8. Statistical and Sequential Intrusion Detection

Simona Buchovecká, Tomáš Čejka

Faculty of Information Technology, CTU in Prague

simona.buchovecka@fit.cvut.cz, tomas.cejka@fit.cvut.cz

November 19, 2018

Introduction

- Based on network traffic (statistical) distribution
- The simplest way to build a statistical model: to compute the parameters of a probability density function for each known class of network traffic and test an unknown sample to determine which class it belongs to.
- Parametric vs Non-Parametric:

  Parametric assume knowledge of underlying distribution and estimate parameters from the given data

  Non-Parametric do not generally assume knowledge of the underlying distribution

- Batch vs Sequential processing:

  Batch process independent time slots

  Sequential process a data stream, time series

Non-statistical approach

Non-statistical features of network intrusions:

- Network protocols are deterministic and well understood
- Protocol anomalies can be detected by stateful analysis
- Many ad-hoc methods work well to detect various attacks

Example of a good deterministic ad-hoc detection rule

- Suspect a host is using P2P transfers if it:
  - uses network flows via port 6881
  - uses ports above 50000, with many port changes
  - connects to many IPs, most of them inaccessible
  - at the end many connection finish at the same time

Statistical Approach

Statistical features of network intrusions:

- Network intrusions occur randomly
- Intrusions occur at unknown points in time
- Intrusions lead to changes of statistical properties of some observable characteristics

Attack detection viewed as a change-point detection (CPD):

- Detect changes in the distributions (models, parameters)
- With fixed delays (batch-sequential approach)
- Or with minimal average delays (sequential approach)
- While maintaining the false alarm rate at a given level

Section 1

Change-Point Detection

Change-Point Detection Methodology

Observed sequence of random variables (or vectors): $X_1, X_2, ...$

- $X_1, X_2, ..., X_n$ represent some network characteristics observed at times $t_1, t_2, ...$
- Examples: numbers of deauthentication frames, numbers of failed connections, levels of link saturation etc.

A change in distribution occurs at an unknown index λ

- The change corresponds to a network traffic anomaly at time t_λ
- P_k and $\boldsymbol{E}_k$ denote the probability and the expectation when $\lambda = k$
- P_0 and $\boldsymbol{E}_0$ correspond to the pre-change and the no-change distribution

CPD Methodology

Sequential CPD procedure:

- Stopping time τ is the time of alarm
  (i.e., detection of a distribution change)
- Detection delay: $\mathrm{ADD}_\lambda(\tau) = \boldsymbol{E}_\lambda(\tau - \lambda | \tau \geq \lambda)$
- False alarm rate: $\mathrm{FAR}(\tau) = \frac{1}{\boldsymbol{E}_0(\tau)}$

Classical Optimal CPD Procedures

Conditional probability density function (pdf) for $X_1, X_2, ...$:

- Before change ($n < \lambda$): $p_0(X_n|X_1, ..., X_{n-1})$ (baseline distribution)
- After change ($n \geq \lambda$): $p_1(X_n|X_1, ..., X_{n-1})$ ("attack" scenario)

Log-likelihood ratio (LLR):

$$Z_{n,\lambda} = \sum_{k=\lambda}^{n} \log \frac{p1(X_k|X_1, ..., X_{k-1})}{p0(X_k|X_1, ..., X_{k-1})}$$

Classical Change-Point Detection

- Uses a threshold of a statistic based on LLR
- Detection when the statistic first exceeds a given threshold
- It is in fact testing hypotheses that the change occurred at the point λ versus that there is no change at all ($\lambda = \infty$)

Classical Optimal CPD Procedures

Shiryaev-Robers-Pollak (SR) procedure:

- Motivated by Bayesian considerations
- "Average LLR" Statistic:

$$R_n = \sum_{\lambda=1}^{n} \exp\{Z_{n,\lambda}\}$$

- Stopping time:

$$\tau_{\mathrm{SP}}(h) = \min\{n \geq 1 : \log R_n \geq h\}$$

Page's cumulative sum (CUSUM) procedure

- Motivated by a maximum likelihood argument
- Maximum LLR statistic:

$$\max_{1 \leq \lambda \leq n} Z_{n,\lambda}$$

- Stopping time.

$$\tau_{\mathrm{CU}}(h) = \min\{n \geq 1 : U_n \geq h\}$$

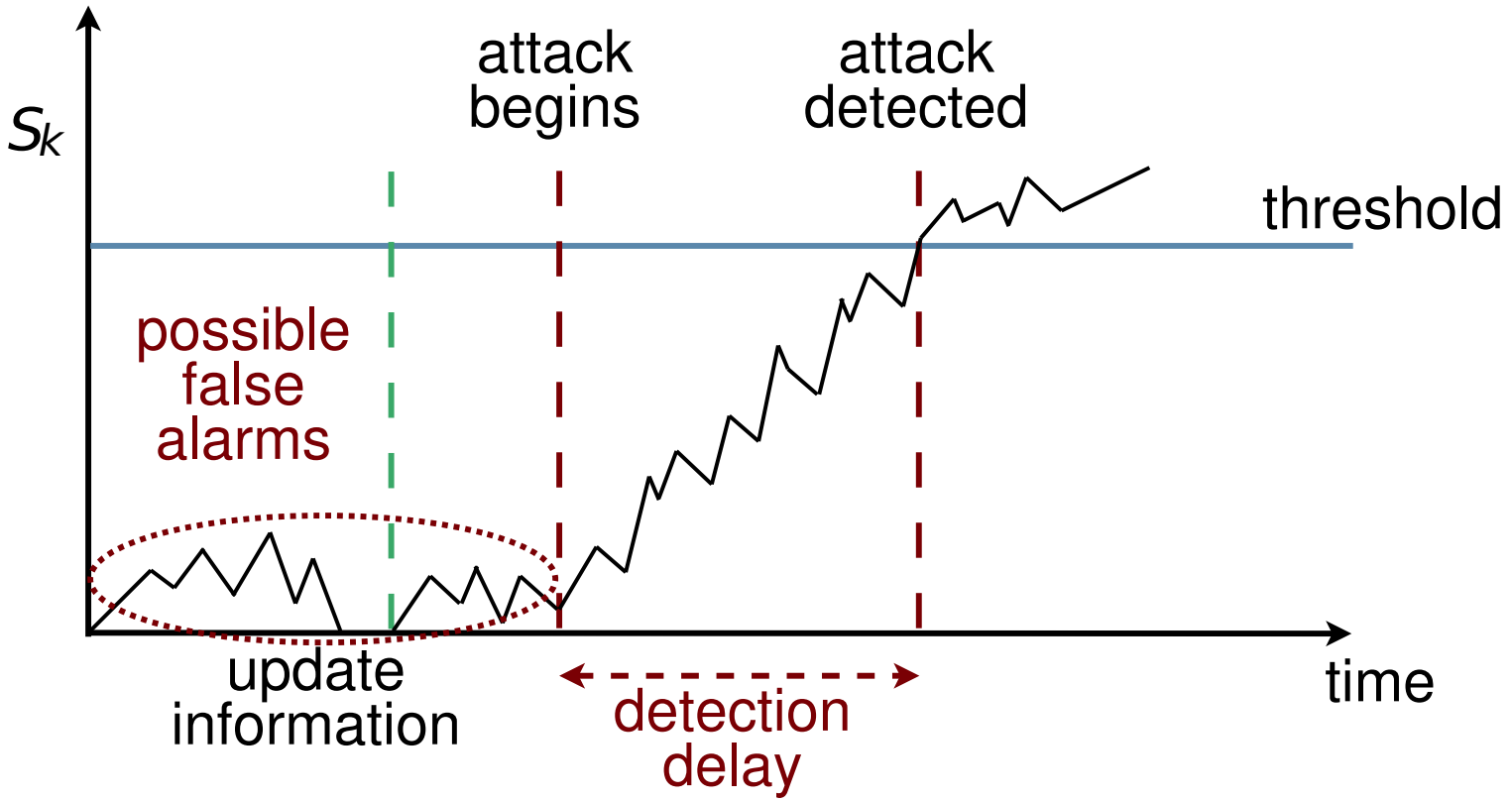
Classical Optimal CPD Procedures

If observations are i.i.d. (independent, identically distributed)

- Both methods minimize the worst-case average detection delay $\sup_\lambda \mathrm{ADD}_\lambda(\tau)$
- Minimization among all methods for which the FAR is fixed ($\mathrm{FAR}(\tau) \leq \varphi$)
- Thresholds should be chosen from the conditions $\boldsymbol{E}_0(\tau) = 1/\varphi$
- A preliminary threshold $h = \log(1/\varphi)$ guarantees $\mathrm{FAR}(\tau) \leq \varphi$ (for both methods)
- U_n can be replaced with

$$\tilde{U}_n = \max\left\{0, \tilde{U}_{n-1} + \log\frac{p_1(X_n)}{p_0(X_n)}\right\}$$

CPD Example


S_k
attack begins
attack detected
threshold
possible false alarms
update information
detection delay
time

Generalized CPD Procedures

The i.i.d. assumption is very restrictive for intrusion detection

- Network data are usually correlated and non-stationary, even bursty, due to substantial temporal variability
- Recent CPD advances: CUSUM and SR are also optimal for general statistical models when the FAR is low (φ is small).

Classical optimal CPD methods require complete prior knowledge of the pre-change and post-change distributions

Generalized CPD Procedures

Parametric modifications:

- Generalized likelihood ratio (LR), LR mixtures, and adaptive LR
- Do not solve the problem when the distributions are not known
- Procedures based on signs or ranks are not quite computationally efficient

Non-Parametric NP-CUSUM procedure (by Tartakovsky et al.):

- Inspired by the CUSUM statistic $\tilde{U}_n = \max\left\{0, \tilde{U}_{n-1} + \log \frac{p_1(X_n)}{p_0(X_n)}\right\}$
- Asymptotically optimal: ADD is nearly minimized for low FAR
- Has manageable computational complexity

Generalized Non-Parametric Sequential Statistical Detection

$$S_n = \max\left\{0, S_{n-1} + f_n(X_n)\right\}, S_0 = 0$$

If we knew the exact probabilistic models for the pre- and post-attack scenarios, then $f_n(X_n)$ would be the Log-Likelihood Ratio.

Non-Parametric Sequential Statistical Learning

$$S_n = \max\left\{0, S_{n-1} + X_n - \mu - \varepsilon\hat{\theta}_n\right\}, S_0 = 0,$$

where:

- X_n is an observed network characteristic in the n^{th} time interval (e.g., number of UDP, TCP SYN, ICMP, or ARP packets),
- μ is a historical estimate of $\boldsymbol{E}(X_n)$,
- ε is a tuning parameter,
- $\hat{\theta}_n$ is an estimate of $\boldsymbol{E}(X_n)$ under attack.

Section 2

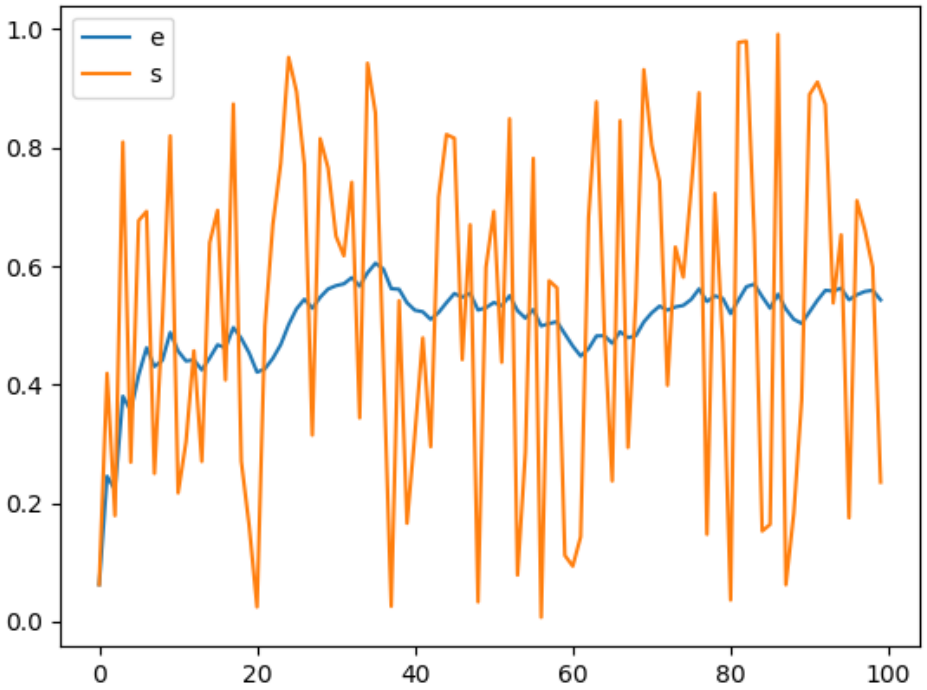
Smoothing and Predicting

Exponential Weighted Moving Average (EWMA)

- Short-term peaks might produce false alerts
- Smoothing of the observed characteristic

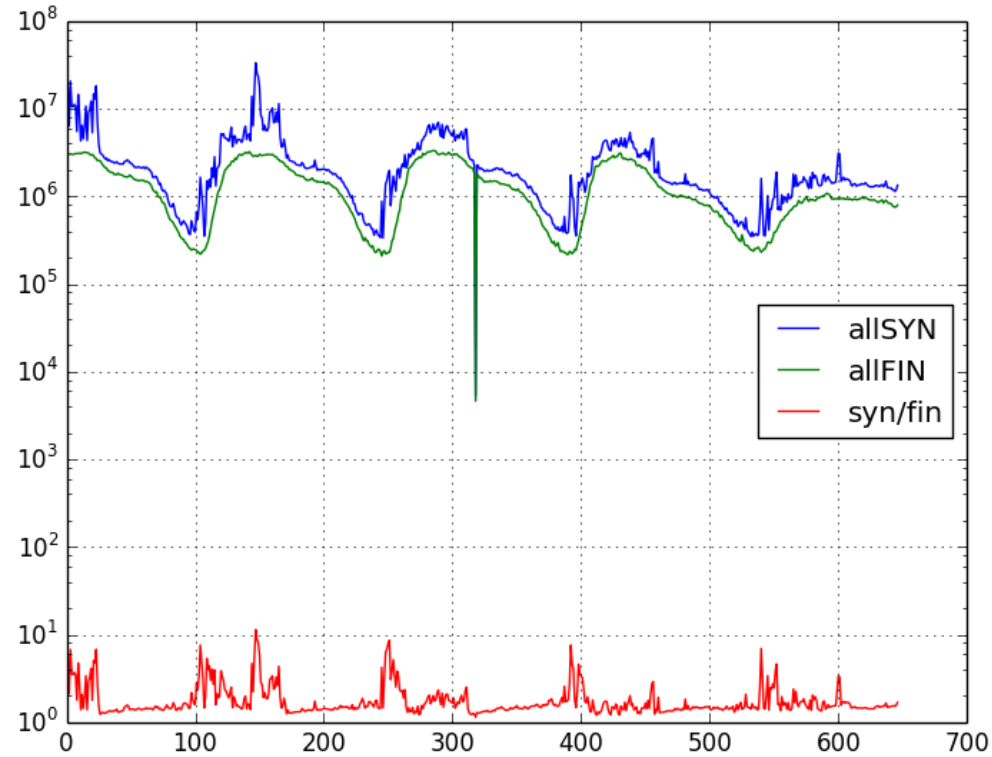
$$z_i = \alpha x_i + (1 - \alpha) z_{i-1}, z_0 = x_0,$$

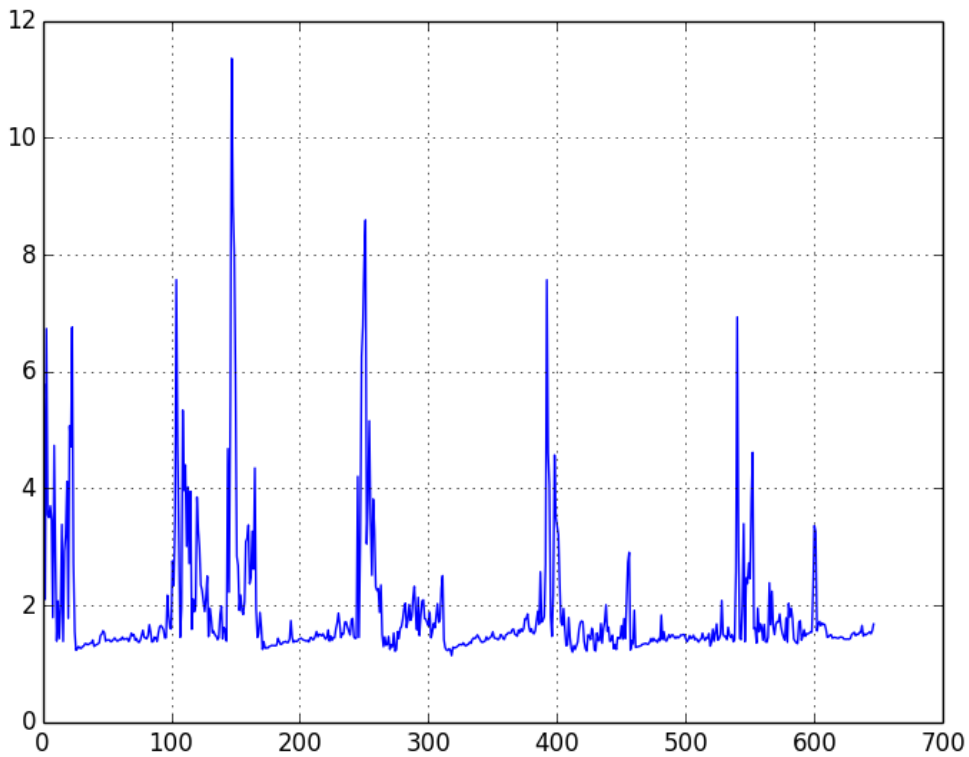
where α is the coefficient of smoothing.
Example with $\alpha = 0.05$:



Smoothing effect of aggregation 1 min

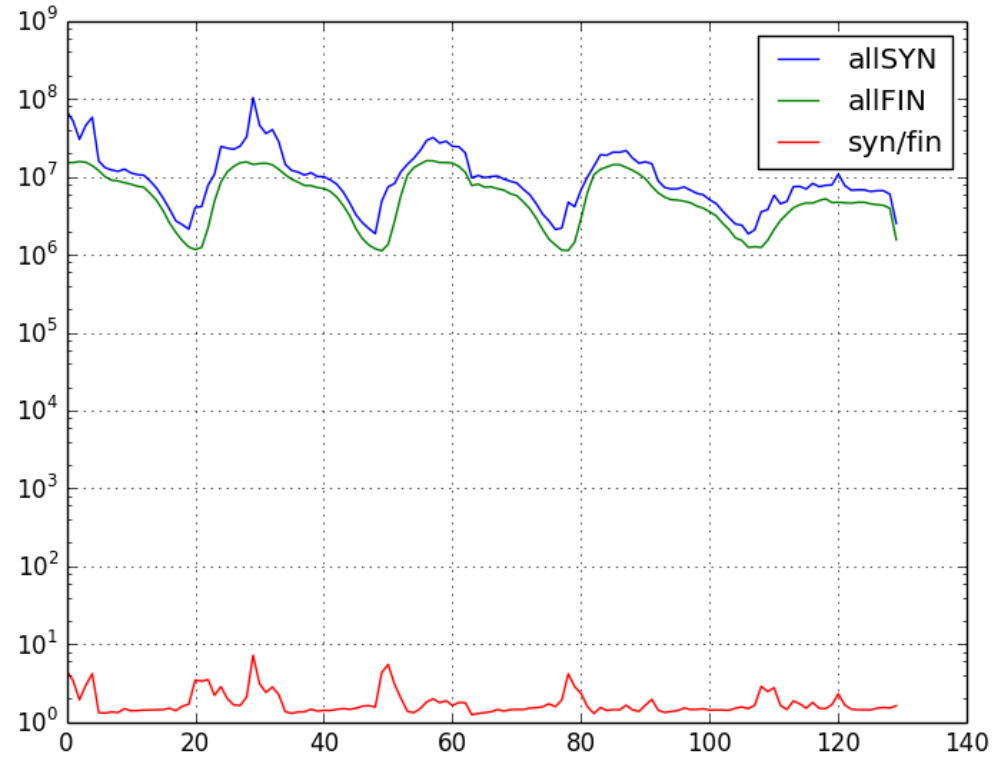
Observed number and SYN/FIN ratio

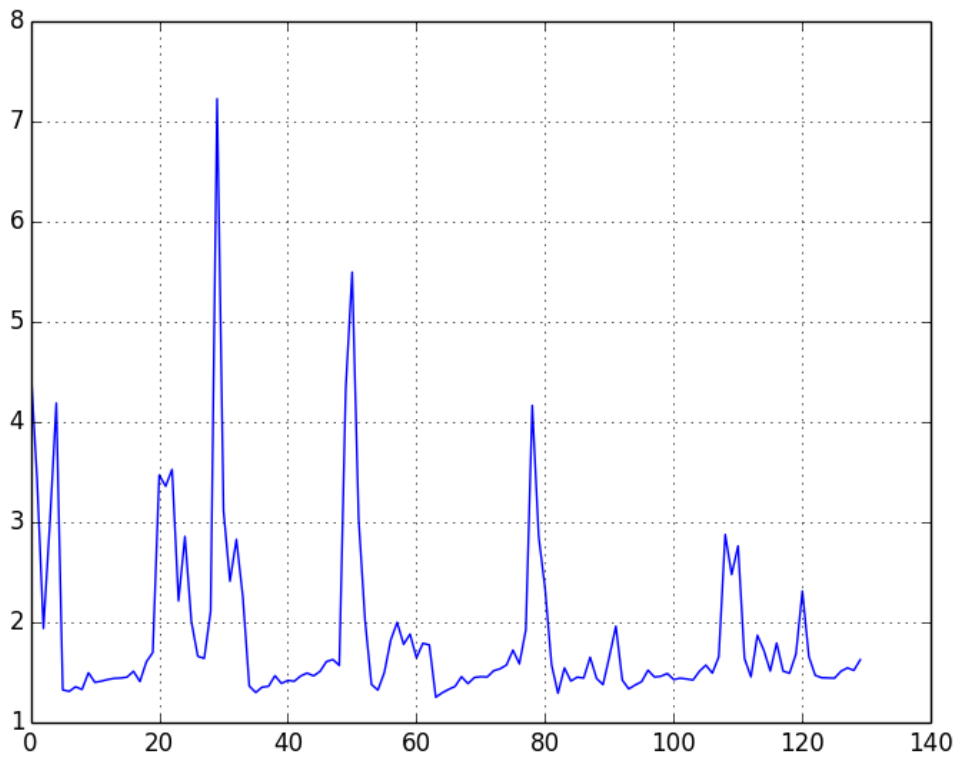




Smoothing effect of aggregation 5 min

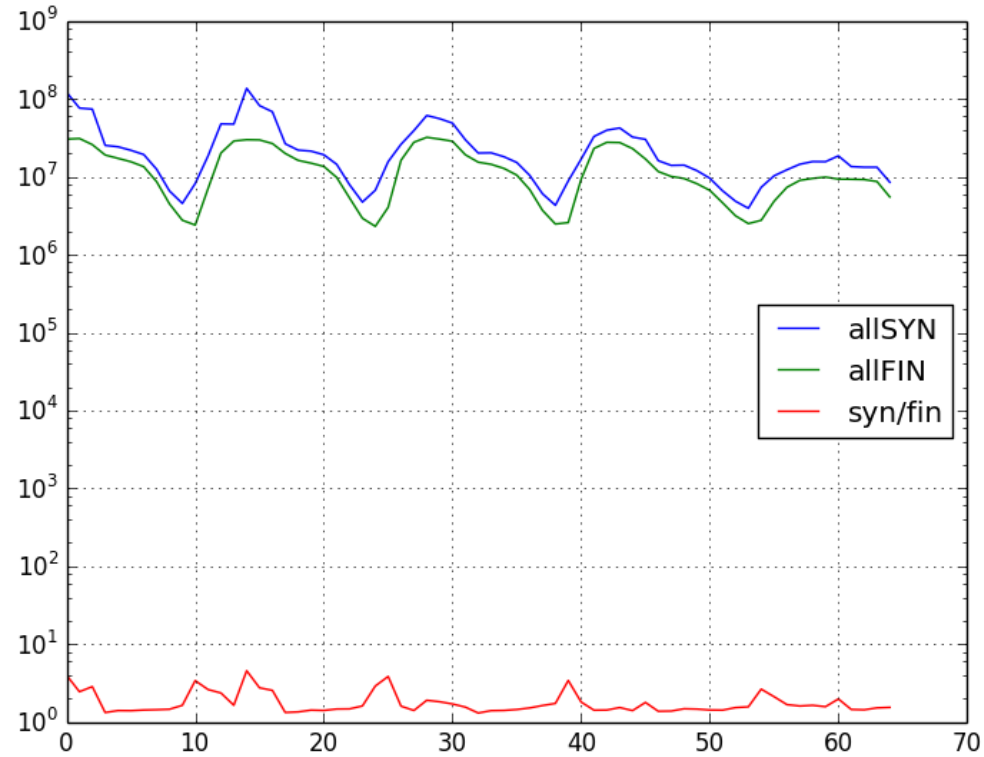
Observed number and SYN/FIN ratio

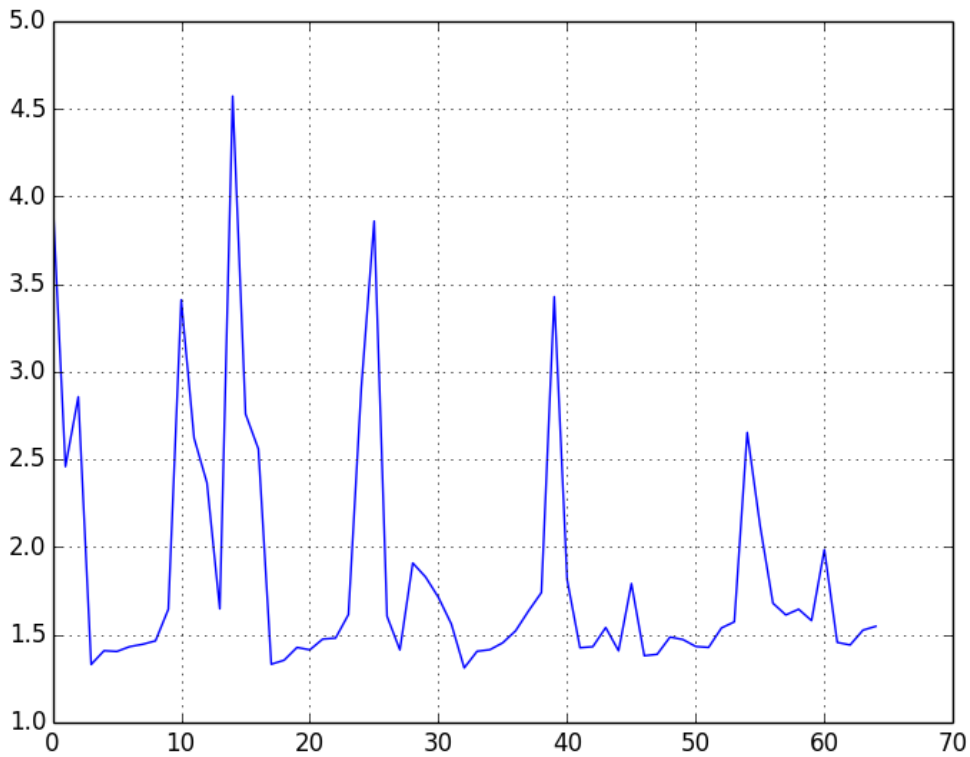




Smoothing effect of aggregation 10 min

Observed number and SYN/FIN ratio





Holt-Winters — Prediction and Seasonality

$$\hat{y}_{i+1} = a_i + b_i + c_{i+1-m} \tag{1}$$

where a, b, c are components defined as:

$$\begin{aligned} a_i &= \alpha(y_i - c_{i-m}) &&+(1-\alpha)(a_{i-1} + b_{i-1}) \\ b_i &= \beta(a_i - a_{i-1}) &&+(1-\beta)b_{i-1} \\ c_i &= \gamma(y_i - a_i) &&+(1-\gamma)c_{i-m} \end{aligned} \tag{2}$$

a_i "baseline" or "intercept",

b_i "linear trend" or "slope",

c_i "seasonal trend",

α, β, γ adaptation parameters, $0 < \alpha, \beta, \gamma < 1$.

The detection is based on a confidence band $(\hat{y}_i - \delta_- d_{i-m},\ \hat{y}_i + \delta_+ d_{i-m})$ where $\hat{y}_i$ is a predicted value and δ_-, δ_+ are scaling factors for the width of the confidence band.

Future Prediction

- Currently, scope of Machine Learning
- Having historical data, a predictor can be trained
- Random forests, XGBoost (Extreme Gradient Boosting), Neural Networks, ...

Section 3

Common Performance Metrics

Test Power and Probability of False Alert

The performance of the sequential detection procedures can be measured using various of criteria: Test power (PWR) and Probability of False Alert (PFA):

- We desire high power and low probability of false alerts.
- For a fixed decision time k we can define:

$$\text{PWR}^k = P(\tau \leq k | \lambda \leq k)$$

$$\text{PFA}^k = P(\tau \leq k | \lambda > k) = P_0(\tau \leq k)$$

- In long-term network monitoring the change-point λ (i.e., an intrusion) may occur very late.
- For large k, any detection procedure will have PFA^k nearly 1 or at least relatively high.

PWR and PFA in practice

- It is more practical to consider conditional PFA for a sliding time interval of length T

$$\text{PFA}_T^k = P(\tau < k + T | \tau \geq k, \lambda \geq k + T) = P_0(\tau < k + T | \tau \geq k)$$

$$\text{PFA}_T = \sup_{1 \leq k \leq \infty} \{P_0(\tau < k + T | \tau \geq k)\}$$

- The condition $\tau \geq k$ corresponds to false alarms shortly after (within) period T
  - The IDS was inspected and found OK at time k
  - The IDS was started or restarted after an alert ($k = 0$)
  - An upper bound on PFA_T works for all of these situations

Run Length and False Alert Rate

- The average run length before the change

$$\mathrm{ARL}^0 = \boldsymbol{E}_0 \tau$$

- The average run length after the change

$$\mathrm{ARL}^1 = \boldsymbol{E}_1 \tau$$

- We desire quick detection (low ARL^1) and infrequent false alerts (high ARL_0)
- The average False Alert Rate

$$\mathrm{FAR}(\tau) = \frac{1}{\boldsymbol{E}_0 \tau}$$

  is often used instead of ARL_0
- Low FAR is a very important and practical requirement!

Detection Delay

- Average detection delay, assuming a change at a fixed $\lambda = k$

$$\boldsymbol{E}_k(\tau - k)^+$$

  - As k increases, the delay often approaches 0.
  - For a random λ we can summarize: $\boldsymbol{E}(\boldsymbol{E}_\lambda(\tau - k)^+))$
- Conditional average detection delay

$$\mathrm{ADD}_\lambda(\tau) = \boldsymbol{E}_\lambda(\tau - \lambda | \tau \geq \lambda)$$

  - Very important in continual surveillance.
  - Often approaches a constant for increasing $\lambda = k$ (stable detection system)

Utility Function

- There is often cost associated with
  - False alerts and detection delays
  - Communication overhead in network security
- If the cost (utility function) is a linear function of FAR and delay, an optimal procedure can in some cases be obtained by fixing the FAR and minimizing the ADD

Evaluation Criteria

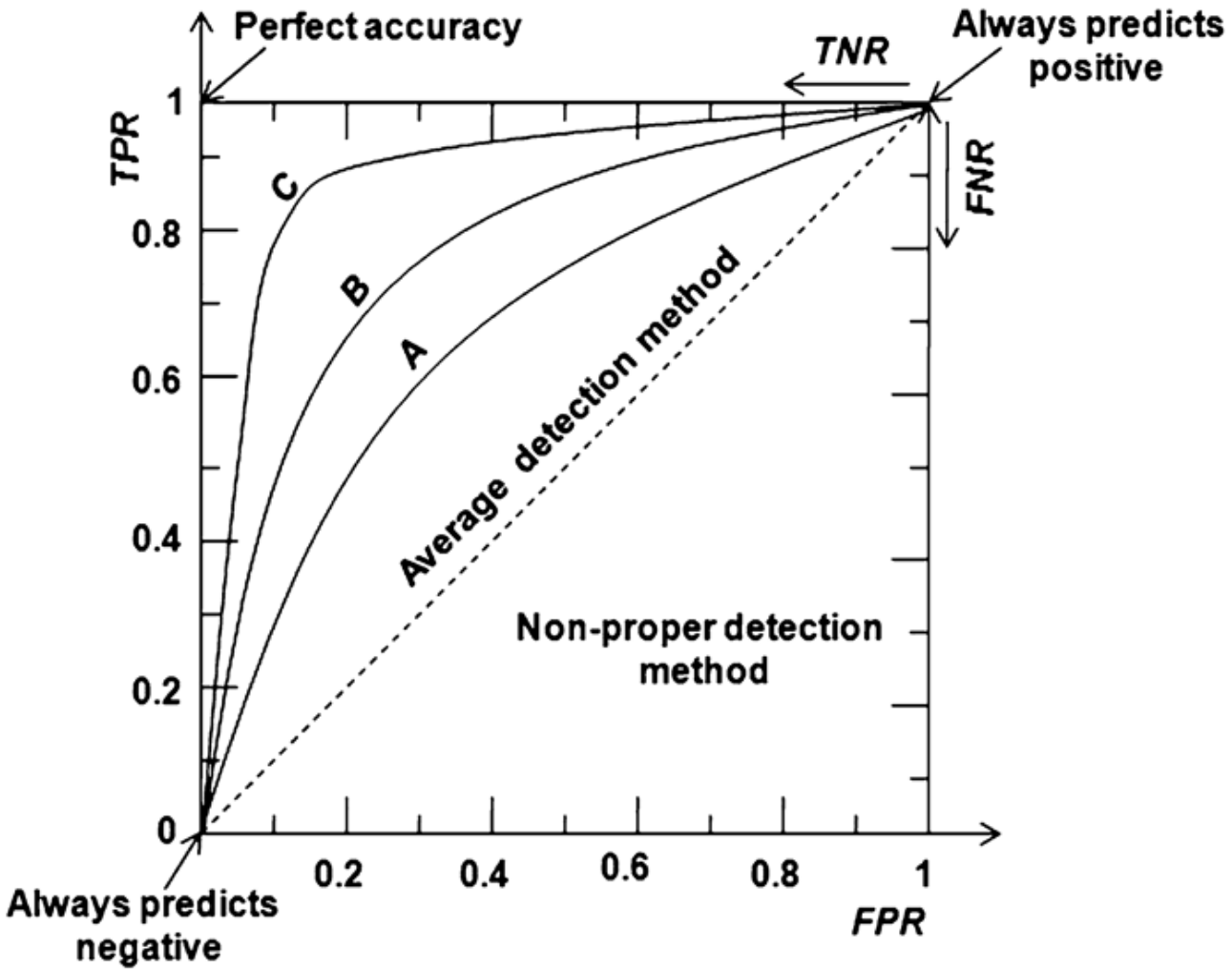
- Accuracy: measures how correctly an IDS works, percentage of detection and failure, false alarms
  (90 % accuracy means correct classification of 90 instance of 100 as belonging to their actual classes)
- Taxonomy of evaluation measures:
  - Data Quality (Quality, Validity, Completeness, Reliability)
  - Correctness (P-R and F-measures, ROC Curves, Misclassification Rate, Confusion Matrix, AUC Area, Sensitivity and specificity)
  - Efficiency (Stability, Timeliness, Unknown Attack, Update Profile, Interoperability, Performance, Generate Alert)

Data Quality

Quality	reliability/legitimacy of source, good selection of samples (unbiased), good sample size (neither over nor under-sampling), time of data, complexity of data
Reliability	accuracy, consistency, expected purpose
Validity	valid data, e.g., good values in expected ranges
Completeness	represent the space of the vulnerabilities and attacks that can be covered by an IDS

Correctness

ROC Curve Receiver Operating Characteristics, originates from signal processing



See ROC Example at https://medium.com/wwblog/

AUC Area Under Curve, computed from ROC, result is withing the range 0.5–1.0

Precision, Recall, and F-Measure defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{Pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F-measure} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

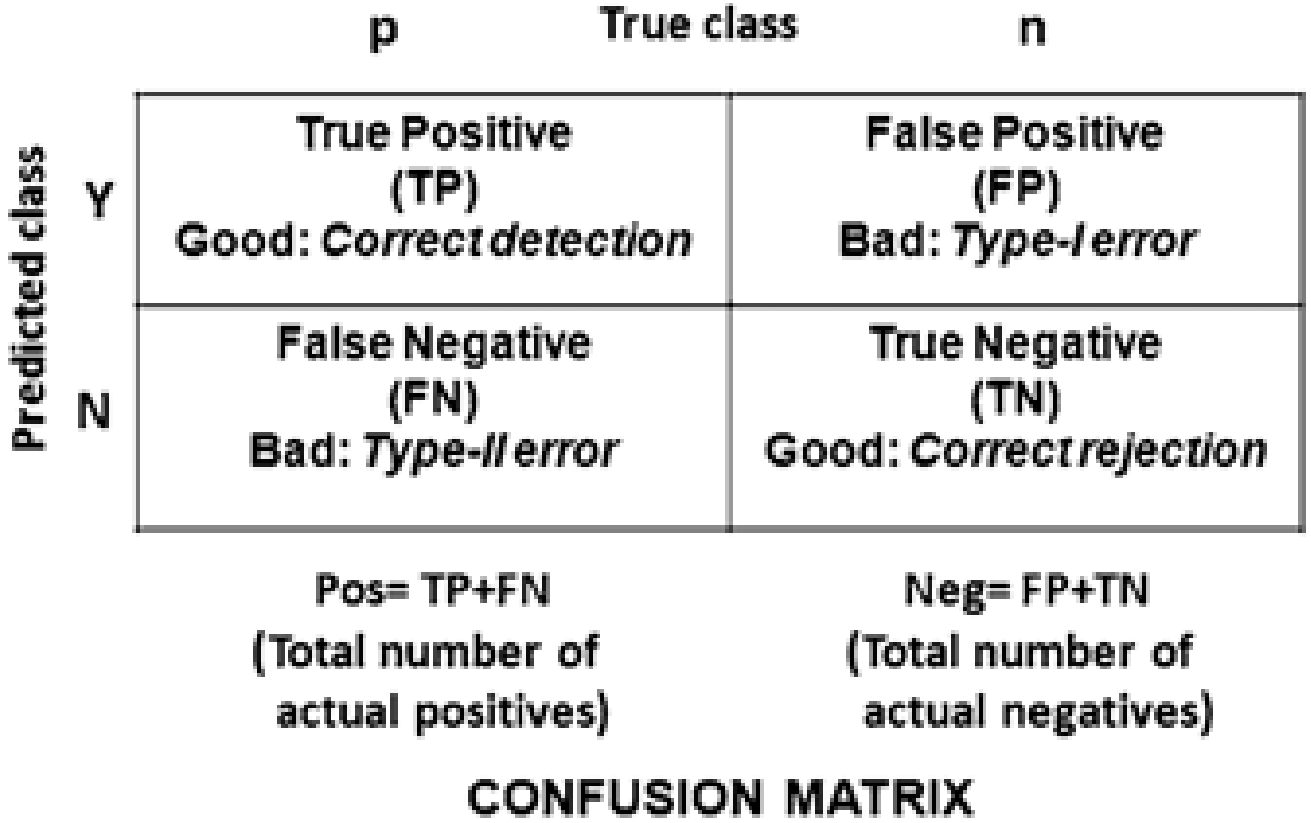
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Pos} + \text{Neg}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{Neg}} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{FPR}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{Pos}} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{TPR}$$

Confusion Matrix is composed of TP, FP, FN, TN:

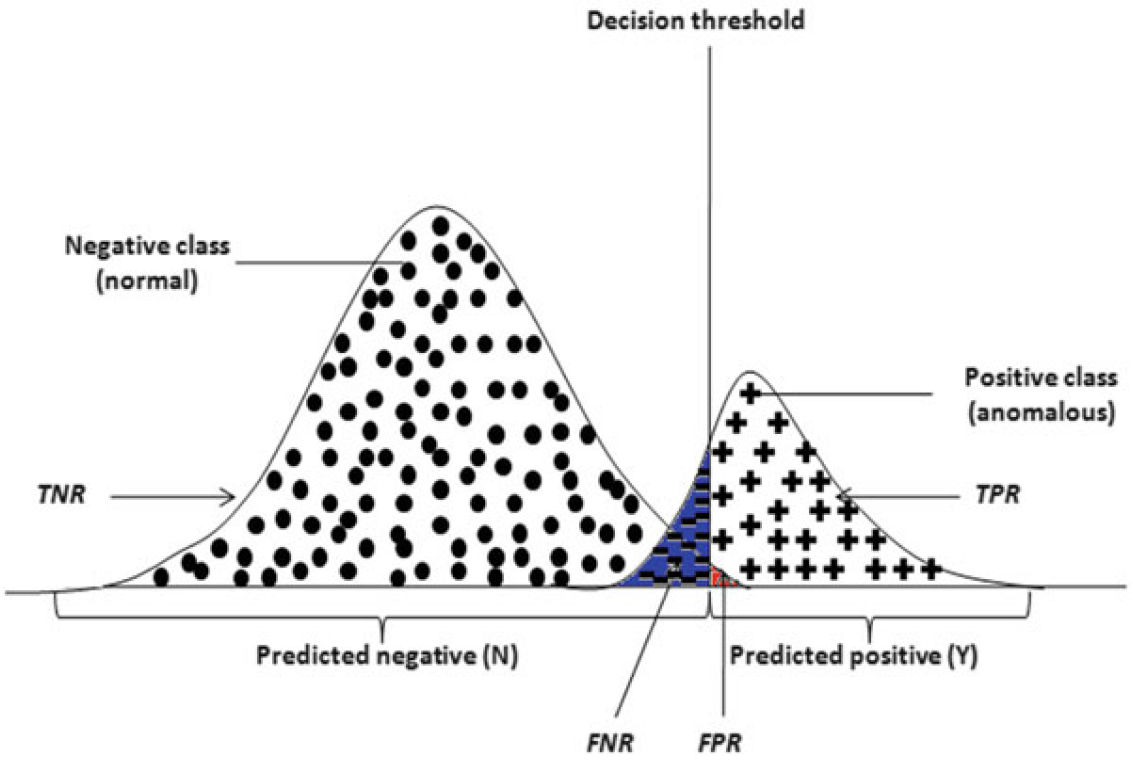


Misclassification Rate

$$\frac{FN + FP}{TP + FP + FN + TN}$$

Sensitivity and Specificity

- TPR is also known as *sensitivity*
- TNR is also known as *specificity*
- TPR, FPR, TNR, and FNR can be defined for the normal class



Tradeoffs in continual surveillance:

Tuning Adjustment	**Effects on FAR and ADD**	**PFA and PWR in a time interval of given size**	**Effects on overall PFA**
Higher Threshold	Smaller FAR Shorter ADD	Smaller PFA Lower Power	PFA approaches 1
Lower Threshold	Higher FAR Shorter ADD	Higher PFA Higer Power	PFA approaches 1
Optimization Strategy A	Limit FAR $\leq \gamma$ Minimize ADD	Limit PFA $\leq \alpha$ Maximize Power	Not applicable
Optimization Strategy B	Limit ADD $\leq K$ Minimize FAR	Guarantee Power $\geq \beta$ Minimize PDA	Not applicable

Efficiency (related generally to detection systems)

Stability	detection performance is consistent in different network scenarios
Timeliness	total delay between t_{attack} and t_{response}
Performance	throughput of the detection method (e.g., how many packets/second without loss), CPU and memory usage $\rightarrow$ computational and memory complexity
Update Profile	possibility to add new or modified profiles or signatures accurately
Interoperability	capability to correlate information from multiple sources
Unknown Attack	ability to detect unknown or modified intrusion patterns

Detection Performance Metrics Tradeoffs

- Algorithms cannot maintain all metrics at prescribed levels
- Optimization of detection procedures balances the tradeoffs

A standard optimization strategy:

- Prescribe the bounds for some metrics
- Use some other metrics as optimization criteria

The selection of these metrics has practical considerations

- Classical approach: maximize the test power among all tests with a fixed prescribed low level of PFA
- Continual surveillance: minimizing the detection delay for a prescribed low FAR

Section 4

Closing Words

References (1/2)

1. Bhuyan, Monowar H., Dhruba K. Bhattacharyya, and Jugal K. Kalita. *Network traffic anomaly detection and prevention: concepts, techniques, and tools.* Springer, 2017.
2. R. Blažek, H. Kim, B. Rozovskii, and A. Tartakovsky, *A novel approach to detection of 'denial-of-service' attacks via adaptive sequential and batch- sequential change-point detection methods,* Proc. 2nd IEEE Workshop on Systems, Man, and Cybernetics, West Point, NY, 2001.
3. A. Tartakovsky, B. Rozovskii, R. B. Blažek, and H. Kim, *Response of authors to discussions on detection of intrusions in information systems by sequential change-point methods,* Statistical Methodology, vol. 3, pp. 329–340, July 2006.
4. A. Tartakovsky, B. Rozovskii, R. B. Blažek, and H. Kim, *Detection of intrusions in information systems by sequential change-point methods,* Statistical Methodology, vol. 3, pp. 252–293, July 2006.

References (2/2)

5. B. Rozovskii, A. Tartakovsky, R. B. Blazek, and H. Kim, *A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods,* IEEE Transactions on Signal Processing, vol. 54, pp. 3372–3382, September 2006.
6. `https://www.symantec.com/connect/articles/statistical-based-intrusion-detection`
7. `https://medium.com/wwblog/evaluating-anomaly-detection-algorithms-with-receiver-ope`
8. `https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/`

Questions?