Open discussion on the potential standard dataset

Albert Cabellos
IETF Berlin, (proposed) NML WG
July 2016
Scope & Context

• In NML#3 (Athens) we discussed about the standardization of datasets for Network ML
• In this presentation:
  1. Bring attention and discuss the importance of datasets in NML
  2. Discuss the importance of datasets in other ML fields
  3. Trigger discussion about standardization of datasets for NML
How important are datasets?

<table>
<thead>
<tr>
<th>Year</th>
<th>Breakthroughs in AI</th>
<th>Datasets (First Available)</th>
<th>Algorithms (First Proposed)</th>
</tr>
</thead>
</table>

Average No. of Years to Breakthrough: 3 years

18 years

Table from: Datasets over Algorithms (SpaceMachine, March 2015)
http://www.spacemachine.net/views/2016/3/datasets-overalgorithms

How important are datasets?

Datasets might be the key limiting factor to the development of new AI techniques.

Table from: Datasets over Algorithms (SpaceMachine, March 2015)
http://www.spacemachine.net/views/2016/3/datasets-overalgorithms
Benefits of public datasets for Network ML

• ML-based algorithms do not provide guarantees (as opposed to traditional networking algorithms)
  – How can we make sure that our newly trained AI algorithm will work in different (untrained) scenarios?
• Provides a benchmark
  – Is the new algorithm better than the old one?
• Encourages research
• Allows for reproducible research
Datasets in AI and Networking

• Several AI fields have well-known public datasets, examples:
  – IMAGENET hosts 14M images for computer Vision. IMAGENET Challenge
  – Yahoo News Feed including 20M anonymized user-data

• The networking field has also a long tradition of public datasets, examples:
  – The CAIDA Anonymized Internet Traces 2012 Dataset
  – RIPE Atlas
  – CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth
Open discussion

• Should this WG promote public datasets for NML?
• How NML datasets are different from already existing network datasets?
• What are the privacy implications of such datasets?
  – Are there other associated risks?
• Can we help by developing a standard?
• If so, what are the relevant aspects of such standard?
  – Traffic features
  – Benchmark
  – Anonymization techniques