Measuring Machine Learning Model Interpretability

Felix Feldmann

19th April 2018

„Interpretability is the degree to which a human can understand the cause of a decision.“

Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.”

Approach 1: Create Simple Models

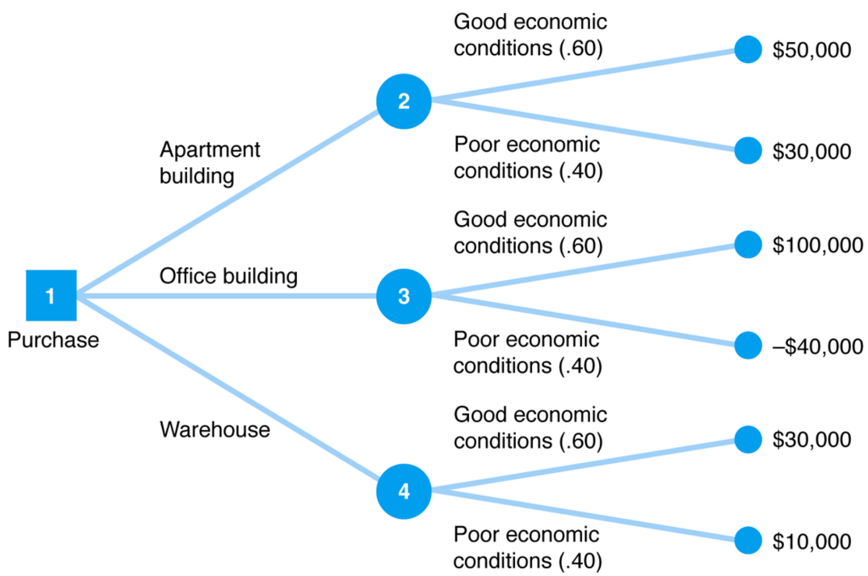
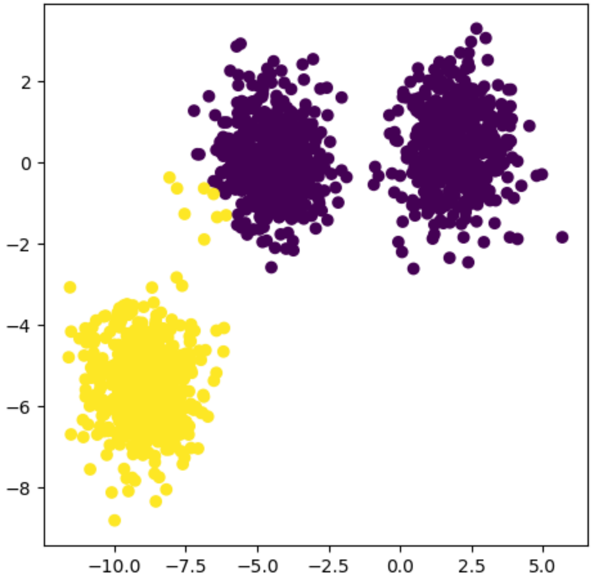


Image: http://pooptronica.com/decision-tree-diagram.html

Small decision tree

Approach 2: Design Simple Explanations



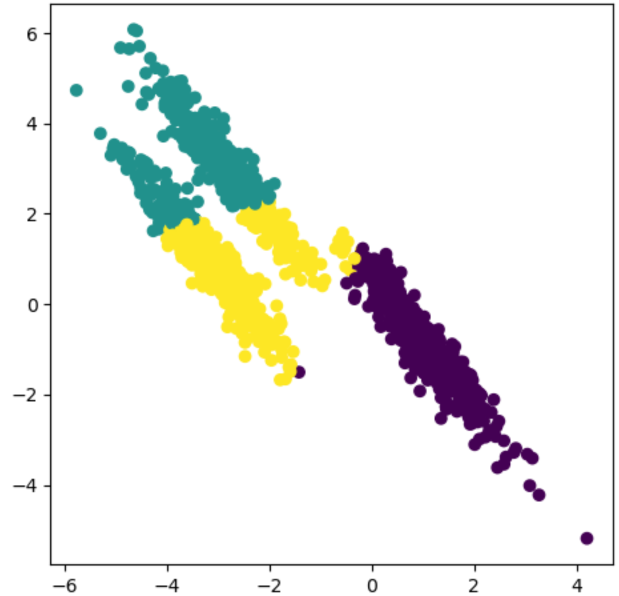
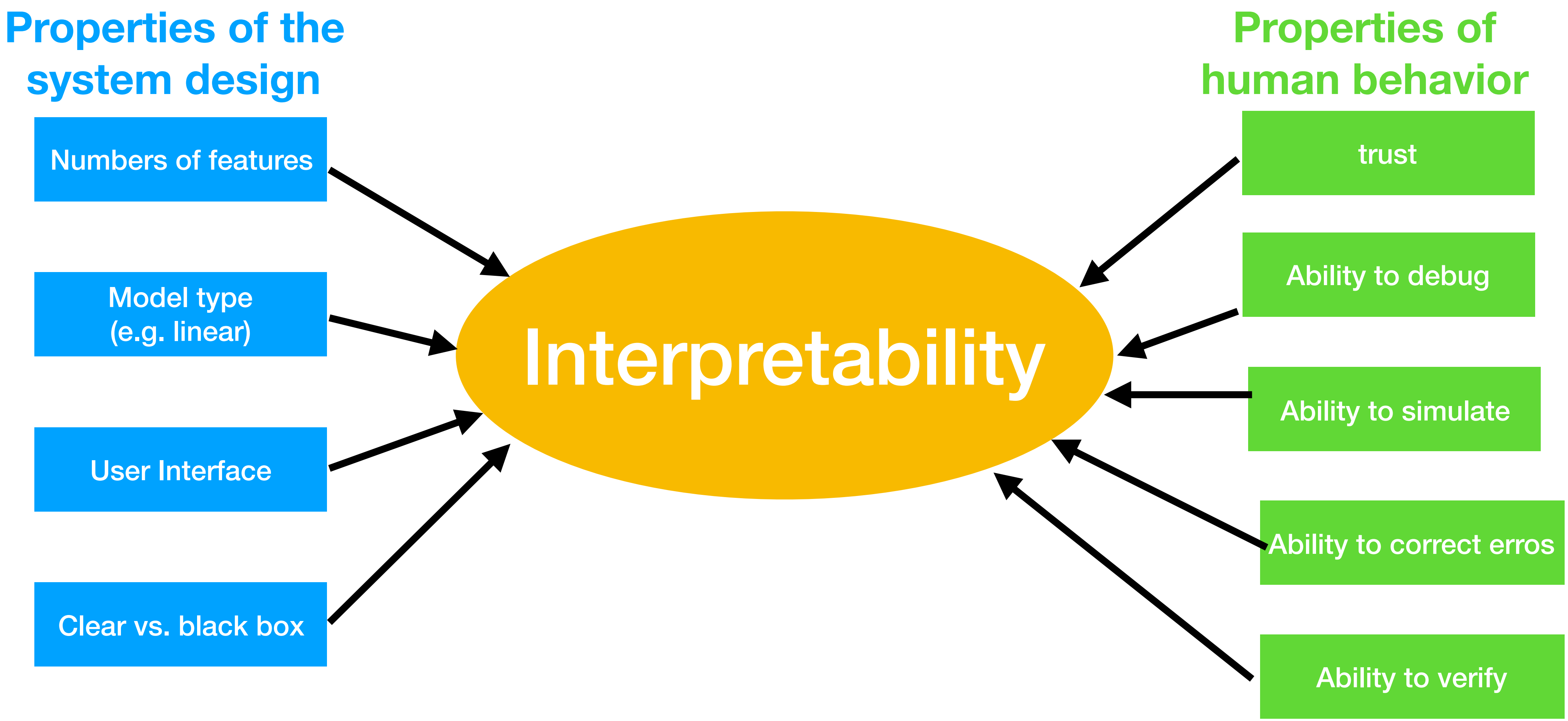


Image: sklearn k-means examples

Visualization of Complex Problems

What is Interpretability for a Machine Learning model?

Interpretability as a Latent Property



Interpretability is not a purely computational problem.

Interdisciplinary approaches necessary to address it.

Legal Necessity

Image: https://woocommerce.com/2017/12/gdpr-compliance-woocommerce/

The data controller shall provide „**meaningful information about the logic involved**, as well as the significance and the envisaged consequences of such processing.“

Different Users Different Needs

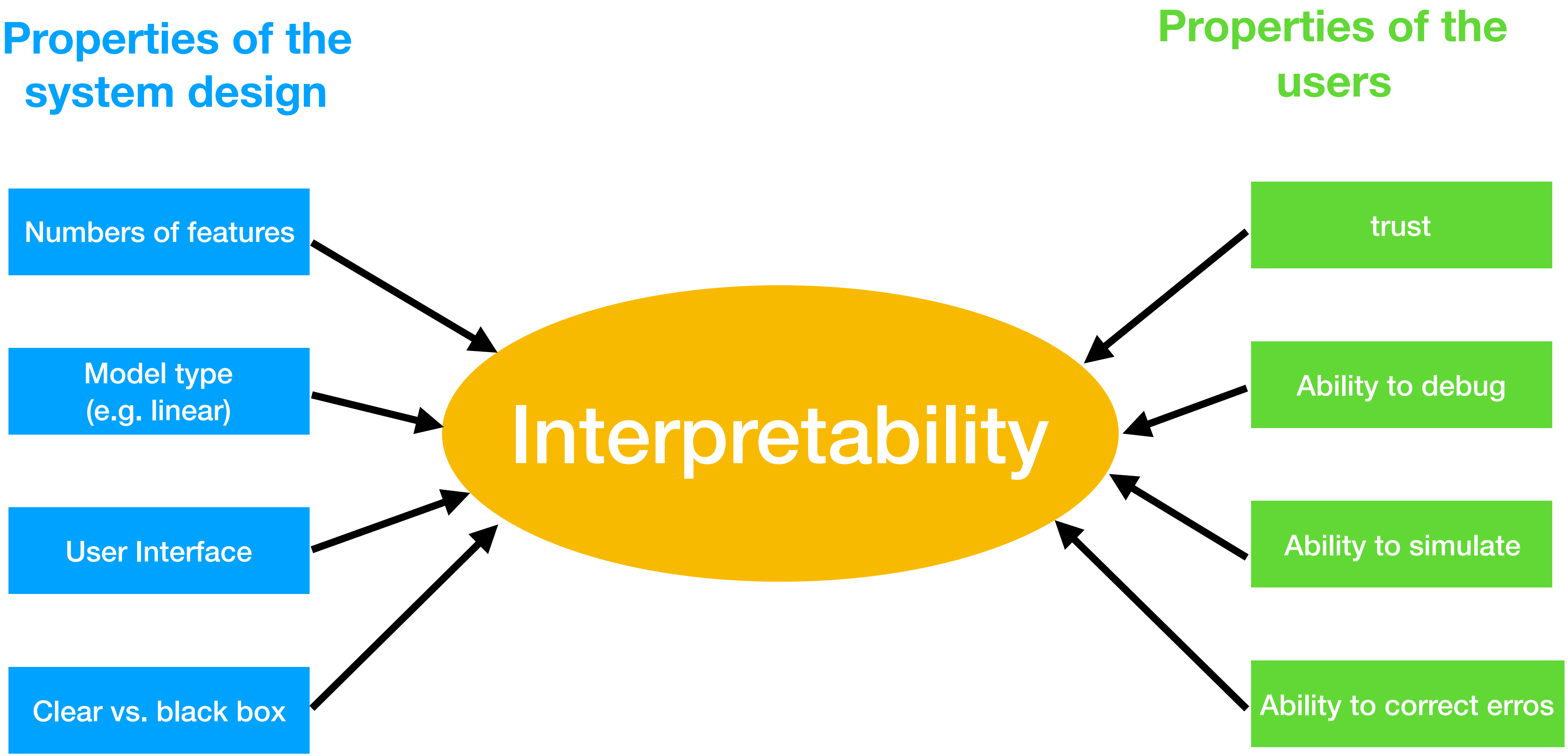
	Explain prediction	Make better decisions	Debug model
CEO		Approach A	
Data scientists			Approach C
Lay people	?	?	?
Regulators	Approach B		

Papers

[1] **Manipulating and Measuring Model Interpretability, February 2018**
Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, Hanna Wallach

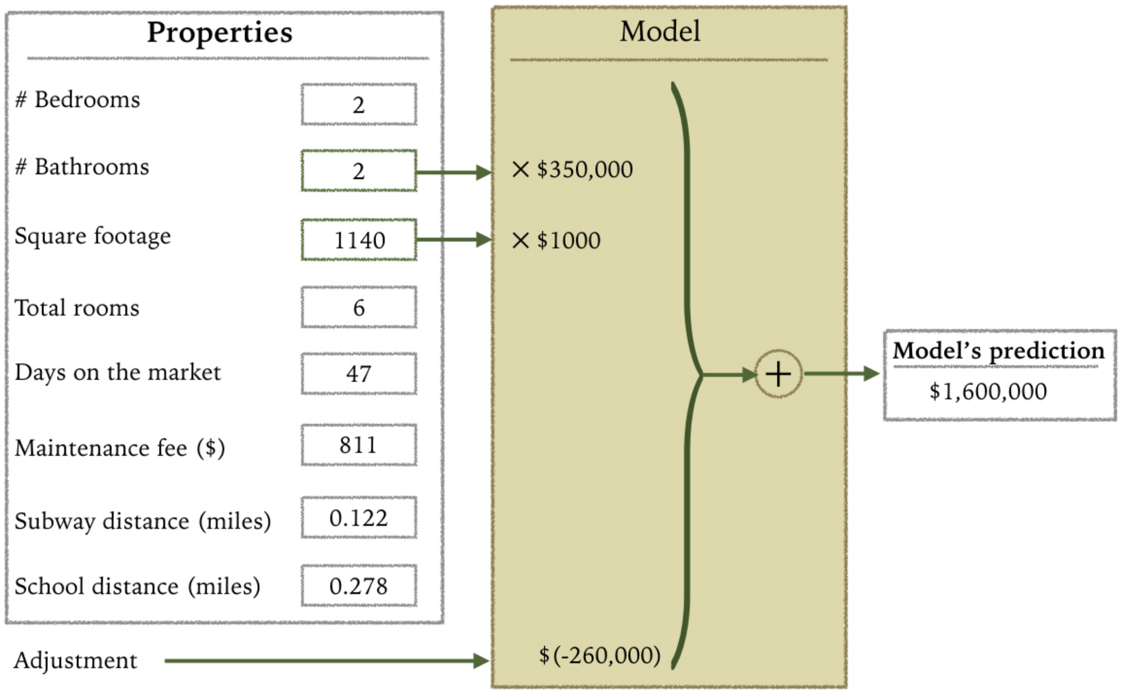
[2] **How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation, February 2018**
Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, Finale Doshi-Velez

Goal of paper [1]: Apply approach to understand the fundamental properties of human behavior relevant to interpretability.


Properties of the system design
Properties of the users
Numbers of features
Model type (e.g. linear)
User Interface
Clear vs. black box
Interpretability
trust
Ability to debug
Ability to simulate
Ability to correct erros

Predictive Tasks

- Participants asked to predict the prices of apartments in New York with the help of a (linear regression) model

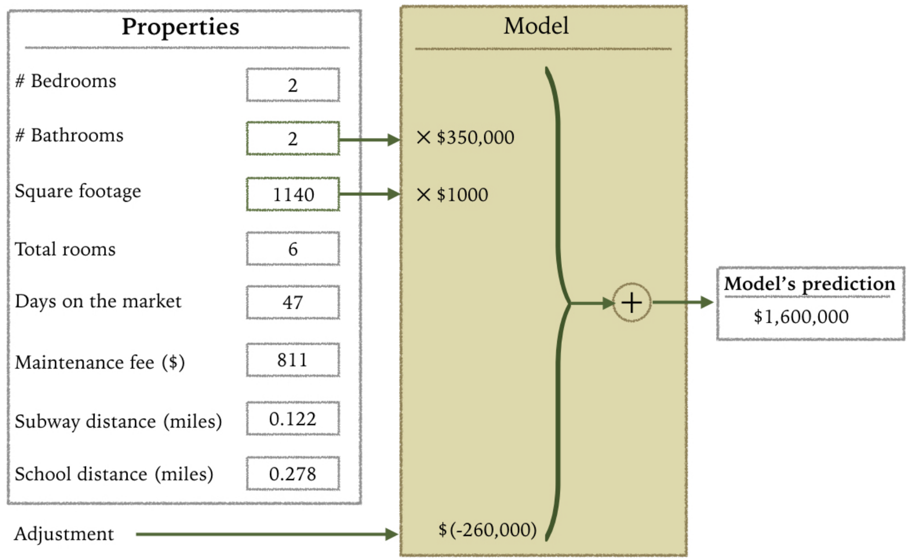


Experiment [1]

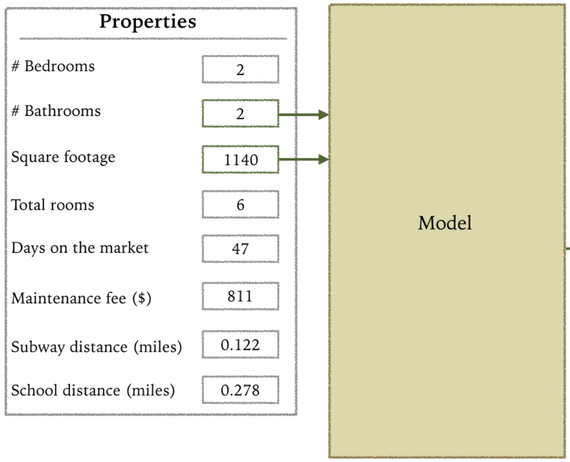
- 1250 participants from Amazon Mechanical Turk

- **Variation of**
  - Number of features
  - Black-box vs. clear models

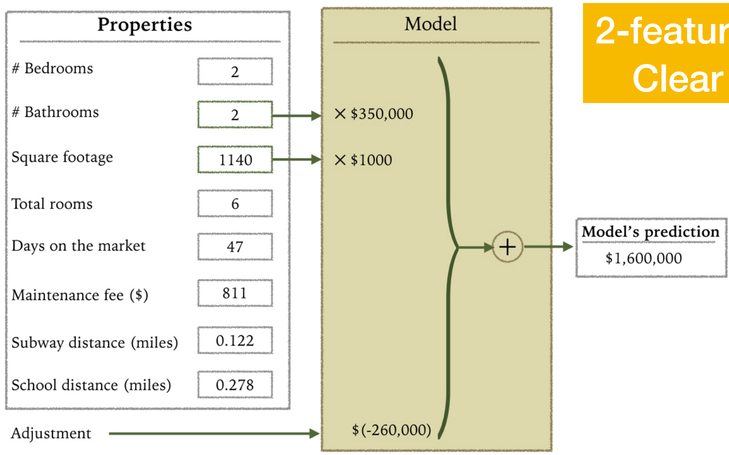
- **Measurements taken**
  - Trust in the model
  - Simulatability
  - Error of the user's predictions



Experimental Conditions

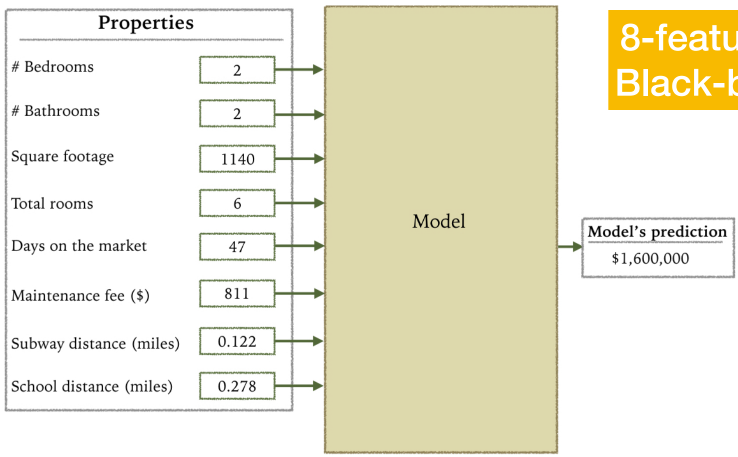


2-feature, Black-box

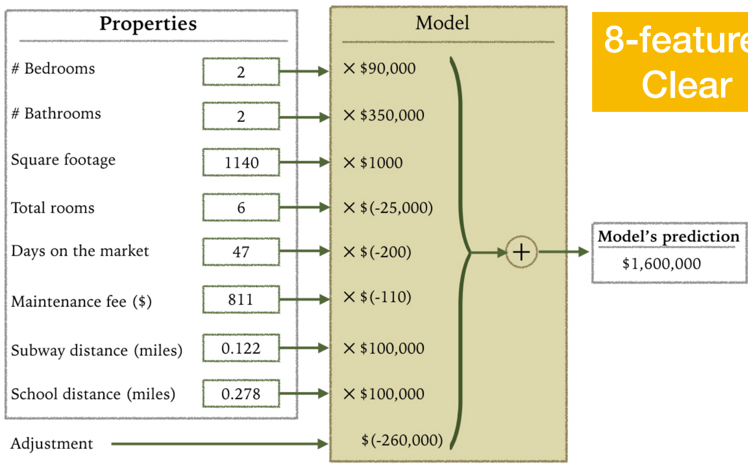


2-feature, Clear

Baseline with no model

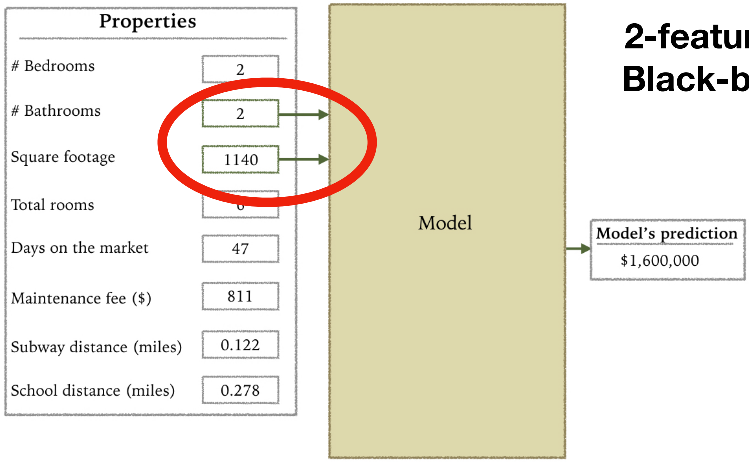


8-feature, Black-box

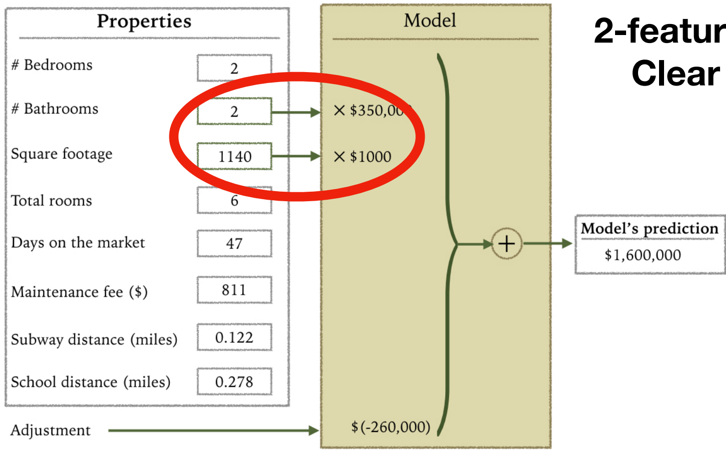


8-feature, Clear

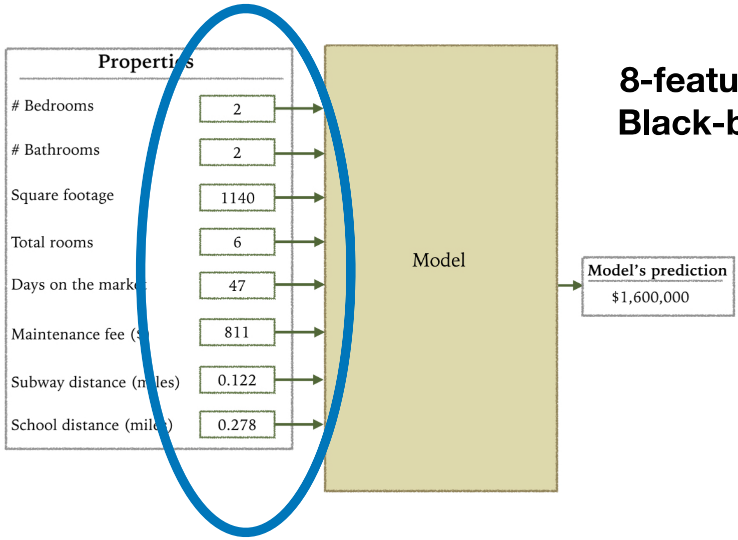
Experimental Conditions



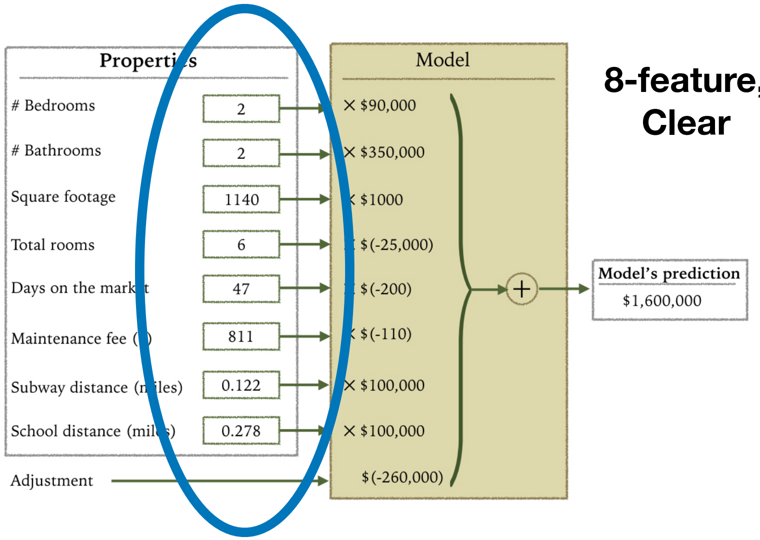
2-feature, Black-box



2-feature, Clear

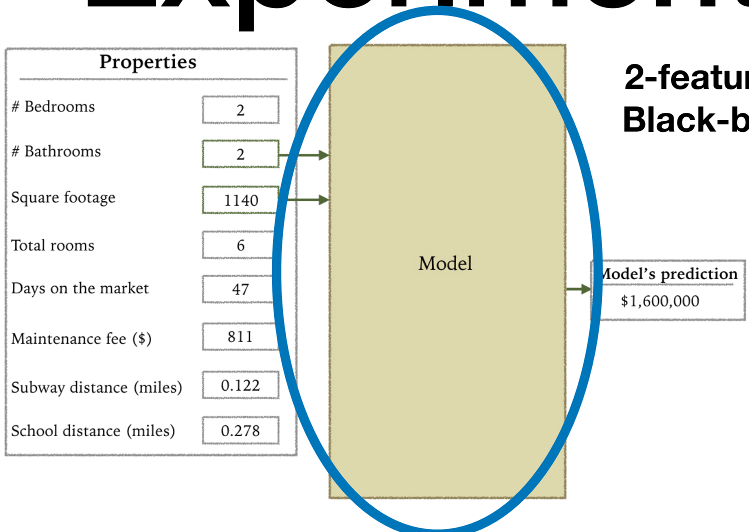


8-feature, Black-box

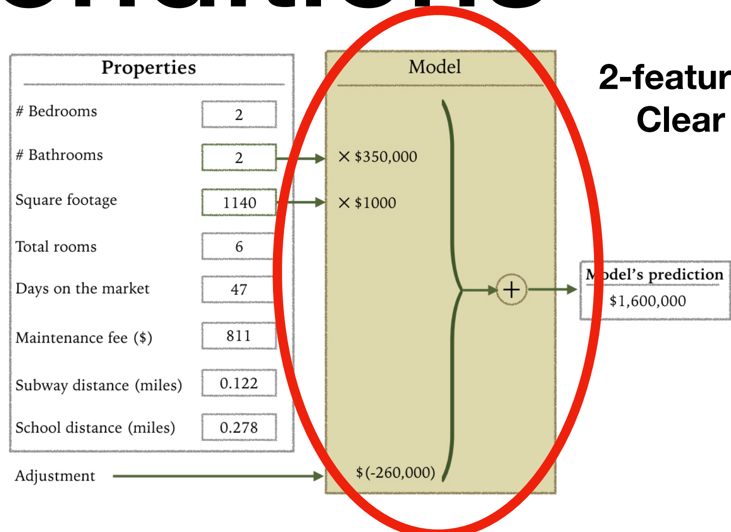


8-feature, Clear

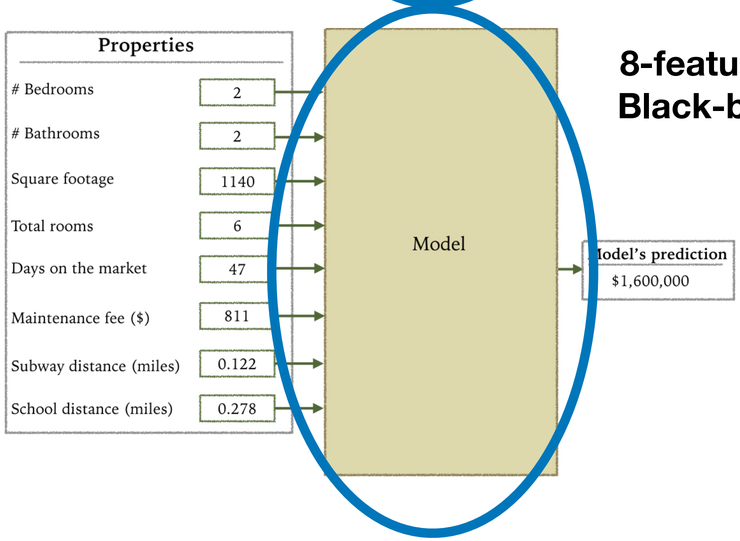
Experimental Conditions



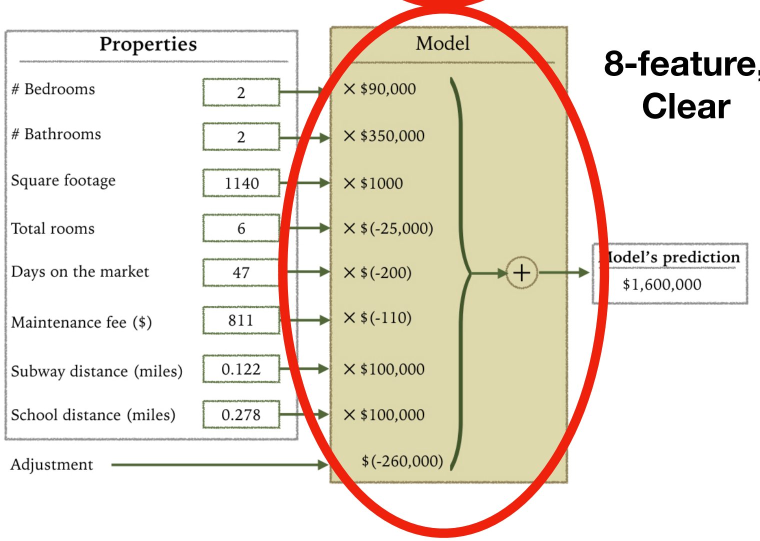
2-feature, Black-box



2-feature, Clear

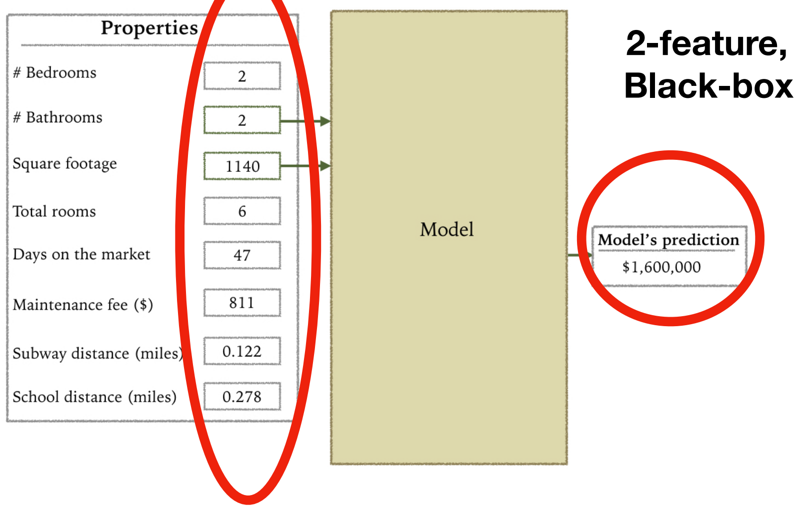


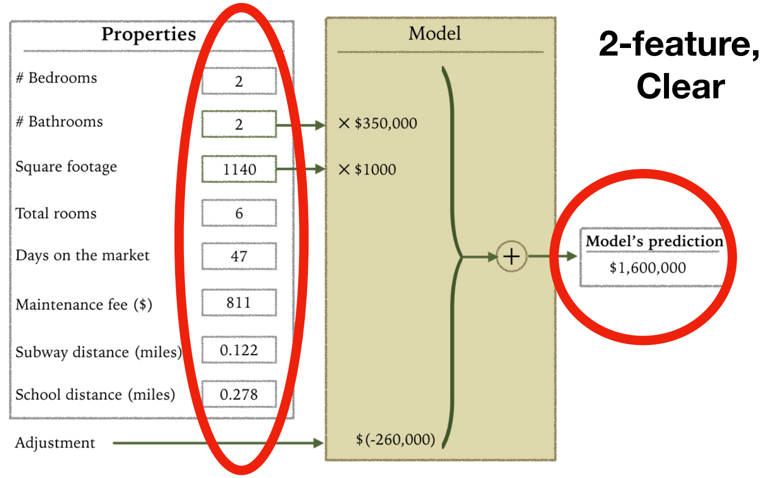
8-feature, Black-box

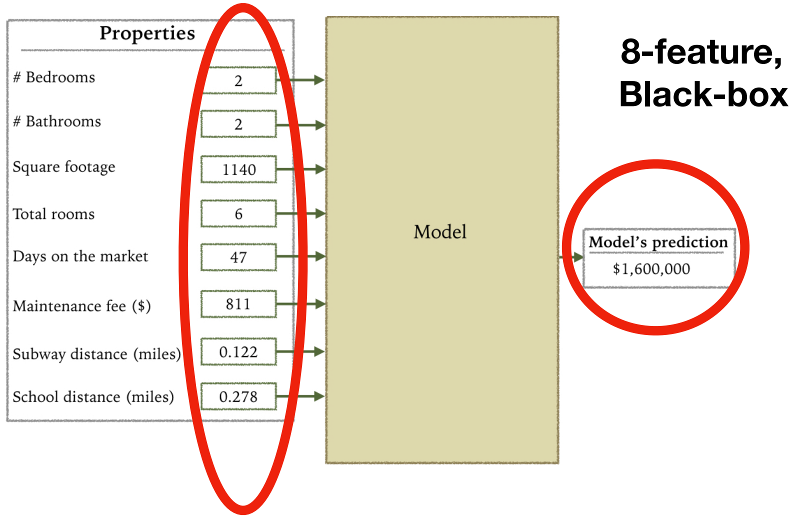


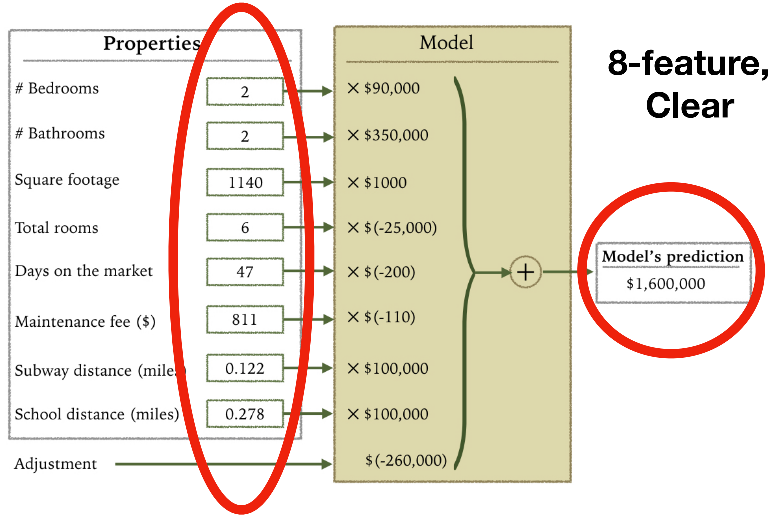
8-feature, Clear

Experimental Conditions

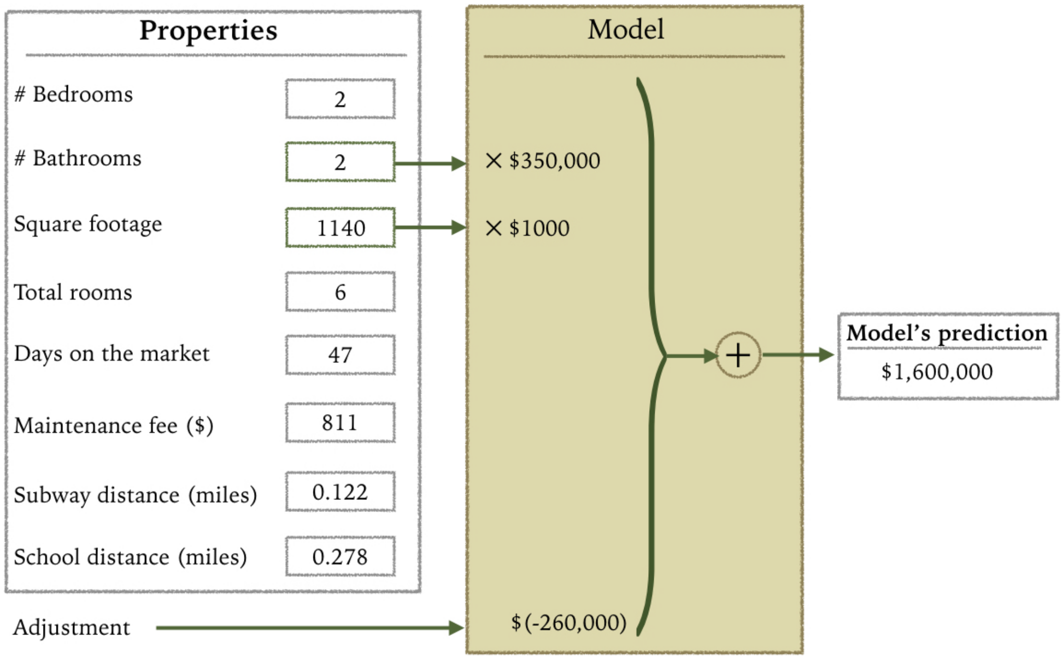








Training Phase - Experimental Interface



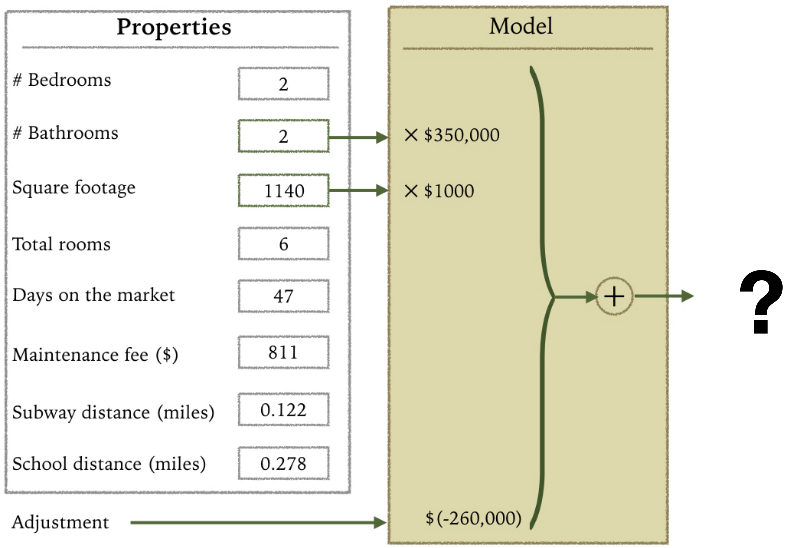
10 Apartments for each user

1. Participants were shown the apartment and their prediction

2. Participants had to make their own predictions

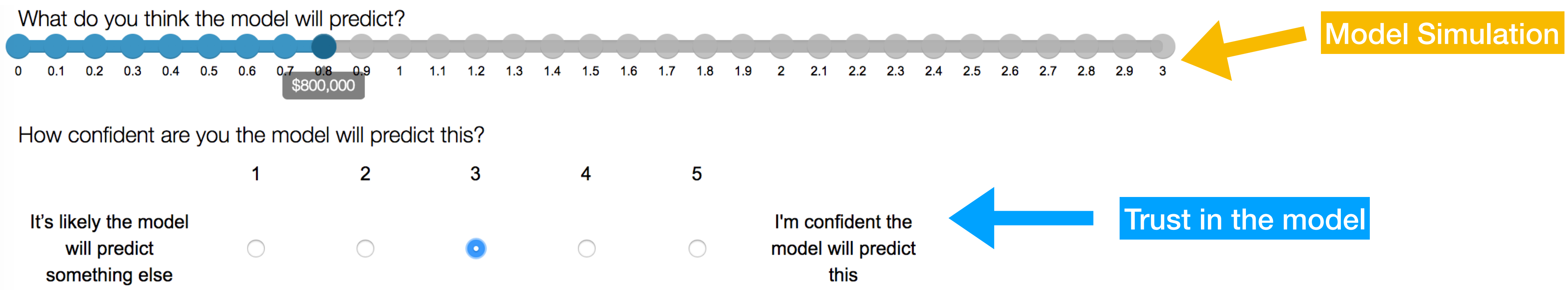
3. Participants were shown the real values

Test Phase - Experimental Interface \1

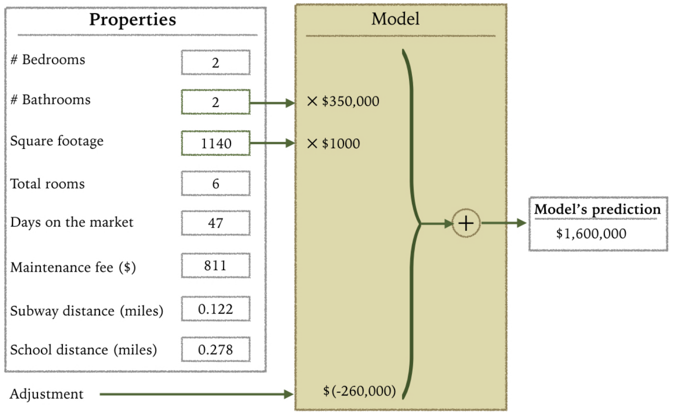


1. Participants were asked to guess what the model will predict (simulatablity).

2. Participants were asked in their confidence in their prediction.

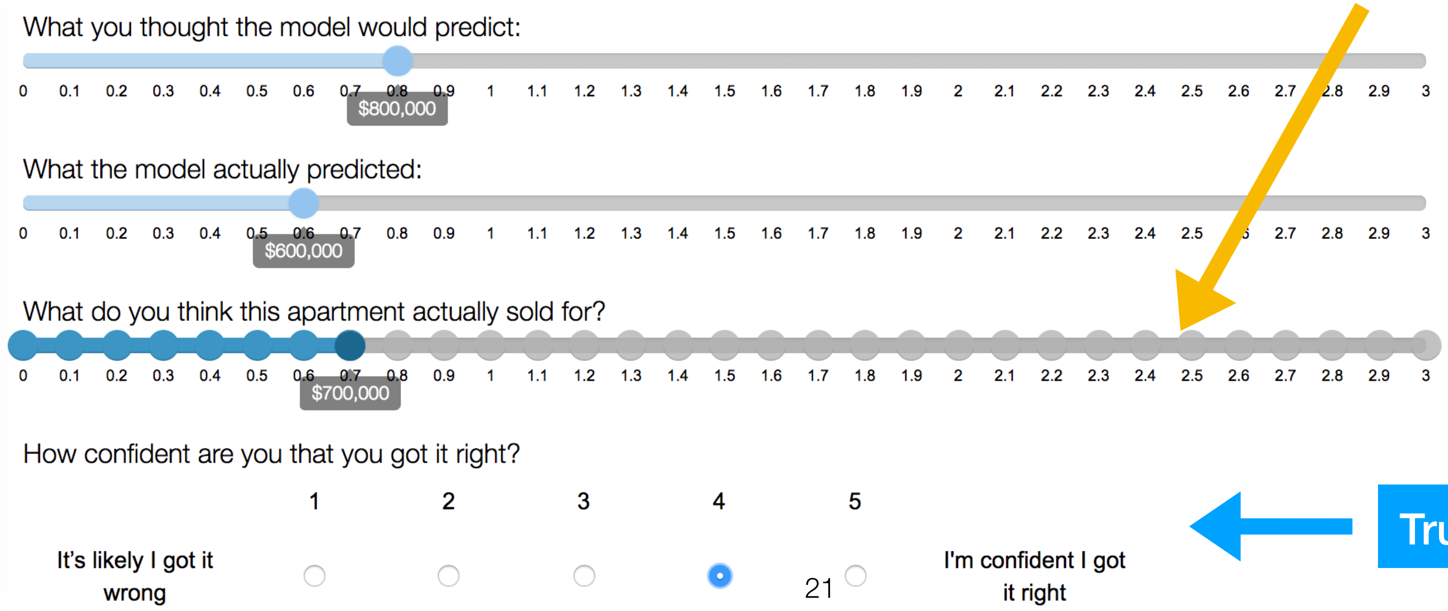


Test Phase - Experimental Interface \2



3. What was the apartment actually sold for? (trust in the model, ability to make good prediction, based on the model)

Final prediction



Trust in the model

(Pre-registered) Hypotheses

1. The clear, 2-feature model will be easiest for participants to simulate.

2. Participants will follow the clear, 2-feature model more than the black-box, 8-feature model.

3. Behavior will vary across conditions when an unusual example leads a model to make a highly inaccurate prediction. (later)

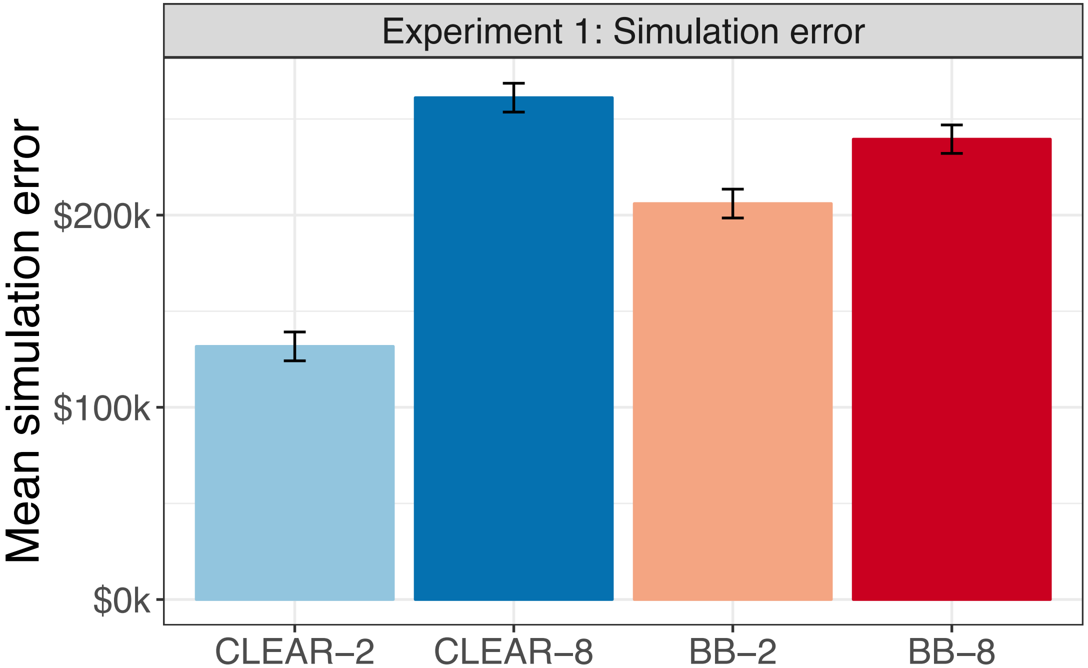
(Pre-registered) Hypotheses

1. **The clear, 2-feature model will be easiest for participants to simulate.**

2. Participants will follow the clear, 2-feature model more than the black-box, 8-feature model.

3. Behavior will vary across conditions when an unusual example leads a model to make a highly inaccurate prediction. (later)

Result: Simulation Error



- Simulation Error: **|model prediction - users guess of model prediction|**
- As hypothesized: lower simulation error in CLEAR-2 model than others.
- Not only transparency, also number of features relevant!

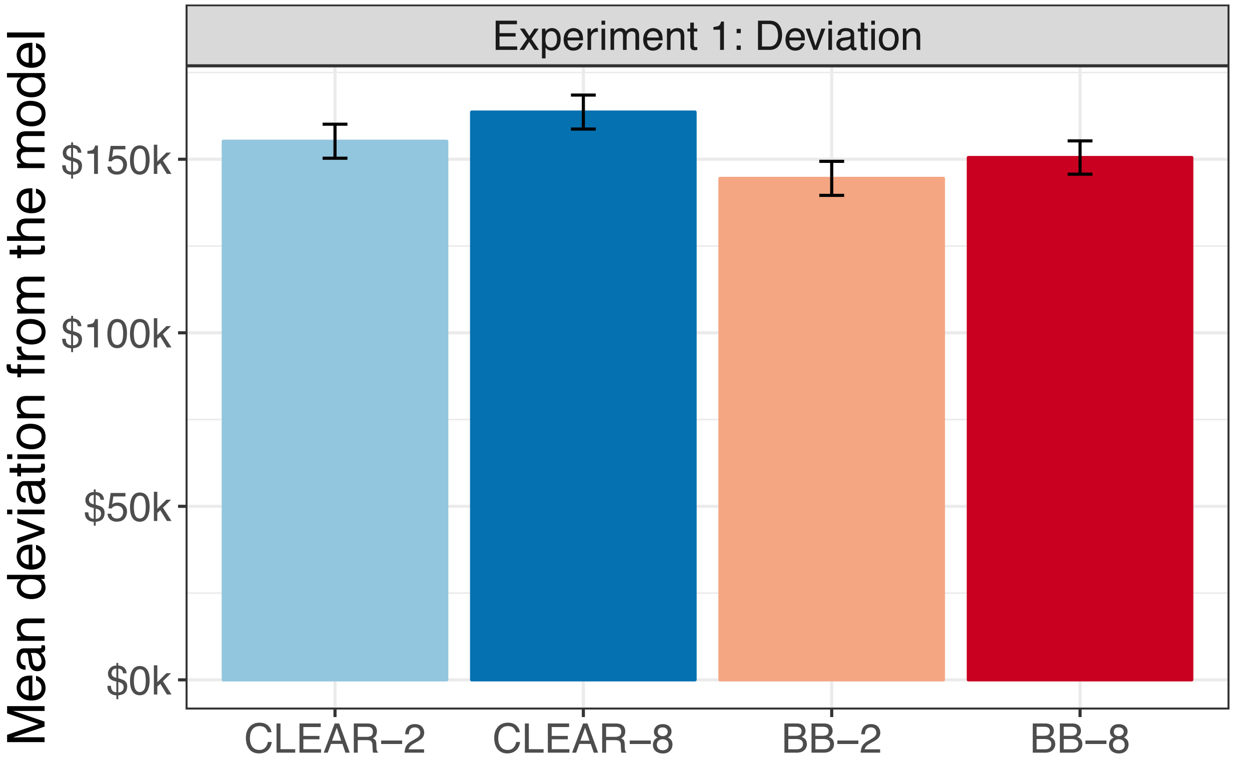
(Pre-registered) Hypotheses

1. The clear, 2-feature model will be easiest for participants to simulate

2. **Participants will follow the clear, 2-feature model more than the black-box, 8-feature model.**

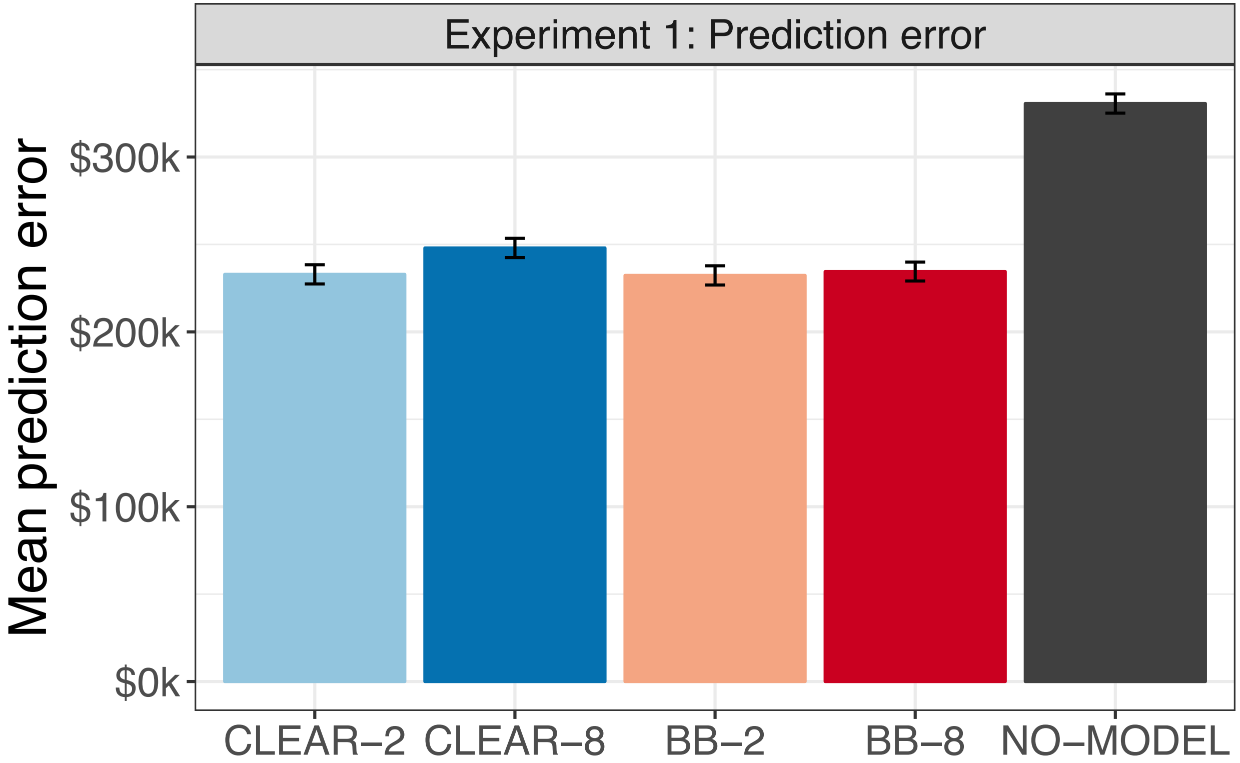
3. Behavior will vary across conditions when an unusual example leads a model to make a highly inaccurate prediction.

Result: Deviation error



- Deviation error: **|model prediction - participant's final prediction|**.
- Smaller value indicates higher trust in the model.
- Obviously hypothesis does not hold.
- All have the same impact on peoples predictions.

Prediction error

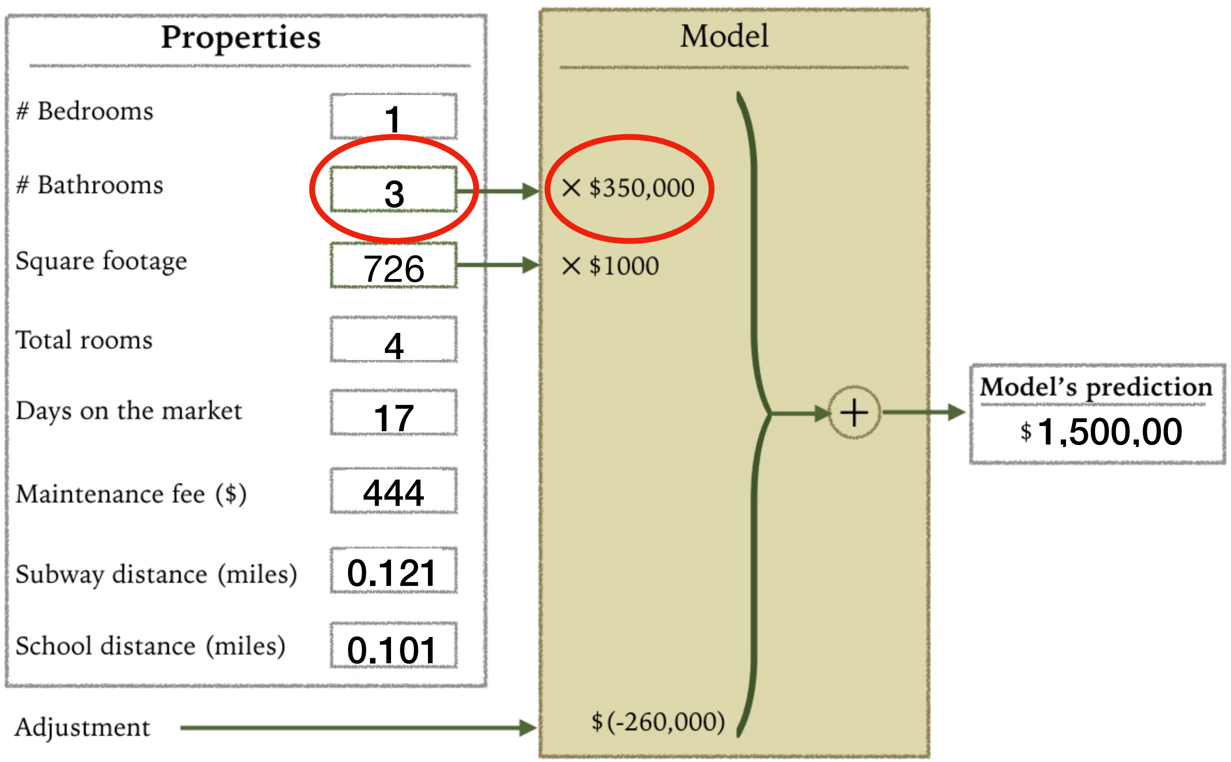


- Prediction error: **|actual price - participant's prediction|**
- No significant difference between the four models.
- Baseline condition with no model much higher, model helps making better predictions.

(Pre-registered) Hypotheses

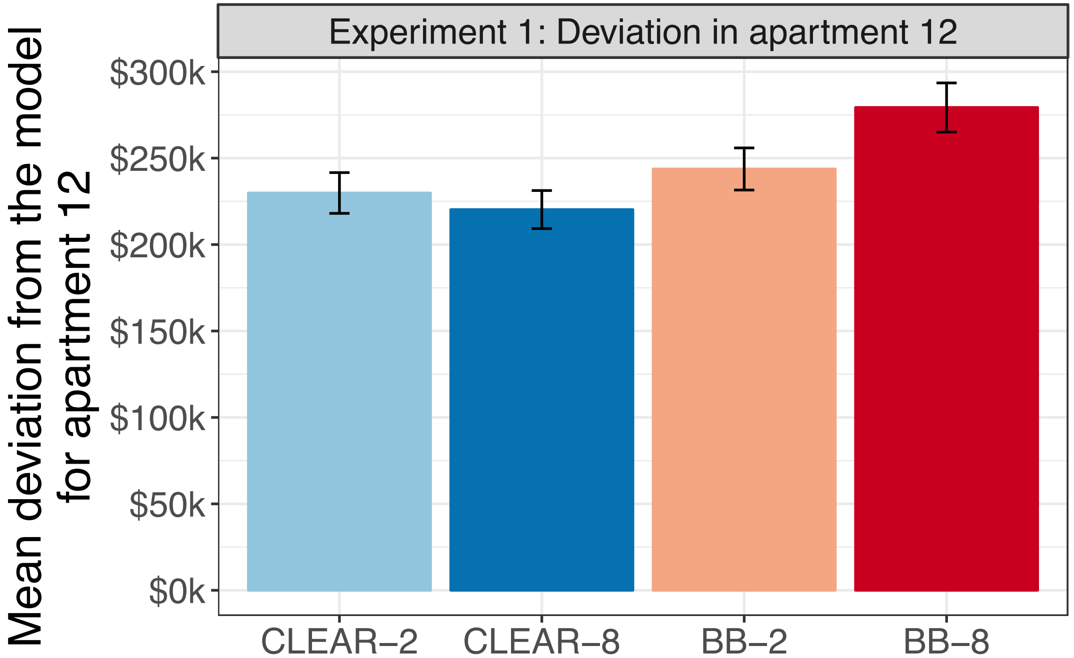
1. The clear, 2-feature model will be easiest for participants to simulate
2. Do they trust the clear 2 feature model more than the black-box, 8-feature model?
3. **Behavior will vary across conditions when an unusual example leads a model to make a highly inaccurate prediction.**

A bad prediction



- Linear regression model uses high weight for a bathroom
- Two apartments with a high number of bathrooms
- **Are participants, which can see the internals, able to spot the mistakes?**

Do people differ, if the model is „bad“?



- If people know when not to trust a model, we should see a larger deviation or higher bars for the clear models.
- Visibility has no impact

Possible Problem: New York City prices are exceptionally high.

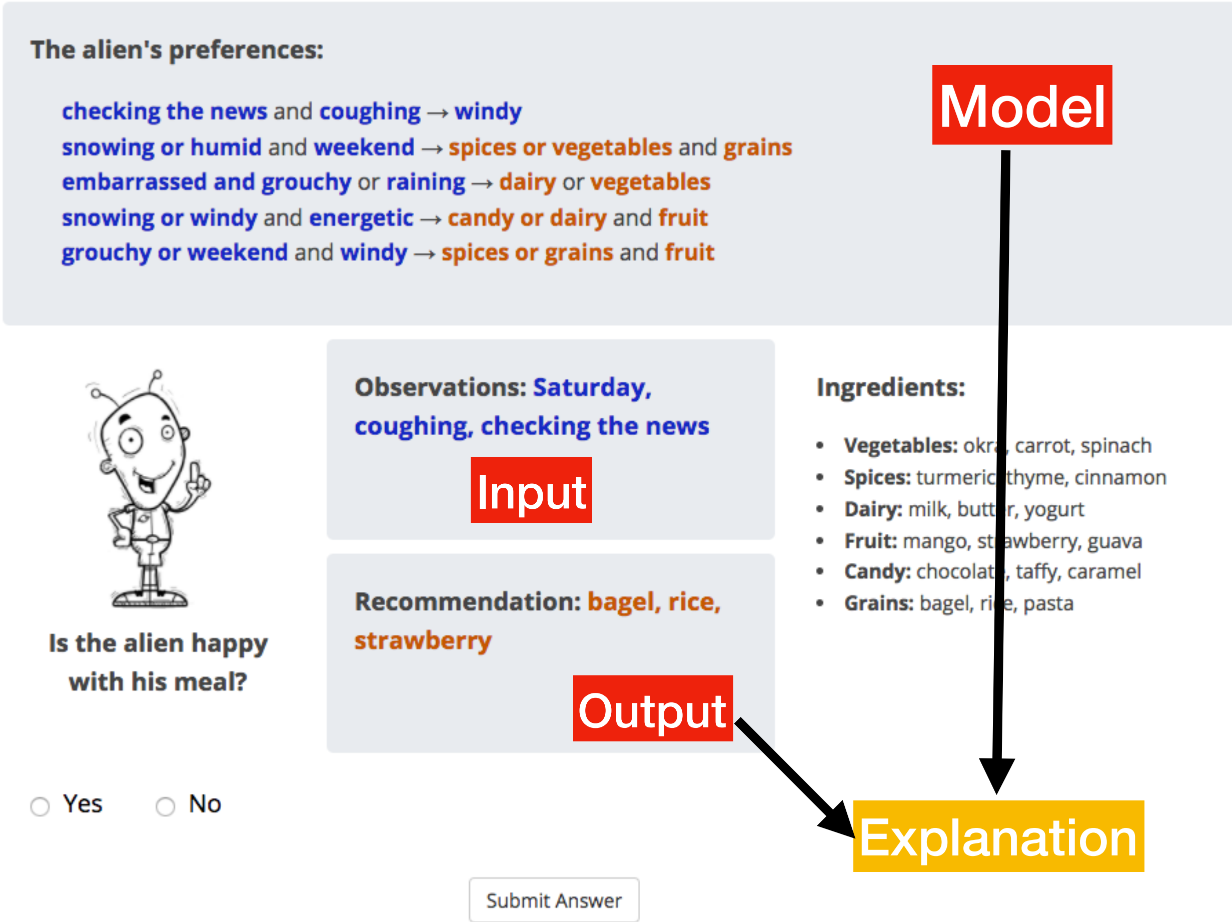
Summary of results

- Participants are better able to simulate the clear, 2-feature model compared with the black-box, 8 feature model.
- No difference in participants' deviation from the model across different conditions (New York prices).
- Transparent models do not help the users make better predictions
- When the model is wrong, participants in the clear conditions deviate less than those in black-box.

Goal of paper [2]: What kind of explanation are truly human interpretable and which are poorly understood?

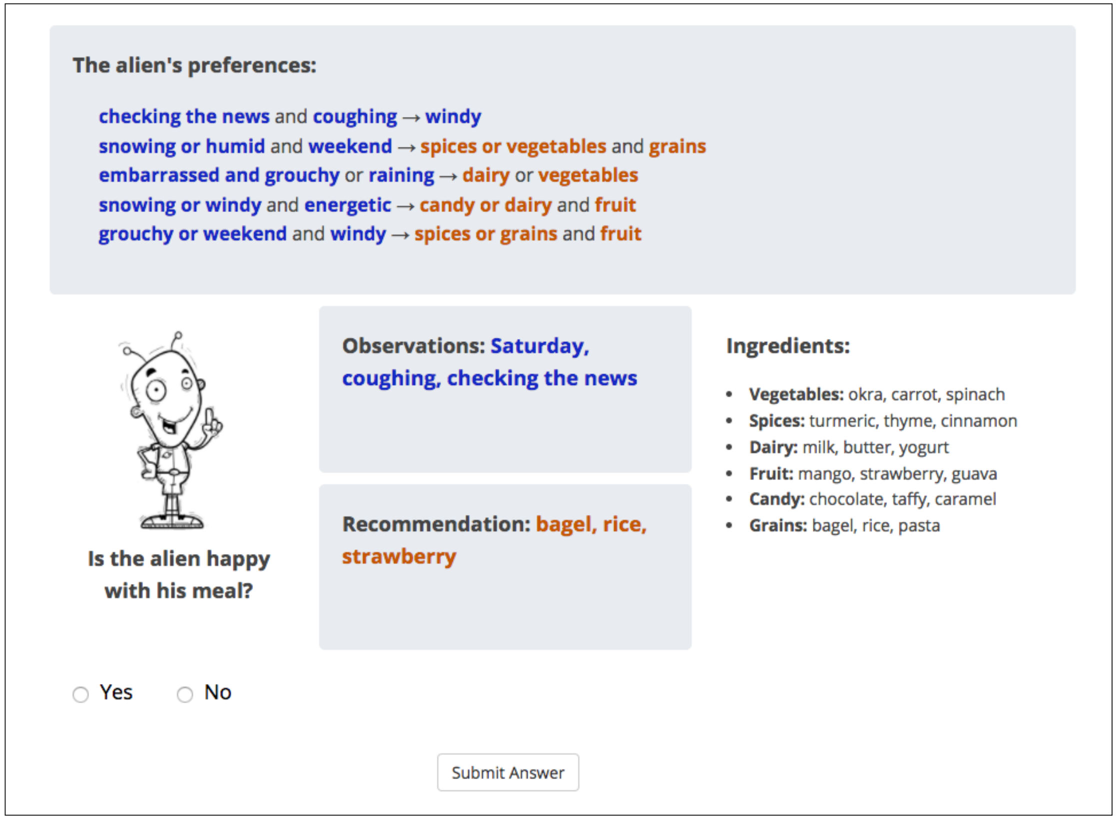
Experiment [2]

- 600 participants from Amazon Mechanical Turk
- Data was generated by humans
  (could be generated by a machine)
- **Variation of**
  - Explanation size
    (length of explanation and output)
  - New Types of Cognitive Chunks
  - Repeated Terms in an Explanation
  - Domain Variation (Recipe, Clinical)
- **Measurements taken**
  - Response time
  - Accuracy
  - Subjective satisfaction
    (rating of the explanation)



Hypotheses and Interface

- Increasing the size of the explanation either preferences or recommendations would increase the time to perform the task.
- Adding cognitive chunks increases the time required to process an explanation.
- If an input condition appeared in several lines of the explanation, it increases the time too find the correct rule.
- Similar results for the clinical domain.



Explicit vs. Implicit

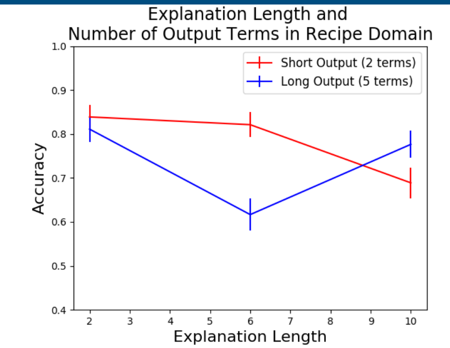
- Variations from explicit to implicit
- **Checking the news** and **coughing** -> **windy**
- **gouchy or weekend** and **windy**

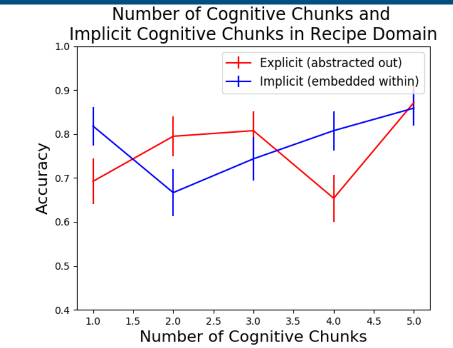
The alien's preferences:

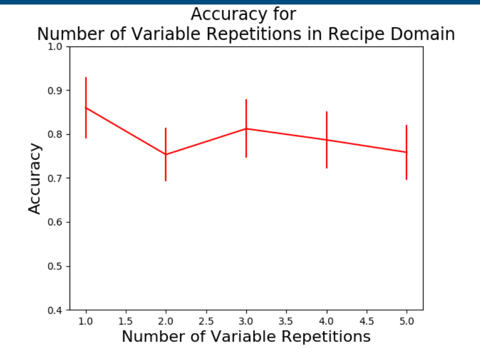
checking the news and **coughing** → **windy**
snowing or humid and **weekend** → **spices or vegetables** and **grains**
embarrassed and grouchy or **raining** → **dairy** or **vegetables**
snowing or windy and **energetic** → **candy or dairy** and **fruit**
grouchy or weekend and **windy** → **spices or grains** and **fruit**

	Explanation Size	New Cognitive Chunks	Variable Repetition

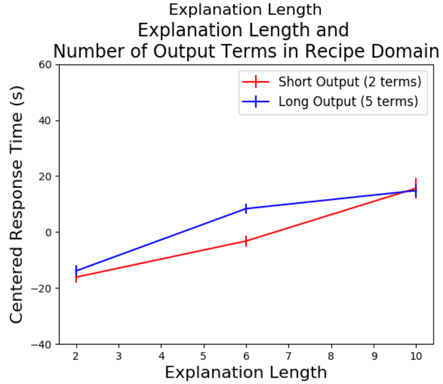
Accuracy

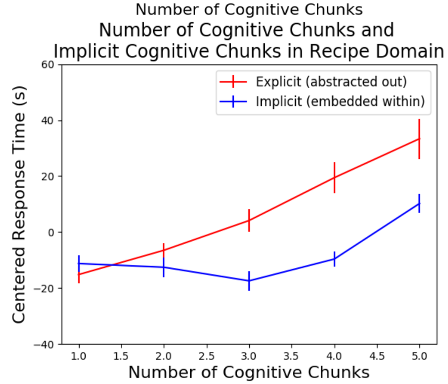


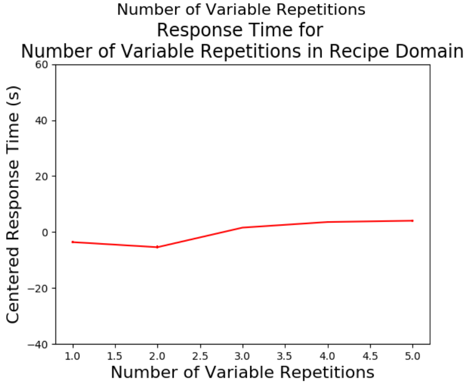




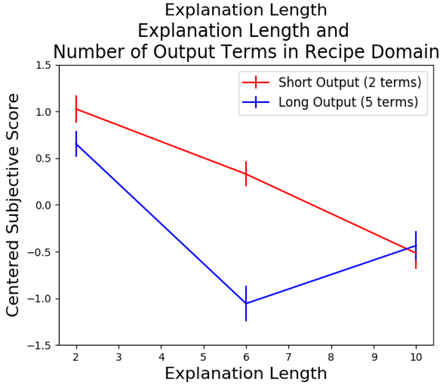
Response Time

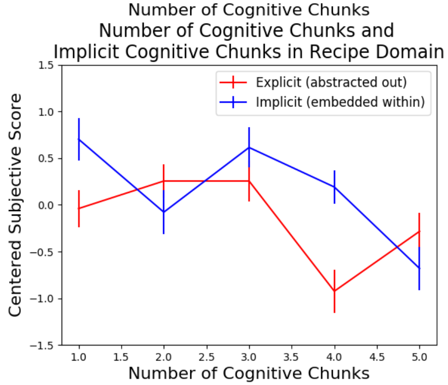


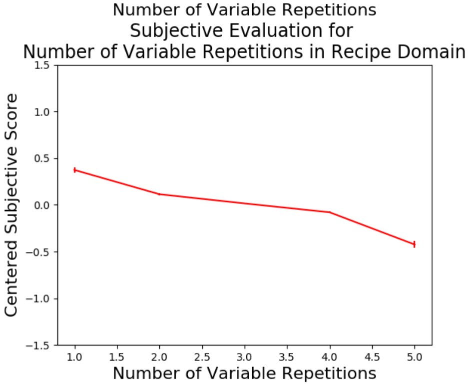




Satisfaction (Subjective evaluation)







Summary of results

- Increase in complexity increases response time.
- Increase in complexity and response time, less satisfaction.
- New Cognitive Chunks increase response time more than variable repetition.
- Response time increased, when new cognitive chunks were made explicit rather than implicit.

Conclusion/Future works

- Both Approaches: Identifying factors which affect ability to interpret machine learning models.
- What factors have the largest/smallest effect on interpretability?
- Recent publish papers, topic emerged in 2017 (also due to GDPR).
- Some values taken from the „system design“, some from the „humans behavior“, more values to be evaluated.
- Focus only on lay people, no specific group (e.g. regulators).
- User biased due to mechanical turk?
- What kind of explanation are best in what context? (Decision tree, Pseudocode)
  Different approaches need to be tested.

Thank you!