Dataset and quality of data for AI

Jérôme François

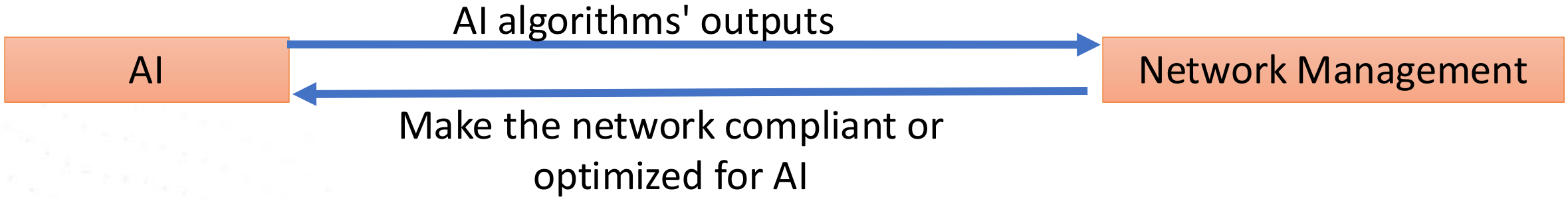
UNIVERSITY OF LUXEMBOURG

Research Challenges in Coupling Artificial Intelligence and Network Management

Jérôme François, Alexander Clemm, Dimitri Papadimitriou, Stenio Fernandes, Stefan Schneider

How was this document was built and structured?

- Coupling Network Management and AI



- The document is about challenges
  - Identify and classify challenges
  - No exhaustivity claimed
  - No solution provided (only examples for illustration purposes)
  - Current SotA overview and gap analysis

Sdsqd

"Data challenges" highlights

- Network data as input for ML algorithms
  - Data for AI-based NM solutions: data definition, mapping to a problem, representation, encoding, external parameters
  - Data collection: integrate AI-based solutions requirements into the collection process requirements
  - Usable data: lack of datasets for training and validation --> administrative, legal and ethical issues... and **quality problems**

Make AI more acceptable for networking

- Favor the most simple algorithms -> help in their explainability
  - Only use most sophisticated algorithms if there is a real added value

- Better to well tune a given algorithm rather than just select one after benchmarking multiple ones blindly
  - Starting with features also (much network data is neither images nor text…)

- Make the ML models robust and generalizable

  --> start with challenges related to data: we need high quality datasets

Always discussing that without significant progress

Data quality

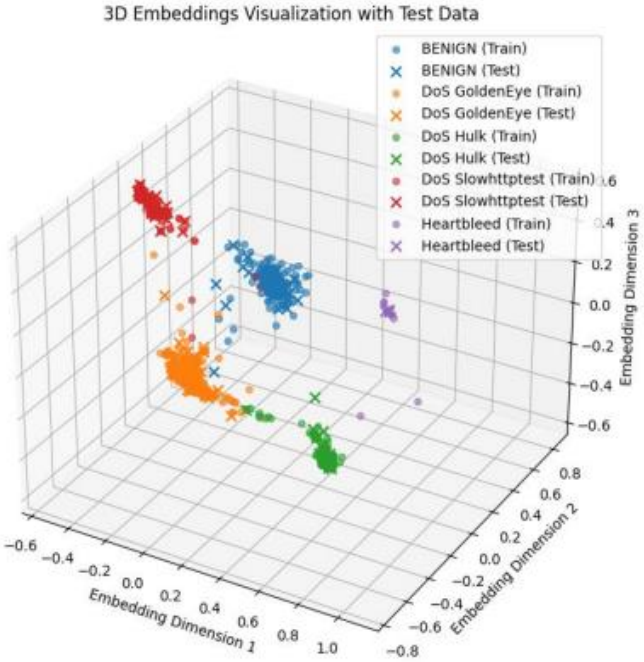
- Inherent problems
  - Hard to get access real data for privacy and legal reasons
  - Hard to label them properly
  - → (Small and) unrealistic datasets

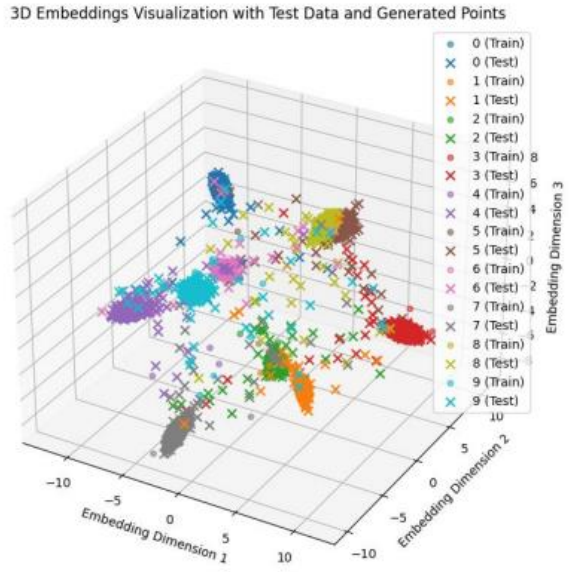
- But necessity to compare and reproduce results
  - Can we generate good data (with simulators or digital twins…)?
  - How good or realistic is a dataset?
- Regulations will also adds constraints (e.g. European AI act)

- Example: How bad is it for network intrusion?
  - CIC-IDS datasets have been extensively used since 2018
  - Easy to classify attacks in these datasets: Random Forest reaches 0.97 F1-score
    - Is it worth to get 0.99 with DL?
    - “Undetected” overfitting issue due to lack of data variability

Data quality measure

- Complexity of data, volume, neuron coverage...
- Feature-based, linearity, neighborhood measures (How Complex Is Your Classification Problem? A Survey on Measuring Classification Complexity. ACM Comput. Surv. 2020)



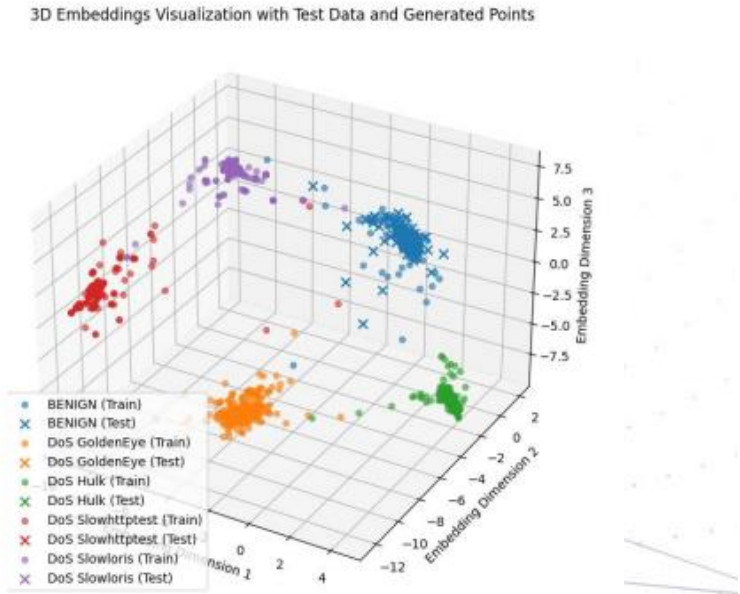


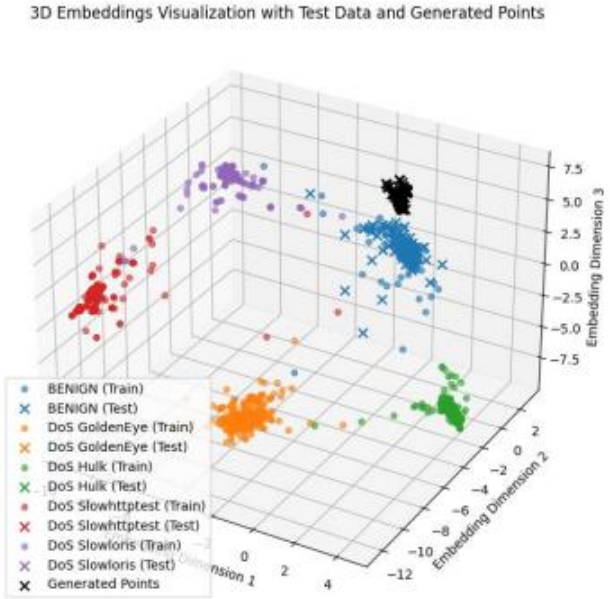
Measures	IDS	CIFAR-10
F1	0.0177	0.8029
F2	0.0000	0.8524
F3	0.0000	0.9998
F4	0.0078	0.2660

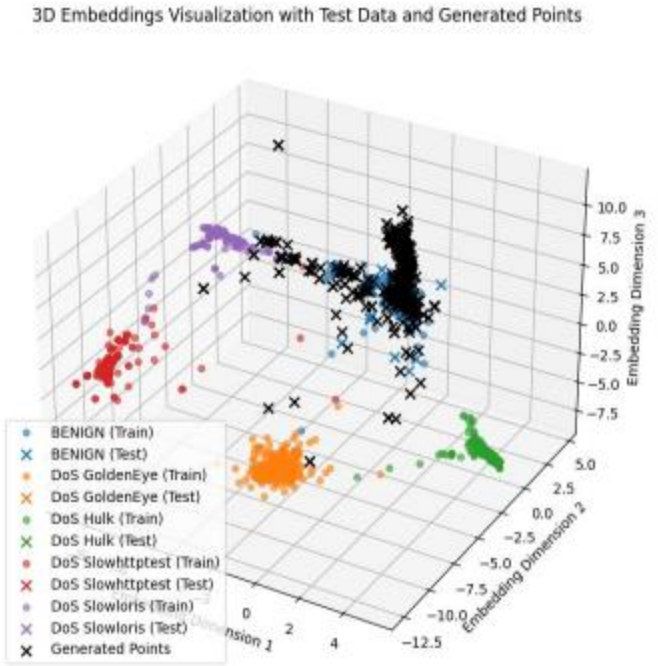
- Those are purely data-related metrics, ok for a classification problems like IDS but not easy to map to other more complex objectives (actions, configurations, etc.)
  - Network-specific quality metric based on the final objective?
  - Guide the collection/generation of new dataset through these metrics?

How to generate data

- Many AI methods exists (e.g. GNN, Evolutionary Algorithm)
- But these are data-centered methods –-> generated data cannot be realistic (e.g. even not compliant with protocol specification)
- Our approach (my research group): do not generate data directly, generate environment to generate data







Without tc

With tc

Research questions

- Assessment of the (re)usability and quality of data **according to a specific goal** ?
  - Derive existing metrics and methodology
  - Avalanche effect on degrading "raw" quality to the final objective
  - Characterize existing and future datasets in a systematic way

- Generation of data
  - For training + For testing --> different requirements
  - **Automation of data generation to be enough realistic** (again this depends on how data will be used)

- Relationship with NDTs to be fed with "high quality" data or to generate data...

- Discussion at IETF 121 meeting https://datatracker.ietf.org/meeting/121/session/nmrg
- To be continued with interested people