Delivering Security Insights with Data Analytics and Visualization

ACSAC Orlando

November 2017

Raffael Marty
VP Security Analytics

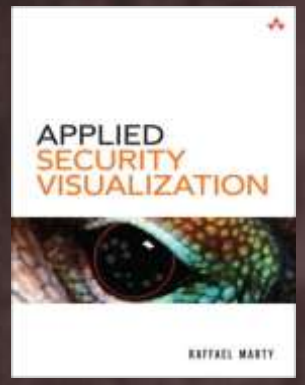
SOPHOS

Disclaimer

"This presentation was prepared solely by Raffael Marty in his personal capacity. The material, views, and opinions expressed in this presentation are the author's own and do not reflect the views of Sophos Ltd. or its affiliates."

Raffael Marty

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



- SecViz
- Logging
- Big Data
- ML & AI
- SIEM
- Leadership
- Zen

The master of Kennin temple was Mokurai. He had a little protégé named Toyo who was only twelve years old. Toyo saw how students entered the masters room each day and received instructions and guidance in Zen. The young boy wished to do zazen (meditation) as well. Upon convincing Mokuri, he went in front of the master who gave him the following koan to ponder:

"You can hear the sound of two hands when they clap together," said Mokurai. "Now show me the sound of one hand."

Outline

- Big Data for Security
- A Security (Big) Data Journey
- Machine Learning and Artificial Intelligence
- Data Visualization
- Solving Security Problems with Data
- A Glimpse Into the Future
- My 5 Security Big Data Challenges

Big Data For Security


“**memory** has become the new hard disk,
hard disks are the **tapes** of years ago.”
-- unknown source

Security Data

Data

- infrastructure / network logs (flows, dns, dhcp, proxy, routing, IPS, DLP, …)
- host logs (file access, process launch, socket activity, etc.)
- HIPS, anti virus, file integrity
- application logs (Web, SAP, HR, …)
- metrics
- configuration changes (host, network equipment, physical access, applications)
- indicators of compromise (threat feeds)
- physical access logs
- cloud instrumentation data
- change tickets
- incident information

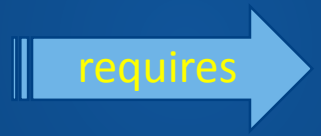
Context

- asset information and classification
- identity context (roles, etc.)
- information classification and location (tracking movement?)
- HR / personnel information
- vulnerability scans
- configuration information for each machine, network device, and application

Big Data Systems – A Complex Ecosystem

Use-cases

- Situational awareness / dashboards
- Alert triage
- Forensic investigations
- Incident management
- Reports (e.g., for compliance)
- Data sharing / collaboration
- Hunting
- Anomaly detection
- Behavioral analysis
- Pattern detection
- Scoring



Storing any kind of data

- Schema-less but with schema on demand
- Storing event data (time-series data, logs)
- Storing metrics

Data **access**

- Fast random access
- Ad-hoc analytical workloads
- Search
- Running models (data science)

Data **processing** needs

- Metric generation from raw logs
- Real-time matching against high volume threat feeds
- Anonymization
- Building dynamic context from the data
- Enrichment with entity information

Are Today's Systems Ready For Big Data Use Cases?

Data Sources

- Haven't been built with analysis in mind
- Logs are incomplete
- Log formats are not standardized

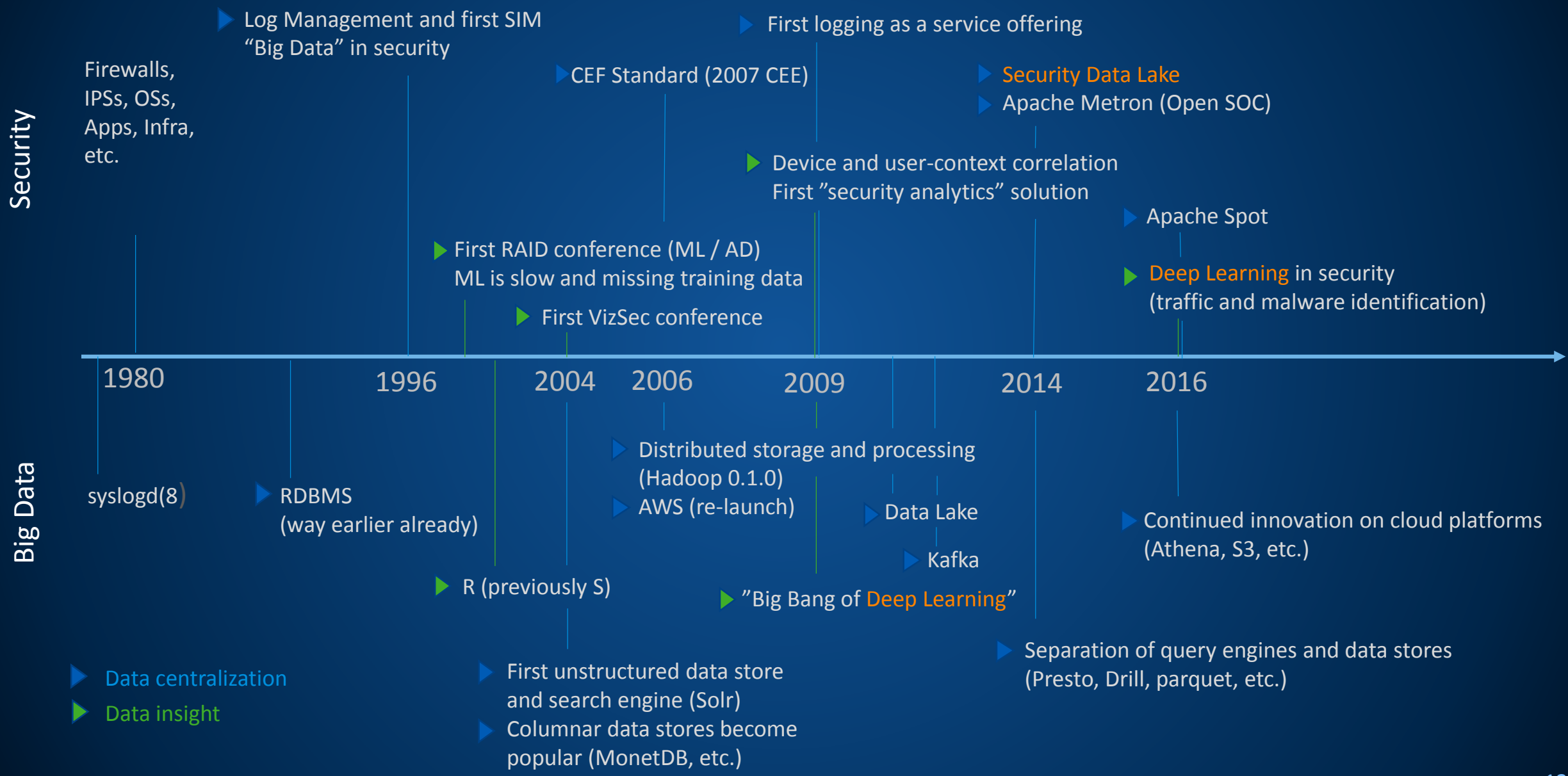
Log mgmt | SIEM | "Big Data Lakes"

- Don't scale well to volumes, variety, and velocity
- No standard data pipelines – results in point to point integrations that are imperfect
- No standard storage concepts – results in data duplication
- No standard use-cases – results in 'spaghetti architectures'

Image credit: http://journeyofhealth.org/

(Incomplete) Security Data History

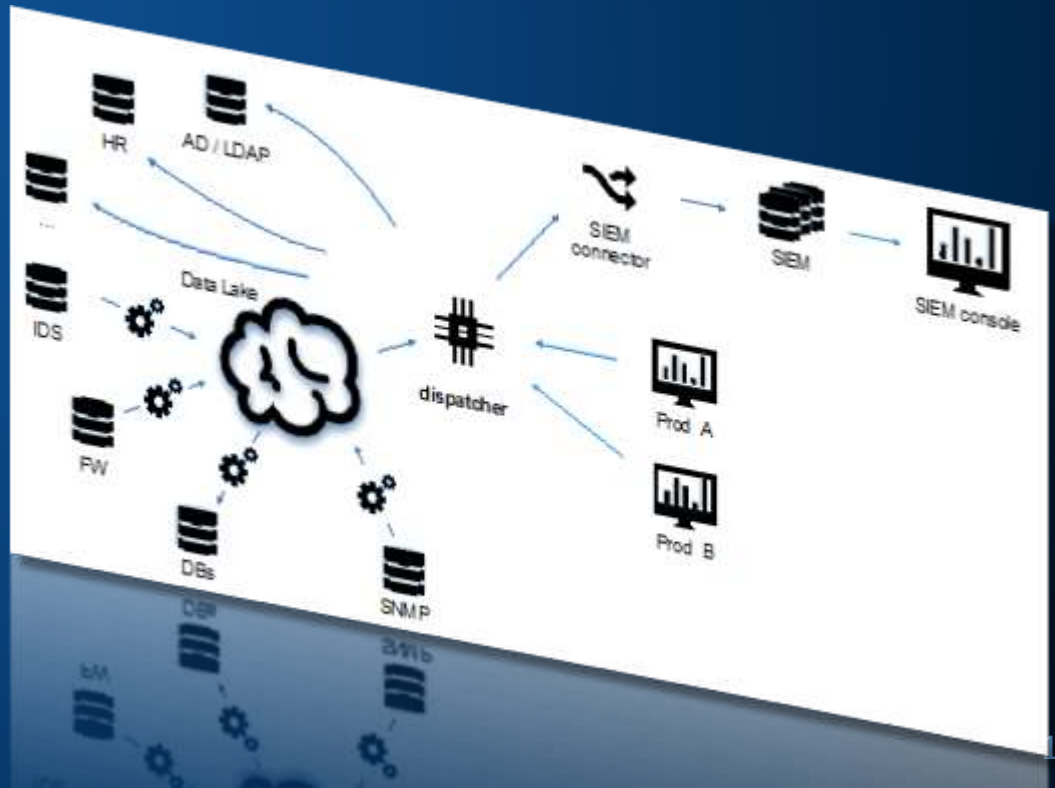
"Big Data Is An Old Problem in Security"



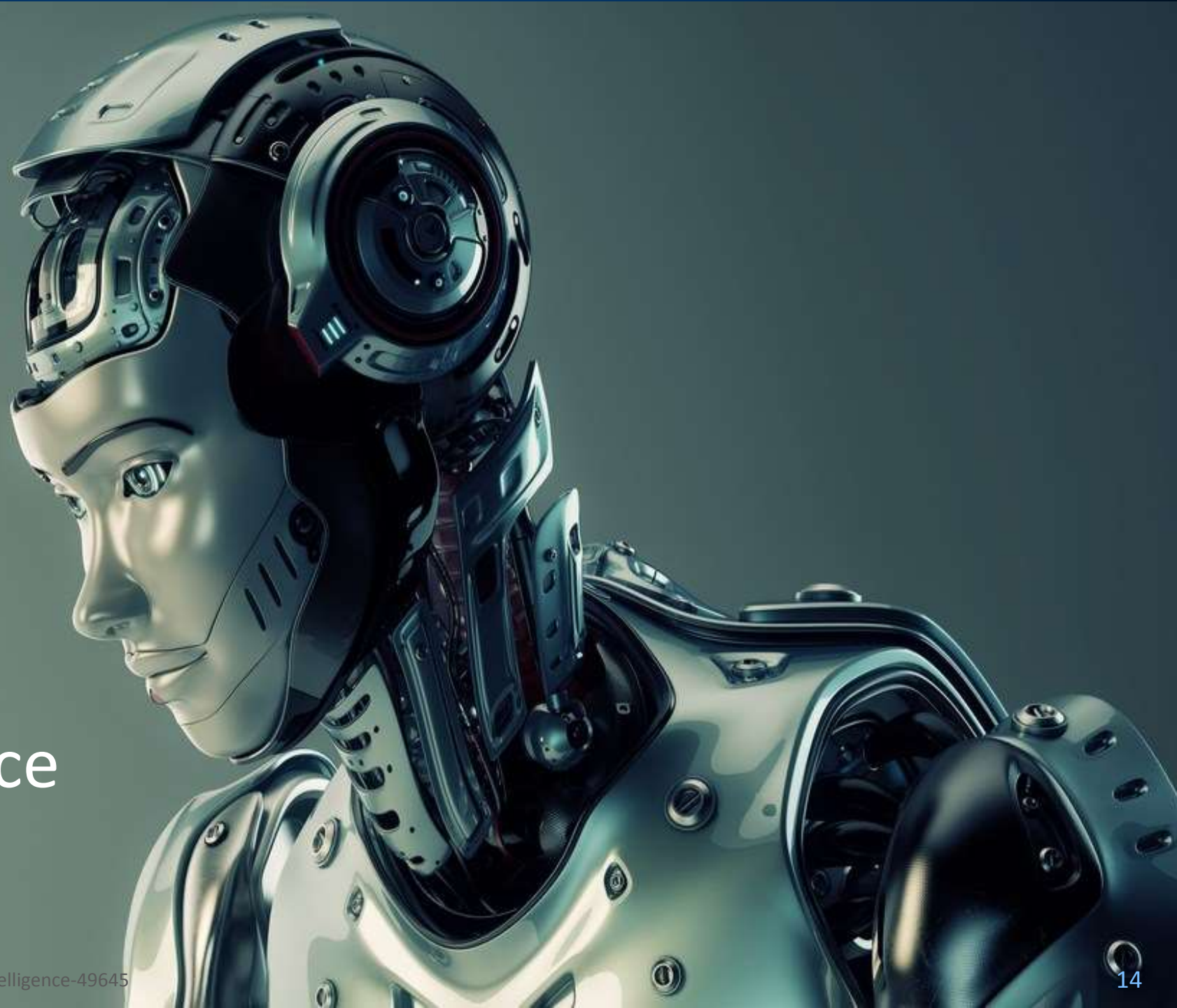
Security Data – The State Today

- *"Security Data Lakes – an excuse to collect anything without having to think about schemas and access patterns."*

- Data and infrastructure challenges to overcome
  - Data standardization (parsing, schemas)
    - Meaning of log entries and fields within
    - When is a log generated, when not?
  - Data infrastructure
    - One architecture for all use-cases
    - Self maintaining and healing
  - Building 'content' across customers?
    - Different policies
    - Different data sources and configurations
  - Data Privacy



Data Science



Data Mining
Machine Learning
Artificial Intelligence

http://theconversation.com/your-questions-answered-on-artificial-intelligence-49645

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
  - Verifiability issues
- **Data Mining** – Methods to explore data – automatically and interactively
- **Artificial Intelligence** – “Just calling something AI doesn’t make it AI.”

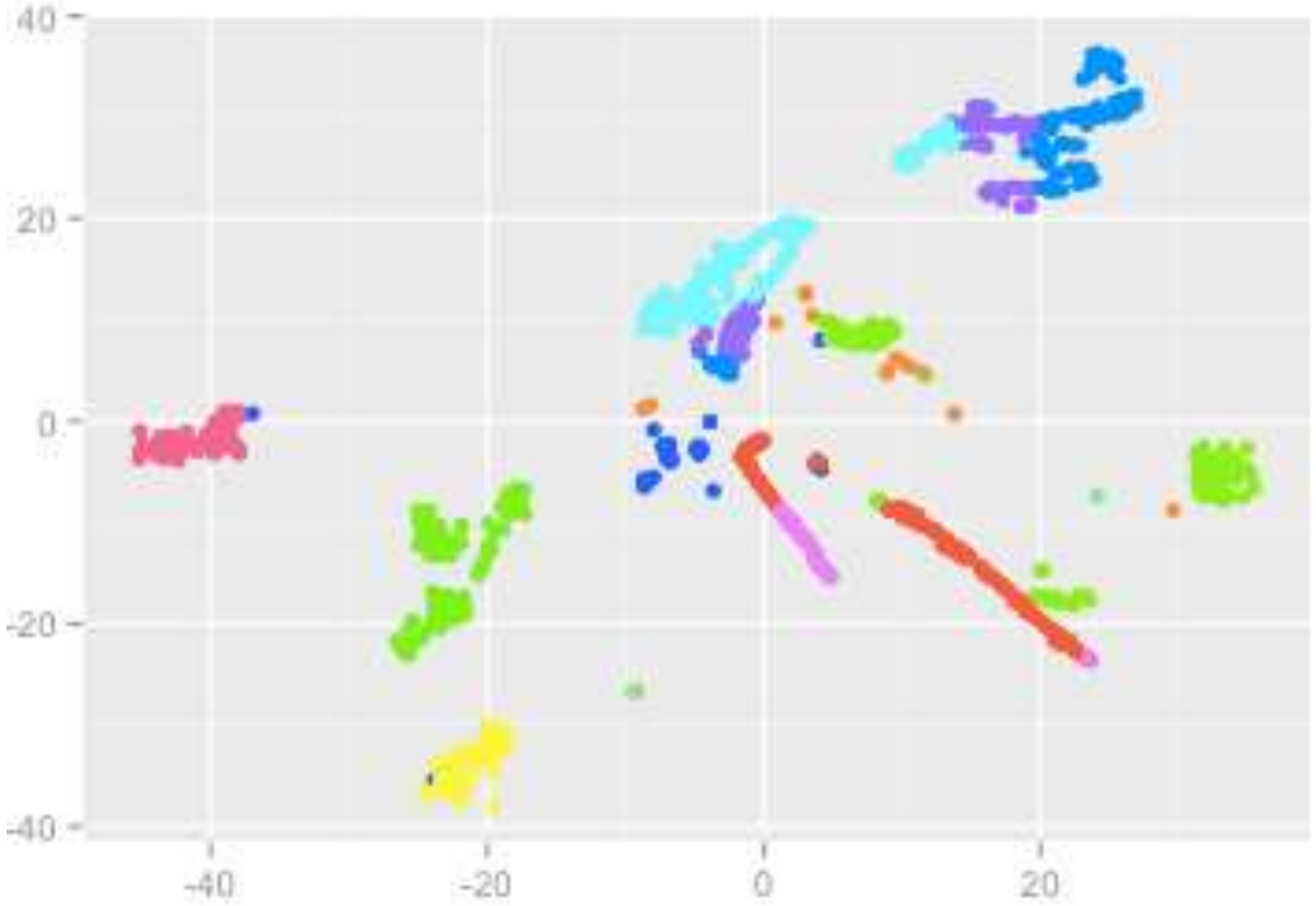
*”A program that doesn't simply classify or compute model parameters, but comes up with **novel knowledge** that a security analyst finds insightful.”*

Machine Learning 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**
- Unsupervised
  - Tier 1 analyst automation (reducing workload from 600M events to 100 incidents)*
  - User and Entity Behavior Analytics (UEBA)
    - Uses mostly regular statistics and rule-based systems

* See Respond Software Inc.

Application of Machine Learning - Anomaly Detection



Objective : Find ‘security incidents’ in the data – deviations from the ‘norm’

- What’s “**normal**”?
- Needs **explainability** for clusters
- Observe clusters **over time** (requires stable ‘incremental’ clustering)
- Even 0.01% of **false positives** are too high (1m log records -> 100 anomalies)

Limits of Machine Learning

"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"

Machine Learning Challenges

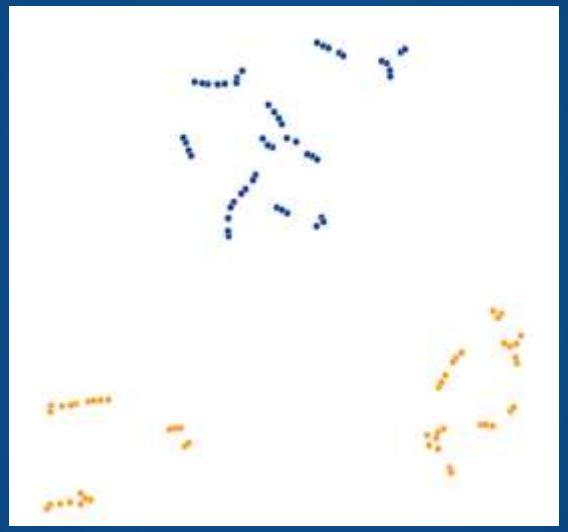
- An algorithm is not he answer. It's the **process** around it (find the best fit algorithm for the data and use-case, feature engineering, supervision, drop outs, parameter choices, etc.)
- Even in deep learning, it's not just about using tensorflow. **Features matter** (e.g., independent bytes versus program flow)
- The algorithms are only as good as **the data** and the knowledge of the data
  - Common data layers / common data models
  - Enriched data
  - Clean data (e.g, source/destination confusions)
- How do we build systems that incorporate expert knowledge?

Illustration of Parameter Choices and Their Failures

- t-SNE clustering of network traffic from two types of machines

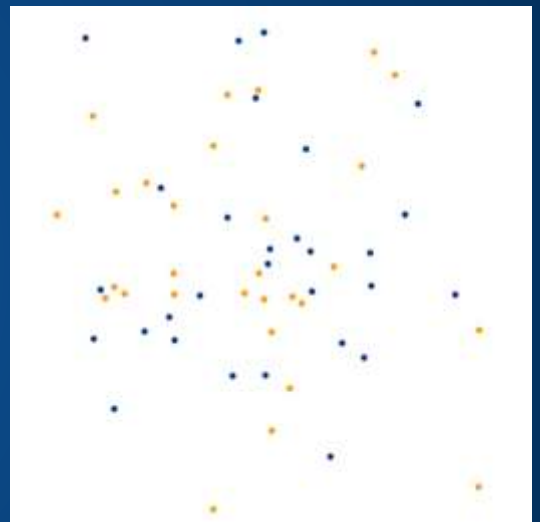
perplexity = 3
epsilon = 3

No clear separation



perplexity = 3
epsilon = 19

3 clusters instead of 2

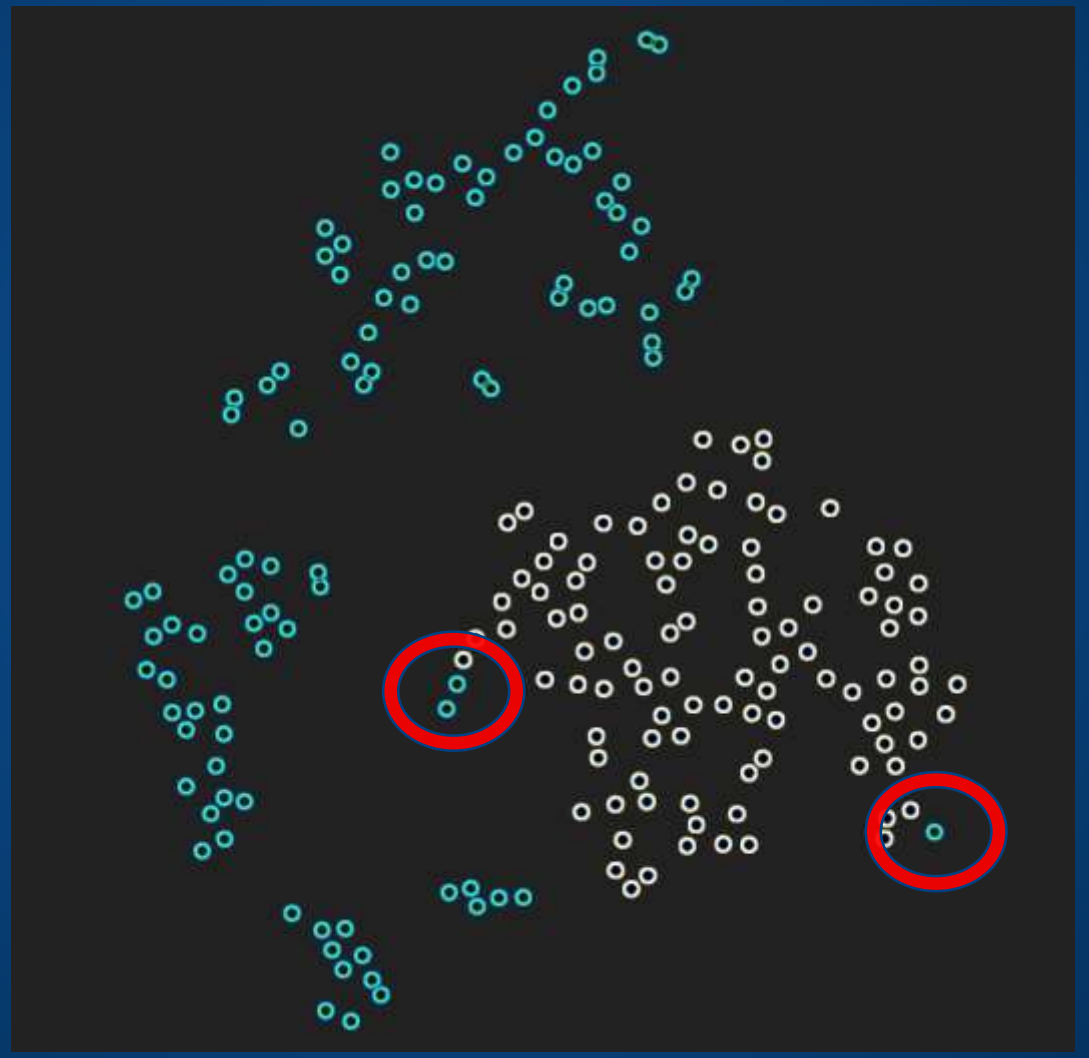


perplexity = 93
epsilon = 19

What a mess

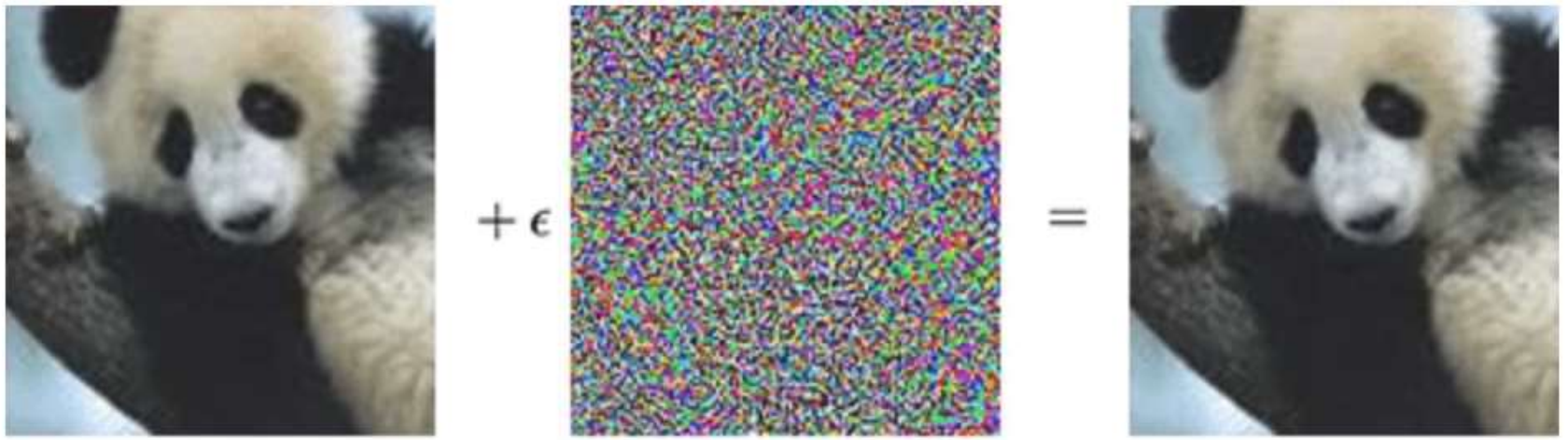
Illustration of Parameter Choices and Their Failures

- Dangerous clusters

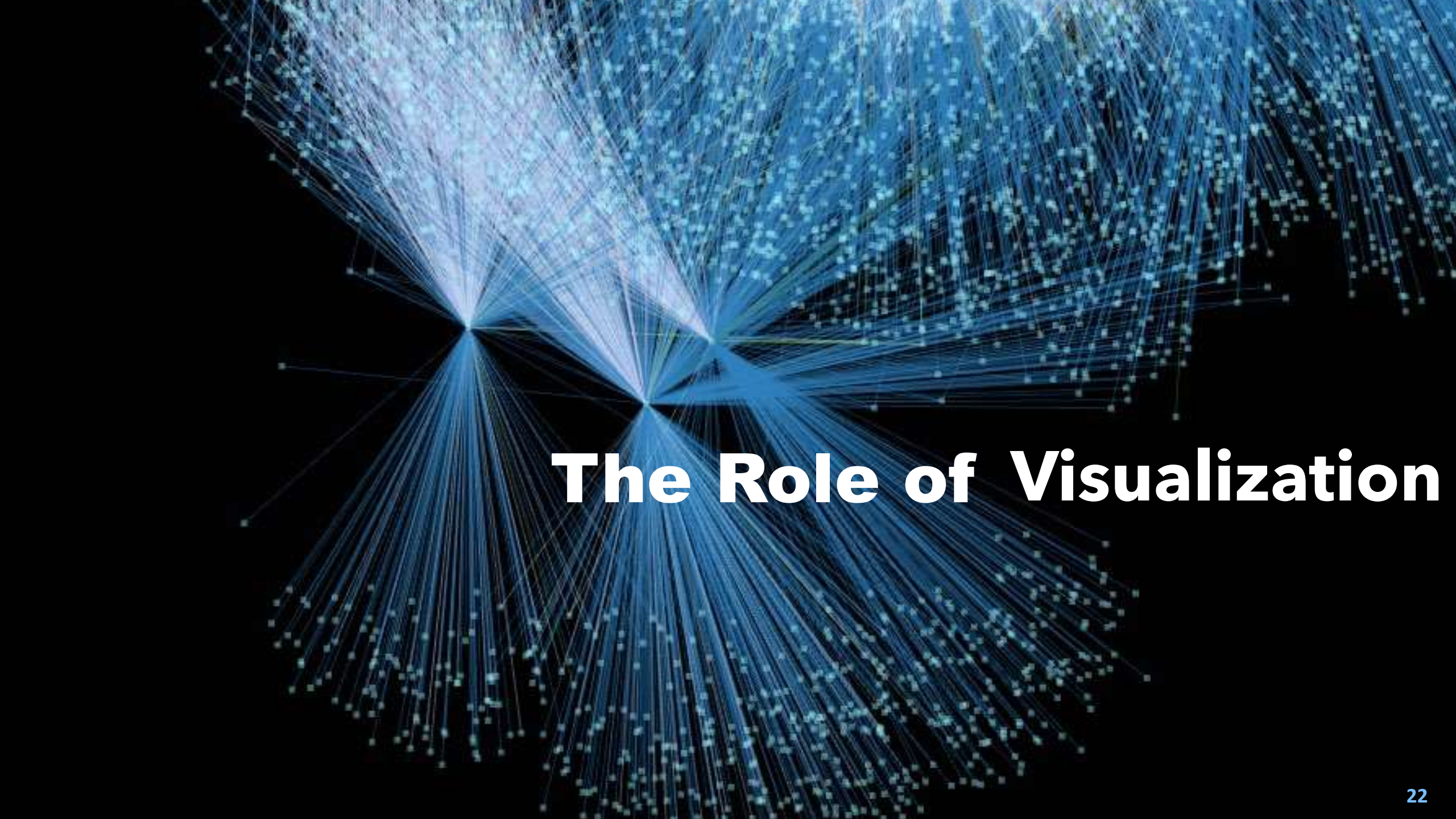


Adversarial Machine Learning

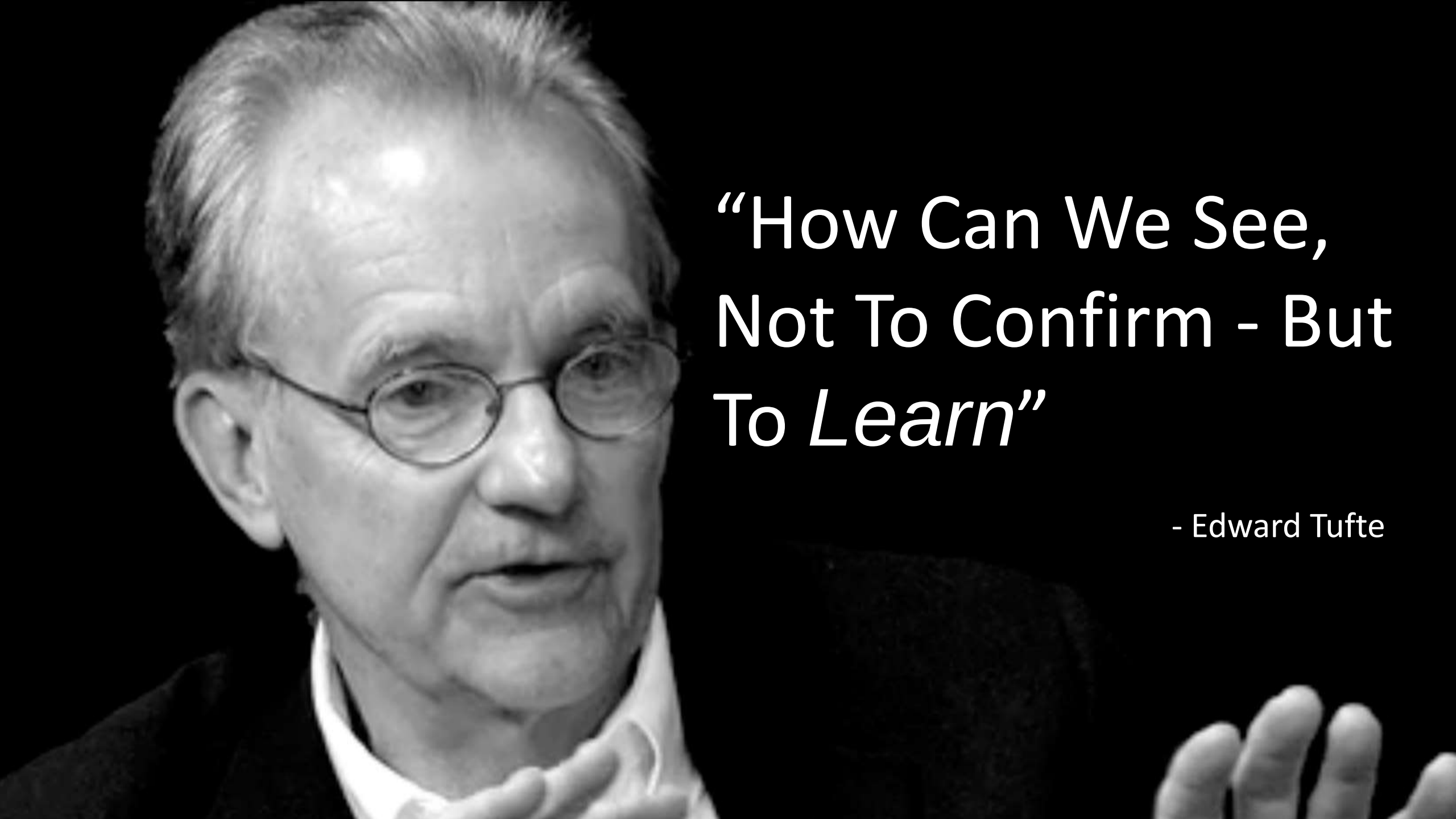
- An example of an attack on deep learning



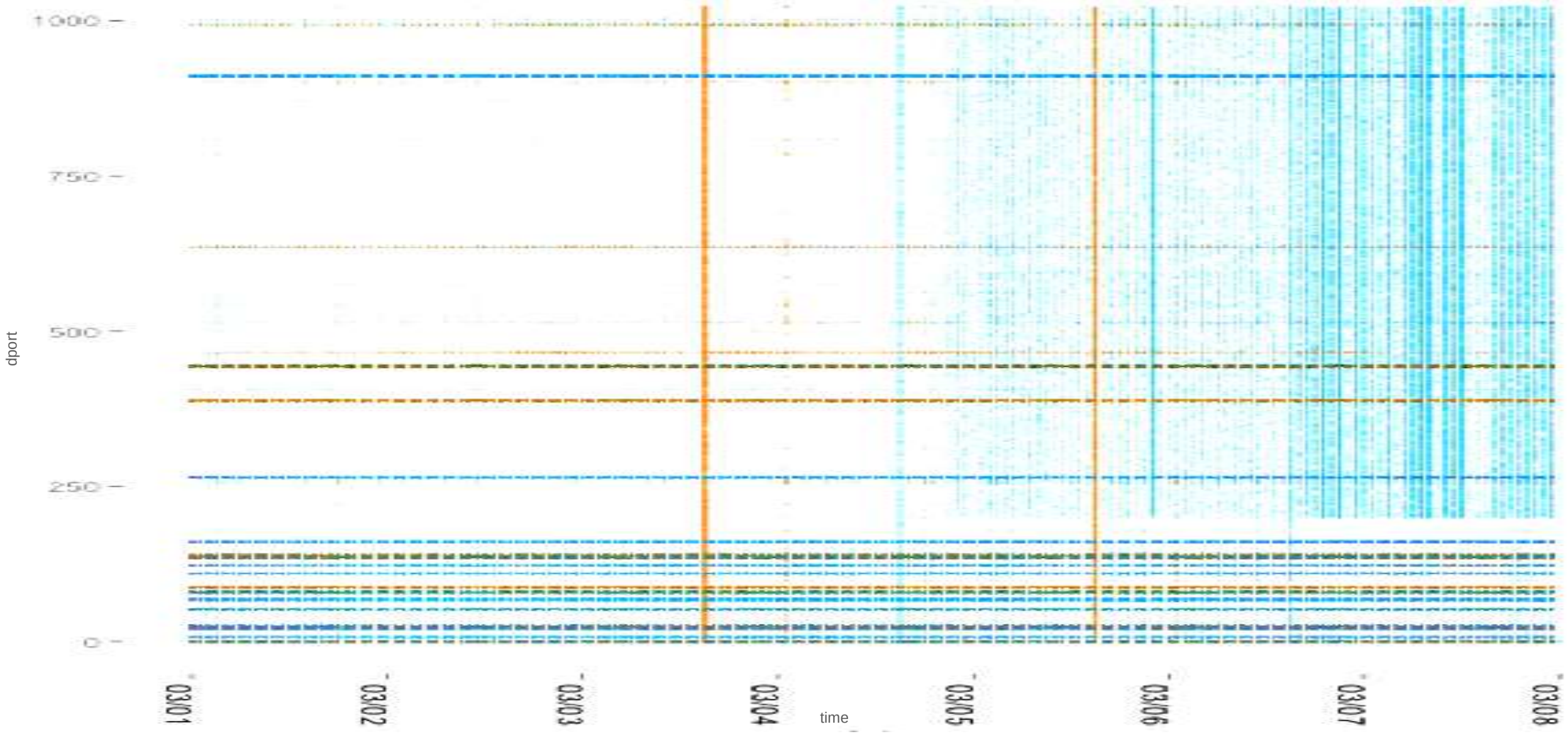
Above: Image Credit: Ian Goodfellow



The Role of Visualization


“How Can We See, Not To Confirm - But To *Learn*”
- Edward Tufte

Why Visualization?

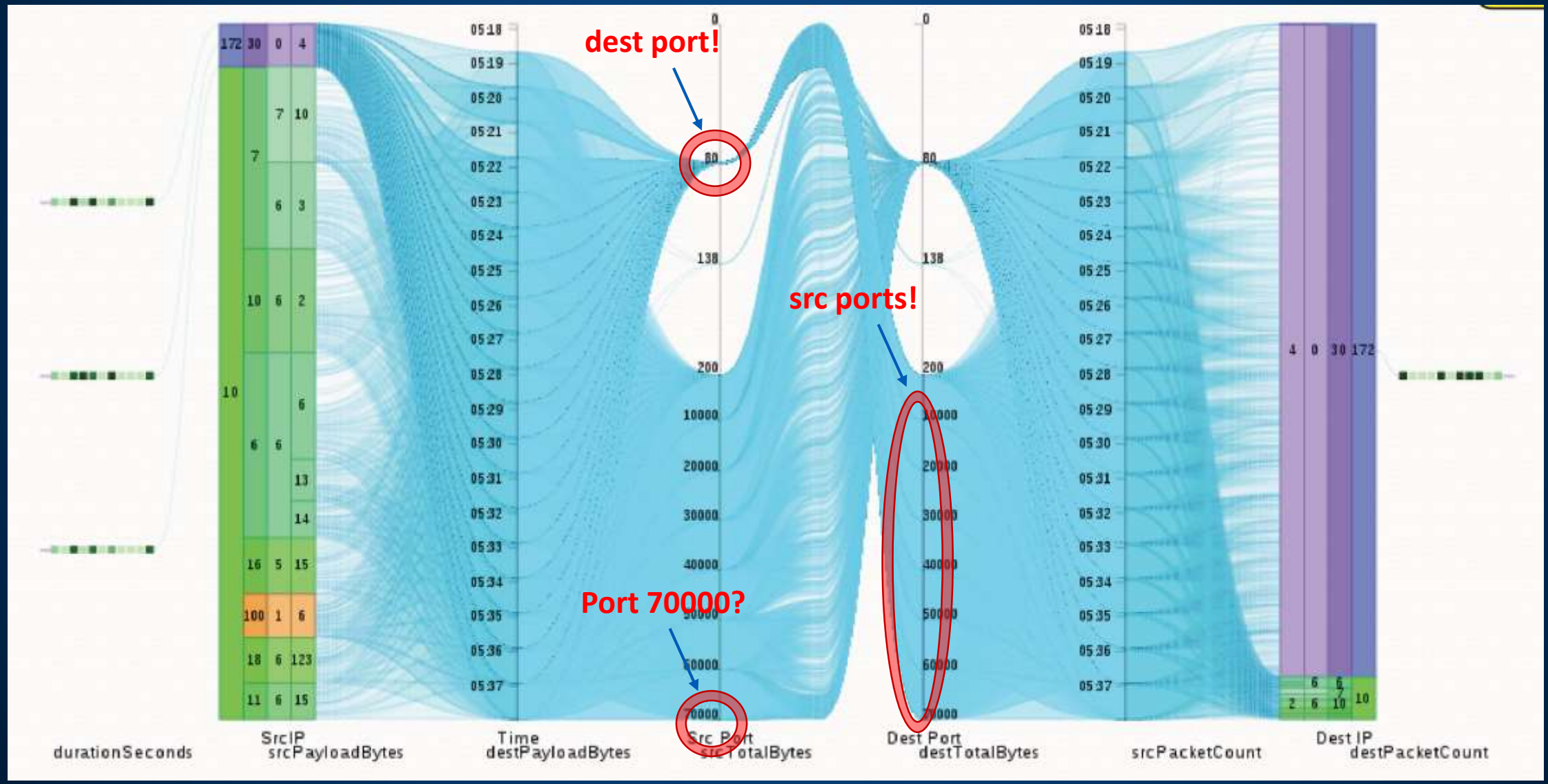


Visualization Overview

- Why?
  - Verify output of machine generated intelligence
  - Focus experts where they are most useful, rather than having them build tools / queries to understand the data
  - Enable exploration and hunting
- What are the limitations?
  - Data is always a problem – we need clean, enriched data
  - Visualization of large data sets
  - Interpretation is hard
    - "And the single port with no traffic is port 0, which is reserved [24]" found in "Visualization of large scale Netflow data" by Nicolai H Eeg-Larsen
    - "… and the destinations are Internet Web Server or DNS server or both with the port 0."
    - ".. so many TCP port scans are distributed in the whole day that most of them can be considered as false positives."

      https://www.researchgate.net/publication/257686749_IDSRadar_A_real-time_visualization_framework_for_IDS_alerts

VAST Challenge 2013 Submission – Spot the Problems?



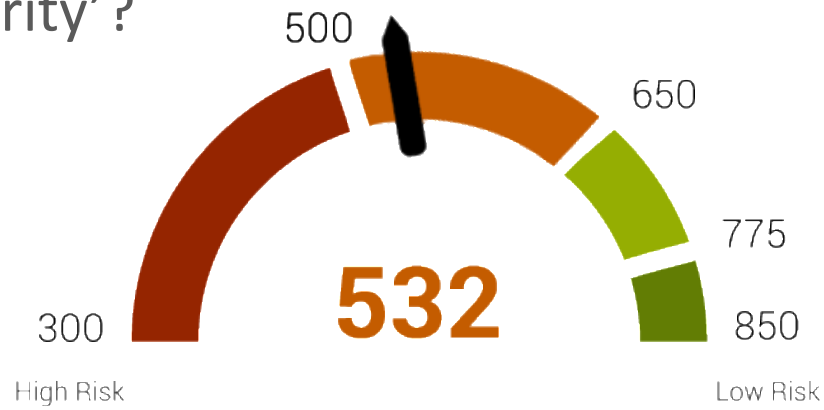
http://vis.pku.edu.cn/people/simingchen/docs/vastchallenge13-mc3.pdf

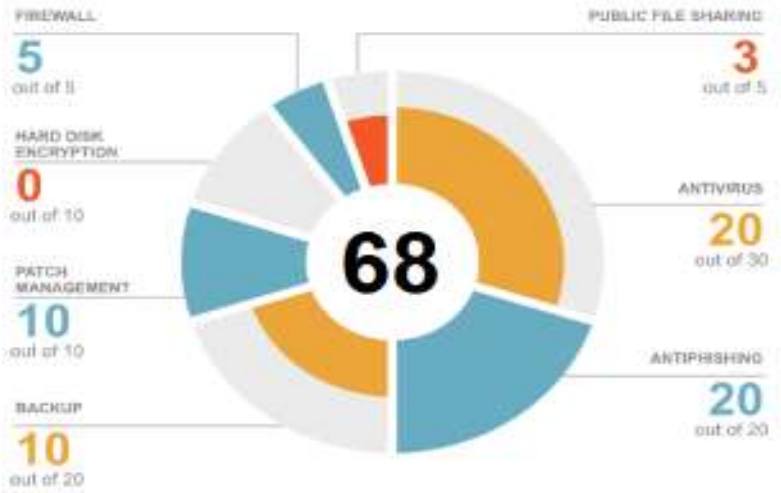
Visualization Challenges

- Backend
  - Super quick data access in any possible way (search, scan, summarize)
  - Ability to ingest any data source - intelligent parsing anyone?
- User Interface
  - The right visualization paradigms
  - How to visualize 1m records?
  - The right data abstractions / summarizations / aggregations
  - Easy to use and still flexible enough
- Data Science
  - Make the machine help us interpret the data
- How to encode domain knowledge?

Visualization Challenges - Security Metrics

- How to quantify ‘security’?





- Provide context

Solving Security Problems With Data

Solving Security Problems With Data

Objective: Automatically detect “problems” / attacks with data

Solution: *Not* ML or AI – the right process for the problem at hand

- Any data science approach:
  - Encode domain knowledge – leverage trained experts (e.g., malware classification with n-grams, or URLs)
  - Involve the right ‘entities’ (e.g., push problems out to the end user)
  - Collect the right data for the given use-cases – don’t forget context and cleaning
  - Plan for expert feedback / validation loop
  - Build solutions for actual problems with real data that produce **actionable insight**
  - Share your insights with your peers – security is not your competitive advantage
- Supervised:
  - Be selective on the problems that have good, large training data sets
- Unsupervised:
  - We need good distance functions. Ones that encode domain knowledge!

Applications of Data in Security

Data → Data Operations → Applications

- Prioritize event and entity data
- Rule-based correlations
- Behavior modeling
- Risk / exposure / threat computation
- Configuration assessments
- Data classification
- Data abstraction
- Cross ‘boundary’ data sharing
- Cross ‘customer’ analytics
- Crowd intelligence
- Enable free-form exploration

- Identify and attribute attacks
- Incident response
- Improve prevention
- Allocate / prioritize work / resources
- Situational awareness
  - Understand exposure
  - Risk inventory
- Spam, malware detection
- Feedback loop on initiatives
- Simplify security
- Continuous attestation
- Micro segmentation
- Risk informed, dynamic enforcement (automation)

Data is a core driver for many or most security use-cases

A Glimpse Into The Future

http://www.aberdeenessentials.com/techpro-essentials/business-leaders-can-utilize-data-even-without-technology-background/

My Magic 8 Ball

- Data is distributed across the edge and (a) central data store
  - We will have a (data lake)++ in every company with all security data (likely in the cloud)
  - Centralize data for correlation (could we get a decentralized correlation system?)
  - Keep raw sensor data at the edge and access through federated query system
  - Threat intelligence will be tailored to your organization and exchanged in real-time
- APIs will be everywhere to let products integrate with each other
- Security Analytics as a product category, as well as orchestration will merge with the data platforms (SIEM++)
- Algorithms take a back seat – insights are key
  - Nobody cares whether you call something artificial intelligence or machine learning. It's about actual results
  - Products will learn from users more and more
- Startups will deliver innovation, but only large organizations will be able to deliver on the overall security promise
- Detection is great. Protection is key. Closing the loop between insight and action.
  - Continuous attestation
  - Risk-based defense
- No 3D visualizations

Thoughts on How We Get There

- Focus on three types of users
  - Data scientists and hunters – that now how to program, have security domain knowledge, and can find complex insights
  - Security analysts – that are using product interfaces to deal with security issues that the system couldn't deal with automatically
  - Non security experts – that need insight into what is happening, but don't know enough to intervene
- AWS will productize the 'all encompassing data backend' (others will contribute the technology)
  - Abstracting the data storage layer
  - Self-optimizing and monitoring query engine
- Hire and train good UX people
- Hire and train security domain experts
  - "A course doesn't make you a data scientist – not a good one at least". It's about the domain knowledge!
- Use deep belief networks rather than deep learning
- Build systems that help analysts and exports be more effective
  - Don't try to replace them - let them do the interesting work
  - Don't make up use-cases. Go into organizations and learn what the real problems are
  - Understand the user personas you are catering to
  - Stop building islands of products – SA is a feature – how do we build that on top of a common platform?
  - Move away from algorithm thinking into use-cases and workflows
- Collect all your data (network and endpoint) in one data store

My 5 Challenges

https://play.google.com/store/apps/dev?id=5029488271380967378

My 5 Challenges

- Establish a pattern / algorithm / use-case sharing effort
- Define a common data model everyone can buy into (CIM, CEF, CEE, Spot, etc.)
  - Including a semantic component for log records, not just syntax
- Build a common entity store
  - Hooked up to a stream of data it automatically extracts entities and creates a state store
  - Allows for fast enrichment of data at ingest and query time
  - Respects and enforces privacy
- Design a great CISO dashboard (framework)
  - Risk and “security efficiency” oriented, actionable views
- Develop systems that ’absorb’ expert knowledge non intrusively

"You can hear the sound of two hands when they clap together," said Mokurai. "Now show me the sound of one hand."

Questions?

http://slideshare.net/zrlram
@raffaelmarty