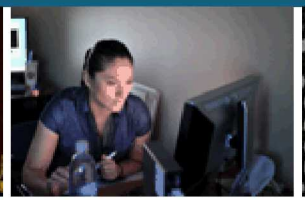
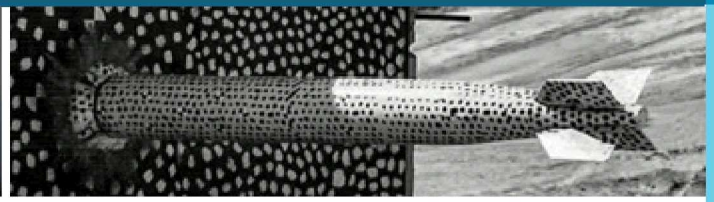
SAND2020-9237C

Deep Learning Models are Only As Good as Their Users; Or: *How Bad Models can Be Good, and Better Models can Be Worse*





Presented By

Zoe Gastelum

Research Team: Laura Matzen, Mallory Stites, Aaron Jones, Michael Trumbo

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

- Motivation
- A little about human performance.
- Visualization experiments: Bad DL is better than no DL, with one exception
- Accuracy experiments: As DL model performance goes up, human performance goes up, with exceptions
- Discussion of future work: Explainability, three ways.

Motivation & Introduction

Deep Learning and International Safeguards

Safeguards "running" towards DL in multiple aspects of verification

Deep learning models are getting very good, but what happens when they are wrong?

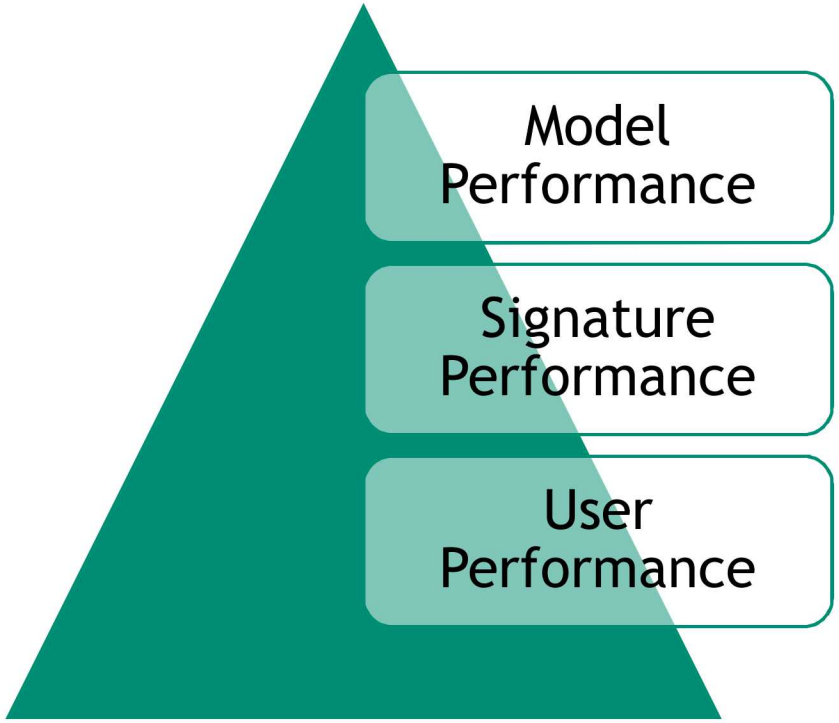
No deep learning algorithms were harmed in the course of this work.

Experimental data:

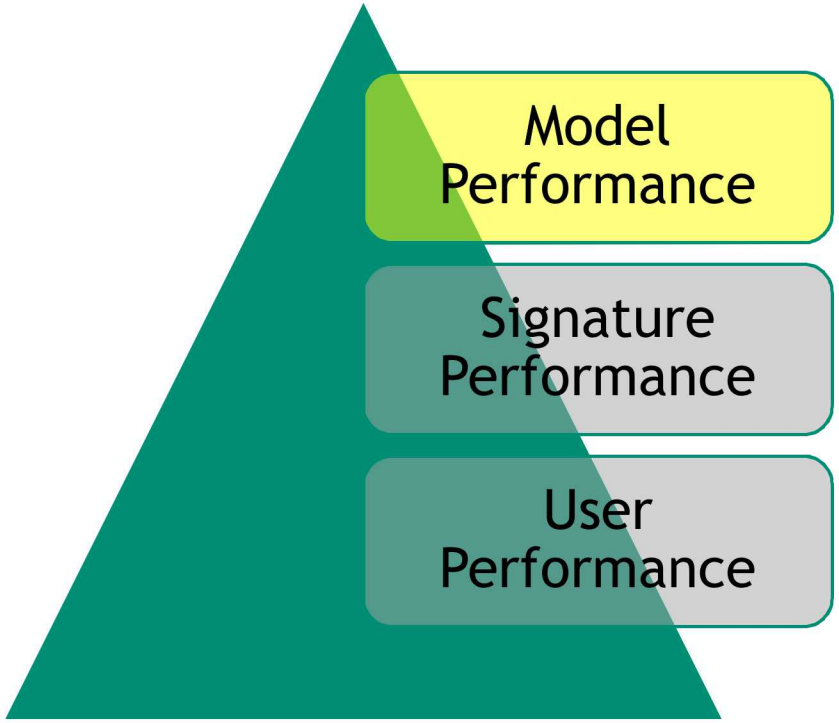
1) Visual search psychology dataset
2) Open source imagery dataset

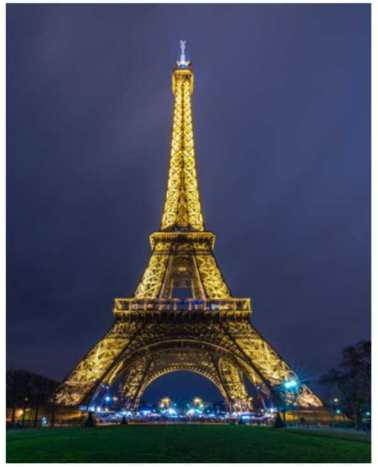
A little about humans…

Levels of performance



How good is my model at identifying the defined signature?



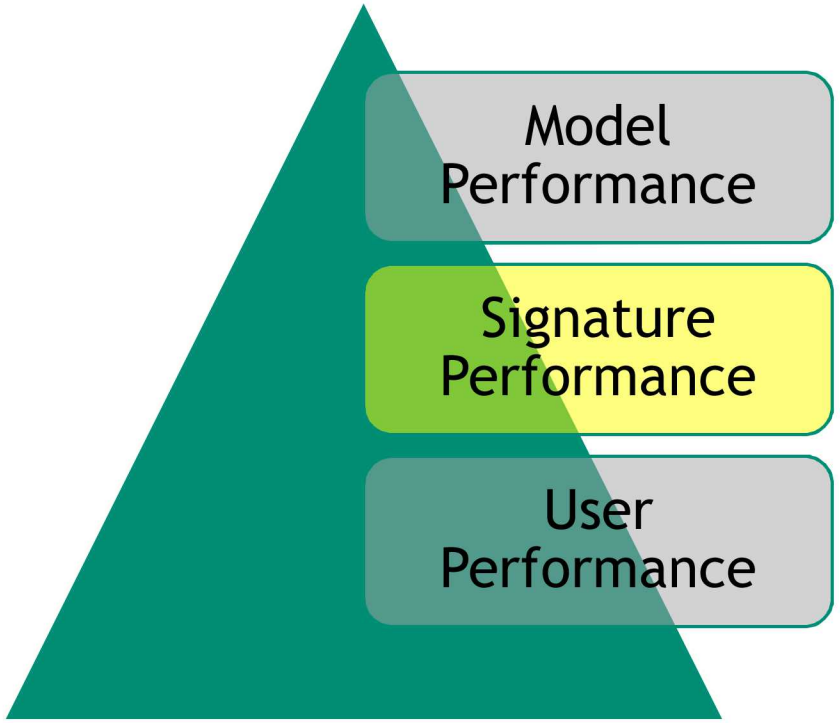


Not a cooling tower

Cooling tower no steam

Cooling tower with steam

How good is my signature at identifying the event?

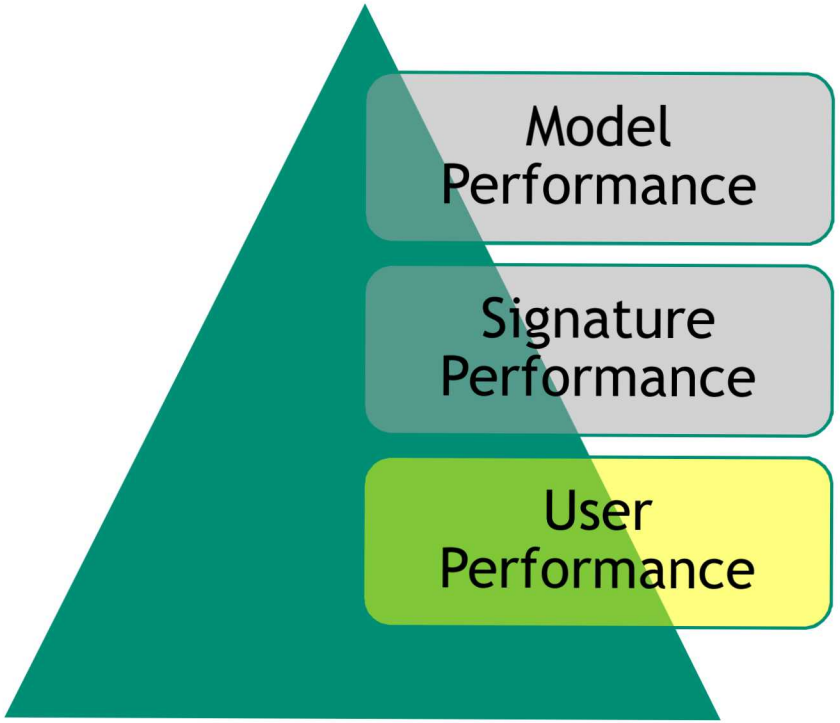


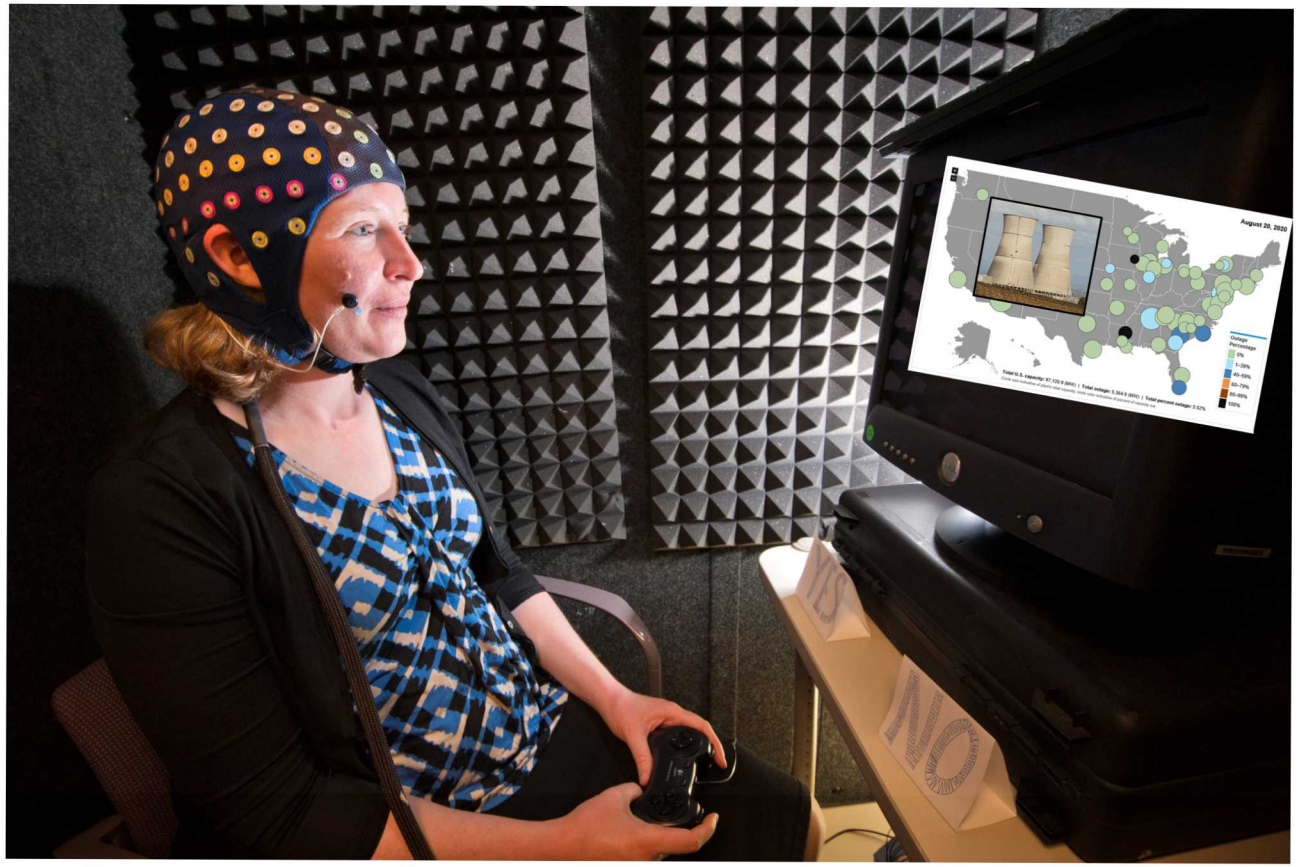
August 20, 2020

0%
1–39%
40–59%
60–79%
80–99%
100%

Total U.S. capacity: 97,[illegible] MW Total outage: 5,364 MW Total percent outage: 5.52%

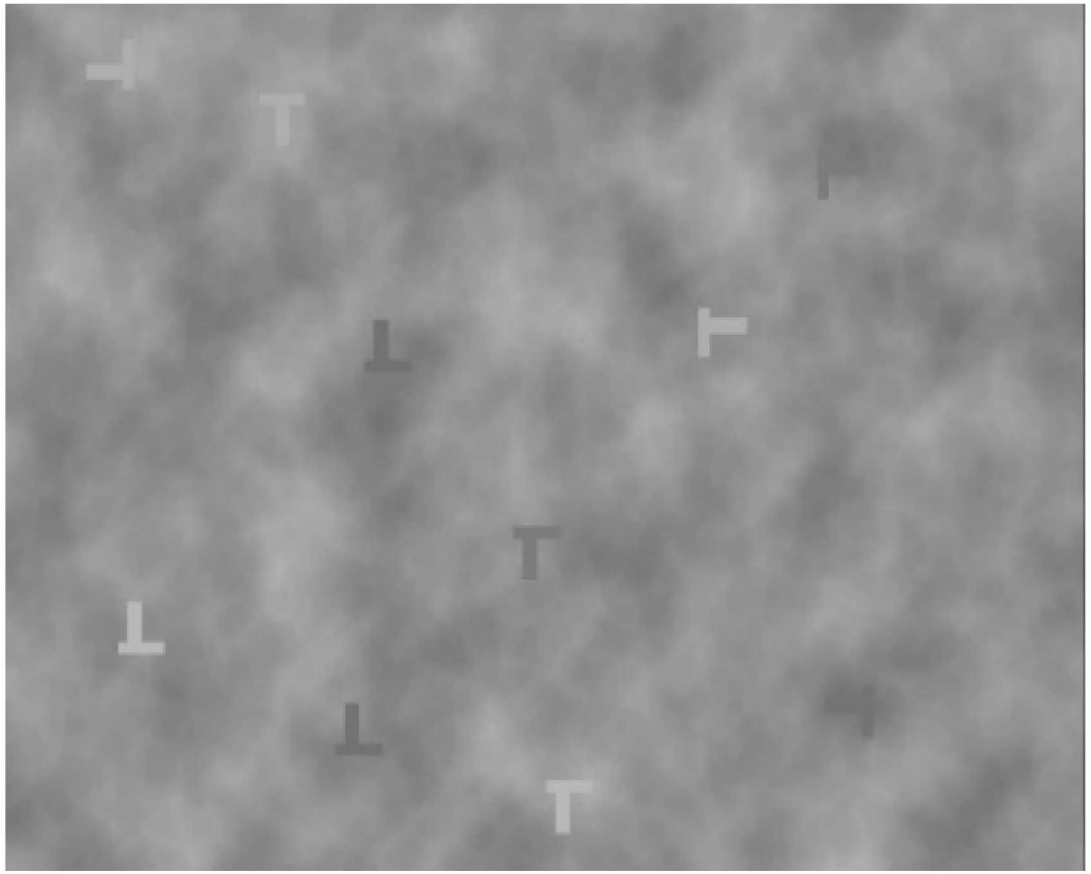
How good is my human at interpreting the model's identification – or misidentification - of the signature?

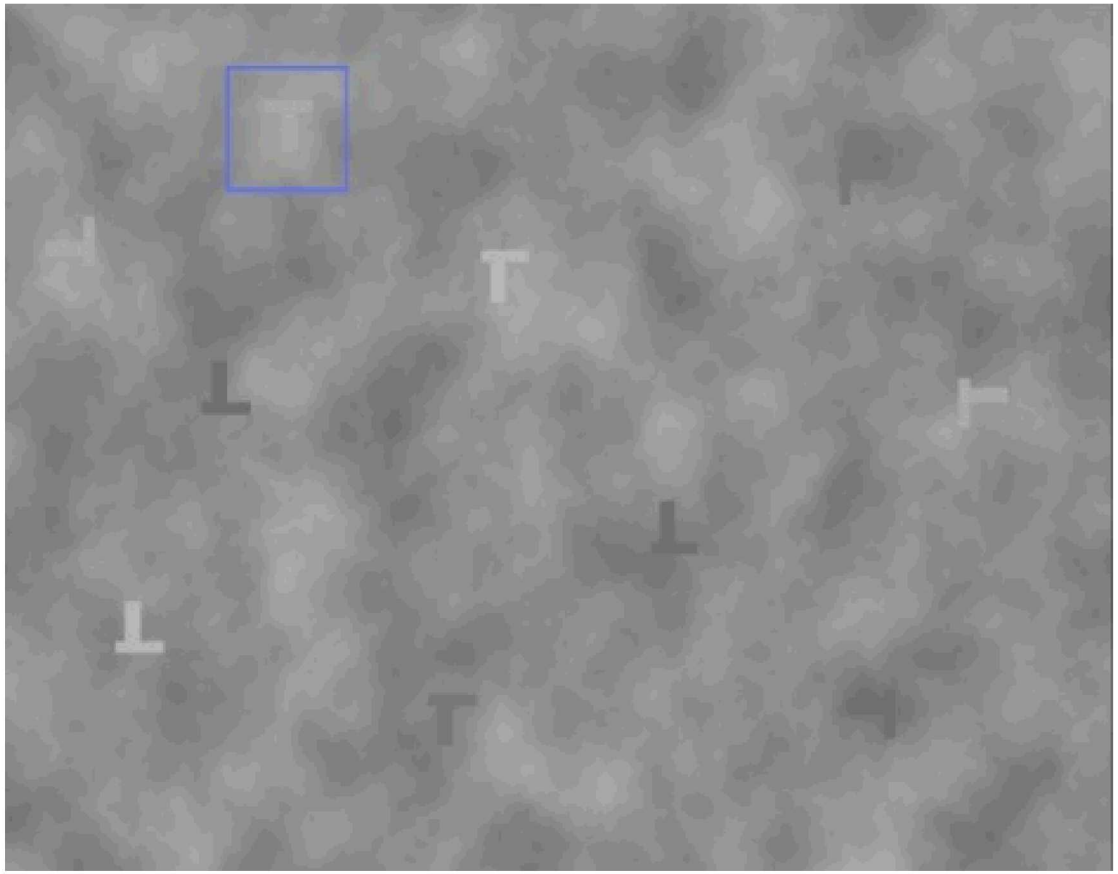




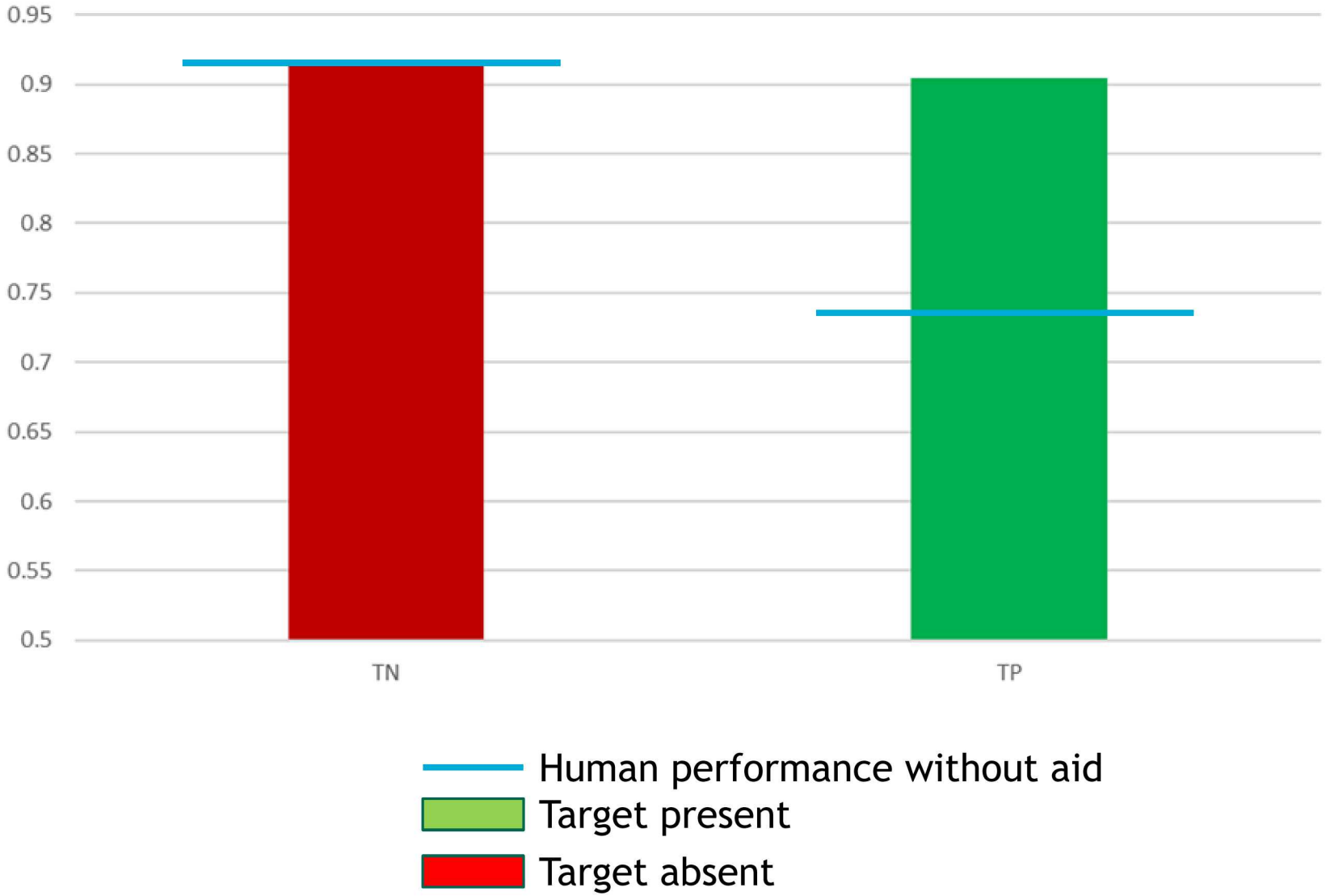
1) Even "bad" DL is better than "no" DL (most of the time)

Is there a “Perfect T” present?





Participant Accuracy for Correct Model Indicators

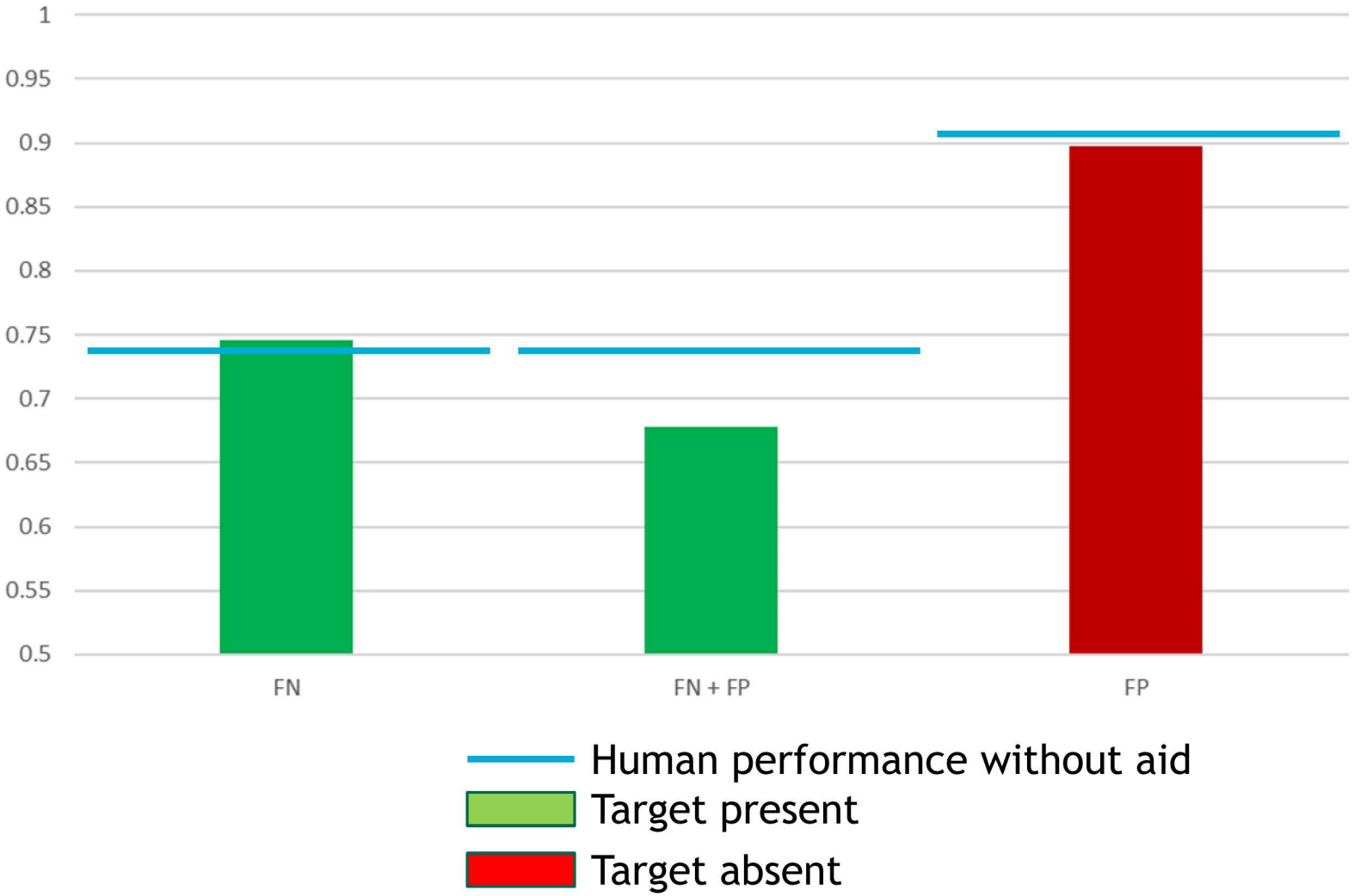


- TP: Accuracy increases from 74% to over 90%

- TN: Accuracy remains about the same at 91%.

- TN presentation is an absence of indicator for all but text conditions, so similar performance on TN and no aid is expected

Human accuracy when the model is wrong



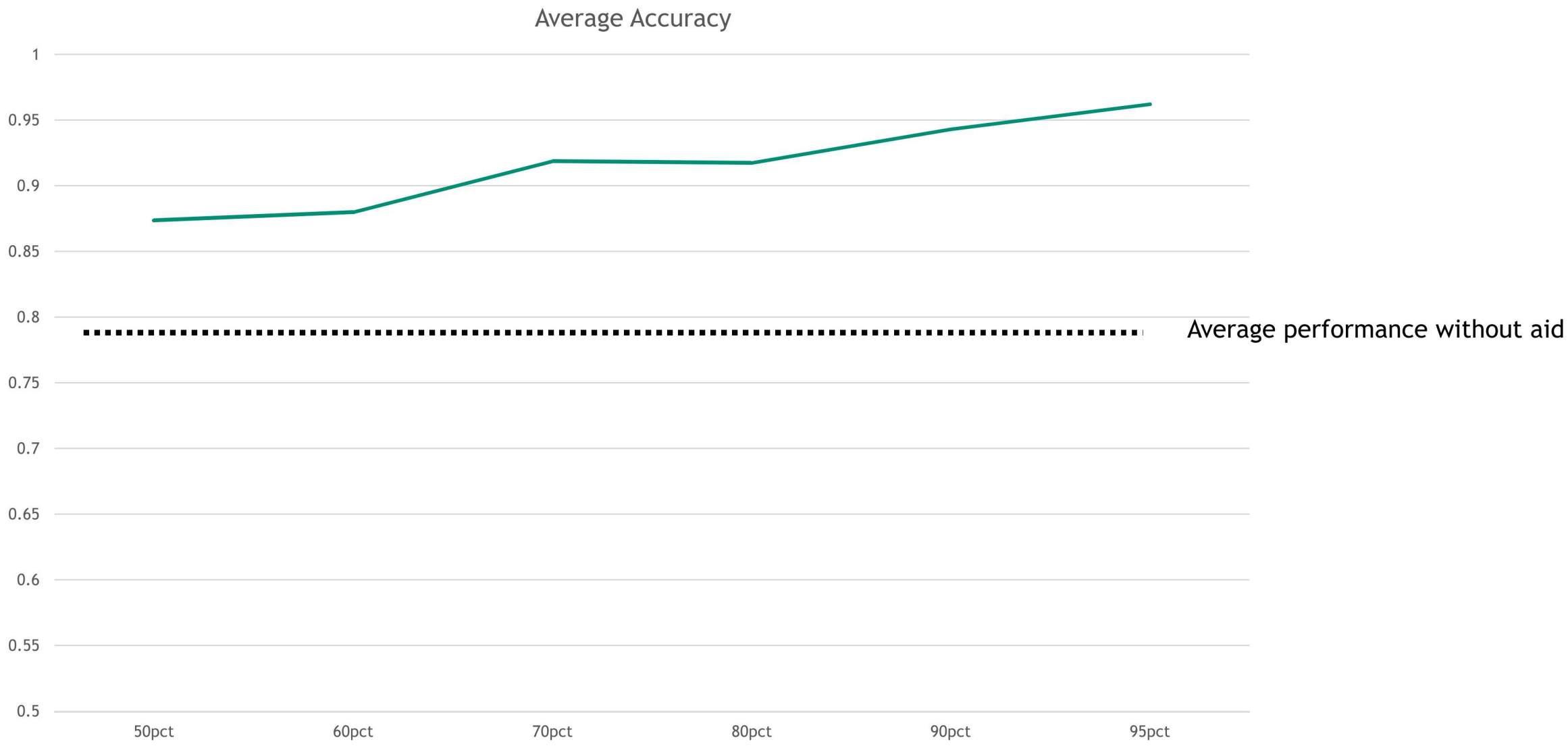
- FN: Accuracy remains about the same, up slightly from 74% no aid to 75%.

- FP: Results stay about the same as no aid to identify the absence of a target

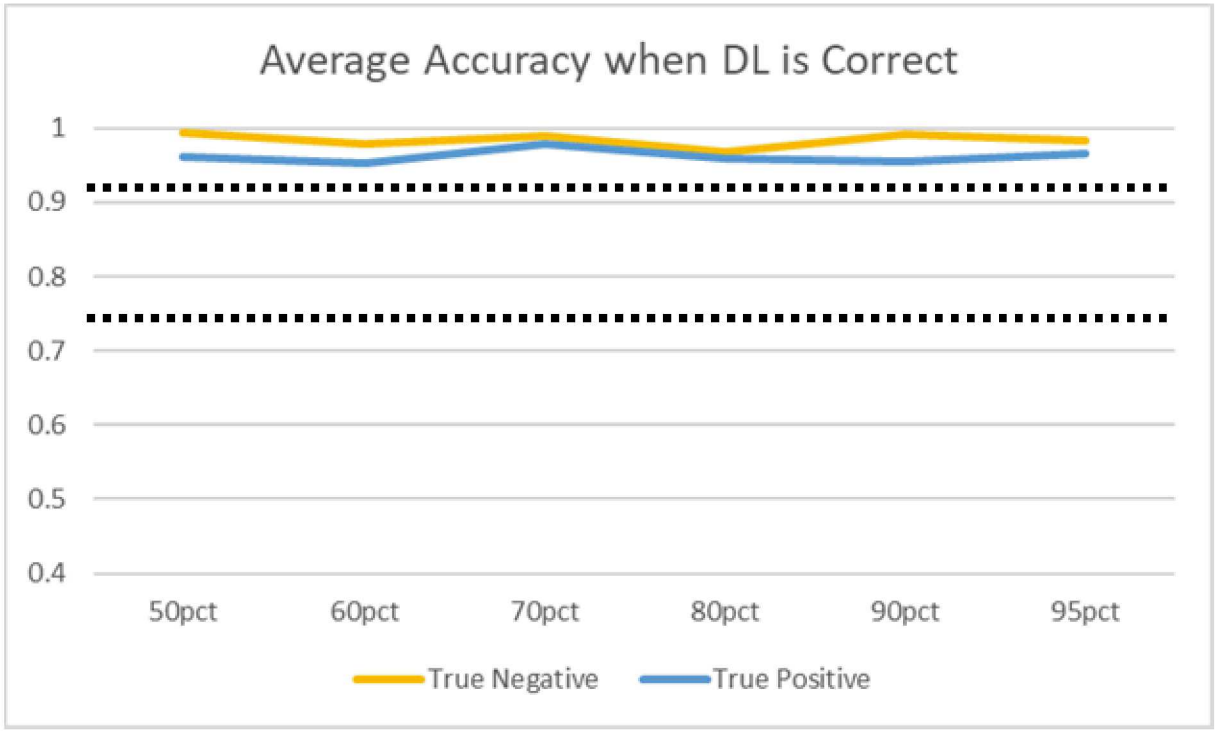
- FN + FP: Performance decreases from the no aid condition, from 74% to about 68%.

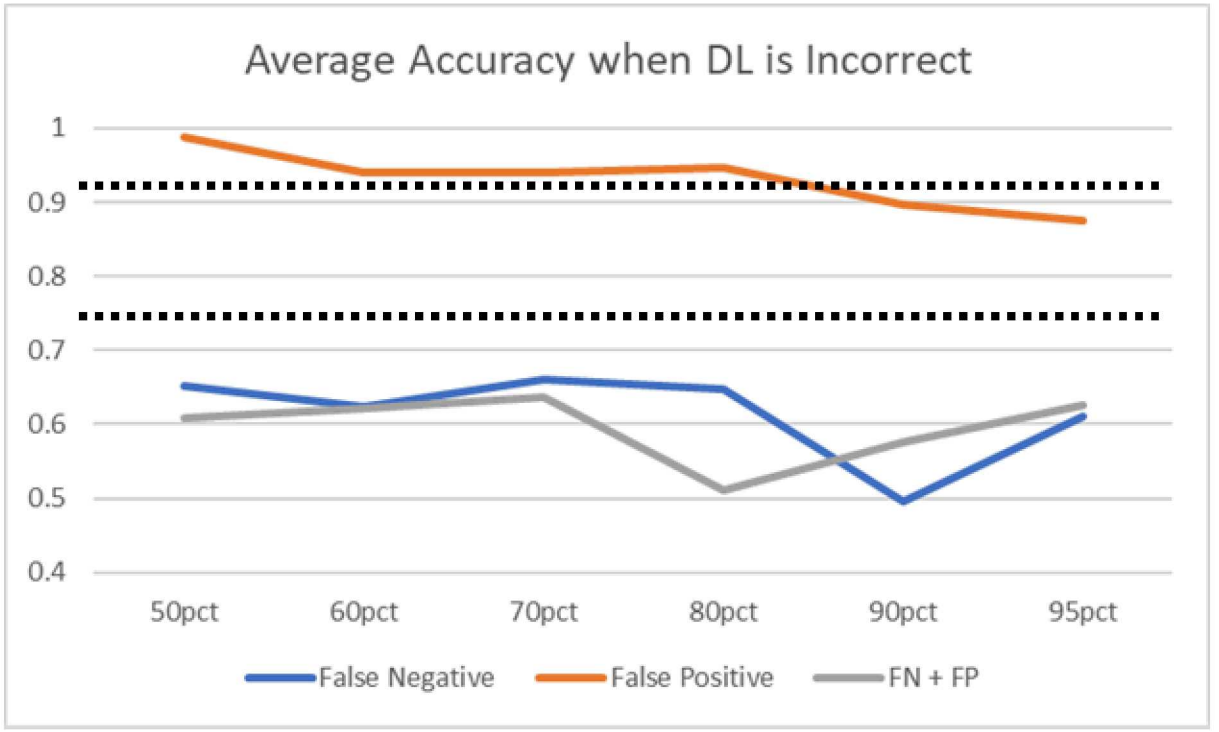
2) As DL model performance goes up, so does human performance (most of the time)

Participant Accuracy as a Function of DL Accuracy



Participant Accuracy as a Function of DL Accuracy

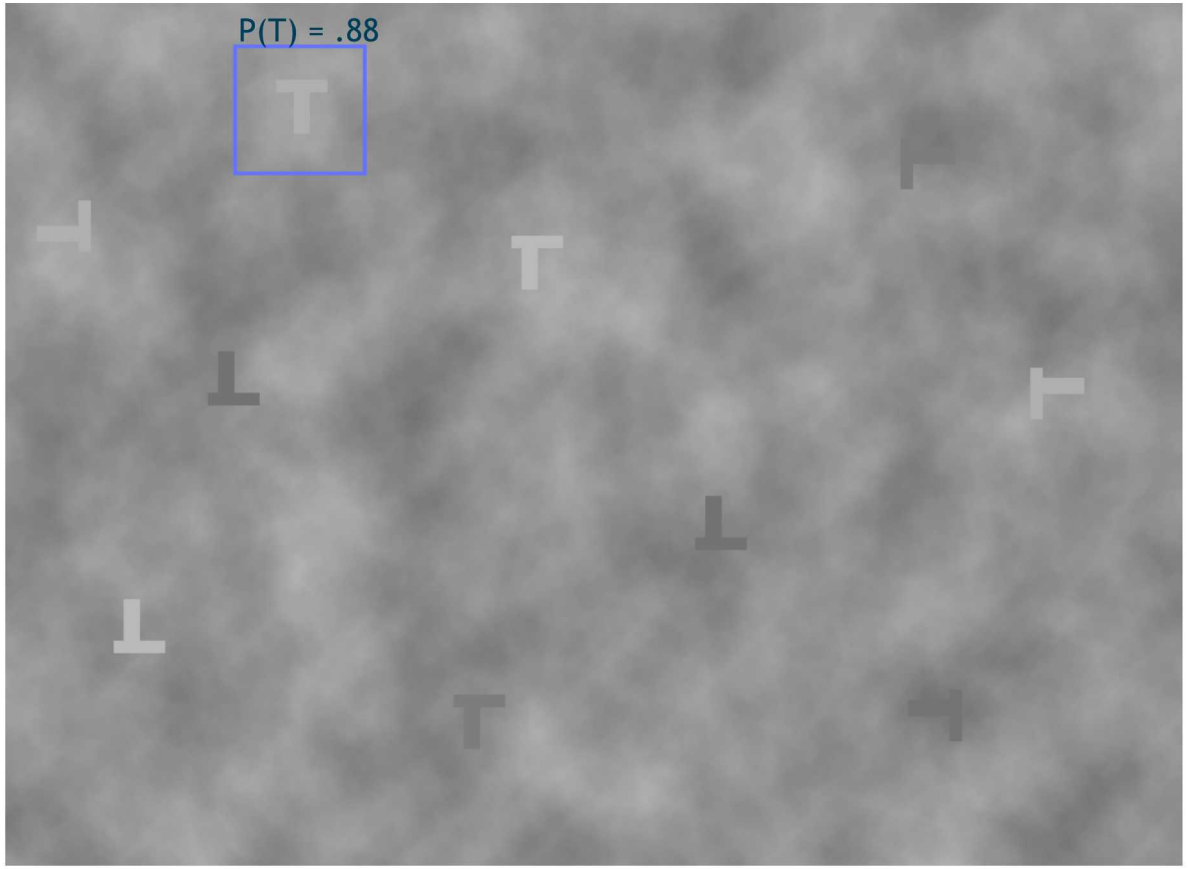


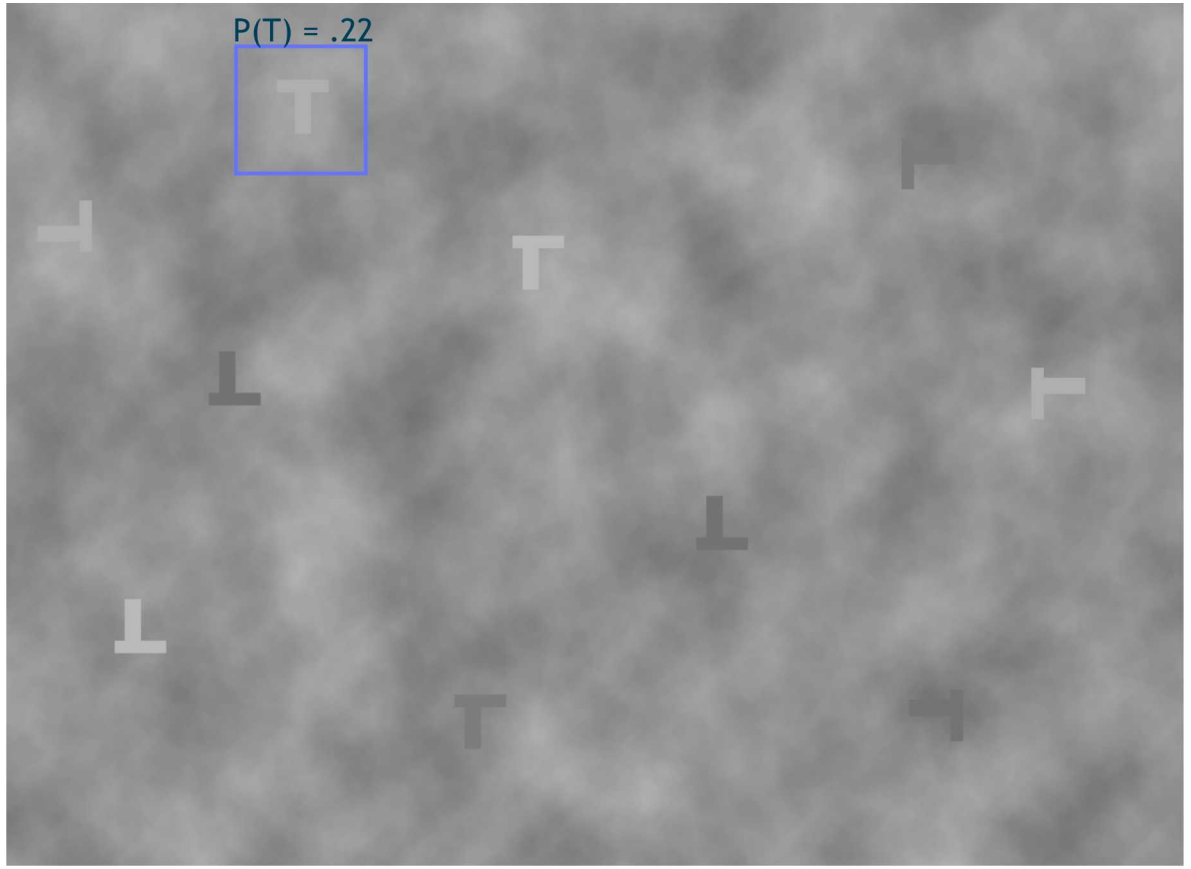


····· (at 74%) Unaided detection of “target present”
····· (at 91%) Unaided detection of “target absent”

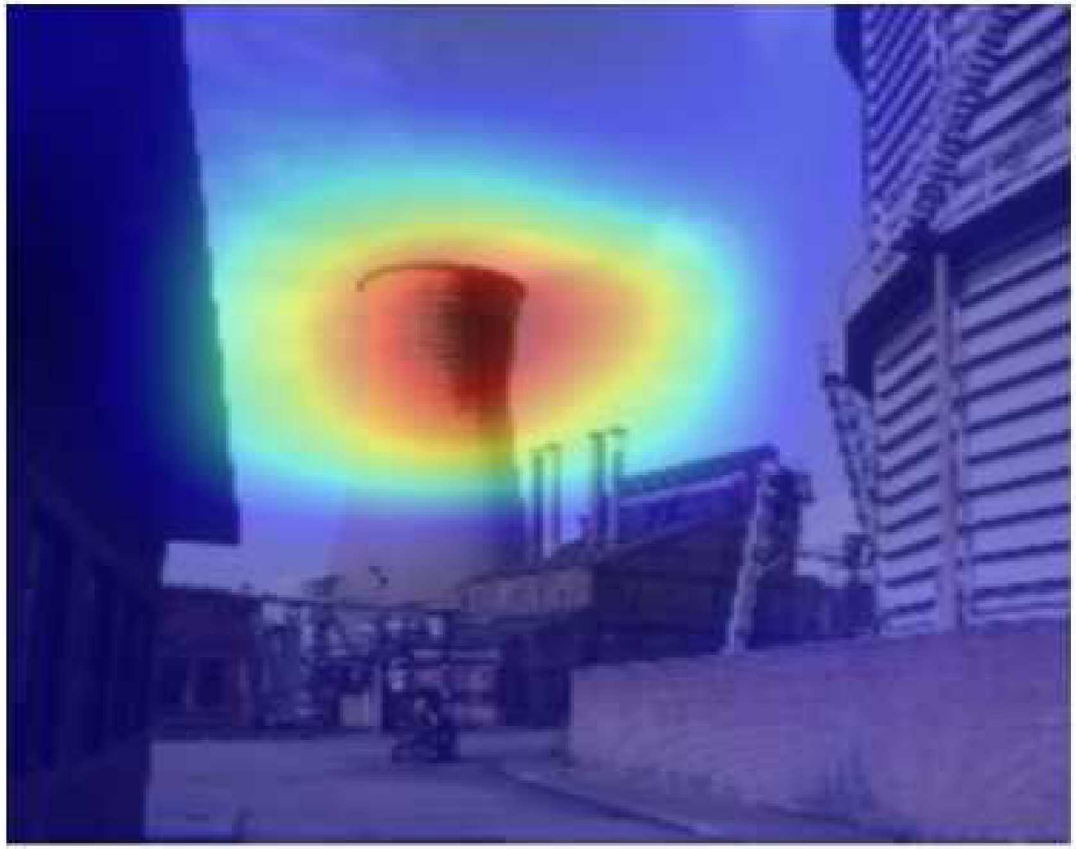
3) Explainability, three ways

Confidence





Salience



$p\ cooling\ tower = 1.00000$

$p\ cooling\ tower = 0.96455$

Transparency

Which model would you trust?

TensorFlow

Google Brain-developed libraries
Image: TensorFlow.org

YOLO v3
Image: Redmon, https://pjreddie.com

Zoe Gastelum

zgastel@sandia.gov

(505) 401-6959

This work was funded by Sandia National Laboratories Laboratory-Directed Research & Development program's Computing and Information Sciences Investment Area, under LDRD project #218306.