Responsible AI:
Addressing Biases in Datasets and Models

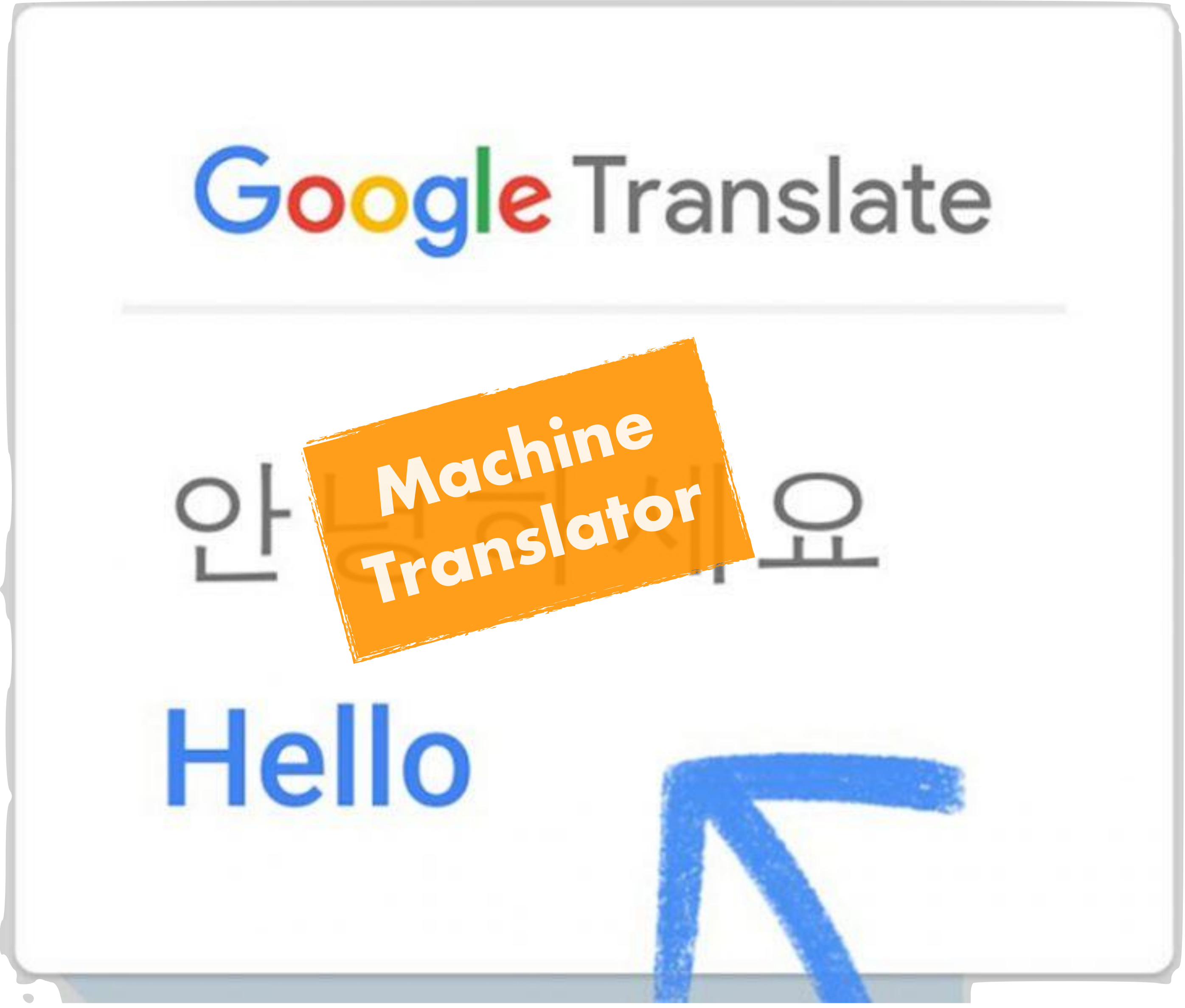
Swabha Swayamdipta
Postdoctoral Investigator, Allen Institute for AI
Nov 2nd, 2020

Responsible AI:
Addressing Biases in Datasets and Models

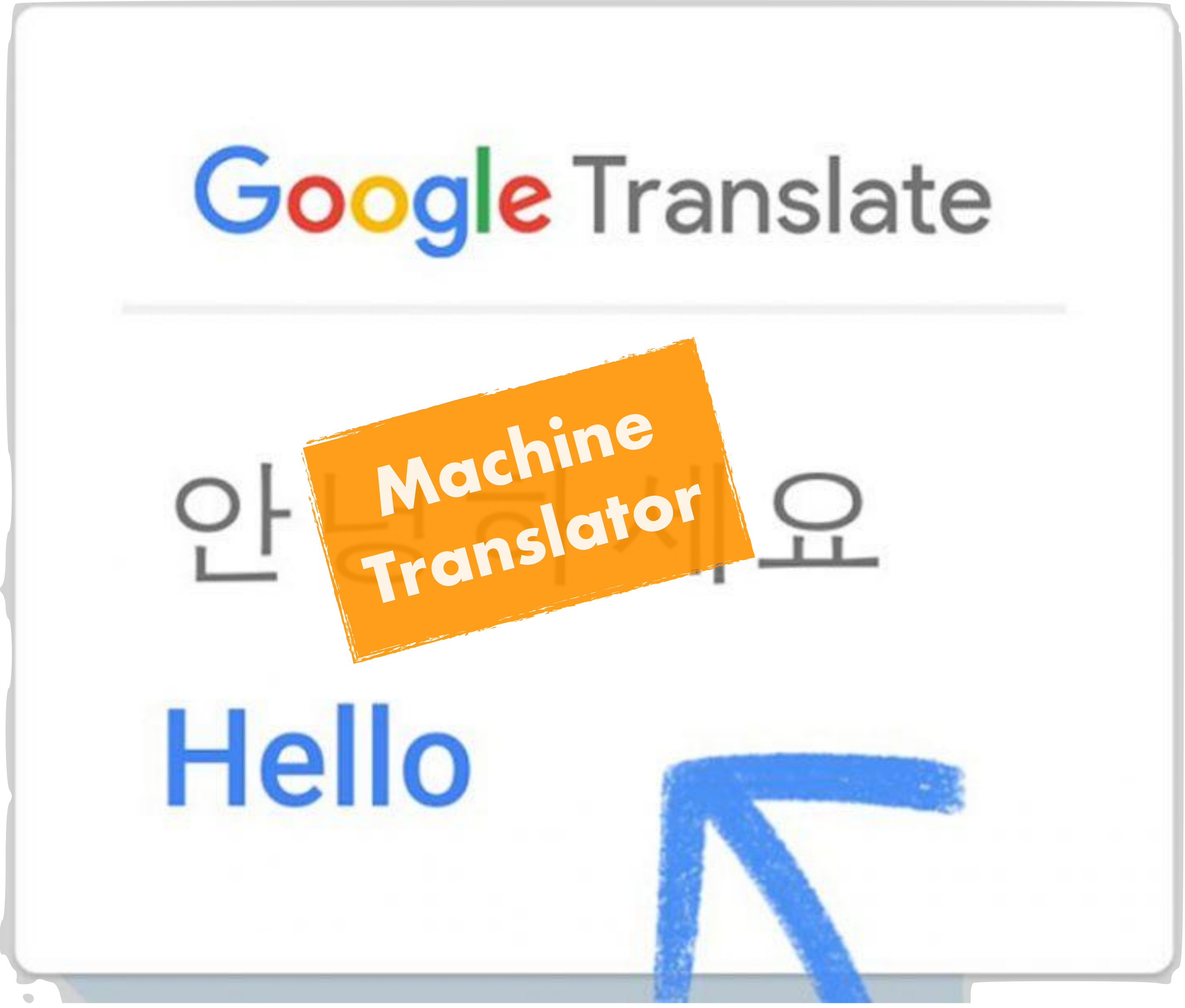
Swabha Swayamdipta
Postdoctoral Investigator, Allen Institute for AI
Nov 2nd, 2020


echo
"Alexa, turn it up."
Automated Assistants


echo
"Alexa, turn it up."
Automated Assistants
amazon music
npr
PHILIPS hue
ring


Google Translate
Machine Translator
Hello


echo
"Alexa, turn it up."
Automated Assistants


Google Translate
안녕하세요
Machine Translator
Hello


Recommender Systerms
NETFLIX


echo
"Alexa, turn it up."
Automated Assistants


Google Translate
Machine Translator
Hello


Social Media

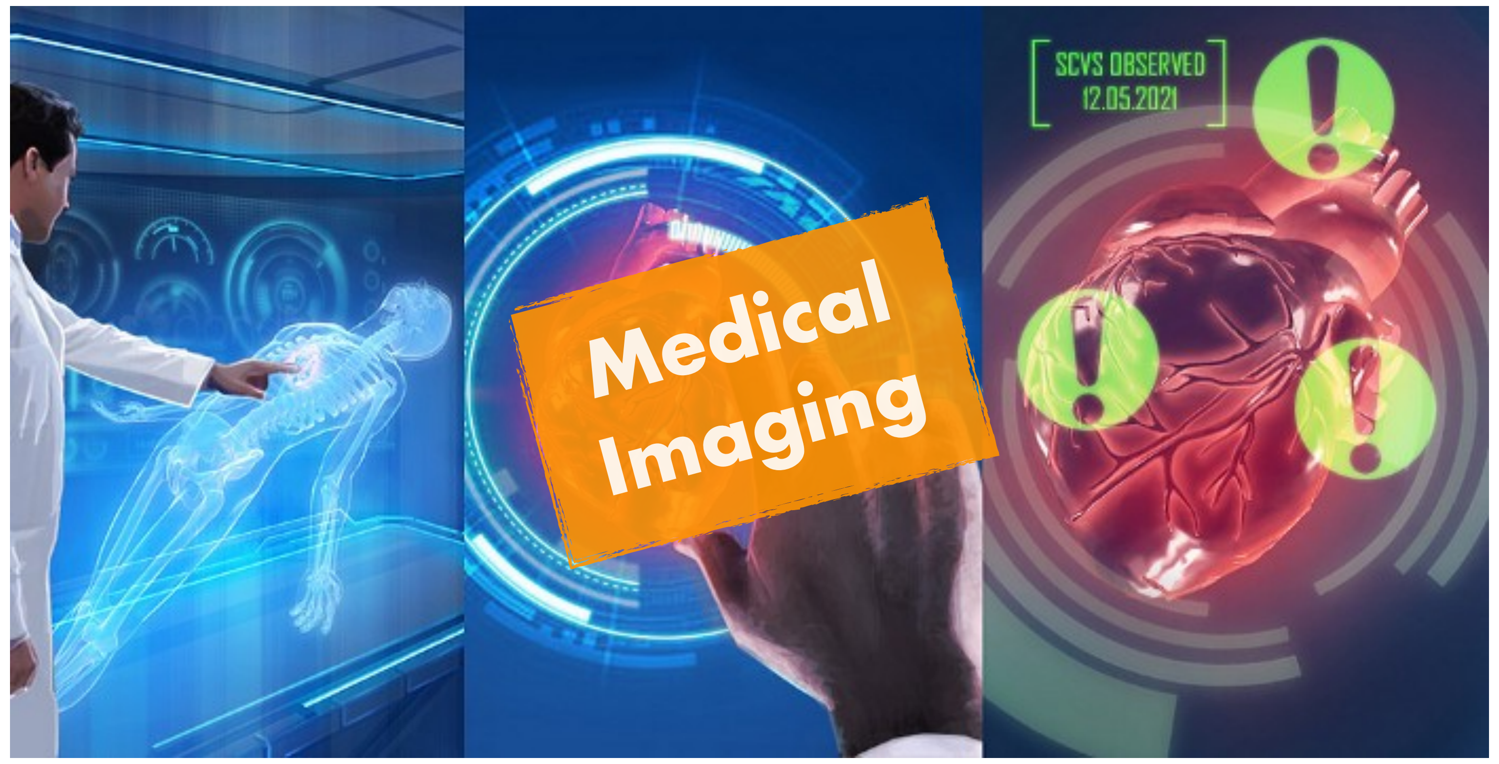

Recommender Systerms
NETFLIX


echo
"Alexa, turn it up."
Automated Assistants


Google Translate
Machine Translator
Hello


LinkedIn
foursquare
LIVEJOURNAL
vimeo
YouTube
Social Media
Blogger
twitter
flickr
Behance
tumblr.
yelp
Like


Recommender Systems
NETFLIX

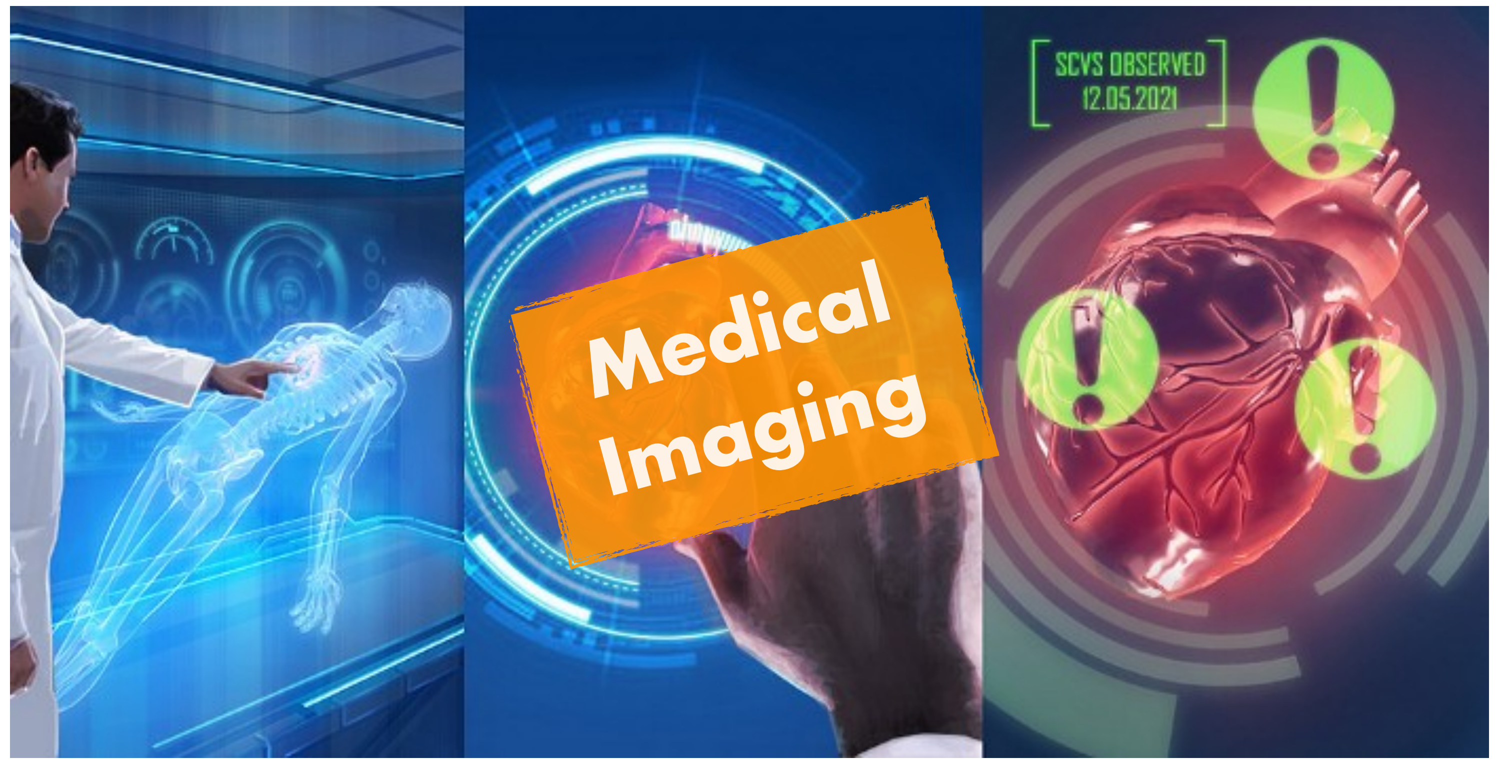

Medical Imaging


echo
"Alexa, turn it up."
Automated Assistants


Google Translate
Machine Translator
Hello


Social Media

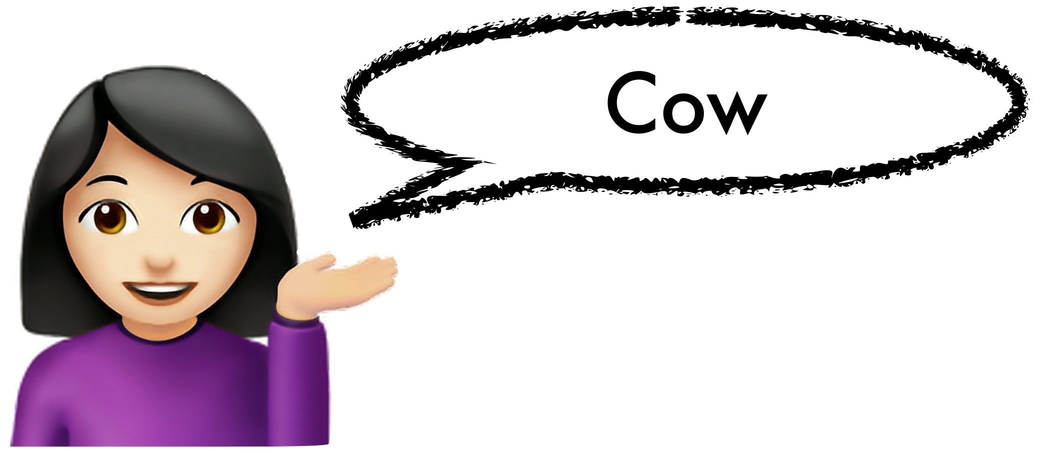

Recommender Systems
NETFLIX


Medical Imaging

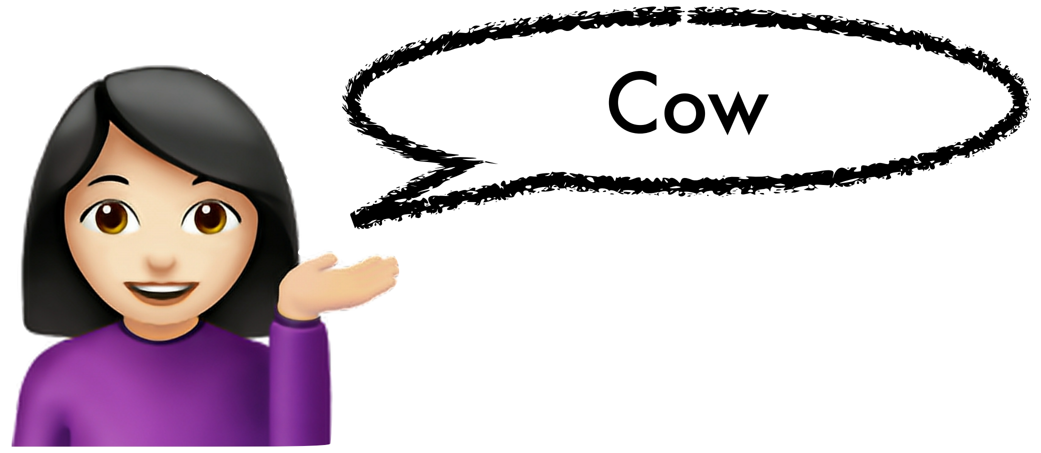

Self-Driving Cars

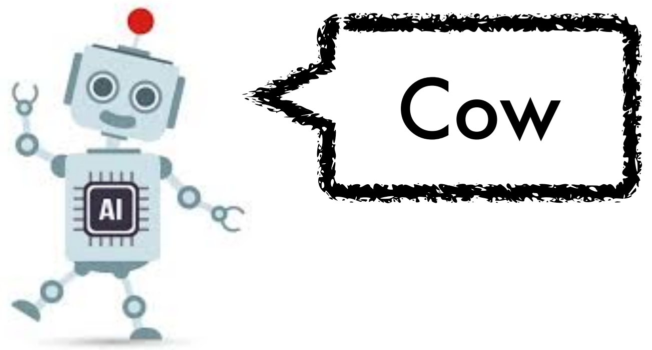

"WITH ARTIFICIAL INTELLIGENCE WE ARE SUMMONING THE DEMON."
-ELON MUSK
HUFF POST
"The development of full artificial intelligence could spell THE END OF THE HUMAN RACE."
-Stephen Hawking
HUFF POST

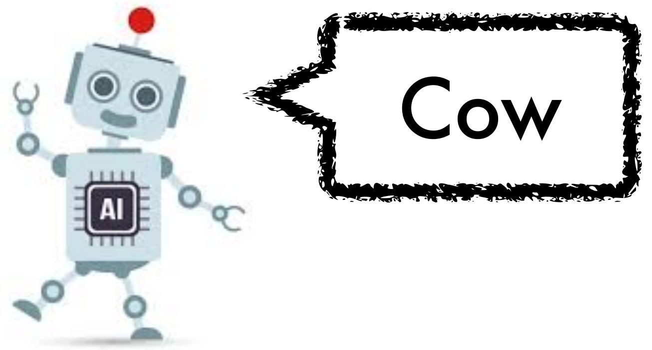
Example from Beery et al. [2019]



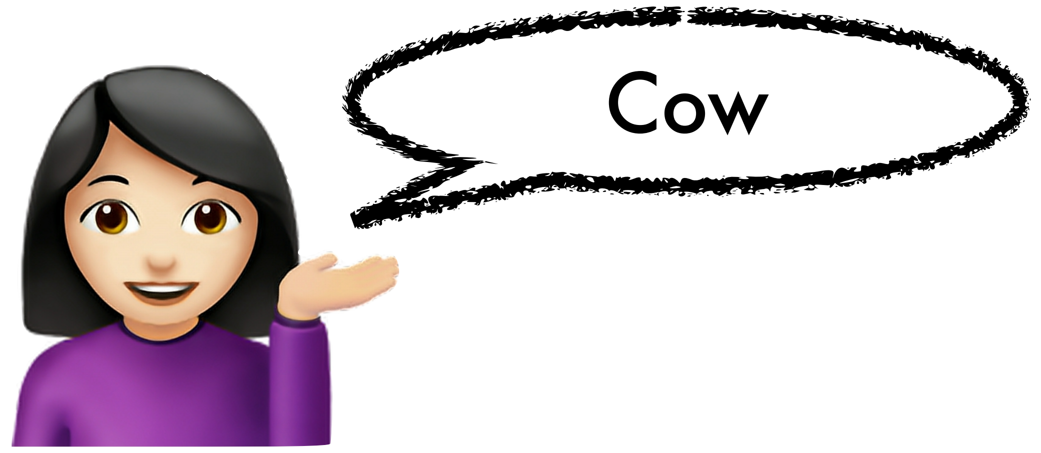
Example from Beery et al. [2019]

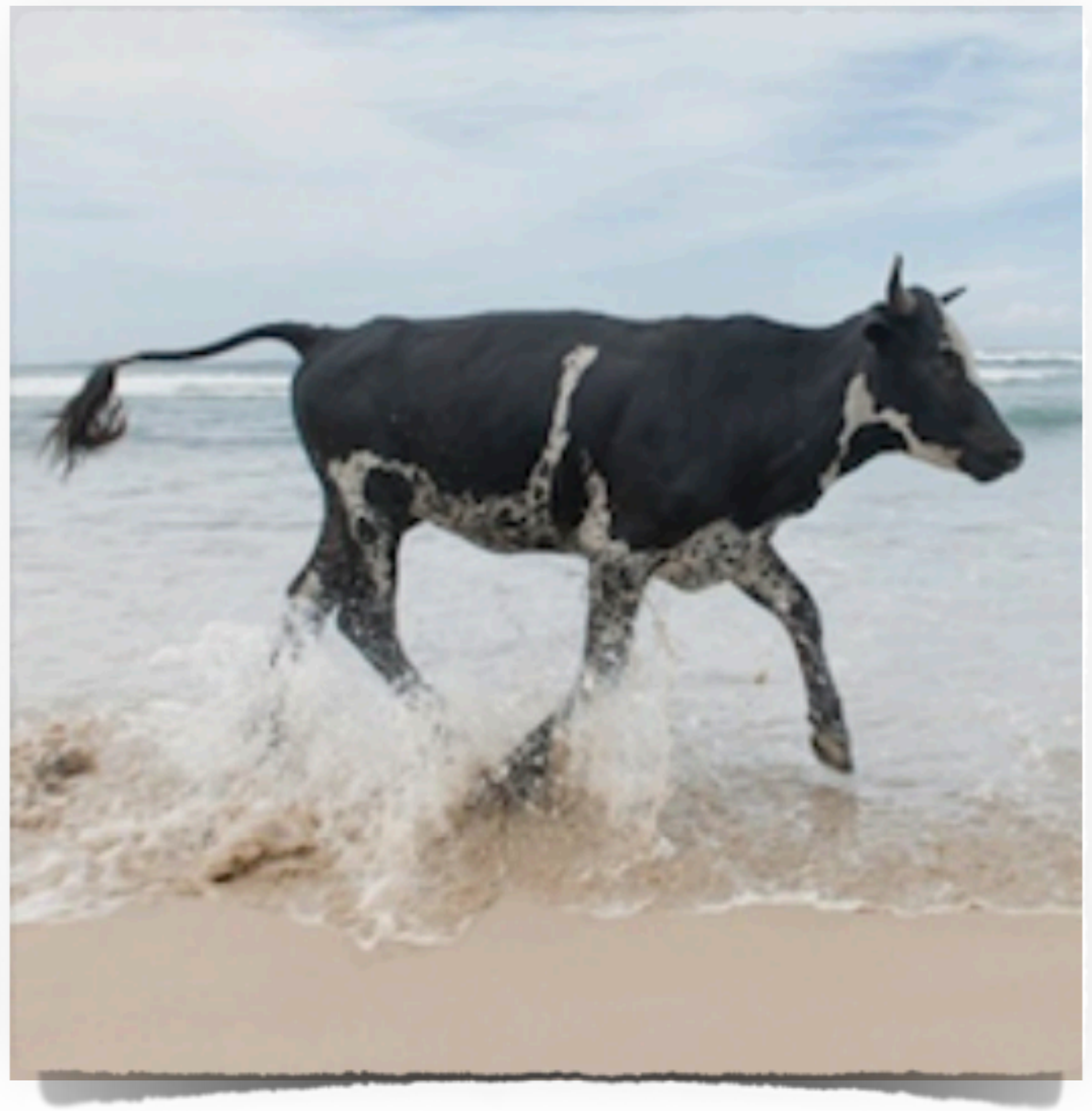


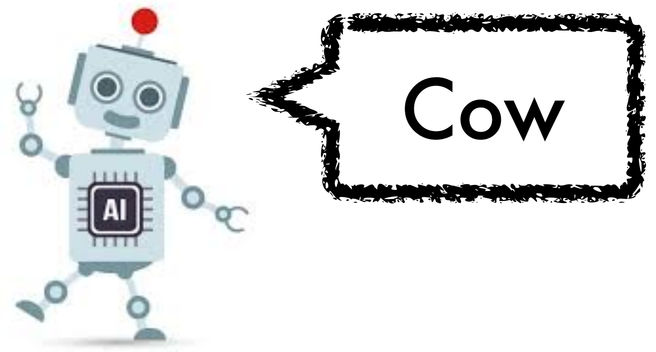


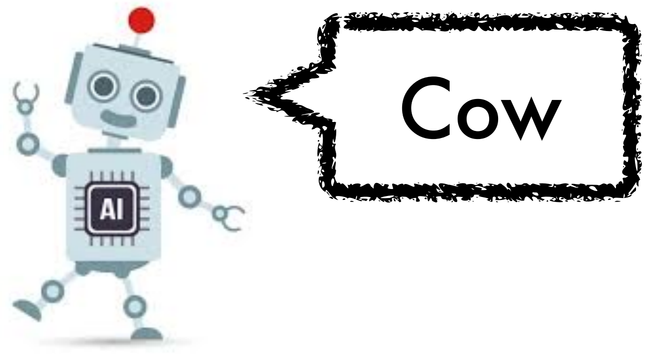


Example from Beery et al. [2019]

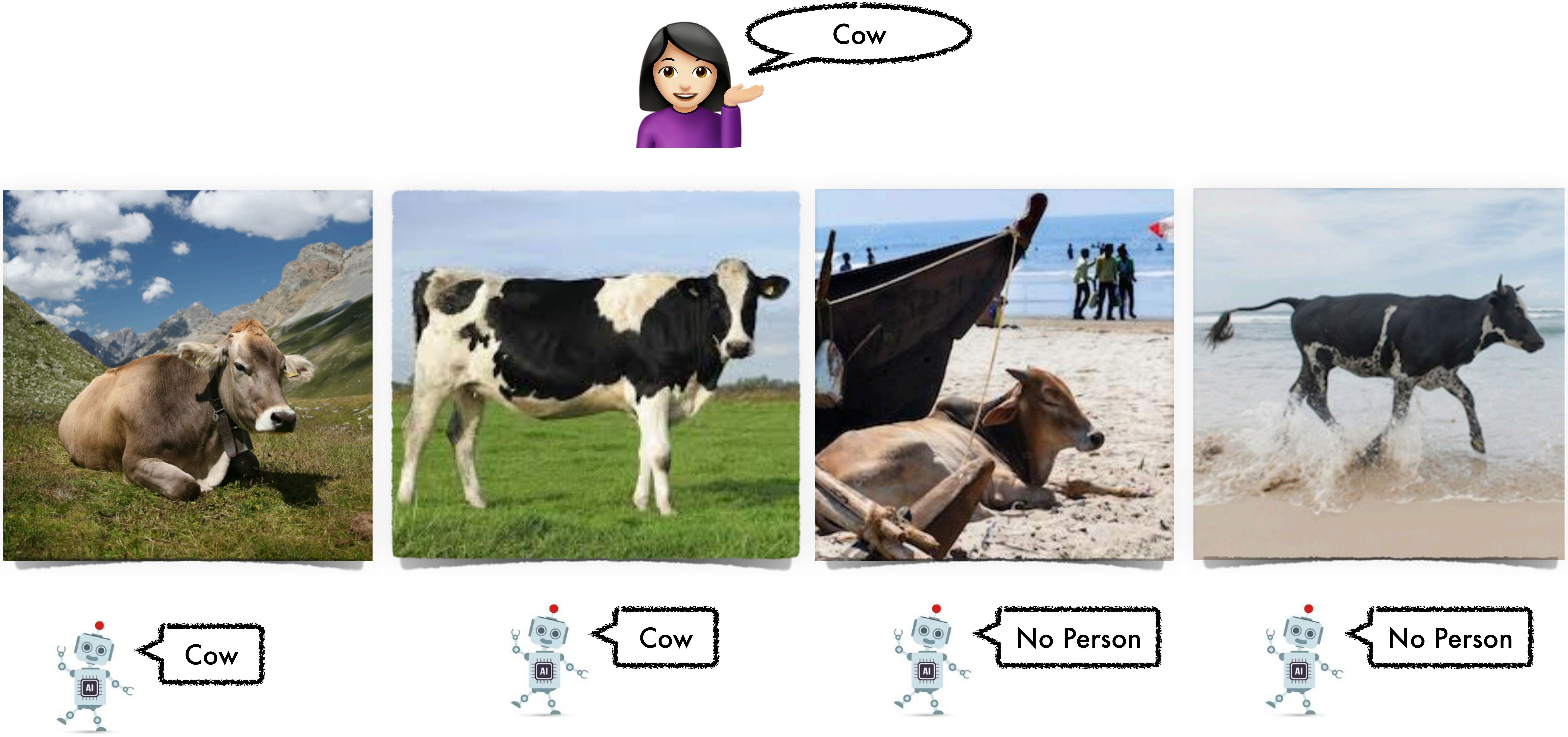




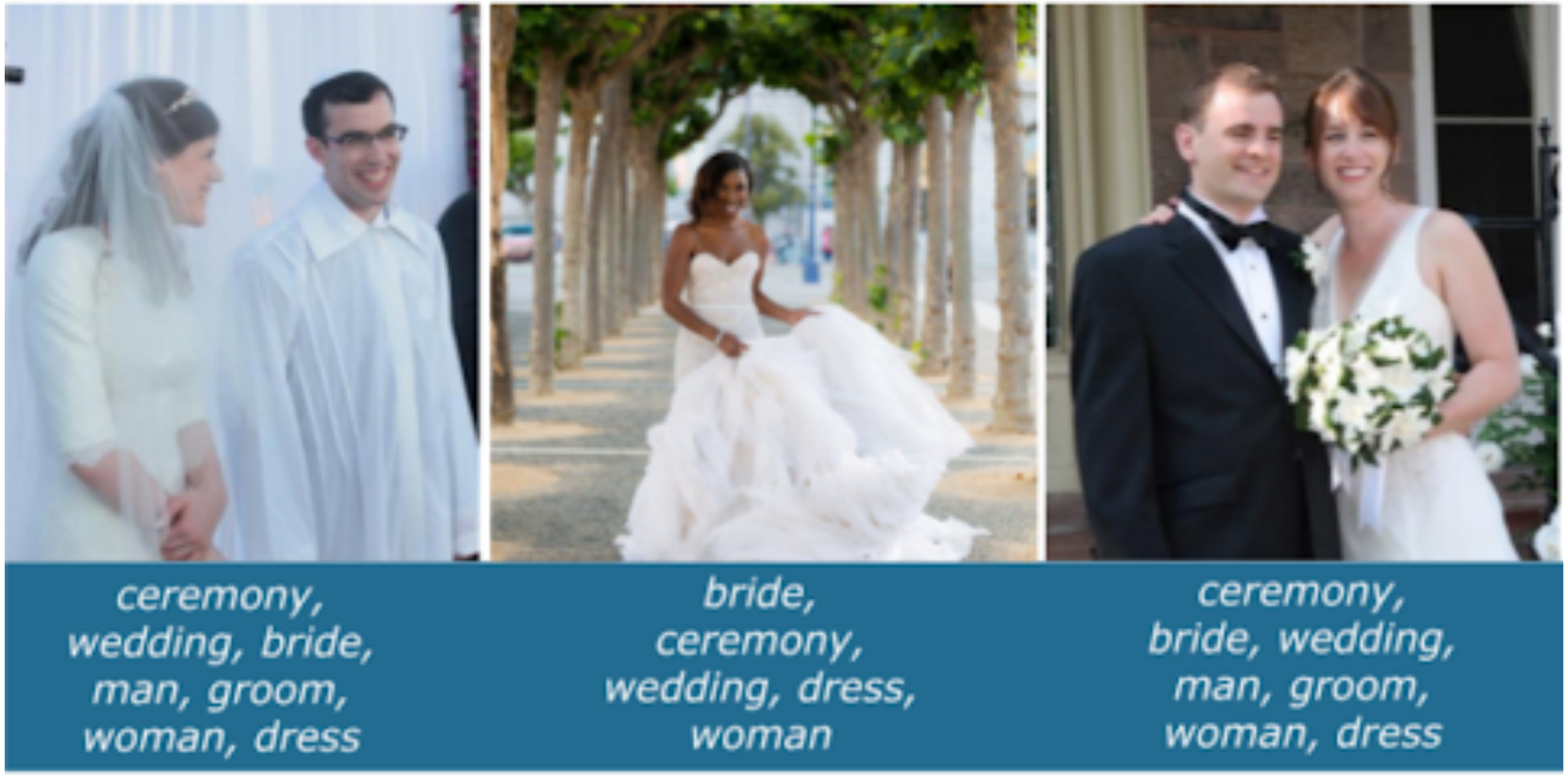


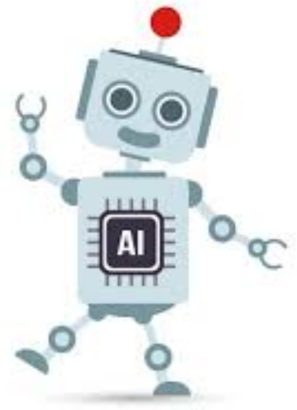


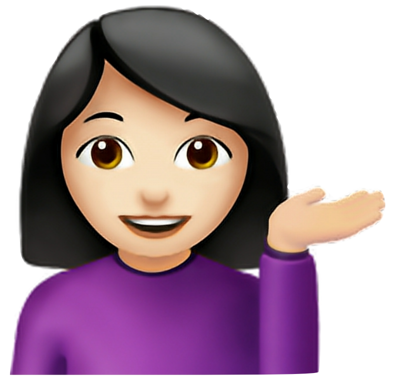
Example from Beery et al. [2019]



Example from Beery et al. [2019]

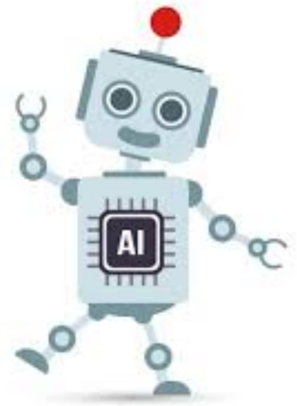






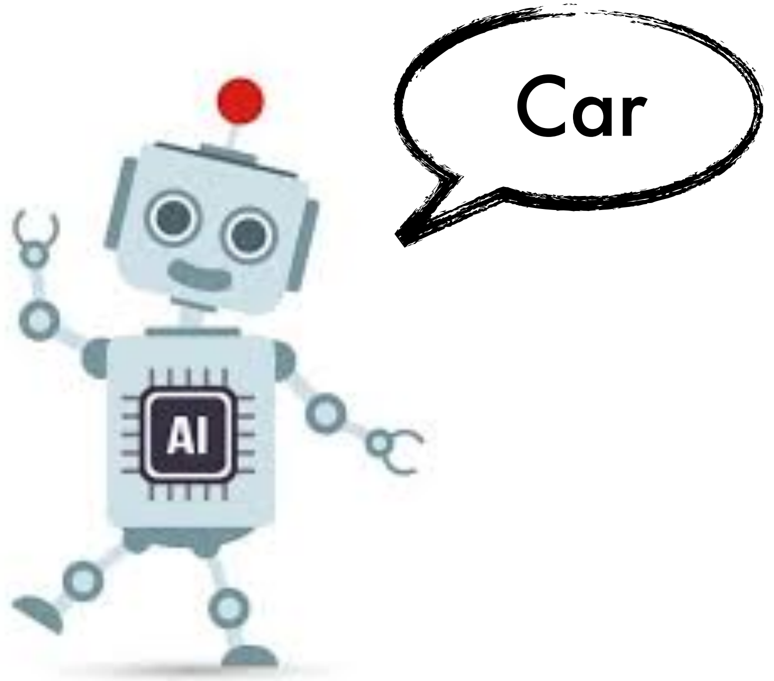
Wedding

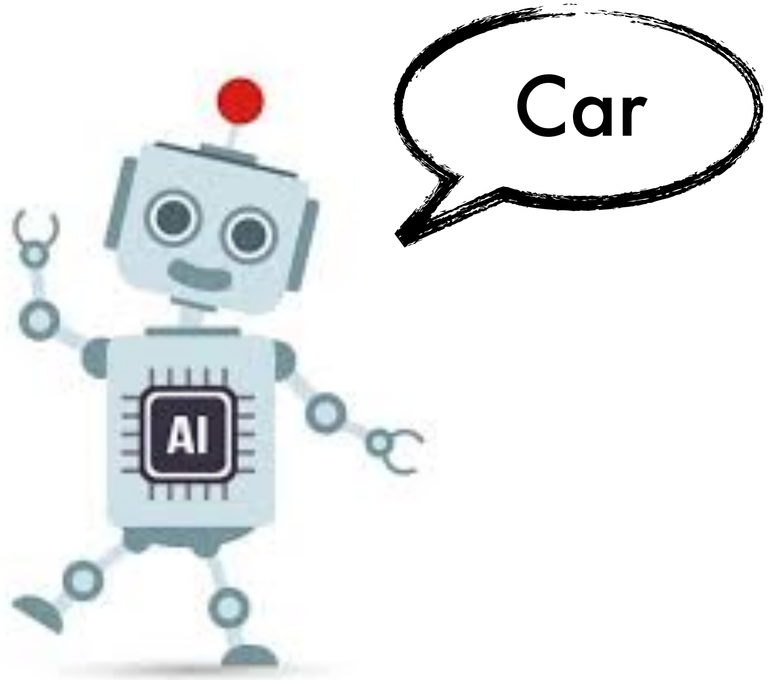
Google Inclusive Images Competition



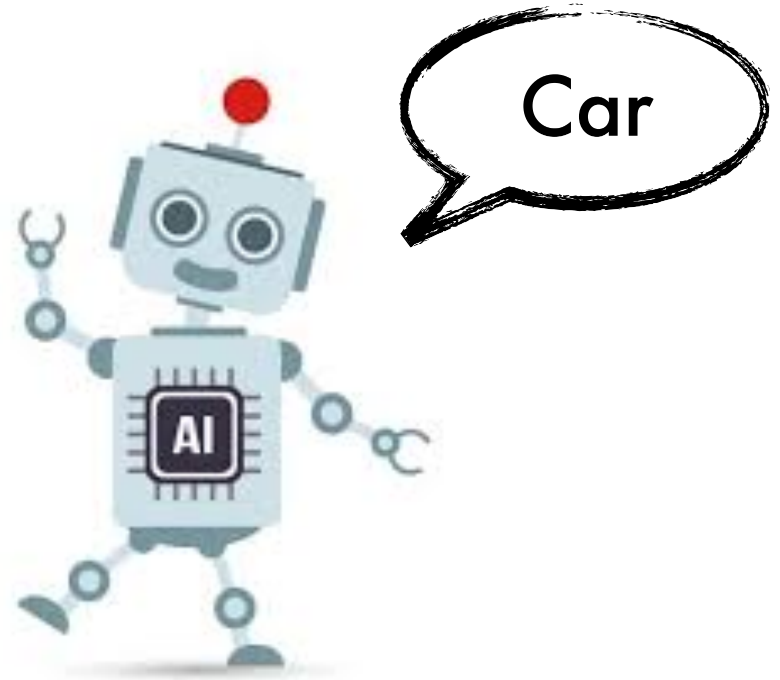
Google Inclusive Images Competition

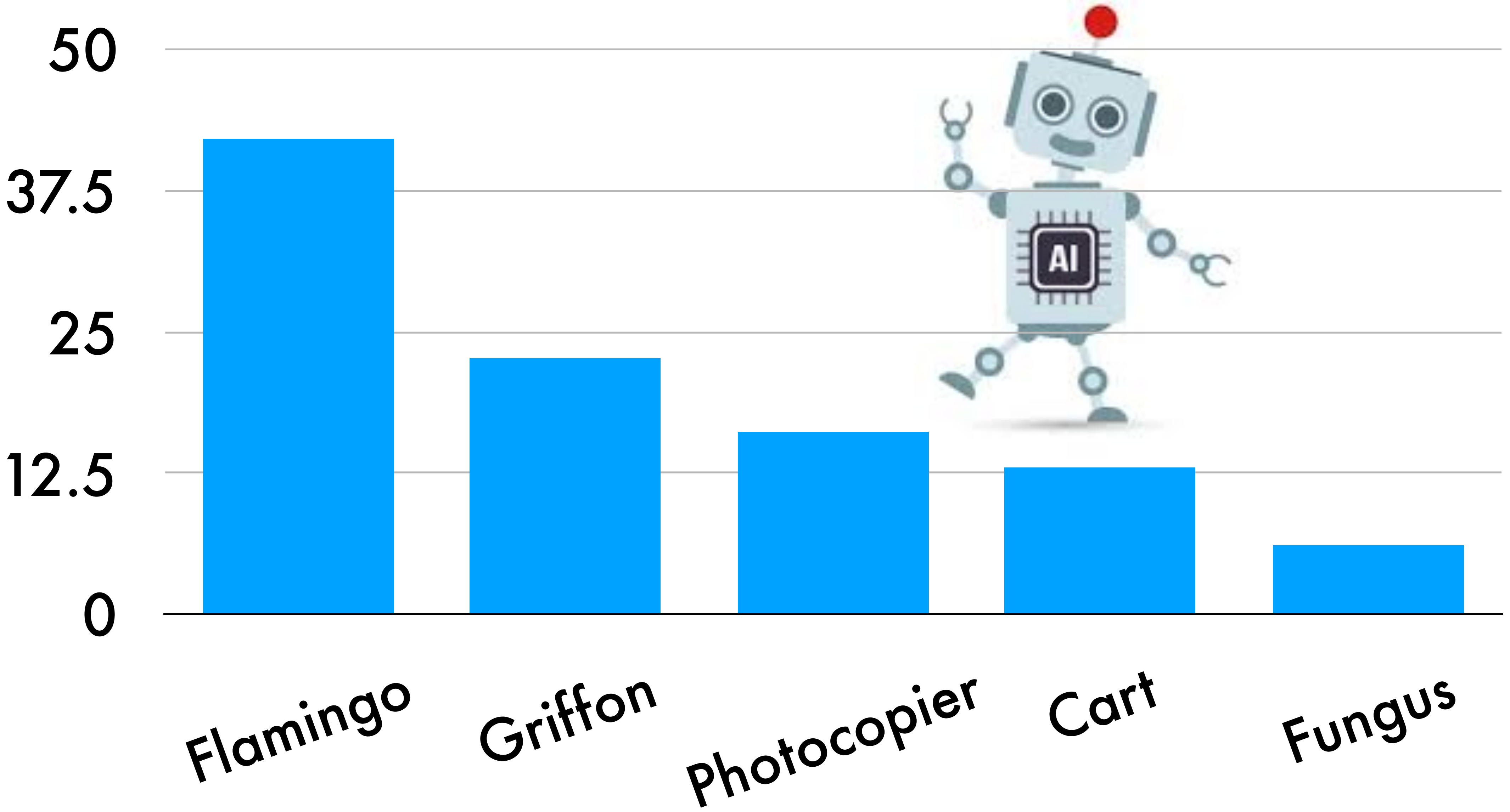

Car


Car
AI



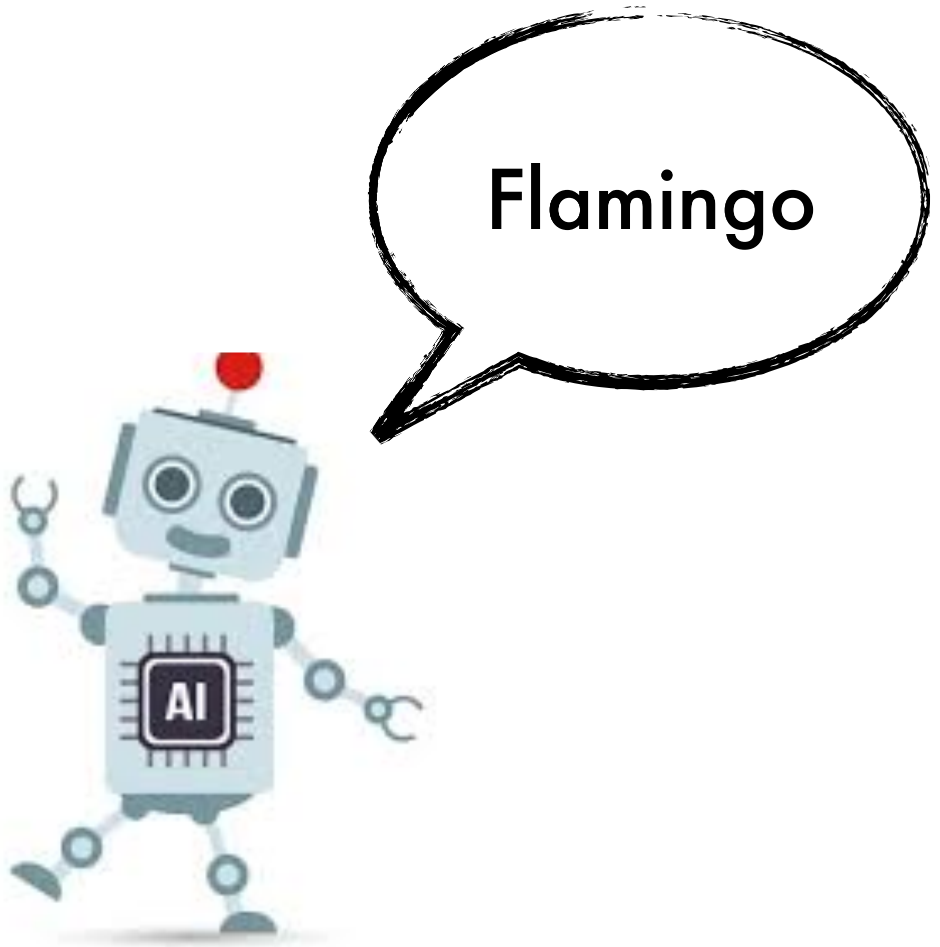
Example courtesy @hardmaru [2019]

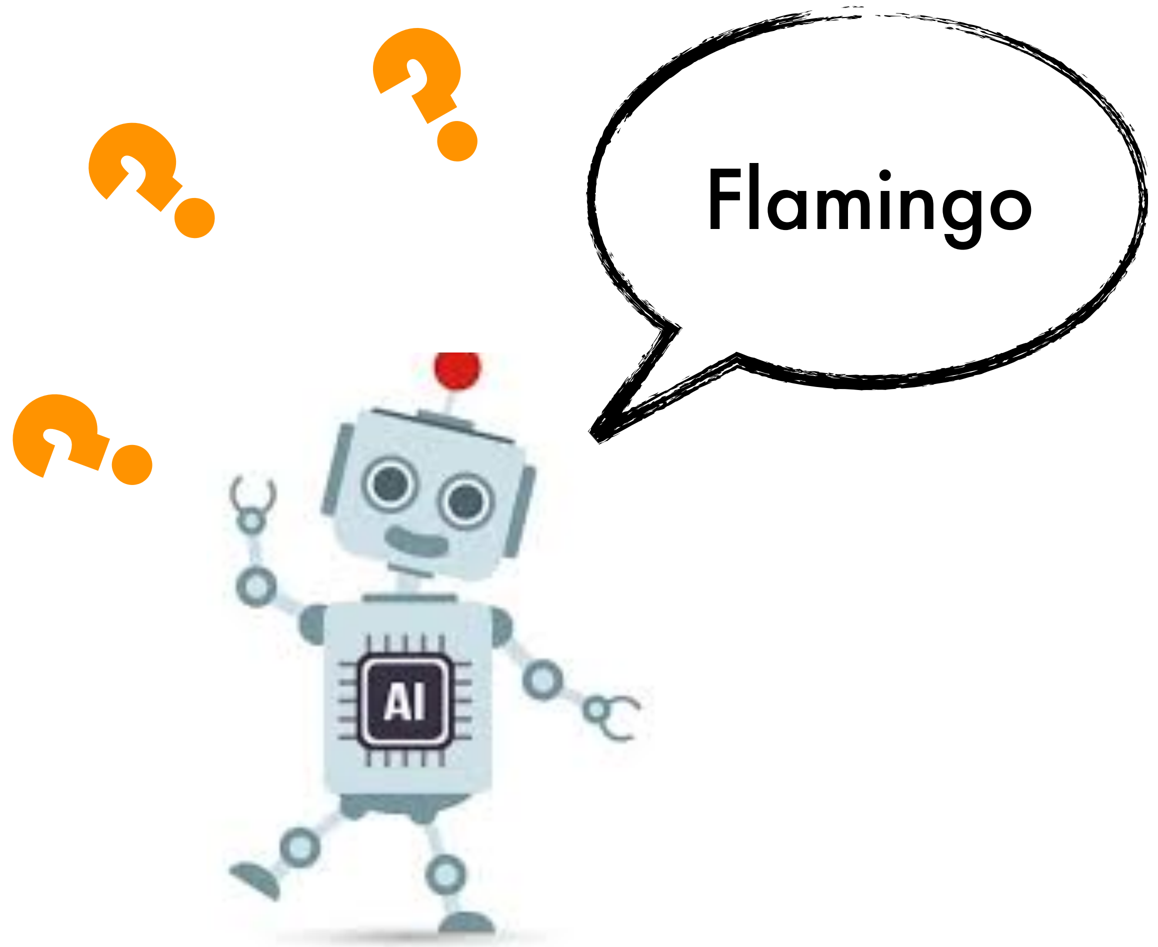




Example courtesy @hardmaru [2019]


Self-Driving
Cars

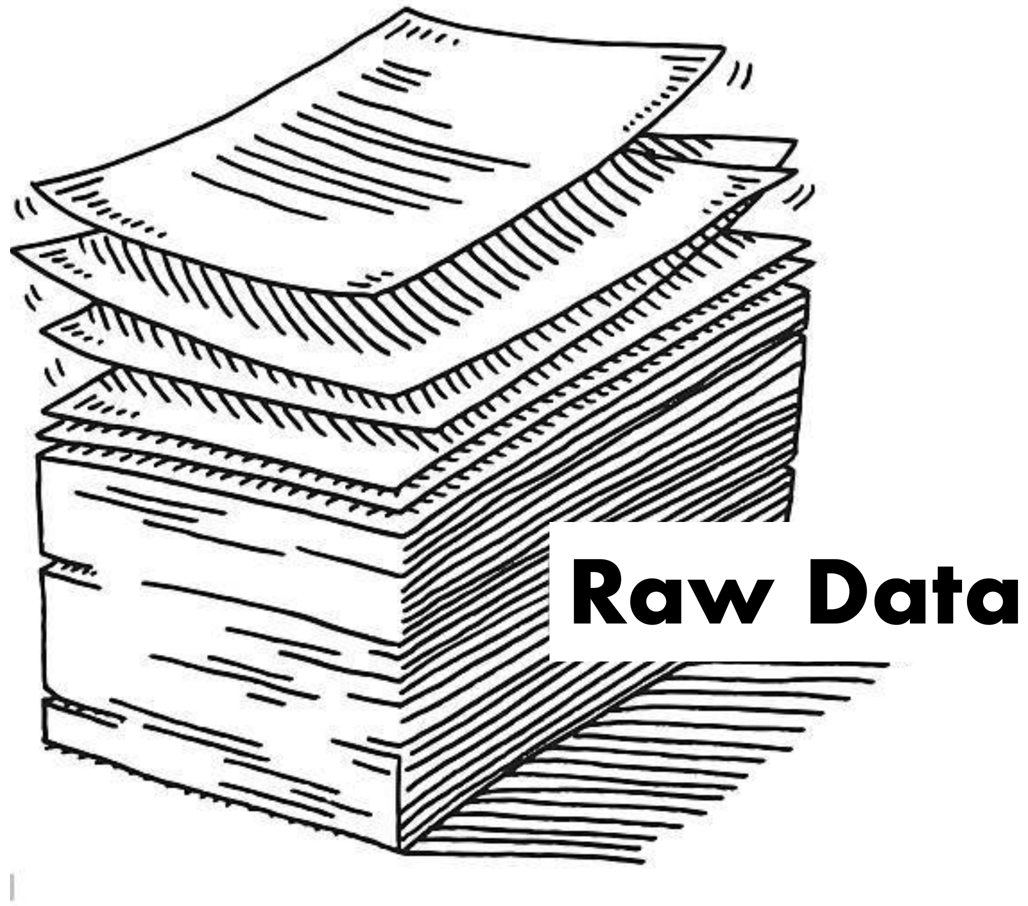

Flamingo
AI

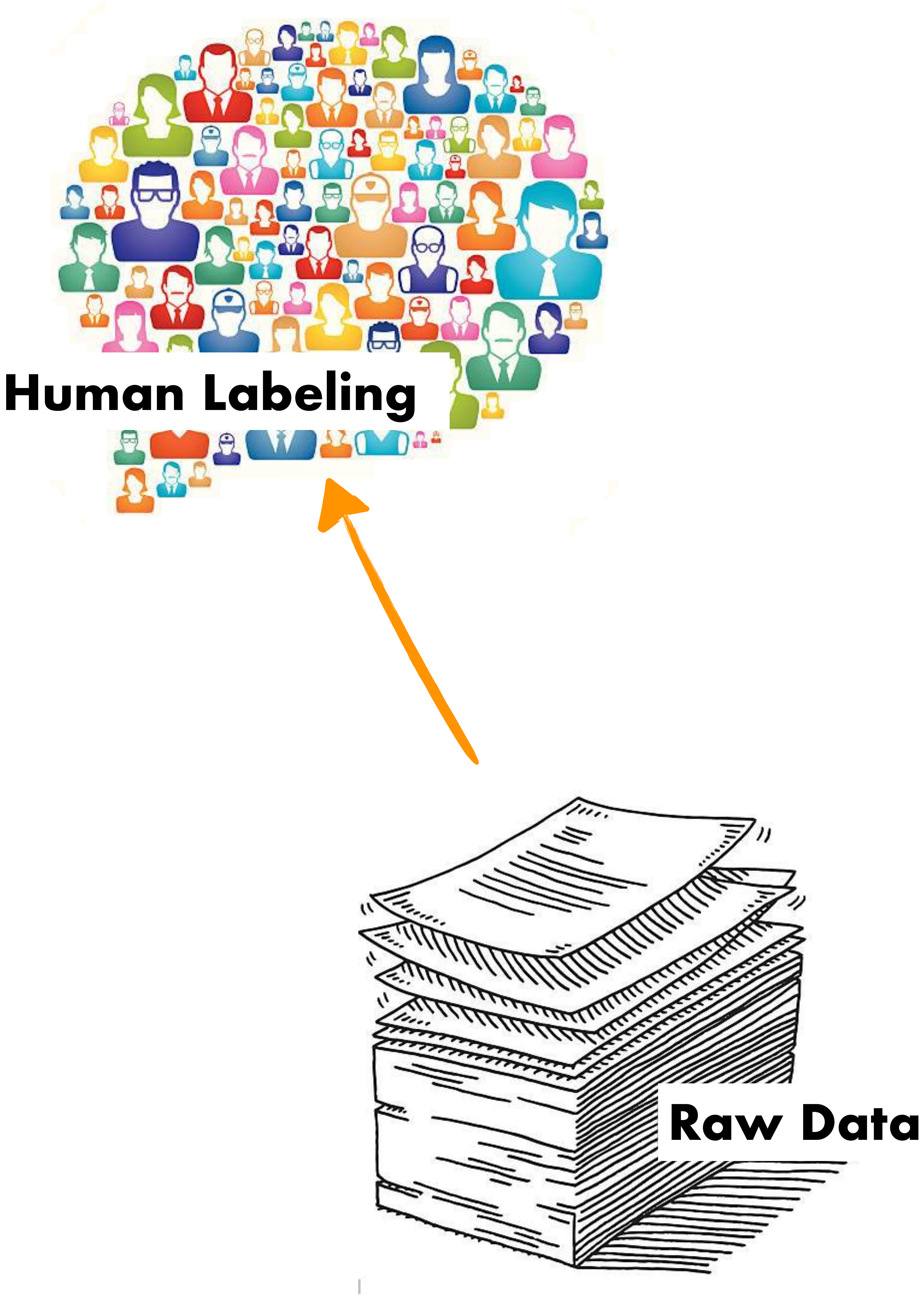


Why does AI, so successful in many applications, still make embarrassing mistakes?

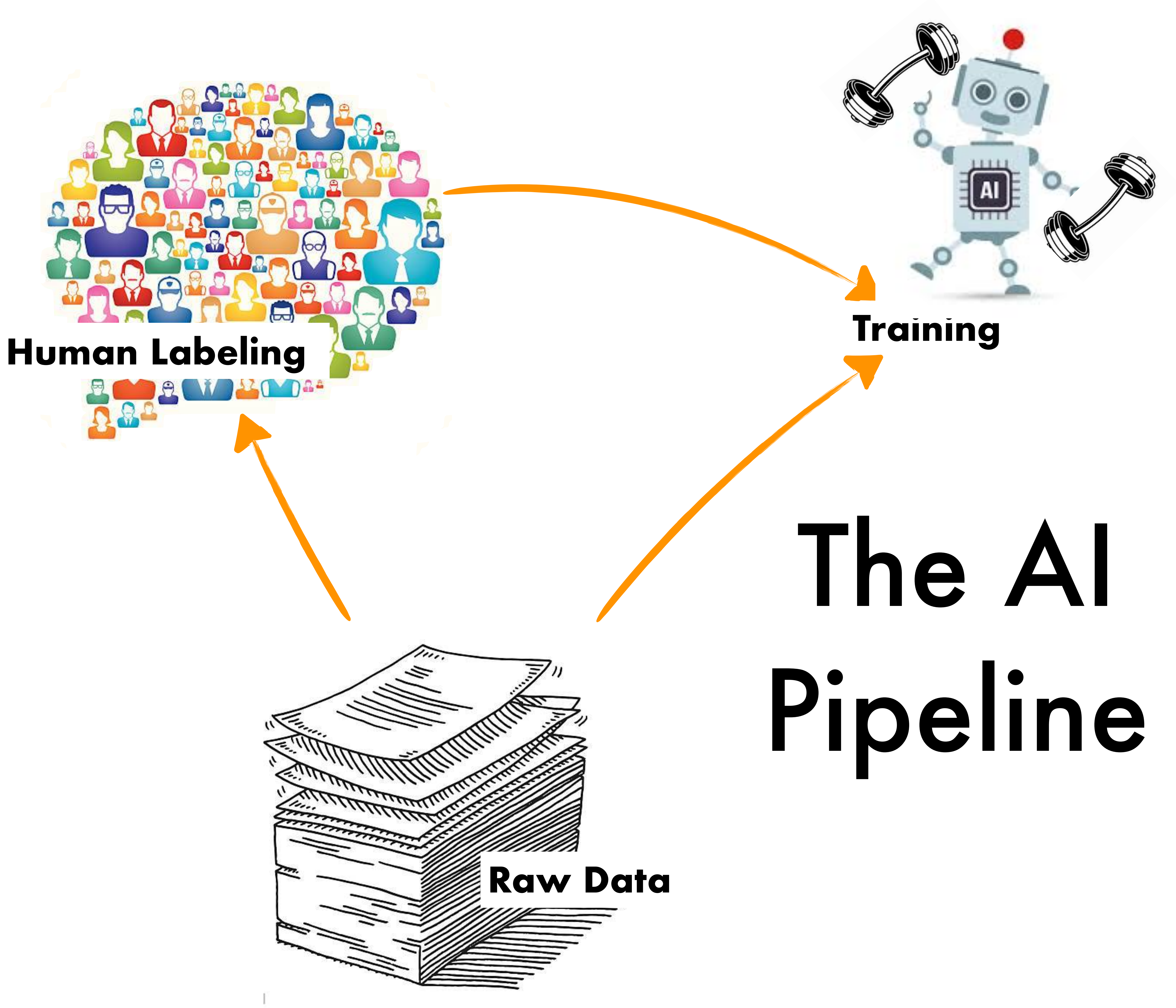
The AI Pipeline

The AI Pipeline



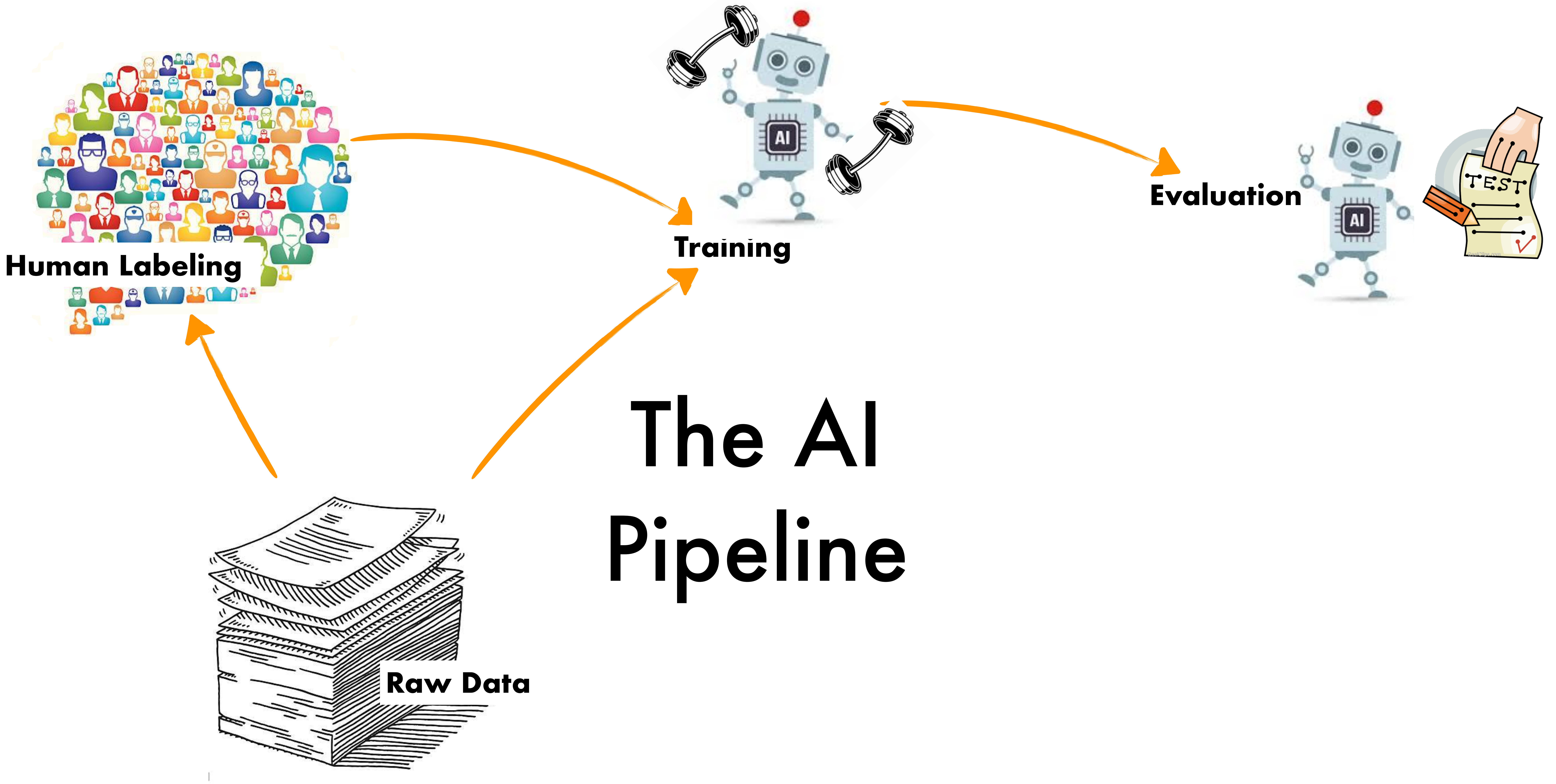


The AI Pipeline

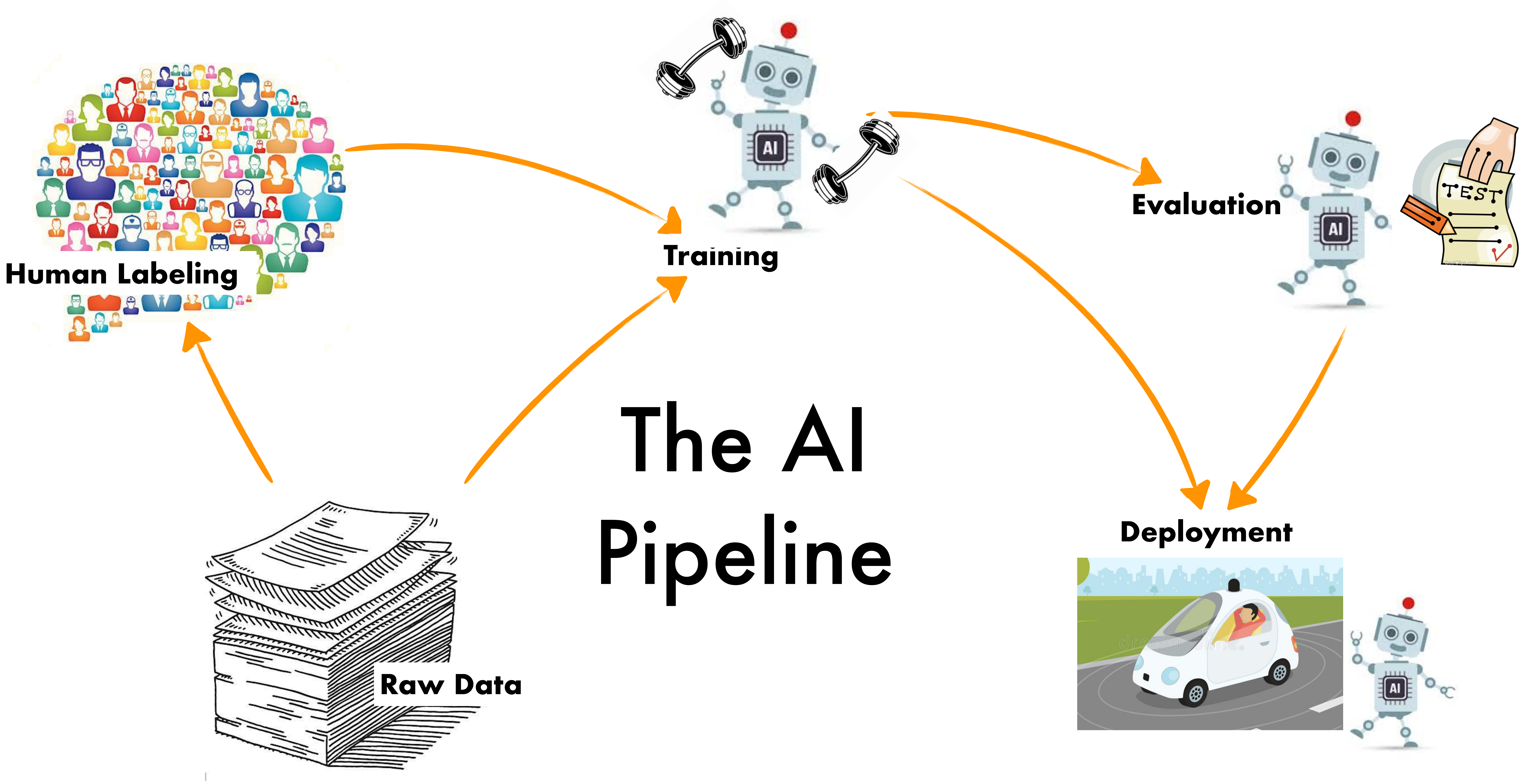


The AI Pipeline

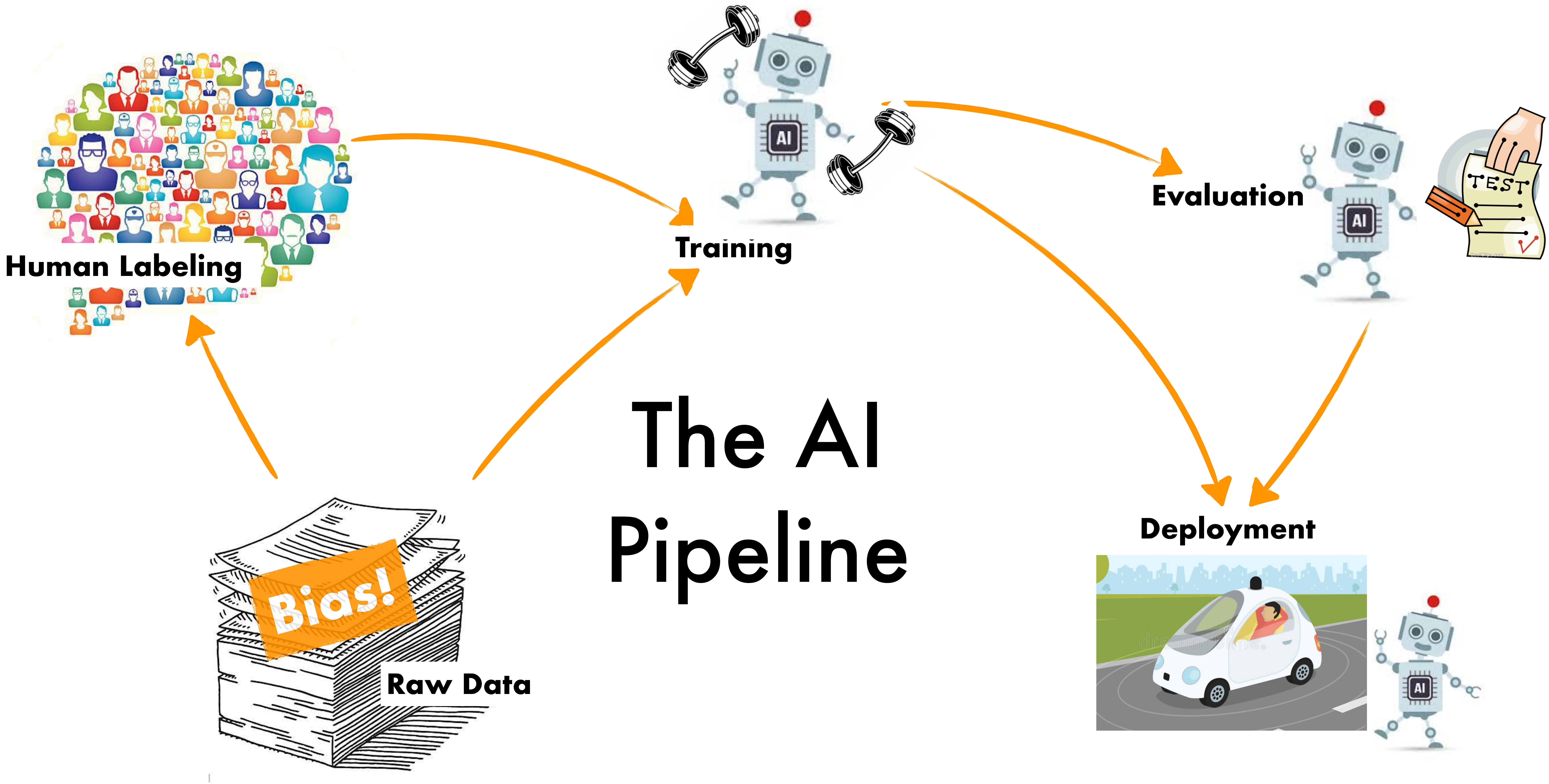
The AI Pipeline



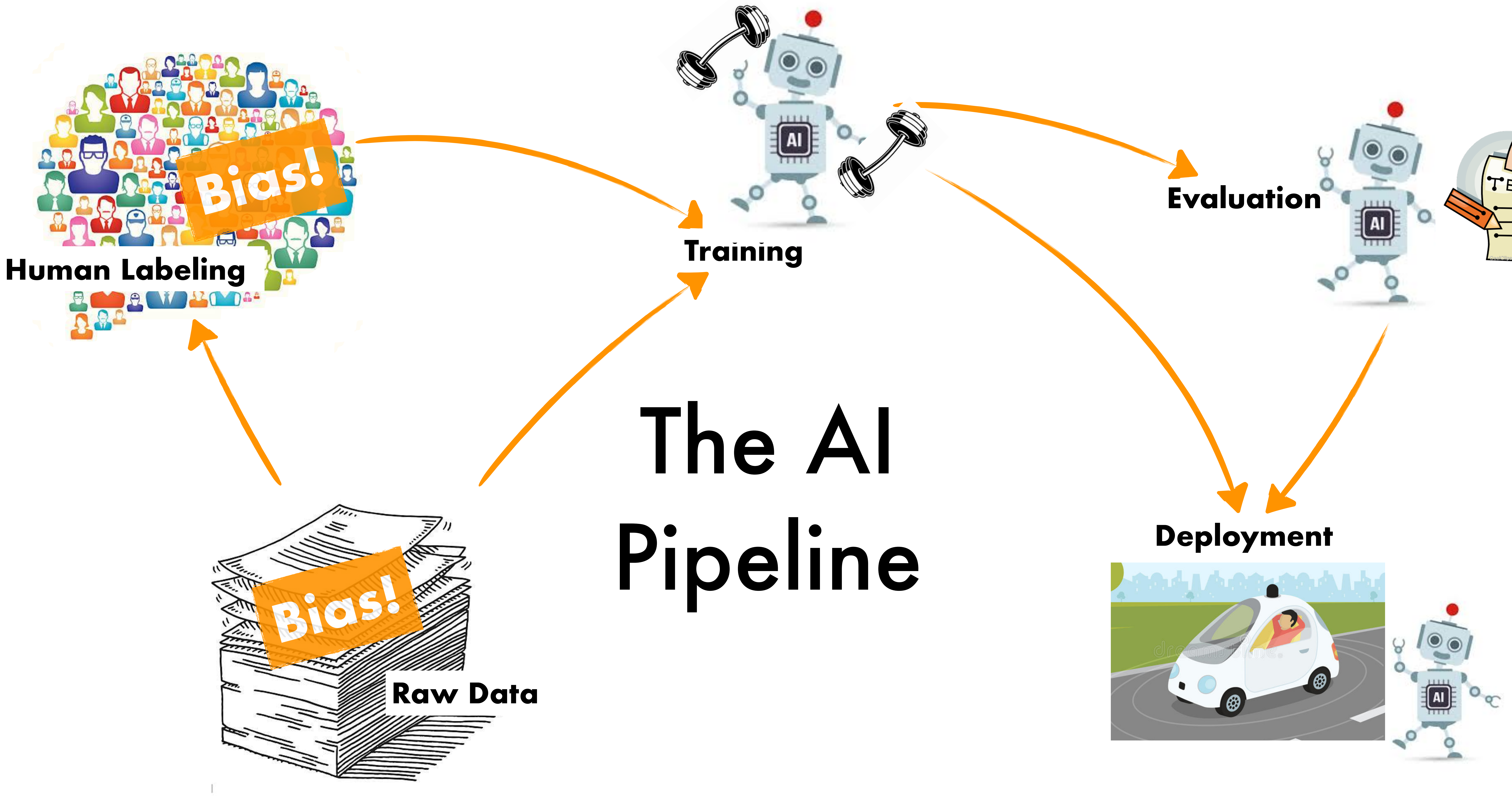
The AI Pipeline



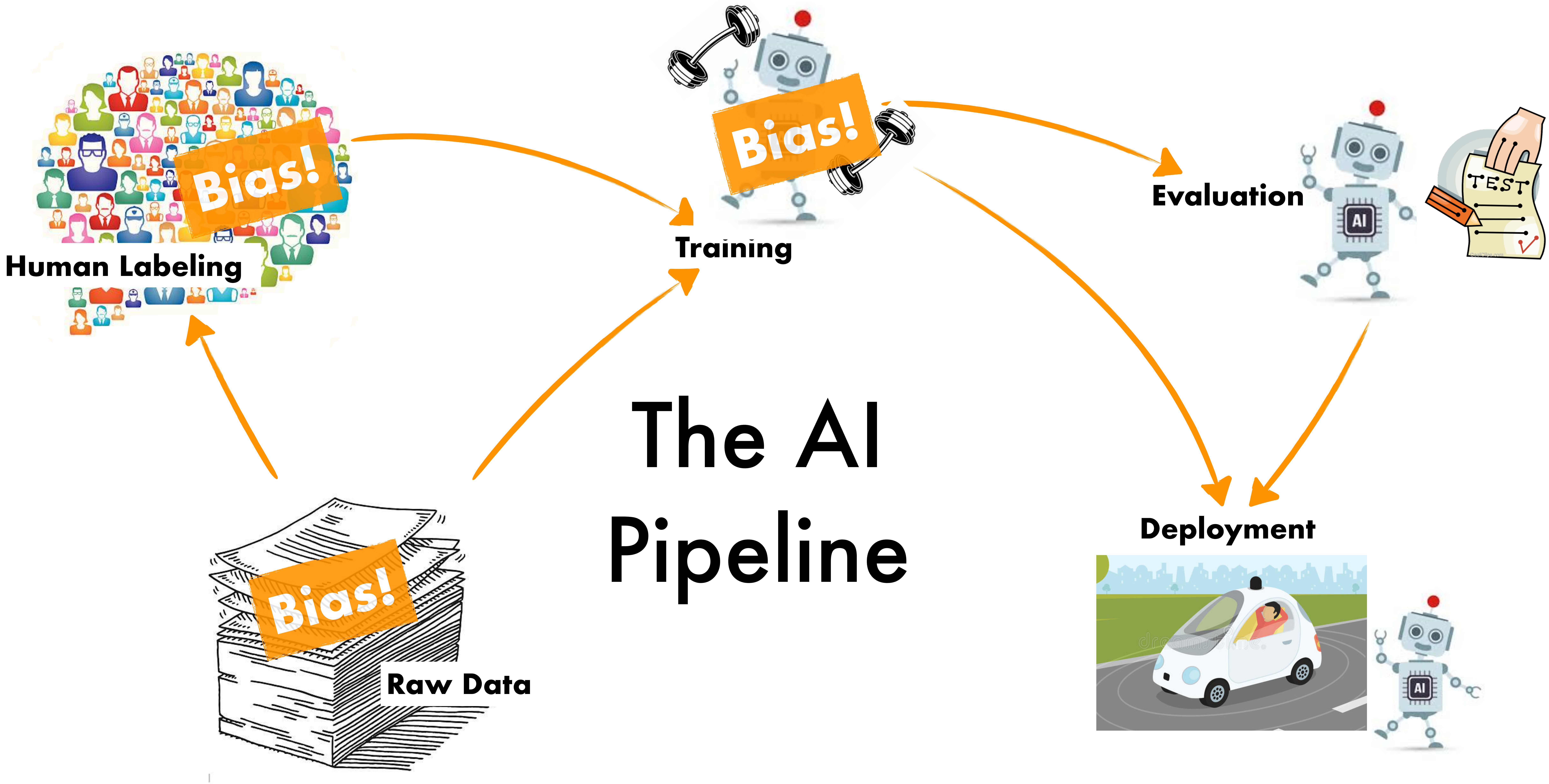
The AI Pipeline

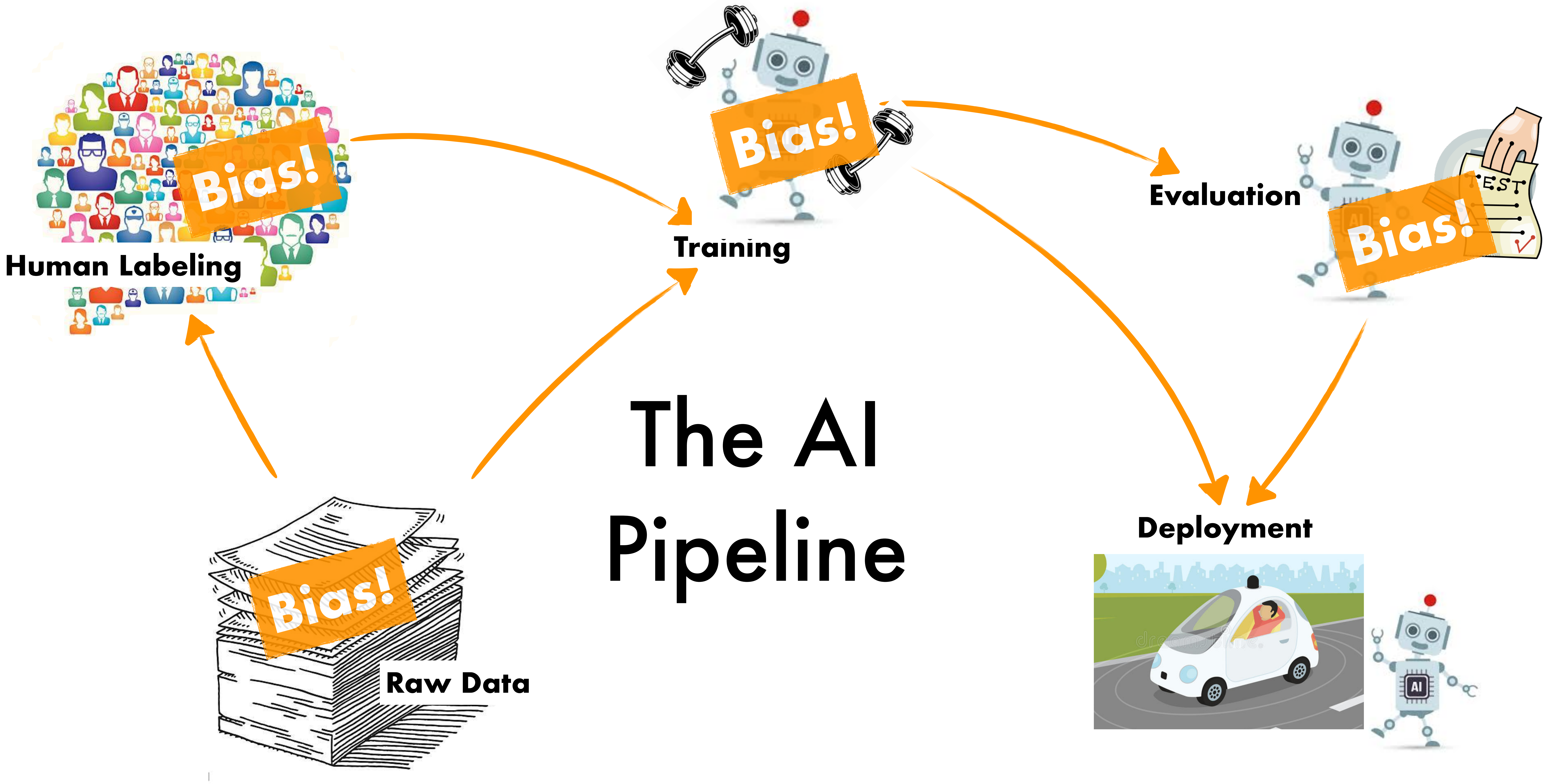


The AI Pipeline

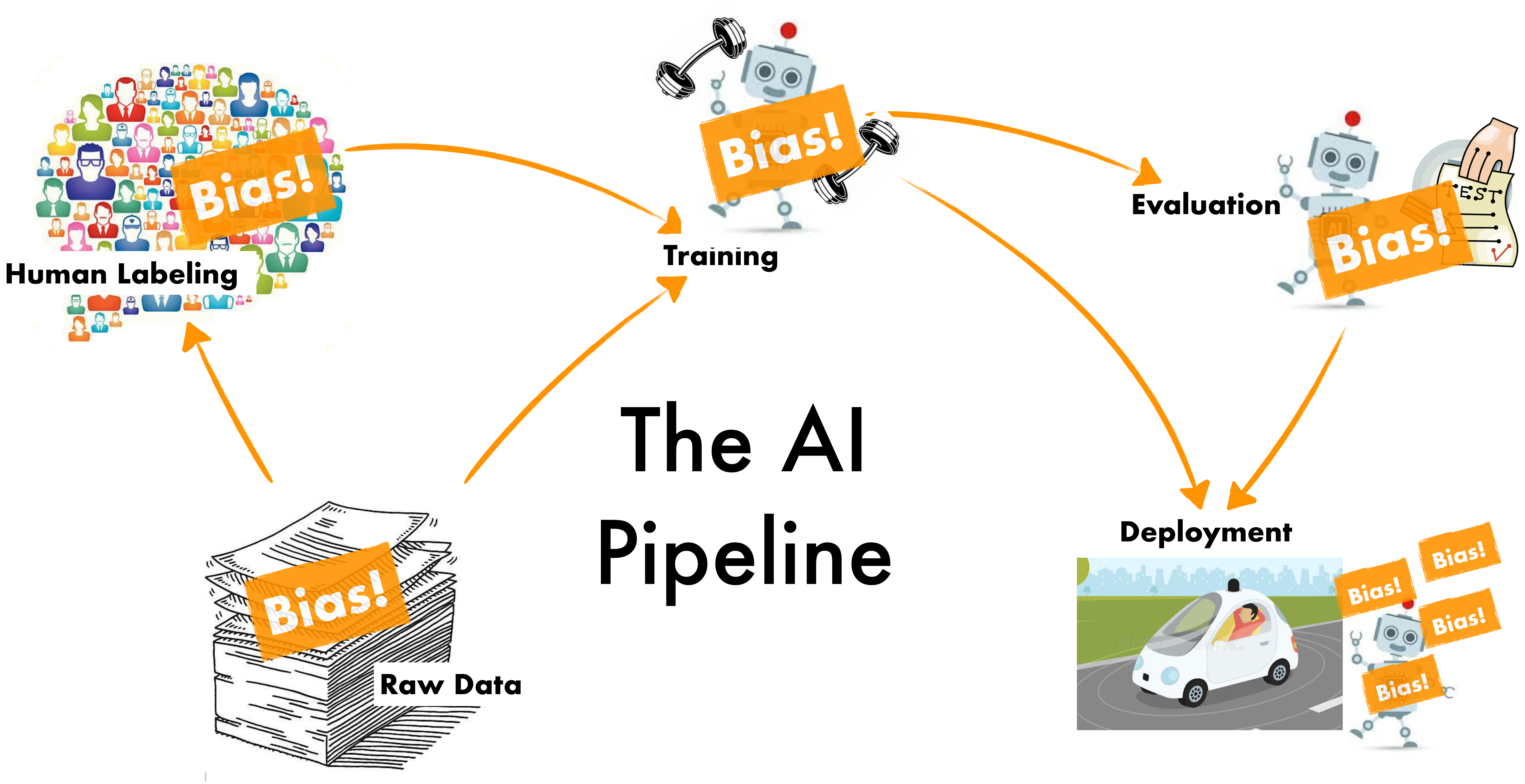


The AI Pipeline




Human Labeling
Bias!
Training
Bias!
Evaluation
Bias!
TEST
The AI Pipeline
Deployment
Raw Data
Bias!
AI

The AI Pipeline



This Talk

This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

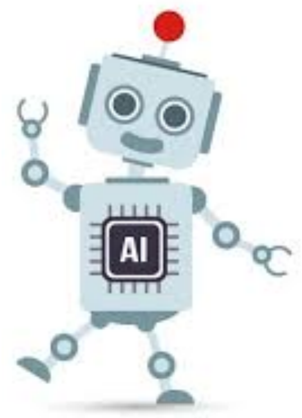
What is Bias?

What is Bias?

- Preference of one decision over another

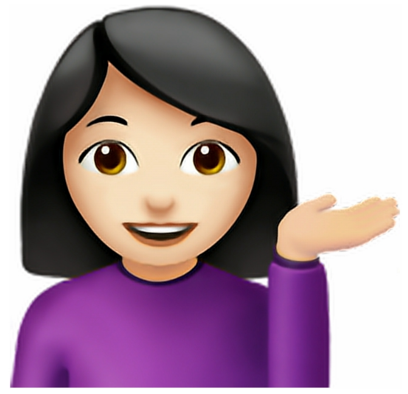
What is Bias?

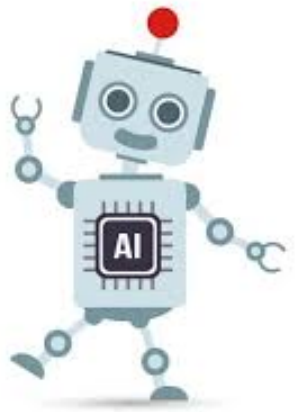
- Preference of one decision over another



What is Bias?

- Preference of one decision over another

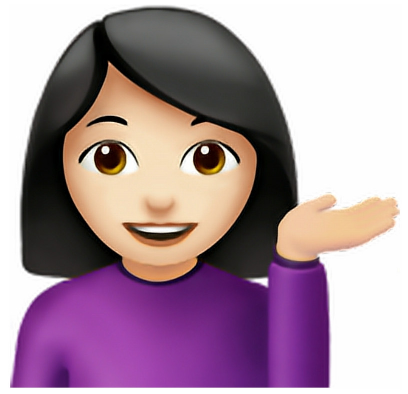


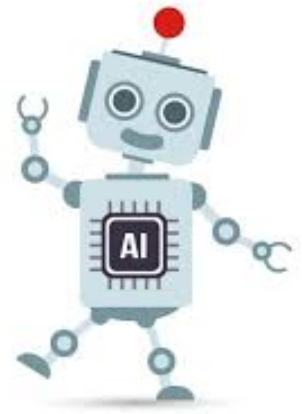


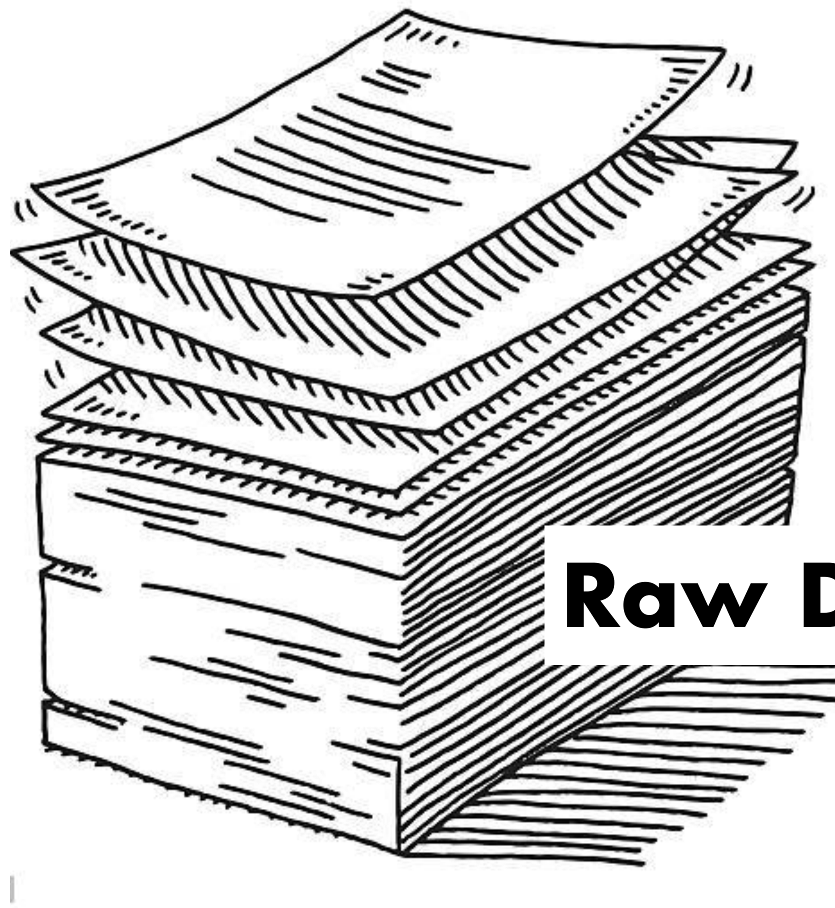
What is Bias?

- Preference of one decision over another

Human biases are reflected in datasets







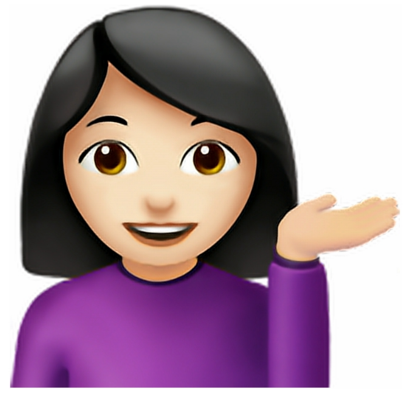
Raw Data

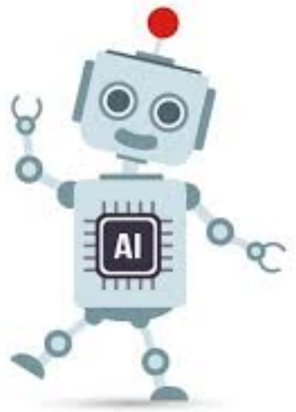
Human Labeling

What is Bias?

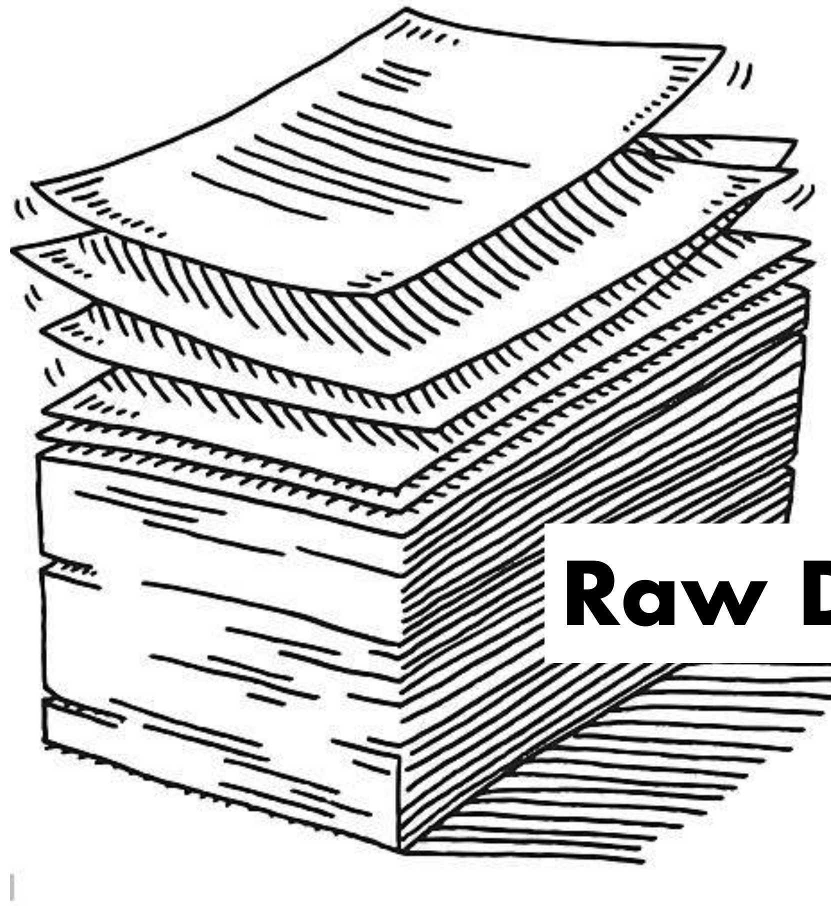
- Preference of one decision over another

Human biases are reflected in datasets



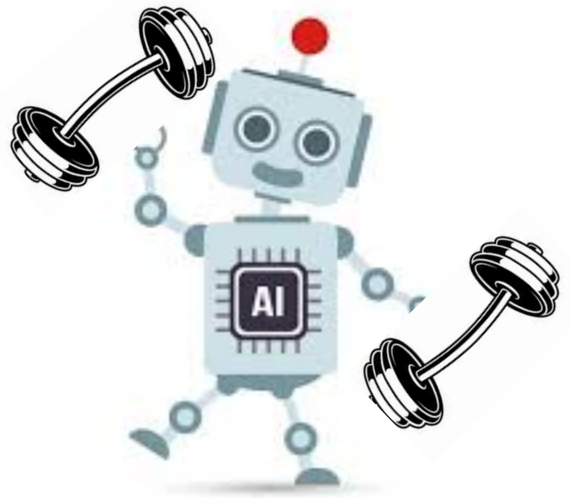


Model biases are reflected in AI decisions



Raw Data

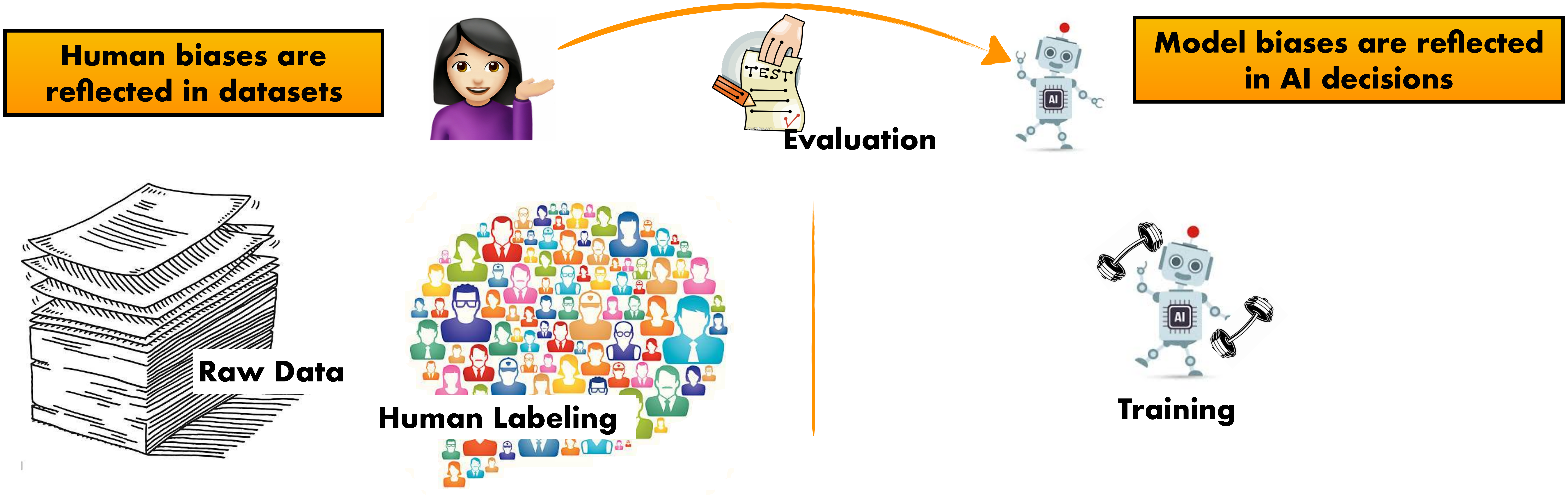
Human Labeling



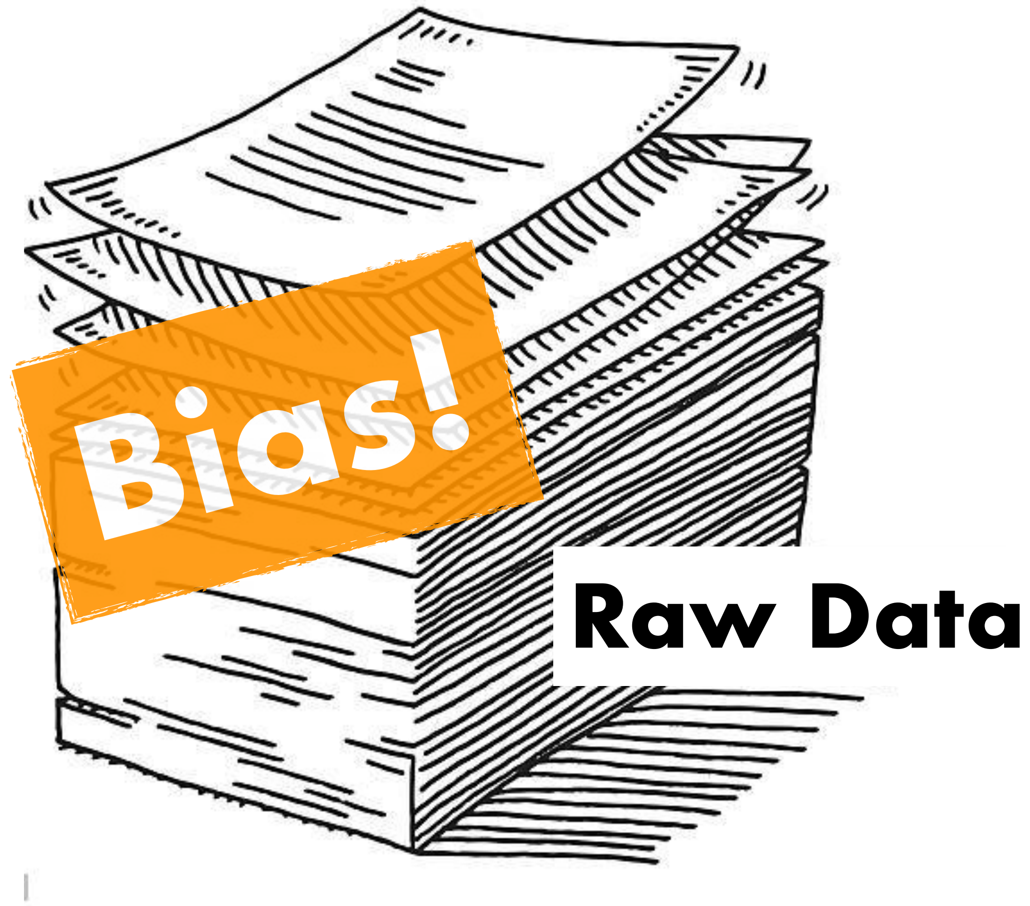
Training

What is Bias?

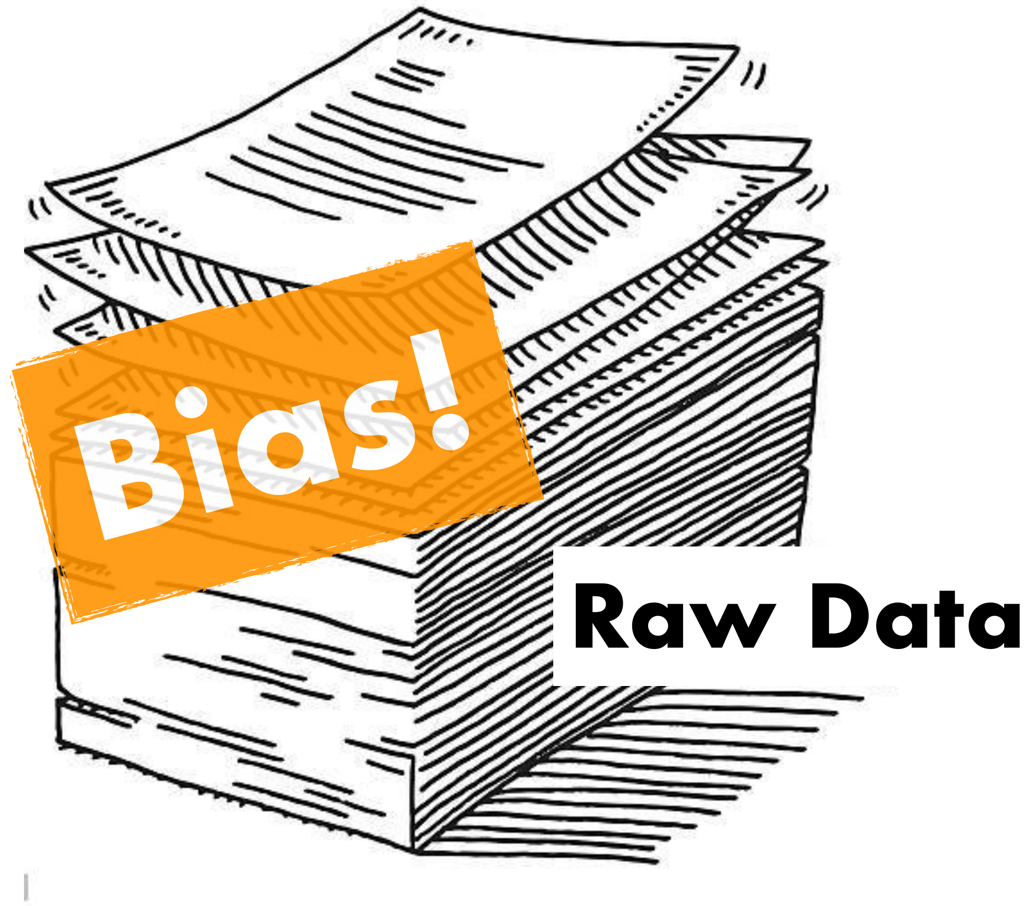
- Preference of one decision over another



Human Biases in Raw Data



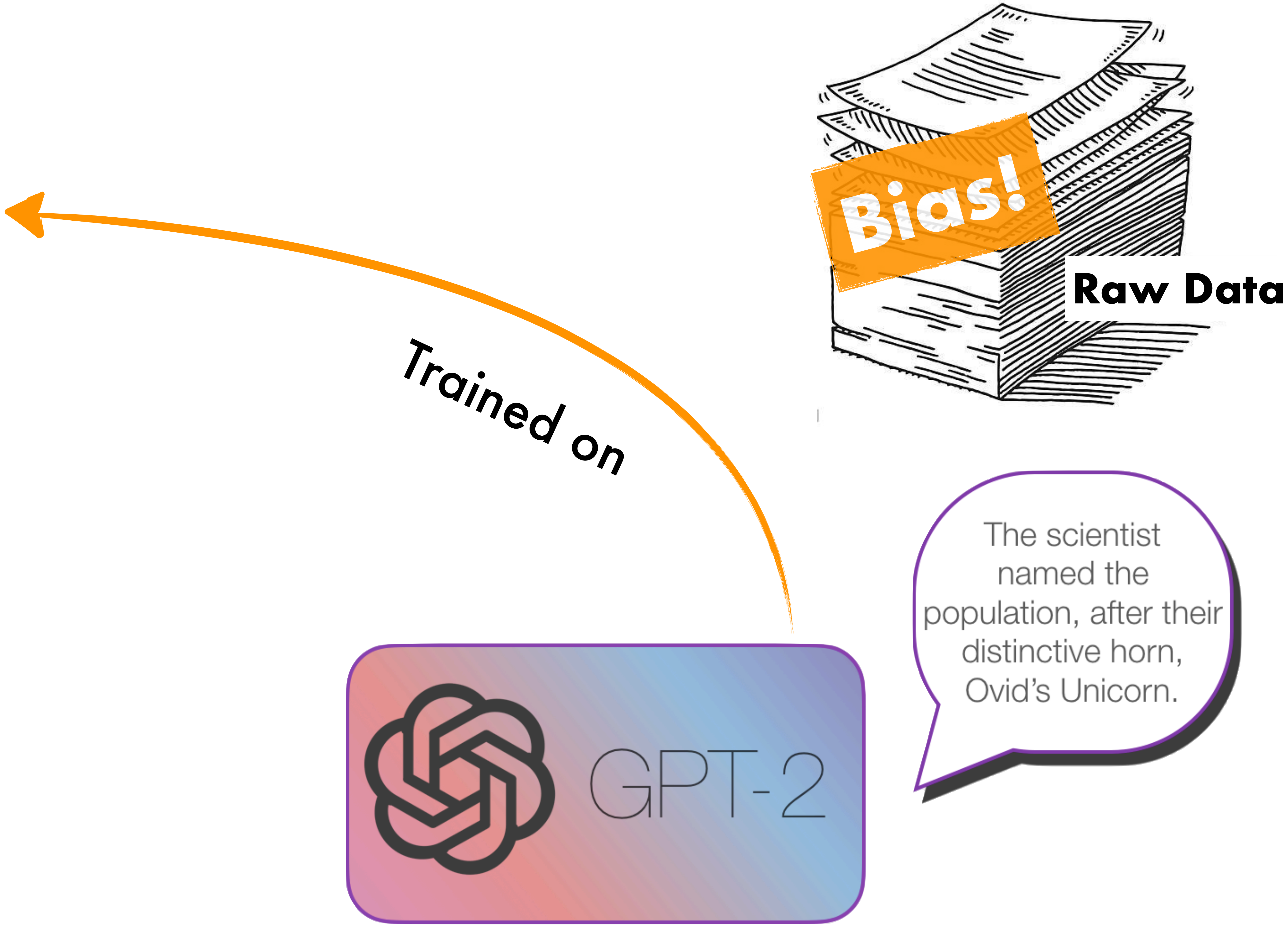
Human Biases in Raw Data



GPT-2

The scientist named the population, after their distinctive horn, Ovid's Unicorn.

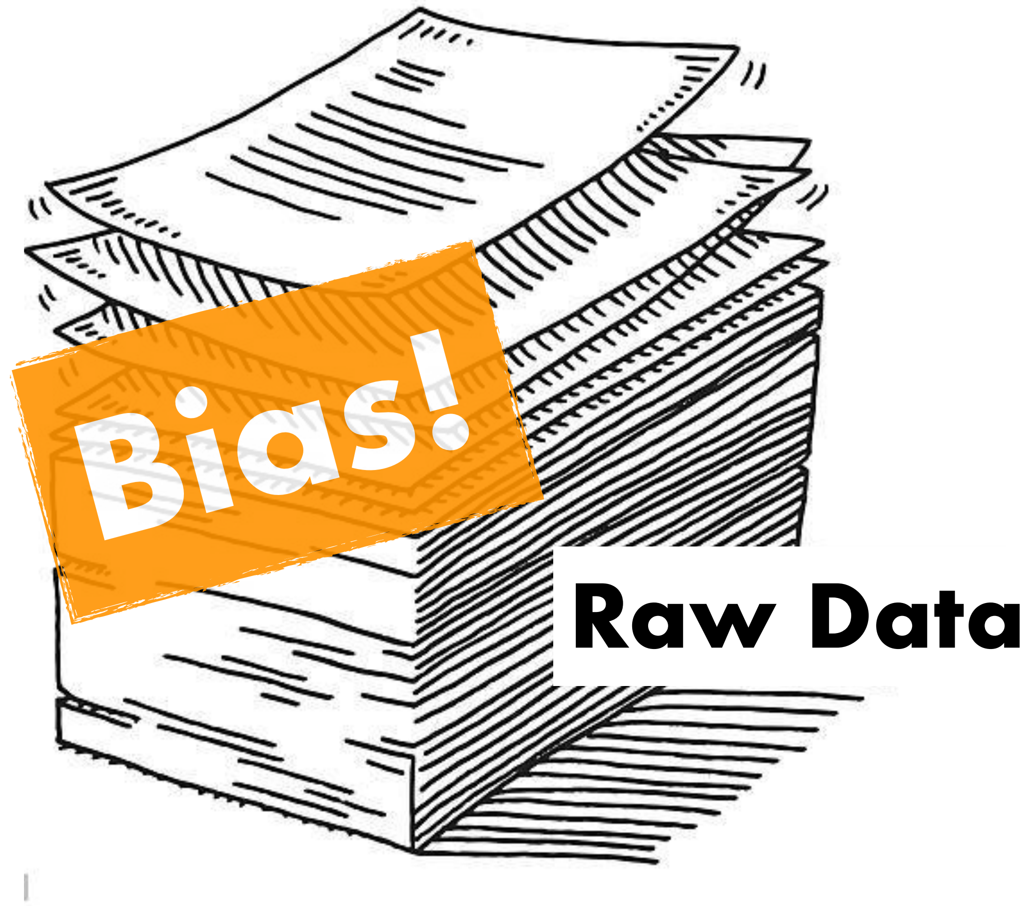
Human Biases in Raw Data



Human Biases in Raw Data

- The Donald
- Breitbart News

Trained on



GPT-2

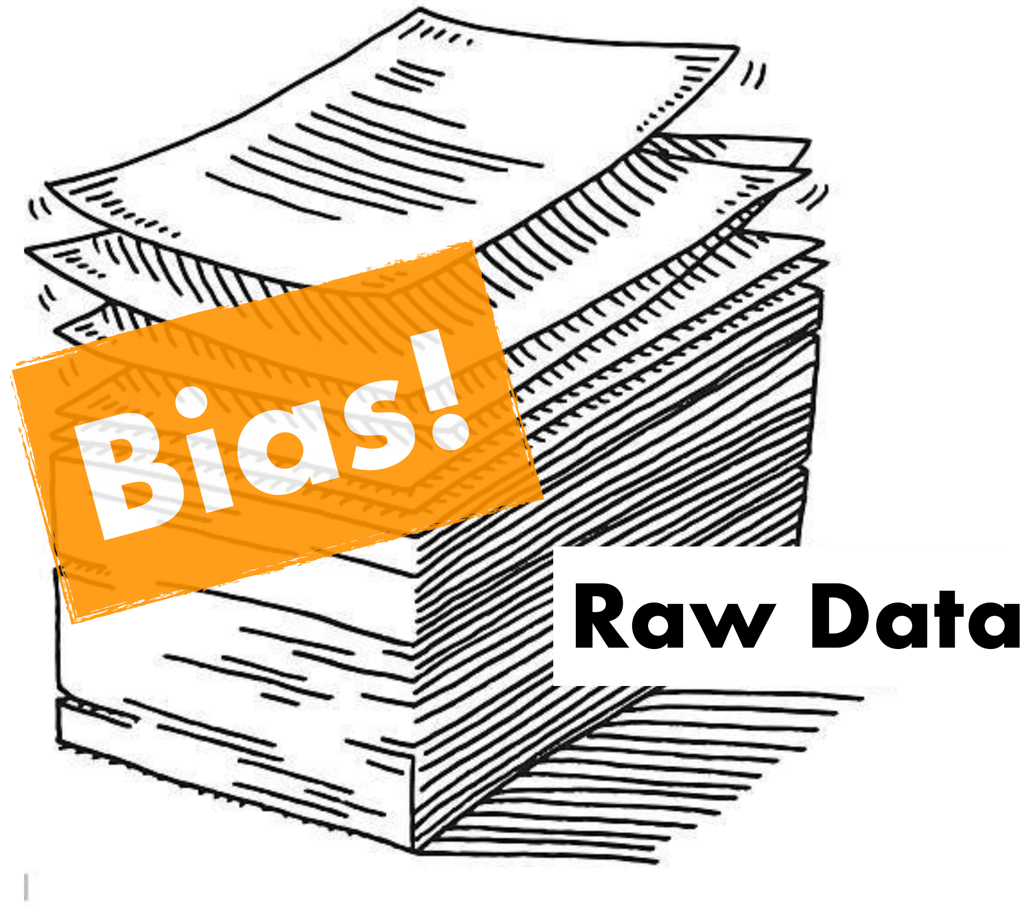
The scientist named the population, after their distinctive horn, Ovid's Unicorn.

Human Biases in Raw Data

- The Donald
- Breitbart News

RealToxicityPrompts [Gehman et al., 2020]

Trained on



GPT-2

The scientist named the population, after their distinctive horn, Ovid's Unicorn.

Human biases in Data Annotation

Human biases in Data Annotation

Example from the Flickr30k Dataset

Credit: van Miltenburg [2016] & Paullada A. [2020] Using Datasets Wisely

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

Example from the Flickr30k Dataset

Credit: van Miltenburg [2016] & Paullada A. [2020] Using Datasets Wisely

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

A worker is being scolded by her boss in a stern lecture.

Example from the Flickr30k Dataset

Credit: van Miltenburg [2016] & Paullada A. [2020] Using Datasets Wisely

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

A worker is being scolded by her boss in a stern lecture.

A manager talks to an employee about job performance.

Example from the Flickr30k Dataset

Credit: van Miltenburg [2016] & Paullada A. [2020] Using Datasets Wisely

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

A worker is being scolded by her boss in a stern lecture.

A manager talks to an employee about job performance.

Sonic employees talking about work.

Example from the Flickr30k Dataset

Credit: van Miltenburg [2016] & Paullada A. [2020] Using Datasets Wisely

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

A worker is being scolded by her boss in a stern lecture.

A manager talks to an employee about job performance.

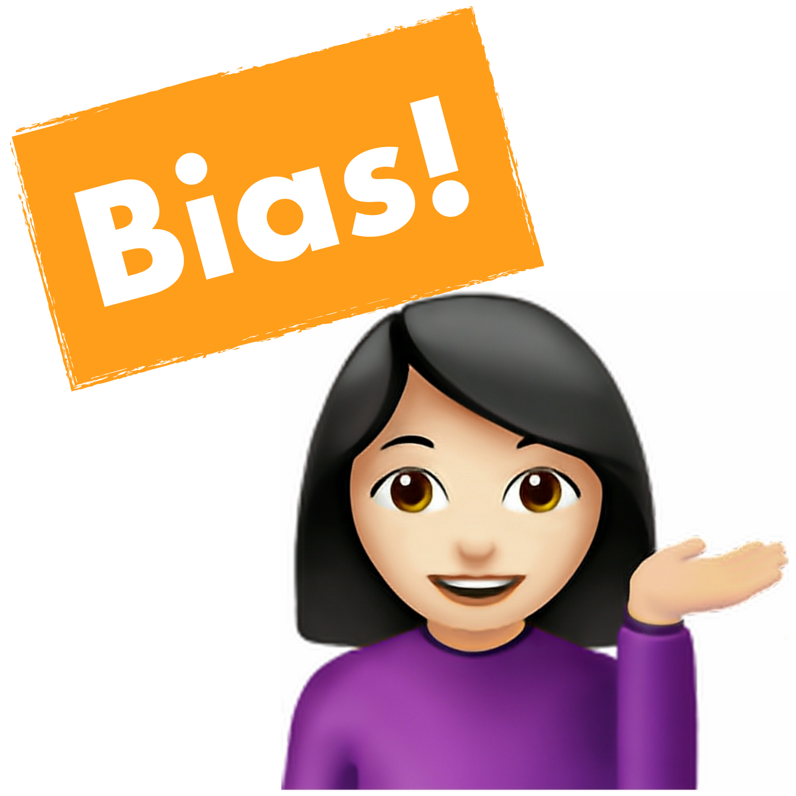
Sonic employees talking about work.

A hot, blond girl getting criticized by her boss.

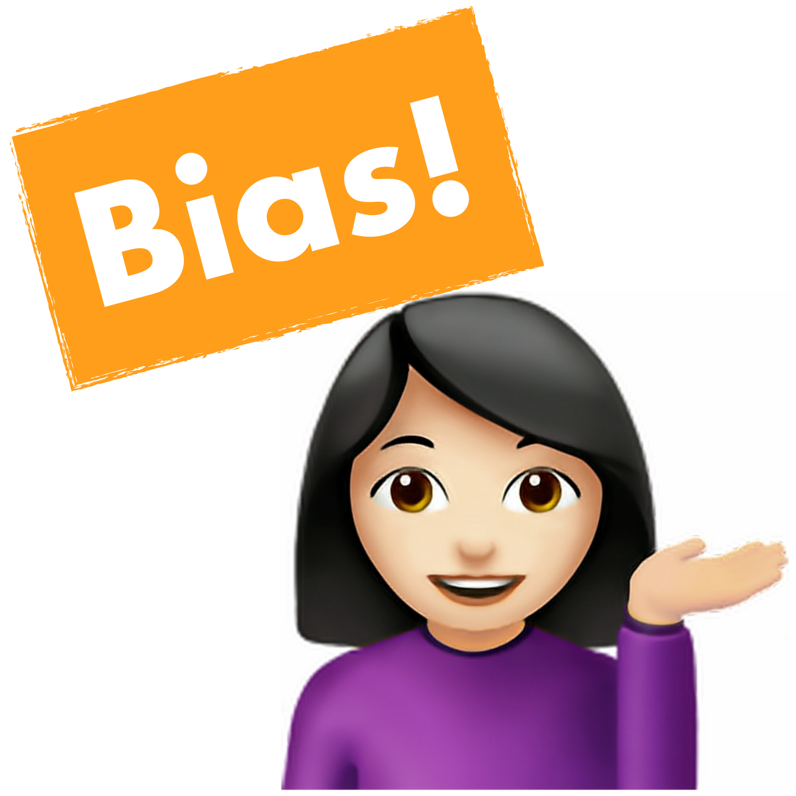
Example from the Flickr30k Dataset

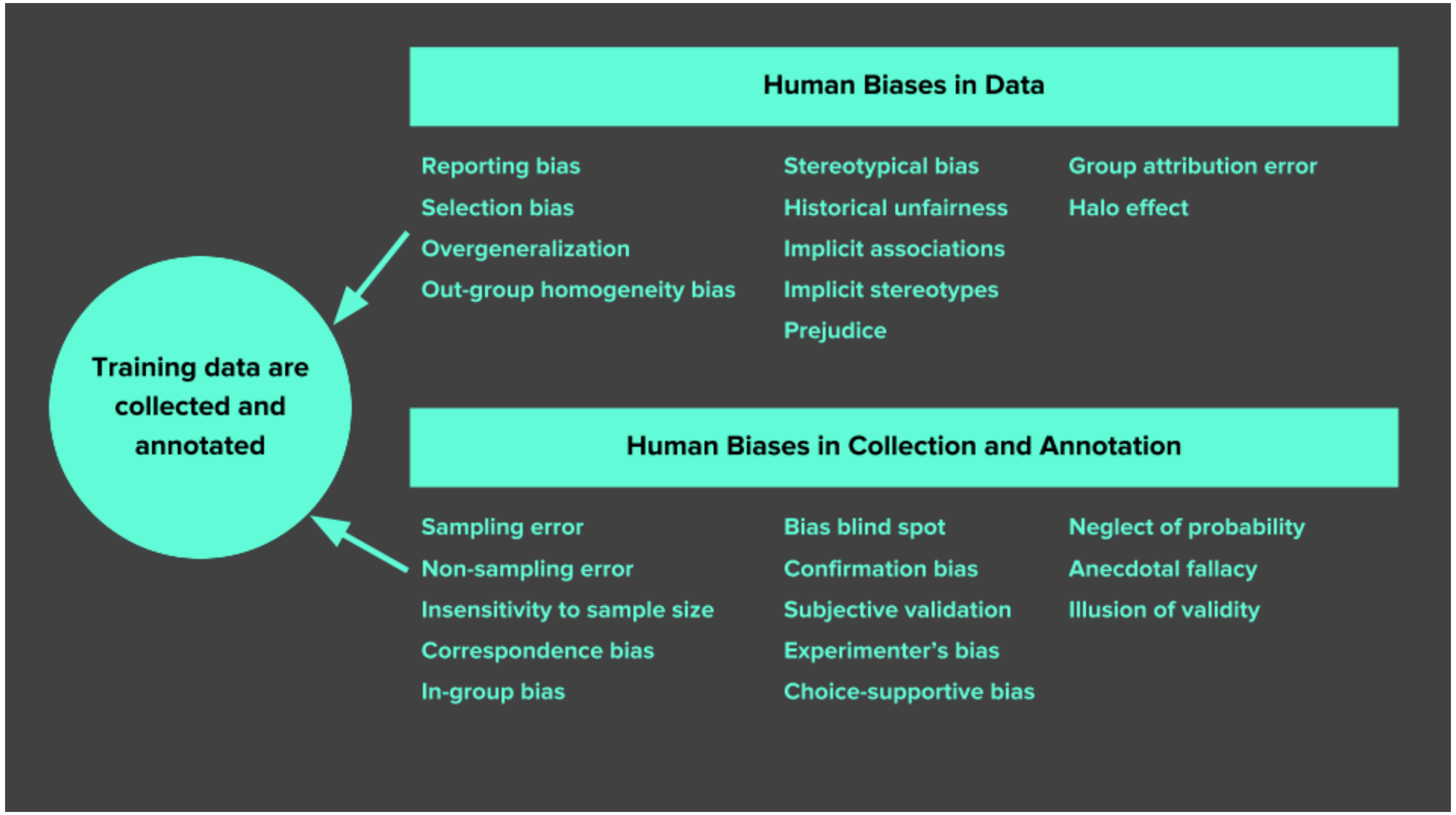
Credit: van Miltenburg [2016] & Paullada A. [2020] Using Datasets Wisely

Human Biases affecting Datasets



Human Biases affecting Datasets





Source: Bias in the Vision and Language of Artificial Intelligence, Mitchell 2019

Case Study: Natural Language Inference

Stanford NLI [Bowman et al., 2015]

14

Case Study: Natural Language Inference

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

Stanford NLI [Bowman et al., 2015]

14

Case Study: Natural Language Inference

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

Premise

A dog is chasing birds on the shore of the ocean.

Hypothesis

The cat is chasing birds.

Stanford NLI [Bowman et al., 2015]

14

Case Study: Natural Language Inference

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

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

Premise

A dog is chasing birds on the shore of the ocean.

Hypothesis

The cat is chasing birds.

Stanford NLI [Bowman et al., 2015]

14

Case Study: Natural Language Inference

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

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

Premise

A dog is chasing birds on the shore of the ocean.

Hypothesis

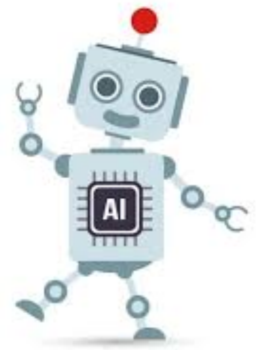
The cat is chasing birds.

Stanford NLI [Bowman et al., 2015]

14

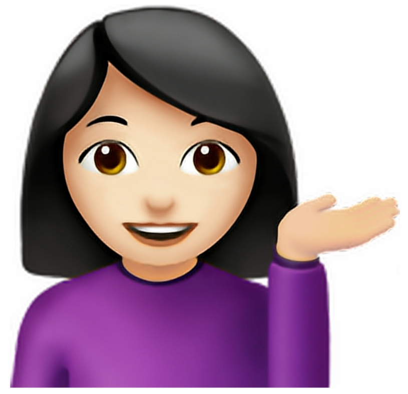
	Premise	Hypothesis
1	A dog is chasing birds on the shore of the ocean.	The cat is chasing birds.
2	Three kids playing with a toy cat in a garden.	There's a toy cat and dog in the garden.
3	A dog and cat are snuggling up during a nap.	A dog and cat are sharing a nap.
4	A few people are staring at something.	The people are staring at a cat.

Premise	A dog is chasing birds on the shore of the ocean.	Three kids playing with a toy cat in a garden.	A dog and cat are snuggling up during a nap.	A few people are staring at something.
Hypothesis	The cat is chasing birds.	There's a toy cat and dog in the garden.	A dog and cat are sharing a nap.	The people are staring at a cat.
	**Contradiction**	**Neutral**	**Entailment**	**Neutral**

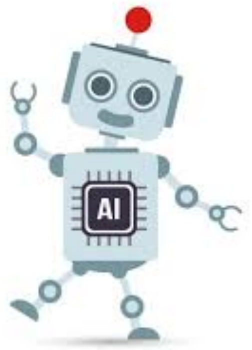
Premise	A dog is chasing birds on the shore of the ocean.	Three kids playing with a toy cat in a garden.	A dog and cat are snuggling up during a nap.	A few people are staring at something.
Hypothesis	The cat is chasing birds.	There's a toy cat and dog in the garden.	A dog and cat are sharing a nap.	The people are staring at a cat.
	**Contradiction**	**Neutral**	**Entailment**	**Neutral**
	**Contradiction**	**Contradiction**	**Contradiction**	**Contradiction**

Premise	A dog is chasing birds on the shore of the ocean.	Three kids playing with a toy cat in a garden.	A dog and cat are snuggling up during a nap.	A few people are staring at something.
Hypothesis	The cat is chasing birds.	There's a toy cat and dog in the garden.	A dog and cat are sharing a nap.	The people are staring at a cat.



Contradiction **Neutral** **Entailment** **Neutral**



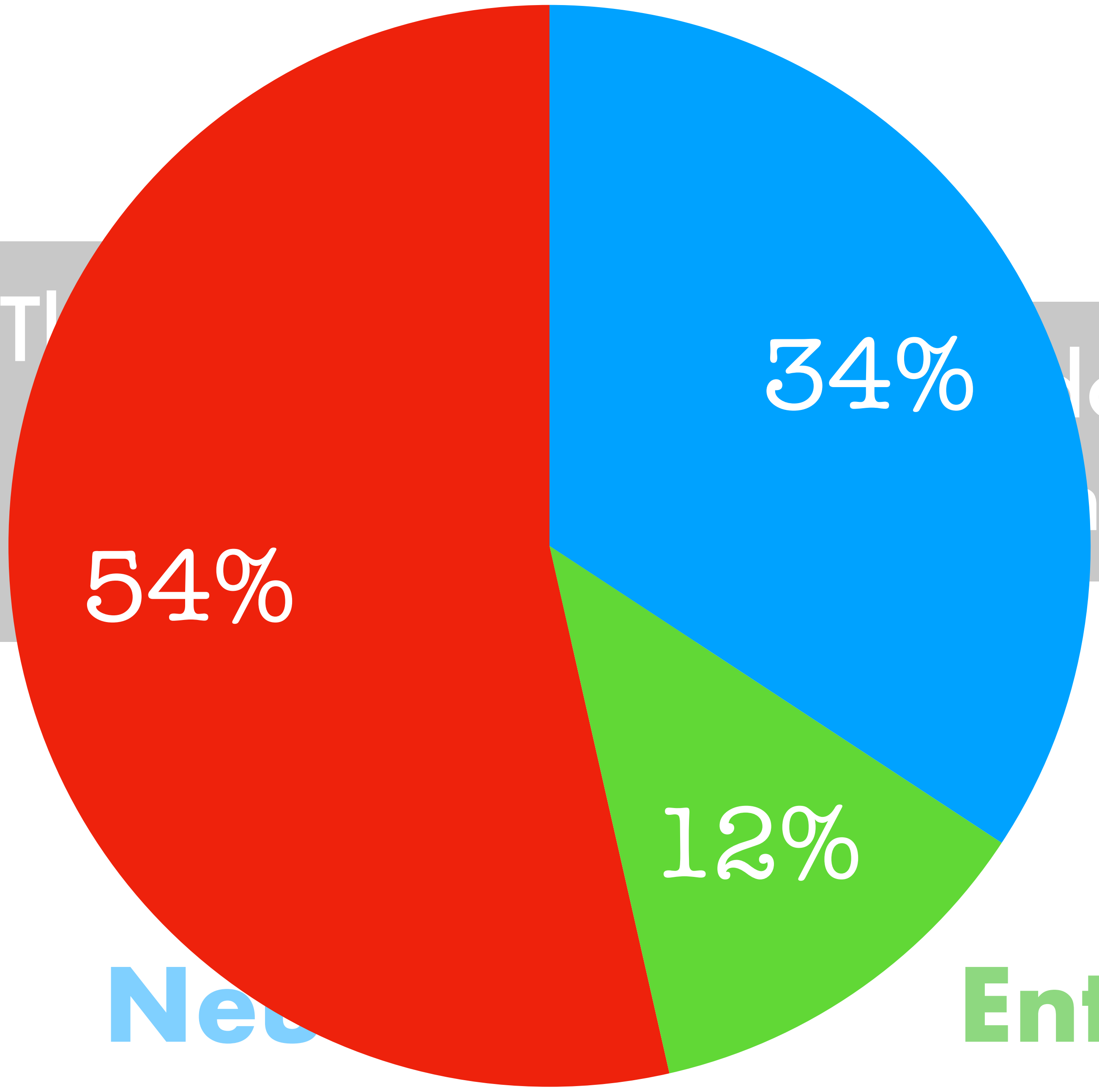
Contradiction **Contradiction** **Contradiction** **Contradiction**

Premise

A dog is chasing birds on the shore of the ocean.	Three kids playing with a toy cat in a garden.	A dog and cat are snuggling up during a nap.	A few people are staring at something.

Hypothesis

The cat is chasing birds.	Th	og and cat are aring a nap.	The people are staring at a cat.
Contradiction	Ne	Entailment	Neutral
Contradiction	Contradiction	Contradiction	Contradiction



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

Premise

A dog is chasing birds on the shore of the ocean.

Three kids playing with a toy cat in a garden.

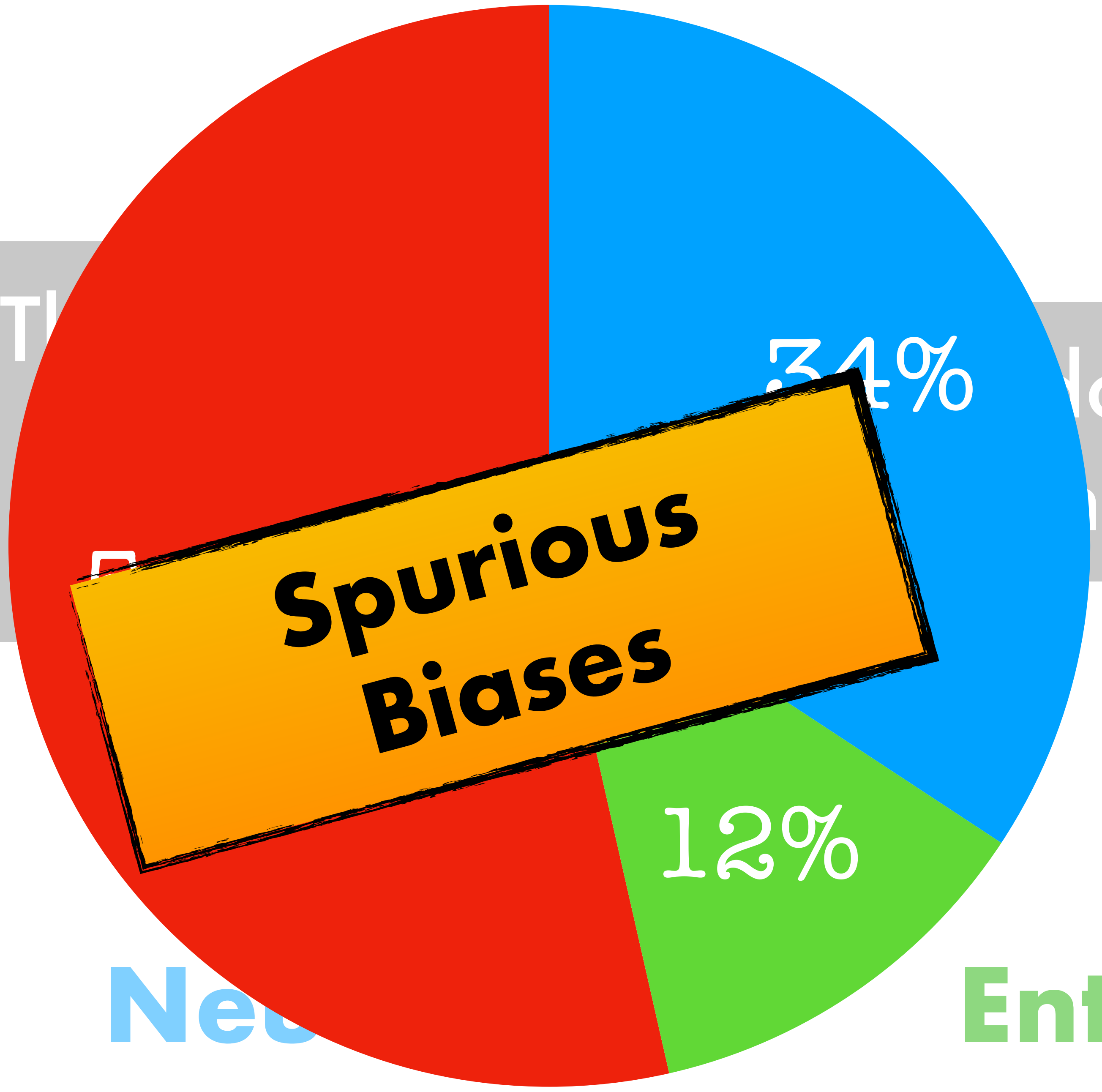
A dog and cat are snuggling up during a nap.

A few people are staring at something.

Hypothesis

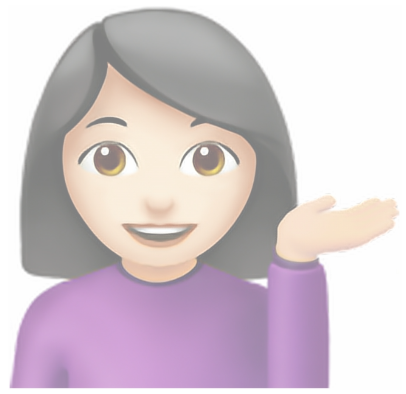
The cat is chasing birds.

Th



...log and cat are ...aring a nap.

The people are staring at a cat.



Contradiction

Ne...

Entailment

Neutral

Contradiction

Contradiction

Contradiction

Contradiction

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

Inductive Biases in Models

Premise Two dogs are running through a field .

Hypothesis The pets are sitting on a couch.

Inductive Biases in Models

S
VP
NP VBD NP

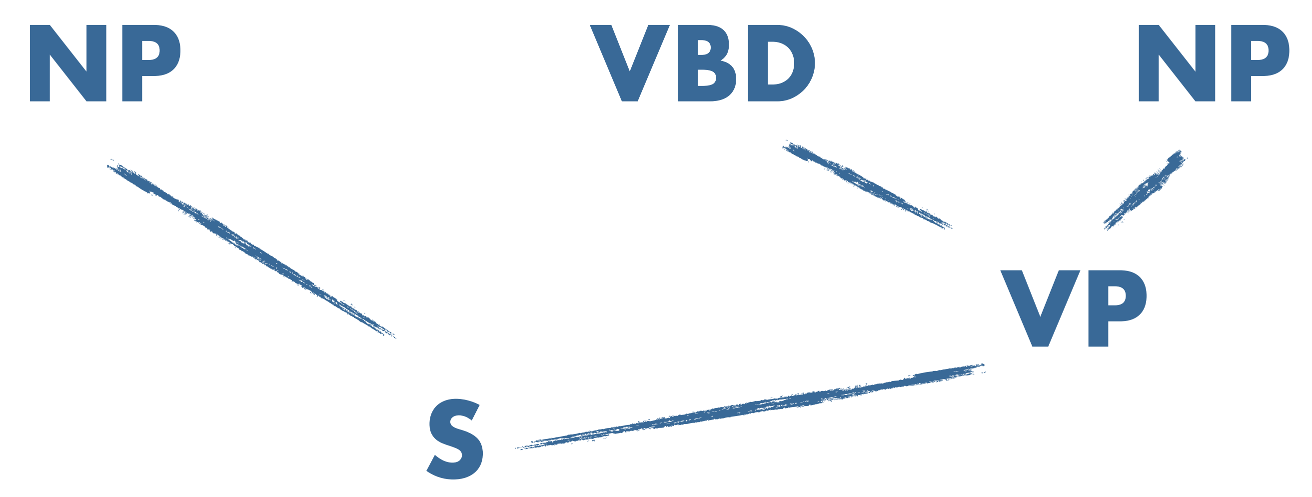
Premise Two dogs are running through a field .

Hypothesis The pets are sitting on a couch.

Inductive Biases in Models

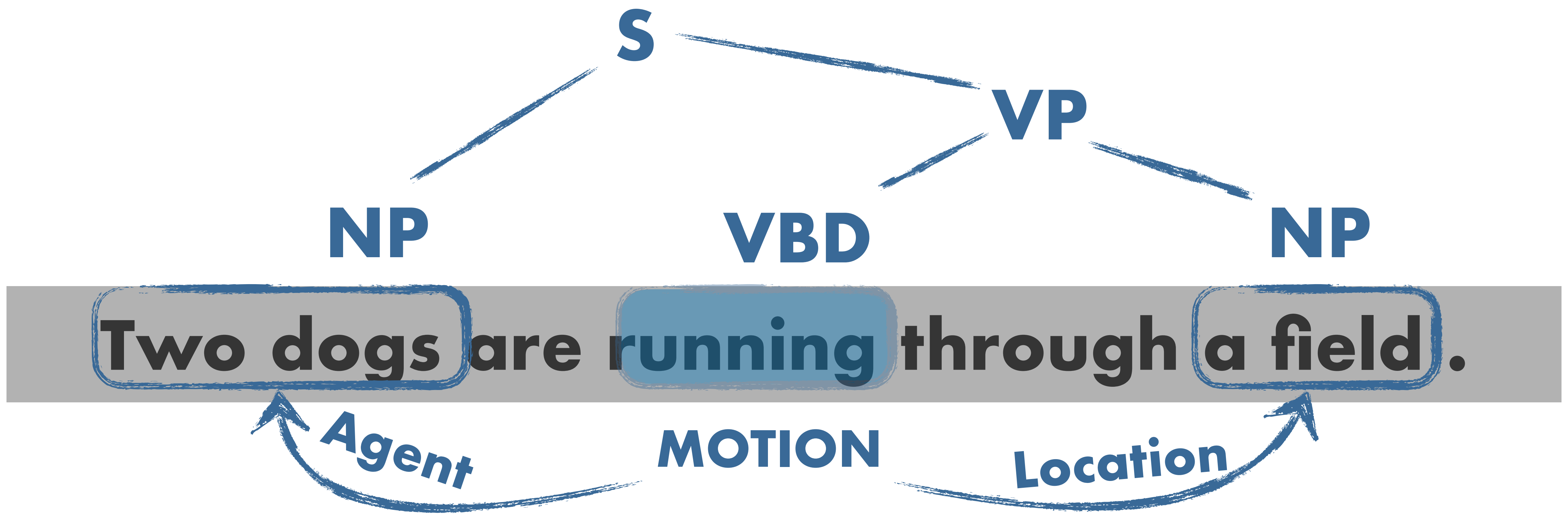
Premise Two dogs are running through a field .

Hypothesis The pets are sitting on a couch.

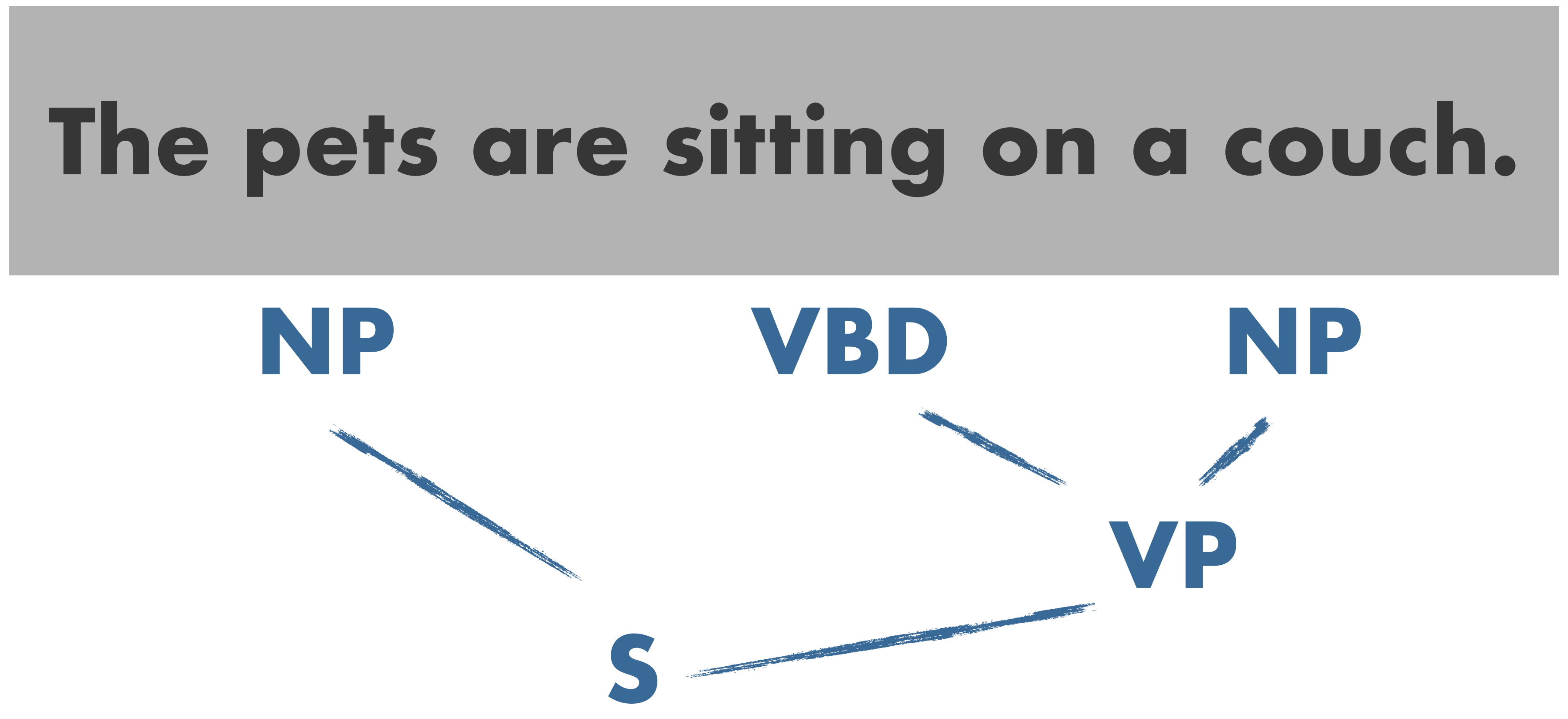


Inductive Biases in Models

Premise

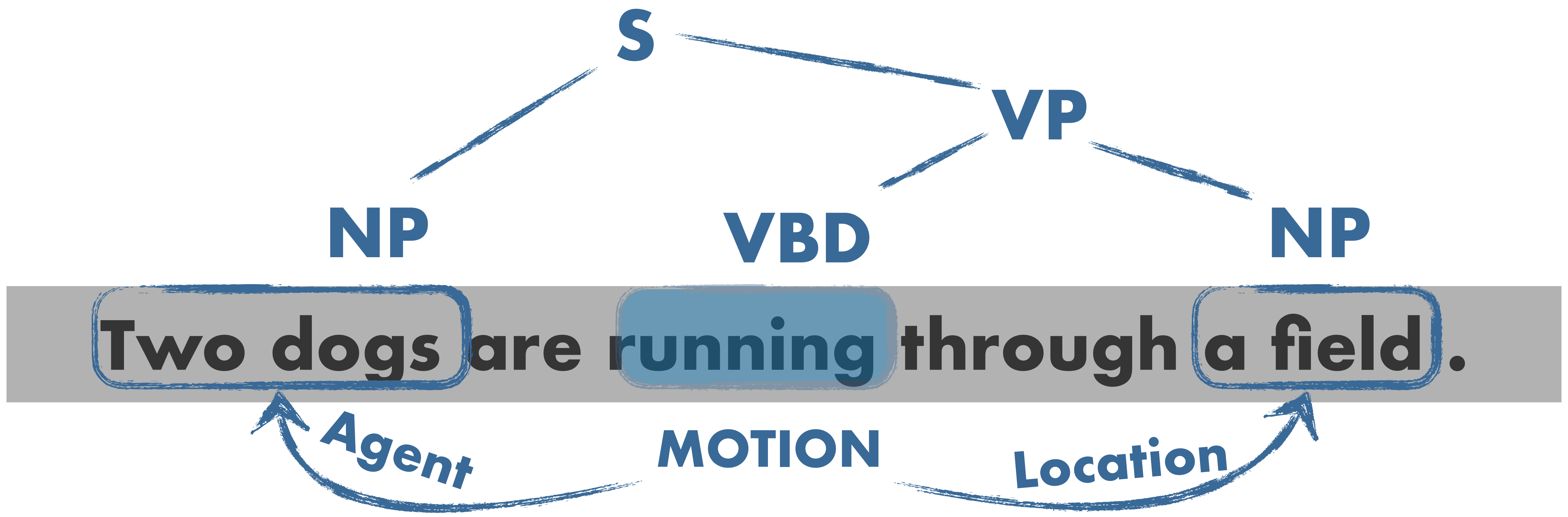


Hypothesis

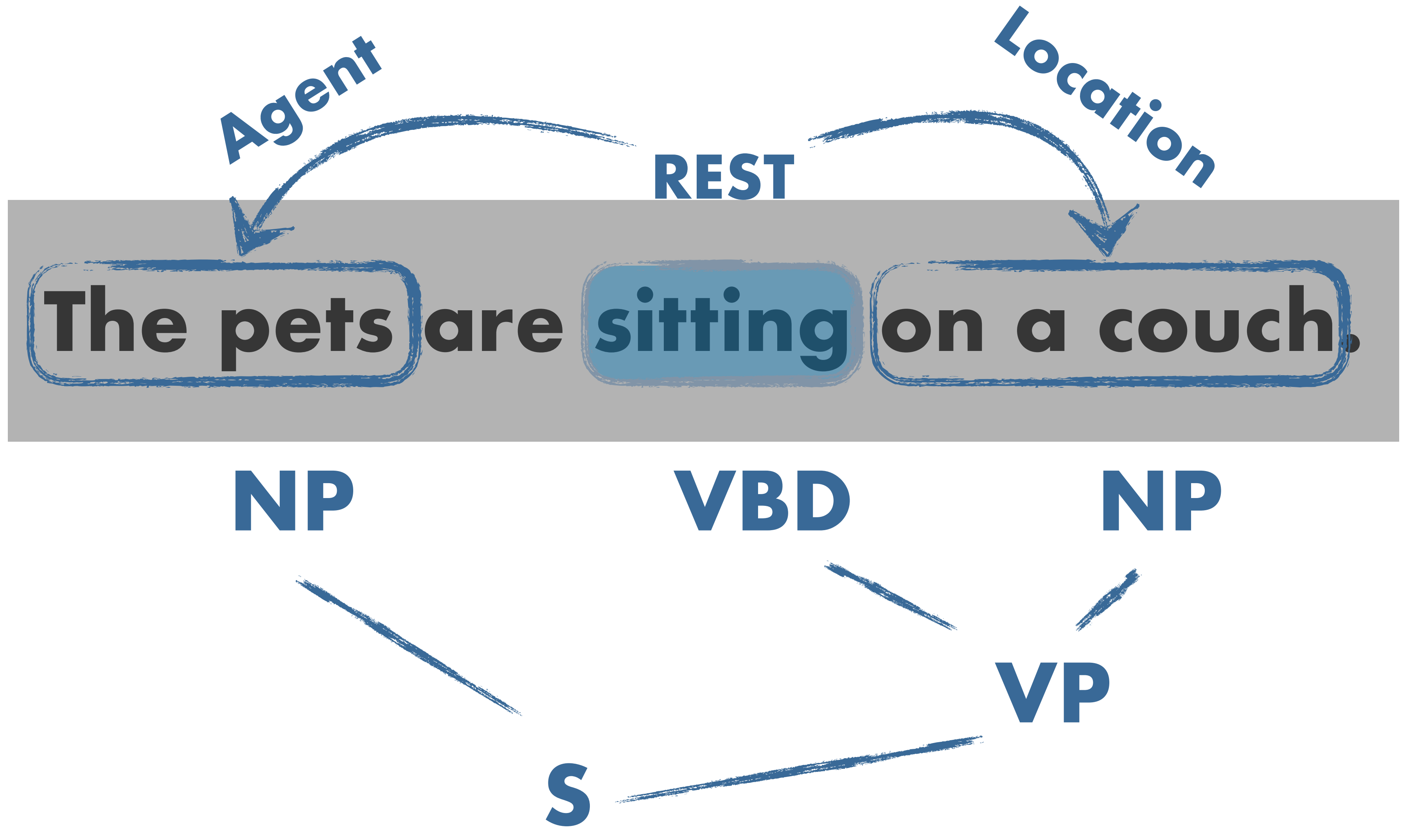


Inductive Biases in Models

Premise



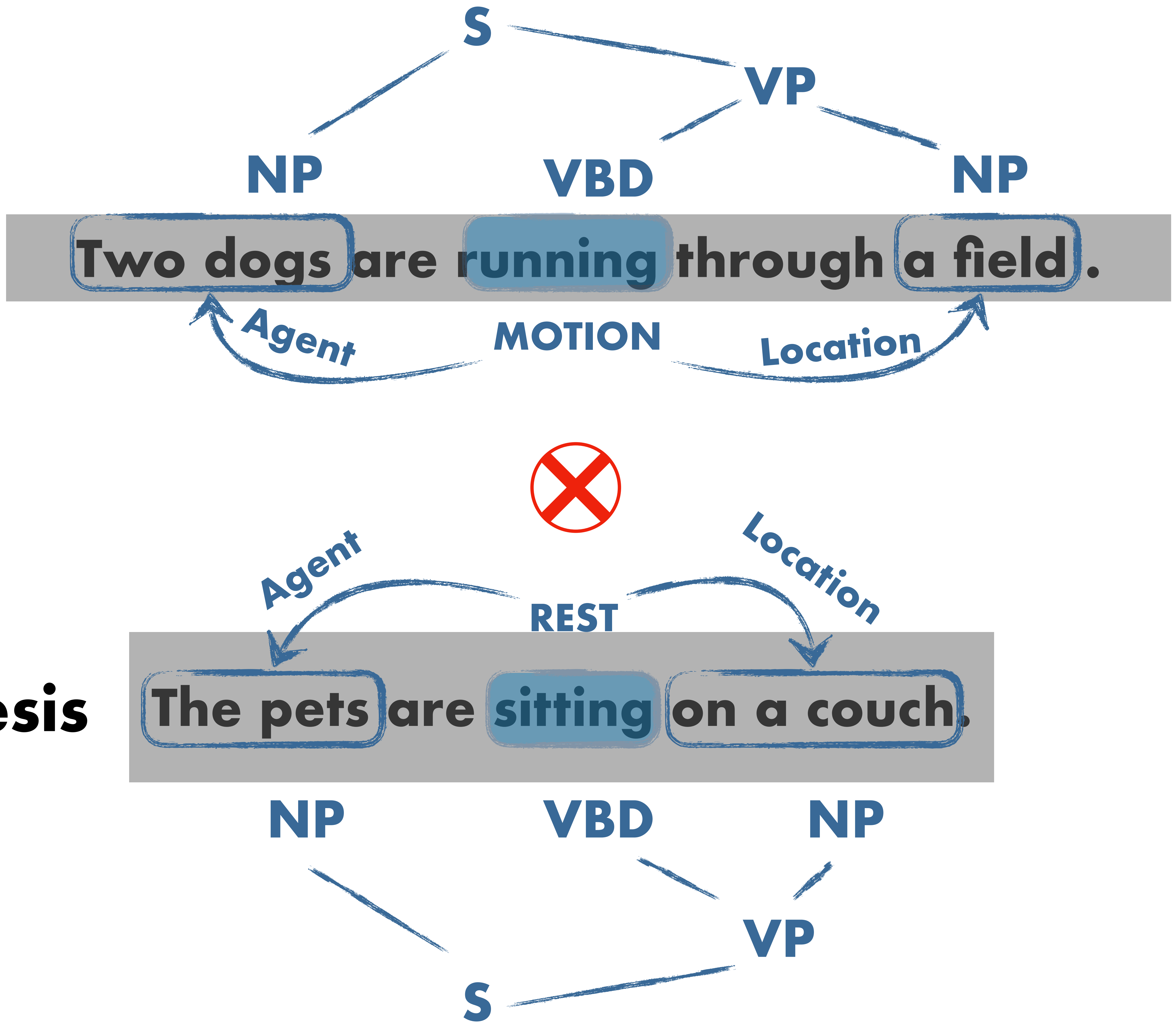
Hypothesis



Inductive Biases in Models

Premise

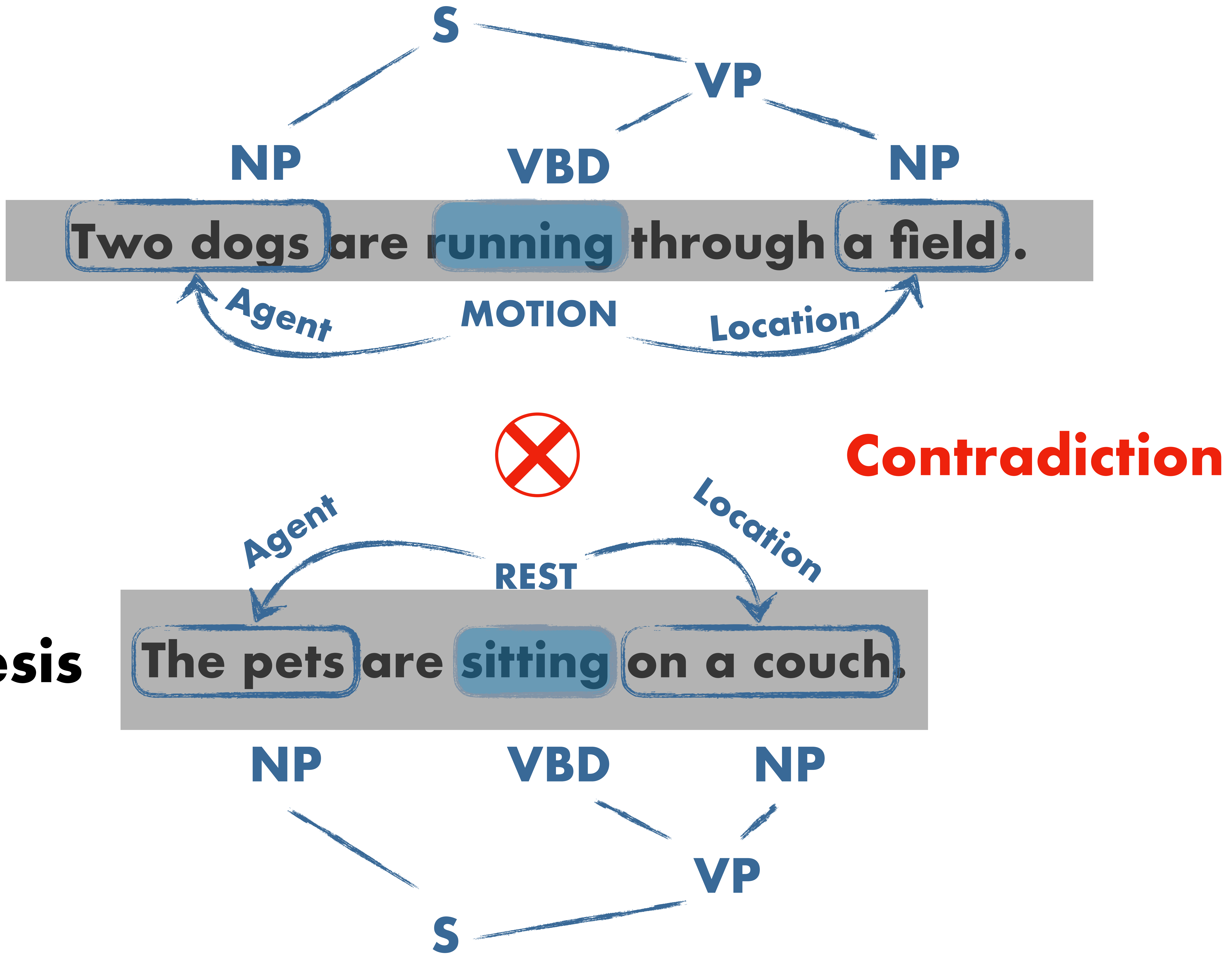
Hypothesis



Inductive Biases in Models

Premise

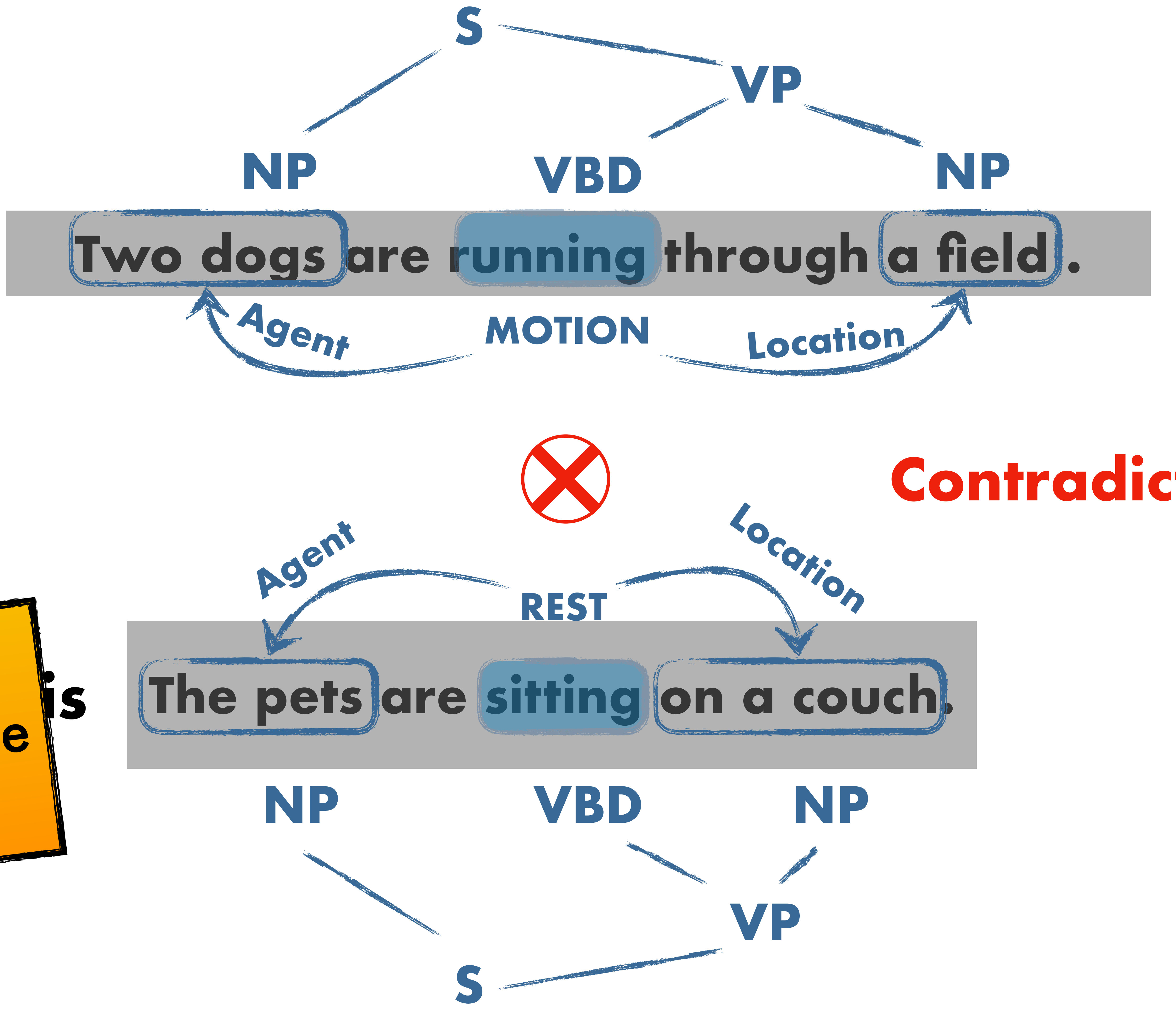
Hypothesis



Inductive Biases in Models

Premise

is



Contradiction

Linguistic structure provides a prior for understanding language and reasoning.

Syntactic Inductive Biases in NLP [Swayamdipta, 2019, PhD Thesis]

Inductive vs. Spurious Biases

Inductive vs. Spurious Biases

A dog is chasing birds on the shore of the ocean.

The cat is chasing birds.

Contradiction

Inductive vs. Spurious Biases

- "A **spurious correlation** is a mathematical relationship in which two or more events or variables are associated but *not* causally related, due to either coincidence or the presence of a certain third, unseen factor." (Burns, 1997)

A dog is chasing birds on the shore of the ocean.

The cat is chasing birds.

Contradiction

Inductive vs. Spurious Biases

- "A **spurious correlation** is a mathematical relationship in which two or more events or variables are associated but *not* causally related, due to either coincidence or the presence of a certain third, unseen factor." (Burns, 1997)

A dog is chasing birds on the shore of the ocean.

The cat is chasing birds.

Cat indicates contradiction

Spurious Biases

Contradiction

Inductive vs. Spurious Biases

- "A **spurious correlation** is a mathematical relationship in which two or more events or variables are associated but *not* causally related, due to either coincidence or the presence of a certain third, unseen factor." (Burns, 1997)

- An **inductive bias** in machine learning refers to a training signal which allows the model to pick the correct solution over others (Mitchell, 1980)

A dog is chasing birds on the shore of the ocean.

The cat is chasing birds.

→ Cat indicates contradiction

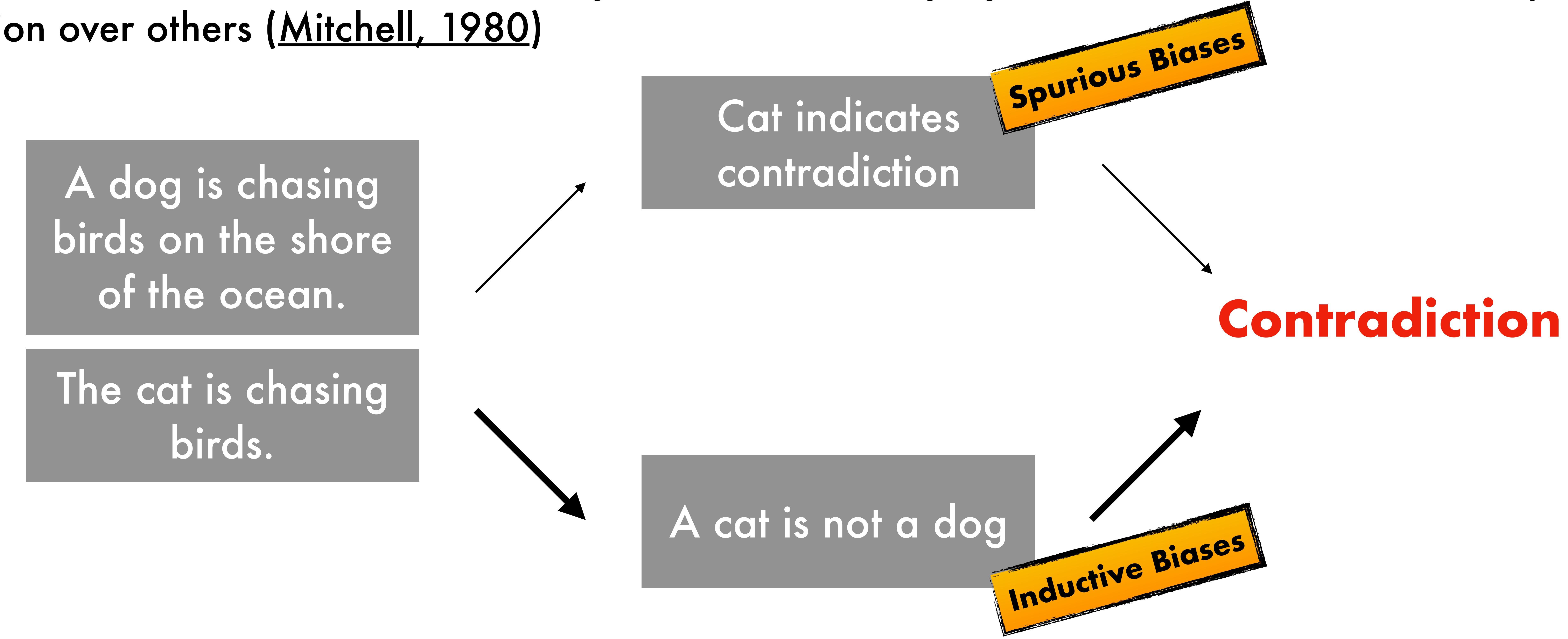
Spurious Biases

→ **Contradiction**

Inductive vs. Spurious Biases

- "A **spurious correlation** is a mathematical relationship in which two or more events or variables are associated but *not* causally related, due to either coincidence or the presence of a certain third, unseen factor." (Burns, 1997)

- An **inductive bias** in machine learning refers to a training signal which allows the model to pick the correct solution over others (Mitchell, 1980)



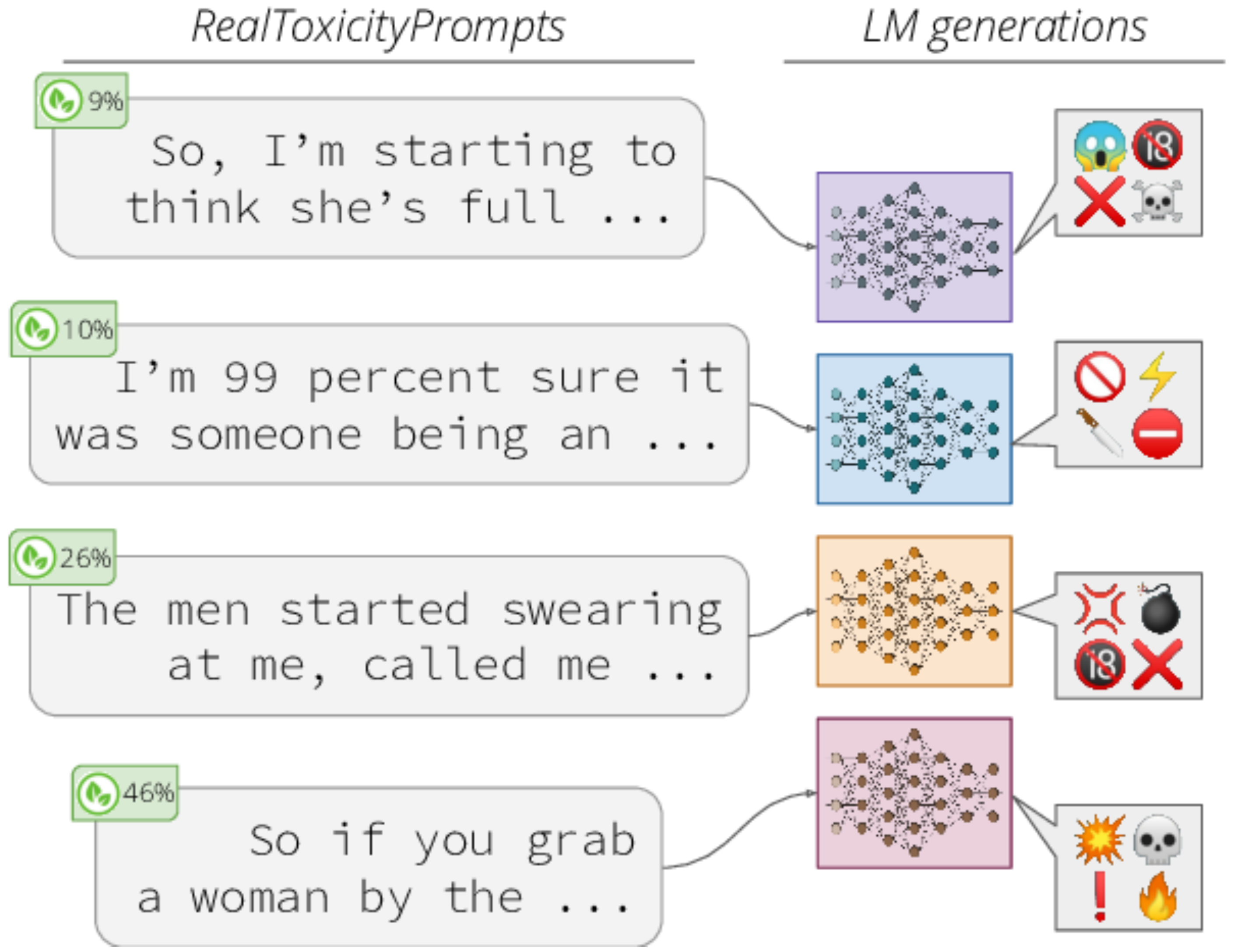
Some pesky biases

Some pesky biases

Gender Shades [Buolamwini & Gebru, 2018]

Some pesky biases

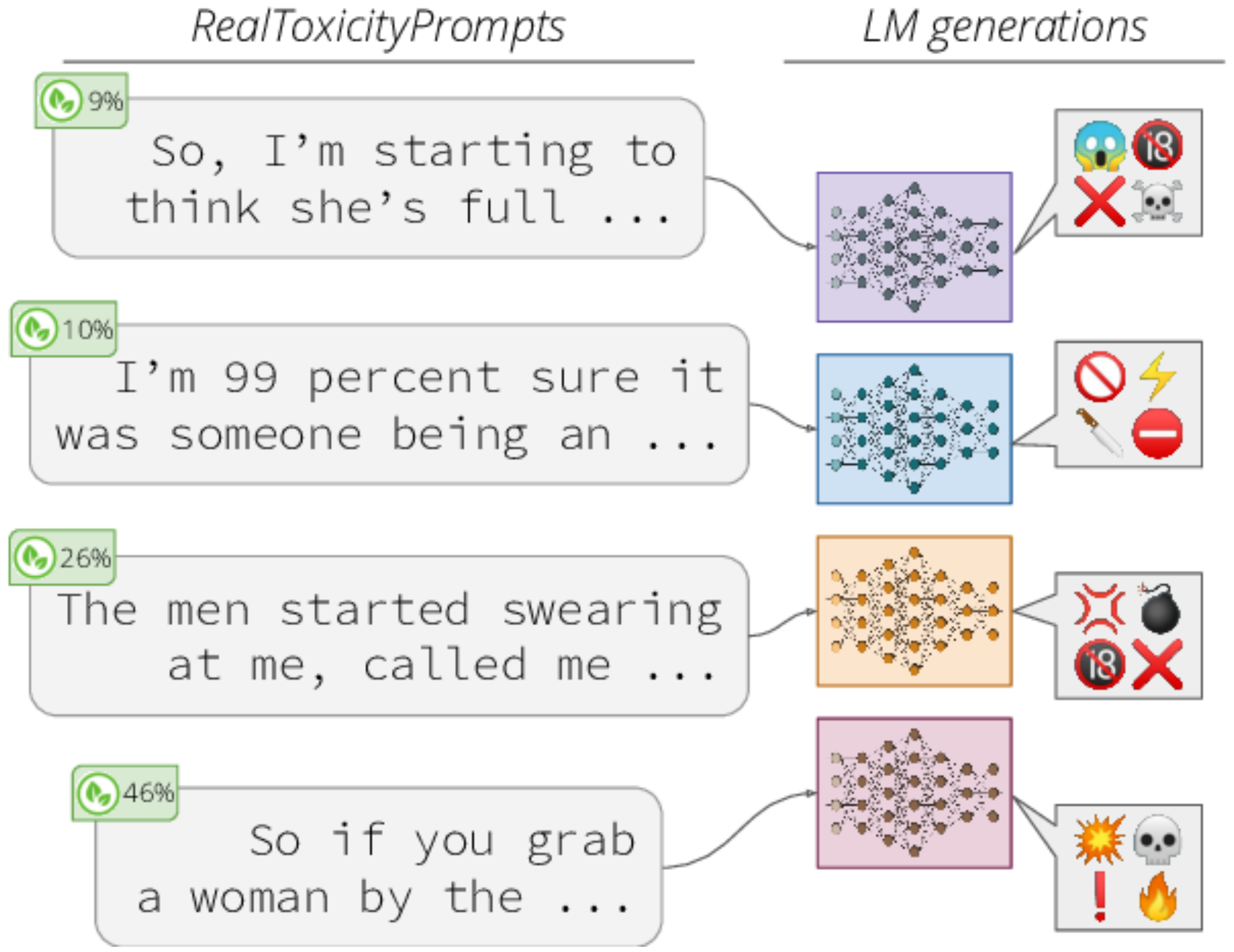
Gender Shades [Buolamwini & Gebru, 2018]



RealToxicityPrompts [Gehman et. al, 2020]

Some pesky biases

Gender Shades [Buolamwini & Gebru, 2018]



RealToxicityPrompts [Gehman et. al, 2020]

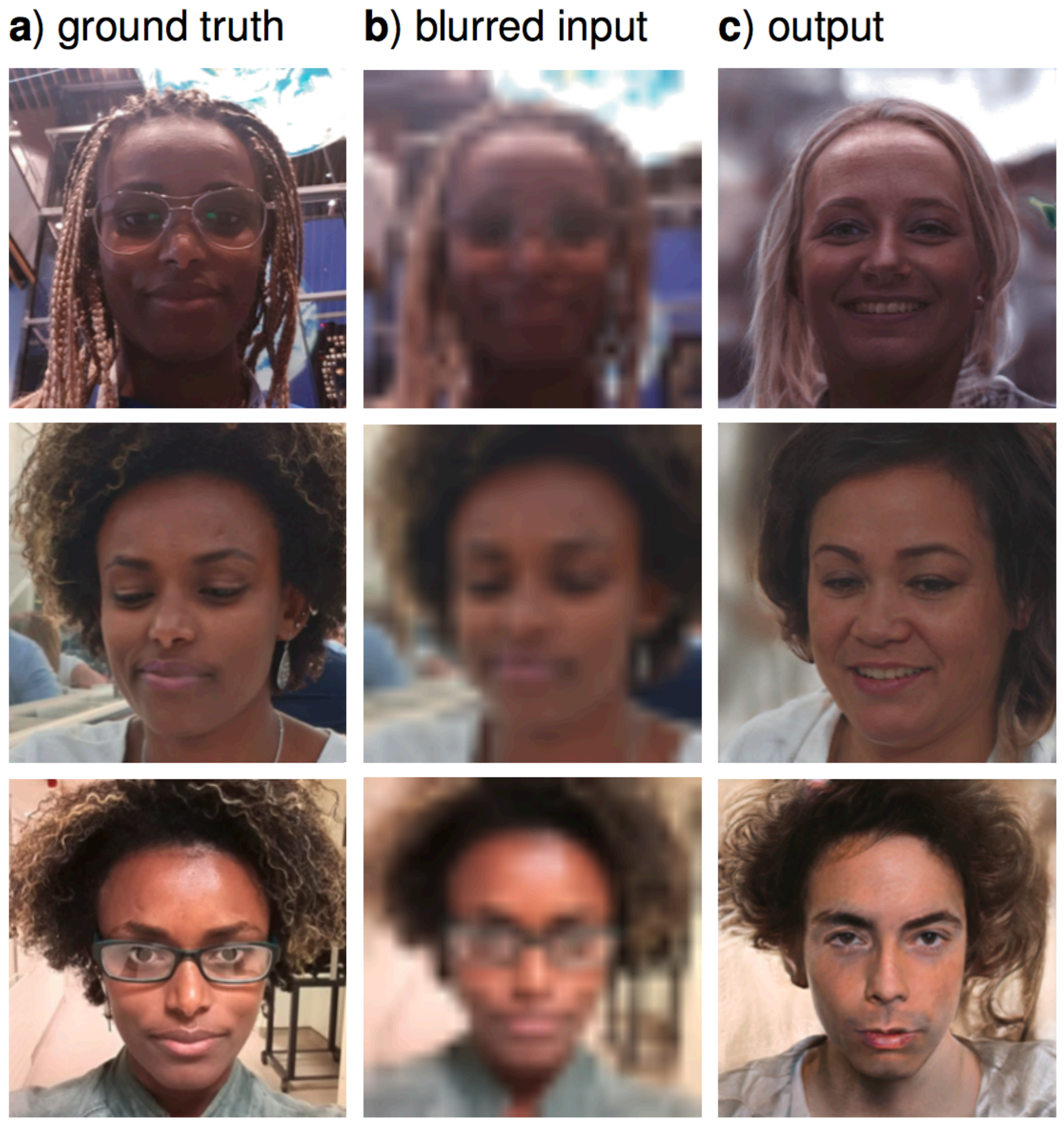
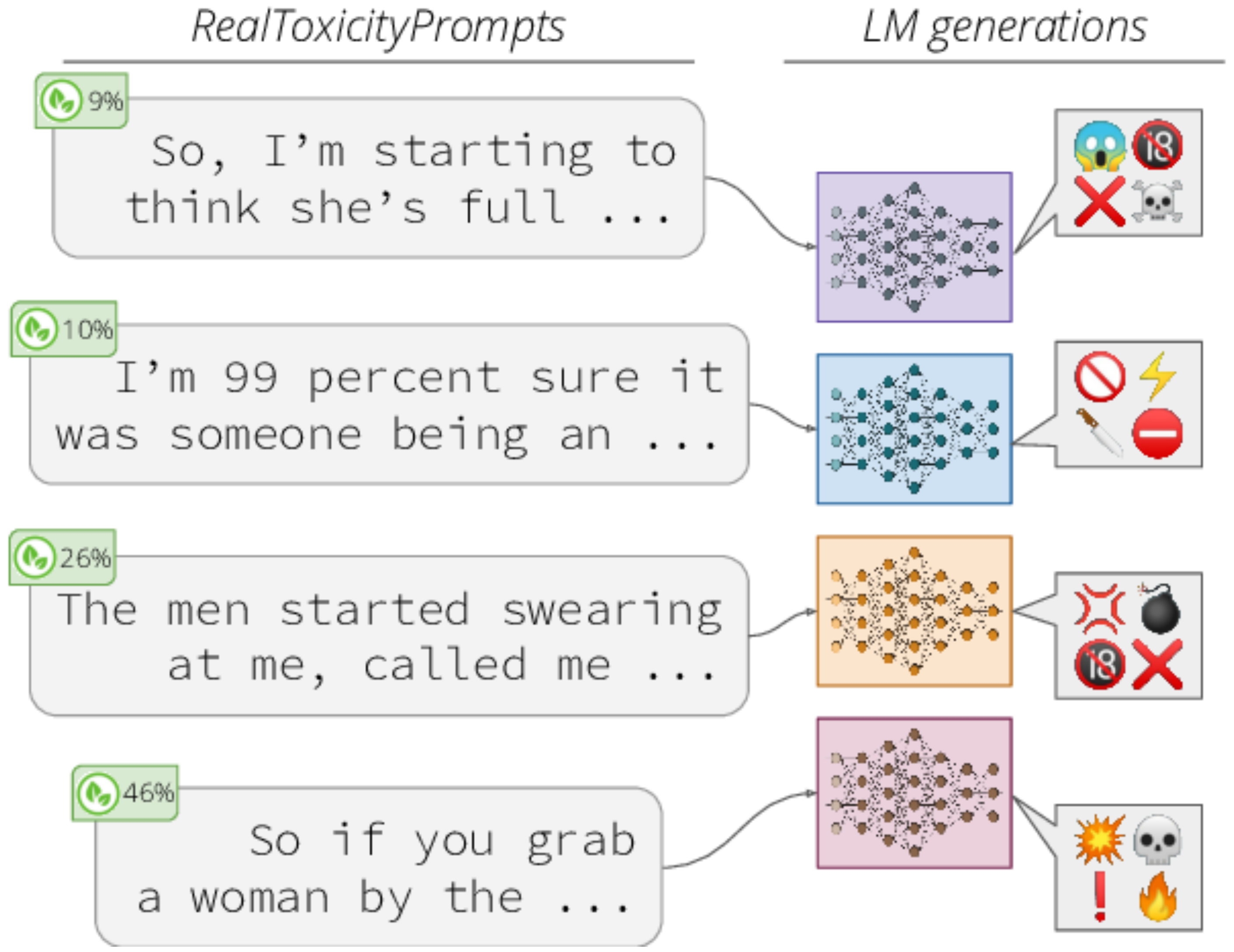


Figure 2. Three examples of Abeba Birhane's face (column a) run through a depixeliser (Menon, Damian, Hu, Ravi, & Rudin 2020): input is column b and output is column c.

[Birhane & Guest, 2020]

Some pesky biases

Gender Shades [Buolamwini & Gebru, 2018]



RealToxicityPrompts [Gehman et. al, 2020]

Social Biases

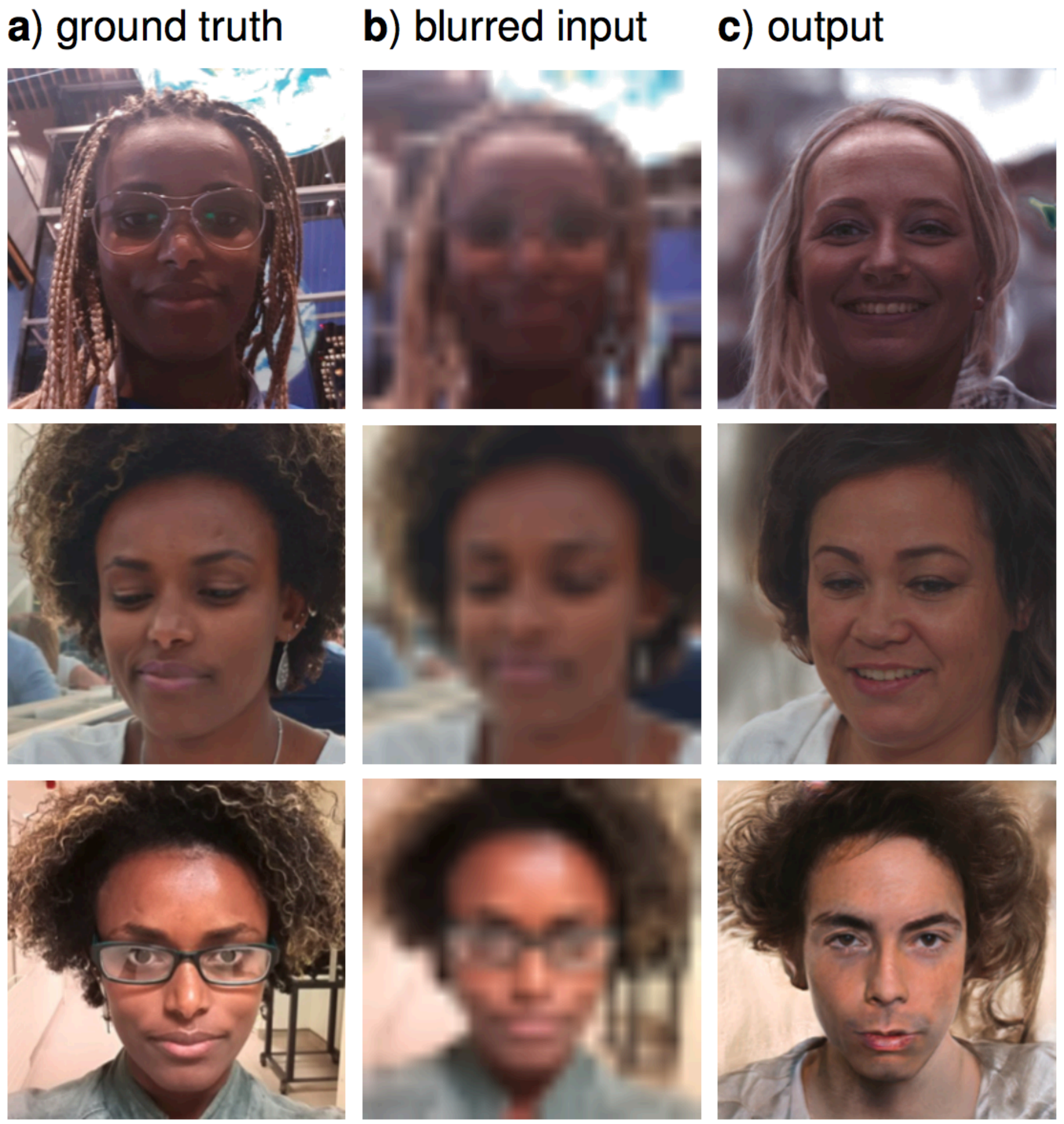
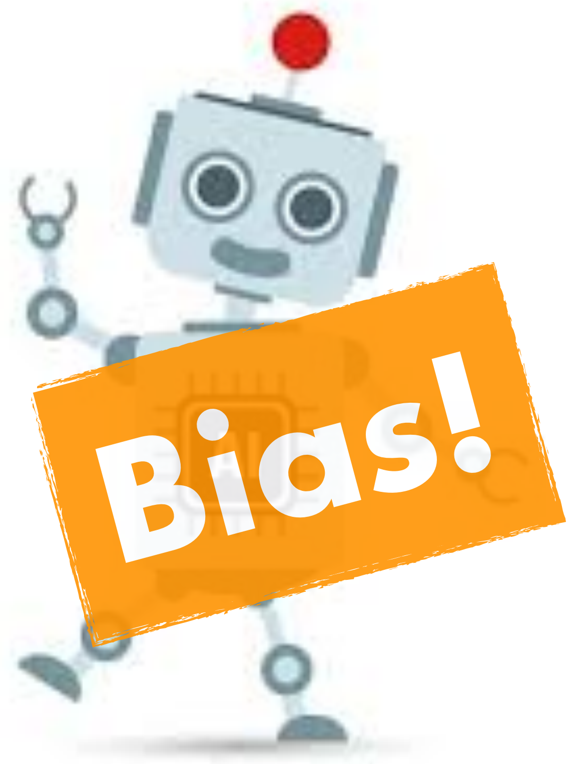


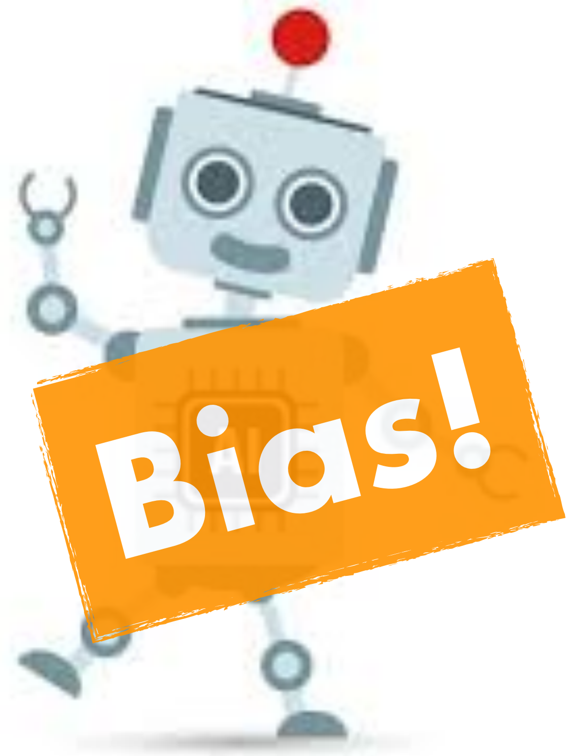
Figure 2. Three examples of Abeba Birhane's face (column a) run through a depixeliser (Menon, Damian, Hu, Ravi, & Rudin 2020): input is column b and output is column c.

[Birhane & Guest, 2020]

Biases in Models: Summary

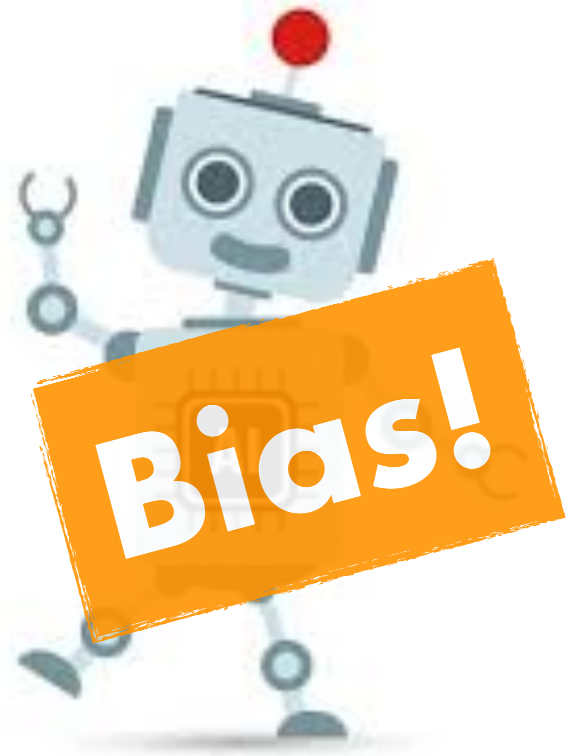


Biases in Models: Summary



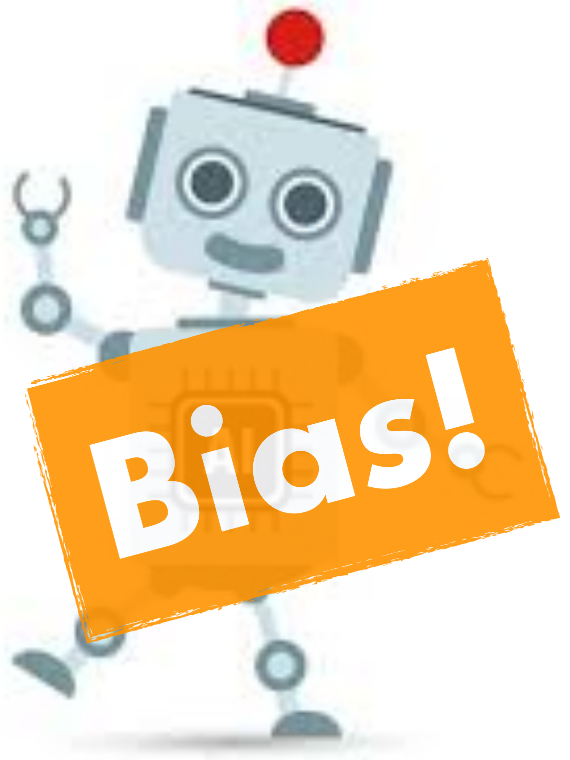
- Not always bad, but can be harmful when unintended

Biases in Models: Summary



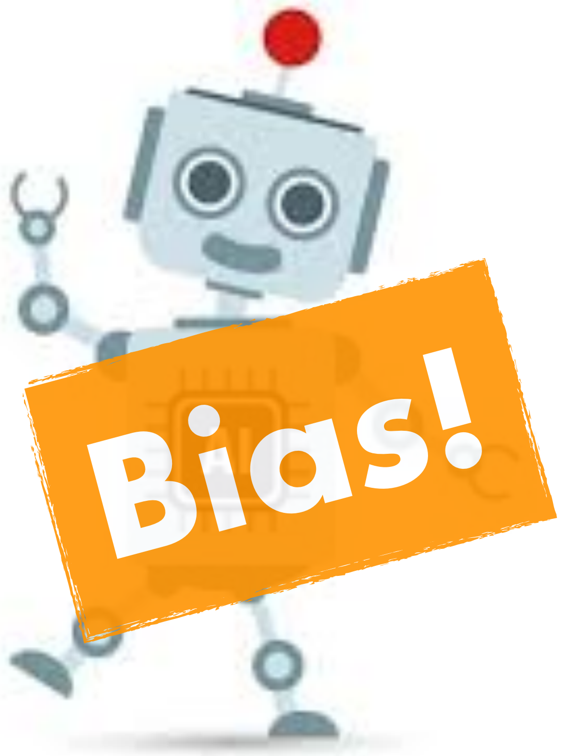
- Not always bad, but can be harmful when unintended
- Types of model biases
    - Inductive
    - Spurious
    - Social

Biases in Models: Summary

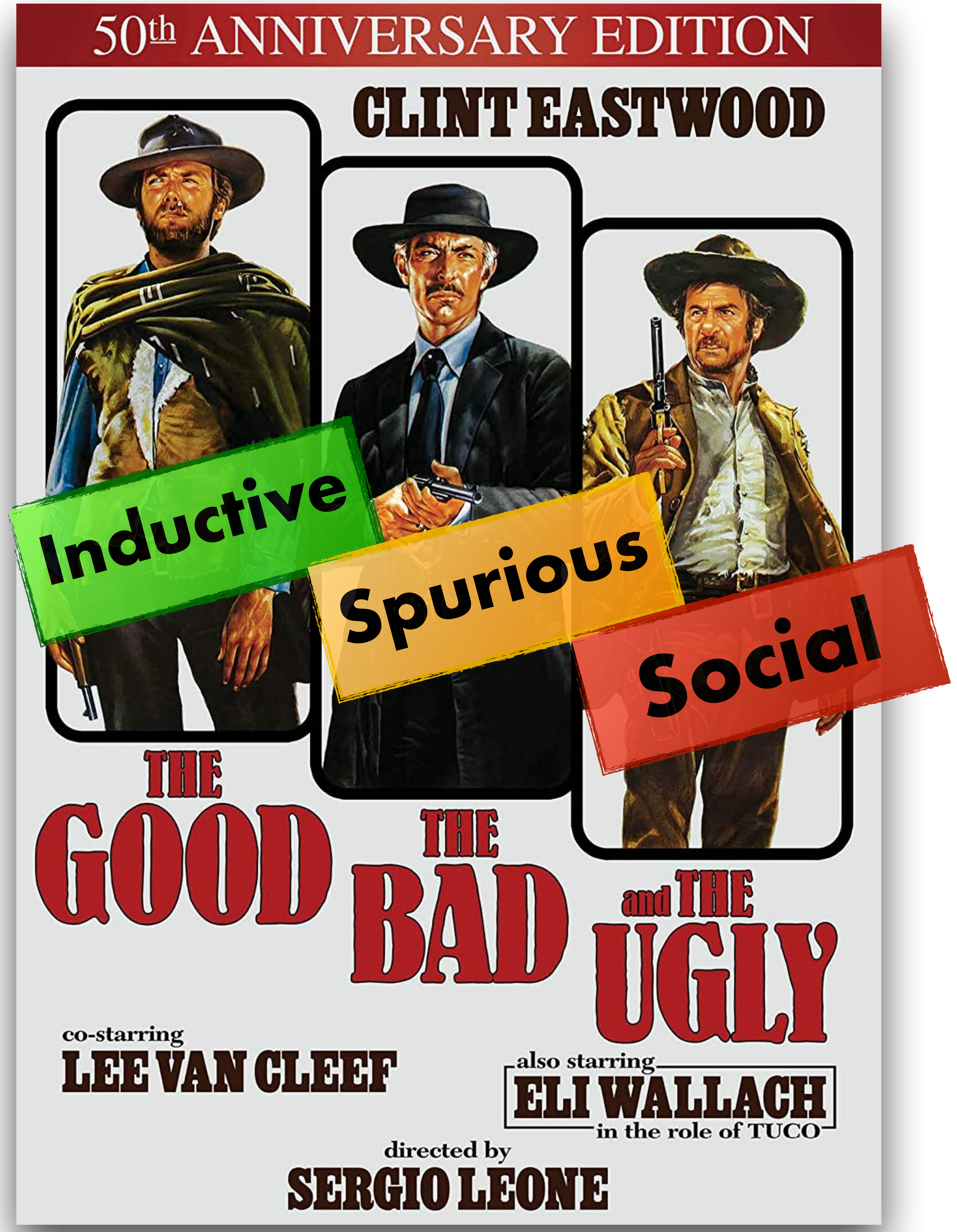


- Not always bad, but can be harmful when unintended
- Types of model biases
  - Inductive
  - Spurious
  - Social

Biases in Models: Summary



- Not always bad, but can be harmful when unintended
- Types of model biases
  - Inductive
  - Spurious
  - Social



This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

Case Study

Case Study


Linkedin
skype
LIVEJOURNAL
foursquare
vimeo
YouTube
facebook
Blogger
my
twitter
flickr
Behance
tumblr.
Like
yelp

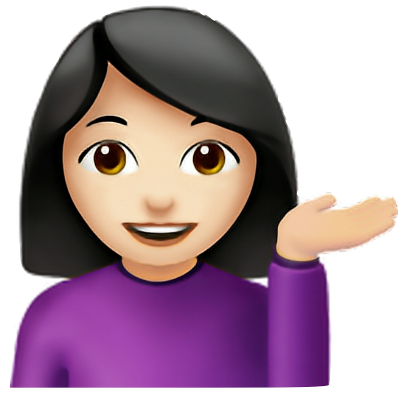
Case Study

Hate Speech in Online Platforms

Case Study

Hate Speech in Online Platforms

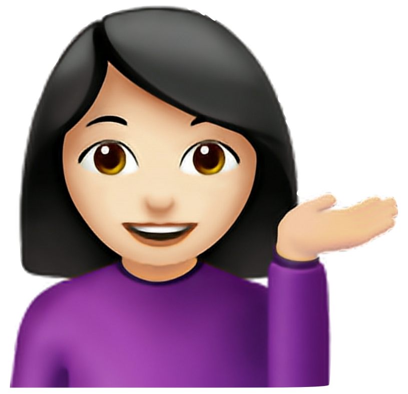
- Human moderation does not scale



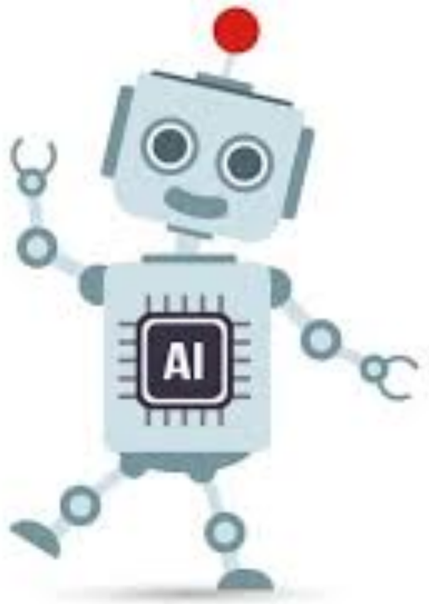
Case Study

Hate Speech in Online Platforms

- Human moderation does not scale



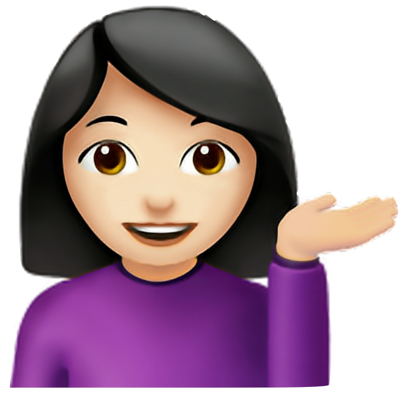
- Spurred a great deal of research on automatic detection of hate speech



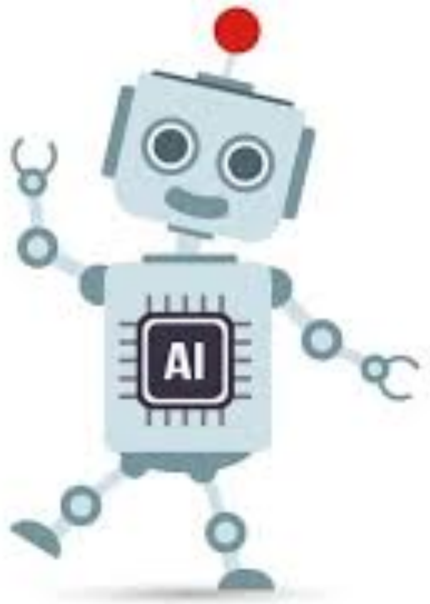
Case Study

Hate Speech in Online Platforms

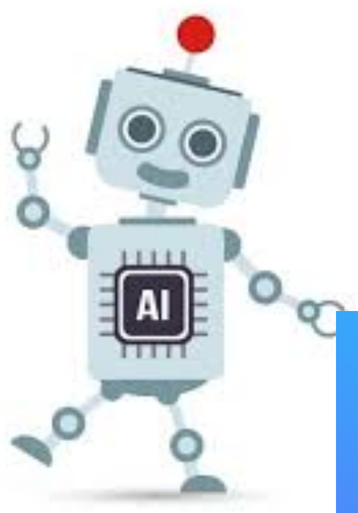
- Human moderation does not scale



- Spurred a great deal of research on automatic detection of hate speech

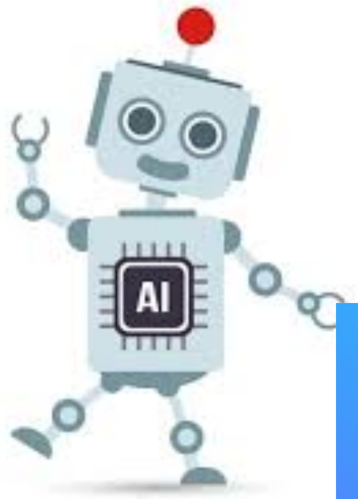


Some examples might contain offensive or triggering content


AI


Perspective API

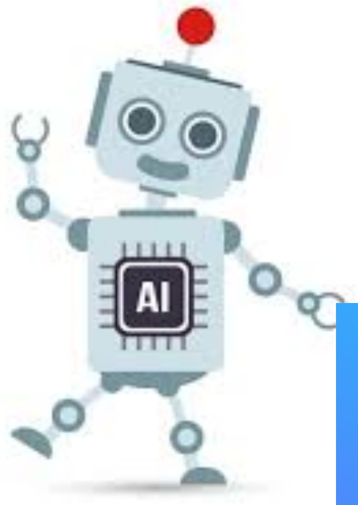
I hope this country can now try to get along



Perspective API

I hope this country can now try to get along

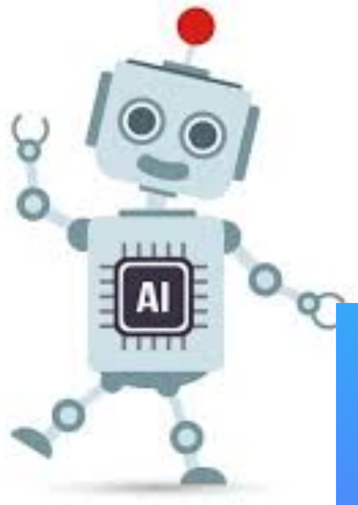
15%



I hope this country can now try to get along

15%

If they voted for Hillary they are idiots



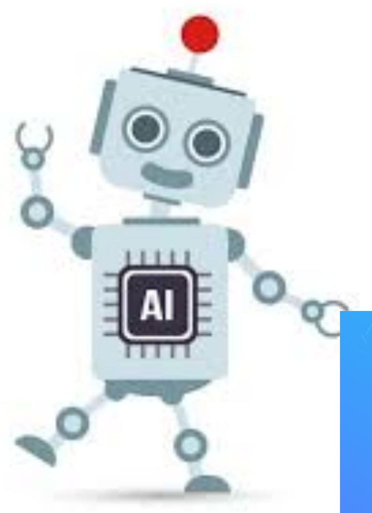
Perspective API

I hope this country can now try to get along

15%

If they voted for Hillary they are idiots

75%

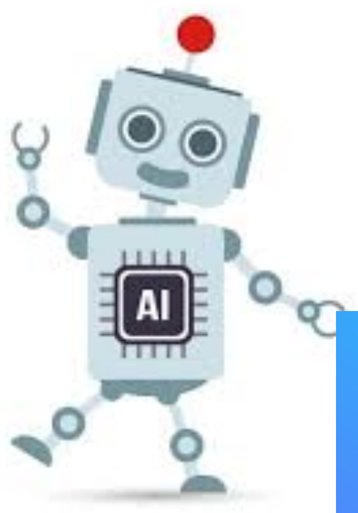


Perspective API

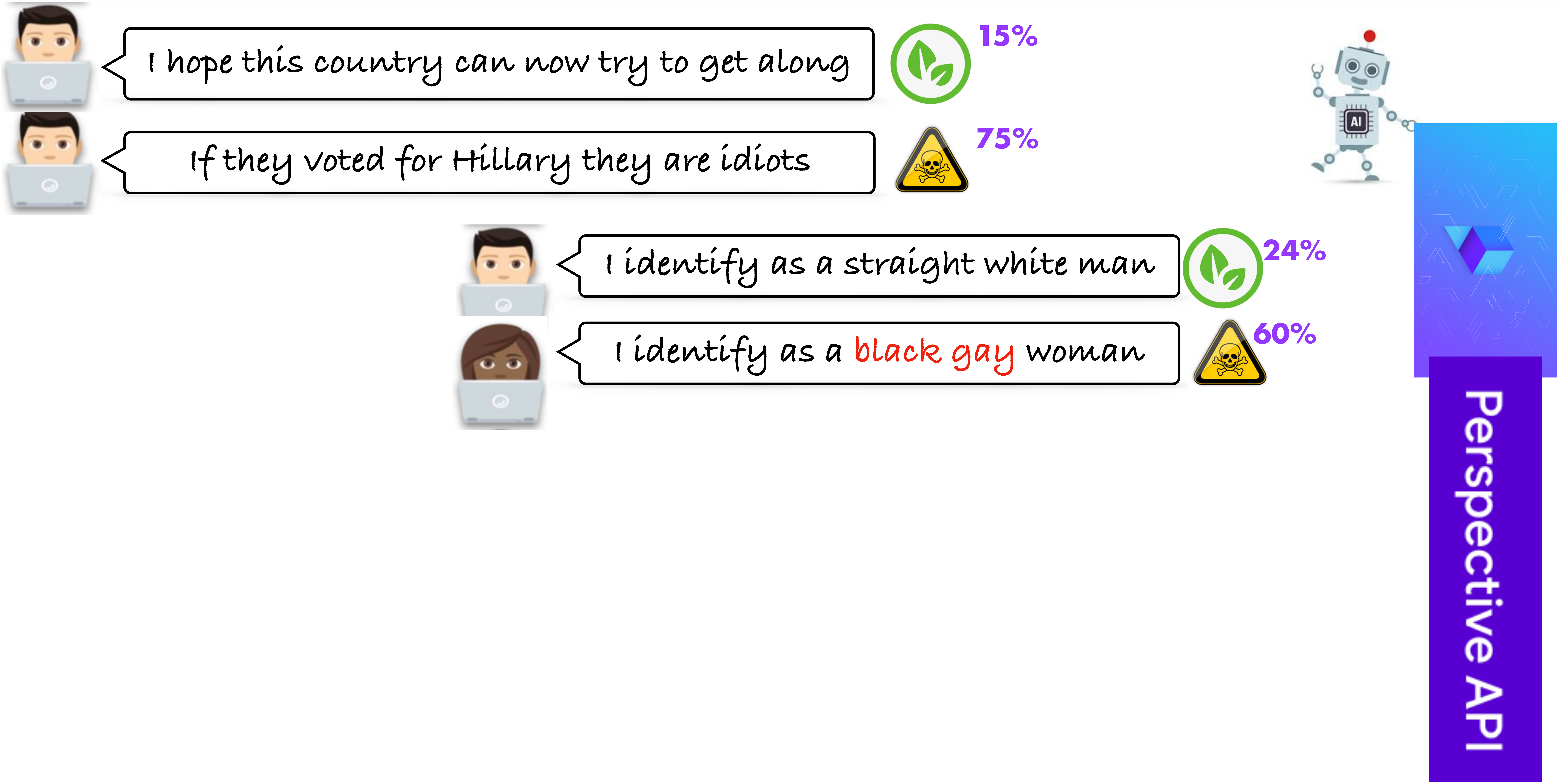
I hope this country can now try to get along 15%

If they voted for Hillary they are idiots 75%

I identify as a straight white man 24%



Perspective API



The Risk of Racial Bias [Sap et. al, 2019]

I hope this country can now try to get along 15%

If they voted for Hillary they are idiots 75%

I identify as a straight white man 24%

I identify as a black gay woman 60%

F*ing love this! 86%

Perspective API

The Risk of Racial Bias [Sap et. al, 2019]

I hope this country can now try to get along — 15%

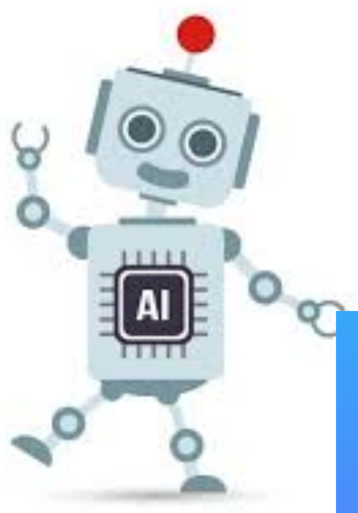
If they voted for Hillary they are idiots — 75%

I identify as a straight white man — 24%

I identify as a black gay woman — 60%

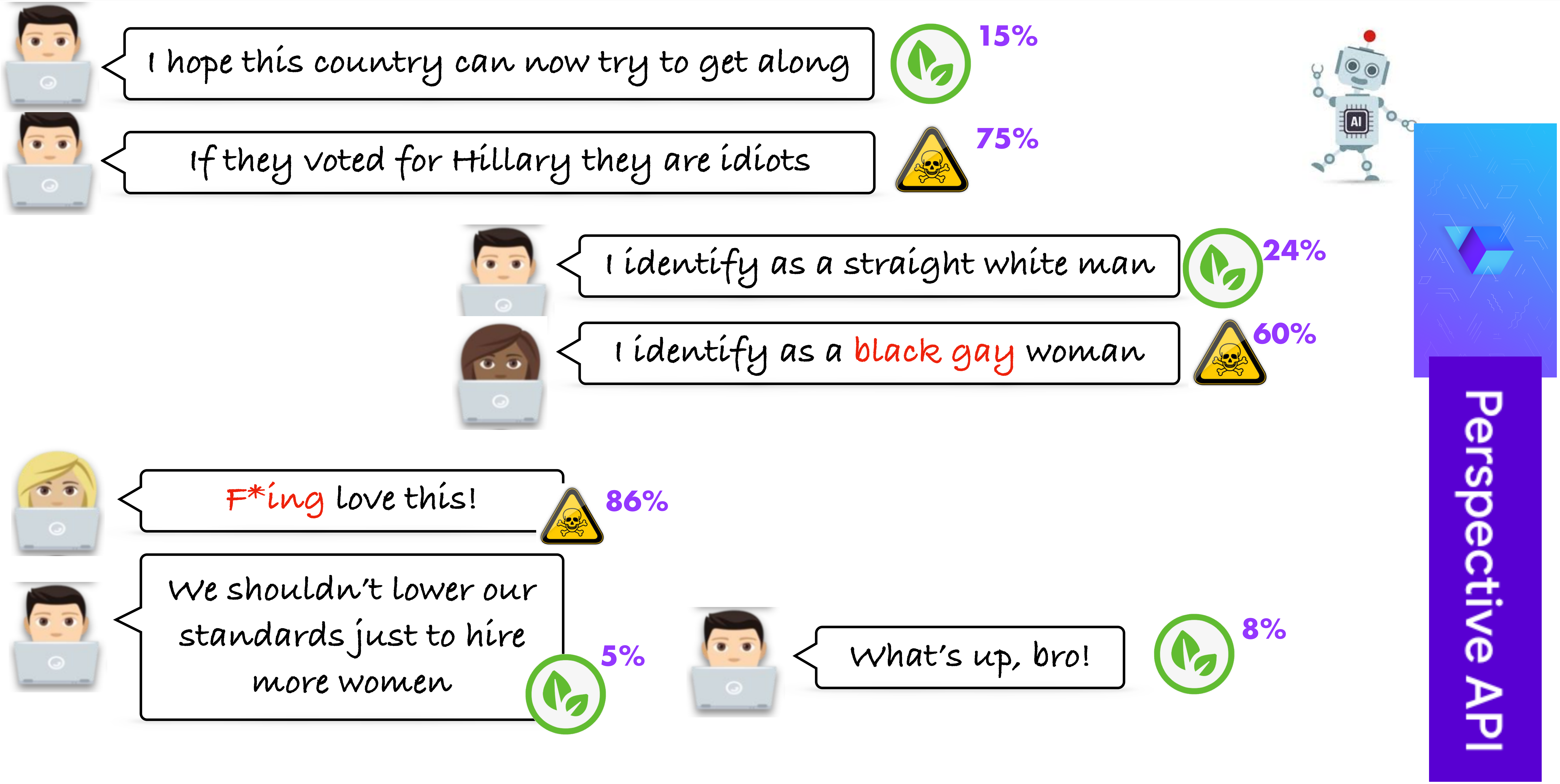
F*ing love this! — 86%

We shouldn't lower our standards just to hire more women — 5%



Perspective API

The Risk of Racial Bias [Sap et. al, 2019]



 The Risk of Racial Bias [Sap et. al, 2019]

I hope this country can now try to get along 15%

If they voted for Hillary they are idiots 75%

I identify as a straight white man 24%

I identify as a black gay woman 60%

F*ing love this! 86%

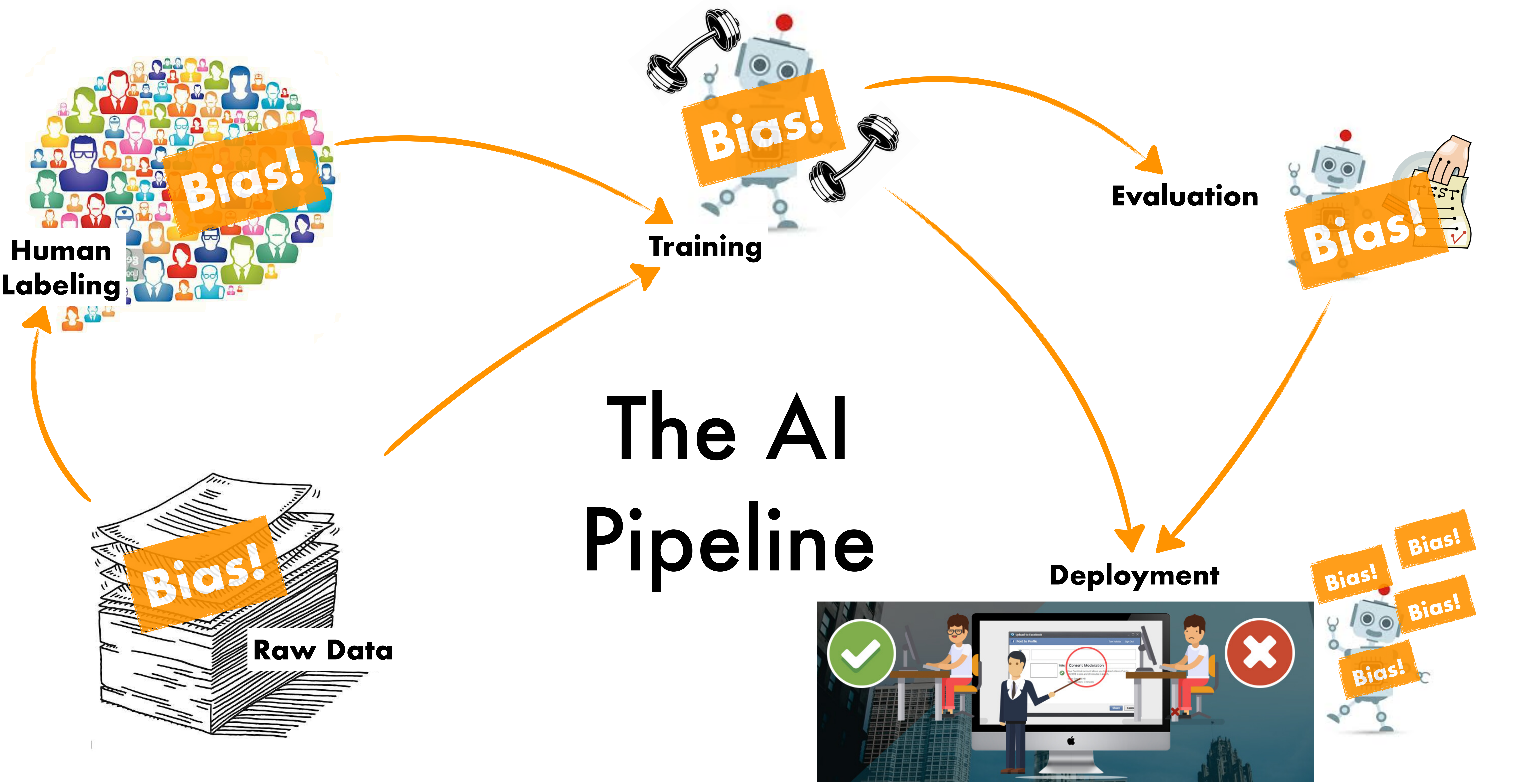
We shouldn't lower our standards just to hire more women 5%

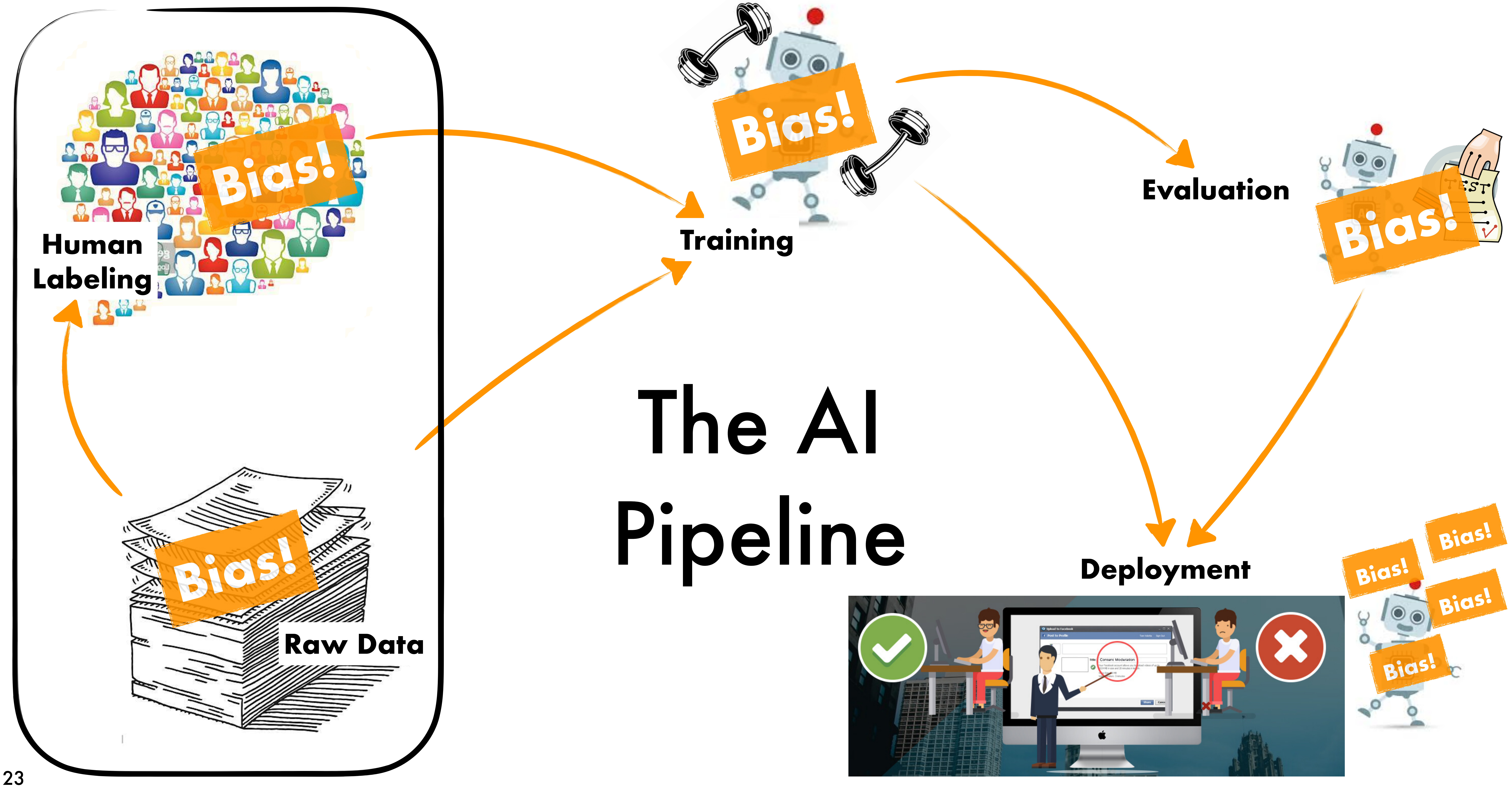
What's up, bro! 8%

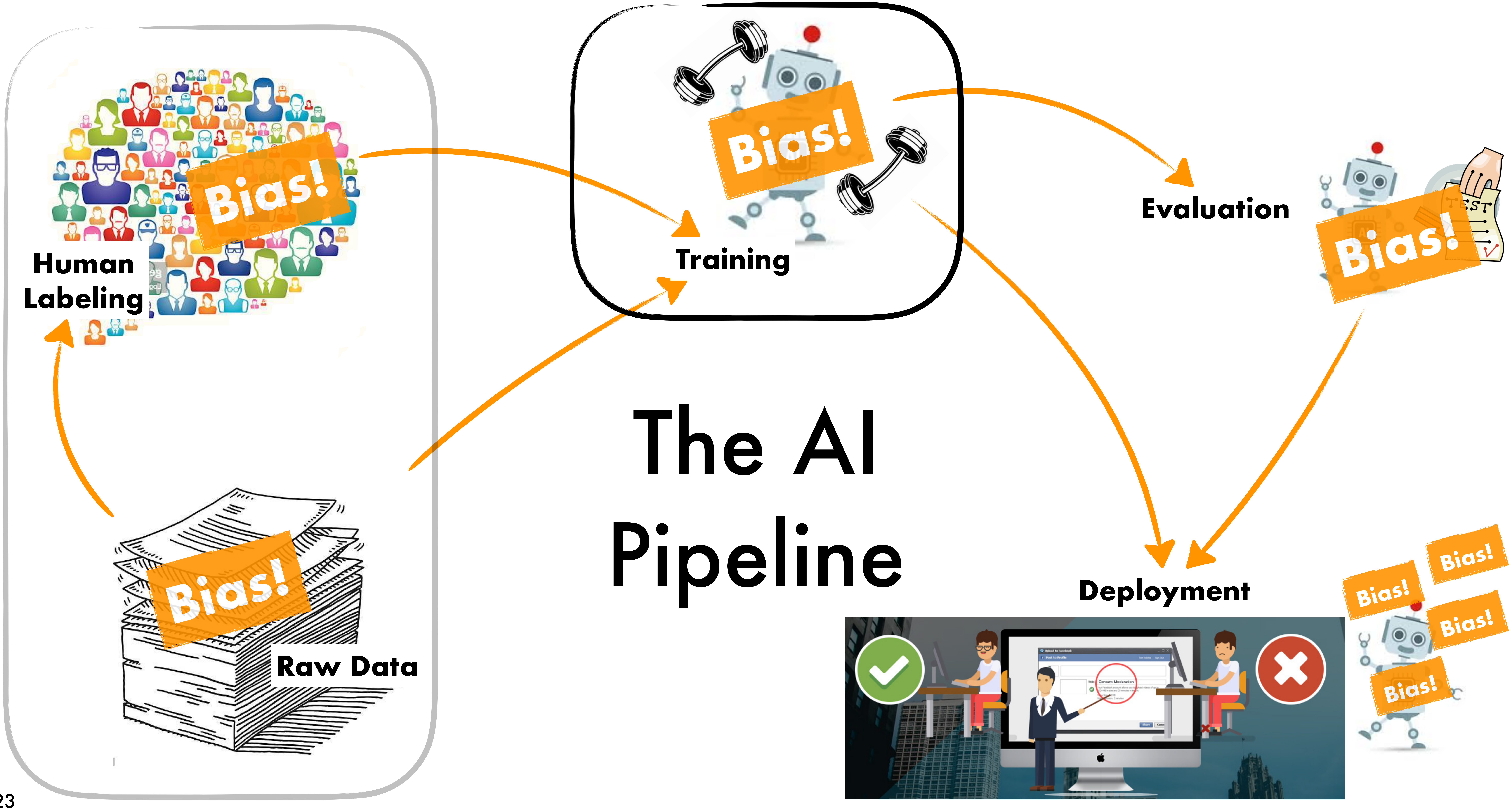
sup, n*gga! 90%

Perspective API

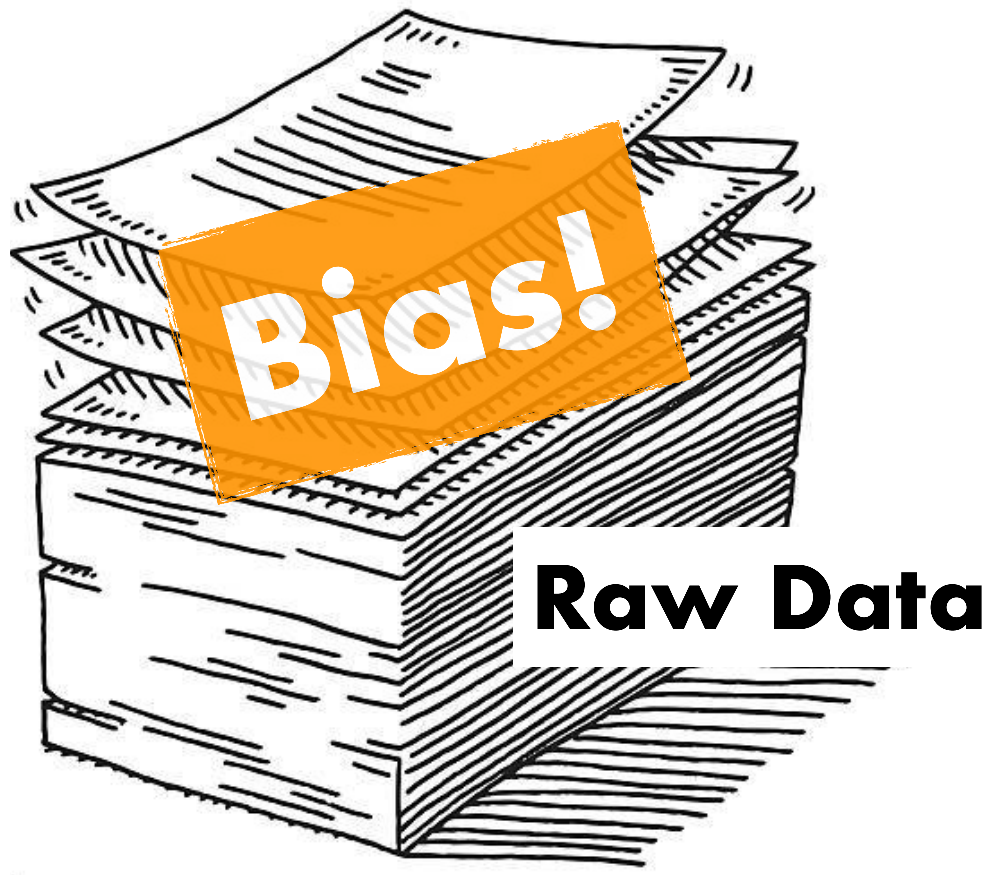
The Risk of Racial Bias [Sap et. al, 2019]


Human Labeling
Bias!
Training
Bias!
Evaluation
Bias!
The AI Pipeline
Raw Data
Bias!
Deployment
Bias!
Bias!
Bias!
Bias!


Bias!
Human
Labeling
Bias!
Raw Data
Bias!
Training
Evaluation
Bias!
The AI
Pipeline
Deployment
Bias!
Bias!
Bias!
Bias!


Human Labeling
Bias!
Training
Bias!
Evaluation
Bias!
The AI Pipeline
Raw Data
Bias!
Deployment
Bias!
Bias!
Bias!
Bias!

Addressing Biases: Datasets



Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]

The Risk of Racial Bias [Sap et. al, 2019]

Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases

The Risk of Racial Bias [Sap et. al, 2019]

Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases
  - Profanity Biases

The Risk of Racial Bias [Sap et. al, 2019]

Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases

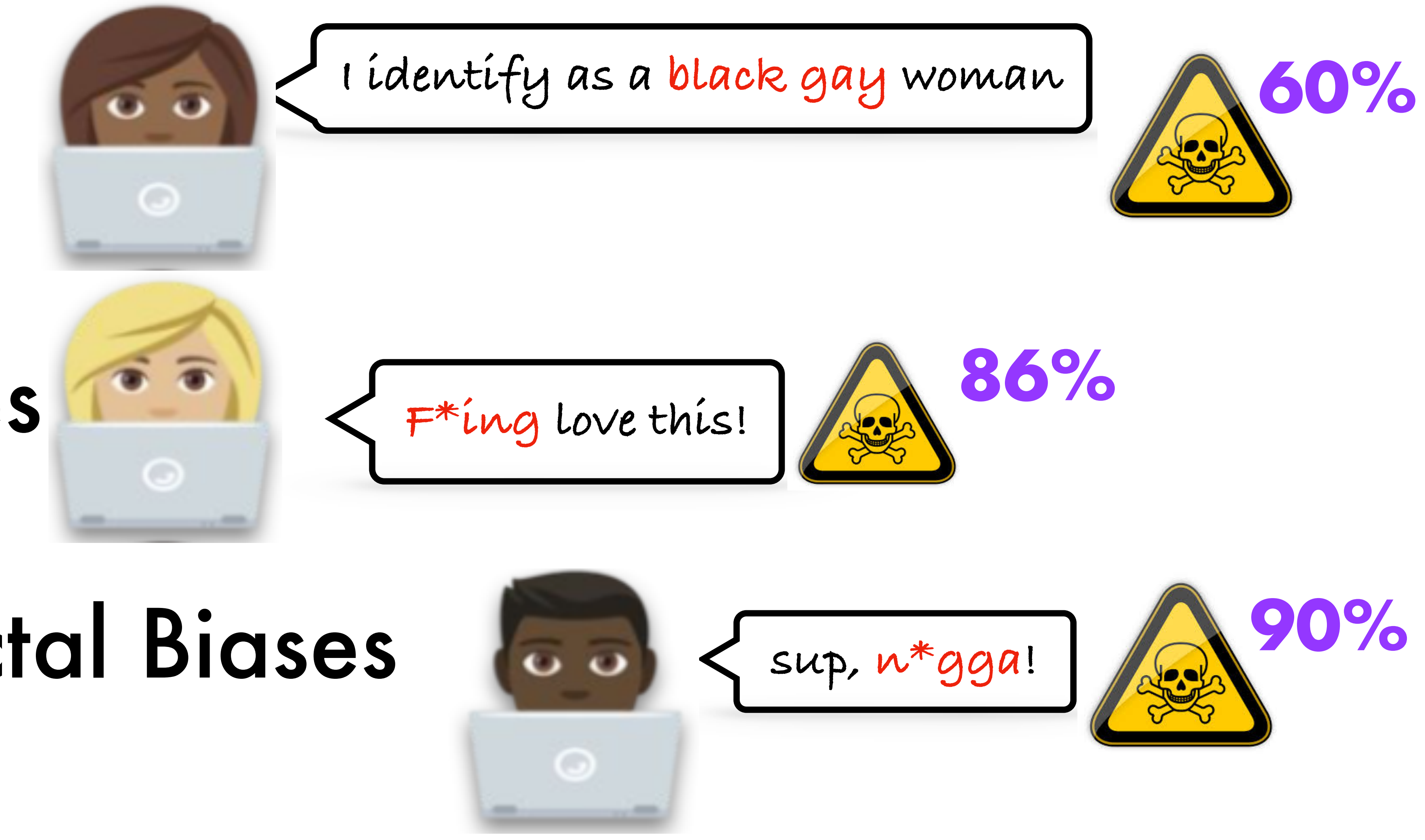
  - Profanity Biases

  - Racial / Dialectal Biases

The Risk of Racial Bias [Sap et. al, 2019]

Addressing Biases: Datasets

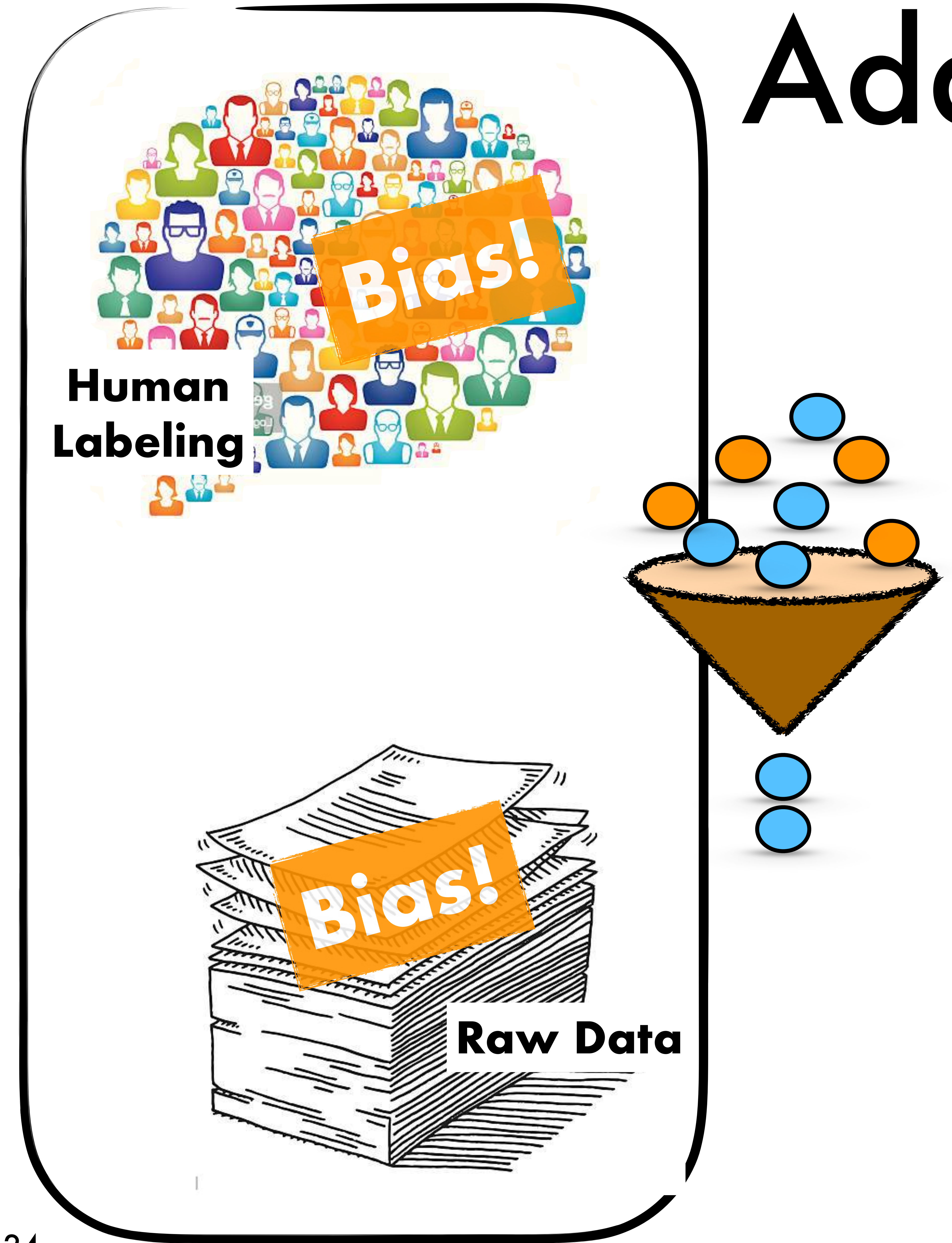
- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases
  - Profanity Biases
  - Racial / Dialectal Biases



- One solution: Filtering / Downsampling

The Risk of Racial Bias [Sap et. al, 2019]

Addressing Biases: Datasets



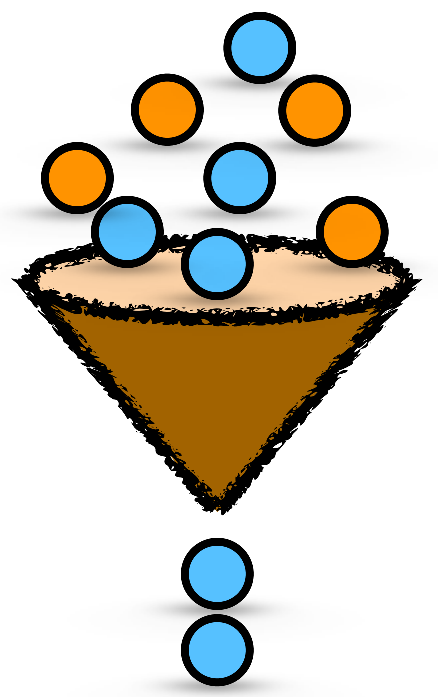
- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases

  - Profanity Biases

  - Racial / Dialectal Biases

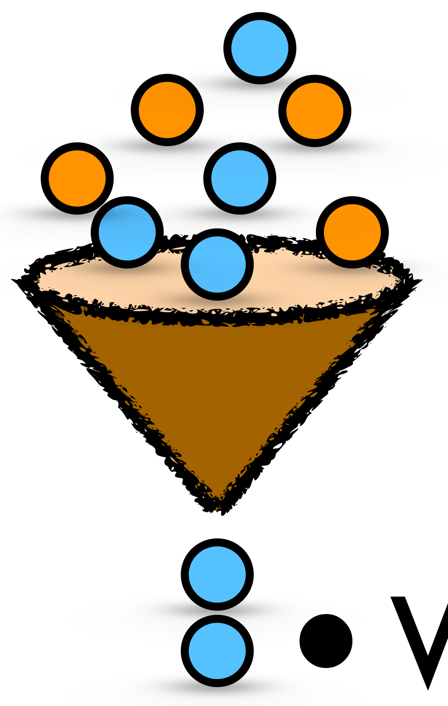
- One solution: Filtering / Downsampling

The Risk of Racial Bias [Sap et. al, 2019]



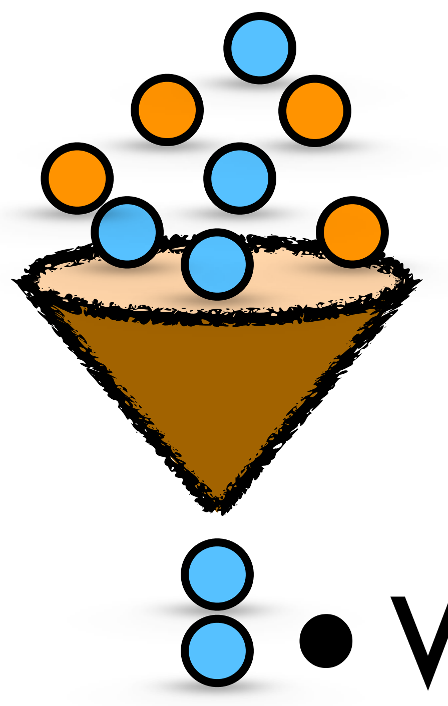
Dataset Filtering

Dataset Filtering



- What instances to filter?

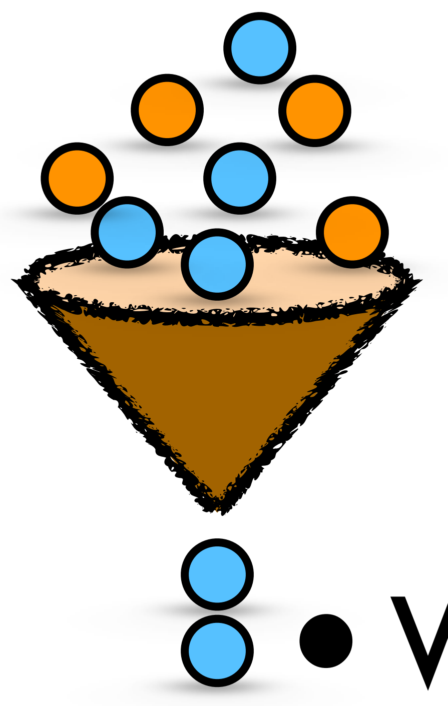
Dataset Filtering



- What instances to filter?

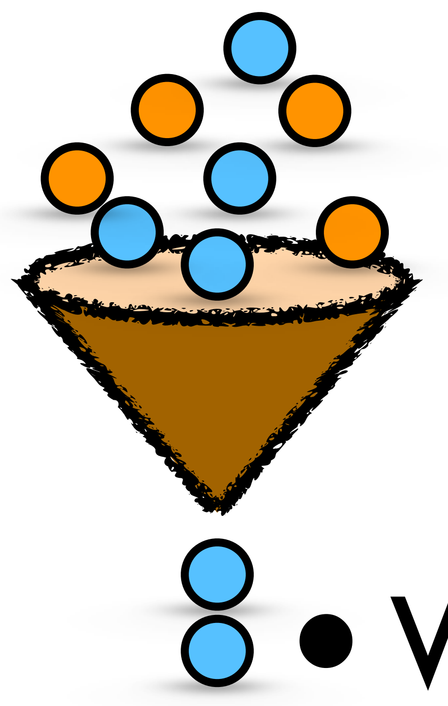
  - Key intuition: Examples which are relatively easy for a model might contain spurious correlations

Dataset Filtering



- What instances to filter?
    - Key intuition: Examples which are relatively easy for a model might contain spurious correlations
- Easy examples can be detected:

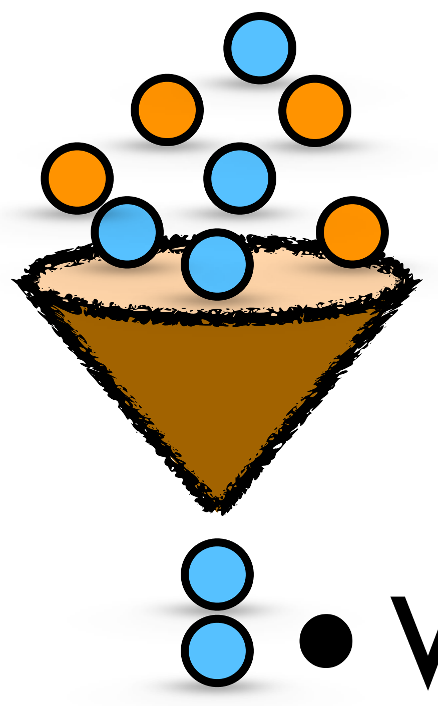
Dataset Filtering



- What instances to filter?
    - Key intuition: Examples which are relatively easy for a model might contain spurious correlations
- Easy examples can be detected:
    - By simple model architectures

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C., 2020]

Dataset Filtering



- What instances to filter?
    - Key intuition: Examples which are relatively easy for a model might contain spurious correlations
- Easy examples can be detected:
    - By simple model architectures
    - Early in the training process

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C., 2020]
Dataset Cartography [Swayamdipta et al., 2020]

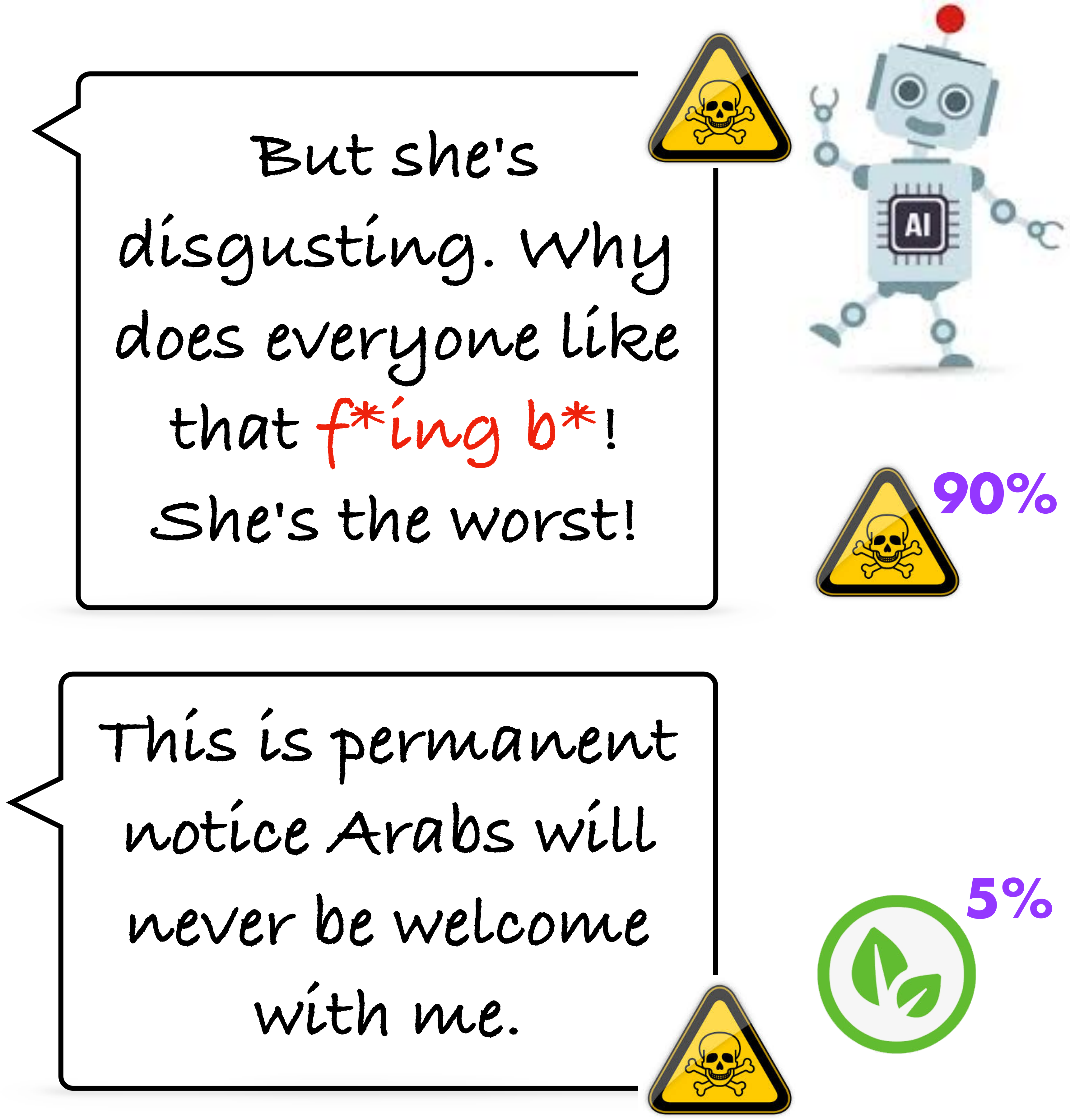
Algorithmic Dataset Filtering

But she's disgusting. Why does everyone like that f*ing b*! She's the worst!

This is permanent notice Arabs will never be welcome with me.

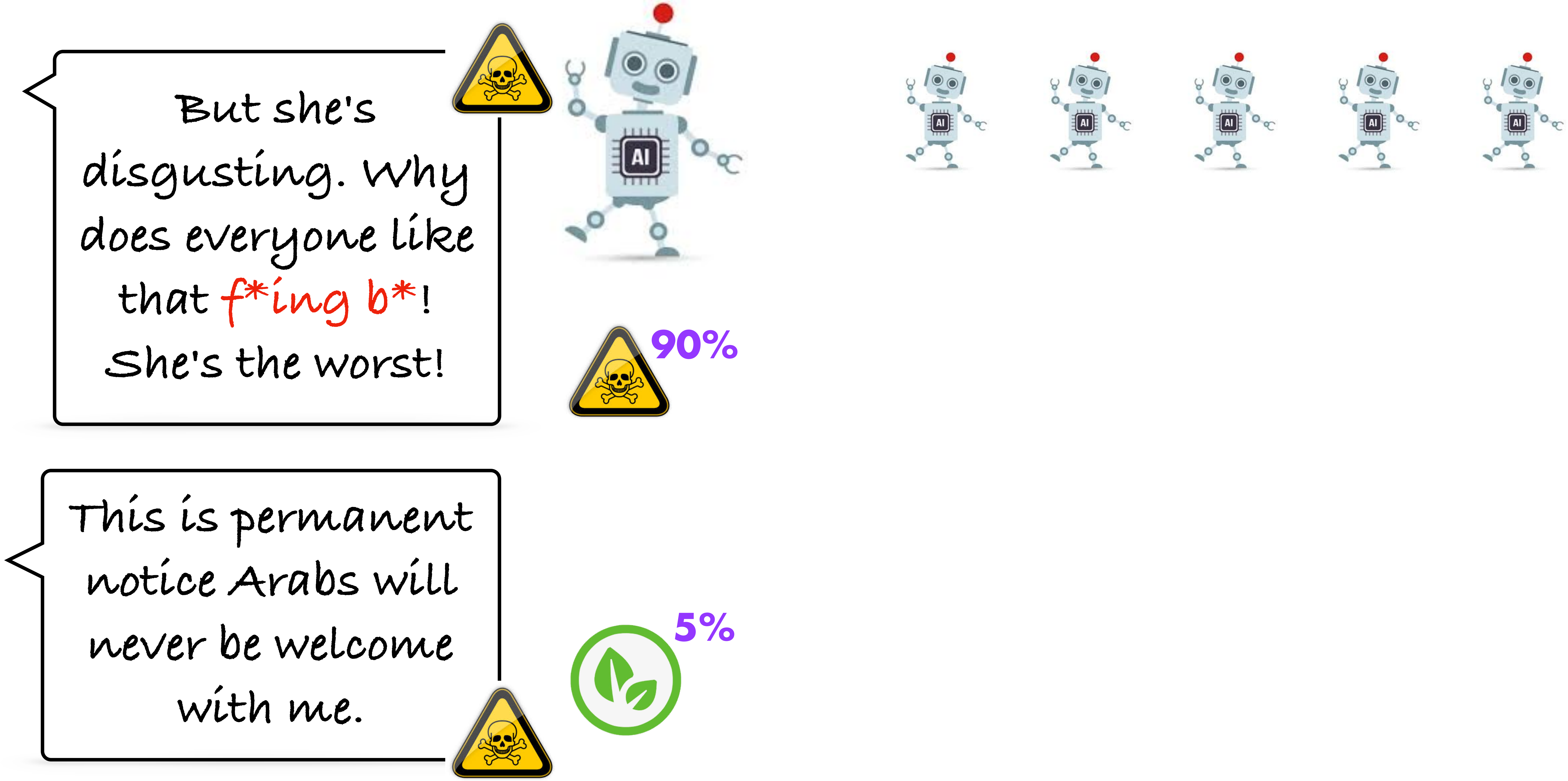
Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]

Algorithmic Dataset Filtering



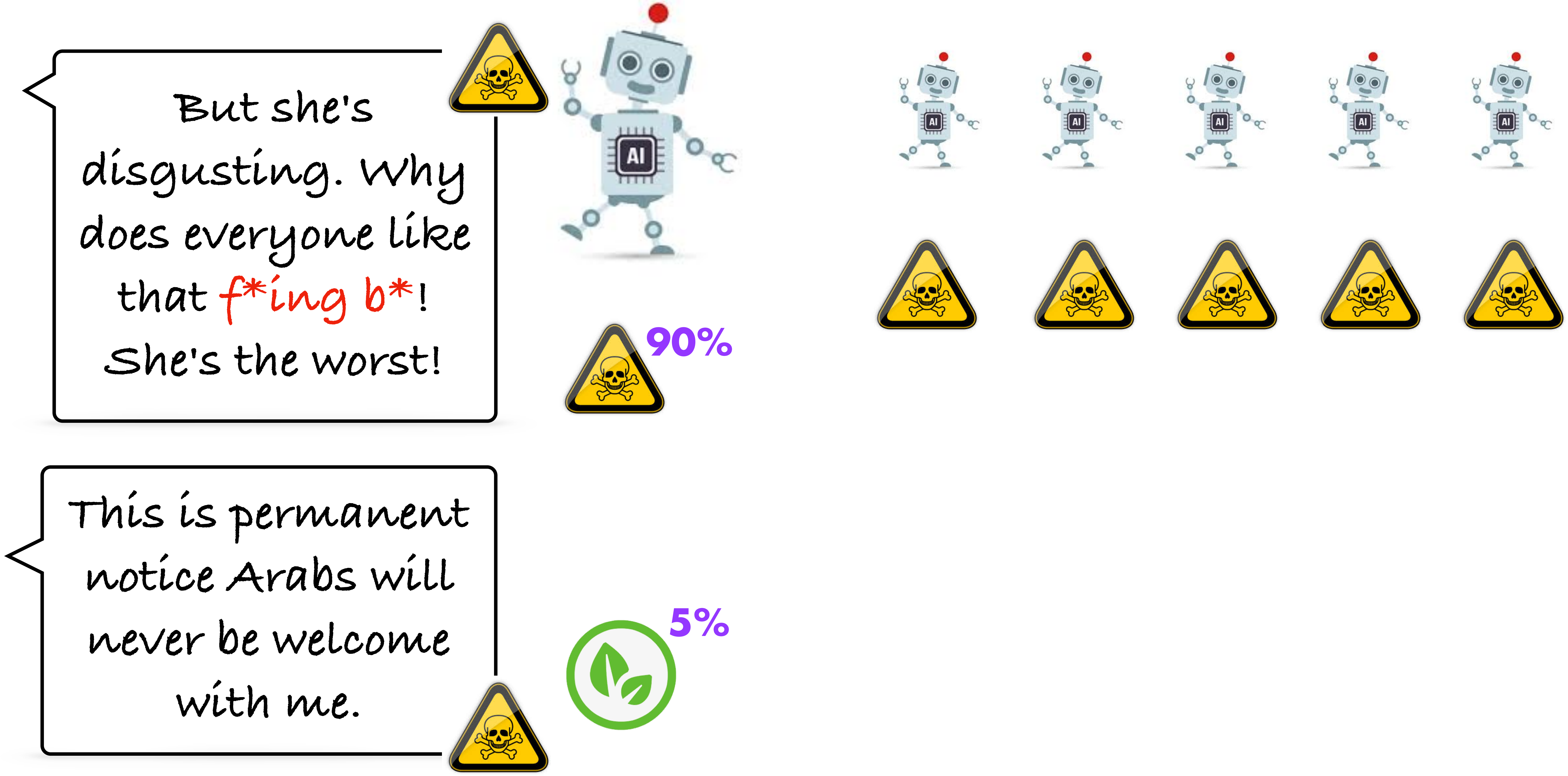
Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]

Algorithmic Dataset Filtering



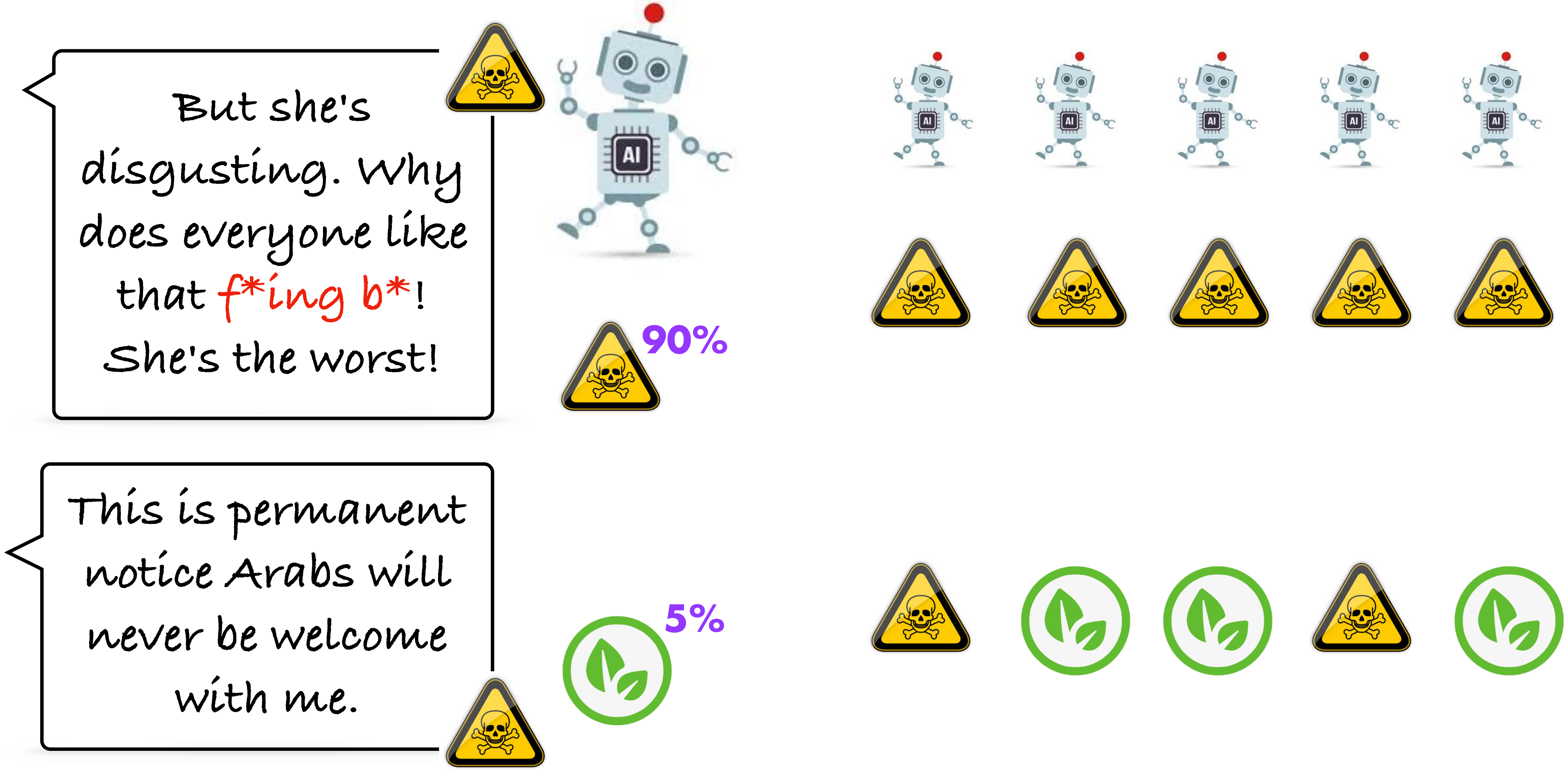
Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]

Algorithmic Dataset Filtering



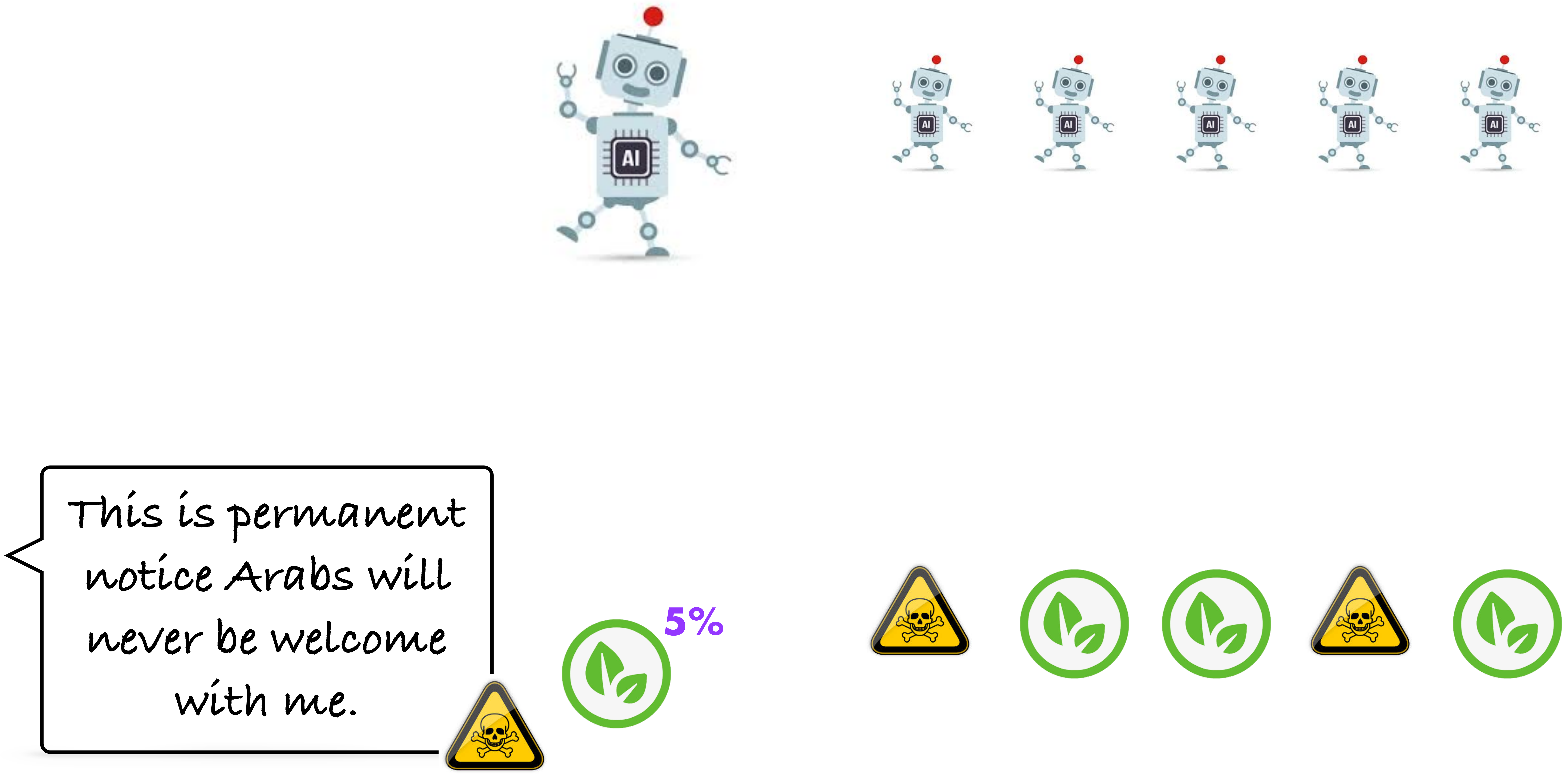
Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]

Algorithmic Dataset Filtering



Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]

Algorithmic Dataset Filtering

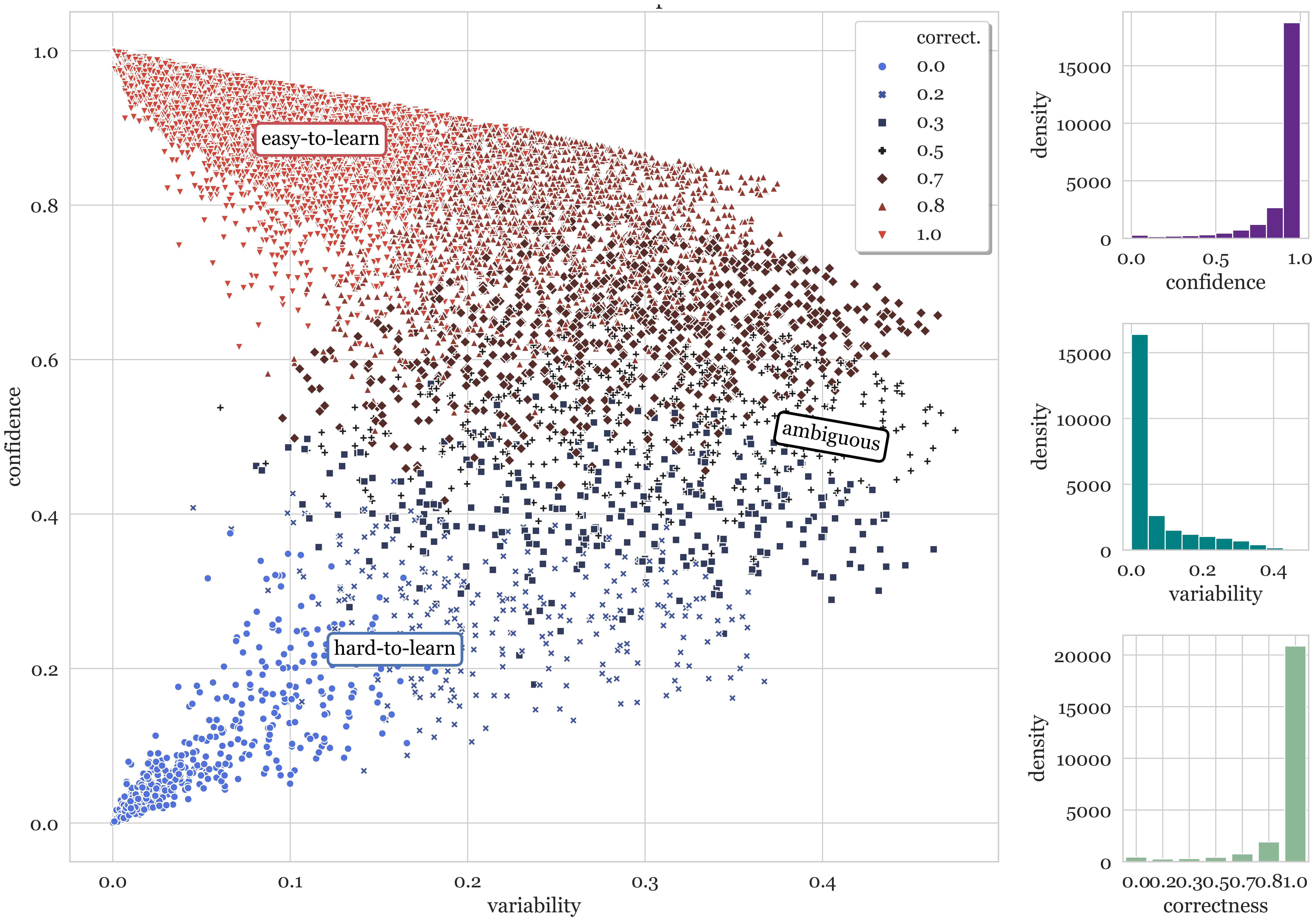


Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]

Data Maps

 Dataset Cartography [Swayamdipta et. al, 2020]

Data Maps

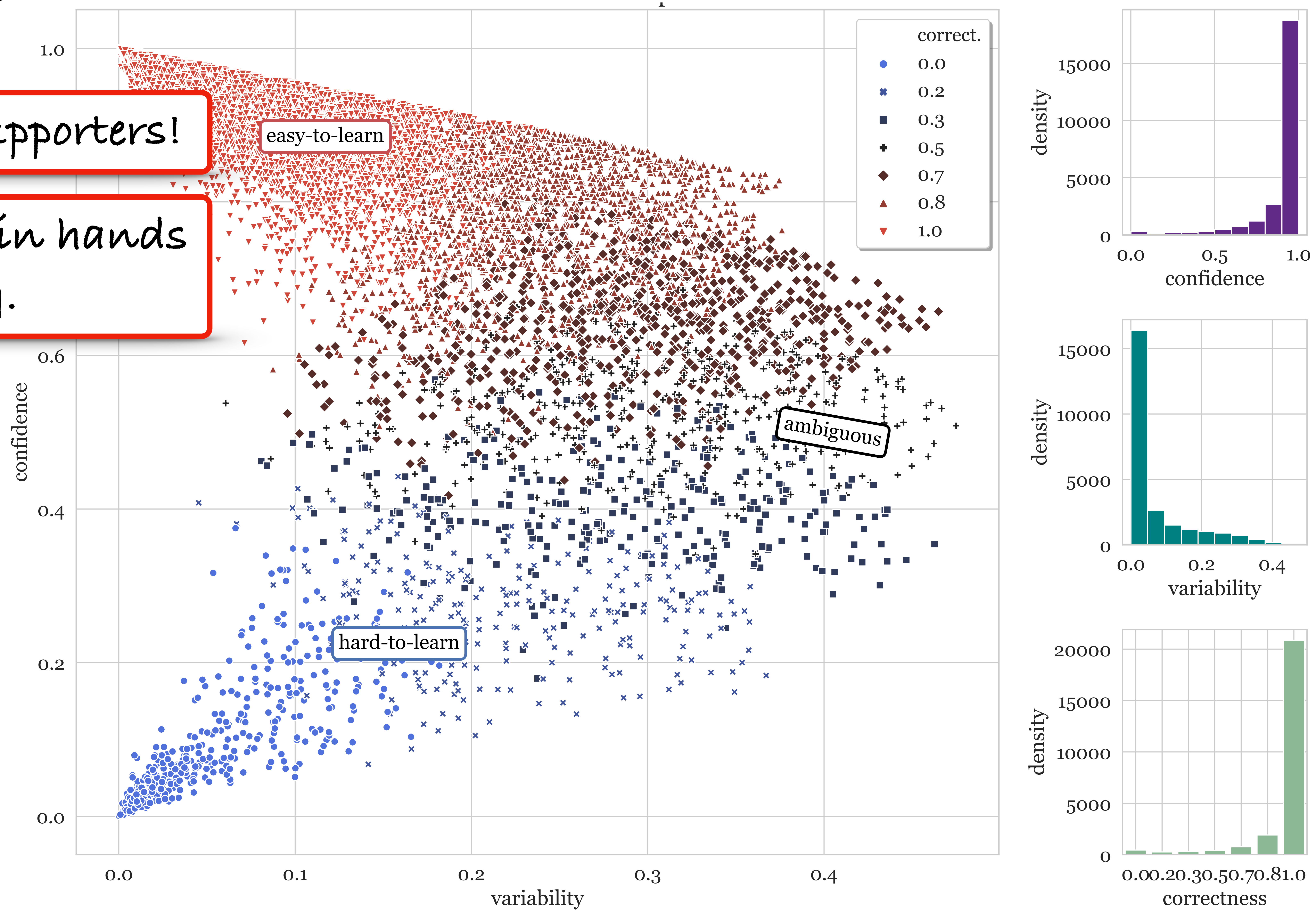


 Dataset Cartography [Swayamdipta et. al, 2020]

Data Maps

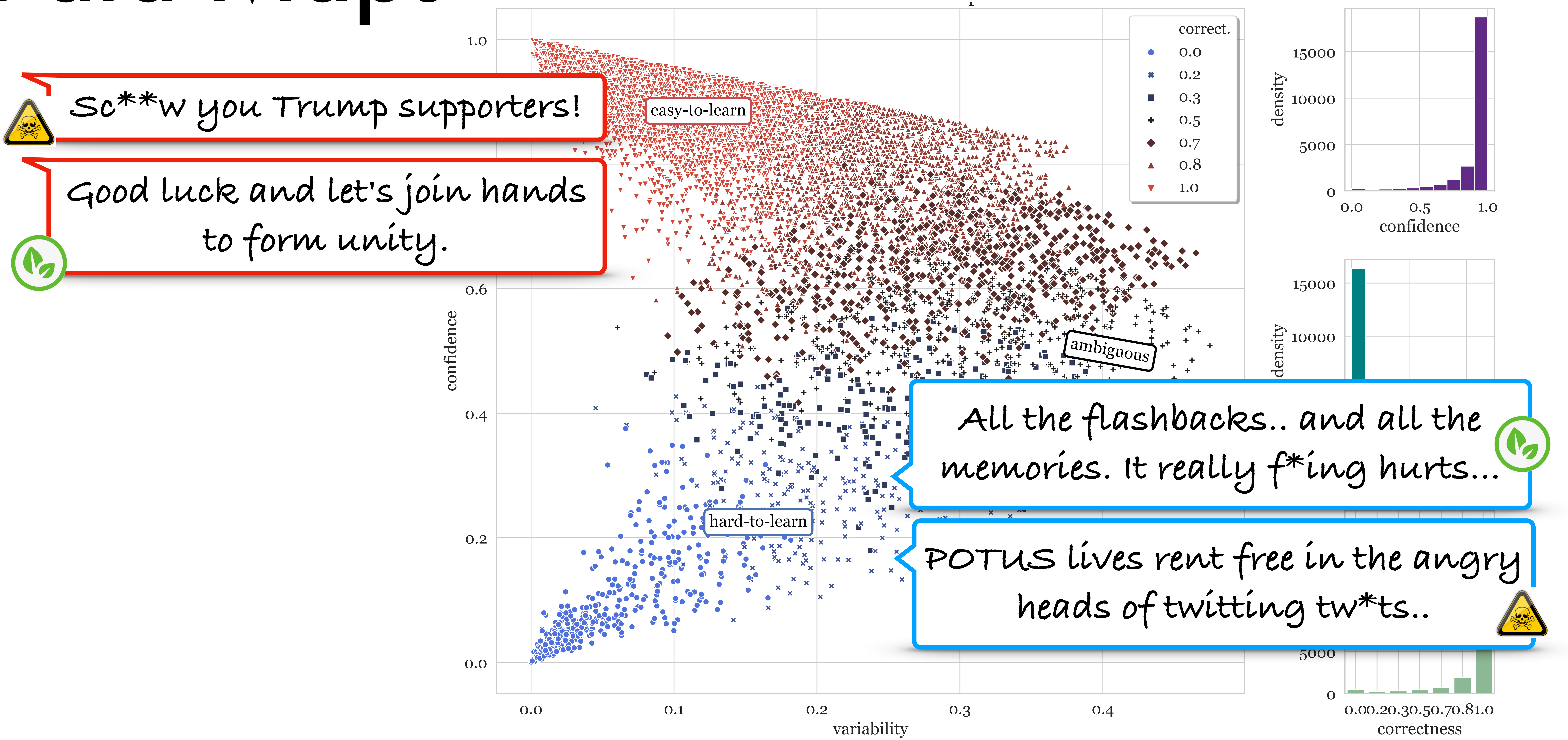
Sc**w you Trump supporters!

Good luck and let's join hands to form unity.

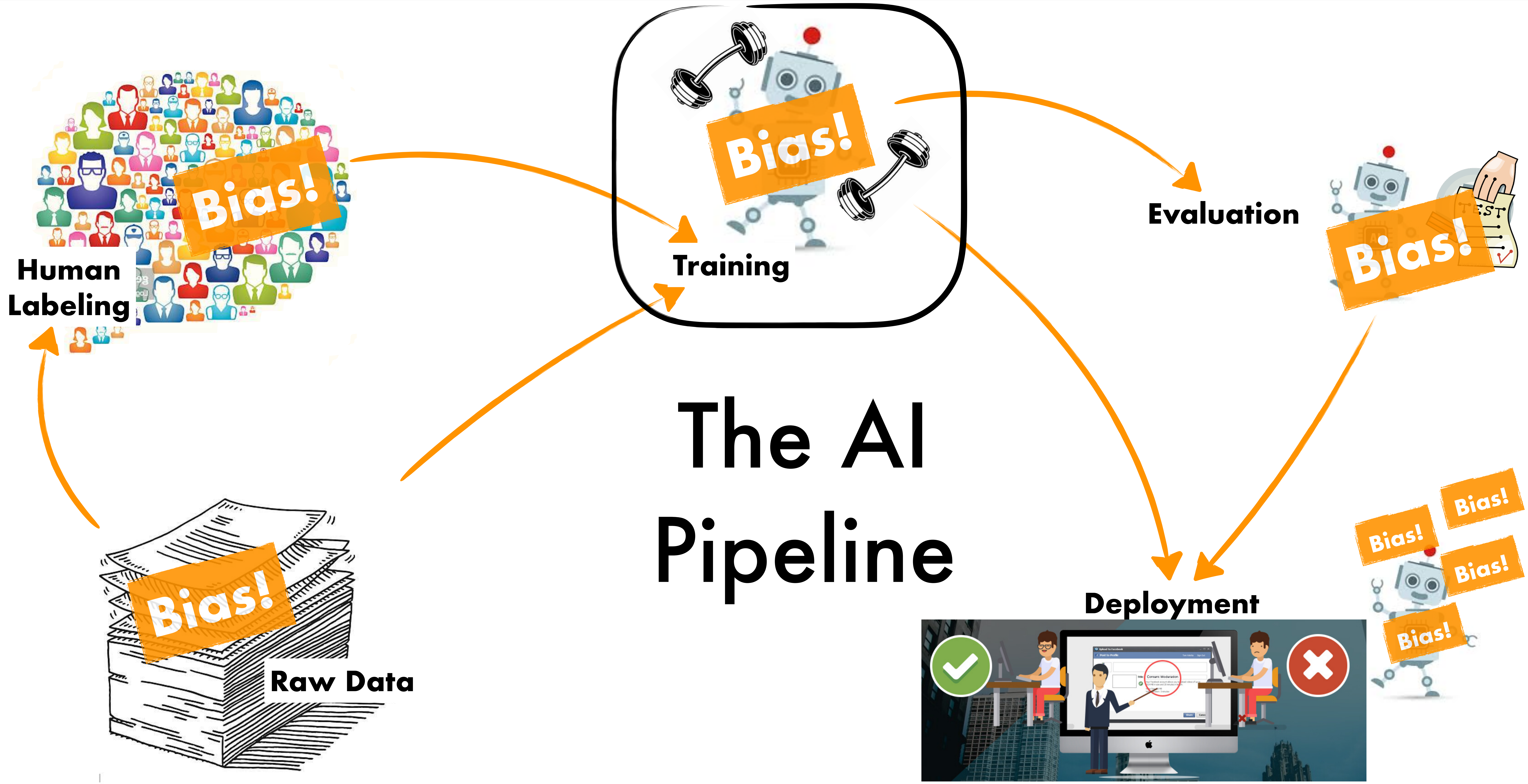


 Dataset Cartography [Swayamdipta et. al, 2020]

Data Maps



 Dataset Cartography [Swayamdipta et. al, 2020]


Human Labeling
Bias!
Training
Bias!
Evaluation
Bias!
The AI Pipeline
Bias!
Raw Data
Deployment
Bias!
Bias!
Bias!
Bias!

Addressing Biases: Models

 [Clark et al., 2019]

Addressing Biases: Models

- Can be used to reduce known biases

[Clark et al., 2019]

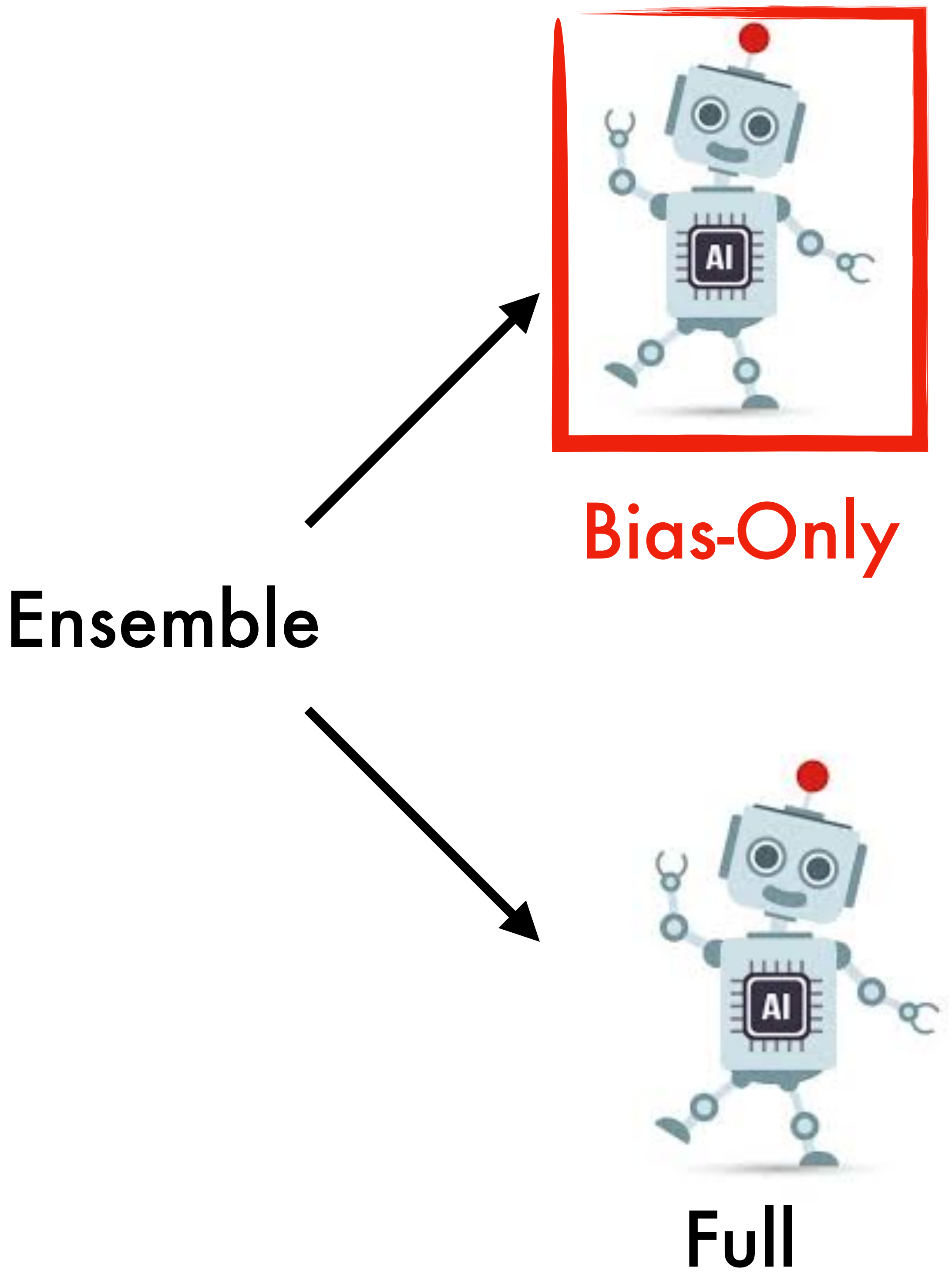
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities

[Clark et al., 2019]

Addressing Biases: Models

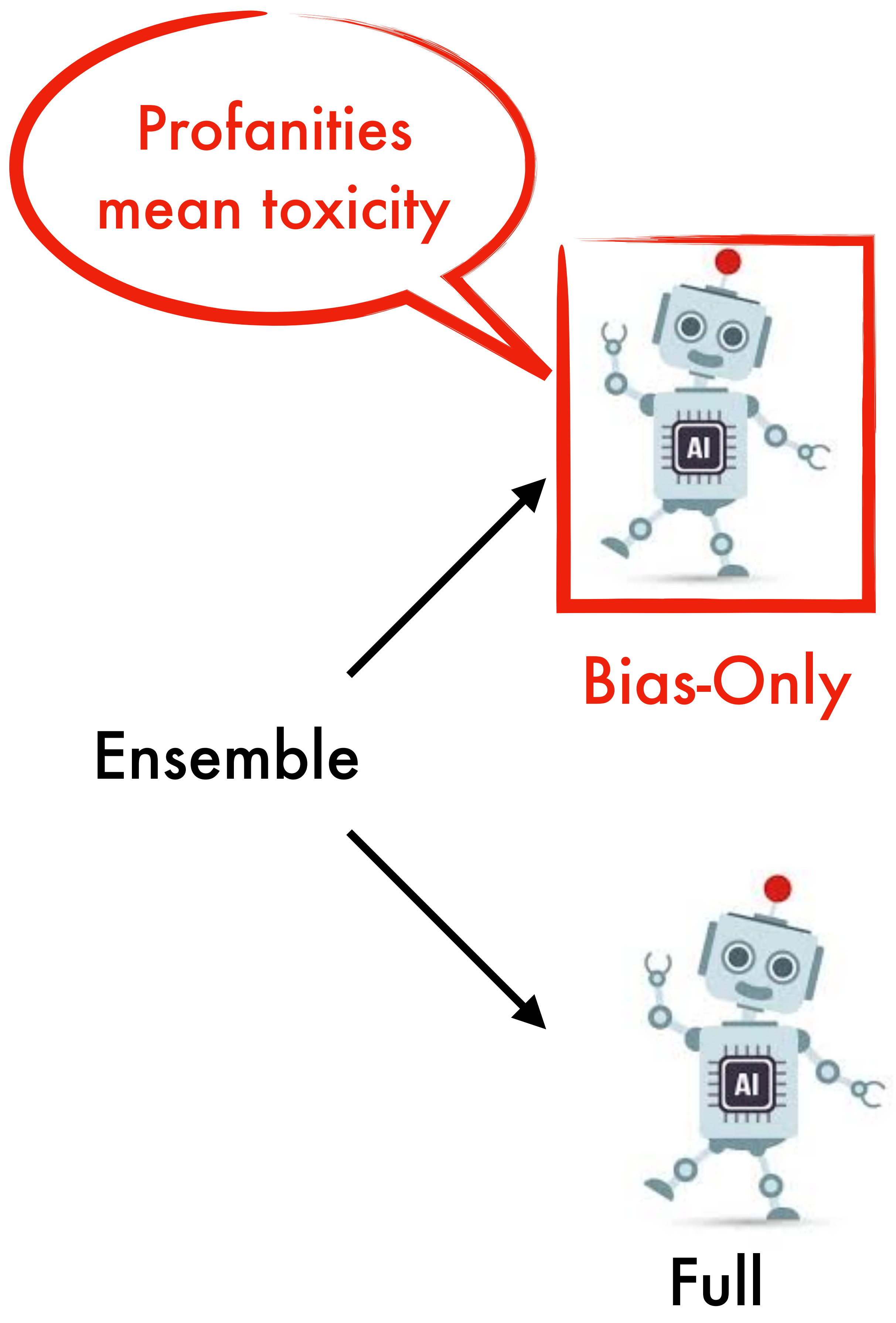
- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model



[Clark et al., 2019]

Addressing Biases: Models

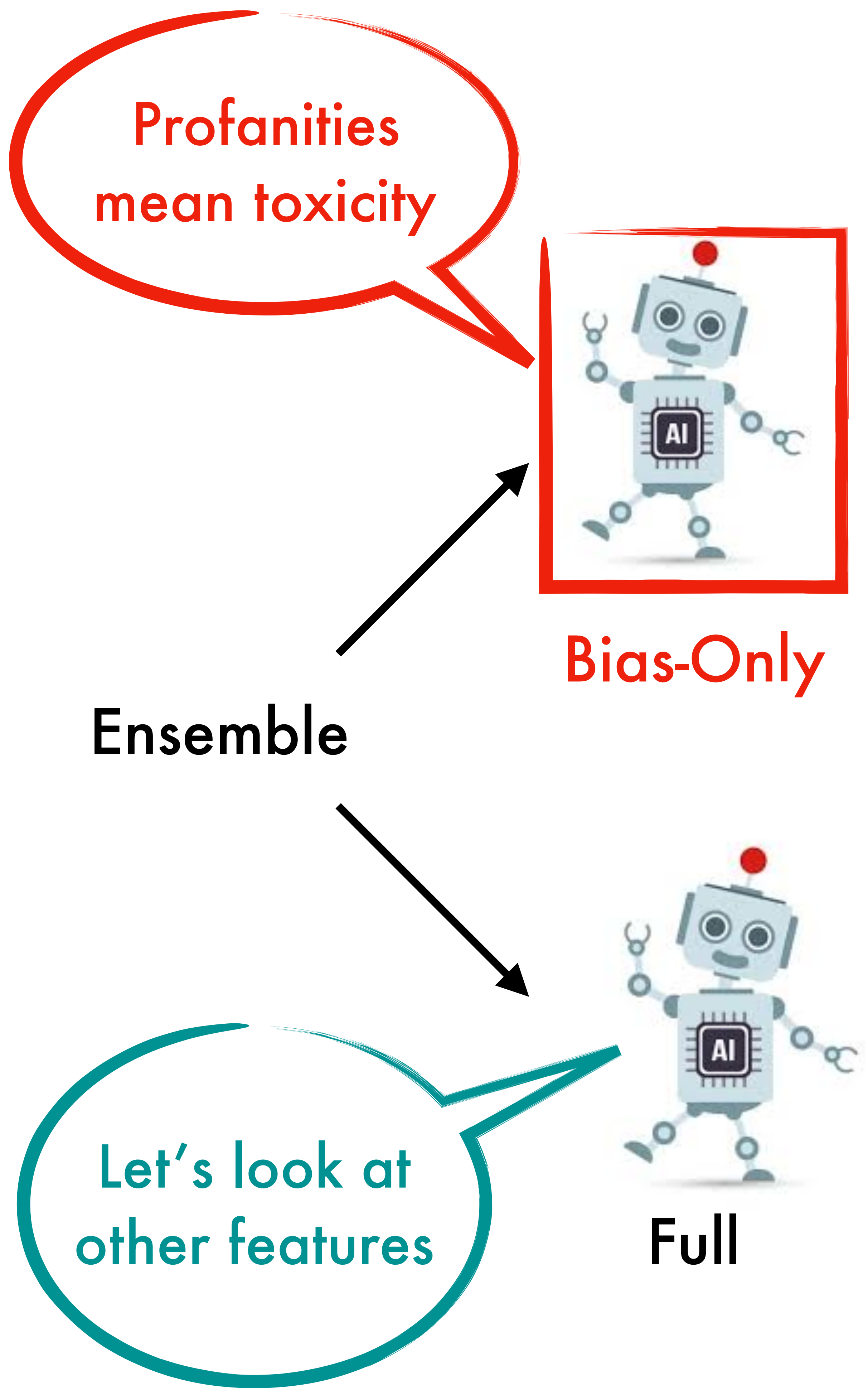
- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases



 [Clark et al., 2019]

Addressing Biases: Models

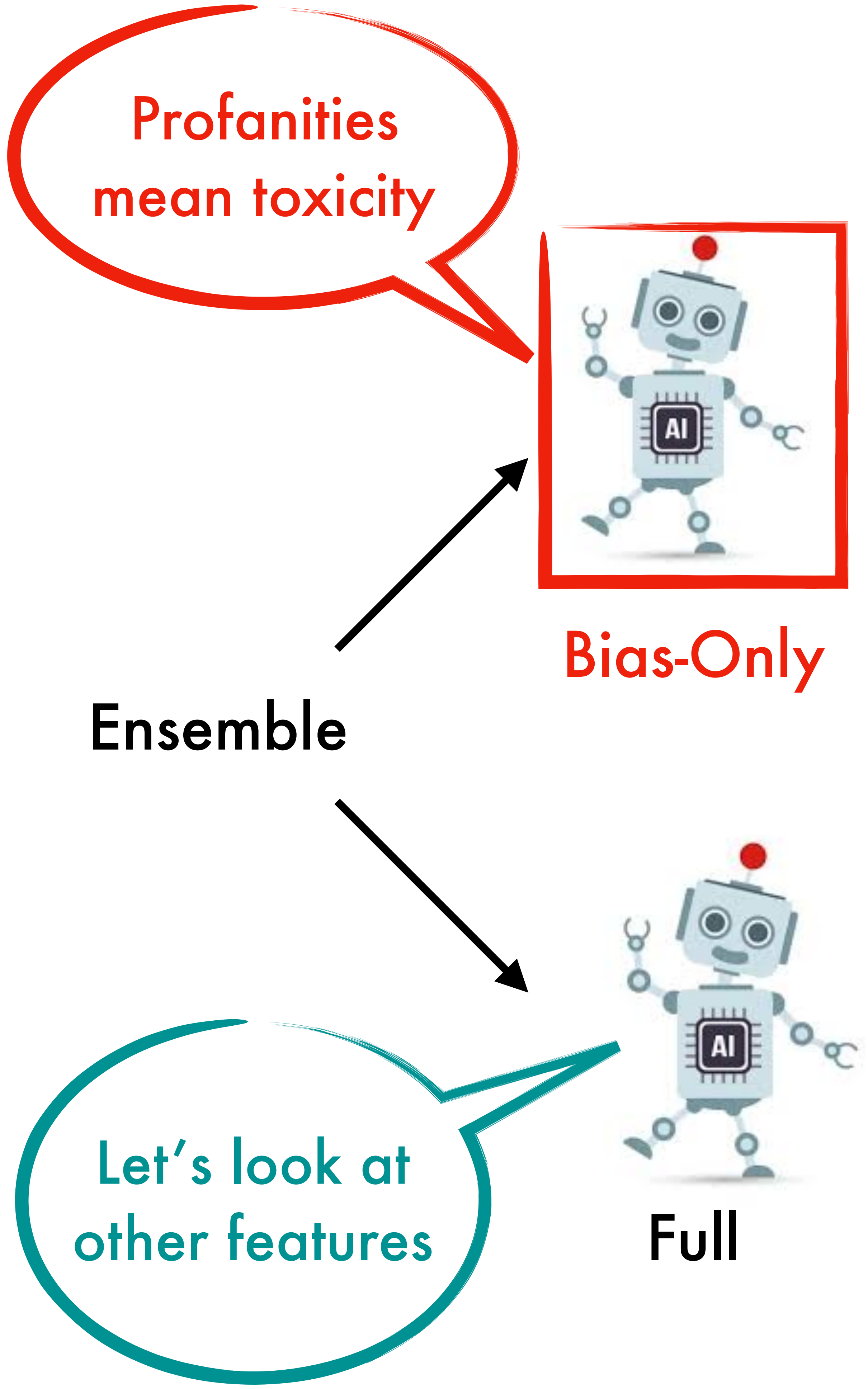
- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases



 [Clark et al., 2019]

Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases

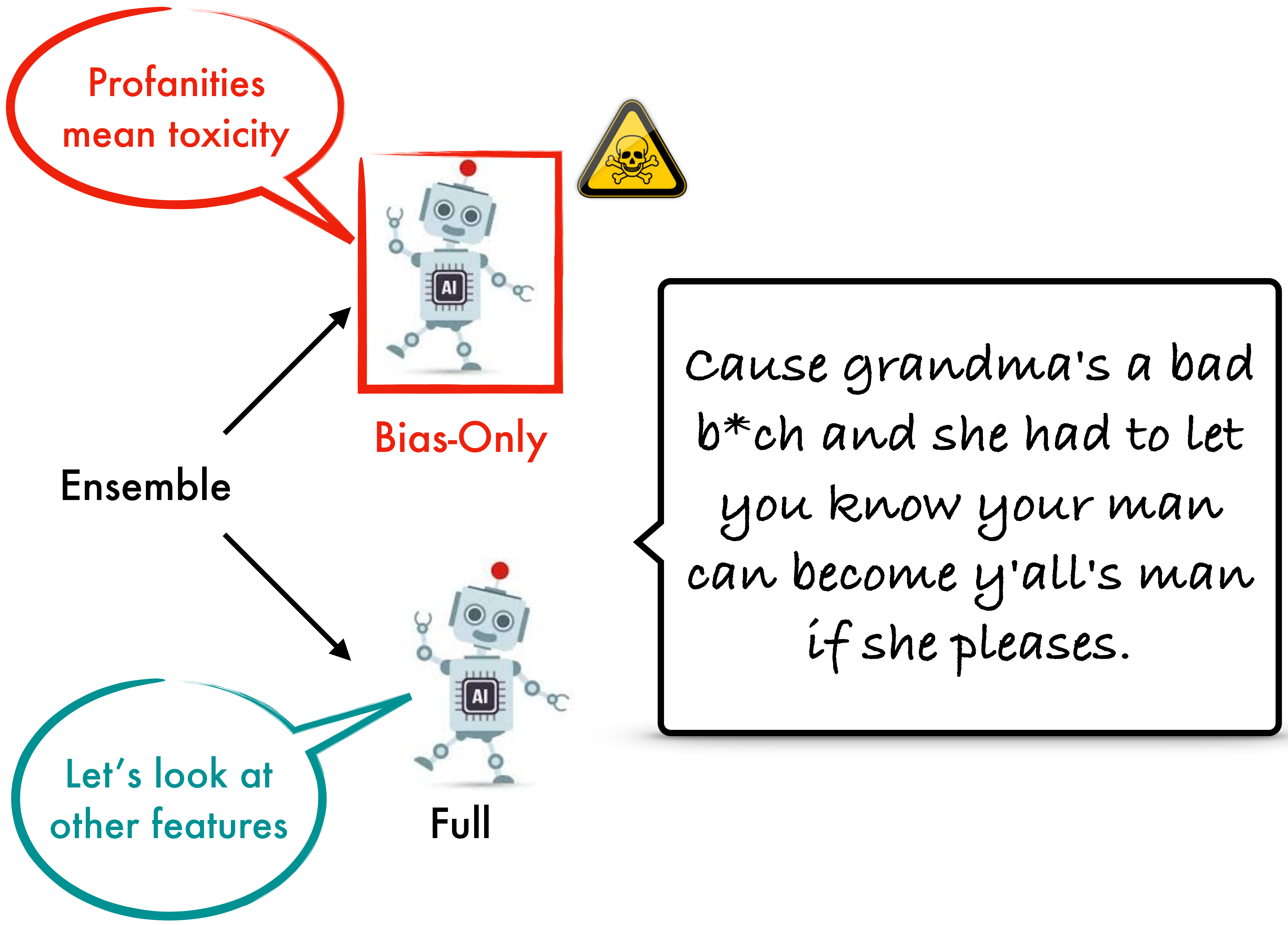


Cause grandma's a bad b*ch and she had to let you know your man can become y'all's man if she pleases.

 [Clark et al., 2019]

Addressing Biases: Models

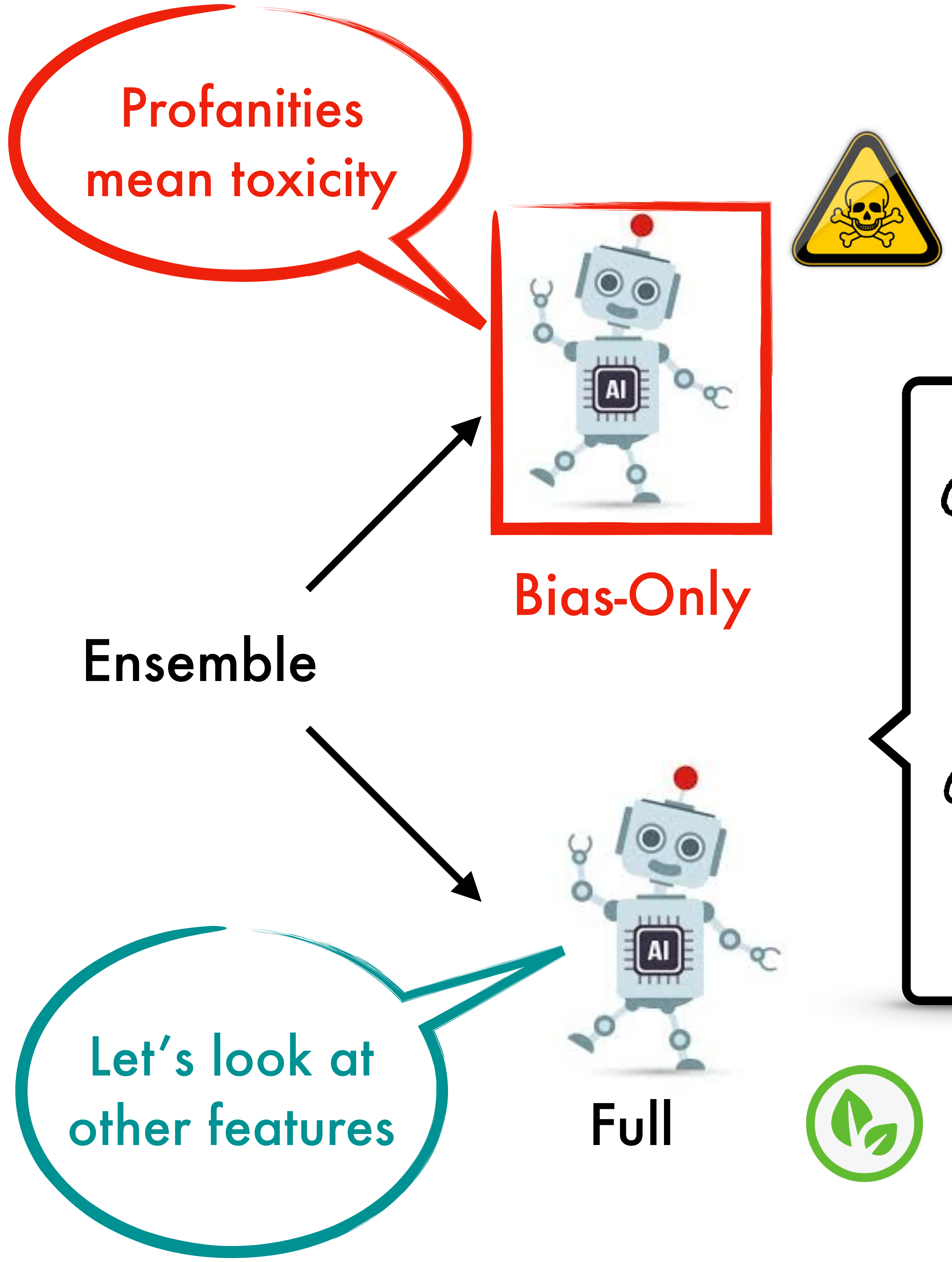
- Can be used to reduce known biases
    - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases



[Clark et al., 2019]

Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases

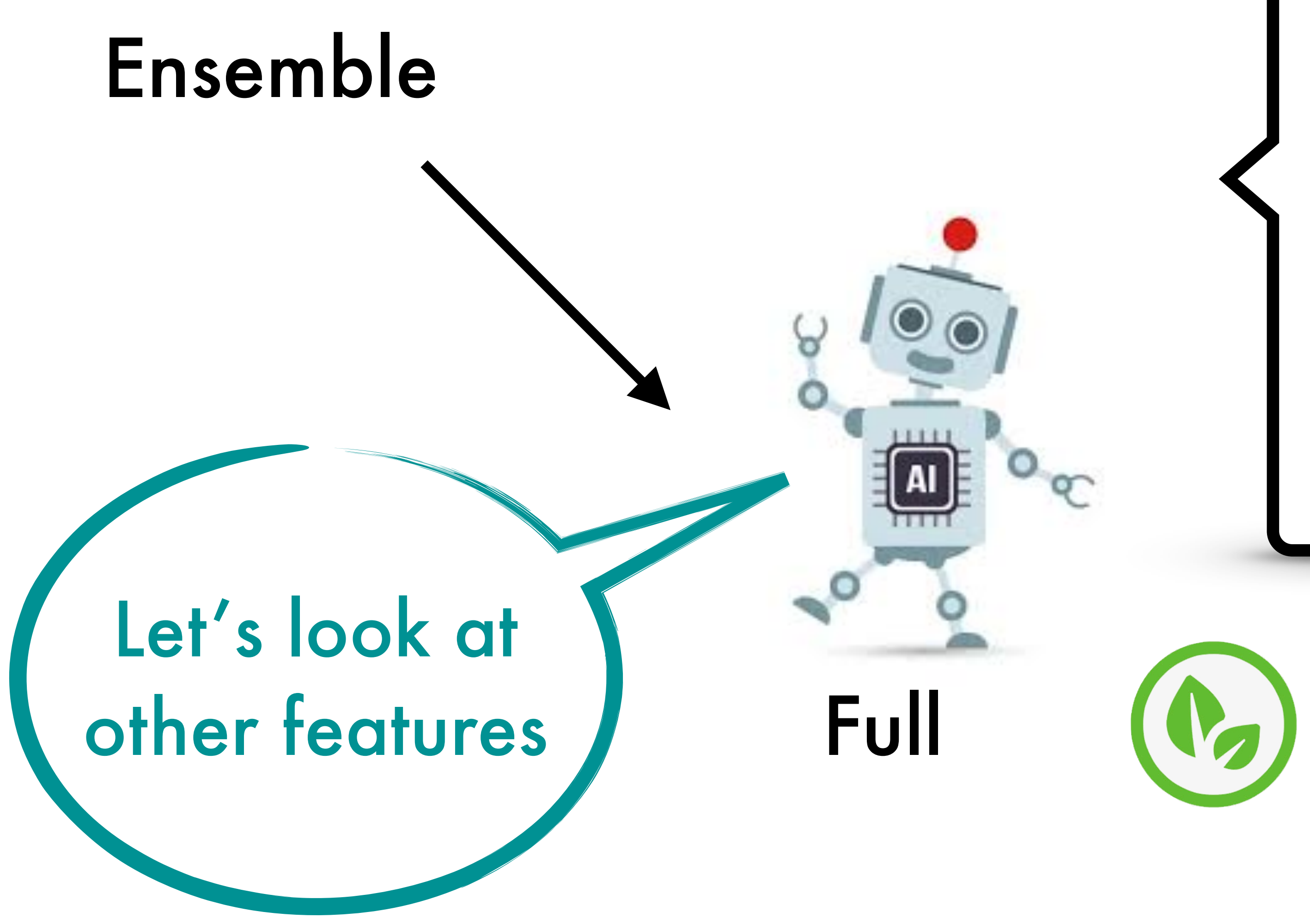


Cause grandma's a bad b*ch and she had to let you know your man can become y'all's man if she pleases.

 [Clark et al., 2019]

Addressing Biases: Models

- Can be used to reduce known biases
    - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases



Cause grandma's a bad b*ch and she had to let you know your man can become y'all's man if she pleases.

 [Clark et al., 2019]

Findings

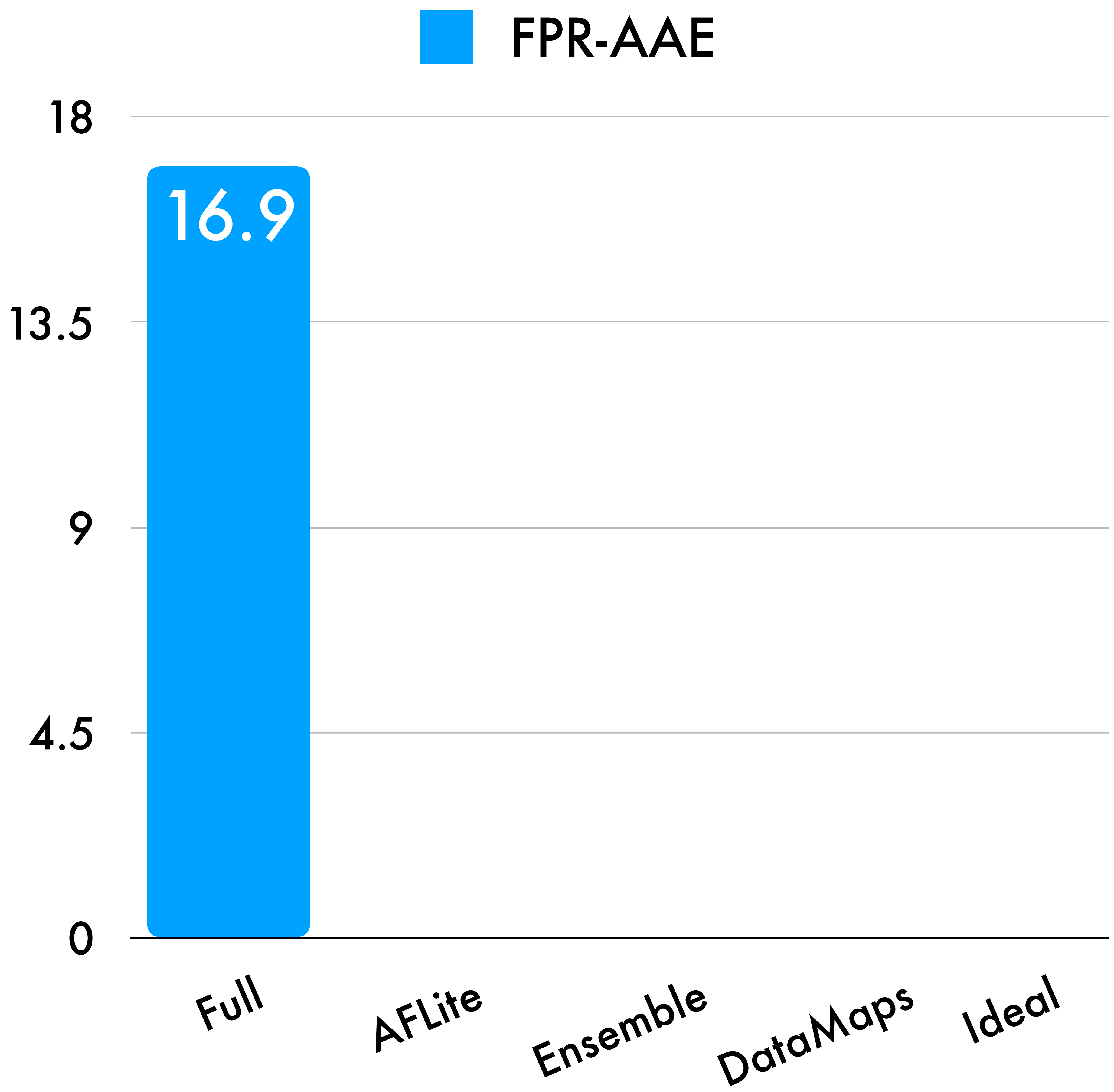
Findings

- Dataset : Founta et al. (2018)
- False Positive Rate on tweets in African American English (AAE)
- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)

The Ineffectiveness of Algorithmic Debiasing on Toxic Language Detection [Z., S., Swayamdipta, C., S., 2020] Under Review

Findings

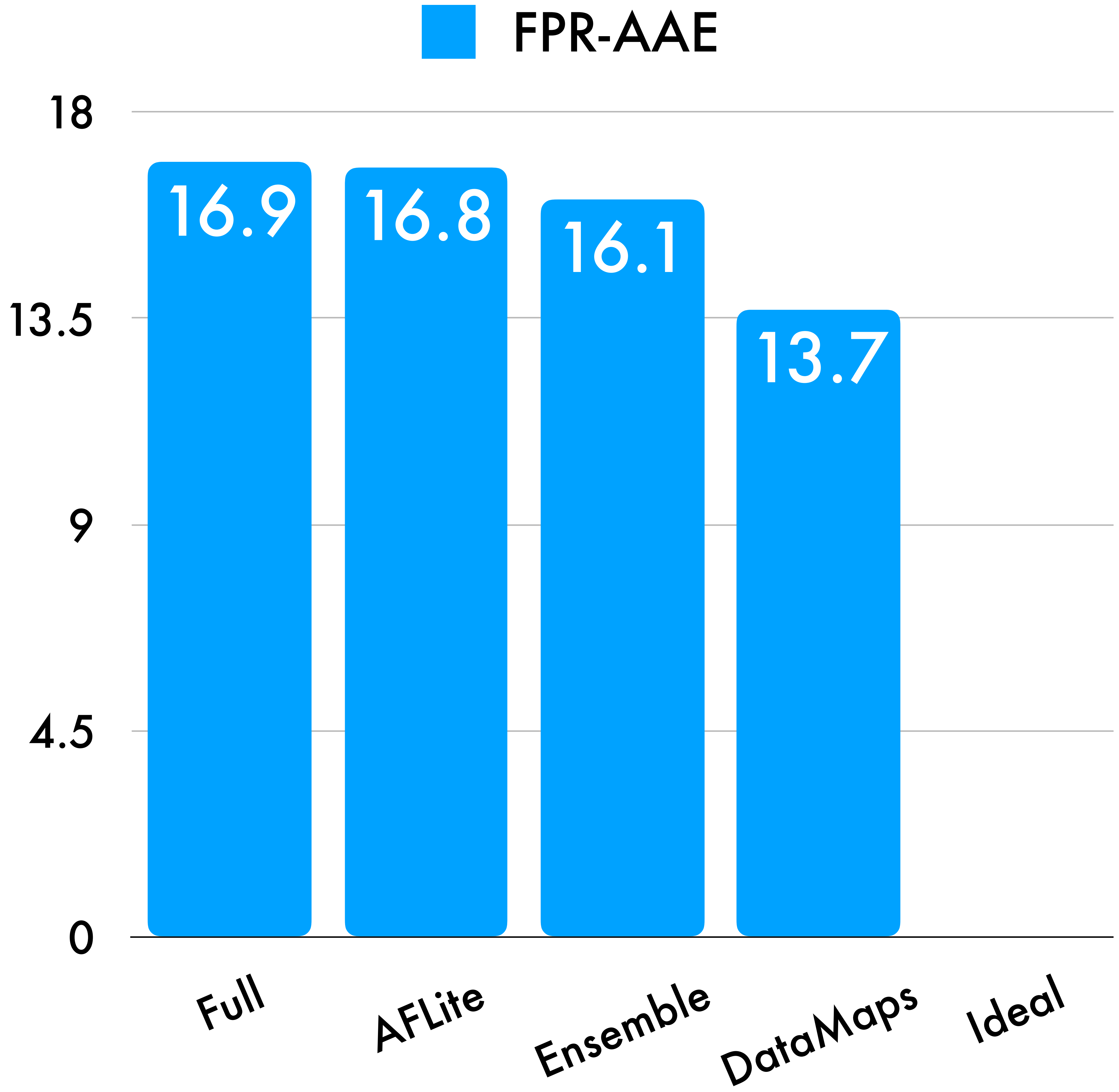
- Dataset : Founta et al. (2018)
- False Positive Rate on tweets in African American English (AAE)
- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)



The Ineffectiveness of Algorithmic Debiasing on Toxic Language Detection [Z., S., Swayamdipta, C., S., 2020] Under Review

Findings

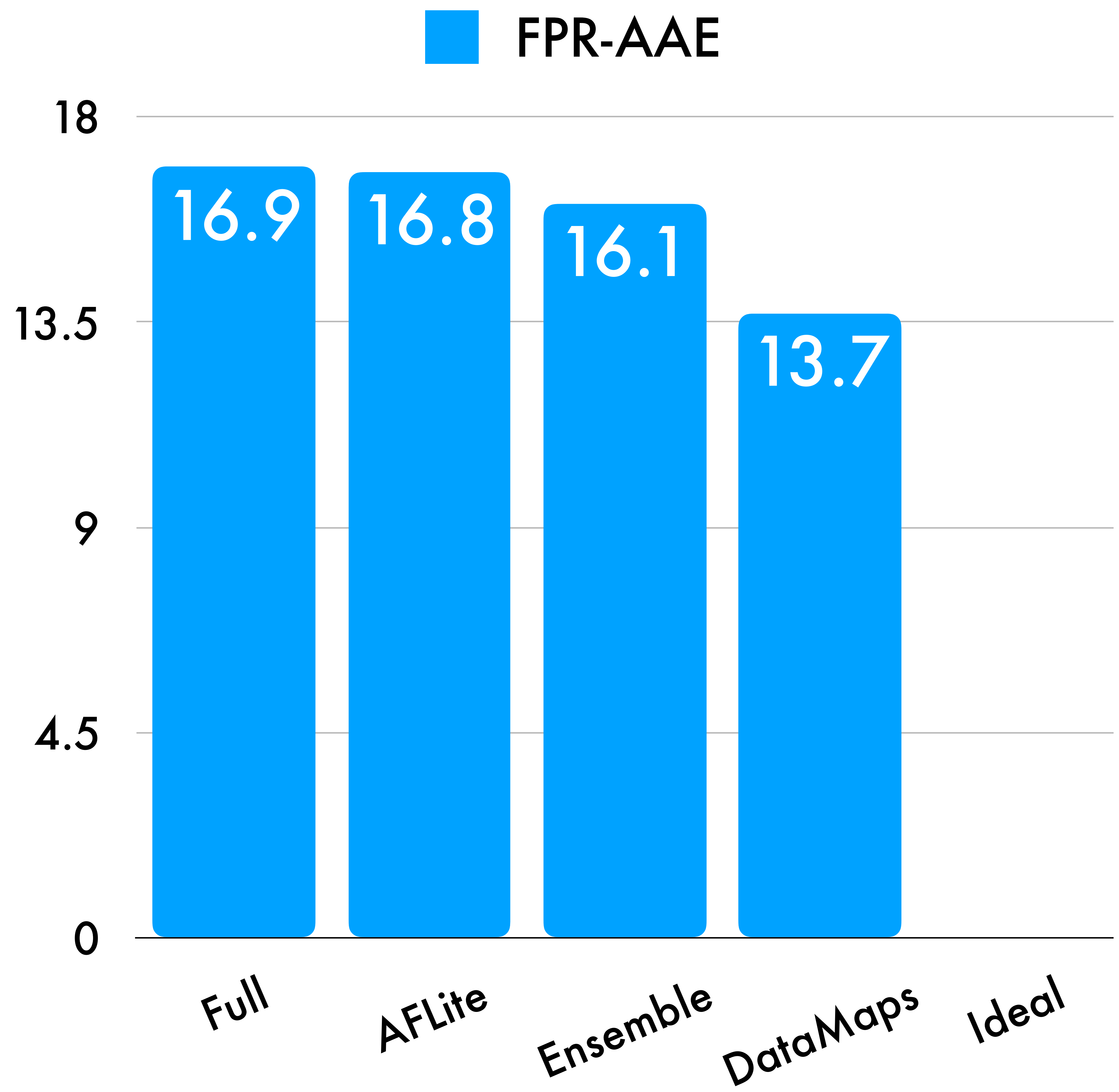
- Dataset : Founta et al. (2018)
- False Positive Rate on tweets in African American English (AAE)
- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)



The Ineffectiveness of Algorithmic Debiasing on Toxic Language Detection [Z., S., Swayamdipta, C., S., 2020] Under Review

Findings

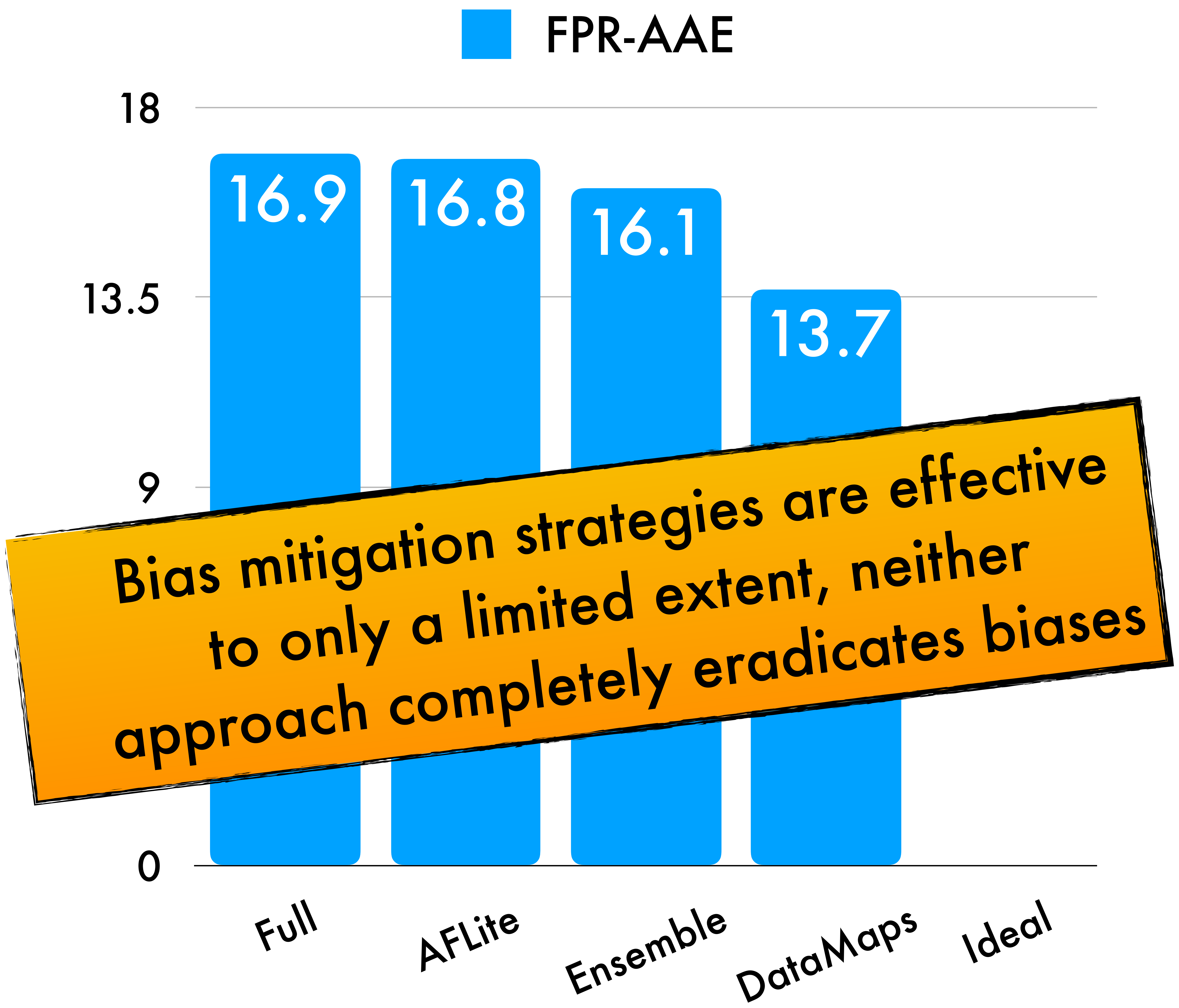
- Dataset : Founta et al. (2018)
- False Positive Rate on tweets in African American English (AAE)
- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)



The Ineffectiveness of Algorithmic Debiasing on Toxic Language Detection [Z., S., Swayamdipta, C., S., 2020] Under Review

Findings

- Dataset : Founta et al. (2018)
- False Positive Rate on tweets in African American English (AAE)
- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)



The Ineffectiveness of Algorithmic Debiasing on Toxic Language Detection [Z., S., Swayamdipta, C., S., 2020] Under Review

Takeaways: Addressing Biases in Hatespeech Detection

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot *remove* racial biases, however

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot <u>*remove*</u> racial biases, however
    - Bias Mitigation instead of *Debiasing*

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot *remove* racial biases, however
    - Bias Mitigation instead of *Debiasing*
- An uncomfortable truth:

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot *remove* racial biases, however
  - Bias Mitigation instead of *Debiasing*
- An uncomfortable truth:

  "Bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed." - [Waseem et al. 2020]

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot *remove* racial biases, however
    - Bias Mitigation instead of *Debiasing*
- An uncomfortable truth:

    "Bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed." - [Waseem et al. 2020]

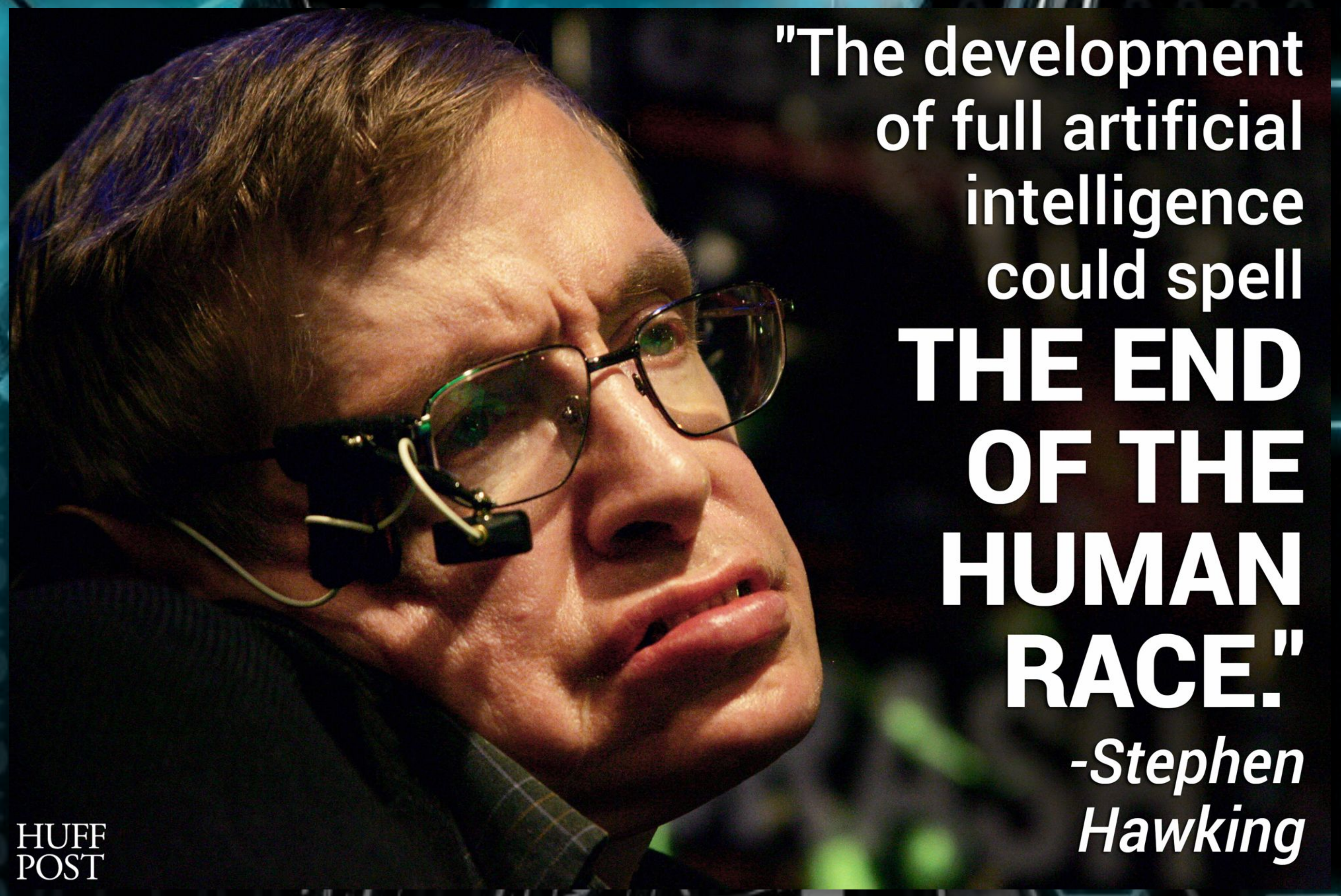
Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot *remove* racial biases, however
  - Bias Mitigation instead of *Debiasing*
- An uncomfortable truth:

  "Bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed." - [Waseem et al. 2020]

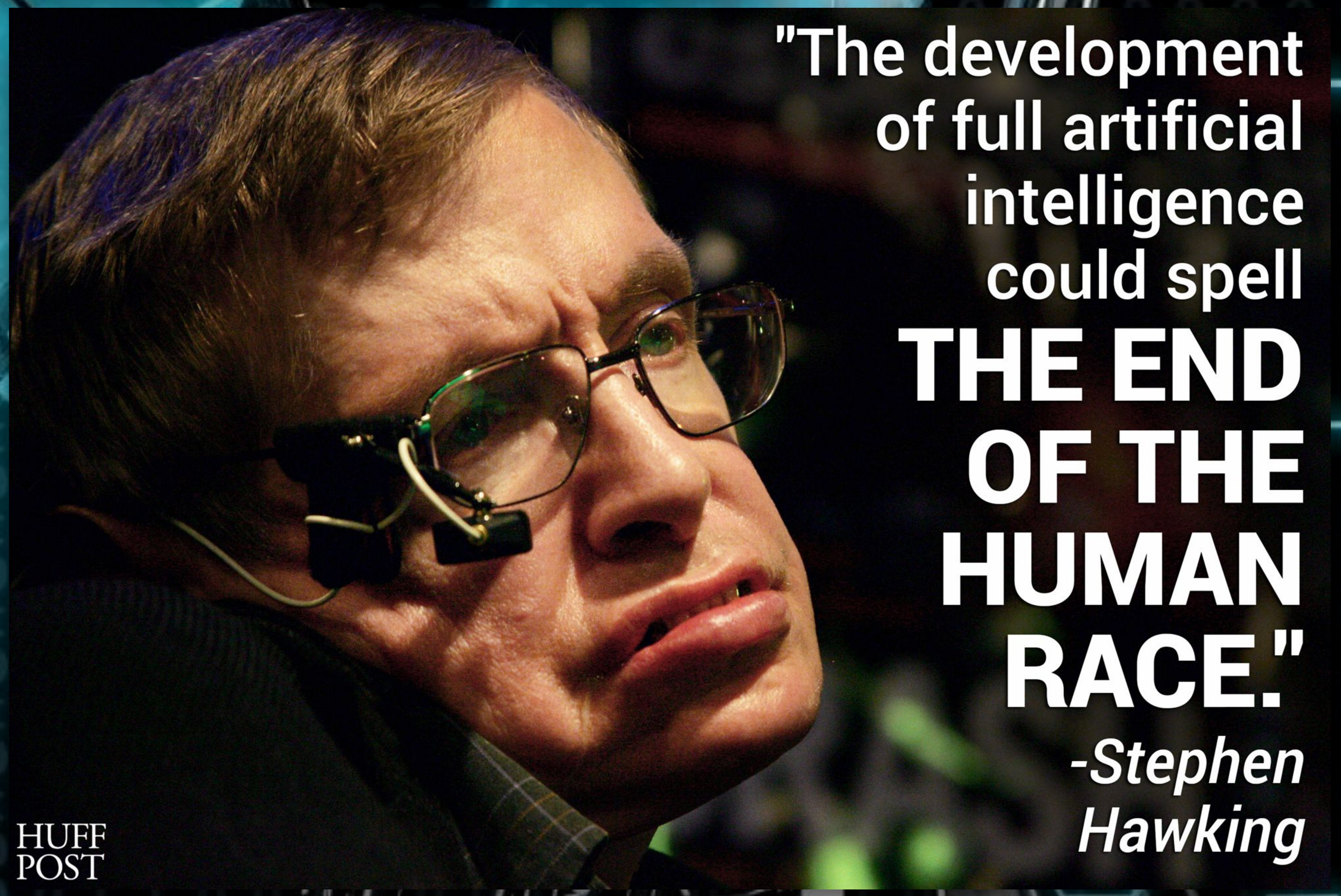

"WITH ARTIFICIAL INTELLIGENCE WE ARE SUMMONING THE DEMON."
-ELON MUSK
HUFF POST
"The development of full artificial intelligence could spell THE END OF THE HUMAN RACE."
-Stephen Hawking
HUFF POST

- AI can affect our lives in many “micro” ways rather than one big “macro” way



- AI can affect our lives in many "micro" ways rather than one big "macro" way

- But ... it is not as hopeless as it seems!



This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

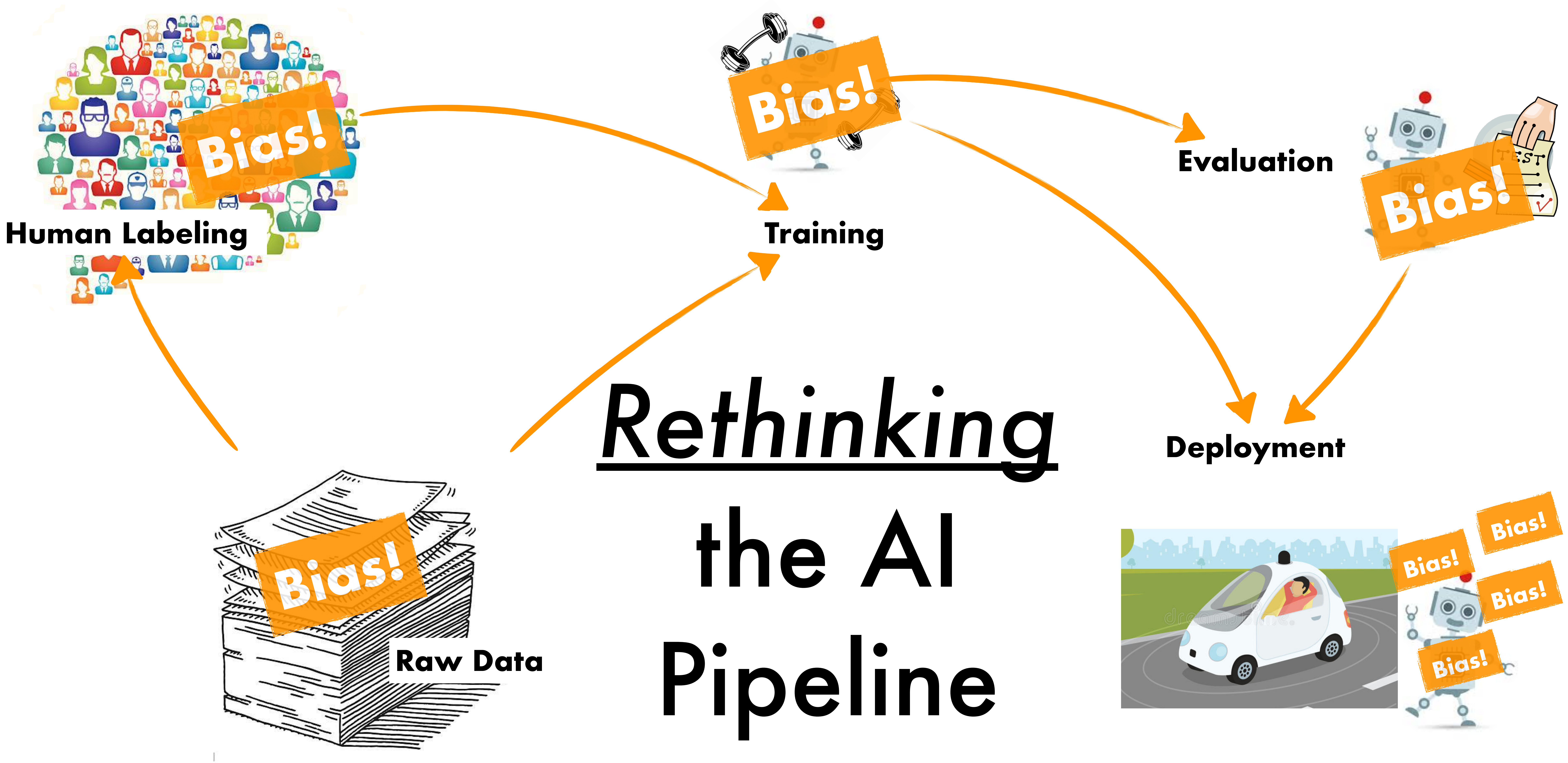
Addressing Biases

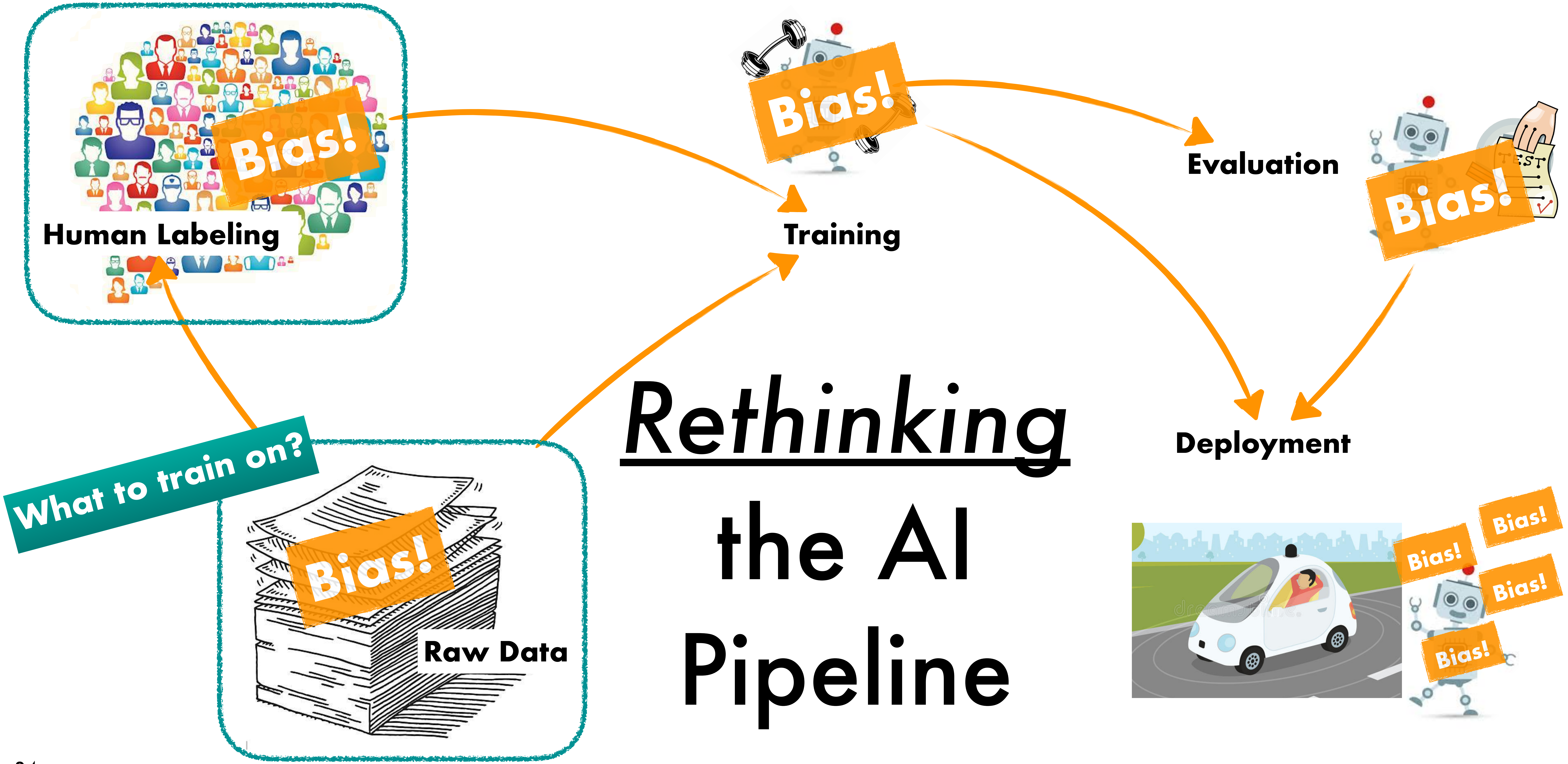
- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

Rethinking the AI Pipeline



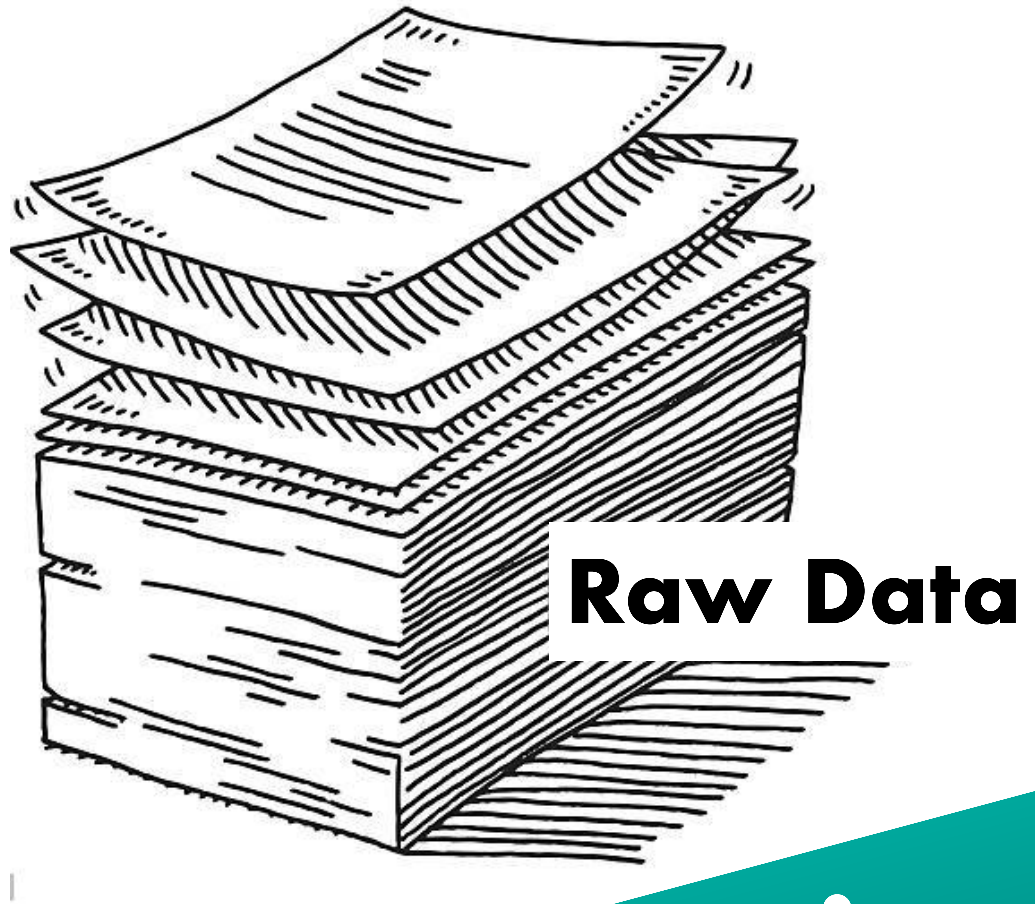

Human Labeling
Bias!
Training
Bias!
Evaluation
Bias!
Deployment
Bias!
Bias!
Bias!
Bias!
What to train on?
Bias!
Raw Data
Rethinking
the AI
Pipeline

Rethinking the AI Pipeline

Rethinking the AI Pipeline

Rethinking the AI Pipeline

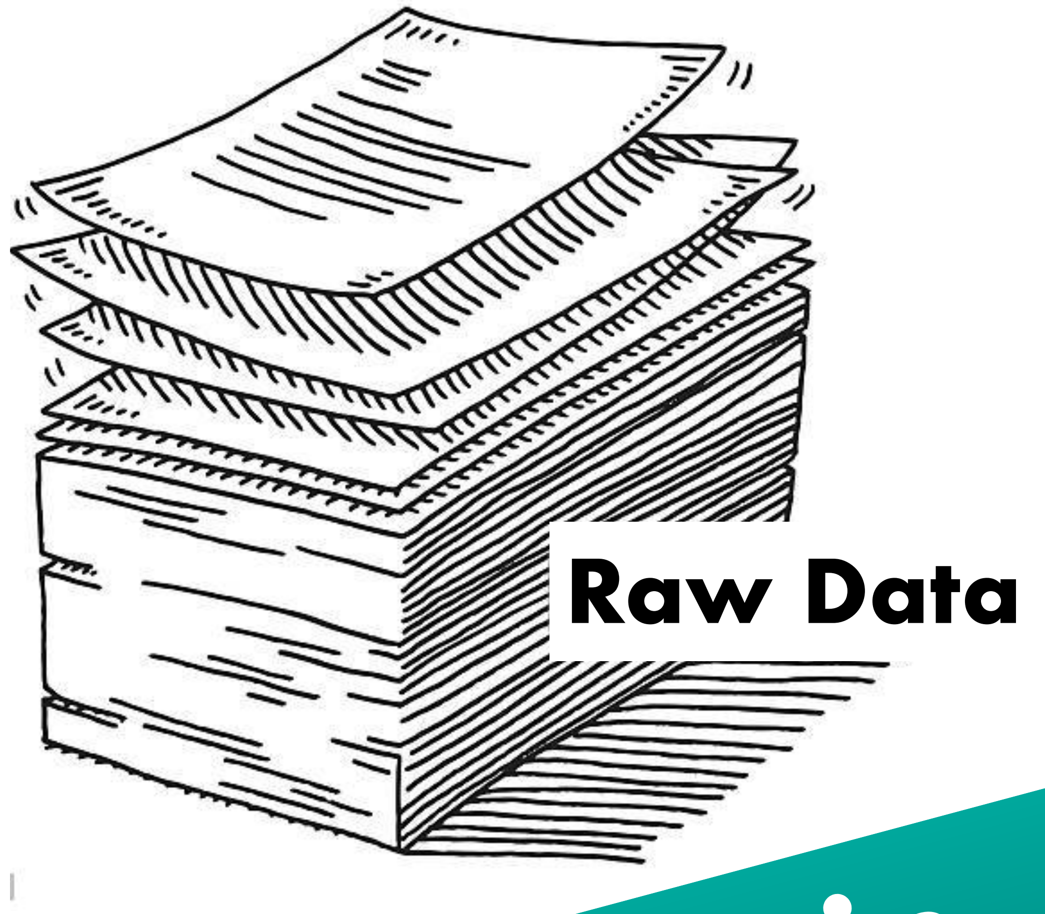
Educating AI: Raw Data



What to train on?

Educating AI: Raw Data

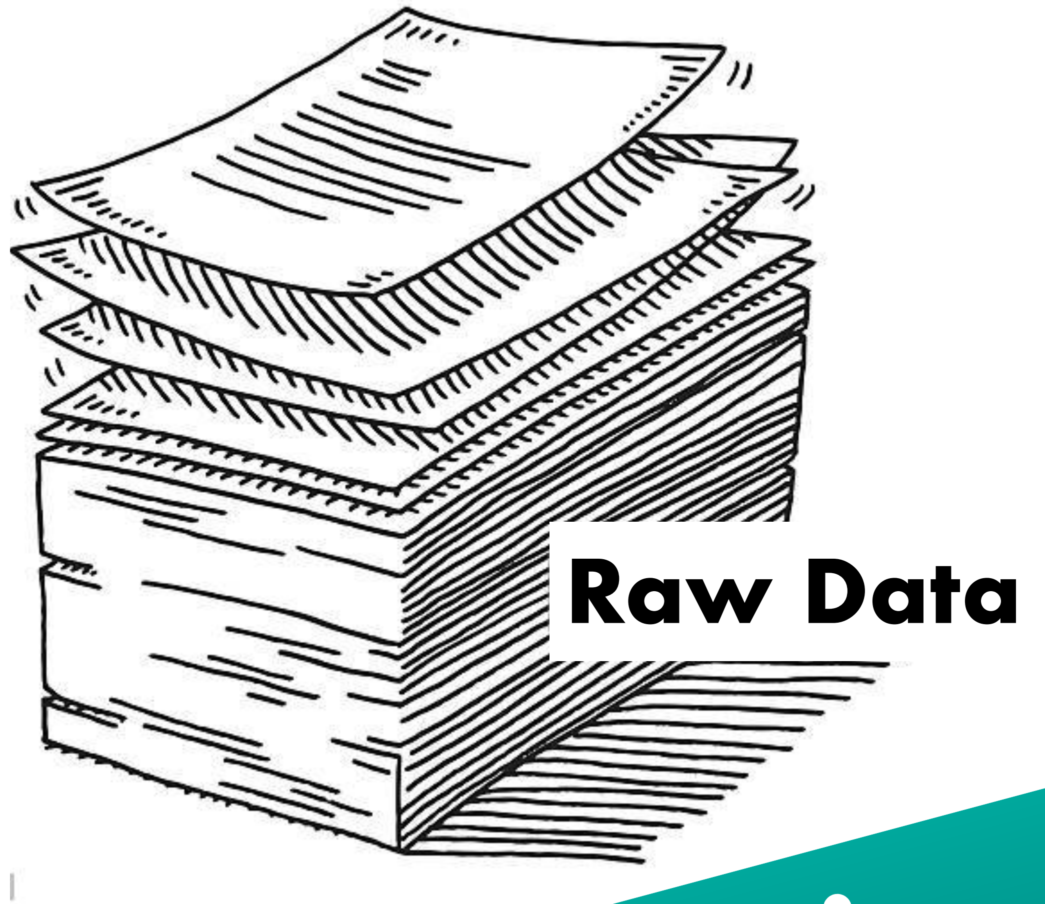
- Curate data with care



What to train on?

Educating AI: Raw Data

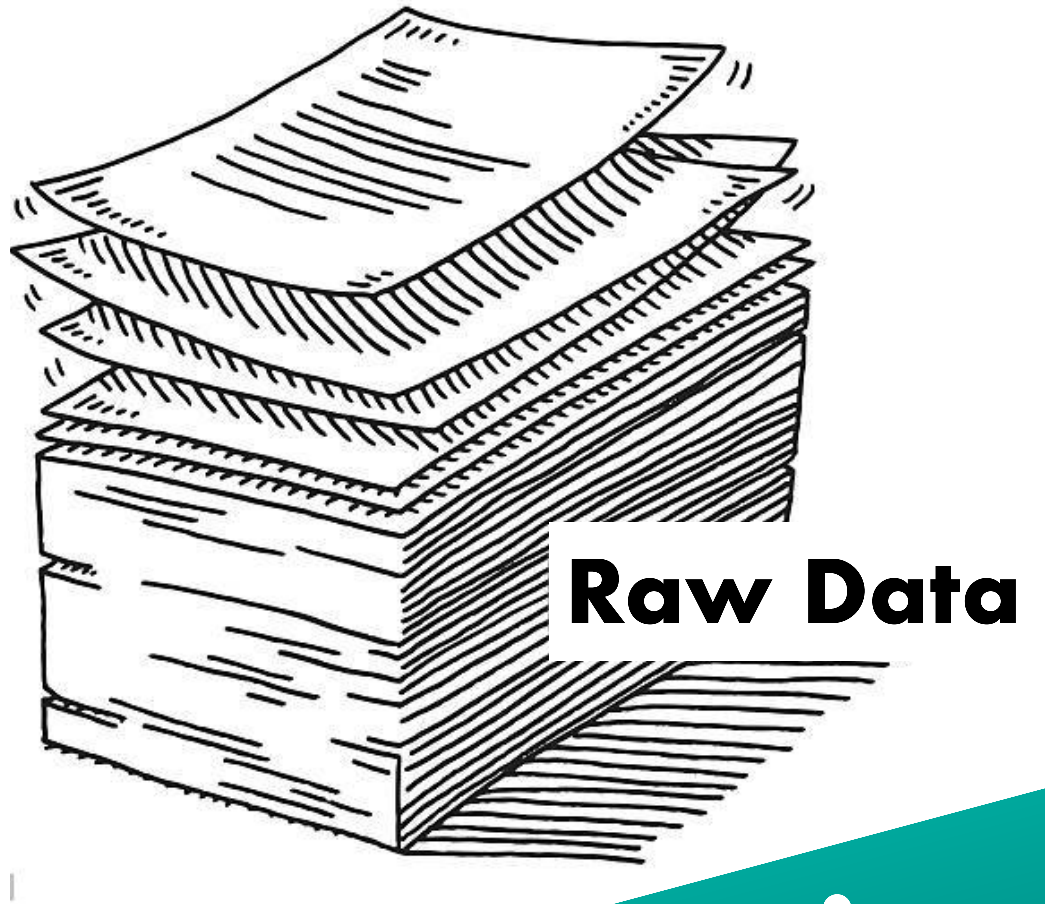
- Curate data with care
    - Sampling Biases. e.g. data containing only white / majority populations



What to train on?

Educating AI: Raw Data

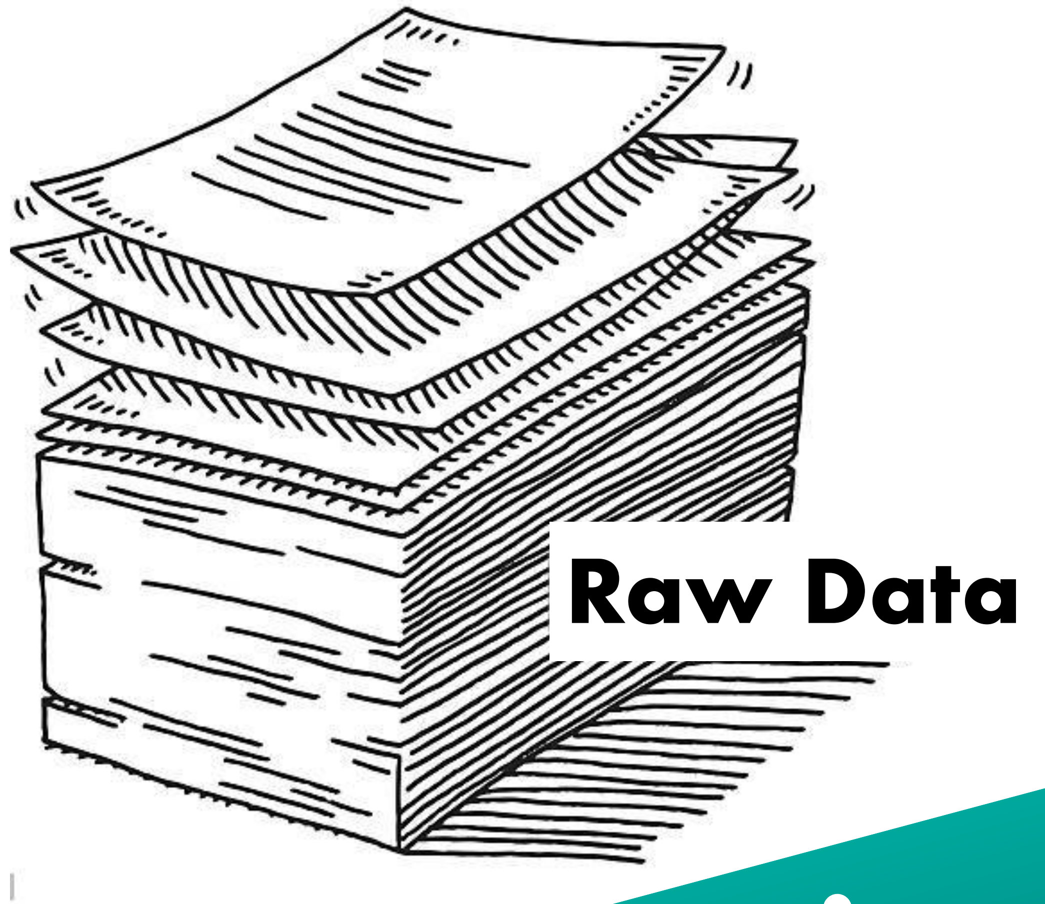
- Curate data with care
  - Sampling Biases. e.g. data containing only white / majority populations
- Dynamic Datasets and Benchmarks



What to train on?

Educating AI: Raw Data

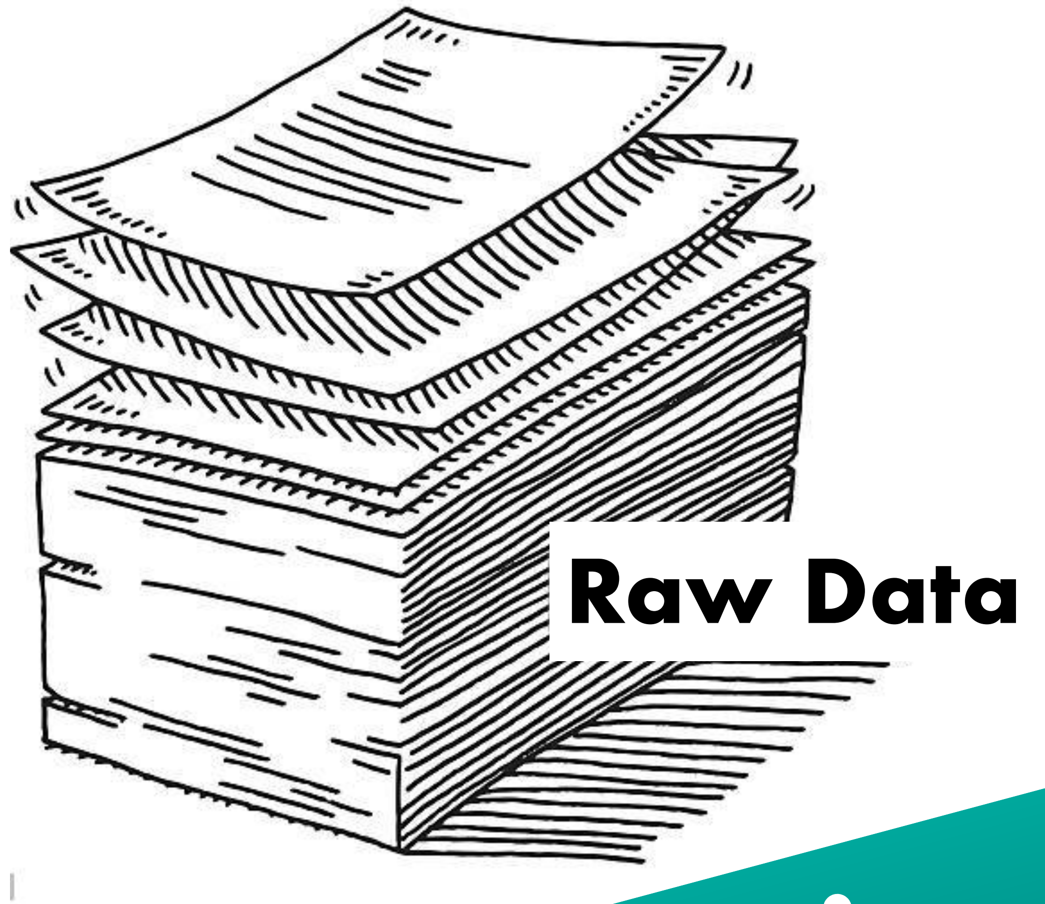
- Curate data with care
    - Sampling Biases. e.g. data containing only white / majority populations
- Dynamic Datasets and Benchmarks
    - Periodic Iterations on Data and Annotations



What to train on?

Educating AI: Raw Data

- Curate data with care
  - Sampling Biases. e.g. data containing only white / majority populations
- Dynamic Datasets and Benchmarks
  - Periodic Iterations on Data and Annotations
  - e.g. Dynabench



What to train on?

Educating AI: Human Labeling

What to train on?

Educating AI: Human Labeling

What to train on?

- Annotator Training to avoid inconsistencies (recall bias)

Educating AI: Human Labeling

- Annotator Training to avoid inconsistencies (recall bias)
    - Avoid stereotyping biases

Educating AI: Human Labeling

What to train on?

- Annotator Training to avoid inconsistencies (recall bias)
    - Avoid stereotyping biases
- Whose voice matters?

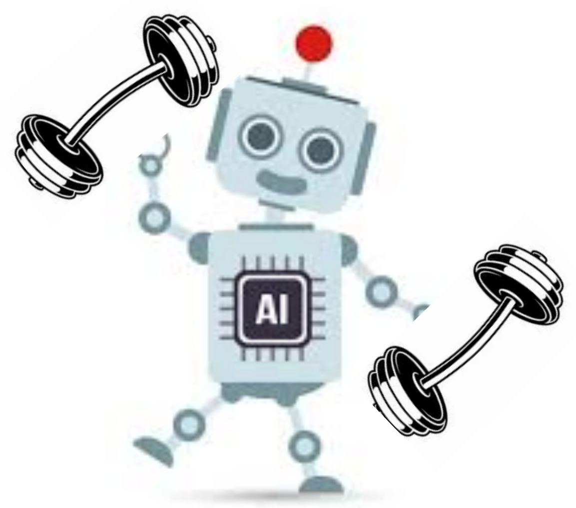
Educating AI: Human Labeling

- Annotator Training to avoid inconsistencies (recall bias)
    - Avoid stereotyping biases
- Whose voice matters?
- Reannotation using a diverse annotator pool / the most affected users

A democratized view of toxic language [V., S., Z., Swayamdipta - In Prep]

Whose perspective is it anyway? [R., P., B., G., Swayamdipta - In Prep]

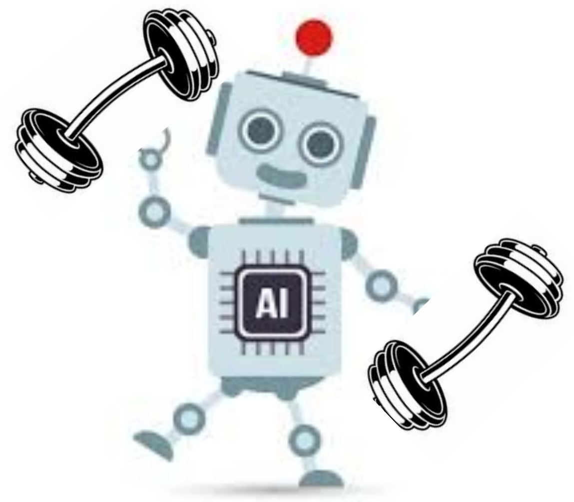
Educating AI: Training



Training

How to train?

Educating AI: Training

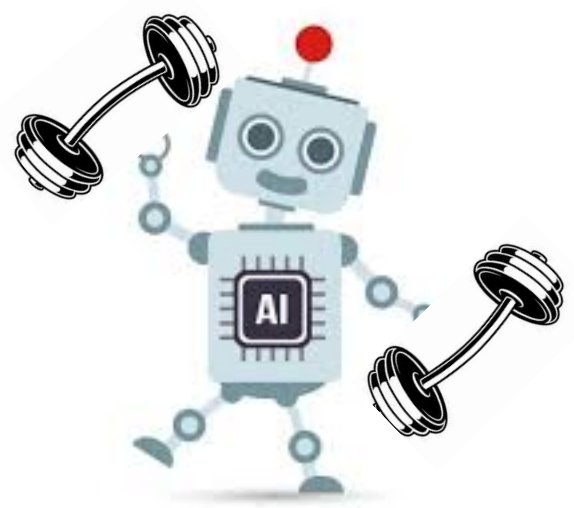


Training

How to train?

Inductive Biases to fight spurious correlations

Educating AI: Training



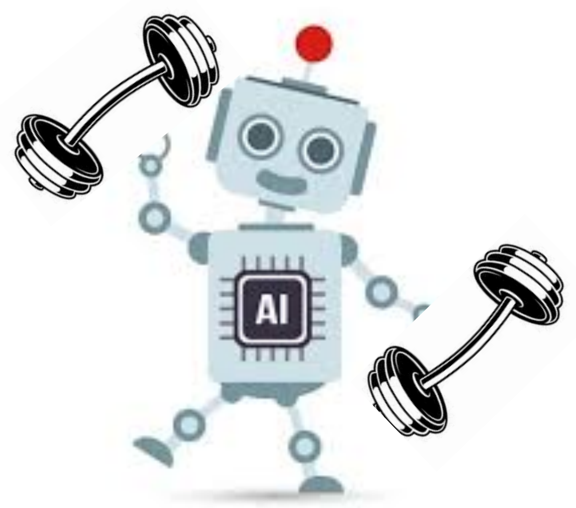
Training

How to train?

Inductive Biases to fight spurious correlations

Inductive Biases to fight Social Biases

Educating AI: Training



Training

How to train?

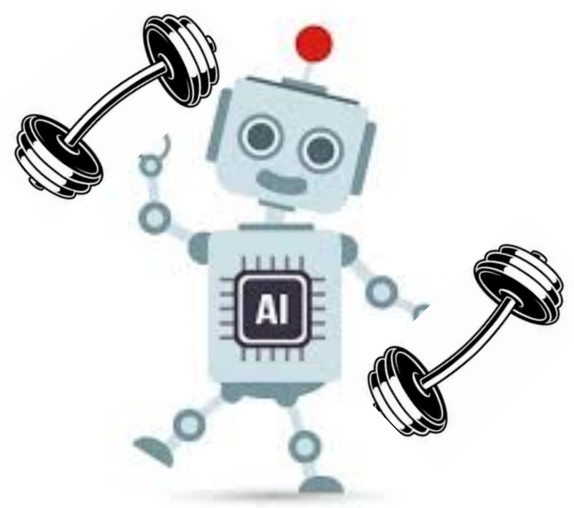
Inductive Biases to fight spurious correlations

Inductive Biases to fight Social Biases

Common Sense

Choi et al. ACL 2020

Educating AI: Training



Training

How to train?

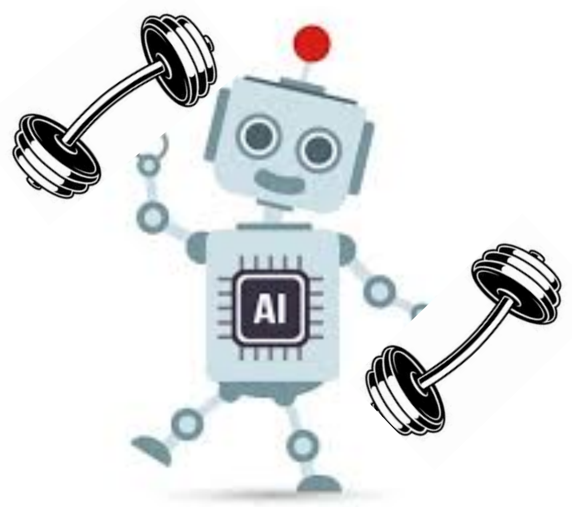
Inductive Biases to fight spurious correlations

Inductive Biases to fight Social Biases

Common Sense

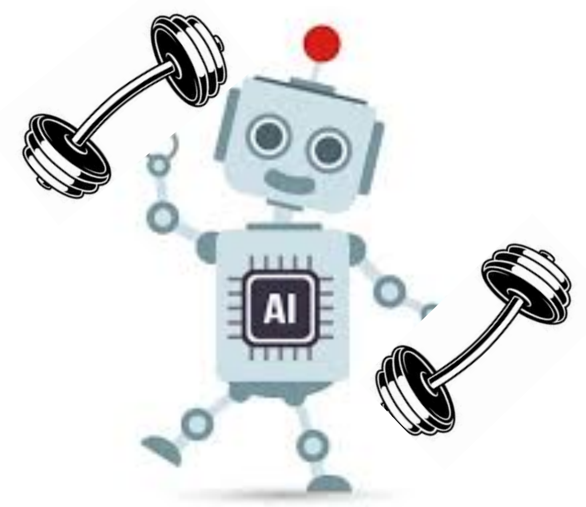
Social Dynamics

Choi et al. ACL 2020



Training

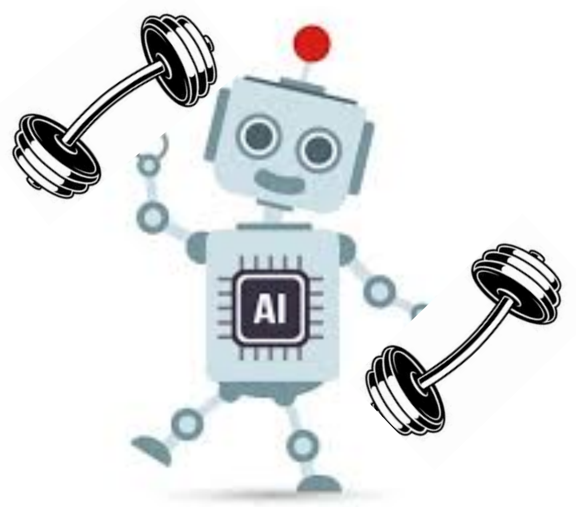
Educating AI: Training



Training

Educating AI: Training

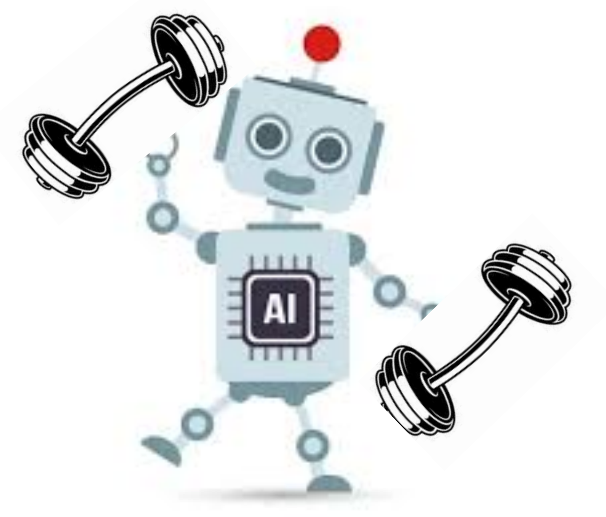
- How to learn? ITERATE!



Training

Educating AI: Training

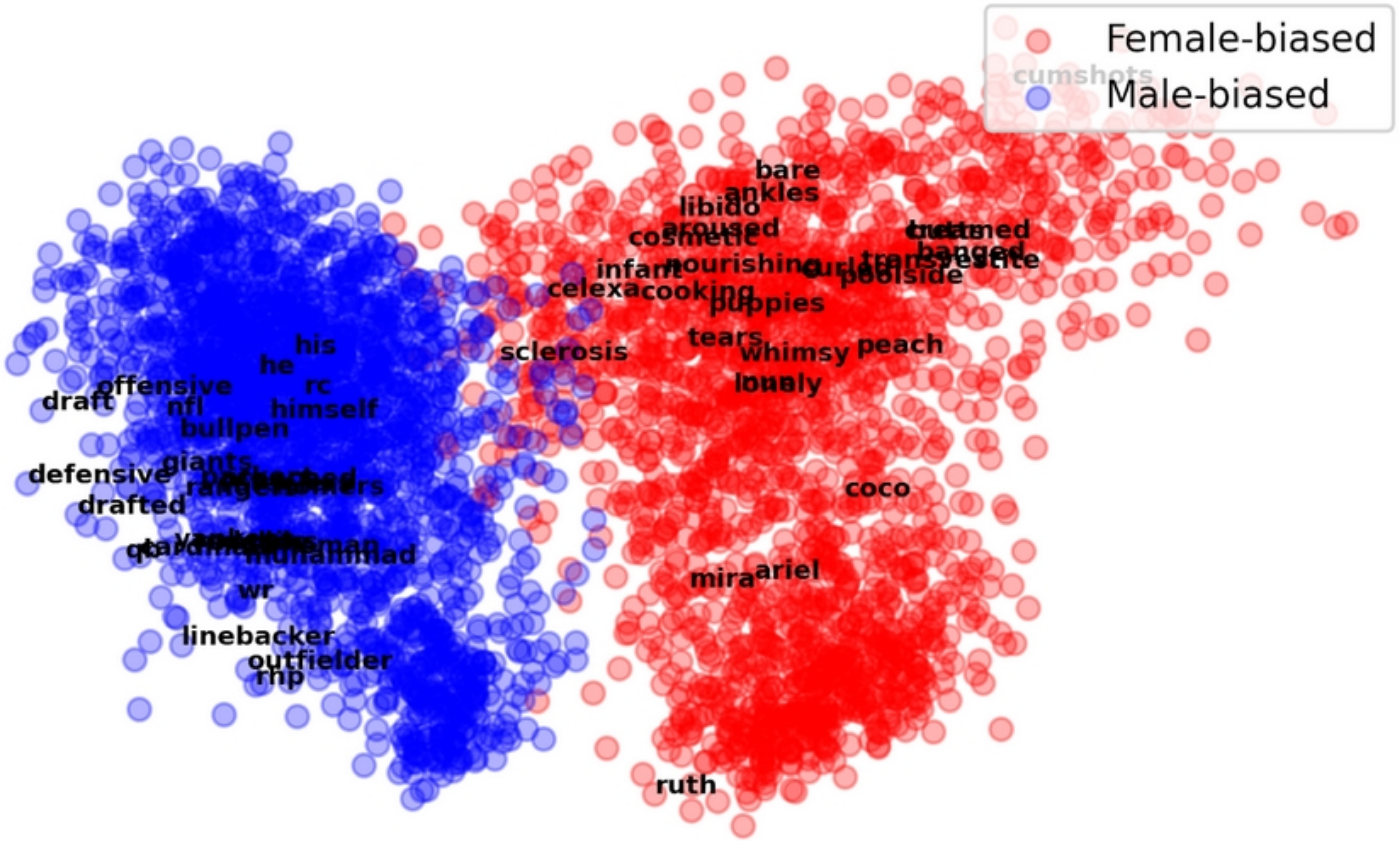
- How to learn? ITERATE!
  - e.g. Removing Gender Bias from Word Embeddings



Training

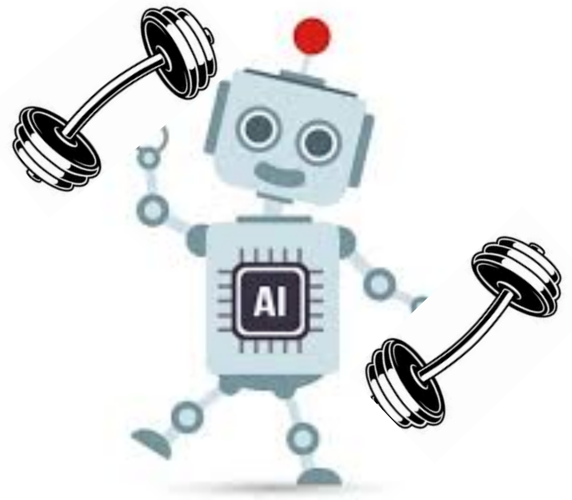
Educating AI: Training

- How to learn? ITERATE!
    - e.g. Removing Gender Bias from Word Embeddings



Bolukbasi et al., 2016; Swinger et al., 2019

Iterative Nullspace Projection for Protected Attribute Removal [Ravfogel et al., 2019]



Training

Educating AI: Training

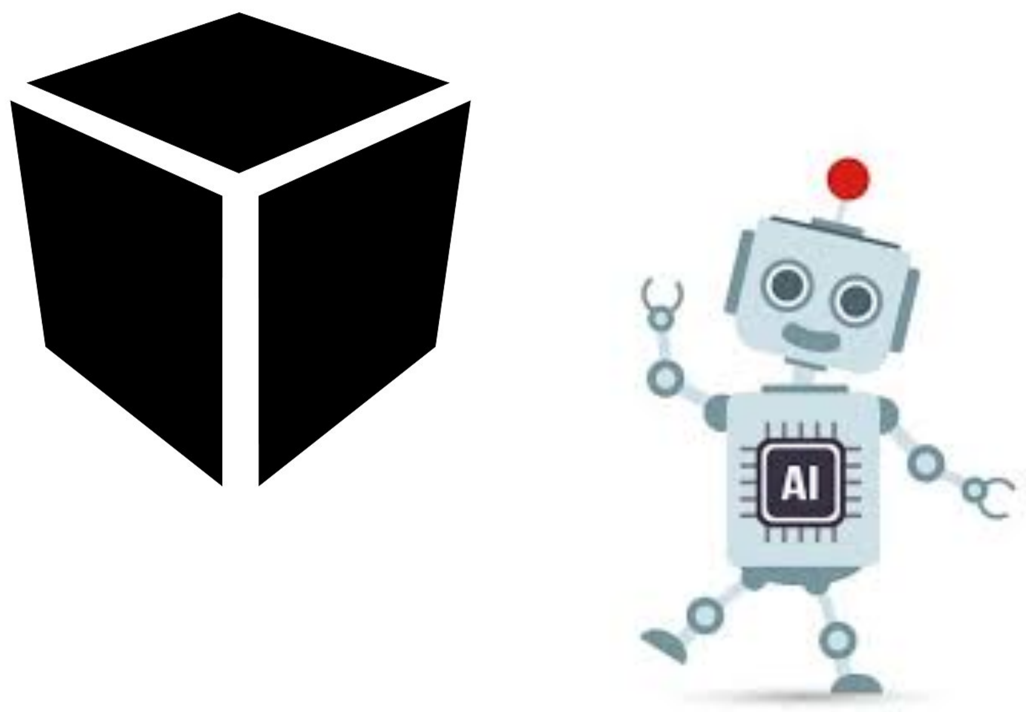
- How to learn? ITERATE!
  - e.g. Removing Gender Bias from Word Embeddings
  - e.g. AFLite

Explanations for evaluating AI

How to evaluate?

Explanations for evaluating AI

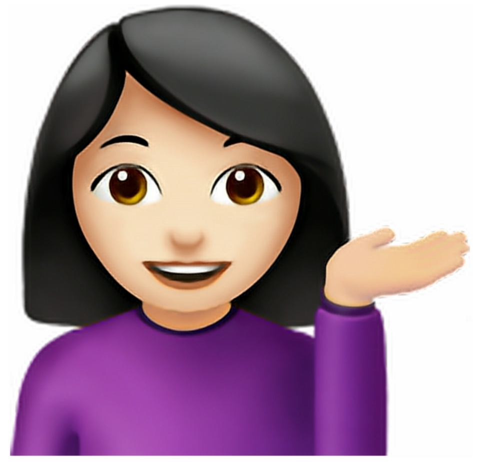
- AI is notorious for being a black box: we cannot simply take an AI decision for granted



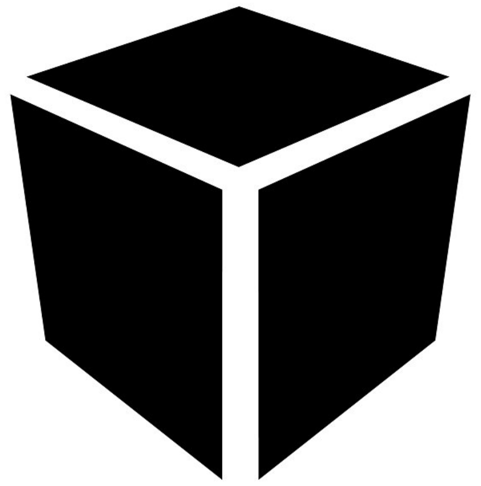
How to evaluate?

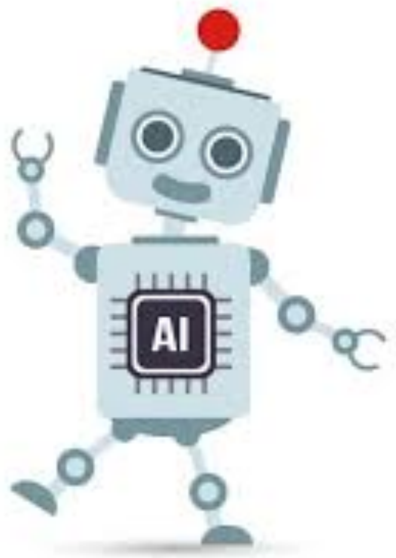
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted



Evaluation

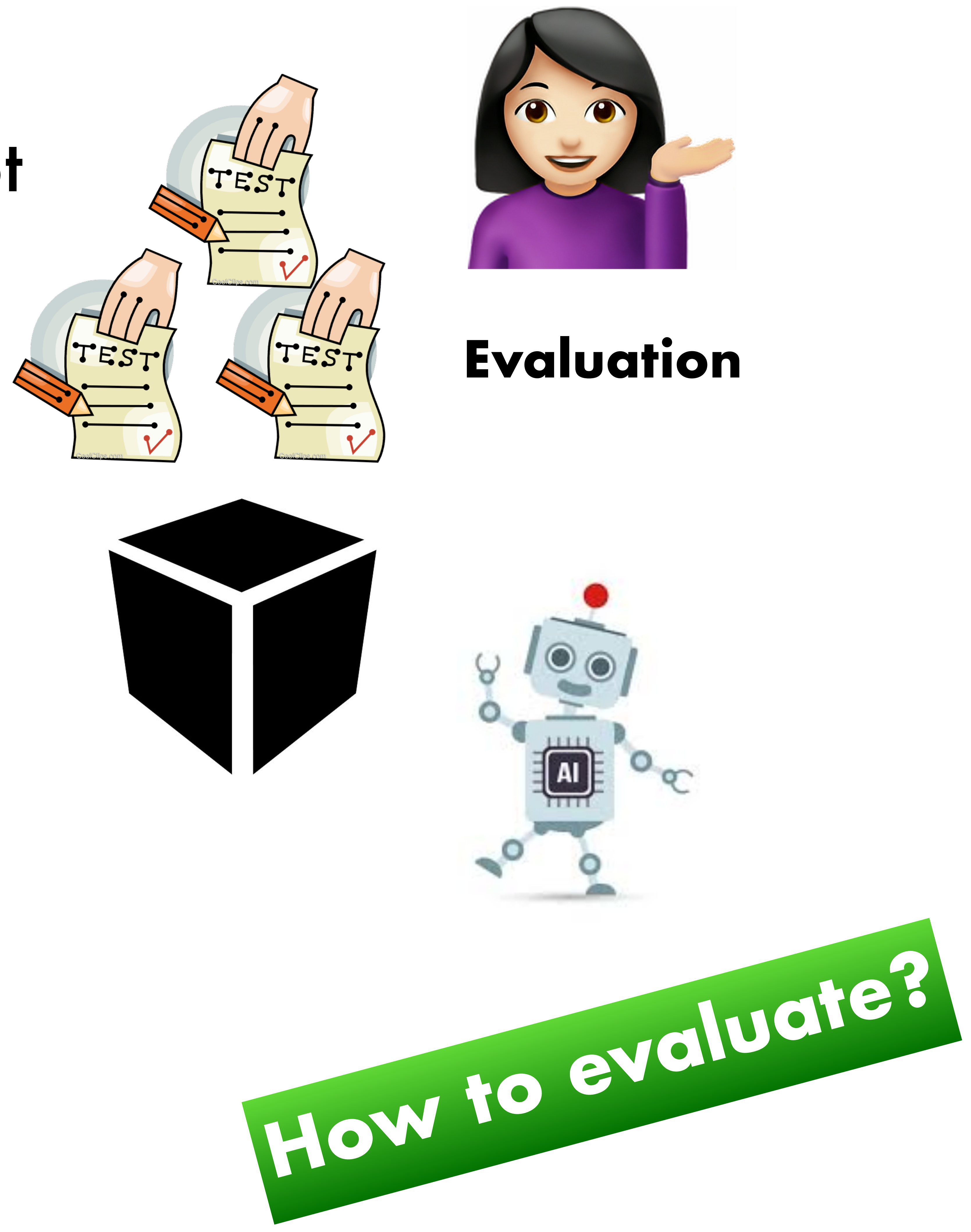




How to evaluate?

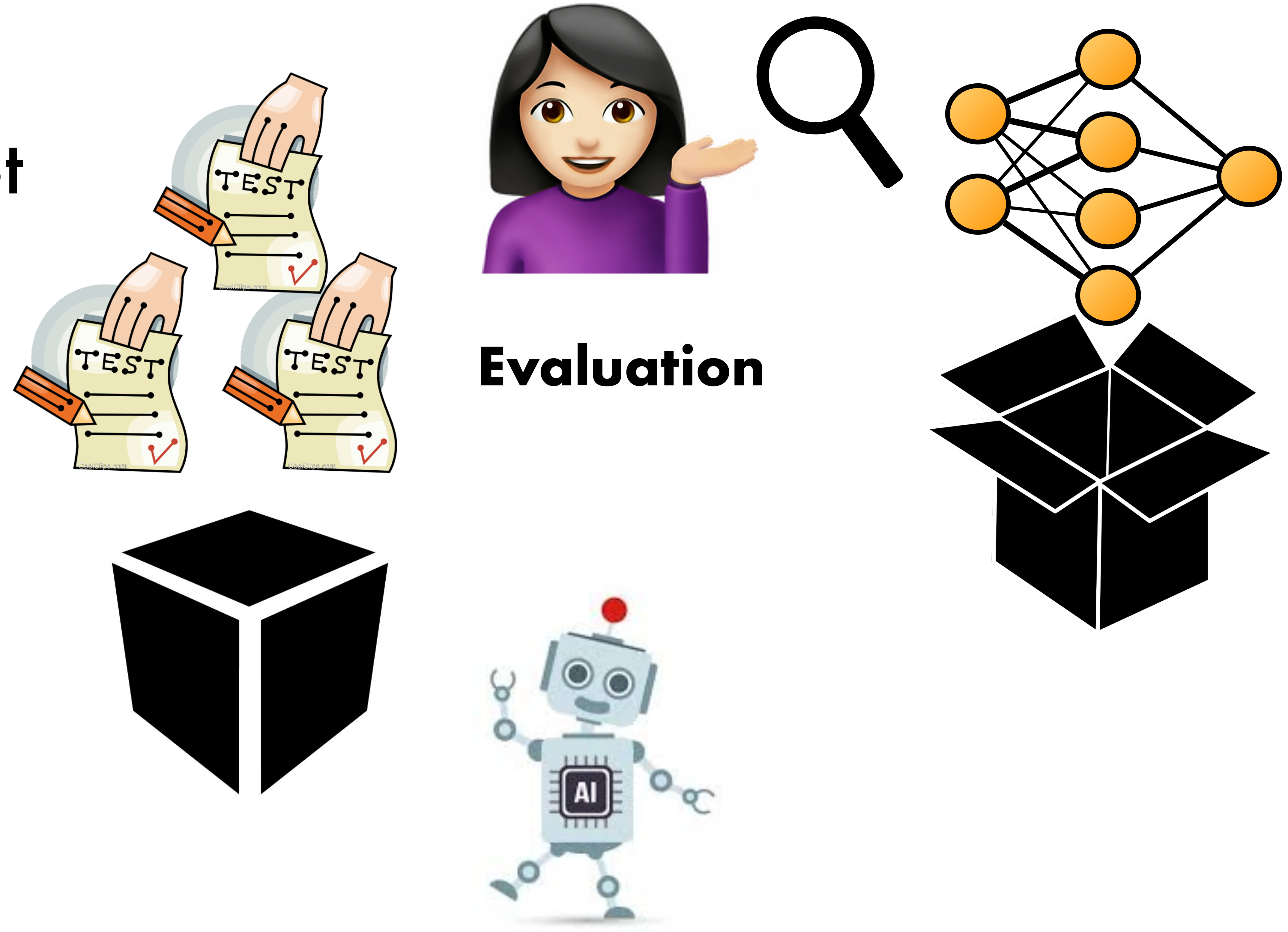
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted
    - Behavioral Testing



Explanations for evaluating AI

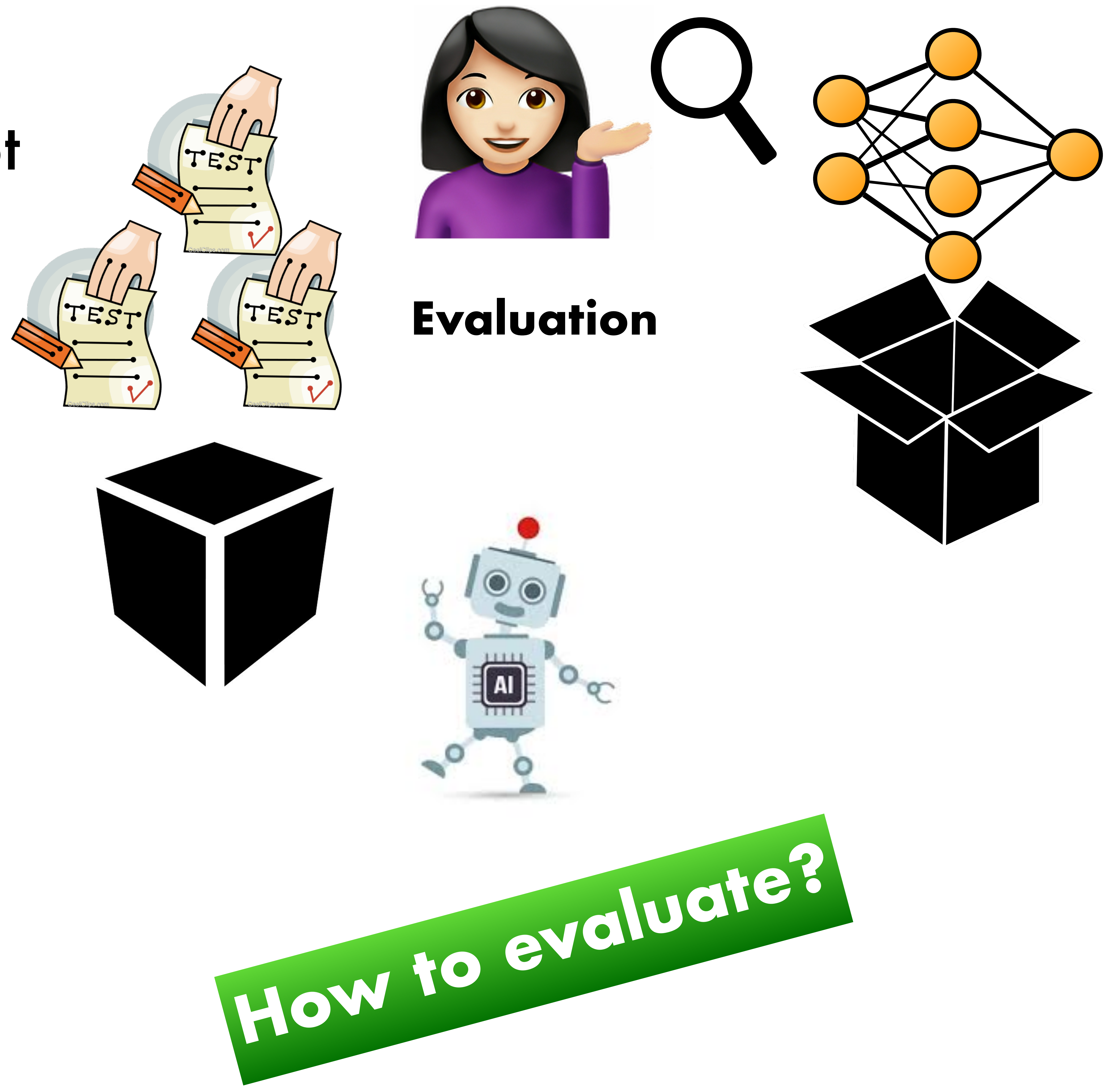
- AI is notorious for being a black box: we cannot simply take an AI decision for granted
  - Behavioral Testing
  - Examining model internals



How to evaluate?

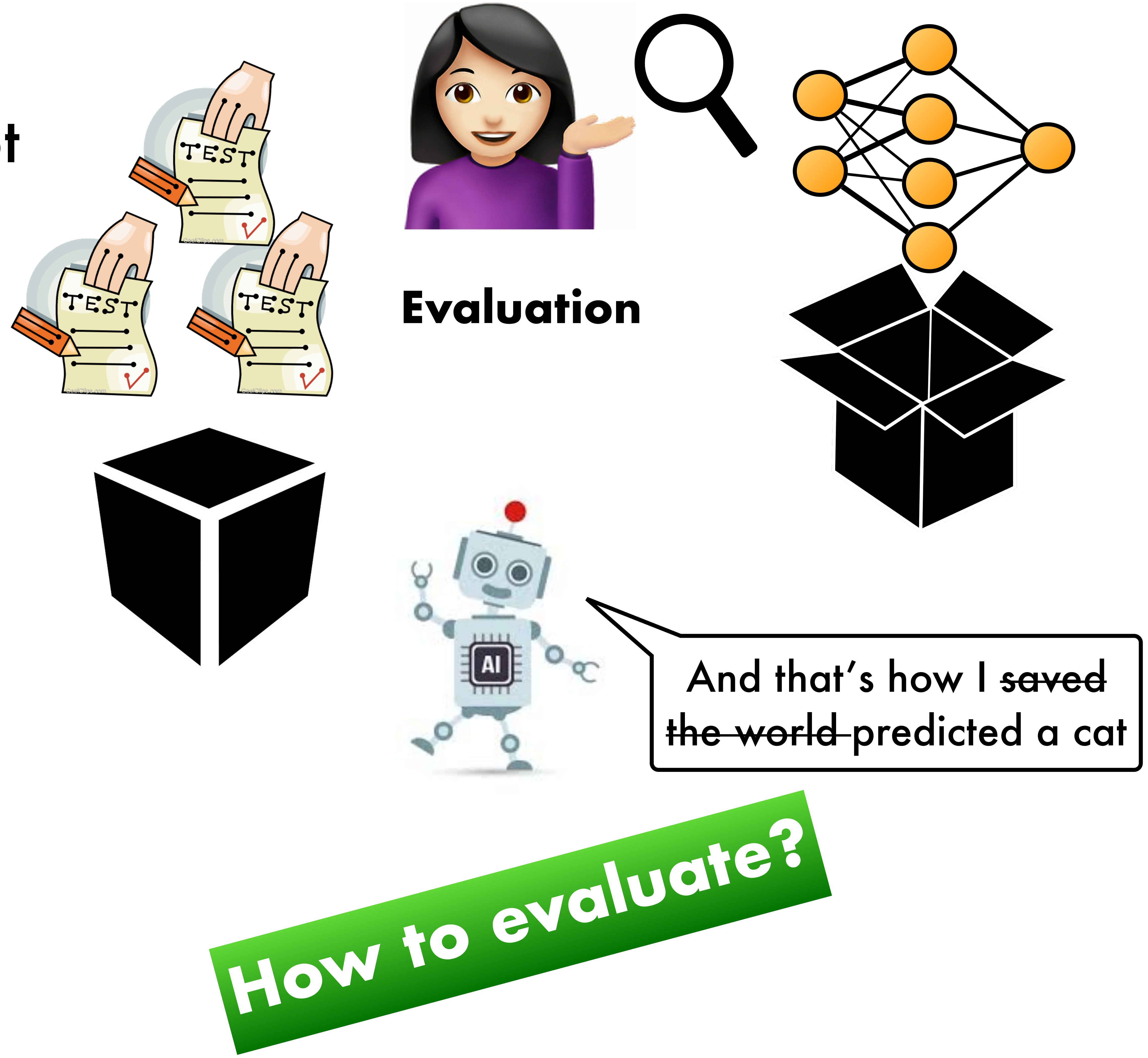
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted
    - Behavioral Testing
    - Examining model internals
- Biases in models can be exposed through explainability



Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted
    - Behavioral Testing
    - Examining model internals
- Biases in models can be exposed through explainability
- Important for building trust (Jacovi et al. 2020)



Contextualize the Decisions

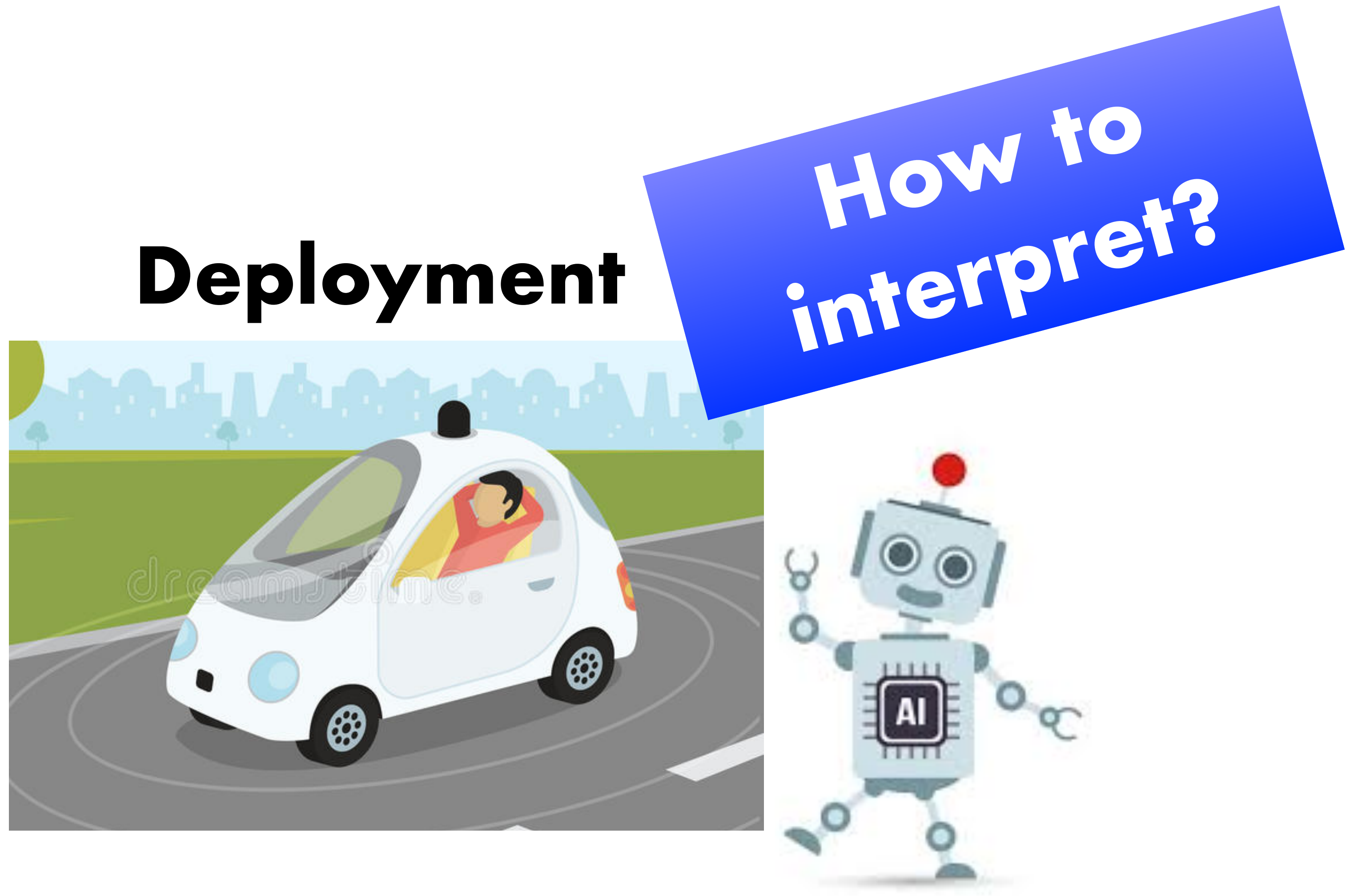
Deployment

How to interpret?

Contextualize the Decisions

- Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]

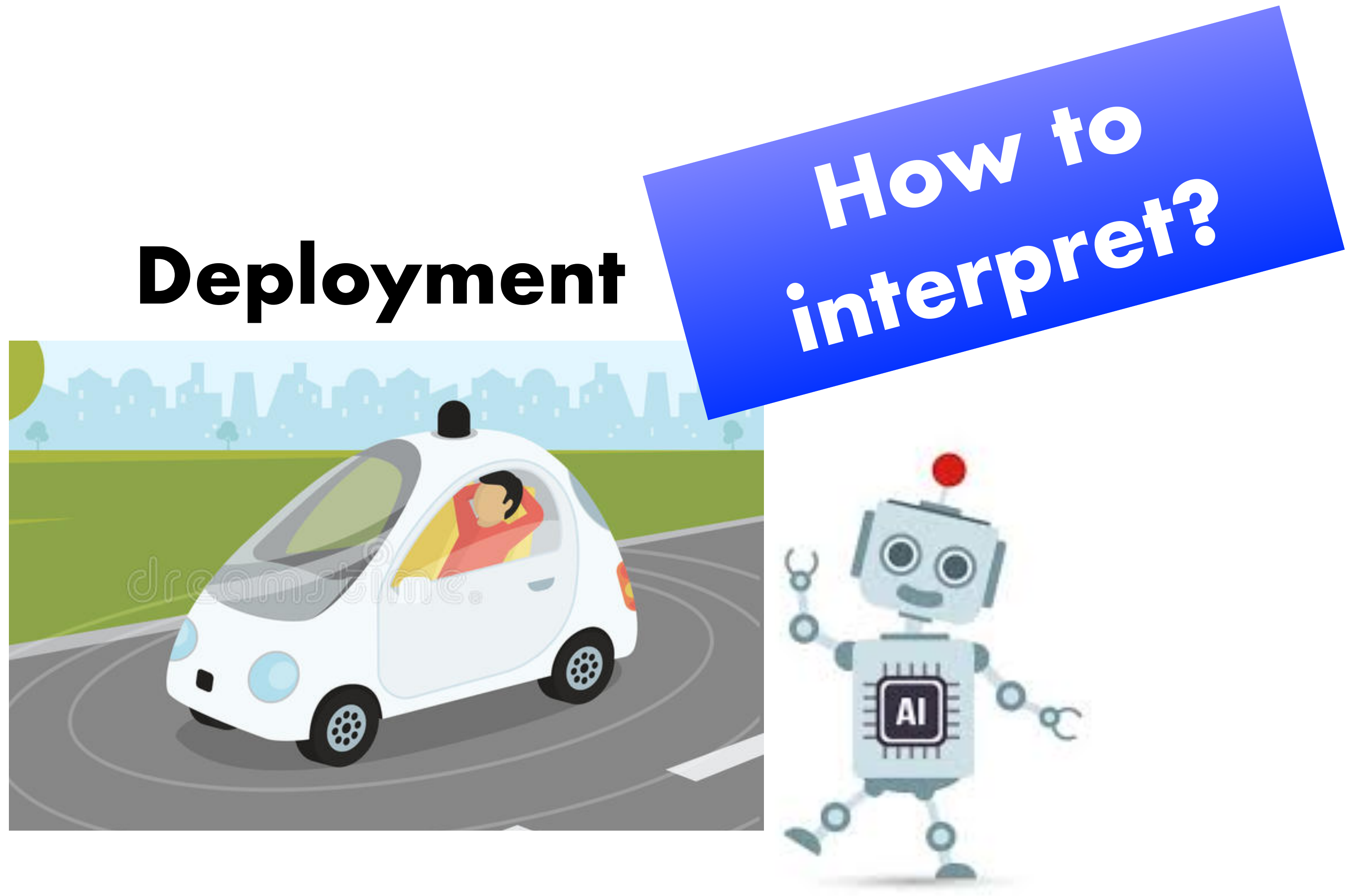
Deployment



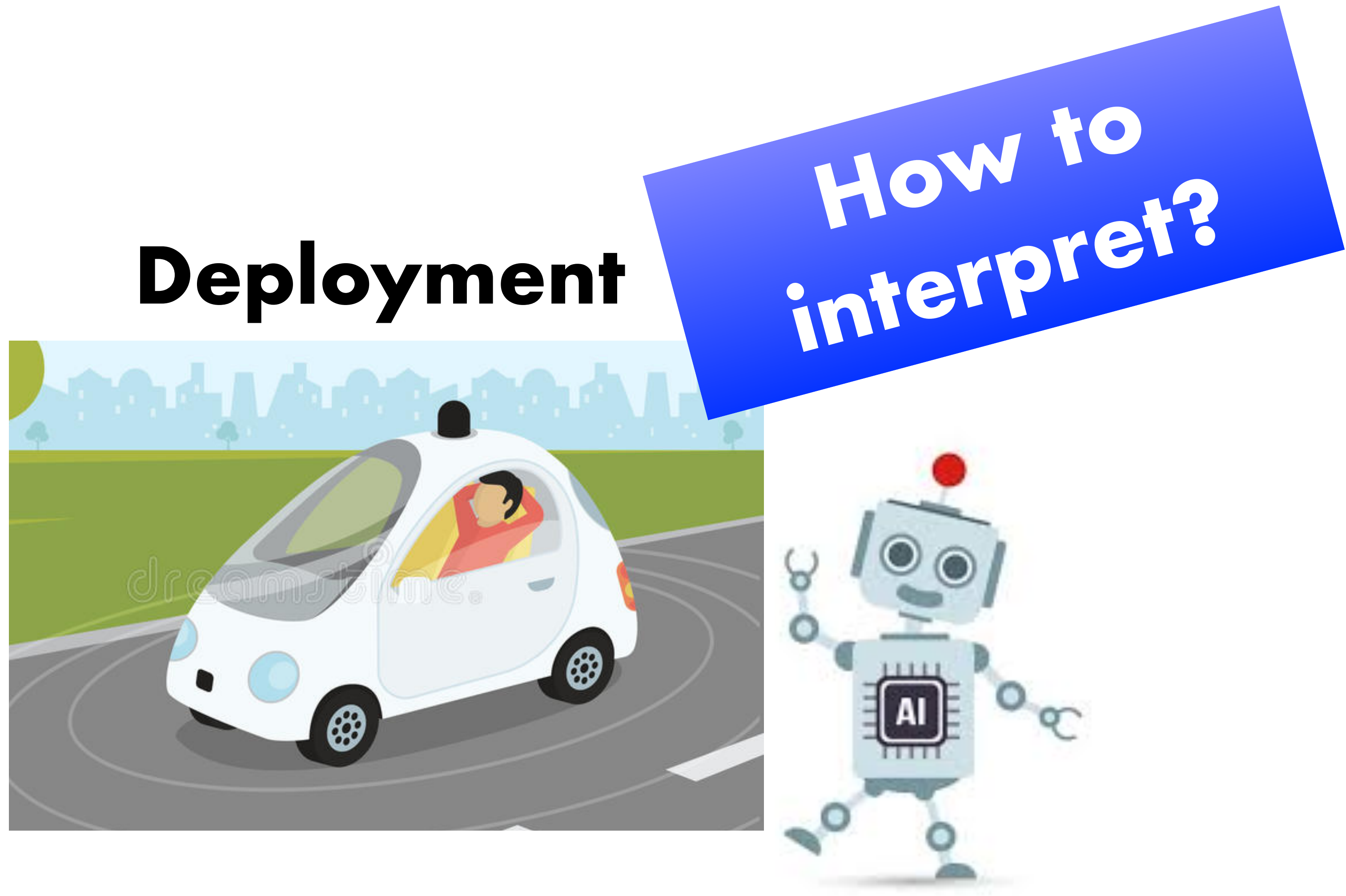
Contextualize the Decisions

- Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]

- Should I trust a decision knowing where it might be coming from?



Contextualize the Decisions



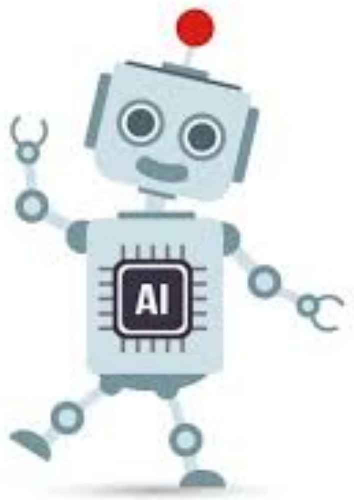
- Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]
- Should I trust a decision knowing where it might be coming from?
    - Datasheets for Datasets [Gebru et al., 2018]

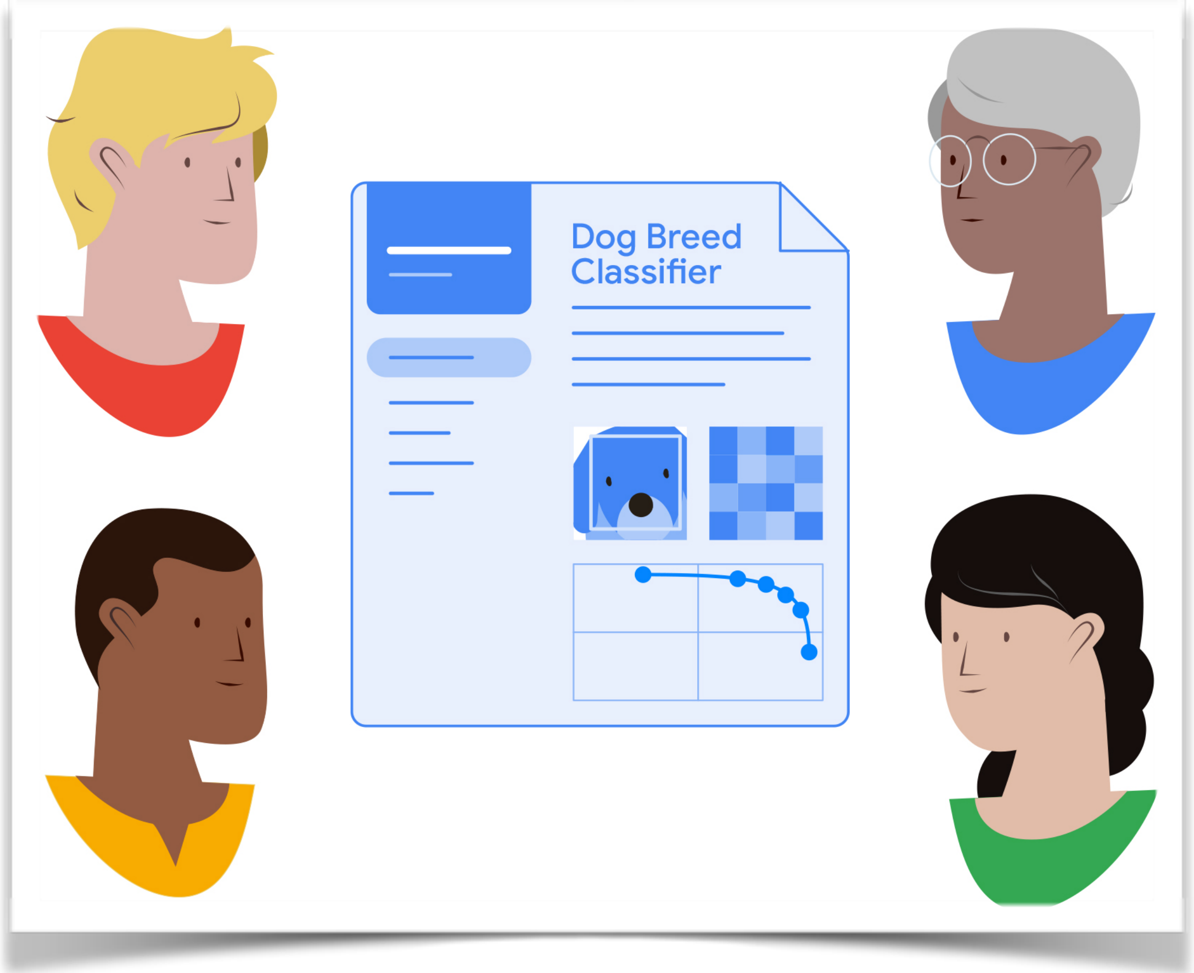
Contextualize the Decisions

- Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]
- Should I trust a decision knowing where it might be coming from?
  - Datasheets for Datasets [Gebru et al., 2018]
  - Model Cards for Model Reporting [Zaldivar et al., 2019]

Deployment

How to interpret?





Take-Home Lessons

Take-Home Lessons

- Educate AI

What to train on?

How to train?

Take-Home Lessons

- Educate AI
- Evaluate AI via Explanations

What to train on?

How to train?

How to evaluate?

Take-Home Lessons

- Educate AI
- Evaluate AI via Explanations
- Contextualize AI Decisions

What to train on?
How to train?
How to evaluate?
How to interpret?

Take-Home Lessons

- Educate AI
- Evaluate AI via Explanations
- Contextualize AI Decisions
- Keep the broader picture in mind: What you do matters!

This Talk: In Summary

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

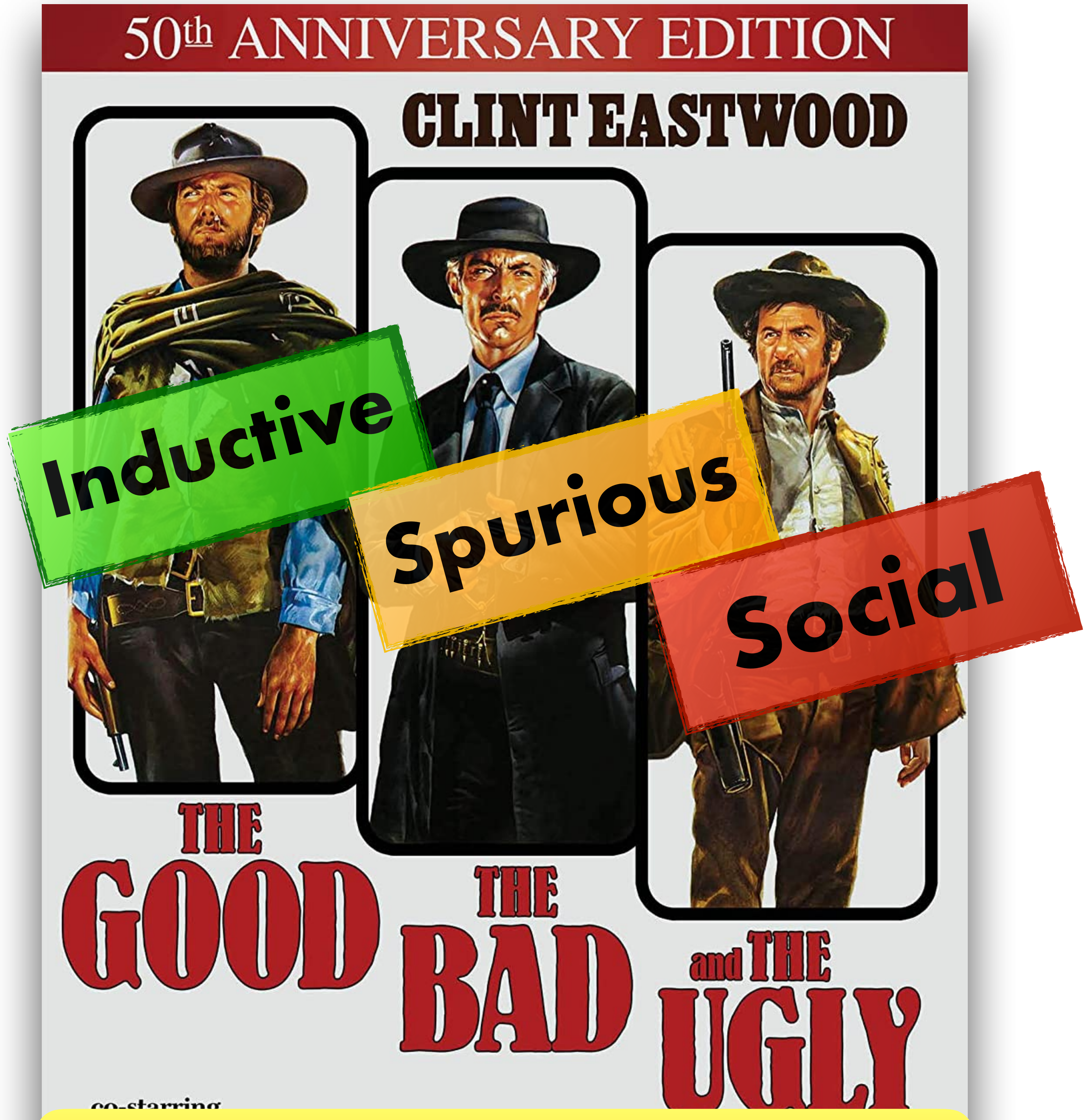
Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

This Talk: In Summary



Data and models may contain different kinds of biases

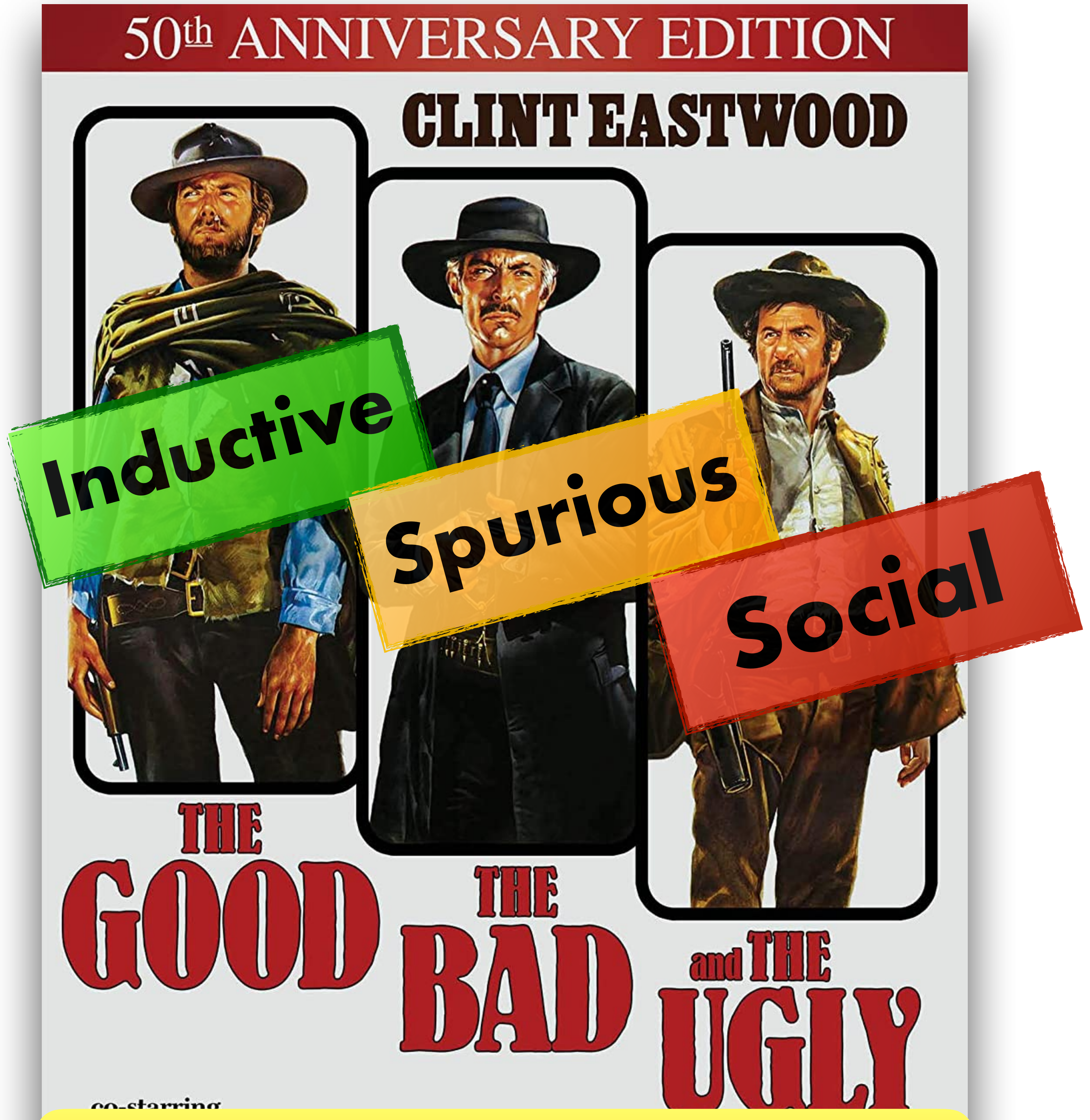
Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize

This Talk: In Summary



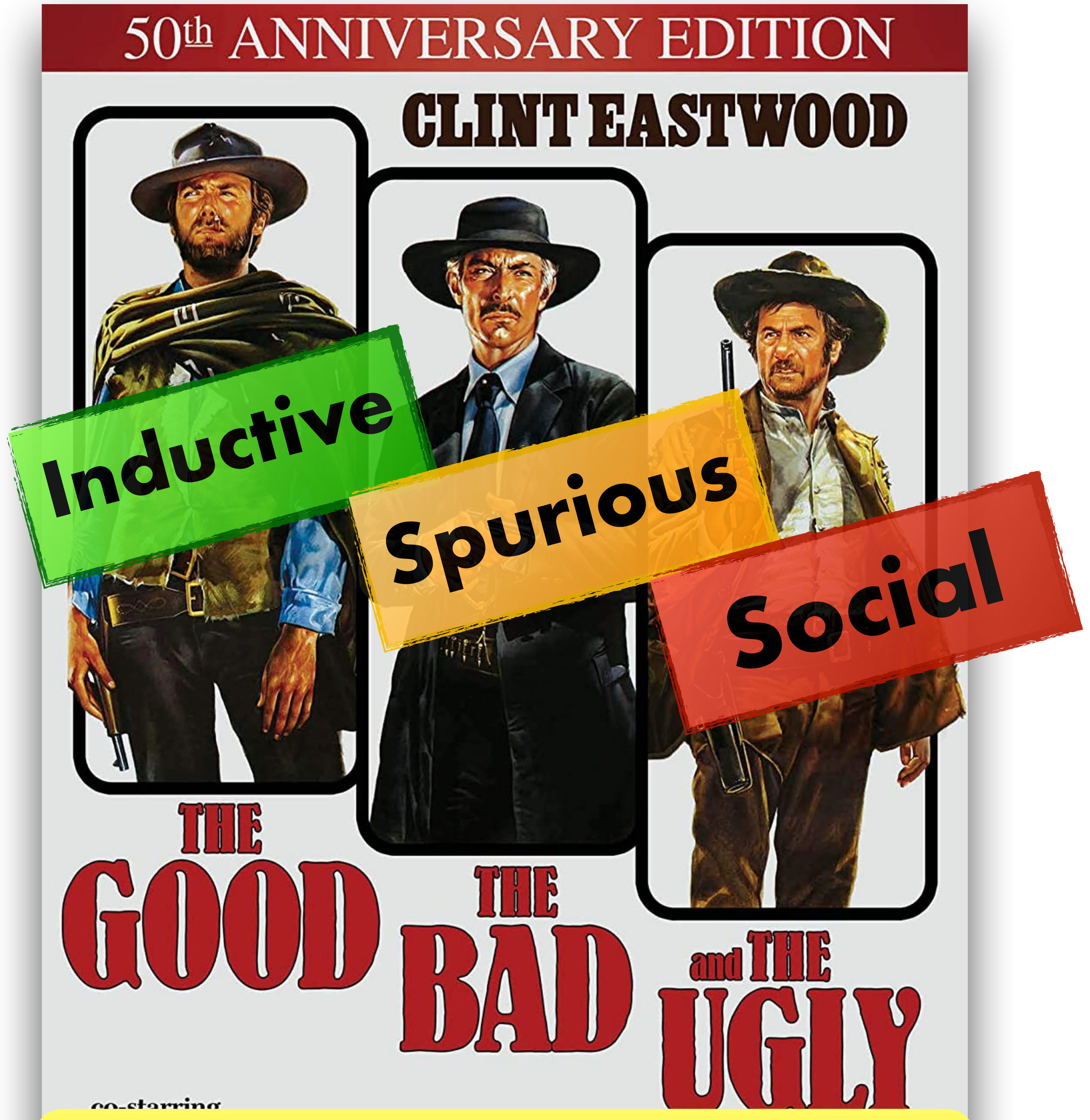
Data and models may contain different kinds of biases

Biases can be extremely tricky to remove

Towards Responsible AI

- Educate
- Explain
- Contextualize

This Talk: In Summary



Data and models may contain different kinds of biases

Biases can be extremely tricky to remove

As the force behind AI, we can really make a difference

Thanks! Questions?

https://swabhs.com

swabhz