Using machine learning for anomaly detection in streaming data and its economic implications

Stefani Majić, Marijana Zekić Sušac, Adela Has
Faculty of Economics in Osijek, University of Osijek

Content

INTRODUCTION

Streaming data

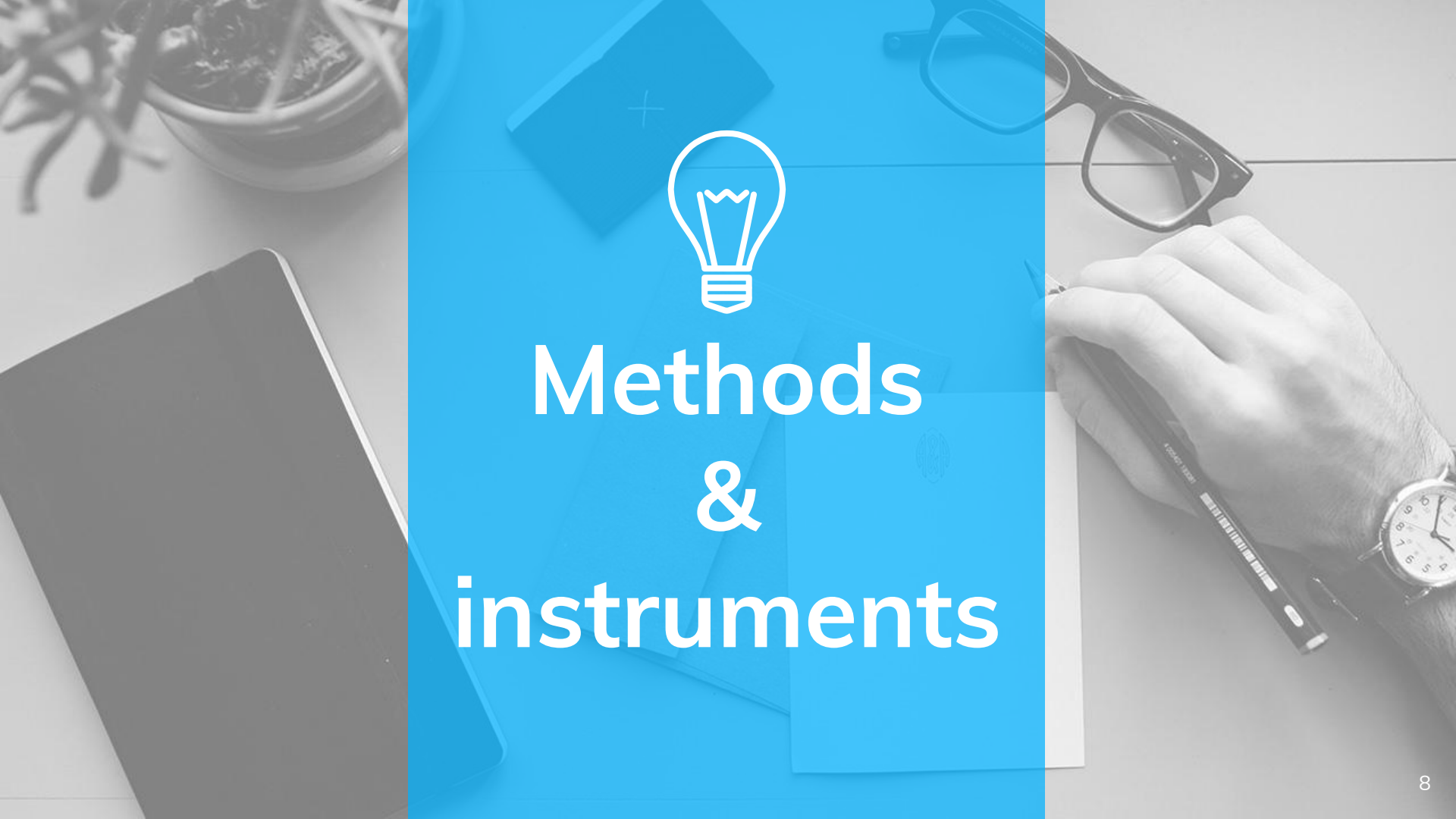
- → sequence of real-time, continuous-stream items ordered either implicitly by arrival time or explicitly by timestamp
- → marked by the inability to control the arrival of items in the sequence and the inability to fully save the whole stream locally
- → sensor data, internet traffic and transaction logs

Challenges

- ➜ Concept drift
- ➜ Storing a relative small subset instead of the whole stream
- ➜ Long-running queries may encounter changes in system conditions throughout their execution
- ➜ Applications that monitor streams in real-time must react quickly to unusual data values

PREVIOUS RESEARCH

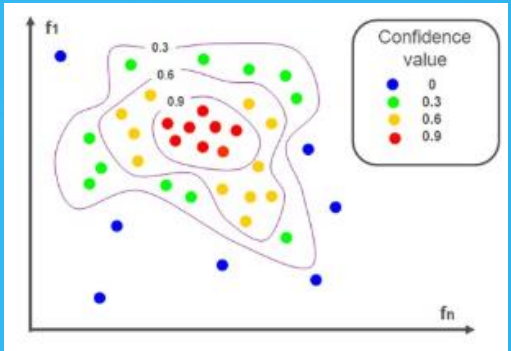
The biggest usage of machine learning is anomaly detection, as evidenced by the interest of many researchers and theoreticians in that area such as Hill (Hill, 2010), Burbeck (Burbeck, 2007) and Davy (Davy, 2006) who explored the possibilities of deploying machine learning algorithms to detect anomalies in data which constantly transmits through the network.



Methods & instruments

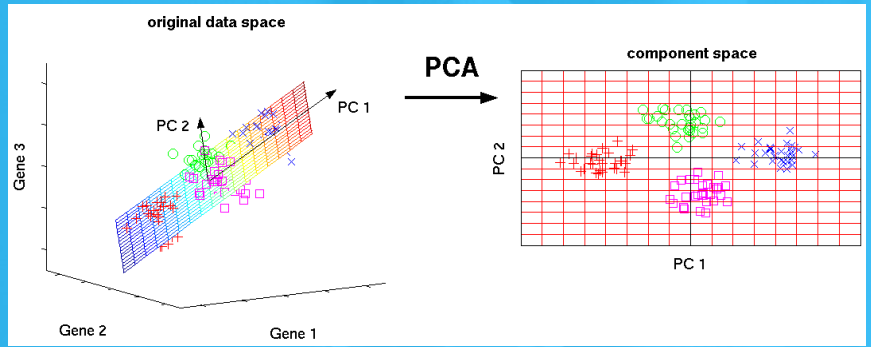
ONE-CLASS SVM

- one-class SVM can be used to define the normal class and points falling outside the given boundaries are defined as anomalies



Principal component analysis

- a technique used to reduce dimensions of datasets by projecting data points into the directions of maximal variance within data space (Microsoft, 2017)



Instruments

Azure ML Studio

a collaborative drag-and-drop tool used to build, test and deploy predictive analytics solution on data

Power BI

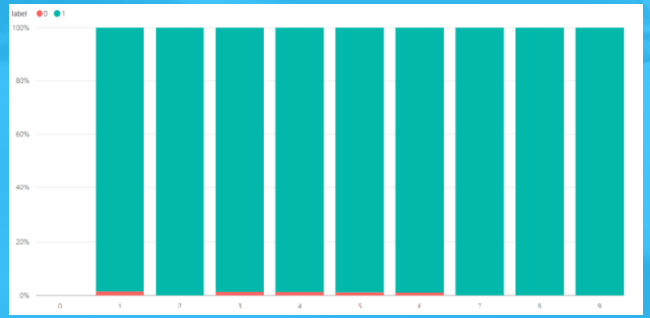
is an evolution of add-ins like Power Pivot, Power Query and Power view with three main points - dashboards, reports and visualizations

SUBJECT OF RESEARCH

Subject of research

Computer Network Traffic dataset

- Dataset from Stanford University and Kaggle.com
- 20803 records and 4 variables
- Info about abnormal activities:
  - 24.08. (IP 1), 04.09. (IP 5), 18.09. (IP 4), 26.09. (IP 3), and 26.09. (IP 6)
  - New categorical variable added – *label (0 – anomaly, 1-normal)*



Computer Network Traffic dataset

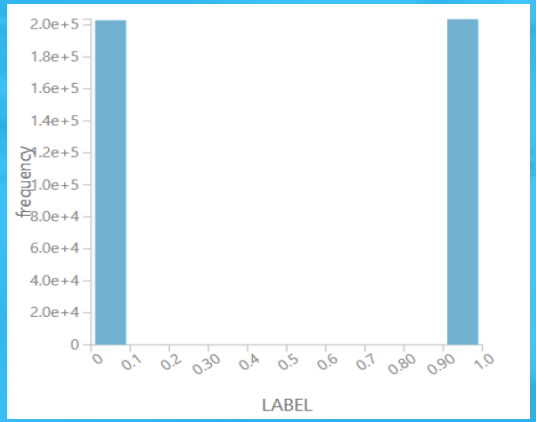
Variables selected for creating model:

- date – record date (01.07.2006 - 30.09.2006, gggg-mm-dd)
- l_ipn – local IP address, integer categorical variable (0 – 9)
- r_asn – remote ASN, integer variable for identification of remote ISP
- f – number of connections per day
- Label – categorical variable (0 or 1)

Subject of research

IoT sensor dataset

- Dataset from Microsoft Cortana Intelligence Gallery
- 406516 records and 9 variables (5 selected)
- Number of normal class records is 203569 and anomalies 202947

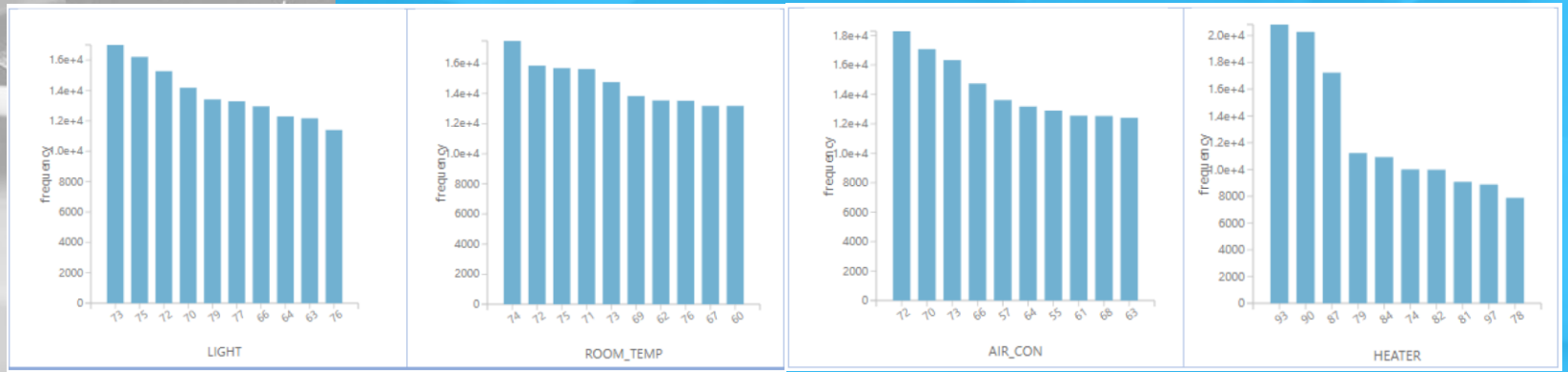


Subject of research

IoT sensor dataset

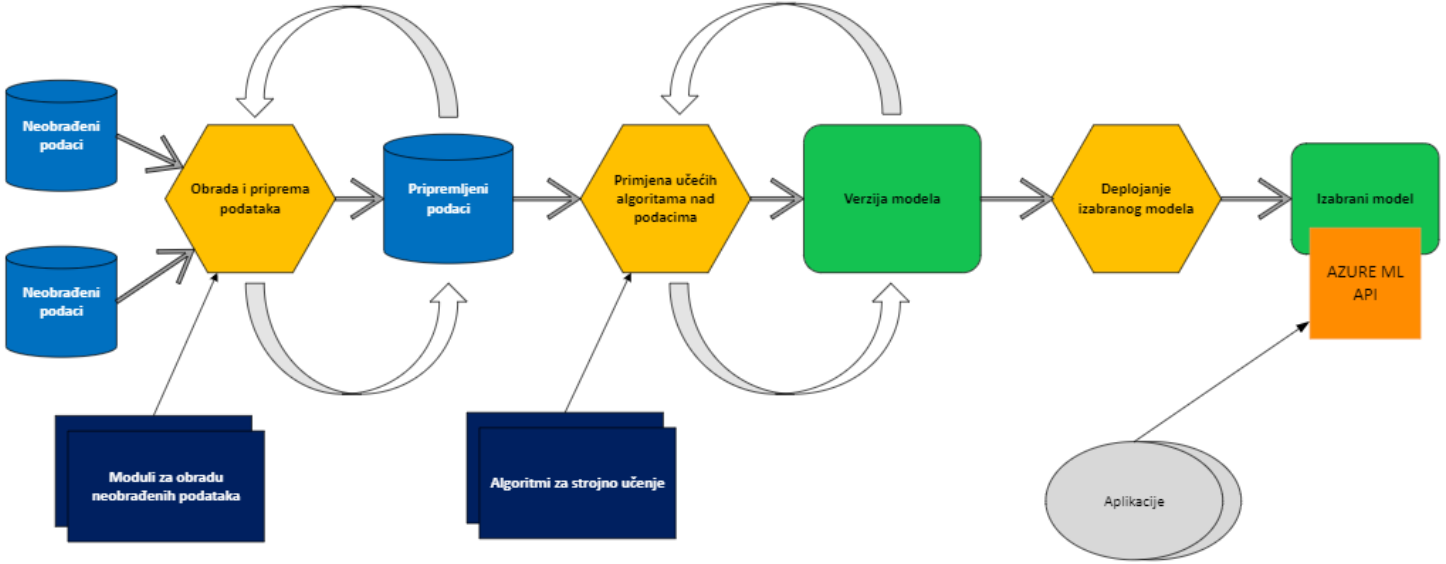
Variables selected for creating model:

- ➜ light – light sensor readings
- ➜ room_temp – temperature sensor readings
- ➜ air_con – air conditioner sensor readings
- ➜ heater – heat sensor readings
- ➜ label – categorical variable (0 or 1)



Machine Learning proces in ML studio

F-score is used for measuring classification performance and RMSE for regression performance



SUBSAMPLING PROCEDURE

Subsample	Computer Network Traffic dataset	IoT sensor dataset
Train subsample	**70% (14552 cases)**	**70% (284491 cases)**
Test subsample	**30% (6236 cases)**	**30% (121925 cases)**
Total sample	**100% (20788 cases)**	**100% (406416 cases)**

FINDINGS

FINDINGS

One-class SVM

Higher accuracy

disadvantage:

slow execution speed

PCA

lower accuracy

Limitations:

linearity, orthogonality

Model	Computer Network Traffic dataset accuracy	IoT sensor dataset accuracy
SVM	**96.4%**	**83.6%**
PCA	**1.1%**	**49.9%**

Findings #2

- ➜ SVM deployed as a web service
- ➜ From initial datasets records were taken out as follows:
  - 15 records – computer network traffic
  - 100 records – IoT sensors

Please note: These records were not introduced to the model in it's lifecycle of learning.

Findings #2

For computer traffic network dataset:

→ in 67% cases – anomaly was correctly identified.
→ in 87% all cases the record's class was correctly determined.
→ in 2 out of 15 cases, the model was incorrect.

For IoT sensors dataset:

→ the model identified the class correctly in 89% cases – observing only anomalies in 82.5% cases and for normal class in 93.3% cases.

ECONOMIC IMPLICATIONS & DISCUSSION

Economic implications

- ➜ Possible applications:
  - Smart house, gaming industry, logistic, marketing, DevOps, Identity managment
- ➜ Economic implications:
  - reducing the cost of data processing, since data with anomalies have a lower rate of usefulness.
  - Avoiding some costly activities by detecting and removing the anomaly on on time
  - Significant economic implications in the area of data processing and decision making.

DISCUSSION & CONCLUSSION

Both methods show great potential in this area.

spotted deficiency of SVM – speed of execution

Further research: to explore how one-class SVM performs in area of computer vision and robotics

If computer network dataset had more anomalies, presumably the SVM model accuracy could be higher

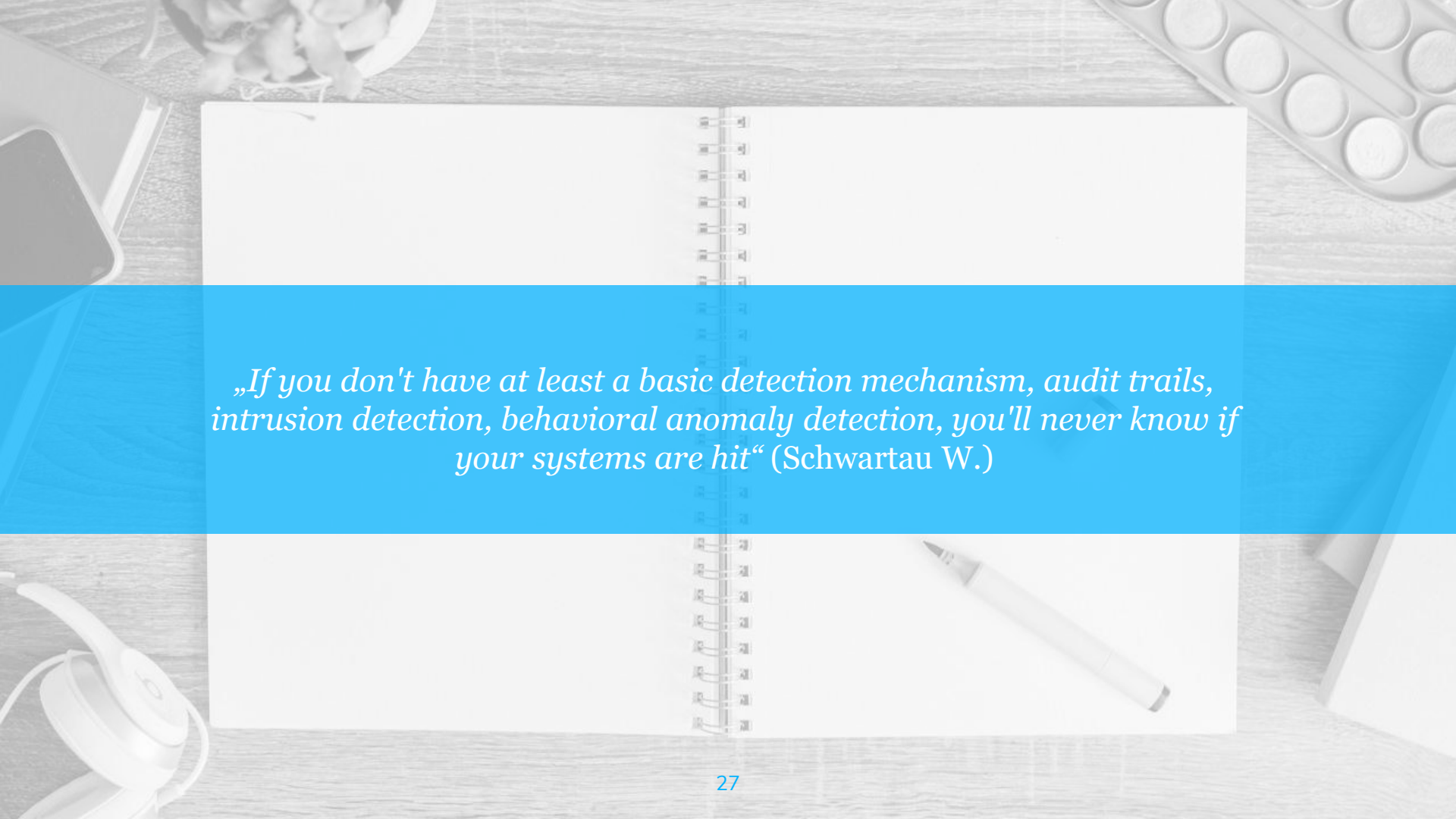
Remaining questions: the possibility of model optimization and how would the created models work on new data in real-life scenarios

References

- Barga R., Fontama V., Tok W. H. (2014). *Predictive analytics with Microsoft Azure Machine Learning: Build and deploy actionable solutions in minutes*. Apress.
- Burbeck K, Nadjm-Tehrani S. (2007). Adaptive real-time anomaly detection with incremental clustering. *Information Technology Report, 12.* Retrieved from: https://www.ida.liu.se/labs/rtslab/publications/2007/BurbeckNadjm07.pdf
- Davy, M., Desobry, F., Gretton, A., and Doncarli, C. (2006). An online support vector machine for abnormal events detection. *Signal processing, 86(8)*, 2009-2025. Retrieved from: https://dl.acm.org/citation.cfm?id=1159461
- Golab L., Tamer Özsu M. (2003). *Issues in Data Stream Management*. Retrieved from: http://www.mathcs.emory.edu/~cheung/papers/StreamDB/SlidingWindows/2003-Issues-in-data-stream-management.pdf
- Hill, D. J., and Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software, 25(9)*, 1014–1022. Retrieved from: https://dl.acm.org/citation.cfm?id=1808816
- Kaggle, Computer Network Traffic Dataset, https://www.kaggle.com/crawford/computer-network-traffic
- Lee H., Roberts S.J. (2007). *On-line Novelty Detection: Using the Kalman Filter and Extreme Value Theory*. Retrieved from: http://www.robots.ox.ac.uk/~sjrob/Pubs/LeeRoberts_EVT.pdf

References

- Mayer D. (2017). *Support Vector Machines,* FH Technikum Wien, Austria. Retrieved from: https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf
- Microsoft, Azure Machine Learning Studio, Retrieved from: https://studio.azureml.net/
- Power BI microsoft.com. Retrieved from: https://powerbi.microsoft.com/en-us/documentation/powerbi-service-dashboards/
- PCA-based Anomaly Detection, MSDN microsoft.com. Retrieved from: https://msdn.microsoft.com/library/azure/c3822fa5-1095-4c72-bd1e-fd43d285153a
- Schölkopf B., Platt J. C., Shawe-Taylor J, Smola A. J. (2000). *Estimating the Support of a High-Dimensional Distribution.* Retrieved from: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-99-87.pdf
- Stanford, Computer Network Traffic Dataset, http://statweb.stanford.edu/~sabatti/data.html
- Tsymbal A. (2004). *The problem of concept drift: definitions and related work.* Retrieved from: https://www.scss.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf
- Tan, S. C., Ting, K. M., and Liu, T. F. (2011). Fast anomaly detection for streaming data. *In IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 22*, pp. 1511.
- Zekić-Sušac, M. (2017). Machine learning in energy consumption management. Zadnik-Stirn, L., Drobne, S. (Eds.), *Proceedings of the 14th International Symposium on Operations Research in Slovenia* (pp. 7-17).
- Zheng A. (2015). *Evaluating Machine Learning Models*, O'Reilly Media



„If you don't have at least a basic detection mechanism, audit trails, intrusion detection, behavioral anomaly detection, you'll never know if your systems are hit“ (Schwartau W.)


Thank you