USING DEEP NEURAL NETWORKS TO DETECT COMPROMISED HOSTS IN LARGE SCALE NETWORKS

ANGEL KODITUWAKKU

THE UNIVERSITY OF TENNESSEE, KNOXVILLE

EBONI THAMAVONG

BOOZ ALLEN HAMILTON

OUTLINE

- Objective: Improving analyst daily workflow and data understanding
- Analyst-Data Science Challenge
- Motivation
- Scope
- Introduction to Deep Learning
- Dataset creation
- Technical methodology
- Model training and tuning
- Results
- Case studies

THE ANALYST-DATA SCIENCE CHALLENGE

- Understanding the ”why”?
- Asking the right questions
- Iteration of ideas and solutions
- Model generation and validation
- Use the model to prioritize the actionable events
- Start analyst workflow

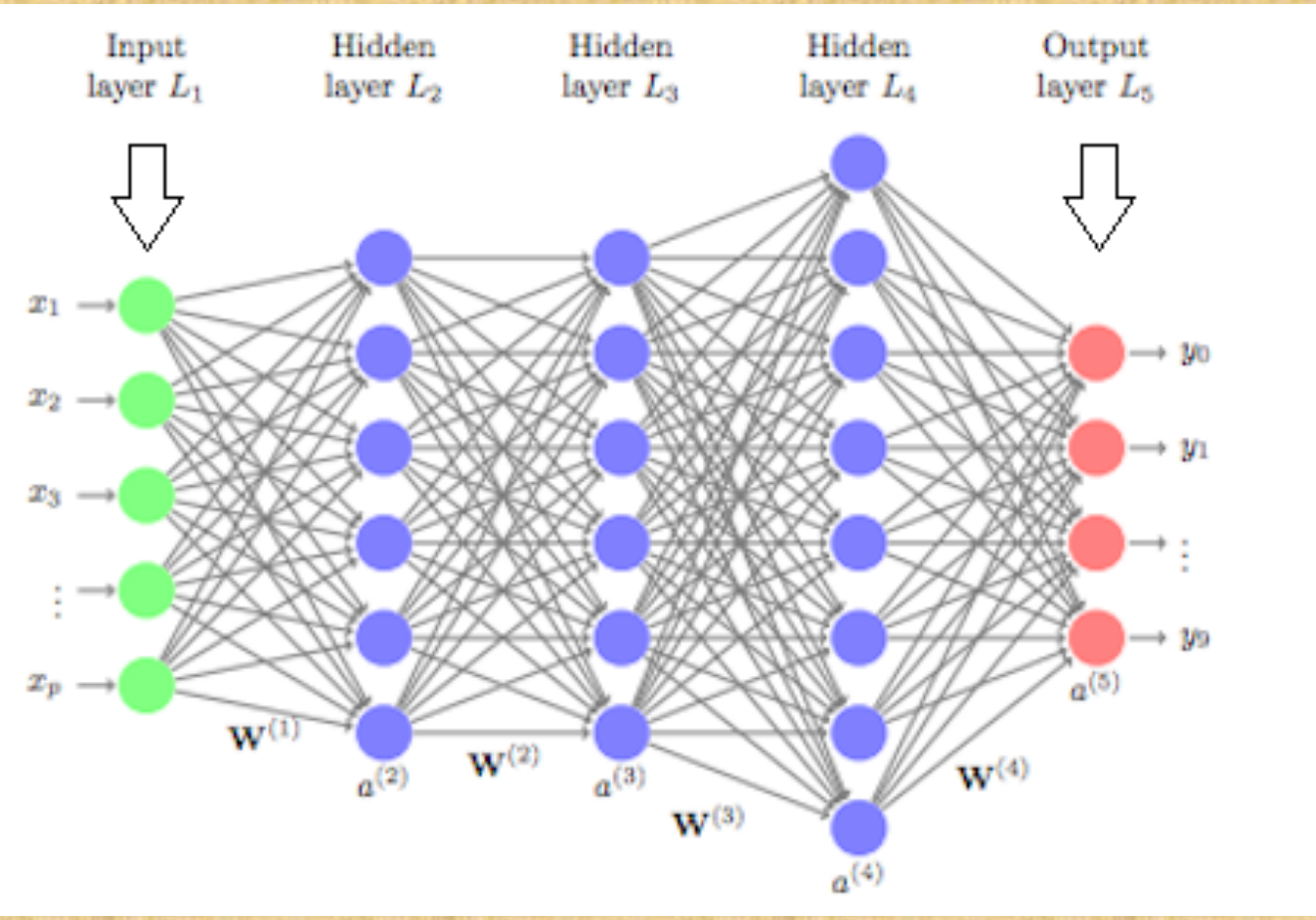
MOTIVATION

- Issues with current datasets.
  - Outdated.
  - Simulated.
  - Does not reflect current threats.
  - Can’t collect the same features from real networks.

- Issues with current statistical models
  - Unusable in practice
  - Static models can’t be used for streaming data

DEEP LEARNING

- Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on artificial neural networks.



- Learn from training data.
- Find progressively higher-order patterns in input data.
- Requires
  - Need quality training data.
  - Tuning of hyperparameters.
  - Testing.

IMAGE SOURCE: https://orbograph.com/deep-learning-how-will-it-change-healthcare/

DATA SOURCES

- Flow data from GLORIAD research & education network
    - Global Ring Network for Advanced Applications Development
    - National Science Foundation sponsored project 2006-2015
    - Closest to real-world global scale flow data
- Maxmind Geolocation data
- Emerging Threats Compromised IPs blacklist
- Global Science Registry (maps IP blocks to institutions)

SCOPE OF RESEARCH

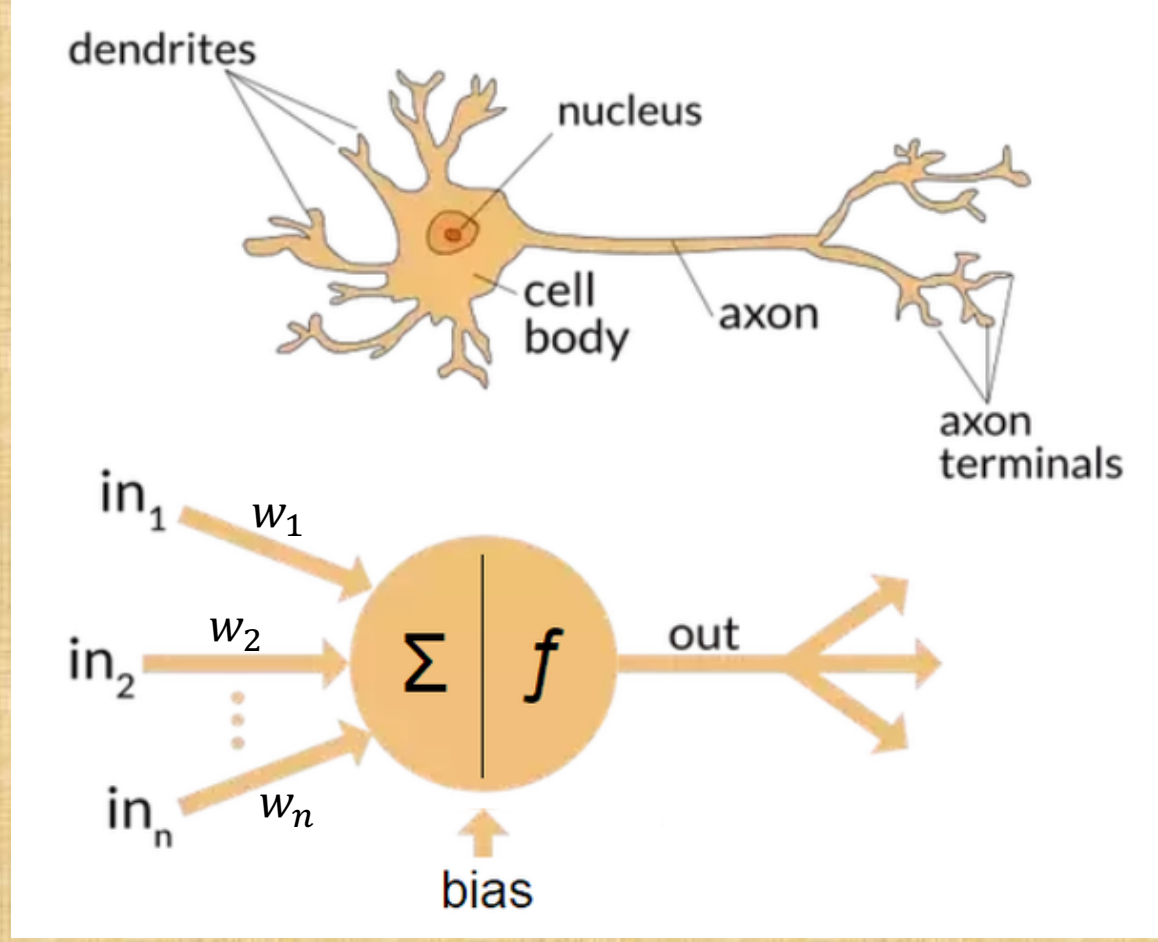
- Why Emerging Threats' Known Compromised Hosts Threat-list ?
  - Can use to enrich streaming flow data on-the-fly
  - Signatures that we can train a statistical model to detect
  - Freely available
  - Actively kept up-to-date

- What is the Known Compromised Hosts Threat-list?

"Hosts that are known to be compromised by bots, phishing sites, etc, or known to be spewing hostile traffic. These are not your everyday infected and sending a bit of spam hosts, these are significantly infected and hostile hosts."

METHODOLOGY

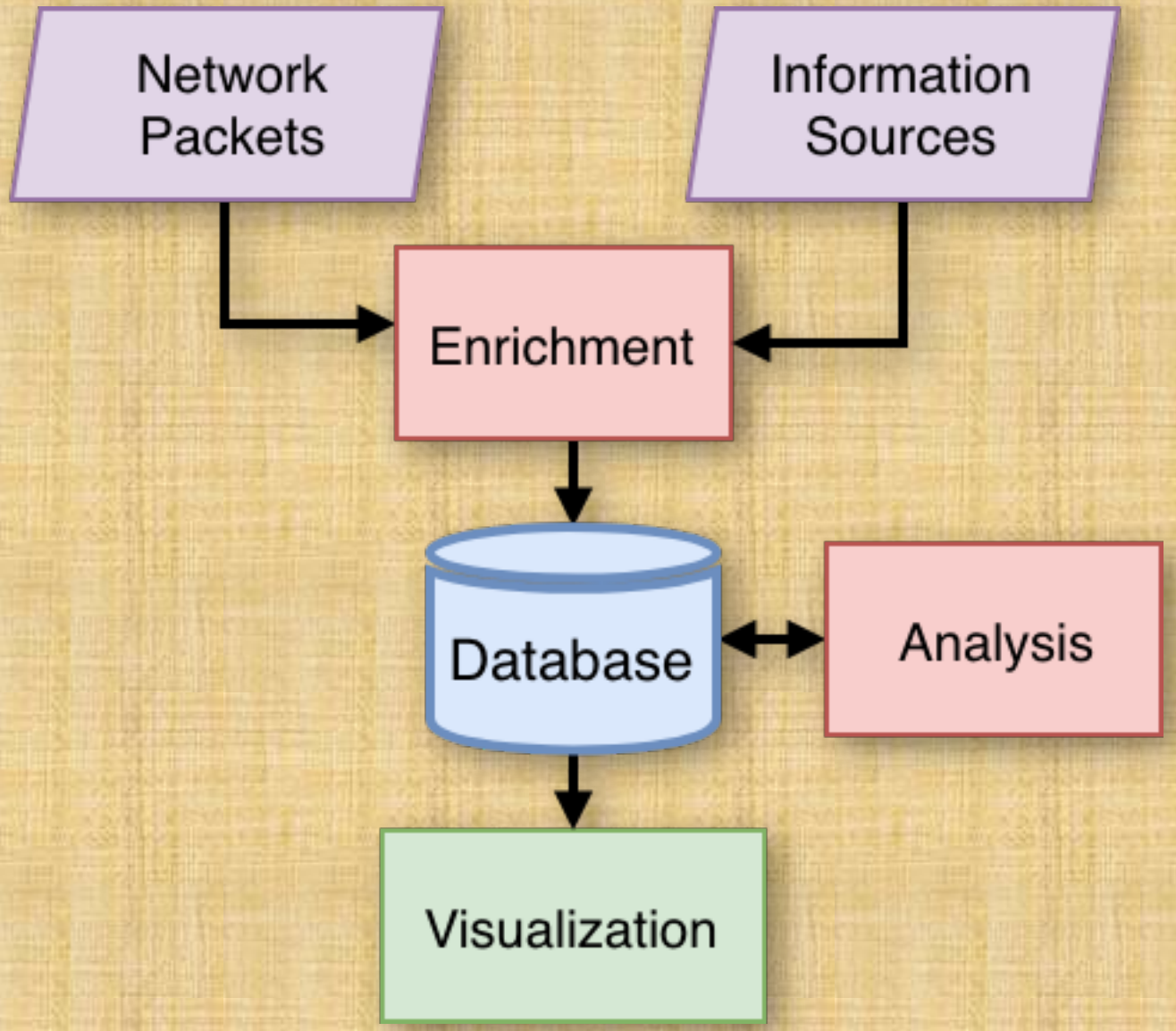
Objective: Find the best weights and the bias for each neuron.



$$Out = \sum_{i=1}^{n} in_i w_i + bias$$

- Use the supervised dataset to train the model.
- Tune the hyper-parameters for optimum accuracy.
- Implement the model as a plug-in module for InSight2 platform.

DATASET CREATION

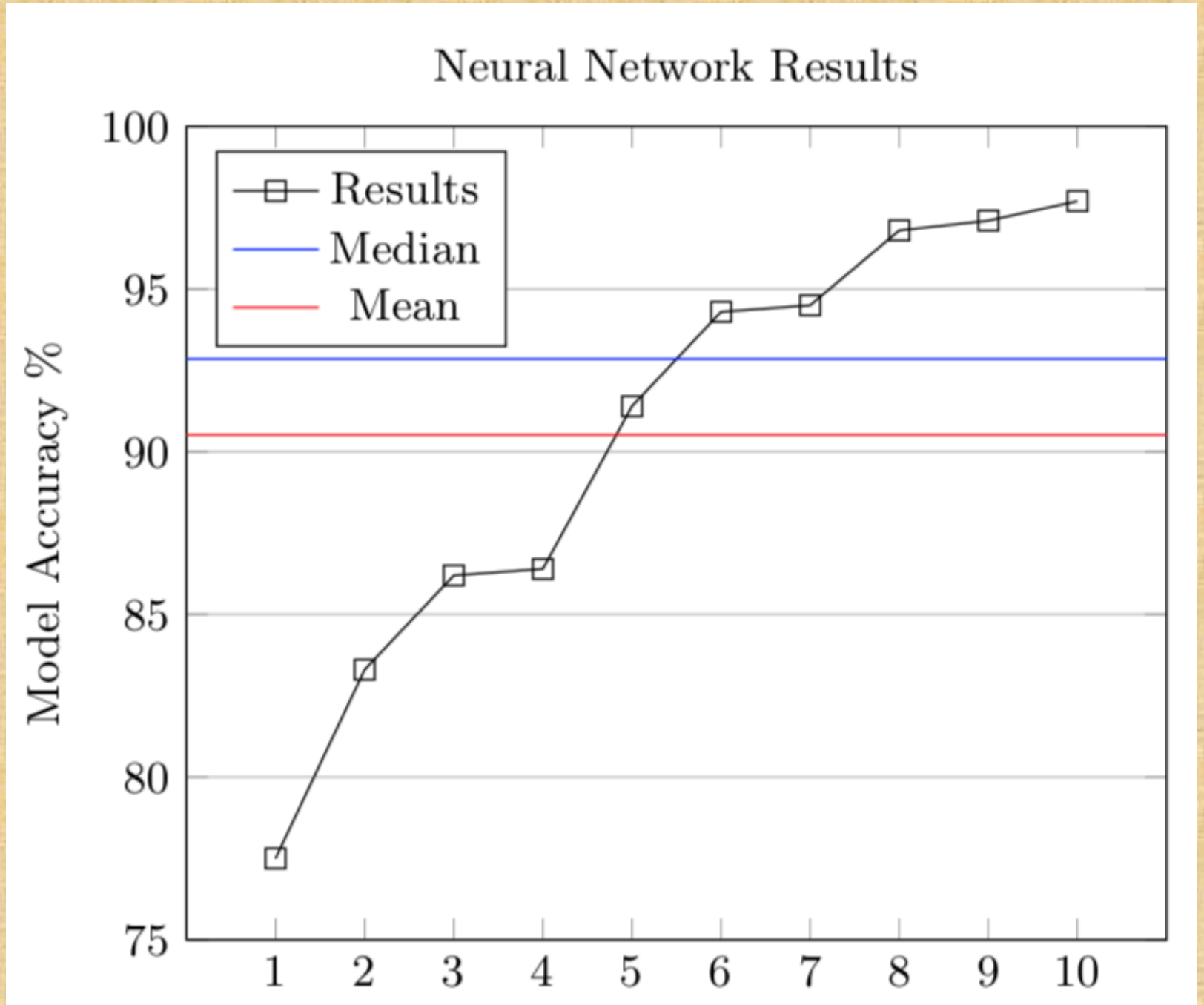


- Collect flow data.
- Enrich the data with,
  - ET's compromised hosts
  - Geolocation
  - GSR
- Extract the supervised dataset

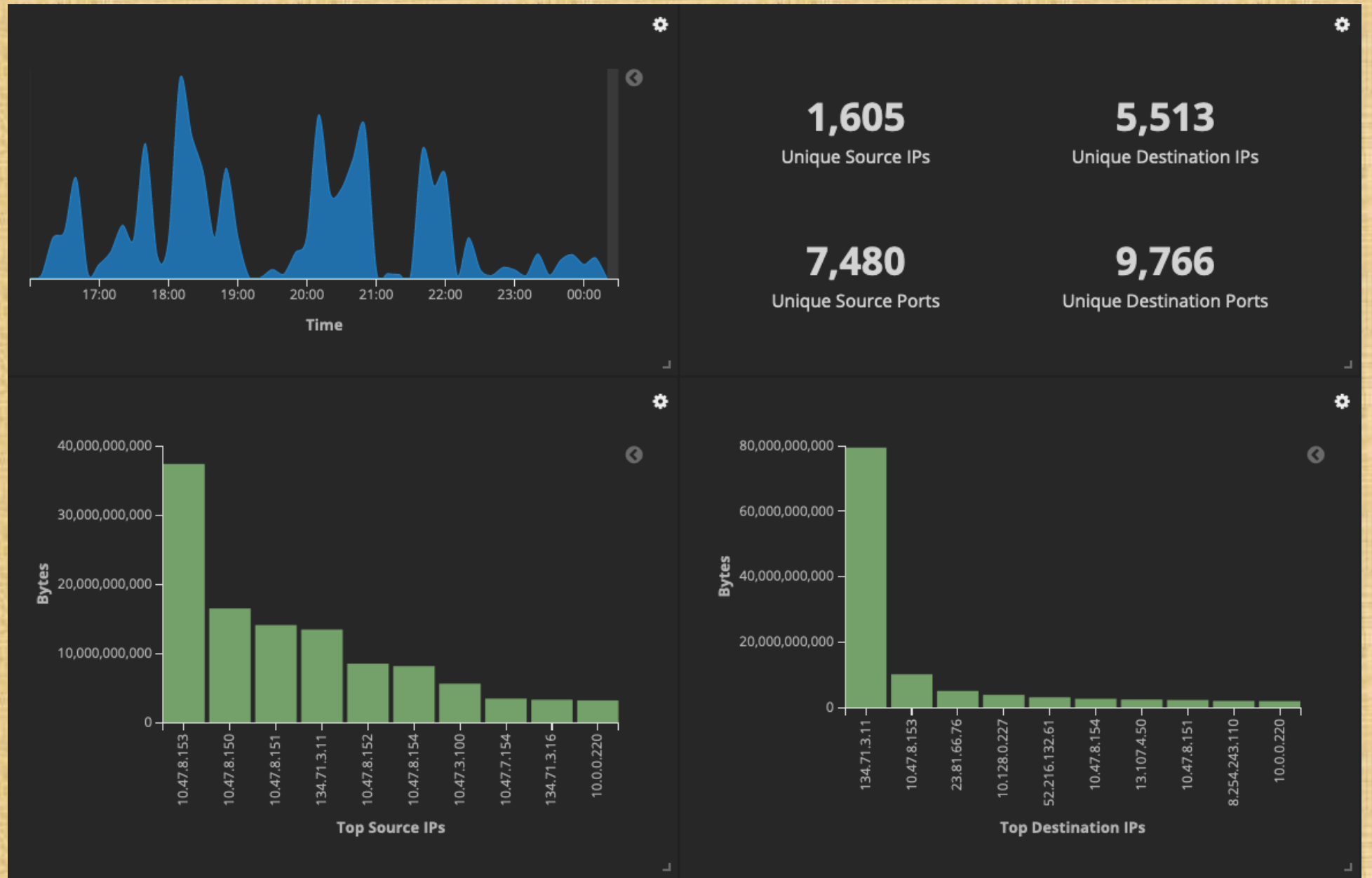
HYPER-PARAMETERS

1. Train/test: 75/25
2. Shuffle: on
3. Learning rate: 0.1
4. Classes: 2
5. Hidden units 40,30,20,10,5
6. One hot encoding
7. Batch size: 100
8. Train steps: 1000

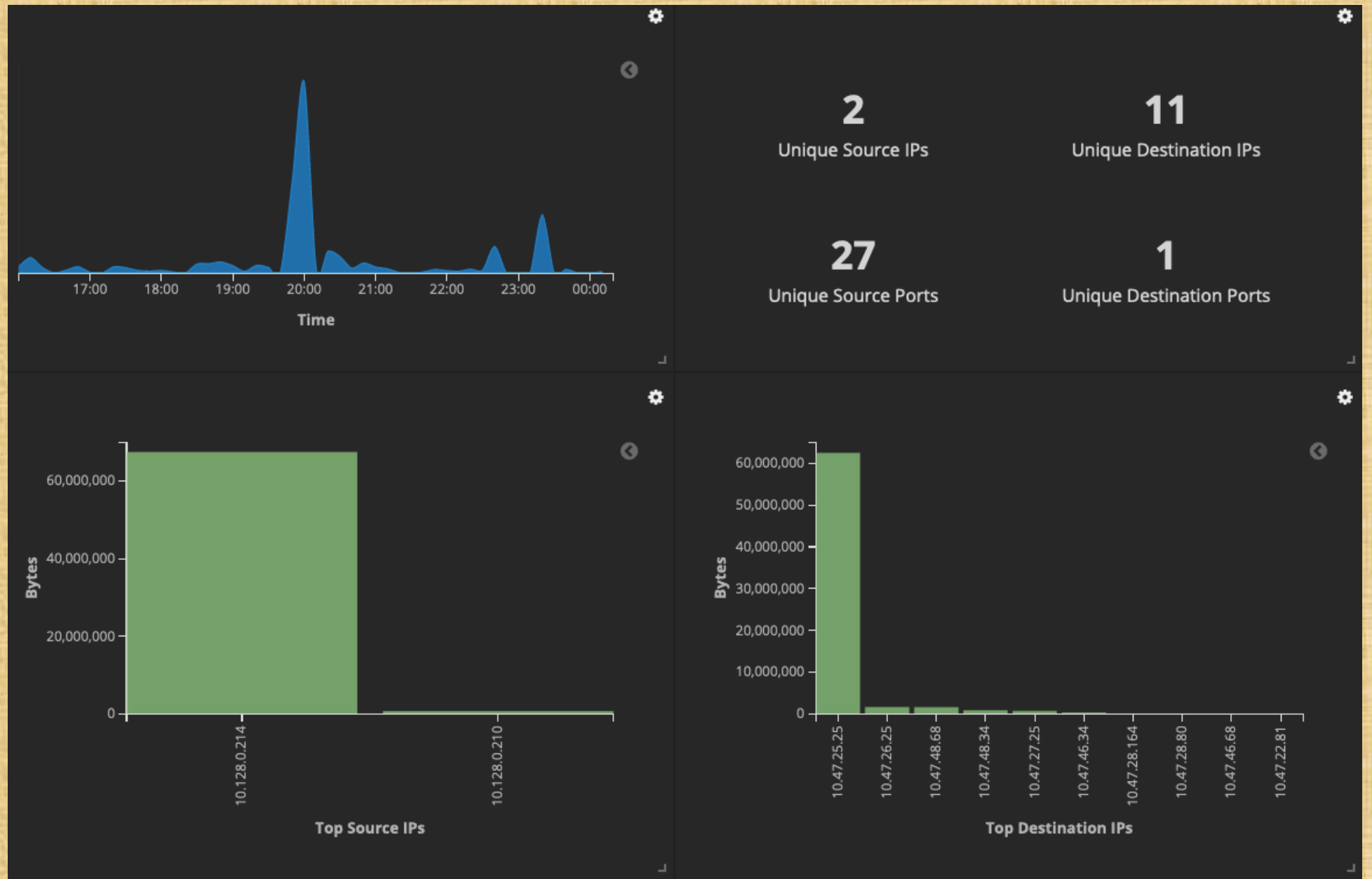
RESULTS


Neural Network Results
Results
Median
Mean
Model Accuracy %
100
95
90
85
80
75
1
2
3
4
5
6
7
8
9
10

CASE-STUDIES (WRCCDC)



CASE-STUDIES (WRCCDC)



Q&A

- Contact
  - Angel Kodituwakku, The University of Tennessee, angelk@utk.edu
  - Eboni Thamavong, Booz Allen Hamilton