Machine Learning for User Behavior Anomaly Detection

EUGENE NEYOLOV, HEAD OF R&D
The only AI-driven SAP & Oracle Cybersecurity Provider

Vulnerabilities Reported

500+ SAP
318

43 AWARDS

INDUSTRIES

Business Service & Products 14.1%
Gas & Oil 10.6%
Security Systems 7.0%
Manufacturing 7.0%
Energy 6.3%
Telecommunications services 4.2%
Banks, Brokers & Finances 4.2%
Science & Education 4.2%
Retail 3.5%
Oil & Gas Operations 3.5%
Software & Programming 2.8%

10 000 AI SECURITY CHECKS COVERED
2 x AVERAGE DEAL SIZE GROWTH

200 DEPLOYMENTS WORLDWIDE

UNIQUE

159

50+ PARTNERS
35 COUNTRIES

US OFFICE

Palo Alto

EMEA OFFICE

Amsterdam

R&D OFFICE

Prague

MACHINE LEARNING LAB

Tel Aviv

120+ CONFERENCES

120 AS SPEAKERS

60+ EMPLOYEES

40 RESEARCH EXPERTS

REPORTS

70+
Eugene Neyolov
HEAD OF R&D

Security engineer and analyst leading applied research projects in security monitoring, threat detection and user behavior analytics.

Current Interests
• Building products for
• Cyber security with
• Data science and
• Hype
• Why
  o ERP Security
  o User Behavior Analytics
  o Machine Learning

• What
  o Static Anomalies
  o Temporal Anomalies

• How
  o Data Preparation
  o Security Analytics
  o Security Data Science
  o Machine Learning
  o Anomaly Detection
ERP
Security
ERP SECURITY

Blind Spot

• Endpoint security
• Network security
• Application security
• Intrusion detection
• Identity and access governance
• Business applications security

Infrastruture focused prevention/detection

Where a real ERP attack happens
ERP SECURITY

Sweet Target

HR Management
Financial Accounting
Sales and Distribution
Materials Management
Quality Management
Production Planning
Plant Maintenance
Supply Chains
...

Enterprises  Attackers
User Behavior
Analytics
Why?

• Legacy threat models
  o Users are the easiest attack vector

• Legacy incident monitoring
  o Infrastructure security focused analysis

• Legacy security alerts analysis
  o No business context enrichment
USER BEHAVIOR ANALYTICS

What?

• User security monitoring
• User-focused alert prioritization
• Advanced context enrichment
• User behavior vs. fraud analysis
  o UBA is about facts in the technical context
    - Developer must work with development server A but have accessed server B owned by the finance department
  o Fraud is about intentions in a business context
    - Salesman signs a contract with company A and not company B, because A is managed by a friend
USER BEHAVIOR ANALYTICS

How?

• Create a user-centered threat model
• Identify user-related data sources
• Build a user behavior baseline
• ???
• PROFIT!!!
Machine Learning
MACHINE LEARNING

Why?

- Escape postmortem rules and signatures
- Self-adjusted dynamic behavior patterns
- Find hidden patterns in user behavior
MACHINE LEARNING

What?

• **ML tasks**
  o Clustering
  o Regression
  o Classification
  o Anomaly detection
  o ...

• **Learning patterns from data**
  o Supervised learning with labeled data
  o Unsupervised learning without labeled data
  o Semi-supervised learning with tips from data or humans
  o Reinforcement learning with a performance feedback loop
  o ...

• ...
MACHINE LEARNING

What?

• ML model
  o Codebase
  o Features structure
  o Model parameters (learned)
  o Model hyperparameters (architecture)

• ML features
  o Categorical (classes)
  o Statistical (counts)
  o Empirical (facts)
  o Continuous
  o Binary
  o ...
MACHINE LEARNING

How?

Data Preparation
- Collect event data
- Normalize events
- Enrich events

Security Analytics
- Categorize events
- Build threat models
- Map events to threats

Security Data Science
- Map threats to algorithms
- Select and encode features
- Define quality requirements

Machine Learning
- Build a model
- Train a model
- Optimize model parameters

Incident Analysis
- User behavior analysis
- Peer group analysis
- Threat classification

Anomaly Detection
- Feed a real data
- Detect anomalies
- Prioritize anomalies
Data Preparation
DATA SOURCES

- APIs
- Log files
- Databases
- Log archives
- Log management tools
- Security monitoring tools
- ...

DATA FORMATS

• Syslog
• Custom mess
• Random key-value
• Proprietary key-value (CEF, LEEF, ...)
• Other terrible options (JSON, CSV, ...)
DATA NORMALIZATION

• Understand that mess
  o When, Who, did What, Where from, Where to, on What

• Bring all formats to the same convention
  o Implement a built-in convertor for each format as a part of the solution (inside)
  o Create a separate convertor tool and treat it as the data source for the model (outside)
  o Build event storage that allows event fields mapping, like Splunk or ELK (infrastructure)

• Find duplicates and missing fields
  o One action generates several entries
  o System doesn’t identify itself in its own logs
  o User’s name is recorded, but not its IP (or vice versa)
### DATA NORMALIZATION: BEFORE

#### SAP Security Audit Log ABAP

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>User</th>
<th>Activity</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>2AU520180313113209000030400001D1nsalab SAP*</td>
<td>SAPMSSY1 0001F&amp;0</td>
<td>nsalab 2AU520180313113209000030400001D1nsalab SAP*</td>
<td></td>
</tr>
<tr>
<td>2AU220180313114609002315800004D4MacBook-SAP*</td>
<td>SESSION_MANAGER SAPMSYST</td>
<td>MacBook-Pro-Nursulta2AU120180313114703002315800004D4MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0001A&amp;1</td>
<td>MacBook-Pro-Nursulta2AU120180313114703002315800004D4MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td>2AUW20180313114703002315800004D4MacBook-SAP*</td>
<td>SESSION_MANAGER RSRZLLG0</td>
<td>MacBook-Pro-Nursulta2AUW20180313114703002315800004D4MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0011RSRZLLG0&amp;</td>
<td>MacBook-Pro-Nursulta2AUW20180313114703002315800004D4MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td>0011RSRZLLG0&amp;</td>
<td>MacBook-Pro-Nursulta2AUW20180313114703002315800004D4MacBook-SAP*</td>
<td>MacBook-Pro-Nursulta2AUW20180313114703002315800004D4MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td>0011RSRZLLG0&amp;</td>
<td>MacBook-Pro-Nursulta2AU3201803131151520023162000008D8MacBook-SAP*</td>
<td>MacBook-Pro-Nursulta2AU3201803131151520023162000008D8MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td>0011SE16</td>
<td>SE16 SAPLSMTR_NAVIGATION</td>
<td>MacBook-Pro-Nursulta2DU92018031311550023162000008D8MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MacBook-Pro-Nursulta2DU92018031311550023162000008D8MacBook-SAP*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MacBook-Pro-Nursulta2DU92018031311550023162000008D8MacBook-SAP*</td>
<td></td>
</tr>
</tbody>
</table>
### DATA NORMALIZATION: AFTER

#### SAP Security Audit Log ABAP

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>User</th>
<th>Device</th>
<th>Action</th>
<th>Context 1</th>
<th>Context 2</th>
<th>Context 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/13/18 11:32</td>
<td>RFC/CPIC Logon Successful</td>
<td>SAP*</td>
<td>nsalab</td>
<td>AU5</td>
<td>F</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:32</td>
<td>Successful RFC Call</td>
<td>SAP*</td>
<td>nsalab</td>
<td>AUK</td>
<td>SLO6</td>
<td>SAPLSLO6</td>
<td>RSAU_READ_FILE</td>
</tr>
<tr>
<td>3/13/18 11:46</td>
<td>Logon Failed</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>AU2</td>
<td>A</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:47</td>
<td>Logon Successful</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>AU1</td>
<td>A</td>
<td>0</td>
<td>P</td>
</tr>
<tr>
<td>3/13/18 11:51</td>
<td>Transaction Started</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>AU3</td>
<td>SE16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:51</td>
<td>Read Table</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>DU9</td>
<td>USR02</td>
<td>2</td>
<td>passed</td>
</tr>
</tbody>
</table>
ERP SECURITY LOGGING

• Common business application logging
  o Event time
  o Event type
  o Server info
  o User info
  o ...

ERP SECURITY LOGGING

• SAP tracks 50+ fields across 30+ log formats
  o SAP system ID (*business entity*)
  o client number (*company sandbox inside a system*)
  o names of processes, transactions, programs or functions (*runtime data*)
  o affected user, file, document, table, program or system (*context data*)
  o amount of inbound and outbound traffic (*network data*)
  o severity, outcome and error messages (*status data*)
  o device forwarded the event (*infrastructure data*)
  o ...
ERP SECURITY LOGGING

SAP Security Audit Log ABAP

• Short list of important fields
  o Time
  o Event type, class
  o System type (log source)
  o System ID, server hostname and IP
  o User name, device hostname and IP
  o Executed program name (transaction, report, remote call)
THREAT MODEL

Use Cases

• **10+ Categories (why)**
  o Data Exfiltration, Account Compromise, Regular Access Abuse, Privileged Access Abuse, ...

• **30+ Classes (what)**
  o Data Transfer, Account Sharing, Password Attack, Privilege Escalation, Lateral Movement, ...

• **100+ Scenarios (how)**
  o Login from multiple hosts, User upgrades its own privileges, Cover tracks via user deletion, ...
Security
Data Science
ANOMALY TYPES

• **Static anomalies**
  - Unusual action (new or rare event)
  - Unusual context (server, device, ...)
  - ...

• **Temporal anomalies**
  - Unusual time
  - Unexpected event
  - Huge events volume
  - ...

• Many anomalies are not malicious
• Anomalies are statistical deviations
• Big infrastructures always have anomalies
<table>
<thead>
<tr>
<th>Threat Model</th>
<th>Category</th>
<th>Temporal Anomalies</th>
<th>Static Anomalies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class</td>
<td>Unusual action</td>
<td>Unusual time</td>
</tr>
<tr>
<td>Regular Access Abuse</td>
<td>Unauthorized Access</td>
<td>high</td>
<td>medium</td>
</tr>
<tr>
<td></td>
<td>Account Sharing</td>
<td>low</td>
<td>medium</td>
</tr>
<tr>
<td>Account Compromise</td>
<td>Password Attack</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Privilege Escalation</td>
<td>high</td>
<td>medium</td>
</tr>
<tr>
<td></td>
<td>Access Enumeration</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Data Exfiltration</td>
<td>Data Transfer</td>
<td>low</td>
<td>medium</td>
</tr>
</tbody>
</table>
Static
Anomalies
STATIC ANOMALY DETECTION

Plan

- Context building
- Context matching
- Anomaly analysis
CONTEXT BUILDING

- Whitelist known values for all users
- Define anomaly scores for all fields
• **Problem**
  o Log poisoning attacks
  o Anomalies in user context

• **Solution**
  o Importance amplification
  o Mean of squared values

<table>
<thead>
<tr>
<th>IP</th>
<th>Mean</th>
<th>IP</th>
<th>Mean</th>
<th>Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.16.100.11</td>
<td>320</td>
<td>172.16.100.11</td>
<td>320</td>
<td>102400</td>
</tr>
<tr>
<td>172.16.100.118</td>
<td>308</td>
<td>172.16.100.118</td>
<td>308</td>
<td>94864</td>
</tr>
<tr>
<td>172.16.100.137</td>
<td>30</td>
<td>172.16.100.137</td>
<td>30</td>
<td>900</td>
</tr>
<tr>
<td><strong>Threshold</strong></td>
<td><strong>219</strong></td>
<td><strong>Threshold</strong></td>
<td><strong>28</strong></td>
<td><strong>8,258</strong></td>
</tr>
</tbody>
</table>
• Compare new events with the user context field by field
• Assign individual anomaly scores for unknown fields
ANOMALY ANALYSIS

- Get a total event anomaly score from all its fields
- Get a total user anomaly score from all its events
Temporal Anomalies
Establish a normal behavior baseline
Train to predict normal user actions
Analyze incorrectly predicted actions
FEATURE ENGINEERING

- Feature selection
- Feature encoding
## FEATURE SELECTION

### Data

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>User</th>
<th>Device</th>
<th>Action</th>
<th>Context 1</th>
<th>Context 2</th>
<th>Context 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/13/18 11:32</td>
<td>RFC/CPIC Logon Successful</td>
<td>SAP*</td>
<td>nsalab</td>
<td>AU5</td>
<td>F</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:32</td>
<td>Successful RFC Call</td>
<td>SAP*</td>
<td>nsalab</td>
<td>AUK</td>
<td>SLO6</td>
<td>SAPLSLO6</td>
<td>RSAU_READ_FILE</td>
</tr>
<tr>
<td>3/13/18 11:46</td>
<td>Logon Failed</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>AU2</td>
<td>A</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:47</td>
<td>Logon Successful</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>AU1</td>
<td>A</td>
<td>0</td>
<td>P</td>
</tr>
<tr>
<td>3/13/18 11:51</td>
<td>Transaction Started</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>AU3</td>
<td>SE16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:51</td>
<td>Read Table</td>
<td>SAP*</td>
<td>MacBook-Pro-Nursulta</td>
<td>DU9</td>
<td>USR02</td>
<td>2</td>
<td>passed</td>
</tr>
</tbody>
</table>
### Feature Encoding

#### Vector

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>User</th>
<th>Device</th>
<th>Action</th>
<th>Context 1</th>
<th>Context 2</th>
<th>Context 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/13/18 11:32</td>
<td>RFC/CPIC Logon Successful</td>
<td>SAP*</td>
<td>nsalab</td>
<td>AU5</td>
<td>F</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3/13/18 11:32</td>
<td>Successful RFC Call</td>
<td>SAP*</td>
<td>nsalab</td>
<td>AUK</td>
<td>SLO6</td>
<td>SAPLSLO6</td>
<td>RSAU_READ_FILE</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
0.19248842592592594 & 0.7110773240660063 & 0.8366013071895425
\end{bmatrix}
\]
FEATURE ENCODING

Knowledge Base

- On-the-fly KB
- Security-focused KB
- Application-focused KB
  - Static (1/100000 scale)
  - Mapping (1/100 scale)
Machine Learning
• Find the right algorithm for a task
• Implement a model and its environment
• Optimize the model for the best accuracy
• **Recurrent neural networks**
  - Simple RNN
    - Forgets longer dependencies
  - Long Short-Term Memory
    - Proven track record
  - Gated Recurrent Unit
    - LSTM simplified
  - Neural Turing Machine
    - RNN on steroids
  - ...

![Diagram of Recurrent Neural Network](image)
MODEL DESIGN

Architecture

- Input
  - Features
- LSTM
  - Time
  - Action
  - Program
- Output
  - Predict
MODEL PARAMETERS

- **Architecture**
  - Layers number, Neurons number, Activation function, Loss function, Optimizer, ...

- **Data**
  - Features, Knowledge base, Sequence length, Normalization, ...

- **Training**
  - Epochs, Batch size, Threshold, Distance, Smoothing, ...

---

Events Storage → Features Encoding → RNN Engine → Anomaly Detection → Anomalies Storage

Model Training → Weights Storage
SEQUENCE LENGTH

- A B C D E F G H A C K E D
- A B C D E F G H A C K E D
- A B C D E F G H A C K E D
- A B C D E F G H A C K E D
- A B C D E F G H A C K E D
KNOWLEDGE BASE SORTING

- **Alphabet**
- **Criticality**
- **Frequency**

Sorted by Alphabet

Sorted by Frequency
ADAPTIVE THRESHOLD

- Error score
  - Distance-based
    - Predicted value (blue)
    - Actual value (green)

- Threshold
  - Max training error score

- Sensitivity
  - As is
  - Coefficient
ANOMALY DETECTION

- Predict a potential user activity
- Report incorrectly predicted events above threshold
ANOMALY DETECTION

Prediction

Time Prediction

Action Prediction
ANOMALY DETECTION

Metrics

- **Accuracy 95%**
  - True Positives 71%
  - True Negatives 97%

- **Errors 5%**
  - False Positives 3%
  - False Negatives 29%
CONCLUSIONS

• Security analytics is more important than machine learning
• ML-driven solutions must help analysts and not replace them
• Adjust accuracy and tolerance to false positives for your situation
• Build an ecosystem of ML models and advanced analytics on top of it
Eugene Neyolov
Head of R&D
neyolov@erpscan.com

Read our blog
erpscan.com/category/press-center/blog/

Join our webinars
erpscan.com/category/press-center/events/

Subscribe to our newsletters
eepurl.com/bef7h1

USA:
228 Hamilton Avenue, Fl. 3, Palo Alto, CA. 94301
Phone 650.798.5255

EU:
Luna ArenA 238 Herikerbergweg, 1101 CM Amsterdam
Phone +31 20 8932892

erpscan.com
inbox@erpscan.com

Štětkova 1638/18, Prague 4 - Nusle, 140 00, Czech Republic