Evaluating Interpretability

CS 282 BR Topics in Machine Learning: Interpretability and Explainability

Ike Lage

02/01/2023

Overview

- Evaluating interpretability in the interpretable ML community:
  - Interpretability depends on human experience of the model
  - Disagreement about the best way to measure it
- These papers:
  - Evaluating factors related to interpretability through user studies

Other Relevant Fields

- Human-Computer Interaction (HCI):
  - Theories for how people interact with technology
- Psychology:
  - Theories for how people process information
- Both have thought carefully about experimental design

Outline

- **Research paper:** “Human Evaluation of Models Built for Interpretability” by Lage et al.
- **Research paper:** “Manipulating and Measuring Model Interpretability” by Poursabzi-Sangdeh et al.
- **Discussion**

Paper 1

Human Evaluation of Models Built for Interpretability

Isaac Lage,[*1] Emily Chen,[*1] Jeffrey He,[*1] Menaka Narayanan,[*1]
Been Kim,[2] Samuel J. Gershman,[1] Finale Doshi-Velez[1]

Contributions

- Research Questions:
  - Which types of decision set complexity most affect human-simulatability?
  - Is relationship between complexity and human-simulatability context dependent?
- Approach:
  - Large scale, carefully controlled user studies

Decision Sets

- Logic-based models are often considered interpretable
- Many approaches for learning them from data

faithful and cold or brave and passive → candy or dairy and fruit

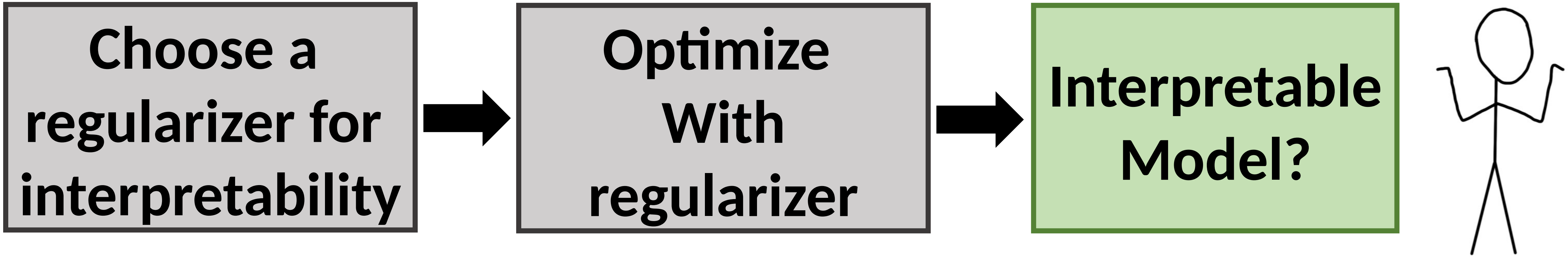
sleepy or patient and obedient → spices and grains or dairy

brave and sleepy or patient or laughing → dairy and fruit or grains

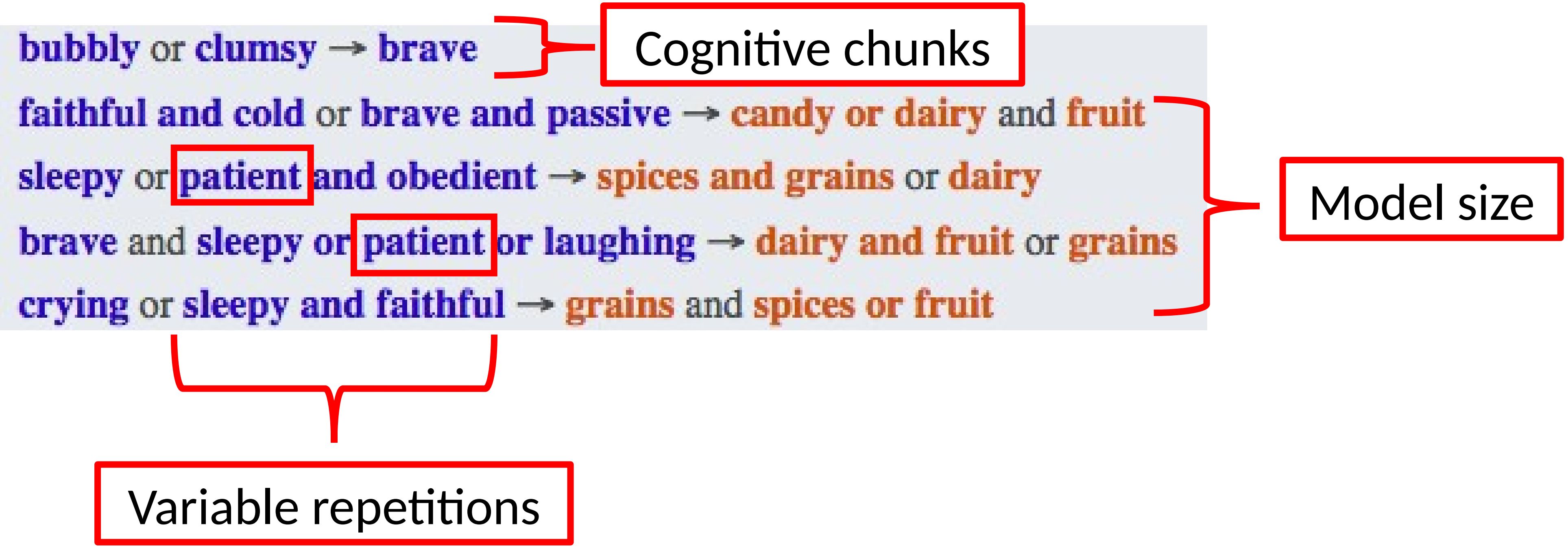
crying or sleepy and faithful → grains and spices or fruit

Regularizers

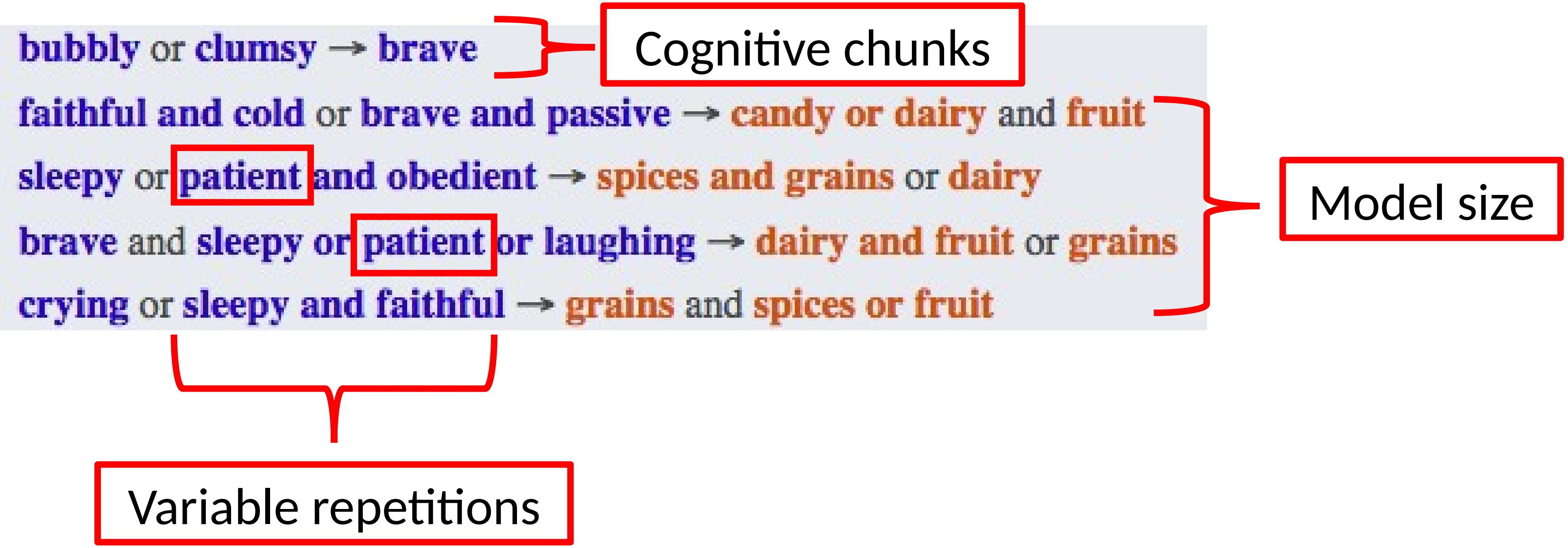
- There are many ways to regularize decision sets that make them less complex
- What kinds of complexity is it most urgent to regularize to learn interpretable models?



Types of Complexity



Types of Complexity



What if we optimized the models with data?

Context: Domains

Ingredients:

- **Vegetables:** okra, carrots, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

Low Risk: Alien meal recommendation

Disease Medications:

- **antibiotics:** Aerove, Adenon, Athoxin
- **painkillers:** Poxin, Parola, Pelapin
- **vitamins:** Vipryl, Vyorix, Votasol
- **stimulants:** Silvax, Setoxin, Soderal
- **tranquilizers:** Trasmin, Tydesol, Texopal
- **laxatives:** Lantone, Lezanto, Lexerol

High Risk: Alien medical prescription

Context: Domains

Ingredients:

- **Vegetables:** okra, carrots, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

Low Risk: Alien meal recommendation

Disease Medications:

- **antibiotics:** Aerove, Adenon, Athoxin
- **painkillers:** Poxin, Parola, Pelapin
- **vitamins:** Vipryl, Vyorix, Votasol
- **stimulants:** Silvax, Setoxin, Soderal
- **tranquilizers:** Trasmin, Tydesol, Texopal
- **laxatives:** Lantone, Lezanto, Lexerol

High Risk: Alien medical prescription

What if we used 2 different real domains?

Context: Tasks

bubbly or clumsy → brave

faithful and cold or brave and passive → candy or dairy and fruit

sleepy or patient and obedient → spices and grains or dairy

brave and sleepy or patient or laughing → dairy and fruit or grains

crying or sleepy and faithful → grains and spices or fruit

Observations: patient, wearing glasses, lazy

Recommendation: milk, guava

- **Simulation:**
  - What would the model recommend the alien?
- **Verification:**
  - Is milk and guava a correct recommendation?
- **Counterfactual:**
  - If patient were replaced with sleepy, would the correctness of the milk and guava recommendation change?

Context: Tasks

bubbly or clumsy → brave

faithful and cold or brave and passive → candy or dairy and fruit

sleepy or patient and obedient → spices and grains or dairy

brave and sleepy or patient or laughing → dairy and fruit or grains

crying or sleepy and faithful → grains and spices or fruit

Observations: patient, wearing glasses, lazy

Recommendation: milk, guava

- **Simulation**:
  - What would the model recommend the alien?
- **Verification**:
  - Is milk and guava a correct recommendation?
- **Counterfactual**:
  - If patient were replaced with sleepy, would the correctness of the milk and guava recommendation change?

What if we used more realistic tasks?

Tradeoff between control and generalizability

- Tradeoff between the ability to tightly control the experiment and running it under realistic conditions (generalizability)

Procedure

- Experiment posted on Mturk
- Takes around 20 minutes
- Participants paid 3 USD
- Excluded participants who could not complete practice questions
  - Total: 50-70 participants out of 150

Instructions ➡ 3-6 practice questions ➡ 15-18 test questions ➡ Payment code

Statistical Analysis: Linear Model

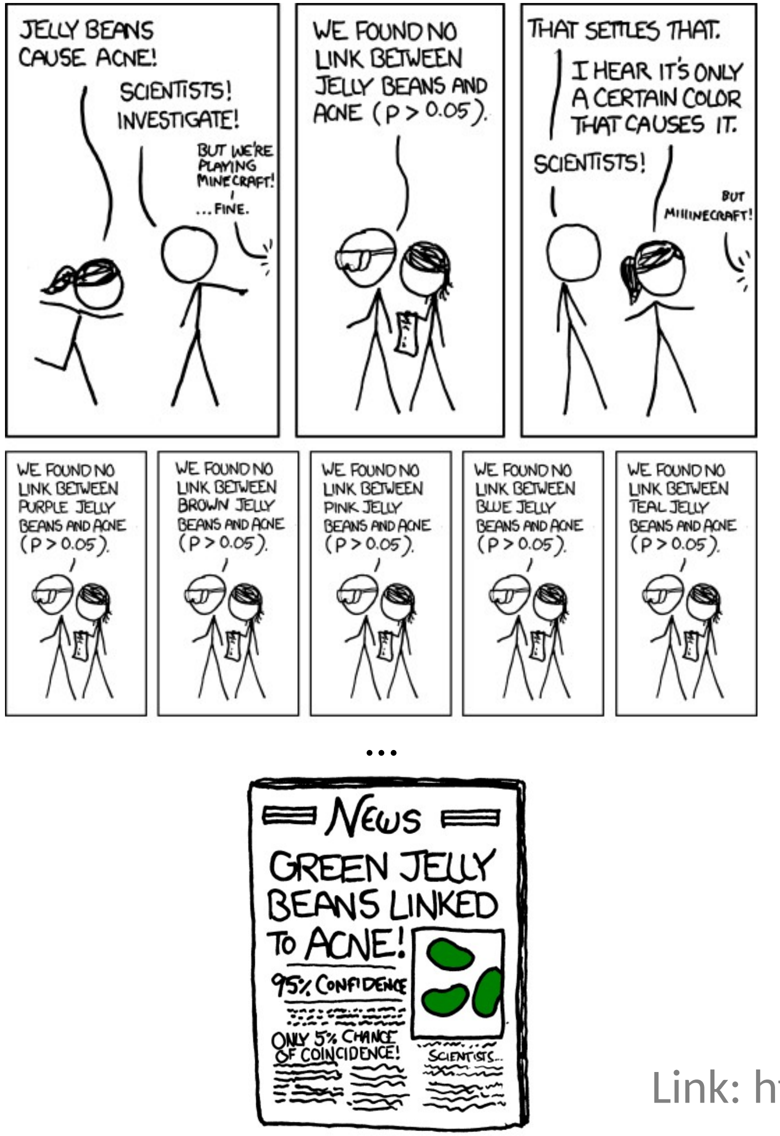
- We use a linear model for each metric in each experiment
  - Response time
  - Accuracy
  - Satisfaction

Statistical Analysis: Linear Model

- We use a linear model for each metric in each experiment
  - Response time
  - Accuracy
  - Satisfaction
- Example – Model Size, Response Time:
  - **Step 1**: Fit linear regression to predict response time from number of lines and number of output terms
  - **Step 2**: Interpret coefficients as effects of number of lines and number of output terms on response time

Stat. Analysis: Multiple Hypothesis Testing

- We use a Bonferroni correction
- Instead of $p < 0.05$, use
  $p < (0.05 / \#\ \text{comparisons})$

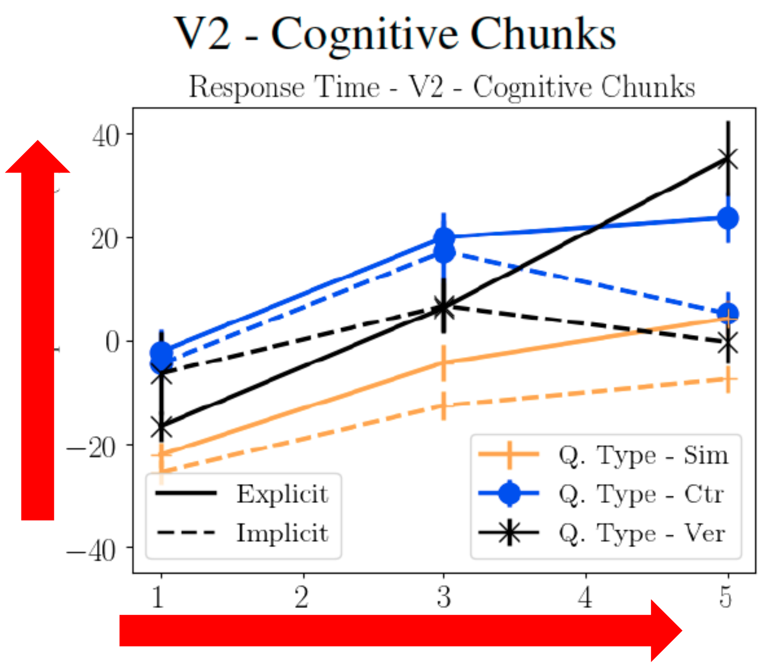


Link: https://xkcd.com/882/

Results: Complexity increases response time

Recipe Domain

Longer response time



Greater Complexity

Greater complexity results in longer response time for all kinds of complexity

Results: Type of complexity matters

		Response Time			
		Clinical		Recipe	
		Weight	P-Value	Weight	P-Value
Model size	# Lines (V1)	01.17	**<.0001**	01.01	.0032
	# Terms (V1)	02.35	**<.0001**	01.57	.0378
	Ver. (V1)	10.50	**<.0001**	04.11	.1210
	Counter. (V1)	21.00	**<.0001**	13.70	**<.0001**
Cognitive chunks	# Chunks (V2)	06.04	**<.0001**	05.88	**<.0001**
	Implicit (V2)	-13.00	**<.0001**	-07.93	**0.0005**
	Ver. (V2)	16.30	**<.0001**	15.40	**<.0001**
	Counter. (V2)	08.56	.0265	19.90	**<.0001**
Variable repetitions	# Rep. (V3)	01.90	.2470	00.88	.4630
	Ver. (V3)	13.00	.0035	13.70	**<.0001**
	Counter. (V3)	20.30	**<.0001**	16.60	**<.0001**

Significant in one domain

Significant in all domains

Significant in neither domain

Response time for: cognitive chunks > model size > repeated terms

Results: Consistency - Domains, Tasks, Metrics

		Response Time			
		Clinical		Recipe	
		Weight	P-Value	Weight	P-Value
Model size	# Lines (V1)	01.17	**<.0001**	01.01	.0032
	# Terms (V1)	02.35	**<.0001**	01.57	.0378
	Ver. (V1)	10.50	**<.0001**	04.11	.1210
	Counter. (V1)	21.00	**<.0001**	13.70	**<.0001**
Cognitive chunks	# Chunks (V2)	06.04	**<.0001**	05.88	**<.0001**
	Implicit (V2)	-13.00	**<.0001**	-07.93	**0.0005**
	Ver. (V2)	16.30	**<.0001**	15.40	**<.0001**
	Counter. (V2)	08.56	.0265	19.90	**<.0001**
Variable repetitions	# Rep. (V3)	01.90	.2470	00.88	.4630
	Ver. (V3)	13.00	.0035	13.70	**<.0001**
	Counter. (V3)	20.30	**<.0001**	16.60	**<.0001**

For example:
Similar effect sizes, both statistically significant

Results consistent across domains, tasks and the response time and subjective difficulty metrics

Results: Counterfactuals are hard

		Response Time			
		Clinical		Recipe	
		Weight	P-Value	Weight	P-Value
Model size	# Lines (V1)	01.17	<**.0001**	01.01	.0032
	# Terms (V1)	02.35	<**.0001**	01.57	.0378
	Ver. (V1)	10.50	<**.0001**	04.11	.1210
	Counter. (V1)	21.00	<**.0001**	13.70	<**.0001**
Cognitive chunks	# Chunks (V2)	06.04	<**.0001**	05.88	<**.0001**
	Implicit (V2)	-13.00	<**.0001**	-07.93	**0.0005**
	Ver. (V2)	16.30	<**.0001**	15.40	<**.0001**
	Counter. (V2)	08.56	.0265	19.90	<**.0001**
Variable repetitions	# Rep. (V3)	01.90	.2470	00.88	.4630
	Ver. (V3)	13.00	.0035	13.70	<**.0001**
	Counter. (V3)	20.30	<**.0001**	16.60	<**.0001**

In all experiments, longer response time than simulation

The counterfactual task is much more challenging than simulation!

Discussion

- Consistent guidelines for interpretability
- Simplified tasks to measure interpretability
- Using Mturk workers as a proxy for domain experts

Paper 2

Manipulating and Measuring Model Interpretability

Forough Poursabzi-Sangdeh
forough.poursabzi@microsoft.com
Microsoft Research

Daniel G. Goldstein
dgg@microsoft.com
Microsoft Research

Jake M. Hofman
jmh@microsoft.com
Microsoft Research

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft Research

Hanna Wallach
wallach@microsoft.com
Microsoft Research

Motivation

- Interpretability as a *latent property* that can be manipulated or measured indirectly
- What are the factors through which it can be manipulated effectively?
- Bring HCI methods to interpretable ML since interpretability is defined by user experience

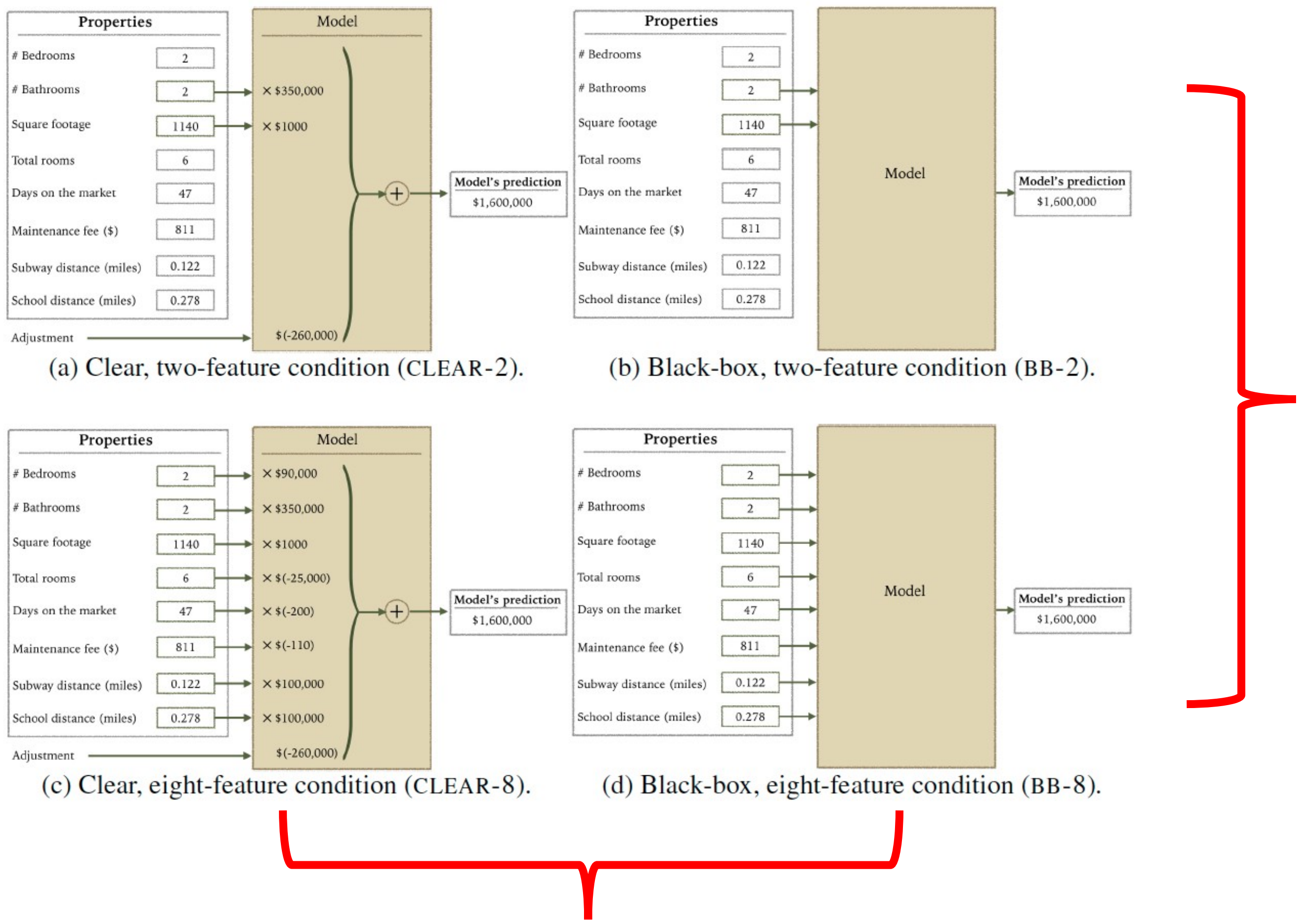
Contributions

- Research Questions:
  - How well can people estimate what a model will predict?
  - How much do people trust a model's predictions?
  - How well can people detect when a model has made a sizable mistake?
- Approach:
  - Large-scale, pre-registered user studies to answer these questions in the context of linear regression models

Comparison to Paper 1

- Studies linear regression models instead of decision sets
- Measures people’s ability to make their own predictions in addition to forward simulation
- Uses real-world housing dataset and models optimized with data

Ways to Manipulate Interpretability



(a) Clear, two-feature condition (CLEAR-2).

(b) Black-box, two-feature condition (BB-2).

(c) Clear, eight-feature condition (CLEAR-8).

(d) Black-box, eight-feature condition (BB-8).

# of Features

Transparency

Procedure

- Participants shown:
  - Training: 10 apartments
  - Testing: 12 apartments (this is the data they use)
- Participants paid 2.5 USD
- 750-1,250 participants per experiment

Each Trial

Forward simulate model's prediction ➡ View model's true prediction ➡ Make own prediction

Statistical Analysis: Participant Specific Effects

- A repeated measures experimental design
  - Each participant makes many predictions
- Use a mixed-effects model to control for correlations between a participant's responses
  - Assumes a random, participant-specific effect

Stat. Analysis: Multiple Hypothesis Testing

- Pre-registering hypotheses corresponds to deciding and publishing which analyses you will run before collecting data
- Reduces the probability that effects were discovered by chance

For example:

4) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

We will use 2-by-2 ANOVA for statistical analysis of the effect of number of features and model clarity on final deviation from model's prediction and simulation error. We will look at the effect of individual factors as well as their interactions.

https://aspredicted.org/xy5s6.pdf

Design choices

- Randomized the order of the first 10 (normal) apartments and fixed the order of the last 2 (unusual)
- All participants are shown an identical set of apartments
- Each participant completed a single condition (between subjects design)

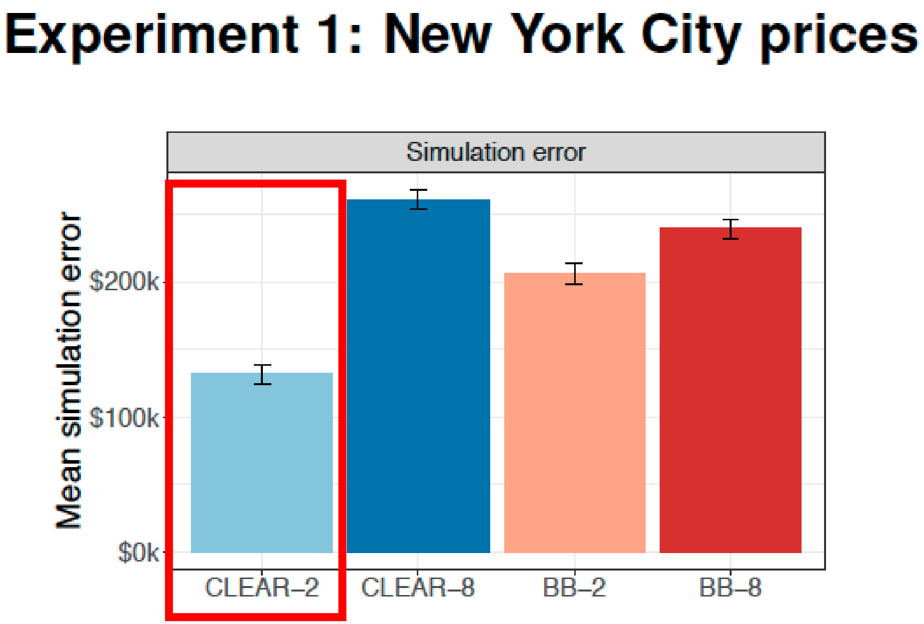
Design choices

- Randomized the order of the first 10 (normal) apartments and fixed the order of the last 2 (unusual)
- All participants are shown an identical set of apartments
- Each participant completed a single condition (between subjects design)

Can introduce bias **Increases variance**

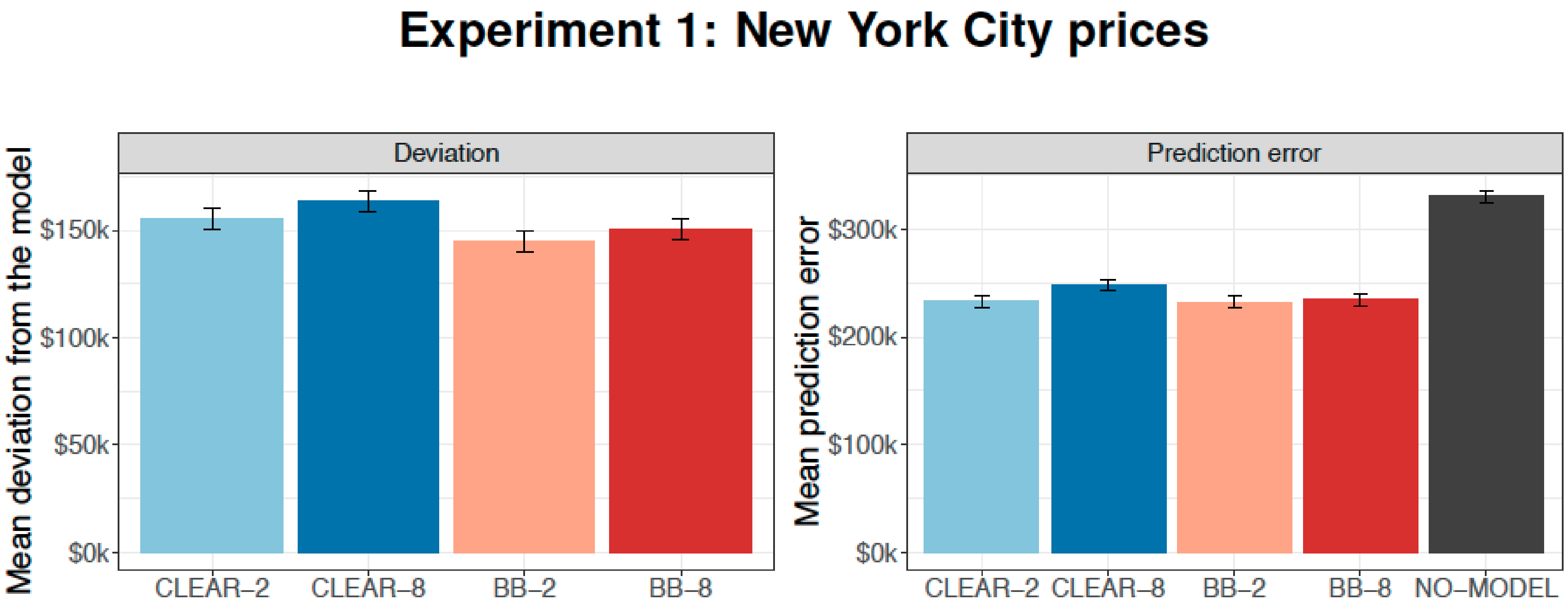
Fix sources of randomness **Randomize as much as possible**

Results: Simulating small, transparent models



Best simulation accuracy with small, transparent models

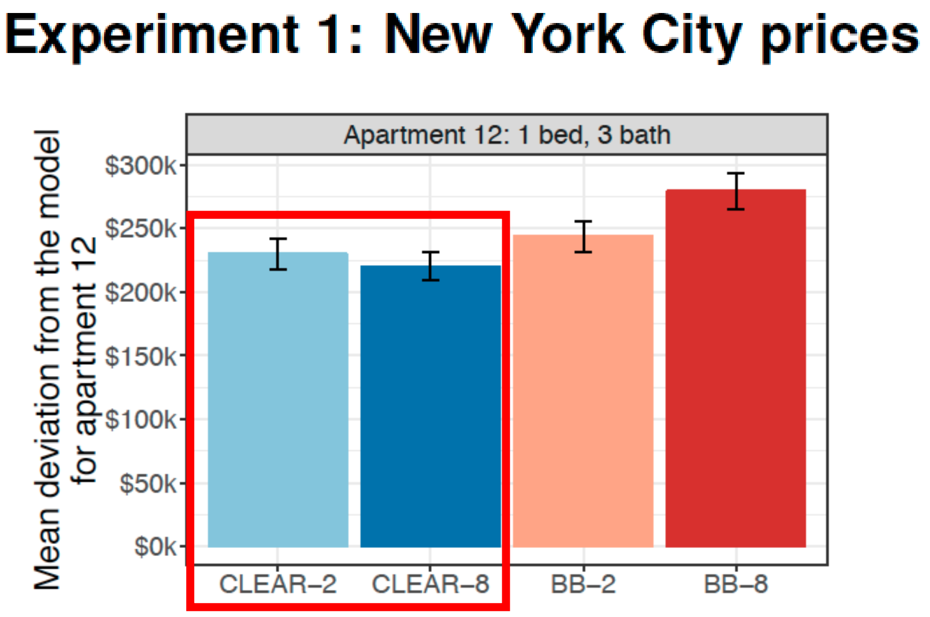
Results: No difference in trust or prediction



None of the conditions are statistically different for trust or prediction error

Results: Clear models make mistakes worse

Deviation
Higher is better



Participants deviate less from the *bad* prediction with clear models

Additional Experiments

- Scaled down prices to better reflect national average
  - Same results

Additional Experiments

- Scaled down prices to better reflect national average
- Better trust metrics
  - No significant different in trust between models

Additional Experiments

- Scaled down prices to better reflect national average
- Better trust metrics
- Attention check for unusual features
  - People catch more errors

Discussion

- Highlighting weird inputs helps catch errors
- Having people predict before seeing the model helped catch errors
- Transparency actually makes people worse at catching errors