PyData Delhi 2018

Social Bias in Machine Learning

Janu Verma

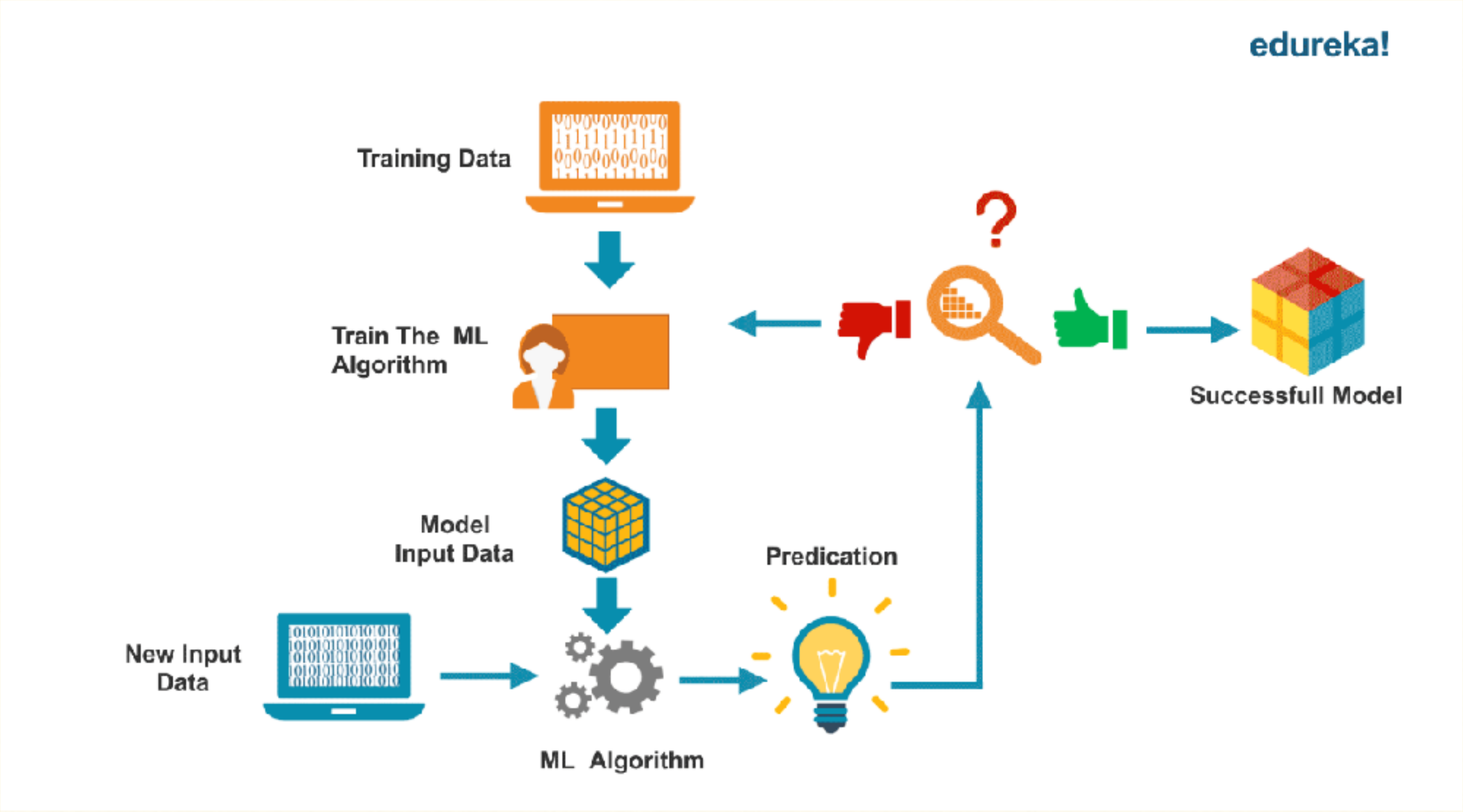
@januverma

About Me

- Data Scientist at Hike, New Delhi.
- Previously -
    - IBM Research, New York
    - Cornell University
    - Kansas State University, Cambridge University
- Researcher in machine learning , data visualization, and mathematics.

Machine Learning

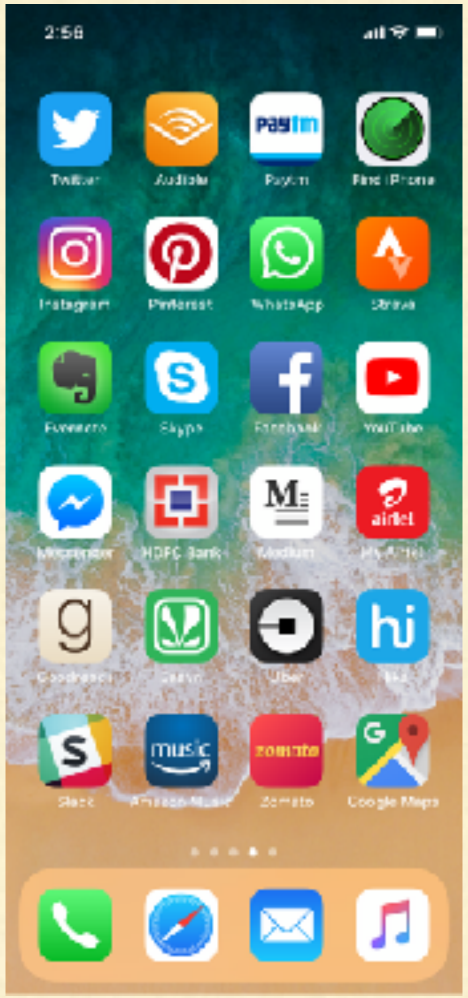
- **Machine learning** is a field of computer science that designs systems with the ability to automatically learn and improve from experience without being explicitly programmed.
- Computer systems that access data and use statistical/mathematical techniques to learn patterns and make inference based on probabilistic responses.
- *Supervised learning* involves providing example inputs and respective outputs to the the program, which 'learns' to make predictions on the outputs for new, unseen inputs.
- e.g. a classification system to categorize the images as cats or dogs.
-  

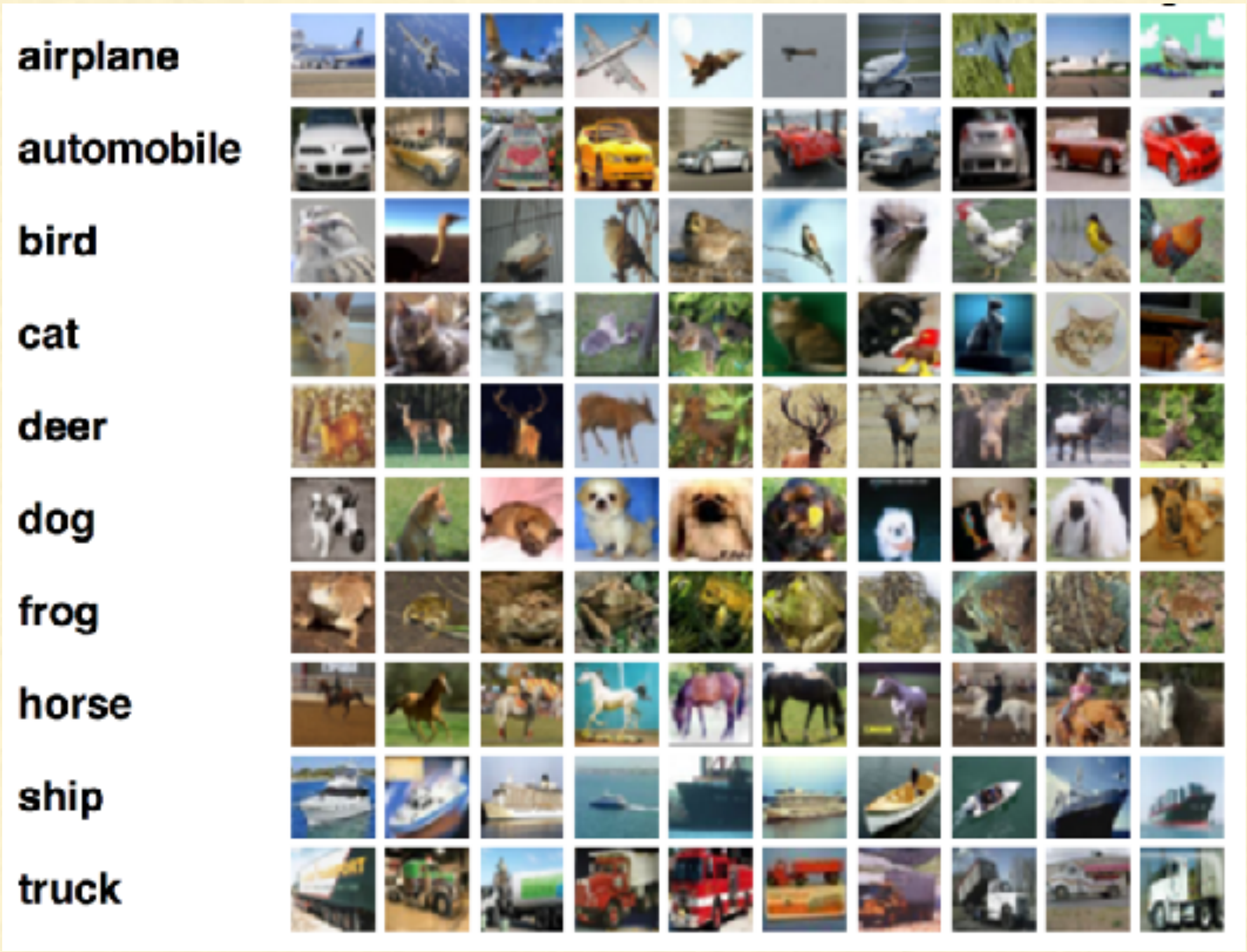


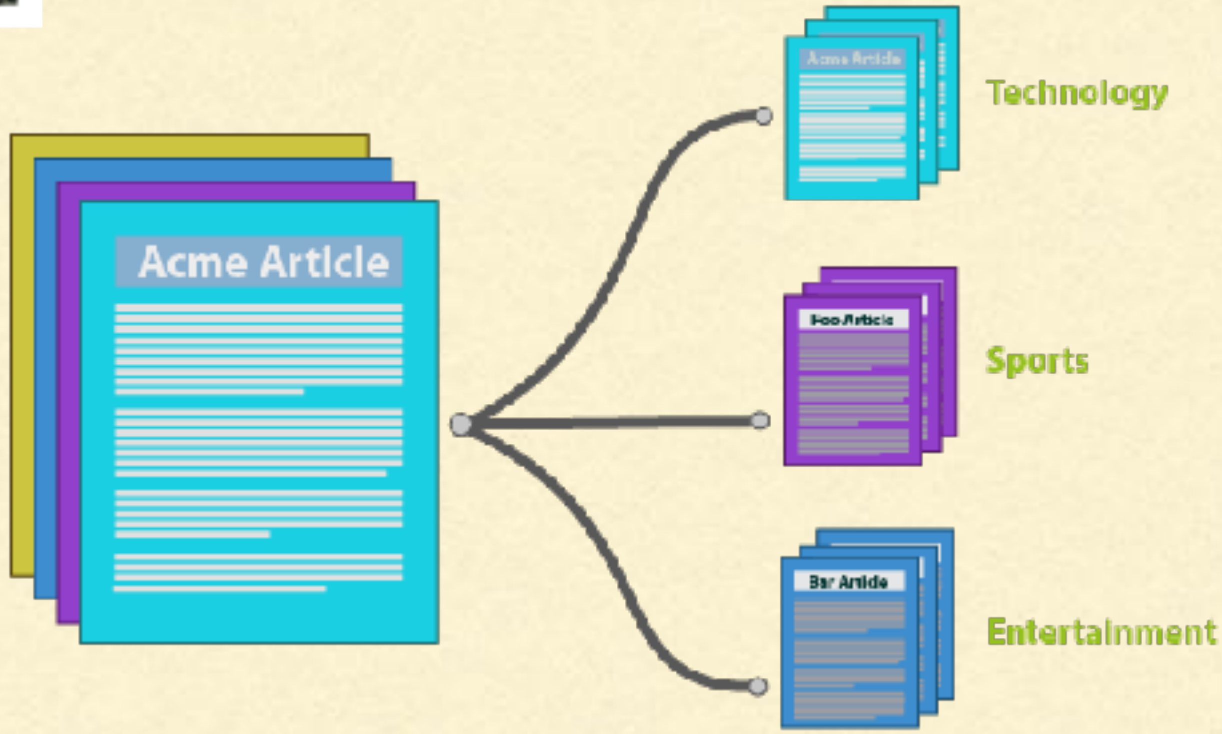
Machine Learning Pipeline

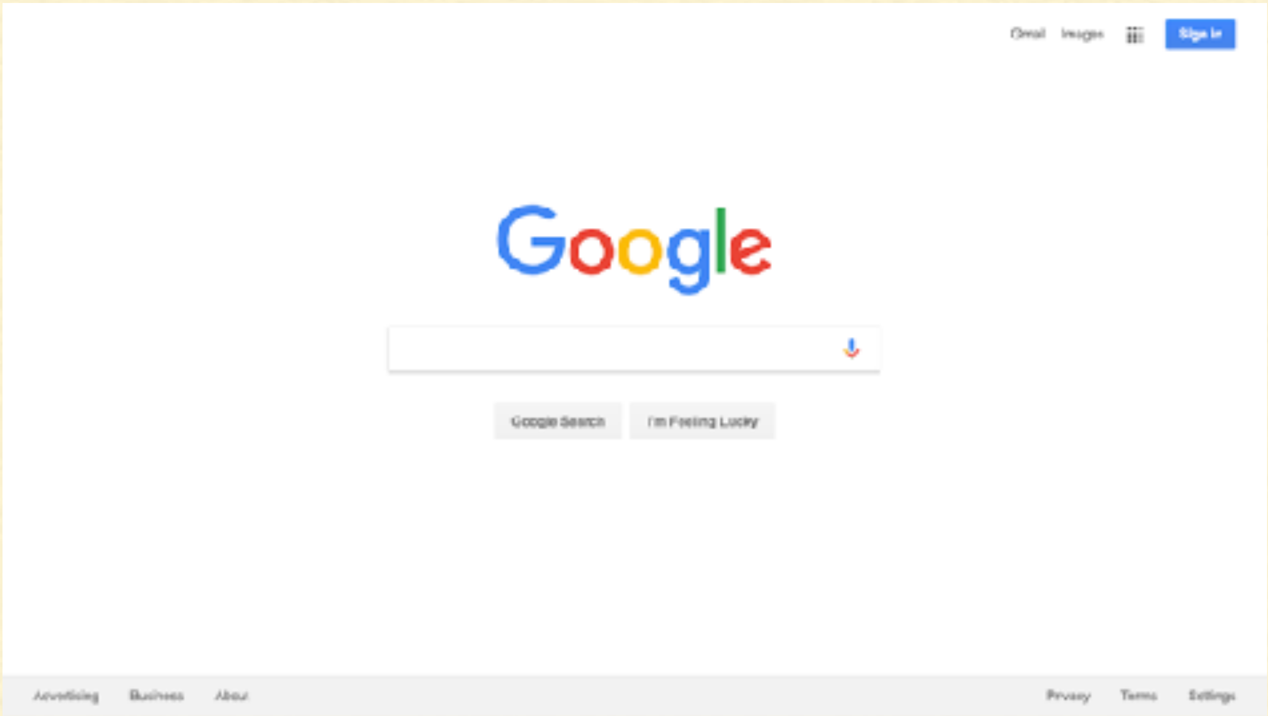
Machine Learning Is Amazing!!

- ML is a remarkable tech that has greatly transformed our lives.
- We use applications of machine learning everyday. Embedded in our cell phones.

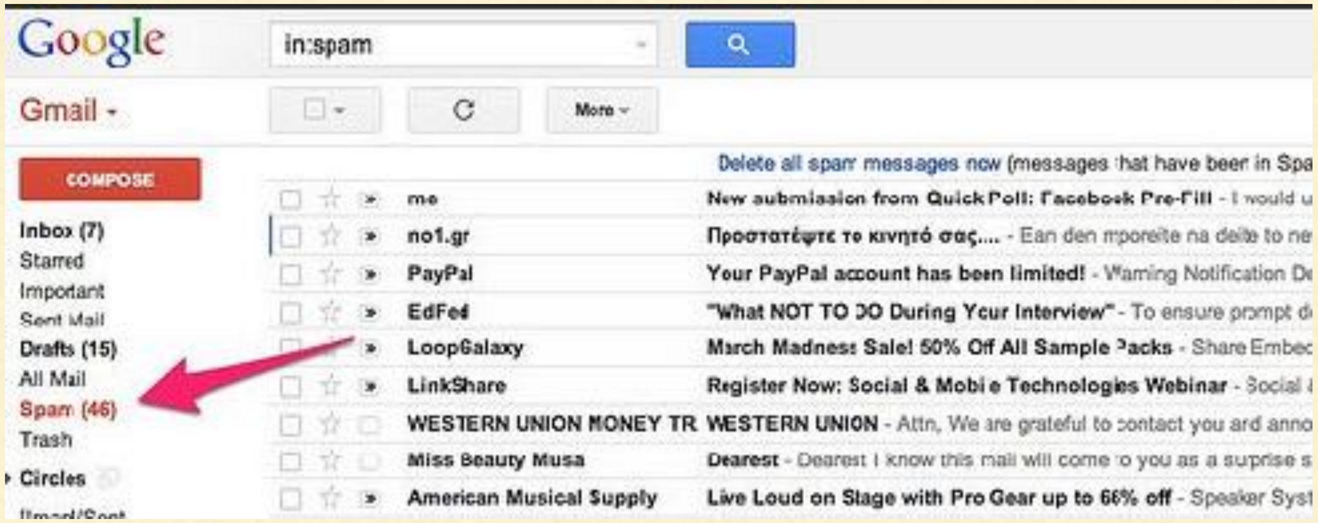



airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck


Acme Article
Technology
Sports
Entertainment

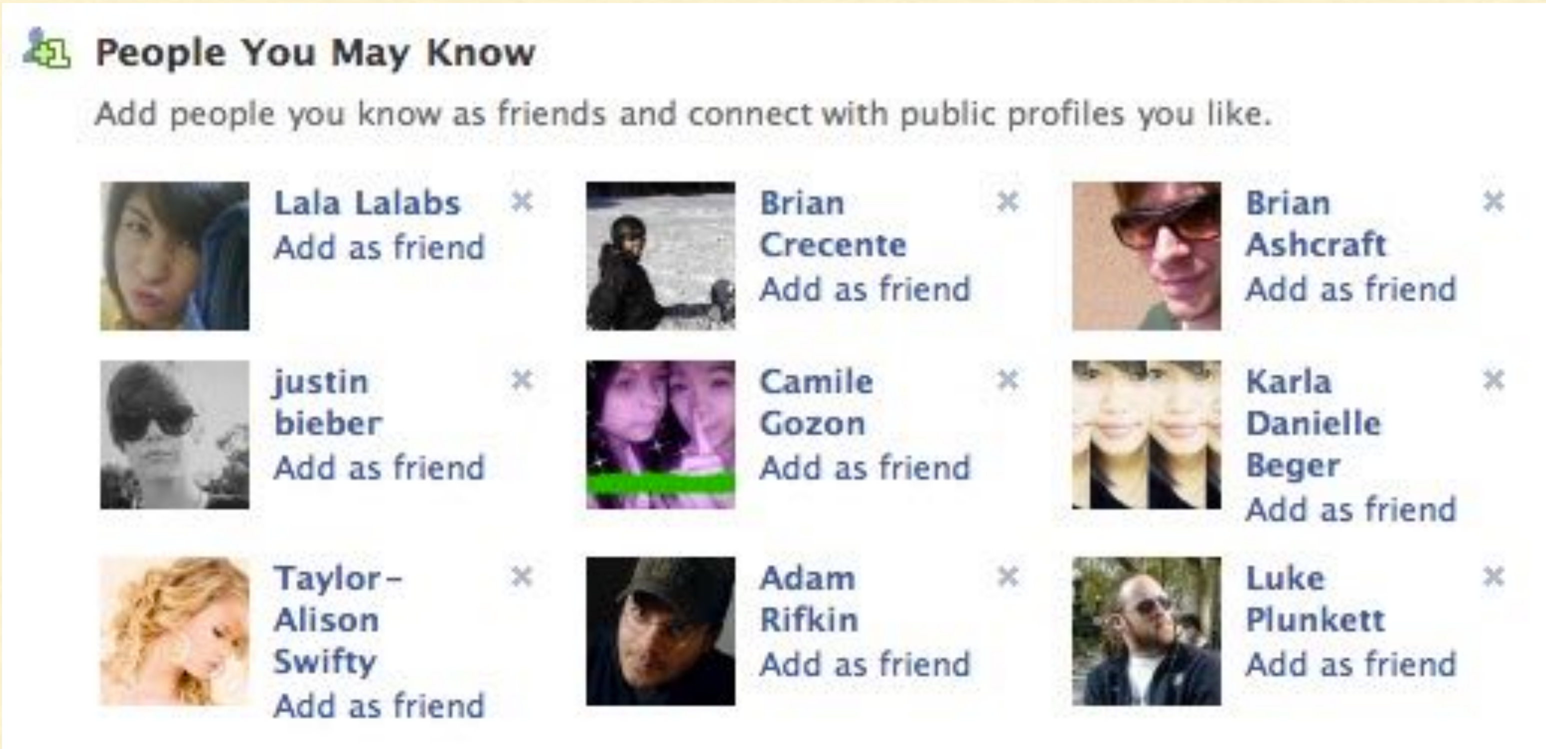

Google


Siri


Google
in:spam
Gmail
More
COMPOSE
Delete all spam messages now
Inbox (7)
Starred
Important
Drafts (15)
All Mail
Spam (46)
Trash
Circles
me
no1.gr
PayPal
EdFed
LoopGalaxy
LinkShare
WESTERN UNION MONEY TR
Miss Beauty Musa
American Musical Supply
Your PayPal account has been limited!
March Madness Sale! 50% Off All Sample Packs
Register Now: Social & Mobile Technologies Webinar
WESTERN UNION - Attn, We are grateful to contact you and anno
Dearest
Live Loud on Stage with Pro Gear up to 66% off

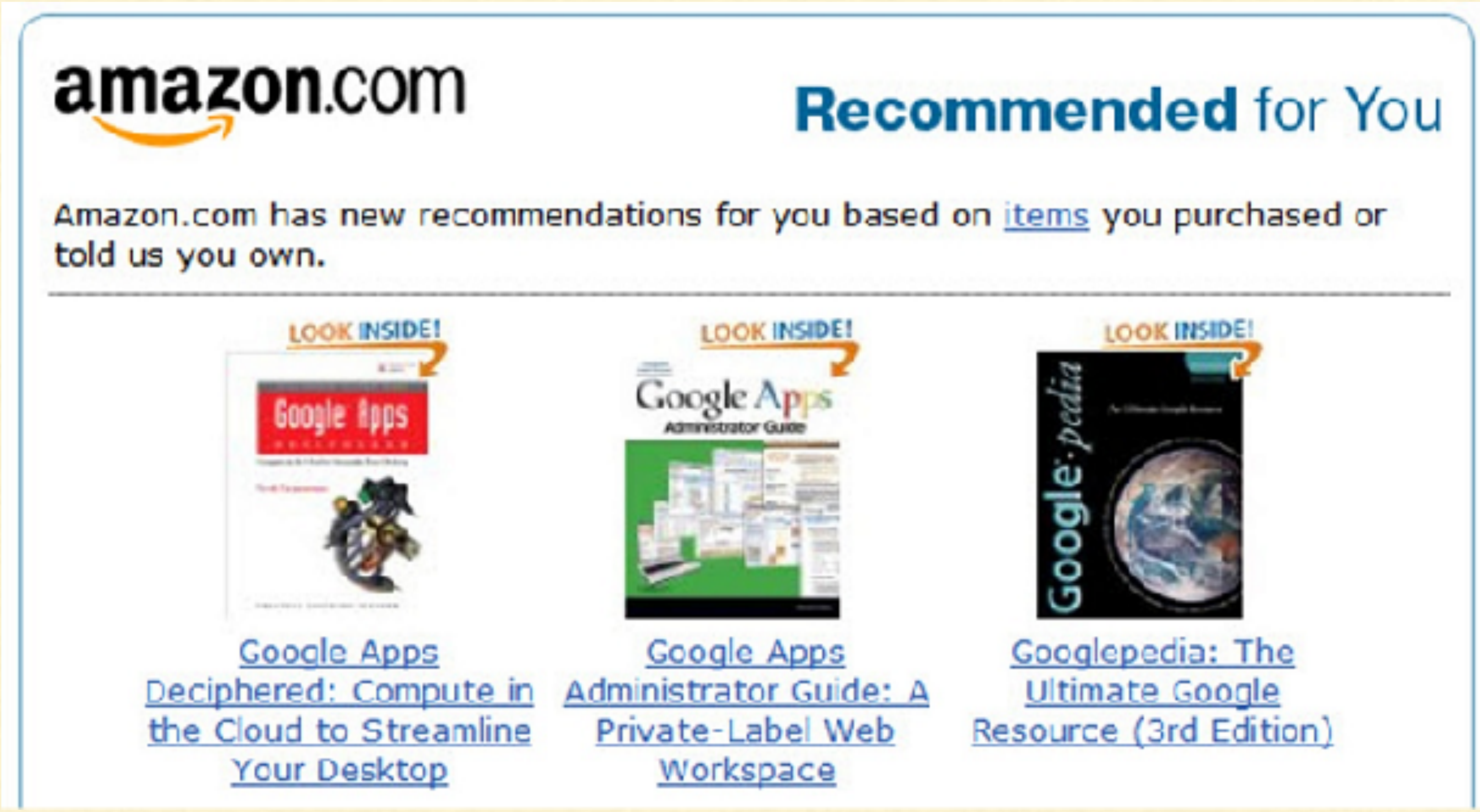

English
Hindi
Apple
सेब
seb
3 more translations


You
Tube


People You May Know
Add people you know as friends and connect with public profiles you like.
Lala Lalabs
Add as friend
Brian Crecente
Add as friend
Brian Ashcraft
Add as friend
justin bieber
Add as friend
Camile Gozon
Add as friend
Karla Danielle Beger
Add as friend
Taylor-Alison Swifty
Add as friend
Adam Rifkin
Add as friend
Luke Plunkett
Add as friend


Who to follow · Refresh · View all
Andrew Robb @AndrewRob...
Followed by EssentialView and ot...
Follow
ABC South East SA @abcsouthe...
Followed by LGA South Australia ...
Follow
ABC
Local
Ian Shuttleworth @Shutts10
Followed by Jason McConnell an...
Follow
Browse categories · Find friends


Nicholas Carlson


amazon.com
Recommended for You
Amazon.com has new recommendations for you based on items you purchased or told us you own.
LOOK INSIDE!
Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop
LOOK INSIDE!
Google Apps Administrator Guide: A Private-Label Web Workspace
LOOK INSIDE!
Googlepedia: The Ultimate Google Resource (3rd Edition)


House of Cards
2013 TV-MA 1 Season
Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.
NETFLIX
Because you watched Orange Is the New Black
HOUSE of CARDS
the L word
ALWAYS SUNNY PHILADELPHIA
New Girl
Because you watched Red Lights
ELECTRIC MIST

Social Bias

- Social Biases are a class of cognitive biases based on how we perceive other human beings. When we try to explain other's behaviour, based on faulty or unfounded preconceptions and prejudices

Social Bias

- Social Biases are a class of cognitive biases based on how we perceive other human beings. when we try to explain other’s behaviour, based on faulty or unfounded preconceptions and prejudices

- e.g. *group attribution error :* Making inferences about an entire group (ethnicity, caste, economic status, region, etc.) based on our observation of a small number of examples. Humans tend to do this even in the face of evidence to the contrary.

Social Bias

- Social Biases are a class of cognitive biases based on how we perceive other human beings. when we try to explain other's behaviour, based on faulty or unfounded preconceptions and prejudices

- e.g. *group attribution error :* Making inferences about an entire group (ethnicity, caste, economic status, region, etc.) based on our observation of a small number of examples. Humans tend to do this even in the face of evidence to the contrary.

- e.g. *just-world hypothesis:* trying to protect our desire for a fundamentally just world by blaming the victim, rather than facing the reality of the arbitrary nature of the incidents.

Social Bias

- Social Biases are a class of cognitive biases based on how we perceive other human beings. when we try to explain other's behaviour, based on faulty or unfounded preconceptions and prejudices

- e.g. *group attribution error :* Making inferences about an entire group (ethnicity, caste, economic status, region, etc.) based on our observation of a small number of examples. Humans tend to do this even in the face of evidence to the contrary.

- e.g. *just-world hypothesis:* trying to protect our desire for a fundamentally just world by blaming the victim, rather than facing the reality of the arbitrary nature of the incidents.

  Experiments have shown that when given a passage about interactions between a male and female that ended in the male raping the female, subjects often attempted to explain the ending as 'inevitable' given the preceding interactions, and to attribute the rape to the female's behaviour, rating her negatively on post-survey questions.

Social Bias

- Social Biases are a class of cognitive biases based on how we perceive other human beings. when we try to explain other's behaviour, based on faulty or unfounded preconceptions and prejudices

- e.g. *group attribution error :* Making inferences about an entire group (ethnicity, caste, economic status, region, etc.) based on our observation of a small number of examples. Humans tend to do this even in the face of evidence to the contrary.

- e.g. *just-world hypothesis:* trying to protect our desire for a fundamentally just world by blaming the victim, rather than facing the reality of the arbitrary nature of the incidents.

  Experiments have shown that when given a passage about interactions between a male and female that ended in the male raping the female, subjects often attempted to explain the ending as 'inevitable' given the preceding interactions, and to attribute the rape to the female's behaviour, rating her negatively on post-survey questions.

- **cf. Thinking Fast Slow - Dan Kahenman**

- We will focus mainly on racial and gender bias in this talk.

Bias In Machine Learning

- Machine learning systems are also prone to social bias and

Bias In Machine Learning

- Machine learning systems are also prone to social bias and
- Machine Learning systems now helps determine who is fired, hired, promoted, granted a loan or insurance, and even how long someone spends in prison.

Bias In Machine Learning

- Machine learning systems are also prone to social bias and
- Machine Learning systems now helps determine who is fired, hired, promoted, granted a loan or insurance, and even how long someone spends in prison.
- Spoiler: The bias in a ML system comes from human biases.

Bias In Machine Learning

- Machine learning systems are also prone to social bias and
- Machine Learning systems now helps determine who is fired, hired, promoted, granted a loan or insurance, and even how long someone spends in prison.
- Spoiler: The bias in a ML system comes from human biases.
- ML systems are not inherently neutral. They reflect the priorities, preferences, and prejudices - *the coded gaze* - of those who have the power to mould artificial intelligence.

Bias In Machine Learning

- Machine learning systems are also prone to social bias and
- Machine Learning systems now helps determine who is fired, hired, promoted, granted a loan or insurance, and even how long someone spends in prison.
- Spoiler: The bias in a ML system comes from human biases.
- ML systems are not inherently neutral. They reflect the priorities, preferences, and prejudices - *the coded gaze* - of those who have the power to mould artificial intelligence.
- More succinctly, the data we use to train ML models has inherent social biases.

Example

- Say we are designing a system that helps companies hire.

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.
- Such a system could (actually is!) be biased:

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.
- Such a system could (actually is!) be biased:
    - Premier institutes like IITs

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.
- Such a system could (actually is!) be biased:
  - Premier institutes like IITs
  - Male candidates

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.
- Such a system could (actually is!) be biased:
    - Premier institutes like IITs
    - Male candidates
    - Urban demography

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.
- Such a system could (actually is!) be biased:
    - Premier institutes like IITs
    - Male candidates
    - Urban demography
    - Economic standing

Example

- Say we are designing a system that helps companies hire.
- The system consumes data on successful hires and builds a model to predict the likelihood of a candidate to be successful.
- Such a system could (actually is!) be biased:
    - Premier institutes like IITs
    - Male candidates
    - Urban demography
    - Economic standing
- Women's participation in the labour force in India is currently at around 27%, is also declining. Thus any data will have an inherent gender bias.

ProPublica

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Prediction Fails Differently for Black Defendants

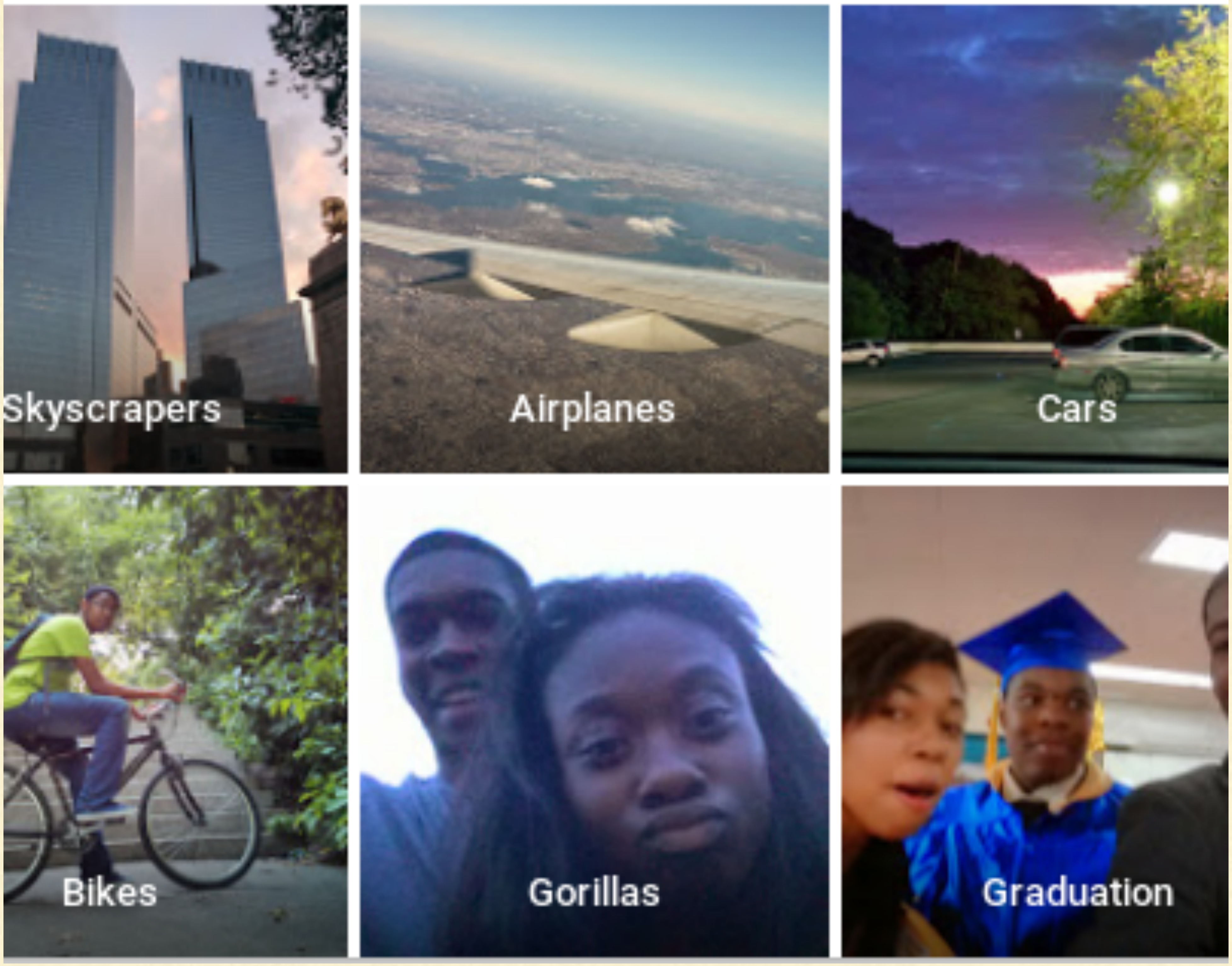
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Race was not a variable in the input data.
Race & gender are latently encoded in MANY other variables.

cf: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

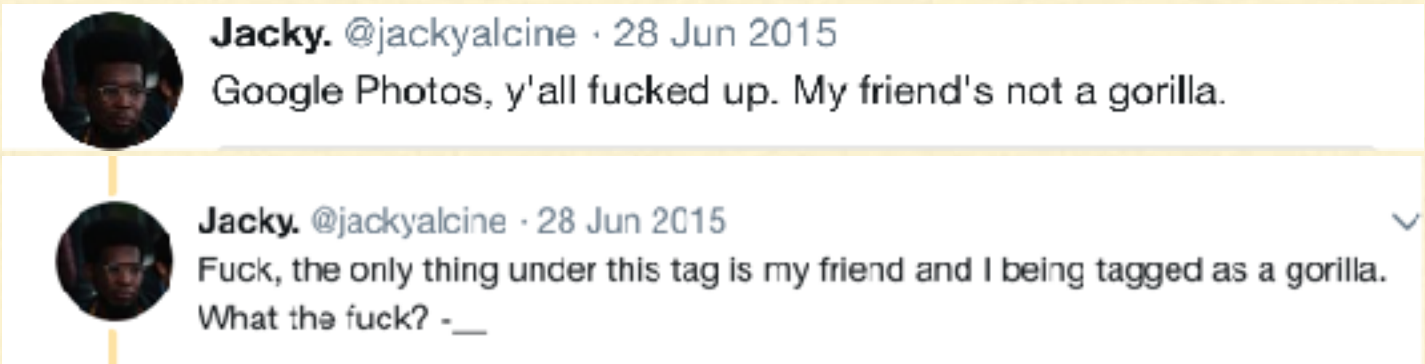
- Muslims, Dalits, and tribals make up 53% of all prisoners in India. (2014 survey)
- In some states, the percentage of Muslims in the incarcerated population was almost thrice the percentage of Muslims in the overall population (NCRB 2016).
- There will be similar statistics for other marginalized groups.

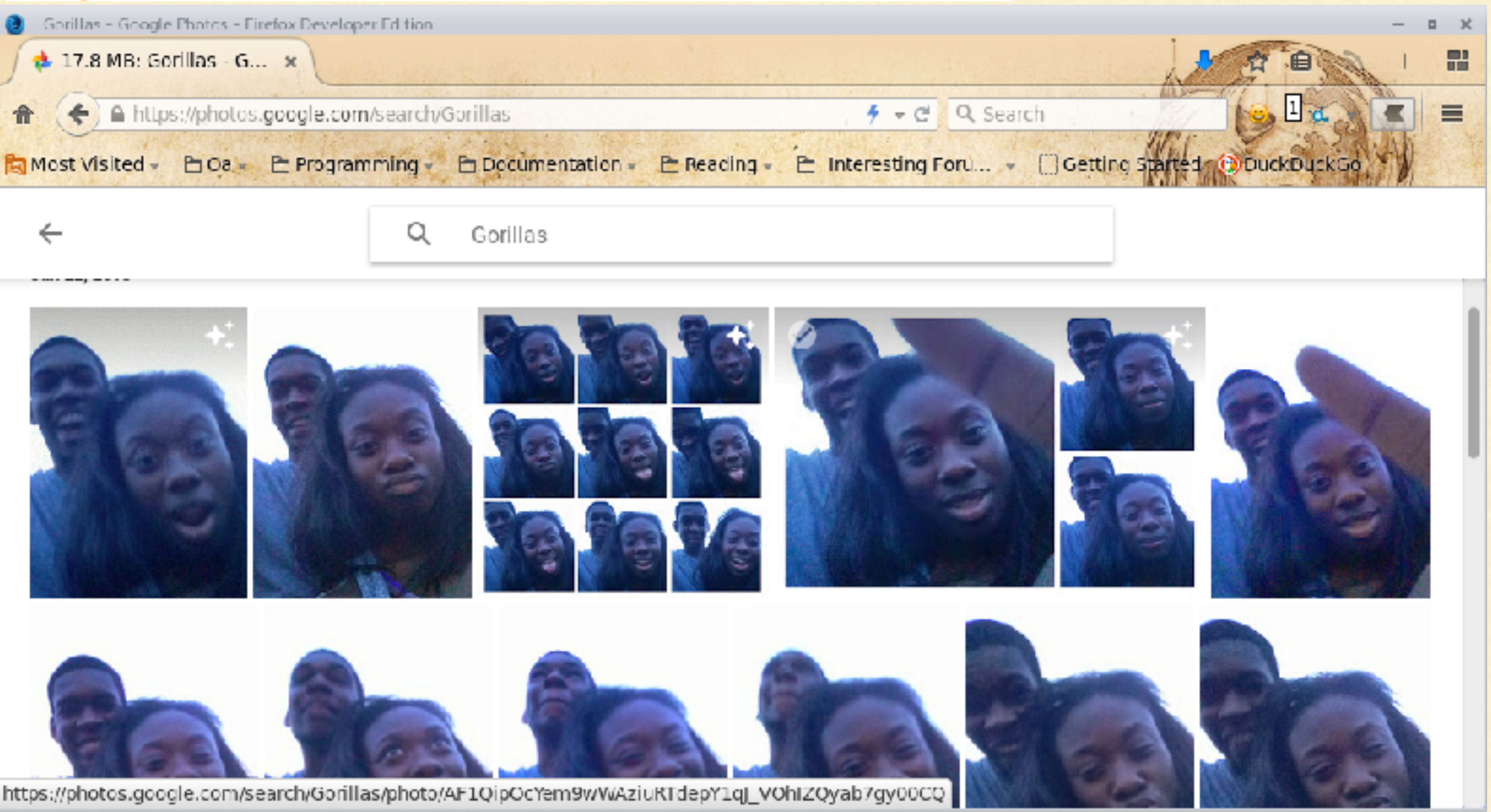
Bias in Image Understanding Software



Google Photos : Bias in Image Recognition

Programmer *Jacky Alcine* posted a series of tweets on this



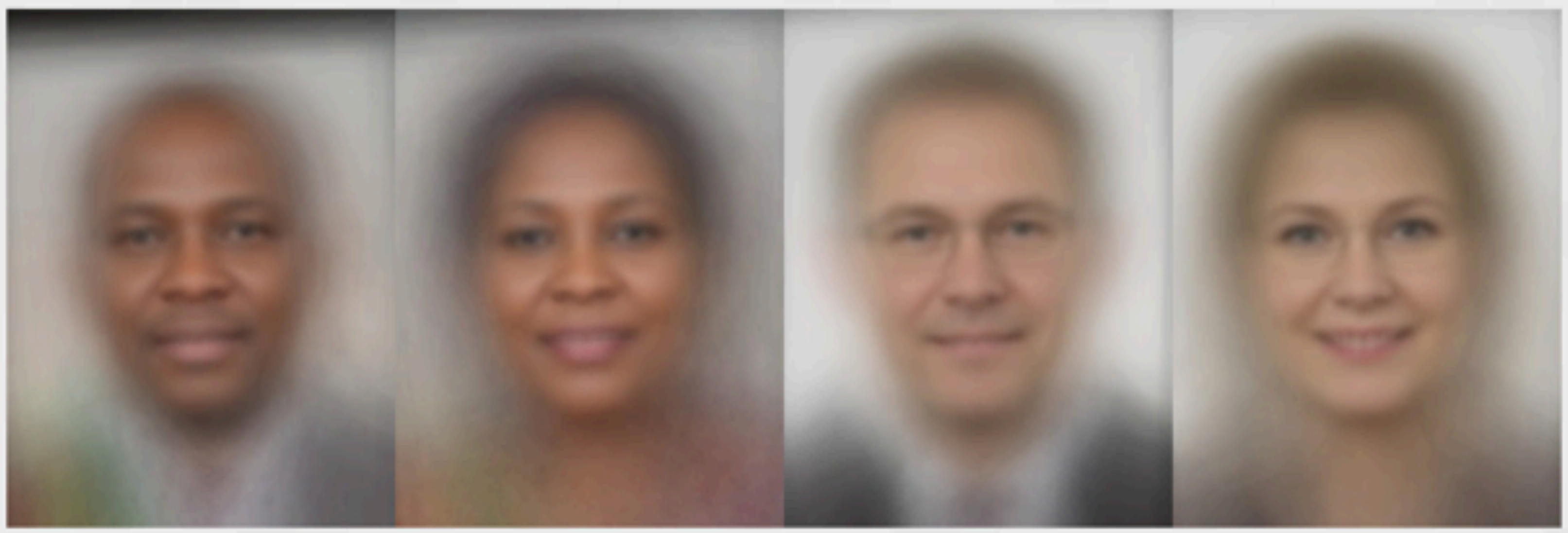


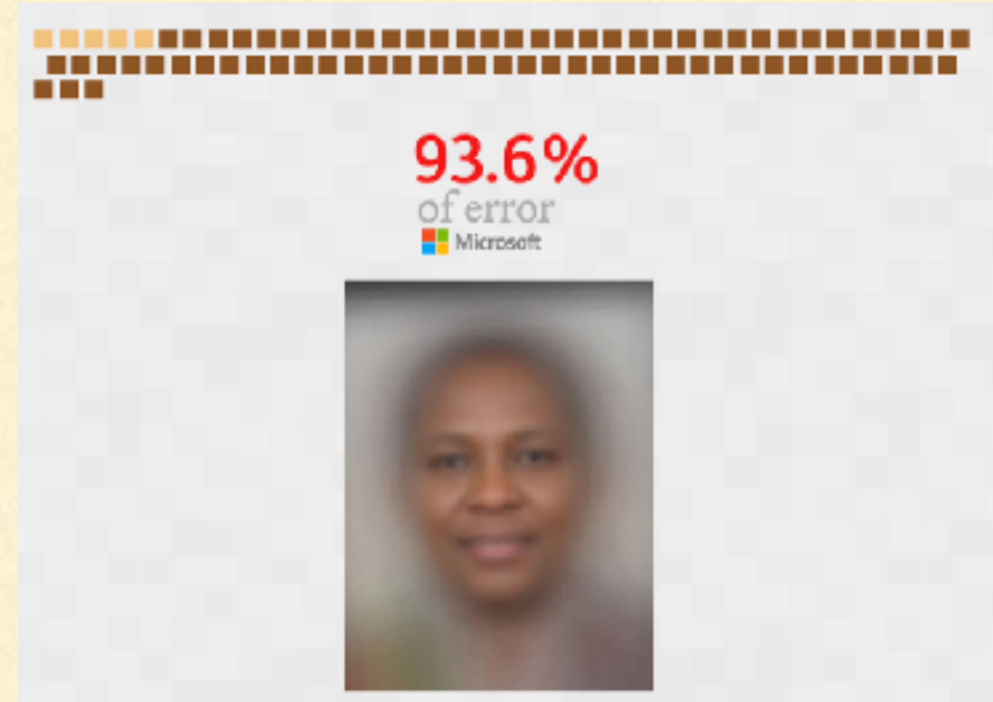
Google has removed the 'gorilla' tag from its new Photos app


FaceApp
FaceApp


Terrance AB Johnson
@tweeterrance
Follow
#faceapp isn't' just bad it's also racist... 🔥 filter=bleach my skin and make my nose your opinion of European. No thanks #uninstalled
11:38 AM - 19 Apr 2017

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%





gendershades.org

Inclusive (gender, skin type, ethnicity, age etc.) product testing and reporting are necessary if the industry is to create systems that work well for all of humanity

Further ethical considerations

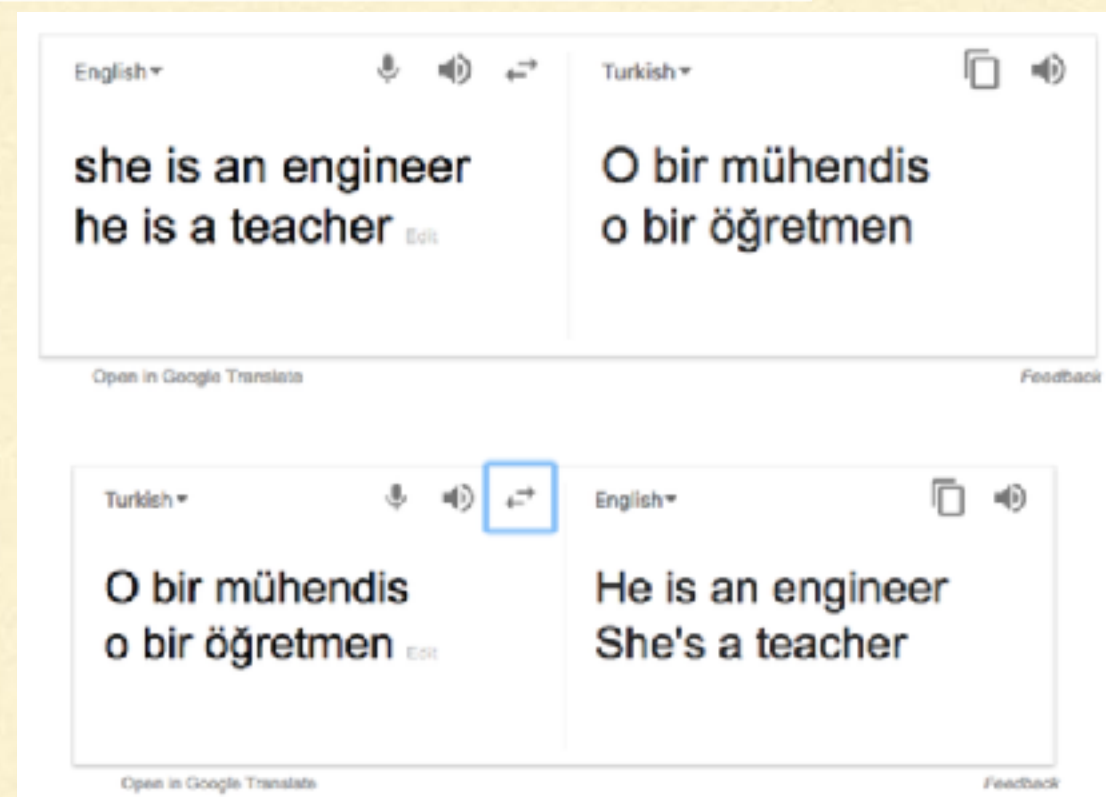
Figure 1: A man waits as his face is scanned at Logan Airport in Boston prior to boarding a flight to Aruba. (Photo: *Boston Globe*, all rights reserved)

China's Xinjiang surveillance is the dystopian future nobody wants

Monitoring tech pioneered in the region is spreading across China and the world.

Bias in Natural Language Understanding (NLU) software

GoogleTranslate



World Embedding

- Word embeddings are a representation of words in a natural language as vectors in a continuous vector space where semantically similar words are mapped to nearby points.
- Assumption: Words that appear in the same context are semantic closer than the words which do not share same context.
- Essentially, we 'embed' words in a vector space. And the weight of the word is distributed across many dimensions which capture the semantic properties of the words.
- Train a neural network on a large corpus of text data e.g. wikipedia dump.

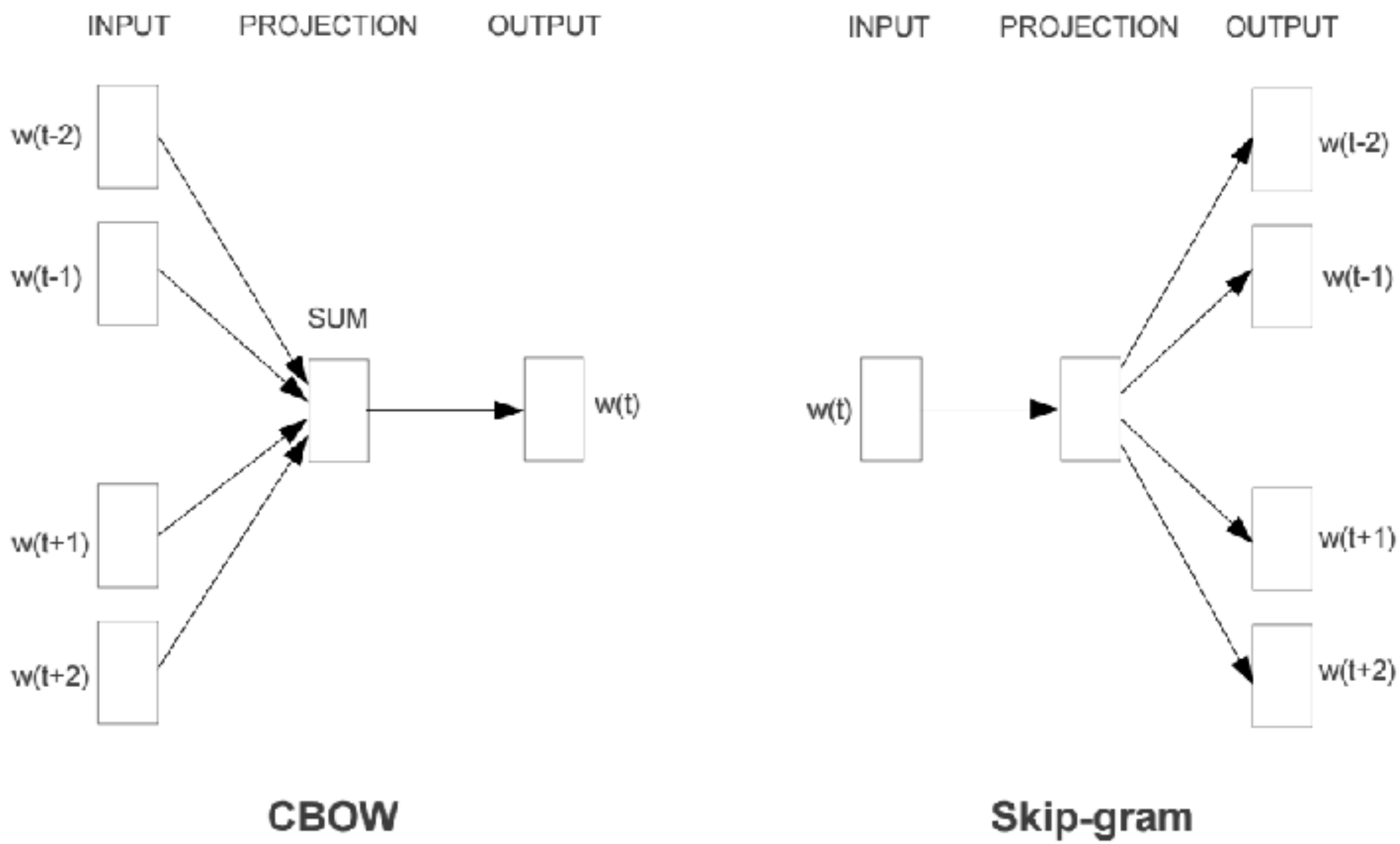


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

A *Somewhat surprisingly, it was found that similarity of word representations goes beyond simple syntactic regularities. Using a word offset technique where simple algebraic operations are performed on the word vectors, it was shown for example that vector("King") – vector("Man") + vector("Woman") results in a vector that is closest to the vector representation of the word Queen.*

–Mikolov et all, Google

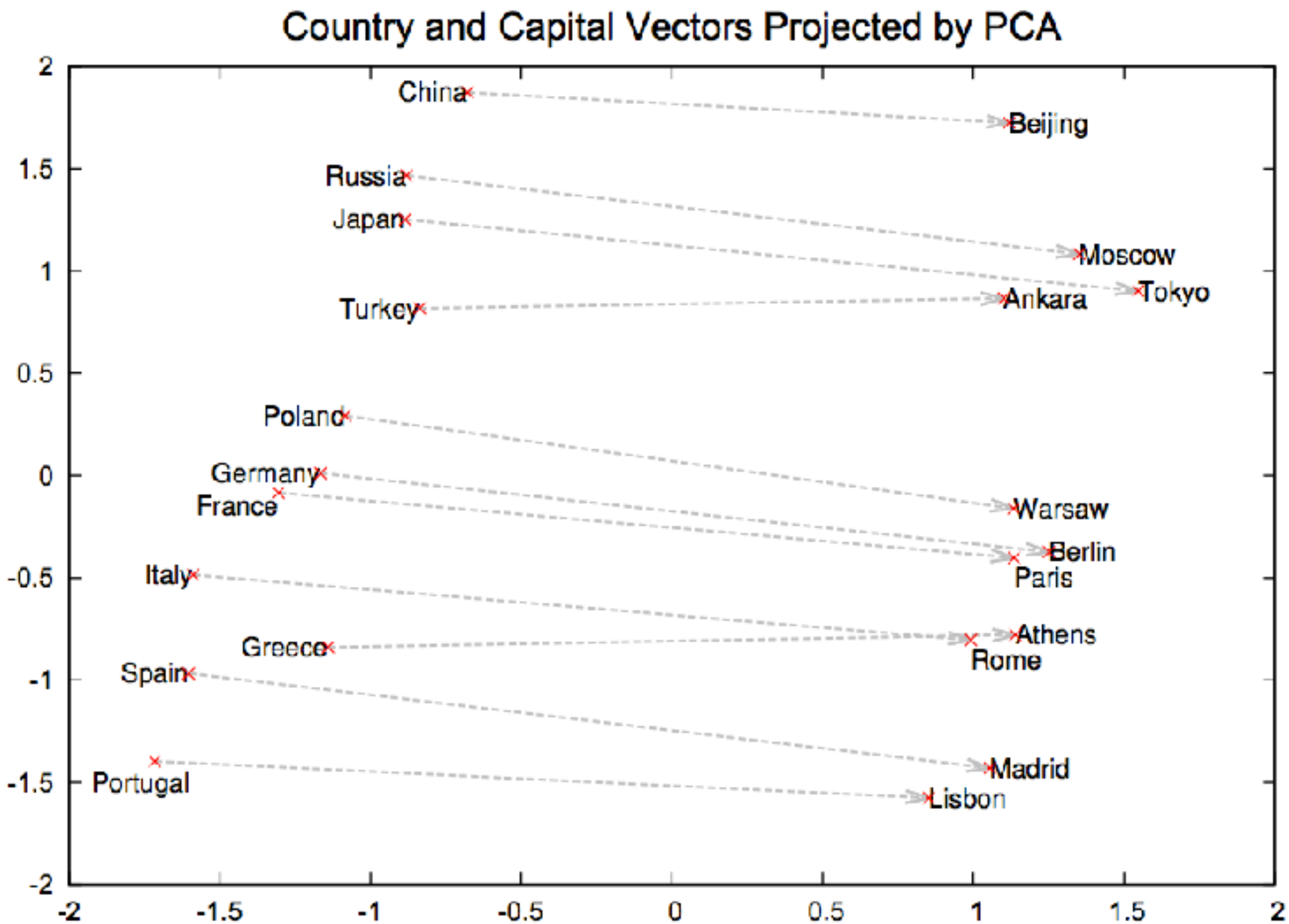


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Table 2: Examples of the analogical reasoning task for phrases (the full test set has 3218 examples). The goal is to compute the fourth phrase using the first three. Our best model achieved an accuracy of 72% on this dataset.

The distance between similar words is low:

```
dist(vecs[wordidx["puppy"]], vecs[wordidx["dog"]])
```

```
0.27636240676695256
```

```
dist(vecs[wordidx["queen"]], vecs[wordidx["princess"]])
```

```
0.20527545040329642
```

And the distance between unrelated words is high:

```
dist(vecs[wordidx["celebrity"]], vecs[wordidx["dusty"]])
```

```
0.98835787578057777
```

```
dist(vecs[wordidx["kitten"]], vecs[wordidx["airplane"]])
```

```
0.87298516557634254
```

Source: Rachel Thomas
@math_rachel

Bias

There is a lot of opportunity for bias:

```
In [20]: dist(vecs[wordidx["man"]], vecs[wordidx["genius"]])
Out[20]: 0.50985148631697985

In [21]: dist(vecs[wordidx["woman"]], vecs[wordidx["genius"]])
Out[21]: 0.6897833082810727
```

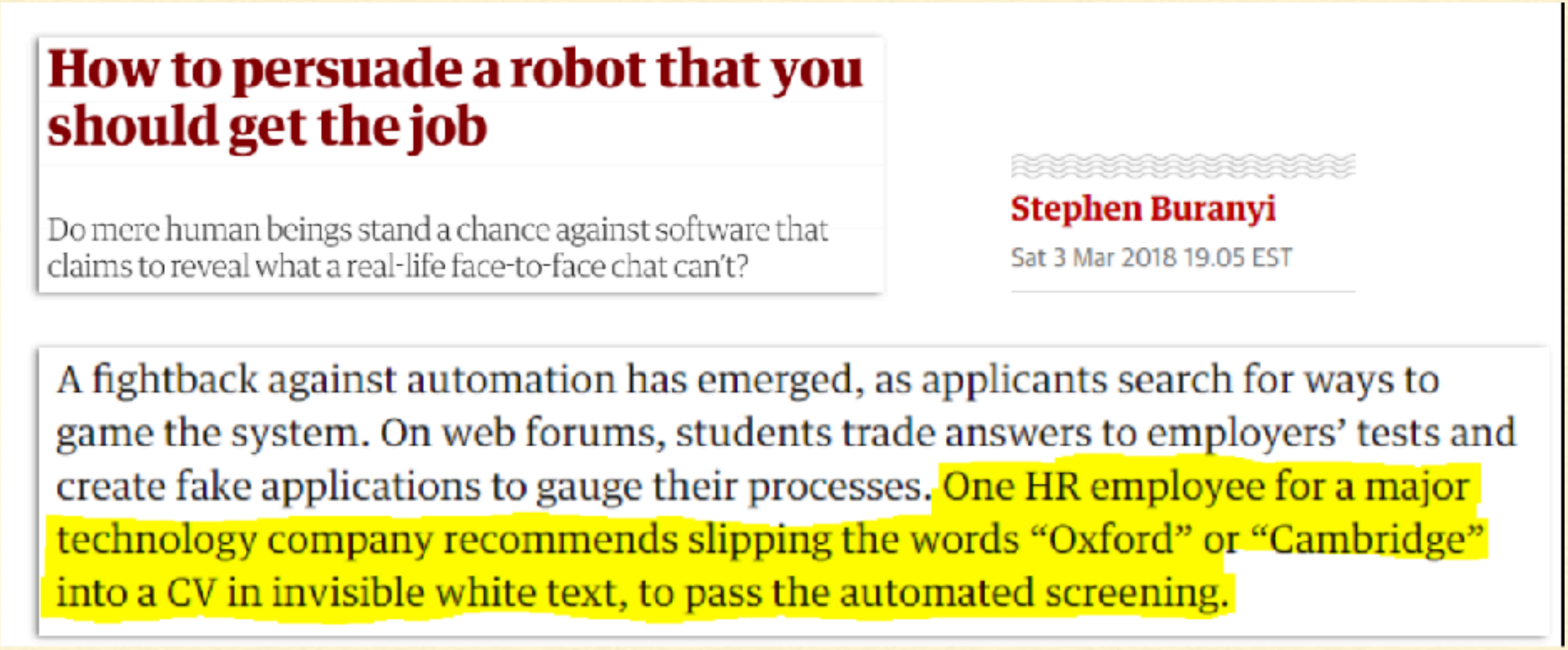
Source: Rachel Thomas
@math_rachel

Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan-Islam[1], Joanna J. Bryson[1,2], and Arvind Narayanan[1]

- Restaurant review app ranked Mexican restaurants lower, because word embeddings had negative connotations with “Mexican”.
- Word embeddings are used in web search engines. What if searching for “machine learning professor” more likely to return male names?

Gaming Machine Learning



How to persuade a robot that you should get the job

Do mere human beings stand a chance against software that claims to reveal what a real-life face-to-face chat can't?

Stephen Buranyi

Sat 3 Mar 2018 19.05 EST

A fightback against automation has emerged, as applicants search for ways to game the system. On web forums, students trade answers to employers' tests and create fake applications to gauge their processes. One HR employee for a major technology company recommends slipping the words "Oxford" or "Cambridge" into a CV in invisible white text, to pass the automated screening.

I’m Just An Engineer ?

“Once The Rockets Are Up, Who Cares Where They Come Down?”

Is a crime scene gang-related? A new computer program may have the answer. ISTOCK.COM/DENISTANGNEYJR

Artificial intelligence could identify gang crimes—and ignite an ethical firestorm

“I think that when you are building powerful things, you have some responsibility to at least consider how could this be used.”
- Blake Lemoine, Google

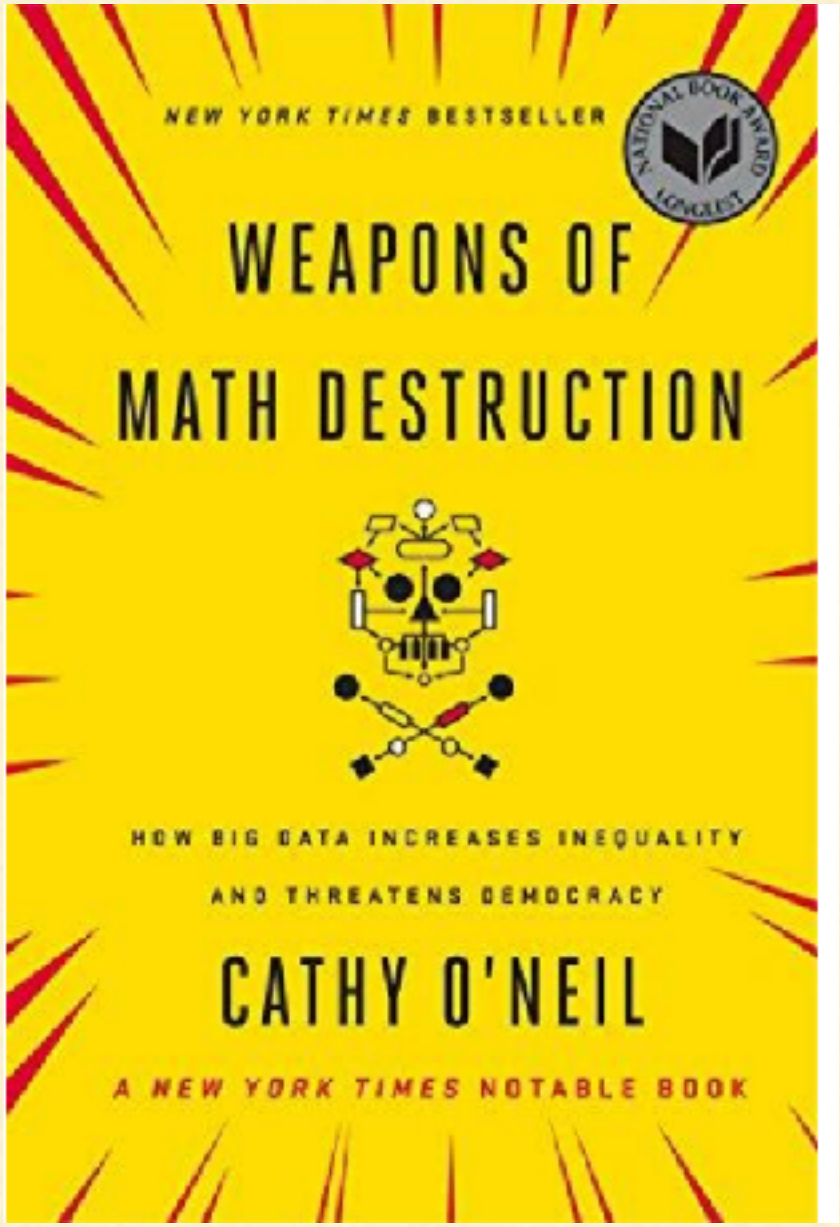
- possible unintended side effects.
- How could the team be sure the training data were not biased to begin with?
- What happens when someone is mislabeled as a gang member?
- The program could do the opposite by eroding trust in communities.
- Predictions could be no better than officers’ intuitions.

Model Harms

Kate Crawford,
Director AI Now,
NYU and MSR
in NIPS 2017 talk

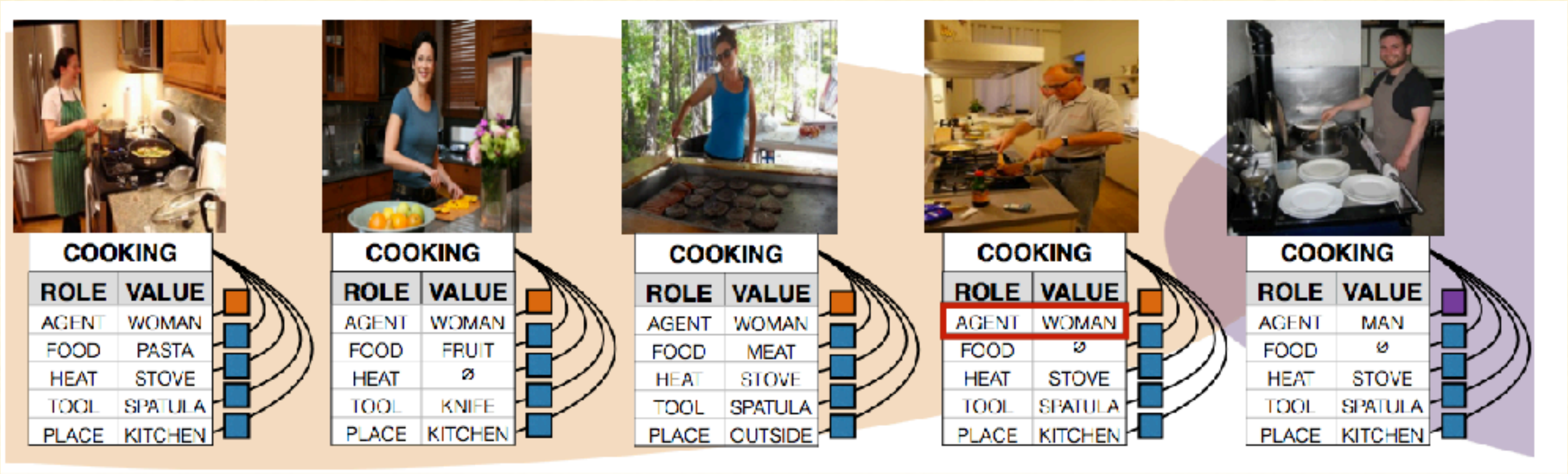
- Allocative harms: resources are allocated unfairly or withheld (transactional, quantifiable).
- Representative harms: systems reinforce subordination/perceived inferiority of some groups (cultural, diffuse, can lead to other types of harm)
    - Stereotypeping
    - Under-representation
    - Recognition

WMD



- WMD is a model which is:
    - Opaque - inscrutable “black box” (often by design).
    - Scalable - capable of exponentially increasing the number of people impacted.
    - Damaging - can ruin people’s lives and livelihoods.

Machine Learning can Amplify the Bias



Training data : 67% of people cooking are women.
Trained model prediction: 84% of people cooking are women.

Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints
https://arxiv.org/abs/1707.09457

What keeps people glued to YouTube? Its algorithm seems to have concluded that people are drawn to content that is more extreme than what they started with — or to incendiary content in general.

The Wall Street Journal conducted an investigation of YouTube content with the help of Mr. Chaslot. It found that YouTube often "fed far-right or far-left videos to users who watched relatively mainstream news sources," and that such extremist tendencies were evident with a wide variety of material. If you searched for information on the flu vaccine, you were recommended anti-vaccination conspiracy videos.

It is also possible that YouTube's recommender algorithm has a bias toward inflammatory content. In the run-up to the 2016 election, Mr. Chaslot created a program to keep track of YouTube's most recommended videos as well as its patterns of recommendations. He discovered that whether you started with a pro-Clinton or pro-Trump video on YouTube, you were many times more likely to end up with a pro-Trump video recommended.

What we are witnessing is the computational exploitation of a natural human desire: to look "behind the curtain," to dig deeper into something that engages us. As we click and click, we are carried along by the exciting sensation of uncovering more secrets and deeper truths. YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales.

YouTube, the Great Radicalizer

Zeynep Tufekci MARCH 10, 2018

Zeynep Tufekci
@zeynep

Matthew Curley
@Curley_Synergy
Follow

Replying to @zeynep

Was literally watching a short video with my daughters on Nelson Mandela yesterday and the next video reccomendation was one where the black people in South Africa are the true racists and criminals. (Don't want to say name of the trashy vid and give it any more visibility)

12:13 PM - 11 Mar 2018

4 Retweets 31 Likes

For Data Scientists

- *The data just reflect the biases in the world* — Can we (as data scientists) just leave it at that:
  - These models greatly affect lives of citizens - from hiring, firing, promotion, financial loans, healthcare, social interactions, imprisonment is steadily going dependent on ML.
  - ‘Blindness’ is not enough.
  - Experiments show that the ML systems can also amplify the biases.
  - Democracies are being undermined due to uncontrolled ML systems.
  - Promote ethical practices in the field.
  - Make world a better place by ethically and properly using ML systems.

Possible Solutions

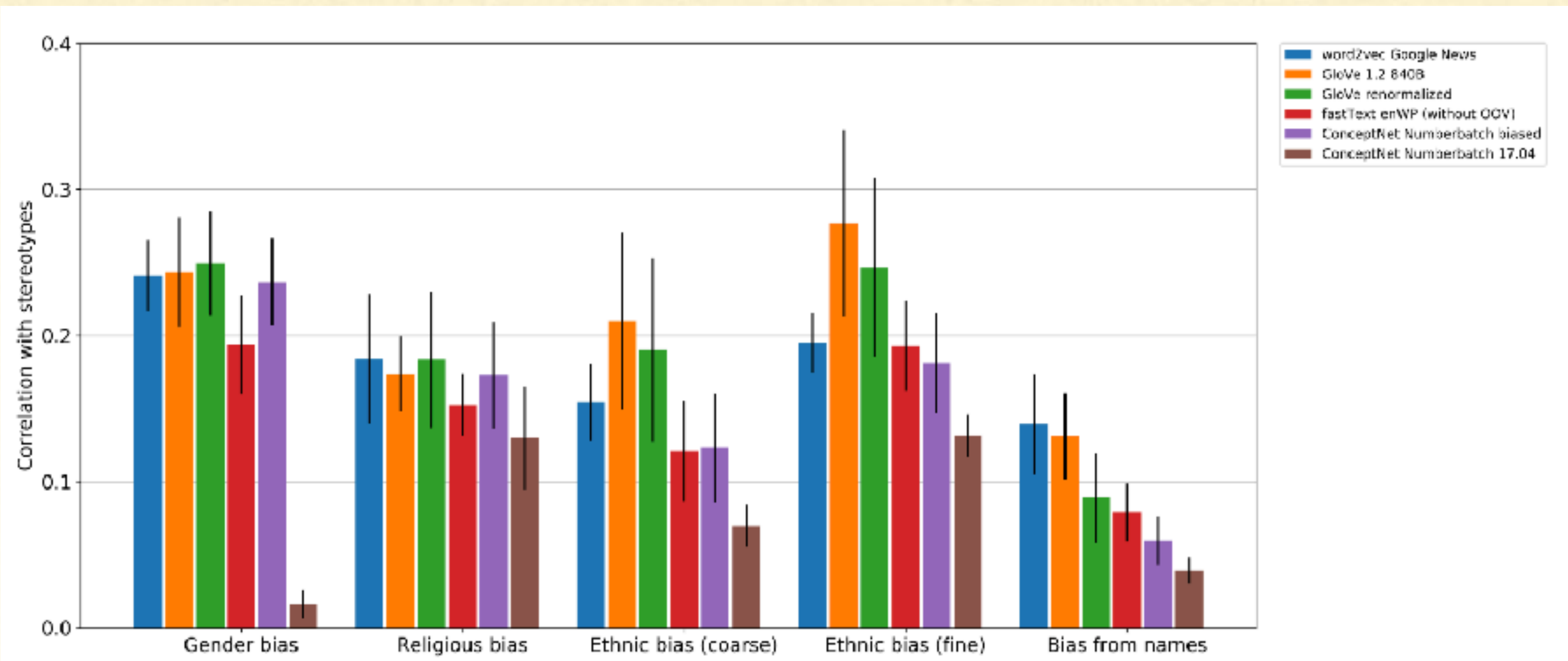
- Actively lookout for bias and find ways to address it. Awareness is better that blindness.
    - e.g. de-bias word embeddings
- More inclusive data collection and usage.
    - e.g. representative of all races, regions
- Think about possible unintended consequences
    - Can authoritarian governments use the system against citizens, trolls/harassers, propaganda/fake news.
- Seek help from domain experts.
    - Linguists can help in achieving more accurate and gender neutral translation.
- Even if you don’t use a feature in your algorithm, the output you get can still be correlated with that feature if the inputs are.
- Research shows diverse teams can help mitigate bias.

Quantifying and Reducing Stereotypes in Word Embeddings

Tolga Bolukbasi[1] TOLGAB@BU.EDU
Kai-Wei Chang[2] KW@KWCHANG.NET
James Zou[2] JAMESYZOU@GMAIL.COM
Venkatesh Saligrama[1] SRV@BU.EDU
Adam Kalai[2] ADAM.KALAI@MICROSOFT.COM
1 Boston University, 8 Saint Mary's Street, Boston, MA
2 Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors

Rob Speer — April 24, 2017



We are all responsible for understanding the systems (including data collection & implementation) our work is a part of and asking questions about ethics

Ask Questions

- What are the possible biases in the data set?
- Is the data open ? How was it collected ? Why was it collected ?
- Were there any methods used to cure for mistakes in data curation?
- Is ML necessary here ?
- What’s the accuracy (metrics) for different subgroups ?
- Do we need a human in the loop?
- What are the consequences of model failure ?
- Is it ethical to build such a model ?

Motivation for Dataset Creation

Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset?

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances are there? (of each type, if appropriate)?

Who was involved in the data collection process? (e.g., students, crowdworkers) and how were they compensated (e.g., how much were crowdworkers paid)?

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame of the instances?

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw
text
resp
dat
for
dat

Do
it a
a la

If t
tior
isti
Is t
ogr
mo
pos

Is t
wh
sta

Datasheets for Datasets*

Timnit Gebru[1], Jamie Morgenstern[2], Briana Vecchione[3], Jennifer Wortman Vaughan[1], Hanna Wallach[1], Hal Daumé III[1,4], and Kate Crawford[1,5]

When will the dataset be released/first distributed?

What license (if any) is it distributed under? Are there any copyrights on the data?

Are there any fees or access/export restrictions?

Any other comments?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

Will the dataset be updated? If so, how often and by whom?

How will updates be communicated? (e.g., mailing list, GitHub)

Is there an erratum?

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

If it relates to people, were they told what the dataset would be used for and did they consent? If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

Machine Learning In India

- ML is growing in India, companies are progressively adapting this technology.
- Research in Universities is also moving at a remarkable pace.
- Our tasks like transportation, banking,
- GoI has just established Artificial Intelligence Task Force for bringing AI in our economic, political and legal procedures.
- Data is being collected by financial institutions, healthcare providers, Govt. Surveys etc. which will facilitate production of stronger ML models.
- It's the right time we also start thinking about bias and implications of ML systems and ethical considerations for building and deploying such systems.

References

- Making small culture changes - Julia Evans

  https://jvns.ca/blog/2017/04/16/making-small-culture-changes/

- ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors

Questions

Contact:

@januverma

http://jverma.github.io/