DHS SCIENCE AND TECHNOLOGY

T&E of ML-based ATR

May 11, 2021

Christopher Smith, PhD
Director, Transportation Security Lab
Office of National Labs
Science and Technology Directorate

Homeland Security
Science and Technology

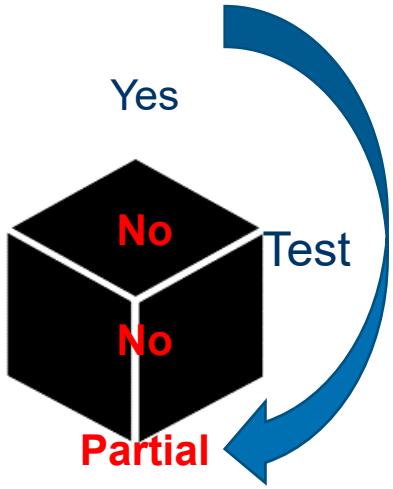
So what? Who Cares?

- The Space: Trustworthy ML Algorithms
- The Problem: How do we test Efficiently?
- The Solution: Some combination of
  - White Box Verification (e.g. DeepXplore)
  - Black Box Verification (e.g. RISE)
  - System-Level Testing (a.k.a Validation)

More White/Black Box Verification makes system-level testing easier and more predictable.

TRL Progression

TRL	Description	Testable? Engineered ATR	Testable? ML-Based ATR
1	Basic principles observed and reported	Yes	Yes
2	Technology concept and/or application formulated.	Yes	Yes
3	Analytical and experimental critical function and/or characteristic proof of concept.	Yes	Yes
4	Component and/or breadboard validation in laboratory environment.	Yes	No
5	Component and/or breadboard validation in relevant environment.	Yes	No
6	System/subsystem model or prototype demonstration in a relevant environment.	Yes	Partial
7	System prototype demonstration in an operational environment.	Yes	Yes
8	Actual system completed and qualified through test and demonstration.	Yes	Yes

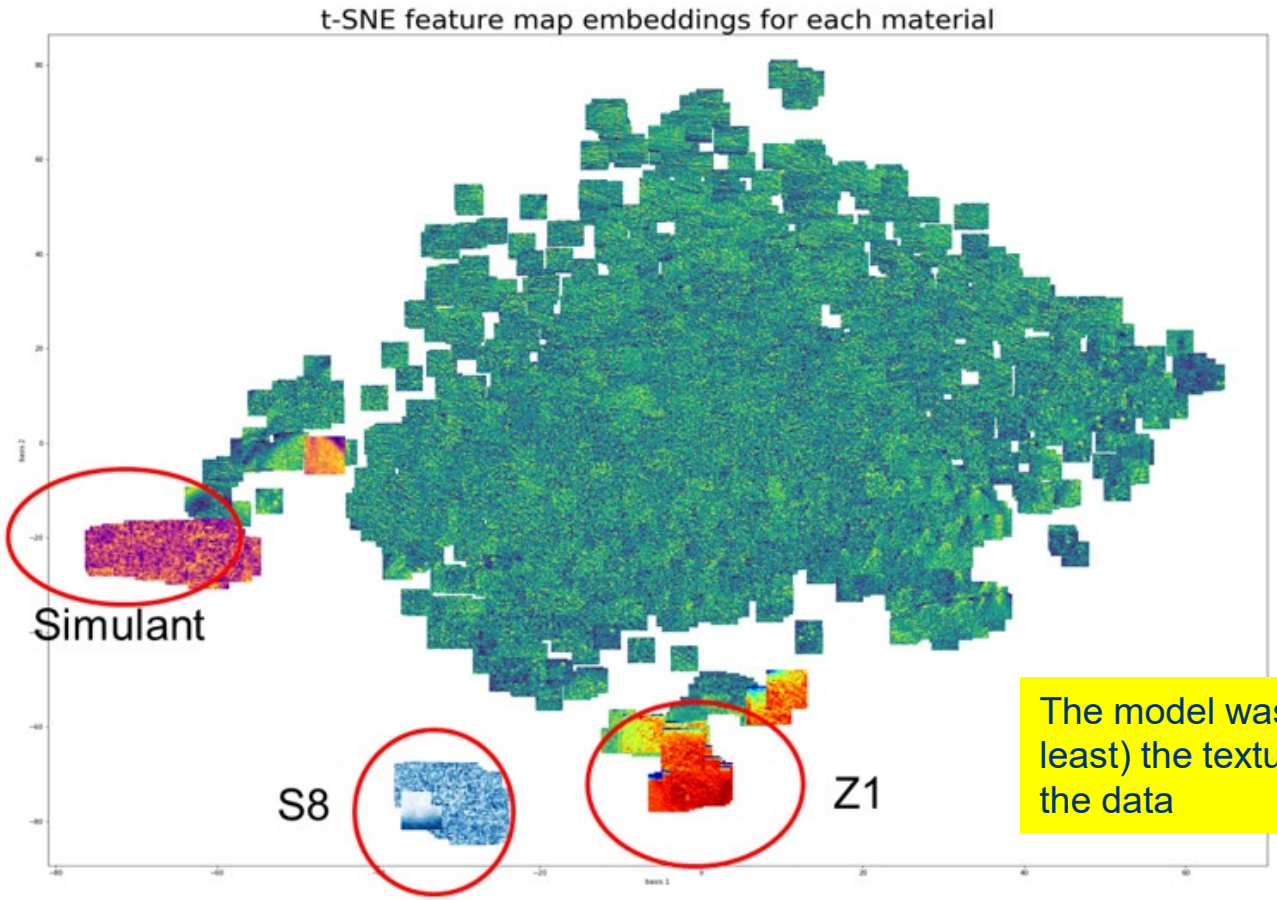


Homeland Security
Science and Technology

Mitigating the problem

ML algorithm vulnerability	Tools to prevent
NN correlates unrelated screening system image artifacts with presence or absence of explosive	TSA restrictions on NN design Rigorous DOE High Fidelity Synthetic Data NN-Validated Simulants Feature Identification Methods Other White-Box **verification** tools Other Black-Box **verification** tools
NN is over-trained (finds very specific threats but can't generalize)	Abundant Digitally-Modified Data TSA Requirements on architecture ("Capacity" restrictions & "Regularization")
NN Training data set is incomplete or insufficiently comprehensive	Comprehensive DOE Digitally-Modified Data
Terrorist manages to create threat that resembles nothing in the training set.	TSA requirements on NN behavior (equivalent to shield alarm)

Looking under the hood



The model was able to learn (at least) the texture features present in the data

How can we use this in practice?

- We can determine if an ATR algorithm is “seeing” texture.
  - If it does we need to include in our test design test articles with representative texture.
  - If not the we probably do not need hi-fidelity textured test articles.
- We can validate simulants (or create synthetic data) based on material properties (average attenuation) and object properties (texture)

Homeland Security

Science and Technology

DIVERSE PERSPECTIVES + SHARED GOALS = POWERFUL SOLUTIONS