

SERC DOCTORAL STUDENT FORUM 2023 | NOVEMBER 14, 2023

Enabling Understanding of Model Behavior by non-AI Expert Decision Makers Through Novel Visualization

Chris Krueger, Justine Manning, Robert Pless, PhD, and Zoe Szajnfarber, PhD

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC



SYSTEMS
ENGINEERING
RESEARCH CENTER

This work is conducted with support from NSF Grant No. 225677 and AIRC task WRT 1071

Motivation

“Do we need to teach decision makers and senior leaders the math behind AI for them to be able to be able to decide which model will work in their application?”

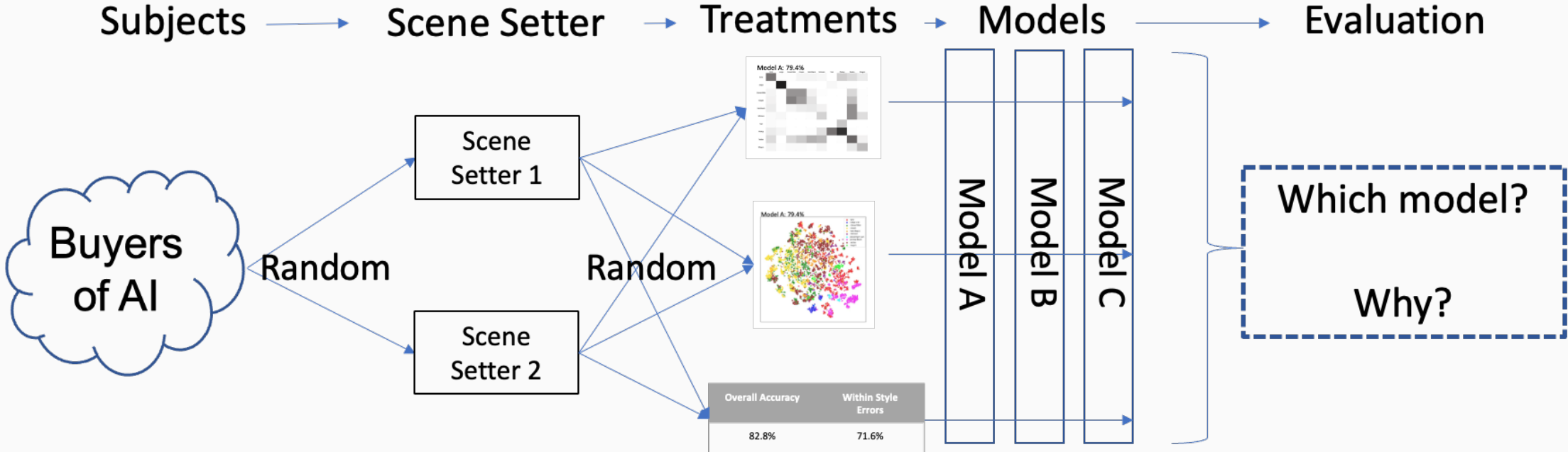


How wrong would these classifications be?

Research Goals

- Discover how to better communicate AI model behavior through visualizations.
- Identify the type and quantity of training for the workforce to better understand the behavior.

Experiment



Subjects

- Sample needs to be representative of acquisitions officers who have advanced degrees in fields other than computer science and AI.
- Pilot: Colleagues
Officers with advanced STEM degrees
- Experiment: Army Acquisition professionals
Civilians with advanced education in subjects such as project management or business and rarely have AI expertise

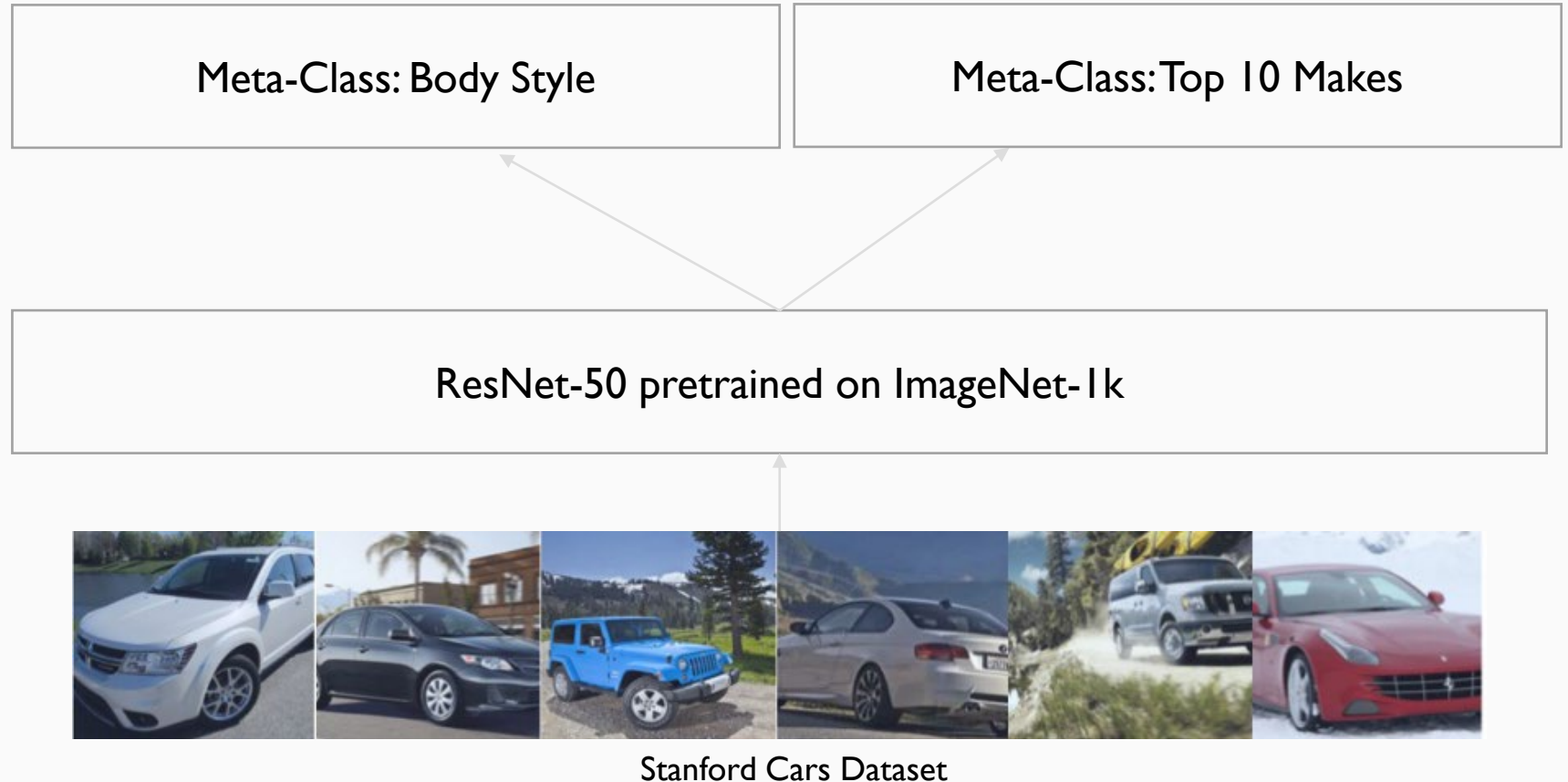
Alternative Models

2 x Alternative Models:

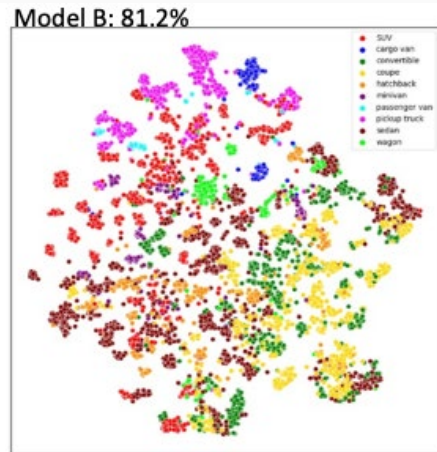
- Pretrained on a meta-class
- Similar accuracy
- Pretraining facilitated deliberate and specific errors induced

1 x Base Model:
Trained for 20 epochs

Total: 3 similarly accurate models each with deliberately different behavior



Visualization Development



t-SNE plot: a dimensionality reduction method that takes the high-dimensional image embeddings and reduces them to two dimensions while preserving the local structure of the embedding neighborhoods

Tighter, more distinct clusters means the model found the most similarities in the dimensions within the specified classes.



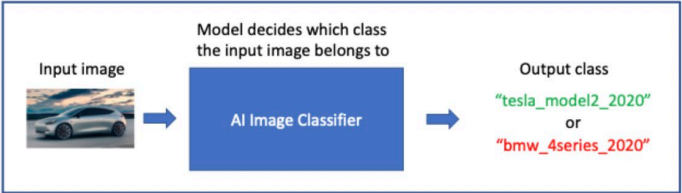
Mistake structure matrix: a modification of the traditional confusion matrix where only the errors within meta-classes are displayed.

The higher the density of error on the diagonal the more errors the model made that at least identified the correct meta-class.

Experiment Training Elements

What is an image classifier?

An image classifier is trained on 1000s of images with the intent to be able to correctly predict the label on a new image.



The accuracy of a model is one way to communicate how often the model correctly labeled an image. It is also useful to know more about the types of mistakes that are made. These population results can be summarized in several different ways. We will be using one of them, mistake structure matrices, in the rest of the experiment.

What is a mistake structure matrix?
(Please watch this short animation.)

This is a mistake structure matrix.

It shows how often a model *confuses* what the class of the input image is for another.

	A	B	C	D
A	■	■	■	■
B	■	■	■	■
C	■	■	■	■
D	■	■	■	■

Your task:

Imagine you work for a City that has introduced variable congestion pricing where the fee is based on the body type (e.g., trucks pay more than sports utility vehicles which pay more than sedans). The Mayor is excited about AI and wants to implement a system that uses an image classifier to automatically charge each vehicle the right rate. Your job is to buy the system that best suits your City's needs. Overall accuracy is important because you want to collect data on which cars people are driving to understand fuel consumption and emissions. However, in this application, distinguishing body types is much more important than distinguishing make for revenue purposes.

Survey

Which model would you choose?

Model Odin

Model Cyclops

Model Argos

Please rank your confidence in your selection of the model.

	1- I am not confident this model is the best.	2	3	4	5- I am highly confident in this model.
On a scale of 1 to 5 (least confident to most) please rank your confidence.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Why did you choose that model?

Rank the models from best to worst suited for your needs. (1 is best.)

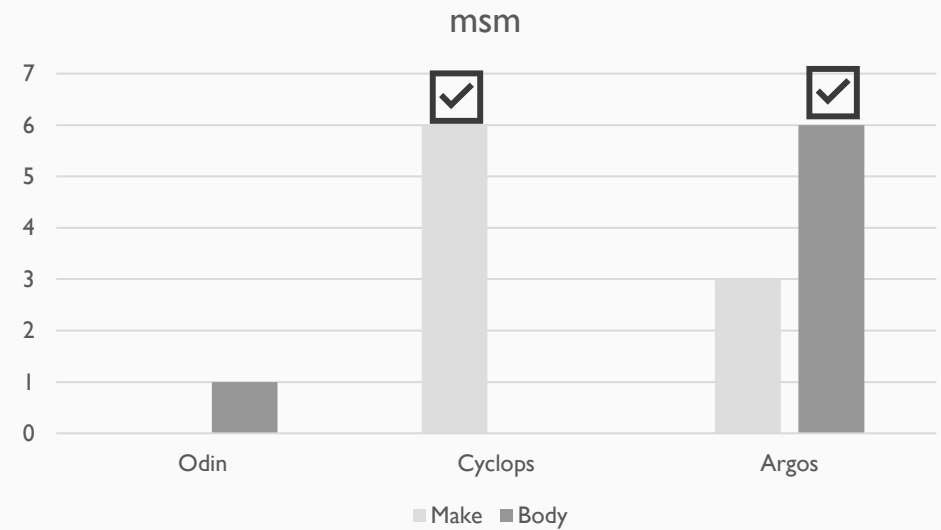
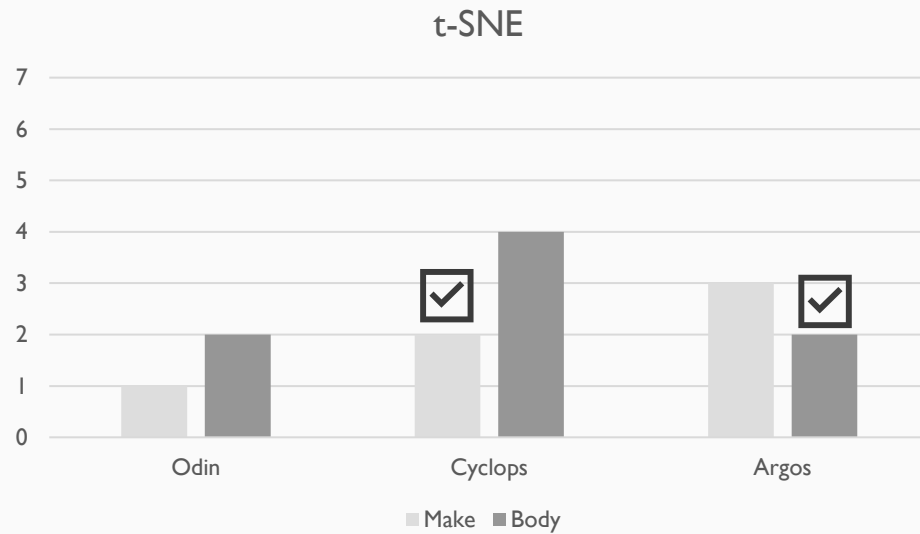
	1	2	3
Model Odin	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Model Cyclops	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Model Argos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Mark the model that is least likely to confuse:

	Odin	Cyclops	Argos
Cargo van and SUV	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Convertible and Coupe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Audi and BMW	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Pilot Analysis

Which model best suits your needs?



Preliminary Results and Future Work

- Need for “explainability” methods that communicate information about behavior (not just positive performance)
 - Pilot data strongly suggests that the type of visualization impacts what decision-makers understand about the model behavior, particularly when it works/fails.
 - Even with a clear organizational preference, most (technically literate) subjects chose the wrong model when presented with standard model metrics and visualizations. Our MSM performs better in this pilot.
 - Remains to be seen if results hold in full experiment.
- Opportunity for “error-aware” explainability
 - Current focus is on when the model is right. Need to also show when (and how) models makes mistakes.
 - We tested a simple modification of a standard approach and saw improvements. Room to explore others including hierarchical scoring and other visualization approaches.
- Opportunity for “expertise-aware” explainability
 - It is known that expertise affects which types of explanations are more compelling; likely holds for AI too.
 - Balance meeting acquirers where “they” are and exploring what training might be needed to improve assessment.

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC