Towards data-driven evaluation of intrusion detection systems

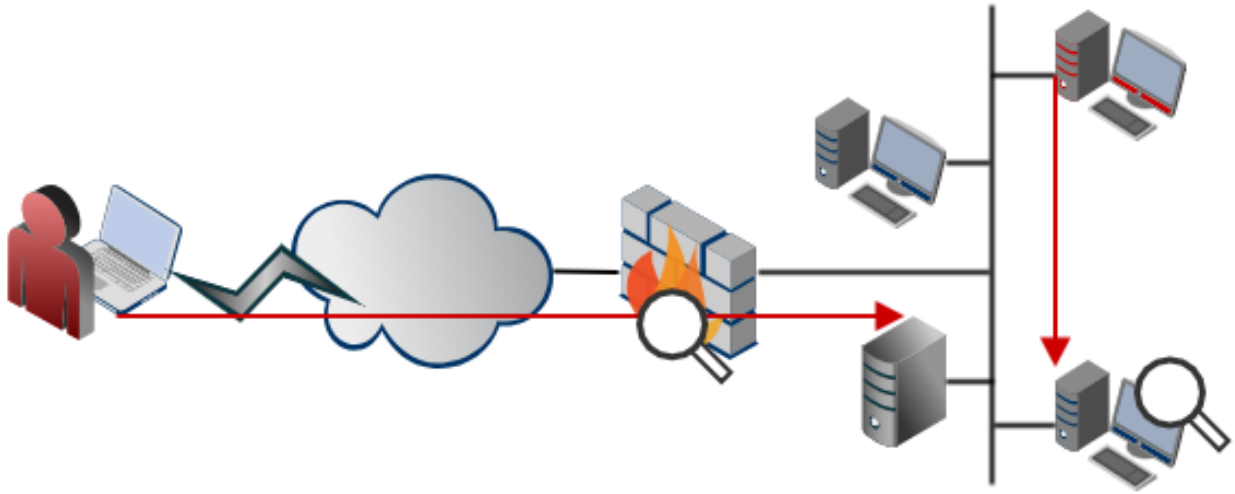
Gregory Blanc
IMT/Télécom SudParis, Institut Polytechnique de Paris

IETF 121 – NMRG session
Dublin, 2024/11/07

$whoami

- faculty at **Télécom SudParis**, an **IMT** school, member of **IP Paris**
- researcher at **SCN** *(Sécurité et Confiance Numérique)*, a team of the **SAMOVAR** lab
- associated member of **LINCS**
- head of the **SSR** *(Sécurité des Systèmes et Réseaux)* specialization curriculum
- leader of the “Evaluation” workpackage of the SuperviZ project

Intrusion Detection



Alert on any **suspicious** activity enabling later filtering or correlation

- What is suspicious?
  - **misuse**: *activity known to be malicious*
  - **anomaly**: *activity deviant from normal*
- How to capture suspicious activities?
  - at the host: process, log, file, etc.
  - in the network: flow, packet headers, payloads, etc.

Huge volume of activities incur *longer* processing time

Misuse detection

Approach	*mostly* attack signatures
Features	packet headers, flow stats, TCP connections, etc.
Trends	data mining and machine learning on labeled traffic datasets
Challenges	■ lack of datasets (existence, diversity, freshness, reliability) ■ frequency of model re-training

Multiclass classification

- Each class encodes a pattern of features, akin to a **signature**
- Model is **limited** to attack classes in the training set
- Alleviates nonetheless the **pain and risk** of manual signature design

Anomaly detection

Approach (normal) behavioural profiles

Learning unsupervised, semi-supervised, supervised

Challenges
- cleanliness of datasets
- accuracy of normal behaviour
- high false positive rate

Binary classification

- Training on **benign data only** yields patterns of normal behaviour
- The trained detection model enables detecting **deviations**
- Lacks precision as an anomaly **does not indicate** malice

Anomaly detection

Approach (normal) behavioural profiles

Learning unsupervised, semi-supervised, supervised

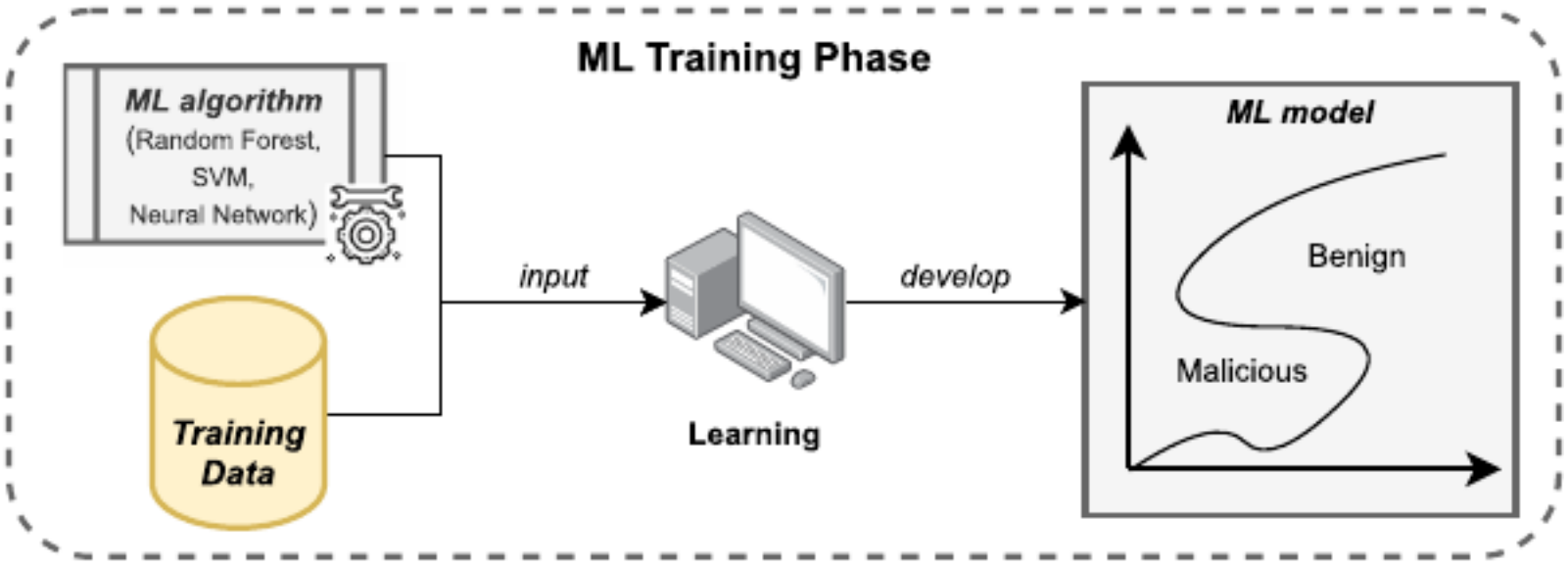
Challenges
- cleanliness of datasets
- accuracy of normal behaviour
- high false positive rate

Binary classification

- Training on **benign data only** yields patterns of normal behaviour
- The trained detection model enables detecting **deviations**
- Lacks precision as an anomaly **does not indicate** malice

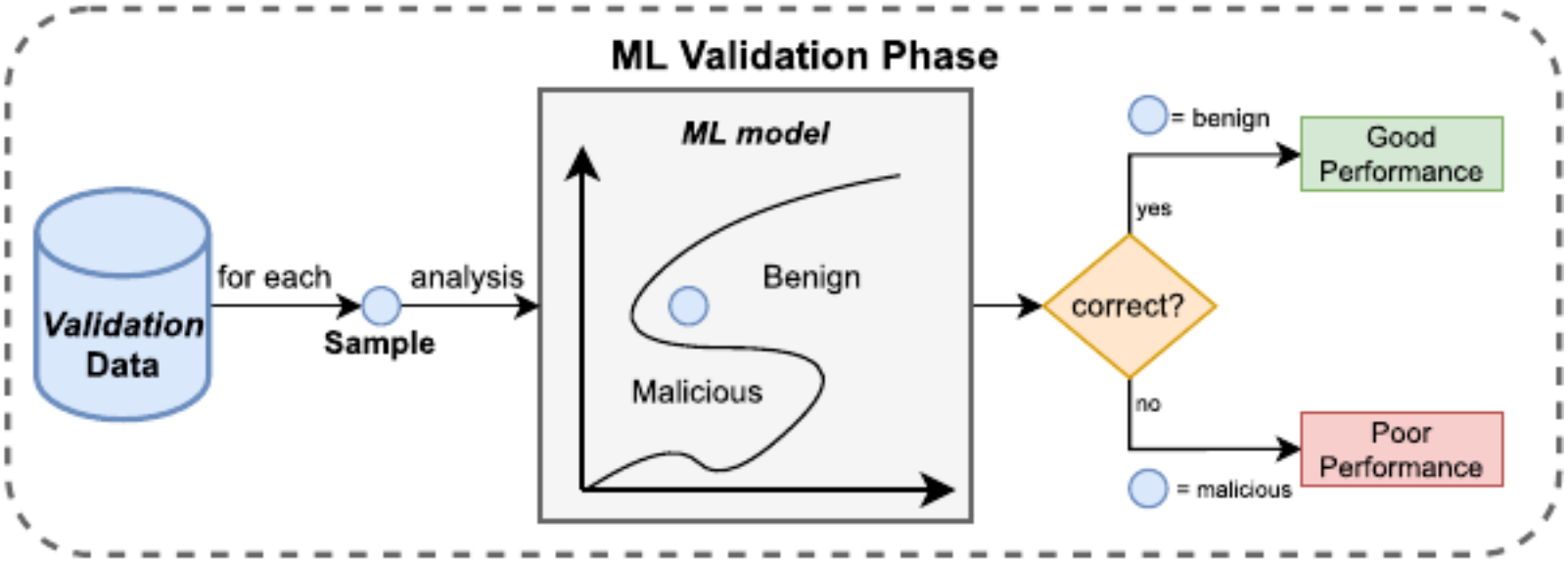
Myth: *contrary to signatures, anomaly-based detection uses ML* [1]

How to Evaluate an ML-based NIDS?



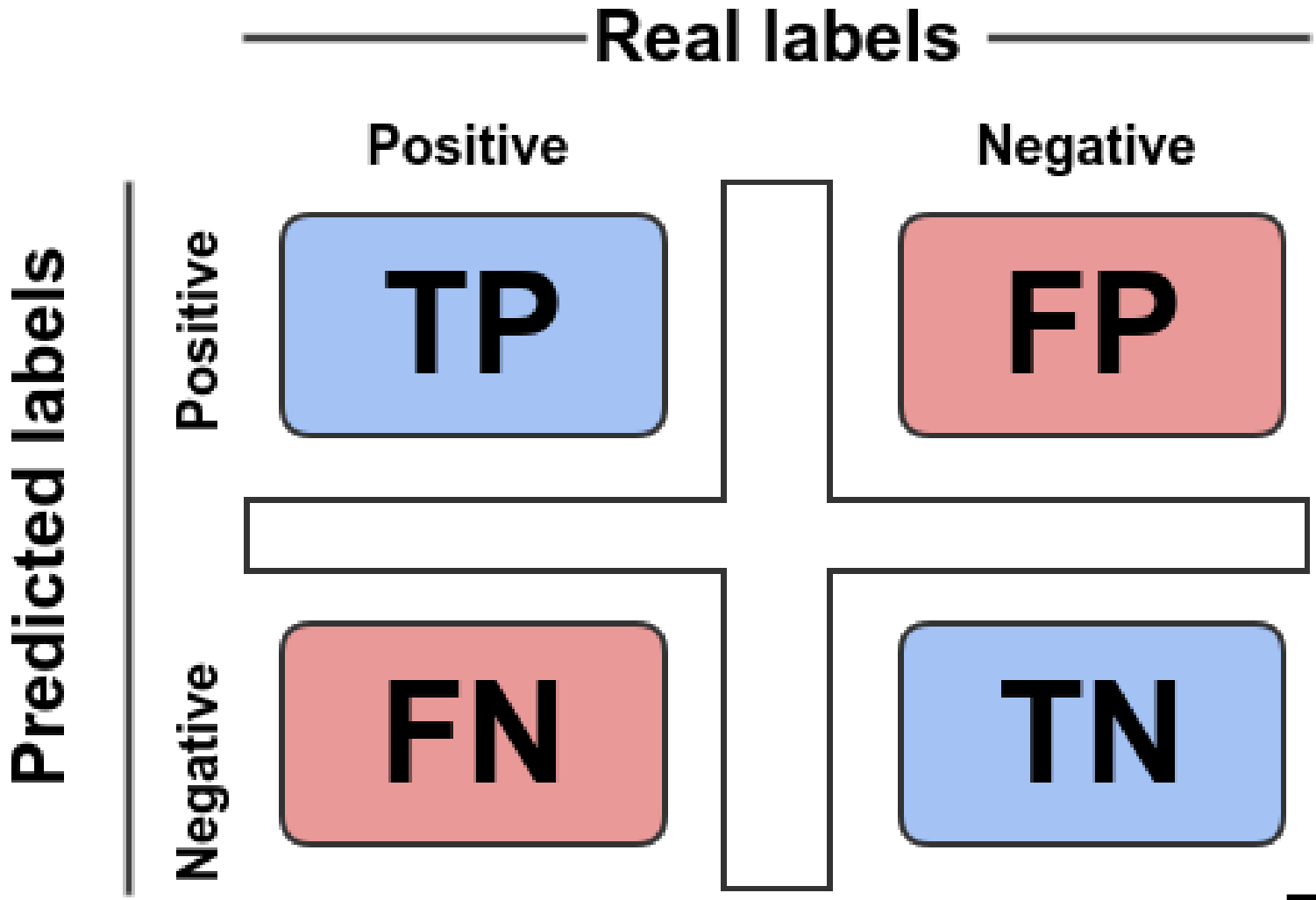
Pictures from Apruzzese et al. [1]

How to Evaluate an ML-based NIDS?



Pictures from Apruzzese et al. [1]

How to Evaluate an ML-based NIDS?



Classification Metrics [2]

Evaluating an IDS is often considered a binary classification problem. Leveraging the confusion matrix, we can measure:

- **Accuracy**: $\frac{TN+TP}{TP+FP+TN+FN)}$ (overall success rate)
- **Precision**: $\frac{TP}{TP+FP}$ (aka *positive predicted value*)
- **Detection Rate**: $\frac{TP}{TP+FN}$ (aka *sensitivity* or *recall*)
- **True Negative Rate**: $\frac{TN}{TN+FP}$ (aka *specificity*)
- **False Positive Rate**: $\frac{FP}{FP+TN} = 1 - TNR$ (aka *fall-out*)
- **F-measure**: $2 \times \frac{precision \times recall}{precision+recall}$
- **Receiver Operating Characteristic curve**: plot of the *sensitivity* as a function of $1-$ *specificity*

Evaluation Metrics

In 2015, IDS evaluation best practices measure (w.r.t. *attack detection*) [3]:

- **Attack detection accuracy**: *accuracy* of an IDS in the presence of *mixed workloads*
- **Attack coverage**: *accuracy* of an IDS in the presence of *pure malicious workloads*
- **Resistance to evasion techniques**:
  - *overlooked* in comparison to above two, as it was considered to be of limited importance from a practical perspective [4]
  - involves *pure malicious* and *mixed* workloads
- Attack detection and reporting speed: relevant for distributed IDS

Other measurements address performance properties of IDS.

Shortcomings

Most ML/DL-based IDS proposals:

- share the same set of metrics
  - **accuracy** instead of *precision* and *recall*
  - fail to use *MCC* when the dataset is **imbalanced**
- use widespread IDS datasets
  - **KDD99** has been over-used
  - many datasets suffer from **shortcut learning** [5] or labeling errors [6, 7]
- propose comparisons
  - experimental protocols differ, e.g., **tasks are different** (supervised classification vs. anomaly detection)
  - experimental settings differ, e.g., same datasets but **different splits**
  - experiments lack temporal/spatial diversity [8]

Datasets

- Packet-based: available in pcap, contains payload, metadata depending on used protocols
- Flow-based: condensed metadata-rich information, no payload, aggregates all packet sharing some properties (e.g., 5-tuple) within a time window
- Other data: hybrid data set (packet/flow, network/host)

Ring et al. [9] surveyed existing datasets and grouped them:

- public? attacks?
- metadata?
- which format
- the volume of data and its duration
- the kind of traffic and the type of network
- balanced? labeled? predefined splits?

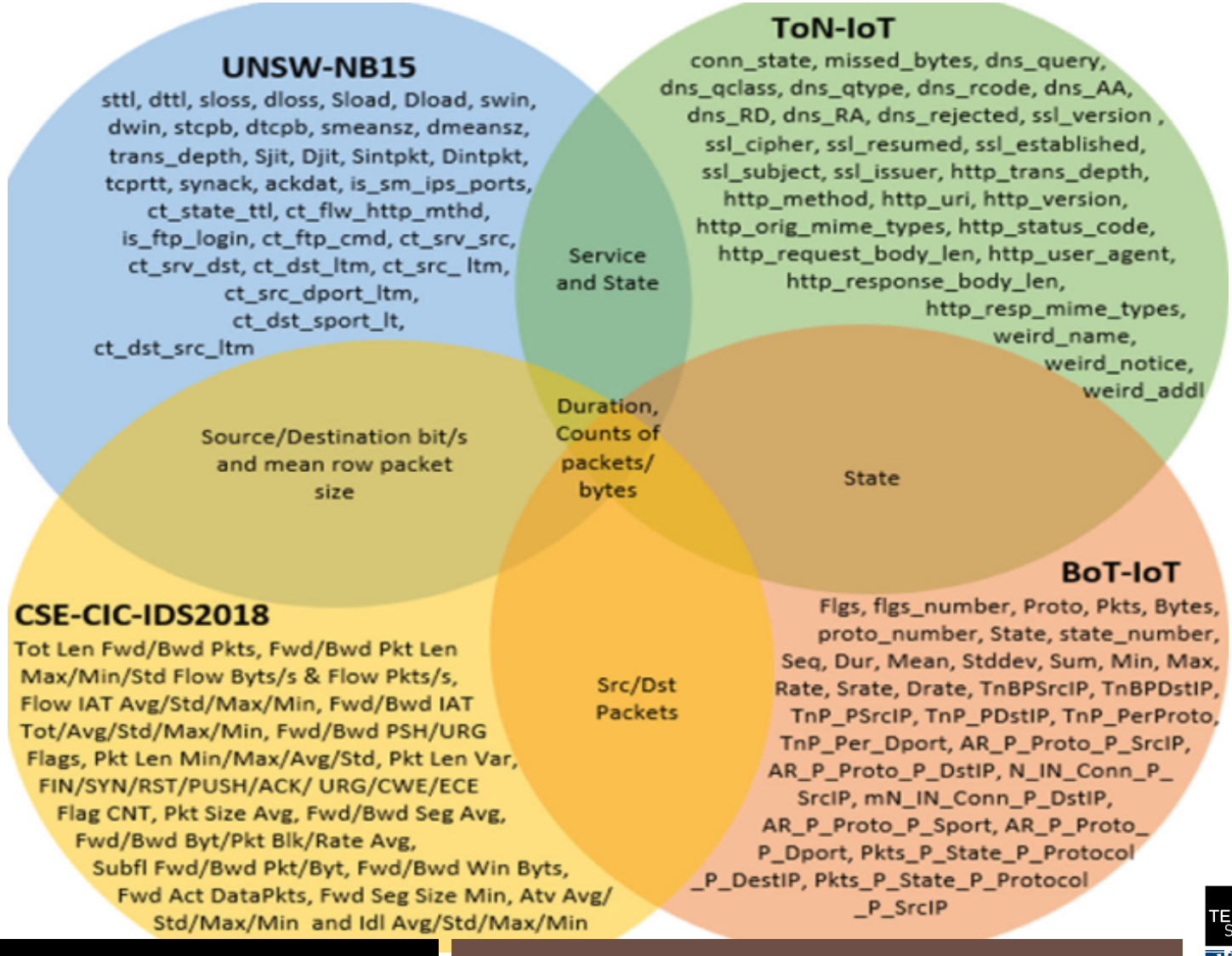
Flow Information

Flow-level datasets are very popular to briefly represent network traffic. Here is a NetFlow [10] based feature set [11].

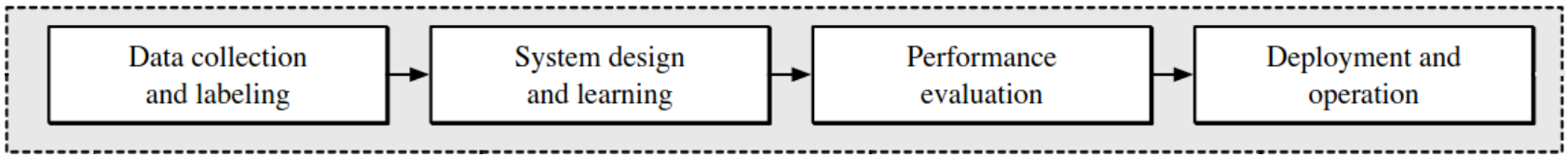
Feature	Description		
IPv4_Src_Addr	–	L7_Proto	–
IPv4_Dst_Addr	–	In_Bytes	Incoming number of bytes
L4_Src_Port	–	Out_Bytes	Outgoing number of bytes
L4_Dst_Port	–	In_Pkts	Incoming number of packets
Protocol	IP protocol identifier	Out_Pkts	Outgoing number of packets
TCP_Flags	Cumulative of all TCP flags	Flow_Duration	Flow duration in milliseconds

Other wider feature sets of dimensions 43 [12] and 83 [13] using NetFlow and CICFlow formats, respectively.

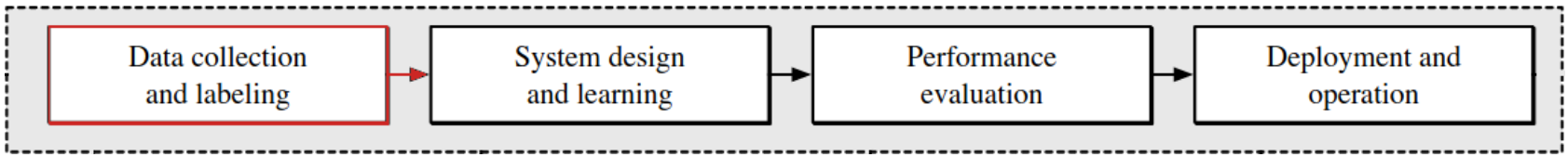
Towards a Standard Feature Set [12]



Common Pitfalls [14]



Common Pitfalls [14]



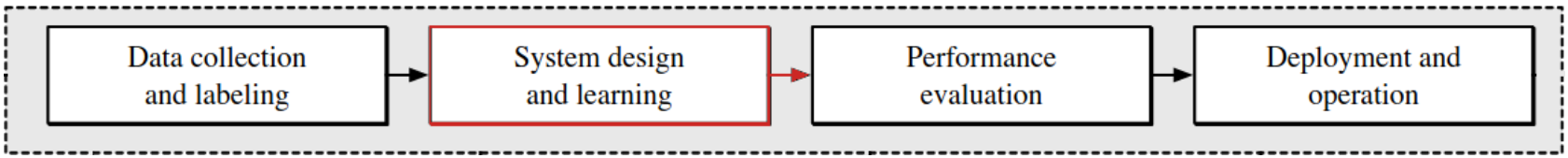
A Sampling bias

- collected data does not sufficiently represent the true data distribution of the underlying security problem

B Label inaccuracy

- labels may suffer from changes in their distribution over time
- labels should be verified manually whenever possible

Common Pitfalls [14]



C Data snooping

- clumsy data splitting yielding information that should not be available at training time

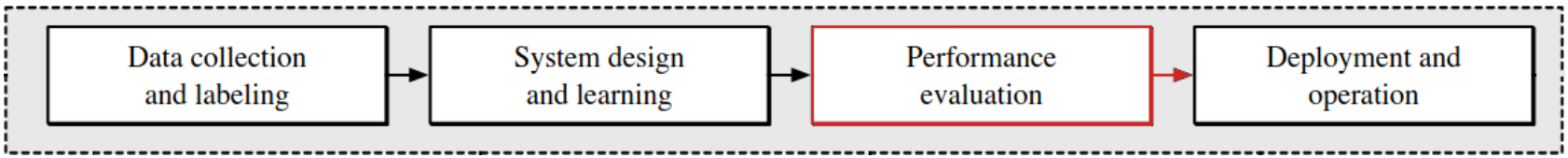
D Spurrious correlations

- artifacts that correlate with the task to solve without being related to it
- need to apply explanation techniques

E Biased parameters

- parameters indirectly depending on the test set

Common Pitfalls [14]



F Inappropriate baselines

- need for a simple baseline to motivate the need for a complex ML system

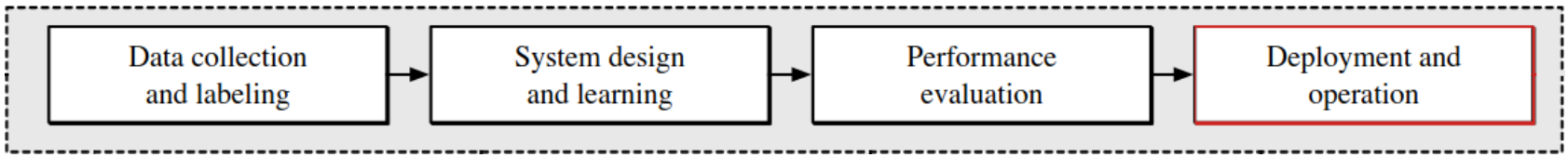
G Inappropriate measures

- evaluation should take into account the data specificities

H Base rate fallacy [15]

- ignoring class imbalance leads to performance overestimation

Common Pitfalls [14]



I Lab-only evaluation
- detection methods evaluated in a *closed world* setting [4]
- e.g., need to consider temporal and spatial relation in the data [8]

J Inappropriate threat model
- security of the detection model (*adaptive adversary* [16]) is not considered
- systematically investigate possible vulnerabilities, focusing on white-box attacks

Mislabelling in CIC-IDS2017 [6]

- CICFlowMeter issue with misordered packets
  - flow processing happens according to the order of packets in the dataset, not the timestamp
  - from 0.028 to less than 0.1% frames are misordered resulting in swapped flows
- CIC-IDS2017 contains duplicated packets (up to 13 times)
  - may be due to port mirroring misconfiguration on the testbed switch
  - more than 4.5% of the packets are duplicated per day
- Further investigation led to the discovery of labeling error
  - 10s of thousands of port scans were wrongly labeled as benign

Concept Drift [17]

Proposed NIDSs assume that the distribution of data is **stationary**. But:

- not all categories of malicious behaviour are represented uniformly across the training set
- well-established traffic features may exhibit a very gradual drift as the user habits change

Andresini et al. outline a few solutions:

- identify which characteristics change and tune the NIDS to traces exhibiting such changes
- train DNN models on historical labeled data and update them to fit unlabeled traces via transfer learning
- past models may be structurally extended to incorporate new model branches

Evaluation of ML-based NIDS: Takeaways

- Lack of a standardized evaluation approach [1]
- Datasets and metrics need to be adapted to the property to assess [18]
- Good quality (legitimate) data is lacking (mostly neglected [19])
- Data, code, hyperparameters are needed to reproduce results [1]
- Baselines are needed to demonstrate the worth of ML/DL [14]
- Comprehensive evaluation is needed in time and space, including unbalanced, non-IID or noisy scenarios

Evaluation of ML-based NIDS: Takeaways

- Lack of a standardized evaluation approach [1]
- Datasets and metrics need to be adapted to the property to assess [18]
- Good quality (legitimate) data is lacking (mostly neglected [19])
- Data, code, hyperparameters are needed to reproduce results [1]
- Baselines are needed to demonstrate the worth of ML/DL [14]
- Comprehensive evaluation is needed in time and space, including unbalanced, non-IID or noisy scenarios

Still: what does evaluation achieve?

Evaluation of ML-based NIDS: Takeaways

- Lack of a standardized evaluation approach [1]
- Datasets and metrics need to be adapted to the property to assess [18]
- Good quality (legitimate) data is lacking (mostly neglected [19])
- Data, code, hyperparameters are needed to reproduce results [1]
- Baselines are needed to demonstrate the worth of ML/DL [14]
- Comprehensive evaluation is needed in time and space, including unbalanced, non-IID or noisy scenarios

Still: what does evaluation achieve? Reproducibility?

Evaluation of ML-based NIDS: Takeaways

- Lack of a standardized evaluation approach [1]
- Datasets and metrics need to be adapted to the property to assess [18]
- Good quality (legitimate) data is lacking (mostly neglected [19])
- Data, code, hyperparameters are needed to reproduce results [1]
- Baselines are needed to demonstrate the worth of ML/DL [14]
- Comprehensive evaluation is needed in time and space, including unbalanced, non-IID or noisy scenarios

Still: what does evaluation achieve? Reproducibility? Generalization?

Evaluation of ML-based NIDS: Takeaways

- Lack of a standardized evaluation approach [1]
- Datasets and metrics need to be adapted to the property to assess [18]
- Good quality (legitimate) data is lacking (mostly neglected [19])
- Data, code, hyperparameters are needed to reproduce results [1]
- Baselines are needed to demonstrate the worth of ML/DL [14]
- Comprehensive evaluation is needed in time and space, including unbalanced, non-IID or noisy scenarios

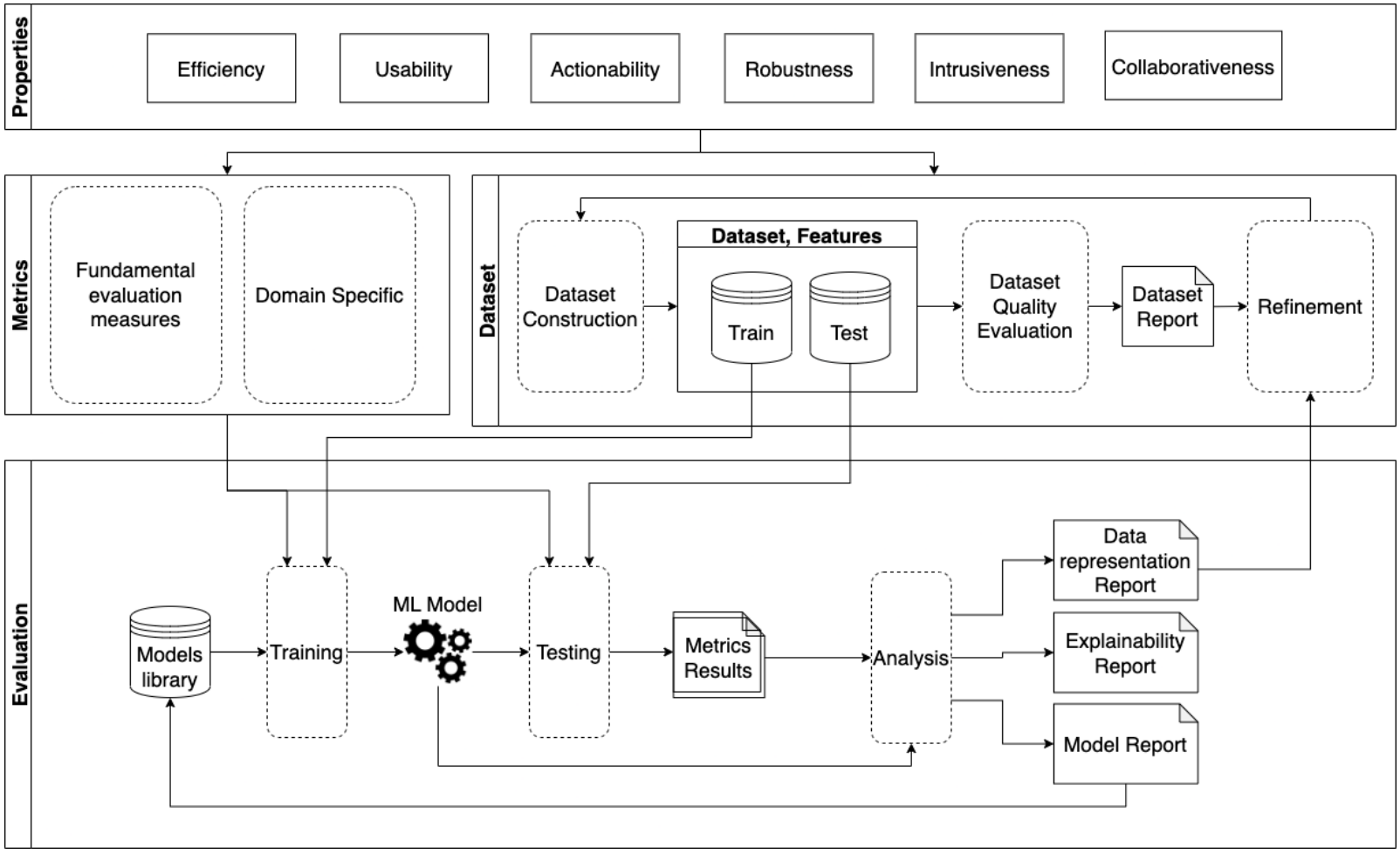
Still: what does evaluation achieve? Reproducibility? Generalization? Comparison?

Evaluation of ML-based NIDS: Takeaways

- Lack of a standardized evaluation approach [1]
- Datasets and metrics need to be adapted to the property to assess [18]
- Good quality (legitimate) data is lacking (mostly neglected [19])
- Data, code, hyperparameters are needed to reproduce results [1]
- Baselines are needed to demonstrate the worth of ML/DL [14]
- Comprehensive evaluation is needed in time and space, including unbalanced, non-IID or noisy scenarios

Still: what does evaluation achieve? Reproducibility? Generalization? Comparison? Robustness?

Framework for Data-driven NIDS Evaluation [18]



Questions and Perspectives

- Can we formalize the relation between data and detection performance?

Questions and Perspectives

- Can we formalize the relation between data and detection performance?
  - to better explain the IDS' decisions
  - needs to include all possible data manipulation in the ML's pipeline

Questions and Perspectives

- Can we formalize the relation between data and detection performance?
  - to better explain the IDS' decisions
  - needs to include all possible data manipulation in the ML's pipeline
- Can we assess the quality of a dataset universally?

Questions and Perspectives

- Can we formalize the relation between data and detection performance?
  - to better explain the IDS' decisions
  - needs to include all possible data manipulation in the ML's pipeline
- Can we assess the quality of a dataset universally?
  - qualitative and statistical metrics
  - coverage/similarity/diversity vs. representation

Questions and Perspectives

- Can we formalize the relation between data and detection performance?
  - to better explain the IDS' decisions
  - needs to include all possible data manipulation in the ML's pipeline
- Can we assess the quality of a dataset universally?
  - qualitative and statistical metrics
  - coverage/similarity/diversity vs. representation
- Can we assess the quality of a dataset with respect to detection?

Questions and Perspectives

- Can we formalize the relation between data and detection performance?
  - to better explain the IDS' decisions
  - needs to include all possible data manipulation in the ML's pipeline
- Can we assess the quality of a dataset universally?
  - qualitative and statistical metrics
  - coverage/similarity/diversity vs. representation
- Can we assess the quality of a dataset with respect to detection?
  - are state-of-the-art metrics enough?
  - needs to generate appropriate workloads for operational scenarios in volume, time and space?
  - needs to challenge the detector (adversarial examples)

Questions and Perspectives

- Can we formalize the relation between data and detection performance?
  - to better explain the IDS' decisions
  - needs to include all possible data manipulation in the ML's pipeline
- Can we assess the quality of a dataset universally?
  - qualitative and statistical metrics
  - coverage/similarity/diversity vs. representation
- Can we assess the quality of a dataset with respect to detection?
  - are state-of-the-art metrics enough?
  - needs to generate appropriate workloads for operational scenarios in volume, time and space?
  - needs to challenge the detector (adversarial examples)
- Can we realistically escape the lab?
  - needs to introduce incidents (loss, delay, noise)
  - needs interactivity?
  - needs end-to-end communication?
  - needs to generate problem-space samples?

Let's further discuss!

Dec 9th, 2024 ARTMAN '24 workshop @ACSAC 40 in Hawai
`https://artman-workshop.gitlab.io/2024/`

Jan 14th, 2025 SuperviZ seminar on IDS evaluation dataset quality @Paris (in French, mostly)

Contact info

`https://cloudgravity.github.io`

`@cloudgravity`

`gregory.blanc@telecom-sudparis.eu`

References I


G. Apruzzese, P. Laskov, E. Montes de Oca, W. Mallouli, L. Brdalo Rapa, A. V. Grammatopoulos, and F. Di Franco, "The role of machine learning in cybersecurity," *Digital Threats: Research and Practice*, vol. 4, no. 1, pp. 1–38, 2023.


N. Moustafa, J. Hu, and J. Slay, "A holistic review of network anomaly detection systems: A comprehensive survey," *Journal of Network and Computer Applications*, vol. 128, pp. 33–55, 2019.


A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating computer intrusion detection systems: A survey of common practices," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, pp. 1–41, 2015.


R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE symposium on security and privacy*, pp. 305–316, IEEE, 2010.

References II

L. D'hooge, M. Verkerken, B. Volckaert, T. Wauters, and F. De Turck, "Establishing the contaminating effect of metadata feature inclusion in machine-learned network intrusion detection models," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pp. 23–41, Springer, 2022.

M. Lanvin, P.-F. Gimenez, Y. Han, F. Majorczyk, L. Mé, and E. Totel, "Errors in the cicids2017 dataset and the significant differences in detection performances it makes," in *International Conference on Risks and Security of Internet and Systems*, pp. 18–33, Springer, 2022.

L. Liu, G. Engelen, T. Lynar, D. Essam, and W. Joosen, "Error prevalence in nids datasets: A case study on cic-ids-2017 and cse-cic-ids-2018," in *2022 IEEE Conference on Communications and Network Security (CNS)*, pp. 254–262, IEEE, 2022.

References III

F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, "{TESSERACT}: Eliminating experimental bias in malware classification across space and time," in *28th USENIX security symposium (USENIX Security 19)*, pp. 729–746, 2019.

M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers & security*, vol. 86, pp. 147–167, 2019.

B. Claise, "Cisco Systems NetFlow Services Export Version 9." RFC 3954, Oct. 2004.

M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, "Netflow datasets for machine learning-based network intrusion detection systems," in *Big Data Technologies and Applications: 10th EAI International Conference, BDTA 2020, and 13th EAI International Conference on Wireless Internet, WiCON 2020, Virtual Event, December 11, 2020, Proceedings 10*, pp. 117–135, Springer, 2021.

References IV

- M. Sarhan, S. Layeghy, and M. Portmann, “Towards a standard feature set for network intrusion detection system datasets,” *Mobile networks and applications*, pp. 1–14, 2022.
- M. Sarhan, S. Layeghy, and M. Portmann, “Evaluating standard feature sets towards increased generalisability and explainability of ml-based network intrusion detection,” *Big Data Research*, vol. 30, p. 100359, 2022.
- D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, “Dos and don’ts of machine learning in computer security,” in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 3971–3988, 2022.
- S. Axelsson, “The base-rate fallacy and the difficulty of intrusion detection,” *ACM Transactions on Information and System Security (TISSEC)*, vol. 3, no. 3, pp. 186–205, 2000.

References V

B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156, 2018.

G. Andresini, F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice, and L. Cavallaro, “Insomnia: Towards concept-drift robustness in network intrusion detection,” in *Proceedings of the 14th ACM workshop on artificial intelligence and security*, pp. 111–122, 2021.

S. Ayoubi, G. Blanc, H. Jmila, T. Silverston, and S. Tixeuil, “Data-driven evaluation of intrusion detectors: a methodological framework,” in *International Symposium on Foundations and Practice of Security*, pp. 142–157, Springer, 2022.

References VI

M. Catillo, A. Pecchia, and U. Villano, "Machine learning on public intrusion datasets: Academic hype or concrete advances in nids?," in *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)*, pp. 132–136, IEEE, 2023.