Datasets and Models for applying ML to networking/security tasks

Datasets:

* [Google Cluster Usage traces (GCUT)](https://drive.google.com/drive/folders/1rP7LPx8vhdp5cMI3ndDpAUaILlyb2vI_?usp=sharing)
* [Wikipedia Web Traffic (WWT)](https://docs.google.com/document/d/1L0VURpZMlMmoBwQwcfudPdHuMSAdECdMdY4f5HvZhQ8/edit?usp=sharing)
* [Measuring Broadband America (MBA)](https://www.fcc.gov/reports-research/reports/measuring-broadband-america/raw-data-measuring-broadband-america-seventh)
* [GWA-T-12 Bitbrains](http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains) (Trace Data):
  + The dataset contains the performance metrics of **1,750 VMs** from a distributed datacenter from **Bitbrains.**
  + request crediting them when publishing results.
* [WILDS.STANFORD.EDU](http://wilds.stanford.edu):
  + [7 Datasets](https://wilds.stanford.edu/datasets/):
  + Paper: [Wilds: A Benchmark of in-the-Wild Distribution Shifts](https://arxiv.org/pdf/2012.07421.pdf)
* [Parallel Workload Archive](https://www.cs.huji.ac.il/labs/parallel/workload/)

Network Downstream Tasks:

|  |  |  |
| --- | --- | --- |
| Name | Year | PDF |
| **Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices** | **CoNEXT 2008** | [**http://alumni.cs.ucr.edu/~dhiman/papers/conext\_08.pdf**](http://alumni.cs.ucr.edu/~dhiman/papers/conext_08.pdf) |
| Multi-classification approaches for classifying mobile app traffic | J. Netw. Comput. Appl. 18 | <https://dl.acm.org/doi/10.1016/j.jnca.2017.11.007> |
| A survey of methods for encrypted traffic classification and analysis | Int. J. Netw. Manag. 2015 | <https://dl.acm.org/doi/10.5555/2885221.2885227> |
| A Deep Learning Approach for IP Hijack Detection Based on ASN Embedding | NetAI 2020 | <https://dl.acm.org/doi/10.1145/3405671.3405814> |
| BGP hijacking classification | IEEE 2019 | <https://ieeexplore.ieee.org/abstract/document/8784511> |