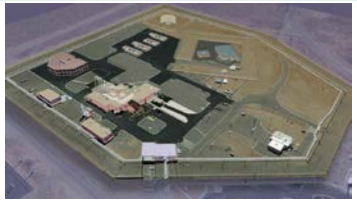
A New Approach to Insider Threat Detection & Mitigation for High Consequence Facilities & Critical Infrastructure :

Artificial Neural Networks & Risk Significance



Adam Williams, Shannon Abbott, Christopher Faucett, Colton Heffington, & Sondra Spence (Sandia)

Presented by: Alan Evans

William Charlton (University of Texas)

Katherine Holt (NNSA's Office of International Nuclear Security)

Introduction to the Role of Artificial Intelligence in Strengthening the Security of Nuclear Facilities| February 6-8, 2024 | Vienna, Austria

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2023-14192C

Roadmap

Introduction

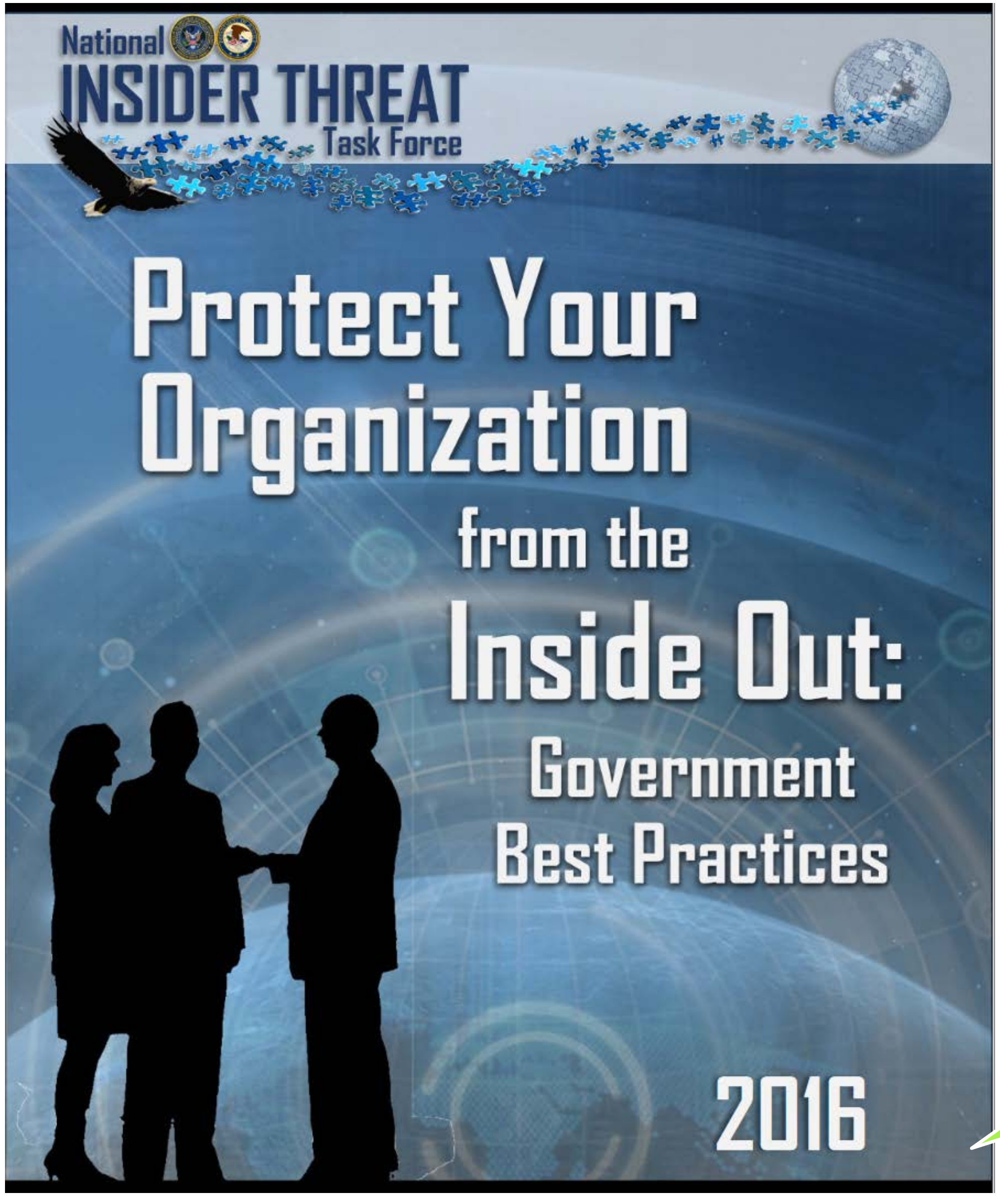
Situating a New Approach to ITDM

Methods & Data Collection

Demonstrating a New Approach to ITDM

Conclusions, Insights & Implications

Introduction



National INSIDER THREAT Task Force

Protect Your Organization from the Inside Out: Government Best Practices

2016

"the **risk** [that] an insider will use their **authorized access**, wittingly or unwittingly, **to do harm** to their organization. This can include theft of proprietary information and technology; damage to company facilities, systems or equipment; actual or threatened harm to employees; or other actions **that would prevent the company** from carrying out its **normal business practices**."

Introduction

~~OFFICIAL USE ONLY—SECURITY-RELATED INFORMATION~~

U.S. NUCLEAR REGULATORY COMMISSION
OFFICE OF NUCLEAR REGULATORY RESEARCH

March 2009
Division 5

REGULATORY GUIDE

REGULATORY GUIDE 5.77
(Draft was issued as DG-5011, dated December 2008)
(New Regulatory Guide)

INSIDER MITIGATION PROGRAM

A. (U) INTRODUCTION

(U) This guide describes an approach that the staff of the U.S. Nuclear Regulatory Commission (NRC) considers acceptable for an insider mitigation program (IMP) at nuclear power reactor facilities. Title 10 of the *Code of Federal Regulations* (10 CFR) Section 73.55, "Requirements for Physical Protection of Licensed Activities in Nuclear Power Reactors against Radiological Sabotage," specifically paragraph (b)(7) states that licensees shall establish, maintain, and follow an access authorization program in accordance with 10 CFR 73.56, "Personnel Access Authorization Requirements for Nuclear Power Plants." The licensee's physical security plan must include descriptions of the access authorization program and the IMP. Furthermore, pursuant to 10 CFR 73.55(b)(9), licensees shall design and implement the IMP to oversee and monitor the initial and continuing trustworthiness and reliability of individuals granted unescorted access or retaining unescorted access authorization to a protected or vital areas. The IMP should use defense-in-depth methodologies to minimize the potential for an insider to adversely affect, either directly or indirectly, the licensee's capability to prevent significant core damage or spent fuel sabotage.

(U) This document provides guidance for an IMP that would meet the requirements in 10 CFR 73.55(b)(7) and (b)(9) and the latest NRC staff endorsed version of the industry's guidance document, Nuclear Energy Institute (NEI) 03-01, "Nuclear Power Plant Access Authorization Program." These sources provide an acceptable approach for an IMP that meets the provisions of 10 CFR 73.55 as part of the licensee's physical security plan. These sources are also consistent with the guidance described in this regulatory guide.

(U) The NRC issues regulatory guides to describe and make available the methods that the NRC staff considers acceptable for use in implementing specific parts of the agency's regulations, techniques that the staff uses in evaluating specific problems or postulated accidents, and data that the staff needs in reviewing applications for permits and licenses. Regulatory guides are not substitutes for regulations, and compliance with them is not required. Methods and solutions that differ from those set forth in regulatory guides will be deemed acceptable if they provide a basis for the findings required for the issuance or continuance of a permit or license by the Commission.

(U) This guide was issued after consideration of comments received from stakeholders.

~~OFFICIAL USE ONLY—SECURITY-RELATED INFORMATION~~

IAEA Nuclear Security Series No. 8-G (Rev. 1)

Implementing Guide

Preventive and Protective Measures against Insider Threats

IAEA
International Atomic Energy Agency

Insider Threat Mitigation Guide

NOVEMBER 2020

Cybersecurity and Infrastructure Security Agency

Insider threat definitions:

- ***NRC*** → "Once an individual has been granted unescorted access to protected and vital areas ... preventing an adverse event becomes dependent on detecting ... and/or denying ... the ***opportunity to commit*** the act"

- ***IAEA*** → "an individual with authorized access to [nuclear material,] associated facilities or associated activities or to sensitive information or sensitive information assets, who ***could commit, or facilitate the commission*** of criminal or intentional unauthorized acts ... [with] an adverse impact on nuclear security"

- ***DHS/CISA*** → "is the ***potential*** for an insider to use their authorized access or special understanding of an organization to harm that organization"

Introduction

~~OFFICIAL USE ONLY—SECURITY-RELATED INFORMATION~~

U.S. NUCLEAR REGULATORY COMMISSION
OFFICE OF NUCLEAR REGULATORY RESEARCH

March 2009
Division 5

REGULATORY GUIDE

REGULATORY GUIDE 5.77
(Draft was issued as DG-5011, dated December 2008)
(New Regulatory Guide)

INSIDER MITIGATION PROGRAM

A. (U) INTRODUCTION

(U) This guide describes an approach that the staff of the U.S. Nuclear Regulatory Commission (NRC) considers acceptable for an insider mitigation program (IMP) at nuclear power reactor facilities. Title 10 of the *Code of Federal Regulations* (10 CFR) Section 73.55, "Requirements for Physical Protection of Licensed Activities in Nuclear Power Reactors against Radiological Sabotage," specifically paragraph (b)(7) states that licensees shall establish, maintain, and follow an access authorization program in accordance with 10 CFR 73.56, "Personnel Access Authorization Requirements for Nuclear Power Plants." The licensee's physical security plan must include descriptions of the access authorization program and the IMP. Furthermore, pursuant to 10 CFR 73.55(b)(9), licensees shall design and implement the IMP to oversee and monitor the initial and continuing trustworthiness and reliability of individuals granted unescorted access or retaining unescorted access authorization to a protected or vital areas. The IMP should use defense-in-depth methodologies to minimize the potential for an insider to adversely affect, either directly or indirectly, the licensee's capability to prevent significant core damage or spent fuel sabotage.

(U) This document provides guidance for an IMP that would meet the requirements in 10 CFR 73.55(b)(7) and (b)(9) and the latest NRC staff endorsed version of the industry's guidance document, Nuclear Energy Institute (NEI) 03-01, "Nuclear Power Plant Access Authorization Program." These sources provide an acceptable approach for an IMP that meets the provisions of 10 CFR 73.55 as part of the licensee's physical security plan. These sources are also consistent with the guidance described in this regulatory guide.

(U) The NRC issues regulatory guides to describe and make available the methods that the NRC staff considers acceptable for use in implementing specific parts of the agency's regulations, techniques that the staff uses in evaluating specific problems or postulated accidents, and data that the staff needs in reviewing applications for permits and licenses. Regulatory guides are not substitutes for regulations, and compliance with them is not required. Methods and solutions that differ from those set forth in regulatory guides will be deemed acceptable if they provide a basis for the findings required for the issuance or continuance of a permit or license by the Commission.

(U) This guide was issued after consideration of comments received from stakeholders.

~~OFFICIAL USE ONLY—SECURITY-RELATED INFORMATION~~

IAEA Nuclear Security Series No. 8-G (Rev. 1)

Implementing Guide

Preventive and Protective Measures against Insider Threats

IAEA
International Atomic Energy Agency

Insider Threat Mitigation Guide

NOVEMBER 2020

Cybersecurity and Infrastructure Security Agency

Insider threat definitions:

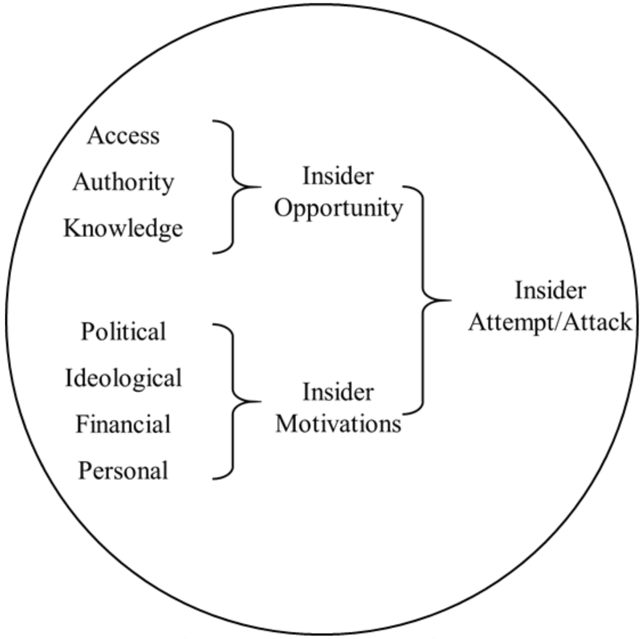
- ***NRC*** → "Once an individual has been granted unescorted access to protected and vital areas … preventing an adverse event becomes dependent on detecting … and/or denying … the ***opportunity to commit*** the act"

- ***IAEA*** → "an individual with authorized access to [nuclear material,] associated facilities or associated activities or to sensitive information or sensitive information assets, who ***could commit, or facilitate the commission*** of criminal or intentional unauthorized acts … [with] an adverse impact on nuclear security"

- ***DHS/CISA*** → "is the ***potential*** for an insider to use their authorized access or special understanding of an organization to harm that organization"

"opportunity" or "could" or "potential" → risk significance

Situating: Operational Patterns & Workplace Rhythms

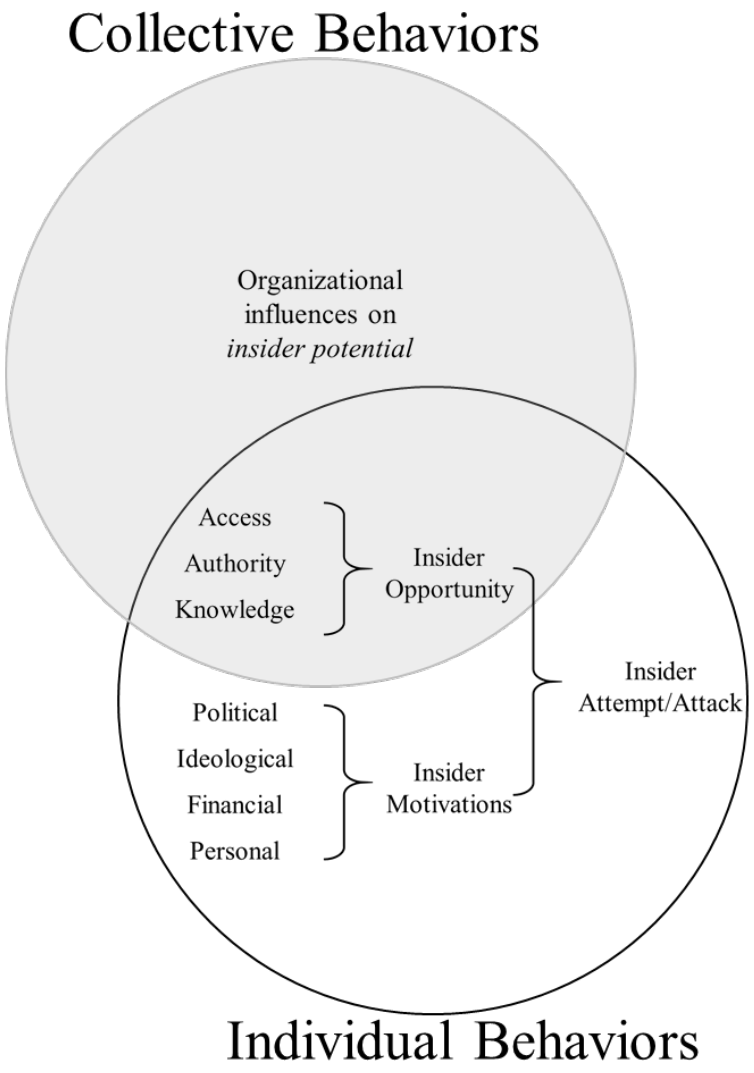


Individual Behaviors

Traditional approaches to Insider Threat Detection & Mitigation (ITDM)

- Focus on individual characteristics
  - Difficult to identify, almost impossible to measure/quantify
- Based on “prevention” and “protection” concepts
  - Best practices, for example
- Struggle to anticipate growing “insider threat potential”
  - Underlying “reactionary” paradigm

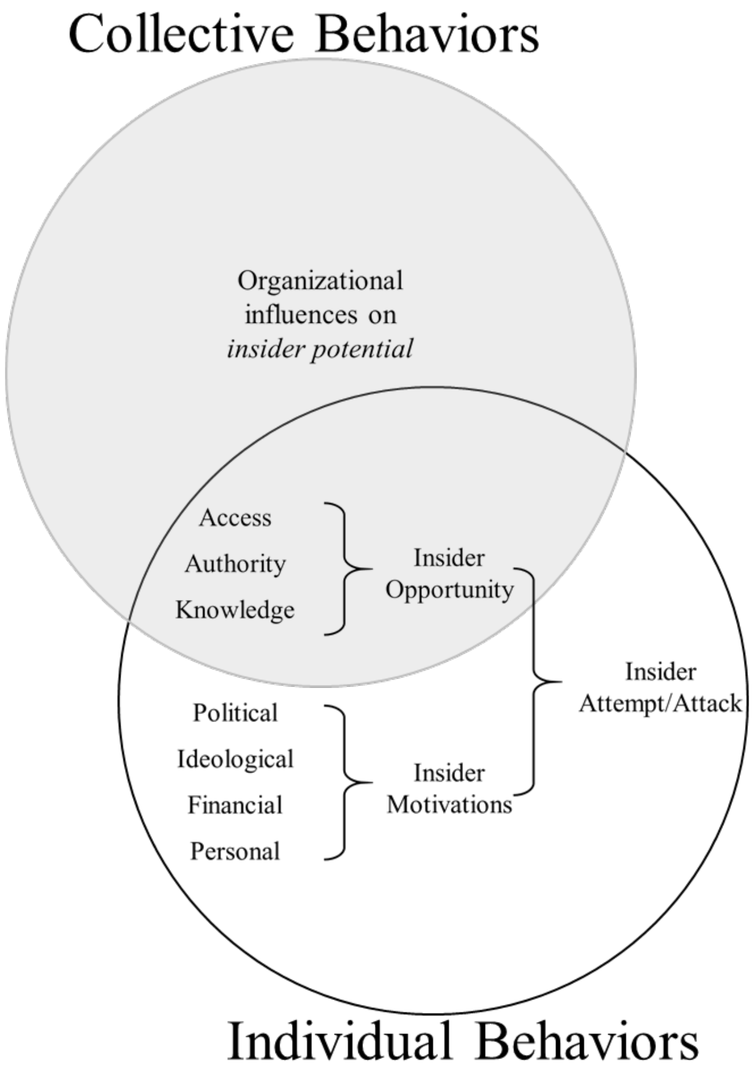
Situating: Operational Patterns & Workplace Rhythms

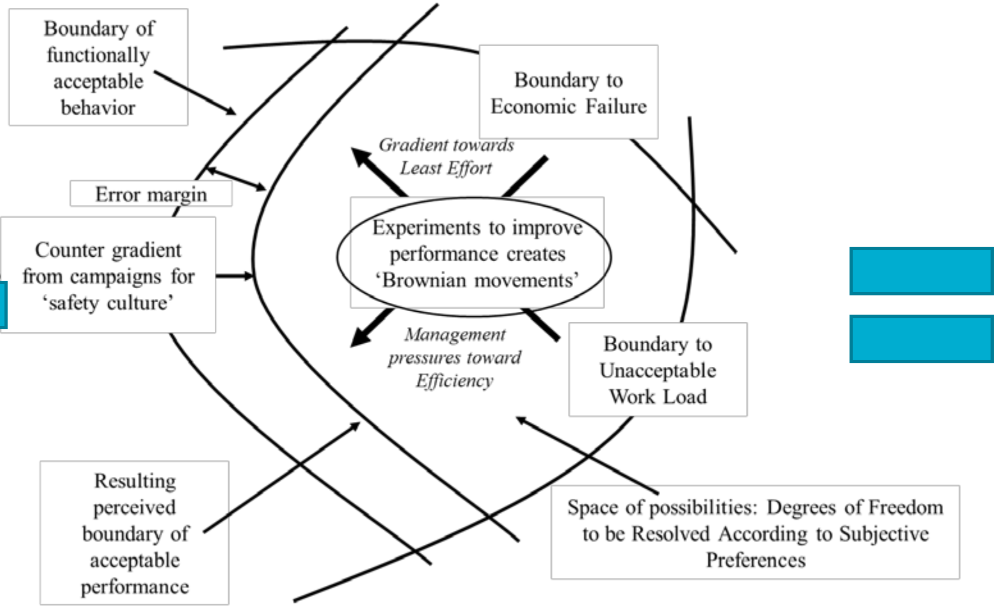


A ***new approach*** for potential improvement, based on several observations:

- People working in nuclear facilities settle into “operational rhythms”
- These rhythms can be described with data/signals already being collected at nuclear facilities
- Recast “preventive” & “protective” approaches as boundaries on these rhythms

Situating: Operational Patterns & Workplace Rhythms




Boundary of functionally acceptable behavior
Boundary to Economic Failure
Error margin
Gradient towards Least Effort
Counter gradient from campaigns for 'safety culture'
Experiments to improve performance creates 'Brownian movements'
Management pressures toward Efficiency
Boundary to Unacceptable Work Load
Resulting perceived boundary of acceptable performance
Space of possibilities: Degrees of Freedom to be Resolved According to Subjective Preferences

A new approach :

- "workplace rhythms"
- data/signals already being collected
- Recast approaches as boundaries on these rhythms

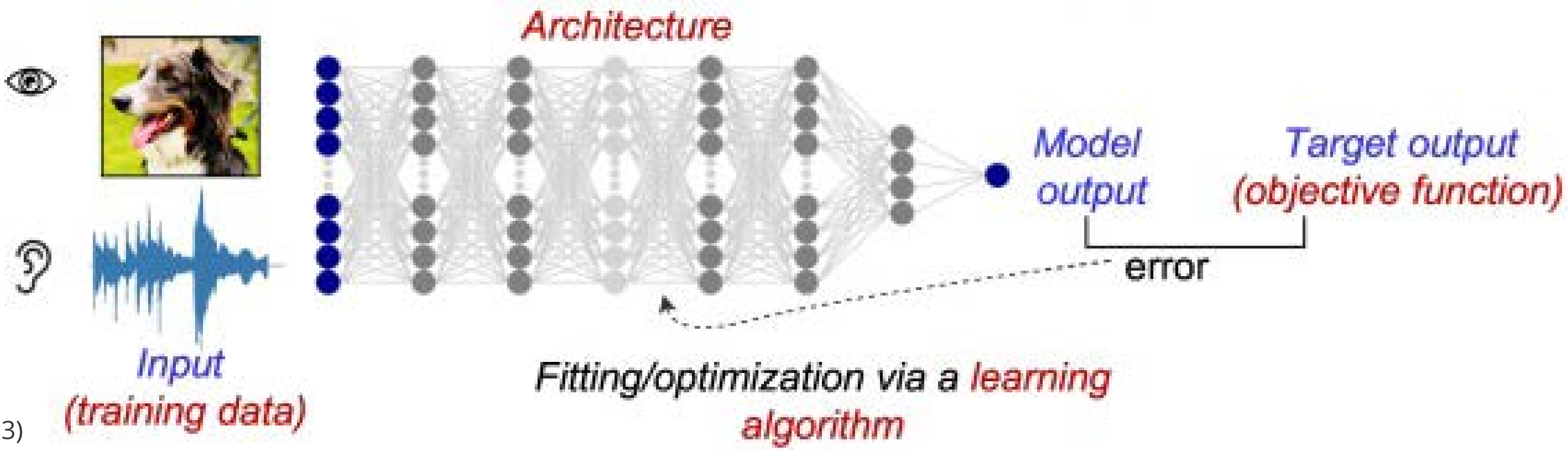
Situating: Artificial Neural Networks

Assumption:

- Insider threat ***attempts*** represent a deviation from these "operational rhythms"

Conclusion:

- Humans are ***creatures of habit*** & ***unpredictable*** – can deviation from normal rhythms ID insiders?
- Anomaly detection ***may*** identify the ***potential*** for an insider opportunity to manifest into action
- ***Artificial neural networks*** (ANNs) can be trained to ID patterns/deviations in operational rhythms

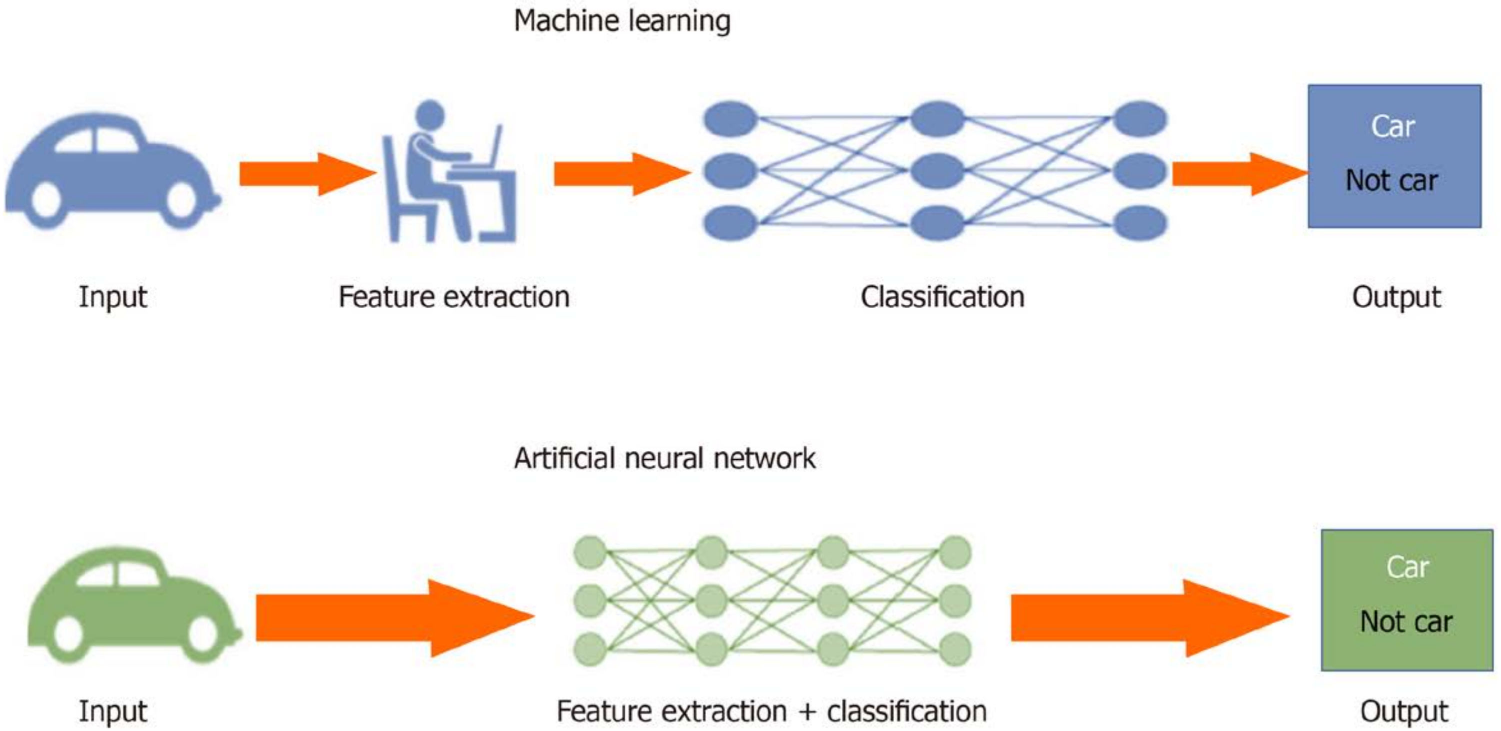

Architecture
Model output
Target output (objective function)
error
Input (training data)
Fitting/optimization via a learning algorithm

Courtesy: (Kanwisher, et. al 2023)

Situating: Artificial Neural Networks

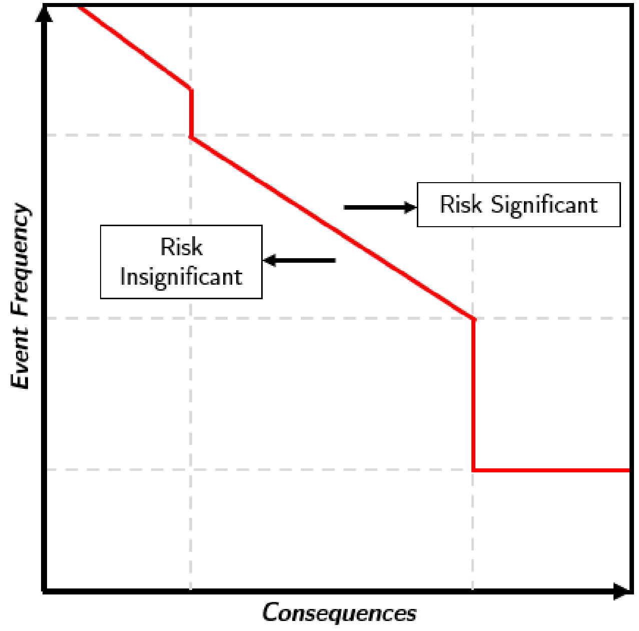
Hypothesis: ANNs can evaluate facility data signals to support ITDM

- Unusual access times as monitored by access control points like badge readers
- Attempts to access physical areas beyond current access level as monitored by access control points
- Increased or routine alarms from personnel radiation portal monitors



Courtesy: (Bao, et. al 2020)

Situating: Risk Significance



Borrowing the concept of ***risk significance*** from nuclear safety:

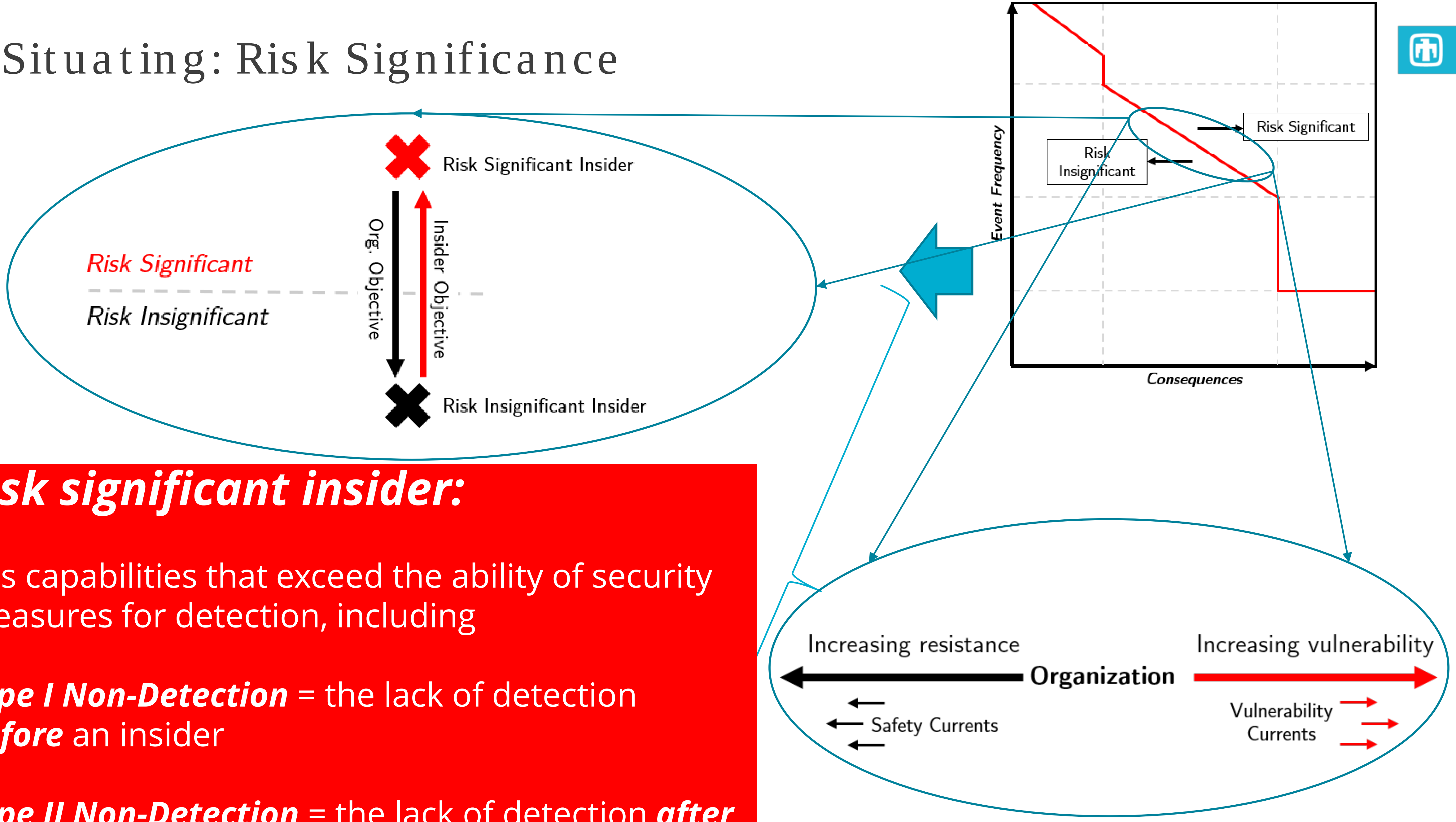
- Risk significance → does an accident sequence exceed a predetermined risk limit?
  - *f* (event frequency, consequences)
- If yes, then those accidents are risk significant

Therefore, ***risk significance for an insider*** considers:

- Best described as a time-variant continuous variable
- Related to the ability to successfully execute an act
- Both individual & facility characteristics
  - Ex: Individuals conduct business according to the access & authority (sometimes knowledge) bestowed by the facility

Workplace rhythms

Situating: Risk Significance



A ***risk significant insider:***

- has capabilities that exceed the ability of security measures for detection, including

- ***Type I Non-Detection*** = the lack of detection ***before*** an insider

- ***Type II Non-Detection*** = the lack of detection ***after*** an insider act

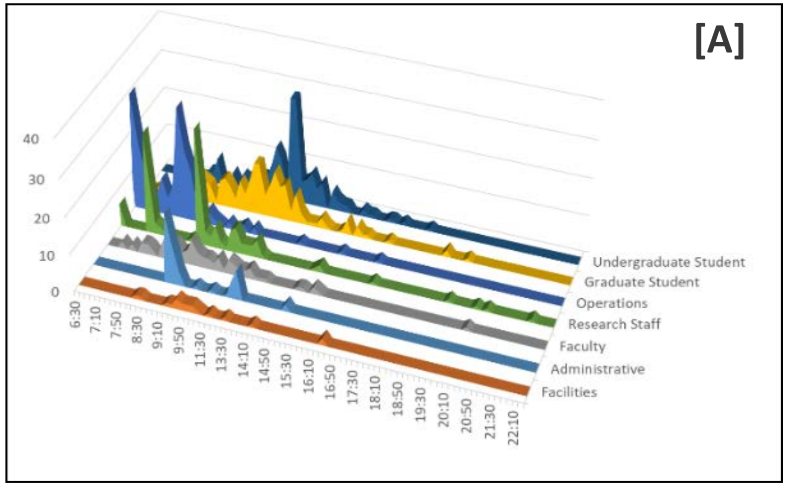
Methods & Data Collection

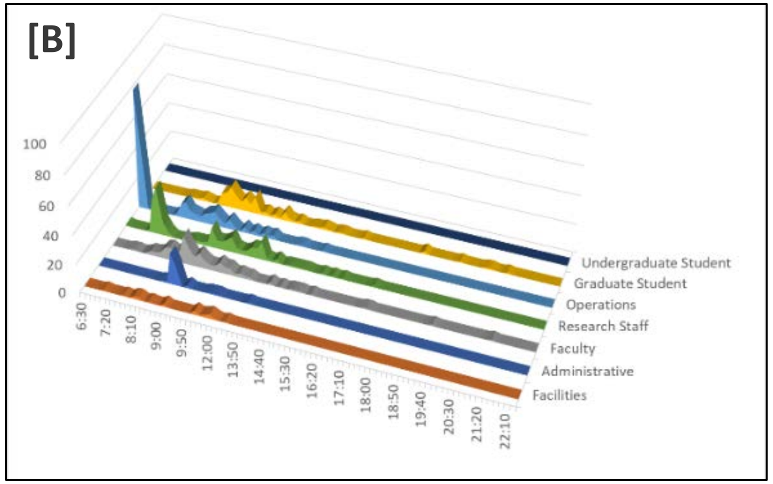
	Description	Implication
Single access point (SAP)	All access control data was organized by sensor location in the facility, date and time of allowed access, and then by identity used for access	Allowed for observation of patterns of accesses in time including bounds for when particular accesses are expected to occur for all individuals as well as for specific individuals
Time-sequenced, multiple access points (TS/MAP)	All access control data was organized by identity used for access, by date and time of allowed access, and then by location in the facility	Allowed for observation of patterns of access by individuals including bounds for when particular individuals would be expected to complete a sequence of access to different locations
Time of access by personnel type	All access control data was organized by access point, date and time of allowed access and then by grouping the identity used for access into a personnel type	Allowed for observation of pattern differences between personnel groups: Facilities, Administrative, Faculty, Research Staff, Operations, Graduate Student, Undergraduate Student

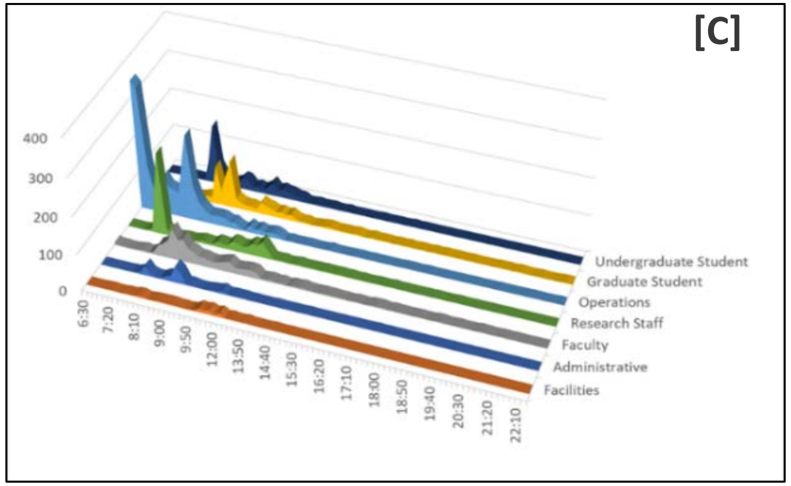
Type	Sensor Type	Data Type	Representative Activity
Access Control	• Badge reader ▪ ORG B entry ▪ Security control panel ▪ Limited area ▪ Reactor control room	• Badge readers: ▪ # authorized attempts ▪ # unauthorized attempts (false negative + false positives) ▪ Time of access attempts	• Personnel arrival to facility • Researchers approaching the reactor • Reactor operator arriving for shift
Intrusion Detection	• Balanced magnetic switch ▪ Limited area ▪ Security control panel ▪ Reactor control room	• Balanced magnetic switches: ▪ # times switch opened ▪ Time at which switch opens	• Researchers approaching the reactor • Maintenance of security control panel • Reactor operator arriving for shift
	• Area motion sensor ▪ Reactor bay ▪ Fuel storage surveillance	• Area motion sensors: ▪ # times change in physical phenomena registered ▪ Time at which change in physical phenomena registered	• Custodial services around the reactor • Transfer of fresh/used fuel into/out of ORG B

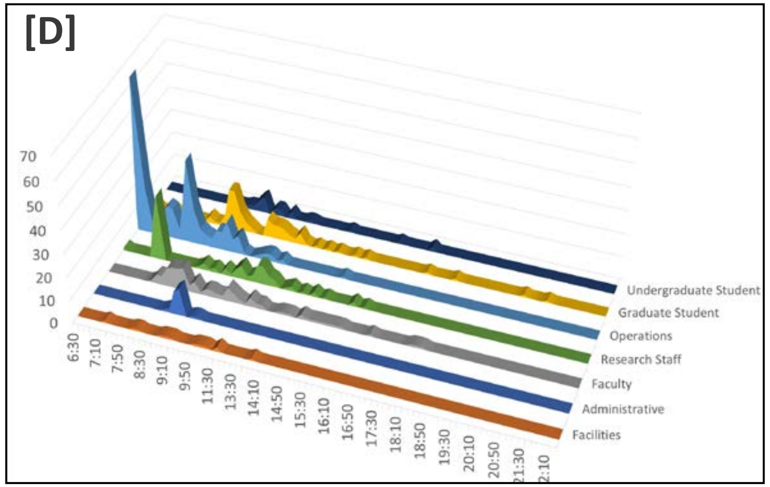
Data Characteristic	Data Set I	Data Set II	Data Set III	Data Set IV
ANN Solution	Tool 1	Tool 1	Tool 1	Tool 2
Date range	10/12/2019 to 03/14/2020	03/15/2020 to 09/25/2020	09/26/2020 to 03/31/2022	03/15/2023 to 09/15/2023
Access control data points	13,653	18,986	74,922	27,653
Intrusion detection data points	694	923	4211	1102
Categories for organizing data points[a]	SAP TSMAP	SAP TSMAP	SAP TSMAP	SAP TSMAP

Demonstrating a New Approach: SAP Frequency







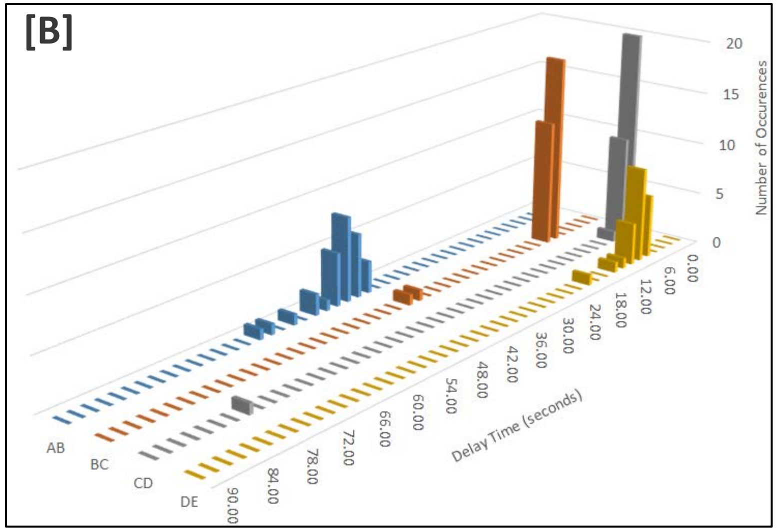


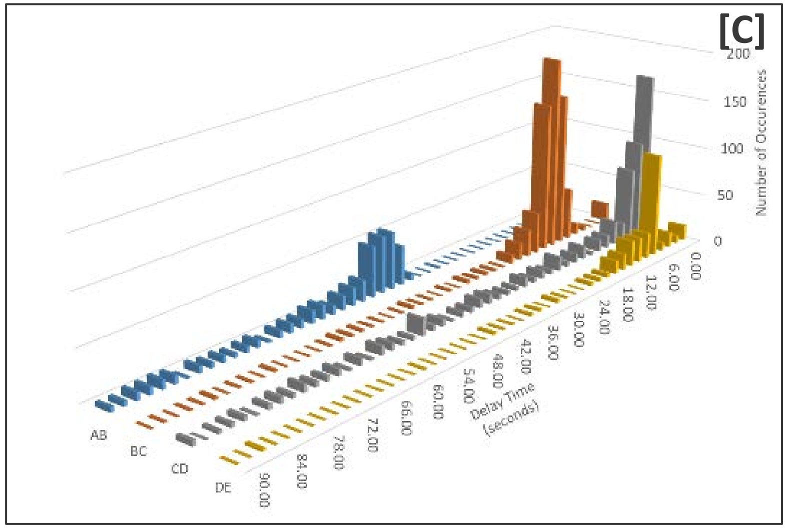
- Somewhat surprising level of regularity
- Time bounds → baseline patterns for ANN
- Key Results:
  - collected data signals can reflect patterns and rhythms in behaviors
  - common patterns and rhythms can form profiles associated with particular personnel categories
  - such personnel category profiles can be used as a baseline of expected behaviors

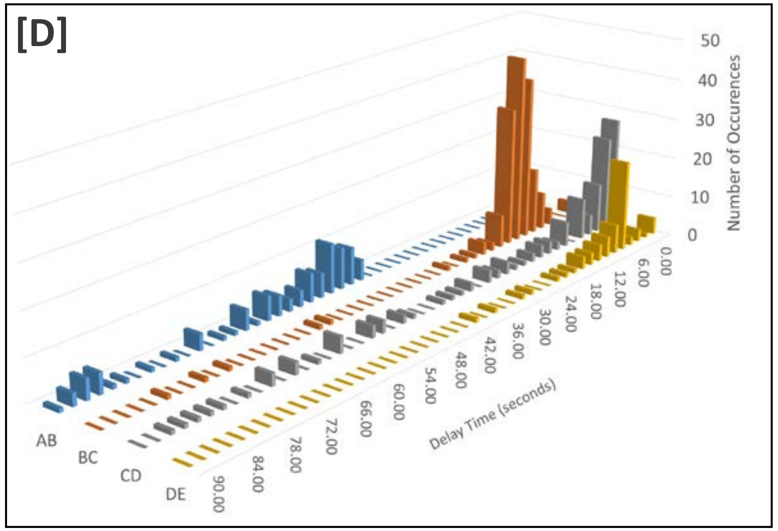
Demonstrating a New Approach: TSMAP Frequency

[A]

These results were not captured for Data Set I







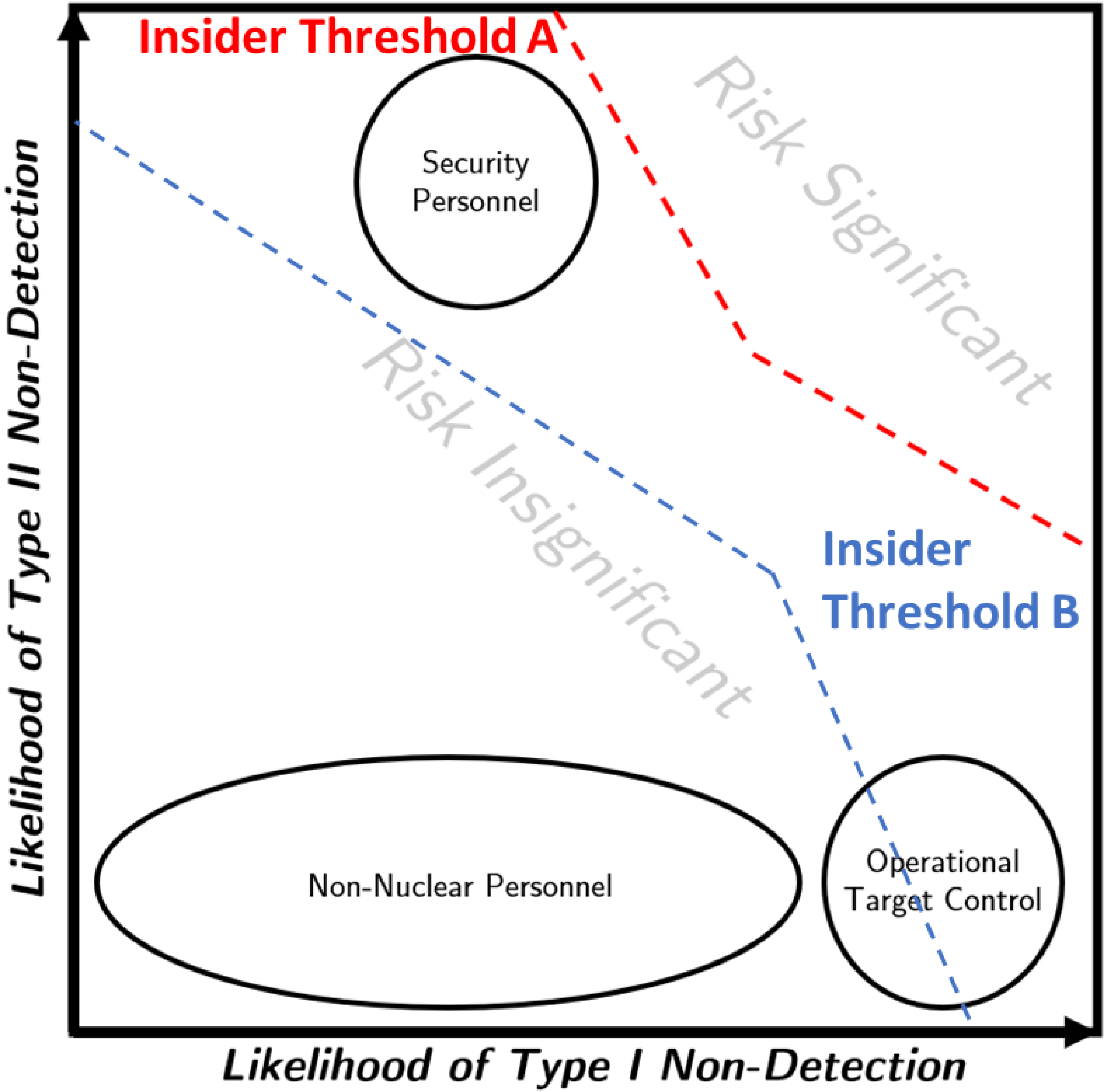
- Pathway-based patterns
  - A→B→C→D→E
- Higher fidelity + more nuanced description of patterns
  - Ex: “this individual is expected to take 42-66 seconds to move from access point A to access point B” (Data Set II)
- Key Results:
  - Higher validity & structure for anomaly detection
  - Captures dynamism of workplace rhythms

Demonstrating a New Approach: Experimental Results

Scenario Name (#)	Test Description (Scenario # & Pathway Name)	Data Set I Results*	Data Set II Results	Data Set III Results	Data Set IV Results
Security Closet Access (1)	Unauthorized Access Attempt (1A)	Detected & Denied in **ALL** Cases [SAP]	Detected & Denied in **ALL** Cases [SAP]	Detected & Denied in **ALL** Cases [SAP]	Detected & Denied in **ALL** Cases [SAP]
	Authorized Access Credentials Used by Unauthorized Individual Who Entered Building Using Their Own Credentials (1B)	Detected & Denied in **MOST** Cases [SAP; TSMAP]	Detected & Denied in **MOST** Cases [SAP; TSMAP]	Detected & Denied in **MOST** Cases [SAP; TSMAP]	Detected & Denied in **NO** Cases [SAP; TSMAP]
	Authorized Access Credentials Used by Unauthorized Individual Who Entered Building Using Authorized Individual's Credentials (1C)	Detected & Denies in **NO** Cases [TSMAP]	Detected & Denies in **NO** Cases [TSMAP]	Detected & Denies in **MOST** Cases [TSMAP]	Detected & Denied in **MOST** Cases [SAP; TSMAP]
Reactor Bay Access (2)	Unauthorized Access to Reactor Bay (2A)	Detected & Denied in **ALL** Cases [TSMAP]	Detected & Denied in **ALL** Cases [TSMAP]	Detected & Denied in **ALL** Cases [TSMAP]	Detected & Denied in **ALL** Cases [TSMAP]
	Early Detection by Motion Sensor (2B)	Not Tested	Detected in **MOST** Cases	Detected in **MOST** Cases	Detected & Denied in **NO** Cases [SAP; TSMAP]
Fuel Storage Surveillance (3)	Insider Surveillance (3A)	*Difficult to Detect Without Additional Sensing Input [TSMAP]*	*Difficult to Detect Without Additional Sensing Input [TSMAP]*	*Difficult to Detect Without Additional Sensing Input [TSMAP]*	Detected & Denied in **NO** Cases [SAP; TSMAP]
	Insider Alarm Testing (3B)	Not Tested	*Difficult to Detect Without Additional Sensing Input [TSMAP]*	*Difficult to Detect Without Additional Sensing Input [TSMAP]*	Detected & Denied in **NO** Cases [SAP; TSMAP]

- Point 1
- Point 2

Demonstrating a New Approach: Risk Significance



- SAP-based or TSMAP profiles → scaffold for functionally unacceptable behaviors or thresholds
  - Or, frame for risk significant insider potential as quantified deviation from expected behaviors

- Benefits:
  - Thresholds derived from ANN-identified patterns
  - Multiple thresholds on same framework (red & blue lines)
  - Clear mapping of different personal categories
  - Provides opportunity for *anticipatory* ITDM

Conclusions, Insights & Implications

Positive results from ongoing data collection & early experiments

Empirical support for theoretical & technical approach to ITDM based on "workplace rhythms"

Shift toward "insider potential" a new, useful framing

Encourages use of facility & system-related data streams; aligns with "workplace rhythms" interpretation

Incorporating risk significance = a data-driven approach

Supports quantitative descriptions of insider potential ***not*** heavily biased with individual psychometric indicators

Incorporating risk significance = inclusive of data already being collected

Leverages wealth of data (e.g., quality assurance) + mitigates common challenges to efficacy of behavioral reporting systems

Incorporating risk significance = streamlines anomaly detection

Helps prioritize deviations in workplace rhythms, with opportunity to anticipate/categorize future deviations in workplace