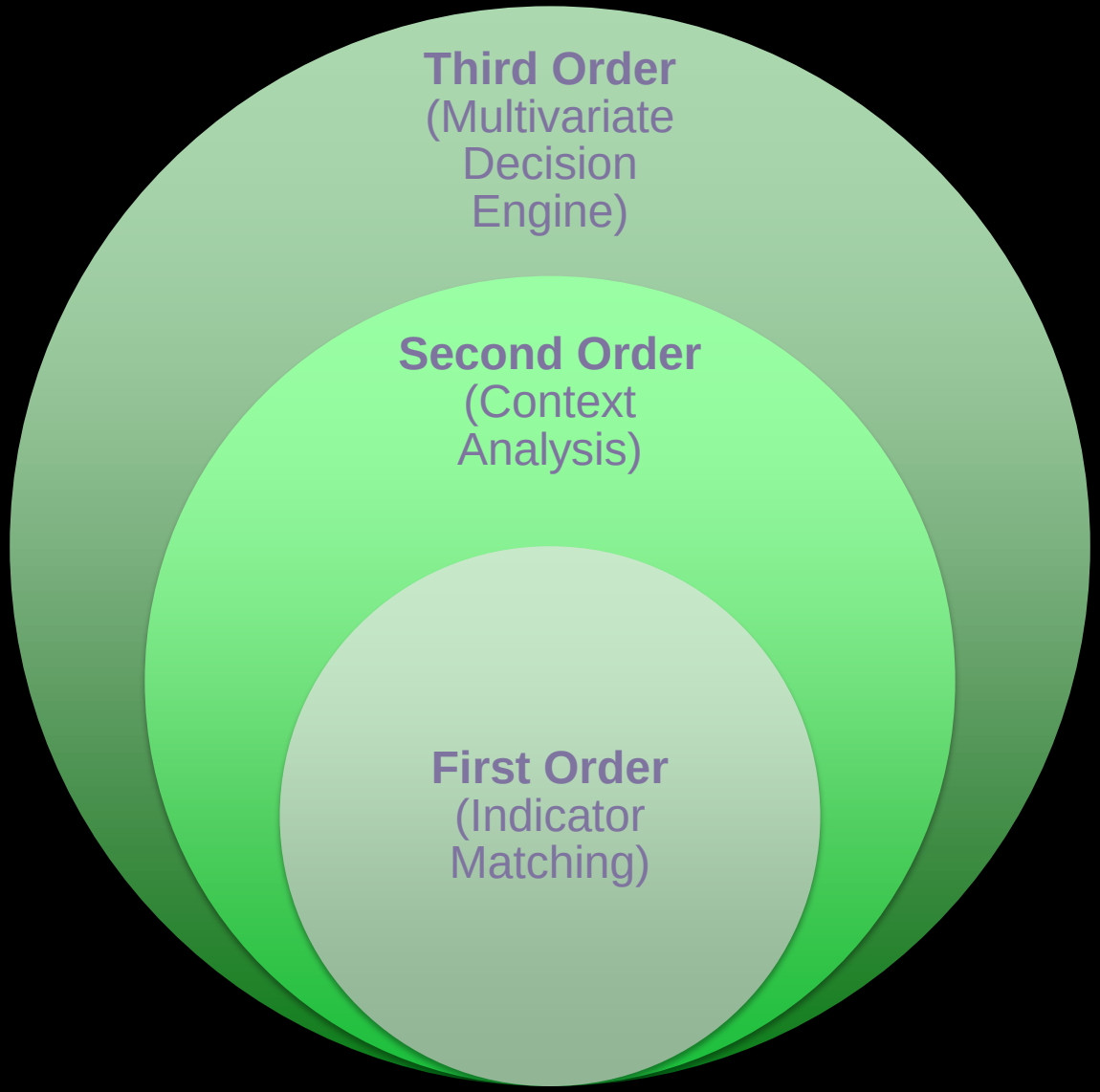
Challenges with the Approach

- How can we best define the cutting scores on all those potential maliciousness ratings?
- How to combine and weight the multivariate composition of these pivoting points?
- Solution is unique per company, including understanding telemetry patterns, risk appetite for FPs / FNs and decision points on when to block and when to alert on something.



The Experience Barrier

Breaking the Experience Barrier



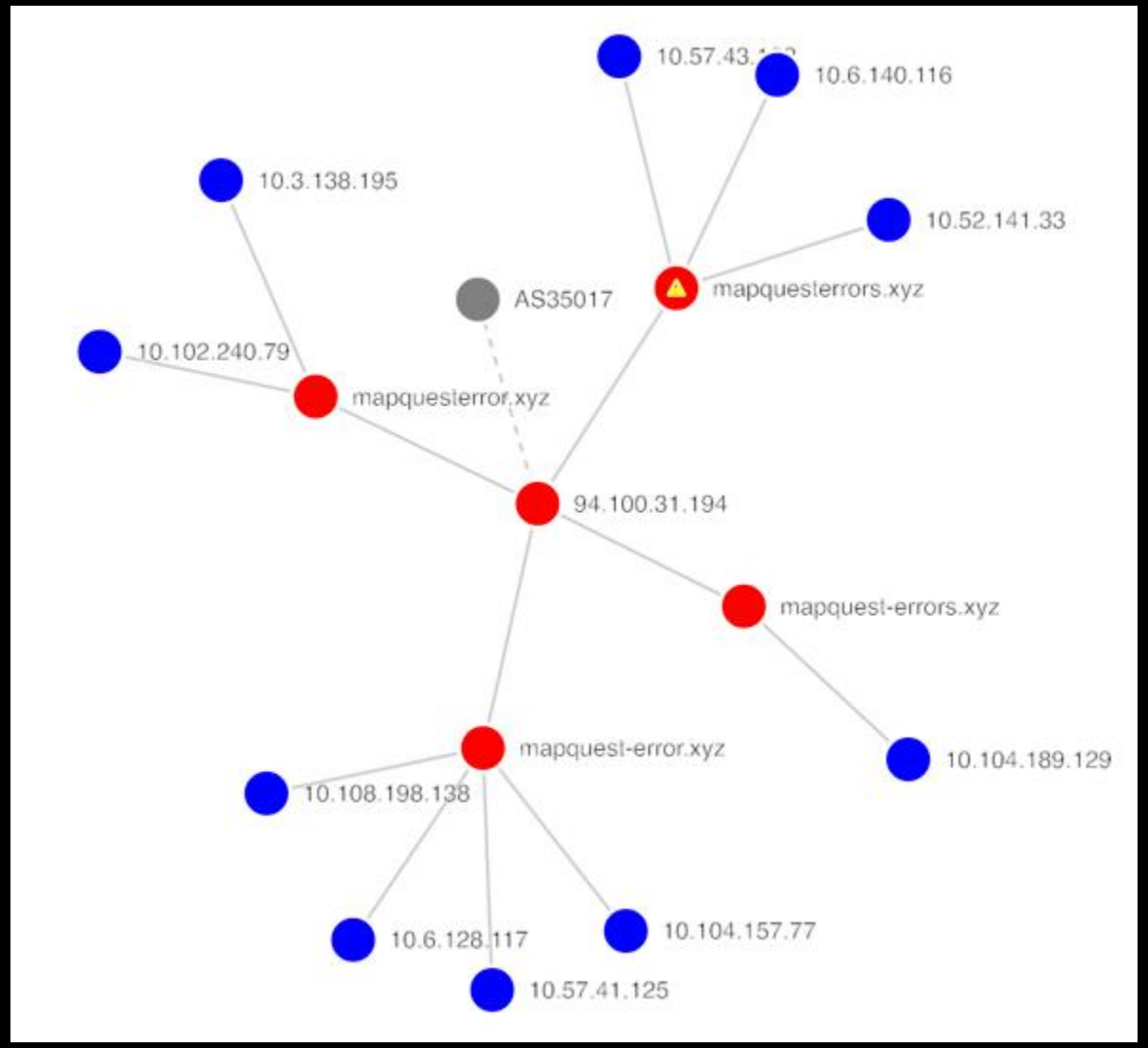
- Combining all the signals from the hunting investigation and making a ”call”:
  - Does being registered in REG-RU and hosted in OVH enough for a conviction?
  - This shady thing is registered in Mark Monitor. Viral legit campaign?

- This ”gut feeling” comes from years and years of knowledge and experience of handling alerts and incidents IRL.

Supervised Machine Learning!!

VS

A More Involved Example (1)



Maliciousness Rating

Country	Minimal (0.38x)
AS	Very High (42.80x)
BGP prefix	Very High (156.74x)
Dst. Host Public Suffix	Very High (15.30x)
Dst. Reverse Host Public Suffix	Very High (4.79x)
Dst. Reverse Host Org. Suffix	Low (71.00x)
Dst. Host SOA Authority	Minimal (0.00x)
Dst. Host SOA E-mail	Minimal (0.00x)
Dst. Host SOA NS	Minimal (0.00x)
Dst. Host WHOIS Registrar	Low (3.65x)
Dst. Host WHOIS Registrant	Low (4.41x)
Dst. Host WHOIS Registrant E-mail	Minimal (0.00x)
Dst. Host WHOIS NS	Very High (75.15x)

Matches

Source	Category	Campaign	Entity
malwaredomains	scam; private	MalwareDomains - scam - 2016-10-05	mapquesterrors.xyz

A More Involved Example (2)

94.100.31.194

BGP Details from October 6, 2016

BGP Prefix	94.100.31.0/24
AS Number	SWIFTWAY-AS Netherlands, ... (35017)

Location Details from October 6th, 2016

Country	Netherlands (NL)

Build the campaign based on the relationships - they all share the same support infrastructure on the IP Address and Name Servers.

WHOIS Details from October 6th, 2016

Authority	mapquesterrors.xyz
Registrar	TLD REGISTRAR SOLUTIONS LTD
Registrar IANA ID	1564
Created	October 5th, 2016
Updated	October 5th, 2016
Expires	October 5th, 2017
Name Servers	ns1.teachmewomen.com ns2.teachmewomen.com
Registrant	WHOIS PRIVACY CORP.
Registrant Street	Ocean Centre, Montagu Foreshore East Bay Street
Registrant City	Nassau
Registrant State	New Providence
Registrant Postal Code	N/A
Registrant Country	BAHAMAS
Registrant E-mail	mapquesterrors.xyz-owner-xhcn@customers.whoisprivacycorp.com
Registrant Phone	15163872248

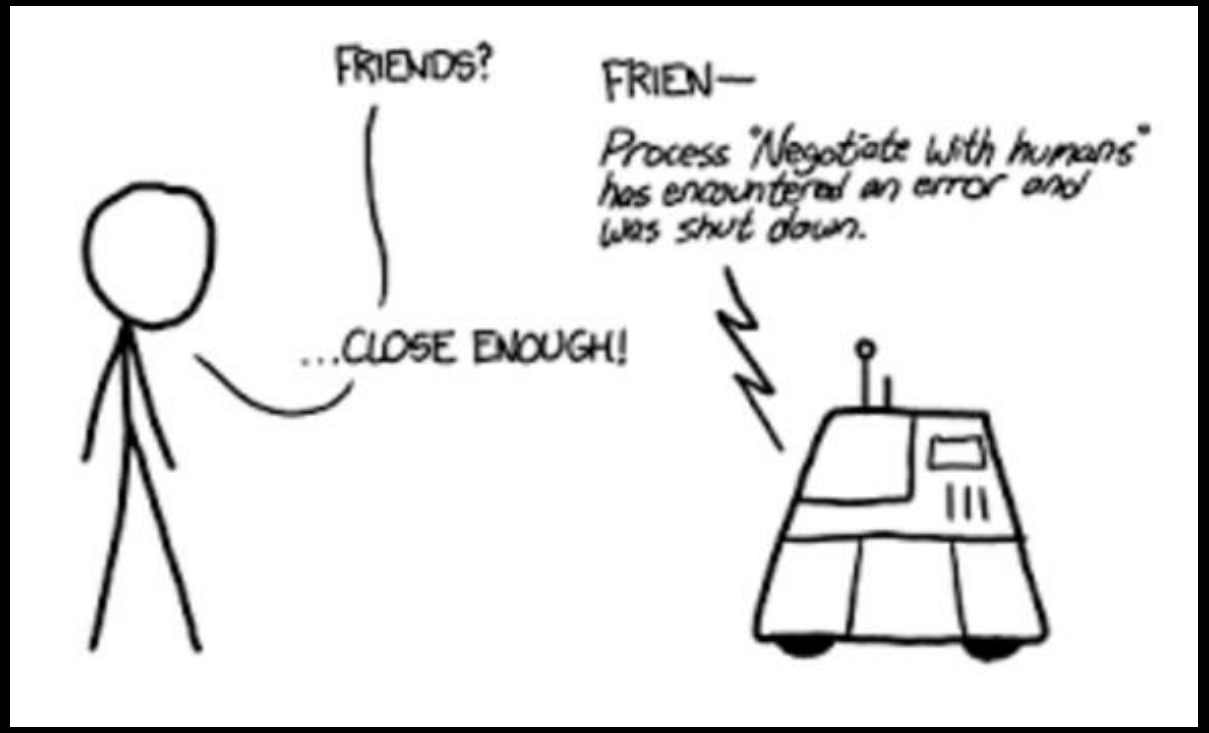
WHOIS Details from October 6th, 2016

Authority	mapquesterror.xyz
Registrar	TLD REGISTRAR SOLUTIONS LTD
Registrar IANA ID	1564
Created	October 5th, 2016
Updated	October 5th, 2016
Expires	October 5th, 2017
Name Servers	ns1.teachmewomen.com ns2.teachmewomen.com
Registrant	WHOIS PRIVACY CORP.
Registrant Street	Ocean Centre, Montagu Foreshore East Bay Street
Registrant City	Nassau
Registrant State	New Providence
Registrant Postal Code	N/A
Registrant Country	BAHAMAS
Registrant E-mail	mapquesterror.xyz-owner-qelm@customers.whoisprivacycorp.com
Registrant Phone	15163872248

The Lee-Bianco Barrier

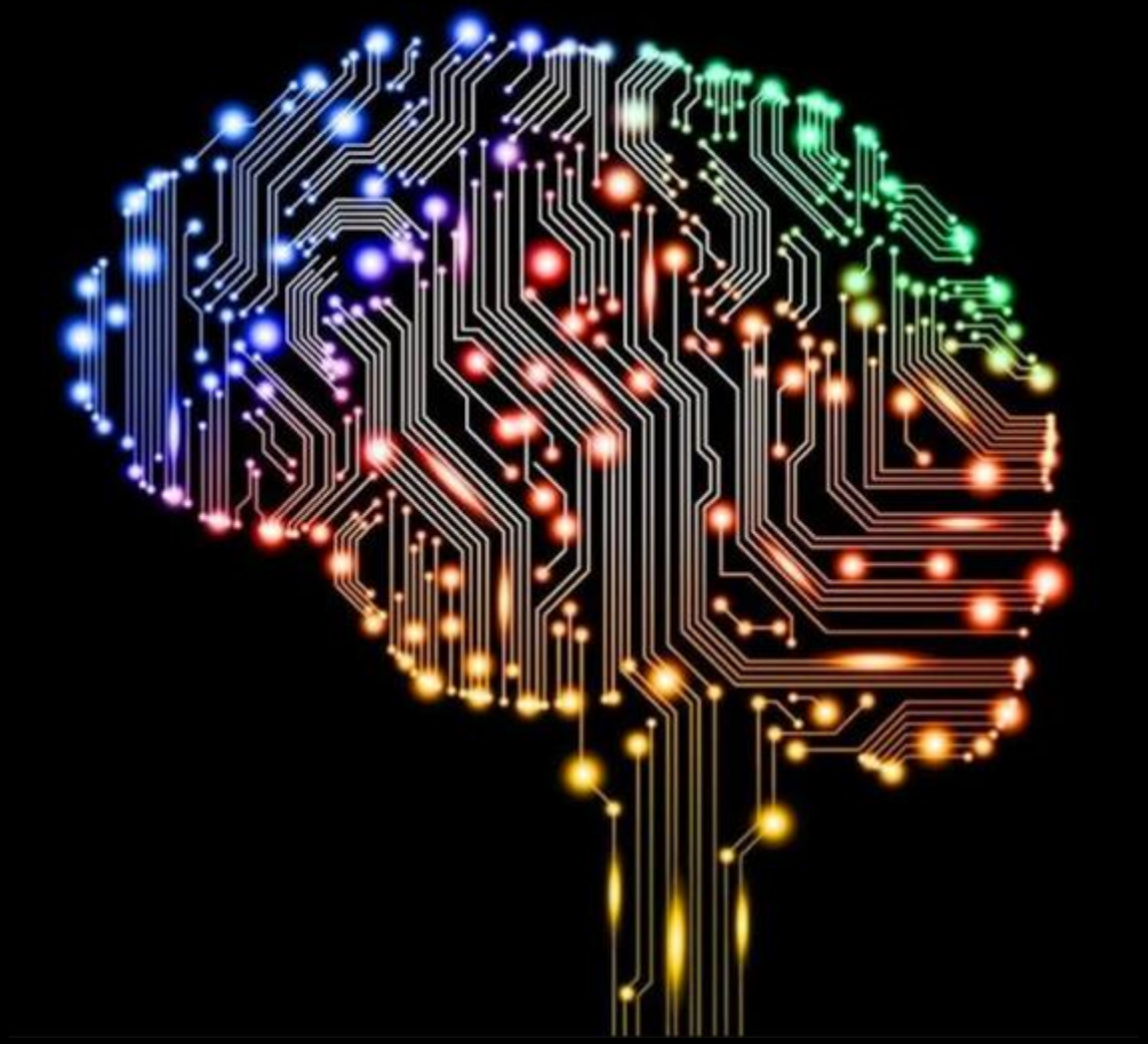
(a.k.a. The Creativity Barrier)

Now what?



- As threats evolve, new types of signals may be necessary for a conviction.
- If the system does not have access to the data that it requires, it cannot evaluate it for decision making.
- Some examples of recent ”new” threats - Domain fronting, IDN phishing
- This is no different from ”Writing a new Runbook” for your team

But what about Deep Learning?

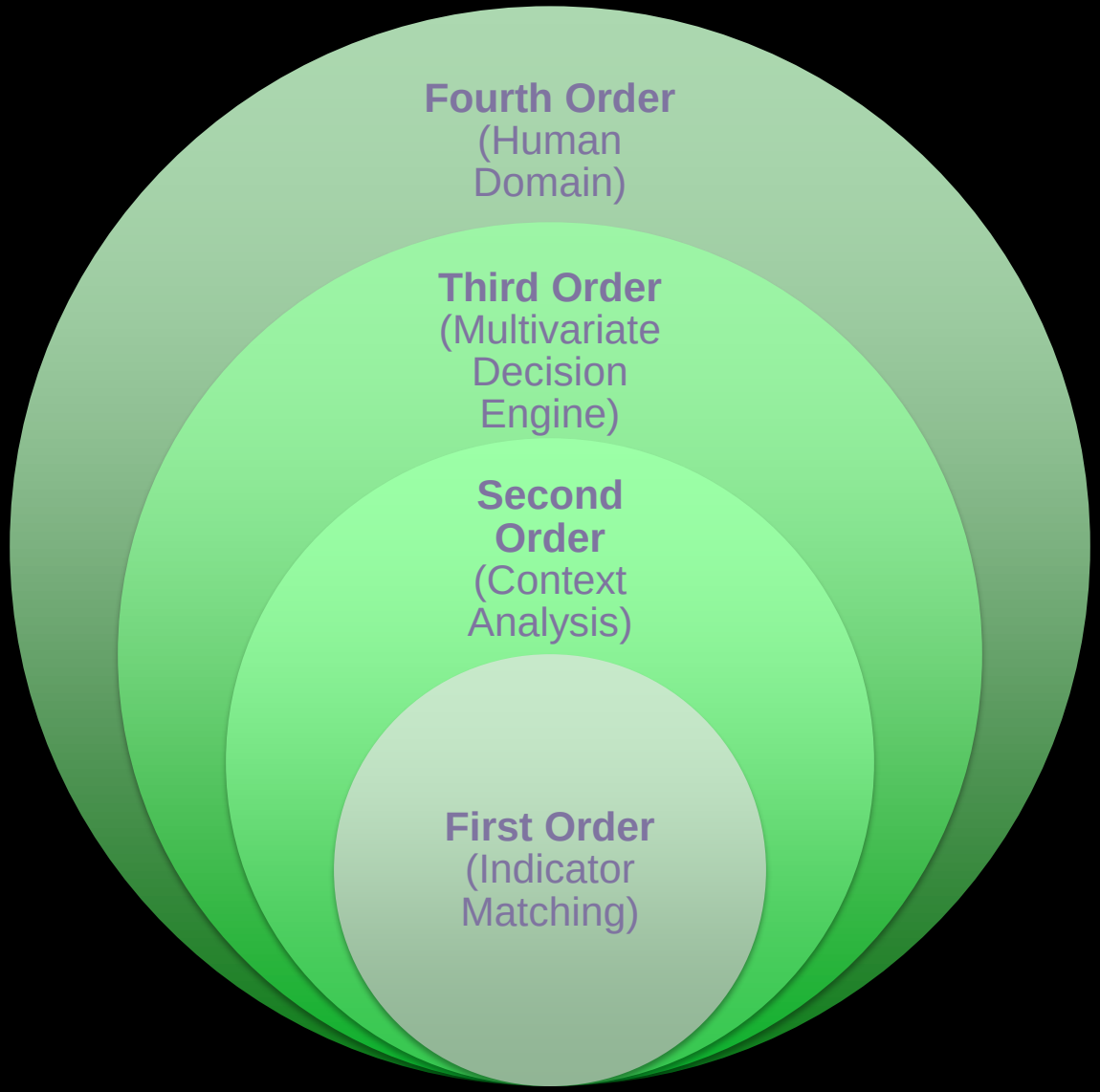


- Convolutional Neural Networks are very good at looking at unstructured data and ”figuring out” what the features should be.
- Great success for image and voice recognition:
  - Needs a lot of samples
  - Trivial to classify by a human
- Neither of these is the case for security – run away from DL vendors

Introducing HAMM

Hunting Automation Maturity Model (#HAMM)

Hunting Automation Maturity Model (#HAMM)

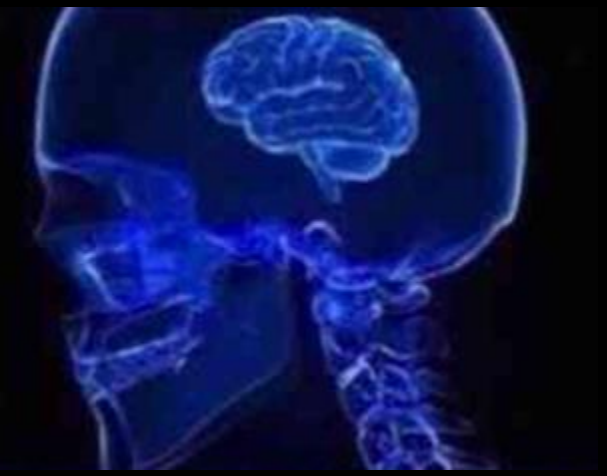


1. Vast majority of “automating hunting” plays - a **signature match**. Incomplete strategy, both prone to a lot of false positives in badly vetted lists and a lot of false negatives because the lists will naturally be incomplete.

2. In this level, a system is evaluating individual **hunting pivoting points** registered or first visited. Identify all the entries that are related to the high maliciousness pivoting points, and **even determine what they are related to** based on the connections to known malicious samples.

3. Multivariate decision making by **prioritizing which ones are the most relevant for detection under specific circumstances**. Third Order systems can decide on the fly which variables from First and Second Order are the most relevant for an environment.

Hunting Automation Maturity Model (#HAMM)



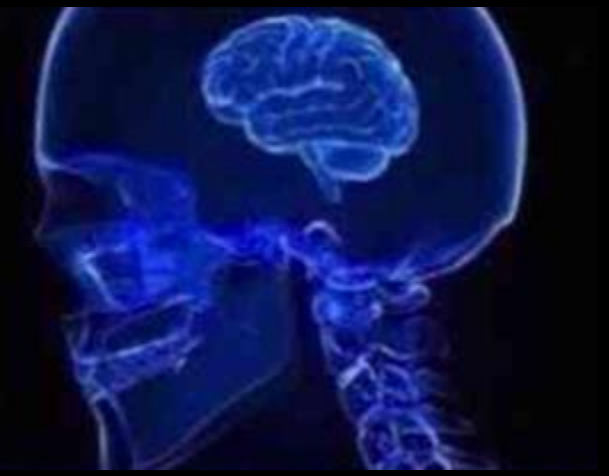
- IOC Matching
- Signatures
- Anti-virus

- Security / Hunting Analytics
- Stats methods
- (Some) UEBA – maybe?

- Supervised machine learning with previous signals

- Rob [M|T] Lee
- David Bianco
- Probably not you

Hunting Automation Maturity Model (#HAMM)



[LAME]

[Predictive Incident Response?]

[Proactive Incident Response?]

[Threat Farming?]

[MAGIC]

NIDDEL™

Share, like, subscribe
Q&A and Feedback please!

Alex Pinto – alexcp@niddel.com
@alexcpsec
@NiddelCorp

"Computers are useless. They can only give you answers." – Pablo Picasso