Machine Learning

for User Behavior Anomaly Detection

EUGENE NEYOLOV, HEAD OF R&D

ERPScan

The only AI-driven SAP & Oracle Cybersecurity Provider

- GARTNER HYPE CYCLE FOR APPLICATION SECURITY
- GARTNER MQ FOR APPLICATION SECURITY
- GARTNER MS FOR SOD TOOLS

VULNERABILITIES REPORTED

500+ / **318 SAP**

43 AWARDS

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

%

INDUSTRIES 40+

US OFFICE — PALO ALTO

EMEA OFFICE — AMSTERDAM

R&D OFFICE — PRAGUE

MACHINE LEARNING LAB — TEL AVIV ...

120+ CONFERENCES

120 AS SPEAKERS

REPORTS 70+

AI **10 000** SECURITY CHECKS COVERED

2x AVERAGE DEAL SIZE GROWTH

60+ EMPLOYEES

40 RESEARCH EXPERTS

200 DEPLOYMENTS WORLDWIDE

UNIQUE 159

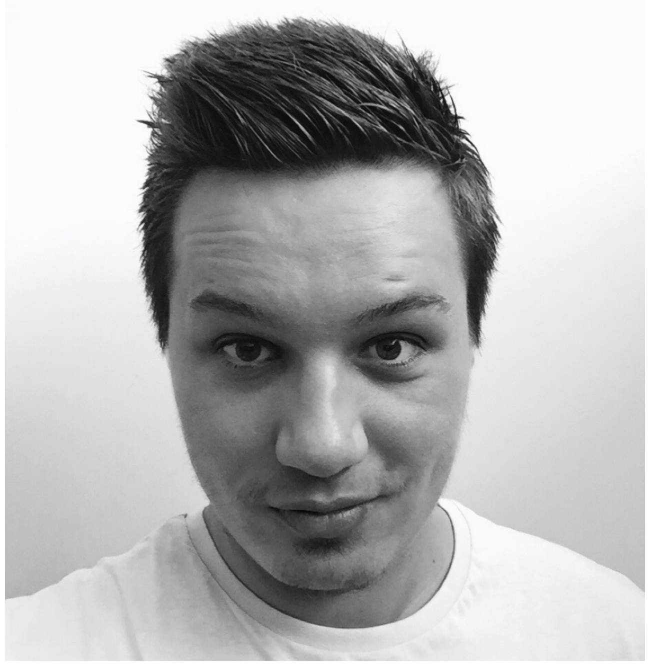
50+ PARTNERS

35 COUNTRIES

READ US IN

WIRED
The Register
DARKReading
International Business Times
MOTHERBOARD
BUSINESS INSIDER
theguardian
Forbes
TechTarget

AUTHOR



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**

OUTLINE

- **Why**
  - ERP Security
  - User Behavior Analytics
  - Machine Learning
- **What**
  - Static Anomalies
  - Temporal Anomalies
- **How**
  - Data Preparation
  - Security Analytics
  - Security Data Science
  - Machine Learning
  - Anomaly Detection

ERP
Security

ERP SECURITY

Blind Spot

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

Infrastructure focused prevention/detection

Where a real ERP attack happens

ERP SECURITY

Sweet Target

Enterprises →

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

→ Attackers

User Behavior Analytics

USER BEHAVIOR ANALYTICS

Why?

- **Legacy threat models**
  - Users are the easiest attack vector
- **Legacy incident monitoring**
  - Infrastructure security focused analysis
- **Legacy security alerts analysis**
  - No business context enrichment

USER BEHAVIOR ANALYTICS

What?

- **User security monitoring**
- **User-focused alert prioritization**
- **Advanced context enrichment**
- **User behavior vs. fraud analysis**
    - 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
    - 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**
  - Clustering
  - Regression
  - Classification
  - Anomaly detection
  - ...

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

MACHINE LEARNING

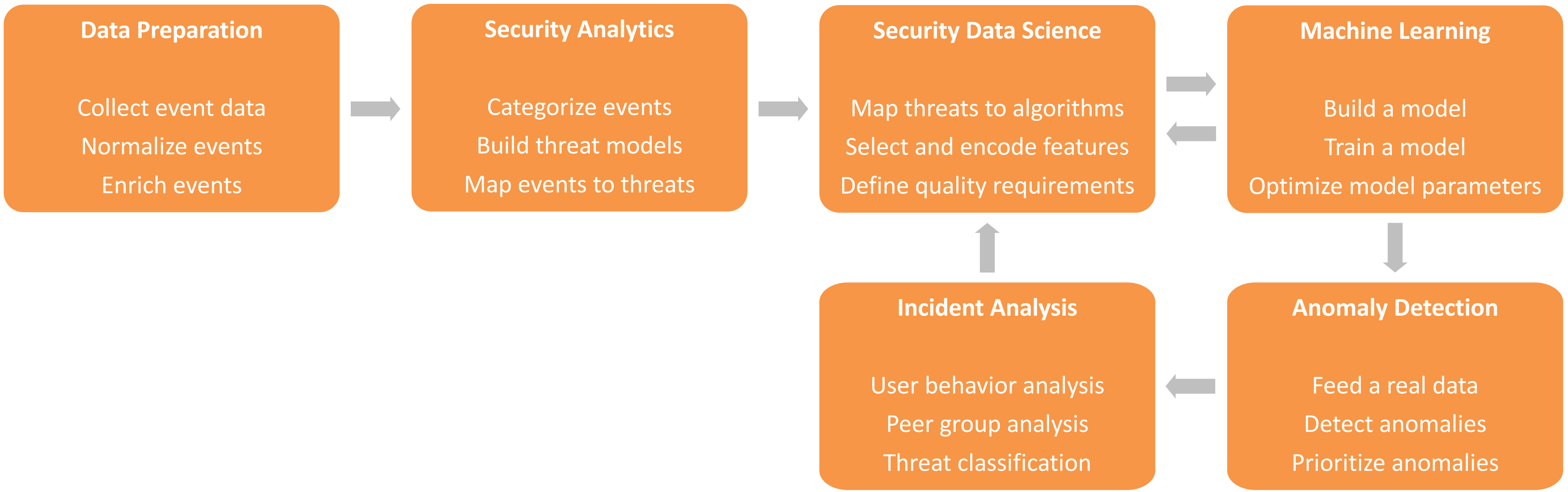
What?

- **ML model**
  - Codebase
  - Features structure
  - Model parameters (learned)
  - Model hyperparameters (architecture)

- **ML features**
  - Categorical (classes)
  - Statistical (counts)
  - Empirical (facts)
  - Continuous
  - Binary
  - ...

MACHINE LEARNING

How?



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**
  - When, Who, did What, Where from, Where to, on What
- **Bring all formats to the same convention**
  - Implement a built-in convertor for each format as a part of the solution (inside)
  - Create a separate convertor tool and treat it as the data source for the model (outside)
  - Build event storage that allows event fields mapping, like Splunk or ELK (infrastructure)
- **Find duplicates and missing fields**
  - One action generates several entries
  - System doesn’t identify itself in its own logs
  - User’s name is recorded, but not its IP (or vice versa)

DATA NORMALIZATION: BEFORE

SAP Security Audit Log ABAP

```
2AU520180313113209000030400001D1nsalab  SAP*                    SAPMSSY1                    0001F&0
                              nsalab          2AUK20180313113209000030400001D1nsalab  SAP*
SAPMSSY1                         0001SLO6&SAPLSLO6&RSAU_READ_FILE                         nsalab
2AU220180313114609002315800004D4MacBook-SAP*      SESSION_MANAGER    SAPMSYST
0001A&1                                              MacBook-Pro-Nursulta2AU120180313114703002315800004D4MacBook-
SAP*      SESSION_MANAGER    SAPMSYST                      0011A&0&P                                  MacBook-
Pro-Nursulta2AUW20180313114703002315800004D4MacBook-SAP*      SESSION_MANAGER    RSRZLLG0
          0011RSRZLLG0&                                    MacBook-Pro-
Nursulta2AUW20180313114703002315800004D4MacBook-SAP*      SESSION_MANAGER    RSRZLLG0_ACTUAL
0011RSRZLLG0_ACTUAL&                               MacBook-Pro-
Nursulta2AU320180313115152002316200008D8MacBook-SAP*      SE16          SAPLSMTR_NAVIGATION
0011SE16                                            MacBook-Pro-Nursulta2DU920180313115155002316200008D8MacBook-
SAP*      SE16          SAPLSETB                     0011USR02&02&passed                                MacBook-Pro-
Nursulta
```

DATA NORMALIZATION: AFTER

SAP Security Audit Log ABAP

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

Security Analytics

ERP SECURITY LOGGING

- **Common business application logging**
  - Event time
  - Event type
  - Server info
  - User info
  - ...

ERP SECURITY LOGGING

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

ERP SECURITY LOGGING

SAP Security Audit Log ABAP

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

THREAT MODEL

Use Cases

- **10+ Categories (why)**
  - Data Exfiltration, Account Compromise, Regular Access Abuse, Privileged Access Abuse, ...
- **30+ Classes (what)**
  - Data Transfer, Account Sharing, Password Attack, Privilege Escalation, Lateral Movement, ...
- **100+ Scenarios (how)**
  - 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
  - ...

ANOMALIES VS. THREATS

- Many anomalies are not malicious
- Anomalies are statistical deviations
- Big infrastructures always have anomalies

ANOMALIES VS. THREATS

Matrix Example

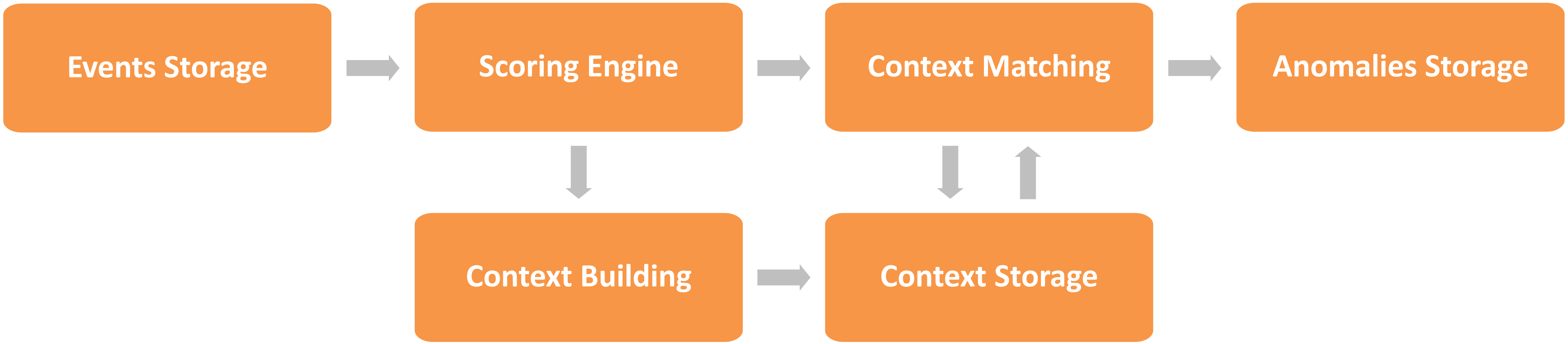
Threat Model		Temporal Anomalies			Static Anomalies		
Category	Class	Unusual action	Unusual time	Unusual volume	New action	New server	New device
Regular Access Abuse	Unauthorized Access	high	medium	low	high	medium	low
	Account Sharing	low	medium	high	low	medium	high
Account Compromise	Password Attack	medium	low	high	low	high	high
	Privilege Escalation	high	medium	low	high	medium	low
	Access Enumeration	high	low	medium	high	medium	low
Data Exfiltration	Data Transfer	low	medium	high	low	high	medium

Static
Anomalies

STATIC ANOMALY DETECTION

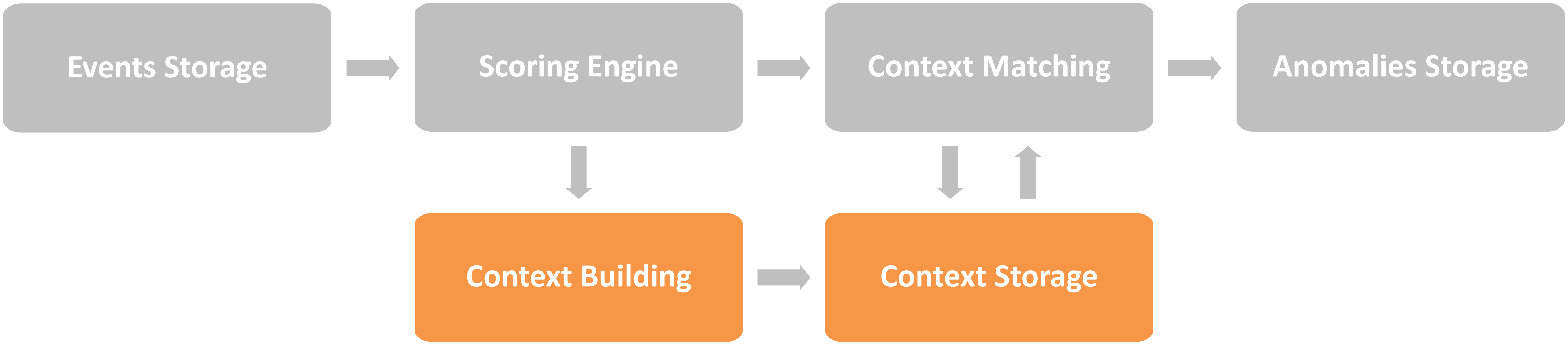
Plan

- Context building
- Context matching
- Anomaly analysis



CONTEXT BUILDING

- Whitelist known values for all users
- Define anomaly scores for all fields



CONTEXT THRESHOLD

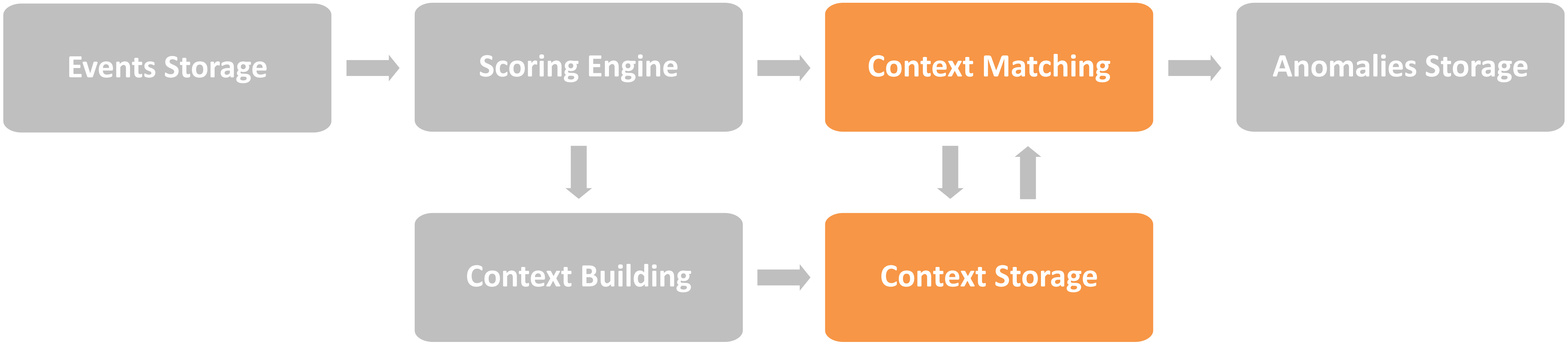
- **Problem**
  - Log poisoning attacks
  - Anomalies in user context
- **Solution**
  - Importance amplification
  - Mean of squared values

IP	Mean
172.16.100.11	320
172.16.100.118	308
172.16.100.137	30
Threshold	**219**

IP	Mean	Squared
172.16.100.11	320	102400
172.16.100.118	308	94864
172.16.100.137	30	900
172.16.100.200	1	1
172.16.100.201	1	1
172.16.100.202	1	[illegible]
172.16.100.203	[illegible]	[illegible]
172.16.100.204	[illegible]	[illegible]
172.16.100.217	[illegible]	1
172.16.10[illegible]	1	1
172.1[illegible].100.219	1	1
172.16.100.220	1	1
Threshold	**28**	**8,258**

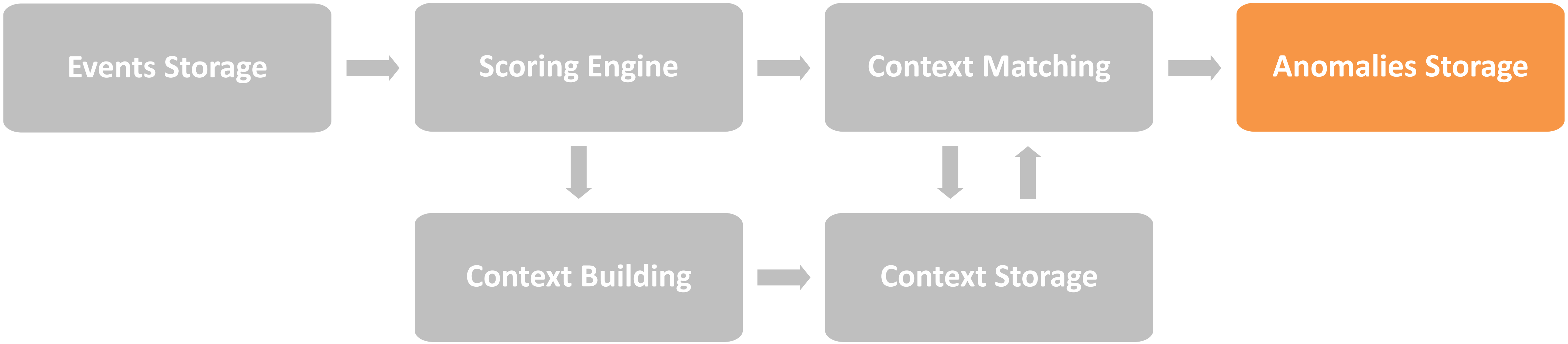
CONTEXT MATCHING

- Compare new events with the user context field by field
- Assign individual anomaly scores for unknown fields



ANOMALY ANALYSIS

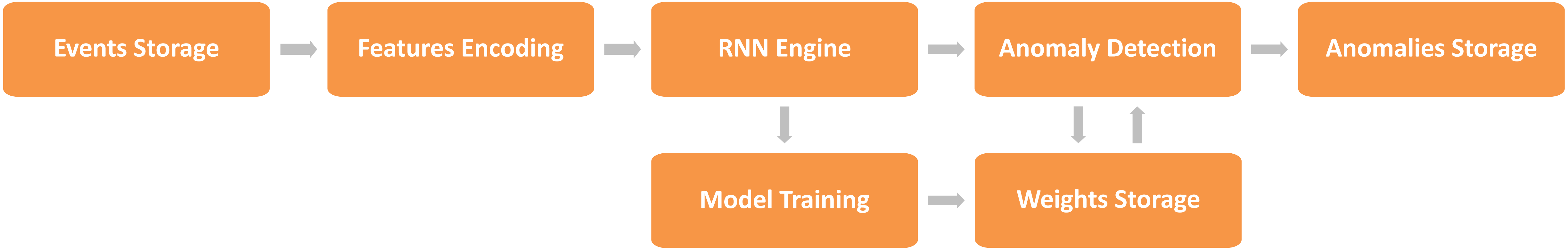
- Get a total event anomaly score from all its fields
- Get a total user anomaly score from all its events



Temporal Anomalies

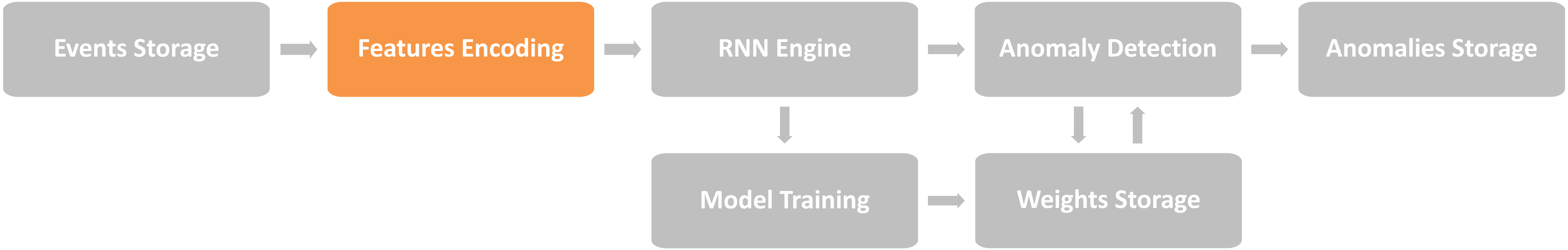
TEMPORAL ANOMALY DETECTION

- Establish a normal behavior baseline
- Train to predict normal user actions
- Analyze incorrectly predicted actions



FEATURE ENGINEERING

- Feature selection
- Feature encoding



FEATURE SELECTION

Data

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

FEATURE ENCODING

Vector

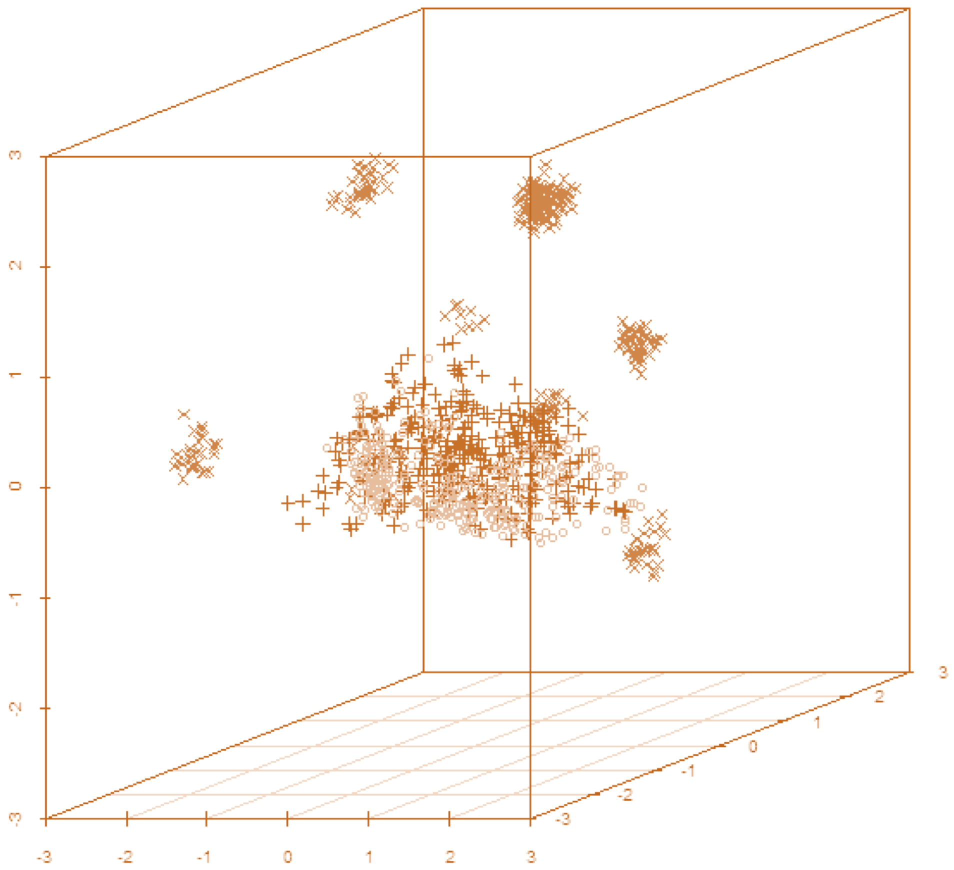
Time	Title	User	Device	Action	Context 1	Context 2	Context 3
3/13/18 11:32	RFC/CPIC Logon Successful	SAP*	nsalab	AU5	F	0	
3/13/18 11:32	Successful RFC Call	SAP*	nsalab	AUK	SLO6	SAPLSLO6	RSAU_READ_FILE

[0.19248842592592594 0.7110773240660063 0.8366013071895425]

FEATURE ENCODING

Knowledge Base

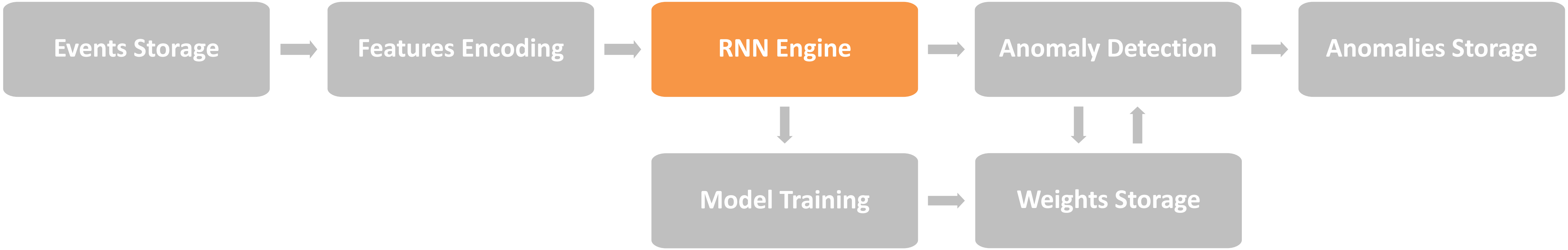
- **On-the-fly KB**
- **Security-focused KB**
- **Application-focused KB**
  - Static (1/100000 scale)
  - Mapping (1/100 scale)



Machine Learning

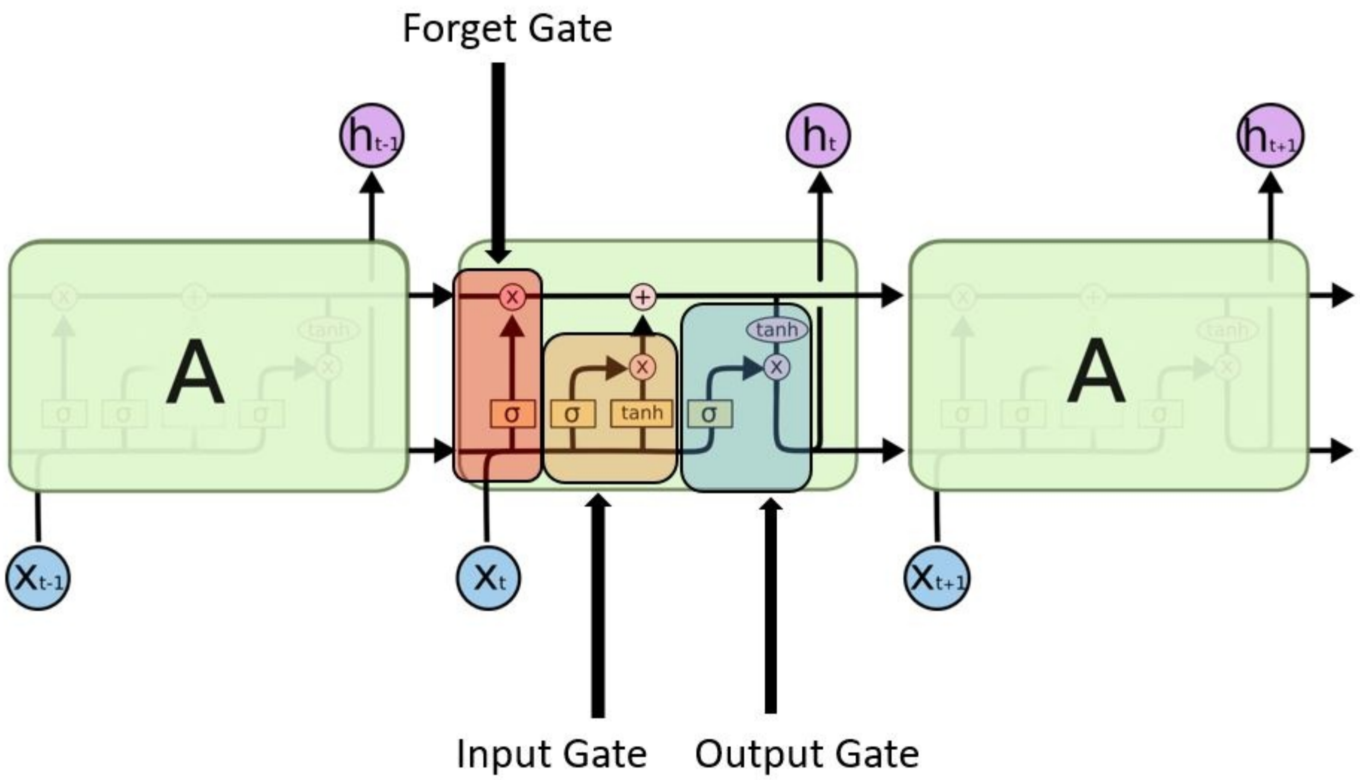
MODEL IMPLEMENTATION

- Find the right algorithm for a task
- Implement a model and its environment
- Optimize the model for the best accuracy



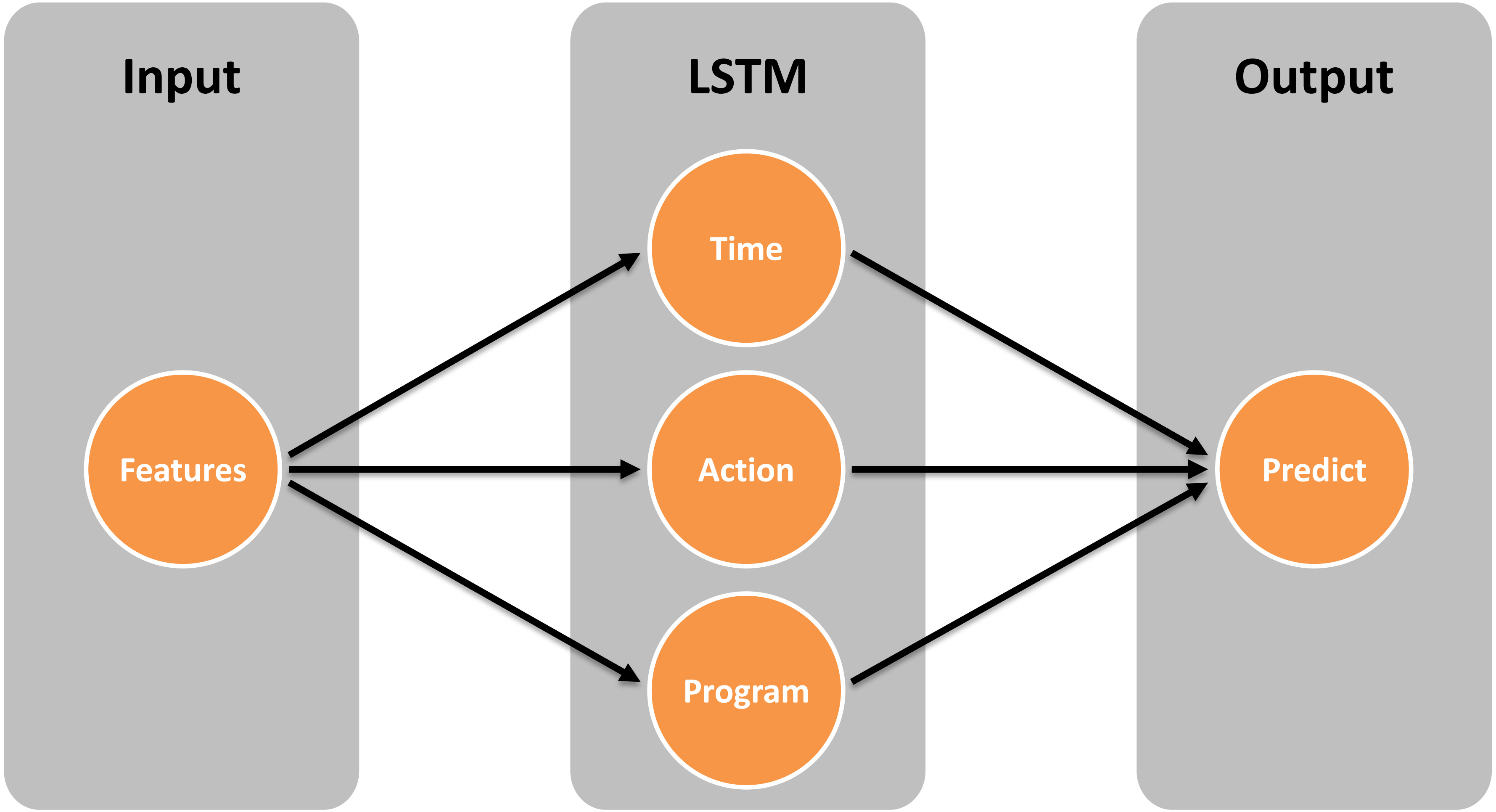
MODEL MEMORY

- **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
  - ...



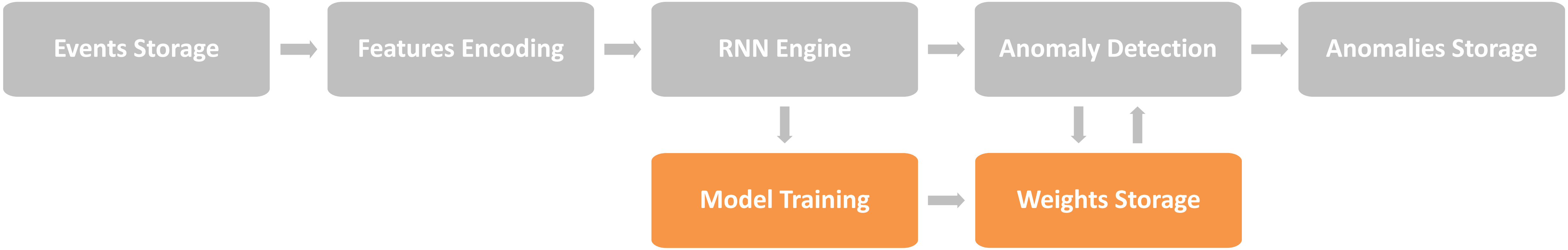
MODEL DESIGN

Architecture



MODEL PARAMETERS

- **Architecture**
  - Layers number, Neurons number, Activation function, Loss function, Optimizer, ...
- **Data**
  - Features, Knowledge base, Sequence length, Normalization, ...
- **Training**
  - Epochs, Bach size, Threshold, Distance, Smoothing, ...

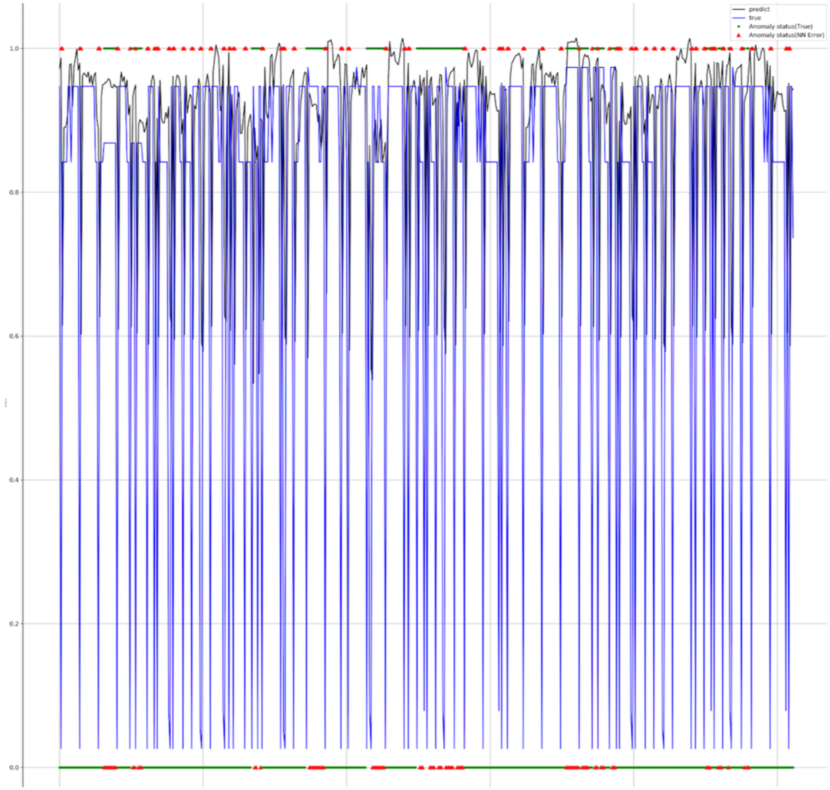


SEQUENCE LENGTH

- 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

KNOWLEDGE BASE SORTING

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

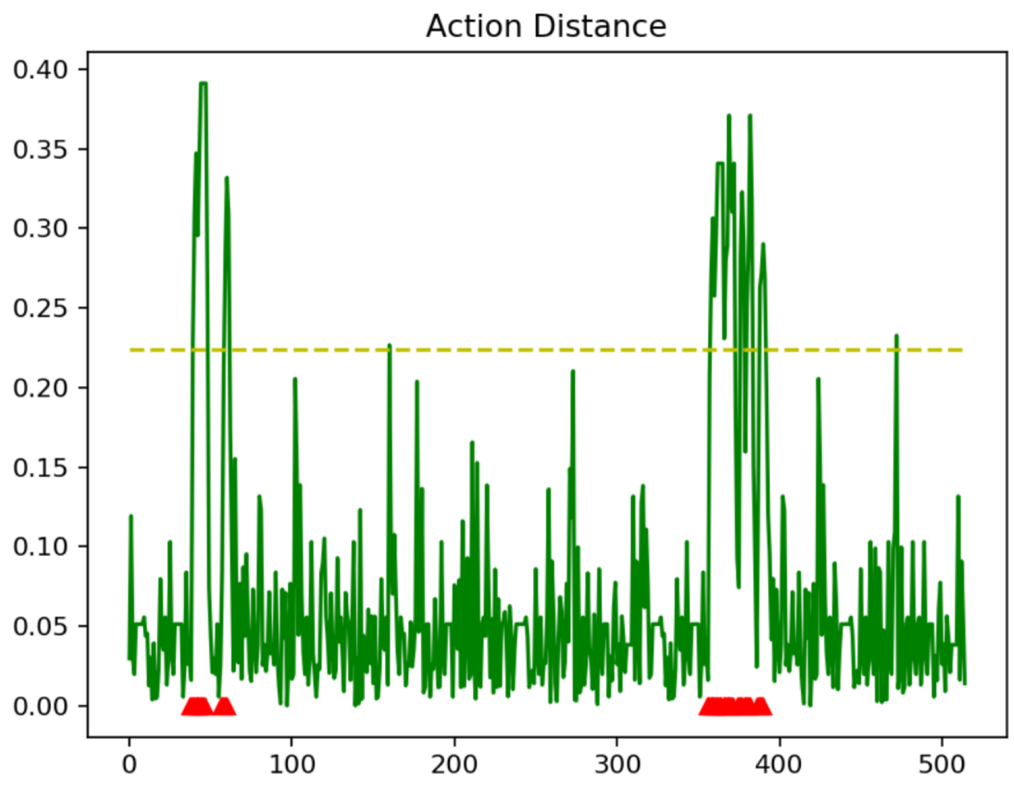


Sorted by Alphabet

Sorted by Frequency

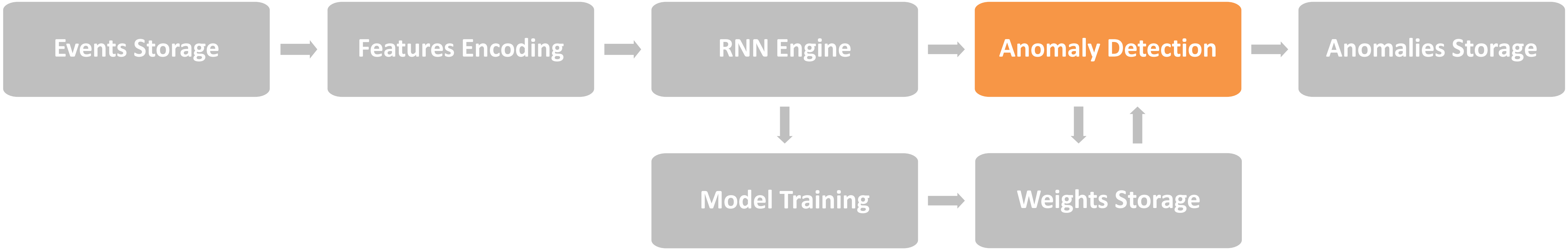
ADAPTIVE THRESHOLD

- **Error score**
  - Distance-based
    - Predicted value (blue)
    - Actual value (green)
- **Threshold**
  - Max training error score
- **Sensitivity**
  - As is
  - Coefficient



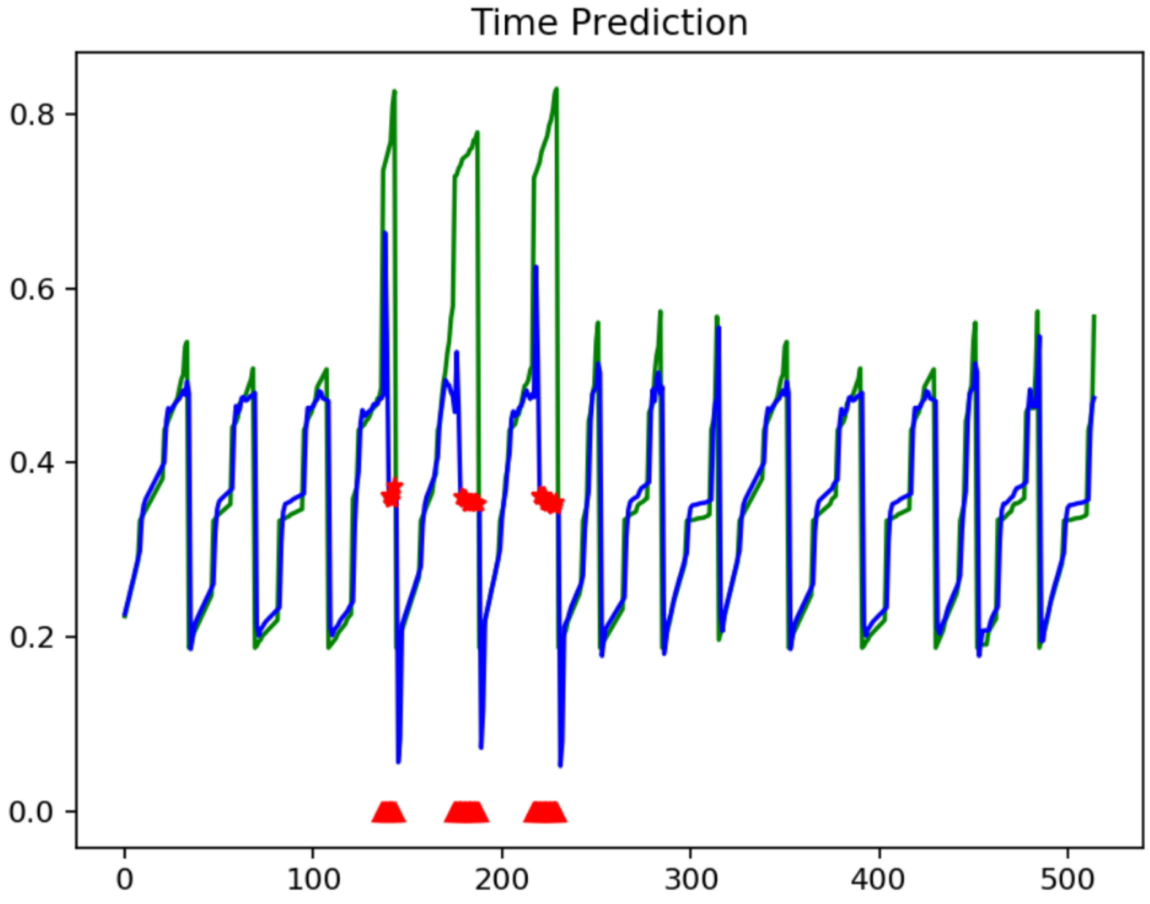
ANOMALY DETECTION

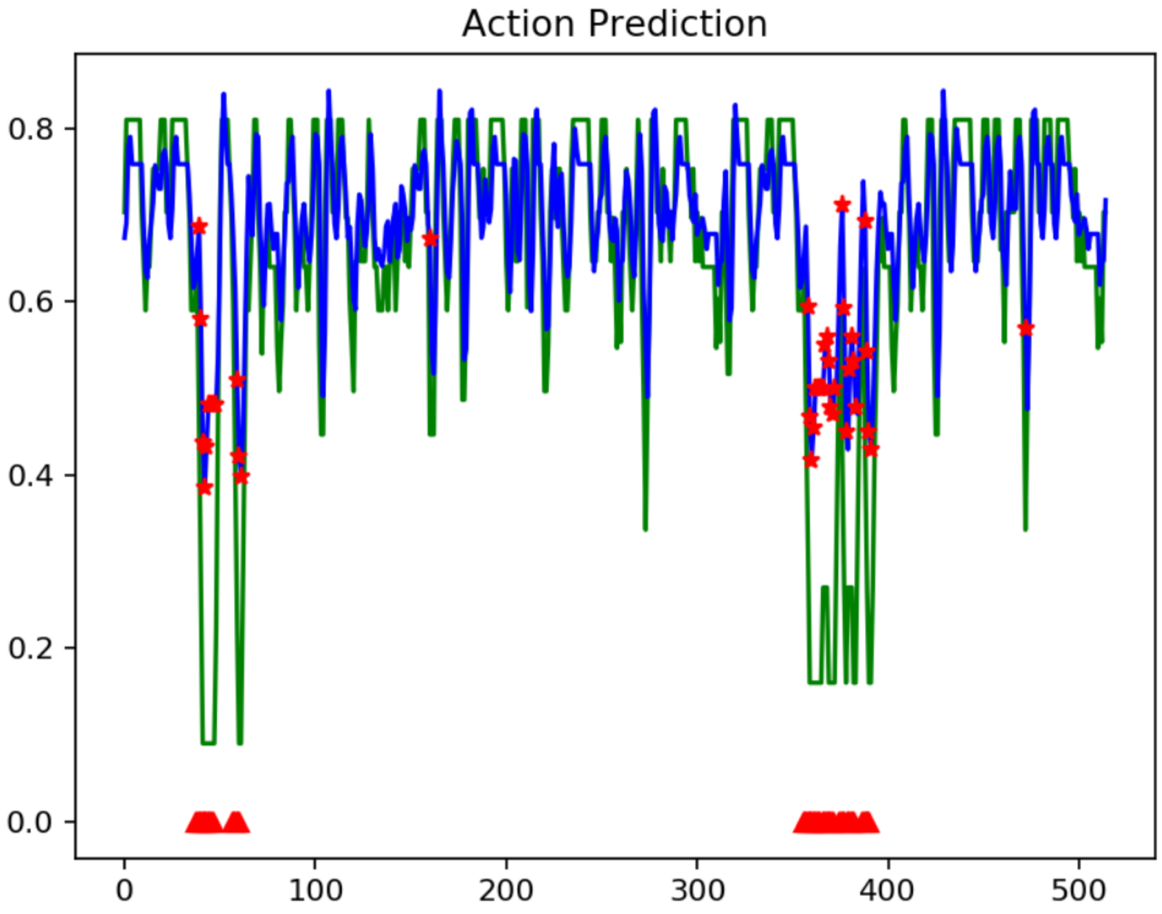
- Predict a potential user activity
- Report incorrectly predicted events above threshold



ANOMALY DETECTION

Prediction

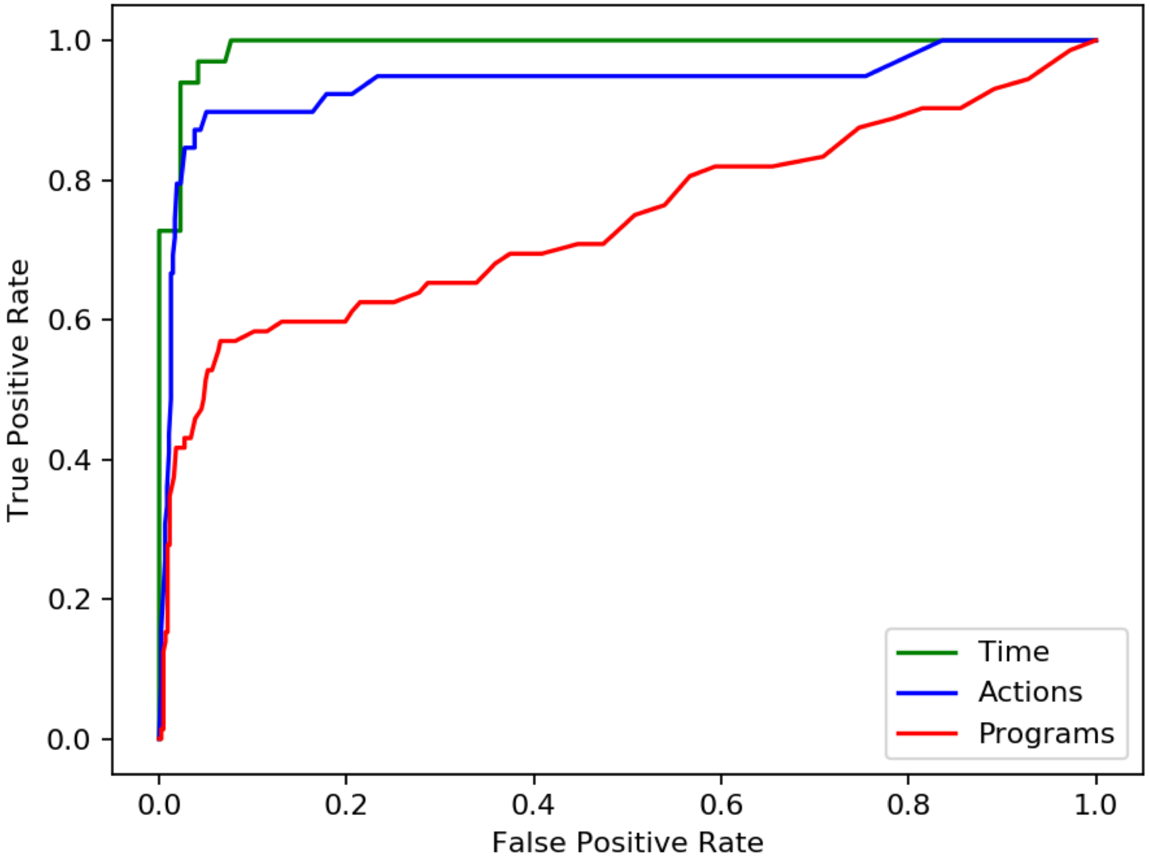




ANOMALY DETECTION

Metrics

- **Accuracy 95%**
  - True Positives 71%
  - True Negatives 97%
- **Errors 5%**
  - False Positives 3%
  - False Negatives 29%



CONCLUSIONS

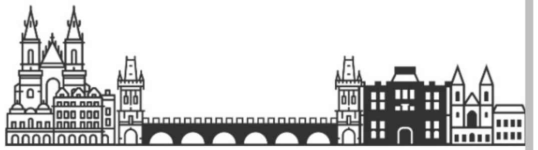
- **Security analytics is more important than machine learning**
- **ML-driven solutions must help analysts and not replace them**
- **Adjust accuracy and tolerance to false positives for your situation**
- **Build an ecosystem of ML models and advanced analytics on top of it**

AI BLESS YOU

Eugene Neyolov
Head of R&D
neyolov@erpscan.com

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

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

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