NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

Enabling SE for AI with Test and Evaluation Harnesses for Learning Systems

Tyler Cody, Ph.D.

Virginia Tech National Security Institute

900 N. Glebe Rd., Arlington, VA, 22203

Executive Summary

1. Test and evaluation (T&E) harnesses relate machine learning *components* to state information from its *subsystem, system*, and *operational environment*.

2. T&E harnesses bridge multiple, discipline-specific orientations to T&E for learning systems to the level of systems engineering processes.

Academic Mission

- I am a systems theorist and machine learning engineer.
- I develop and apply systems theory to bridge systems engineering and artificial intelligence.

My mission is to develop rigorous first-principles for engineering AI-heavy systems.

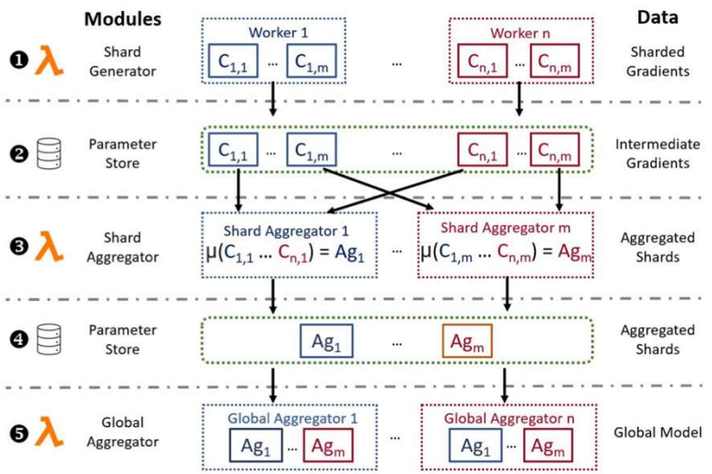
"In order to improve your game, you must study the endgame before everything else, for whereas the endings can be studied and mastered by themselves, the middle and the opening must be studied in relation to the endgame."

- Jose Raul Capablanca

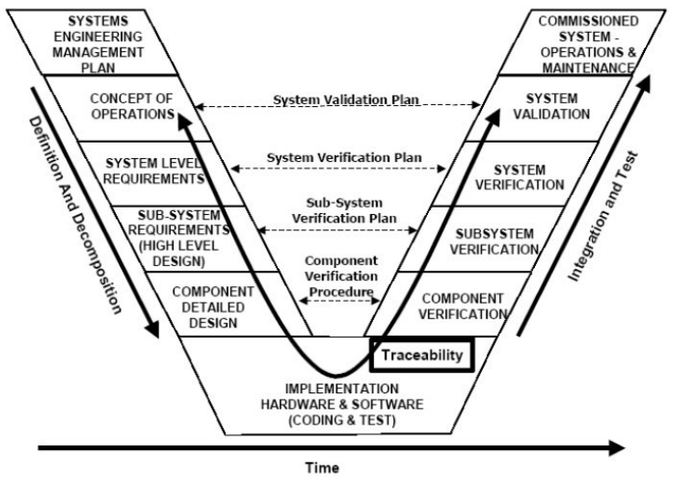
Scope

$$
\begin{aligned}
\textbf{minimize:} \quad & f(\mathbf{x}) \\
& \mathbf{x} \in \Re^n \\
\textbf{subject to:} \quad & \mathbf{g}_L \leq \mathbf{g}(\mathbf{x}) \leq \mathbf{g}_U \\
& \mathbf{h}(\mathbf{x}) = \mathbf{h}_t \\
& \mathbf{a}_L \leq \mathbf{A}_i\mathbf{x} \leq \mathbf{a}_U \\
& \mathbf{A}_e\mathbf{x} = \mathbf{a}_t \\
& \mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U
\end{aligned}
$$

AI Research



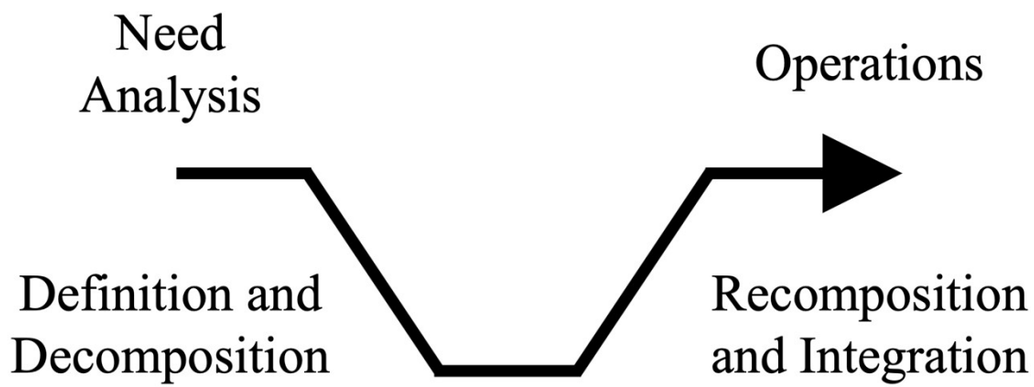
AI Engineering Research



AI Systems Engineering Research

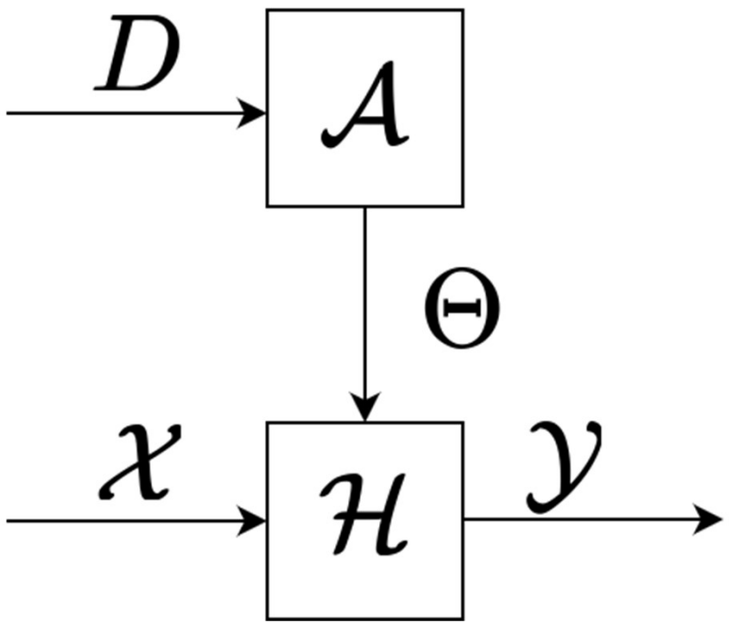
Systems Engineering

1. Need analysis, stakeholder value elicitation, requirements
2. System definition and decomposition
3. Recomposition and integration
4. Operations, Maintenance, Retirement



The canonical systems engineering “V” process.

Learning Systems

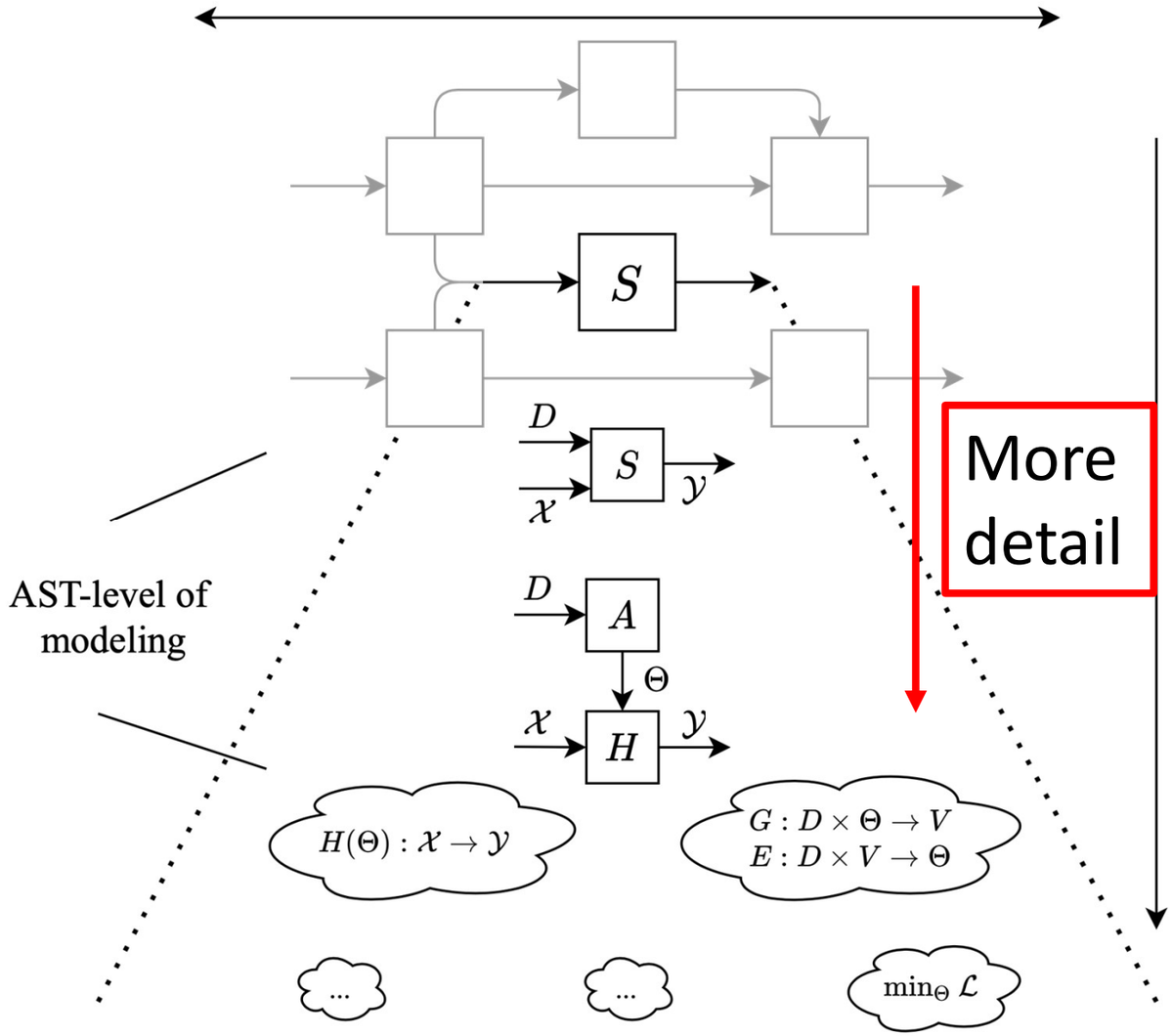


Term	Definition
Learning system	$S \subset \{\mathcal{A}, D, \Theta, \mathcal{H}, \mathcal{X}, \mathcal{Y}\}$
Learning algorithm	$\mathcal{A} : D \to \Theta$
Hypotheses	$\mathcal{H} : \Theta \times \mathcal{X} \to \mathcal{Y}$
Learned model	$\mathcal{H}(\theta) : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta$

Cody, Tyler. “Mesarovician Abstract Learning Systems.” *International Conference on Artificial General Intelligence*. Springer, Cham, 2021.

Stratifying and Specifying Learning Systems

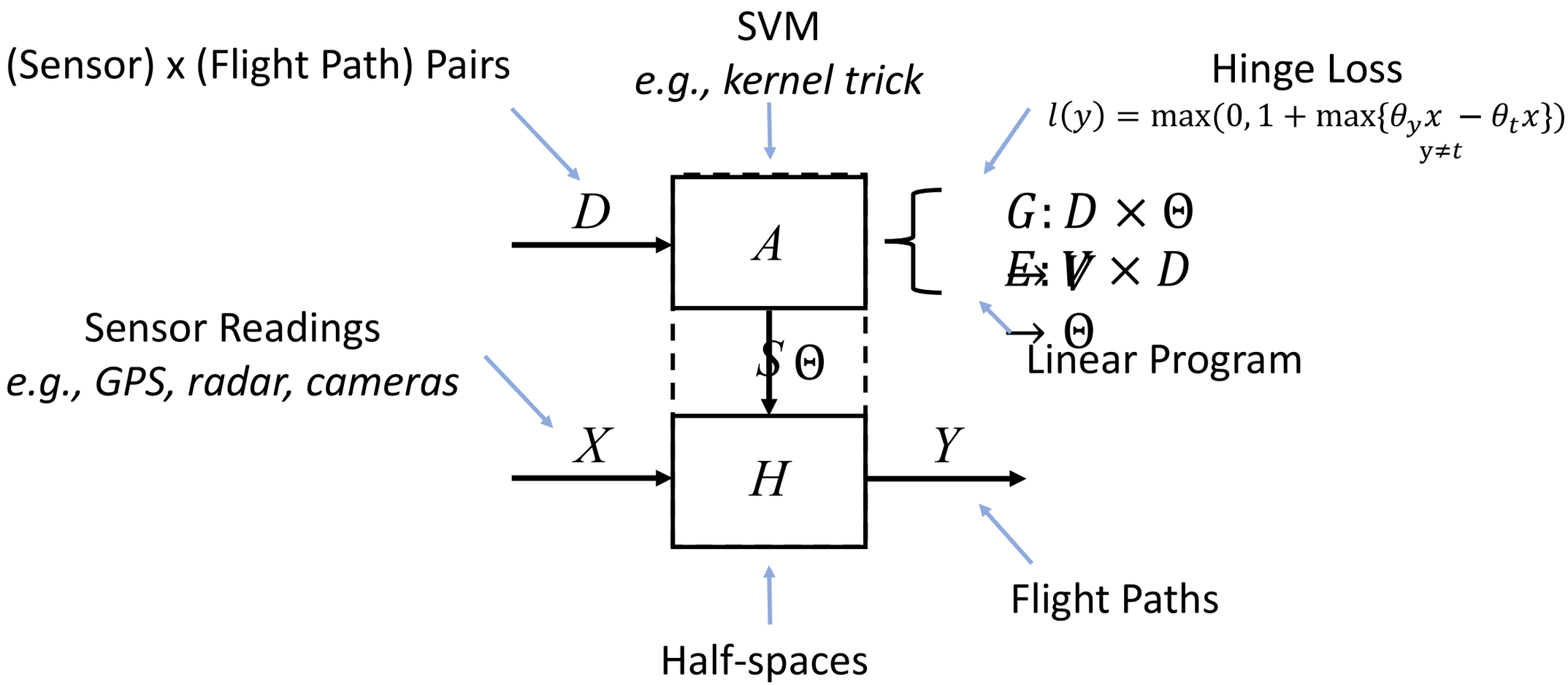
- Use systems theory to model learning systems **top-down**



Cody, Tyler. “Mesarovician Abstract Learning Systems.” *International Conference on Artificial General Intelligence*. Springer, Cham, 2021.

Learning Systems in UAS

Suppose choice of Support Vector Machines (SVMs)



ML-Oriented T&E

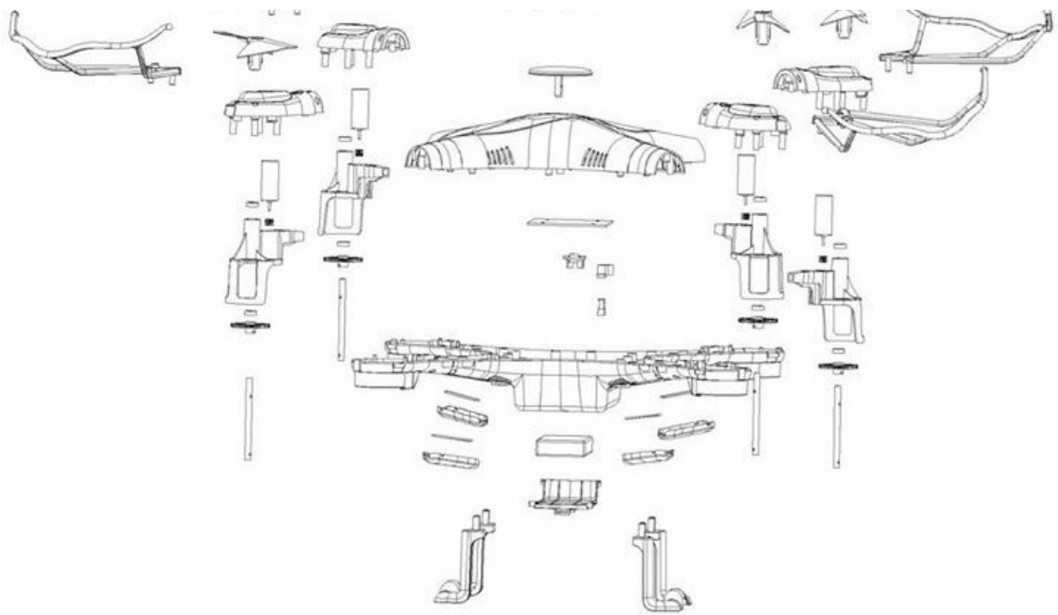
- The “independent and identically distributed” (IID) assumption of statistics bears out in the T&E of learning algorithms in the form of held-out training data and cross-validation
- Learning systems *S* contain learning algorithms *A* and the learned model $H(\theta)$.
- While it may make sense to use IID data to check if learning algorithm *A* is working as intended, it apparently falls short of testing if a given learned model $H(\theta)$ exhibits satisfactory operational performance across a given range of expected operational conditions.

Software-Oriented T&E

- From a software perspective, learning systems are “data + software”
- The software part is treatable as usual; the data part is not
- Software-oriented T&E cares less about IID assumptions and more about the input-output relationships of the learned function (e.g., metamorphic testing)
- Also, emerging topic of data assurance (and quality)
- While better aligned with the concept of functional requirements, it under-emphasizes the cyber-physical nature of learning systems; ML solutions dependent not just on data, but also on the physicality of learning systems and the processes generating their inputs

Hardware-Oriented T&E

- Hardware-oriented perspective increases the scope further, to begin to consider the cyber-physical nature of machine learning
- More narrowly scoped research considers the immediate needs of ML like compute, storage, and power
- More broadly scoped research considers the platform system more generally (e.g., comms, sensing)



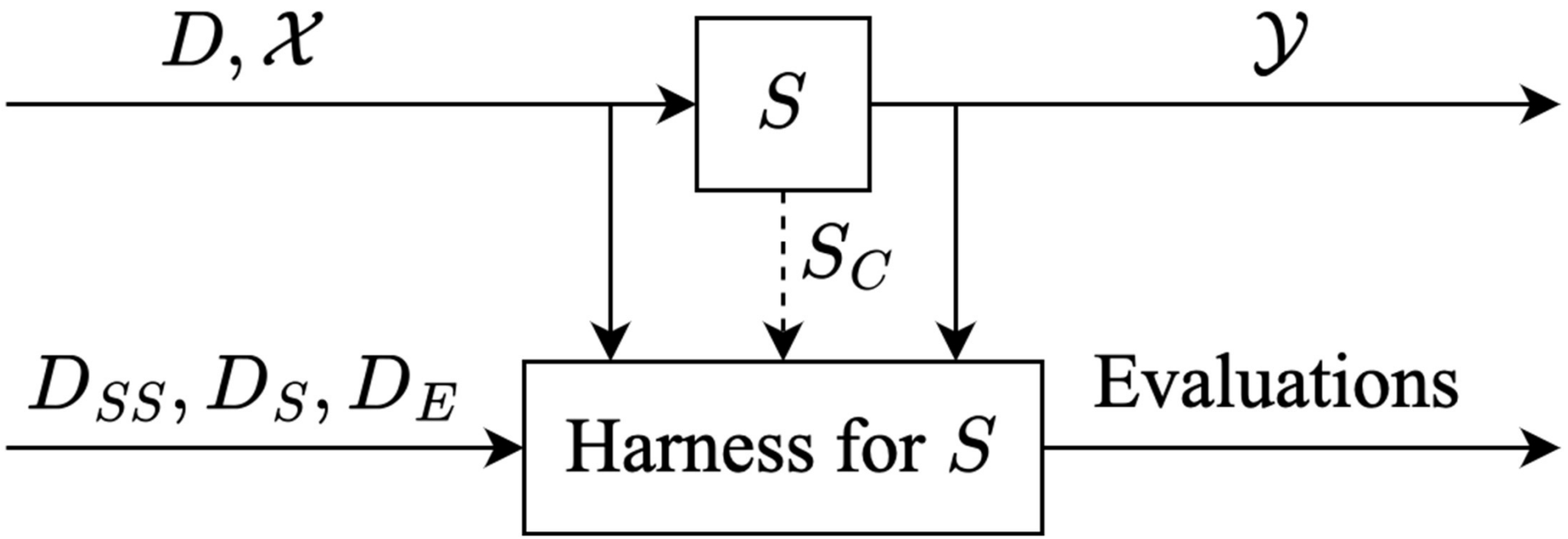
Takeaways

In short, ML-oriented T&E can occur as part of software-oriented T&E, and software-oriented T&E can be replicated, at least partially, during hardware-oriented T&E. **The various orientations are synergistic**.

- The ML-orientation tests a sub-component of the component learning system *S*, the learning algorithm *A*.
- The software-orientation test the component *S* as a whole, including software and data *D*.
- The hardware-orientation considers testing after subsystem integration.

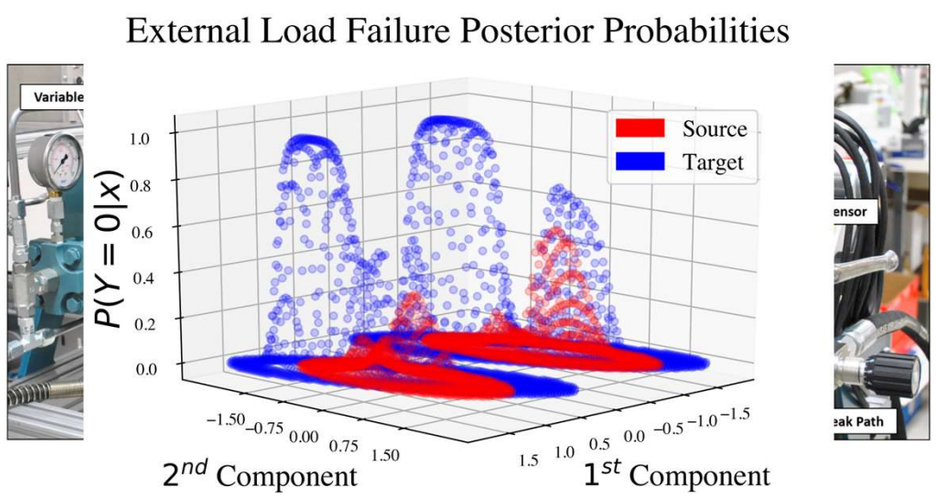
T&E Harnesses for Learning Systems

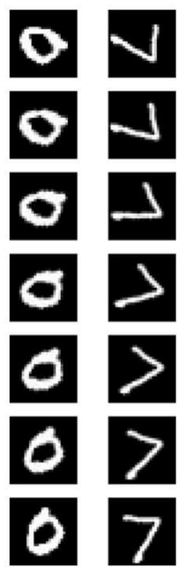
T&E harnesses for learning systems are input-output systems that take state information from the ML solution, its subsystems, systems, and operational environments as input and give evaluations as output.

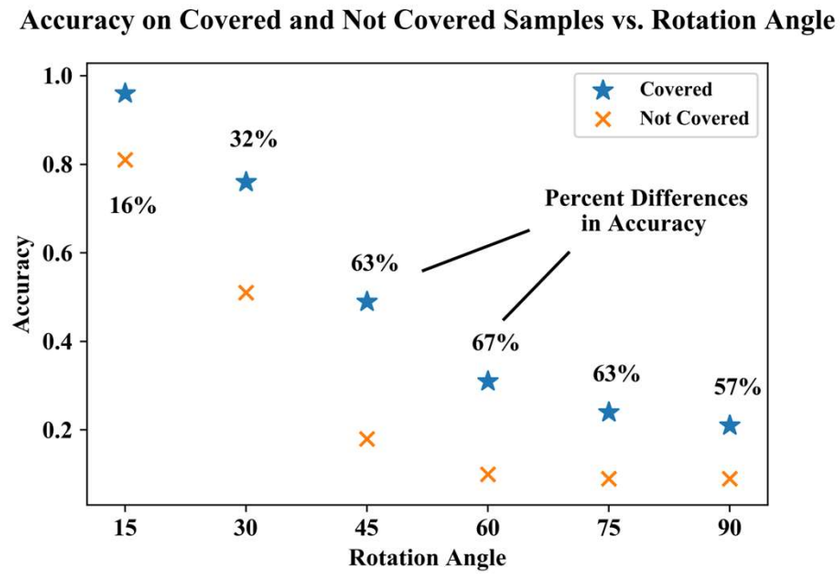


Evidence for Relevance of System State

Lessons Learned: Out-of-distribution performance is expected to be low.





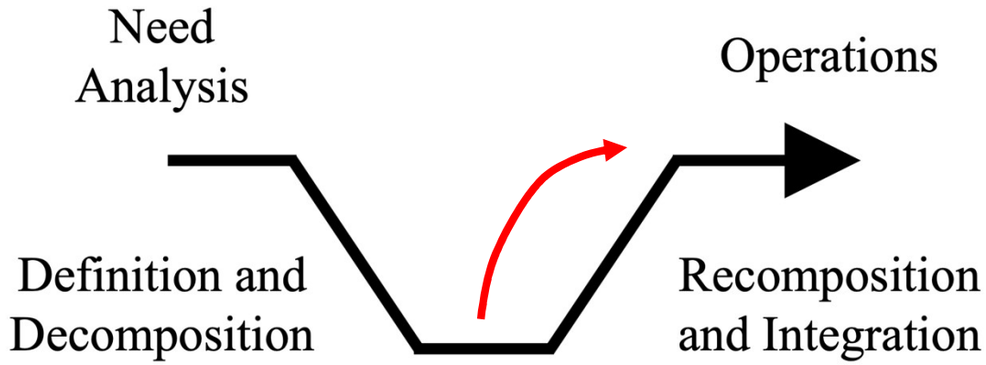


Cody, Tyler, Stephen Adams, and Peter A. Beling. "Empirically measuring transfer distance for system design and operation." *IEEE Systems Journal* (2022).

Cody, Tyler, et al. "Systematic training and testing for machine learning using combinatorial interaction testing." *2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2022.

Reshaping the “V” process

- T&E harnesses enable the flattening of the “V” model by shifting the focus of T&E for learning systems from component-level outputs to systems-level outcomes.
- From a software engineering perspective, this flattening of the “V” model corresponds to the emerging software engineering paradigm termed DevSecOps



State information brings ML solution development closer to “launch”

On Acquisition, Developmental Test

- It is unclear how T&E processes that focus on model accuracy on held-out data are able to give assurance that the model will achieve needed outcomes during operation
- By focusing on held-out data, current practice narrowly scopes themselves to component-level testing, implicitly assuming that if an ML solution meets its component-level functional requirements, then it will provide needed outcomes after aggregation with the rest of the system
- T&E harnesses tether ML solutions to the systems (e.g., platform, mission) wherein they operate, ensuring evaluations are contextualized

On Operations, Operational Test

- The no free lunch theorem of statistical learning theory suggests that no single model can be optimized for all conditions at once
- Domain adaptation is the rule, not the exception, and so-called universal models, e.g., general purpose vision models, are the exception, not the rule
- T&E harnesses provide a construct for coordinating the therefore necessary continuous T&E and continuous re-engineering of ML solutions.

Conclusion and Future Directions

- T&E harnesses anchor ML solutions to the subsystems, systems, and environments within which they operate
- The use of T&E harnesses spans the systems engineering “V”
- T&E harnesses are well-suited for the burgeoning disciplines of digital and model-based systems engineering
- Further studies are needed to develop standard reference architectures for T&E harnesses and to develop meta-models that codify their role in digital engineering processes.

Contact

Tyler Cody, Ph.D.
Research Assistant Professor
Intelligent Systems Division, Virginia Tech National Security Institute
tcody@vt.edu

LinkedIn, ResearchGate as “Tyler Cody”

Cody, Tyler, Peter Beling, and Laura Freeman. “Test and Evaluation Harnesses for Learning Systems.” *2022 IEEE AUTOTESTCON*. IEEE, 2022.