UNIVERSITY OF WARSAW

QED SOFTWARE

Introducing AI/ML Components into Software Systems

Dominik Ślęzak
https://qed.pl/

SOD SECURITY ON-DEMAND

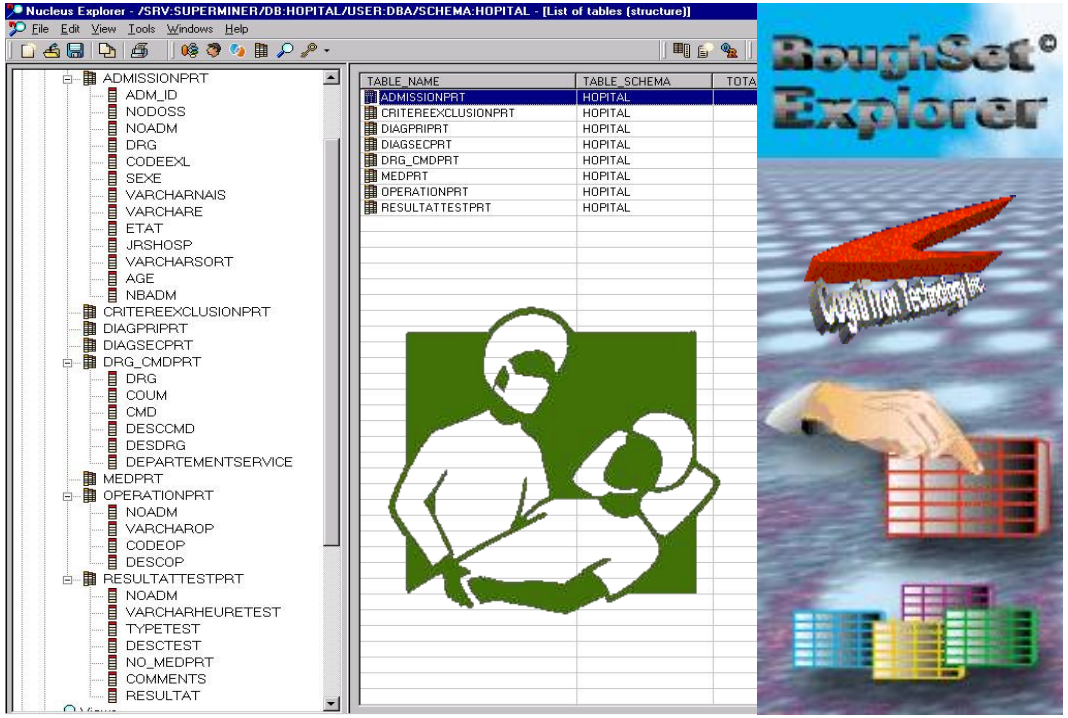
knowledgePit
A Data Challenge Platform

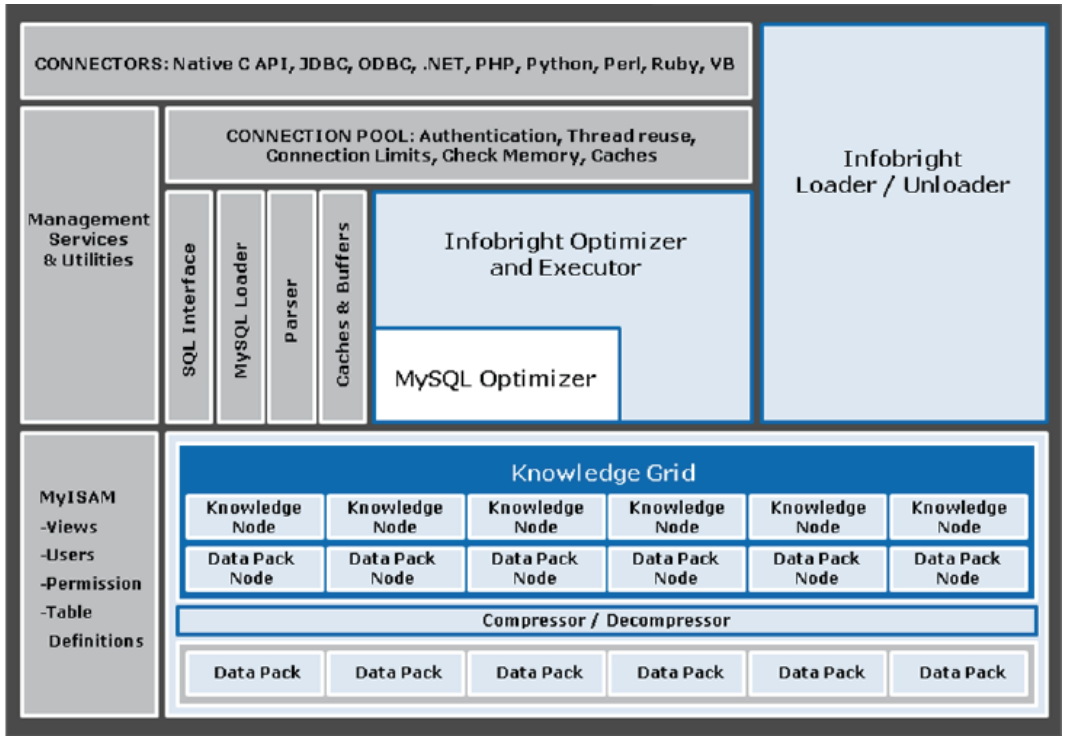
Outline

- Introduction (my personal background)
- Introduction (what I want to say actually)
- ML competitions - what is their purpose?
- Label in the Loop and other [ML] projects
- Concluding remarks (many of them!!!)

Introduction
(my personal background)

- Data Analytics 1995-2005
- Data Processing 2005-2015





INFOBRIGHT

Customers (examples)

Information Builders

CISCO

SOD SECURITY ON-DEMAND

Label in the Loop

BRIGHTBOX

SMALL BIG DATA

CHALLENGES

QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

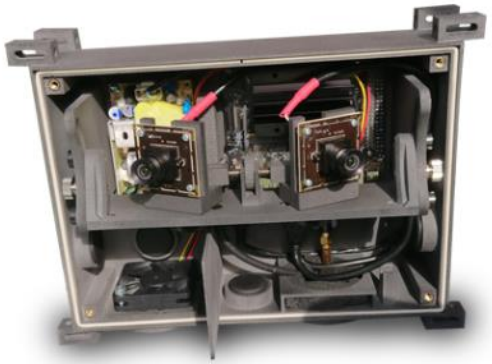
Understand how data quality affects the prediction models

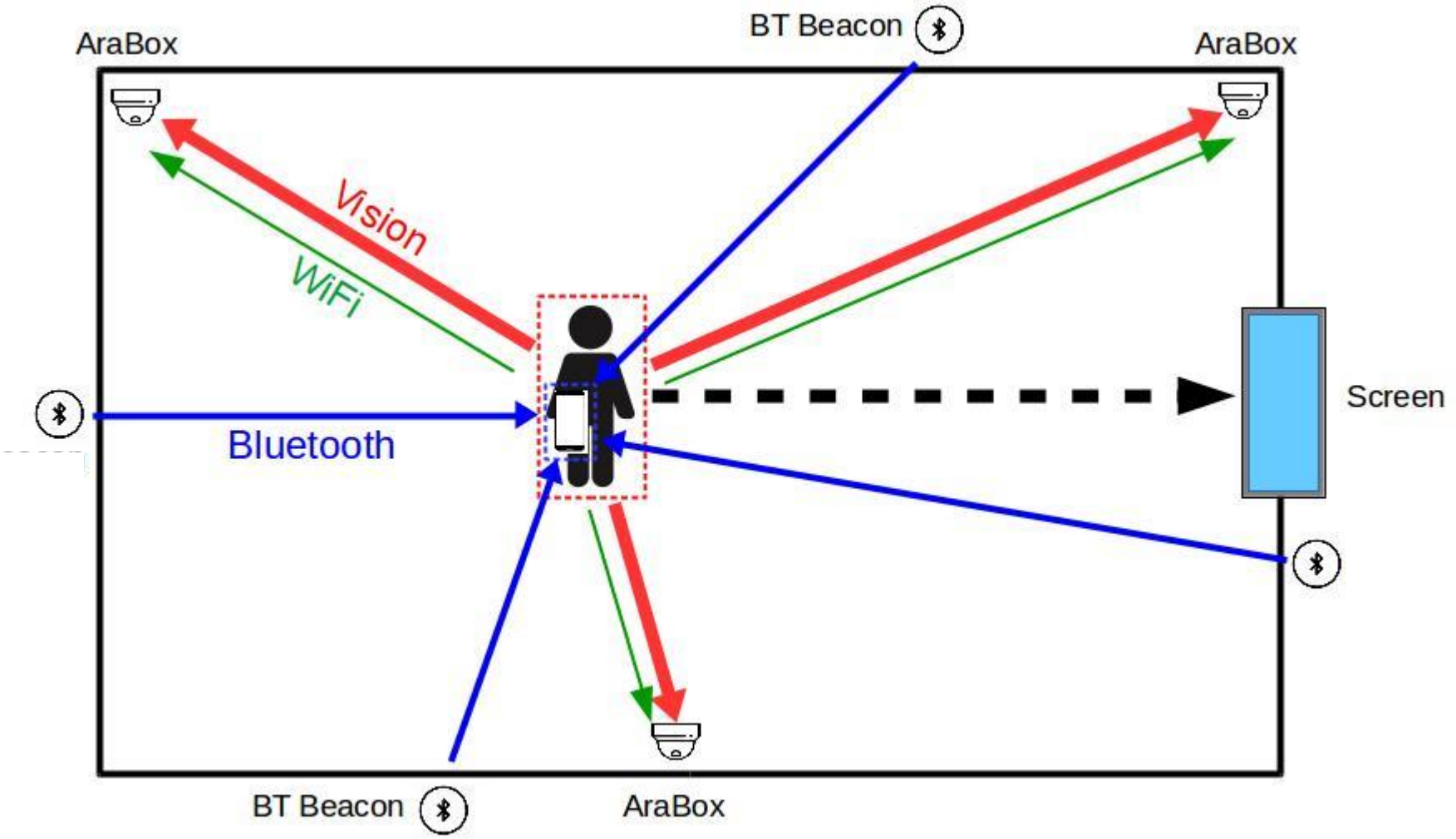
SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows

DED SOFTWARE





jetline

MyLED

SAMSUNG

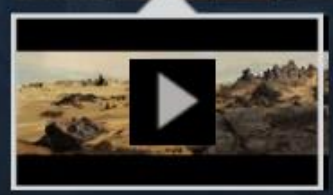
https://arahub.ai/

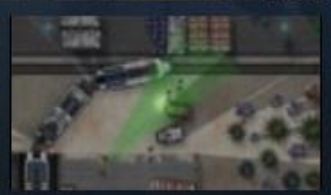
DED SOFTWARE

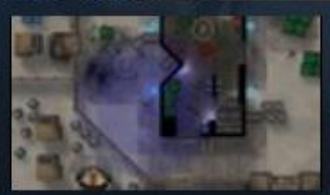
All Games > Strategy Games > Tactical Troops: Anthracite Shift

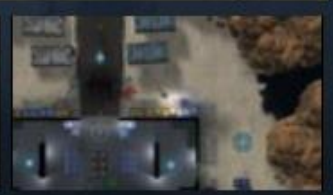
Tactical Troops: Anthracite Shift

Community Hub









Tactical Troops: Anthracite Shift is an indie turn-based tactical science-fiction game set on the beautiful yet dangerous planet of Anthracite. Command a troop of elite soldiers using an advanced teleportation technology and their overwatch technique.

AI ≠ ML

ALL REVIEWS:

RELEASE DATE:

DEVELOPER: QED Games, K-Project

PUBLISHER: Slitherine Ltd.

Popular user-defined tags for this product:

Turn-Based Strategy | Turn-Based Tactics | Sci-fi | +

Introduction
(what I want to say actually)

- Adoption of AI and ML solutions is on the bucket list of many companies which want to harness the power of the data, hence increasing their competitive advantage
- When it comes to actual implementation of these solutions in the company's day to day operation, the major revelation usually occurs:
  - the realization that there is a sea of difference between launching a lab-based data science project aimed at building AI/ML models over historical data and making those models „alive"
    - i.e. working in a production environment, adapting to changes and truly supporting the users

ML competitions - what is their purpose?

- Home
- Compete
- Data
- Notebooks
- Discuss
- Courses
- More

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the Documentation or learn about InClass competitions.

👋 New to Kaggle? Start here!

Our Titanic Competition is a great first challenge to get started.

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 22801 Teams

Knowledge

All Competitions

Active | Completed | InClass

All Categories ▾ Default Sort ▾

OSIC Pulmonary Fibrosis Progression
Predict lung function decline
Featured • 3 months to go • Code Competition • 86 Teams

$55,000

SIIM-ISIC Melanoma Classification
Identify melanoma in lesion images
Featured • a month to go • 1886 Teams

$30,000

ALASKA2 Image Steganalysis
Detect secret data hidden within digital images

$25,000

kaggle

IEEE BigData 2019 Cup: Suspicious Network Event Recognition

QED SOFTWARE

8 months, 3 weeks ago

Suspicious Network Event Recognition is a data mining challenge organized in association with IEEE BigData 2019 conference. The task is to decide which alerts should be regarded as suspicious based on information extracted from network traffic logs. The competition is kindly sponsored by Security On-Demand (**https://www.securityondemand.com/**) and QED Software (**http://qed.pl/**).

Overview

Cyber threat detection and analytics play a pivotal role in providing security to organizations that provide web services, and to their users. Importance of this field is continuously growing due to the increasing abundance of Internet services, wireless networks, smart devices, etc. Since the cybersecurity domain is hugely complex, it is also one of the major challenges of the contemporary world.

In this challenge, the task is to detect truly suspicious events and false alarms within the set of so-called network traffic alerts, that the Security Operations Center (SOC) Team members @ SOD have to analyze on an everyday basis. An efficient classification model could help the SOC Team to optimize their operations significantly. It is worth adding that although the competition sponsor is entirely commercial, the knowledge and experience that can be gathered by the competition participants may be highly beneficial to improve the intelligent cybersecurity modules in many organizations.

BREAKING NEWS: Security On-Demand Launches ThreatWatch® Hunt, Advanced Threat Hunting Service

00 : 18 : 25

hours minutes seconds

TODAY IT TAKES AS LITTLE AS 19 MINUTES FOR AN ATTACKER TO "OWN" YOU

Total Logs Analyzed: 4,141,434 Logs

Last 24 Hours

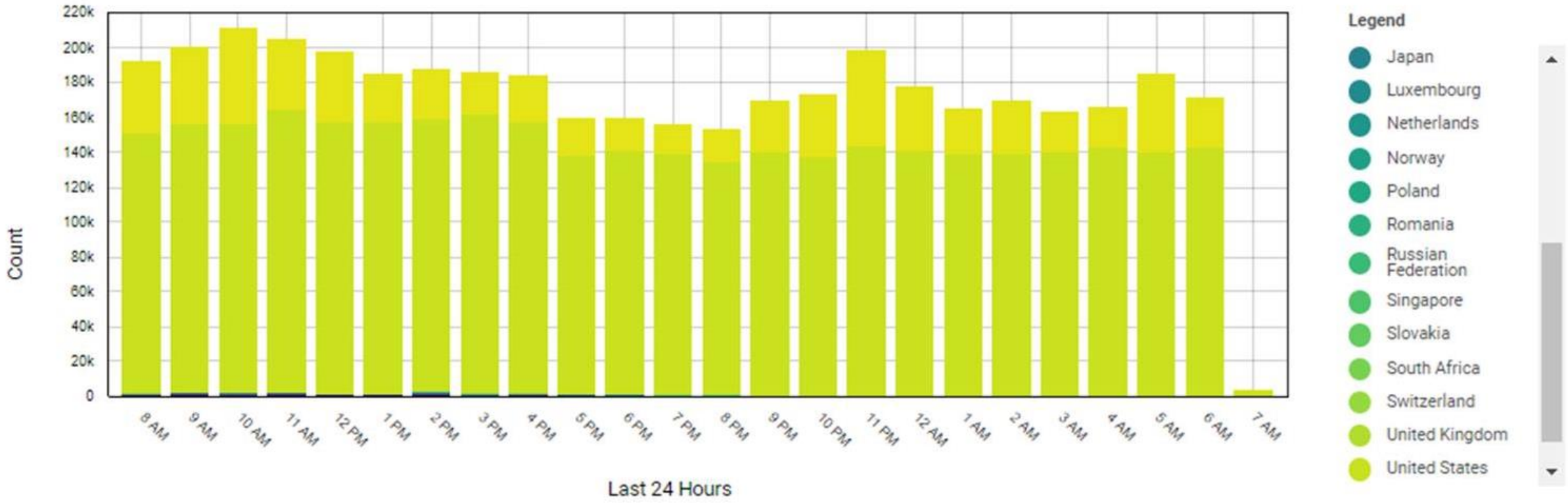
IP Address + | Report Device + | Vendor/Type + | Port + | User + | Country + | Allow/Deny + | Direction +

Filters Allow/Deny = Allow X Destination Port = 443 X

Reset Filters VIEW LOG DETAILS

Log Summary: Last 24 Hours

Destination Country*



Case Study (1)

SOD SECURITY ON-DEMAND

- The **SOD** servers register billions of logs daily
- Millions of atomic alerts are identified
  - Alert identification rules are designed by the **Threat Reconnaissance Unit** and deployed at the ETL level
- Atomic alerts are aggregated into thousands of correlated alerts
- The analysts can investigate only less than 100 of alerts daily
- Customers are informed about roughly 7% of investigated alerts

knowledgePit
A Data Challenge Platform

Case Study (2)

SOD SECURITY ON-DEMAND

- The **SOD** servers register billions of logs daily
- Millions of atomic alerts are identified
  - Alert identification rules are designed by the **Threat Reconnaissance Unit** and deployed at the ETL level
- Atomic alerts are aggregated into thousands of correlated alerts
- The analysts can investigate only less than 100 of alerts daily
- Customers are informed about roughly 7% of investigated alerts

Label in the Loop

What if the data is not good enough?

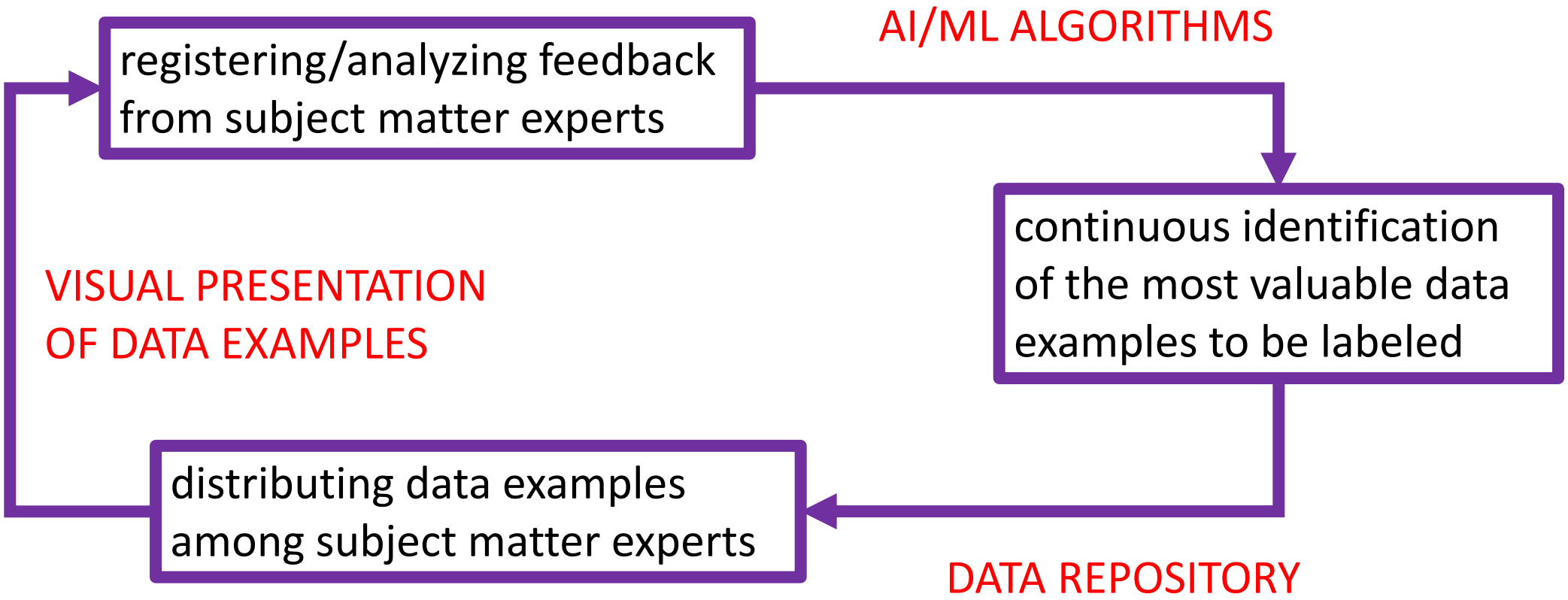
- ML algorithms require the training data
  - What if there are no (sufficient amount of) cases with appropriate labels?

- Could we ask „the crowd” for help?
  - Yes - unless… the labeling process requires highly specialized knowledge…

- **Active Learning**

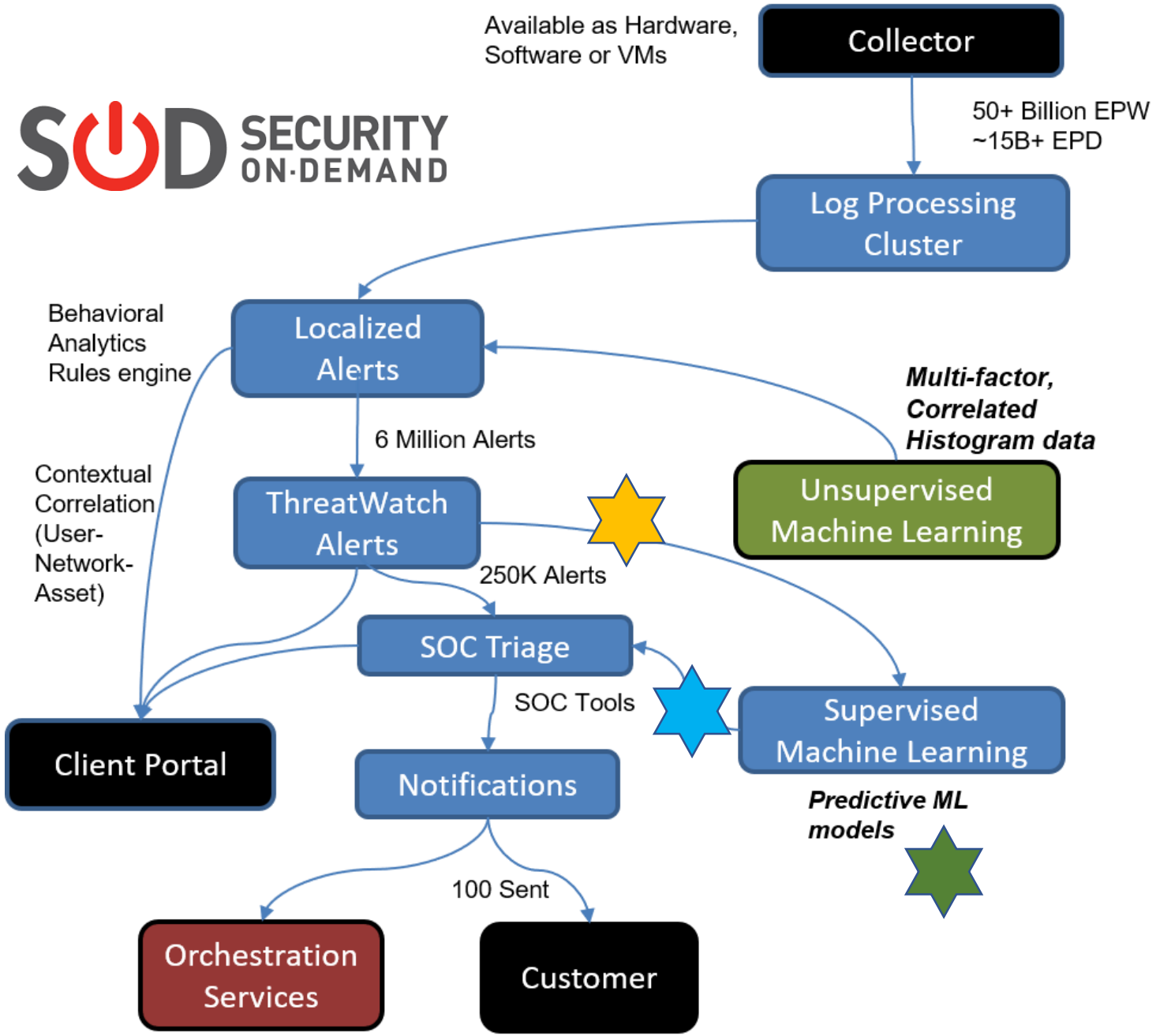
  „is a special case of machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points.”
  https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

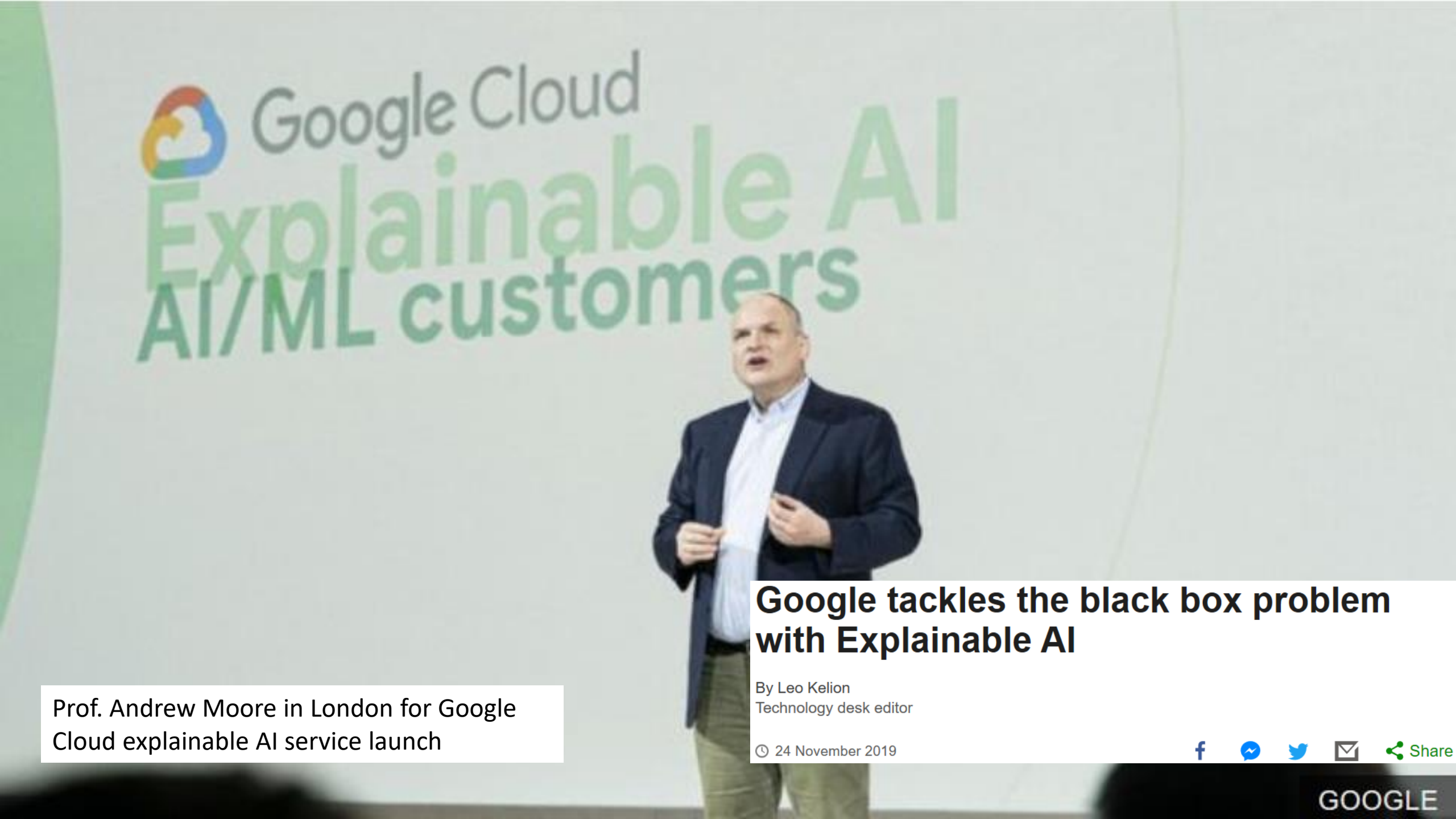
LITL (Label In The Loop)



Case Study (3)

- ML models may require extraction of additional features from far larger data (it may be possible only approximately)
- If the SOC analysts are encouraged to make decisions based on ML advices, they would like to know more about it
- And by the way, there is no guarantee that the initial training data set was correctly labeled!!!





Prof. Andrew Moore in London for Google Cloud explainable AI service launch

Google tackles the black box problem with Explainable AI

By Leo Kelion
Technology desk editor

24 November 2019

Share

There is still a lot to be done…

- Explaining to humans why AI/ML models… are not certain
- Explaining to humans why AI/ML models… make mistakes
- Explaining to AI/ML models what humans want from them

Concluding Remarks
(a screenshot borrowed from the keynote by Frank Buschmann)

IEEE BigData 2020 Cup: Predicting Escalations in Customer Support

Information Builders

2 months, 4 weeks from now

Predicting Escalations in Customer Support is a data mining challenge organized in association with the IEEE BigData 2020 conference. The task is to predict which cases in Information Builders' technical support ticketing system will be escalated in the nearest future by customers. The competition is organized jointly by Information Builders (https://www.informationbuilders.com/) and QED Software (http://www.qed.pl/).

Overview

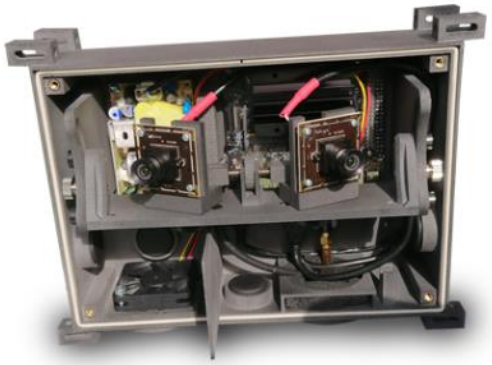
Technical Support Representatives of Information Builders strive to provide the highest quality level of support to their customers. At times, we may encounter situations where our support process and the needs of our customers conflict. When this occurs, undoubtedly, an escalation will arise. Every escalation is very disruptive to the support process. It changes the day to day activities of Technical Support Representatives, and more importantly, we have an upset customer. The ability to predict when an escalation may arise will allow us to react and do what's possible to prevent an escalation, diffuse a potential problem, thus maintaining customer satisfaction. We should be able to predict "when" an escalation occurs, it is also equally important to predict why an escalation is going to arise – is it due to a production outage, duration, technical proficiency, project deadlines or other issues. Depending upon the type of escalation, we will be able to build differing support processes that can be best suited to prevent an escalation.

This competition – aiming at building models that predict whether particular customer success cases are going to escalate in future based on information about their up-to-now history – is an important step for Information Builders to provide their customers with better services relying on modern machine learning solutions.

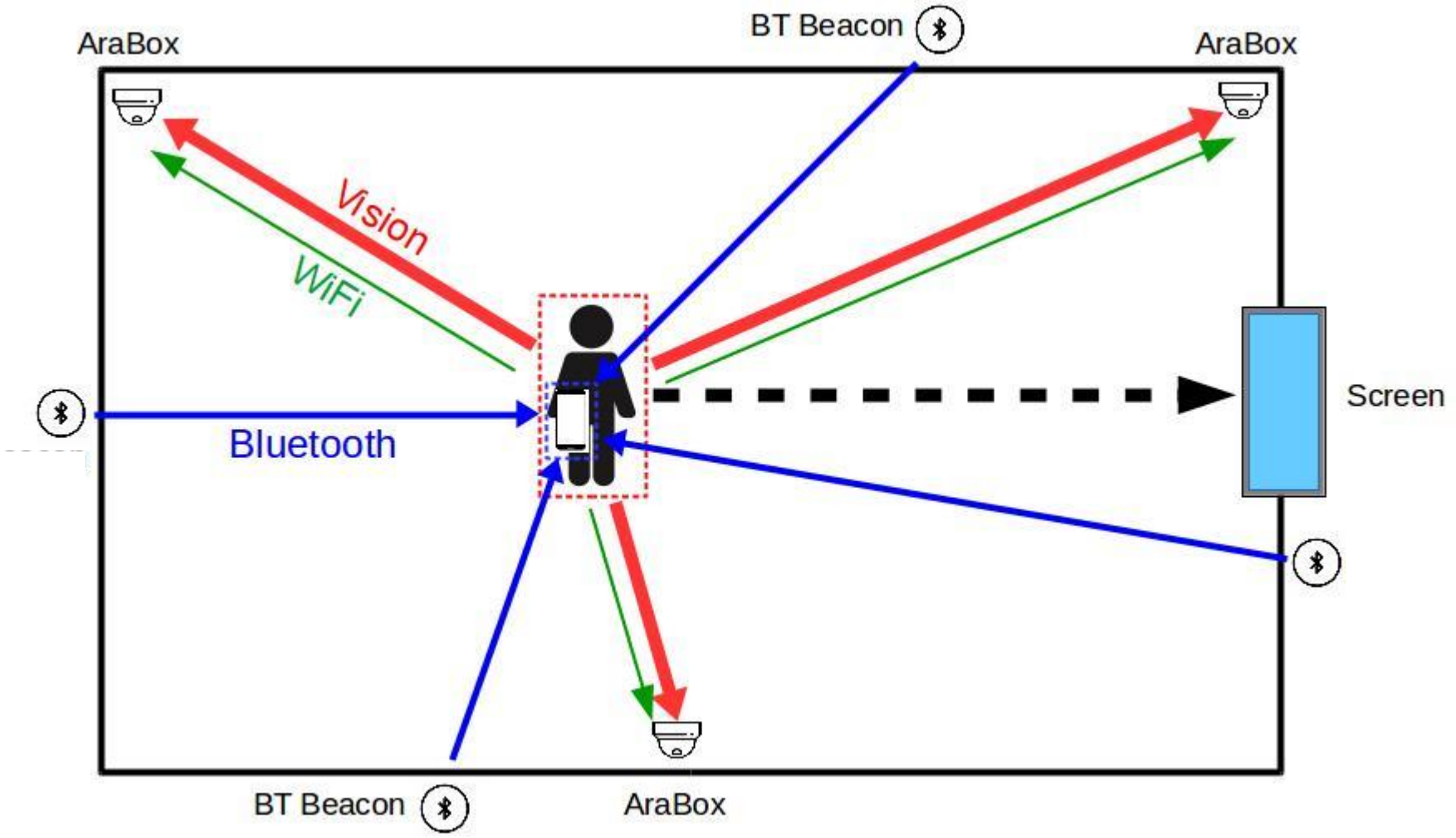
More details regarding the task and the description of the challenge data set can be found in the **Task description** section.

Special track at IEEE BigData 2020. A special session devoted to the challenge will be held at the IEEE BigData 2020 conference. We will invite authors of selected challenge reports to extend them for publication in the conference proceedings (after reviews by Organizing Committee members) and presentation at the conference. The publications will be indexed in the same way as regular conference papers. The invited teams will be chosen based on their final rank, innovativeness of their approach, and quality of the submitted report.


MyLED
FAJNIE, ŻE JESTEŚ
OUT OF HOME
MyLED
flying tiger
INFOLINIA: 22 36 838 88
SOFTWARE
00:09
00:11



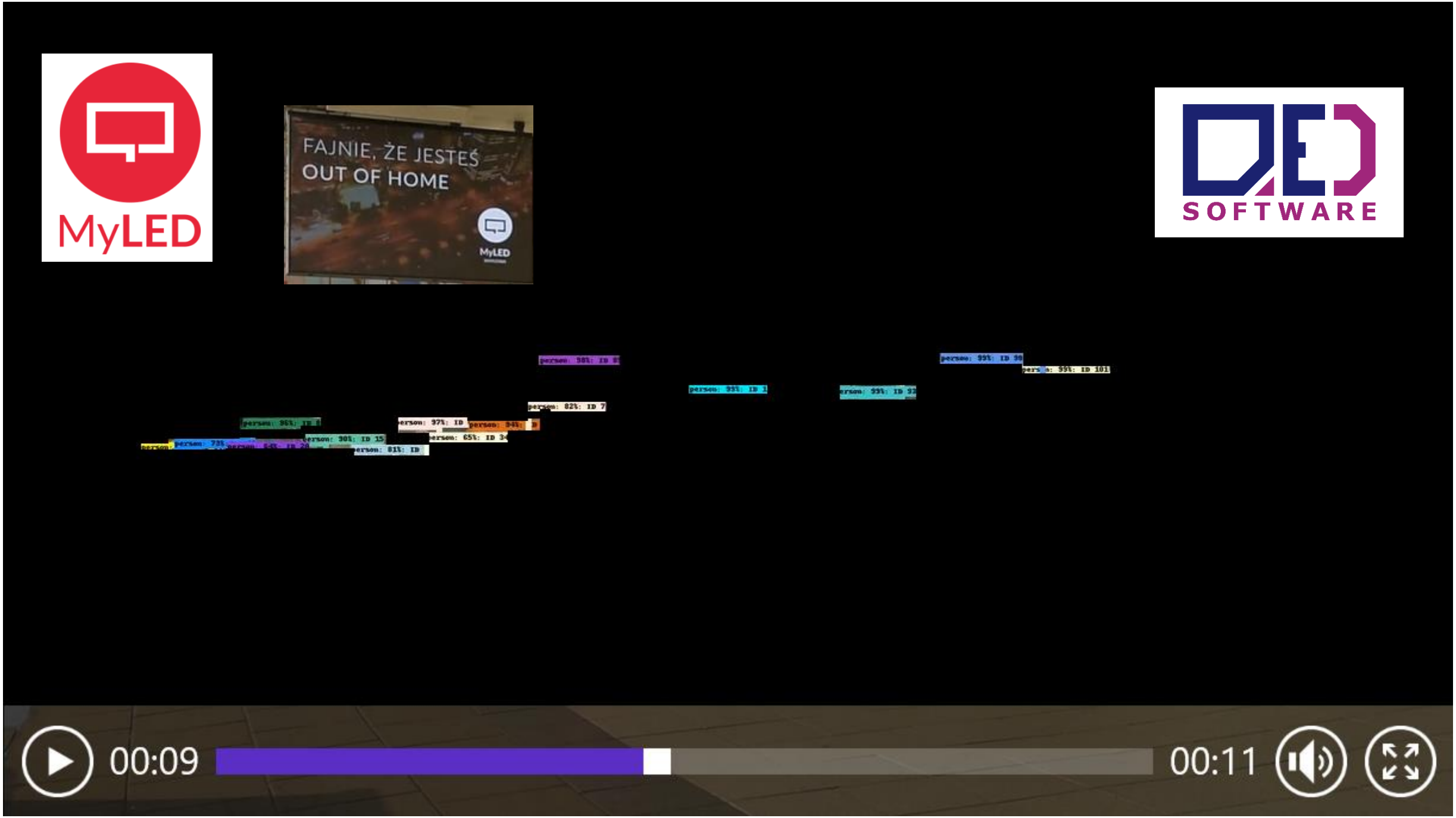
How about encoding video stream on-the-fly?



SMALL BIG DATA

https://arahub.ai/

SOFTWARE


MyLED
FAJNIE, ŻE JESTEŚ
OUT OF HOME
MyLED
SOFTWARE
00:09
00:11

UNIVERSITY OF WARSAW

QED SOFTWARE

Thank you!

slezak@mimuw.edu.pl

dominik.slezak@qed.pl

SOD SECURITY ON-DEMAND

knowledgePit
A Data Challenge Platform