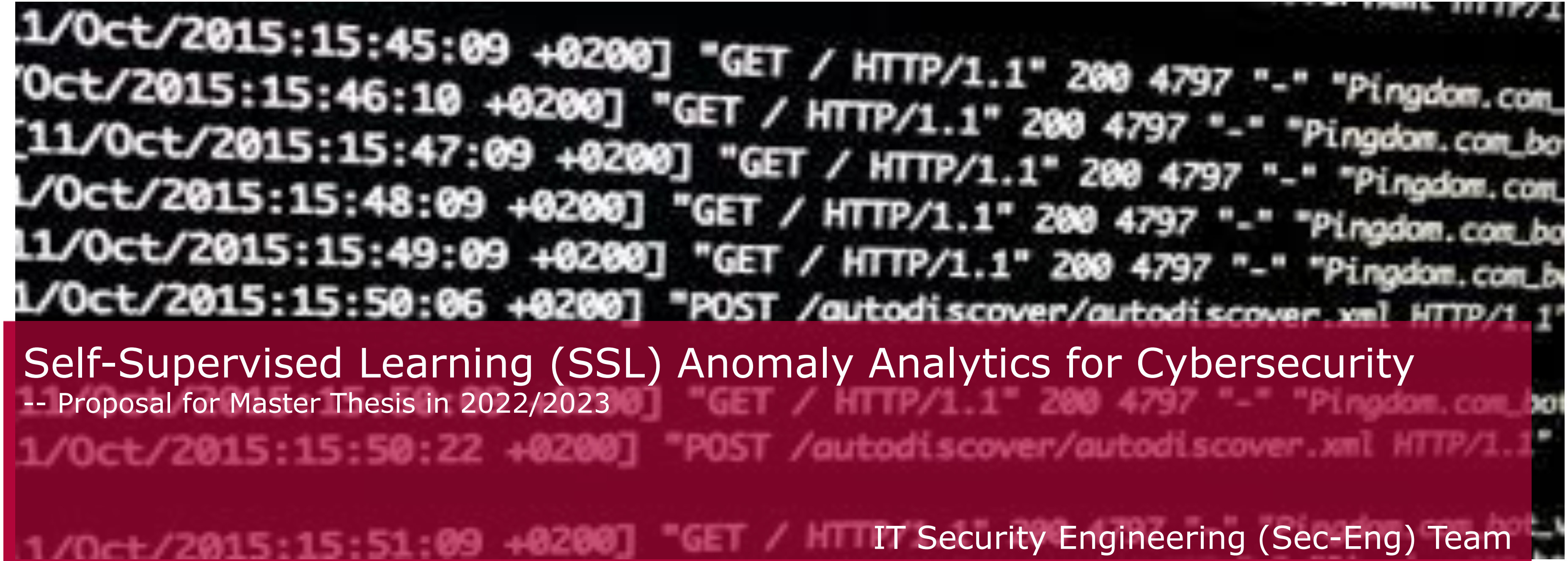


```
1/Oct/2015:15:45:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_
Oct/2015:15:46:10 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bo
[11/Oct/2015:15:47:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_
1/Oct/2015:15:48:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bo
11/Oct/2015:15:49:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_b
1/Oct/2015:15:50:06 +0200] "POST /autodiscover/autodiscover.xml HTTP/1.1"
                     +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bot
1/Oct/2015:15:50:22 +0200] "POST /autodiscover/autodiscover.xml HTTP/1.1"
1/Oct/2015:15:51:09 +0200] "GET / HTT
```

Self-Supervised Learning (SSL) Anomaly Analytics for Cybersecurity

-- Proposal for Master Thesis in 2022/2023

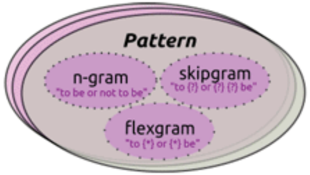
IT Security Engineering (Sec-Eng) Team
Prof. Meinel‘s Chair „Internet Systems and Technologies“

Hasso Plattner Institute, Potsdam, Germany

Motivation

- To Model as much as possible log-based Security-relevant Data using **Self-supervised Learning (SSL)** e.g., Attack log pattern, attack representational, etc.
  - Environmental (infrastructure) data: web\app-server logs, web-request\HTTP logs, ...
  - CTI/OSINT: e.g., vulnerabilities, weaknesses, attack Techniques and Tactics, IOCs, …
  - Runtime data: alerts, web-server logs, traffics, memory snapshots, process lists, ...

- To Establish effective SSL Analytics for log-based Threat Detection/Hunting implement over **web-server logs** with **NLP** annotation approach:
  - SSL-based Reasoning, log augmentation, Outlier/Anomaly Detection, Clustering logs...
  - NLP-based augmentation approach, Tokenization (N-gram, Skip-gram, Flex-gram), Bag-of-Words\ semantic embedding (word2vec, TF-IDF,…), pre-trained NLP model (cyBERT)



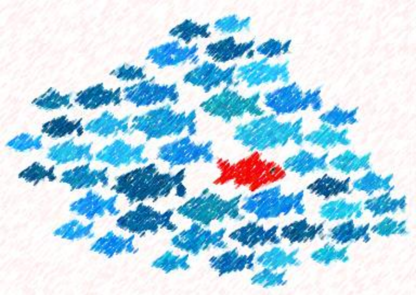
picture Credit to this article

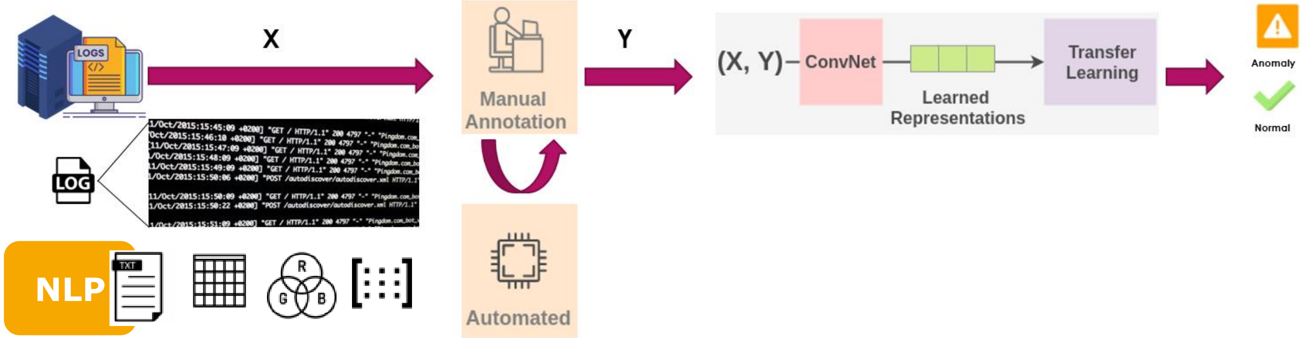
Goals

- study and evaluate the **state-of-the-art theories** and practices of SSL Modeling & Analytics

- investigate and **showcase** the feasibilities and benefits to apply SSL Modeling & Analytics in the domain of cybersecurity

- propose and conceptualize methods to **enhance existing cybersecurit solutions** (e.g., some mainstreaming SIEM systems) using **SSL-based outlier detection** for web-request **Sentiment Analysis**





```
1/Oct/2015:15:45:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_
Oct/2015:15:46:10 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bo
[11/Oct/2015:15:47:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_
/Oct/2015:15:48:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bo
11/Oct/2015:15:49:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_b
1/Oct/2015:15:50:06 +0200] "POST /autodiscover/autodiscover.xml HTTP/1.1"

11/Oct/2015:15:50:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bo
1/Oct/2015:15:50:22 +0200] "POST /autodiscover/autodiscover.xml HTTP/1.1"

1/Oct/2015:15:51:09 +0200] "GET / HTTP/1.1" 200 4797 "-" "Pingdom.com_bot_
```

Research Topic & Questions

- Topic: **Web-server log anomaly detection using SSL with NLP approach**

- Questions:
  - □ What is the best **NLP augmentation technique** to automate annotation?
    - – What is the best data structure to feed to SSL model?
    - – Evaluation of used NLP-based technique for predicted values to matching
  - □ What is efficient SSL model/design is the best for Web-server log data?
    - – SSL **representation** of web-request data
    - – Reasoning, correlations, log parsing, mining, EDA,....
    - – Performance, scalability, ...

Organization

- Requirements:
  - M.Sc. Programs: Cybersecurity, IT Systems Eng., or Data Eng.
  - (**Expected**) knowledge and experiences/skills on:
    - Network/System/Application security, IT/Security operations
    - Web-server logs, (Big) Data science and engineering, Regex patterns, Log Parsing(Templatization), NLP, .
- Deliverables:
  - Master Thesis
  - running prototype
  - Scientific publications on international conferences/journals (**expected**)
- Supervision:
  - Sec-Eng@HPI: Dr. Feng Cheng, Mehryar Majd
  - Cybersecurity/Data Engineering experts from our project partners

References

- [1] Logram: Efficient Log Parsing Using n-Gram Dictionaries
- [2] LogNG: an Online Log Parsing Method Based on N-Gram
- [3] METING: A Robust Log Parser Based on Frequent n-Gram Mining
- [4] Using NLP Techniques for Log Analysis to Recommend Activities For Troubleshooting Processes
- [5] Dynamic N-Gram Based Feature Selection for Text Classification
- [6] Log Posterior Approach in Learning Rules Generated using N-Gram based Edit distance for Keyword Search
- [7] HPM: A Hybrid Model for User’s Behavior Prediction Based on *N*-Gram Parsing and Access Logs
- [8] WhatNext: A Prediction System for Web Requests using N-gram Sequence Models
- [9] Experience Report: Log Mining using Natural Language Processing and Application to Anomaly Detection
- [10] MSc thesis: Log Server Analytics
- [11] MSc thesis: Log Classification using NLP Techniques
- [12] Towards an NLP-based log template generation algorithm for system log analysis
- [13] Toward Semi-Autonomous Information Extraction for Unstructured Maintenance Data in Root Cause Analysis
- [14] LogNG: An Online Log Parsing Method Based on N-gram
- [15]

Thank you
for your attention!

HPI IT-Security Engineering Team
Hasso-Plattner-Institut at University of Potsdam
Campus Griebnitzsee, 14482 Potsdam, Germany
Email:
Online Services: