DEEP in the NET – Deep Learning for TMA

Dr. Pedro Casas

Senior Scientist

Data Science & Artificial Intelligence

AIT Austrian Institute of Technology @Vienna

10th TMA PhD School
June 27, 2022

AI4NETS


While AI has produced major breakthroughs in current data driven landscape...
…its successful application to data communication networks is still at a very early stage

Two decades of AI4NETS

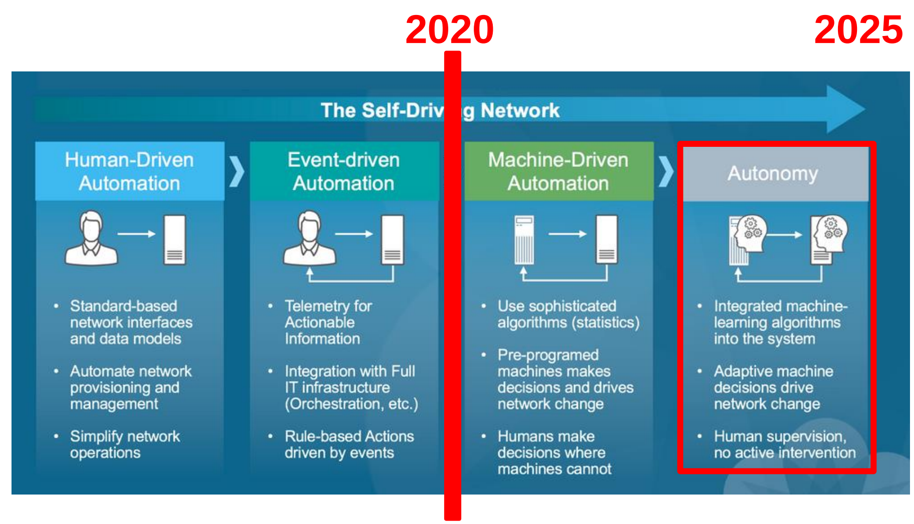
- **Cognitive Networking** (1998): networks with cognitive capabilities which could learn from past observations and behaviors, to better adapt to end-to-end requirements.

- Term **re-furbished along time**, referring to it as self-organizing networks, self-aware networks, **self-driving networks**, intelligent networks, etc.

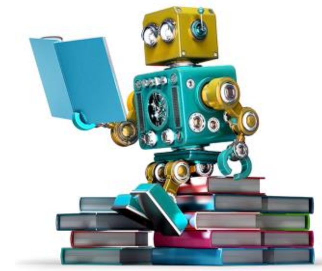
- However, there is a **striking gap** between the **extensive academic research** and **the actual deployments of such AI-based systems in operational environments**.

- **Why?** my take: there are still **many unsolved complex challenges** associated to the **analysis of Networking data through AI/ML.**

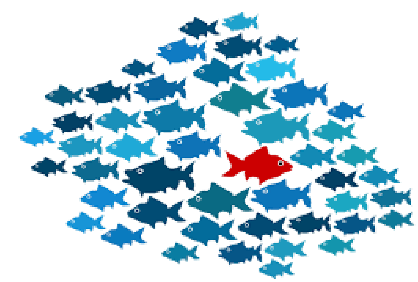
- **Hot Topic in the agenda of main Internet players:**
  - Network Operators
  - **Network Vendors (self-driving networks)**
  - Content Providers: the **Internet business of end-user engagement**

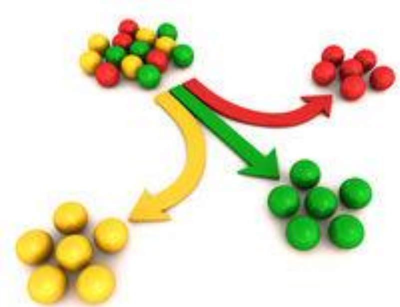


Machine Learning for Network Analytics Use Cases – What do I do @AIT?



- Network Traffic Monitoring & Analysis
- End-User Experience (Internet-QoE)
- Cybersecurity & Anomaly Detection
- Network Performance Forecasting

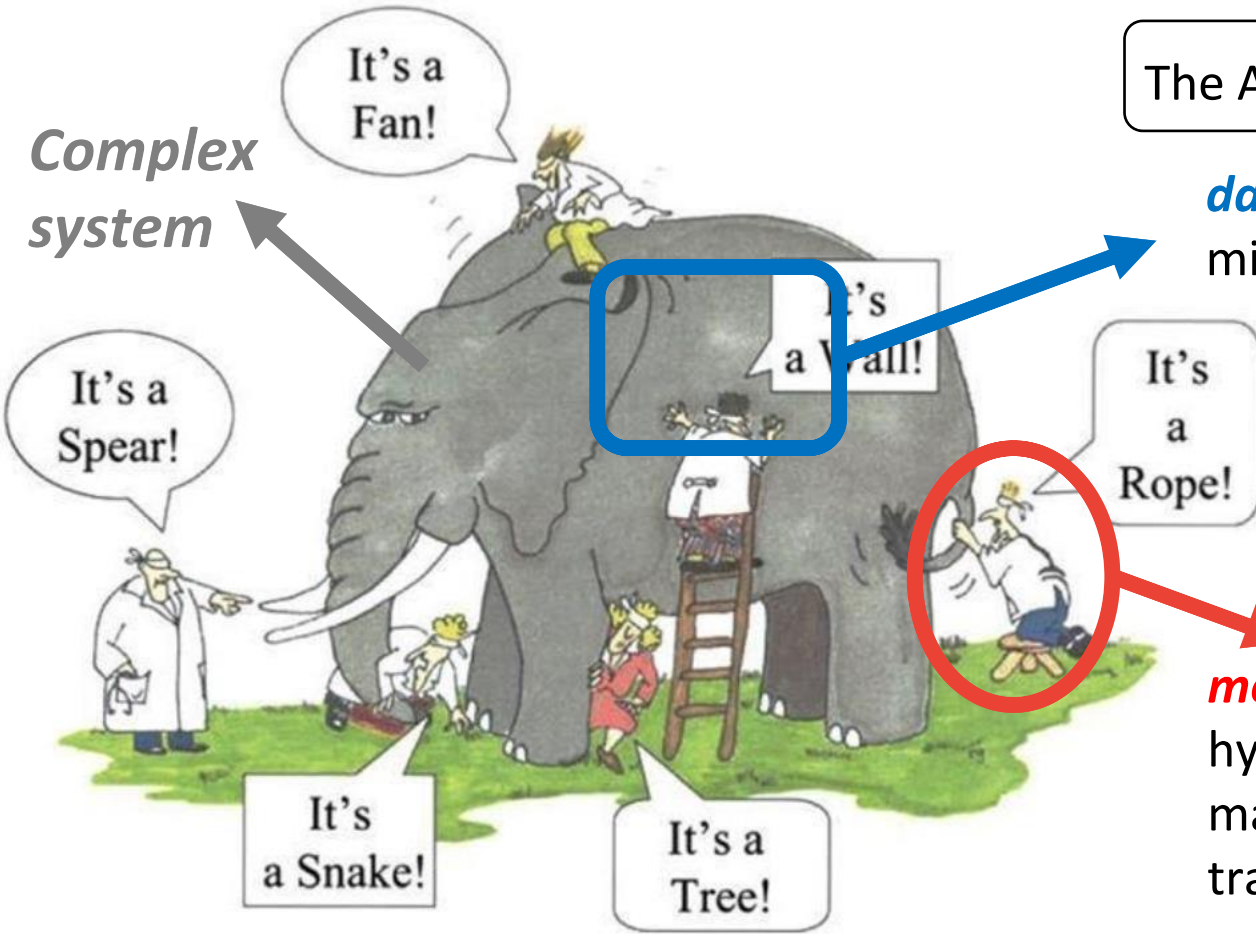




AI/ML and a Data Driven Approach to Science

- ***Big Data*** + Machine ***(Deep) Learning*** = **data driven** analysis of complex systems

Complex system



The Achilles heel of ML/AI: ***BIAS***

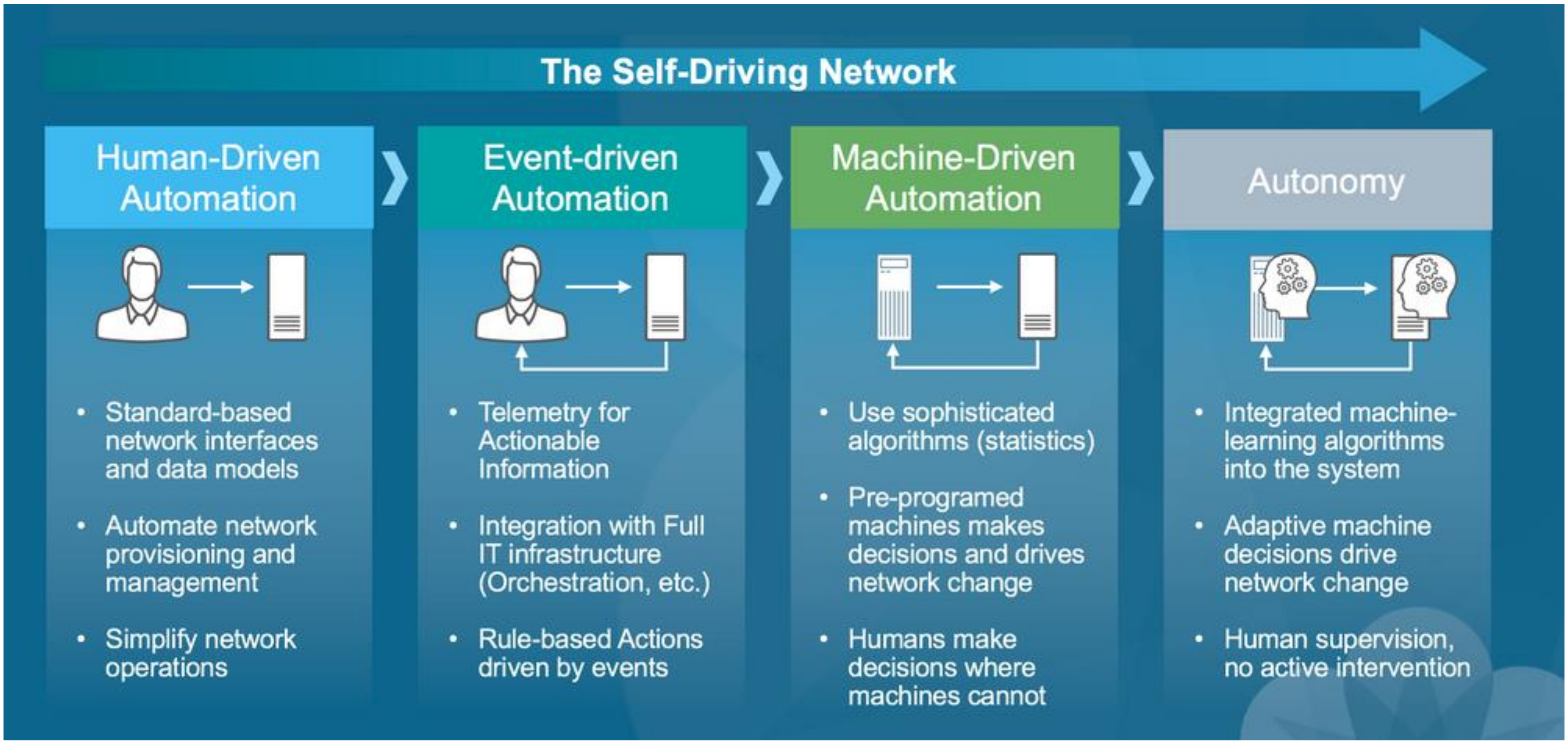
data is biased: partial or misrepresentation of real system

models are biased: assumptions or hypotheses of behavior, mathematical properties, lack of transparency

- The AI/ML model ***user is biased***, or unaware of the limitations of AI/ML: model **evaluation/testing**, model ***certification***, **correlation vs causality**

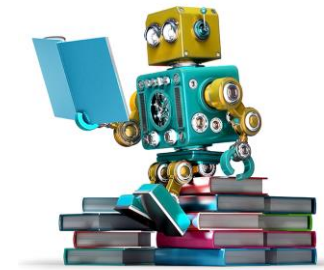
What is Blocking AI Success in Networking?

- ***Lack of Learning Generalization:*** it becomes ***extremely difficult*** in the networking practice to learn ***models which can generalize to operational environments***



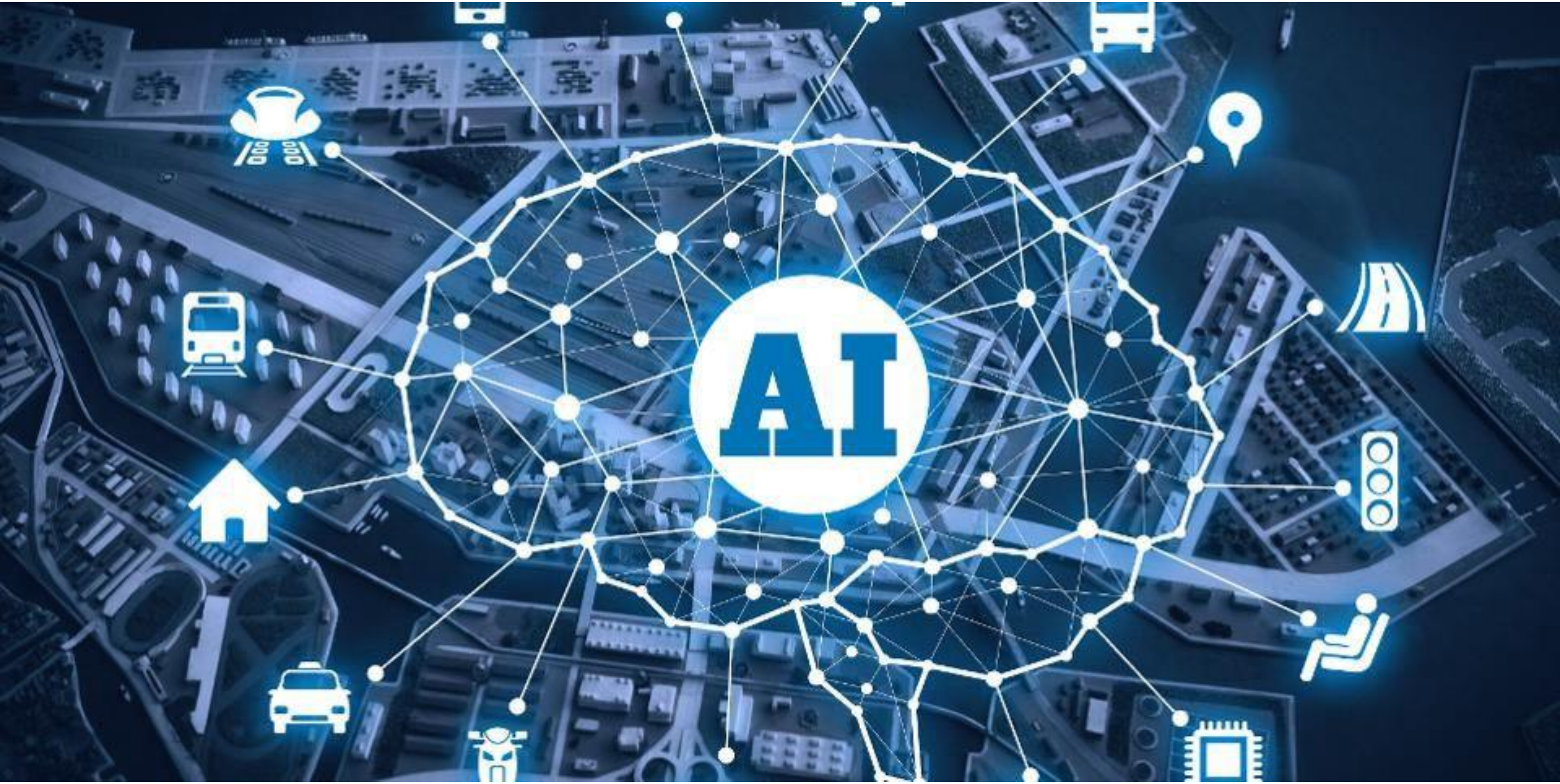
Organization of the Talk

Dealing with Some of the Challenges in AI4NETS



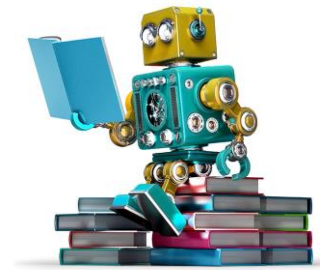
- Deep Learning for Malware Detection – ***Avoid Feature Engineering***
- Generative Models for Anomaly Detection – ***Avoid Traffic Modeling***
- Explainable Artificial Intelligence (XAI) – ***Interpret Model Decisions***
- Super Learning for Network Security – ***Avoid Model Decision***
- Adaptive/Stream Learning for NetSec – ***Deal with Concept Drifts***
- Reinforced Active Learning – ***Deal with Lack of Ground Truth***

Let's take a step back and set a common ground ☺



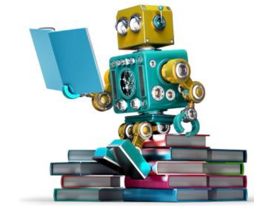
Organization of the Talk

Dealing with Some of the Challenges in AI4NETS



- **Deep Learning for Malware Detection** – ***Avoid Feature Engineering***
- Generative Models for Anomaly Detection – ***Avoid Traffic Modeling***
- Explainable Artificial Intelligence (XAI) **–** ***Interpret Model Decisions***
- Super Learning for Network Security – ***Avoid Model Decision***
- Adaptive/Stream Learning for NetSec – ***Deal with Concept Drifts***
- Reinforced Active Learning – ***Deal with Lack of Ground Truth***

Artificial Intelligence – As Smart as a Donut!

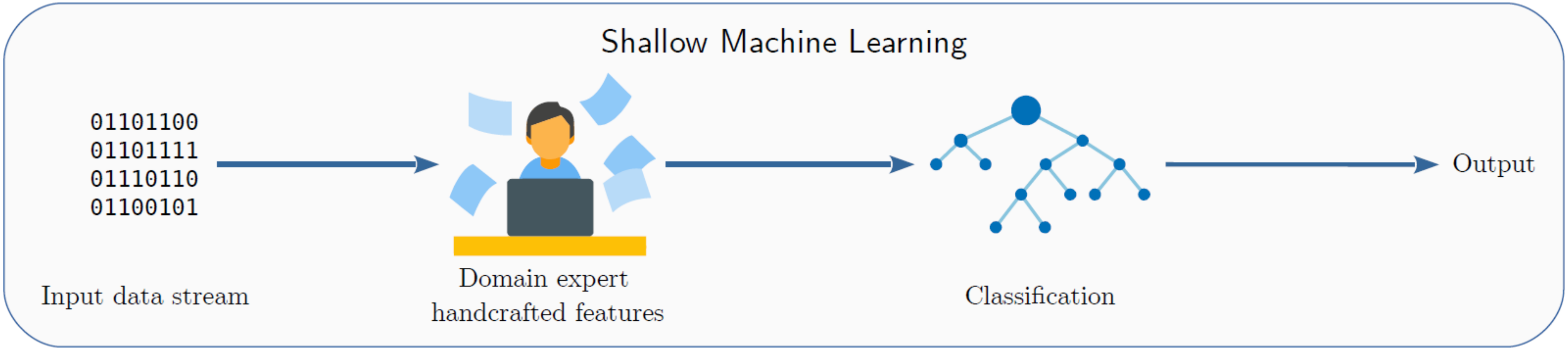


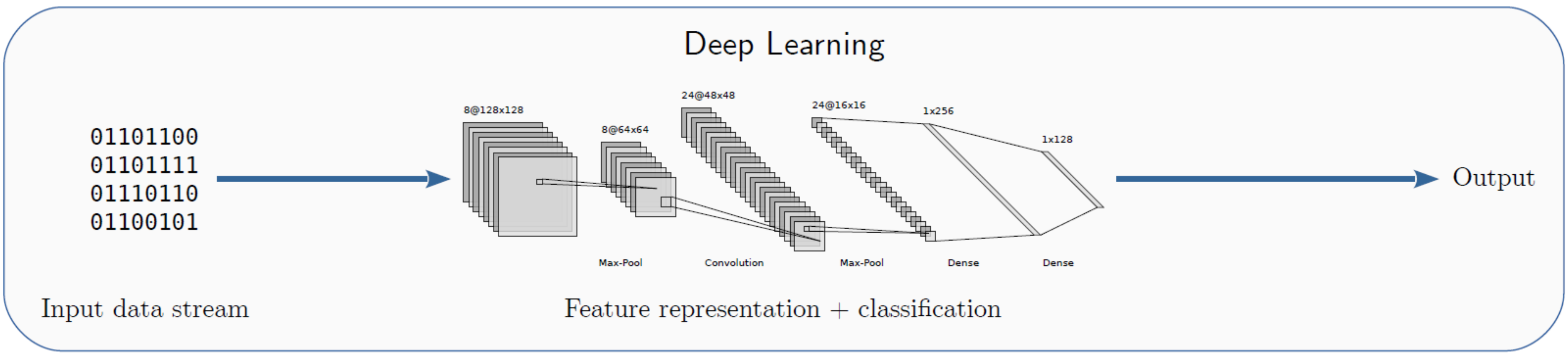
- **Machine Learning** is still **very stupid** – the **big revolution** is on **big data processing** and **data availability**/accessibility
- **Current ML benefits** are fundamentally due to machines ability to **blindly**:
  - compute lots of math operations per second
  - handle large amounts of data
  - deal with data in high-dimensional spaces
- **A lot of data required** to “learn“ simple logical inter-relations
- **Shallow Learning**: less data but ***human expert knowledge required***, to properly guide the ***feature engineering*** process
- **Deep Learning**: ***automated feature engineering*** (representation learning) but ***needs much more data***

RawPower

we explore **deep learning** for ***blind*** malware detection in network traffic

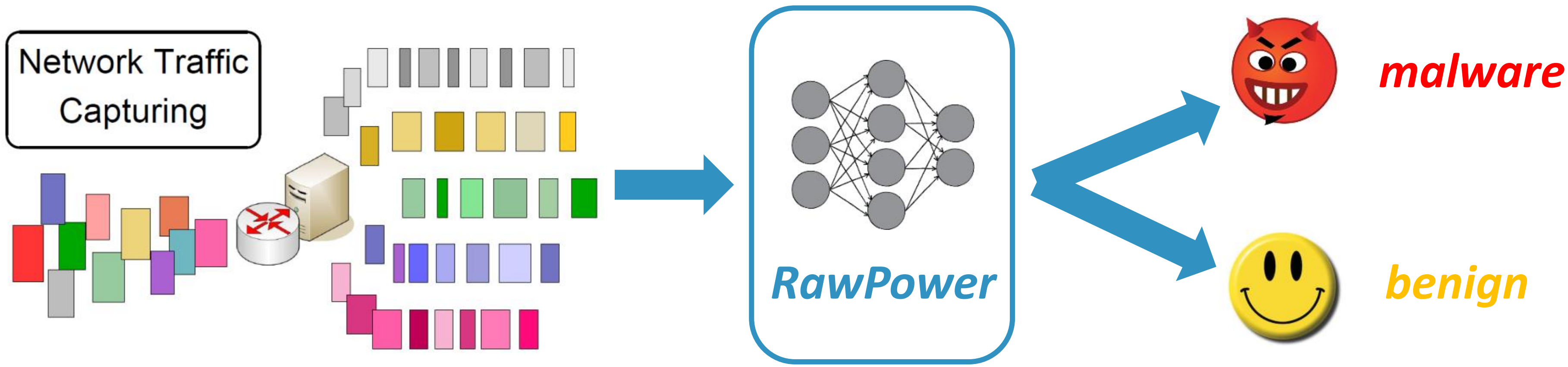
Shallow Learning vs Deep Learning





Basic Concepts of RawPower

- The ***input to the Deep Learning*** model is **RAW** – ***only byte-streams***
- **No need to define** tailored, domain-knowledge-based **input features**



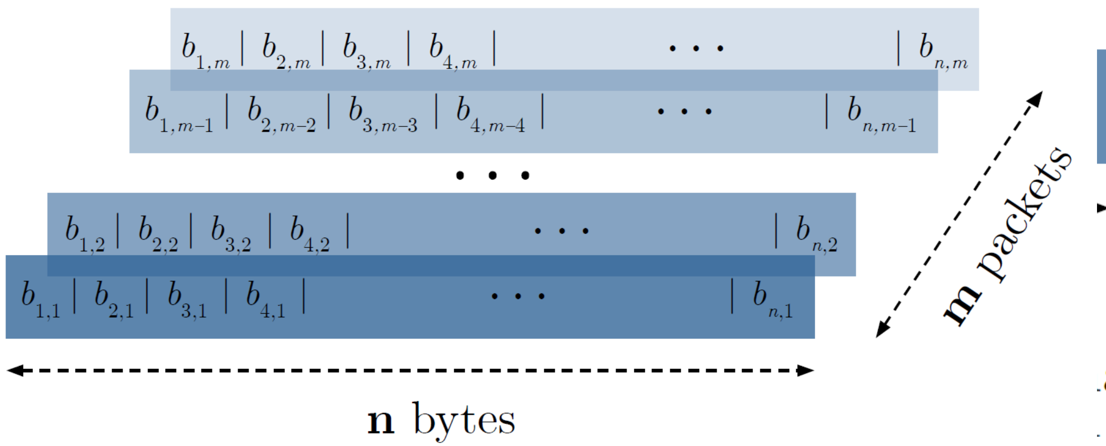
- Different architectures to **analyze both packet-based** and **flow-based byte aggregations**
- Models for **binary malware detection** – fully **supervised-based training**

Raw Input Representations

- **Input representation of the data**, as well as **network architecture**, are both **key elements** to consider when **building a DL model**

- We take **two types of raw input representations:** ***packets*** and ***flows***. Decimal normalized representation of every byte of every packet is a different input

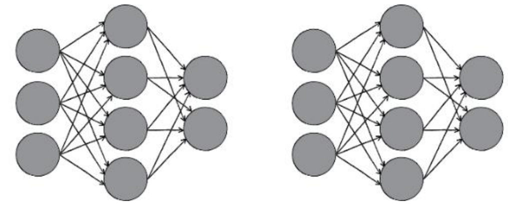
- **Flow representation:** matrix-like input, first **m** packets x first **n** bytes



(a) Pac... ape of the input da... and n the number

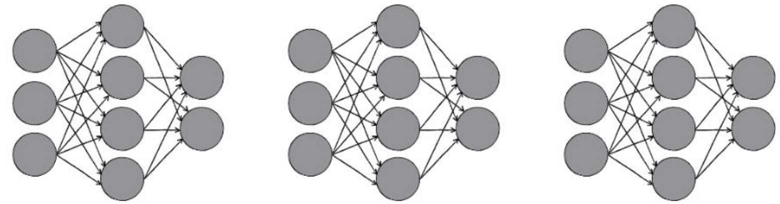
(b) Flow representation for the input data. A tensor of size (N, m, n) where N represents the number of instances –flows–, m the number of channels –packets– and n the number of steps –bytes–.

Deep Learning – Architectural Principles

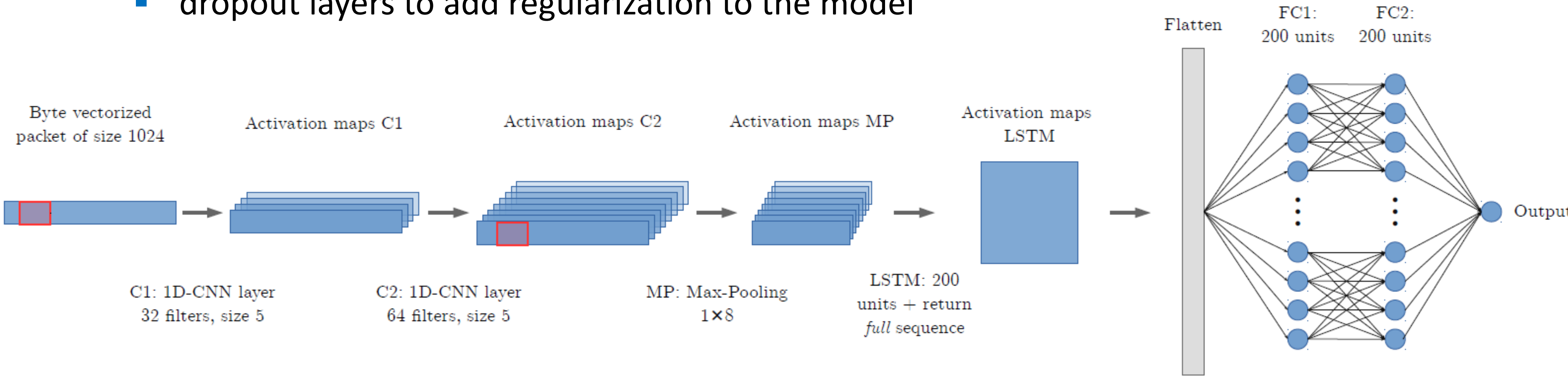


- The **core layers** used for both models are basically two: **convolutional and recurrent**
- Convolutional, to build the feature representation of the spatial data inside the packets and flows
- Recurrent layers are used together with the convolutional ones to allow the model keeping track of temporal information
- **Fully-connected layers** to deal with the different feature combinations
- **Batch Normalization**: layer inputs are normalized for each mini-batch. As a result: higher learning rates can be used, model less sensitive to initialization and also adds regularization
- **Dropout**: randomly drop units (along with their connections) from the neural network during training. A very efficient way to perform **model averaging**

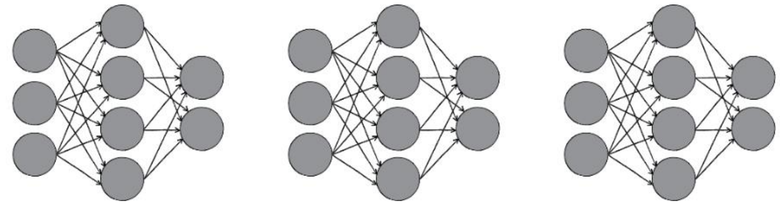
DL Architectures – Packets



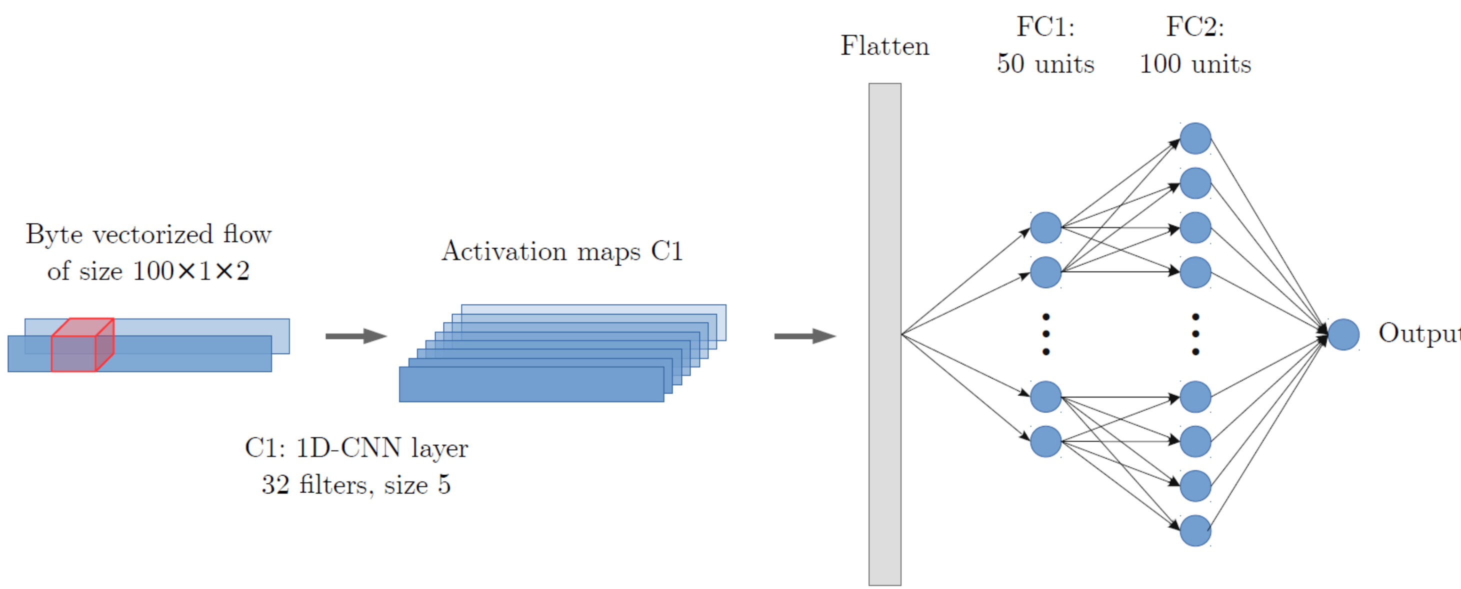
- **Raw Packets Architecture:**
  - **n** is set to first **1024 bytes**
  - two 1D-CNN layers of 32 and 64 filters (size 5) respectively
  - MP - max pooling layer (size 8)
  - LSTM layer with 200 neurons
  - two fully-connected layers of 200 neurons each
  - **binary cross-entropy as loss function**
  - spatial and normal batch normalization layers after each 1D-CNN and FC layers to ease training
  - dropout layers to add regularization to the model



DL Architectures – Flows

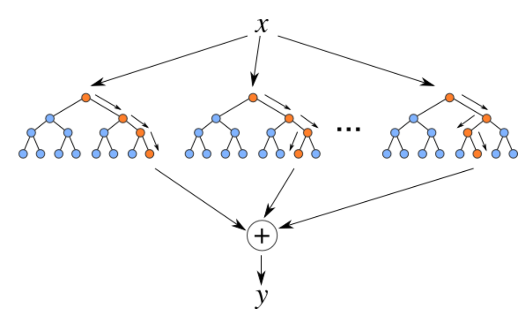


- **Raw Flows Architecture:** we go for a simpler model, with ***less features***
  - **n** is set to first **100 bytes**, and **m** to first **2 packets**
  - one 1D-CNN layers of 32 filter (size 5)
  - two fully-connected layers of 50 and 100 neurons each
  - **binary cross-entropy as loss function**
  - spatial and normal batch normalization layers
  - dropout layers to add regularization to the model



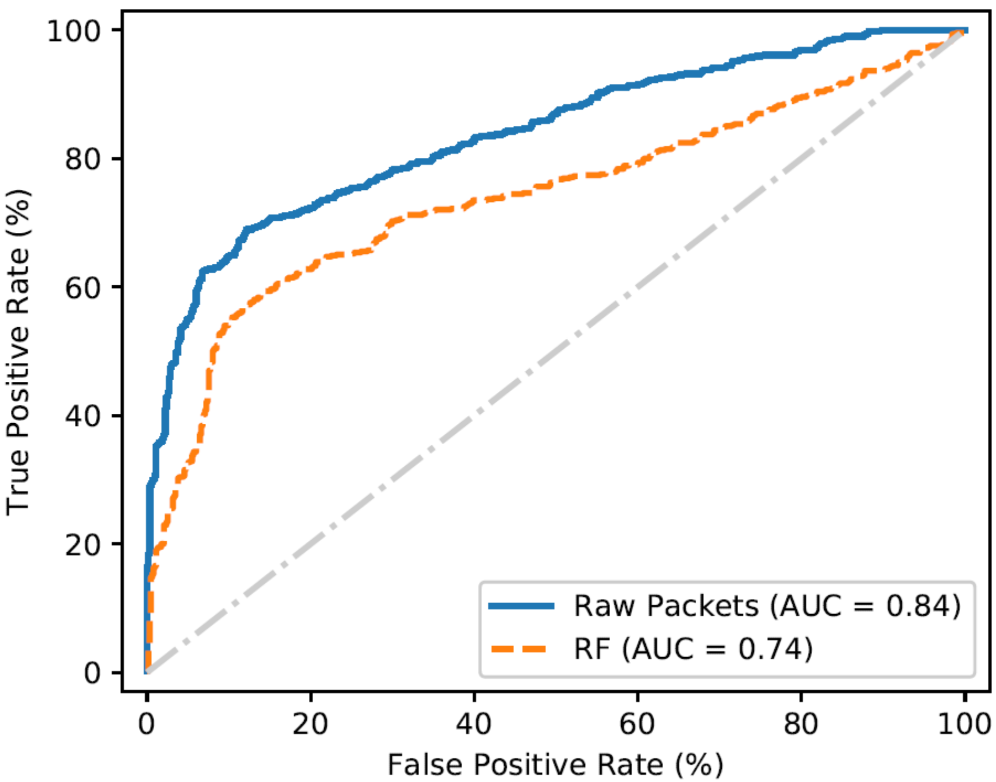
Evaluations

- All evaluations run on top of **Big-DAMA cluster** (distributed CPU)
- **Keras framework** running **on top of TensorFlow**
- **Dataset**: ***malware*** and ***normal*** traffic captures (pcap) performed by the Stratosphere IPS Project of the CTU University of Prague
- **250.000 raw packet** instances, **70.000 raw flow** instances
- **80%** of the samples for **training**, **10%** for **validation** and **10%** for **testing**
- **Compare** performance to **highly expressive Random Forest**:
  - same raw inputs
  - 100 trees
  - max depth and instances per leaf set for high expression
  - **selected based on great outperformance in state of the art**



RawPower – Packet Representation

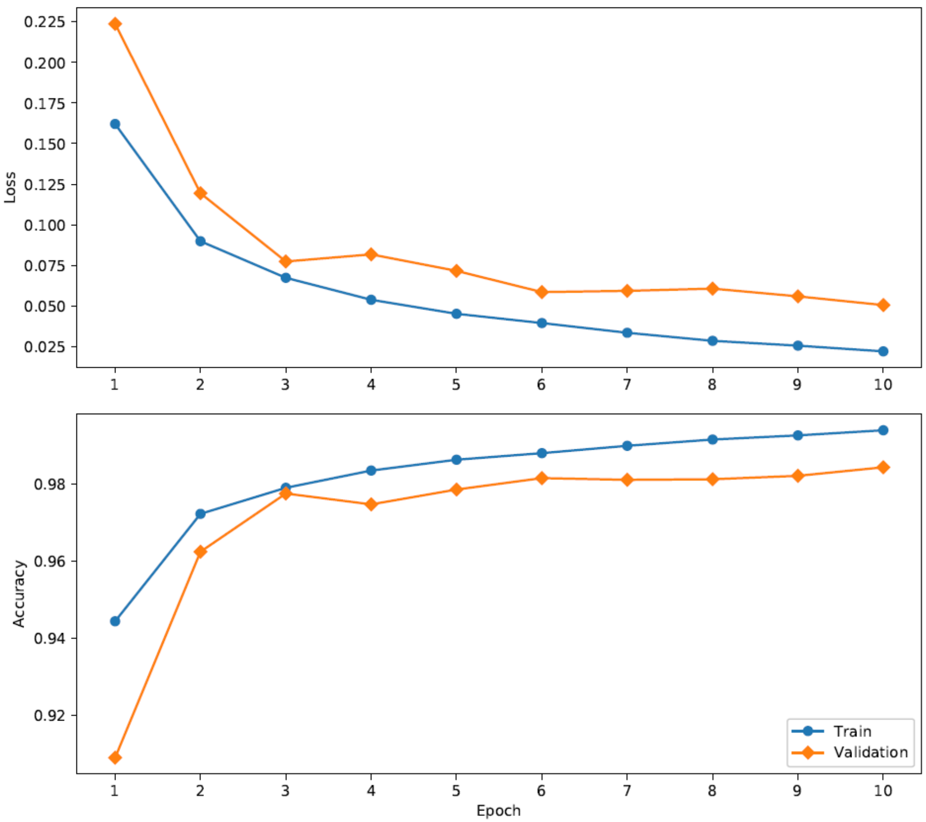
- Malware consists of **10 different malware types**, collected at controlled environment

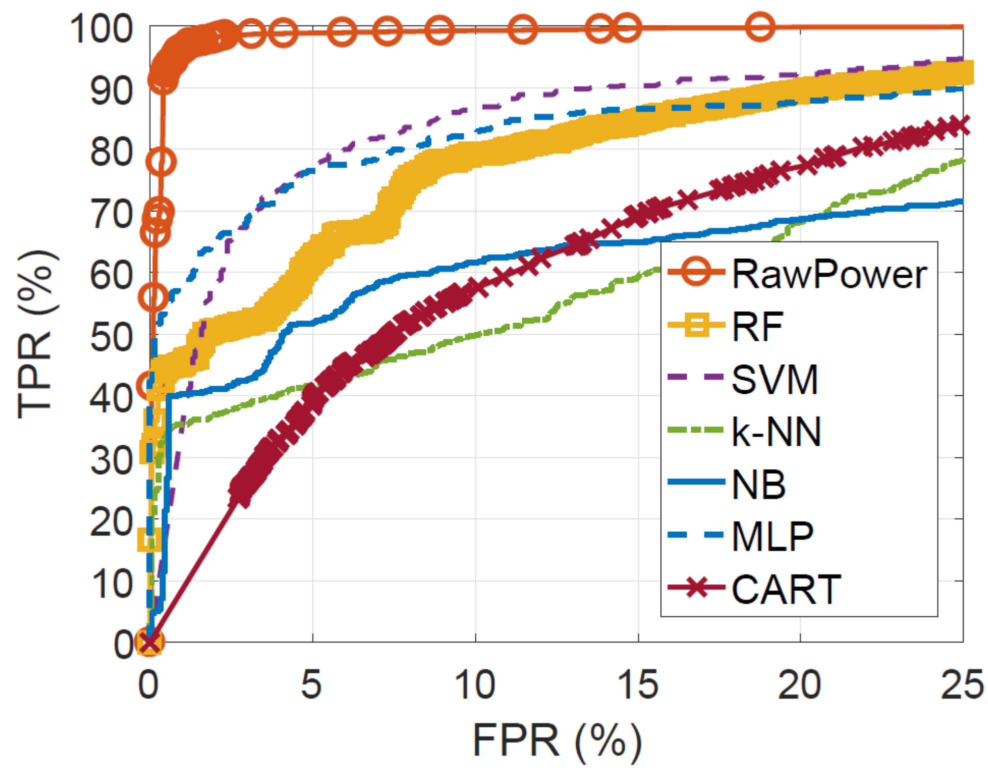


- **ROC curves** for both RawPower and RF
- Both models using the same **raw packet inputs**
- **Performance is not good at the packet-level**
- **Little gain w.r.t. a simple RF model**

RawPower – Flow Representation vs Shallow ML

- **Training and validation** evolution over **10 epochs**
- **Much better performance** at the **flow level**
- RawPower can **detect almost 98% of the malware** flows with a FPR < 0.5%
- **Shallow models not able to capture the underlying relations**

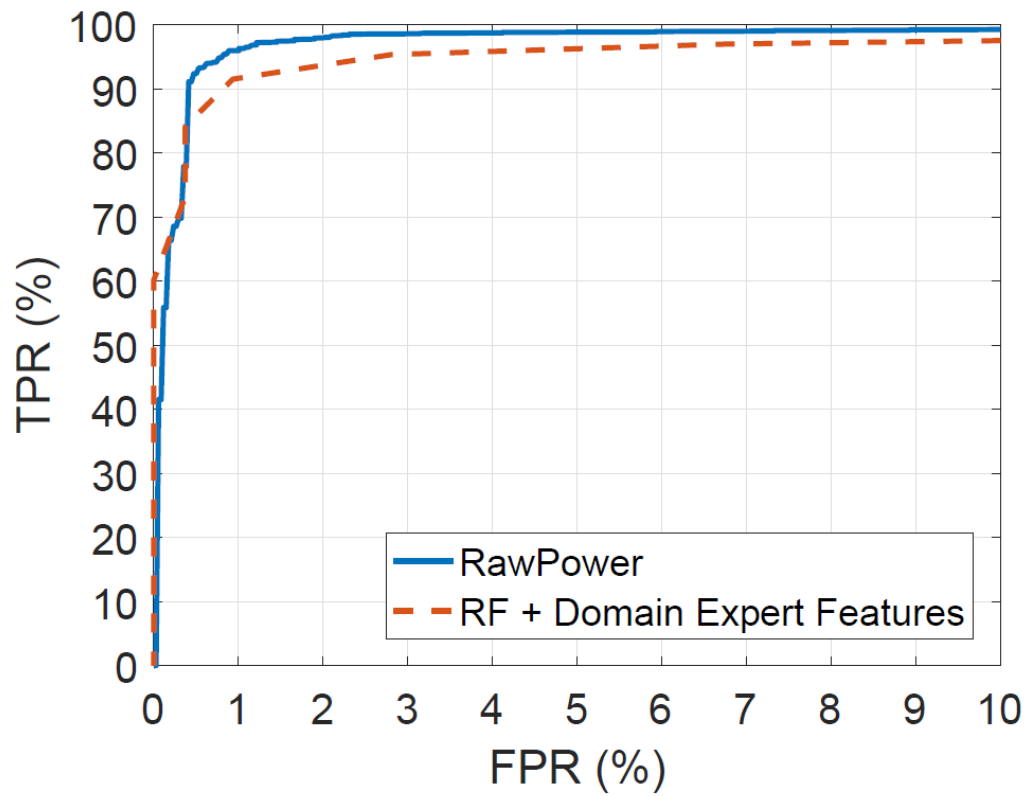




RawPower vs. shallow models.

RawPower – Flow Representation vs Expert Features

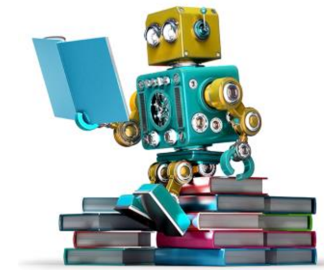
- Comparison **against** traditional **RF-based model**, which uses highly engineered **input features**, extracted from **domain knowledge**
- Both models provide **comparable results**
- The **key advantage of RawPower** is to rely directly on the usage of **bytestream raw data as input**
- **Input representation learning**: no the need for feature engineering



RawPower vs. knowledge-based inputs.

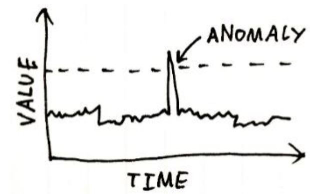
Organization of the Talk

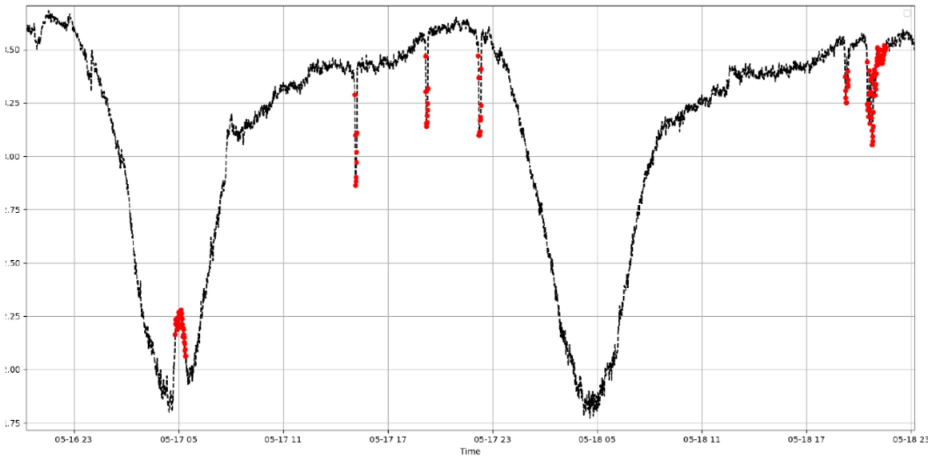
Dealing with Some of the Challenges in AI4NETS

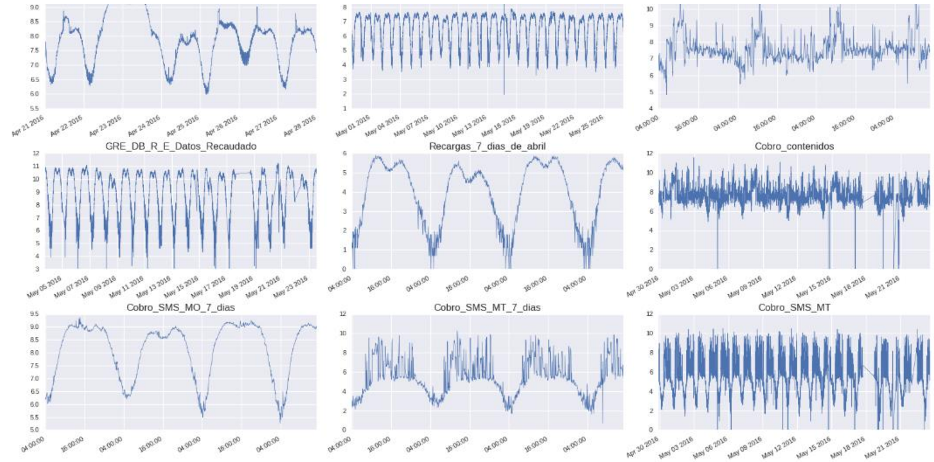


- Deep Learning for Malware Detection – ***Avoid Feature Engineering***
- **Generative Models for Anomaly Detection** – ***Avoid Traffic Modeling***
- Explainable Artificial Intelligence (XAI) **–** ***Interpret Model Decisions***
- Super Learning for Network Security – ***Avoid Model Decision***
- Adaptive/Stream Learning for NetSec – ***Deal with Concept Drifts***
- Reinforced Active Learning – ***Deal with Lack of Ground Truth***

Anomaly Detection in Multivariate Time-Series







Anomalies in an univariate time series

Different univariate time series of the same system

- **Anomaly Detection (AD)** is, by definition, an ***unsupervised process*** (detect what is different from the majority – the ***baseline***)

- **Baseline construction** (i.e., system modeling) is ***complex*** and ***error prone***, especially when dealing with ***multi-dimensional*** system characterization

- Solution: delegate the baseline construction to ***generative models***

Generative Models

- Given training data, ***generate new samples*** from ***same distribution***

Training data ~ $p_{data}(x)$

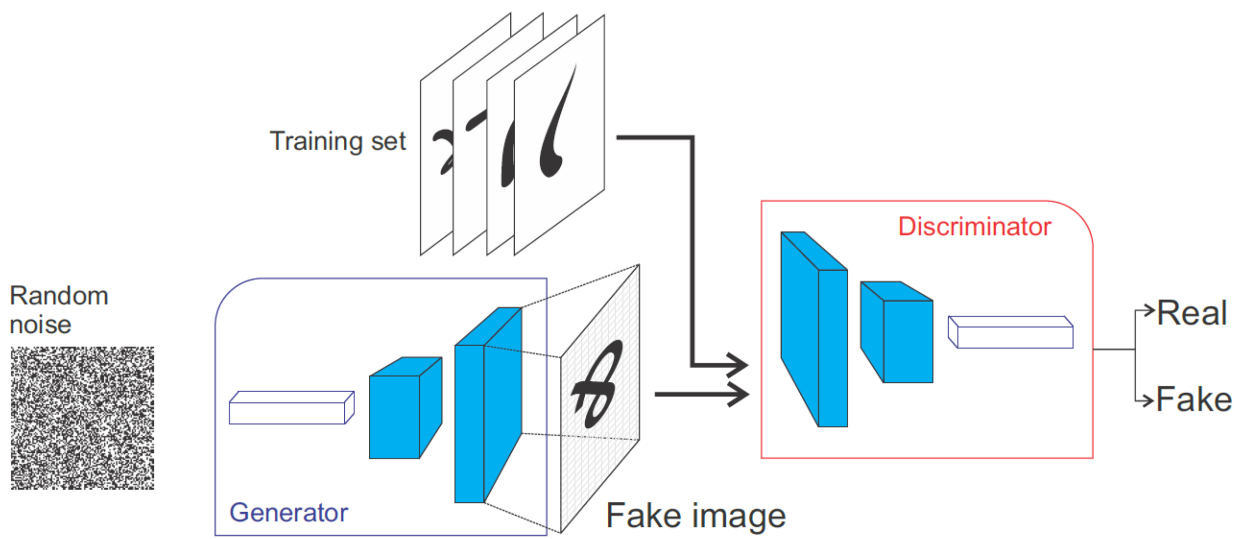
Generated samples ~ $p_{model}(x)$

- The **problem of generative models** is about **learning $p_{model}(x)$ similar to $p_{data}(x)$**

- Generative model learning is about **density estimation:**
  - **Explicit density estimation**: explicitly define and solve for $p_{model}(x)$
  - ***Implicit density estimation***: learn model that can sample from $p_{model}(x)$ w/o explicitly defining it

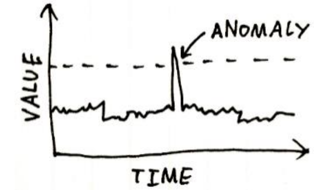
Generative Adversarial Networks (GANs)

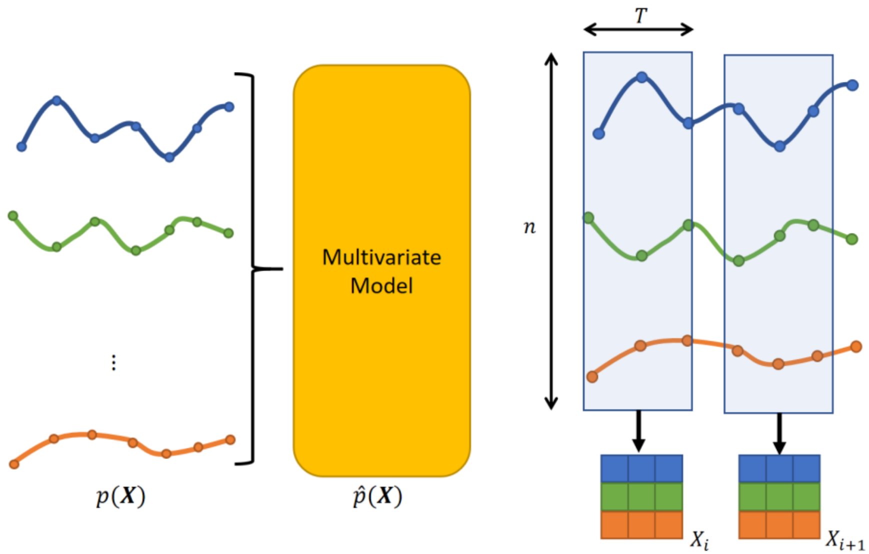
- Implicit density estimation through **game-theoretic approach**
- Learn to generate samples from training distribution through **2-players (minimax) game**
- **Problem:** want to sample from potentially complex, high-dimensional training distribution. No direct way to do this!
- **Solution:** sample from a simple distribution, e.g. **random noise**. Learn transformation to training distribution, **using a neural network**

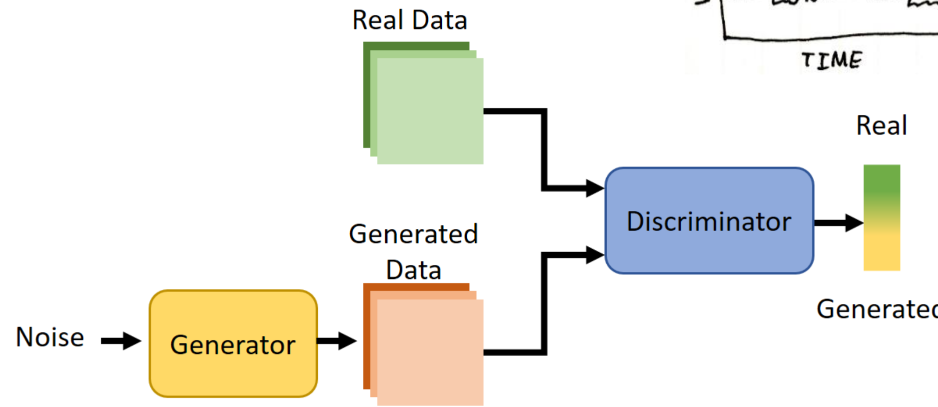


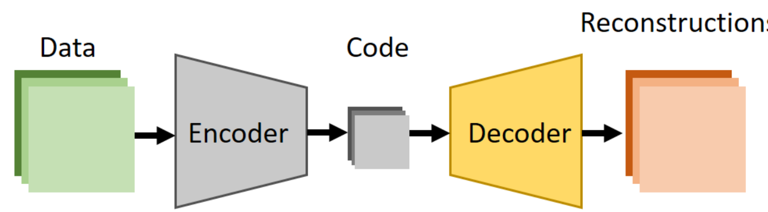
- **Generator** network: tries to fool the discriminator by generating real-looking instances from random noise
- **Discriminator** network: tries to distinguish between real and fake instances

Generative Models for MV-TS Anomaly Detection







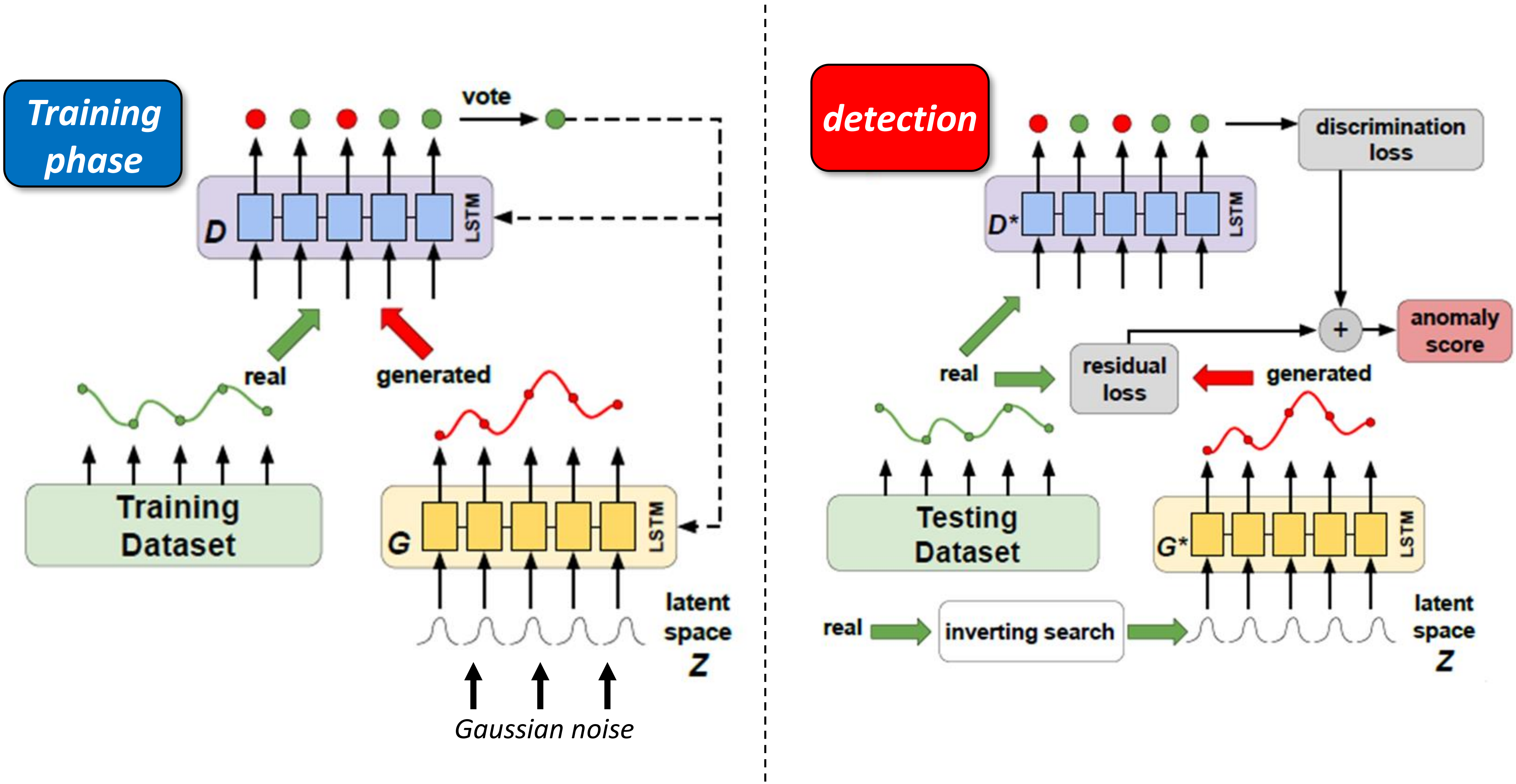


Two different generative models for ***AD in multi-variate time series***

- ***Net-GAN***: Recurrent Neural Networks (LSTM) trained through GANs
- ***Net-VAE***: Variational Auto-Encoders (VAE) using feed-forward NNs
  - VAEs improve Auto-Encoders by ***regularizing the latent-space*** → enabling ***generative process***
- Input samples: matrix with *n* (number of variables) x *T* (length of sequence)

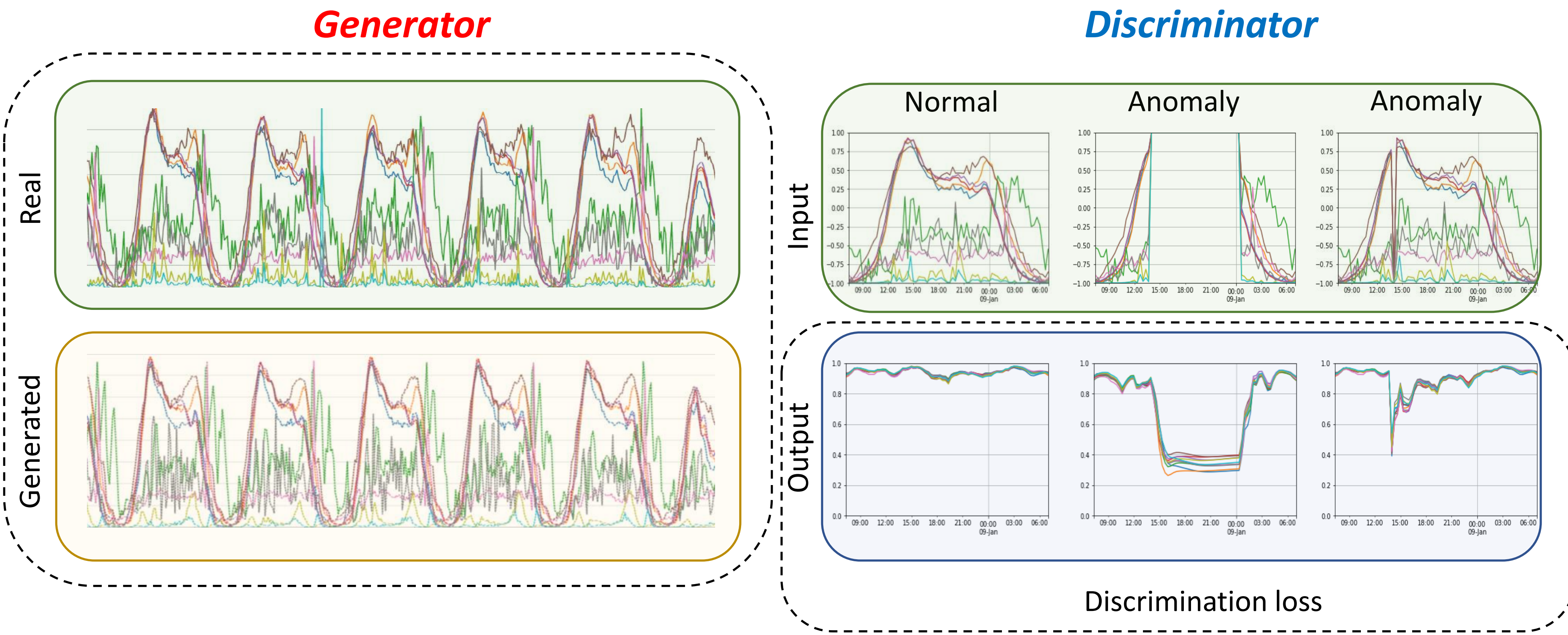
Network Anomaly Detection with Net-GAN

- ***Net-GAN detection*** can be done both through the ***generator (G)*** and the ***discriminator (D)***



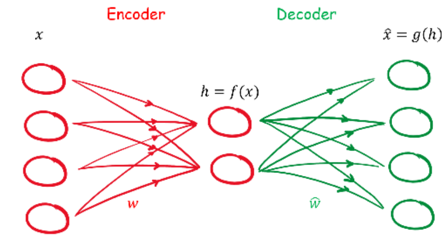
Examples on Real (Mobile) ISP Network Data

Net-GAN application phase

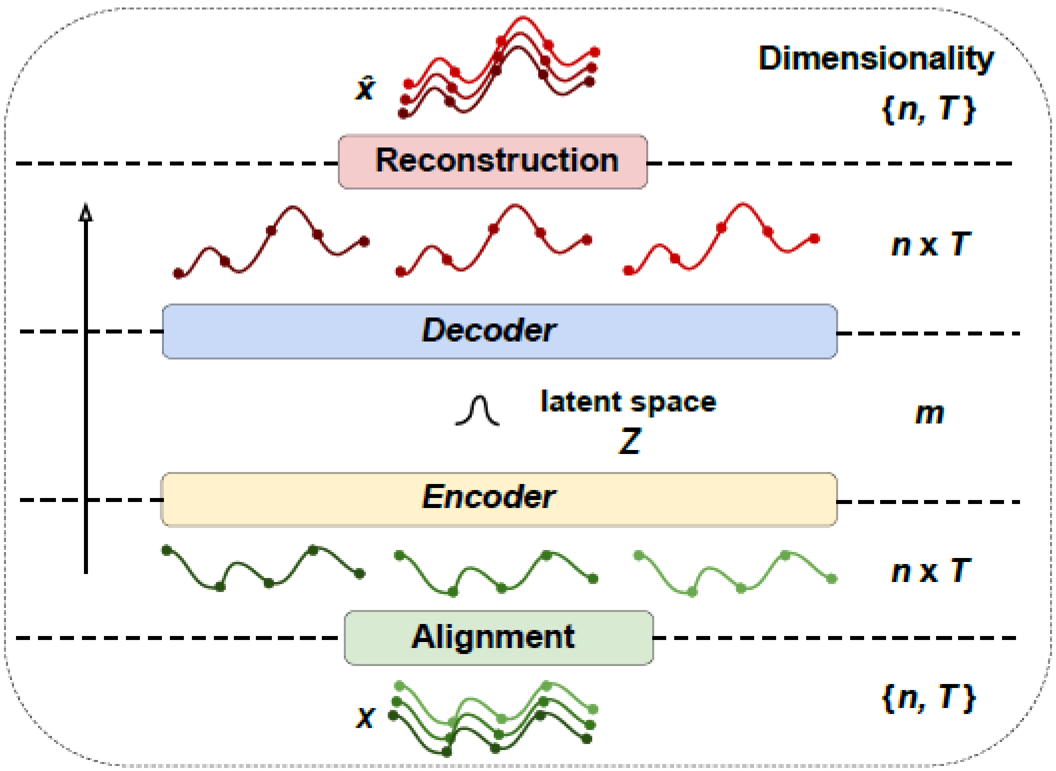


- ***Net-GAN AD generator (G)*** → residual loss
- ***Net-GAN AD discriminator (D)*** → discrimination loss

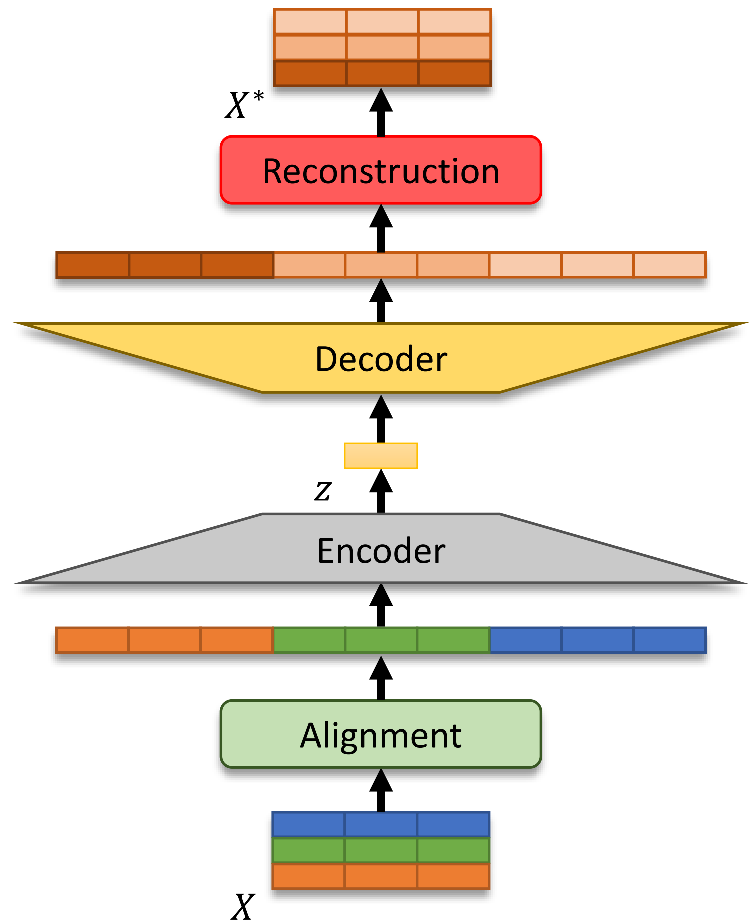
Network Anomaly Detection with Net-VAE



- Net-VAE architecture:
  - standard encoder and decoder functions
  - encoder/decoder using ***3-layer FF networks***
  - detection on ***residual loss***

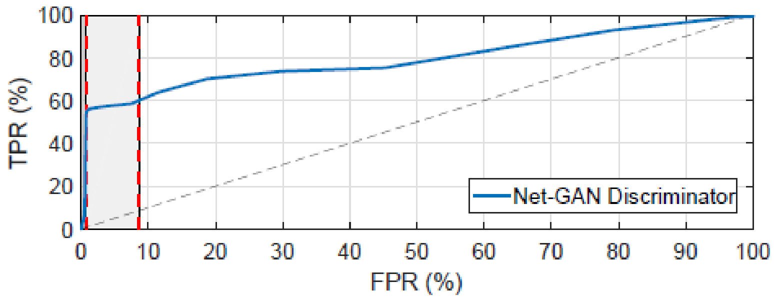


Net-VAE architecture.

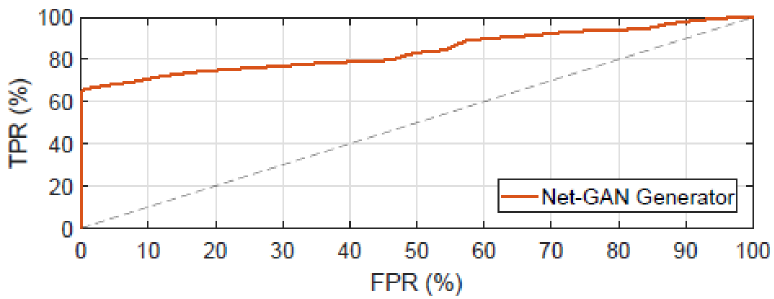


Anomaly Detection with *Net-GAN* and *Net-VAE*

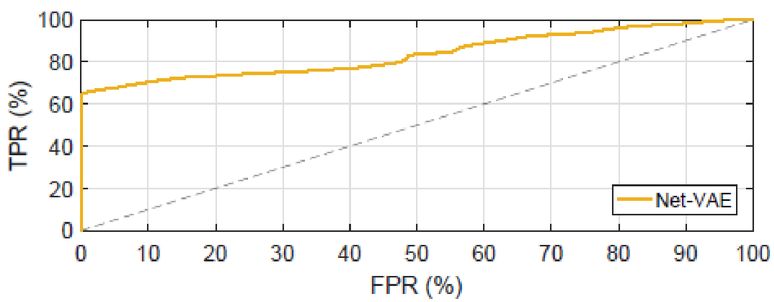
SWaT (CPS measurement)



(a) Detection performance with Net-GAN-D.

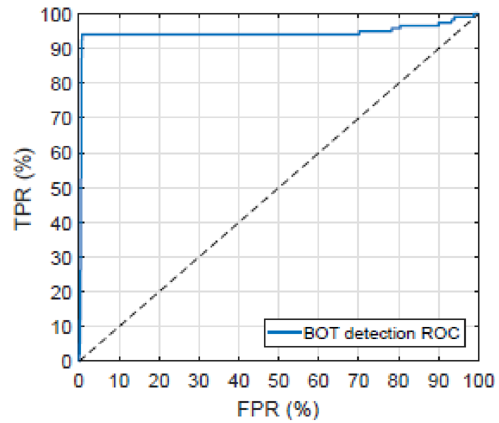


(b) Detection performance with Net-GAN-G.

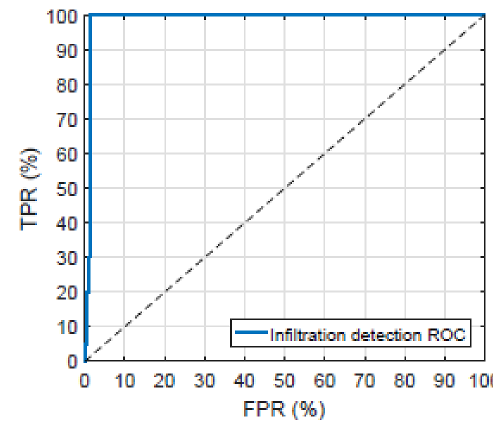


(c) Detection performance with Net-VAE.

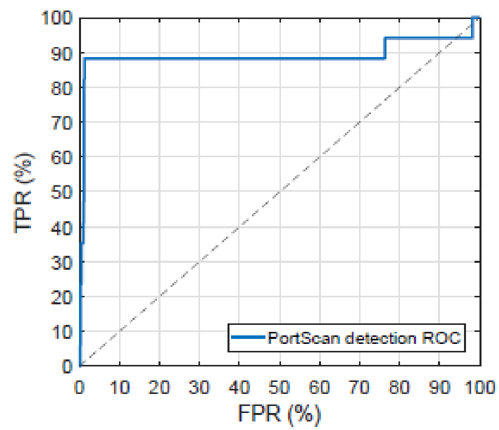
CICIDS2017 (SYN-NET measurements)



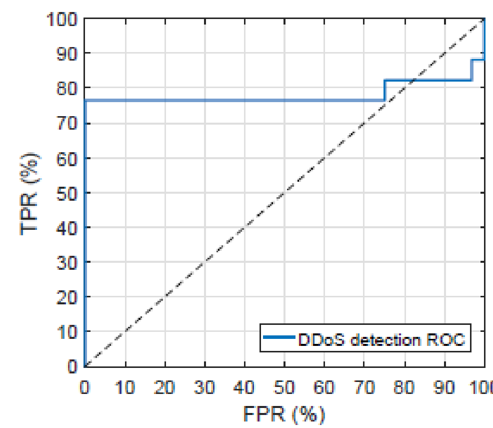
(a) Botnet



(b) Infiltration



(c) Port Scan



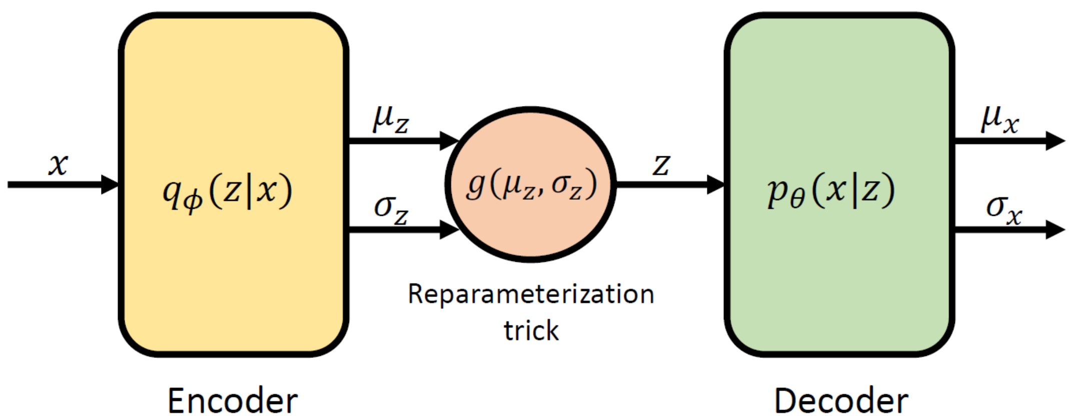
(d) DDoS

Variational Auto-Encoders ***with*** *Dilated Convolutions*

- ***Unsupervised and multivariate*** *approach to* ***anomaly detection*** *in time-series, based on VAEs*

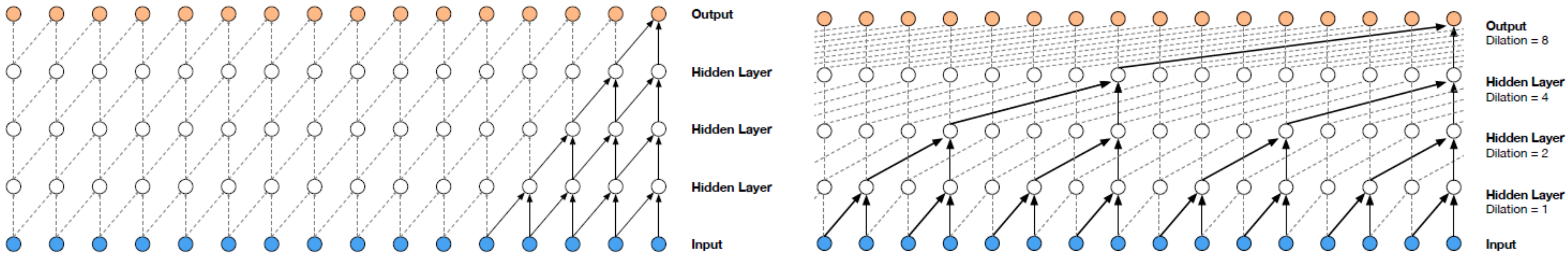
- VAEs are a ***generative version of classical autoencoders***, with the particularity of having, by conception, continuous latent spaces; as such, ***VAEs enable a generative process*** (e.g., GANs)

- To ***exploit the temporal dependencies*** and characteristics of time-series data in a fast and efficient manner, we take a ***Dilated Convolutional Neural Network (DCNN) as the VAE's encoder and decoder architecture***



Variational autoencoder

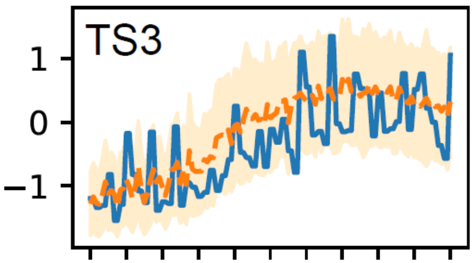
Dilated Convolutions ***for*** *Longer-in-Time* ***Sequences***



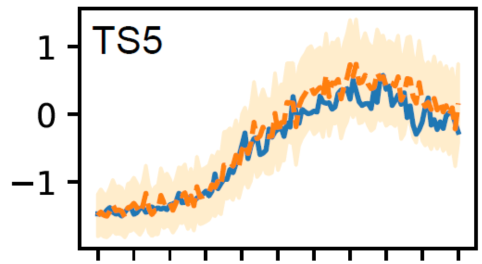
(a) Normal convolution. (b) Dilated convolution.

- *Using* ***CNNs with causal filters requires large filters or many layers to learn from long sequences***
- ***Dilated convolutions*** *improves time-series modeling by increasing the receptive field of the neural network (**longer in the past sequences**)…*
- *…reducing computational and memory requirements, enabling training on long sequences*

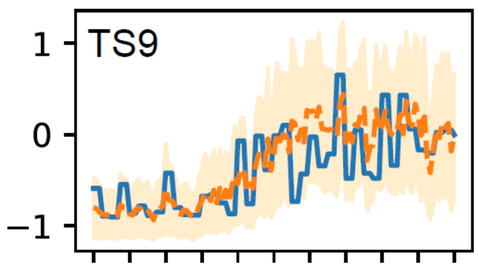
DC-VAE Encoder Architecture



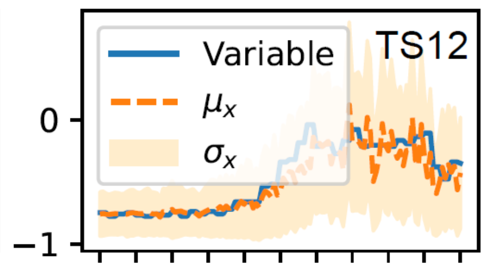
(a) Prediction of time-series TS_3.



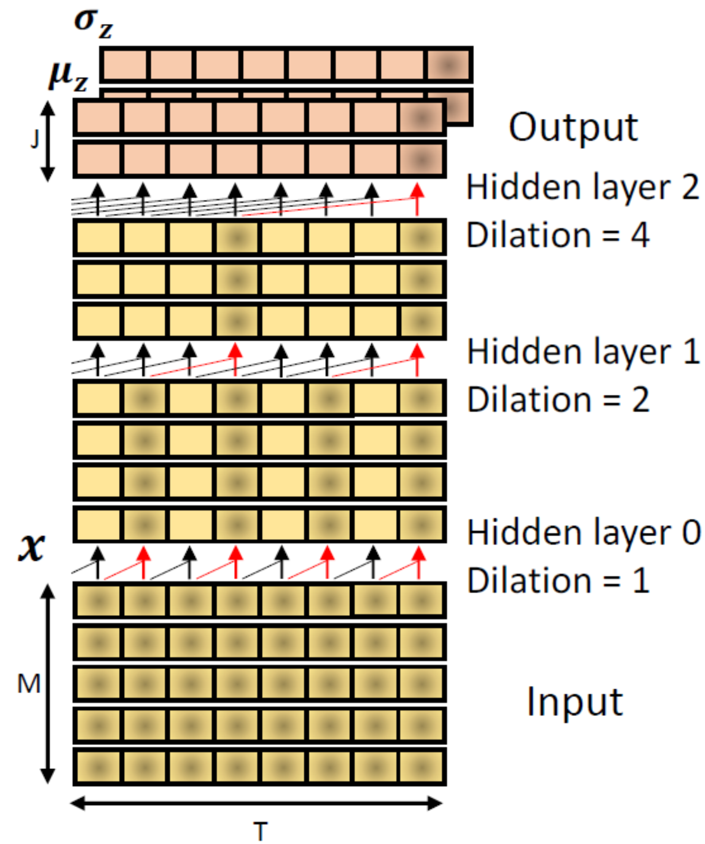
(b) Prediction of time-series TS_5.



(c) Prediction of time-series TS_9.



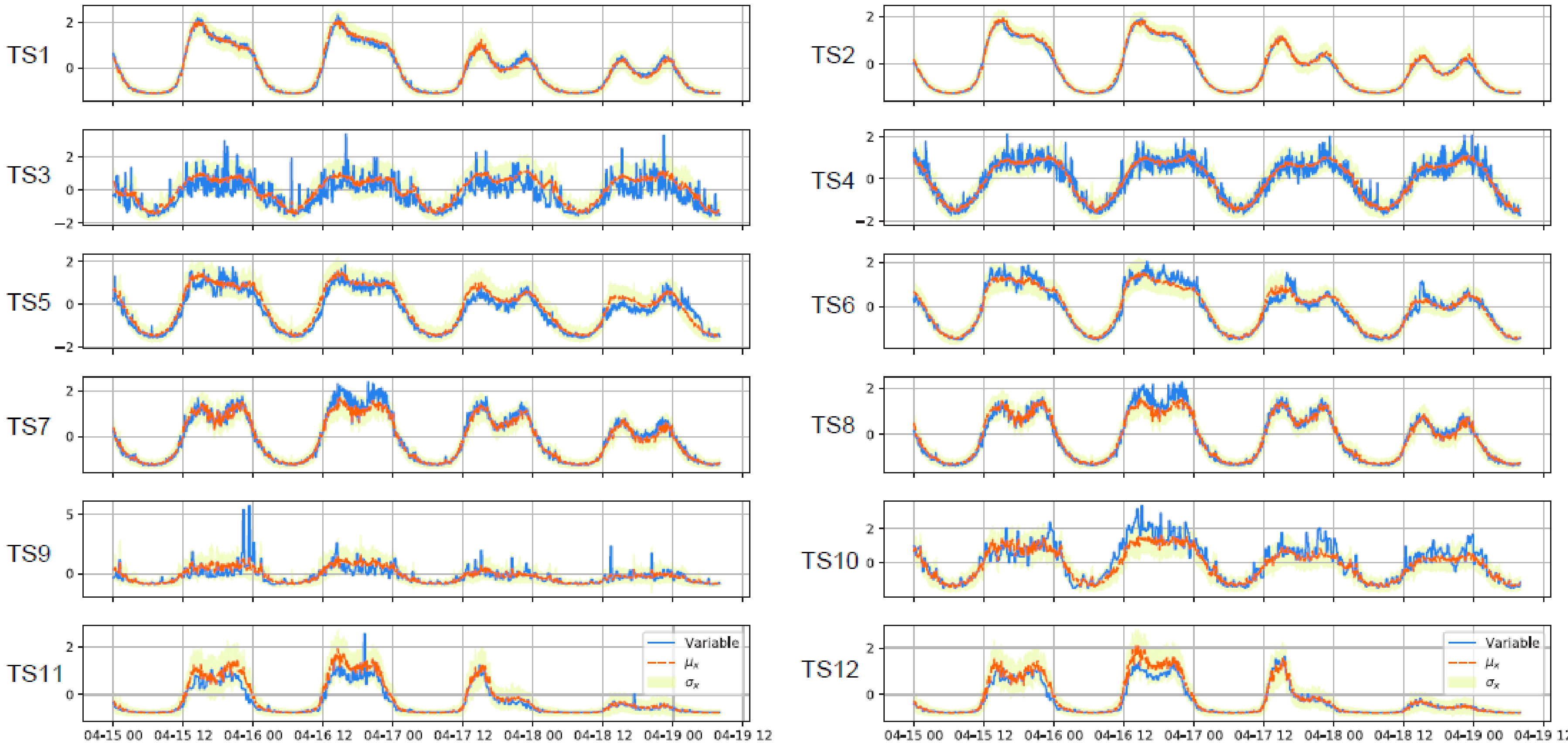
(d) Prediction of time-series TS_{12}.



- Example of time-series analysis through DC-VAE. *The **normal-operation region** is defined by μ_x and σ_x*

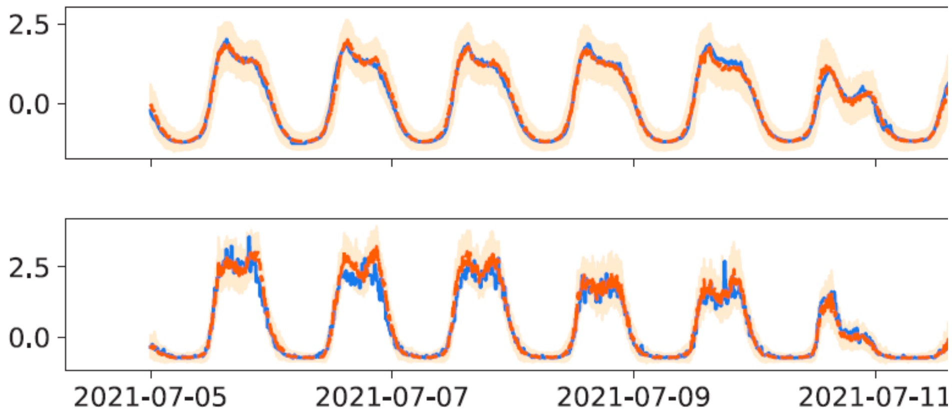
- Encoder architecture using causal dilated convolutions, implemented through a stack of ***1D convolutional layers***

DC-VAE – Application Examples in TELCO Datasets (1/3)

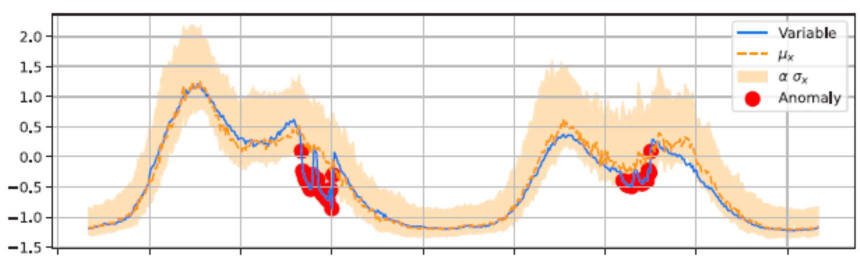


TELCO dataset time-series, for a period of four days, along with the corresponding *DC-VAE* estimations. The temporary receptive field – i.e., length of the rolling time-window, is $T = 512$ samples, spanning about two days of past measurements.

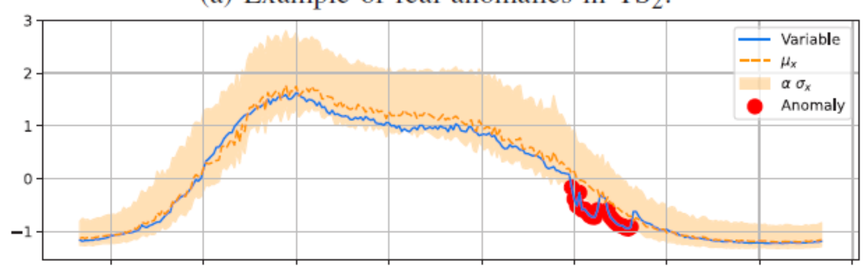
DC-VAE – Application Examples in TELCO Datasets (2/3)



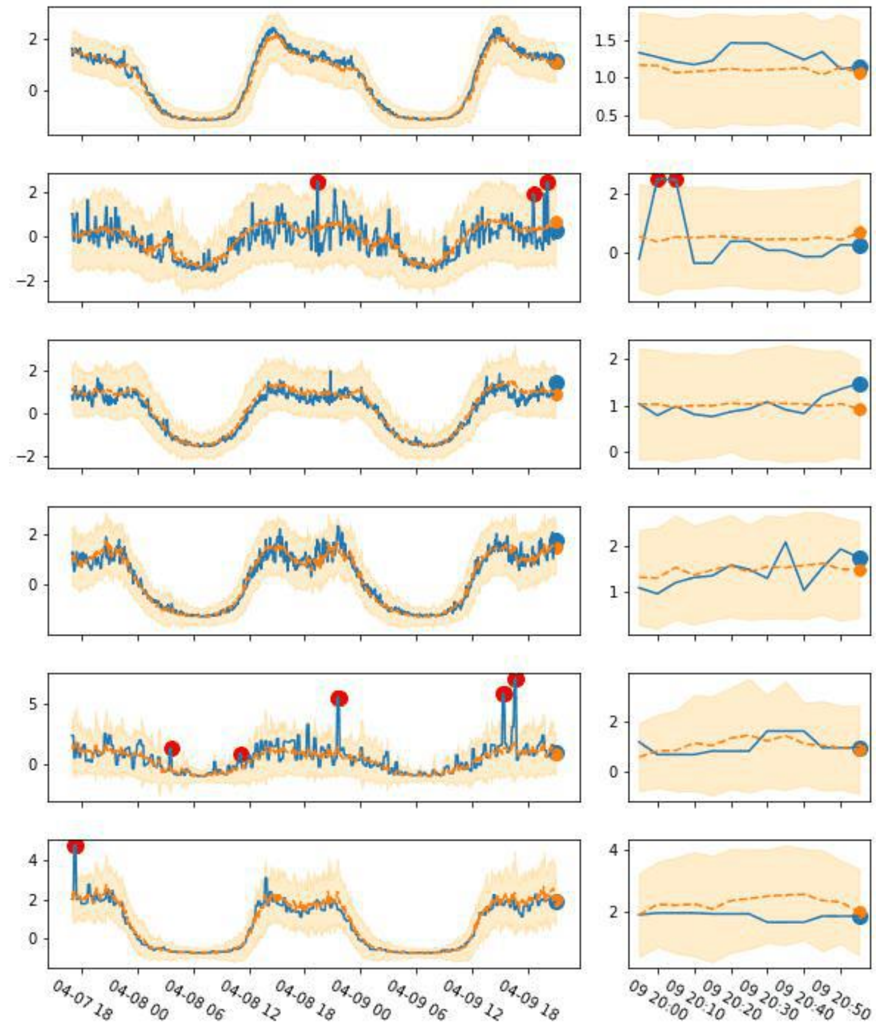
DC-VAE operation for time-series with stationary behavior. Weekly



(a) Example of real anomalies in TS_2.



(b) Example of real anomalies in TS_2.



DC-VAE – Application Examples in TELCO Datasets (3/3)

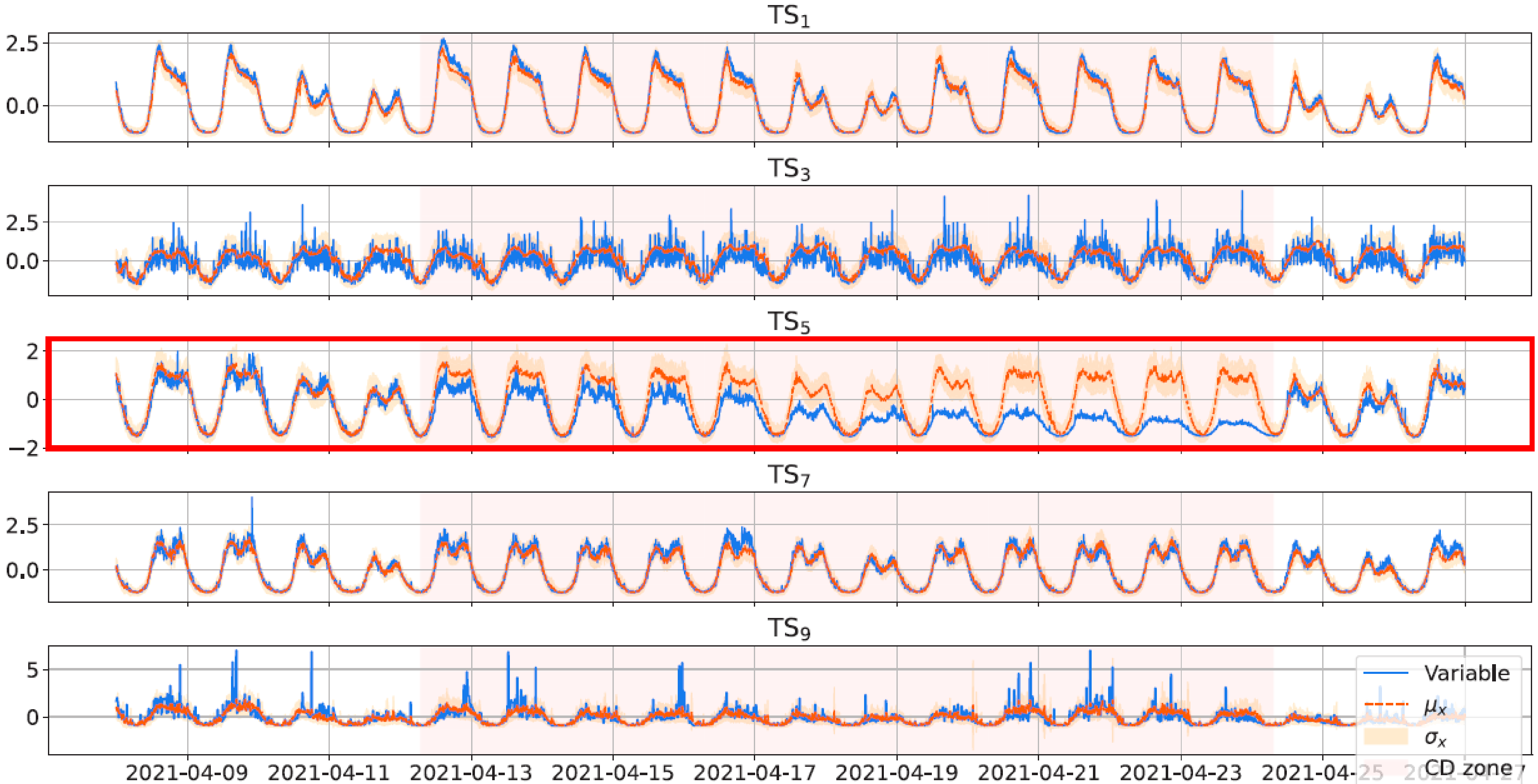
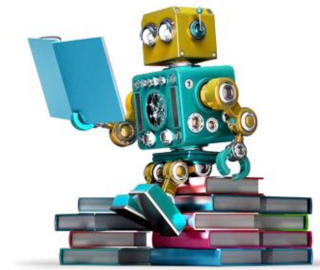


Fig. 14. *DC-VAE* response to univariate concept-drift: a gradual linear fall of the values during the day without affecting night behavior. While the drift does not affect the predictions on the other time-series, it becomes easily detectable at the corresponding time-series.

- Multivariate modeling helps to cope with concept-drift detection

Organization of the Talk

Dealing with Some of the Challenges in AI4NETS



- Deep Learning for Malware Detection – ***Avoid Feature Engineering***
- Generative Models for Anomaly Detection – ***Avoid Traffic Modeling***
- **Explainable Artificial Intelligence (XAI) –** ***Interpret Model Decisions***
- Super Learning for Network Security – ***Avoid Model Decision***
- Adaptive/Stream Learning for NetSec – ***Deal with Concept Drifts***
- Reinforced Active Learning – ***Deal with Lack of Ground Truth***

EXplainable AI (XAI) – Why Should I Trust You?

- **ML models →** mostly are **black boxes** (exceptions: linear models, decision trees, etc.) – e.g.: some popular ML models have 10s of millions of parameters!

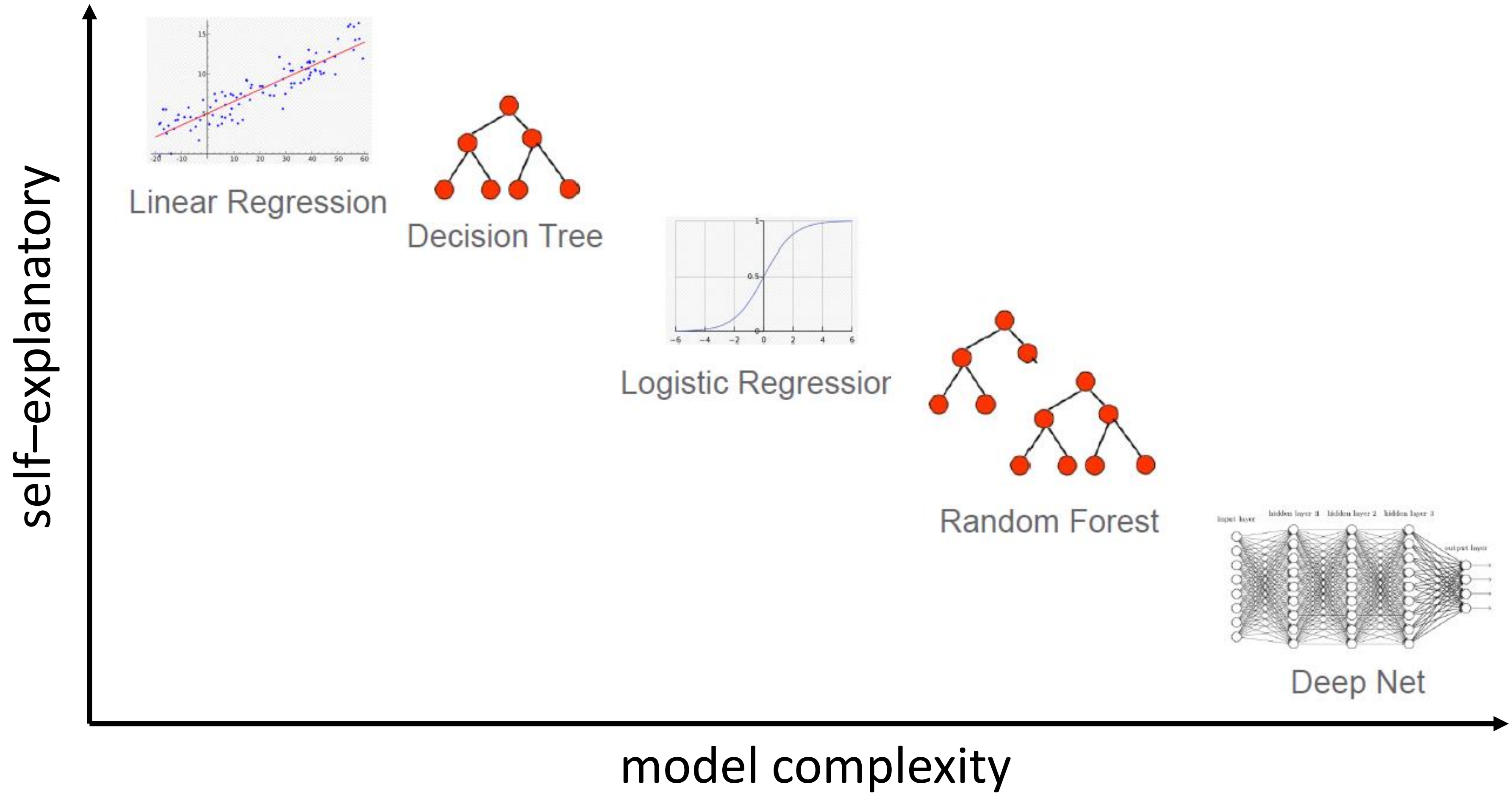
- Models are evaluated off-line before deployment on available test datasets – **data @runtime might change** (concept drift)

- Humans want to **understand model's behavior to gain trust** (applicability in the practice)
  - trusting an individual model's prediction
  - trusting a model (inspect a set of representative individual predictions)

- **Explainable AI:** approaches capable to explain models and individual predictions, by tracking back to the inputs leading to a certain output

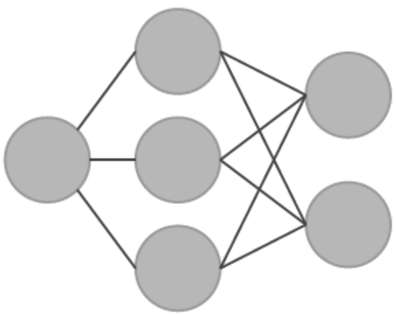
Why XAI?

- **Ideally**, **ML models** should be ***self-explanatory:*** **improve end-user understanding** and **trust**, by offering simple **explanations** of the ***"whys" of certain decision***
- **Only few models are self-explanatory:**



A Simple XAI Example

- **Application Example:** AI-supported disease diagnosis

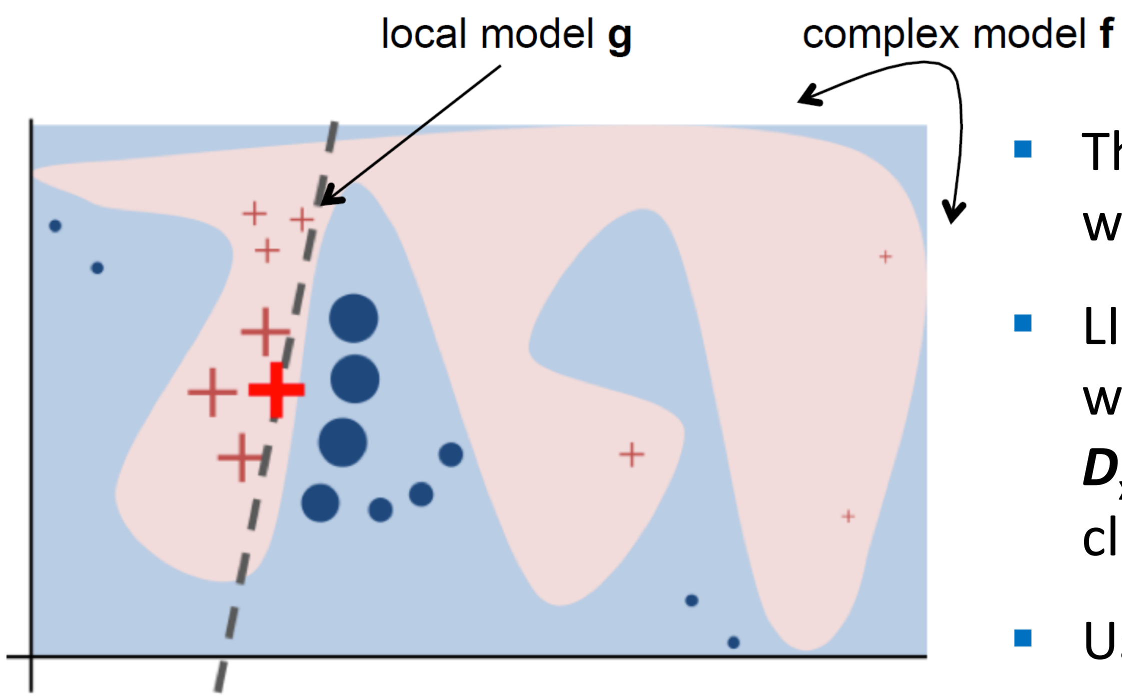


Model

- Explainer: **LIME** - **Local** Interpretable **Model-agnostic** Explanations
- ***LIME approach***: builds an **interpretable model** that is **locally** faithful to the classifier under analysis
- Other approaches: **SHAP**, LRP (NNs), PDP, etc.

LIME in a Nutshell – Sampling for Local Exploration

- Let **f** be an unknown complex decision function (**blue**/**pink** background)

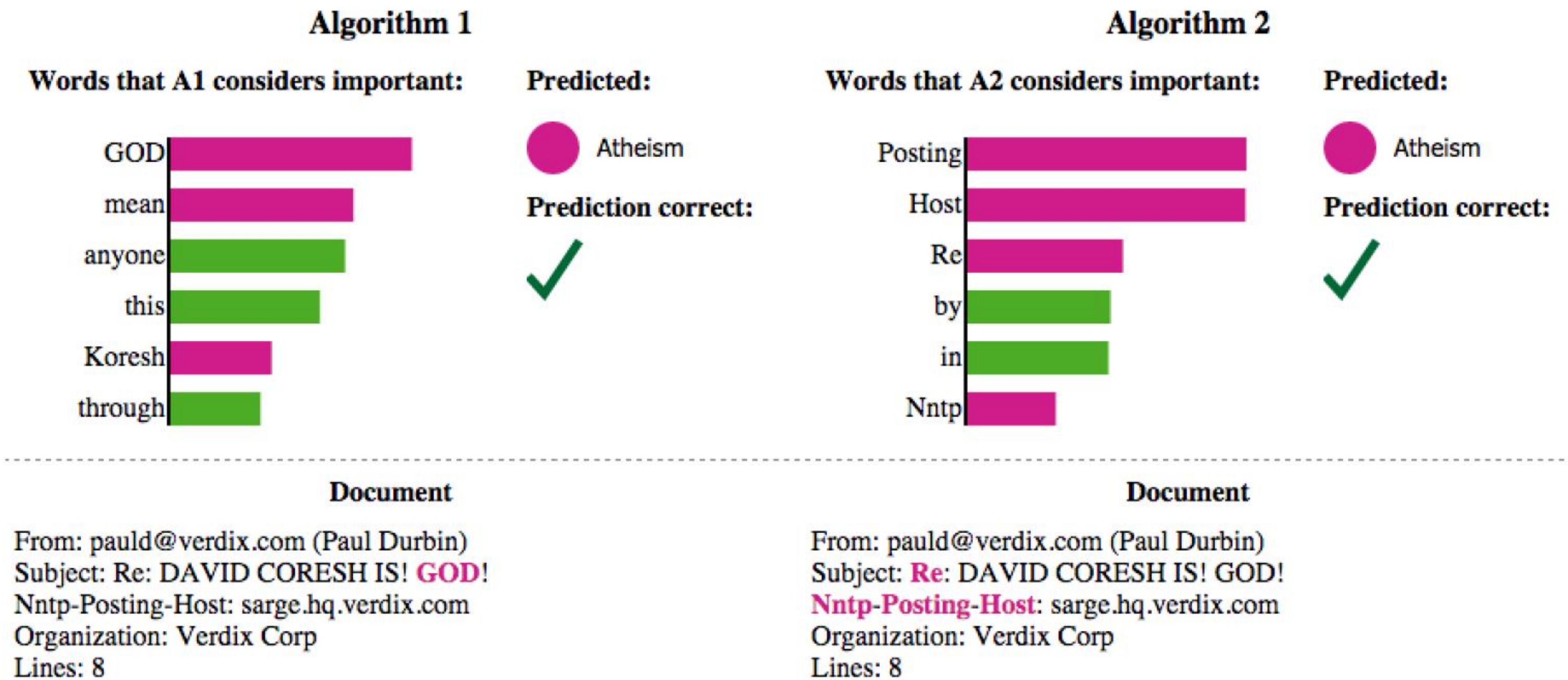


- **g** is *interpretable*, *locally faithful to* **f** (captured by $\boldsymbol{D_x}$), and *model agnostic* (uses **f(z)** as labels)
- *robust to sampling noise*, thanks to $\boldsymbol{D_x}$

- The **bold red-cross (x)** is the instance we want to explain
- LIME samples instances **z** around **x**, weighted by some **similarity measure** $\boldsymbol{D_x} \rightarrow \boldsymbol{D_x}(z)$ is higher for instances closer to x
- Using model **f**, gets the corresponding predictions **f(z)**
- Finally, it uses **z** and **f(z)** to **build an interpretable model g** (e.g, linear) around **x**

LIME Examples (I) – Model Comparison/Selection

- Task: word-based **email classification**, Christianity or Atheism
- 2 models (Algorithm 1 vs Algorithm 2), **which one is better?**



- Algorithm 2 is better than Algorithm 1 in terms of **accuracy in validation…**
- …but Algorithm 2 makes predictions for arbitrary reasons…Algorithm 1 is better
- **Performance metrics should be carefully considered**

LIME Examples (II) – Model Performance Evaluation

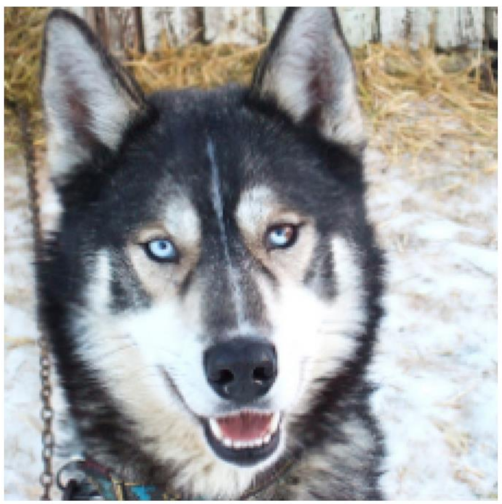
- Task: **image classification**, using Google's pre-trained Inception CNN architecture

(a) Original Image

- Figs. (b,c,d) report super-pixel explanations provided by LIME
- Top 3 classes: ***Electric Guitar*** (p = 0.32), ***Acoustic Guitar*** (p = 0.24), and ***Labrador*** (p = 0.21)
- The image is wrongly classified, but **explanations provide trust** in the model, as they are reasonable

LIME Examples (III) – Discover Biased Data

- Task: train a classifier to distinguish between Wolves and Huskies
- Biased data (e.g., **undesirable strong correlations**) → wrong classifier
- **Hard to identify** by looking at the raw data and predictions



(a) Husky classified as wolf

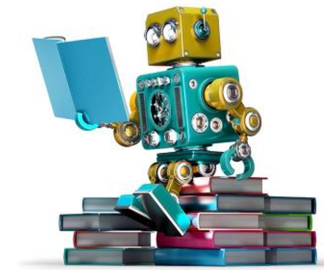
(b) Explanation

- Bias@training: all pictures of Wolves had **snow in background**
- The **classifier performs well** according to **cross-validation** in this **biased dataset…**

- …but **explanations** of individual predictions show that **the model learnt a biased pattern**: if snow → wolf, else → Husky

Organization of the Talk

Dealing with Some of the Challenges in AI4NETS



- Deep Learning for Malware Detection – ***Avoid Feature Engineering***
- Generative Models for Anomaly Detection – ***Avoid Traffic Modeling***
- Explainable Artificial Intelligence (XAI) **–** ***Interpret Model Decisions***
- **Super Learning for Network Security** – ***Avoid Model Decision***
- Adaptive/Stream Learning for NetSec – ***Deal with Concept Drifts***
- Reinforced Active Learning – ***Deal with Lack of Ground Truth***

Ensemble Learning for Network Security

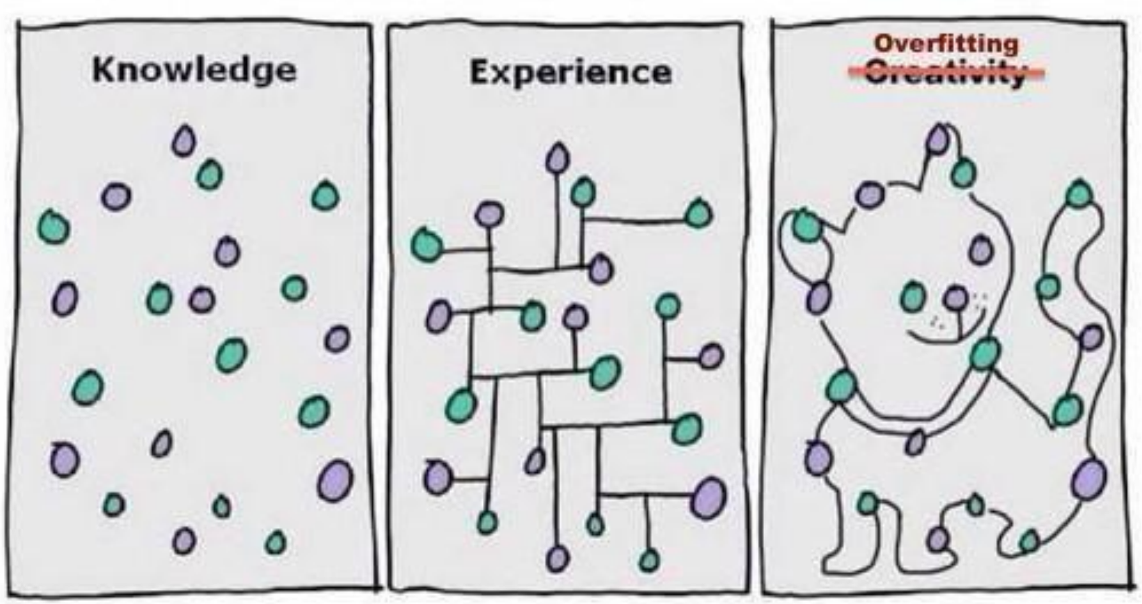
- Which is the **best model** or category of models for a **specific learning task**?
- Deep Learning? Not obvious in the context of Network traffic Monitoring and Analysis (NMA)
- Our claim: "***multiple-eyes principle***" → **ensemble learning** models
- We explore the application of **ensemble learning models** to multiple NMA problems…
- …following a particularly promising model known as the **Super Learner**

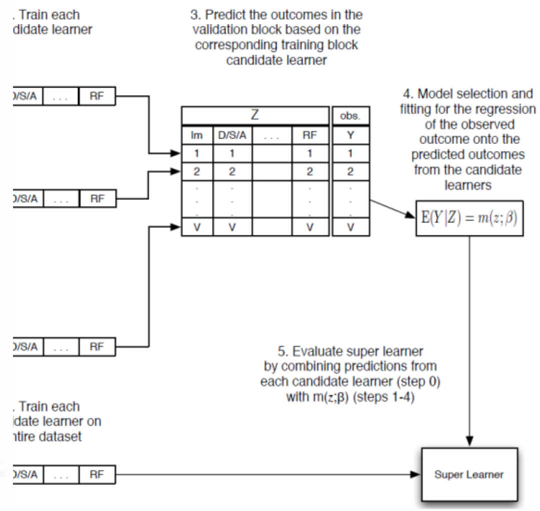
Ensemble Learning for Network Security

ensemble learning: combine multiple (base) learning models to obtain better performance.

- If a set of base learners do not capture the true prediction function (the oracle), **ensembles** can give a **good approximation** to that **oracle function**.
- **Ensembles perform better** than the individual **base algorithms**.
- **Multiple approaches** to ensemble learning, including **bagging** (decrease variance), **boosting** (decrease bias), and **stacking** (improve predictive performance)

Super Learner



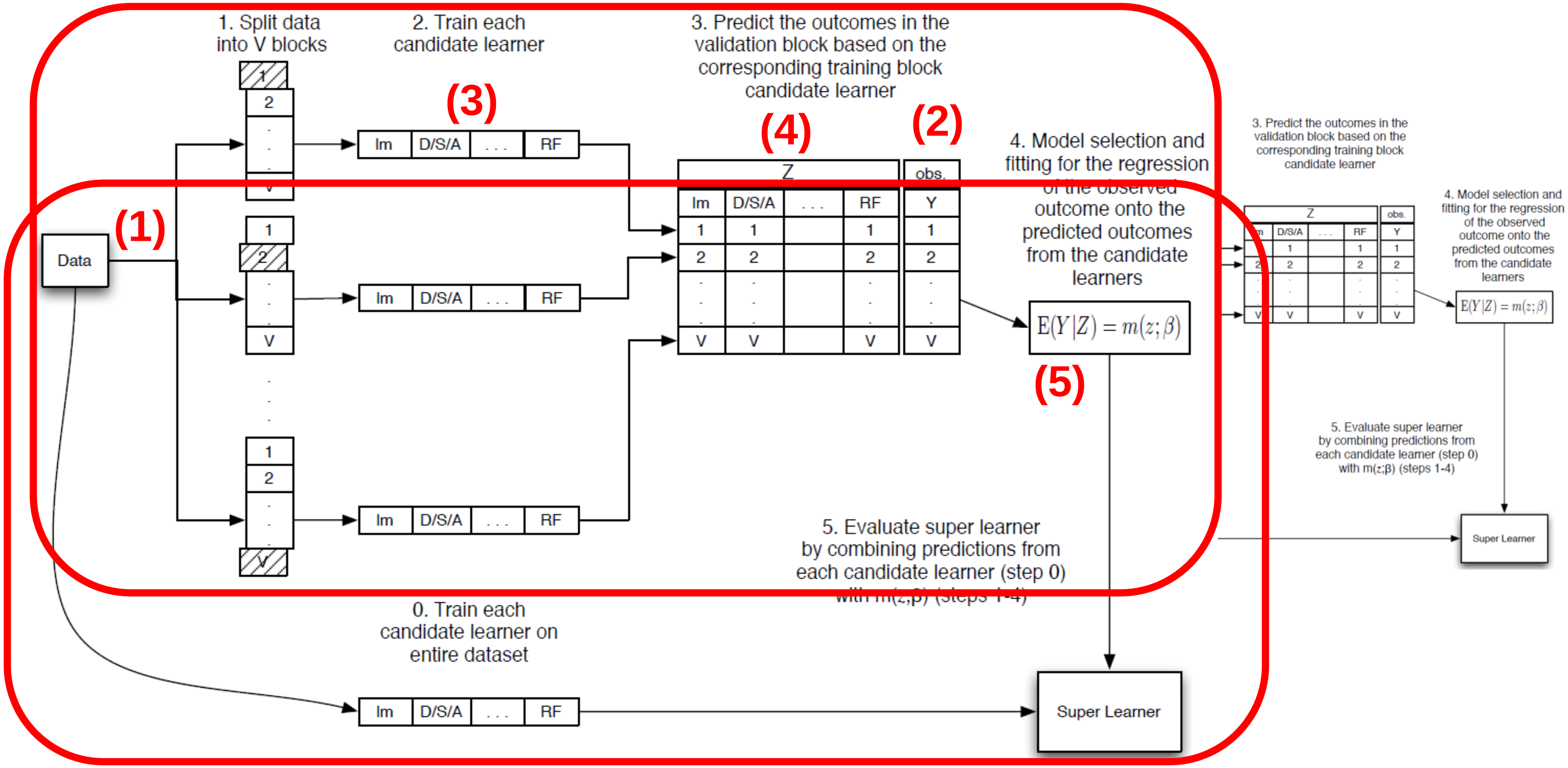


- **General ensemble learning** approaches might be **prone to over-fitting**.

- **Super Learner** [Van der Laan'07]: **stacking** ensemble learning **meta-model** that **minimizes over-fitting likelihood** using a **variant of cross-validation**.

- Finds the **optimal combination of a collection of prediction algorithms** → performs asymptotically as well – or better, than any of the ***base learners***.

Super Learner – How Does it Work?



- **2-steps approach (training and validation of Super Learner):**

1st ➔ *given a* ***(1) dataset X(k,l) with labels Y(k) (2)****, and a set of* ***(3) n base learners*** *(e.g., DTs, ANNs, SVM, etc.), build a* ***(4) new dataset Z(k,n), Y(k)*** *(by cross validation) to* ***(5)*** *train the* ***Super Learner model m(z,β)***

2nd ➔ *train each of the* ***n base learners*** *using training/validation split of {X,Y} and* ***compute predictions*** *(on top of validation set)* ***using meta-model m(z,β) (trained in step 1)***

GML Learning for NMA

- The **Super Learner meta-model** could be **whatever algorithm**
- The **original work** [Van der Laan'07] uses a **simple minimum square linear regression model** as the example **Super Learner**.
- ***Problem***: how to **define weights** to **perform** properly in **every dataset**?
- ***GML Learning***: computes **weights** with an **exponential probability of success, reducing** the **influence** of **poor base learning models**.

$$m_{GML}(X) = \sum_{j=1}^{J} w_j h_j(X)$$

base learners

$$w_j = \frac{e^{\lambda \alpha_j}}{\sum_{i=1}^{J} e^{\lambda \alpha_i}}$$

base learner accuracy

control variable: reduces weight for low accuracy predictors

Models Benchmarking

We compare several models for NMA:

- We take **5 standard base learning models**: linear **SVM, CART, *k*-NN, ANN (MLP)** and **Naïve Bayes**

  We **build 4 different Super Learners:**

  1. **Logistic regression** (binary output 0/1)
  2. **Weighted Majority Voting (MV):**
     - **MVuniform:** same weight to each base learner
     - **MVaccuracy:** weights are computed using base learner accuracy
  3. **Decision Tree meta-learner** (CART)

- **Boosting** (ensemble learning)**: AdaBoost tree**
- **Bagging** (ensemble learning): **Bagging tree** and **Random Forest**
- ***GML Learning***

Multiple NMA Problems

Five network measurement problems for model benchmarking:

1. **NS –** ***detection of network attacks*** in **WIDE/MAWI traffic** (transpacific links)
2. **AD –** ***detection of smartphone-apps anomalies*** in cellular networks (data captured at core cellular network)
3. **QoE-P –** ***QoE prediction*** in cellular networks (data captured at smartphones)
4. **QoE-M –** ***QoE-modeling for video streaming*** (smartphones public datasets)
5. **PPC –** ***Internet-paths*** dynamics tracking – ***prediction of path changes*** (M-Lab traceroute measurements)

(some) Evaluation Datasets

- We focus on two NMA problems:
  - Detection of **Network Attacks** in **WIDE/MAWI** network traffic
  - Detection of **App-related Anomalies** in an **Operational Cellular Network**

- **WIDE Network traffic using MAWI labels**
  - traffic traces captured daily on backbone link be
  - **MAWI labels**: uses a combination of four traditi
  - label the collected traffic by h
  - 5 attack classes: **DDoS, flashc**
  - The dataset spans a **full week**
    traces are split in **consecutive**
  - **245 features** describe the traf
  - These include throughput, pac
    addresses and ports, transpor
    (empirical distributions, samp

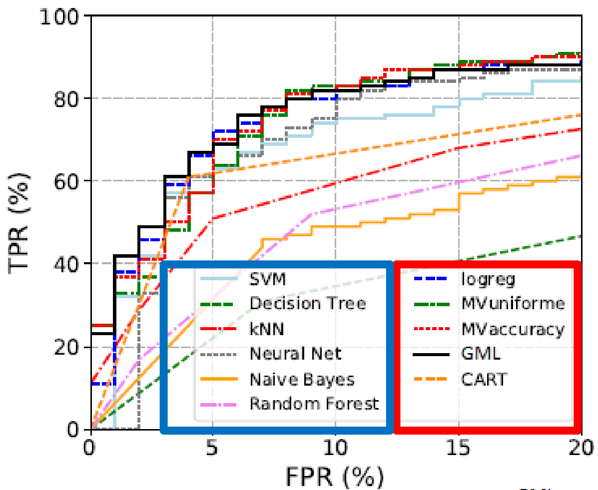
- **Synthetically generated dataset for AD in cellular networks**
  - **derived from real cellular ISP measurements** (traffic measurements collected during 6-months in 2014)
  - **Anomaly Templates**: derived f
    the cellular traffic → in this pa
  - **Evaluation labelled dataset**: 1
    different anomaly instances o
    intensity (number of involved
  - **36 features** describing 10' tim
  - These include FQDNs, DNS err
    manufacturer (empirical distri
    and more

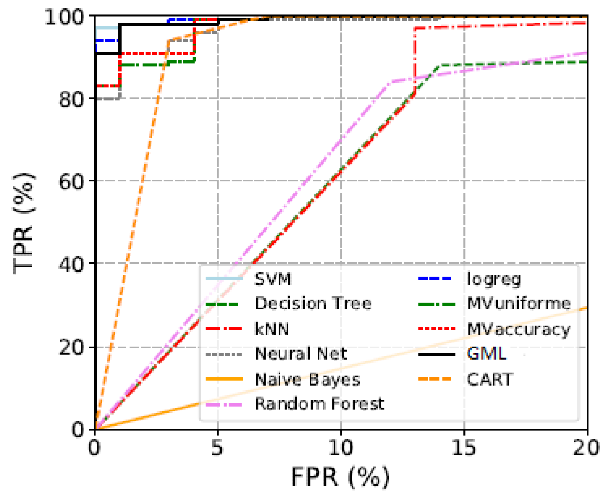
Field	Feature	Description
Tot volume	# pkts	num. packets
	# bytes	num. bytes
PKT size	pkt_h	H(PKT)
	pkt_{min,avg,max,std}	min/max/std, $\overline{\text{PKT}}$
	pkt_p{1,2,5,...95,97,99}	percentiles
IP Proto	# ip_protocols	num. diff. IP protocols
	ipp_h	H(IPP)
	ipp_{min,avg,max,std}	min/max/std, $\overline{\text{IPP}}$

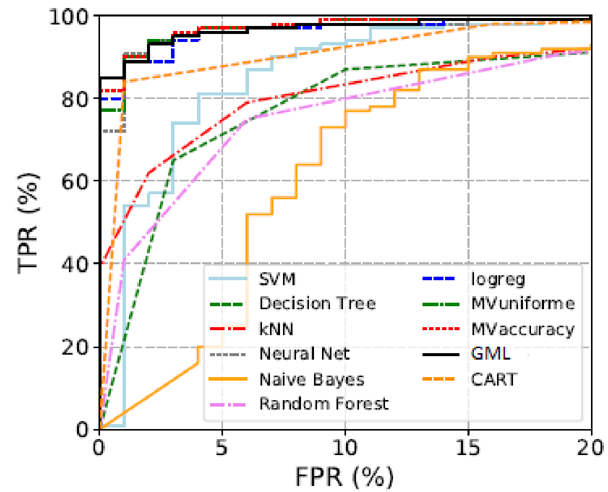
Type	E_1	E_2	E_3
Start time t_1	9:00	13:00	18:00
Duration d	2h	1 day	1h
Involved devices D	10%	5%	3%
Back-off time	5 sec	180 sec	20 sec
Manufacturer	single popular	multiple	multiple
OS	single	single	multiple
Error flag	+5% timeout	—	—
FQDN	top-2LD	top-2LD	top-2LD

Table III
ANOMALOUS DNS TRAFFIC FEATURES FOR TYPES E_1, E_2, E_3.

Benchmark for Network Security







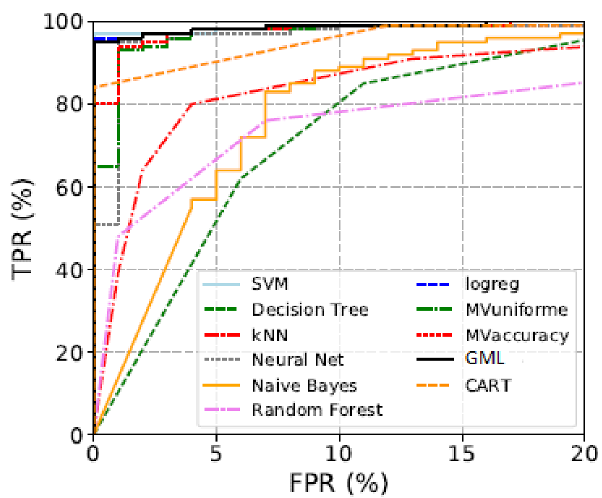
GML

Base Learners ***Super Learners***

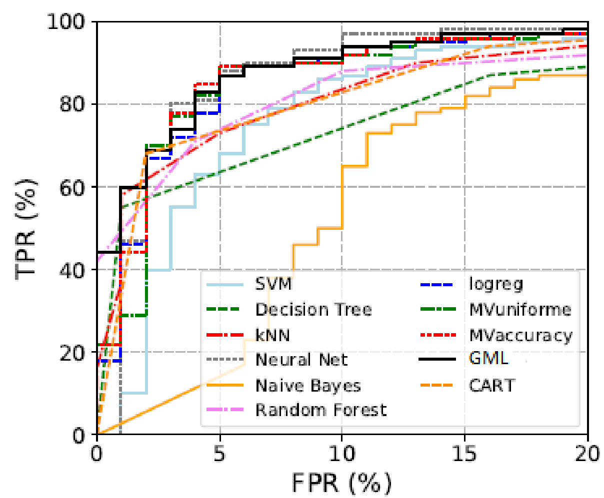
(a) DDoS.

(b) HTTP Flashcrowd (MPTP-la).

(c) Ping Flood.

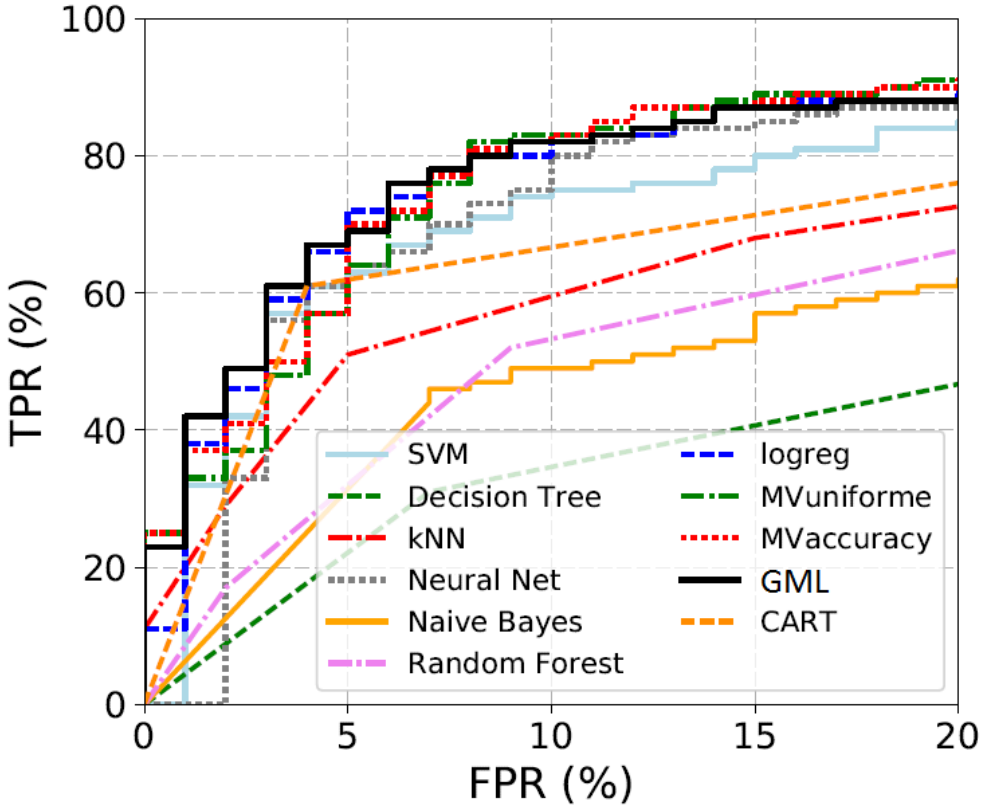


(d) Netscan UDP.



(e) Netscan TCP-ACK.

Benchmark for Network Security



(a) DDoS.

- **Super Learners (SLs) outperform** both **base learners**, as well as the **RF model**
- The CART SL performs the worst → **regression-based models are more accurate for SL**
- **GML slightly outperforms other SLs**

Benchmark for Network Security

	DDoS
CART	0.745
Naïve Bayes	0.730
MLP	0.907
SVM	0.883
kNN	0.720
Random Forest	0.827
Bagging Tree	0.823
AdaBoost Tree	0.892
logreg	0.926
MVaccuracy	0.924
MVuniform	0.923
CART	0.867
GML	**0.935**

- We take the **Area Under the ROC Curve** (AUC) as **benchmarking metric**
- **SLs performance increase is higher when base learners perform worse**
- Even if slightly, the **GML model systematically outperforms other models**

Benchmark for Anomaly Detection

	E1	E2	E3
CART	0.993	0.873	0.978
Naïve Bayes	0.956	0.861	0.959
MLP	0.997	0.944	0.996
SVM	0.996	0.944	0.995
kNN	0.995	0.859	0.963
Random Forest	**0.999**	0.876	0.993
Bagging Tree	0.996	0.885	0.983
AdaBoost Tree	0.998	0.945	0.995
logreg	**0.999**	0.952	0.996
MVaccuracy	**0.999**	0.948	0.996
MVuniform	**0.999**	0.945	0.996
CART	0.997	0.924	0.994
GML	**0.999**	**0.963**	**0.997**

- Similar observations are drawn from the AD benchmark
- **Anomalies E1 and E3 are easier to detect,** and **base learners** provide already **very accurate** results
- **E2 anomalies are stealthier** (long duration, small volume), and ***GML provides a clear performance increase***

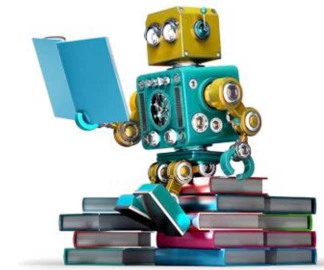
Full Benchmark in multiple NMA Problems

	AD	NS	QoE-P	QoE-M	PPC	ALL
CART	0.948 (**3.9%**)	0.872 (**11.1%**)	0.956 (**3.7%**)	0.952 (**4.4%**)	0.966 (**1.9%**)	0.935 (**5.4%**)
Naïve Bayes	0.925 (**6.2%**)	0.826 (**15.8%**)	0.752 (**24.2%**)	0.754 (**24.3%**)	0.924 (**6.3%**)	0.819 (**17.1%**)
MLP	0.979 (**0.7%**)	0.970 (**1.1%**)	0.887 (**10.7%**)	0.882 (**11.5%**)	0.964 (**2.1%**)	0.929 (**6.0%**)
SVM	0.978 (**0.8%**)	0.955 (**2.6%**)	0.786 (**20.8%**)	0.790 (**20.7%**)	0.886 (**10.1%**)	0.869 (**12.1%**)
kNN	0.939 (**4.8%**)	0.892 (**9.1%**)	0.788 (**20.6%**)	0.793 (**20.4%**)	0.920 (**6.7%**)	0.854 (**13.6%**)
Random Forest	0.956 (**3.1%**)	0.903 (**7.9%**)	0.983 (**1%**)	0.978 (**1.8%**)	0.969 (**1.6%**)	0.957 (**3.2%**)
Bagging Tree	0.954 (**3.2%**)	0.895 (**8.7%**)	0.976 (**1.7%**)	0.975 (**2.1%**)	0.973 (**1.3%**)	0.953 (**3.6%**)
AdaBoost Tree	0.979 (**0.7%**)	0.930 (**5.2%**)	0.982 (**1.1%**)	0.984 (**1.2%**)	0.875 (**11.2%**)	0.954 (**3.5%**)
logreg	0.982 (**0.4%**)	0.960 (**2.1%**)	0.981 (**1.1%**)	0.978 (**1.9%**)	0.941 (**4.5%**)	0.970 (**1.9%**)
MVaccuracy	0.981 (**0.5%**)	0.974 (**0.7%**)	0.984 (**0.9%**)	0.991 (**0.6%**)	0.972 (**1.3%**)	0.981 (**0.8%**)
MVuniform	0.980 (**0.6%**)	0.973 (**0.8%**)	0.980 (**1.3%**)	0.984 (**1.2%**)	0.980 (**0.5%**)	0.979 (**1.0%**)
CART	0.971 (**1.5%**)	0.946 (**3.6%**)	0.956 (**3.6%**)	0.960 (**3.6%**)	0.968 (**1.8%**)	0.959 (**3.0%**)
GML	**0.986**	**0.981**	**0.993**	**0.996**	**0.985**	**0.989**

- ***GML*** does not only **outperforms the most accurate first level learners**…
- …but **also outperforms other ensemble-learning models** based on bagging, boosting and stacking
- The **GML model performs the best for all scenarios**, suggesting a potentially good approach to go for by default in similar NMA problems

Organization of the Talk

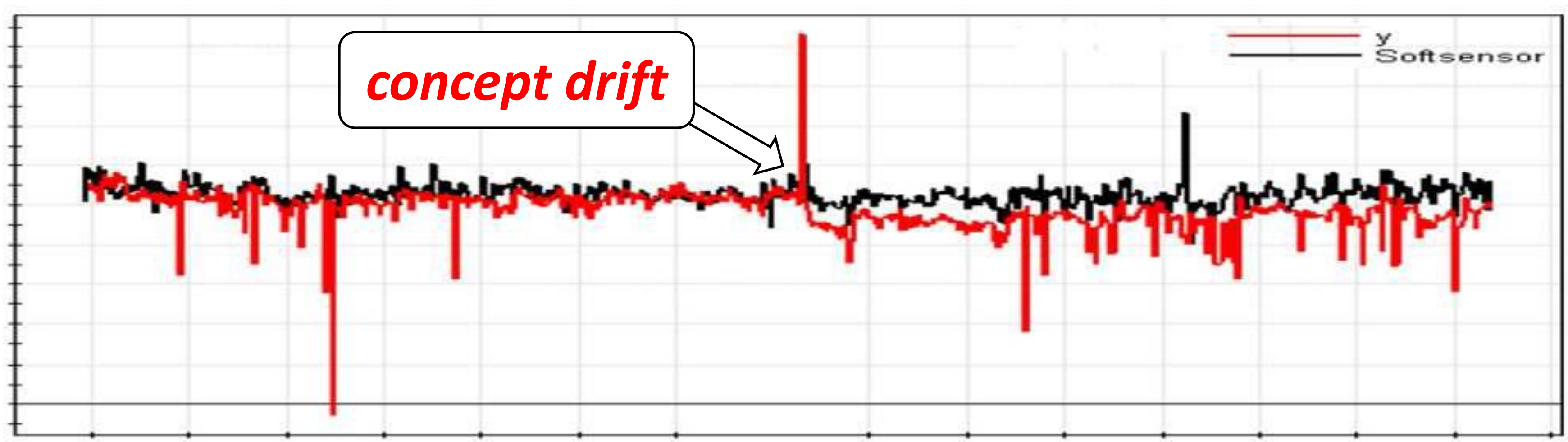
Dealing with Some of the Challenges in AI4NETS



- Deep Learning for Malware Detection – ***Avoid Feature Engineering***
- Generative Models for Anomaly Detection – ***Avoid Traffic Modeling***
- Explainable Artificial Intelligence (XAI) **–** ***Interpret Model Decisions***
- Super Learning for Network Security – ***Avoid Model Decision***
- **Adaptive/Stream Learning for NetSec** – ***Deal with Concept Drifts***
- Reinforced Active Learning – ***Deal with Lack of Ground Truth***

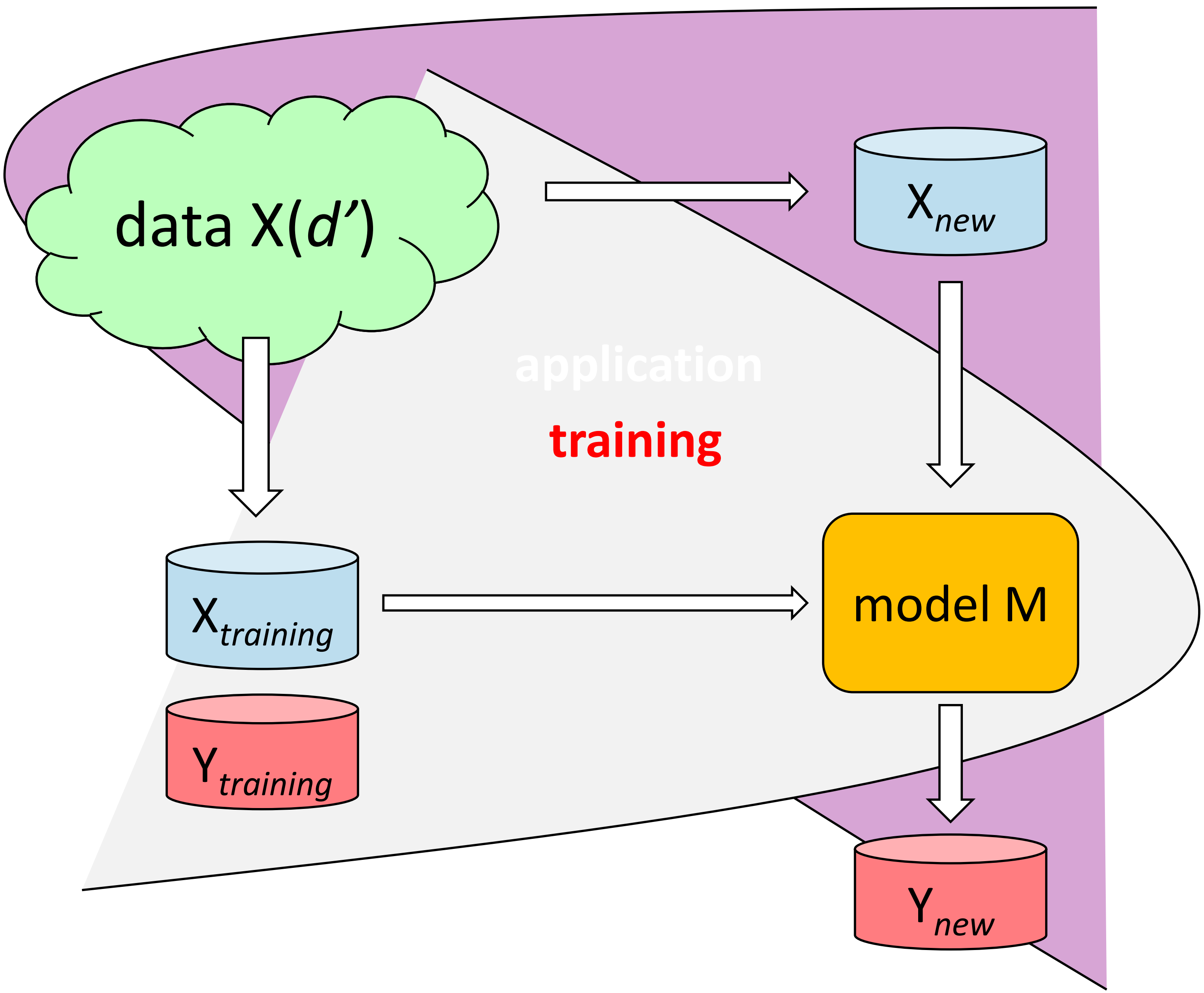
Adaptive or Stream-based Learning (*credits to Albert Bifet*)

- Let us go a bit deeper into the problem of **concept drift in supervised learning**
- And overview the main principles **how to deal with concept drift**



- ***Concept Drift (non-stationarity)***: the **statistical properties** defining the relationships between **input data and output target** **change over time**.
- This causes problems because the **predictions become less accurate as time passes**.

Concept Drift: a Trap for (off-line) Supervised Learning



training: learn a mapping function $\mathbf{Y}_{training} = \mathbf{M}\,(\mathbf{X}_{training})$

application: use learnt function/model on newly, unseen data $\mathbf{Y}_{new} = \mathbf{M}\,(\mathbf{X}_{new})$

…but what happens if/when $\mathbf{X}_{new}$ is **derived from a different distribution** $d' \neq d$?

(off-line) Supervised Learning under Concept Drifts

- Detection of **network attacks** in MAWI – WIDE network
- 10-fold cross-validation, **high detection performance with low FPR**…

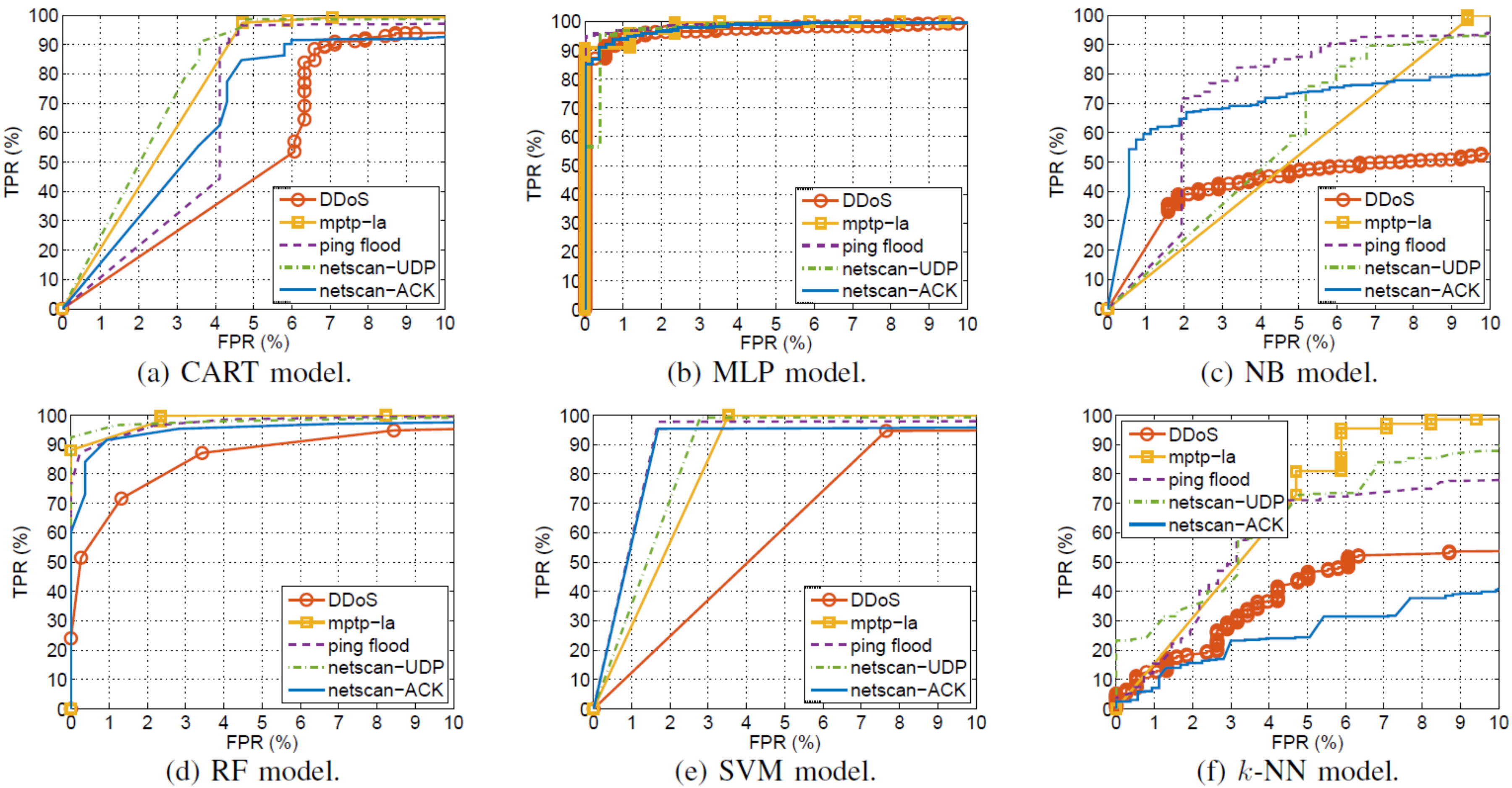
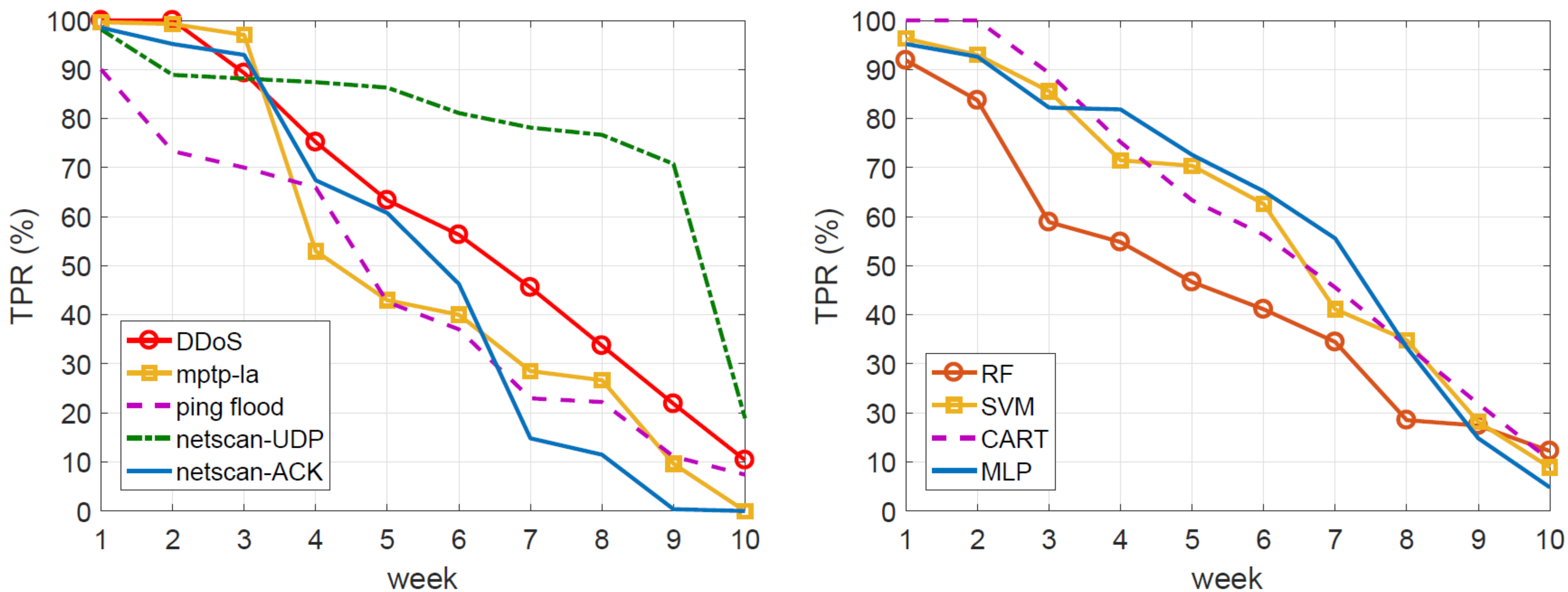


Figure 1: Detection performance (ROC curves) achieved by the different models for detection of network attacks.

(off-line) Supervised Learning under Concept Drifts

- …accuracy remains high for the first 3 weeks (**training on first 3 days**)…
- …but **models accuracy start to rapidly degrade over time**

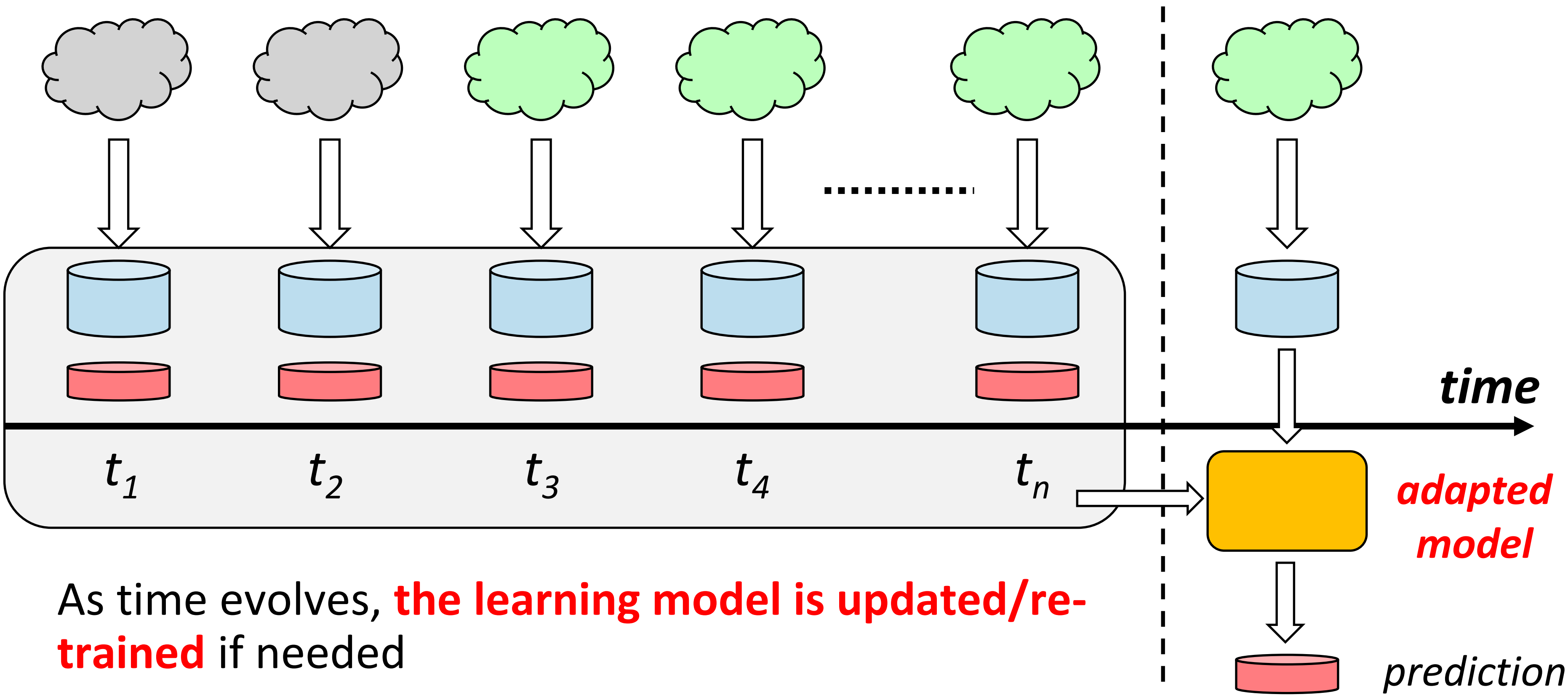


(a) CART model. (b) Detection of DDoS attacks.

Figure 2: Performance drift for the off-line trained models along time. Training is done on the first 3 days of data.

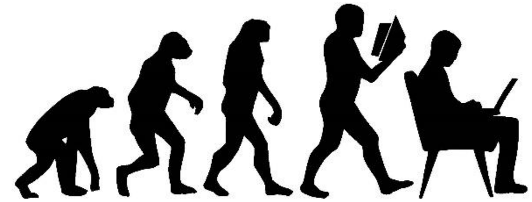
Learning in an Online Setting – Stream/Adaptive Learning

- In an **online setting**, **data** arrives continuously, as a **stream of samples**
- **Adaptive learning** consists of **learning from continuous data** in efficient way, using a **limited amount of memory**
- Adaptive learning approaches work in a **limited amount of time**

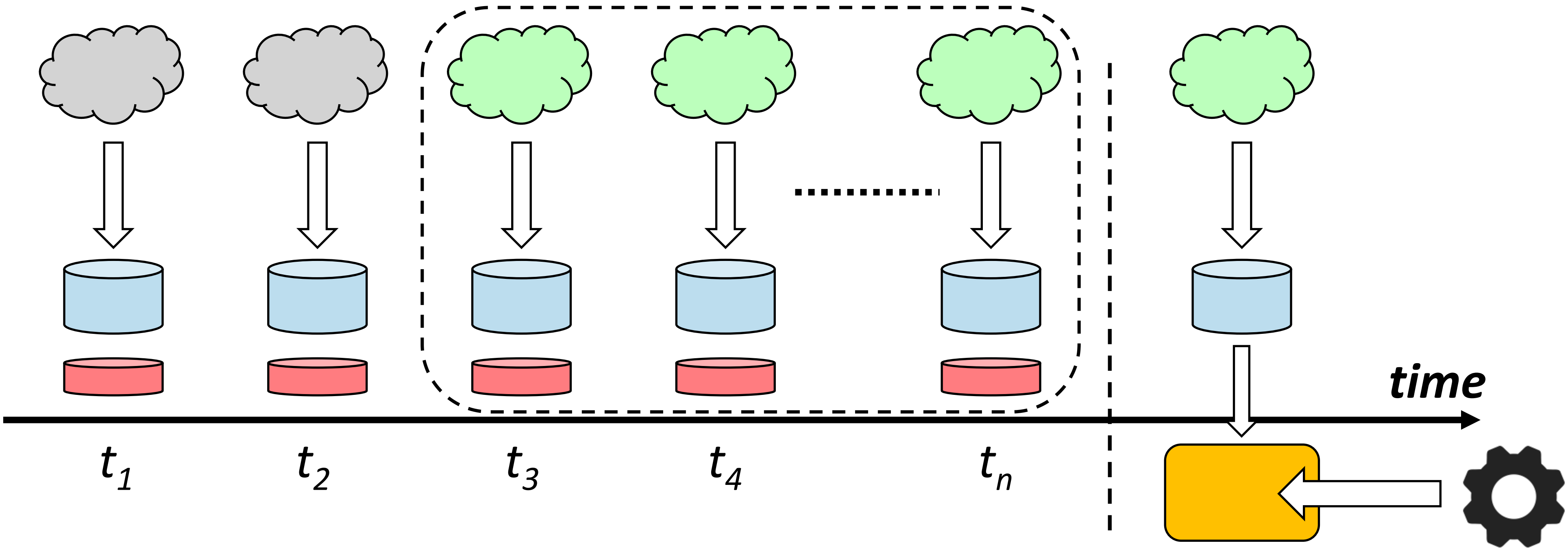


As time evolves, **the learning model is updated/re-trained** if needed

Adaptation Strategies



- Two main approaches for adaptation:
  - **re-train the model** by carefully selecting the ***best*** data
  - **adjust** the previously learnt **model incrementally**

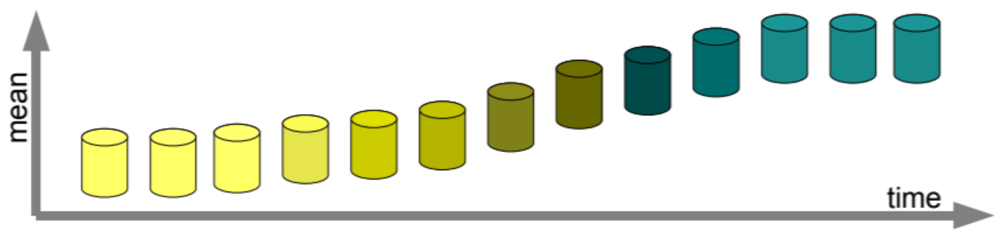


Desired Properties of a System to Handle Concept Drift

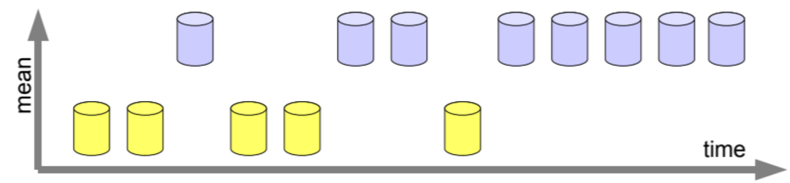
- Adapt **fast** to concept drift
- **Robust** to noise, but **adaptive to changes**
- Capable to **deal with reoccurring contexts** (avoid catastrophic forgetting)
- **Use limited resources in terms of time and memory**

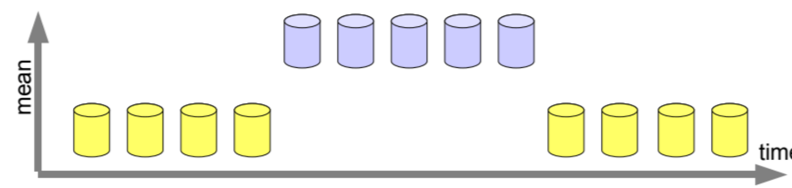
What types of Concept Drift can we get?

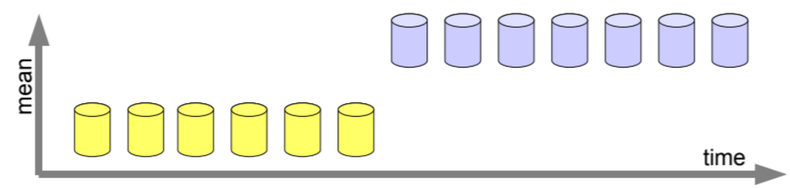
- The **change** to the data **could take any form**
- It is **conceptually easier** to consider the case where there is some **temporal consistency** to the change
- **Incremental drift:** one could assume that data collected within a specific time period show the same relationship and that this **changes smoothly over time**



- But of course, other types of changes may include:
  - A **gradual drift** over time
  - A **recurring** or cyclical **drift**
  - A sudden or **abrupt drift**







Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

strategy

memory

change detection and follow up
triggering

adapt at every step
evolving

reactive forgetting
single model

ensemble
maintain memory

Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

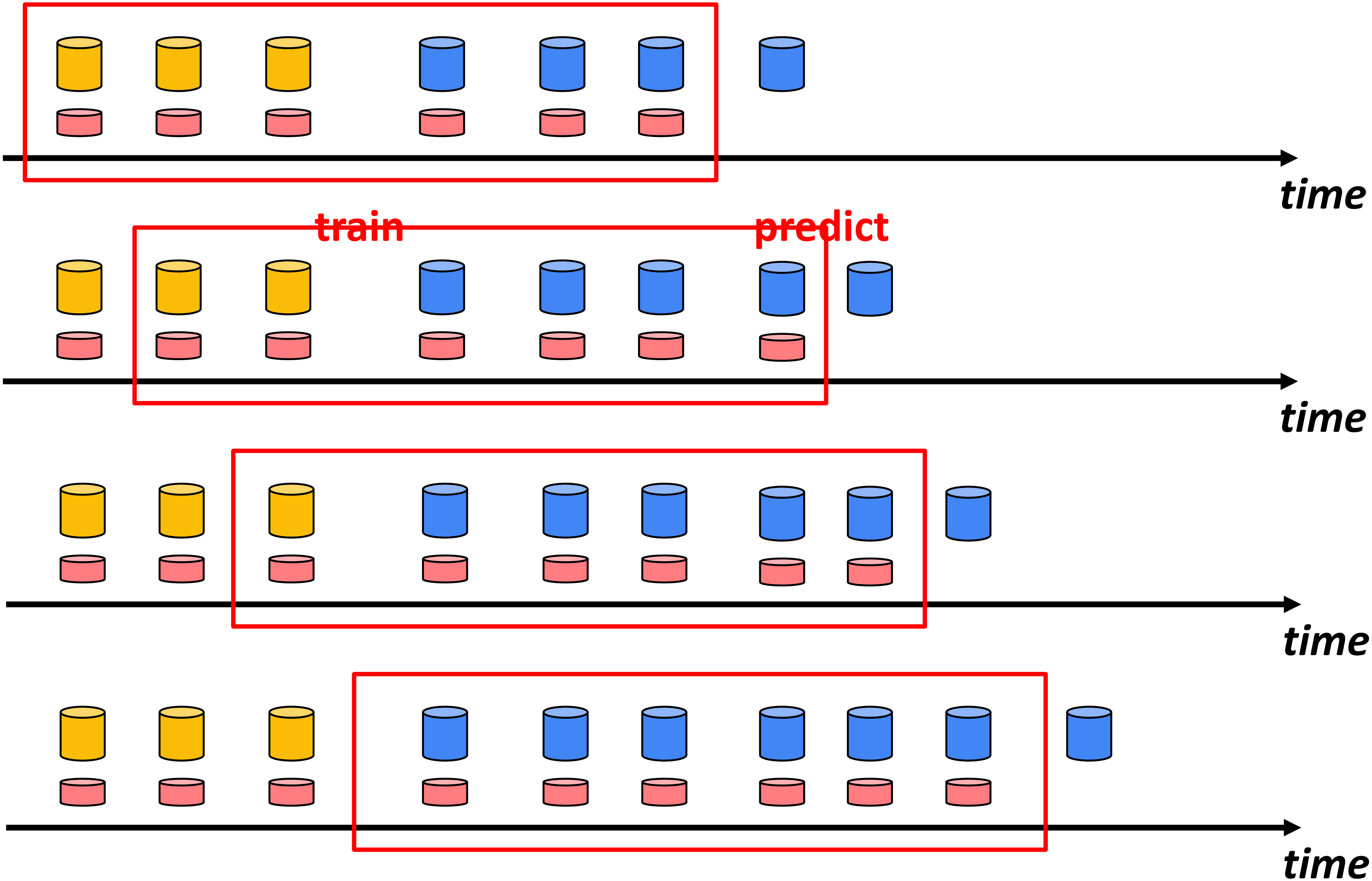
memory \ strategy	triggering	evolving
single model		
ensemble		

Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

strategy / memory	triggering	evolving
single model		**forgetting** • forget old data • re-train at fixed rate • fixed windows • instance weighting
ensemble		

Fixed-size Training Window

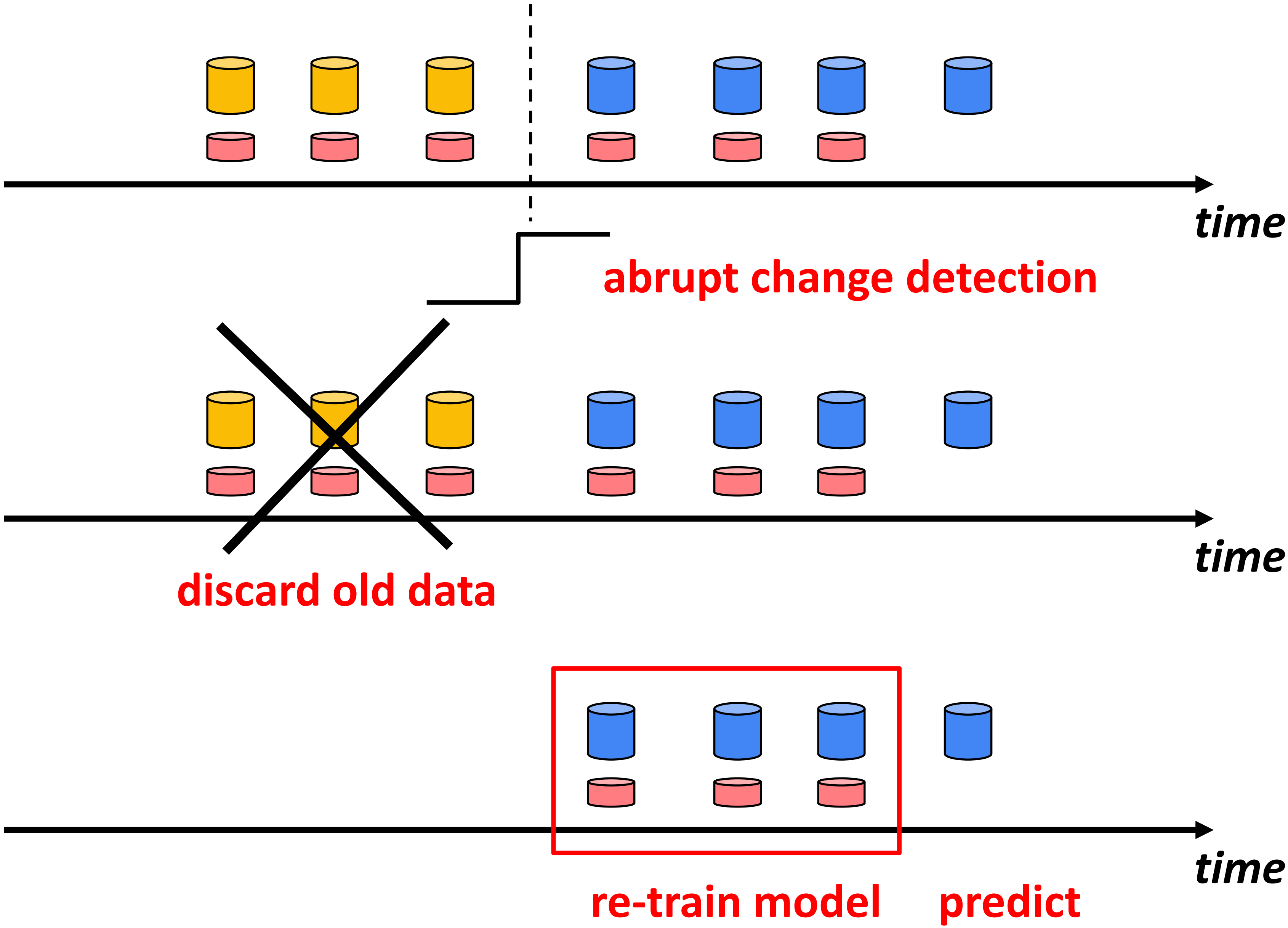


Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

strategy / memory	triggering	evolving
single model	**detectors** • **detect a change and discard the past** • **variable windows**	
ensemble		

Variable Training Window, Change Detection and Cut

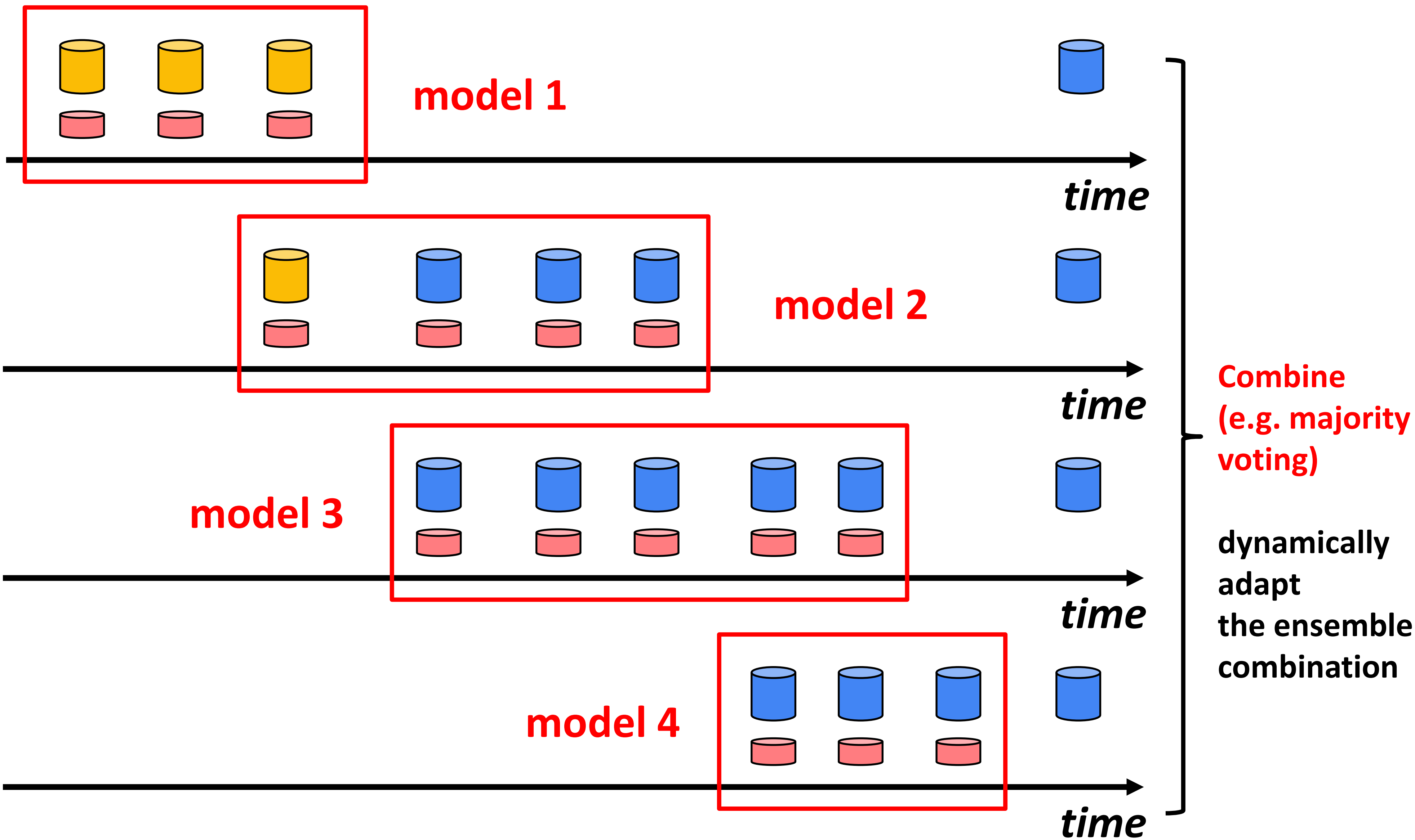


Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

strategy / memory	triggering	evolving
single model		
ensemble		**dynamic ensemble** • **build many models** • **dynamically combine** • **dynamic combination rules**

Dynamic Ensemble Learning

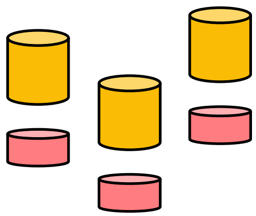


Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

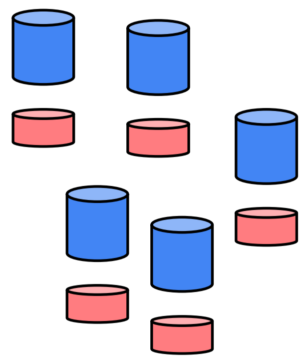
strategy / memory	triggering	evolving
single model		
ensemble	**contextual** • **build many models** • **switch among them based on input** • **meta–learning**	

Contextual (Meta) Approaches

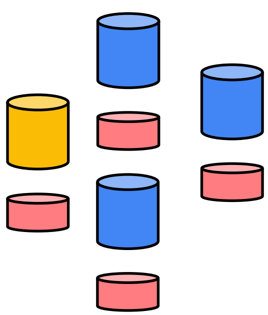


set 1 → model 1

partition training data to build multiple models

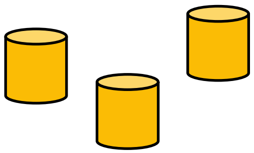


set 3 → model 3

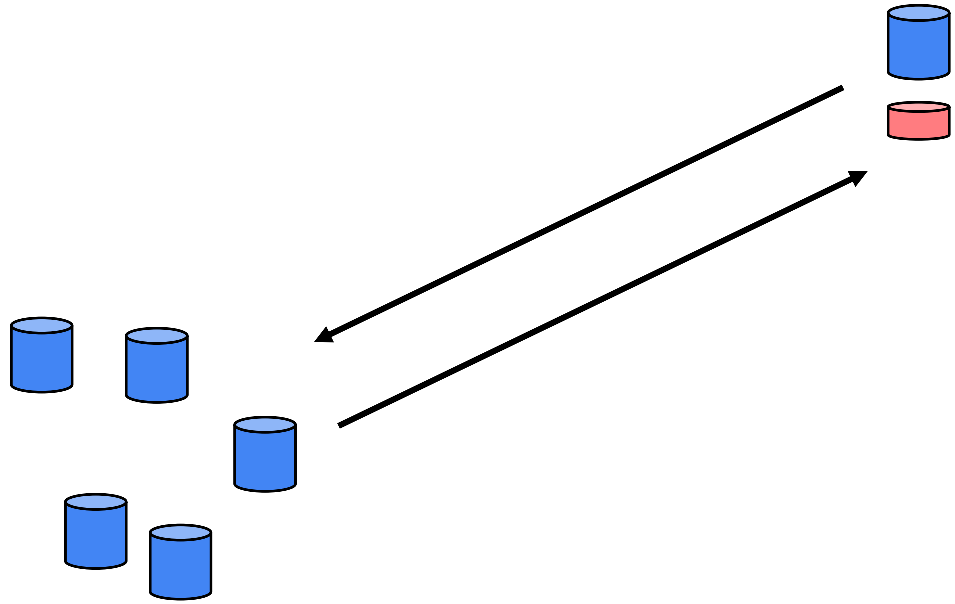


set 2 → model 2

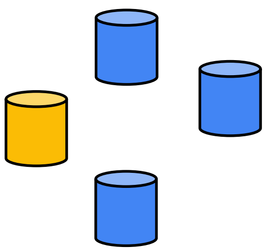
Contextual (Meta) Approaches



set 1 → model 1



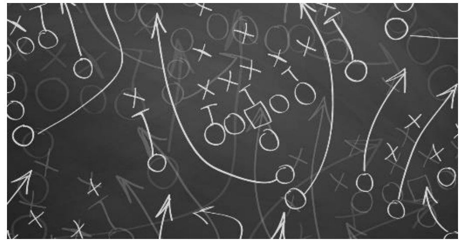
set 3 → model 3



set 2 → model 2

find which partition better represents the new instance, and use the corresponding model

Adaptation Strategies to Concept Drift



- A taxonomy of approaches (***A. Bifet, J. Gama***)

memory \ strategy	triggering	evolving
single model	**detectors** • detect a change and discard the past • variable windows	**forgetting** • forget old data • re-train at fixed rate • fixed windows • instance weighting
ensemble	**contextual** • build many models • switch among them based on input • meta–learning	**dynamic ensemble** • build many models • dynamically combine • dynamic combination rules

Adaptation Strategies to Concept Drift

- A taxonomy of approaches (***A. Bifet, J. Gama***)

memory \ strategy	triggering	evolving
single model	**detectors** **abrupt drift**	**forgetting** **abrupt drift**
ensemble	**contextual** **recurring drift**	**dynamic ensemble** **gradual drift**

Adaptive/Stream Learning Models for NetSec

- Implement an adaptive approach using ***single models*** and a ***change-detection*** algorithm to detect concept drifts

- Take ADWIN (**Adaptive WINdowing**) to detect changes

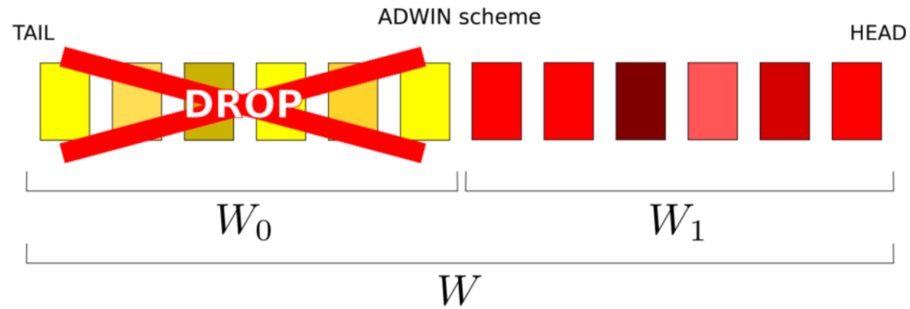
- ADWIN automatically **grows the learning window** when **no change** is apparent, and ***shrinks it when concept drifts*** are detected

- **Properties:** automatically adjusts its **window size** to the **optimum** balance point **between reaction time** and **small variance**

Adaptive WINdowing algorithm

The idea of ADWIN is straightforward:

- it keeps a sliding window W with the most recently observed data
- whenever two ***large enough*** sub-windows of W exhibit ***distinct enough*** averages, **the older portion of the window is dropped**.

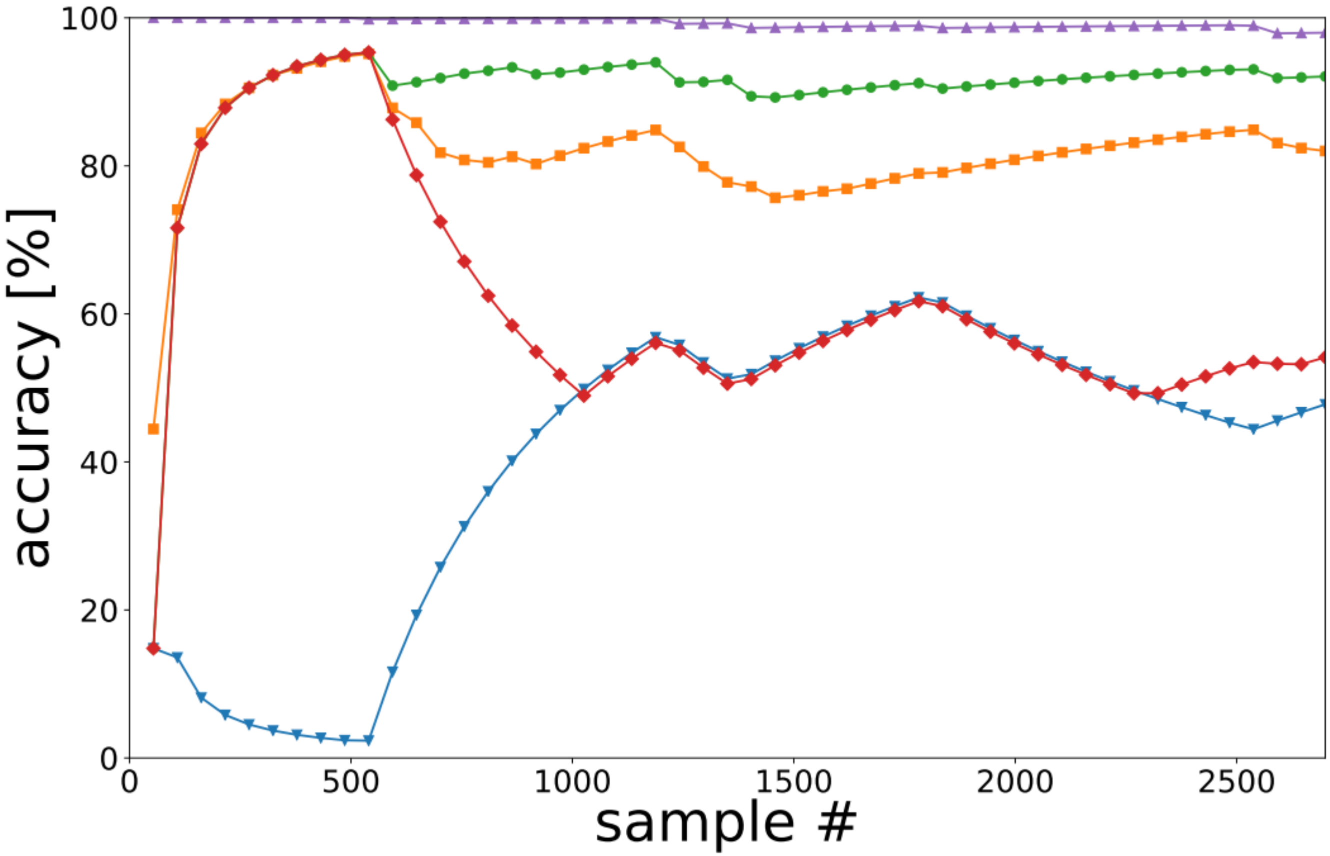


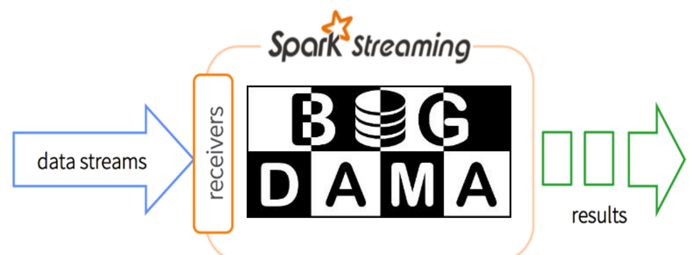
initialize window W
for each $t > 0$ **do**
    $W \leftarrow W \cup \{x_t\}$ (add x_t to the head of W)
    repeatdrop instances from the tail of W
    until $\|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}\| \geq \epsilon$ for every split of $W = W_0 \cdot W_1$
    return $\hat{\mu}_W$
end for

where $\hat{\mu}_{W_0}$ and $\hat{\mu}_{W_1}$ are the averages of the instances in W_0 and W_1 respectively.

Adaptive/Stream Learning Models for NetSec

- Adaptive learning algorithms **trained on labelled data, using ADWIN**





Stream-based Learning Models Performance

- Multiple stream machine learning models, using ADWIN
- Detection accuracy, ***normalized to batch-based algorithms*** performance

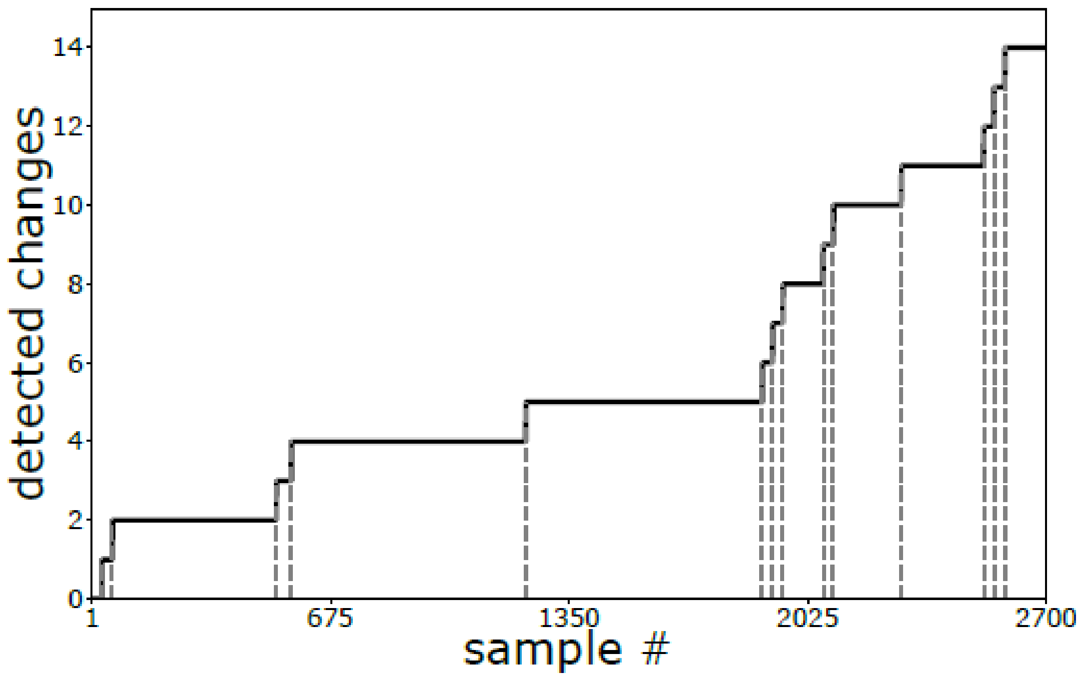
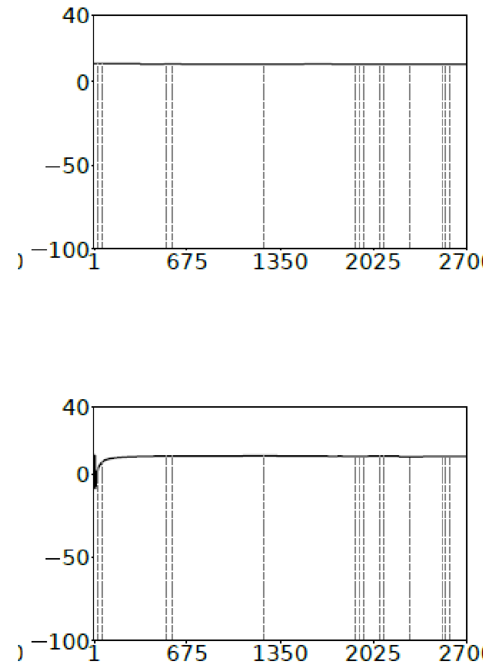


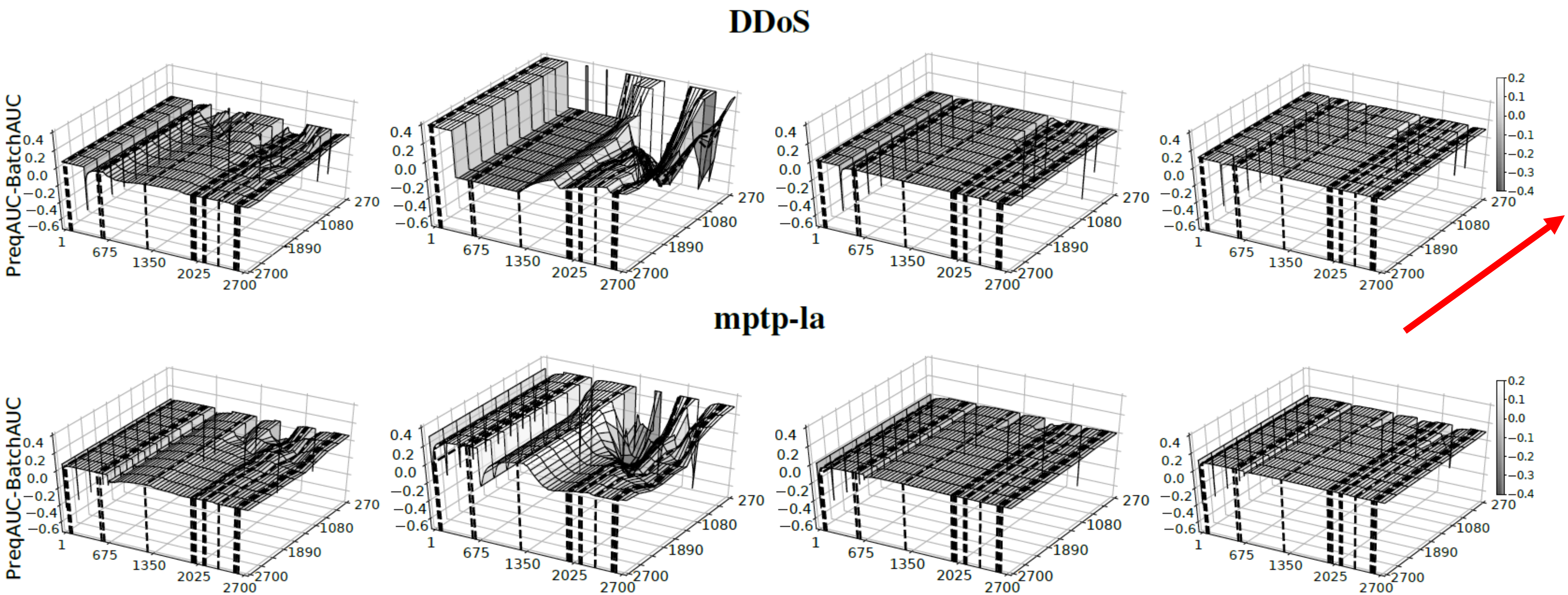
Figure 1: ***Page-Hinkley Concept Drift Detection.*** **Changes in the dataset distribution detected by the Page-Hinkley test. Detected changes are marked with dashed lines.**



(d) *SGD*

Stream-based Learning Models Performance

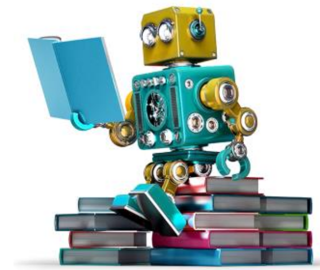
- Multiple stream machine learning models, using ***fixed windowing***
- AUC (ROC curve), ***normalized to batch-based algorithms*** performance
- **Different window sizes tested**



(a) *k-NN* (b) *HAT* (c) *ARF* (d) *SGD*

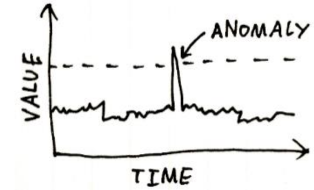
Organization of the Talk

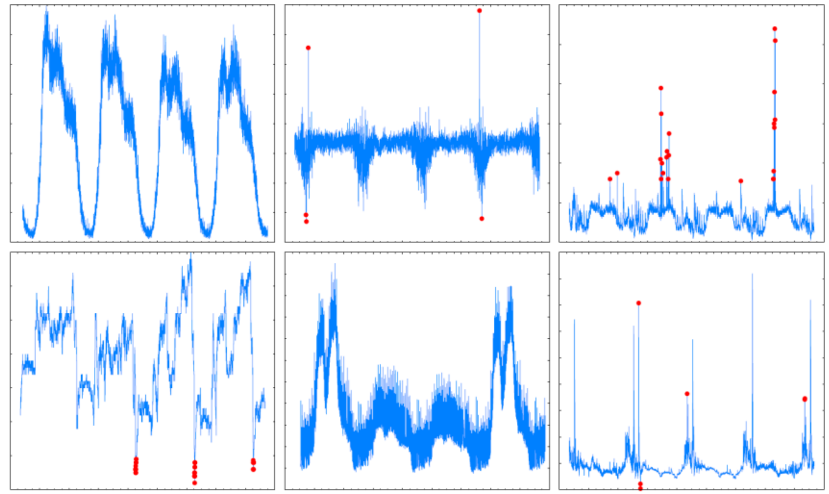
Dealing with Some of the Challenges in AI4NETS



- Deep Learning for Malware Detection – ***Avoid Feature Engineering***
- Generative Models for Anomaly Detection – ***Avoid Traffic Modeling***
- Explainable Artificial Intelligence (XAI) **–** ***Interpret Model Decisions***
- Super Learning for Network Security – ***Avoid Model Decision***
- Adaptive/Stream Learning for NetSec – ***Deal with Concept Drifts***
- **(Reinforced) Active Learning** – ***Deal with Lack of Ground Truth***

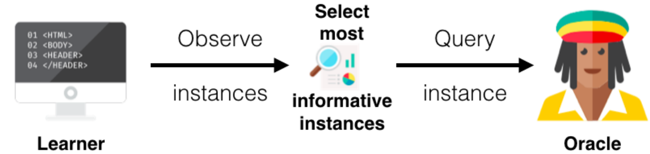
Time-Series Anomaly Detection and Active Learning





- **Anomaly Detection (AD)** is, by definition, an ***unsupervised process*** (detect what is different from the majority)
- The ***limitation of AD*** is on the **high false alarm rates** (difficult to construct and track the baseline)
- **Supervised Learning** provides **better accuracy and lower false alarm rates**, but ***requires large amounts of labeled data***….
- ….and **in the case of anomaly detection**, it has to ***deal with heavily unbalanced classes and concept drifts***
- Solution: **semi-supervised learning**, relying on ***human labeling*** (**through AL**) to tailor and ***improve detection performance*** over time

What is it and Why Active Learning?

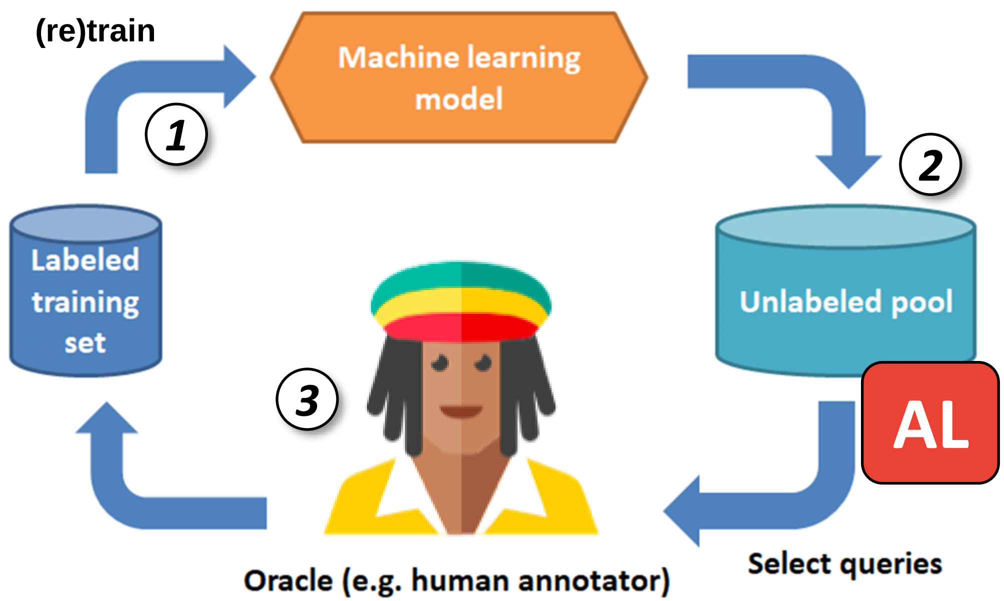


- Lets assume we want to ***build a "small" labeled dataset*** for semi-supervised learning → ***e.g., by manually labeling some samples***
- ***How to cherry-pick the best samples to label?*** → ***Active Learning (AL)***
- The ***learning algorithm itself chooses*** the data it wants to learn from
- Using AL, learning models ***can perform as good as*** (***or better** than*) ***traditional methods with substantially less data for training***

General AL step-by-step

1. ***Bootstrapping:*** train an initial model ***M*** from a small labeled training dataset D_0
2. ***Sample selection:*** apply ***M*** to unlabeled data, and ***select*** best ***samples to label***, based on a ***query strategy*** (e.g., model uncertainty) → D_i
3. ***Oracle labelling & retraining:*** label selected samples and retrain model ***M with*** $\{D_0 + D_i\}$

Stopping criteria: query budget, model performance, number iterations, YNI…



ALADDIN – Building Blocks

- Detection Model → ***RCET 100*** (Random Committee of 100 Extremely Randomized Trees), or simple ***RF100*** model

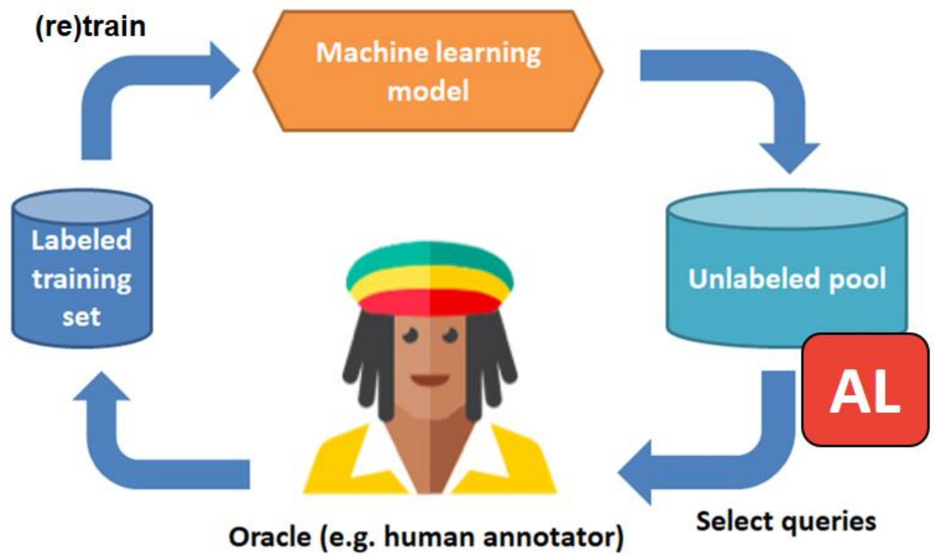
- AL Sampling – Pool based, using ***uncertainty sampling*** – in particular keep the ***least confident samples***, with a **predefined querying budget**

- Required ML model output: **per-class prediction probabilities** (could also plug-in other relevant info, e.g., anomaly scores)
    - mean terminal leaf probability across all trees
    - fraction of trees voting either class
    - softmax activation layer

- ***Labeled training set (bootstrap model)***

- ***Unlabeled pool (AL operation)***

- ***Testing set***



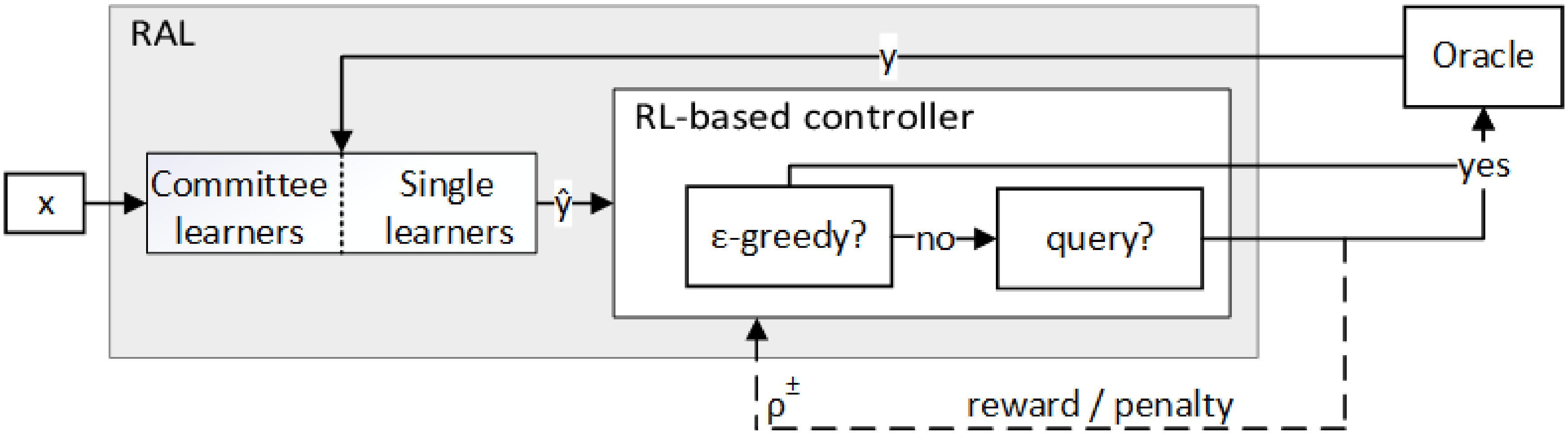
Improving Stream-based Active Learning by Reinforcement (RAL)

- How do we deal with the **limited amount of labeled data?**
- **Active Learning (AL):** aims at labelling only the most informative samples
- **AL** can be applied to the **streaming scenario**, to complement previous approaches and **reduce the amount of labeled data**

- RAL – improves stream-based AL by Reinforcement Learning (RL)
  - AL bases its decisions based EXCLUSUVELY on model uncertainty
  - **RAL permits to additionally learn in a feedback loop**, based on the **effectiveness of the requested labels**
  - **Reward** in case asking oracle was informative (models would have predicted wrong label)
  - **Penalty** otherwise

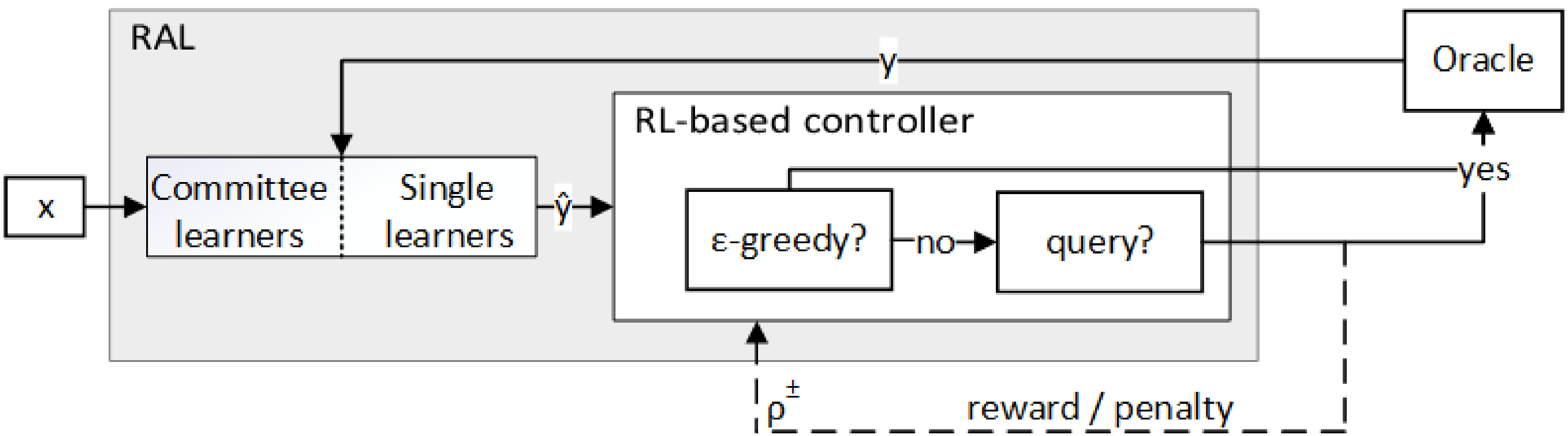
RAL Principles and Components

- RAL is based on an ensemble of models
- RAL makes use of **contextual-bandit algorithms** (EXP4) to tune the decision powers of the different models depending on their behavior
- RAL uses a **ε-greedy approach** to handle concept drift and **improve the exploration/exploitation trade-off**



RAL Principles and Components

- The **querying decision** (ask or not for a label) is taken **based on model prediction uncertainty** and a **threshold**
- Each algorithm in the ensemble (committee) gives its advice, based on its prediction uncertainty
- RAL takes into account the decisions of the members + their decision power
- Obtained **feedback influences the querying threshold**:
  - In case of **penalty**, the threshold decreases.....otherwise, it **slightly increases**



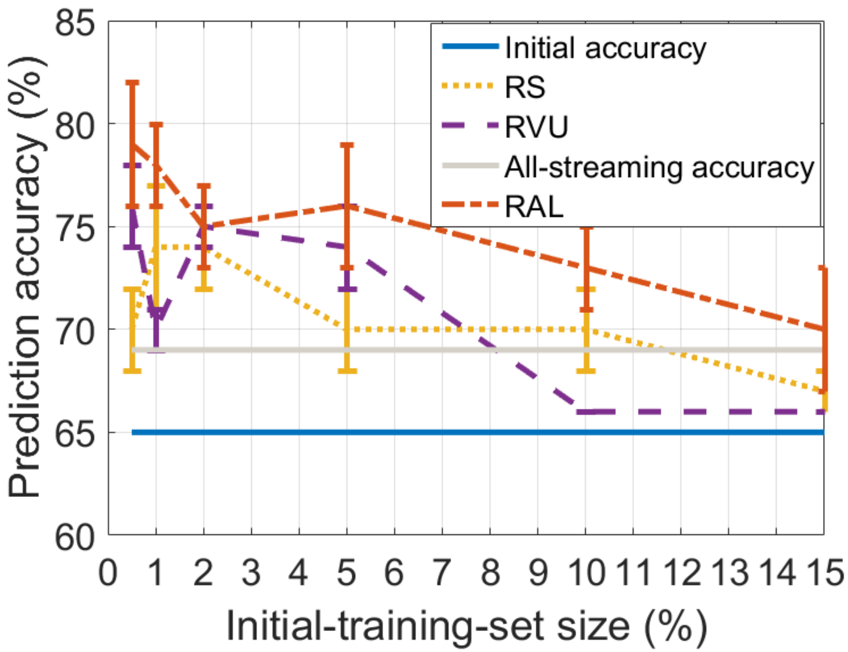
RAL Evaluation vs. State of the Art

- **RAL vs** RVU (**Randomized Variable Uncertainty**) and simple **random sampling** (RS)

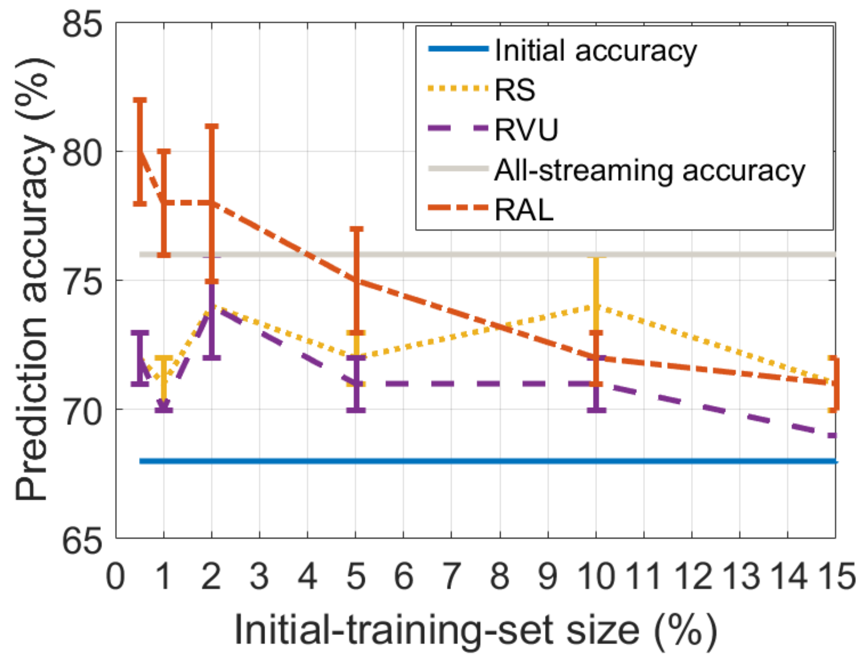
- Evaluation on data extracted from **MAWILab – in the wild network security**

- We divide each dataset into three consecutive parts:
  - **Initial training set** (variable size)
  - **Validation set** (last 30%), to evaluate the classifiers
  - **Streaming set** (remaining part of the dataset), for picking samples to learn from

RAL Evaluation vs. State of the Art – Prediction Accuracy

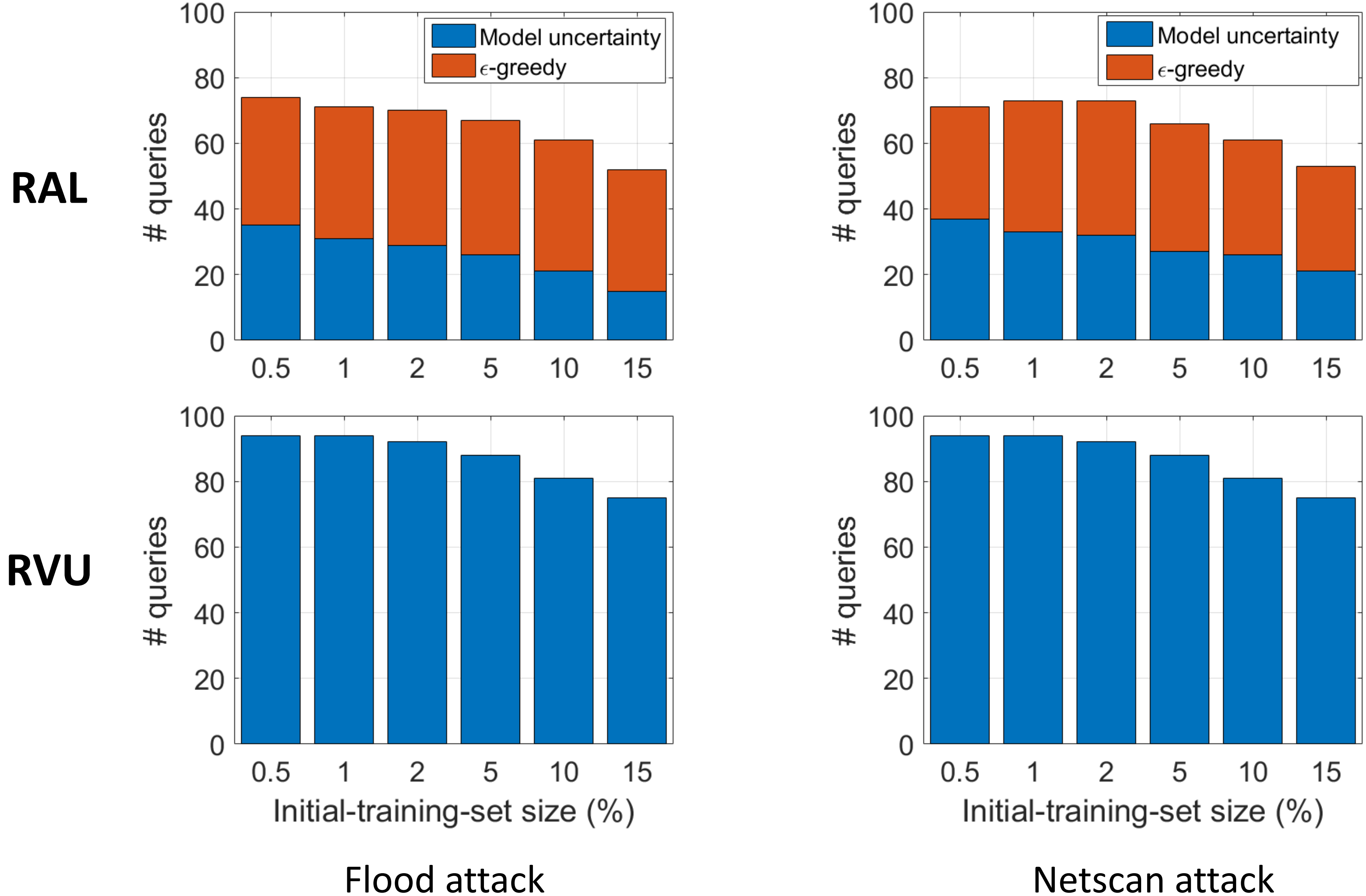


Flood attack



Netscan attack

RAL Evaluation vs. State of the Art – Querying Cost



So What's Next?

- We're still far from ***making AI immediately applicable***
  - Limitations of learning process, data, models
  - Lack of generalization
  - Continual learning challenges – catastrophic forgetting and transfer
  - Lack of real knowledge generation – building simple mappings is *easy*
  - Portability of models to real deployments – *plug & play?*
- ***Effective Machine Learning*** – a mix of interesting challenges:
  - Transfer learning
  - Explainable AI (XAI)
  - Multi-task learning
  - Meta learning and hierarchical learning
  - Graph Neural Networks
- And back right to the start: **the successful application of AI to network measurement problems is still on an early stage**

Thanks

Dr. Pedro Casas

Data Science & Artificial Intelligence

AIT Austrian Institute of Technology @Vienna

pedro.casas@ait.ac.at

http://pcasas.info