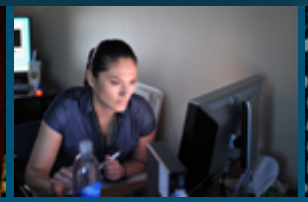
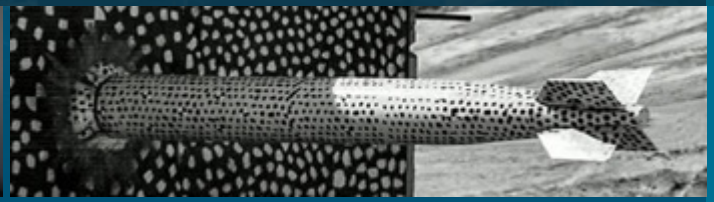
SAND2021-8880C

Sage Advice? The Impacts of Explanations for Machine Learning Models on Human Decision-Making in Spam Detection





Presented by

Mallory Stites, Megan Nyre-Yu, Blake Moss, Charles Smutz, & Michael R Smith

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.

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.

Introduction

Goal: increase efficiency and effectiveness of cybersecurity analysts by providing information about which features of the email an ML model used to classify it as malicious or benign

Experimental Question: How can machine learning explanations be displayed to users to improve efficiency and effectiveness of decision-making?

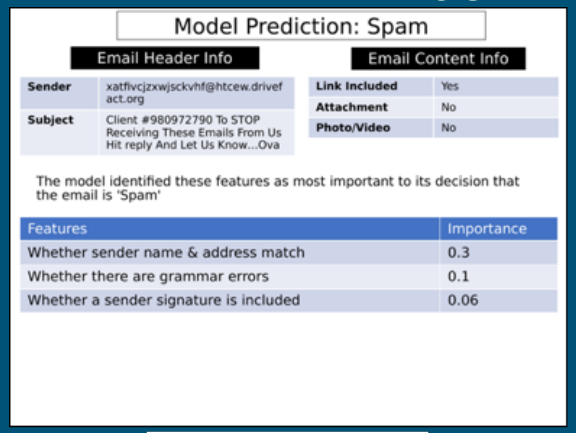
We focused on the following aspects of the ML model:

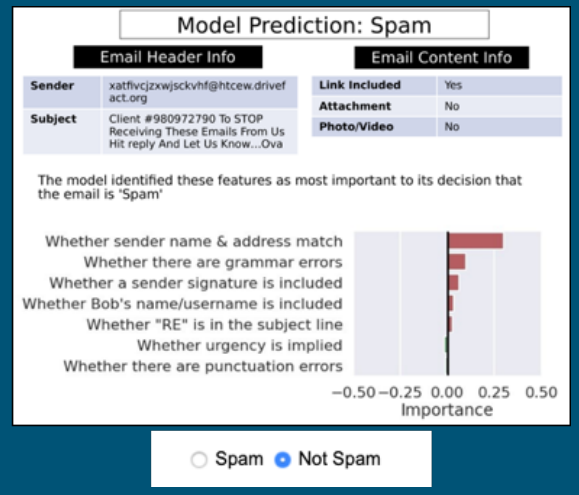
- Number of features displayed from model
- Visualization of feature importance
- Overall model accuracy

Prediction: More understandable model outputs would improve user performance relative to baseline

Current Experiment: Task

Simulated Spam Detection Task





Trust in Automation Survey (Hoff & Bashir, 2015)

1	I am confident in the [tool]. I feel that it works well.
2	The outputs of the [tool] are very predictable.
3	The tool is very reliable. I can count on it to be correct all the time.
4	I feel safe that when I rely on the [tool] I will get the right answers.
5	The [tool] is efficient in that it works very quickly.
6	I am wary of the [tool].** [Reverse scored.]
7	The [tool] can perform the task better than a novice human user.
8	I like using the system for decision making.

Stimuli

Gathered initial set of 102 emails (50 spam, 52 benign)

Assigned each email a binary yes/no value on feature set

Trained a Random Forest ensemble classifier

Utilized SHAP library to generate importance values for each feature in model's classification

Model achieved 97% classification accuracy
- 48/50 Spam, 51/52 Benign

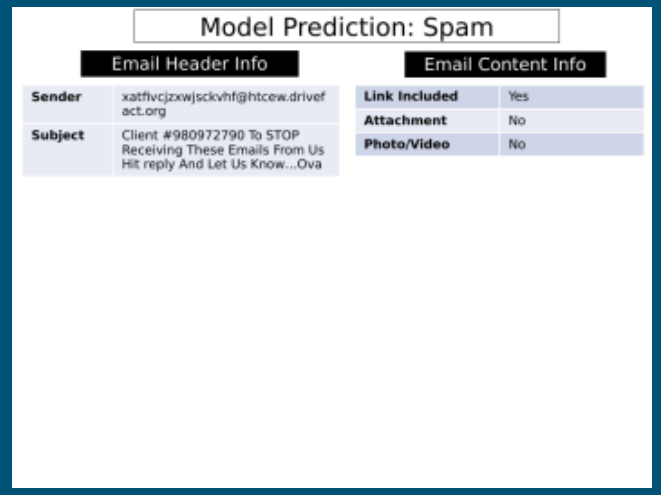
Selected 40 Spam, 40 Benign that were accurately classified

Feature Name	Percentage of Stimuli with a Positive Value on Feature	
	Spam	Not Spam
Email contains Link*	0.88	0.28
Email contains Attachment*	0.05	0.18
Email contains Photo*	0.48	0.30
Email Sender + Address match	0.20	1.00
Email contains Spelling errors	0.25	0.08
Email contains Grammatical errors	0.53	0.13
Email contains Punctuation errors	0.38	0.08
Email recipient name is mentioned	0.23	0.50
Email contains Symbols (e.g., Greek letters)	0.10	0.05
Email contains high count of "!"	0.08	0.08
Email contains high count of # sign	0.03	0.03
Email contains signature	0.13	0.30
Email urgency is implied	0.60	0.05

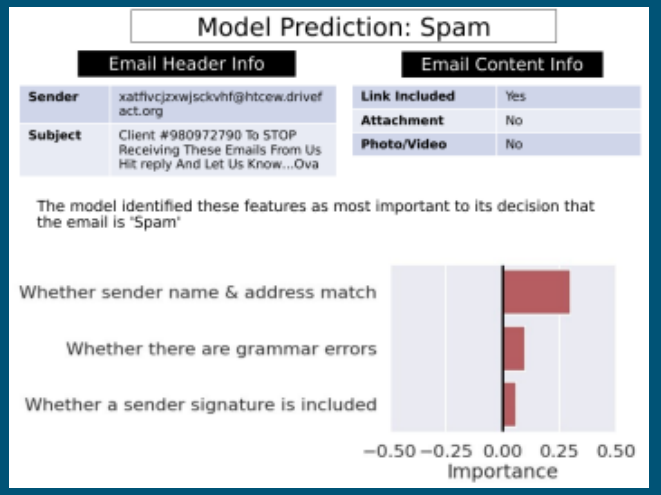
Method

Factors Manipulated: Within Subjects

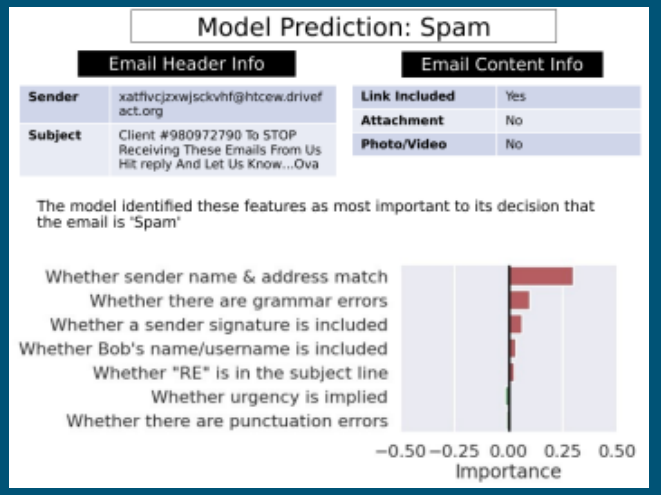
Number of Features from ML model shown on each trial



0 features (Baseline)



3 features

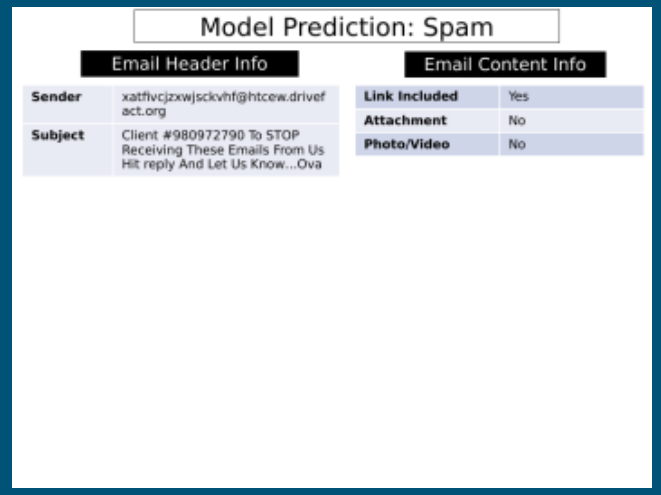


7 features

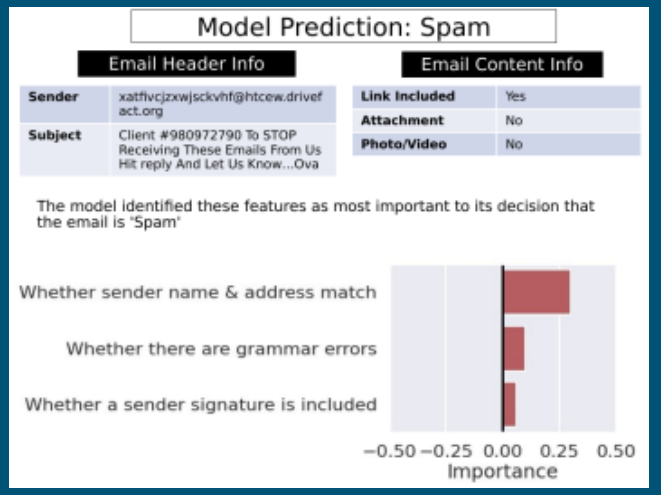
Method

Factors Manipulated: Within Subjects

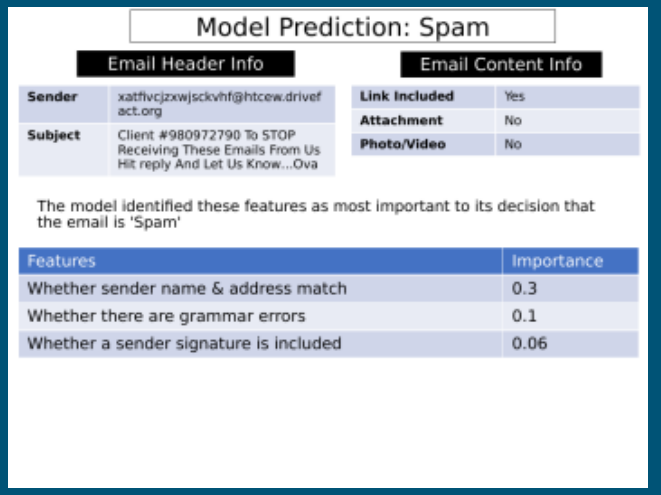
Feature Importance Visualization



No visualization (Baseline)



Graph

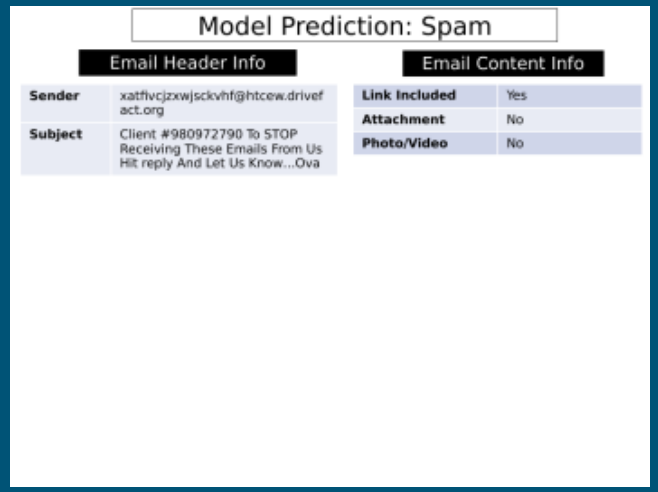


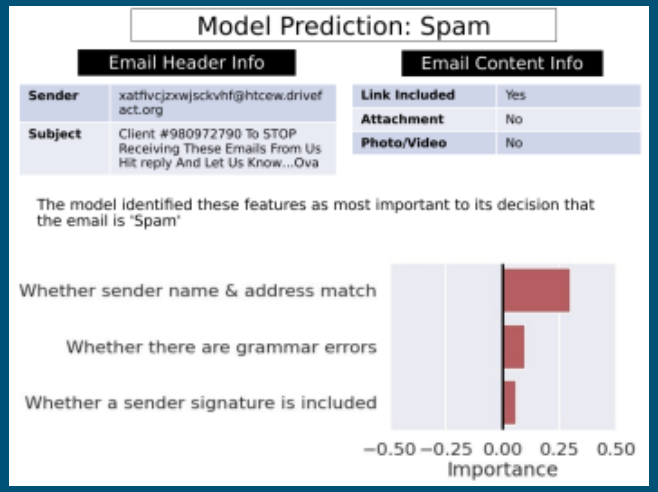
Table

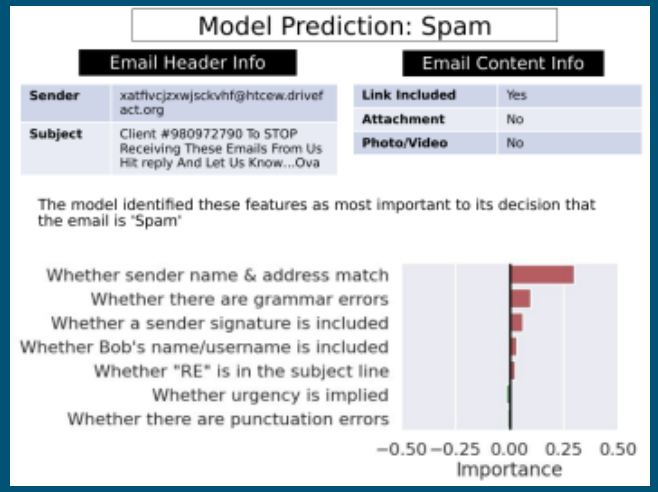
Method

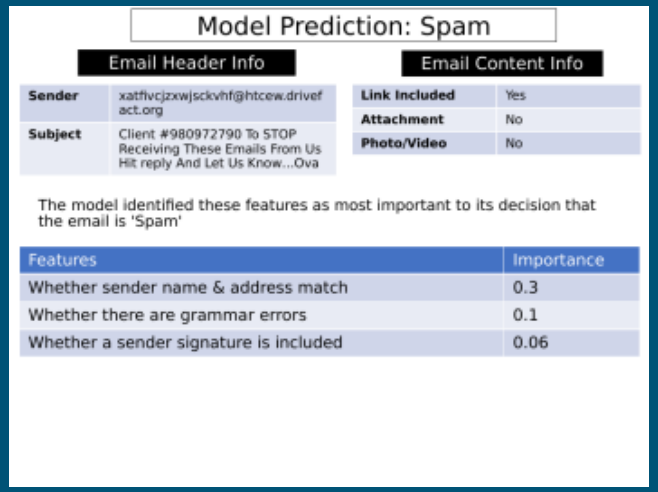
Factors Manipulated: Within Subjects

Each stimulus could appear in All 5 visualization conditions

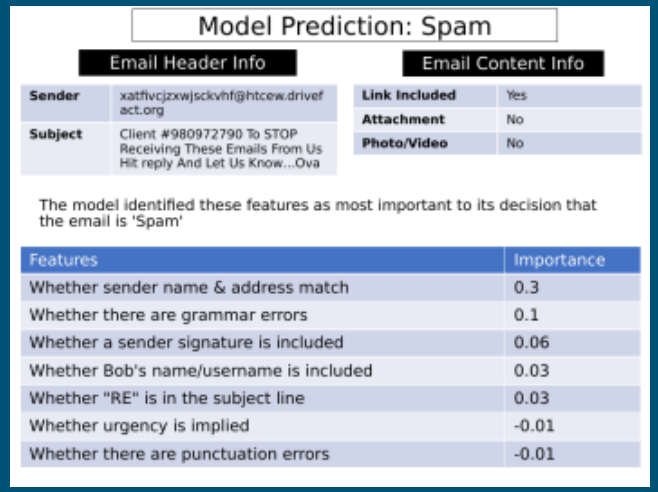








Features	Importance
Whether sender name & address match	0.3
Whether there are grammar errors	0.1
Whether a sender signature is included	0.06



Features	Importance
Whether sender name & address match	0.3
Whether there are grammar errors	0.1
Whether a sender signature is included	0.06
Whether Bob's name/username is included	0.03
Whether "RE" is in the subject line	0.03
Whether urgency is implied	-0.01
Whether there are punctuation errors	-0.01

Method

Factors Manipulated: Between Subjects

Experiment-wide model accuracy

50%

		Ground Truth	
		Spam	**Not Spam**
Model Decision	**Spam**	Hit 50%	False Alarm (FA) 50%
	Not Spam	Miss 50%	Correct Rejection (CR) 50%

88%

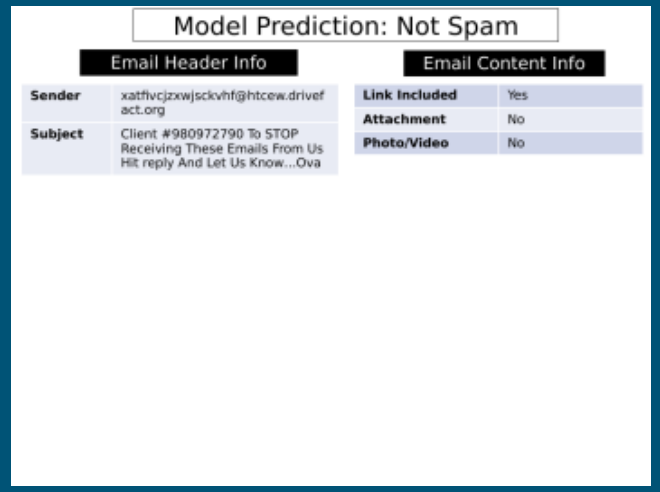
		Ground Truth	
		Spam	**Not Spam**
Model Decision	**Spam**	Hit 88%	False Alarm (FA) 12%
	Not Spam	Miss 12%	Correct Rejection (CR) 88%

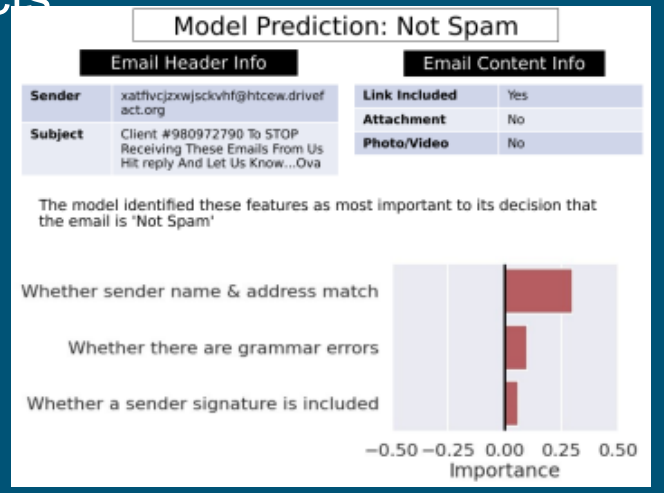
Method

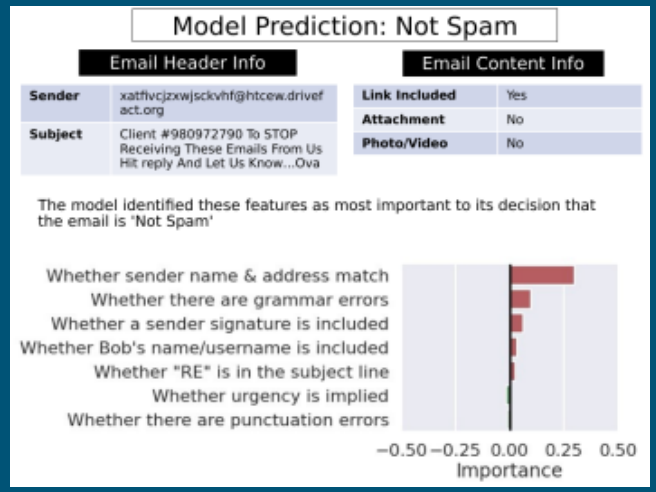
Factors Manipulated: Within Subjects

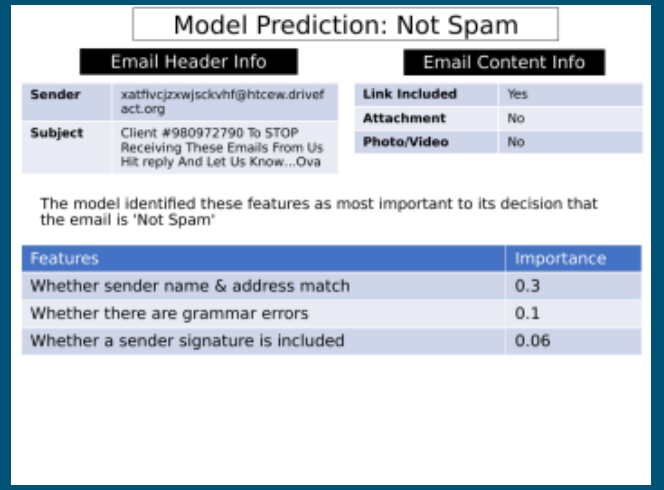
For ‘incorrect’ trials, the features and values remained the same

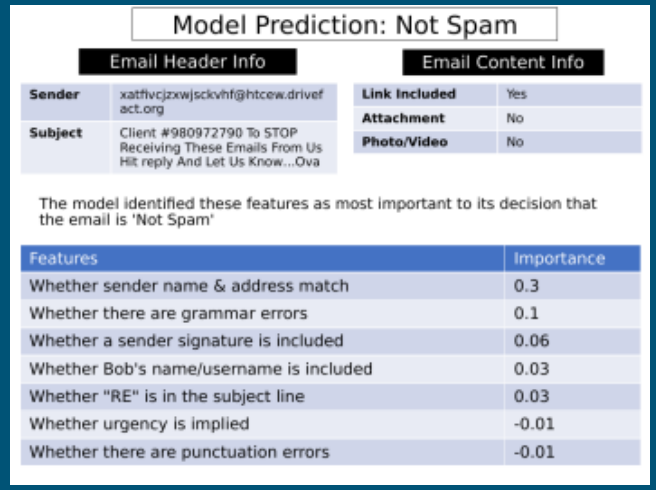
but the decision was reversed











Procedure

200 people participated via Amazon Mechanical Turk
- With approval from Sandia Human Studies Board

Each participant completed 80 unique trials
- ~~40 spam, 40 not spam~~

Measures collected include:
- Trial level accuracy
- Trial response time (RT) in milliseconds (ms)
- Trust in Automation Scale

50%

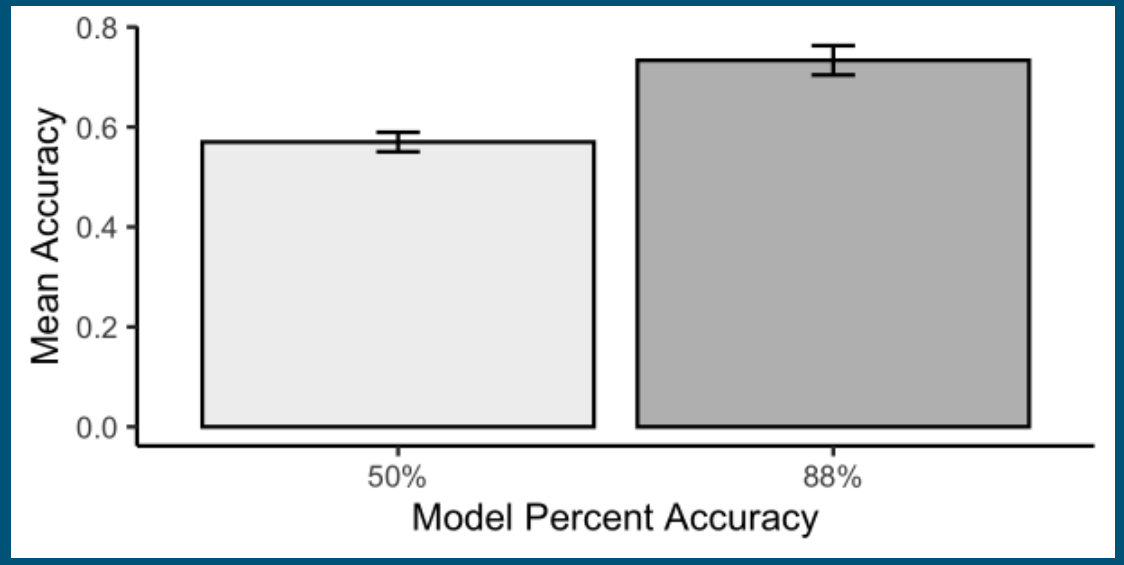
Model Decision		Ground Truth: Spam	Ground Truth: Not Spam
	Spam	**Hits: 4 each** Baseline 3-graph 3-table 7-graph 7-table	**FA: 4 each** Baseline 3-graph 3-table 7-graph 7-table
	Not Spam	**Miss: 4 each** Baseline 3-graph 3-table 7-graph 7-table	**CR: 4 each** Baseline 3-graph 3-table 7-graph 7-table

88%

Model Decision		Ground Truth: Spam	Ground Truth: Not Spam
	Spam	**Hits: 7 each** Baseline 3-graph 3-table 7-graph 7-table	**FA: 1 each** Baseline 3-graph 3-table 7-graph 7-table
	Not Spam	**Miss: 1 each** Baseline 3-graph 3-table 7-graph 7-table	**CR: 7 each** Baseline 3-graph 3-table 7-graph 7-table

Results

Question 1: Did participant accuracy difer by visualization type, model accuracy, or their interaction?



No main effect of visualization type

Significant main effect of model accuracy

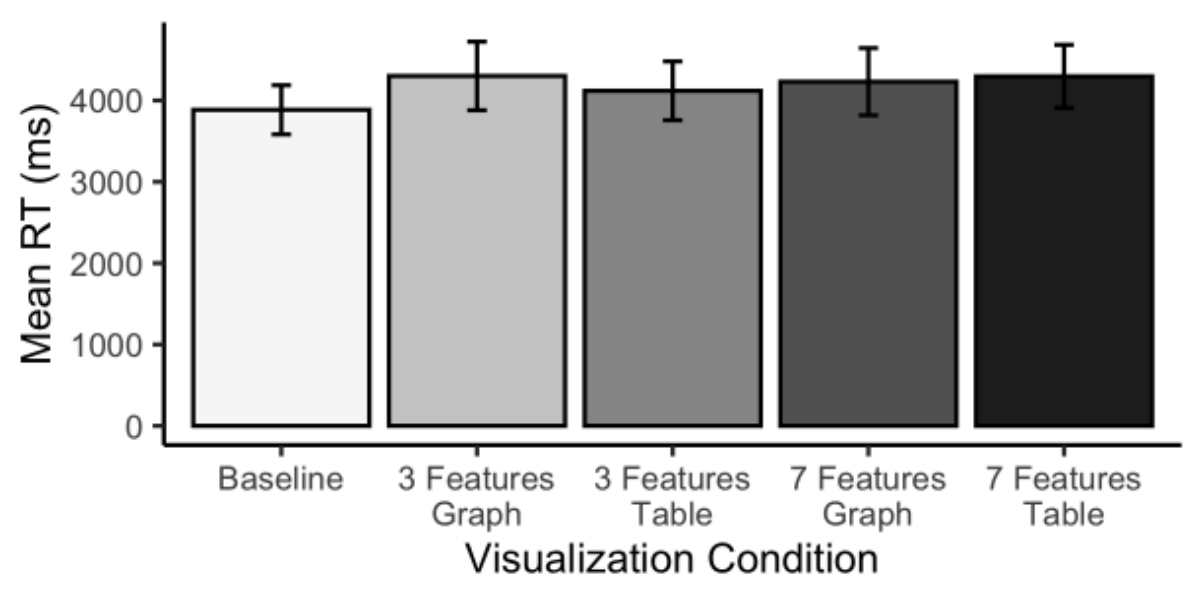
No interaction

Participants were more accurate when the model was more accurate

Our prediction that more model information would improve accuracy was not supported

Results

Question 2: Did RT differ by visualization type, model accuracy, or their interaction?



Significant main effect of visualization type

No main effect of model accuracy

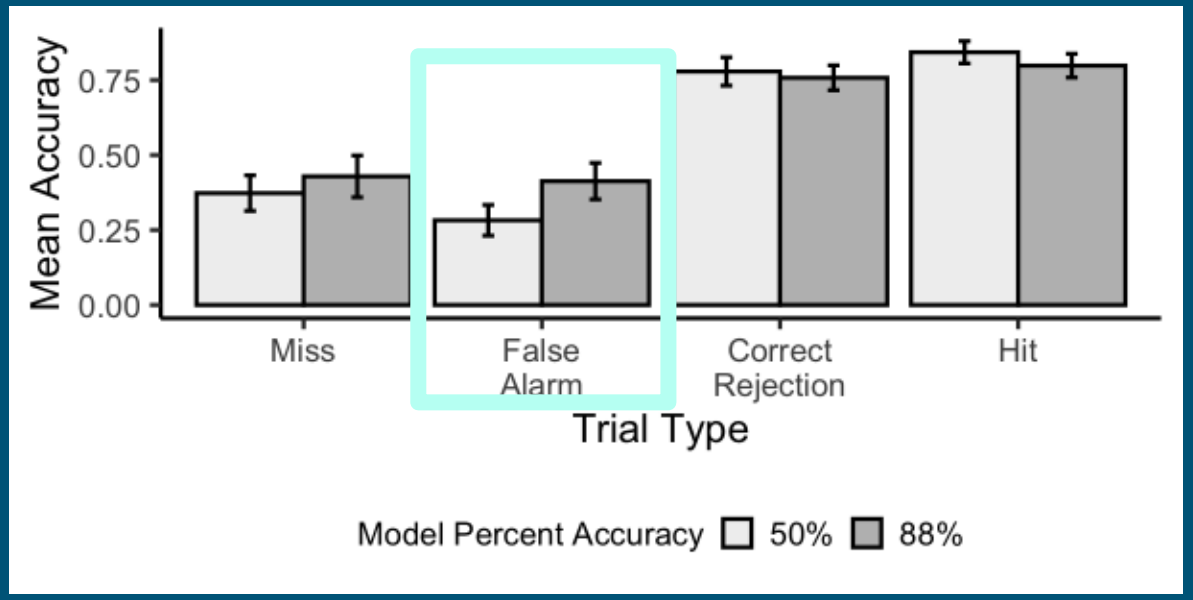
No interaction

Participants were faster in baseline than 7-feature table condition

Our prediction graph format would be easier to read than table format was not supported

Results

Question 3: Did participant accuracy differ by model prediction type (e.g., hit, FA) and/or overall model accuracy?



Marginal main effect of model prediction type

Significant main effect of model accuracy

Significant interaction

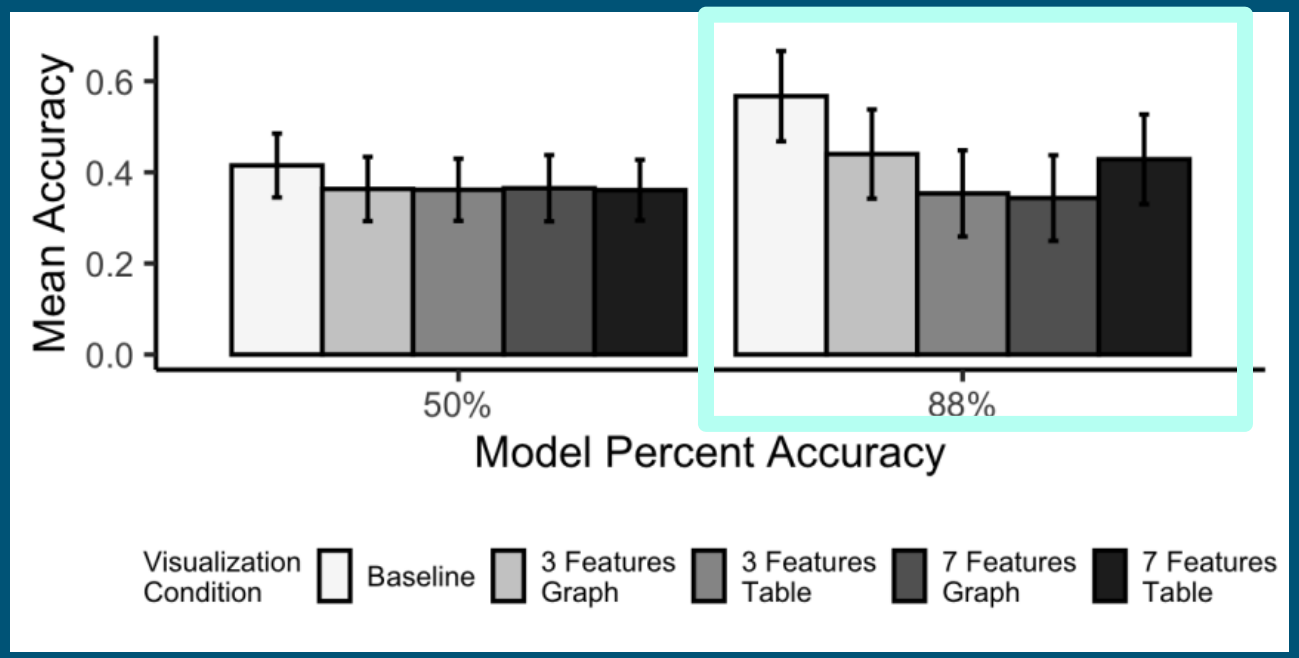
No pair-wise differences between groups for correct model predictions (Hits, CR)

Significantly higher accuracy on FA trials in 88% model accuracy condition

- Interesting because: people tend to miss rare errors
- FAs were rare in this condition
- Maybe generally accurate model helped people LEARN to identify them better?

Results

Question 4: Did participant accuracy differ by model prediction type (e.g., hit, FA), visualization type, and/or overall model accuracy?



Participant Accuracy for **Miss Trials** Only
(Spam trials that model predicted Not Spam)

Significant 3-way interaction

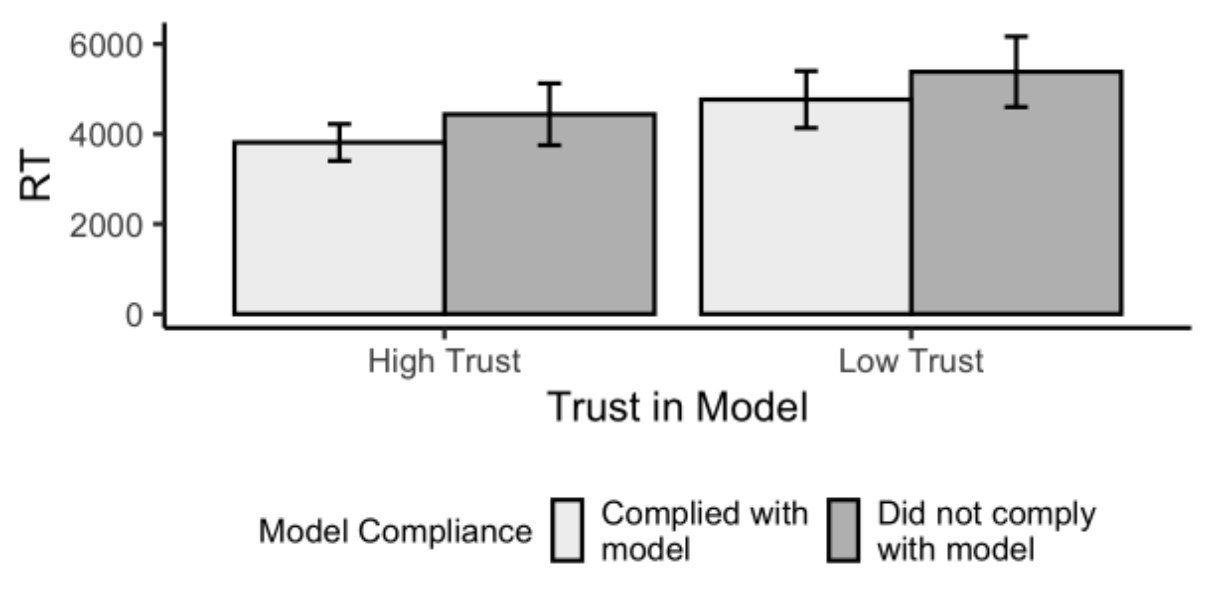
When the model MISSED a target, people were more likely to agree with the model's incorrect prediction if the model gave them MORE information

- This was especially evident in the 88% model condition

This is a risk: as models improve in accuracy, we must find a way to help human decision-makers notice these rare but potentially high-consequence model errors

Results

Question 5: Did individual differences in Trust in Automation impact participant's compliance with model prediction?



Significant main effect of compliance

Significant main effect of trust

No interaction

All participants were slower to respond when they did not comply with the model's prediction

Low trust individuals were slower overall, regardless of their compliance

Longer RTs suggest that participants were taking more time to evaluate evidence / make decision

Conclusions

Higher model accuracy produced higher task accuracy
- Possible benefit:
  - people were better able to correctly identify Non-Spam trial when model made False Alarm
- Possible risk:
  - people were LESS able to correctly identify Spam target when model missed—especially when model provided more feature information

Highly accurate models could produce over-reliance on decisions
- As ML models continue to improve, the human decision-makers must be considered

Graph vs table feature importance visualization did not impact performance

Individual differences in Trust in Automation impacted time to make decision but not likelihood of model compliance

Caveats / Future Directions

We used a domain general task and collected data from the general population

- Domain expertise, previous experience with model will likely impact reliance on ML output

Our task differed from real cybersecurity context

'Incorrect' model predictions used same features/weights as correct model predictions

Future work should continue to explore what explainable ML means, and continue to refine measurement of these concepts to enable better comparison across studies

Efficacy of ML outputs for decision-making need to take into account:

- Nature of task
- Model accuracy
- Type of errors likely to be made