January 20, 2021

Technology Innovation for Fission Batteries: Autonomous Controls and Operation

11:00	Fission Battery Initiative and Workshop Overview	Youssef Ballout (INL)
11:15	Challenges in Achieving Autonomy in Advanced Reactors	Nam Dinh/ Linyu Lin (NCSU)
11:40	R&D Opportunities to Achieve Autonomous Operation for Fission Batteries	Yasir Arafat (INL)
12:05	Covert Cognizance (C2): Novel Modeling and Monitoring Paradigm for Critical Systems	Abdel-Khalik Hany (Purdue)
12:30	Break	
12:45	Dispatchable, Base-Load Nuclear: The Case for a Fission Thermal Battery	Anthonie Cilliers (Kairos)
1:10	Failures in AI and ML: Insights and Mitigations	Charmaine Cecilia Sample (INL)
1:35	Resilient Fission Battery Control: Challenges & Opportunities	Michael W. Sievers (JPL/NASA)
2:00	Panel Session	

INL National University CONSORTIUM

January 13, 2021

Youssef Ballout, Ph.D.
Director of the Reactor Systems Design and Analysis Division

Fission Battery Initiative

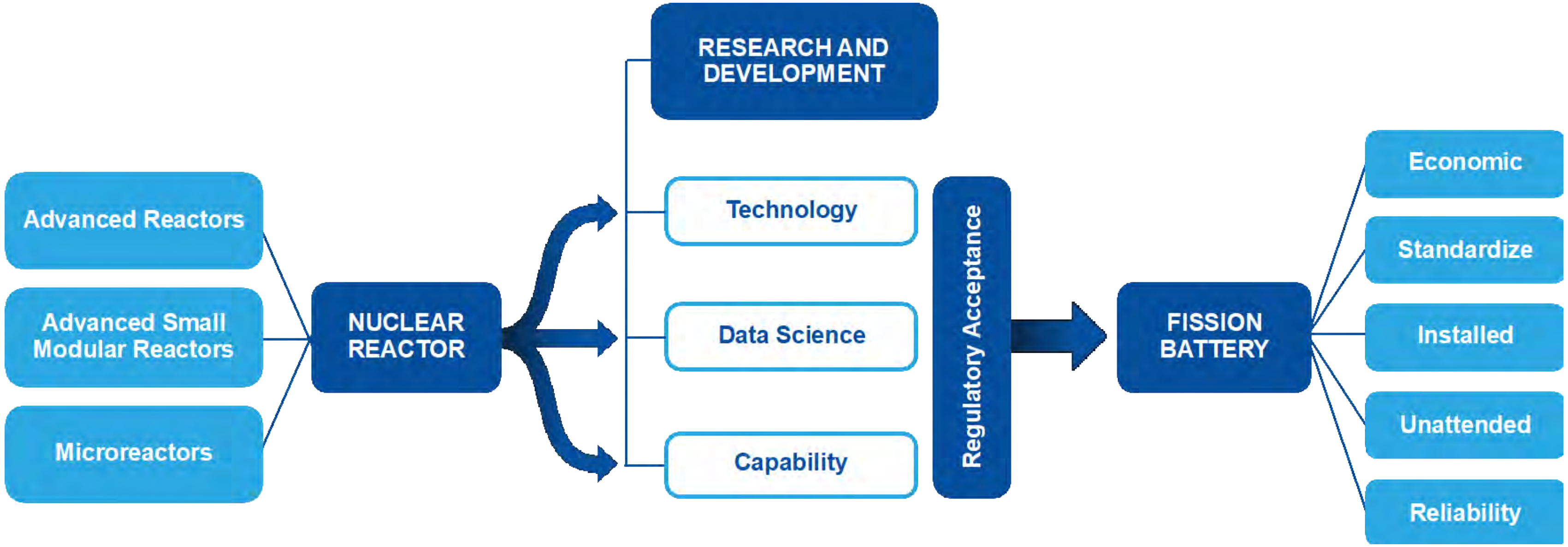
Nuclear Science and Technology

Idaho National Laboratory

Fission Battery Initiative

Vision: Developing technologies that enable nuclear reactor systems to function as batteries.

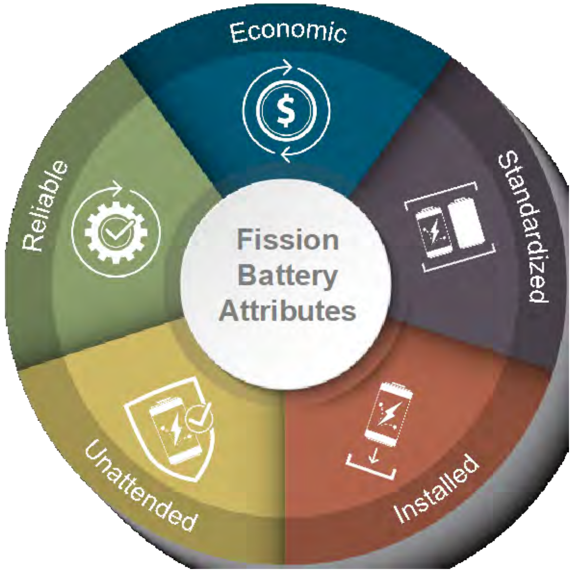
Outcome: Deliver on research and development needed to provide technologies that achieve key fission battery attributes and expand applications of nuclear reactors systems beyond concepts that are currently under development.



Research and development to enable nuclear reactor technologies to achieve fission battery attributes

Fission Battery Attributes

- **Economic** – Cost competitive with other distributed energy sources (electricity and heat) used for a particular application in a particular domain. This will enable flexible deployment across many applications, integration with other energy sources, and use as distributed energy resources.
- **Standardized** – Developed in standardized sizes, power outputs, and manufacturing processes that enable universal use and factory production, thereby enabling low-cost and reliable systems with faster qualification and lower uncertainty for deployment.
- **Installed** – Readily and easily installed for application-specific use and removal after use. After use, fission batteries can be recycled by recharging with fresh fuel or responsibly dispositioned.
- **Unattended** – Operated securely and safely in an unattended manner to provide demand-driven power.
- **Reliable** – Equipped with systems and technologies that have a high level of reliability to support the mission life and enable deployment for all required applications. They must be robust, resilient, fault tolerant, and durable to achieve fail-safe operation.



Fission Battery Workshop Series

- **Jointly INL and National University Consortium are organizing workshops across five areas:**
  - Market and Economic Requirements for Fission Batteries and Other Nuclear Systems
  - Technology Innovation for Fission Batteries
  - Transportation and Siting for Fission Batteries
  - Security Scoping for Fission Batteries
  - Safety and Licensing of Fission Batteries

- **Expected outcomes:**
  - Each workshop outcomes are expected to outline the goals of each fission battery attribute

Idaho National Laboratory

Department of NUCLEAR ENGINEERING

Challenges in Achieving Autonomy in Advanced Reactors

Nam Dinh, Linyu Lin, Edward Chen and Paridhi Athe

Department of Nuclear Engineering

North Carolina State University

NC STATE UNIVERSITY COLLEGE OF ENGINEERING

Outline

- Background
- Digital twins and artificial intelligence
- Issues and solution approaches
- Concluding remarks

New Paradigm in Control Requirements

New Operating Conditions:

- Dynamic & drastic load following vs steady state power generation
- Long-term Operating conditions vs yearly maintenance & fuel swap

Different risk profiles:

- No pumps
- Self contained heat pipes OR submerged in coolant
- Atmospheric Operation
- …

Paradigm Shift in Operation and Control Requirements

- Remote operation
- Long-term operation & maintenance
- Reduced power
- Dynamic load following
- Different risk profiles

→

- Reduced reliance for direct human oversight
- Accurate virtual representations
- Dynamic decision-making system
- Continuous monitoring and learning

Autonomy Enabled by Digital Twin and Artificial Intelligence

Levels of Automation

Department of NUCLEAR ENGINEERING

Level 0 – No Automation
- Manual control, human makes all decisions & actions

Level 1 – Driver Assistance
- SINGLE automated system i.e. speed through cruise control
- Human still controls majority of time

Level 2 – Partial Automation
- Multiple functions are combined and automated (i.e. steering & acceleration).
- Human still monitors all tasks and can control @ any time

Level 3 – Conditional Automation
- Environmental detection, most tasks are autonomous
- Human override is still required.

Level 4 – High Automation
- Performs all task autonomously under SPECIFIC conditions. Geo-fencing required.
- Human override optional.

Level 5 – Full Automation
- All task autonomous, no oversight required.
- Zero human attention & interaction

Human Centered Control (Levels 0–2): **Current generation reactors**

Automated system monitors the environment (Levels 3–5): **Target range of autonomy for advanced reactors**

Advanced reactors increasingly rely on automation systems in O&M

Characteristics of High-Level Automation

Intelligence → minimal to no reliance on human intervention. Whole system control, implies **embedded decision-making & planning authority**.

Robustness → accounts for uncertainties & unmodeled dynamics. Fault management **(avoidance, removal, tolerance, & forecasting)**

Optimization → rapid response, minimal target deviation & efficient actuator actions

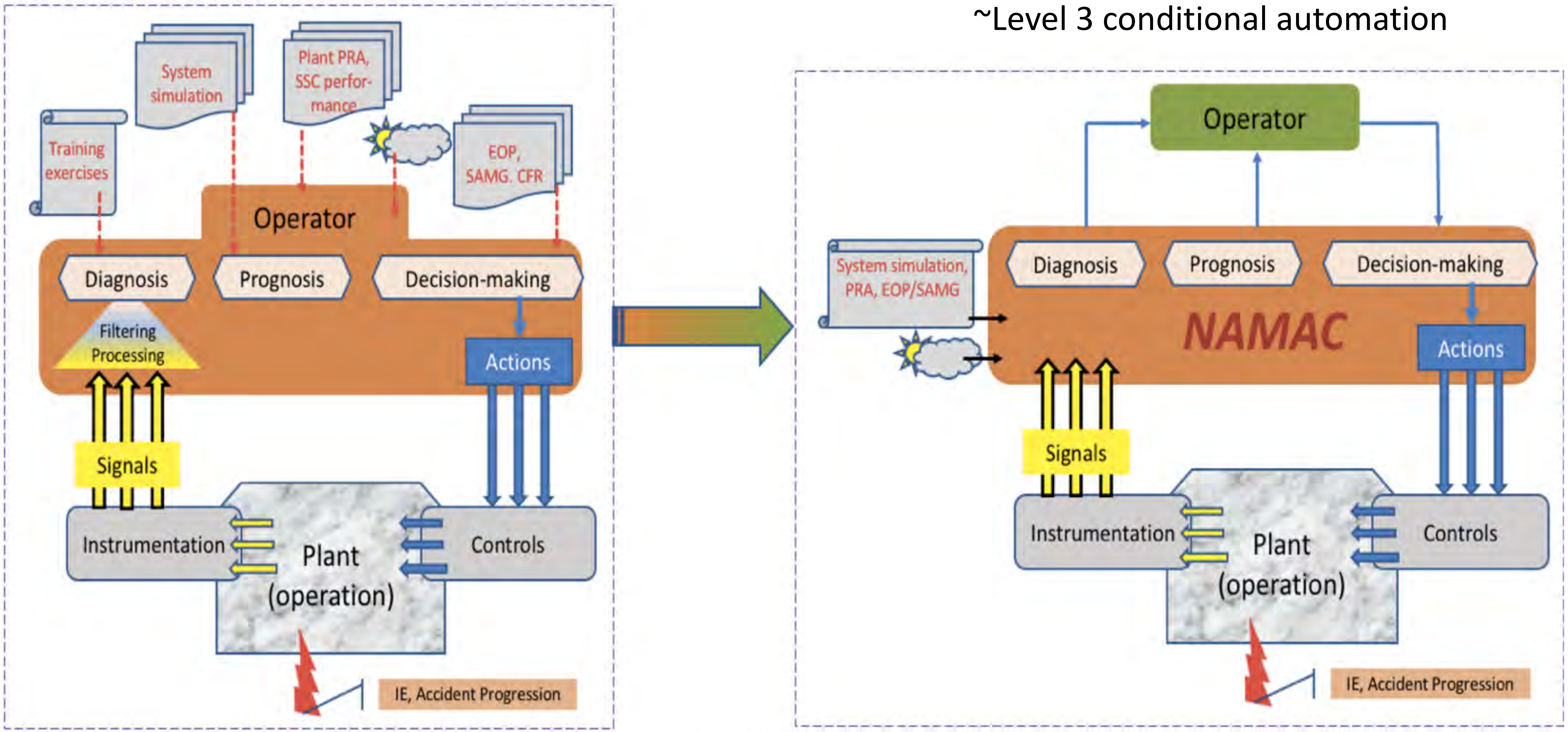
Flexibility & Adaptability → diverse measurements, multiple communication options, & **alternate control solutions**

Higher degrees of autonomy are characterized by greater fault detection and diagnosis, more embedded planning and goal setting, learning and even self-healing

R.T. Wood, et al., "An Autonomous Control Framework for Advanced Reactors", Nuclear Engineering and Technology, 2017

NAMAC as Nearly Autonomous Management and Control

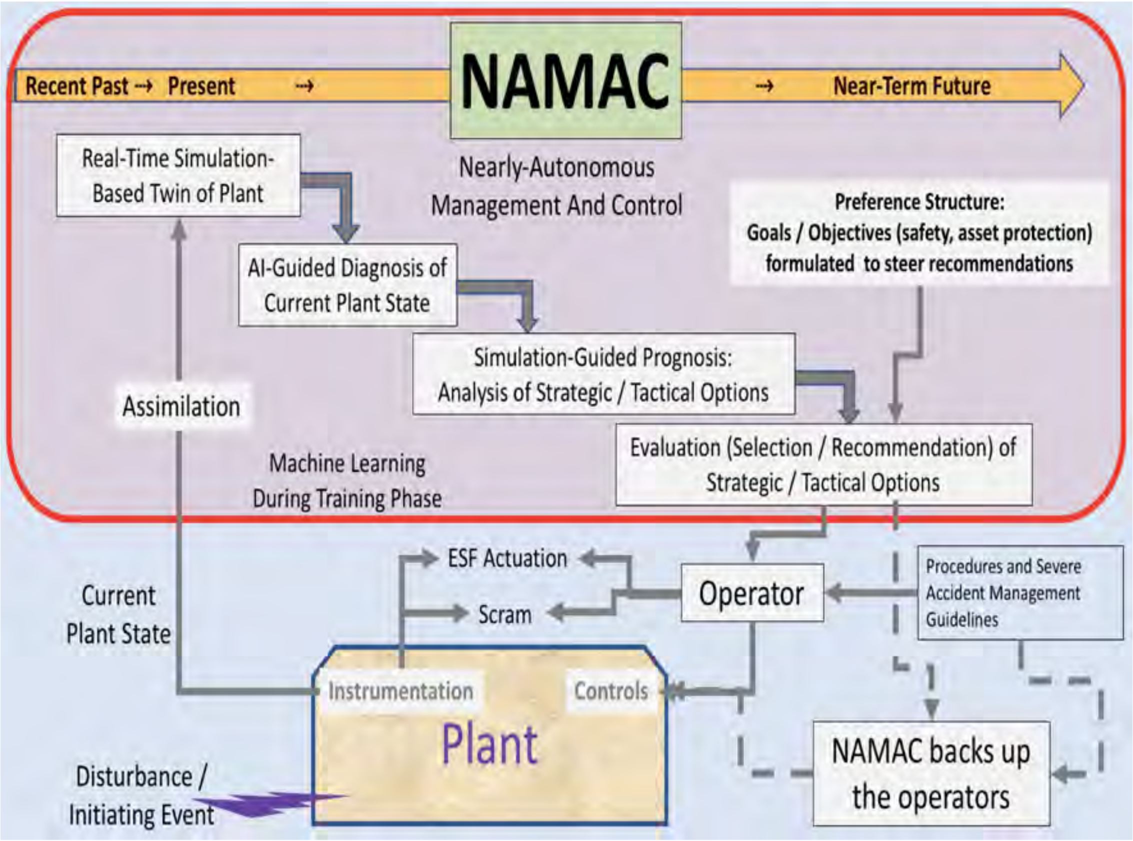
~Level 3 conditional automation



Transition from Operator-Centric Plant Control Architecture to NAMAC-enabled Plant Control Architecture

Nearly Autonomous Management and Control (NAMAC)

- A comprehensive control system to assist plant operations
  - Knowledge integration
    - Scenario-based model of plant (systems, success paths)
    - plant operating procedures, tech. specs., etc.
    - Real-time measurements
  - Digital twin technology
    - Power of AI/ML



- NAMAC
  - Diagnoses the plant state
  - Searches for all available mitigation strategies
  - Projects the effects of actions and uncertainties into the future behavior
  - Determines the best strategy considering plant safety, performance, and cost.

Digital Twin and Artificial Intelligence are key enabling technologies of NAMAC

Outline

- Background
- Digital twins and artificial intelligence
- Issues and solution approaches
- Concluding remarks

Digital Twin (DT)

- Digital Twin technology - construct a digital replica (twin) for the real reactors and transients for the intended use
- DTs provide insights equivalent to Modeling and Simulation (M&S) **BUT**
  - Needs to learn and provide insights faster than the development and uses of M&S
- But DTs are tightly coupled with operation
  - Assimilating and adapting to real-time information from the operating environment
  - Interacting with user for specific objectives

Digital twins need to be adequately modelled for a specific function in a specific operating environment

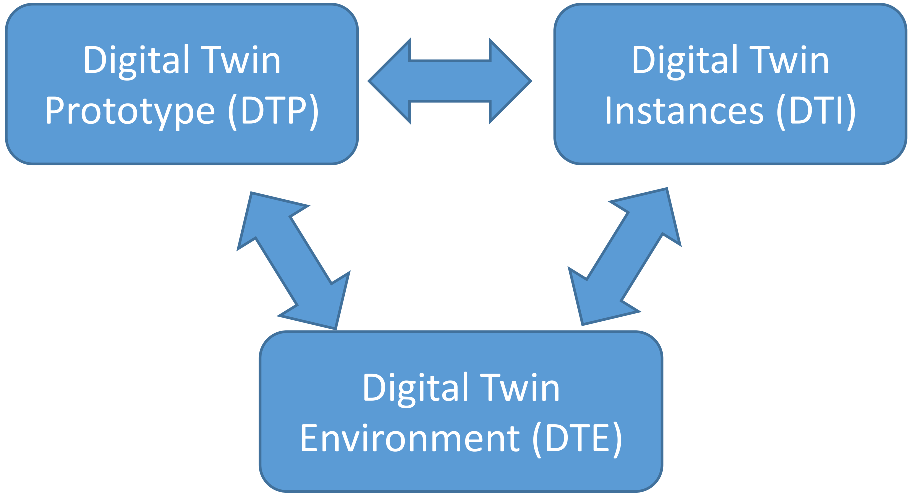
Definitions for DTs [1]

MODEL

- Data-driven model
- Mechanistic model
- Reasoning-based model

INTERFACE

- API
- I/O
- User Interface



FUNCTION

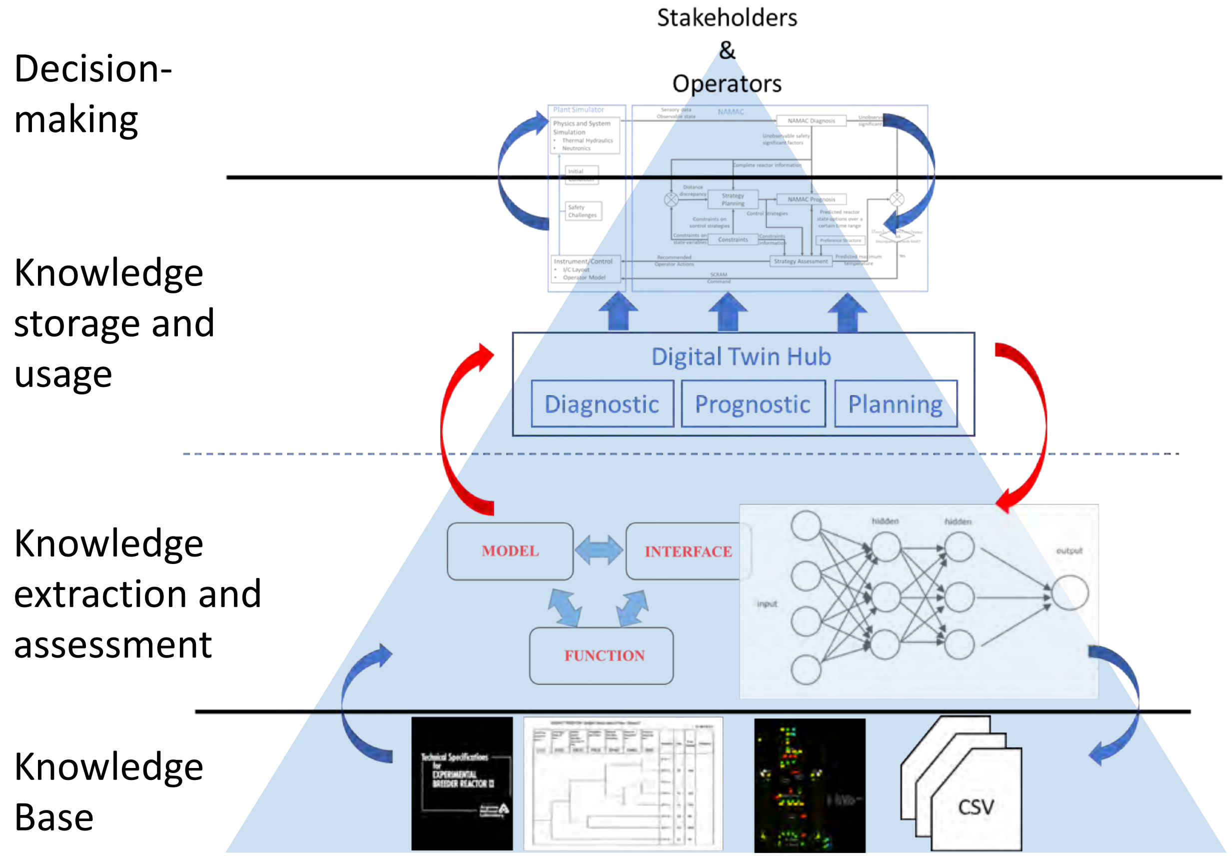
- Use cases
- Objectives
- Output types

[1] F. Kahlen, et al., “Transdisciplinary perspectives on complex systems - new findings and approaches”, Springer, 2017

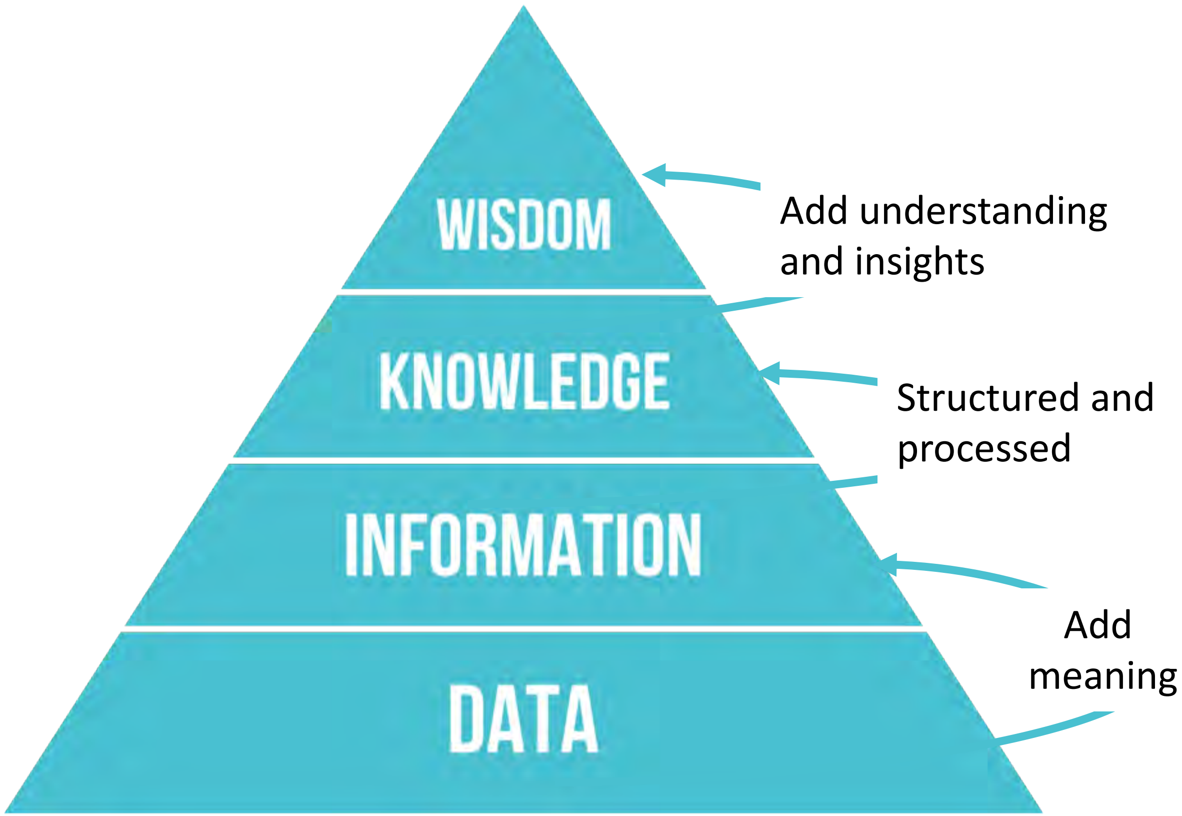
Artificial Intelligence

AI adds meaning to raw data with typical machine learning algorithms like artificial neural networks, fuzzy logics, etc.

NAMAC development process



DIKW Pyramid as a definition and representation for intelligent system [1]

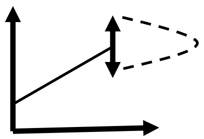
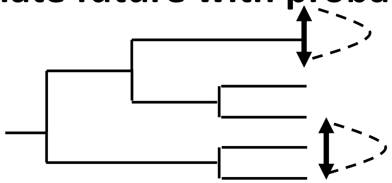
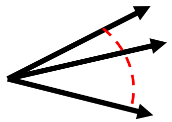


[1] J. Rowley, "The Wisdom Hierarchy: Representations of the DIKW hierarchy", Journal of information Science, pp. 163-180, 2006

Outline

- Background
- Digital twins and artificial intelligence
- Issues and solution approaches
- Concluding remarks

Impact of Digital Twin Uncertainty

Complete Certainty		Level 1	Level 2	Level 3	Total ignorance
	Scenarios' Future States	**A clear future with sensitivity** 	**Alternate future with probabilities** 	**A multiplicity of plausible futures** 	
	Digital Twins	A single set of digital twins with fixed form and parameter	Alternative digital twins with alternative forms and parameters where weights and uncertainties can be sufficiently characterized by probability distributions	Alternative digital twins with alternative forms and parameters where weights and uncertainties are known imprecisely	
	Appropriate target	High-consequence systems where decision making is fundamentally based on DTs, e.g., quantification or final O&M support	Moderate consequence systems with some reliance on DTs, e.g., preliminary O&M support	Low-consequence systems with little reliance on DTs, e.g., scoping studies or conceptual O&M support	

Challenge

Digital Twin uncertainty needs to be evaluated

Digital Twin Development and Assessment Process (DT-DAP)

- DT-DAP to identify major sources of uncertainty and to avoid biases due to implicitness

- The DAP is conducted iteratively, and the corresponding elements are refined until an acceptable set of DTs are delivered

  Element 1: Refined requirements

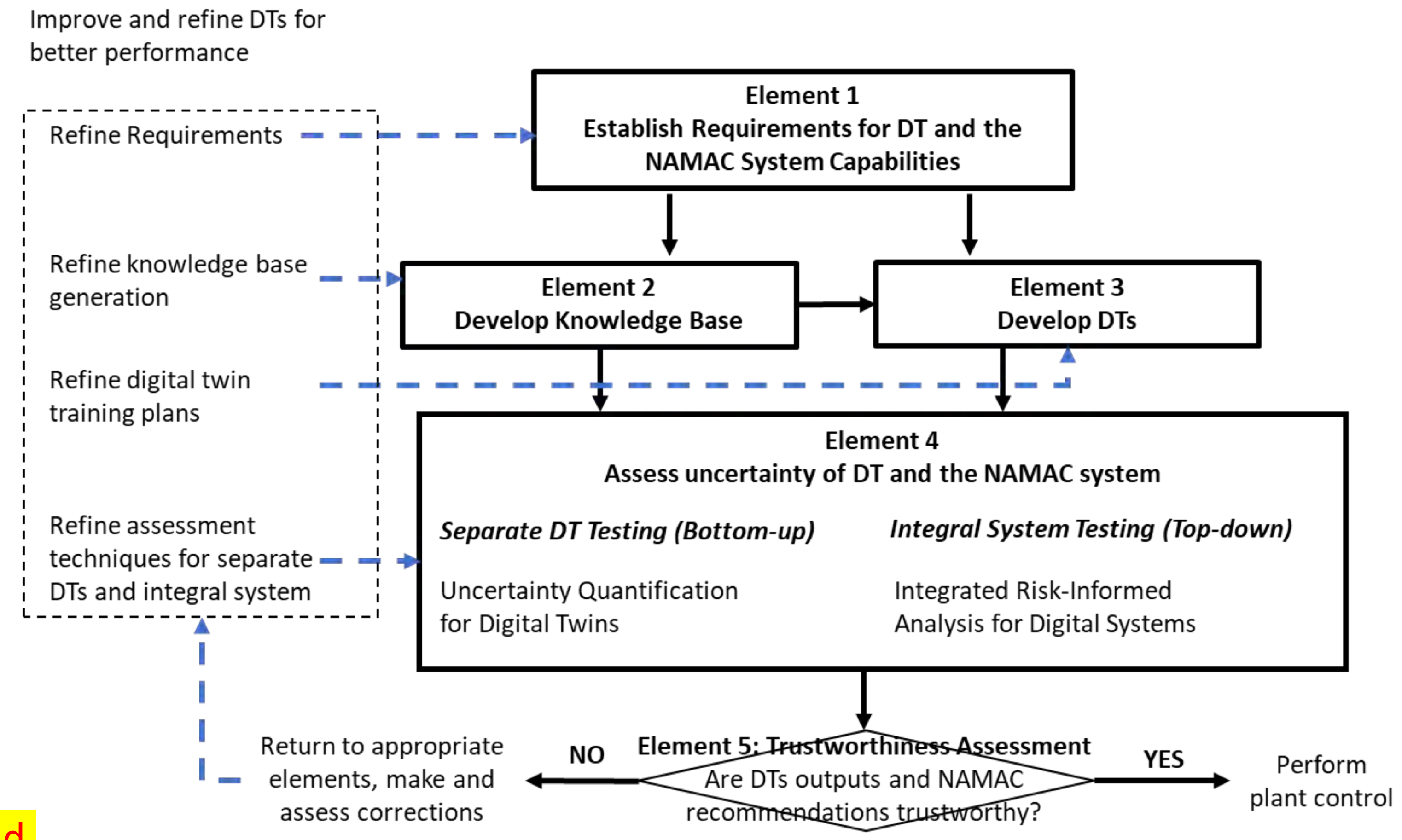
  Element 2: More complex and more realistic knowledge base

  Element 3: Different machine-learning algorithms, hyperparameter tunning

  Element 4: ML uncertainty quantification, software reliability analysis

Challenge in DT-DAP

Digital Twin Trustworthiness needs to be defined and evaluated in a transparent, consistent, and improvable manner



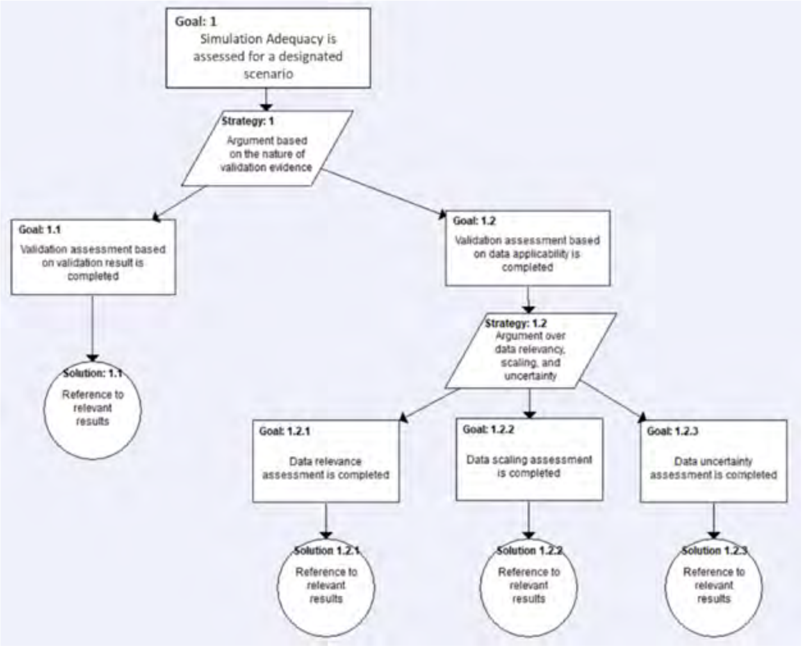
Adopted from U.S. NRC RG 1.203 “Transient and Accident Analysis Methods”

Trustworthiness Assessment

MATURITY / ELEMENT	Maturity Level 0	Maturity Level 1	Maturity Level 2	Maturity Level 3
Representation and Geometric Fidelity				
Physics and Material Model Fidelity				
Code Verification				
Solution Verification				
Model Validation				
Uncertainty Quantification and Sensitivity Analysis				

- For model-based approaches, the trustworthiness, also known as credibility, can be technically assessed by six attributes [1]:
  - Representation and geometric fidelity
  - Physics and material model fidelity
  - Code verification
  - Solution verification
  - Model validation
  - Uncertainty quantification and sensitivity analysis

- For ML-based digital twins, the trustworthiness could depend on
  - Accuracy, Security, Robustness, Explainability, Reliability [2]
  - and more…



Challenges in DT Trustworthiness Assessment

- DT trustworthiness needs to be evaluated by integrating information (evidence) from different sources and heterogeneous types of data
- Complex relations, priority, and trade-off between different attributes of Trustworthiness

[1] W.L. Oberkampf, et al., "predictive capability maturity model for computational modeling and simulation (SAND2007-5948)", Sandia National Laboratory, 2007
[2] NIST, "Fundamental and applied research and standards for AI technologies (FARSAIT)", 2018

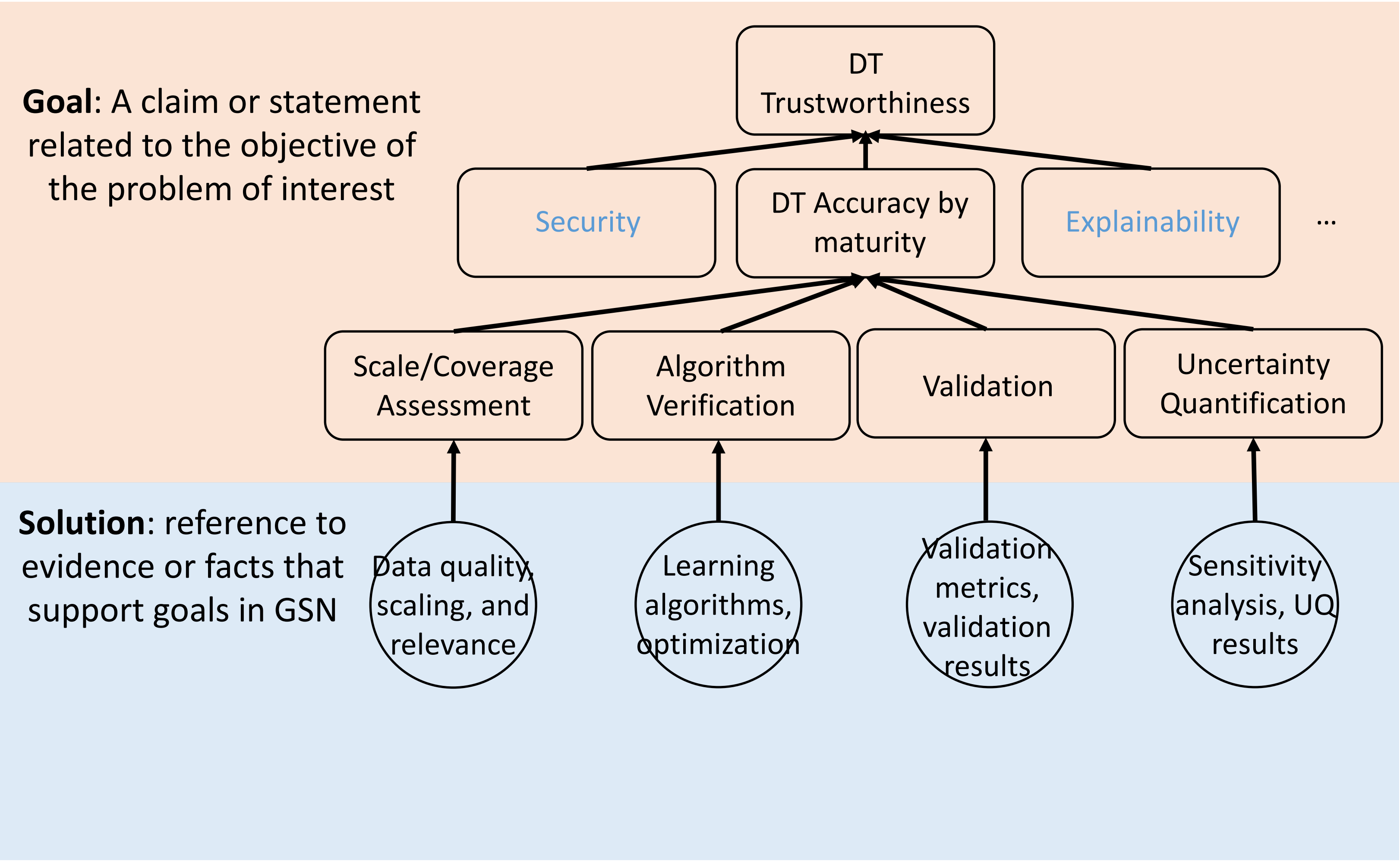
Trustworthiness Assessment

- Accuracy (VVUQ) is one of the major attributes of trustworthiness
- The trustworthiness assessment framework is developed based on assurance case that aims to
  - Justifies if DT is acceptably mature in a structured argument, supported by evidence, for a specific application in a specific operating environment

Challenges in DT Assurance Case

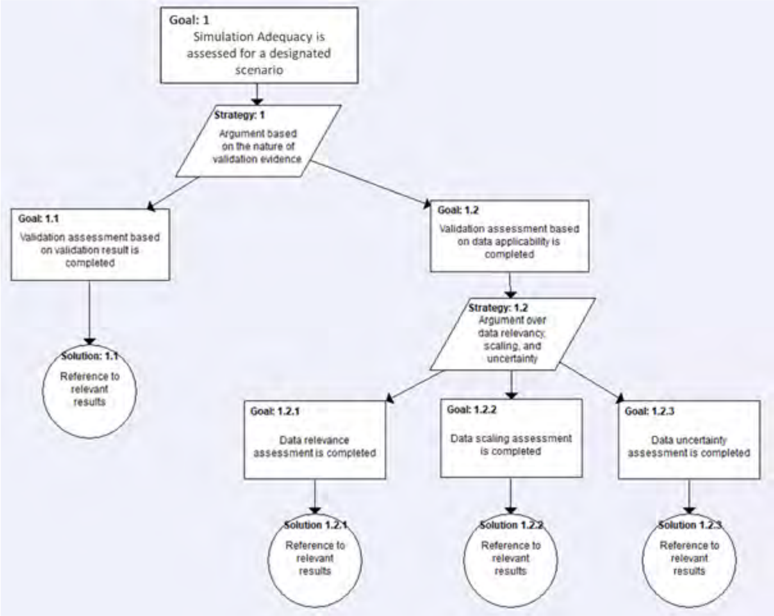
- Define DT maturity
- Collect and integrate evidence
- Online maturity evaluation and real-time deviation detection

Trustworthiness Assessment Framework



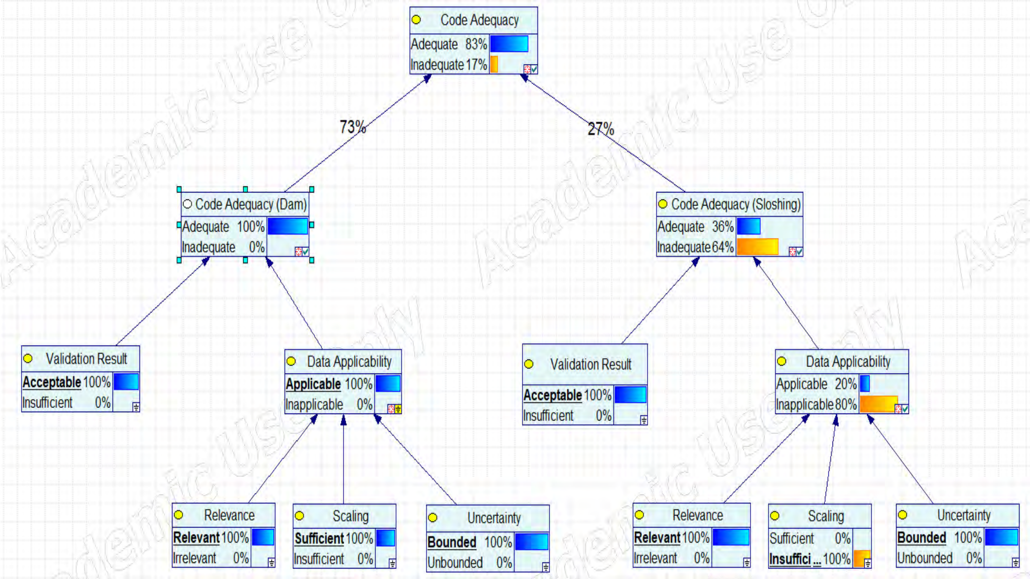
Predicted Capability Maturity Quantification (PCMQ)

- Similar techniques, named predictive capability maturity quantification by Bayesian network (PCMQBN), are developed to evaluate the adequacy (maturity or credibility) of a computational fluid dynamic (CFD) code in simulating an external-flooding scenario [1]



Transfer the argument to a belief network

$$\text{Simulation Adequacy} = \{\text{Scenario, Belief, Maturity Levels}\}$$



Challenges in adapting PCMQ to DT trustworthiness assessment

- Quantifying evidence and maturity
- Dependency among different evidence and goal nodes
- Relating accuracy results with risk analysis

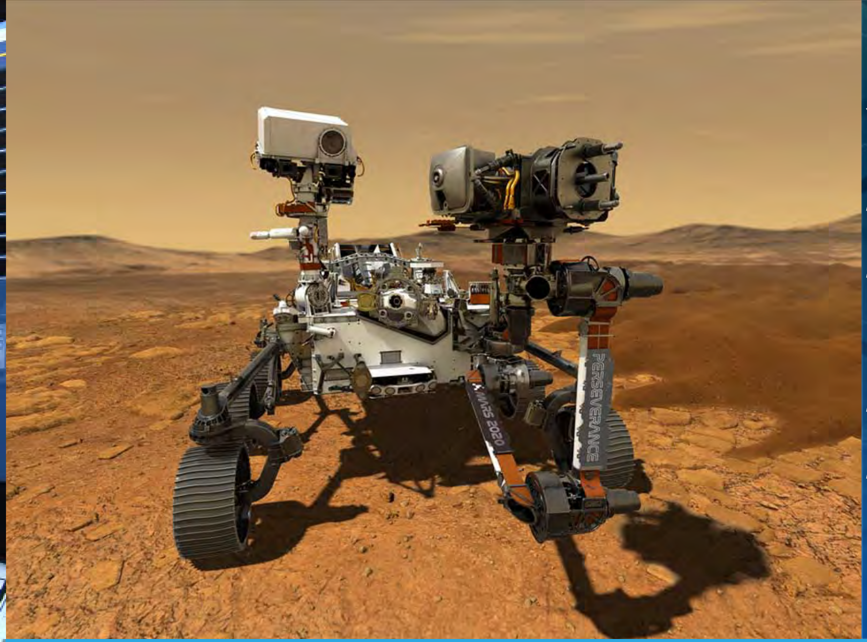
[1] L. Lin, N. Dinh, "Predictive capability maturity quantification using Bayesian network", Journal of Verification, Validation, and Uncertainty Quantification, 2020

Summary

- Advanced Reactor design offers opportunity and challenges for advanced control strategies
  - The ideal levels of automation are to be adapted, but expected to be high-level, risk-informed and data-driven
  - Characteristics of autonomy are largely conceptual, and their relations/trade-off need to be evaluated
- From NAMAC's experience, digital twin and artificial intelligence are key enabling technologies of autonomous control systems
  - Digital twins' uncertainty presents a major challenge, and we suggest dealing with it through a formal framework
- In the digital twin development and assessment process, the trustworthiness is a critical element
  - It is a challenge to collect and integrate heterogenous types and sources of evidence, and we suggest an accuracy assessment framework by software assurance case
  - We suggest adapting the predictive capability maturity quantification (PCMQ) framework for assessing the maturity of DTs and AI.

Acknowledgement

- The NAMAC system is developed with the support of ARPA-E MEITNER program under the multi-organizational (NCSU-NMSU-OSU-INL-ORNL-ANL-TP-ZNE) collaborative project entitled:” Development of a Nearly Autonomous Management and Control System for Advanced Reactors”



Yasir Arafat

Microreactor Technical Lead

Nuclear Science and Technology (NS&T)

MARVEL Project & Technical lead,

DOE Microreactor Program | NRIC

Fission Batteries

R&D Opportunities to achieve Autonomous Operation

January 2021– Idaho Falls, ID

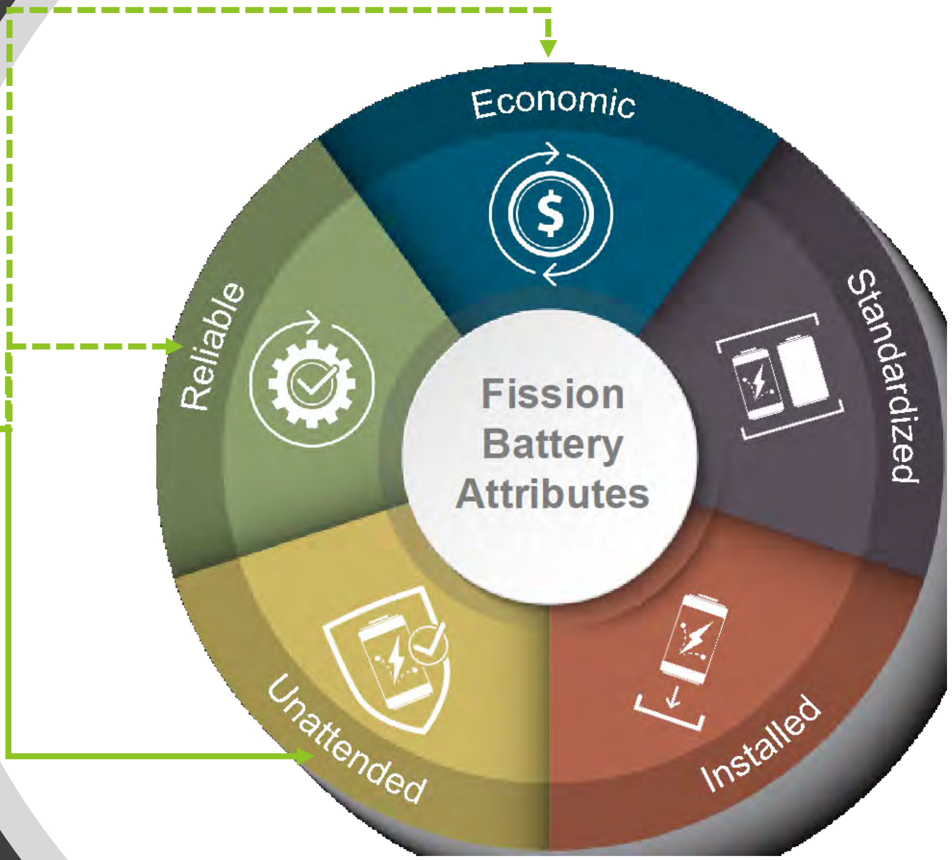
Idaho National Laboratory

Autonomous Operation

- **Autonomous control systems** are designed to perform well under significant uncertainties in the system and environment for extended periods of time, and they must be able to compensate for system failures without external intervention"

Vs.

- **Automation**, which is often defined as a process or procedure performed with minimal human assistance



Why seek autonomous operation?

- Operators for a fission battery is a **significant cost driver**
- **Staffing requirements** during operations
  - Constraint: design, regulations, end user
  - % contribution to LCOE by # staff

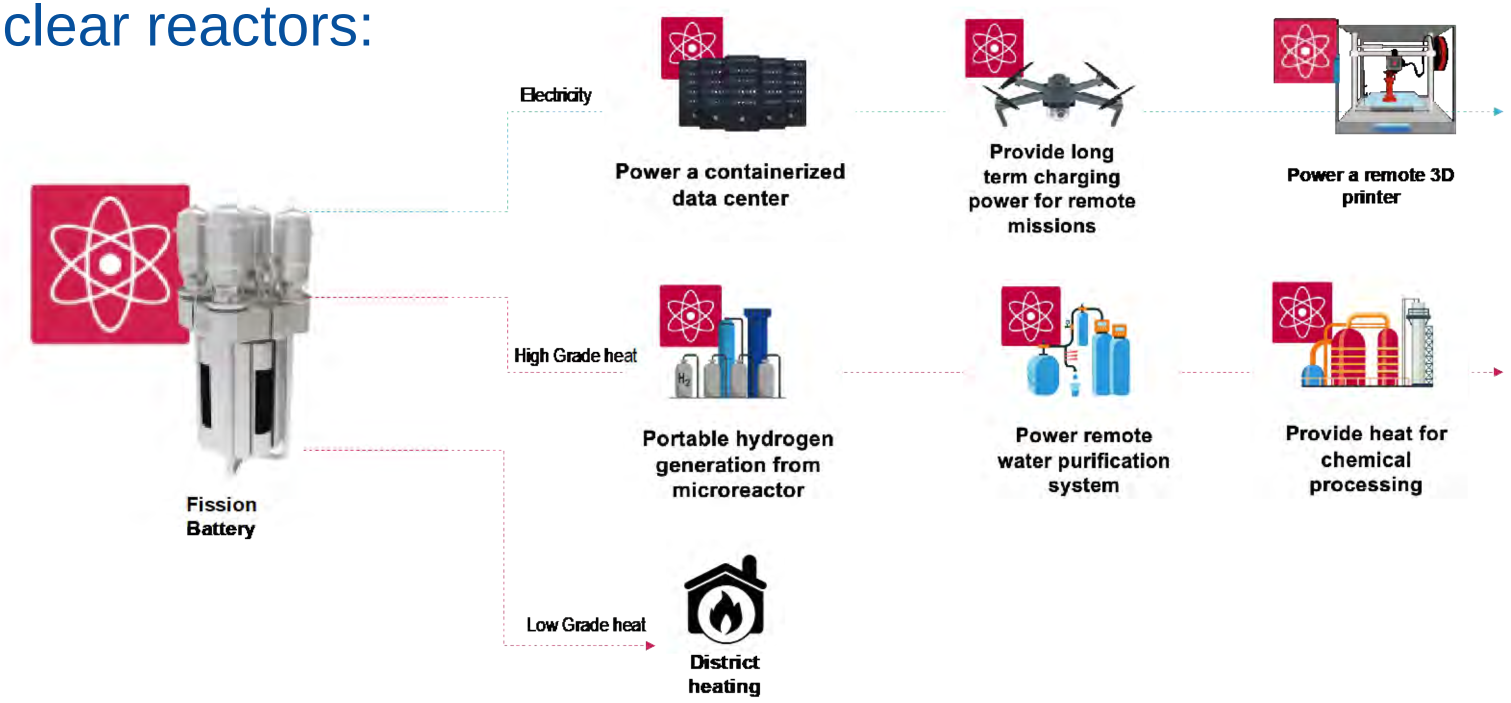
*Assuming 5 years of operation

MWe	# staff	$/hr	$/MWh	LCOE	% LCOE	Max CAPEX of Autonomous systems *
1.5	1	100	67	$ 450	15%	$4.3M
1.5	2	100	67	$ 450	30%	$8.6M
3	1	100	33	$ 450	7%	$4.3M
3	2	100	66	$ 450	14%	$8.6M
10	2	100	20	$ 200	10%	$8.6M
30	2	100	3	$ 200	2%	$8.6M

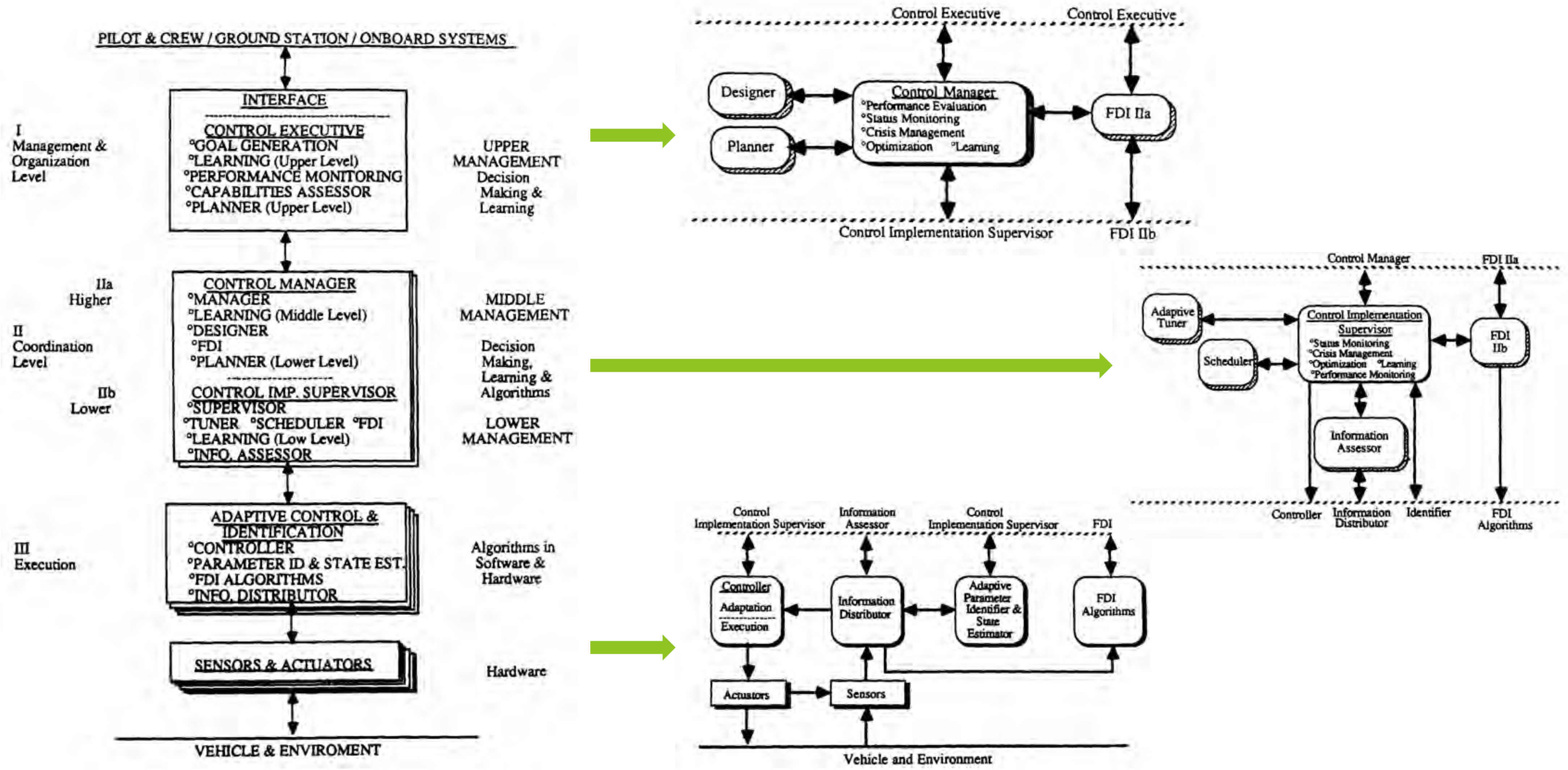
- Maximum CAPEX for autonomous systems is independent of reactor size
  - Autonomous system replaces hourly rate of staff for x amount of years

Operation & Maintenance in Nuclear Power Plants

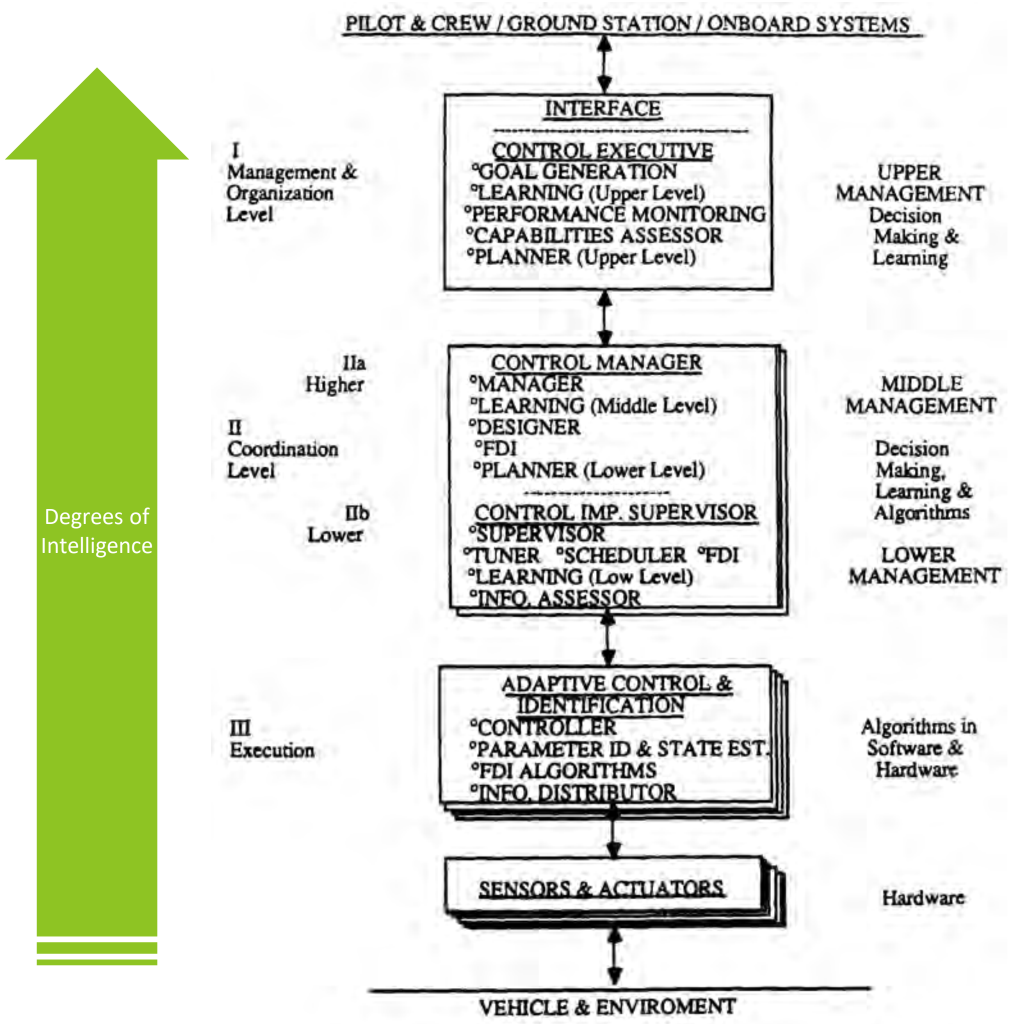
- Nuclear reactors are **complex systems** that utilize sophisticated controllers, trained operators to achieve desired performance for
  - Operability (match demand and supply of electricity)
  - Safety (ensure no radiological impact to people/environment)
- **Functions** of “people” in today’s nuclear reactors:
  - Reactor Startup & Shutdown
  - Evaluate Plant Performance
  - Fault-detection & diagnosis
  - Emergency Operation
  - Fuel reload
  - Load Management
  - Demand Management
  - Maintenance
  - Repair

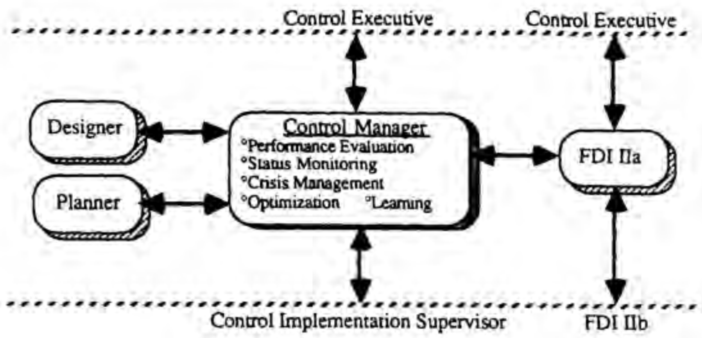


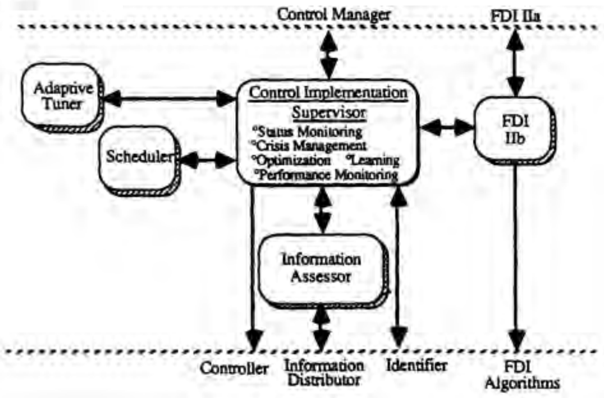
Architecture Schematic of Autonomous Controller

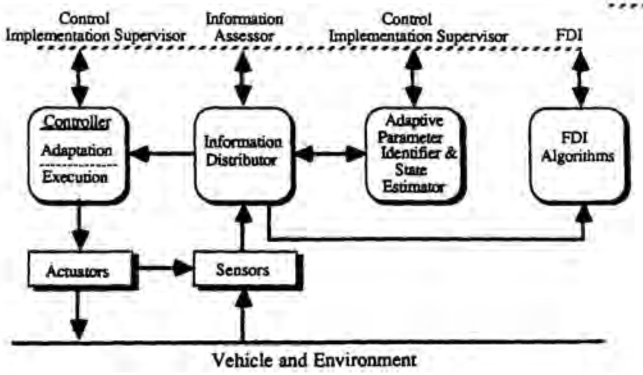


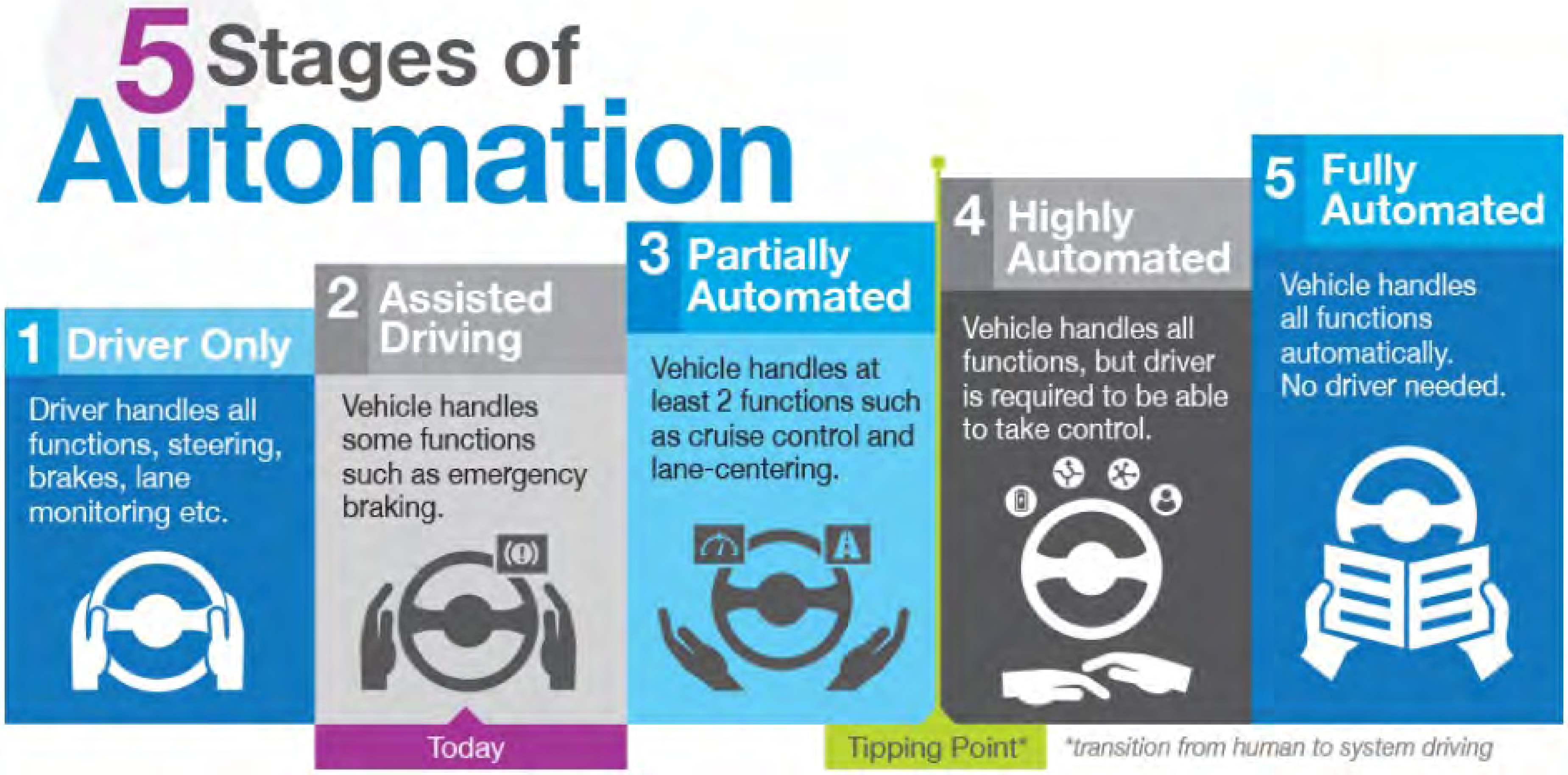
Architecture Schematic of Autonomous Controller





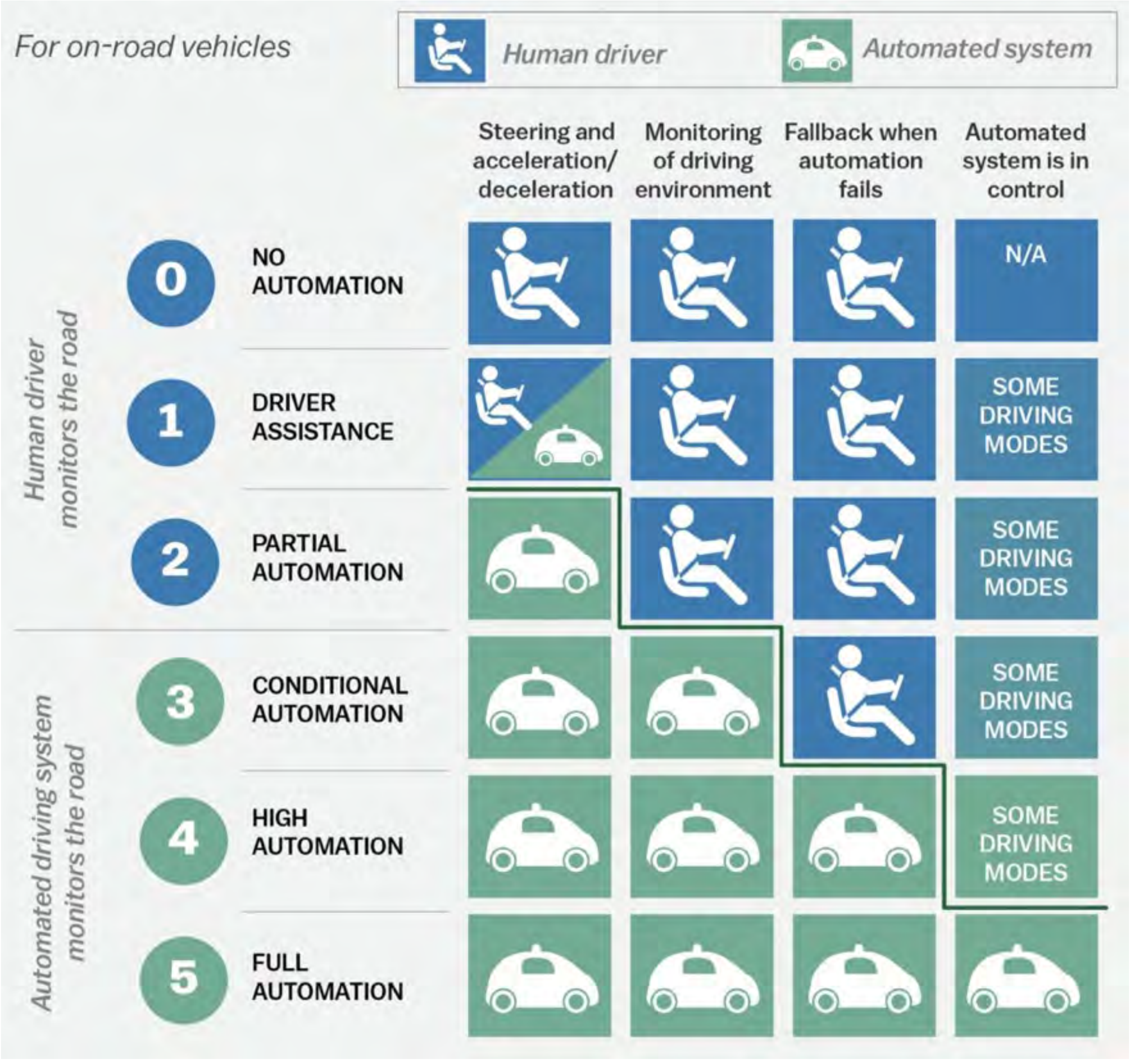




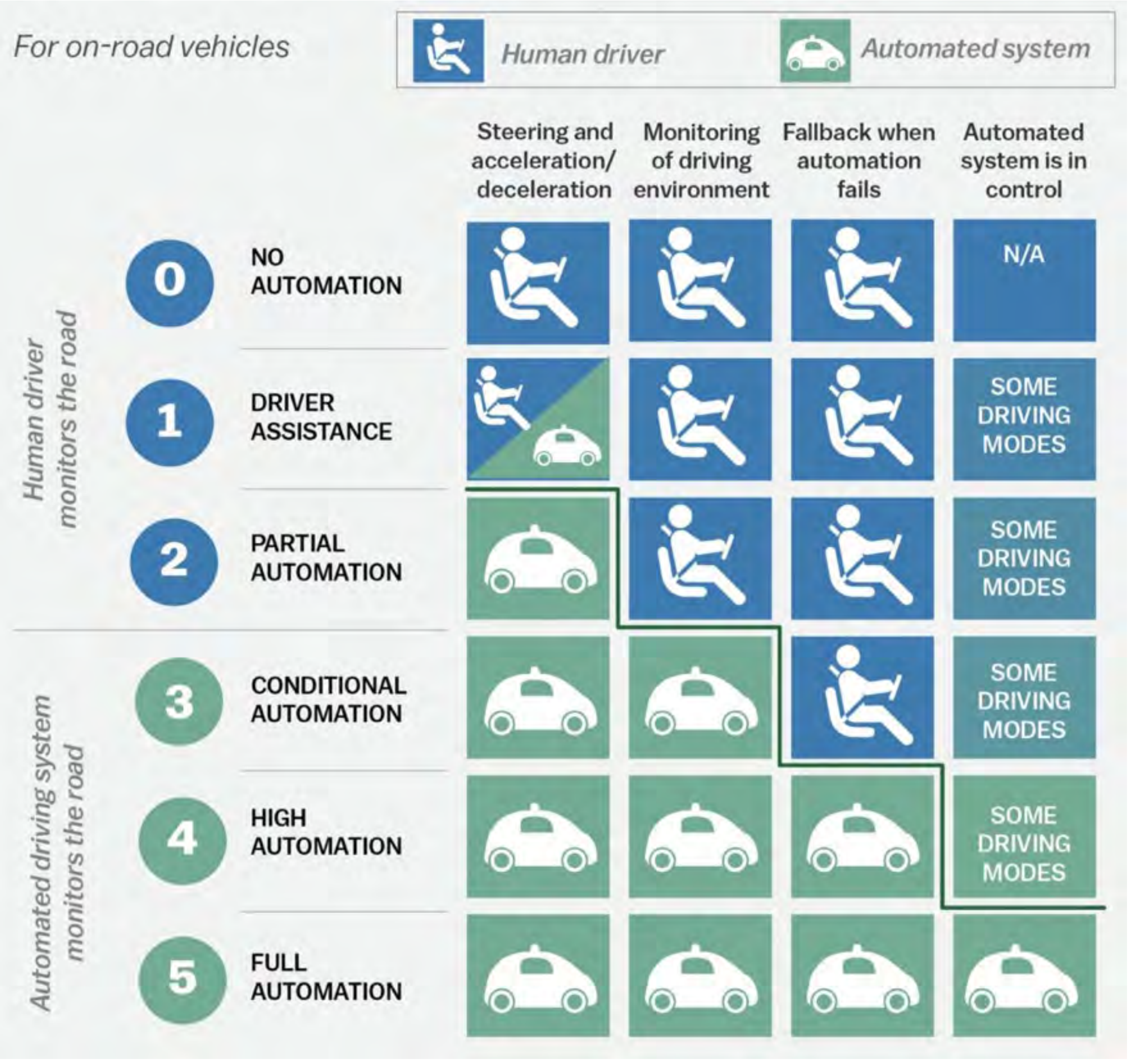


What does it mean for nuclear reactors?

Levels of Autonomy



Levels of Autonomy



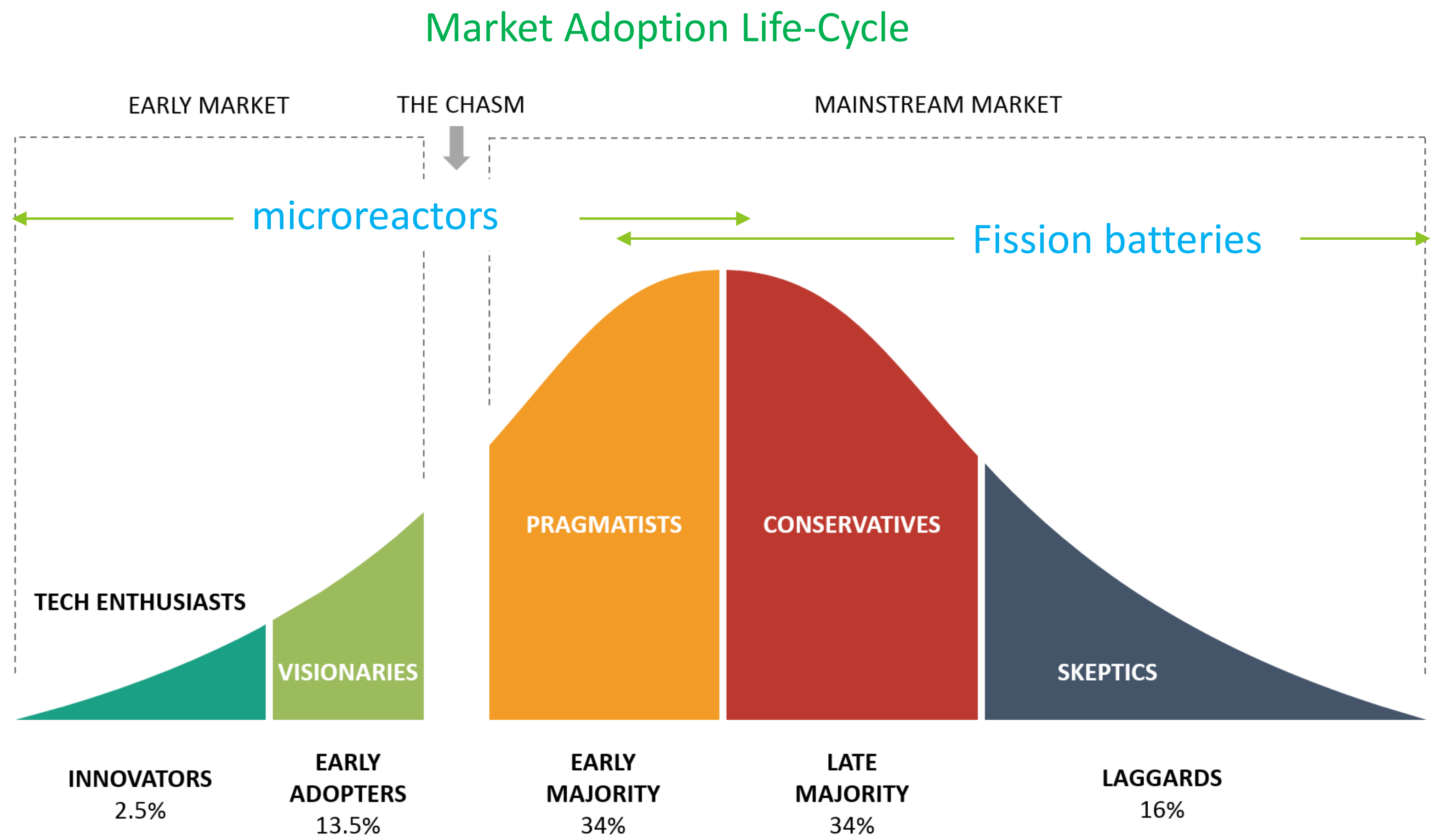
Microreactors

- Gen I Microreactors
- Staffed reactors with remote monitoring
- Staffed reactors, with remote monitoring & control
- Unstaffed reactors, with remote monitoring & full control

Fission Batteries

- Unstaffed reactors, with remote monitoring & partial control
- Unstaffed reactors, with remote monitoring & no operator control

Who are the end users of autonomous fission batteries?

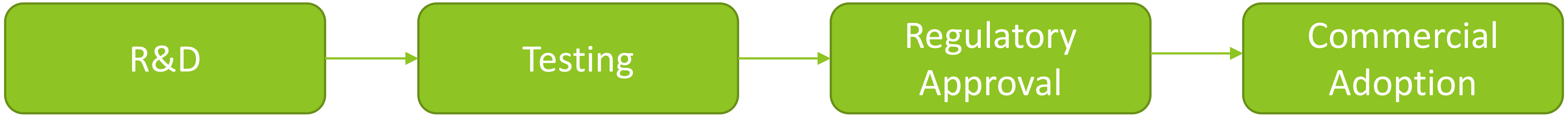


Microreactor Markets: Industry/DOE/DoD Demo* | Eilson* AFB | Remote Applications* | Grid-Edge* | Centralized Grid*

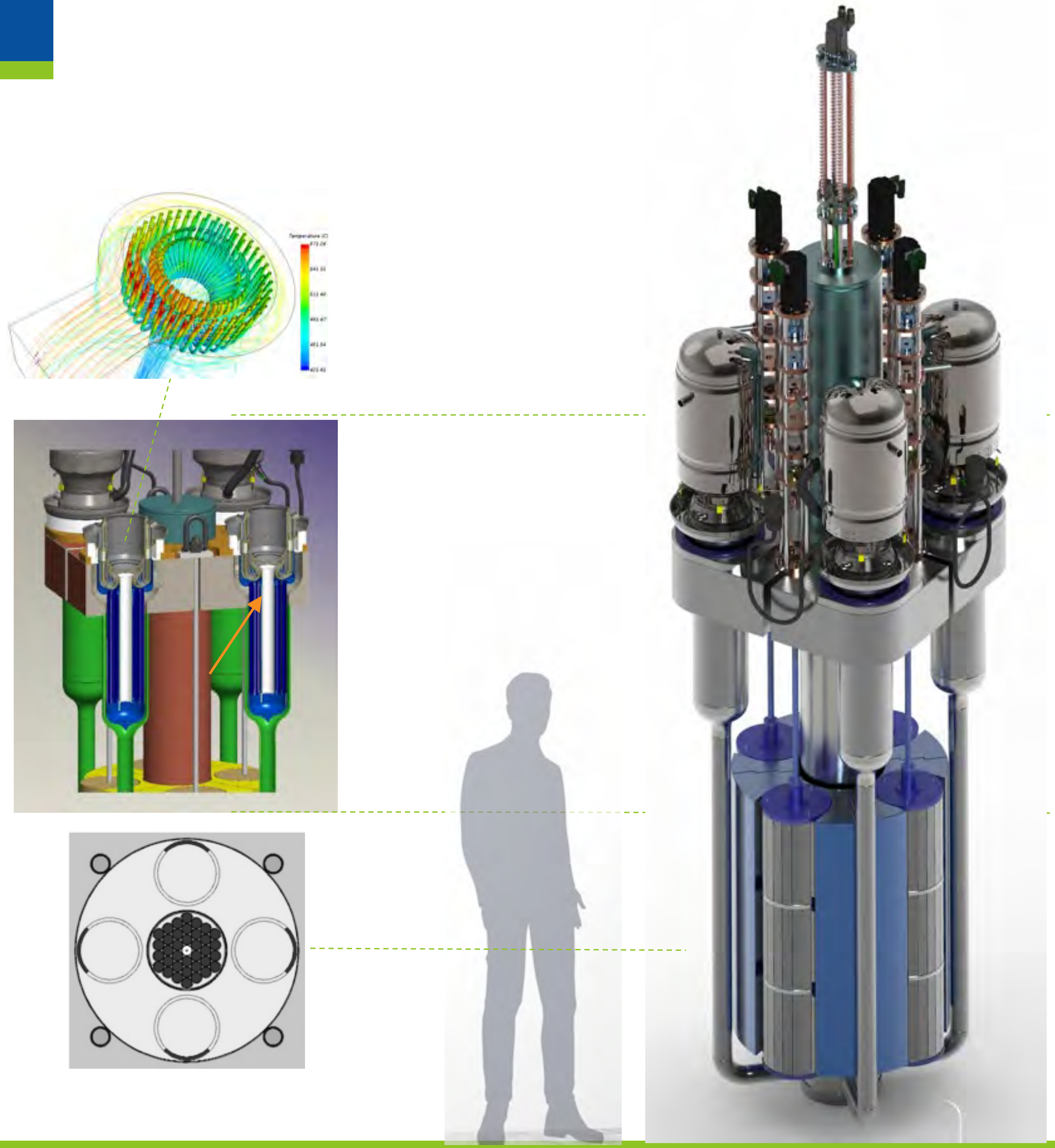
* Not validated

Today’s Challenges of Autonomous Operation

- Real autonomous control systems are only feasible
  - with the availability of cheap sensors,
  - the capacity to handle enormous amounts of data, and
  - the processing capacity and methods to perform the necessary decision algorithms
  - Cyber threats → especially with remote control
- Significant regulatory hurdles to license for Autonomous operation
  - Environment: High consequence to failure
  - Performance: Lack of testing data
  - Reliability: Manually operated microreactors/fission batteries must be deployed first
- Some suggest that conventional control has a better, more established track record than techniques from intelligent control, which are relatively new and in a very early stage of development.



MARVEL- Testbed for Autonomous Control Systems



- Thermal Power- 100 kWth
- Electrical Output ~20 kWe
- Max High Grade heat ~ 45 kWth @ 450 °C
- Max Low Grade heat ~ 75 kWth @ 50 °C
- Modified TRIGA fuel- UZrH1.7 (made in INL)
- Inspired by SNAP 10A core geometry: 36 pins
- Four helium Stirling engines @ 400-500 C inlet T
- Air is ultimate heat sink for primary and decay heat removal

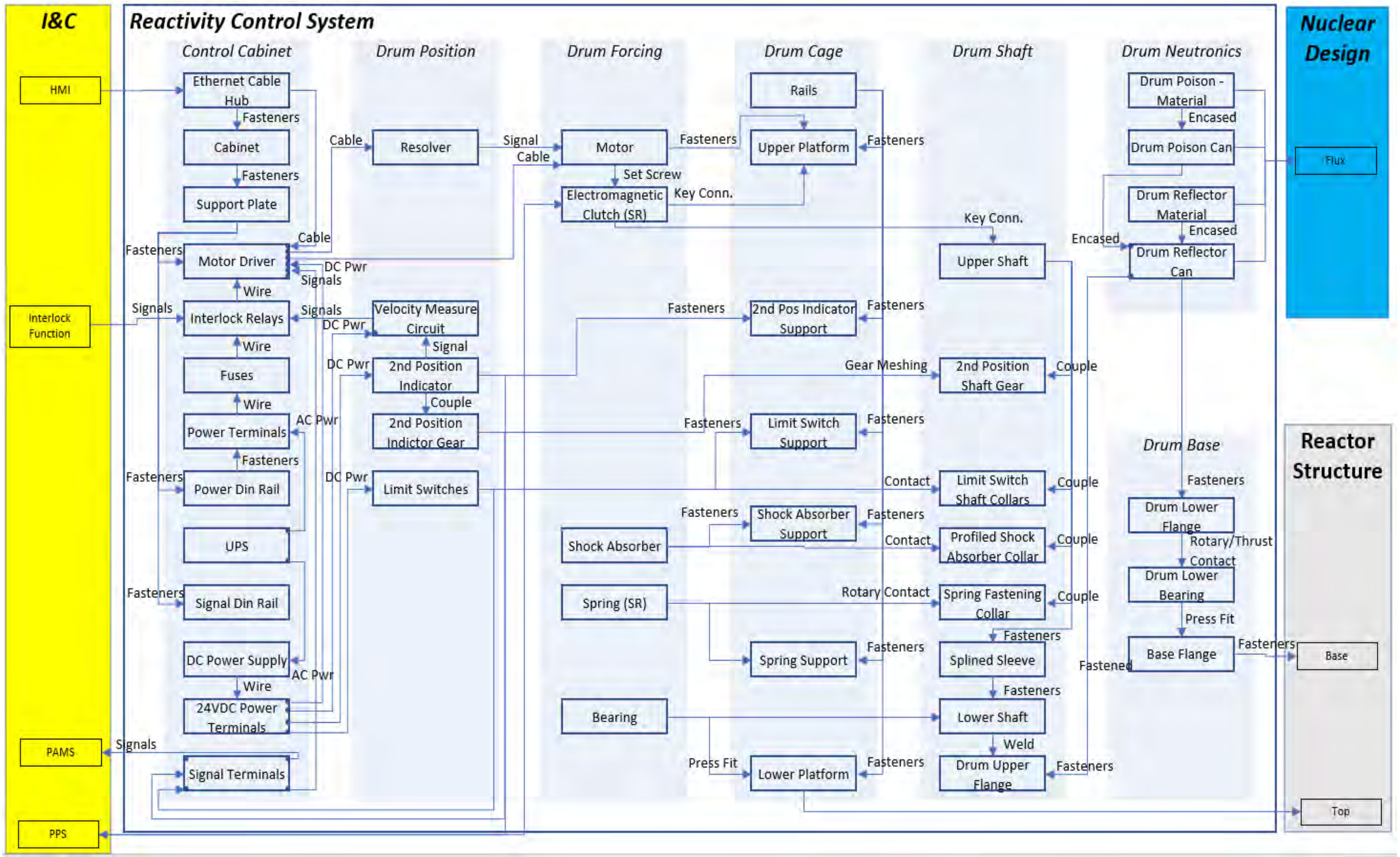
Site: TREAT Storage Pit (8’x12’x10’) and TREAT control room

Storage pit → T-REXC
TREAT microReactor EXperiment Cell

Control Room

MARVEL Operation & Maintenance

- Current Criticality Target: June 2022
- 4 years operation;
- < 50% capacity factor
- Manual Operation; 2 operators (SRO, RO)
- Remote monitoring (power only)
- Microgrid Controller & renewable generation interface
- Planned maintenance- minimum
- Unplanned maintenance/repairs- spares


I&C
HMI
Interlock Function
PAMS
PPS
Reactivity Control System
Control Cabinet
Ethernet Cable Hub
Fasteners
Cabinet
Fasteners
Support Plate
Fasteners
Motor Driver
Cable
DC Pwr
Signals
Wire
Signals
Interlock Relays
Signals
DC Pwr
Wire
Fuses
Wire
Power Terminals
AC Pwr
Fasteners
Fasteners
Power Din Rail
UPS
Fasteners
Signal Din Rail
DC Power Supply
AC Pwr
Wire
24VDC Power Terminals
Signals
Signal Terminals
Drum Position
Cable
Resolver
Velocity Measure Circuit
Signal
DC Pwr
2nd Position Indicator
Couple
2nd Position Indictor Gear
DC Pwr
Limit Switches
Drum Forcing
Signal Cable
Motor
Set Screw
Electromagnetic Clutch (SR)
Key Conn.
Shock Absorber
Spring (SR)
Bearing
Press Fit
Drum Cage
Rails
Fasteners
Upper Platform
Fasteners
Fasteners
2nd Pos Indicator Support
Fasteners
Fasteners
Limit Switch Support
Fasteners
Fasteners
Shock Absorber Support
Fasteners
Spring Support
Fasteners
Lower Platform
Fasteners
Drum Shaft
Key Conn.
Upper Shaft
Gear Meshing
2nd Position Shaft Gear
Couple
Contact
Limit Switch Shaft Collars
Couple
Contact
Profiled Shock Absorber Collar
Couple
Rotary Contact
Spring Fastening Collar
Couple
Fasteners
Splined Sleeve
Fasteners
Lower Shaft
Weld
Drum Upper Flange
Fasteners
Drum Neutronics
Drum Poison - Material
Encased
Drum Poison Can
Drum Reflector Material
Encased
Encased
Drum Reflector Can
Drum Base
Fasteners
Drum Lower Flange
Rotary/Thrust Contact
Drum Lower Bearing
Press Fit
Base Flange
Fastened
Fasteners
Nuclear Design
Flux
Reactor Structure
Base
Top

R&D Pathway to achieve Autonomous Operation using MARVEL

To achieve full autonomous operation, we have to…

Start Small, Dream Big

	Remote Monitoring & Control	Operator control	Machine Control
Phase 0	No	Full	No
Phase 1	**Yes**	Full	No
Phase 2	Yes	**Partial**	**Partial**
Phase 3	Yes	**No**	**Full**

R&D Pathway to achieve Autonomous Operation using MARVEL

To achieve full autonomous operation, we have to…

Start Small, Dream Big

	Remote Monitoring & Control	Operator control	Machine Control
Phase 0	No	Full	No
Phase 1	**Yes**	Full	No
Phase 2	Yes	**Partial**	**Partial**
Phase 3	Yes	**No**	**Full**

Operation Functions

- Reactor Startup & Shutdown
- Evaluate Plant Performance
- Fault-detection & diagnosis
- Emergency Operation
- ~~Fuel reload~~
- Load Management
- Demand Management
- Maintenance
- Repair

Fuzzy Logic

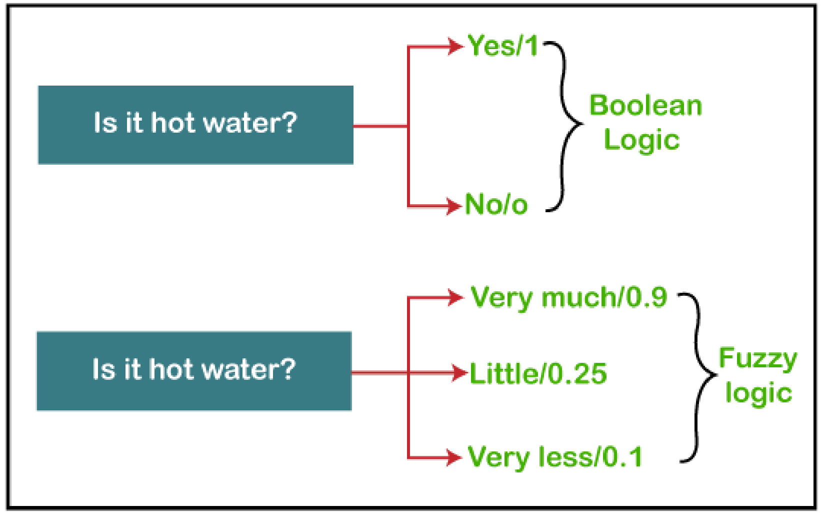
Advance Sensors

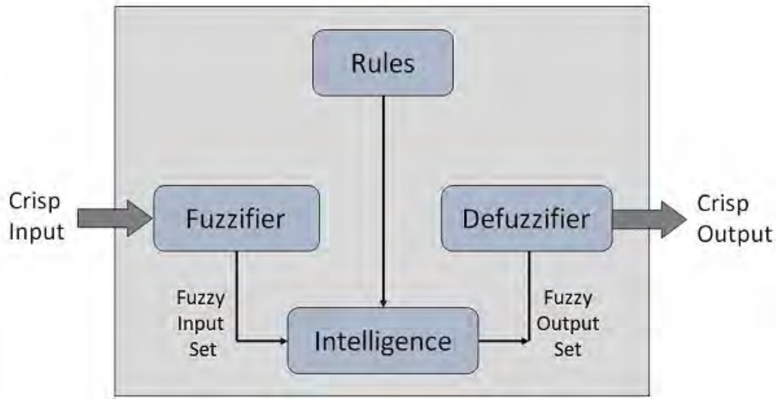
Digital Twin

Artificial Intelligence

Reduced Order Models

Phase 2: Partial Operator Function (Fuzzy Logic)

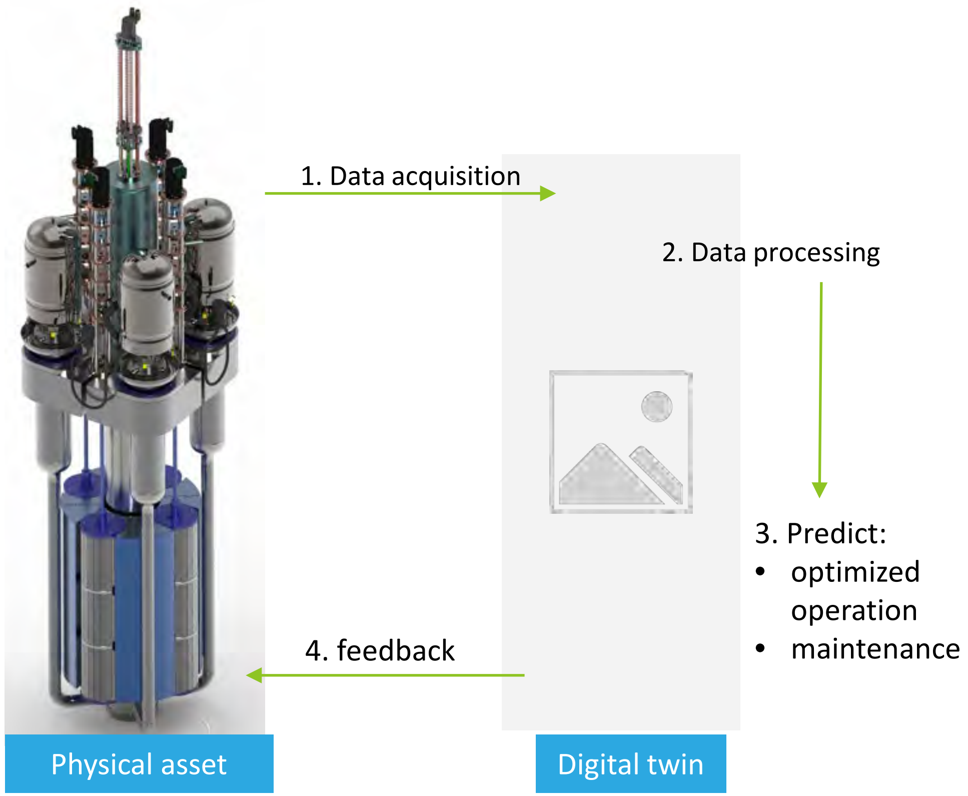




- Example:
  - Reactors have design limits on structural materials, fuel, coolant, etc
  - In a postulated accident condition, if these design limits are reached, reactors need shut down to prevent any catastrophic failure
  - With fuzzy logic, we don’t necessarily have to shut down the reactor, rather operate at lower power or avoid thermal cycling
- Benefits: Make better/faster safety & operability decisions, Improve availability → reduce operator functions
- Some reactors like MARVEL are ideal to test fuzzy logic, because of safety pedigree, i.e. strong reactivity feedback

Phase 2: Partial Operator Function (Digital Twin)

- A digital twin is a digital/virtual copy of physical asset or product



Phase 2: Partial Operator Function (Digital Twin)

- A digital twin is a digital/virtual copy of physical asset or product

1. Data acquisition

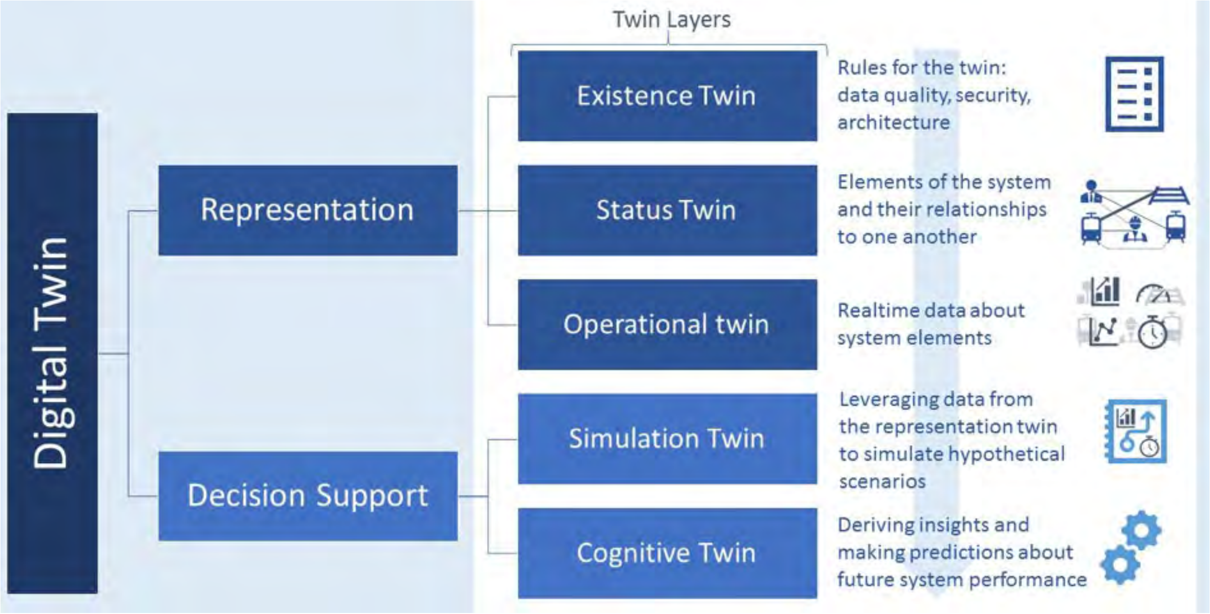
2. Data processing

3. Predict:
- optimized operation
- maintenance

4. feedback

Physical asset

Digital twin



What is there?

What is it doing?

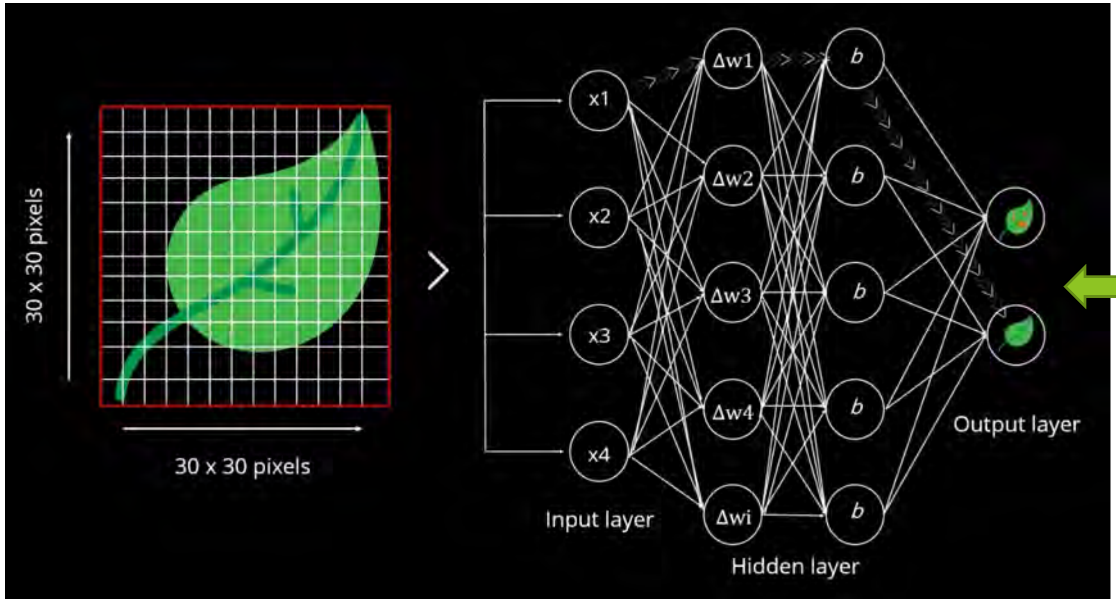
What is it doing NOW?

What might it do IF ___ changes?

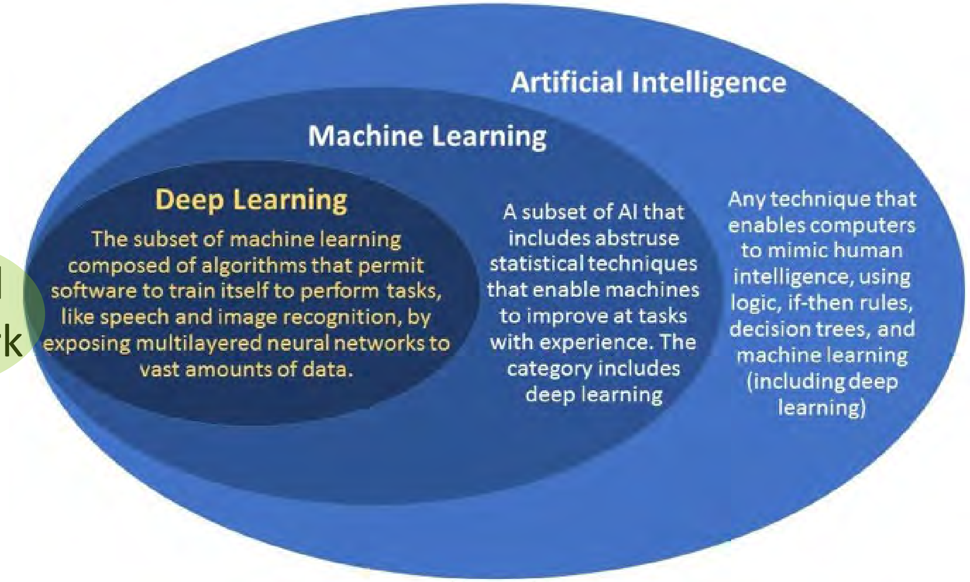
What might is it going to do?

Reference: https://www.rssb.co.uk/what-we-do/insights-and-news/blogs/digital-twins-and-the-railway-one-framework-many-implementations

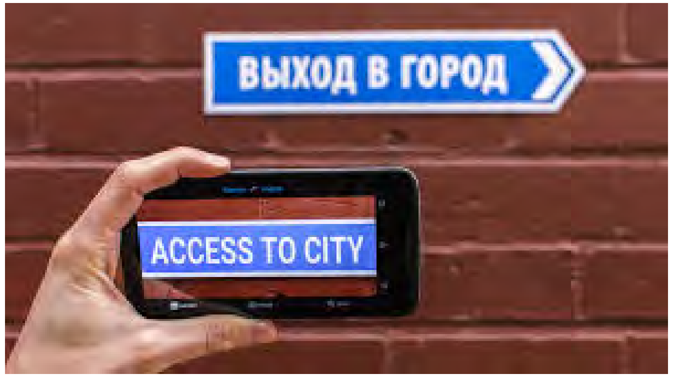
Phase 3: Neural Network



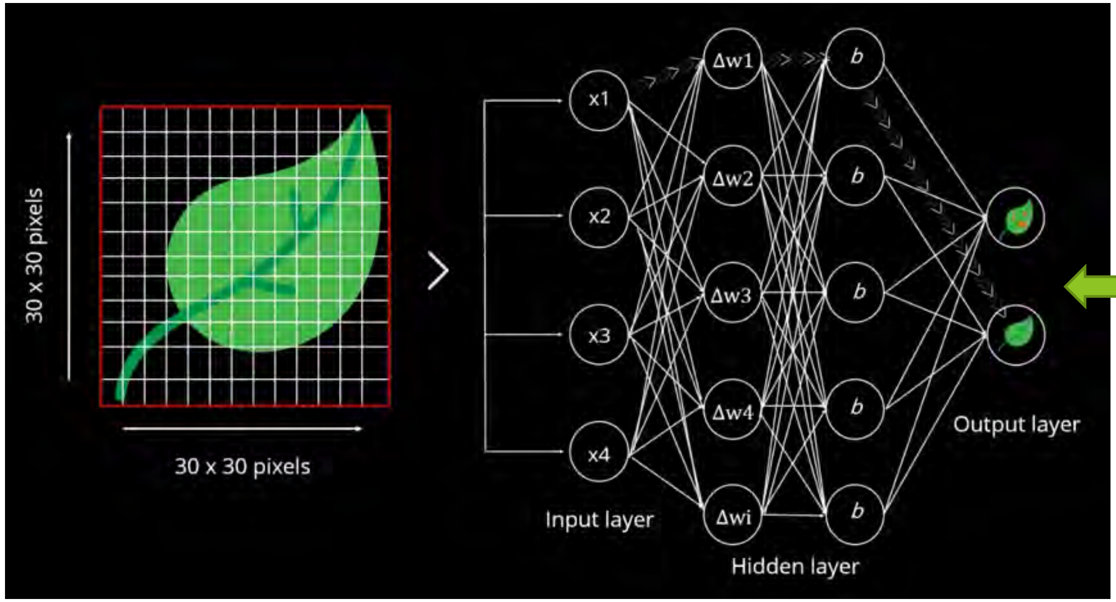
Neural Network



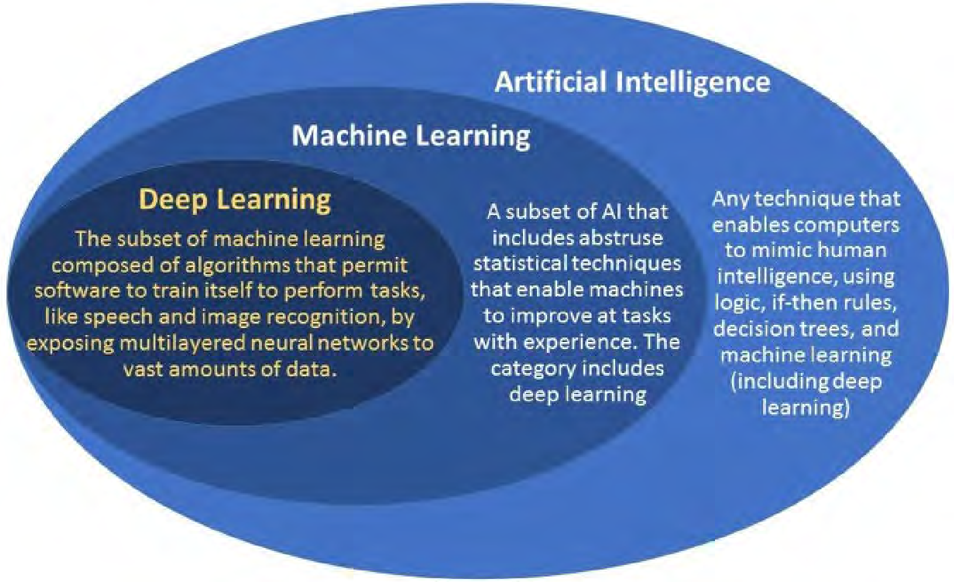
<u>Current Applications:</u>
Live Google Translate using camera



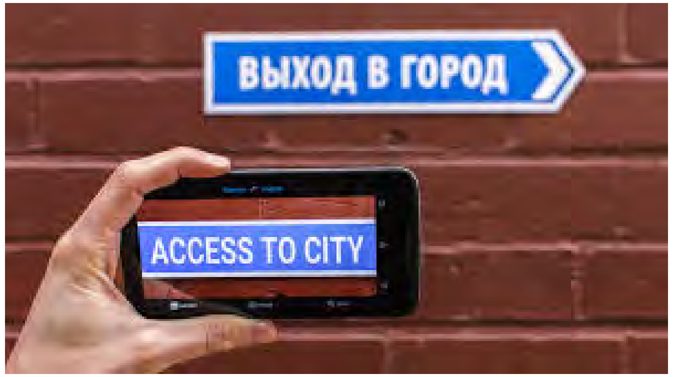
Phase 3: Neural Network



Neural Network



<u>Current Applications:</u>
Live Google Translate using camera



- Can we use neural networks to teach a reactor to make instant decisions?
- Can we make an AI based Instrumentation & Control system & replace people?
- Can we ever obtain an operating license of a fission battery from NRC?

Thank you!

What other technologies and development efforts are needed to achieve Autonomous Control → fission batteries?

Contact Information
Yasir Arafat
yasir.arafat@inl.gov
412-736-4886 (cell)
208-526-3074 (office)
https://www.linkedin.com/in/yasirarafatinl

Idaho National Laboratory

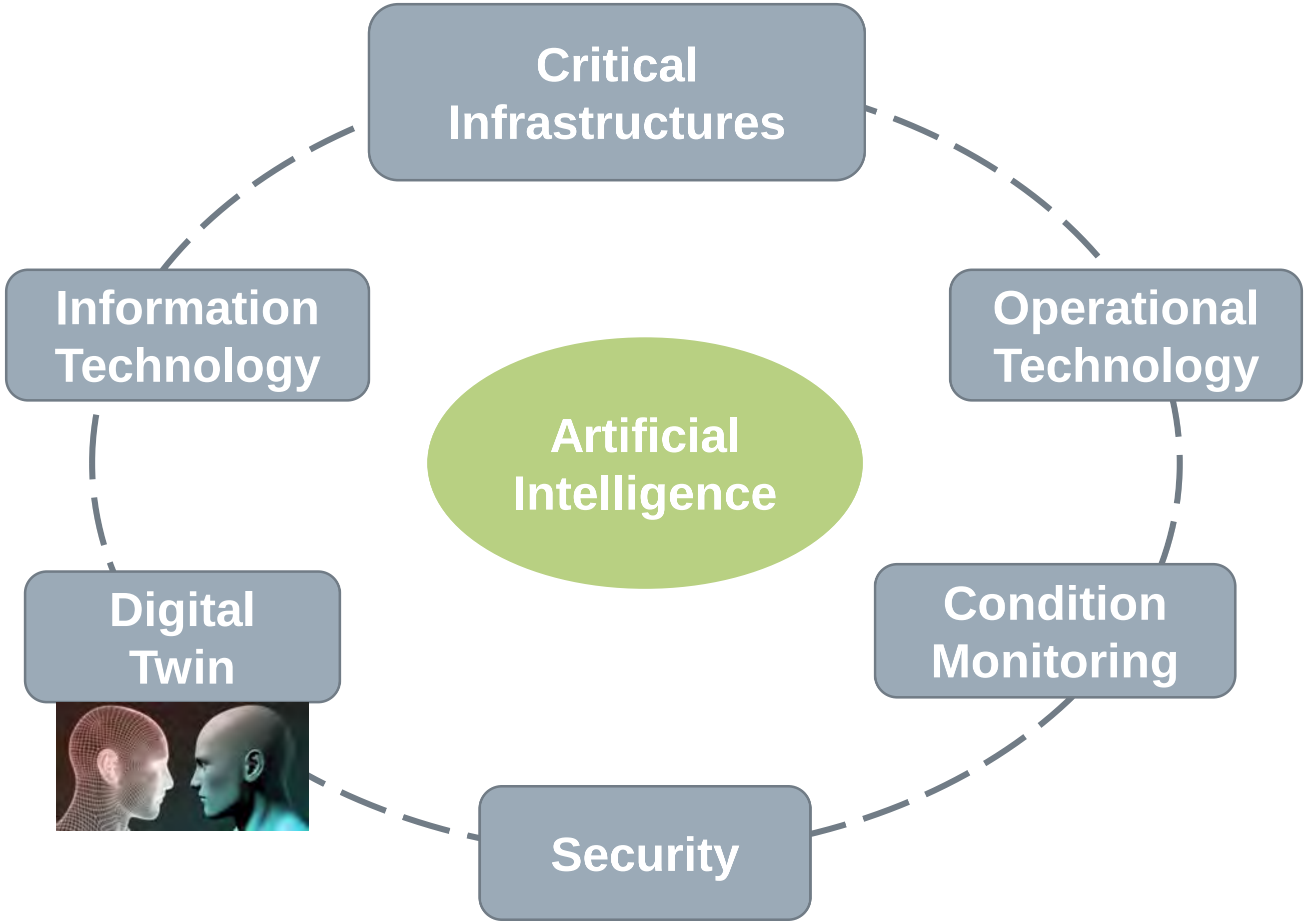
Covert Cognizance (C^2):

Novel Modeling and Monitoring Paradigm for Critical Systems

Hany Abdel-Khalik, Associate Professor,
School of Nuclear Engineering

Fission Battery Workshop, Jan 20, 2021

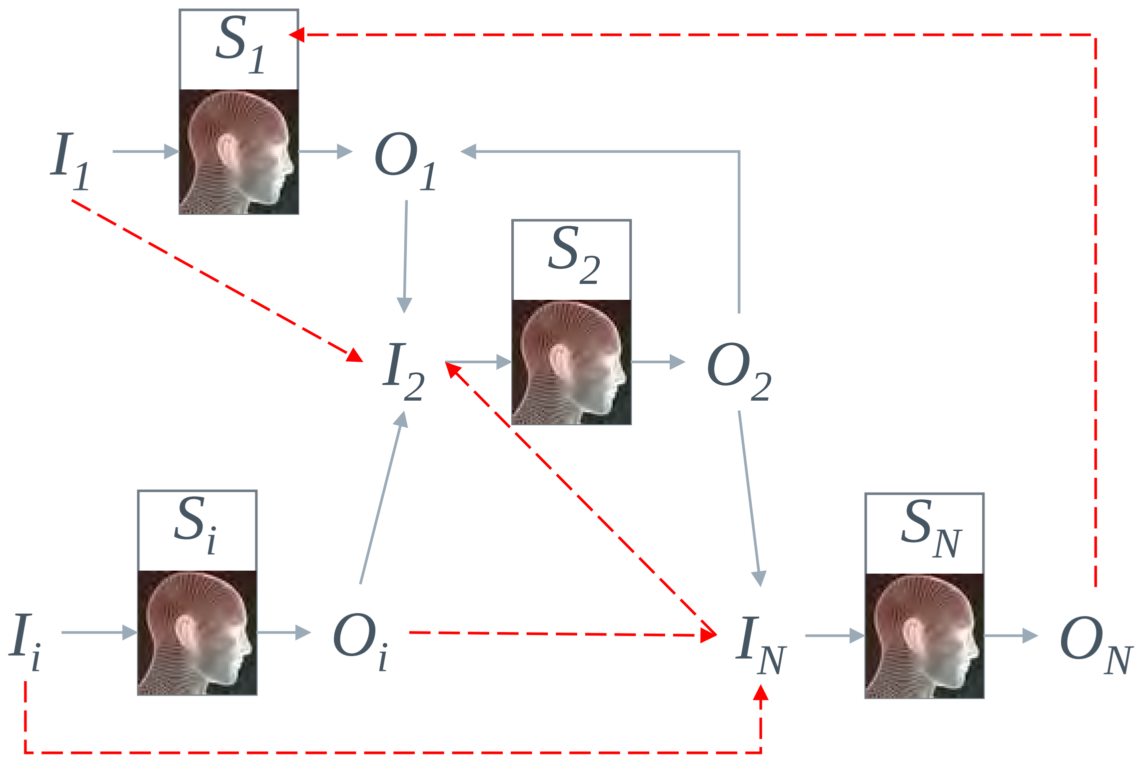
Computerized Decision Making Capability @ Center of 21st Science and Engineering Challenges



https://cipher.com/blog/the-16-sectors-of-critical-infrastructure-cybersecurity/
https://newsroom.cisco.com/feature-content?type=webcontent&articleId=1923920

C^2 Paradigm

- How to develop global self-awareness?
  - Pinpoints problems in a non-probabilistic manner.
  - Cannot be evaded by Adversarial AI



Current R&D efforts:

- Digital Twinning, providing unprecedented levels of details for diagnosis and control
  - Challenge: Digital Twins to have unavoidable uncertainties
  - Challenge: Modeling of critical systems is well-understood
- AI/ML, seeking to develop continuous learning platform for integrating digital twins models with measured data
  - Challenge: It is not clear when and how AI fails

C^2 Inspired by Active Monitoring

To find out what happens to a system when you interfere with it, you **have to interfere** with it (not just **passively** observe it).

George Box,
"Use and Abuse of Regression," Technometrics, Nov. 1966

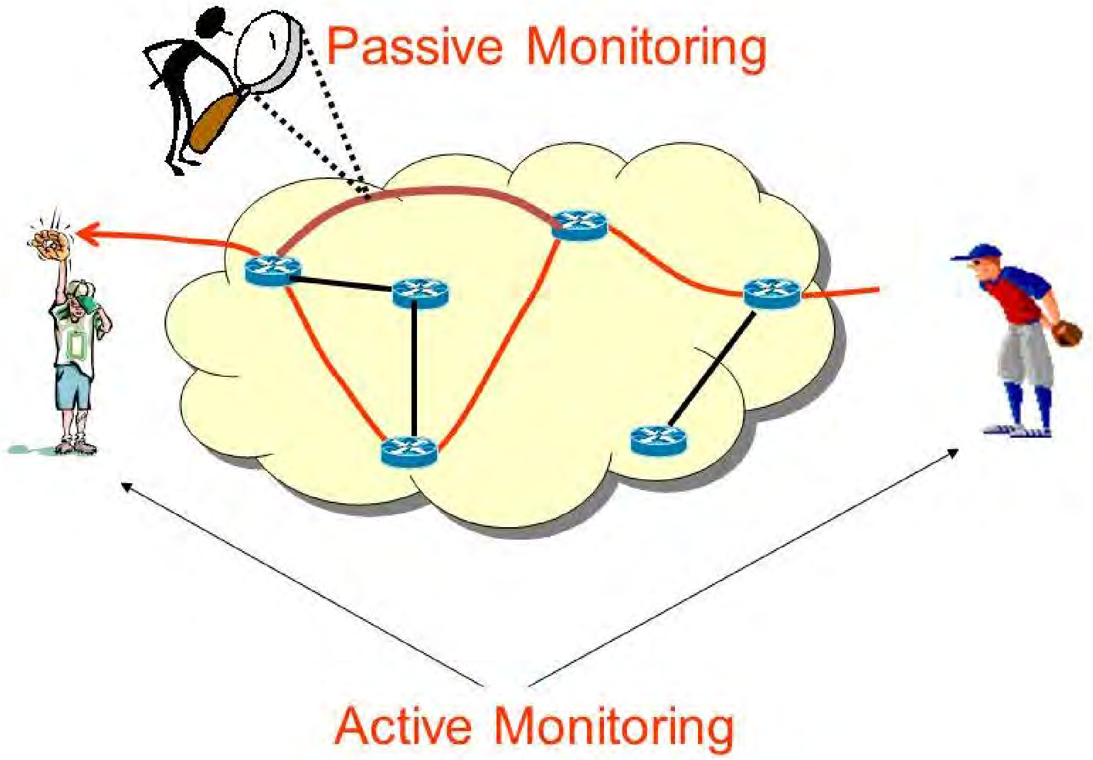
Passive vs. Active Monitoring

PASSIVE

Monitoring without Interfering

ACTIVE

Interfering for better Monitoring



Straightforward active monitoring algorithms are not suitable for unattended operation, because they were not designed with adversarial scenarios in mind.

https://slideplayer.com/slide/7395148/

State-of-the-art Monitoring vs. C^2 Paradigm

Survival Bias

Data analyzed are initially selected based on some unquestionable criterion

State-of-the-art
(Dominance) often selected as criterion for majority of AI techniques

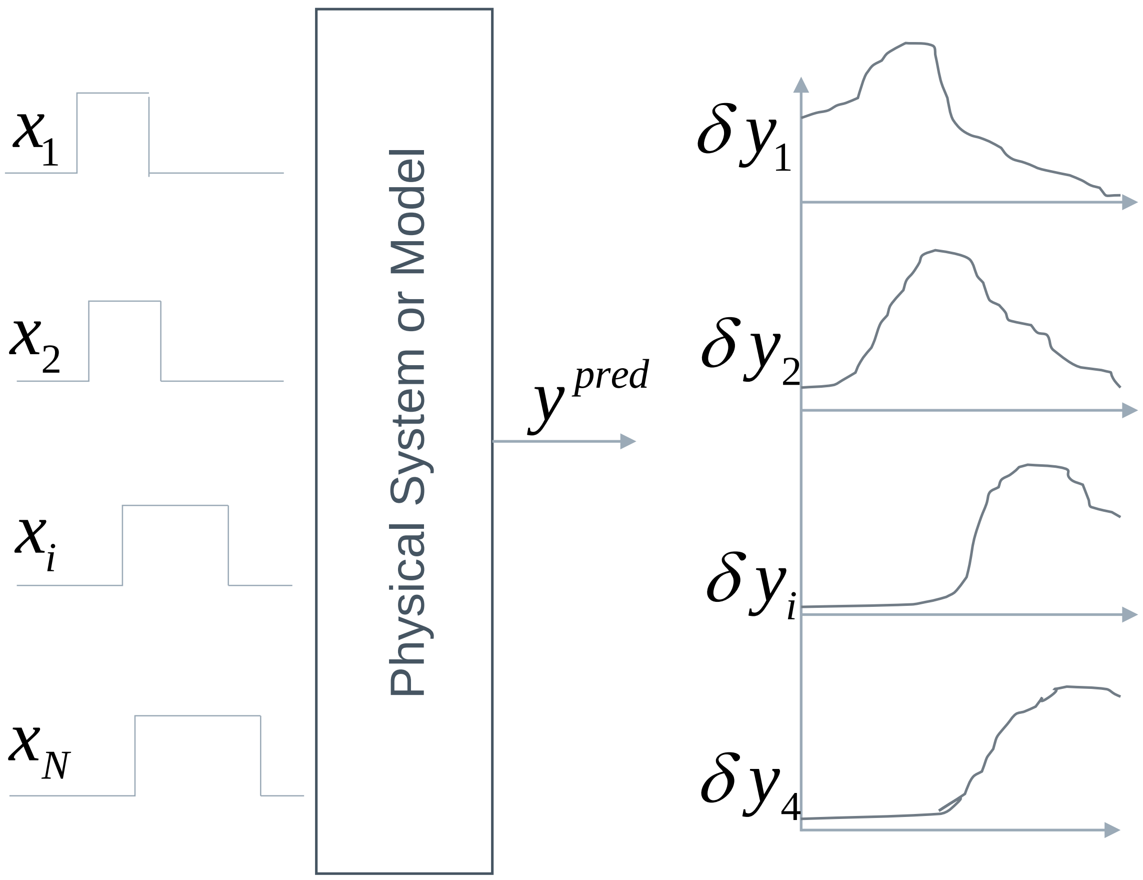
Patterns with weaker Impact (i.e, higher order patterns) recently proposed to improve classification-ability of AI techniques

Zero-Impact Patterns (Non-Patterns) treated as noise.

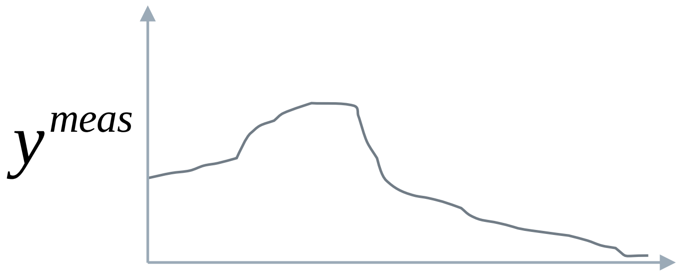
Provide huge space to store cognizance (self-awareness) information that blinds AI techniques

State-of-the-art Monitoring Paradigm

Discovery/Learning Mode



Inference Analysis



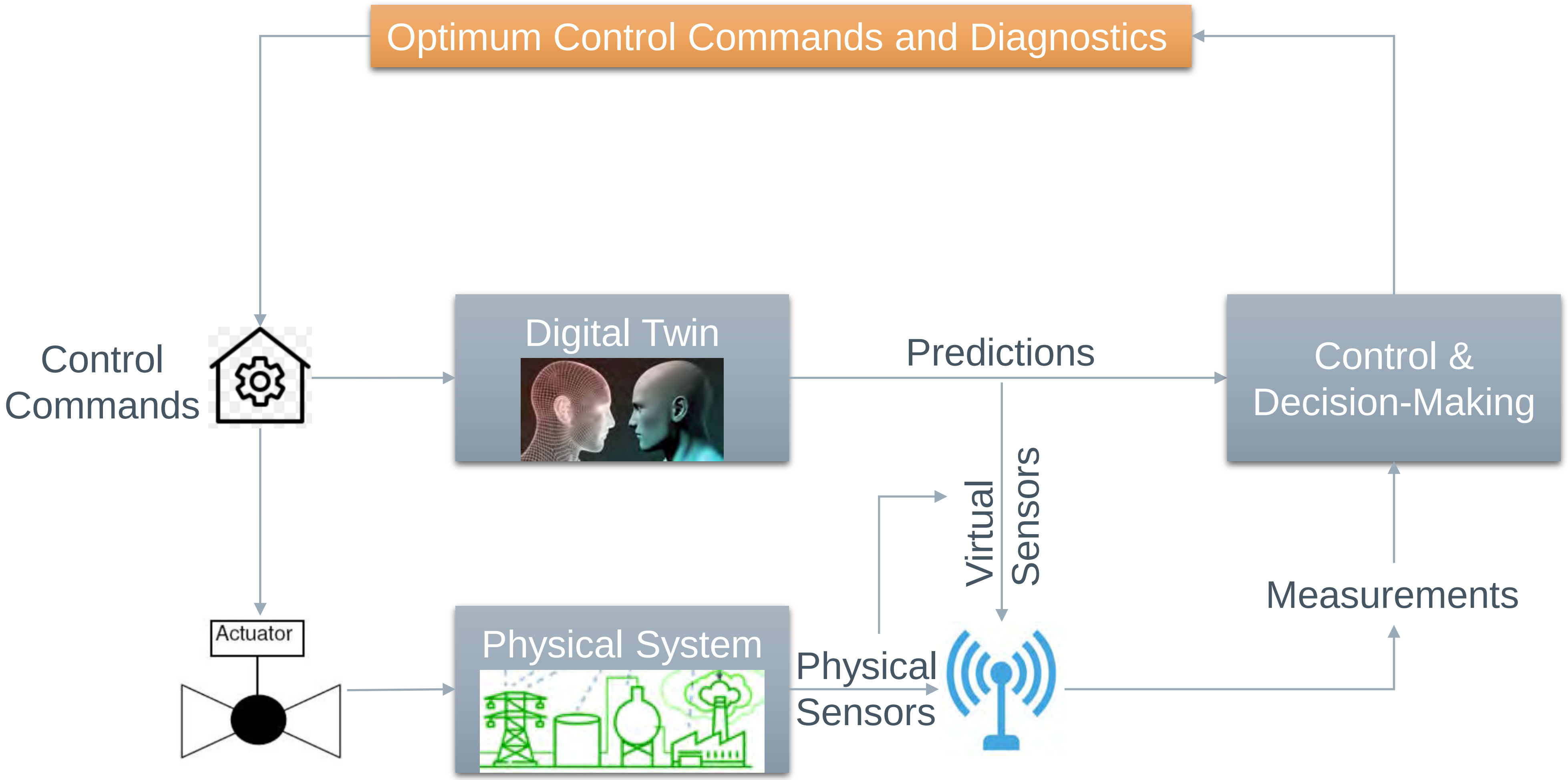
$$y^{meas} - y^{pred} = \sum \delta x_i^{inf} \delta y_i$$

Many solutions exist, described probabilistically

Mathematical criterion enforced to select single solution; criterion is system specific and cannot be generalized for multi-physics systems

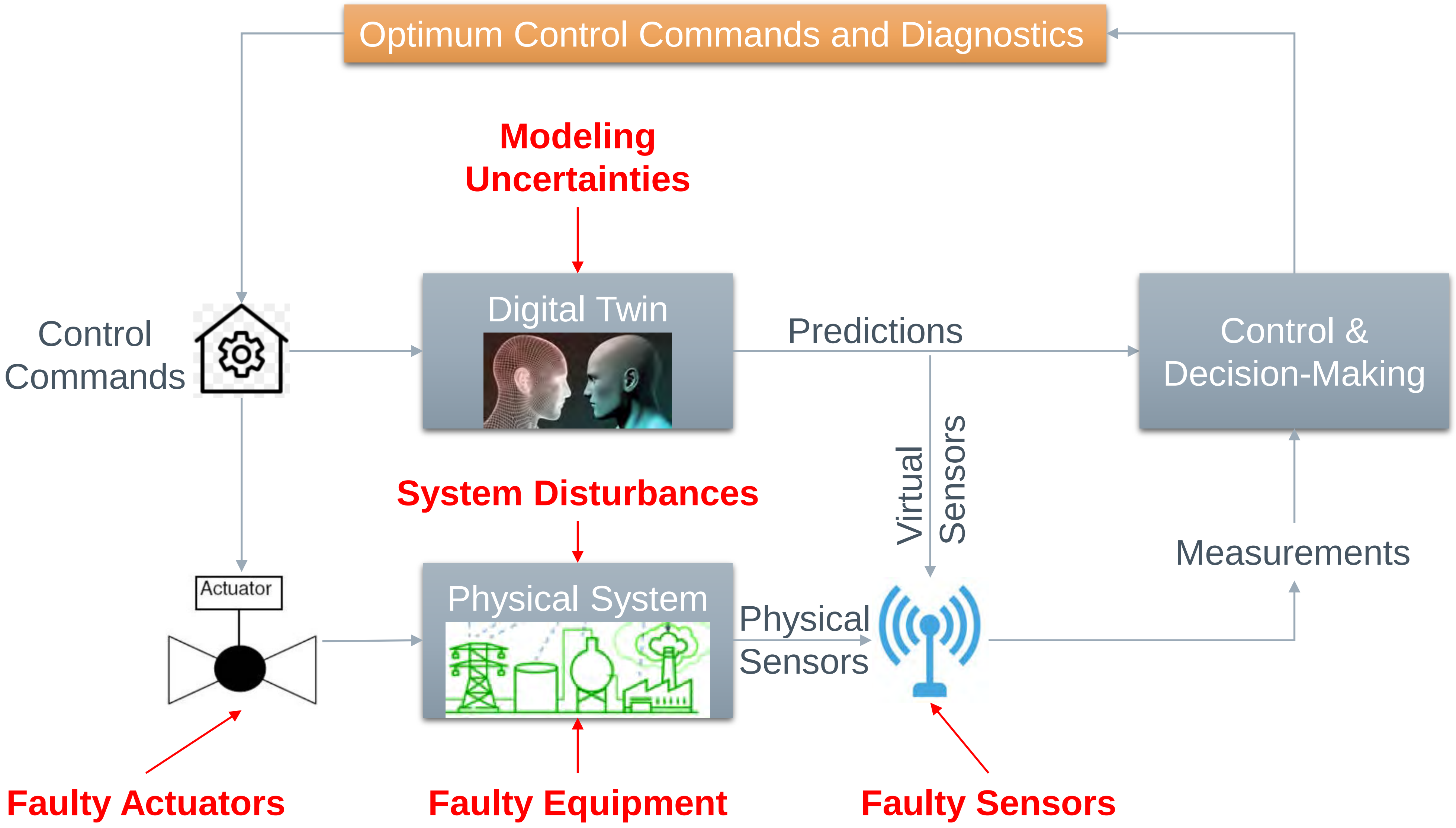
Superposition-based Inference/Discovery Algorithms limited in pinpointing cause-effect relationships

Unattended Operation: Problem Setup

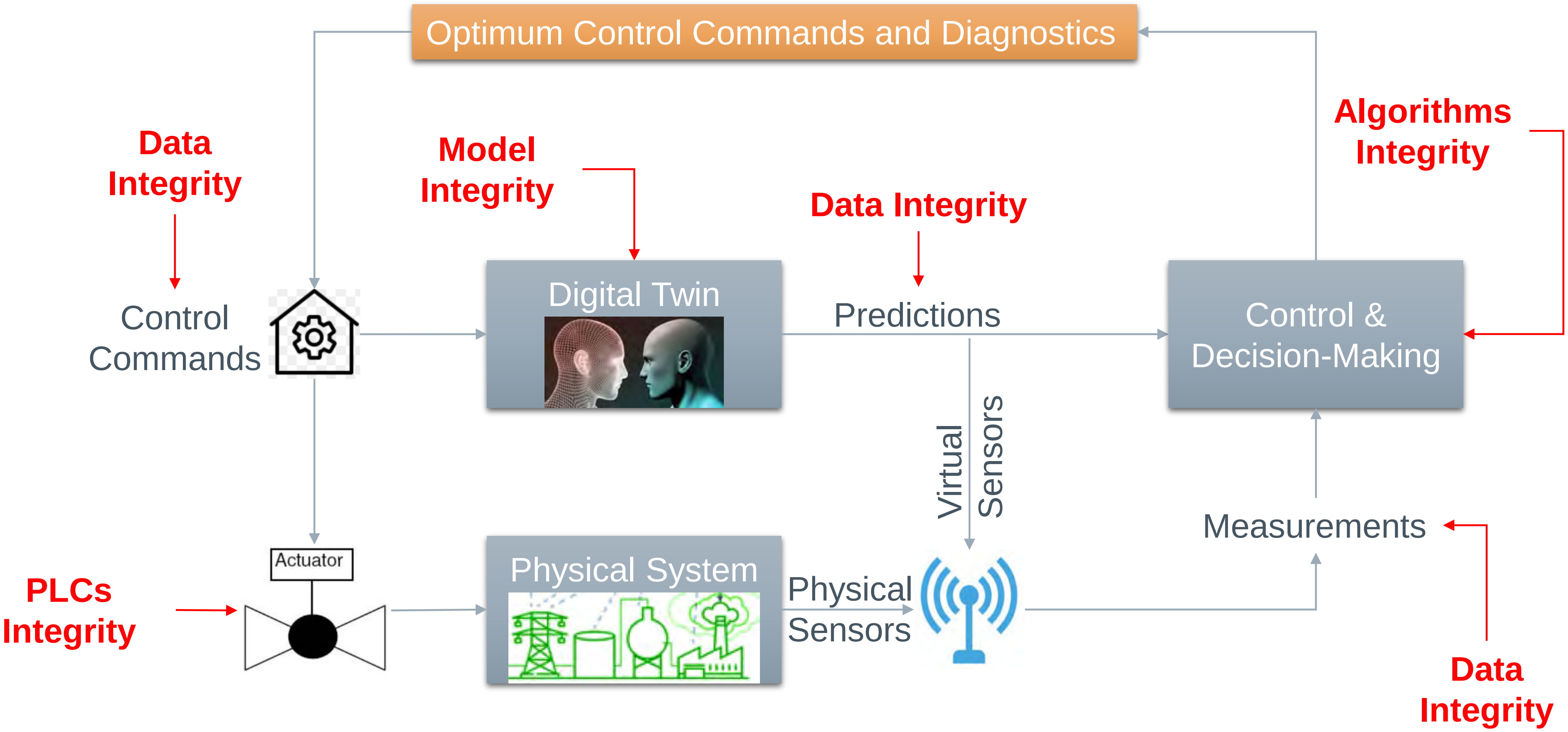


Images from RSC Publishing – Royal Society of Chemistry, EnggCyclopedia, Icons8, Free Icons Library

Unattended Operation – Uncertainty Challenges (1)



Unattended Operation – Integrity Challenges (2)



Unattended Operation – Uncertainty Challenges (1)

ΔK

Many Known (considered) Influencers

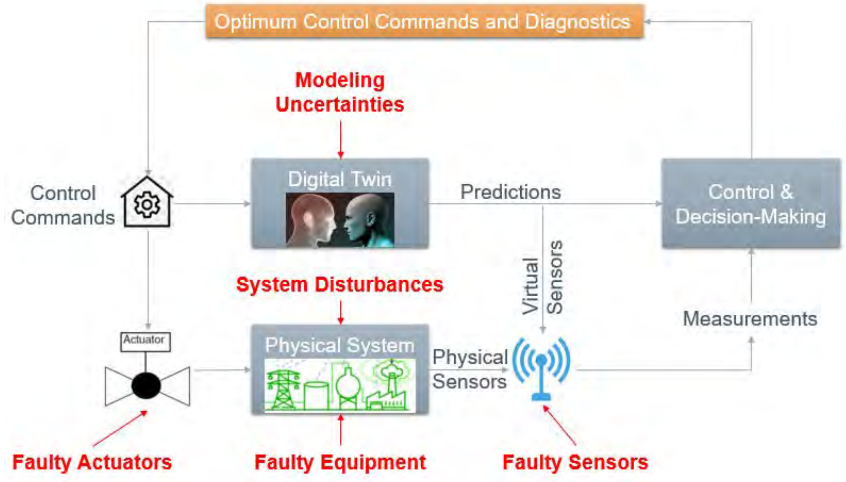
ΔU

Many Unknown (unconsidered) Influencers

Process Exhibiting Unknown (to-be-discovered) Behavior, e.g., Black-Box Model, Physical Process

ΔO

Observable



Unattended Operation – Uncertainty Challenges (1)

ΔK

Many Known (considered) Influencers

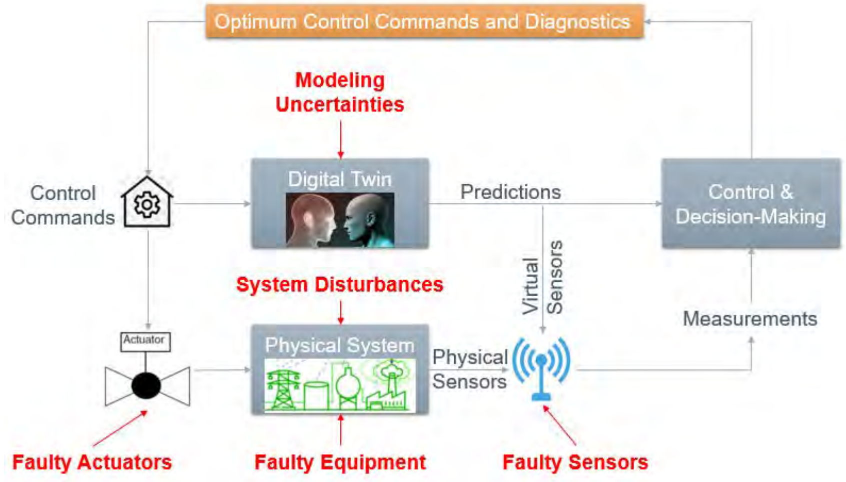
ΔU

Many Unknown (unconsidered) Influencers

Process Exhibiting Unknown (to-be-discovered) Behavior, e.g., Black-Box Model, Physical Process

ΔO

Observable



$$\Delta O \simeq \frac{\partial O}{\partial K}\Delta K + \frac{\partial O}{\partial U}\Delta U$$

Premise is that the *U* influences average out

Discovery Model

Unattended Operation – Uncertainty Challenges (1)

ΔK

Many Known (considered) Influencers

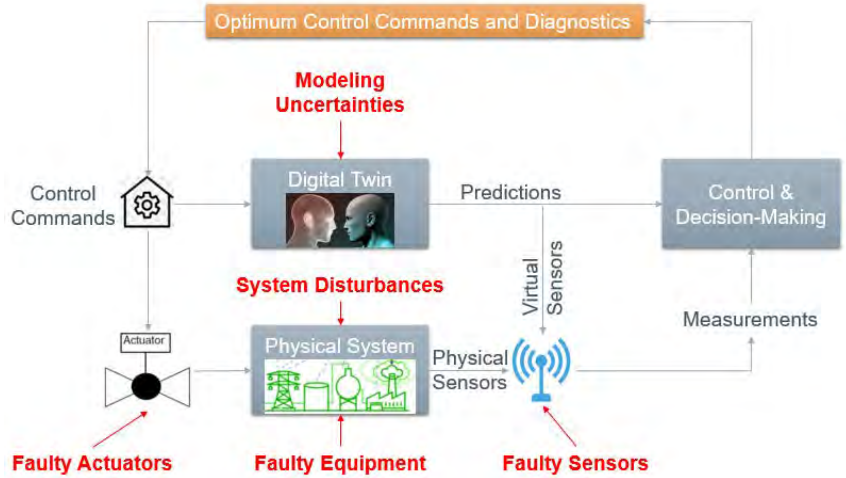
ΔU

Many Unknown (unconsidered) Influencers

Process Exhibiting Unknown (to-be-discovered) Behavior, e.g., Black-Box Model, Physical Process

ΔO

Observable



$$\Delta O \simeq \frac{\partial O}{\partial K}\Delta K + \frac{\partial O}{\partial U}\Delta U$$

Premise is that the *U* influences average out

Discovery Model

Forward Analysis

$$\Delta O \simeq \frac{\partial O}{\partial K}\Delta K + \varepsilon$$

Inference Analysis

Unattended Operation – Uncertainty Challenges (1)

ΔK

Many Known (considered) Influencers

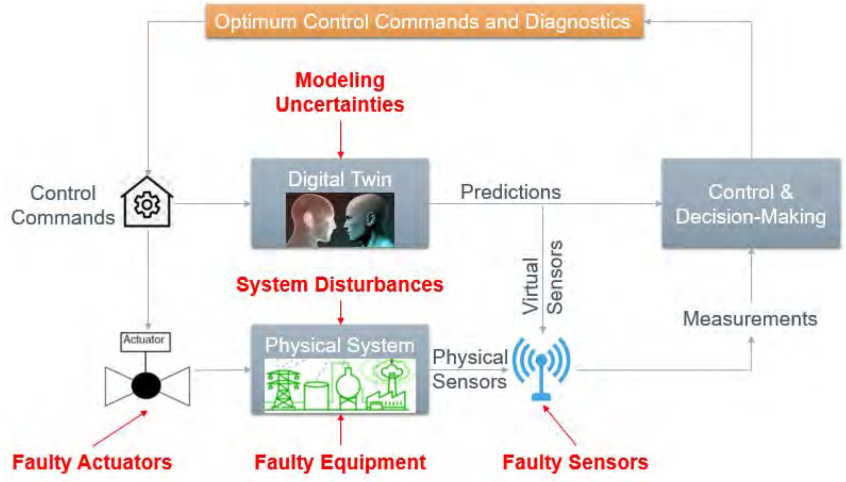
ΔU

Many Unknown (unconsidered) Influencers

Process Exhibiting Unknown (to-be-discovered) Behavior, e.g., Black-Box Model, Physical Process

ΔO

Observable



$$\Delta O \simeq \frac{\partial O}{\partial K}\Delta K + \frac{\partial O}{\partial U}\Delta U$$

Premise is that the *U* influences average out

Discovery Model

Forward Analysis

$$\Delta O \simeq \frac{\partial O}{\partial K}\Delta K + \varepsilon$$

Inference Analysis

Many plausible ΔK satisfy forward discovery model, forcing inference analyses to be probabilistic

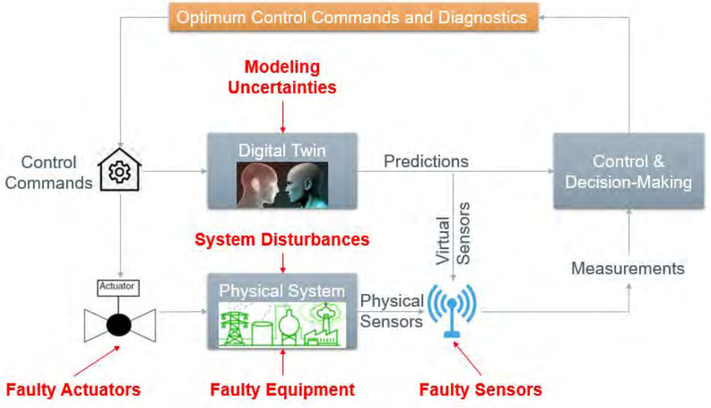
Unattended Operation – Uncertainty Challenges (1)

› Implications:

– Discovery Models are vulnerable to integrity attacks via careful data manipulation

– Inference analysis requires many samples (or high fidelity failure models) for high success rate (i.e., low FP/TN)

– Inference analysis performance more vulnerable to integrity attacks, decreasing its reliability for fault identification and isolation



Discovery/Learning

Probabilistic

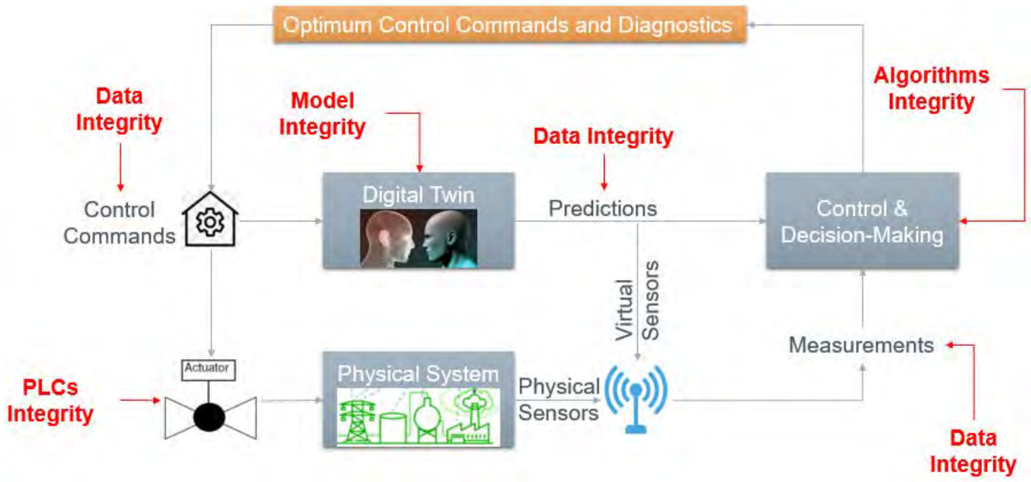
Fingerprinting

Unattended Operation – Integrity Challenges (2)

Encryption

Watermarking

Scrambling



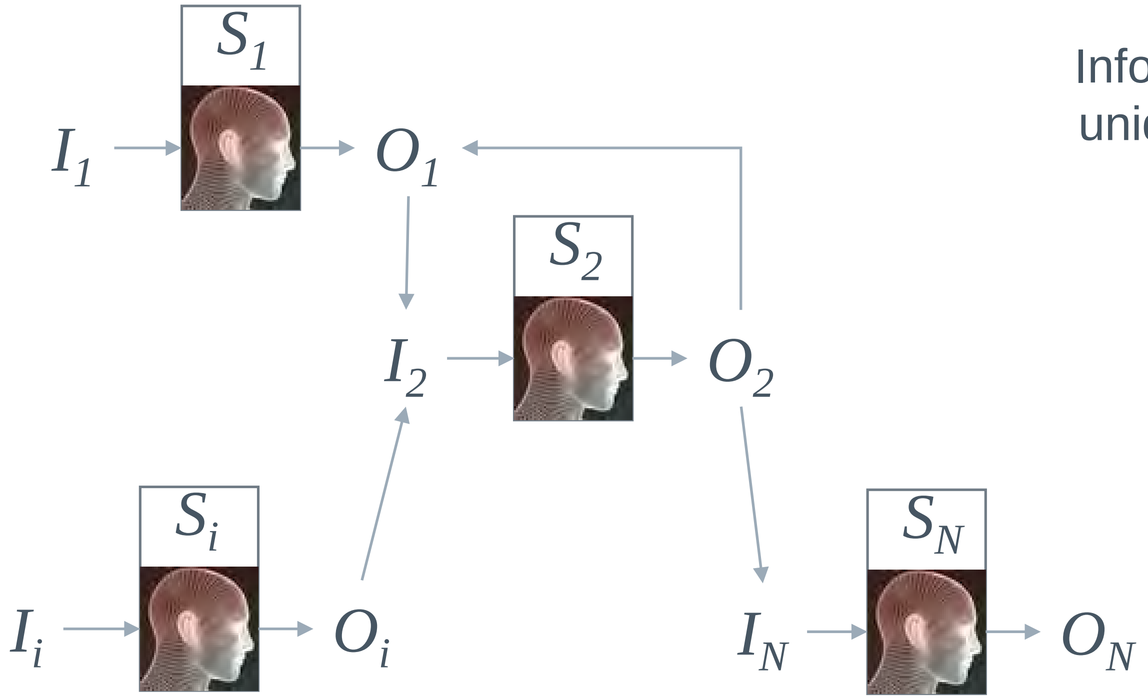
Paradigm Shift needed, shifting from overt access-prevention to covert zero-impact-while-under-attack methods

Gatekeeping

Fingerprint Auditing

Medium.com, PaintShop Pro, CyberExperts.com

Covert Cognizance (C^2) Paradigm
How to covertly develop global self-awareness?

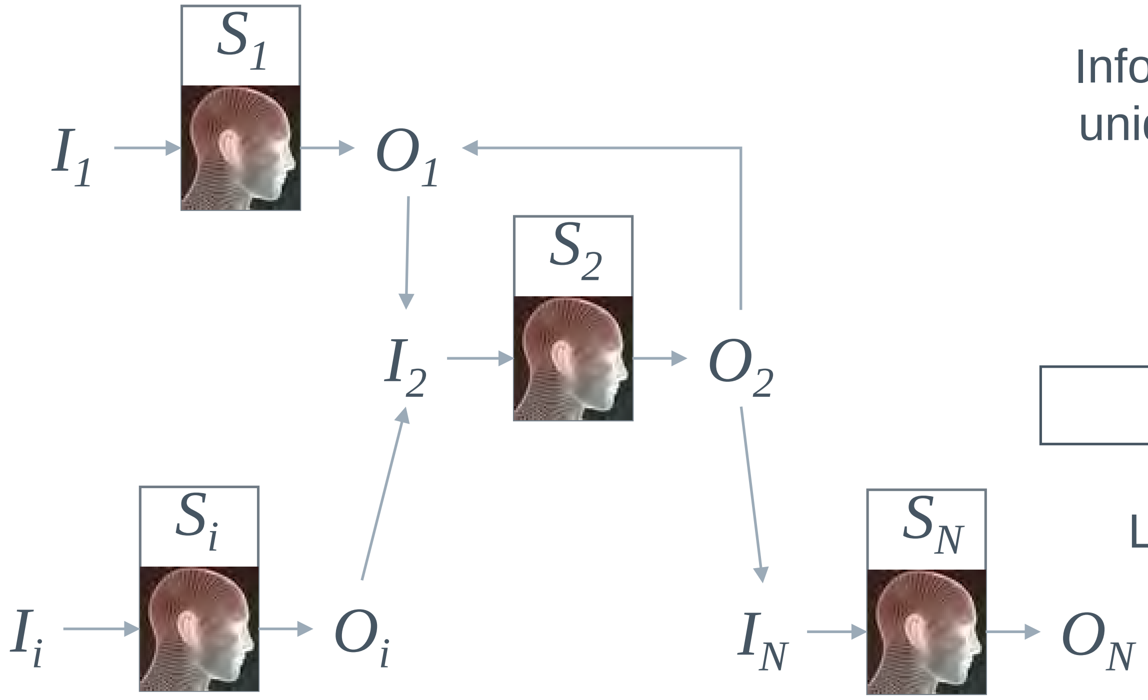


Global (System-wide) Awareness

Information about multiple sub-systems can be uniquely derived as features from a given sub-system state and its I/O data stream.

Covert Cognizance (C²) Paradigm

How to covertly develop global self-awareness?



Global (System-wide) Awareness

Information about multiple sub-systems can be uniquely derived as features from a given sub-system state and its I/O data stream.

Existing Awareness Paradigm

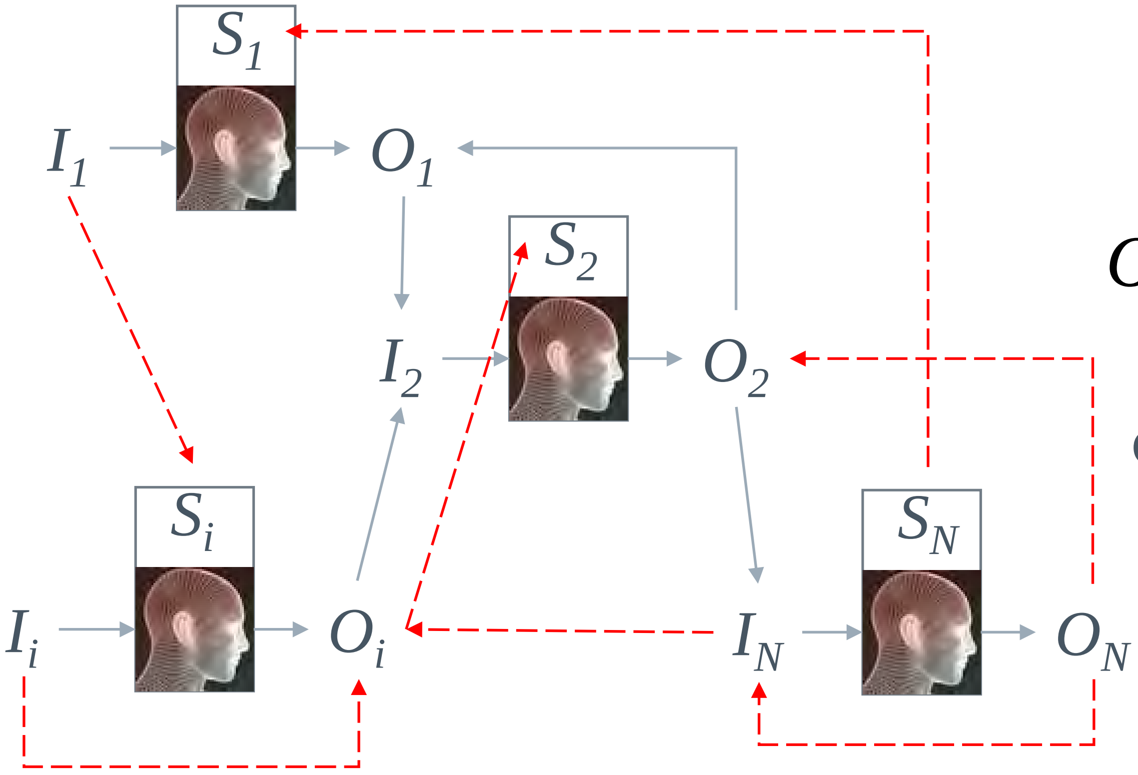
Local Awareness:

$$O_i = \Omega_i\left(S_i, I_i\right)$$

Existing paradigm extends local reach to multiple sub-systems via **Correlation-based AI**, forcing explainable, causal, inference analyses to be probabilistic

Covert Cognizance (C^2) Paradigm

How to covertly develop global self-awareness?



Global (System-wide) Awareness

Embeds delta terms consistent with noise/uncertainties and with zero impact

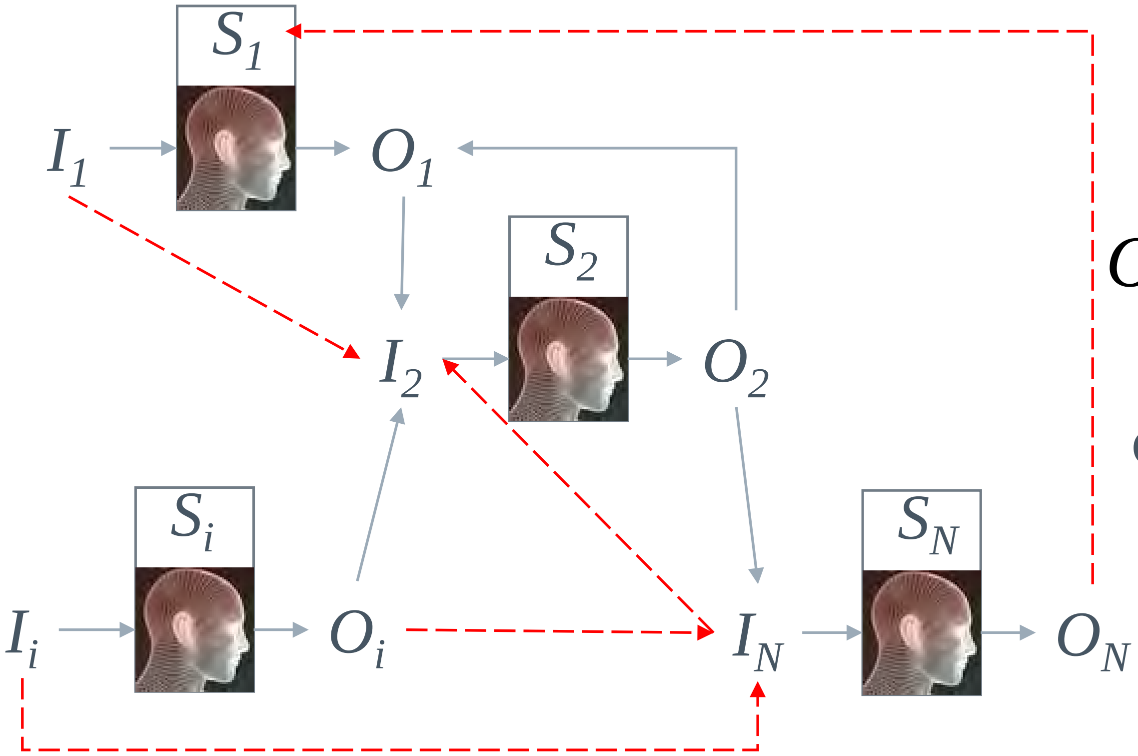
$$O_i + \Delta O_i = \Omega_i \left(S_i + \Delta S_i , I_i + \Delta I_i \right)$$

C^2 information randomly generated/channeled, requiring no additional variables, with same entropy, and undiscoverable via AI

Decoy C^2 information (designed to be discoverable via AI) to track attackers and determine their goals.

Covert Cognizance (C^2) Paradigm

How to covertly develop global self-awareness?



Global (System-wide) Awareness

Embeds delta terms consistent with noise/uncertainties and with zero impact

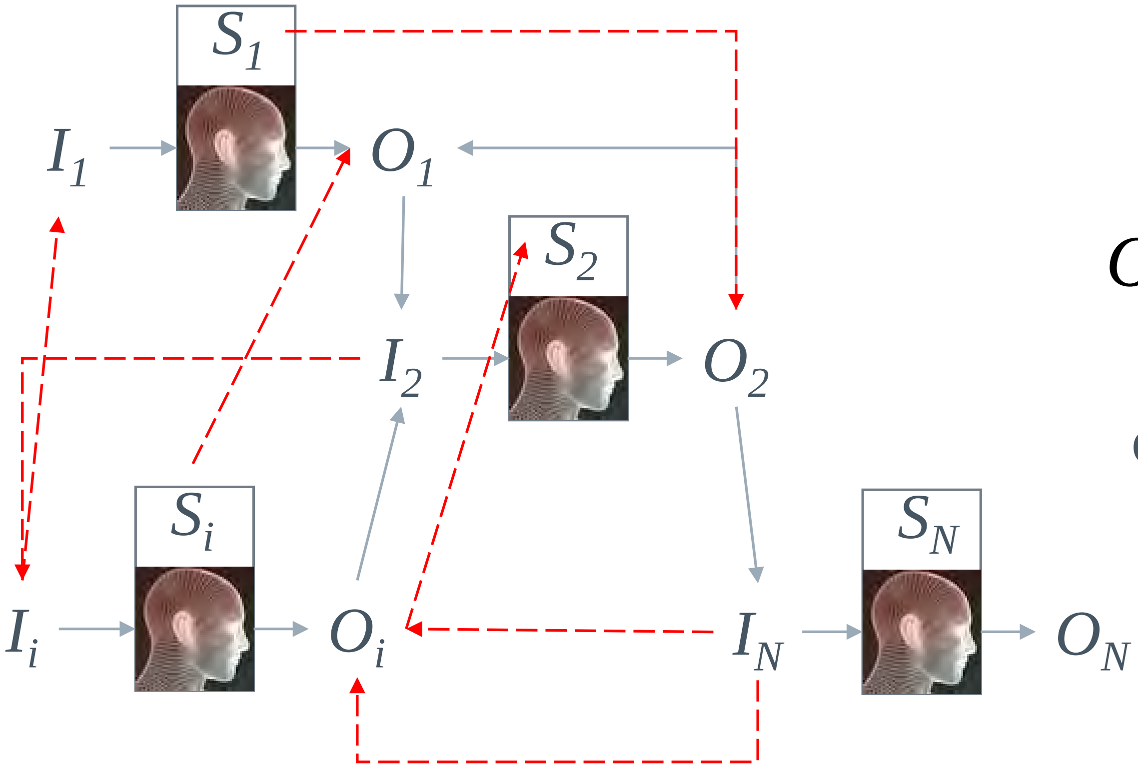
$$O_i + \Delta O_i = \Omega_i \left(S_i + \Delta S_i , I_i + \Delta I_i \right)$$

C^2 information randomly generated/channeled, requiring no additional variables, with same entropy, and undiscoverable via AI

Decoy C^2 information (designed to be discoverable via AI) to track attackers and determine their goals.

Covert Cognizance (C^2) Paradigm

How to covertly develop global self-awareness?



Global (System-wide) Awareness

Embeds delta terms consistent with noise/uncertainties and with zero impact

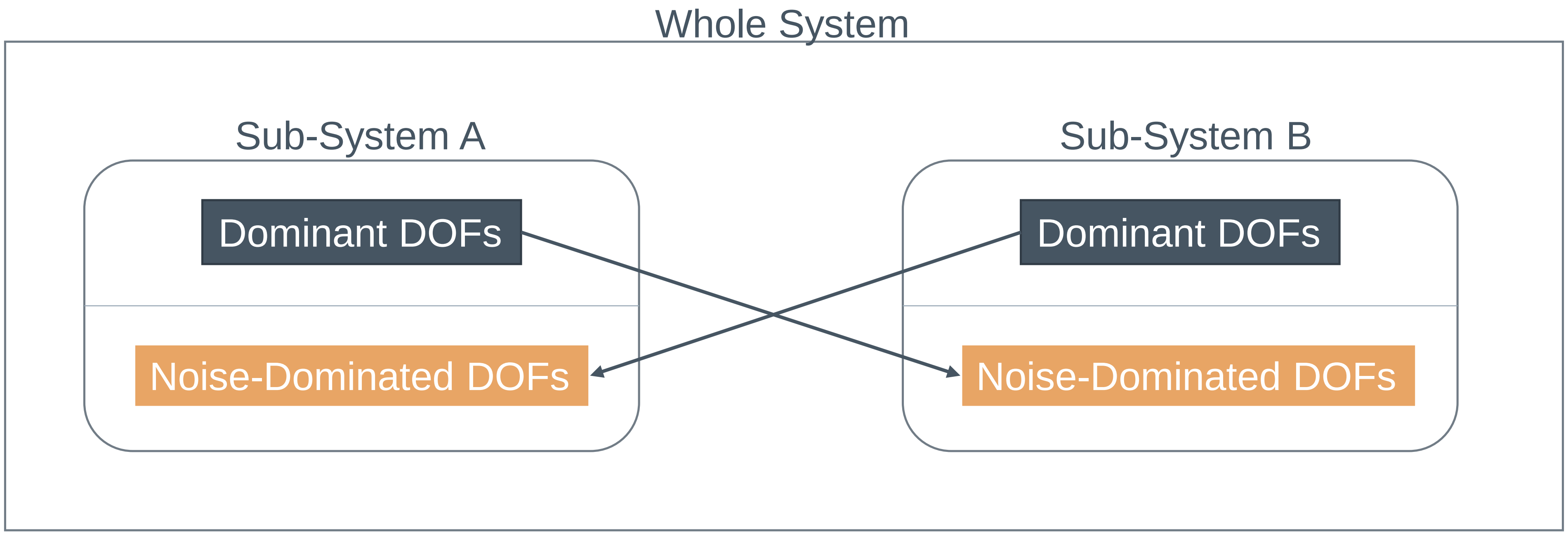
$$O_i + \Delta O_i = \Omega_i \left(S_i + \Delta S_i, I_i + \Delta I_i \right)$$

C^2 information randomly generated/channeled, requiring no additional variables, with same entropy, and undiscoverable via AI

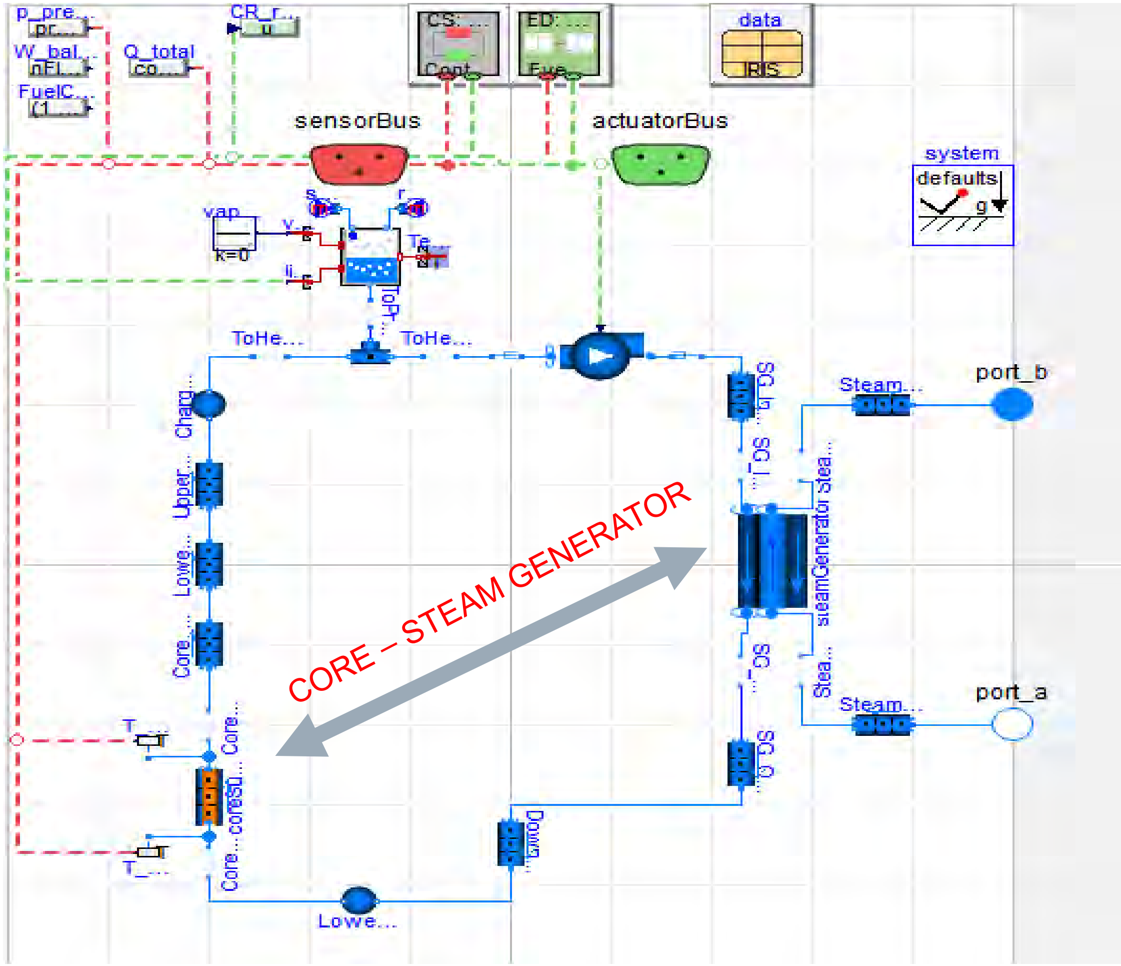
Decoy C^2 information (designed to be discoverable via AI) to track attackers and determine their goals.

Why C² Possible?

- Complex Systems are reducible, implying that:
  - Dominant behavior can be described using small no. DOFs,
  - Leaving huge number of “un-used” noise-dominated DOFs, that can serve as carrier variables



Covert Cognizance (C^2) Example

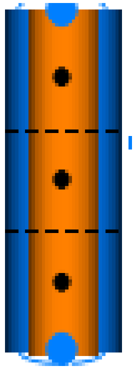


Generated using Dymola

- **Cognizance:** The core and the SG are mutually aware of each other state functions, i.e., they carry information about each other.
- **Covertness:** The information is embedded in the process variables via randomized mathematical transformations
  - Embedded in real-time along noisy non-observable components for zero-impact on system state and control strategies.
  - One-time pad representation immune to AI learning.

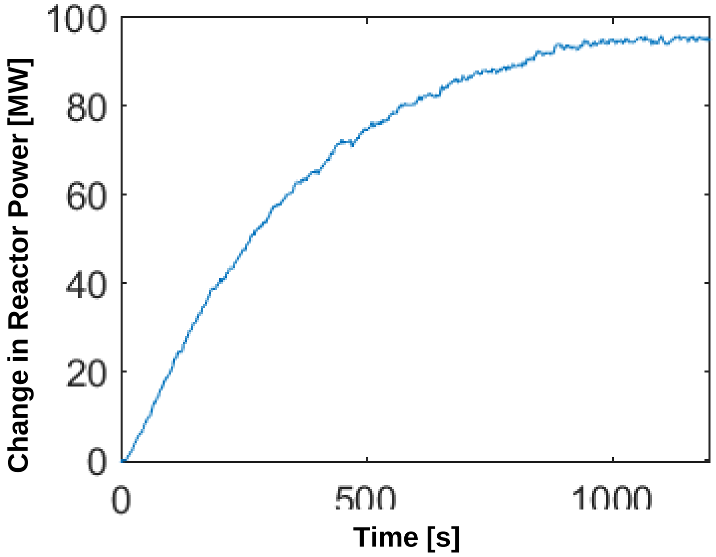
Covert Cognizance Example

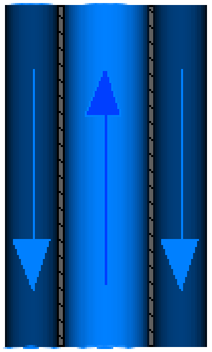
Core Dynamics



$$\frac{d}{dt}\begin{bmatrix}\delta P(t)\\ \delta C(t)\\ \delta T_f(t)\\ \delta T_m(t)\end{bmatrix}=\begin{bmatrix}-\frac{\beta}{\tau_p} & \lambda & \frac{\beta}{\tau_p}P_0\alpha_f & \frac{\beta}{\tau_p}P_0\alpha_m\\ \frac{\beta}{\tau_p} & -\lambda & 0 & 0\\ \frac{1}{\rho_f c_f A_r V_{cell}} & 0 & -\frac{1}{\tau_{hx}} & \frac{1}{\tau_{hx}}\\ 0 & 0 & \frac{LNR}{\rho_m c_m V_m} & \frac{-LNR-2\dot{m}c_m}{\rho_m c_m V_m}\end{bmatrix}\begin{bmatrix}\delta P(t)\\ \delta C(t)\\ \delta T_f(t)\\ \delta T_m(t)\end{bmatrix}+\begin{bmatrix}0\\ 0\\ 0\\ \frac{2\dot{m}c_m}{\rho_m c_m V_m}\end{bmatrix}\delta T_{in}(t)$$

Core Response





SG Dynamics

$$\rho_{pr}c_{pr}V_{pr}\frac{d\delta T_{pr}}{dt}=c_{pr}W_{pr}\widehat{\delta T_m}-\alpha_{pr}A_{pr}(\delta T_{pr}-\delta T_{me})$$

$$\rho_{me}c_{me}V_{me}\frac{d\delta T_{me}}{dt}=\alpha_{pr}A_{pr}\delta T_{pr}-(\alpha_{pr}A_{pr}+\alpha_{se}A_{se})\delta T_{me}+\alpha_{se}A_{se}\frac{\partial T_{sat}}{\partial P}\delta Pr$$

$$\frac{k_4}{v_d}\frac{d\delta L_d}{dt}+\frac{x_r V_{st}}{v^{p2}}\frac{\partial v^p}{\partial Pr}\frac{d\delta Pr}{dt}=-\frac{W_{st}}{x_M^2}\delta x_M-\delta W_d$$

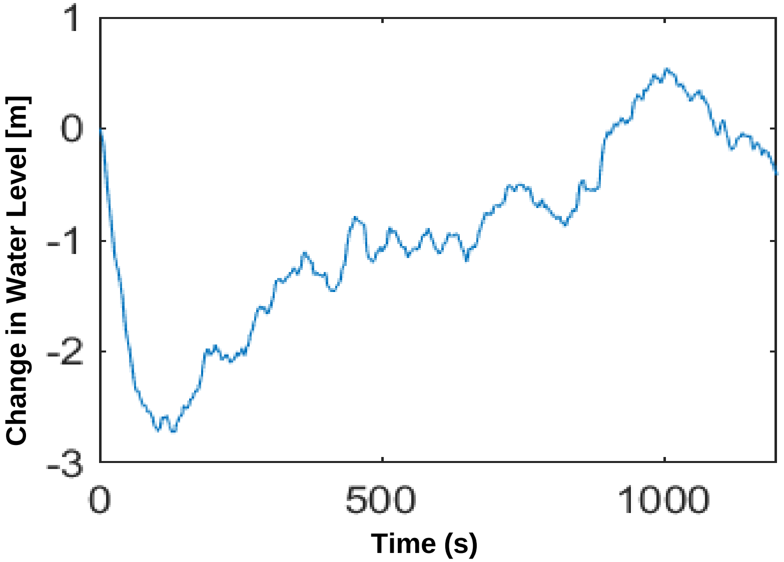
$$\frac{h_d k_4}{v_d}\frac{d\delta L_d}{dt}+\frac{V_d}{v_d}\frac{d\delta h_d}{dt}+\frac{h^f x_r V_d}{v^{p2}}\frac{\partial v^p}{\partial Pr}\frac{d\delta Pr}{dt}=-\frac{h^f W_{st}}{x_M^2}\delta x_M+x_r W_{st}\frac{\partial h^f}{\partial P}\delta Pr-h_d\delta W_d-W_d\delta h_d$$

$$-\frac{k_1 V_s}{v_s^2 v_g}\frac{d\delta x_M}{dt}-\frac{V_s}{v_s^2}\left(\frac{\partial v^f}{\partial Pr}+k_1 x_M\frac{\partial v_g}{\partial Pr}\right)\frac{d\delta Pr}{dt}=\delta W_d+\frac{W_{st}}{x_M^2}\delta x_M$$

$$\frac{k_1 V_s}{v_s}\left(r-\frac{h_s v_g}{v_s}\right)\frac{d\delta x_M}{dt}-\frac{V_s}{v_s}\left[\frac{h_s}{v_s}\left(\frac{\partial v^f}{\partial Pr}+k_1 x_M\frac{\partial v_g}{\partial Pr}\right)-\frac{\partial h^f}{\partial Pr}-k_1 x_M\frac{\partial r}{\partial Pr}\right]\frac{d\delta Pr}{dt}=$$
$$\alpha_{se}A_{se}\delta T_{me}+W_d\delta h_d+h_d\delta W_d+\frac{W_{st}}{x_M^2}h^f\delta x_M-\left[\frac{W_{st}}{x_M}\left(\frac{\partial h^f}{\partial Pr}+x_M\frac{\partial r}{\partial Pr}\right)+\alpha_{se}A_{se}\frac{\partial T_{sat}}{\partial Pr}\right]\delta Pr$$

$$\delta W_d=\frac{k_8}{2}\left(\frac{L_d}{v_d}-\frac{L_w}{v^f}\right)^{-\frac{1}{2}}\left\{\frac{\delta L_d}{v_d}+\frac{V_s}{A_w v_s}\left[1+k_1(1-x_M)\frac{v_g}{v_s}\right]\delta x_M+\frac{V_s}{A_w}\frac{(1-x_M)}{v_s^2}\left(\frac{\partial v^f}{\partial Pr}+k_1 x_M\frac{\partial v_g}{\partial Pr}\right)\delta Pr\right\}$$

SG Response



Other Applications for C^2

- Allow software to develop cognizance about its own execution history
- Employ C^2 to stop software reverse-engineering
- Develop born-secured ROM models

Covert Cognizance (C^2) Paradigm

- Paradigm to develop global (system-wide) self-awareness in a covert manner without impacting system performance
  - Relies on active rather than passive monitoring
  - Fingerprint-based vs. Probabilistic-Correlation-based Awareness
  - Embedding is a form of "active interference"; however ROM research proved that complex systems have too many redundant noise-dominated degrees-of-freedom (denoted by non-patterns), representing perfect carrier of C^2 information.
  - Embedding C^2 information along non-patterns ensures zero system impact, does not require additional carrier variable, ensures non-discoverability via Adversarial Intelligence

Acknowledgement

- The ideas presented have been inspired/supported by R&D work sponsored by several institutions over past five years, including
  - Sandia National Laboratory
  - Department of Energy, NEUP
  - Army Research Lab
  - Idaho National Laboratory

Kairos Power

Dispatchable, base-load nuclear: The case for a fission thermal battery

DR. ANTHONIE CILLIERS

JANUARY 2020

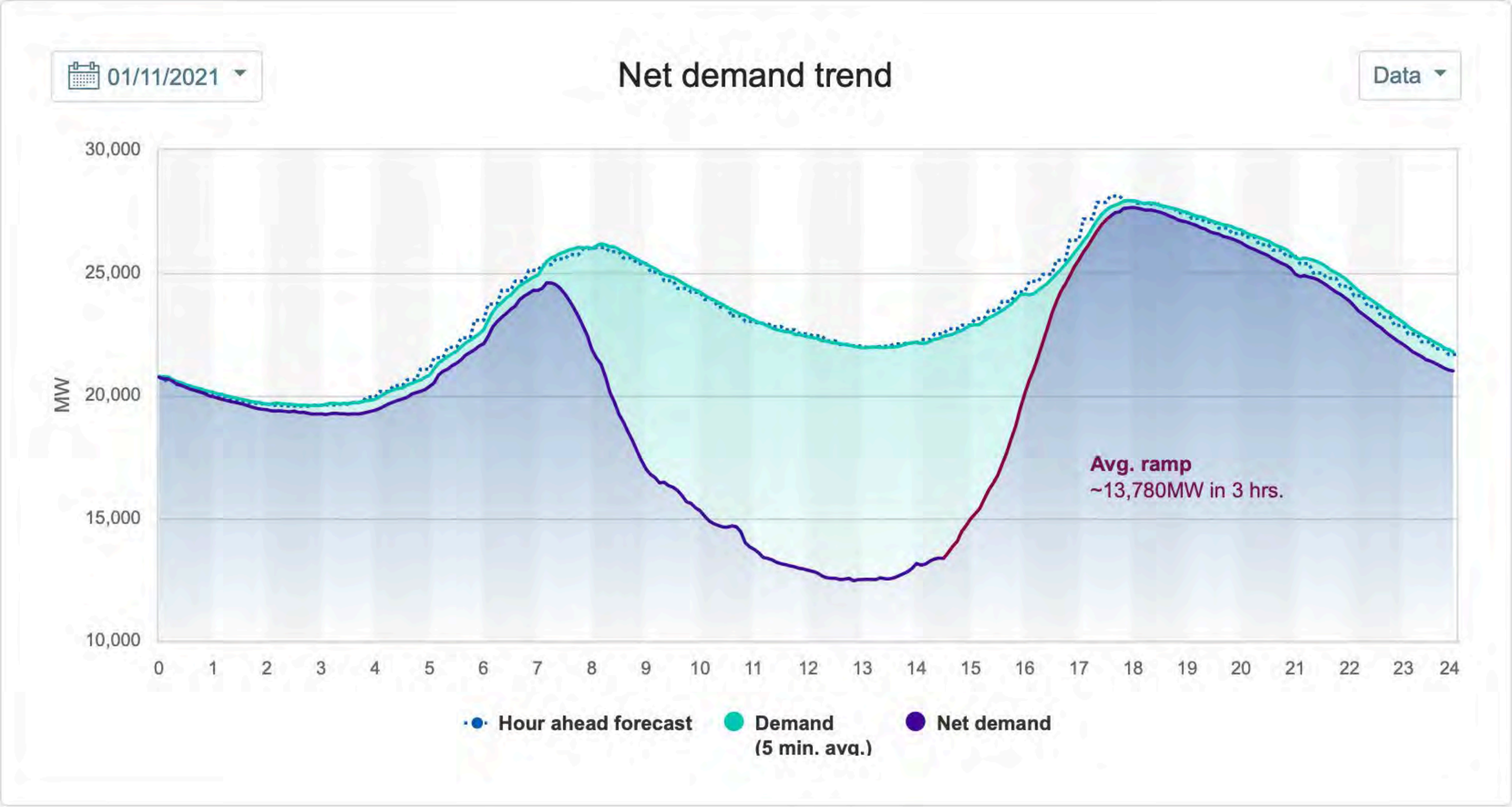
Kairos Power’s mission is to enable the world’s transition to clean energy, with the ultimate goal of dramatically improving people’s quality of life while protecting the environment.

In order to achieve this mission, we must prioritize our efforts to focus on a clean energy technology that is *affordable* and *safe*.

Conventional Nuclear Power Plant

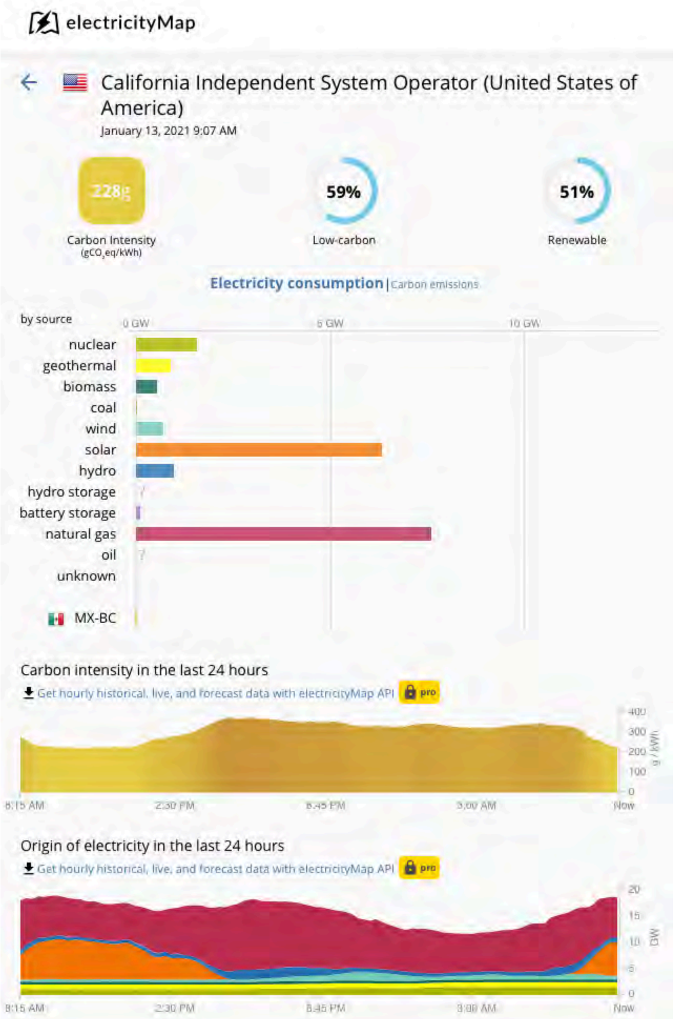
- High Capital cost, long construction times incurring interest during construction.
- Low fuel cost
- Perfect for baseload power supply
- High capacity-factor of ~92% provides optimal Levelized Cost of Electricity (LCOE) for High Capex low Opex power plants.
- Allows power ramping of up to 10% per minute.
- Why energy storage or a fission battery?

Net Demand: California ISO



California 24-hour electricity supply sources

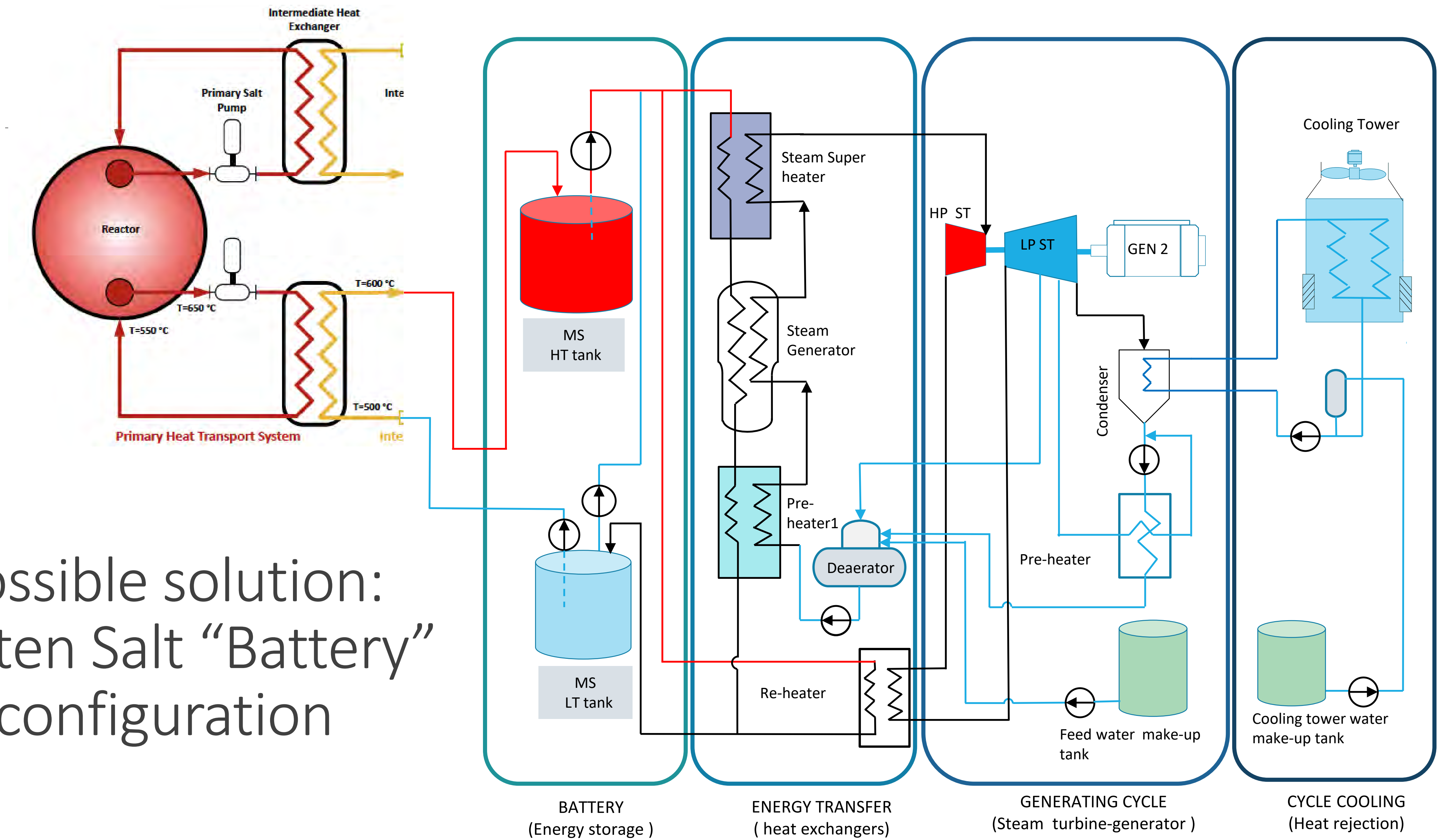
- Nuclear power provides baseload supply of 1.56GW (6-12% of demand)
- Geothermal provides baseload supply of 879MW (4-7% of demand)
- Biomass provides baseload supply of 535MW (2-4% if demand)
- Hydro provides flexible supply (5-10% of demand)
- Wind power provides intermittent seasonal supply
- Solar power provides variable supply during the day peaking at up to 43% of demand.
- Natural gas fills in the gaps up to 73% of demand during peak low solar times.
- What are the low carbon alternatives to fill the gaps?



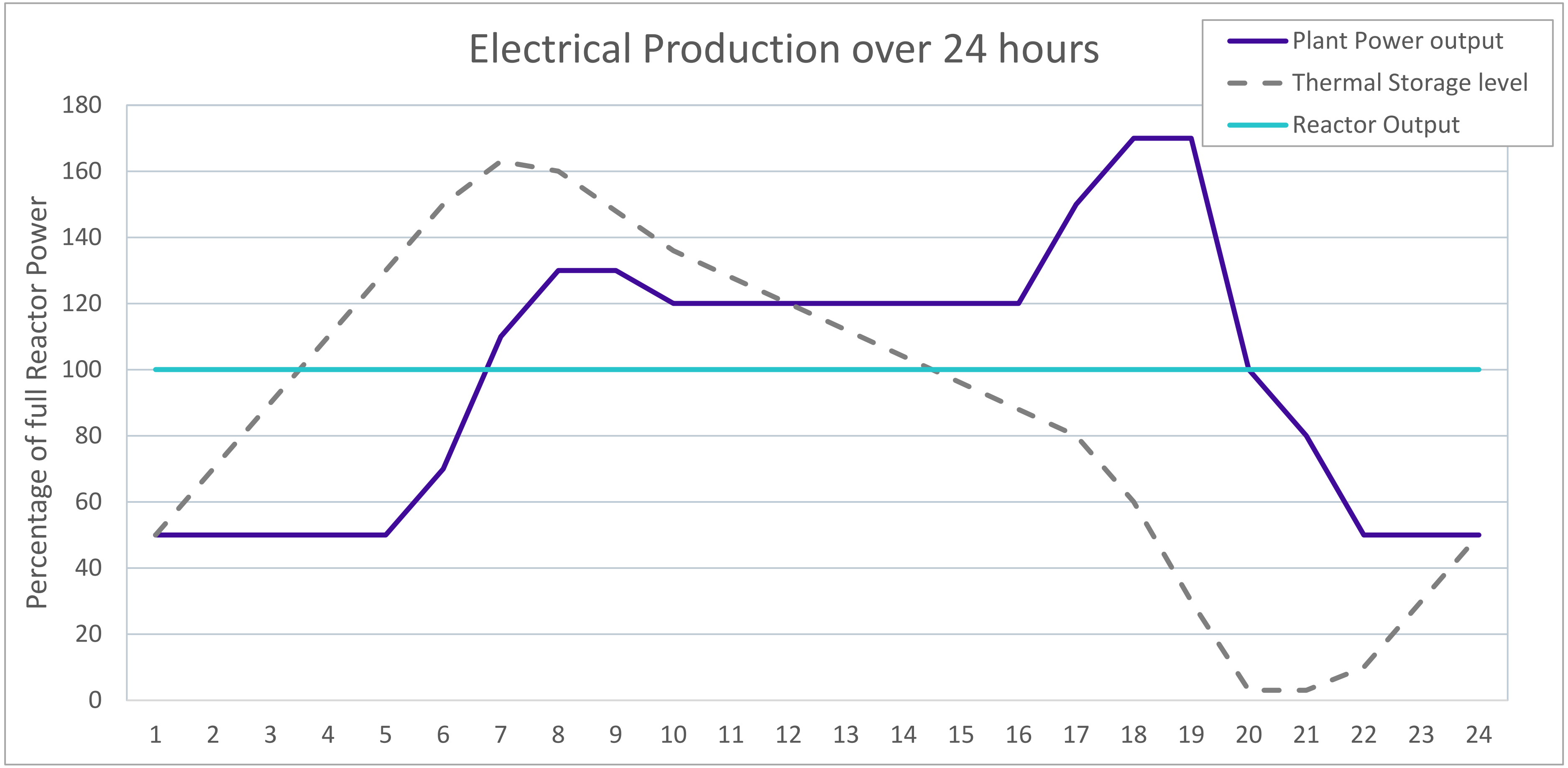
Implications of current and future grid supplies

- Baseload supplies will be limited to 20-25% of demand:
  - Low carbon options are Nuclear, Hydro, Geothermal.
  - Rest of demand will be from flexible dispatchable sources with intermittent and variable wind and solar.
- Flexible supplies lowers the capacity factor, increasing LCOE.
  - Low Capex, high fuel cost sources work well as flexible supplies.
  - No low carbon flexible supply options universally available.
- More penetration of intermittent variable sources, we will see more curtailment of supply and negative supply value during low demand times.
  - Justifies cost of storage and dispatch during peak demand times.
- Grid needs affordable, clean, dispatchable energy sources.

Possible solution: Molten Salt "Battery" configuration

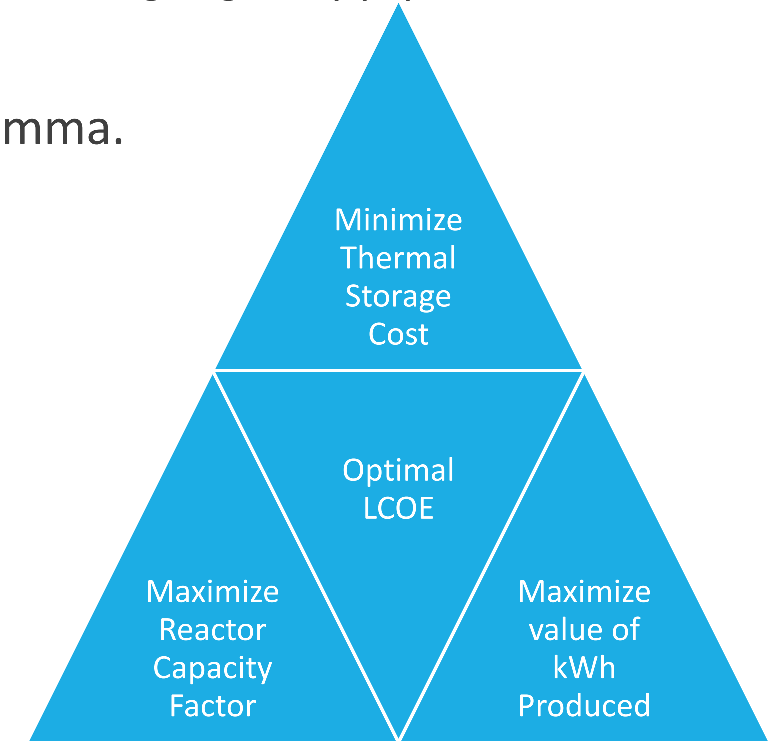


Impact of Molten Salt “Battery”

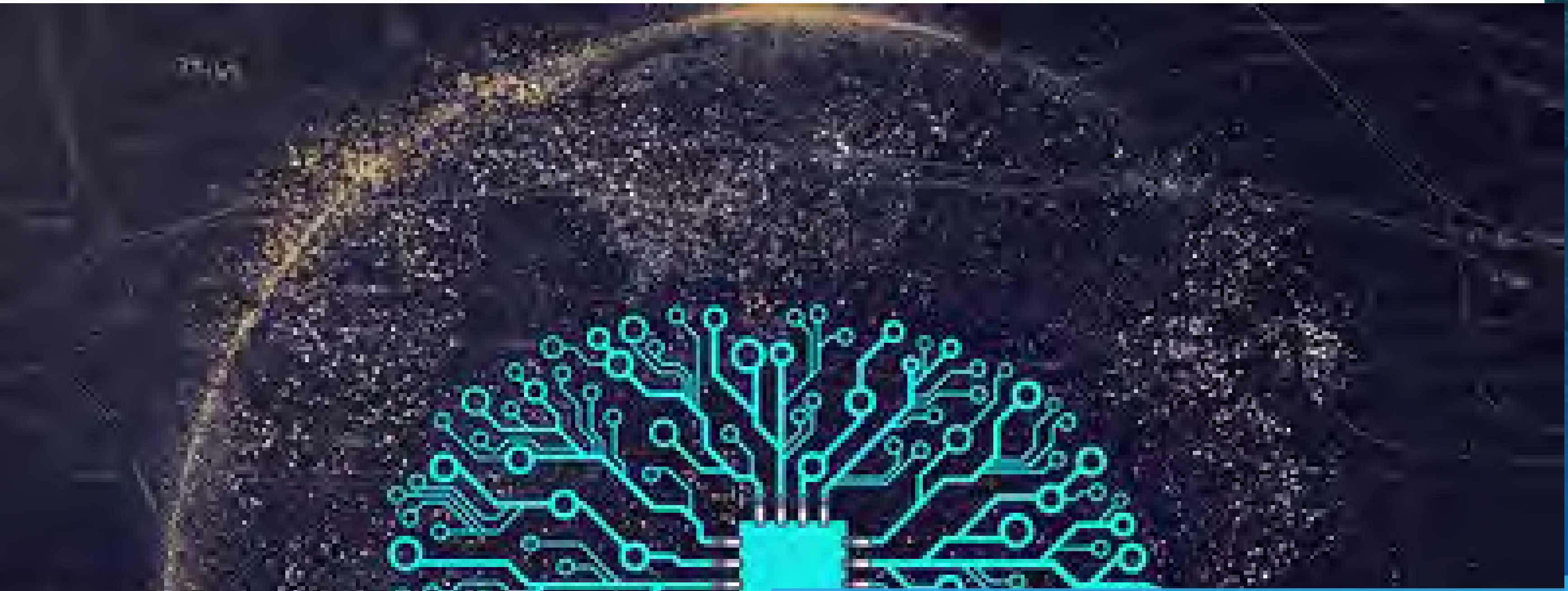


Implications of using a fission thermal battery

- Increases capital cost – molten salt storage battery comes at a cost – LCOE goes up.
- Increases capacity factor when used with intermitted sources – LCOE comes down.
- Provides energy storage for intermitted sources during high supply low demand times – increases value of supplied electricity unit.
- Problem to solve: The generation/demand trilemma.



Thank You



Dr. Char Sample
Chief Scientist – Cybercore Division
January 2021

Failures in AI and ML

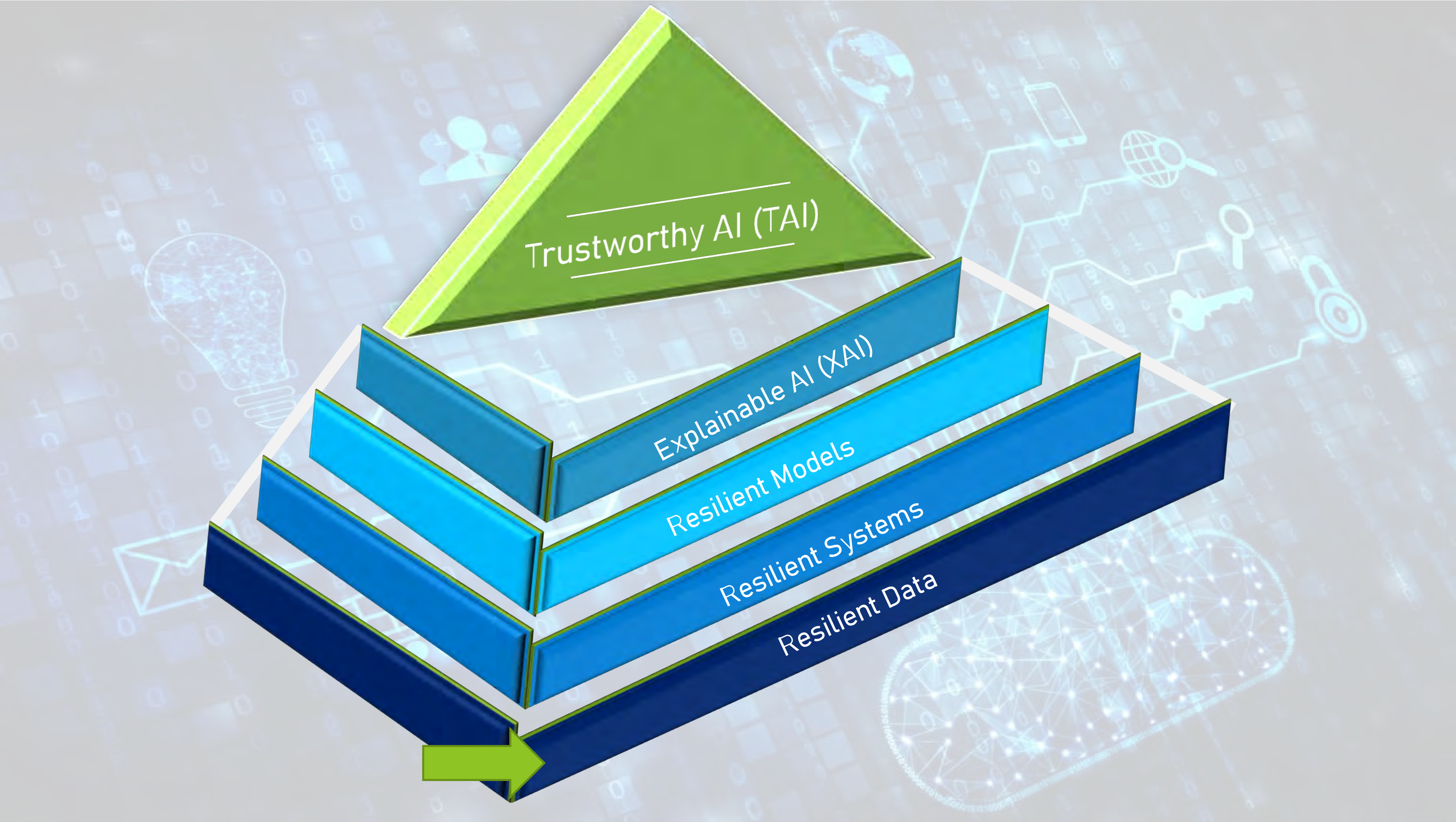
Insights and Mitigations

Introduction & Background

- **AI** – technology that performs tasks which mimic human intelligence [1].
- **Machine learning (ML)**
  - Powers AI
  - Algorithms capable of generalizing lessons learned from a limited data set to allow for abstraction of lessons to a larger environment [2].

Problem

- **AI/ML introduces problems of a breadth and nature that are difficult for humans to envision.**
  - Traditional security problems
  - AI/ML unique problems
  - Rapidity


Trustworthy AI (TAI)
Explainable AI (XAI)
Resilient Models
Resilient Systems
Resilient Data

Specific Problem

- **Problem 1: Data corruption**
  - *Description*: This group of attacks includes data poisoning, data perturbations, environmental corruptions, side effects, common corruption.
  - Effects
    - Misclassifications
    - Inaccurate results

Explanation

Data corruption

Data poisoning: Attacker contaminates training data. Introduction of a significant amount of erroneous data to trick the ML algorithm to think the data is normal.

Explanation

Data corruption

- *Data perturbation*: Attacker modifies a query to attain a desired response. Introducing an electronic disturbance during training to change the transcribing process.

Explanation

Data corruption:

- *Environmental corruption*: By making a change to background data. Shown to fool autonomous vehicles

Explanation

Data corruption:

- *Common corruption*: Changes to lighting, angles, zooming, noisy images. Example: image recognition software becomes less accurate when light changes, foggy conditions etc.

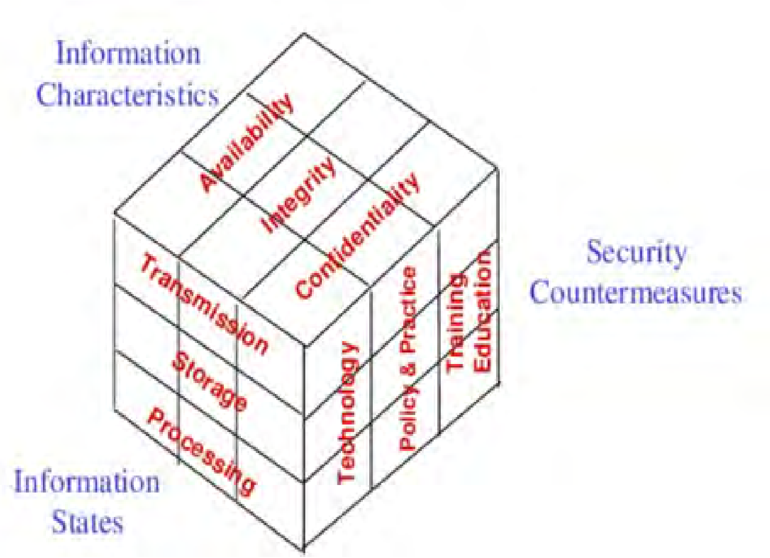
Explanation

Data corruption:

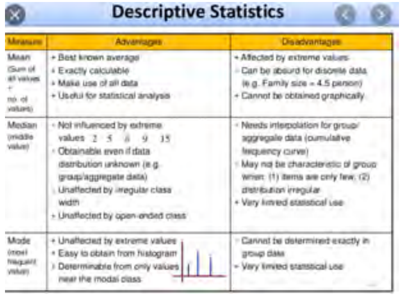
- *Side effects*: Seen when the environment may interfere with the goal of the system. System disrupts the environment, e.g. robots running over plants in the garden to scare intruders.

Background – Information Security & Information Theory

Information Security – McCumber Model



Information Theory



Descriptive Statistics

Measure	Advantages	Disadvantages
Mean (Sum of all values ÷ no. of values)	+ Best known average + Exactly calculable + Make use of all data + Useful for statistical analysis	+ Affected by extreme values - Can be absurd for discrete data (e.g. Family size = 4.5 person) + Cannot be obtained graphically
Median (middle value)	- Not influenced by extreme values 2 5 8 9 15 - Obtainable even if data distribution unknown (e.g. group/aggregate data) - Unaffected by irregular class width + Unaffected by open-ended class	- Needs interpolation for group/ aggregate data (cumulative frequency curve) - May not be characteristic of group when: (1) items are only few; (2) distribution irregular + Very limited statistical use
Mode (most frequent value)	+ Unaffected by extreme values + Easy to obtain from histogram + Determinable from only values near the modal class	- Cannot be determined exactly in group data + Very limited statistical use

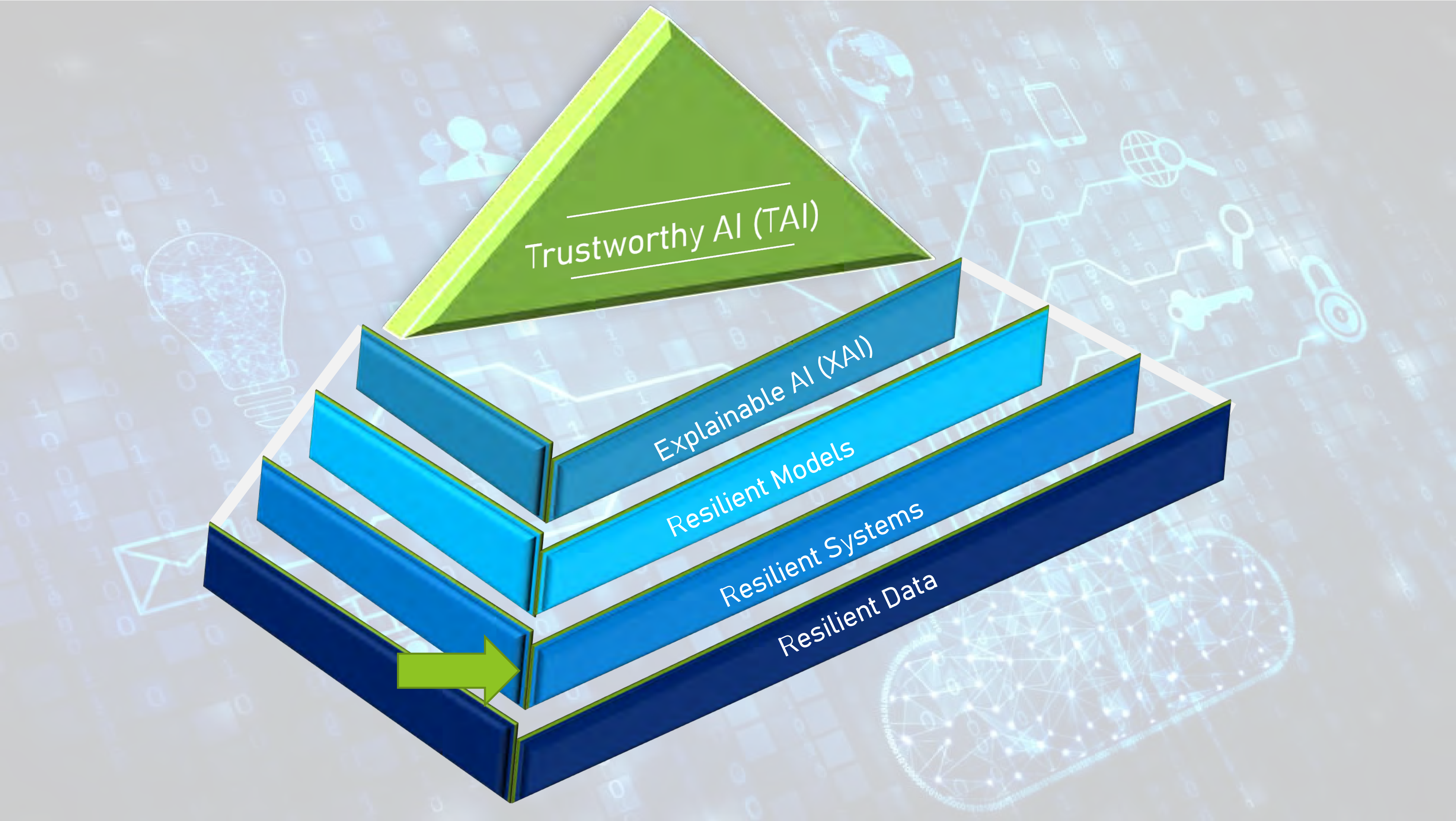
Additional context

- **6 processing meta states**
  - Start-up
  - Idle
  - Normal
  - Busy
  - Failing
  - Failed
- **Calendar profiles**
  - Holidays
  - Weekends
  - Workdays
  - Time of year

Resilient Data examples

- **Network data**
  - Historical SIEM data
  - QoS data (capacity, bandwidth usage, # of connections, fluctuations, state data etc., client and server hardware data.)
- **ICS data**
  - ”Physics” data obtained by sensors– (temperature, state, flow rate, valve position, container info etc.)

- Specific-based intrusion detection vs anomaly detection


Trustworthy AI (TAI)
Explainable AI (XAI)
Resilient Models
Resilient Systems
Resilient Data

Specific Problem

- **Problem 2: System** corruption
  - Description: Reprogramming ML, malicious ML provider recovering training data, reward hacking, backdoor ML, software dependencies exploitation, AI supply chain attacks
  - Effects:
    - Misclassification
    - Improper groupings
    - Data loss
    - DoS

Explanation

System corruption

Description:
*Reprogramming M*L –
Reprogram ML system for an unintended purpose. Specially crafted query can be re-programmed to perform a task outside of the original purpose.

Explanation

System corruption

Description: *Malicious ML provider recovering training data*, Malicious provider queries client model recovering customer training data.

Explanation

System corruption

Description: *Reward hacking.* Algorithm reward system reward gap between stated and true rewards. Typically done in reinforcement learning.

Explanation

System corruption

Description: *Backdoor ML*. ML provider has back doors into algorithms allowing for various assorted problems such as time bombs, logic bombs, etc.

Explanation

System corruption

Description: *Software dependencies exploitation*, Traditional software exploits, e.g. buffer overflows, etc.

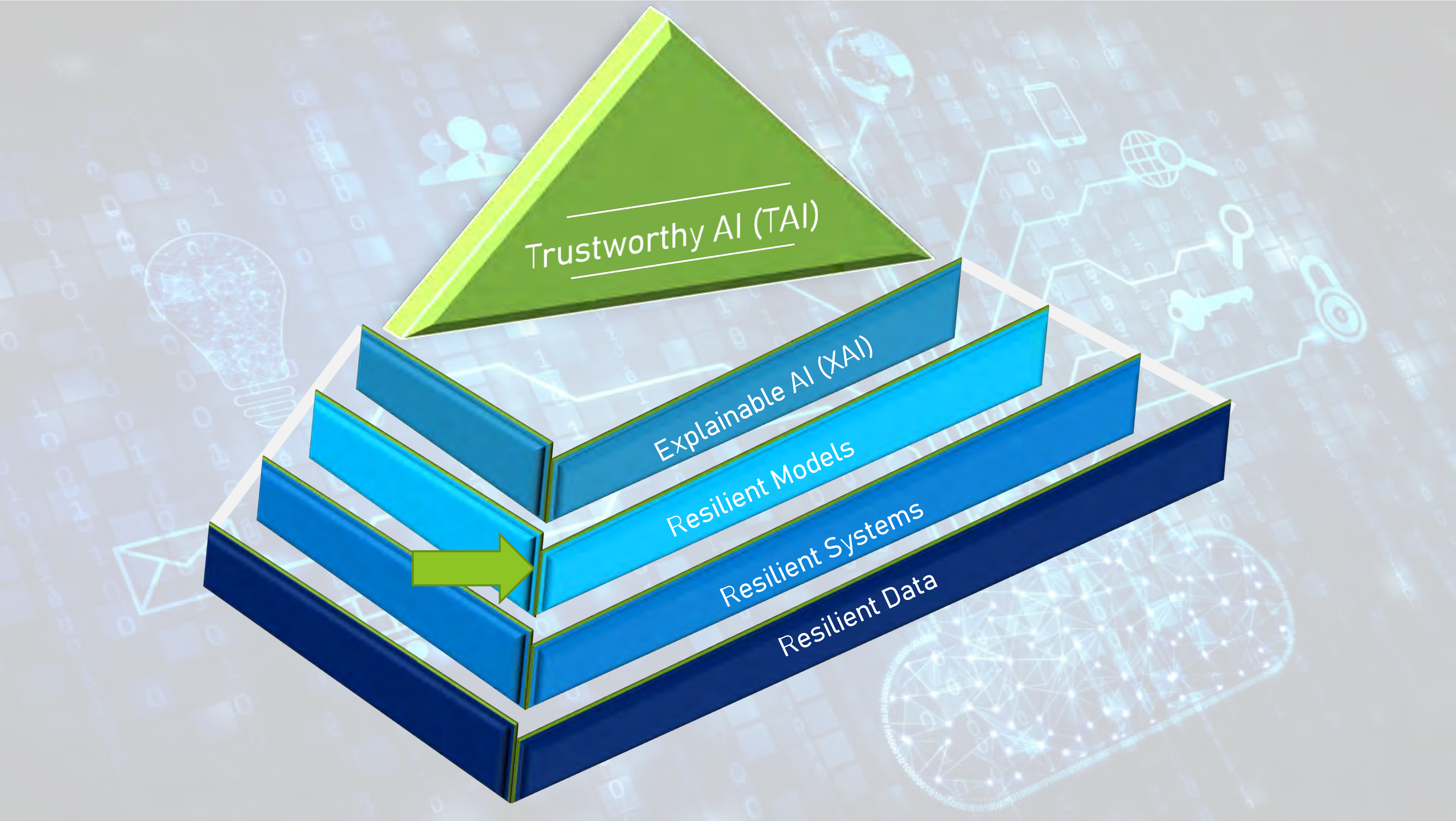
Explanation

System corruption

- Description: AI supply chain attacks. Attacker compromises ML models during downloading for use.

Research Area: Systems Resilience

- Creating Resilient Systems
  - Red Teaming AI
  - Malware discovery in binaries
  - Self-healing solutions
  - Supply chain research


Trustworthy AI (TAI)
Explainable AI (XAI)
Resilient Models
Resilient Systems
Resilient Data

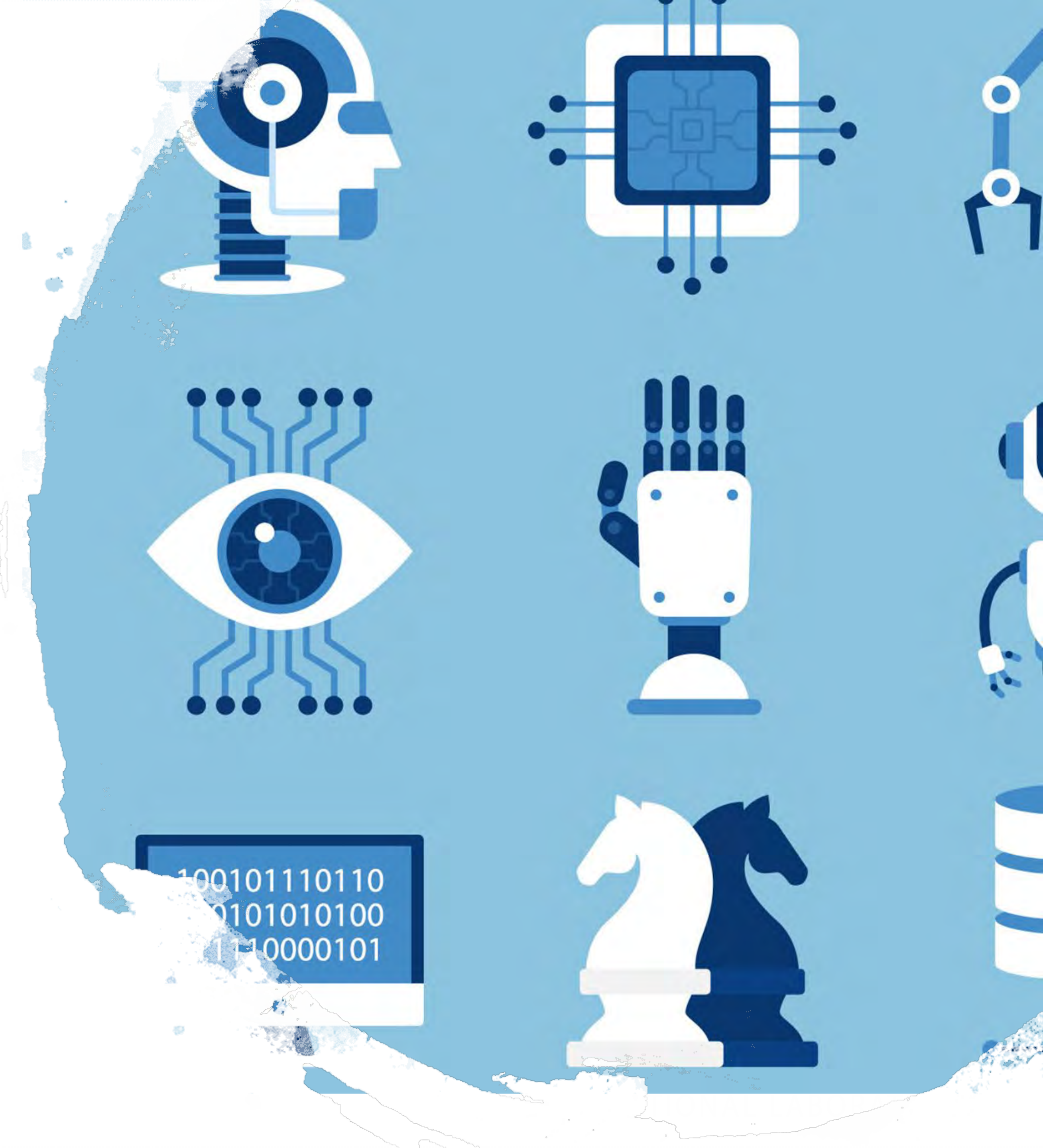
Explanation

- **Problem 3: Model corruption**

  Description: Membership inference, model stealing, model inversion, distributional shifts

  Effects:

  - Data loss
  - Algorithm manipulation
  - Algorithm anticipation
  - Data grouping manipulation

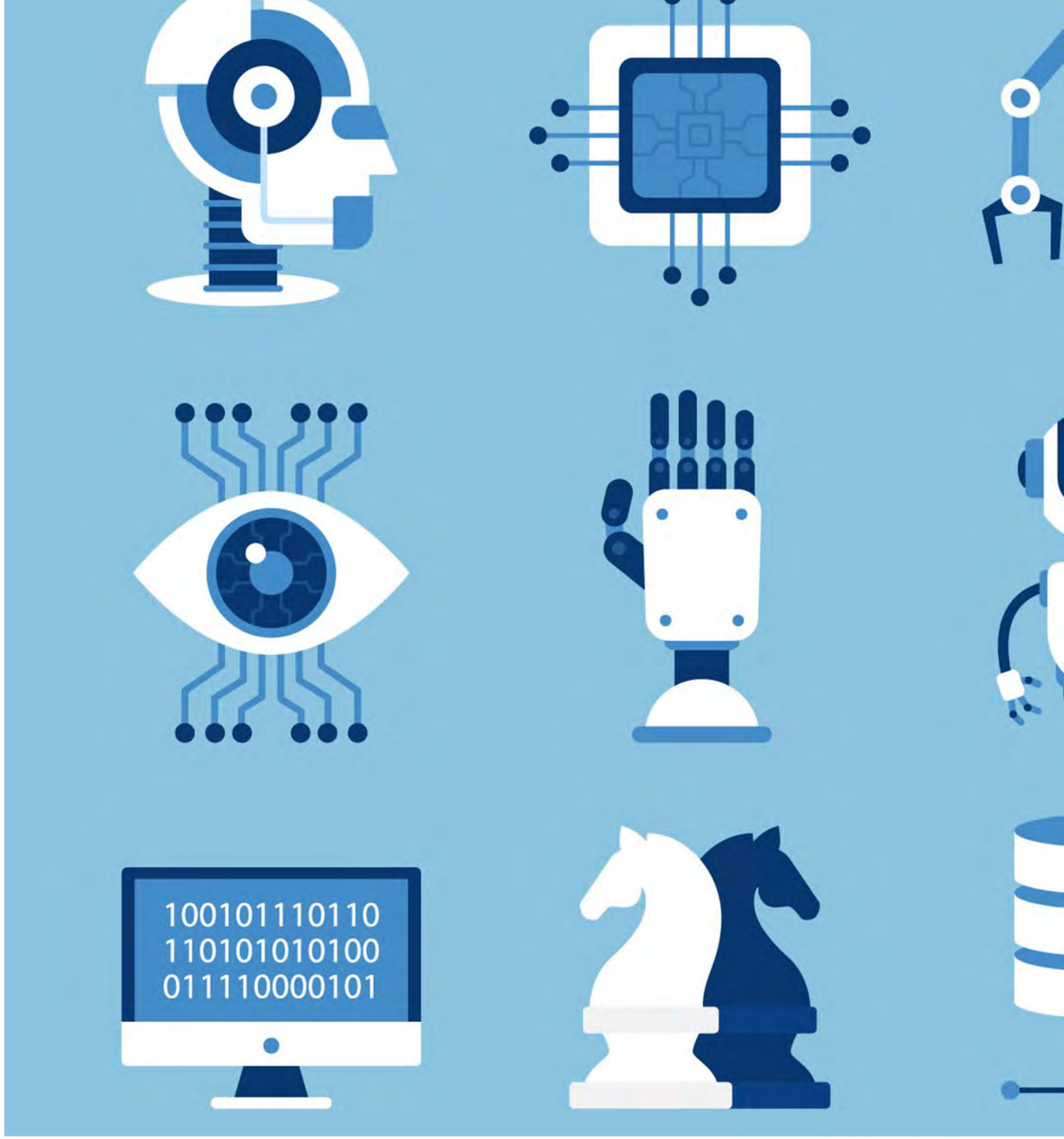


Explanation

Problem 3: Model corruption

Description: *Membership inference* – attacker determines whether a record is part of the training data used.

Attackers make accurate predictions based on specific attributes.

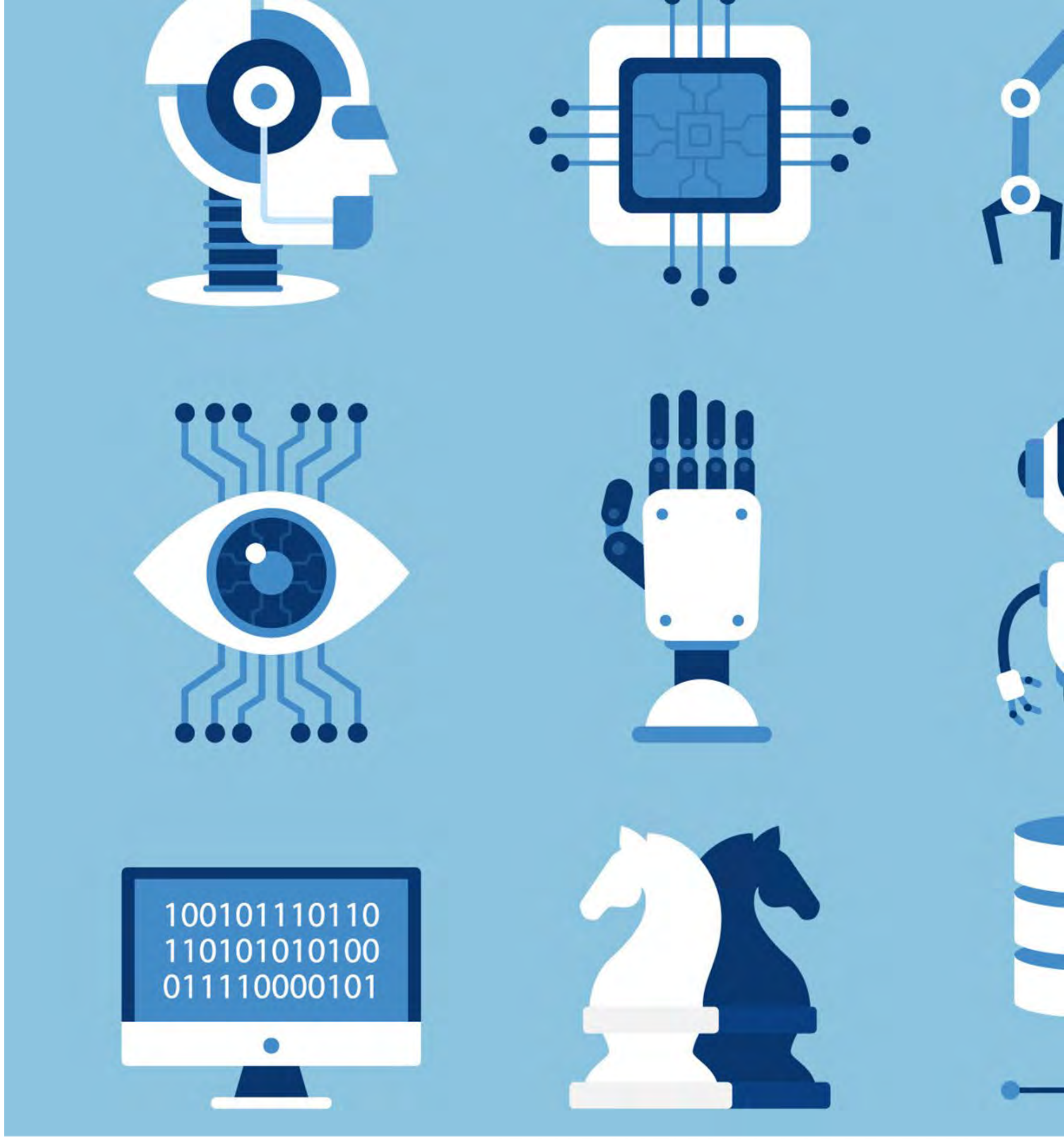


Explanation

Problem 3: Model corruption

Description: *Model stealing*.

Attacker recovers the model through carefully crafted legitimate queries, can rebuild a twin model, making possible response prediction.

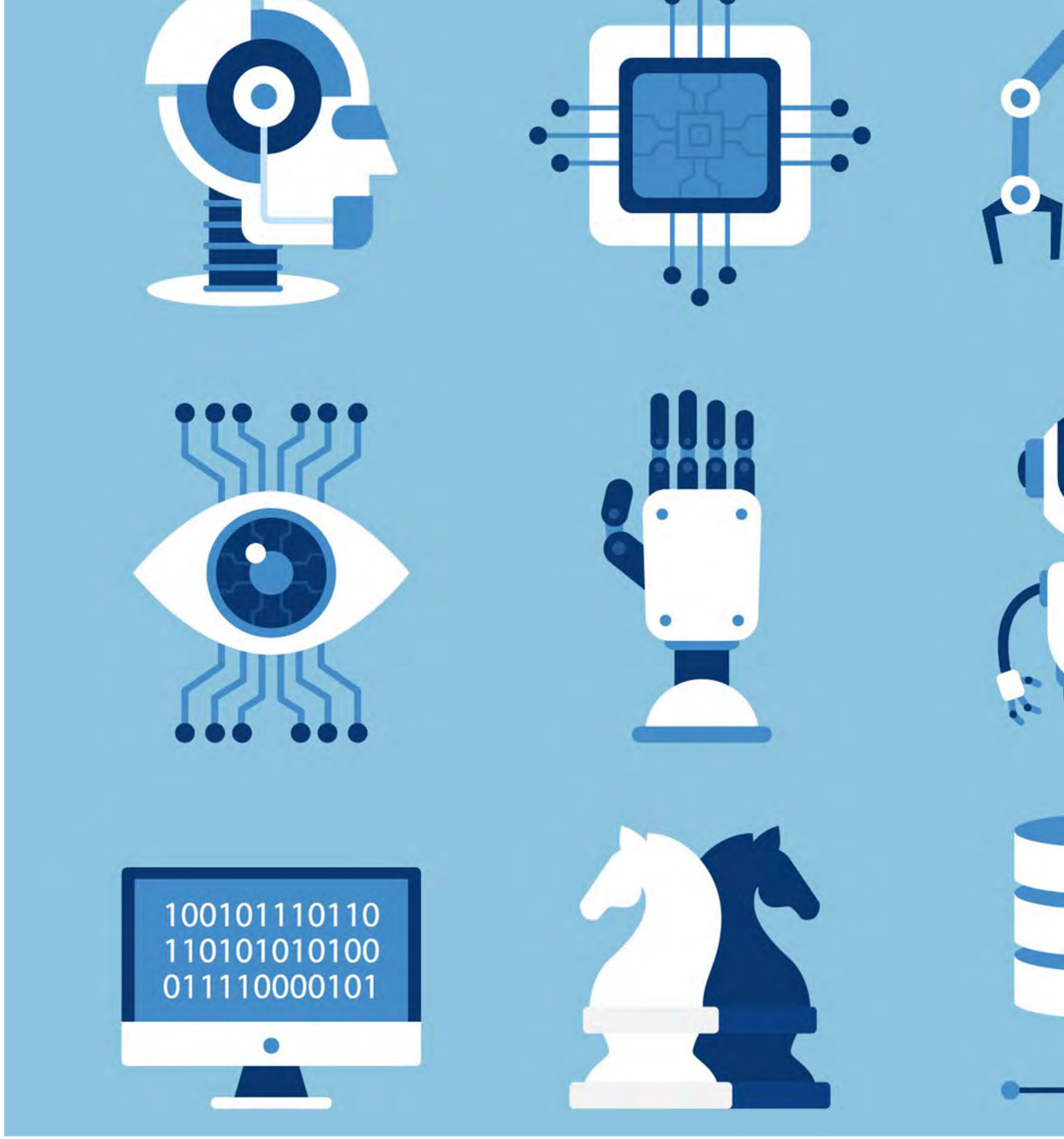


Explanation

Problem 3: Model corruption

Description: *Model inversion* - attacker discovers private features used in the model through careful queries.

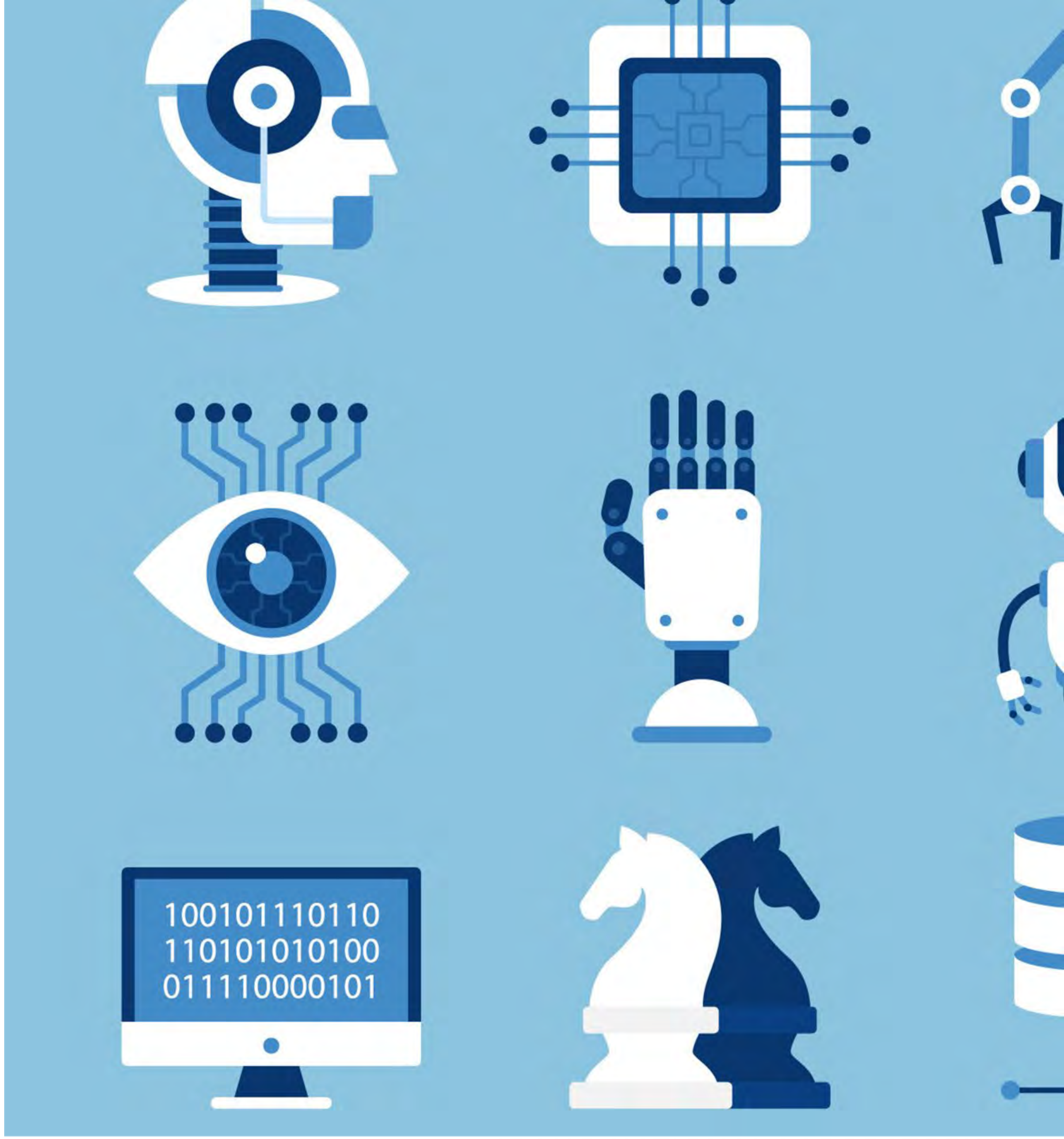
Attackers recover private training data, allowing for reconstruction of complete outputs.



Explanation

Problem 3: Model corruption

Description: *Distributional shifts.* System is tested in one environment but deployed in a different environment and the system can not adjust accordingly.



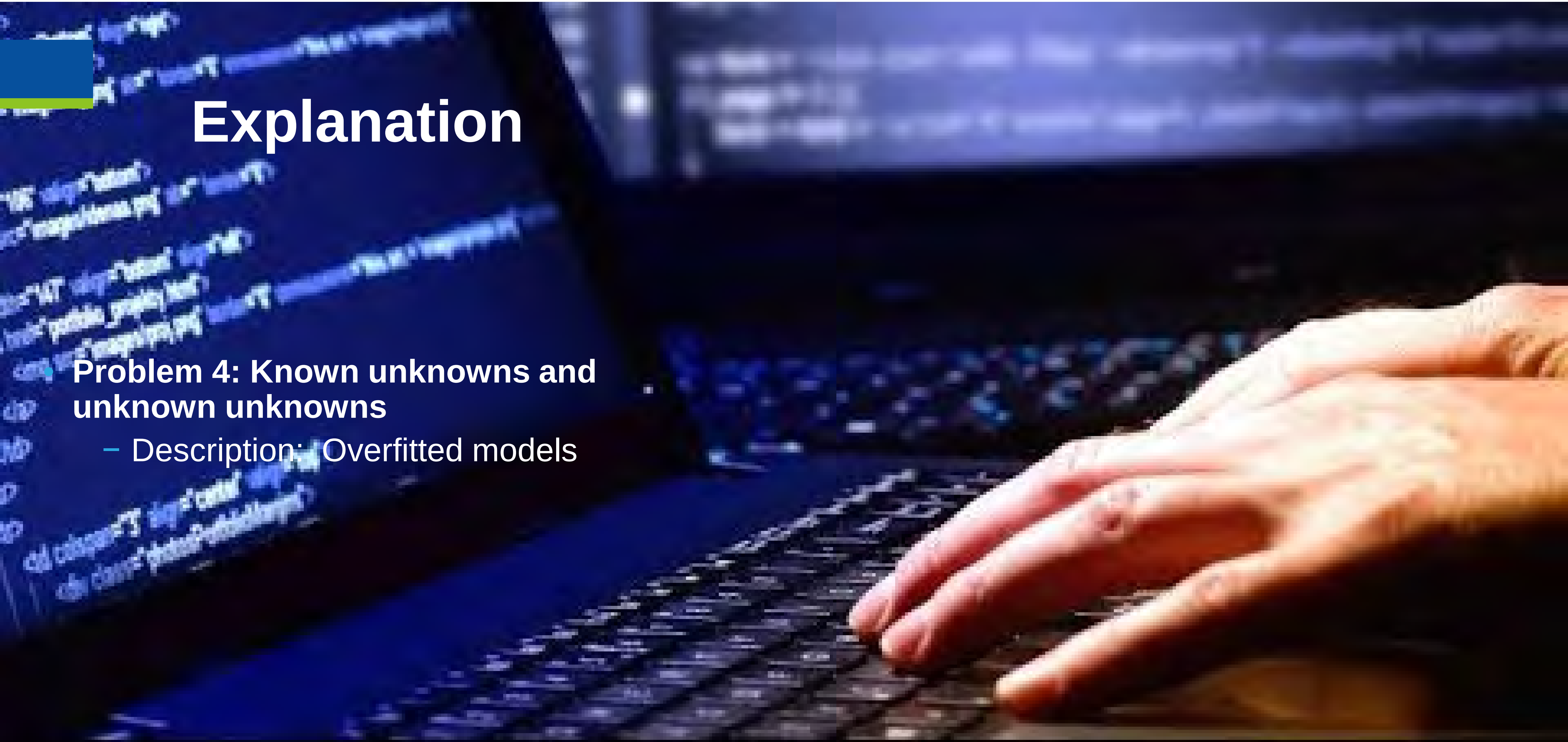
Research Area: Proposal for Model Resilience

- **Fingerprinting and countering fingerprinting efforts**
  - Time
  - Queries
  - Enforcement of Byzantine behaviors
  - Inconsistent deception

Specific Problem

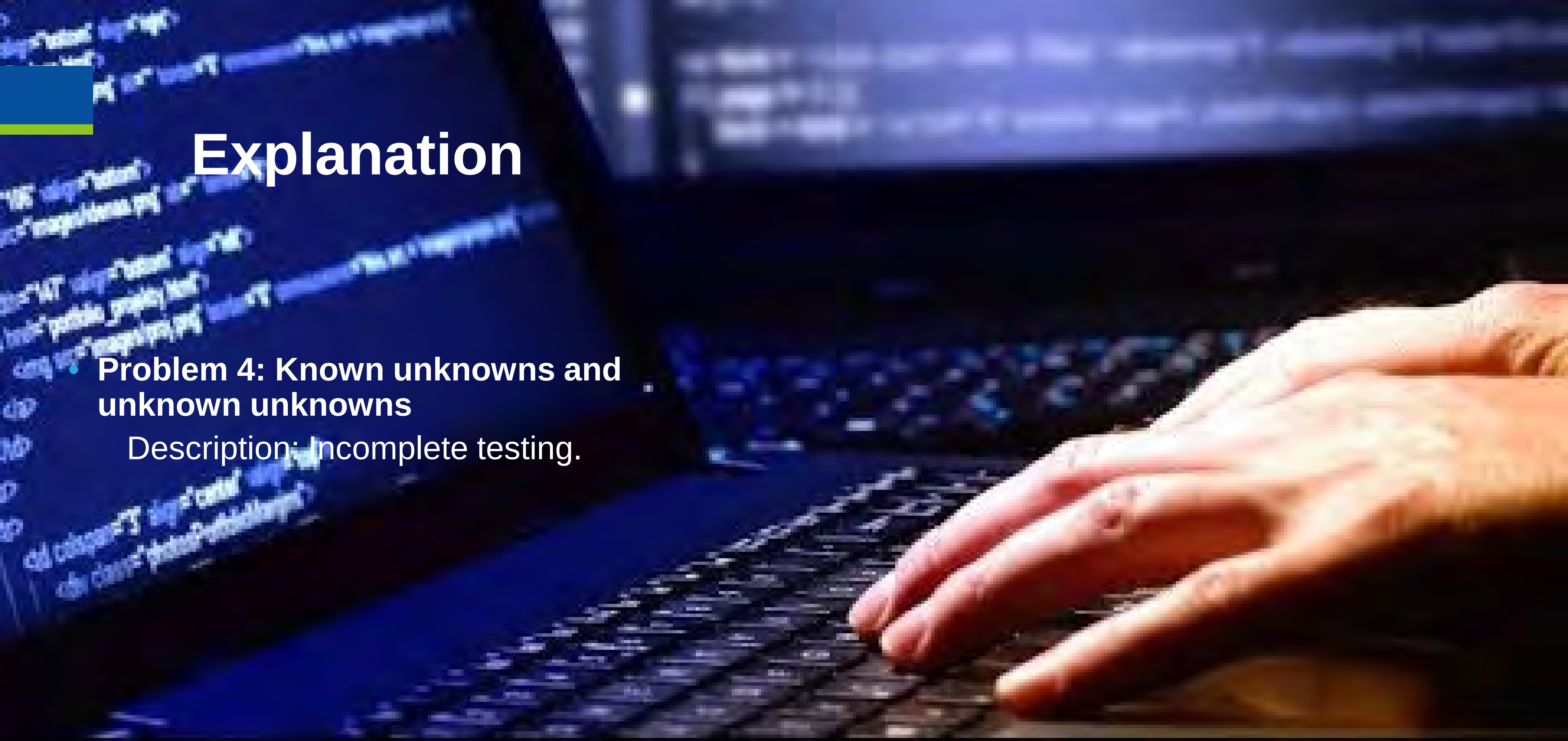
Problem 4: Known unknowns and unknown unknowns

- Description: Natural adversarial examples, overfitted models, incomplete testing, MUAI.
- Effects:
  - Algorithms prioritization schemes are inappropriate or inaccurate
  - Algorithms behave in unanticipated, unintended manner
  - Algorithm confusion



Explanation

- **Problem 4: Known unknowns and unknown unknowns**
  - Description: Overfitted models



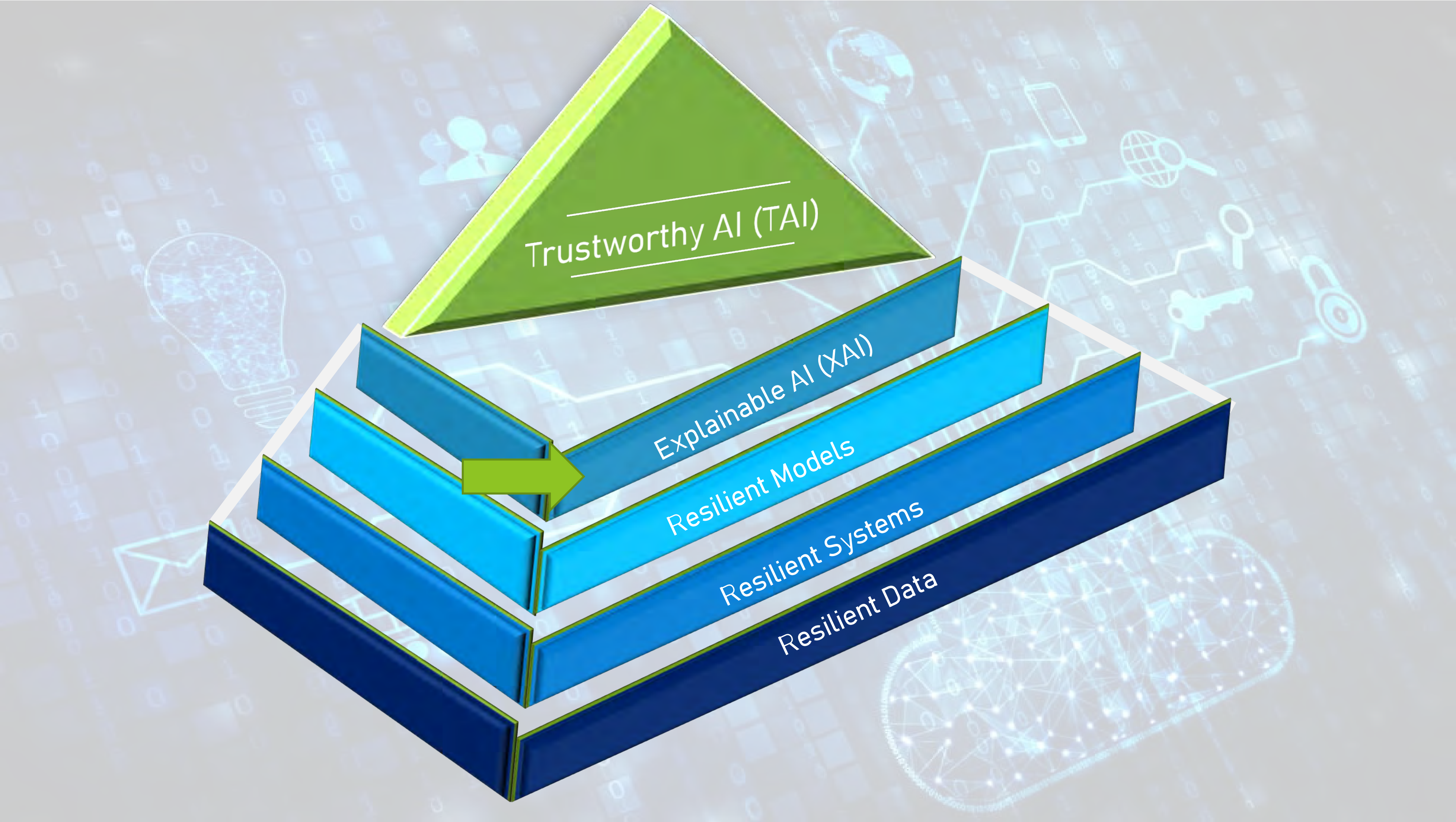
Explanation

- **Problem 4: Known unknowns and unknown unknowns**

  Description: Incomplete testing.

Explanation

- **Problem 4: Known unknowns and unknown unknowns**
  - Description: MUAI.


Trustworthy AI (TAI)
Explainable AI (XAI)
Resilient Models
Resilient Systems
Resilient Data

Research Area: Explainable AI (XAI)

- **Characteristics of XAI**
  - Evidence based output
  - User specific explanation
  - Consistently accurate
  - Knowledge limits
  - Resilient

Conclusions

- AI disruption will transform many of the workflows in our current lives.
- AI disruption will introduce unforeseeable problems.
- Humans will need to remain in the loop for the foreseeable future.
- Significant research into all aspects of intelligence.

Questions & Answers

Contact: Char Sample

e-mail: Charmaine.Sample@inl.gov

Cell: 301.346.9953

References

1. National Security Commission on Artificial Intelligence. Available: https://www.nscai.gov/about/faq

2. A. Parisi, August 2019. *Hands-on: Artificial Intelligence for Cybersecurity*, Packt Publishing Ltd., Birmingham, UK, ISBN: 978-1-78980-402-7.

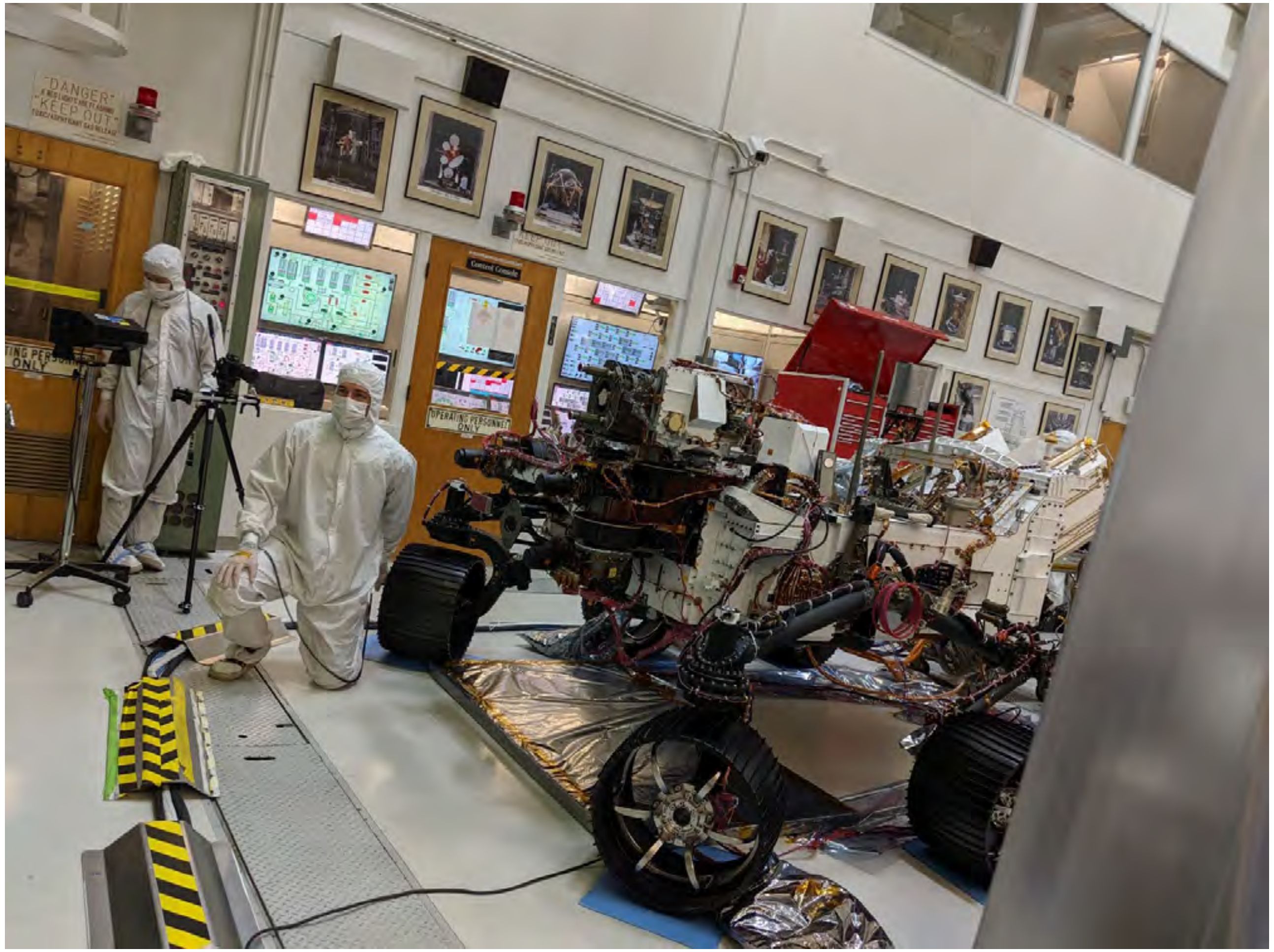
3. R. Shankar Siva Kumar, J. Snover, D. O'Brien, K. Albert, and S. Viljoen, November 2019. "Failure modes in machine learning". Available: https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

4. M. Brundage , S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." February 2018.

5. C. Sample, S.M. Loo and M. Bishop. *Resilient Data: An Interdisciplinary Approach,* Proceedings of IEEE Resilience Week, October 2020.

Other publications

- M. Bishop, M. Carvalho, R. For, and L.M. Mayron, 2011. “Resilience is More than Availability”, In Proceedings of the 2011 New Security Paradigms Workshop, pp. 95 – 104, ACM.
- C. Sample , T. Watson, S. Hutchinson, B. Hallaq, J. Cowley, and C. Maple, “Data Fidelity: Security’s Soft Underbelly”, in Proceedings of the 11th IEEE International Conference on Recent Challenges for Information Science, pp. 315 – 321, May 10–12, 2017.
- M. DeLucia, S. Hutchinson, and C. Sample. “Data Fidelity in the Post-Truth Era Part 1: Network Data”, in proceedings of the *International Conference on Cyber Warfare and Security,* pp. 149 – 159, March 2018.
- H.S. Che, A.S. Abdel-Khalik, O. Dordevic, and E. Levi. “Parameter Estimation of Asymmetrical Six-Phase Induction Machines Using Modified Standard Tests.” *IEEE Transactions on Industrial Electronics,* 64(8), pp. 6075-6085, 2017.
- K. Chan, K. Marcus, L. Scott, and R. Hardy. “Quality of Information Approach to Improving Source Selection in Tactical Networks.” In *IEEE 18th International Conference on Information Fusion,* pp. 566-573, 2015.
- DARPA GARD program website: https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception



RESILIENT FISSION BATTERY CONTROL

CHALLENGES & OPPORTUNITIES

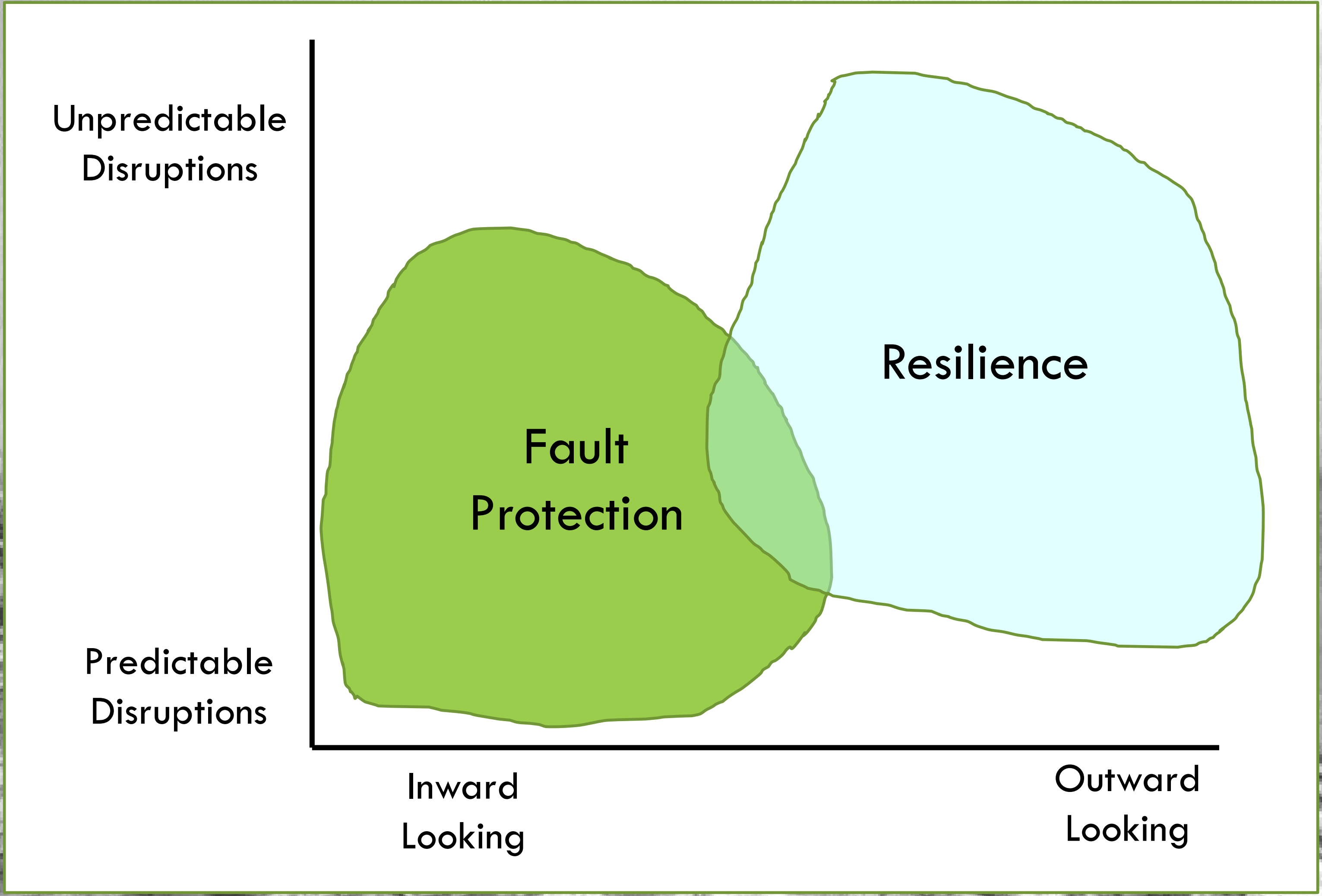
MICHAEL SIEVERS*
JET PROPULSION LABORATORY

*THIS WORK WAS DONE AS A PRIVATE VENTURE AT THE UNIVERSITY OF SOUTHERN CALIFORNIA AND NOT IN THE AUTHOR'S CAPACITY AS AN EMPLOYEE OF THE JET PROPULSION LABORATORY, CALIFORNIA INSTITUTE OF TECHNOLOGY.

SO THAT WE'RE ON THE SAME PAGE…

- Resiliency is a property associated with system behavior
  - Enables continued useful service despite disruptive events
- Three general categories of disruption:
  - External disruption – caused by factors outside the control of the system such as a natural disaster
  - Systemic disruption – a service interruption due to an internal fault
  - Human agent-triggered disruption – the result of human error or misuse of the system
- Resilient systems are trusted, adaptable, and effective in spite of unknown-unknowns
  - How do we protect against unpredictable disruptions if we don't know what to look for?

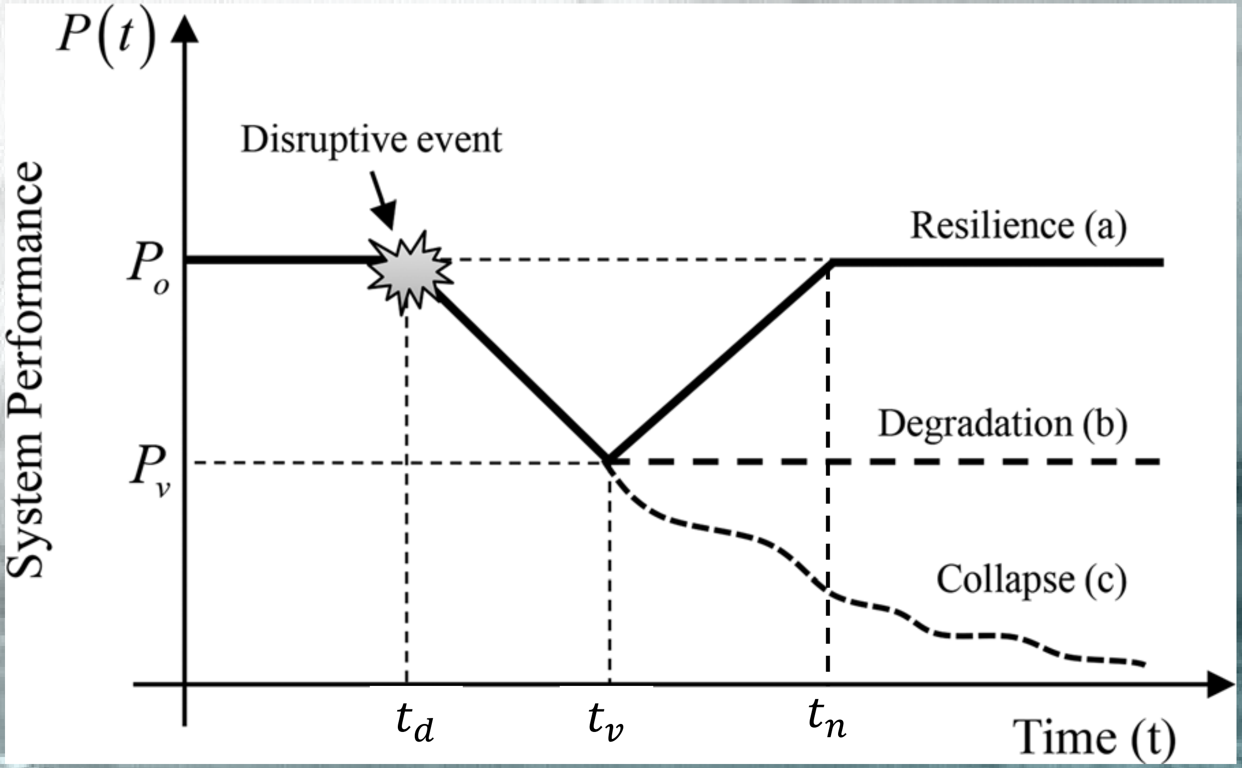
RELATIONSHIP TO FAULT-PROTECTION



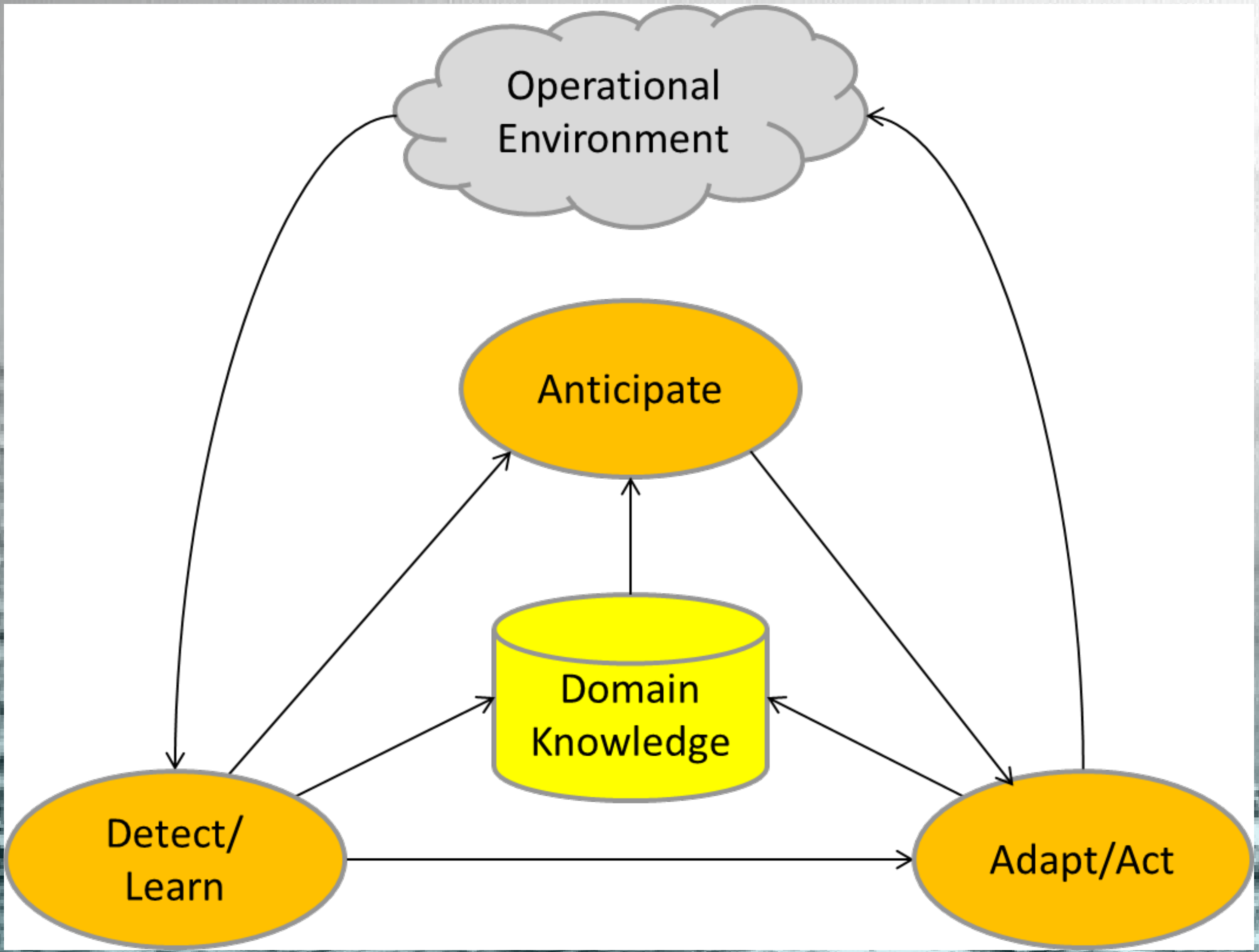
TYPICAL RESILIENCE CURVE

- System performance is normal until a disruptive event occurs
- System performance drops to a minimum until recovery occurs
- If recovery doesn't occur or isn't successful, then system drops below acceptable performance
- System may recover to full performance, may end up degraded until repair actions take places, or may collapse
- Loss of resilience ψ_{loss} is approximated as the integral of the degradation over the interval $[t_d, t_n]$

$$\psi_{loss} = \int_{t_d}^{t_n} [P_o(t_o) - P(t)]dt$$



RESILIENCE:
COMPONENTS AND RELATIONSHIPS



Resilience: Avoid, withstand, adapt to, recover from perturbations & surprises including *unknown-unknowns*

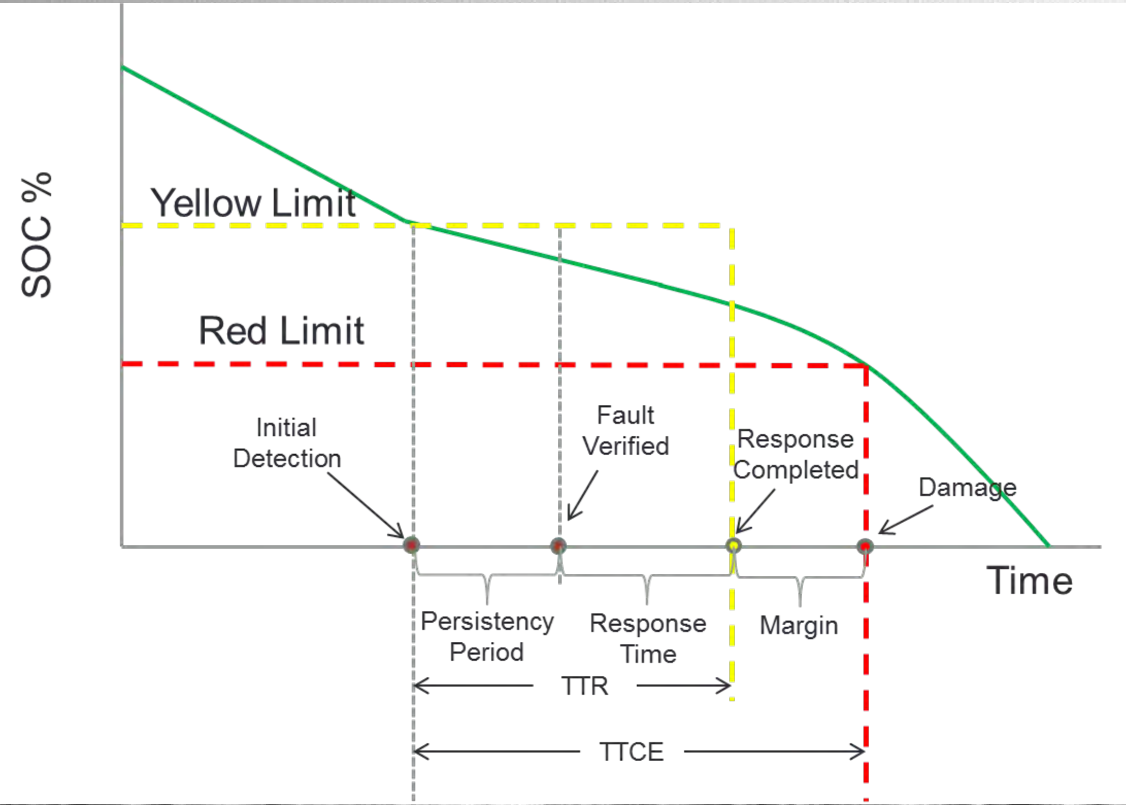
KNOWN AND UNKNOWN UNKNOWNS

- Known unknowns are potential risks that we are aware of and plan for
  - Most spacecraft are protected against known unknows
    - Safehold, redundancy and cross-strapping, fault containment, error correcting codes, …
  - Analyses determine risk likelihood and impact
    - FMECA, FTA, FFA, PRA….
  - Unfortunately, many mission-ending spacecraft failures result from overlooked or incorrectly assessed risks…
- Unknown unknowns are so completely unexpected events that would not be not considered
  - "I knew failures were possible so I included redundancy, but I just didn't think my subsystem would fail *there*!"
    - Subsystem engineer's statement at a spacecraft failure review board

TRADITIONAL FAULT-PROTECTION

- Traditional fault protection focuses on risks we know or suspect
  - Usually implemented hierarchically in which higher-level protection covers potential gaps in lower-level protection
  - Each higher level of protection takes more drastic measures to stabilize a fault condition
    - We often use “safety-net” measures at the highest level, e.g., puts a spacecraft into survival operation
  - In most cases, actions taken by fault-protection do not restore operation
    - Recovery is usually under ground control… But…
  - An issue often overlooked in traditional is *time-to-critical-effect (TTCE)* - a factor of fault coverage (the probability that a system recovers given that a fault has occurred)
    - $\rightarrow$ Fault responses must complete within TTCE or permanent damage or degradation occurs

TIME-TO-CRITICAL-EFFECT EXAMPLE



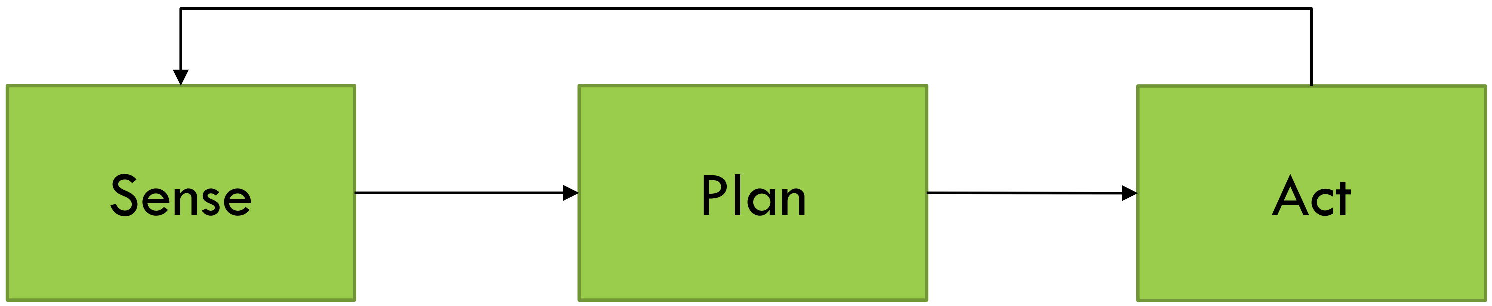
$$
\begin{aligned}
risk_{failure} &= P(fault \cap \overline{TTR < TTCE}) \\
&= P(\overline{TTR < TTCE | fault})P(fault) \\
&= (1 - c)P(fault)
\end{aligned}
$$

FISSION BATTERY CONTROLLER RESILIENCE CHALLENGES

- We can assume that some faults are managed by conventional fault-protection (stabilization & ground recovery), but others will need more urgent attention, and some might be unknowable – until they happen
- We also know that not all states or parameters are observable – so knowing system state with certainty isn't always possible
- Summarizing the challenges:
  - Unknown-unknowns
  - Potentially short TTCEs that are inconsistent with ground intervention
  - Partial observability
  - And one we didn't mention yet: taking actions may make bad situations worse

OVERCOMING RESILIENCE CHALLENGES

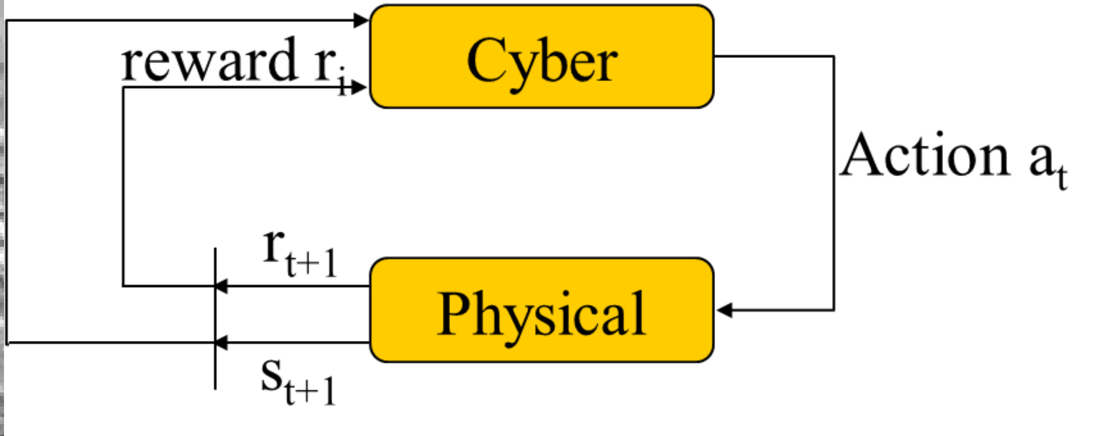
- Several methods have looked at creating resilient systems that similar to feedback control systems



- Sensing is the easy part, but how do we plan and what actions should we take?

IMPLEMENTING RESILIENCE USING REINFORCEMENT LEARNING

- RL is a machine learning construct in which software learns which actions to take actions based on maximizing a cumulative discounted reward function
  - Future actions have discounted value due to uncertainty in whether they can be used and in their effectiveness
- E.g., a Markov Decision Process (MDP) defines an environment for reinforcement learning (RL)
  - All forms of RLs can be represented as an MDP



RL Cycle

MARKOV DECISION PROCESS (MDP)

- A MDP comprises:
  - A set of possible states, S
  - A set of possible actions, A
  - A reward function $R(s, a)$
  - Transition probabilities, T, that depend on state and action
  - A belief state that is the probability distribution over the system states
- Markov property: the effects of an action taken in a state depend only on the current state and not previous states
- Two types of actions:
  - Deterministic actions: $T: S \; x \; A \; \rightarrow S$ for each state and action
  - Stochastic actions: $T: S \; x \; A \; \rightarrow Prob(S)$ for each state and action define a probability distribution over next states, i.e., $P(s'|s, a)$

POLICY

- A policy, π, is a mapping from S to A, $\pi: s \in S \rightarrow a \in A$
- I.e., when in state s, execute the action, π(s)
- An action transitions the system to state s'
- Important caveat: this assumes full observability, i.e., we know the state we've transitioned to
- The "goodness" of a policy can be established for deterministic actions by totaling the *discounted* rewards from state *s* – but that might require an infinite number of iterations
- For stochastic actions we evaluate the *expected reward* – which might also be infinite

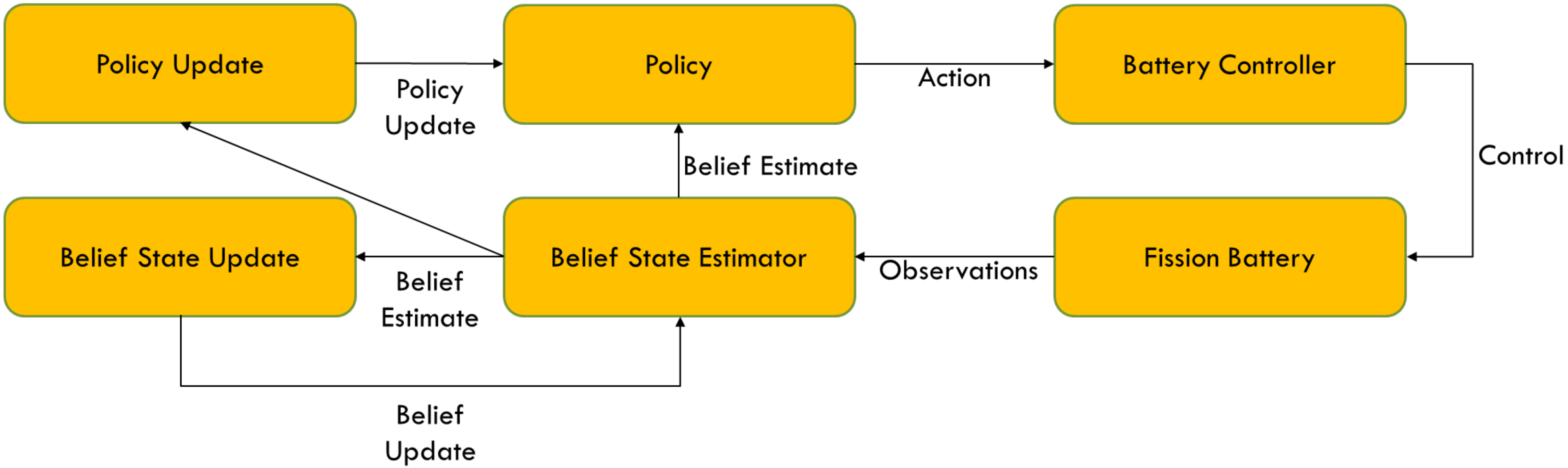
FINITE HORIZON BELLMAN EQUATIONS

- After determining an optimal policy then the Markov model is readily analyzed
- The nuance though is that we now take an action based on the state we're in
- If we find ourselves either in a known bad state or heading into a bad state, then our actions must create a *trajectory* either to a working state or to a safe state
- The optimal policy may change with time if we discover that actions do not help
  - We might have made the wrong assumptions up-front
  - The system, usage, environment, disruptions… may have changed
- That is, we must adapt to new realities by evaluating the effectiveness of actions

PARTIAL OBSERVABILITY

- A *Partially Observable Markov Decision Process* (POMDP) is a MDP comprising hidden states and observables
- State transition and emission probabilities as a function of actions taken are *learned* during system testing and updated during operation
  - E.g., using the Viterbi Algorithm
  - We want to know that the system transitions to the belief state arrived at after taking an action is “correct”
  - But since “correct” cannot be determined with certainty, what we want to know is whether the Pr of transitioning to the expected state is >> than any other state

CONCEPTUAL CONTROL ARCHITECTURE



- State space explosion is a major issue
- Approximations, paring, hierarchies, and heuristics help

REFERENCES

- A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic Concepts and Taxonomy of Dependable and Secure Computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, pp. 11-33, 2004
- K. Åström, "Optimal Control of Markov Processes with Incomplete State Information, *Journal of Mathematical Analysis and Applications* 10. p.174-205, 1965
- E. Hollnagel, D.D. Woods, and Nancy Leveson, eds., *Resilience Engineering Concepts and Precepts*, Ashgate Publishing Company, ISBN-0-7546-4641-6, 2006
- R. Patriarca, J. Bergström, G., Di Gravio, and F. Costantino, "Resilience engineering: Current status of the research and future challenges," Safety Science, Volume 102, pp. 79-100, 2018
- A., Righi, T. A. Saurin, and P. Wachs. "A Systematic Literature Review of Resilience Engineering: Research Areas and a Research Agenda Proposal." Reliab. Eng. Syst. Saf. 141 pp.142-152, 2015
- N. Roy, G. Gordon, and A. Thrun, "Finding Approximate POMDP Solutions Through Belief Compression," Journal of Artificial Intelligence Research 23 (2005) 1-40
- M. Sievers, A. Madni, and P. Pouya, "Assuring Spacecraft Swarm Byzantine Resilience," SciTech 2019
- M. Sievers, A.M. Madni, and P. Pouya, "Assuring Spacecraft Swarm Resilience," in *Proc. AIAA Scitech*, San Diego, 2018.
- M. Sievers and A.M. Madni, "Defining Credible Faults, A Risk-Based Approach," AIAA Space, Pasadena California, August 2015
- M. Sievers and A. M. Madni, "Contract-Based Byzantine Resilience in Spacecraft Swarms," AIAA Scitech 2017
- A.M. Madni and S. Jackson. "Towards a conceptual framework for resilience engineering," IEEE Systems Journal, 3.2, 181-191, 2009
- P. Smyth, "Clustering Sequences with Hidden Markov Models," *Advances in Neural Information Processing Systems*, Cambridge, MA, USA: MIT Press, 1997, pp. 648–654
- H. Kimura, K. Miyazaki, S. Kobayashi, "Reinforcement Learning in POMDPs with Function Approximation," Proc 14th ICML, pp. 152-160, 1997
- M. Igl, L. Zintgraf, T. Anh Le, F. Wood, and F. Whiteson, "Deep Variational Reinforcement Learning for POMDPs," Proceedings of the35thInternational Conference on MachineLearning, Stockholm, Sweden, PMLR 80, 2018