Sprucing up a Dataset:
Adversarially Filtering Dataset Artifacts

Swabha Swayamdipta

Oct 18th, 2019

UWNLP

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

UW Linguistics Colloquium

Linguistics

Linguistics

Machine Learning

Linguistics

Machine Learning

NLP

Linguistics

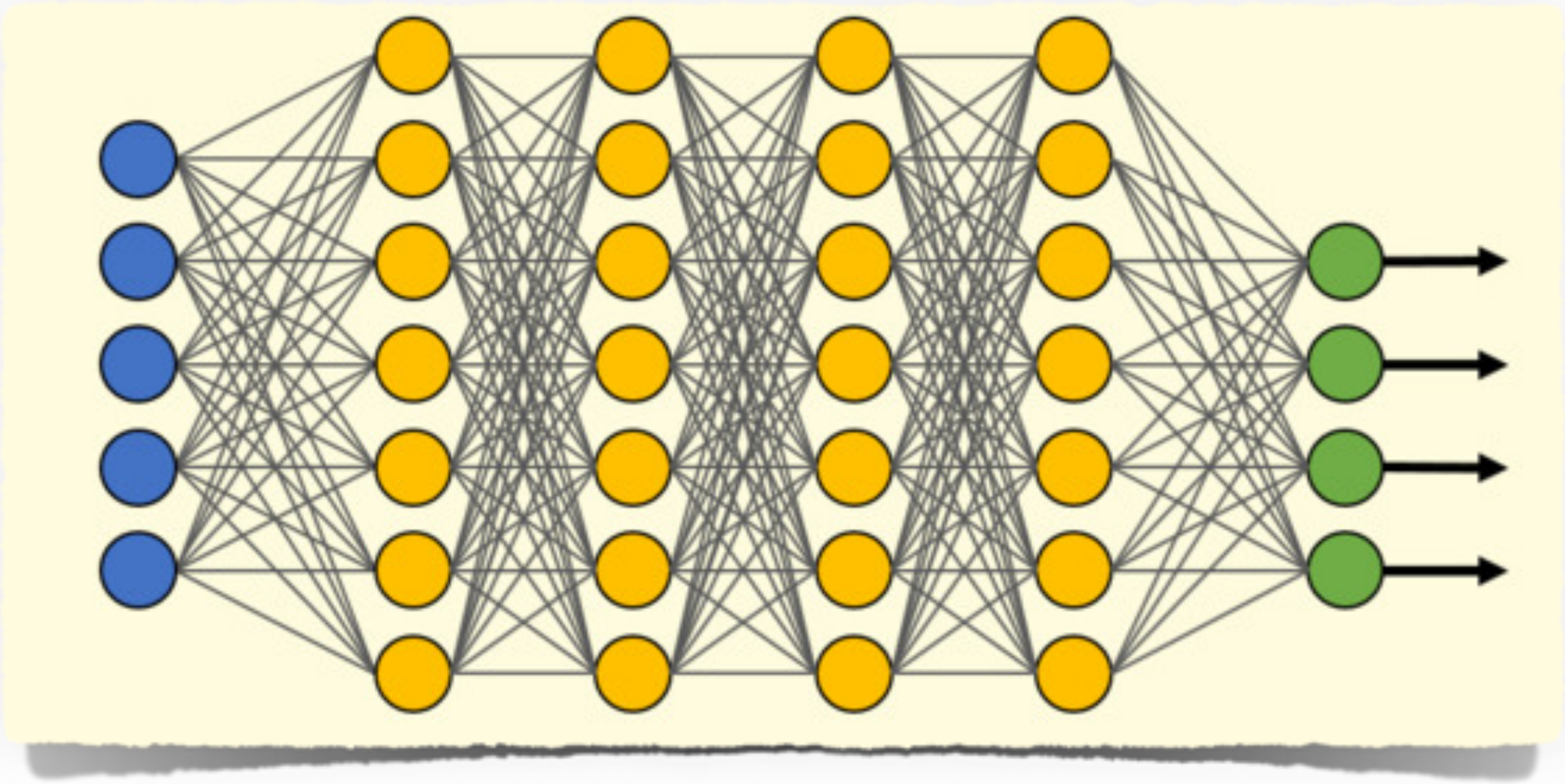
Machine Learning

NLP

Linguistics

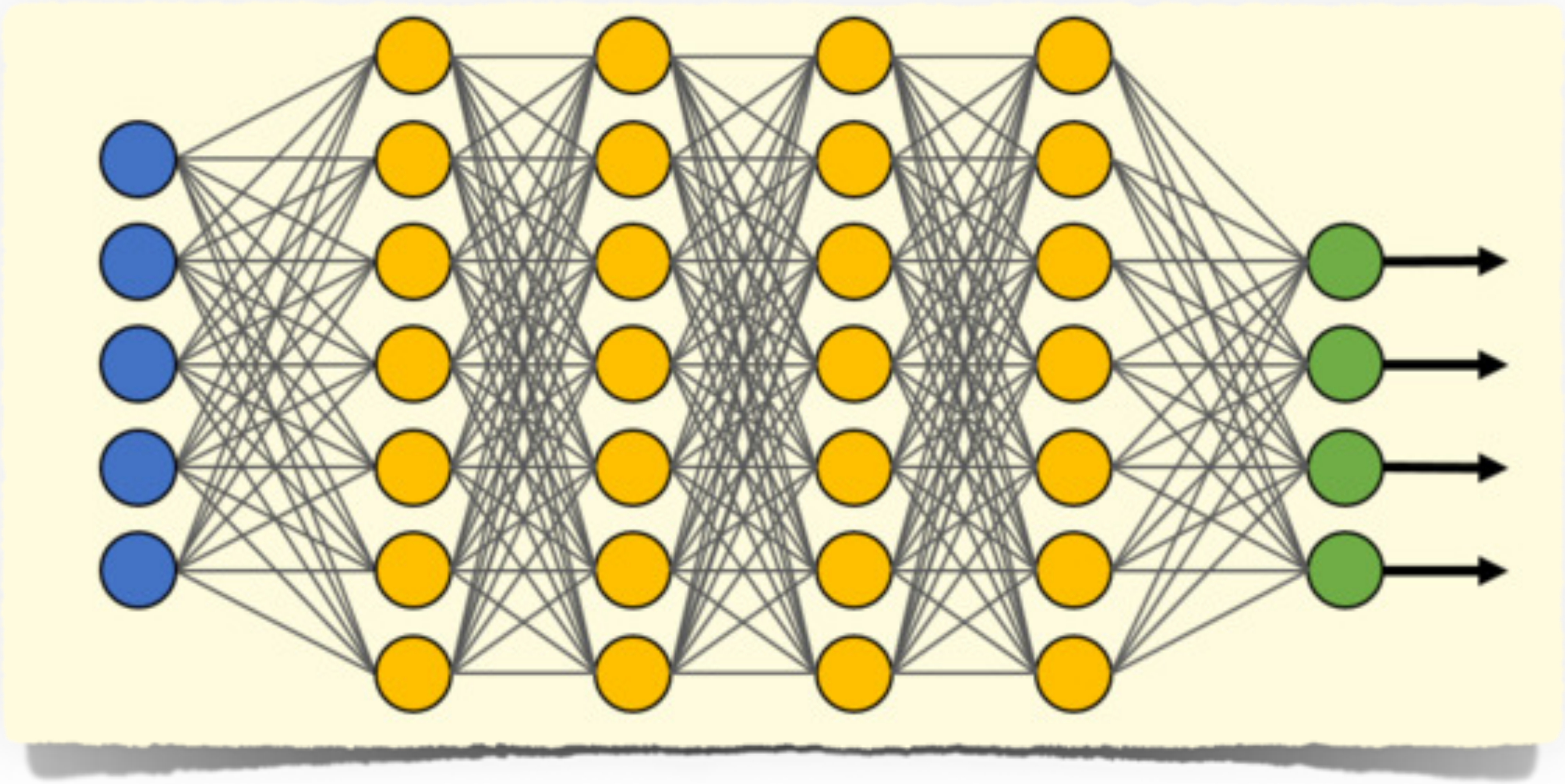
Machine Learning

NLP

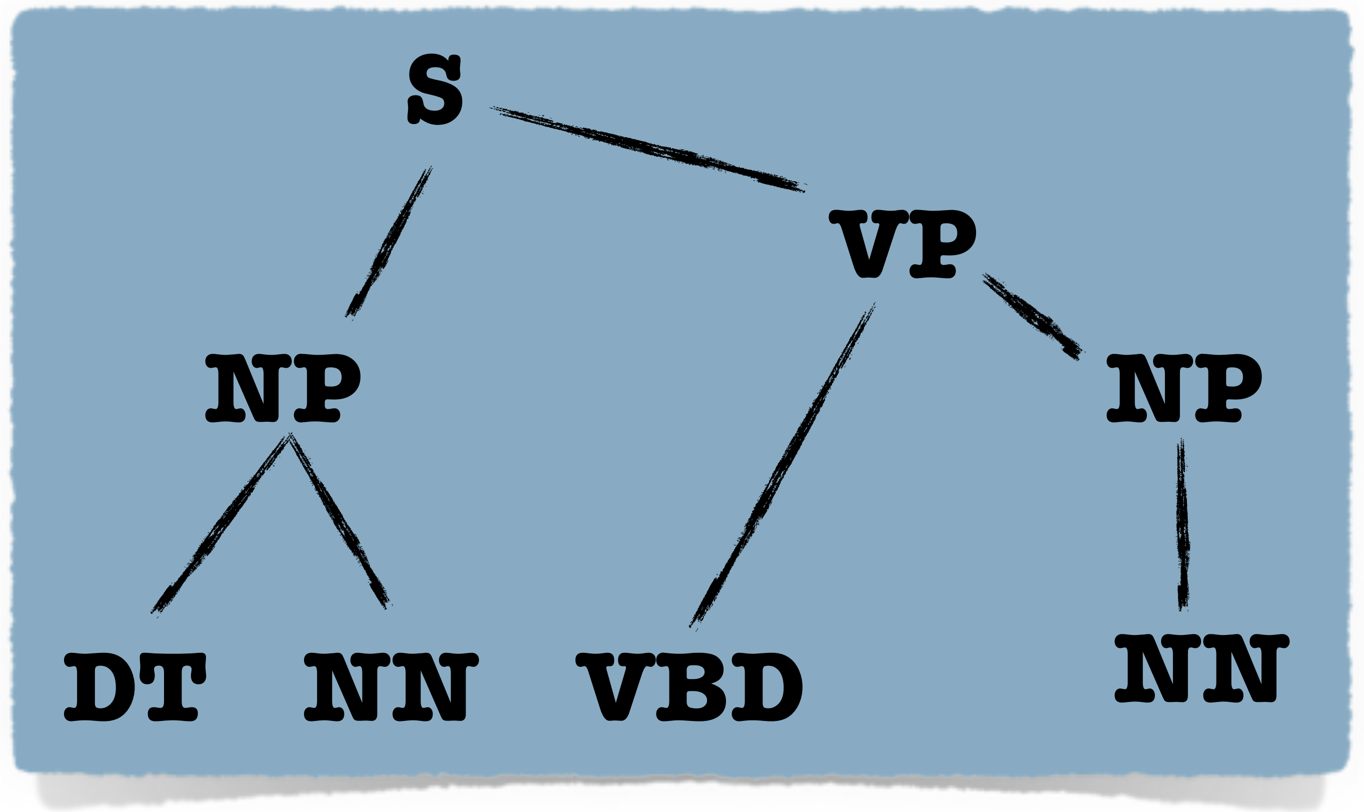


Linguistics

Machine Learning

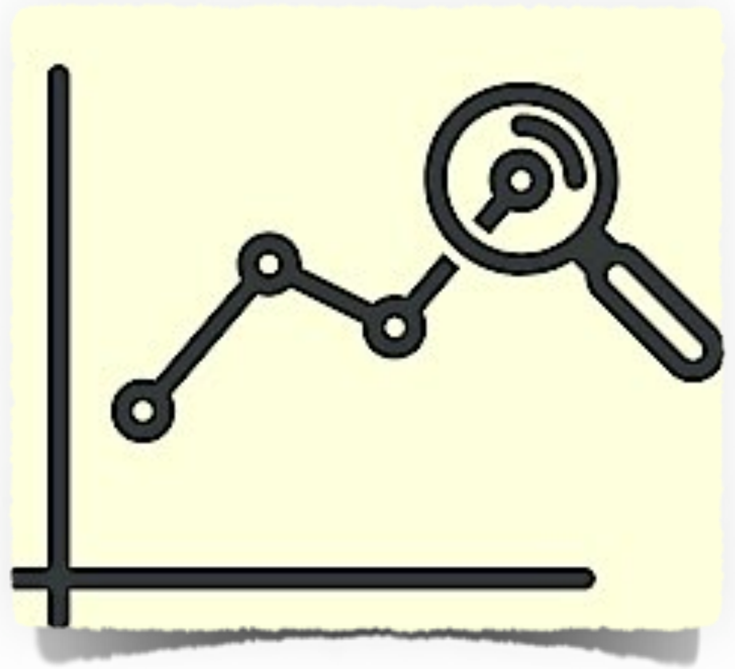


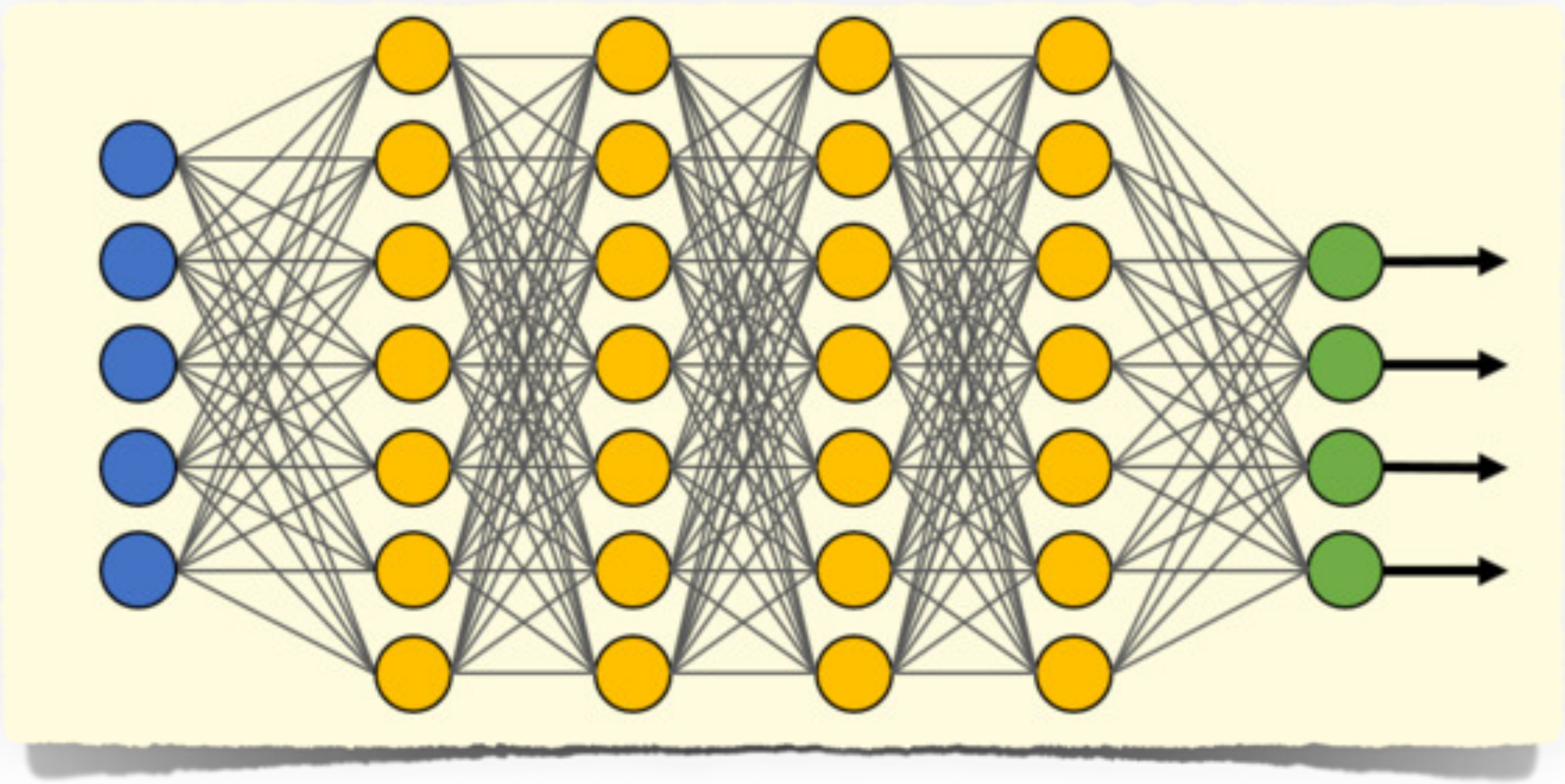
NLP



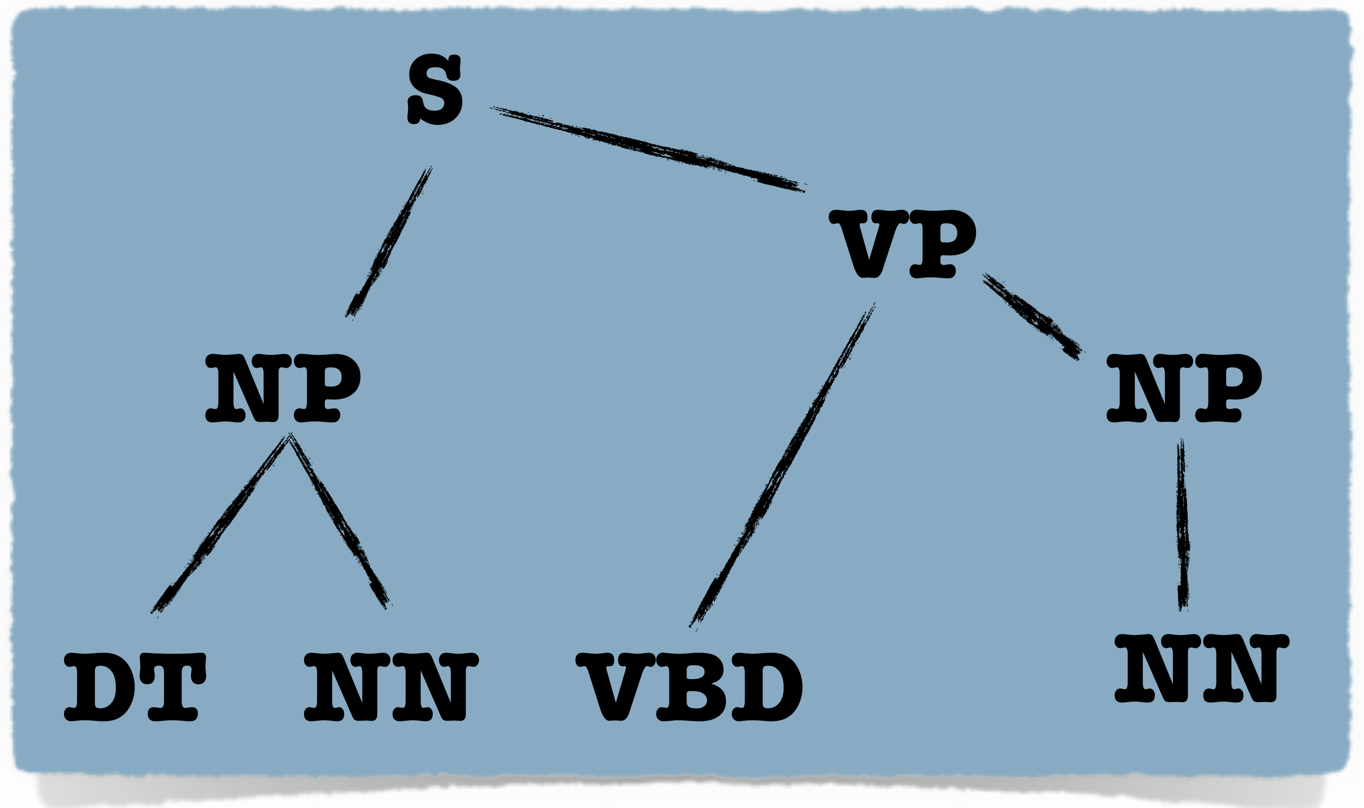
Linguistics

Machine Learning





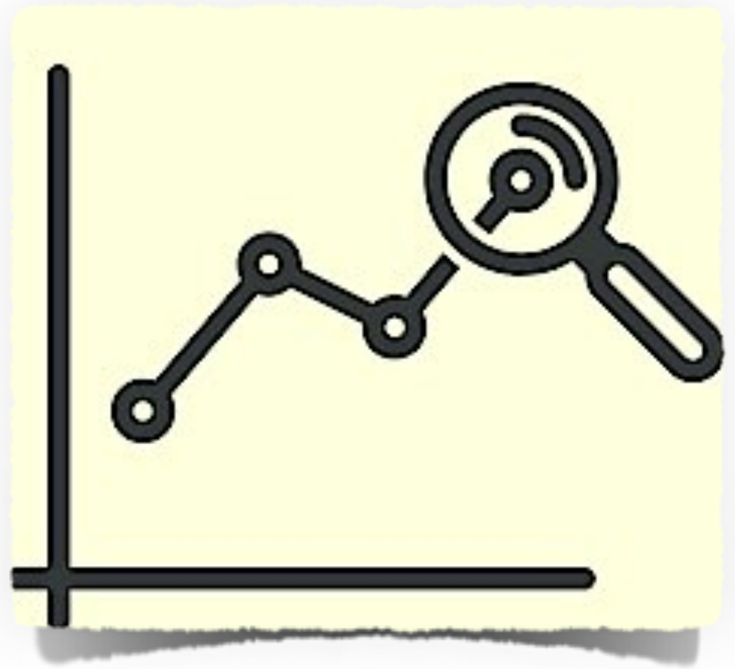
NLP



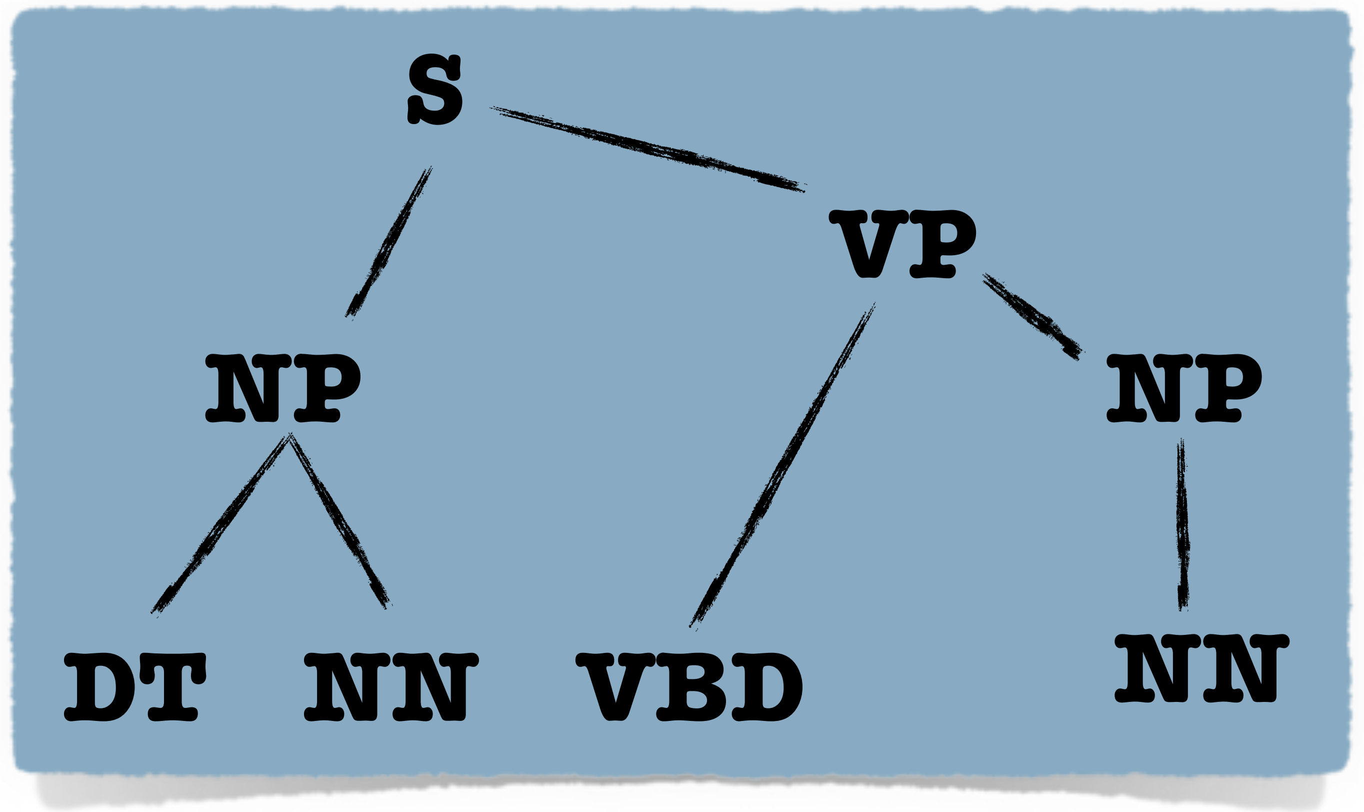
Data

Linguistics

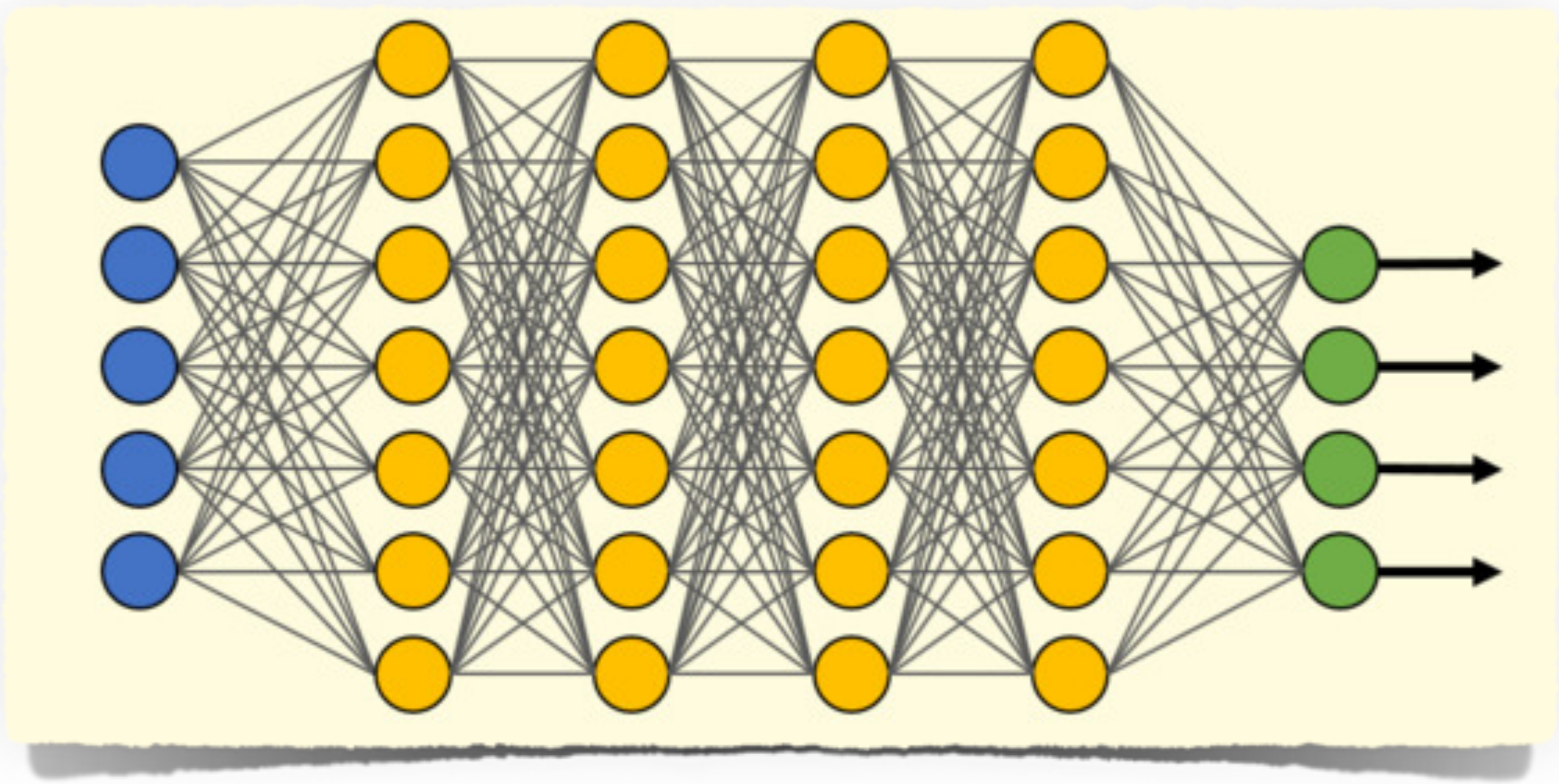
Machine Learning



NLP



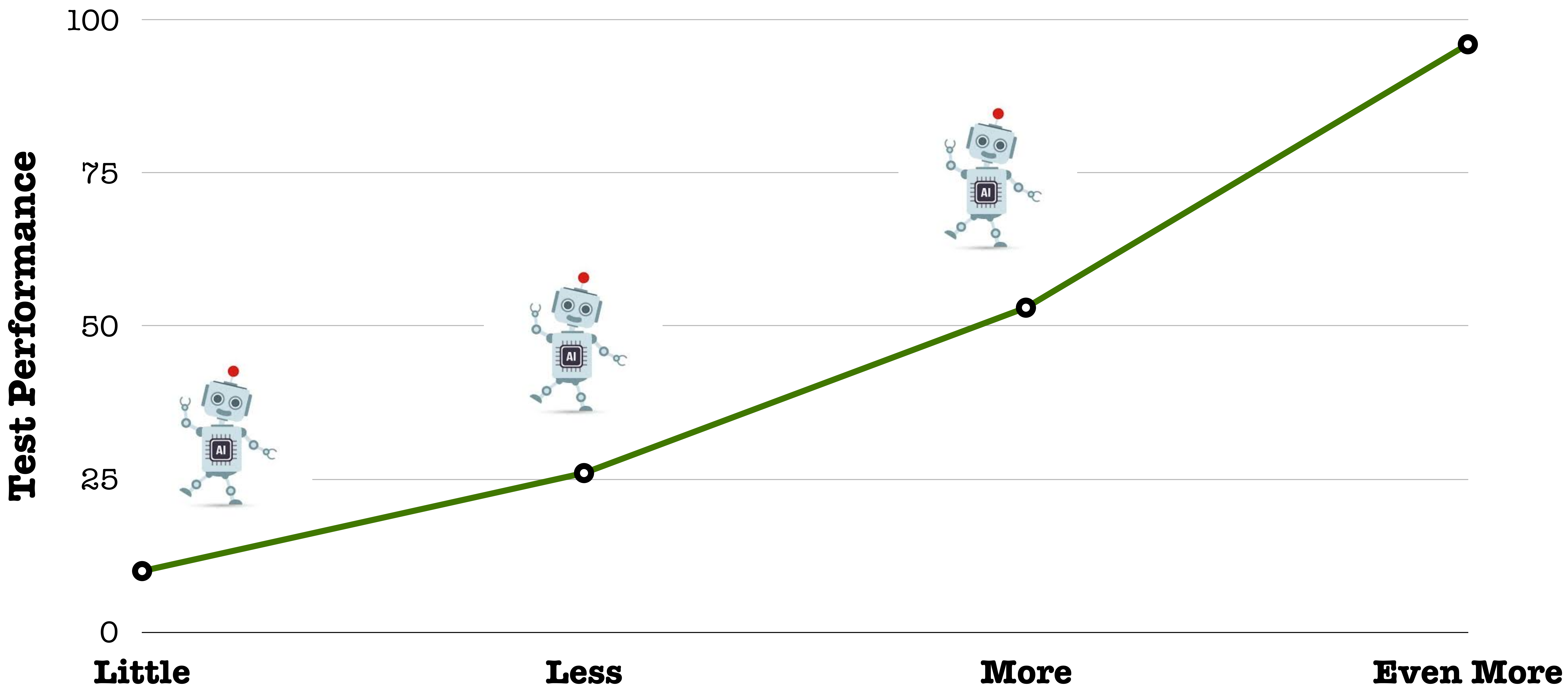
Data



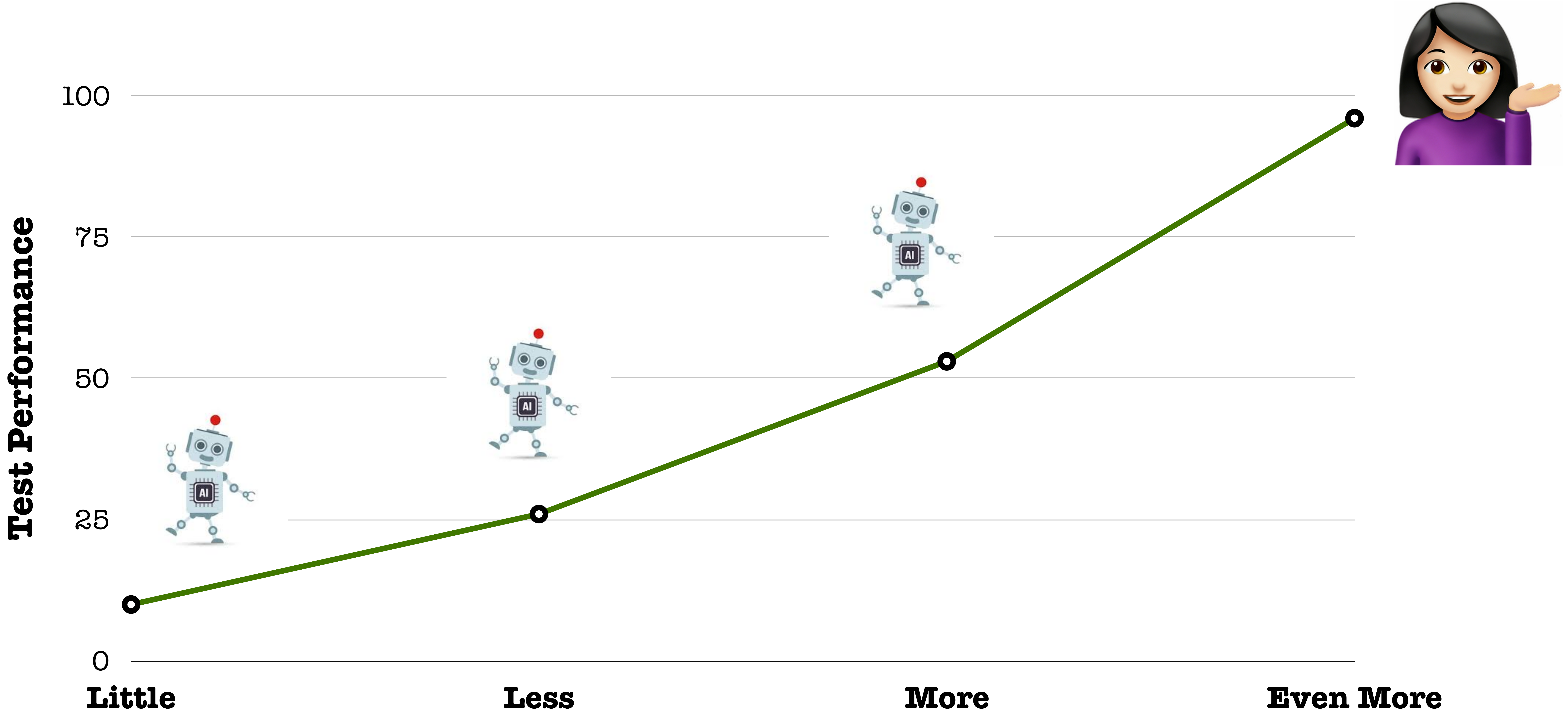
Linguistics

More data, better prediction?

More data, better prediction?



More data, better prediction?



More data, better prediction?

Datasets abound!

Datasets abound!

SWAG: A Large-Scale Adversarial Dataset

SQuAD
The Stanford Question Answering Dataset

QuAC
Question Answering in Context

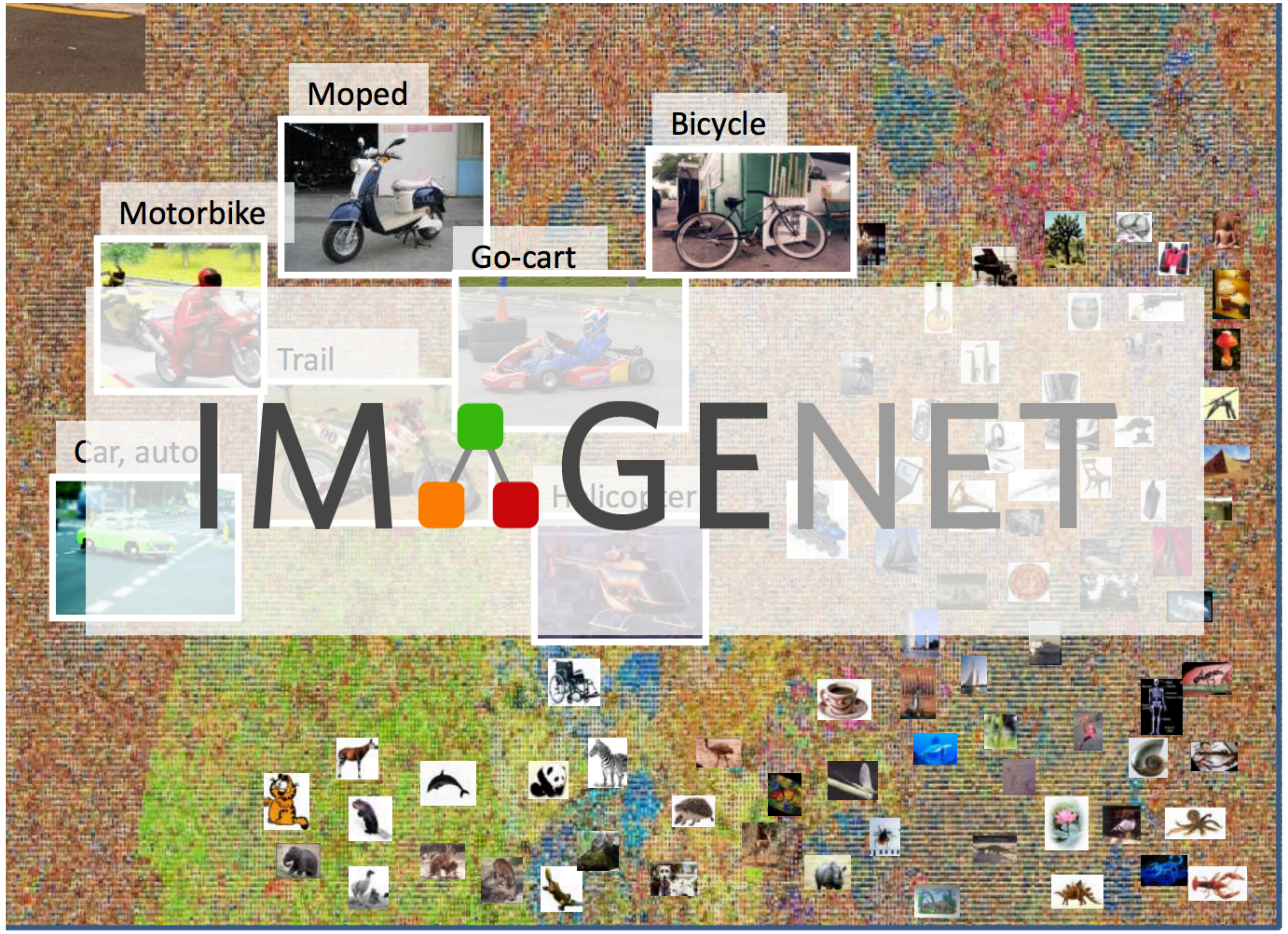
Sentiment Analysis

The Stanford Natural Language Inference (SNLI) Corpus

MultiNLI

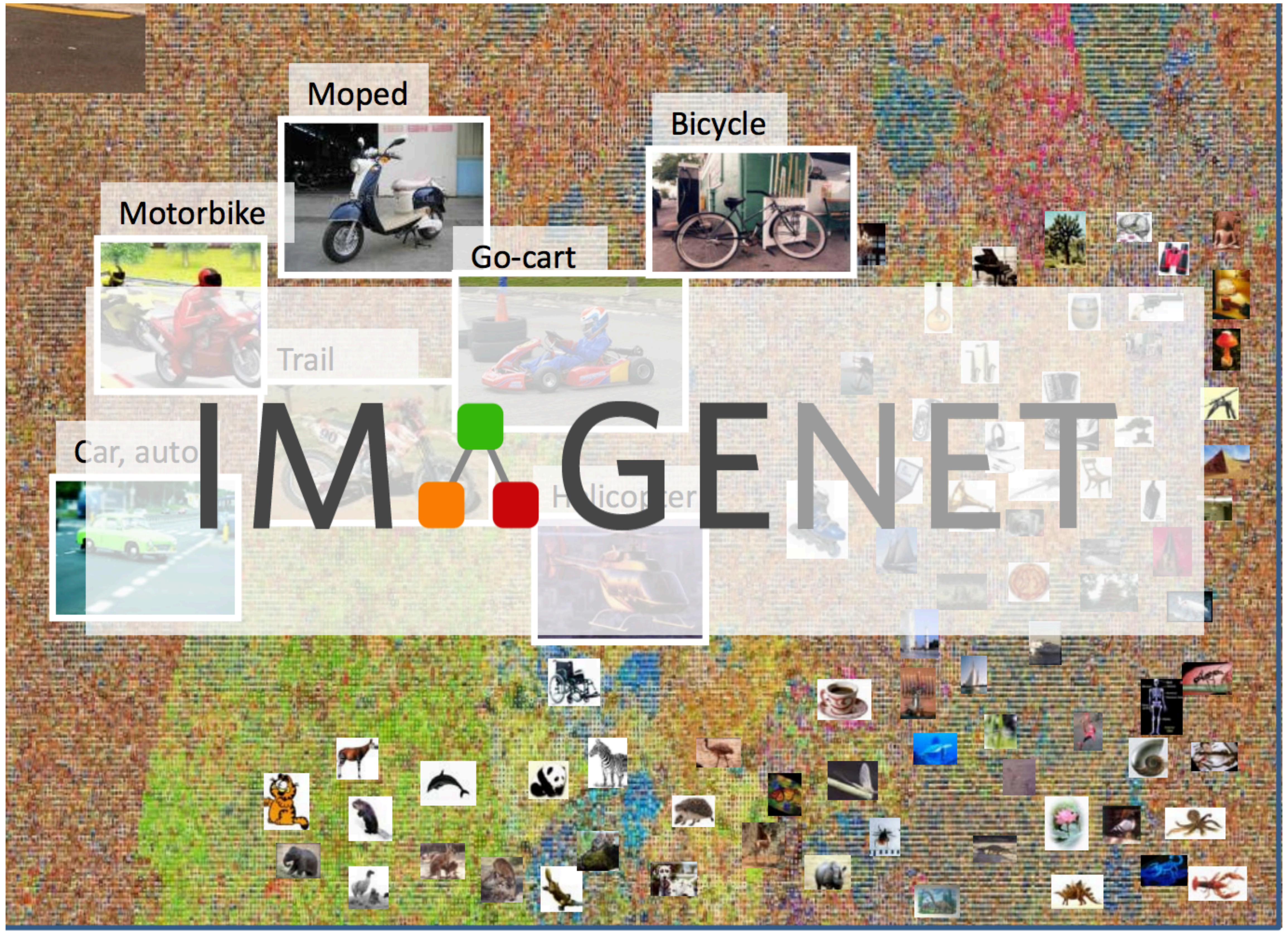
Story Cloze Test and ROCStories Corpora

Classifying images

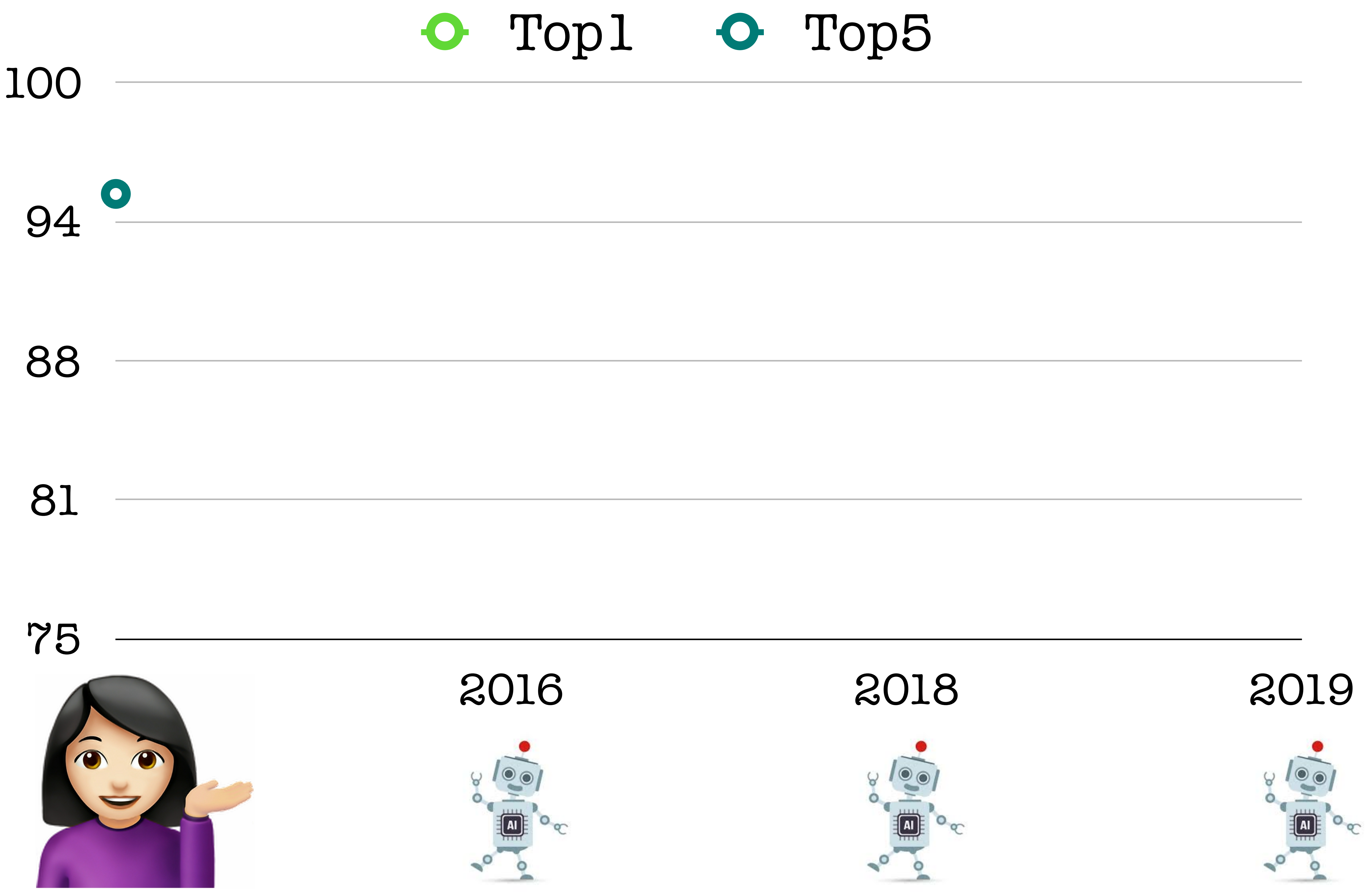


IMAGENET [Deng et al., CVPR 2009]

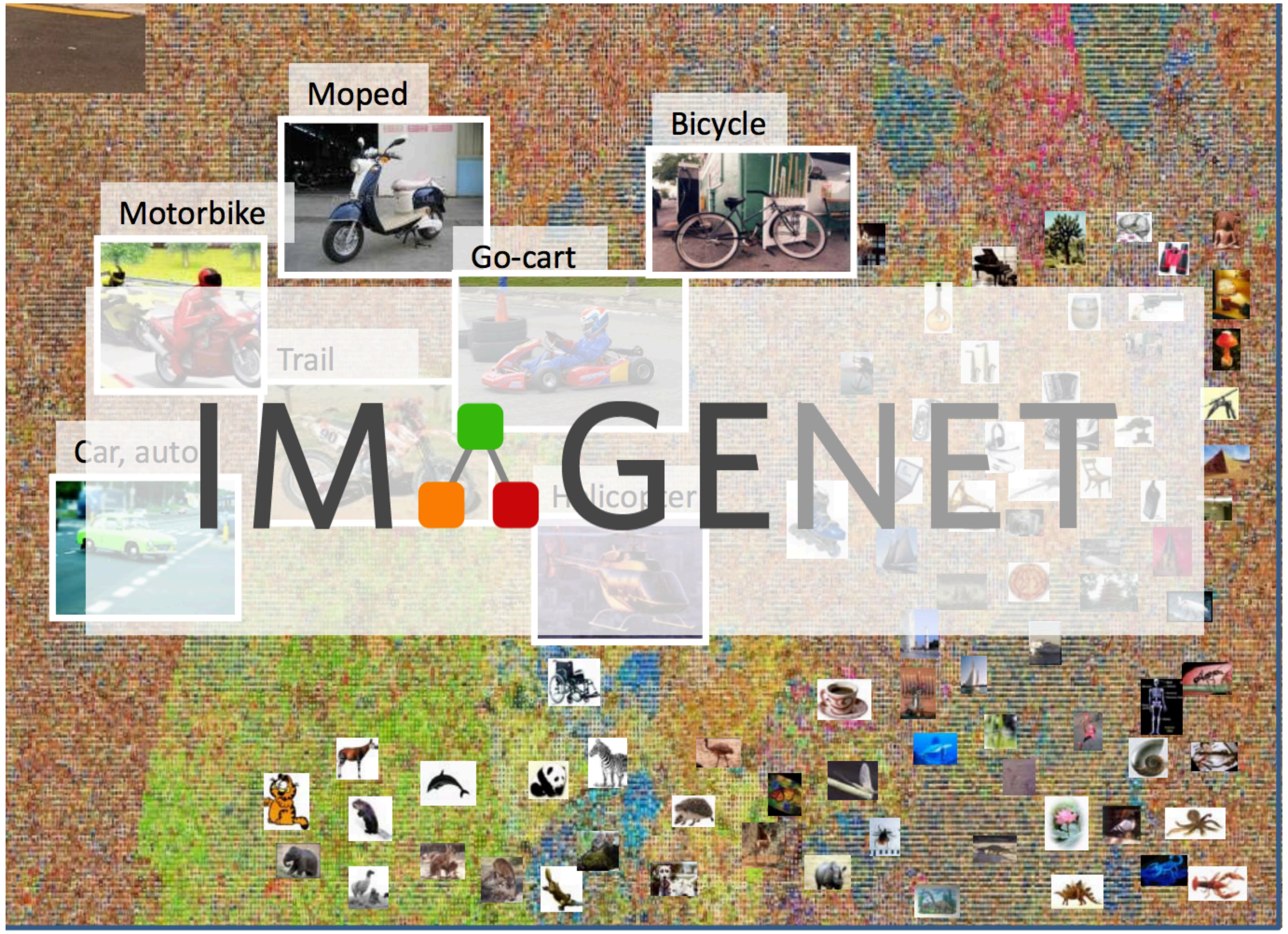
Classifying images



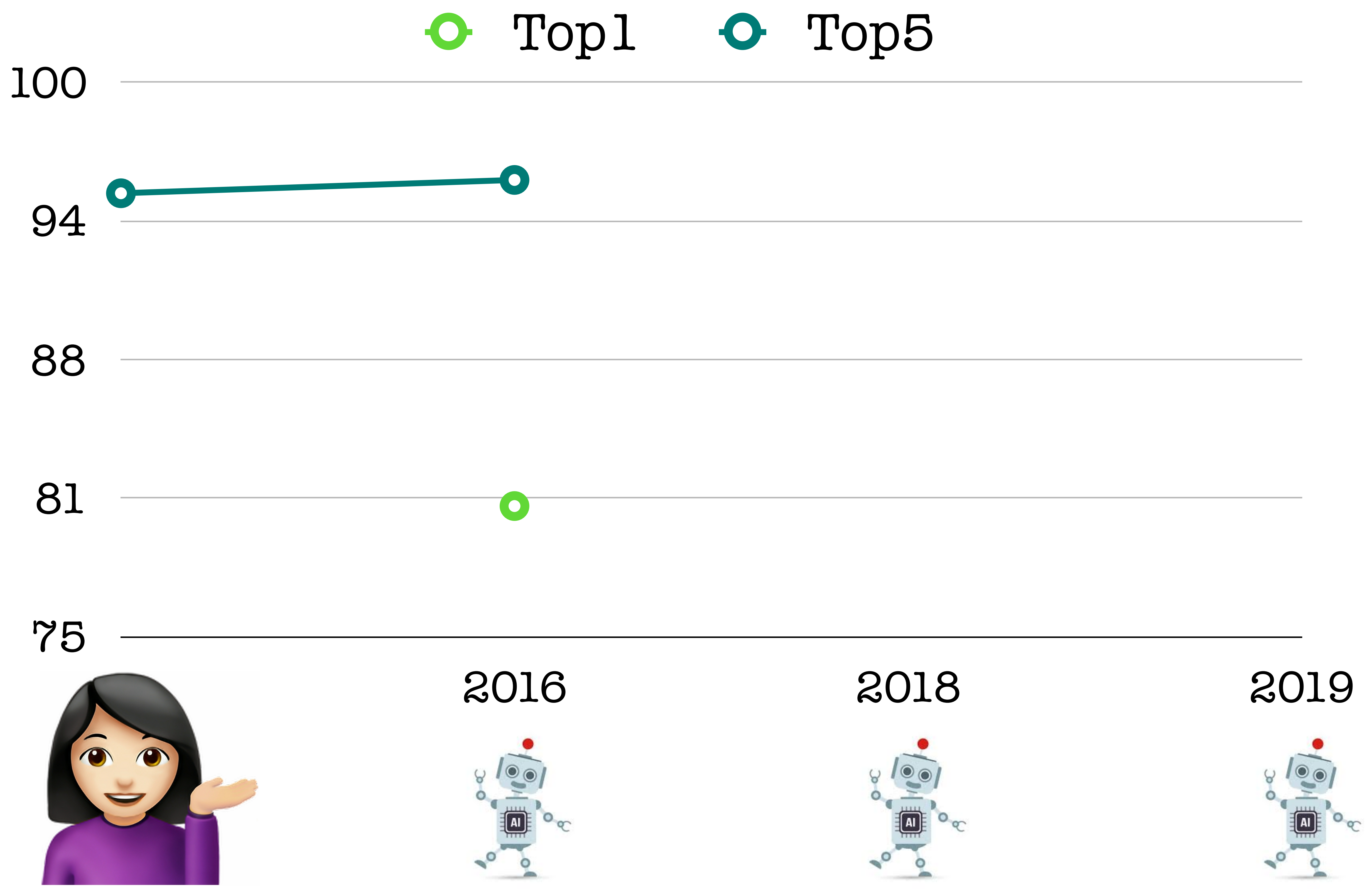
IMAGENET [Deng et al., CVPR 2009]



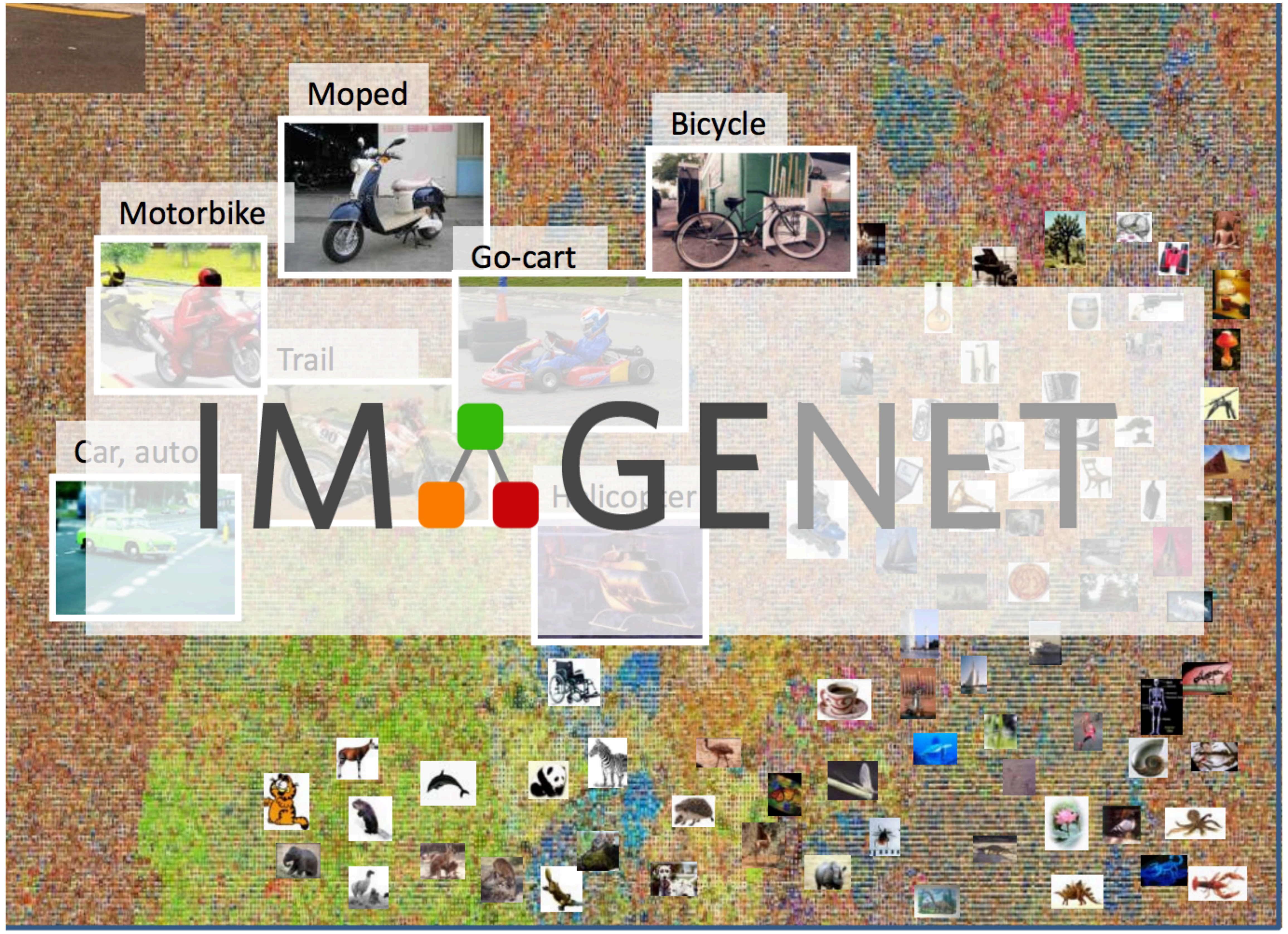
Classifying images



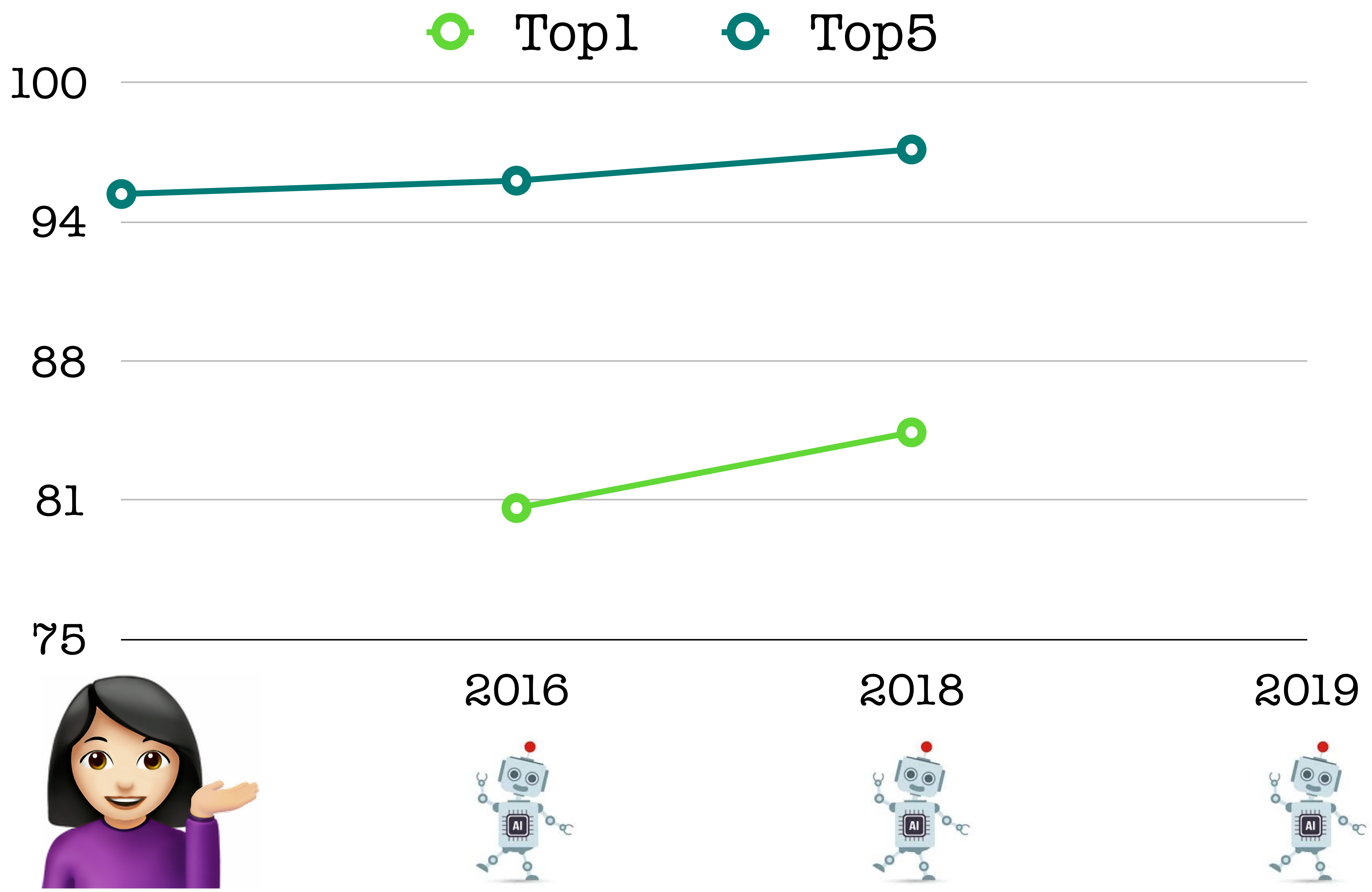
IMAGENET [Deng et al., CVPR 2009]



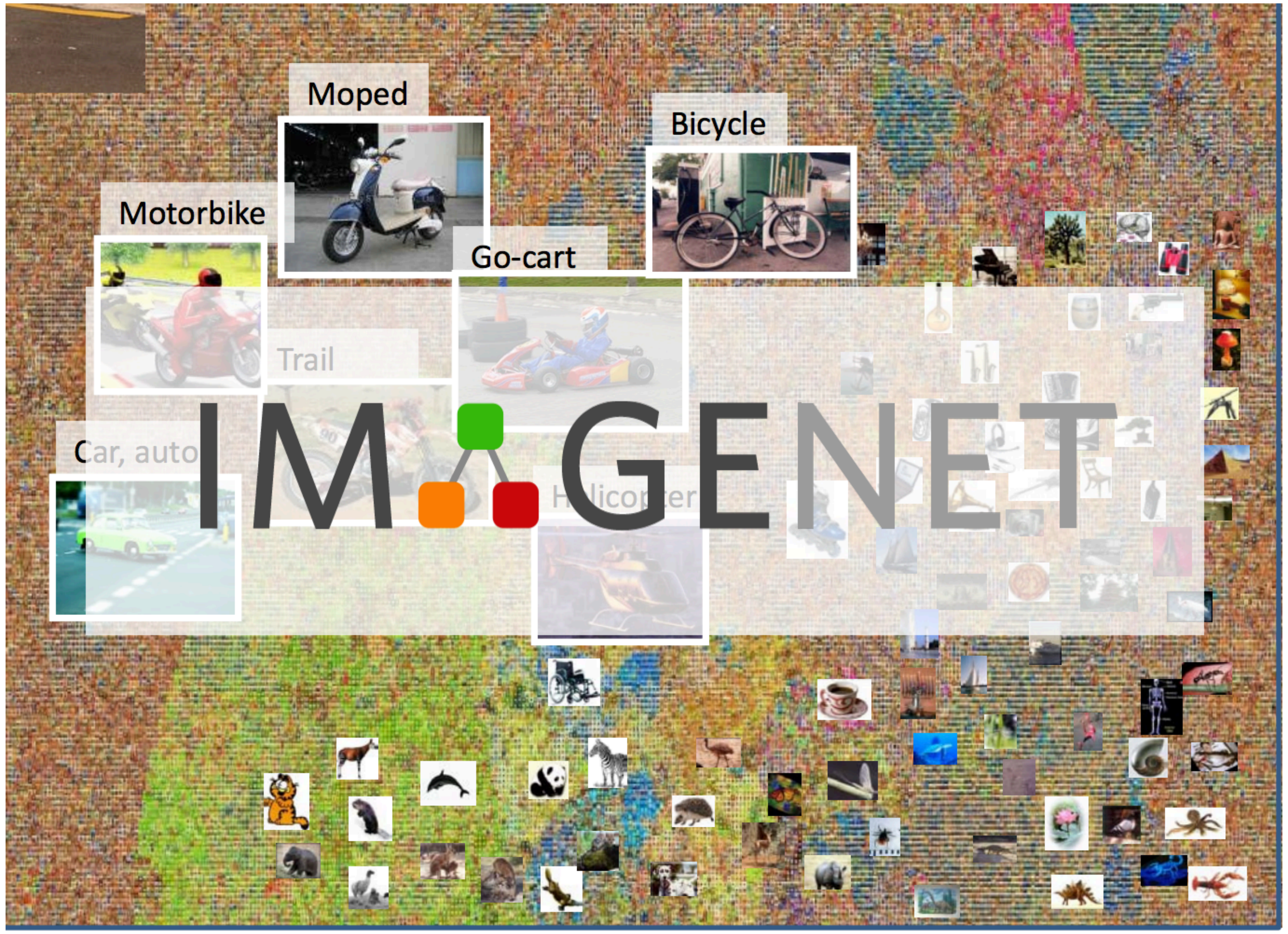
Classifying images



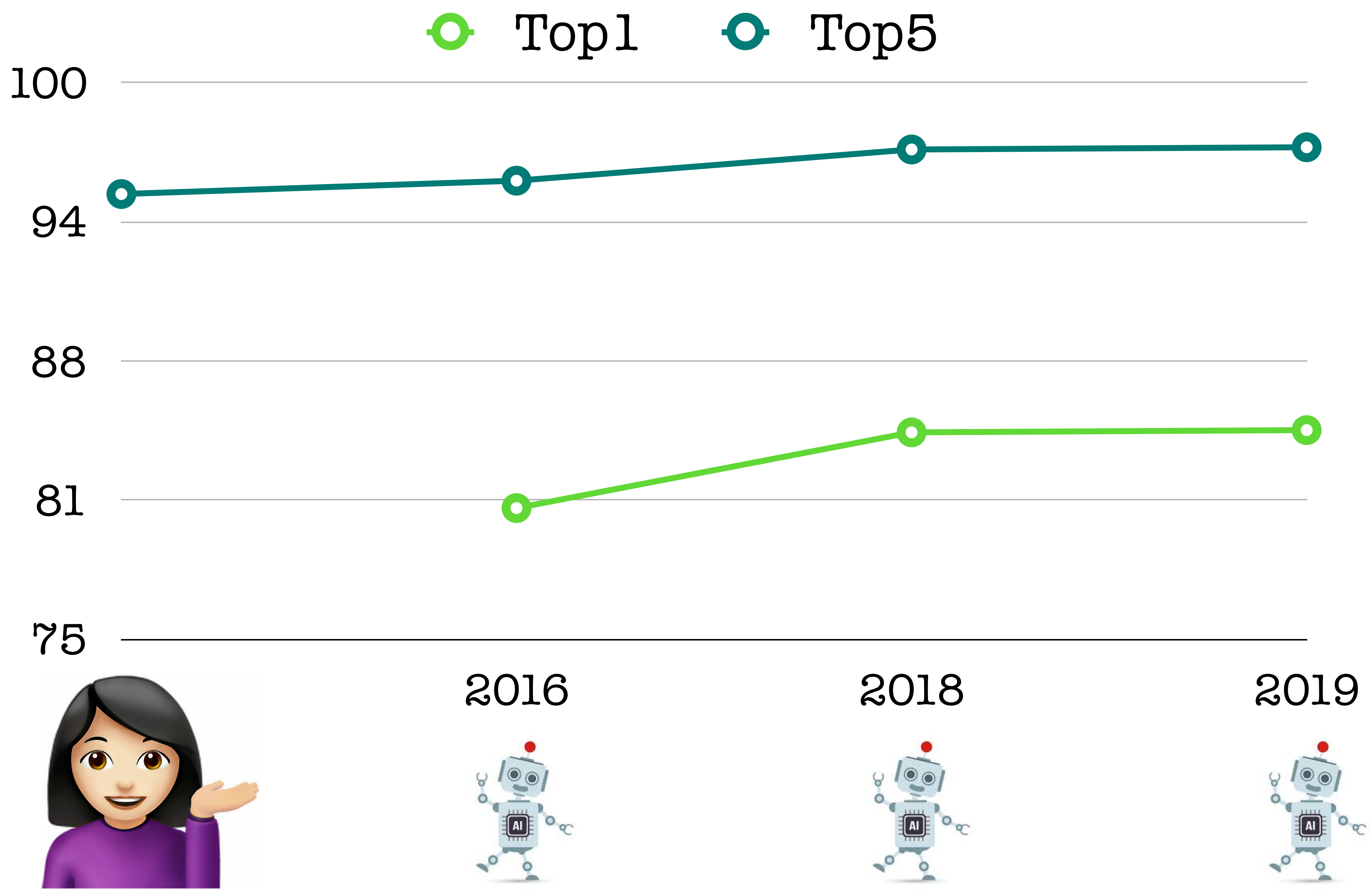
IMAGENET [Deng et al., CVPR 2009]



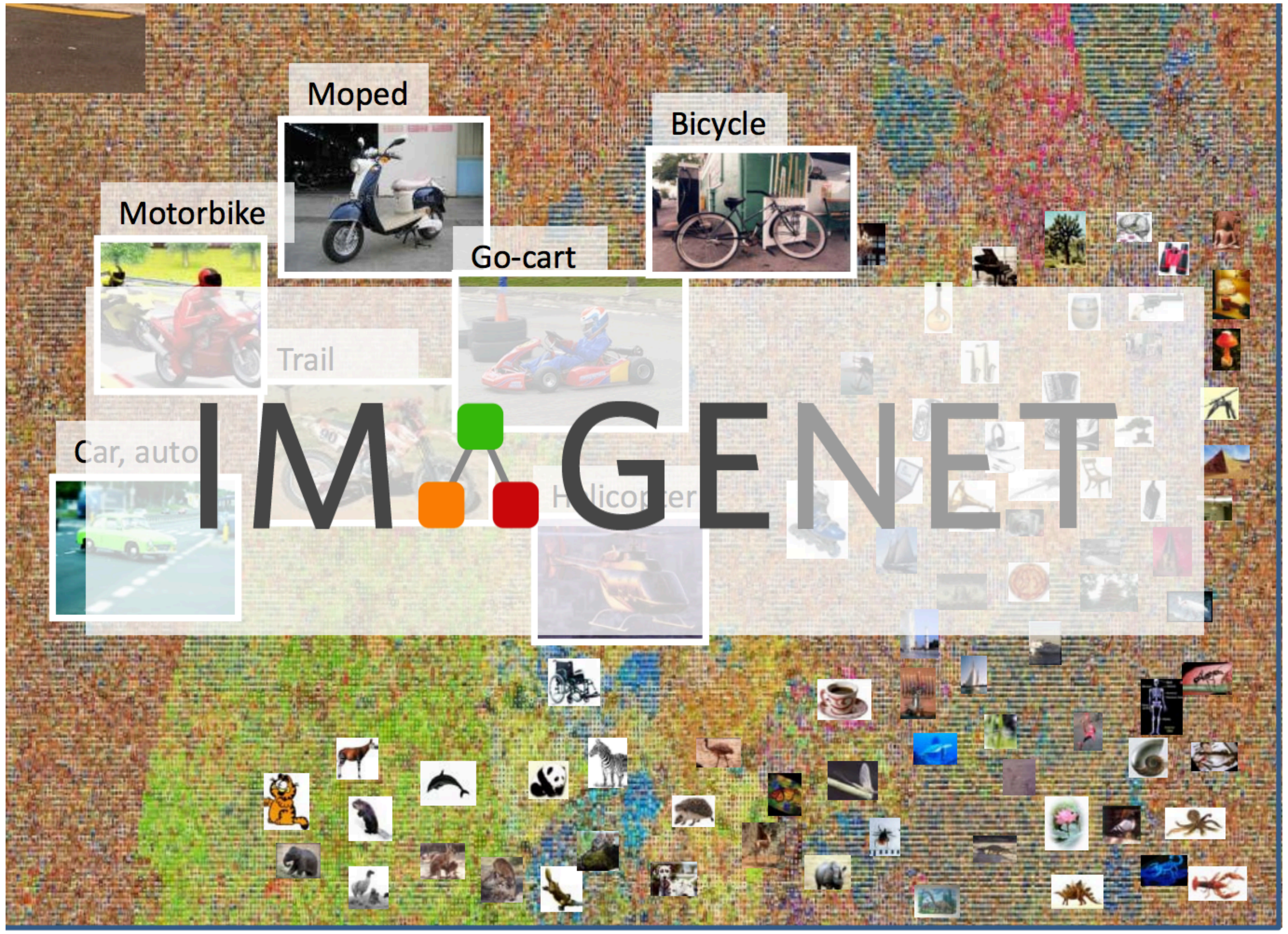
Classifying images



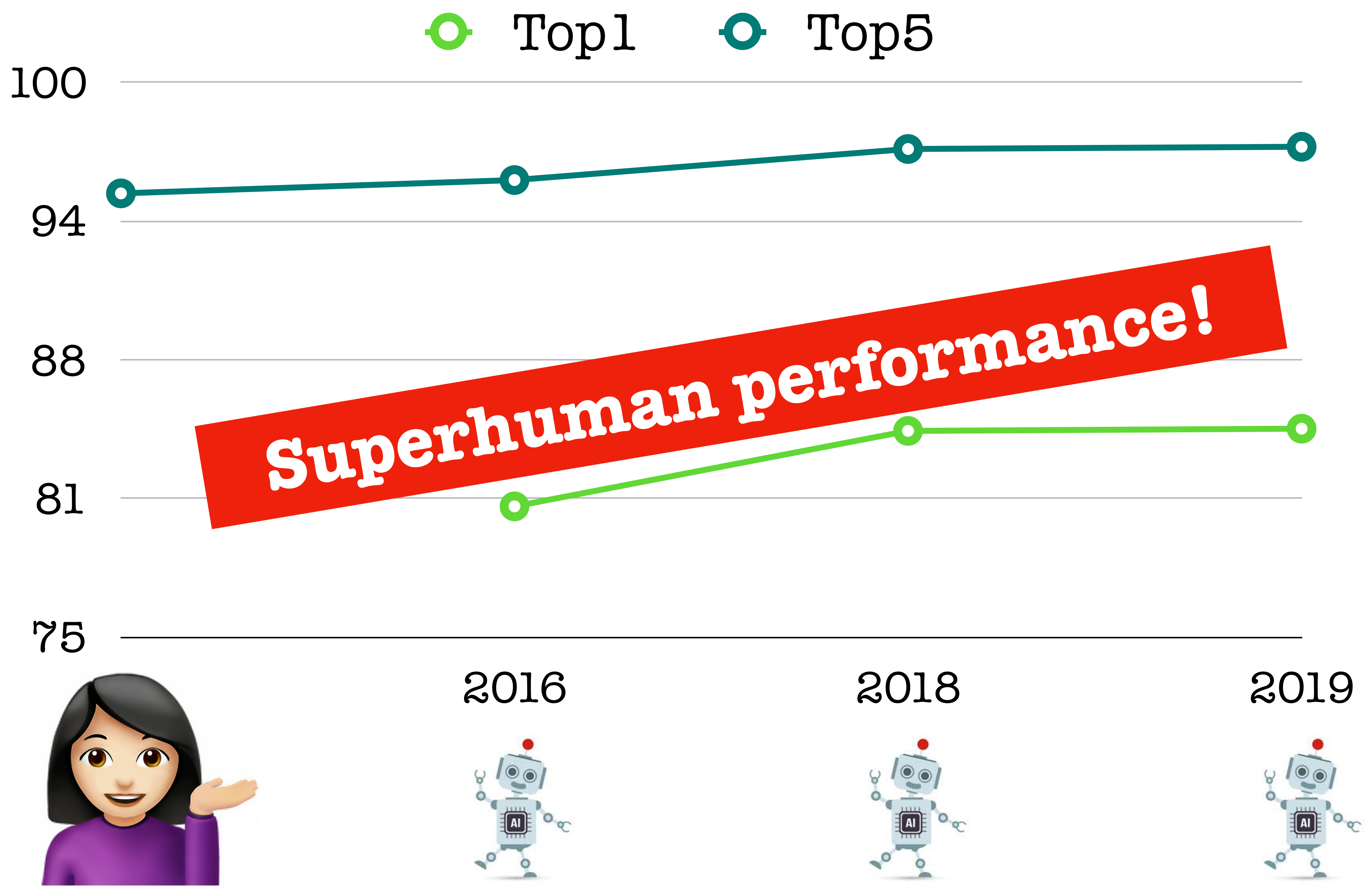
IMAGENET [Deng et al., CVPR 2009]



Classifying images



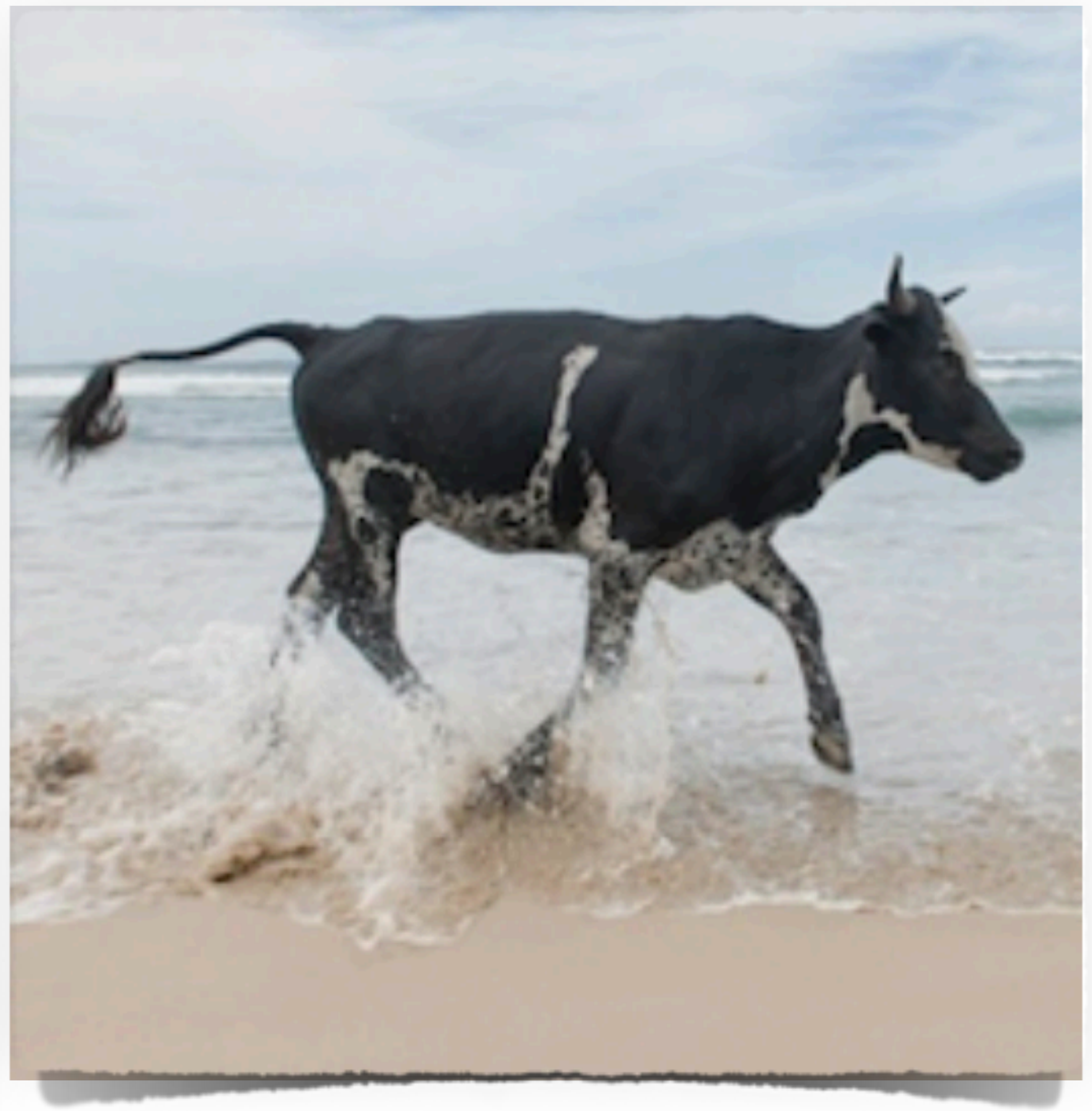
IMAGENET [Deng et al., CVPR 2009]



A simple quiz

A simple quiz



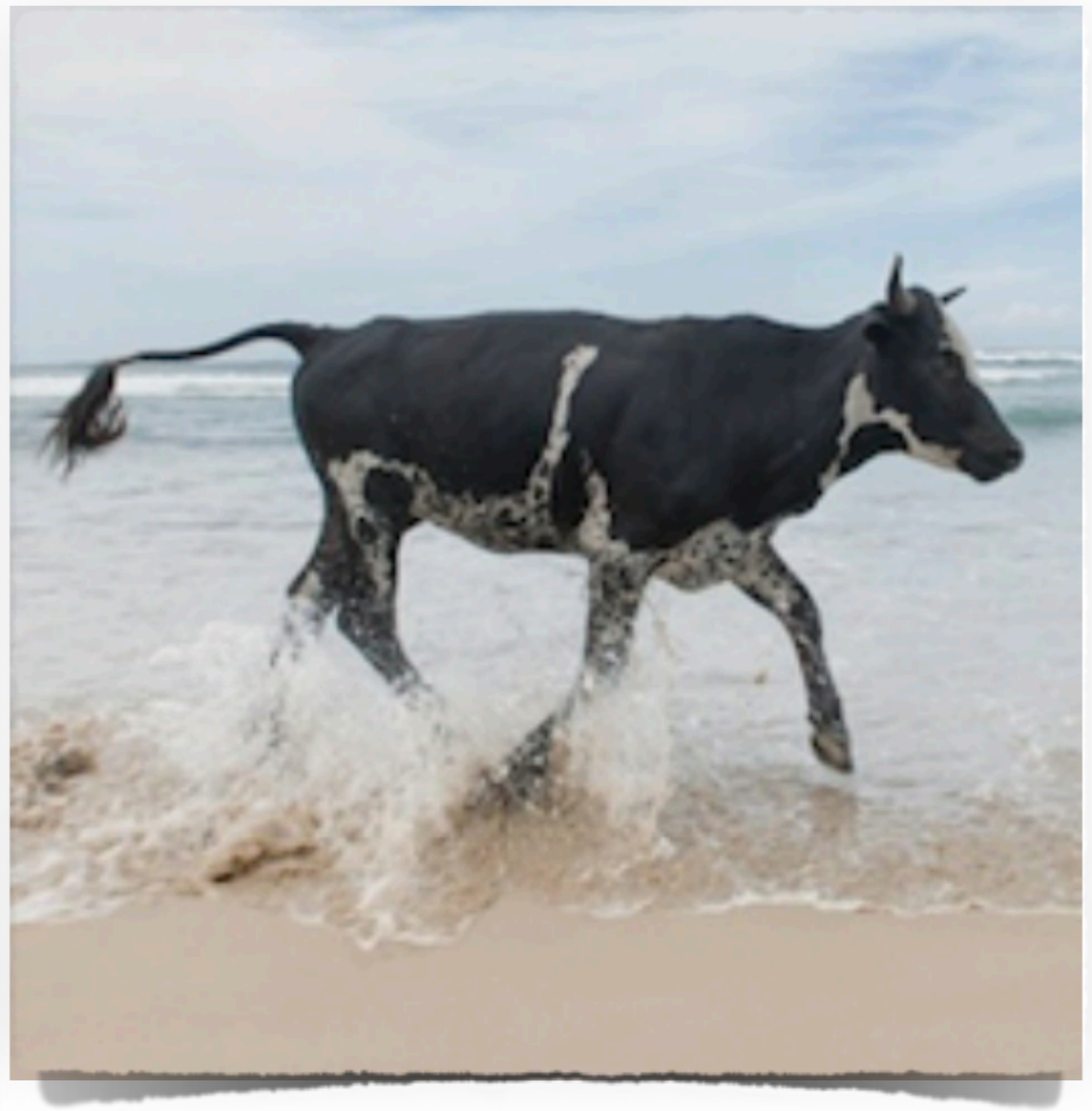



A simple quiz

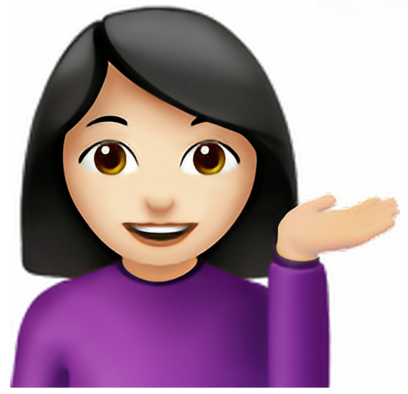
Cow

Cow

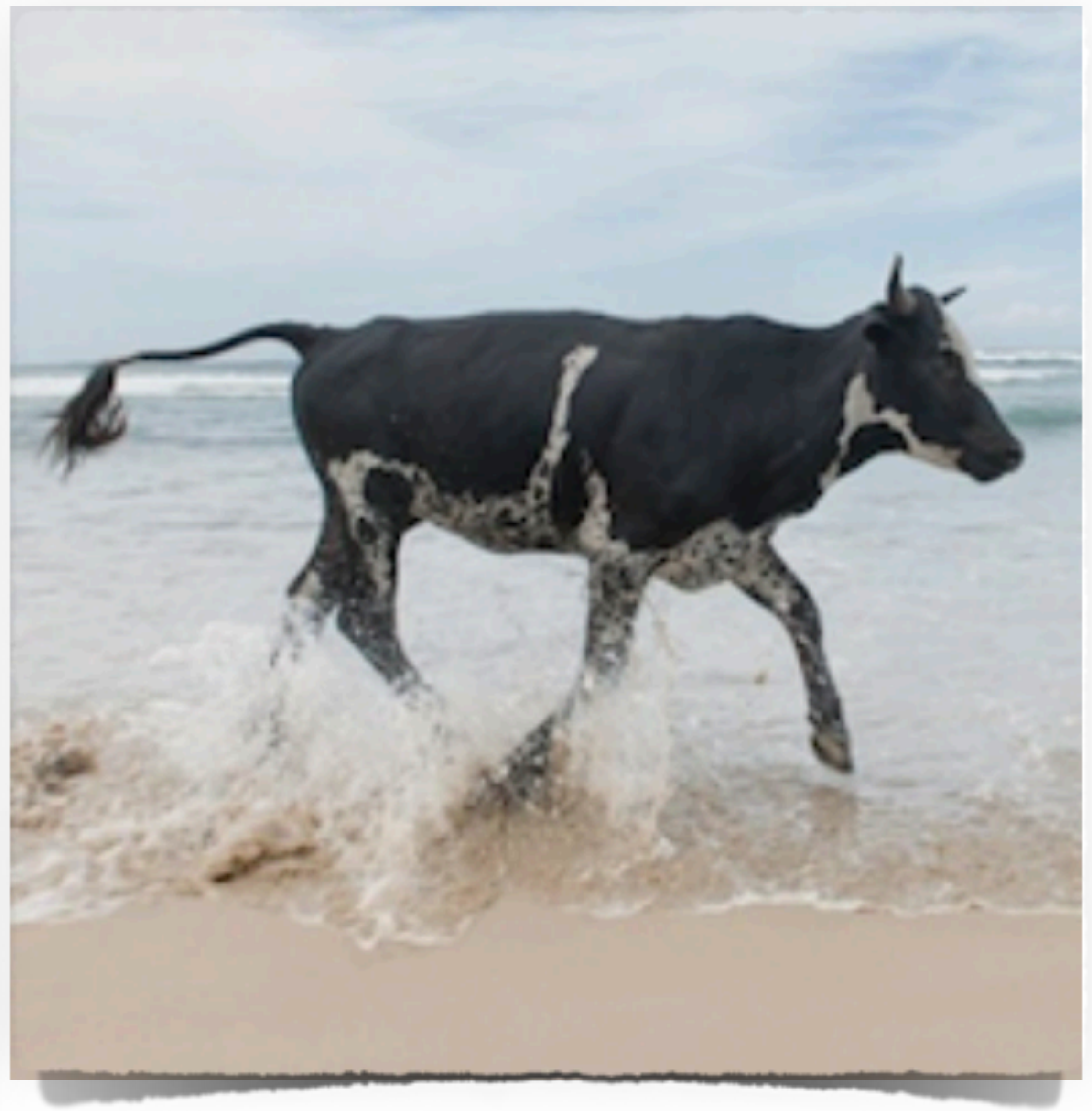
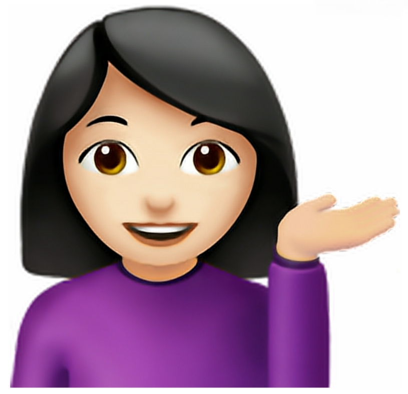
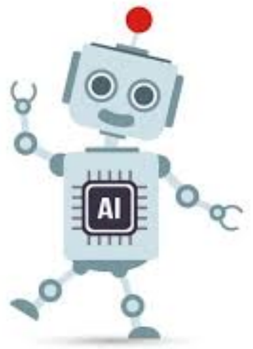
Cow



Cow

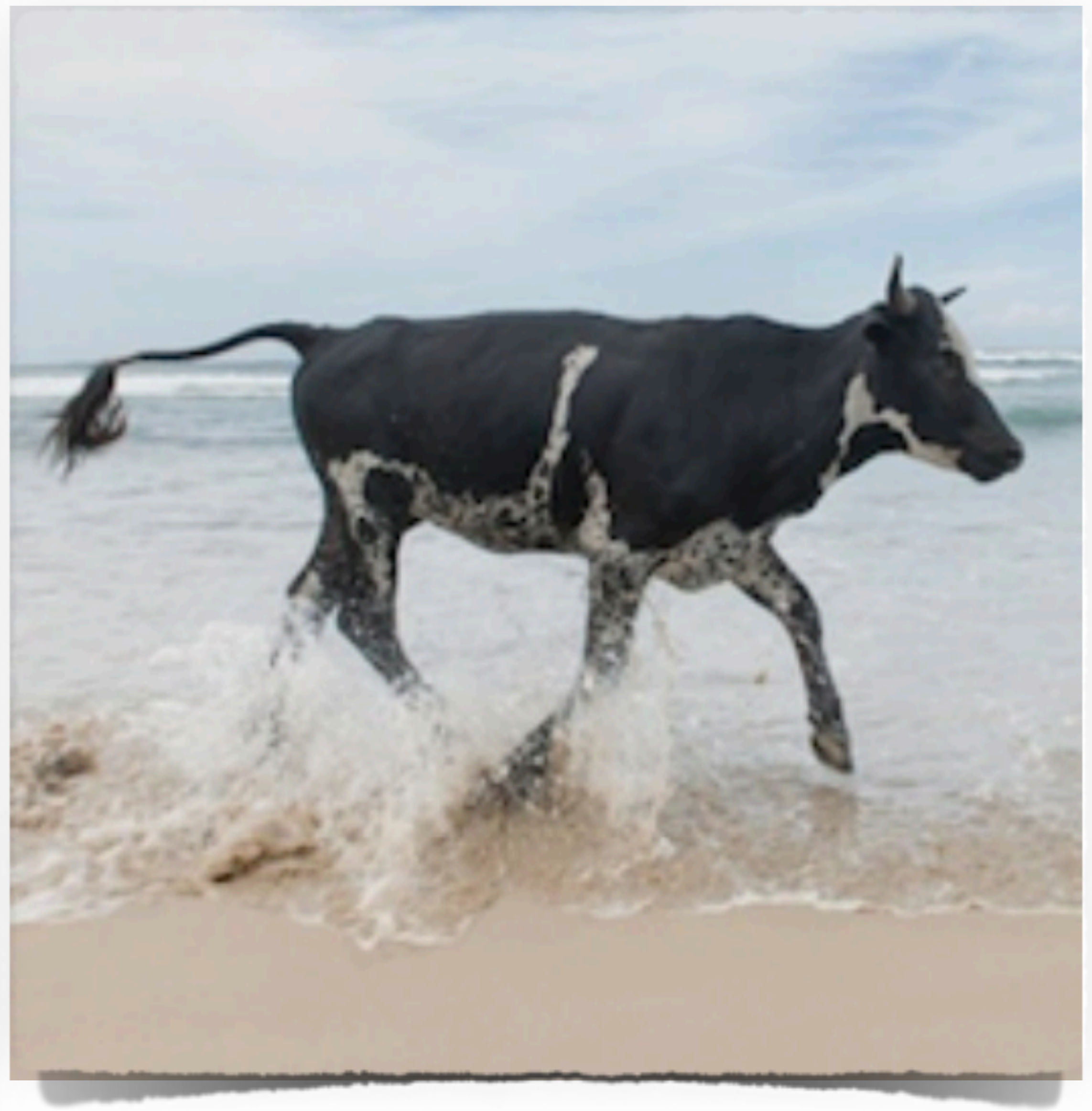
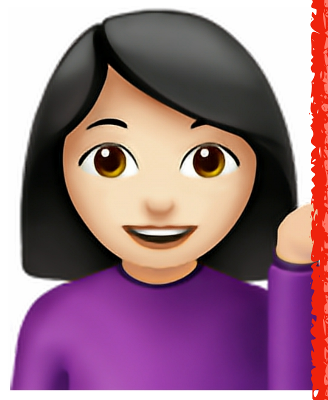
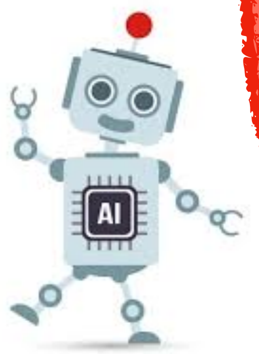


A simple quiz

				
	Cow	Cow	Cow	Cow
	Cow	Cow	No person	No person

Beery et al., 2018

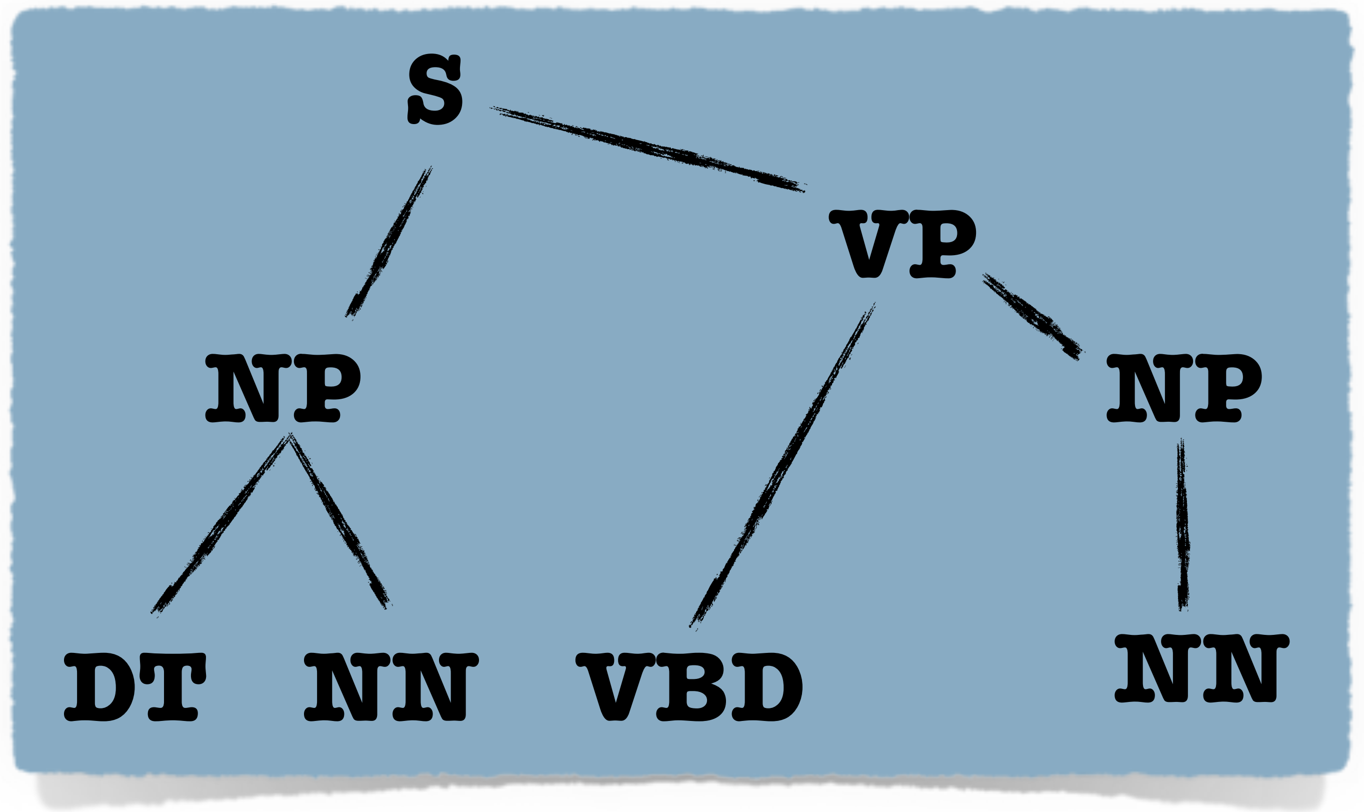
A simple quiz

				
	Cow	Cow	Cow	Cow
	Cow	Cow	No person	No person

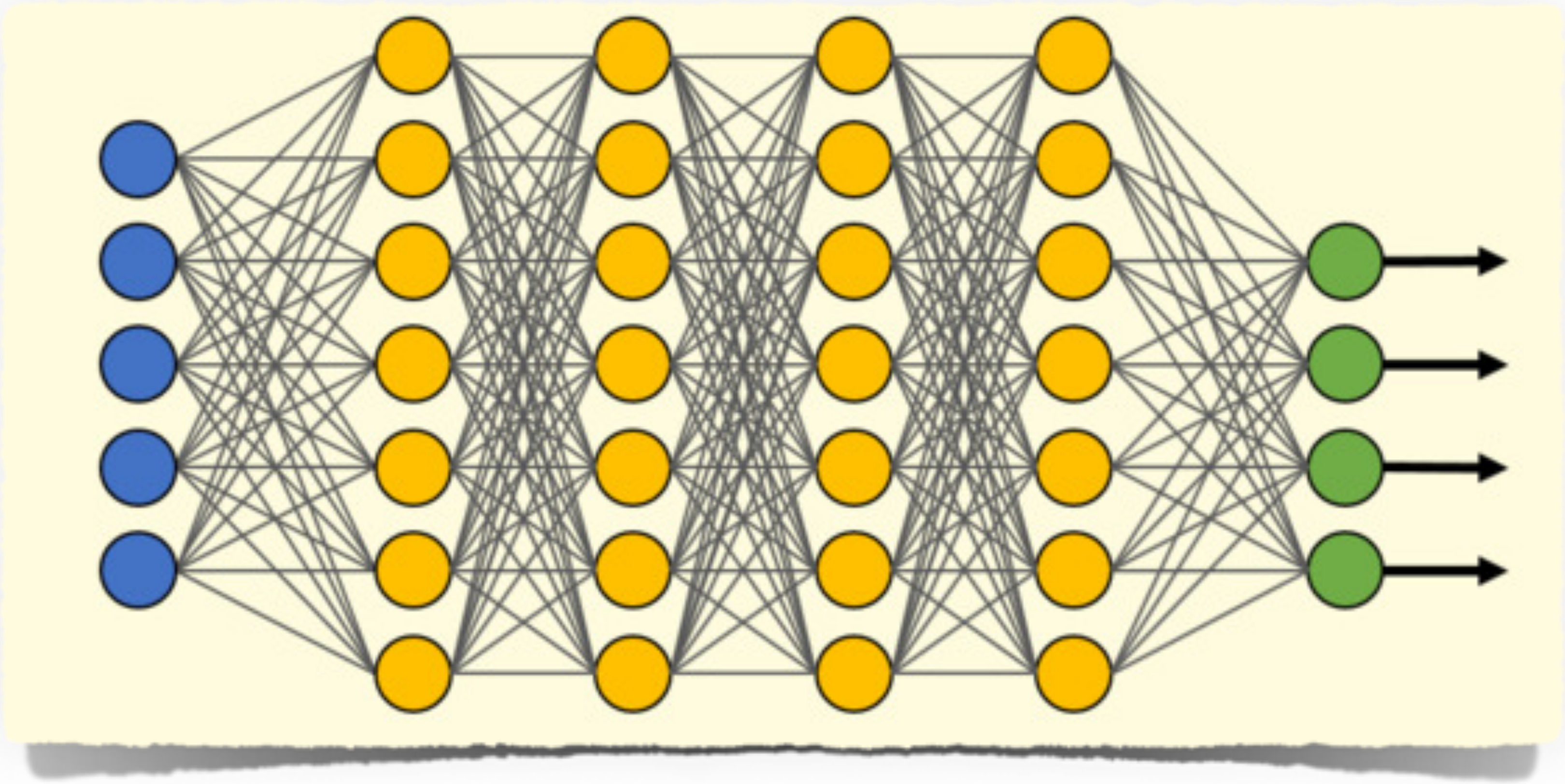
Beery et al., 2018

Machine Learning

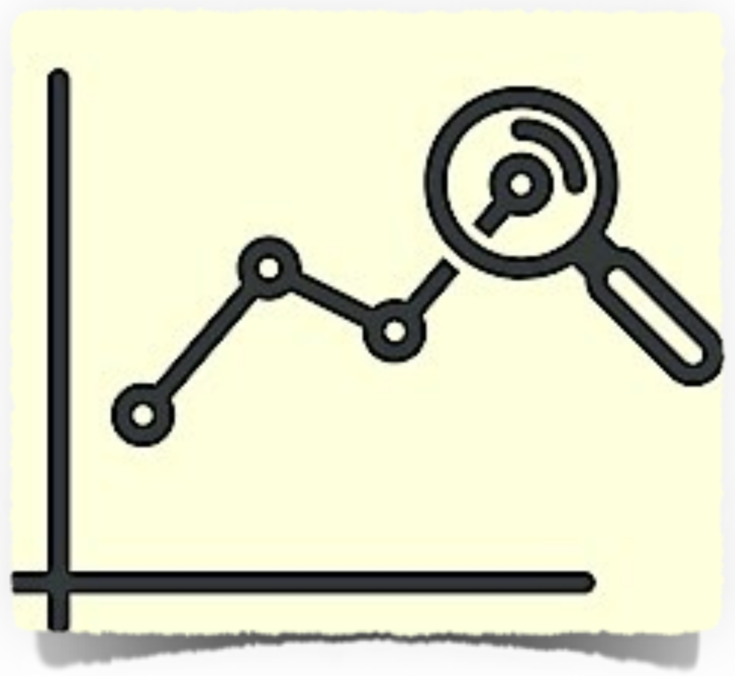
NLP



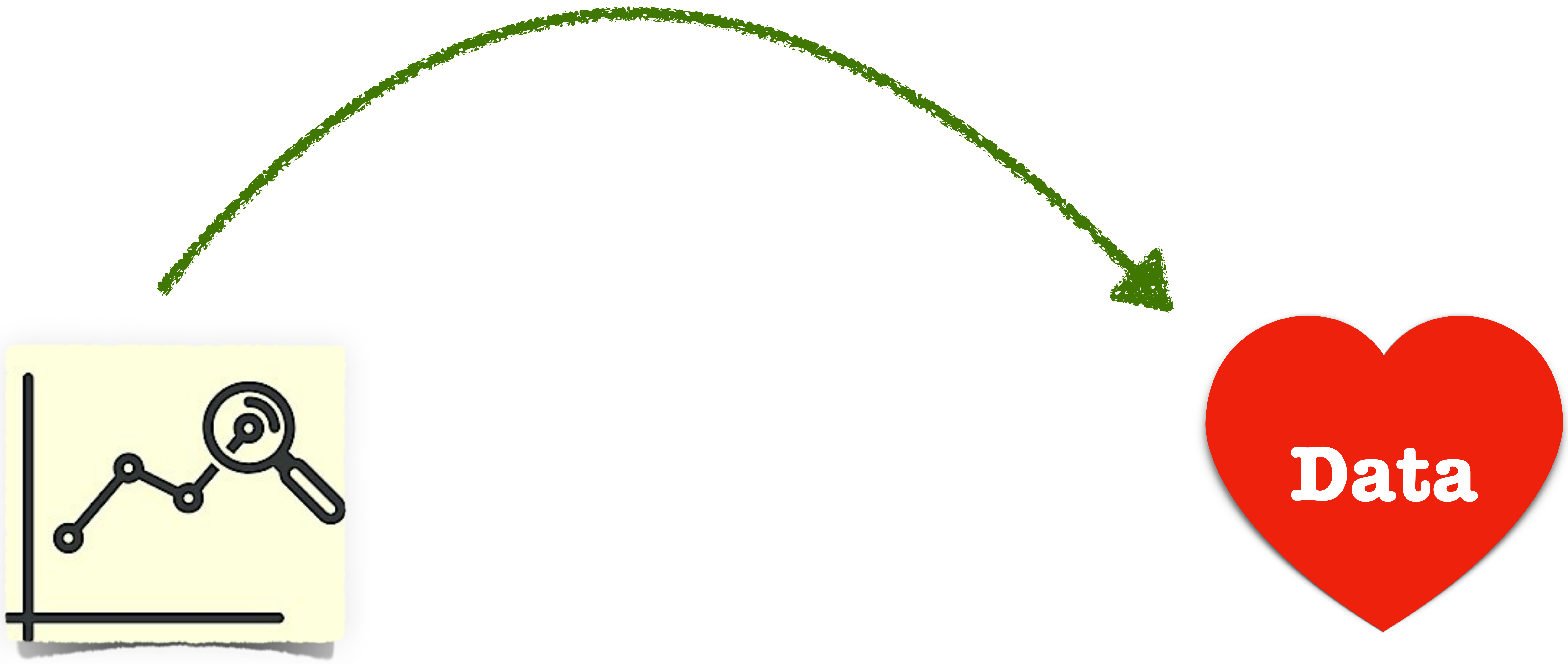
Data



Linguistics




Data


Data

Question #1

What do predictions tell us about the data?

Natural Language Inference (NLI)

Natural Language Inference (NLI)

Given a premise, is a hypothesis true, false or neither?

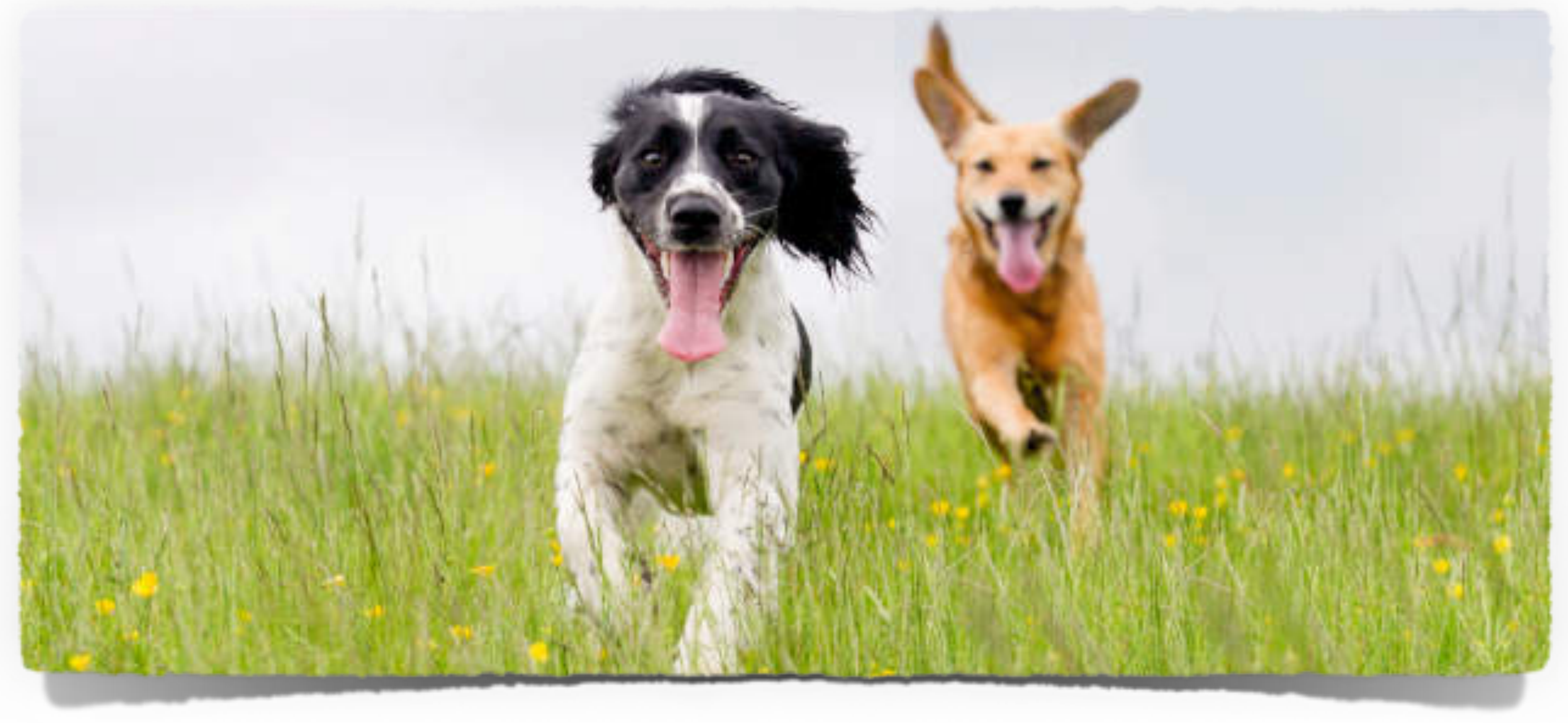
Natural Language Inference (NLI)

Premise Two dogs are running through a field.

Hypothesis The pets are sitting on a couch.

Given a premise, is a hypothesis true, false or neither?

Natural Language Inference (NLI)



Premise Two dogs are running through a field.

Hypothesis The pets are sitting on a couch.

Given a premise, is a hypothesis true, false or neither?

- True → **Entailment**
- False → **Contradiction**
- Cannot Say → **Neutral**

Natural Language Inference (NLI)

Premise Two dogs are running through a field.

Hypothesis The pets are sitting on a couch.

Given a premise, is a hypothesis true, false or neither?

- True → **Entailment**
- False → **Contradiction**
- Cannot Say → **Neutral**

Natural Language Inference (NLI)

Premise Two dogs are running through a field.

Hypothesis The pets are sitting on a couch.

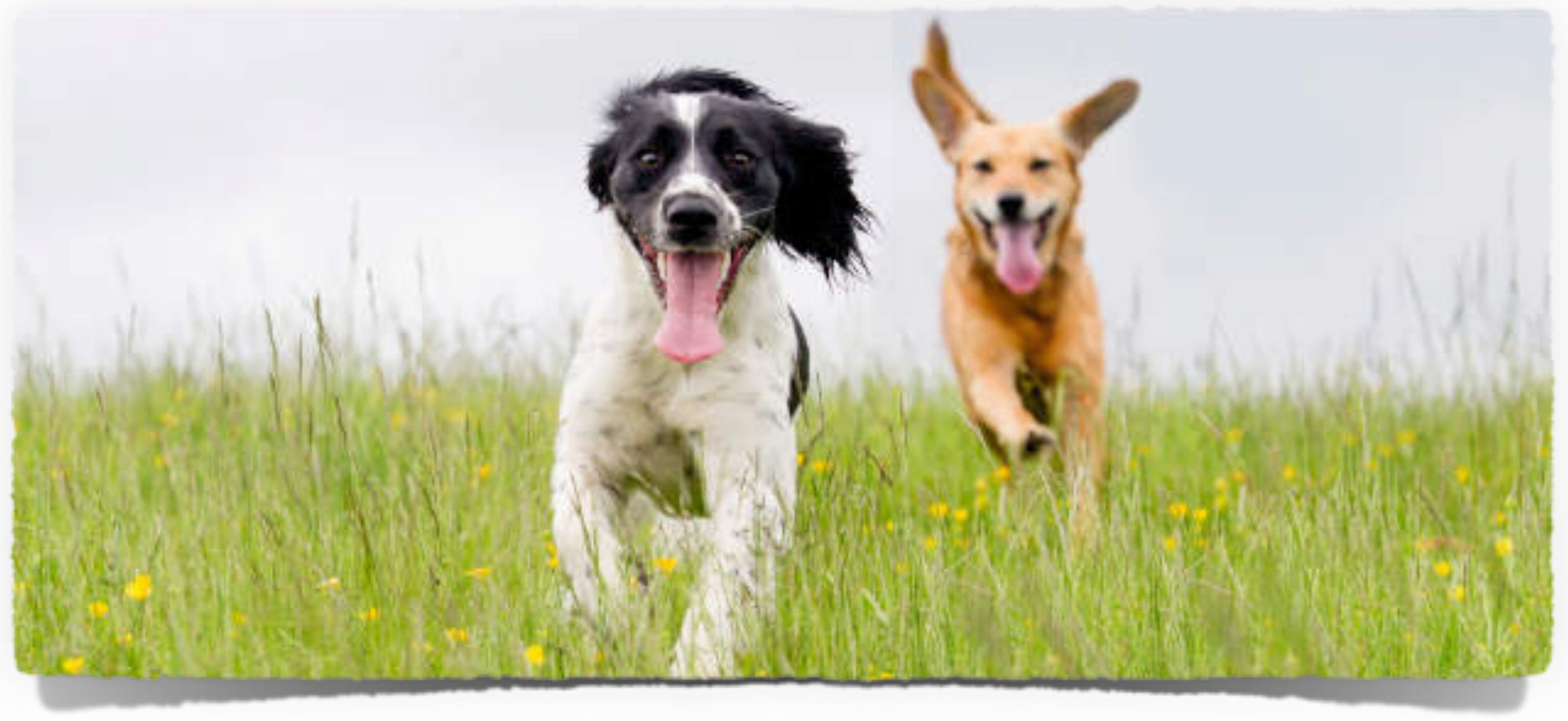
Given a premise, is a hypothesis true, false or neither?

- True → **Entailment**
- False → **Contradiction**
- Cannot Say → **Neutral**

[Katz, 1972; van Benthem, 2008; Dagan et al., 2006]

NLI Dataset Creation Process

NLI Dataset Creation Process



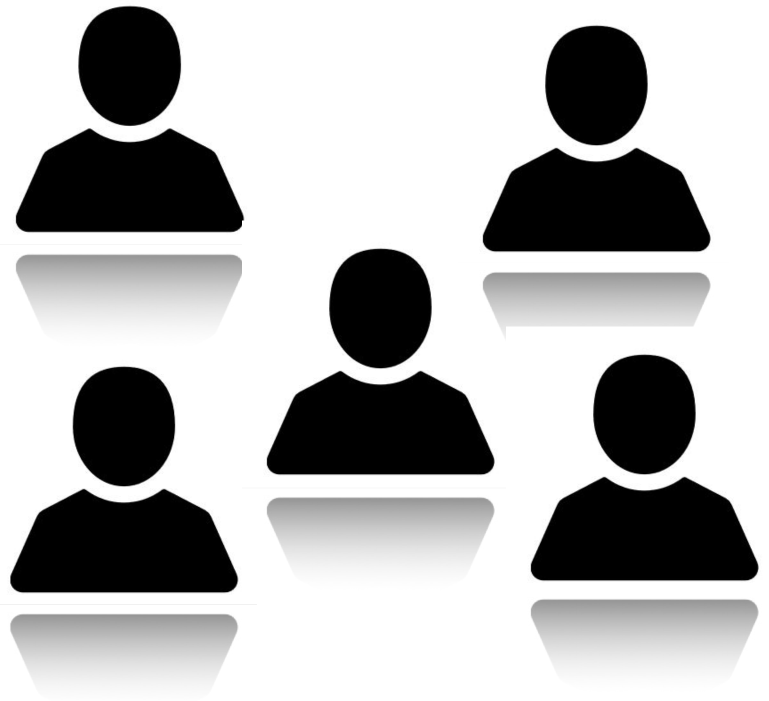
Two dogs are running through a field.

Premise

NLI Dataset Creation Process

Two dogs are running through a field.

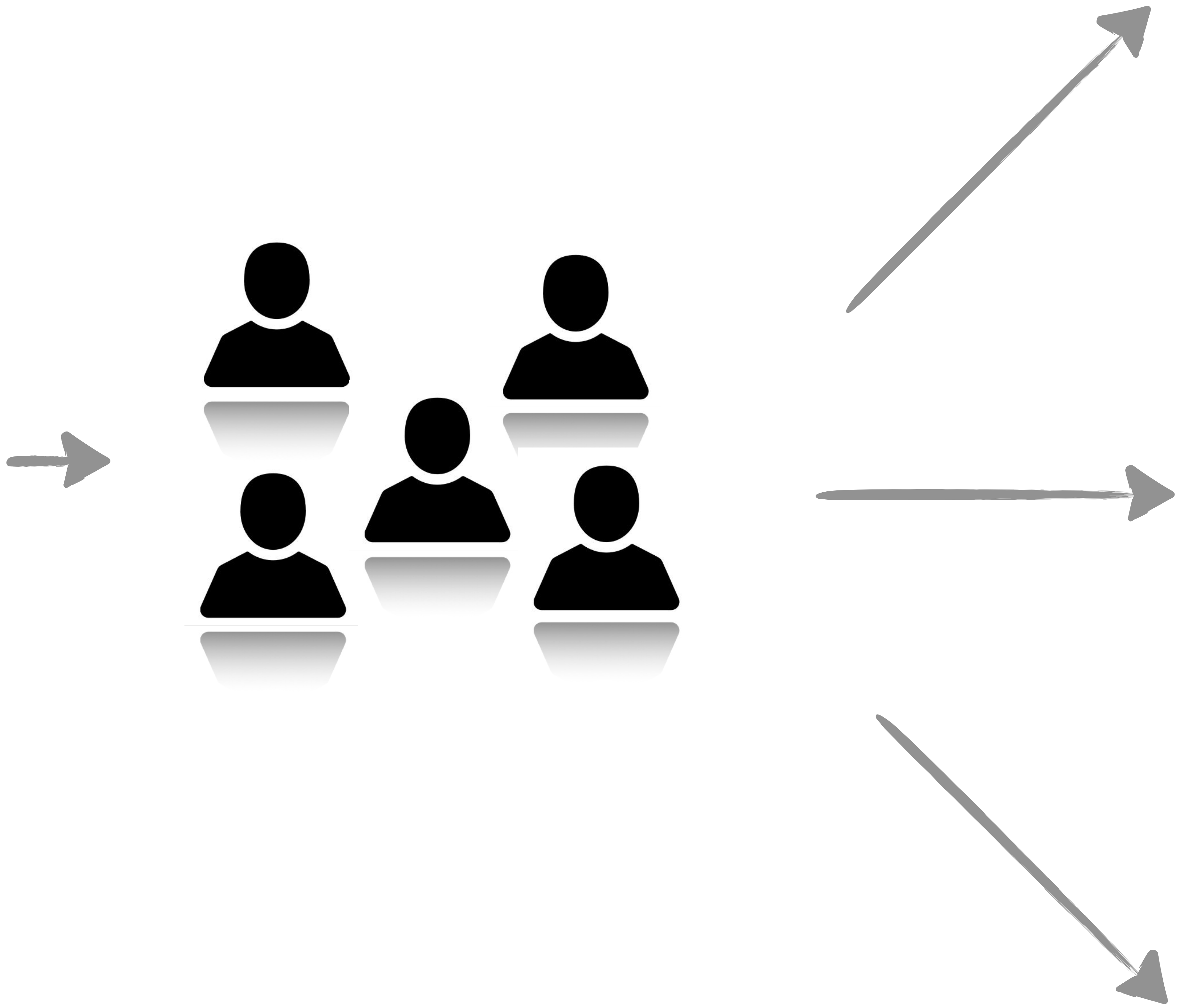
Premise



NLI Dataset Creation Process

Two dogs are running through a field.

Premise



NLI Dataset Creation Process

Two dogs are running through a field.

Premise

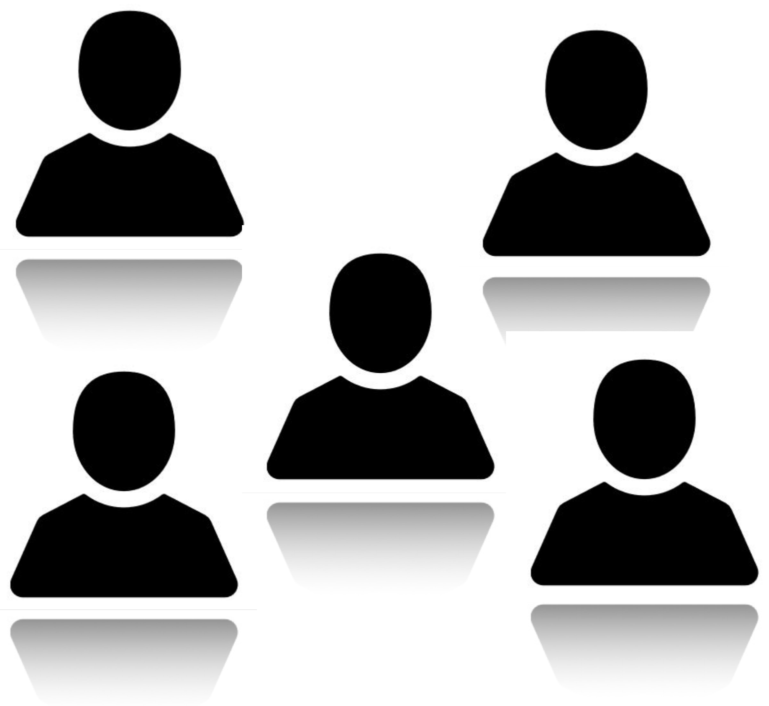
Entailment

There are animals outdoors.

NLI Dataset Creation Process

Two dogs are running through a field.

Premise



Entailment

There are animals outdoors.

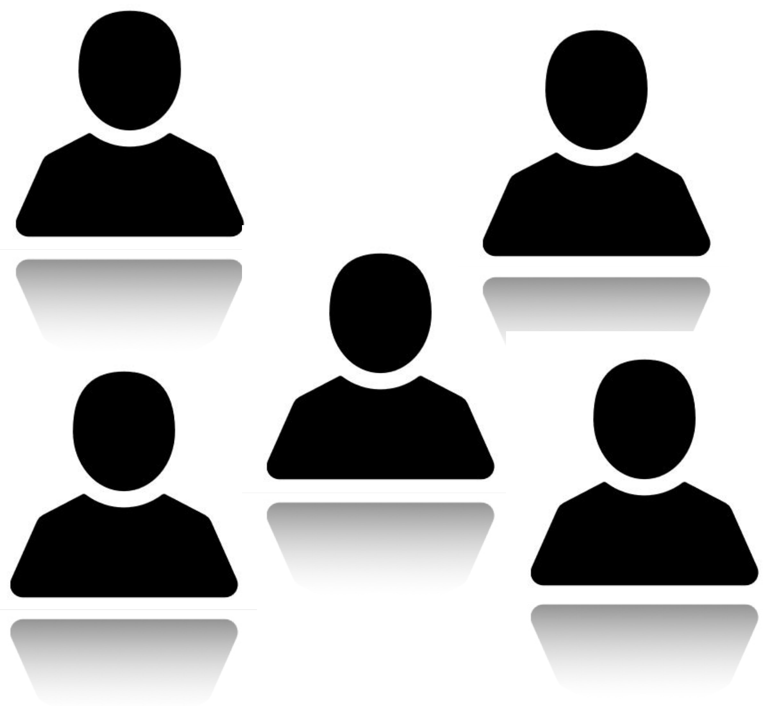
Neutral

Some puppies are running to catch a stick.

NLI Dataset Creation Process

Two dogs are running through a field.

Premise



Entailment → There are animals outdoors.

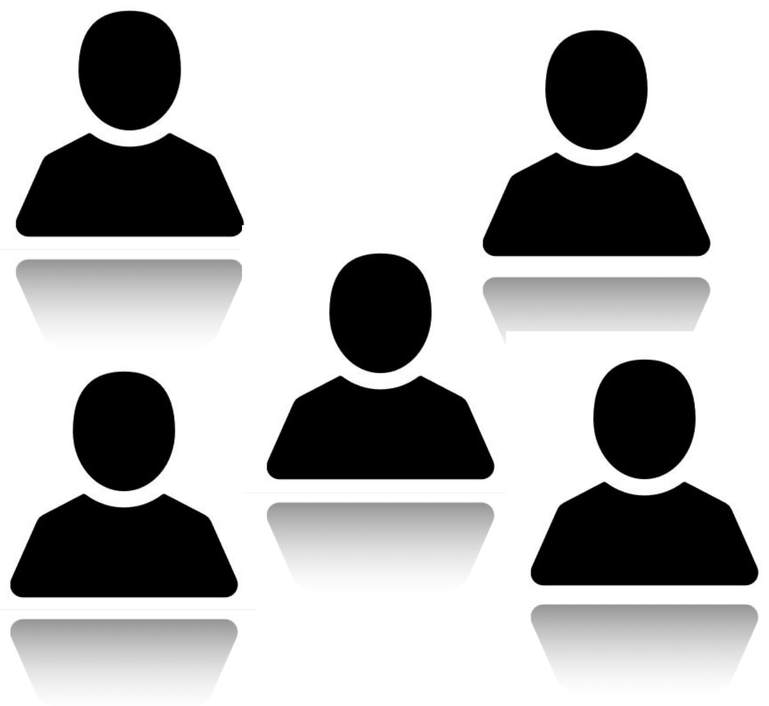
Neutral → Some puppies are running to catch a stick.

Contradiction → The pets are sitting on a couch.

NLI Dataset Creation Process

Two dogs are running through a field.

Premise



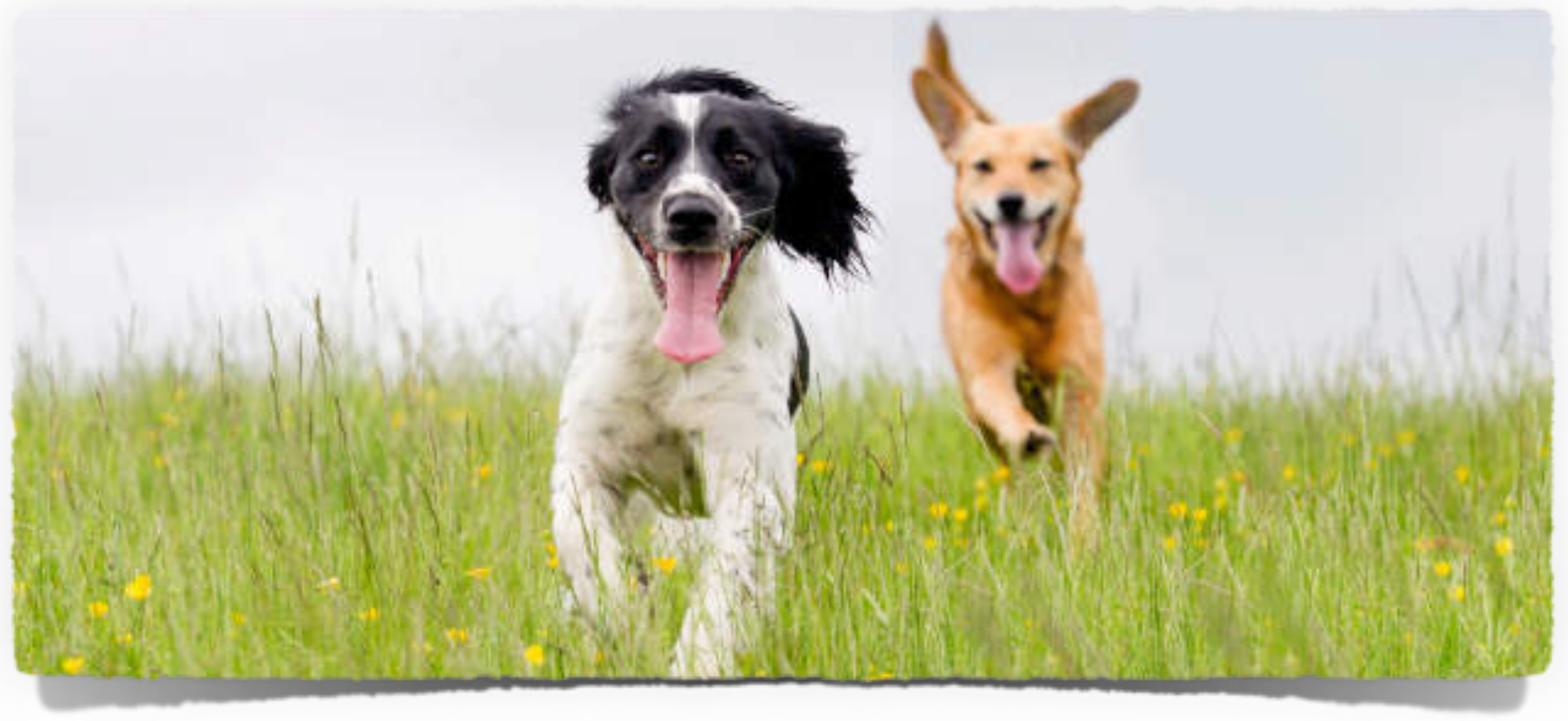
Entailment → There are animals outdoors.

Neutral → Some puppies are running to catch a stick.

Contradiction → The pets are sitting on a couch.

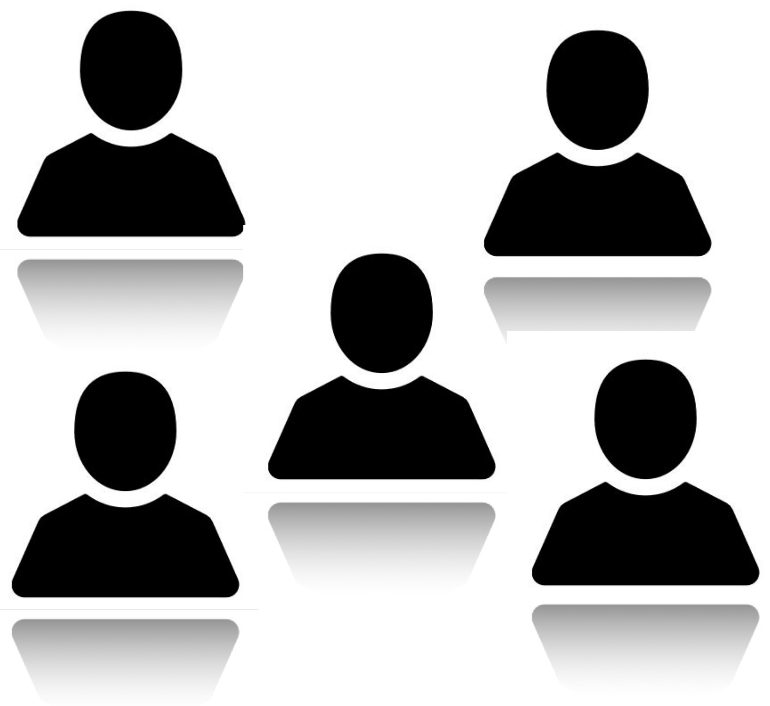
- **Stanford NLI** [Bowman et. al, 2015] 570 K

NLI Dataset Creation Process



Two dogs are running through a field.

Premise



Entailment

There are animals outdoors.

Neutral

Some puppies are running to catch a stick.

Contradiction

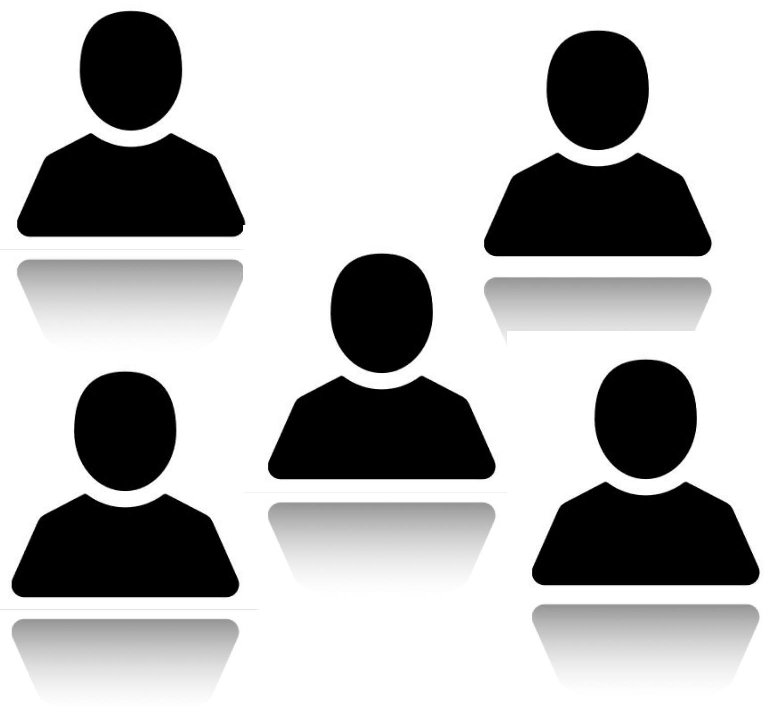
The pets are sitting on a couch.

- **Stanford NLI** [Bowman et. al, 2015] 570 K
- **Multi-genre NLI** [Williams et. al., 2017] 433 K

NLI Dataset Creation Process

Two dogs are running through a field.

Premise



Entailment → There are animals outdoors.

Neutral → Some puppies are running to catch a stick.

Contradiction → The pets are sitting on a couch.

- **Stanford NLI** [Bowman et. al, 2015] 570 K
- **Multi-genre NLI** [Williams et. al., 2017] 433 K
  - Matched and Mismatched Test Sets

Leaderboard progress

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	**bgm**			0.90557	4	14d
2	**Haoming Jiang**			0.87923	10	1mo
3	**Xiaodong Liu**			0.86443	4	10mo
4	**Anonymous**			0.86351	2	1y
5	**anonymous11111**			0.85177	18	1mo
6	**Ariel**			0.85065	41	5mo
7	**sherry77**			0.85034	17	5mo
⋮						
	Bidirectional LSTM			0.67507		
104	**gabrielalmeida**			0.67313	5	8mo
105	**Zippy**			0.67160	2	1y
106	**kudkudak**			0.66435	2	1y
107	**Shawn Tan**			0.65271	1	6d
	CBOW			0.65200		

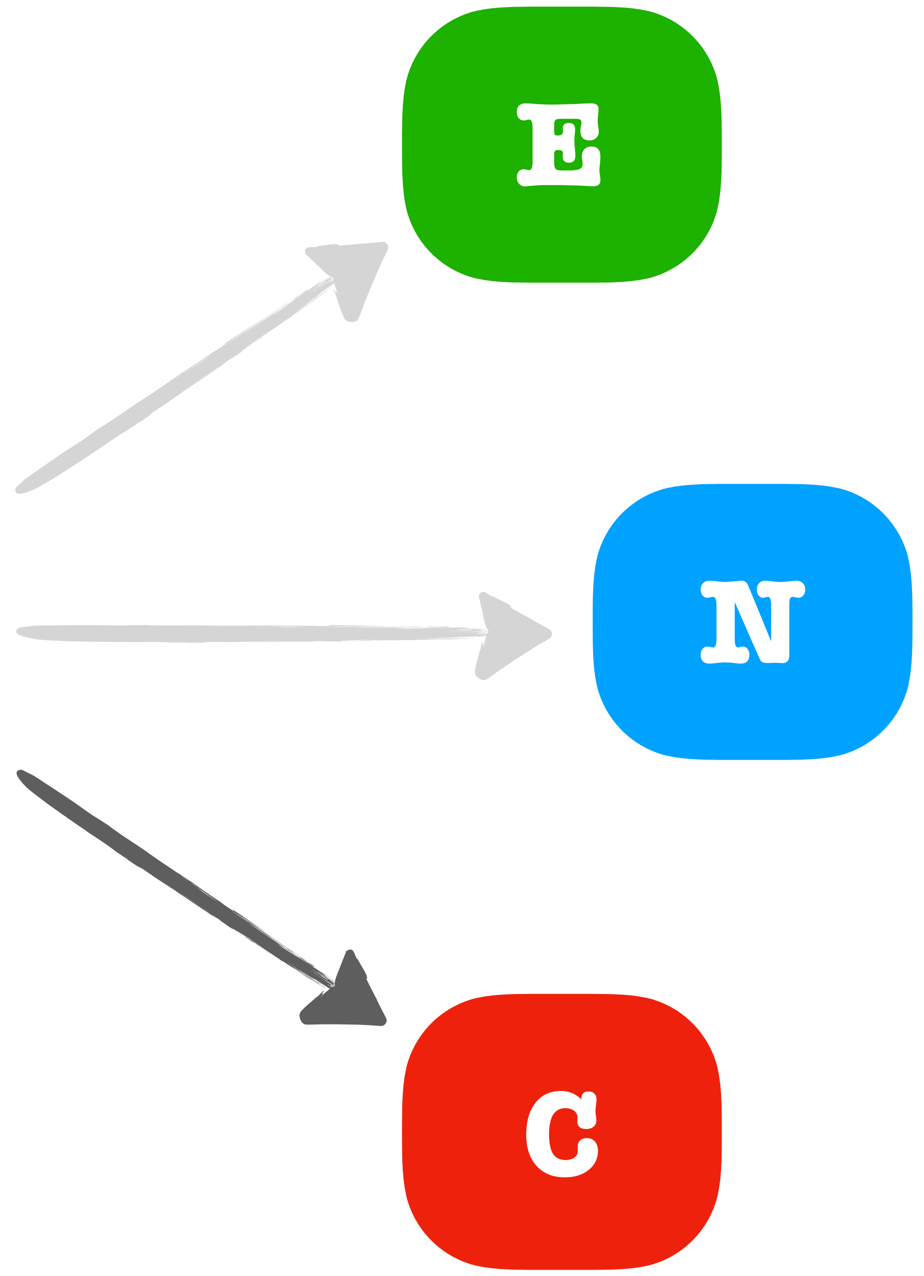
NLI as Text Classification

Two dogs are running through a field.

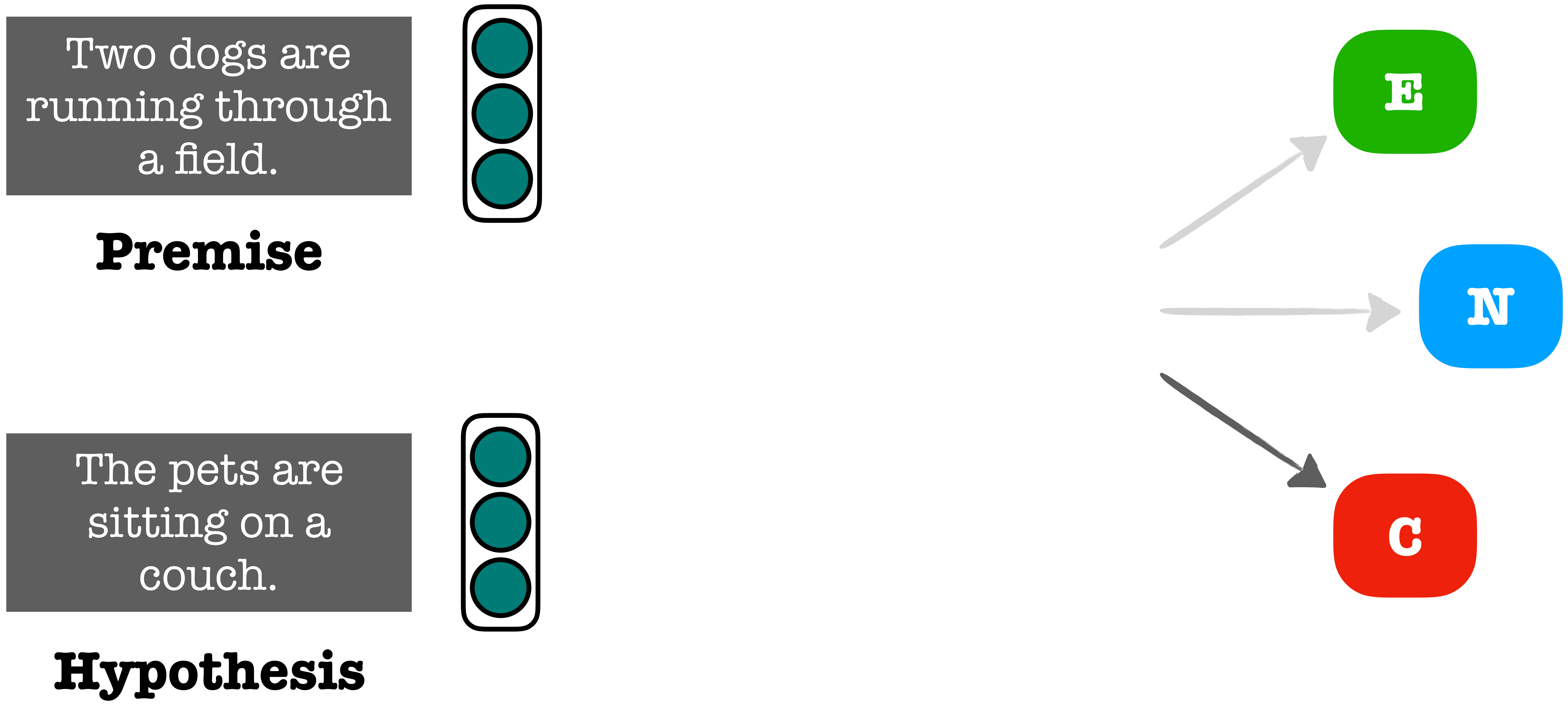
Premise

The pets are sitting on a couch.

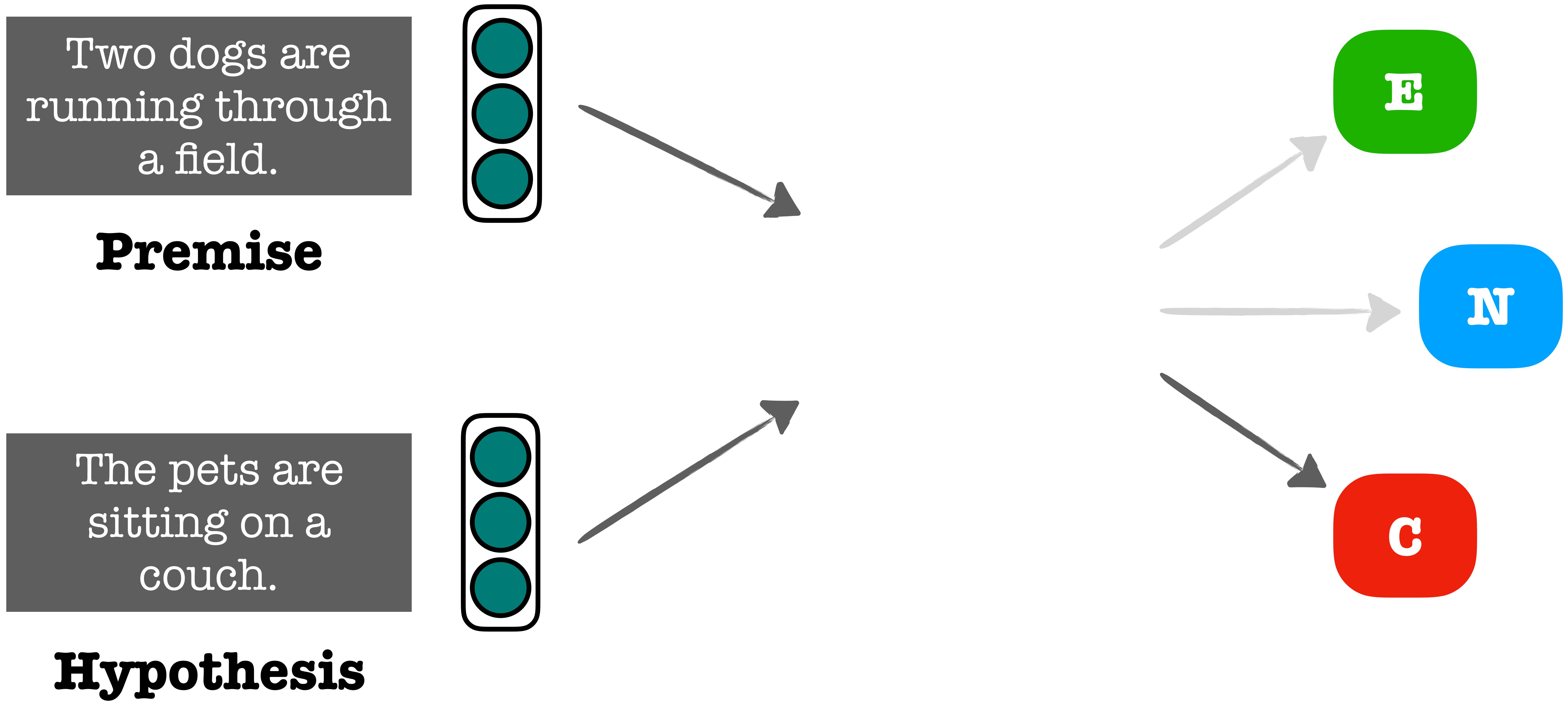
Hypothesis



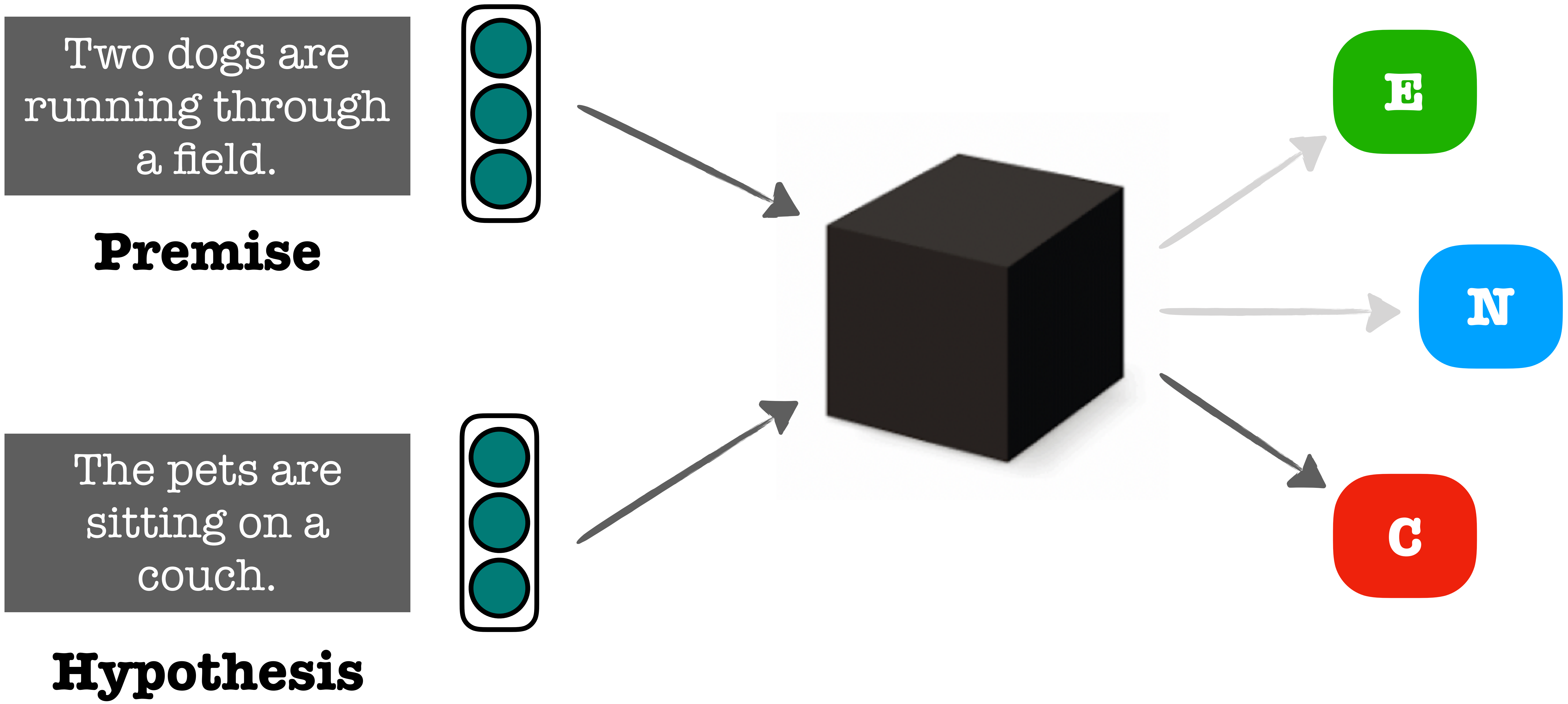
NLI as Text Classification



NLI as Text Classification

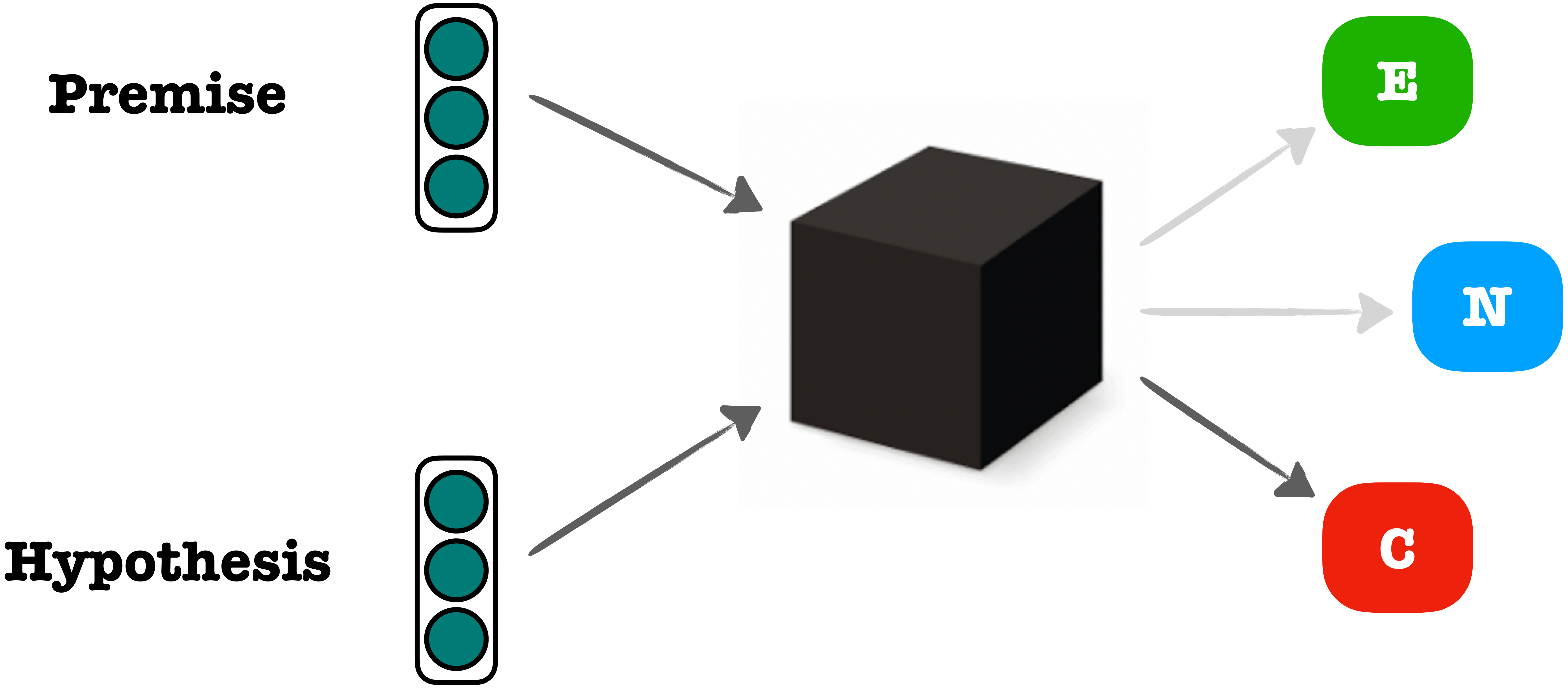


NLI as Text Classification



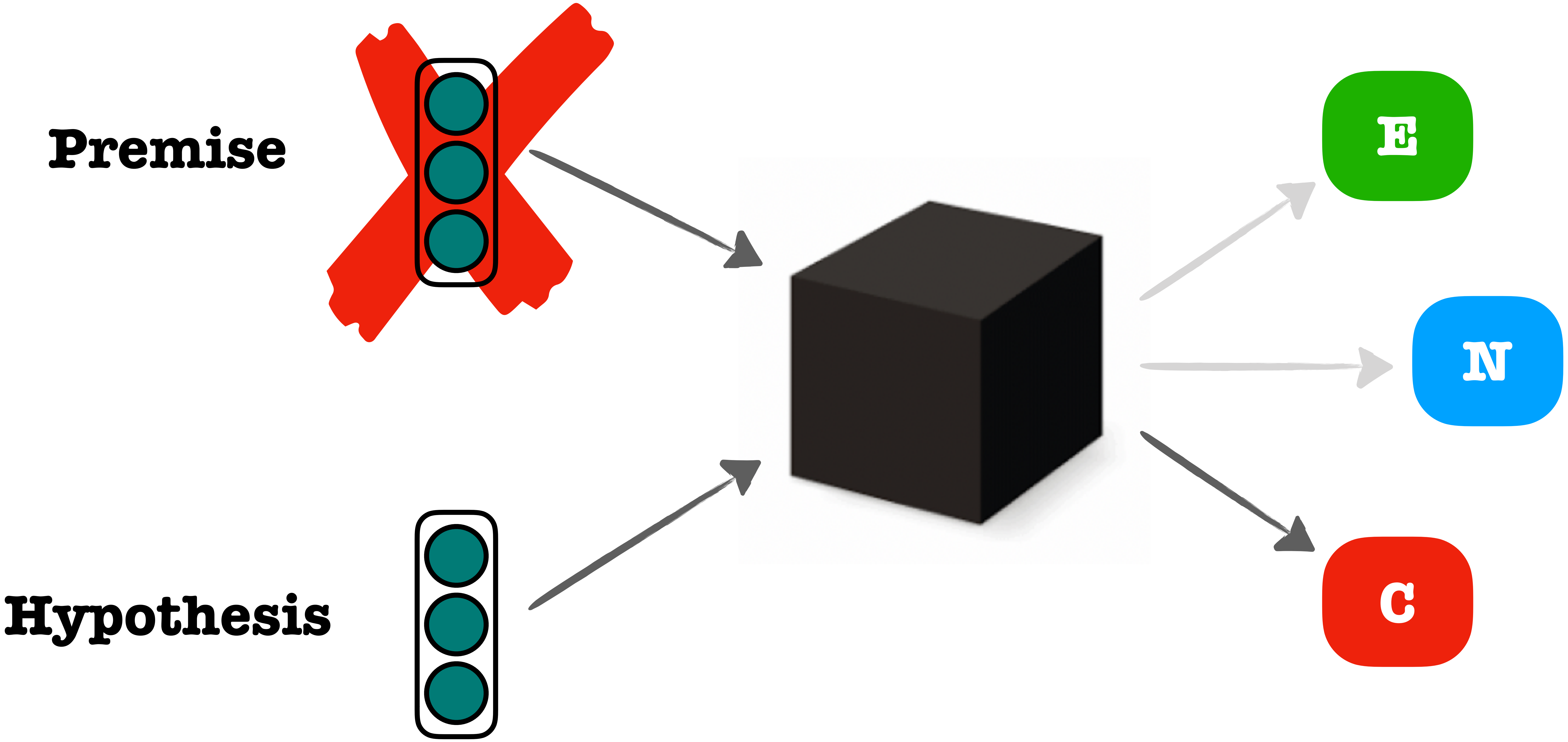
DAM - Decomposable Attention Model (Parikh et. al. 2016)
ESIM - Enhanced Sequential Inference Model (Chen et. al., 2017)
DIIN - Densely Interactive Inference Network (Gong et. al. 2018)

A simple experiment



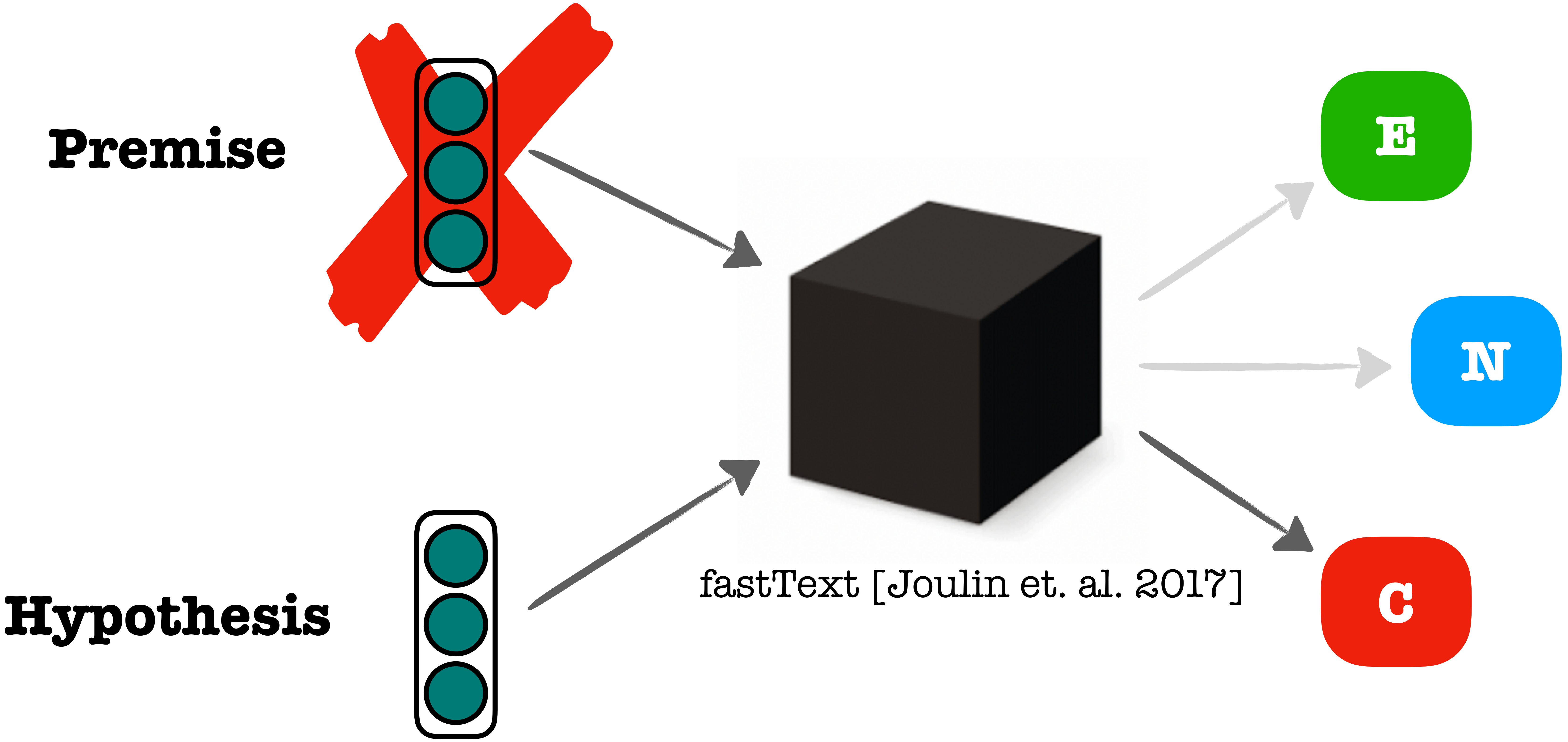
Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

A simple experiment



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

A simple experiment



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Intuitively…

Given **no** premise, is a hypothesis true, false or neither?

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

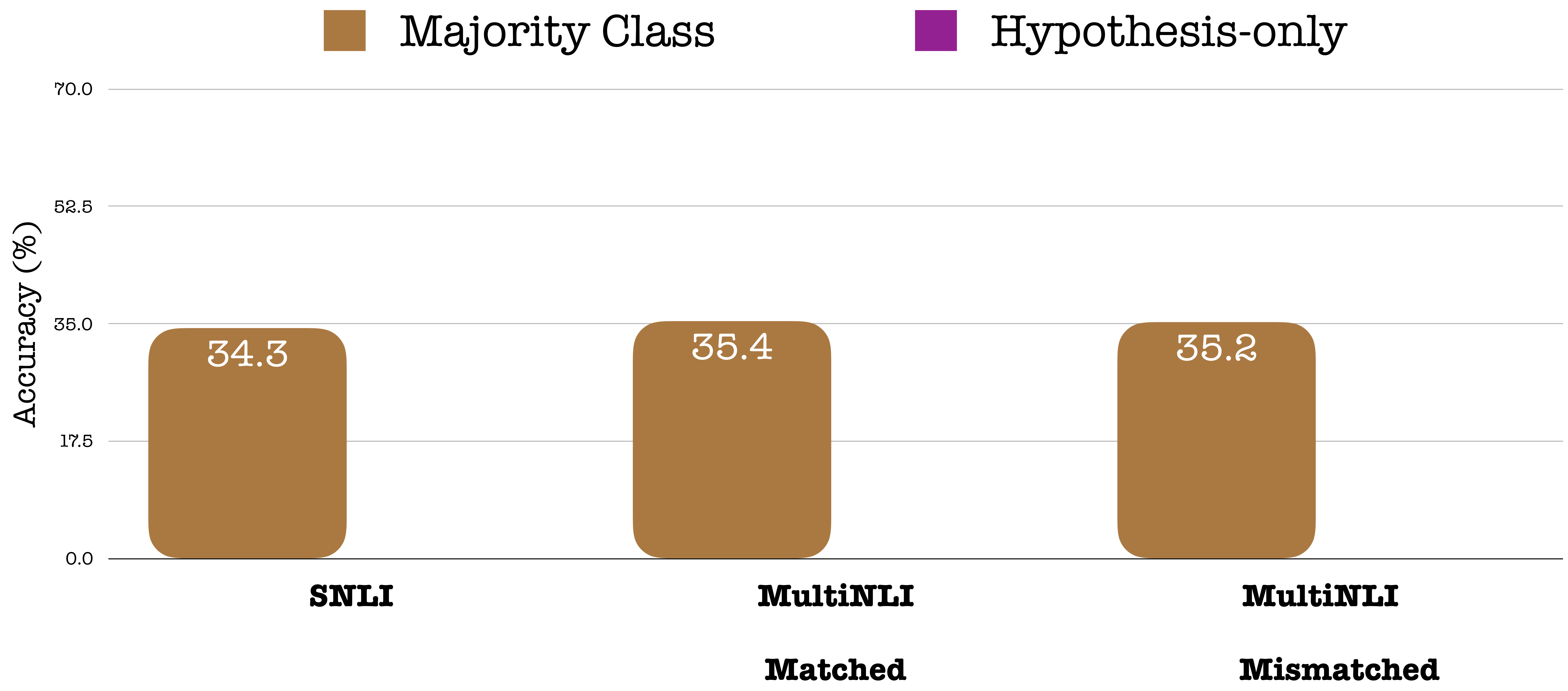
Intuitively…

Hypothesis The little boy is diving off the diving board because he is an excellent swimmer.

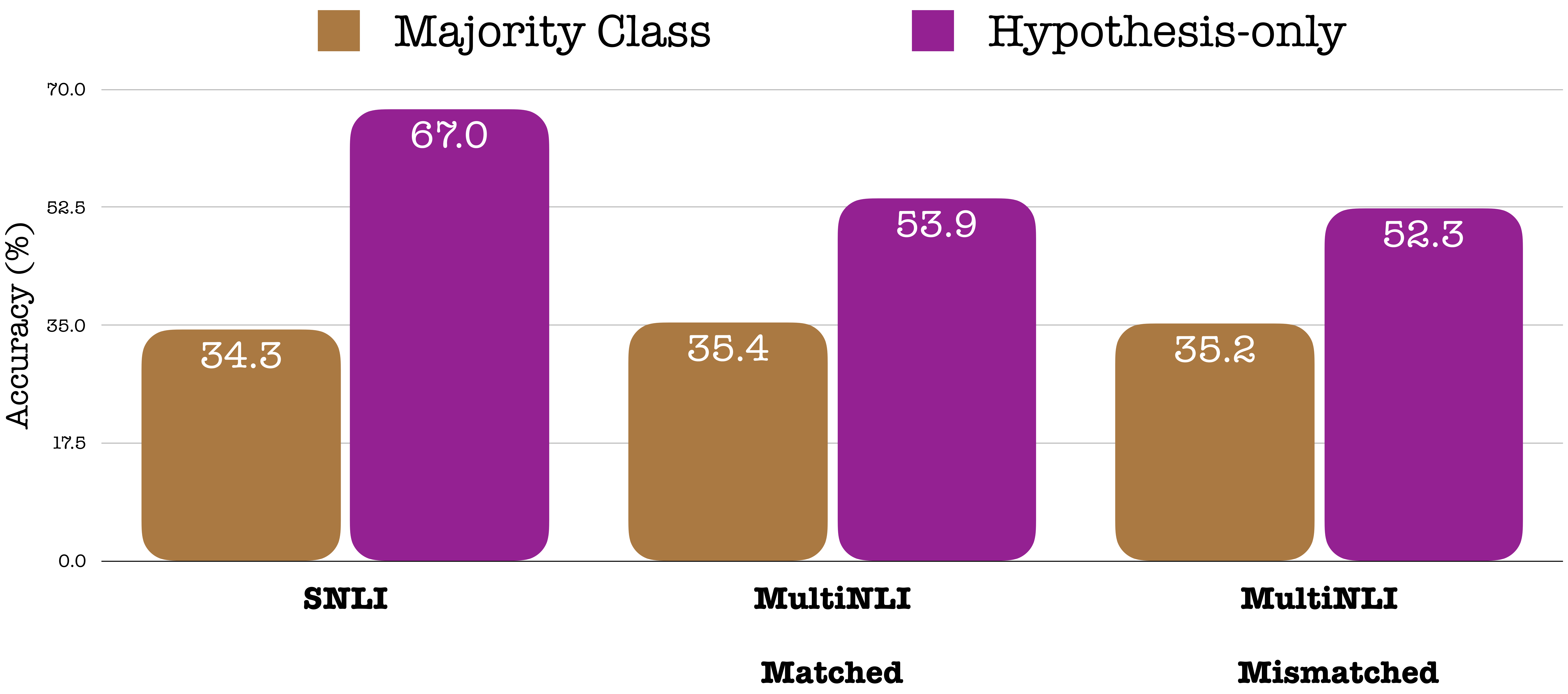
Given **no** premise, is a hypothesis true, false or neither?

- True → **Entailment**
- False → **Contradiction**
- Cannot Say → **Neutral**

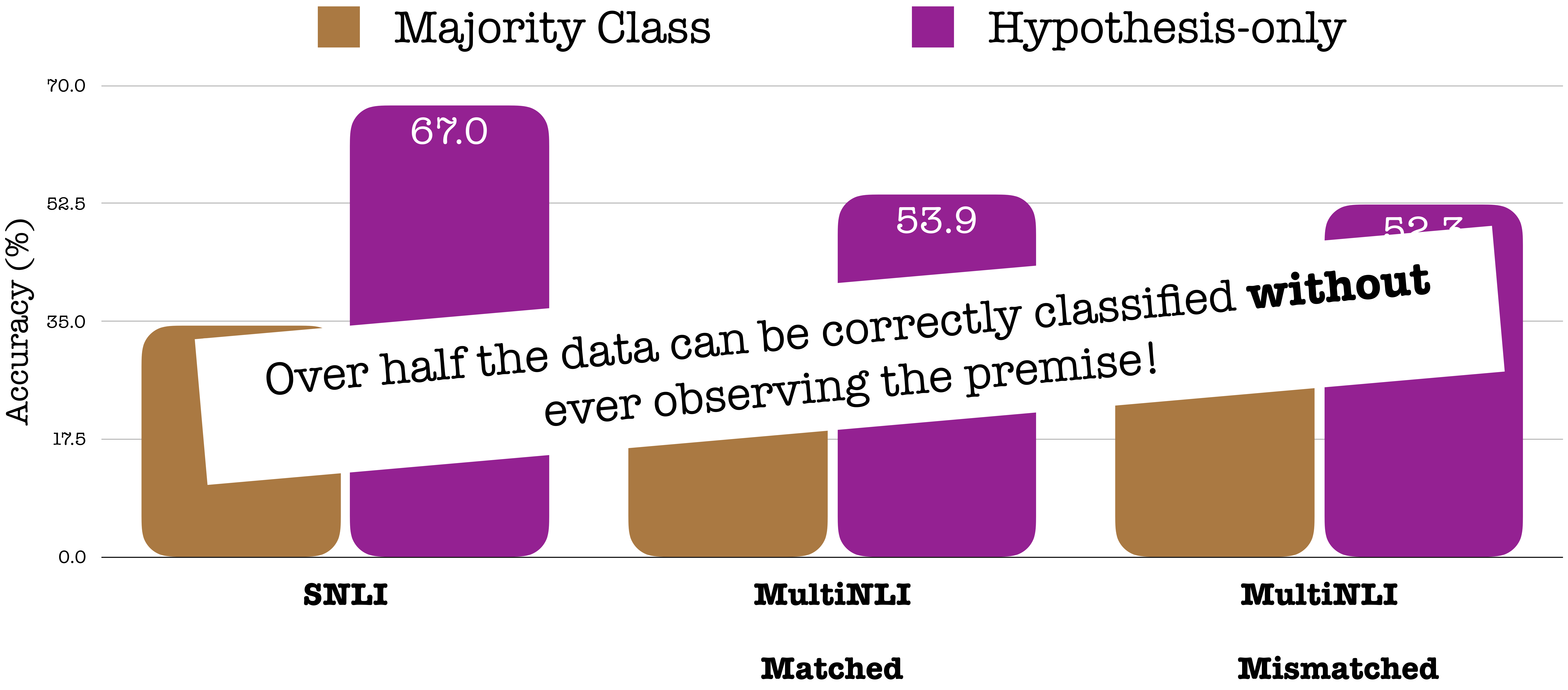
Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

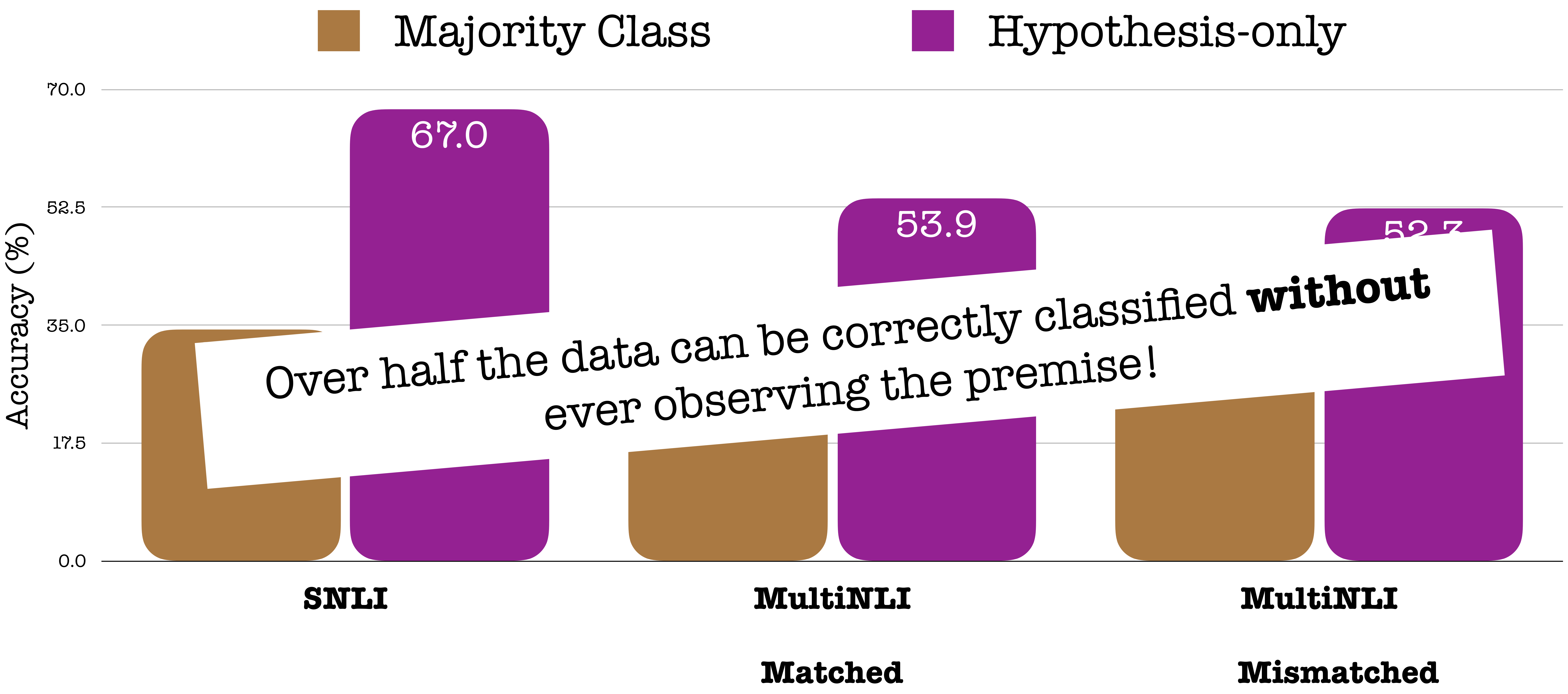


Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]



Over half the data can be correctly classified **without** ever observing the premise!

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]



Poliak et. al., 2018,
Glockner et. al., 2018

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Digging Deeper

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Digging Deeper

- **Annotation Artifacts**: Clues which give away the correct prediction without any reasoning.

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Digging Deeper

- **Annotation Artifacts**: Clues which give away the correct prediction without any reasoning.
- Hypothesis-only artifacts are class-specific.

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Digging Deeper

- **Annotation Artifacts**: Clues which give away the correct prediction without any reasoning.
- Hypothesis-only artifacts are class-specific.
- Word-class association via PPMI:

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Digging Deeper

- **Annotation Artifacts**: Clues which give away the correct prediction without any reasoning.

- Hypothesis-only artifacts are class-specific.

- Word-class association via PPMI:

$$\max\{0, \log \frac{p(\textbf{word}, \textbf{class})}{p(\textbf{.}, \textbf{class})p(\textbf{word}, \textbf{.})}\}$$

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Digging Deeper

- **Annotation Artifacts**: Clues which give away the correct prediction without any reasoning.
- Hypothesis-only artifacts are class-specific.
- Word-class association via PPMI:

$$\max\{0, \log \frac{p(\textbf{word}, \textbf{class})}{p(., \textbf{class})\,p(\textbf{word}, .)}\}$$

N	C	E
Outdoors	Nobody	Tall
Least	Sleeping	First
Instrument	No	Competition
Outside	Tv	Sad
Animal	Cat	Favorite

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Entailment Artifacts

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Entailment Artifacts

Some men and boys are playing frisbee in a grassy area.

Premise

Generalization

People play frisbee **outdoors**.

Hypothesis

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Entailment Artifacts

Some men and boys are playing frisbee in a grassy area.

Premise

Generalization

People play frisbee **outdoors**.

Hypothesis

A person in a red ~~**shirt**~~ is mowing the grass with a ~~**green**~~ riding mower.

Premise

Simplification

A person in red is cutting the grass on a riding mower.

Hypothesis

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Neutral Artifacts

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Neutral Artifacts

Modifiers

A man is doing work on a **black** Amtrak train.

Hypothesis

A middle-aged man works under the engine of a train on rail tracks.

Premise

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Neutral Artifacts

Modifiers

A man is doing work on a **black** Amtrak train.

Hypothesis

A middle-aged man works under the engine of a train on rail tracks.

Premise

A group of female athletes are huddled together and excited.

Premise

Purpose Clauses

They are huddled together **because** they are working together.

Hypothesis

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Contradiction Artifacts

Contradiction Artifacts

Older man with white hair and a red cap painting the golden gate bridge on the shore with the golden gate bridge in the distance.

Premise

Negation

Nobody wears a cap.

Hypothesis

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Contradiction Artifacts

Older man with white hair and a red cap painting the golden gate bridge on the shore with the golden gate bridge in the distance.

Premise

Negation

Nobody wears a cap.

Hypothesis

Three dogs racing on racetrack.

Premise

Cats!

Three **cats** race on a track.

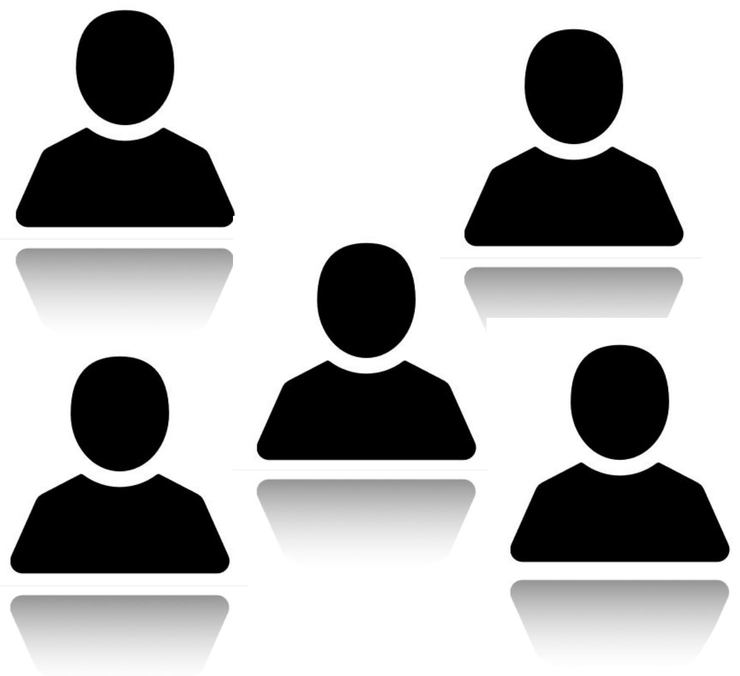
Hypothesis

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

A possible explanation

Two dogs are running through a field.

Premise



Entailment → There are **animals** outdoors.

Neutral → Some puppies are running **to catch a stick**.

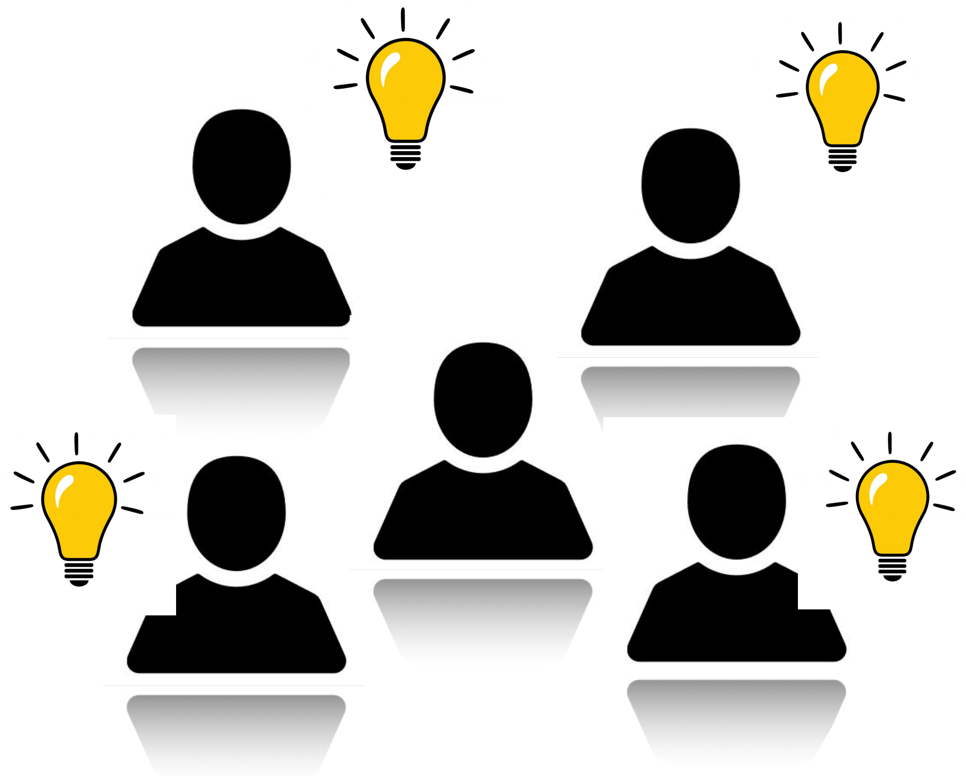
Contradiction → The pets are sitting on a couch.

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

A possible explanation

Two dogs are running through a field.

Premise



Entailment → There are **animals** outdoors.

Neutral → Some puppies are running **to catch a stick**.

Contradiction → The pets are sitting on a couch.

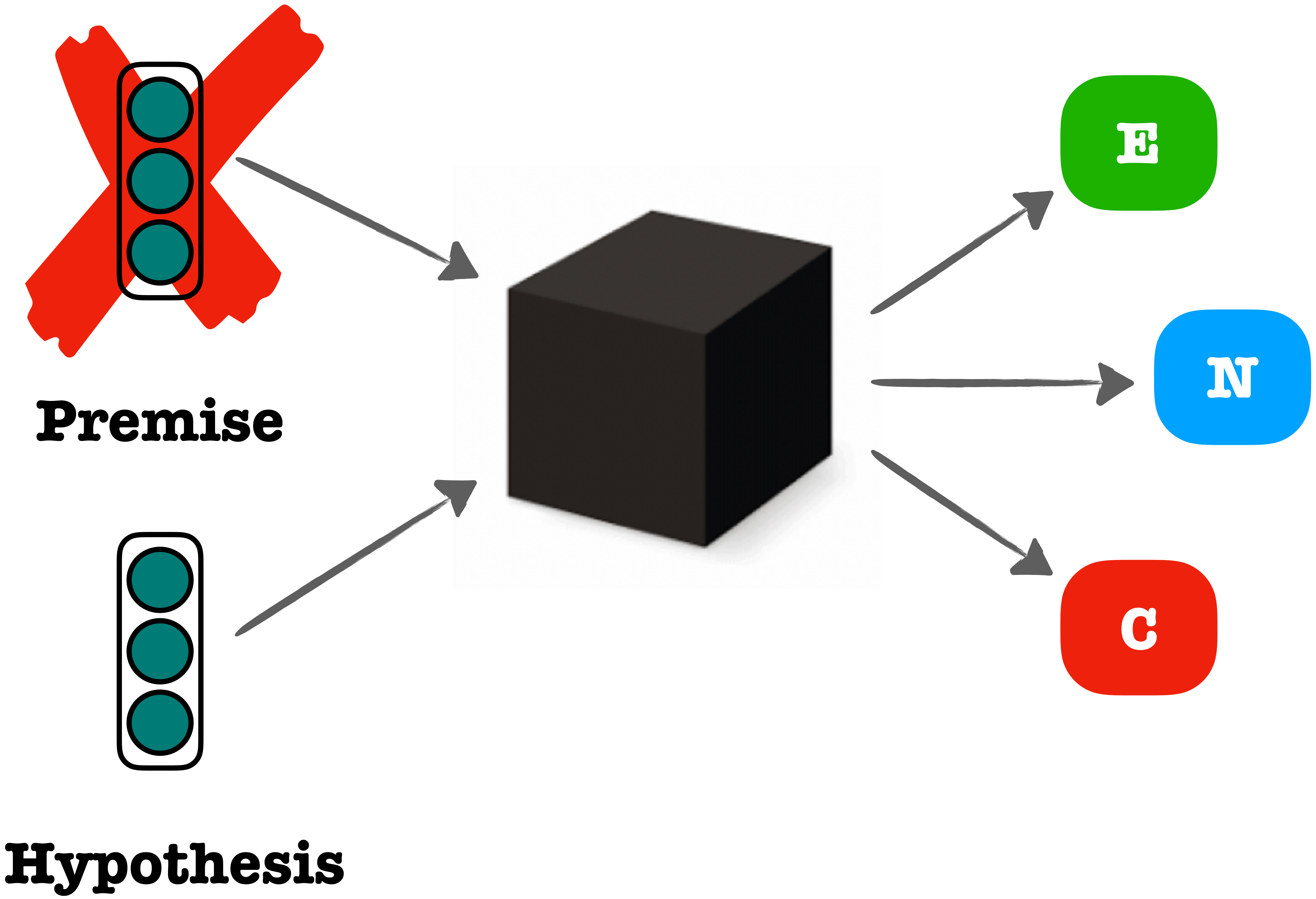
Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Are seed examples responsible?

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.

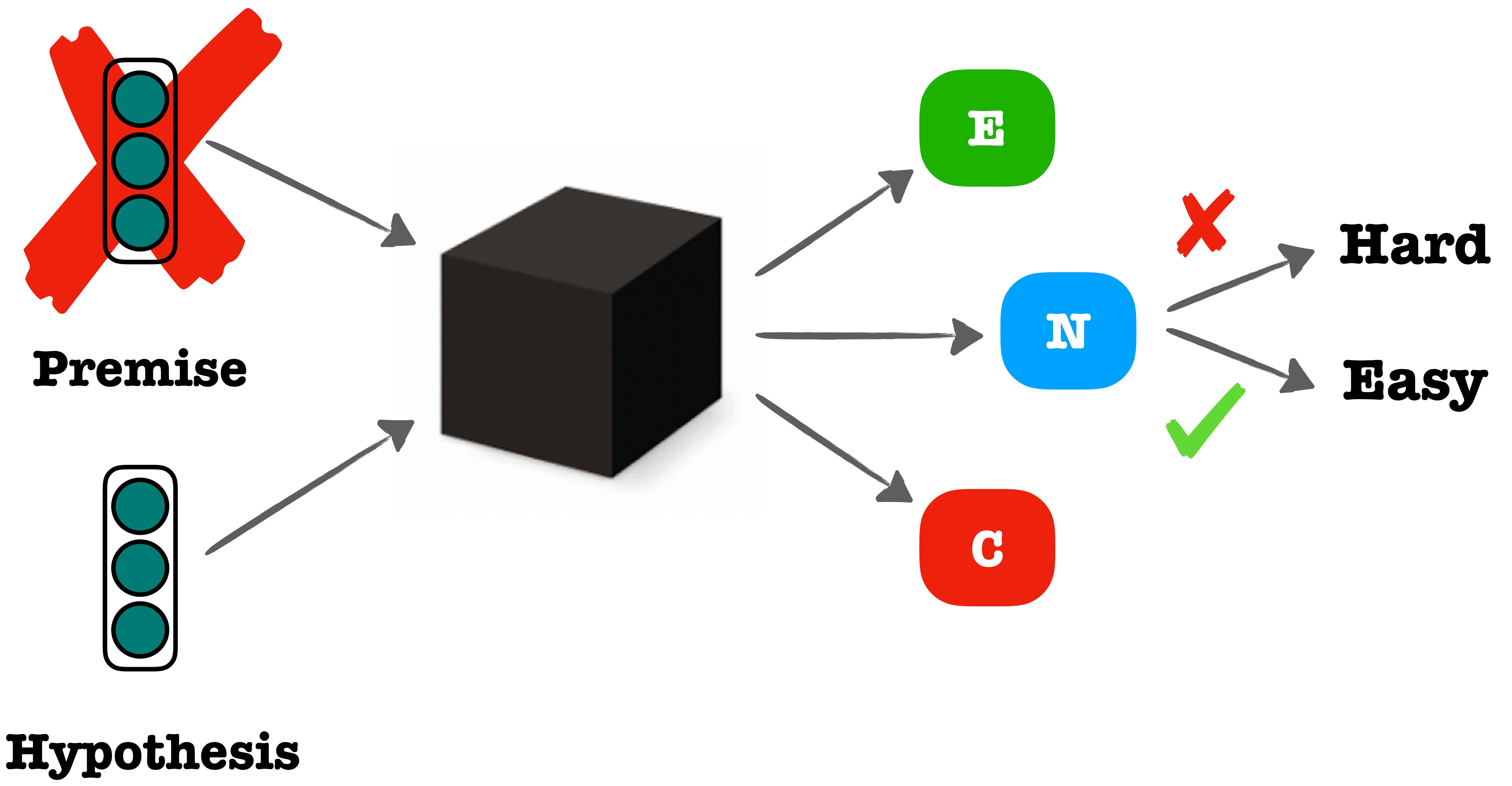
Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Identifying examples with artifacts



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Identifying examples with artifacts

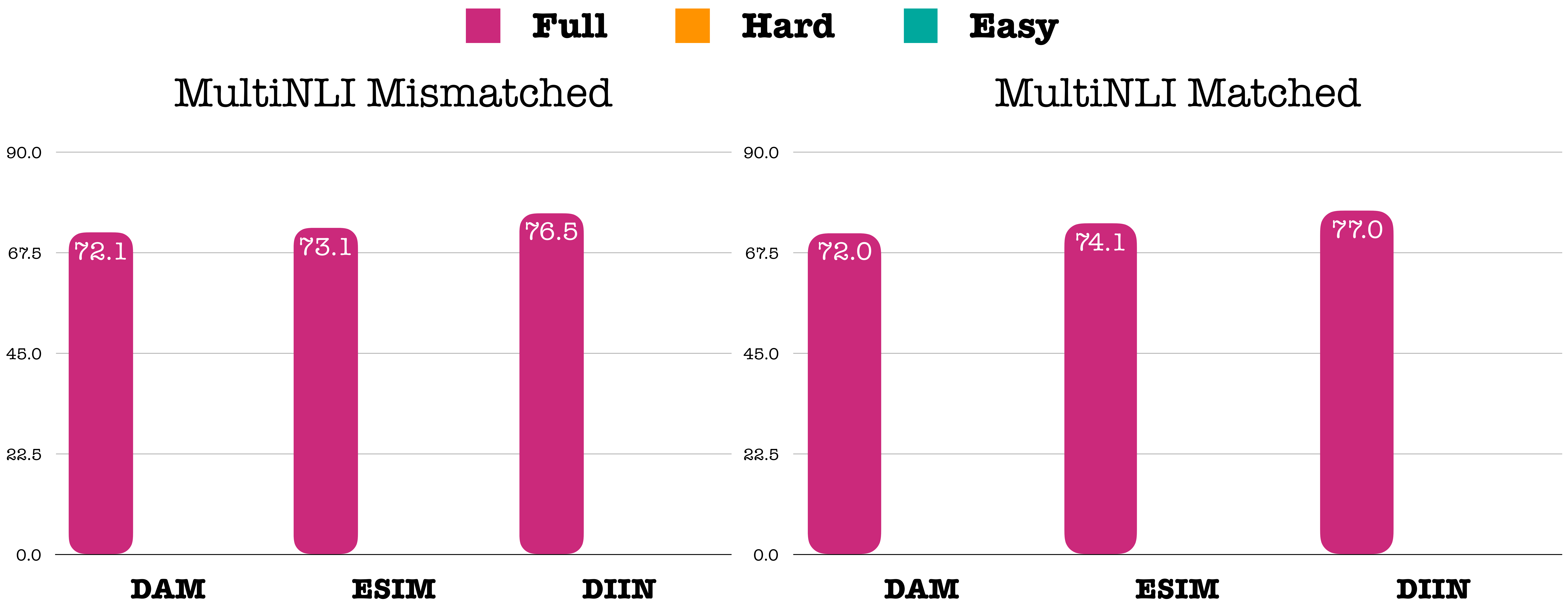


Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Revisiting NLI models

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

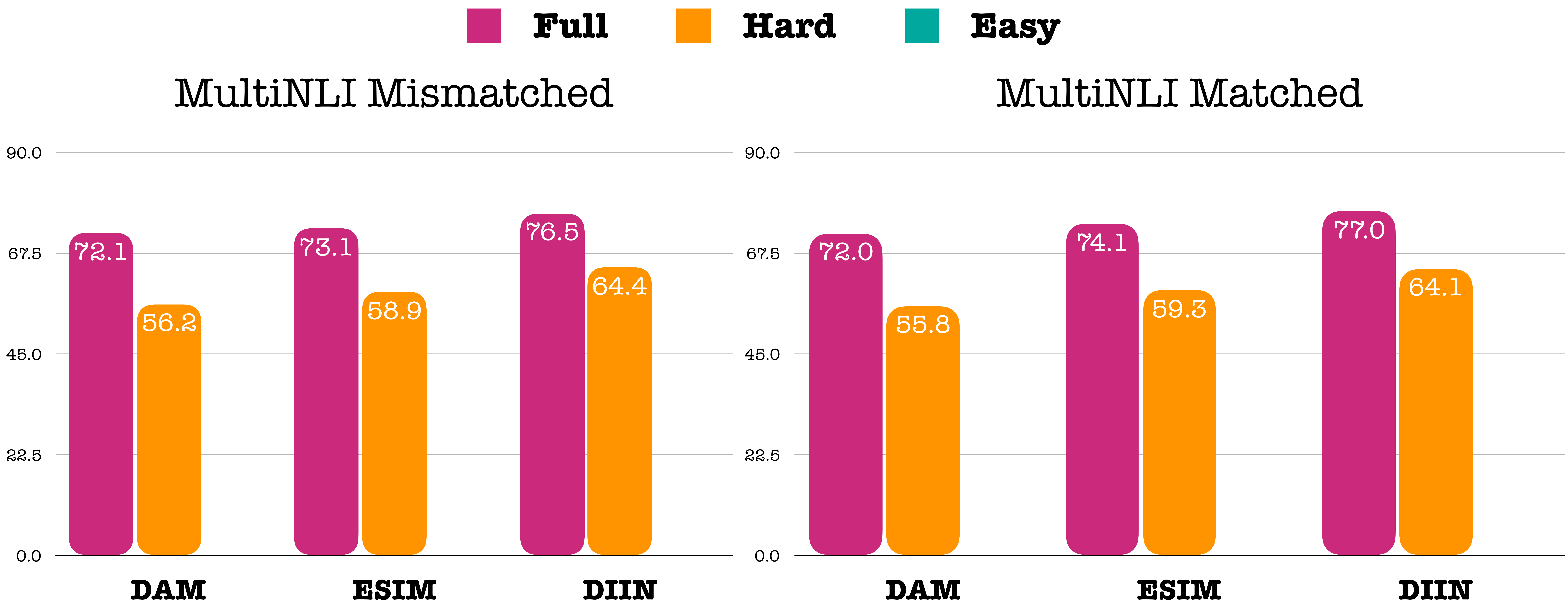
Revisiting NLI models



DAM - Decomposable Attention Model [Parikh et. al. 2016]
ESIM - Enhanced Sequential Inference Model [Chen et. al., 2017]
DIIN - Densely Interactive Inference Network [Gong et. al. 2018]

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

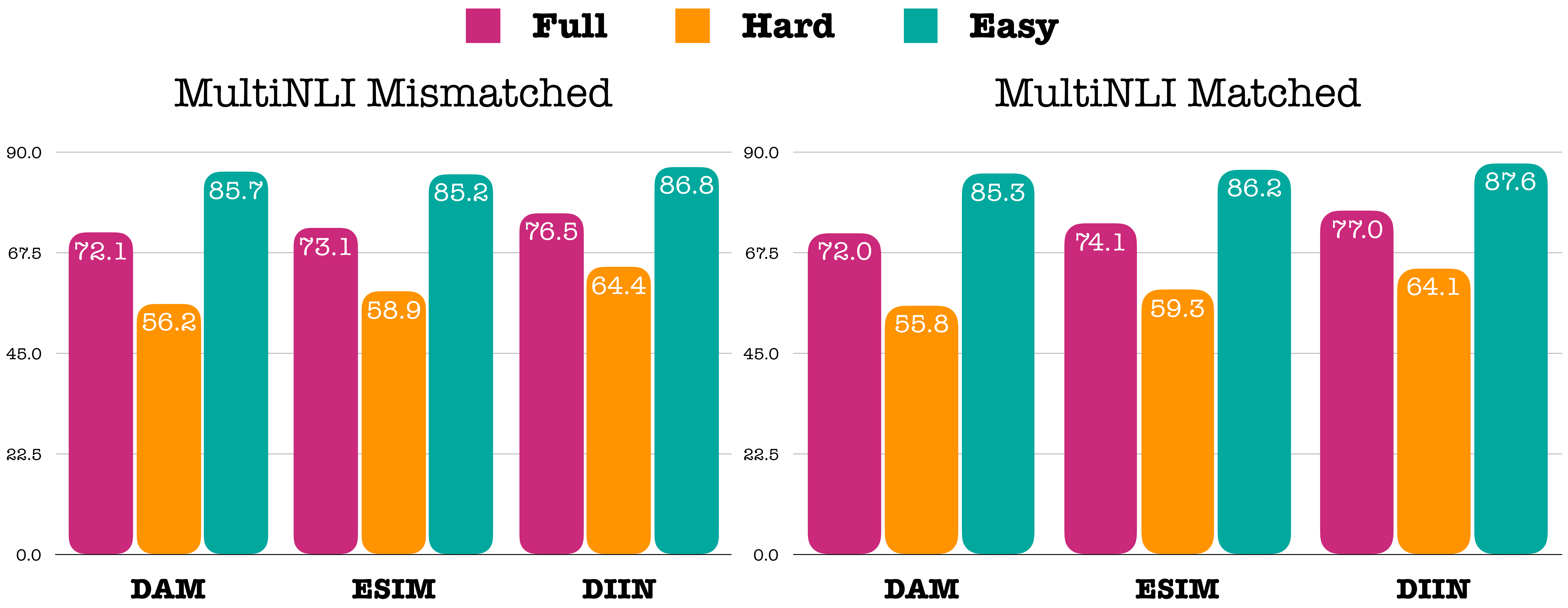
Revisiting NLI models



DAM - Decomposable Attention Model [Parikh et. al. 2016]
ESIM - Enhanced Sequential Inference Model [Chen et. al., 2017]
DIIN - Densely Interactive Inference Network [Gong et. al. 2018]

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

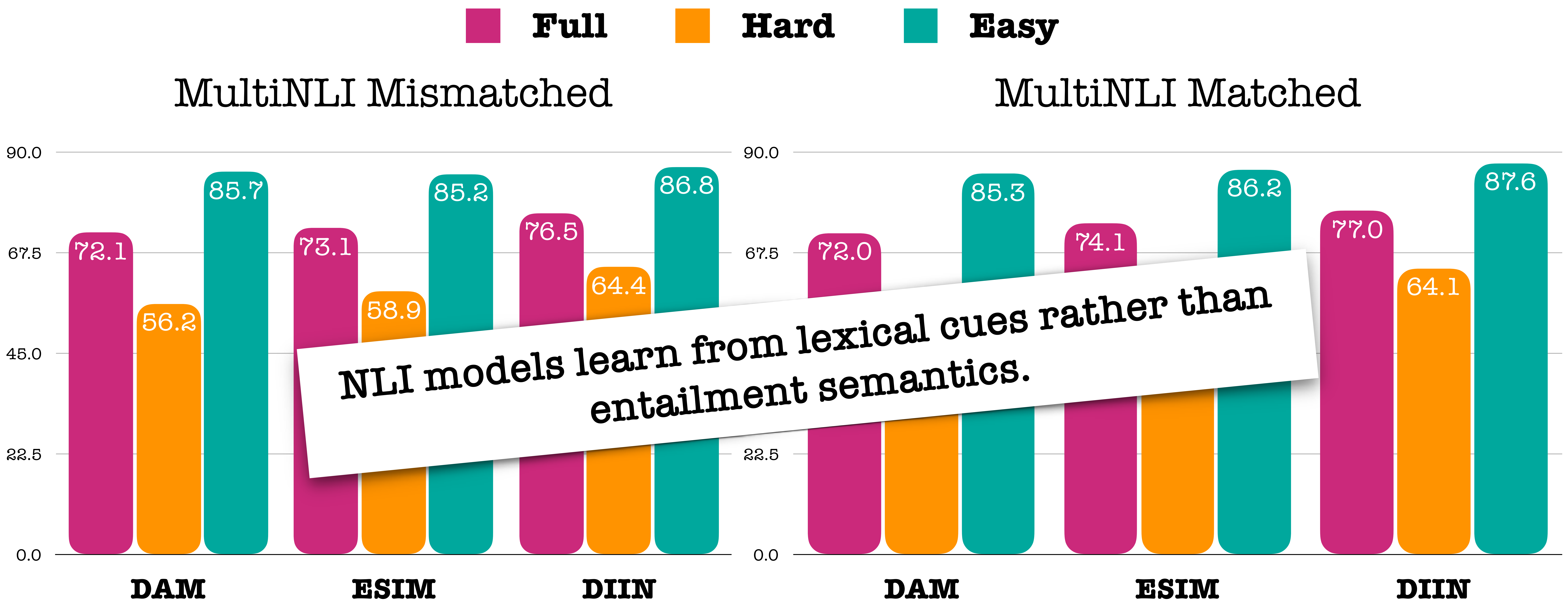
Revisiting NLI models



DAM - Decomposable Attention Model [Parikh et. al. 2016]
ESIM - Enhanced Sequential Inference Model [Chen et. al., 2017]
DIIN - Densely Interactive Inference Network [Gong et. al. 2018]

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Revisiting NLI models



DAM - Decomposable Attention Model [Parikh et. al. 2016]
ESIM - Enhanced Sequential Inference Model [Chen et. al., 2017]
DIIN - Densely Interactive Inference Network [Gong et. al. 2018]

Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Not unique to NLI...

Not unique to NLI...

SQuAD

The Stanford Question Answering Dataset

Jia & Liang et al., 2017

Not unique to NLI…

Story Cloze Test and ROCStories Corpora

Schwartz et al., 2017;
Cai et al., 2017

SQuAD
The Stanford Question Answering Dataset

Jia & Liang et al., 2017

Not unique to NLI…

Story Cloze Test and ROCStories Corpora

Schwartz et al., 2017;
Cai et al., 2017

SQuAD
The Stanford Question Answering Dataset

Jia & Liang et al., 2017

VQA Visual Question Answering

Jabri et al., 2016

Not unique to NLI…

Story Cloze Test and ROCStories Corpora

Schwartz et al., 2017;
Cai et al., 2017

SQuAD
The Stanford Question Answering Dataset

Jia & Liang et al., 2017

VQA Visual Question Answering

Jabri et al., 2016

cnn_dailymail

Chen et al., 2017

Looking Ahead: Improved Data Collection

Looking Ahead: Improved Data Collection

- Partial input baselines. E.g.

Looking Ahead: Improved Data Collection

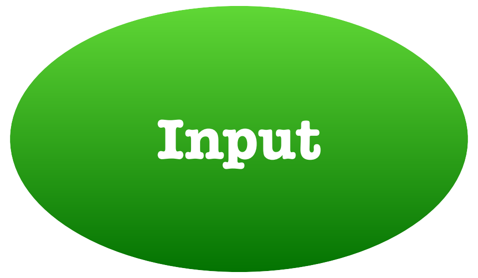
- Partial input baselines. E.g.
  - SWAG [Zellers et. al., 2018],
  - DROP [Dua et. al., 2019],
  - Diverse NLI [Poliak et. al., 2018]

Looking Ahead: Improved Data Collection

- Partial input baselines. E.g.
  - SWAG [Zellers et. al., 2018],
  - DROP [Dua et. al., 2019],
  - Diverse NLI [Poliak et. al., 2018]
- Alternatives to human elicitation for building datasets?

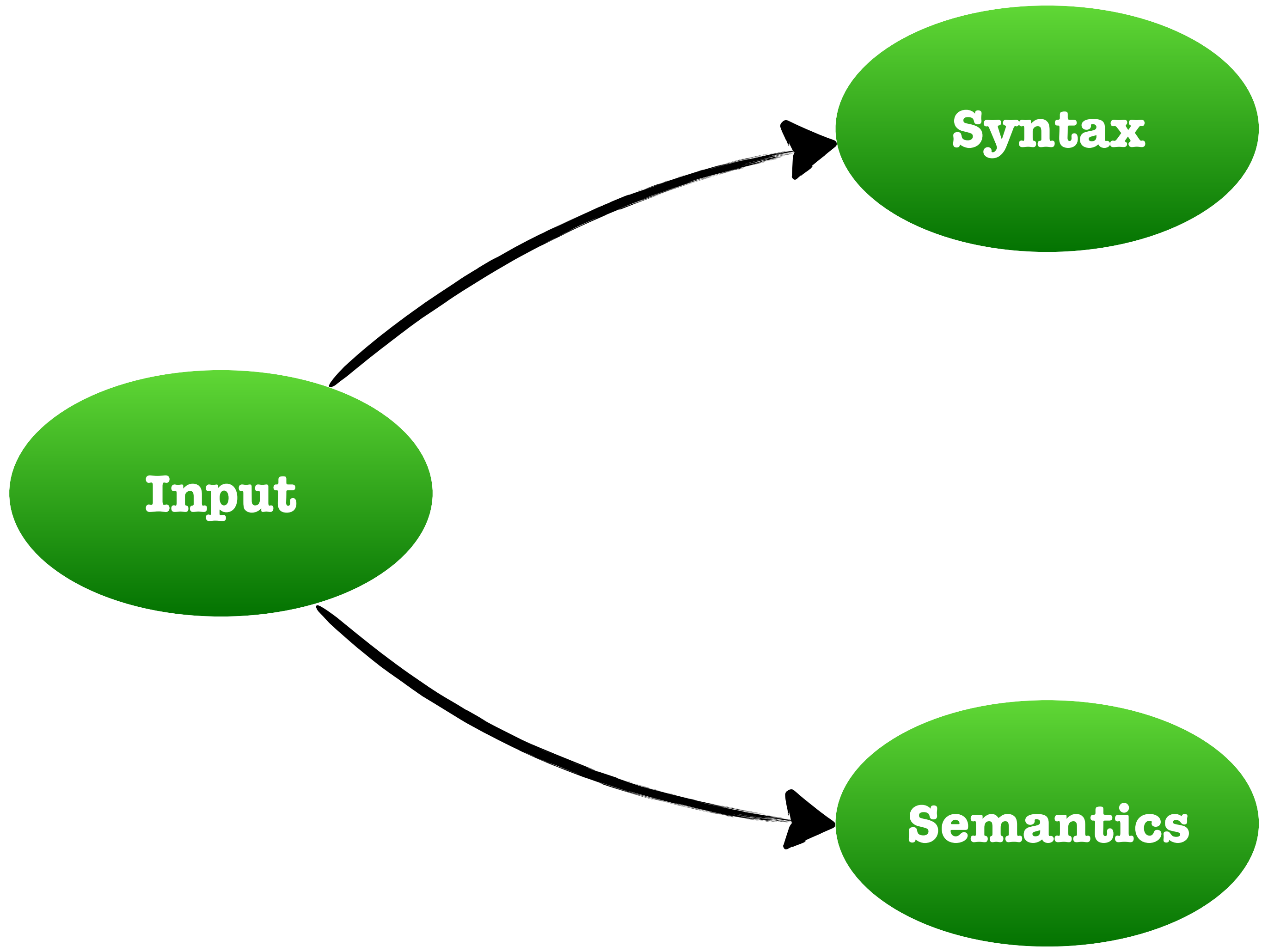
Looking Ahead: Improved Data Collection

- Partial input baselines. E.g.
  - SWAG [Zellers et. al., 2018],
  - DROP [Dua et. al., 2019],
  - Diverse NLI [Poliak et. al., 2018]
- Alternatives to human elicitation for building datasets?



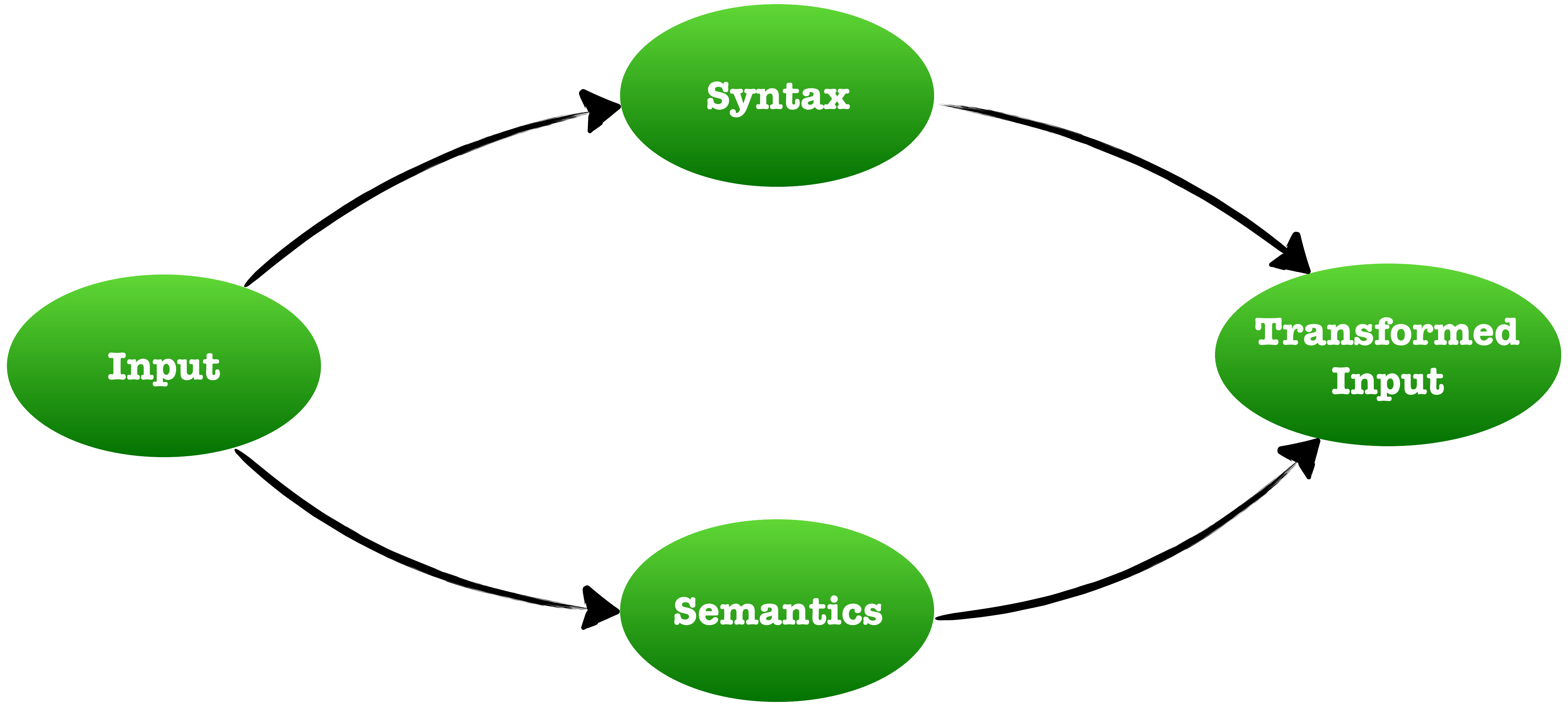
Looking Ahead: Improved Data Collection

- Partial input baselines. E.g.
  - SWAG [Zellers et. al., 2018],
  - DROP [Dua et. al., 2019],
  - Diverse NLI [Poliak et. al., 2018]
- Alternatives to human elicitation for building datasets?

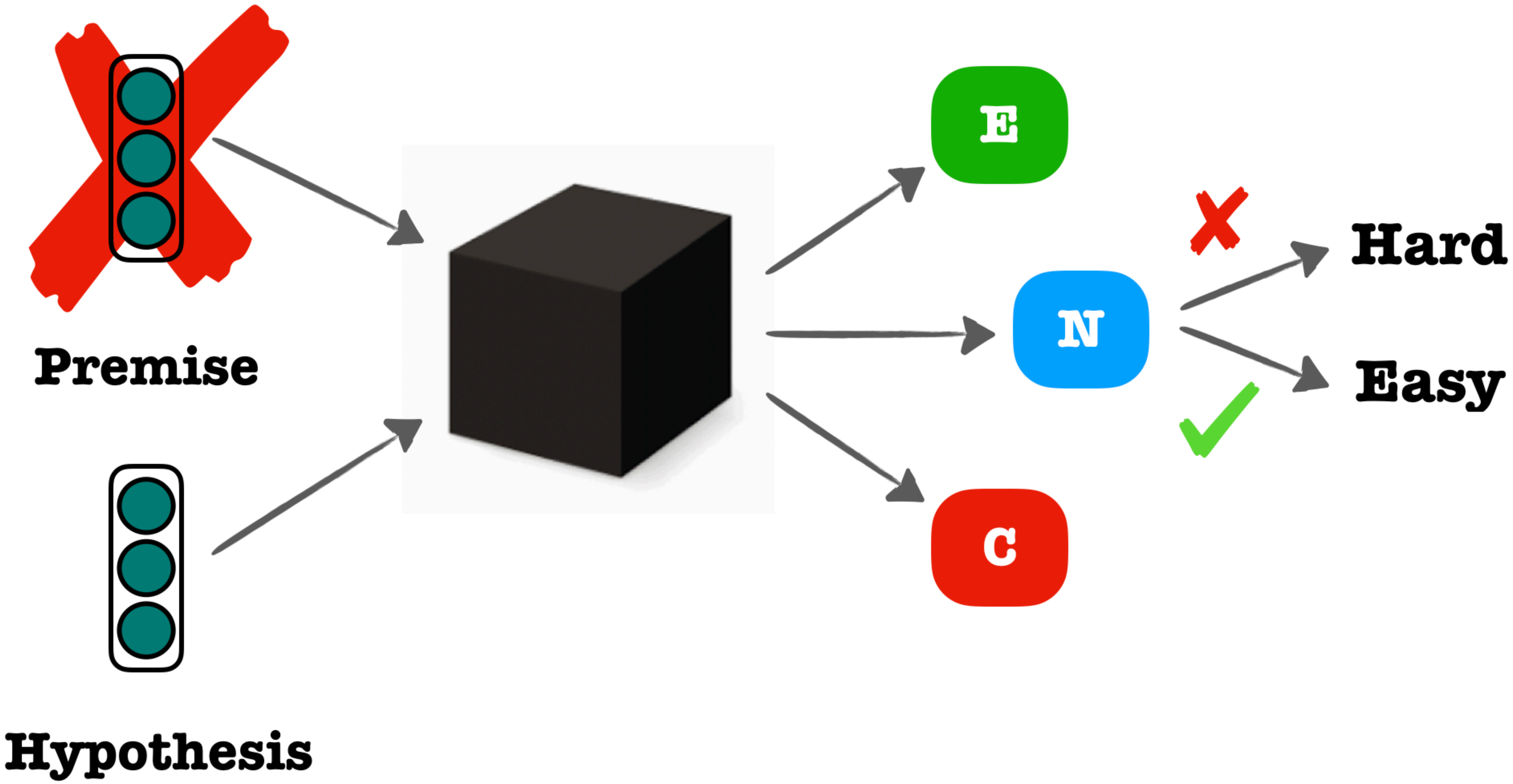


Looking Ahead: Improved Data Collection

- Partial input baselines. E.g.
  - SWAG [Zellers et. al., 2018],
  - DROP [Dua et. al., 2019],
  - Diverse NLI [Poliak et. al., 2018]
- Alternatives to human elicitation for building datasets?



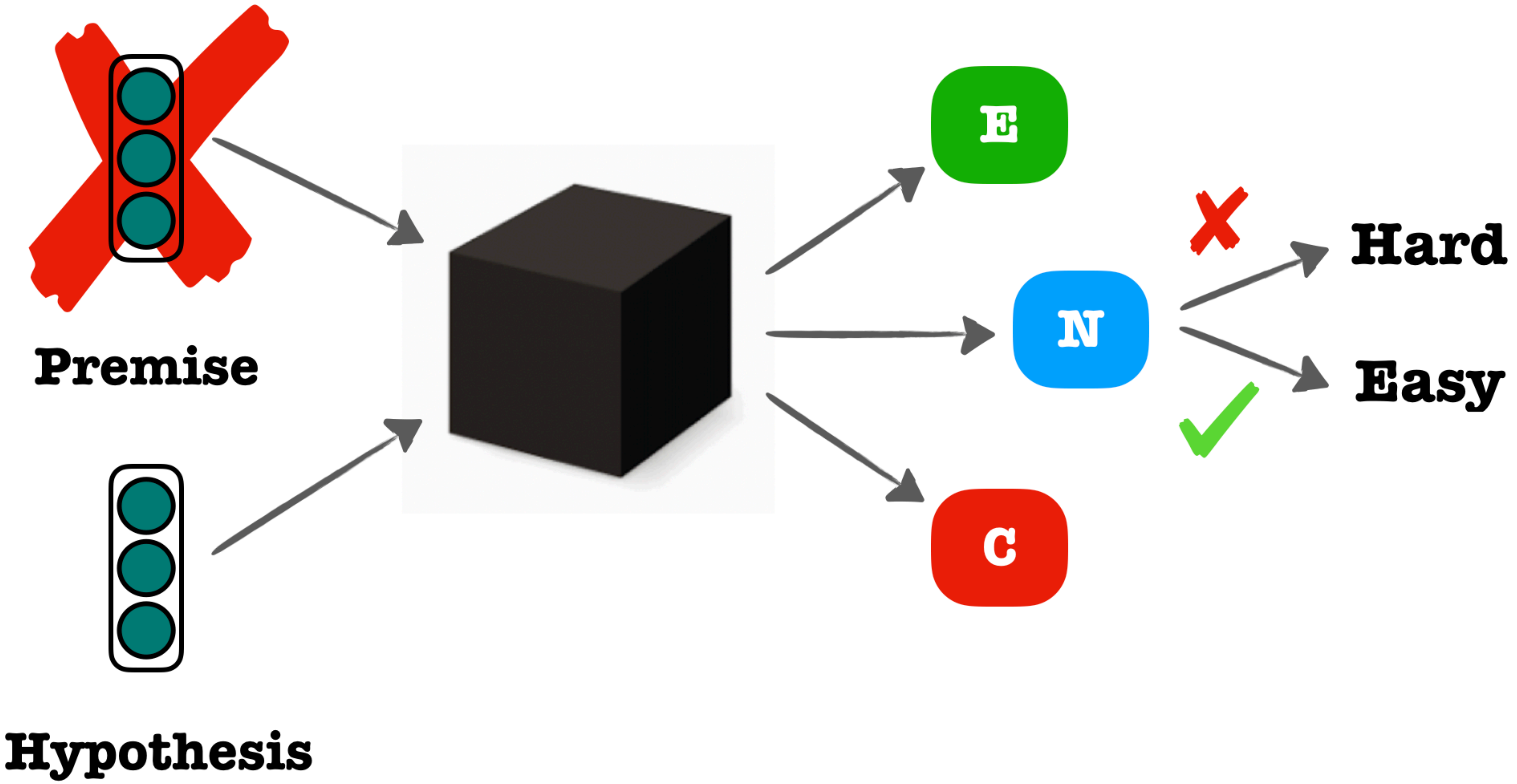
Filtering hypothesis-only artifacts



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Filtering hypothesis-only artifacts

- Other kinds of artifacts. For e.g. shortening the premise.

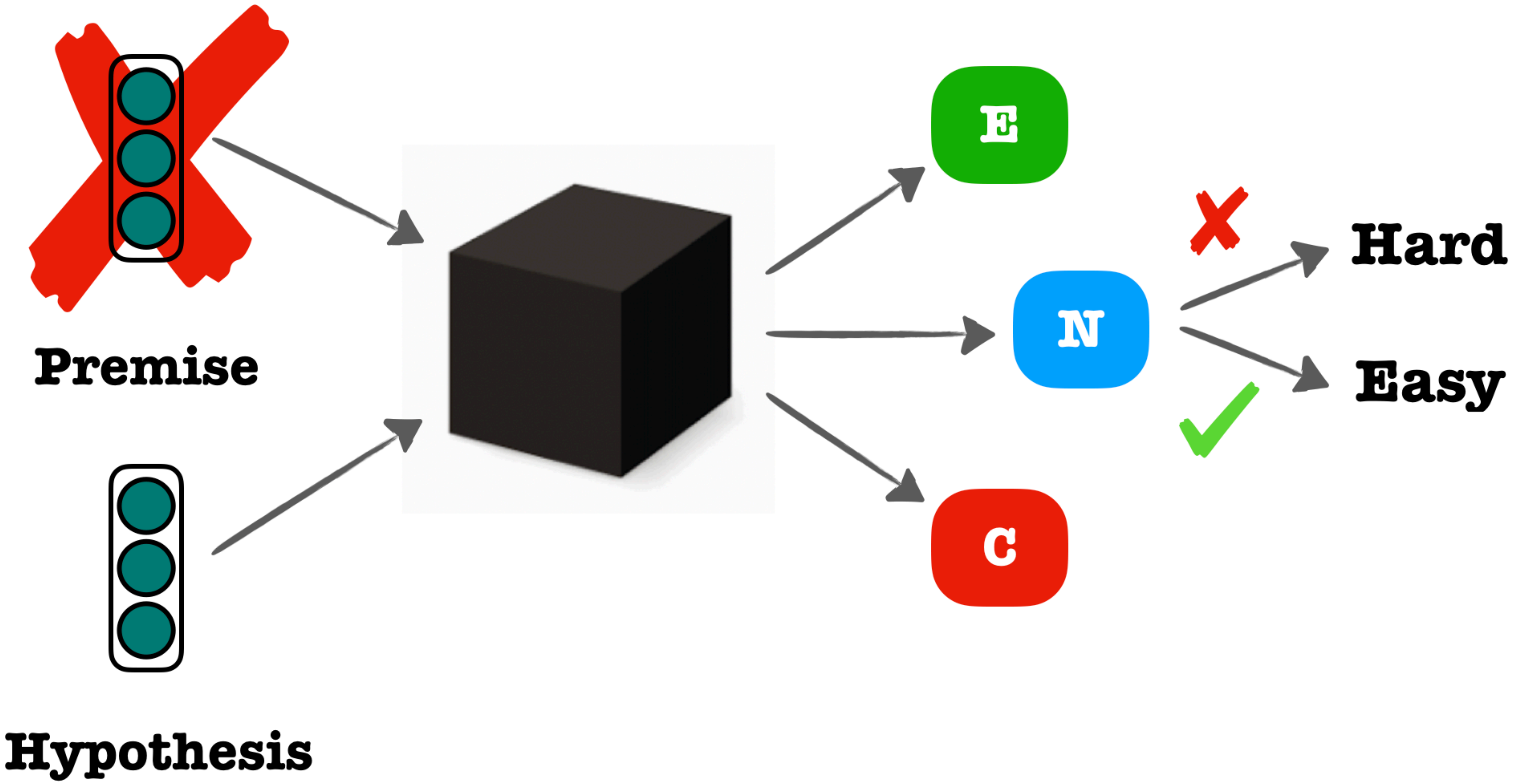


Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

Filtering hypothesis-only artifacts

- Other kinds of artifacts. For e.g. shortening the premise.

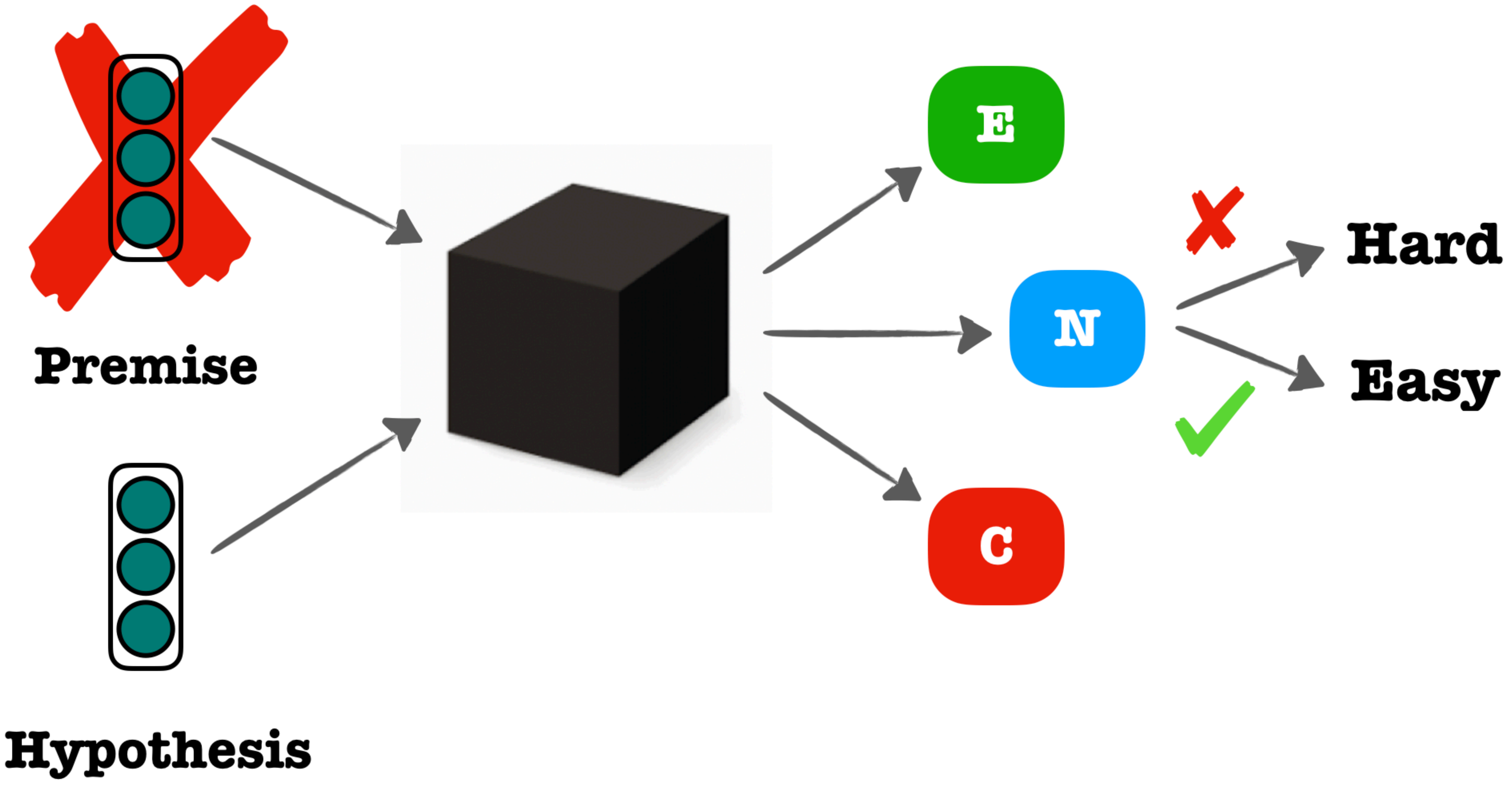
- Examples with artifacts are still valid examples…



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

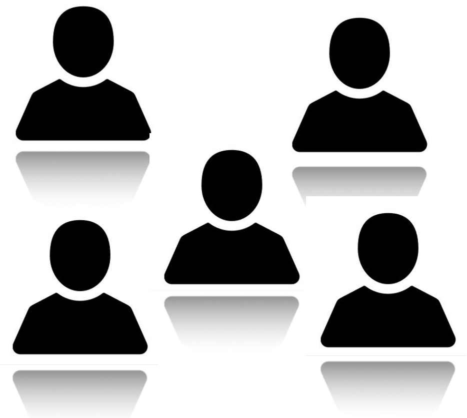
Filtering hypothesis-only artifacts

- Other kinds of artifacts. For e.g. shortening the premise.
- Examples with artifacts are still valid examples…
- Hard examples exhibit their own artifacts!



Annotation Artifacts in NLI Data [G*., Swayamdipta*, L., S., B., & S., NAACL 2018]

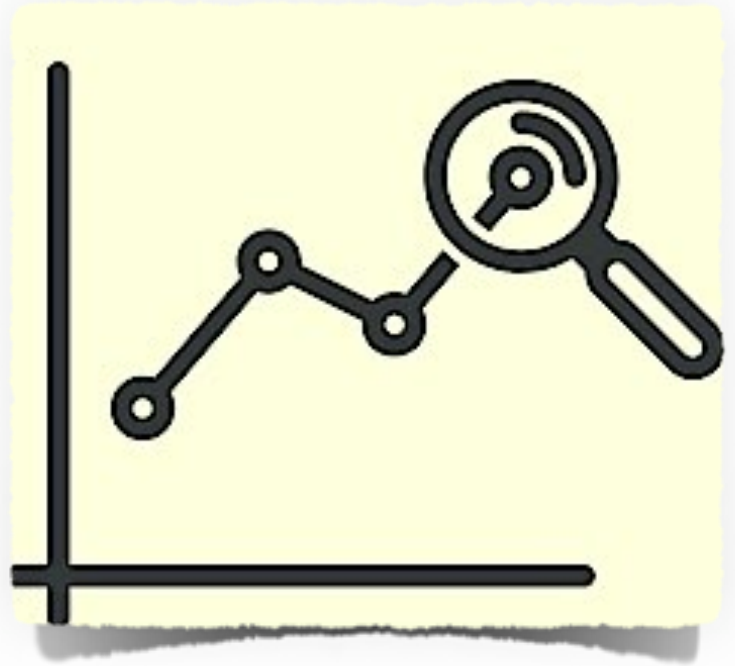
What do predictions tell us about the data?



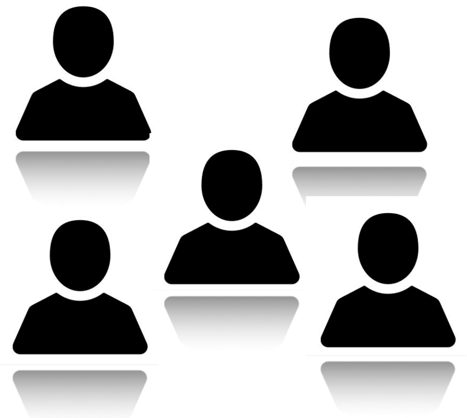
Annotation Artifacts Abound!

Data

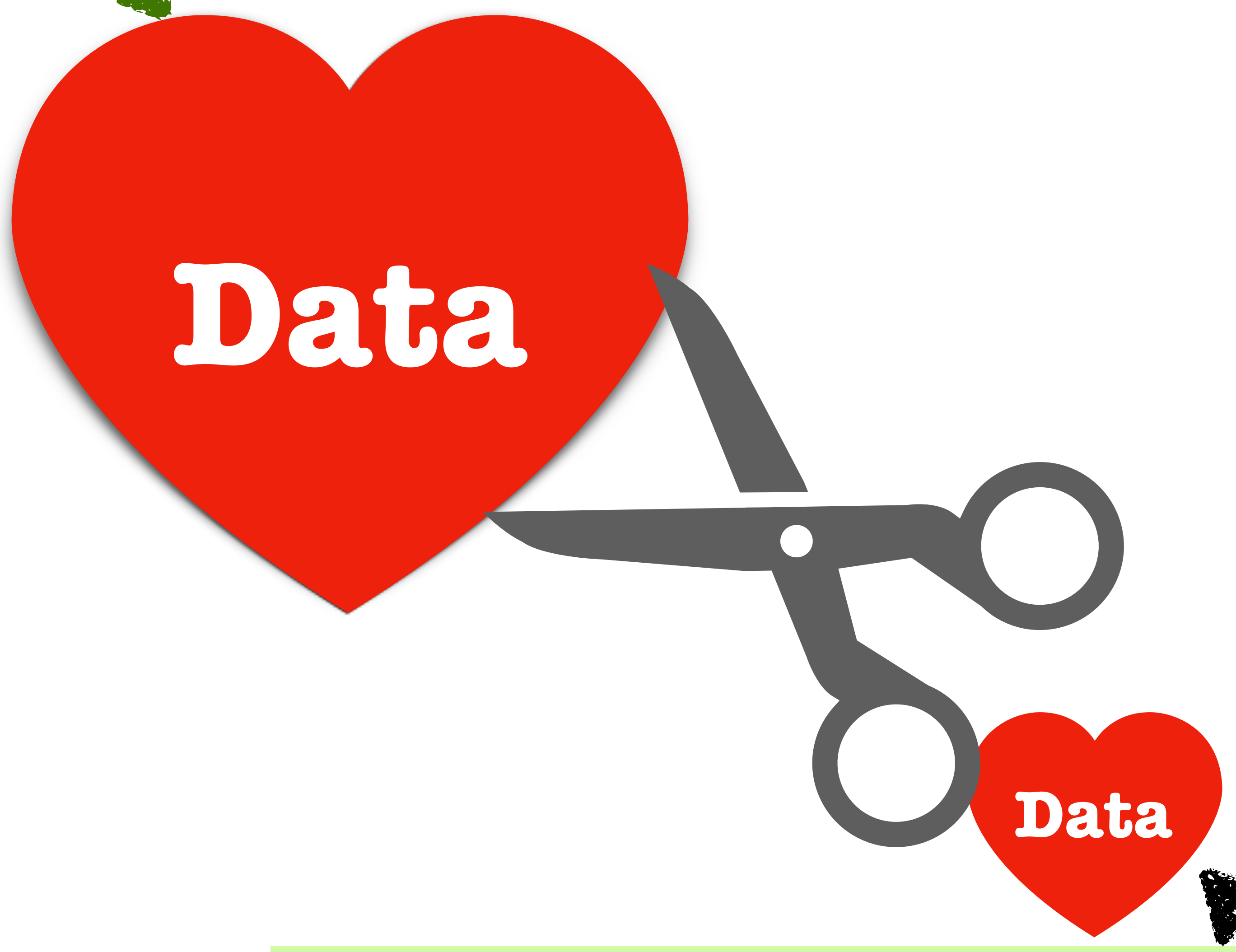
Question #2



What do predictions tell us about the data?

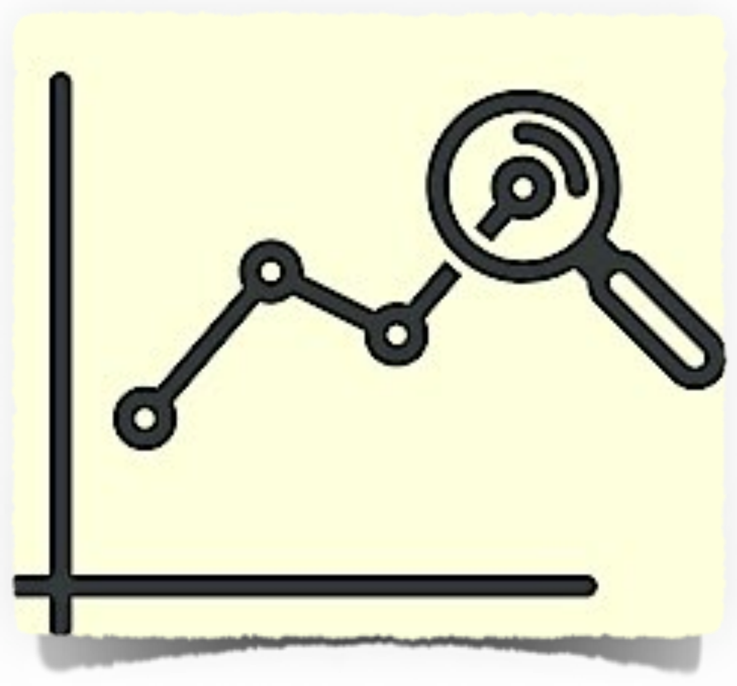


Annotation Artifacts Abound!

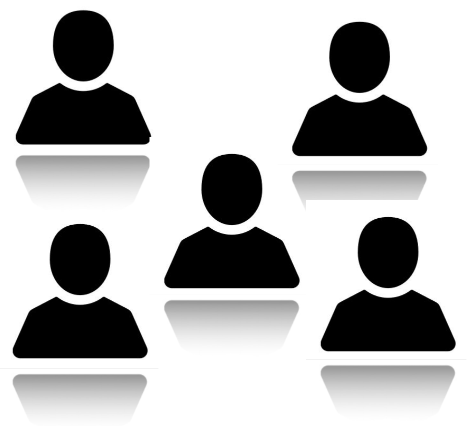


How can we use this information to spruce up our datasets?

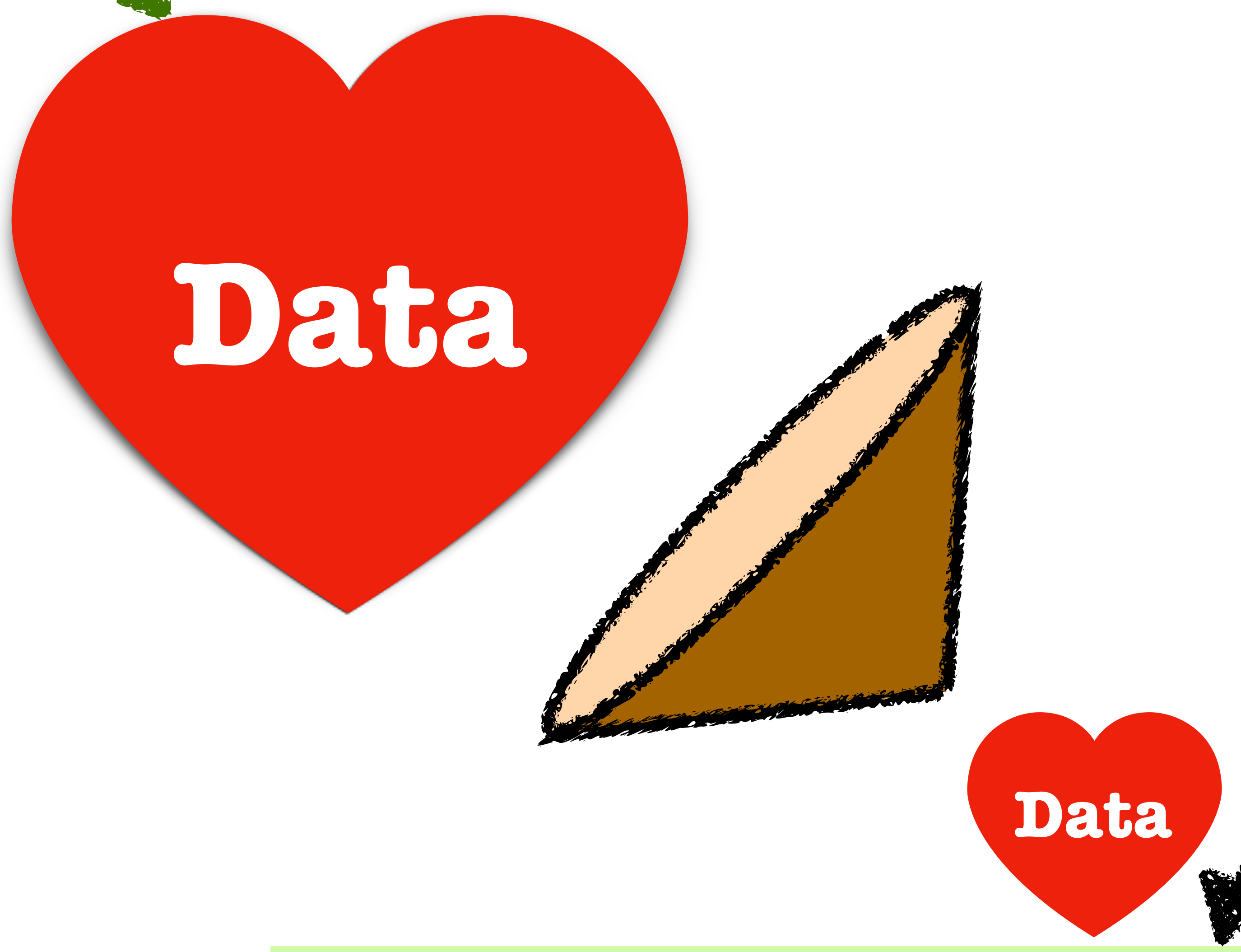
Question #2



What do predictions tell us about the data?



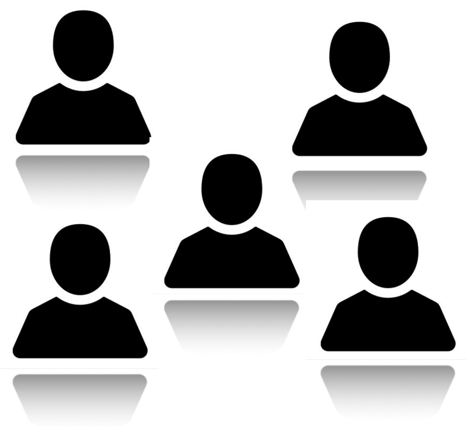
Annotation Artifacts Abound!



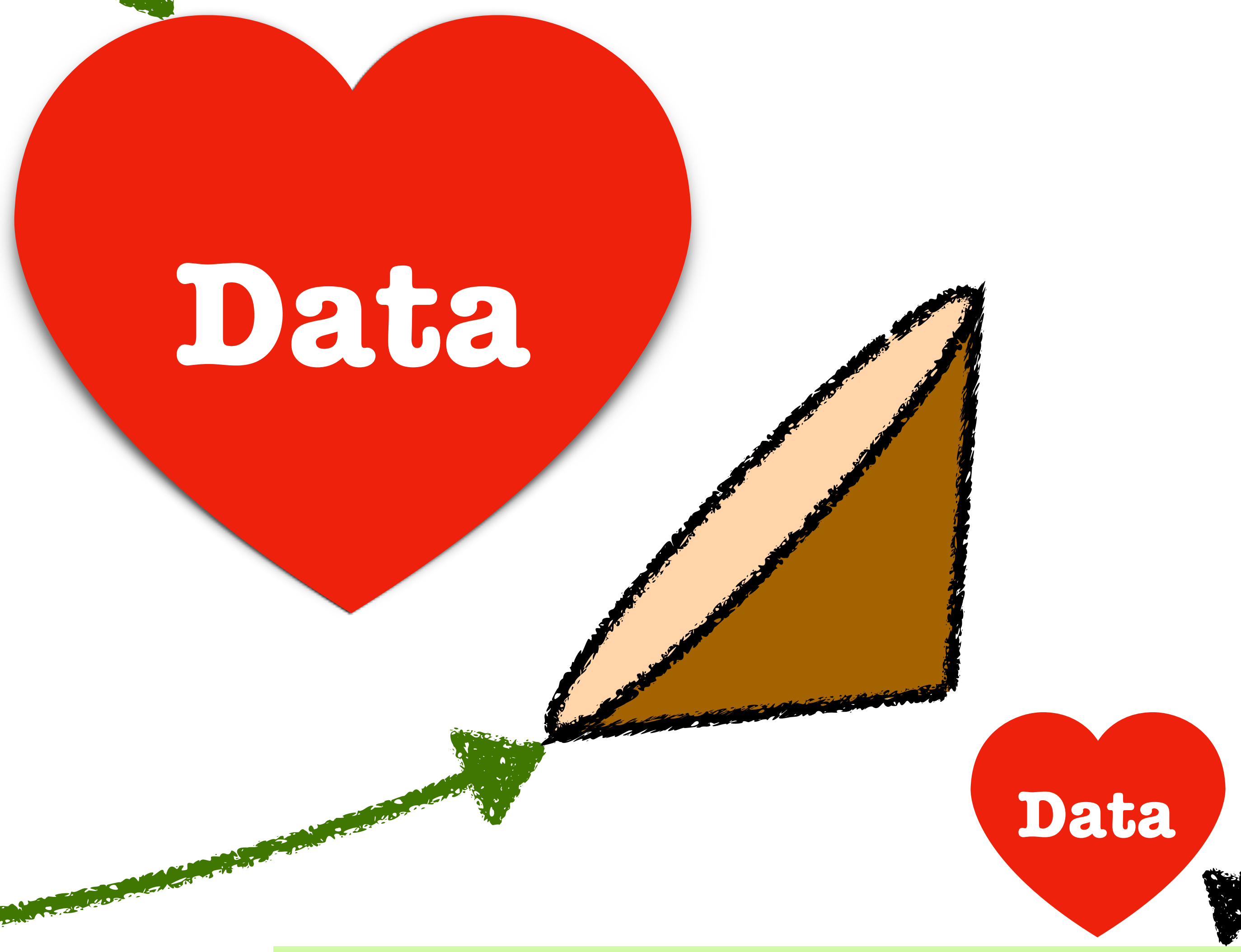
How can we use this information to spruce up our datasets?

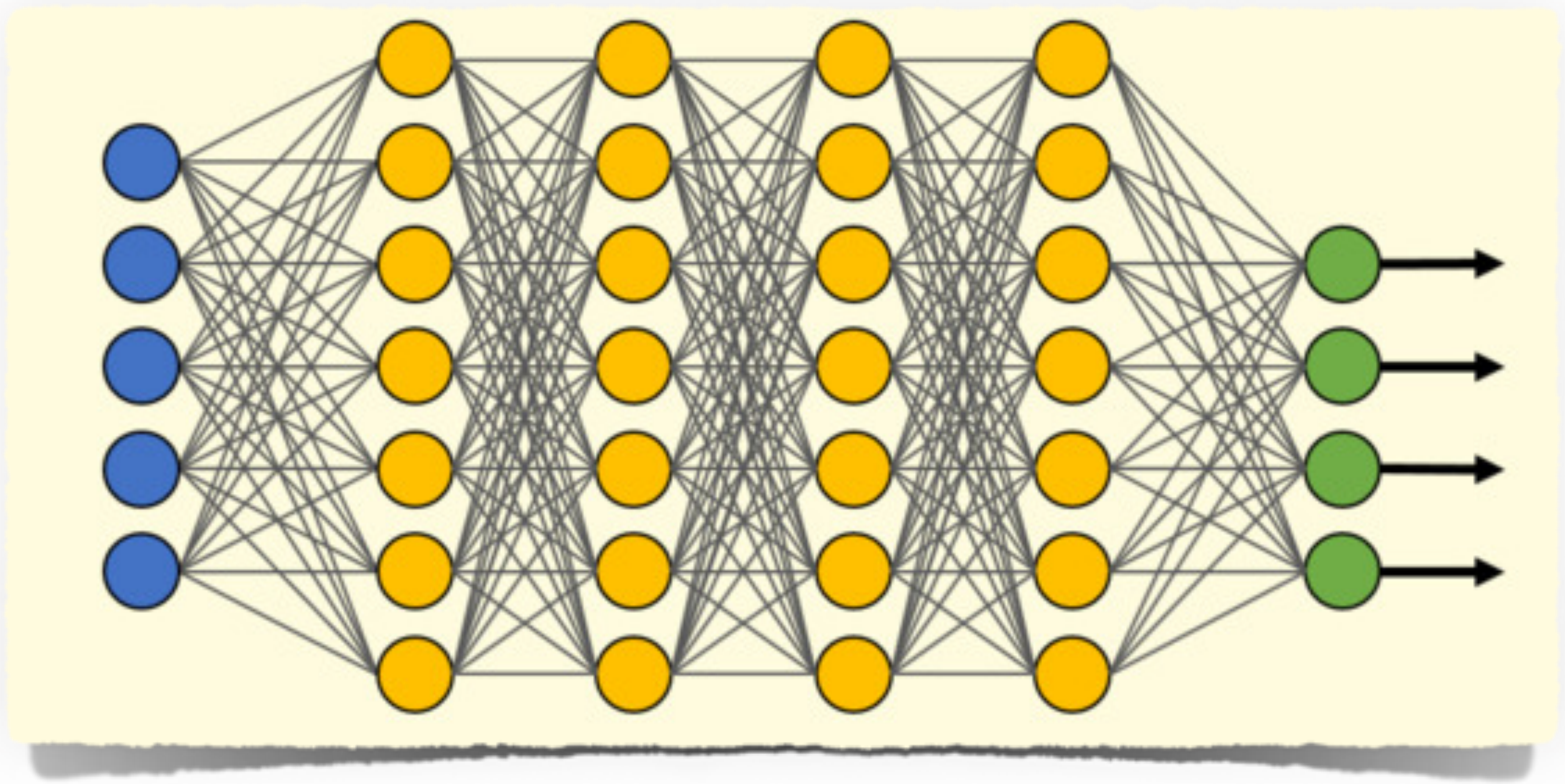
Question #2

What do predictions tell us about the data?



Annotation Artifacts Abound!





How can we use this information to spruce up our datasets?

Goal

Goal

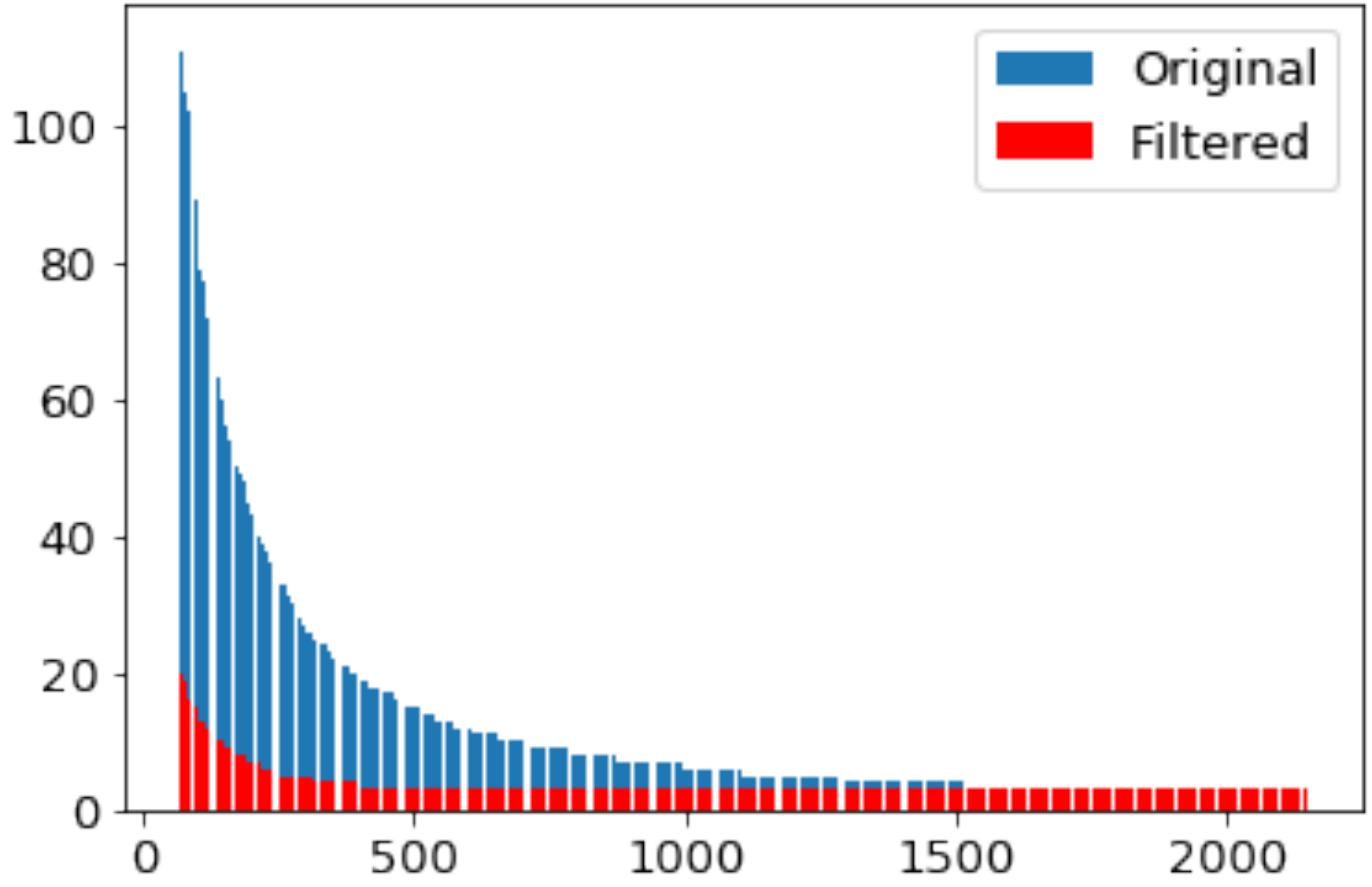
- **Balance out** the occurrence of different phenomena in the dataset
    - Filter out a majority of the samples which exhibit artifacts.

Goal

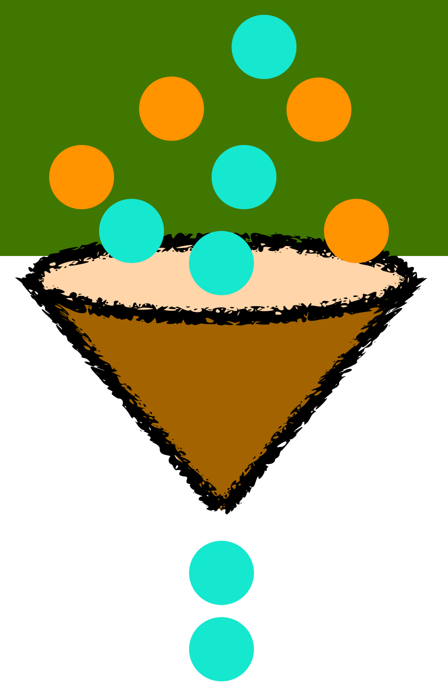
- **Balance out** the occurrence of different phenomena in the dataset
    - Filter out a majority of the samples which exhibit artifacts.

Goal

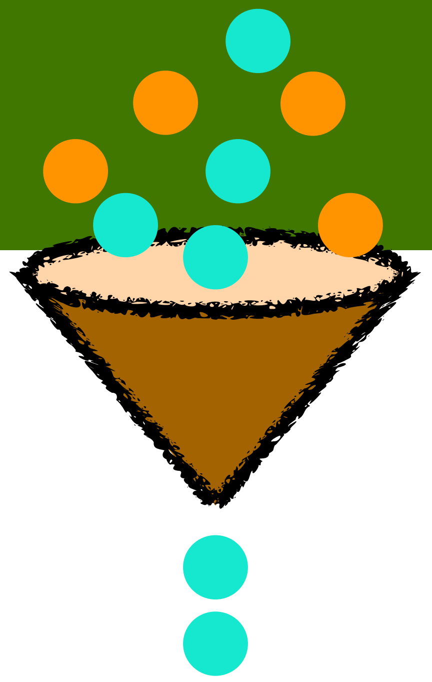
- **Balance out** the occurrence of different phenomena in the dataset
  - Filter out a majority of the samples which exhibit artifacts.
- Avoid **head phenomena redundancy**



Insights

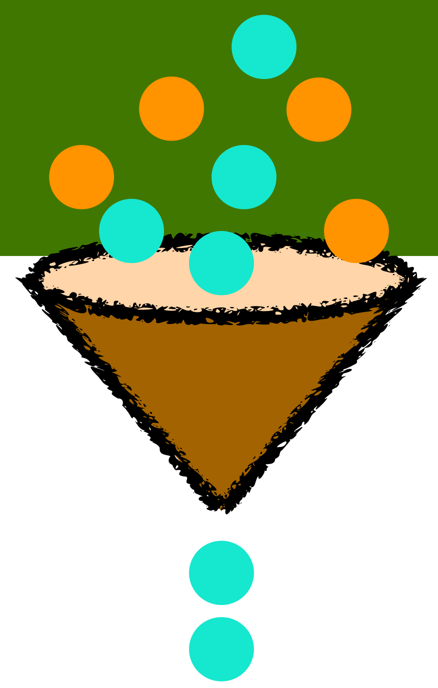


Insights



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

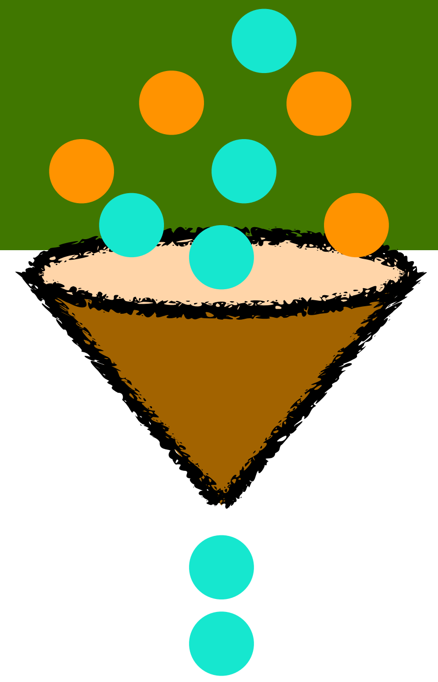
Insights



i. The models that exploit artifacts, can be used to **detect artifacts**! Better than manual identification... (e.g. hypothesis-only artifacts)

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Insights



i. The models that exploit artifacts, can be used to **detect artifacts**! Better than manual identification… (e.g. hypothesis-only artifacts)

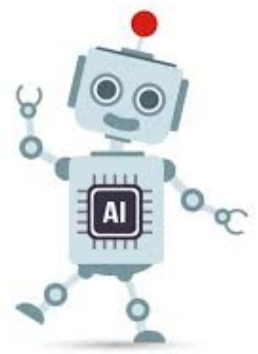
ii. Examples with artifacts can be classified correctly by **multiple** models.

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

How predictable is a sample?

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

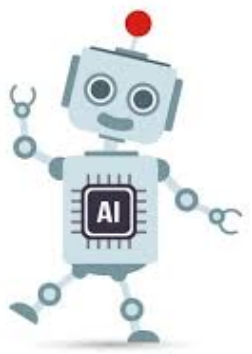
How predictable is a sample?

- Can it be predicted by a simple model? 

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

How predictable is a sample?

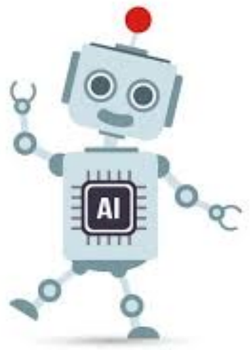
- Can it be predicted by a simple model?



- How much training does it take to predict it?

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

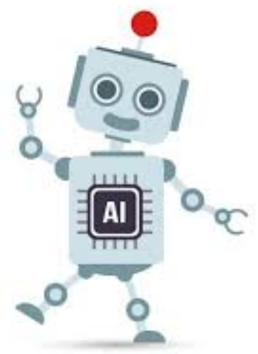
How predictable is a sample?

- Can it be predicted by a simple model? 
- How much training does it take to predict it? 
- How confident is the model? 

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

How predictable is a sample?

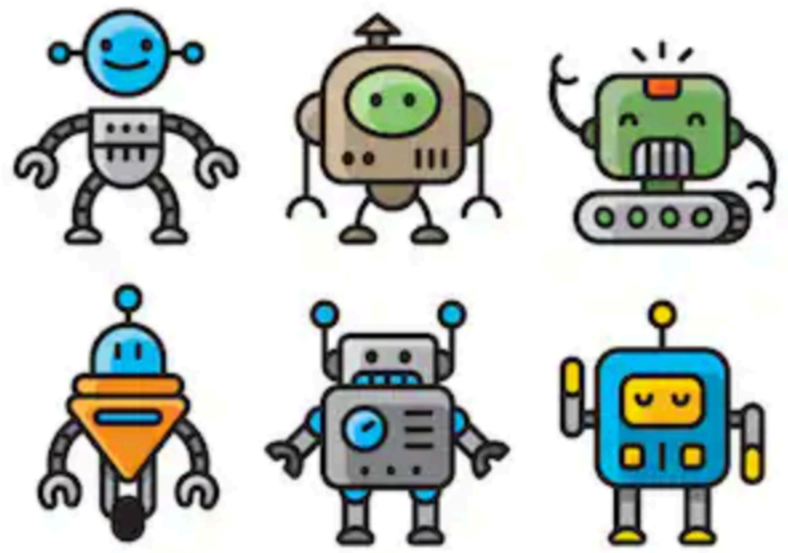
- Can it be predicted by a simple model?



- How much training does it take to predict it?

- How confident is the model?

- Can it be predicted by **several** simple models?

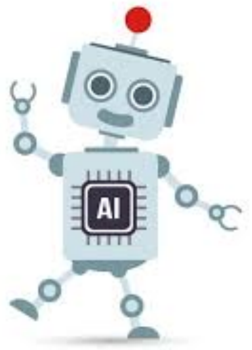


Ensembles of
Linear Classifiers

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

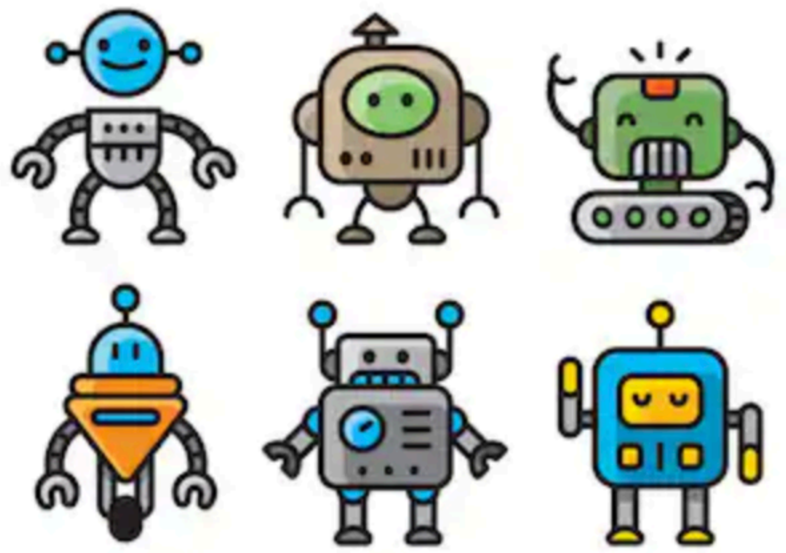
How predictable is a sample?

- Can it be predicted by a simple model?



- How much training does it take to predict it?

- How confident is the model?
- Can it be predicted by **several** simple models?



Predictability Score

Ensembles of
Linear Classifiers

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Predictability Scores

Ribiero et al., 2017

Predictability Scores

Ribiero et al., 2017

Iteration 1	0.99	0.93	0.97	0.87	0.89	0.78

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Predictability Scores

Ribiero et al., 2017

Iteration 1	0.99	0.93	0.97	0.87	0.89	0.78
Iteration 2						

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Predictability Scores

Ribiero et al., 2017

Iteration 1	0.99	0.93	0.97	0.87	0.89	0.78
Iteration 2				0.83	0.99	0.54

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Predictability Scores

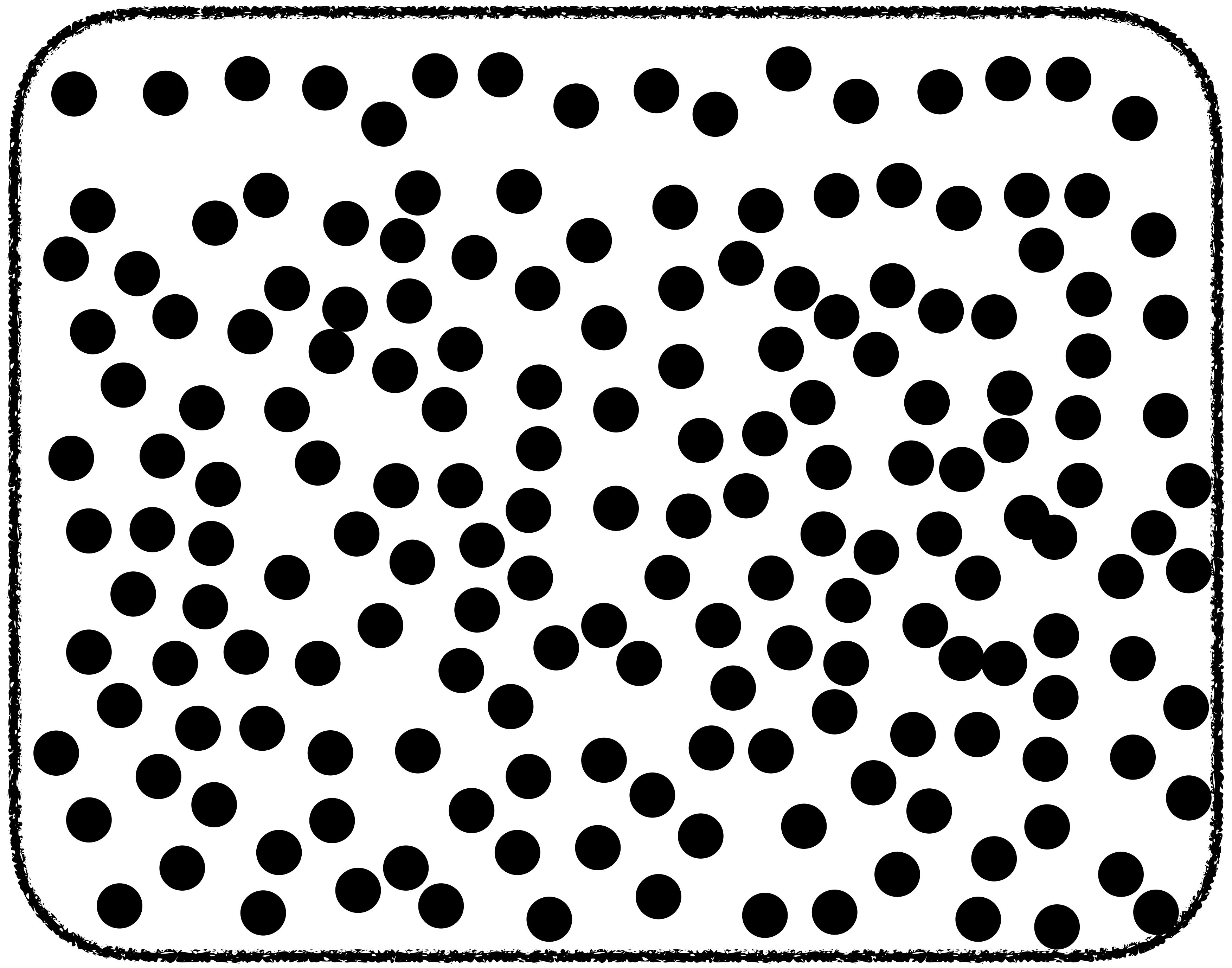
Ribiero et al., 2017

Iteration 1	0.99	0.93	0.97	0.87	0.89	0.78
Iteration 2				0.83	0.99	0.54

Iterative and Greedy!

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

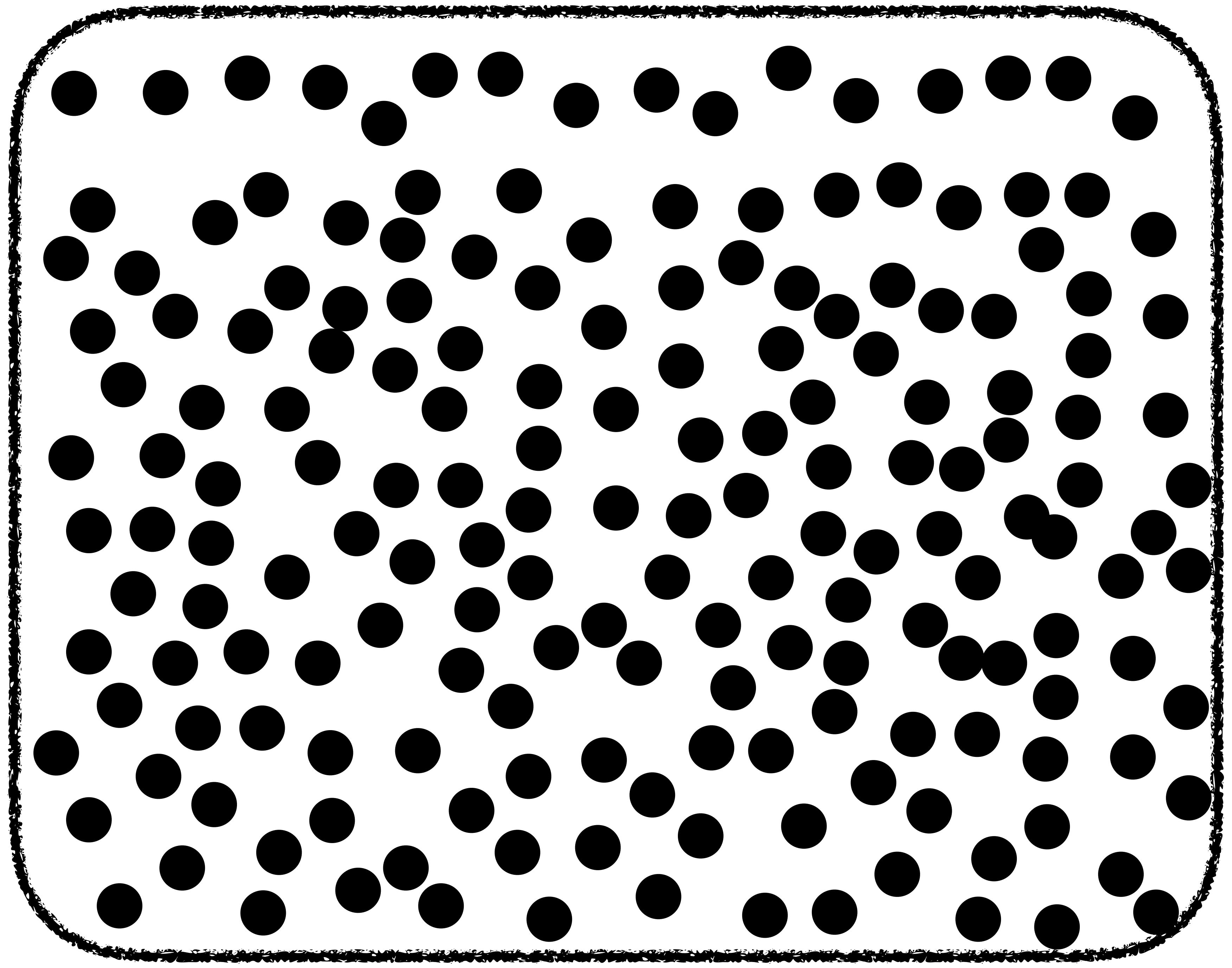
Algorithm



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

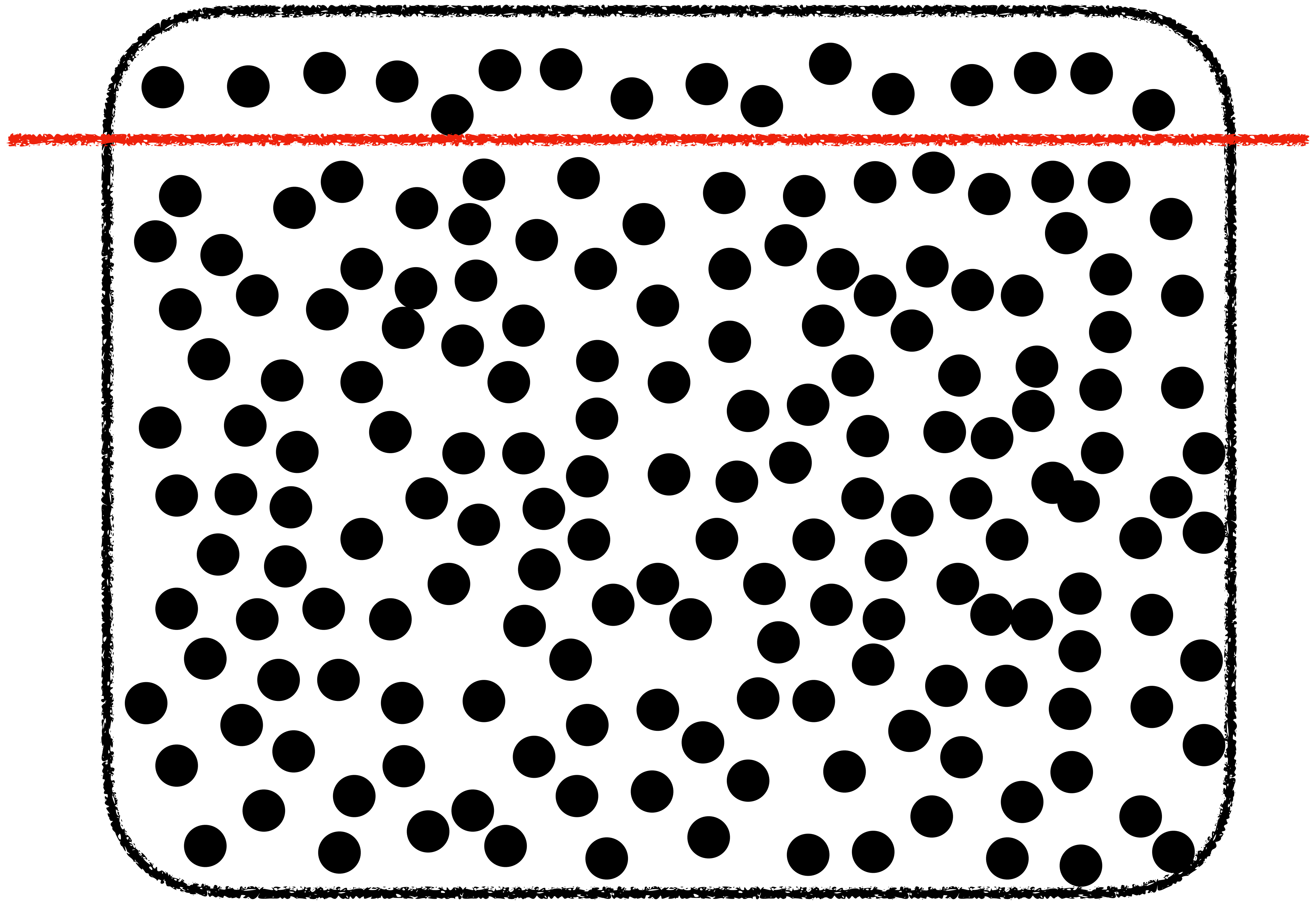
- Start with an initial feature representation, ϕ



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

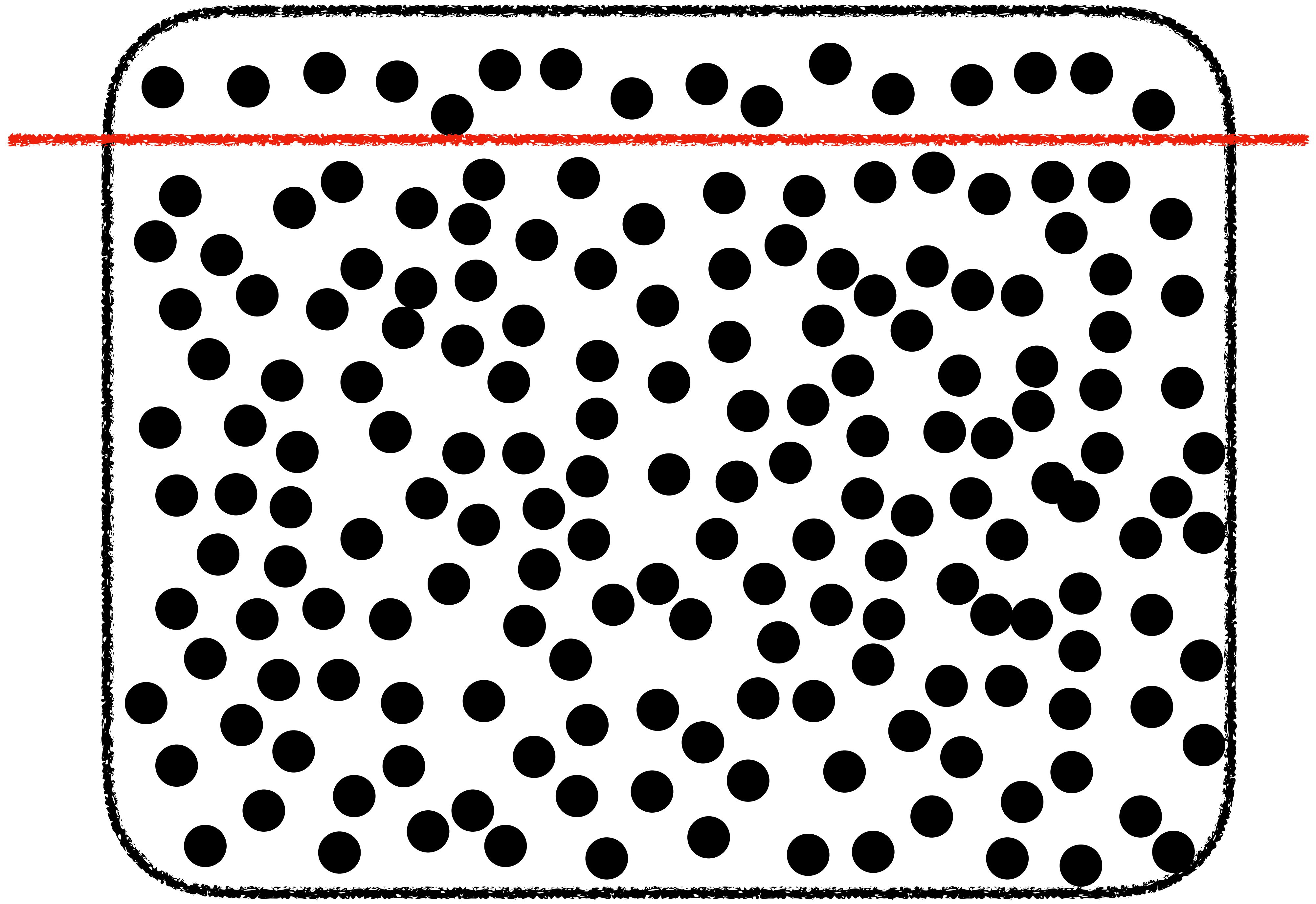
- Start with an initial feature representation, ϕ



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

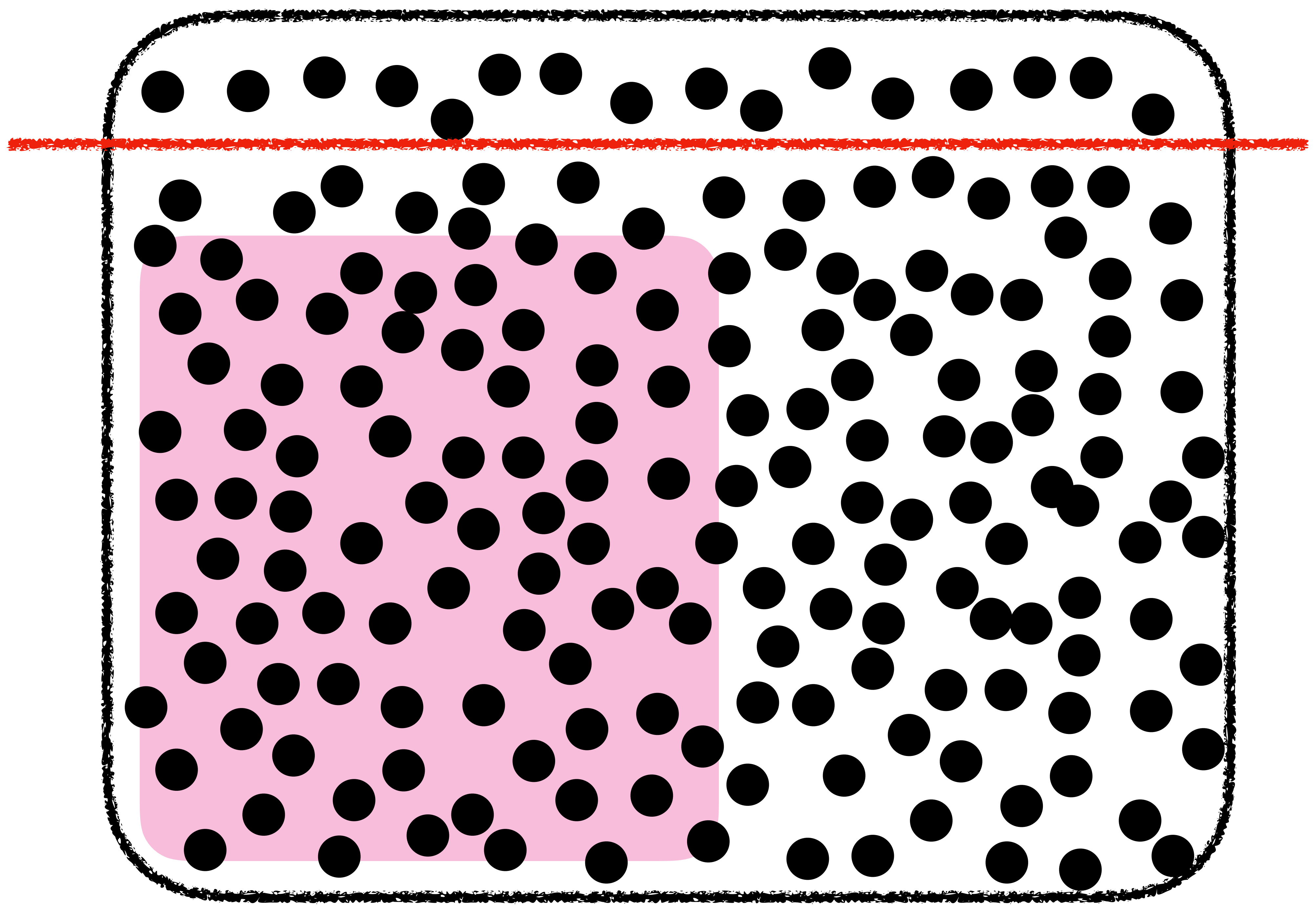
- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

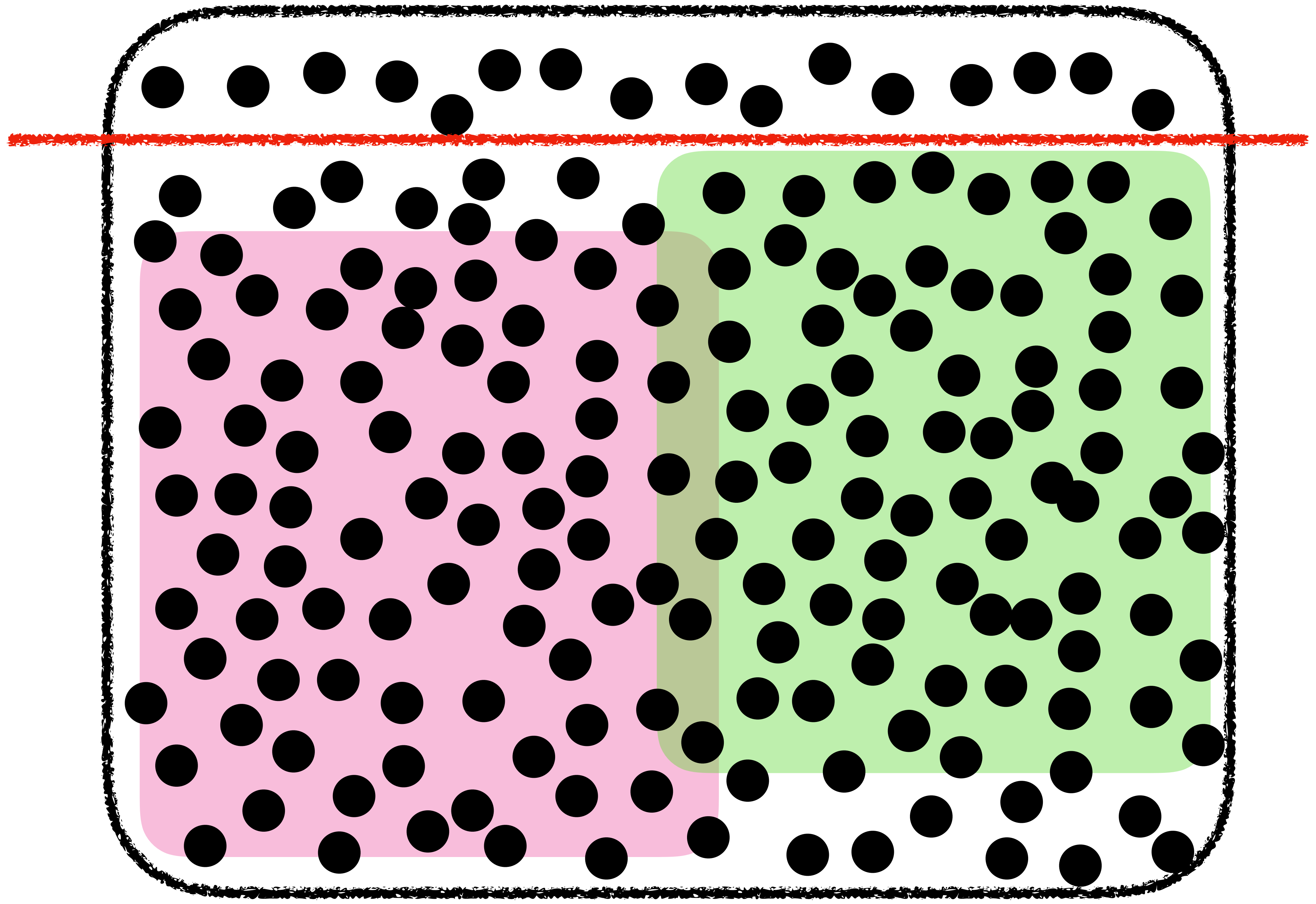
- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

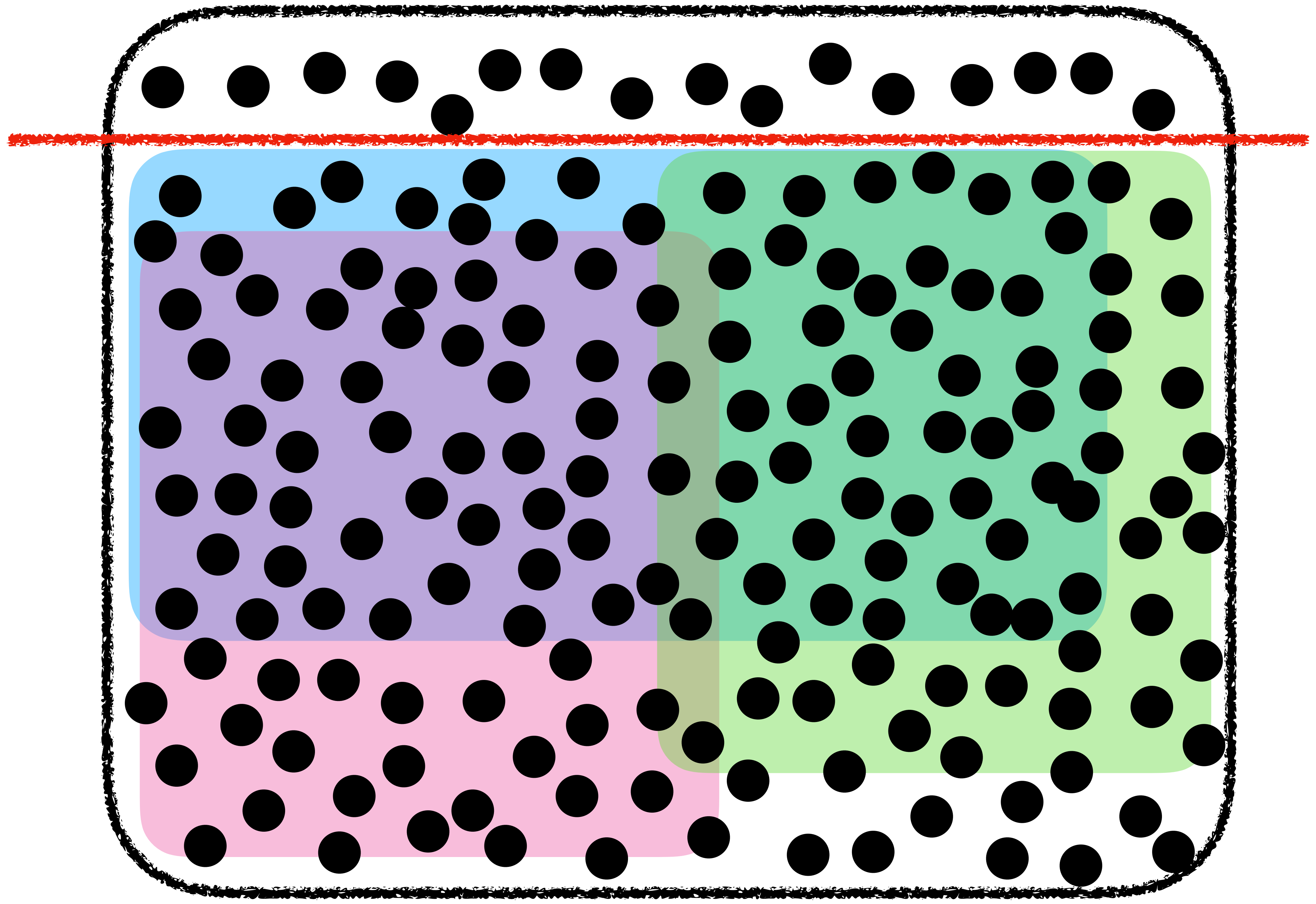
- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

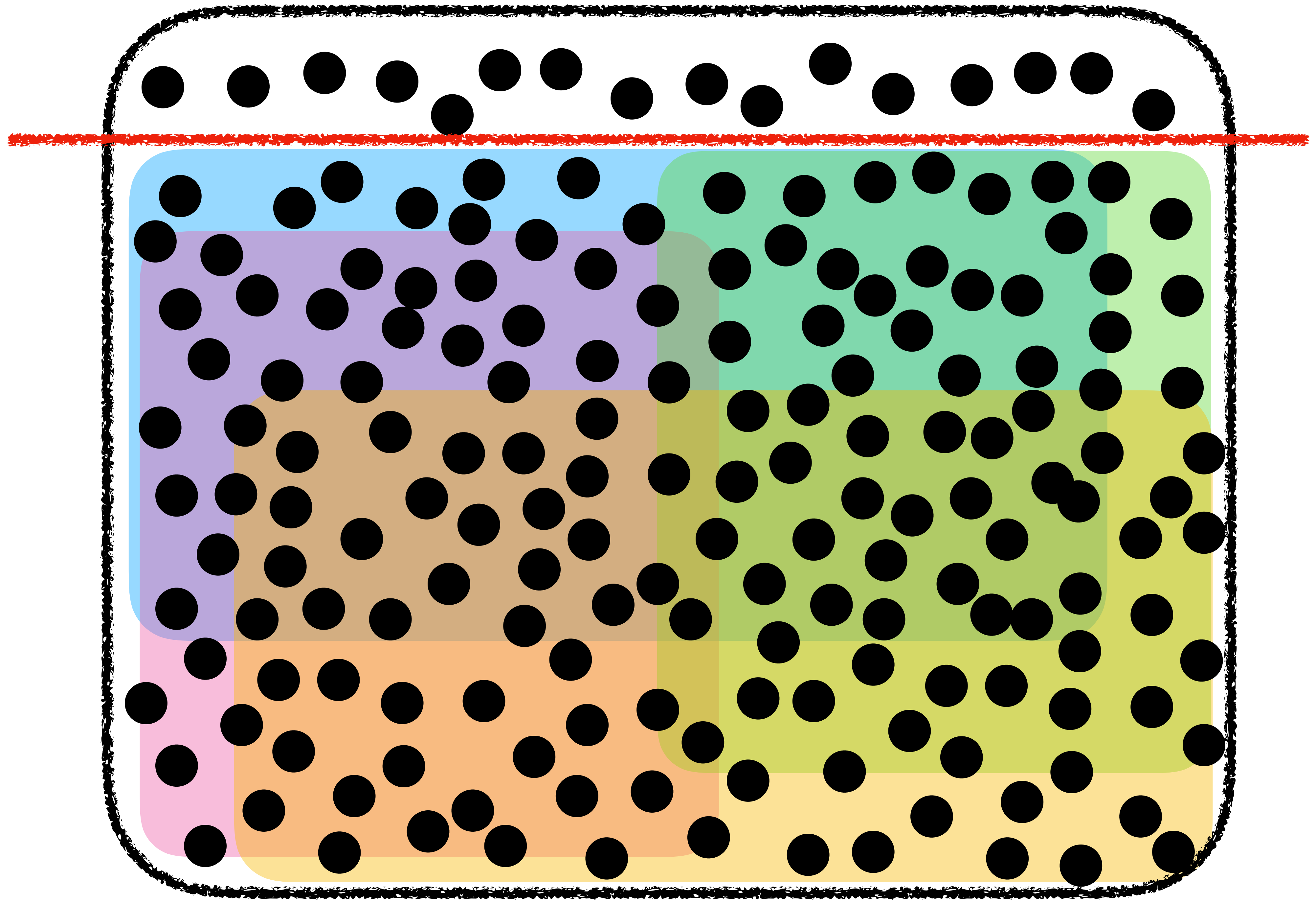
- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.

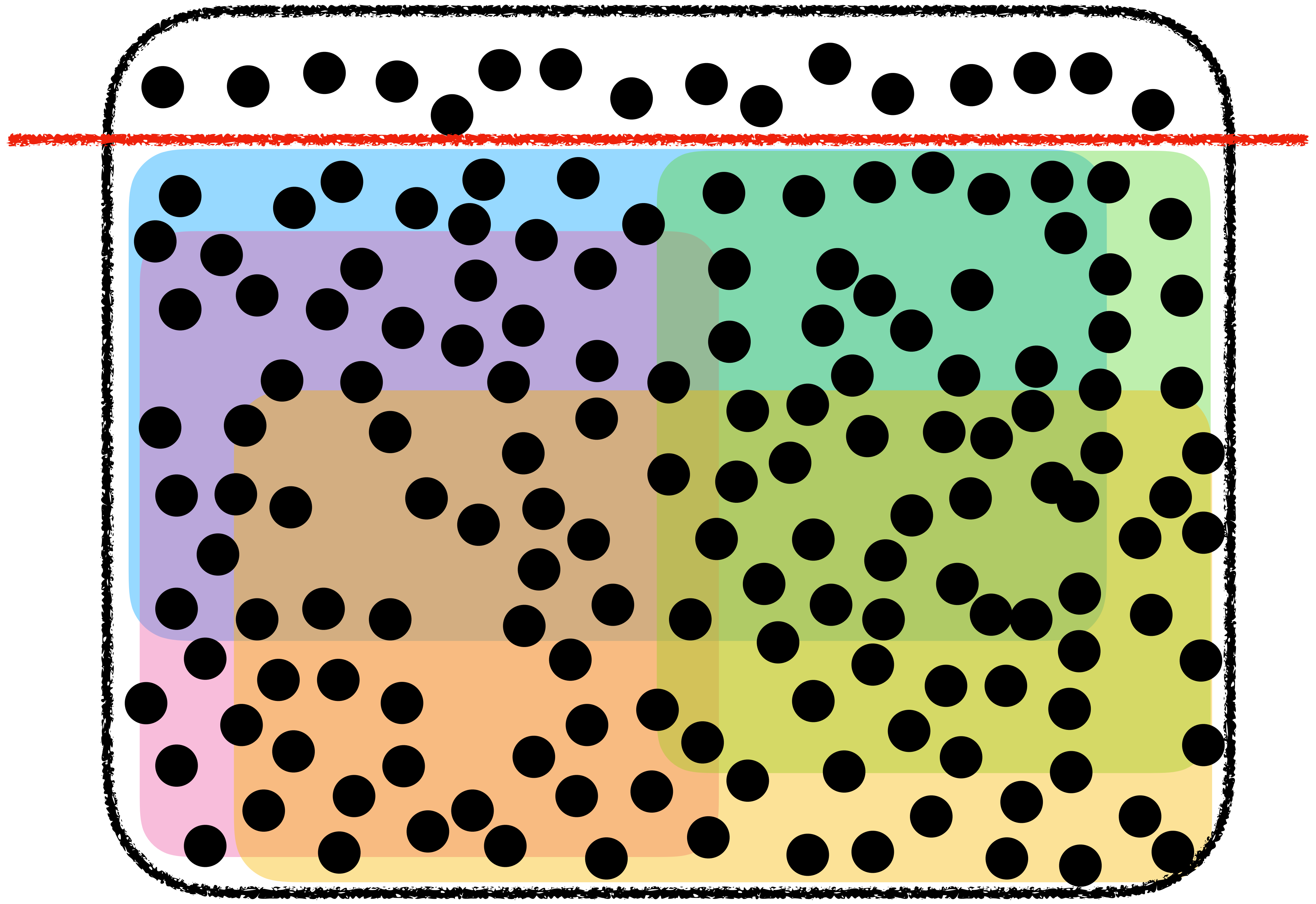


Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.

- Discard the top-k examples which are correctly identified by most models, iteratively, till the ensemble is no longer confident.



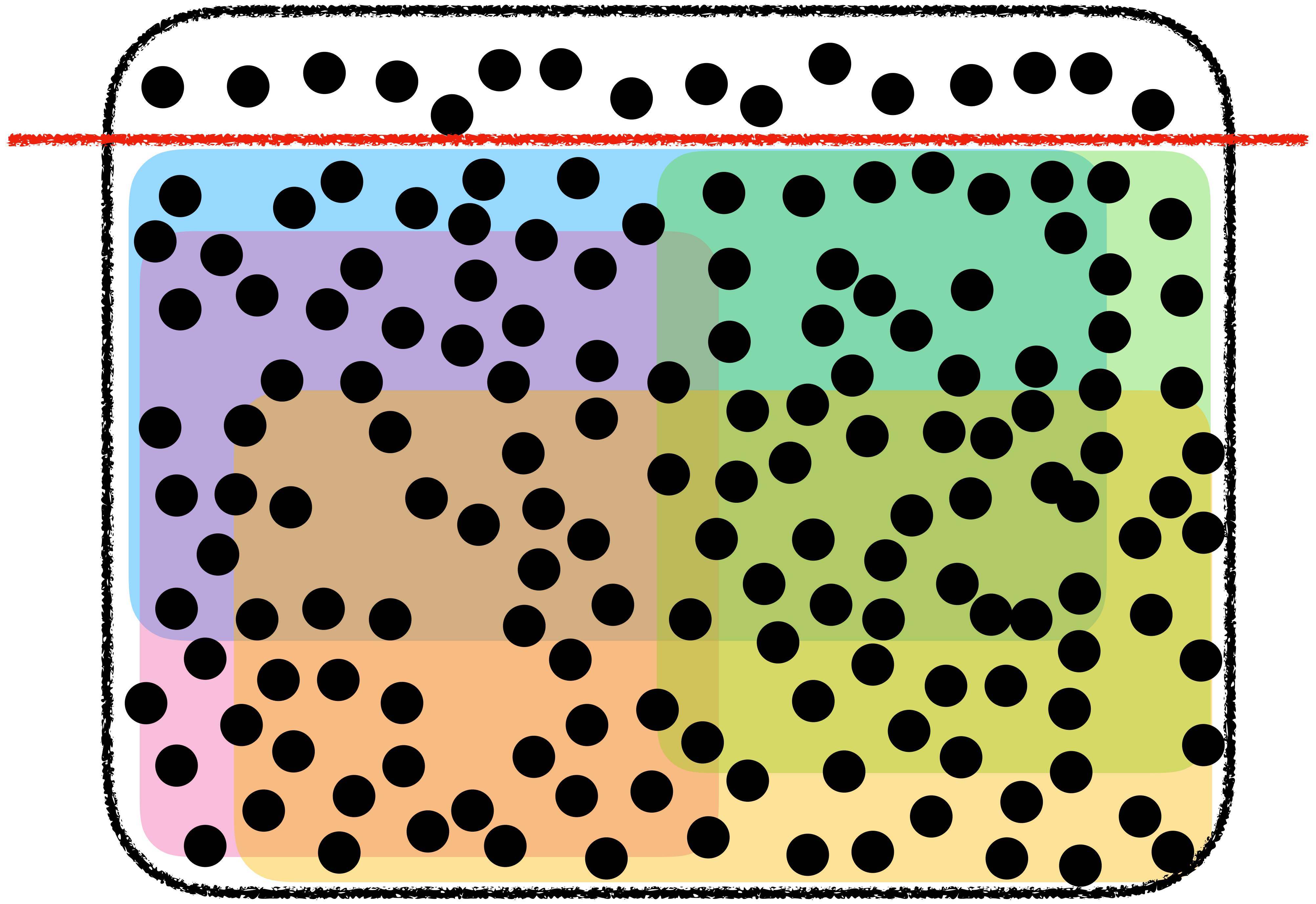
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Algorithm

Precursor: Zellers et al., 2018; 2019

- Start with an initial feature representation, ϕ
- Train multiple models on random partitions of the remaining data.

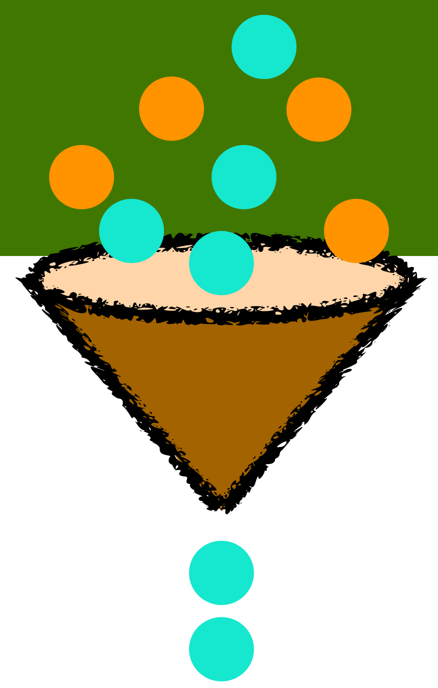
- Discard the top-k examples which are correctly identified by most models, iteratively, till the ensemble is no longer confident.



- Lightweight Adversarial Filtering (**AFLite**) :
  - linear models
  - fixed feature representation, ϕ.

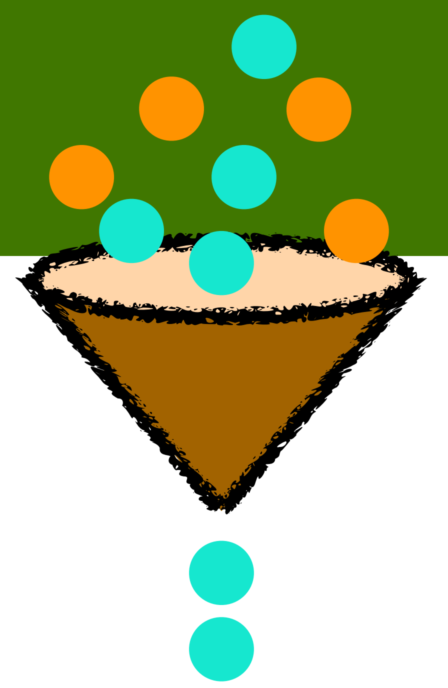
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Follow-up questions



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

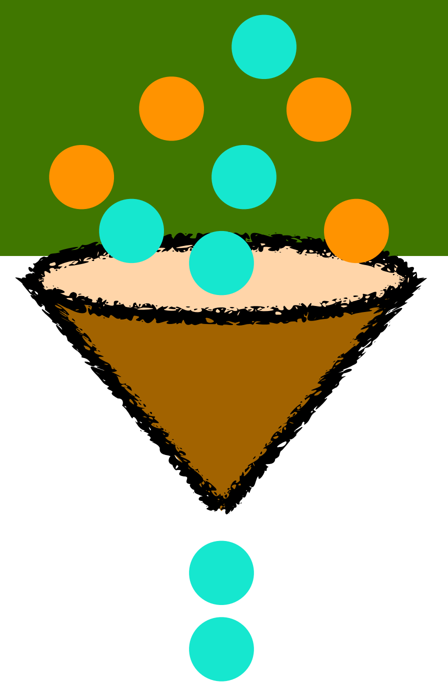
Follow-up questions



- Is AFLite optimal?
  - No! It's a greedy procedure

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

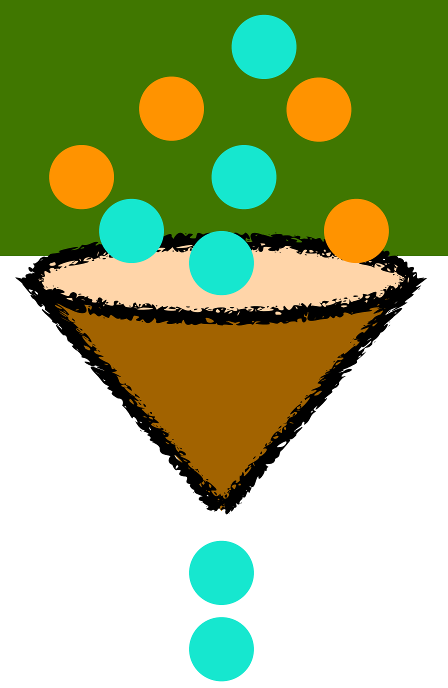
Follow-up questions



- Is AFLite optimal?
  - No! It's a greedy procedure
- Is there an optimal variant?
  - Yes! But intractable — subset selection problem.

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Follow-up questions



- Is AFLite optimal?
  - No! It's a greedy procedure
- Is there an optimal variant?
  - Yes! But intractable — subset selection problem.

- How would it work?
  - Find the smallest subset, any train-test split of which achieves high accuracy, but does not generalize to held-out data outside the subset.
  - Mini-max optimization problem

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

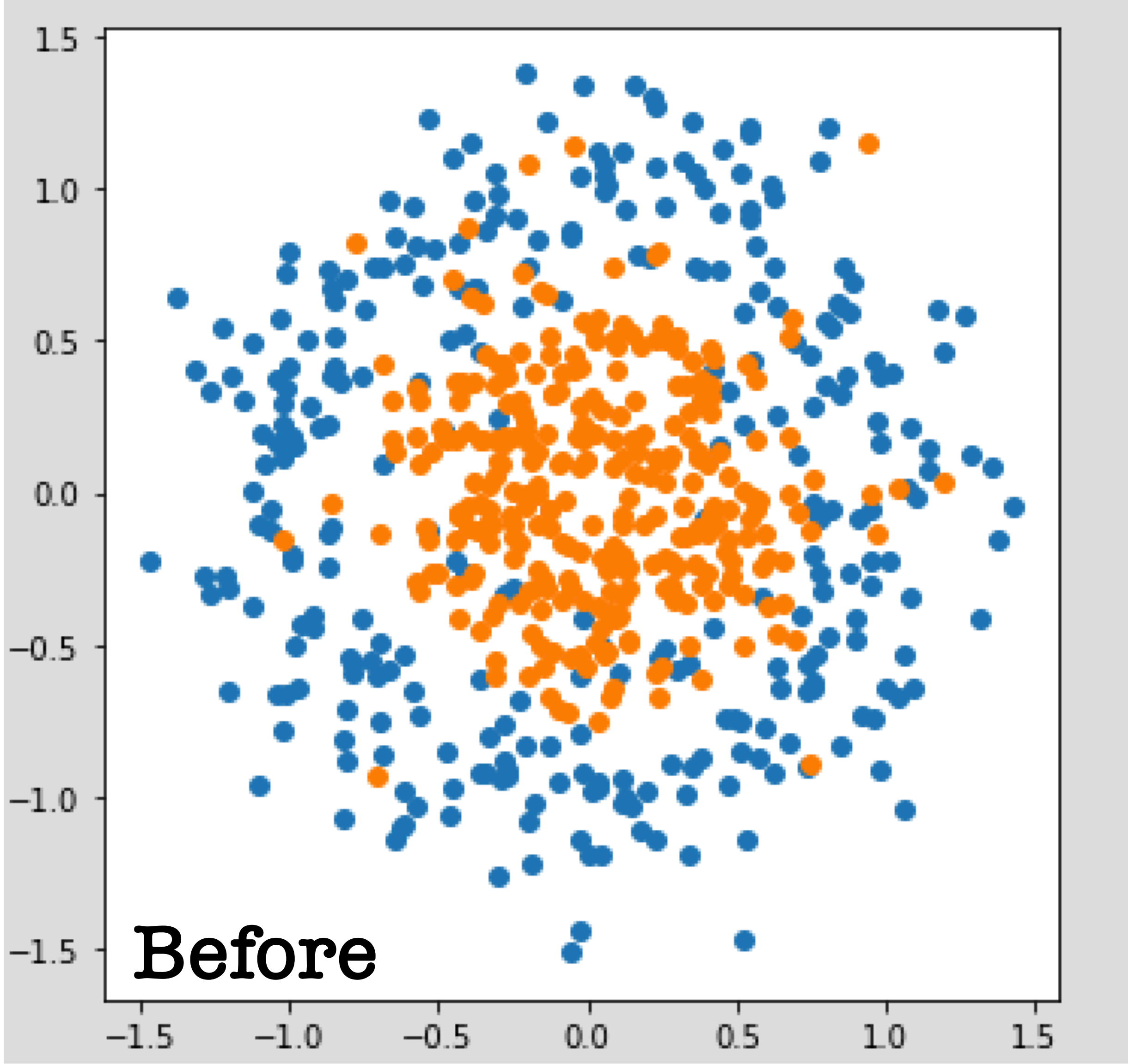
Evaluation Setting

- Extrinsic Evaluation:
  - Model performance on test before / after filtering.
  - Training data also changes to account for distributional differences
- Intrinsic Evaluation:
  - Filtered dataset properties

	Unfiltered Train	Filtered Train
Unfiltered Test	✔	
Filtered Test		✔

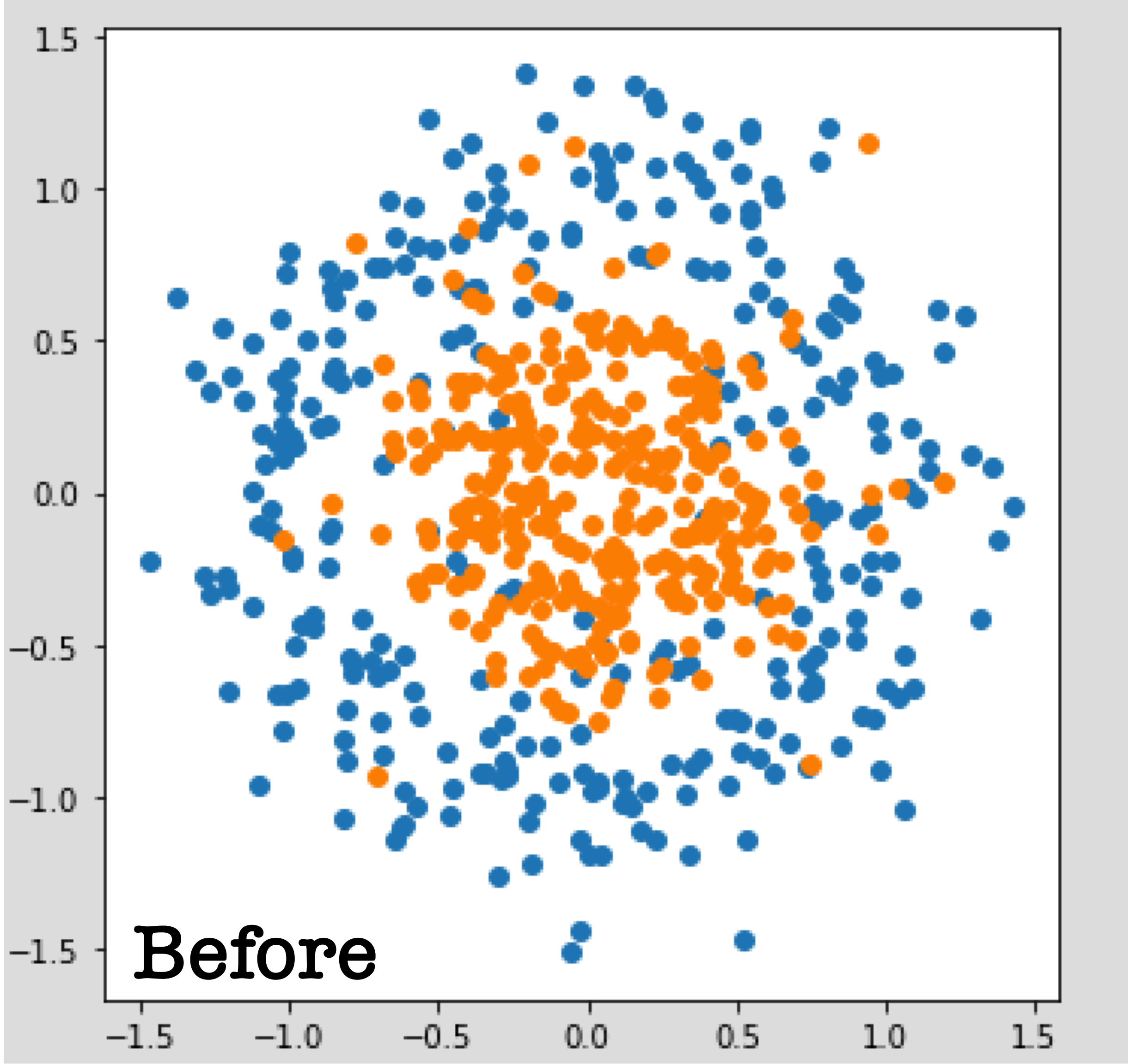
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

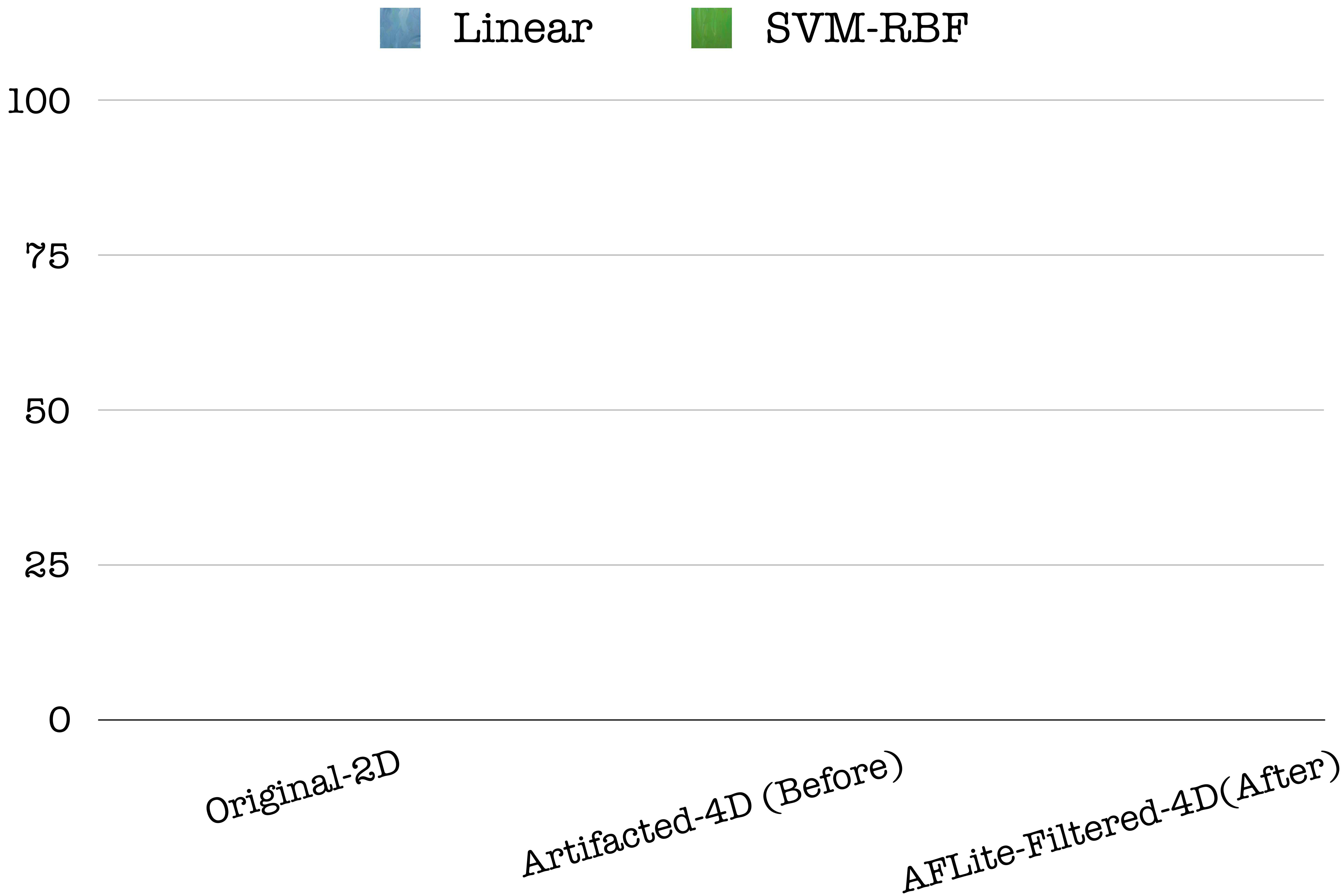
Task 0: Synthetic



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

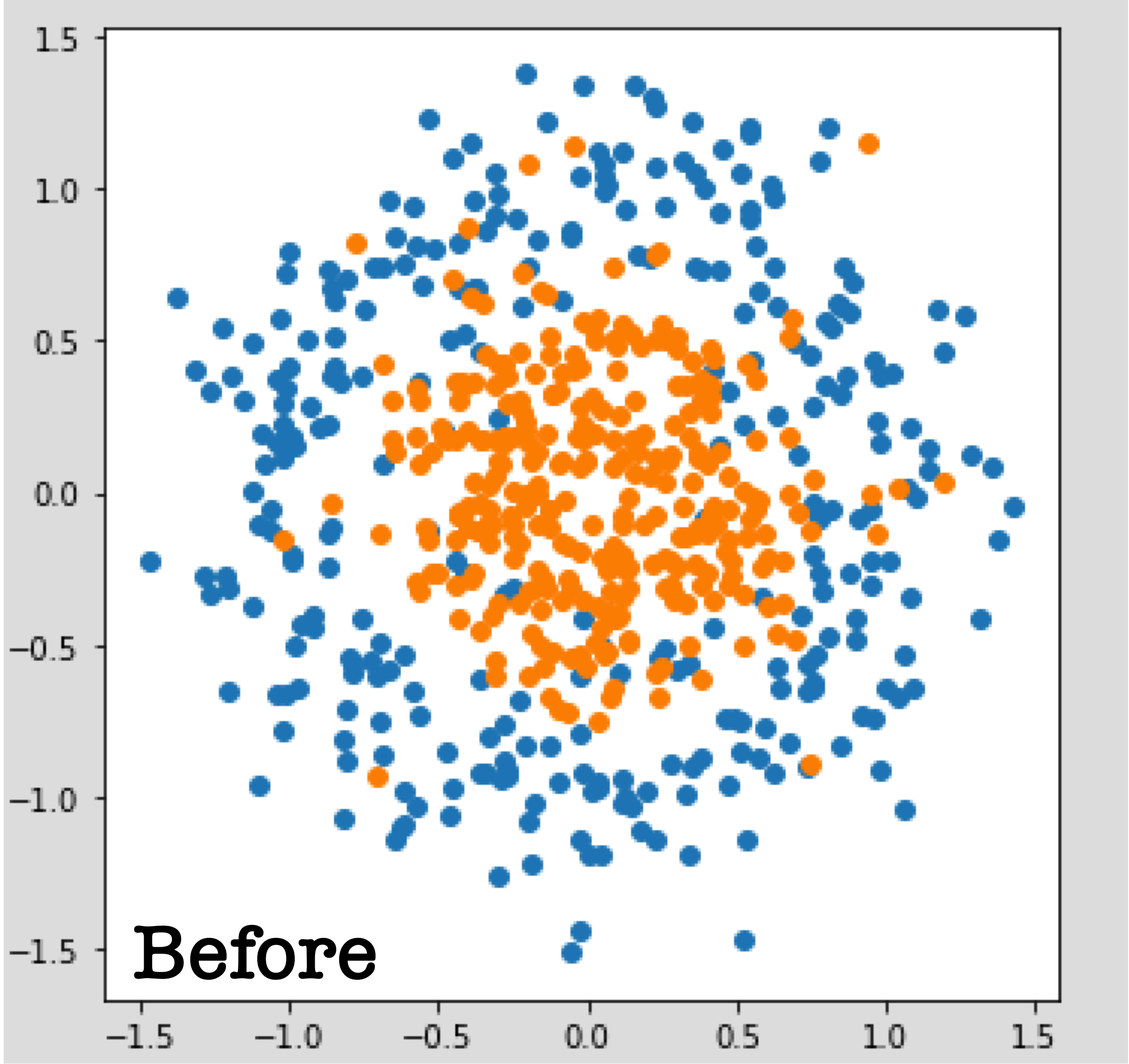
Task 0: Synthetic

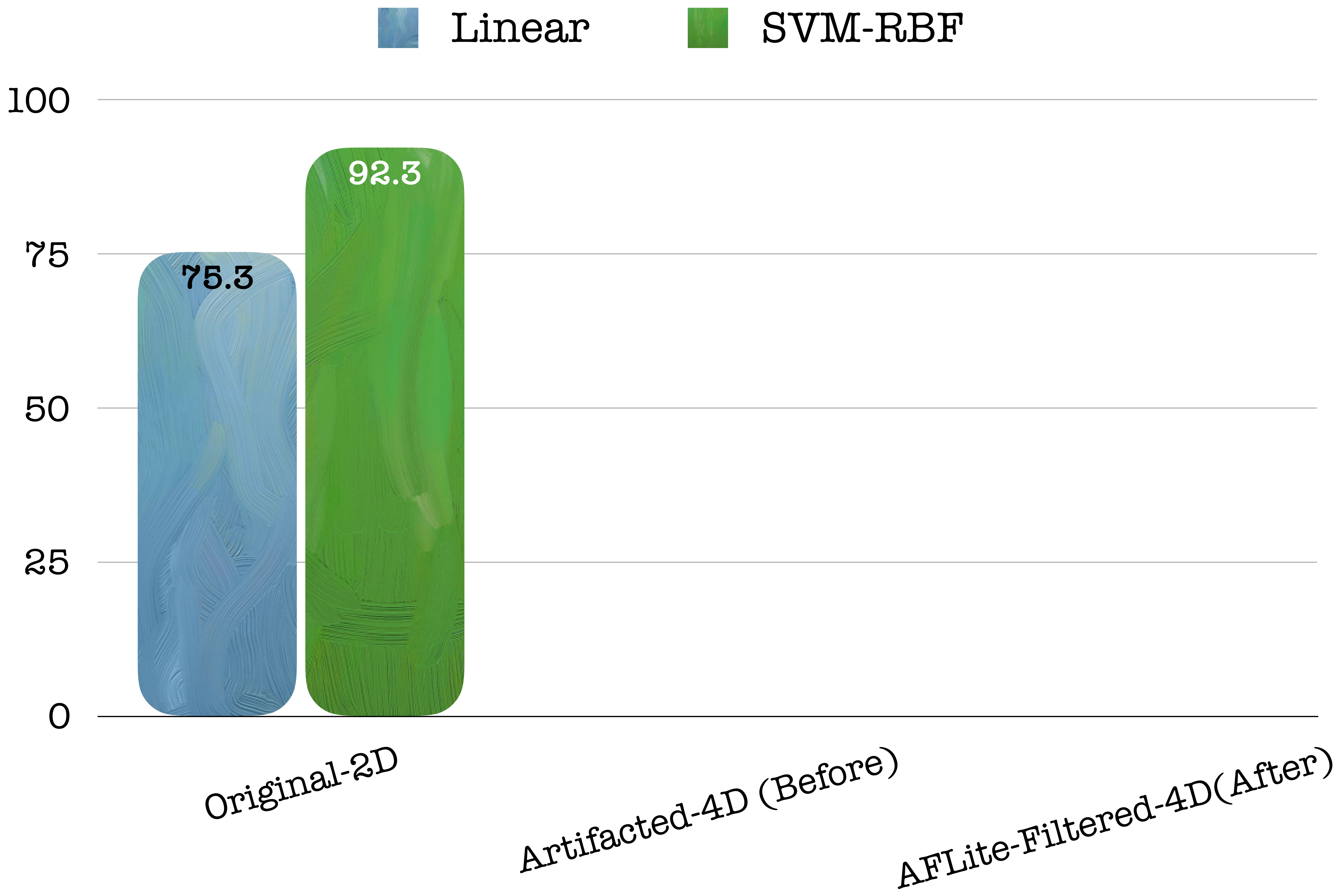




Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

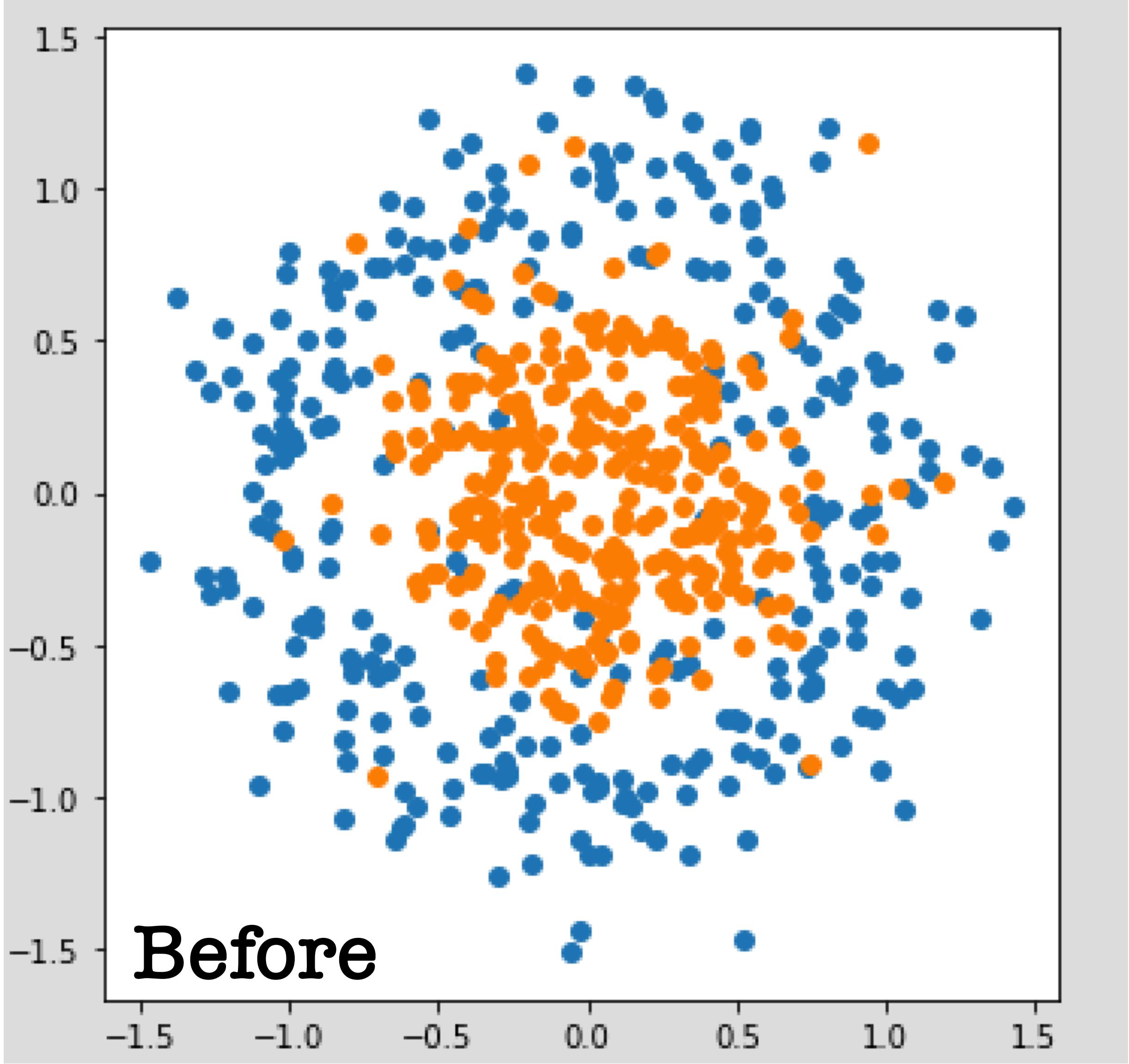
Task 0: Synthetic

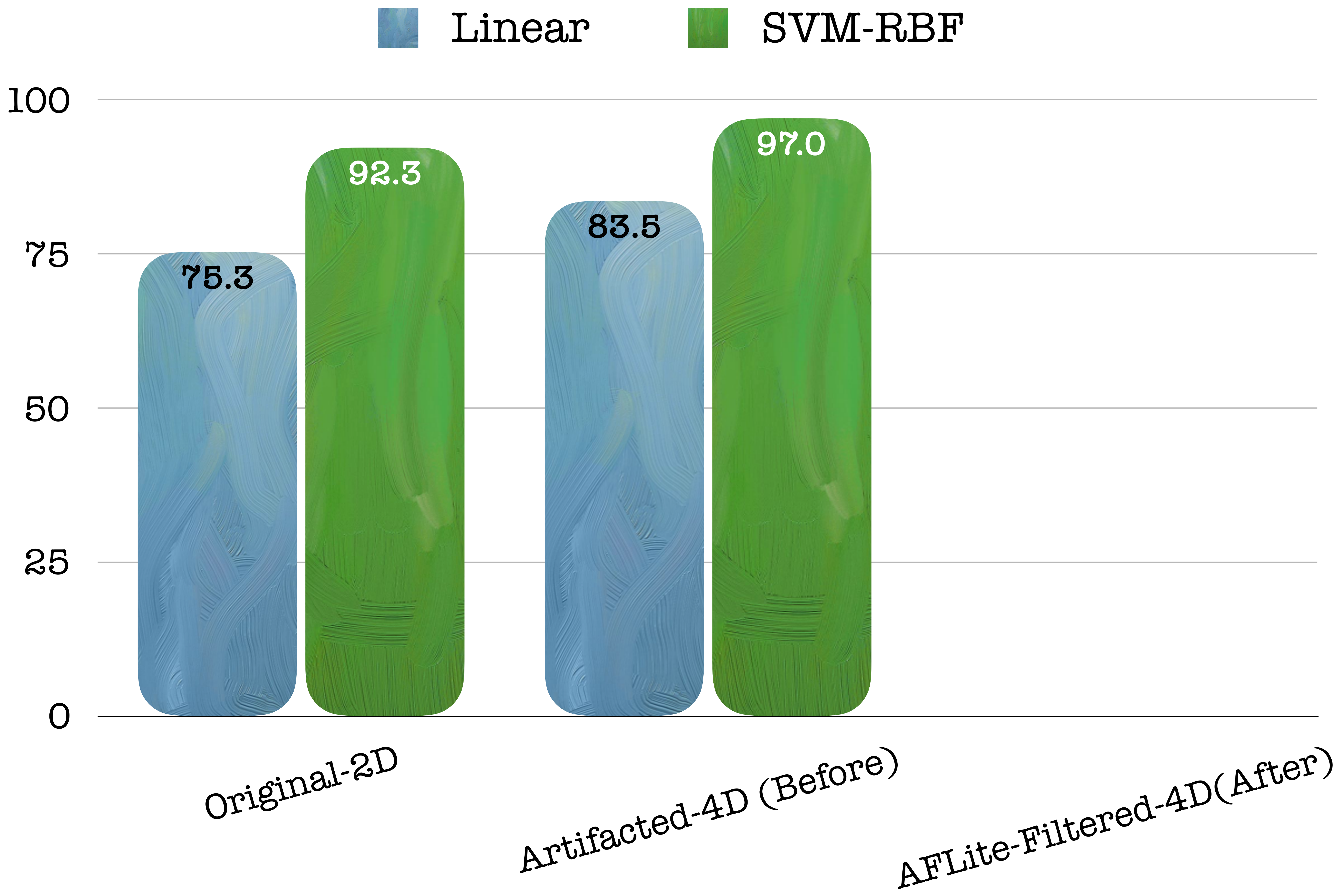




Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

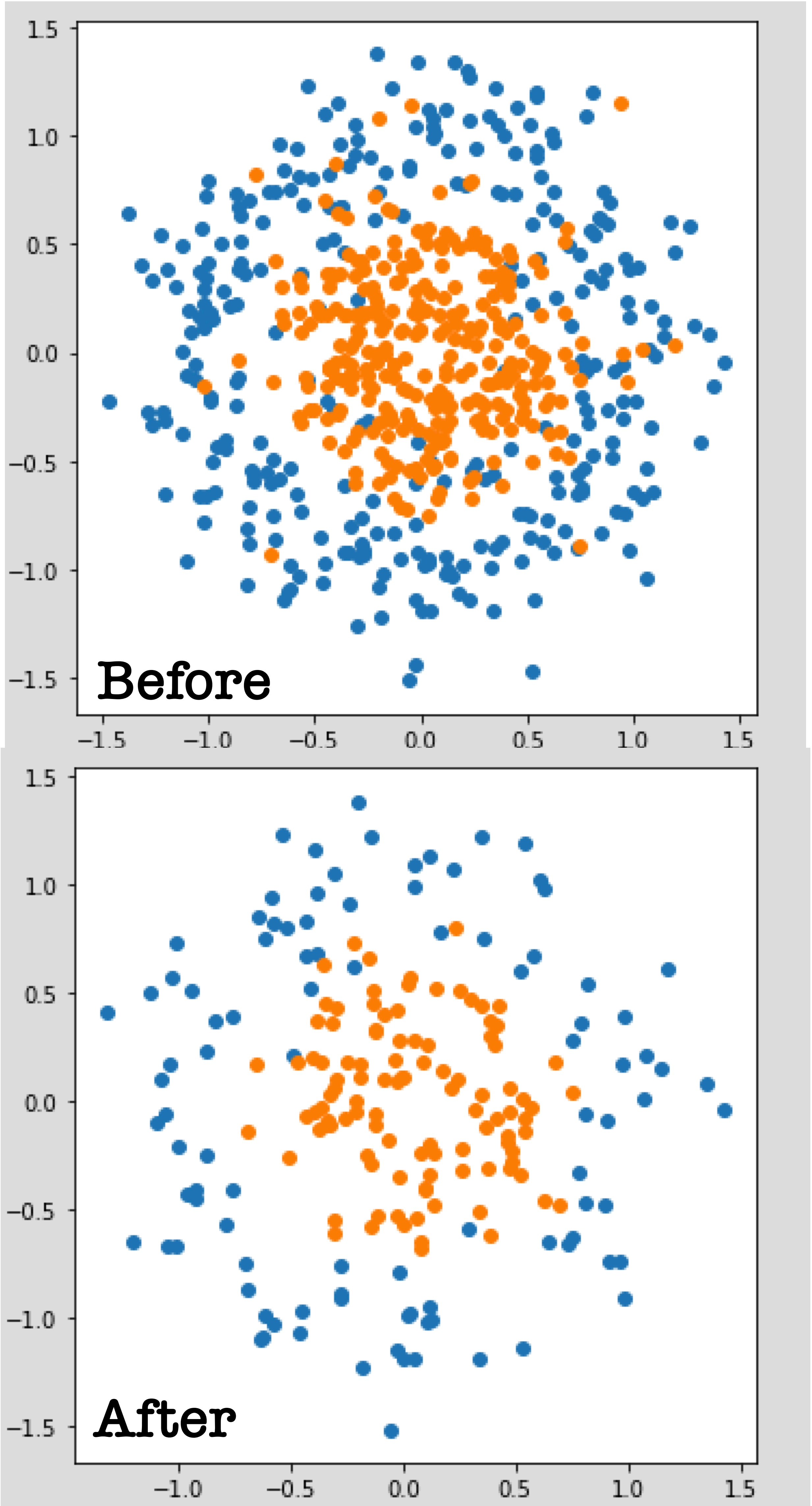
Task 0: Synthetic

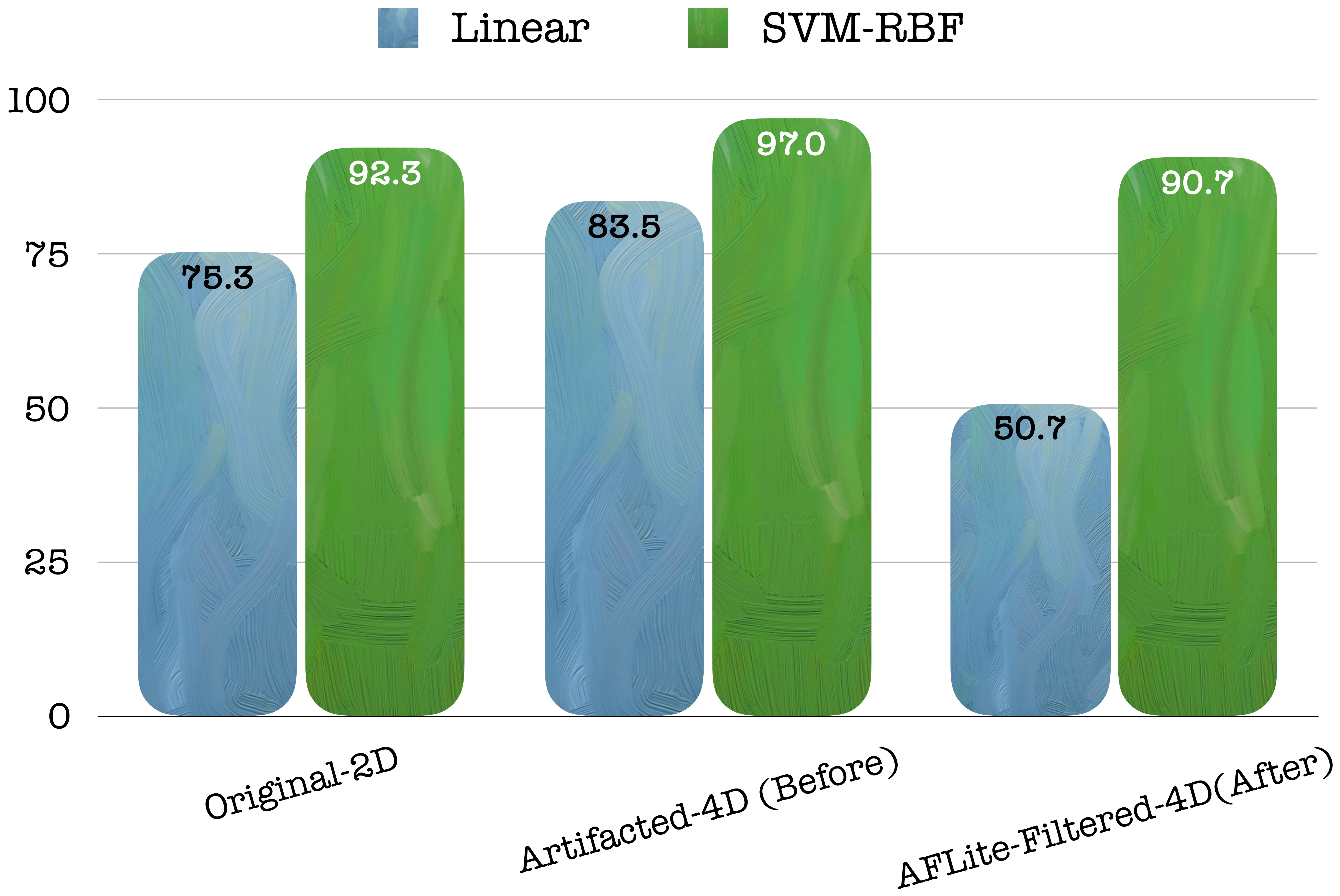




Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 0: Synthetic



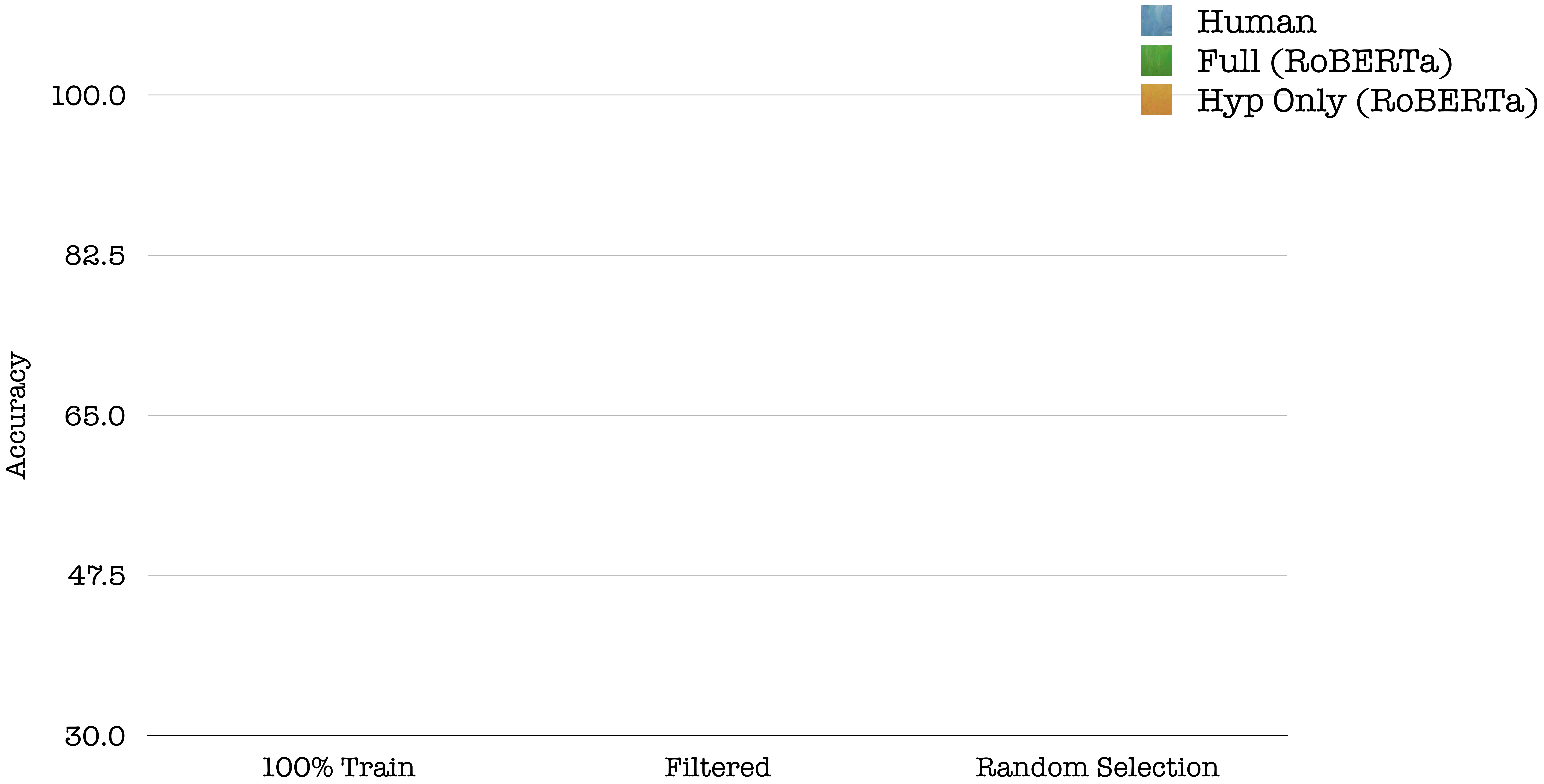


Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 1: SNLI performance

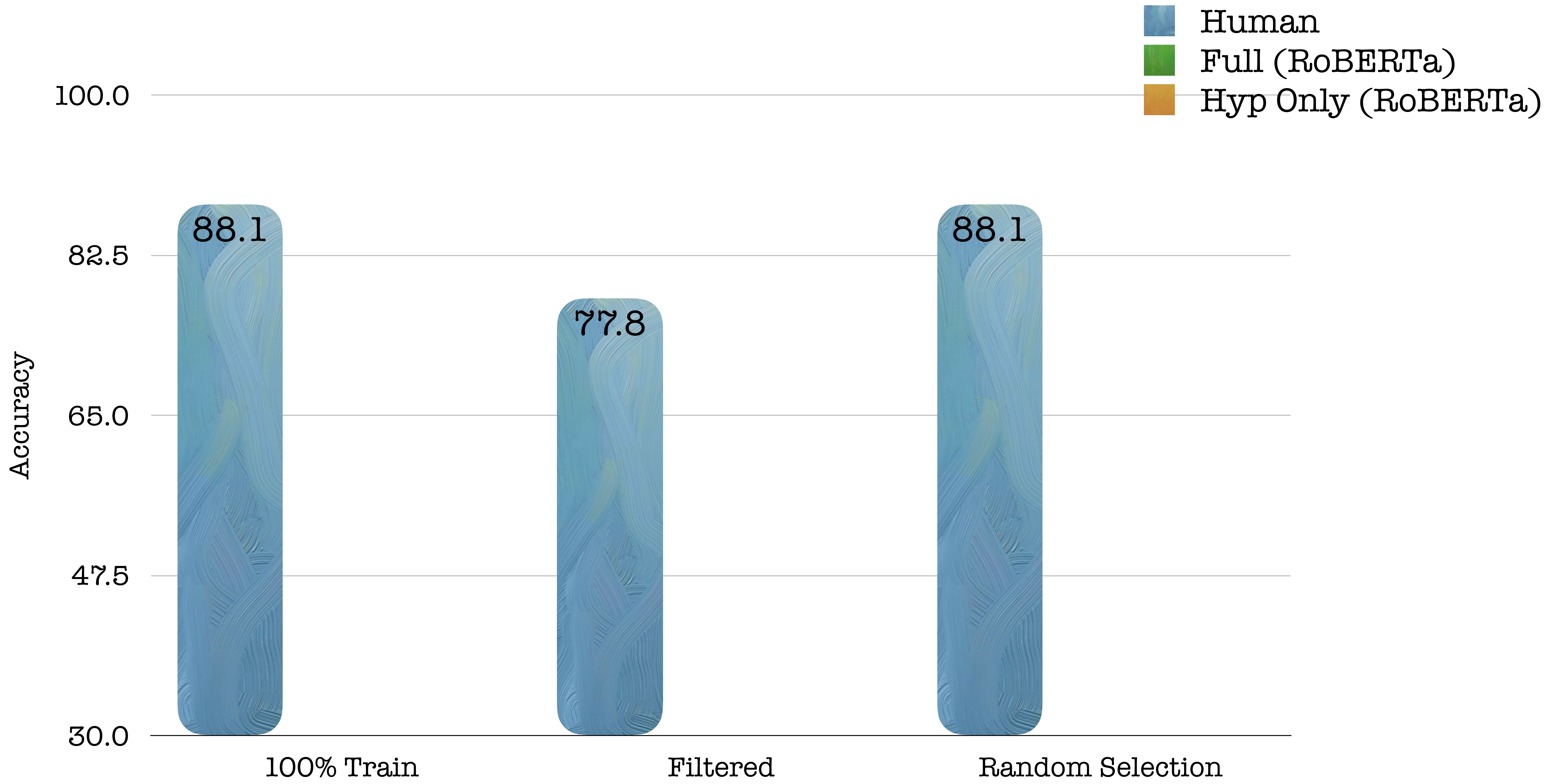
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 1: SNLI performance



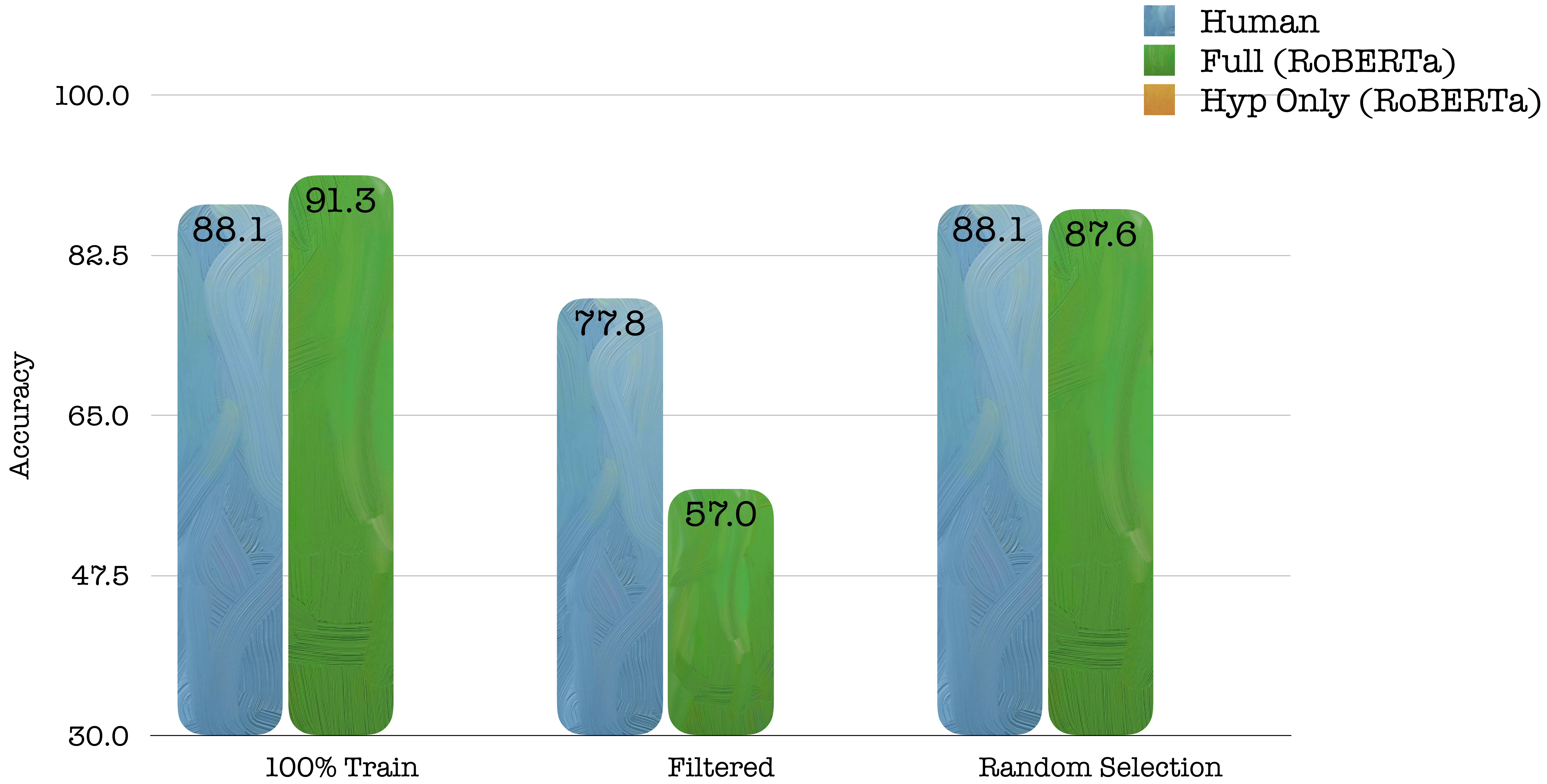
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 1: SNLI performance



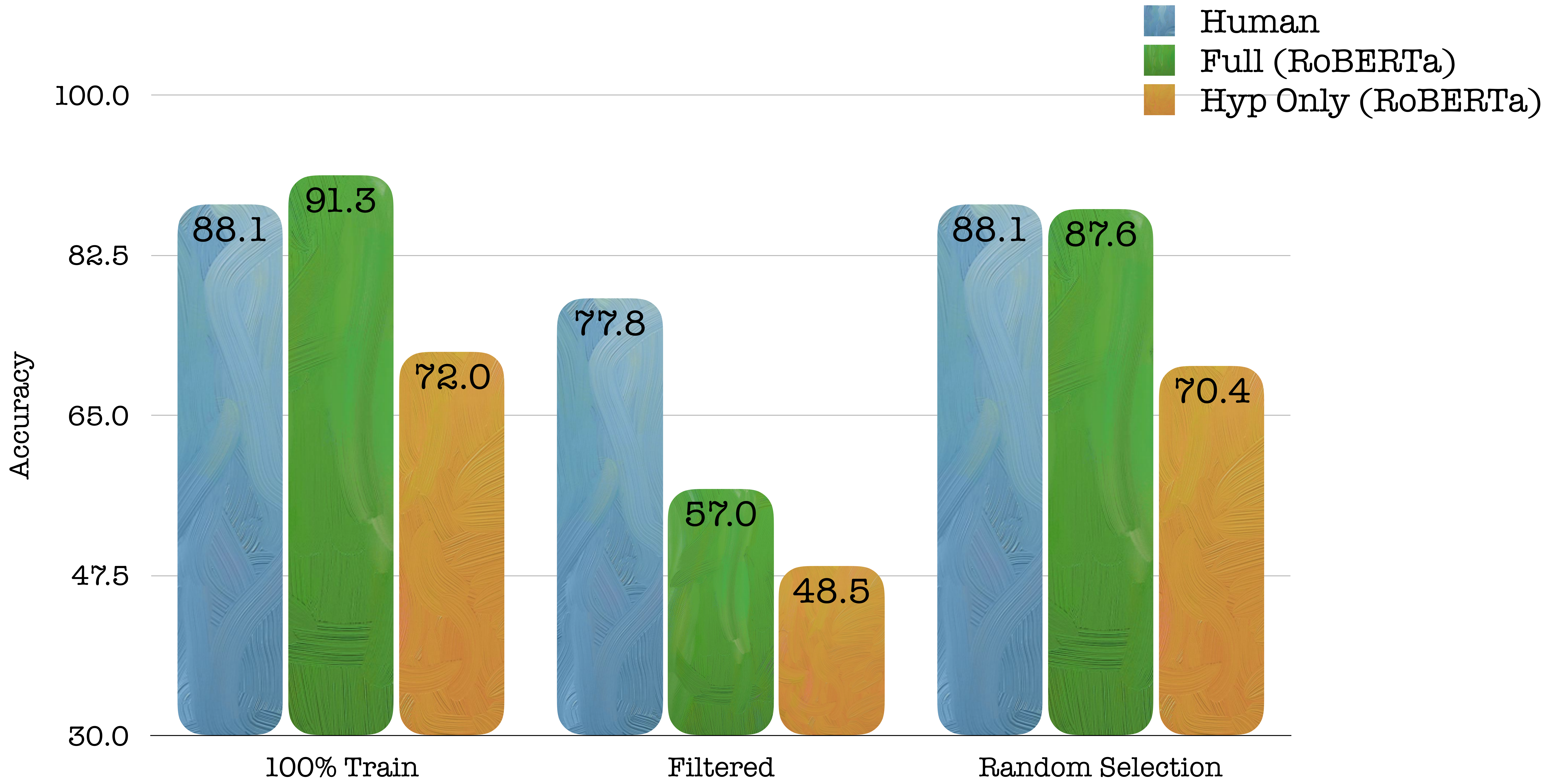
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 1: SNLI performance



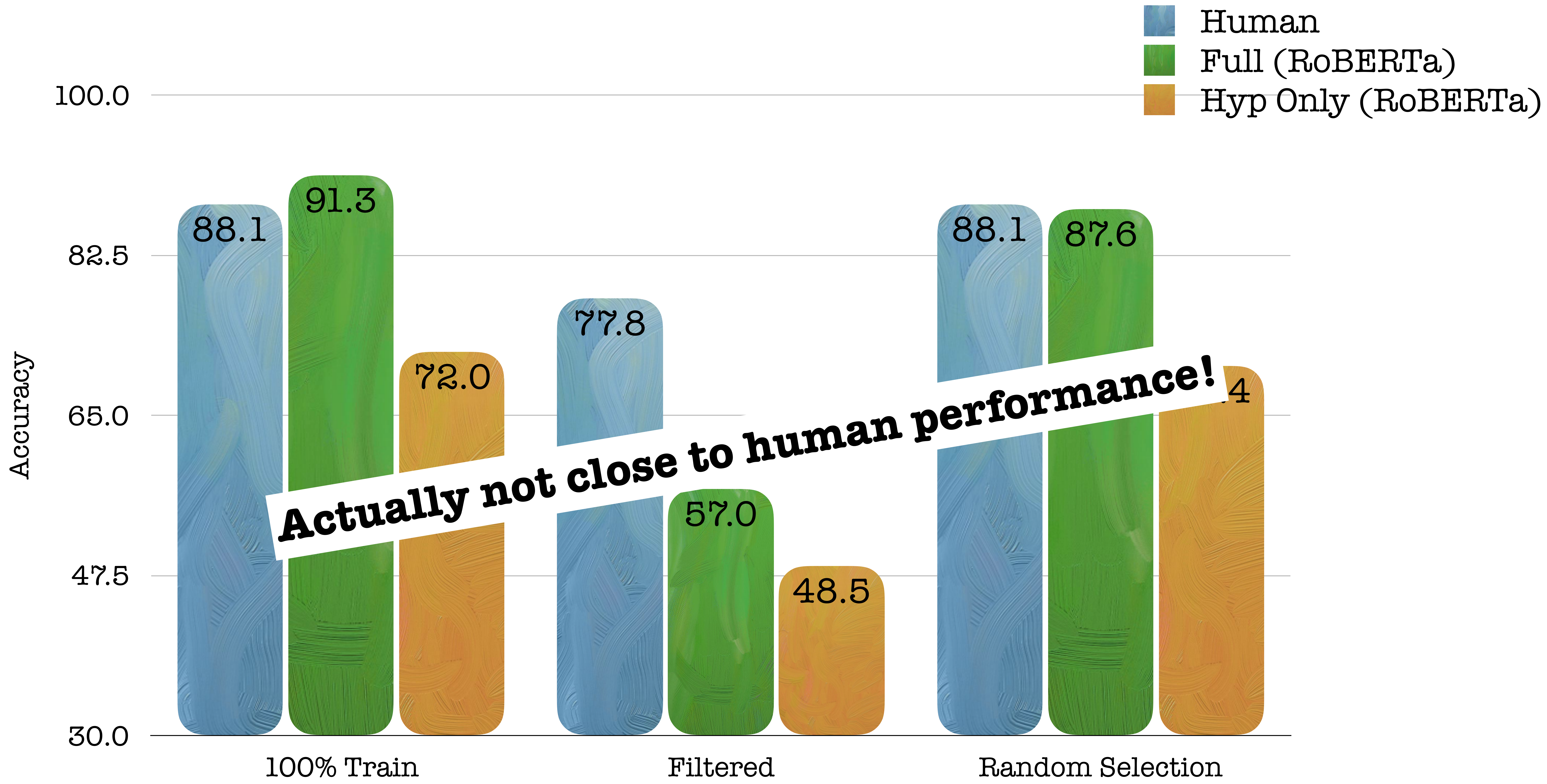
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 1: SNLI performance



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

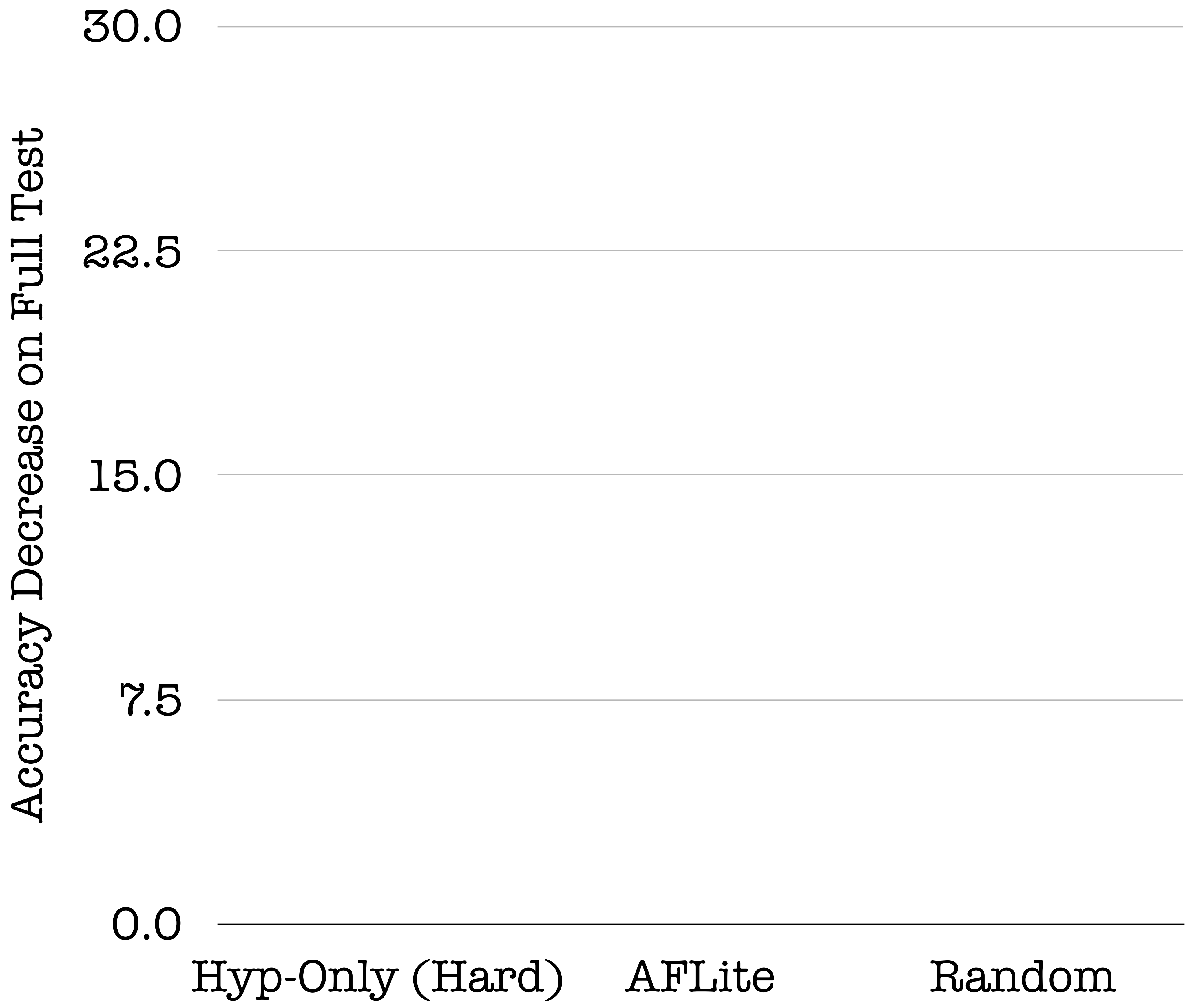
Task 1: SNLI performance



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

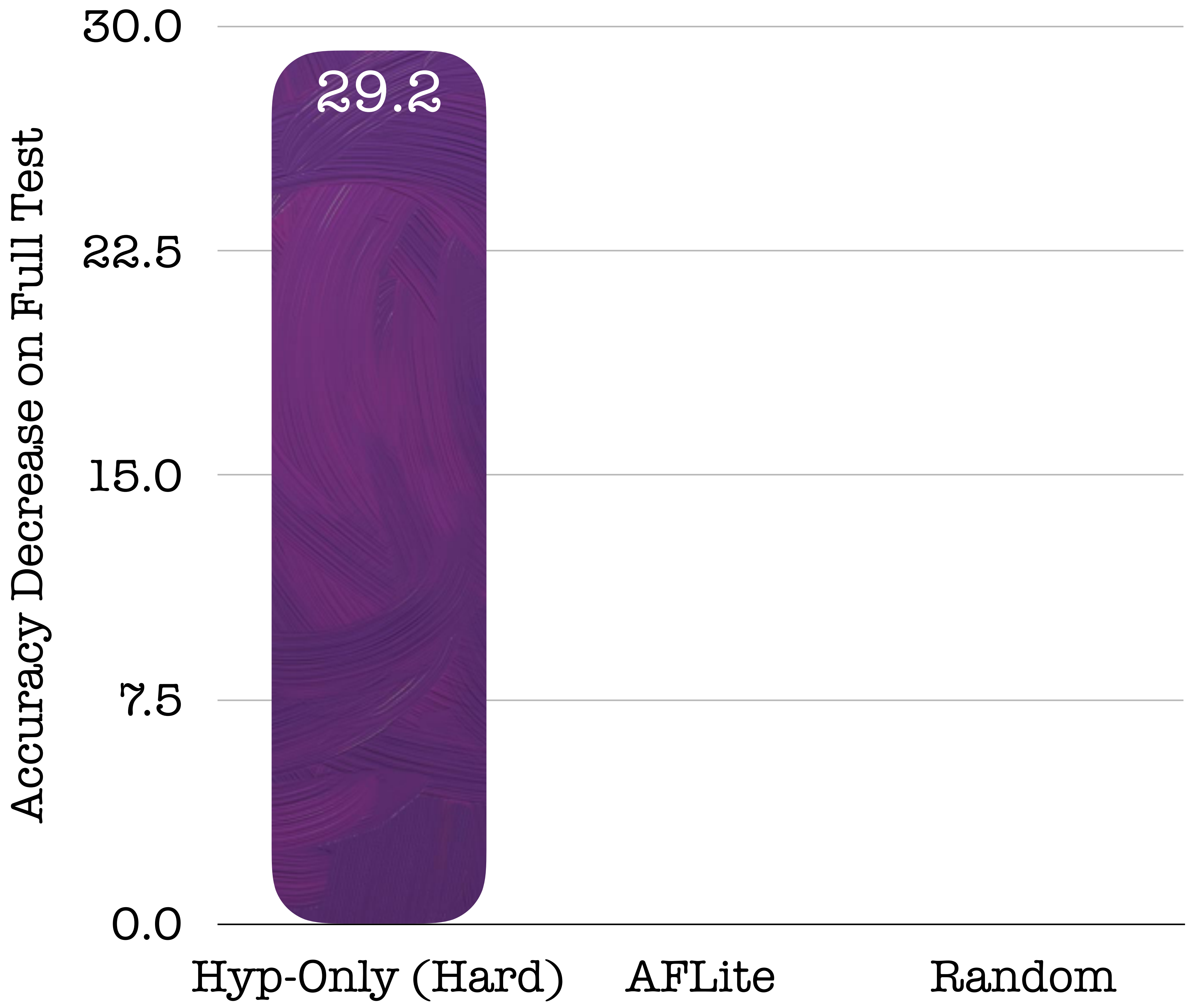
Comparison with Hyp-Only Filtering

Comparison with Hyp-Only Filtering



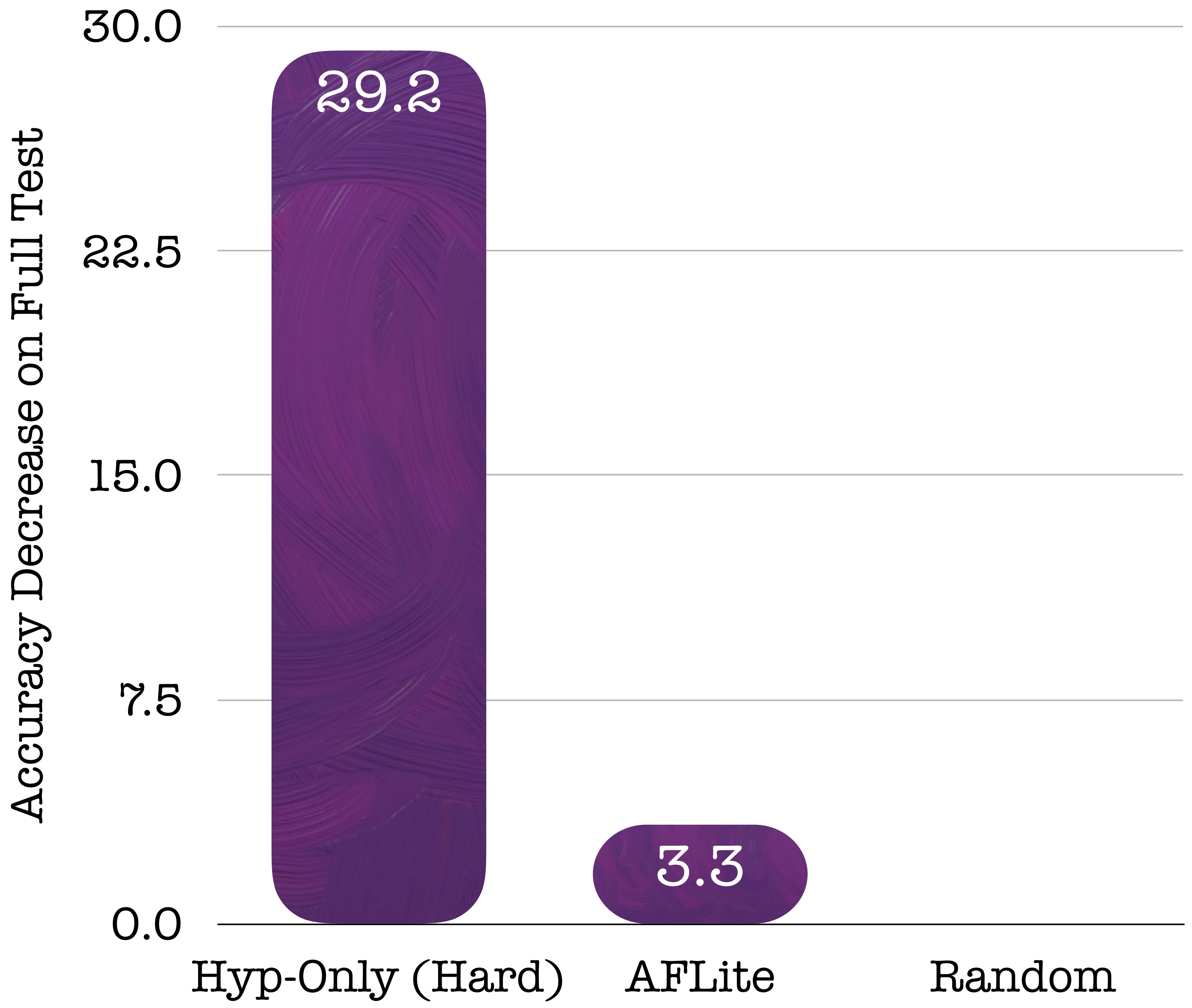
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Comparison with Hyp-Only Filtering



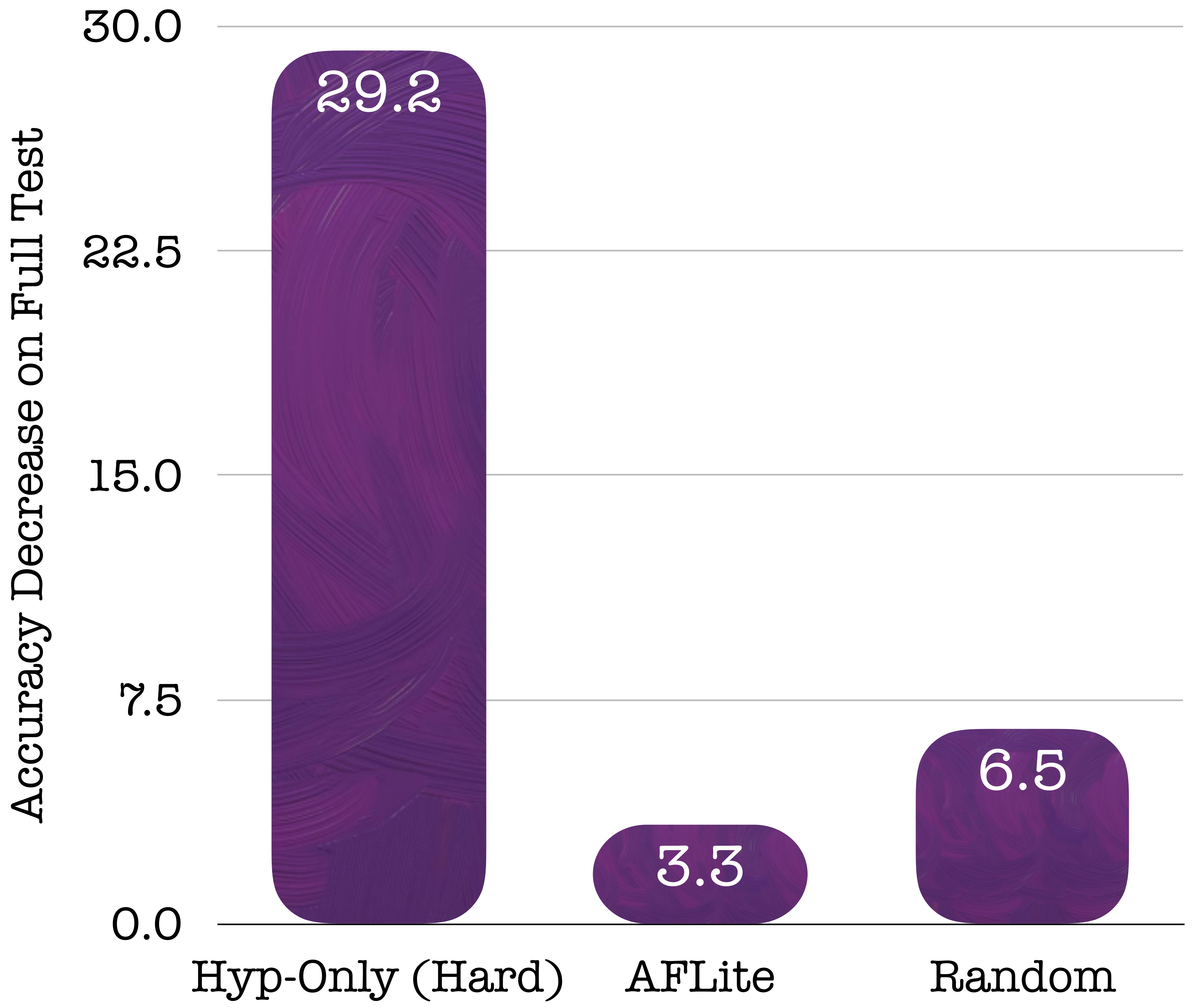
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Comparison with Hyp-Only Filtering



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

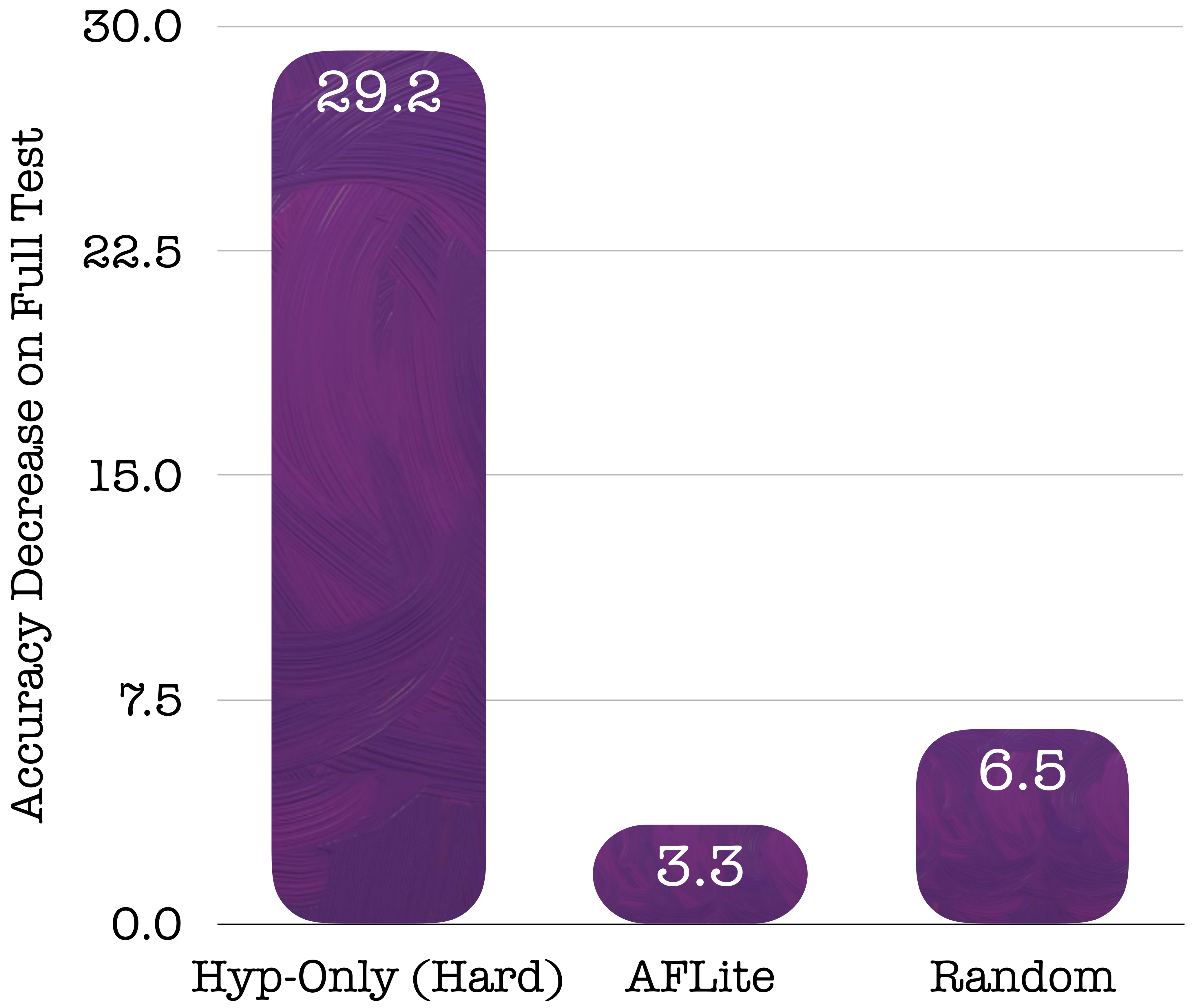
Comparison with Hyp-Only Filtering



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Comparison with Hyp-Only Filtering

- AFLite retains generalizability to many examples.
- Manually detecting artifacts can only get rid of some
- Manual filtering for **balancing** artifacts might not be effective.



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Digging Deeper Again

- Different word-class associations.

E	N	C
Transportation	Nearby	Nobody
Least	Pretending	Stealing
Awake	Backyard	Aliens
Sport	Dalmation	Sandbox
Multiple	Farmer	Stole

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Digging Deeper Again

- Different word-class associations.

E	N	C
Transportation	Nearby	Nobody
Least	Pretending	Stealing
Awake	Backyard	Aliens
Sport	Dalmation	Sandbox
Multiple	Farmer	Stole

E	N	C
Outdoors	Tall	Nobody
Least	First	Sleeping
Instrument	Competition	No
Outside	Sad	Tv
Animal	Favorite	Cat

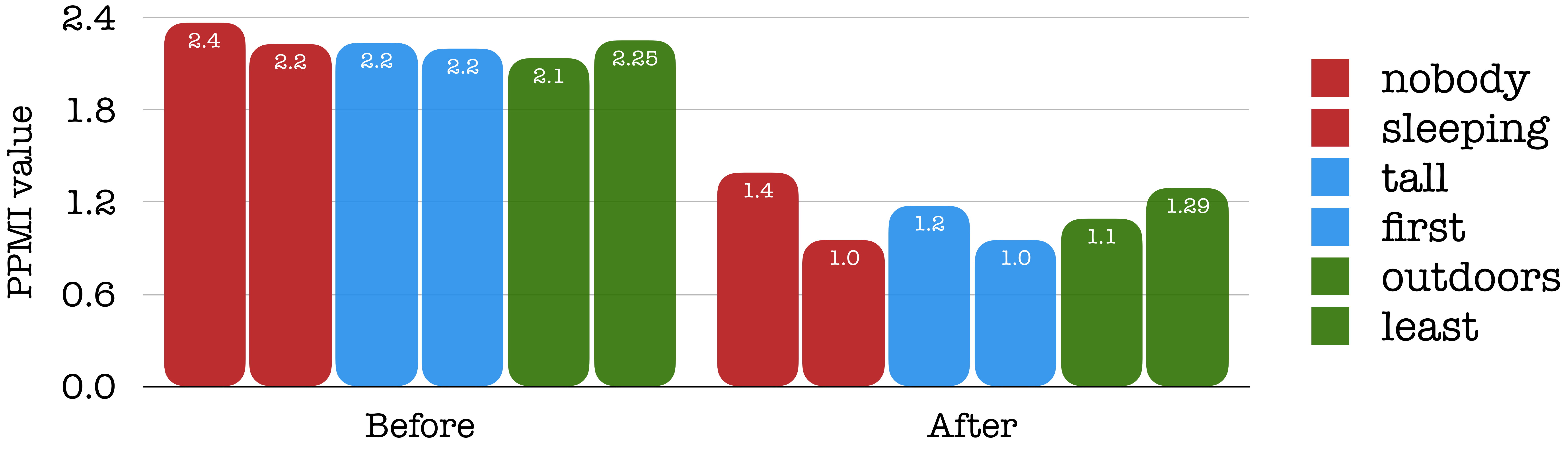
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 1: Word-Class Association

- Overall word-class association decreases

Effect 1: Word-Class Association

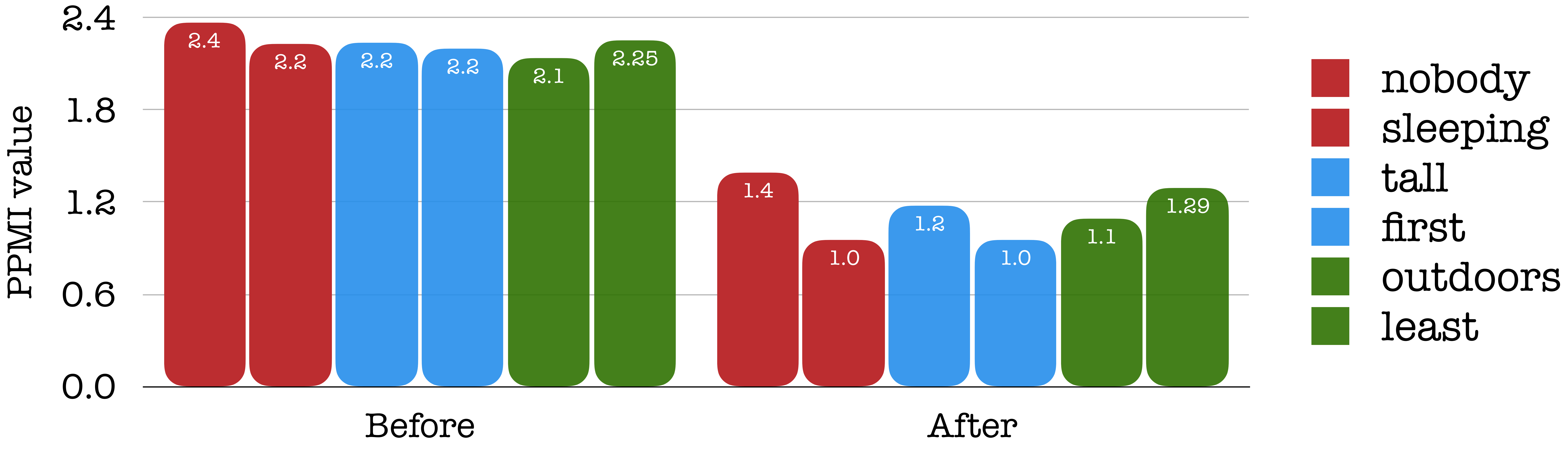
- Overall word-class association decreases



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 1: Word-Class Association

- Overall word-class association decreases

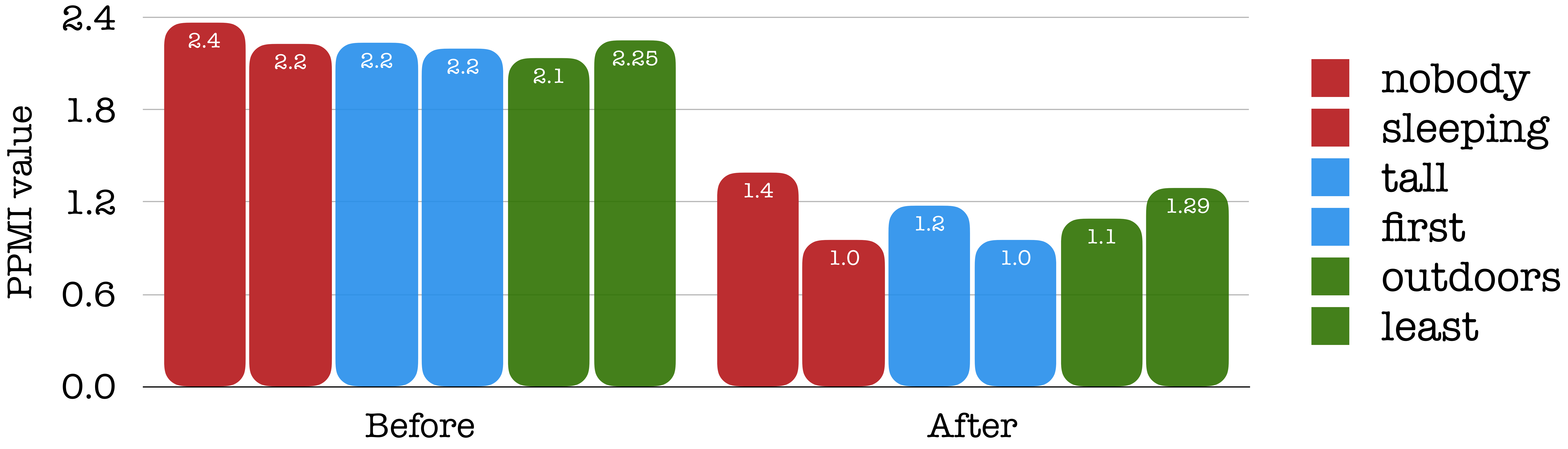


- If word association was the only indicator

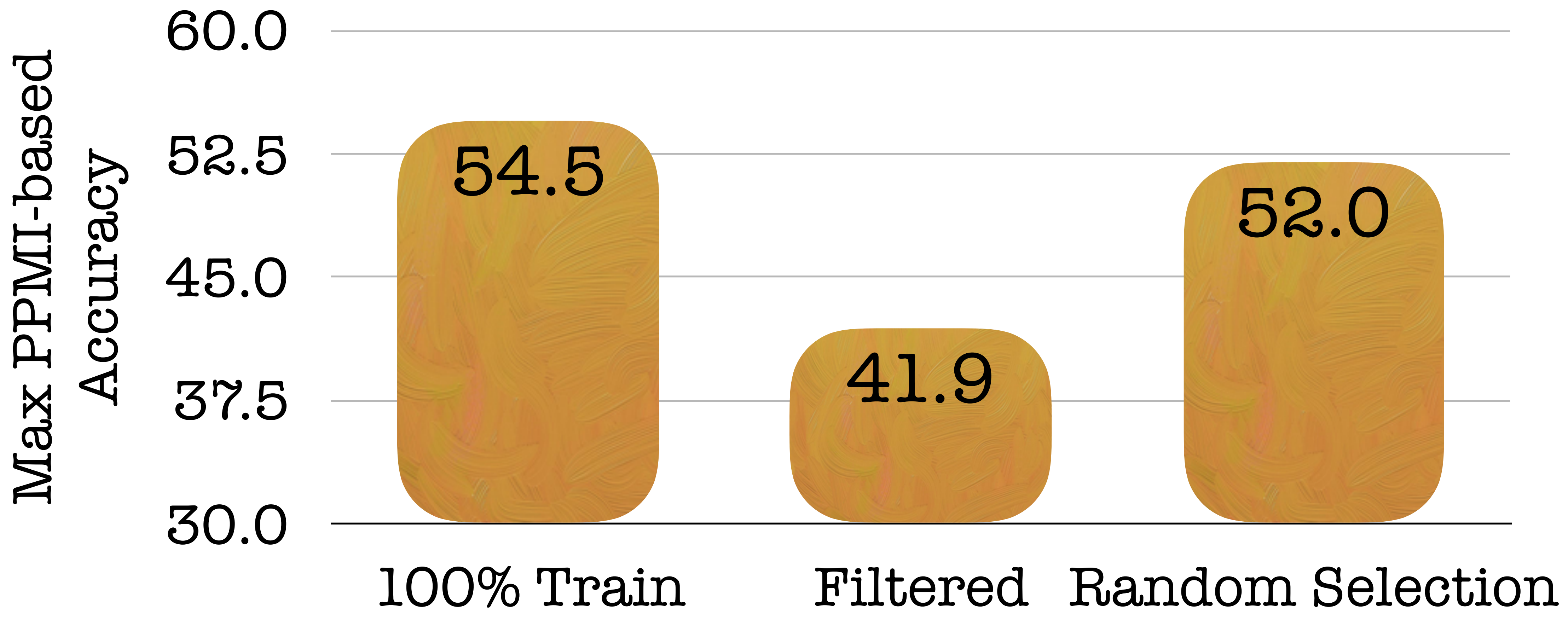
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 1: Word-Class Association

- Overall word-class association decreases



- If word association was the only indicator



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 2: Newer, less common artifacts

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 2: Newer, less common artifacts

Premise		Hypothesis	
Punk rocker playing the drums and singing.	**Appearance** →	A boy is **pretending** he is a musician	**N**

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 2: Newer, less common artifacts

Punk rocker playing the drums and singing.

Premise

Appearance

A boy is **pretending** he is a musician

Hypothesis

N

In this photo, a little boy is smiling, and wearing what seems to be a cape.

Premise

Irrelevance

The birds were short **necked.**

Hypothesis

C

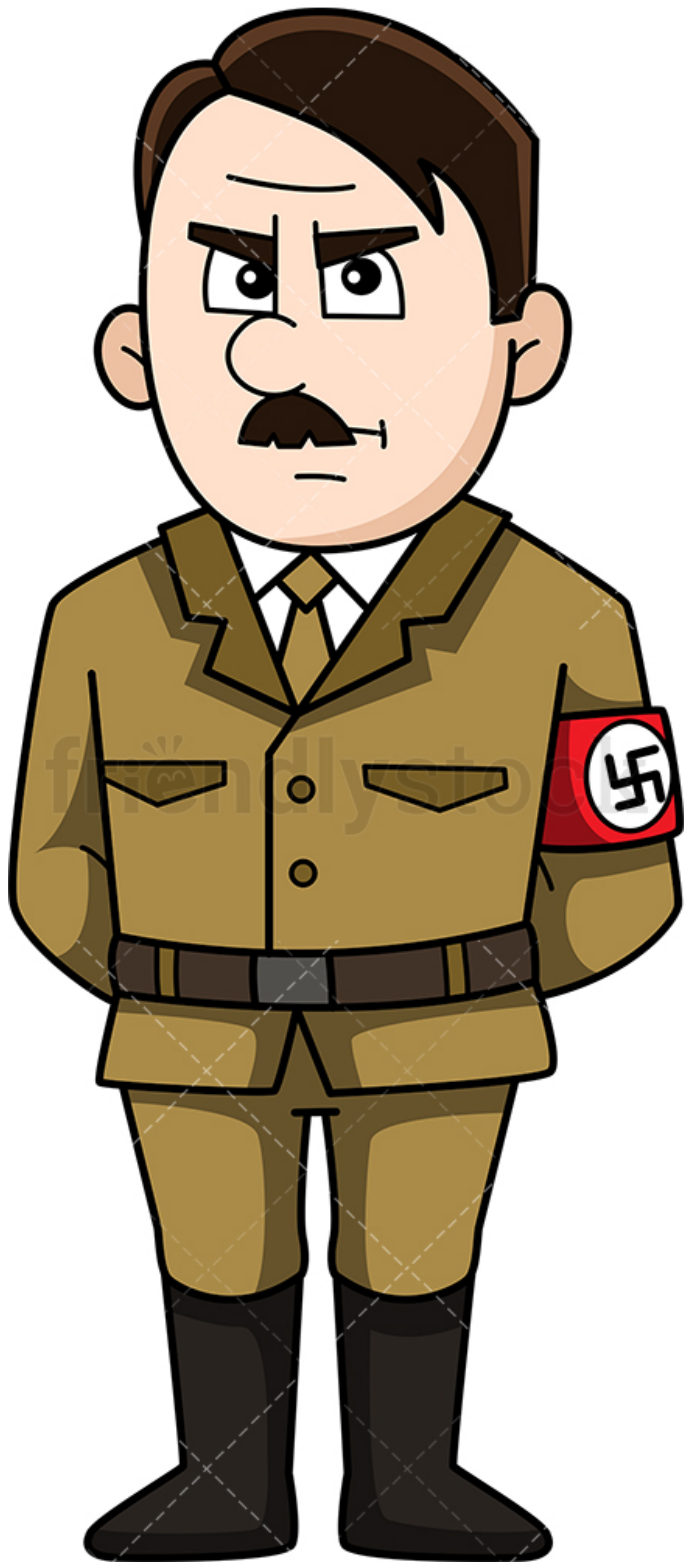
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 2: Newer, less common artifacts

Premise		Hypothesis	
Punk rocker playing the drums and singing.	**Appearance** →	A boy is **pretending** he is a musician	**N**
In this photo, a little boy is smiling, and wearing what seems to be a cape.	**Irrelevance** →	The birds were short **necked.**	**C**
A woman is standing up on a bus.	**Wokeness?** →	The woman is **awake.**	**E**

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 3: Ambiguous Artifacts



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Effect 3: Ambiguous Artifacts

Two men sit casually in folding chairs, gesturing, and speaking to one another.

Premise

Neutral

Hitler and Stalin share tea and crumpets.

Hypothesis

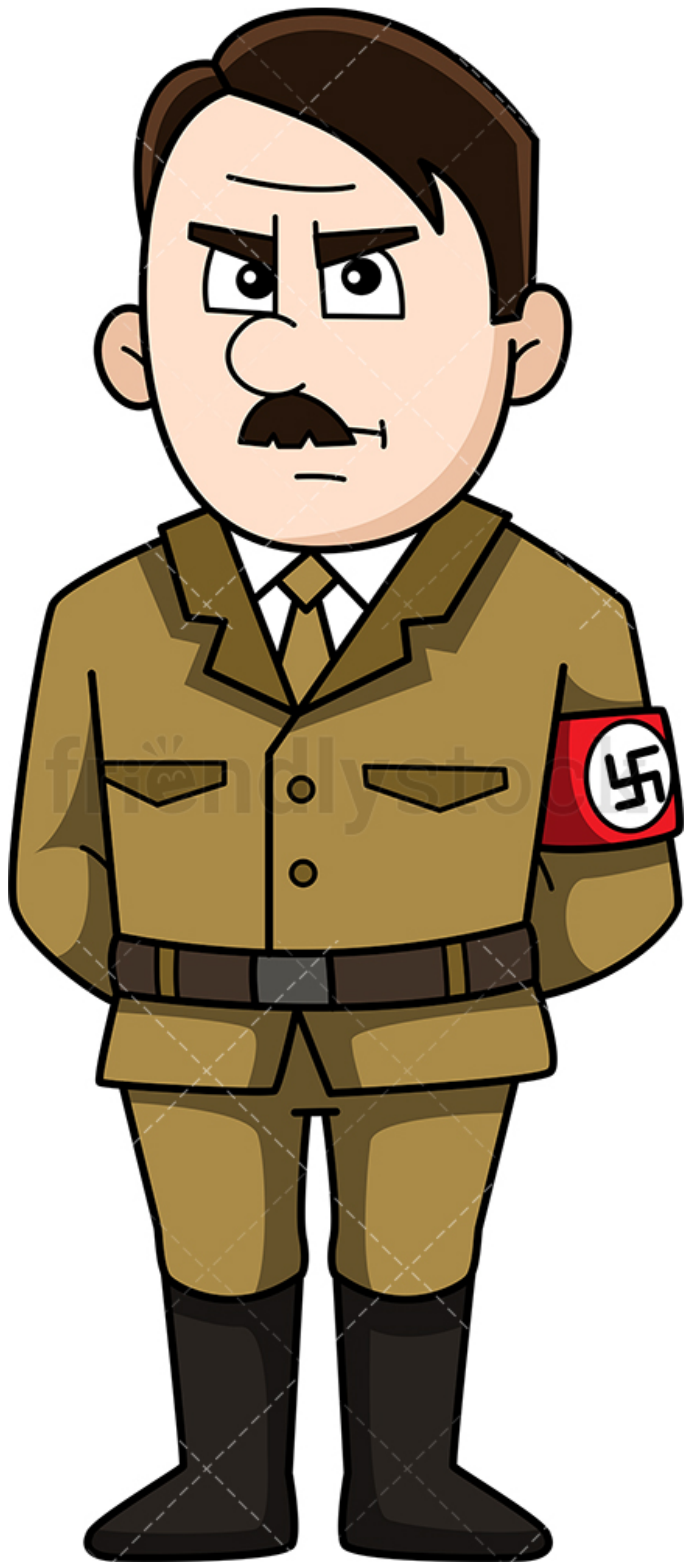
A man in a tie is holding a microphone while the people around him are cheering.

Premise

Contradiction

Hitler gives a speech.

Hypothesis

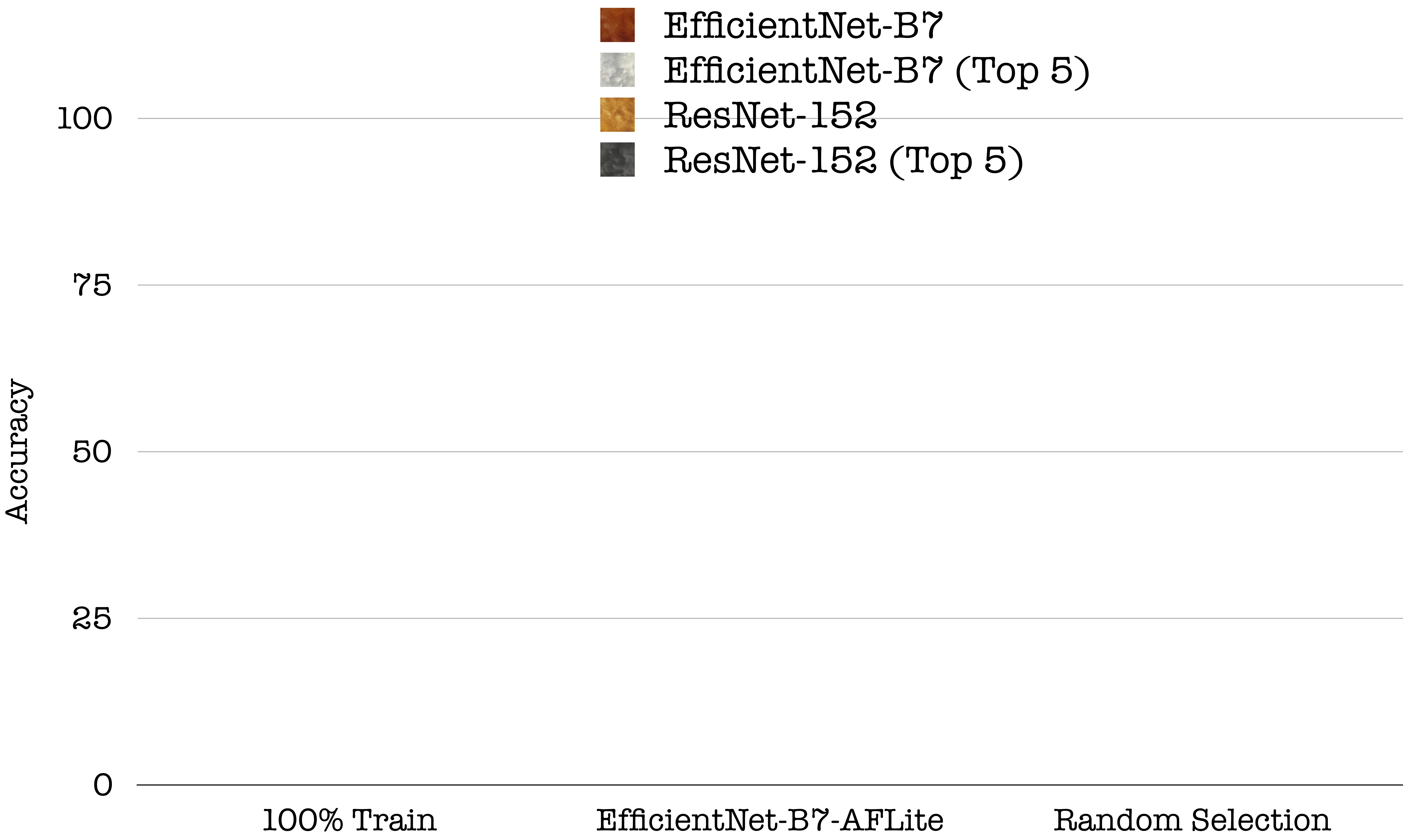


Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 4: Image Classification

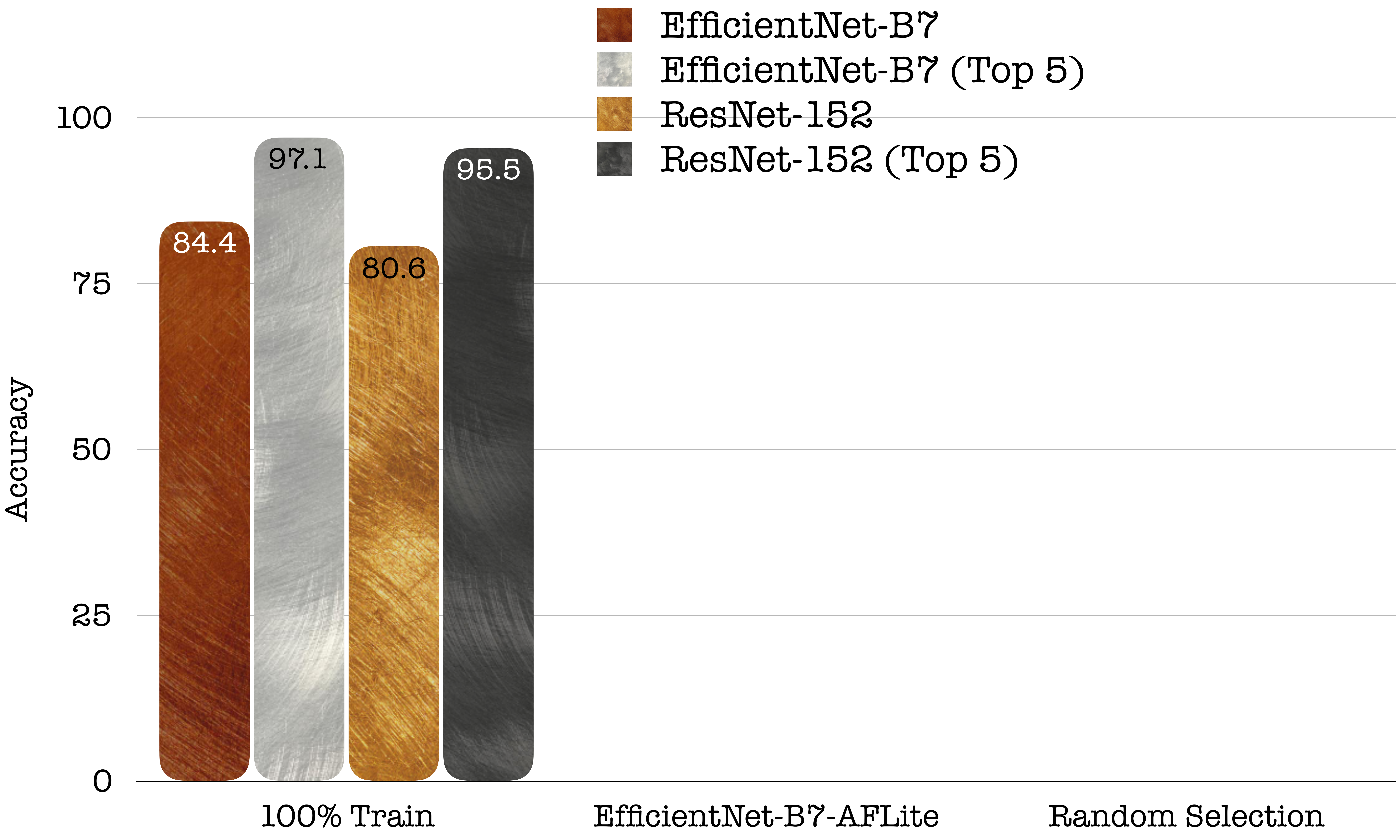
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 4: Image Classification



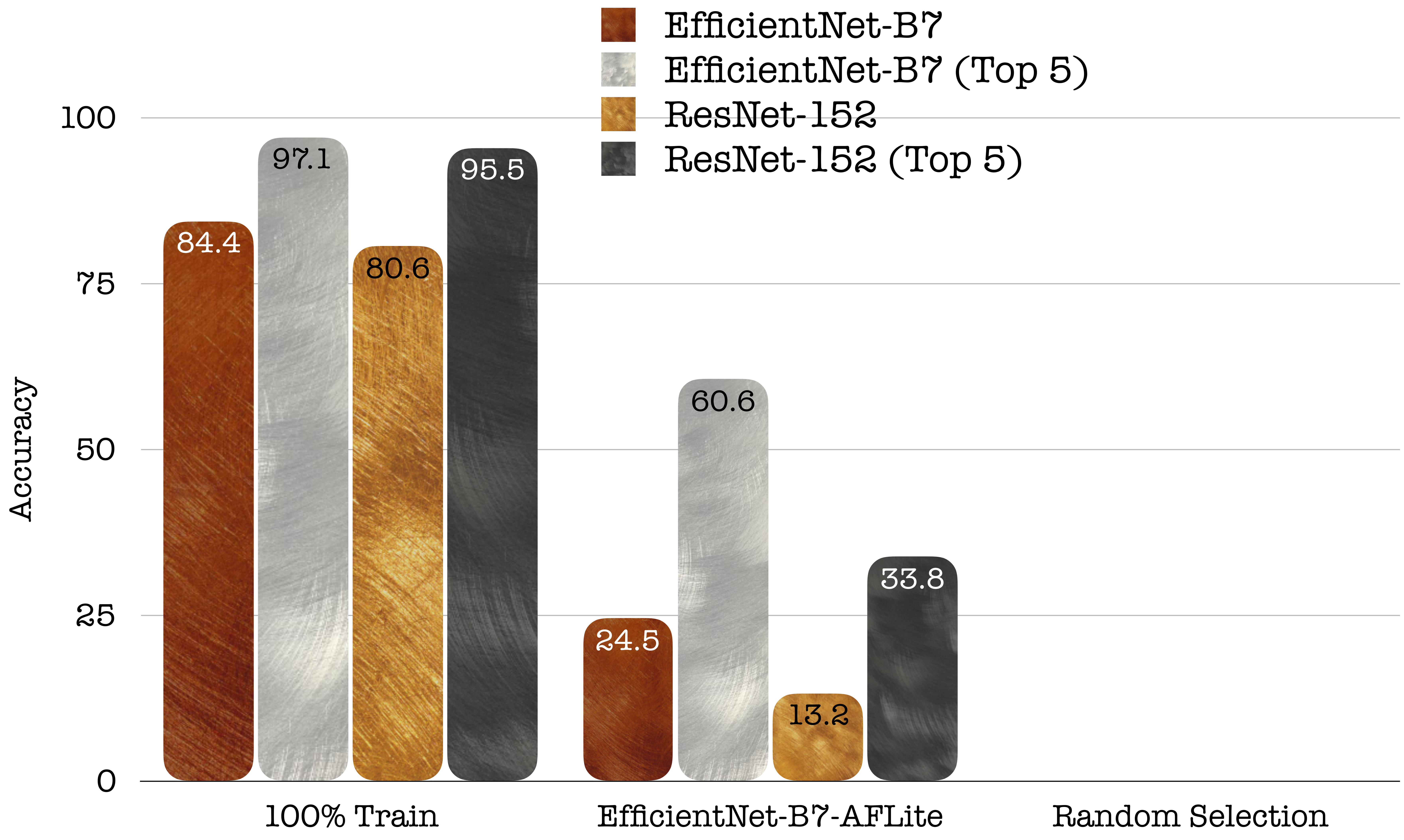
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 4: Image Classification



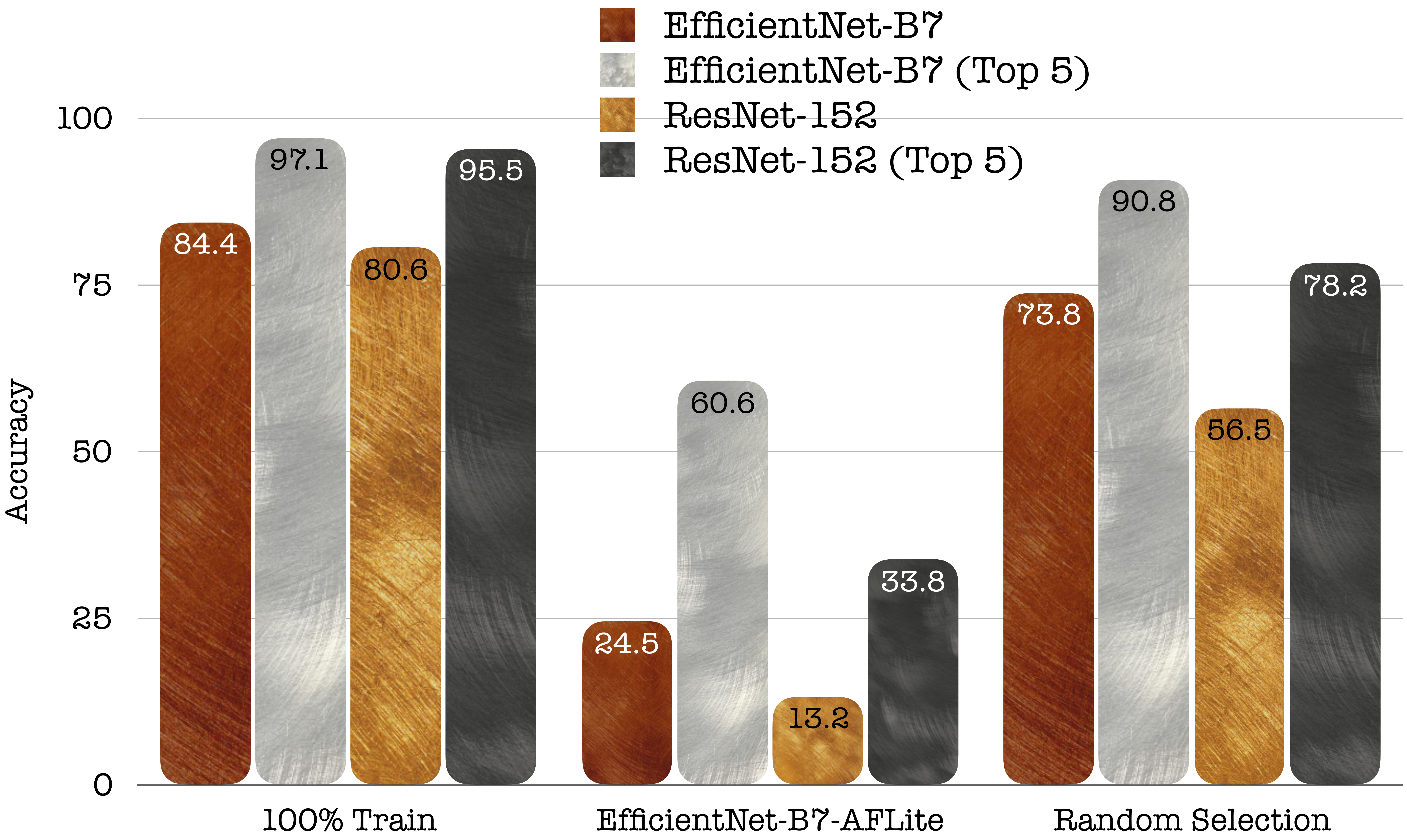
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 4: Image Classification



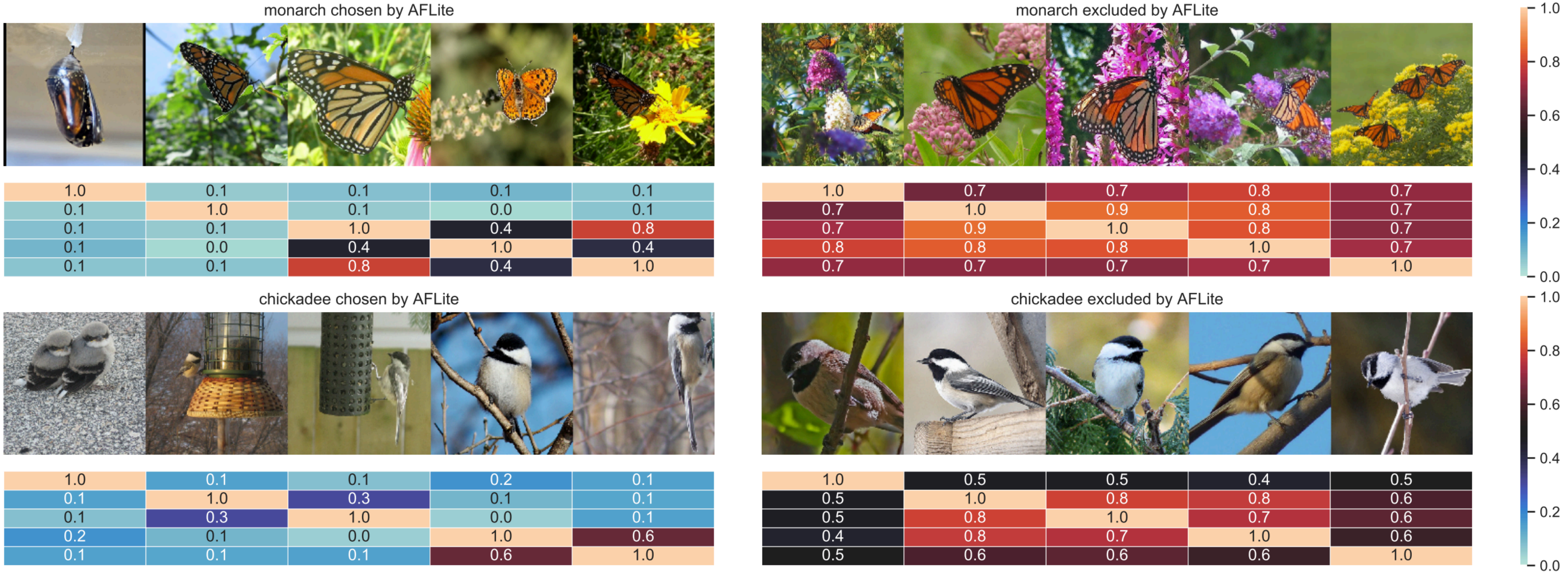
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Task 4: Image Classification



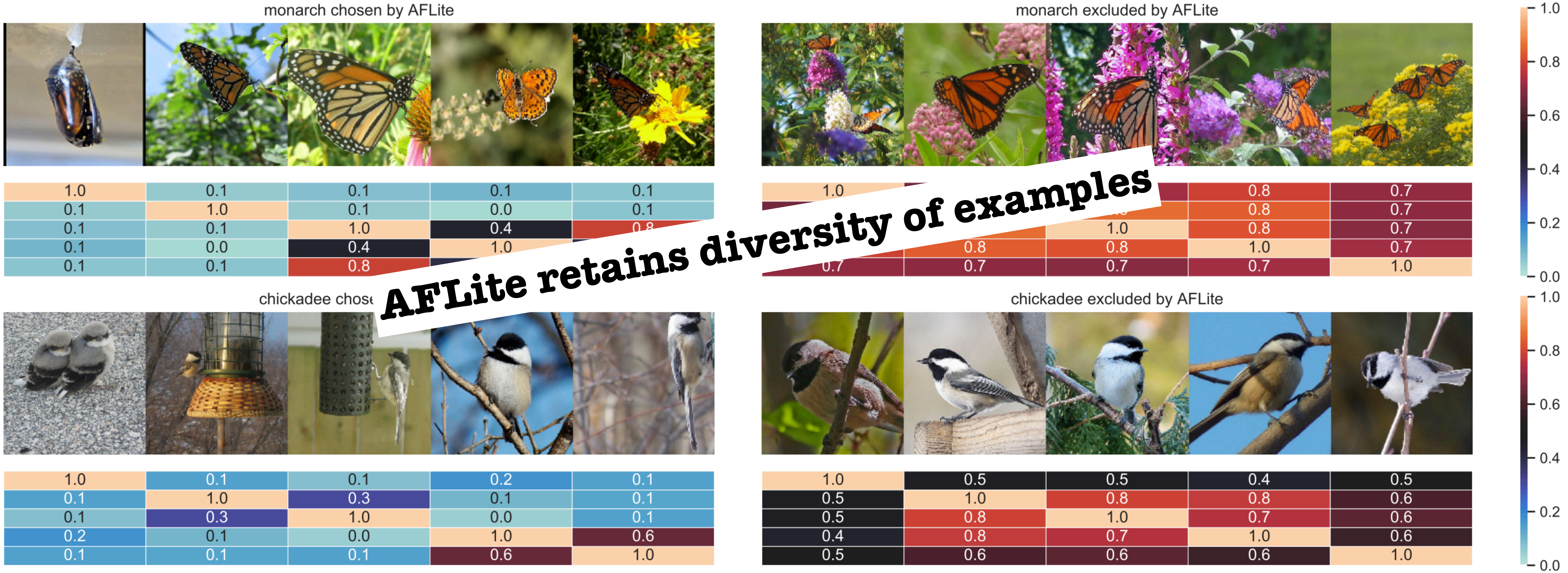
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

A nearest neighbor perspective



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

A nearest neighbor perspective

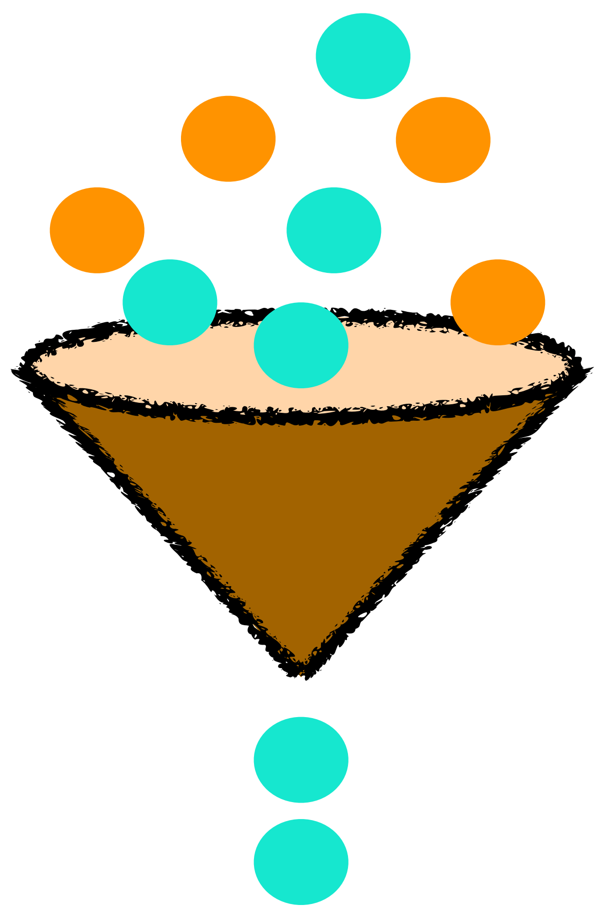


Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

Takeaways

Takeaways

Performance on some large-scale datasets could be misleading.

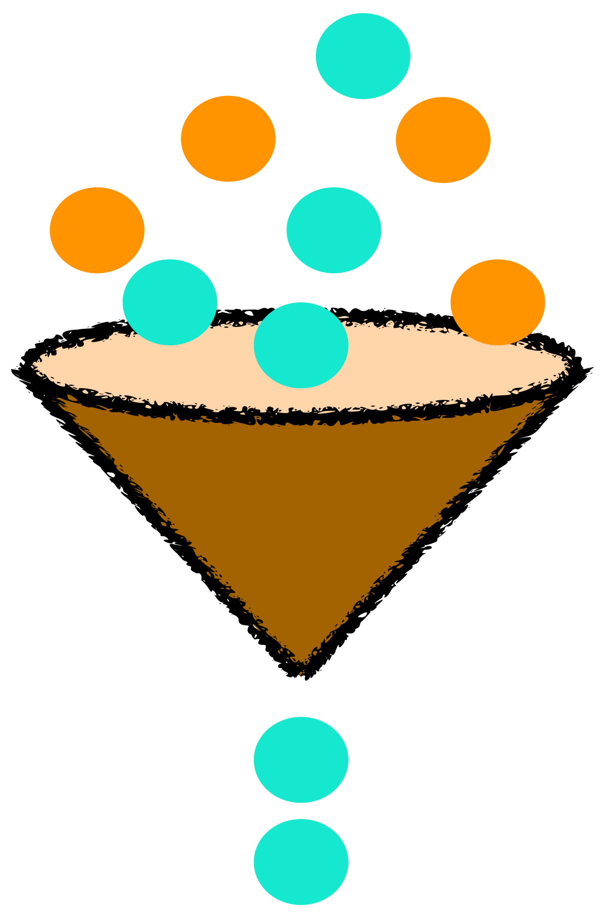


English-only #BenderRule

Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

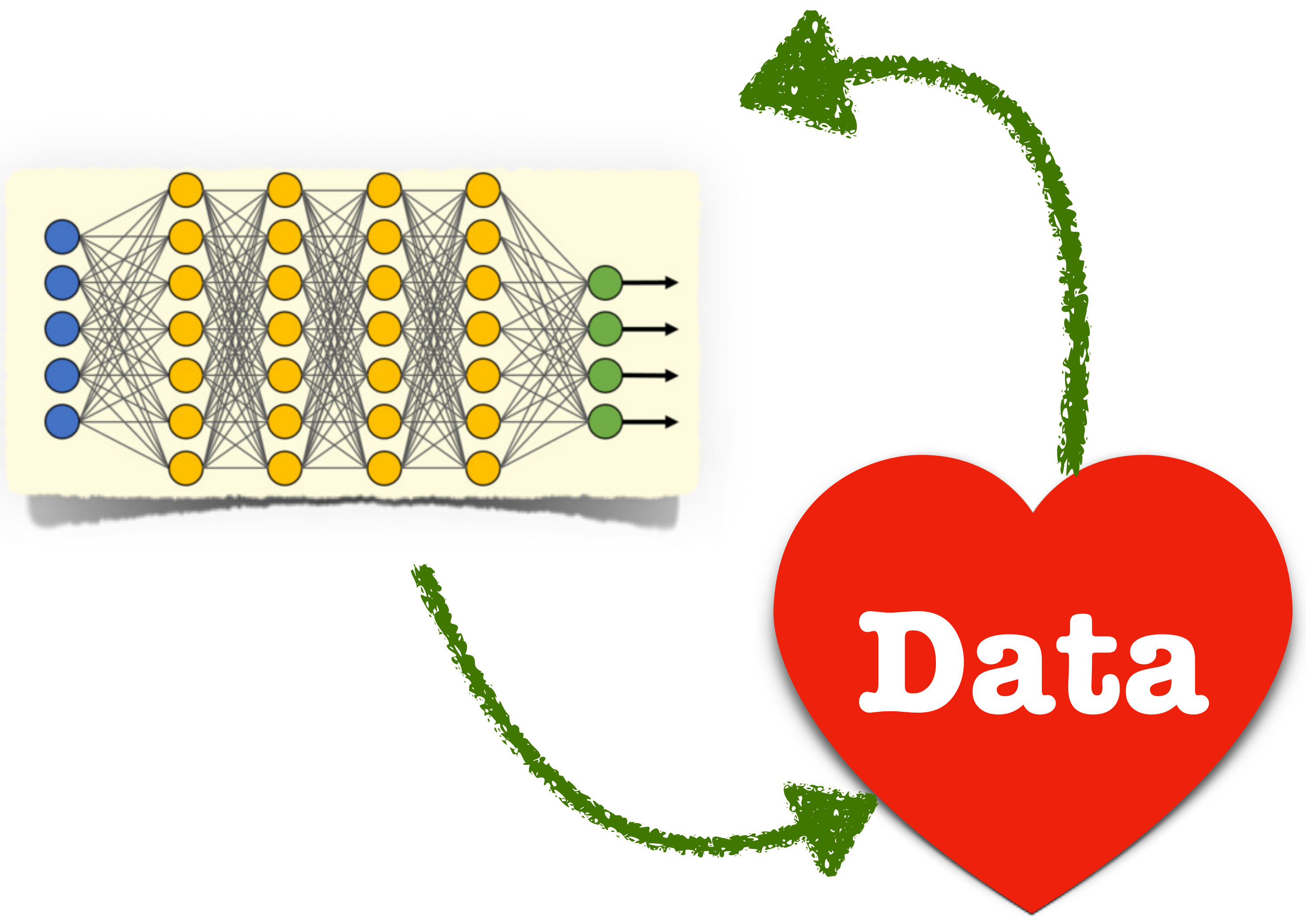
Takeaways

Performance on some large-scale datasets could be misleading.



English-only #BenderRule

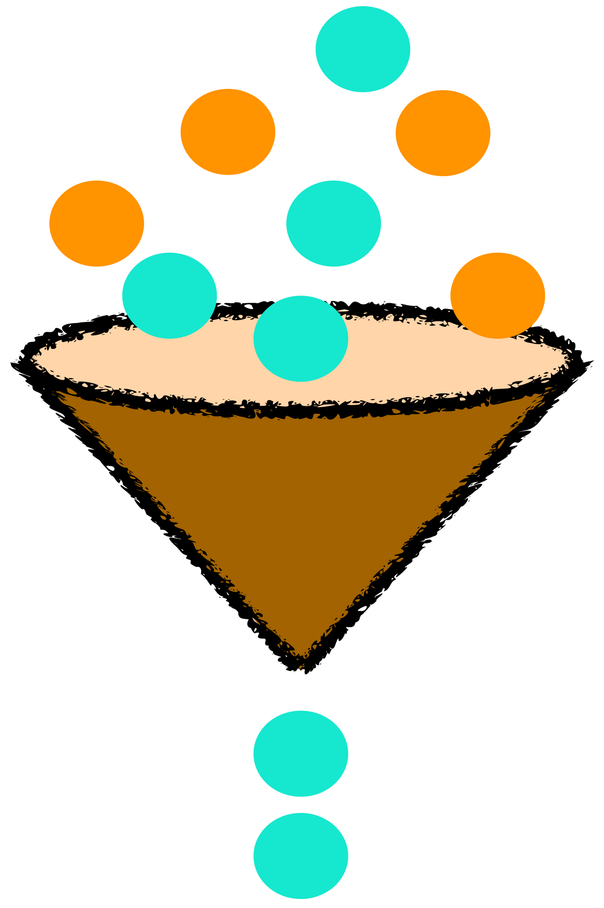
What makes a good feature representation for adversarial filtering? How to reduce model dependence?



Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

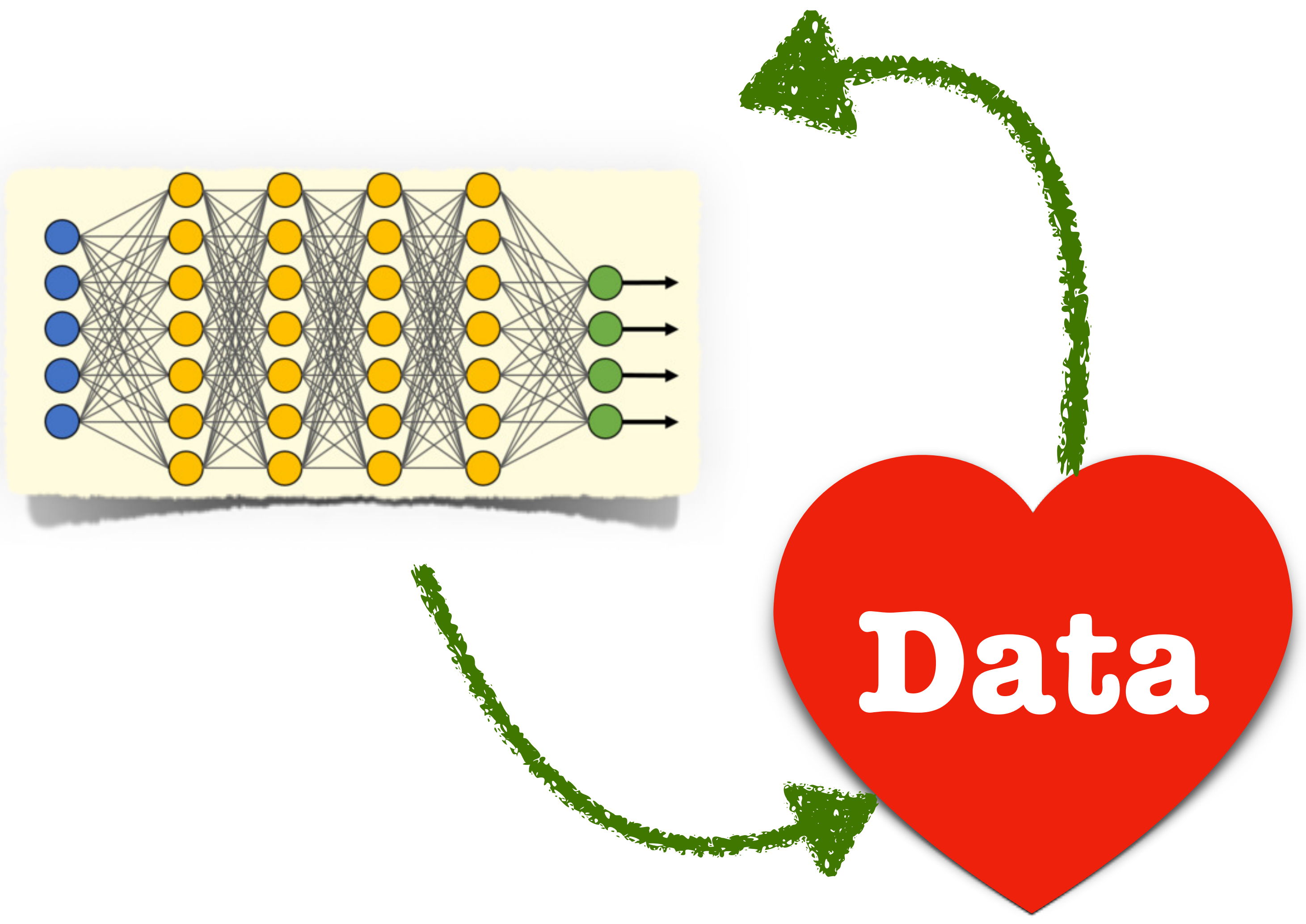
Takeaways

Performance on some large-scale datasets could be misleading.



English-only #BenderRule

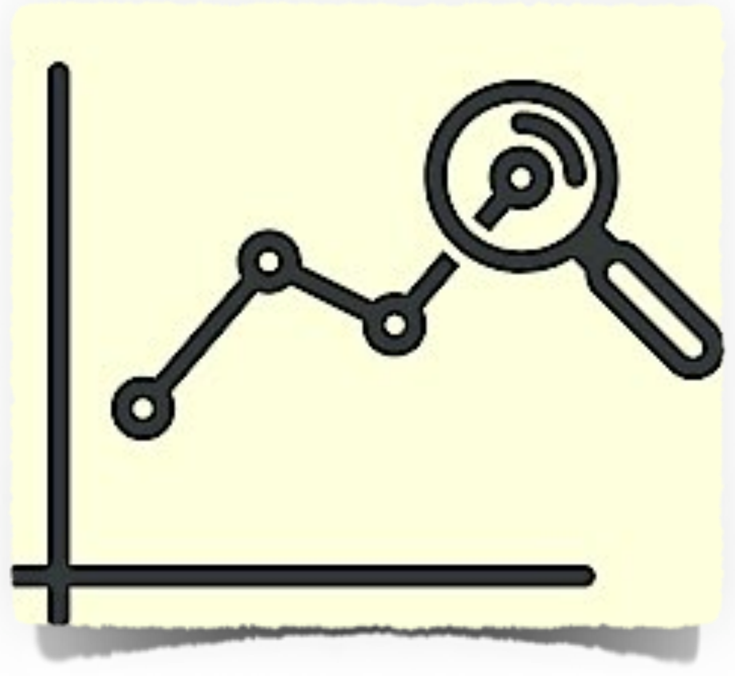
What makes a good feature representation for adversarial filtering? How to reduce model dependence?



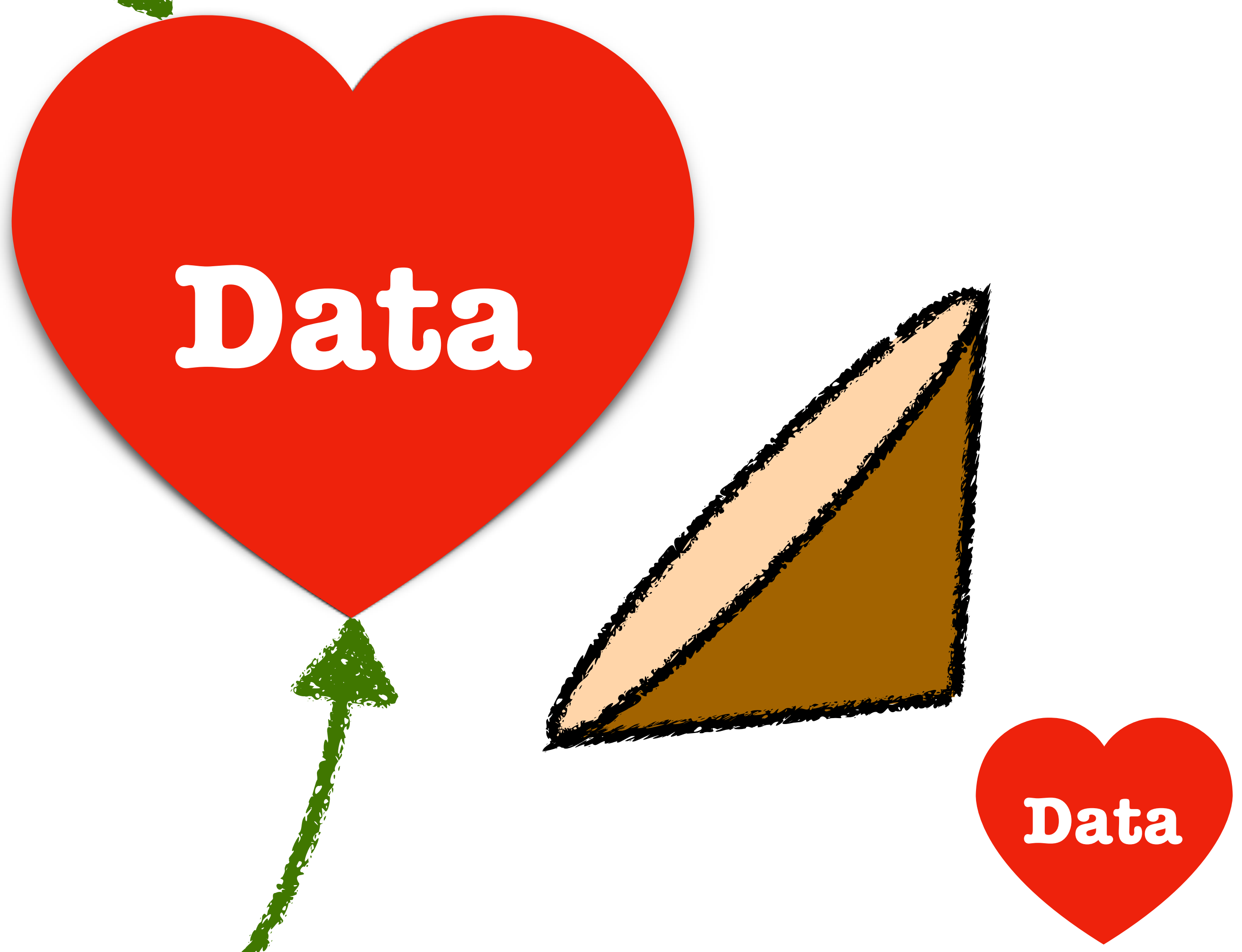
Looking forward: Can we reuse filtered out data?

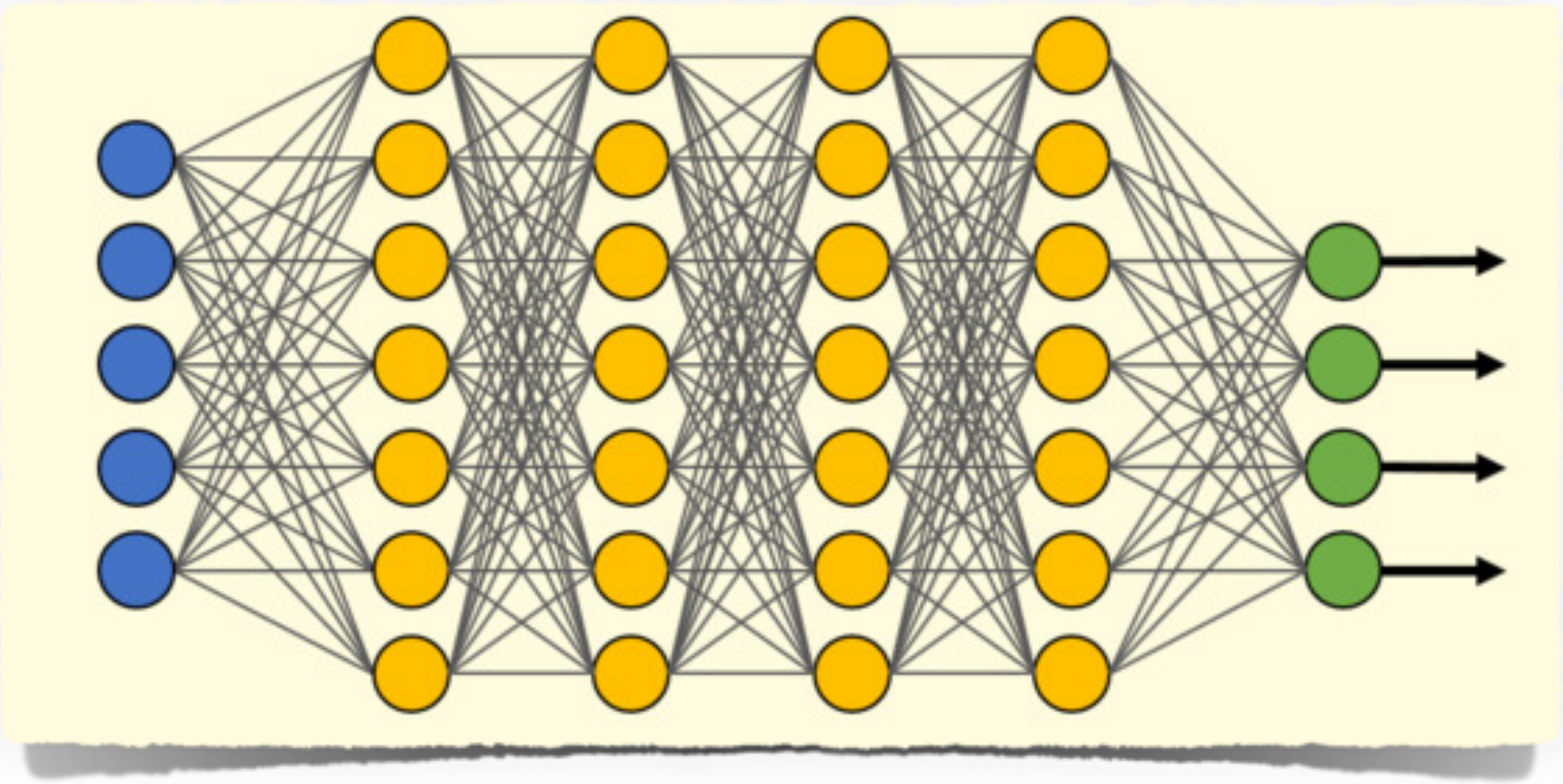
Adversarial Filtering of Dataset Biases [L., Swayamdipta, B., Z., S., P. & C. (in submission)]

In Summary



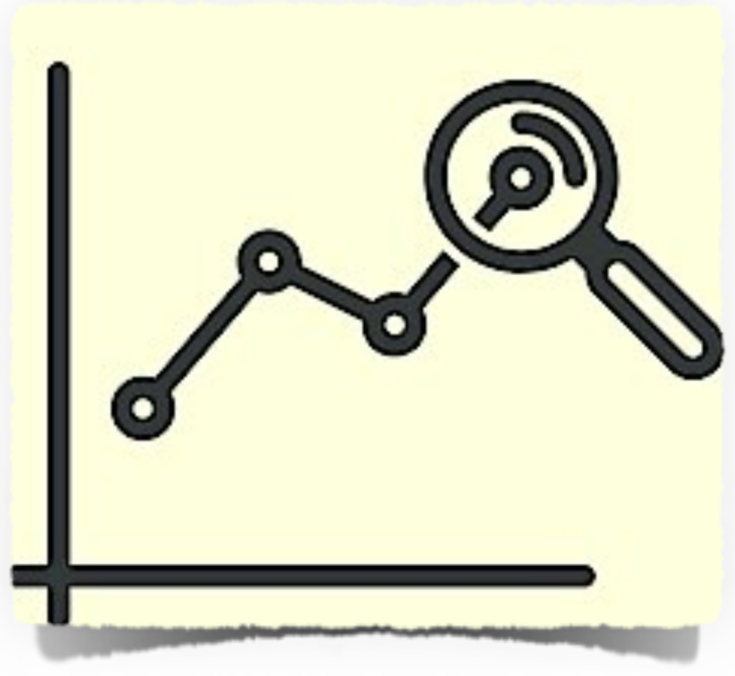
What do predictions tell us about the data?



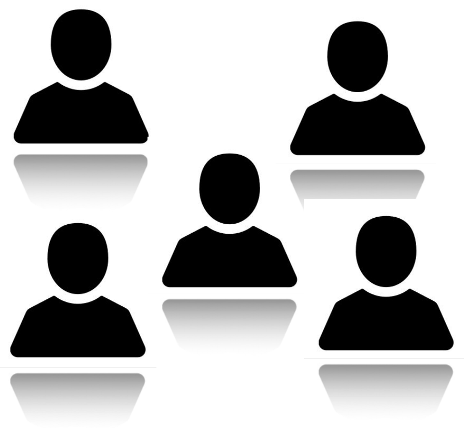


How can we use this information to spruce up our datasets?

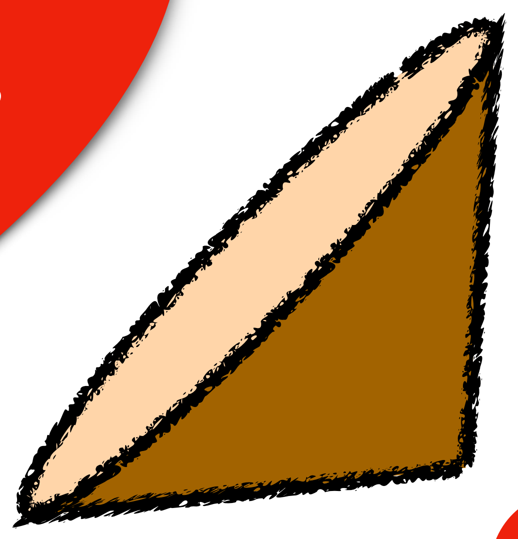
In Summary

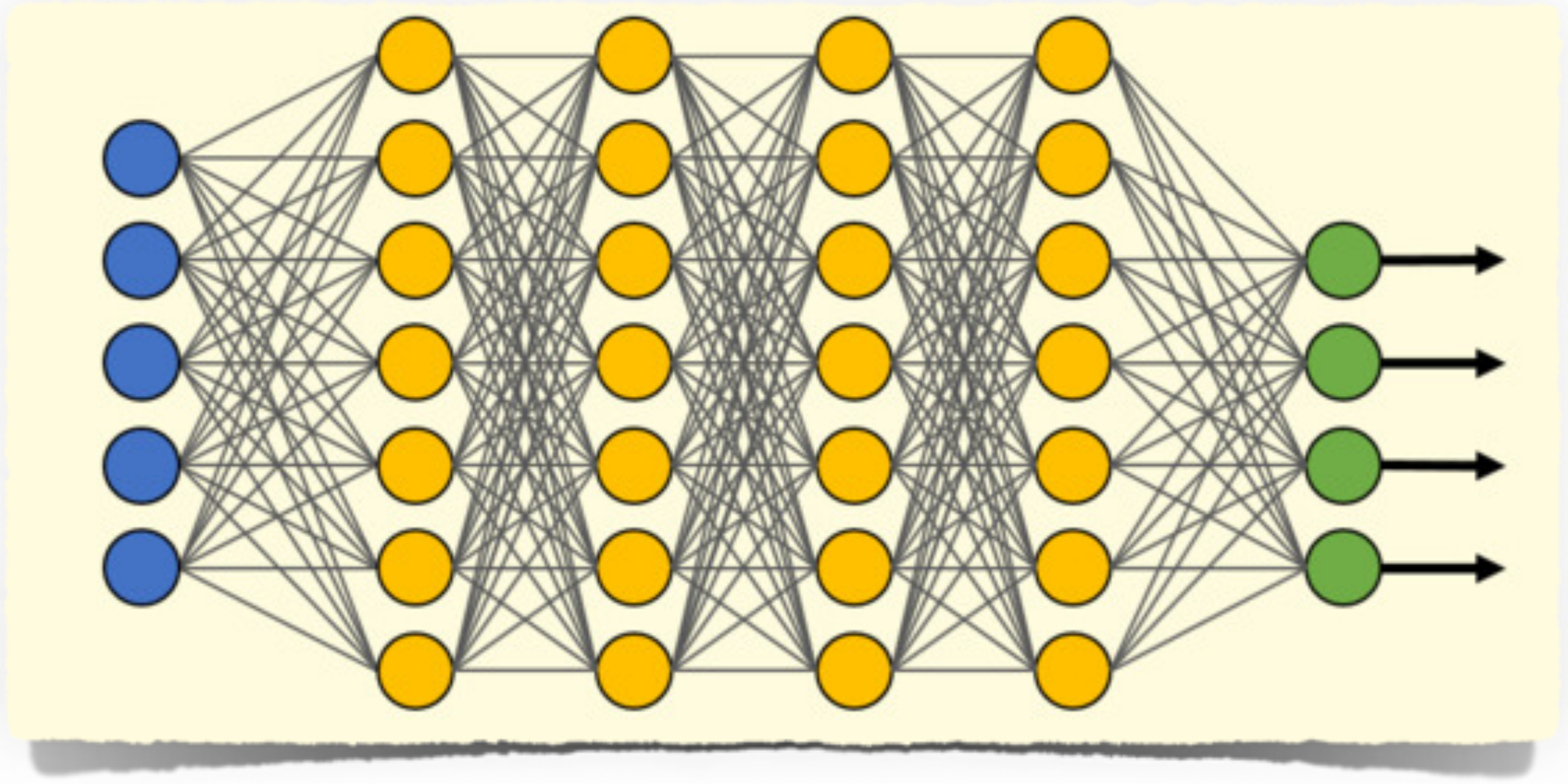


What do predictions tell us about the data?



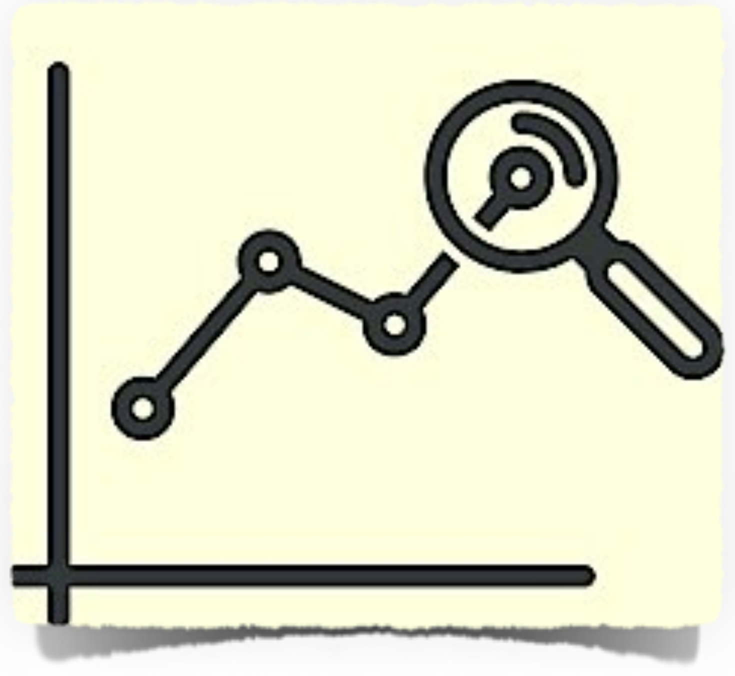
Annotation Artifacts Abound!



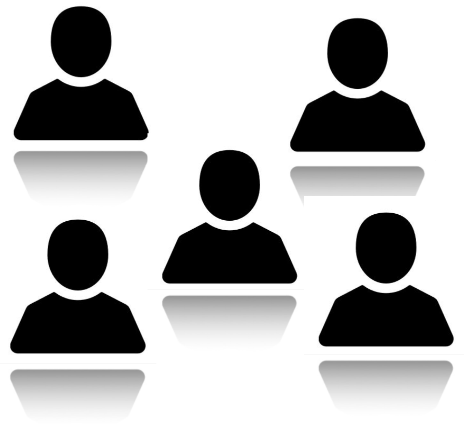


How can we use this information to spruce up our datasets?

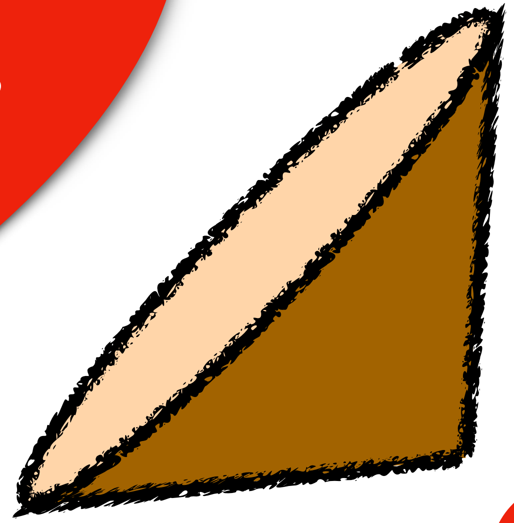
In Summary



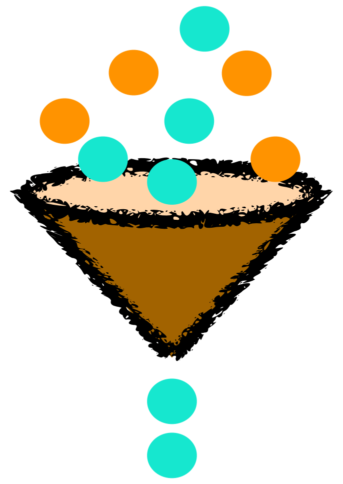
What do predictions tell us about the data?

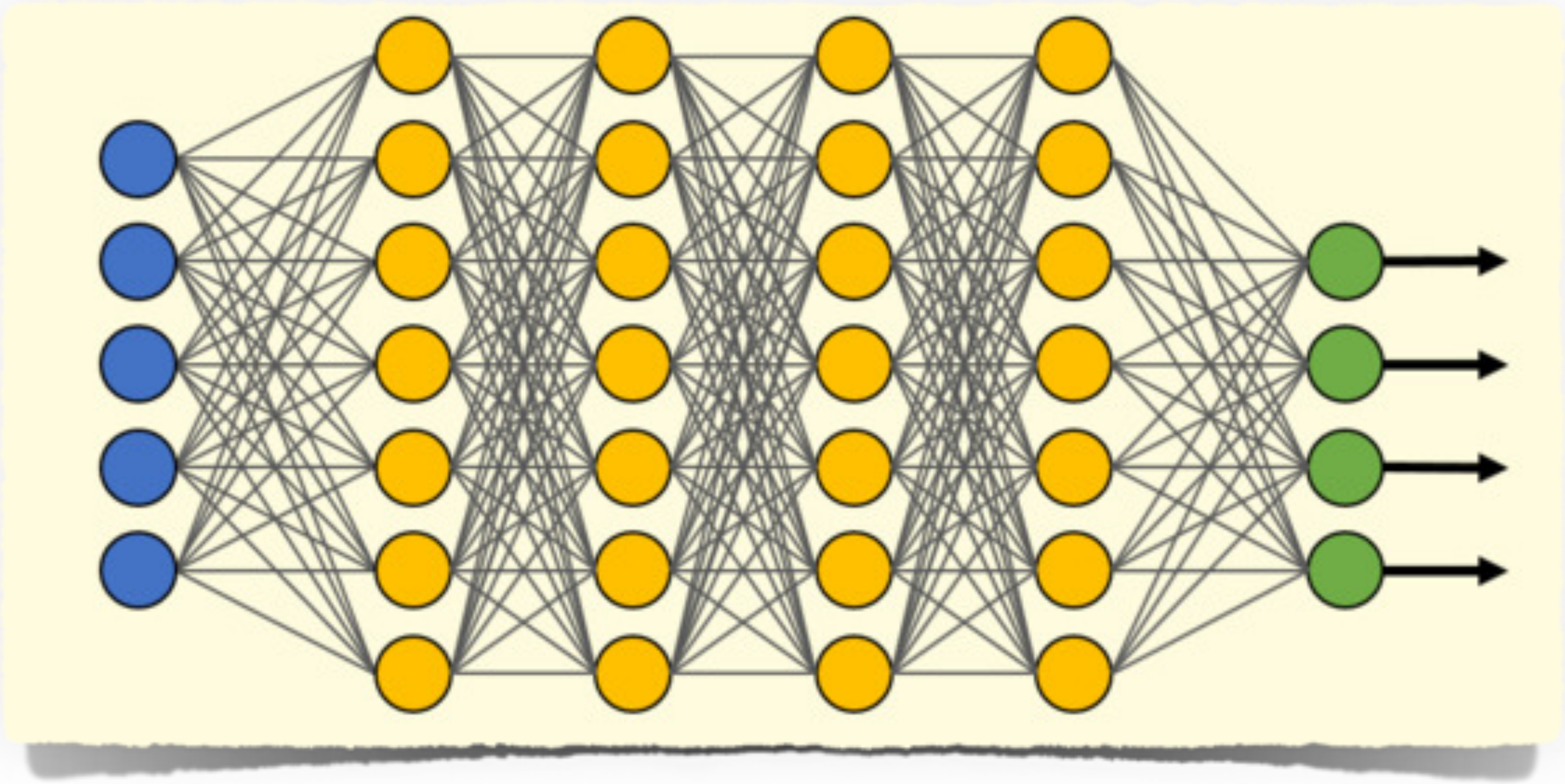


Annotation Artifacts Abound!



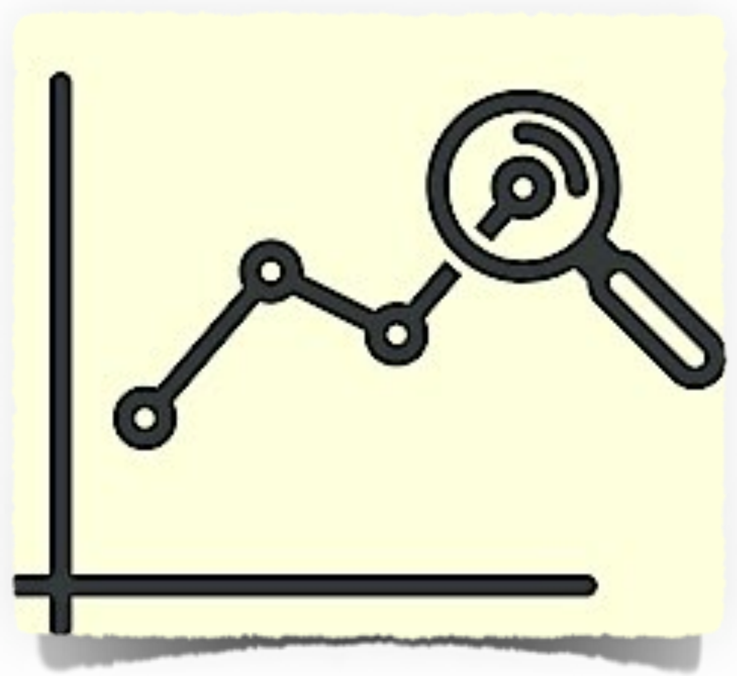
Lightweight Adversarial Filtering



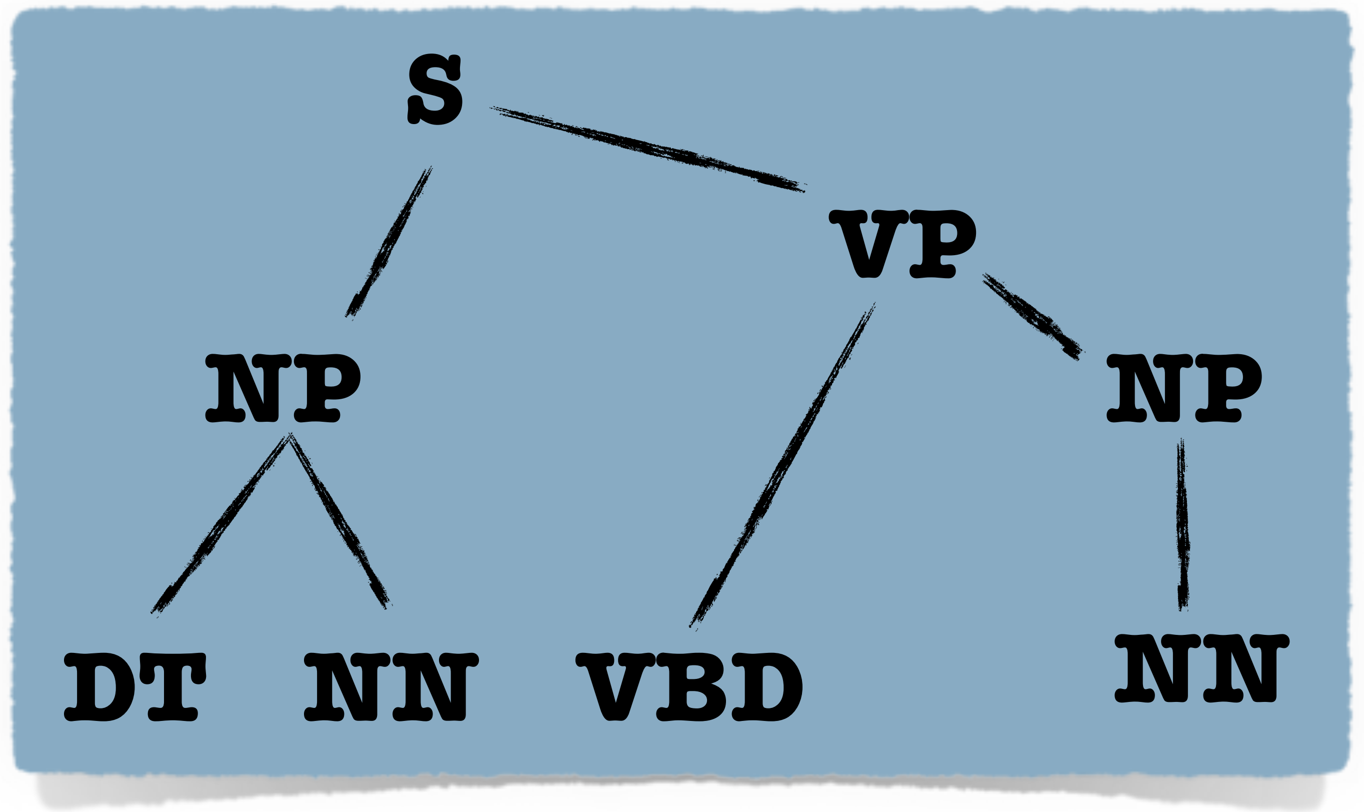


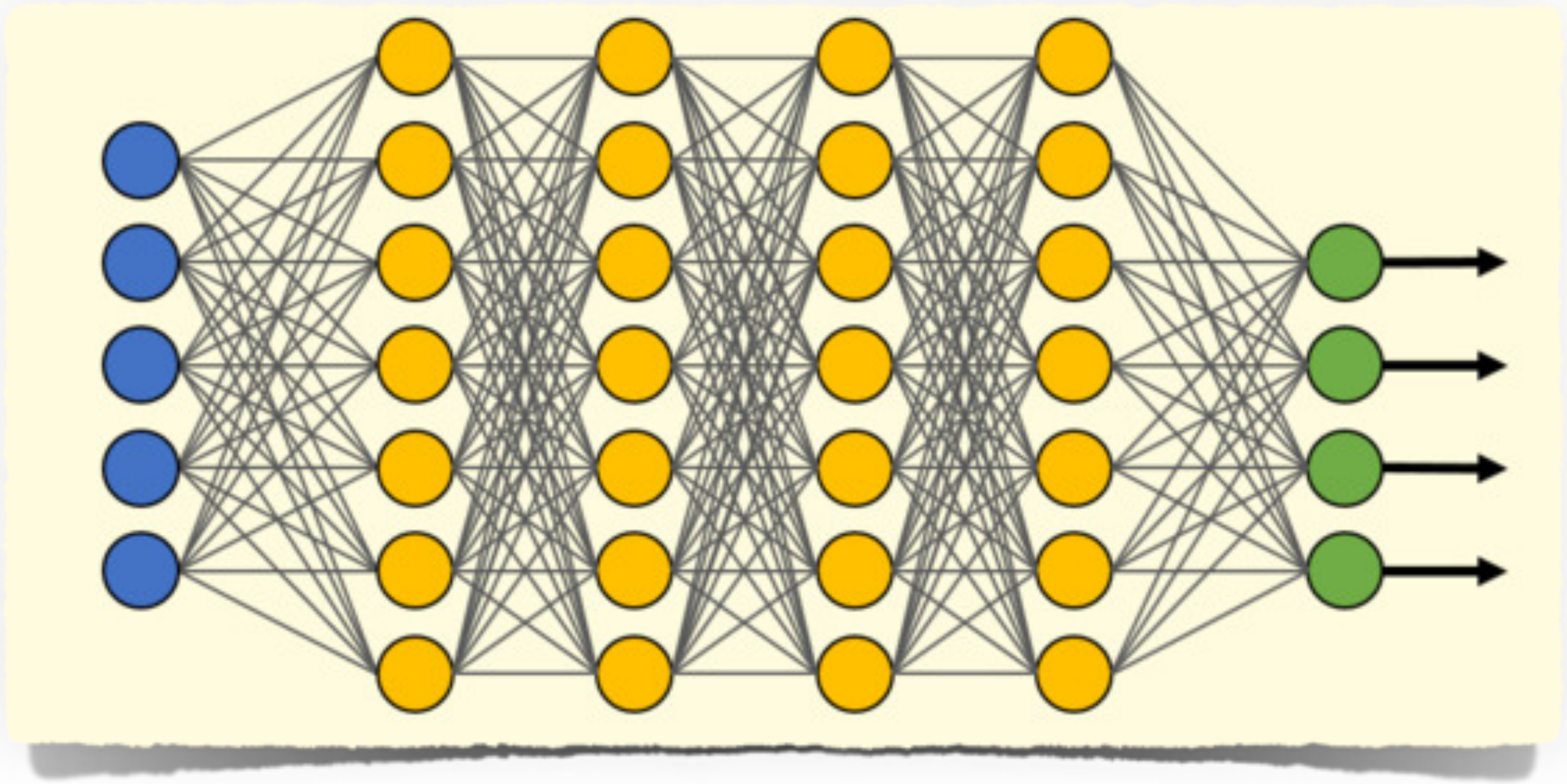
How can we use this information to spruce up our datasets?

Machine Learning



NLP

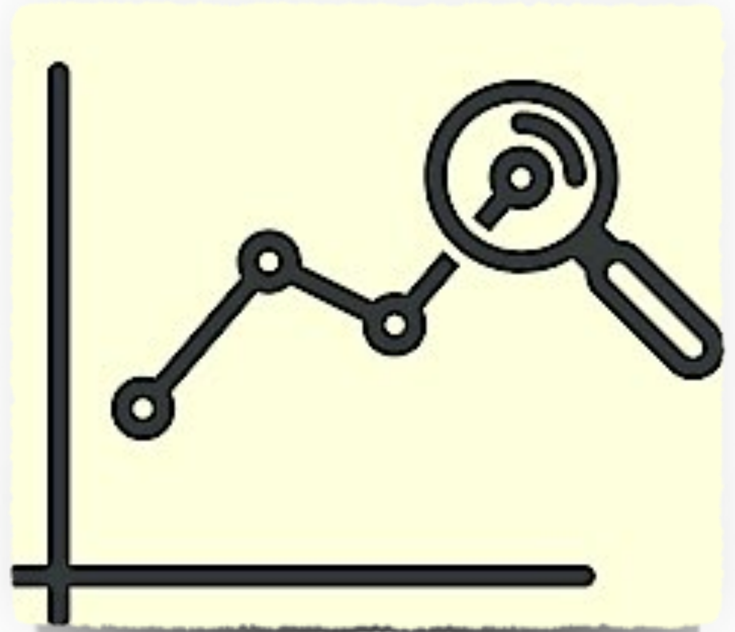


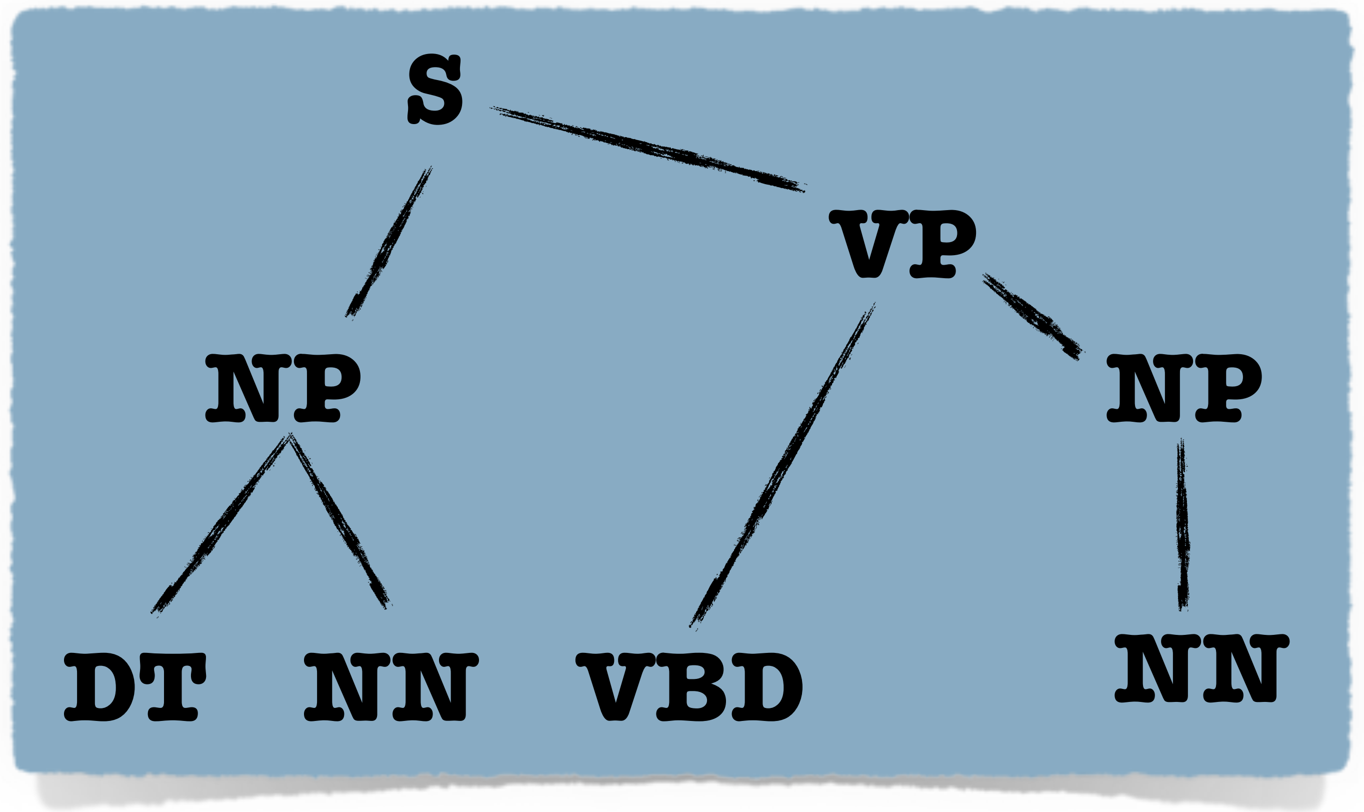


Linguistics

Machine Learning

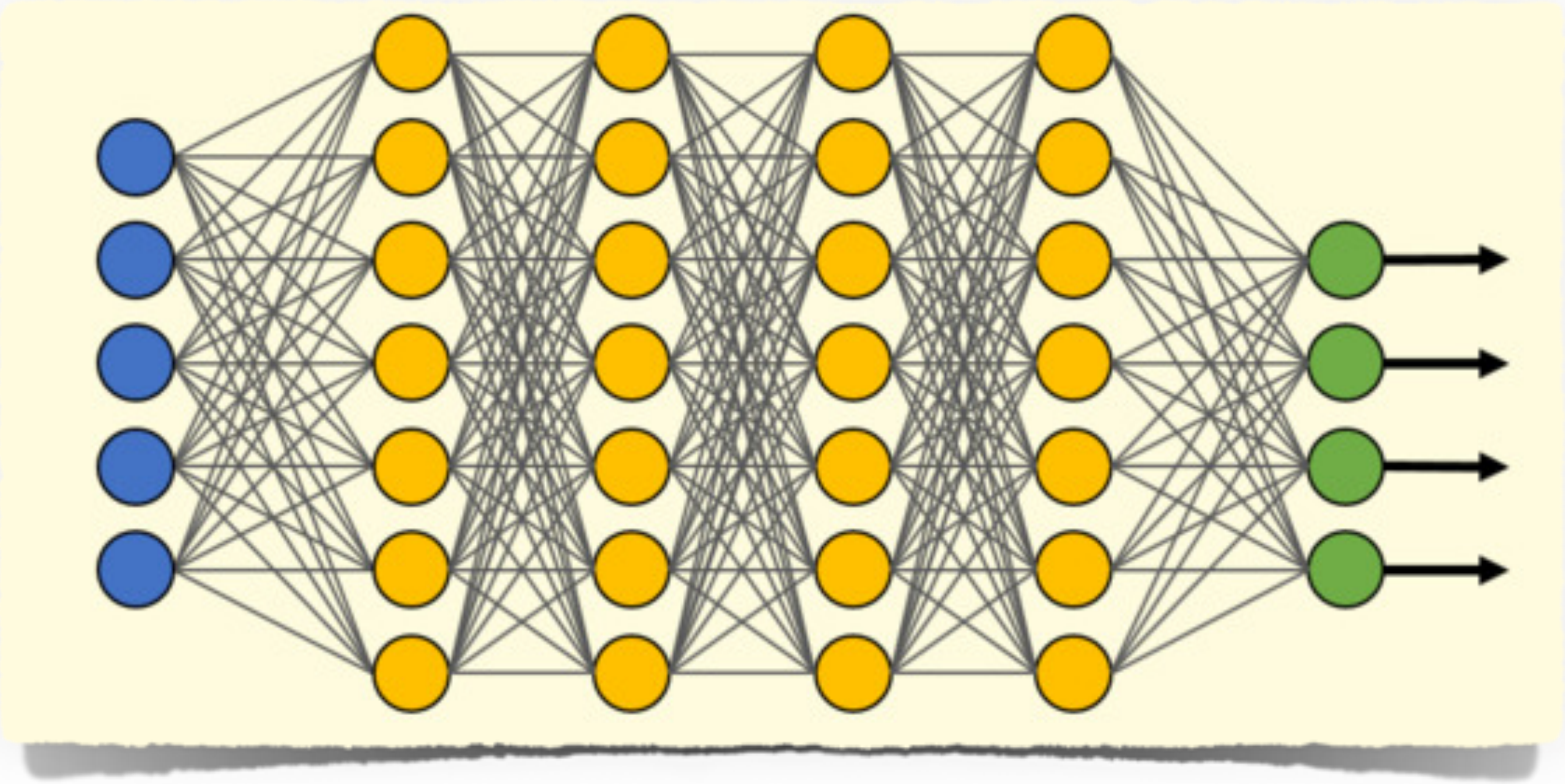
What does our knowledge of language tell us about models?





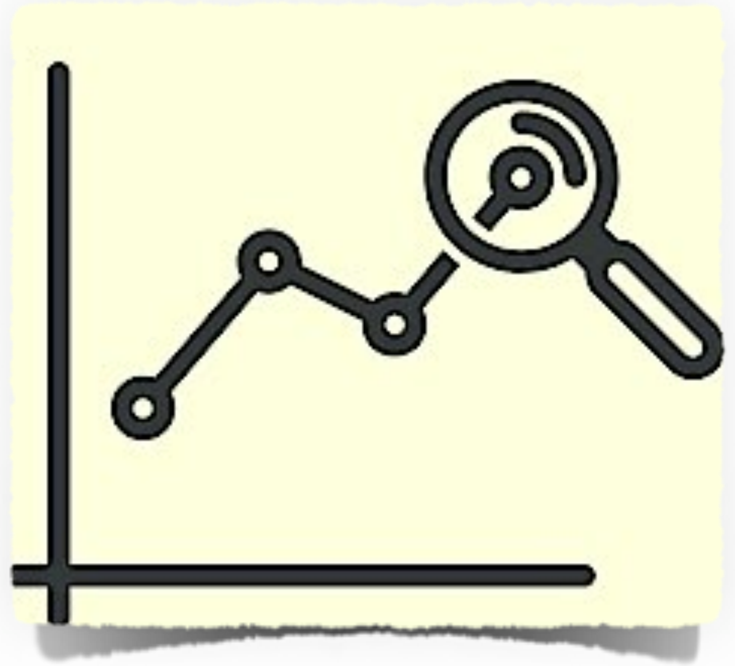
NLP

Data

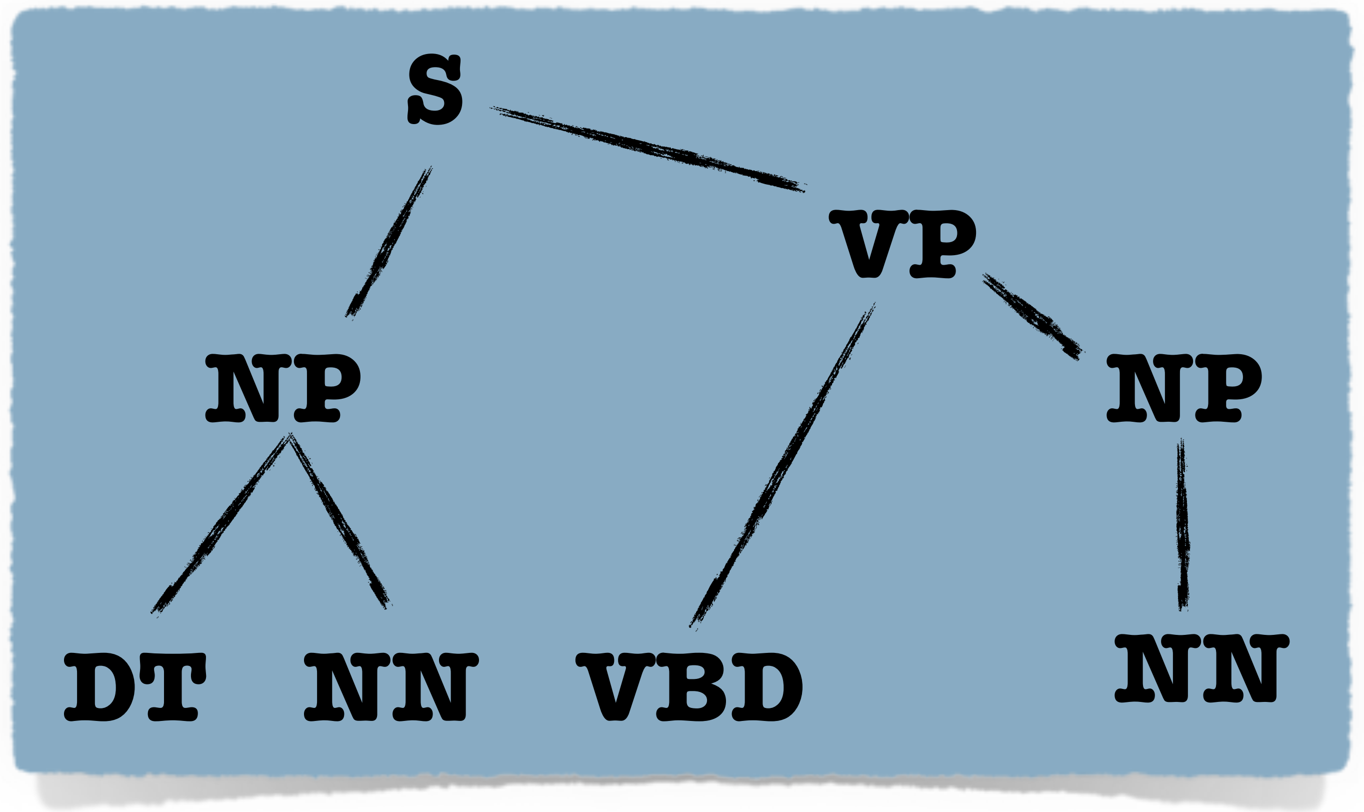


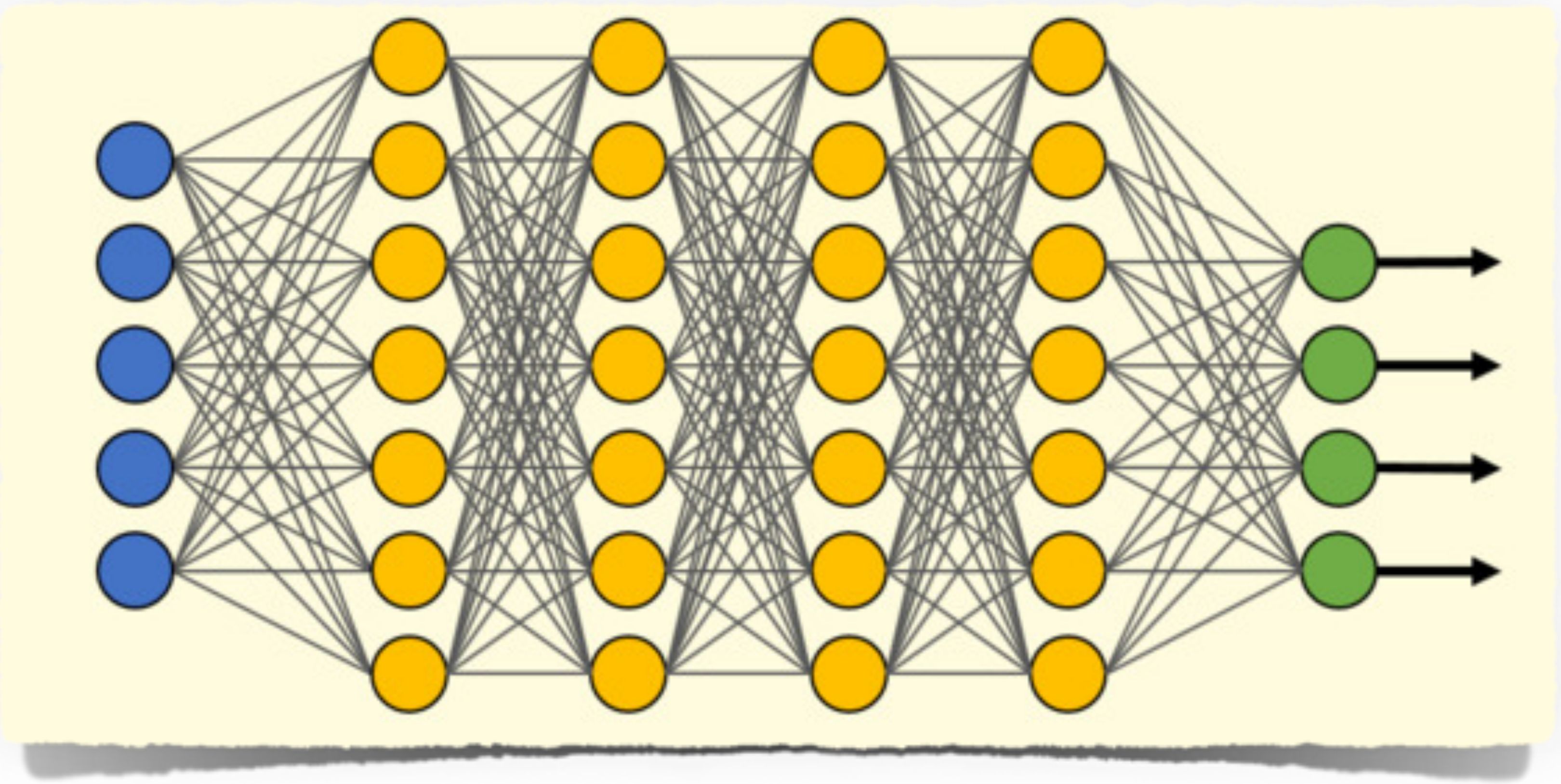
Linguistics

Machine Learning



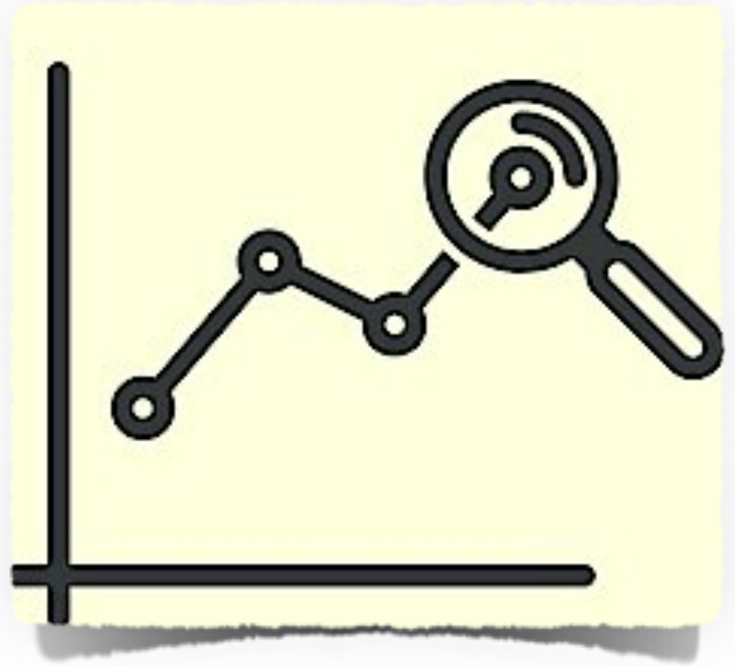
NLP



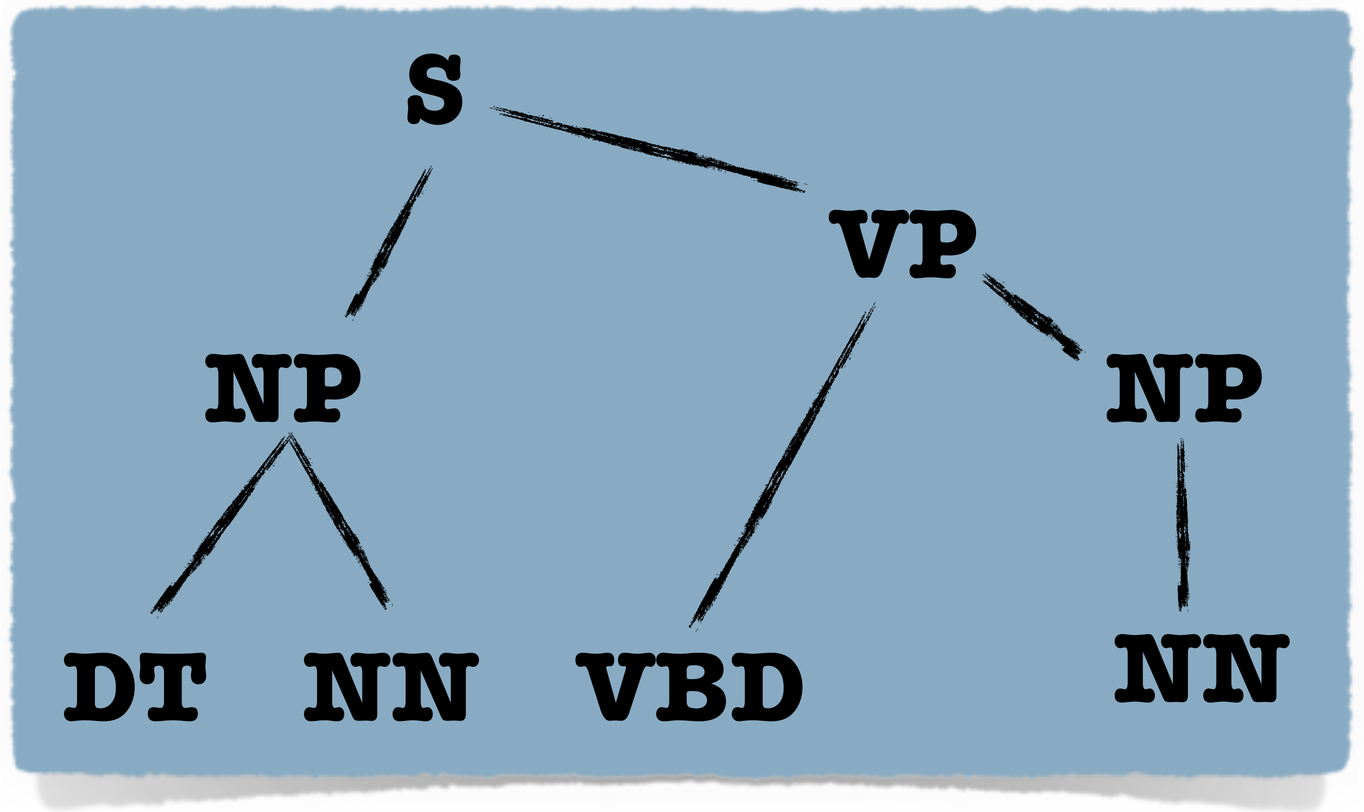


Linguistics

Machine Learning

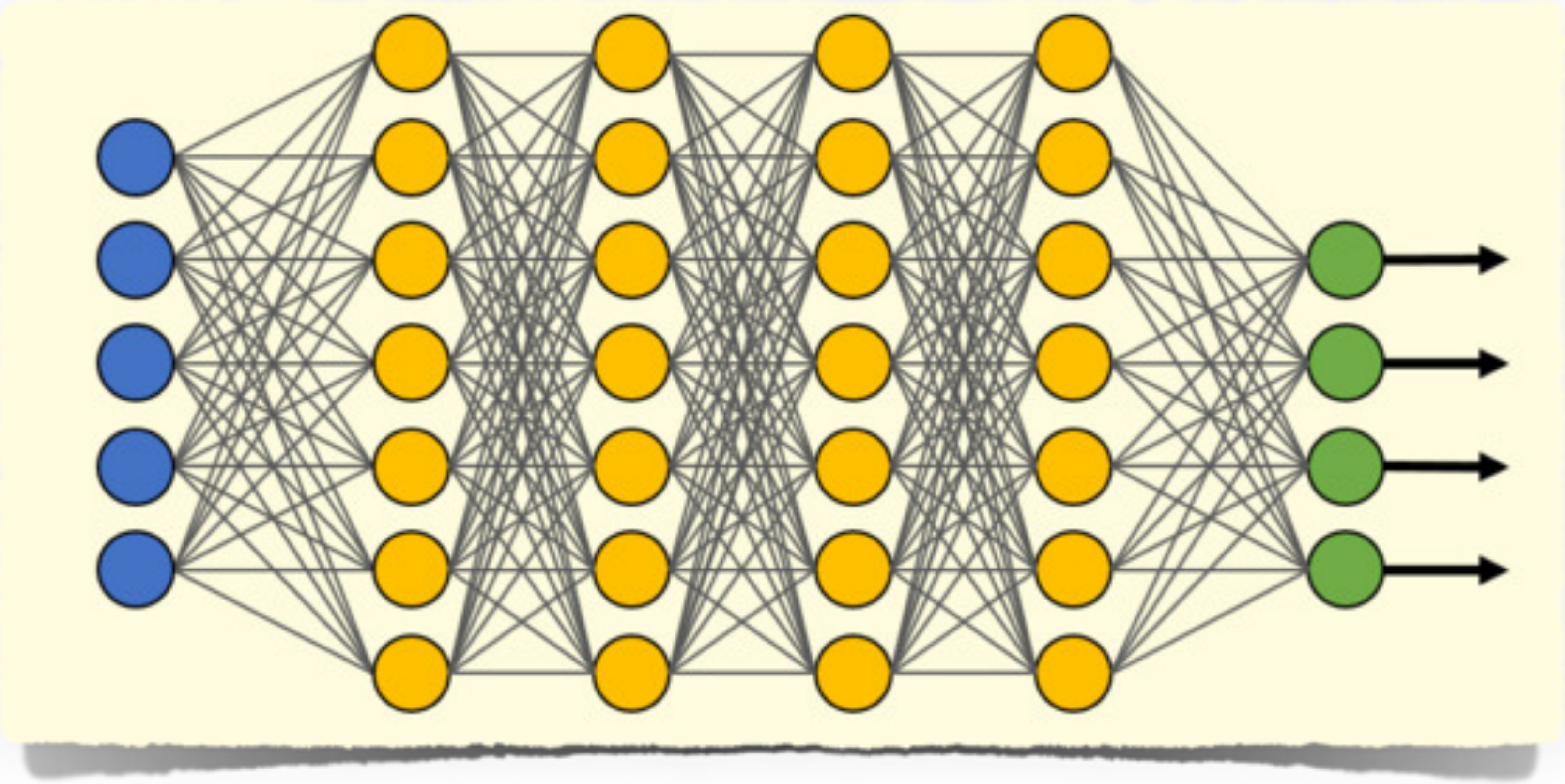


NLP



Data

Can we improve performance by accounting for linguistic structure?



Linguistics

Sam
Bowman

Suchin
Gururangan

Matt
Peters

Noah A.
Smith

Chandra
Bhagavatula

Ronan
LeBras

Ashish
Sabharwal

Rowan
Zellers

Yejin
Choi

Omer
Levy

Roy
Schwartz

Thanks!

swabhs.com

swabhs

swabhz