Promises, Promises: AI and Certification

Bill Scherlis
scherlis@cmu.edu

SCC
Annapolis
15 May 2025

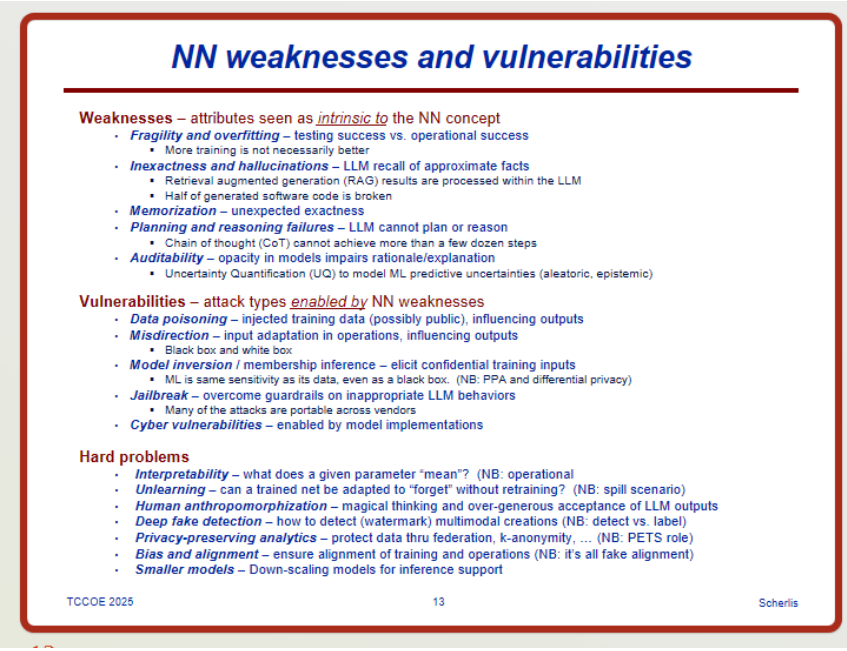
AI

Modern AI
- **Statistical predictions**
- **Not logic based**

Dimensions of untrustworthiness
- **Reactive**
- **Statistical errors**
- **Opacity**
- **Skew**
- **Cannot actually plan or reason**
- **Lacks persistent memory**
- **Learning**

Symbolic AI and hybridization
- **Auditability**
- **Representation and reasoning**
- **Heuristic power**
- **Partner to humans**

Beyond
- **Grounding**
- **Expert behavior**



NN weaknesses and vulnerabilities

Weaknesses – attributes seen as *intrinsic to* the NN concept
- ***Fragility and overfitting*** – testing success vs. operational success
  - More training is not necessarily better
- ***Inexactness and hallucinations*** – LLM recall of approximate facts
  - Retrieval augmented generation (RAG) results are processed within the LLM
  - Half of generated software code is broken
- ***Memorization*** – unexpected exactness
- ***Planning and reasoning failures*** – LLM cannot plan or reason
  - Chain of thought (CoT) cannot achieve more than a few dozen steps
- ***Auditability*** – opacity in models impairs rationale/explanation
  - Uncertainty Quantification (UQ) to model ML predictive uncertainties (aleatoric, epistemic)

Vulnerabilities – attack types *enabled by* NN weaknesses
- ***Data poisoning*** – injected training data (possibly public), influencing outputs
- ***Misdirection*** – input adaptation in operations, influencing outputs
  - Black box and white box
- ***Model inversion*** / membership inference – elicit confidential training inputs
  - ML is same sensitivity as its data, even as a black box. (NB: PPA and differential privacy)
- ***Jailbreak*** – overcome guardrails on inappropriate LLM behaviors
  - Many of the attacks are portable across vendors
- ***Cyber vulnerabilities*** – enabled by model implementations

Hard problems
- ***Interpretability*** – what does a given parameter "mean"? (NB: operational
- ***Unlearning*** – can a trained net be adapted to "forget" without retraining? (NB: spill scenario)
- ***Human anthropomorphization*** – magical thinking and over-generous acceptance of LLM outputs
- ***Deep fake detection*** – how to detect (watermark) multimodal creations (NB: detect vs. label)
- ***Privacy-preserving analytics*** – protect data thru federation, k-anonymity, … (NB: PETS role)
- ***Bias and alignment*** – ensure alignment of training and operations (NB: it's all fake alignment)
- ***Smaller models*** – Down-scaling models for inference support

TCCOE 2025 13 Scherlis

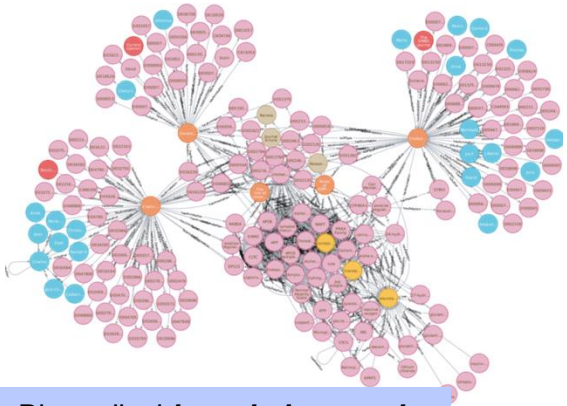
AI

Modern AI
- **Statistical predictions**
- **Not logic based**

Dimensions of untrustworthiness
- **Reactive**
- **Statistical errors**
- **Opacity**
- **Skew**
- **Cannot actually plan or reason**
- **Lacks persistent memory**
- **Learning**

Symbolic AI and hybridization
- **Auditability**
- **Representation and reasoning**
- **Heuristic power**
- **Partner to humans**

Beyond
- **Grounding**
- **Expert behavior**



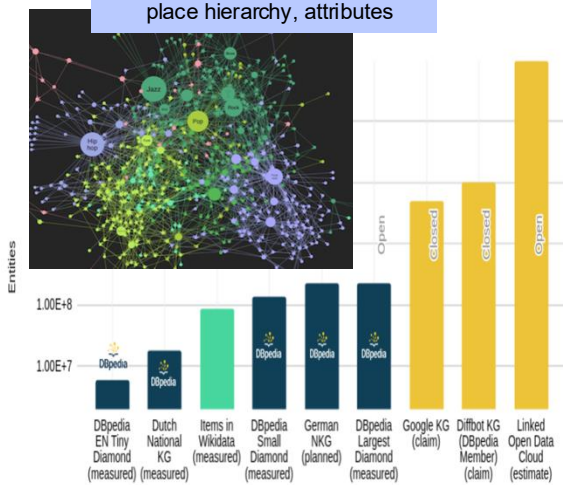
Biomedical ***knowledge graphs***

KG: {< uid, node, edge, node >}

[Term]
id: **DOID:0060013**
name: **X-linked severe combined immunodeficiency**
alt_id: DOID:5811
def: "A severe combined immunodeficiency that is a X-linked SCID that has_material_basis_in mutations in genes encoding common gamma chain proteins shared by the interleukin (IL-2, 4, 7, 9, 16 and 21) receptors resulting in a non-functional gamma chain, defective interleukin signalling, minimal or absent T- and NK cells and non-functional B-cells."
[**url:http://en.wikipedia.org/wiki/Severe_combined_immunodeficiency**, url:https://ghr.nlm.nih.gov/condition/x-linked-severe-combined-immunodeficiency#synonyms]
comment: OMIM mapping confirmed by DO. [LS].
subset: DO_rare_s
subset: NCIthesaur
synonym: "gamma
synonym: "SCID-X
synonym: "thymic e
synonym: "XSCID"
xref: GARD:5618
xref: MESH:D053632
xref: MIM:300400
xref: NCI:C4682
xref: SNOMEDCT_US_2023_03_01:203592006
xref: UMLS_CUI:C1279481
is_a: DOID:627 ! severe combined immunodeficiency

Node from a 170kloc human disease ***ontology***

Geoname ***vocabulary***
(25m placenames)
place hierarchy, attributes



Google ***knowledge graph***
approaching 10B nodes (2020)

```
way 2:        transport a box with a truck
precondition: truck-at(t,l_f),adjacent(l_f,l_n),box-at(b,l_f),
              in-city(l_f,c_f),in-city(l_n,c_n),same-city(c_f,c_n)
task network: (load-truck(b,t,l_f),drive(t,l_f,l_n),
              unload-truck(b,t,l_n),deliver(b,l_n,l_t))
```

Robot ***planning rules***

AI

Modern AI

- **Statistical predictions**
- **Not logic based**

Dimensions of untrustworthiness

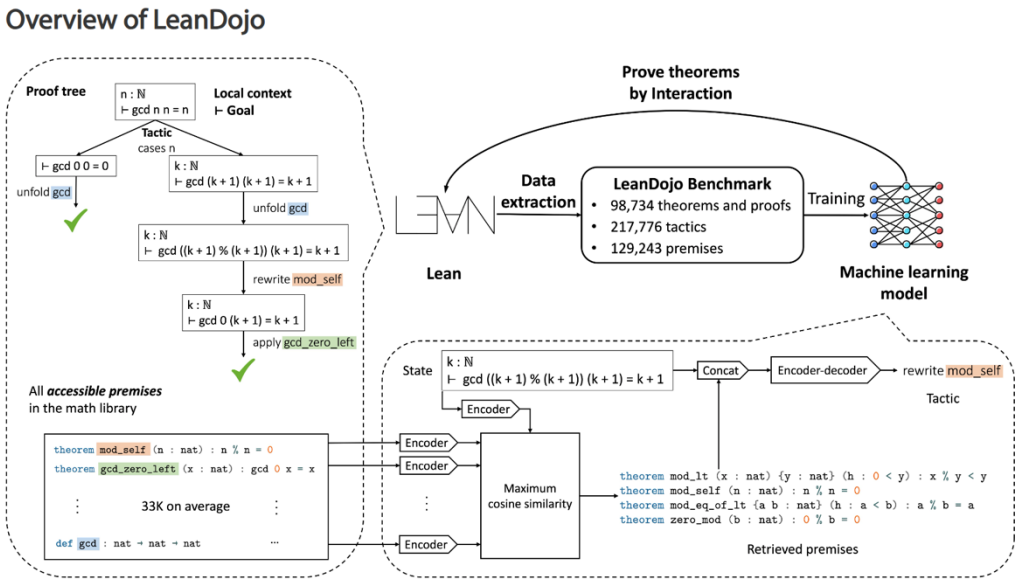
- **Reactive**
- **Statistical errors**
- **Opacity**
- **Skew**
- **Cannot actually plan or reason**
- **Lacks persistent memory**
- **Learning**

Symbolic AI and hybridization

- **Auditability**
- **Representation and reasoning**
- **Heuristic power**
- **Partner to humans**

Beyond

- **Grounding**
- **Expert behavior**



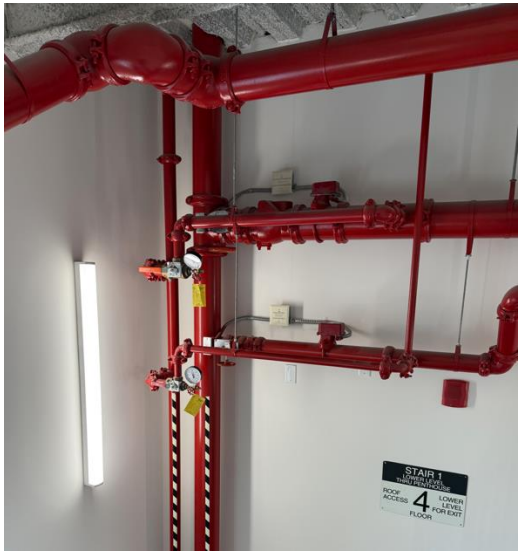
AI

Modern AI
- **Statistical predictions**
- **Not logic based**

Dimensions of untrustworthiness
- **Reactive**
- **Statistical errors**
- **Opacity**
- **Skew**
- **Cannot actually plan or reason**
- **Lacks persistent memory**
- **Learning**

Symbolic AI and hybridization
- **Auditability**
- **Representation and reasoning**
- **Heuristic power**
- **Partner to humans**

Beyond
- **Grounding**
- **Expert behavior**



AI

Modern AI
- **Statistical predictions**
- **Not logic based**

Dimensions of untrustworthiness
- **Reactive**
- **Statistical errors**
- **Opacity**
- **Skew**
- **Cannot actually plan or reason**
- **Lacks persistent memory**
- **Learning**

Symbolic AI and hybridization
- **Auditability**
- **Representation and reasoning**
- **Heuristic power**
- **Partner to humans**

Beyond
- **Grounding**
- **Expert behavior**

Lessons

- **Dimensions of scale**
  - Scope.
  - Complexity and size.
  - Practice.
  - Evolution.
  - Adoptability.
- **Use case categories**
  - Major systems that matter.
  - Big results.
  - Ecosystems.
  - Broad use.
- **Industry uses**
- **Architecture matters**
- **Productivity can be a selling point**
- **Standards can drive**

Accepting proofs

- **Social process – but on what?**
- **Physical stuff – the Fallacy of Identification**

Accepting contingencies

- **What's not modeled**
- **Incorrect behaviors**
- **Validation**
- **Risks**

Table 3-1. Participant Use Cases

Point(s) of Contact	Use Cases	Dimensions of Scale
• Clark Barrett	Solvers for Boolean satisfiability and Satisfiability Modulo Theories (SMT) as industrial workhorses	• Complexity and the size of systems • Range of properties and qualities • Effectiveness and efficiency of formal methods-related modeling and tooling • Ease of use

[3] See https://cps-vo.org/group/FMatScale/report. Note: This site requires a user name and password

4

Formal Methods at Scale: 2019 Workshops Report

Point(s) of Contact	Use Cases	Dimensions of Scale
• Darren Cofer • Matt Wilding	Large aerospace and defense systems	• Effectiveness and efficiency of formal methods-related modeling and tooling
• Michael Collins • Kristin Giammarco	Modeling behavior of complex systems concepts supporting reasoning methods critical to national cyber and cryptologic missions	• Range of properties and qualities • Ease of use
• Takeaways from presentation by Byron Cook	Cloud Services foundational assurance (e.g. cryptography, virtualization, storage)	• Complexity and size of systems • Range of properties and qualities • Ability to rapidly co-evolve systems and associated evidence within continuous integration/deployment (CI/CD) • Ease of use
• Mike Dodds • John Launchbury • Stephen Magill	Cryptography & Formal Methods as a Service	• Range of properties and qualities • Effectiveness and efficiency of formal methods-related modeling and tooling • Ability to rapidly co-evolve systems and associated evidence within CI/CD • Ease of use
• Kathleen Fisher	Cyber-retrofit of DoD platforms	• Range of properties and qualities • Ability to co-evolve systems
• Warren A. Hunt, Jr. • J. Strother Moore	Creation, analysis, and maintenance of models of computational systems, assist with creating, analyzing, and maintaining models of computational systems developed by in use by companies including AMD, ARM, Centaur Technology, IBM, and Intel	• Effectiveness and efficiency of formal methods-related modeling and tooling • Range of properties and qualities • Complexity and the size of systems • Ability to rapidly co-evolve systems and associated evidence within CI/CD • Ease of use
• Peter O'Hearn	Scaling the impact of analysis of apps for Android and iOS, Facebook Messenger, Instagram, and other apps	• Complexity and size of systems • Range of properties and qualities • Effectiveness and efficiency of formal methods-related modeling and tooling • Ability to rapidly co-evolve systems and associated evidence within CI/CD • Ease of use
• Ray Richards	DoD military systems	• Complexity and size of systems • Effectiveness and efficiency of formal methods-related modeling and tooling • Ability to rapidly co-evolve systems and associated evidence within CI/CD • Ease of use

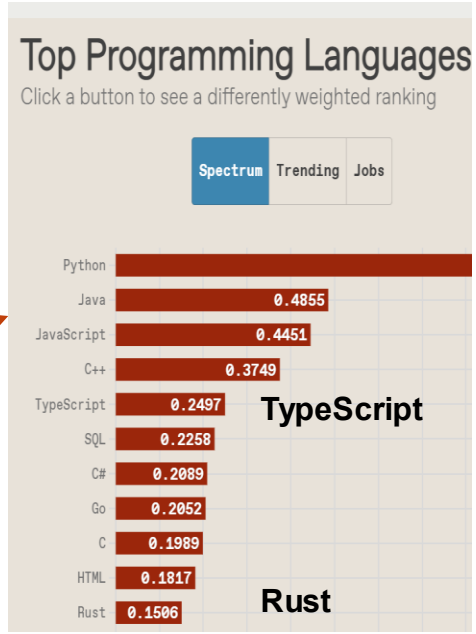
5

Formal Methods at Scale: 2019 Workshops Report

Point(s) of Contact	Use Cases	Dimensions of Scale
• Natarajan Shankar	Designing, analyzing, and creating computer systems	• Complexity and size of systems • Range of properties and qualities • Effectiveness and efficiency of formal methods-related modeling and tooling • Ability to rapidly co-evolve systems and associated evidence within CI/CD • Ease of use
• Nikhil Swamy	Components in HTTPS ecosystem, including transport layer security, the main protocol at the heart of HTTPS, as well as the main underlying cryptographic algorithms, such as AES, SHA2 or X25519	• Complexity and size of systems • Range of properties and qualities • Effectiveness and efficiency of formal methods-related modeling and tooling • Ability to rapidly co-evolve systems and associated evidence • Ease of use

Lessons

- **Dimensions of scale**
  - Scope.
  - Complexity and size.
  - Practice.
  - Evolution.
  - Adoptability.
- **Use case categories**
  - Major systems that matter.
  - Big results.
  - Ecosystems.
  - Broad use.
- **Industry uses**
- **Architecture matters**
- **Productivity can be a selling point**
- **Standards can drive**



Accepting proofs

- **Social process – but on what?**
- **Physical stuff – the Fallacy of Identification**

Accepting contingencies

- **What's not modeled**
- **Incorrect behaviors**
- **Validation**
- **Risks**

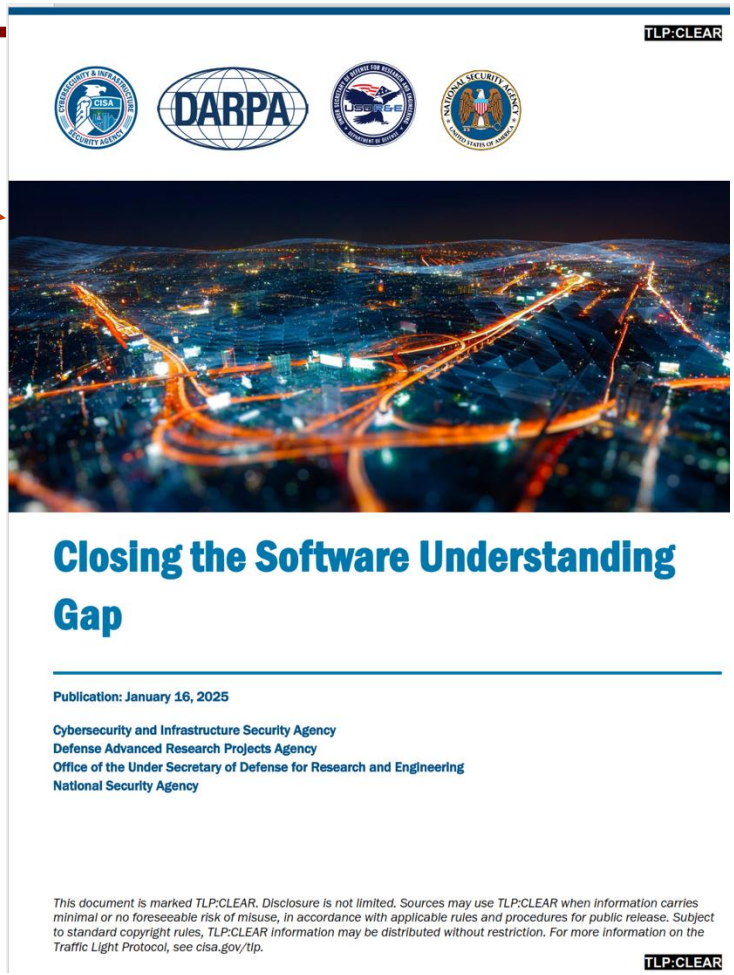
Information Loss

Software Understanding
- **NB: Report and workshop**
- **Broadly inclusive problem**

Nature of evidence
- **Code and beyond**
- **Rules of the road**
- **Higher level models and patterns**
- **Development artifacts, informal and formal**
- **Argumentation**

Tightening feedback loops
- **Defer stimulus**
- **Advance response**

Roles for FM and heuristic AI


TLP:CLEAR

DARPA

Closing the Software Understanding Gap

Publication: January 16, 2025

Cybersecurity and Infrastructure Security Agency
Defense Advanced Research Projects Agency
Office of the Under Secretary of Defense for Research and Engineering
National Security Agency

This document is marked TLP:CLEAR. Disclosure is not limited. Sources may use TLP:CLEAR when information carries minimal or no foreseeable risk of misuse, in accordance with applicable rules and procedures for public release. Subject to standard copyright rules, TLP:CLEAR information may be distributed without restriction. For more information on the Traffic Light Protocol, see cisa.gov/tlp.

TLP:CLEAR

Information Loss

Software Understanding
- **NB: Report and workshop**
- **Broadly inclusive problem**

Nature of evidence
- **Code and beyond**
- **Rules of the road**
- **Higher level models and patterns**
- **Development artifacts, informal and formal**
- **Argumentation**

Tightening feedback loops
- **Defer stimulus**
- **Advance response**

Roles for FM and heuristic AI

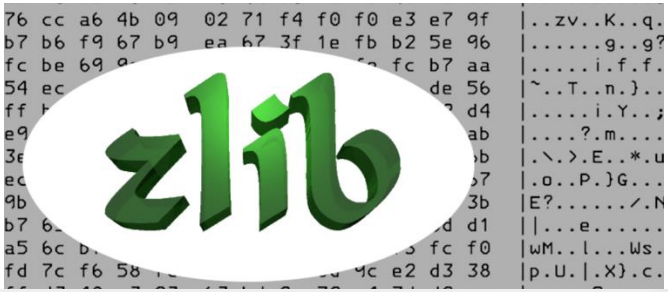
Risk Modeling and Mitigation

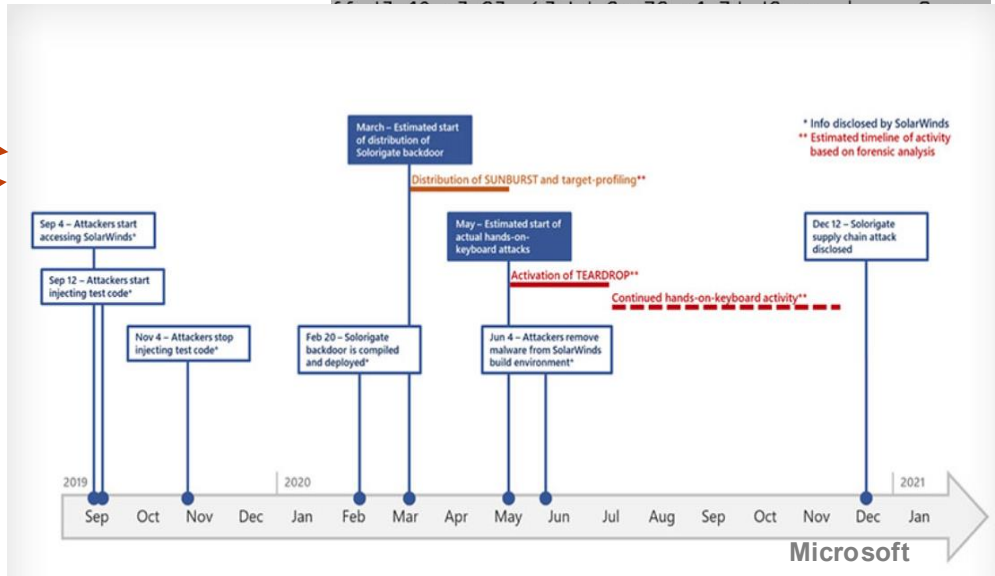
Integration context for overall risk management
- **Evidence role**

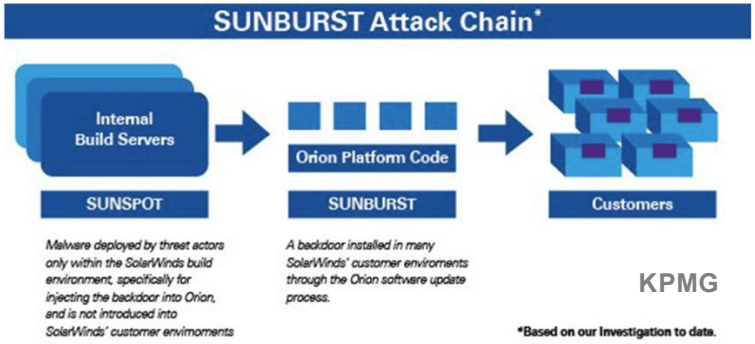
The inadequacy of traditional models
- **Actuarial risk**

Cyber difficulties
- **Adversaries**
- **Opacity**
- **Hidden correlations**
- **Context of system operations**
- **Vulnerabilities**
- **Attack surfaces**

AI difficulties
- **Abundant attack surfaces**
- **Dimensions of consequence**
- **Opacity and auditing**







Opportunities for Advancing FM and AI

AI

- AI for FM
- Modeling and proving
- Proof maintenance
- Hybrid AI for high impact
  - Auditability and trustworthiness

FM

- Usability and integration into process
- Reduce information loss -- the role of evidence
- Find the math; hide the math
- AI assist for model development and provers
- Model scope and expressiveness

Architecture

- Design for verifiability
- Resilience
- Maximizing architecture benefits
- Modeling and analysis -- move to left

Attack surface modeling

- Modeling
- Mitigation
- Work factor interventions