# New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Datasets

Dr Nour Moustafa

# Dr Nour Moustafa

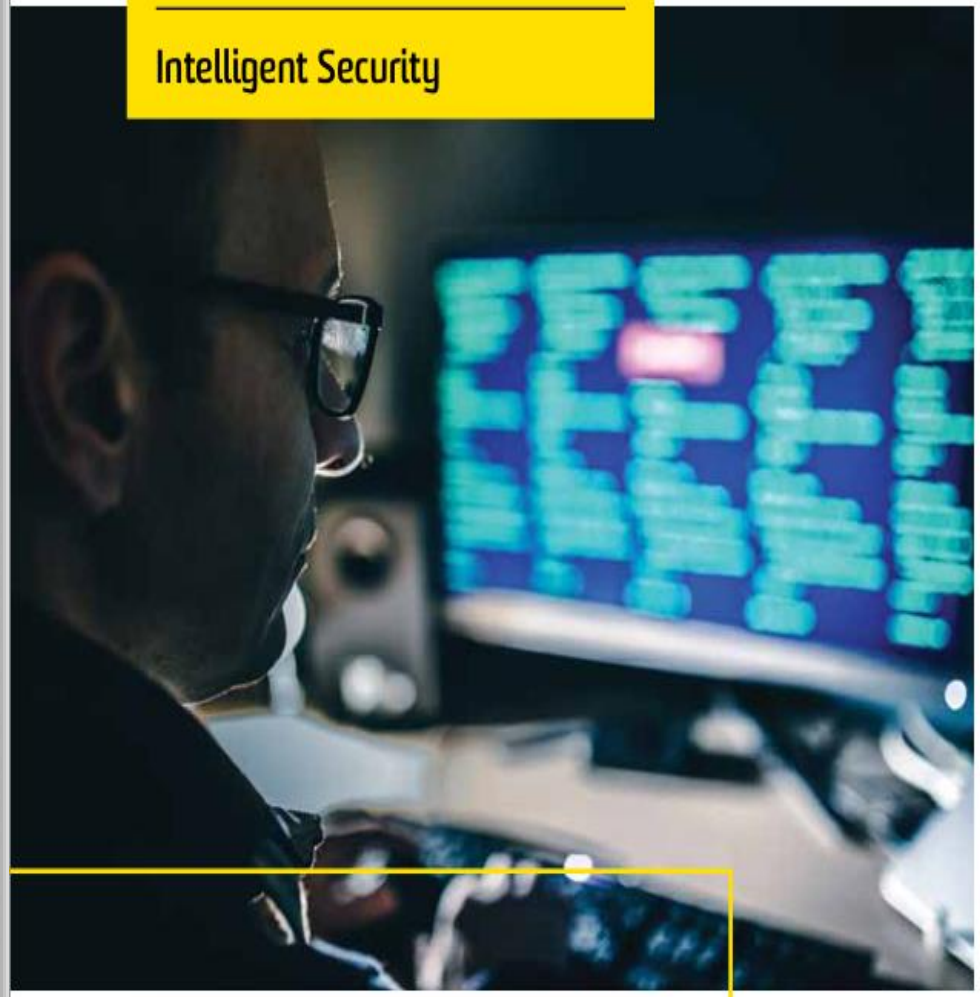
Lecturer, Theme Lead of Offensive Security & Postgraduate Discipline Coordinator (Cyber) at School of Engineering and Information Technology, UNSW Canberra Cyber, the Australian Defence Force Academy (ADFA)

Associate Editor at IEEE Access, Future Internet, BIGDATA EAI International Conference & Reviewer of many high-tier journals and conferences such as IEEE Transactions on Information Forensics and Security (TIFS) , IEEE Communications Magazine (ICM), Future Generation Computer Systems (FGCS) & Network and Computer Applications (JNCA)

My Research interests include developing intrusion detection, threat intelligence, Privacy Preservation, digital forensics, big data collections and analytics in IoT and IIoT networks using Statistical Learning, Machine and Deep Learning algorithms

Have several grants between 2018 and 2019 for solving real-world problems related to Cyber Security applications and Industry 4.0 systems: the Australian Cyber Security Collaborative Research Centre (CSCRC), Australian Federal Police, Australian Army, Systems Capability Centre (ADFA), UNSW Canberra, and Australian Research Data Commons (ARDC)

## Intelligent Security




More information

**Dr Nour Moustafa**

School of Engineering and Information Technology

**T:** +61 (0) 416 817 811 | **E:** nour.moustafa@unsw.edu.au


**Development of intelligent methods—such as adversarial machine learning and cyber threat intelligence—for automatically detecting, responding to, and preventing advanced persistent threats.**

### Competitive advantage

- Development of Cyber threat intelligence models such as intrusion detection, privacy-preserving, and digital forensics-based statistics, machine and deep learning models
- Development of automated penetration testing methods based on AI planning
- Design of new testbed architectures for Industry 4.0 networks
- Leading analysis of how AI could develop automated cyber applications, for the Australian Army, Australian Federal Police (AFP), and the Cyber Security Cooperative Research Centre (CSCRC)
- Advanced threat intelligence models for deterring cyber threats and reducing financial losses and critic infrastructure damages

### Impact

The increase in everything-connected, online systems that both sense from and interact with the physical world poses a security risk. The extent to which countries such as Australia are already dependent on cyber-physical systems – which is projected to increase – means that the impact of any disruption is potentially catastrophic.

### Successful applications

- Evaluating Network Intrusion Detection based Deep Learning and Graph Models
- A Collaborative Host-Network Anomaly Detection Framework for Internet of Things
- A new intelligent wargaming web service-based Machine Learning for the Australian Army to understand human influences and behaviours

### Capabilities and facilities

- Cyber Range Labs
- Digital Forensics Lab
- IoT Lab

### Our partners

- Australian Federal Police (AFP)
- Data 61 CSIRO
- CyberCRC
- Australian Army
- Oracle
- Cyber Center for Security and Analytics at UTSA USA

# AGENDA

- **Project Overview**
- **Key Issues & Lessons Learnt**
- **FAIR Principles & TON_IoT Datasets**
- Participants & Collaborators

# Project Overview

**Research challenges**

- Collect Industrial Internet of Things (IIoT) data for Cyber Security Applications
- Analyse and filter IIoT data in standard formats, agreed protocols and properties
- Launch hack and normal events

**Research methods**

- Design IoT network testbed
- Use big data analysis tools for examining collected datasets
- Apply machine/deep learning algorithms for initial evaluation

**Cyber Applications**

- Intrusion Detection
- Threat intelligence
- Digital Forensics
- Privacy Preservation

**Environments**

- Internet of Things (IoT)
- Network systems
- Industrial IoT
- Cloud and Fog

**Future plan**

- Maintaining Data FAIR principles
- Developing an Industry 4.0 Cyber Security Platform for training and research purposes


Australian Research Data Commons

# Key Issues & Lessons Learnt

Development of many IoT services and integration of them into cloud and network systems are one of the issues of scalability and operability.

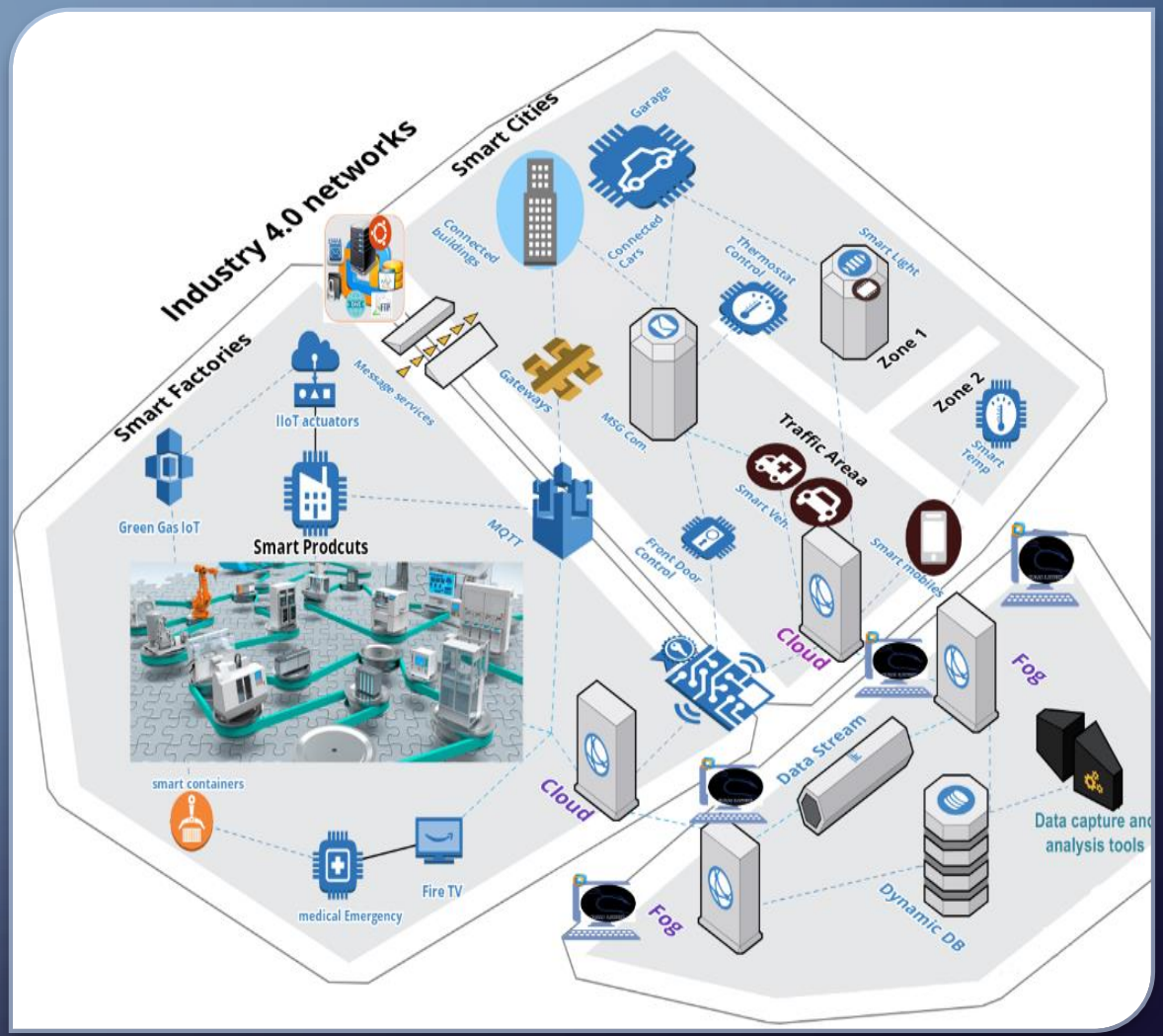
- Managing IoT sensors and network elements into separate layers of edge/fog, cloud and network

Collection of heterogeneous data sources, along with ensuring the correctness of security events, demanded the integration of data science and network security skills.

- Handling the challenge of collecting structured and unstructured data sources and processing their high dimensional space in real-time

Launching hacking scenarios to multiple systems of IoT, network, Windows and Linux is an arduous task because a cyber-attack that could exploit a system vulnerability is different at every system involved in the testbed network in most cases.

- Applying a standard cyber threat framework for various systems. A cyber-kill chain model was utilised to making homogenous vulnerabilities for breaching the systems by homogenous exploits.




An architectural design for generating datasets from the Industry 4.0/IIoT network at the Cyber Range labs at UNSW Canberra

# TON_IoT Datasets For Cybersecurity Applications

- The TON_IoT datasets are new generations of Industry 4.0/Internet of Things (IoT) and Industrial IoT (IIoT) datasets for evaluating the fidelity and efficiency of different cybersecurity applications based on Artificial Intelligence (AI) and Machine/Deep Learning algorithms.
- The datasets have been called ‘ToN_IoT’ as they include heterogeneous data sources collected from Telemetry datasets of IoT and IIoT sensors, Operating systems datasets of Windows 7 and 10 as well as Ubuntu 14 and 18 TLS and Network traffic datasets. The datasets were collected from a realistic and large-scale network designed at the Cyber Range and IoT Labs of the UNSW Canberra Cyber, the School of Engineering and Information technology (SEIT), UNSW Canberra @ the Australian Defence Force Academy (ADFA).
- The testbed was deployed using multiple virtual machines and hosts of windows, Linux and Kali Linux operating systems to manage the interconnection between the three layers of IoT, Cloud and Edge/Fog systems. A set of IoT devices and sensors, such as green gas IoT and industrial IoT actuators, is connected to MQTT gateways to publish and subscribe to various topics, such as measuring temperature and humidity. The datasets were gathered in a parallel processing to collect several normal and cyber-attack events from IoT networks.

# TON_IOT DATASETS

- Different hacking techniques, such as DoS, DDoS and ransomware against, were launched against web applications, IoT gateways and computer systems across the IIoT network. The directories of the TON_IoT datasets include the following:
  - **Raw datasets**: IoT/IIoT datasets were logged in log and CSV files, where more than 10 IoT and IIoT sensors such as weather and Modbus sensors were used to capture their telemetry data.
  - **Network datasets:** were collected in the packet capture **(pcap)** formats, **log** files and **CSV** files of the Bro tool.
  - **Linux datasets:** were collected by running a tracing tool on Ubuntu 14 and 18 systems, especially atop, for logging desk, process, processor, memory and network activities. The data were logged in **TXT** and **CSV** files.
  - **Windows datasets** were captured by executing dataset collectors of the Performance Monitor Tool on Windows 7 and 10 systems. The raw datasets were collected in a **blg** format opened by Performance Monitor Tool to collect activities of desk, process, processor, memory and network activities in a **CSV** format.

# TON_IOT DATASETS (CONT.)

- **Processed datasets:** The four datasets were filtered to generate standard features and their label. The entire datasets were processed and filtered in the format of **CSV** files to be used at any platform. The new generated features of the four datasets were described in the **'Description_stats_datasets'** folder, and the number of records including normal and attack types is also demonstrated in this folder.
- **Train_Test_datasets:** This folder involves samples of the four datasets in a CSV format that were selected for evaluating the fidelity and efficiency of new cyber security application-based AI and machine learning algorithms. The number of records including normal and attack types for training and testing the algorithms are listed in the **'Description_stats_datasets'** folder.
- **Description_stats_datasets:** This folder includes the description of the features of the four processed dataset (the folder of processed datasets) and the statistics (i.e., the number of rows of normal and attack types).
- **SecurityEvents_GroundTruth_datasets:** This folder includes the security events of hacking happened in the four datasets and their timestamp (ts). The datasets were labelled based on tagging IP addresses (192.168.159.30-39) and their timestamps in the four datasets. These IP addressed were used for Kali Linux systems to launch and exploit the systems of the four environments of IoT/IoT systems such as Cloud gateways, MQTT protocols, web applications of Node Red, Linux, Windows and network services.
- The datasets can be used for validating and testing various Cybersecurity applications-based AI such as intrusion detection systems, threat intelligence, malware detection, fraud detection, privacy-preservation, digital forensics, adversarial machine learning, threat hunting. The dataset was sponsored by the Australian Research Data Commons (ARDC) and UNSW Canberra.

# FAIR Principles & TON_IoT Datasets

The new datasets, named TON_IoT (**T**elemetry data of **IoT** devices, **O**perating system, **N**etwork data), and their metadata have been publicly published through the sustained Research Data Australia (ResData) at https://doi.org/10.26190/5d7ac9bfe8487 .

The ResData contains a link to the data on CloudStor at https://cloudstor.aarnet.edu.au/plus/s/ds5zW91vdgjEj9i .

The raw and processed datasets and their tools of collection and analysis have been publicly published to enable developers and researchers in other domains, such as data science and general machine learning applications, for using the datasets.

The TON_IoT have been integrated with our existing datasets, UNSW_NB15 and Bot-IoT, which have been widely used in academia and industry, notably anomaly detection systems of Oracle and Microsoft.

The datasets have been stored in files of a Giga-byte size at maximum to assert the download persistence at any Internet speed.

The UNSW ICT department has a regular update to the datasets and keeps all UNSW datasets available.

The department has a backup plan to ensure that the UNSW public datasets will be available all the time to support researchers and developers for easily downloading the datasets at anywhere and anytime.

# CONCLUSION

- **Free use of the TON_IoT datasets for academic research purposes is hereby granted in perpetuity. Use for commercial purposes is allowable after asking the author, Dr Nour Moustafa, who has asserted his right under the Copyright. The datasets was sponsored by grants from the Australian Research Data Commons, https://ardc.edu.au/news/data-and-services-discovery-activities-successful-applicants/,and UNSW Canberra.**
- For more information about the datasets, please contact the author, Dr Nour Moustafa, on his email: nour.moustafa@unsw.edu.au or eng.nourmosuatafa@hotmail.com.
- More information about Dr Nour Moustafa is available at:
  - https://www.unsw.adfa.edu.au/our-people/dr-nour-moustafa
  - https://research.unsw.edu.au/people/dr-nour-moustafa-abdelhameed-moustafa
  - https://www.linkedin.com/in/nour-moustafa-0a7a7859/
- Links of TON_IoT datasets:
  - https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-ton-iot-Datasets/
  - https://cloudstor.aarnet.edu.au/plus/s/ds5zW91vdgjEj9i