Integrating Human Reasoning and Machine Learning for Causal Learning Applied to Defense Applications

Dr. Ying Zhao
Naval Postgraduate School

Invited Talk at the Supsec 3rd workshop: AI for Supervision
https://supsec.github.io/
September 19th, 2022
Inria Rennes, France

About the Naval Postgraduate School (NPS)

- US Navy-owned university, located in Monterey, California, USA.
- Faculty and facilities including US DoD connections and classified labs
- Esteemed military experts, strategists, and policy influencers.
- Collaborates with DoD research labs and word-class research organizations
- NPS students are experienced warfighters and government civilian engineers Recently formed the Naval Warfare Studies Institute (NWSI) to accelerate and advance NPS educational and applied research activities to Naval and Marine Corps priority operational problems.

DISTRIBUTION STATEMENT A, APPROPRIATE FOR PUBLIC RELEASE

Myself

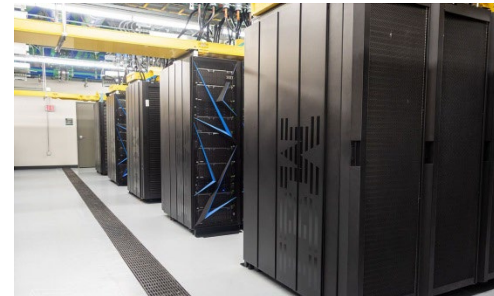
http://faculty.nps.edu/yzhao/

- Ph.D. on Mathematics and Artificial Intelligence from MIT
- Industrial experience: principal researcher
  - Defense contractors: BBN
  - IBM research and global consulting
  - Database marketing and data mining applications, banking and insurance problems
- Co-founded Quantum Intelligence, Inc.: DoD small business innovation research (SBIR), many Phase I, 3 Phase II projects, 4 patents, participated Trident Warrior exercise, joined NPS after that
- Joined NPS in 2009
  - Research Professor
    - The Naval Research Program (NRP) Research Group: Data Sciences Meet Machine Learning and Artificial Intelligence for Military Applications
    - Social and semantic network analysis, Navy recruiting, Navy acquisition research, online persona
    - PI for DoD funded projects of big data analytics applied to combat ID, logistics, cyber, wargaming
    - Students thesis projects
- Current an ESEP scholar at the Defence Science and Technology Lab at the UK

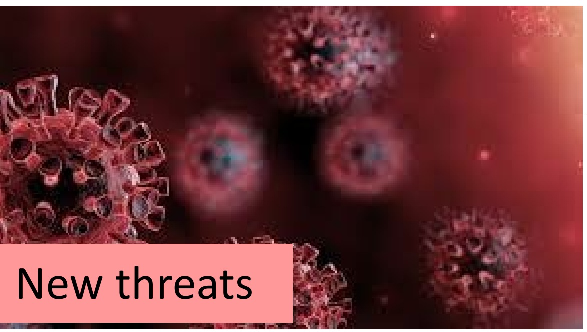
Face Overwhelming Challenges and Opportunities

No data or bad/fake data

Generation: xVs



Data storage, cloud, parallel computing GPU, TPU,



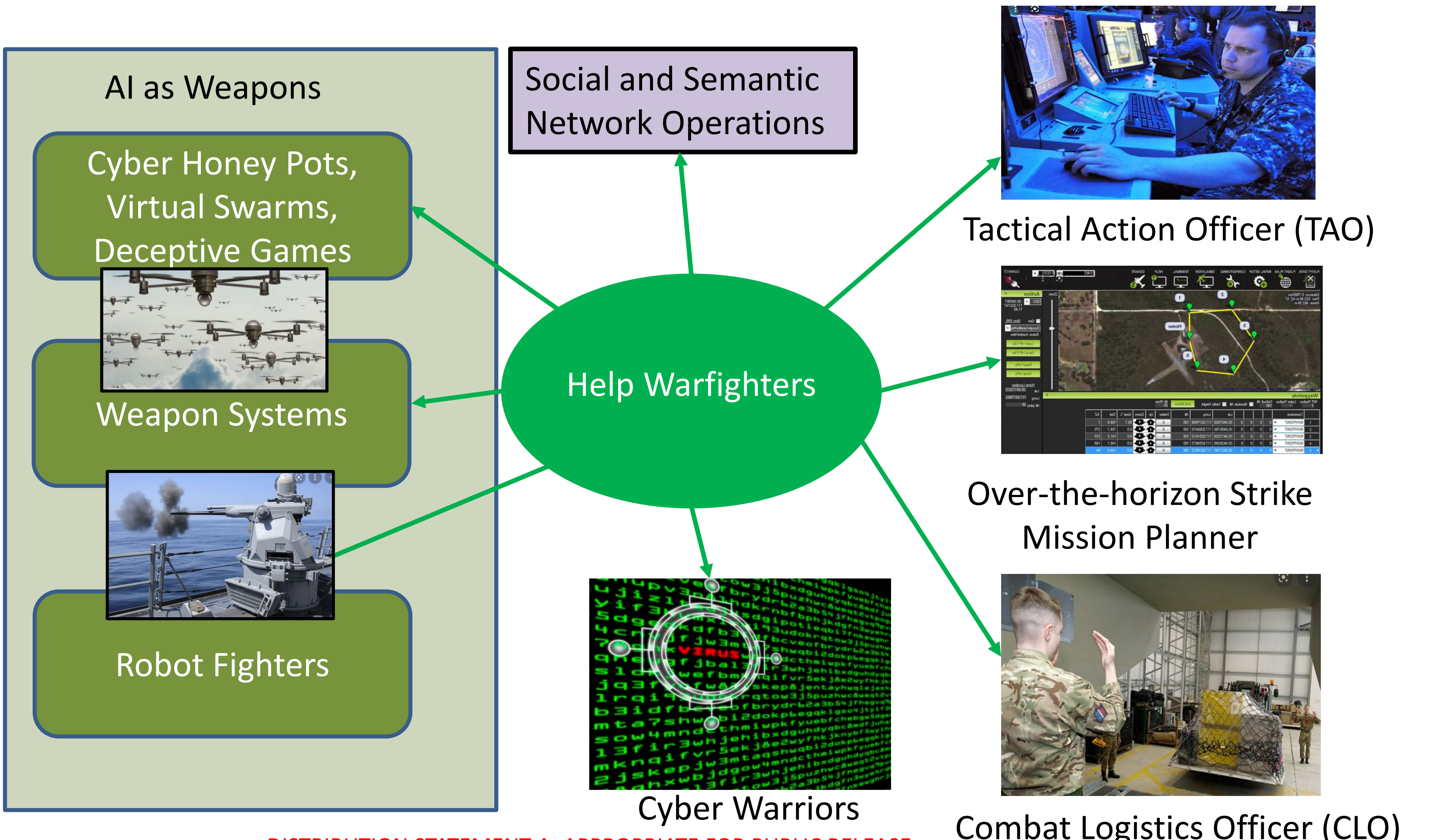
New threats

New challenges

Deep Analytic Algorithms

- **Statistics**
- **Business Intelligence**
- **Deep Learning**
- **Machine Learning**
- **Optimization**
- **Game Theory**
- **Complex System Theory**
- …

Warfighters Need Automation Tools and Trusted AI Used in Different Levels of Applications and Operations



Importance of Studying Integrated Human Reasoning, Machine Learning and Causal Learning

- Focus on the visualization and causal inference aspects of human reasoning
- Provide trusted and safe automation and AI tools
- Provide explainable and actionable information to human operators
  - AI/ML models and simulation models,
    - consistent
    - explainable, no black boxes
    - test theories for a range of users in a wide range of applications such
      - campaign/mission planning
      - future warfighting concepts designing and simulation
      - warfighter training, etc., allow different questions to be asked easily
- Link to ML/AI advancement, and Turing tests
  - Important for studying AI, cognition, and metacognition
  - Artificial General intelligence (AGI): a knowledge system is always with us, learns itself and helps us learn

Use Case 1: Los Alamos National Laboratory's corporate, internal computer network (https://csr.lanl.gov/data/cyber1/)

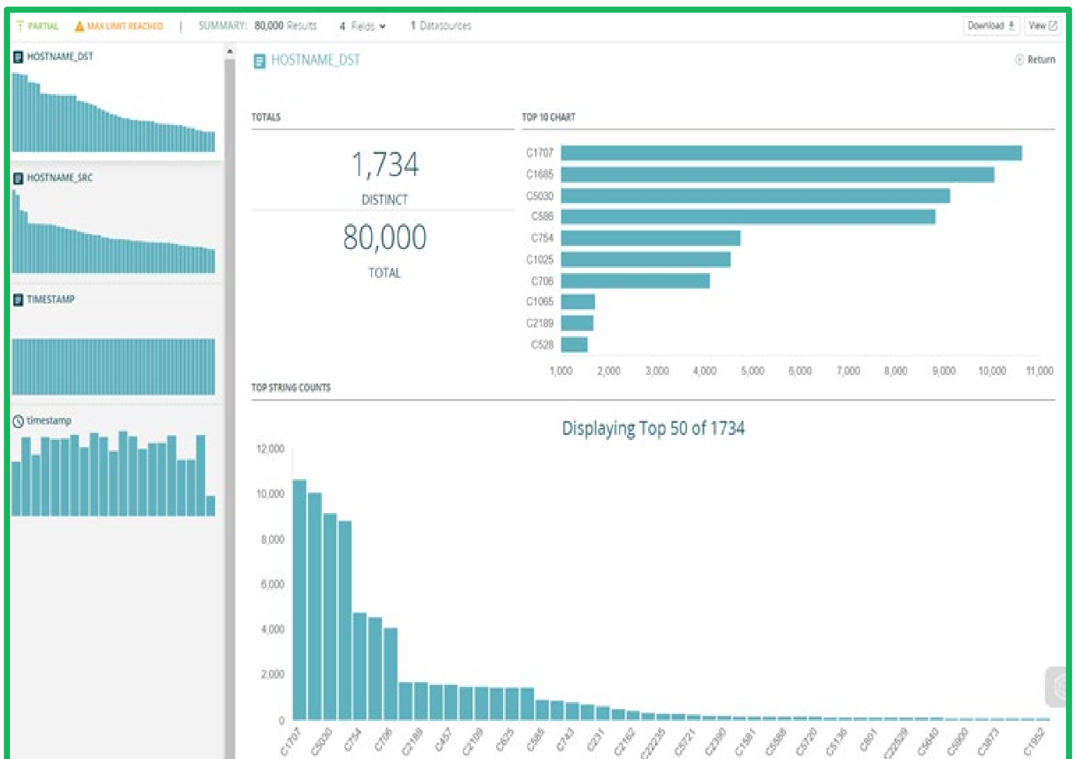
- 58 consecutive days de-identified windows-based authentication events, ~1.6 billion events
  - individual computers
  - centralized Active Directory domain controller servers
  - process start and stop events from individual Windows computers
  - Domain Name Service (DNS) lookups on internal DNS servers
  - network flow data at several key router locations
- ~15,000 computers
- ~12,000 users
- ~60,000 processes
- 12 gigabytes compressed
- ~2% of the computers were hacked or hacking
- The goal is to accurately classify the hacked or hacking computers from the rest of the normal ones.

Data Cleaned

- De-identified
  - Some of the well-known ports (e.g., http port 80 and 443),
  - Some protocols (e.g., 6 for Transmission Control Protocol),
  - Some users (e.g., SYSTEM or Local Service) are left identified
- Time is captured in one-second intervals, starting with a time epoch of (1).

Human Analyst's Approaches

- Visualize using the Big Data Platform (BDP) and other tools

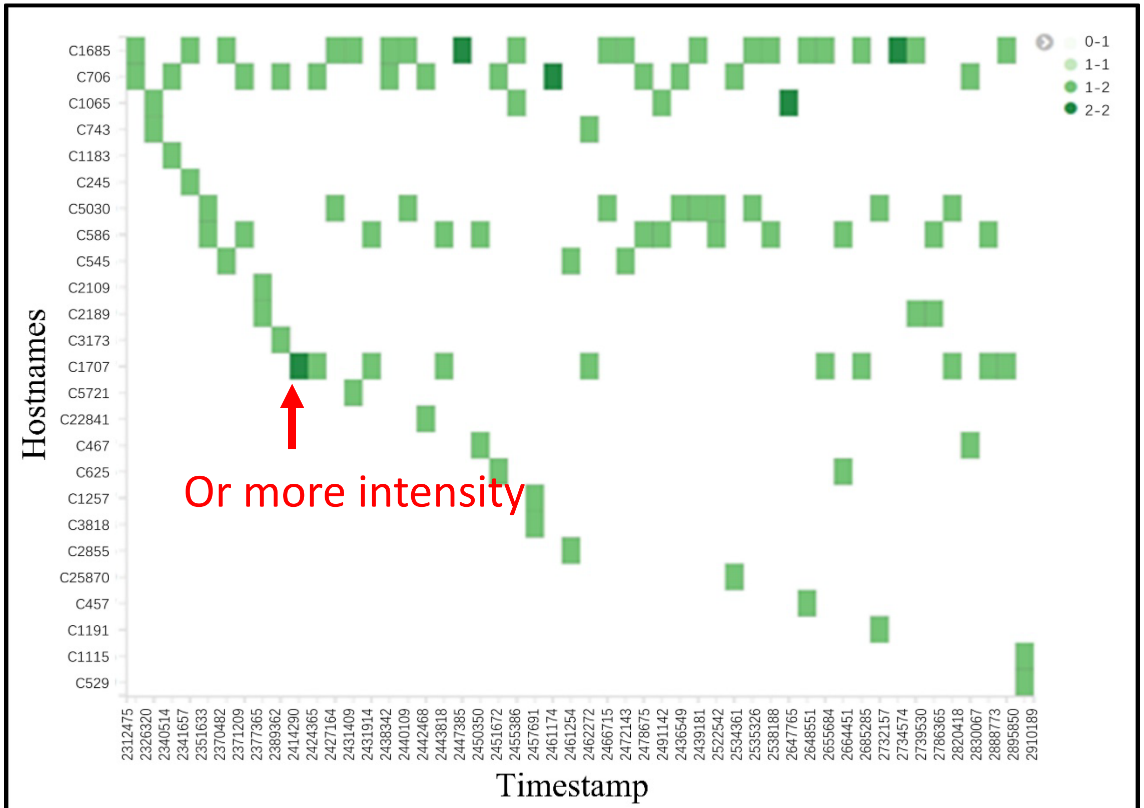


Anomalies

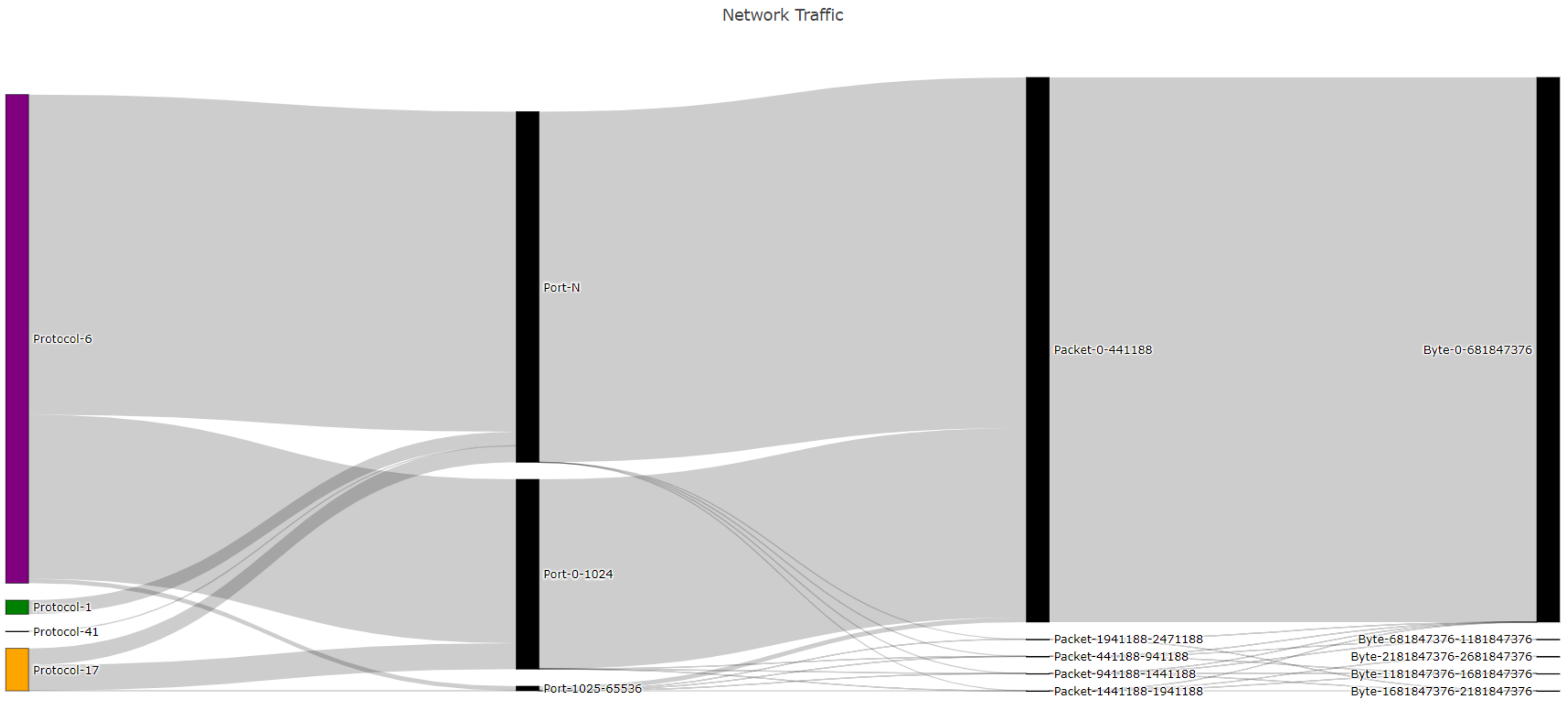
- The Defense Information Systems Agency (DISA)
- Cyber Situational Awareness Analytic Capabilities (CSAAC)
- Amazon Web Services (AWS) big data tools
  - Apache Spark Apache Storm Hadoop Map/Reduce, Kibana
  - NodeJS
  - R-Shiny

Heat Maps

Used for a longer time



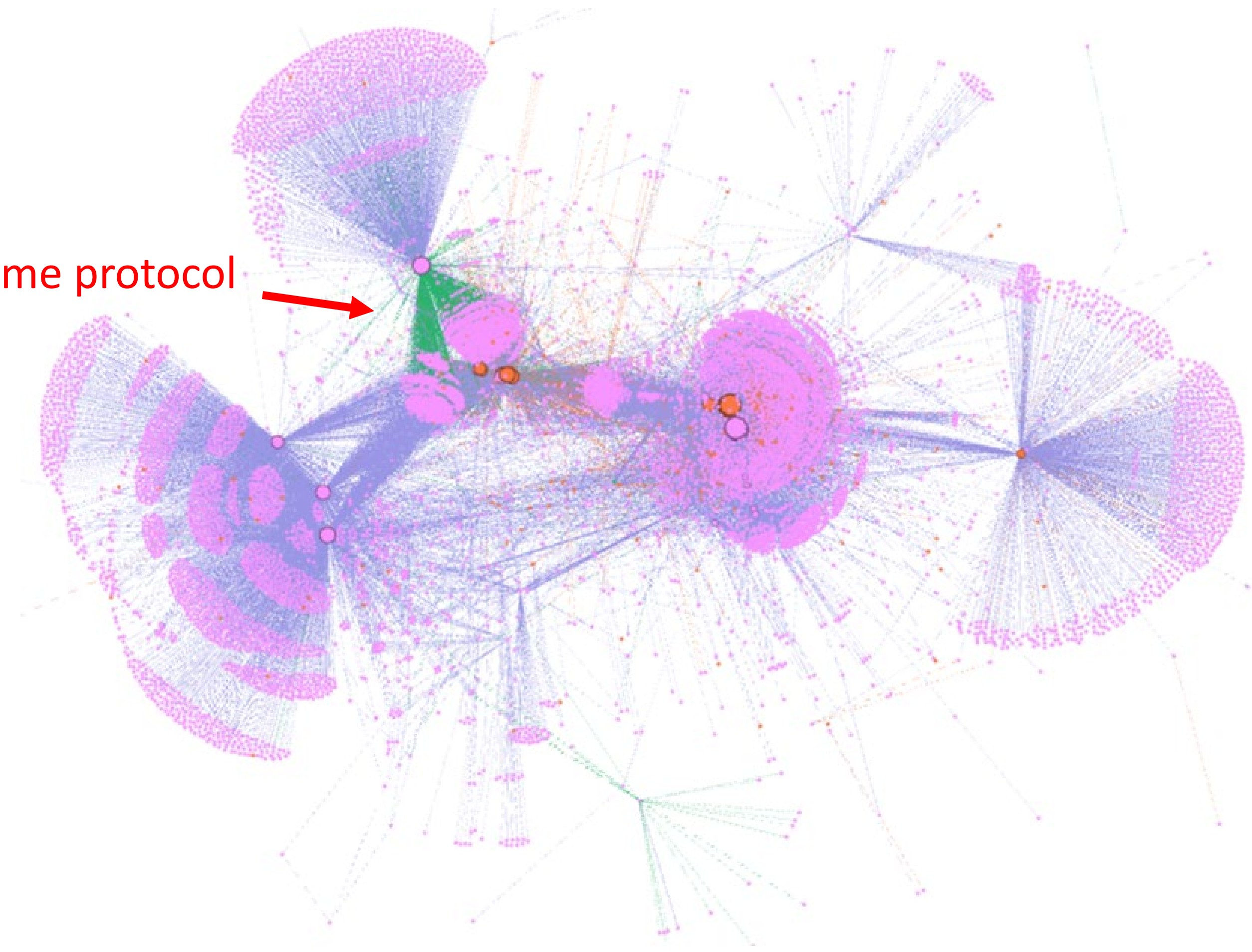
Sankey Network View



Anomalies Anomalies Anomalies

Gephi Network Visualization

Some protocol

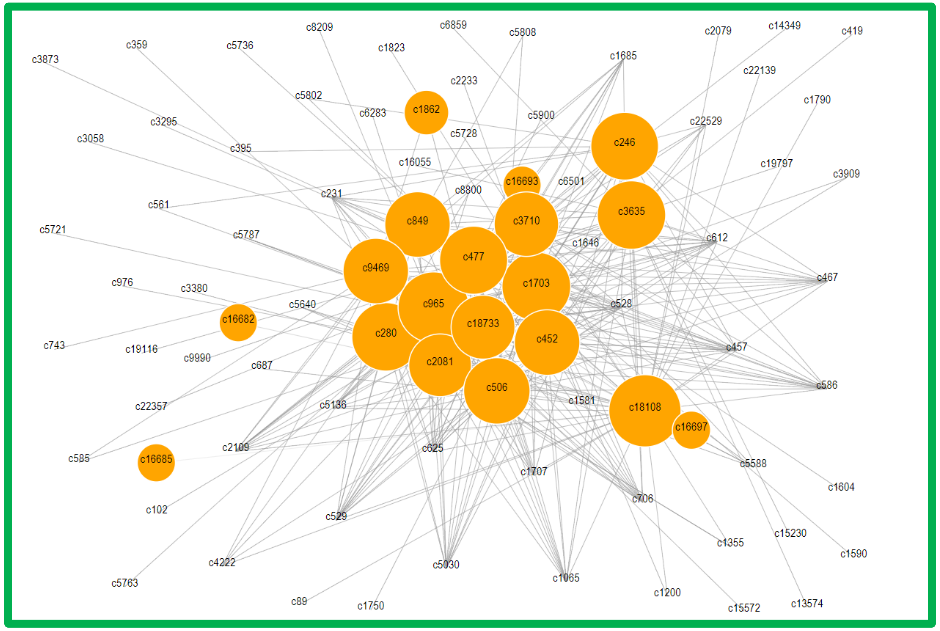


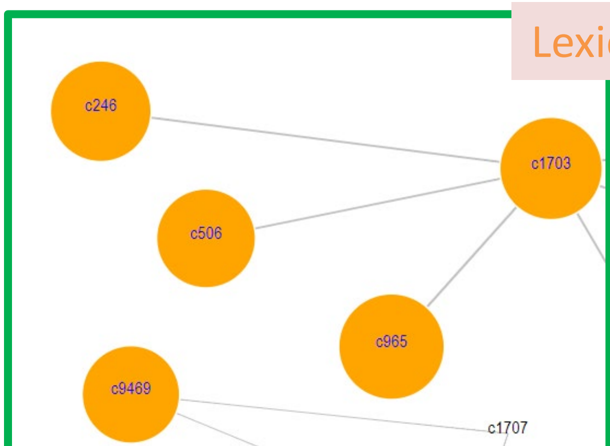
- Nodes are hacking or hacked
- Green connections are “protocol-1,” which relate to the hacked computers

Challenges

- Difficult for multi-dimensional analysis
- Difficult for classification and prediction using ML/AI methods

Features and Derived Features for ML





Lexical Link Analysis

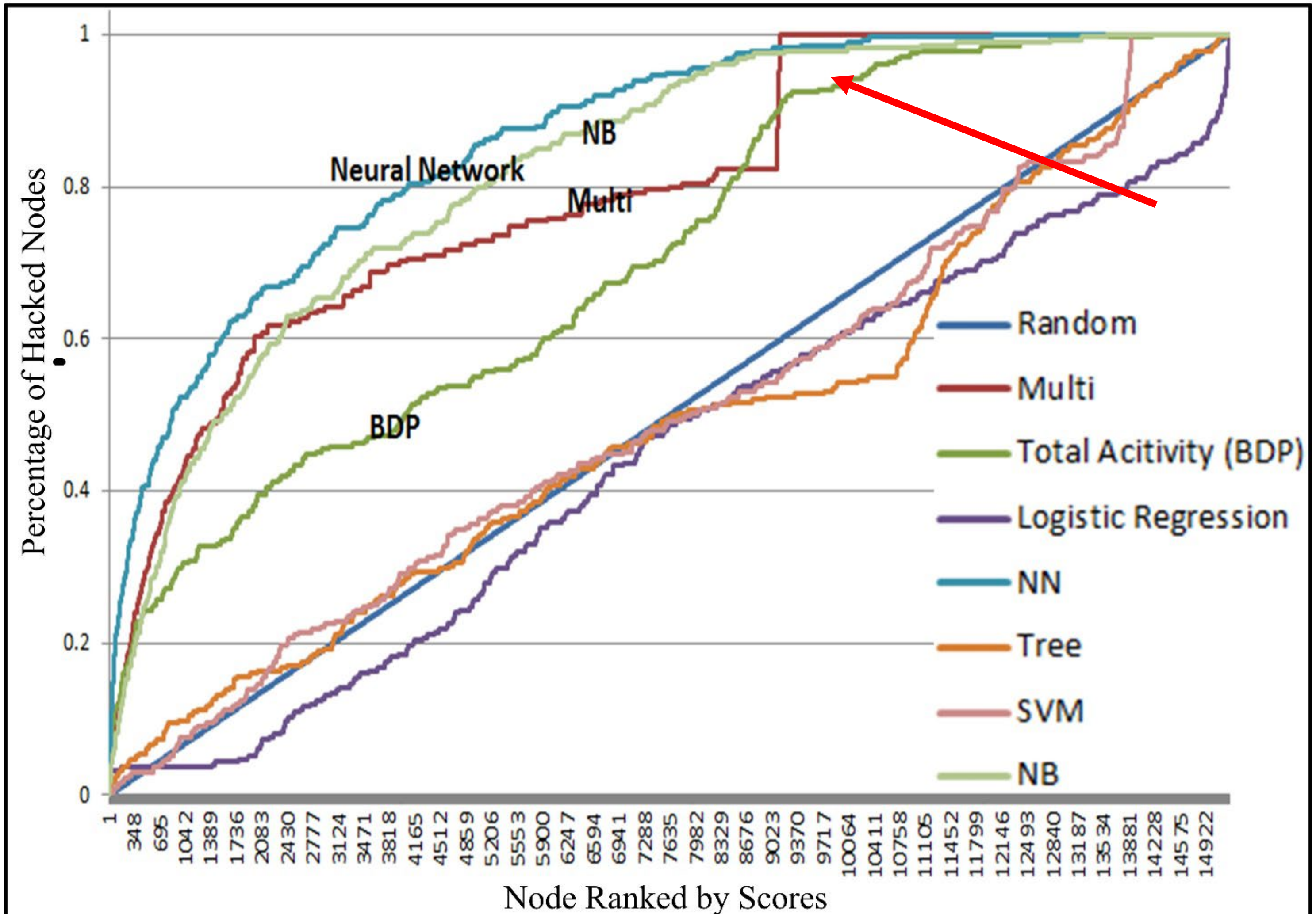
computer ID
hacked/hacking or not
degree
betweenness
degree in
degree out
degree in*degree out
number of unique processes • number of total Processes
total number of destinations
total number of authorization
total number of successful logon
number of authorization types
number of logon types
number of orientations
number of connections
number of source ports
number of destination ports
total duration of connections
total packets of connections
total bytes of connections

ML: Neural network (NB)
Nearest neighbor (NB), decision tree, logistic regression, support vector machine

Human's Causal Inference

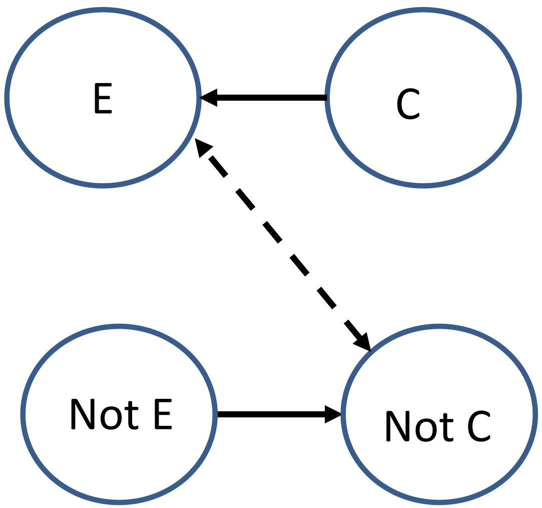
- If a computer is hacked or hacking other computers, its activity, which can be measured in various ways, e.g., total activity from BDP, has to increase.
- If a computer is a normal domain name server, it should not request any name lookups to other computers.
- If a computer is a normal computer, it should not perform any name lookups from other computers.

Results Compare with ML Methods: Gains Chart

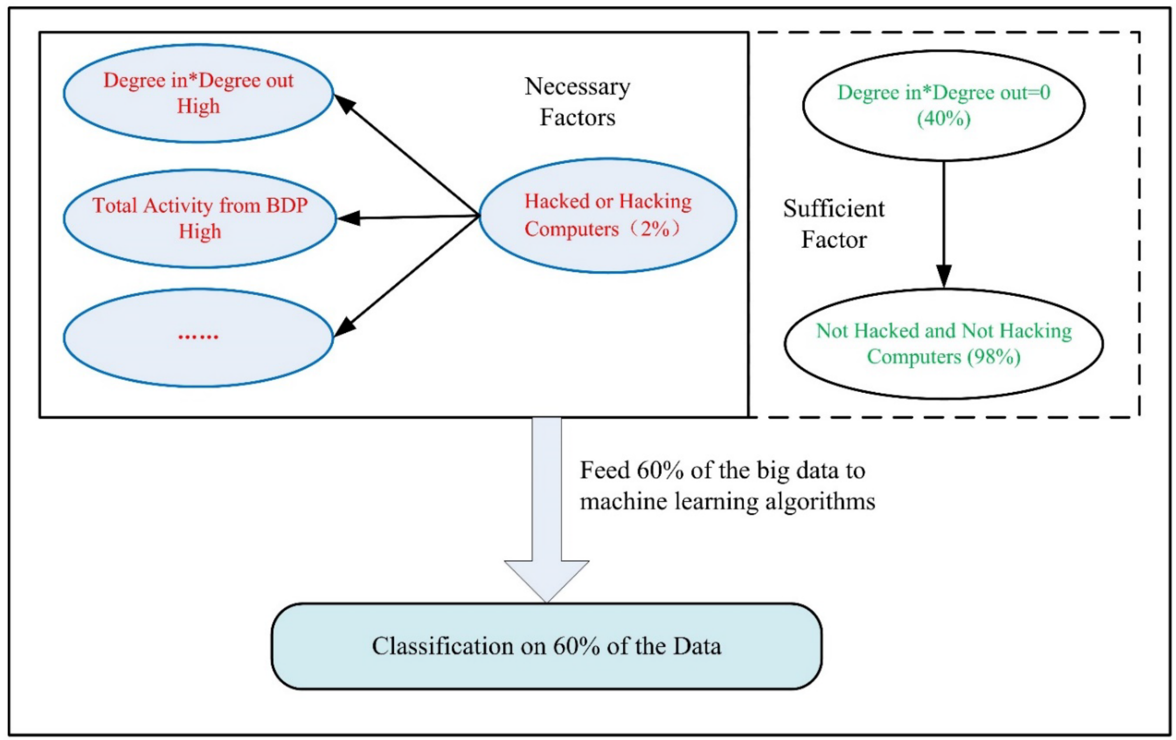


Causal Inference cut 40% data Multi=degree_in* degree_out

Human’s Causal Inference



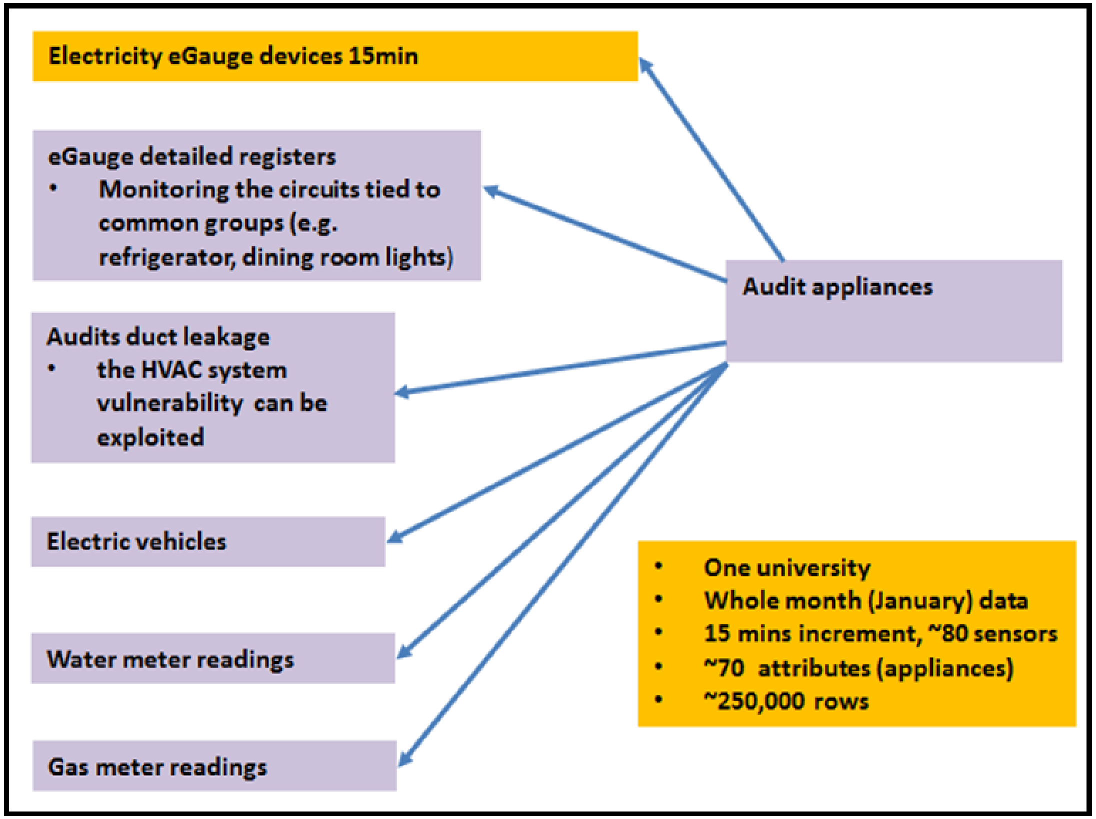
Counterfactuals
Max P[E|do(C)] –P[E|do(not C)]



- **Reduce big data and focus on smaller areas**

Use Case 2: Deep Analytics for Management and Cybersecurity of the National Energy Grid

(https://link.springer.com/chapter/10.1007/978-3-030-50426-7_23)



The Pecan Street organization [12]. Pecan Street collects energy usage for a smart city

- a conscious and curated effort to record the right data for energy consumption in a methodical manner.
- 750 million records are collected daily as circuit-level electricity usage data (per kWh) with 67 fields listing various equipment used on site (e.g., furnace, kitchen, lights, dish washer, dryer, etc.).
- One month of data consisting of 250,000 records in 15 min data blocks for 100 participants (users or data ids) as follows:

– air1: air conditioner 1

– air2: air conditioner 2

– air3: air conditioner 3

– aquarium1: aquarium 1

– bathroom1: bathroom 1

– bathroom2: bathroom 2

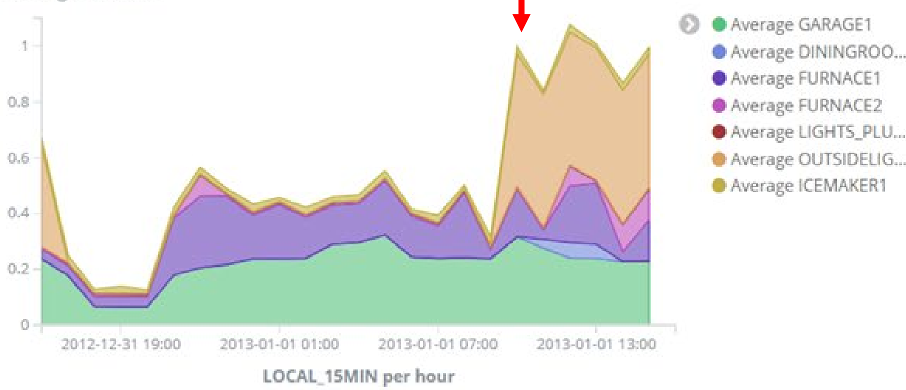
– bedroom1: bedroom 1 **–** ...

- **Anomaly detection: E.g., unauthorized running of energy grid servers in January from the air conditioner usage**.

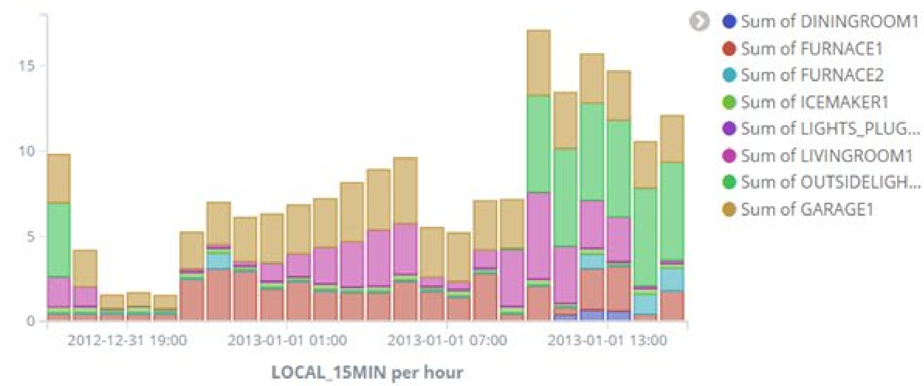
BDP Visualization

Specific time and area

AverageElecArea



Bar



New Visualization

Overall Average of Sum of GARAGE1	Lower Standard Deviation of GARAGE1	Upper Standard Deviation of GARAGE1	Min GARAGE1	Max GARAGE1
2.66	0.077	0.367	0.065	0.342

metrics

0.135
Average OUTSIDELIGHTS_PLUGS1

0
Min OUTSIDELIGHTS_PLUGS1

0.489
Max OUTSIDELIGHTS_PLUGS1

-0.294
Lower Standard Deviation of OUTSIDELIGHTS_PLUGS1

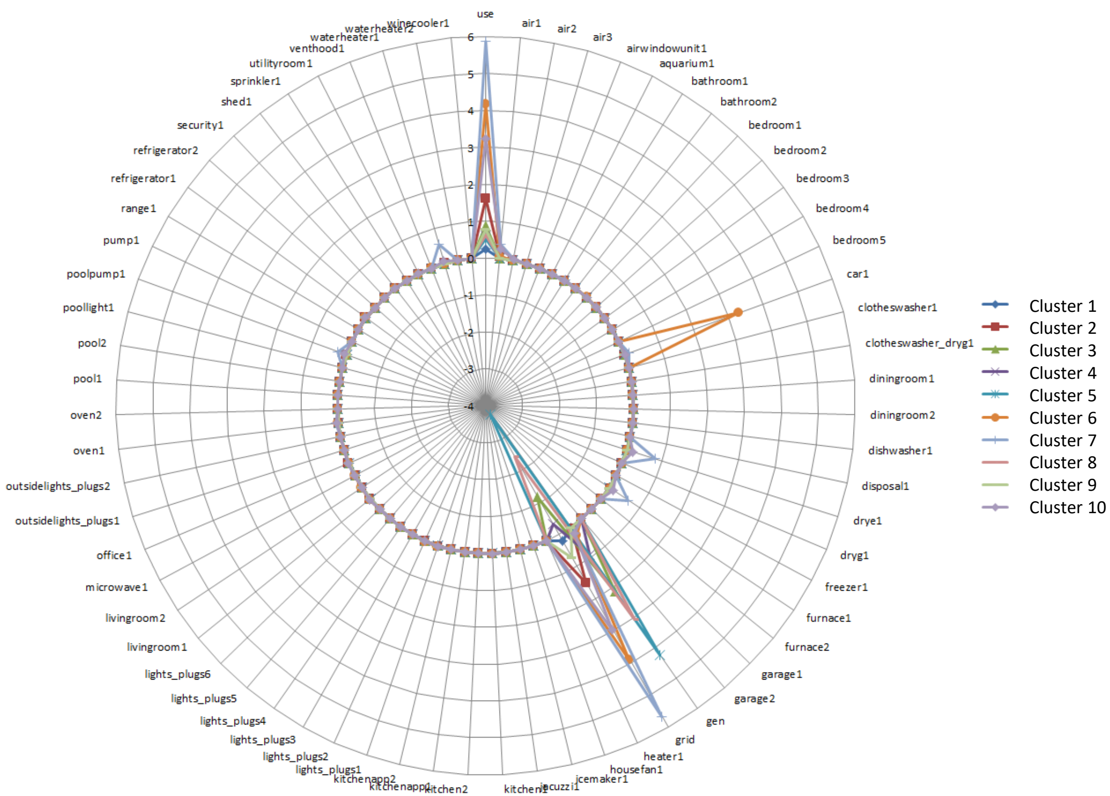
0.564
Upper Standard Deviation of OUTSIDELIGHTS_PLUGS1

0
25th percentile of OUTSIDELIGHTS_PLUGS1

0
50th percentile of OUTSIDELIGHTS_PLUGS1

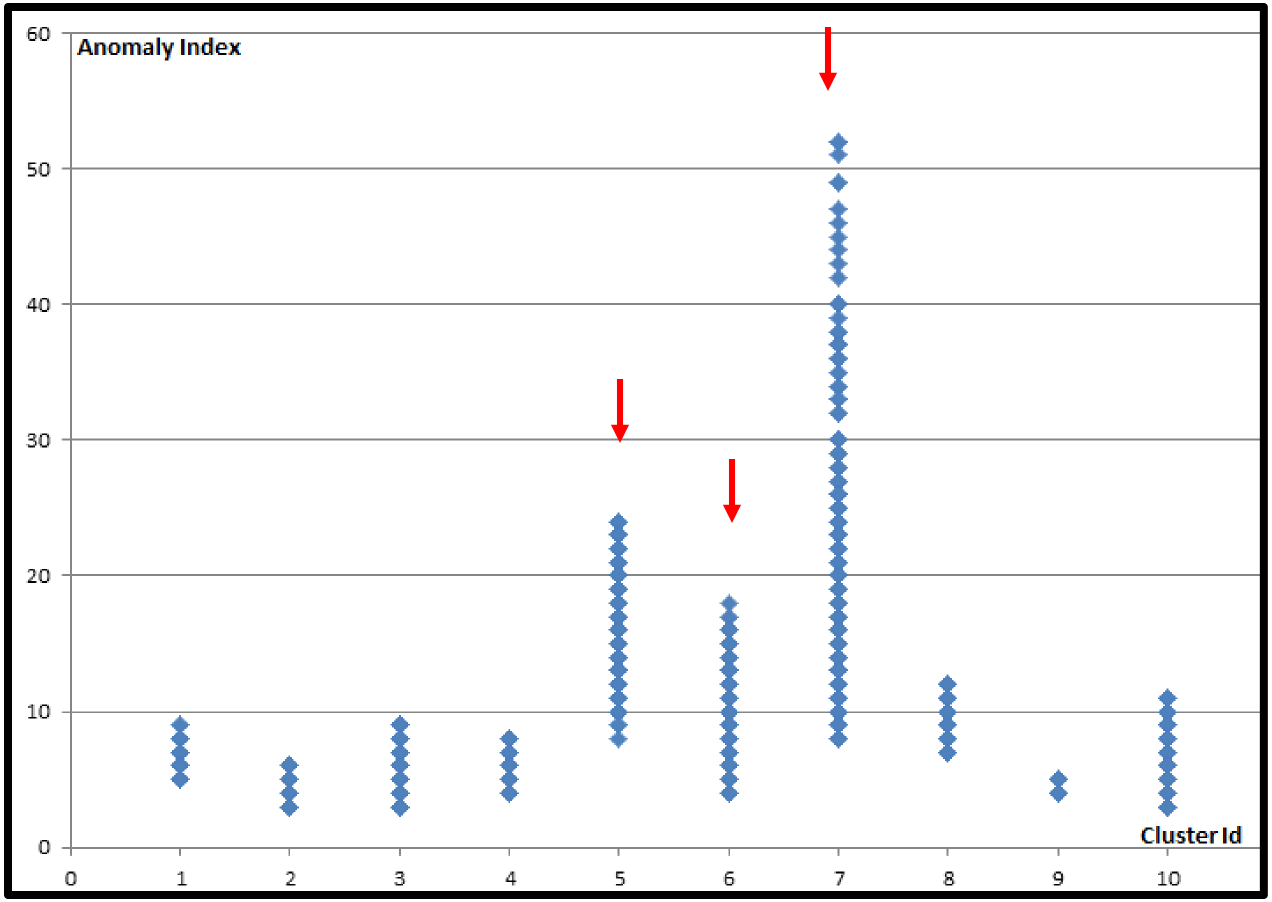
0.474
75th percentile of OUTSIDELIGHTS_PLUGS1

Unsupervised ML: Kmeans Clustering



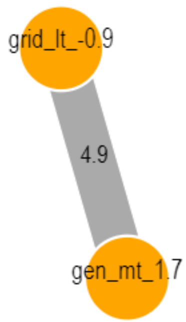
Cluster 7: Average high usages within the cluster attribute to the areas of “use”, “grid”,”drye1”, “furnace1”, “poollight1”,and “waterheater1”.
Cluster 6: Average high usages attribute to the areas of “use”, “car1”, “gen”, and “grid”.
Cluster 5: Average high usages attribute to the areas of “gen” and “grid” (negative – giving back to the grid).

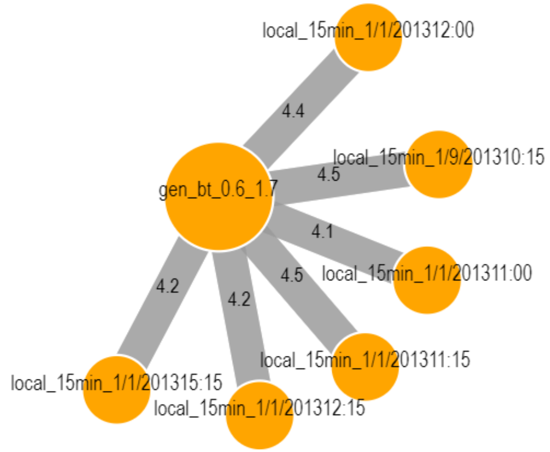
Anomaly Index



- The distances to the 10 cluster centers

Examples

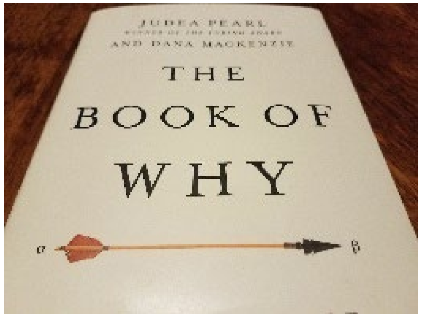




- "gen" means there is a generator at home and negative "grid" means the generator gives energy back to the grid

Causal Inference

- Causes and effects
  - Effects are often observable data, e.g., "total activity from BDP high" or "degree in*degree out high"
  - Causes: hacked or not
    - E= Total Activity from BDP High
    - C= Hacked or Hacking Computers
- Three pillars
  - Association/Correlation, posterior probability or maximum likelihood
    - P(C|E)=P(Hacked or Hacking Computers|Total Activity from BDP High)
  - Intervention
    - P(E|C), ensure C is actionable or P(E|do(C))
  - Counterfactuals
    - What if I had acted differently?
    - Compare P(E|C) and P(E|Not C)



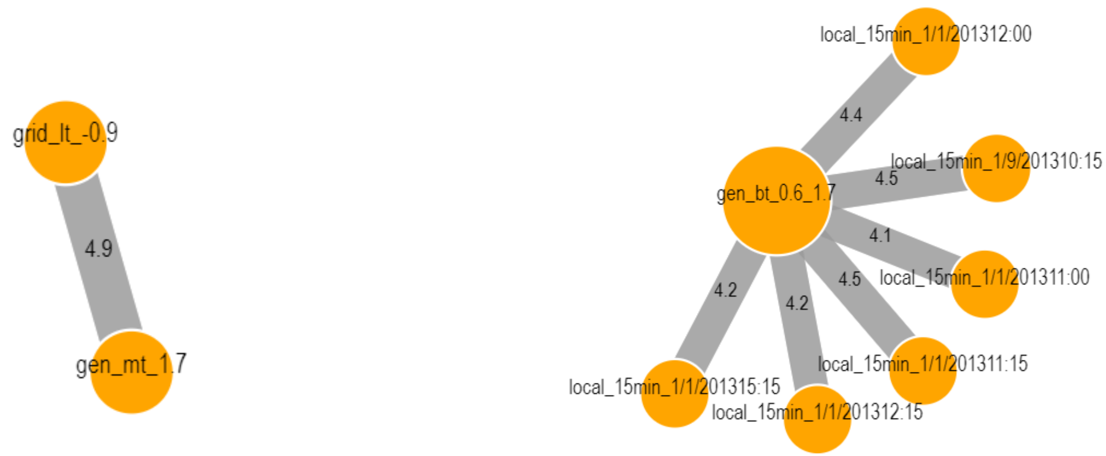
Judea Pearl (2018)

Causal Learning

Theme Id	Theme Keywords
6(P)	disposal1_bt_0.0_0.1,microwave1_bt_0.0_0.1,kitchenapp1_bt_0.0_0.1
7(P)	lights_plugs1_bt_-0.0_0.1 lights_plugs2_bt_-0.0_0.1,range1_bt_0.0_0.2
10(E)	car1_bt_-0.6_0.7,dryg1_bt_-0.0_0.1,bathroom1_bt_0.0_0.1
14(E)	livingroom1_bt_-0.0_0.1 security1_mt_0.0,kitchenapp2_bt_0.0_0.1 livingroom1_bt_-0.0_0.1
4(E)	kitchen1_bt_-0.1_0.1,oven1_bt_0.0_0.3,grid_mt_1.8
2(A)	bedroom1_bt_0.0_0.1 bedroom2_bt_-0.0_0.1,bedroom2_bt_-0.0_0.1 lights_plugs1_mt_0.2
11(A)	livingroom2_bt_-0.0_0.1 outsidelights_plugs1_bt_0.0_0.1,livingroom1_bt_0.1_0.3
13(A)	clotheswasher_dryg1_bt_-0.0_0.1 waterheater1_bt_-0.4_0.8,clotheswasher_dryg1_bt_-0.0_0.1 refrigerator1_mt_0.2,refrigerator1_mt_0.2 waterheater1_bt_-0.4_0.8
12(A)	bedroom4_bt_0.0_0.1 bedroom5_bt_0.0_0.1,bedroom5_bt_0.0_0.1 office1_bt_0.1_0.3,bedroom4_bt_0.0_0.1 office1_bt_0.1_0.3
8(A)	local_15min_1/1/201312:30 gen_bt_0.6_1.7,local_15min_1/1/201311:15 gen_bt_0.6_1.7,local_15min_1/1/201312:15 gen_bt_0.6_1.7,local_15min_1/1/201311:00 gen_bt_0.6_1.7,local_15min_1/1/201312:00 gen_bt_0.6_1.7,local_15min_1/9/201310:15 gen_bt_0.6_1.7,local_15min_1/1/201315:15 gen_bt_0.6_1.7

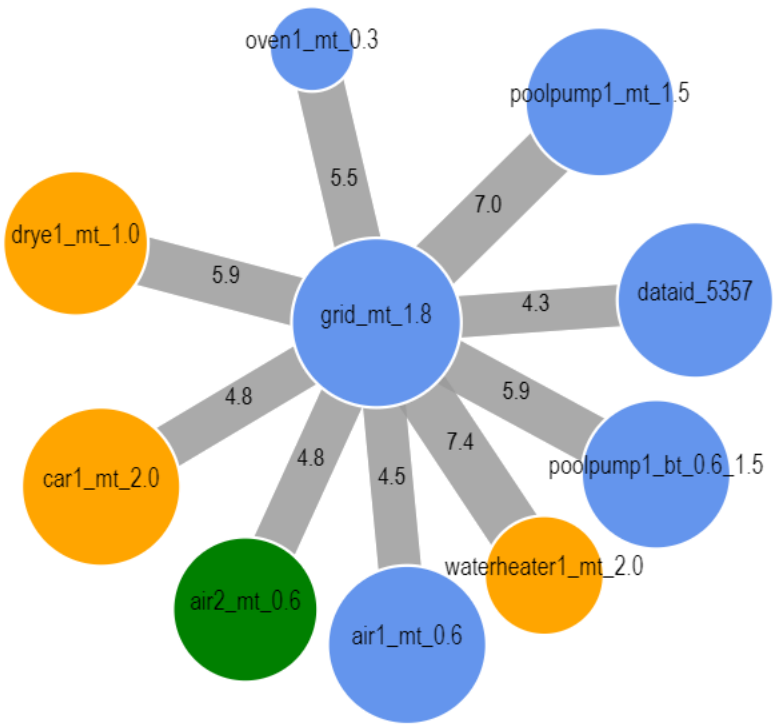
- Themes and causal links discovered by lexical link analysis
- Human analysts validate the causal relations

Examples



- “gen” means there is a generator at home and negative “grid” means the generator gives energy back to the grid

Causal Level 1



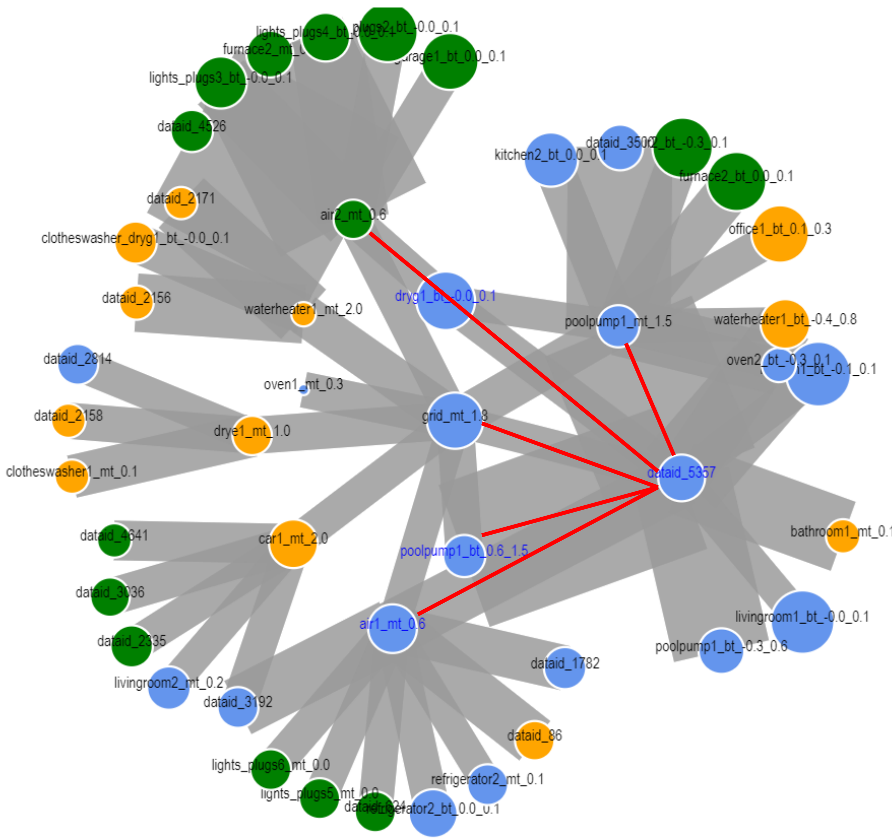
Link weights computed using Lift

$$prob_{ij} = \frac{word\ features\ i, j\ together}{word\ feature\ j} \quad (1)$$

$$prob_i = \frac{word\ feature\ i}{all\ word\ features} \quad (2)$$

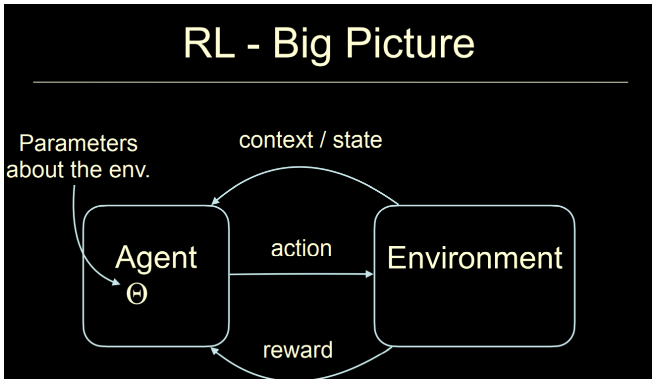
$$lift_{ij} = \frac{prob_{ij}}{prob_i} \quad (3)$$

Causal Level 2



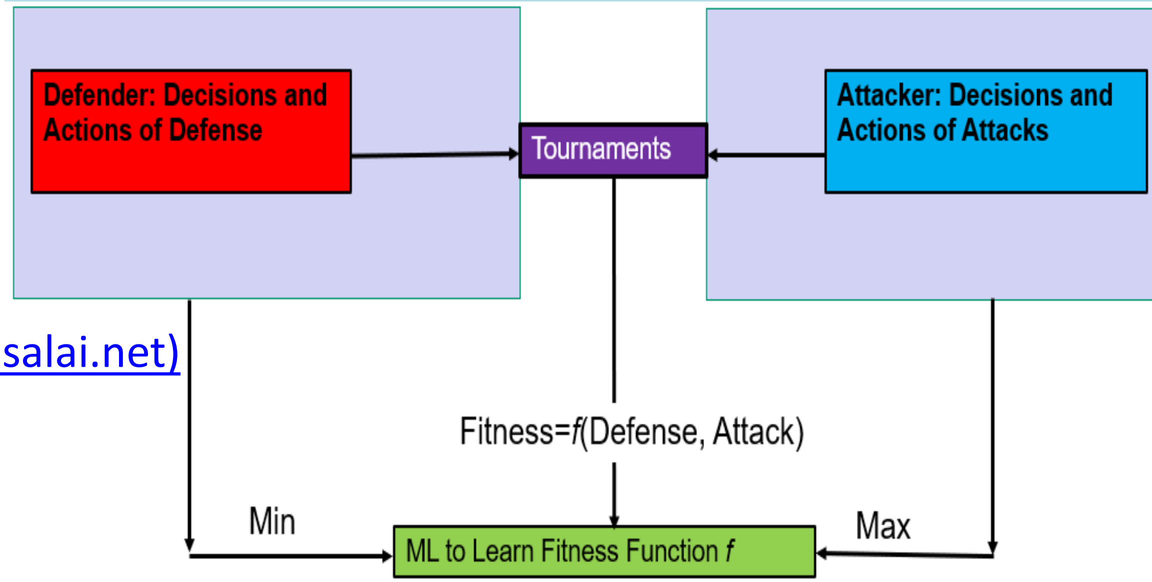
- Only *dataid* 5357 directly links
  - *grid mt* 1.8 and the first level features
  - *poolpump*1 *mt* 1.5,
  - *poolpump*1 *bt* 0.6 1.5,
  - *air*1 *mt* 0.6, and *air*12 *mt* 0.6.
- *dataid* 5357 is a real cause
- *waterheater*1 *mt* 2.0, *drye*1 *mt* 1.0, *car*1 *mt* 2.0, and *oven*1 *mt* 0.3 may be independent causes with no confounders.

Importance of Causal Learning and Human Knowledge in Future Data Sciences and Wargaming



[Causal Reinforcement Learning (causalai.net)]

Need Casual Graphs for Defenders



Opponent agent

- Case 1: Environmental (neutral)
- Case 2: Strategic complementary factors
- Case 3: Strategic competitive factors

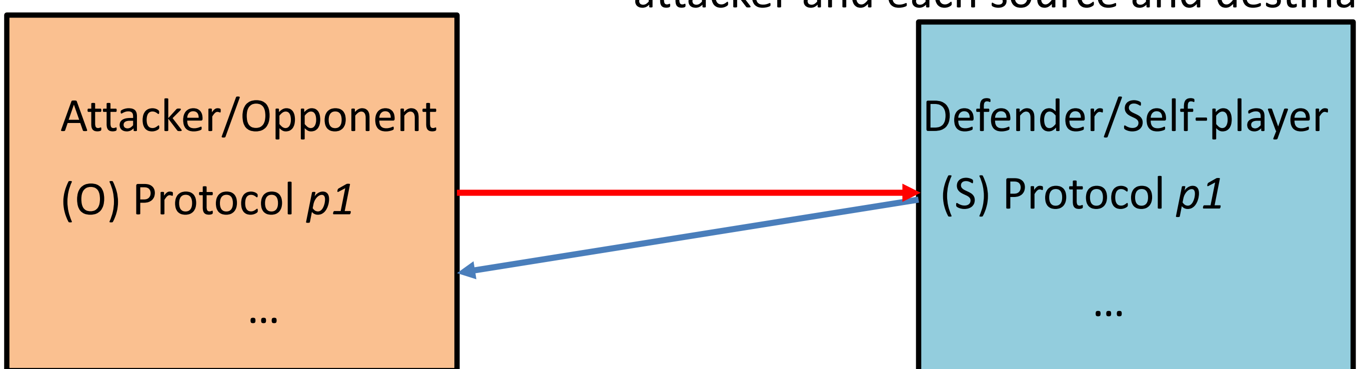
- Causal Inference
- Adversarial Patch
- Control
- Deception

Use Case 3: Threat and Capability Coevolutionary Wargame (TCCW) Applied to Advanced Persistent Threats, funded by OUSD(R&E) as part of Cyber Agreements for Resilient Machines through Augmented AI (CARMA-AI) Project
(Presented at the Naval Annual Machine Learning (NAML) Conference 2022)

Objective: What are the characteristics of effective decoys? How can ML/AI methods inform configuration of more effective decoys?
Initial Data: Network traffic generated during cyber deception experimentation with human attackers and decoy systems

Attacker ID	Source IP	Destination IP	Packet Count	Protocol	Timestamps
…	…	…	…	…	…

Transformation → "**Tournament**:" number of protocols source to destination and destination to source for each attacker and each source and destination



Conclusions

- Human reasons and knowledge
  - Provide explainable automation and new information
    - Visualization
    - Interface
  - Reduce big data
  - Validate causal relations discovery
  - Speed up and guide search, and perform defense and control more effectively
- Should one incorporate complex ontologies into ML/AI algorithms?
  - Discover the “sweet spots” of exploration (machine intelligence) and exploitation (causal inference including human knowledge)

Acknowledgments and Disclaimer

I would like to thank

- the Naval Postgraduate School (NPS)'s Naval Research Program (NRP) for supporting the research
- the Office of Naval Research (ONR)'s Naval Enterprise Partnership Teaming with Universities for National Excellence (NEPTUNE 2.0) program

The views presented are those of the authors and do not necessarily represent the views of the U.S. Government, Department of Defense (DoD), or their Components.

THANK YOU!