I See What you See: Real Time Prediction of Video Quality from Encrypted Streaming Traffic

Sarah Wassermann, Michael Seufert*, Pedro Casas

AIT Austrian Institute of Technology @Vienna

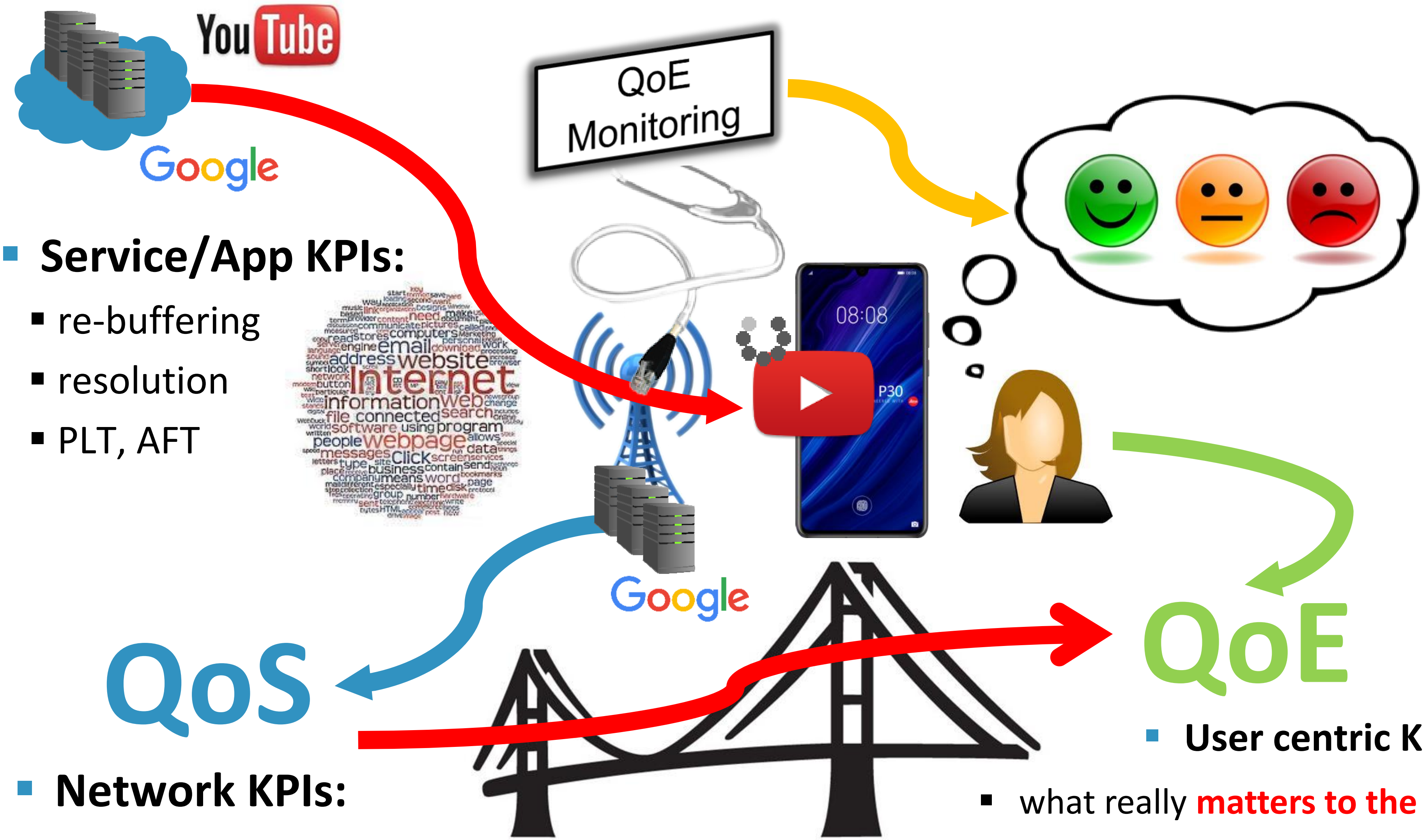
Institute of Computer Science, Würzburg University

Li Gang, Kuang Li

Huawei Technologies R&D @Shenzhen

A Bit of Context – QoE Monitoring (ISP PoV)

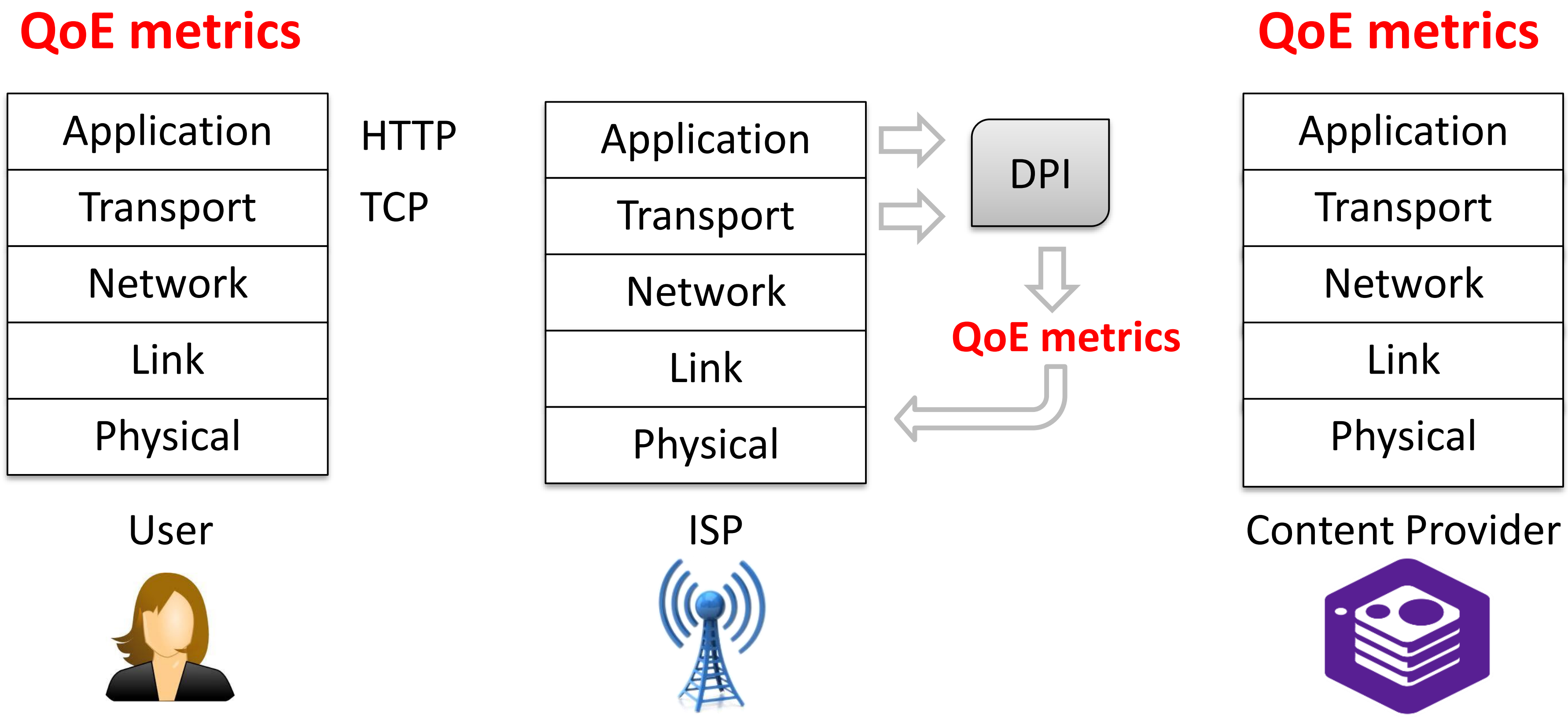
- **Service/App KPIs:**
  - re-buffering
  - resolution
  - PLT, AFT



- **Network KPIs:**
  - throughput, latency, packet loss, jitter

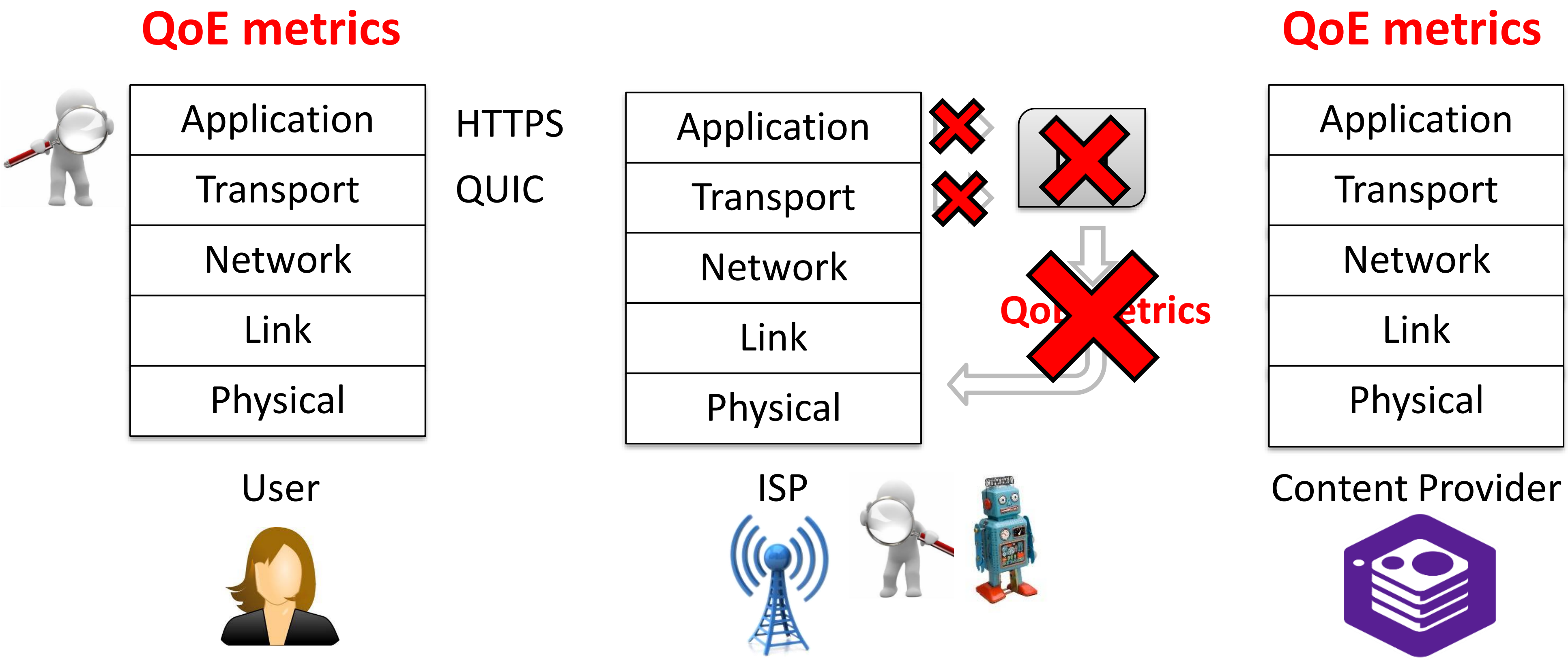
- **User centric KPIs:**
  - what really **matters to the end-user**
  - **responsiveness**, **interactivity**, **availability**, **acceptability**, **satisfaction**, **engagement**, **churning**

The Rise of End-2-End Encryption



- QoE monitoring approach: with non-encrypted traffic, DPI-based approaches:
  - ***"YOUQMON: A System for On-line Monitoring of YouTube QoE in Operational 3G Networks"***
  - ***"Monitoring YouTube QoE: Is Your Mobile Network Delivering the Right Experience to your Customers?"***
  - ***"Passive YouTube QoE Monitoring for ISPs"***

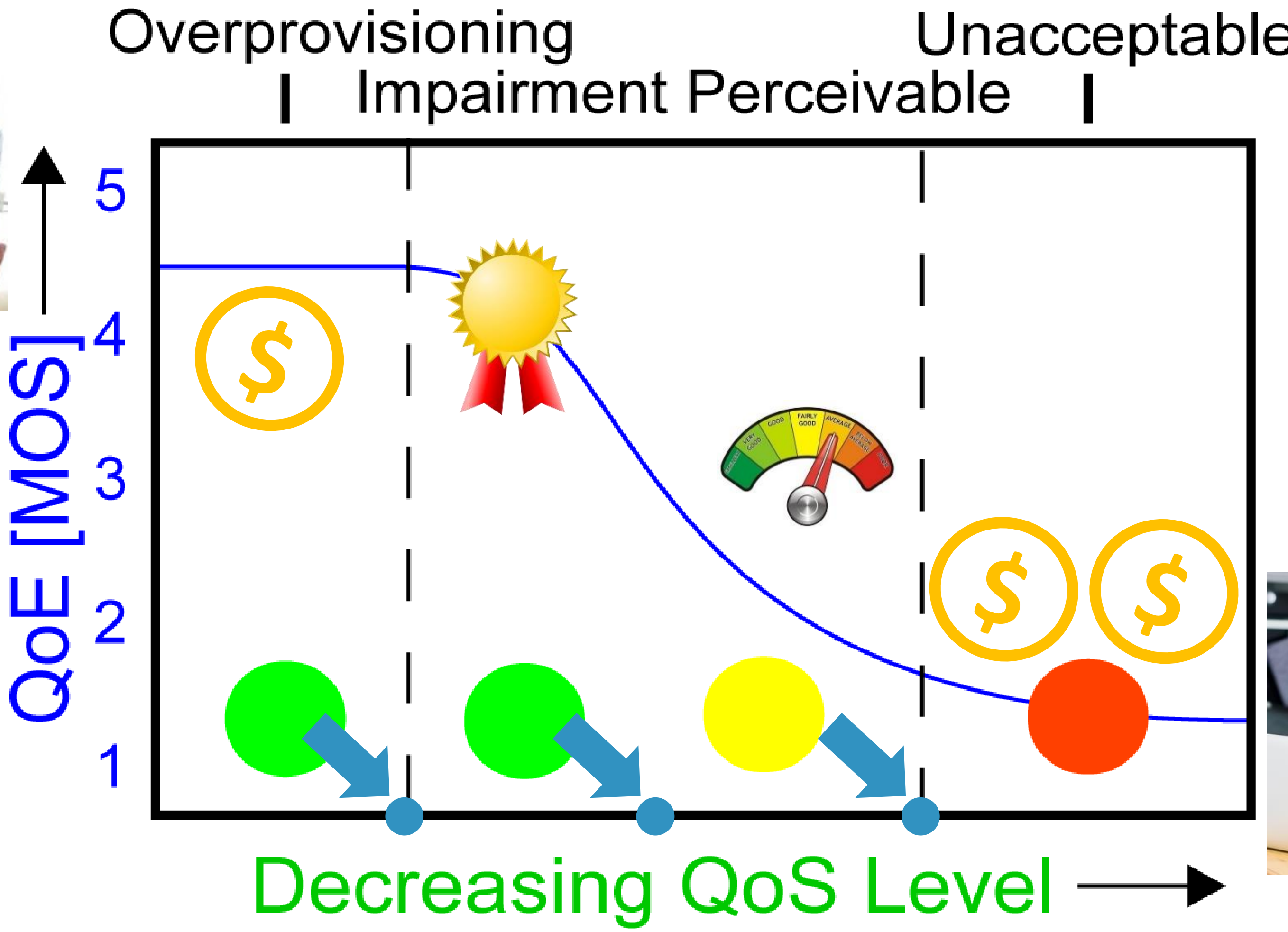
The Rise of End-2-End Encryption

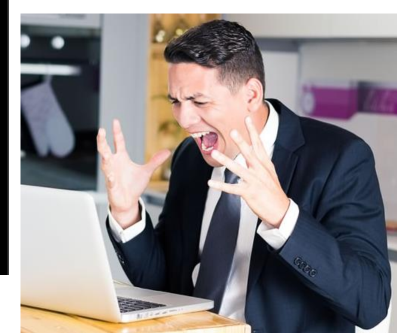


- **HTTPS** and **QUIC** turn previous approaches no longer applicable – **lack of visibility for ISPs**
  - Solution I – **monitoring directly at the end devices**
  - Solution II – **monitoring at the core, relying on Machine Learning (ML) approaches**

Why is QoE so Relevant?

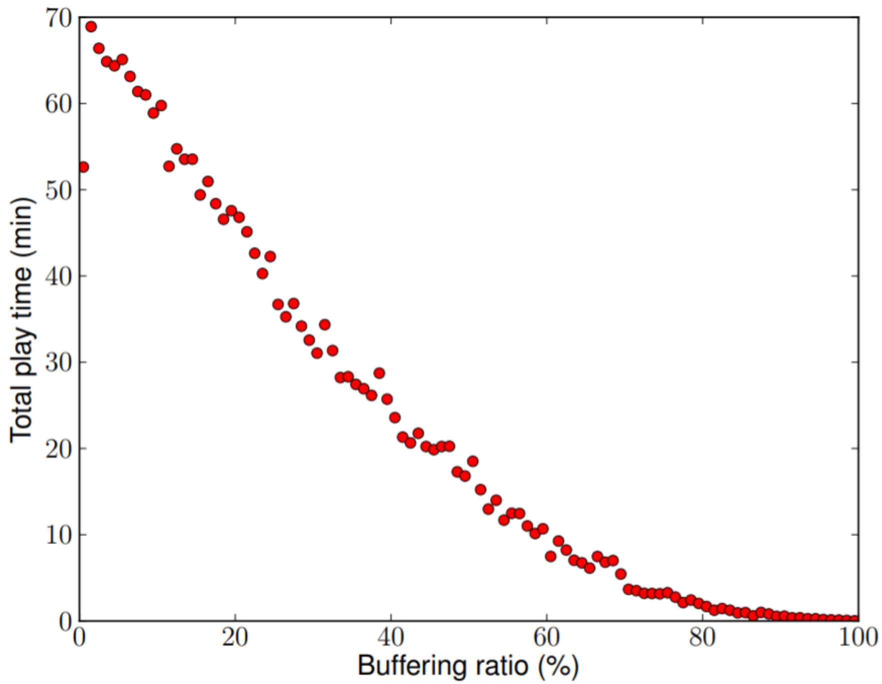
Dimensioning & Operation





Non-linearities **and** ***saturation*** effects = **typical for QoE**

Why is QoE so Relevant?
User Engagement



Total video play time vs.
re-buffering ratio

"Understanding the Impact of Video Quality on User Engagement"
@SIGCOMM'11

- **Poor QoE** significantly **reduces user engagement**
- **Increase of the buffering ratio** of only **1%** can lead to more than **three minutes of reduction** in the user engagement

Why is QoE so Relevant?
Customer Experience

- **Marketing driver:** intensifying competition in telecom markets
- **Customer perception** and judgement becoming **increasingly relevant**

- Avoid **customer churn** for quality dissatisfaction
- **Attract new customers** with better service provisioning **(NPS vs. MOS)**
- Understand **what matters the most to customers**

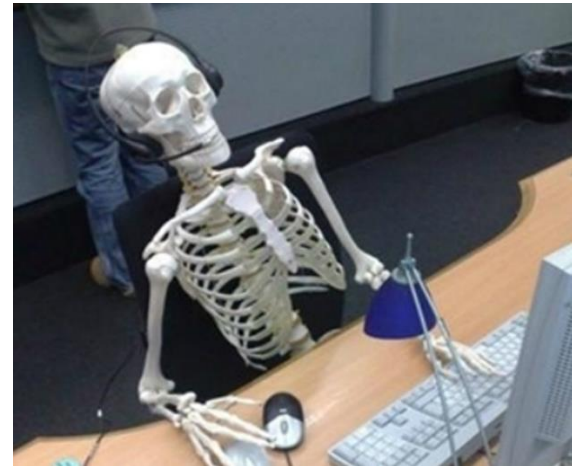
What Happens when QoE Degrades?

- An example: **what happens when latency increases too much in web browsing?**

- *Google* – **Inter-domain routing changes** cause more than **40% of the cases in which clients experienced** a **latency increase of at least 100 ms**

- *Amazon* – **every additional 100 ms of page load time** could cost them **1%** of their **sales**, and **a page load slowdown of just one second** could turn into a **$1.6 billion loss in sales each year**

- *Google* – **slowing search results down by 400 ms**, they could **loose 8 million searches per day** → Google Ads!



What do we Need from the E2E Network?

- Video Streaming
- 360º Streaming

- QoS – ***downlink bandwidth***
- User-perceived – ***re-buffers***

- Web Browsing

- QoS – ***latency***
- User-perceived – ***ATF time***

What do we Need from the E2E Network?

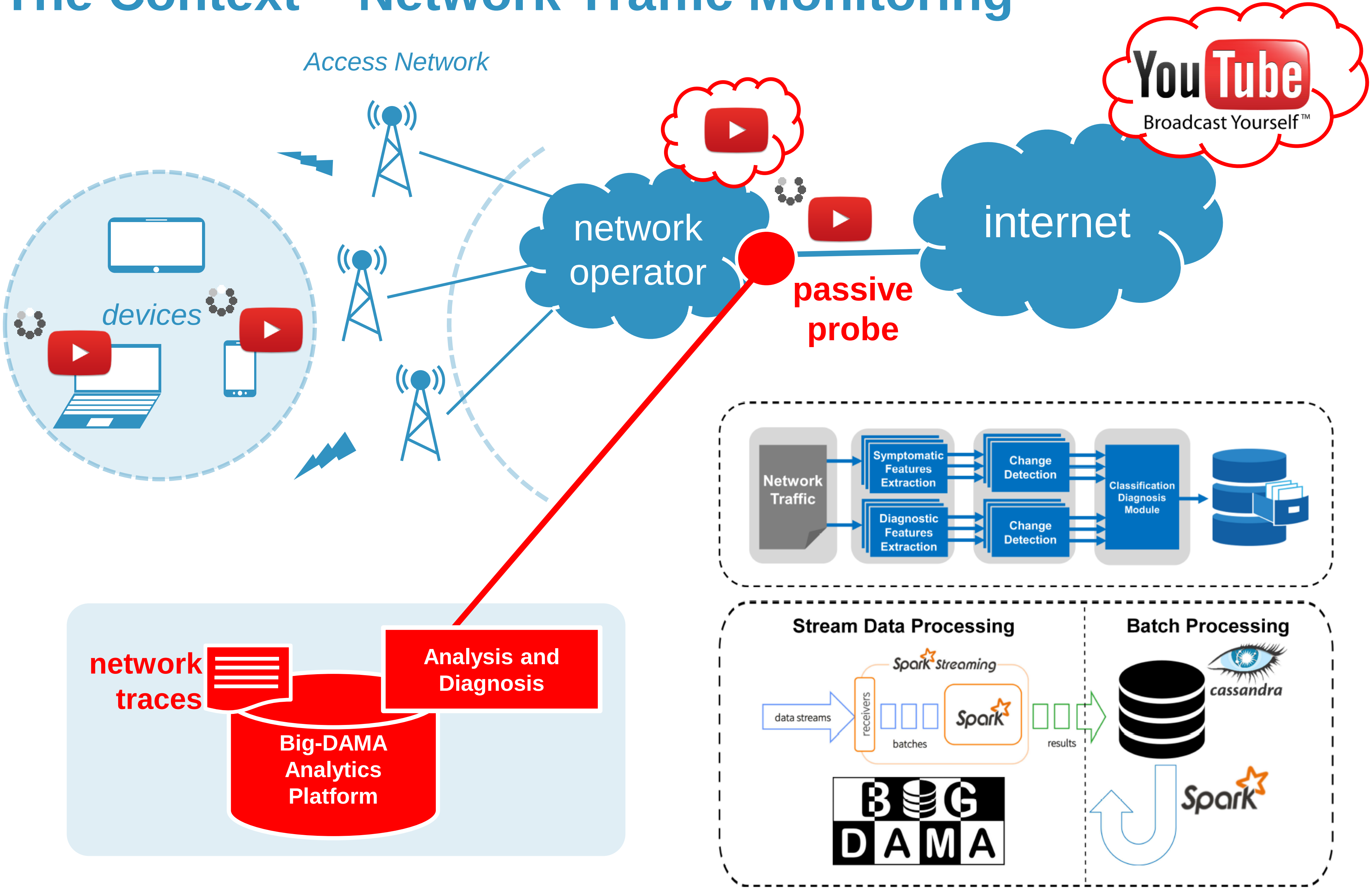
- Cloud Services
- QoS – ***downlink bandwidth/latency***
- User-perceived – ***responsiveness***

File Sync and Storage	E-Mail	CRM	Virtual office	Remote Virtual Desktop	Tele-presence	Gaming	Virtual Life

Low → **High**

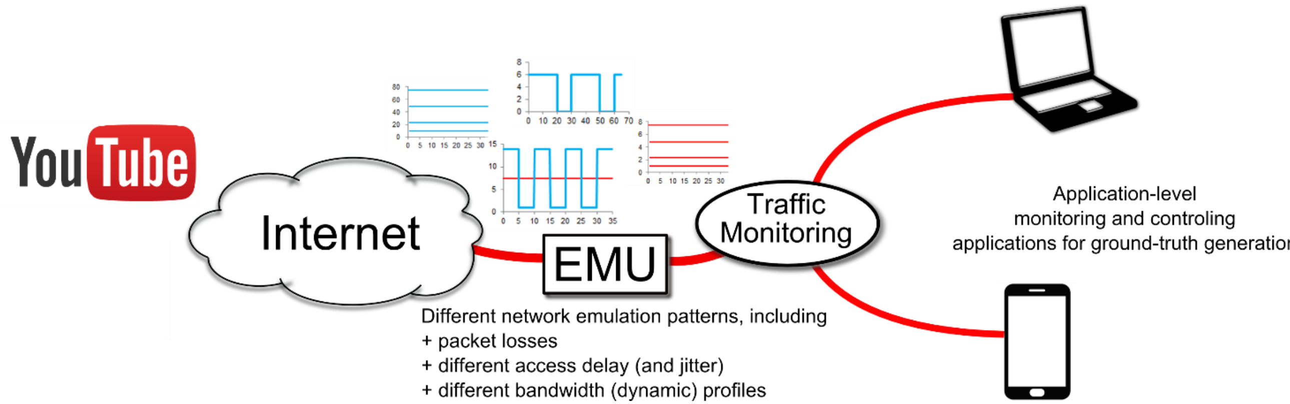
Degree of Interactivity

The Context – Network Traffic Monitoring



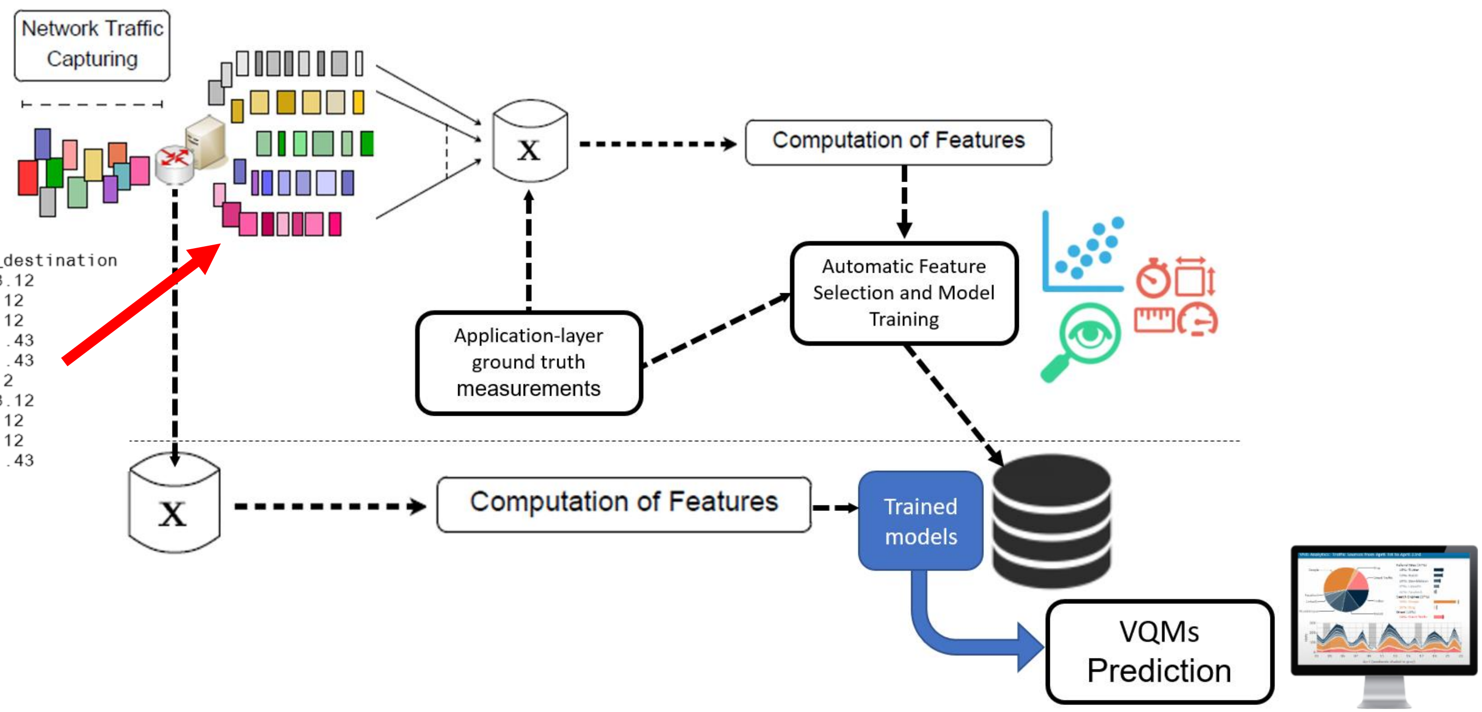
Methodology – Data Generation, Model Training and Execution

- **Fully controlled testbed:** generating and measuring all relevant metrics at the different layers of the communications stack.

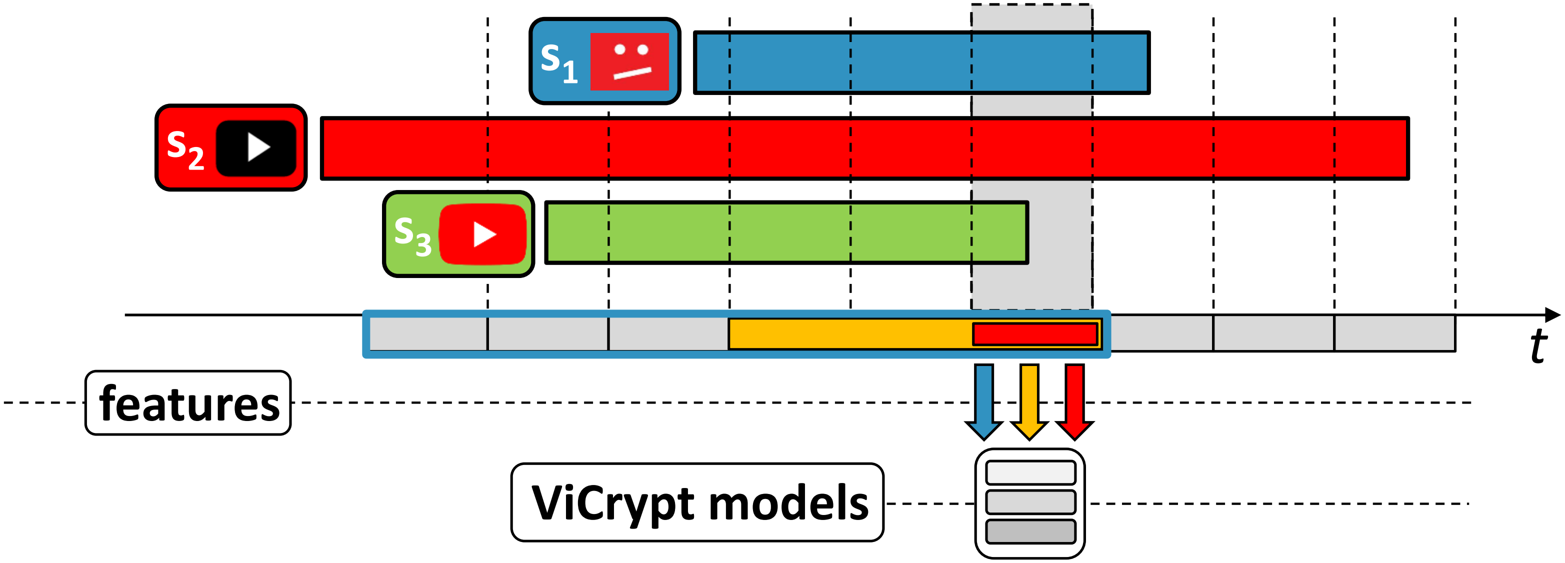


- **Using the generated datasets** to **build, train** and later on **execute different machine learning based models** for VQM prediction and monitoring.

```
time, packet_size (bytes), protocol, ip_source, ip_destination
1519892311.852817000,1396,17,192.168.1.43,212.247.8.12
1519892311.853004000,982,17,192.168.1.43,212.247.8.12
1519892311.853093000,700,17,192.168.1.43,212.247.8.12
1519892311.872293000,1396,17,212.247.8.12,192.168.1.43
1519892311.872889000,1396,17,212.247.8.12,192.168.1.43
1519892311.873728000,84,17,192.168.1.43,212.247.8.12
1519892311.880288000,1396,17,192.168.1.43,212.247.8.12
1519892311.880296000,978,17,192.168.1.43,212.247.8.12
1519892311.880315000,696,17,192.168.1.43,212.247.8.12
1519892311.895292000,1396,17,212.247.8.12,192.168.1.43
```



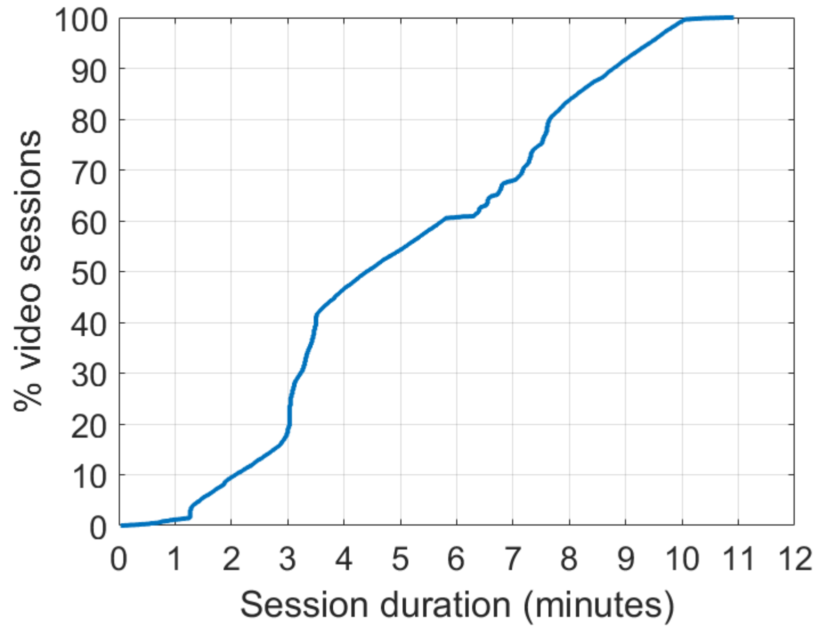
Stream-based Prediction of YouTube QoE

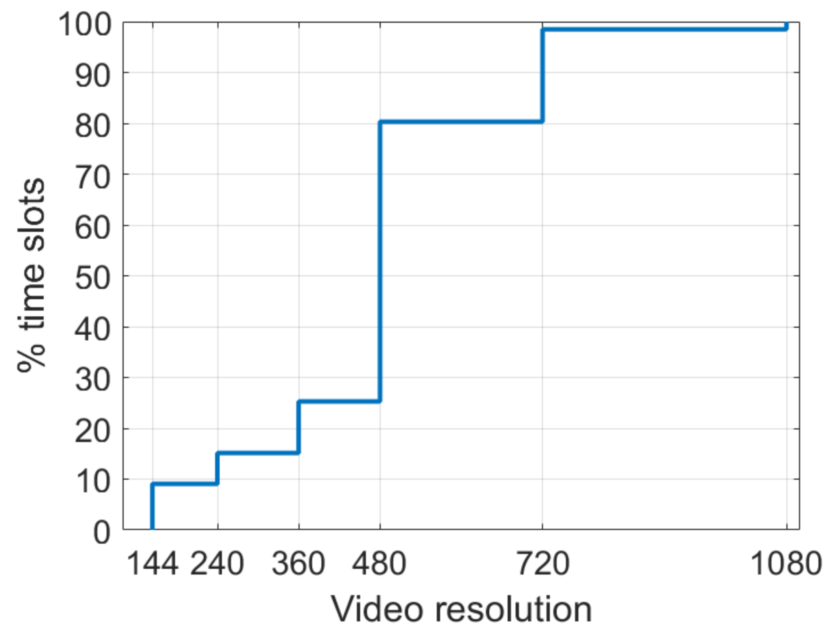


- **Video stream-based analysis,** using multiple sliding windows, capturing different temporal phenomena (**current time**, **short-term trend**, **session-aggregated**)
- **Analysis is done in real time:** for every video session and **for every new time slot of 1 second**, we consider the following sets of features (**207 in total**):
  - Features extracted from **current time slot (C) – 69 features**
  - Short-memory (trend) based features, extracted from **last T (3) slots (CT) – 69 features**
  - Cumulative based features, extracted from **all past traffic for this video session (CS) – 69**
- **Feature computation is done continually, in constant-memory boundaries, using sketches**
- ***Machine learning models*** trained for prediction of **re-buffering events**, **video resolution**, **video bitrate**

Dataset Description

- **15.000+ YouTube video sessions** streamed and recorded in summer 2018
- **JavaScript-based monitoring** script to measure ground truth
- Home and corporate WiFi networks, LTE mobile networks
- **QUIC and TCP sessions**
- Bandwidth limitations: 20Mbps, 5Mbps, 3Mbps, 1Mbps, 300kbps + fluctuations
- Different ISPs, different geographic locations (Italy, Austria, Germany)
- **Prediction task:** per second video resolution, ***6-classes classification*** – 144p, 240p, 360p, 480p, 720p, 1080p





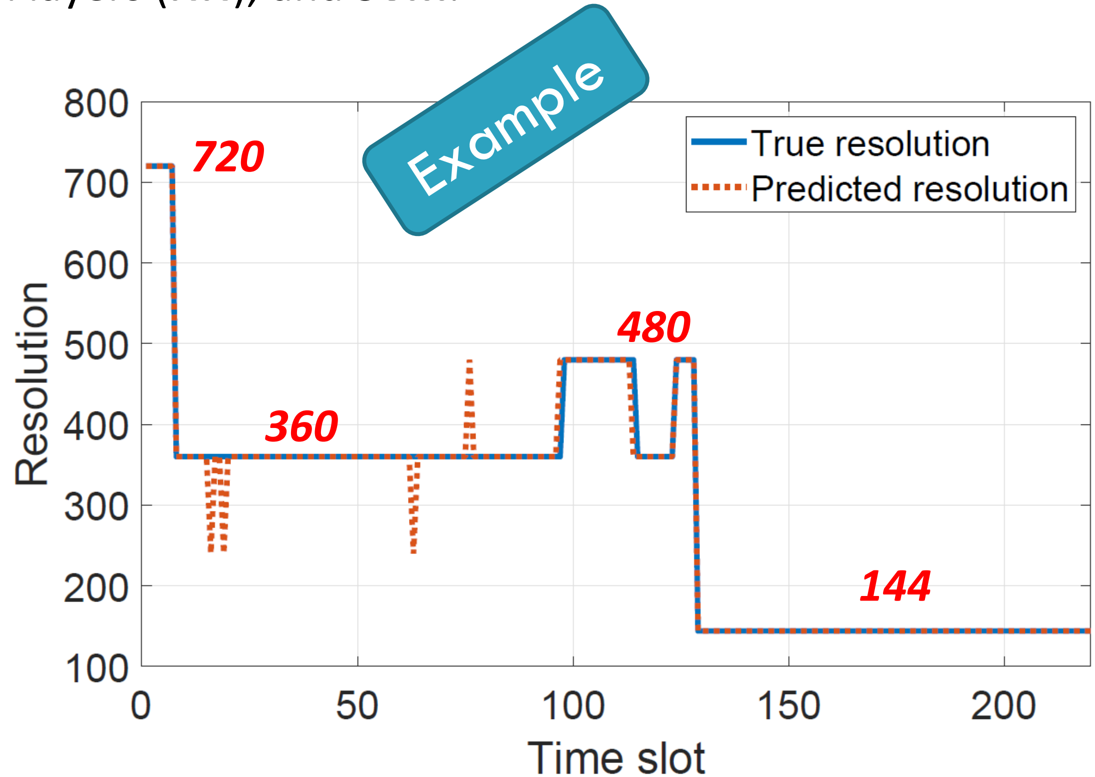
On-line Prediction of Video Resolution

- More than 4.6M individual, 1 second slots for training (***5-fold cross validation***)
- **Benchmarking of 9 ML models:** decision trees (**DT**), random forests with 10 trees (**RF10**), Adaboost using 50 trees (**ADA**), an ensemble with 10 extremely randomized trees (**ERT10**), bagging with 10 trees (**BAGGING**), Naïve Bayes (**BAYES**), k-nearest neighbors with k= 5 (**KNN**), feed forward neural networks with 3 hidden layers (**NN**), and **SVM**.

	Training time (min)	Accura
DT	43	92
RF10	2	92
ADA	125	68
ERT10	1	90
BAGGING	37	95
BAYES	1	42
KNN	9	73
NN	507	58
SVM	194	54

Benchmarking of different ML models

- ***For the sake of speed, we use RF10 as underlying model***

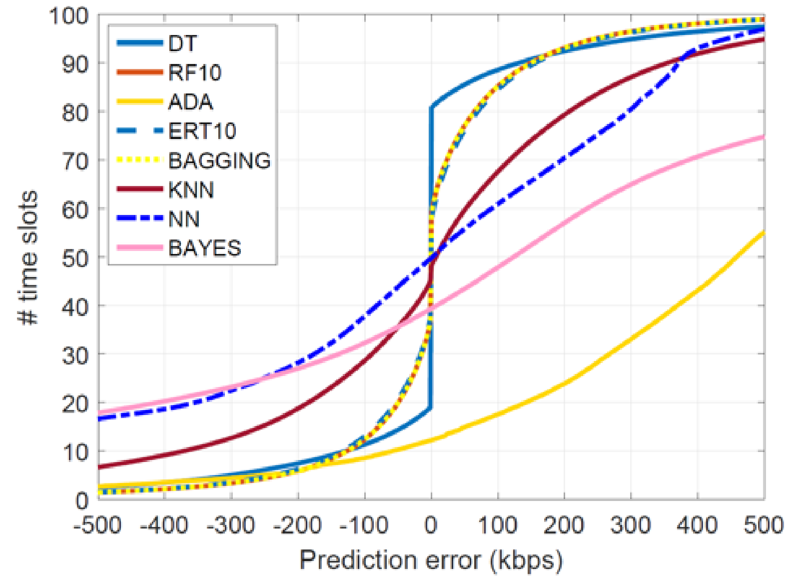


Video-resolution prediction.

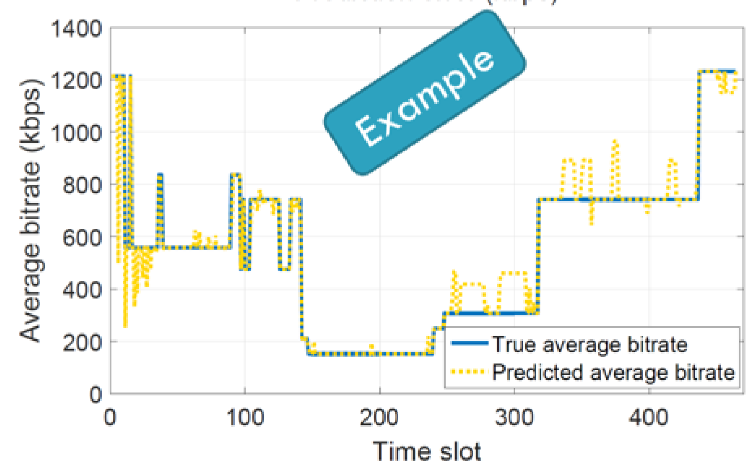
On-line Prediction of Video Bit-Rate

- **Regression task:** estimation of per second average video bitrate

	5-CV time (minutes)	MAE (kbps)	RMSE (kbps)	MRE (%)	PLCC
DT	31	94	246	18	0.88
RF10	36	89	179	18	0.93
ADA	126	492	573	130	0.59
ERT10	7	93	182	19	0.93
BAGGING	22	89	179	17	0.93
BAYES	3	2,540	6,530	545	-0.14
KNN	6	229	353	42	0.70
NN	305	333	489	70	0.20
SVM	143	10^{23}	$2 \cdot 10^{23}$	$2 \cdot 10^{23}$	0.12



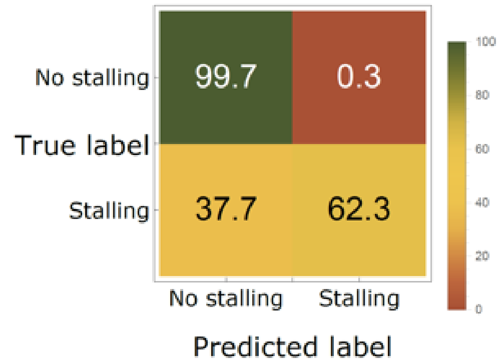
- ERT10 & BAGGING realize **MAE below 100kbps**, and **RMSE below 190kbps** (penalizes larger errors)
- **80% of the slots** are estimated with **errors below 100kbps**
- Predictions are ***highly correlated*** with the target (***PLCC = 0.93***)

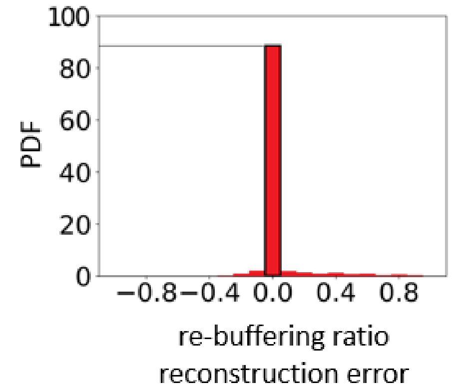


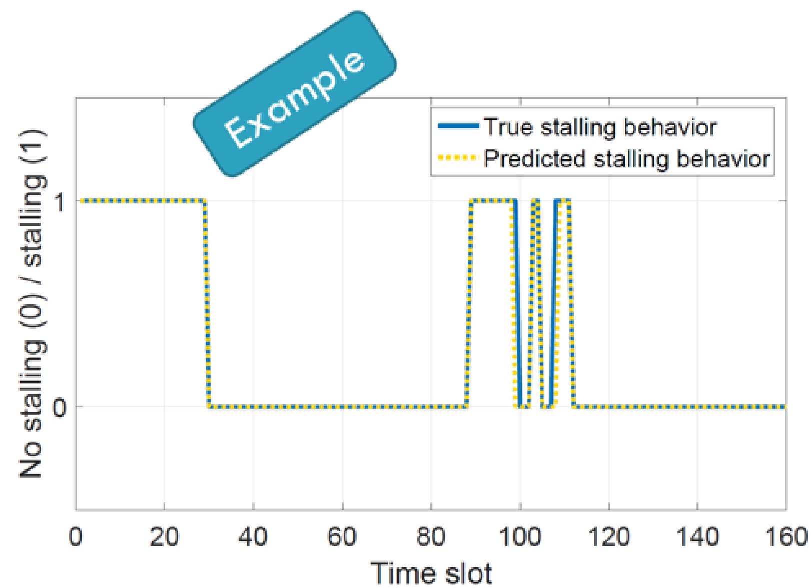
On-line Prediction of Video Stalling

- **Binary classification task:** playback stalled/not-stalled at every new slot

	Accuracy (%)	Recall (%)	Precision (%)	5-CV time (minutes)
DT	96	64	68	57
RF10	97	55	88	3
ADA	95	29	61	154
ERT10	97	54	88	1
BAGGING	97	65	87	63
BAYES	50	86	9	1
KNN	96	48	71	10
NN	94	0	0	600
SVM	84	62	21	36
ISO10	86	13	8	4
LOF	86	11	6	46







- **per-slot re-buffering** estimation errors are high, **stalling slots under-estimated…**
- …but **estimation of re-buffering ratio** is ***perfect for almost 90% of the videos***

Impact of Feature Selection

- Impact of different feature sets on classification performance
- ***F_C*** *– features in current slot,* ***F_T*** *– last T (3) slots,* ***F_S*** *– cumulative session slots*
- ***$F_{DOWN/UP}$*** *– all features downstream/upstream*
- ***F_{TOP20}*** *– top-20 features by feature selection*

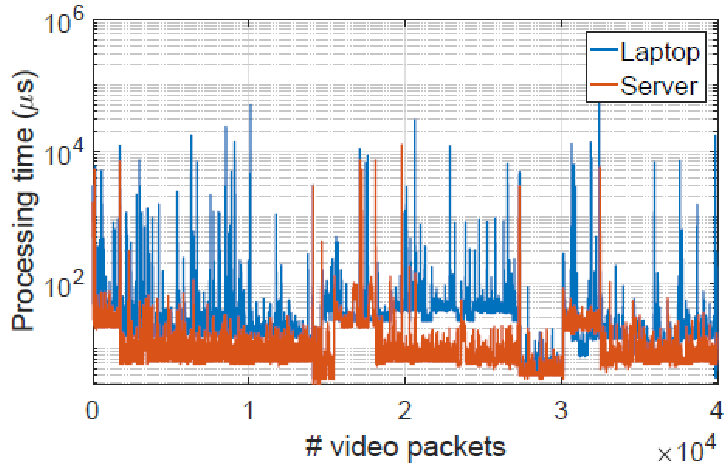
Features	# Features	Accuracy (%)
F_C	69	70
F_T	69	73
F_S	69	96
F_{DOWN}	81	90
F_{UP}	81	90
F_{TOP20}	20	95

All features	*207*	*92 %*

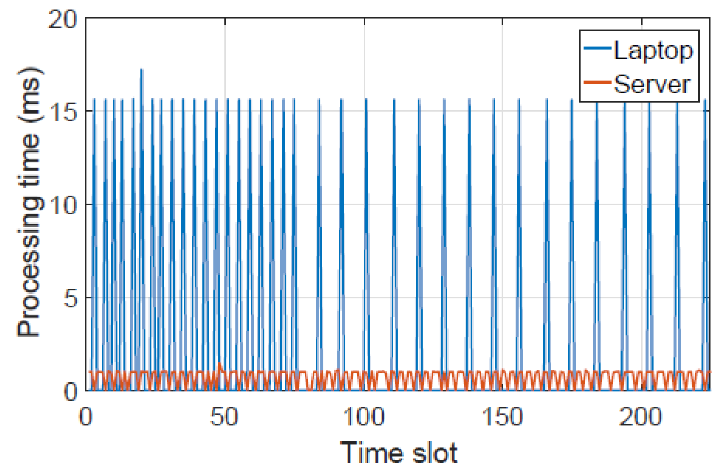
- The top-20 features provide the best trade-off

- The **longer the memory** for feature computation, the **higher the accuracy**
- *Cumulative session-based features (**F_S**)* are the most relevant feature set, ***improving by 4% the performance obtained by all 207 features***

Computational Time Analysis – RF10 Real Time



Features update time at each new packet.



Prediction time.

- Evaluation of ***full feature set update time*** (done for every new incoming packet) and ***prediction time*** (done for every 1s slot), using an upper bound with all 207 features.
- **Laptop** (i5 CPU, 8GB RAM) vs. **Server** (Xeon Silver, 48 cores, 128GB RAM)
- On server, **average duration of full feature set update is 13 µs**, **prediction time below 1.4ms**
- On laptop, **average feature update duration takes 37 µs**, **prediction time below 16ms**
- ViCrypt ***can perform video-resolution predictions in real time***, with an end-to-end computational delay way below the time slot length of 1 s

https://bigdama.ait.ac.at/

http://mobiqoe.ait.ac.at/

Q&A…

AIT Austrian Institute of Technology @Vienna

pedro.casas@ait.ac.at

http://pcasas.info