A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models

[work in progress]

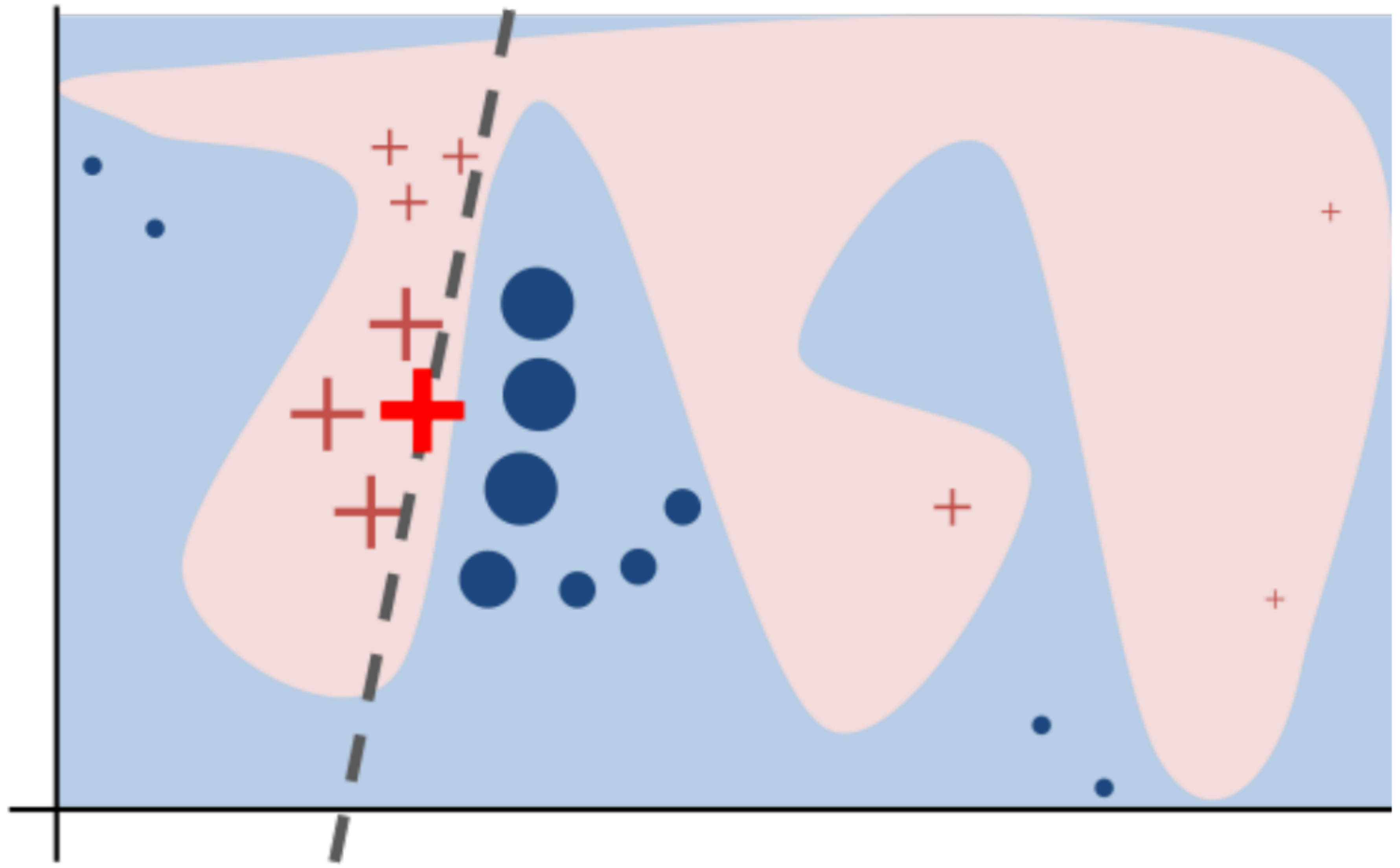
Josua Krause*, Adam Perer**, Enrico Bertini*

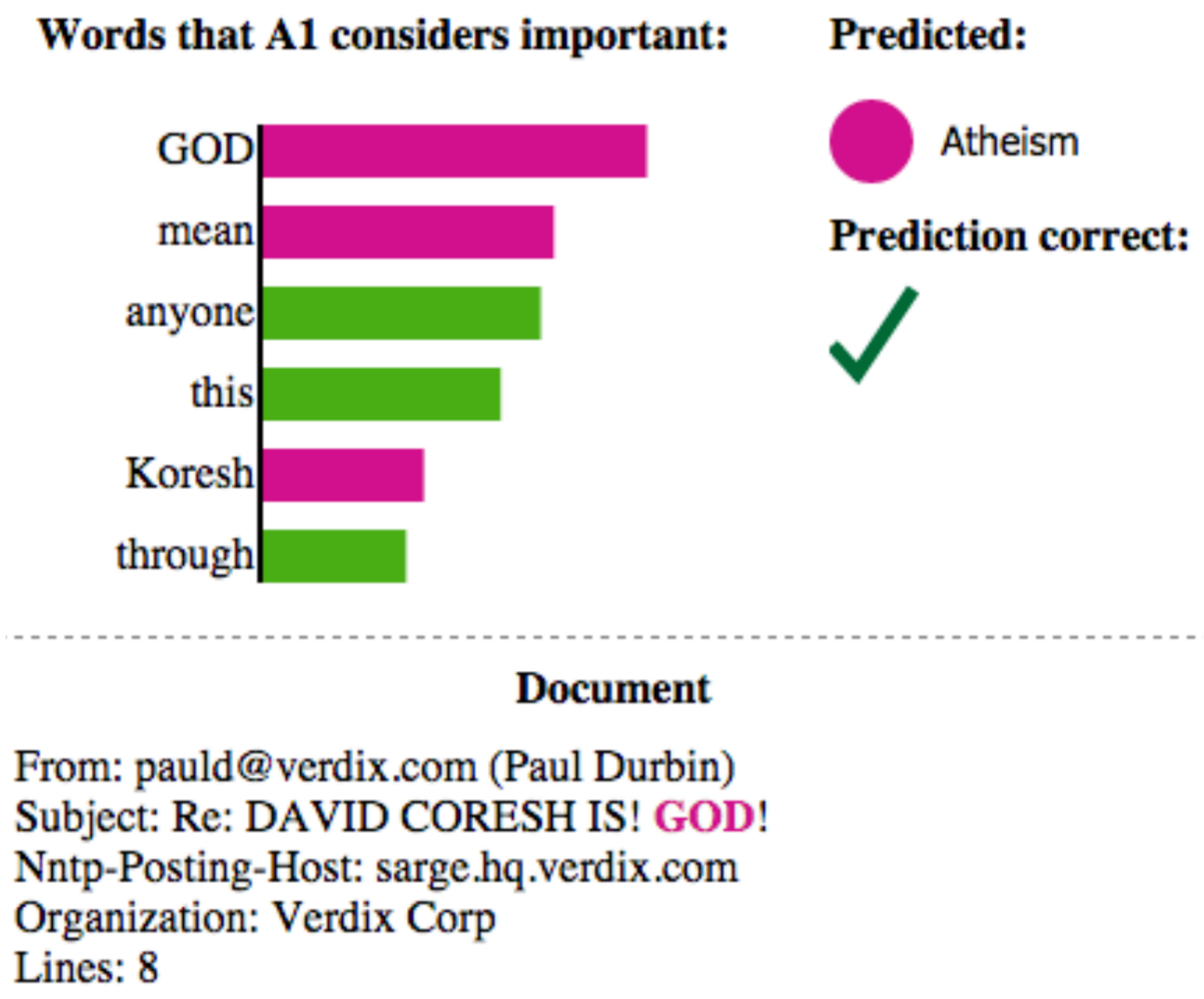
Mon, August 20th 2018

* NYU TANDON SCHOOL OF ENGINEERING

** IBM

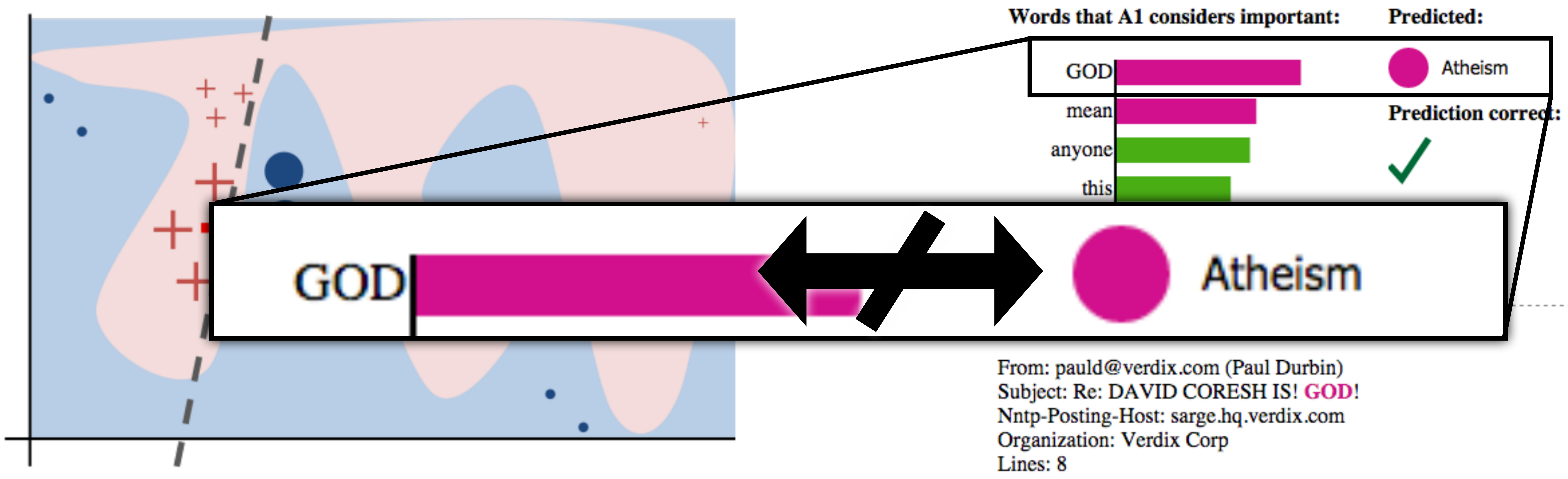
Instance Explanations





"Why Should I Trust You?" Explaining the Predictions of Any Classifier
Marco Riberio, Sameer Singh, Carlos Guestrin
International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2016)

Finding Data Biases



"Why Should I Trust You?" Explaining the Predictions of Any Classifier
Marco Riberio, Sameer Singh, Carlos Guestrin
International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2016)

Problem:

Inspecting single instances
does not scale well

Solution:

Aggregating data and explanations

Ground Truth	Positive	vs.		Negative	
Prediction	Positive	vs.		Negative	
	Correct	vs.		Incorrect	

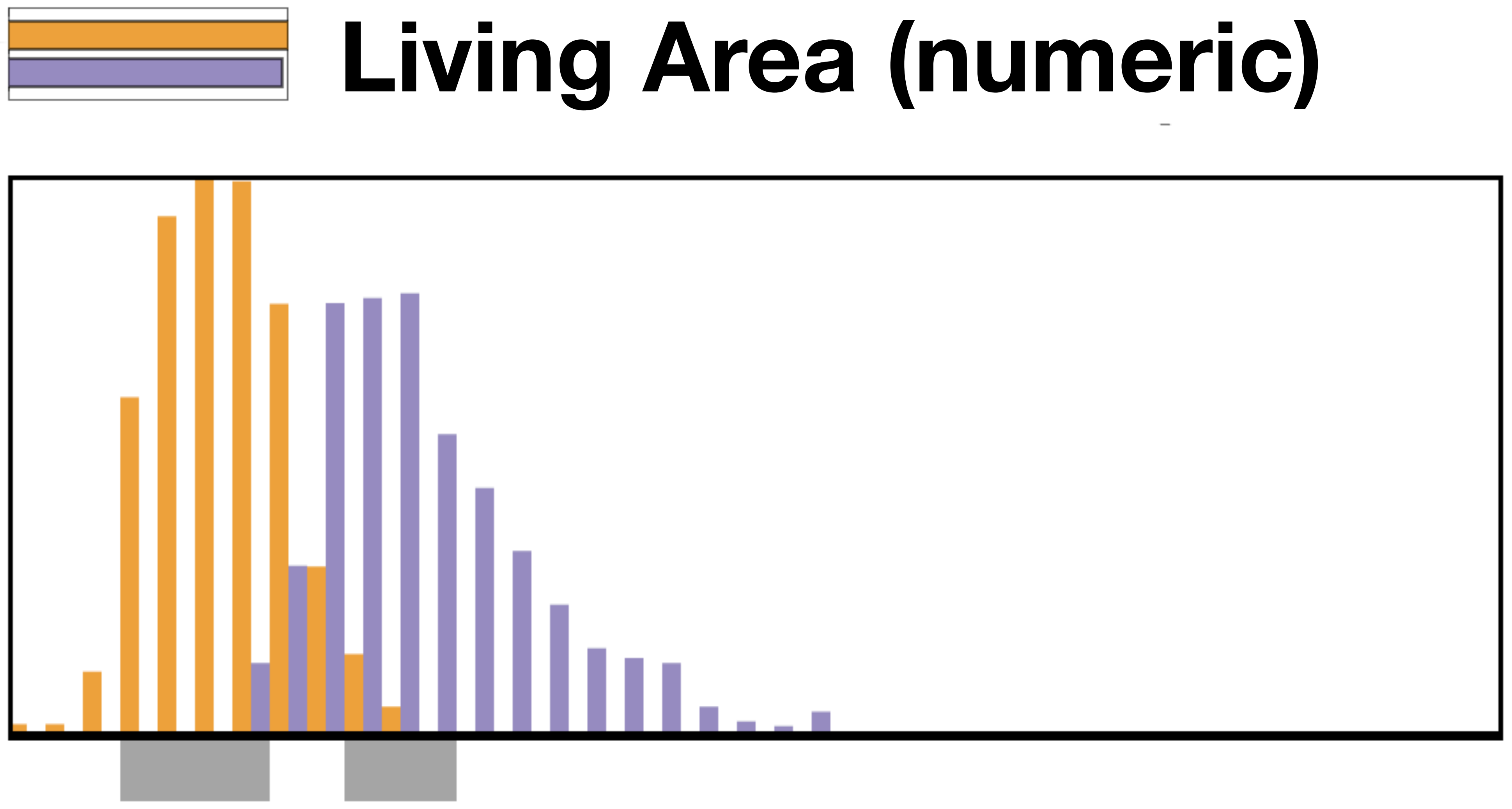
Solution:

Aggregating data and explanations

Ground Truth	**Positive**	vs.	**Negative**
Prediction	**Positive**	vs.	**Negative**
	Correct	vs.	**Incorrect**

Solution:

Aggregating data and explanations

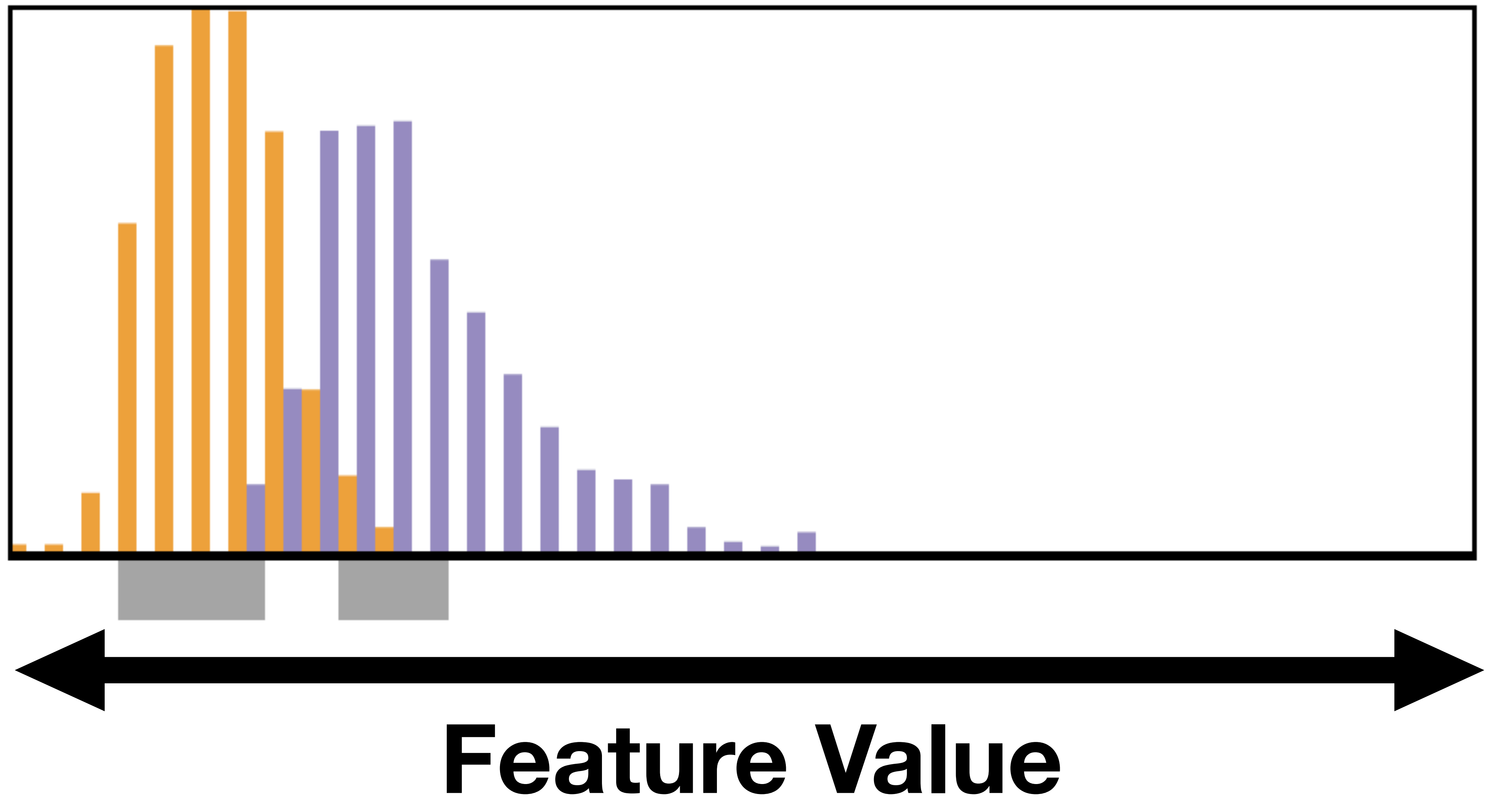


Ground Truth Positive vs. Negative

Prediction Positive vs. Negative

Correct vs. Incorrect

Living Area (numeric)



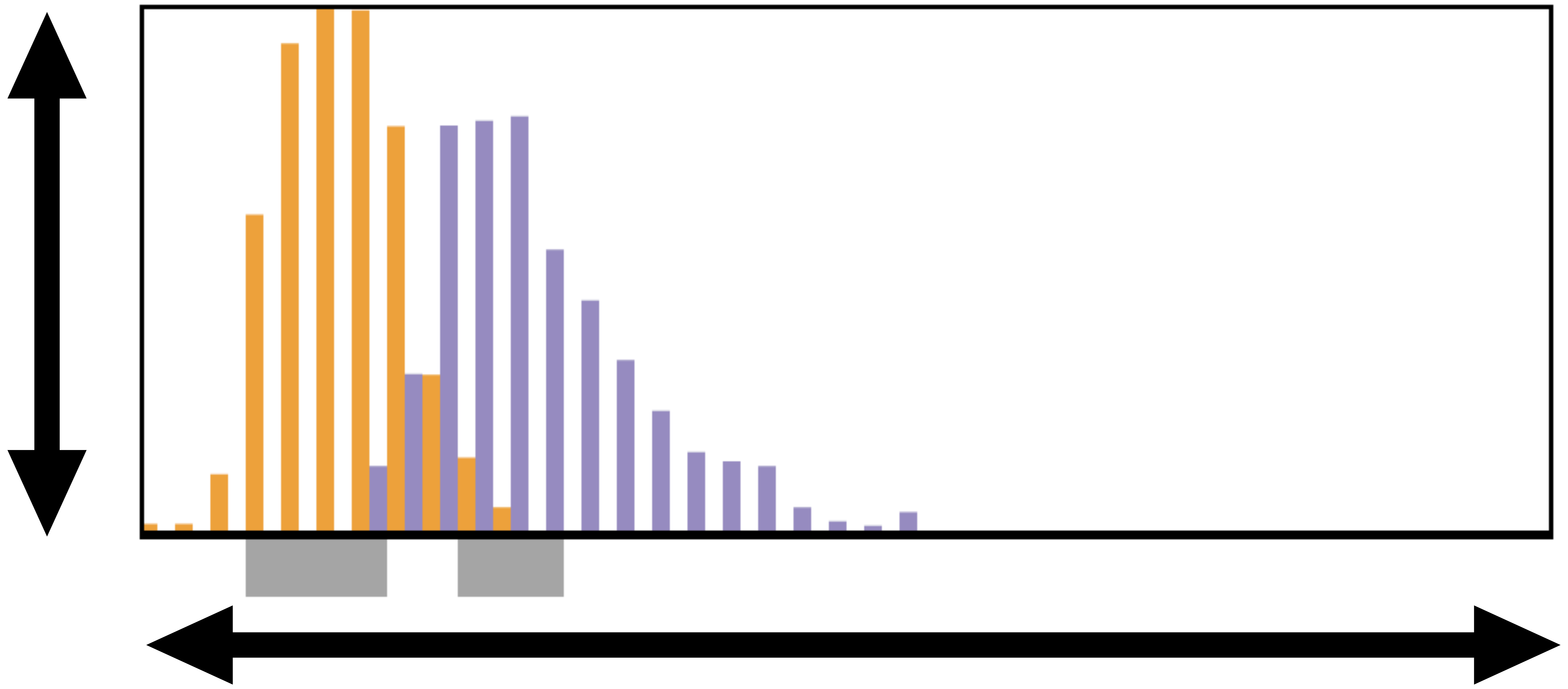
Ground Truth Positive vs. Negative

Prediction Positive vs. Negative

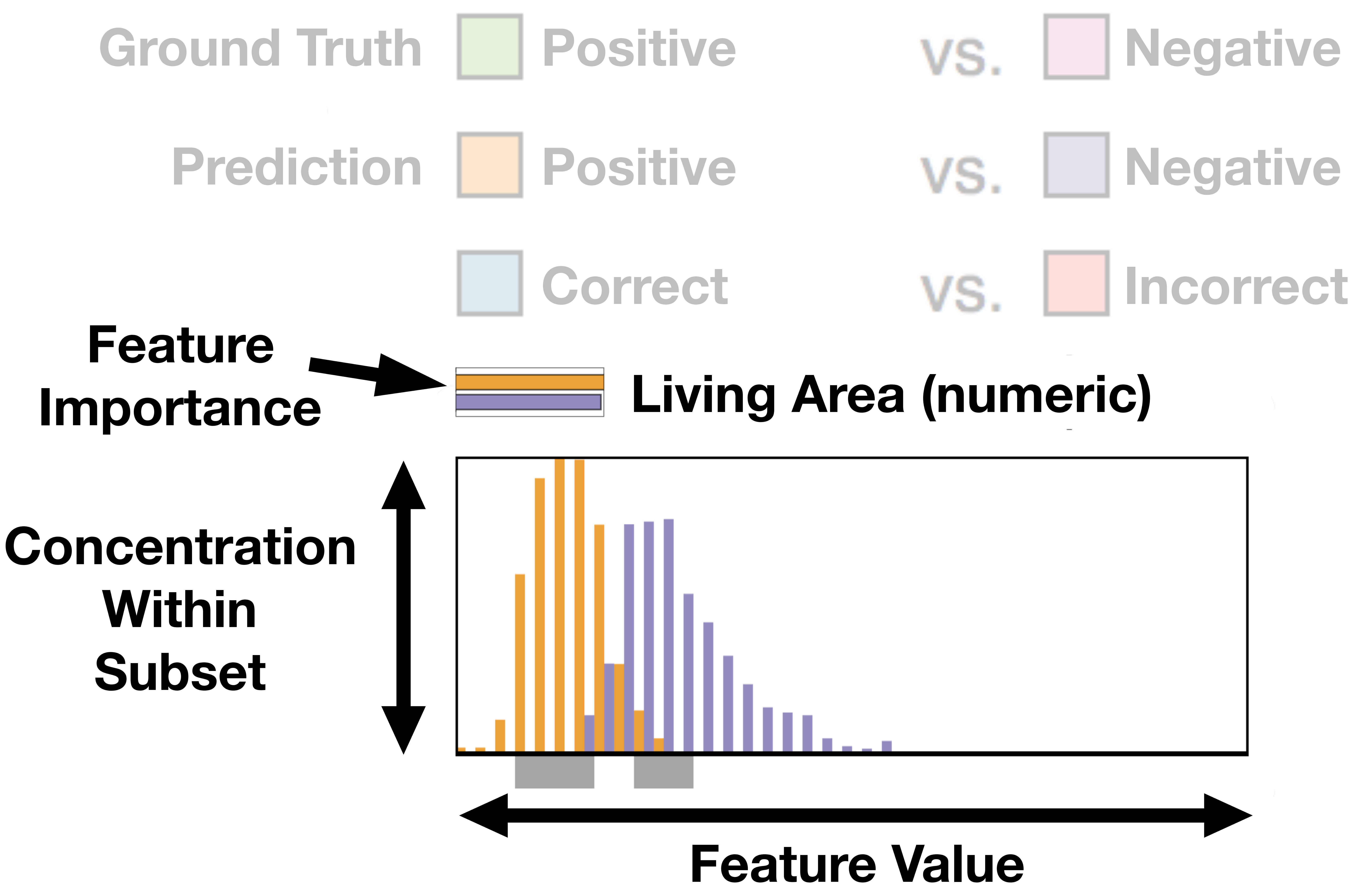
Correct vs. Incorrect

Living Area (numeric)

Concentration Within Subset



Feature Value



Ground Truth
Positive
vs.
Negative
Prediction
Positive
vs.
Negative
Correct
vs.
Incorrect
Feature Importance
Living Area (numeric)
Concentration Within Subset
Feature Value


Sorted by
Importance
Living Area (nu…
Room Count (n…
Neighborhood …
Overall Quality…
Overall Conditi…
Foundation (cat)
House Style (c…
Month Sold (nu…
Garage (cat)

What is the impact of aggregation?

What is the impact of instance-level explanations?

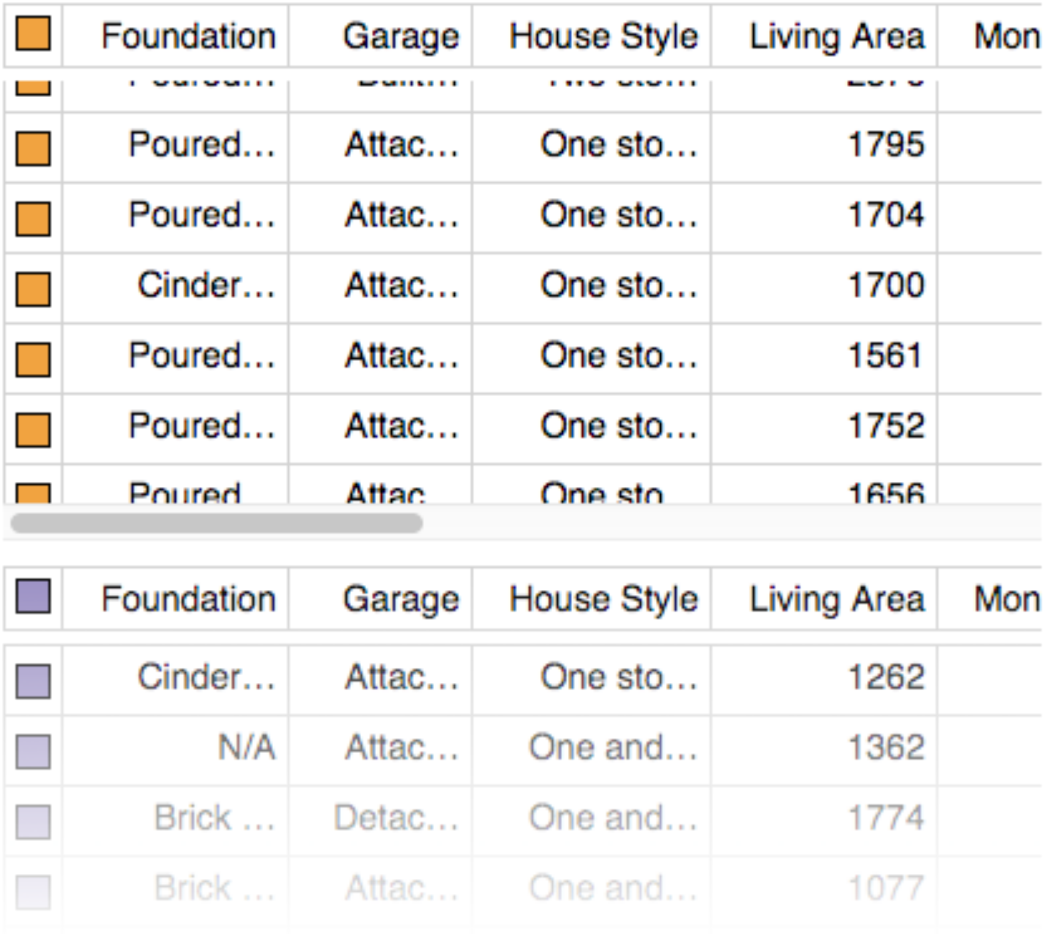
How do those settings affect the ability to detect biases in the data?

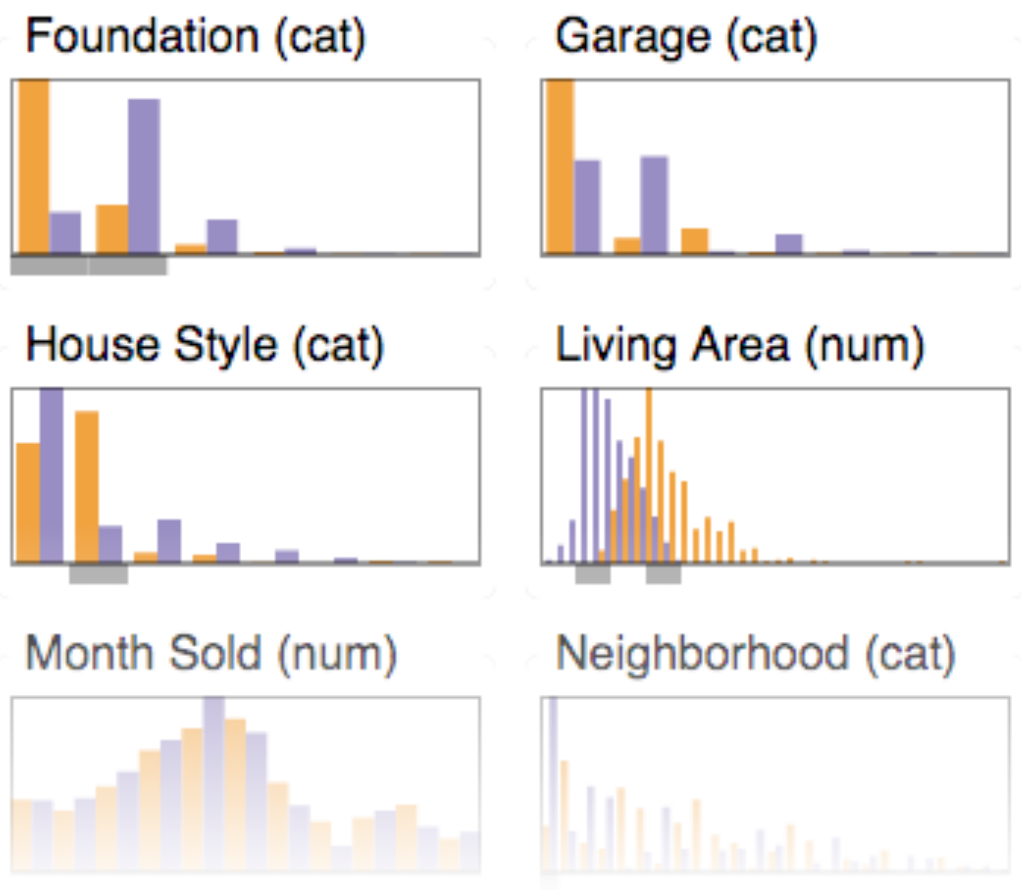
Four Conditions

Table

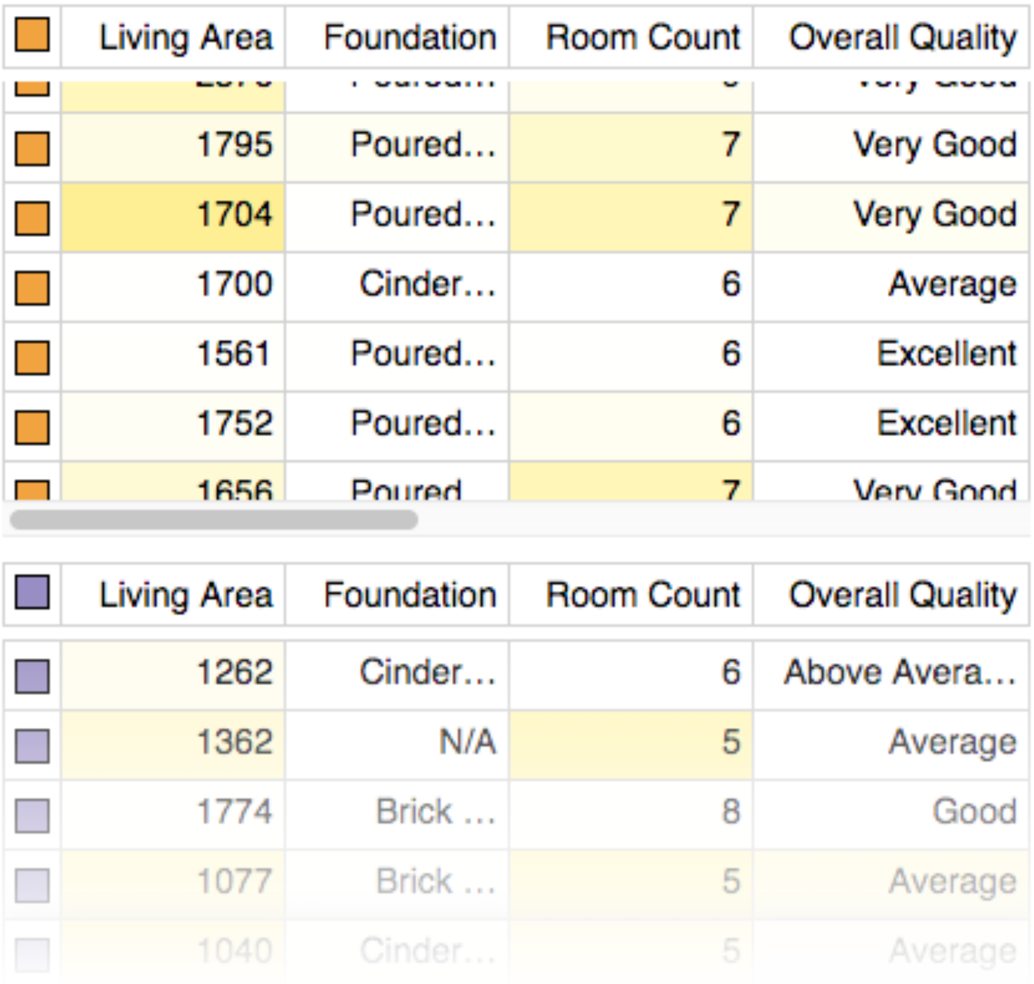
Histogram

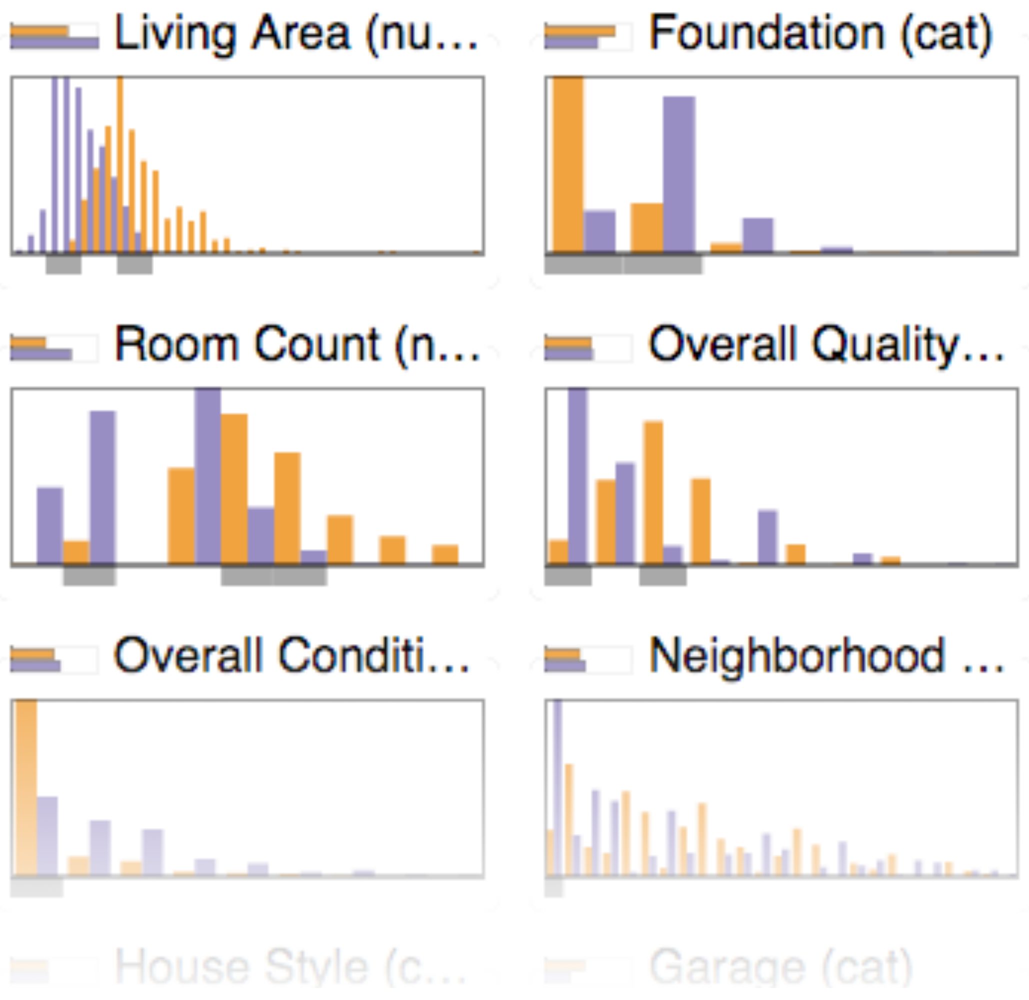
No Explanation





Explanation



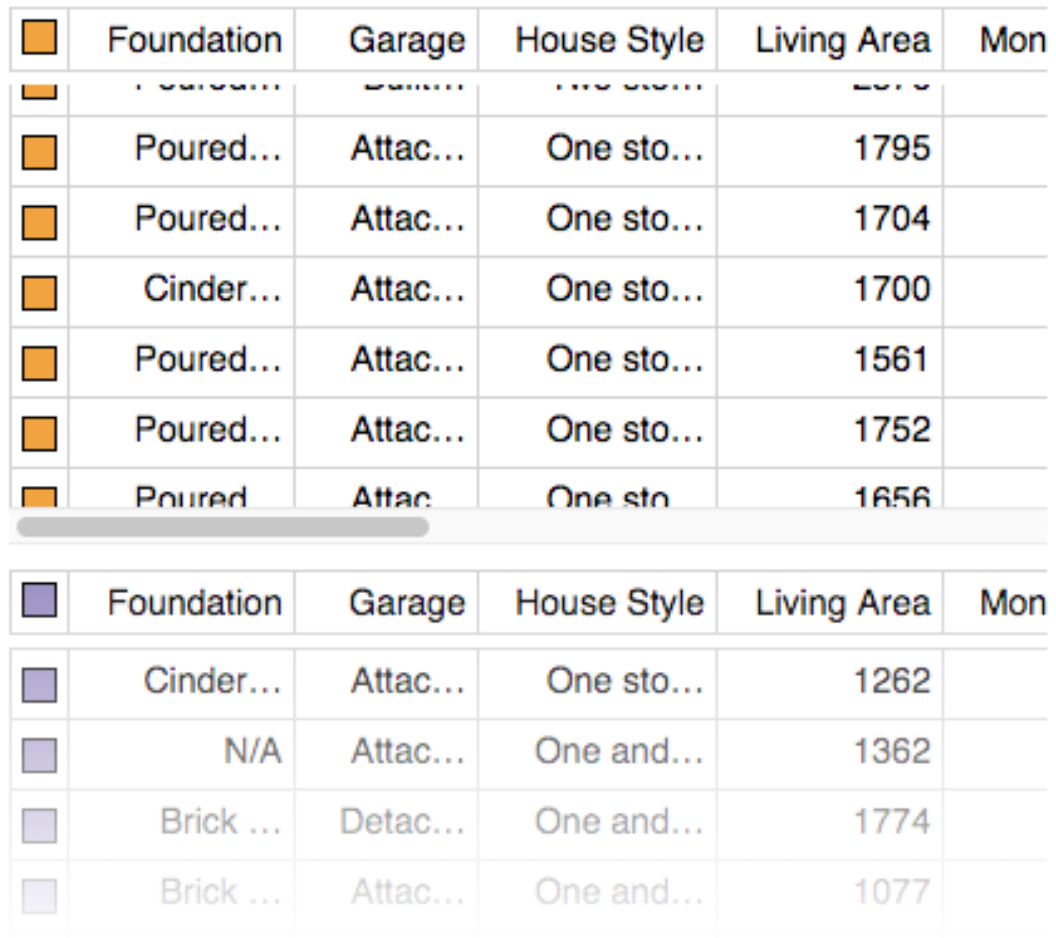


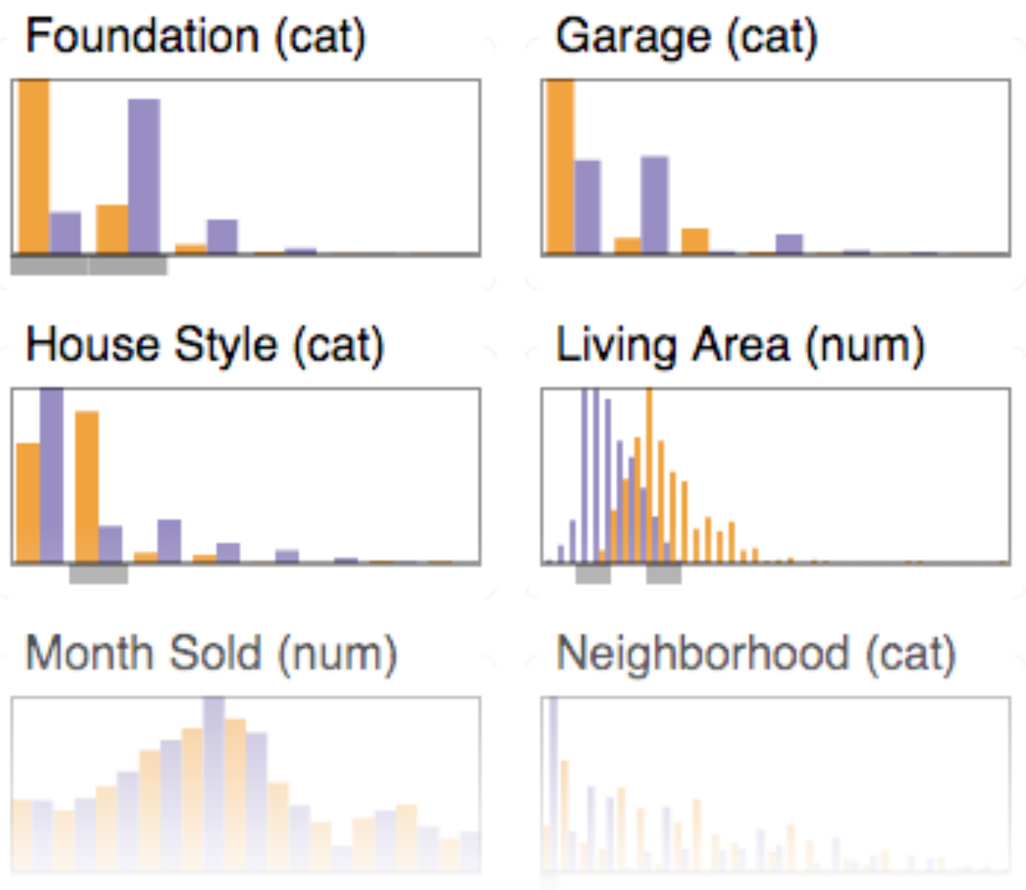
Four Conditions

Table

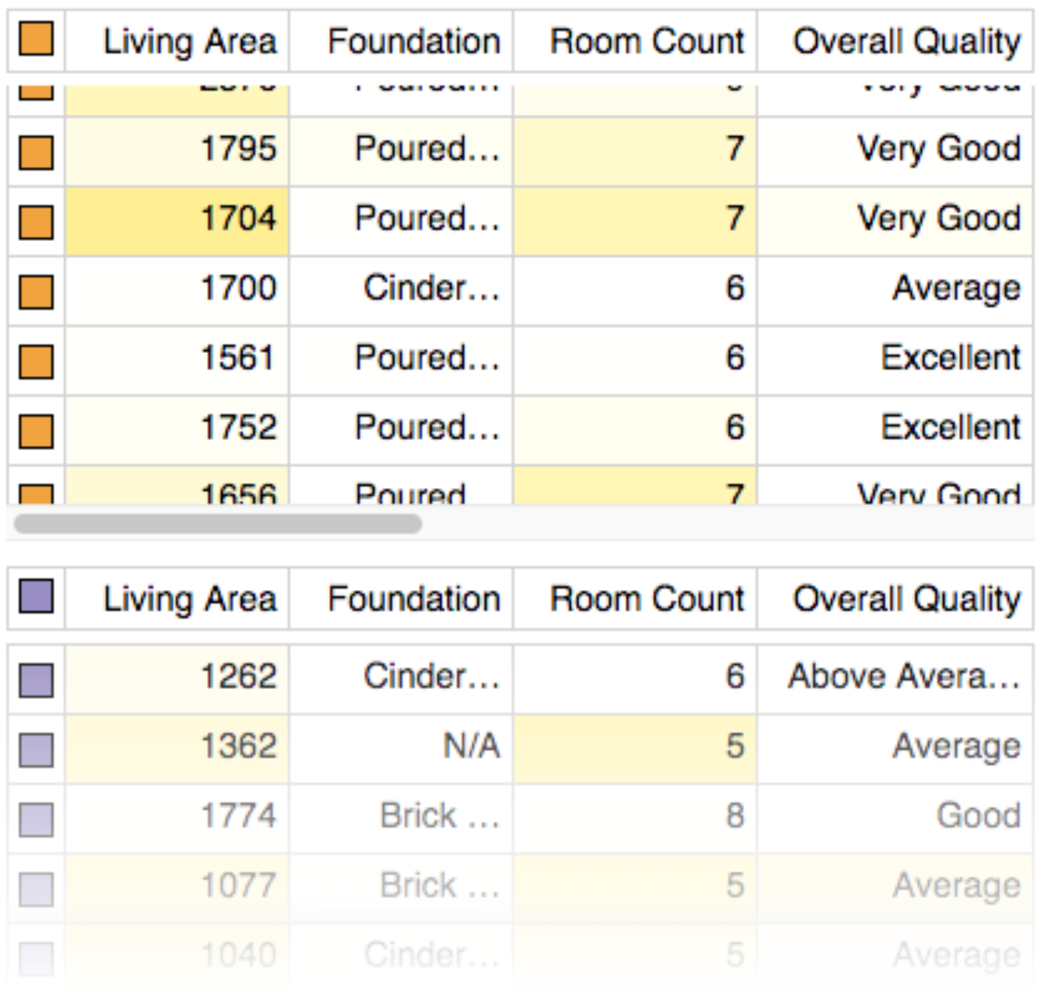
Histogram

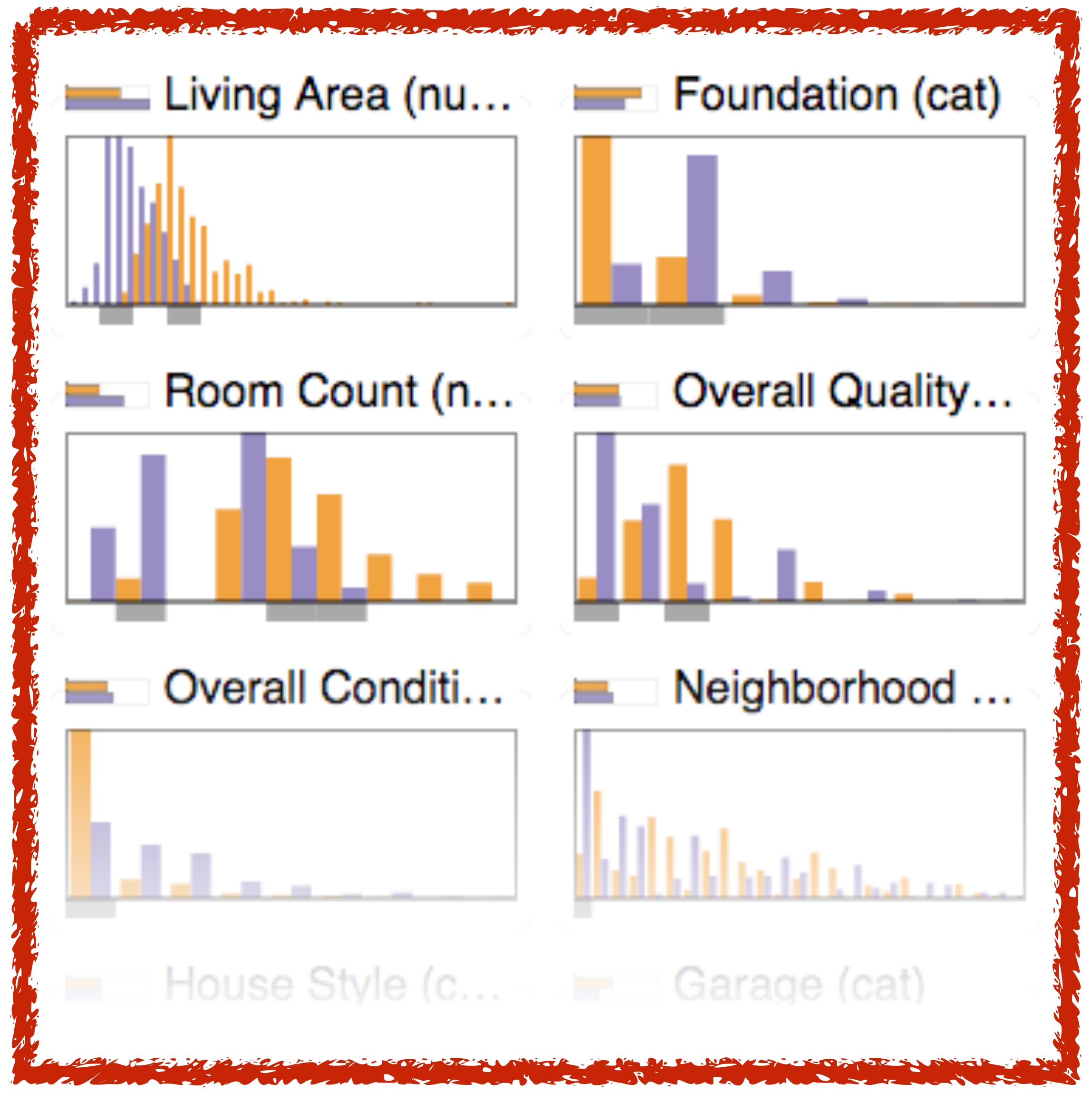
No Explanation





Explanation





Four Conditions

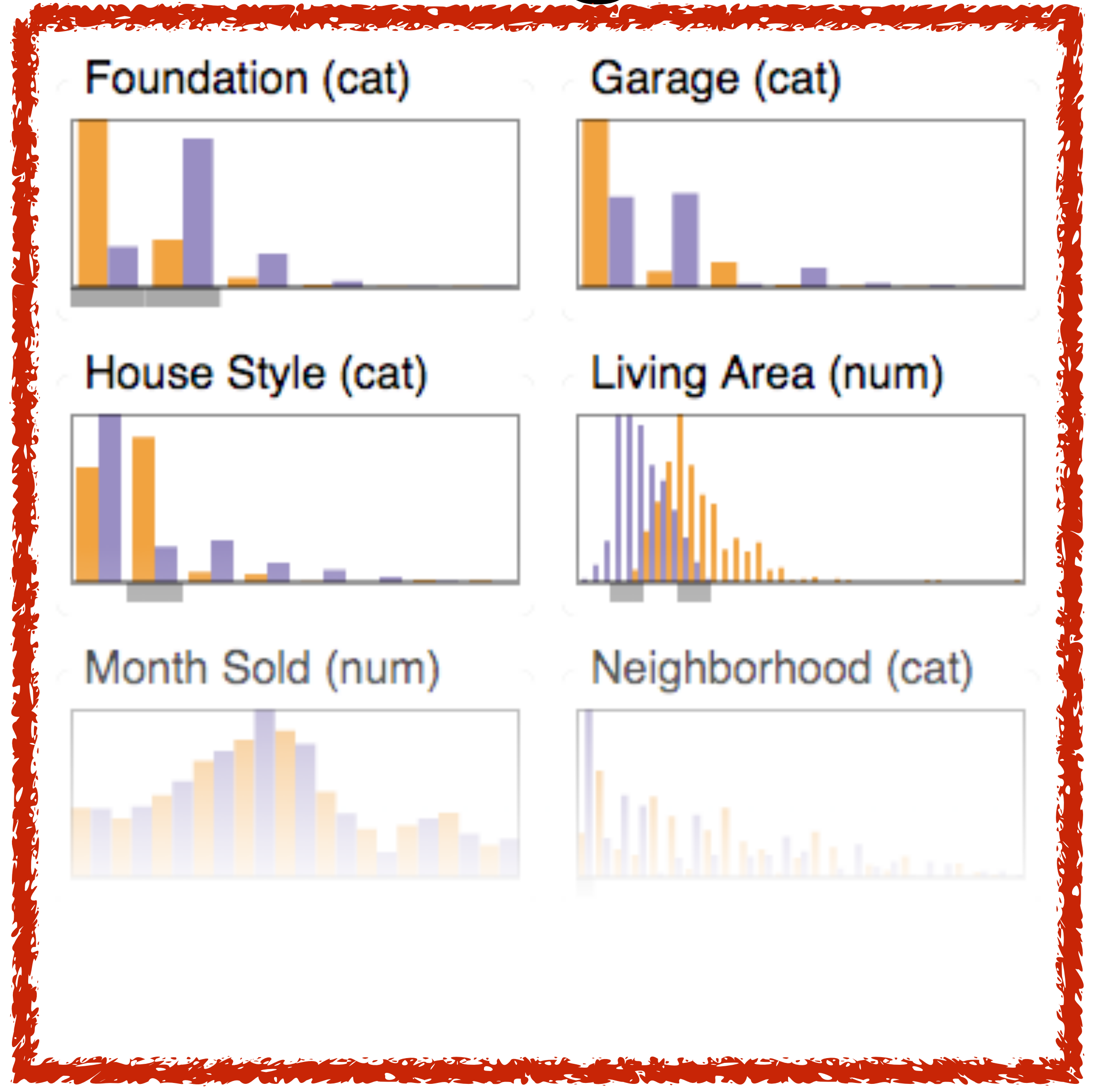
Table

Histogram

No Explanation

	Foundation	Garage	House Style	Living Area	Mon
	Poured...	Attac...	One sto...	1795	
	Poured...	Attac...	One sto...	1704	
	Cinder...	Attac...	One sto...	1700	
	Poured...	Attac...	One sto...	1561	
	Poured...	Attac...	One sto...	1752	
	Poured	Attac	One sto	1656	

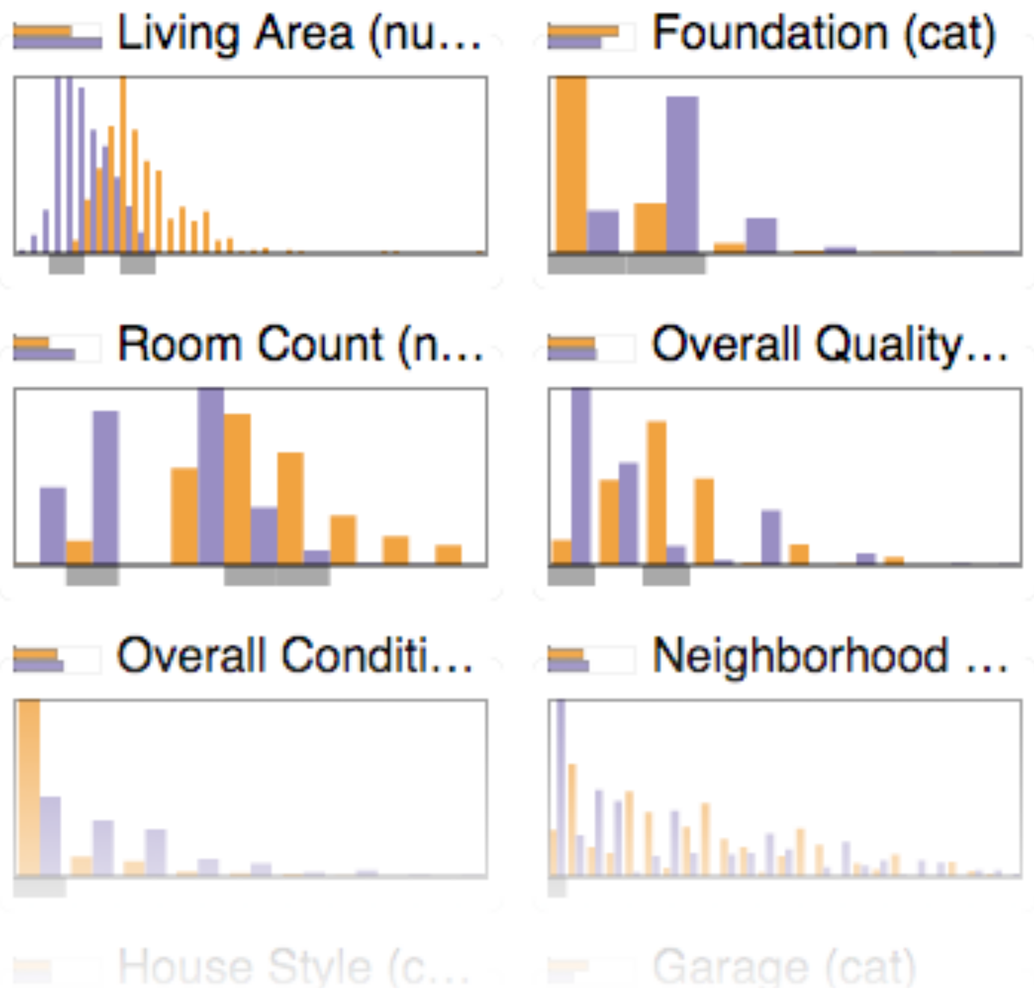
	Foundation	Garage	House Style	Living Area	Mon
	Cinder...	Attac...	One sto...	1262	
	N/A	Attac...	One and...	1362	
	Brick ...	Detac...	One and...	1774	
	Brick ...	Attac...	One and...	1077	



Explanation

	Living Area	Foundation	Room Count	Overall Quality
	1795	Poured...	7	Very Good
	1704	Poured...	7	Very Good
	1700	Cinder...	6	Average
	1561	Poured...	6	Excellent
	1752	Poured...	6	Excellent
	1656	Poured	7	Very Good

	Living Area	Foundation	Room Count	Overall Quality
	1262	Cinder...	6	Above Avera...
	1362	N/A	5	Average
	1774	Brick ...	8	Good
	1077	Brick ...	5	Average
	1040	Cinder...	5	Average

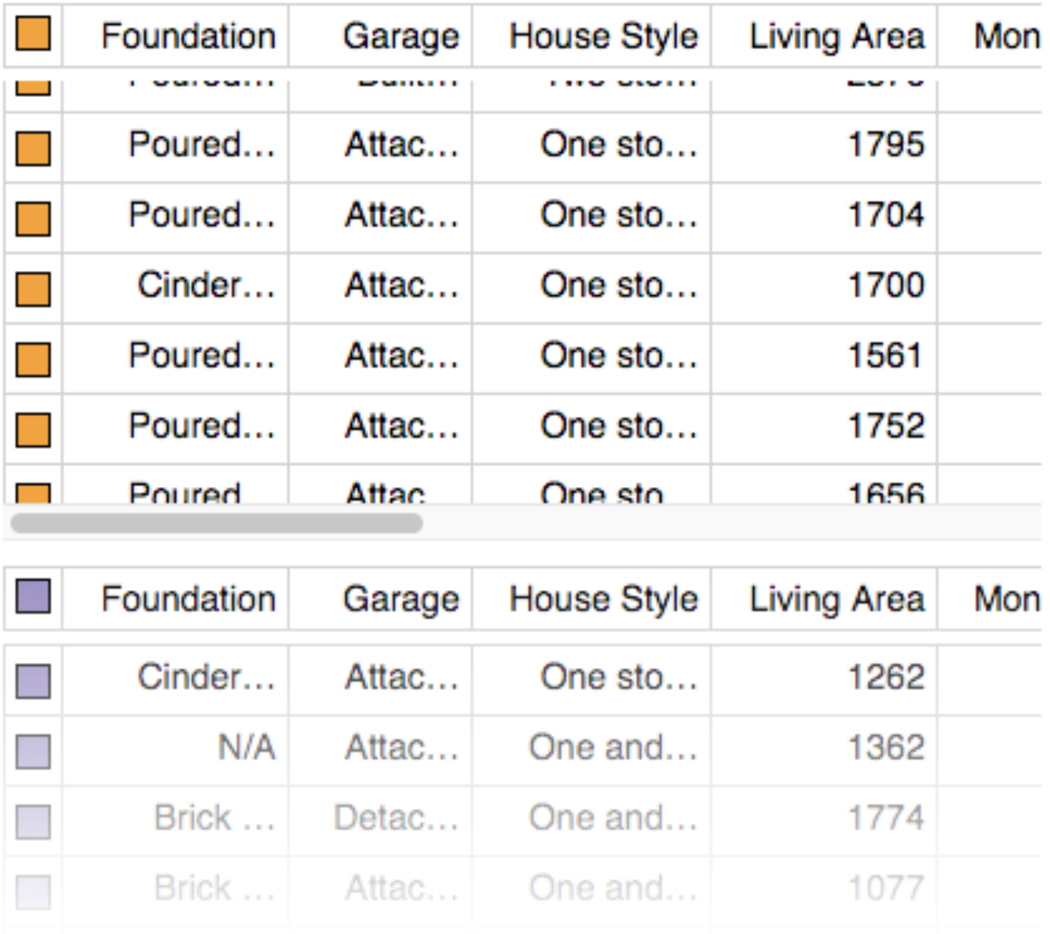


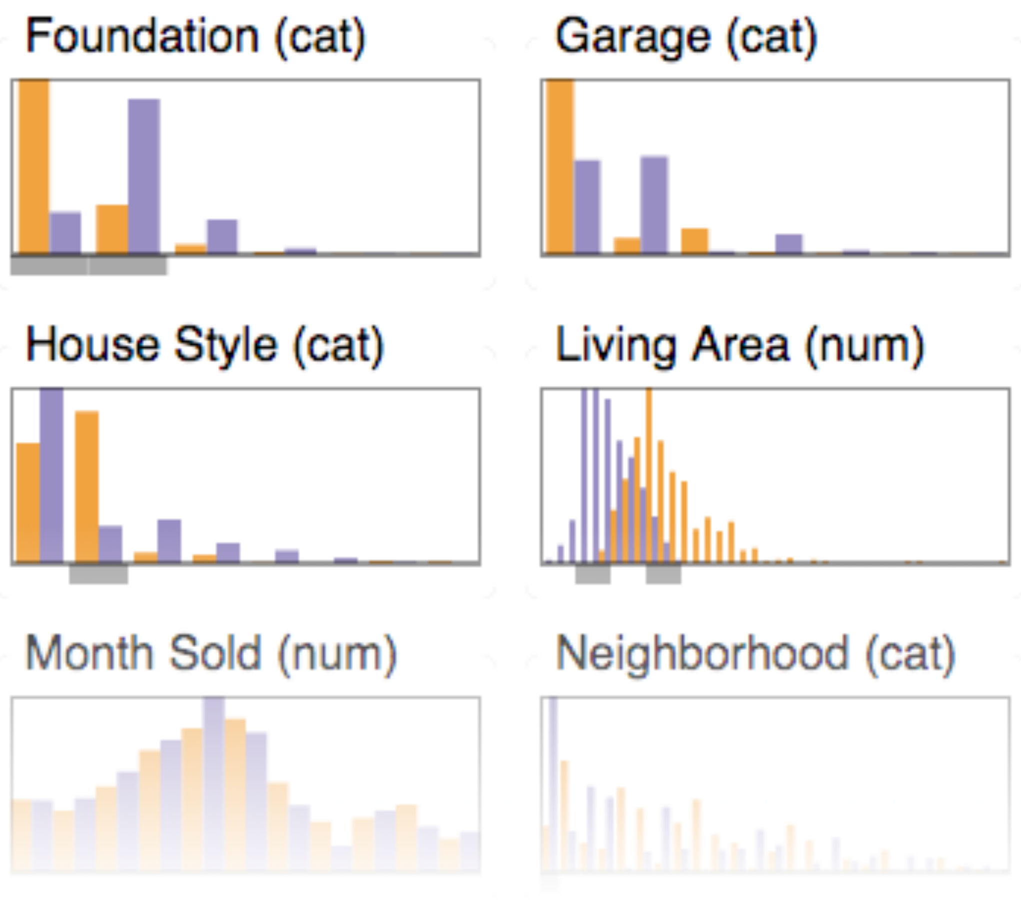
Four Conditions

Table

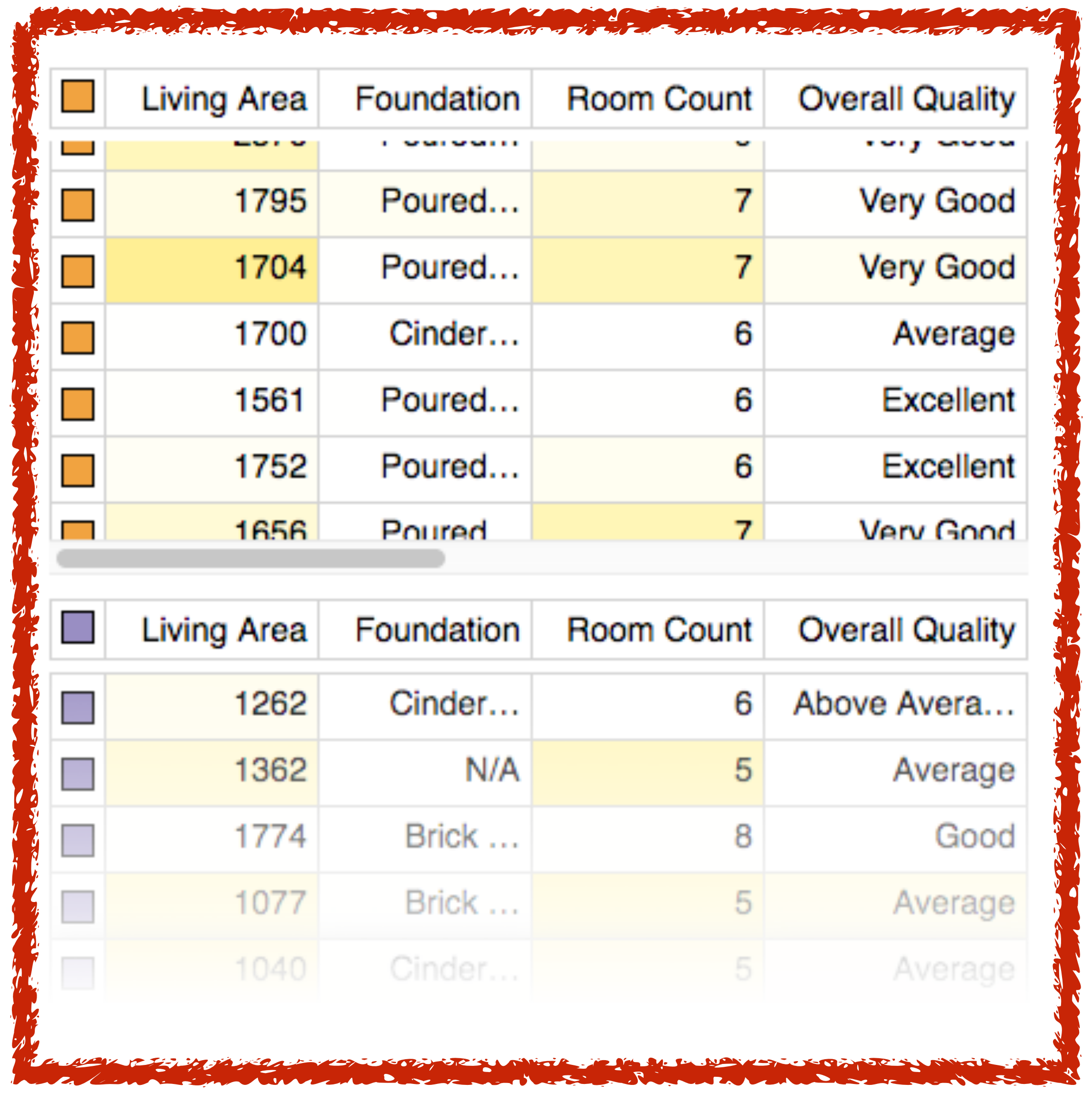
Histogram

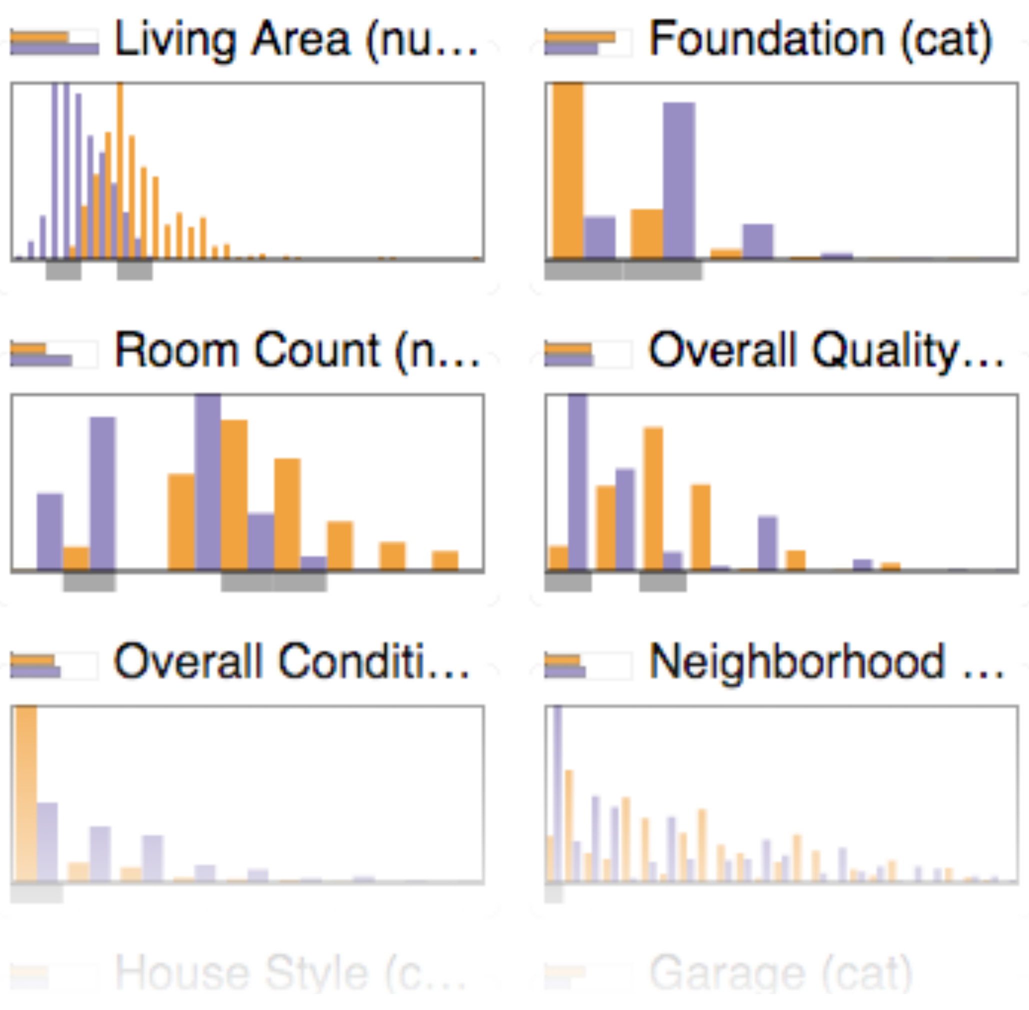
No Explanation





Explanation



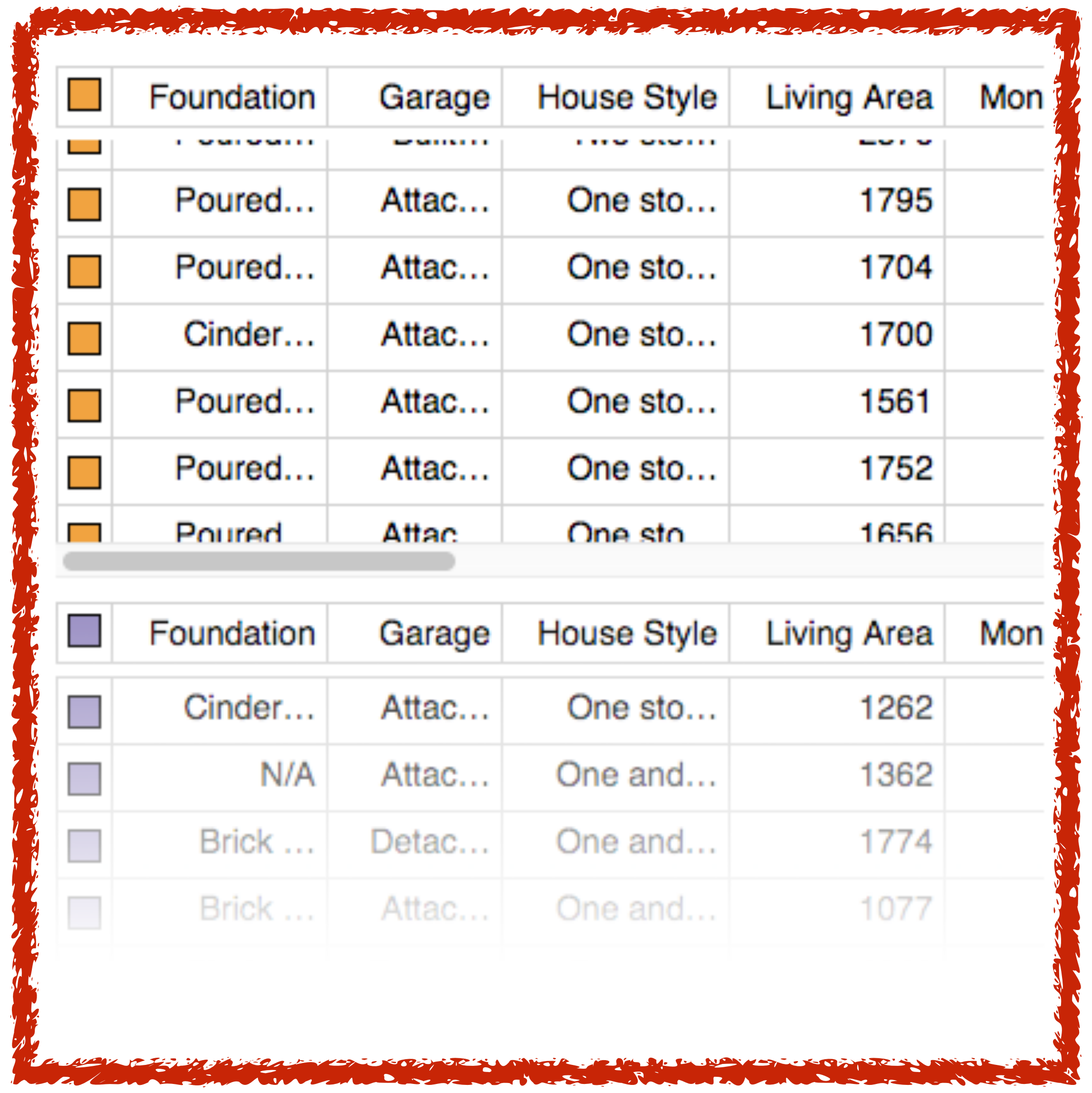


Four Conditions

Table

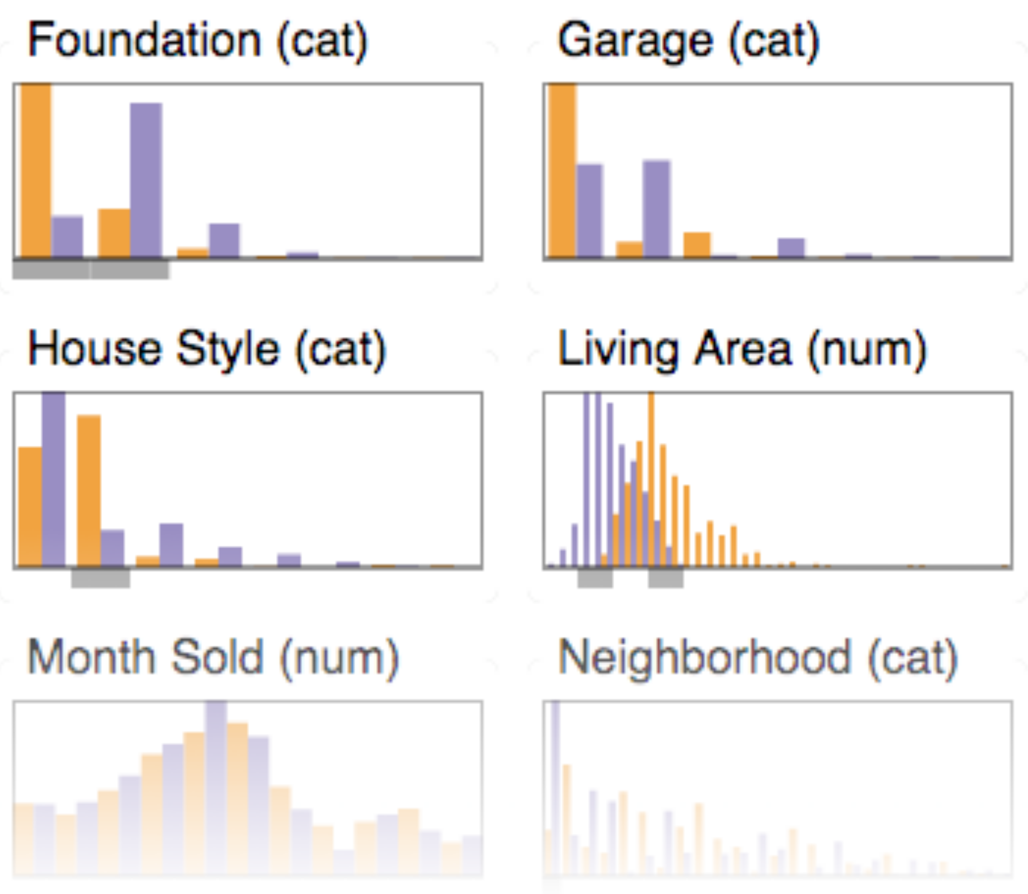
Histogram

No Explanation

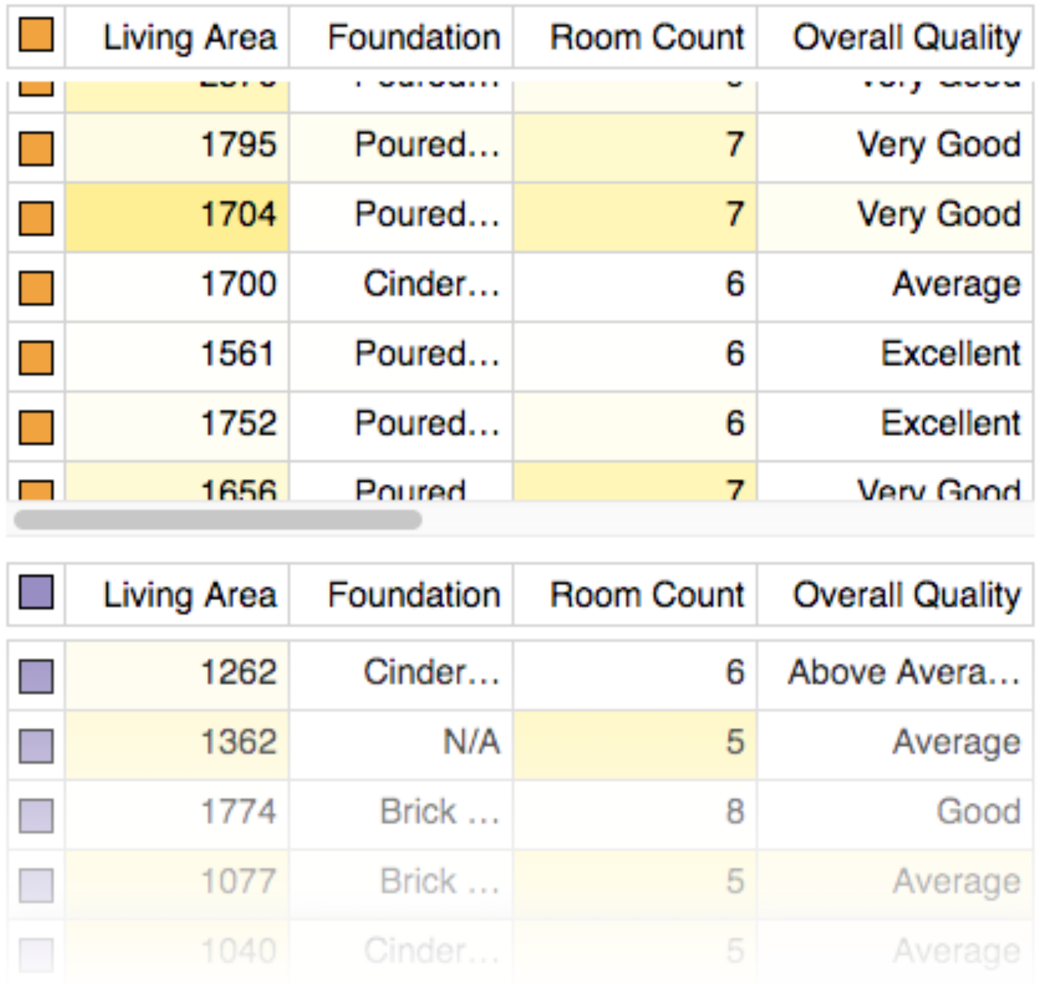


	Foundation	Garage	House Style	Living Area	Mon
	Poured…	Attac…	One sto…	1795	
	Poured…	Attac…	One sto…	1704	
	Cinder…	Attac…	One sto…	1700	
	Poured…	Attac…	One sto…	1561	
	Poured…	Attac…	One sto…	1752	
	Poured	Attac	One sto	1656	

	Foundation	Garage	House Style	Living Area	Mon
	Cinder…	Attac…	One sto…	1262	
	N/A	Attac…	One and…	1362	
	Brick …	Detac…	One and…	1774	
	Brick …	Attac…	One and…	1077	

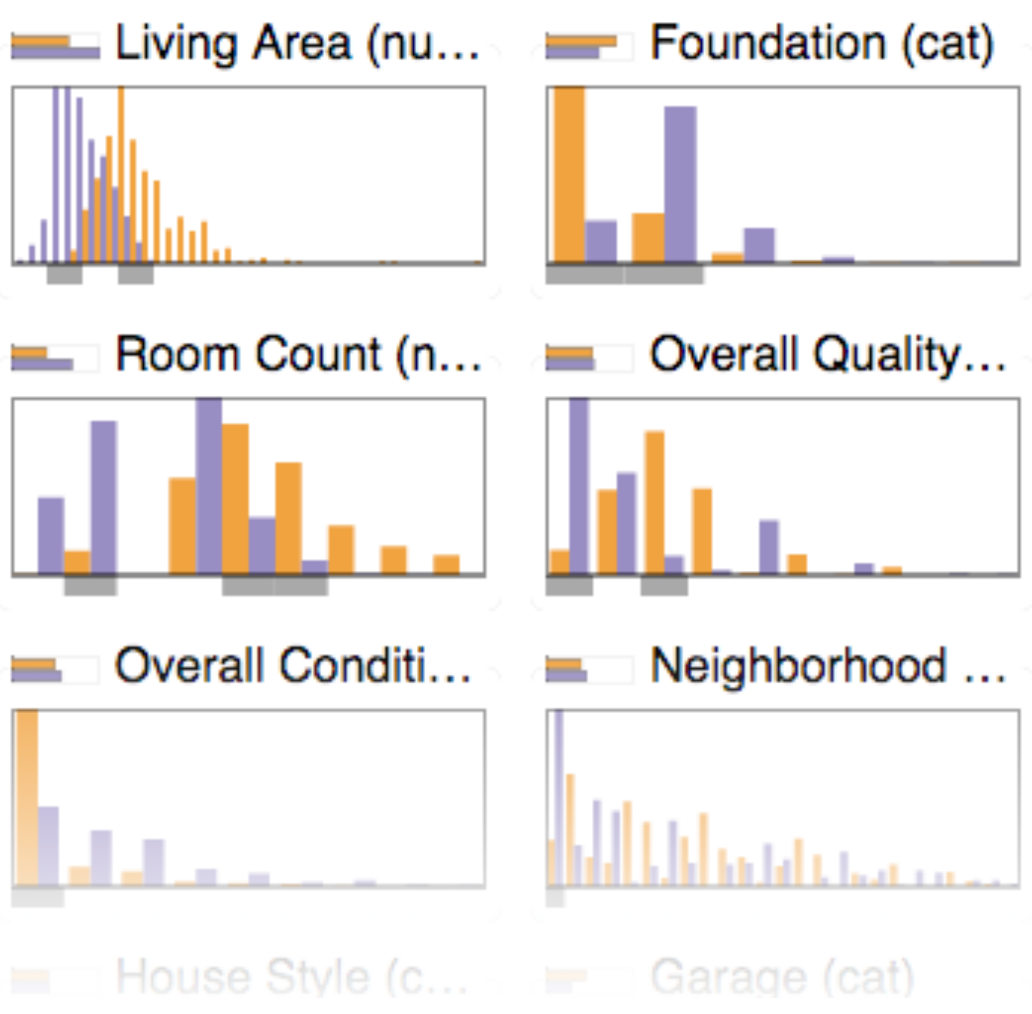


Explanation



	Living Area	Foundation	Room Count	Overall Quality
	1795	Poured…	7	Very Good
	1704	Poured…	7	Very Good
	1700	Cinder…	6	Average
	1561	Poured…	6	Excellent
	1752	Poured…	6	Excellent
	1656	Poured	7	Very Good

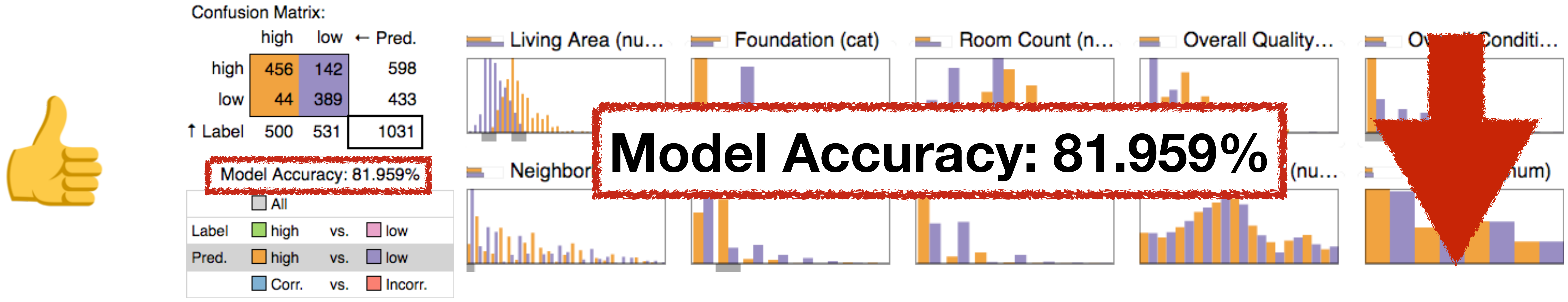
	Living Area	Foundation	Room Count	Overall Quality
	1262	Cinder…	6	Above Avera…
	1362	N/A	5	Average
	1774	Brick …	8	Good
	1077	Brick …	5	Average
	1040	Cinder…	5	Average

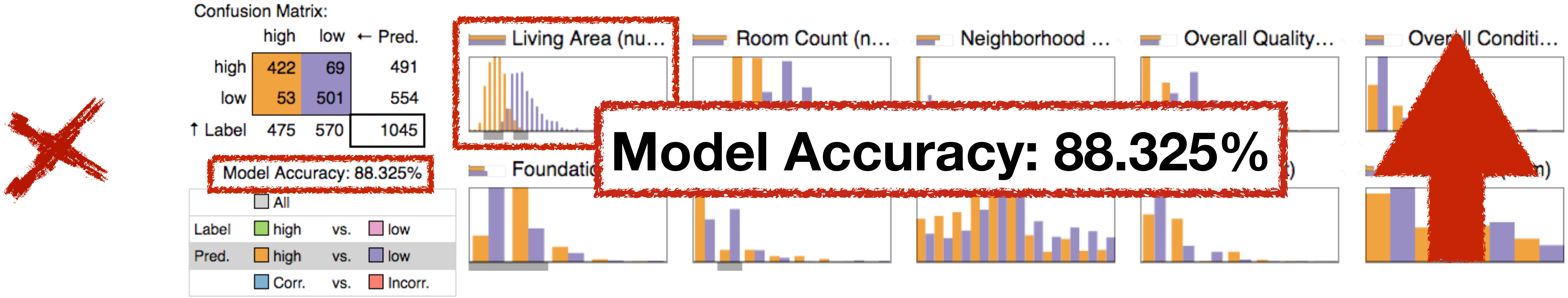


Two Data Sets

Two Data Sets

Two Data Sets





Questions

Individual models:

- Do you think the predictions of the model **make sense**?

  5 point Likert scale (Not at all – Very much)
- How well does the model perform in terms of **accuracy**?

  5 point Likert scale (Not much – Very well)
- How much do you **trust** the model?

  5 point Likert scale (Not at all – Very much)
- Why do you trust or not trust this model?

  Free text answer

Summary:

Which model do you prefer?

Multiple choice and text answer

Study

100 participants

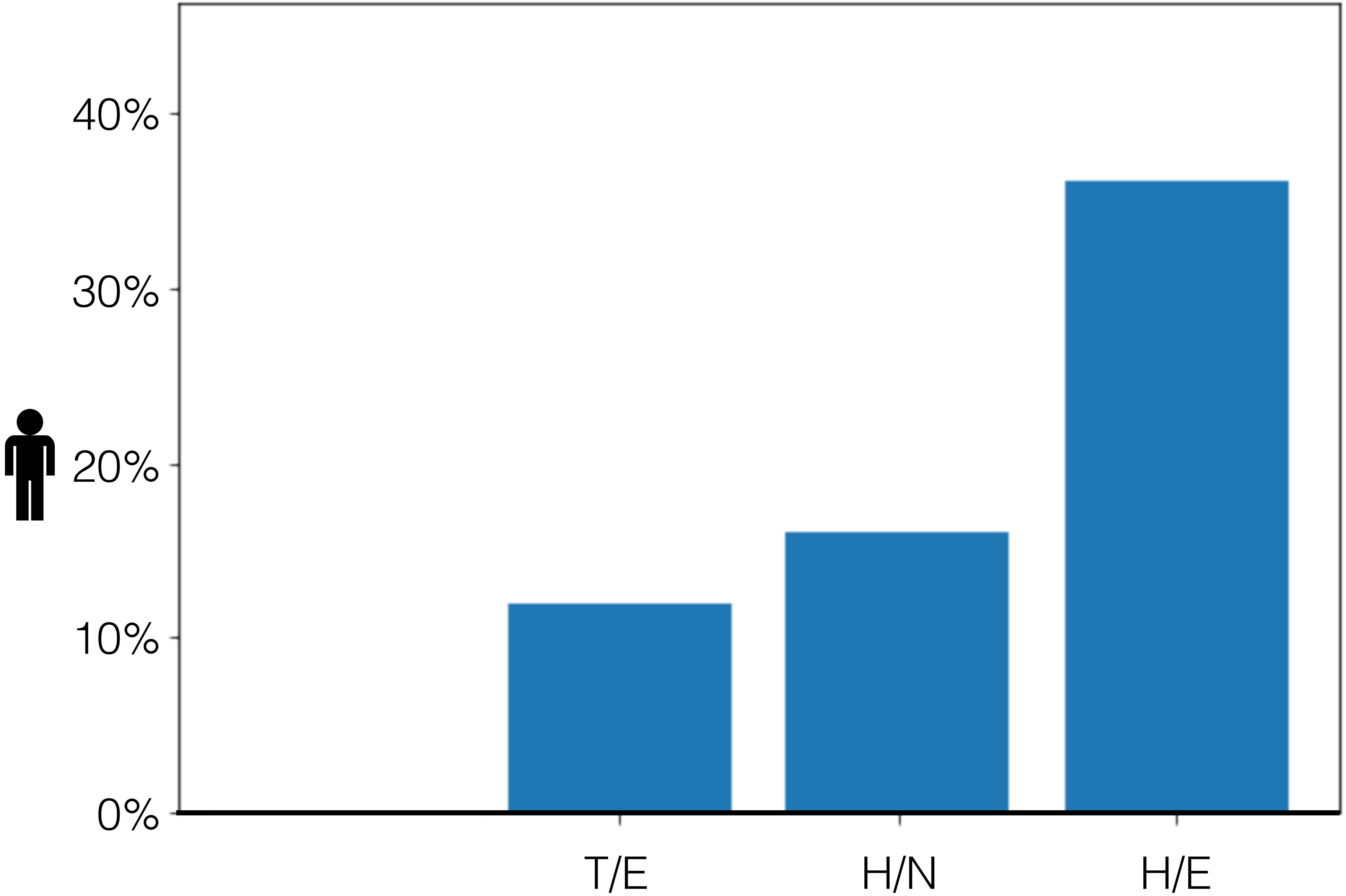
4 conditions (25 each):

- Table without Explanations (**T/N**)
- Table with Explanations (**T/E**)
- Histogram without Explanations (**H/N**)
- Histogram with Explanations (**H/E**)

Random model order
Correctly identified more accurate model

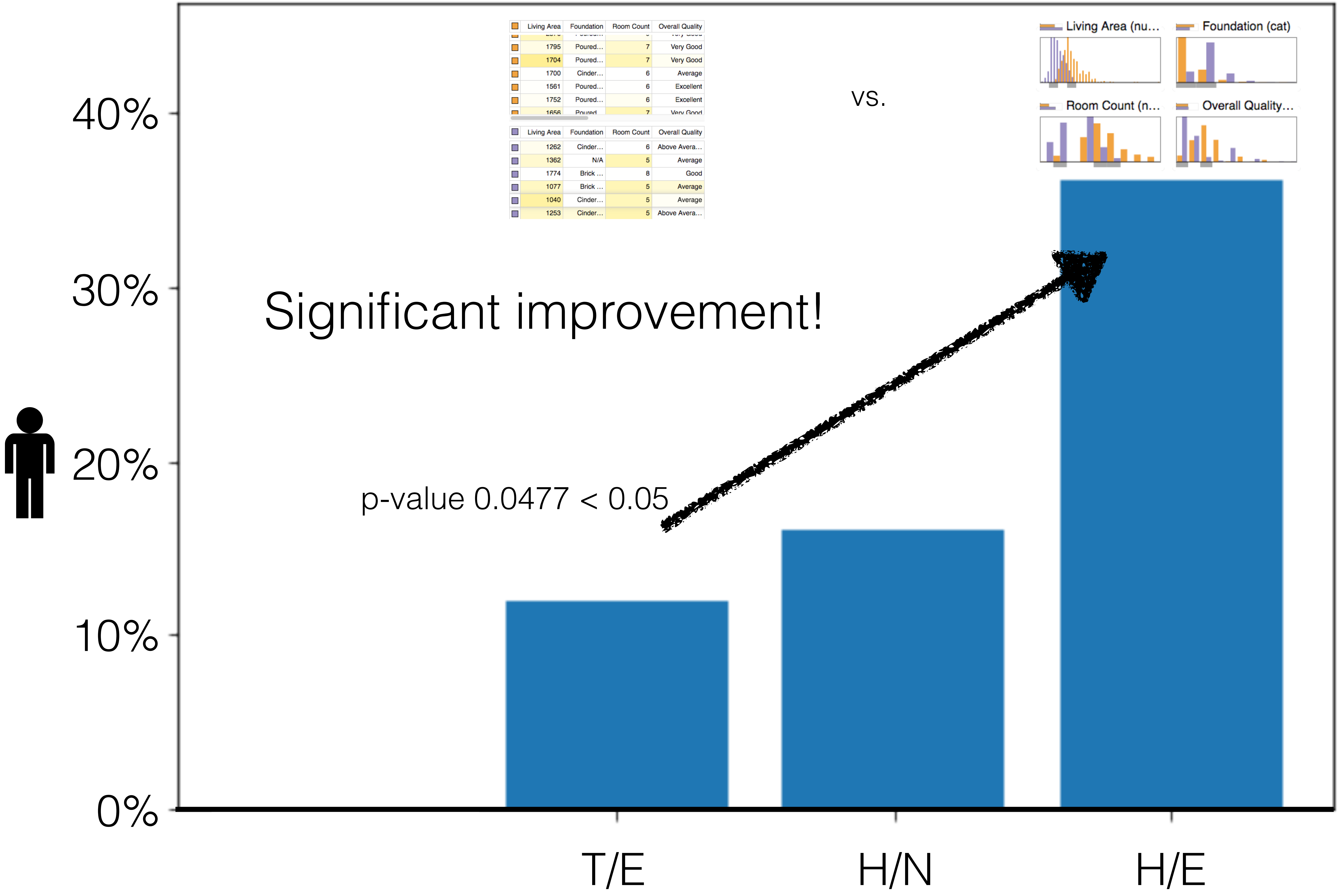
Evaluation metrics:
Model preference (trust)
Bias detection

Participants Who Trusted the Correct Model



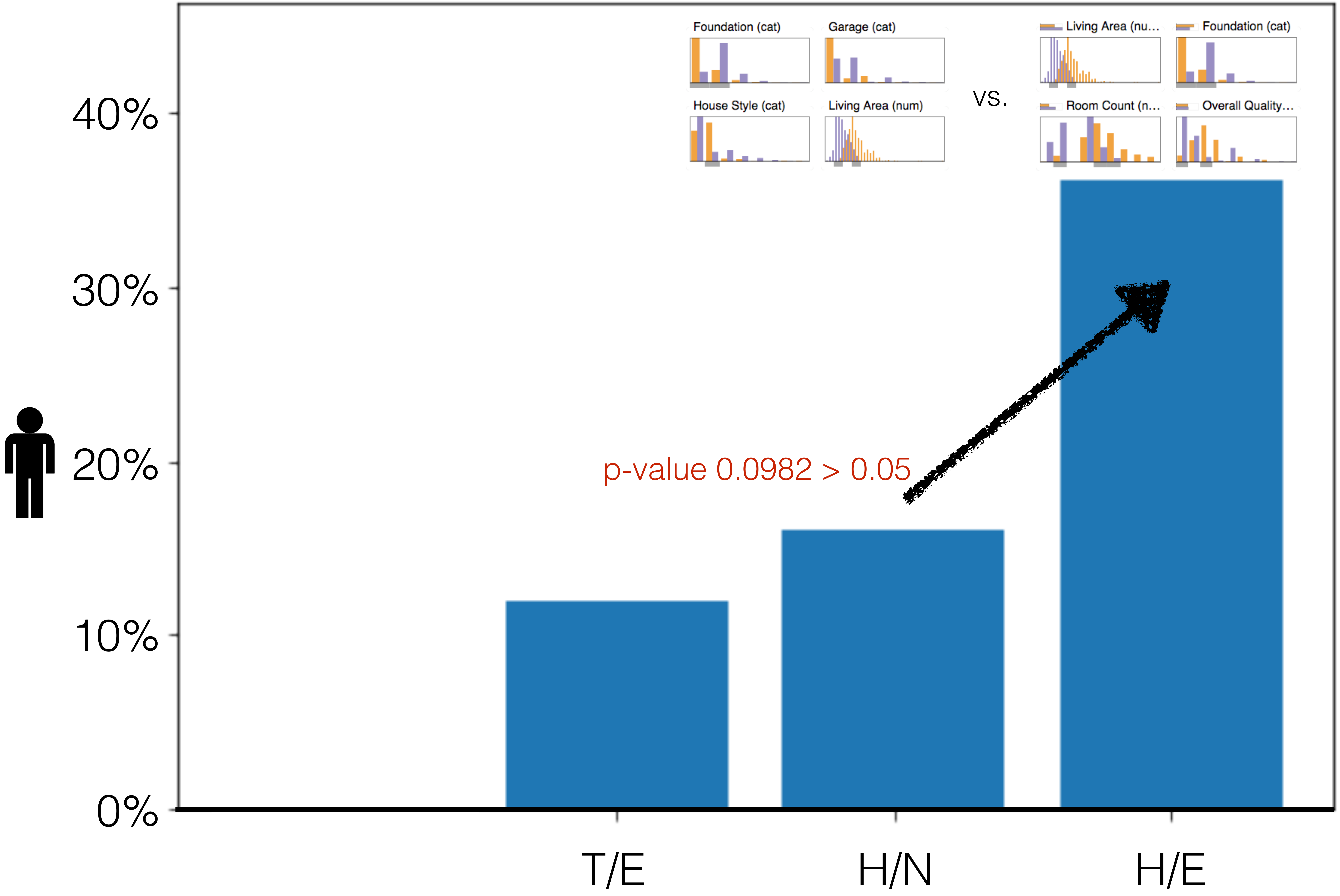
T: Table H: Histogram E: Explanation N: No Explanation

Participants Who Trusted the Correct Model



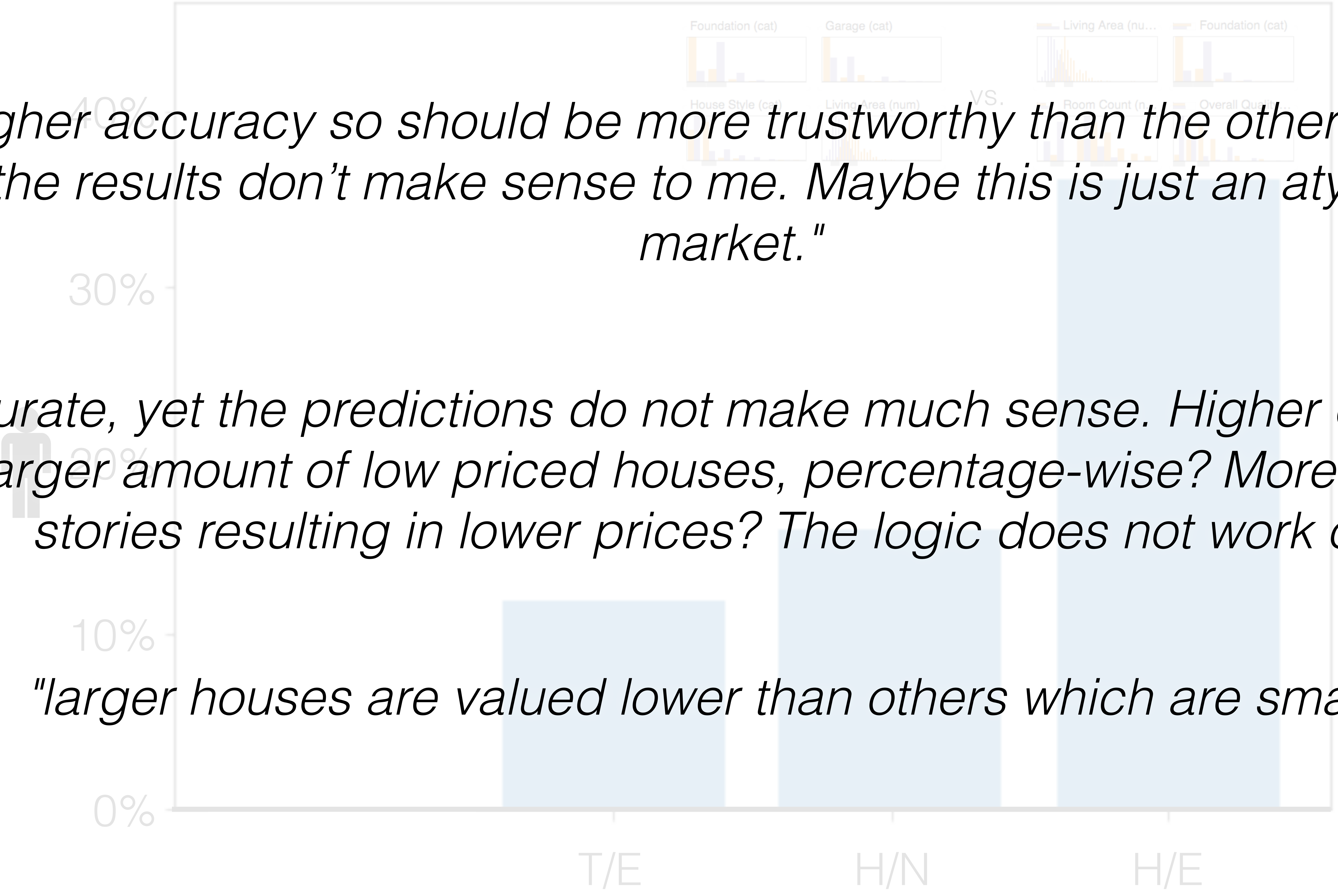
T: Table H: Histogram E: Explanation N: No Explanation

Participants Who Trusted the Correct Model



T: Table H: Histogram E: Explanation N: No Explanation

Participants Who Trusted the Correct Model



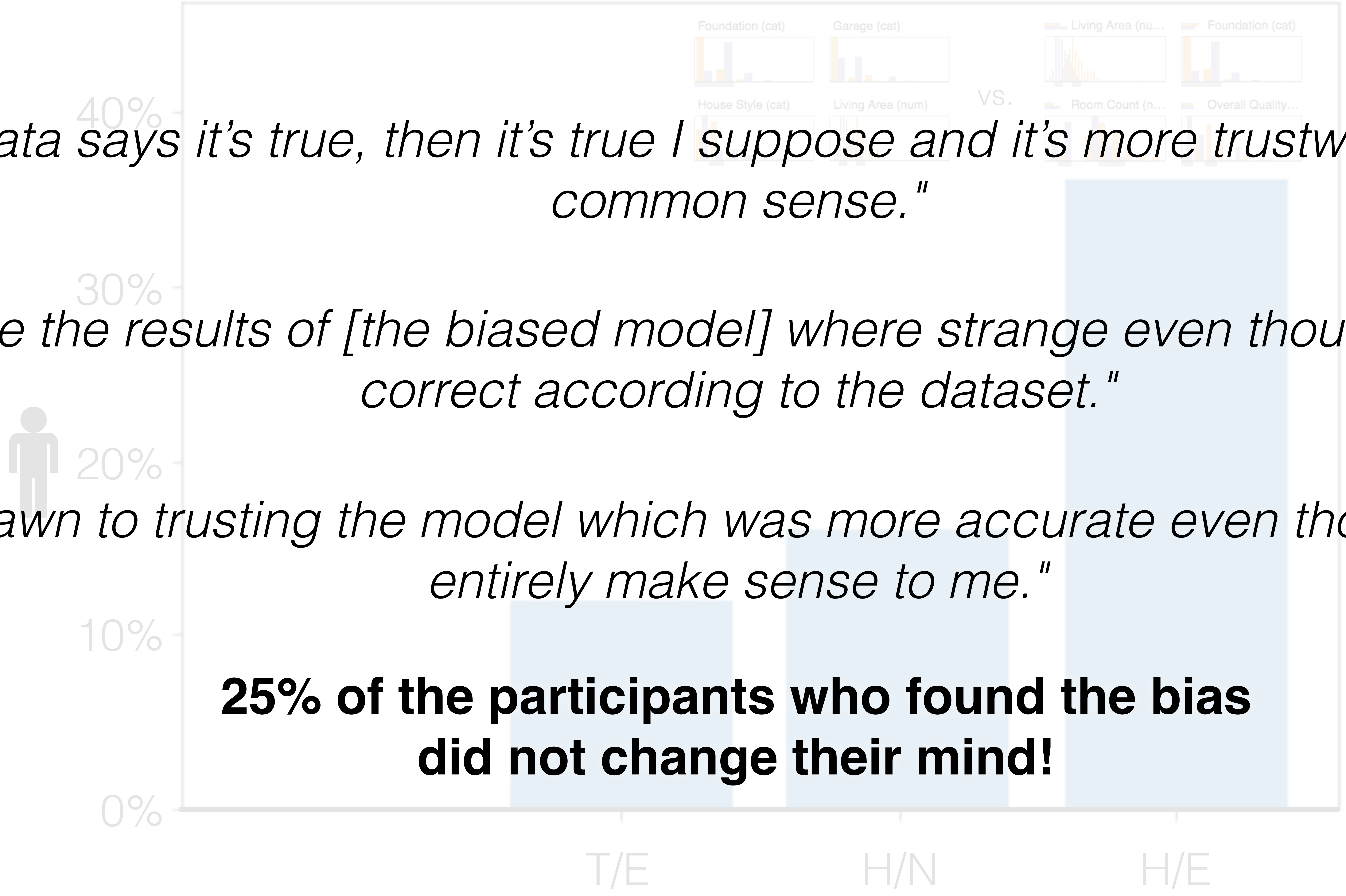
"It has higher accuracy so should be more trustworthy than the other one. However some of the results don’t make sense to me. Maybe this is just an atypical property market."

"It is accurate, yet the predictions do not make much sense. Higher quality houses having a larger amount of low priced houses, percentage-wise? More rooms, area, or stories resulting in lower prices? The logic does not work out."

"larger houses are valued lower than others which are smaller"

T: Table H: Histogram E: Explanation N: No Explanation

Participants Who Trusted the Correct Model



"If the data says it's true, then it's true I suppose and it's more trustworthy than my common sense."

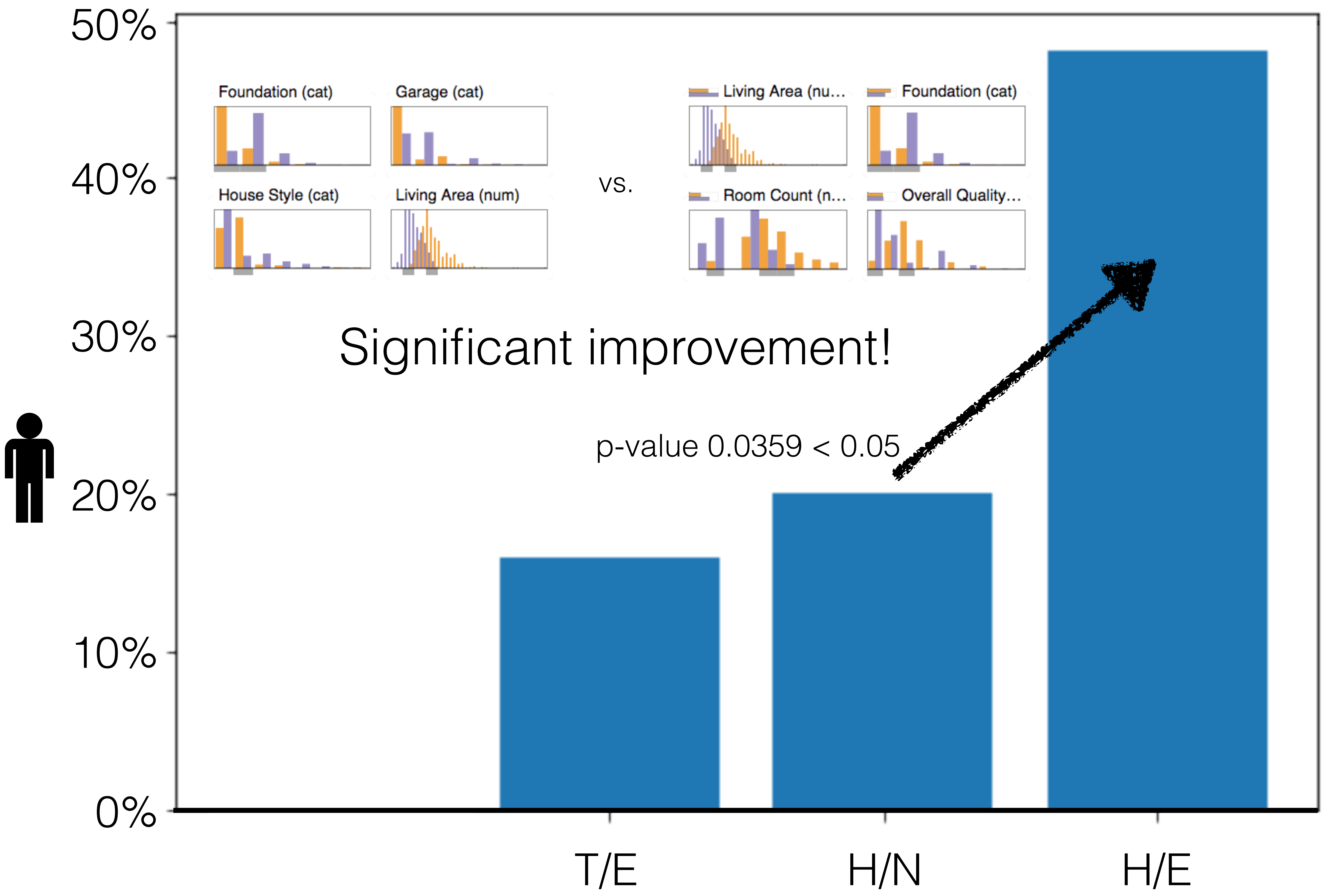
"I feel like the results of [the biased model] where strange even though they where correct according to the dataset."

"I'm drawn to trusting the model which was more accurate even though it didn't entirely make sense to me."

25% of the participants who found the bias did not change their mind!

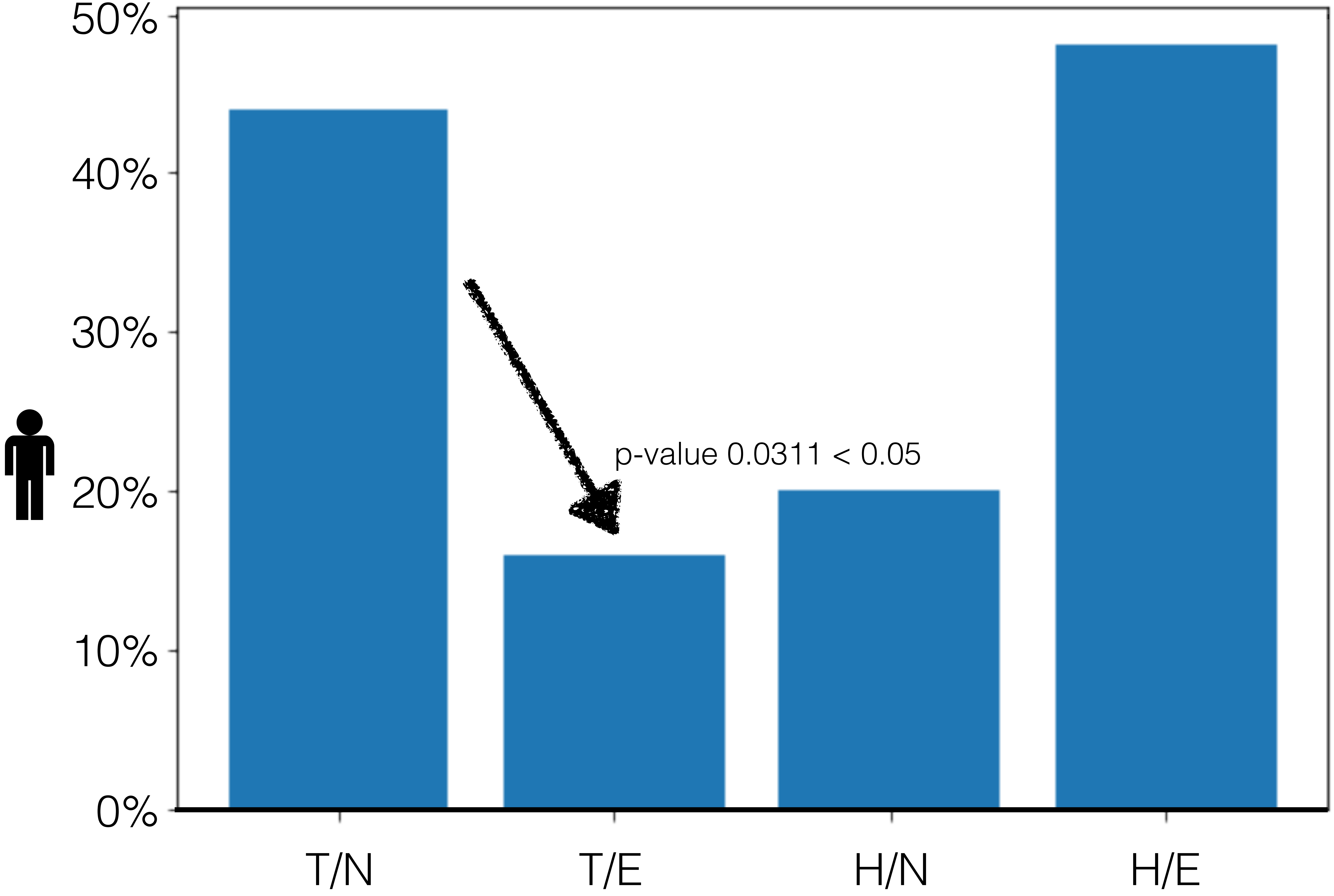
T: Table H: Histogram E: Explanation N: No Explanation

Participants Who Detected the Bias

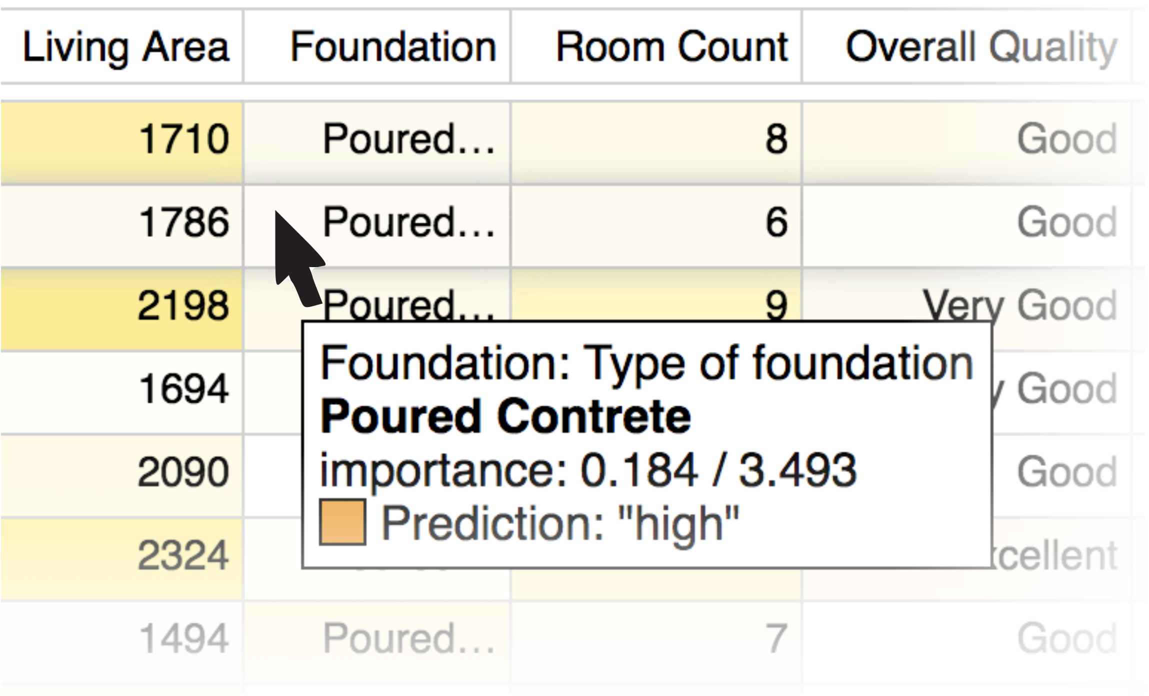


T: Table H: Histogram E: Explanation N: No Explanation

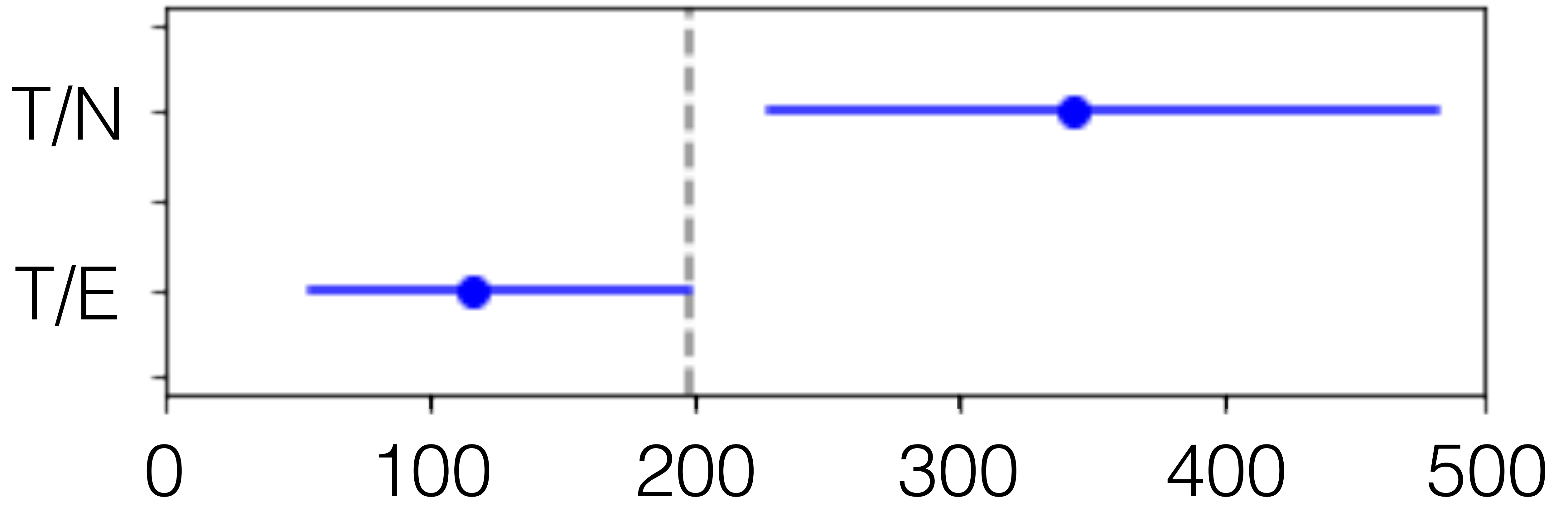
Participants Who Detected the Bias

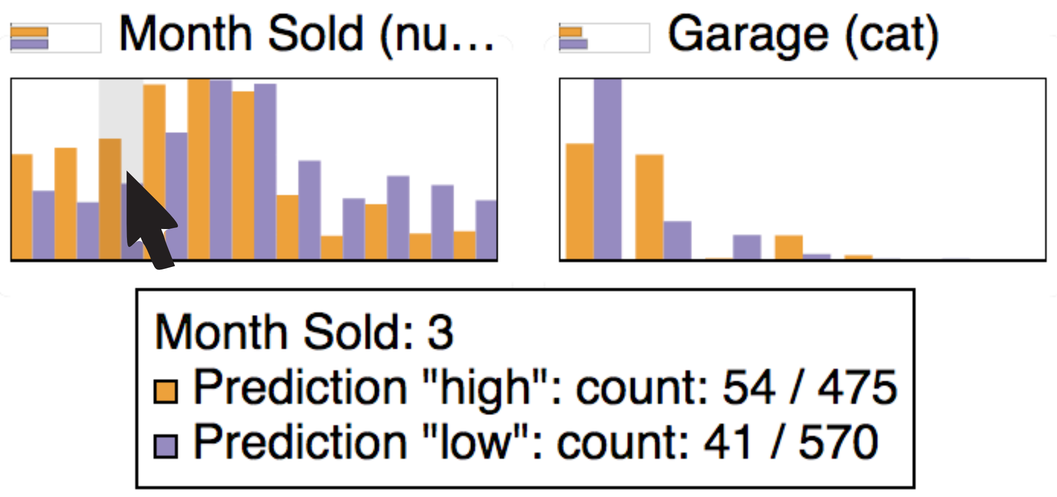


T: Table H: Histogram E: Explanation N: No Explanation

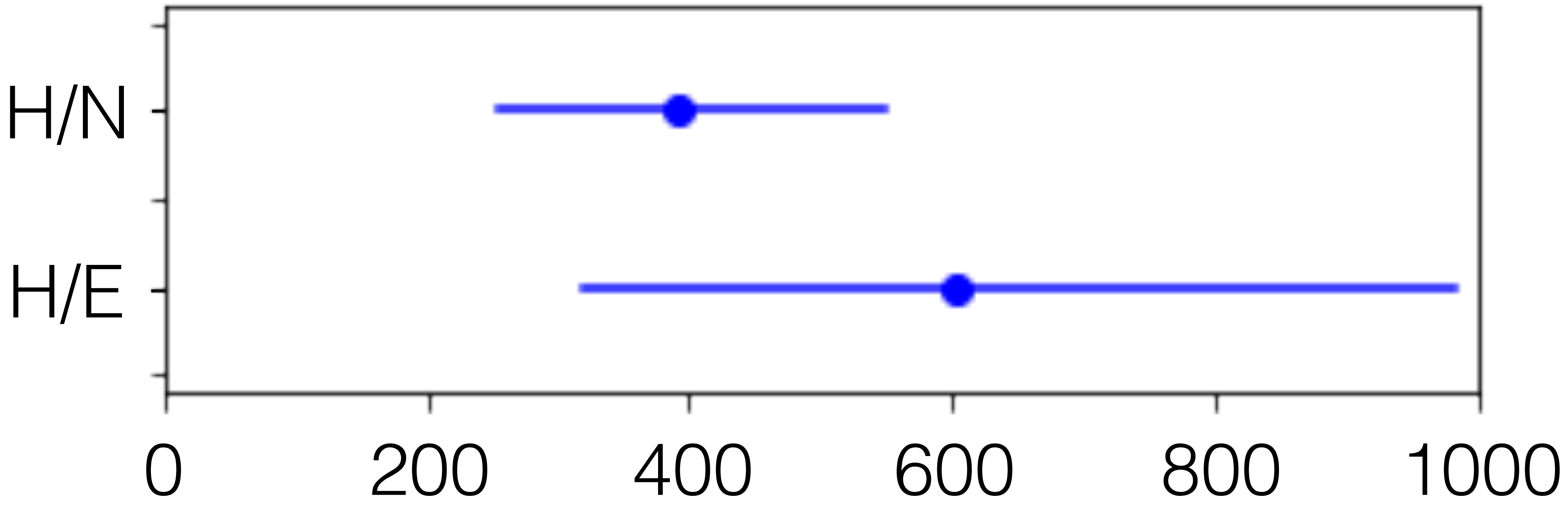


Number of Hovered Cells





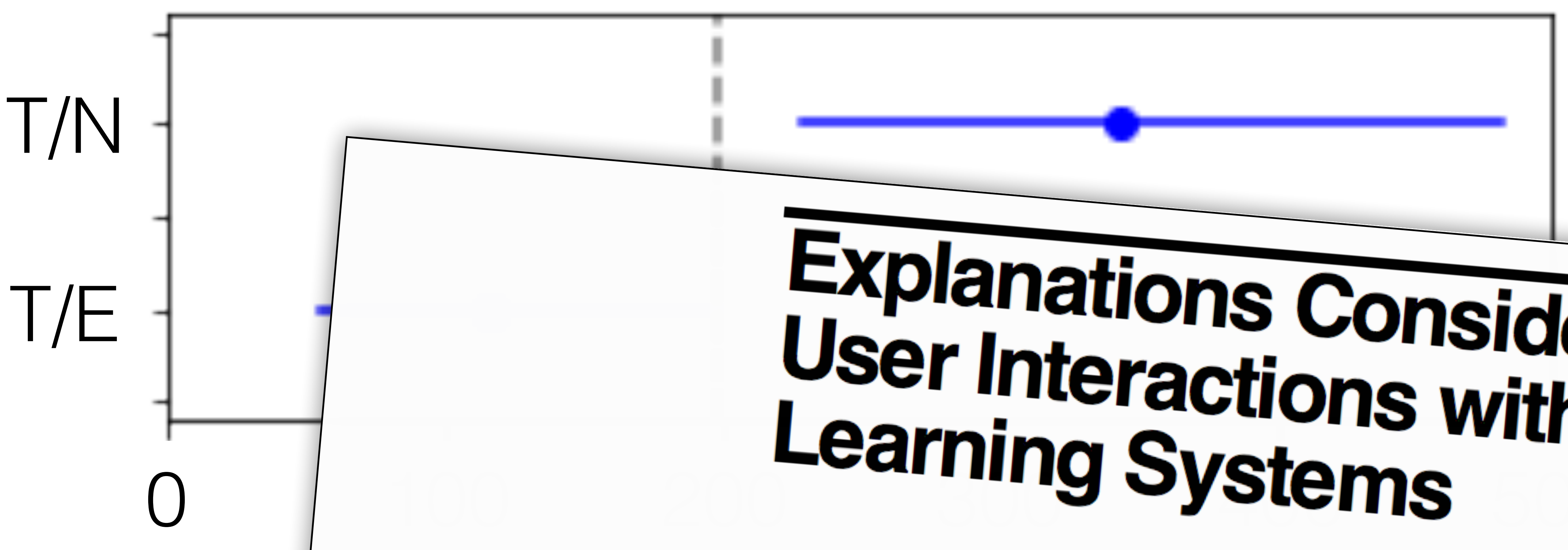
Number of Hovered Bars



Bootstrapped 95% Confidence Intervals

T: Table H: Histogram E: Explanation N: No Explanation

Number of Hovered Cells



Explanations Considered Harmful? User Interactions with Machine Learning Systems

Simone Stumpf
Adrian Bussone
Dympna O'Sullivan
Centre for Human Computer
Interaction Design
School ...

Abstract
It has been suggested that the intelligibility of machine learning system behavior is an important factor in ensuring that users can identify that the system has erred, understand how the system operates and that thereby they are better able to provide appropriate feedback to the machine learning system to improve its accuracy. There has been increasing research into how to make machine learning intelligible to users without a background in AI, and it has been shown that providing explanations of a system's reasoning has many benefits. In this paper we review recent work in this area but also point to instances when explanations might have less desirable effects. Further work is warranted to understand how best to expose the reasoning of machine learning systems to improve their usability.

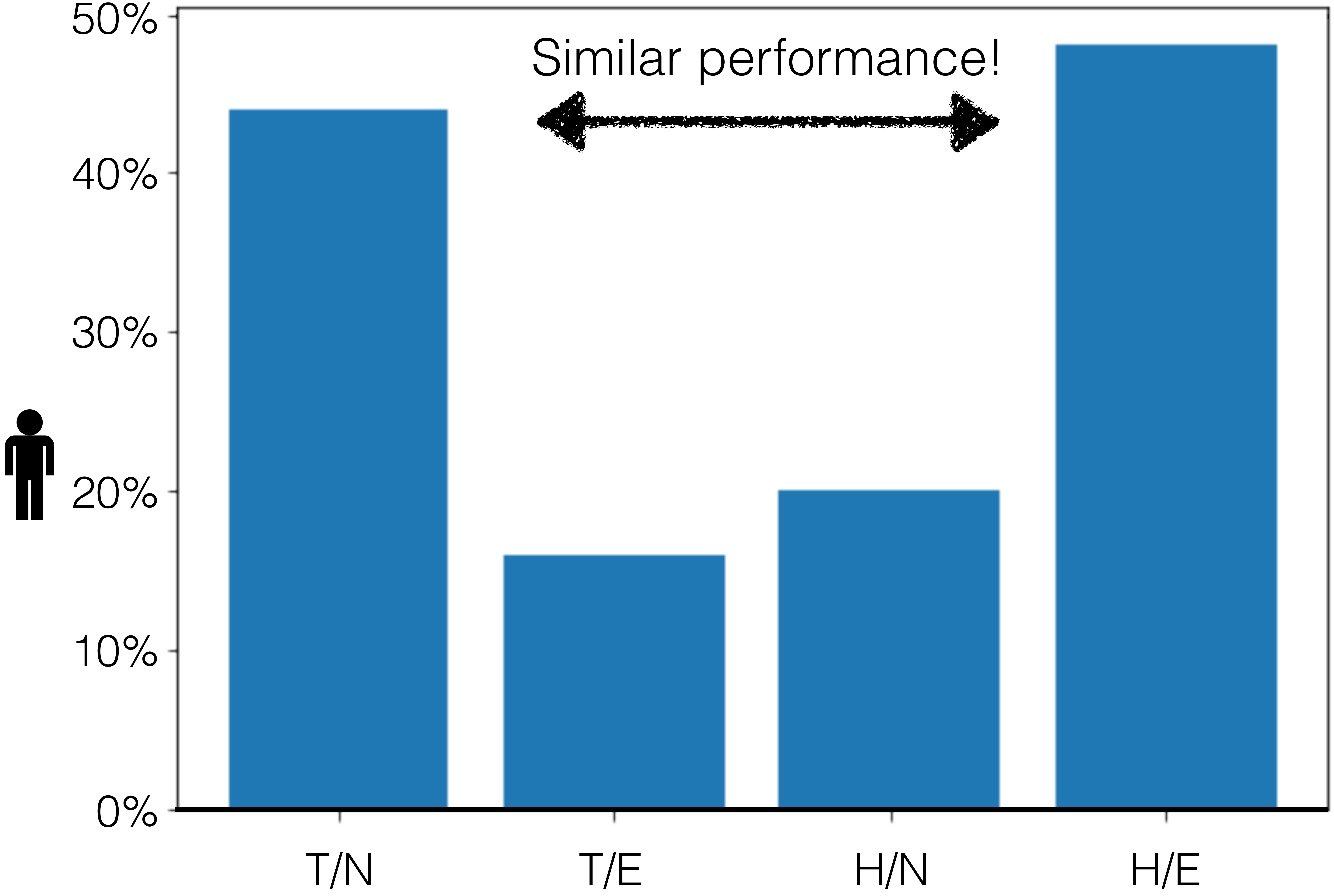
Author Keywords
Machine learning; explanations; reliability; intelligibility.

Benefitting InfoVis with Visual Difficulties

Jessica Hullman, *Student Member, IEEE*, Eytan Adar, and Priti Shah

Abstract—Many well-cited theories for visualization design state that a visual representation should be optimized for quick and immediate interpretation by a user. Distracting elements like decorative "chartjunk" or extraneous information are avoided so as not to slow comprehension. Yet several recent studies in visualization research provide evidence that non-efficient visual elements may benefit comprehension and recall on the part of users. Similarly, findings from studies related to learning from visual displays in various subfields of psychology suggest that introducing cognitive difficulties to visualization interaction can improve a user's understanding of important information. In this paper, we synthesize empirical results from cross-disciplinary research on visual information representations, providing a counterpoint to efficiency-based design theory with guidelines that describe how *visual difficulties* can be introduced to benefit comprehension and recall. We identify conditions under which the application of visual difficulties is appropriate based on underlying factors in visualization interaction like active processing and engagement. We characterize effective graph design as a trade-off between efficiency and learning difficulties in order to provide Information Visualization (InfoVis) researchers and practitioners with a framework for organizing explorations of graphs for which ... and recall are crucial. We identify implications of this view for the design and evaluation of information ... dividual differences.

Participants Who Detected the Bias

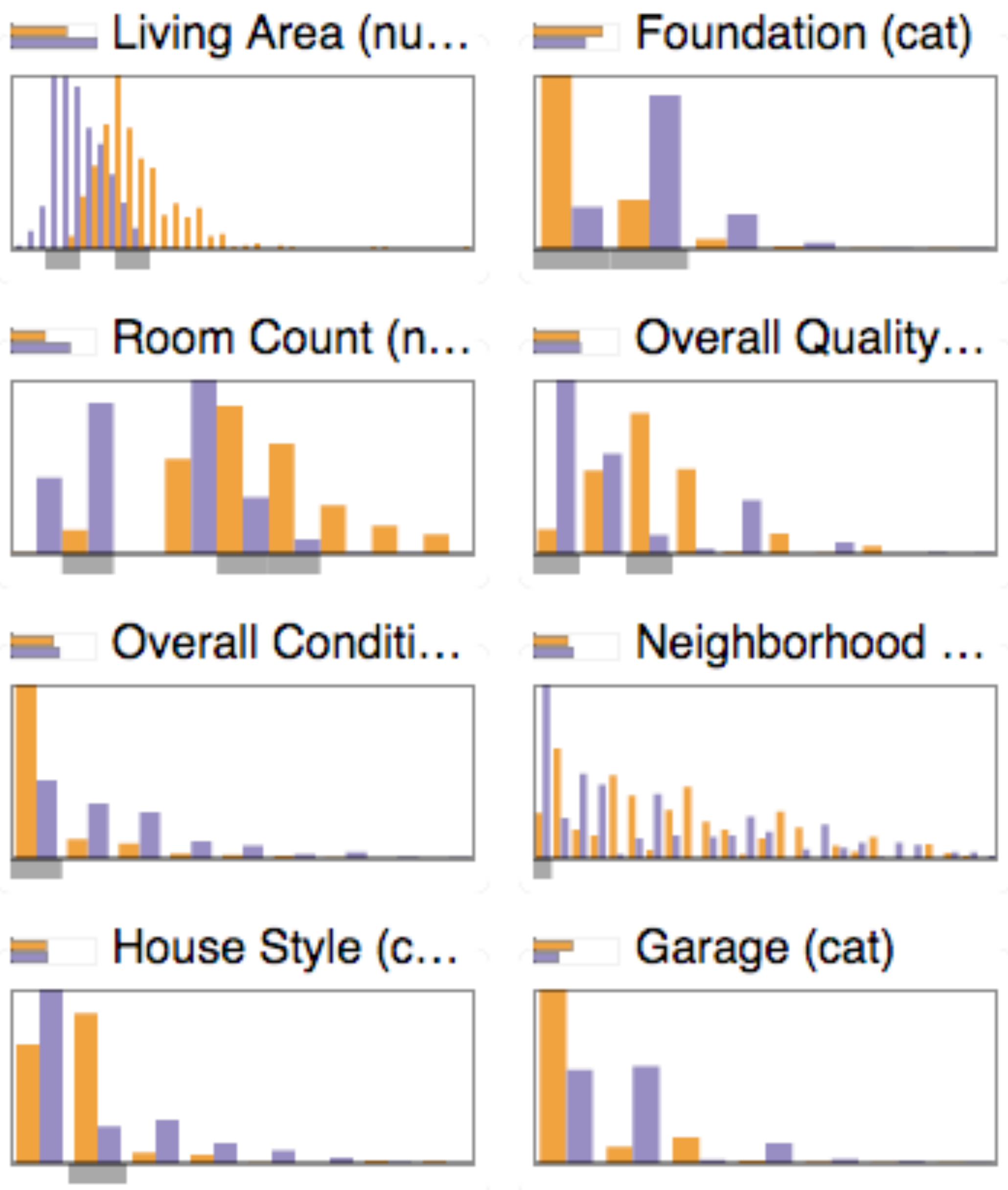


T: Table H: Histogram E: Explanation N: No Explanation

Foundation	Garage	House Style	Living Area	Mon
Poured…	Attac…	One sto…	1795	
Poured…	Attac…	One sto…	1704	
Cinder…	Attac…	One sto…	1700	
Poured…	Attac…	One sto…	1561	
Poured…	Attac…	One sto…	1752	
Poured	Attac	One sto	1656	

Foundation	Garage	House Style	Living Area	Mon
Cinder…	Attac…	One sto…	1262	
N/A	Attac…	One and…	1362	
Brick …	Detac…	One and…	1774	
Brick …	Attac…	One and…	1077	
Cinder…	Detac…	One sto…	1040	
Cinder…	Attac…	One sto…	1253	

vs.



Note that the task was chosen in a way that under **all conditions** it was possible to find the bias.

Histograms scale better to larger data sets or more complex errors in the data. In tables you have to extrapolate...

Lessons Learned

People trust accuracy (too much).

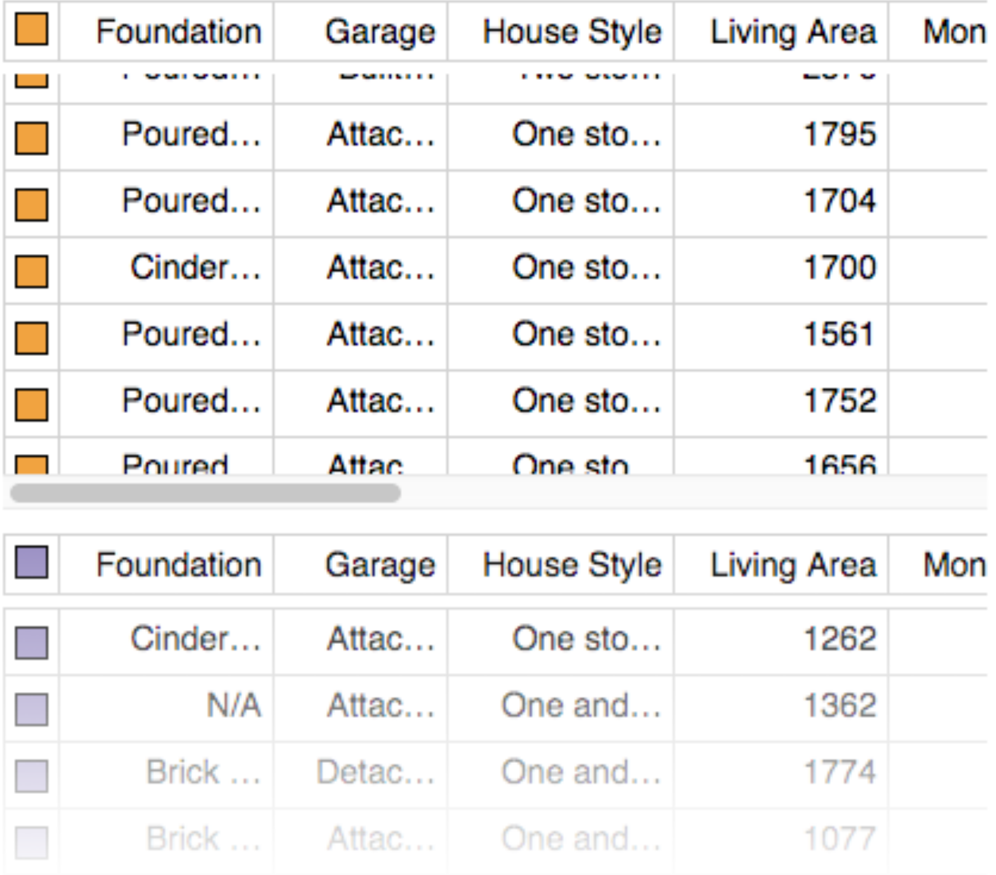
Aggregating instance-level explanations significantly helps detecting biases compared to individual explanations.

Individual instance-level explanations may hurt performance.

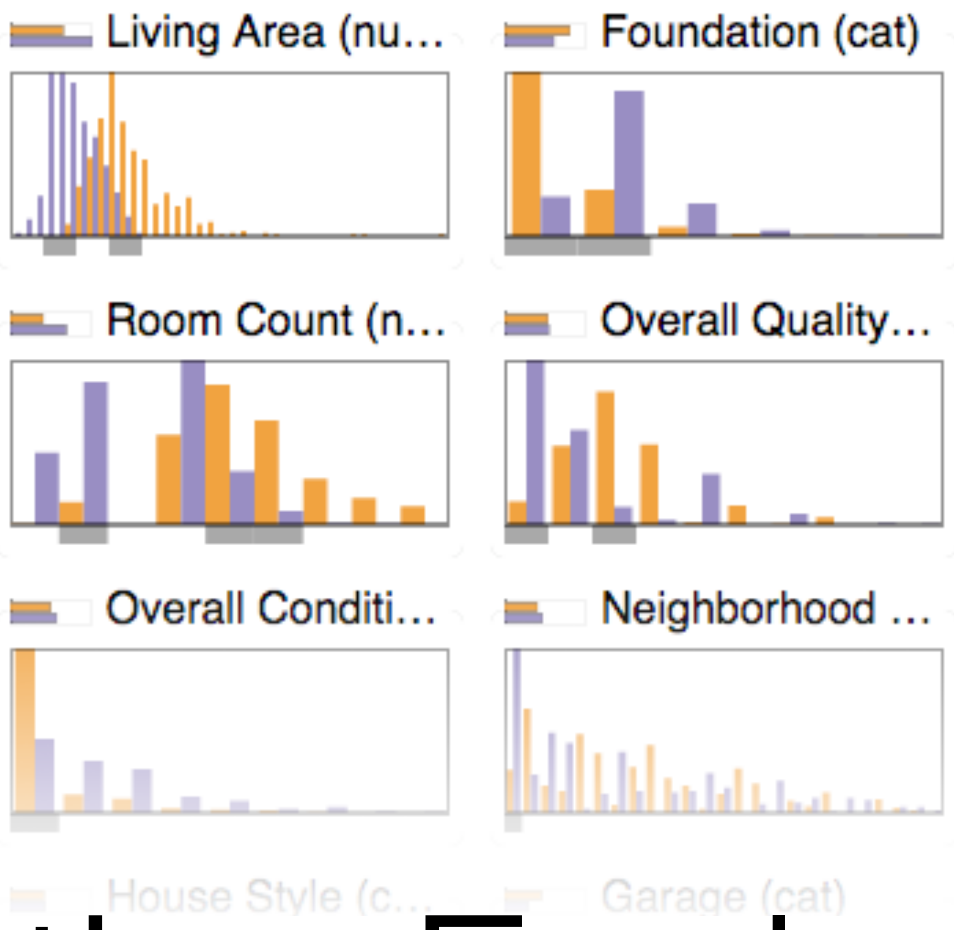
Further Work

More targeted studies
to confirm hypotheses

Different results for expert users?



Thank You!



A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models

[work in progress]

Josua Krause*, Adam Perer**, Enrico Bertini*

* NYU TANDON SCHOOL OF ENGINEERING

** IBM