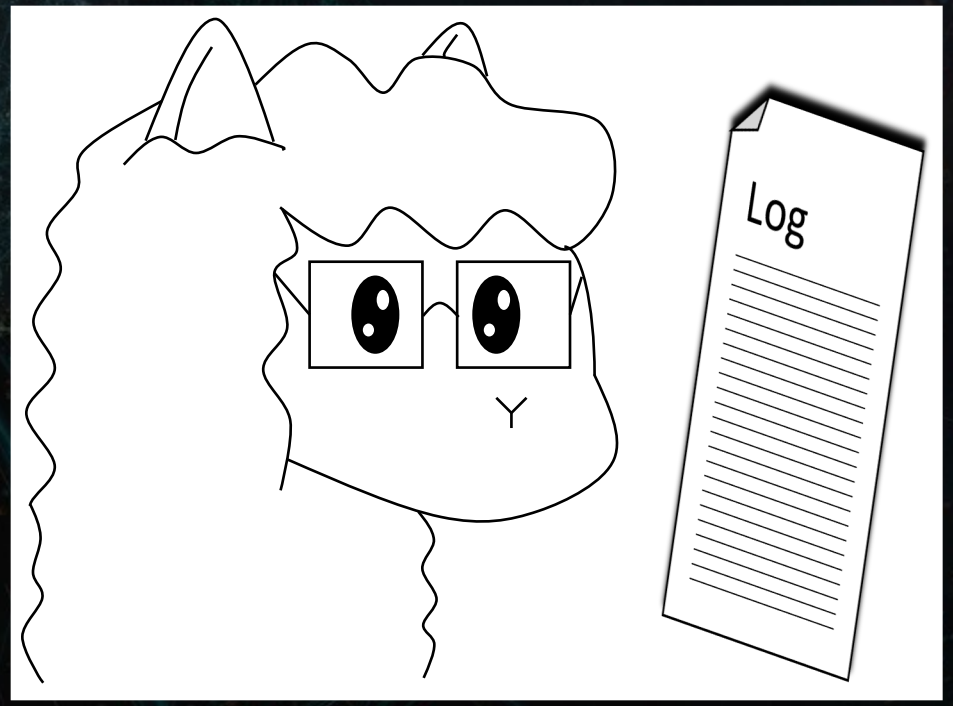
Log Analysis Machine Learner (LAMaLearner)



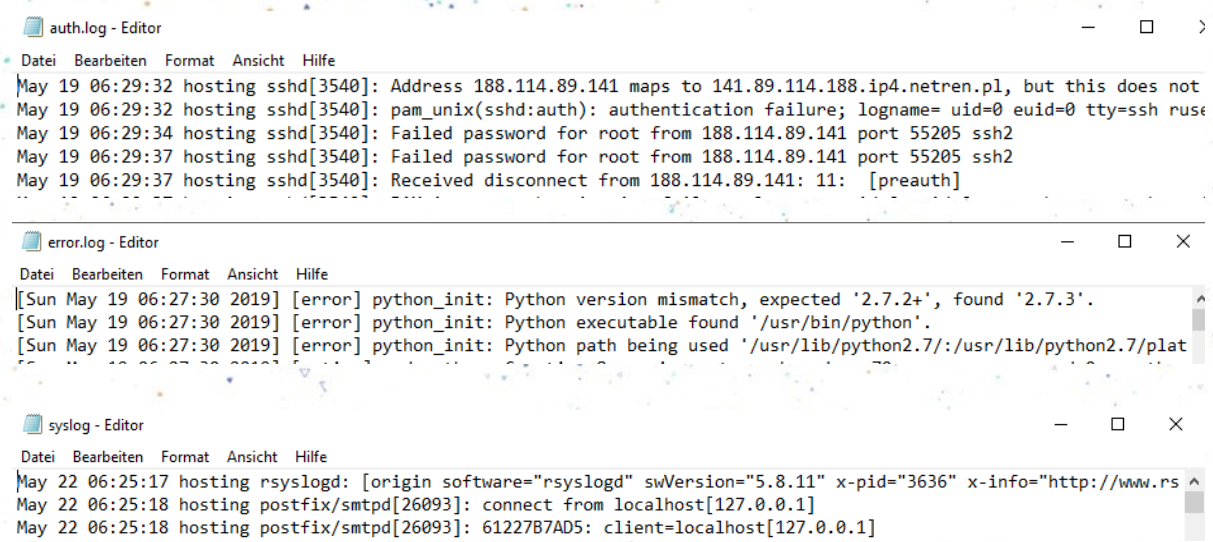
Agenda

- Log Management
- Security Information and Event Management (SIEM)
- Herausforderungen
- Wie LAMaLearner hilft
- MAD2 & LAMaLearner
- Anwendungsfälle
- Demo

SCHOELLER ONTEC

Log Management

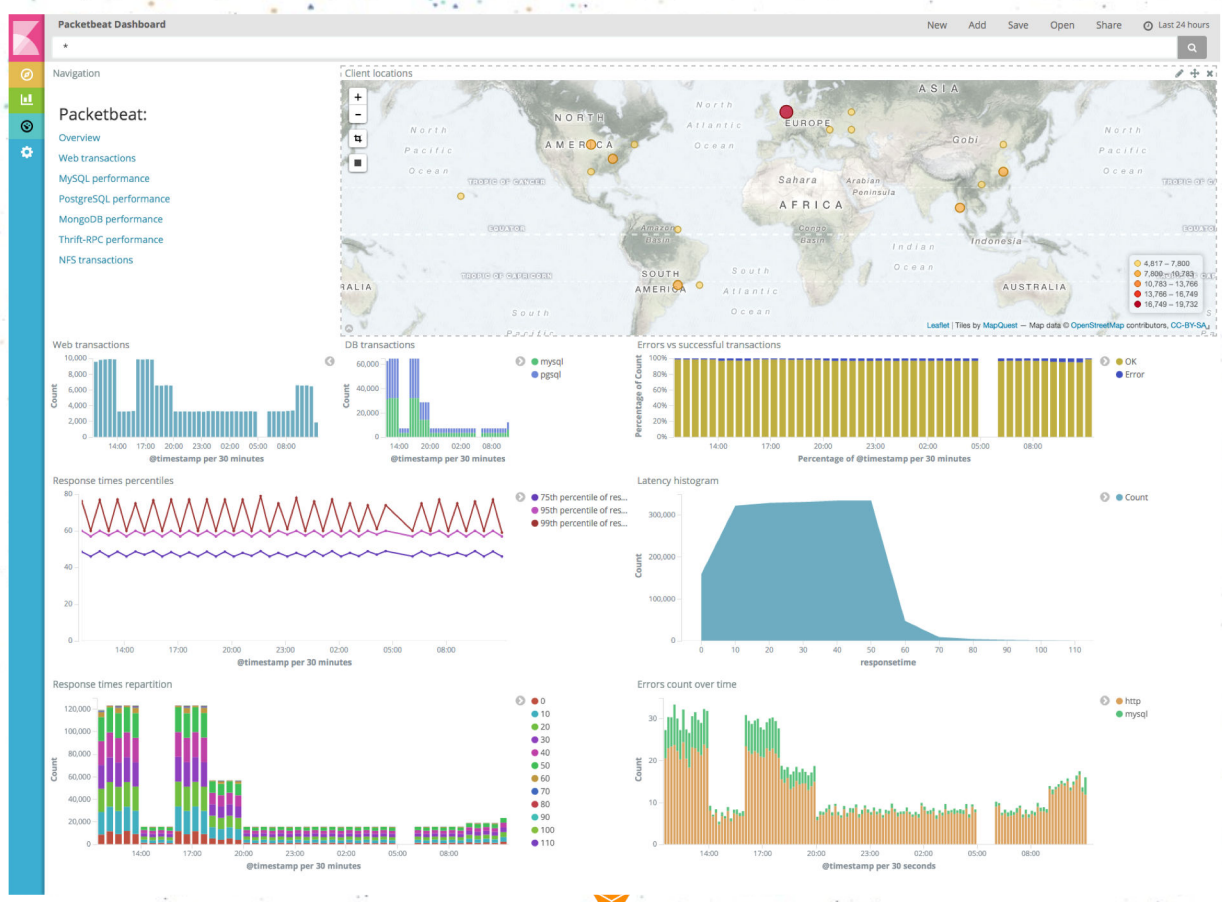
- Computer und Netzwerkkomponenten schreiben Log Files
- Diese können ausgewertet werden
  - Forensisch nach einen Fehler oder Vorfall
  - In Nahezu-Echtzeit während ein System arbeitet um Fehler zu bemerken
- Log Files können zentral gesammelt werden
  - Server mit Potentiell mehr Speicherplatz als auf individuellen Computern
  - Analyse auch möglich wenn Quell-Computer offline ist
- Kann für Compliance notwendig sein.



SCHOELLER ONTEC

Security Information and Event Management (SIEM)

- Security zentrische Erweiterung zum normalen Log Management
- Sammelt Logs aus heterogenen Komponenten
- Ermöglicht Visualisierung von Event Häufigkeiten
  - Anzahl fehlgeschlagener Log-Ins je Zeitintervall
- Ermöglicht Definition von Alarmierungsregeln. Z.B.:
  - 3 fehlgeschlagene Log-Ins auf einer Maschine in einer Minute
  - 15 Firewall Rejections in einer Minute
  - 1 Malware erkannt …



SCHOELLER ONTEC

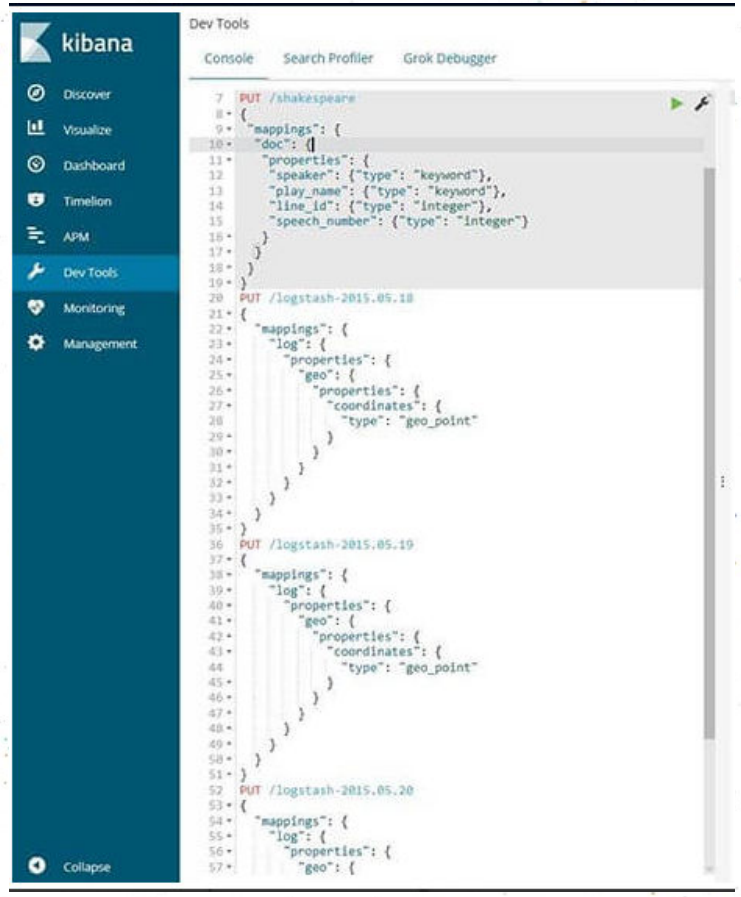
SIEM und Log Management Herausforderungen

- Volume & Velocity
  - Log Daten einer Organisation können schnell hunderte GB erreichen
  - Sammeln und Auswerten kann Systeme überfordern
- Variety
  - Es gibt keine verbindliche Normen für Logs
  - Jedes System schreibt Logs nach anderen Mustern z.B.:
    - Name, IP, Status, Meldungstext
    - Meldungstext, Zeit, Name, Status
  - Aktuelle SIEM Systeme funktionieren nur mit manuell definierten Mustern
    - Meldungstexte können völlig individuell sein
    - Namen und IPs sind Organisationsspezifisch

SCHOELLER ONTEC

Beispiel dafür Variety Herausforderungen

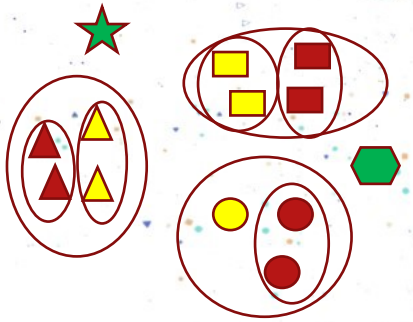
- Damit man in ElasticStack Events aggregiert werden können, muss ein entsprechendes Muster vorgegeben werden.
- Dafür muss das Muster bekannt sein.
- Das ist schwer, wenn man nicht vorher weiss, wonach man eigentlich sucht.

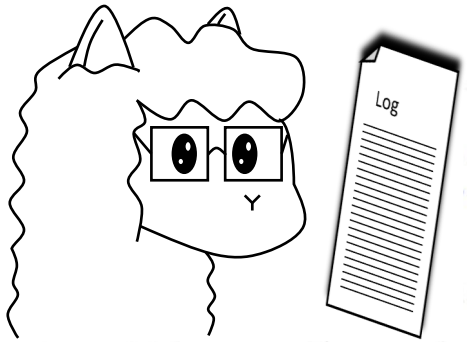


SCHOELLER ONTEC

LAMaLearner lernt selbst und hilft

- LAMaLearner analysiert Log Einträge mit unbekanntem Format
- LAMaLearner erzeugt Cluster inhaltlich ähnlicher Log Einträge
- Cluster können hierarchisch sein
- LAMaLearner erzeugt gemeinsame Muster für erkannte Cluster





LAMaLearner erzeugt hierarchische Cluster aus Logs

- 0 - 2019.05.16 - 09:58:00 - Router myrouter2 interface 10.10.10.12 down
  Delete
- 1 - 2019.05.16 - 09:58:01 - Router myrouter1 interface 192.168.13.1 down
  Delete
- 2 - 2019.05.16 - 09:58:02 - Router myrouter5 interface 192.168.22.33 down
  Delete
- 3 - 2019.05.16 - 09:58:03 - Router myrouter6 interface 192.168.22.31 down
  Delete
- 4 - 2019.05.16 - 09:58:04 - Router myrouter6 interface 192.168.22.31 down
  Delete
- 5 - 2019.05.16 - 09:58:05 - Connection from 192.168.1.1
  Delete
- 6 - 2019.05.16 - 09:58:06 - RSA key-gen for Tobi completed
  Delete
- 7 - 2019.05.16 - 09:58:07 - Password authentication for Willi accepted
  Delete
- 8 - 2019.05.16 - 09:58:08 - Interface 5 shut down
  Delete
- 9 - 2019.05.16 - 09:58:09 - RSA key-gen for Manfred completed
  Delete
- 10 - 2019.05.16 - 09:58:10 - Connection from 192.168.3.5
  Delete
- 11 - 2019.05.16 - 09:58:02 - Router myrouter5 interface 192.168.22.33 down
  Delete

Available clusters:

- Id: 0: Router interface down Total Risk Score: 0 Events Labels:
  - Id: 1: Router myrouter5 interface 192.168.22.33 down Total Risk Score: 0 Events Labels:
  - Id: 2: Router myrouter6 interface 192.168.22.31 down Total Risk Score: 0 Events Labels:
- Id: 3: Connection from Total Risk Score: 0 Events Labels:
- Id: 4: RSA key-gen for completed Total Risk Score: 0 Events Labels:

Events that are not in clusters:

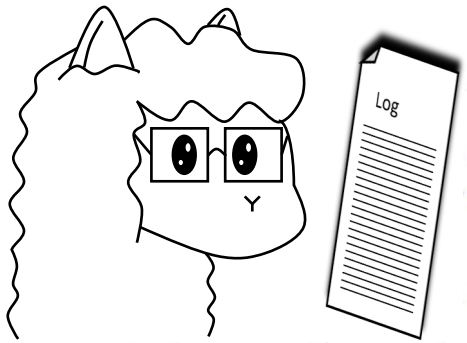
Show in Events

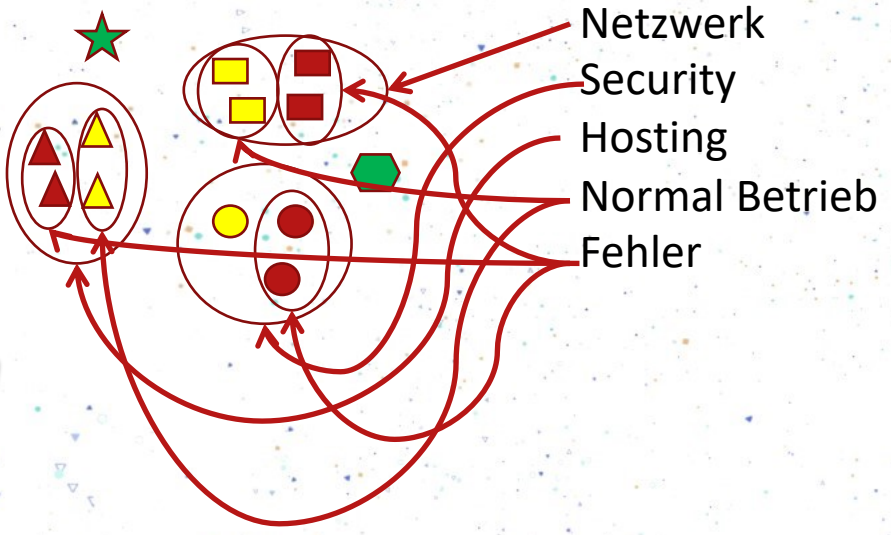
- Id: 7
- Id: 8

SCHOELLER ONTEC

LAMaLearner lernt selbst und hilft

- LAMaLearner ermöglicht Annotation von Clustern mit beliebigen Labeln
- Mehrere Label können pro Cluster vergeben werden





SCHOELLER ONTEC

Cluster werden durch frei wählbare Labels annotiert

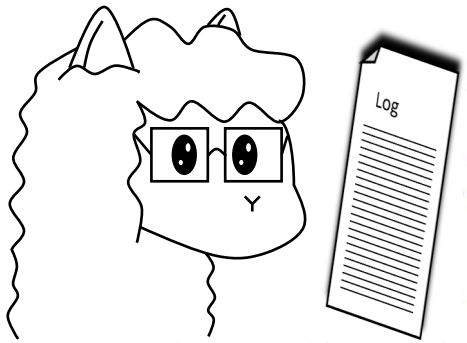
Available clusters:

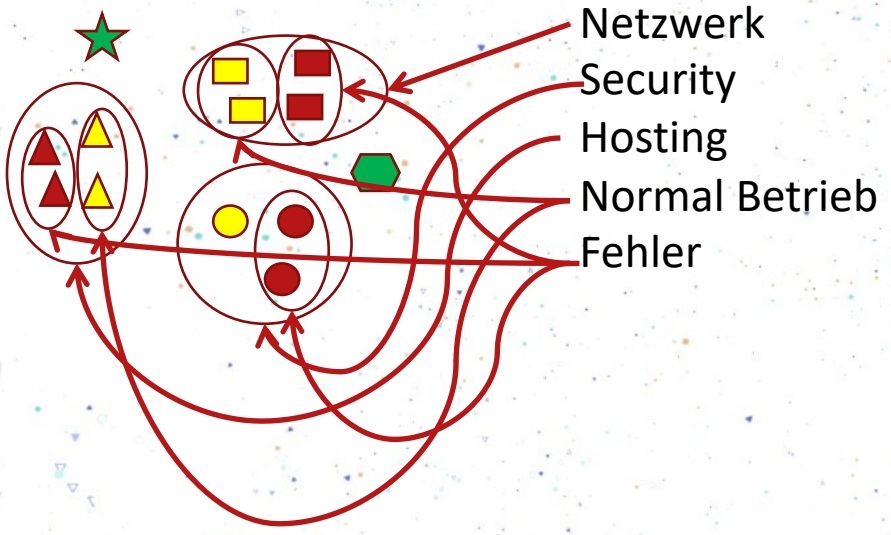
- Id: 0: Router interface down Total Risk Score: 5 Events Labels:
  failure
  network
  success ▾ add
  - Id: 1: Router myrouter5 interface 192.168.22.33 down Total Risk Score: 5 Events Labels:
    failure
    network
    success ▾ add
  - Id: 2: Router myrouter6 interface 192.168.22.31 down Total Risk Score: 5 Events Labels:
    failure
    network
    success ▾ add
- Id: 3: Connection from Total Risk Score: 0 Events Labels:
  success ▾ add
- Id: 4: RSA key-gen for completed Total Risk Score: -3.5 Events Labels:
  success
  security
  failure ▾ add

SCHOELLER ONTEC

LAMaLearner lernt selbst und hilft

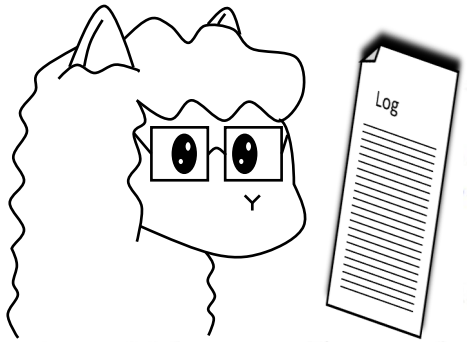
- Sind genügend Cluster annotiert kann LAMaLearner selbst lernen Cluster zu unterscheiden.
  Z.B.:
  - Viereck → Netzwerk
  - Dreieck → Hosting
  - Kreis → Security
  - Rot → Fehler
  - !Rot → Normal Betrieb
- Dieses Modell wird komplett eigenständig gelernt.

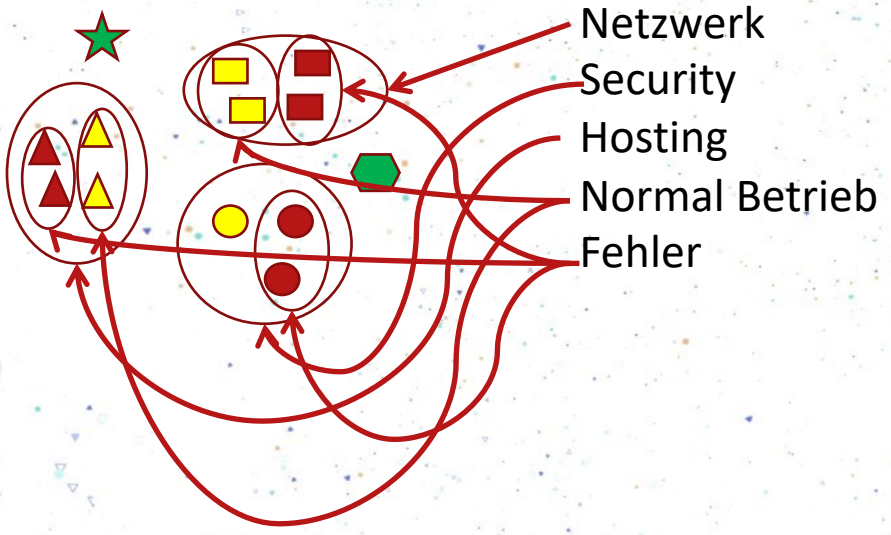


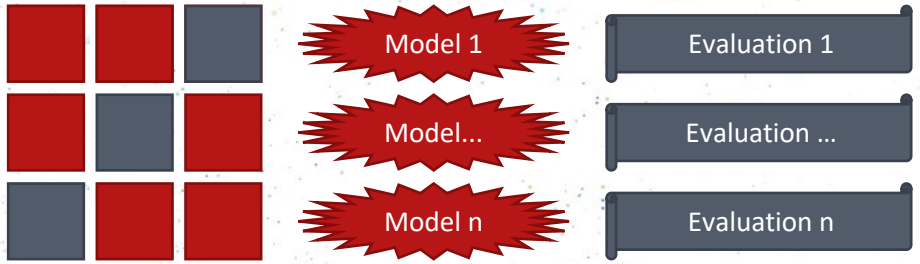


So lernt LAMaLearner und testet sich

- LAMaLearner testet sich selber mit n-fold cross-validation
  - n ist eine natürliche Zahl. Als Beispiel hier n=3.
  - Alle bekannten Zuordnungen werden in n Teilmengen unterteilt
  - Dann in n Iterationen:
    - n-1 Teilmengen zum lernen
    - 1 Teilmenge zum testen
- Bestes Modell wird ausgewählt und gespeichert







SCHOELLER ONTEC

LAMaLearner generiert Modelle

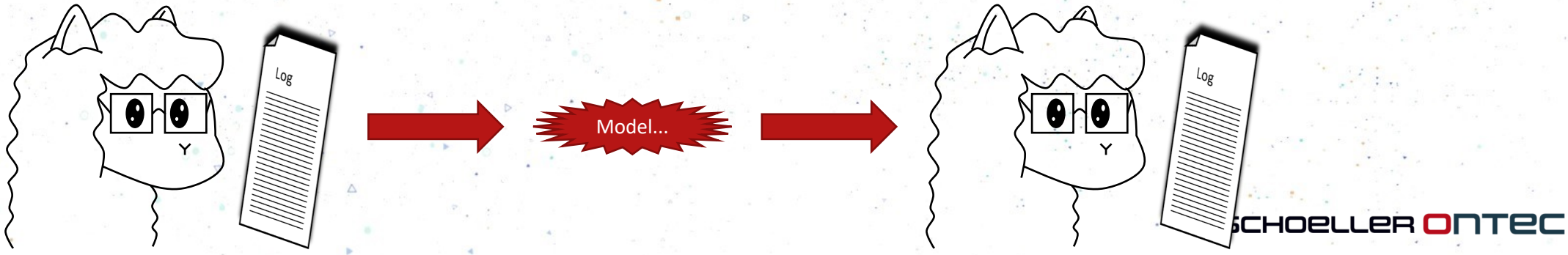
- Modelle speichern Labels und Regeln diese zu identifizieren
- Modelle können gespeichert und zwischen LAMaLearner Instanzen übertragen werden.
- Evaluationsdaten werden als Metadaten ins Modell integriert

- Creator: Tobias Eljasik-Swoboda
- Log Type: Windows Security Event Log
- Usage: Tells failures from successes
- Creation Date: 2019-09-02 13:50:48.149
- Version: 0.10.1-SNAPSHOT
- Training description: Trained and evaluated on 744 clusters using 3-fold cross-validation. Selection Policy: MicroaverageF1

Fold:	0						
Microaverage Precision:	1	Microaverage Recall:	1	Microaverage F1:	1		
Macroaverage Precision:	1	Macroaverage Recall:	1	Macroaverage F1:	1		
Category Id:	Category Label:	True Positives	False Positives	False Negatives	Precision	Recall	F1
0	success	123	0	0	1	1	1
1	failure	4	0	0	1	1	1

Fold:	1						
Microaverage Precision:	1	Microaverage Recall:	1	Microaverage F1:	1		
Macroaverage Precision:	1	Macroaverage Recall:	1	Macroaverage F1:	1		
Category Id:	Category Label:	True Positives	False Positives	False Negatives	Precision	Recall	F1
0	success	124	0	0	1	1	1
1	failure	2	0	0	1	1	1

Fold:	2						
Microaverage Precision:	1	Microaverage Recall:	1	Microaverage F1:	1		
Macroaverage Precision:	1	Macroaverage Recall:	1	Macroaverage F1:	1		
Category Id:	Category Label:	True Positives	False Positives	False Negatives	Precision	Recall	F1
0	success	125	0	0	1	1	1
1	failure	1	0	0	1	1	1



Wie wird Effektivität gemessen?

- True Positive (TP): Getesteter Cluster hatte tatsächlich Label das es haben sollte.
- False Positive (FP): Getesteter Cluster wurde dem falschen Label zugeordnet
- False Negative (FN): Getester Cluster hätte dieses Label haben sollen, hatte aber ein anderes
- Precision: Verhältnis TP zu FP: Wieviel vom gefundenen richtig ist.
- Recall: Verhältnis TP zu FP: Wieviel vom zu Findenden gefunden wurde
- F1: Geometrisches Mittel aus Precision und Recall
- Microaverage: Gesamtwerte aus einzelnen TP, FP und FN berechnen
- Macroaverage: Durchschnittswerte

- Creator: Tobias Eljasik-Swoboda
- Log Type: Windows Security Event Log
- Usage: Tells failures from successes
- Creation Date: 2019-09-02 13:50:48.149
- Version: 0.10.1-SNAPSHOT
- Training description: Trained and evaluated on 744 clusters using 3-fold cross-validation. Selection Policy: MicroaverageF1

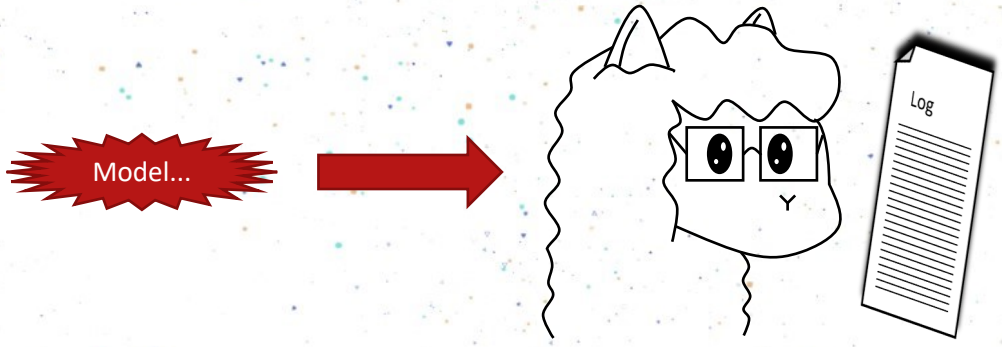
Fold:	0						
Microaverage Precision:	1	Microaverage Recall:	1	Microaverage F1:	1		
Macroaverage Precision:	1	Macroaverage Recall:	1	Macroaverage F1:	1		
Category Id:	Category Label:	True Positives	False Positives	False Negatives	Precision	Recall	F1
0	success	123	0	0	1	1	1
1	failure	4	0	0	1	1	1

Fold:	1						
Microaverage Precision:	1	Microaverage Recall:	1	Microaverage F1:	1		
Macroaverage Precision:	1	Macroaverage Recall:	1	Macroaverage F1:	1		
Category Id:	Category Label:	True Positives	False Positives	False Negatives	Precision	Recall	F1
0	success	124	0	0	1	1	1
1	failure	2	0	0	1	1	1

Fold:	2						
Microaverage Precision:	1	Microaverage Recall:	1	Microaverage F1:	1		
Macroaverage Precision:	1	Macroaverage Recall:	1	Macroaverage F1:	1		
Category Id:	Category Label:	True Positives	False Positives	False Negatives	Precision	Recall	F1
0	success	125	0	0	1	1	1
1	failure	1	0	0	1	1	1

LAMaLearner macht, was das gelernte Modell sagt

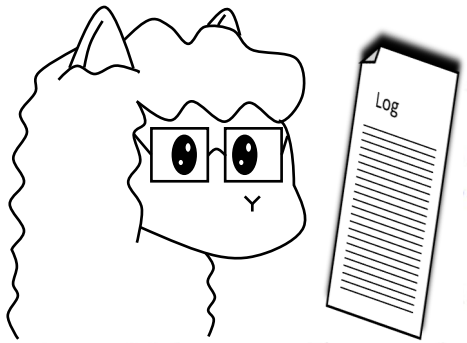
- Das heisst, LAMaLearner Instanzen müssen nicht bei 0 Anfangen.
- Neue Instanzen könnten mit vorhandenen Modellen / Wissen gestartet werden.
- So kann LAMaLearner Events mit gelernten Labels annotieren.
- Dies kann z.B. einen Pattern Matcher in Log Beats ersetzen ohne selber Pattern definieren zu müssen.

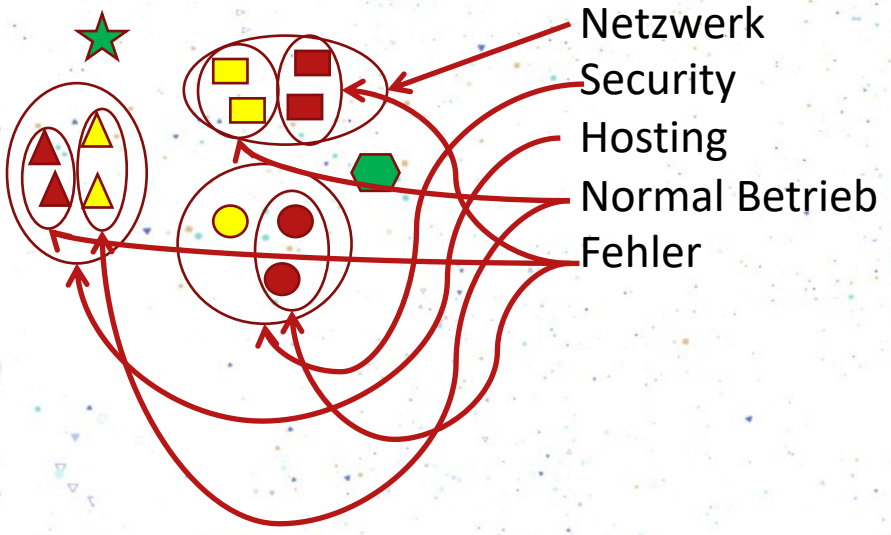


SCHOELLER ONTEC

Alternativ: Regel basierte Labels

- Neben dem Machine Learning Model kann LAMaLearner auch Label nach gefundenen Begriffen zuordnen.
- Diese können variabel mit Regeln kombiniert werden:
  - Nur Modell
  - Modell AND Regeln
  - Modell OR Regeln
  - Nur Regeln





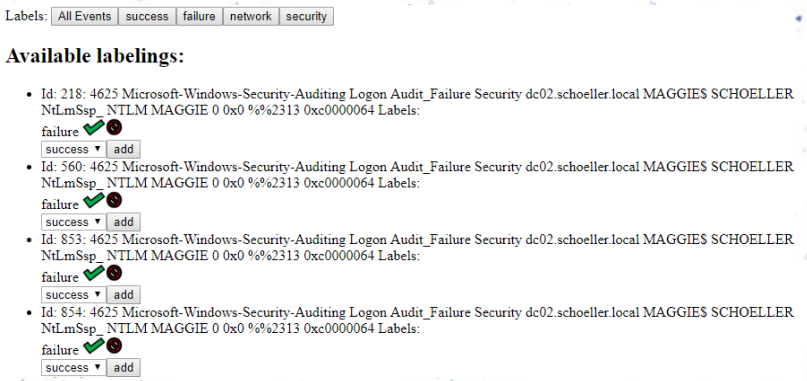
My rules:
Down → Fehler
Success → Normal Betrieb
Router → Netzwerk

SCHOELLER ONTEC

LAMaLearner lernt selbst und hilft

- LAMaLearner mit Regeln und/oder gelerntem Modell kann
  - Einzelne Events mit völlig unbekanntem Format Labeln zuordnen.
  - Neue Cluster finden und diese auch Labeln zuordnen.
- LAMaLearner kann alarmieren
  - Labels haben zugeordneten Risk Score.
  - Wenn Gesamt Risk Score eines Events oder Clusters > Threshold, wird externes Kommando aktiviert.

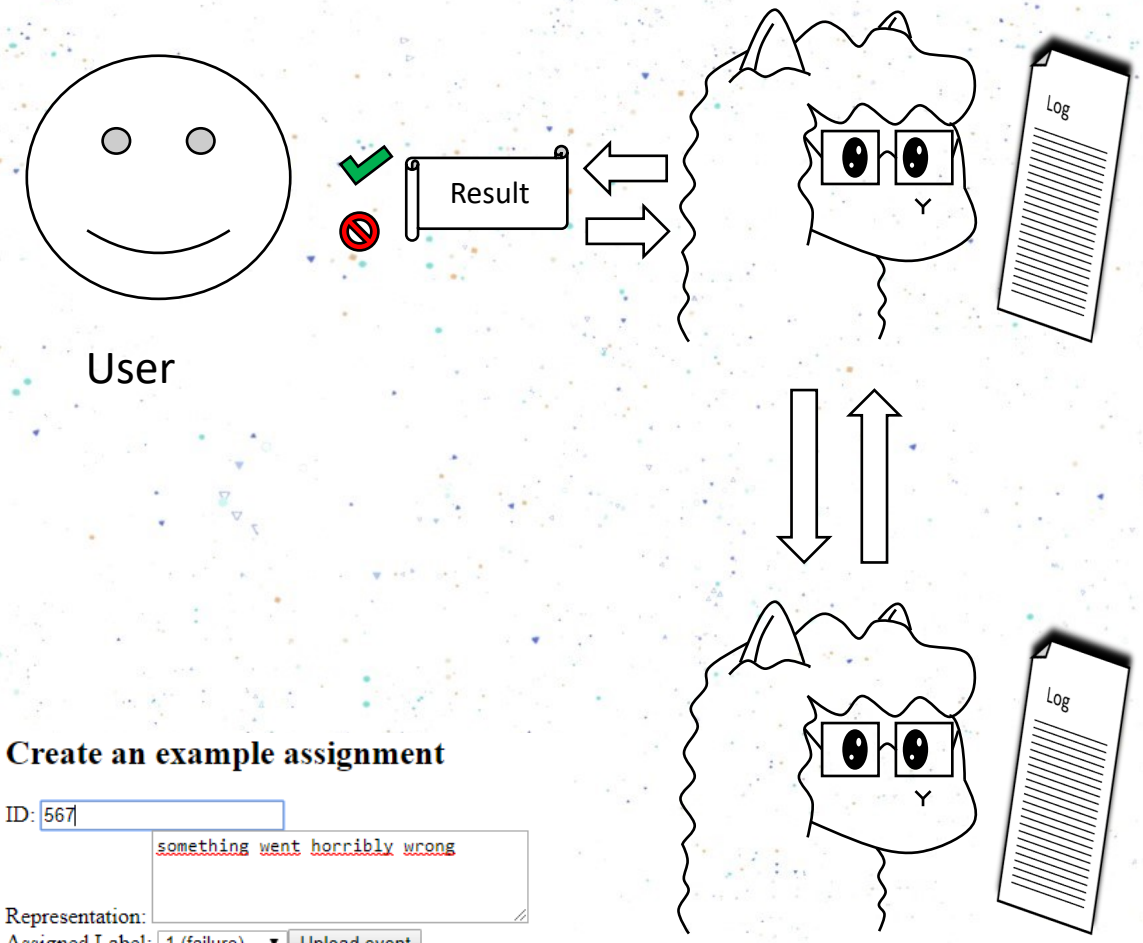
My rules:
Down → Fehler
Success → Normal Betrieb
Router → Netzwerk



SCHOELLER ONTEC

LAMaLearner lernt immer weiter

- User kann LAMaLearner Ergebnisse immer bestätigen oder korrigieren.
- Jeder Hinweis wird als Faktenwissen in LAMaLearner Instanz gespeichert
  - Event 1 hat Label A
  - Cluster 3 hat Label B
  - …
- Fakten können zwischen LAMaLearner Instanzen kopiert und zwischengespeichert werden.
- Fakten können editiert und manuell angelegt werden



Create an example assignment

ID: 567

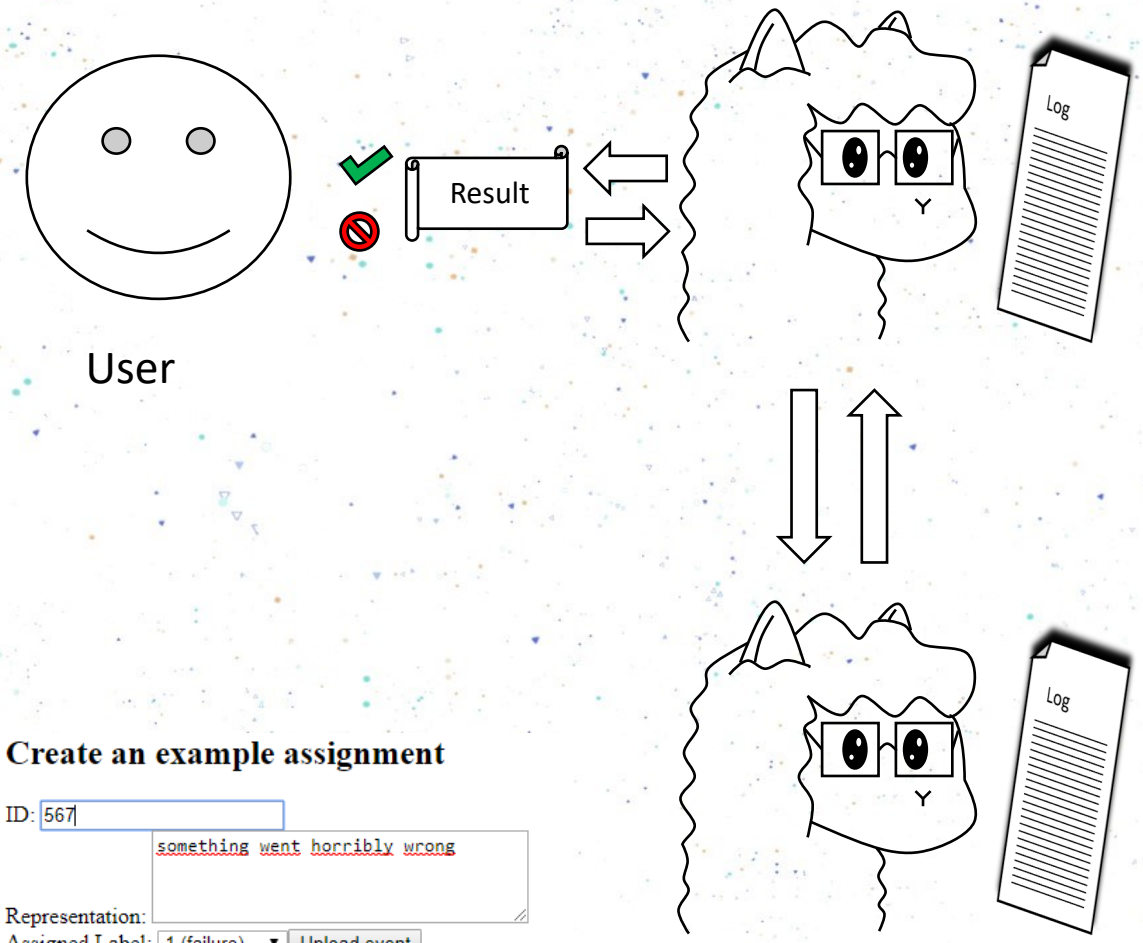
Representation: something went horribly wrong

Assigned Label: 1 (failure) Upload event

- Clear

LAMaLearner lernt immer weiter

- Aus mehr Fakten können bessere Modelle generiert werden.
- Bessere Modelle bringen bessere Ergebnisse.
- Modelle können zwischen LAMaLearner Instanzen kopiert werden.



Create an example assignment

ID: 567

Representation: something went horribly wrong

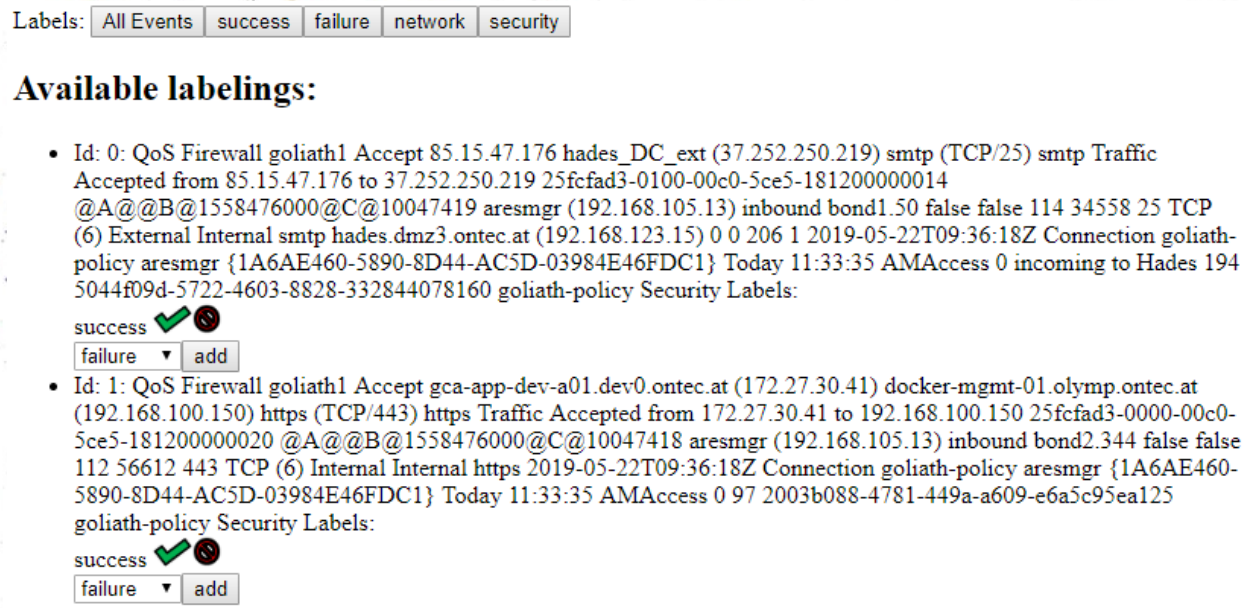
Assigned Label: 1 (failure) Upload event

- Clear

SCHOELLER ONTEC

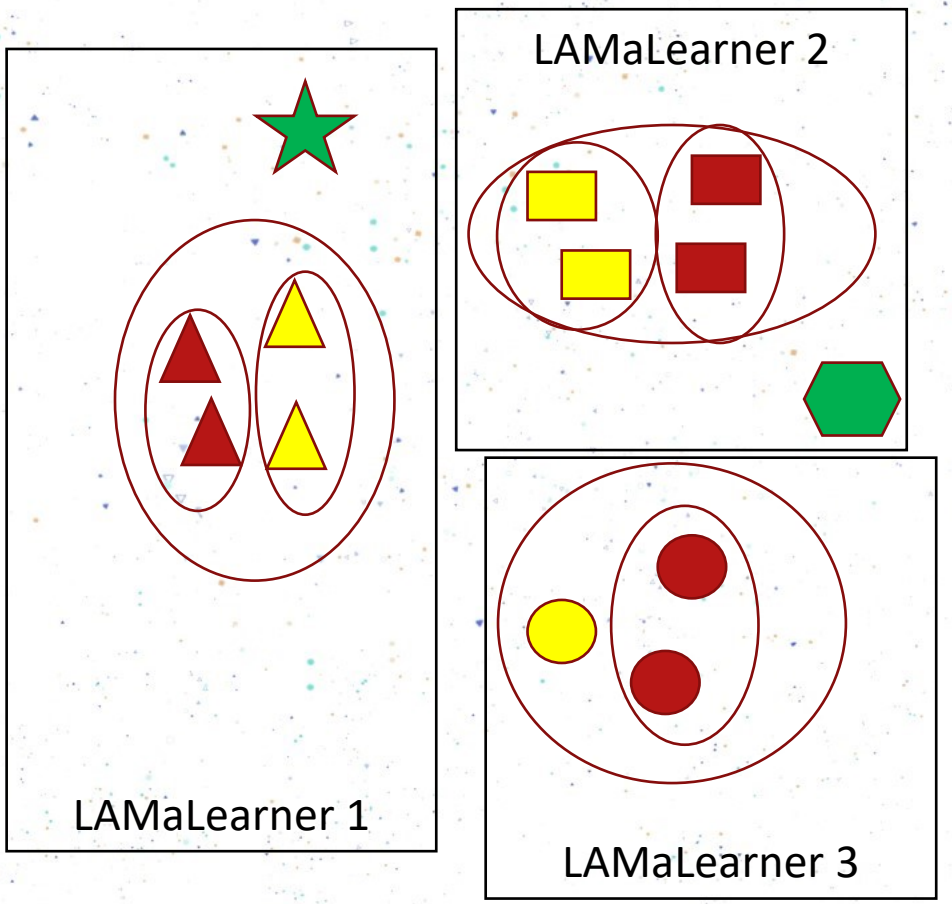
LAMaLearner vereinfacht Variety Herausforderungen

- LAMaLearner lernt Muster aus Log Events: Keine manuelle Definition notwendig.
- LAMaLearner macht Logs Übersichtlicher.
  - Millionen Events
  - 10 – 100 Cluster
  - 2-5 Labels
- LAMaLearner kann auch neue Formate Labeln
  - Z.B. Modell aus bekannten Logs erzeugen und dann auf neues Format anwenden.
  - Hier: Gelernt aus Windows Security Event Log – Gelabelt: Firewall Log



LAMaLearner ist Ready für Volume & Velocity

- Events können auf mehrere LAMaLearner Instanzen verteilt werden.
- Einzelne Instanzen können separat
  - Cluster finden
  - Cluster labeln
  - Events labeln
- Ergebnisse einzelner LAMaLearner Instanzen können wieder zusammengeführt werden.



SCHOELLER ONTEC

LAMaLearner ist mobil und flexibel

- LAMaLearner braucht keine Verbindung in eine Cloud
- LAMaLearner läuft überall wo es Java gibt
  - Getestet mit Java 9.0.1
  - auf Windows 10, Ubuntu Linux und Mac OS X
- LAMaLearner kommuniziert über eine einfache REST/JSON API
  - Port in Config File konfigurierbar
  - Mehrere Instanzen auf gleichem Rechner möglich
- Bringen Sie LAMaLearner zu Ihren Daten, nicht Ihre Daten zum LAMaLearner
  - Alle Funktionen in 24 MB Software
  - Model Dateien < 1 MB
- Scale Out Anywhere: Last Abhängig können LAMaLearner Work Loads auf beliebig viele Maschinen verteilt werden.
  - Auf Laptop mit Windows 10, Intel i7-7820 (Quad Core, 2.9 GHz) / 16 GB RAM:
    - Clustering: 1000 Events in ca. 3 Sekunden. 336 Events pro Sekunde
    - Event Labeling: 1000 Events in 100 ms. 10000 Events pro Sekunde

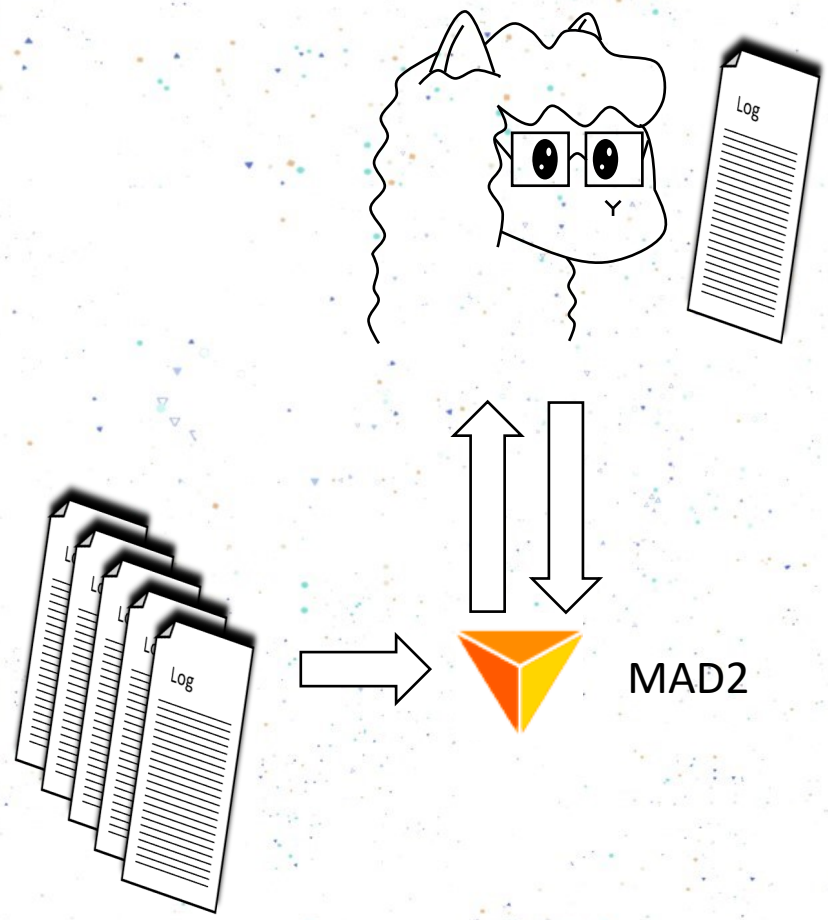
SCHOELLER ONTEC

Log files als REST / JSON ?!

- Log Files werden normalerweise nicht als JSON file geschrieben.
- Im ELK Stack werden die Events in Elastic Search gespeichert.
- Elastic Search hat auch eine REST / JSON Schnittstelle.
- Logstash sammelt Log Files für Elastic Search
- LAMaLearner nutzt zu diesem Zweck MAD2

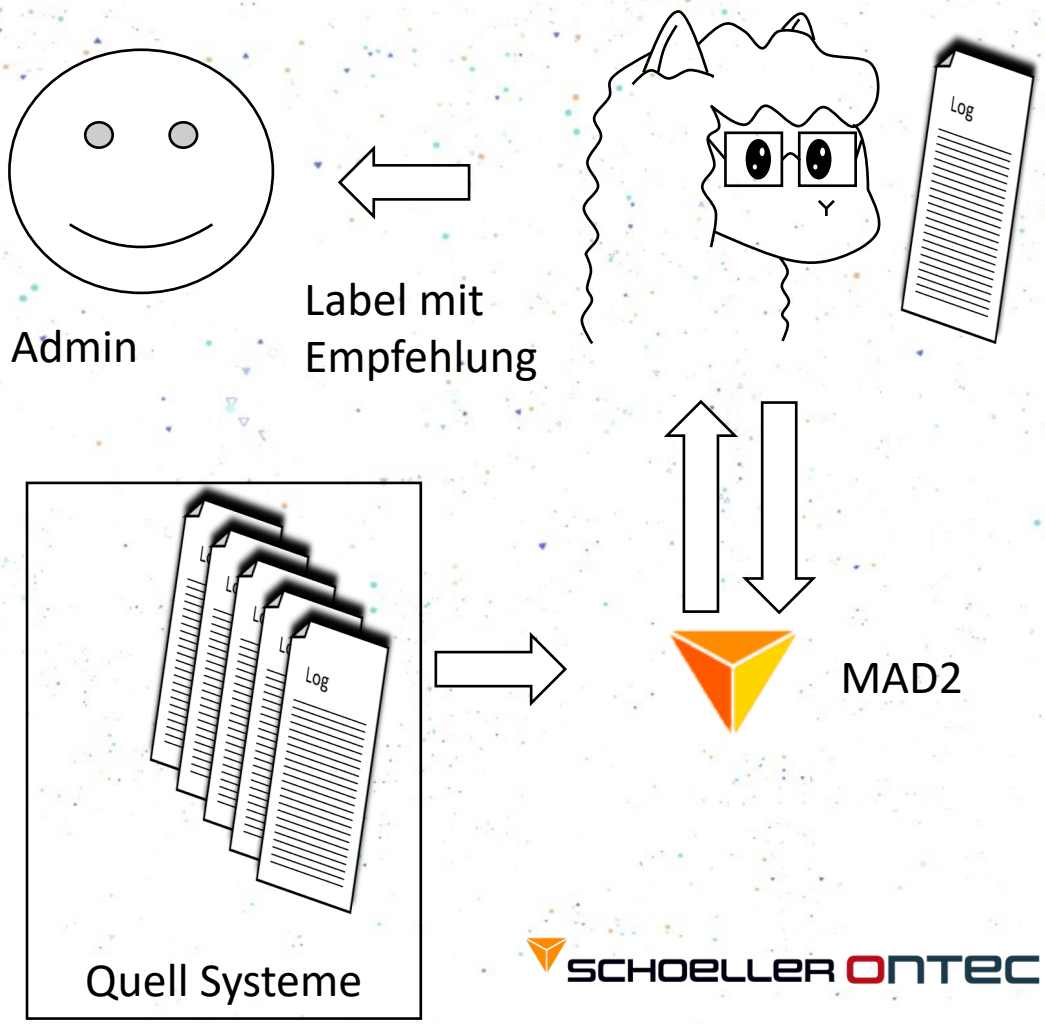
LAMaLearner und MAD2

- Das Modular Abstract Data processing Tool (MAD2)
  - Kann Log Files aus diversen Quellsystemen sammeln
  - Speichert Log Files zentral
  - Verarbeitet 10.000 Events pro Sekunde / Instanz
  - Kompressionsrate: 3% der Original Log Größe
  - Generiert Reports nach fixen Regeln
- Bestehende Schnittstelle zur Verwendung von LAMaLearner aus MAD2
- Zwei Tools zum intelligenten Log Management von SCHOELLER NETWORK CONTROL



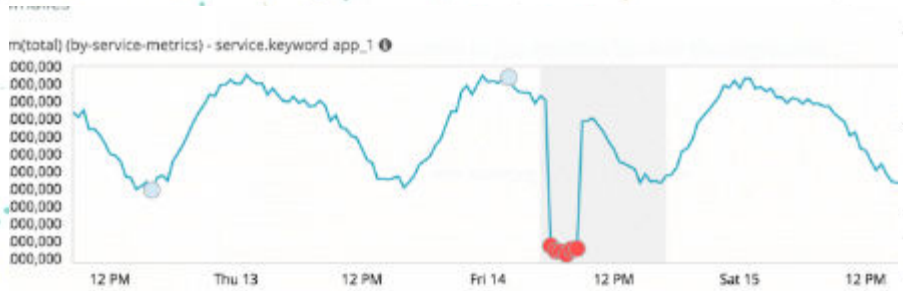
LAMaLearner und MAD2

- Geschriebene Logs können mittels MAD2 und LAMaLearner ausgewertet werden.
- Bestimmte gefundene Labels können bestimmte Handlungen implizieren.
- Diese können Administratoren vorgeschlagen werden.
- Darauf aufbauende SIEM Features
  - Visualisierung Event Label Häufigkeiten pro Zeitintervall
  - Alarmierungsregeln
  - Abweichungsberechnung (Unterschiede zu „normalen“ Tagen)



Haben Log Management Lösungen nicht schon AI?

- Existierende AI Features verarbeiten aggregierte Time-Series Daten.
  - Input: Kennlinien über Zeit
  - Output: Wo die Kennlinien nicht normal aussehen
- Das bedingt, dass man die Kennlinien hat
- Dafür muss richtiges Pattern Matching stattfinden.
- LAMaLearner ersetzt die manuelle Pattern Definition.

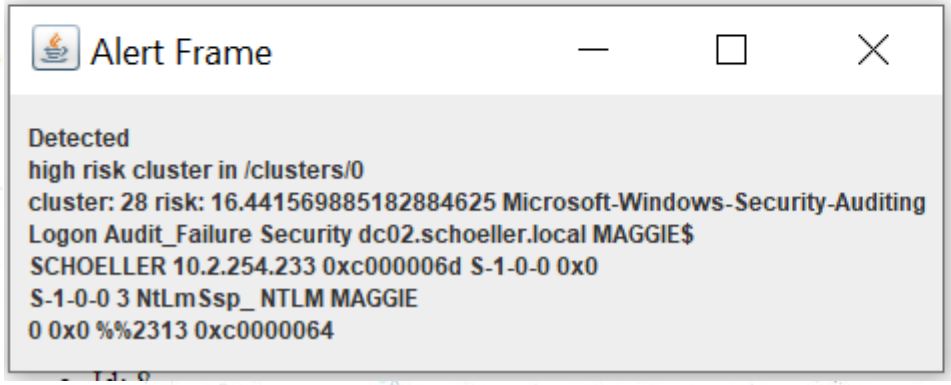


Beispiel Anwendungsfälle LAMaLearner

Alarmieren, wenn Fehler im Windows Event Log auftauchen

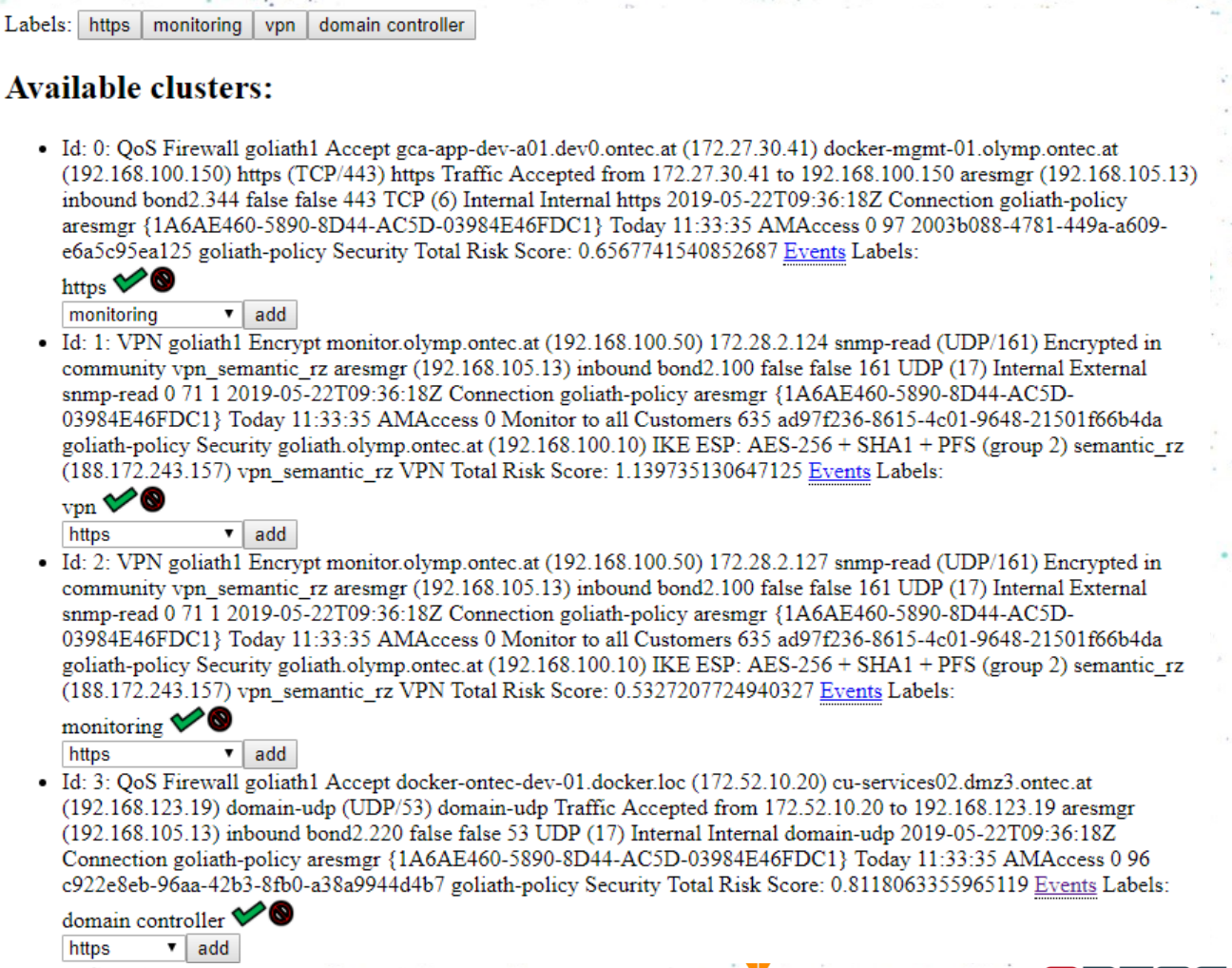
Available clusters:

- Id: 28: 4625 Microsoft-Windows-Security-Auditing Logon Audit_Failure Security dc02.schoeller.local MAGGIE$ SCHOELLER 10.2.254.233 0xc000006d S-1-0-0 0x0 S-1-0-0 3 NtLmSsp_ NTLM MAGGIE 0 0x0 %%2313 0xc0000064
  Total Risk Score: 16.44156988518288 Events Labels:
  failure
  security
  success add



Herausfinden, was über die FW läuft

- Unsupervised clustering
- Labels aus Clustern ablesen:
  - 14 Cluster statt 50 Meldungen
  - Zum Beispiel:
    - https
    - Monitoring (snmp / check_mk)
    - VPN
    - Domain Controller
  - Model für Label Identifikation



Nach Updates richtig labeln

- Version X loggt in spezifischem Format.
- Daraus kann LAMaLearner ein Modell generieren das Fehler von Normal unterscheidet.
- Version X+1 loggt plötzlich ganz anders.
- LAMaLearner Modell funktioniert immer noch, wo Parser Regeln versagen.


```
events:
  0:
    id:       0
    time:     ""
    message:  "Component router 1 with IP 192.168.13.1 is down 0"
  1:
    id:       1
    time:     ""
    message:  "Component router 1 with IP 192.168.13.1 is working correctly 7"
  2:
    id:       2
    time:     ""
    message:  "Component router 1 with IP 192.168.13.1 is >79% traffic capacity 10"
events:
  0:
    id:       0
    time:     ""
    message:  "System 192.168.13.1 (router 1) down 0"
  1:
    id:       1
    time:     ""
    message:  "System 192.168.13.1 (router 1) working correctly 7"
  2:
    id:       2
    time:     ""
    message:  "System 192.168.13.1 (router 1) >79% traffic capacity 10"
```


SCHOELLER ONTEC

Neue Logformate verstehen

Available clusters:

- Id: 0: ::1 - - "GET /server-status?auto HTTP/1.1" 200 441 "-" "Python-urllib/2.7" Total Risk Score: 0 Events Labels:
  ▾ add
- Id: 1: 127.0.0.1 - - "GET / HTTP/1.1" 403 202 "-" "Mozilla/5.0 (ISPConfig monitor)" Total Risk Score: 0 Events Labels:
  ▾ add
- Id: 2: ::1 - - "GET /server-status?auto HTTP/1.1" 200 439 "-" "Python-urllib/2.7" Total Risk Score: 0 Events Labels:
  ▾ add
- Id: 3: ::1 - - "GET /server-status?auto HTTP/1.1" 200 440 "-" "Python-urllib/2.7" Total Risk Score: 0 Events Labels:
  ▾ add
- Id: 4: ::1 - - "GET /server-status?auto HTTP/1.1" 200 437 "-" "Python-urllib/2.7" Total Risk Score: 0 Events Labels:
  ▾ add
- Id: 5: ::1 - - "GET /server-status?auto HTTP/1.1" 200 438 "-" "Python-urllib/2.7" Total Risk Score: 0 Events Labels:
  ▾ add
- Id: 6: ::1 - - "GET /server-status?auto HTTP/1.1" 200 436 "-" "Python-urllib/2.7" Total Risk Score: 0 Events Labels:
  ▾ add

SCHOELLER ONTEC

Aufgaben Mischen

- Log Typ
- Nominal & Failure

- Id: 54: 4625 Microsoft-Windows-Security-Auditing Logon Audit_Failure Security dc02.schoeller.local MG MG-NB 10.2.200.54 0xc000006d S-1-0-0 0x0 S-1-0-0 3 NtLmSsp_ NTLM MG-NB 0 0x0 %%2313 0xc000006a Total Risk Score: 0.532842053868346 Events Labels:
  Windows Security Log
  failure
  FW Log add
- Id: 55: 4769 Microsoft-Windows-Security-Auditing Kerberos_Service_Ticket_Operations Audit_Success_ Security dc02.schoeller.local bb@SCHOELLER.LOCAL SCHOELLER.LOCAL DC02$ S-1-5-21-831815799-2714527801-2835520152-1108 0x40810000 0x12 ::ffff:10.2.200.49 0x0 Total Risk Score: 0.08708915947350236 Events Labels:
  Windows Security Log
  nominal
  FW Log add

- Id: 99: 4648 Microsoft-Windows-Security-Auditing Logon Audit_Success_ Security dc02.schoeller.local tecadmin SCHOELLER 10.2.100.124 22249 {00000000-0000-0000-0000-000000000000} S-1-5-18 DC02$ SCHOELLER 0x3e7 0x200 C:\Windows\System32\lsass.exe {00000000-0000-0000-0000-000000000000} localhost localhost Total Risk Score: 0.06499146357122207 Events Labels:
  Windows Security Log
  nominal
  FW Log add
- Id: 102: 127.0.0.1 - - "GET / HTTP/1.1" 403 202 "-" "Mozilla/5.0 (ISPConfig monitor)" Total Risk Score: 0.06698409963412978 Events Labels:
  Access Log
  nominal
  Windows Security Log add

“Eigentlich Egal“ Events identifizieren

- Häufig stehen in Log Files viele irrelevante Meldungen die nicht über einzelne Felder voneinander unterschieden werden können
- LAMaLearner kann lernen relevant von irrelevant zu unterscheiden.

SCHOELLER ONTEC

Denglische Meldungen verstehen

- Für LAMaLearner ist die natürliche Sprache der Log Meldungen irrelevant.
- Solange Cluster manuell mit Labels annotiert werden, kann LAMaLearner lernen Events Labels zuzuordnen.

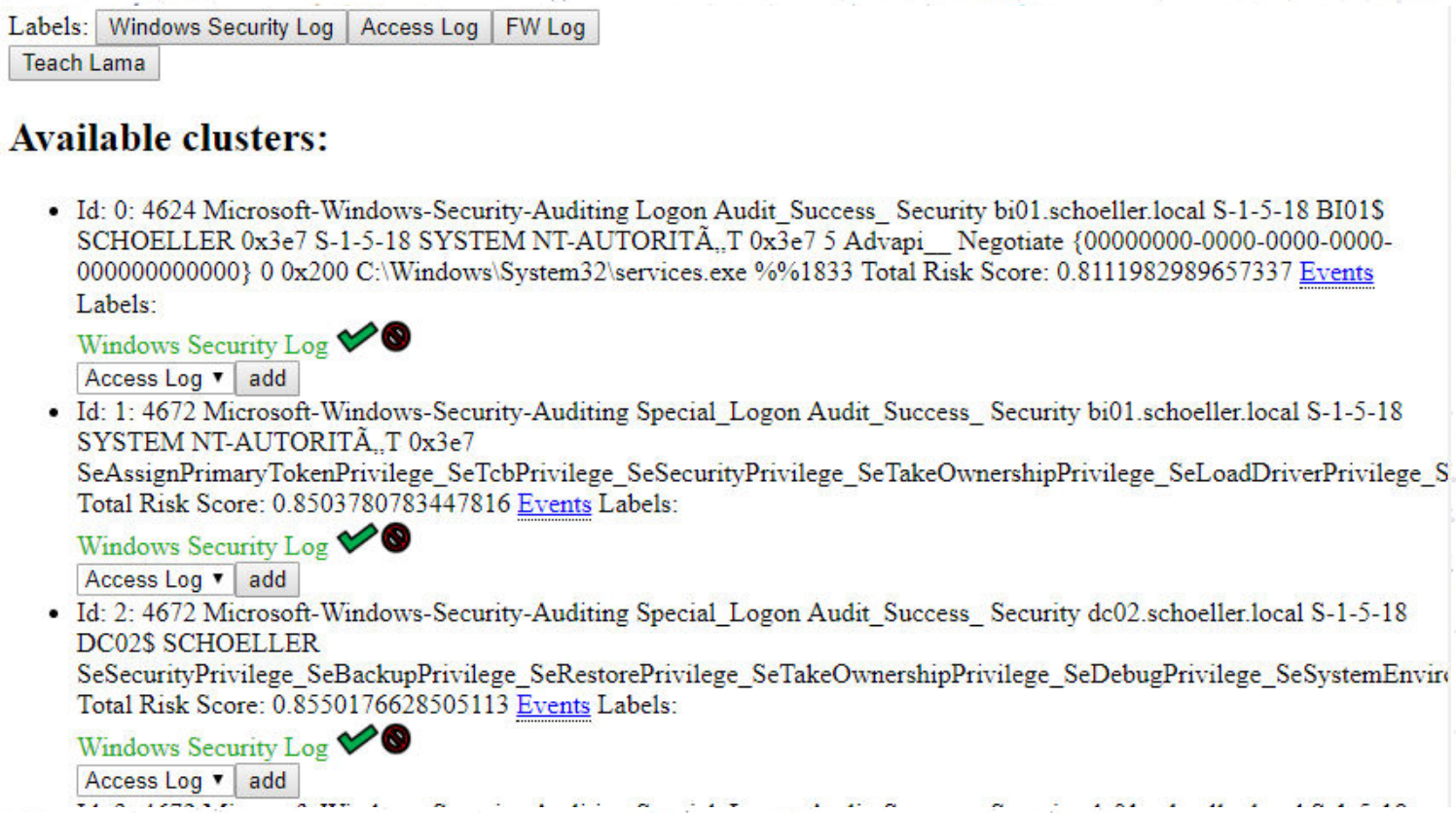
Label Typen pro Zeitintervall aggregieren

- MAD2 überträgt Events über definierte Zeiträume
- LAMaLearner annotiert Labels für den jeweiligen Zeitraum
- MAD2 erzeugt Reports über die Label Häufigkeiten pro Zeitraum.
  Z.B.:
  - Anzahl fehlgeschlagener Log Ins
  - Anzahl Firewall Rejections
  - Anzahl erkannte Malware
  - …
- Durch Machine Learning basiertes Modell können Labels auch in veränderten Log Formaten erkannt werden.
- Basierend auf erkannten Labels pro Event oder Anzahl Label pro Zeitraum können Alarmierungen durchgeführt werden
- Auf Aggregationen können Trendprognosen berechnet und Anomalien festgestellt werden.

SCHOELLER ONTEC

Log Typen auseinanderhalten

- 47 Cluster statt 18110 Events (Aufmerksamkeitsschwelle 5)
- 56 Cluster statt 18110 Events (Aufmerksamkeitsschwelle 2)
- Model für Cluster mit Aufmerksamkeitsschwelle 2:
  - 4 FP/FN
  - 52 TP
  - 93% F1



LAMaLearner

Tobias Eljasik-Swoboda

SCHOELLER NETWORK CONTROL GmbH, ONTEC AG
Ernst-Melchior-Gasse 24/DG, A-1020 Wien, Austria

Tel.: +43 1 20 55 20-0, Fax: +43 1 20 55 20-20
office@ontec.at, www.ontec.at