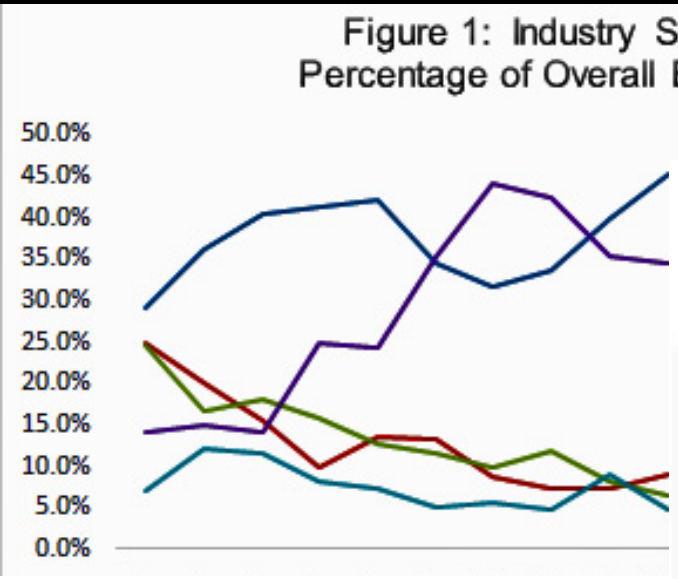
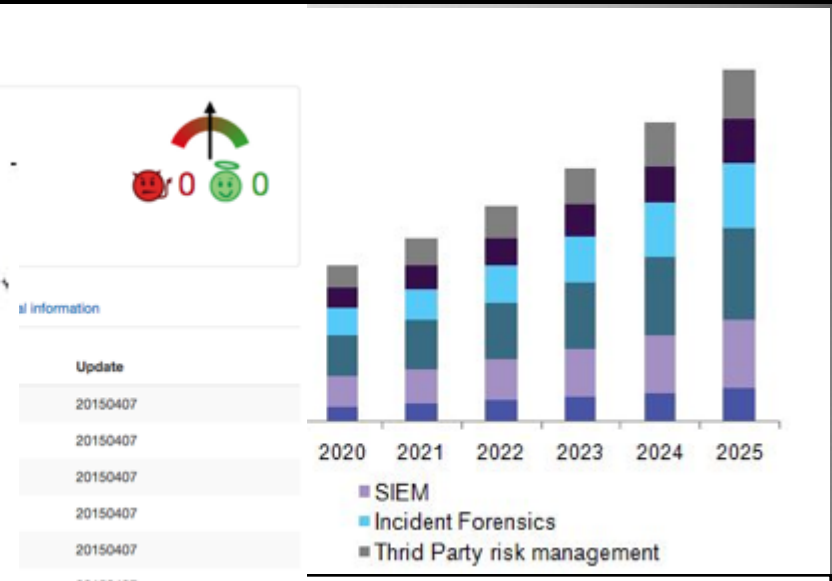
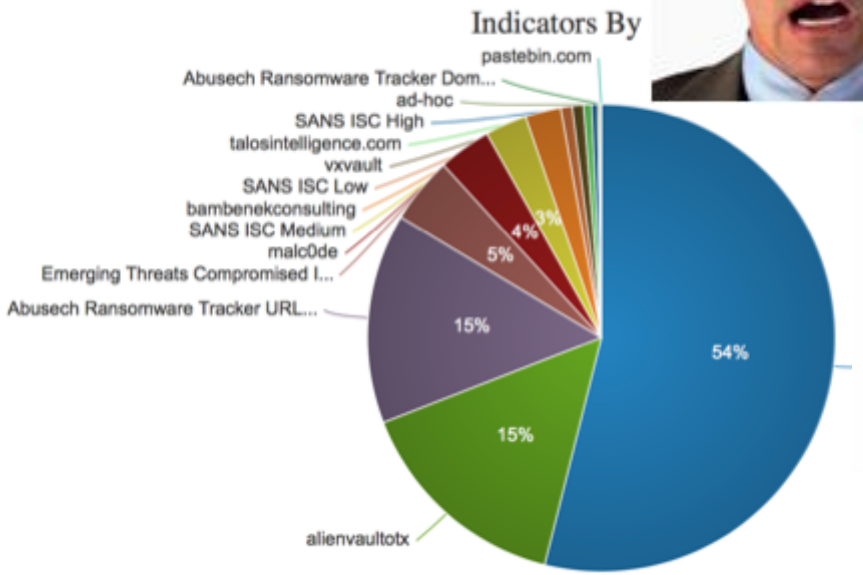
Practical Statistics for Threat Intelligence

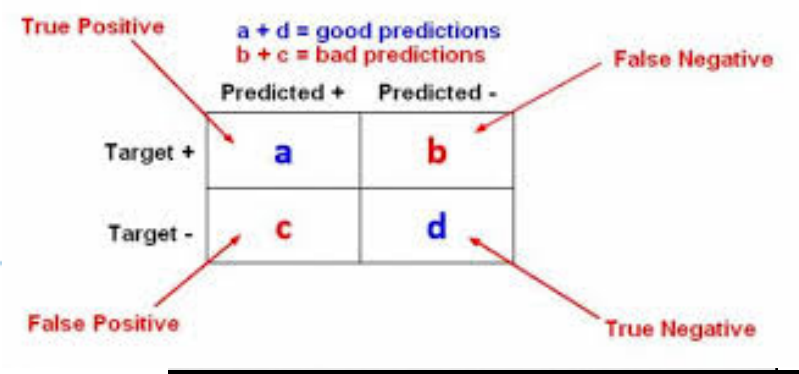


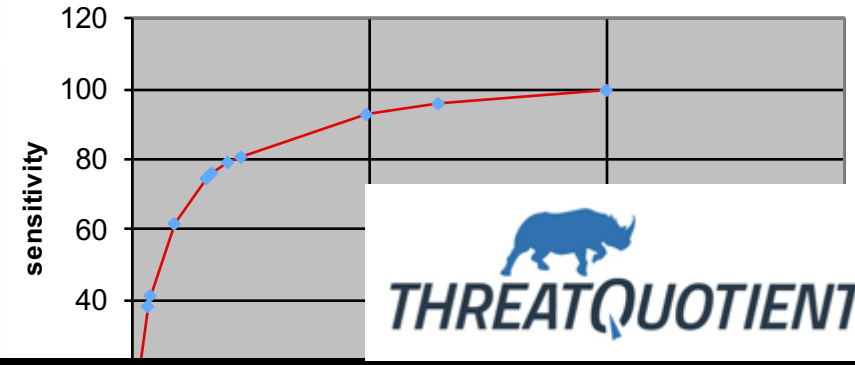
$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}.$$









ROC Curve

THREATQUOTIENT

Nir Yosha – Threat Intel Engineer

Agenda

- What is Threat Intel?
- Main Challenges with Threat Intel
- Can Statistics Help?
- Confidence Level
- Conditional Probability
- Correlation
- Scoring
- Summary

Nir Yosha

Israeli Intelligence Corps

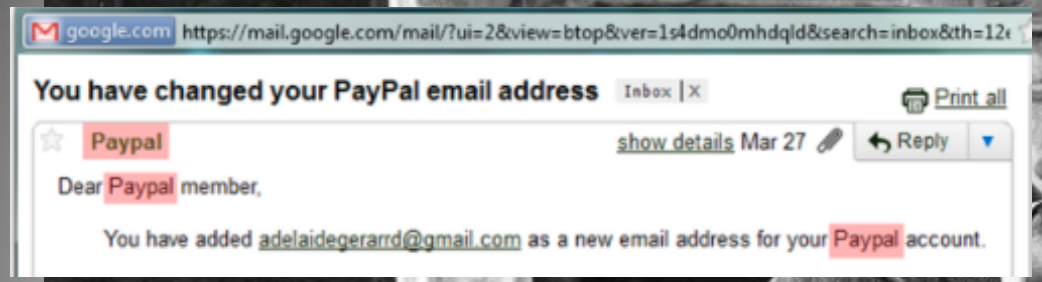
Technion
Israel Institute of Technology

Threat Intel Platform

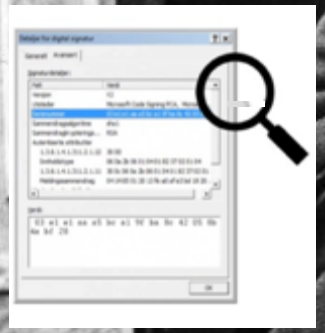
Network Security

User Behavior Analytics

Indicators of Compromise (IOCs)



Domain name
haartezenglish.strangled.net
wallanews.sytes.net
ynet.sytes.net



MD5
A5DE87646EE943CD1F448A67FDBE2817
F982401E46864F640BCAEDC200319109
EC5B360F5FF6251A08A14A2E95C4CAA4
97576FA7A236679DBE3ABE1A4E852026
C1EC435E97A4A4C5585392D738B5879F
2559FE4EB88561138CE292DF5D0E099F
0ABF3FA976372CBC8BF33162795E42A8

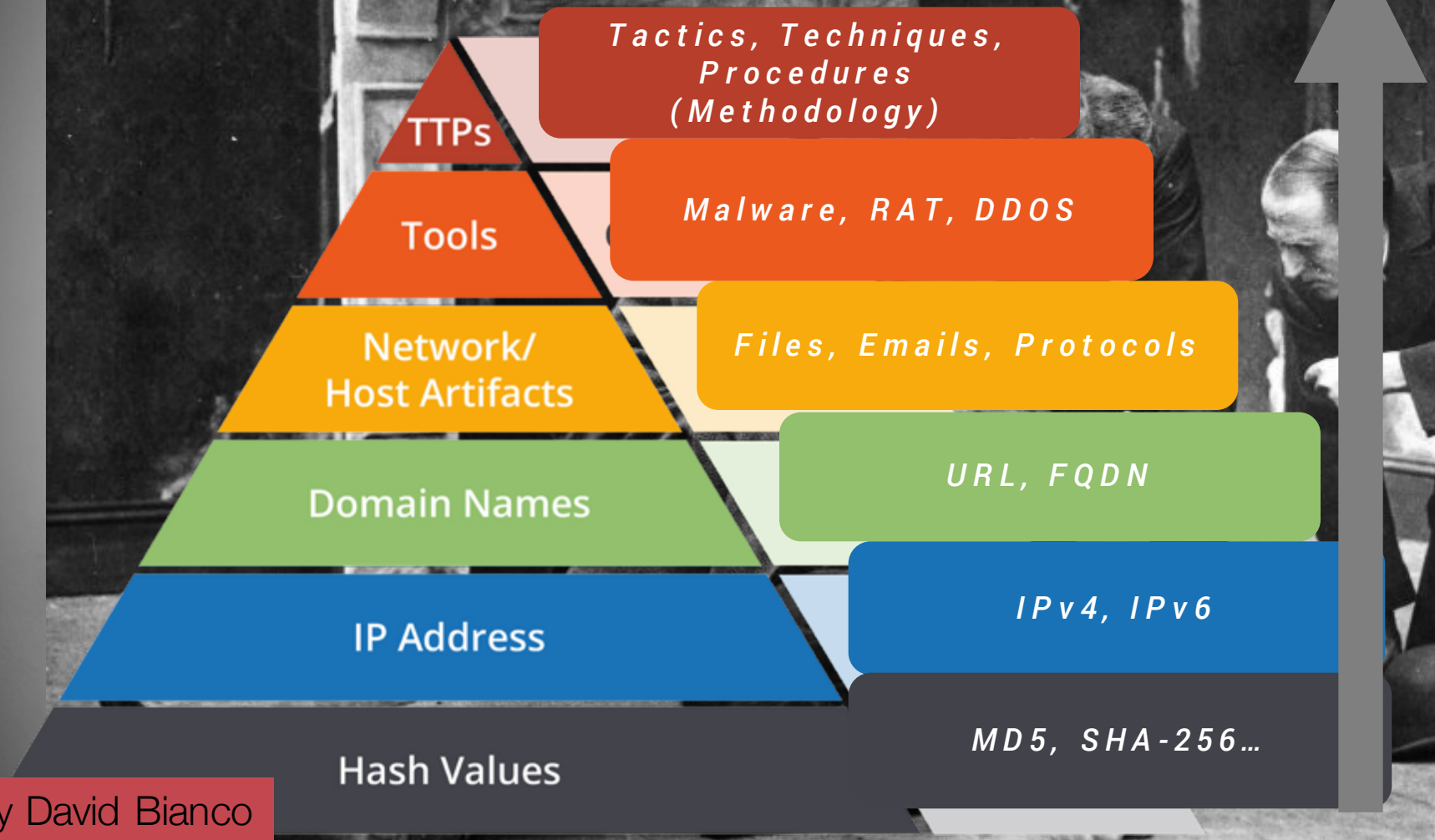
204.152.203.99

ATTRIBUTES	
First Seen	2010-04-19 15:20:55
Last Seen	2016-05-29 20:50:18
Resolutions	145
Network	204.152.192.0/19
AS	8100 (IPTELLIGENT - IPTelligent LLC)

Base64 string
2YXYrtin2KjYsdin2Ko=
2KjYp9iz2KjZiNix2K/Yp9iq
Y3YuZG9j
157Xktei15nXnQ==
2LPZitix2Kkg2LDYp9iq2YrYqQ==
cGFzc3dvcmRz
16HXmdeh157XkNeV16o=
INeR15nXmNeX15XXnyDXpNeg15nXnQ==
d29ybQ==
bXljZXJ0
LnBmeA==

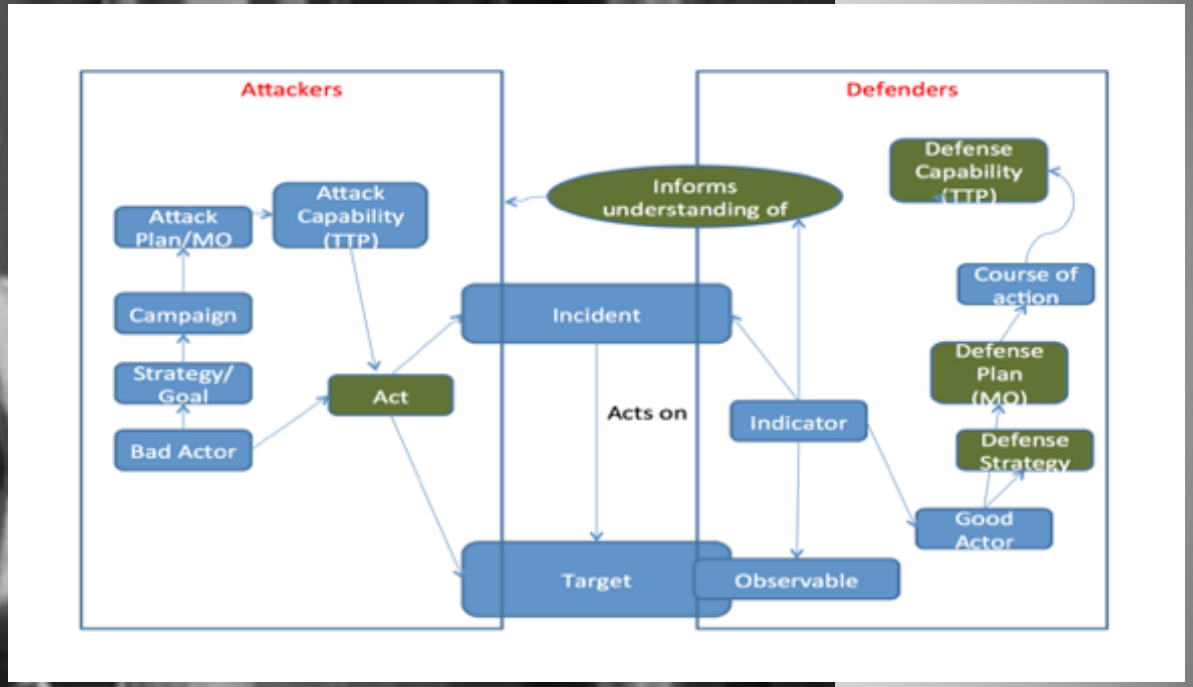
- RECONNAISSANCE
- WEAPONIZATION
- DELIVERY
- EXPLOITATION
- INSTALLATION
- COMMAND & CONTROL
- ACTION ON OBJECTIVES

Pyramid of Pain



by David Bianco

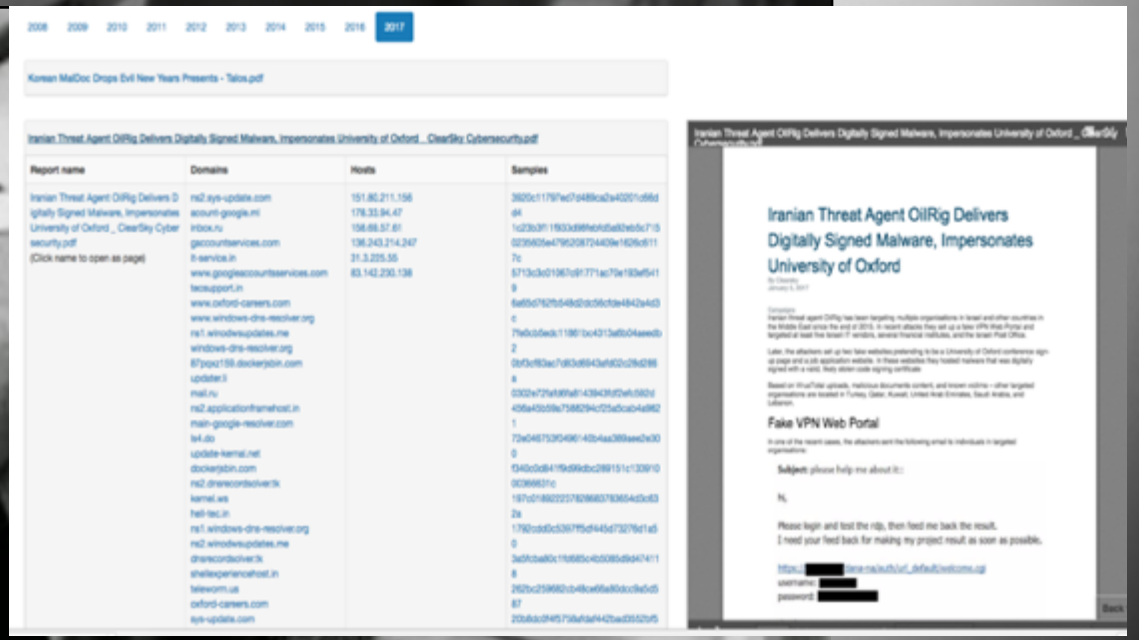
Threat Intelligence



Set of data collected, assessed and applied regarding security threats, threat actors, exploits, malware, vulnerabilities, indicators of compromise.

Threat Info ≠ Threat Intel

Threat Information	Threat Intelligence
Raw, unfiltered feed	Processed information
Not enriched	Normalized & enriched
Not actionable	Actionable



Multiple sources of data

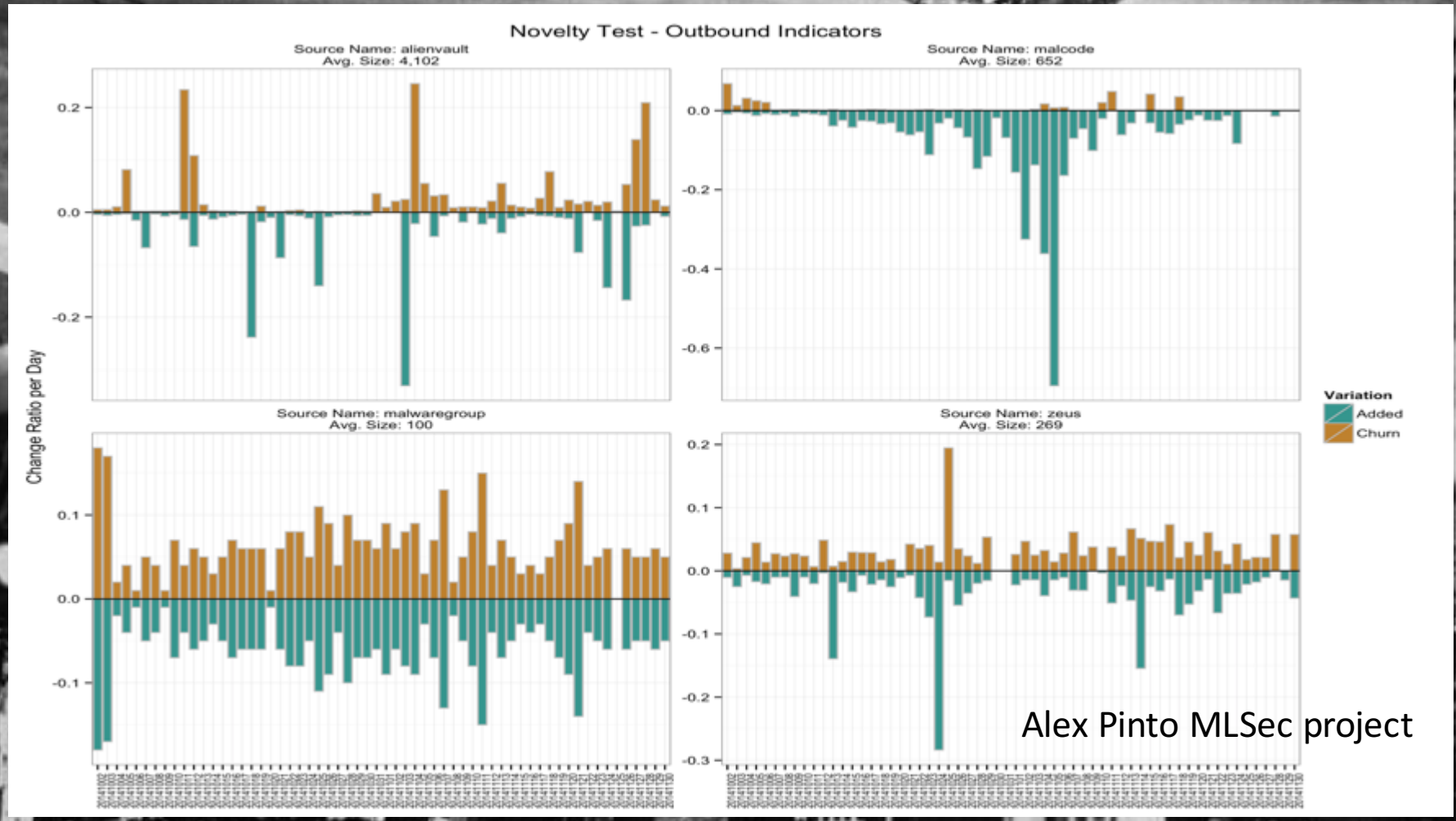
- OS
- Community
- Commercial
- Enrichment
- Internal

- TTP
- Victims
- Infrastructure
- IOCs (Indicators)
- Adversary (Actor)

Main Challenges with Threat Intel

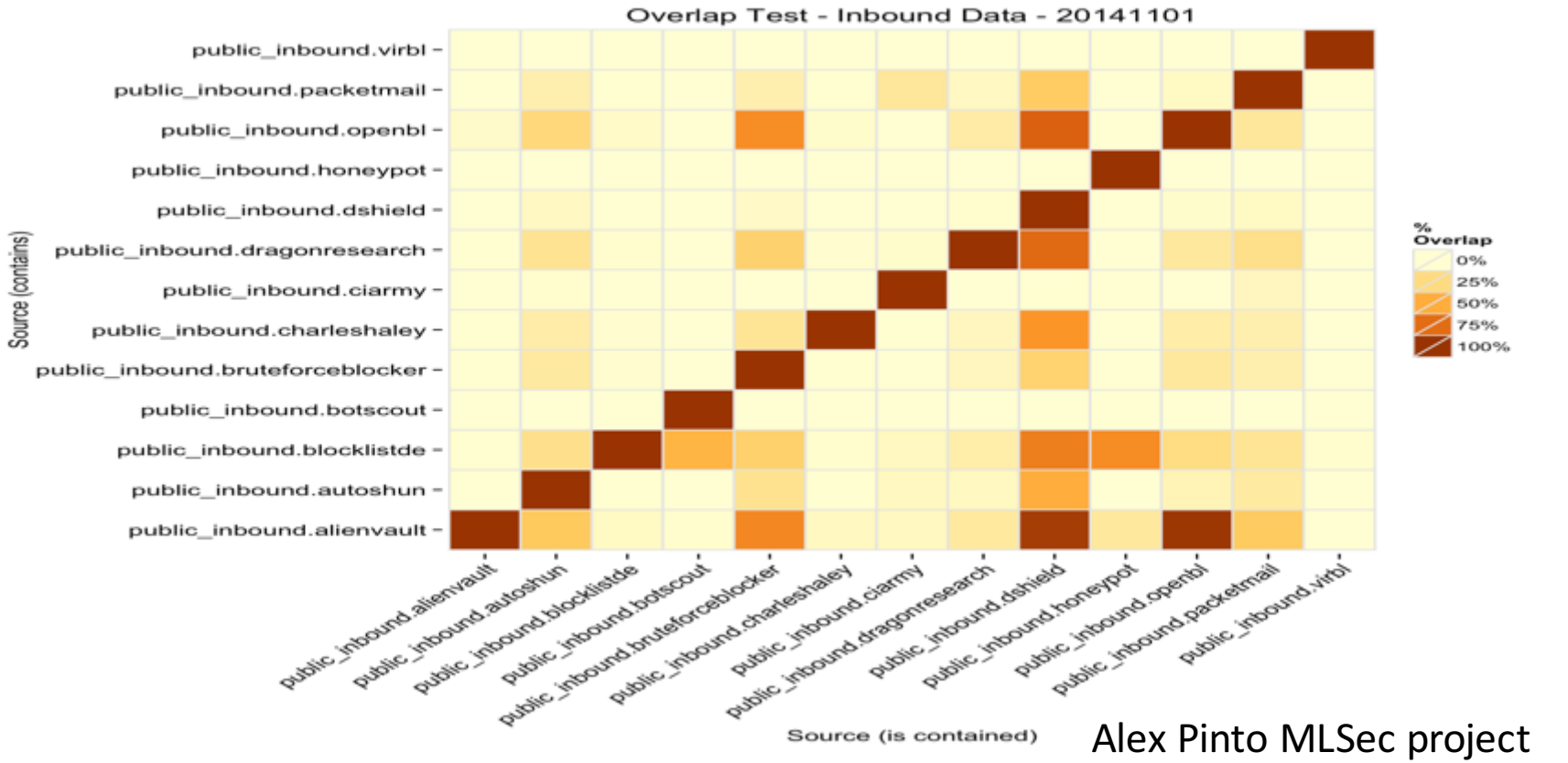
- Low confidence in data sources
- Non relevant information
- Old and duplicated data
- Data with no context (IP of what??)
- Different ‘opinions’ across Threat Intel Sources

Novelty test



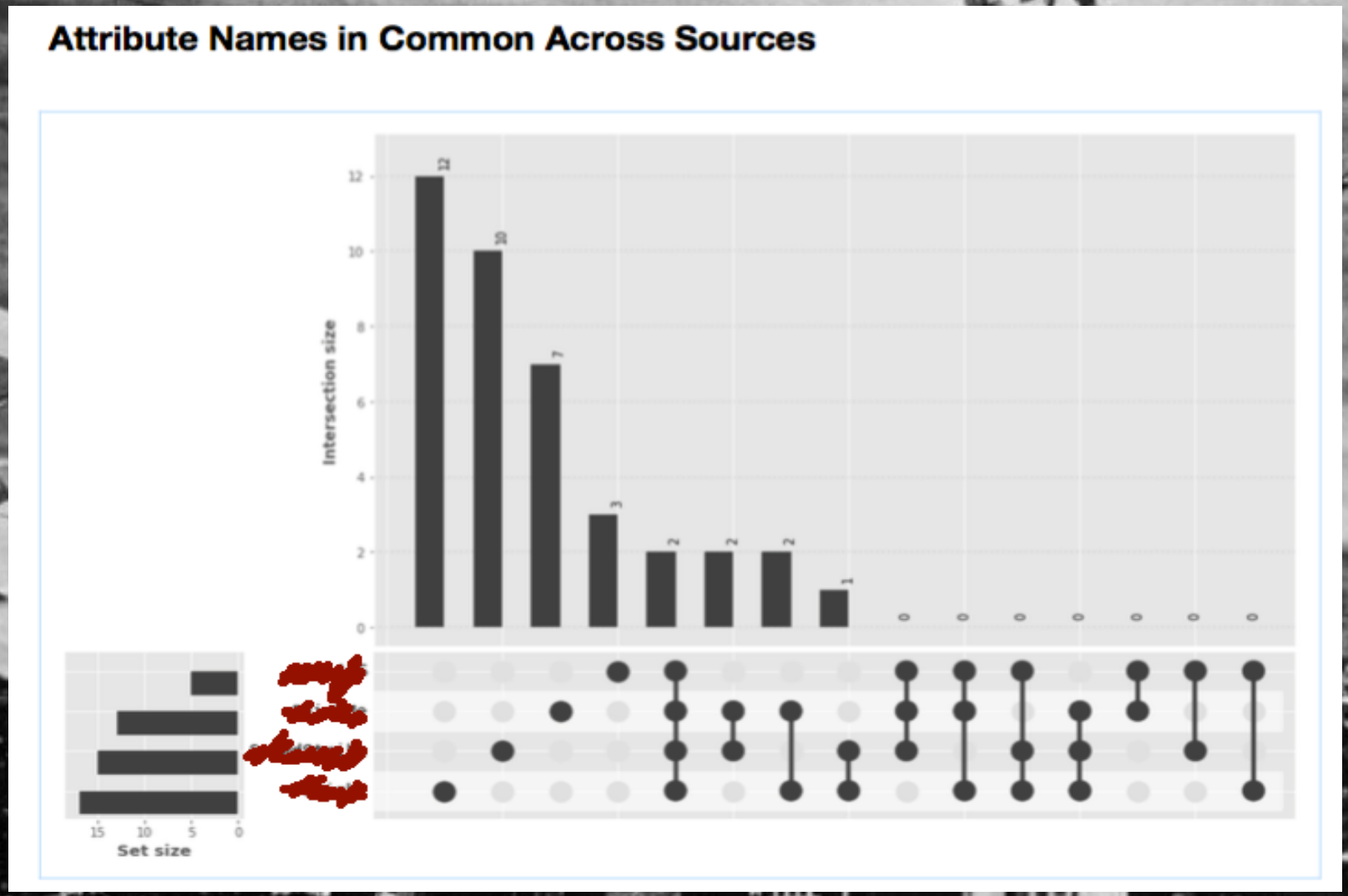
Alex Pinto MLSec project

Overlap test

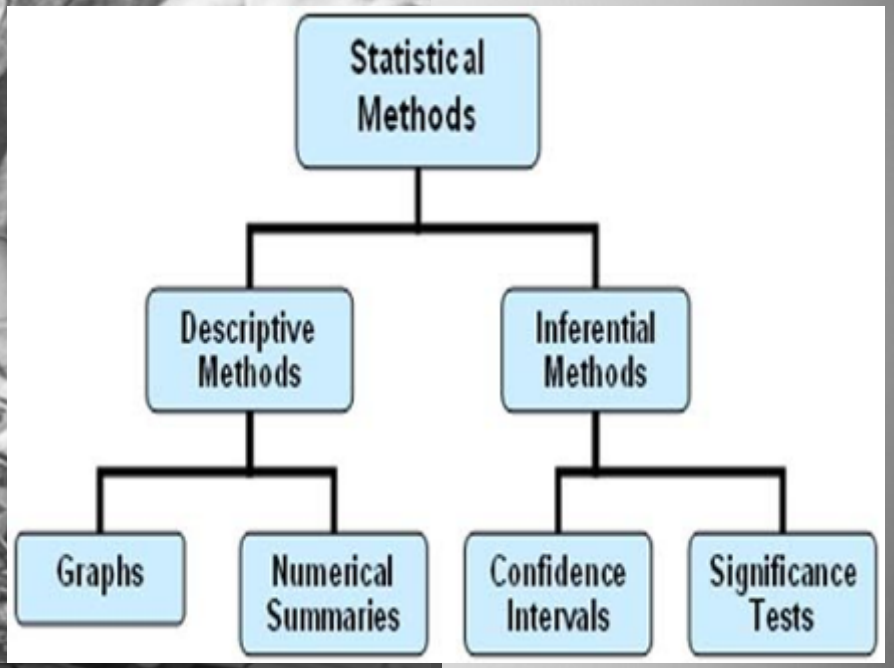


Alex Pinto MLSec project

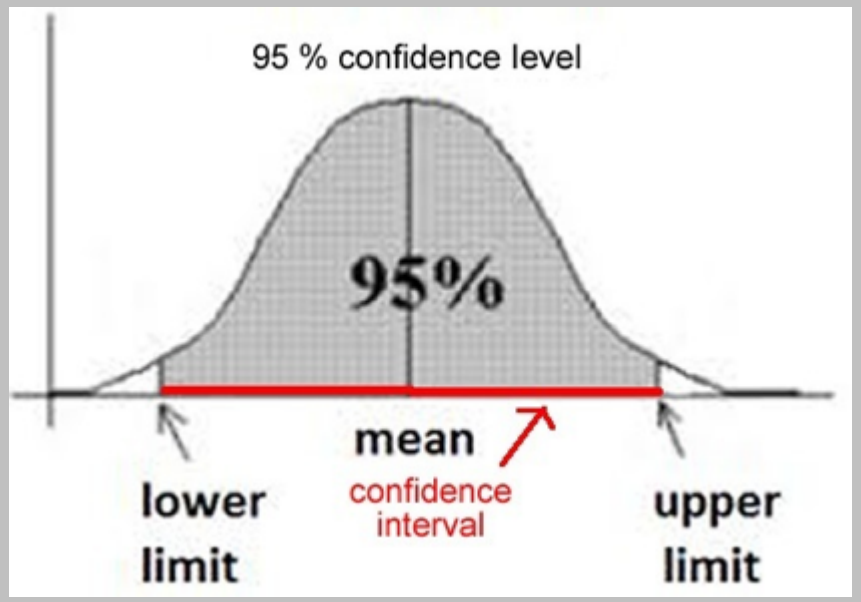
Vendor specific terminology



Can Statistics Help Threat Intel?



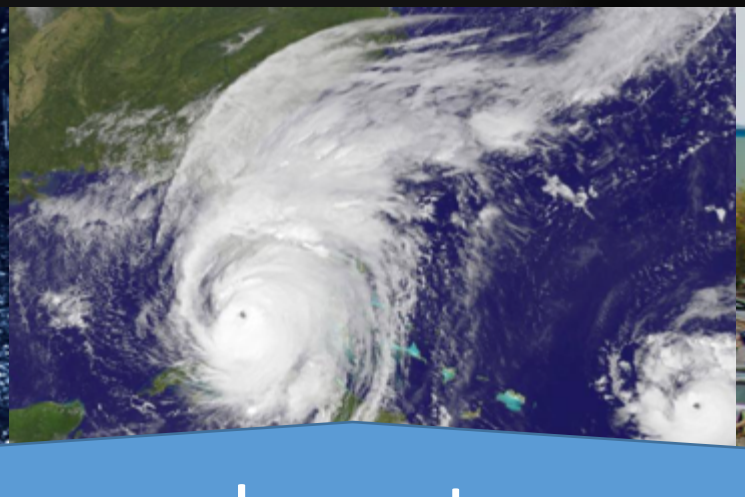
Collection, analysis, interpretation, and presentation of masses of numerical data.


95 % confidence level
95%
mean
confidence interval
lower limit
upper limit


IT'S OVER.
AFTER TWENTY MONTHS IT'S FINALLY OVER.
I DON'T HAVE TO BE AN ELECTION JUNKIE ANYMORE.
I DON'T HAVE TO CARE ABOUT OPINION POLLS, EXIT POLLS, MARGINS OF ERROR, ATTACK ADS, GAME-CHANGERS, TRACKING POLLS, SWING STATES, SWING VOTES, THE BRADLEY EFFECT, OR <NAME> THE <OCCUPATION>.
I'M FREE.
Google
2012 polling statistics

Confidence

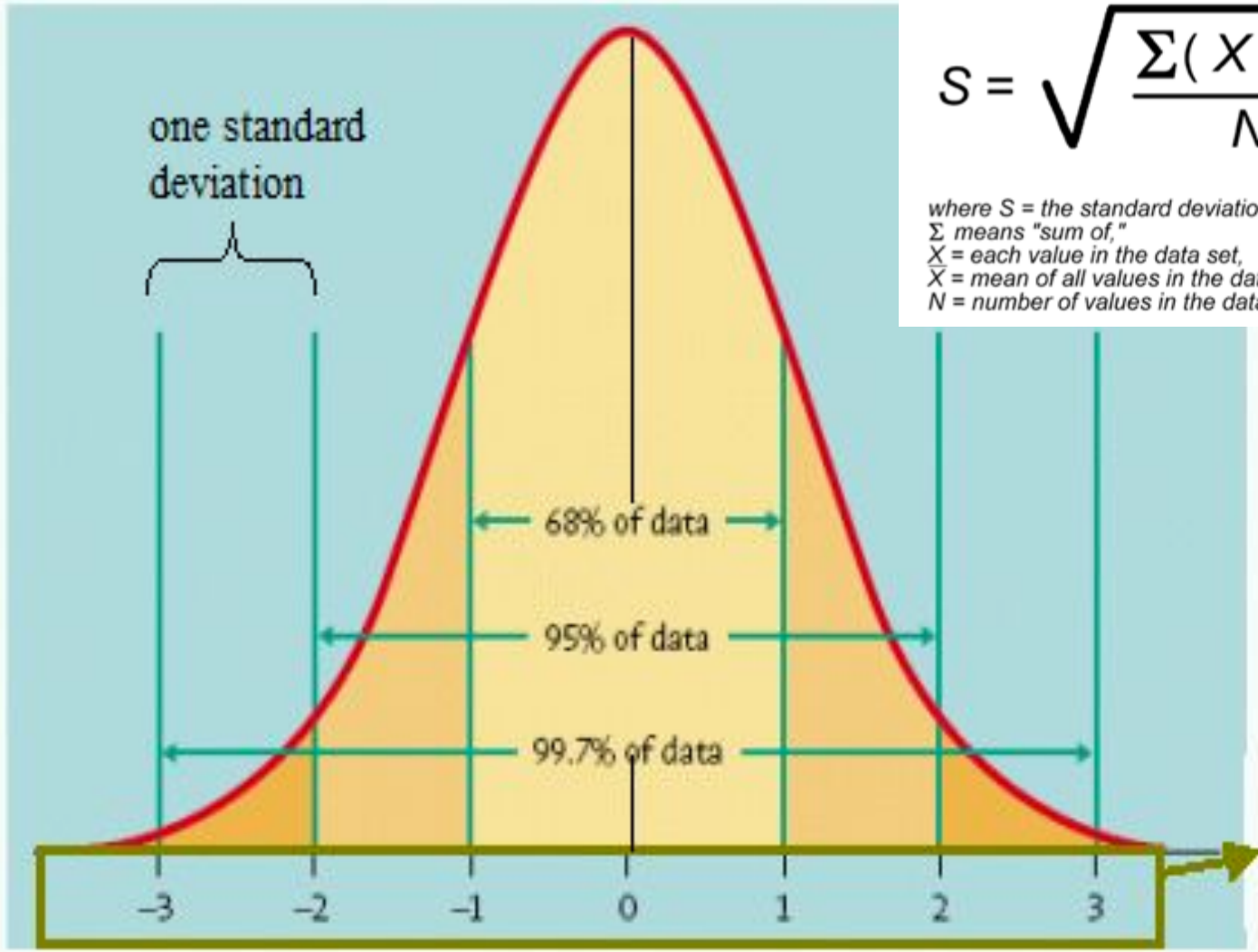
Confidence ≠ Impact



Impact

Margin of Error

Z-score is the number of standard deviations from the mean a data point is.

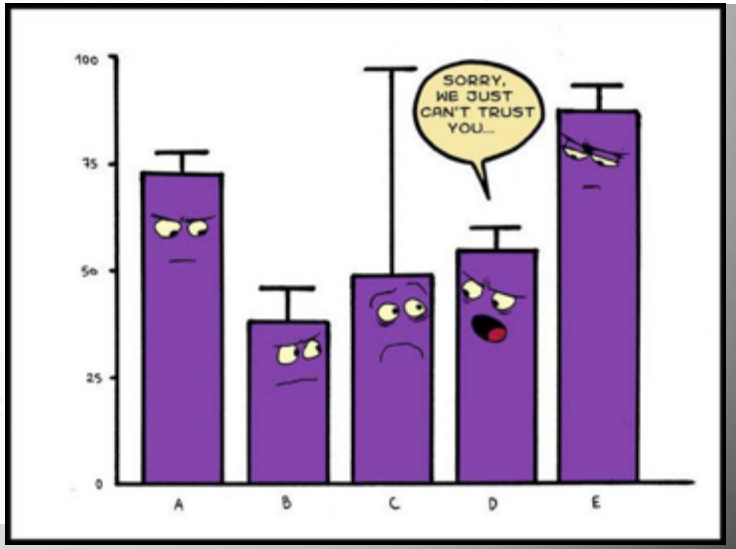


$$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\bar{X}$ = mean of all values in the data set,
N = number of values in the data set.

Confidence Level	Confidence Coefficient, $1-\alpha$	z value, $z_{\alpha/2}$
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

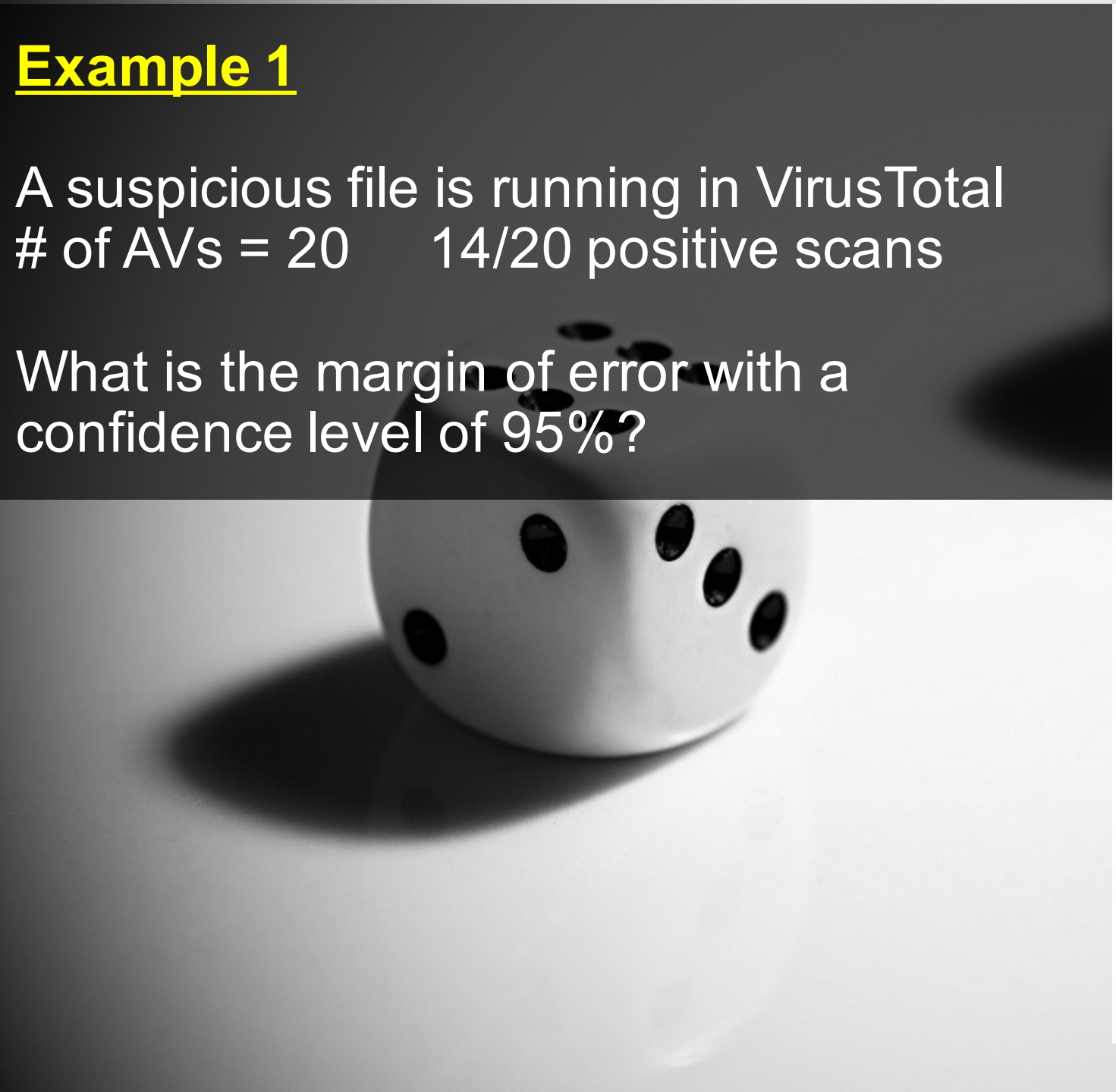
$$z = \frac{x - \mu}{\sigma}$$



Example 1

A suspicious file is running in VirusTotal
# of AVs = 20 14/20 positive scans

What is the margin of error with a confidence level of 95%?



virustotal

SHA256: b459f10d09a24918704c12d62af9b467e2f7d9726c6ebc632122b88990c1bec1

File name: 5804993f8cf083173707dfa1f6911b9d.exe

Detection ratio: 2 / 56

Analysis date: 2015-04-07 05:01:26 UTC (3 days, 11 hours ago)

0 0

Analysis | File detail | Additional information | Comments 0 | Votes | Behavioural information

Antivirus	Result	Update
AVG	Opera Software ASA	20150407
Sophos	Mal/EncPk-AAK	20150407
ALYac	✓	20150407
AVware	✓	20150407
Ad-Aware	✓	20150407
AegisLab	✓	20150407

Margin of Error Formulas

$$ME = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

$$ME = 1.96\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

for 95% confidence interval

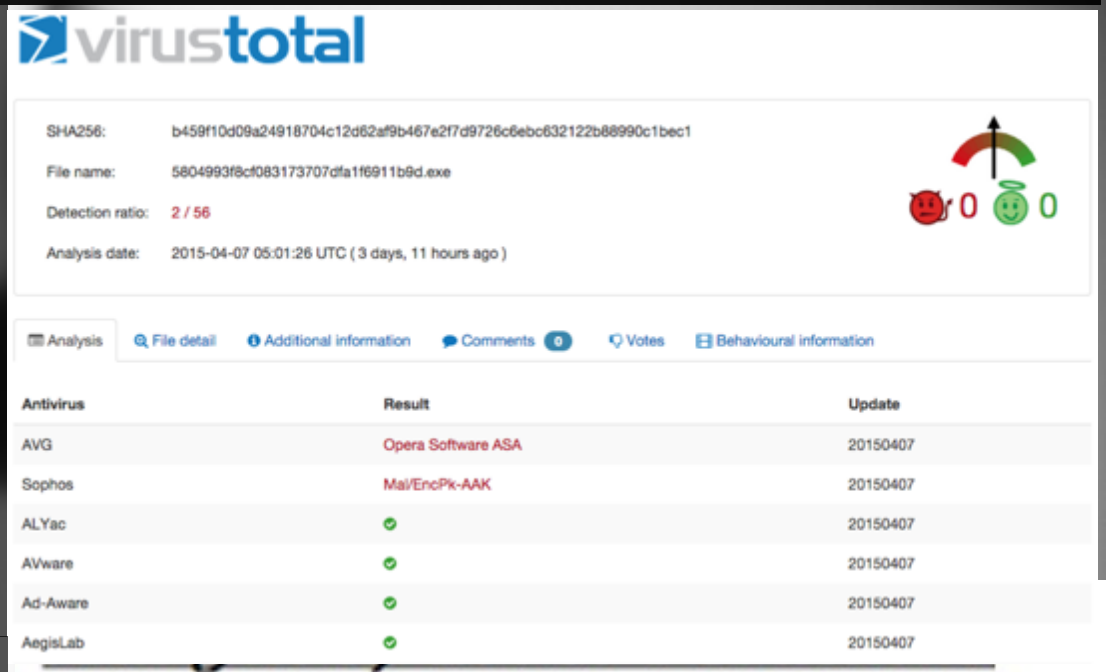
Size does matter!

Solution 1

P(Malware) = 14/20 = 0.7
Z Score of 95% = 1.96 (normal dis.)
# of AVs = 20
ME = $1.96\sqrt{(0.3x0.7/20)}$ = 0.2 = 20%
50% < P(Malware) < 90%

What if we increase the sample size?

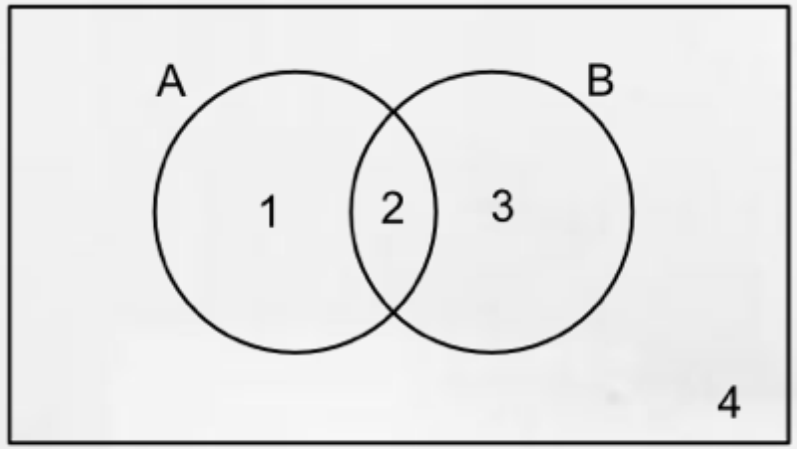
P = 140/200 = 0.7
Z Score of 95% = 1.96 (normal dis.)
# of AVs = 200
ME = $1.96\sqrt{(0.3x0.7/200)}$ = 0.063= 6.3%
63.7% < P(Malware) < 76.3%



$$ME = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

$$ME = 1.96\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

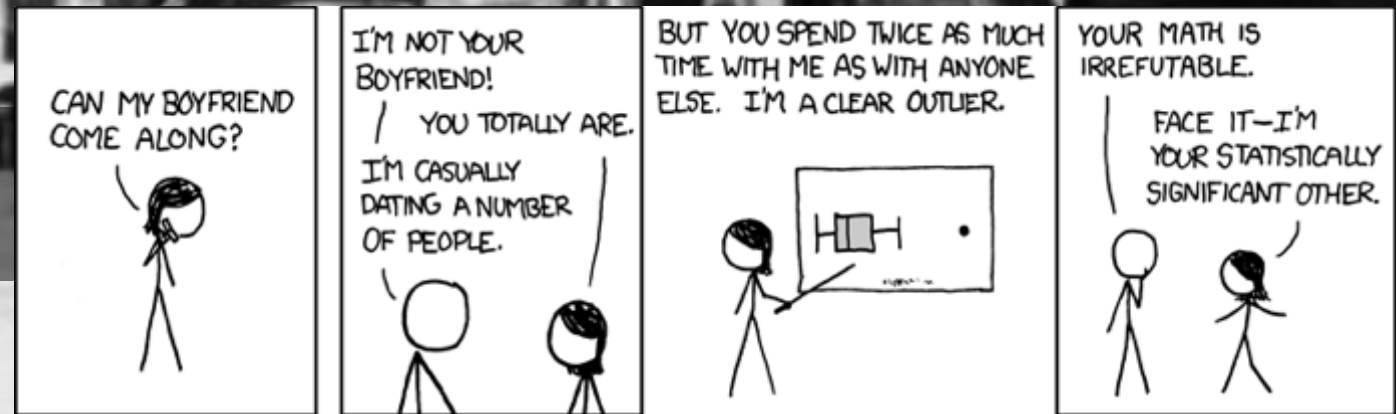
for 95% confidence interval



P (A | B) is A *given* B

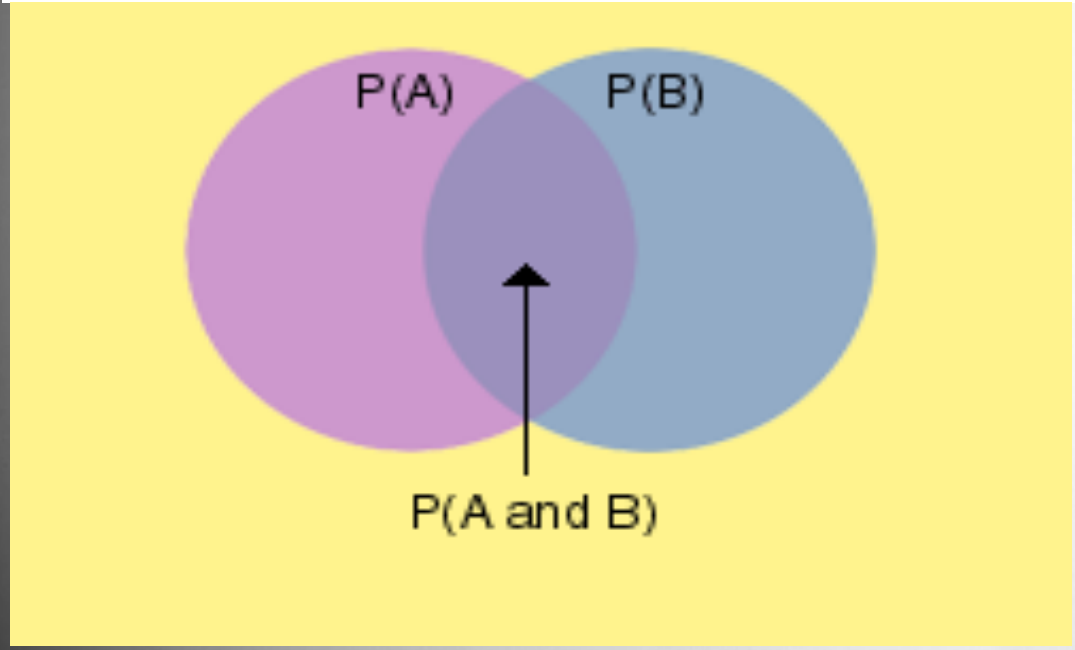
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{2+1} = \frac{2}{3}$$

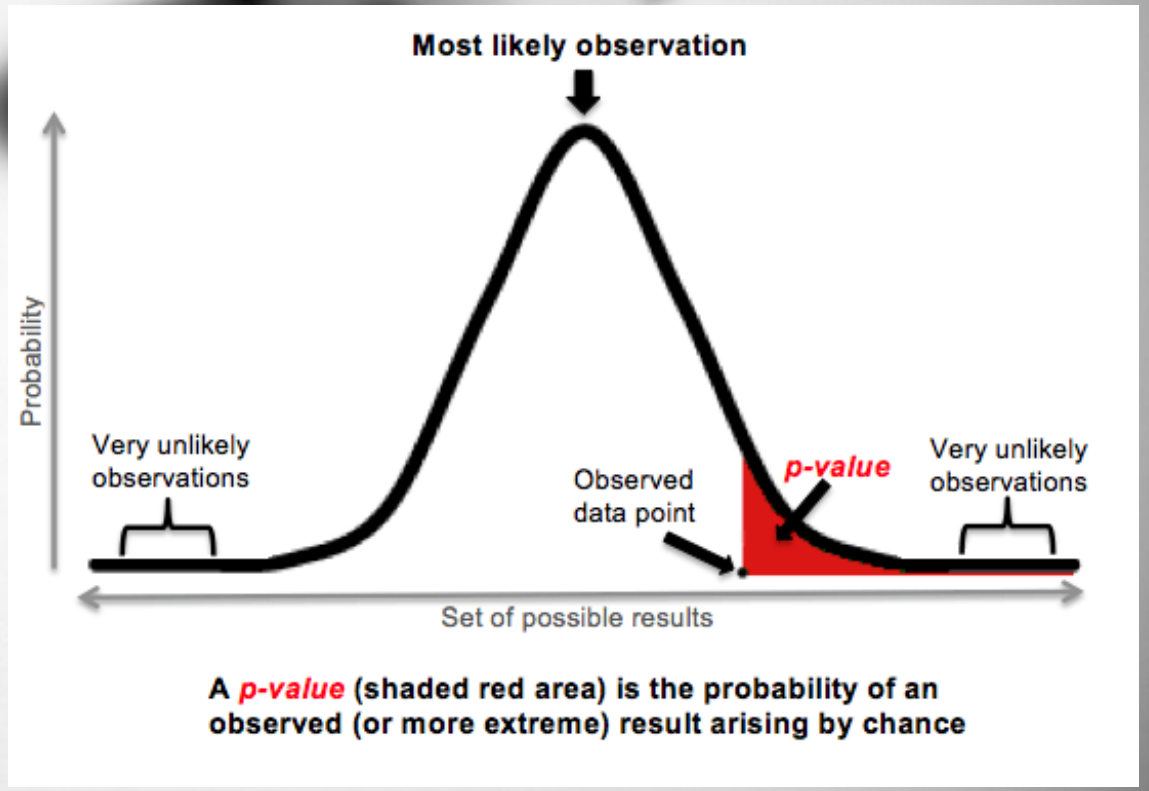
Probability



Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

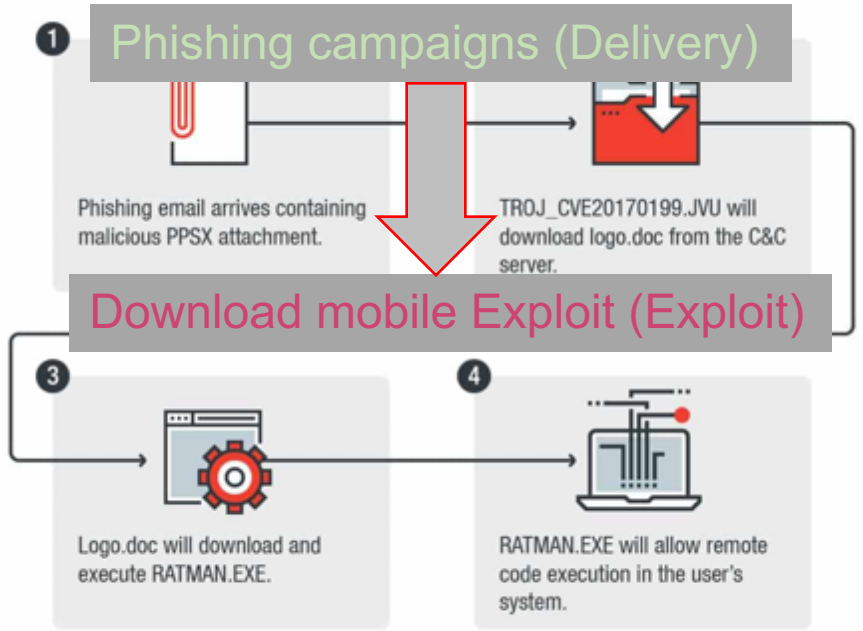


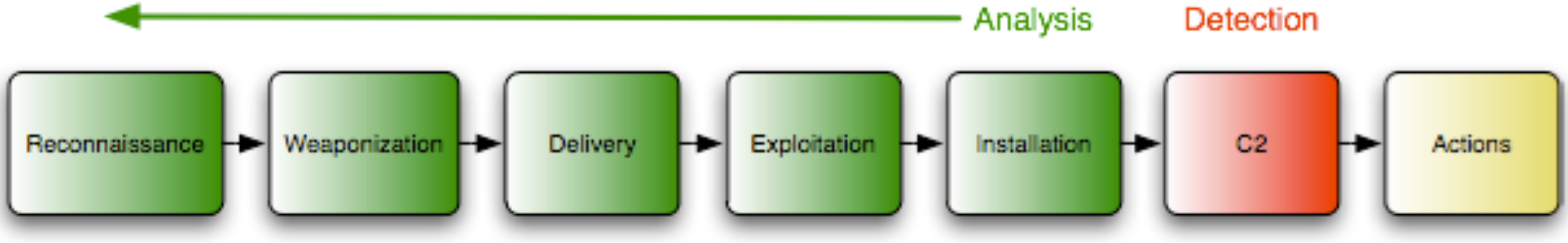


Example 2

Out of 100 incidents under investigation 40 are a result of phishing campaigns (Delivery) and 30 are mobile Exploit (Exploit).

What is the probability of a mobile malware giving that its delivery method is phishing campaign?





Improve accuracy with time

Solution 2

P(A and B) = P(A) x P(B) =>

0.3 x 0.4 = 0.12 = 12%

P(A|B) = P(A AND B)/P(B) =>

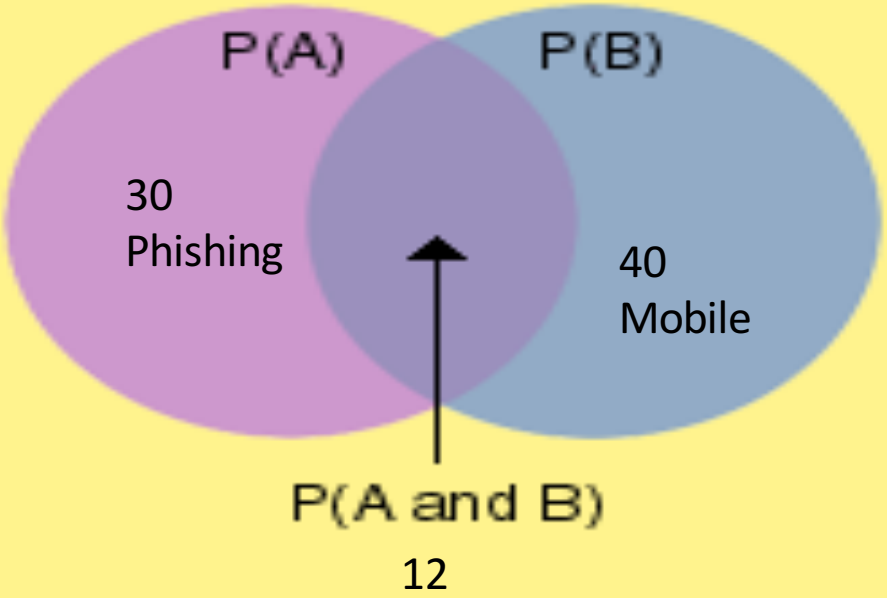
12/40 = 30%

What if a sample data includes the following?

P(A AND B) = 20% (known)

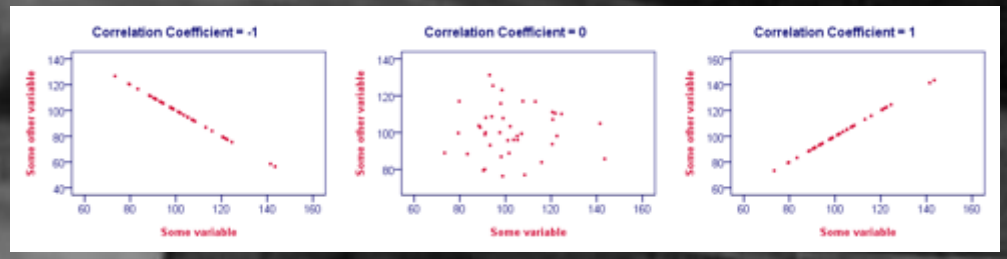
P(A|B) = P(A AND B)/P(B) =>

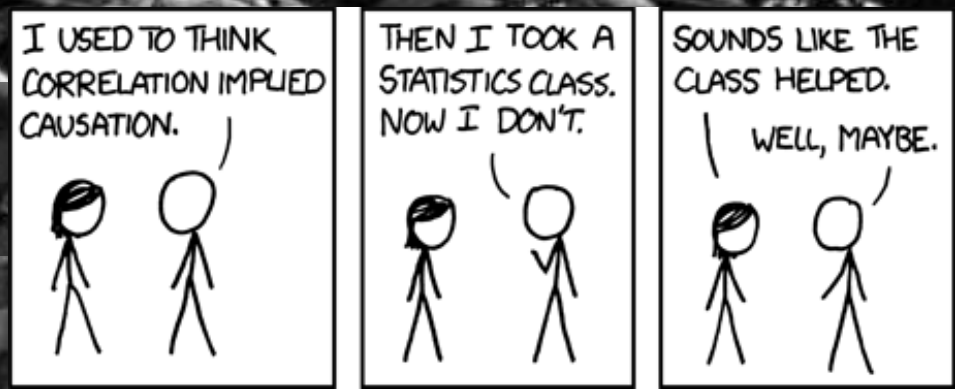
20%/40% = 50%



$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Correlation





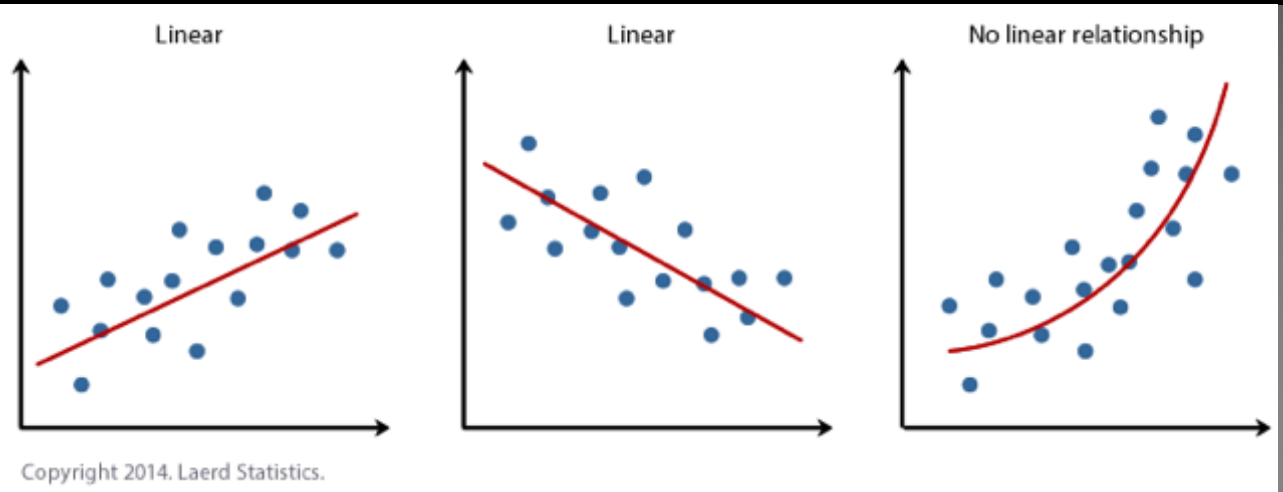
Regression and Forecasting

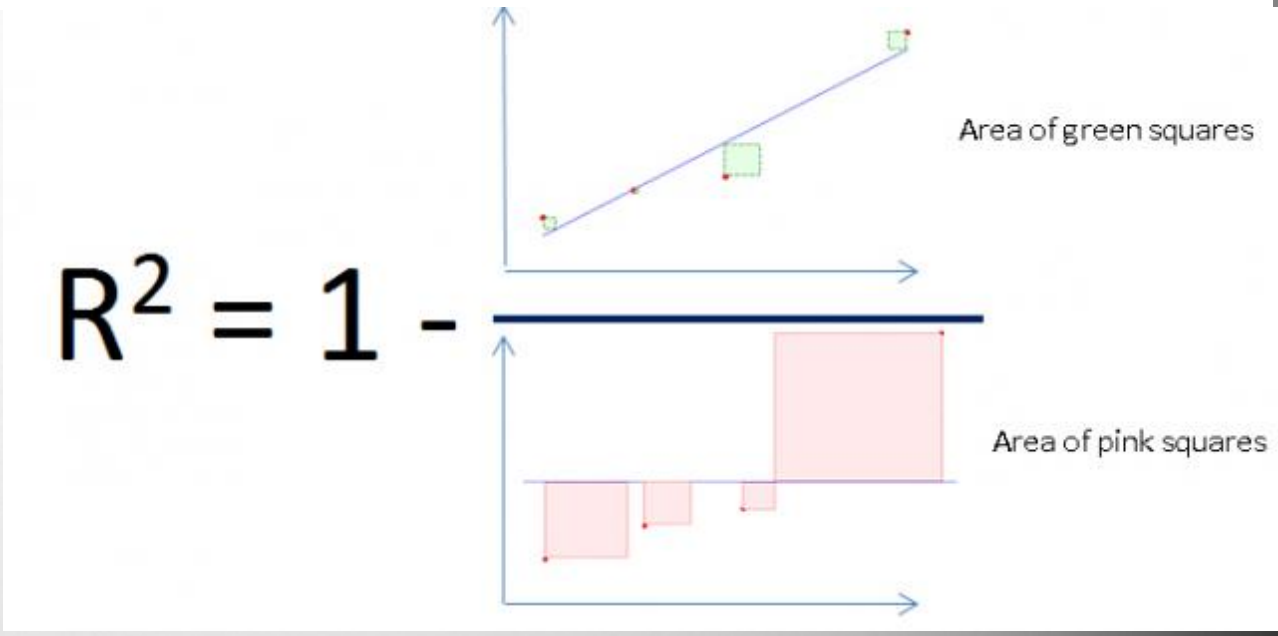
Regression Statistics	
Multiple R	0.85
R Square	0.73
Adjusted R Square	0.68
Standard Error	7.83
Observations	15.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00	1984.27	992.13	16.19	0.00
Residual	12.00	735.33	61.28		
Total	14.00	2719.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	419.95	37.40	11.23	0.00	338.46	501.45
Price	-42.80	8.30	-5.15	0.00	-60.89	-24.70
Cprice	4.39	9.74	0.45	0.66	-16.82	25.60

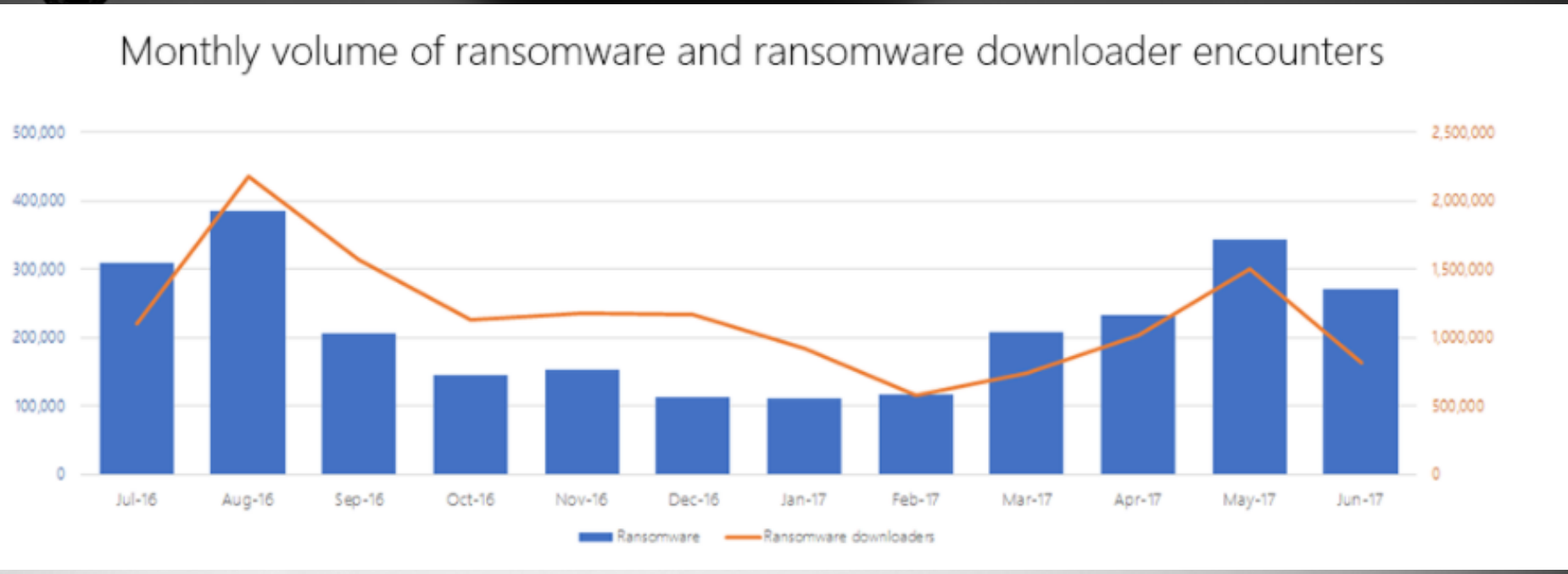




Example 3

Predict the number of incidents in July based on the past 6 month. Use moving average, smoothing, seasoning and regression.

Month	Incidents
Jan	*1,325*
Feb	*1,353*
Mar	*1,305*
Apr	*1,275*
May	*1,210*
Jun	*1,195*
Jul	*?*



Solution 3

include trend analysis in exponential smoothing.

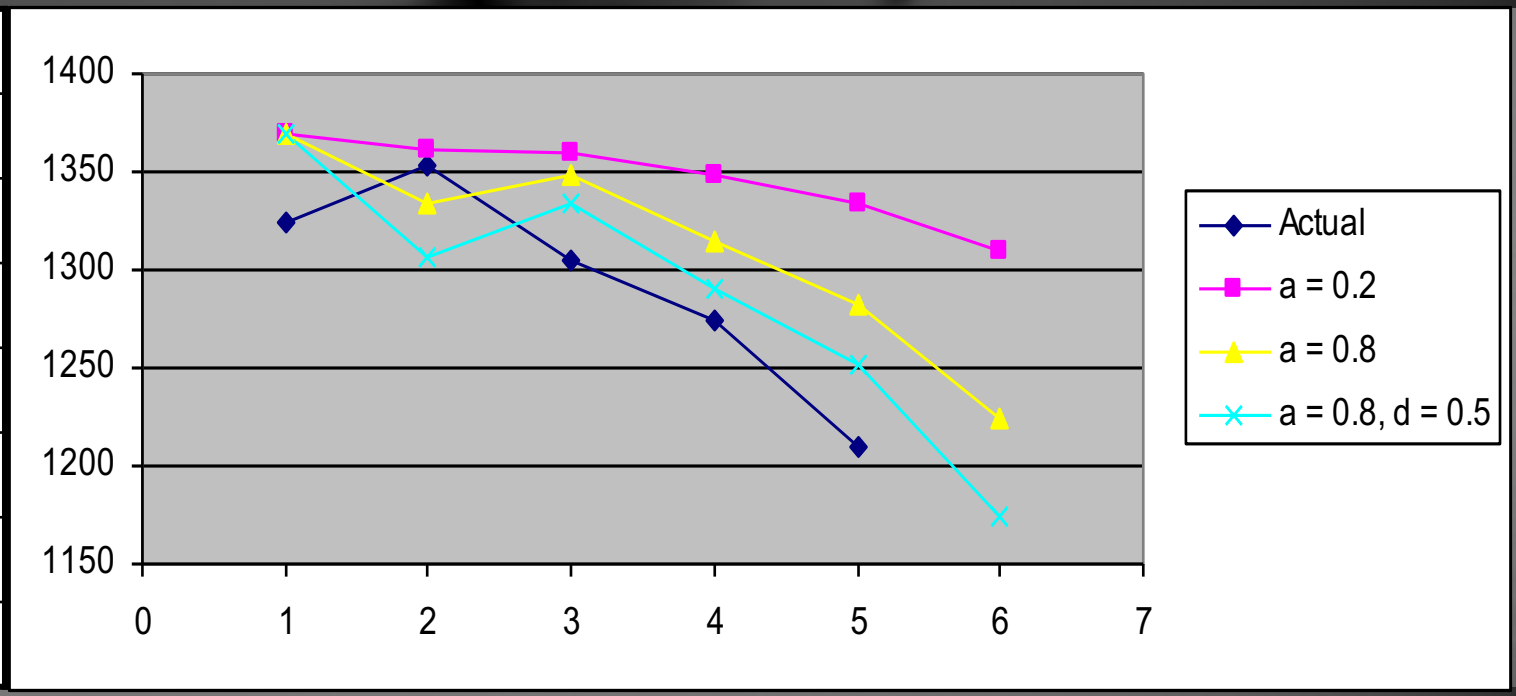
$$T_t = T_{t-1} + \delta(F_t - FIT_{t-1})$$

FIT: Forecast including trend
δ: Trend smoothing constant

$$FIT_t = F_t + T_t$$

$$F_t = FIT_{t-1} + \alpha(A_{t-1} - FIT_{t-1})$$

Month	Incidents
Jan	*1,325*
Feb	*1,353*
Mar	*1,305*
Apr	*1,275*
May	*1,210*
Jun	*1,195*
Jul	*?*



Regular exponential smoothing will always lag behind the trend.

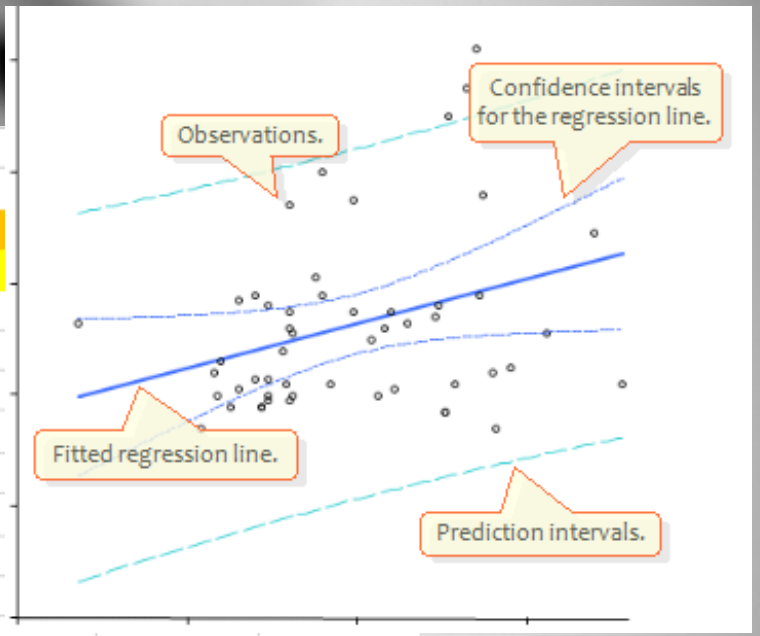
F is the forecast without trend and T is the trend component

Regression

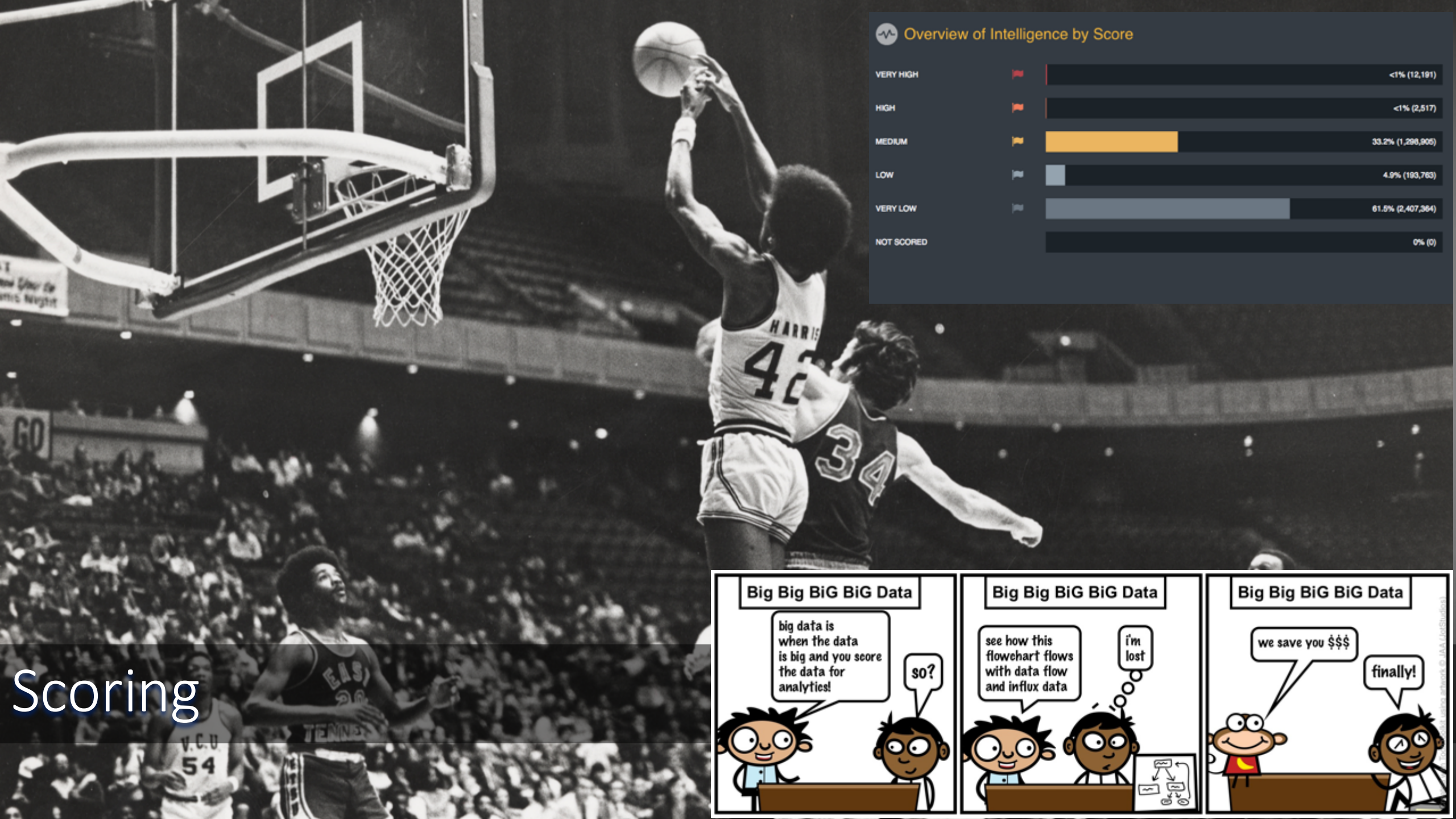
Influence Score based on:

Indicator Type | Indicator Source | Attributes | Adversary Relationship

Regression Statistics		Linear Regression for Indicator Score						
Multiple R	0.645712437	Predictors - # of Attribtues, # of Sources, # of related adversaries						
R Square	0.416944552							
Adjusted R Square	0.412105917							
Standard Error	1.120229993							
Observations	487							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	432.5435862	108.135897	86.16988086	3.48716E-55			
Residual	482	604.8691448	1.25491524					
Total	486	1037.412731						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.32248956	0.280393065	22.548666	2.10687E-77	5.771545821	6.8734333	5.77154582	6.8734333
attrb_total	-0.583131574	0.036721225	-15.87996	5.59042E-46	-0.655285031	-0.5109781	-0.655285	-0.5109781
related_sources	3.140066752	0.250619197	12.5292347	2.25903E-31	2.647625623	3.63250788	2.64762562	3.63250788
related_adv	-0.304786318	0.053510731	-5.695798	2.13846E-08	-0.409929441	-0.1996432	-0.4099294	-0.1996432
related_ind	0.005577054	0.00177405	3.1436849	0.001771244	0.002091227	0.00906288	0.00209123	0.00906288
Eg. Estimated score for indicator with 6 attributes, 2 sources and 20 related indicators -> Y = 6.32 - 0.58X6 + 3.14X2 + 0.005X20						=>	9.12	



Scoring


Overview of Intelligence by Score
VERY HIGH
<1% (12,191)
HIGH
<1% (2,517)
MEDIUM
33.2% (1,298,905)
LOW
4.9% (193,763)
VERY LOW
61.5% (2,407,364)
NOT SCORED
0% (0)

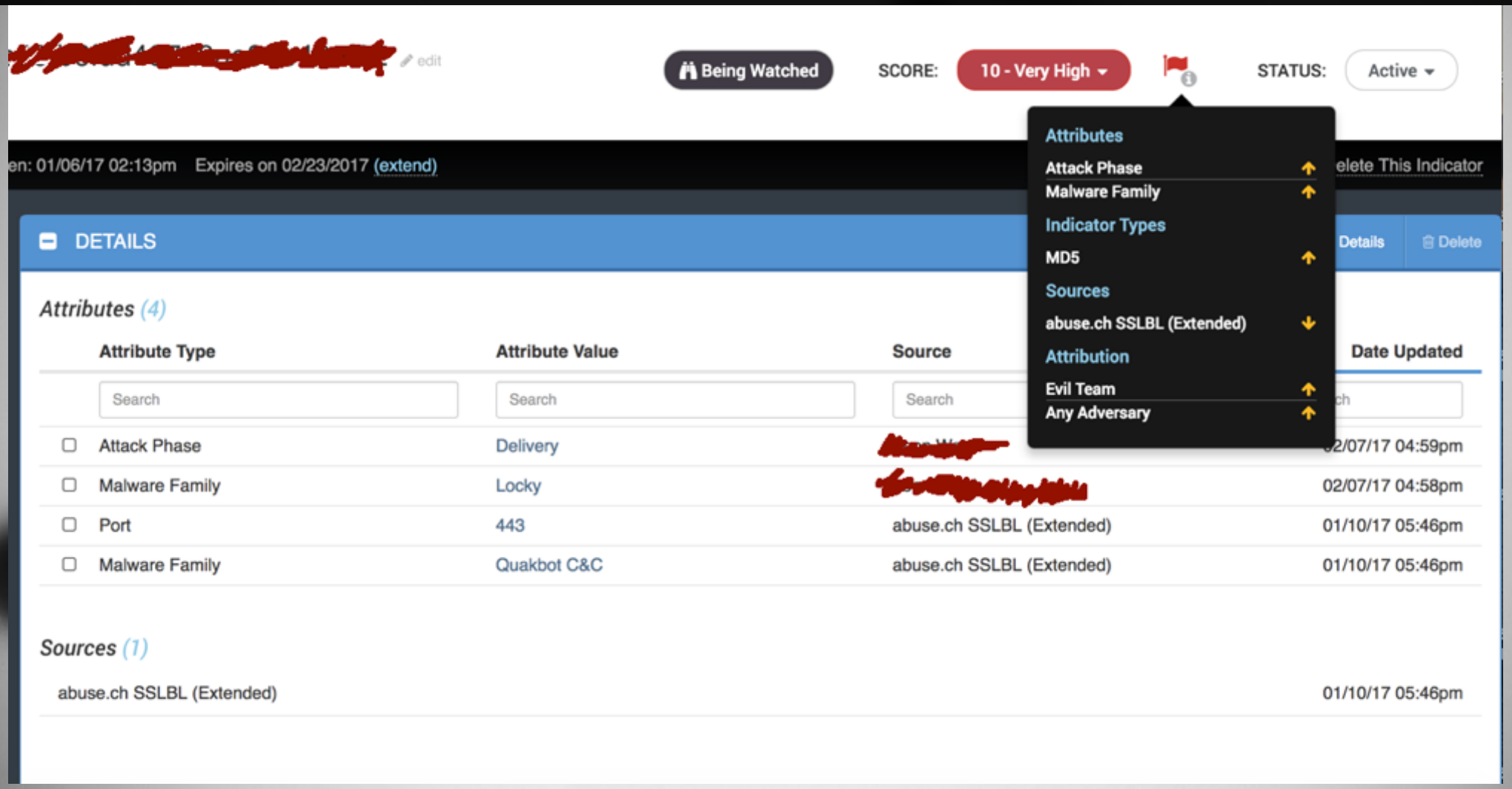
Big Big BiG BiG Data
big data is when the data is big and you score the data for analytics!
so?
Big Big BiG BiG Data
see how this flowchart flows with data flow and influx data
i'm lost
Big Big BiG BiG Data
we save you $$$
finally!

Scoring Categories

Threat agent factors					Vulnerability factors			
Skill level	Motive	Opportunity	Size		Ease of discovery	Ease of exploit	Awareness	Intrusion detection
5	2	7	1		3	6	9	2
Overall likelihood=4.375 (MEDIUM)								

Technical Impact					Business Impact			
Loss of confidentiality	Loss of integrity	Loss of availability	Loss of accountability		Financial damage	Reputation damage	Non-compliance	Privacy violation
9	7	5	8		1	2	1	5
Overall technical impact=7.25 (HIGH)					Overall business impact=2.25 (LOW)			

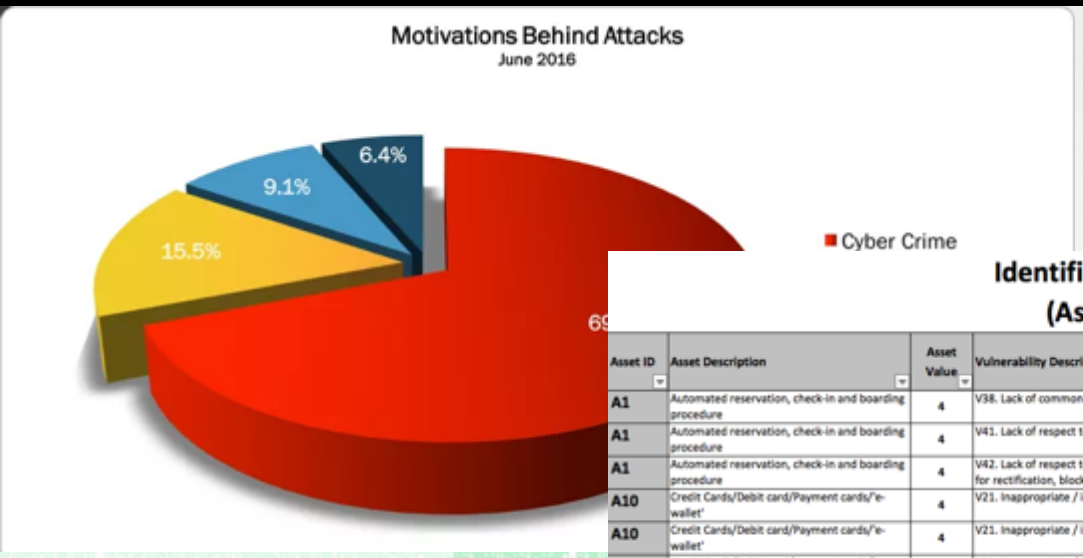
Scoring IOCs



IOC Classification

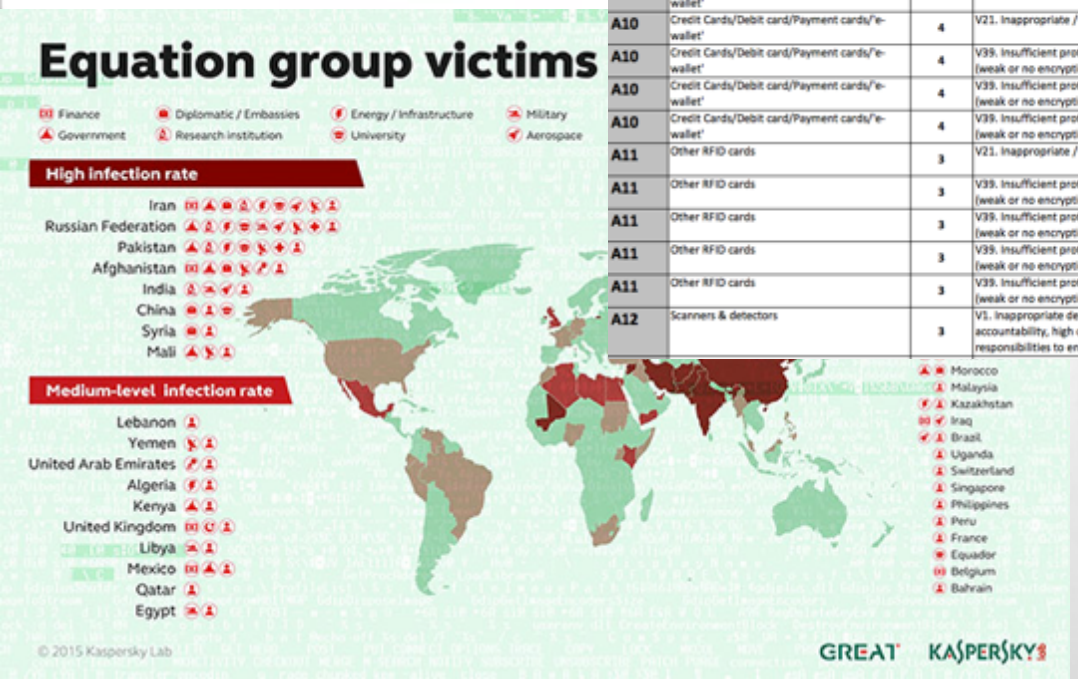
Classifiers	Example	Why
Indicator Type	IP address, URL, Domain, File	Support different security devices
Attributes	Language, Country, Malware family	Used by most vendors
Adversary	Industry, Government, PII, PHI, PCI	Relevant adversaries to your vertical / endpoints
Incidents or Events	Age, Owner, ASN, Day of Week, User ID	Connect the dots across the kill chain
Relevance	CVE, OS, User, Brand	Relevant to your environment

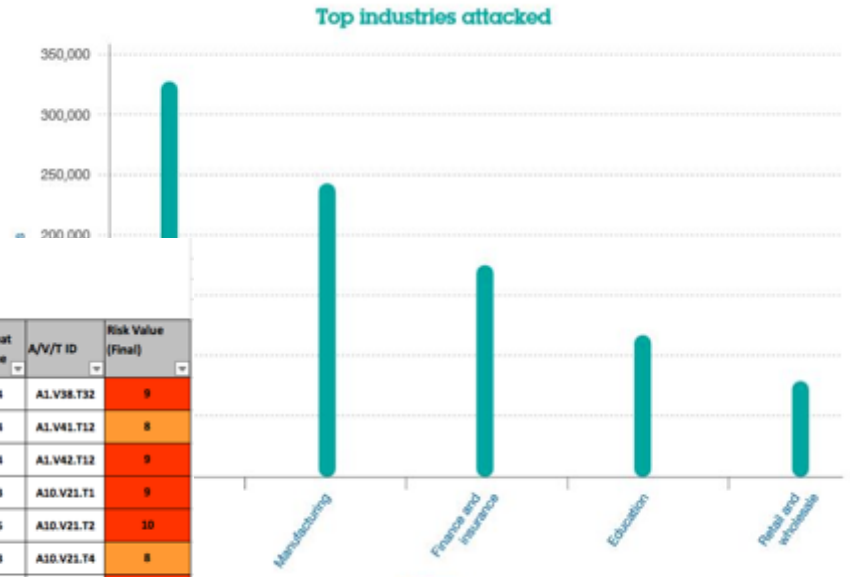
Threat Intel Final Scoring

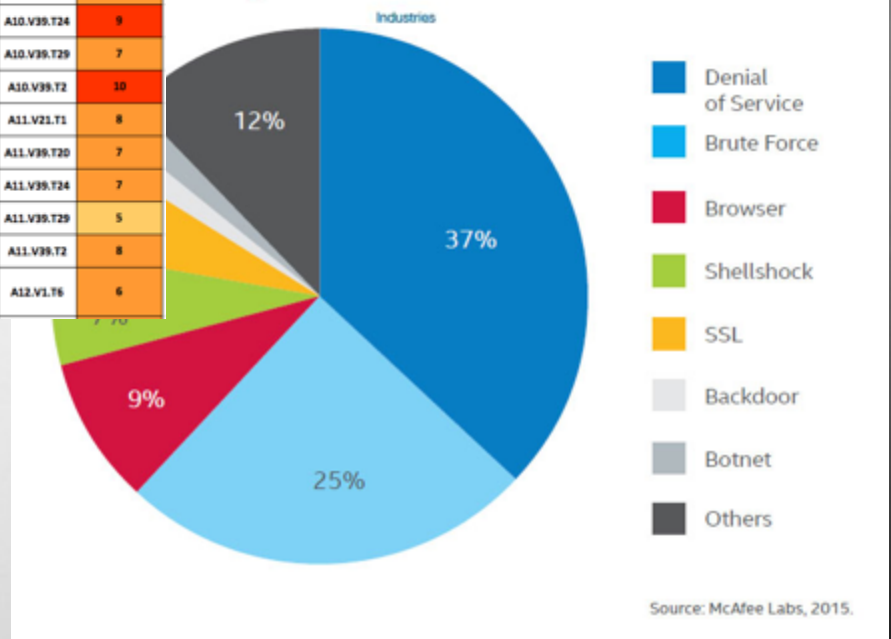


Identification & Assessment of Risks (Asset-Threat-Vulnerability)

Asset ID	Asset Description	Asset Value	Vulnerability Description	A/V Value	Threats	Threat Value	A/V/T ID	Risk Value (Final)
A1	Automated reservation, check-in and boarding procedure	4	V38. Lack of common or harmonised legislation in EU Member States	4	T32. Profiling	4	A1.V38.T32	9
A1	Automated reservation, check-in and boarding procedure	4	V41. Lack of respect to the data conservation principle	3	T12. Non-compliance with data protection legislation	4	A1.V41.T12	8
A1	Automated reservation, check-in and boarding procedure	4	V42. Lack of respect to the rights of the data subject (such as the right for rectification, blocking or deletion of data).	4	T12. Non-compliance with data protection legislation	4	A1.V42.T12	9
A10	Credit Cards/Debit card/Payment cards/'e-wallet'	4	V21. Inappropriate / inadequate identity management	4	T1. Denial of service attack / Flood / Buffer overflow	3	A10.V21.T1	9
A10	Credit Cards/Debit card/Payment cards/'e-wallet'	4	V21. Inappropriate / inadequate identity management	4	T2. Spoofing of credentials / bypass authentication	5	A10.V21.T2	10
A10	Credit Cards/Debit card/Payment cards/'e-wallet'	4	V21. Inappropriate / inadequate identity management	4	T4. Traffic analysis / scan / probe	3	A10.V21.T4	8
A10	Credit Cards/Debit card/Payment cards/'e-wallet'	4	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	4	T24. Worms, viruses & malicious code	3	A10.V39.T24	9
A10	Credit Cards/Debit card/Payment cards/'e-wallet'	4	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	4	T29. MANET/Adhoc network routing attack	2	A10.V39.T29	7
A10	Credit Cards/Debit card/Payment cards/'e-wallet'	4	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	4	T2. Spoofing of credentials / bypass authentication	5	A10.V39.T2	10
A11	Other RFID cards	3	V21. Inappropriate / inadequate identity management	4	T1. Denial of service attack / Flood / Buffer overflow	3	A11.V21.T1	8
A11	Other RFID cards	3	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	3	T20. Fake / rogue RFID readers / scanning of RFID reader and /or tag	3	A11.V39.T20	7
A11	Other RFID cards	3	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	3	T24. Worms, viruses & malicious code	3	A11.V39.T24	7
A11	Other RFID cards	3	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	3	T29. MANET/Adhoc network routing attack	2	A11.V39.T29	5
A11	Other RFID cards	3	V39. Insufficient protection of wireless networks and communication (weak or no encryption etc.)	3	T2. Spoofing of credentials / bypass authentication	5	A11.V39.T2	8
A12	Scanners & detectors	3	V1. Inappropriate design of procedures - includes: lack of accountability, high complexity of procedures, assigning extensive responsibilities to end-users (in critical parts of the procedures) etc.	2	T6. Social engineering attack	4	A12.V1.T6	6







Score Evaluation

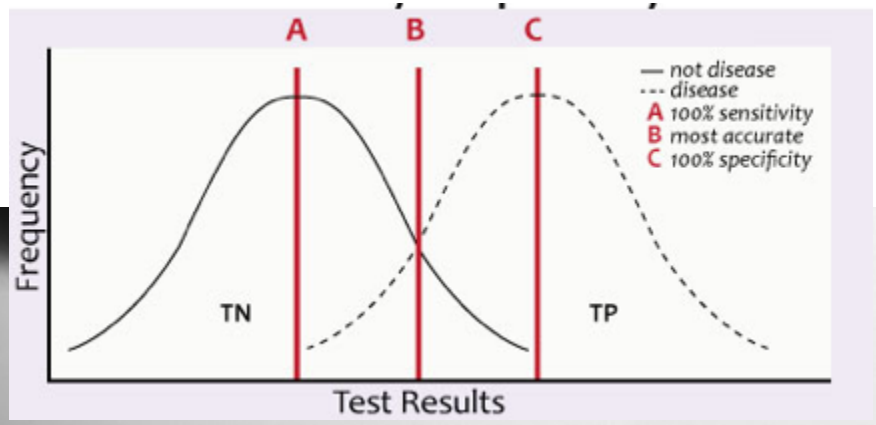
TO ESTIMATE CONFIDENCE INTERVALS FOR SENSITIVITY, SPECIFICITY AND TWO-LEVEL LIKELIHOOD RATIOS:

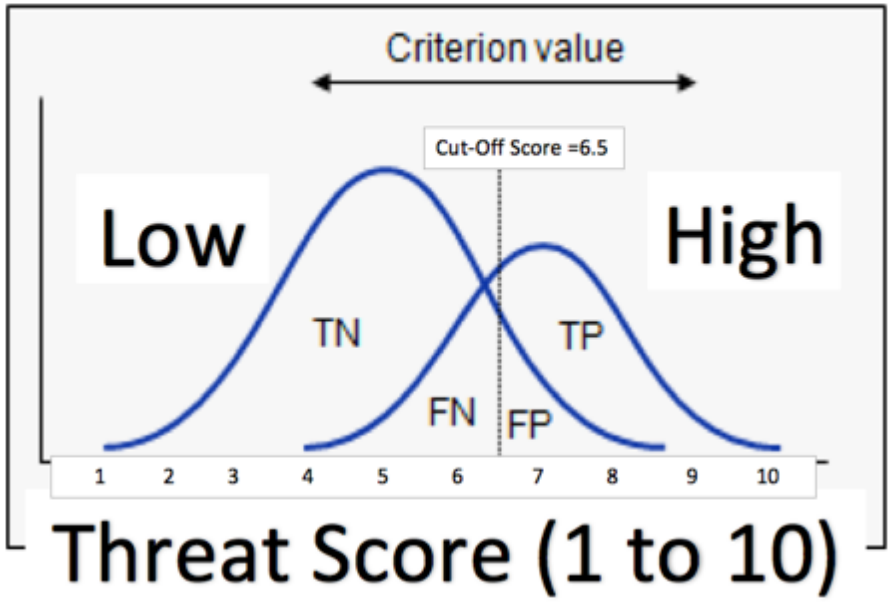
	Found Relevant	Found Not Relevant
Above Cut-off Score	200	200
Below Cut-off Score	20	450

Enter the required confidence interval: 90

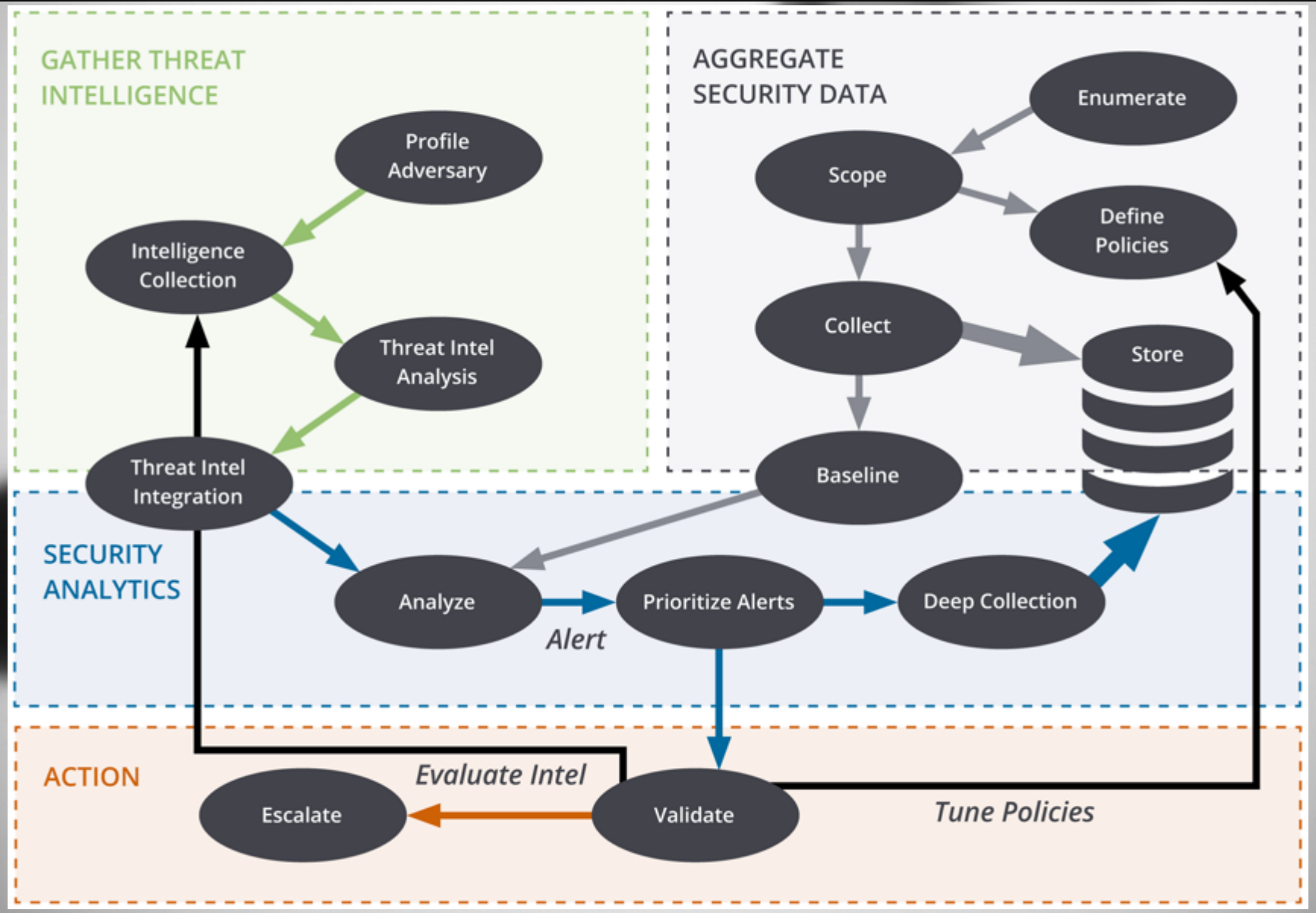
RESULT:

Sensitivity:	0.9091	CI:	0.872	to	0.9362
Specificity:	0.6923	CI:	0.6618	to	0.7212
Positive likelihood ratio:	2.955	CI:	2.666	to	3.275
Negative likelihood ratio:	0.131	CI:	0.092	to	0.187
Diagnostic odds ratio:	22.500	CI:	14.928	to	33.914





Putting it all together



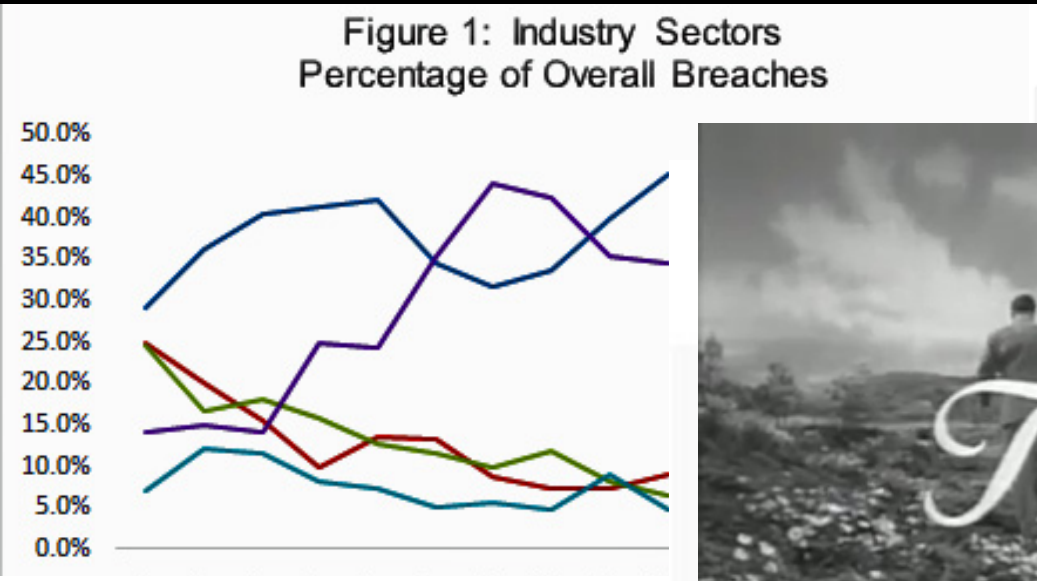
Where should we start?

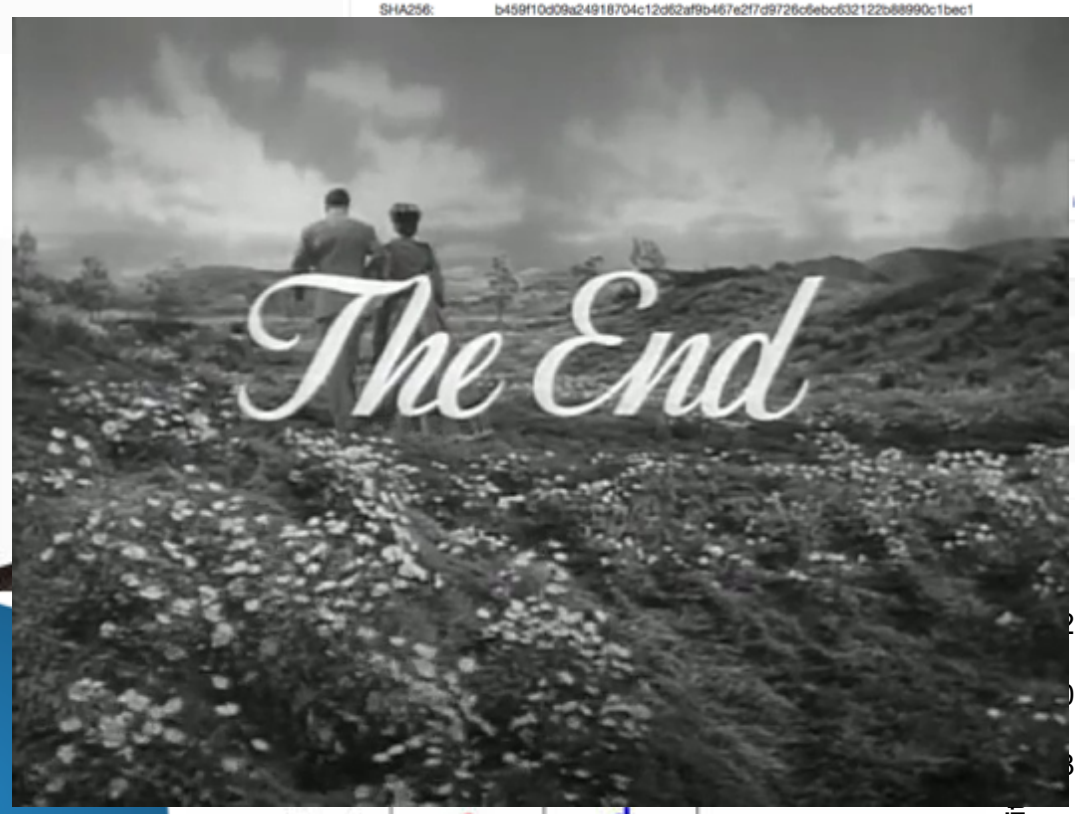
- Aggregate, Normalize and Deduplicate your threat intelligence data (TIP)
- Consult with a data scientist on your current dataset
- Use your intuition and experience when building a model (Scoring)
- Use false positive as feedback to your model (False positives)
- Keep your statistical models aligned with your organizational goals

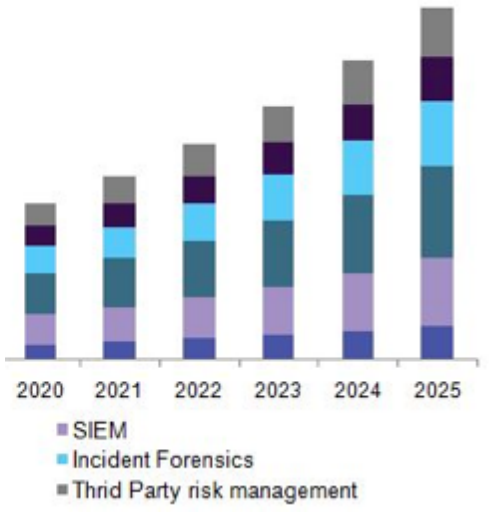
Summary

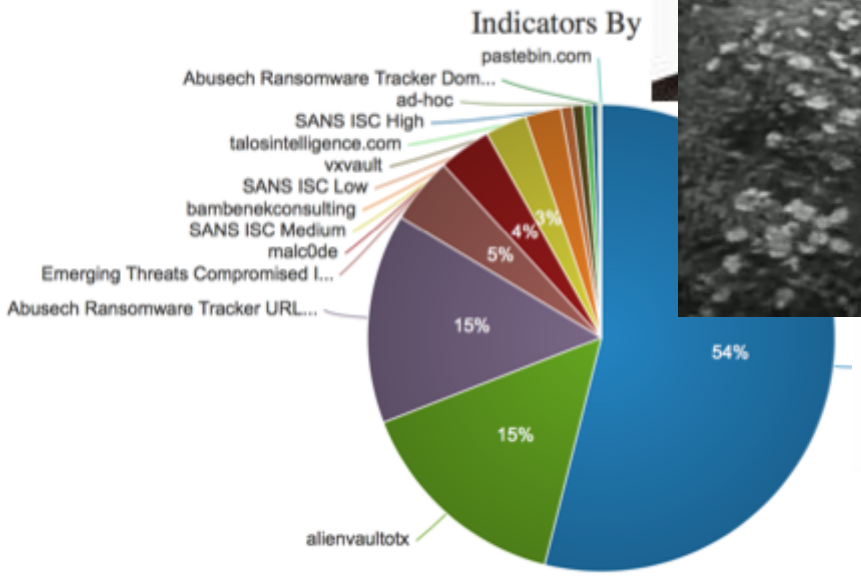
- The higher on the Pyramid of Pain, the longer indicators stay relevant.
- Indicators can help not only detect, but also prevent future attacks.
- Current threat intelligence feeds has lots of noise.
- Statistics can help reduce the noise by:
  - Move from subjective to objective decisions
  - Correlate internal and external attributes
  - Reduce the number of false positives
  - Identify threat related trends

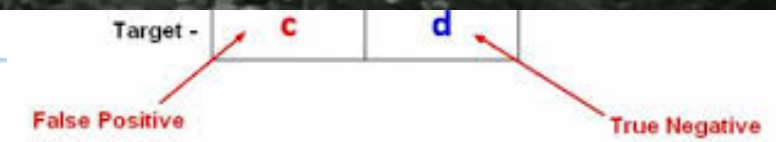
Practical Statistics for Threat Intelligence

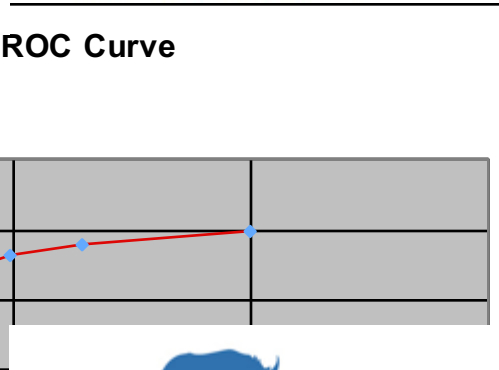












Nir Yosha – Threat Intel Engineer