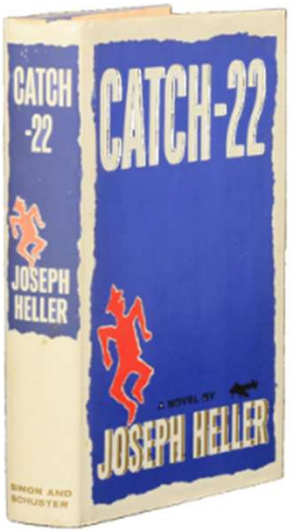
IDA

A HellerVVA Problem: The Catch-22 for Simulated Testing of Fully Autonomous Systems



Dr. Daniel J. Porter

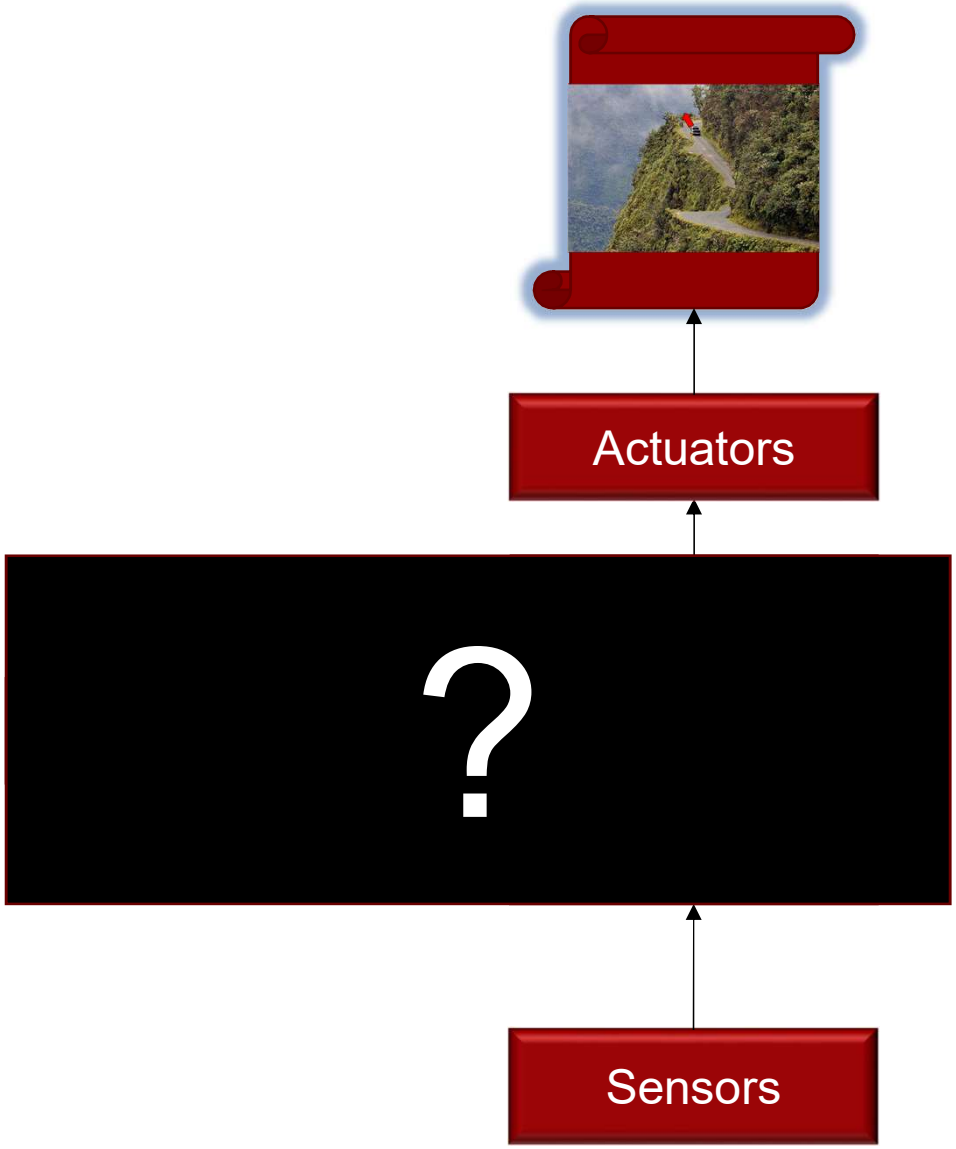
May 15th, 2020

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

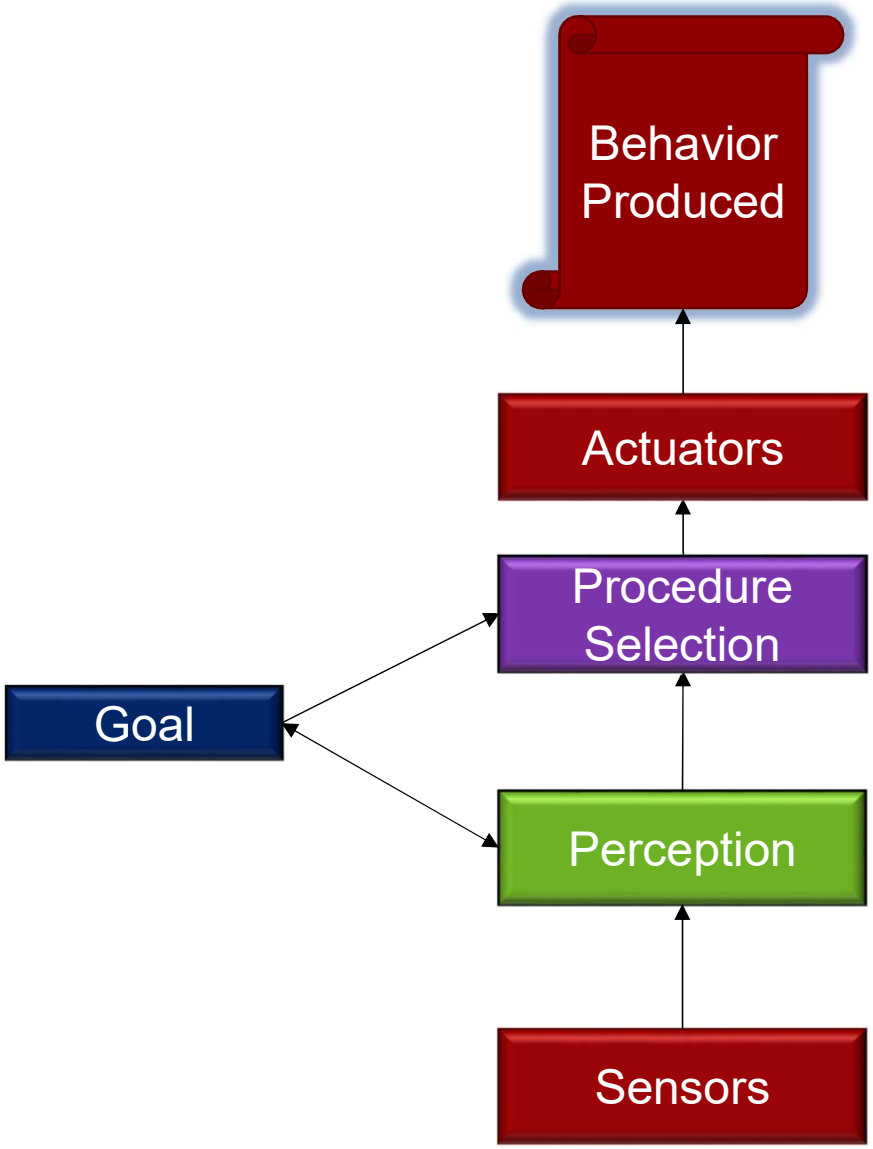
The ability to make valid inferences is the best defense against unintended behaviors.

Inferring behavior requires understanding the decisions that causally drive those behaviors

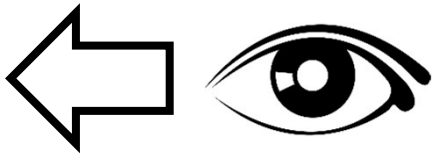
We cannot generalize behavior from black boxes



Perception, goals, and procedure selection are the basic decisions that drive behaviors



Diagnosing unintended behavior will require unobtrusive instrumentation on decision processes



"Cognitive Instrumentation"

We need to instrument these across entire system lifecycle

Actuators

Procedure Selection

Goal

Perception

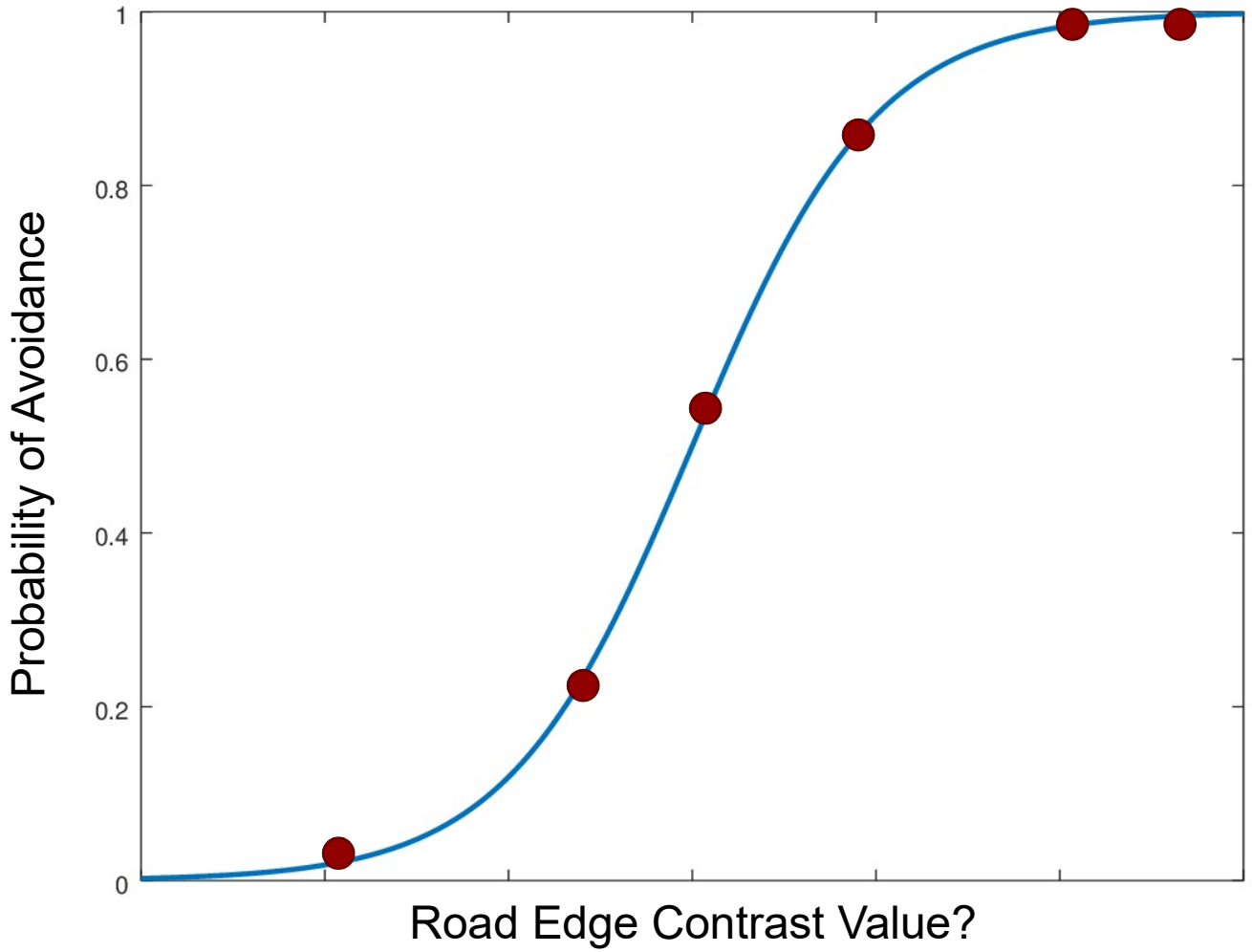
Sensors

We're used to instrumenting these in test

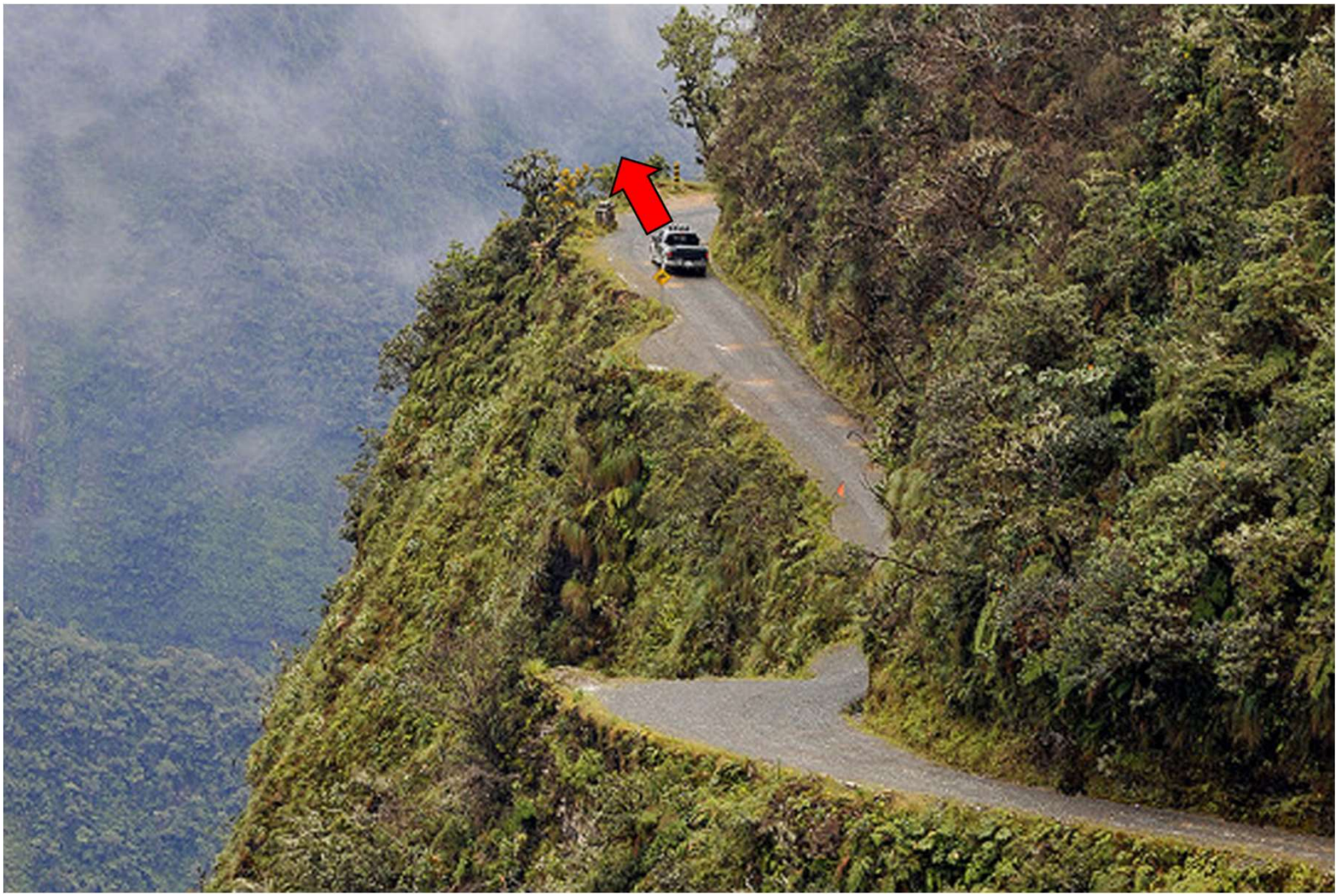
Correlation == Causation

(as least to Machine Learning)

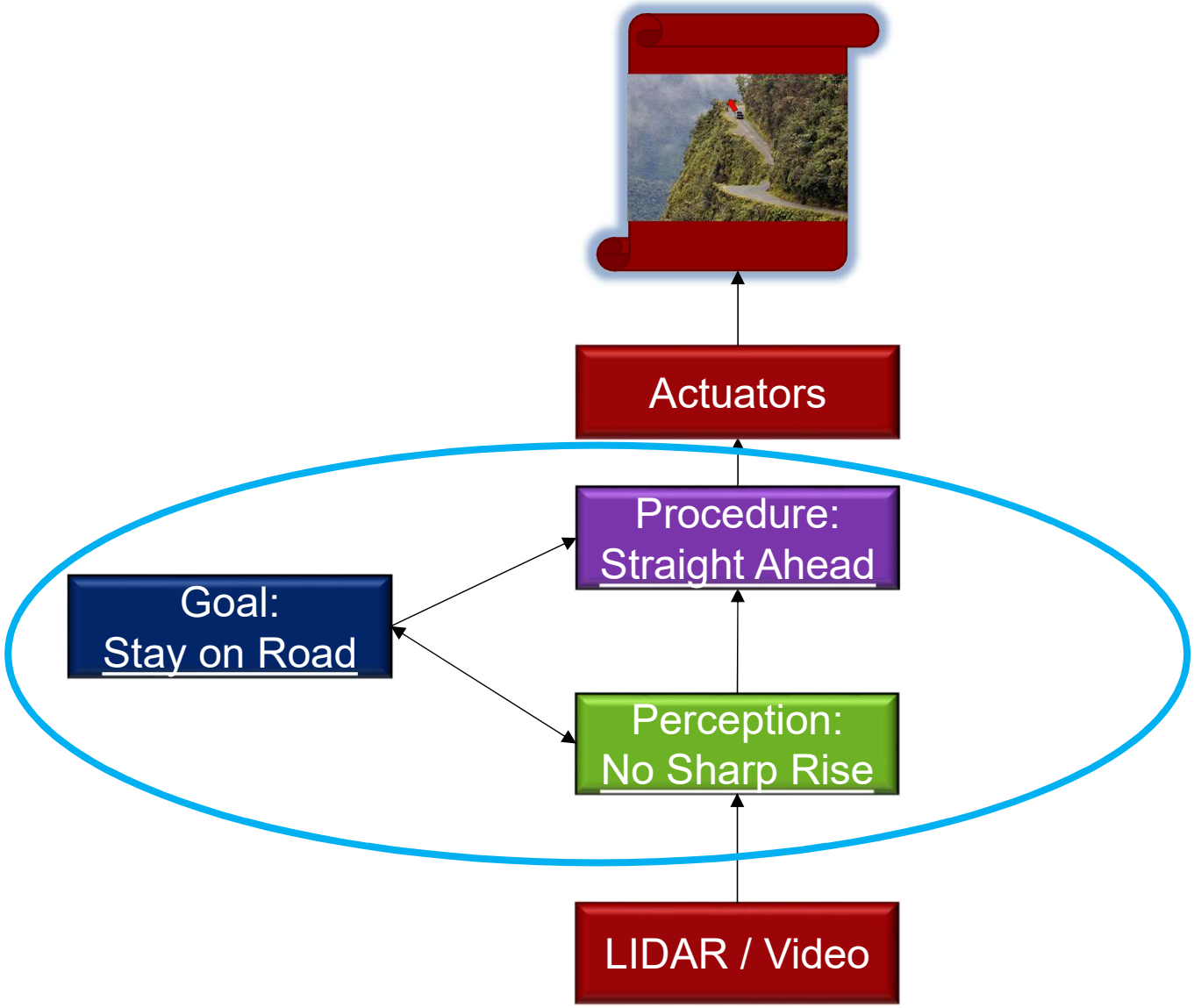
We ultimately want to validly generalize across information dimensions to avoid unintended behaviors



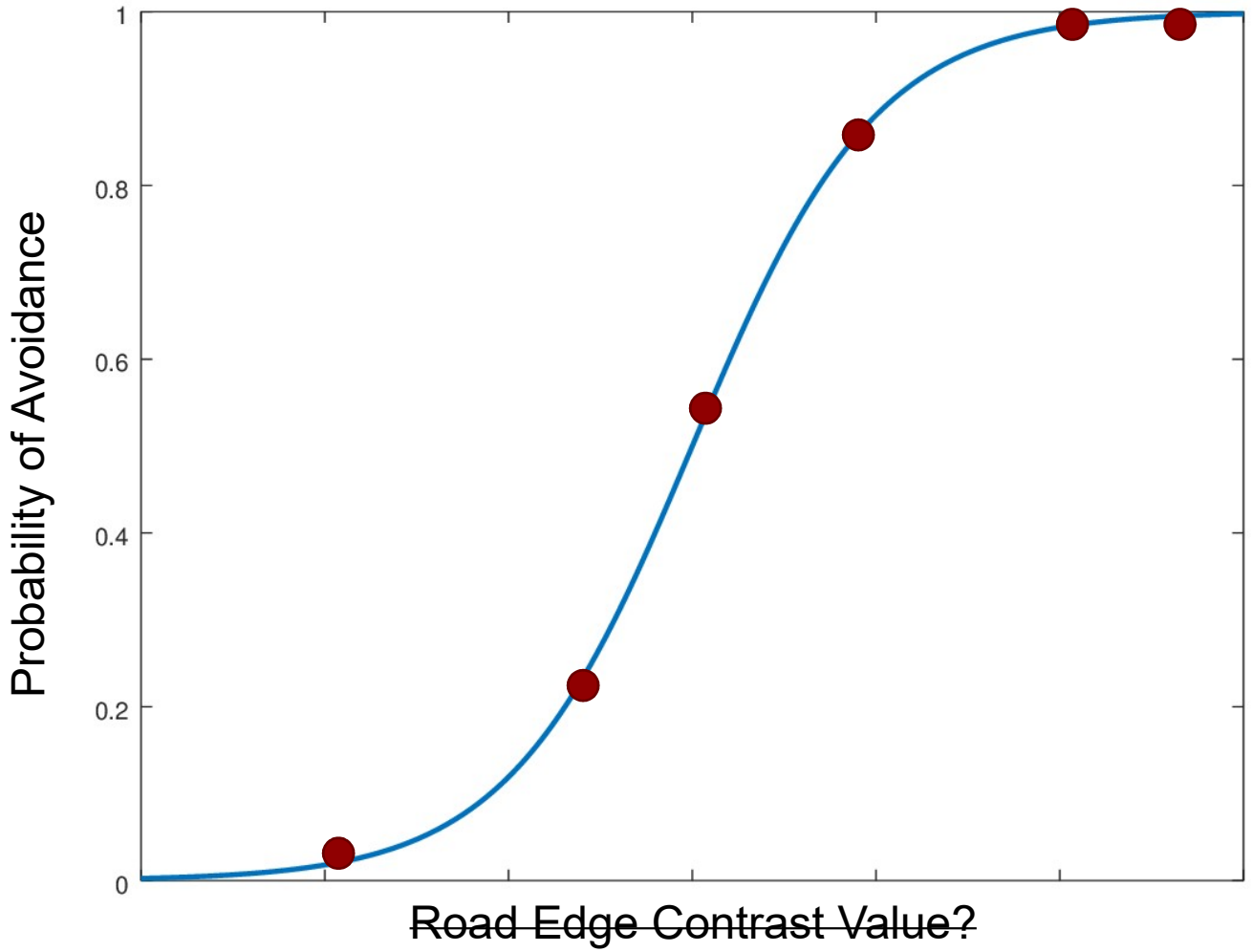
Test points can help invalidate assumptions about decision making processes



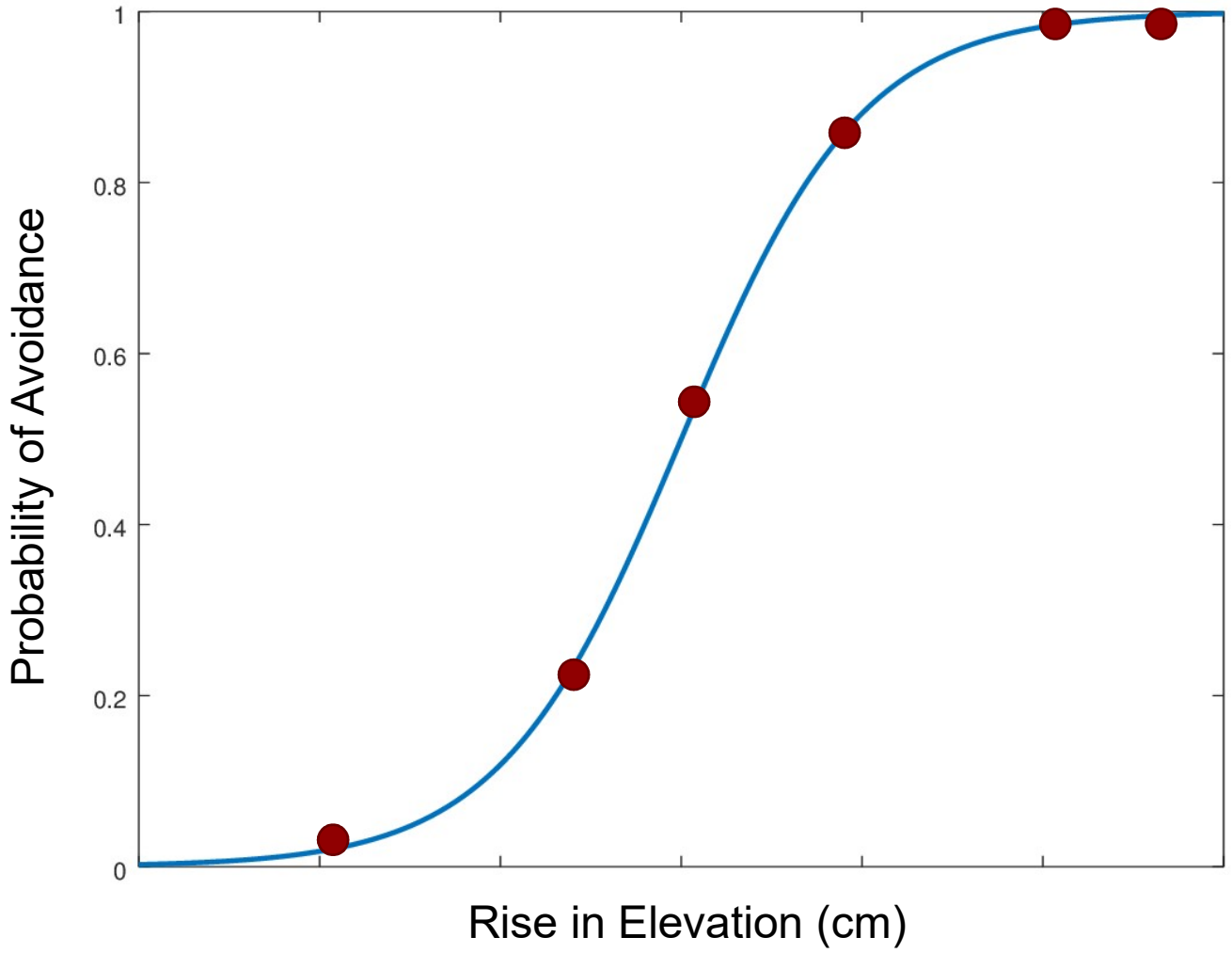
We have to obtain, verify, validate, and accredit models of system decision making



We need to ensure the information dimensions varied in test are the causal drivers and not just correlated

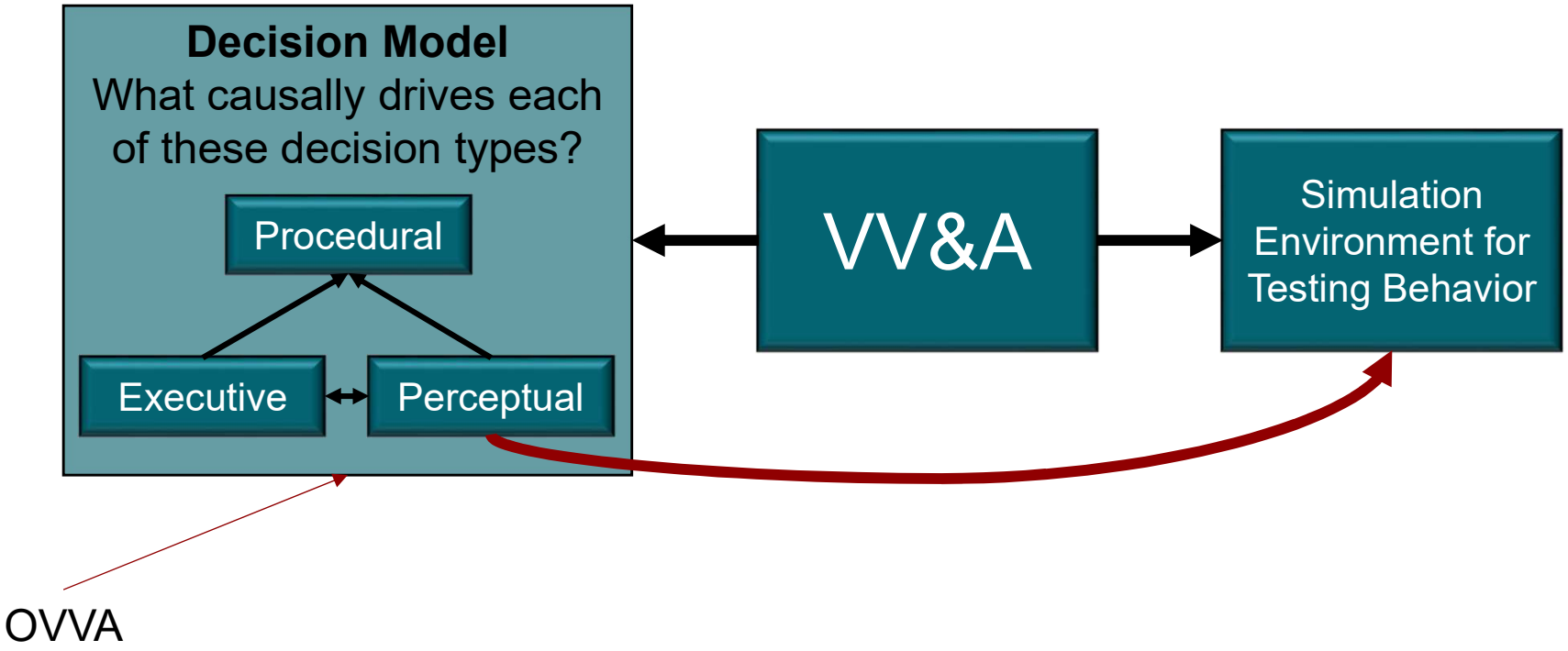


We need to ensure the information dimensions varied in test are the causal drivers and not just correlated



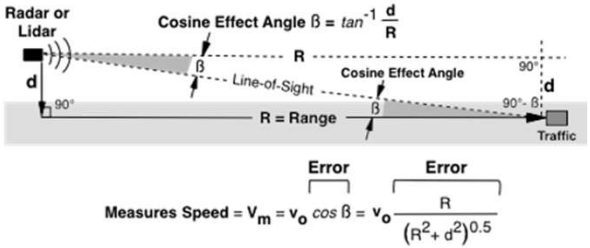
How to obtain, verify, validate, and accredit system decision models

VV&A needs to happen for more than one thing



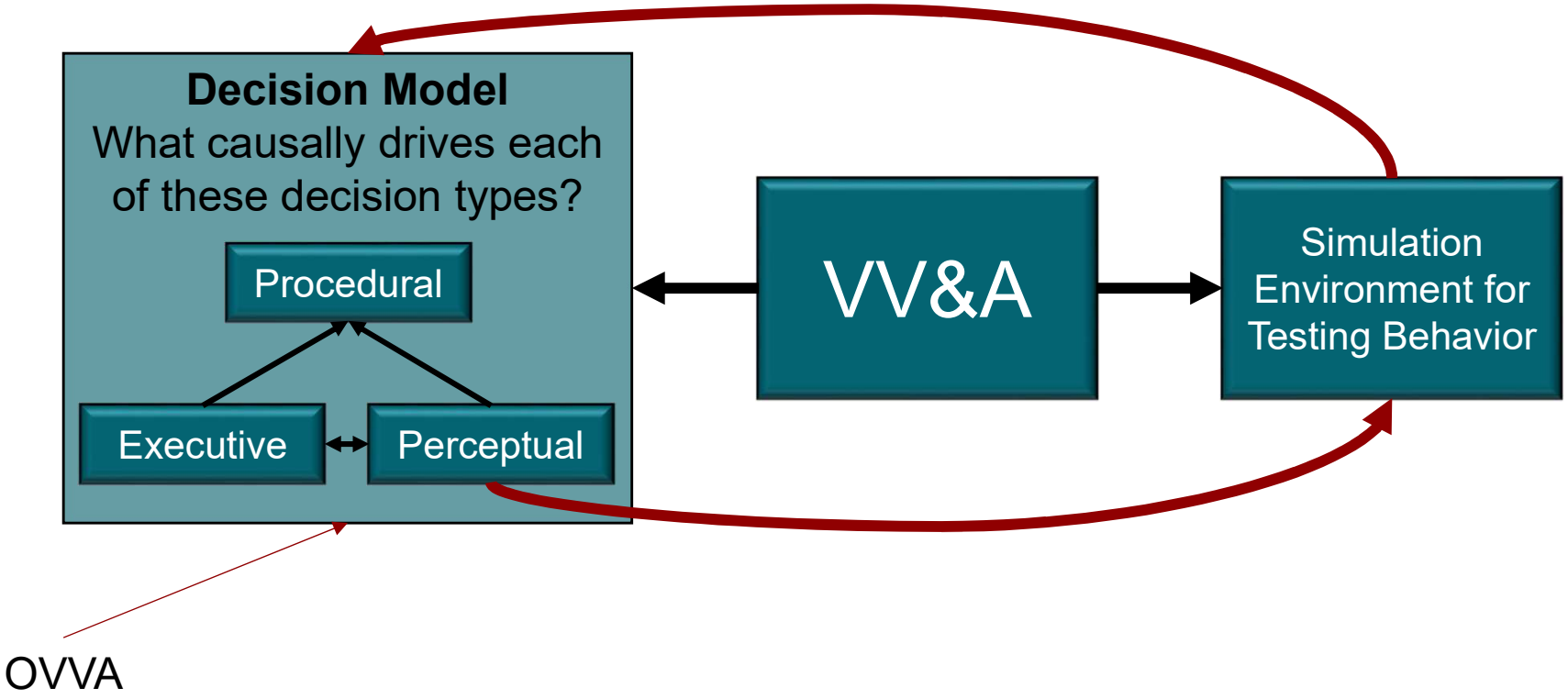
Sensor physics can be valid without the environmental features being valid and representative

θ is a causal driver of threat perception



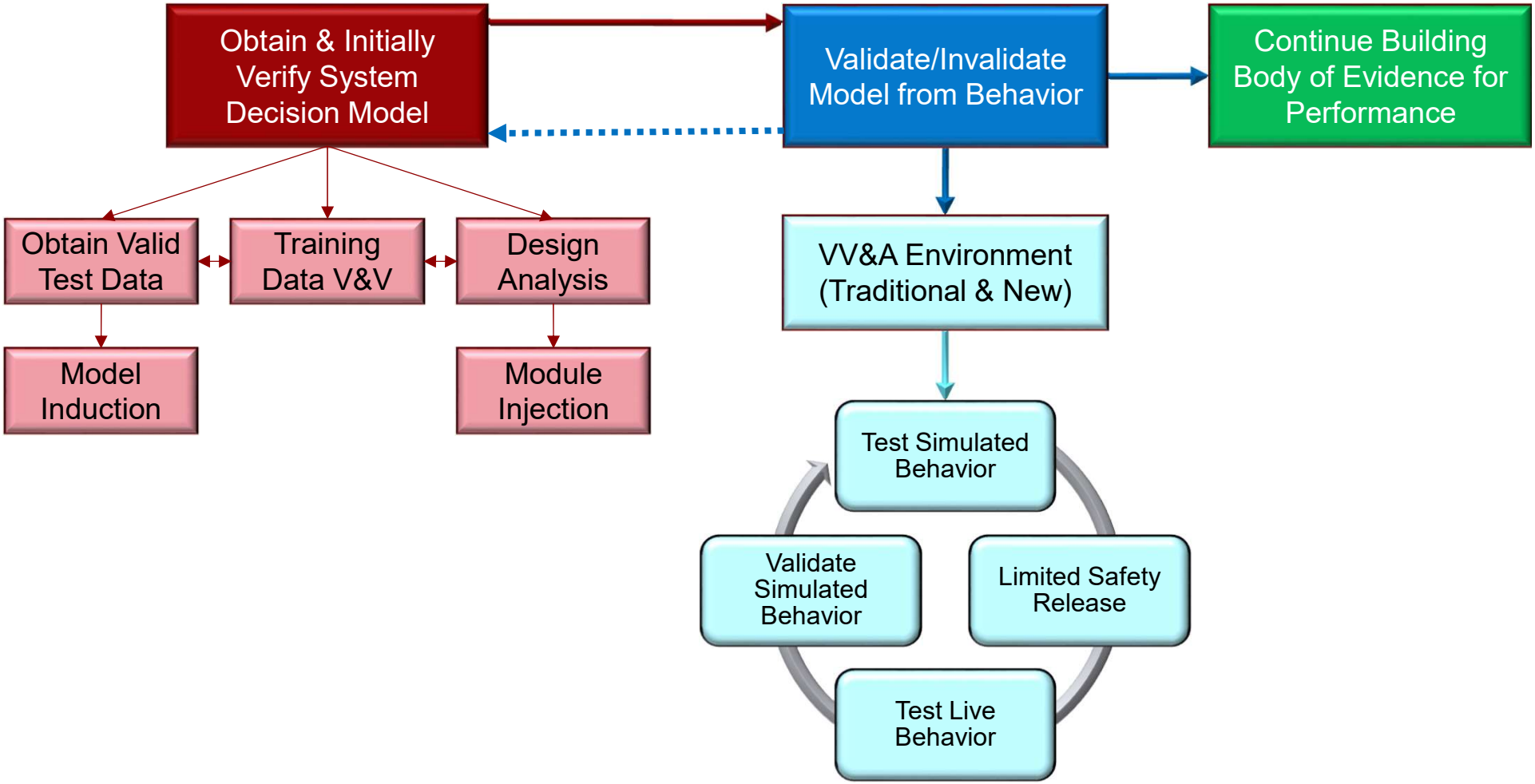
Behavioral sim doesn't vary barrel angle

VV&A needs to happen for more than one thing

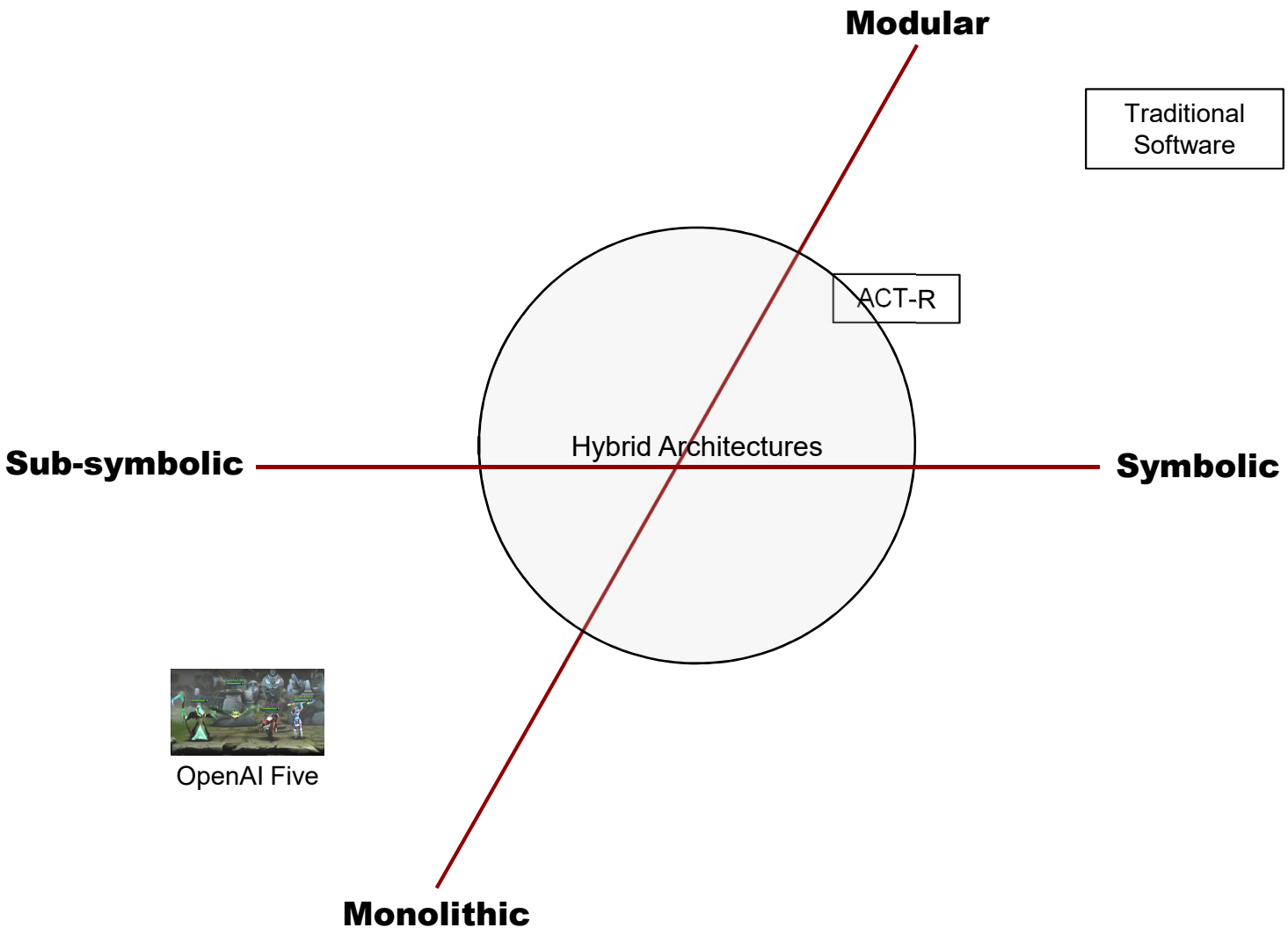


OVVA requires iterative test and evaluation

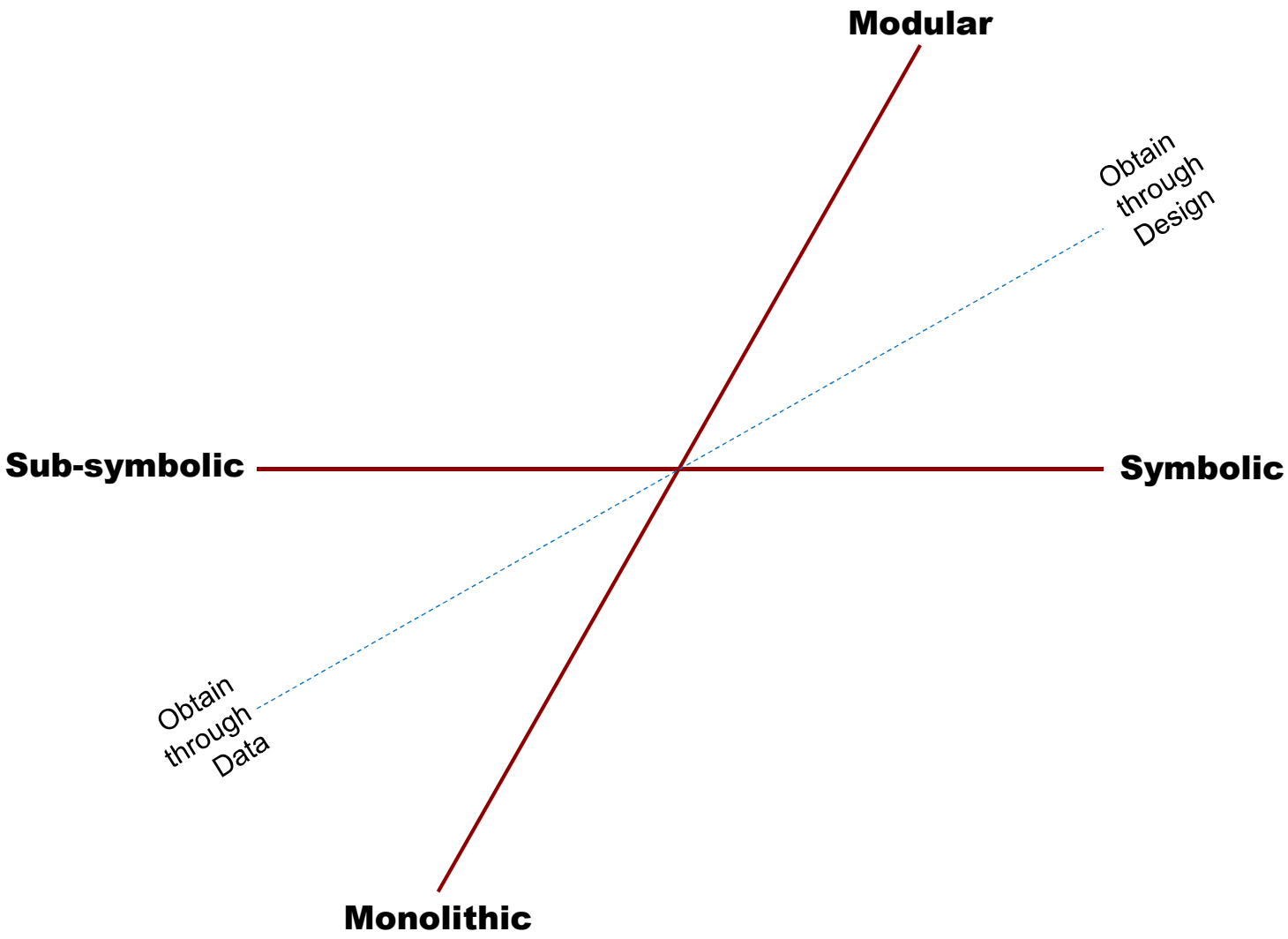
OVVA requires iterative test and evaluation



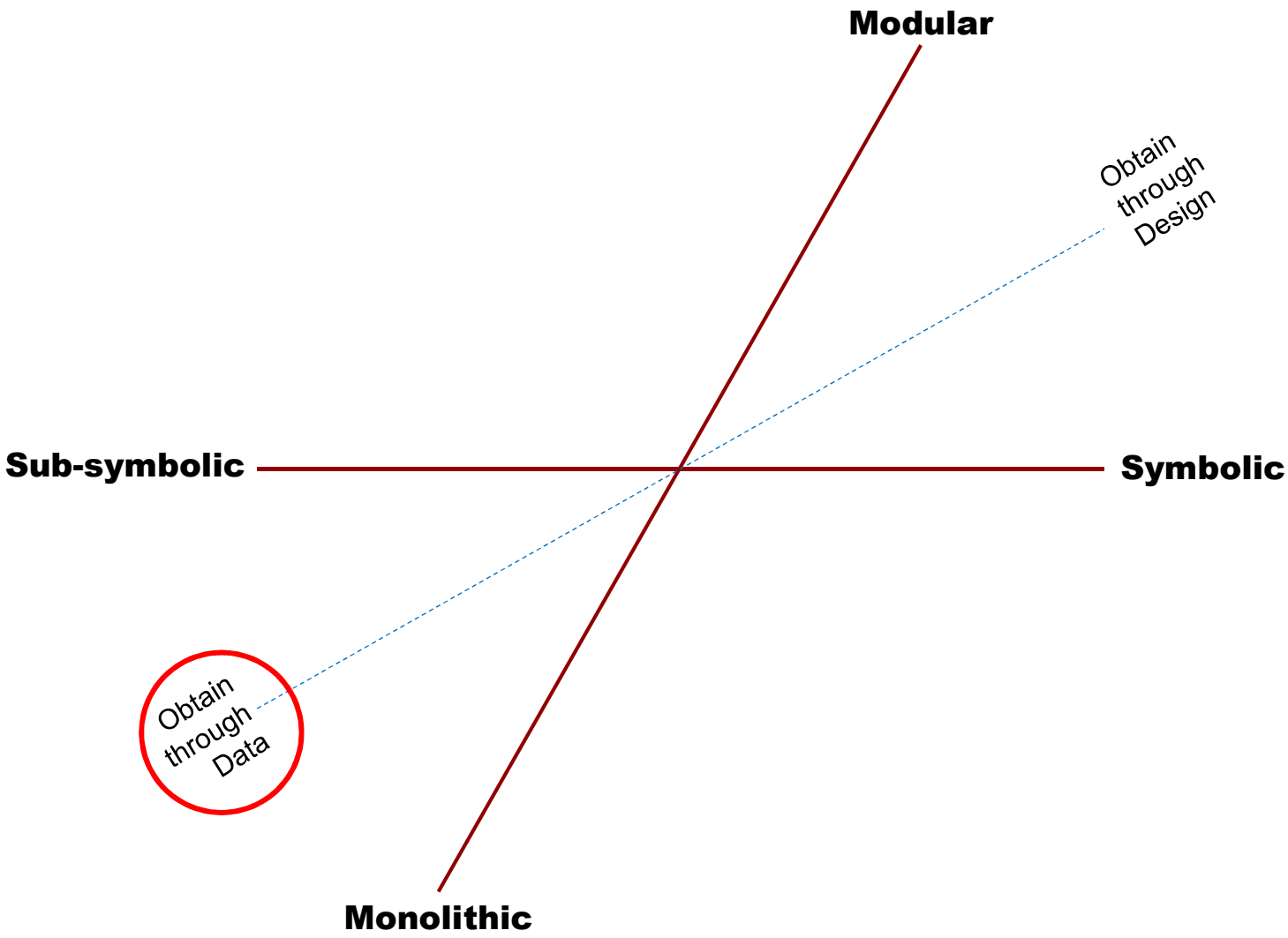
System design alters how to obtain decision model



System design alters how to obtain decision model

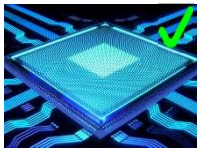


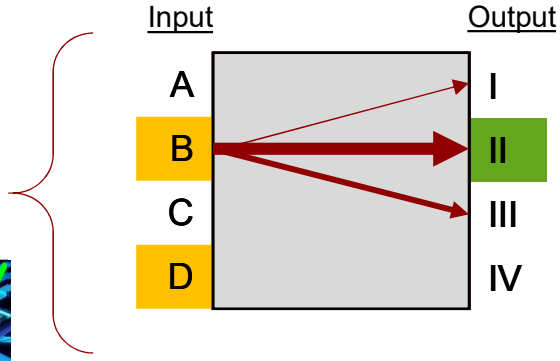
System design alters how to obtain decision model



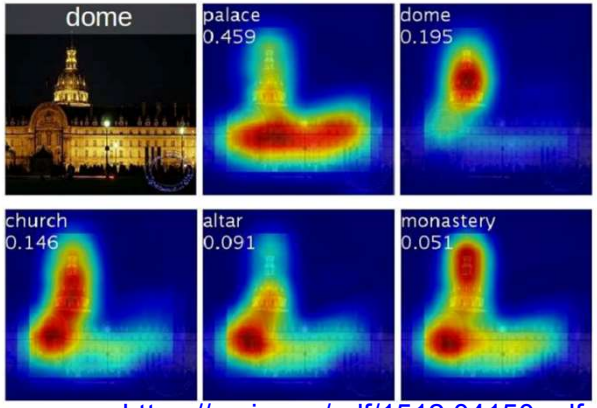
Model induction has promising data-driven techniques, but may be insufficient for embedded full autonomy

Assumption:
Have valid inputs



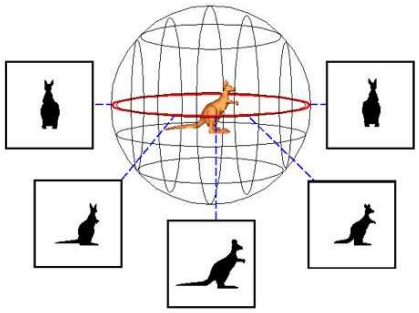


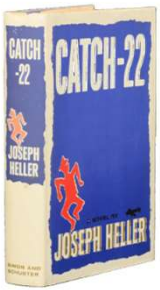
E.g., Saliency Mapping



https://arxiv.org/pdf/1512.04150.pdf

Full autonomy can change the information acquired

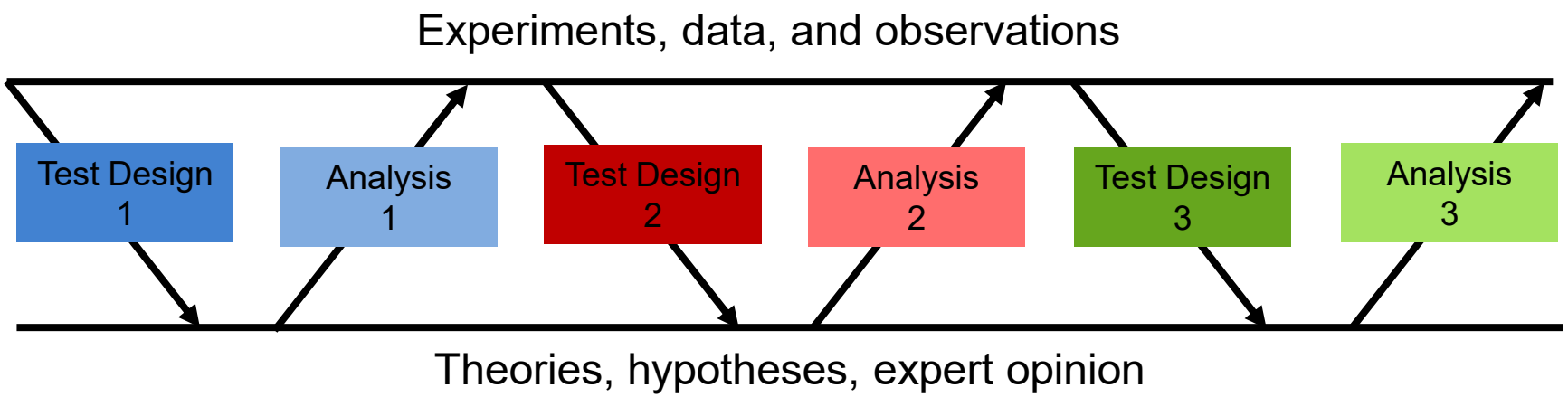


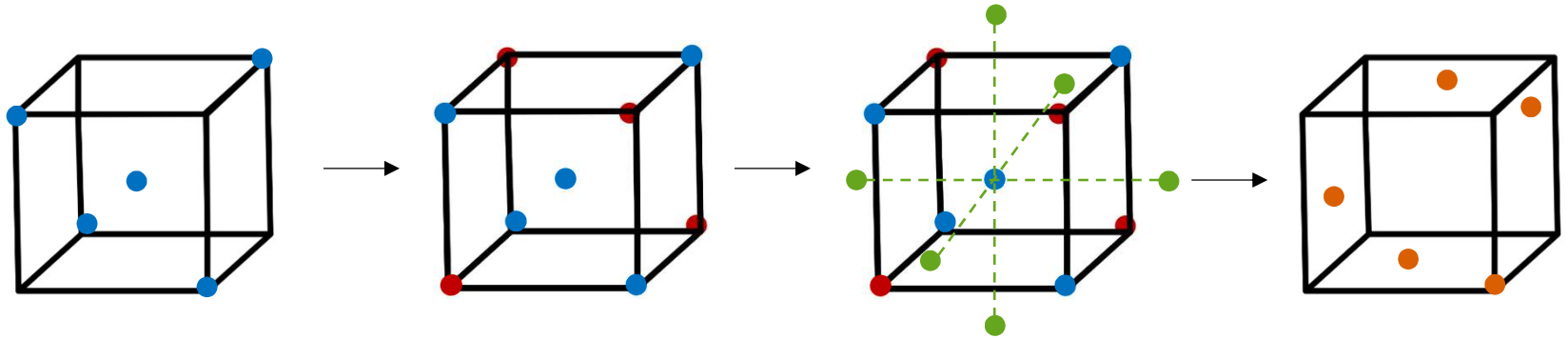


Need model to VV&A sim

Need sim for safety release

Sequential experimentation is (likely) the most efficient method for model induction





Screening Designs

Interaction Designs

Higher-Order Designs

Validation Designs


STOP
20319
CONSOLIDATED BUS TRANSIT

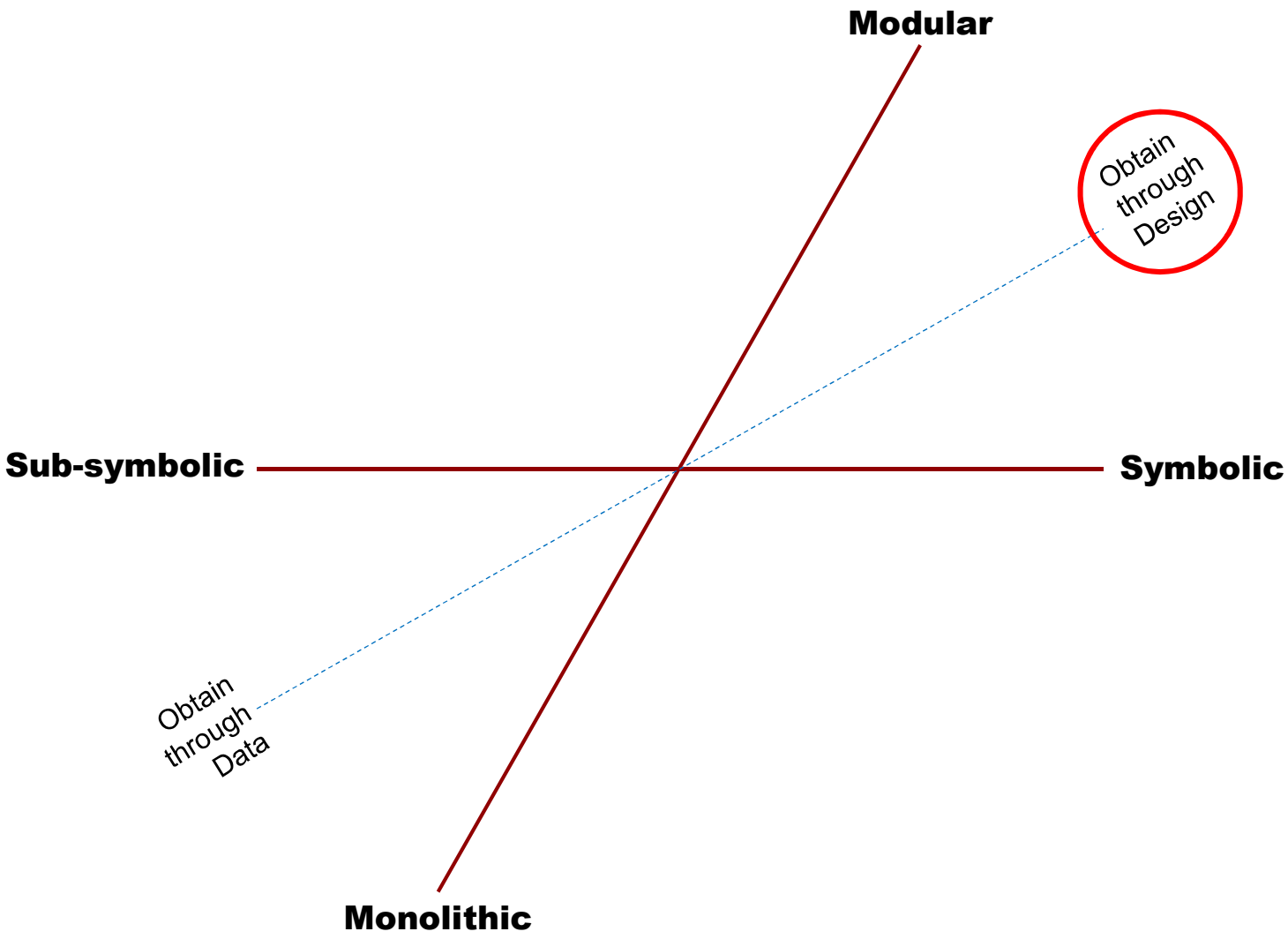

623G
CAT


POLICE

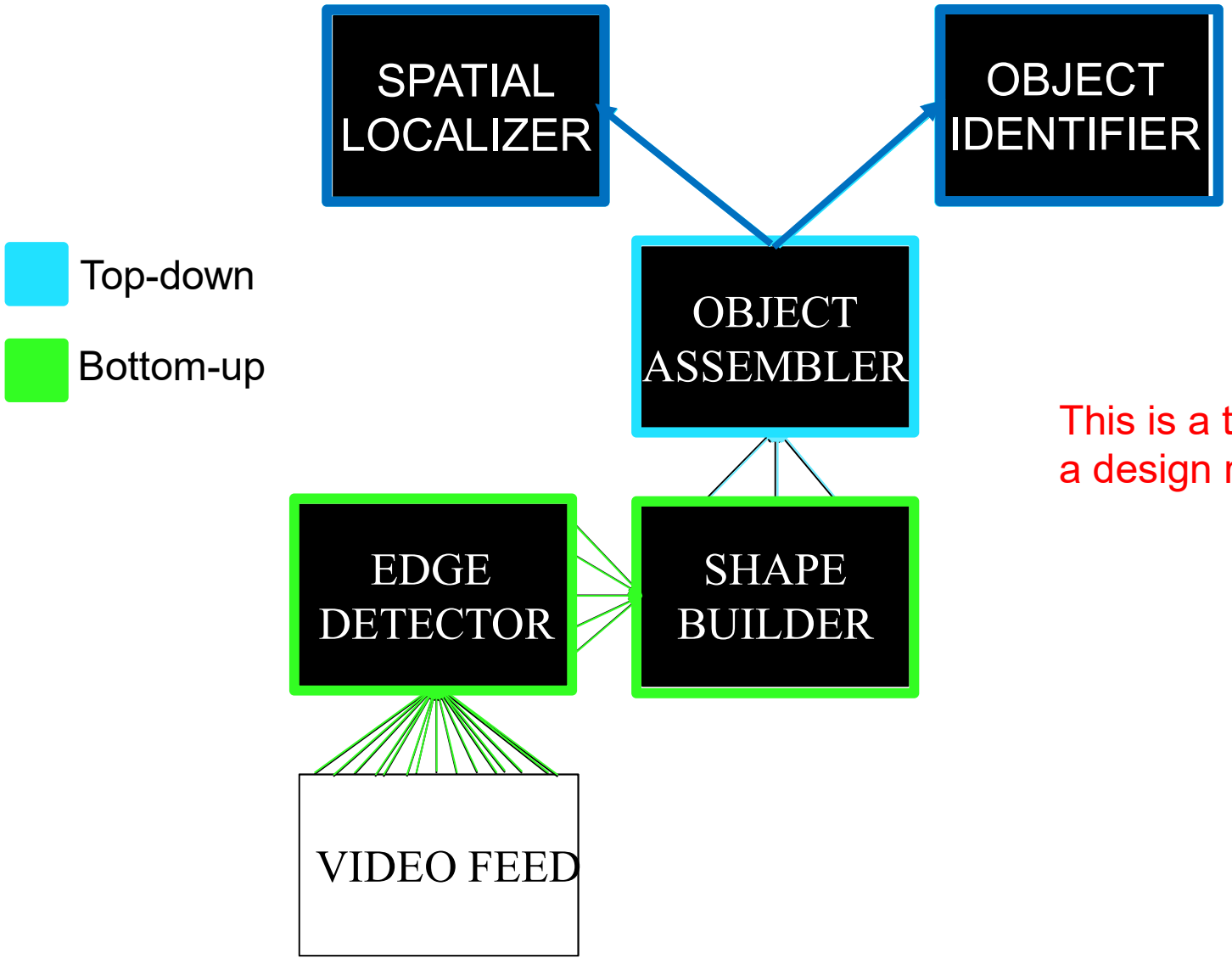
Sub-symbolic, monolithic systems will demand much greater quantities of data to obtain decision models.

These data may be expensive for both time and resources.

System design alters how to obtain system model

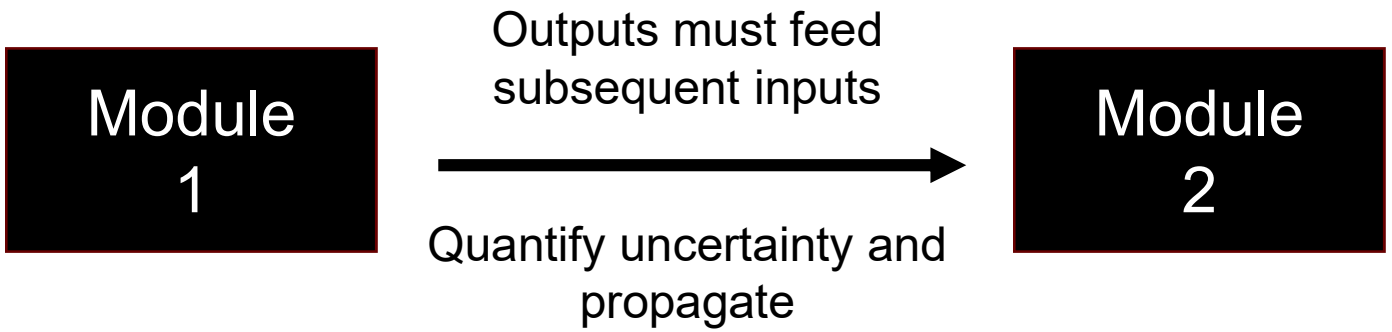


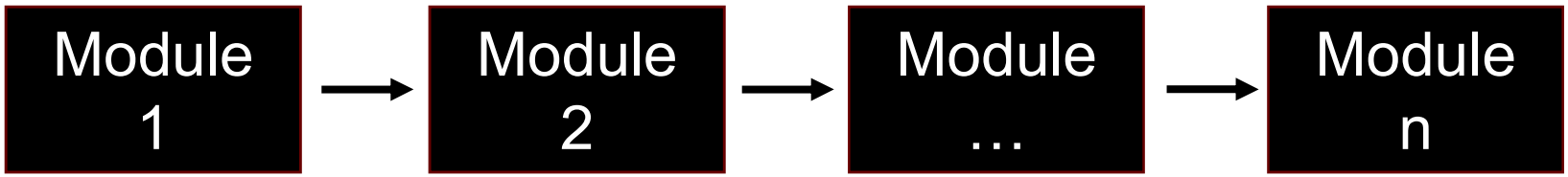
Modular architectures' decision models can be initially verified through cascading compositional verification



This is a toy example, not a design recommendation

Bayesian network models can quantify uncertainty in decision making across distributed modules





Propagating uncertainty across multiple modules provides uncertainty estimates in all or part of the decision model, supporting verification