The complex interplay between

Data and Model Quality

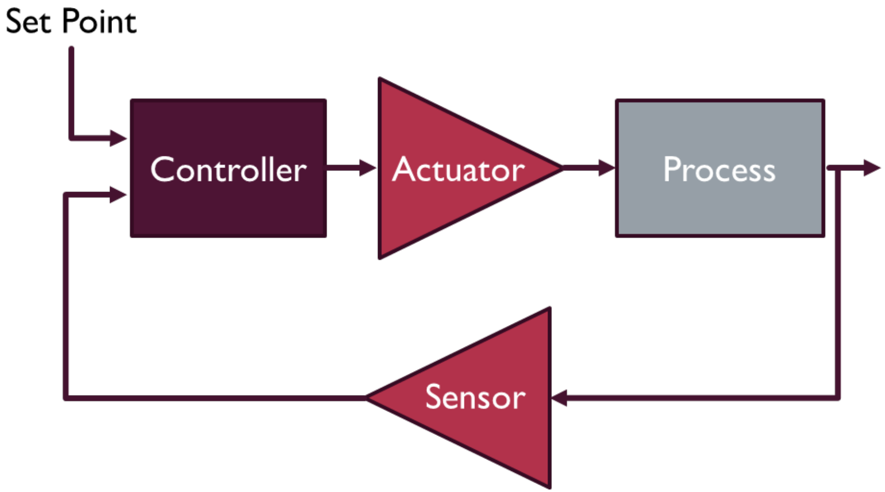
a case study in traffic anomaly detection

IETF 121 NMRG

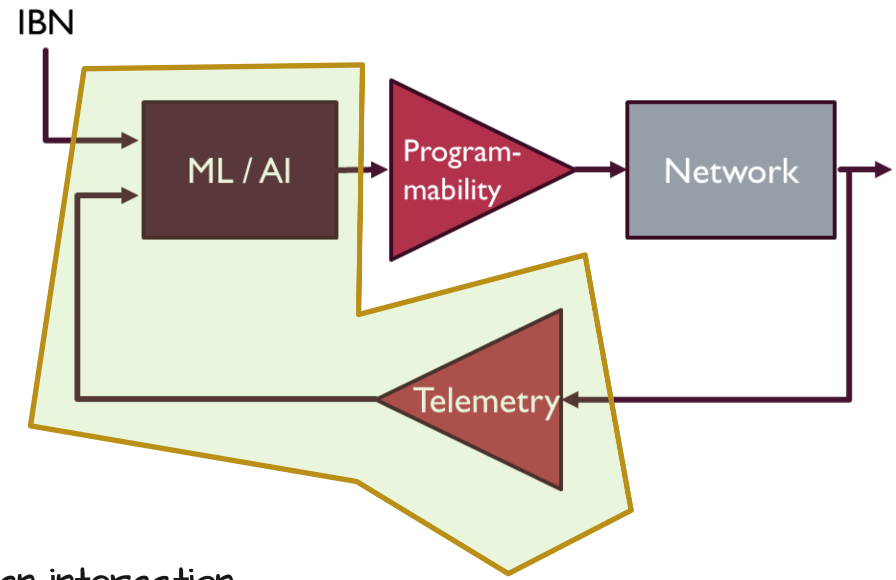
Prof. José Camacho
josecamacho@ugr.es
Computational Data Science Laboratory (CoDaS Lab)

The closed loop analogy

Industrial Feedback Controller



Autonomic Network

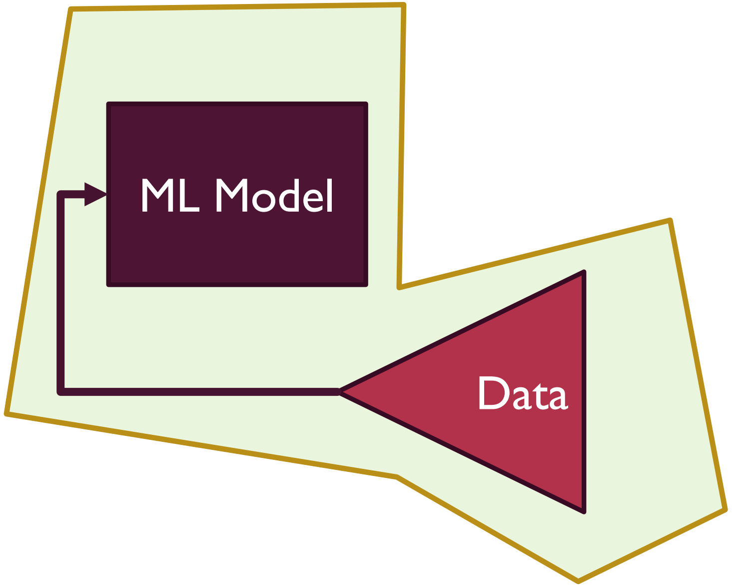


Minimize human interaction

(Data) Quality in / (Model) Quality out

Can we **automatically ASSESS** data-model quality interaction? How?

Use this to correct for low quality data / optimize data quality



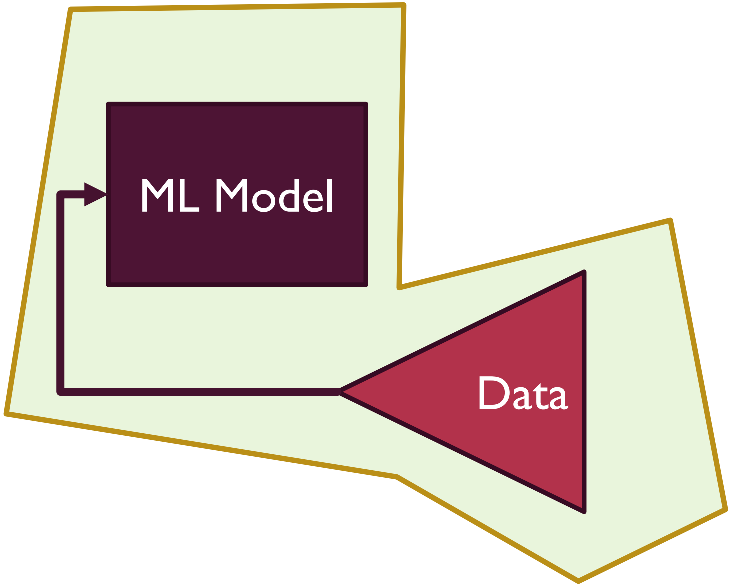
DQ is multifaceted

accuracy, completeness, validity, timeliness, consistency, correctness, uniqueness, reliability, ...

(Data) Quality in / (Model) Quality out

ML prespective:

DQ defined as "the best training data to **optimize model performance**"

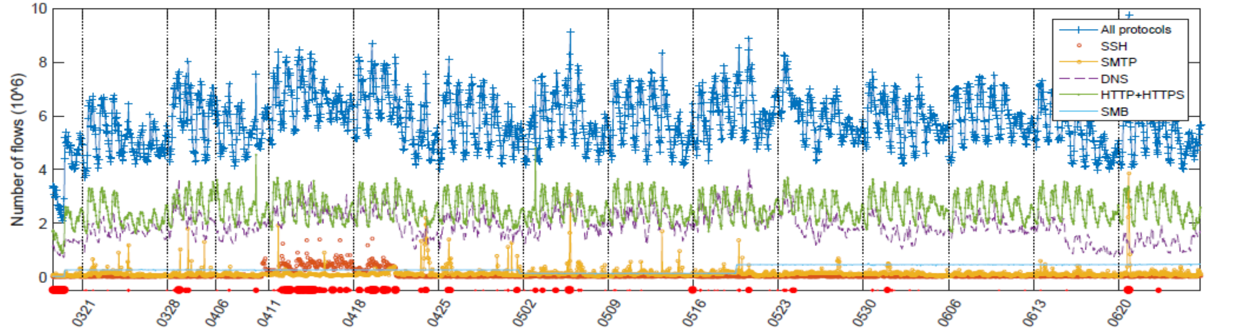


UGR'16[1]

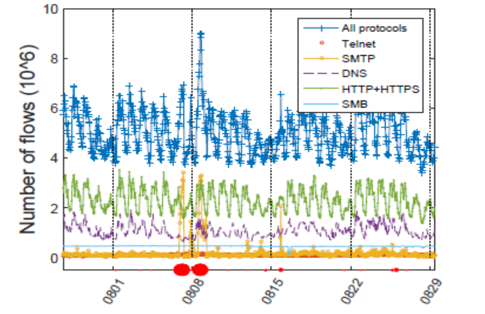
Performance defined as "the **most accurate** identification of traffic anomalies" (<u>perceived</u>)

[1] Macia-Fernandez, G., Camacho, J., Magan-Carrion, R., Garcia-Teodoro, P., Theron, R. UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs. Computers & Security **2018**, 73, 411–424.

UGR'16



(a) Calibration Set



(b) Test Set

TABLE IV
FEATURES OF THE CALIBRATION AND THE TEST SETS.

Feature	Calibration	Test
Capture start	10:52h 03/18/2016	13:43h 07/27/2016
Capture end	18:27h 06/26/2016	09:27h 08/29/2016
Attacks start	N/A	00:00h 07/28/2016
Attacks end	N/A	12:00h 08/09/2016
Number of files	17	6
Size (compressed)	181GB	55GB
# Connections	≈ 13,000M	≈ 3,900M

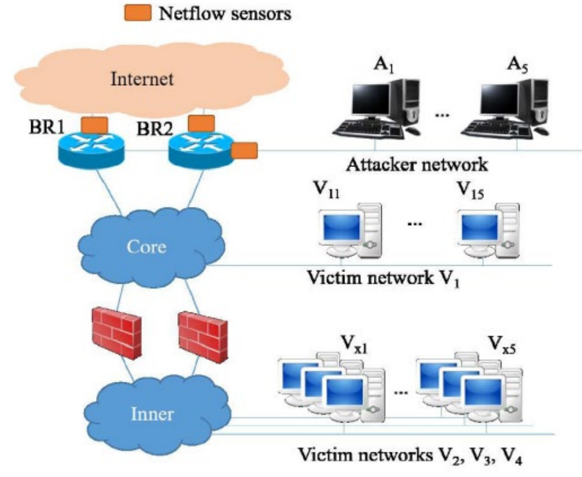


Fig. 3 – Topology of the network.

Our Goal

Understand

challenges

for autonomic optimization of data quality using a

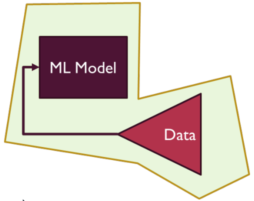
benchmark

for anomaly detection: the UGR'16 dataset, where

perceived detection performance

is the goal

Automate the ML PIPELINE



DATA COLLECTION

- Sensor placement
- Data capture, transm. & storage

PRE-PROCESSING

- Feature engineering
- Data cleaning

ML TRAINING

- ML choice
- Meta-par. optimization
- Data labelling

MODEL EVALUATION

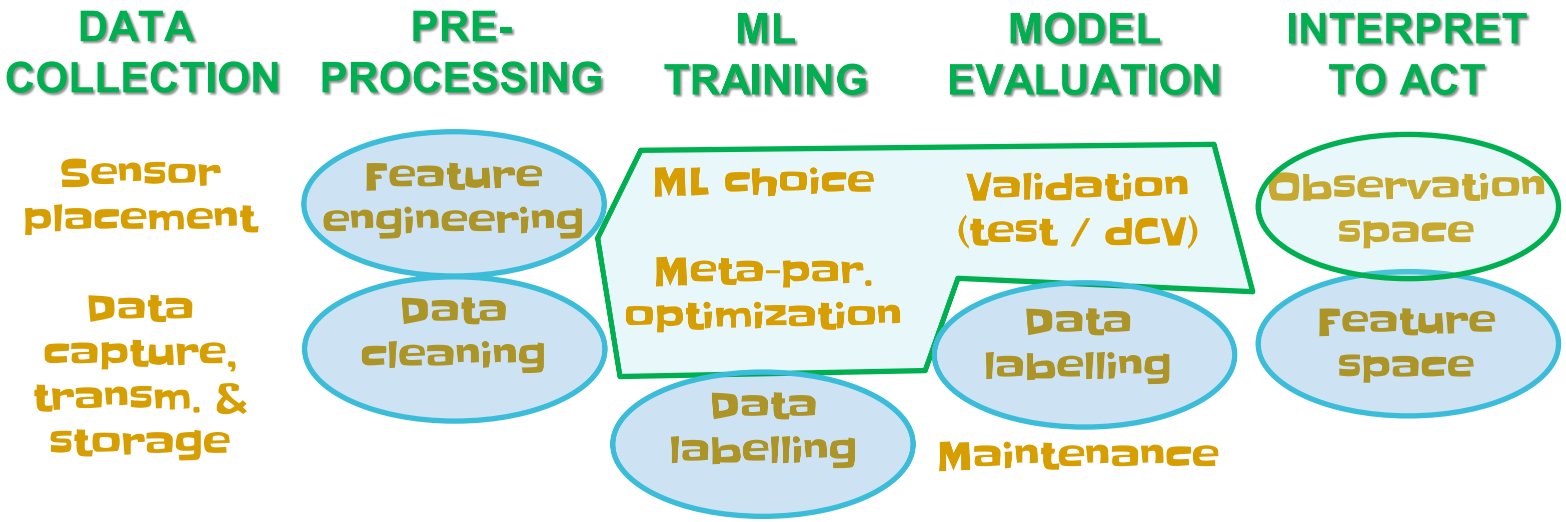
- Validation (test / dCV)
- Data labelling
- Maintenance

INTERPRET TO ACT

- Observation space
- Feature space

TABLE II
UGR'16 DATASET VARIANTS.

Label	Training	Type of flows	Anonymized flows
UGR'16v1	March to June	Unidirectional	No
UGR'16v2	March to May	Unidirectional	No
UGR'16v3	March to May	Bidirectional	Yes
UGR'16v4	March to May	Unidirectional	Yes



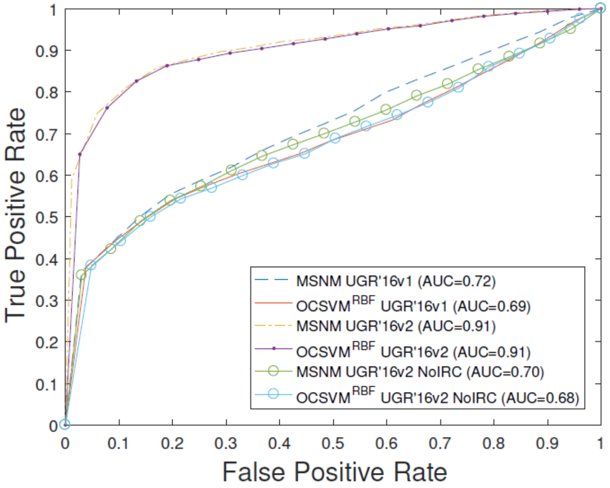
Camacho, J., Rodríguez-Gómez, R. Data quality tools to optimize an anomaly detection benchmark. Submitted to Data, 2024.

TABLE II
UGR'16 DATASET VARIANTS.

Label	Training	Type of flows	Anonymized flows
UGR'16v1	March to June	Unidirectional	No
UGR'16v2	March to May	Unidirectional	No
UGR'16v3	March to May	Bidirectional	Yes
UGR'16v4	March to May	Unidirectional	Yes

Data cleaning & labelling

Observation space



Data cleaning Labelling → +0.2 AUC

ML Method → +0.03 AUC

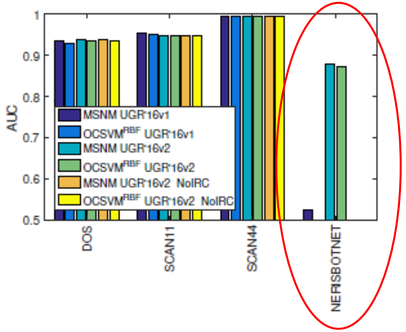
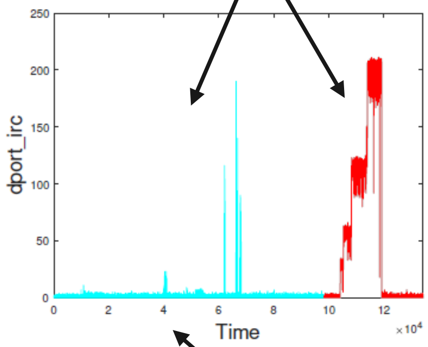


TABLE II
UGR'16 DATASET VARIANTS.

Label	Training	Type of flows	Anonymized flows
UGR'16v1	March to June	Unidirectional	No
UGR'16v2	March to May	Unidirectional	No
UGR'16v3	March to May	Bidirectional	Yes
UGR'16v4	March to May	Unidirectional	Yes

Training

UGR'16v1



UGR'16v2

Feature space

ML detection pattern

UGR'16v1 UGR'16v2

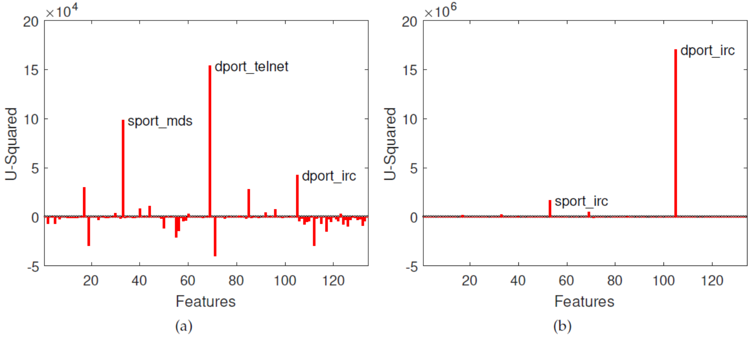
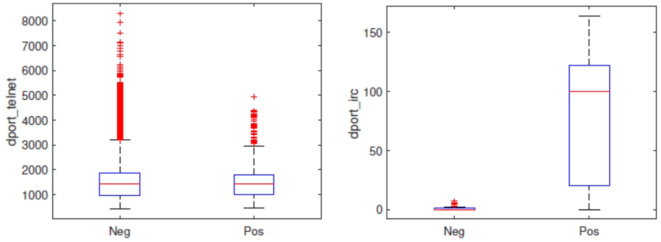


Figure 2. Comparison of U-Squared statistics for the NERISBOTNET attack using as a reference UGR'16v1 (a) and UGR'16v2 (b).

Test



(a) Ttest $p\_value = 0.47$ (b) Ttest $\mathbf{p\_value} < \mathbf{0.01}$

TABLE II
UGR'16 DATASET VARIANTS.

Label	Training	Type of flows	Anonymized flows
UGR'16v1	March to June	Unidirectional	No
UGR'16v2	March to May	Unidirectional	No
UGR'16v3	March to May	Bidirectional	Yes
UGR'16v4	March to May	Unidirectional	Yes

Feature engineering

Observation space

Feature Engineering → +0.1 AUC

ML Method → +0.01 AUC

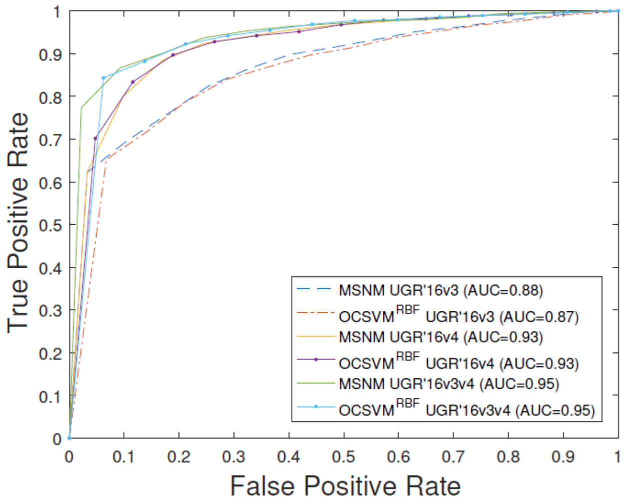
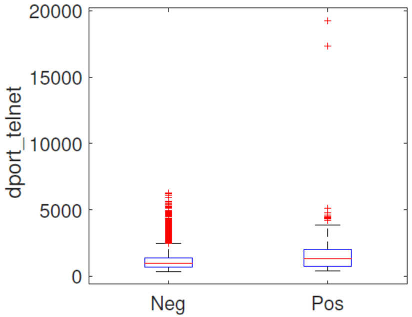


TABLE II
UGR'16 DATASET VARIANTS.

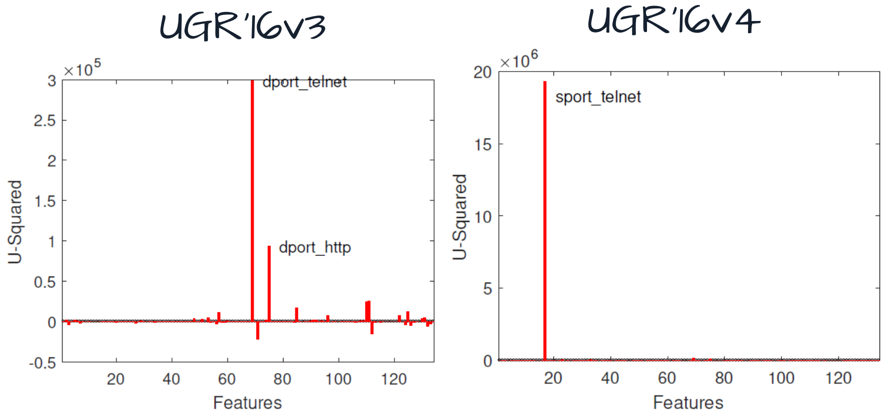
Label	Training	Type of flows	Anonymized flows
UGR'16v1	March to June	Unidirectional	No
UGR'16v2	March to May	Unidirectional	No
UGR'16v3	March to May	Bidirectional	Yes
UGR'16v4	March to May	Unidirectional	Yes

Feature engineering



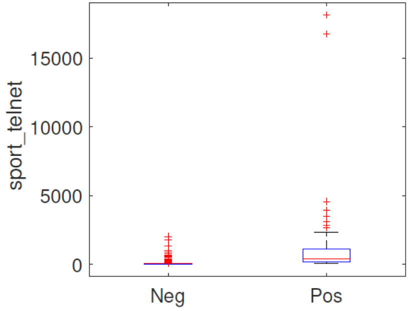
(a) Ttest **p_value** $<$ **0.01**

Feature space



(a) (b)

Figure 6. Comparison of U-Squared statistics for the DOS attack using as a reference UGR'16v3 (a) and UGR'16v4 (b).



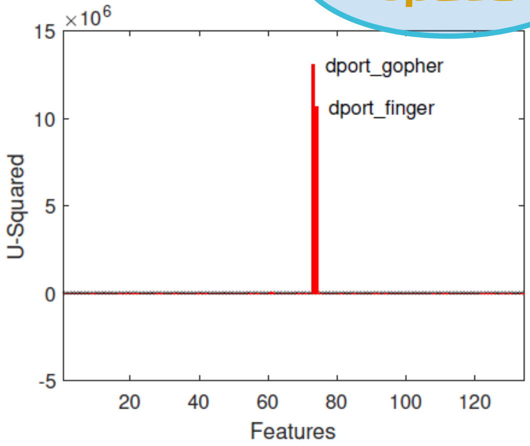
Ttest **p_value** $<$ **0.01**

TABLE IV
FEATURES OF THE CALIBRATION AND THE TEST SETS.

Feature	Calibration	Test
Capture start	10:52h 03/18/2016	13:43h 07/27/2016
Capture end	18:27h 06/26/2016	09:27h 08/29/2016
Attacks start	N/A	00:00h 07/28/2016
Attacks end	N/A	12:00h 08/09/2016
Number of files	17	6
Size (compressed)	181GB	55GB
# Connections	≈ 13,000M	≈ 3,900M

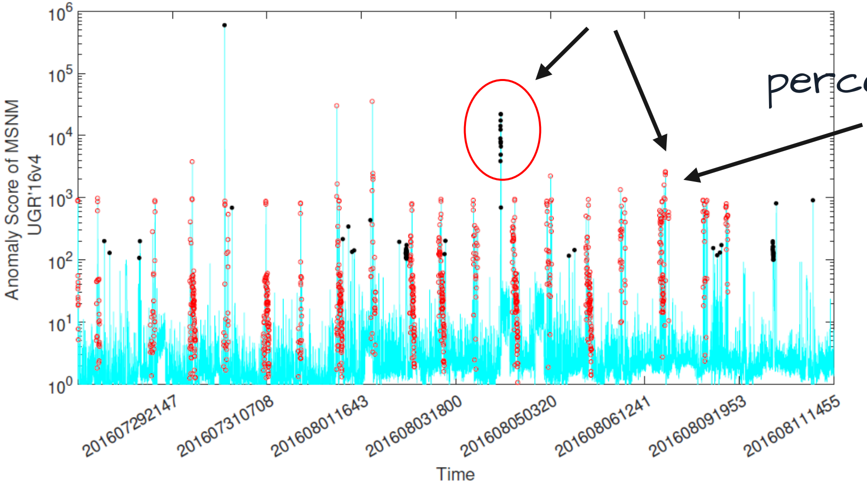
Test Labelling

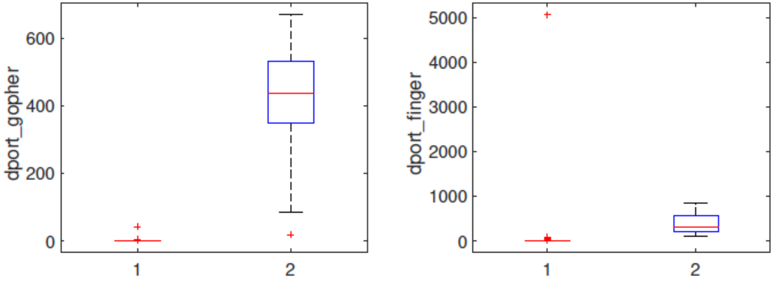
Feature space



real

perceived





(a) Ttest p_value < 0.01 (b) Ttest p_value < 0.01

"ML performance depends on all pipeline steps"

"Good understanding of the impact in performance"

"NMRG to reflect on the entire pipeline for representative case studies"

- TAKEAWAYs

MATERIALS

Some related materials you can freely use:

- The original **UGR'16**: https://nesg.ugr.es/nesg-ugr16/
- New datasets, tutorials & code: https://codas.ugr.es/animalicos/en/results
- A Data Analysis as a Service (DAaaS): https://codas.ugr.es/animalicos/en/daaas
- Video tutorials for the DAaaS: https://www.youtube.com/@TheCoDaSLab
- Software tools: https://github.com/CoDaSLab

ACKNOWLEDGEMENTS: This work was supported by the Agencia Estatal de Investigación in Spain, MCIN/AEI/ 10.13039/501100011033, grant No PID2020-113462RB-I00

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

THANKS

Prof. José Camacho
josecamacho@ugr.es
Computational Data Science Laboratory (CoDaS Lab)

Computational Data
Science Lab

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Stories**

Please keep this slide for attribution