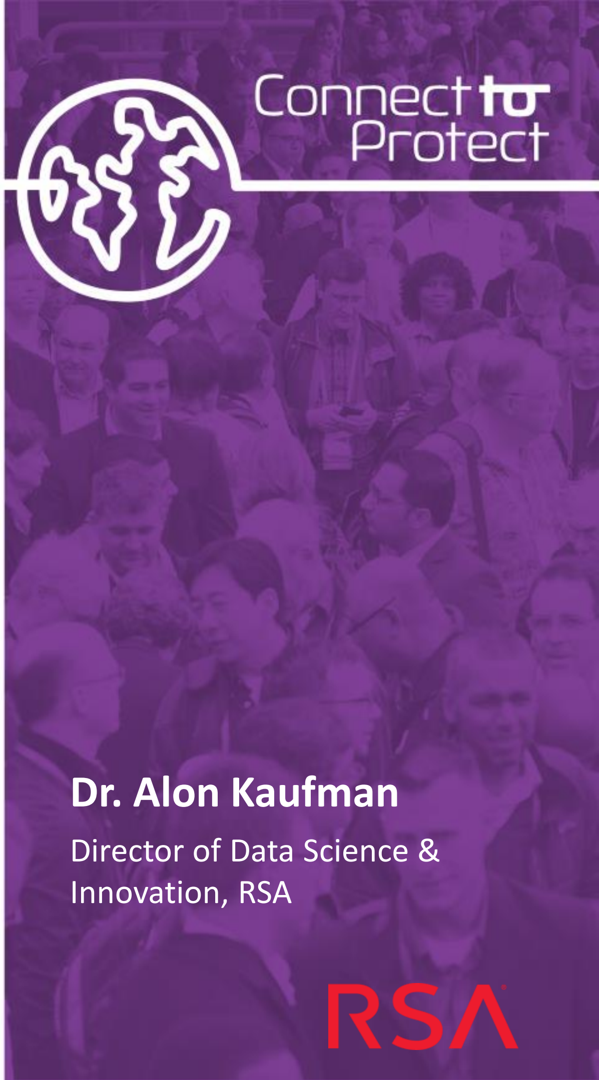
RSAConference2016

San Francisco | February 29 – March 4 | Moscone Center

Connect to Protect

SESSION ID: STR-R02

Data Science Transforming Security Operations



Dr. Alon Kaufman

Director of Data Science & Innovation, RSA

#RSAC

RSA

Data Science & Security Operation?

Who uses data science in their security practice?

In what processes throughout your security operations do you use data science?

Have you seen a significant value come out of your data science solutions?

Do you see data science playing in role in the Cybersecurity market shift: “By 2020, 60% of enterprise information security budgets will be allocated for rapid detection and response approaches, up from less than 20% in 2015 (Gartner) ”

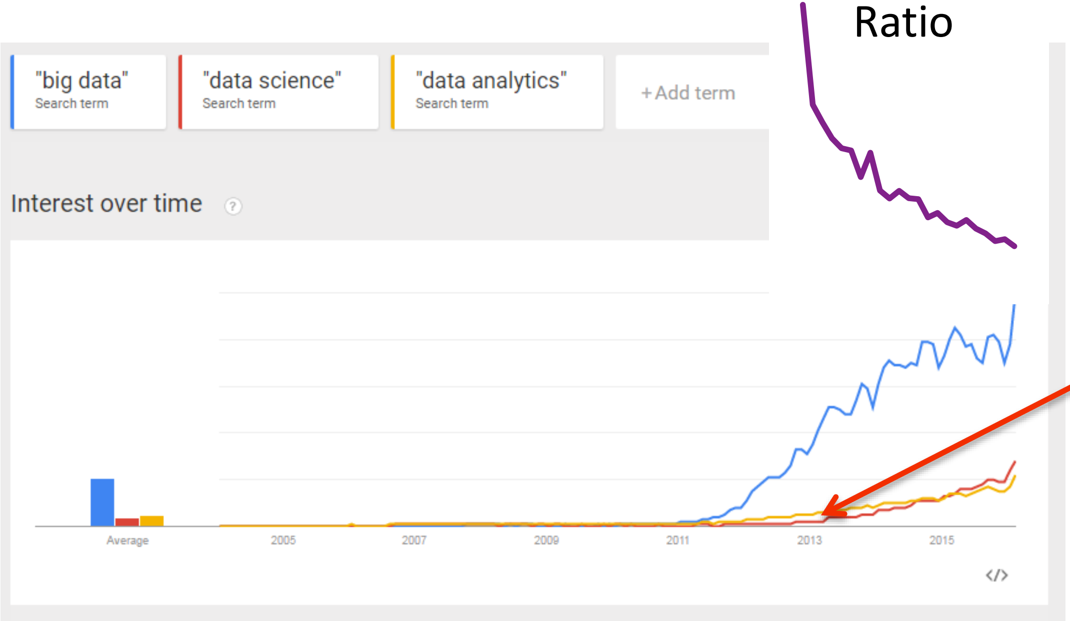
Data Science has way more to offer than prevention & detection... It can and should be used as a key methodology and technology spanning all processes in security operations....

Agenda

- What is data science, and why in security?
  - You should know by now ;)
  - What's special about data science in security
- 5 Maturity levels of data science in security operations
  - Data science goes way beyond the prevention & detection in the entry level…
- DS maturity survey
  - Where is your organization/product in terms of DS maturity?
- Building a security data science practice in house, Yes or No?
- Summary

What is Data Science – in 1 Sentence

- Making sense out of big data…
  - Getting the data we collect to work for us



The demand is just growing…

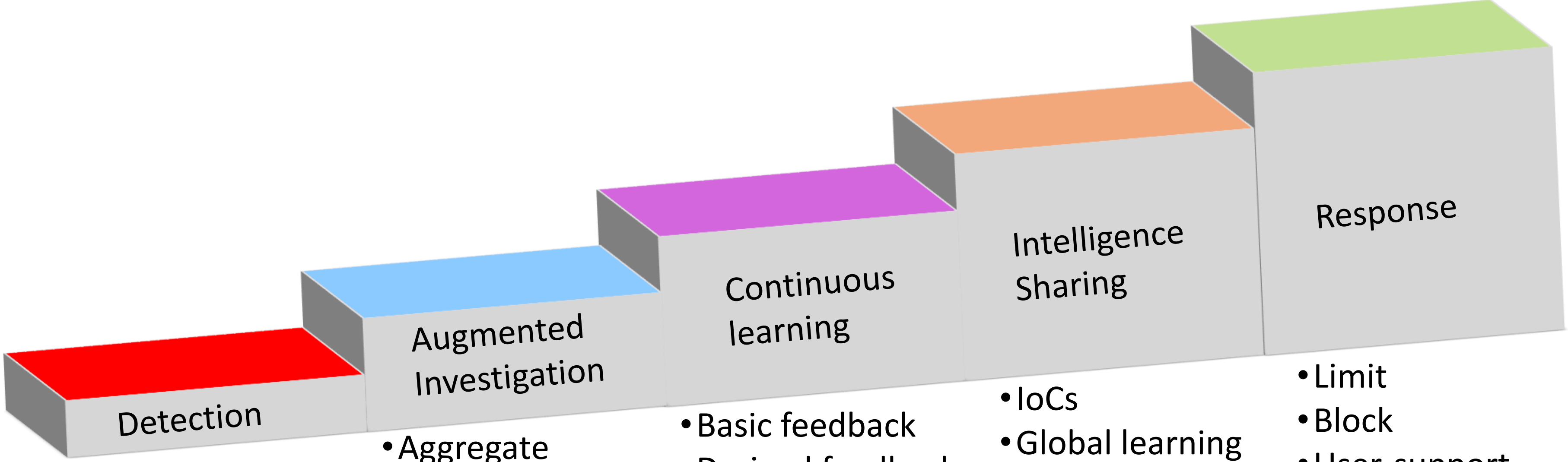
Why Data Science in Security?

- We have all (most) of the data already….. Yet still being breached… while the attacks are hidden in our data
- Security operations are getting too complex for humans alone… and we are facing a huge staffing gap…
- Other industries demonstrated huge value with DS, given a hard problem and the relevant data at hand:
  - Retail recommendation systems, up-sells, cross-sell
  - Bio-informatics
  - Image object recognition
  - Voice recognition
  - Self driving cars
  - …

What's Special About Data Science in Security?

- Dealing with a hostile dynamic world!
- Human/Machine synergy
- High price of *False-Negative* errors
- Gathering/Sharing data
- Lack of labeled attacks for training and learning
- In security detection is just the beginning….

5 Levels of Data Science Maturity



Detection
- Known bad
- Adaptive learning
- Integrated scoring

Augmented Investigation
- Aggregate
- Prioritize
- Automate & Recommend

Continuous learning
- Basic feedback
- Derived feedback
- Learning from analyst actions

Intelligence Sharing
- IoCs
- Global learning
- Policies

Response
- Limit
- Block
- User-support

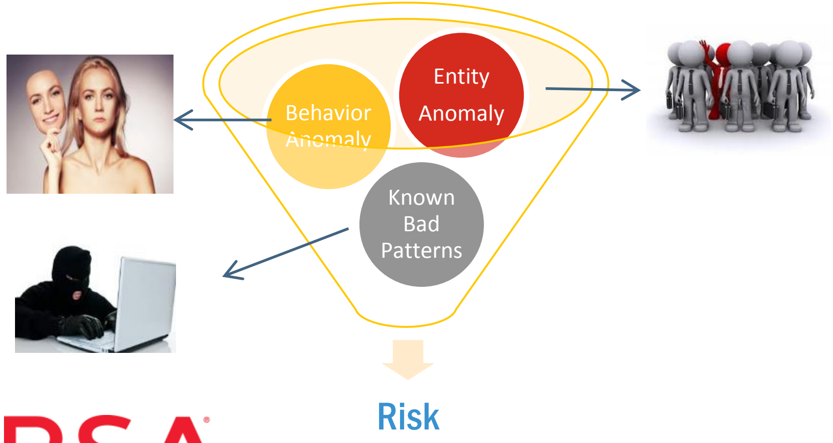
Key message: Data science is a key methodology and technology, not a plug-in feature…

Detection: The Holy Grail of Data Science….

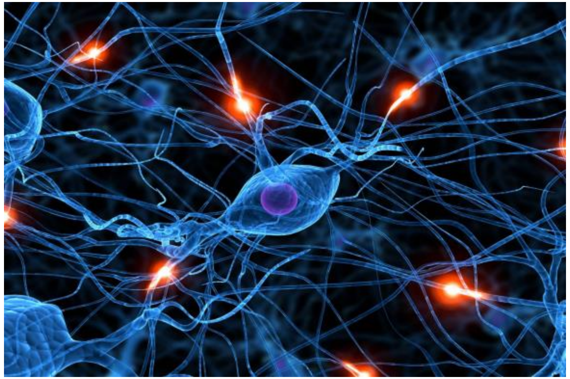
- The data exists, and so also endless point solutions for detection

- The key to success is:

Compressive Risk Scoring



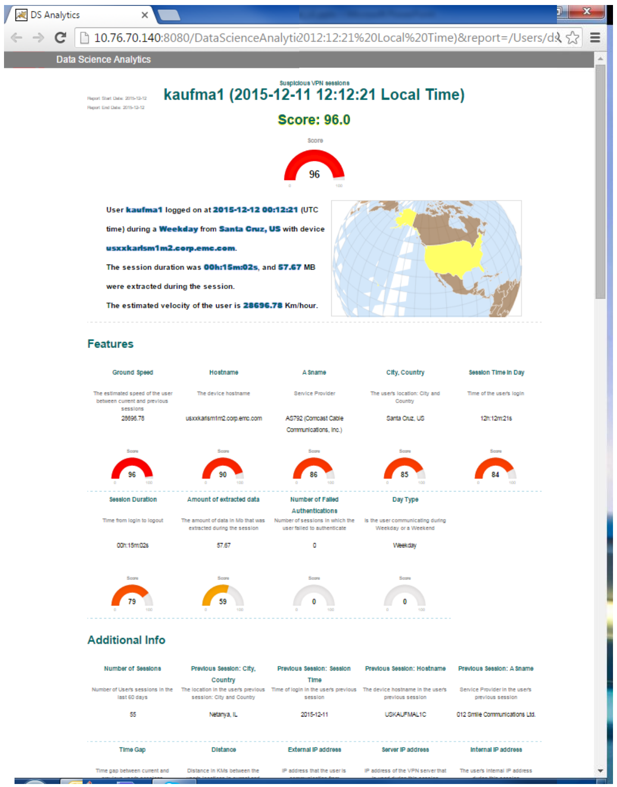
Integrated Approach



Comprehensive Risk Score - Example

Suspicious User Login Detection

- Multivariate Machine Learning algorithm to detect login impersonation
  - Multiple inputs from multiple sources:
    - Hostname, location, server, duration, auth, time of day, data tx/rx,….
  - Model output
    - Risk score (combined measure of how risky the behavior is)
  - Modeling concept:
    - **Known bad**: blocked users, unrealistic ground-speed, authentication
    - **User anomaly**: base line per feature and detect deviation from norm
    - **Peer group anomaly:** Prior knowledge, new user, acceptable behavior changes


DS Analytics
10.76.70.140:8080/DataScienceAnalyti2012:12:21%20Local%20Time)&report=/Users/d
Data Science Analytics

Suspicious VPN sessions

Report Start Date: 2015-12-12
Report End Date: 2015-12-12

kaufma1 (2015-12-11 12:12:21 Local Time)

Score: 96.0

User **kaufma1** logged on at **2015-12-12 00:12:21** (UTC time) during a **Weekday** from **Santa Cruz, US** with device **usxxkarlsm1m2.corp.emc.com**.

The session duration was **00h:15m:02s**, and **57.67** MB were extracted during the session.

The estimated velocity of the user is **28696.78** Km/hour.

Features

Ground Speed	Hostname	A Sname	City, Country	Session Time In Day
The estimated speed of the user between current and previous sessions	The device hostname	Service Provider	The user's location: City and Country	Time of the user's login
28696.78	usxxkarlsm1m2.corp.emc.com	AS792 (Comcast Cable Communications, Inc.)	Santa Cruz, US	12h:12m:21s
Score 96	Score 90	Score 86	Score 85	Score 84

Session Duration	Amount of extracted data	Number of Failed Authentications	Day Type
Time from login to logout	The amount of data in Mb that was extracted during the session	Number of sessions in which the user failed to authenticate	Is the user communicating during Weekday or a Weekend
00h:15m:02s	57.67	0	Weekday
Score 79	Score 59	Score 0	Score 0

Additional Info

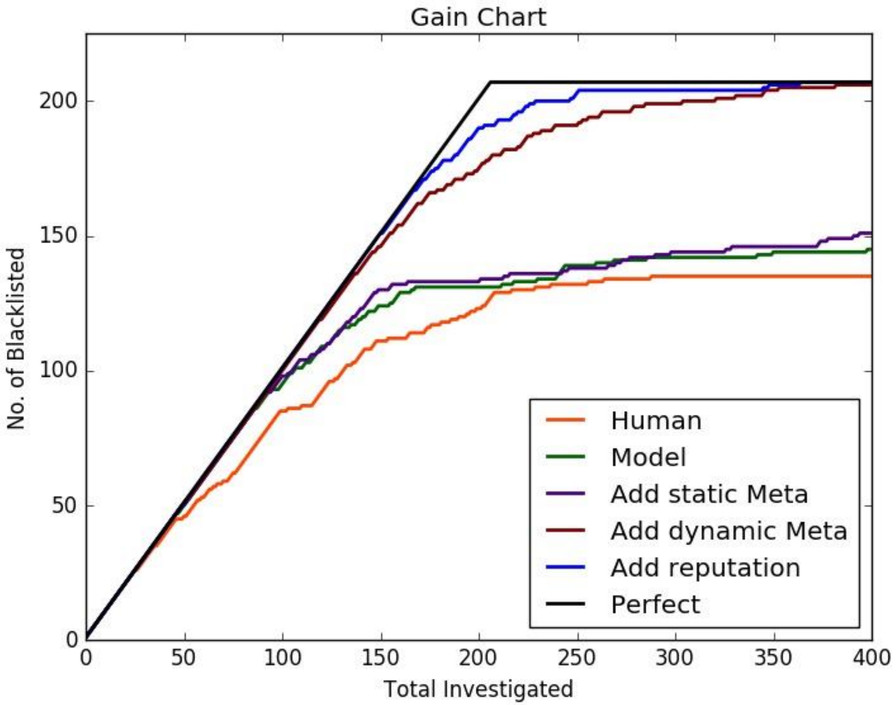
Number of Sessions	Previous Session: City, Country	Previous Session: Session Time	Previous Session: Hostname	Previous Session: A Sname
Number of User's sessions in the last 60 days	The location in the user's previous session: City and Country	Time of login in the user's previous session	The device hostname in the user's previous session	Service Provider in the user's previous session
55	Netanya, IL	2015-12-11	USKAUFMAL1C	012 Smile Communications Ltd.

Time Gap	Distance	External IP address	Server IP address	Internal IP address
Time gap between current and	Distance in KMs between the	IP address that the user is	IP address of the VPN server that	The user's internal IP address

Integrating Different Approaches - Example

Endpoint Malware Detection

- The market is highly fragmented with endless point solutions
- Each vendor/solution takes a different valid approach with pros and cons
- Combining them provides enhanced performance:
  - Human
  - Static analysis
  - Dynamic analysis
  - Community reputation



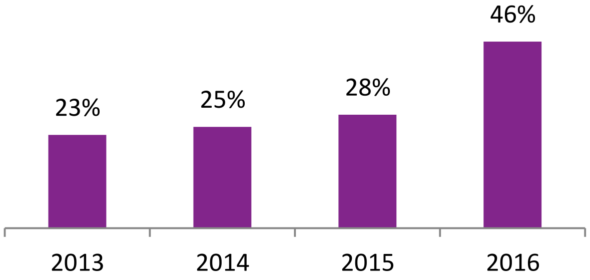
Augmented Investigation

- The goal is not replace the analysts but augment them and simplify their work:
  - Shortage of cybersecurity skills continues to grow
  - Most of analysts’ time goes on selecting what alerts to investigate
  - Attacks typically trigger multiple alerts throughout the different attack phases
  - 70% of the procedures done by analysts are repeatable
- The Key to success:
  - Prioritize
  - Aggregate
  - Automate & Recommendation

Augmented Investigation

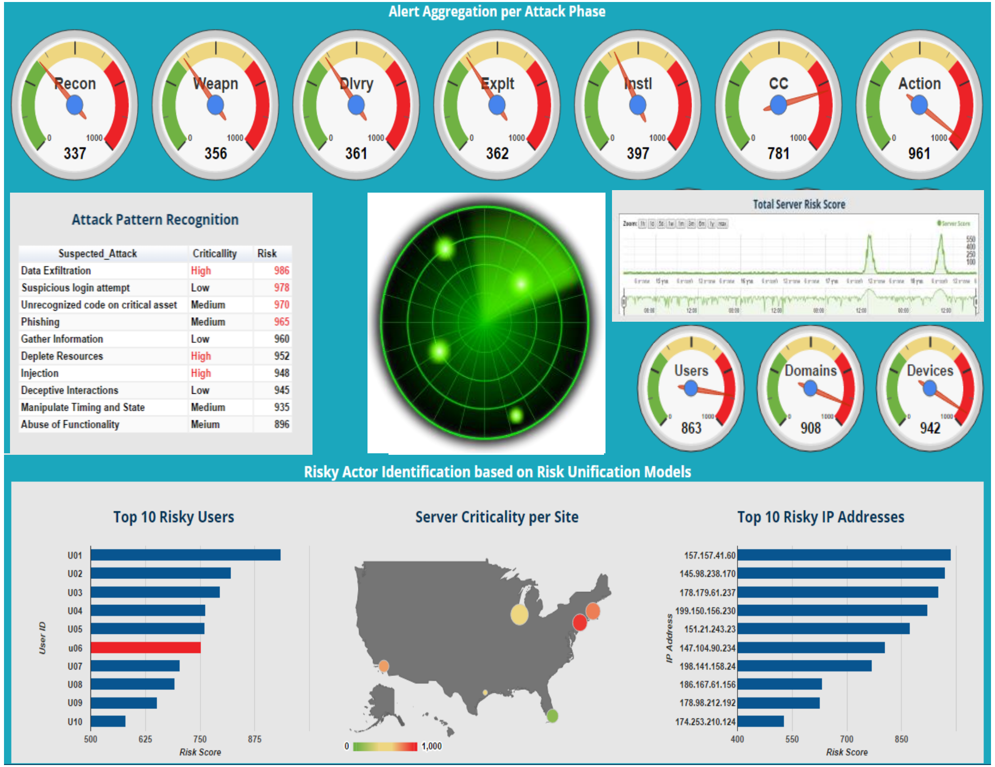
Shortage in CyberSecurity Skills

(ESG, 2016)



Augmented Investigation - Example

- Top-down Hierarchical approach
- Pre-fetch all supporting data
- Risk scoring prioritization
- Aggregate across entities (user, devices, application, …)
- Moving from alerts to attack vectors
- Guide the analyst with recommendations

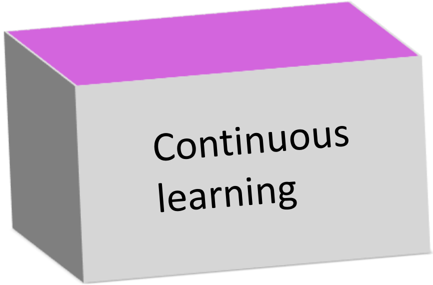


Attack Pattern Recognition

Suspected_Attack	Criticallity	Risk
Data Exfiltration	High	986
Suspicious login attempt	Low	978
Unrecognized code on critical asset	Medium	970
Phishing	Medium	965
Gather Information	Low	960
Deplete Resources	High	952
Injection	High	948
Deceptive Interactions	Low	945
Manipulate Timing and State	Medium	935
Abuse of Functionality	Meium	896

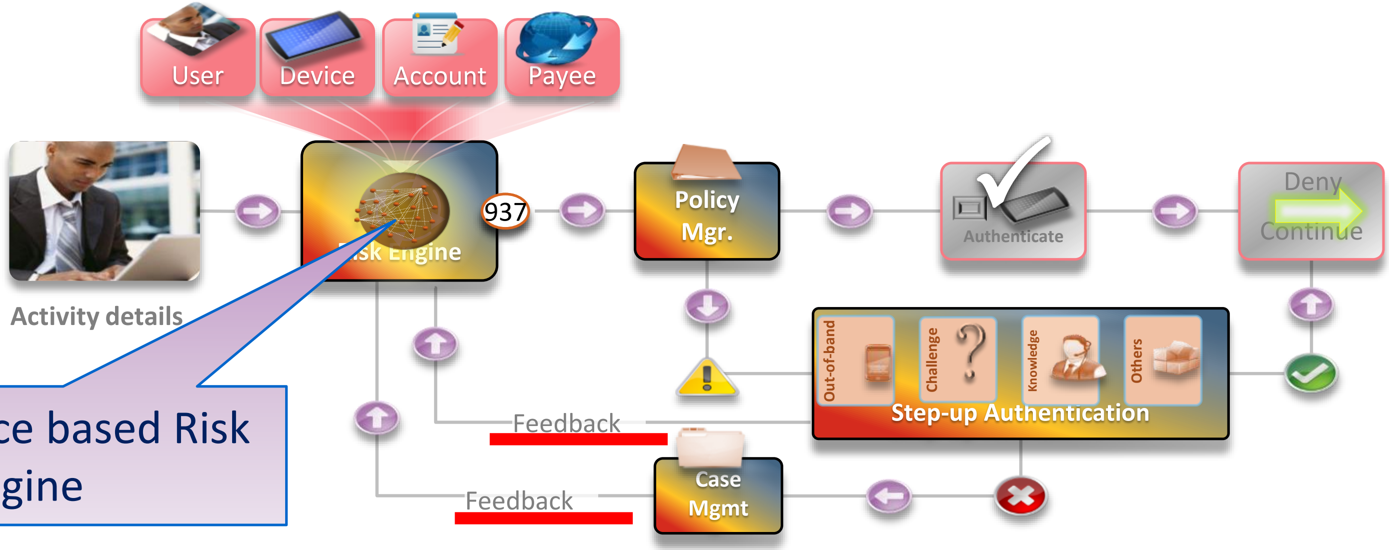
Continuous Learning

- As in any learning “teachers” are beneficial – supervised learning
  - Feeding back results to the learning engine
  - When direct feedback is lacking it can be derived
  - Learning from analyst behavior and actions


Continuous learning

Leaning and Self-Improving Detection - Example

- Ongoing, automatic self-learning fraud detection model



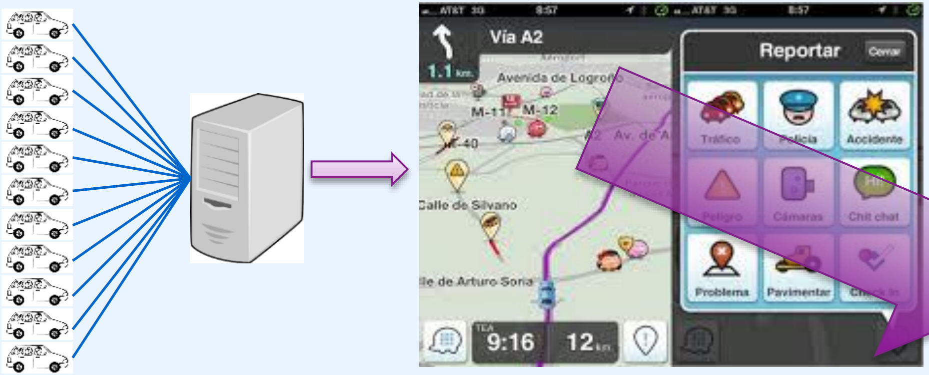
Data Science based Risk Engine

Intelligence Sharing

Tiny part of the road from each

Analytics

Map + prediction + navigation instruction

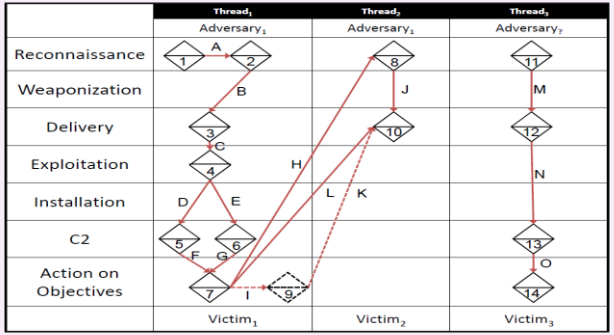


Waze. Outsmarting traffic, Together.

Intelligence Sharing

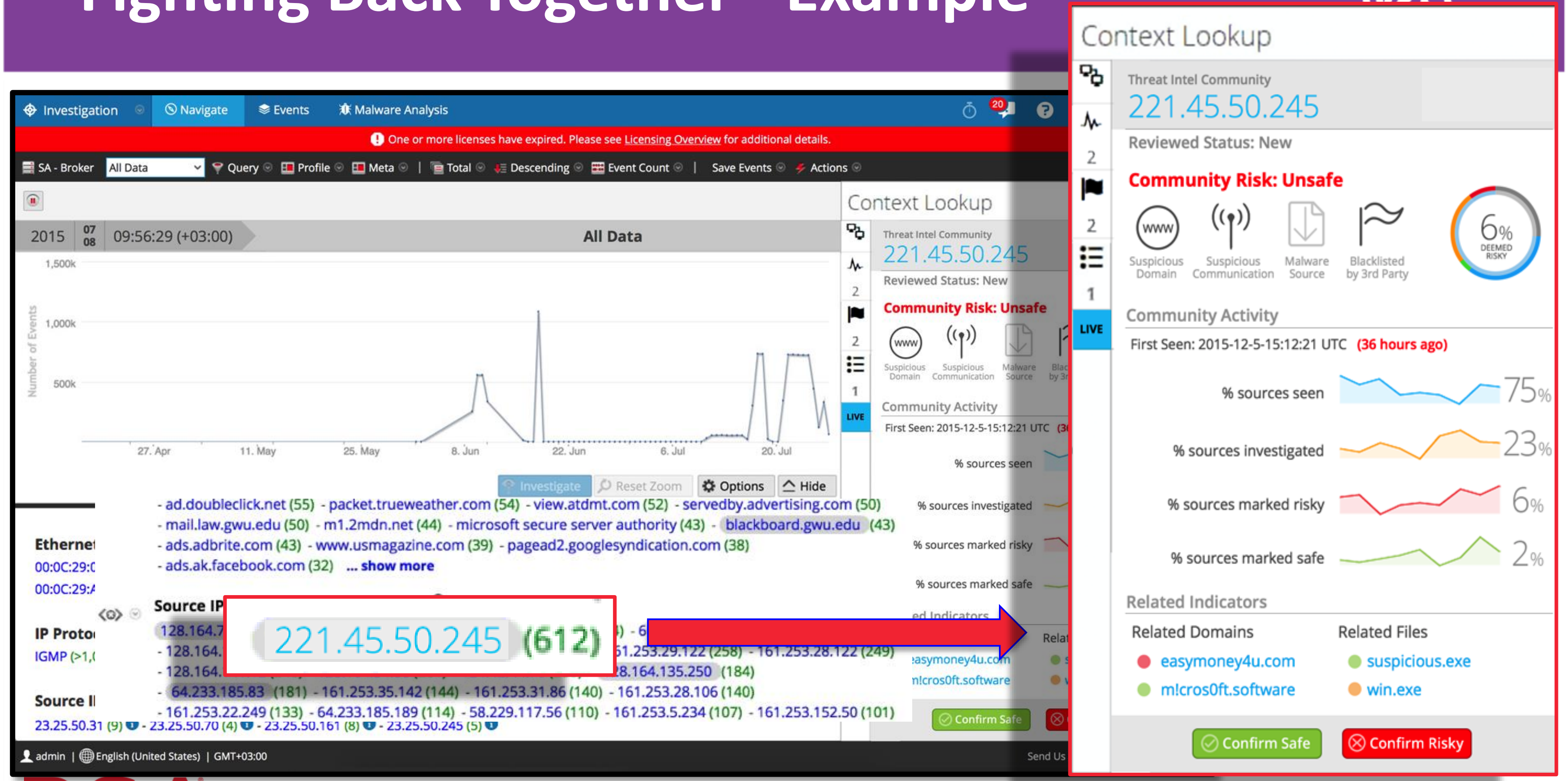
Crowdsourced security intel’

Security map + predictions + mitigation instructions



To date the industry state of the art sharing is around ***IoCs***, next phase is to share, learn and crowdsource ***policies***, ***procedures*** & ***mitigations***

Fighting Back Together - Example



One or more licenses have expired. Please see Licensing Overview for additional details.

2015 07 08 09:56:29 (+03:00) All Data

- ad.doubleclick.net (55) - packet.trueweather.com (54) - view.atdmt.com (52) - servedby.advertising.com (50)
- mail.law.gwu.edu (50) - m1.2mdn.net (44) - microsoft secure server authority (43) - blackboard.gwu.edu (43)
- ads.adbrite.com (43) - www.usmagazine.com (39) - pagead2.googlesyndication.com (38)
- ads.ak.facebook.com (32) ... show more

Ethernet
00:0C:29:0
00:0C:29:A

Source IP
128.164.7
- 128.164.
- 128.164.

221.45.50.245 (612)

) - 6
61.253.29.122 (258) - 161.253.28.122 (249)
28.164.135.250 (184)
- 64.233.185.83 (181) - 161.253.35.142 (144) - 161.253.31.86 (140) - 161.253.28.106 (140)
- 161.253.22.249 (133) - 64.233.185.189 (114) - 58.229.117.56 (110) - 161.253.5.234 (107) - 161.253.152.50 (101)

IP Proto
IGMP (>1,(

Source I
23.25.50.31 (9) - 23.25.50.70 (4) - 23.25.50.161 (8) - 23.25.50.245 (5)

admin | English (United States) | GMT+03:00

Context Lookup

Threat Intel Community
221.45.50.245

Reviewed Status: New

Community Risk: Unsafe

Community Activity

First Seen: 2015-12-5-15:12:21 UTC (36 hours ago)

% sources seen
% sources investigated
% sources marked risky
% sources marked safe

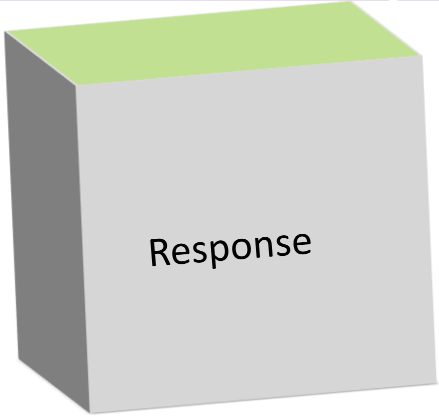
Related Indicators

Related Domains
- easymoney4u.com
- m!cros0ft.software

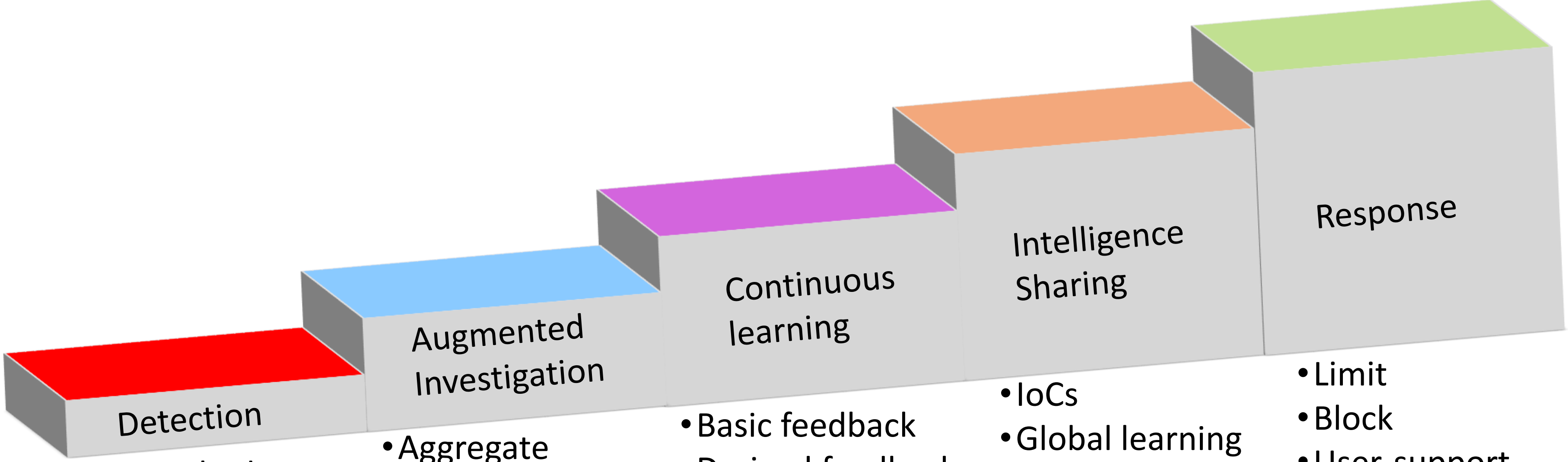
Related Files
- suspicious.exe
- win.exe

Response

- Taking automatic actions based on insights:
  - Limit access / Require additional input
    - Risk based authentication
    - Partial blocking
  - Automatic blocking
  - Guide the analyst through investigation
    - Pre-fetch all required data
    - Recommend next action



5 Levels of Data Science Maturity



Detection

- Known bad
- Adaptive learning
- Integrated scoring

Augmented Investigation

- Aggregate
- Prioritize
- Automate & Recommend

Continuous learning

- Basic feedback
- Derived feedback
- Learning from analyst actions

Intelligence Sharing

- IoCs
- Global learning
- Policies

Response

- Limit
- Block
- User-support

Key message: Data science is a key methodology and technology, not a plug-in feature…

Survey: How DS-Mature Are Your Operations?

(How many fields? (5), Overall score? (22 points))

Detection

- ✓ Do you use advanced, adaptive, analytics for detection?
- ✓ Can you bake into the analytics engines your human insights?
- ✓ Do you have your various products integrated at the analytics level?

Augmented Investigation

- ✓ Can you combine multiple alerts into some attack description?
- ✓ Do you have one integrated priority queue?
- ✓ Do you utilize automatic enrichments, hints, guidance or recommendation to assist analysts?

Continuous Learning

- ✓ Do you leverage analysts decision for operations improvement?
- ✓ Do you have any level of automatic, self learning from feedback?
- ✓ Do your overall operations improve based on your analysts work?

Intelligence Sharing

- ✓ Do you utilize community data to improve operations?
- ✓ Do your systems “learn” from data outside of your system?
- ✓ Do you have a mechanism to improve human actions based on the community?

Response

- ✓ Do you use automatic response based on analytics?
- ✓ Are any decisions or actions fed back to analysts as a results of the risk?

Building a Security Data Science Practice in House, Yes or No?

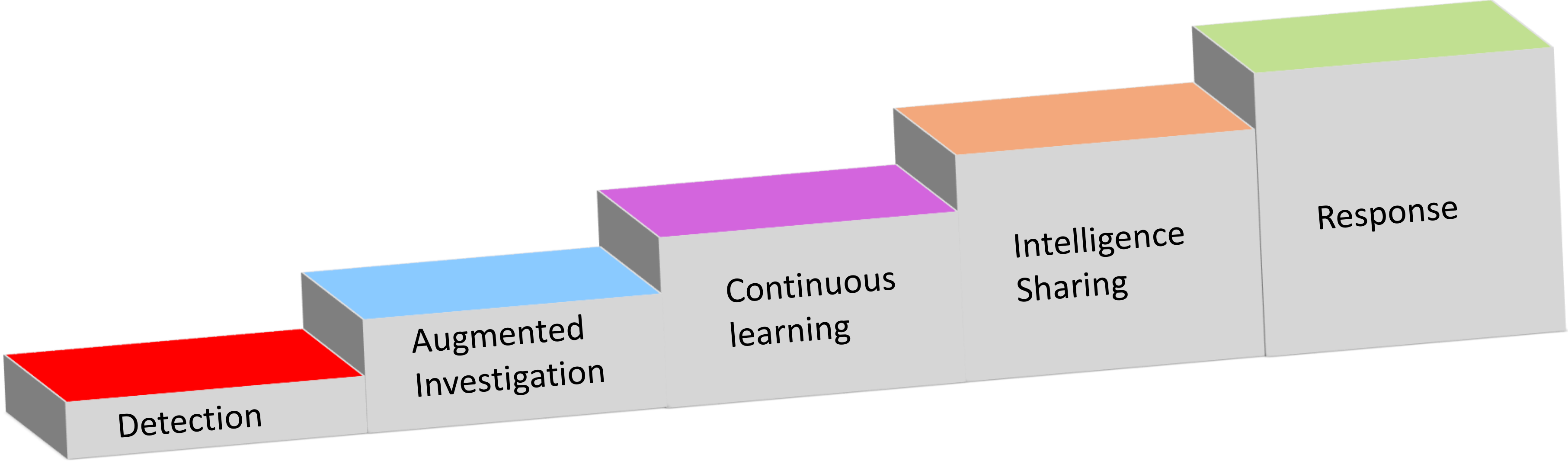
- Applying Data Science requires joint effort between data scientists, security experts and the business owners ↔ Alignment from stakeholders
- To date hiring people with a data science background is hard, nevertheless with security domain knowledge ↔ Invest in staffing and diverse backgrounds
- From research to an operational process/product – long journey from the proof-of-signal to an operational system ↔ Organization & operational breadth
- Data, Data, Data…. ↔ Collaborate / share
- You don't want data science… you actually want data science backed into your solution in an intuitive, easy to use manner ↔ Integrated home grown solution

Applying What You Have Learned Today

- Take the survey and assess how advanced is your DS strategy
- Identify gaps, and in what area focus is needed
- Work up the DS stairs:
  - Detection -> Investigation -> continuous learning -> Intl Sharing -> Automatic response (Risk based response)
- Data Science in house:
  - Alignment cross-org
  - Staff wisely
  - Be prepared for a long (and expensive) journey
- Constantly strive to see how DS augments your analysts, and not try replace them!

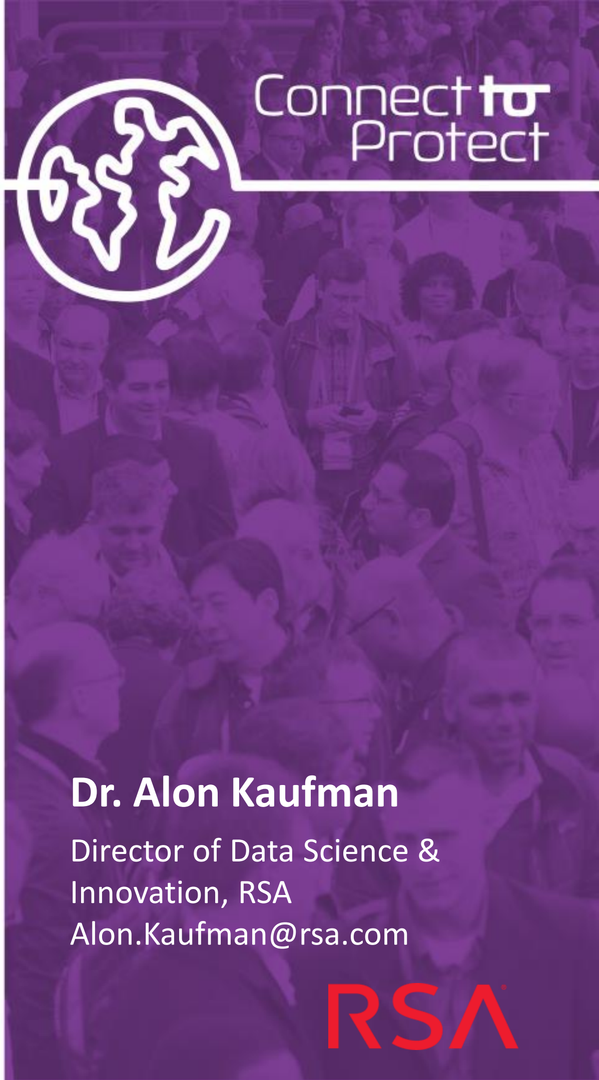
Summary

**Data Science has way more to offer than prevention & detection ...
It can and should be used as a key methodology and technology spanning all processes in security operations…**



RSAConference2016

San Francisco | February 29 – March 4 | Moscone Center



Connect to Protect

SESSION ID: STR-R02

Data Science Transforming Security Operations

Dr. Alon Kaufman

Director of Data Science & Innovation, RSA
Alon.Kaufman@rsa.com

RSA

#RSAC