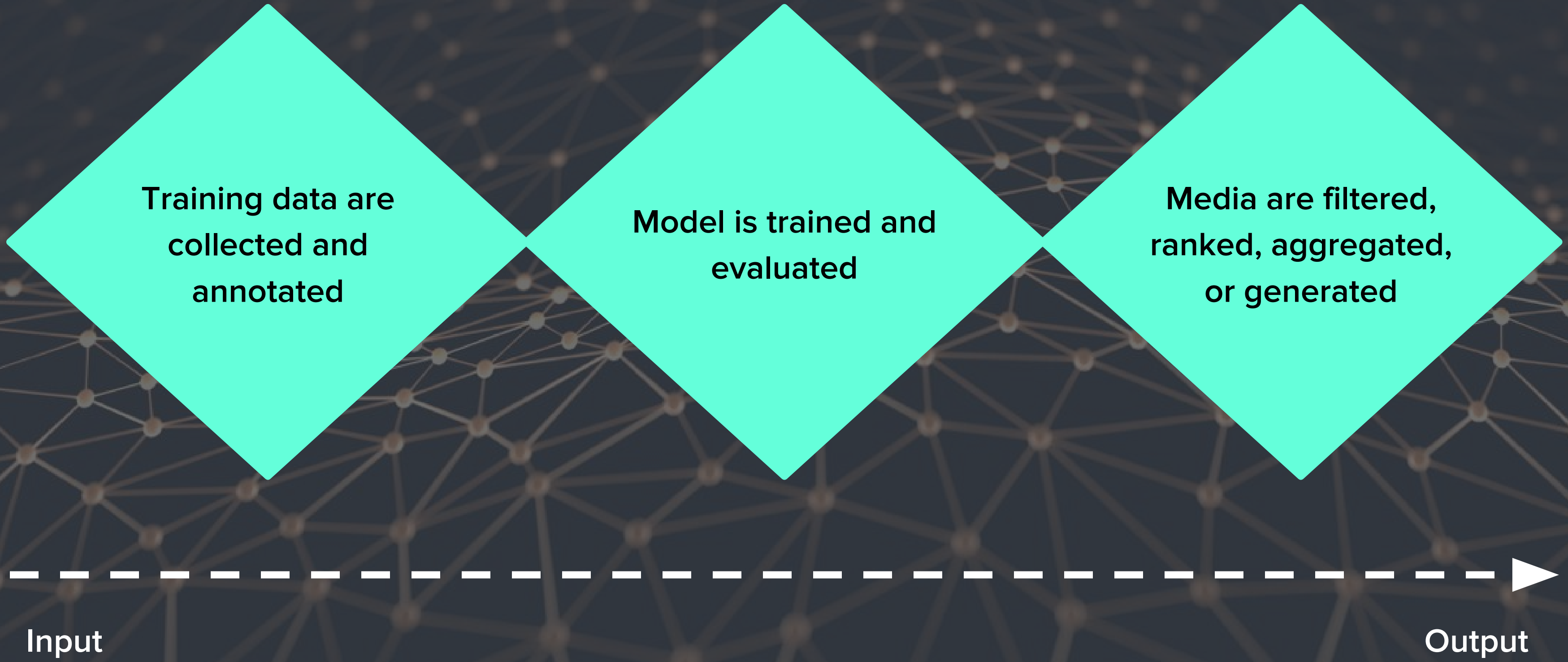
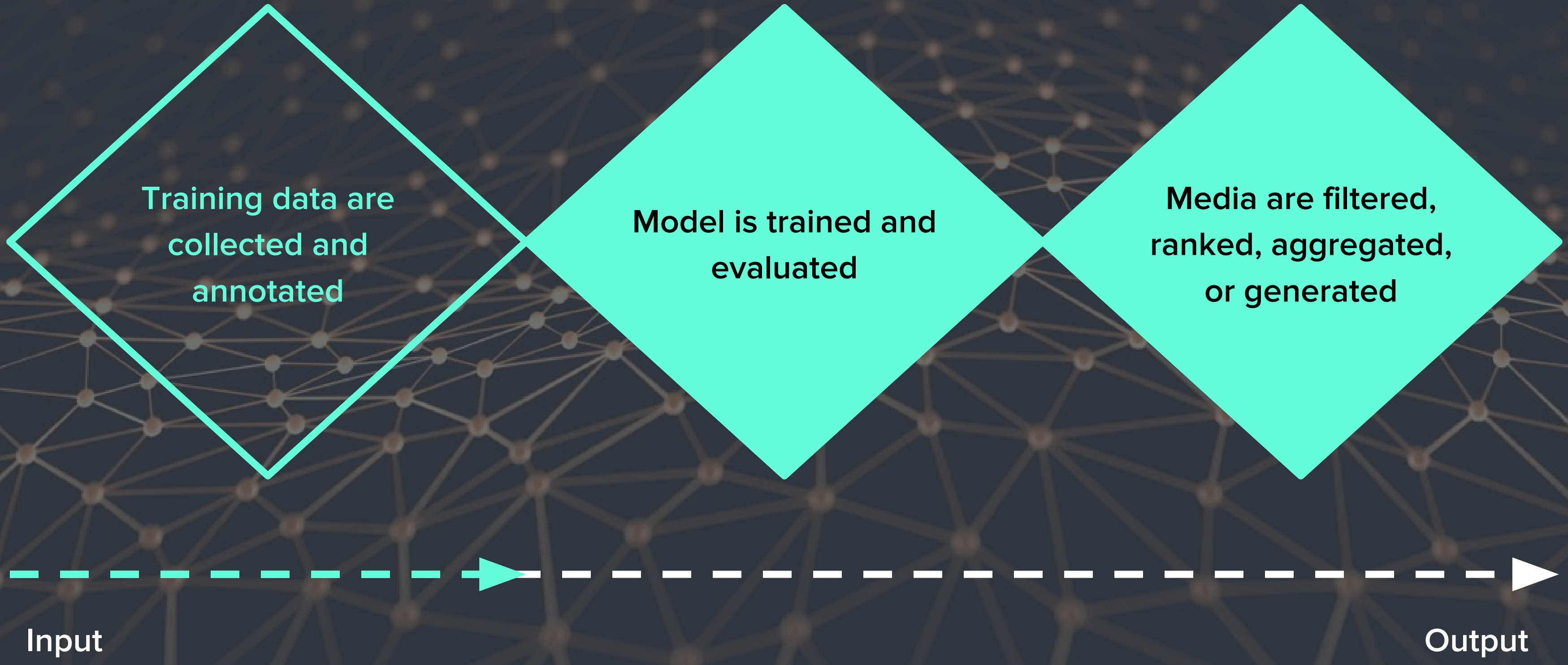
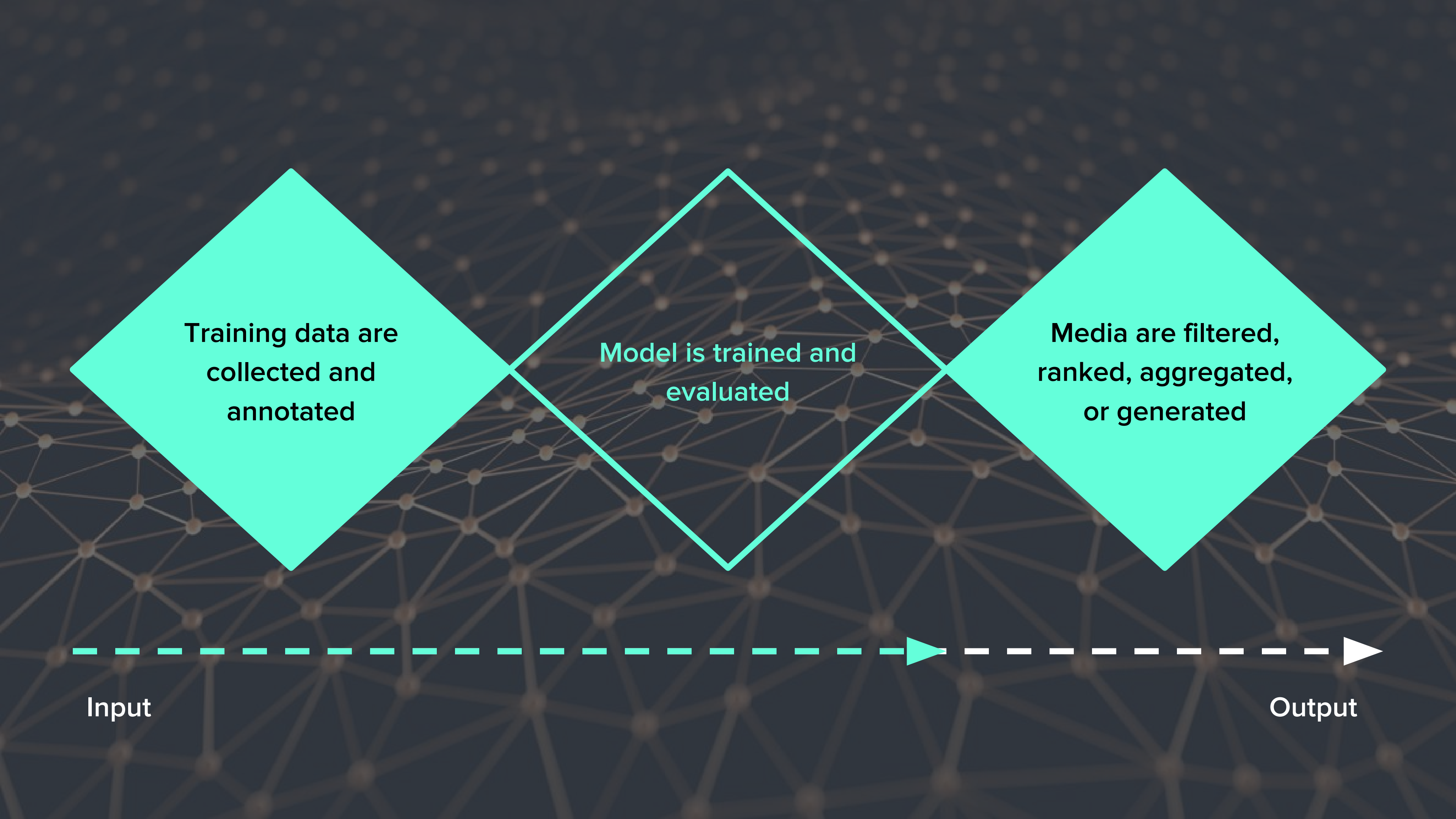
Seeds of Change

Root causes of algorithmic unfairness, and a path forward

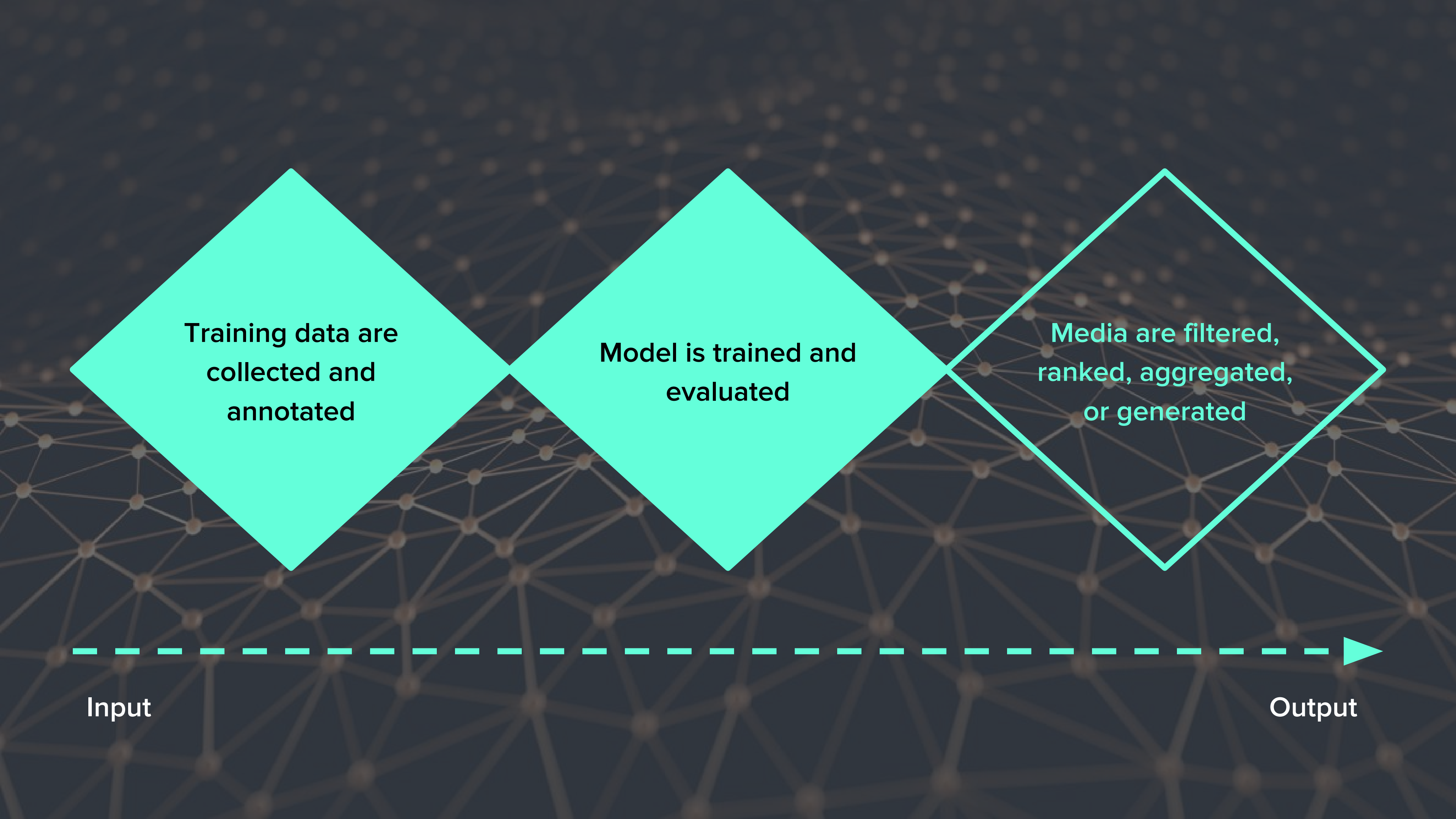
A quick recap


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


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



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


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

At a high level, where is **unfairness** creeping in?

Within the data

Reporting bias

Selection bias

Overgeneralization bias

Out-group homogeneity bias

Stereotypical bias

Historical Unfairness

Implicit associations

Implicit stereotypes

Prejudice

Group Attribution error

Halo effect

Within the data

Reporting bias
Selection bias
Overgeneralization bias
Out-group homogeneity bias
Stereotypical bias
Historical Unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group Attribution error
Halo effect

Data 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

Within the data

Reporting bias
Selection bias
Overgeneralization bias
Out-group homogeneity bias
Stereotypical bias
Historical Unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group Attribution error
Halo effect

Data 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

Training and evaluation

Evaluation metric
Features
Objective Function
Model architecture
Variables
Tasks
Hyperparameters

Within the data

Reporting bias
Selection bias
Overgeneralization bias
Out-group homogeneity bias
Stereotypical bias
Historical Unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group Attribution error
Halo effect

Data 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

Training and evaluation

Evaluation metric
Features
Objective Function
Model architecture
Variables
Tasks
Hyperparameters

We use data to estimate how likely different things are

Stereotypical bias

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.

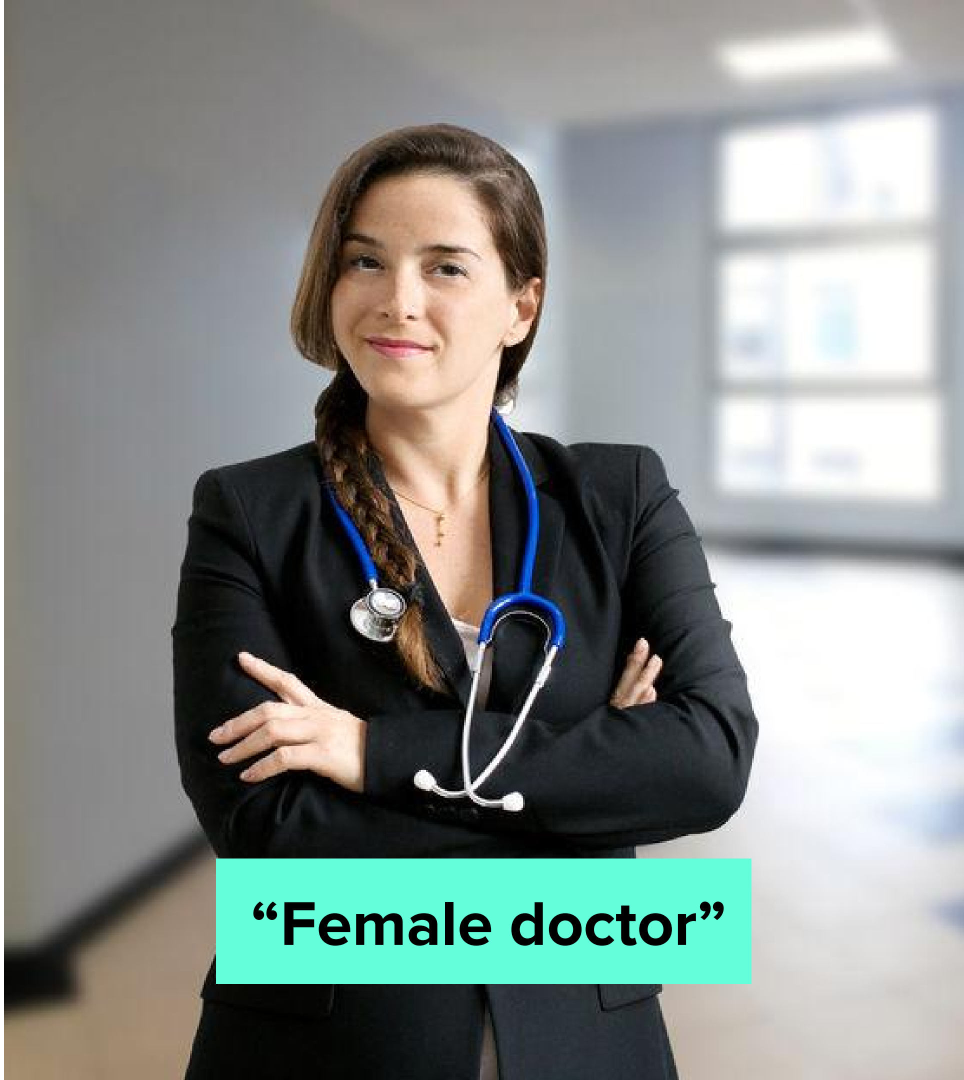
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

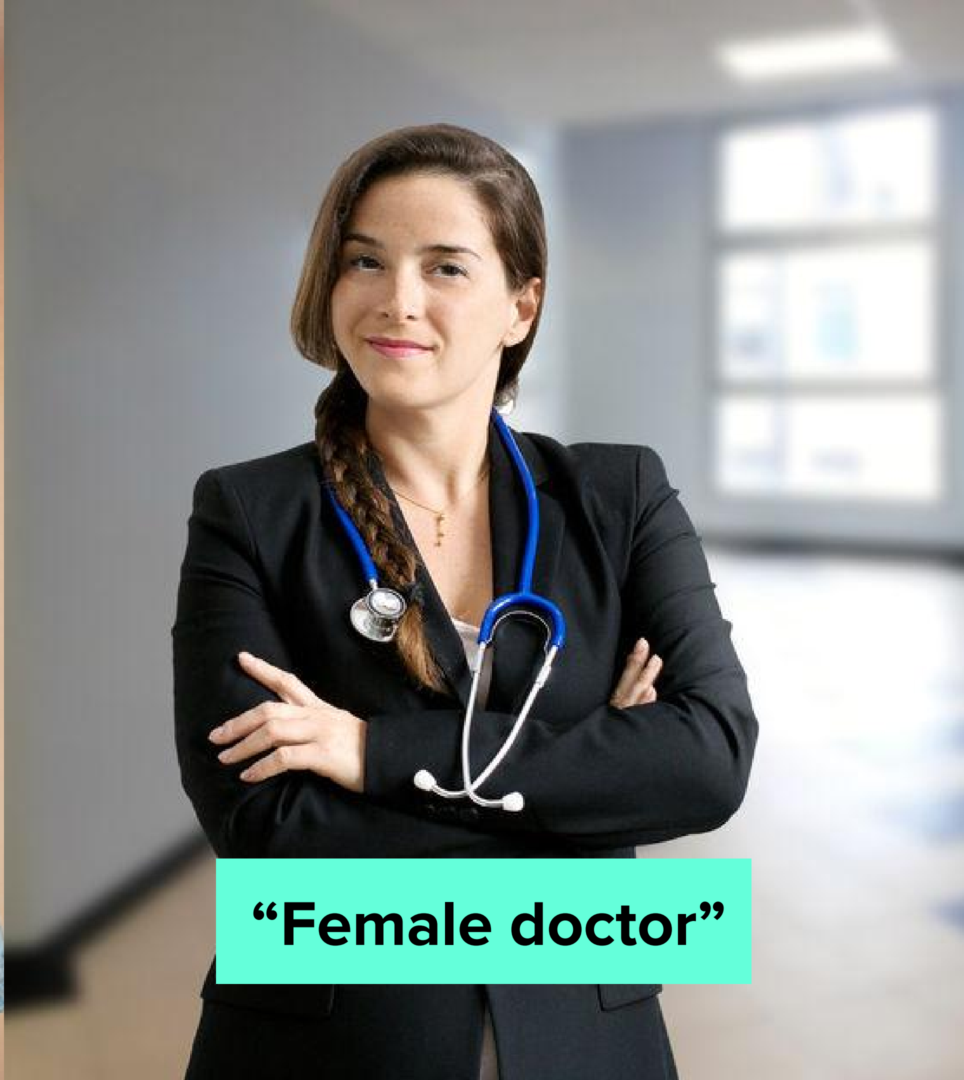
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

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

How could this be?



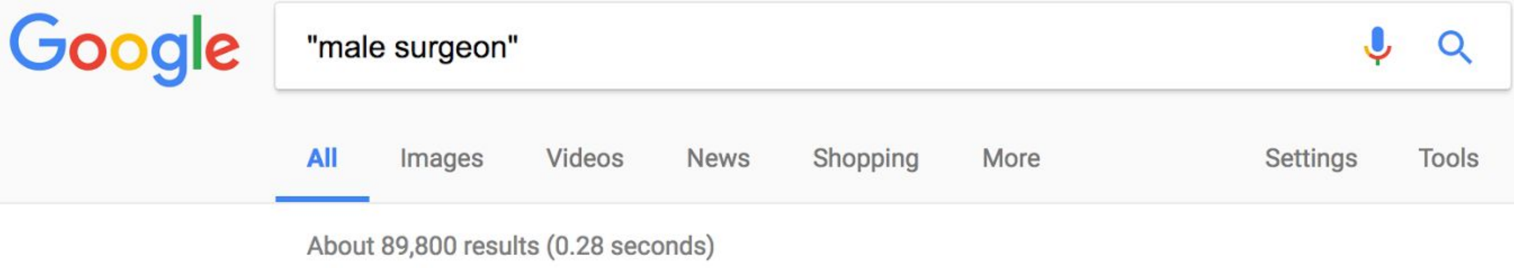

“Doctor”

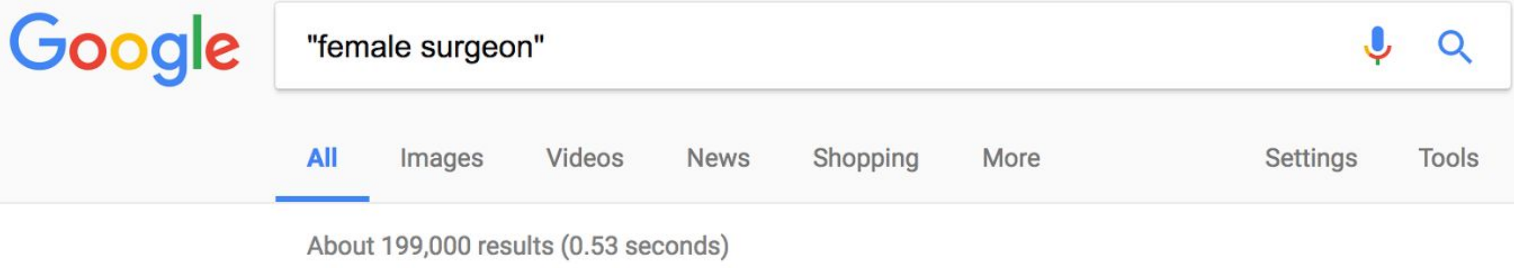

“Female doctor”

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

Reporting bias

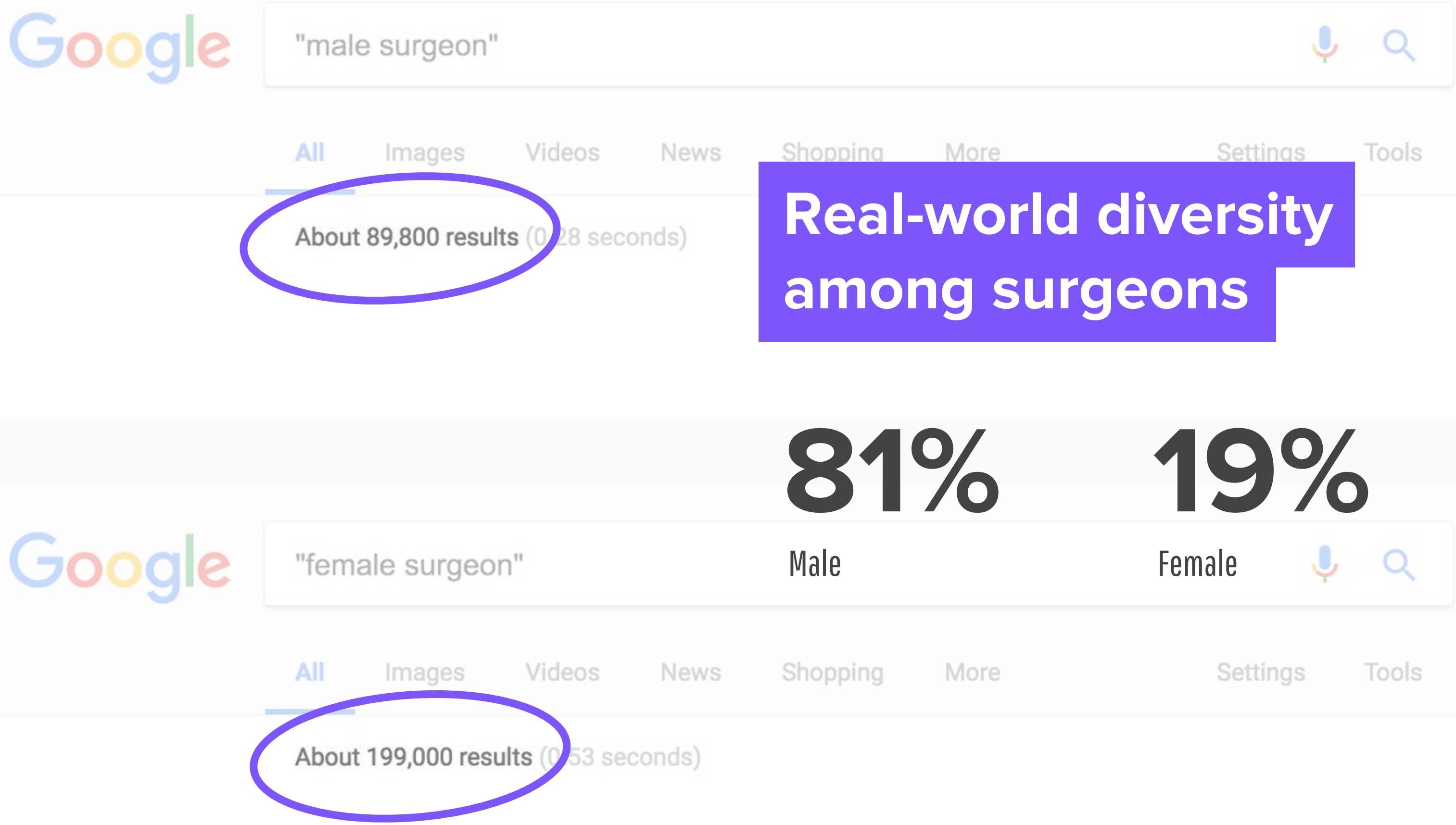

Google
"male surgeon"
All
Images
Videos
News
Shopping
More
Settings
Tools
About 89,800 results (0.28 seconds)


Google
"female surgeon"
All
Images
Videos
News
Shopping
More
Settings
Tools
About 199,000 results (0.53 seconds)

Real-world diversity among surgeons

81%
Male

19%
Female



SOURCE

[Statistics on the Number of Women Surgeons in the United States]

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

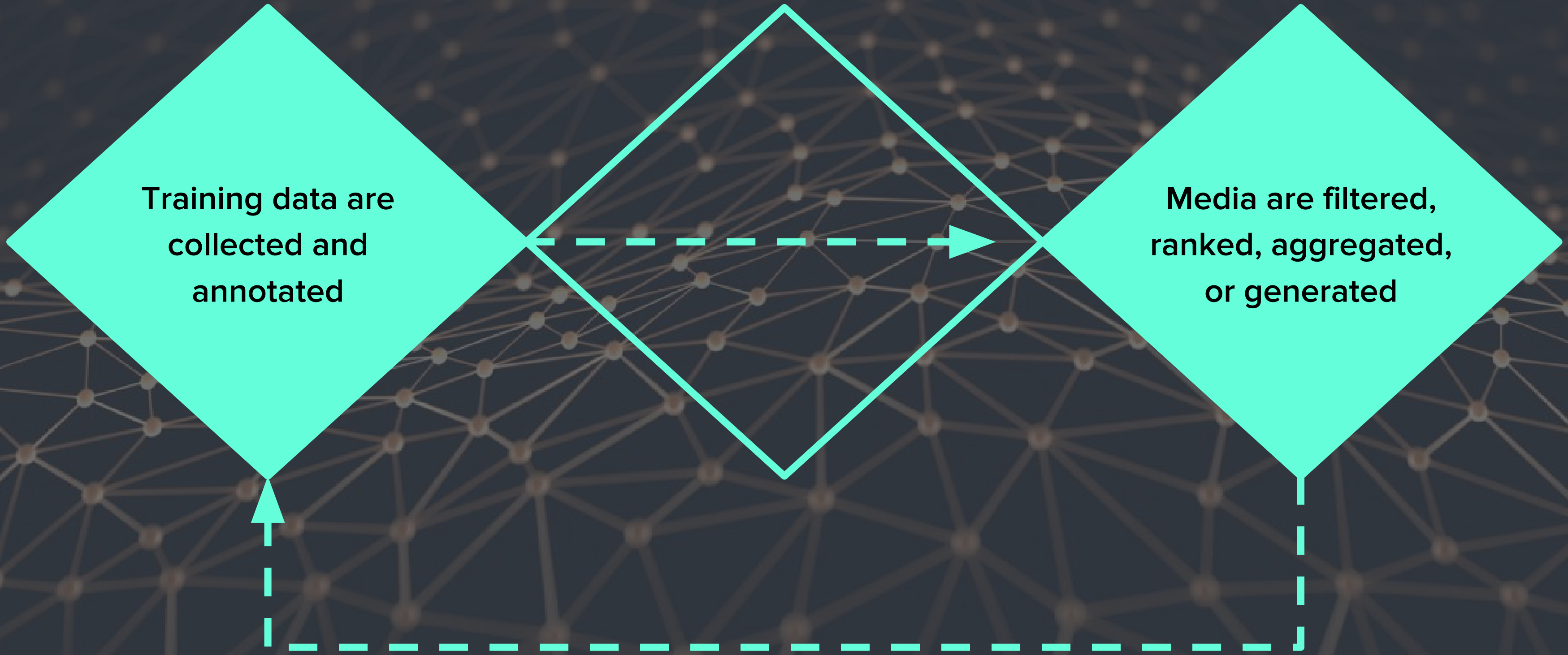
World learning from text

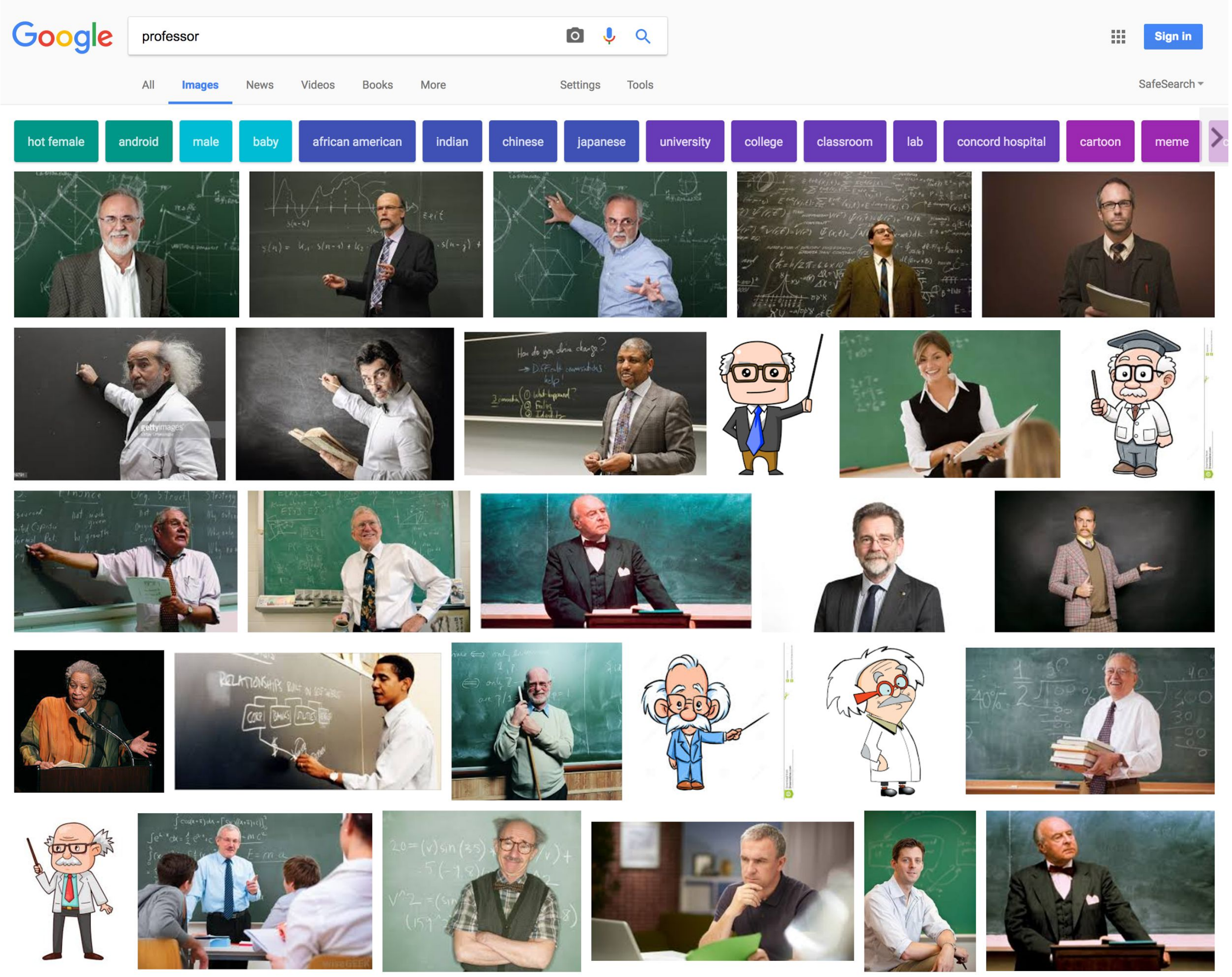
Gordon and Van Durme, 2013

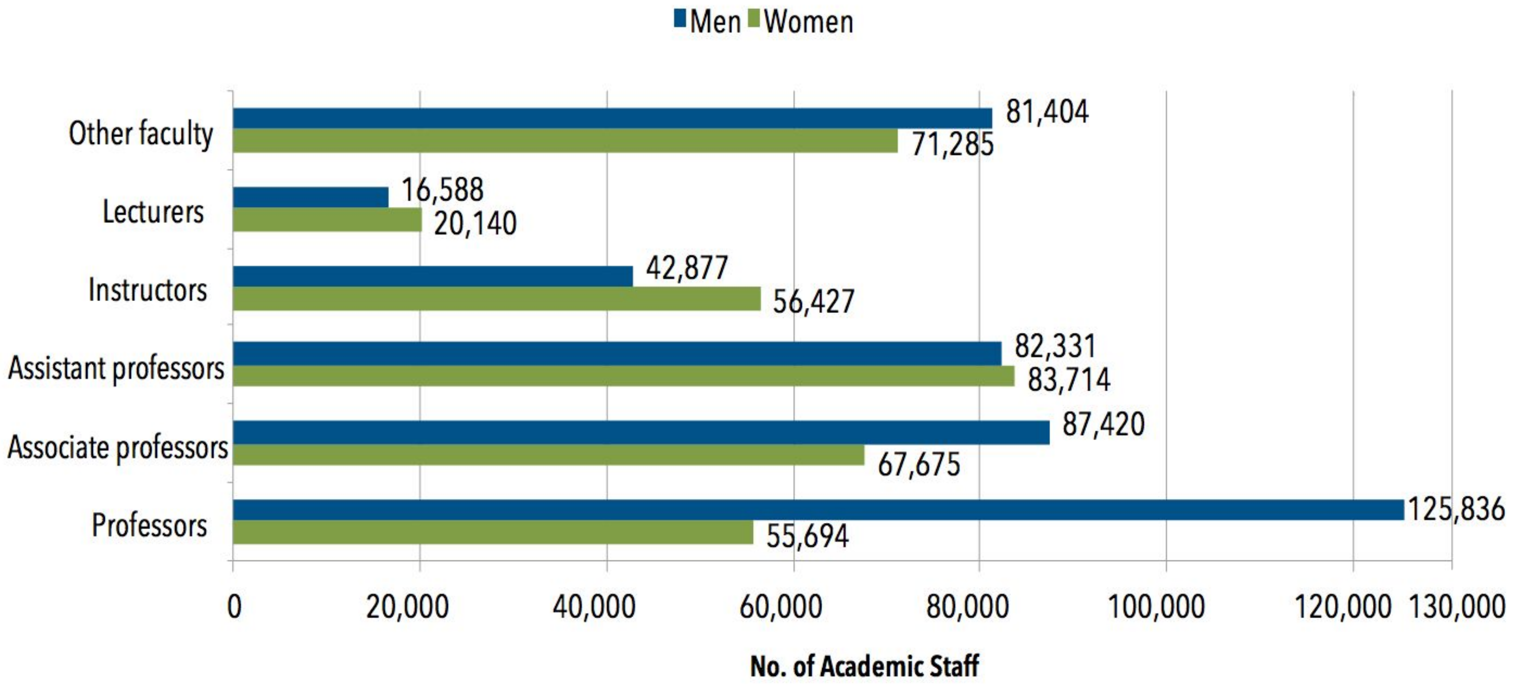
Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

Top results show historical unfairness, implicit associations, and implicit stereotypes reflected in Reporting Bias

We tend to mention and share things that are outside of our expectation of day-to-day norms; ignoring the things that “go without saying”.


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

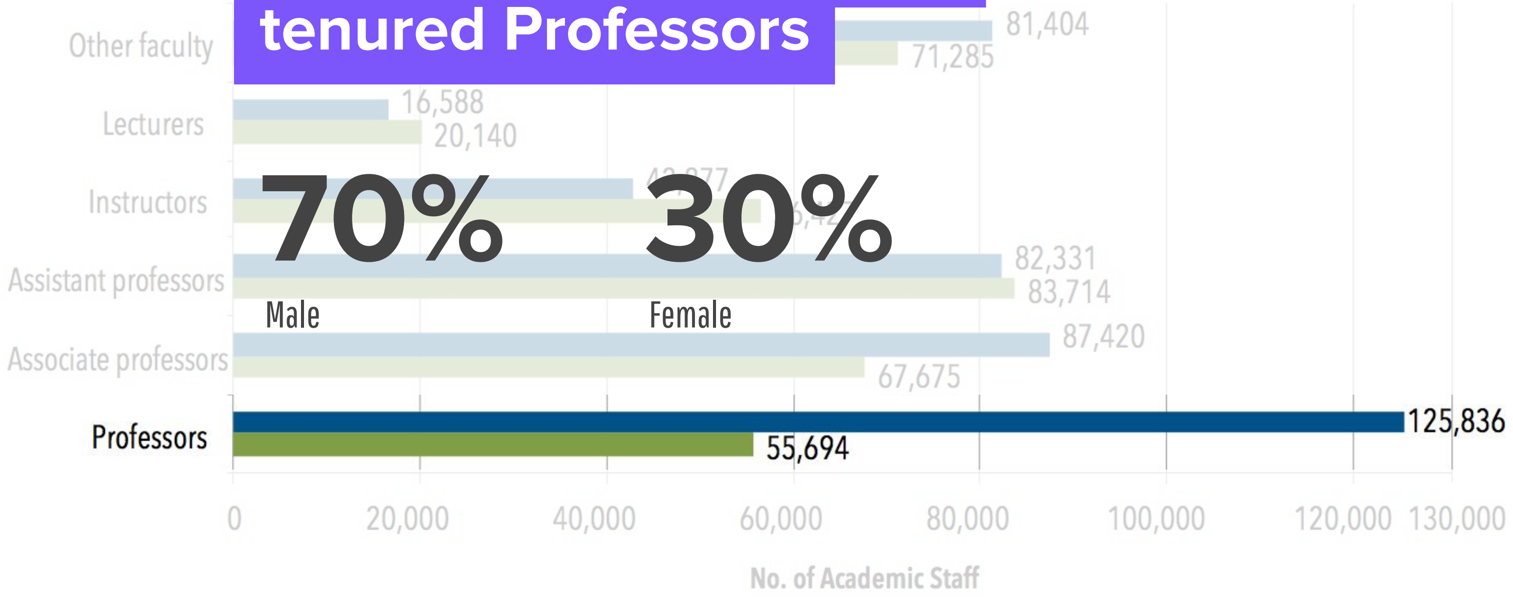

professor
Sign in
All
Images
News
Videos
Books
More
Settings
Tools
SafeSearch
hot female
android
male
baby
african american
indian
chinese
japanese
university
college
classroom
lab
concord hospital
cartoon
meme



SOURCE

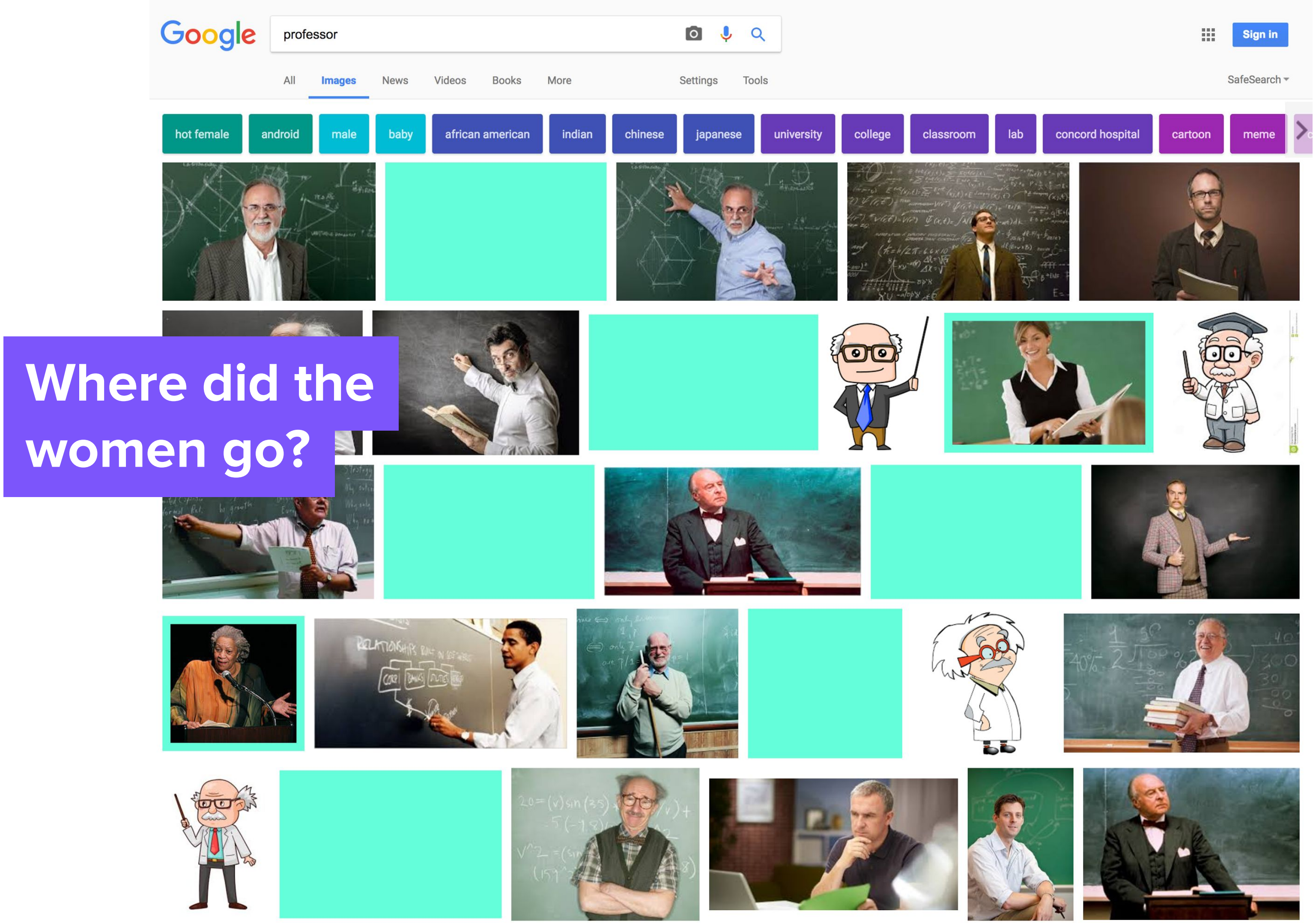
Johnson, Heather L. 2016. Pipelines, Pathways, and Institutional Leadership: An Update on the Status of Women in Higher Education. Washington, DC: American Council on Education

Gender diversity among tenured Professors

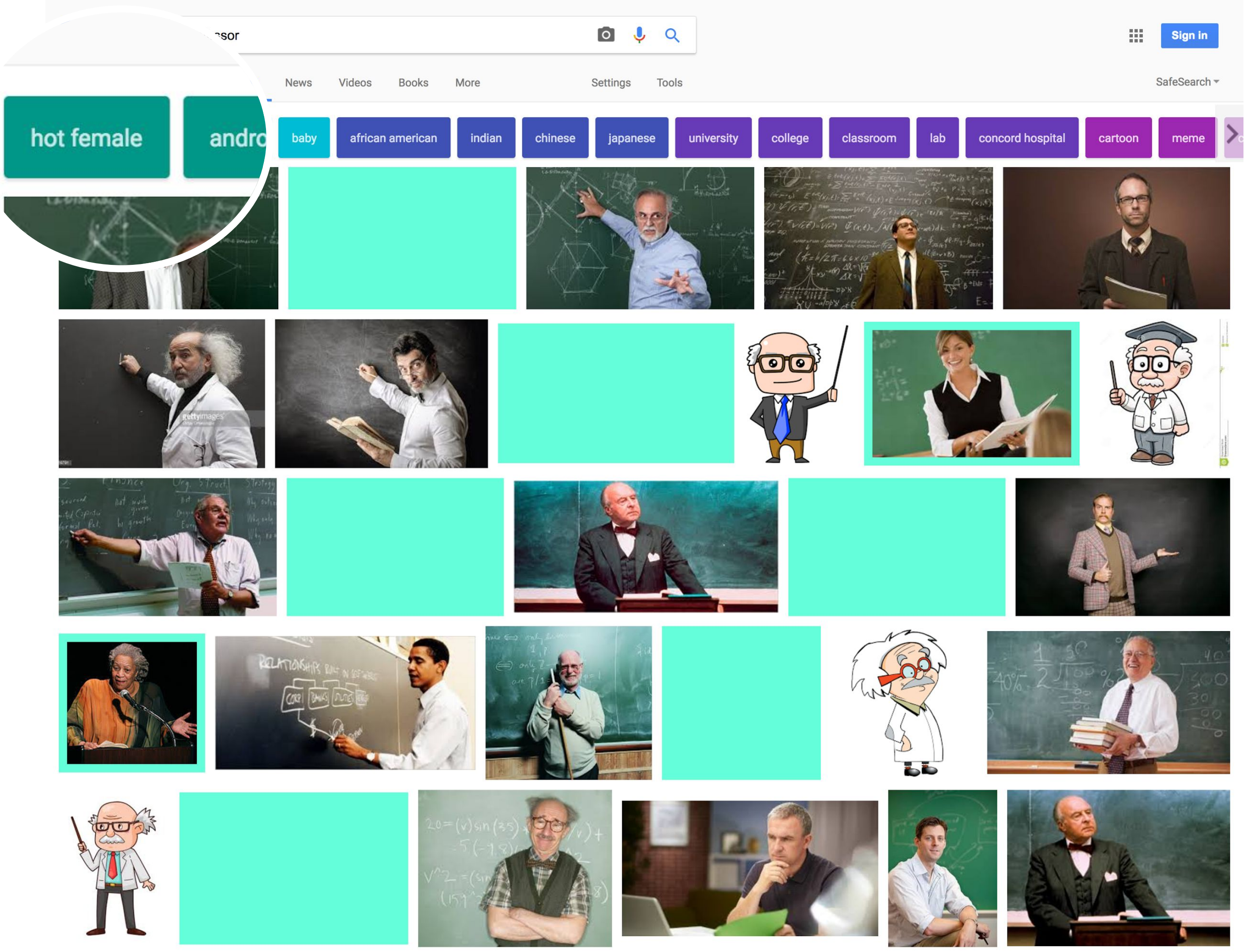


SOURCE

Johnson, Heather L. 2016. Pipelines, Pathways, and Institutional Leadership: An Update on the Status of Women in Higher Education. Washington, DC: American Council on Education



Google
professor
Sign in
All
Images
News
Videos
Books
More
Settings
Tools
SafeSearch
hot female
android
male
baby
african american
indian
chinese
japanese
university
college
classroom
lab
concord hospital
cartoon
meme
Where did the women go?



sor
News
Videos
Books
More
Settings
Tools
Sign in
SafeSearch
hot female
andro
baby
african american
indian
chinese
japanese
university
college
classroom
lab
concord hospital
cartoon
meme

Training data are collected and annotated

Model is trained and evaluated

Media are filtered, ranked, aggregated, or generated

INSIGHT: EVALUATION METRIC

The Confusion Matrix

Evaluation Metric Insights: The Confusion Matrix

Predictions

References

Evaluation Metric Insights: The Confusion Matrix

Predictions

References

Create for each (subgroup, prediction) pair.
Compare across subgroups.

Evaluation Metric Insights: The Confusion Matrix

Predictions

References

Create for each (subgroup, prediction) pair.
Compare across subgroups.

Example: women, face detection
men, face detection

Evaluation Metric Insights: The Confusion Matrix

		Predictions	
		Positive	Negative
References	Positive		
	Negative		

Evaluation Metric Insights: The Confusion Matrix

		Predictions	
		Positive	**Negative**
References	**Positive**	Reference says something exists Model predicts it **True Positives**	Reference says something exists Model doesn’t predict it **False Negatives** *Type II Error*
	Negative	Reference says something doesn’t exist Model predicts it **False Positives** *Type I error*	Reference says something doesn’t exist Model doesn’t predict it **True Negatives**

Evaluation Metric Insights: The Confusion Matrix

		Predictions	
		Positive	**Negative**
References	**Positive**	Reference says something exists Model predicts it **True Positives**	Reference says something exists Model doesn’t predict it **False Negatives** *Type II Error*
	Negative	Reference says something doesn’t exist Model predicts it **False Positives** *Type I error*	Reference says something doesn’t exist Model doesn’t predict it **True Negatives**

The Problem Areas

Evaluation Metric Insights: The Confusion Matrix

		Predictions		
		Positive	**Negative**	**Calculate**
References	**Positive**	Reference says something exists Model predicts it **True Positives**	Reference says something exists Model doesn't predict it **False Negatives** *Type II Error*	**True Positive Rate/ Sensitivity/ Recall** **False Negative Rate/ Miss Rate**
	Negative	Reference says something doesn't exist Model predicts it **False Positives** *Type I error*	Reference says something doesn't exist Model doesn't predict it **True Negatives**	**False Positive Rate/ Fallout** **True Negative Rate/ Specificity**
		Precision / Positive Predictive Value, False Discovery Rate	**Negative Predictive Value, False Omission Rate**	**LR+, LR-**

Evaluation Metric Insights: The Confusion Matrix

		Predictions: Positive	**Predictions: Negative**	**Calculate**
References	**Positive**	Reference says something exists Model predicts it **True Positives**	Reference says something exists Model doesn't predict it **False Negatives** *Type II Error*	**True Positive Rate/ Sensitivity/ Recall** **False Negative Rate/ Miss Rate**
References	**Negative**	Reference says something doesn't exist Model predicts it **False Positives** *Type I error*	Reference says something doesn't exist Model doesn't predict it **True Negatives**	**False Positive Rate/ Fallout** **True Negative Rate/ Specificity**
		Precision / Positive Predictive Value, False Discovery Rate	**Negative Predictive Value, False Omission Rate**	**LR+, LR-**

Evaluation Metric Insights: The Confusion Matrix

		Predictions		
		Positive	**Negative**	**Calculate**
References	**Positive**	Reference says something exists Model predicts it **True Positives**	Reference says something exists Model doesn't predict it **False Negatives** *Type II Error*	**True Positive Rate/ Sensitivity/ Recall** **False Negative Rate/ Miss Rate**
	Negative	Reference says something doesn't exist Model predicts it **False Positives** *Type I error*	Reference says something doesn't exist Model doesn't predict it **True Negatives**	**False Positive Rate/ Fallout** **True Negative Rate/ Specificity**
		Precision / Positive Predictive Value, False Discovery Rate	**Negative Predictive Value, False Omission Rate**	**LR+, LR-**

Evaluation Metric: Error trade-offs

False Positive

(Type I error)

False Negative

(Type II Error)

Error trade-offs

Real World Example:

Error trade-offs

Real World Example:

- Project working with clinicians for mental health

Error trade-offs

Real World Example:

- Project working with clinicians for mental health
- Trying to detect suicide risk

Error trade-offs

Real World Example:

- Project working with clinicians for mental health
- Trying to detect suicide risk
- For patient trust (and sanity), important not to have False Positives

Error trade-offs

Real World Example:

- Project working with clinicians for mental health
- Trying to detect suicide risk
- For patient trust (and sanity), important not to have False Positives
    - Predicting suicide risk when there is not a risk

Error trade-offs

Real World Example:

- Project working with clinicians for mental health
- Trying to detect suicide risk
- For patient trust (and sanity), important not to have False Positives
  - Predicting suicide risk when there is not a risk
- Prioritize True Positive Rate at a low False Positive Rate

Choose your evaluation metrics in light of acceptable tradeoffs between False Positives and False Negatives.

TOOL: EVALUATION METRICS

Lantern: Guided Model Analysis

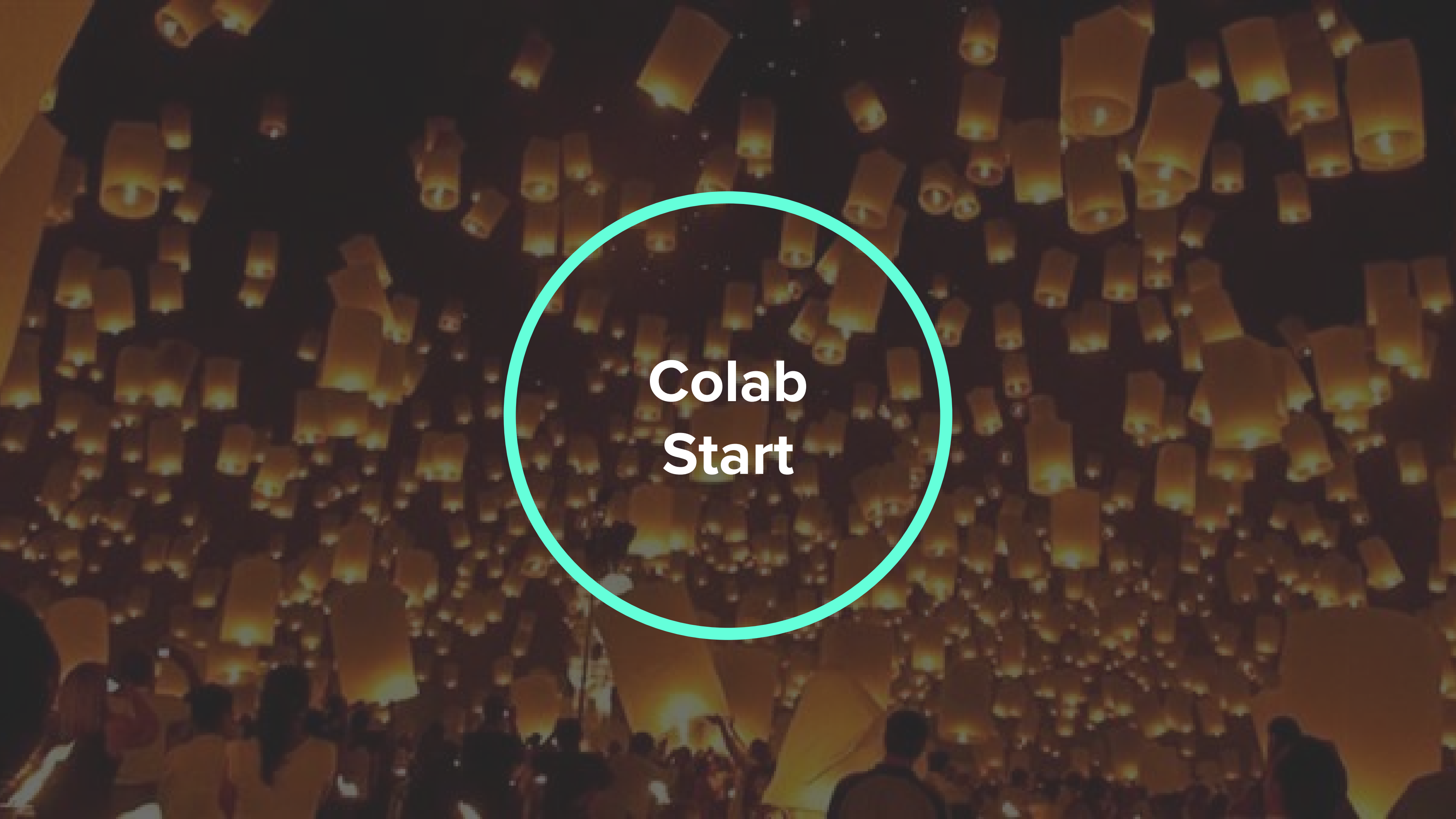
Model Evaluation

+

Data slicing

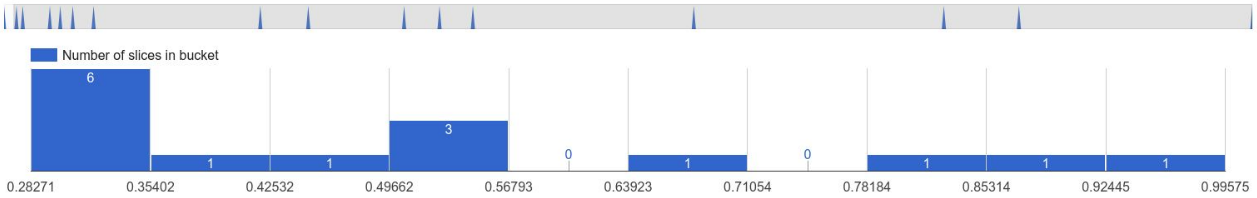
=

Better Understanding of Disproportionate Outcomes



Colab
Start

go/lantern-eval-colab



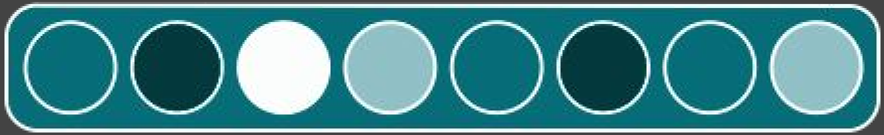
feature	totalWeightedExamples	averageLabel	averageRefinedPrediction	calibration	logisticLoss	auc	auprc	binaryConfusionMatricesFromRegression (Threshold: 0.30000)				
								F1 Score	Precision	Recall	Accuracy	Specificity
education:10th	933	0.06645	0.21602	3.25074	0.32892	0.85757	0.27181	0.16842	0.24242	0.12903	0.91533	0.97130
education:11th	1175	0.05106	0.21899	4.28859	0.28501	0.86086	0.43380	0.39216	0.47619	0.33333	0.94723	0.98027
education:12th	433	0.07621	0.22183	2.91067	0.31049	0.89189	0.46742	0.36735	0.56250	0.27273	0.92841	0.98250
education:5th-6th	333	0.04805	0.21693	4.51487	0.28830	0.83517	0.38840	0.28571	0.33333	0.25000	0.93994	0.97476
education:7th-8th	646	0.06192	0.22079	3.56580	0.31755	0.75342	0.23235	0.17949	0.18421	0.17500	0.90093	0.94884
education:9th	514	0.05253	0.21513	4.09547	0.30427	0.71621	0.19608	0.12500	0.14286	0.11111	0.91829	0.96304
education:Assoc-acdm	1067	0.24836	0.24023	0.96728	0.50756	0.83630	0.63774	0.30904	0.67949	0.20000	0.77788	0.96883
education:Assoc-voc	1382	0.26122	0.24748	0.94741	0.52772	0.81889	0.61423	0.31643	0.59091	0.21607	0.75615	0.94711
education:Bachelors	5355	0.41475	0.27218	0.65624	0.67363	0.82166	0.75993	0.35000	0.80291	0.22377	0.65528	0.96107
education:Doctorate	413	0.74092	0.32741	0.44190	0.99575	0.74769	0.89145	0.36220	0.92000	0.22549	0.41162	0.94393
education:HS-grad	10501	0.15951	0.22960	1.43942	0.42488	0.82322	0.47113	0.26211	0.47567	0.18090	0.83754	0.96216
education:Masters	1723	0.55659	0.30197	0.54254	0.81734	0.81346	0.83589	0.37001	0.84701	0.23670	0.55136	0.94634
education:Preschool	51	0.00000	0.22628	0.00000	0.54768	0.00000	0.00000	0.00000	0.00000	0.00000	0.96078	0.96078
education:Prof-school	576	0.73438	0.39011	0.53122	0.86115	0.82715	0.92702	0.48687	0.93919	0.32861	0.49132	0.94118
education:Some-college	7291	0.19023	0.23249	1.22211	0.45257	0.85802	0.57256	0.28229	0.57813	0.18673	0.81937	0.96799

INSIGHT: FEATURES

Word embeddings

Common ML Feature: Word Embeddings

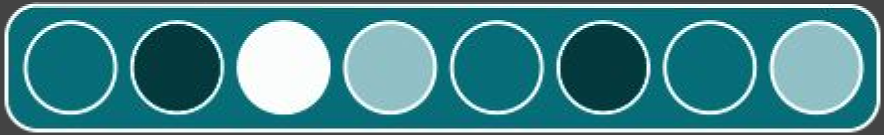
Word embeddings represent each word as a vector.

$\overrightarrow{\text{woman}}$: 

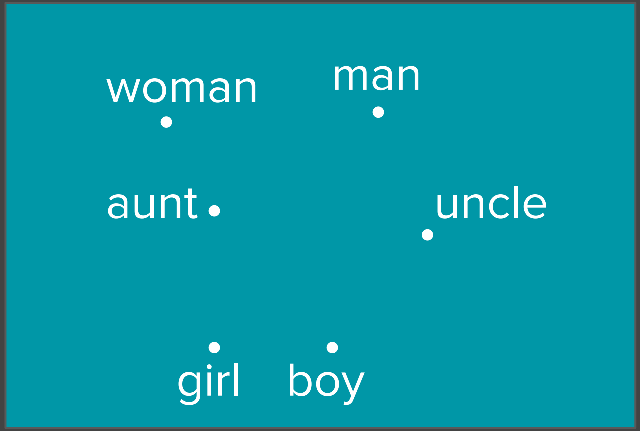
Common ML Feature: Word Embeddings

Word embeddings represent each word as a vector.

$\overrightarrow{\text{woman}}$:



Allows us to calculate similarity between words.



Common ML Feature: Word Embeddings

Word embeddings represent each word as a vector.

Similarities between embeddings can be found using **cosine distance**:

$$\cos(\vec{man}, \vec{woman}) = \frac{\vec{man} \cdot \vec{woman}}{\|\vec{man}\| \cdot \|\vec{woman}\|}$$

Common ML Feature: Word Embeddings

Word embeddings represent each word as a vector.

Similarities between embeddings can be found using cosine distance.

Similarities between the **difference** between vectors can also be calculated using cosine distance.

$\vec{g} = \vec{man} - \vec{woman}$

$\vec{r} = \vec{king} - \vec{queen}$

$$\cos(\vec{g}, \vec{r}) = \frac{\vec{g} \cdot \vec{r}}{||\vec{g}|| \cdot ||\vec{r}||}$$

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016). "Man is to Computer Programmer as Woman is to Homemaker?: Debiasing Word Embeddings". Proceedings of NIPS.

Common ML Feature: Word Embeddings

Word embeddings represent each word as a vector.

Similarities between embeddings can be found using cosine distance.

Similarities between the difference between vectors can also be calculated using cosine distance.

This can show us roughly equivalent relationships between words.

$$\overrightarrow{\textbf{man}} - \overrightarrow{\textbf{woman}} \approx \overrightarrow{\textbf{king}} - \overrightarrow{\textbf{queen}}$$

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016). "Man is to Computer Programmer as Woman is to Homemaker?: Debiasing Word Embeddings". Proceedings of NIPS.

Common ML Feature: Word Embeddings

Word embeddings represent each word as a vector.

Similarities between embeddings can be found using cosine distance.

Similarities between the difference between vectors can also be calculated using cosine distance.

This can show us roughly equivalent relationships between words … including unfairness.

$$\vec{\textbf{man}} - \vec{\textbf{woman}} \approx \vec{\textbf{king}} - \vec{\textbf{queen}}$$

$$\vec{\textbf{man}} - \vec{\textbf{woman}} \approx \textbf{computer } \vec{\textbf{programmer}} - \vec{\textbf{homemaker}}$$

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016). "Man is to Computer Programmer as Woman is to Homemaker?: Debiasing Word Embeddings". Proceedings of NIPS.

Potential Solution: Debias your embeddings

High-Level:

1. Calculate the representation of a concept, like “gender”, using word embeddings.
2. Subtract this representation from learned word embeddings.
3. Use a hyperparameter to define how much this subtraction effects the embedding.

Link to Code

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016). “Man is to Computer Programmer as Woman is to Homemaker?: Debiasing Word Embeddings”. Proceedings of NIPS.

TECHNIQUE: EMBEDDINGS

Embeddings with Tensorflow

Embeddings
reveal words used in
similar contexts within
your dataset.

Colab Start

go/tf-embedding-colab

Id	L2 Distance↑	L2 Norm	Adjust	Word
2000	0.000000	1.000000	Remove	teacher
1736	0.707160	1.000000	Add	teachers
44702	0.732374	1.000000	Add	guidance_counselor
6229	0.740699	1.000000	Add	elementary
105512	0.791613	1.000000	Add	paraprofessional
371401	0.795801	1.000000	Add	paraeducator
13229	0.798513	1.000000	Add	Teacher
931	0.829719	1.000000	Add	student
198	0.833520	1.000000	Add	school
4825	0.837015	1.000000	Add	classroom

https://g3doc.corp.google.com/engedu/ml/mldays/g3doc/embeddings_demo.md

Embeddings Demo

THE JOURNEY CONTINUES

Fairness-Relevant Tools

Google-internal

go/mlx	Suite of tools useful for different aspects of fairness/bias. Some key tools also listed below.
go/tfx Codelab	Computes statistics over data for visualization and example validation; anomaly detection; etc.
go/mlx tools	Great list of tools to help visualize different aspects of your model.
go/mlx-lantern Codelab	Computes evaluation metrics and loss for slices of your data with visualization. Interested in adding further support relevant to fairness in particular. Use with go/tfx or Sibyl.
go/ml-dash	Compare metrics; visualize loss over time; etc.
go/wide-n-deep	Combine the benefits of wide models and deep models (deep learning).
go/multitask	Support multitask (multi-headed) learning. Predicting several tasks at once can be useful for the tasks to mutually benefit one another.
go/glassbox	Interpretable machine learning.
go/bias	Report biased Google products.

Google-internal

Embedding Projector 📕	View how different strings of text pattern with other strings in a high-dimensional space.
go/mledu-in-embeddings 📕	View word relationships in embedding space.
Rank Lab □ Recipes & Best Practices	Supports feature ablation experiments, shuffling.
Fast Feature Ablation □	Fast Feature Ablation (FFA) adapts the feature ablation process cpop/jpg developed for SmartASS to an implementation suitable for Tensorflow and TF.Learn specifically.
Chain □□ Codelab	Provides easy handling for moving from detection to evaluation. Includes a face attribute client: Age/Gender/UHS estimates (common in semantic scene understanding).
Affective Computing □□	Label images for affective states, emotions, etc.
VSEval □□ Codelab	Flexible infrastructure to acquire, store, and share high-quality ground truth, as well as by offering insightful statistics and visualization tools to support such research.
Learning Arbiter □□ Codelab	The Arbiter Perception Eval system is in development! It aims to be a modular service oriented ecosystem built to ease up the evaluation of machine perception models.

Thanks!

dsculley@
mmitchellai@

ML Fairness

Machine Learning, Subgroup Discovery

go/ml-fairness-tools
go/ml-fairness-metrics

References

Benton, Adrian; Mitchell, Margaret; Hovy, Dirk (2017). "Multi-task learning for Mental Health Conditions with Limited Social Media Data". Proceedings of EACL.

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016). "Man is to Computer Programmer as Woman is to Homemaker?: Debiasing Word Embeddings". Proceedings of NIPS.

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

Kay, Matthew; Matuszek, Cynthia; Munson, Sean A. (2015). "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations". Proceedings of CHI.

Misra, Ishan; Girshick, Ross; Mitchell, Margaret; Zitnick, Larry (2016). "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels". Proceedings of CVPR.

Wapman, Mikaela; Belle, Deborah (2014). "A Riddle Reveals Depth of Gender Bias". Boston University. As reported by Barlow, Rich. BU Today. https://www.bu.edu/today/2014/bu-research-riddle-reveals-the-depth-of-gender-bias/

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

THE JOURNEY CONTINUES

Additional Slides

INSIGHT: TASKS

Leverage multiple tasks to improve performance across different subgroups

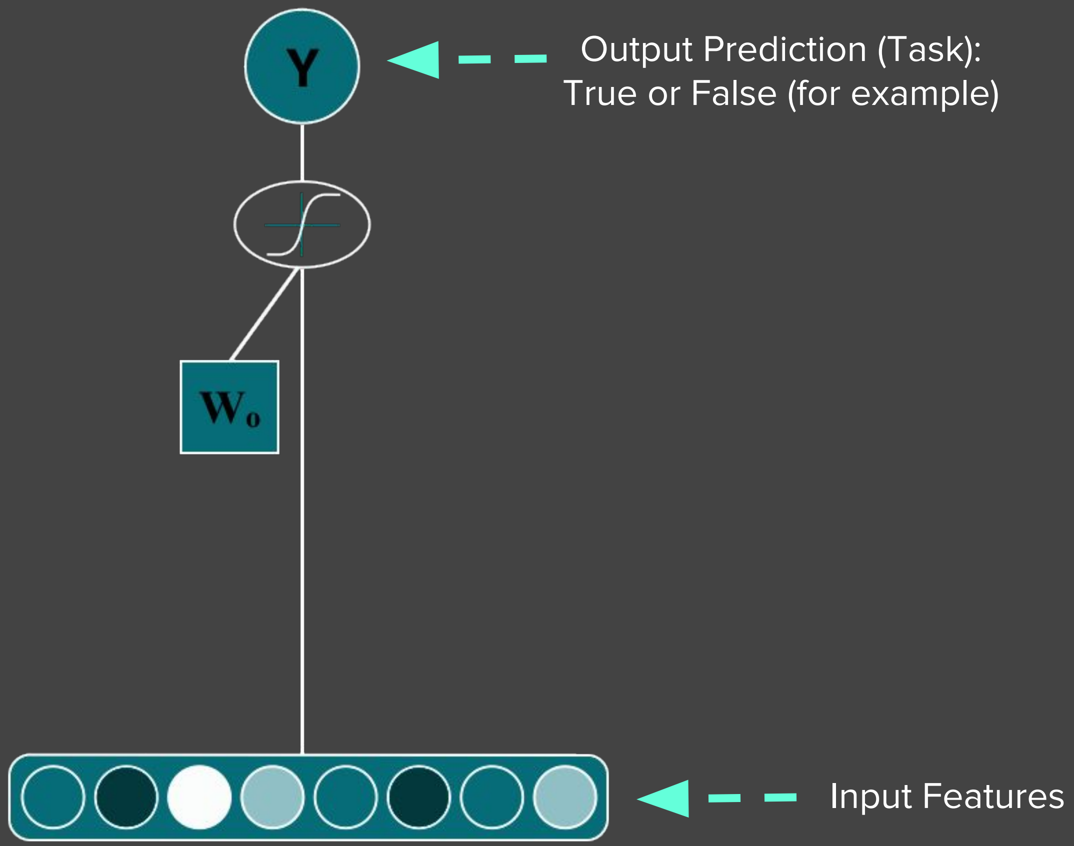
go/tf-multitask

Motivation from "The Karate Kid"

Single-task Learners
(STL)

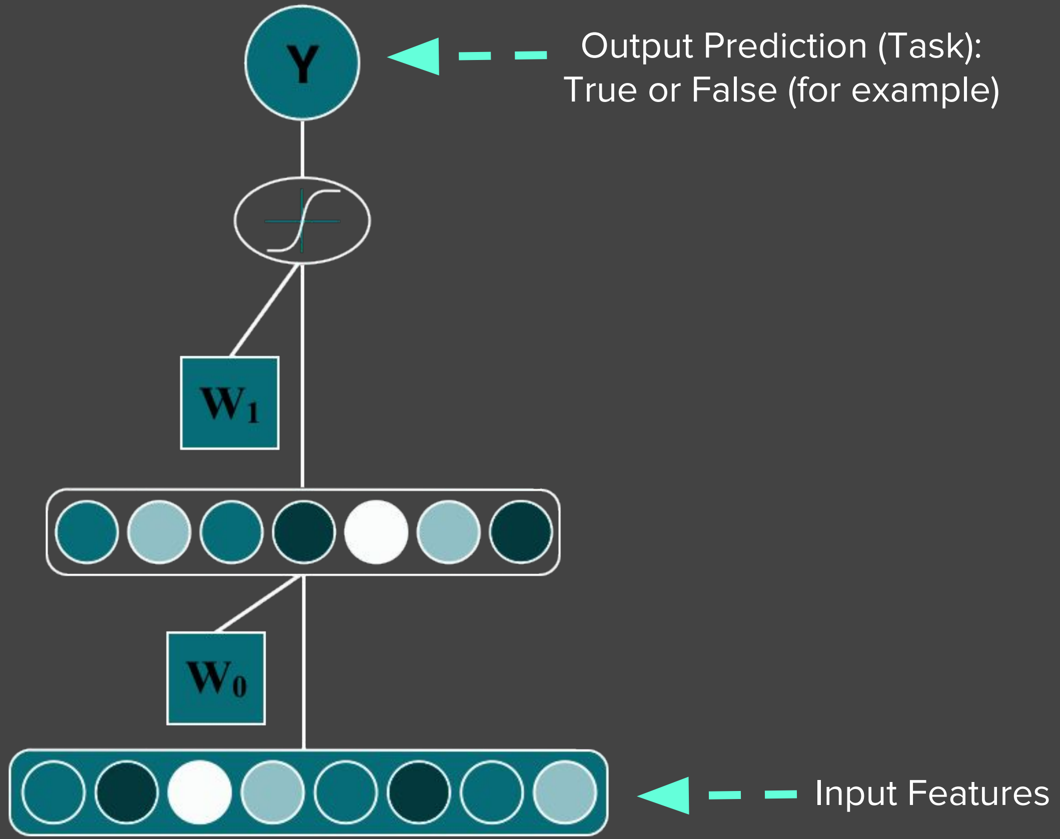
Multitask Learner
(MTL)

Single-Task: Logistic Regression

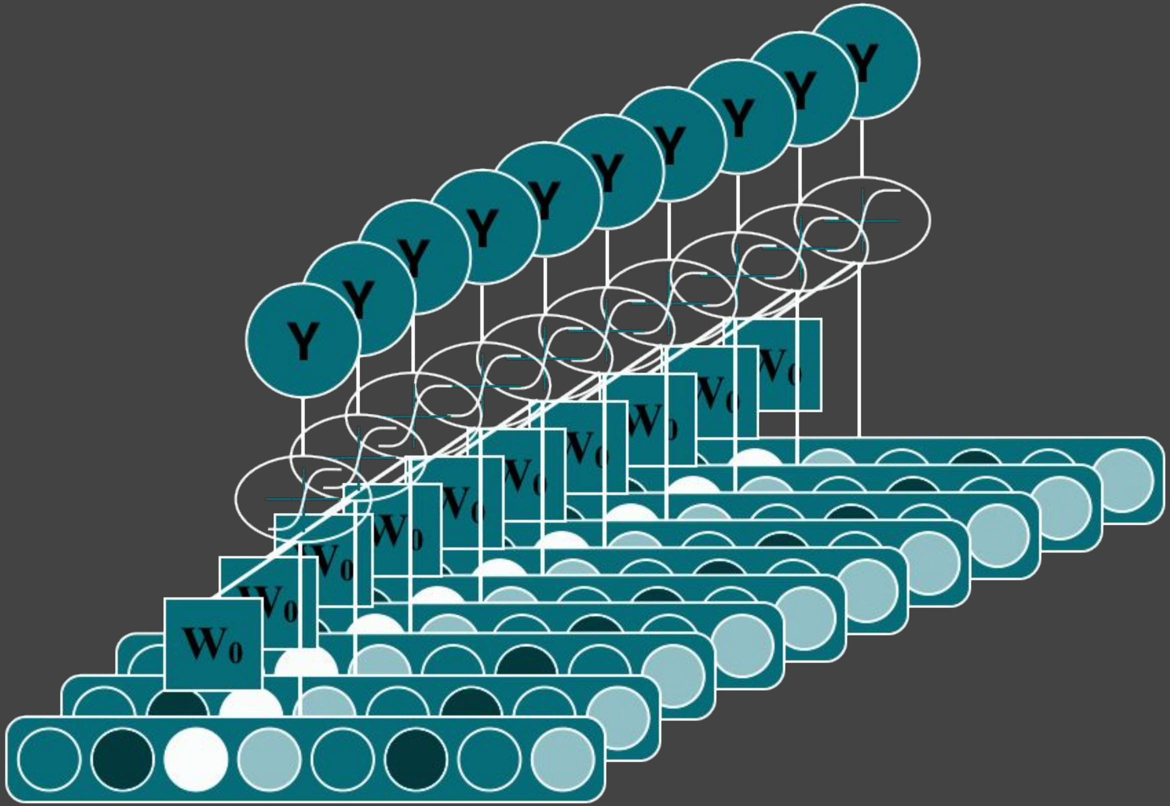


Single-Task: Deep Learning

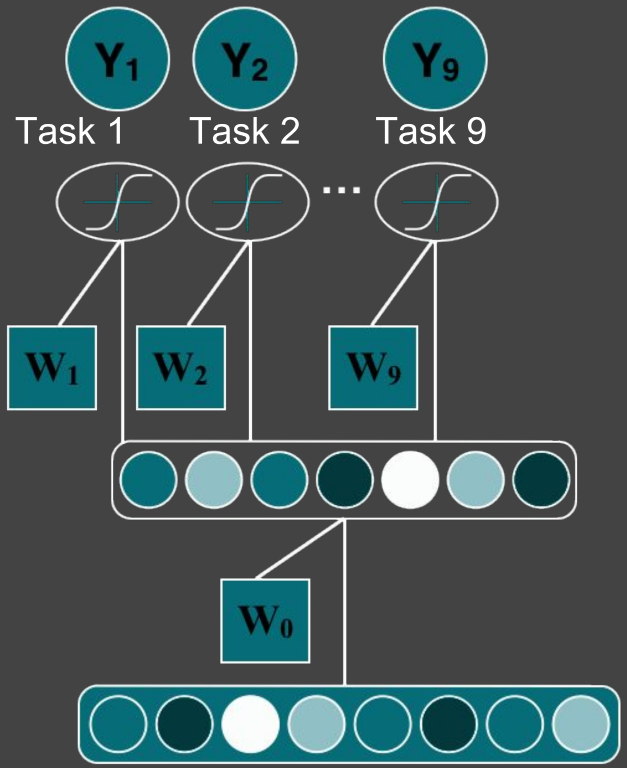
Fancier!!



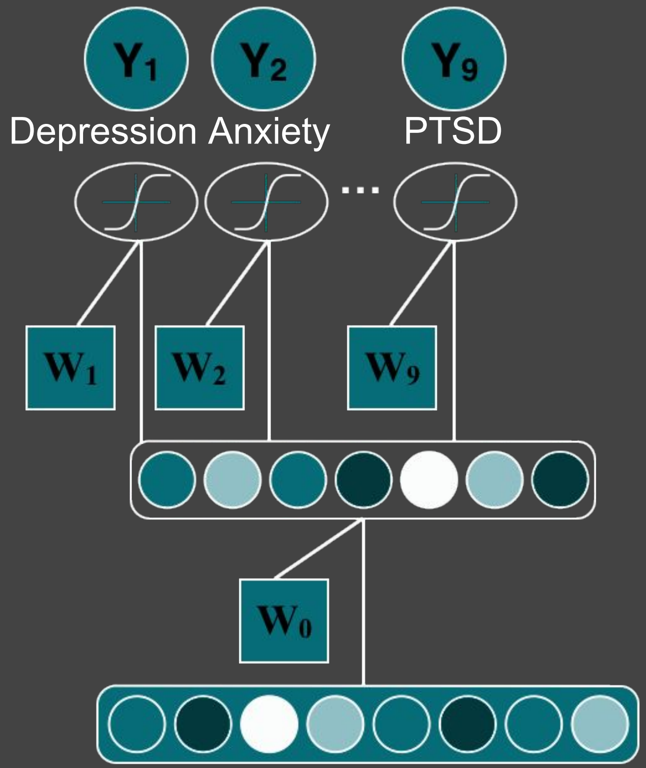
Multiple Tasks with Basic Logistic Regression



Multiple Tasks + Deep Learning: Multi-task Learning



Multiple Tasks + Deep Learning: Multi-task Learning Example

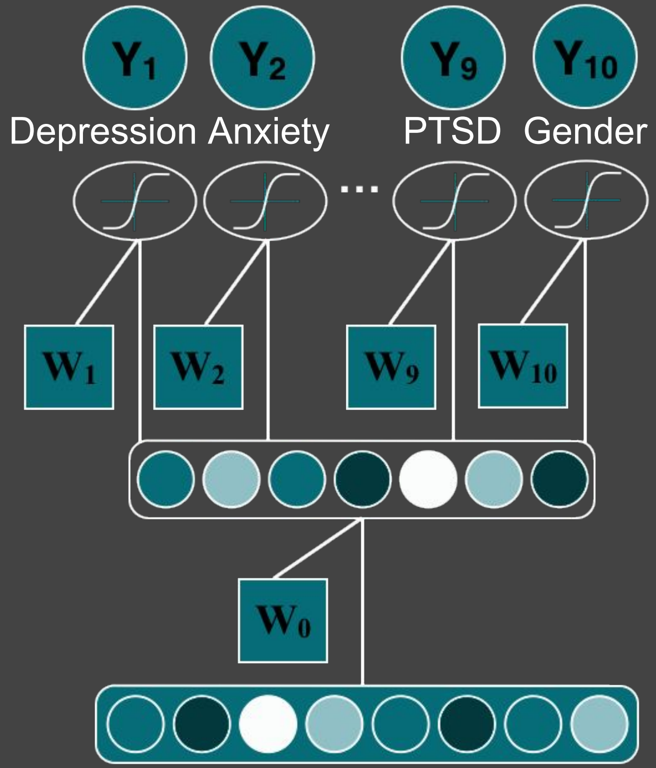


Task	N
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	**9611**

<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

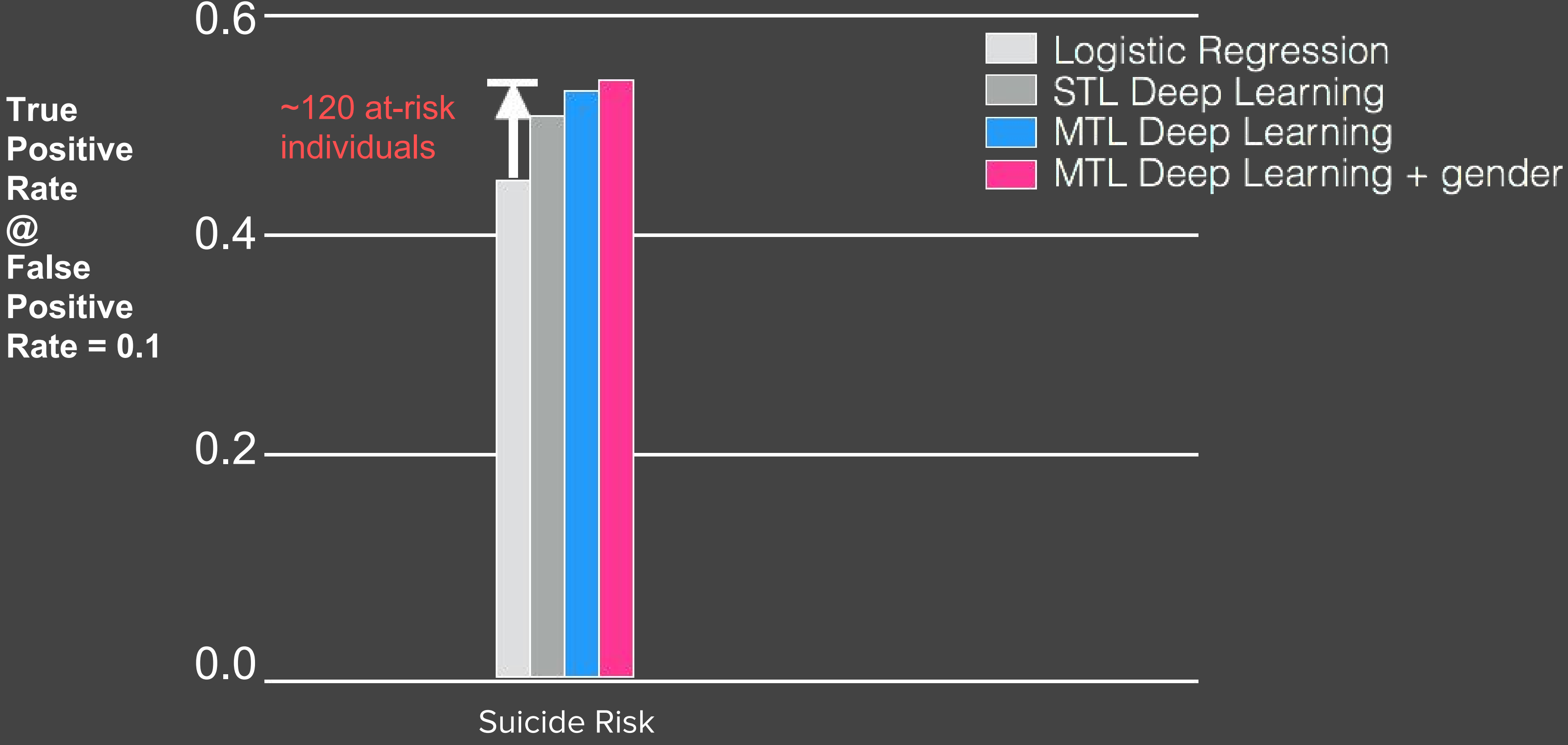


Task	N
Gender	1101
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	**9611**

<5% positive examples

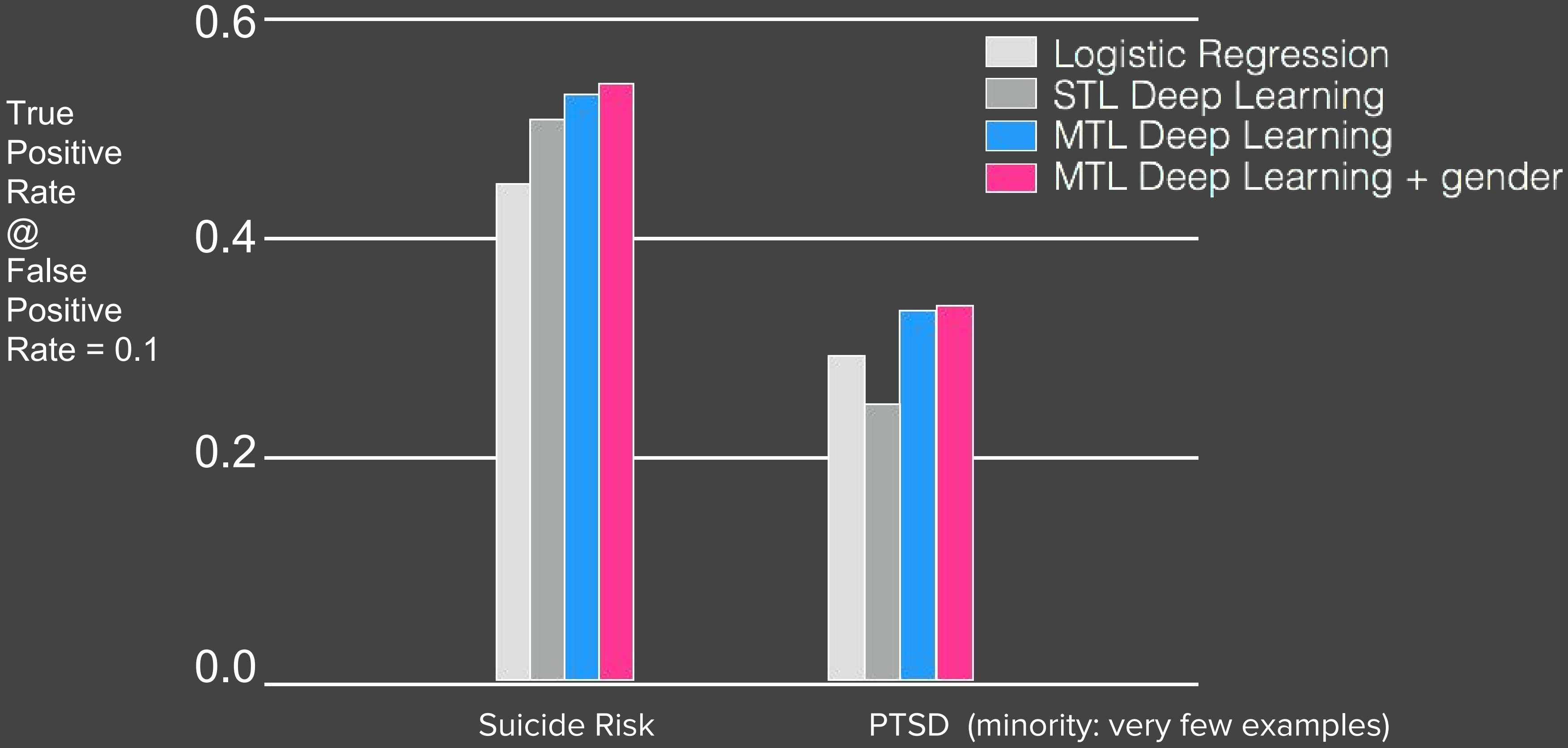
[Benton, Mitchell, Hovy. 2017. “Multi-task learning for Mental Health Conditions with Limited Social Media Data”]

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”

Lantern: Guided Model Analysis, including Multi-Task!

Includes offline model evaluations, computation of metrics on different slices of the data

feature	auc	auprc	averageLabel	averageRefinedPrediction	binaryConfusionMatricesFromRegression			
					Threshold: 0.75000			
					F1 Score	Precision	Recall	Accuracy
age:19	0.66929	0.70809	0.55309	0.55243	0.40000	0.10000	0.30000	0.20000
age:18	0.68247	0.73486	0.62338	0.46560	0.41000	0.11000	0.31000	0.21000
age:20	0.66872	0.70736	0.55309	0.56765	0.43000	0.13000	0.33000	0.23000
age:22	0.71525	0.77450	0.62338	0.46510	0.44000	0.14000	0.34000	0.24000

[go/mlx-lantern]

[Source Document for Multi-Task Models]

INSIGHT: OBJECTIVE FUNCTION

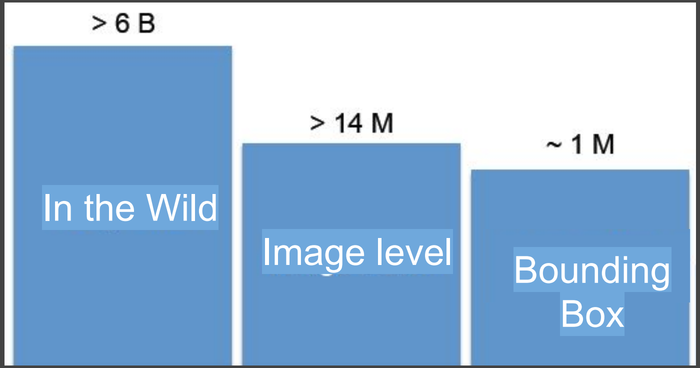
Visual presence + Relevance

Data data everywhere ...

Facebook

300 Million images uploaded everyday

flickr



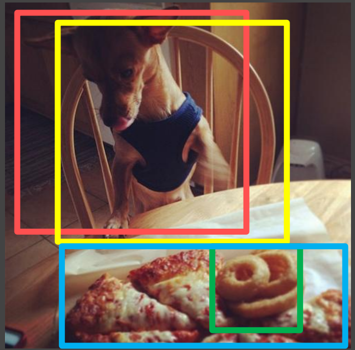
YouTube

100 hours of video every minute

#dog #hungry

OMG Frodo is sitting eating pizza and donuts.

dog, chair, pizza, donut



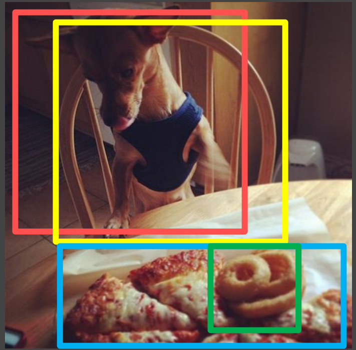
dog, chair, pizza, donut

Data data everywhere ...

But not many labels to train

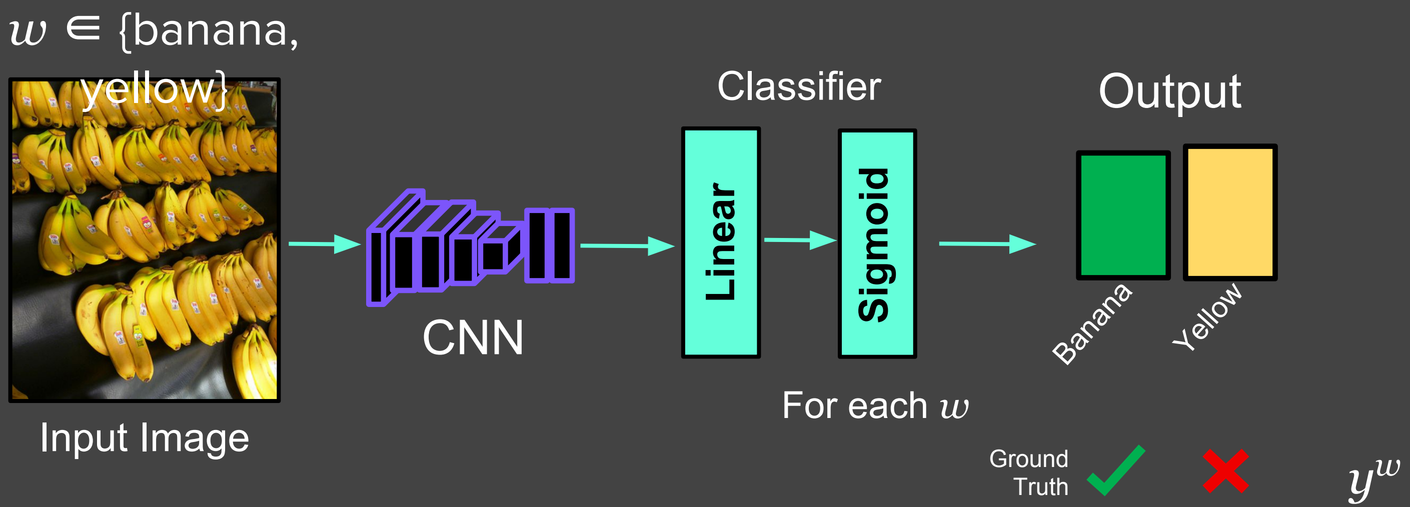
Exhaustively annotated data is expensive

dog, chair, pizza, donut



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

Simple Image Classification



"Gold standard" Annotation: Human-biased label $y^w \in \{0, 1\}$

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

Factoring in Reporting Bias: Idea

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

Factoring in Reporting Bias: Idea

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

$w \in$ {banana, yellow}

Factoring in Reporting Bias: Idea

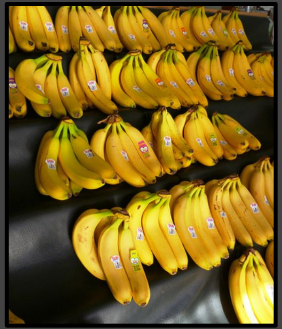
- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the concept visually present?*
  - Relevance r – *Is the concept relevant for a human?*

$w \in$ {banana,
yellow}

Factoring in Reporting Bias: Idea

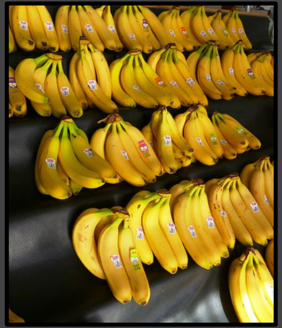
- A human-biased prediction h can be factored into two terms
  - Visual presence v – *Is the concept visually present?*
  - Relevance r – *Is the concept relevant for a human?*

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



Factoring in Reporting Bias: Idea

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



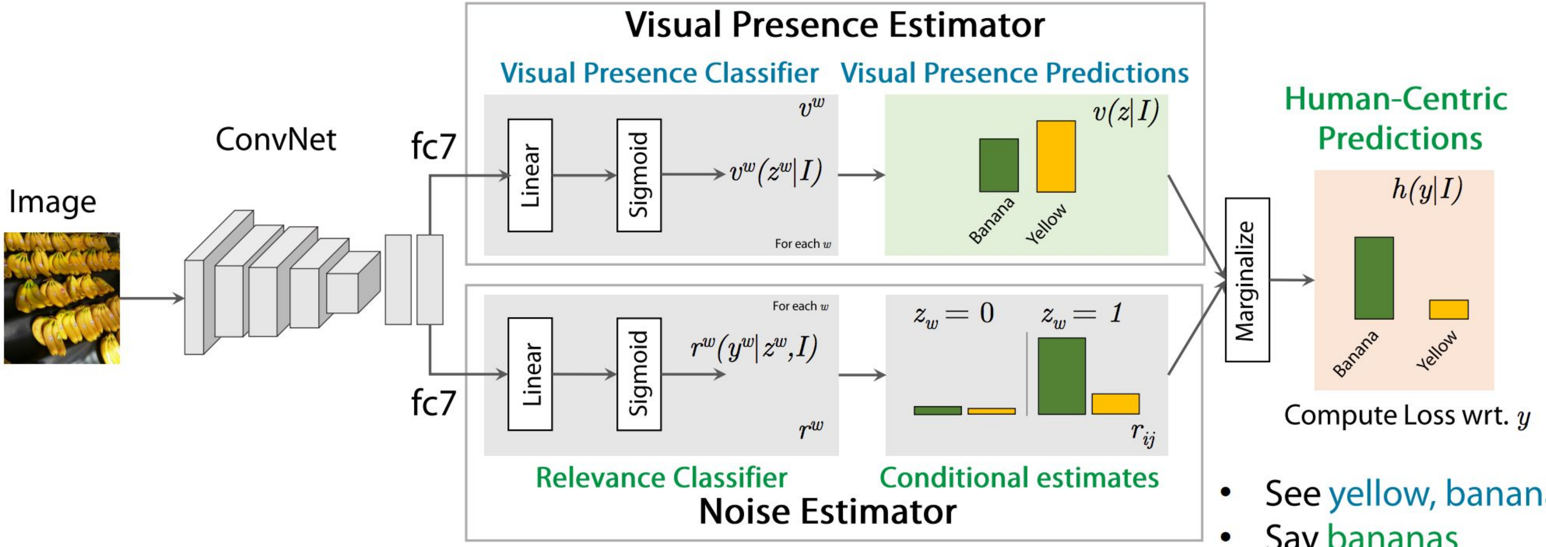
Given visual presence, is concept relevant?

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-biased**)	y	h

End-to-End Approach



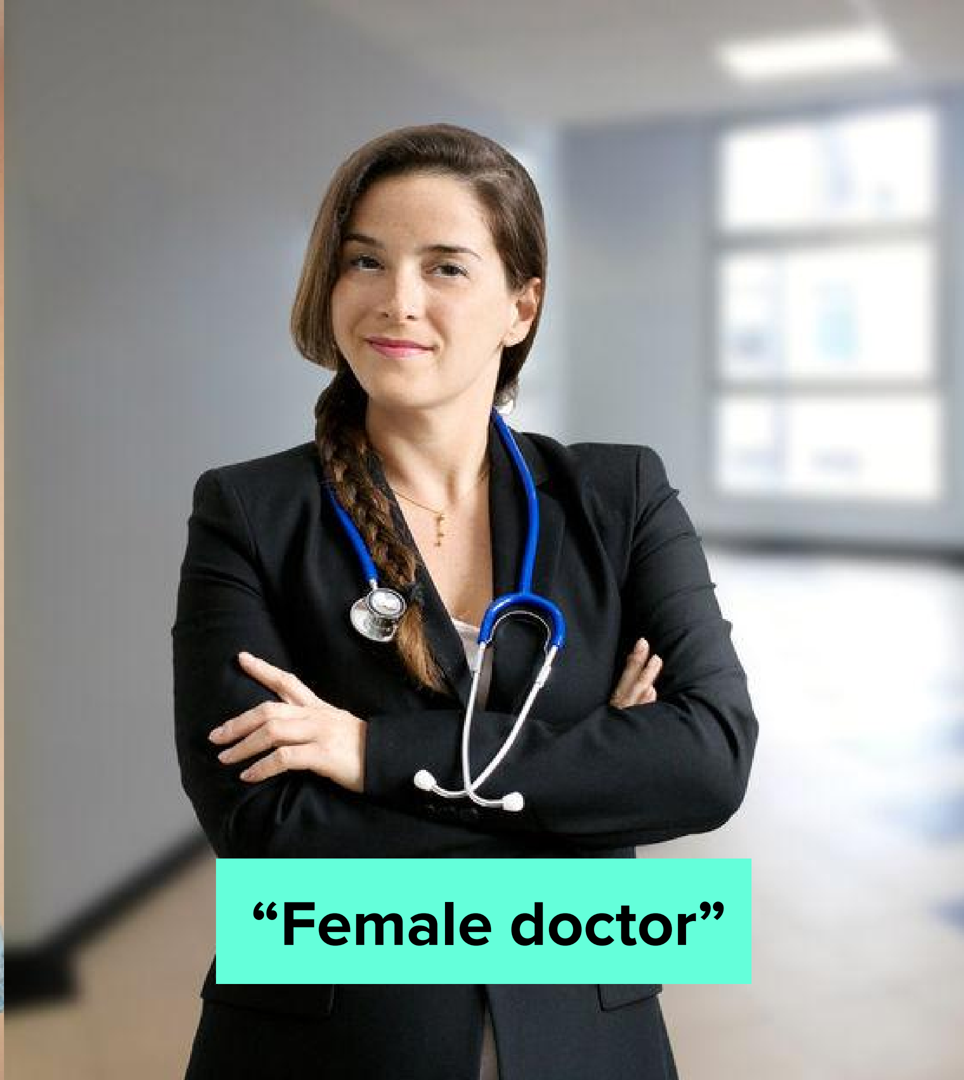
- See yellow, bananas
- Say bananas

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

SOURCE

Misra, Ishan; Girshick, Ross; Mitchell, Margaret; Zitnick, Larry (2016). "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels". Proceedings of CVPR.


“Male doctor”


“Female doctor”

Thanks!

mmitchellai@

ML Fairness

Machine Learning, Subgroup Discovery

go/ml-fairness-tools
go/ml-fairness-metrics