INCS-CoE *Digital Trust* Forum Day 1

Tuesday, November 30, 2021

7:00AM-8:30AM EST; 12:00PM-1:30PM GMT; 9:00PM-10:30PM JST

Hosted by:

Northeastern University

United States

UMBC
Northeastern University

United Kingdom

Imperial College London
Royal Holloway University of London
University of Cambridge

Japan

Keio University
Kyushu University

France

University of Limoges

Israel

Technion Israel Institute of Technology
Ben-Gurion University

Australia

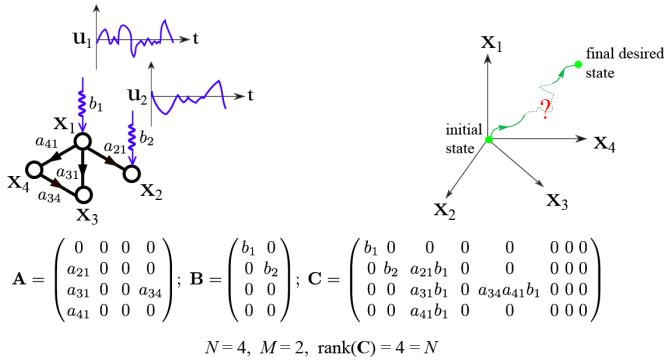
Edith Cowan University

Intervention for Collective Pro-Social Behavior

Babak Heydari, Northeastern University

Do what engineers have been doing for decades **[hierarchical design, control]**

What about autonomy?

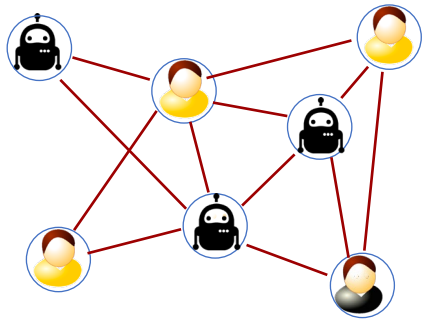


Liu, Slotine, Barabasi(2011)

Do what economists would do **[Incentives]**

What about system's architecture and design aspects?

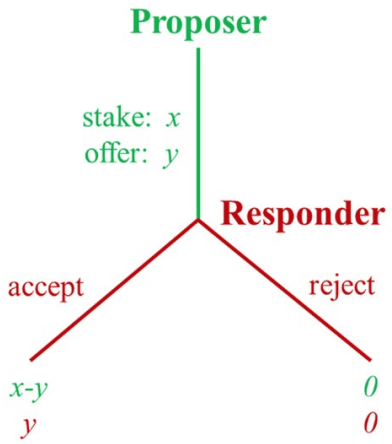
Interaction and communication structure [network architecture and composition]

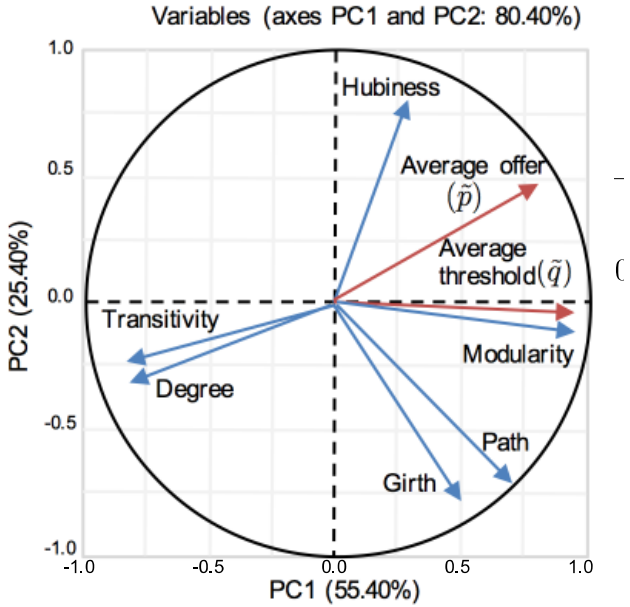


Network Drivers or Fairness and Equity of Decentralized Resource Distribution

Evolutionary Agent-Based Simulation on Networks

Pairwise Ultimatum Game





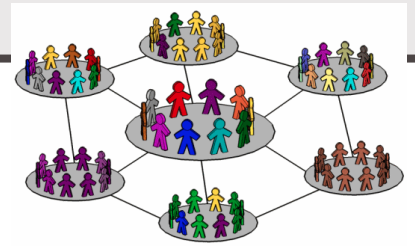
Human has a sense of fairness!

Experimental results of UG:

offer: 30-50%, demand: 25-40%

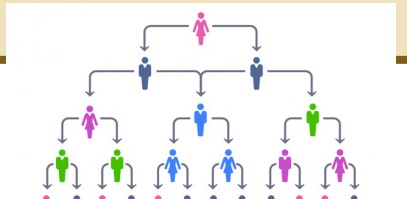
Network Modularity (community structure)

- Size of a group's payoff from cooperation in economic production *(Payoffs to Cooperation)*
- Scale of cooperative units ***(Henrich et. al, 2001)***

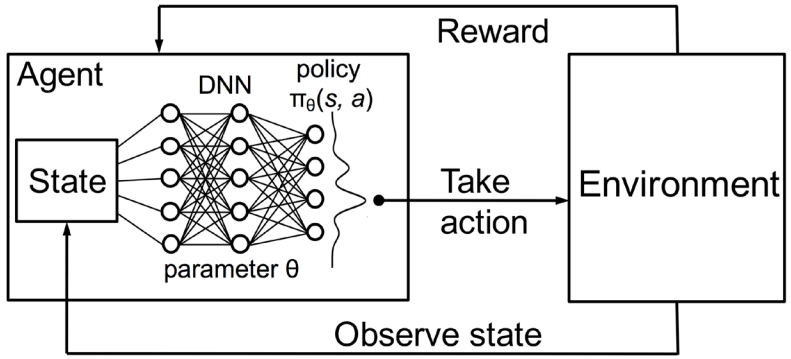


Hubiness (hierarchal structure)

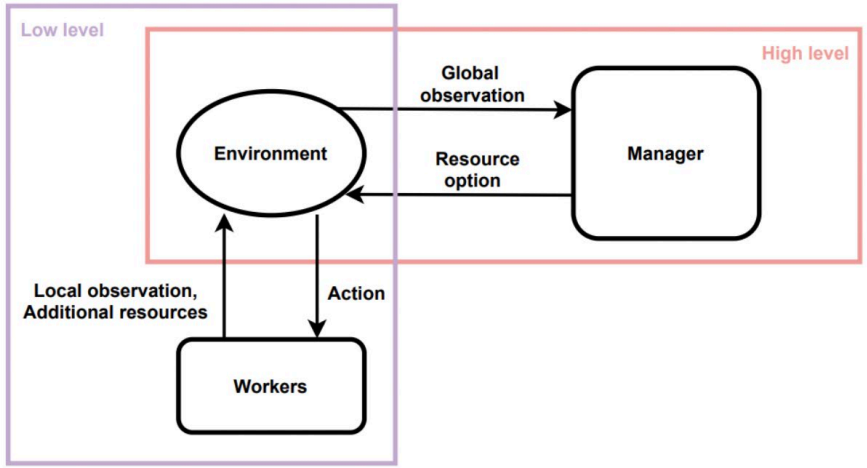
- Governance structure (single vs. multiple interacting chiefdoms)
- Market Integration **(Paciotti and Hadley,2003)**



Mosleh and Heydari (2016 and 2017)



Dynamic Incentive and Communication Structure Using Deep Reinforcement Learning



Algorithm 1 Learning process for workers

Initialize Replay Buffer D_i, actor μ_i and critic $Q^i(s,a)$ for each agent.
for $Epoch = 1, 2, \ldots$ **do**
  for $agent = 1, 2, \ldots, N$ **do**
    Run policy μ_θ in environment for T time steps
    store each agent's history $s_i^t, a_i^t, r_i^t, s_i'^t$ to D_i
  end for
  In training process, each agent i will sample experience $s_i^t, a_i^t, r_i^t, s_i'^t$ randomly from D_i
  Update actor μ_θ using $Equation3$
  Update critic $Q^i(s,a)$ using $Equation1$
end for

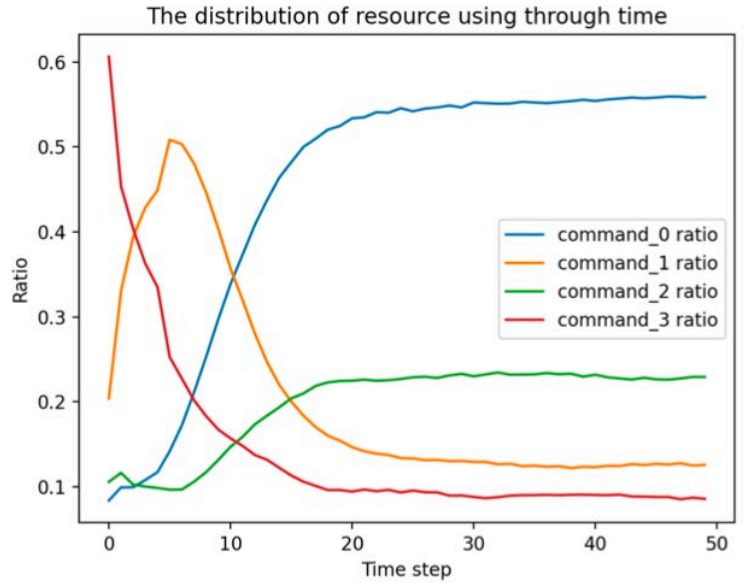
$Q^\pi(s,a)$ can be updated using the method described in Q learning section, this method become the actor-critic algorithm[15]. If we extend the policy t deterministic policies $\mu_\theta : O \rightarrow A$, we get DDPG[5], the gradient of the objective changes as:

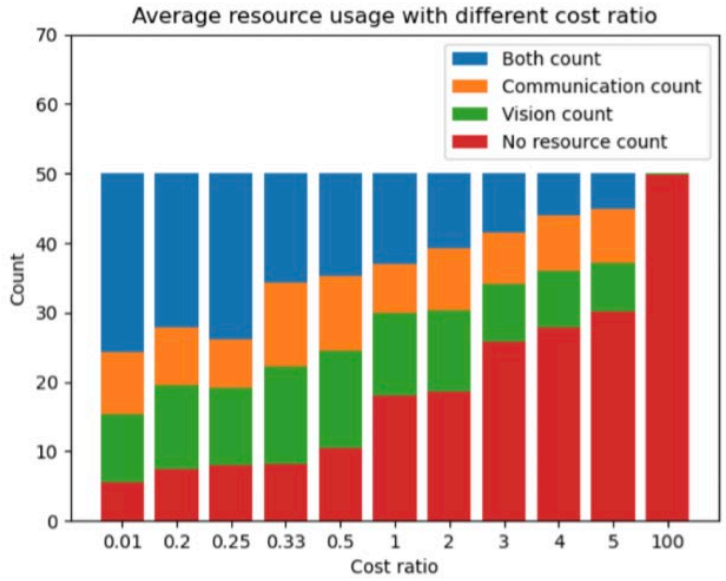
$$\nabla_\theta J(\theta) = E_{s,a}[\nabla_\theta \mu_\theta(a|s) \nabla_a Q^\mu(s,a)|a = \mu_\theta(s)] \quad (3)$$

Also, the representation of $\mu_\theta(a|s)$ is using Deep neural etwork.

lgorithm 2 Learning process for manager

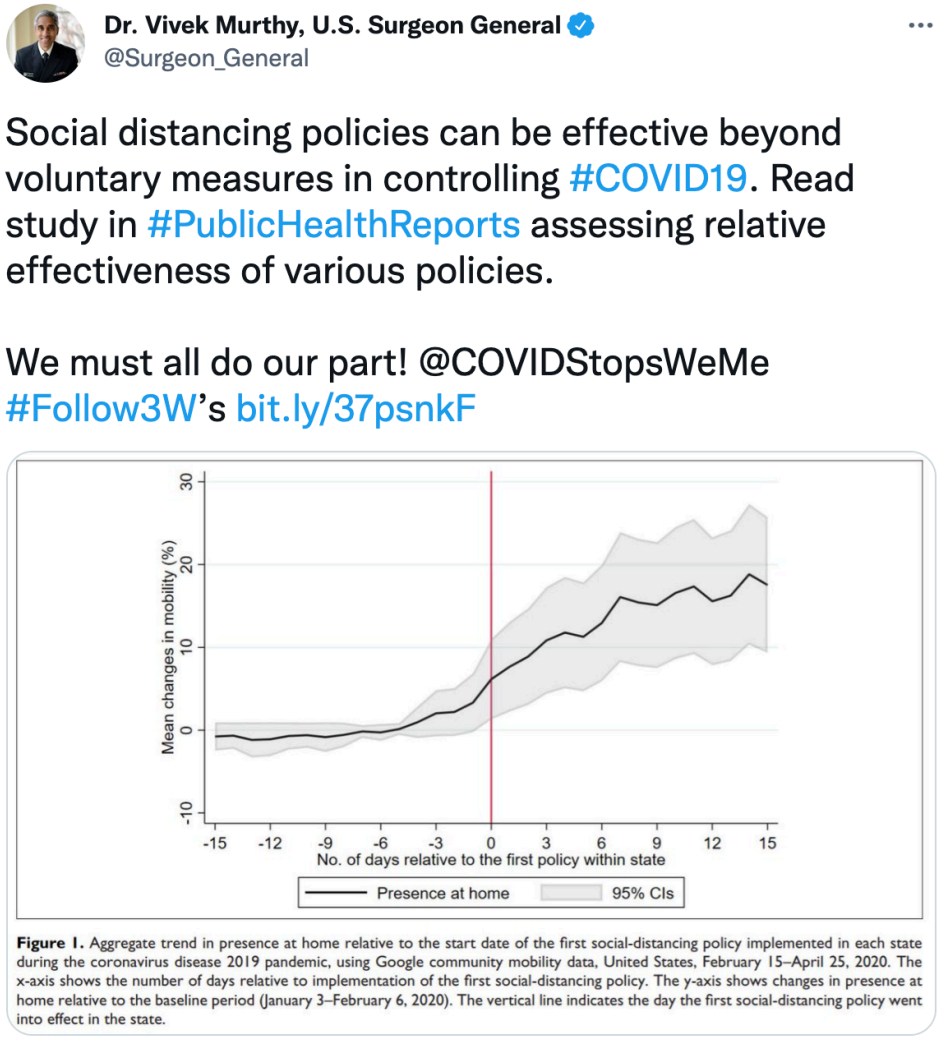
Initialize Replay Buffer D, action value function $Q(s,a)$
for $Epoch = 1, 2, \ldots$ **do**
  for time step=1,2,$\ldots, T$ **do**
    Choose action a at each time step using $a = \arg\max_{a^t} Q(s = s^t, a = a^t)$
    store each manager's history s^t, a^t, r^t, s' to D
  end for
  In training process, manager will sample experience s^t, a^t, r^t, s' randomly from D
  Update action value function $Q(s,a)$ using $Equation1$
end for



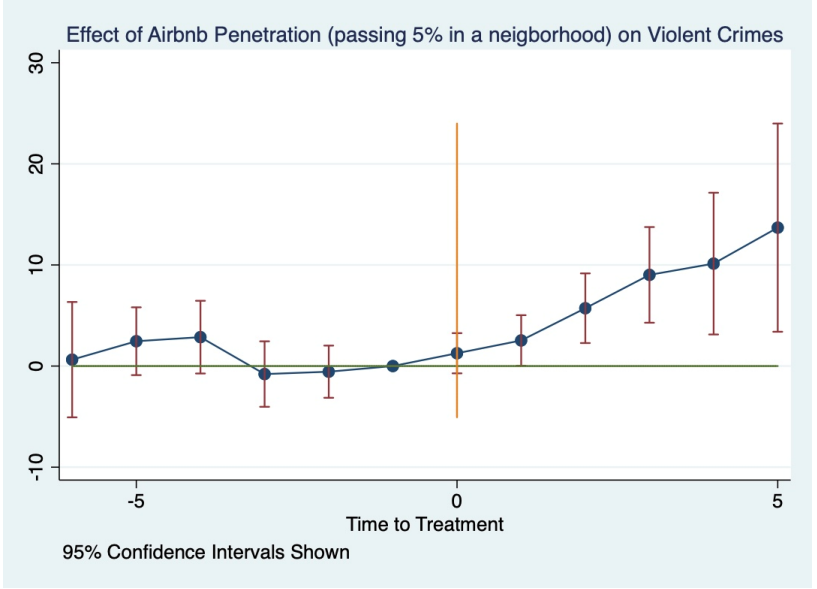


Empirical Methods to measure intervention effectiveness

Separating Voluntary vs. Policy-driven Social Distancing for COVID19
(Abouk, Heydari, 2021)



Airbnb and Social Organization
(Ke, O'Brien, Heydari, 2021)



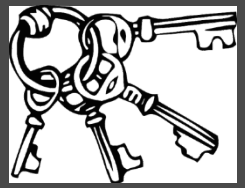
Difference-in-Difference Causal Identification

Naghmeh Karimi
Assistant Prof.
Director of SECRETS (SECure, REliable and Trusted Systems) lab
CSEE Department
University of Maryland Baltimore County

Hardware Security & Trust
Hardware Assisted Cyber Security

Integrated Circuits

Dependability



Security &Trust

Security & Trust

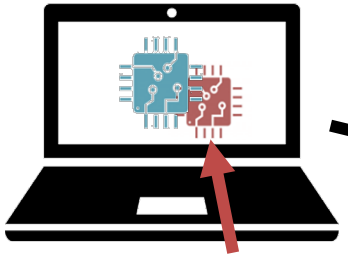
- **Confidentiality**
- **Availability**
- **Integrity**

- Ensuring the security and trust in Integrated circuits (ICs) is of outmost importance.
- Securing the software alone is not a solution as an untrusted hardware infrastructure can compromise the whole system.
- The problem is exacerbated for the data sensitive chips, e.g., cryptographic modules.

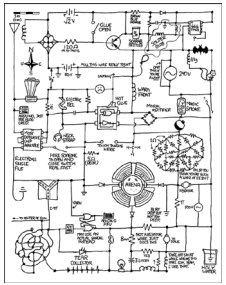
The IC Landscape

Distributed Manufacturing Process (Multiple Companies)

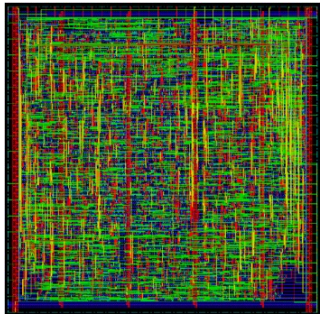
Design



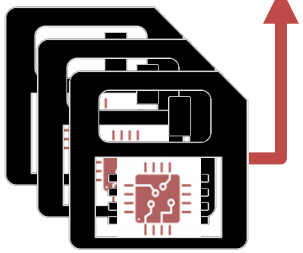
Synthesis



Place & Route



IP Integration



Other Companies

Third Party Fabrication

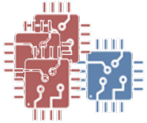
Attack Motivations & Related Countermeasures

$$$

Counterfeit Devices

Recycled

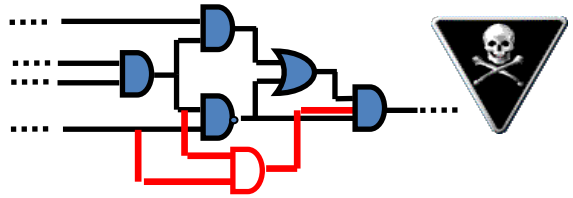
Cloned



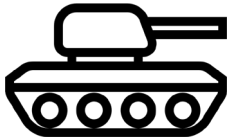
Overproduced

Tampered Devices

Trojan- Infested



Denial of service
System malfunction
Leaking sensitive data



Our group expertise in approaching Trust & Security concerns

- Hardware-assisted authentication protocols in IoT frameworks
- Hardware based tampered-device detection
- Obfuscation schemes to protect the Intellectual Properties (IPs)
- Countermeasures against side-channel analysis attacks to preserve data secrecy

UMBC

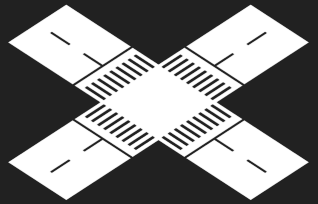
Thanks

Roberto Yus, Assistant Professor

UMBC
CSEE Department
ryus@umbc.edu

Data Management

Knowledge Representation



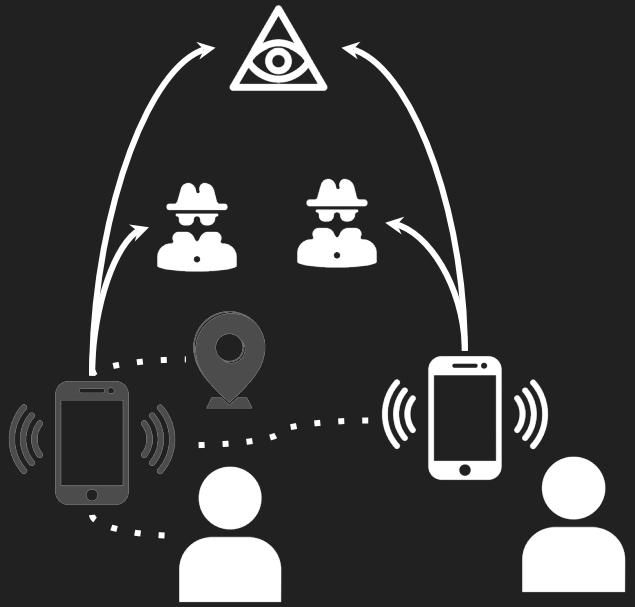
Privacy

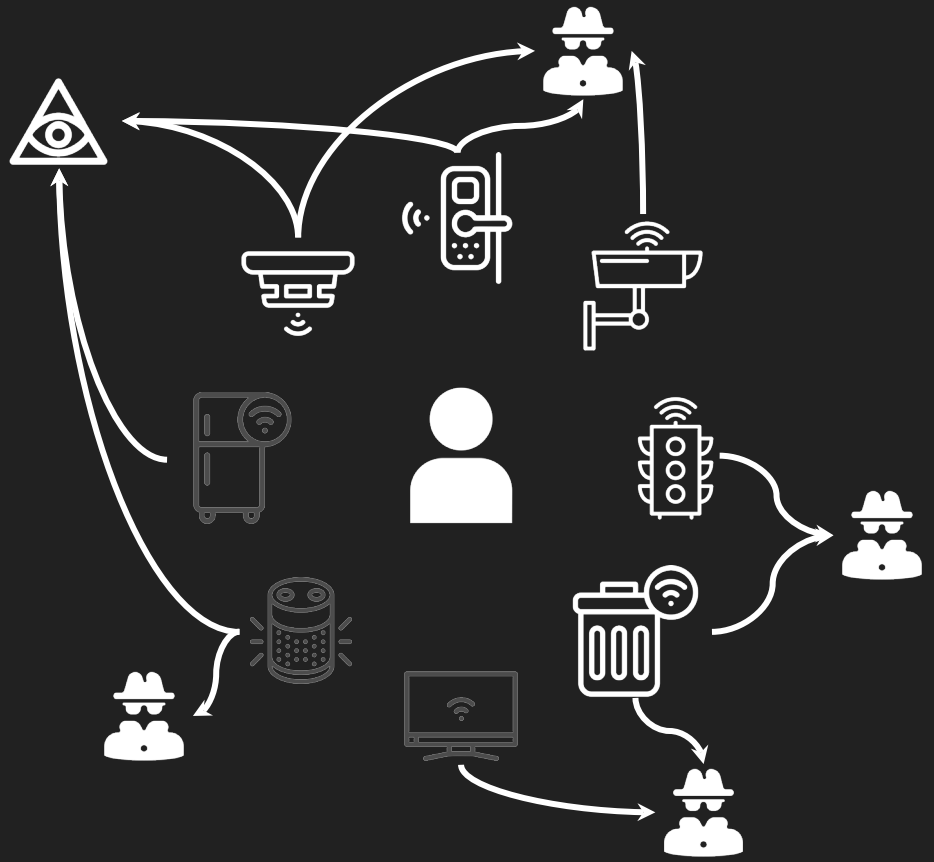
Internet of Things

Design **semantic data management** solutions to empower **IoT systems** to understand user information requirements, as well as their **privacy preferences**, and tailor their operations to those.

Spaces are Becoming Smarter...

… and More Intrusive



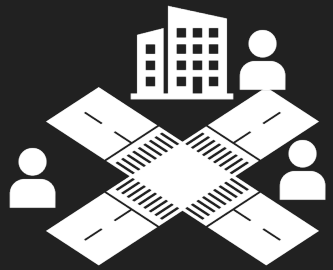


Roberto Yus (https://robertoyus.com)
UMBC

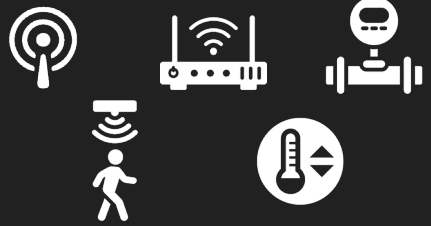
Semantic & Privacy-Aware IoT Data Management

"Do not track my location"

"Do not share my social interactions with applications"



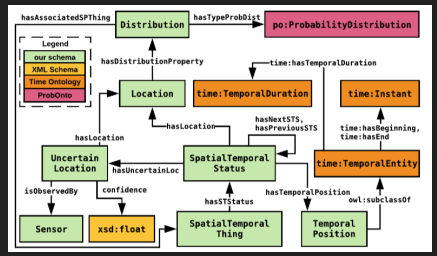
SEMANTIC GAP

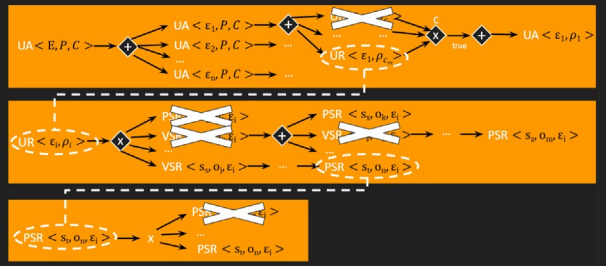


IoT System

- **Understand both worlds and their interrelations**
- **Translate requests/ commands/ policies**
- **Enrich raw data**
- **Enforce privacy preferences**

Efficiency





Real-world deployments

NAVWAR

UNIVERSITY OF CALIFORNIA · IRVINE

BALL STATE UNIVERSITY

[1] Sieve: A Middleware Approach to Scalable Access Control for Database Management Systems. Proc. VLDB Endow.
[2] SmartBench: A Benchmark For Data Management In Smart Spaces. Proc. VLDB Endow.
[3] Abstracting Interactions with IoT Devices Towards a Semantic Vision of Smart Spaces. BuildSys19

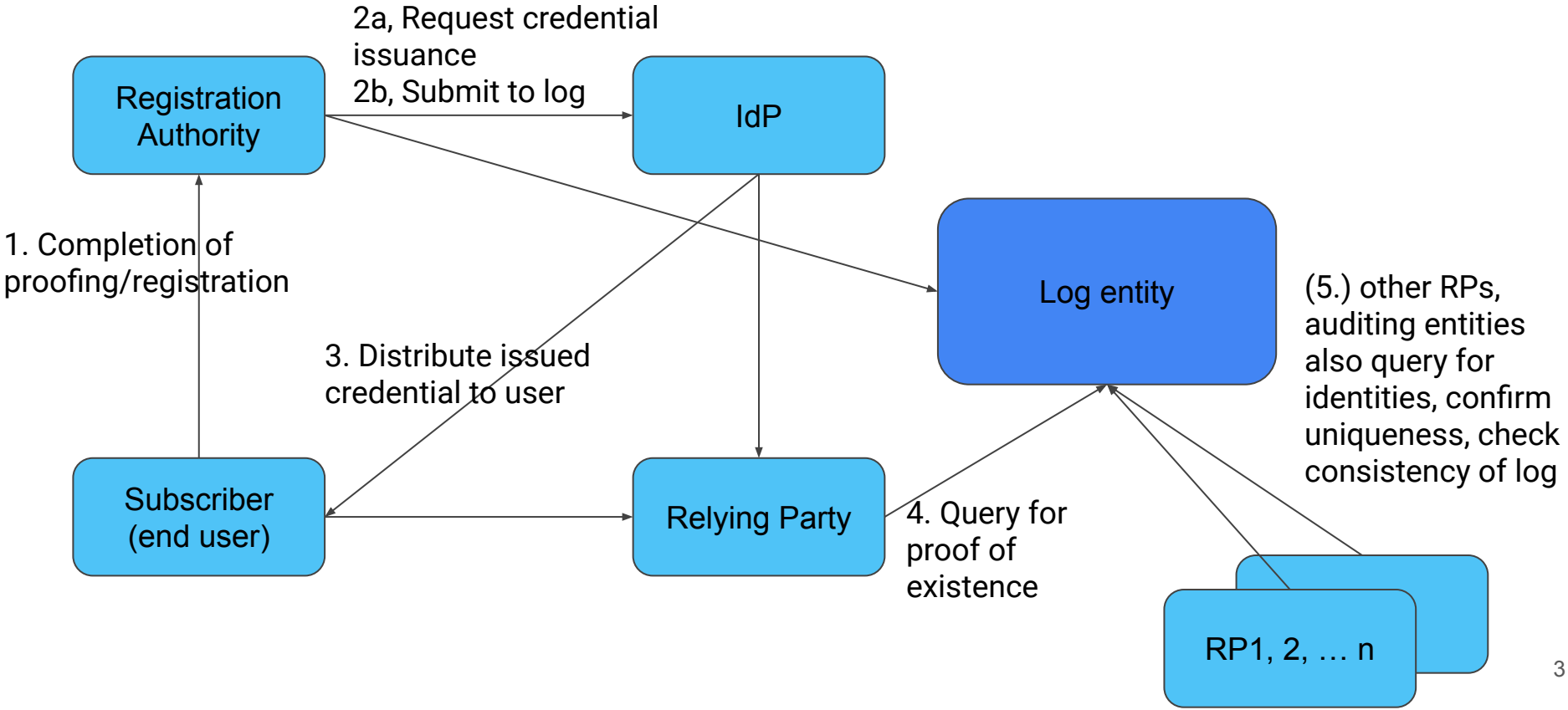
Improving visibility into provider credential issuance

Korry Luke (koluke@sfc.wide.ad.jp)
Keio University

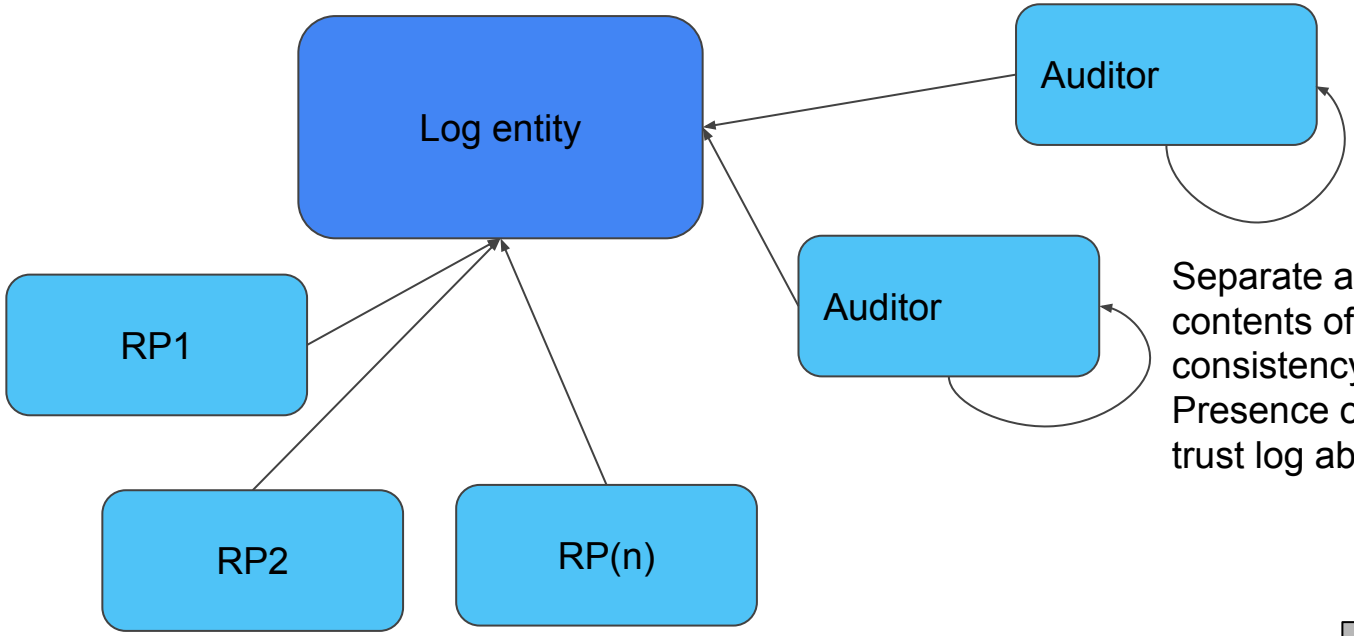
Parallels between IdPs and CAs: disconnect between technical and real world

- CAs in Web PKI used to be somewhat blindly trusted
  - Covered by periodic audits, operational practices, etc., but limited technical constraints
  - DigiNotar 2011 incident: misissued but technically valid certificates emphasized importance of monitoring trusted entities
  - CAs now required to comply with Certificate Transparency (logging system)
- IdP world has similar characteristics
  - IdPs in a federated system are generally trusted/trustable
  - However, they may have imperfect operations or be susceptible to attacks both before and after the enrollment process
  - As part of federation setup, IdPs provide CPS, operational statements, audit compliance, etc.
  - Federation, single-sign on growing in use in private, public sectors
  - Ticking time bomb?
  - Operational best practices, some form of standards exist, but don’t address underlying unilateral blind trust problem

Peer-based log audit mechanism for credential issuance



Log operation/monitoring



Separate auditor entities can download contents of the entire log entity, confirm consistency of data
Presence of auditors removes need to trust log absent-mindedly

1. RPs query for validity of IDs they receive
2. Logs provide proof of inclusion for each entry, which logs verify against log's public key

Contact info:
Korry Luke
Keio University
koluke@sfc.wide.ad.jp

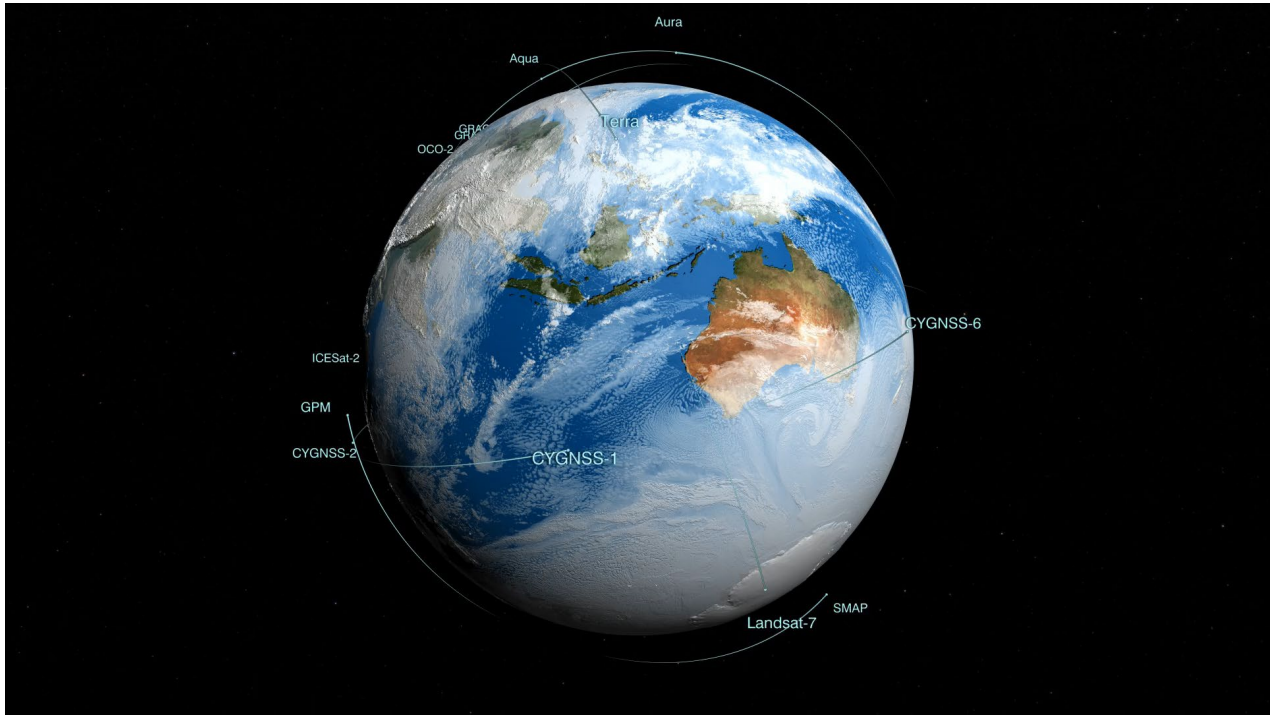
Digital Trust

Kostas Research Institute (KRI) at Northeastern University

Cordula A Robinson, Ph.D., E: c.robinson@kri.neu.edu, 857.319.6650

OVERVIEW

- KRI knowledge prototypes for u*se-inspired research and operational outcomes.*
- G-MAP –capability as a substratum tracking and listening engine to understand subtle cues and to predict propagating behaviors. The obvious first use case is the current pandemic, also considering many other applications including spread of chem-bio agents and information operations also be relevant to DoD.
- MEAD – dual use to protect our data assets being automatically mineable sources, thus, prevent adversaries from noticing patterns of strategic importance.
- Utilize massive amounts high-dimensional data for detection, tracking, forecasting.
- Data is an unreliable friend



M.E.A.D- Manipulating and Exploiting AI Data

- Cordula A. Robinson, Ph.D. KRI
  - Prof Paul Hand and Shelley Lin @Northeastern University,
  - and Richard J. Wood, Ph.D. SSCI

Systems in the absence of trust

STRATEGIC ADVANTAGE

- US Intelligence require as much strategic advantage as kinetic and non-kinetic battlefields continue to merge.
- Understanding adversarial capabilities in intercepting US geospatial data sources in near real-time requires new approaches to address those threats.
- Digital arms race implies the side that “leverages data to gain military advantage will be the side to win wars at speed and scale” (Pentagon’s new data strategy).
- Our goal - offer proactive defensive means to increase an intelligence analysts’ ability to operate in the wide-open via purposeful obfuscation techniques and maintain a low-profile.

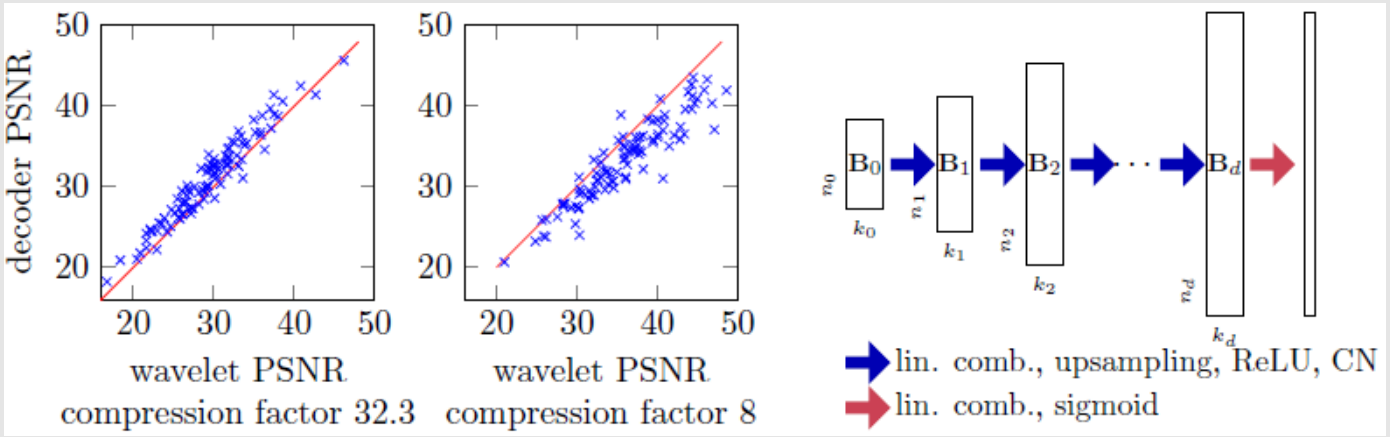
Dual-use AI

- Geospatial Cyber: Data poisoning; Cyber and evasion attacks (model fooling); Extraction attacks (model stealing); and/or Model inversions.

- Dual-use AI:
- Should be effective against automatic scraping and classification (imperceptible, universal and resistant tainting methods).
- Should protect our data assets as automatically mineable sources, thus, prevent adversaries from noticing patterns of strategic importance.

- Prototyping MEAD relevant to 1) assure data access/integrity in offensive end-to-end data delivery and data assurance; and 2) modernize offensive analytic workflows.

Why MEAD? Why Kostas Research Institute?



- Adversarial images are close to natural as possible.
- Leverages advances of the field of image priors, including the field of compressed sensing.
- Natural image perturbations more resistant to image/video degradation than unnatural (static or other high-frequency).
- Larger magnitude perturbations can be admitted before natural perturbations are visible to humans or automated detectors
- Cordula A Robinson, Ph.D. c.robinson@kri.neu.edu

International Digital Trust Forum

Dr Li Zhang, Reader, Department of Computer Science

ROYAL HOLLOWAY UNIVERSITY OF LONDON

Research Interest and Expertise

ROYAL HOLLOWAY UNIVERSITY OF LONDON

My research expertise

-- Deep Learning, Machine Learning, Computer Vision, Biometrics and Intelligent Robotics

Related topics

-- Mechanisms for improving digital trust

-- Human and societal aspects of digital trust

Applications

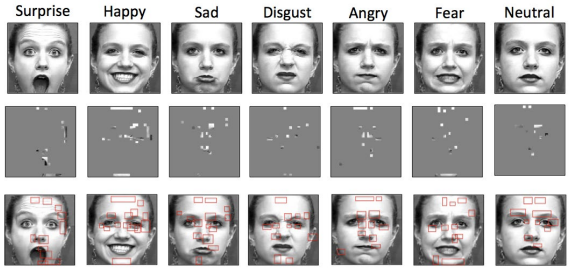
-- AI and robotics systems

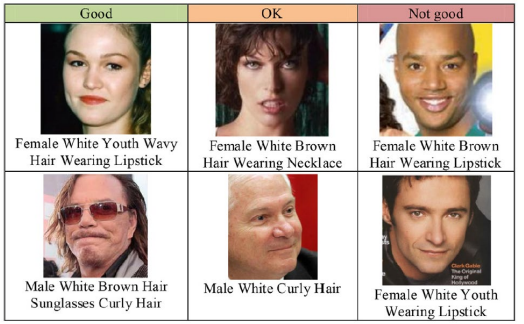
-- Facial expression/gesture recognition

-- Face, fingerprint and Iris recognition

-- Deepfake detection

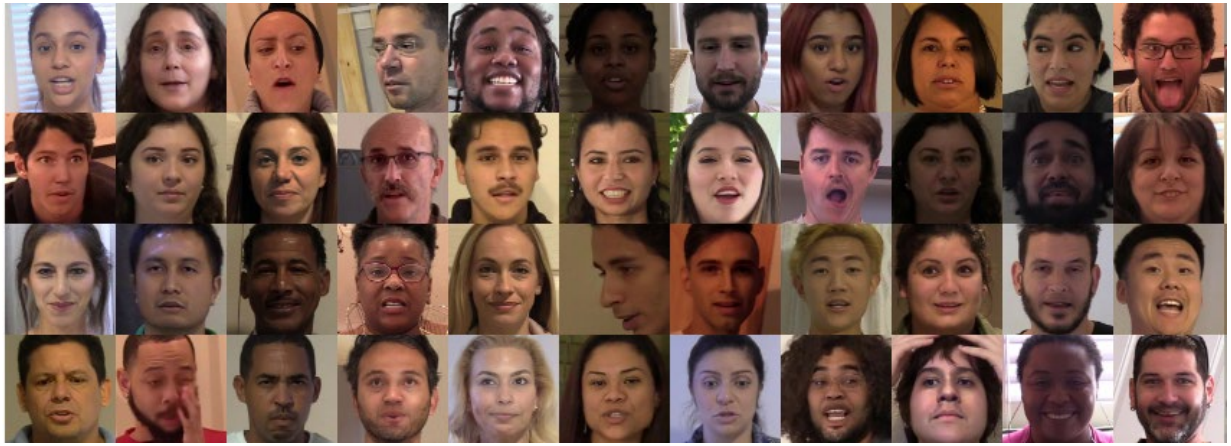
-- Detection of P2P Botnet and online phishing emails

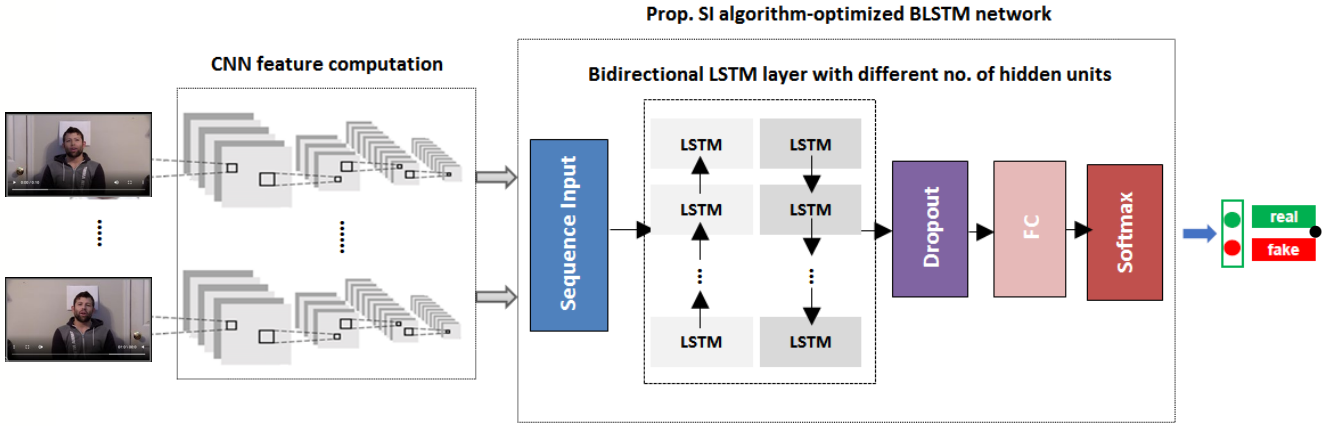




- **(PI) European Regional Development Fund (ERDF)** – Intensive Industrial Innovation Programme (IIIP). Image Description Generation for Monitoring Well-being in the Elderly (2019-2023)
- **(PI) ERDF – IIIP** – Video Surveillance for Health and Safety Monitoring in Retail Stores (2018-2022)
- **(Co-I) London Tech Bridge, APEX Undersea Challenge –** Remote and Accurate Detection of Underwater Obstacles (06/2021-02/2022)
- **(Co-I) Innovate UK (ISCF Manufacturing made smarter)** (03/2021-09/2021)

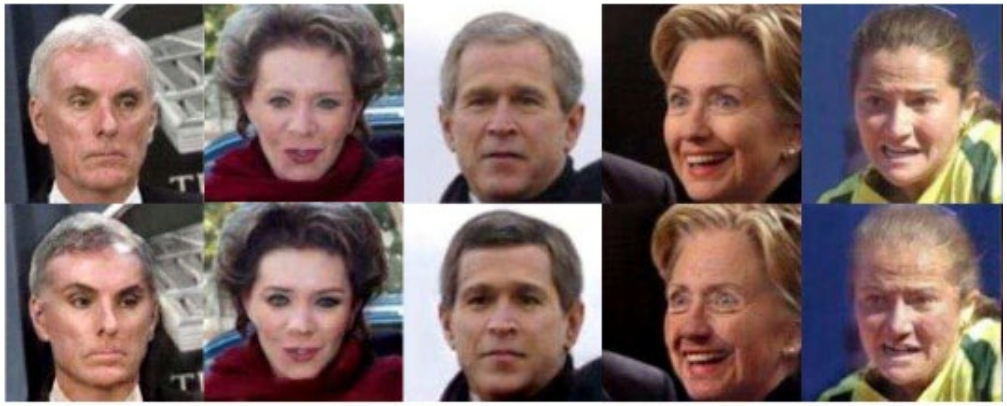
Real-World Face Forgery Detection



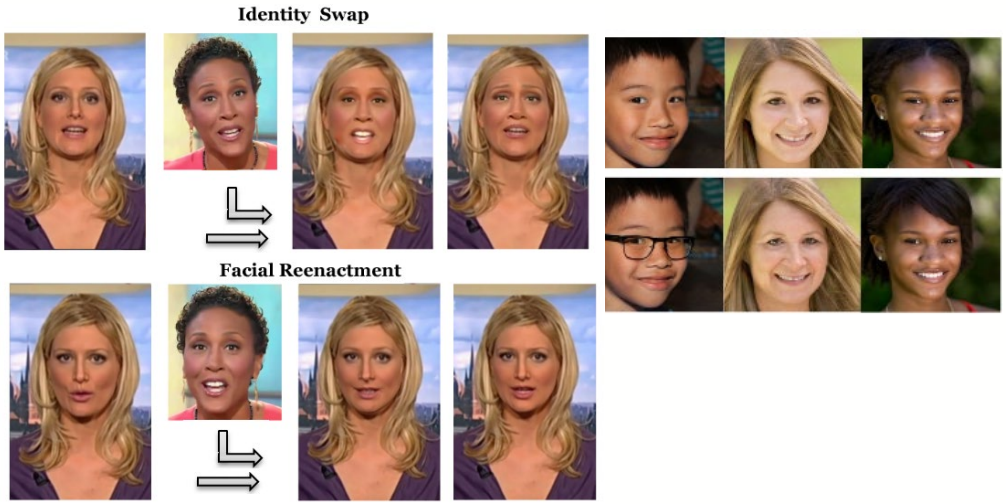


- Identify videos with facial or voice manipulations
- AI-generated synthetic media becomes the most significant cyber threat.
- The DeepFake Detection Challenge (DFDC) Dataset – over 100,000 total clips with 3,426 actors produced with diverse face swap methods [Dolhansky et al., 2020]

Videos in indoor and outdoor settings, with a variety of real-world lighting conditions and pose variations



Learning Residual Images for Face Attribute Manipulation [Shen and Liu, 2017]



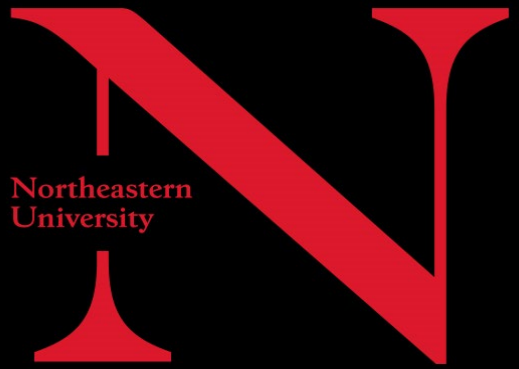
DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection [Tolosana et al., 2020]
FaceForensics++ [Rossler et al., 2019]

- Other types of attacks – identity and expression modification [Rossler et al., 2019]
- Facial expression manipulation (i.e. facial reenactment techniques) - Transfer facial expressions of one person to another person in real time
- Identity manipulation - Replace the face of a person with the face of another person using lightweight models running on smartphones.
- Other face manipulations include attribute modification, e.g. altering hair styles, adding glasses, ageing effects etc.

Other Face Manipulation Detection

INCS-CoE Digital Trust Forum

Nov. 30, 2021



Institute for

Experiential AI

AI and Digital Trust – some thoughts

Usama Fayyad
Executive Director, IEAI
Professor of the Practice, Khoury College for Computer Sciences

Institute for
Experiential AI
Experiential AI

Northeastern University

Four Key Ideas

How can AI and Data Science contribute to Digital Trust?

1. ***Monitoring is a key general challenge:*** without AI the problem is tedious, difficult, and hopeless

2. ***Too Many Monitoring Tools and too many false alarms:*** how do we change the dynamic?

3. ***Sharing is Caring:*** How do you establish trust in how data and signals are shared?

4. ***Context is Key:*** understanding and modelling context of events ad entities is hard and data-intensive

Institute for Experiential AI

Experiential AI

Northeastern University

Monitoring Problems

Manifest in many domains and require much attention

- ***Healthcare and Health***
- ***Surveillance and physical security***
- ***Cybersecurity***
- ***Manufacturing***
- ***Operations***
- ***Public services (transport, traffic, crowd management)***
- ***Network performance***
- ***Fraud detection and prevention***

Institute for Experiential AI

Experiential AI

Northeastern University

An Example Application Area:

Monitoring in Cybersecurity

How the SOC can benefit from embedding human-in-the-loop AI to gain efficiencies

Institute for Experiential AI

Experiential AI

Northeastern University

Motivations

Cybersecurity is an urgently needed, emergent area, rich for applications of Data Science and AI

- Opportunity to create shared data repositories
- Opportunity to create a focused area for training and experimentation
- Major area of investment for all enterprises, government, and many small startups
- Northeastern has a strong presence in this area

Formulation

Cybersecurity systems and monitoring tools generate an overwhelming volume of false alarms

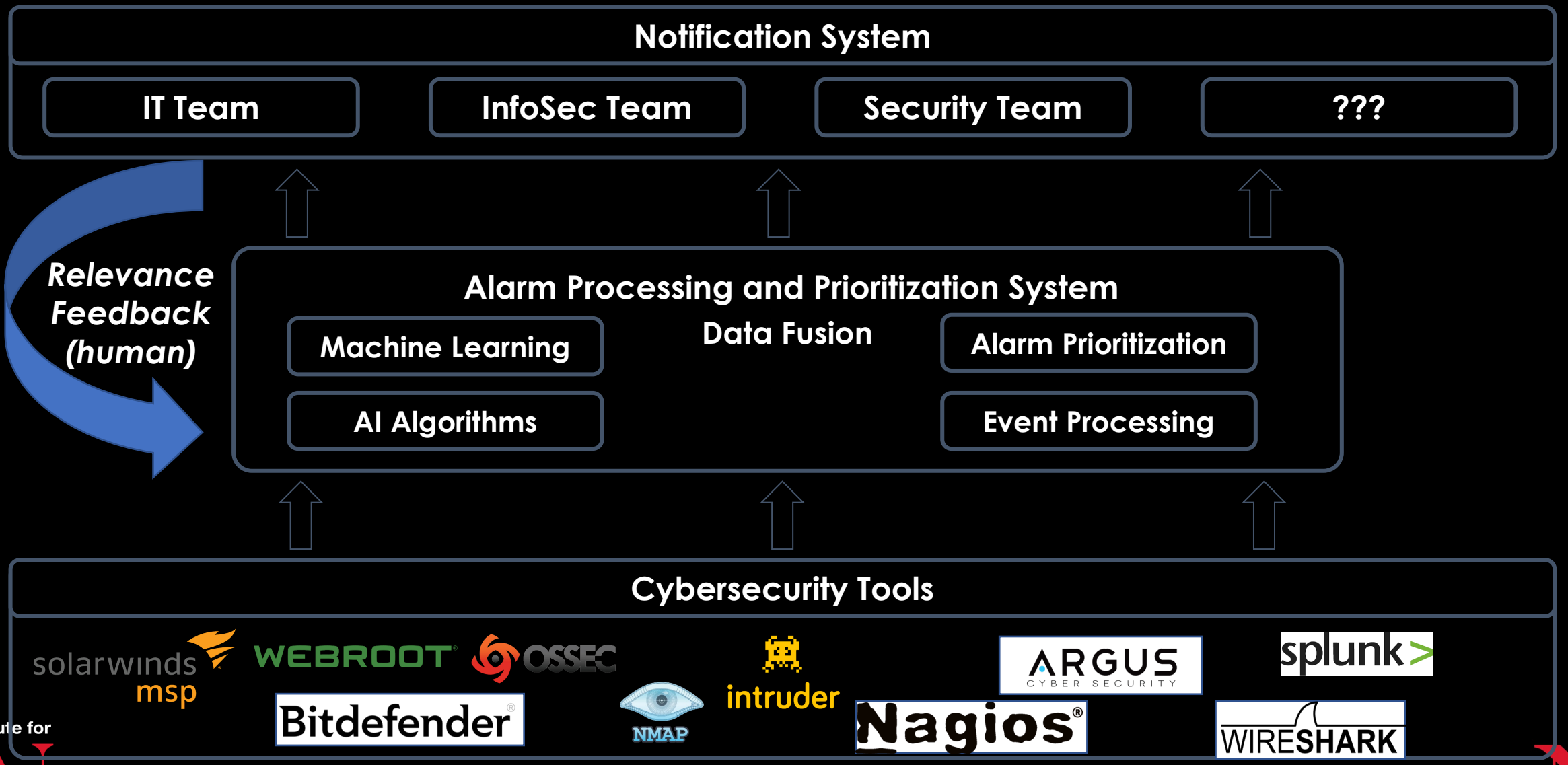
- ***Problem:*** Organizations use many tools and programs for cybersecurity. These generate too many false alarms
- ***Solution:*** Leverage AI/ML/DS to prioritize alarms for focus of attention, then enable context understanding for human review/relevance feedback. Enable faster data gathering for investigations
- ***Adapt:*** each relevance feedback adds to smarter training of system to refine future alarm prioritization

Institute for Experiential AI

Experiential AI

Northeastern University

Cyber Security Data Fusion



Institute for Experiential AI

Experiential AI

Northeastern University

The Traditional Approach To Cybersecurity

THREAT CENTRIC

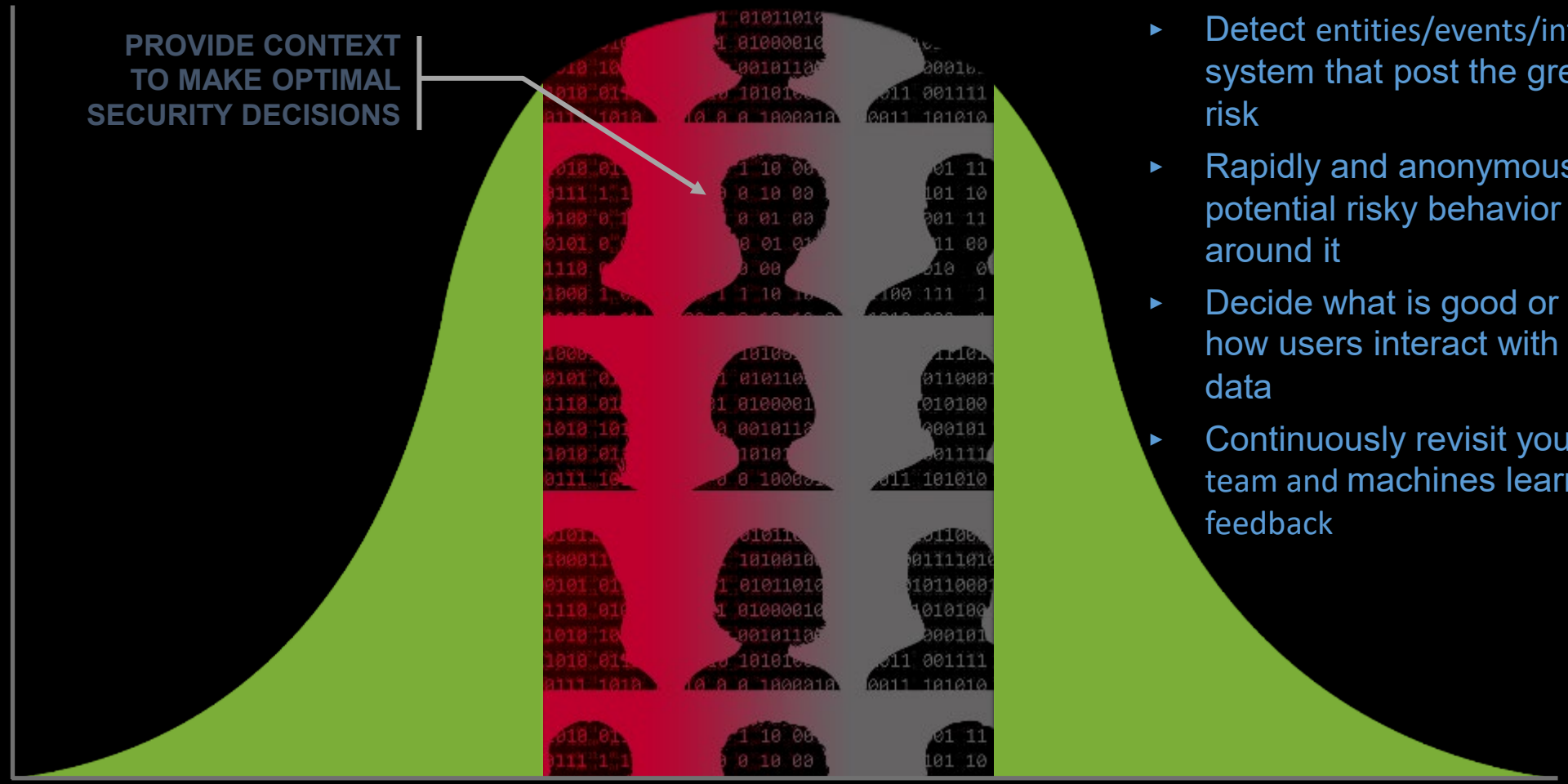
- Trusting static policies in a dynamic environment
- Decide what is good or bad at a single point in time
- Configure your defenses to stop the bad from entering and allow the good to pass through

Institute for Experiential AI

Experiential AI

Northeastern University

A New Paradigm: Human-centric Cybersecurity



BEHAVIOR CENTRIC

- Detect entities/events/interactions with system that post the greatest potential risk
- Rapidly and anonymously understand potential risky behavior and context around it
- Decide what is good or bad based on how users interact with most sensitive data
- Continuously revisit your decisions as team and machines learn from event feedback

Institute for Experiential AI

Experiential AI

Northeastern University

Sharing Data

- If your neighbor is attacked and you were spared, just wait!
- Sharing data and signals can help
  - Recognize attacks faster
  - Prevent future attacks
  - Understand how to create counter-measures
- Cybersecurity data should not be treated as a competitive advantage?
- E.g. CDA – Cyber Data Alliance at Barclays – we got 9 competing EU banks to share data

The Challenges:

- How do you insure privacy?
  - Strong anonymization
  - Differential Privacy
  - Restrict usage through contracts
- How do you preserve secrecy and security?
- How do you define events of interests and map them to effective recognition?

Questions/Discussion

U.Fayyad@northeastern.edu

Institute for Experiential AI

Northeastern University