AI & ML IN CYBERSECURITY

Why Algorithms Are Dangerous

RAFFAEL MARTY
Vice President of Corporate Strategy, Forcepoint

BlackHat USA | August 2018

FORCEPOINT

A BRIEF SUMMARY

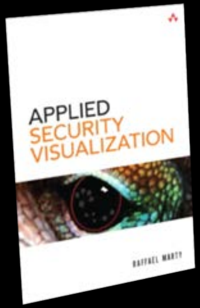
- We don’t have artificial intelligence (yet)
- Algorithms are getting ‘smarter’, but experts are more important
- Stop throwing algorithms on the wall - they are not spaghetti
- Understand your data and your algorithms
- Invest in people who know security (and have experience)
- Build systems that capture “export knowledge”
- Think out of the box, history is bad for innovation
- Focus on advancing insights

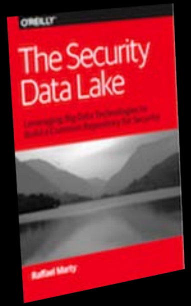
RAFFAEL MARTY

FORCEPOINT

- Sophos
- PixlCloud
- Loggly
- Splunk
- ArcSight
- IBM Research

- Security Visualization
- Big Data
- ML & AI
- SIEM
- Corp Strategy
- Leadership
- Zen





OUTLINE

01 STATISTICS, MACHINE LEARNING & AI
Defining the concepts

02 THE ALGORITHMIC PROBLEM
Understanding the data and the algorithms

03 AN EXAMPLE
Let’s get practical

01

STATISTICS MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

*“Everyone calls their stuff ‘machine learning’ or even better ‘artificial intelligence’ - It’s not cool to use **statistics**!”*

*“Companies are throwing **algorithms** on the wall to see what sticks (see security analytics market)”*

ML AND AI – WHAT IS IT?

MACHINE LEARNING

Algorithmic ways to "describe" data

- **Supervised**
  - We are giving the system a lot of training data and it learns from that
- **Unsupervised**
  - We give the system some kind of optimization to solve (clustering, dim reduction)

DEEP LEARNING

A "newer" machine learning algorithm

- Eliminates the feature engineering step
- Explainability / verifiability issues

DATA MINING

Methods to explore data – automatically

ARTIFICIAL INTELLIGENCE

"Just calling something AI doesn't make it AI."

"A program that doesn't simply classify or compute model parameters, but comes up with novel knowledge that a security analyst finds insightful."

We don't have artificial intelligence (yet)

WHAT "AI" DOES TODAY

KICK A HUMAN'S ASS AT GO

DESIGN MORE EFFECTIVE DRUGS

MAKE SIRI SMARTER

MACHINE LEARNING USES IN SECURITY

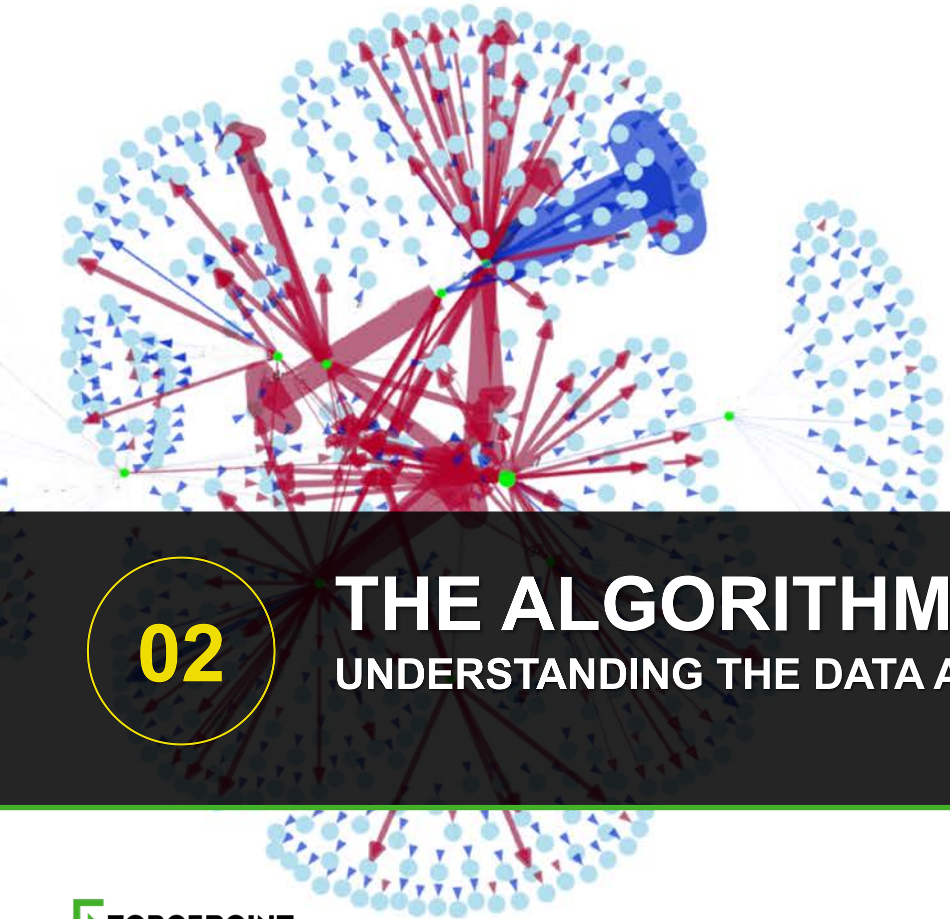
SUPERVISED

- Malware classification
  - Deep learning on millions of samples - 400k new malware samples a day
  - Has increased true positives and decreased false positives compared to traditional ML
- Spam identification
- MLSec project on firewall data
  - Analyzing massive amounts of firewall data to predict and score malicious sources (IPs)

UNSUPERVISED

- DNS analytics
  - Domain name classification, lookup frequencies, etc.
- Threat Intelligence feed curation
  - IOC prioritization, deduplication, …
- Tier 1 analyst automation
  - Reducing workload from 600M raw events to 100 incidents*
- User and Entity Behavior Analytics (UEBA)
  - Uses mostly regular statistics and rule-based approaches

* See Respond Software Inc.



02

THE ALGORITHMIC PROBLEM

UNDERSTANDING THE DATA AND THE ALGORITHMS


WARNING
ALGORITHMS ARE DANGEROUS

FAMOUS AI (ALGORITHM) FAILURES

December 2009 | Hewlett-Packard investigates instances of so-called "racist camera software" which had trouble recognizing dark-skinned people

March 2015 | A Carnegie Mellon University study determines that some personalized ads from sites such as Google and Facebook are gender-biased

July 2015 | A Google algorithm mistakenly captions photos of black people as "Gorillas"

March 2016 | Microsoft shuts down AI chatbot Tay after it starts using racist language

May 2016 | ProPublica investigation finds that a computer program used to track future criminals in the US is racially biased

September 2016 | Machine-learning algorithms used to judge an international beauty contest displays bias against dark-skinned contestants

PENTAGON - AI FAIL

http://neil.fraser.name/writing/tank/

WHAT MAKES ALGORITHMS DANGEROUS?

ALGORITHMS MAKE **ASSUMPTIONS** ABOUT THE **DATA**

- Assume 'clean' data (src/dst confusion, user feedback, etc.)
- Often assume a certain type of data and its distribution
- Generally don't deal with outliers
- Machine learning assumes enough, representative data
- Need contextual features (e.g., not just IP addresses)
- Assume all input features are 'normalized' the same way

ALGORITHMS ARE **TOO EASY** TO USE THESE DAYS (TENSORFLOW, TORCH, ML ON AWS, ETC.)

- The process is more important than the algorithm (e.g., feature engineering, supervision, drop outs, parameter choices, etc.)

ALGORITHMS DO NOT TAKE **DOMAIN KNOWLEDGE** INTO ACCOUNT

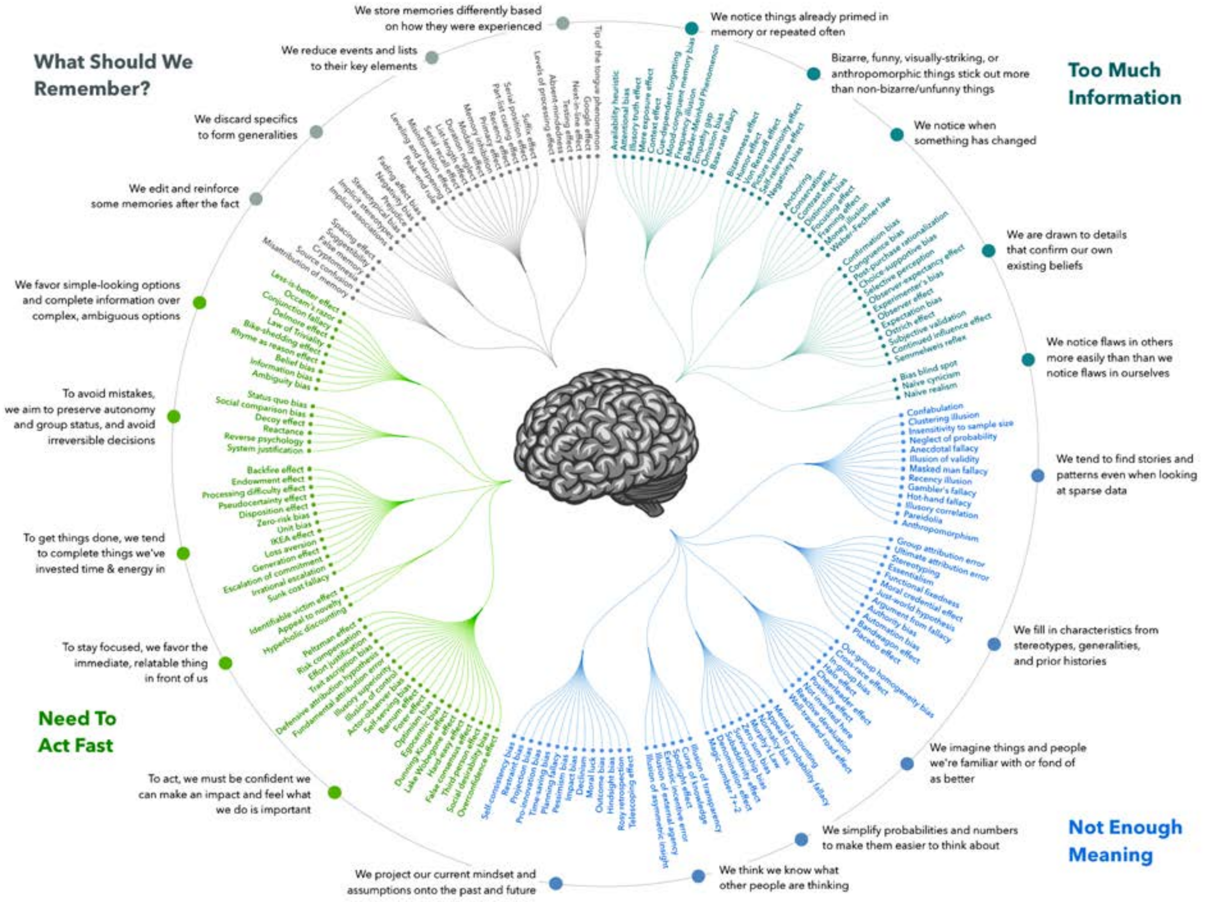
- Defining meaningful and representative distance functions, for example
- e.g., each L4 protocol exhibits different behavior. Train it separately.
- e.g., interpretation is often unvalidated - beware of overfitting and biased models.
- Ports look like integers, they are not, same is true for IPs, processIDs, HTTP return codes, etc.

WHAT MAKES ALGORITHMS DANGEROUS?

SAMPLE BIAS

KNOWLEDGE BIAS

HISTORY BIAS



CONFIRMATION BIAS

AVAILABILITY BIAS

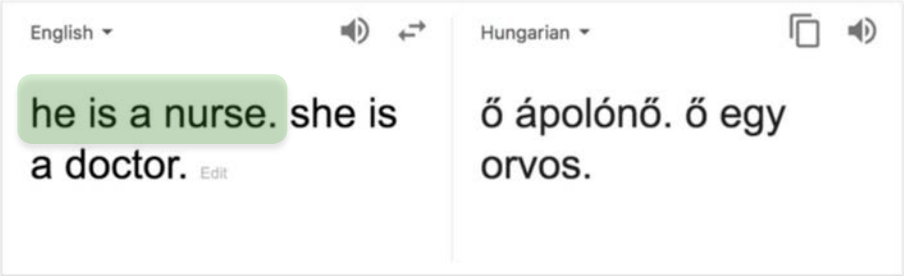
CORRELATION BIAS

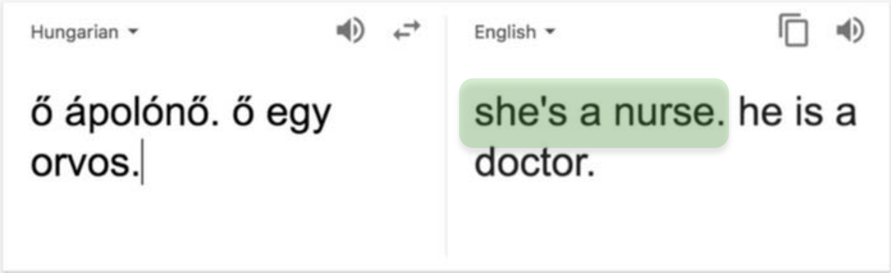
https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18

COGNITIVE BIASES

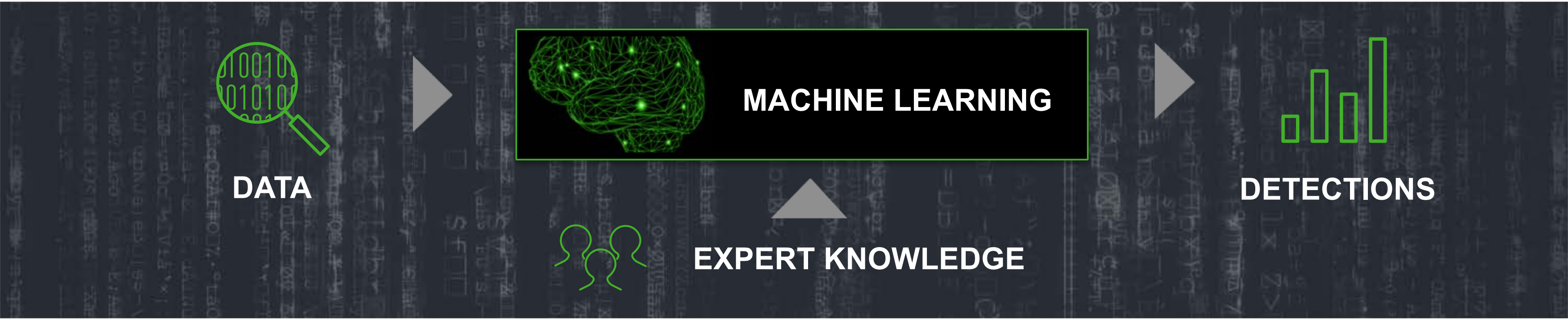
How biased is your data set? How do you know?

- Only a single customer’s data
- Learning from an ‘infected’ data set
- Collection errors
- Missing data (e.g., due to misconfiguration)
- What’s the context the data operates in?
- FTP although generally considered old and insecure, isn’t always problematic
- Don’t trust your IDS (e.g. “UDP bomb”)





THE DANGERS WITH DEEP LEARNING – WHEN NOT TO USE IT



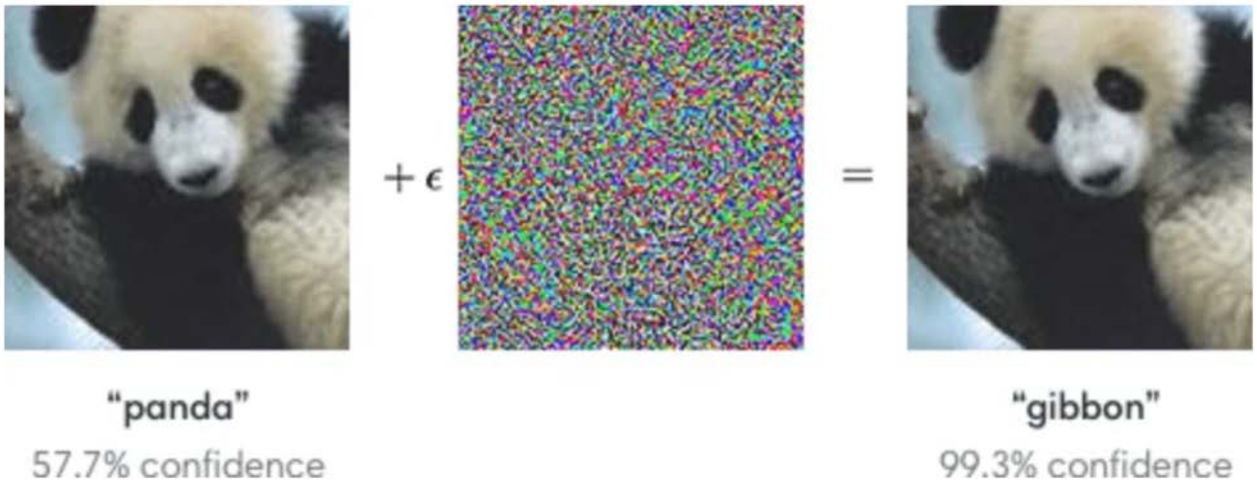
- Not enough or no quality **labelled data**
- Data **cleanliness** issues (timestamps, normalization across fields, etc.)
- Bad understanding of the data to engineer meaningful **features** (e.g., byte stream for binaries)
- Data is prone to **adversarial** input

- No well trained **domain experts** and data scientists to oversee the implementation
- A need to understand what ML actually learned (**explainability**)

- **Verifiability** of output
- **Interpretation** of output

ADVERSARIAL MACHINE LEARNING

An example of an attack on deep learning



Above: Image Credit: Ian Goodfellow

03

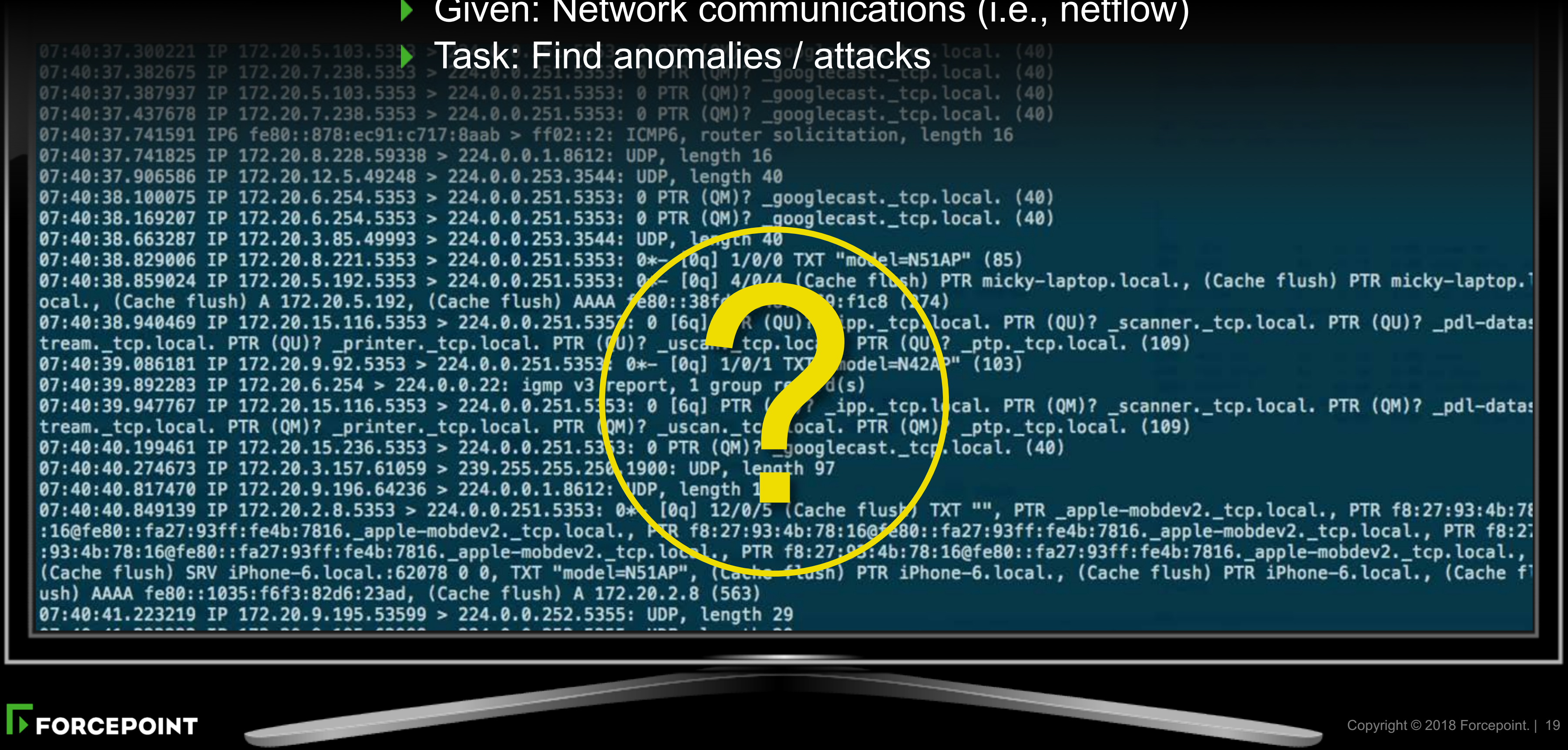
EXAMPLE

LET’S GET PRACTICAL

FORCEPOINT

FINDING ANOMALIES / ATTACKS IN NETWORK TRAFFIC

- Given: Network communications (i.e., netflow)
- Task: Find anomalies / attacks



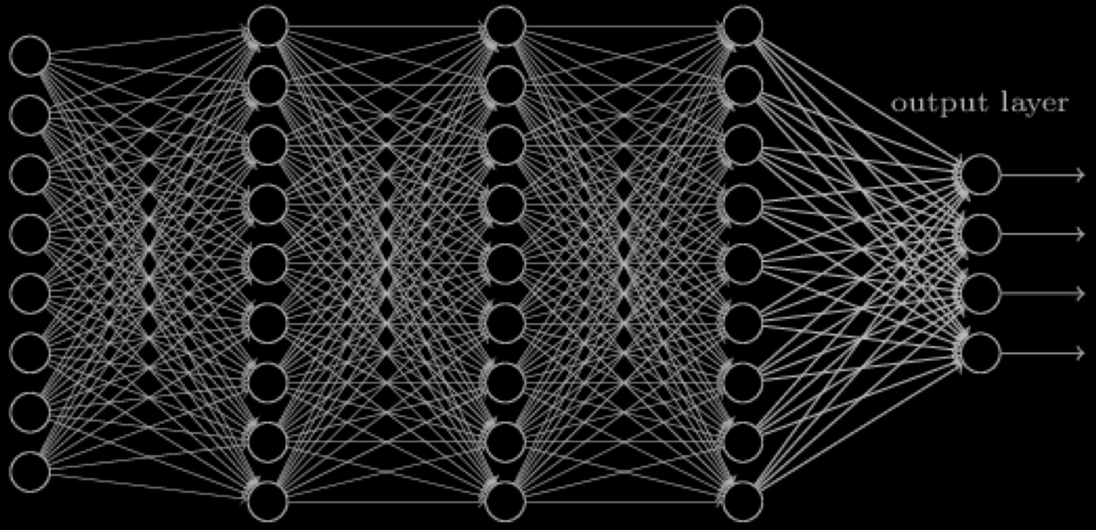
DEEP LEARNING – THE SOLUTION TO EVERYTHING

DEEP LEARNING PROMISES A FEW THINGS:

- ‘Auto’ feature extraction
- High accuracy of detections

AND WE SATISFY SOME REQUIREMENTS:

- Lots of data available
- BUT: A single record does not indicate good/bad
- BUT: Not enough ‘information’ within flows – need **context**
- BUT: No **labels** available



MOST SECURITY PROBLEMS CAN’T BE SOLVED WITH DEEP LEARNING
or supervised methods in general

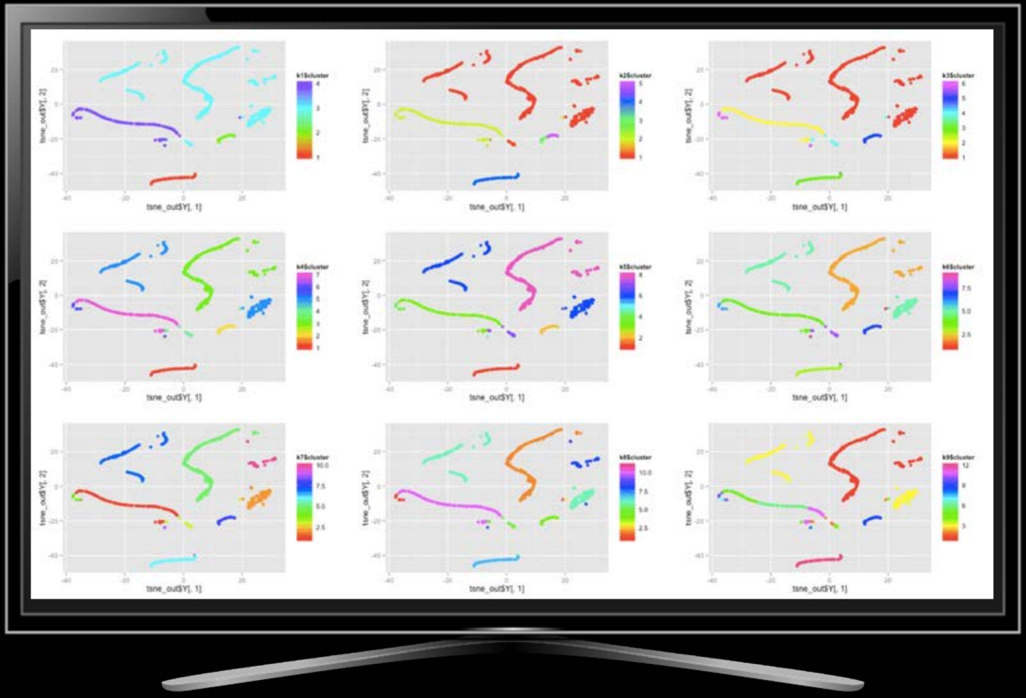
FORCEPOINT

UNSUPERVISED TO THE RESCUE?

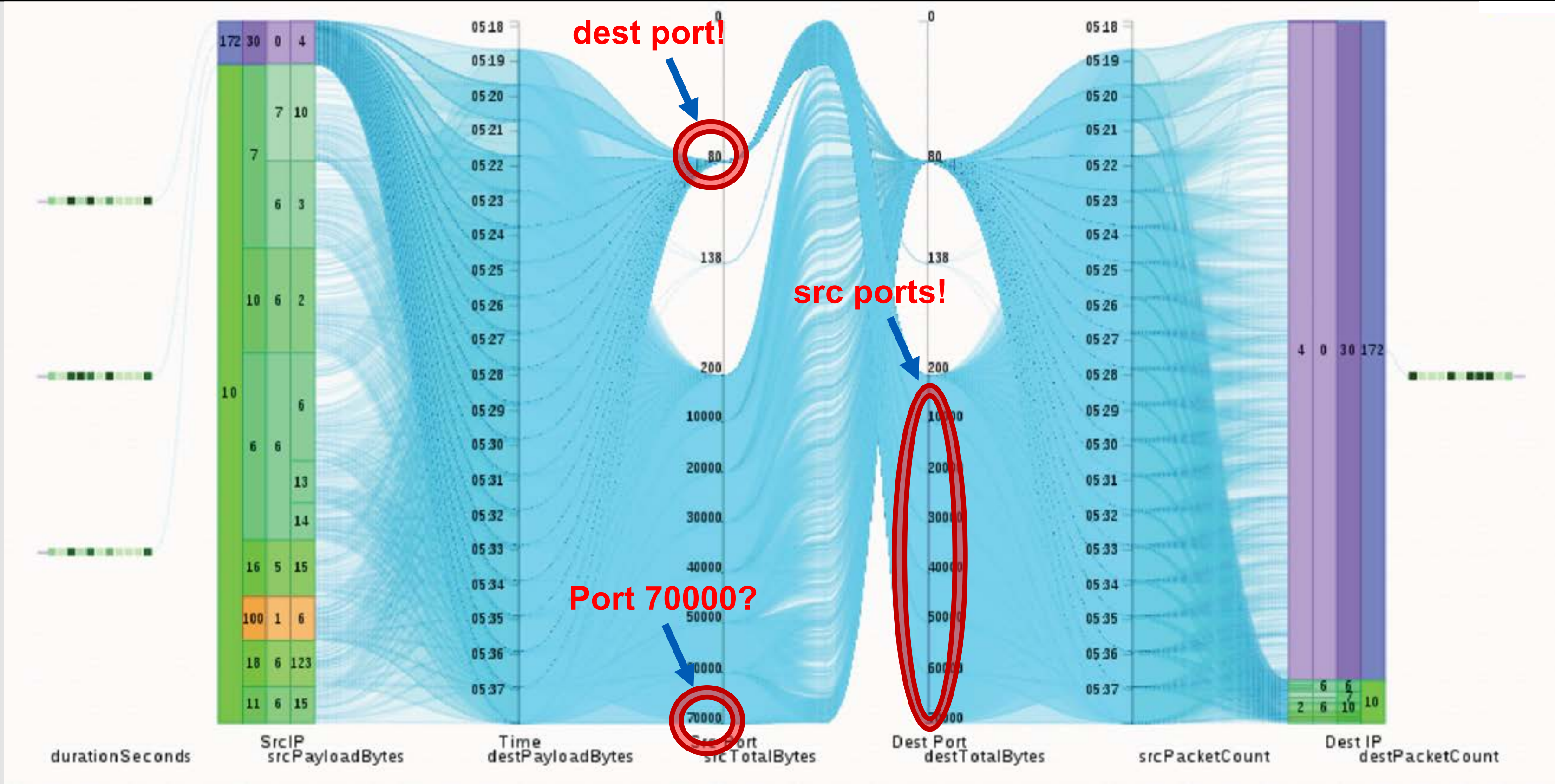
Can we exploit the inherent structure within the data to find anomalies and attacks?

Clustering traffic to find outliers

1. Clean the data
2. Engineer distance functions
3. Figure out the right algorithm
4. Apply the correct algorithmic parameters
5. Data interpretation



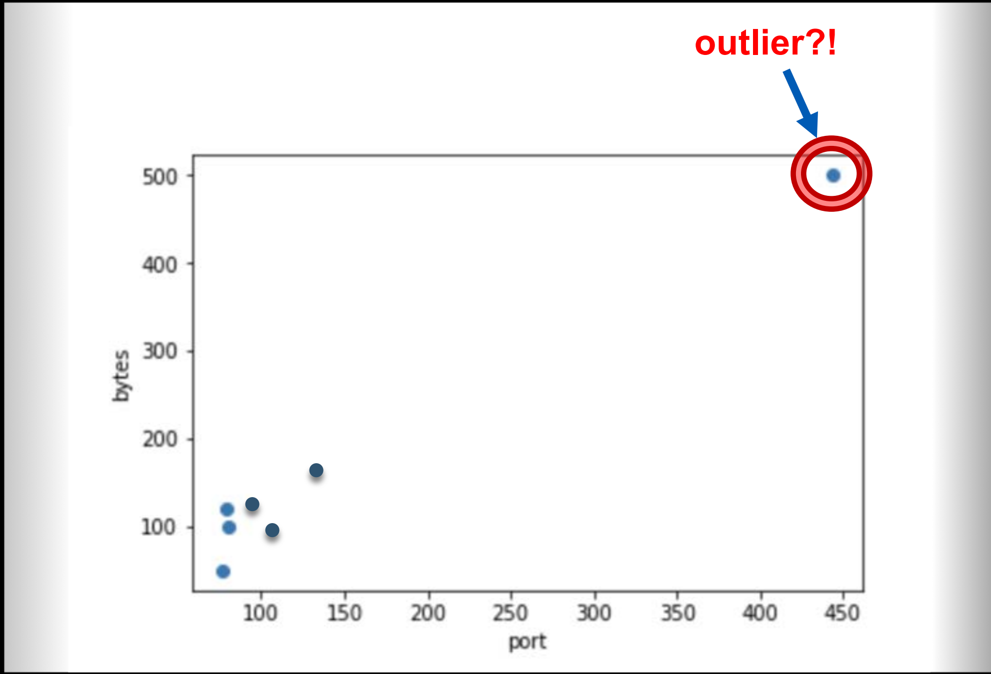
1. UNDERSTAND AND CLEAN THE DATA



http://vis.pku.edu.cn/people/simingchen/docs/vastchallenge13-mc3.pdf

2. ENGINEERING DISTANCE FUNCTIONS

- Distance functions define the similarity of data objects
- Need domain-specific similarity functions
  - URLs (simple levenshtein distance versus domain based?)
  - Ports (and IPs, ASNs) are NOT integers
  - Treat user names as categorical, not as strings



FORCEPOINT

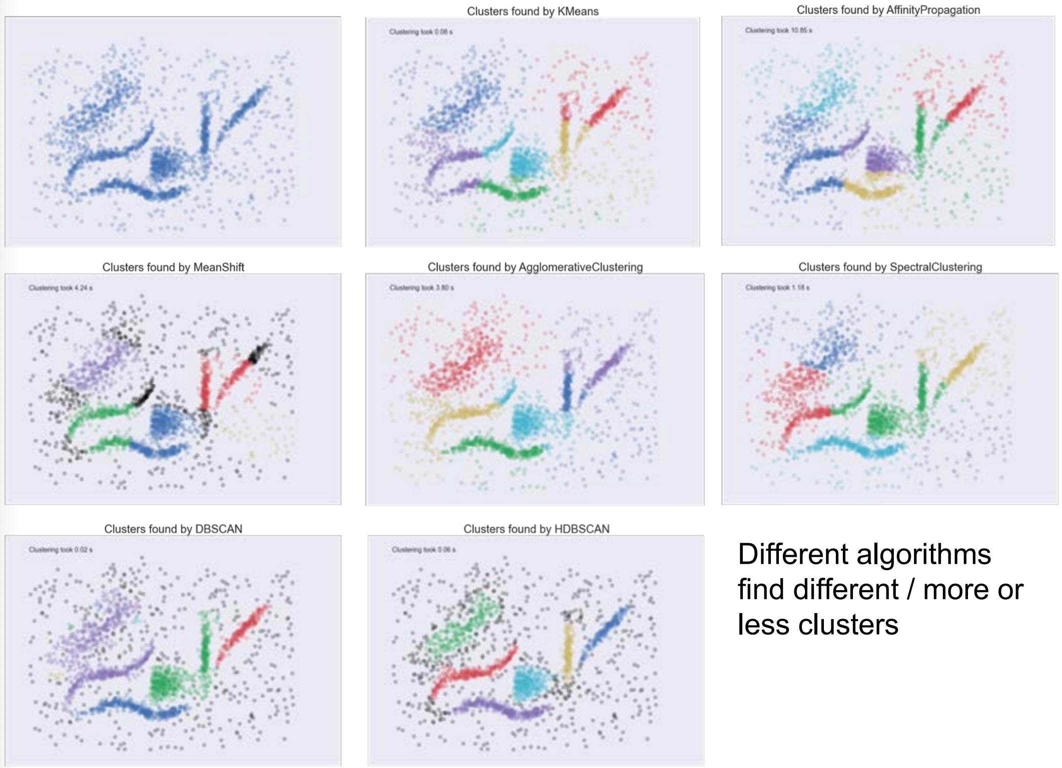
3. CHOOSING THE RIGHT UNSUPERVISED ALGORITHM

CLUSTERING ALGORITHMS

- K-means
- Affinity Propagation (AP)
- DBScan
- t-SNE

CRITERIA TO CHOOSE AN ALGORITHM

- Dimensionality of data
- “Shape” of data
- Intrinsic algorithm workings
- Algorithm convergence or speed

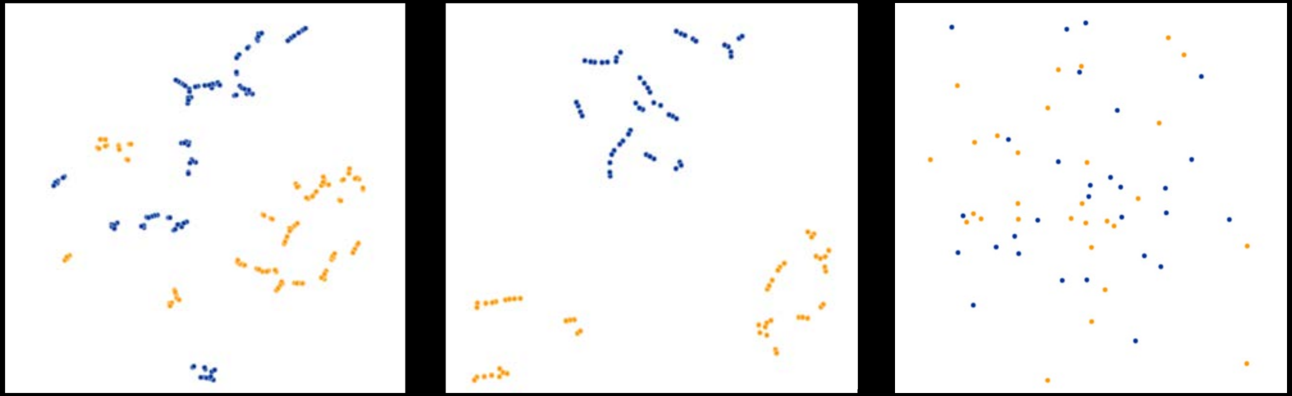


Different algorithms find different / more or less clusters

http://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

4. CHOOSING THE CORRECT ALGORITHM PARAMETERS

- The dangers of not understanding algorithmic parameters
- t-SNE clustering of network traffic from two types of machines



perplexity = 3
epsilon = 3

No clear separation

perplexity = 3
epsilon = 19

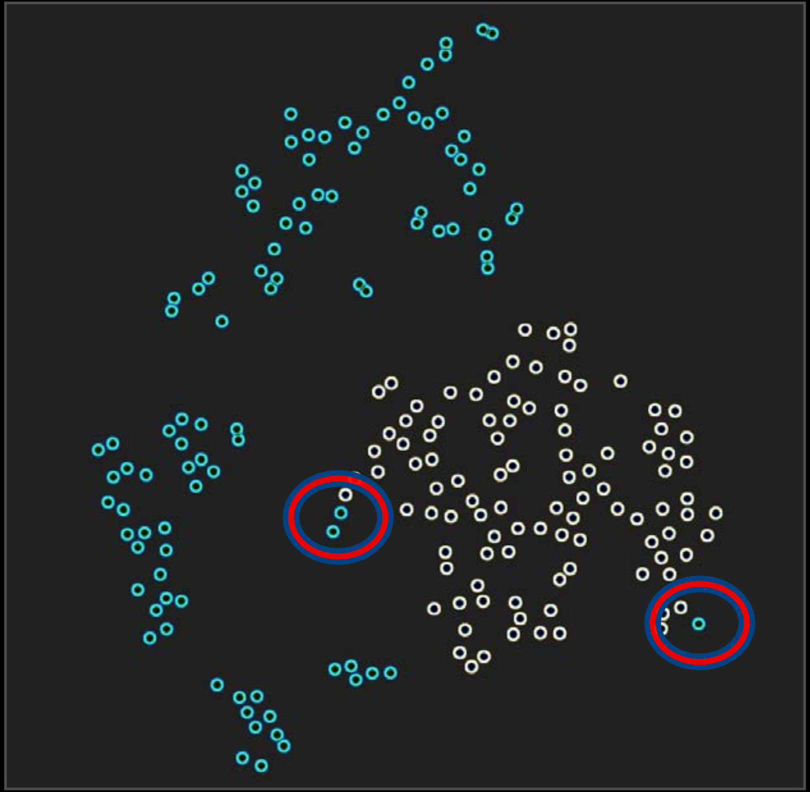
3 clusters instead of 2

perplexity = 93
epsilon = 19

What a mess

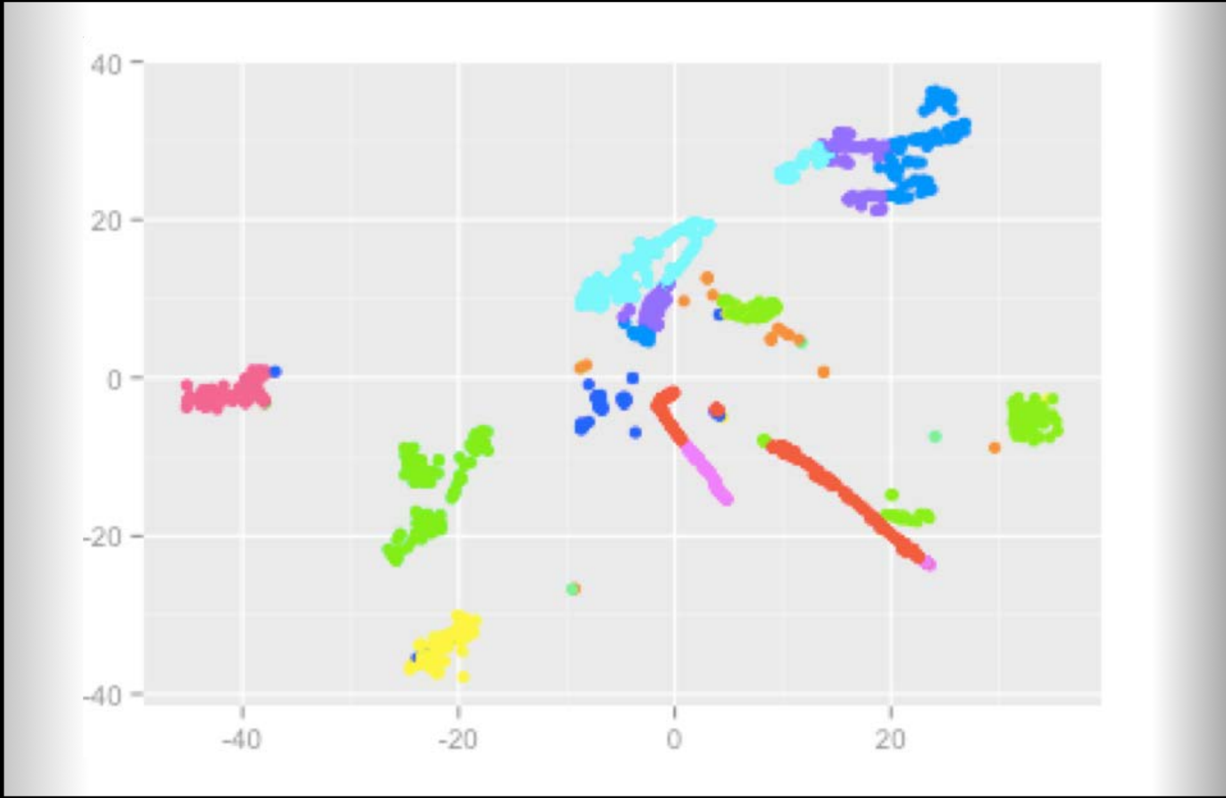
4. CHOOSING THE CORRECT ALGORITHM PARAMETERS

- And this is when it gets dangerous
- Access decisions / enforcements based on cluster membership



5. INTERPRETING THE DATA

We analyze network traffic. The graph shows an abstract space (X and Y axes have no specific meaning). Each dot represents a device on the network. Colors represent machine-identified clusters.



Interpretation questions:

- What are these clusters?
- What are good clusters?
- What's anomalous?
- Where are compromised machines / attackers?

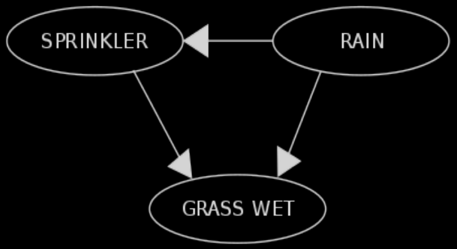
A DIFFERENT APPROACH - PROBABILISTIC INFERENCE

Rather than running algorithms that model the shape of data, we need to take **expert knowledge / domain expertise** into account

Introducing Belief Networks

- Models that represent the **state** of the 'world'
- Helps us make predictions and **reason** about the world
- A graph rather than huge joint distribution tables across all states
- Using **Bayes** theorem to calculate 'belief'
- Could use ML to learn graph structure (nodes and edges), but it'll get too unwieldy and non-interpretable!

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

"What is the probability that it is raining, given the grass is wet?": 35.77%

BAYESIAN BELIEF NETWORK 1ST STEP – BUILD THE GRAPH

1. **What's our objective?**
2. **What behaviors can we observe?**
   - What are observable factors that reduce uncertainty of the central inference (of device compromised)
   - Observations should not be locally dependent – they should be true across all customers / environments
   - Do we have that data?
   - Do we need context for it?

Device is Compromised

Open port 53

Is using port 23?

Protocol mismatch

New protocol seen

Not seen for a week

Shows up with new OS

Has known vulnerabilities

Mistake in IP classification

Connecting to suspicious IP

Machine got update to new OS

Connecting from suspicious IP

Device is in maintenance mode

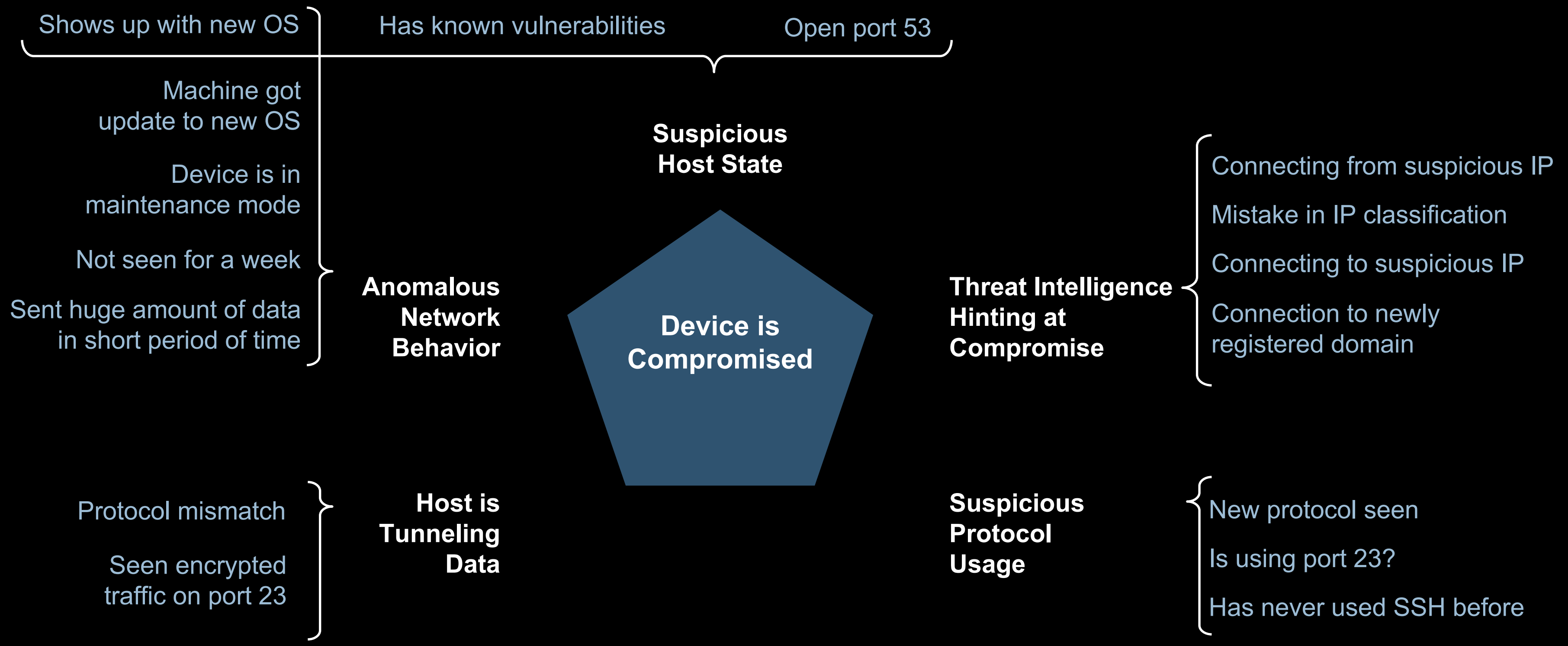
Seen encrypted traffic on port 23

Connection to newly registered domain

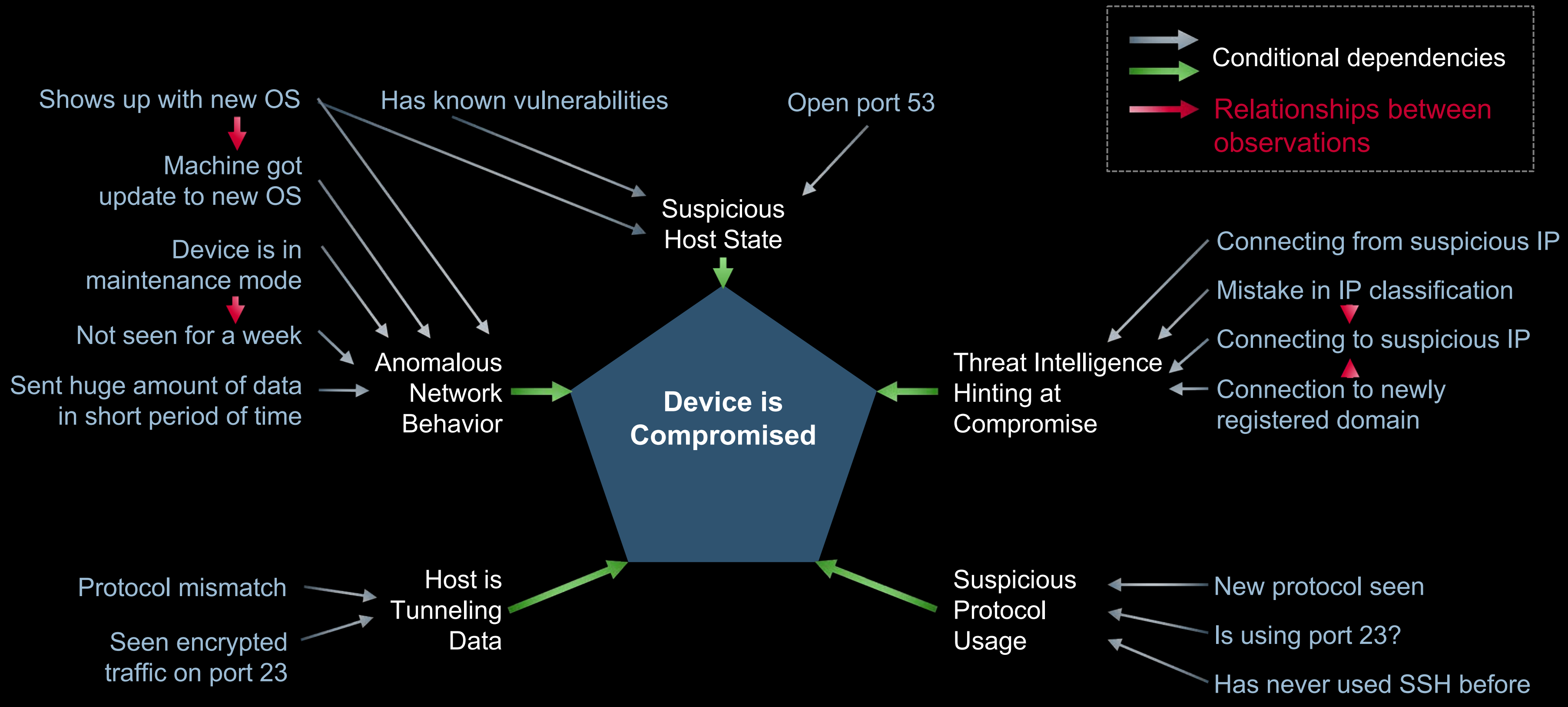
Sent huge amount of data in short period of time

BAYESIAN BELIEF NETWORK 2ND STEP – GROUP NODES

Complexity of this network is too high. We cannot computer all the conditional probabilities.
Therefore we need to introduce “grouping nodes”.



BAYESIAN BELIEF NETWORK 3RD STEP – INTRODUCE DEPENDENCIES



BAYESIAN BELIEF NETWORK 4TH STEP – ESTIMATE PROBABILITIES

NODE PROBABILITIES

- P(Protocol mismatch) = 0.01 OR “very low”
- P(Seen encrypted traffic on port 23) = 0.01 OR “very low”
- P(Host is Tunnelling Data) = 0.01 OR ”very low”

Expert Knowledge

CONDITIONAL PROBABILITIES

- Our belief network teaches us: “Tunnelling is not independent of seeing port 23 traffic”
- P(Tunnelling | Enc. Port 23 Traffic) = (P(Enc. Port 23 | Tunnelling) * P(Tunnelling)) / P(Enc. Port 23)

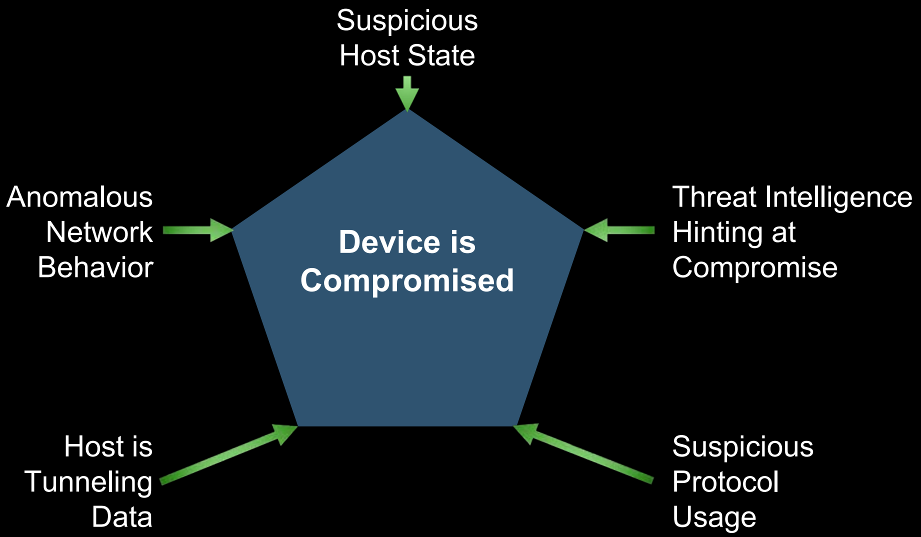
JOINT PROBABILITIES

- Multiple factors lead to Tunnelling, not just one
- P(Tunnelling | Enc. Port 23 **AND** Proto mismatch) = (P(Enc. Port 23 AND Proto mismatch | Tunnelling) * P(Tunnelling)) / P(Enc Port 23 AND Proto mismatch)

More precise than in pervious formula

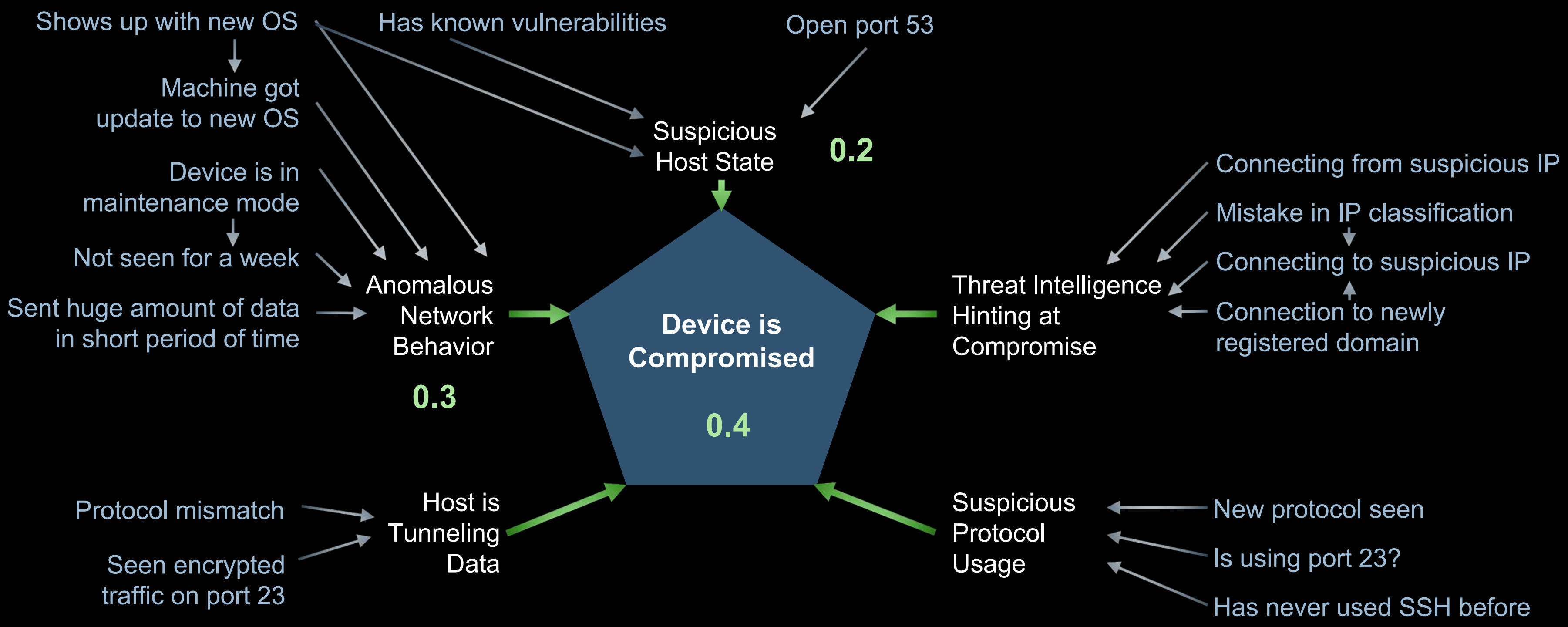
BAYESIAN BELIEF NETWORK 5TH STEP – GOAL COMPUTATION

The probability that we have a compromised device is the joint and conditional probability over all the 'group nodes'



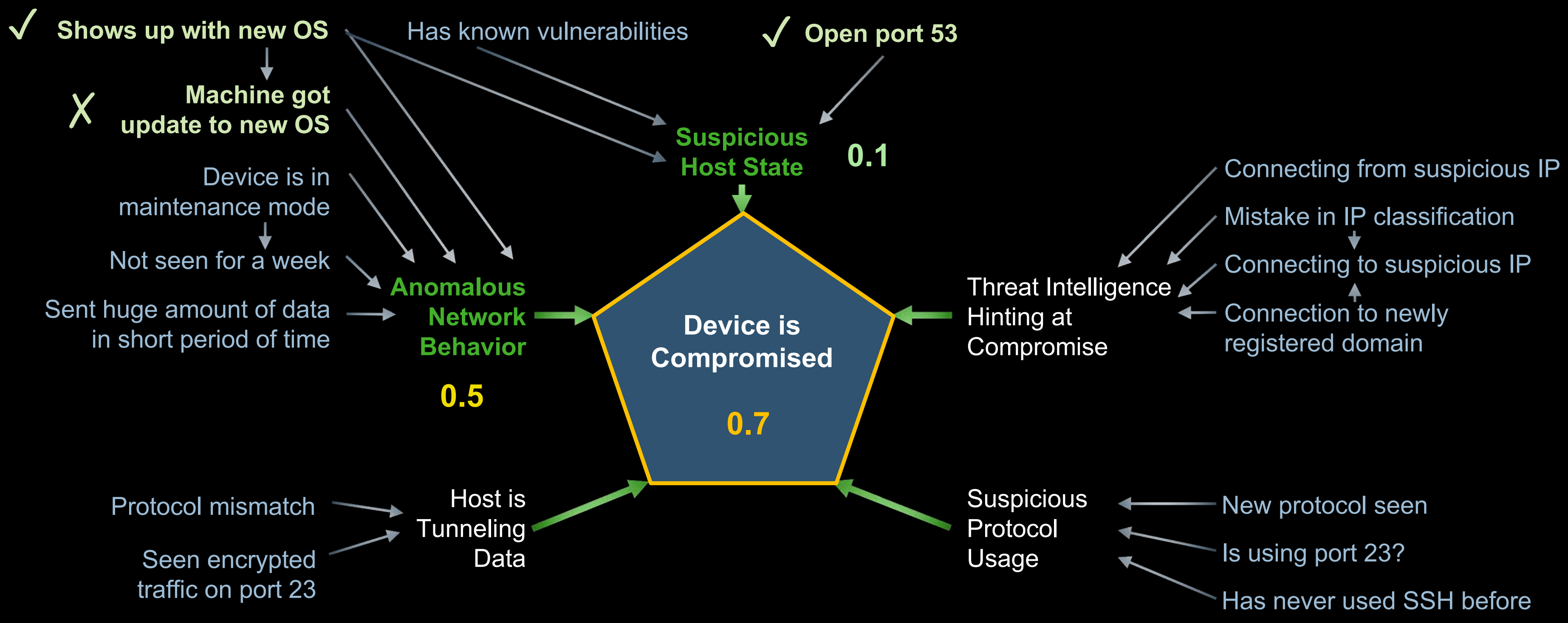
FORCEPOINT

BAYESIAN BELIEF NETWORK 6TH STEP – OBSERVE ACTIVITIES



BAYESIAN BELIEF NETWORK 6TH STEP – OBSERVE ACTIVITIES

1. Update the '**observation nodes**' in the network with observation (what we find in the logs)
2. **Re-compute** probabilities on the connected nodes



BAYESIAN BELIEF NETWORK 7TH STEP – EXPERT INPUT

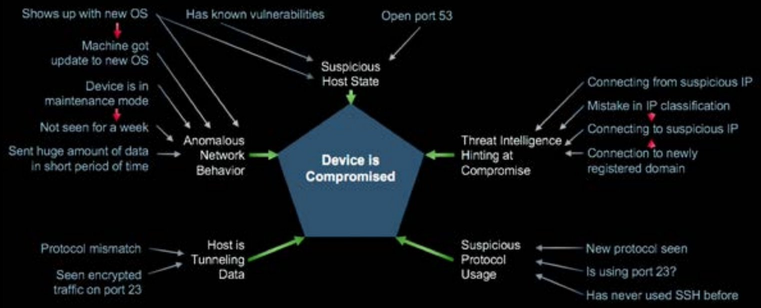
Strengthen the network by introducing expert knowledge

- Pose any combinations of 'observations' and 'group' nodes as questions to experts
- *Asking meaningful questions is an art and requires expert knowledge*
- You will find that it matters how you named your nodes to define good questions

Question	Expert Answer
What's the probability that device is compromised and I have highly suspicious network behavior and nothing on threat intelligence	0.3
Probability that host is in suspicious state, given that port 53 is open, brand new OS	0.1
How likely is it that we see a connection to a newly registered domain and we see port 23 traffic?	0.01
Etc.	

Note how this is not a full joint probability over only a subset of the group nodes.

We can have questions across observational nodes of different groups as well



FORCEPOINT

BELIEF NETWORKS – SOME OBSERVATIONS

> Biggest benefit of belief networks is that the learned knowledge can be **verified** and **extracted**!

- Iterative process of adding more nodes, grouping, adding expert input, etc.
- Graph allows for answering many questions – e.g., sensitivity analysis
- Do not determine the probabilities on the observation nodes with historic data. It is only accurate for scenarios that were included in data – how do you know your data covered all scenarios?
- Each problem requires the definition of a graphs based on expert input
- A generic "Network Traffic" graph is hard to build and train
  - Not every FTP is bad
  - Poor network practice -> e.g., using unencrypted protocols like FTP

Thanks Chris @ respond-software.com for all your help!

IN SUMMARY

FORCEPOINT

RECOMMENDATIONS

- Start with defining your use-cases, not choosing an algorithm
- ML is barely ever the solution to your problem
- Use ensembles of algorithms
- Teach the algos to ask for input – if it's unsure, have it ask an expert rather than making a decision on its own
- Make sure models keep up with change and forget old facts that are not relevant anymore
- Do you need white lists / black lists for your algos to not go haywire?
- Verify your models - use visualization to help with that
- Share your insights with your peers – security is not your competitive advantage
- GDPR – transparency on what data is collected and used for decisions

"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

BLACK HAT SOUNDBITES

black hat USA 2018

"Algorithms are getting 'smarter', but experts are more important"

"Understand your data, your algorithms, and your data science process"

"History is not a predictor – but knowledge is"

QUESTIONS?

http://slideshare.net/zrlram
@raffaelmarty

