The Role of Data in Building Robust Models

Roy Schwartz
The Hebrew University of Jerusalem
Google, 08/2023

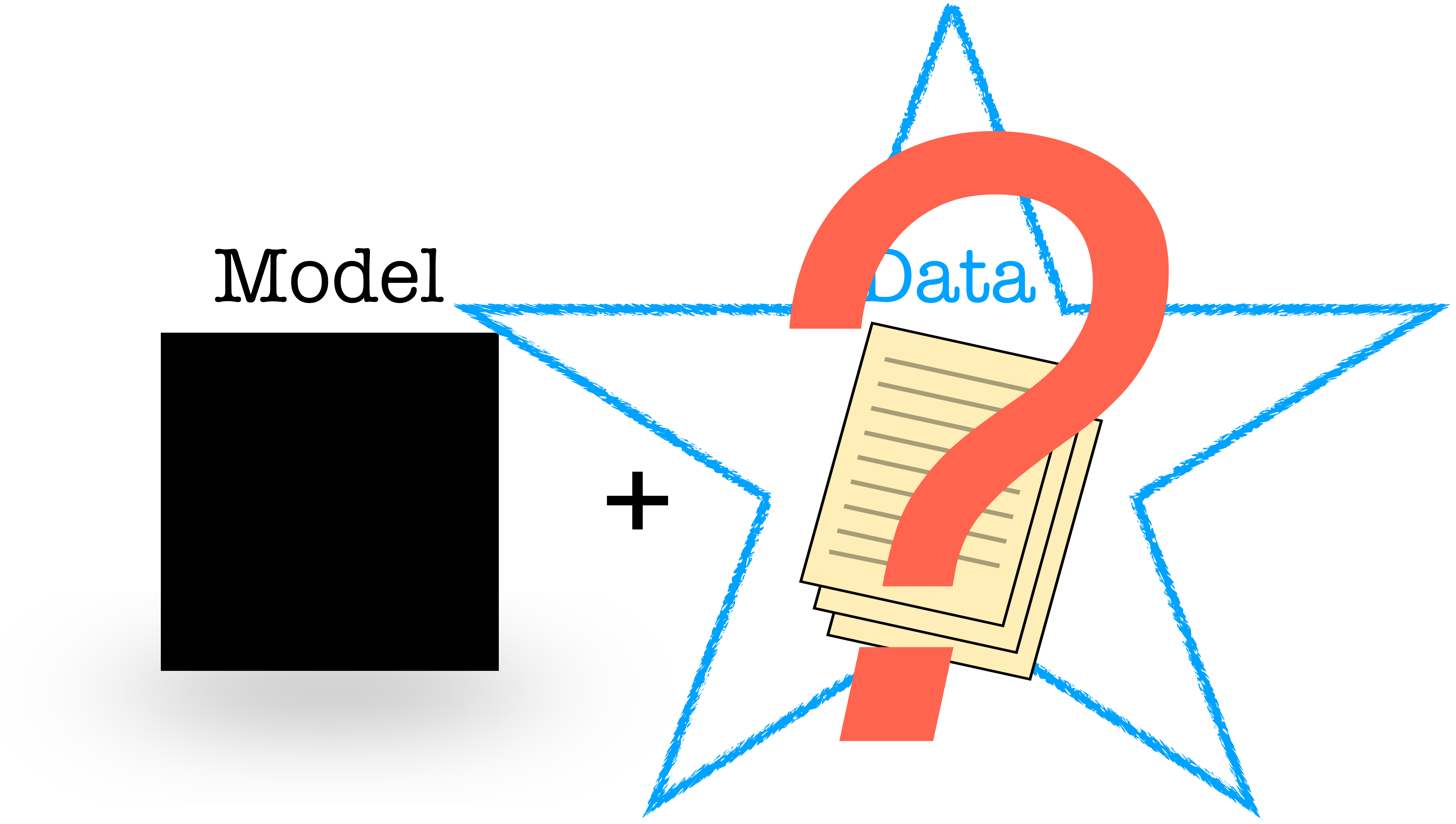
THE HEBREW UNIVERSITY OF JERUSALEM

Model Robustness

Wang et al. (2022)

Distribution shifts

Adversarial Attacks



Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases

Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases

Visual Question Answering

- VQA dataset
  - Antol et al. (2015)
- Input: an image and a question
  - What sport is this man playing?
  - Do you see a shadow?
- Output: answer
  - Tennis, yes

Spurious Correlations in VQA

- 40% of the questions in VQA starting with "***What sport is this***" are answered with "***tennis***"
- "***yes***" is the answer to 87% of the questions in the VQA dataset starting with "***Do you see a***"
  - Zhang et al. (2016); Goyal et al. (2017)

ROC Story Cloze Task

Mostafazadeh et al. (2016)

Context	Right Ending	Wrong Ending
Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.

- A story comprehension task
- The task: given a story prefix, distinguish between the coherent and the incoherent endings

Spurious Correlations in ROC

S. et al. (2017); Cai et al. (2017)

- Train a binary classifier on **the endings only**
    - Ignoring the story prefix

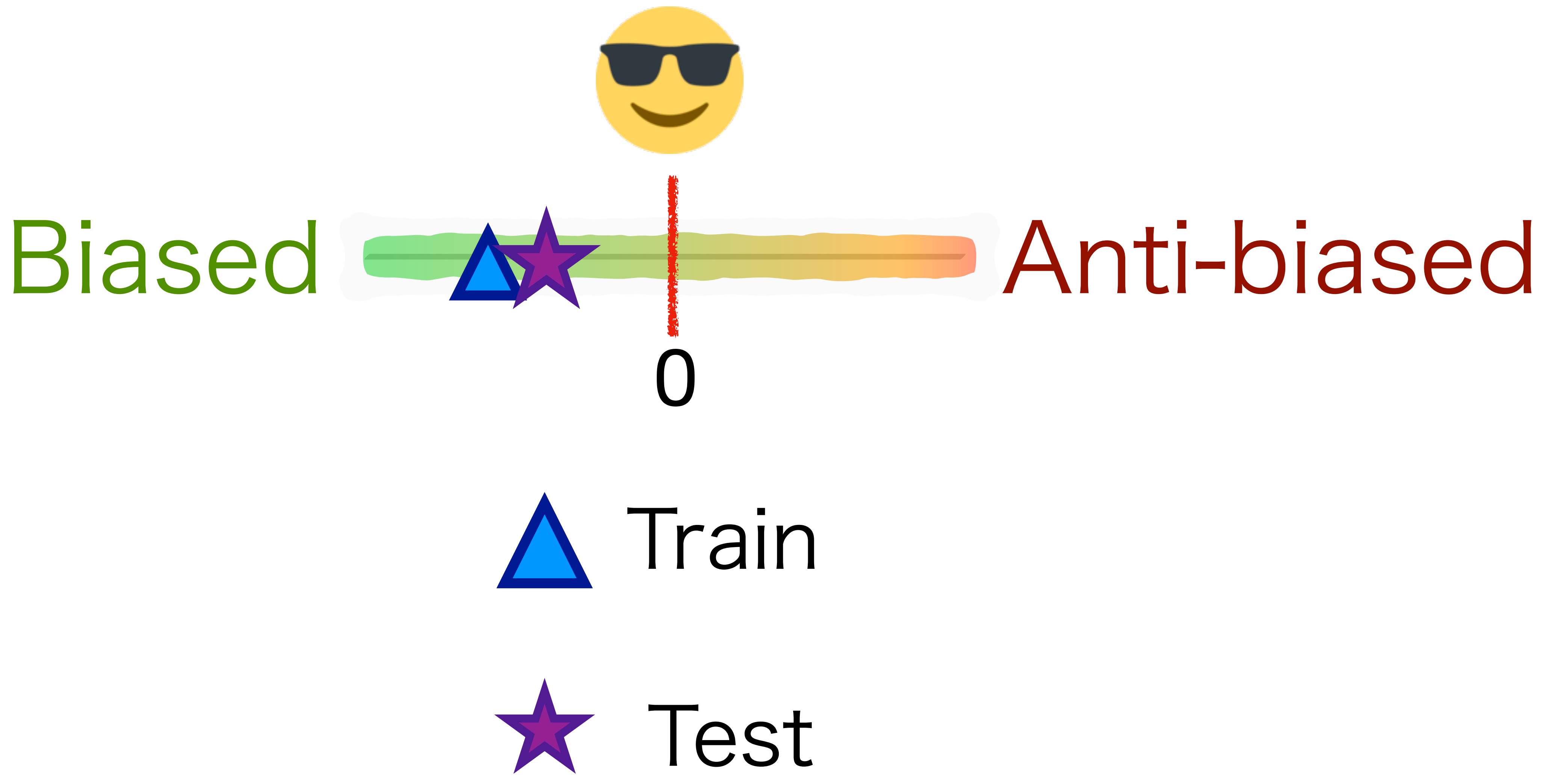
Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Model	Acc.
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
†EndingsOnly (Cai et al., 2017)	0.725
cogcomp	0.744
HIER,ENCPLOTEND,ATT (Cai et al., 2017)	0.747
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Right	Weight	Freq.	***Wrong***	Weight	Freq.
‘ed .’	0.17	6.5%	START NNP	0.21	54.8%
‘and ’	0.15	13.6%	NN .	0.17	47.5%
JJ	0.14	45.8%	NN NN .	0.15	5.1%
to VB	0.13	20.1%	VBG	0.11	10.1%
‘d th’	0.12	10.9%	START NNP VBD	0.11	41.9%

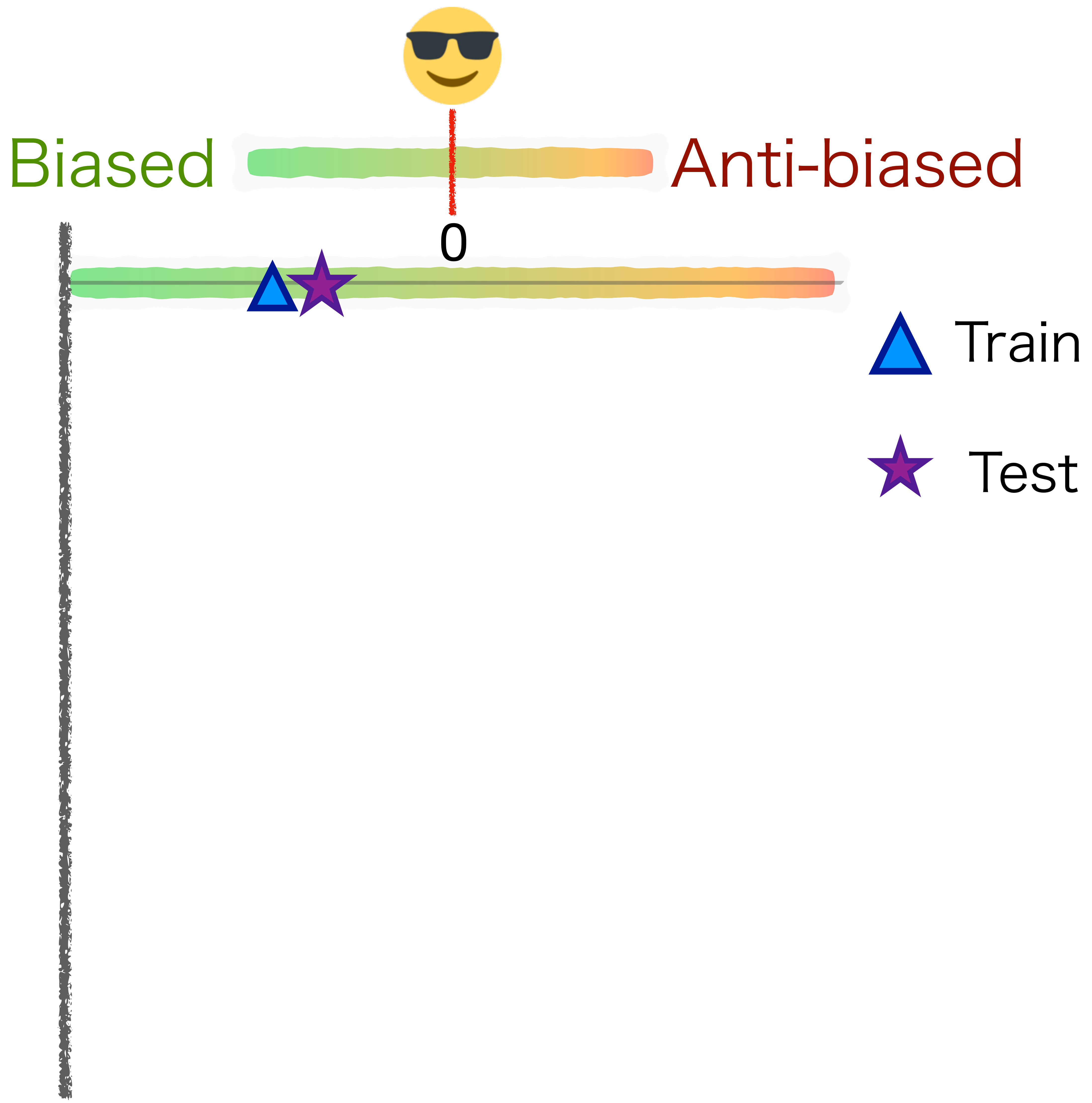
Other Spurious Correlations

- Other tasks
  - NLI (Gururangan, …, S. et al., 2018; Poliak et al., 2018; Tsuchiya, 2018)
  - Question answering (Kaushik & Lipton, 2018)
  - Winograd Schema (Elazar et al., 2021)
  - …



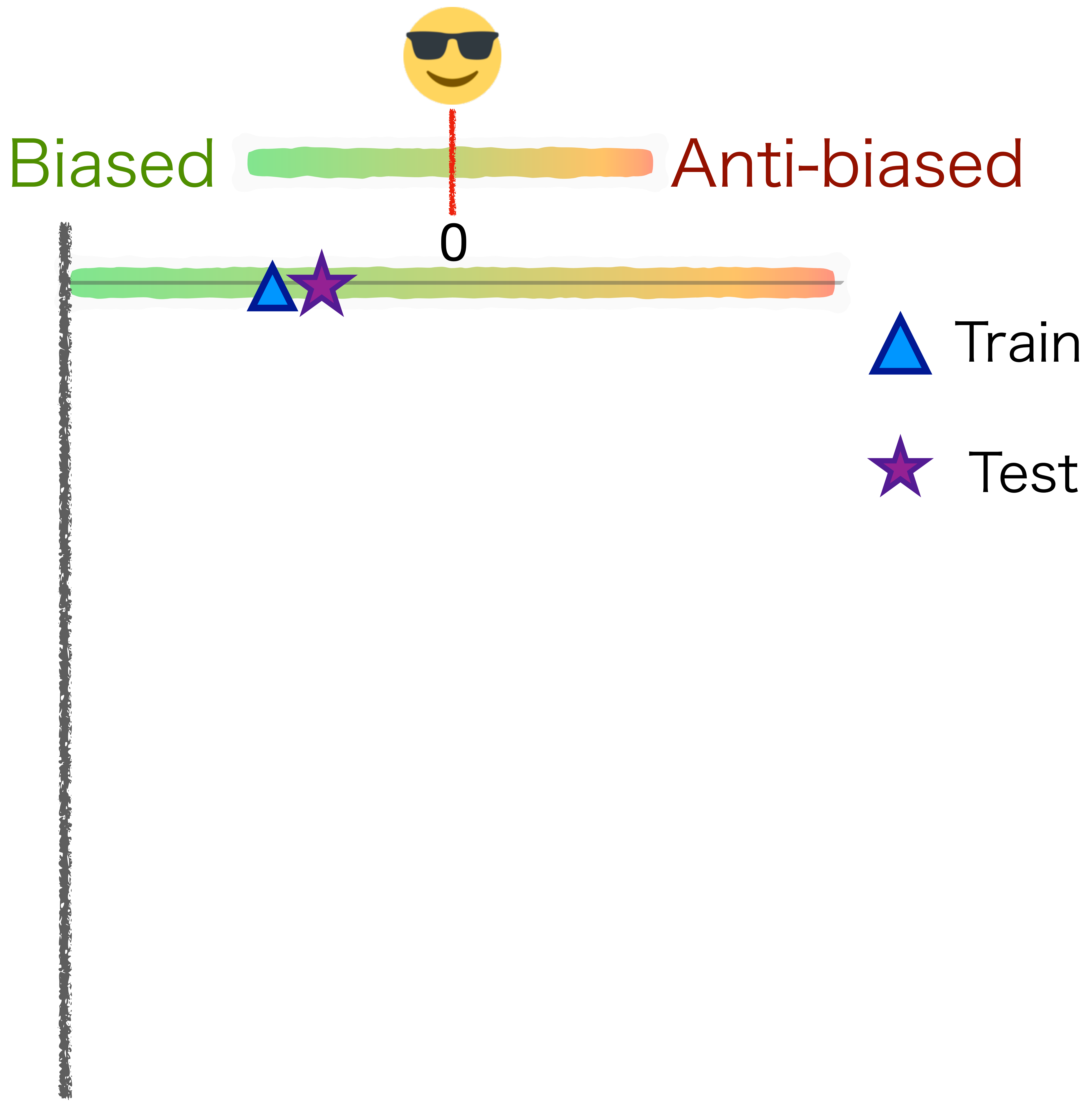
Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Dataset Balancing

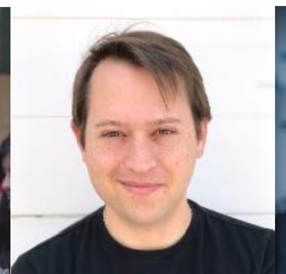
Augmentation

- The key idea: balance-out spurious correlations
- Vision and Language datasets
  - VQA 2.0 (Goyal et al., 2017)
  - GQA (Hudson and Manning, 2019)
- Language only
  - ROC stories cloze task 1.5 (Sharma et al., 2018)

Filtering

Zellers, Bisk, S. & Choi (2018); Sakaguchi et al. (2020)

- Filter-out “easy” examples from existing datasets
  - Typically using an adversarial model
- A widely used approach
  - SWAG (Zellers, Bisk, S. & Choi (2018); Record (Zhang et al., 2018); WinoGrande (Sakaguchi et al., 2020)


Filtering as Balancing

- As the adversarial model grows, models will pick up *subtler* correlations
- At the extreme, the result is a fully *balanced* dataset

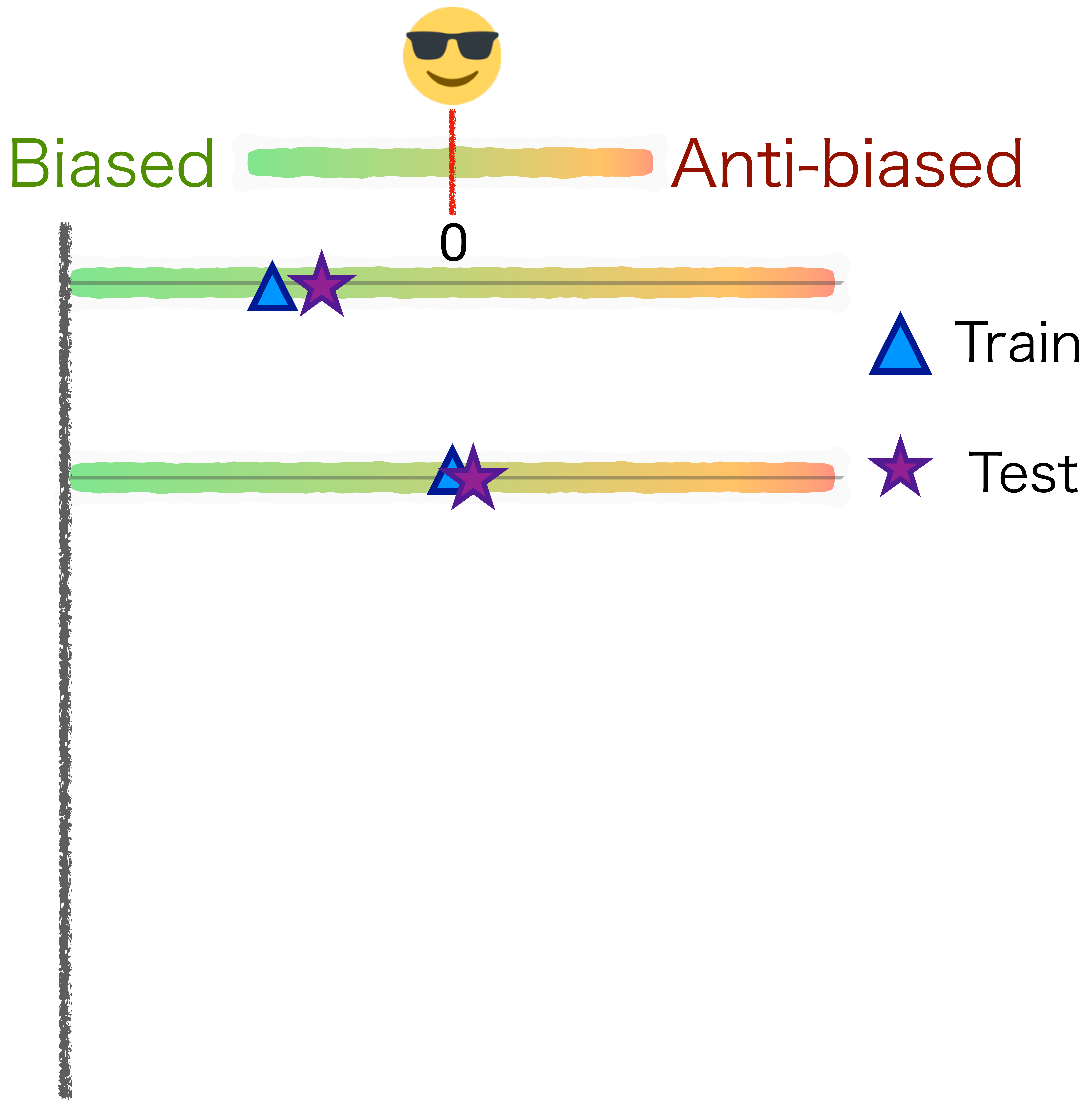
Every Correlation is Spurious!

Gardner et al. (2021)

- *Every* simple correlation between single word features and output labels is spurious
- *Competent* datasets: the marginal probability for every feature is uniform over the class label
  - $\forall x_i, y \in Y, p(y \mid x_i) = \dfrac{1}{|Y|}$

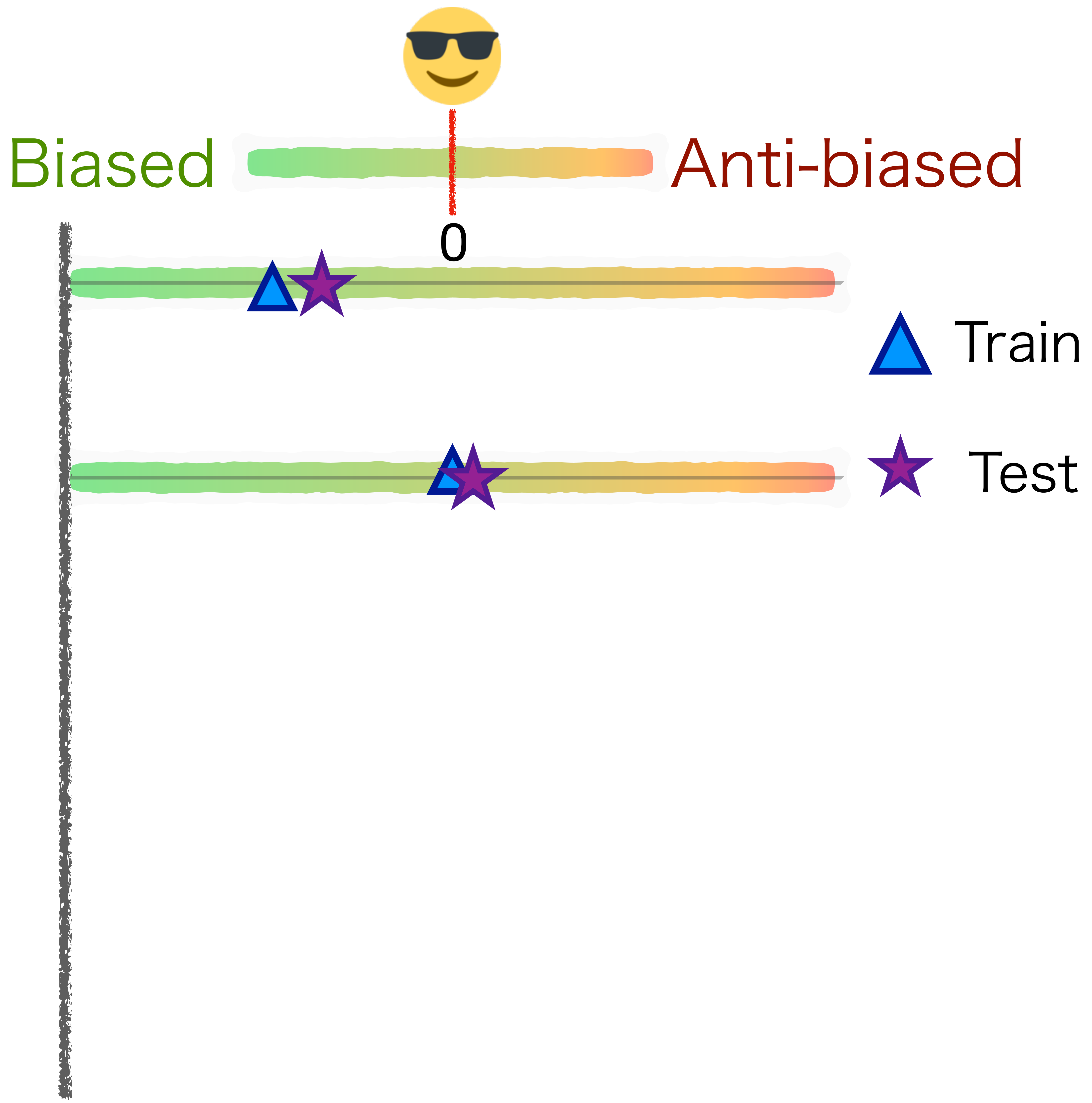
Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases

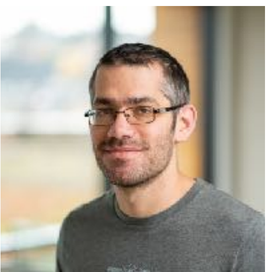


Reality

Benchmark	Baseline	Shortly after
SWAG (Zellers, Bisk, S. & Choi, 2018)	52%	86% (Devlin et al., 2018)
DROP (Dua et al., 2019)	47 F1	88 F1 (Chen et al., 2020)
HellaSWAG (Zellers et al., 2019)	47%	93% (He et al., 2020)
WinoGrande (Sakaguchi et al., 2020)	53% AUC	88% AUC (Raffel et al., 2020)

On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations

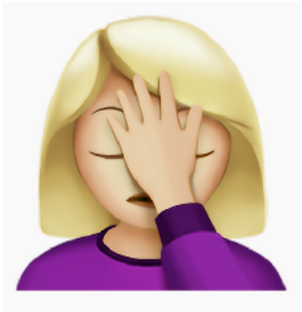
Roy Schwartz **Gabriel Stanovsky**
School of Computer Science, The Hebrew University of Jerusalem
{roy.schwartz1,gabriel.stanovsky}@mail.huji.ac.il



Balancing too Little is Insufficient

Toy Example

The dataset is balanced for unigrams



But still contains spurious **bigrams** features

- E.g., “*very good*”, as “***not** very good*” yields negative sentiment

Split	Text	Label
Train	very good	+
	very bad	−
	not good	−
	not bad	+
Test	not very good	−
	good	+

Balancing too Little is Insufficient

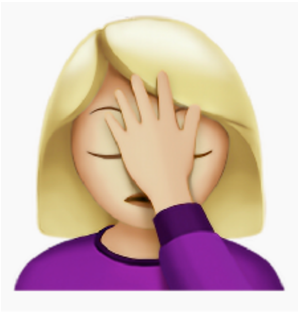
Natural Language

- The same example can apply with larger *n*’s
- More broadly, any phrase or feature combination can alter its meaning in some context
  - Negation, sarcasm, humor, …
- As a result, balancing too little is **insufficient** for mitigating all spurious correlations

Too much Balancing Leaves Nothing

Toy Example

The dataset is also balanced for unigrams



But if we balance it for bigrams, we are left with **no learnable signal**

Original Train Set	
Input	**Label**
0 0	0
0 1	1
1 0	1
1 1	0

Too much Balancing Leaves Nothing

More Broadly

- Consider an NLP dataset D with maximal length n
- By definition, balancing any combination of up to n features (including) leaves no learnable signal in D
- Conclusion: *balancing too much* is not helpful either

Does a sweet-spot exist between balancing too little and too much?

Is Balancing even Desired?

- Dataset balancing prevents models from having a fallback option in cases of uncertainty
  - As these would evidently cause it to make mistakes on some inputs
- But fallback meanings are crucial for language understanding, as contexts are often underspecified
  - Graesser (2013)

Is Balancing even Desired?

- Especially relevant for world knowledge and common-sense knowledge
  - Joe Biden is the president of the US
  - A person is typically happy when they receive a present
- As a result, dataset balancing is **undesired**

Who is the president of the U.S.?

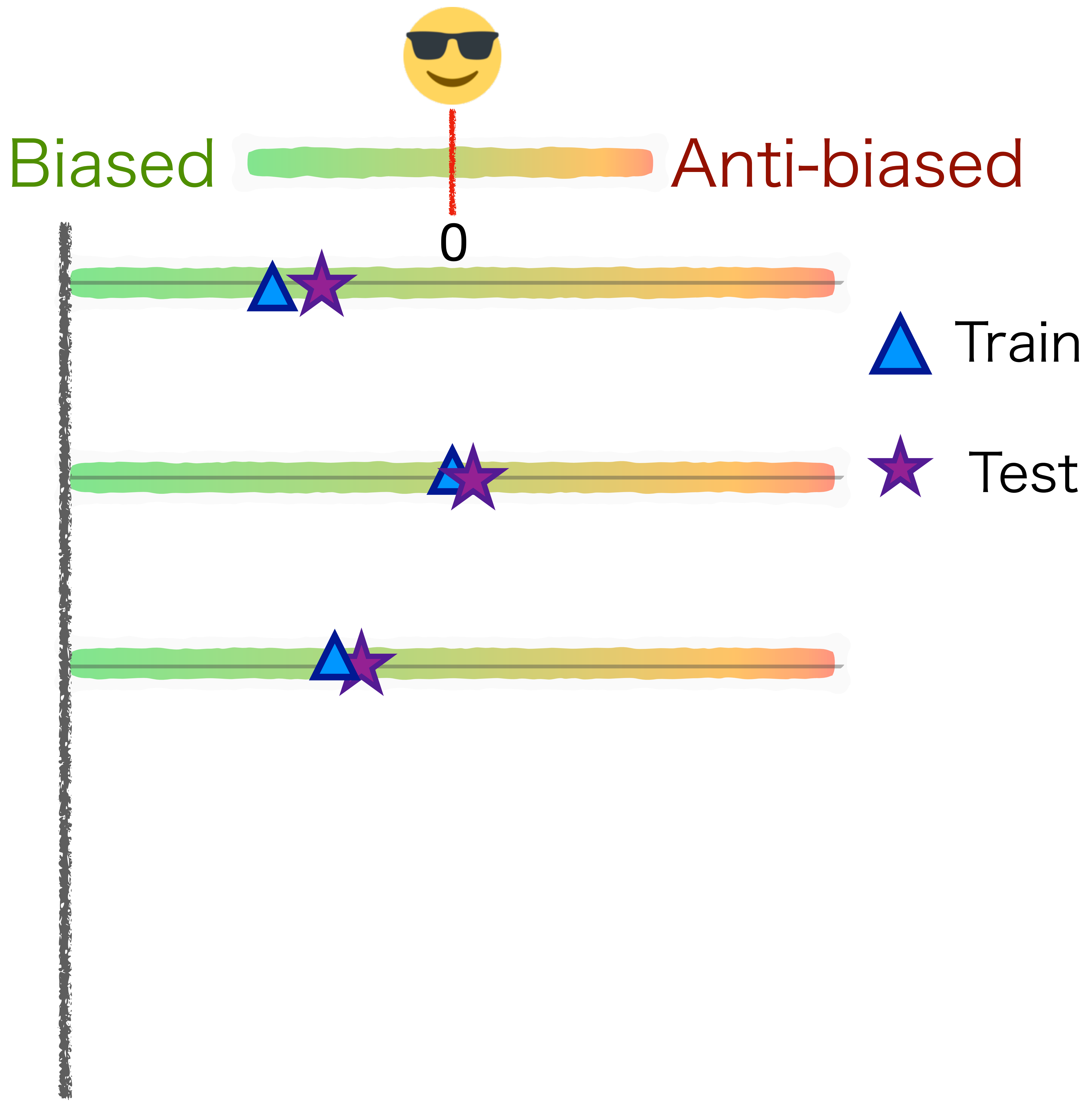
Context	Answer
$\emptyset$	Joe Biden
The year 2019	Donald Trump
The West Wing, season 1	Josiah “Jed” Bartlet

Is dataset balancing the right way forward?

Filtered sets:

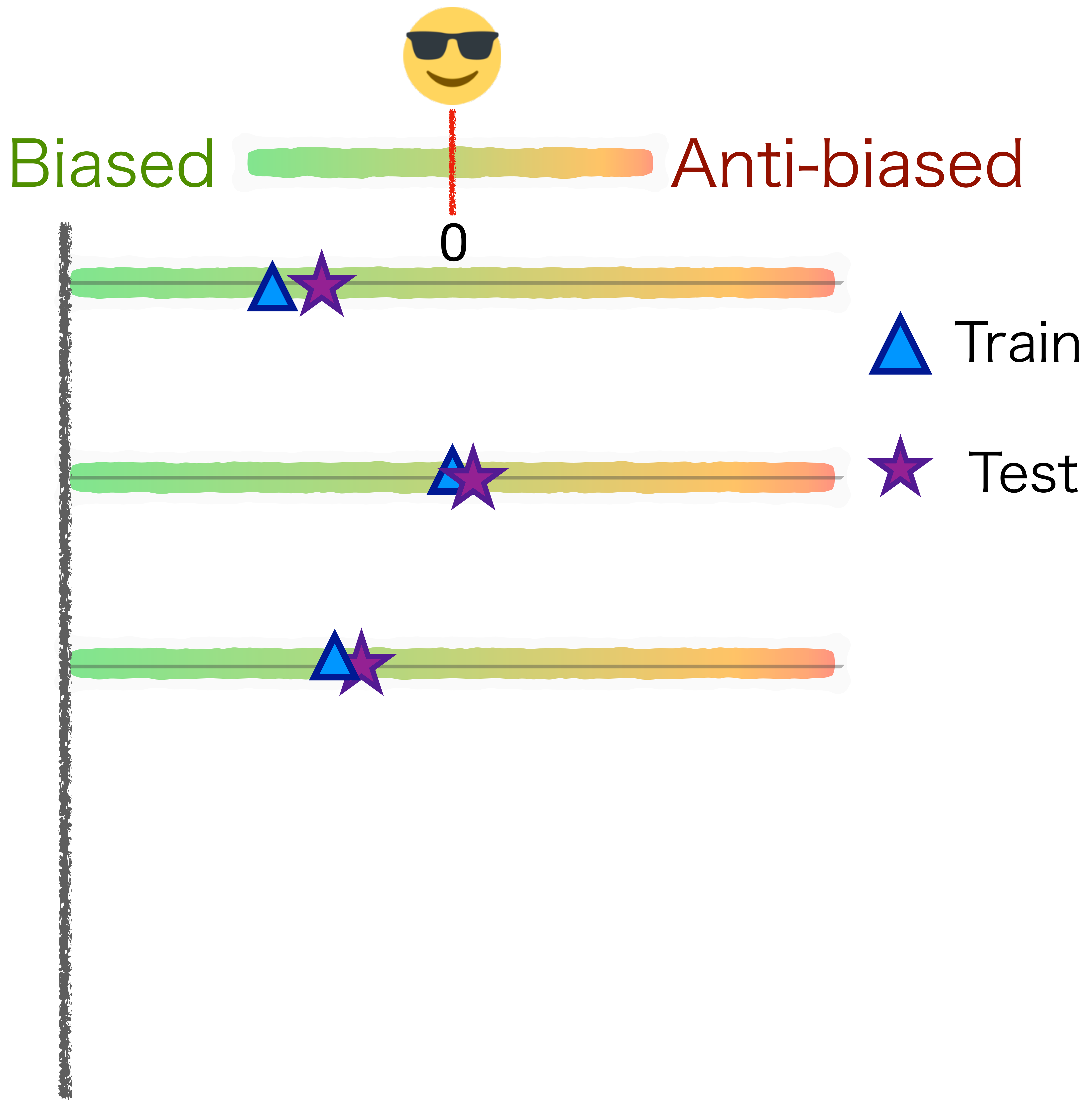
Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Mitigating Spurious Correlations

- Modify the **model**
  - Adversarial networks (Belinkov et al., 2019; Grand and Belinkov, 2019; Wang et al., 2019; Cadene et al., 2019)
  - Model ensembles (Clark et al., 2019,2020; He et al., 2019; Bahng et al., 2020)
- Integrate **causality** into our models
  - Eisenstein (2022); Joshi et al. (2022)
- Build better **benchmarks**

Challenge Sets

- Challenge dataset (aka *adversarial datasets*) intentionally aim to mislead the model
  - The goal is to uncover specific model weaknesses

HANS

McCoy et al. (2019)

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Challenge Sets

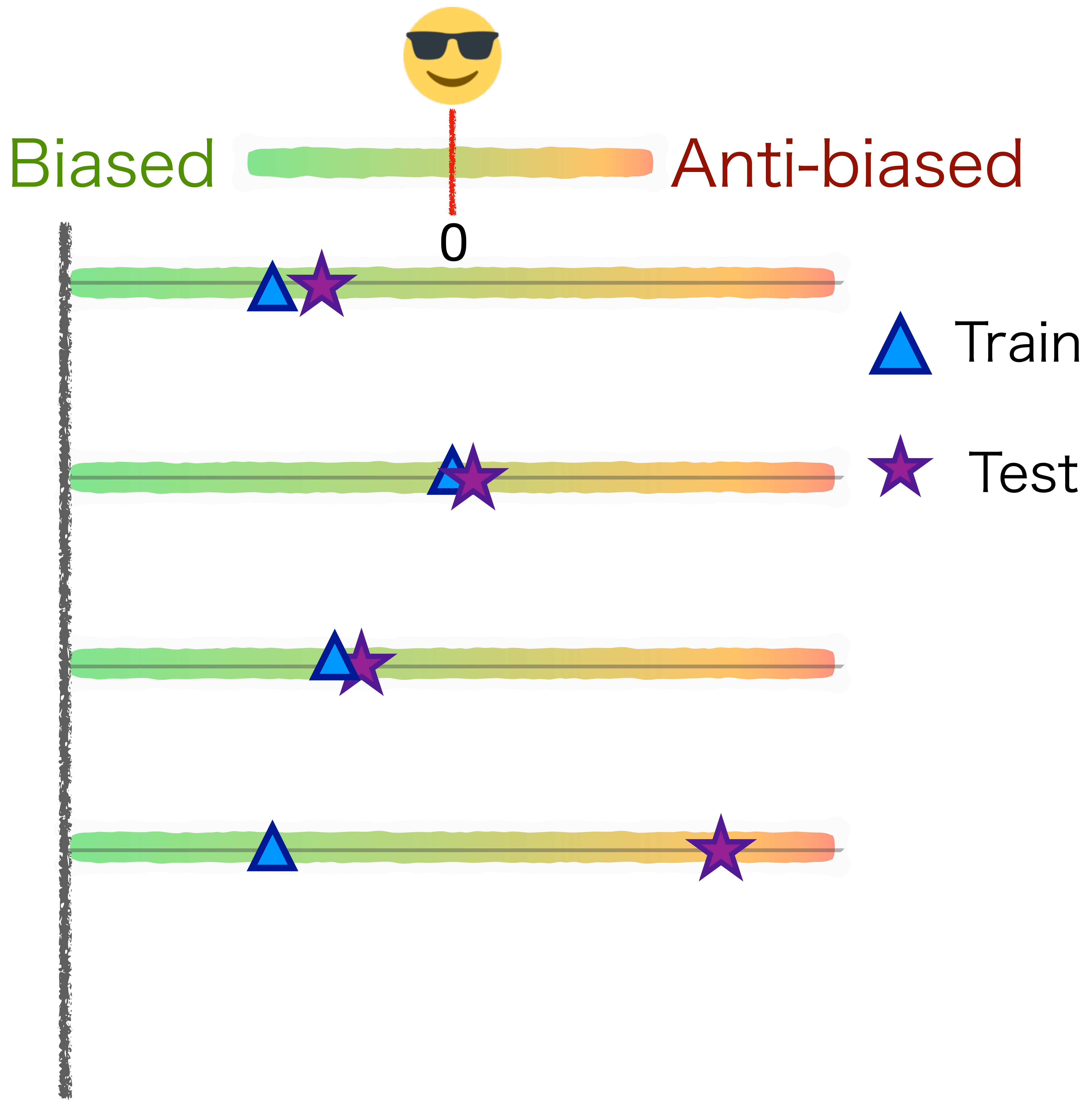
- Test various Types of Capabilities
  - Shift in distribution
  - Ignoring noise
  - Handling misspellings
  - Handling negation
  - Handling temporal modifications

- Typically **manually created**

- Applied to a Range of NLP Tasks
  - NLI
  - (Visual-/)Question answering
  - Machine Translation
  - Text classification
  - …

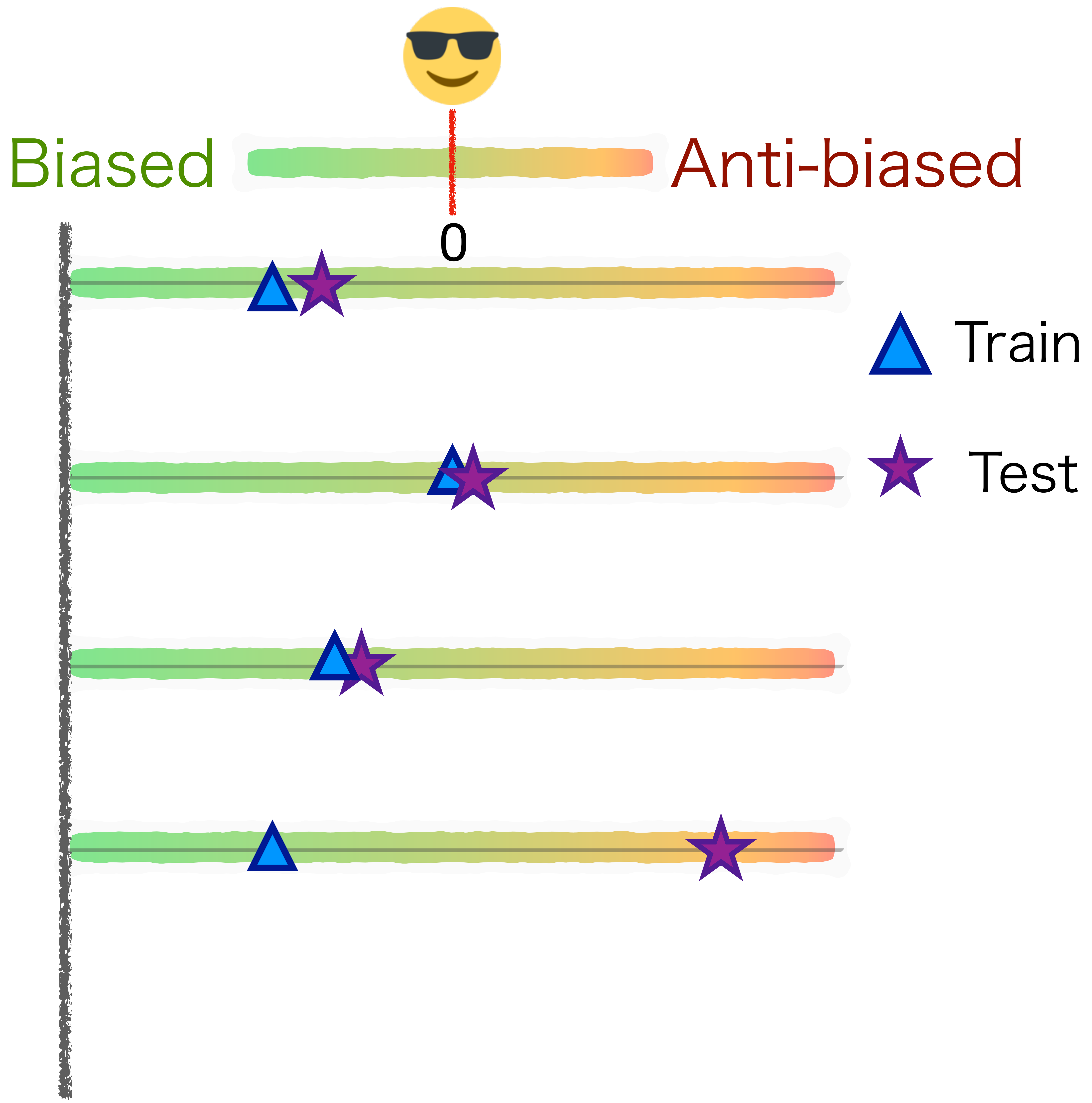
Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Outline

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases



Fight Bias with Bias

Reif & S. (Findings of ACL 2023)

- Balancing only hides the problem
    - Some biases remain hidden in the data
- We want models that are robust to such biases
- Let's ***amplify*** the biases in the data

Amplify Biases???

- Could we ever create datasets that don't contain exploitable biases?
  - Linzen et al. (2020); S. & Stanovsky (2022)
- Biases "hide" in hard, filtered training sets
  ⇒ Harder to evaluate impact on models
- Datasets with amplified biases will create a better testbed to develop methods for *mitigating them*

Don’t Filter, Amplify

Bias-amplified Splits: *Biased* Training, *Anti-biased* Test

Discussion

- I am not sure about the practicality of this setup (easy train and hard test sets) in reality because learning solutions from easy examples only and expecting it to generalize to hard examples is like a dream in ML. Easy examples have heuristics that strong models can easily learn and achieve zero training loss. Then how could we expect them to learn harder patterns?! How can debiasing methods actually help if there are no non-easy sample in the training set?

  I also do not agree with the saying that most models fail - of course they fail, they were only trained on biased data. I'll give an example from gender bias: say that *all* nurses in the world were women. Could you "blame" a model that was trained on such data for being biased? Thus, when you only keep biased samples, it's weird to say that it fails to generalize, because during training there really isn't any difference between the "spurious" features and "robust" features. The paper also doesn't propose (as possible directions) ways of solutions: "models should instead be evaluated on datasets with amplified biases, such that only true generalization will result in high performance" - as I see it, there's not really a way for improving a model trained only on biased samples. Instead, I think we should concentrate on making the models generalize from the little hard examples they do have.

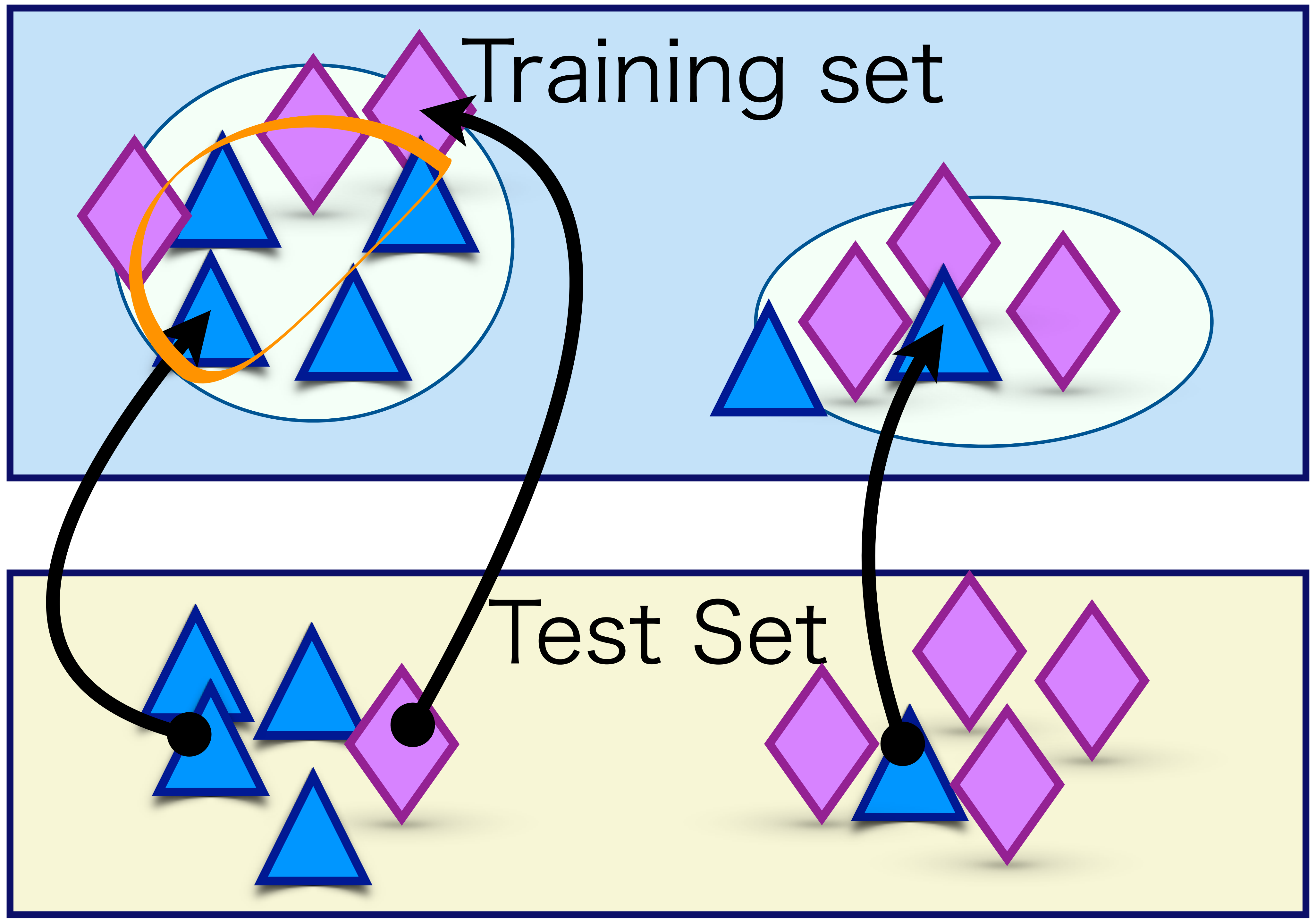
Detour

Definitions of *Biased* and *Anti-biased*

- Dataset cartography
  - Swayamdipta, S. et al. (2020)
- Partial-input baselines
  - Gururangan, …, S. et al (2018); Poliak et al. (2018)
- Minority examples
  - A method we introduce to detect minority examples

Detecting Minority Examples

- Cluster training set using the model representation
- Detect majority labels within each cluster
    - Use them as our new “biased” training set
- Deduce test set minority examples by nearest neighbor in the training set
    - Use them as our new test set



Results

MultiNLI; RoBERTa-large

- Most validation data is **biased**
- Automatic challenge sets are **hard**
- Bias amplification makes data harder
- Automatic challenge sets are as hard as HANS

Train	Val.	Cart.	ParIn.	Mino.	HANS
full	$90.4_{0.2}$	$59.9_{0.7}$	$79.7_{0.6}$	$71.9_{0.3}$	$78.2_{0.5}$
biased	$88.4_{0.7}$	$51.7_{0.5}$	$68.2_{0.3}$	$50.5_{1.2}$	$51.4_{0.4}$

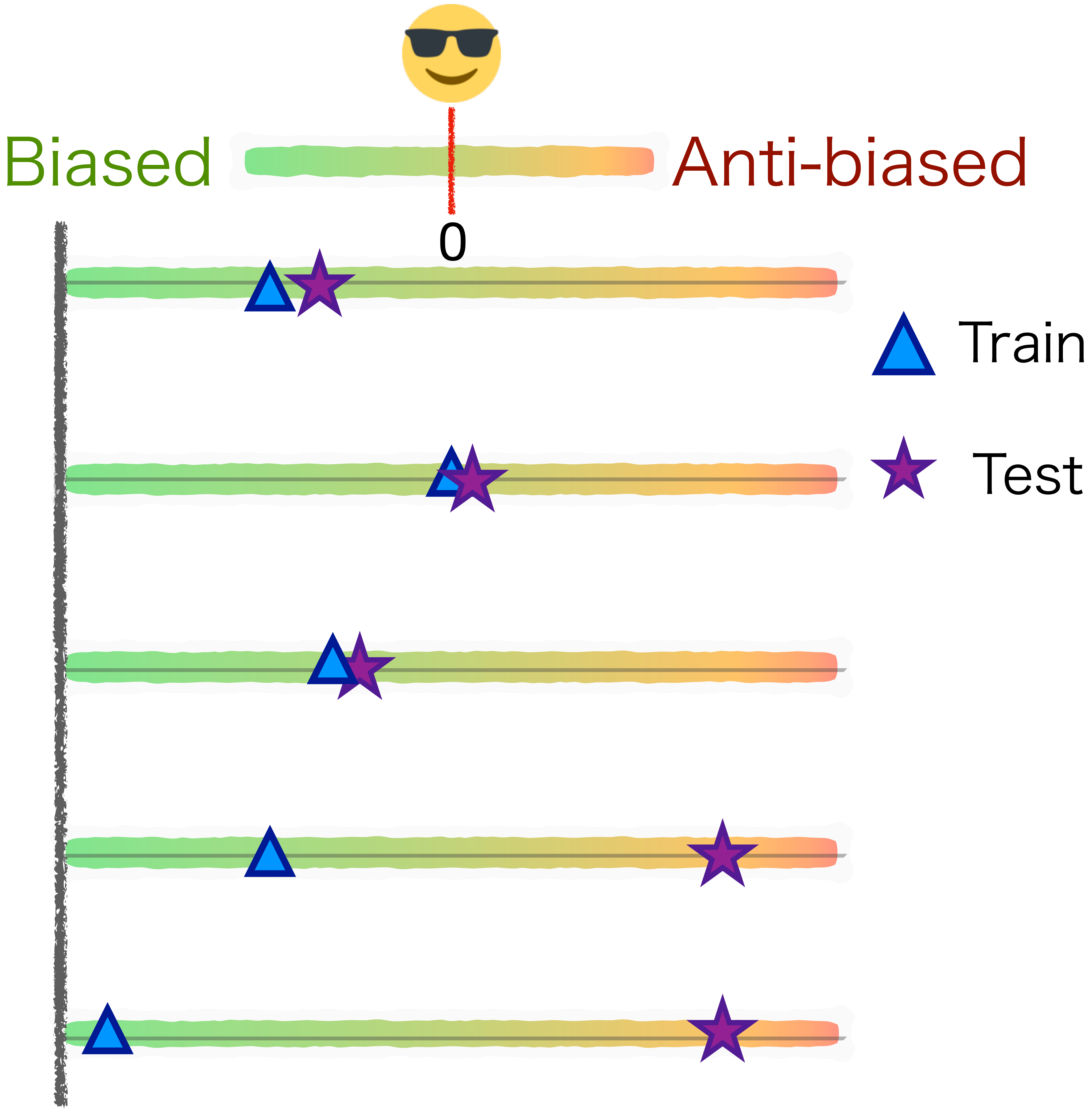
Test

What about LLMs?

- The web is biased too!
  - Birhane et al., 2021; Dodge et al., 2021
- Robustness is a major issue in LLMs too
  - Liu et al. (2021); Lu et al. (2022); Maus et al. (2023)
- Balancing is even less practical there
- We need **robust modeling**!

Summary

- Models are not robust
  - **Spurious correlations** in NLP datasets
- Fixing the training set
  - **Balancing** and **filtering**
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Changing the test set
  - **Challenge/adversarial** sets
- A new evaluation framework
  - **Amplified** biases

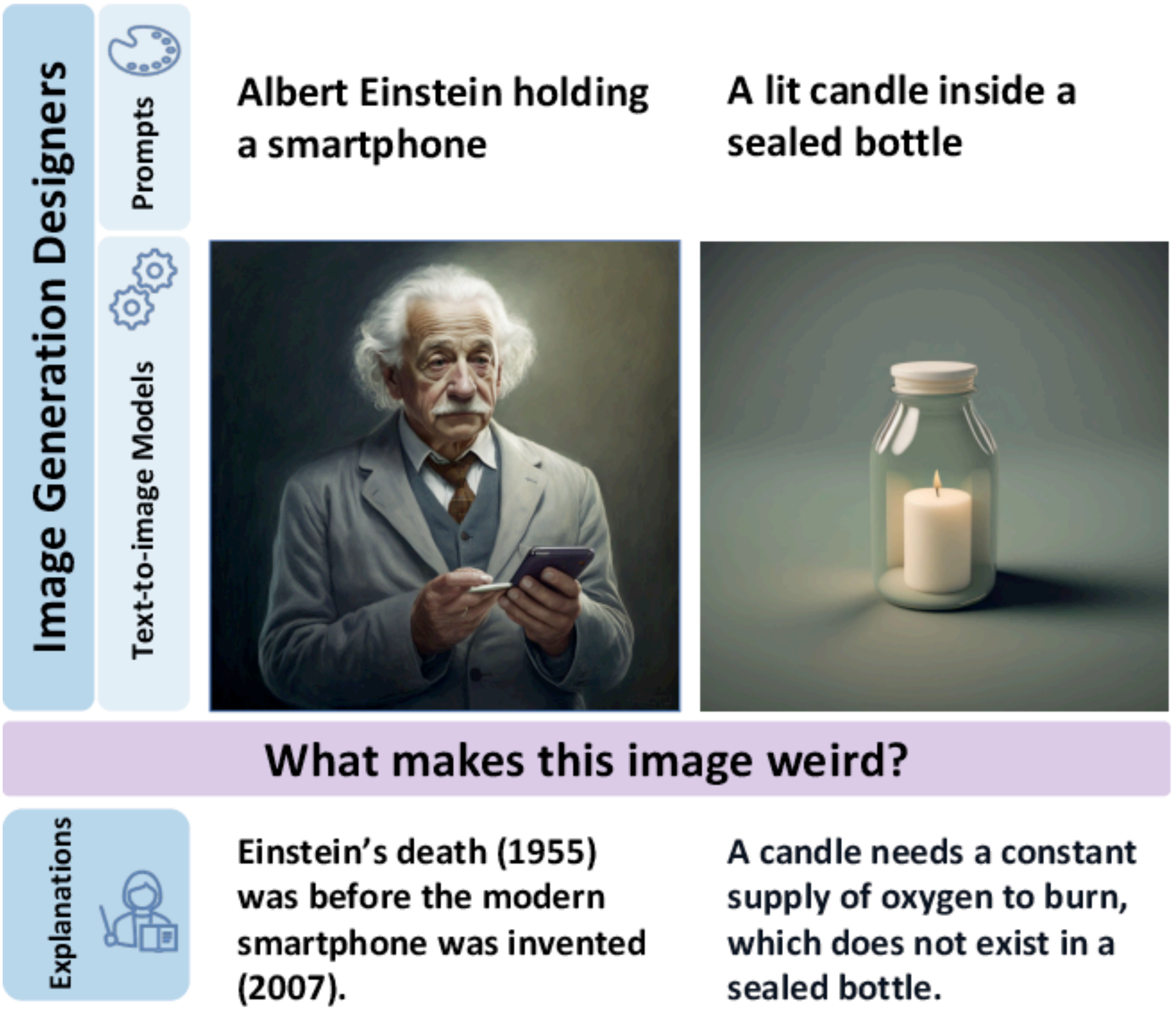


Detour

WHOOPS!

Bitton-Guetta, Bitton, … , S. (ICCV 2023)

- A dataset of “weird” images
  - Generated by designers using image generation tools
- Humans both
  - **Easily understand** what’s going on in the image
  - Can **generate explanations** of what’s weird in the image
  - Machines do much poorly



Back