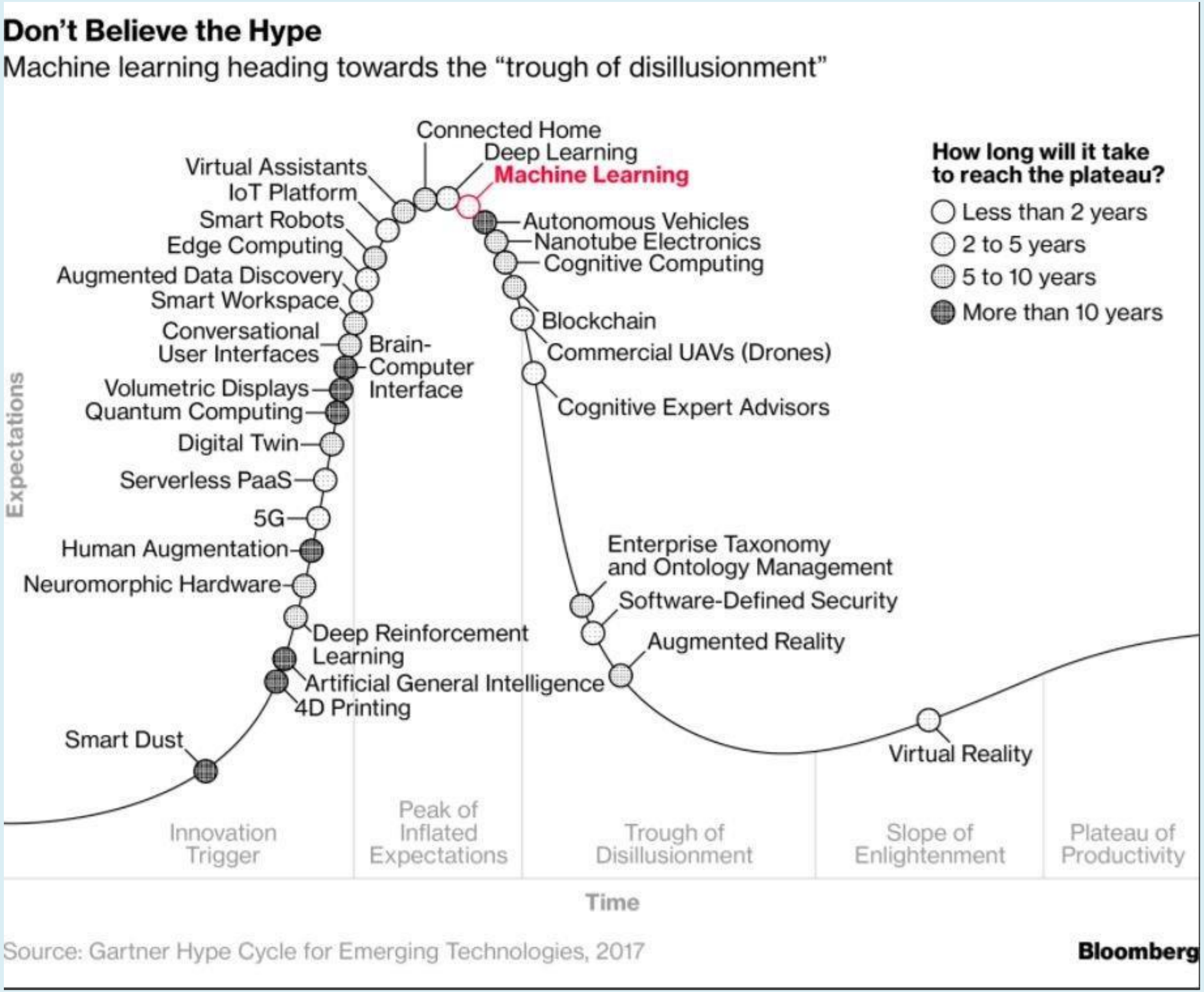
Machine Learning with Networking Data

Andreas Dewes, Katharine Jarmul –KIProtect
Andreas Lehner – DCSO GmbH (German Cyber Security Organization)

RIPE77

AI…

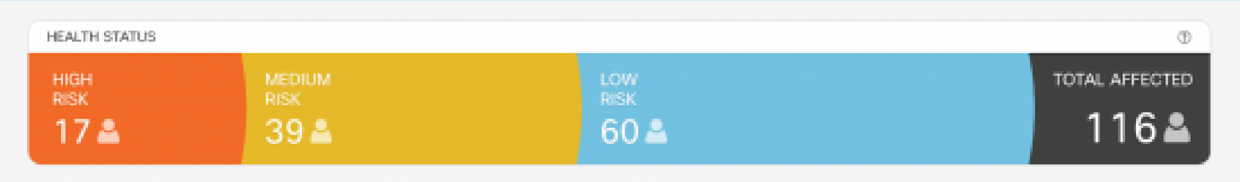
Is it Hype?



Don't Believe the Hype
Machine learning heading towards the "trough of disillusionment"
Connected Home
Deep Learning
Machine Learning
Virtual Assistants
IoT Platform
Smart Robots
Edge Computing
Augmented Data Discovery
Smart Workspace
Conversational User Interfaces
Brain-Computer Interface
Volumetric Displays
Quantum Computing
Digital Twin
Serverless PaaS
5G
Human Augmentation
Neuromorphic Hardware
Deep Reinforcement Learning
Artificial General Intelligence
4D Printing
Smart Dust
Autonomous Vehicles
Nanotube Electronics
Cognitive Computing
Blockchain
Commercial UAVs (Drones)
Cognitive Expert Advisors
Enterprise Taxonomy and Ontology Management
Software-Defined Security
Augmented Reality
Virtual Reality
How long will it take to reach the plateau?
Less than 2 years
2 to 5 years
5 to 10 years
More than 10 years
Expectations
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Source: Gartner Hype Cycle for Emerging Technologies, 2017
Bloomberg

AI…

Is it Hype?



https://blogs.cisco.com/security/closing-one-learning-loop-using-decision-forests-to-detect-advanced-threats

Why Use Machine Learning?

- Effective and adaptive pattern mining
  - “Learn” as the Data or Patterns Change
  - Scale with Your Data
- Feature-extraction
  - Network Engineer Knowledge
  - Security Research
  - Statistical Variables
- Wide Variety of Algorithms and Architectures
  - Supervised, Semisupervised and Unsupervised
  - Ability to Adapt Your Target

What Networking Problems Can ML Help?

- Network Security
  - Malicious Traffic Detection
  - Malware Identification
  - Data Loss Prevention
- Traffic Classification
  - Application Identification
  - QoS Policies
  - Traffic Engineering
- Optimization / Predictive Maintenance
- Log Analysis

Advisory Board

- Decision on DCSO strategy
- Specification of DCSO services by cross-company working groups
- Trustworthy information exchange
- Max. 30 members

Customers

- Advisory Board members
- Enterprises and their supply chain partners
- Public sector

DCSO ENGINEERING SECURITY

- Founded 11/2015 in Berlin by

Allianz · BAYER · BASF We create chemistry · VOLKSWAGEN AKTIENGESELLSCHAFT

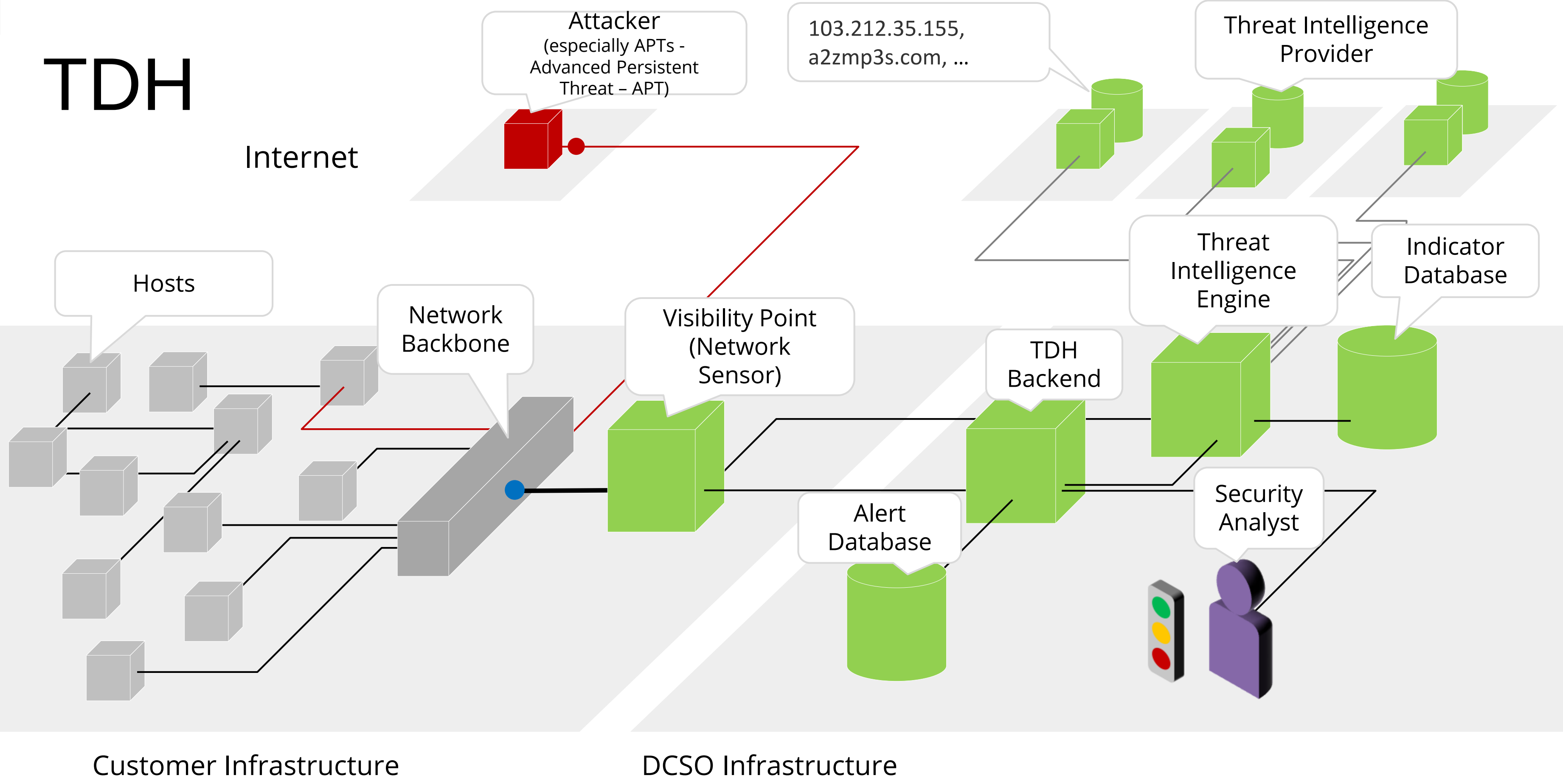
- Not aiming at profit maximization
- ~100 employees

Council

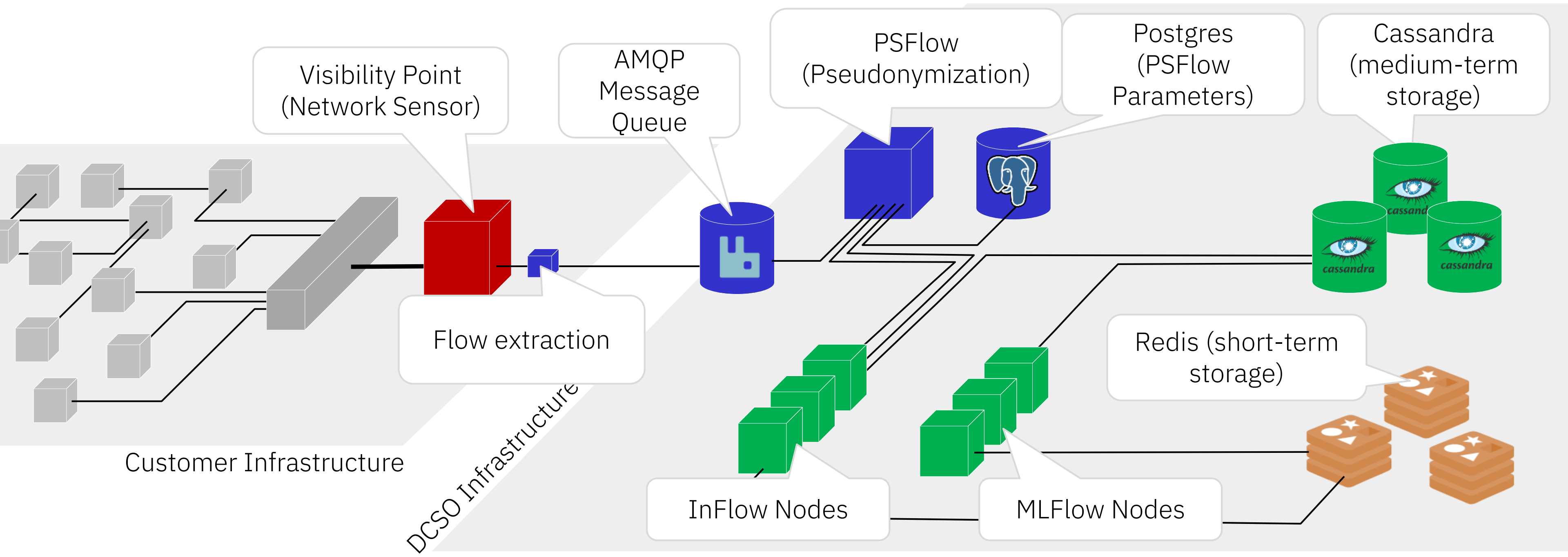
- Governance of security critical and highly sensitive information in the area of cyber security

Partners

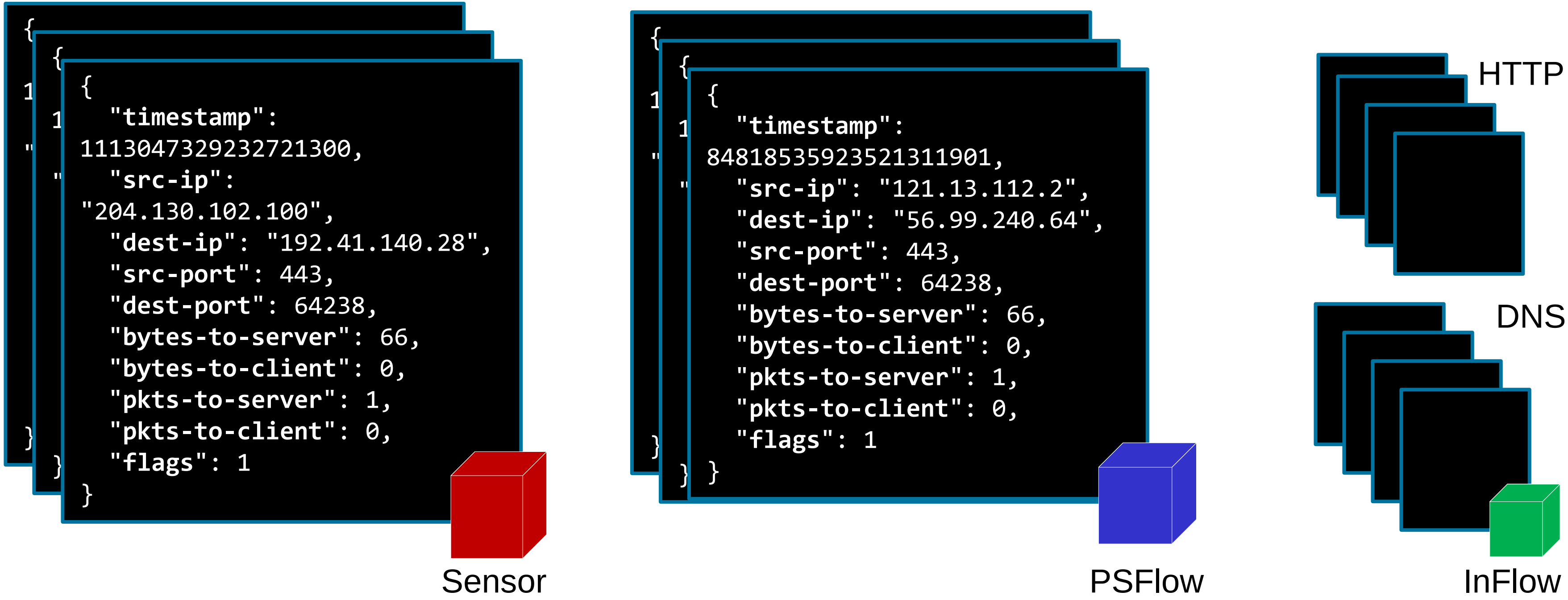
- Community
- (Research) partners
- Computer Emergency Response Teams (CERTs) groups
- NGOs


TDH
Internet
Attacker
(especially APTs -
Advanced Persistent
Threat – APT)
103.212.35.155,
a2zmp3s.com, ...
Threat Intelligence
Provider
Threat
Intelligence
Engine
Indicator
Database
Hosts
Network
Backbone
Visibility Point
(Network
Sensor)
TDH
Backend
Alert
Database
Security
Analyst
Customer Infrastructure
DCSO Infrastructure

Setup For ML-Based Flow Analysis



Feature Engineering – Part 1

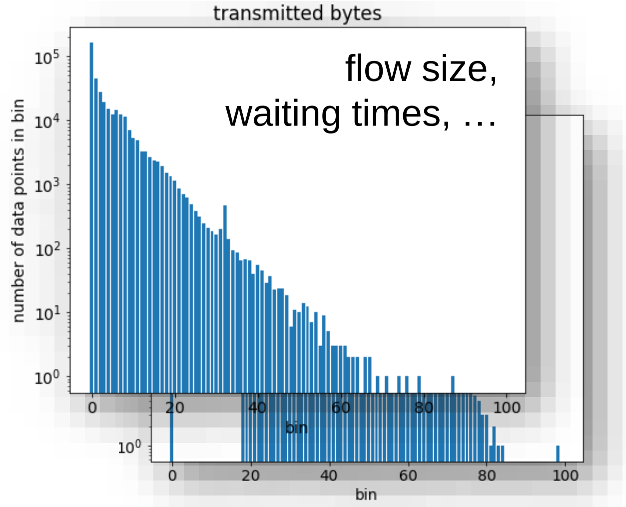


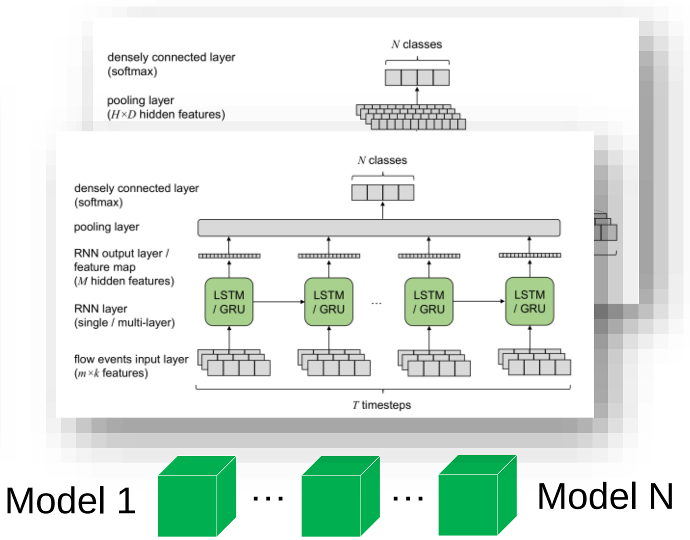
Collect flows from network sensors / endpoints

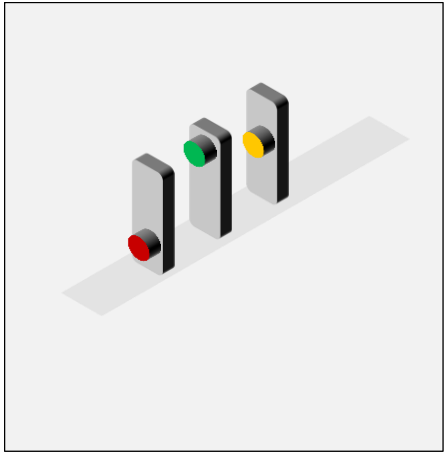
Pseudonymize/anonymize flows on the edge / gateway

Aggregate pseudonymized flows (e.g. by host, protocol, communication pairs, …)

Feature Engineering – Part 2







Convert flow sequences to appropriate features, e.g. using one-hot encoding / discretization

Train / execute on a suitable deep-learning model (e.g. for a specific protocol, malware, …)

Classify flows based on models and feed results back into IDS

Preliminary Results: Protocol Classification

Training with labeled flow data of finite length (e.g. 128 time steps).

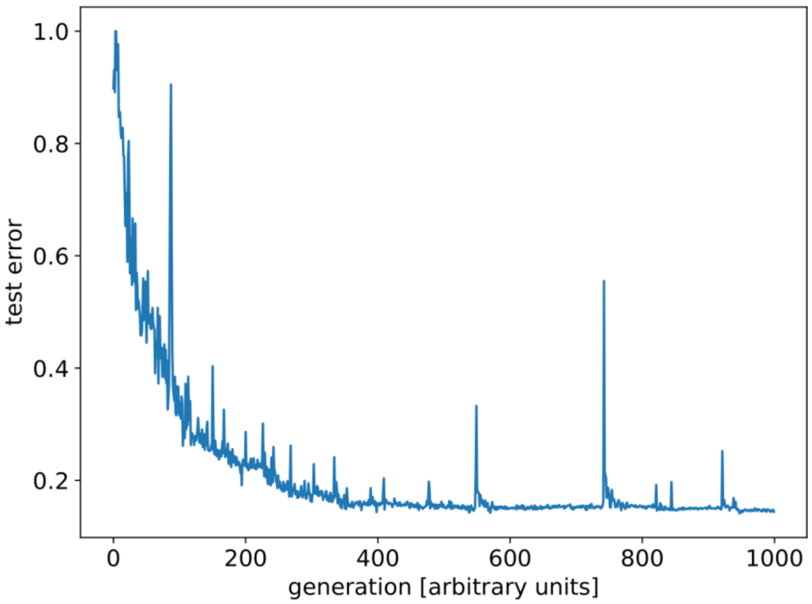
Architecture is able to learn characteristics of individual protocols. Error rate can be asymptotically reduced by averaging over time.

Comparable performance to statistics-based approaches, but more flexible.

<u>**So what?**</u>

To build real-world models, large data sets of labeled flows are necessary.

We need more & better data!

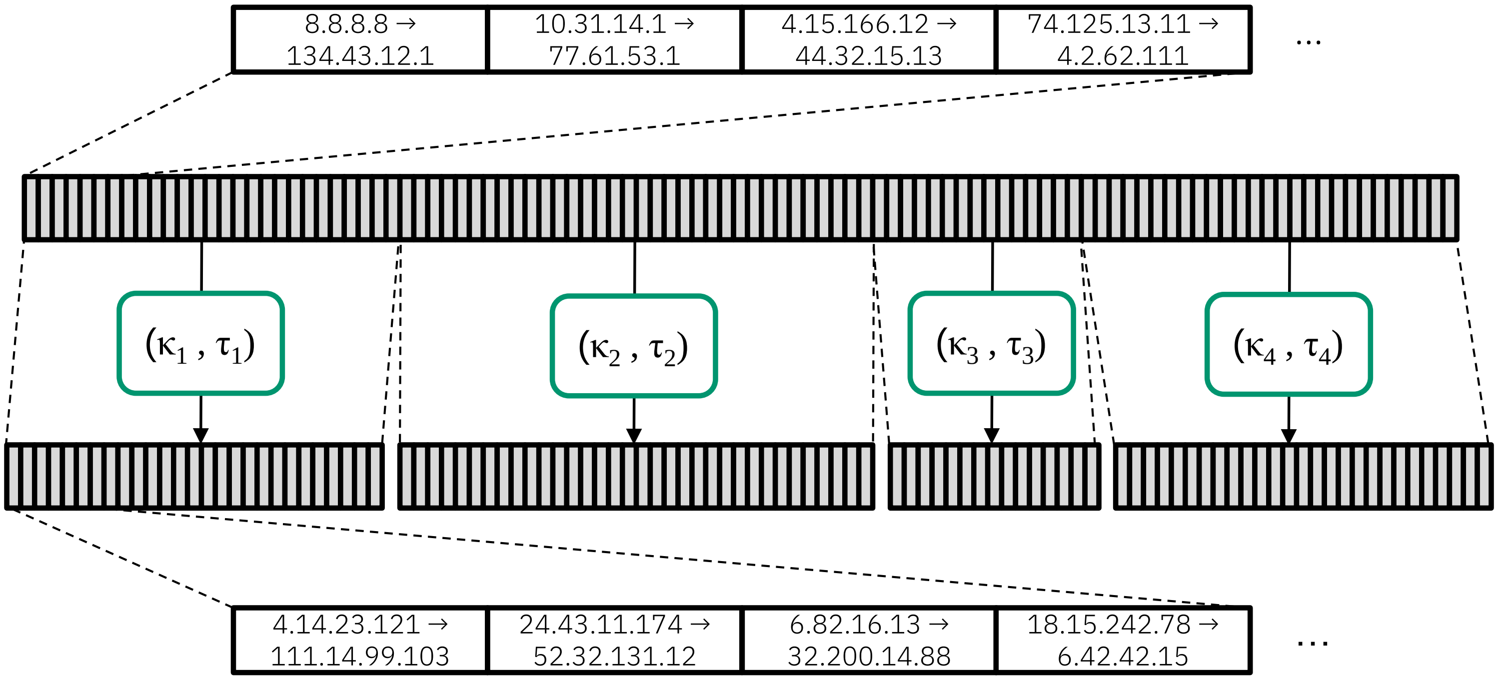


(detailed analysis & paper coming 2019)

Privacy Concerns

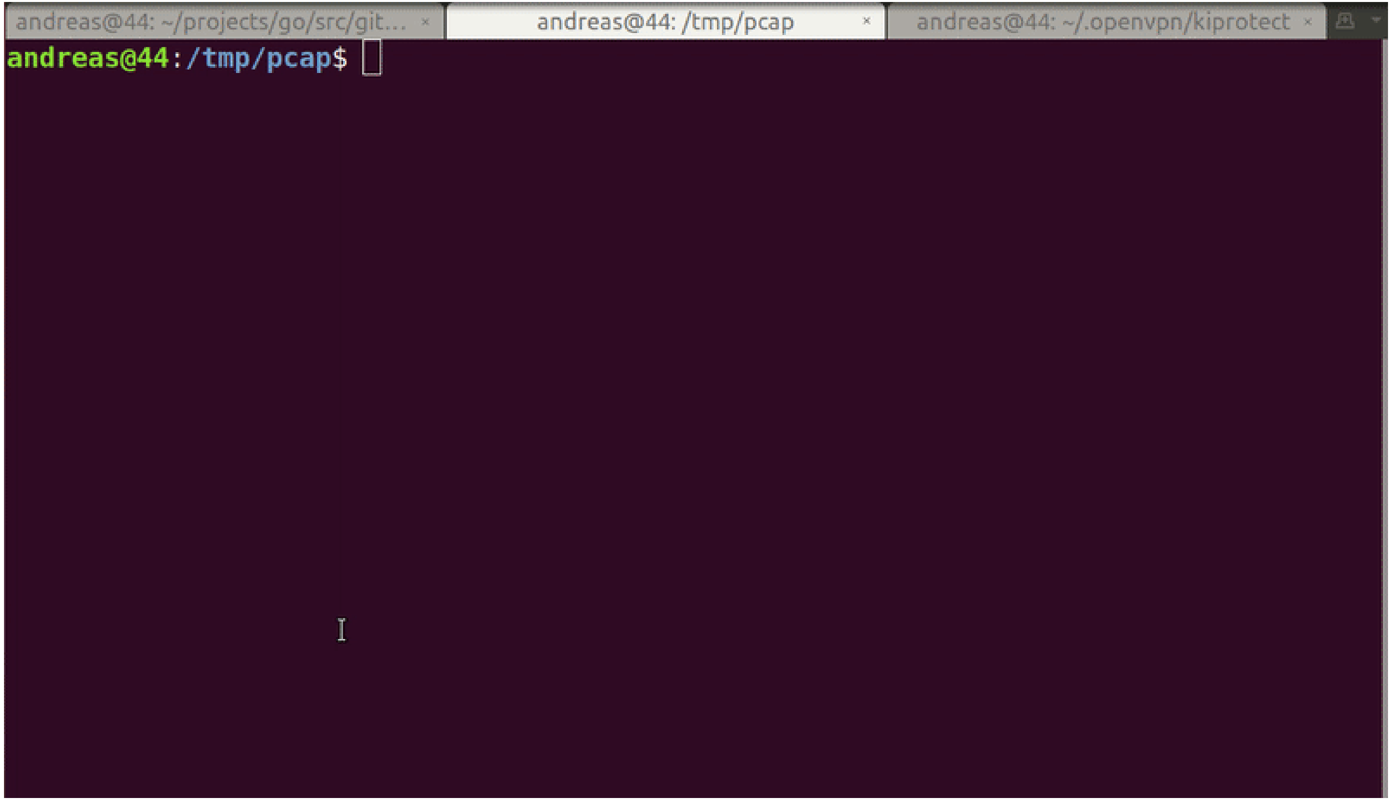
- Ability to Recover Secrets from Machine Learning Models
- Sharing with Other Networks / Providers
- Utilizing Cloud Data Analysis tools and vendors
- GDPR

Cryptographic Flow Pseudonymization



(κ,τ) – anonymized flow data

Secure PCAP Sharing



https://kiprotect.com/product/ipprotect.html

ML for Networks: Yes, We Can!

- Despite the hype, Machine Learning can help with real networking problems
- Defining your problem, determining what algorithms to use and gathering data (and, if needed, labeling the data) are required
- Pseudonymization is an effective privacy-preserving method for IP addresses, and using a structure-preserving pseudonymization allows for data utility

Thank you!

Questions? We'd Love to hear them!

Or reach out anytime:

info@kiprotect.com
@KIProtect (Twitter)
https://github.com/kiprotect

Andreas Dewes
andreas@kiprotect.com
@japh44 (Twitter)

Katharine Jarmul
katharine@kiprotect.com
@kjam (Twitter)

Andreas Lehner
andreas.lehner@dcso.de
@DCSO_de (Twitter)