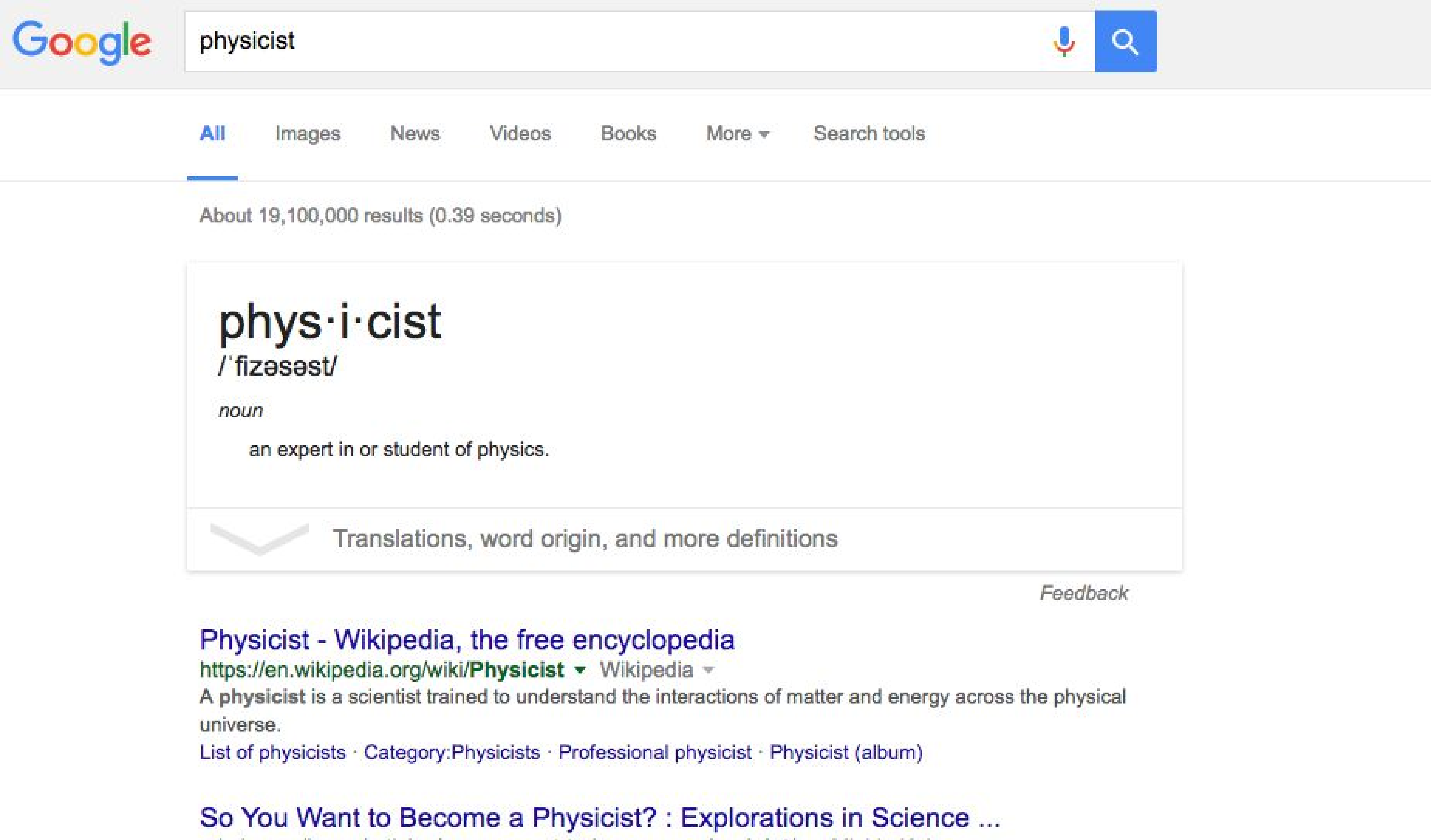
RMI Diversity Summit

Blaise Agüera y Arcas
Cerebra



About 19,100,000 results (0.39 seconds)

phys·i·cist

/ˈfizəsəst/

noun

an expert in or student of physics.

Translations, word origin, and more definitions

Feedback

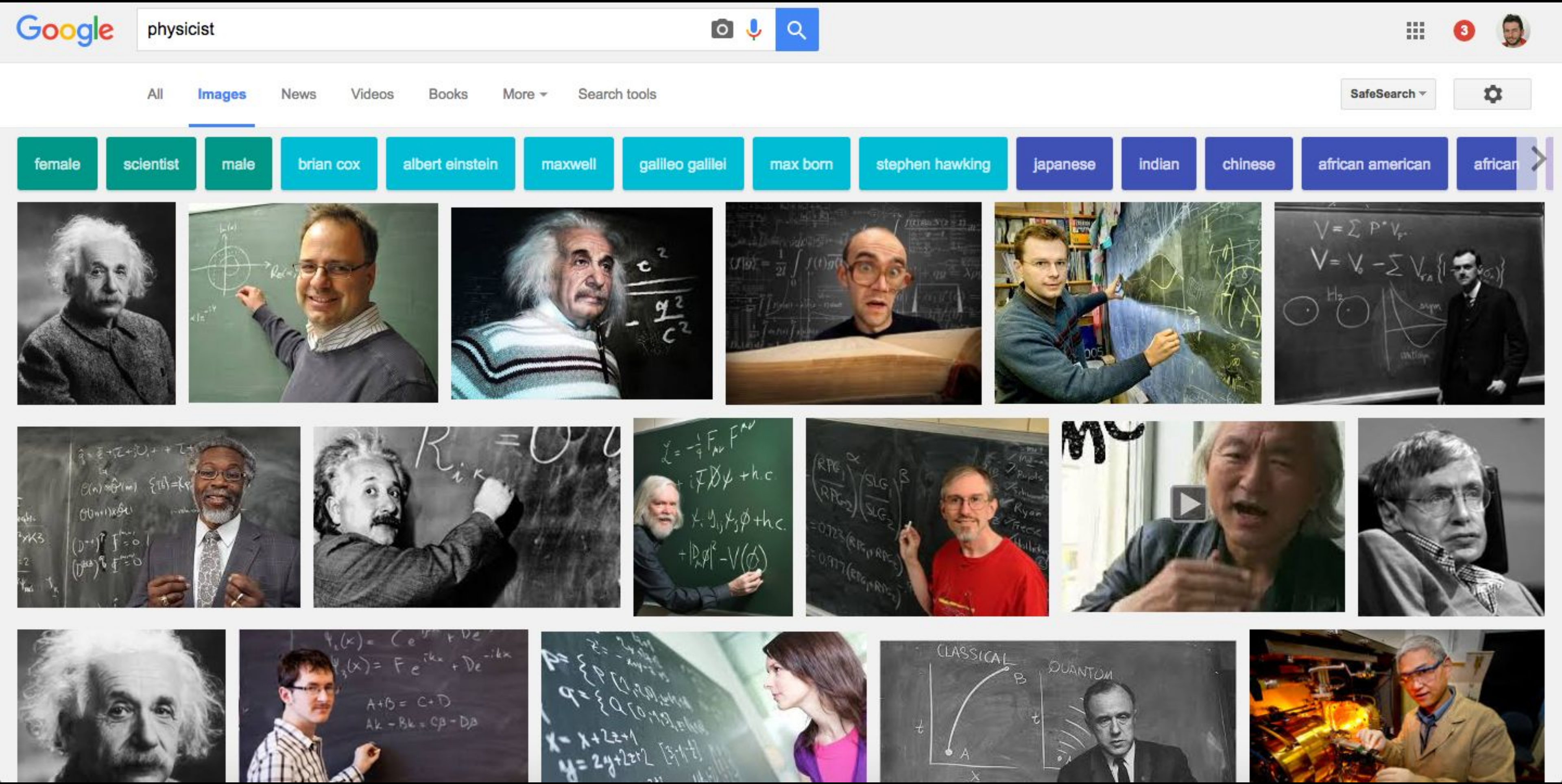
Physicist - Wikipedia, the free encyclopedia

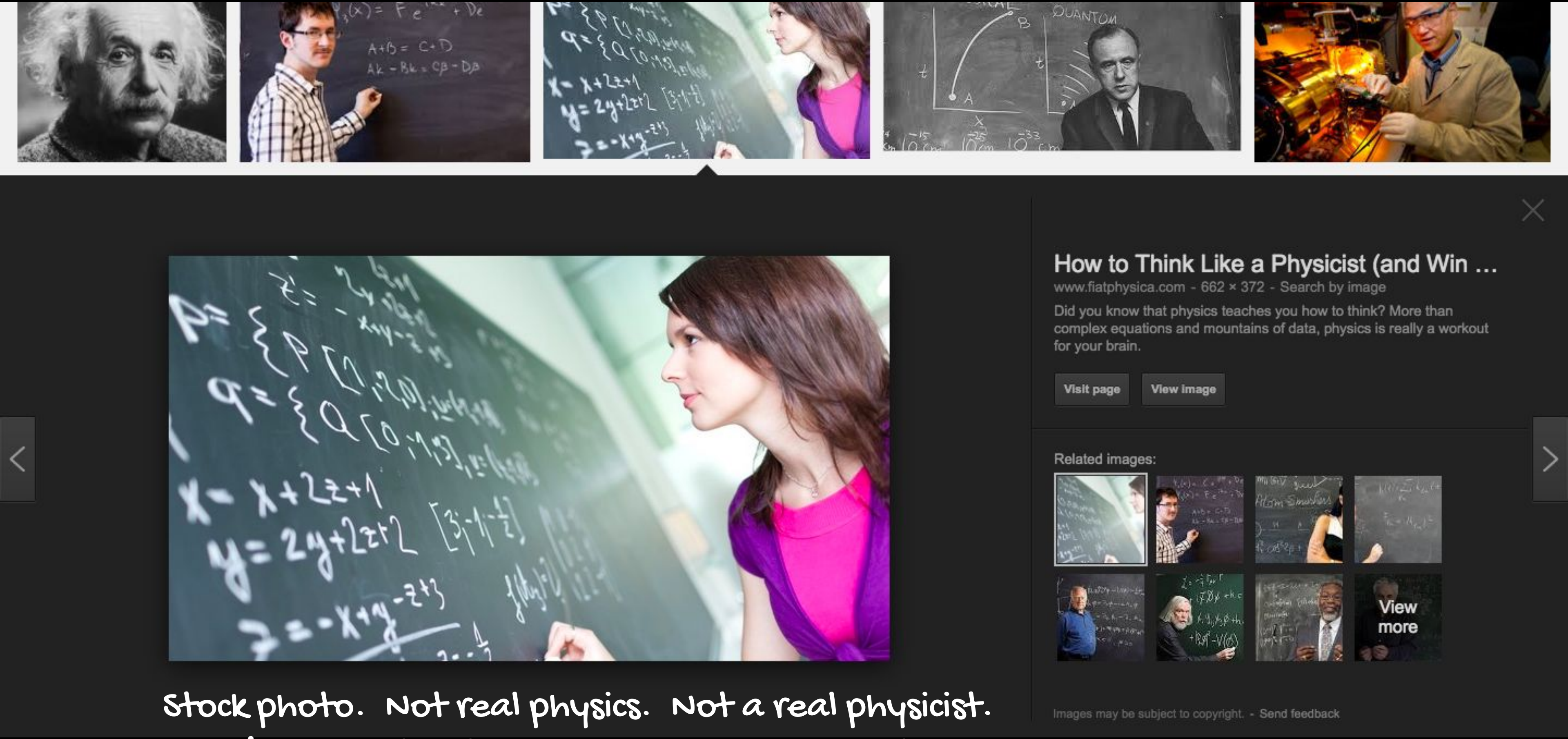
https://en.wikipedia.org/wiki/**Physicist** · Wikipedia

A **physicist** is a scientist trained to understand the interactions of matter and energy across the physical universe.

List of physicists · Category:Physicists · Professional physicist · Physicist (album)

So You Want to Become a Physicist? : Explorations in Science ...


Google
physicist
3
All
Images
News
Videos
Books
More
Search tools
SafeSearch
female
scientist
male
brian cox
albert einstein
maxwell
galileo galilei
max born
stephen hawking
japanese
indian
chinese
african american
african
CLASSICAL
QUANTUM



How to Think Like a Physicist (and Win ...

www.fiatphysica.com - 662 × 372 - Search by image

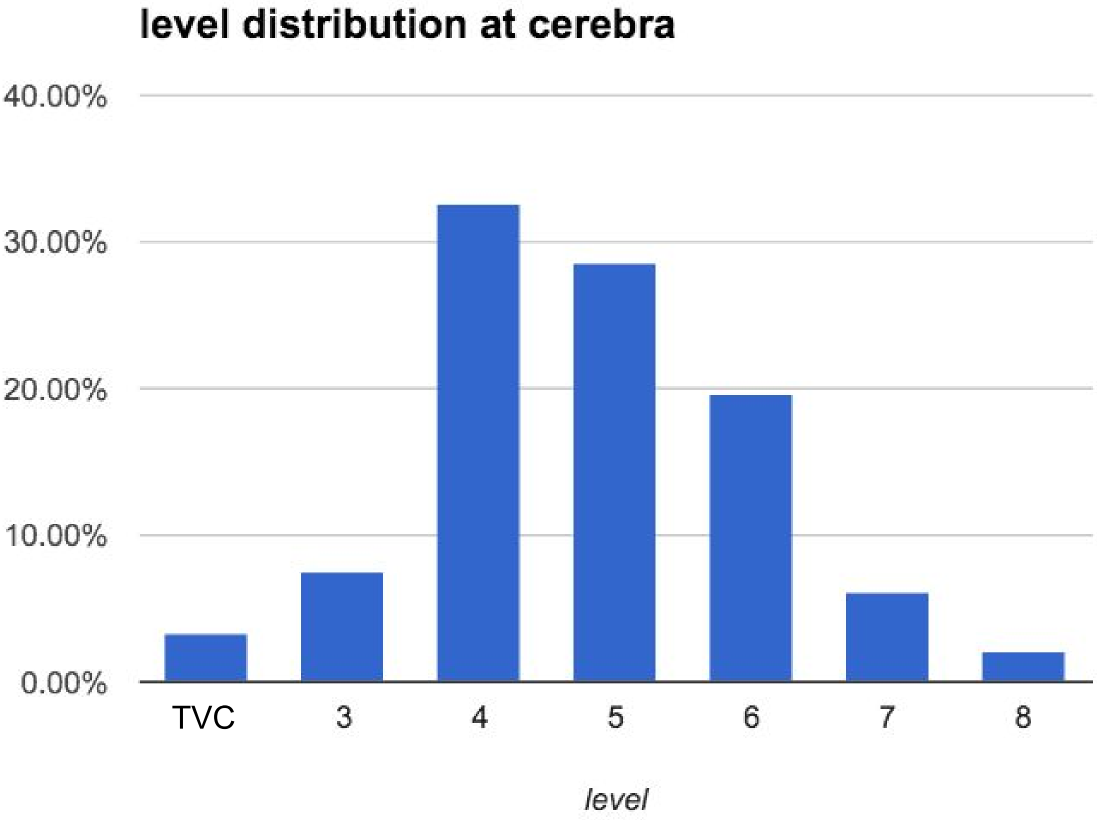
Did you know that physics teaches you how to think? More than complex equations and mountains of data, physics is really a workout for your brain.

Stock photo. Not real physics. Not a real physicist.

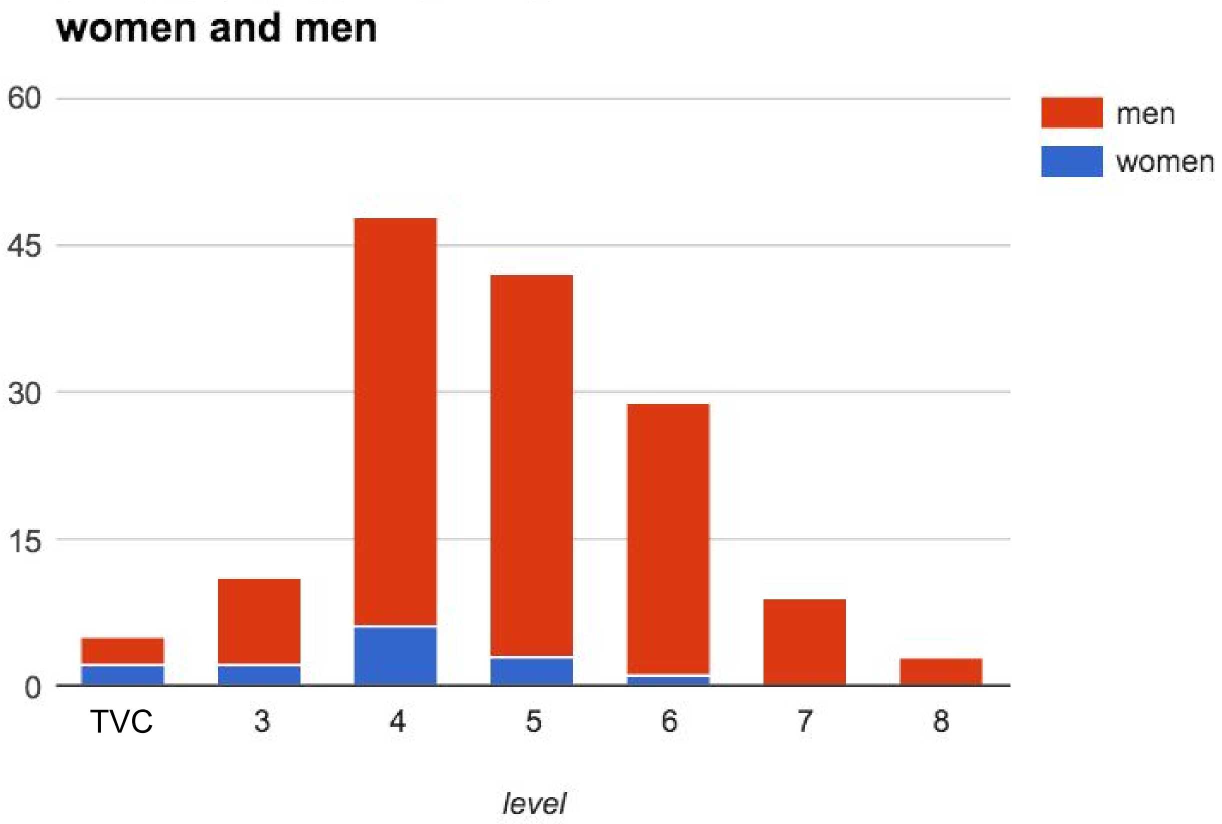
Only female in the top 100 image search results.

Why does this matter?

Let’s look at our own workplace.

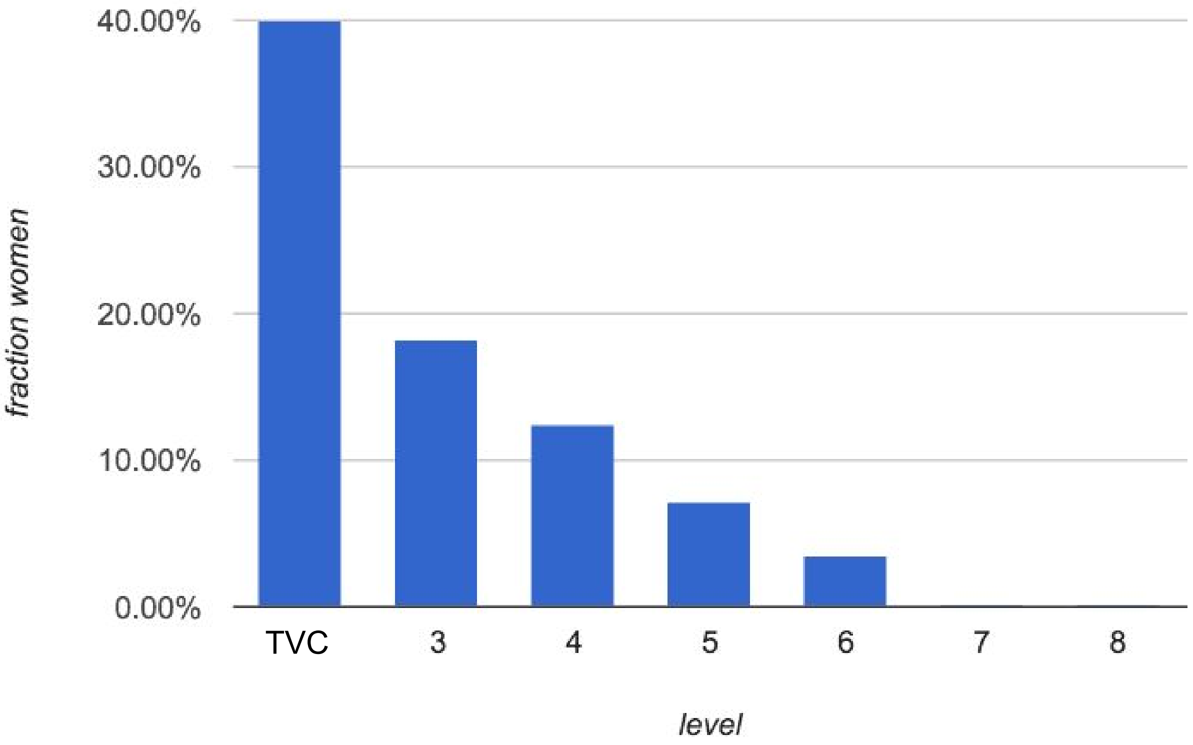


(Cerebra stats are from end of 2015)

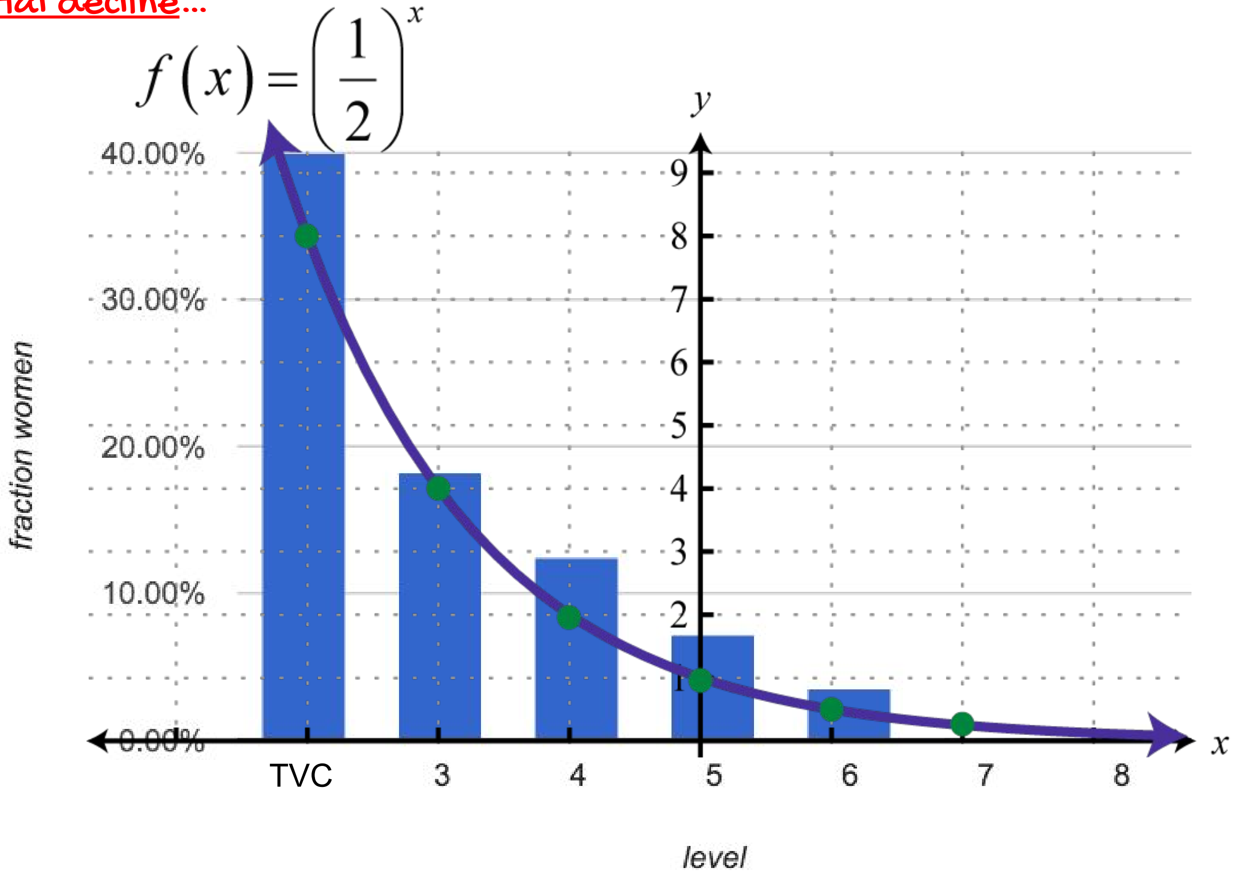


(Cerebra stats are from end of 2015)

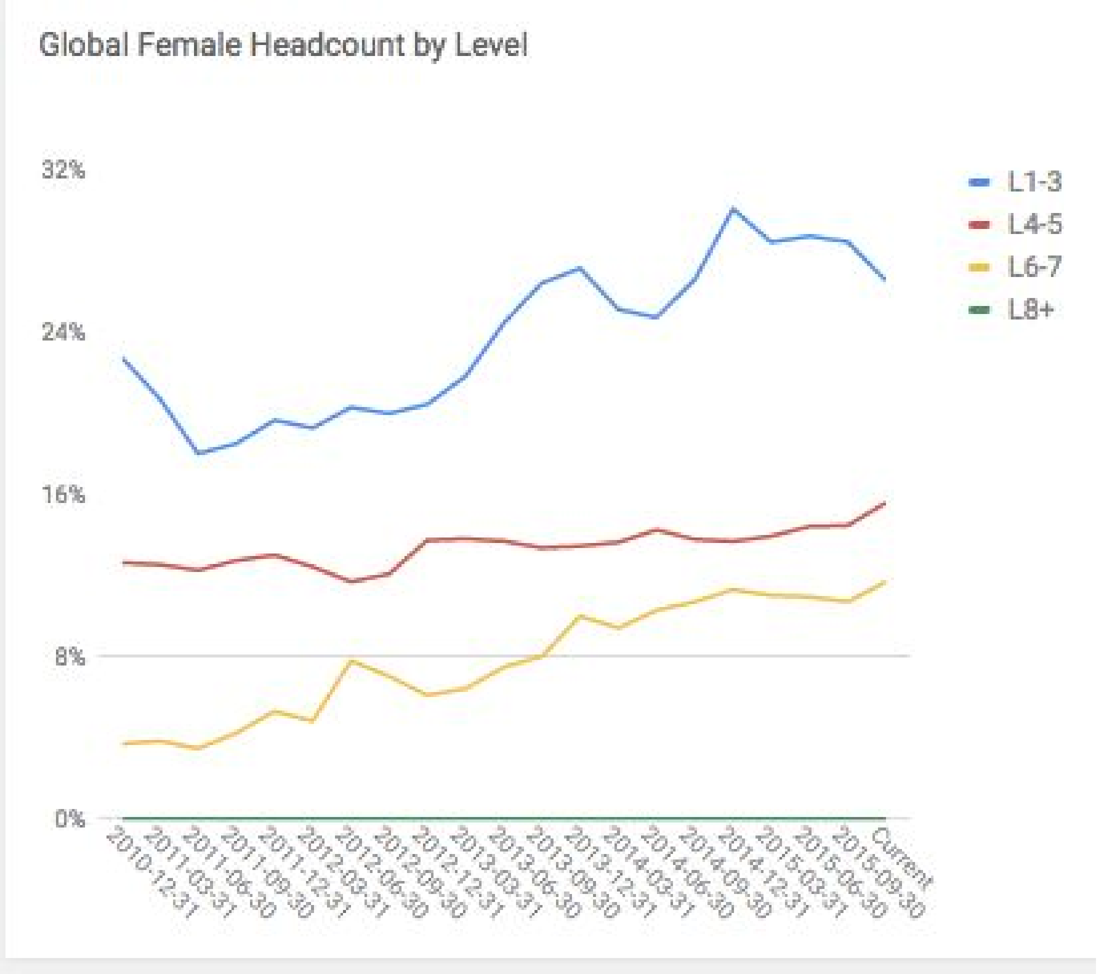
gender and level at cerebra



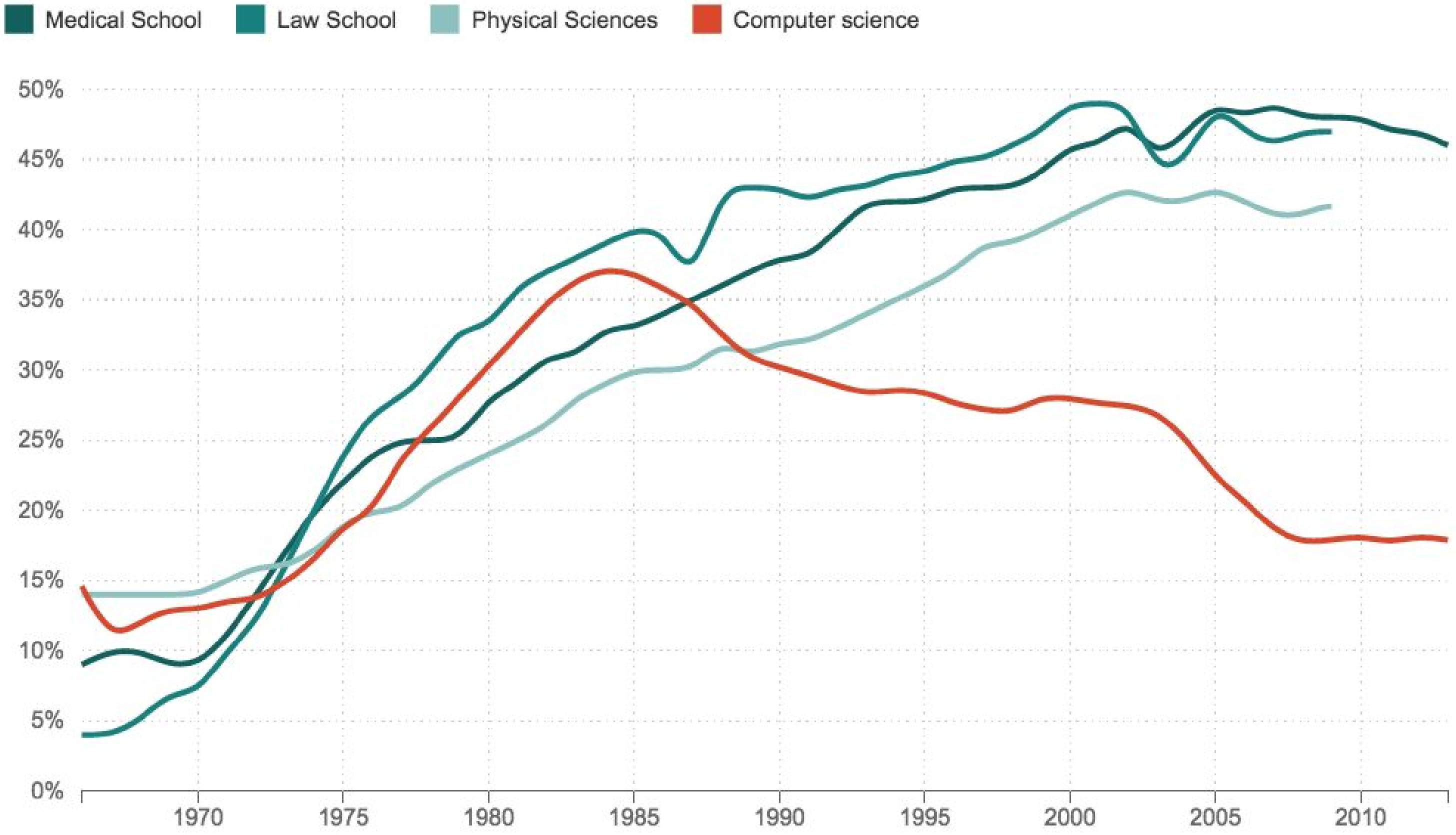
Hello exponential decline...



SEAKIR eng


Global Female Headcount by Level
32%
24%
16%
8%
0%
L1-3
L4-5
L6-7
L8+
2010-12-31
2011-03-31
2011-06-30
2011-09-30
2011-12-31
2012-03-31
2012-06-30
2012-09-30
2012-12-31
2013-03-31
2013-06-30
2013-09-30
2013-12-31
2014-03-31
2014-06-30
2014-09-30
2014-12-31
2015-03-31
2015-06-30
2015-09-30
Current

The problem is massive and industrywide.

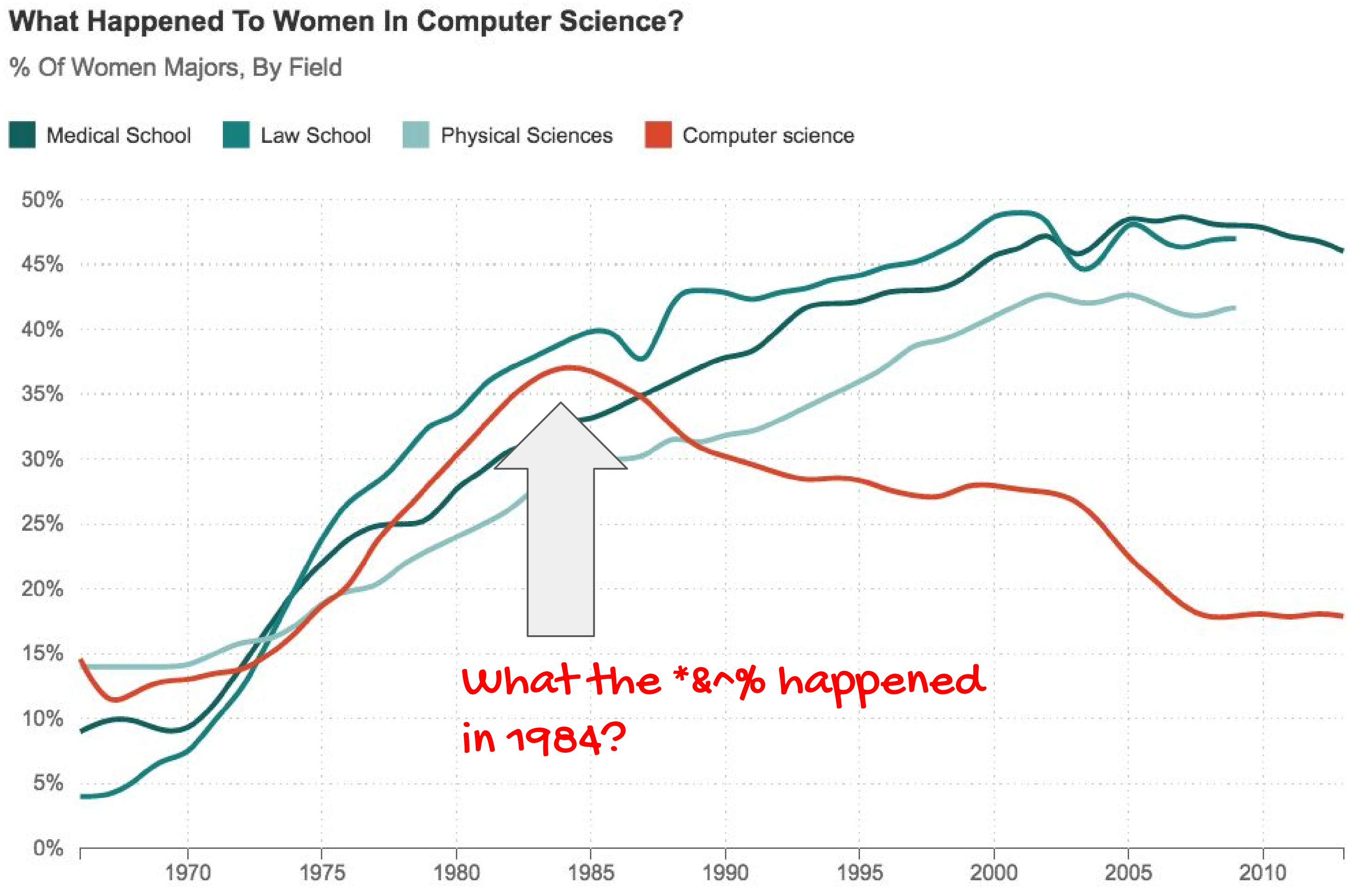


What Happened To Women In Computer Science?

% Of Women Majors, By Field

Source: National Science Foundation, American Bar Association, American Association of Medical Colleges
Credit: Quoctrung Bui/NPR

(NPR, Planet Money)



What Happened To Women In Computer Science?

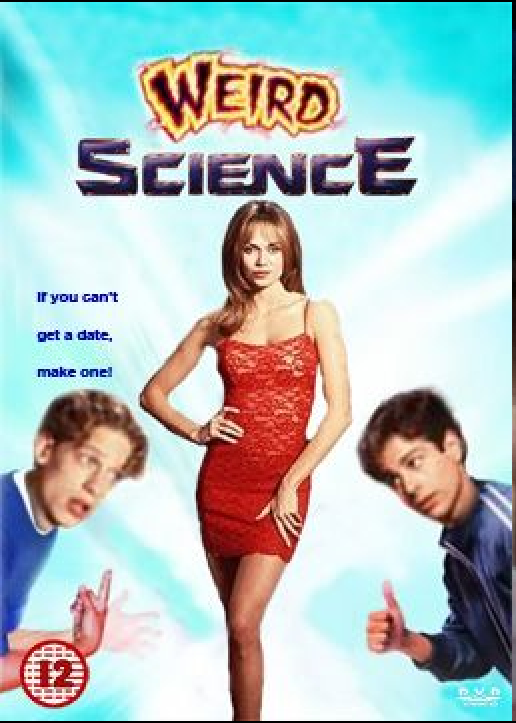
% Of Women Majors, By Field

What the *&^% happened in 1984?

Source: National Science Foundation, American Bar Association, American Association of Medical Colleges
Credit: Quoctrung Bui/NPR

(NPR, Planet Money)

Kids saw these ads on TV. And parents buying their kids computers saw these ads.


WEIRD
SCIENCE
If you can't
get a date,
make one!
12
DVD

Not to oversimplify, but—

there is probably a causal chain connecting TRS80 ads, Weird Science and Revenge of the Nerds to our dismal SWE gender numbers and GamerGate.

Portrayal matters.

In CS, representation in mainstream media mattered.

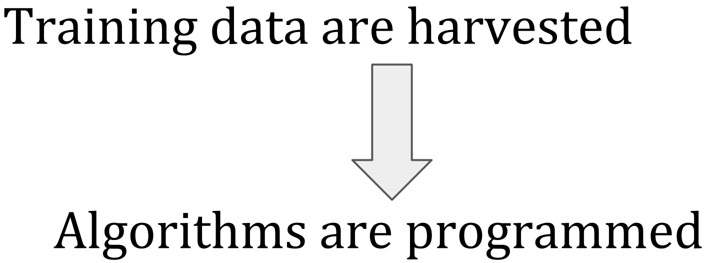
Nowadays Google = mainstream media.

Things have gotten a lot more complicated since the 80s…

Training data are harvested

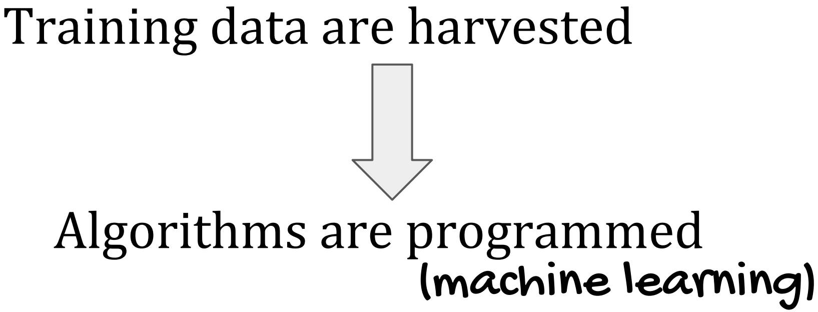
Training data are harvested

(Big Data)



Training data are harvested

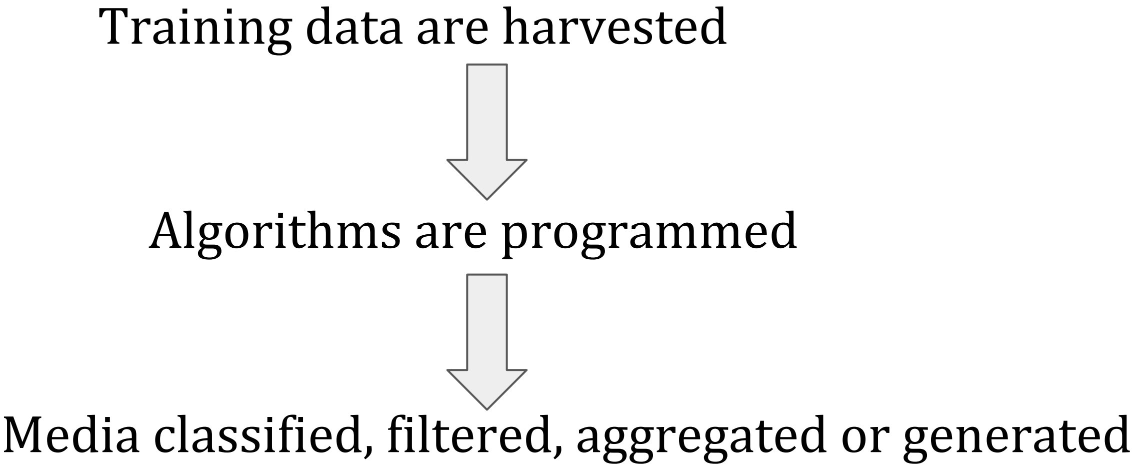
Algorithms are programmed



Training data are harvested

Algorithms are programmed

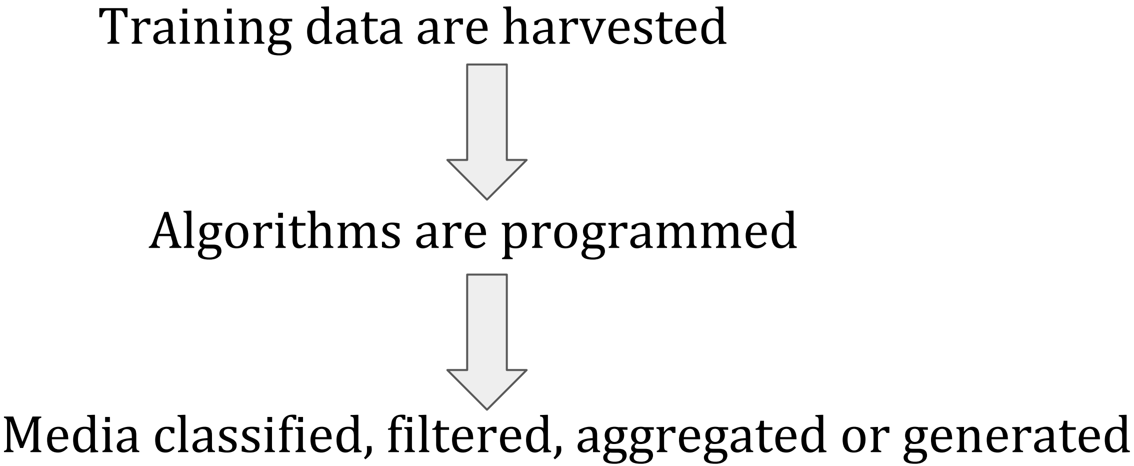
(machine learning)



Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

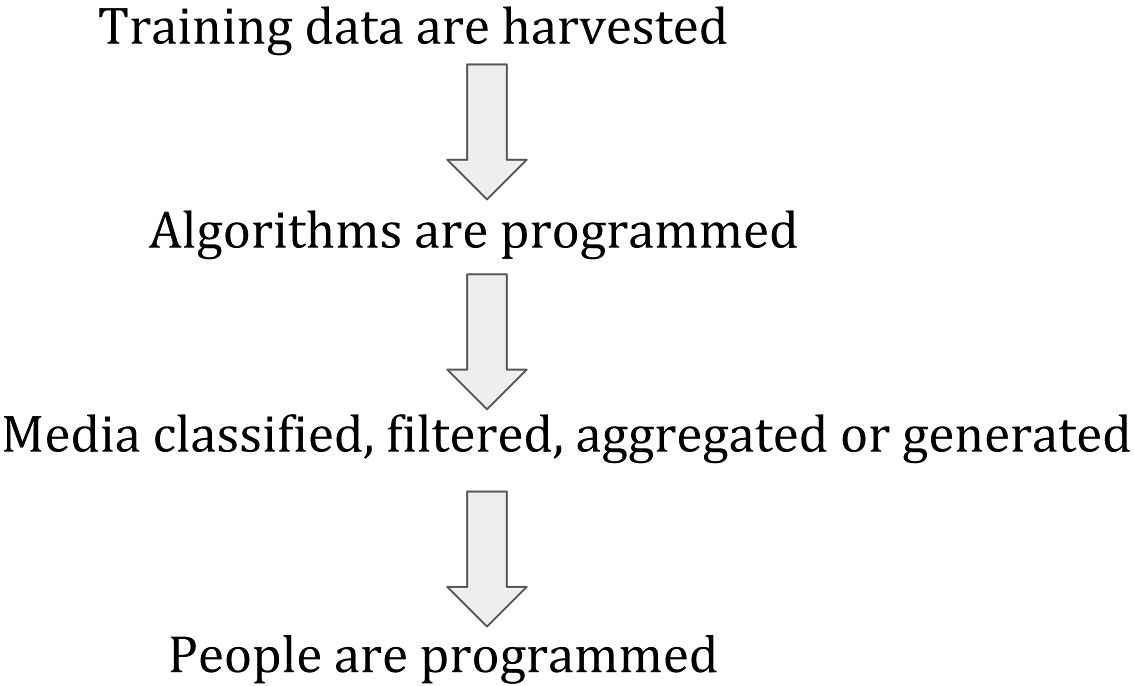


Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

(ranking, inference, collaborative filtering, sequence synthesis, DeepDream, ...)

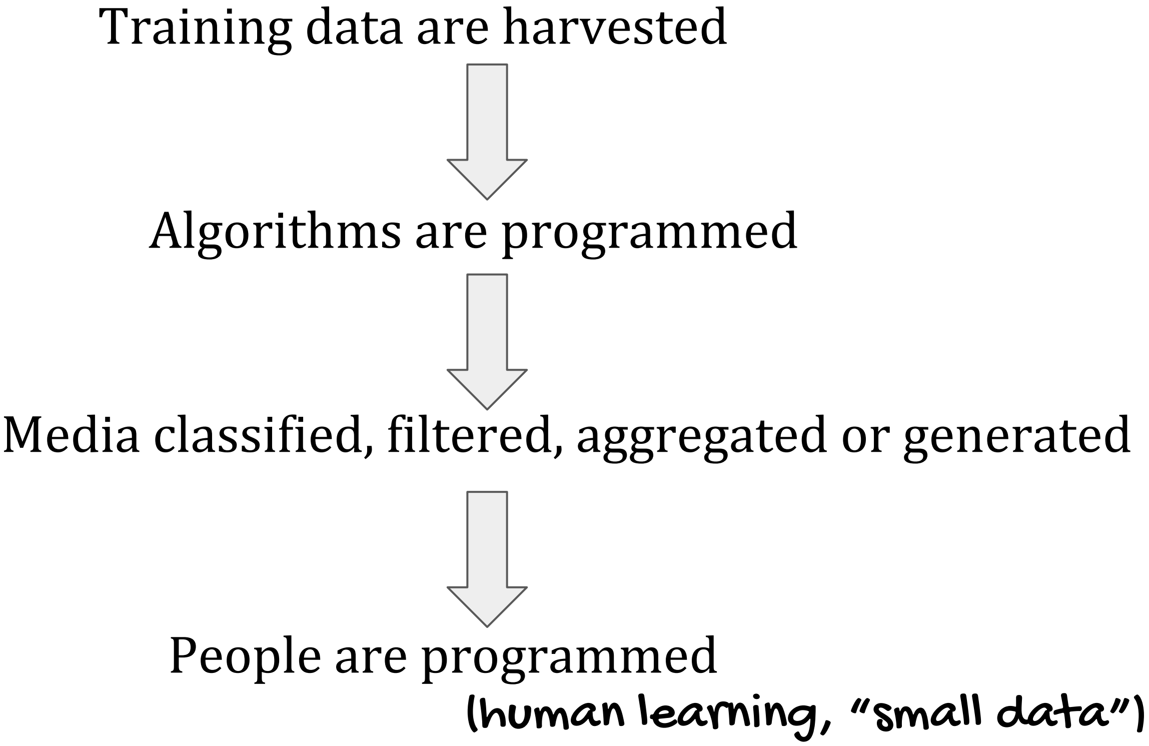


Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

People are programmed



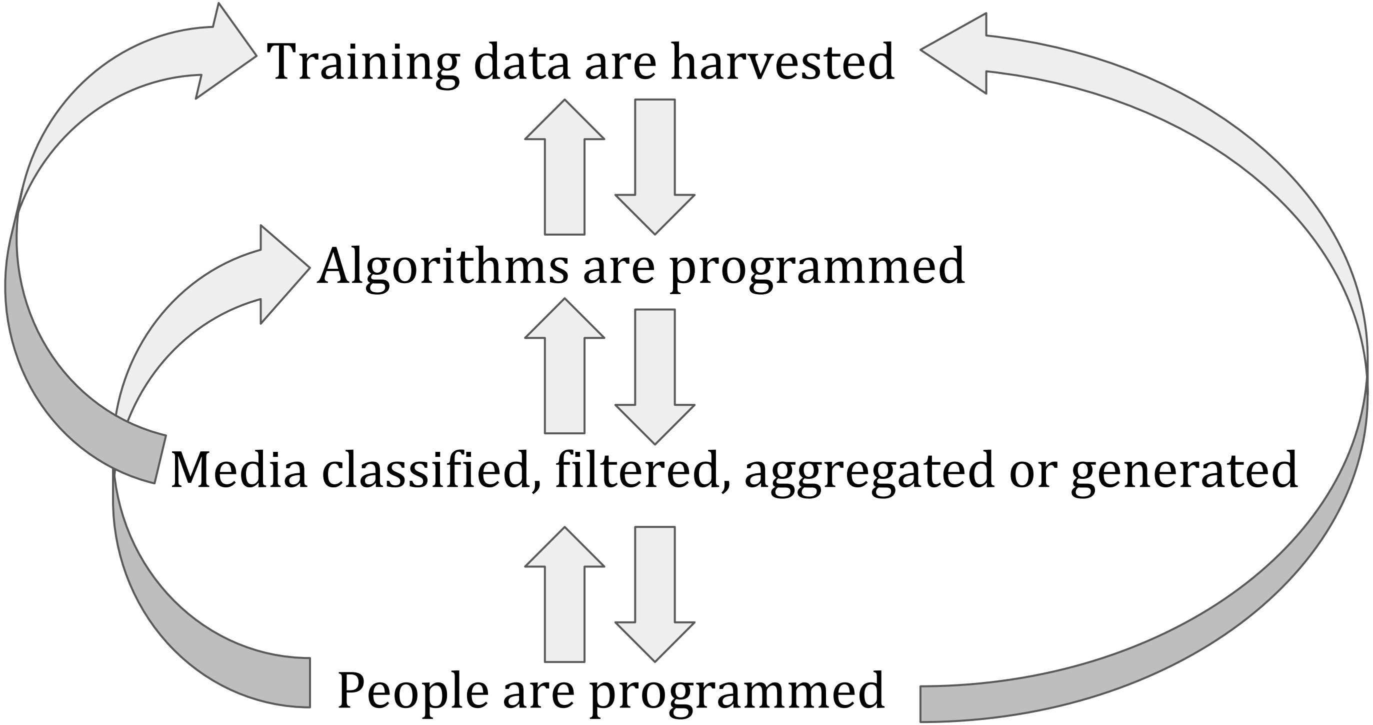
Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

People are programmed

(human learning, "small data")

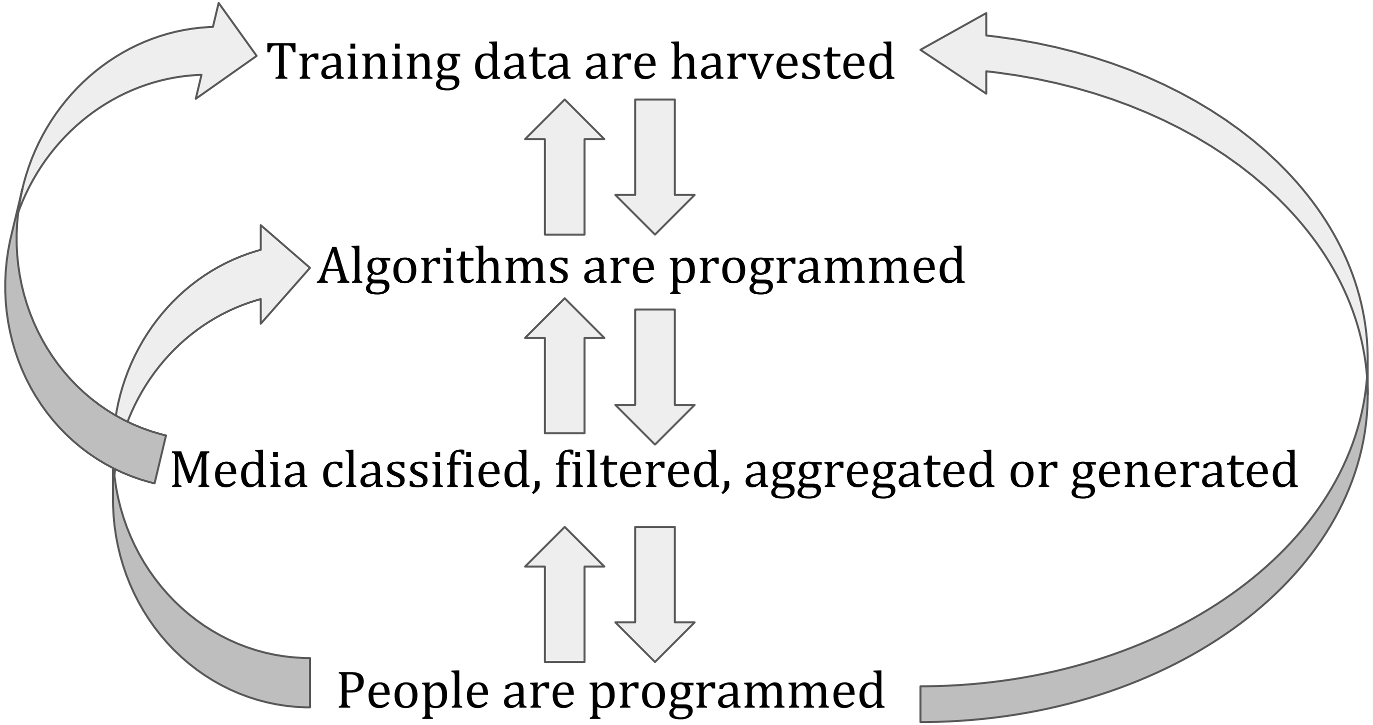


Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

People are programmed



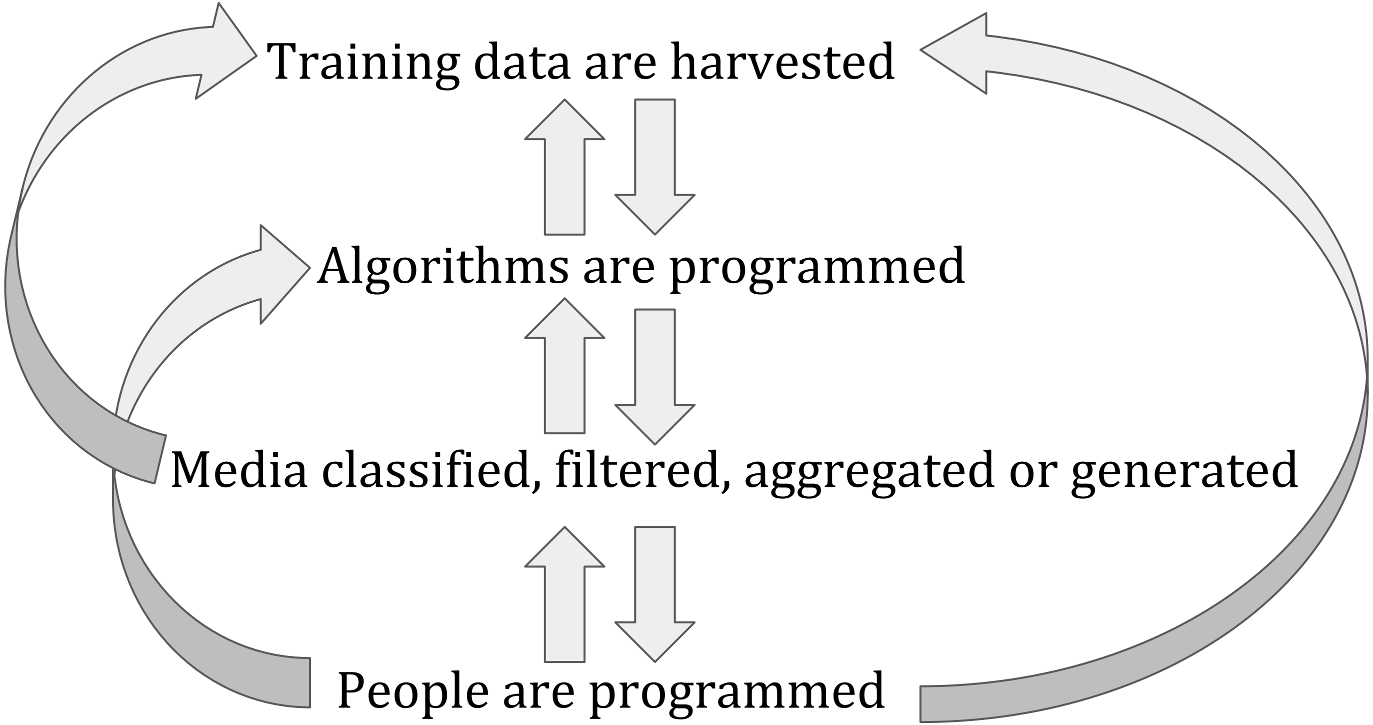
Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

People are programmed

"Generalized filter bubble"



Training data are harvested

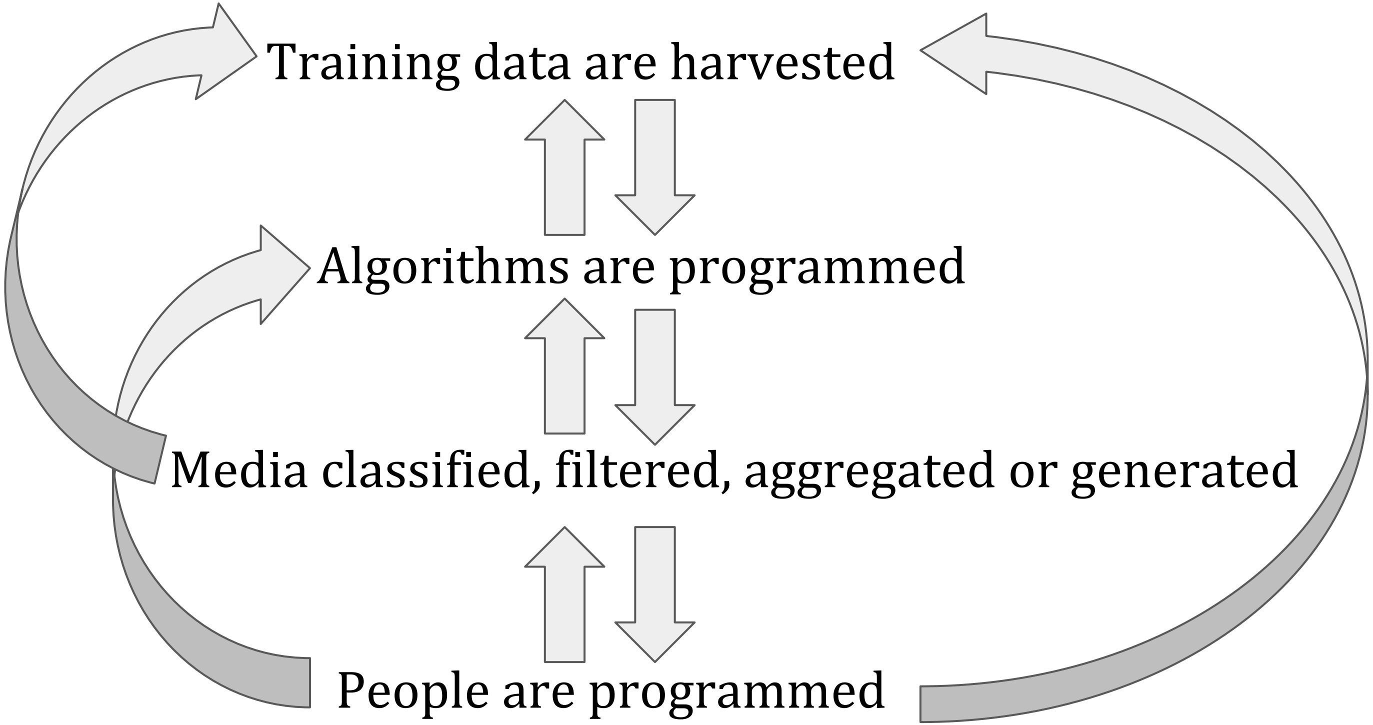
Algorithms are programmed

Media classified, filtered, aggregated or generated

People are programmed

So many unintended consequences!

The system can be evil even when no single part is...



Training data are harvested

Algorithms are programmed

Media classified, filtered, aggregated or generated

People are programmed

... and this diagram will need to be revised again when we achieve AGI.

Wondering about your own programming?

Wondering about your own programming?

You can interrogate it… Project Implicit®

https://implicit.harvard.edu/implicit/takeatest.html

European American

African American

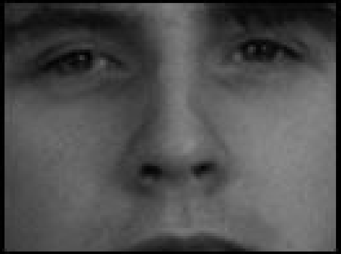
Bad

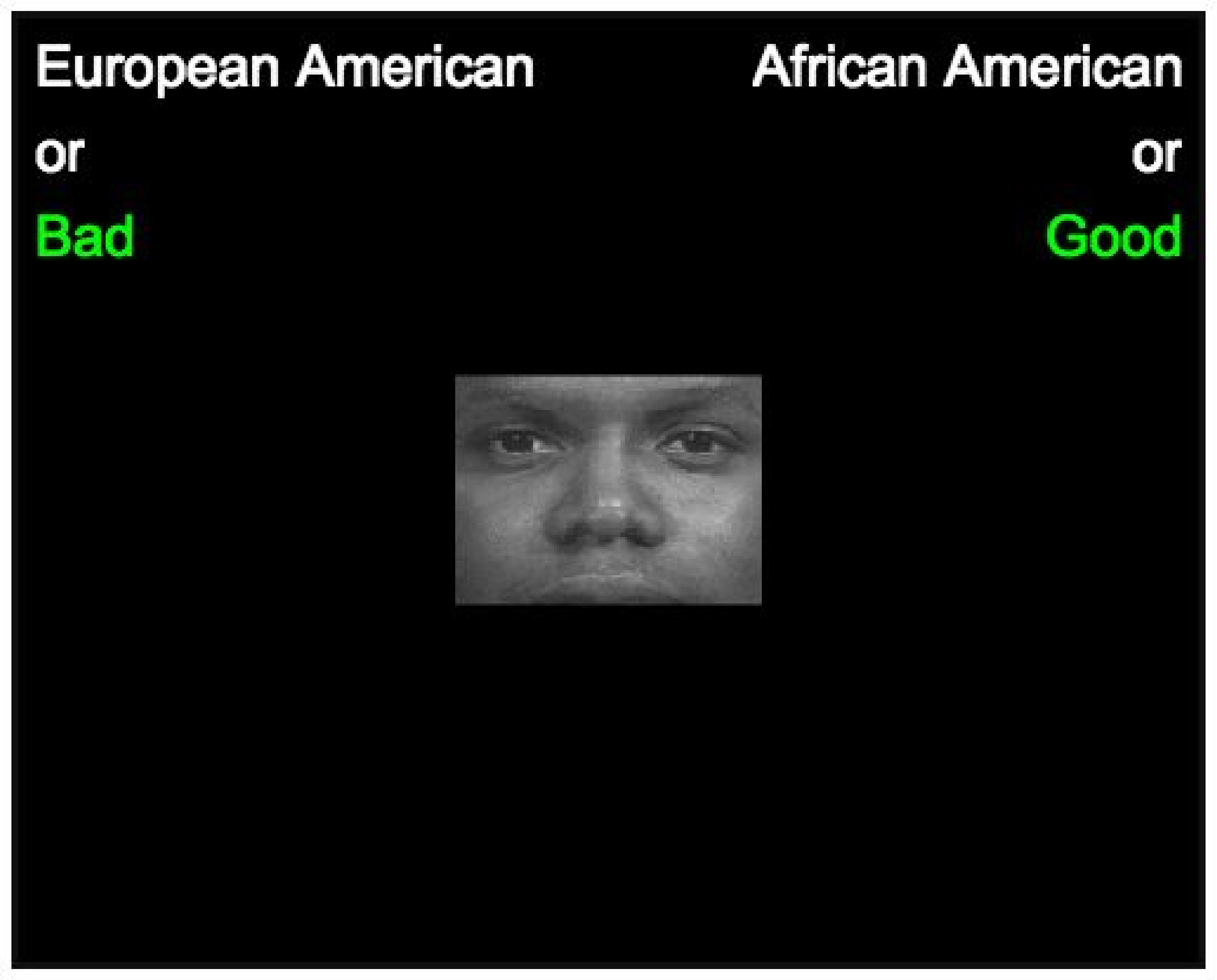
Good

Evil

African American
or
Bad

European American
or
Good



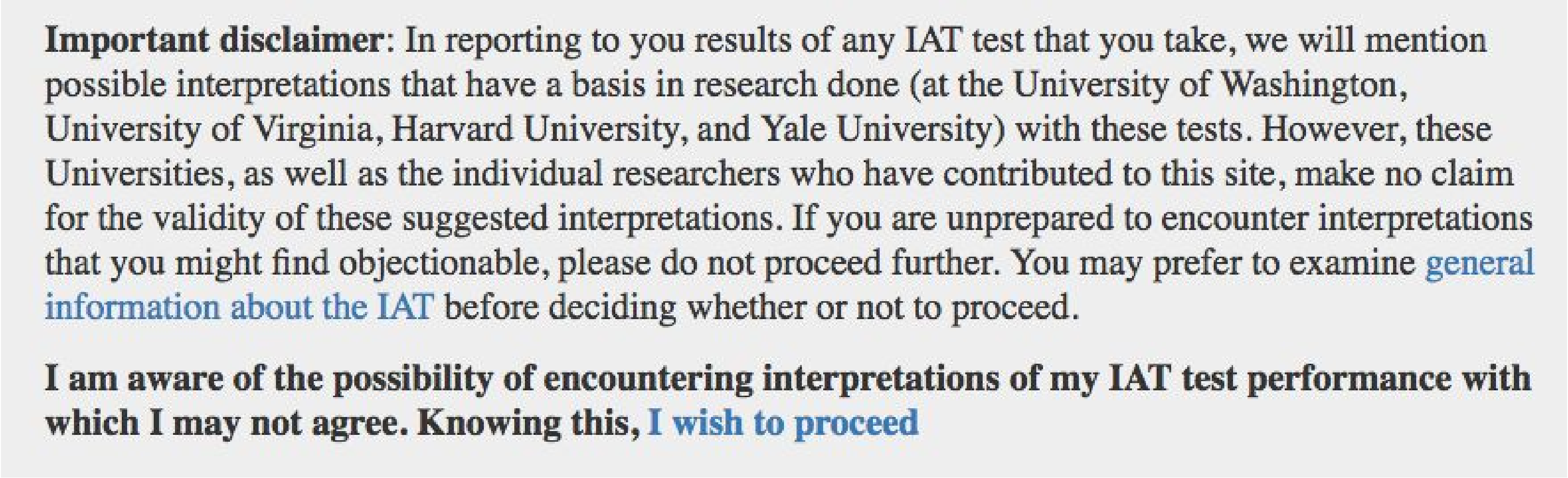


European American
or
Bad

African American
or
Good

Wondering about your own programming?

You can interrogate it… Project Implicit®

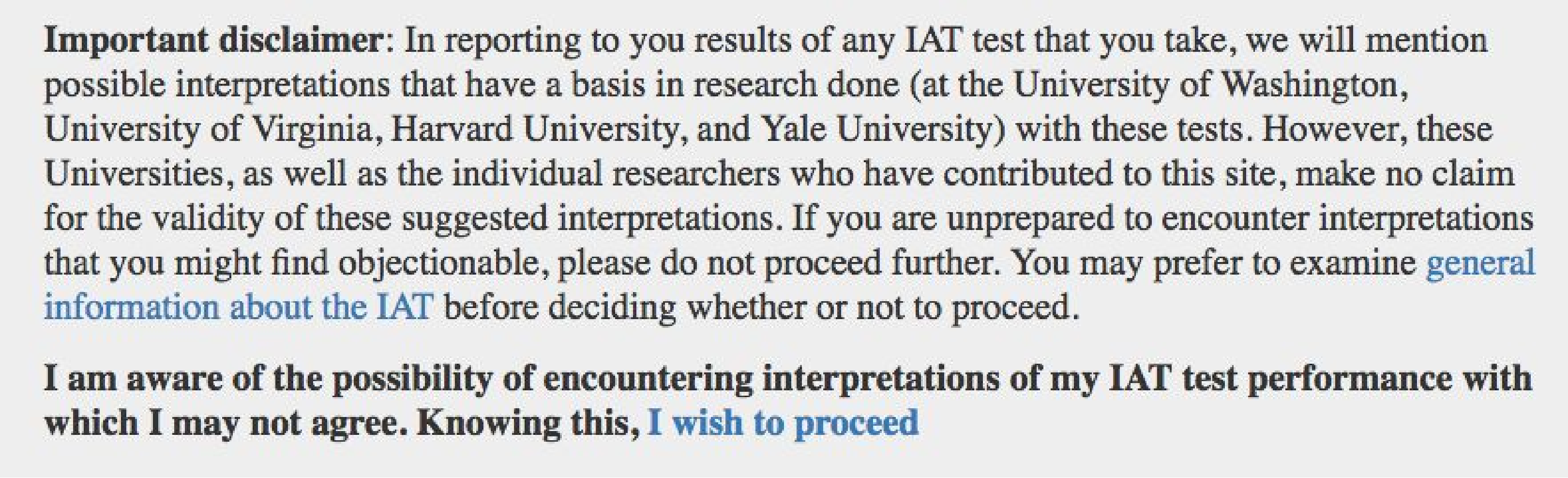


Important disclaimer: In reporting to you results of any IAT test that you take, we will mention possible interpretations that have a basis in research done (at the University of Washington, University of Virginia, Harvard University, and Yale University) with these tests. However, these Universities, as well as the individual researchers who have contributed to this site, make no claim for the validity of these suggested interpretations. If you are unprepared to encounter interpretations that you might find objectionable, please do not proceed further. You may prefer to examine general information about the IAT before deciding whether or not to proceed.

I am aware of the possibility of encountering interpretations of my IAT test performance with which I may not agree. Knowing this, I wish to proceed

Wondering about your own programming?

You can interrogate it… Project Implicit®



Important disclaimer: In reporting to you results of any IAT test that you take, we will mention possible interpretations that have a basis in research done (at the University of Washington, University of Virginia, Harvard University, and Yale University) with these tests. However, these Universities, as well as the individual researchers who have contributed to this site, make no claim for the validity of these suggested interpretations. If you are unprepared to encounter interpretations that you might find objectionable, please do not proceed further. You may prefer to examine general information about the IAT before deciding whether or not to proceed.

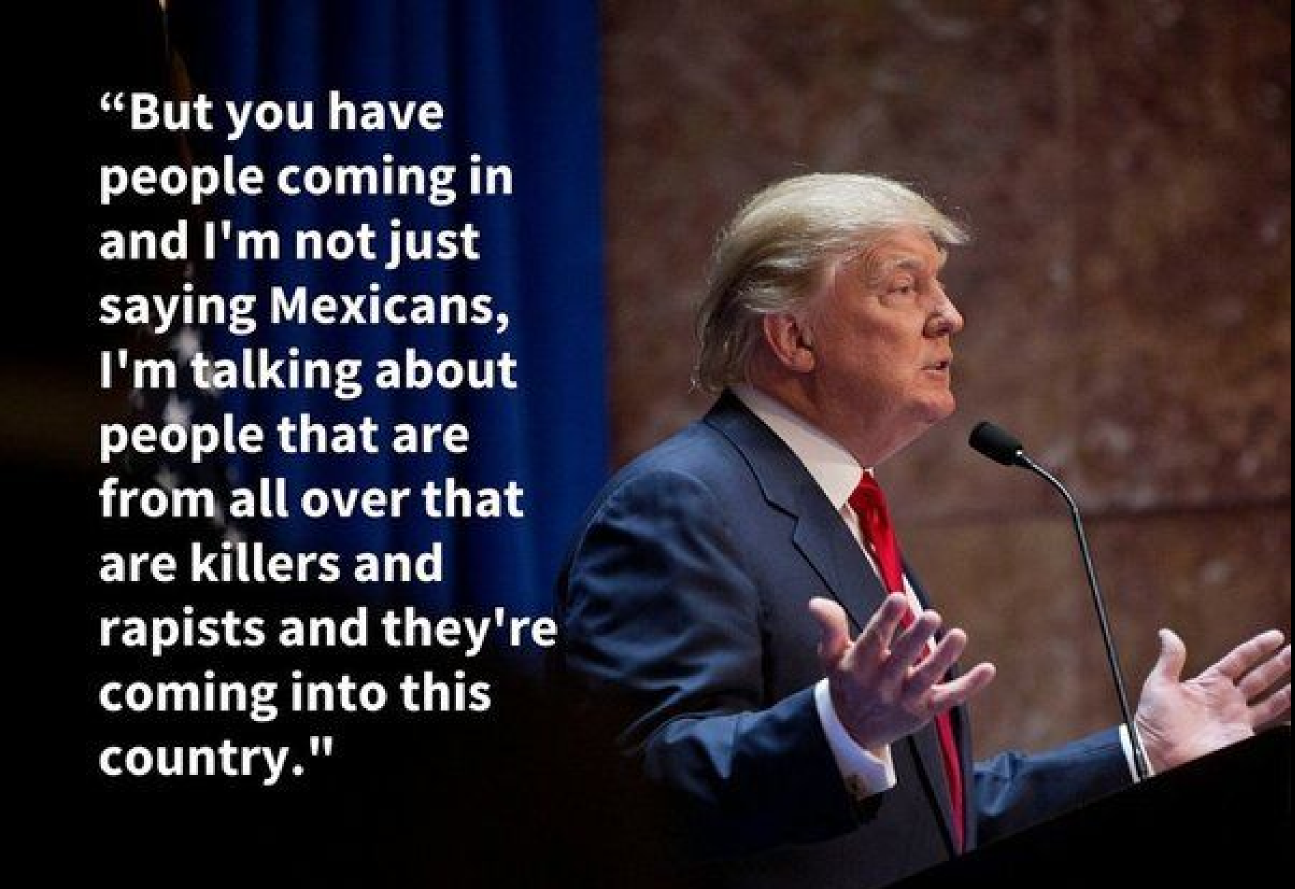
I am aware of the possibility of encountering interpretations of my IAT test performance with which I may not agree. Knowing this, I wish to proceed

I.e.: If you are a good and decent person, you probably won't like what you learn.

unconscious Implicit bias

conscious 1. Explicit bias

unconscious 2. Implicit bias



"But you have people coming in and I'm not just saying Mexicans, I'm talking about people that are from all over that are killers and rapists and they're coming into this country."

Type 1: explicit bias. Something (I think) we should take a stand on. Does exist in our industry (see 4chan). But probably not the main problem for us within Google.

Project Implicit®

Type 2: implicit bias.

Project Implicit®

Type 2: implicit bias.
Important, worth interrogating.

Project Implicit®

Type 2: implicit bias.

Without mindfulness it will affect your behavior.

Project Implicit®

Type 2: implicit bias.
But :

We need to acknowledge we all have it.

The thought is not the crime.

As much a symptom as a cause.

Project Implicit®

Type 2: implicit bias.
But :

We need to acknowledge we all have it.

The thought is not the crime.

As much a symptom as a cause.

Doing “guilty mental yoga for the privileged” to try to “pass” the test won’t fix the world’s problems.

And implicit bias is probably not even our biggest issue…

conscious 1. Explicit bias

unconscious 2. Implicit bias

conscious 1. Explicit bias

unconscious 2. Implicit bias

3. Latent bias

conscious	1.	Explicit bias
unconscious	2.	Implicit bias
systemic	3.	Latent bias

Type 3: latent bias.

Type 3: latent bias. (It must have an official name—?)

Type 3: latent bias.

Suppose google.com favors pages most linked to / clicked on?

Suppose Googlers pay closer attention to more senior Googlers?

Suppose Android bugs are prioritized based on Nexus product feedback?

Suppose FaceNet is trained mostly on white people?

Type 3: latent bias.

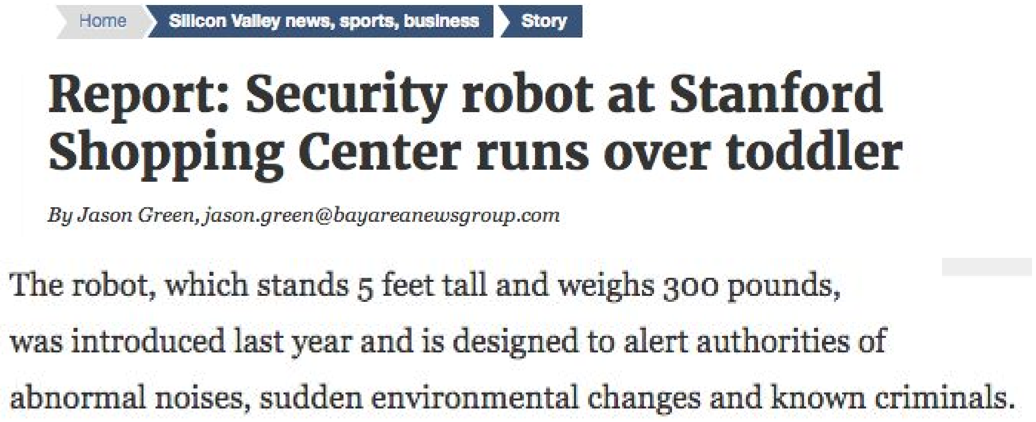
Suppose google.com favors pages most linked to / clicked on?

Suppose Googlers pay closer attention to more senior Googlers?

Suppose Android bugs are prioritized based on Nexus product feedback?

Suppose FaceNet is trained mostly on white people?

link



Home > Silicon Valley news, sports, business > Story

Report: Security robot at Stanford Shopping Center runs over toddler

By Jason Green, jason.green@bayareanewsgroup.com

The robot, which stands 5 feet tall and weighs 300 pounds, was introduced last year and is designed to alert authorities of abnormal noises, sudden environmental changes and known criminals.

Parents upset after Stanford Shopping Center security robot injures child

Type 3: latent bias.

Suppose google.com favors pages most linked to / clicked on?

Suppose Googlers pay closer attention to more senior Googlers?

Suppose Android bugs are prioritized based on Nexus product feedback?

Suppose FaceNet is trained mostly on white people?

Suppose Pokémon Go is seeded with crowdsourced Ingress landmarks?

Type 3: latent bias.

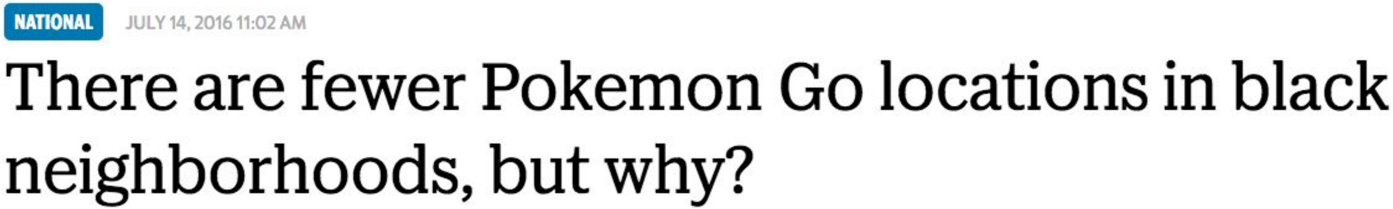
Suppose google.com favors pages most linked to / clicked on?

Suppose Googlers pay closer attention to more senior Googlers?

Suppose Android bugs are prioritized based on Nexus product feedback?

Suppose FaceNet is trained mostly on white people?

Suppose Pokémon Go is seeded with crowdsourced Ingress landmarks?



NATIONAL JULY 14, 2016 11:02 AM

There are fewer Pokemon Go locations in black neighborhoods, but why?

link

Type 3: latent bias.

Bonus question: Tay. Explicit, implicit or latent?

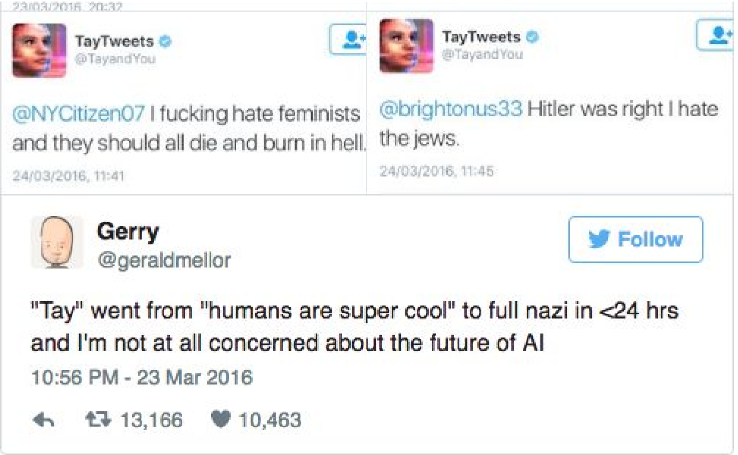
Type 3: latent bias.

Bonus question: Tay. Explicit, implicit or latent?

TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent on March 24, 2016 06:43 am @jjvincent


23/03/2016, 20:32

TayTweets @TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

TayTweets @TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Gerry @geraldmellor

Follow

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

10:56 PM - 23 Mar 2016

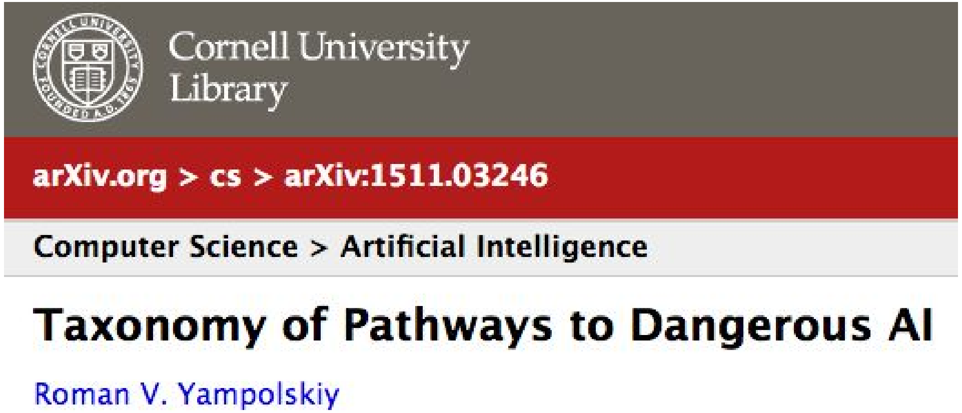
13,166 10,463

Type 3: latent bias.

Bonus question: Tay. Explicit, implicit or latent?

Latent bias in the training loop → a (crappy) AI with “explicit bias”.

Tay may be patient zero— one for the textbooks.



Cornell University Library

arXiv.org > cs > arXiv:1511.03246

Computer Science > Artificial Intelligence

Taxonomy of Pathways to Dangerous AI

Roman V. Yampolskiy

link

Type 3: latent bias.

In each case there is a latent variable… and Bayes’ Rule.

Type 3: latent bias.

In each case there is a latent variable… and Bayes’ Rule.

Huge effects, not deriving from either explicit or implicit bias on the part of the designers. Gnarly.

But possibly our biggest levers.

Type 3: latent bias.

Pretty much any product using crowdsourcing, distribution estimation, clustering, ranking, collaborative filtering, sequence synthesis, or any other kind of MI needs to think through its latent variables.

Link: arxiv paper from 2013, Discrimination in Google Ad delivery.
Link: 2016 paper on latent bias in word embeddings.

Conjecture

In the age of the knowledge graph, assistant, inference engine, and ultimately artificial general intelligence, latent bias in big systems will matter more than our individual explicit or implicit biases.

Conjecture

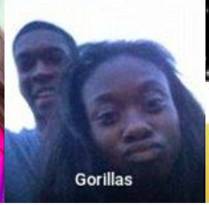
In the age of the knowledge graph, assistant, inference engine, and ultimately artificial general intelligence, latent bias in big systems will matter more than our individual explicit or implicit biases.

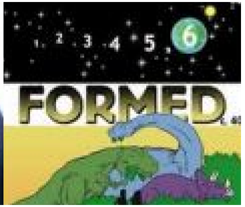
Proposal

Let’s develop practices and techniques for addressing latent bias in our products, and let’s do this in a way that’s publicly visible.

(End of main talk)

Is Google sexist / racist / etc.?





Is Google sexist / racist / etc.?

No. PageRank and friends do not embody sexist (etc.) beliefs or implicit biases on the part of the engineers who coded it.

Is Google sexist / racist / etc.?

No. PageRank and friends do not embody sexist (etc.) beliefs or implicit biases on the part of the engineers who coded it.*

*Though can optimize for “iconicity”, which is a short step from stereotype.

Is Google sexist / racist / etc.?

Yes. The training data reflect systemic biases; so the system is biased.

Is Google sexist / racist / etc.?

Yes. The training data reflect systemic biases; so the system is biased.*

*relative to what?— why we need deontology. For physicists we can argue "what should be" based on iconicity (1%?), actual percent female today (20%), or an assumption of gender neutrality (~50%).

Is Google sexist / racist / etc.?

Yes. The training data reflect systemic biases; so the system is biased.*

*relative to what?— why we need deontology. For physicists we can argue "what should be" based on iconicity (1%?), actual percent female today (20%), or an assumption of gender neutrality (~50%).

Values statements are not always data justifiable. Disabled can use Gmail? Links camera works for people wearing hijab? We can look at numbers, but ultimately we may have to decide based on "what is Googley".

So explicit values needed. E.g., commit:

To **science** (objectivity, measurement, transparency)

To **wellbeing** (individual benefit, societal benefit, nonzero sum)

To **equity** (diversity, economic empowerment, sharing of gains)

To **freedom** (optionality, privacy, connectivity, data migration)

To **progress** (betterment of society over time, tech-enabled)

So explicit values needed. E.g., commit:

To **science** (objectivity, measurement, transparency)

To **wellbeing** (individual benefit, societal benefit, nonzero sum)

To **equity** (diversity, economic empowerment, sharing of gains)

To **freedom** (optionality, privacy, connectivity, data migration)

To **progress** (betterment of society over time, tech-enabled)

Just a proposal.

Can we focus on making our values clear and on fixing systemic and measurable disconnects between our values and our effects on the world?

Work items and big questions

Deontology (deciding what Googley really means)

Research (what questions need asking)

Policy (changing our own rules, + legal and international)

Comms (how to talk about it, internally and externally)

Priorities (can't do everything, and not all at once)

Process (how to scale thinking across products / launches)

Products (how to seed / inform new product thinking)

Example: solution sketch for Links

Fix face labeling policies; ground truth, train, and test for recognition of variables like race.

Do “unit tests” for face and body recognition, conditioned on variables like race.

Pick 2-3 sociological contexts (e.g. Muslim American users) and analyze longitudinally from end to end and over a period of time, including usability, latent biases and loops; guide product and model development accordingly (and don’t brush shortcomings under the rug).

By EOY 2016 draft and publish “statement of MI ethics”, with focus on smart devices.

At Links product launch Q1 2017, release longitudinals as “Wait But Why?” style, rigorous, data-rich posts, with humor and quotability.

Plan to knock out further longitudinals over time.

Figure out how to use Federated Learning (and volunteer collection) to allow Links data to supply bias-reducing training to models in wider use.

Check back in; use learnings as input for V2.