Spurious Correlations: Challenges, Solutions, and Opportunities

Roy Schwartz
The Hebrew University of Jerusalem
CLunch Seminar, UPenn, 03/2023

THE HEBREW UNIVERSITY OF JERUSALEM

Motivation

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)

Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
    - Practical and conceptual limitations
- Alternatives to **dataset balancing**
    - Fight bias with bias

Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge**/**adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
    - Practical and conceptual limitations
- Alternatives to **dataset balancing**
    - Fight bias with bias

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%

Natural Language Inference (NLI)

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

SNLI (Bowman et al., 2015); MNLI (Williams et al., 2018)

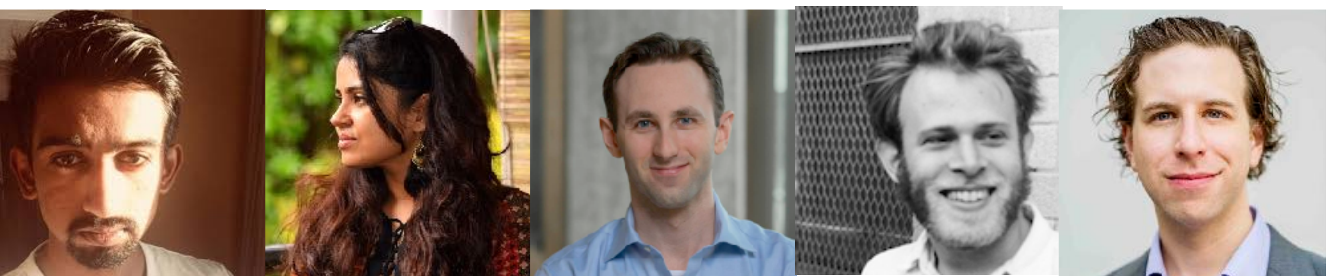
Spurious Correlations in NLI Datasets

Gururangan, Swaymdipta, Levy, S., Bowman, Smith (2018); Poliak et al. (2018); Tsuchiya (2018)

- Train a hypothesis-only classifier
  - No premise

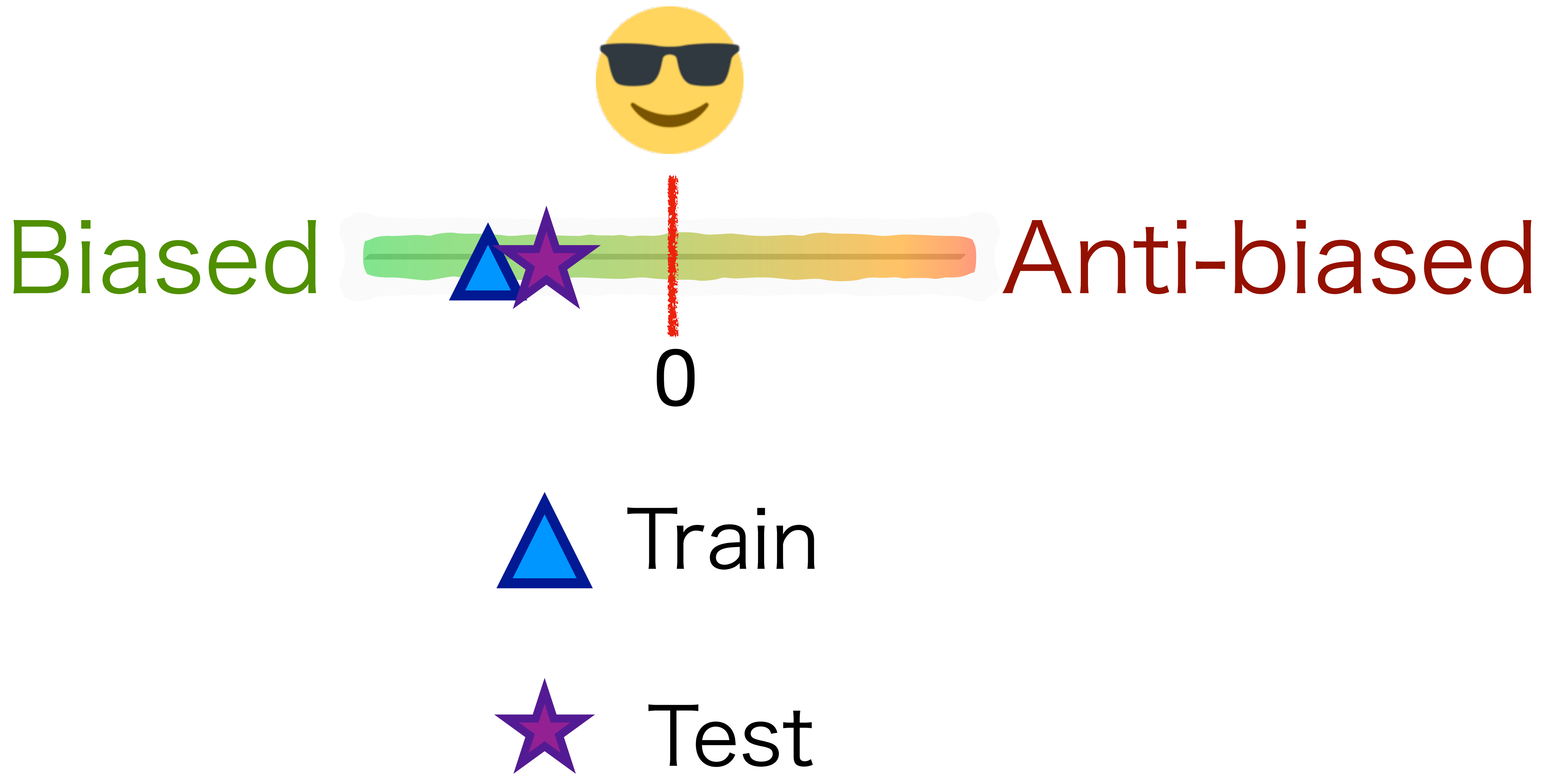
Model	SNLI	MultiNLI Matched	MultiNLI Mismatched
majority class	34.3	35.4	35.2
`fastText`	**67.0**	**53.9**	**52.3**

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%



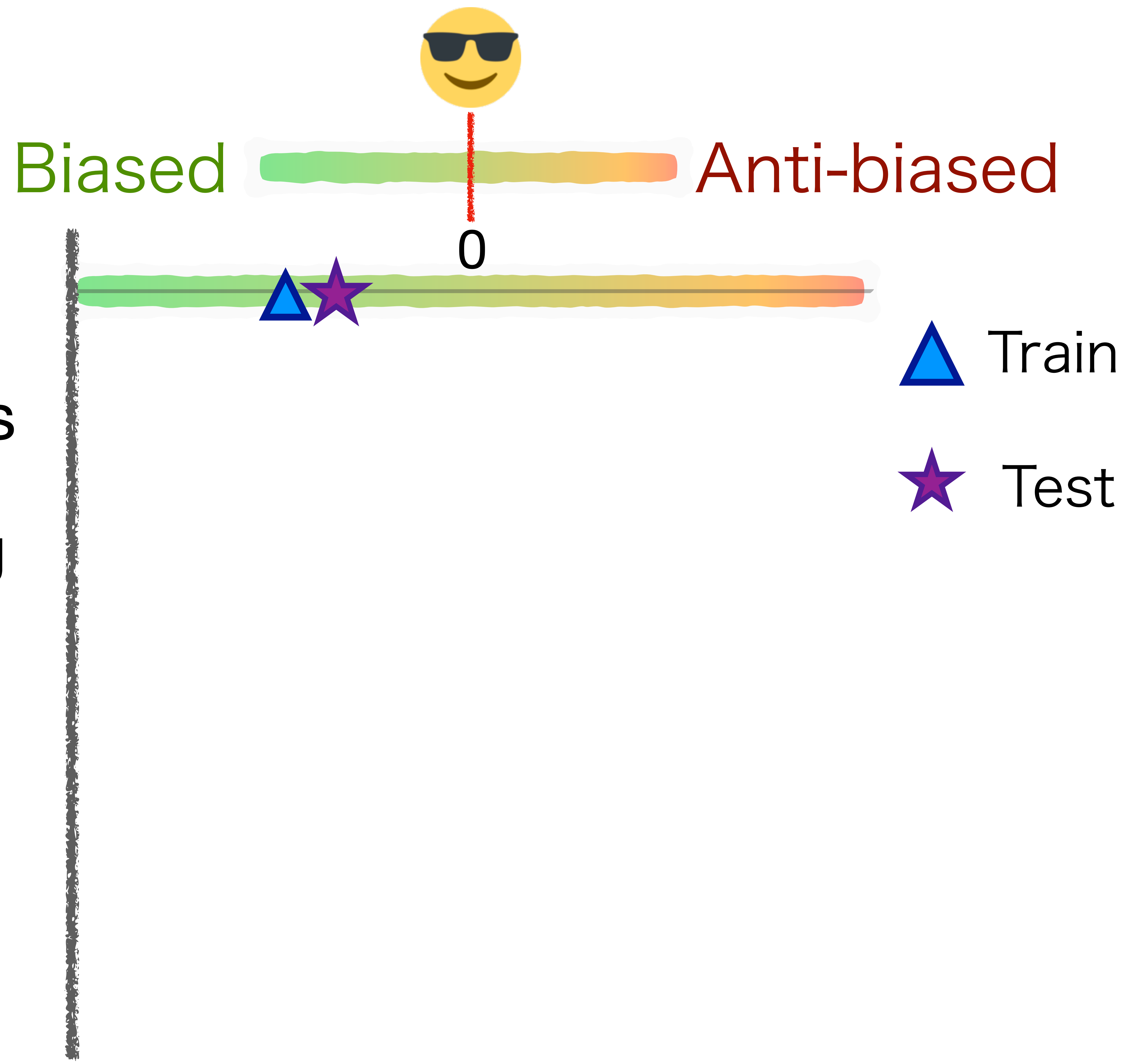
Other Spurious Correlations

- Other tasks
  - Question answering (Kaushik & Lipton, 2018)
  - Winograd Schema (Elazar et al., 2021
  - ...



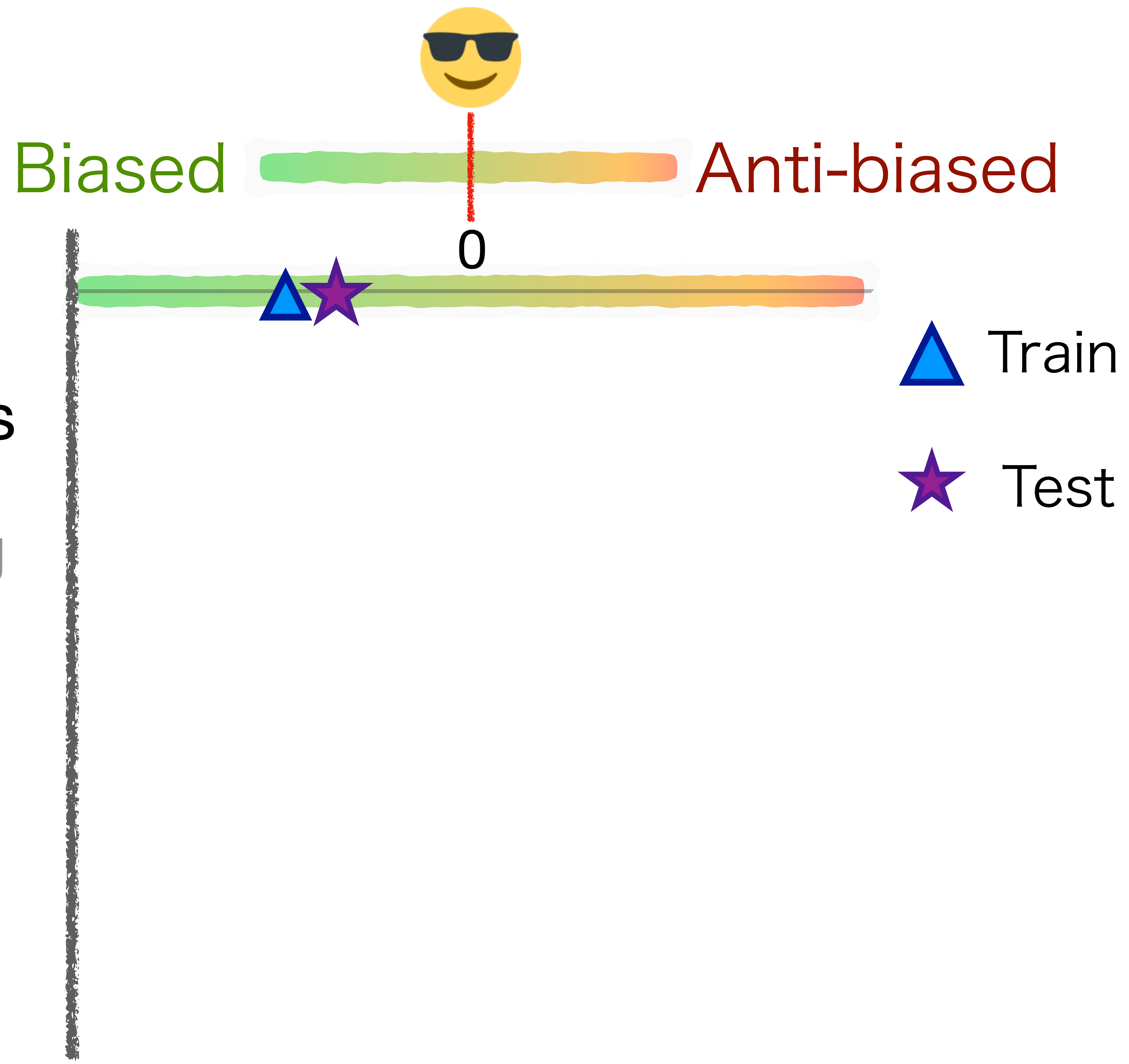
Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
  - Practical and conceptual limitations
- Alternatives to **dataset balancing**
  - Fight bias with bias



Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
    - Practical and conceptual limitations
- Alternatives to **dataset balancing**
    - Fight bias with bias



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)
- **Modify the data**

Challenge Sets

- NLP models are very sensitive to their training domain
- Testing a model on a different distribution often leads to reduced performance
  - Fixing this problem is one of the key challenges in NLP and AI in general
- Challenge dataset (aka *adversarial datasets*) intentionally aim to mislead the model
  - The goal is to uncover specific model weaknesses

Adversarial SQuAD

Jia et al. (2017)

SQuAD1.1 Leaderboard (Rajpurkar et al., 2016)

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.202**	**95.379**
2 May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
3 Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903

Article: Super Bowl 50
Paragraph: *"Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.*

Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*

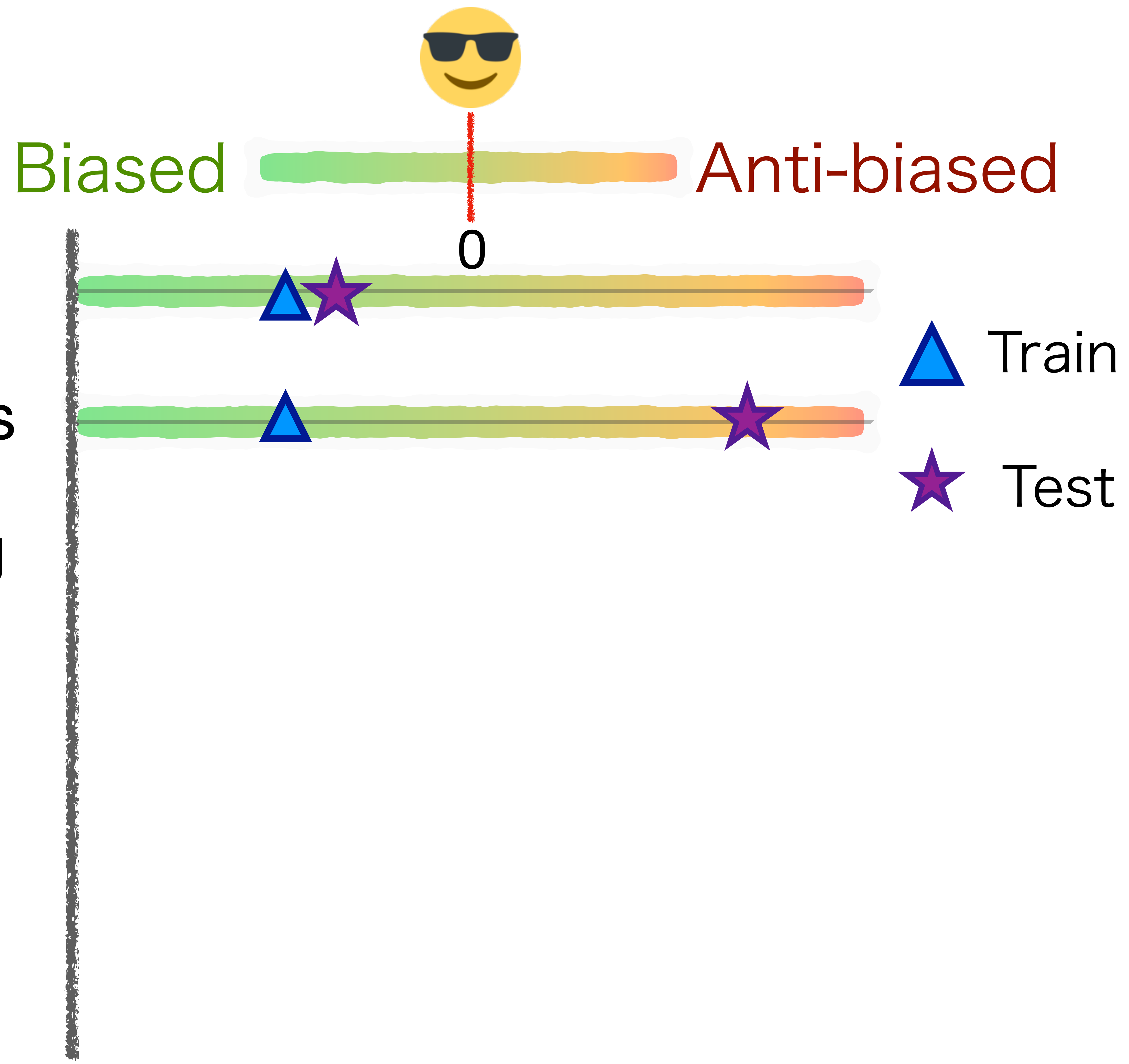
Challenge Sets

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

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

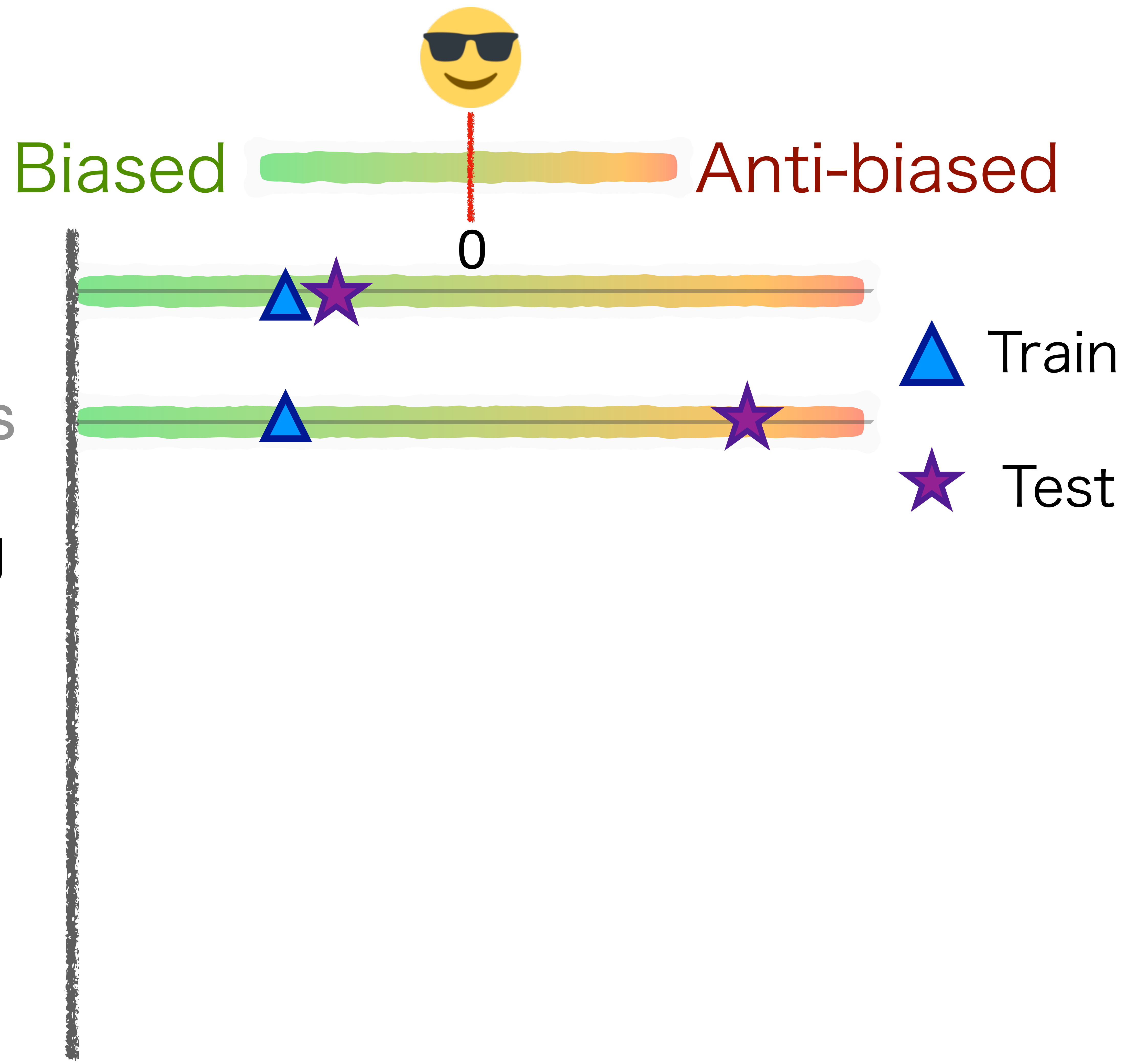
Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
    - Practical and conceptual limitations
- Alternatives to **dataset balancing**
    - Fight bias with bias



Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
    - Practical and conceptual limitations
- Alternatives to **dataset balancing**
    - Fight bias with bias



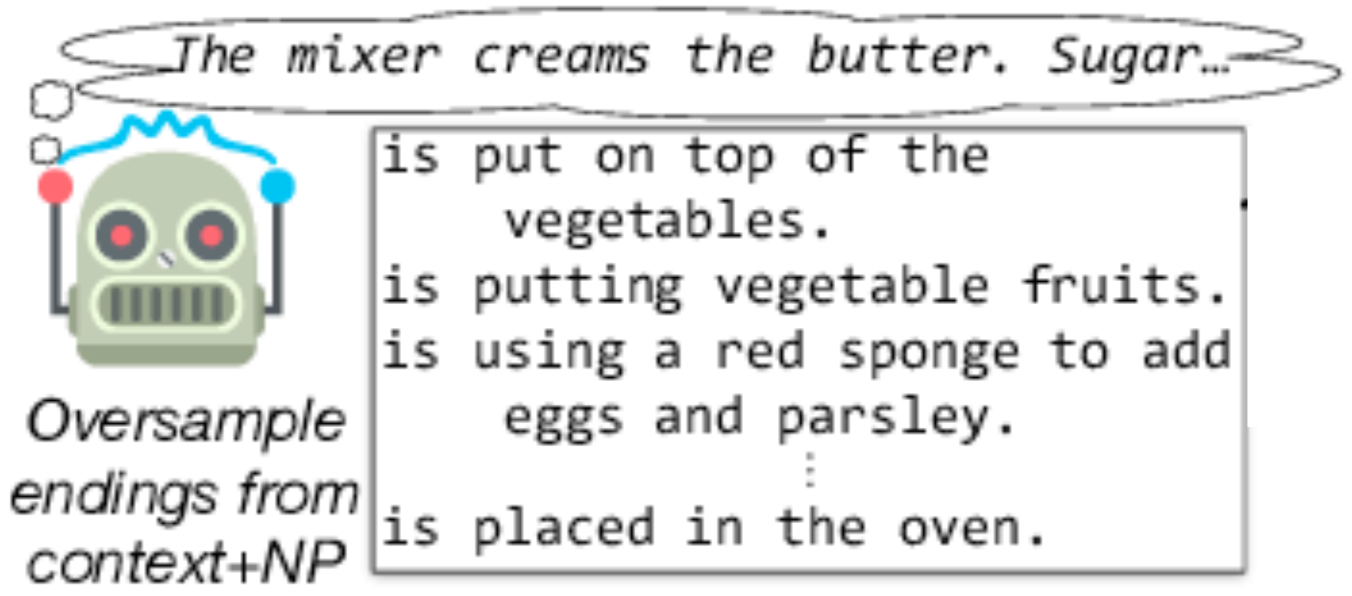
Dataset Balancing
Augmentation

- 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)

Adversarial Filtering (AF)

Zellers, Bisk, S. & Choi (2018)

- A multi-choice setting
  - Assume a human-generated input passage
- An LM generates *many* possible continuations
- A discriminator trained to identify the machine-generated options
- Iteratively until convergence:
  - Select easily-identifiable options
  - Replace them with other (harder) options
- Validate resulting data with human experts



Adversarial Filters of Dataset Biases (AFLite)

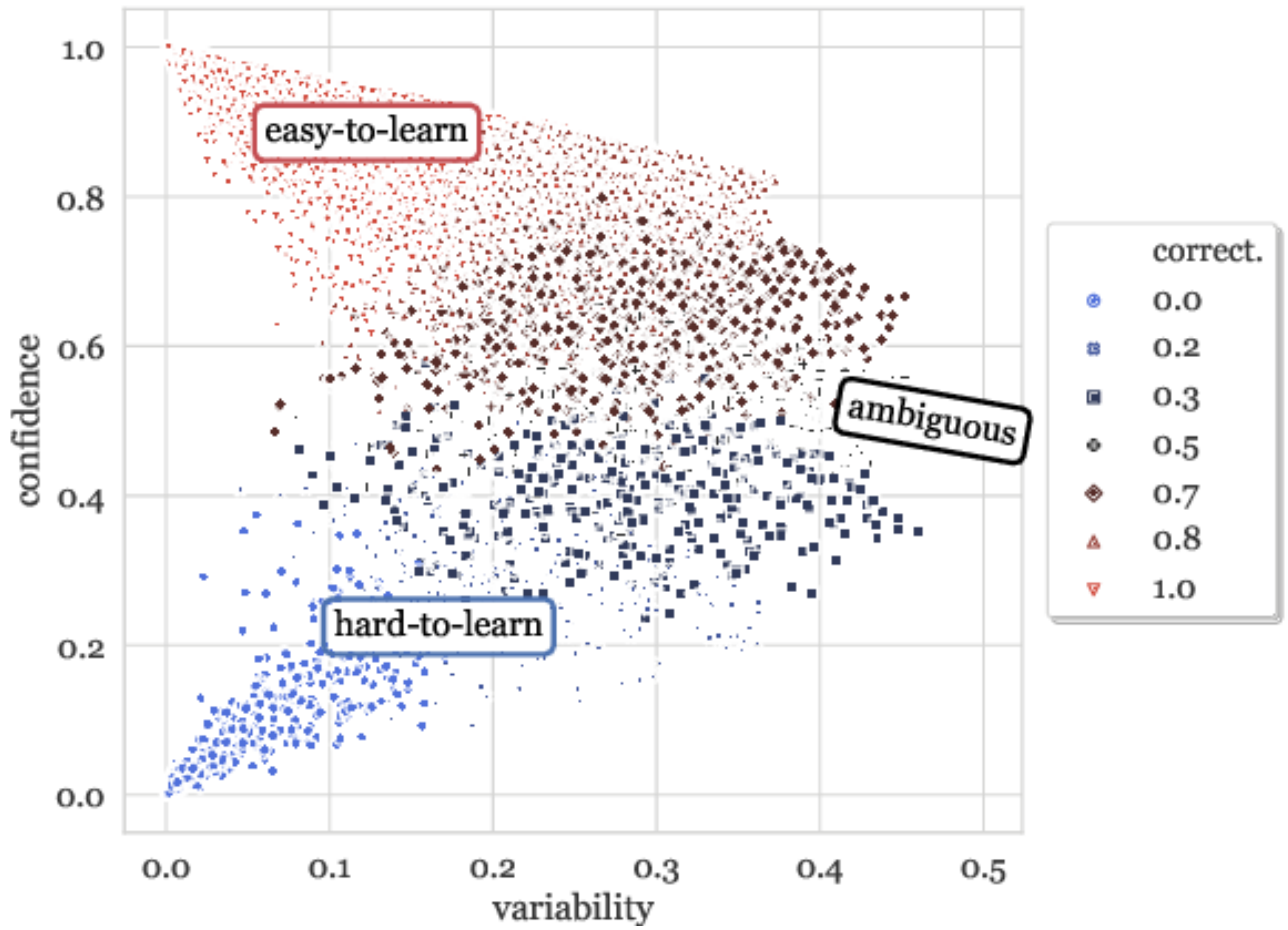
<u>Sakaguchi et al. (2020)</u>

- Start from a collected dataset *D*
- Iteratively
  - Randomly break *D* into *n* different train/test splits
  - Train a classifier on each training split
  - Filter out the instances that are solved by most models
- Return filtered dataset

Dataset Cartography

Swayamdipta, S. et al. (2020)

- Identify different regions in datasets
- Most examples are *easy-to-learn*
- Training on the most ambiguous examples leads to **better generalization**



Filtering is Widely Adopted

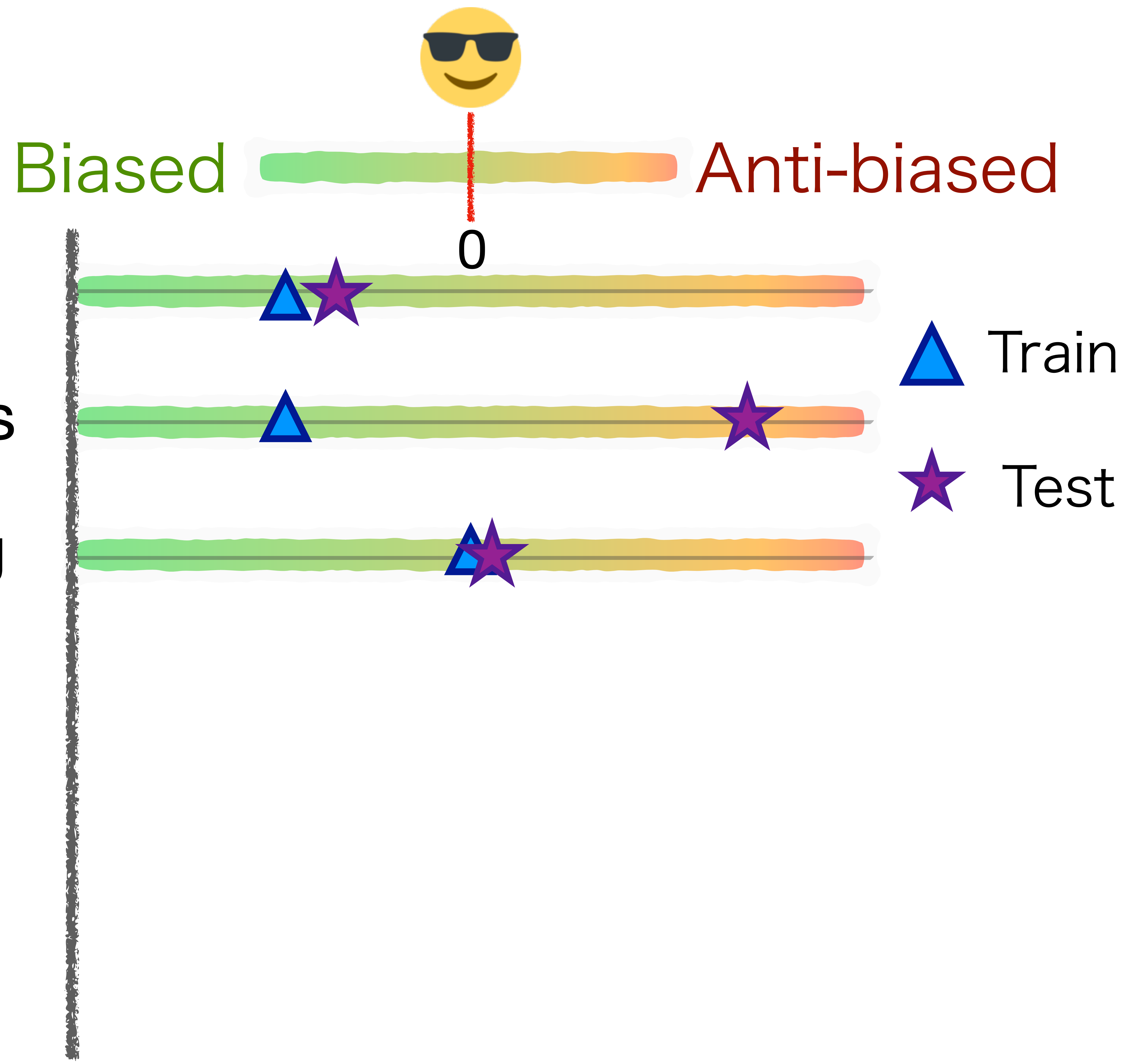
- Record (Zhang et al., 2018)
- DROP (Dua et al., 2019)
- HellaSWAG (Zellers et al., 2019)
- αNLI (Bhagavatula et al., 2019)
- WinoGrande (Sakaguchi et al., 2020)
- …

Filtering as Balancing

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

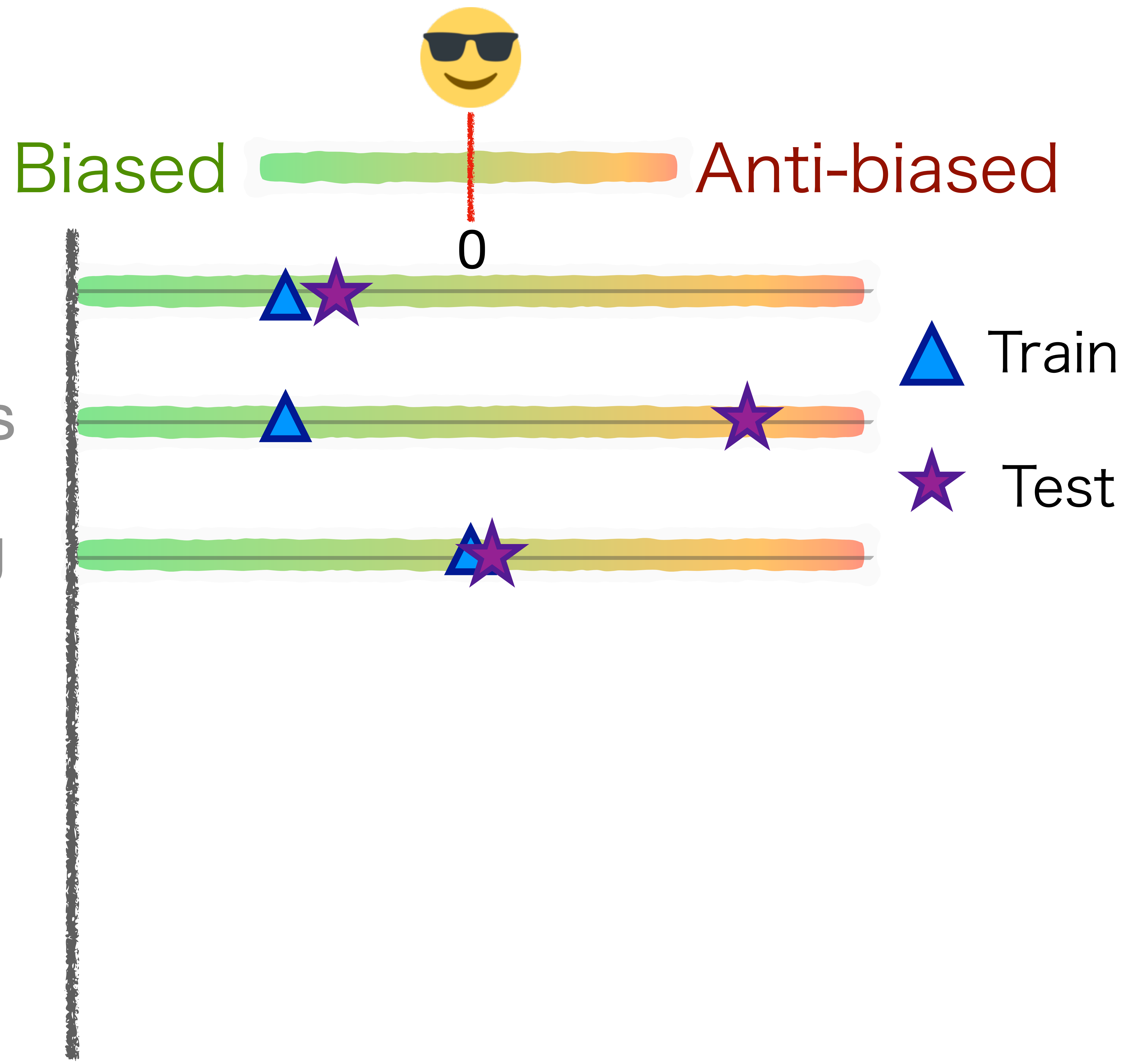
Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
  - Practical and conceptual limitations
- Alternatives to **dataset balancing**
  - Fight bias with bias



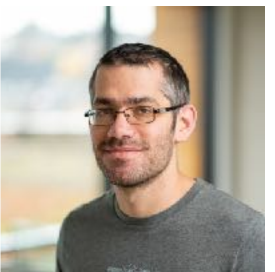
Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
  - Practical and conceptual limitations
- Alternatives to **dataset balancing**
  - Fight bias with bias



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



Everything 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
  - $\forall x_i, y \in Y, p(y | x_i) = \frac{1}{|Y|}$

The Balancing Approach

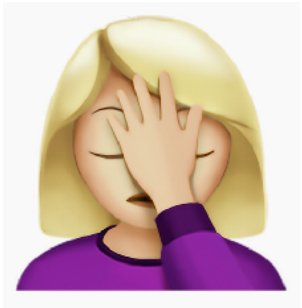
- Gardner et al. (2021):
  - For each feature *f*:
    - if (*f* contains information):
      - => *f* can be exploited

- Balancing/Filtering:
  - => To avoid exploitation, for each feature *f*, **eliminate information** in *f*

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

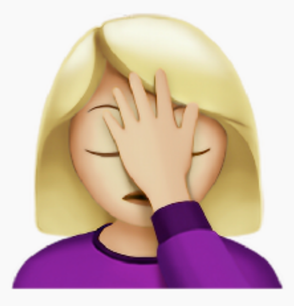
Natural Language

- 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?

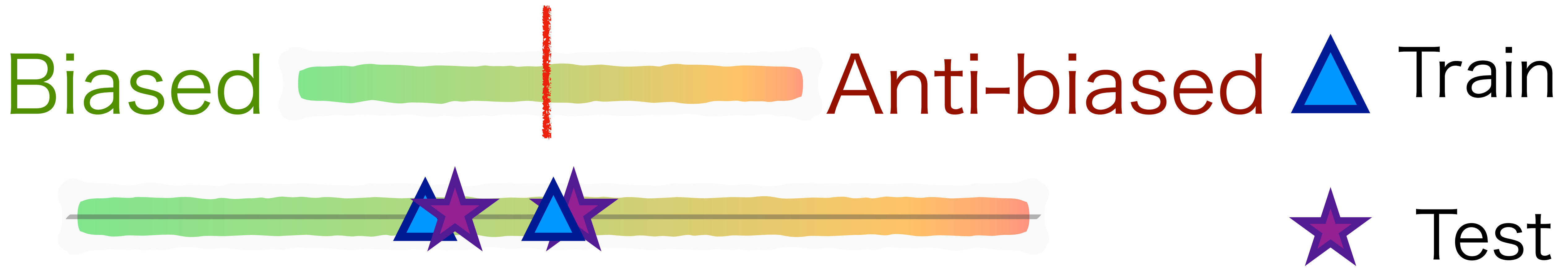
- Especially relevant for world knowledge and common-sense knowledge
  - Joe Biden is the president of the US
  - A person is typically happy when they receive a present
- As a result, dataset balancing is **undesired**

Who is the president of the U.S.?

Context	**Answer**
∅	Joe Biden
The year 2019	Donald Trump
The West Wing, season 1	Josiah "Jed" Bartlet

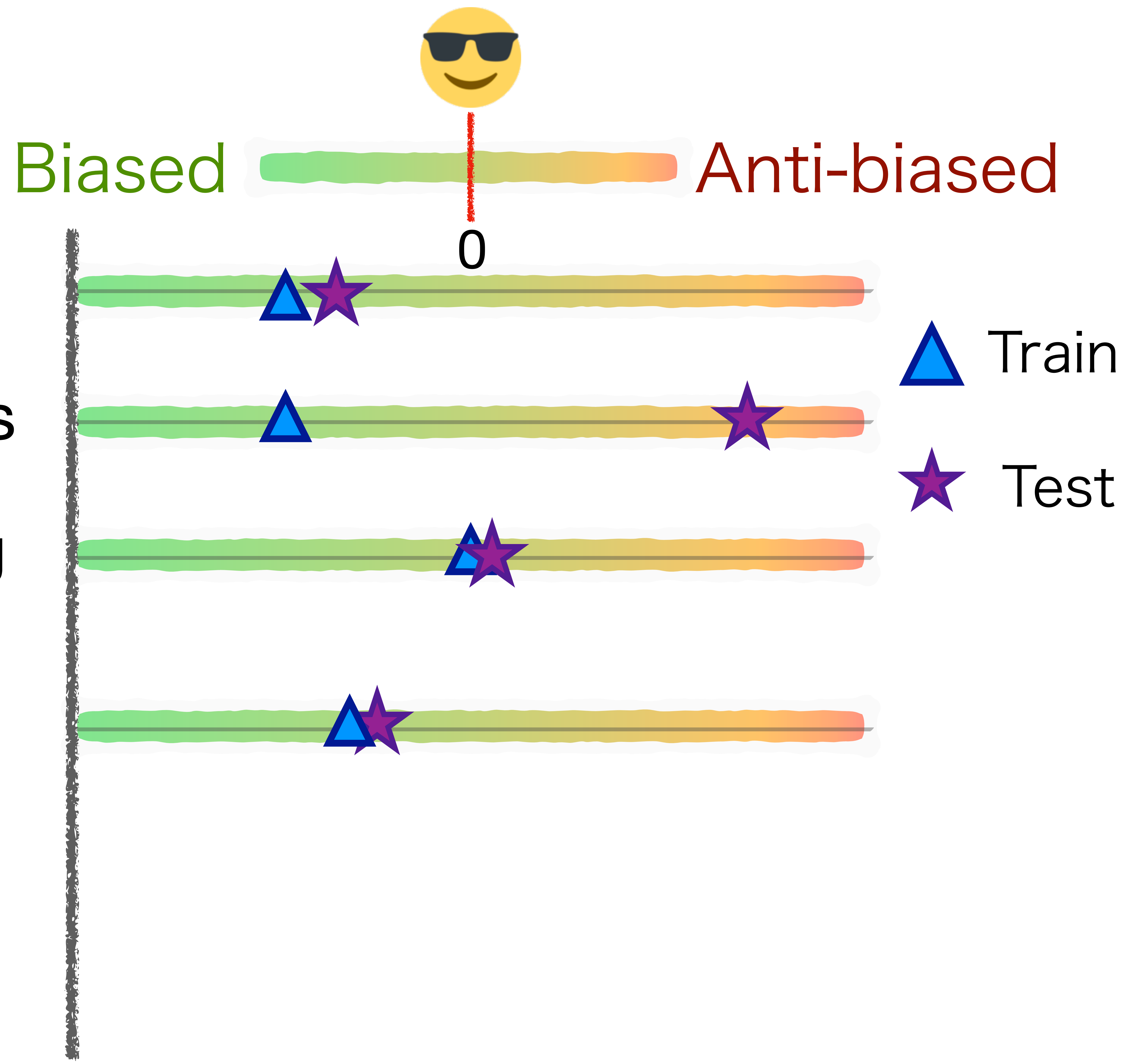
Is dataset balancing the right way forward?

Filtered sets:



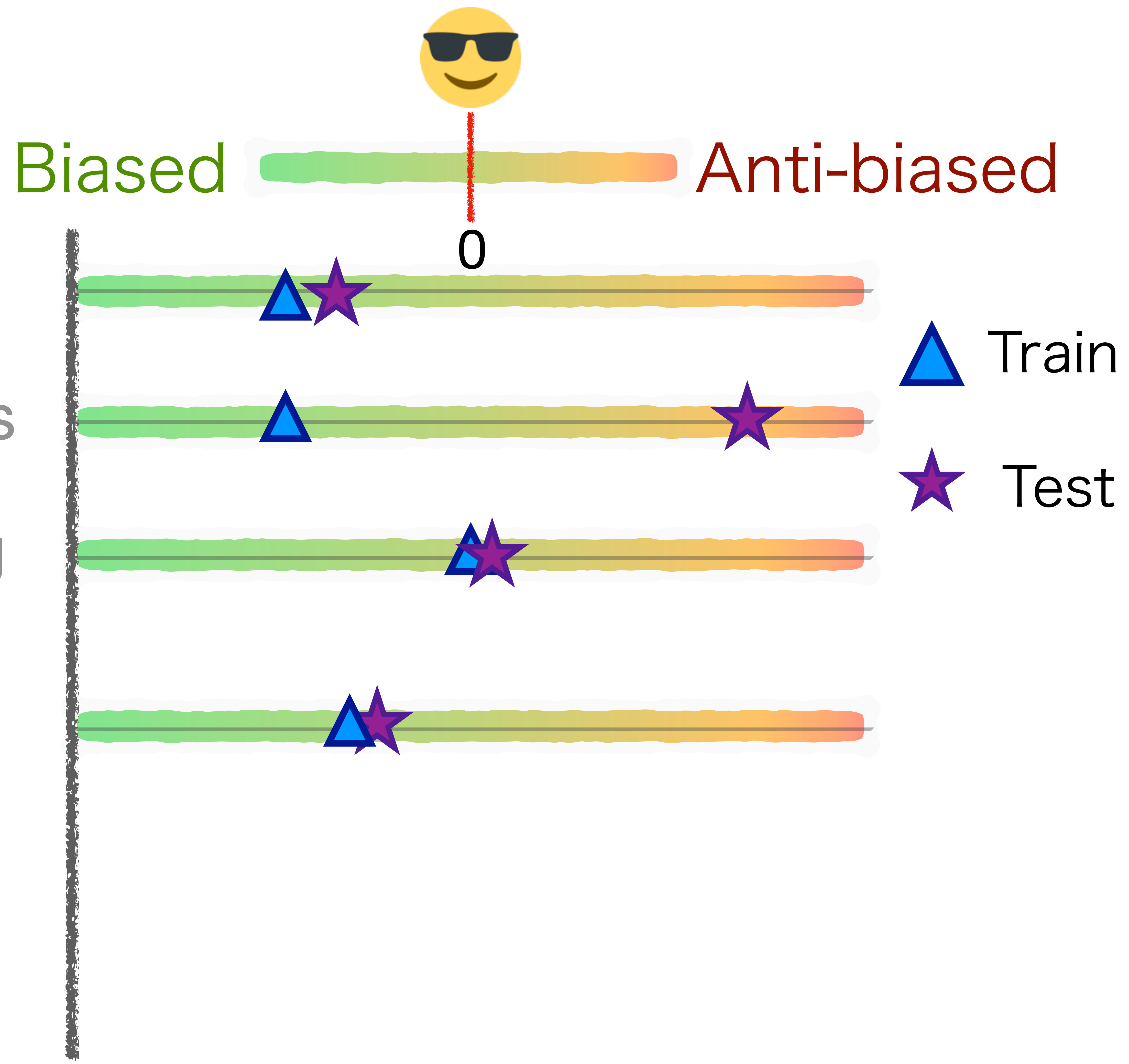
Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
  - Practical and conceptual limitations
- Alternatives to **dataset balancing**
  - Fight bias with bias



Outline

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
    - Practical and conceptual limitations
- Alternatives to **dataset balancing**
    - Fight bias with bias



Suggested Alternatives

- Instead of balancing, augment datasets with *richer contexts*
- Instead of a closed label set, support *abstention/interaction*
- Instead of large-scale fine-tuning, move to *few-shot learning*

*How can we encourage the development of models **robust** to spurious correlations?*

Fight Bias with Bias

Reif & S. (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

Definitions of *Biased* and *Anti-biased*

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

ACL Reviews

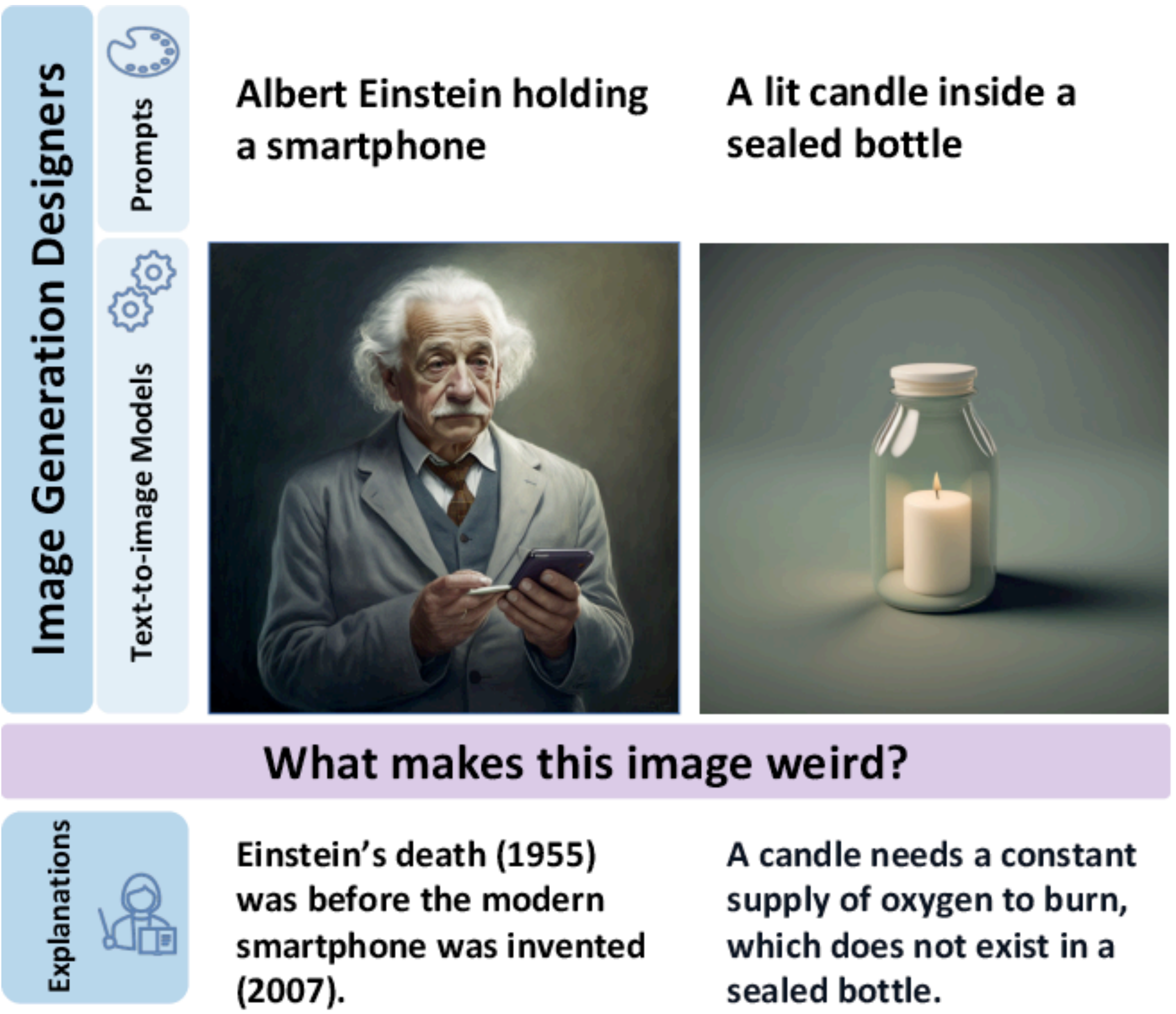
- I am not sure about the practicality of this setup (easy train and hard test sets) in reality because learning solutions from easy examples only and expecting it to generalize to hard examples is like a dream in ML. Easy examples have heuristics that strong models can easily learn and achieve zero 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.

WHOOPS!

A Vision-and-Language Benchmark of Synthetic and Compositional Images
Bitton-Guetta, Bitton, Hassel, Schmidt, Elovici, Stanovsky & S. (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



Results

MultiNLI; RoBERTa-large

- Most validation data is **biased**
  - Training on biased data leads to small differences on standard validation set
- Training on **all data** and testing on **anti-biased data** leads to large performance drops
- Training on **biased data** and testing on **anti-biased data** leads to additional large drops

Train	Val.	Cart.	ParIn	Mino.
full	$90.4_{0.2}$	$59.9_{0.7}$	$79.7_{0.6}$	$71.9_{0.3}$
Biased	$88.4^{*}_{0.7}$	$51.7_{0.5}$	$68.2_{0.3}$	$50.5_{1.2}$

Test

Summary

- **Spurious correlations** in NLP datasets
- Fix the test set: **challenge/adversarial** sets
- Fix the training set: **balancing** and **filtering**
- On the limitations of **dataset balancing**
  - Practical and conceptual limitations
- Alternatives to **dataset balancing**
  - Fight bias with bias