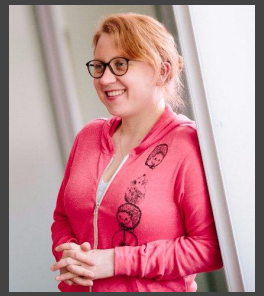
The Seen and Unseen Factors Influencing Knowledge in AI Systems



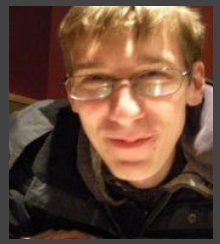
Margaret Mitchell
Google

Ishan Misra
CMU

Larry Zitnick
FAIR

Ross Girshick
FAIR

Adrian Benton
JHU



Dirk Hovy
U. Copenhagen

Josh Lovejoy
Google

Hartwig Adam
Google

Blaise Agüera
y Arcas - Google

What do you see?

What do you see?

- Bananas

What do you see?

- Bananas
- Dole Bananas

What do you see?

- Bananas
- Dole Bananas
- Bananas at a store

What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves

What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them

What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

What do you see?

- Bananas
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...We don’t tend to say

Yellow Bananas

What do you see?

Green Bananas

Unripe Bananas

What do you see?

Ripe Bananas

Bananas with spots

What do you see?

Ripe Bananas

Bananas with spots

Bananas good for banana bread

What do you see?

Yellow Bananas

Yellow **is prototypical for bananas**

Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to** behaviourally and **cognitively usable proportions**

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)

May also store exemplars (Wu & Barsalou, 2009)

Fruit

Bananas
“Basic Level”

Unripe Bananas, Cavendish Bananas

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

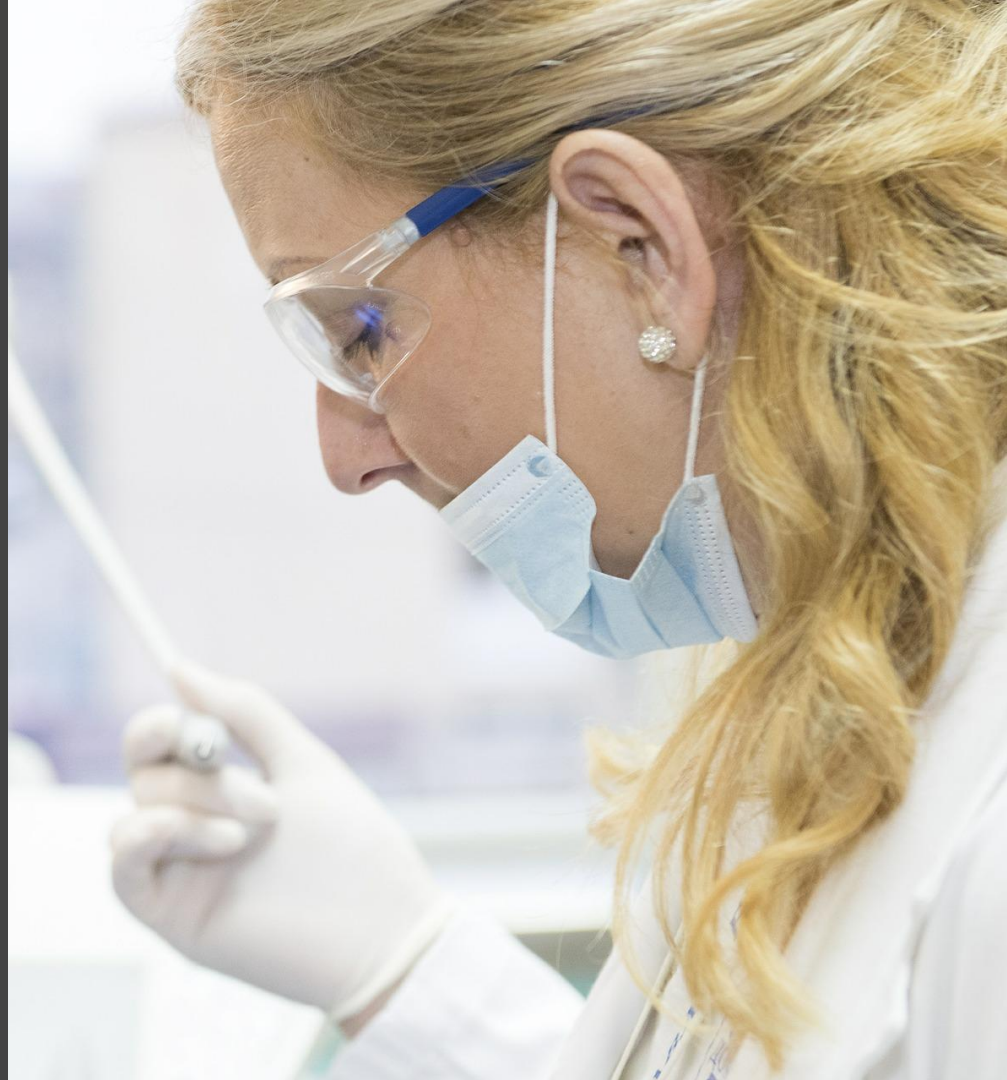
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

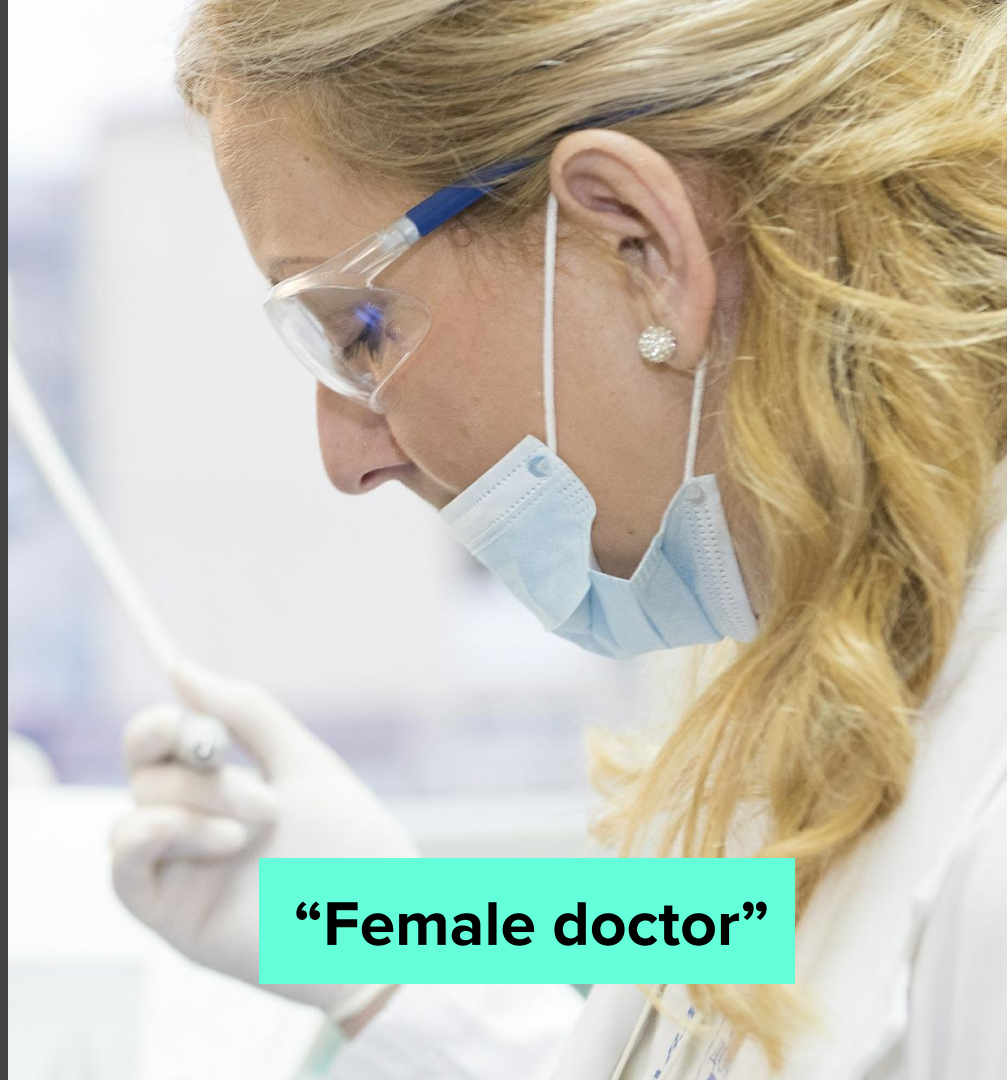
How could this be?



A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?




“Doctor”
“Female doctor”

The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

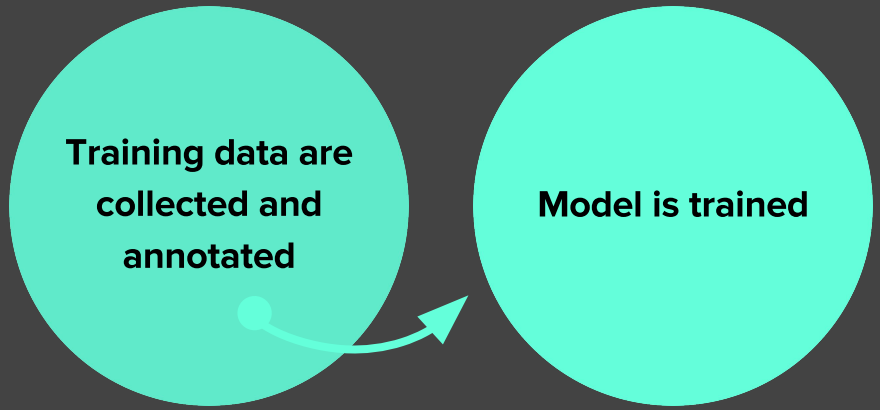
Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

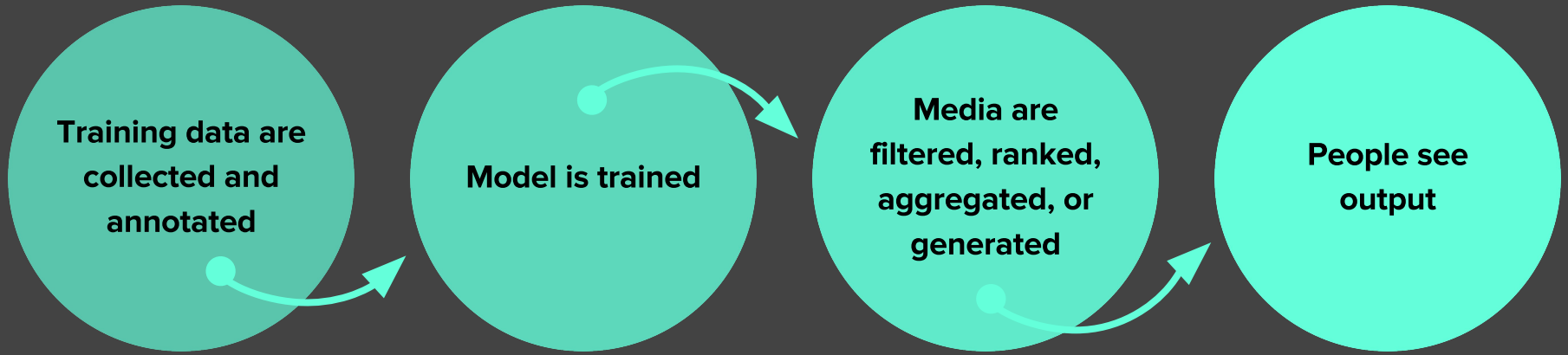
Also called “Black Sheep” problem, “Giraffes problem” in vision/language

Photographer Bias: Natural tendency of photographers to place object of interest in the center

Training data are collected and annotated


Training data are collected and annotated
Model is trained

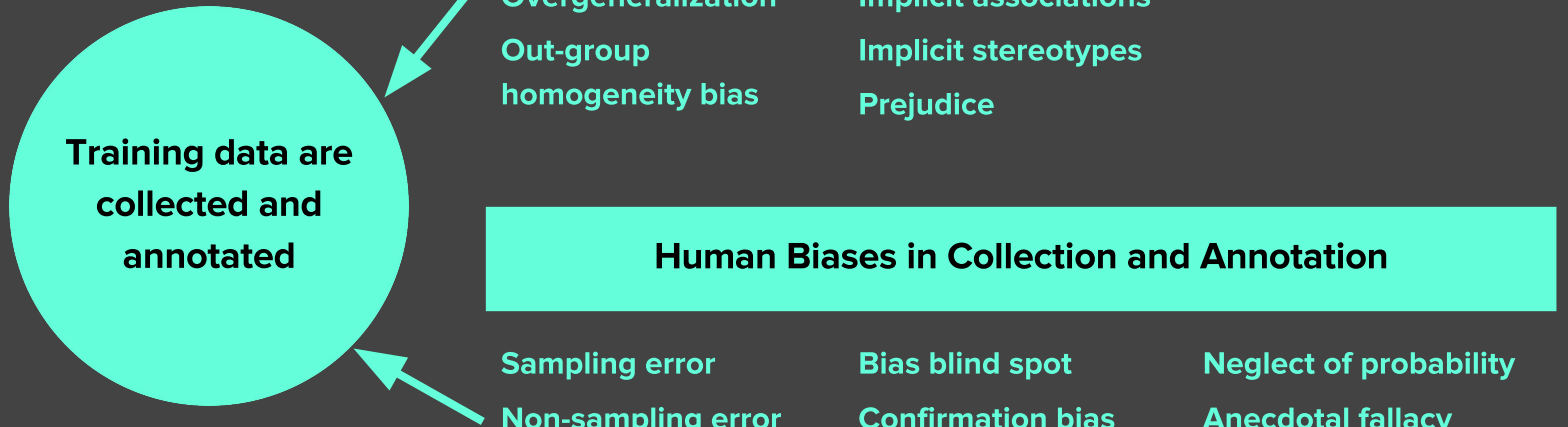

Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated


Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output

Human Biases in Data

Training data are collected and annotated

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect
- Stereotype threat


Training data are collected and annotated
Human Biases in Data
Reporting bias
Selection bias
Overgeneralization
Out-group homogeneity bias
Stereotypical bias
Historical unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group attribution error
Halo effect
Human Biases in Collection and Annotation
Sampling error
Non-sampling error
Insensitivity to sample size
Correspondence bias
In-group bias
Bias blind spot
Confirmation bias
Subjective validation
Experimenter's bias
Choice-supportive bias
Neglect of probability
Anecdotal fallacy
Illusion of validity
Automation bias

Reporting bias: What people share is not a reflection of real-world frequencies

Selection Bias: Selection does not reflect a random sample

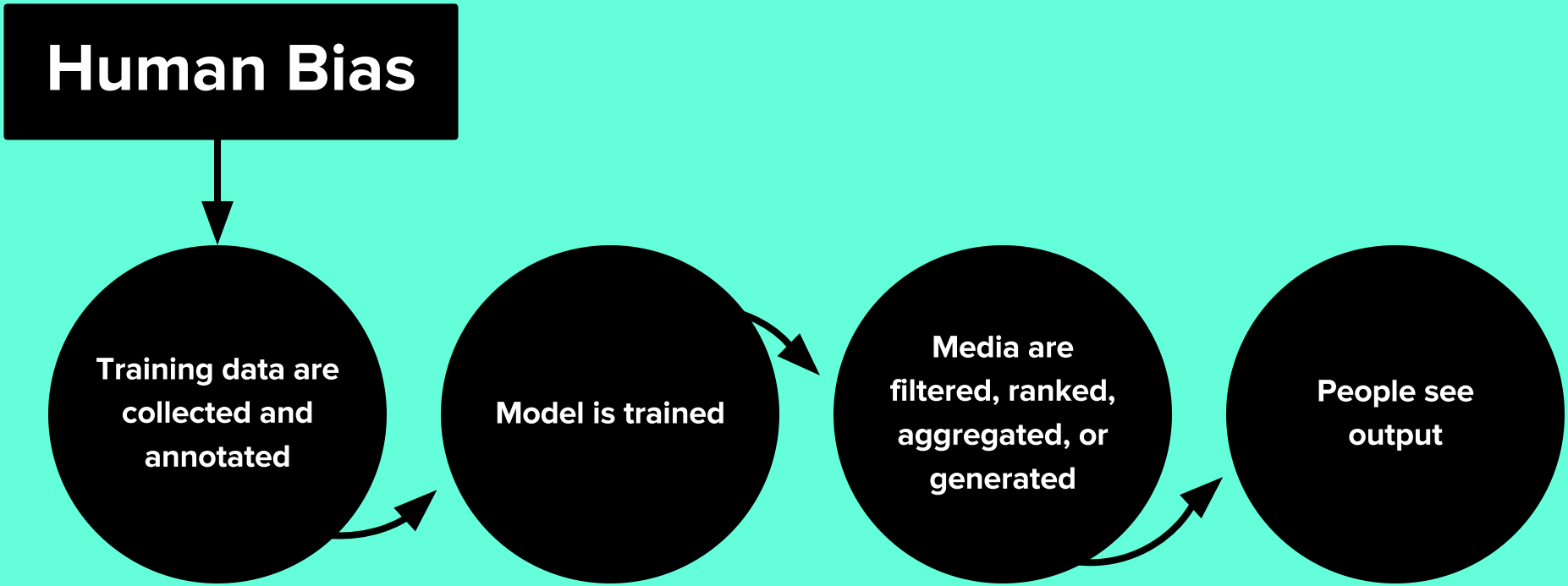
Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough

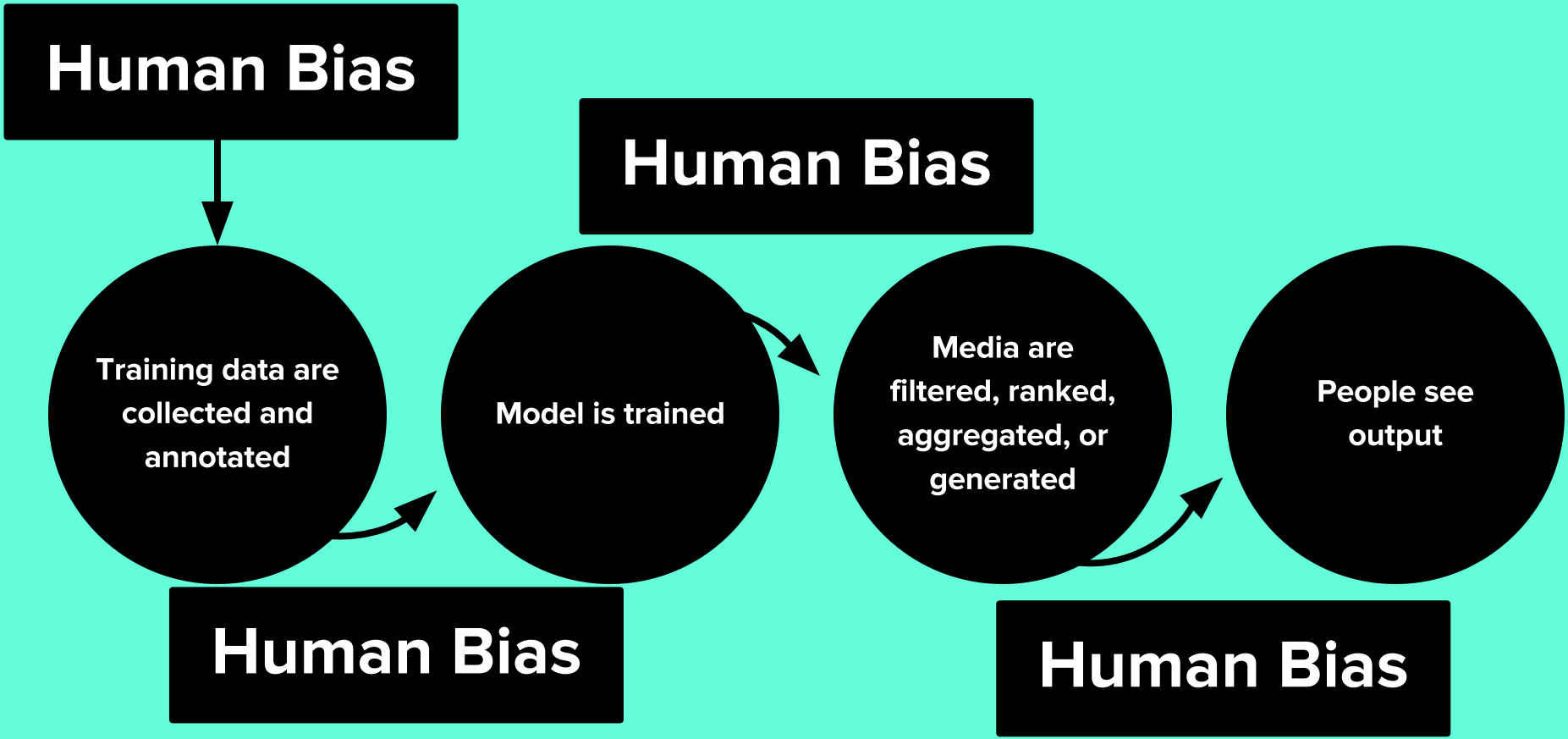
Out-group homogeneity bias: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

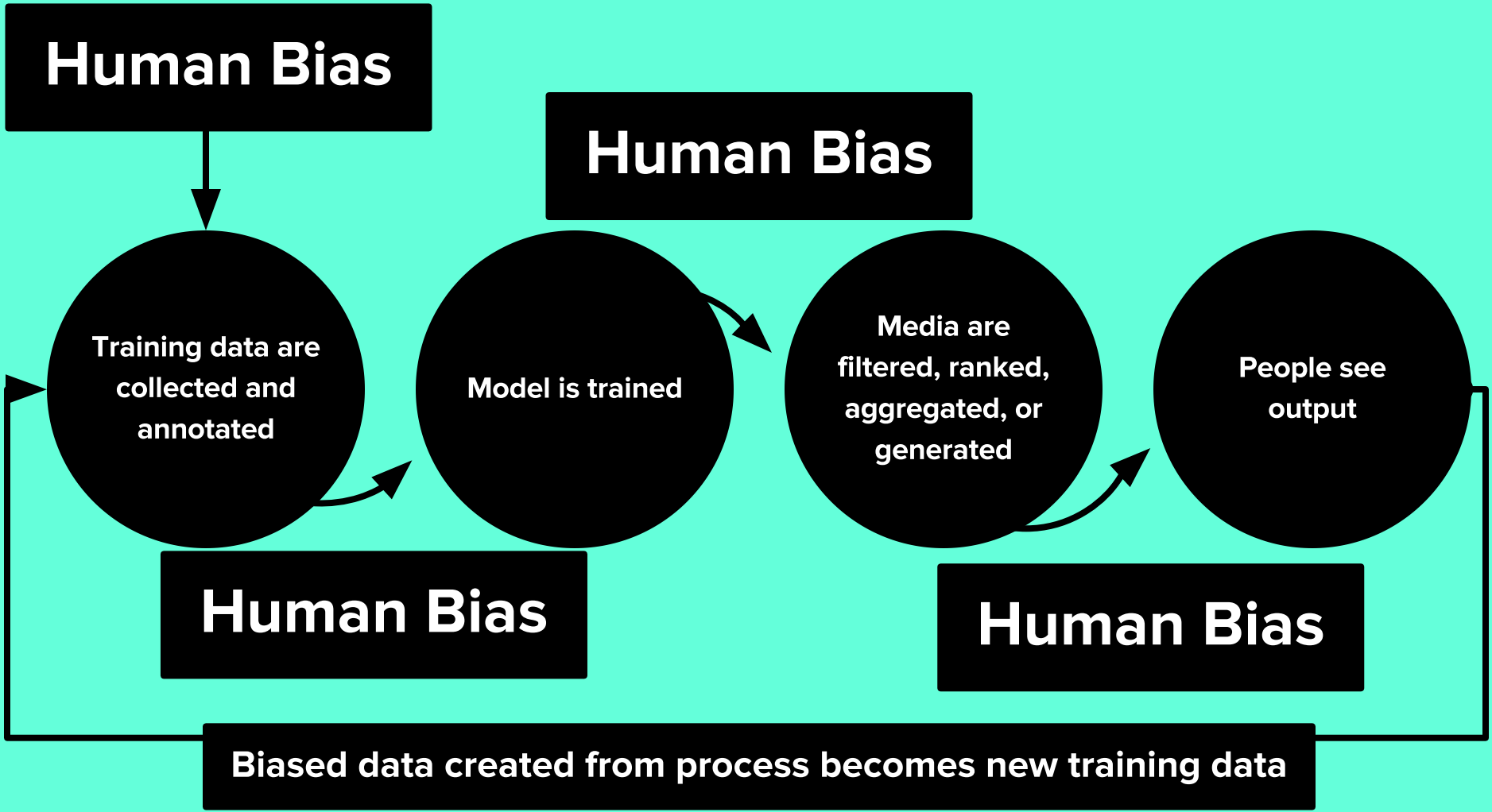
Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

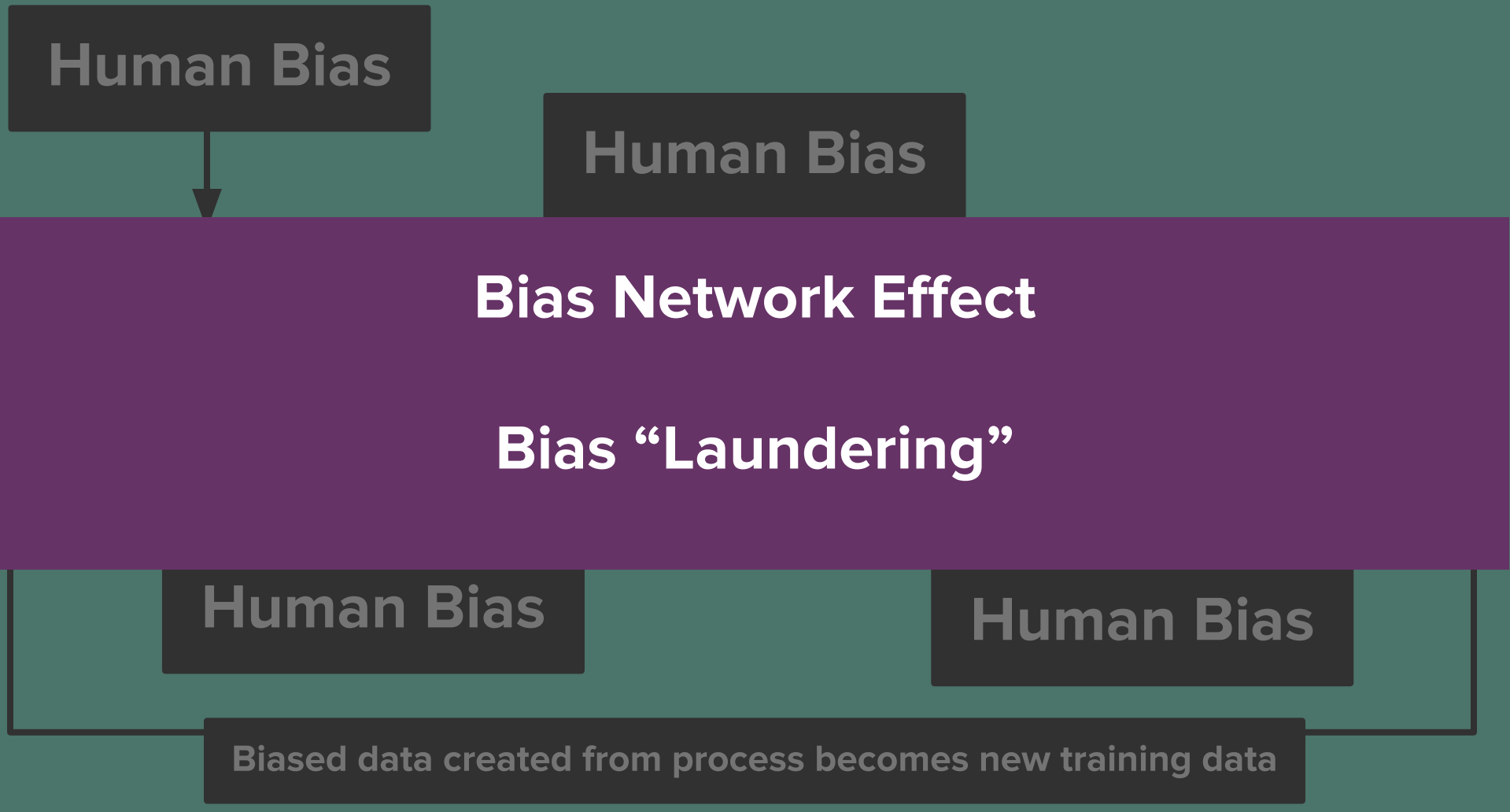
Automation bias: Propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct (Cummings, 2004)


Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output


Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output


Human Bias
Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output
Human Bias
Human Bias


Human Bias
Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output
Human Bias
Human Bias
Biased data created from process becomes new training data


Human Bias
Human Bias
Bias Network Effect
Bias “Laundering”
Human Bias
Human Bias
Biased data created from process becomes new training data

Human data perpetuates human biases.

As ML learns from human data, the result is a bias network effect.

“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian

CREDIT

The Guardian view on machine learning: people must decide

“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian

CREDIT

The Guardian view on machine learning: people must decide

01

Information technology can amplify misinformation

Trending news, Ranked results, Autocomplete

SOURCES

The Guardian

International Business Times

02

Legal software can propagate discrimination

Predictive policing, Risk assessment, Assessing criminality

SOURCES

Physiognomy's New Clothes

MIT Technology Review

ProPublica

The Intercept

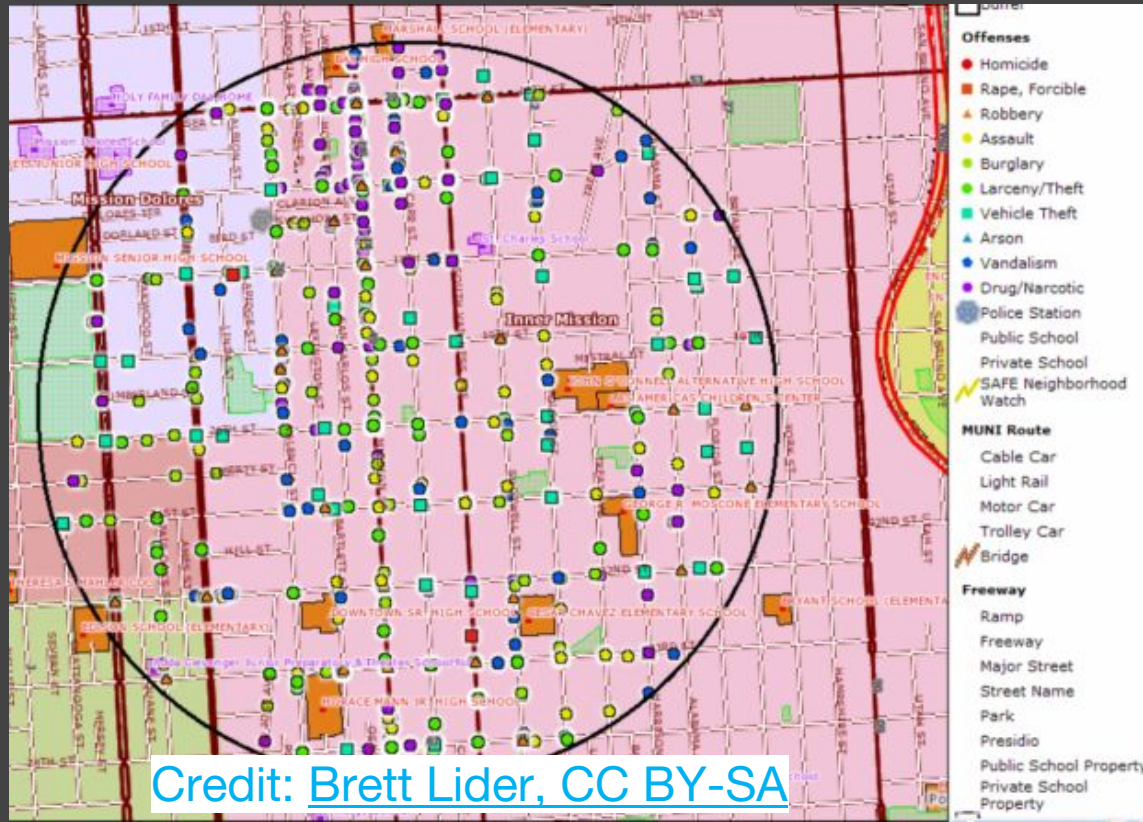
Predictive Policing

Identifies fine-grained areas of potential criminal activity

Used to help law enforcement decide where to deploy

Focus on gun, domestic violence

Mixed results



Credit: Brett Lider, CC BY-SA

Predictive Sentencing

Northpointe: Risk in criminal sentencing (ProPublica, 2016)

The likelihood of each committing a future crime is predicted.

Borden — who is black — was rated a high risk. Prater — a more seasoned criminal, who is white — was rated a low risk.

2 years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Predictive Criminality

An Israeli startup, Faception, who has not published any details about their methods, sources of training data, or quantitative results:

> *“Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial image.”*

Offering specialized engines for recognizing “High IQ”, “White-Collar Offender”, “Pedophile”, and “Terrorist” from a face image.

Main clients are in homeland security and public safety.

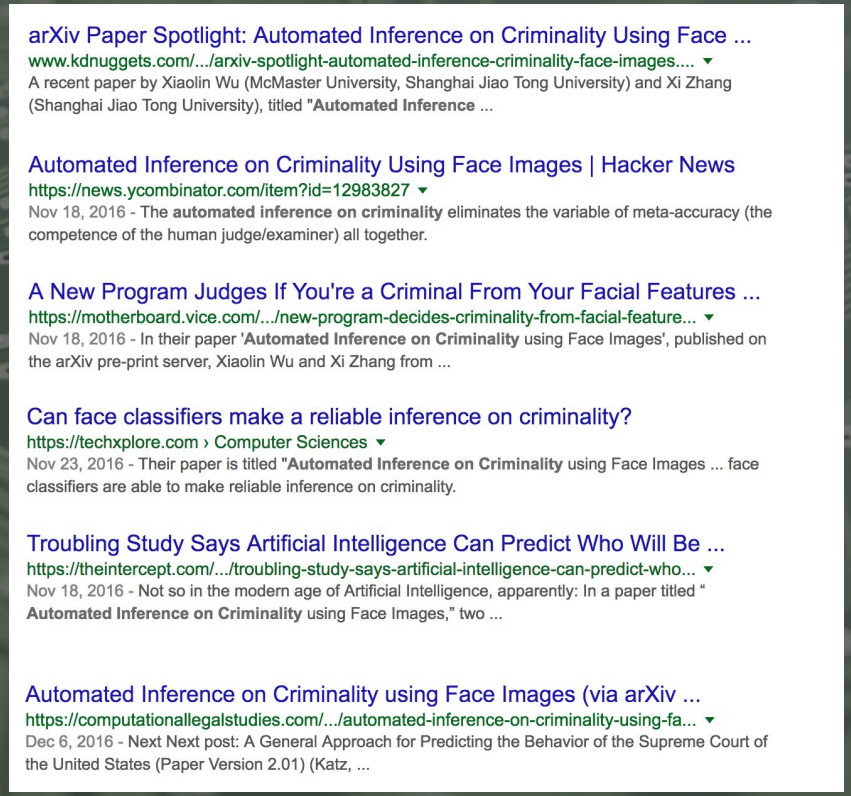
Predictive Criminality

"Automated Inference on Criminality using Face Images" Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces -- "wanted suspect" pictures from specific areas, rest from crawling web.

"[…] angle θ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ..."

Predictive Criminality - The Media Blitz



arXiv Paper Spotlight: Automated Inference on Criminality Using Face ...
www.kdnuggets.com/.../arxiv-spotlight-automated-inference-criminality-face-images....
A recent paper by Xiaolin Wu (McMaster University, Shanghai Jiao Tong University) and Xi Zhang (Shanghai Jiao Tong University), titled "**Automated Inference** ...

Automated Inference on Criminality Using Face Images | Hacker News
https://news.ycombinator.com/item?id=12983827
Nov 18, 2016 - The **automated inference on criminality** eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.

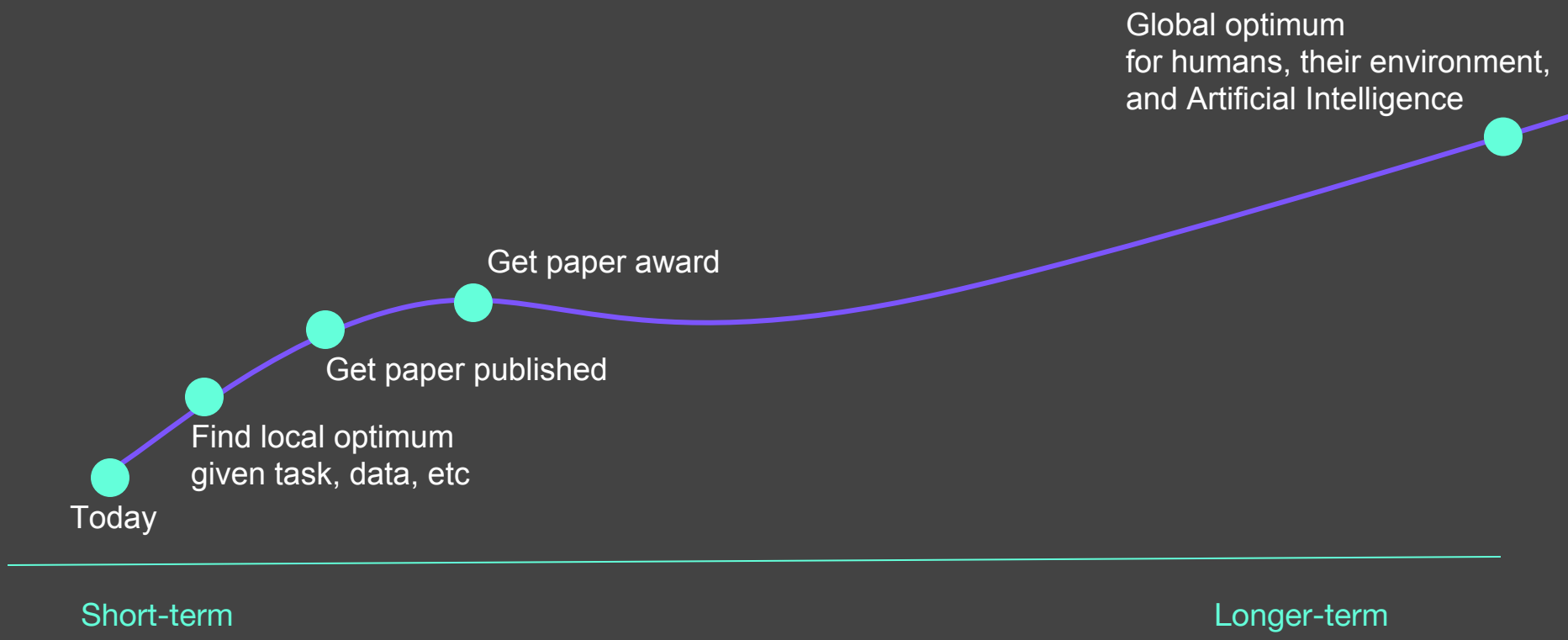
A New Program Judges If You're a Criminal From Your Facial Features ...
https://motherboard.vice.com/.../new-program-decides-criminality-from-facial-feature...
Nov 18, 2016 - In their paper '**Automated Inference on Criminality** using Face Images', published on the arXiv pre-print server, Xiaolin Wu and Xi Zhang from ...

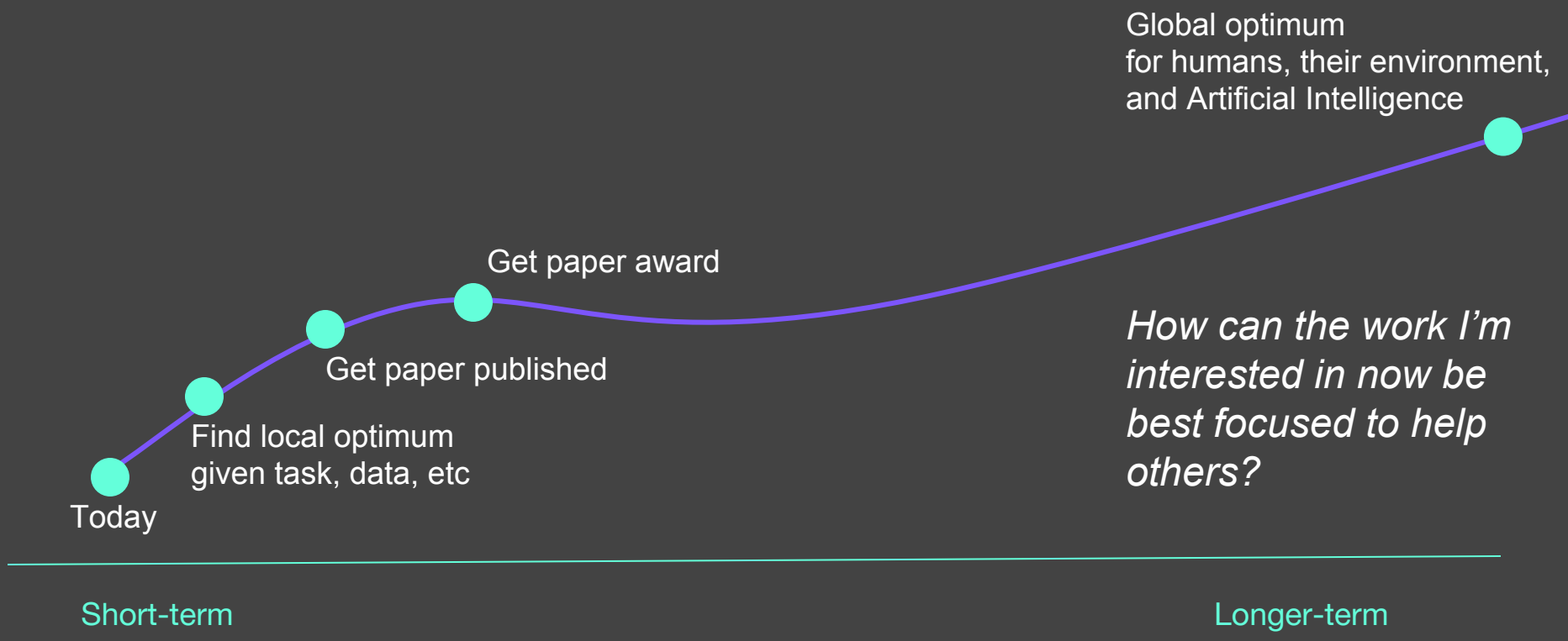
Can face classifiers make a reliable inference on criminality?
https://techxplore.com › Computer Sciences
Nov 23, 2016 - Their paper is titled "**Automated Inference on Criminality** using Face Images ... face classifiers are able to make reliable inference on criminality.

Troubling Study Says Artificial Intelligence Can Predict Who Will Be ...
https://theintercept.com/.../troubling-study-says-artificial-intelligence-can-predict-who...
Nov 18, 2016 - Not so in the modern age of Artificial Intelligence, apparently: In a paper titled " **Automated Inference on Criminality** using Face Images," two ...

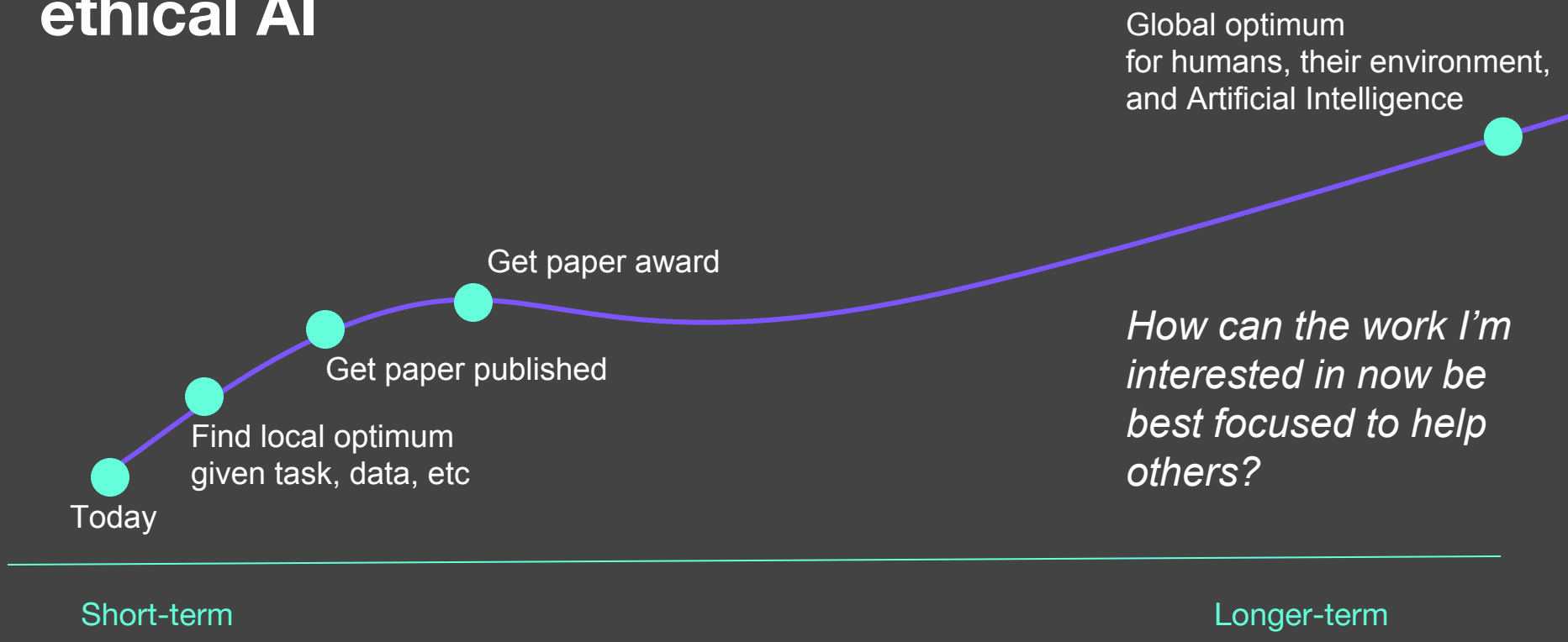
Automated Inference on Criminality using Face Images (via arXiv ...
https://computationallegalstudies.com/.../automated-inference-on-criminality-using-fa...
Dec 6, 2016 - Next Next post: A General Approach for Predicting the Behavior of the Supreme Court of the United States (Paper Version 2.01) (Katz, ...

But it’s up to **us** to influence how AI evolves.


Global optimum
for humans, their environment,
and Artificial Intelligence
Get paper award
Get paper published
Find local optimum
given task, data, etc
Today
Short-term
Longer-term


Global optimum
for humans, their environment,
and Artificial Intelligence
Get paper award
Get paper published
Find local optimum
given task, data, etc
Today
How can the work I'm interested in now be best focused to help others?
Short-term
Longer-term

Begin tracing out paths for the evolution of ethical AI



Rest of Talk

1. Modeling world knowledge (and biases!) with latent variables
2. Focus on best performance across **groups** of people

- Working with experts and those affected to better understand what's needed
- Contextualizing work for public

Modeling World Knowledge with Latent Variables:
A case study in vision-to-language

Modeling World Knowledge with Latent Variables:
A case study in vision-to-language

Those affected: People who are blind

Modeling World Knowledge with Latent Variables:

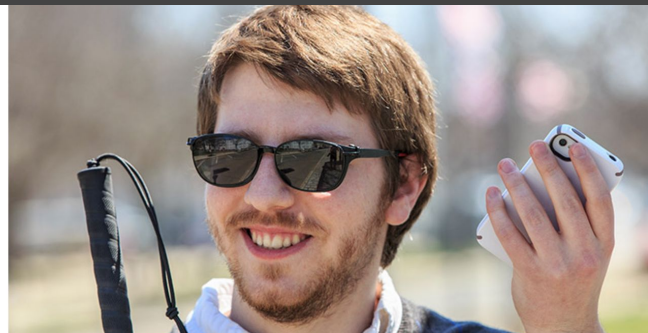
A case study in vision-to-language

Those affected: People who are blind

Seeing AI: Microsoft research project that brings together the power of the cloud and AI to deliver an intelligent app, designed to help you navigate your day.

Turns the visual world into an audible experience

With this intelligent camera app, just hold up your phone and hear information about the world around you



Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

Also called “Black Sheep” problem, “Giraffes problem” in vision/language

Photographer Bias: Natural tendency of photographers to place object of interest in the center

Bias: A systematic deviation from full ground truth.

Bias: A systematic deviation from full ground truth.

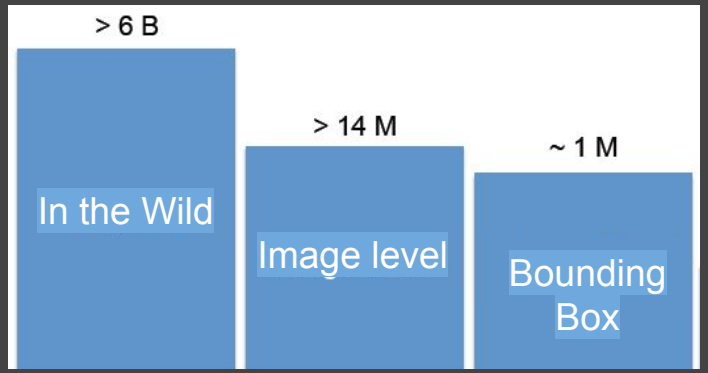
Can be used as **a signal.**

Data data everywhere …

Facebook

300 Million images uploaded everyday

flickr



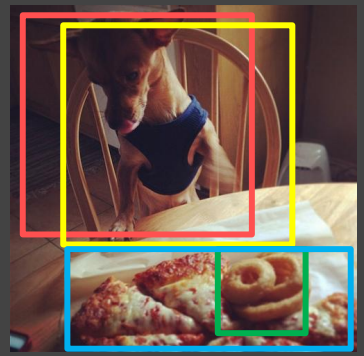
YouTube™

100 hours of video every minute

#dog #hungry

OMG Frodo is sitting eating pizza and donuts.

dog, chair, pizza, donut



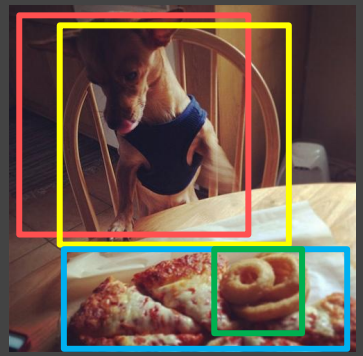
dog, chair, pizza, donut

Data data everywhere …

But not many labels to train

Exhaustively annotated data is expensive

dog, chair, pizza, donut



dog, chihuahua, brown, chair, table, wall, space heater, pizza, greasy, donut 1, donut 2, pizza slice 1, pizza slice 2…

“In the Wild” Labeled Images: Why?

- “Freely” available: Image tags, descriptions on social media
- Fast way to gather data beyond typical categories
- Annotations for an image are on a “per-image” basis [mostly]

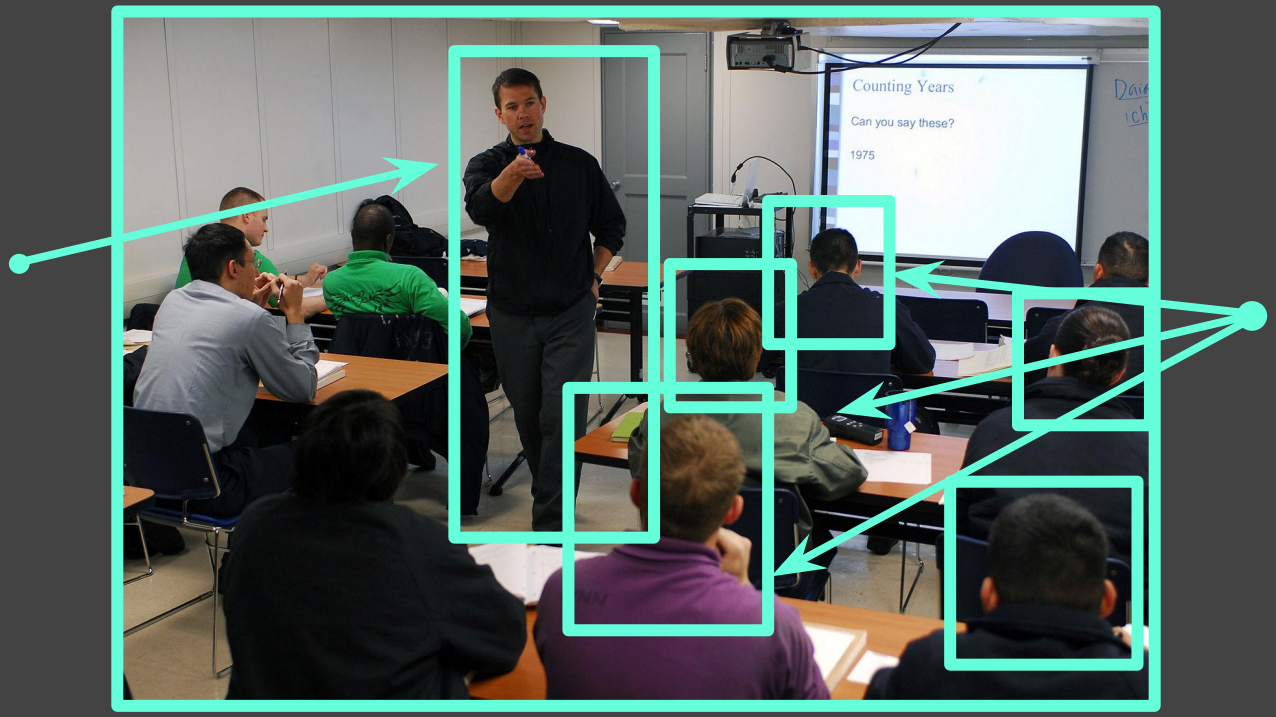
#dog #food #hungry

A hungry dog looks at the food on the table

Challenges with WILD Labeled Images

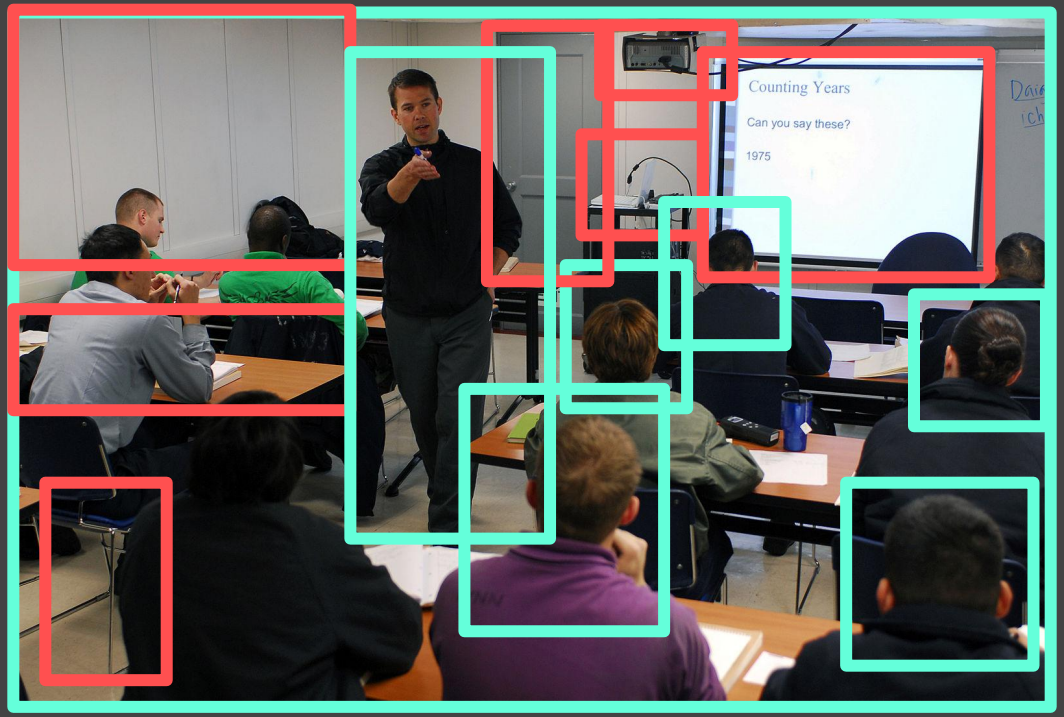
Teacher in a classroom talking to students

Challenges with WILD Labeled Images



Teacher in a classroom talking to students

Challenges with WILD Labeled Images



Teacher in a classroom talking to students

Challenges with WILD Labeled Images

Teacher in a classroom talking to students

Humans subjectively decide ***what to report*** and ***what not to report*** in an image

Challenges with WILD Labeled Images

Teacher

Humans subjectively decide ***what to report*** and ***what not to report*** in an image

Teacher in a classroom talking to students

Challenges with WILD Labeled Images

Teacher
Male
White
Early 30s
Wearing Pants

Humans subjectively decide ***what to report*** and ***what not to report*** in an image

Teacher in a classroom talking to students

Using Labeled Images In the Wild

How do we train visually correct classifiers from wild data?

Input Wild Data

#dog #food #hungry

Expected Output

dog, brown, chihuahua, chair, pizza, donut

Problem setup

- **Input:** Image, human-biased labels
- **Goal:** Learn visually correct classifiers
- **Challenge:** Do not have access to ground truth; have access to what humans have reported

Input

Teacher in a classroom talking to students

Goal

teacher, man, standing, early 30s, white, classroom, high school, students, projector, door, wall, podium, learning

Human-Biased Labels

- Highly dependent on the input image
  - Example: *bicycle*

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Human Biased Labels

- Highly dependent on the input image
- The outcome of a complex systematic process [human judgment]
  - Humans are fairly systematic in such labeling

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Human-Biased Labels

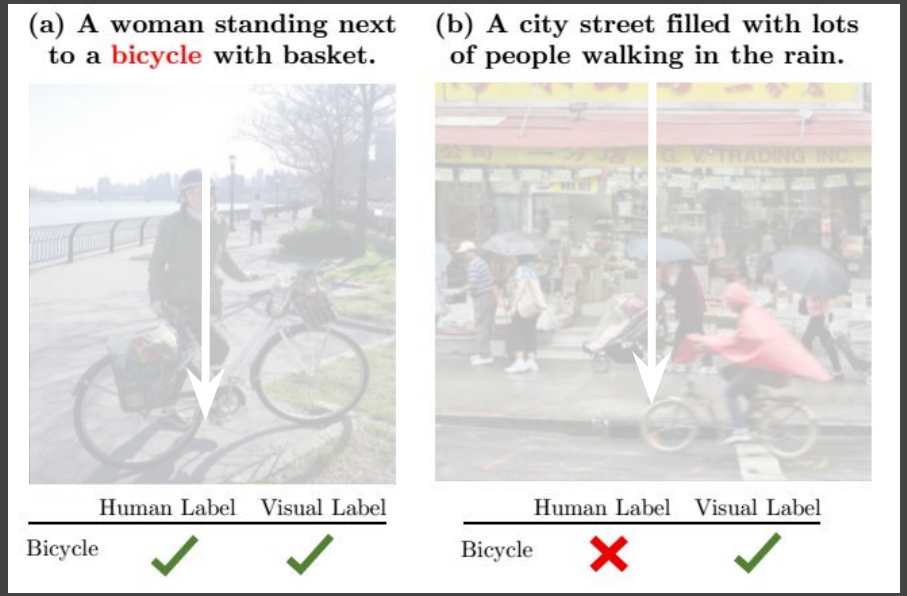
- Highly dependent on the input image
- The outcome of a complex systematic process [human judgment]
  - Humans are fairly systematic in such labeling
  - Humans refer to object properties when it helps distinguishability, conversation etc.

    [Gregory 1966], [Rosch 1973], [Sedivy et al., 2003], [Koolen et al., 2011]

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Related Work: Modeling label noise

- Assumes the noise is conditionally independent of the input image
  [Mnih and Hinton 2012], [Natarajan et al., 2013], [Reed et al., 2014], [Sukhbaatar et al., 2015], [Izadinia et al., 2015]



Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Related Work: Modeling label noise

- Assumes the noise is conditionally independent of the input image
  [Mnih and Hinton 2012], [Natarajan et al., 2013], [Reed et al., 2014], [Sukhbaatar et al., 2015], [Izadinia et al., 2015]
- Assumes that estimating noise requires access to exhaustively labeled data
  [Xiao et al., 2015]

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Notation

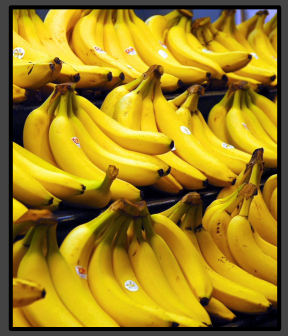
$w \in$ {banana, yellow}

Input Image

	Banana	Yellow	Label
Visually correct ground truth (**Unknown**)	✓	✓	z^w
Available Ground truth (**human-biased**)	✓	✗	y^w

Simple Image Classification

$w \in$ {banana, yellow}



Input Image

CNN

Classifier

Linear

Sigmoid

For each *w*

Output

Banana

Yellow

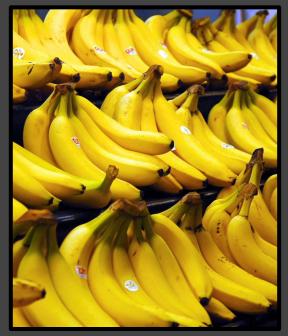
Ground Truth

y^w

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Simple Image Classification

$w \in$ {banana, yellow}



Input Image

CNN

Classifier

Linear

Sigmoid

For each *w*

Output

Banana

Yellow

Ground Truth

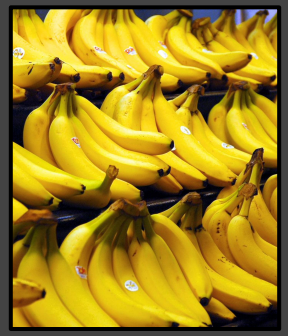
y^w

Human-biased label $y^w \in \{0, 1\}$ (Gold Standard)

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Simple Image Classification

$w \in$ {banana, yellow}



Input Image

CNN

Classifier

Linear

Sigmoid

For each w

Output

Banana

Yellow

Ground Truth

y^w

Human-biased label $y^w \in \{0, 1\}$ (Gold Standard)

Prediction $h^w(y^w|I)$

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Factoring in label bias: Idea

- A human-biased prediction h can be factored into two terms

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Factoring in label bias: Idea

- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the object visually present?*

$w \in$ {banana, yellow}

✓ ✓

visually correct ground truth (unknown): z

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Factoring in label bias: Idea

- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the object visually present?*
  - Relevance r – *Is the object relevant for a human?*

$w \in$ {banana, yellow}

✓ ✗

Use z to predict human-biased label y

	Label	Prediction
Visually correct ground truth (**Unknown**)	z	v
Available ground truth (**human-centric**)	y	h

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Factoring in label bias: Idea

- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the object visually present?*
  - Relevance r – *Is the object relevant for a human?*

$$h = f(r, v)$$

	Label	Prediction
Visually correct ground truth (**Unknown**)	z	v
Available ground truth (**human-centric**)	y	h

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Factoring in label bias: Idea

- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the object visually present?*
  - Relevance r – *Is the object relevant for a human?*

Given visual presence, is concept relevant?

Is concept present?

$$h(y|I) = \sum_{j \in \{0,1\}} r(y|z=j,I)v(z=j|I)$$

	Label	Prediction
Visually correct ground truth (**Unknown**)	z	v
Available ground truth (**human-centric**)	y	h

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Factoring in label bias: Idea

- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the object visually present?*
  - Relevance r – *Is the object relevant for a human?*

Given visual presence, is concept relevant? Is concept present?

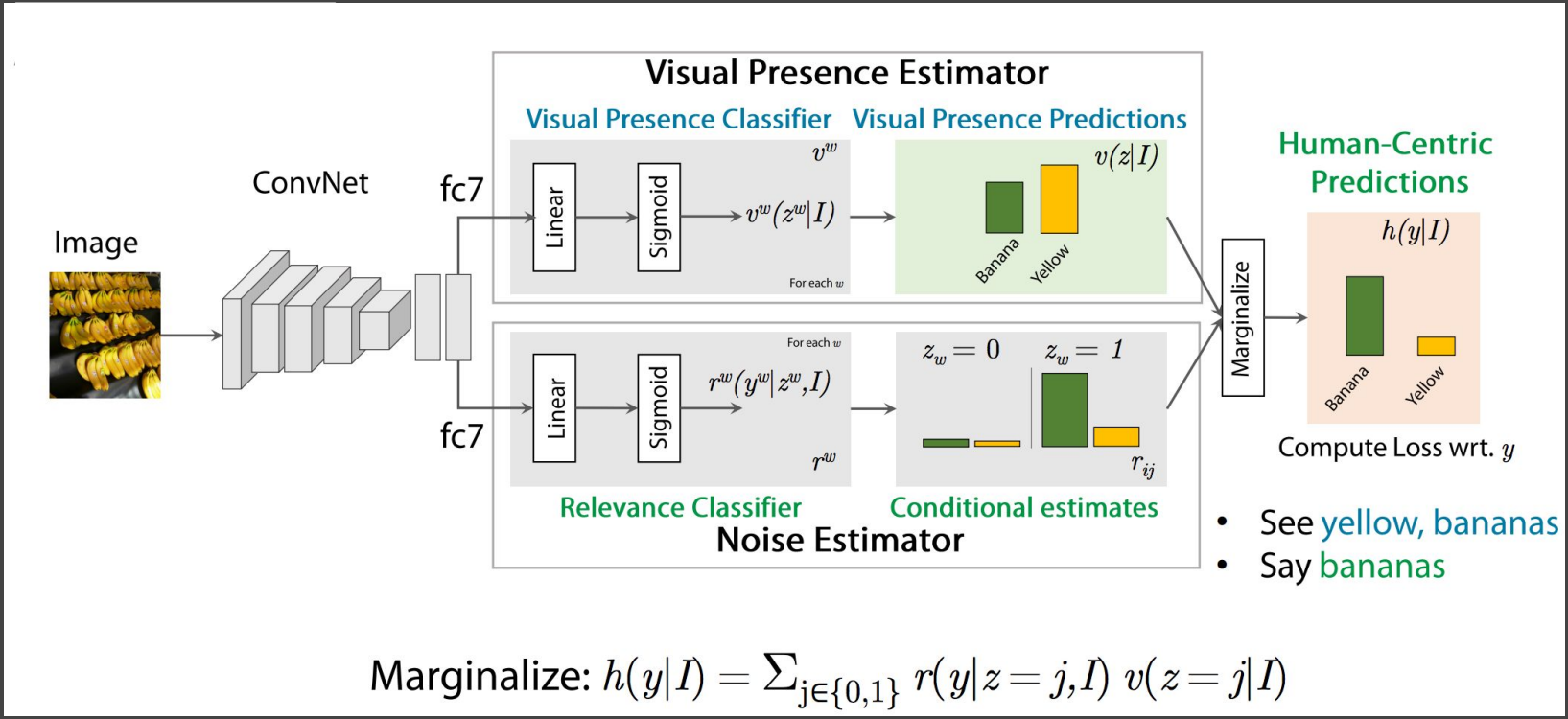
$$h(y|I) = \sum_{j \in \{0,1\}} r(y|z=j,I)v(z=j|I)$$

- Allows classifier to not get penalized for correct predictions

	Label	Prediction
Visually correct ground truth (**Unknown**)	z	v
Available ground truth (**human-centric**)	y	h

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

End-to-End Approach



Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

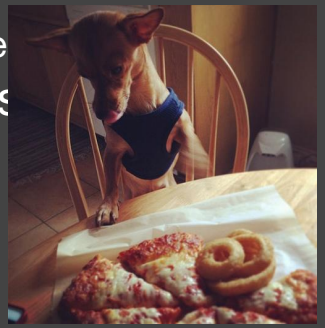
Results

- Evaluate both *v* and *h* predictions
- Microsoft COCO dataset
  - Human-biased labels *y* from Captions [1000 categories]
  - Visually correct labels *z* from Detection Bounding boxes [73 categories]
  - #images: 80k train, 20k test

- YFCC100M
  - Yahoo Flickr images with tags [1000 categorie
  - Random subset #images: 75k train, 15k tes

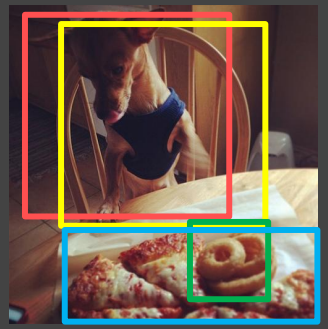
	Label	Prediction
Visually correct ground truth (**Unknown**)	z	v
Available ground truth (**human-centric**)	y	h

y labels



A hungry dog looks at the food on the table.

z labels: Only for testing



dog, chair, pizza, donut

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Evaluating using y (human-biased) and z (annotated)

Evaluation using observed caption concepts

Method	Prob	Mean Average Precision NN	VB	JJ	DT	PRP	IN	Others	All
COCO Dataset. 1000 Visual concepts from Captions									
MILVC	-	41.6	20.7	23.9	33.4	20.4	22.5	16.3	34.0
MILVC + Multiple fc8	-	41.1	20.9	23.7	33.6	21.1	22.8	16.8	33.8
MILVC + Latent	v	42.9	21.7	24.9	33.1	19.6	23.0	16.2	35.1
MILVC + Latent	h	**44.3**	**22.3**	**25.8**	**34.4**	**21.8**	**23.6**	**17.3**	**36.3**
Classif.	-	34.9	18.1	20.5	32.8	19.2	21.8	16.3	29.0
Classif. + Multiple fc8	-	34.2	17.7	19.9	32.6	19.0	21.5	15.9	28.4
Classif. + Latent	v	37.7	19.6	22.0	32.6	20.2	22.0	16.3	31.2
Classif. + Latent	h	**38.7**	**20.1**	**22.6**	**33.8**	**21.2**	**23.0**	**17.5**	**32.0**
YFCC100M: Flickr images with tags (90k images, 1k tags)									
MILVC	-	5.7	9.2	5.2	-	3.8	8.8	6.1	5.7
MILVC + Multiple fc8	-	4.6	6.2	3.8	-	2.7	7.3	3.1	4.5
MILVC + Latent	v	9.8	15.1	8.9	-	8.3	12.4	12.4	9.8
MILVC + Latent	h	**11.2**	**15.4**	**9.9**	-	**8.2**	**16.3**	**12.5**	**11.2**

All methods use VGG16. Trained using binary cross-entropy loss.
MILVC: Fang et al., 2015; Classif.: Simple classification baseline;
Multiple-fc8: Same # parameters as our model.

Evaluation using annotated concepts

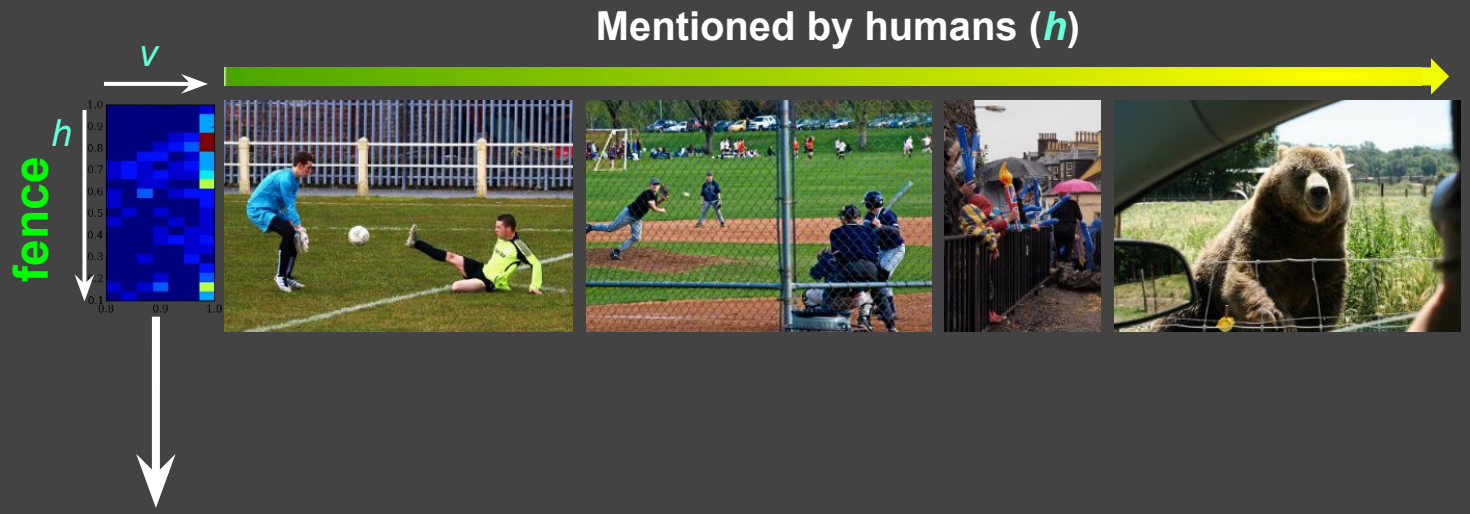
COCO Dataset. 73 annotated concepts from Bounding Boxes.

	MILVC	v	h	Using ground truth
mAP	63.7	**66.8**	66.5	76.3

NN: Nouns
VB: Verbs
JJ: Adjectives
DT: Determiner
PRP: Pronouns
IN: Prepositions

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Qualitative Results



Threshold at $v \geq 0.85$

Shows how much *v* and *h* are decoupled

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Qualitative Results

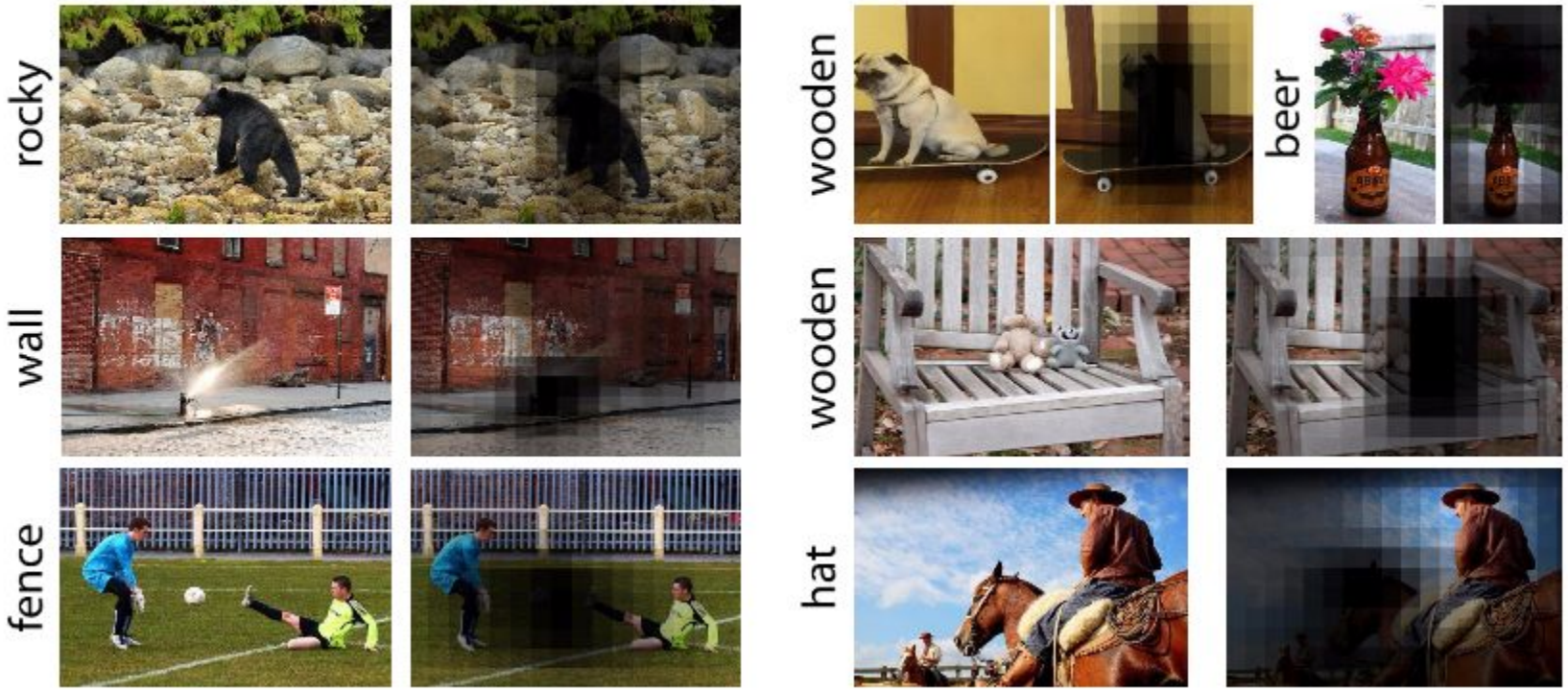
Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Corrected Error Modes

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

When to mention it?

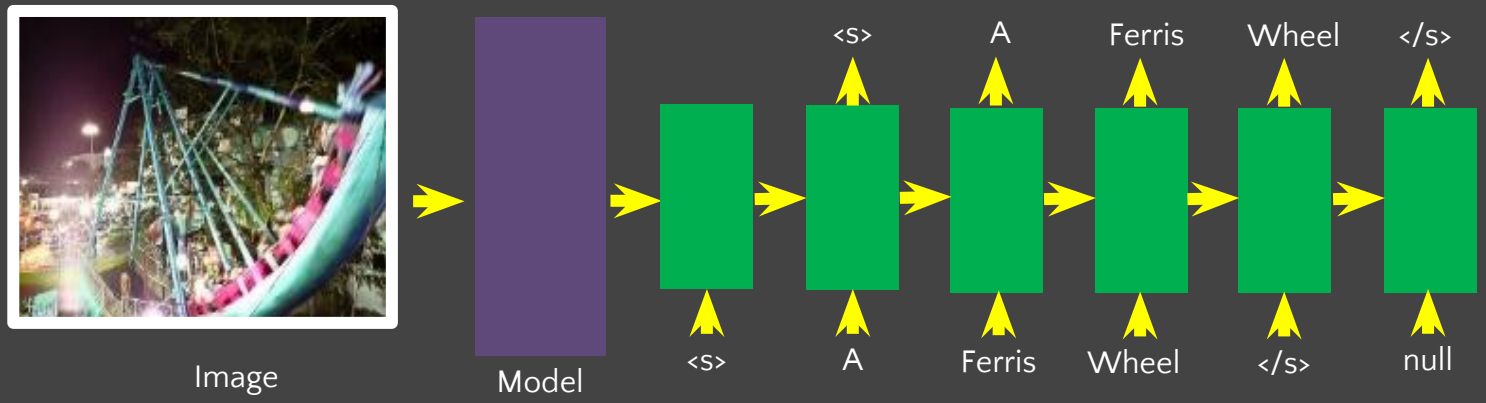
When would you mention something not worth mentioning?



Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Improvement in Downstream Applications: Image Captioning

	Prob	BLEU-4	ROUGE	CIDEr
MILVC	-	27.7	51.8	89.7
MILVC + Latent	h	**29.2**	**52.4**	**92.8**



Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Rest of Talk

1. Modeling world knowledge (and biases!) with latent variables
2. Focus on best performance across **groups** of people
   - Working with experts and those affected to better understand what's needed
   - Contextualizing work for public

Fair is Fair: For all groups of people

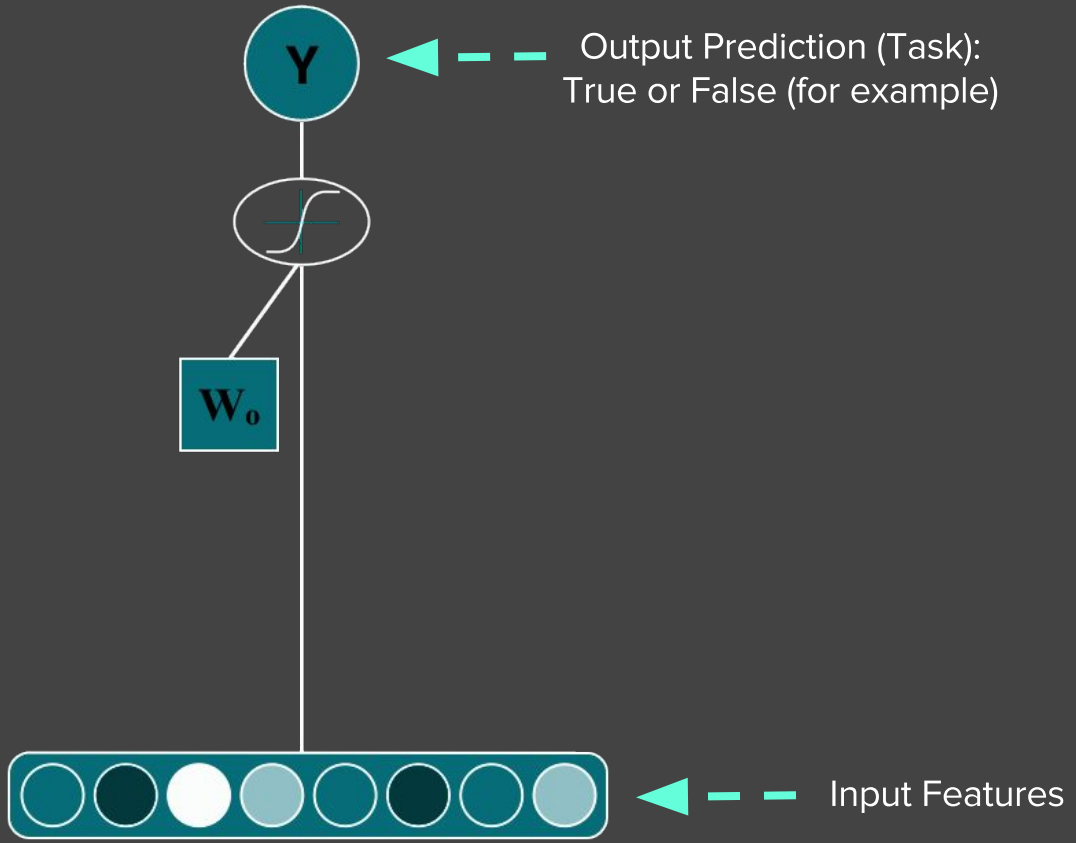
Those affected: People with neuroatypicality, clinicians

Motivation from “The Karate Kid”

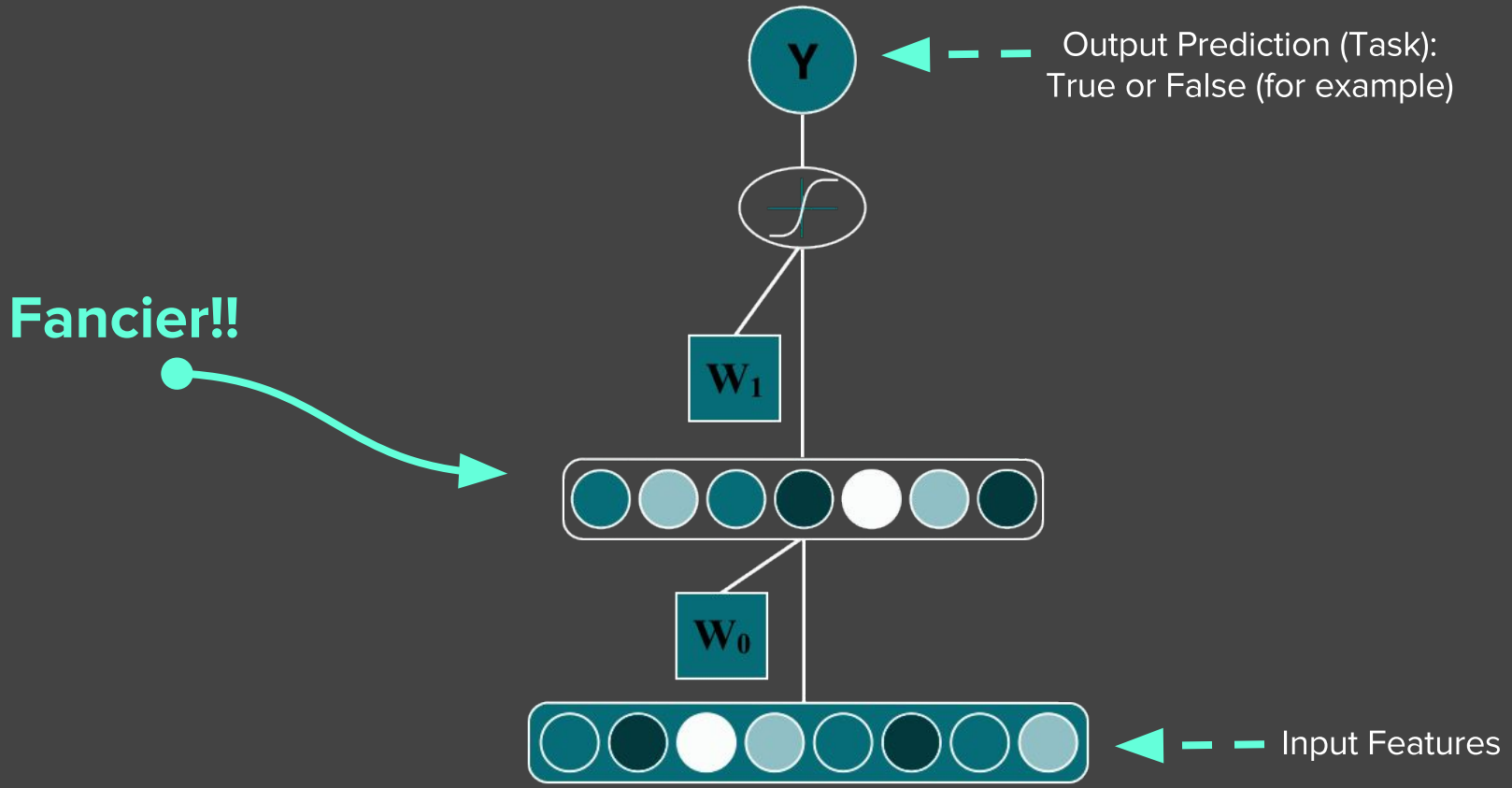
Single-task Learners
(STL)

Multitask Learner
(MTL)

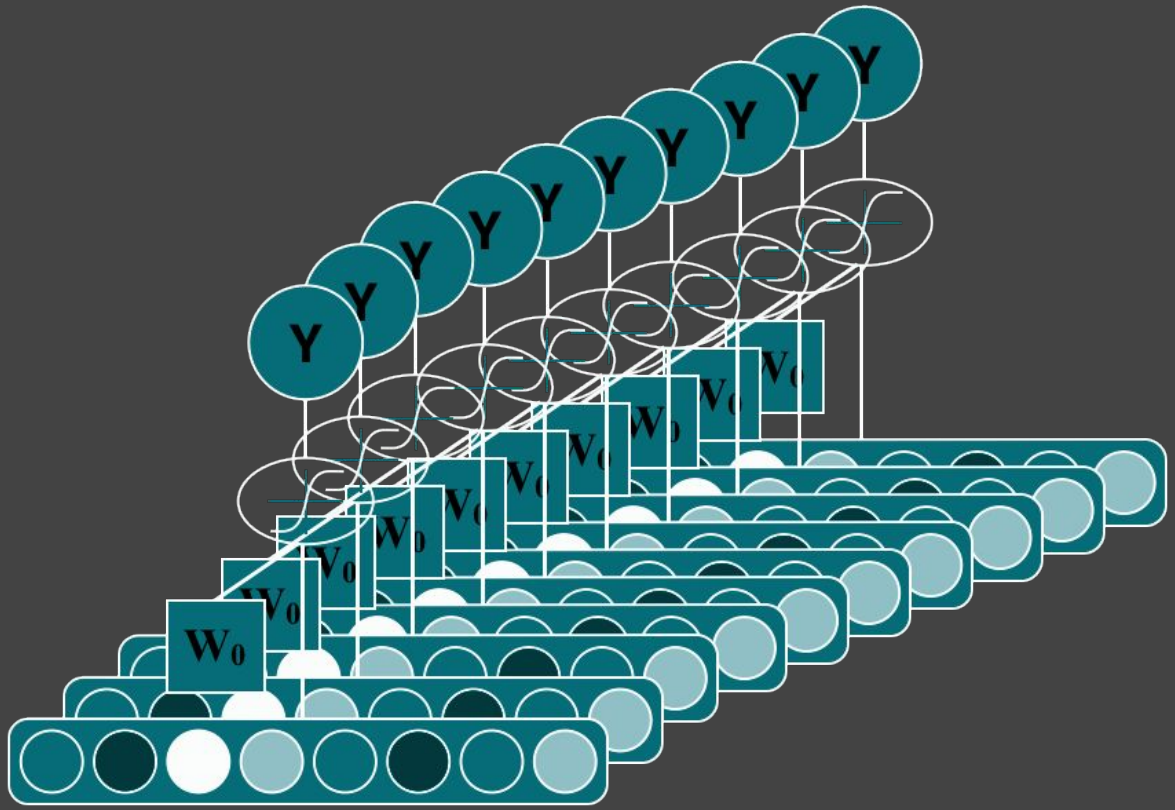
Single-Task: Logistic Regression



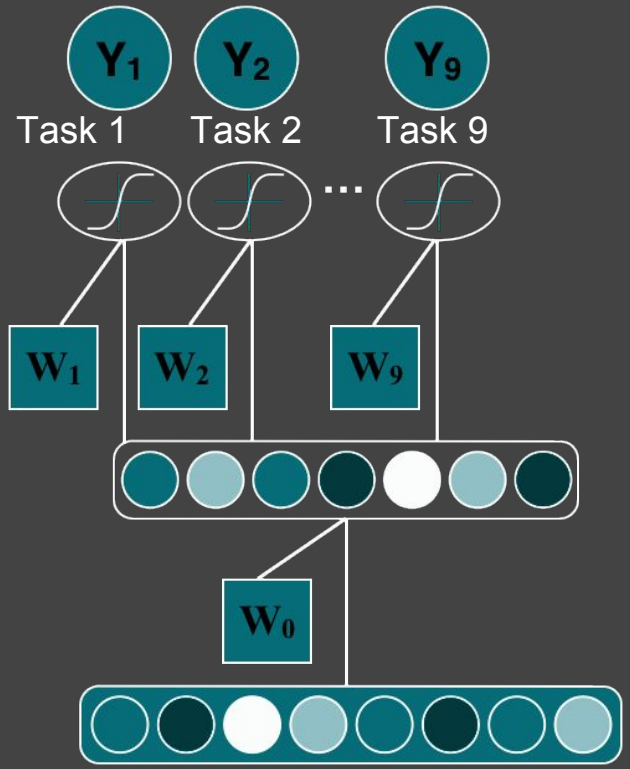
Single-Task: Deep Learning



Multiple Tasks with Basic Logistic Regression

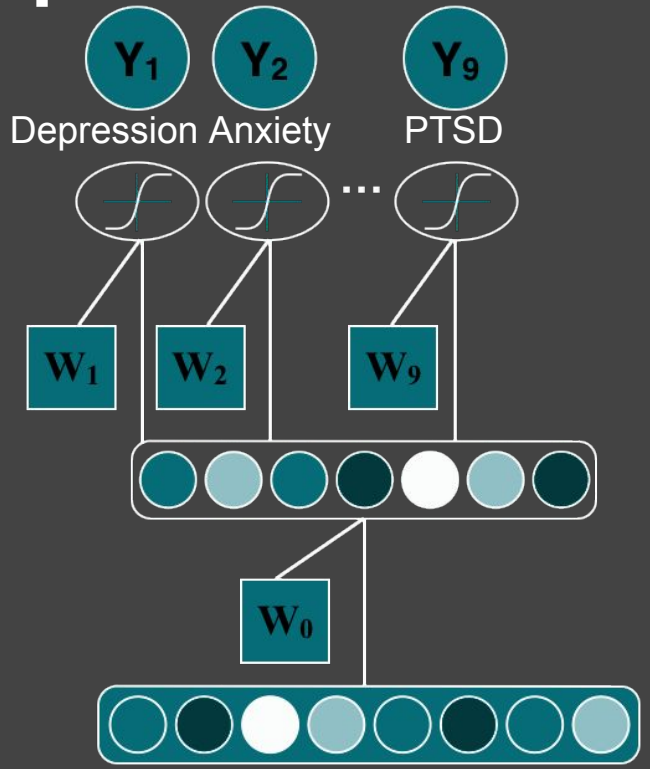


Multiple Tasks + Deep Learning: Multi-task Learning



Benton, Mitchell, Hovy. 2017. "Multi-task learning for Mental Health Conditions with Limited Social Media Data"

Multiple Tasks + Deep Learning: Multi-task Learning Example

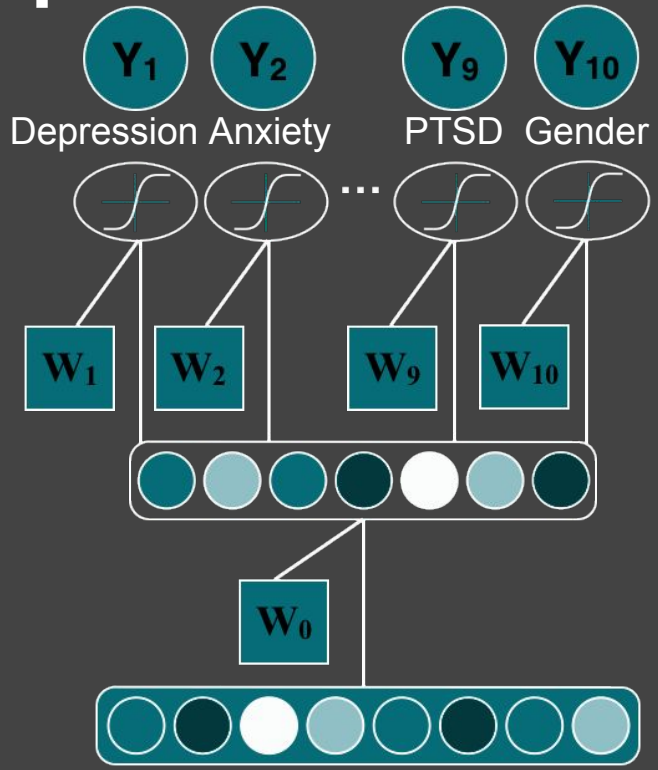


Task	N
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	**9611**

<5% positive examples

Benton, Mitchell, Hovy. 2017. "Multi-task learning for Mental Health Conditions with Limited Social Media Data"

Multiple Tasks + Deep Learning: Multi-task Learning Example

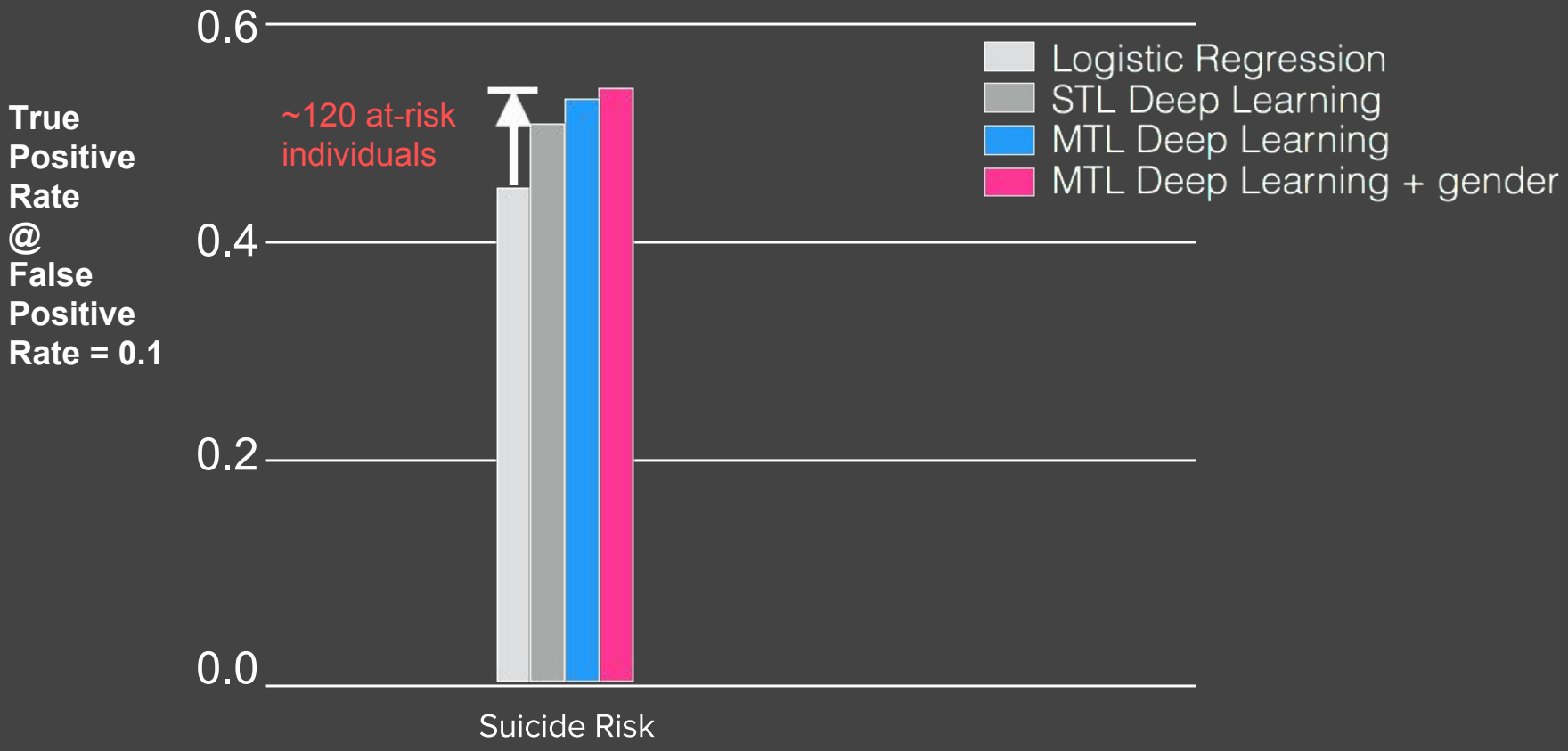


Task	N
Gender	1101
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	**9611**

<5% positive examples

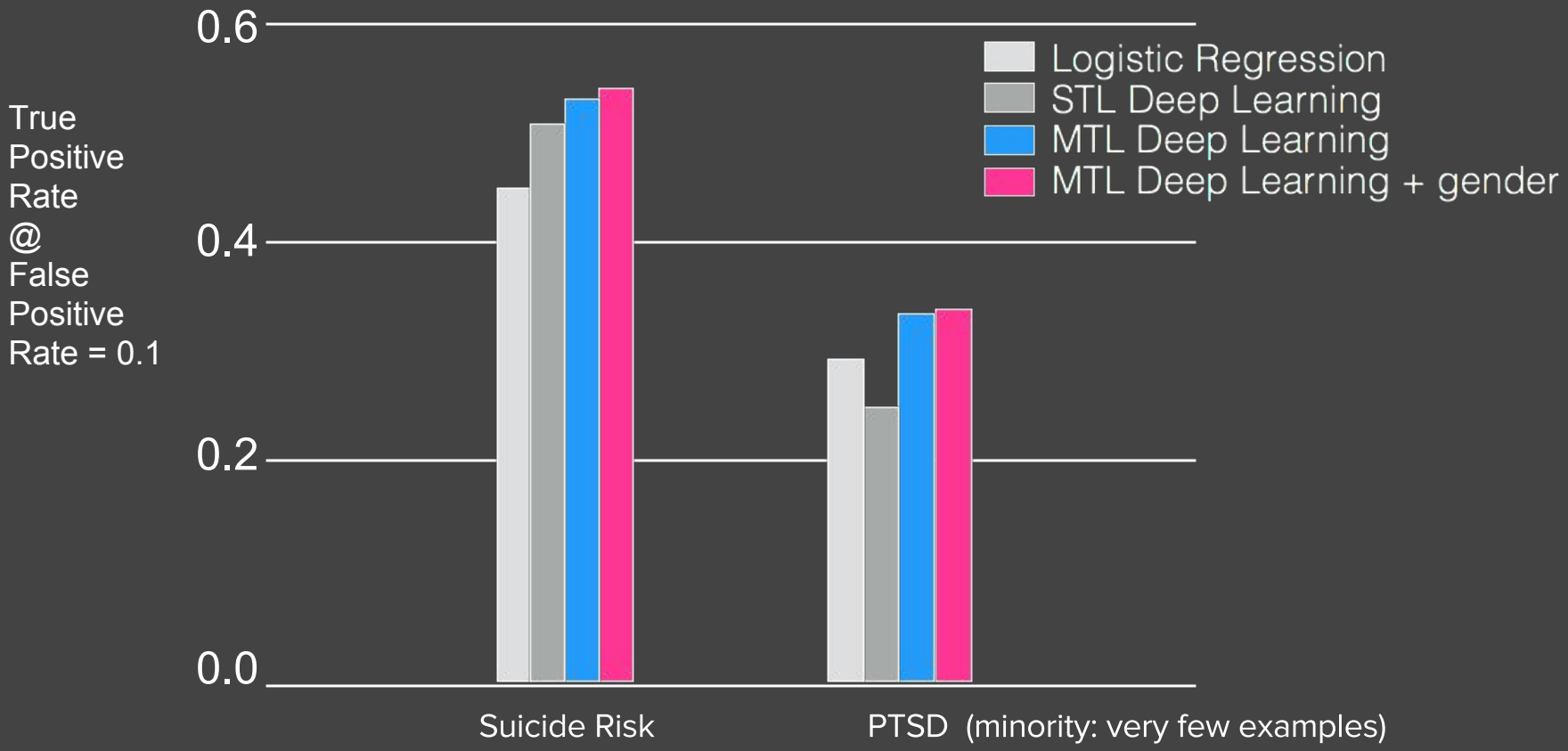
Benton, Mitchell, Hovy. 2017. "Multi-task learning for Mental Health Conditions with Limited Social Media Data"

Improved Performance across Subgroups



Benton, Mitchell, Hovy. 2017. "Multi-task learning for Mental Health Conditions with Limited Social Media Data"

Improved Performance across Subgroups



Benton, Mitchell, Hovy. 2017. "Multi-task learning for Mental Health Conditions with Limited Social Media Data"

Reading for the masses….

Multi-Task Learning for Mental Health using Social Media Text

Adrian Benton
Johns Hopkins University
adrian@cs.jhu.edu

Margaret Mitchell
Microsoft Research*
mmitchellai@google.com

Dirk Hovy
University of Copenhagen
mail@dirkhovy.com

Contextualizing and considering ethical dimensions

2 Disclaimer

As with any author-attribute detection, there is the danger of abusing the model to single out people (*overgeneralization*, see Hovy and Spruit (2016)). We are aware of this danger, and sought to minimize the risk. For this reason, we don't provide a selection of features or representative examples. The experiments in this paper were performed with a clinical application in mind, and use carefully matched (but anonymized) data, so the distribution is not representative of the population as a whole. The results of this paper should therefore *not* be interpreted as a means to assess mental health conditions in social media in general, but as a test for the applicability of MTL in a well-defined clinical setting.

Benton, Mitchell, Hovy. 2017. "Multi-task learning for Mental Health Conditions with Limited Social Media Data"

Reading for the masses….

Science

Me, Me, Me: People Who Overuse The First-Person Singular Are More Depressed

A new study links first-person singular pronouns to relationship problems and higher rates of depression.

By Rose Pastore *May 3, 2013*

Contextualizing and considering ethical dimensions

PHASE 01

Consider the problem

How will the model be affected when a blind spot is found in the training data?

PHASE 02

Ask experts for answers

What do the experts say is most useful or necessary? What do they think of the problem you're working on?

PHASE 03

Engage with Policy

Serve as consultant for your senator, congressperson; be part of legal and policy meetings relevant to your work.

PHASE 04

Design the human input

Is this an unambiguous task? How will you verify that crowdworkers are performing tasks “correctly”? How will you best leverage human biases?

PHASE 05

Engage diverse crowdworkers well

Speed and Agreement are bedrock measures of "click-workers", and their goal is different from yours. Without consideration of speed/pay tradeoffs, and crowdworker diversity, click-work turns into exponential groupthink that bakes cultural biases directly into training data.

PHASE 06

Train the models to account for bias

What does an outlier use case look like, and how does the model handle it? What implicit assumptions might be helpful to model?

PHASE 07

Interpret outcomes

Is the ML overgeneralizing? If a human were to perform this task, what would appropriate social behavior look like? What interpersonal cues might be relevant that are missing from the input or interface? E.g. body language, tone of voice.

PHASE 08

Publish with context

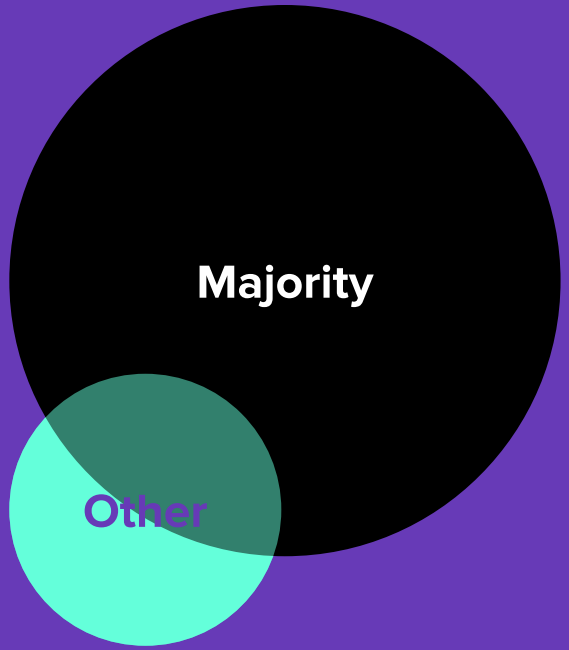
Are you sharing examples? Why or why not?
How should this technology be directly used?
What are some easy misconceptions that we can avoid?

PHASE 09

Publish for reproducibility

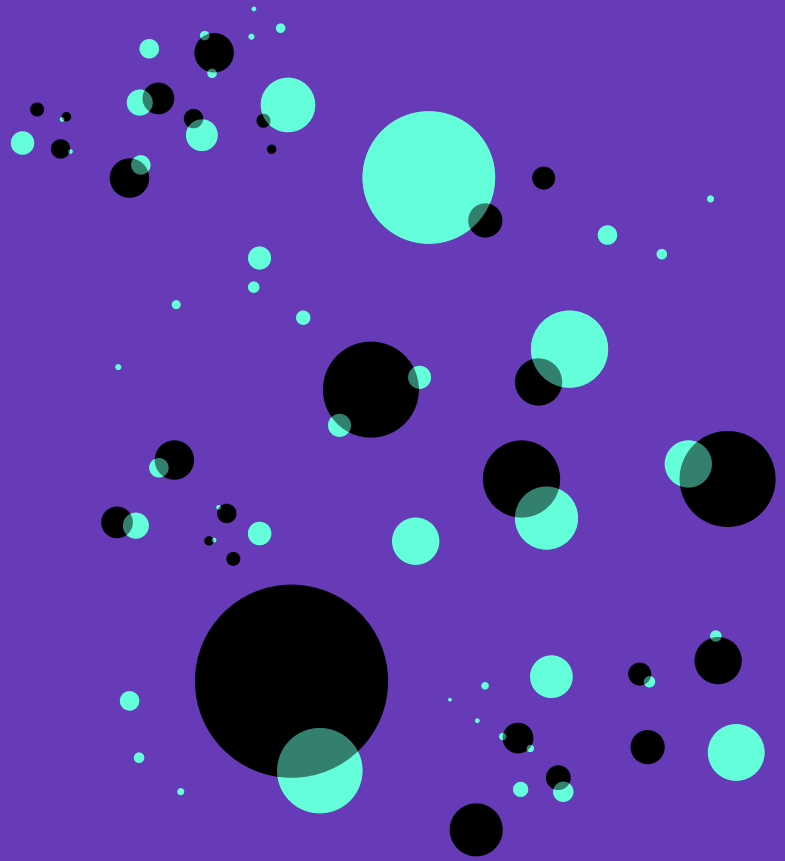
Scientific claims should be possible to reproduce given enough information, and access to data (when applicable).

Moving from majority representation...



Moving from majority representation...

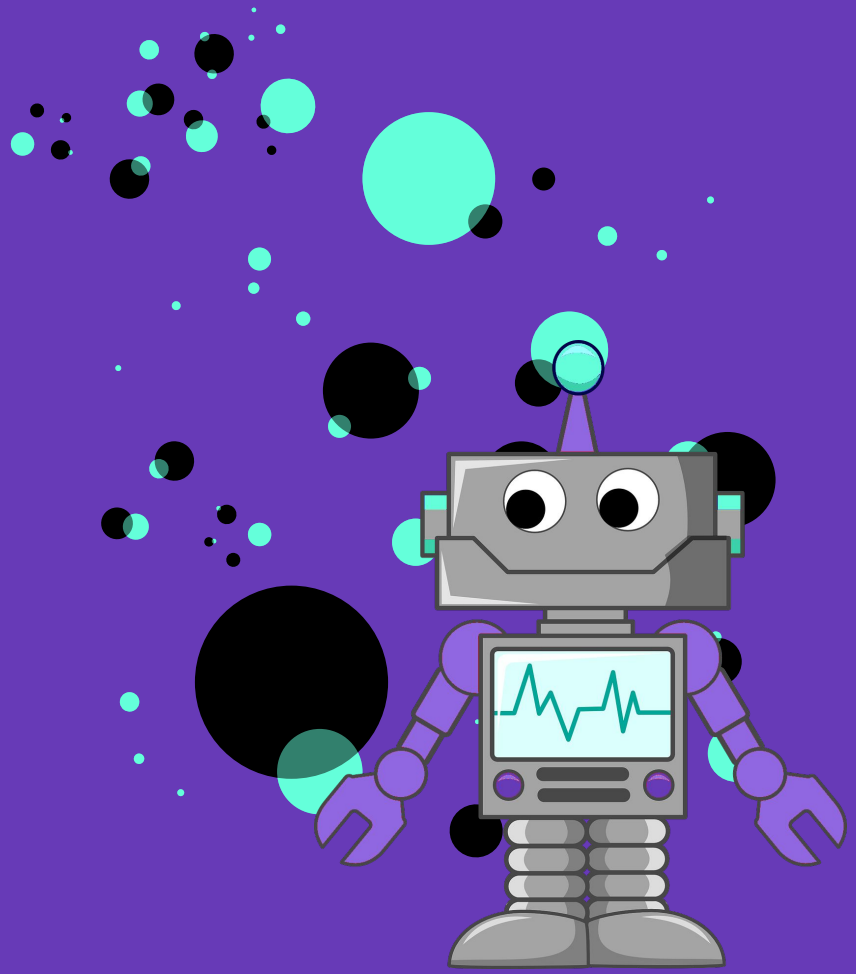
...to diverse representation



Moving from majority representation...

...to diverse representation

...for ethical AI




“Male doctor”
“Female doctor”

Thanks!

margarmitchell@gmail.com

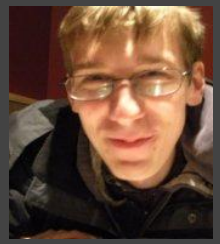
Margaret Mitchell
Google

Ishan Misra
CMU

Larry Zitnick
FAIR

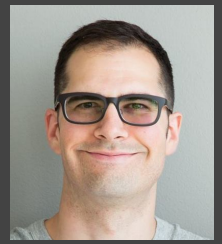
Ross Girshick
FAIR

Adrian Benton
JHU



Dirk Hovy
U. Copenhagen

Josh Lovejoy
Google



Hartwig Adam
Google

Blaise Agüera
y Arcas - Google

References

https://www.quirks.com/articles/9-types-of-research-bias-and-how-to-avoid-them

KDD Tutorial: http://francescobonchi.com/algorithmic_bias_tutorial.html

https://dub.washington.edu/djangosite/media/papers/unequalrepresentation.pdf

A. Romei and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, pp 582-638.

Benton, Mitchell, Hovy. 2017. “Multi-task learning for Mental Health Conditions with Limited Social Media Data”

Fang et al., “From Captions to Visual Concepts and Back”. CVPR 2015

Gordon, Jonathan; Van Durme, Benjamin (2013). "Reporting Bias and Knowledge Acquisition". Proceedings of the 2013 workshop on Automated knowledge base construction: 25–30.

References

Misra, et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR.

Rosch, E. (1975). Cognitive representation of semantic categories. Journal of Experimental Psychology, 104 , 192–233.

Rosch, E., C. Mervis, C., W. Gray, W., Johnson, D., & Braem, P. B. (1976). Basic objects in natural categories. Cognitive Psychology, 8 , 382–439.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7 , 573–605.

Aguera y Arcas, B, Mitchell, M & Todorov, A. “Physiognomy’s New Clothes”. Medium, 2017

Cummings, Mary (2004). "Automation Bias in Intelligent Time Critical Decision Support Systems" (PDF). *AIAA 1st Intelligent Systems Technical Conference*. ISBN 978-1-62410-080-2. doi:10.2514/6.2004-6313.