

THREAT MODELING
CONNECT

HACKATHON

APRIL 1-21, 2024

Brought to you by: **IriusRisk** **SHOSTACK**
+ ASSOCIATES

KEYNOTE SPEECH

Threat Modeling in the Age of AI

April 2, 11:00am ET

Speaker

Adam Shostack

President @ Shostack + Associates



HACKATHON

APRIL 1-21, 2024

Threat modeling in the age of AI

Threat Modeling Connect Hackathon 2024

Adam Shostack

adam@shostack.org

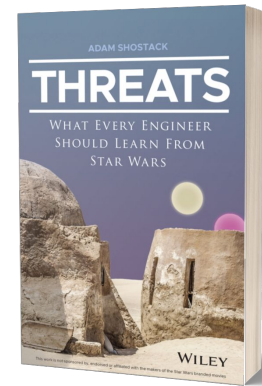
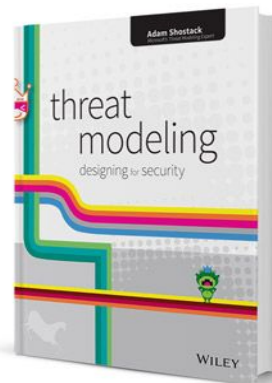
infosec.exchange/@adamshostack

About Adam Shostack

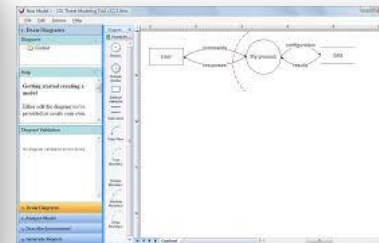
SHOSTACK

+ A S S O C I A T E S

W
UNIVERSITY *of* WASHINGTON

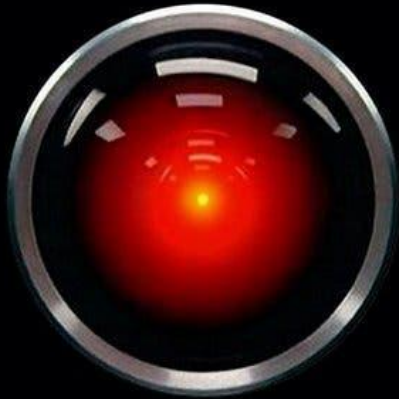



black hat[®]



IriusRisk

The age of AI



The age of AI

AD

write me a talk abstract about threat modeling in the age of ai



Title: "Navigating Cybersecurity Challenges: Threat Modeling in the Age of AI"



Abstract:

In the rapidly evolving landscape of technology, the integration of Artificial Intelligence (AI) has revolutionized various sectors, providing unprecedented opportunities and efficiencies. However, this advancement is accompanied by new and sophisticated cybersecurity threats. This abstract delves into the critical need for adapting and enhancing traditional threat modeling approaches to effectively address the evolving threat landscape in the age of AI. We will explore the intricacies of AI-powered systems, potential vulnerabilities, and innovative threat modeling strategies tailored to safeguard AI-enabled environments. By examining the convergence of AI and cybersecurity, this abstract aims to shed light on the imperative of proactive threat modeling as a fundamental component of a comprehensive cybersecurity strategy.



Orion-Cygnus

YOU ARE HERE

Perseus

Outer

Norma

3kpc

Crux-Scutum

Carina-Sagittarius

Threat modeling in 2024

- 1999 Kohnfelder + Garg create STRIDE
- 2014 *Threat Modeling: Designing for Security*
- 2020 Threat Modeling Manifesto
- 2024
 - Competition between commercial vendors
 - Coopetition in community spaces
 - OWASP Appsec dominated by TM talks!
 - ThreatModCon (Lisbon, San Francisco!)

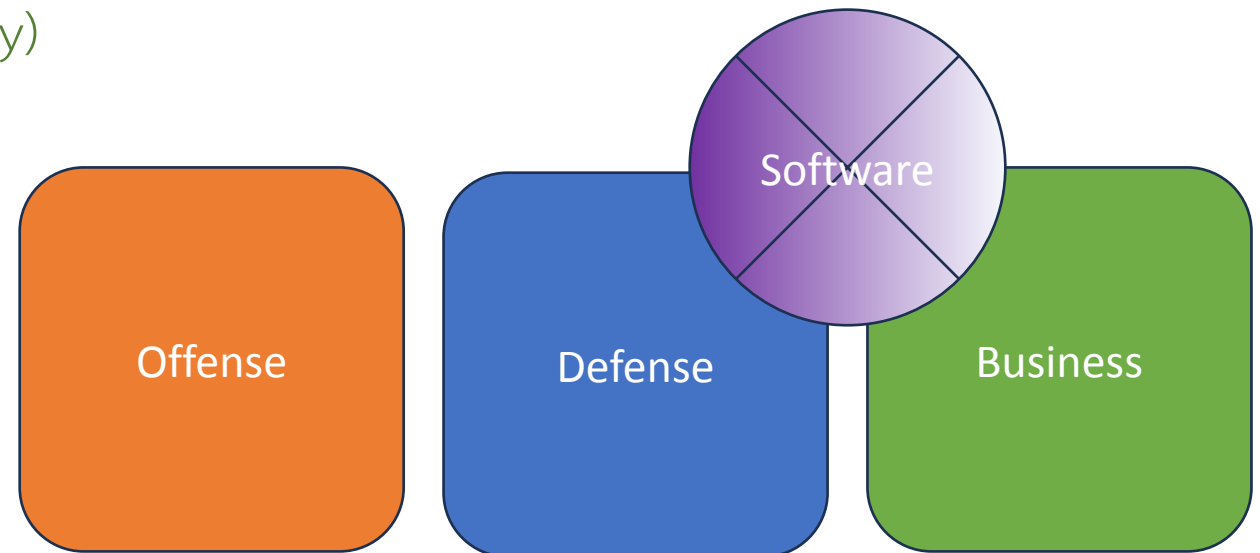
Last Year's Hackathon

- Great participation
- Very challenging judging
- Team of judges had some strong opinions

Threat modeling AI

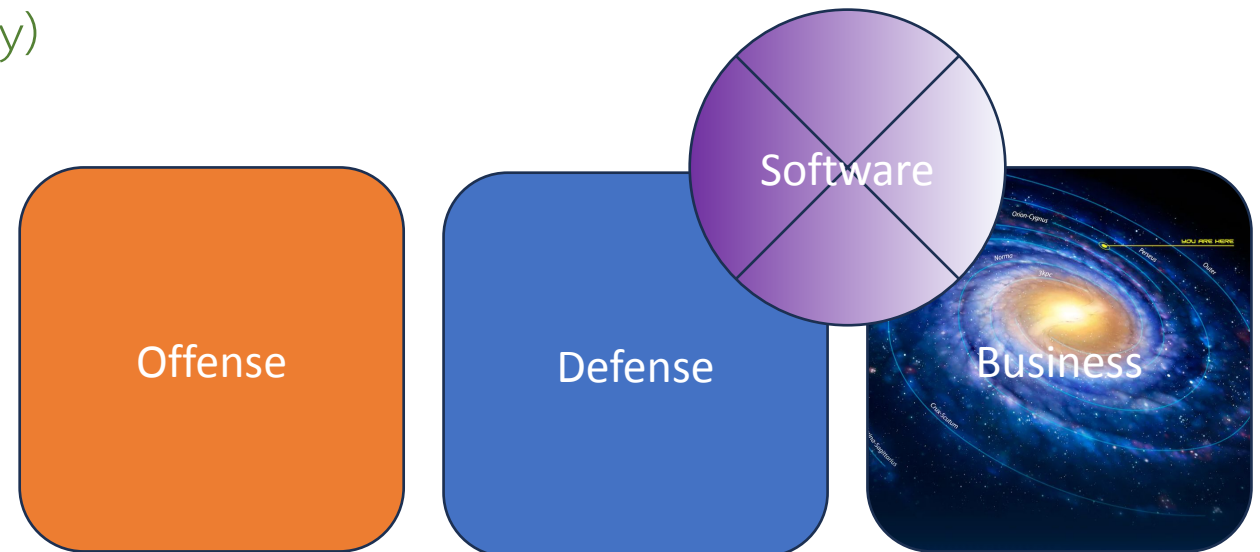
Four scenarios

- AI for offense (write me a phishing email/malware/etc)
- AI for defense (anti-spam, Microsoft defender copilot)
- AI for business (main focus today)
- AI for software development (2nd focus today)



Four scenarios

- AI for offense (write me a phishing email/malware/etc)
- AI for defense (anti-spam, Microsoft defender copilot)
- AI for business (main focus today)
- AI for software development (2nd focus today)



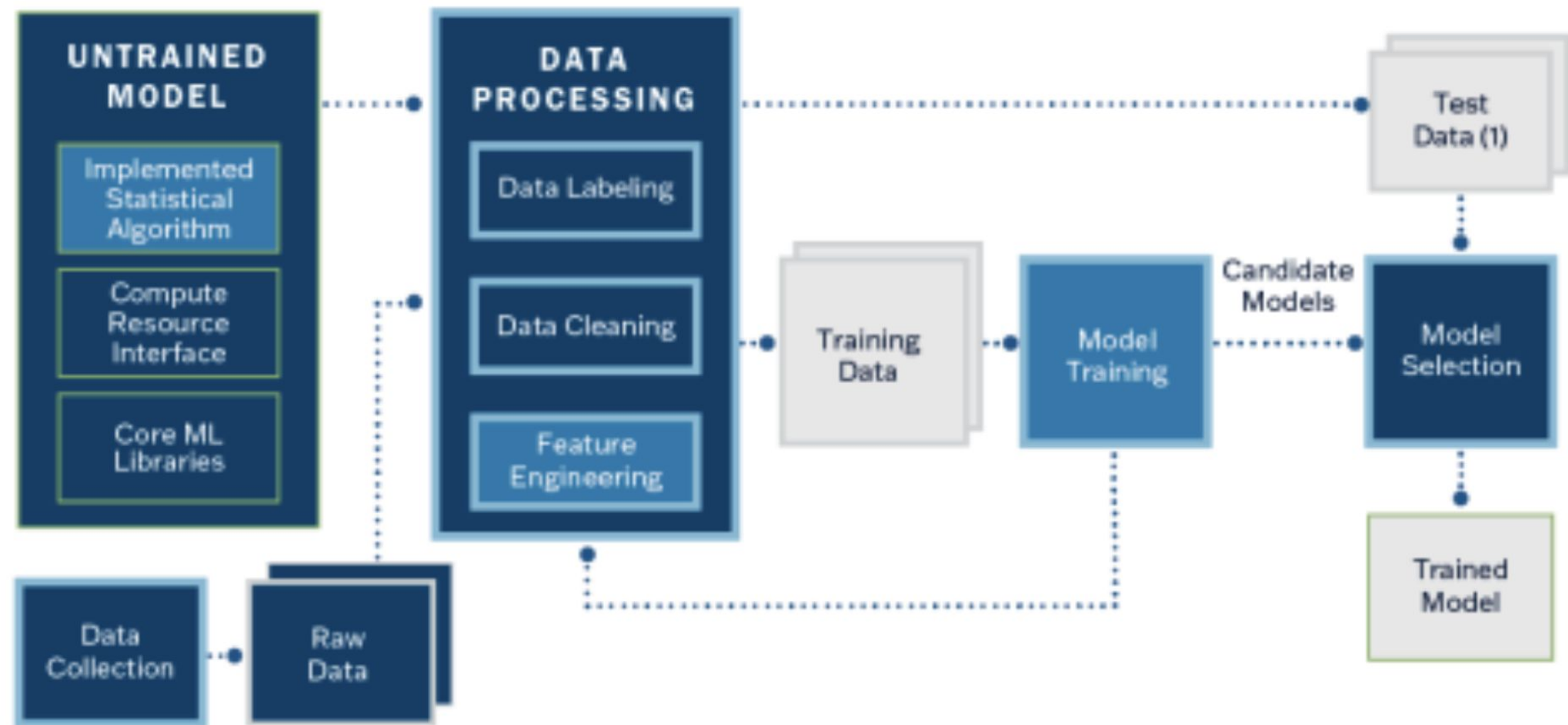
What are we working on *with AI*?

- “Adding an LLM to our business”
- We work on (next slides)*
 - Model building and validation
 - Model deployment
 - Operational environment
- We should ask “what can go wrong” with each!
 - That is, we should threat model each

* Christine Lai, Jonathan Spring, Software Must Be Secure by Design, and Artificial Intelligence Is No Exception, CISA Blog, Aug 18, 2023,
<https://www.cisa.gov/news-events/news/software-must-be-secure-design-and-artificial-intelligence-no-exception>

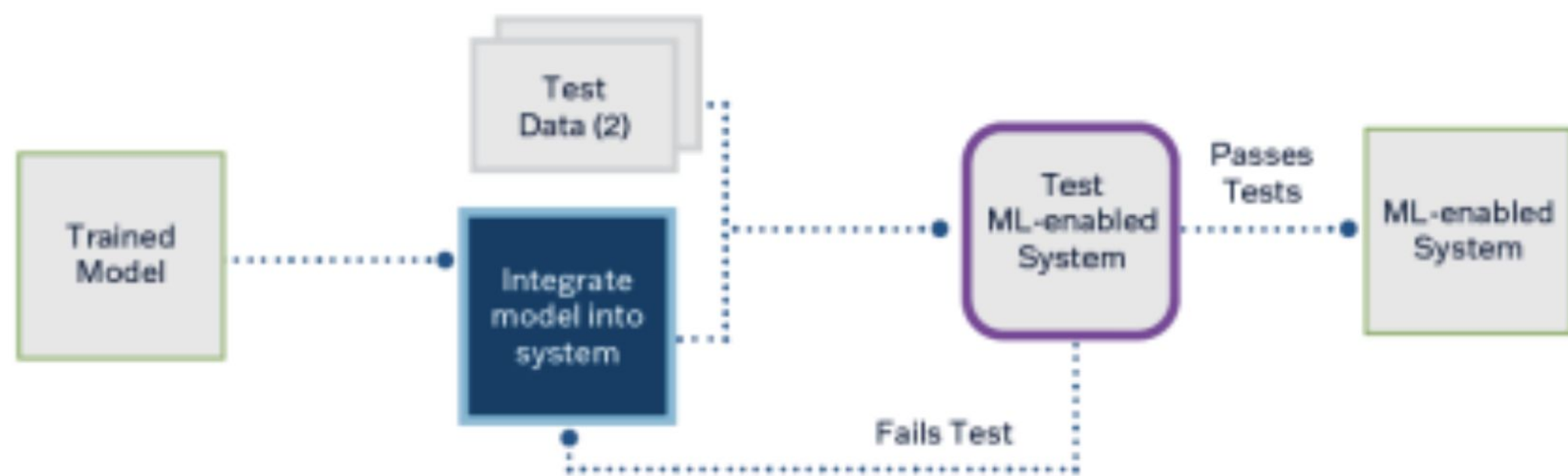


MODEL BUILDING AND VALIDATION



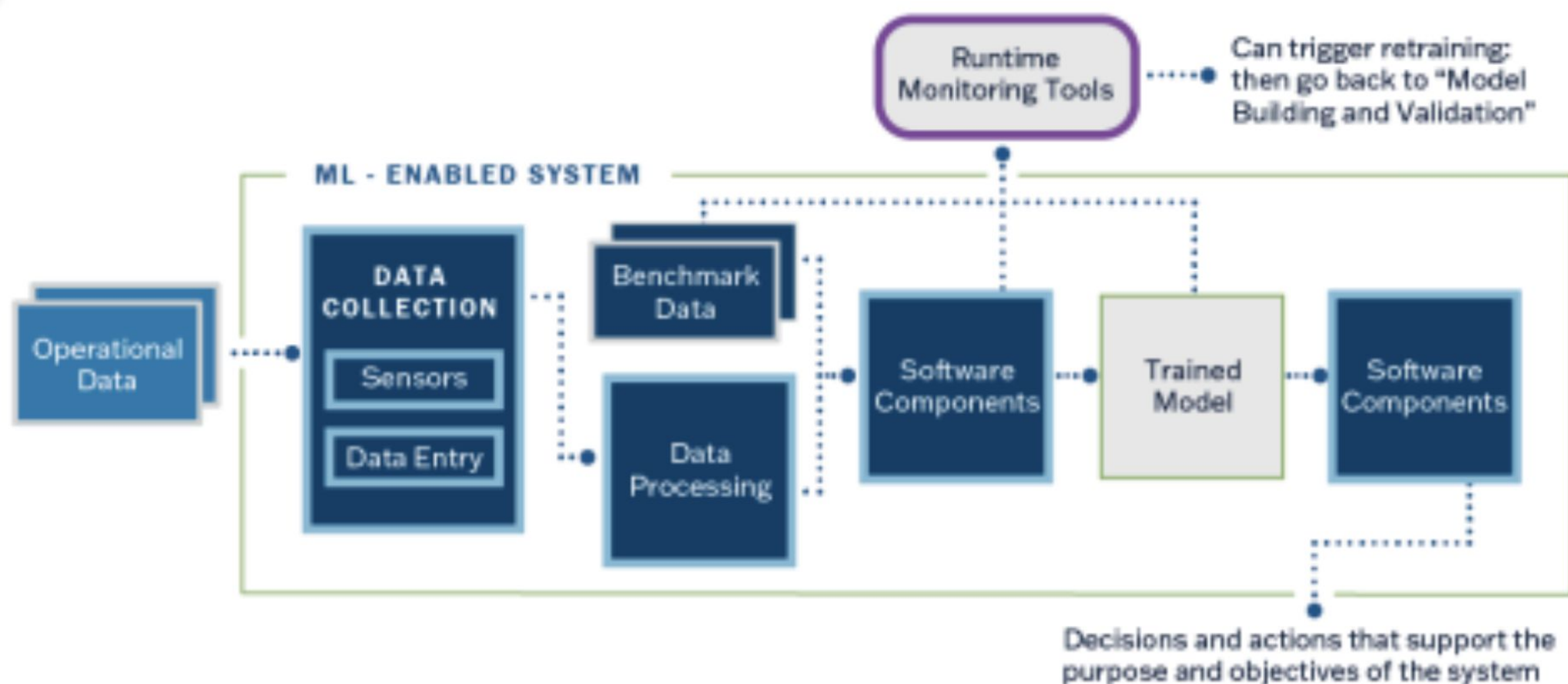


MODEL DEPLOYMENT





OPERATIONAL ENVIRONMENT +



Importance of training data

- What training data are we working on?
 - Pre-selected + curated?
 - Live internet data? (Hi Tay! 😄)
 - Customer interaction?
- How frequently do we re-train/tune/adapt?
- Where are those adapted results visible?
- Different answers allow for very different threats

Ways to discover what can go wrong *with AI*

- OWASP top ten (+ Adam's variant)
- Berryville's tools
- Microsoft's lists
- Emily Bender's De-jargoning

- Don't forget STRIDE + Kill chains still apply

OWASP Top 10 for LLM Applications

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

Adam's Top Ten (or so) candidates

- Prompt injection
 - Data leakage (OWASP's sensitive info disclose)
 - Training data poisoning
 - Over-reliance on LLM-generated content
-

- Hallucination
- Inexplicability
- Bias
- Insecure development + deployment

These match OWASP
Top10llm.com

These don't 🙄

Berryville Institute of Machine Learning

- Think tank of security + ML experts studying machine learning sec <https://berryvilleiml.com/>
- Taxonomy of threats (2019)
 - Manipulation of input, data, models
 - Extraction of input, data, models
- Architectural Risk Analysis (ARA) of a generic ML system (2023)
- ARA for LLM (2024)
- Annotated bibliography

Microsoft's Lists

- Threat Modeling AI/ML Systems and Dependencies
- Failure Modes in Machine Learning
- Securing the Future of Artificial Intelligence and Machine Learning at Microsoft
- (Links, comments: <https://shostack.org/blog/tmt-machine-learning/>)

Emily Bender's De-Jargoning

- What is being automated?
- Who's automating it and why?
- Who benefits from that automation?
- How well does the automation work in its use case that we're considering?
- Who's being harmed?
- Who has accountability for the functioning of the automated system?
- What existing regulations already apply to the activities where the automation is being used?

<https://medium.com/@emilynbender/opening-rehab-us-ai-in-the-workplace-new-crisis-or-longstanding-challenge-eb81d1bee9f>



- Here we are at the kickoff of the 2nd Threat Modeling Hackathon
- Meet your mentors soon
- Looking forward to meeting the winners

Thank you!

Questions?

Now or

adam@shostack.org