RSAConference2019
San Francisco | March 4–8 | Moscone Center

SESSION ID: AIR-F01

Use Model to Deconstruct Threats: Detect Intrusion by Statistical Learning

Tao Zhou

Senior Staff Algorithm Engineer
Alibaba Security, Alibaba Group

RSAConference2019

Challenge: Security Data Analysis in Internet Giants

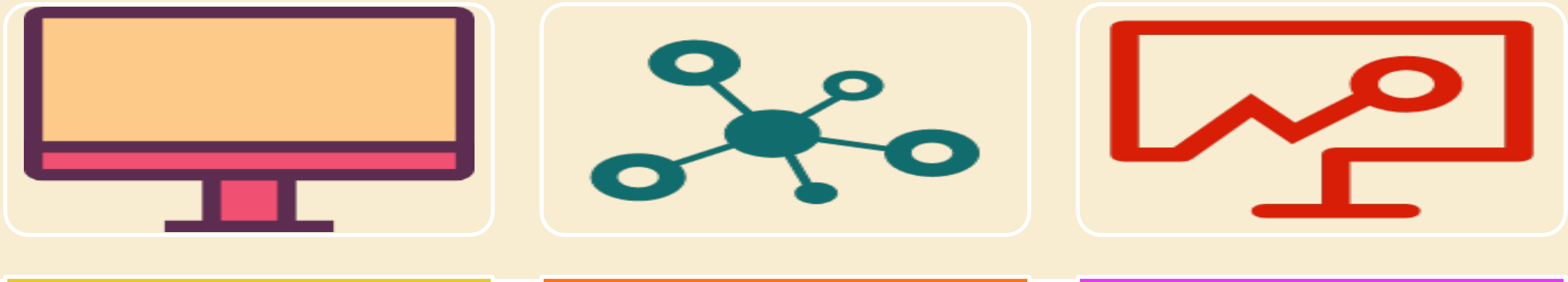
Hacker's Intrusion and Data Breach

- Data leakage brings huge losses to Internet giants.

Attack and Defense is the Contest Between Humans

- What's the advantages on defense side?
  - Control over network infrastructure
  - Ability to deploy security devices and data collection instruments

Data is the Biggest Advantage of Defenders!



Endpoint:

- AV
- EDR
- HIDS
- RASP...

Network:

- NGFW
- IDS
- Access Gateway
- Audit...

Business System:

- DNS
- Mail
- IAM
- VPN...

Gap Between Data Collection and Security Operations

We upgrade the security monitor ability! Now we collect more than 50 kinds of devices, and produce 5 billion records per hour!

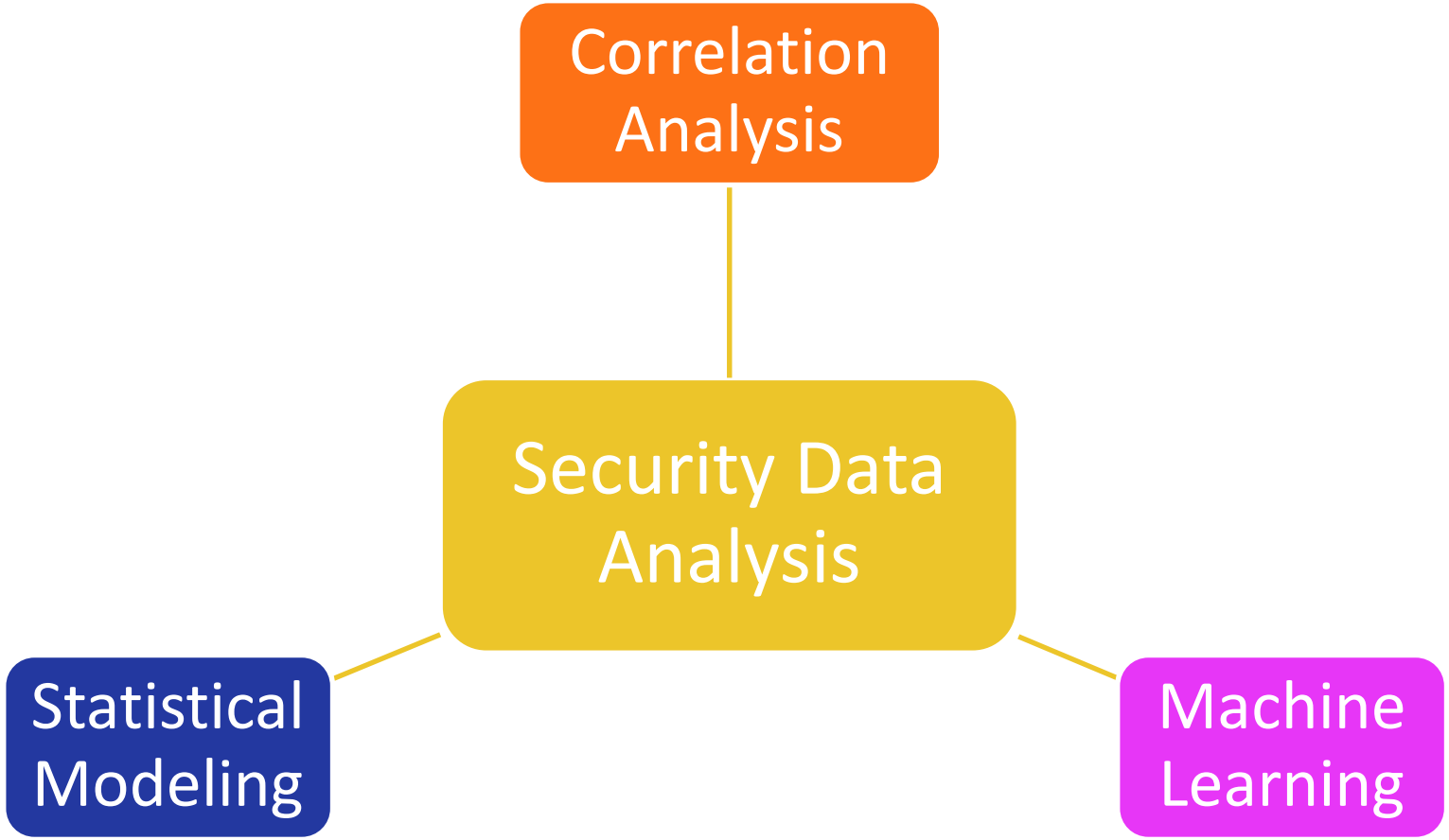
Network administrator

Well, I can handle up to 100 alarms per day…

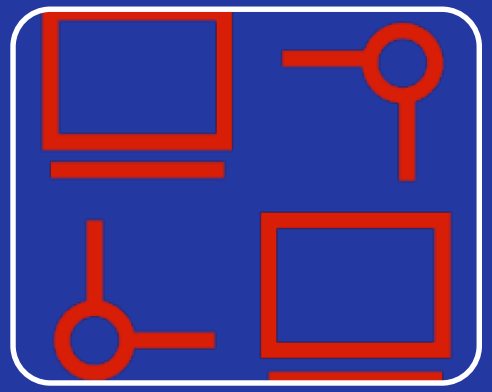
Security Operator

How To Handle the Gap?

Different Modes in Security Data Analysis

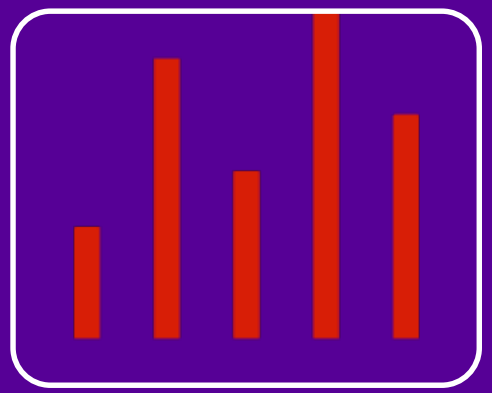


Machine Learning vs. Statistical Modeling



Machine Learning: Rely on data and algorithms

- Large amount of labeled data
- Feature engineering / log embedding
- Proper learning algorithms

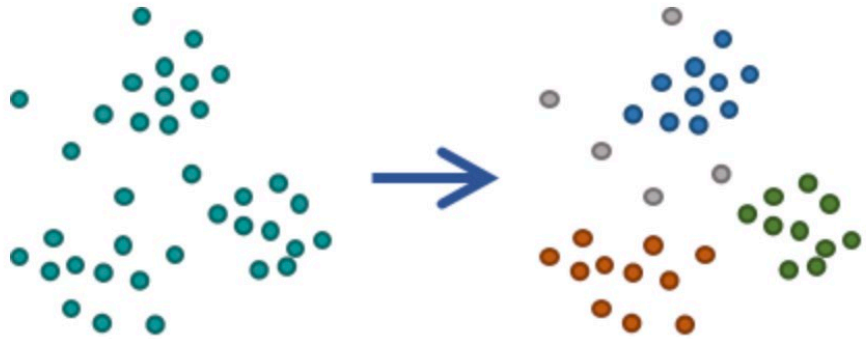


Statistical Modeling: Rely on human's experiences

- Find common trait of attack behavior
- Feature engineering
- Proper statistical algorithms

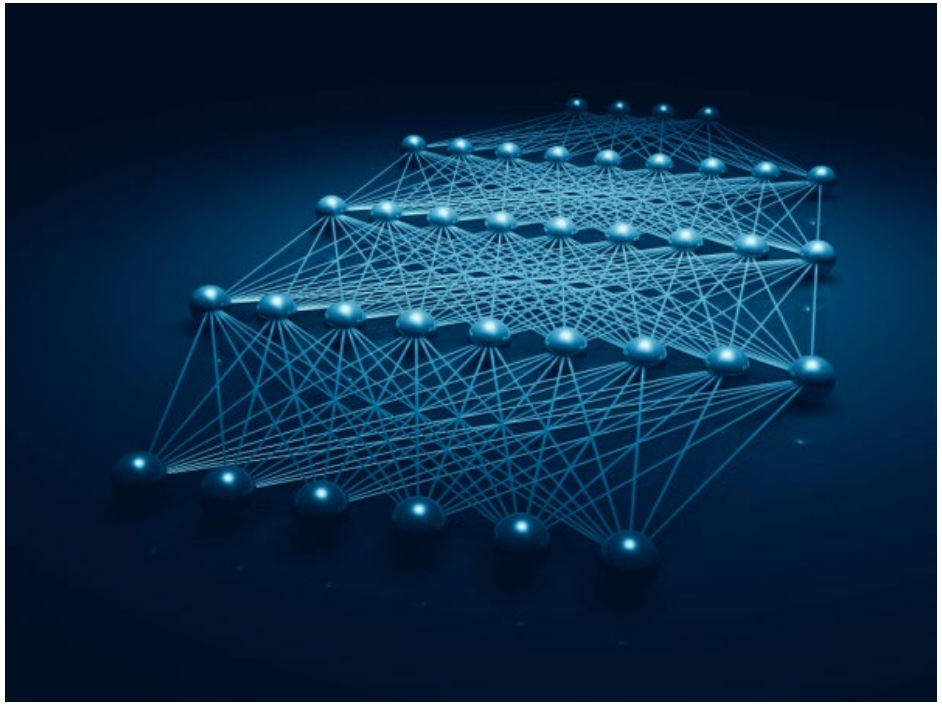
Machine Learning Has Less Opportunity in Intrusion Detection

- Machine Learning is good at finding normal patterns, but intrusion is abnormal behaviors.
  - Can't simply think that abnormal data is left after normal is removed.
- ‘Big data’ is not equivalent to ‘Big labeled data’
  - The accuracy and recall rate of unsupervised learning can’t support security operations.



Machine Learning Has Less Opportunity in Intrusion Detection

- An open domain problem of intrusion detection.
  - It's difficult to define a proper loss function to judge whether a record represent intrusion or not.
- Interpretability of results.
  - Only the answer of 'yes or no' is not enough for security analysis.



Suitable Scenarios for Machine Learning in Intrusion Detection

- Specific area.
- Easy to cumulate labeled data.

Spam detection

DGA domain detection

Web Crawler detection

RSAConference2019

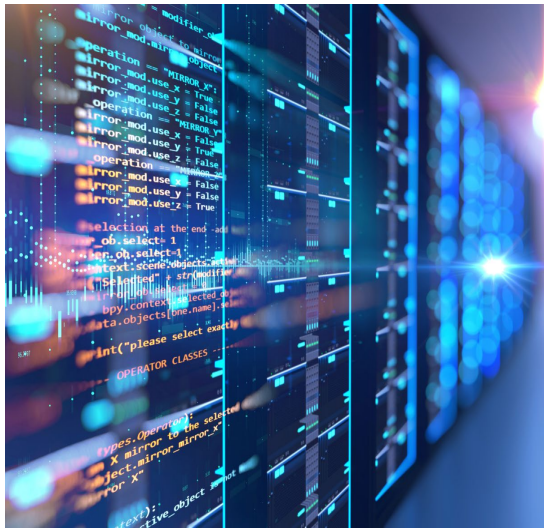
Solution: Use Statistical Modeling to Deconstruct Threats

Key Point: Intrusion Trace Detection

- Rome wasn't built in a day: effective attacks always take a long process.

Long Time

Multiple Stages



Multiple Nodes

The Process of Security Data Analysis

Data preprocessing

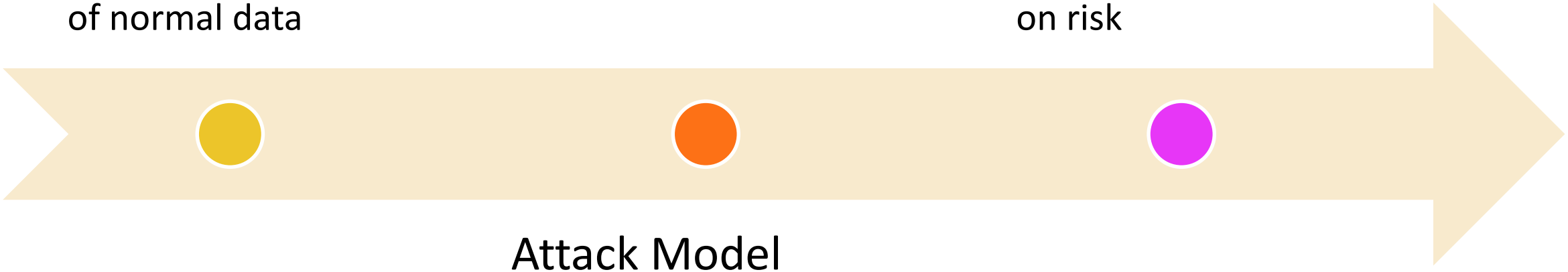
- Remove interference of normal data

Attack Model

- Identify the suspicious behavior

Alarm Correlation

- Prioritize alarms based on risk



Data Preprocessing

Normal-behavior oriented model:
Repetitive behavior is always normal.

Filter out the normal data with the largest proportion.

Recall is the most important indicator.

Demo: the Coarse Filter of HIDS Log

- Input: Log of HIDS
  - including records of process, file and network connection.
- Output: abnormal behavior of host

Object type	Factors Should be consider	How to quantize
Process	Is derived relationship normal?	Transition probability between processes
File	Is R/W operation normal?	Read and Write score of file
Connection	Is the connection normal?	The popularity of TCP connection

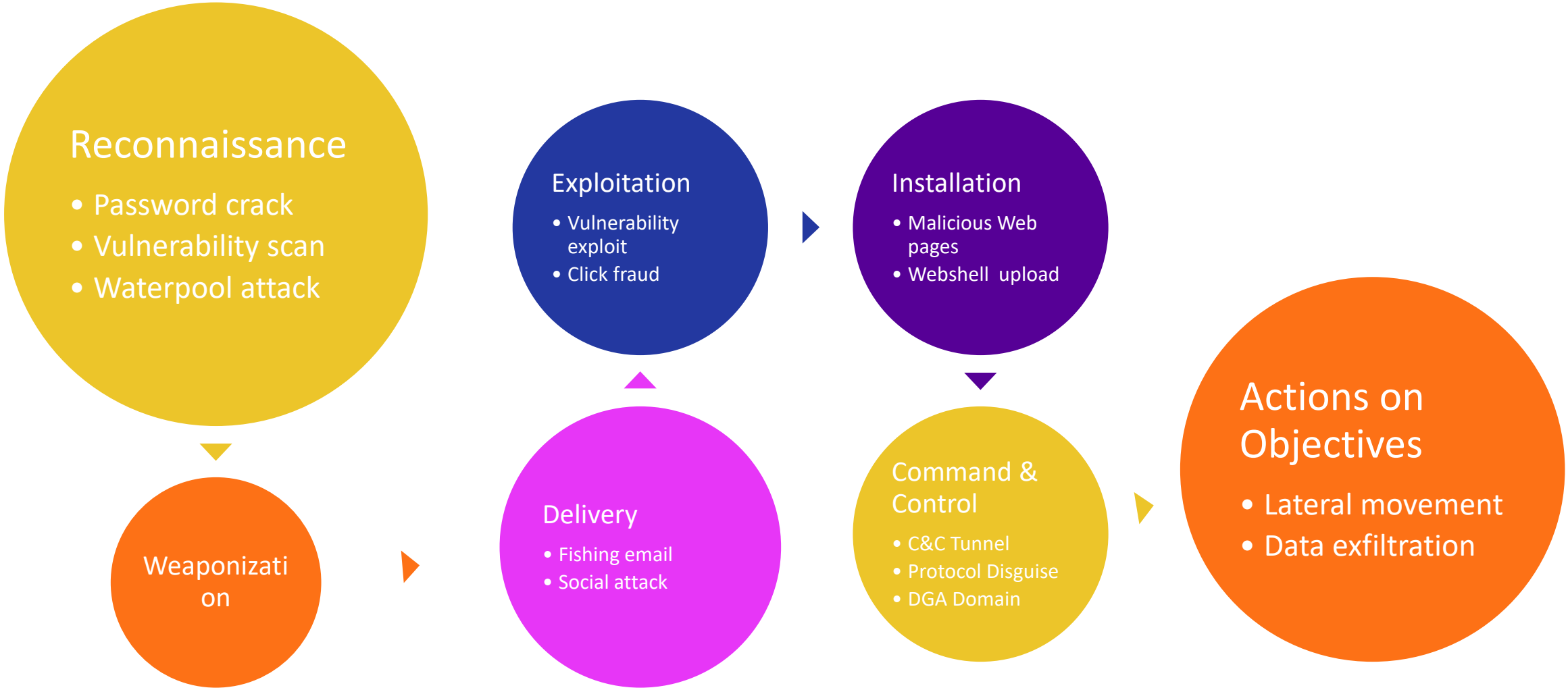
Attack Model by Statistics

Attack oriented model: Same kind of attack tends to have common features.

Retrieve the more attack-like behavior form abnormal behavior.

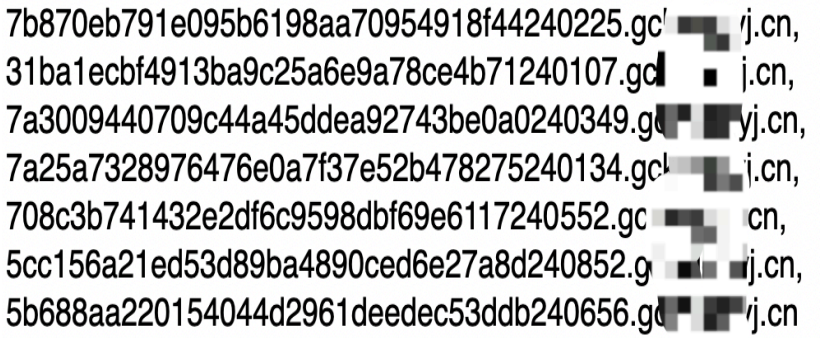
Precision is the most important indicator: False-positive is expensive.

Kill Chain Based Model Construction



Demo: DNS Tunnel Detection

- Input: Log of DNS Server
  - Records of DNS request
- Output: Suspicious DNS tunnel connection
- Common feature
  - Entropy of sub domain name
  - Number of Individual host visits the domain
  - Lifetime of the root domain



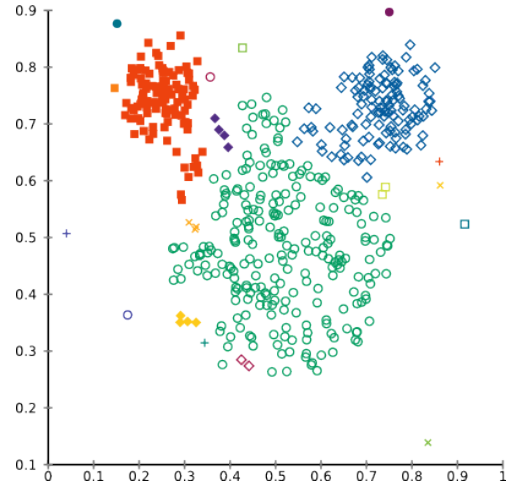
Demo: Command & Control Tunnel Detection

- Input: Network flow record
  - Including source IP/Port, destination IP/Port, time, size, etc.
- Output: Suspicious Trojan victim and controller connection.
- Common feature
  - The heartbeat between victim and controller, always have the same destination IP/Port, same time interval, and same size.
- Detection algorithm
  - stability of time interval sequence, can be achieved by FFT algorithm, or ratio between mean and variance.

Demo: DGA Domain Detection

- Input: Log of DNS Server
  - Records of DNS request
- Output: Suspicious DGA domain name
- Key Point
  - Machine learning only detects the randomness of domain name text, no matter use LSTM, N-Gram, GBDT or something else.
  - Should be re-filtered by behavior feature, such as request number, sub domain number, etc.

earnestnessbiophysicalohax.com
kwtoestnessbiophysicalohax.com
rvcxestnessbiophysicalohax.com
hjbtestnessbiophysicalohax.com
txmoestnessbiophysicalohax.com
agekestnessbiophysicalohax.com
dbzwestnessbiophysicalohax.com
sgjxestnessbiophysicalohax.com
……



Insufficiency of Statistical Modeling

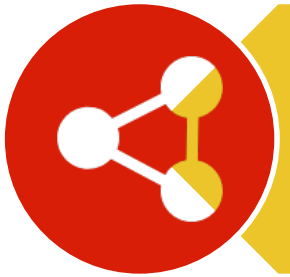
- High false-positive rate
  - Something ‘like an attack’ is not always real attack.
  - Some legal systems resemble attack behaviors in specific aspect.

Adding printer may cause network scanning.

Some cloud-based service may have heart beat behavior.

Anti-virus tools may use DNS tunnel to sample suspicious file.

Enhancement: Post Correlation of Events



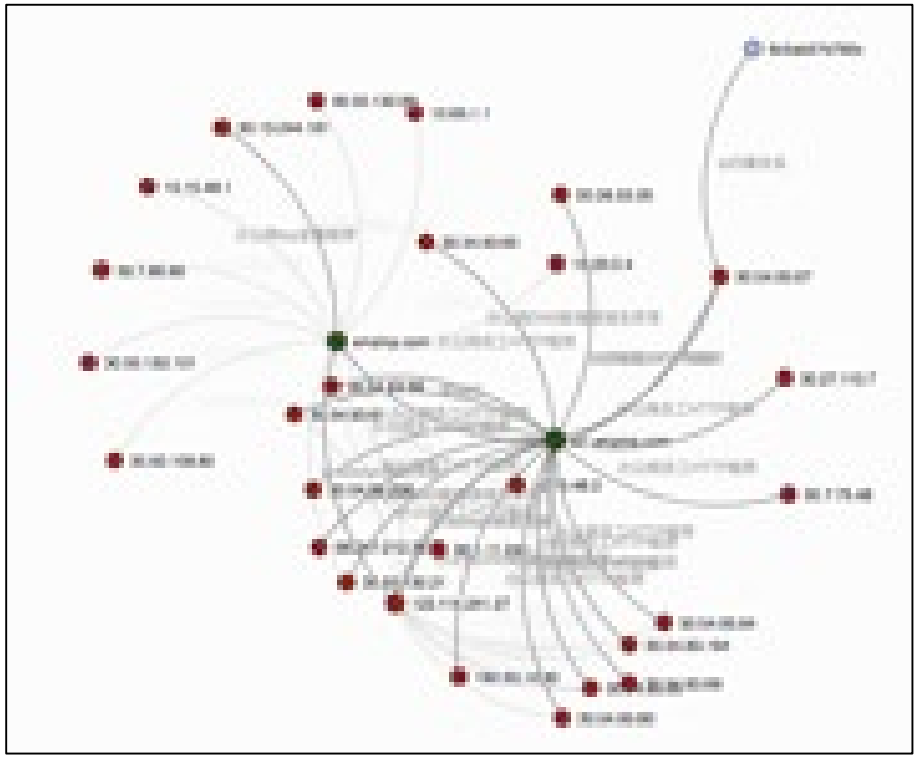
Graph Based attack path discovery and risk prioritization.

Distill high confident events from low accurate alarms, and build complete attack scenario.

Precision is the most important indicator: False-positive is expensive.

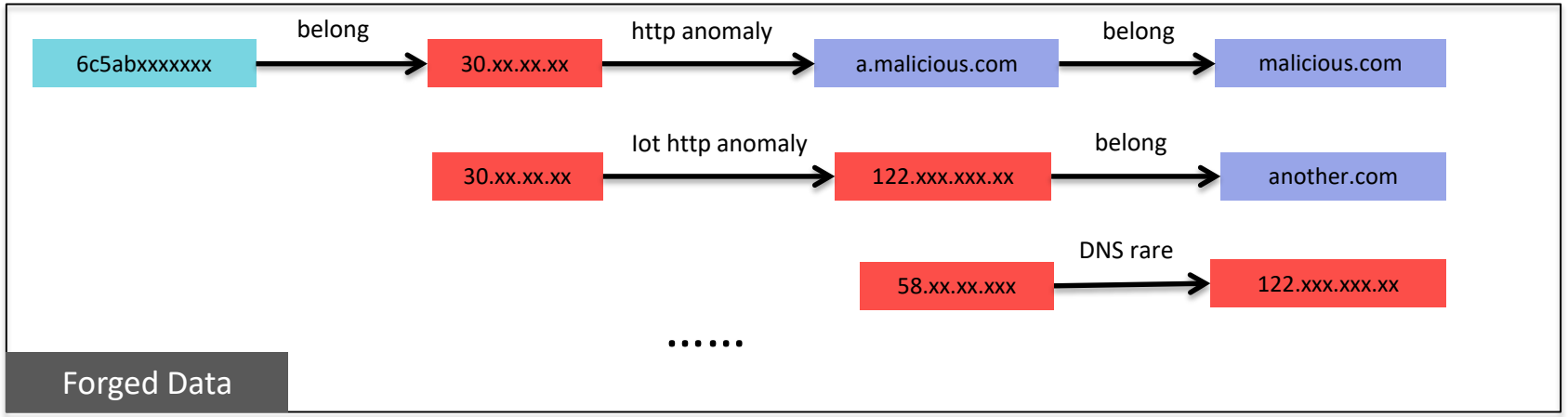
Build Attack Graph From Alarms

- Attack Graph: A directed Graph
  - Node: Asset, IP Address, Website
  - Edge: Attack relationship, or risk's propagation.
  - A pair of nodes can only have one edge, no matter how many alarms exist between them.

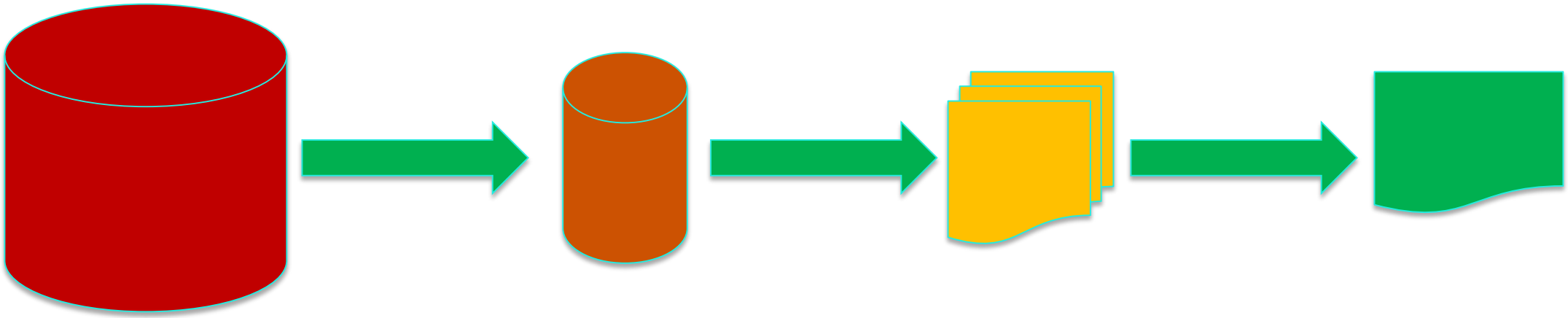


Path Discovery and Risk Prioritization

- Risk of an attack path should be quantized by:
  - Different attack phases
  - Network distribution of assets
  - The risk and accuracy of each alarm



Results and Performance



Raw Data: more than 30 billion per day

Pre-filtered data: about 20 million per day

Alarm of statistical modeling: about 3000 per day

Events after correlation: 100 events per day

- Real attack will always be included in the top 100 events!

Summary

Machine learning is not suitable for detecting advanced threats in big data.

- The intrinsic property of intrusion detection.

Statistical modeling is more applicable for enterprise security data analysis.

- Detect typical behaviors in different attack phases.

Statistical modeling combined with data preprocessing and post correlation.

- The key point to improve precision-recall rate.

Using Statistical Modeling in Your Environment

- Next week you should:
  - Identify the 10 most dangerous attack techniques to your business (From the past network intrusion case, and the Red Team contests) .
  - Attack techniques can be described by ATT&CK knowledge base: https://attack.mitre.org/ .
- In the first three months you should:
  - Collect the end-to-end data relating to the attacks.
  - Build normal behavior baseline of each assets, and filter out at least 99% of the raw data.
- Within six months you should:
  - Build the statistical models to highlight the intrusion trace.
  - Correlate raw alarms and generate high risk events.
  - Calculate the precision-recall rate of the detection results, and adjust the model parameters based on it.

RSAConference2019

Thank you for attending this forum!

Email: devin.zt@alibaba-inc.com