# One App to Rule Them All: Applying Machine Learning to Find Them

SEC1471B

**Ali Çetin**

Senior Security Engineer | Yapı Kredi Bank

**Semih Gelişli**

Head of Cyber Security | Yapı Kredi Bank


splunk> .conf22

# Forward-Looking Statements

This presentation may contain forward-looking statements regarding future events, plans or the expected financial performance of our company, including our expectations regarding our products, technology, strategy, customers, markets, acquisitions and investments. These statements reflect management's current expectations, estimates and assumptions based on the information currently available to us. These forward-looking statements are not guarantees of future performance and involve significant risks, uncertainties and other factors that may cause our actual results, performance or achievements to be materially different from results, performance or achievements expressed or implied by the forward-looking statements contained in this presentation.

For additional information about factors that could cause actual results to differ materially from those described in the forward-looking statements made in this presentation, please refer to our periodic reports and other filings with the SEC, including the risk factors identified in our most recent quarterly reports on Form 10-Q and annual reports on Form 10-K, copies of which may be obtained by visiting the Splunk Investor Relations website at www.investors.splunk.com or the SEC's website at www.sec.gov. The forward-looking statements made in this presentation are made as of the time and date of this presentation. If reviewed after the initial presentation, even if made available by us, on our website or otherwise, it may not contain current or accurate information. We disclaim any obligation to update or revise any forward-looking statement based on new information, future events or otherwise, except as required by applicable law.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. We undertake no obligation either to develop the features or functionalities described, in beta or in preview (used interchangeably), or to include any such feature or functionality in a future release.



splunk> .conf22


**Ali Çetin**

Senior Security Engineer | Yapı Kredi Bank

**David Hourani**

PS Architect | Splunk

**Georgios Glymidakis**

PS Architect | Splunk

**Semih Gelişli**

Head of Cyber Security | Yapı Kredi Bank


splunk> .conf22

# About Yapı Kredi Bank

We are the 3rd largest private bank in Turkey with total assets worth $ 3.04 billion as of the end of 2021

**805**
Branches

**4,590**
ATM's

**~15.5k**
Employees

**12.9m**
Credit Cards

**9.1m**
Digital Customers

**84%**
Digital Banking Active Customer Penetration

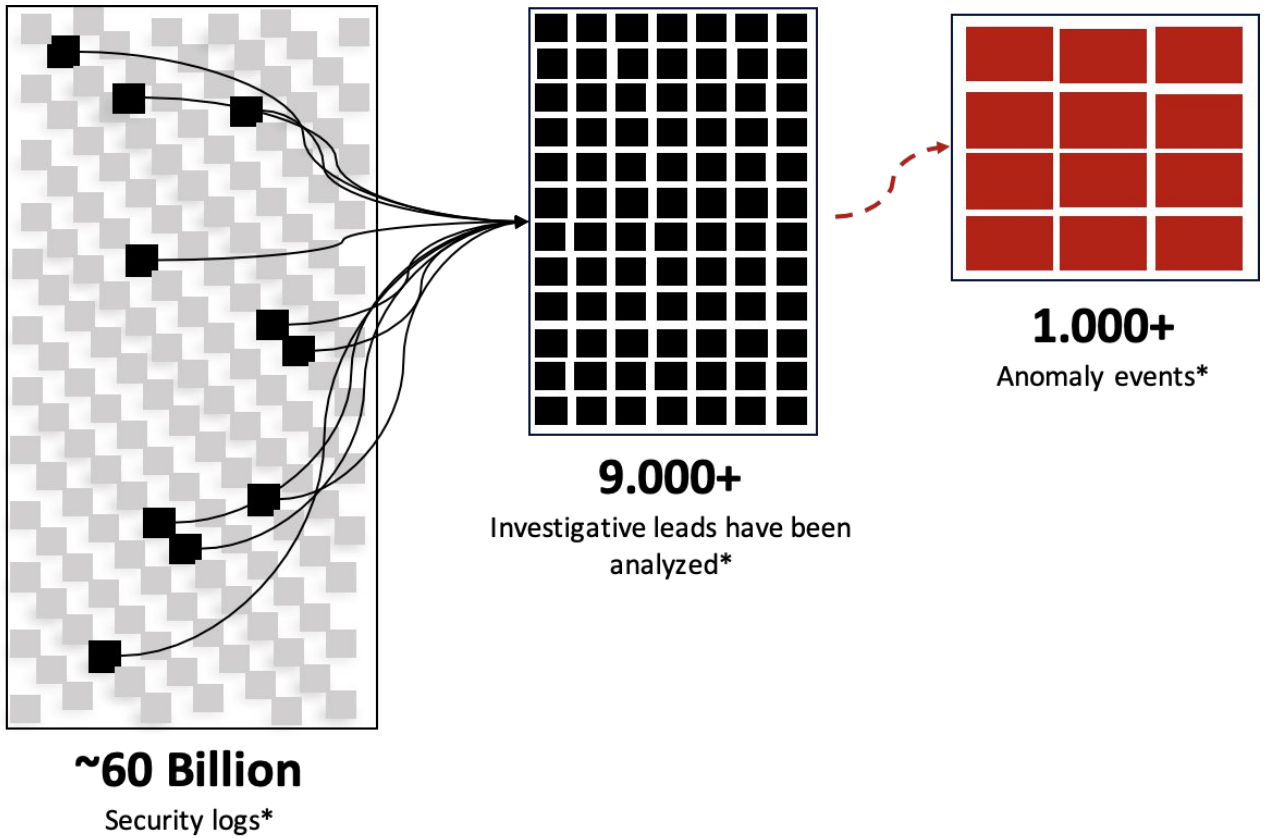
# Yapı Kredi Splunk Infrastructure

Our philosophy on Splunk and Artificial Intelligence is to place technology enabling human to make tasks more effective and efficient with output-driven methodology

**100+**
Data Sources

**50+**
Data Collector

**~%45**
Successfully Filtering out




**%95+**
SLA Success

**100+**
Monthly Alert Per Analyst

**8 min**
Mean time to Detect

*2021 Yearly Statistics

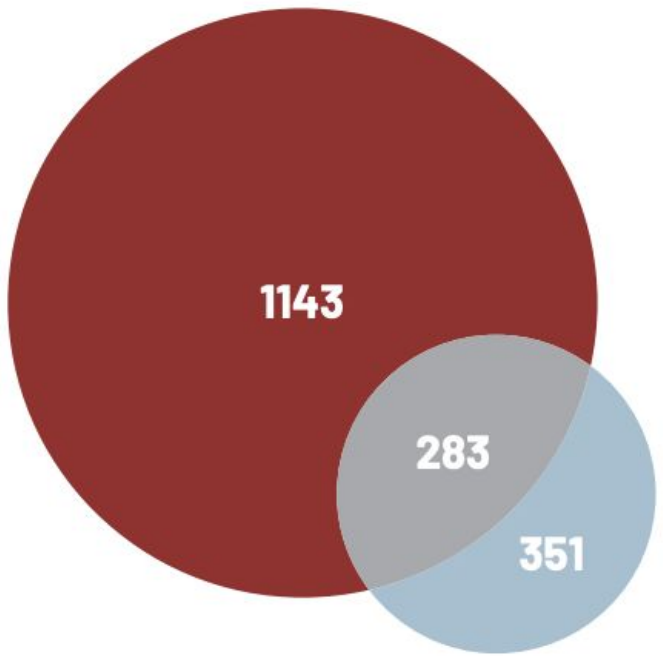
splunk> .conf22

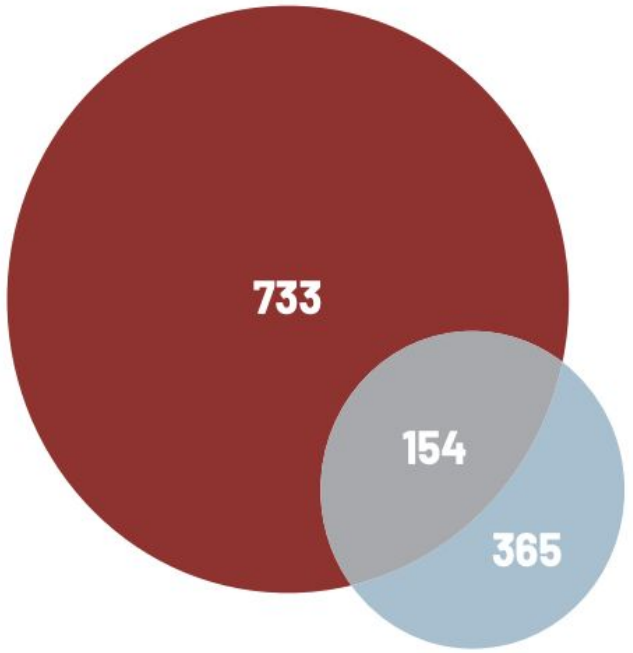
# Agenda

1. What's lurking in your environment?
   - What are we looking for and how can we spot abnormal behaviours within our environment?
2. Non-ML gurus are digging ML
   - The methodology of our use-cases and how did we kick off?
3. Wrestling with Likelihood approaches and ML
   - A bit statistics and ML for Security related data
   - Process Anomaly Detection
   - Service User Account Behaviours
   - Bonus
4. Wrap Up

# What's lurking in your environment?

Mandiant M-Trend 2022 Report shows us day by day new APT groups and Malwares have been increasing significantly




- Newly Tracked Threat Groups in 2021
- Newly Tracked and Observed Threat Groups in 2021
- Observed Threat Groups in 2021

- Newly Tracked Malware Families in 2021
- Newly Tracked and Observed Malware Families in 2021
- Observed Malware Families in 2021

splunk> .conf22


“All we have to decide is what to do with the **data** that is given us.”

**Gandalf**

# My Precious: Right Data & Quality

## Unlock the hidden value of your data

"Data is food for AI, and modern AI systems need not only calories, but also high quality nutrition."
Andrew NG

splunk> .conf22

# Threat Hunting

## Three Steps of Methodology

- The Trigger
- Investigation & Enrichment
- Resolution

# Non-ML Gurus are digging Statistics & ML

Statistics

Fit

Clustering

Features

Apply

Prediction

Data

Classification

splunk> .conf22

# Statistics & Machine Learning Steps

Is this event normal?

**Collect & Aggregate**

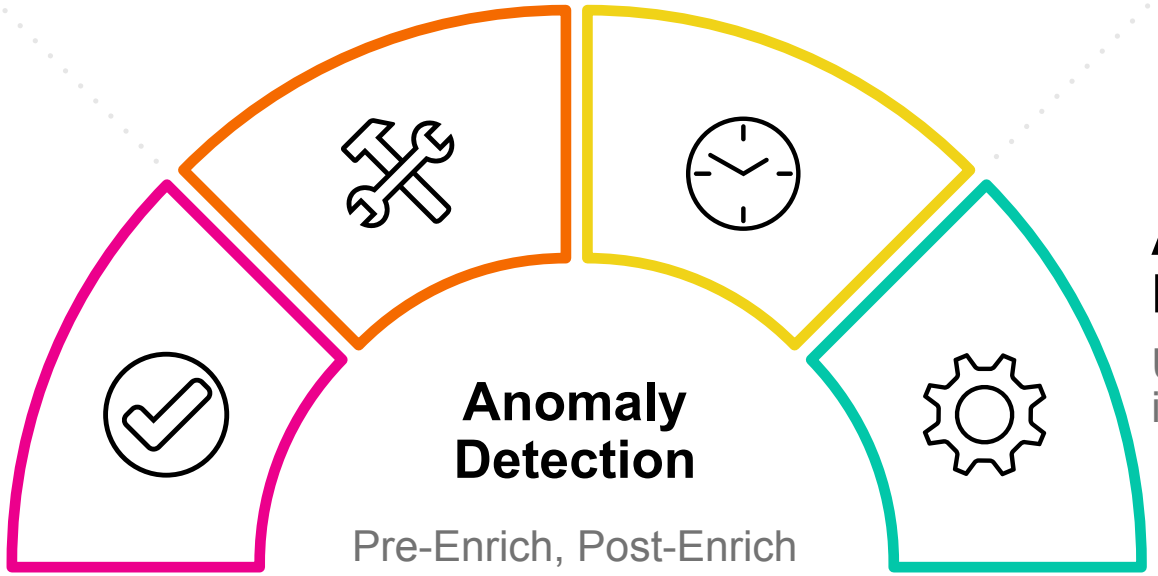
Baseline search

**Feature Selection**

Create numeric fields for statistics and ML

**Statistics | ML**

Decide approach | Choose your model



**Anomaly Detection**

Pre-Enrich, Post-Enrich

**Alert & Dashboard**

Up to your imagination

# Use Case: Process Anomaly Detection

splunk> .conf22

# Analogy

Q: How can we define the model to detect anomalous process creations and first time seen process tree per host & globally?

## Likelihood Approach Methodology

- aggregate and count up number of times that process trees seen
- create features | eventstats & | eval; total count global and local,local and global ratio percent, threshold
- likelihood approach
- run every night e.g. over 30 days
- collect everything into summary index
- create post enrichment steps; clustering, process eps, path eps, Echotrail etc.

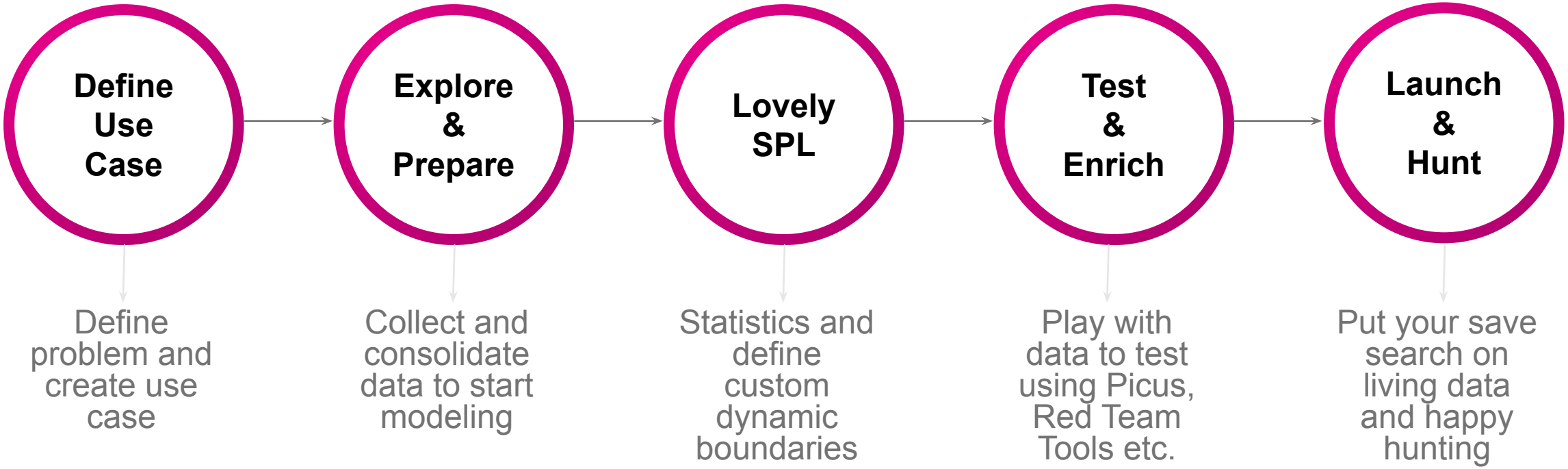
## Results

- outliers of same hash running with multiple process tree
- outliers of same process running with multiple hashes
- first time appearance of process tree

## Challenges

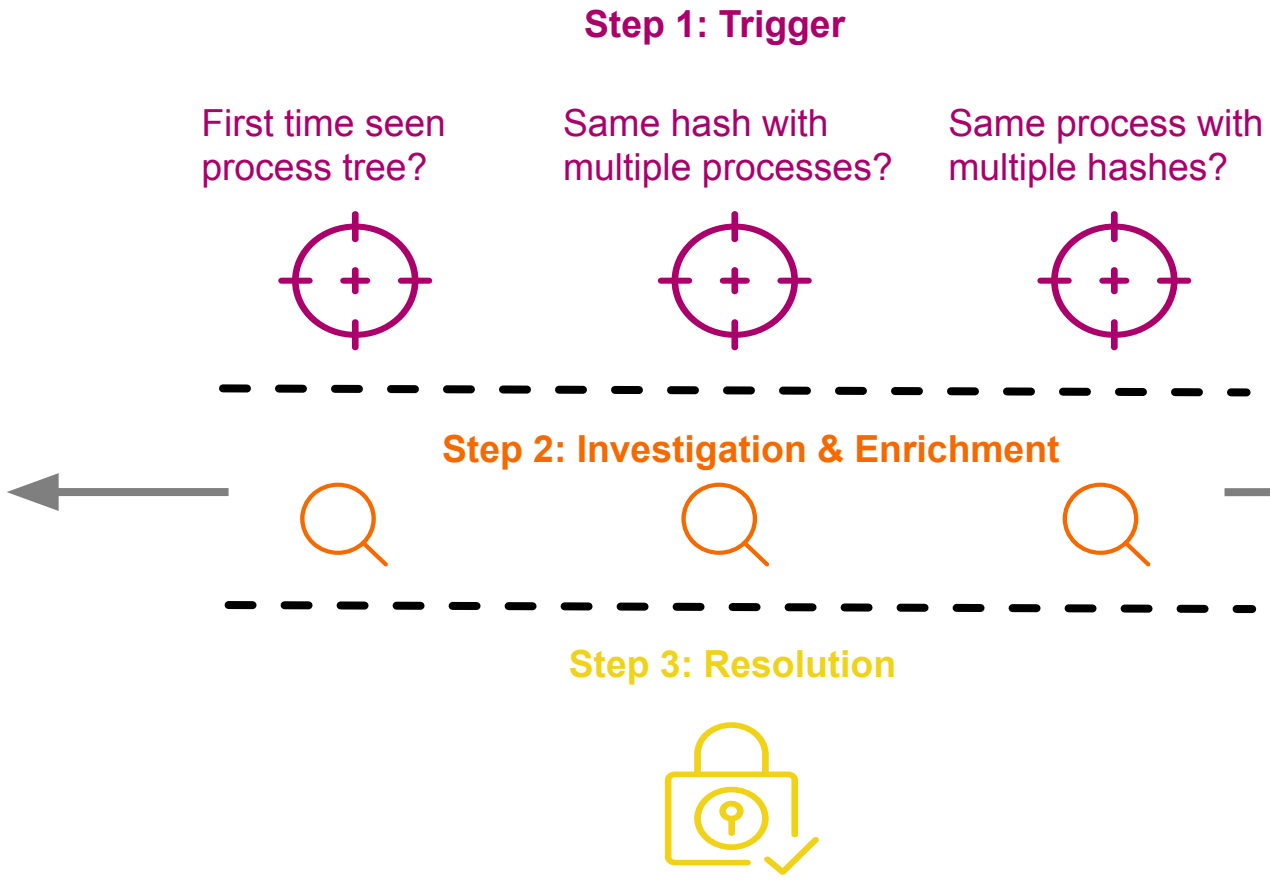
- processes with /$TEMP$/ and /$USER$/ directories are making noise
- legit processes can have multiple different hashes and can be running under multiple paths and may still come as noise

# Likelihood Approach Steps



Anomalous Process Detection

# Detection Diagram




- Calculating path appearance EPS & using Echotrail data
- Clustering processes for reducing noise using e.g. paths, cmdline, parents, TFIDF etc.

- Involving with TI
- Using external dataset e.g. Virustotal
- Clustering processes for reducing noise using e.g. paths, cmdline

**Findings turn into incident OR whitelist!**

Anomalous Process Detection

# Lovely SPL & Likelihood Details

```
| stats sum(count) as total_hour max(_time) as LastTime by dest
  full_process _time,process_hash
| stats sum(total_hour) as total max(_time) as LastTime by full_process
  process_hash dest
| eventstats sum(total) as total_by_dest by dest
| eventstats sum(total) as total_global
| eventstats sum(total) as total_global_process by full_process
| eventstats sum(total) as total_global_hash by process_hash
| eval ratio_perc_local=(total/total_by_dest)*100
| eval ratio_perc_global_process=(total_global_process/total_global)*100
| eval threshold=ratio_perc_global_process/4
| eval ratio_perc_global_hash=(total_global_hash/total_global)*100
| eval recent=relative_time(now(),"-1d@d")
| sort +ratio_perc_global_hash
| where (ratio_perc_global_hash < threshold  AND LastTime > recent) OR
  (ratio_perc_global_process < ratio_perc_global_hash  AND LastTime >
  recent)
| eval signature=if((ratio_perc_global_hash < threshold  AND LastTime >
  recent),"same process with multiple hashes","same hash with different
  processtree")
| eval _time=LastTime
| table full_process process_hash dest total signature _time
| collect index=whatever sourcetype=whatever addtime=true marker
  ="search_name=\"it's not magic, yet here we are\""
```

Creating process tree and aggregate data with | stats

Calculating base feature (total count locally and globally)

Unleash total ratio locally and globally by hash values

Adding likelihood approaches!

Clean table & write outputs

splunk> .conf22

Anomalous Process Detection

# Enrichment Details

```
| eval comb=path_len*dir_count
| eventstats stdev(path_len) as stdev_path_len stdev(parent_len) as
    stdev_parent_len stdev(ut_shannon) as stdev_shannon avg(path_len) as
    avg_path_len avg(parent_len) as avg_parent_len stdev(comb) as stdcomb a
    (comb) as avgcomb by process_name
```

Creating features for cluster

```
|  eval upperBound=avgcomb+stdcomb,lowerBound=avgcomb-stdcomb,upperBoundv2
    =avgcomb+stdcomb*2,lowerBoundv2=avgcomb-stdcomb*2
|  eval cluster=case(comb <= upperBound AND comb >= lowerBound,"1",comb >
    upperBound AND comb < upperBoundv2,"2", comb < lowerBound AND comb >
    lowerBoundv2,"3",(comb > upperBoundv2 OR comb < lowerBoundv2),"-1")
```

Creating custom cluster

```
|  cluster t=0.8 field=process_path_value showcount=true labelonly=t
| lookup process_tracker full_process OUTPUTNEW dest as previous_dest
|  stats values(detected_dest) as detected_dest  dc(previous_dest) as
    count_previous_dest dc(detected_dest) as count_detected_dest values
    (full_process) as full_process by process_name dir_count cluster
    cluster_label cluster_count
```

Double checking with | cluster

```
|  eval first_seen=if(count_previous_dest="0" AND count_detected_dest="1"
    ,"First Seen Globaly","Seen Previously")
|  search first_seen="First Seen Globaly" AND cluster="*"
```

Looking for the first time seen globally

Anomalous Process Detection

# Drilldown Investigation

Enriching output with external data sources

```
4 |  stats count by process_name
5 |  rename process_name as process_info
6 |  lookup echotrail process_info OUTPUT process_eps
7 |  search NOT process_eps="*"
8 | head 250
9 | echotrail echotrail_hash="hash" echotrail_eps="process_eps" echotrail_pat
    ="paths"
```

| process_info | process_path_value | check_path | paths | process_eps |
|---|---|---|---|---|
| taskhostw.exe | C:\\Windows\\System32 | 1 | ["C:\\Windows\\System32","99.89"]<br>["C:\\WINDOWS\\System32","0.11"] | 86.41 |
| explorer.exe | C:\\Windows | 1 | ["C:\\Windows","67.50"]<br>["C:\\Windows\\SysWOW64","32.06"] | 94.72 |

**Server First Seen**

| process_id | parent_process_name | process_name |
|---|---|---|
| 7748 | 456406.exe | net.exe |
| 9060 | 456406.exe | net.exe |
| 11716 | 456406.exe | powershell.exe |
| 2128 | 456406.exe | powershell.exe |

**Detailed Process Tree**

tree

```
picus.agent.service.exe (11676)
|--- 456406.exe (7104)                    2022-04-25 03:20:04 "C:\ProgramData\Picus Security\Picus Agent\Scenarios\57300-456406\4
`` |--- powershell.exe (11716)             2022-04-25 03:20:10 "powershell.exe" -command "if(Test-Path -Path 'HKLM:SOFTWARE\Classe
[0].Replace('WinRAR.exe', 'Rar.exe') 'a -r %%public%%\Downloads\328054.rar %%public%%\Downloads\test\mimi.log'}else{Write-Host 'YOK'}"
`` |--- net.exe (9060)                    2022-04-25 03:20:09 "net" localgroup administrators guest /delete
`` `` |--- net1.exe (5728)                2022-04-25 03:20:10 C:\Windows\system32\net1 localgroup administrators guest /delete
`` |--- net.exe (7748)                     2022-04-25 03:20:09 "net" localgroup administrators guest /add
`` `` |--- net1.exe (11588)               2022-04-25 03:20:09 C:\Windows\system32\net1 localgroup administrators guest /add
`` |--- taskkill.exe (5712)                2022-04-25 03:20:08 "taskkill.exe" /f /im hh.exe
`` |--- cmd.exe (1380)                     2022-04-25 03:20:08 "cmd.exe" /c del C:\Windows\TEMP\CradleTest.txt
`` |--- hh.exe (10368)                     2022-04-25 03:20:08 "hh.exe" http://10.60.68.11:80/529363\CradleTest.txt
`` |--- cmd.exe (11980)                    2022-04-25 03:20:07 "cmd.exe" /c del "C:\wmi.dll"
`` |--- cmd.exe (2276)                     2022-04-25 03:20:07 "cmd.exe" /c whoami &gt; C:\wmi.dll 2&gt;&amp;1
`` `` |--- whoami.exe (10988)             2022-04-25 03:20:07 whoami
`` |--- cmd.exe (11696)                    2022-04-25 03:20:07 "cmd.exe" /c echo \"1q2w3e4r\" &gt;&gt; pass.txt
`` |--- cmd.exe (5164)                     2022-04-25 03:20:06 "cmd.exe" /c echo \"Password1\" &gt;&gt; pass.txt
`` |--- cmd.exe (9348)                     2022-04-25 03:20:06 "cmd.exe" /c echo \"1q2w3e4r\" &gt;&gt; pass.txt
`` |--- cmd.exe (9300)                     2022-04-25 03:20:06 "cmd.exe" /c echo \"Password1\" &gt;&gt; pass.txt
`` |--- cmd.exe (9368)                     2022-04-25 03:20:06 "cmd.exe" /c del dirlog.txt
`` |--- cmd.exe (6840)                     2022-04-25 03:20:05 "cmd.exe" /c dir &gt; dirlog.txt
```

splunk> .conf22

# Use Case: Service Account Anomaly Detection

splunk> .conf22

# Analogy

Q: How can we define the model to profile each service account login behaviours and detect outliers using MLTK?

## MLTK Approach Methodology

- aggregate service accounts by LogonType, dest and src
- create features; dc(dest), values(dest), dc(src), values(src), LogonType
- | fit DensityFunction model on each feature by service account user
- | apply your_model each day to test last day's data
- collect into summary index
- generate risk score for priority
- run every night and happy dashboarding

## Result

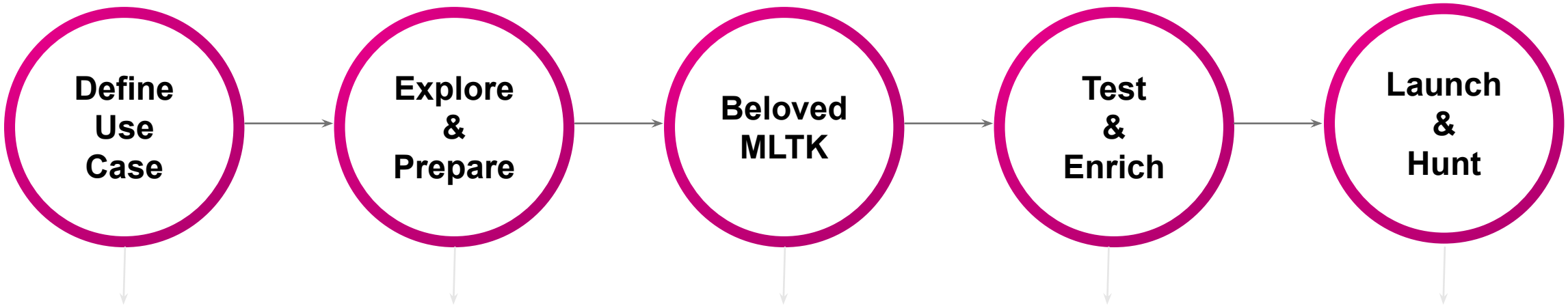
- spots three type of outliers: LogonType anomaly, Destination anomaly and Source Anomaly
- gives you another perspective to look for e.g. Lateral Movement, Credential Theft

## Challenges

- newly added users or servers login sessions can make noise: Enrich data with Active Directory data

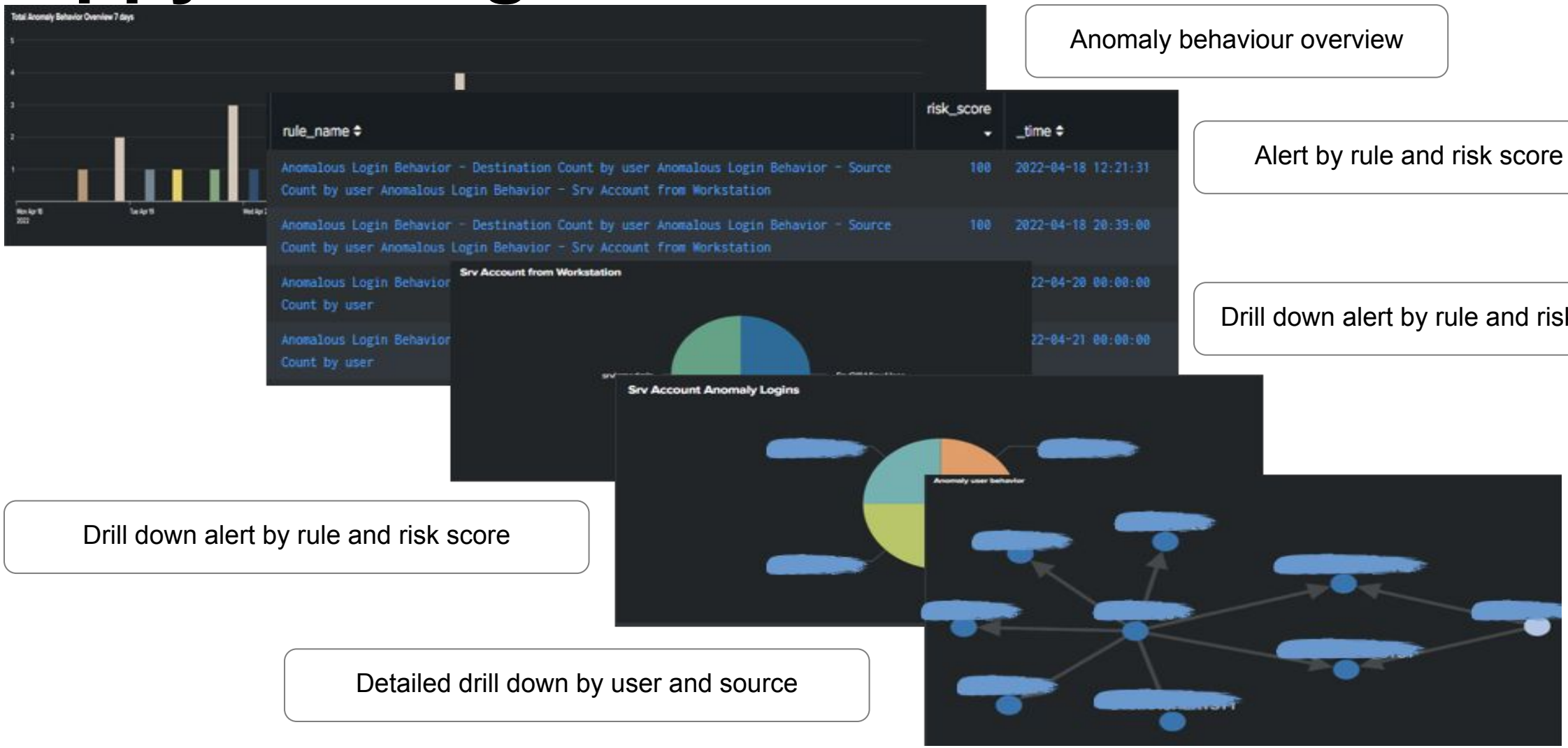
Service Account Anomaly Detection

# MLTK Steps



Service Account Anomaly Detection

# Happy Hunting




Anomaly behaviour overview

Alert by rule and risk score

Drill down alert by rule and risk score

Drill down alert by rule and risk score

Detailed drill down by user and source

splunk> .conf22

# Bonus Use Case: CMDLine Obfuscation Detection

splunk> .conf22

# Analogy

Q: How can we define the model to classify each cmdline behaviours as 1|0 and detect obfuscation using MLTK?

## MLTK Approach Methodology

Aggregation & Feature Selection

- create labeled dataset
- aggregate cmdline with users, host
- create features: proportion of special char, proportion of escape sequences, and TFIDF
- reduce dimension with PCA
- | fit Classification model on each feature by cmdline
- choose best fitted model
- | apply your_model
- collect into summary index and add sigma rules

## Result

- spots obfuscated CMDLine
- gives another perspective to look for e.g. Execution, Defense Eviason

## Challenges

- brand-new obfuscation methods may not be detected
- management of whitelist & blacklist to update model periodically

CMDLine Obfuscation Detection

# Lovely SPL & ML Details

```
| stats count by cmdline
| eval cmdline_len = max(1,len(cmdline))
| rex field=cmdline max_match=0 "(?<spaces>\s)"
| rex field=cmdline max_match=0 "(?<special>[^\w|\s\\\:\.\"]+)"
| rex field=cmdline max_match=0 "(?<upper>[A-Z-\s]+)"
| eval space_ratio=if(isnull(spaces),0.0,mvcount(spaces) / cmdline_len)
| eval special_ratio=if(isnull(special),0.0,mvcount(special) / cmdline_len)
| eval upper_ratio=if(isnull(upper),0.0,mvcount(upper))
| `ut_shannon(cmdline)`
| eval num_obfuscation =
    (mvcount(split(cmdline,"`"))-1)
  + (mvcount(split(cmdline, "^"))-1)
  + (mvcount(split(cmdline, "'"))-1)
  + (mvcount(split(cmdline, "%"))-1)
  + (mvcount(split(cmdline, "*"))-1)
  + (mvcount(split(cmdline, "+"))-1)
  + (mvcount(split(cmdline, "="))-1)
  + (mvcount(split(cmdline, ">"))-1)
```

creating features for classification, more features=higher accuracy

looking for specific characters that mostly using for obfuscation

- cleaning data
- features
- TFIDF
- PCA
- Classification

| | | | |
|---|---|---|---|
| obfuscated | obfuscated | powershell | ""'\""dir c:\windows\system32\calc.exe\""\|&(\""iex\"")'\|&(\""IEX\""); Write-Host \""<>^^\|\&^\" |
| obfuscated | obfuscated | powershell | ""SEt O07sTn ([cHaR[<br>&}3'+'{'+')}2{}0{}2'+'{'+',)'+'<br>epy}2{+}'+'2'+'{t'+'}2{+'+'}2{<br>'+'DERIAPN}2{+'+'}2{U}2{+}2{'+'<br>;txt.a'+'l}2'+'{+}2{b led}2{( ' |
| obfuscated | obfuscated | powershell | ""Write-Host 'this i |
| obfuscated | obfuscated | powershell | ""Write-Host \""this is a test 1\" |
| obfuscated | obfuscated | powershell | ""dir \""c:\windows\system32\ca*c.exe\""" |

| class | predicted | orig_text |
|---|---|---|
| obfuscated | obfuscated | CmD , , ,/R "", (^set G^T=ch)& (^sEt i^J=^e)& (s^Et ^Ma=o %%^te^mp%%)&& , cAll, S^ET C23z=%%i^J%%%%G^T%%%%^Ma%%& |
| obfuscated | obfuscated | CmD , ,,,,,,,,/V:ON,,,,,,,,,/C"",,,,,,,,,(set ` =h)&&(,(,(set } =wr),),)&&(,(,(set ? =c),),)&&(,(,(set , =P%%""" |

splunk> .conf22

# Wrap Up

"Your time will come. You will face the same challenges, and you will defeat it." - Arwen

- lovely SPL and Statistics are still one of the best weapon in arsenal
- risk score can reduce FP
- right data with high quality model allows you to look at hidden
- features and enrichments are the key
- combine threat hunting steps with Math and signature based rules
- we don't know if we never try!
- research, engage and build!

"It's the Statistics & ML that's never started as takes longest to finish." - Samwise

- DGA App for Splunk
- MLTK showcases
- ESCU anomaly use cases
- Splunk Blogs: A Splunk Approach to Baselines, Statistics and Likelihoods on Big data
- SEC1374, SEC1495C, SEC1395A

# Thank You

splunk> .conf22