FINDING NEEDLES IN A NEEDLE FACTORY

Dr. Nur Zincir-Heywood
Computer Science, Dalhousie University
Halifax, NS, Canada

NETWORK INFORMATION MANAGEMENT AND SECURITY (NIMS) LAB

Systems that can Adapt

Identify Different Behaviours

Network / Application Data

Security / Fault

http://www.wiringthebrain.com/2010/10/searching-for-needle-in-needle-stack.html

BEHAVIOUR IDENTIFICATION

- http://www.wiringthebrain.com/2010/10/searching-for-needle-in-needle-stack.html
- http://www.freewtc.com/images/products/needles_pins_2_90270.jpg

FINDING NEEDLES

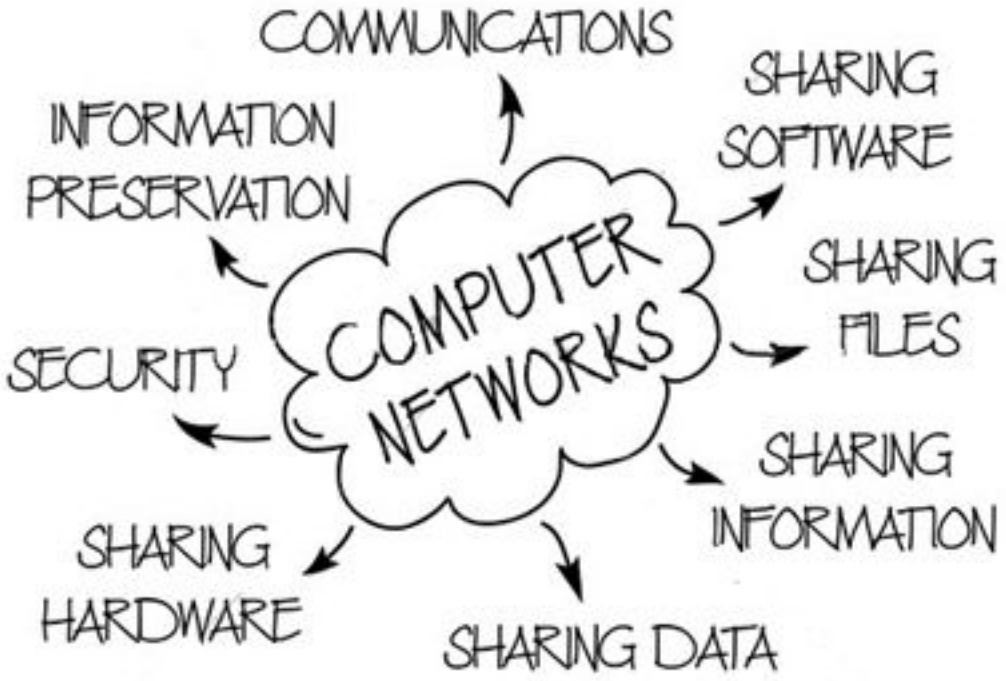
http://compscimiller.com/tag/computer-networks/

CHALLENGES

Superposition of behaviours

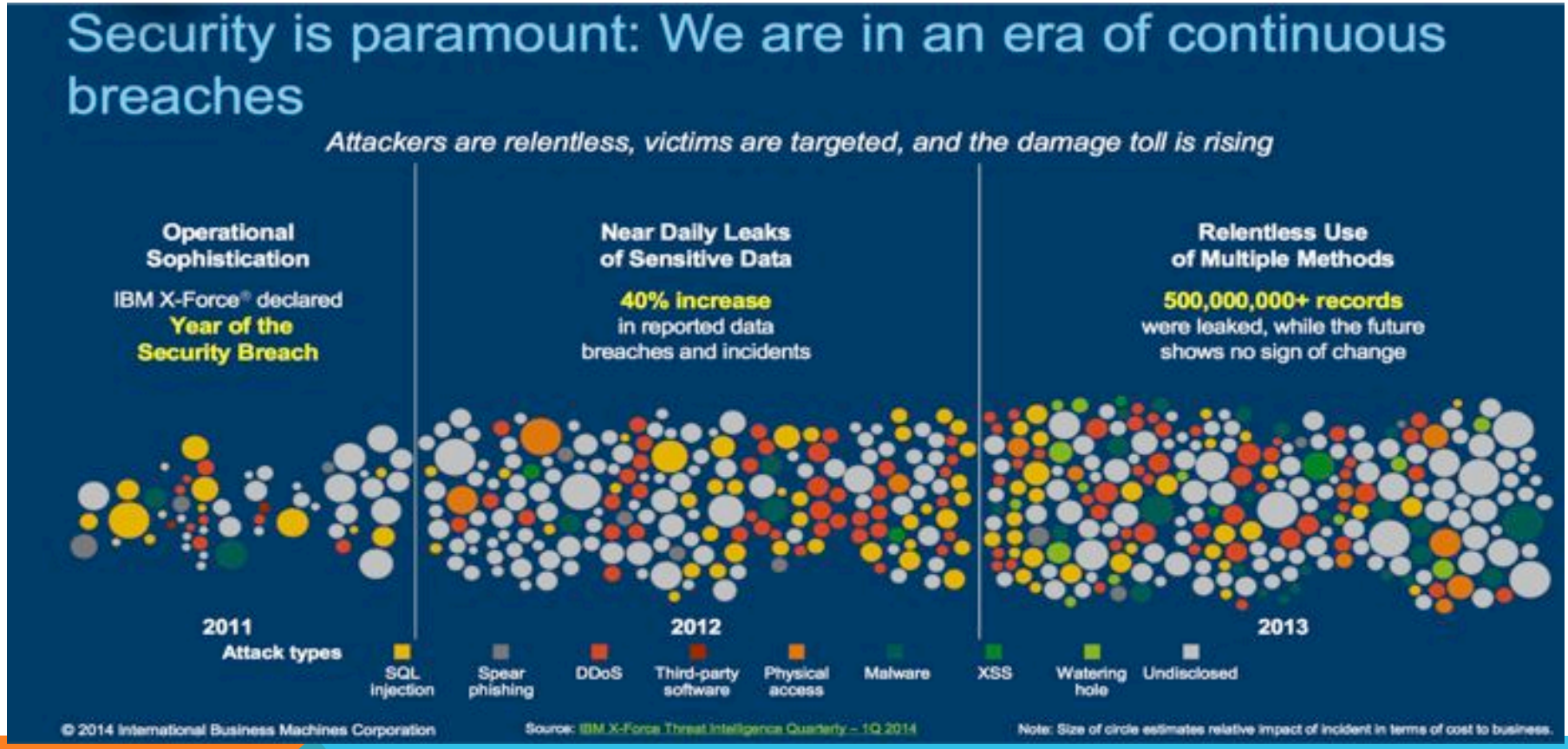
Mix of stationary and non-stationary behaviours

RAPID GROWTH…



http://cikgusuruhbuat.blogspot.ca/2015/07/list-benefit-of-computer-network.html

MISSION CRITICAL INFORMATION SYSTEMS



https://practicalanalytics.files.wordpress.com/2014/02/securityanalytics2.png

CYBER-SECURITY

https://www.symantec.com/security-center/threat-report

E-SERVICES

http://www.all-internet-security.com/wp-content/uploads/2016/09/secure-acess.png

WELL ☺

Good potential for data driven approaches

Need to be careful, though!

- A learning system to recognize tanks
- What did it actually learned?

However, can we think of everything?!

https://www.google.ca/search?q=tanks&rlz=1C5CHFA_enCA711CA711&espv=2&biw=1364&bih=697&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiP8v_-5IfQAhWe0YMKHY__DJQQ_AUIBigB

Data Analytics: Machine Learning Based Approaches

- Techniques
  - Supervised
  - Unsupervised
  - Semi-supervised
- Tools
  - Open source (Weka, Moa, R, …)
  - Commercial (Matlab, IBM, Oracle, Amazon…)

SO WHERE DO WE START?!

How to represent data?

How to sample data?

How to represent objectives to analyze data?

How to measure performance?

How to incorporate visualization?

How much prior knowledge?

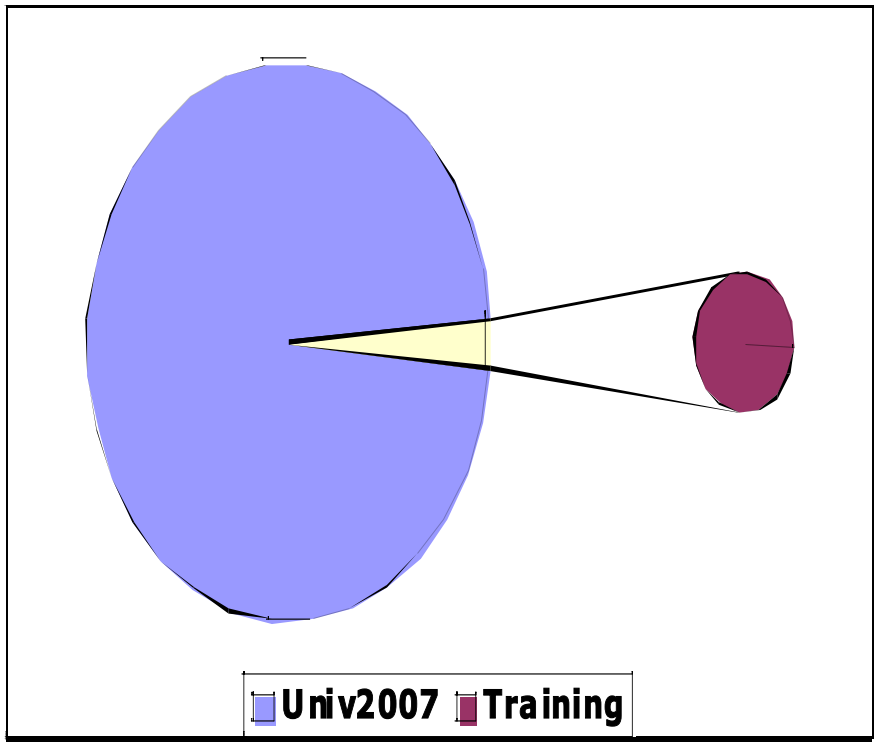
DATA FEATURES

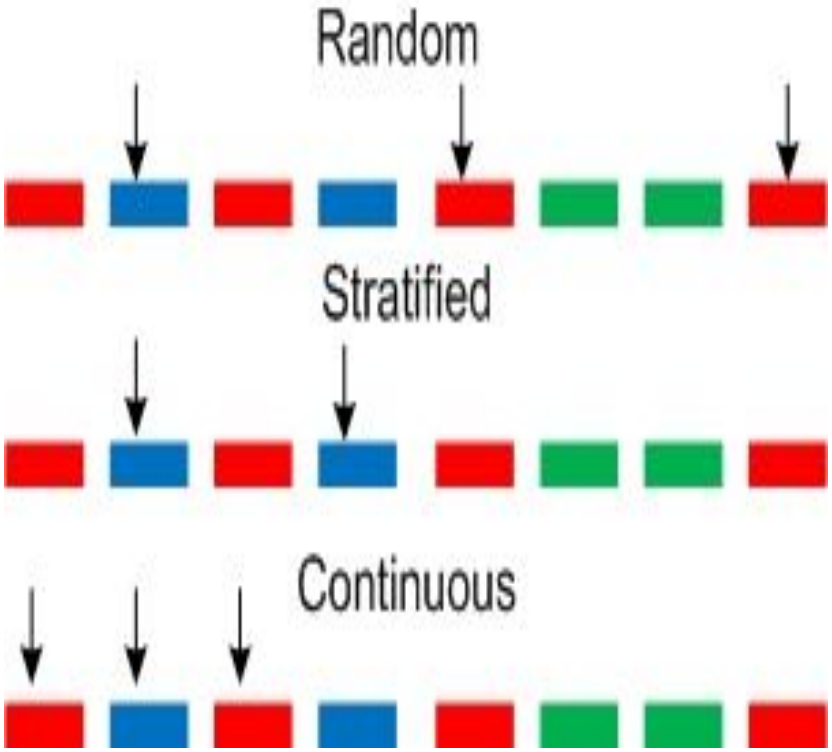
AccelProbe
CellProbe
LightProbe
MagneticProbe
NoiseProbe
RotationProbe
RunningApplications Probe
SystemProbe
WiProbe

SMS and e-mail
Phone Web Apps
Contact Contact URL Name
Date/time Date/ time Date/time Date/time
Duration Duration
Duration Duration
Launches Outgoing or ingoing Referring URL
Sent or received
Word count

IP addresses
Port numbers
Direction
Protocol
Number of packets
Number of bytes
Duration
Inter-arrival time
Other stats

DATA SAMPLING

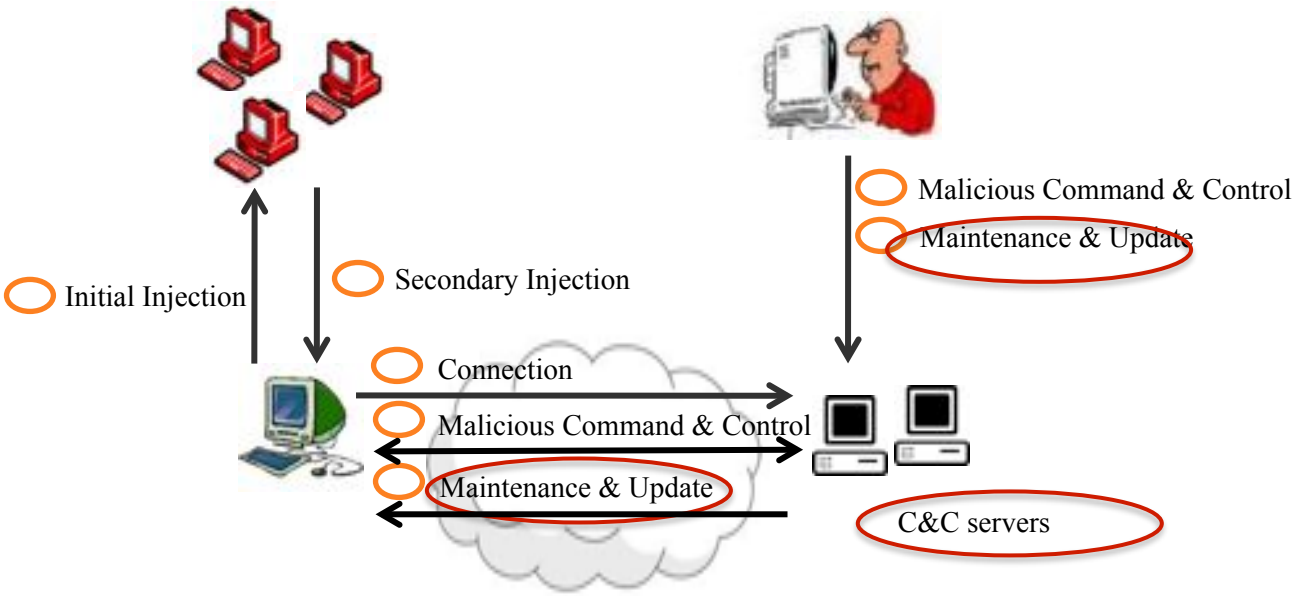




"How Robust Can a Machine Learning Approach Be for Classifying Encrypted VoIP?", R Alshammari, AN Zincir-Heywood, Journal of Network and Systems Management 23 (4), 830-869, 2015

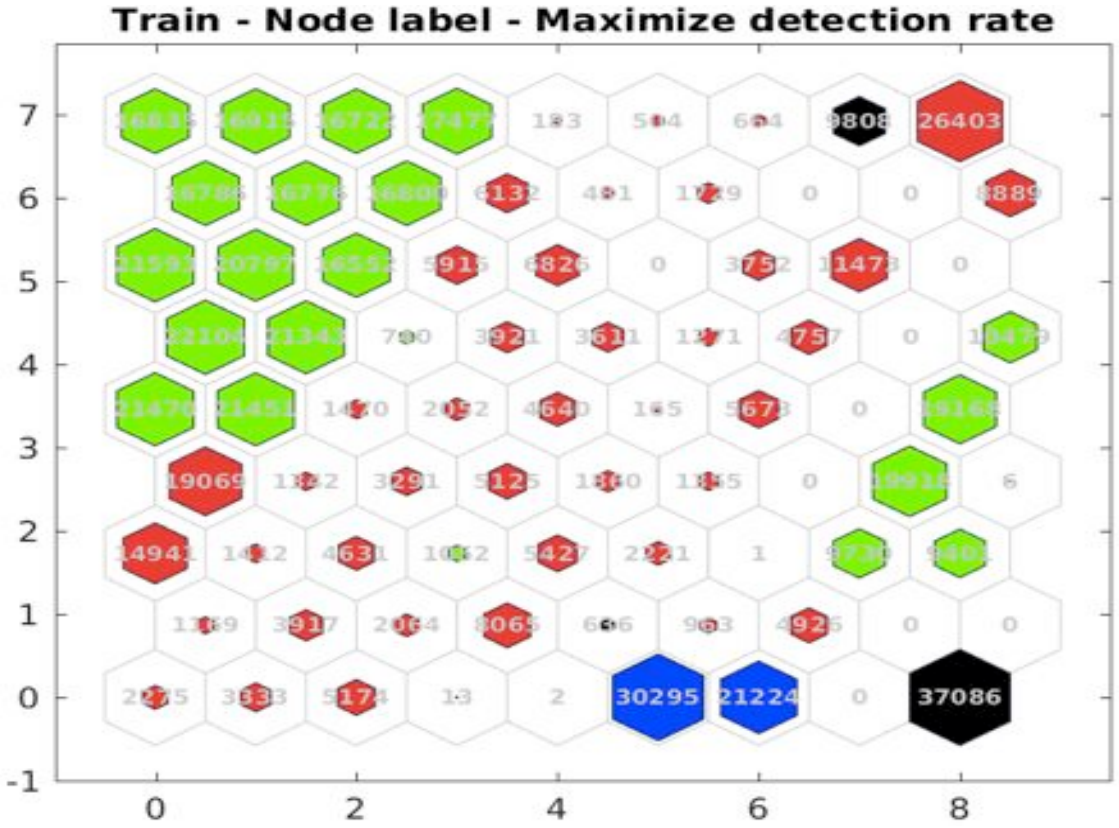
BOTNET

Botnet is a set of compromised hosts (aka bots) that are under the remote control obotmaster



"Benchmarking the Effect of Flow Exporters and Protocol Filters on Botnet Traffic Classification", F Haddadi, AN Zincir-Heywood, IEEE Systems Journal, 2014

SELF ORGANIZING MAP – CTU/CAPTURE9



4 colours:

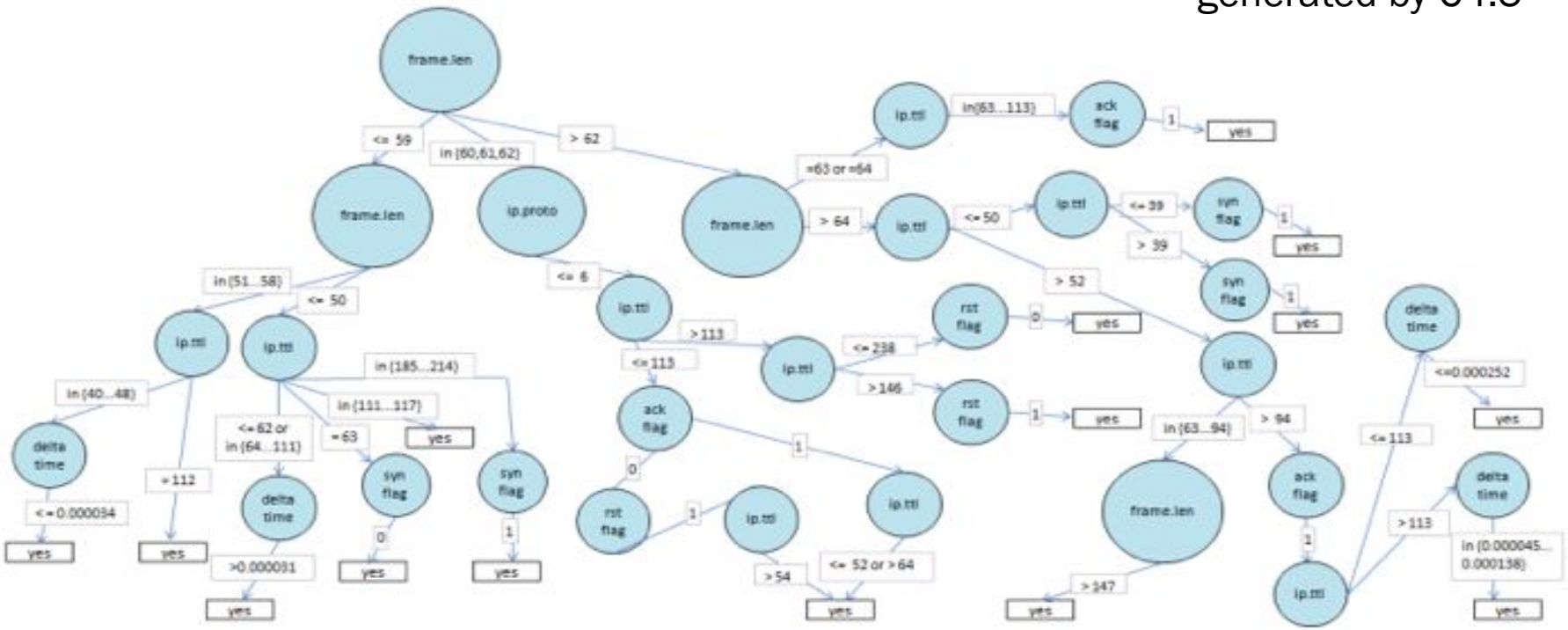
- red - background,
- green - normal,
- blue - botnet C&C,
- black - botnet

- "Data Analytics on Network Traffic Flows for Botnet Behaviour Detection", DC Le, AN Zincir-Heywood, MI Heywood,IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2016
- "A Hierarchical SOM-based Intrusion Detection System", HG Kayacik, AN Zincir-Heywood, MI Heywood, Engineering Applications of Artificial Intelligence, Elsevier Journal, 2007

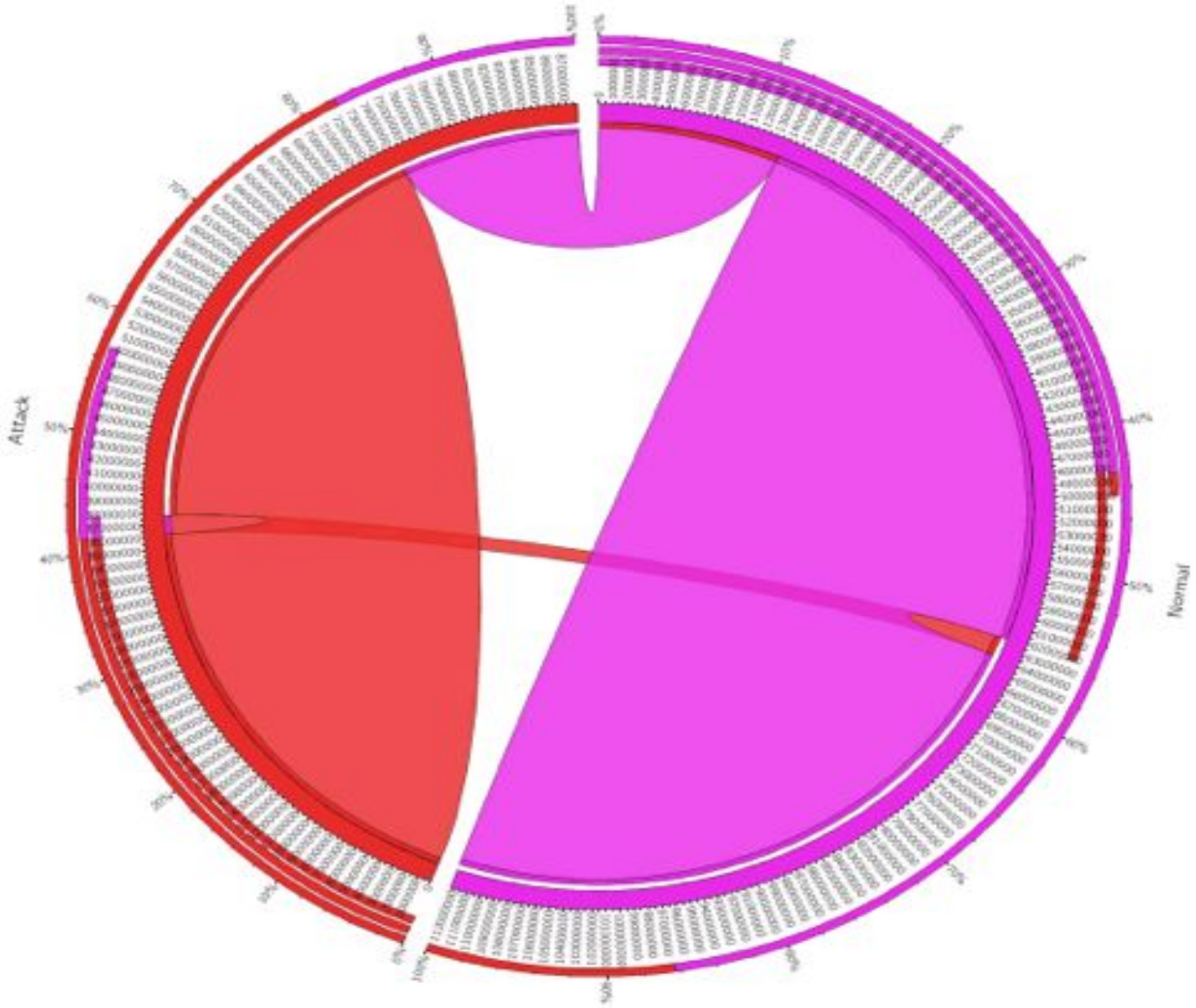
INVESTIGATING THE ROBUSTNESS CAIDA/DARKNET

Root: frame.len

Decision Tree
generated by C4.5



"Feature selection for robust backscatter DDoS detection", E Balkanli, AN Zincir-Heywood, MI Heywood, IEEE Local Computer Networks Conference Workshop on Network Measurements, pp. 611 – 618, 2015.



Circos by Decision Tree

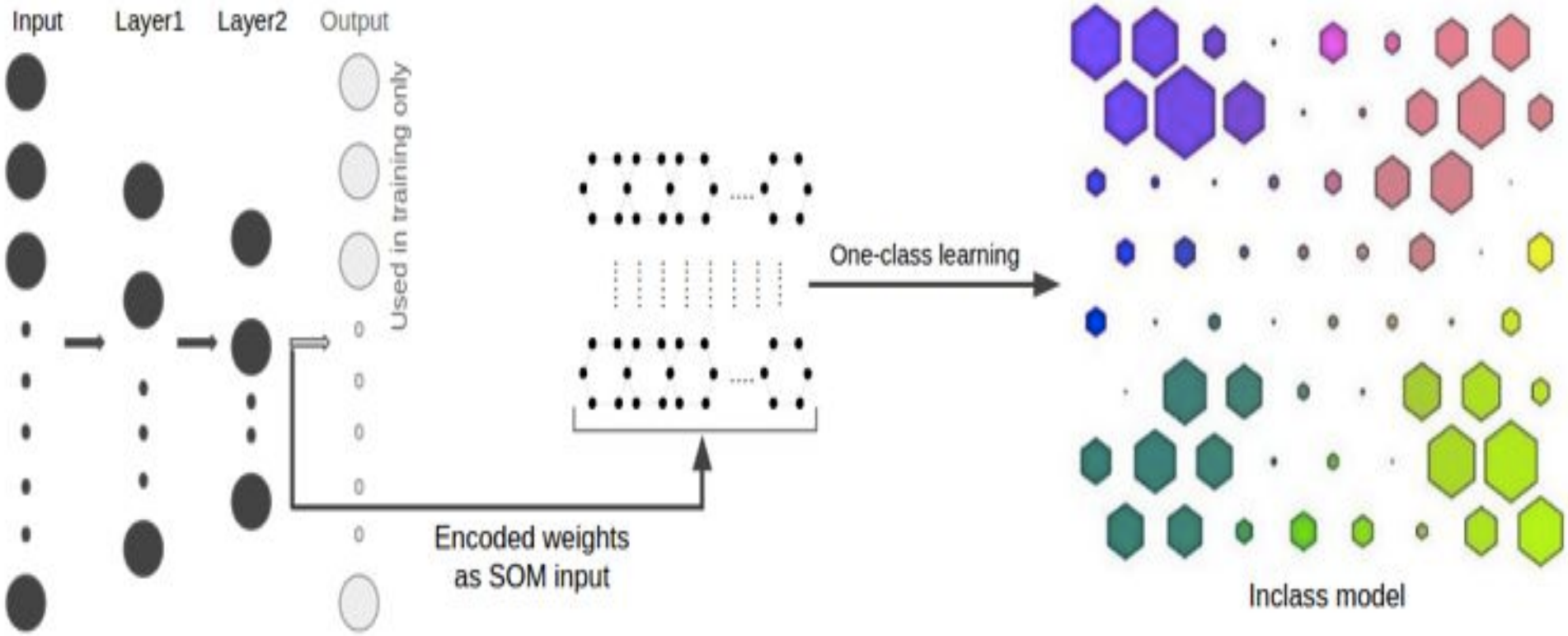
Red: attack
Purple: non-attack

Line purple to red represents false positives (10%)

Line red to purple represents false negatives (1%)

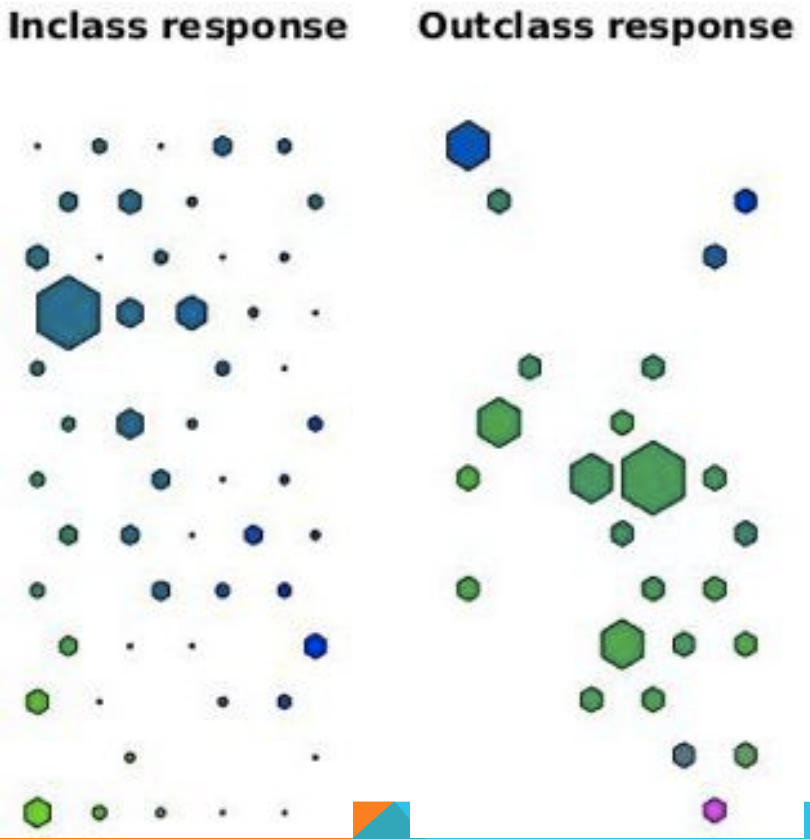
"Feature selection for robust backscatter DDoS detection", E Balkanli, AN Zincir-Heywood, MI Heywood, IEEE Local Computer Networks Conference Workshop on Network Measurements, pp. 611 – 618, 2015.

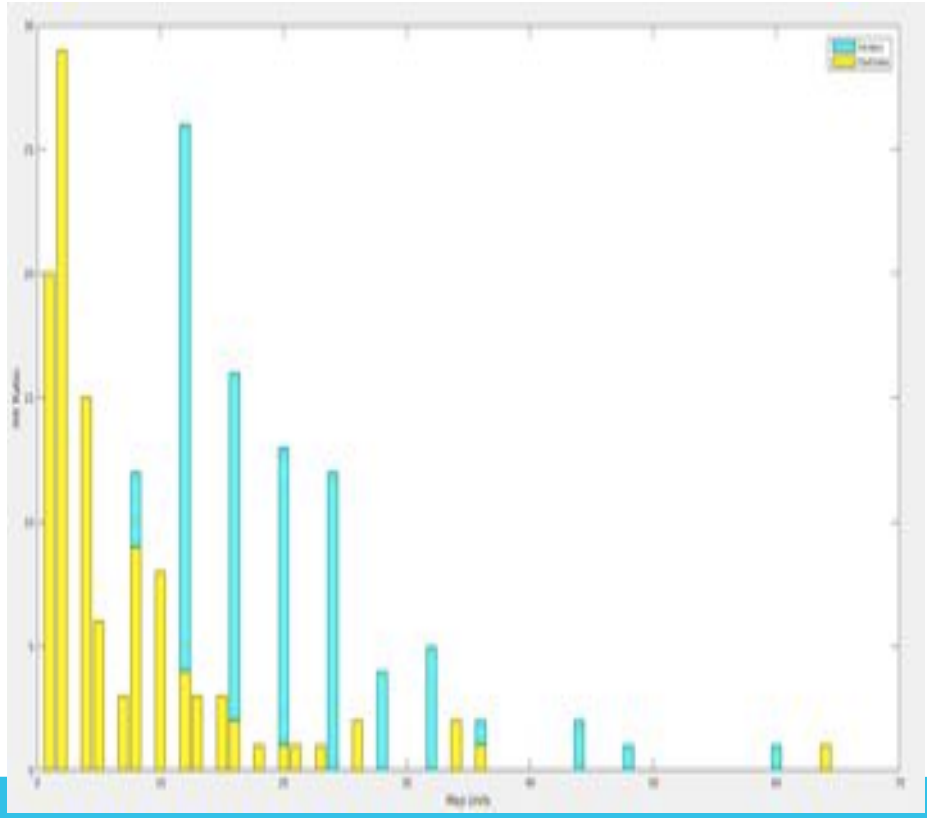
AUTOENCODER BASED SELF ORGANIZING MAP APPROACH



"Smart Phone User Behaviour Characterization based on Autoencoders and Self Organizing Maps", D Rajashekar, AN Zincir-Heywood, MI Heywood, IEEE International Conference on Data Mining Workshop on Data mining for Cyber Security, 2016

SOM MODELLING OF USER BEHAVIOUR





"Smart Phone User Behaviour Characterization based on Autoencoders and Self Organizing Maps", D Rajashekar, AN Zincir-Heywood, MI Heywood, IEEE International Conference on Data Mining Workshop on Data mining for Cyber Security, 2016

HOW MUCH PRIOR KNOWLEDGE?

Data and Objectives

- Constraints search space
- Blind side

What is the cost of providing labels?

What is the deployment environment?

Location, Time, Evasion

WHAT DID WE LEARN?

Data driven

- New insight and knowledge

Input – representation

- Packet / Flow / usage

Generalization

- Time & Location & Evasion

Output – objectives

- Can say “I don’t know”
- Value of certainty

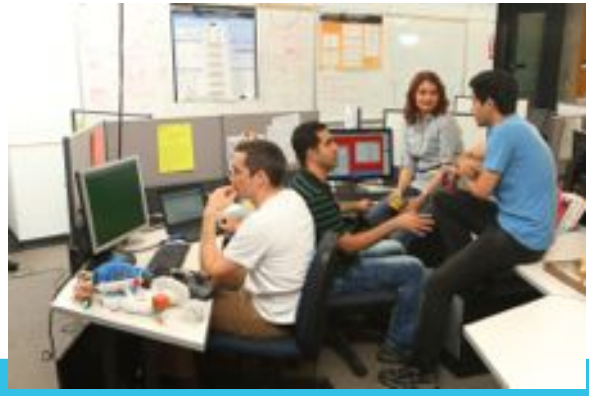
SO WHAT?!

Ever changing cycle

Nothing stays the same

Alert to measuring change!

THANK YOU!
QUESTIONS?



Dalhousie NIMS Lab-- https://projects.cs.dal.ca/projectx/

www.cs.dal.ca/~zincir

zincir@cs.dal.ca

NIMS