

Searching For Bias

Search engines serve as the reference librarians of the web. Individuals and organizations would be ill-served if the world shared the same librarian. Similarly, they would be ill-served by a search engine monopoly

-A Mowshowitz, A. Kawaguchi 2002
Information Processing & MGMT

75% of all desktop/laptop & 87% of all mobile search performed using Google

NetMarketShare.com, November 2018

Our assumptions

- The internet is navigated primarily via IR systems
- Users trust results from these systems as though from newspapers, libraries, and phonebooks
- Much of the information that exists on the internet is unstructured, not vetted, and when it is indexed it is often indexed by the creator
- The makers of successful IR systems earn money somehow
- When errors occur with IR systems, the companies are rarely transparent about cause or fix

4 Different Weather systems showing the 'same' forecast

11 abc CARRBORO NORTH CAROLINA

8:20 AM 99%

63°/50° 42%

SATURDAY, 12/1 DAY Cloudy, a shower in the p.m. 62°/56° 60%

SUNDAY, 12/2 DAY An a.m. shower; mostly cloudy 73°/48° 48%

MONDAY, 12/3 DAY Cloudy, a little rain

Accuweather shows a more positive weather icon

WEDNESDAY, 12/5 DAY Times of clouds and sun 47°/28° 19%

THURSDAY, 12/6 DAY

AccuWeather

Chapel Hill Partly Cloudy

43°

Friday TODAY 61 48

| | | | | | | |
|-----|------|-------|-------|-------|------|------|
| Now | 9 AM | 10 AM | 11 AM | 12 PM | 1 PM | 2 PM |
| 43° | 48° | 52° | 54° | 55° | 58° | 59° |

Saturday 61 55

Sunday 71 51

Monday 62 41

Tuesday 50 33

Wednesday 45 25

Google and Apple both pull from The Weather Channel, yet provided differing results at the same time and from the same location

weather chapel hill

Chapel Hill, NC Fri, 8 AM, Partly Cloudy

41° F | °C

| | | | | | |
|---------|---------|----------------|---------|---------|---------|
| 10 AM | 3 PM | 8 PM | 1 AM | 6 AM | |
| FRI | SAT | SUN | MON | TUE | W |
| 61° 41° | 61° 55° | 71° 51° | 63° 41° | 50° 33° | 44° 25° |

More on weather.com

How might various forecast presentations affect real world decisions and behaviors?

Detailed Forecast

Today A chance of rain, mainly before 1pm. Mostly cloudy, with a high near 59. Southwest wind 5 to 11 mph, with gusts as high as 20 mph. Chance of precipitation is 30%. New precipitation amounts of less than a tenth of an inch possible.

Saturday Night Showers, mainly before 1am. Low around 56. Southeast wind around 6 mph becoming south after midnight. Chance of precipitation is 70%. New precipitation amounts between a tenth and quarter of an inch possible.

Sunday A chance of showers and thunderstorms. Mostly cloudy, with a high near 73. Southwest wind 5 to 11 mph, with gusts as high as 24 mph. Chance of precipitation is 30%. New rainfall amounts of less than a tenth of an inch, except higher amounts possible in thunderstorms.

Sunday Night A slight chance of showers after 1am. Partly cloudy, with a low around 52. Chance of precipitation is 20%.

National weather service. No icon, states a 30% chance of rain

Tuesday Night Mostly cloudy, with a low around 35.

Wednesday Mostly sunny, with a high near 60.

View in Desktop Mode

Bias is disproportionate weight in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.

- From Wikipedia

- Anchoring: The first piece of information given is usually preferred
- Patternicity: We want to see patterns
- Attribution: 'I have my reasons, but you, you're just a bad person'
- Confirmation: 'I knew it!'
- Framing: Your solution sounds better when you define the problem
- Association: Linked to polarizing figure
- Self serving bias: We like info that makes us feel good about ourselves
- Status Quo: Resisting change
- **Conflicts of Interest**

Gender and Racial Bias

Where does the bias originate?

Society?

Developers?

Is search engine bias even possible?

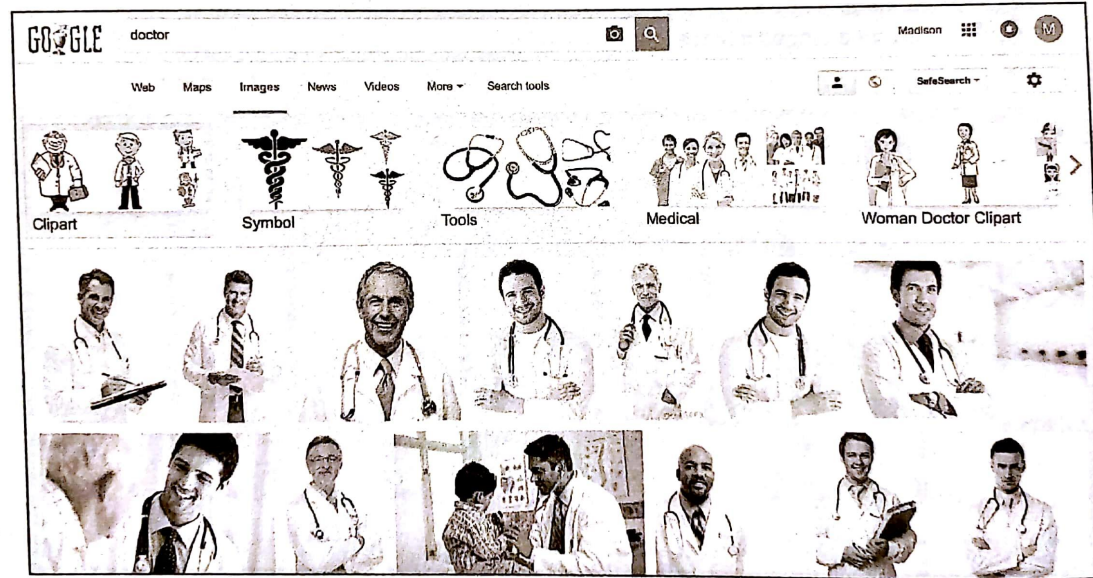


Figure 2.14. Google Images search on “doctor” featuring men, mostly White, as the dominant representation, April 7, 2016.

It definitely exists... **now what?**

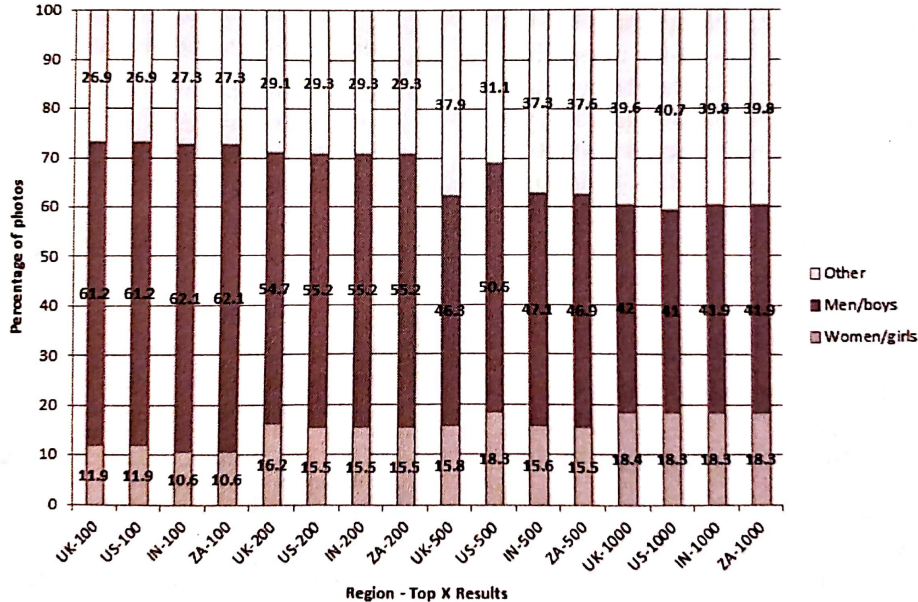


Figure 2: Gender distribution in photos retrieved for "person."

What is a "person"?

- Image results for first 100, 200, 500, 1000 results for query "person"
- No significant difference in region
- Discrepancy in top results
- "Person" retrieves twice as many photos of men as women
- Retrievability bias

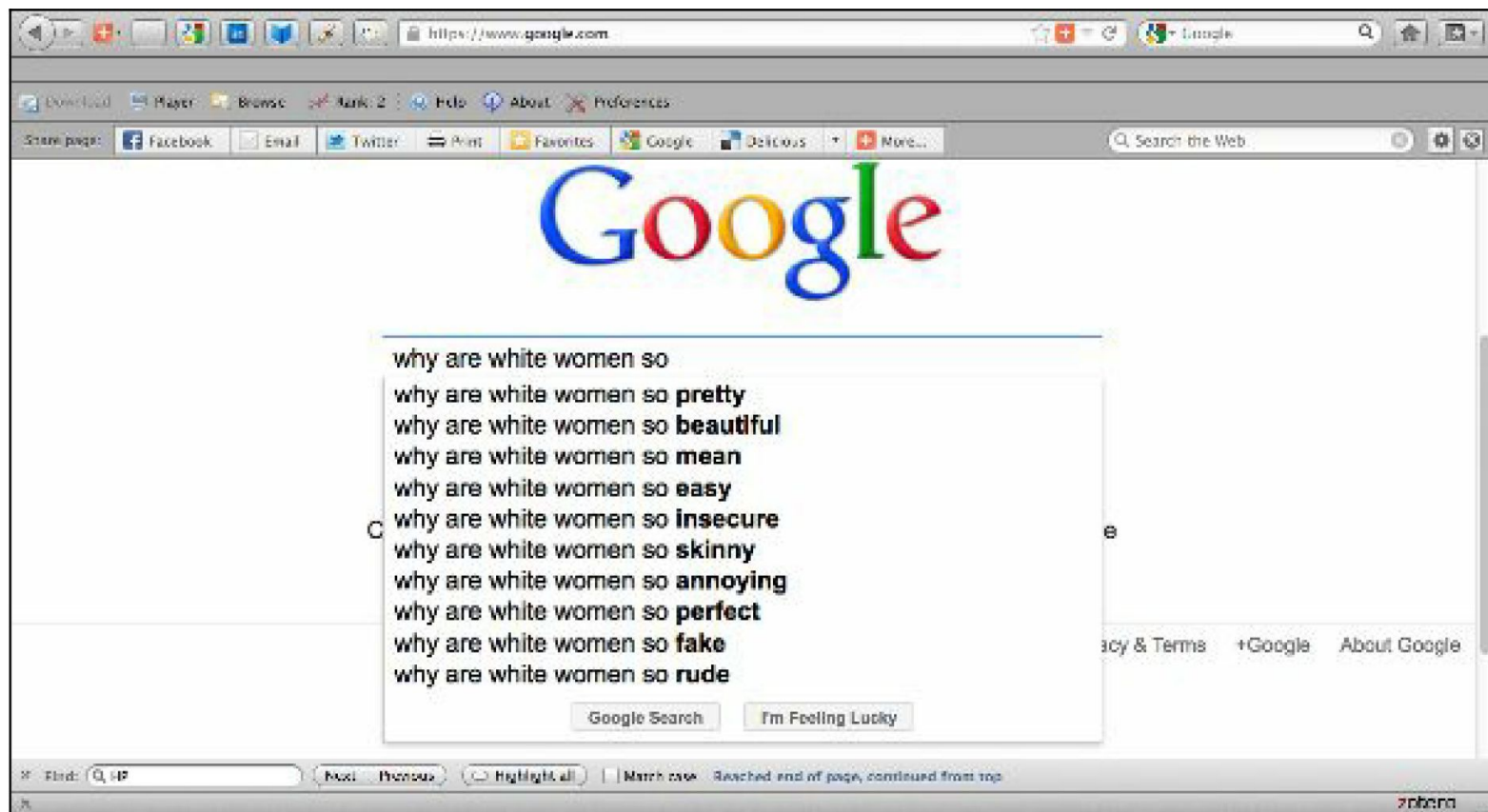


Figure 1.6. Google autosuggest results when searching the phrase “why are white women so,” January 25, 2013.

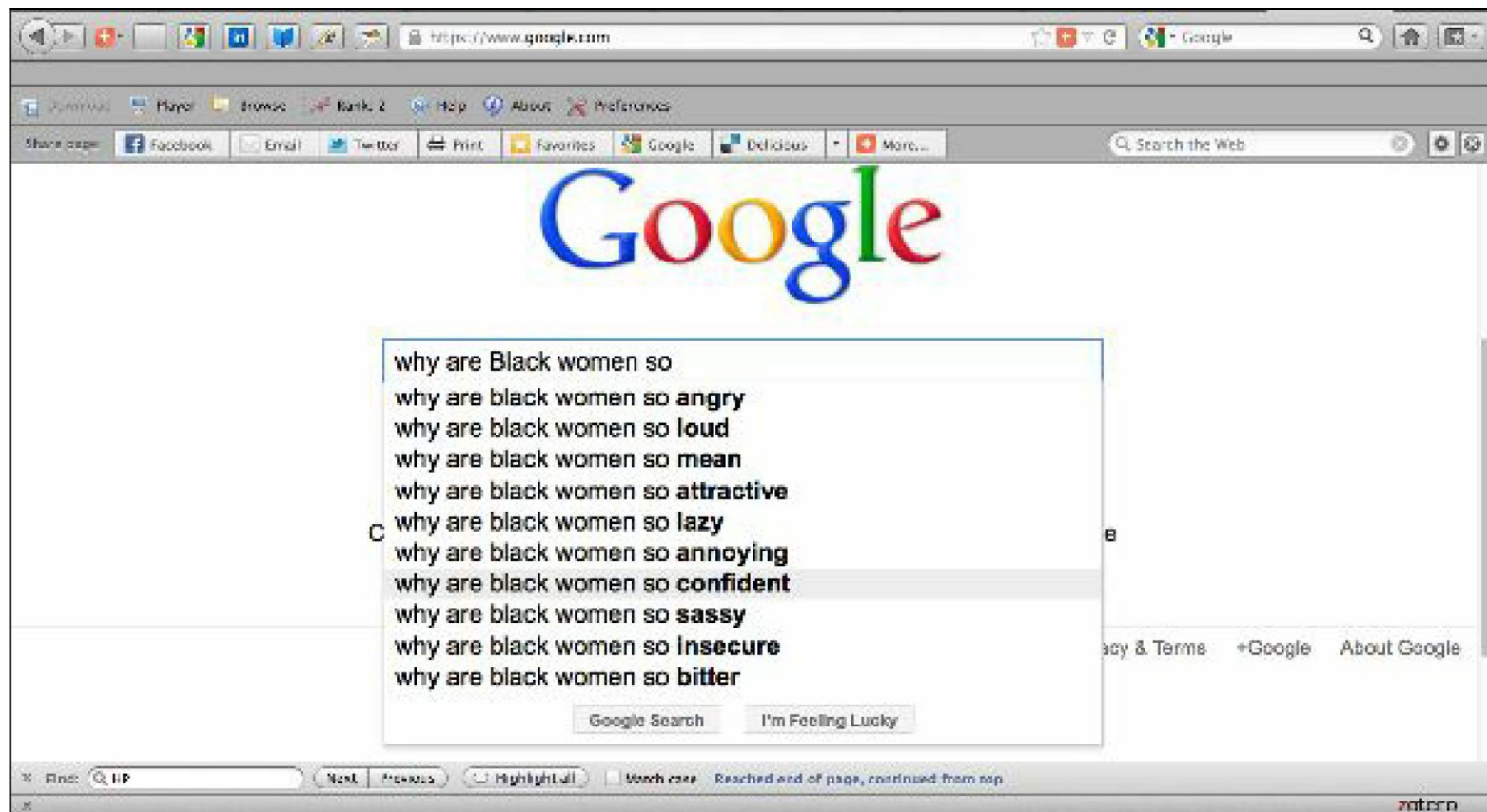


Figure 1.5. Google autosuggest results when searching the phrase “why are black women so,” January 25, 2013.

It definitely exists... **now what?**

Real-world implications

Societal Stress

Reinforcing stereotypes

False data

Cultivation Theory



Figure 2.16. Tweet about Google searches on “unprofessional hairstyles for work,” which all feature Black women, while “professional hairstyles for work” feature White women, April 7, 2016.

Solutions?

- Address society as a whole
- Make people aware of how algorithms work, what's behind search engines
- Inform developers of their cognitive biases and train them to recognize and correct them in their code

Where does the bias come from?

Developers



YOU

Algorithms

Facebook's EdgeRank

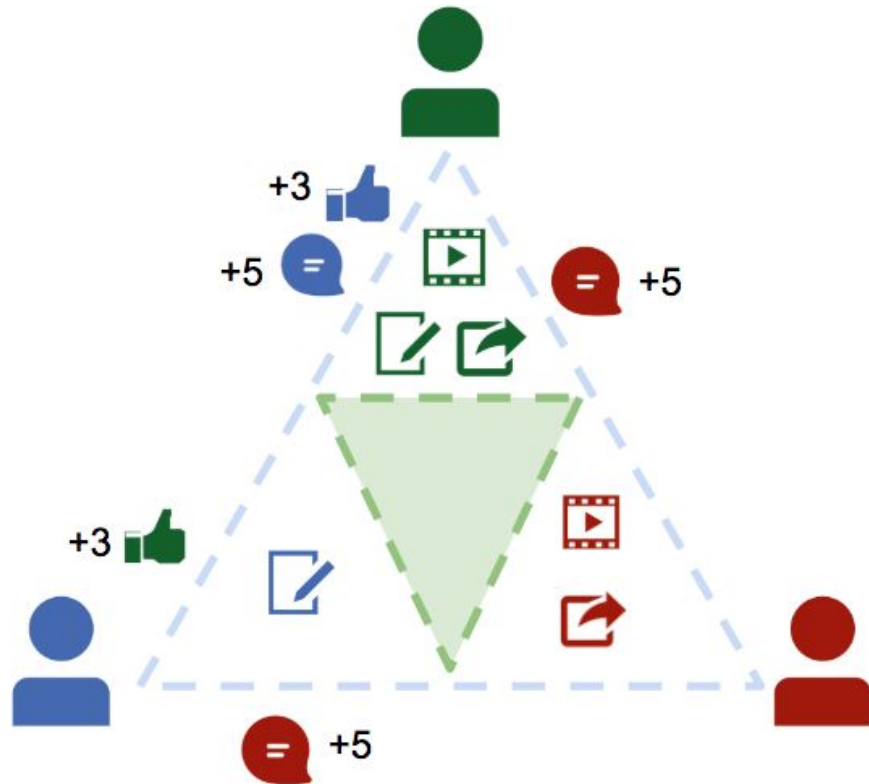
Algorithm

$$\sum_{\text{edges } e} u_e w_e d_e$$

u_e Affinity score between user and edge creator

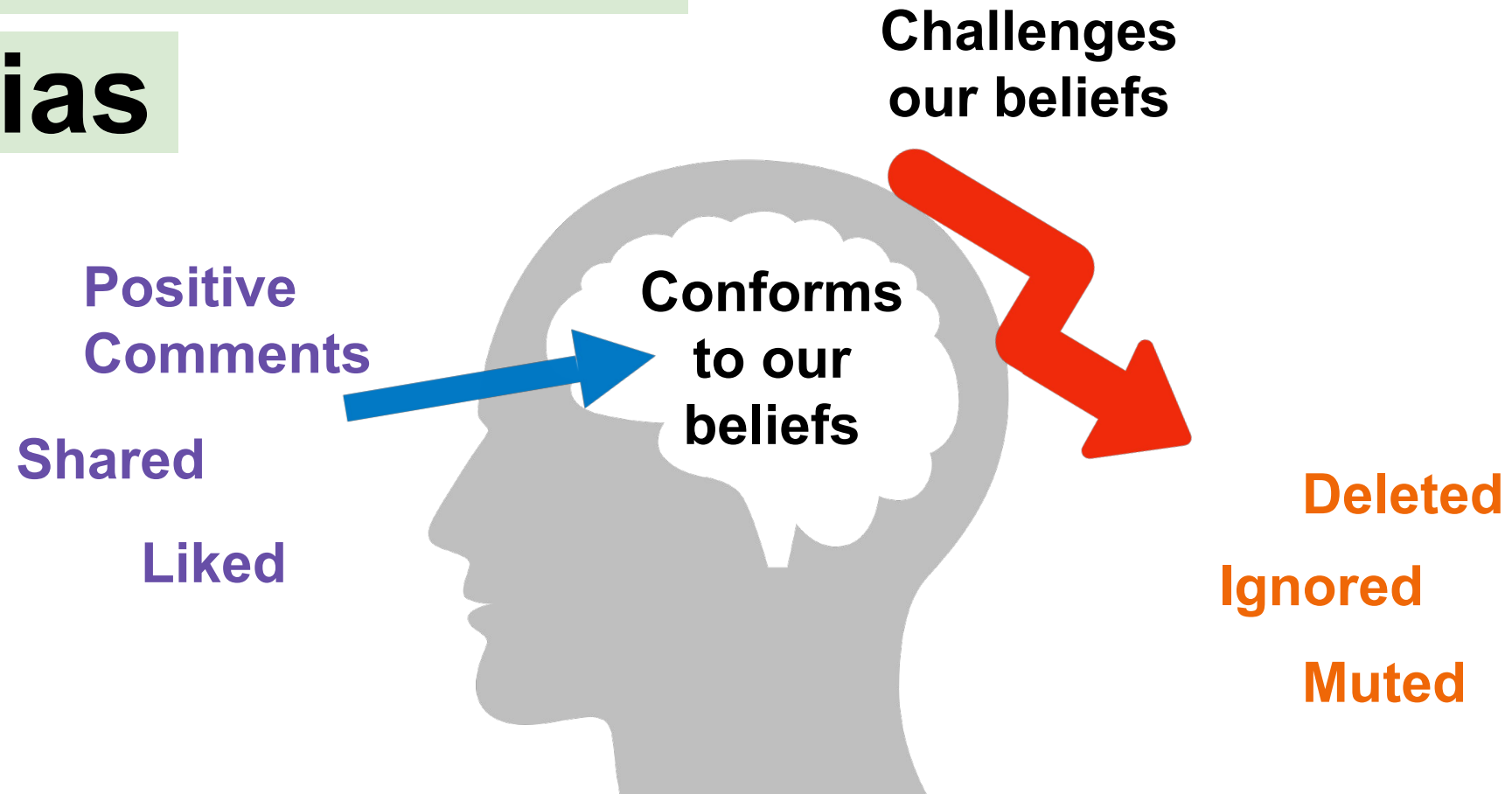
w_e Weight for this edge type (create, connect, like, tag, etc.)

d_e Time decay factor based on how long ago the edge was created



Confirmation

Bias



Possible Solutions

- ❑ Minimize the bad effects and improve the good effects of this technology instead of trying to get rid of it all together
- ❑ Go from purely algorithmic approach to include human curators to help determine trends
- ❑ Create an algorithm that will inform us of who and where the information is coming

Bias on the web

Data bias:

- Geographic
- Economic
- Age
- Gender
- Language

Sampling bias

Algorithmic bias

Selection bias

Activity bias

Presentation bias

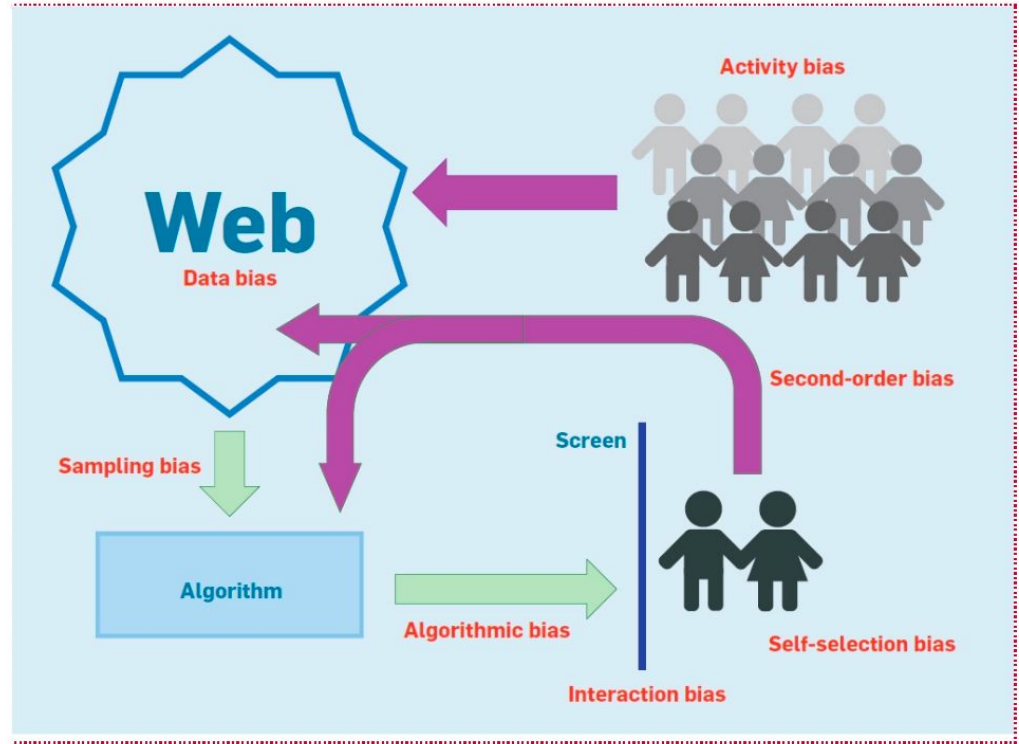
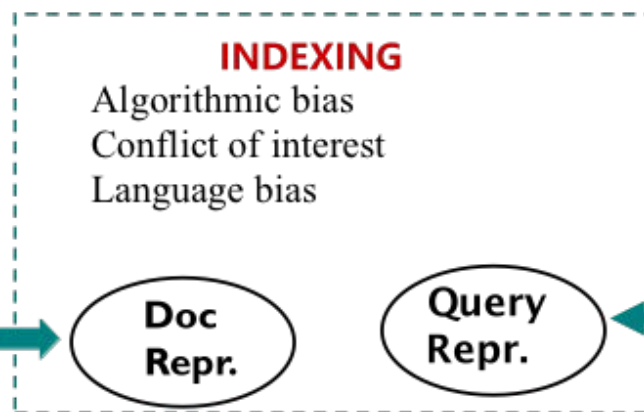


Figure 1. The vicious cycle of bias on the Web.

Potential for bias in a typical IR system

Documents

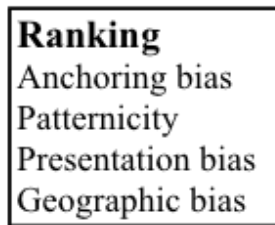
Sampling bias
Activity bias



Query

User

SEARCHING

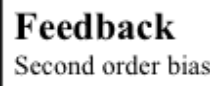


INTERFACE



results

Self selection bias
Confirmation bias



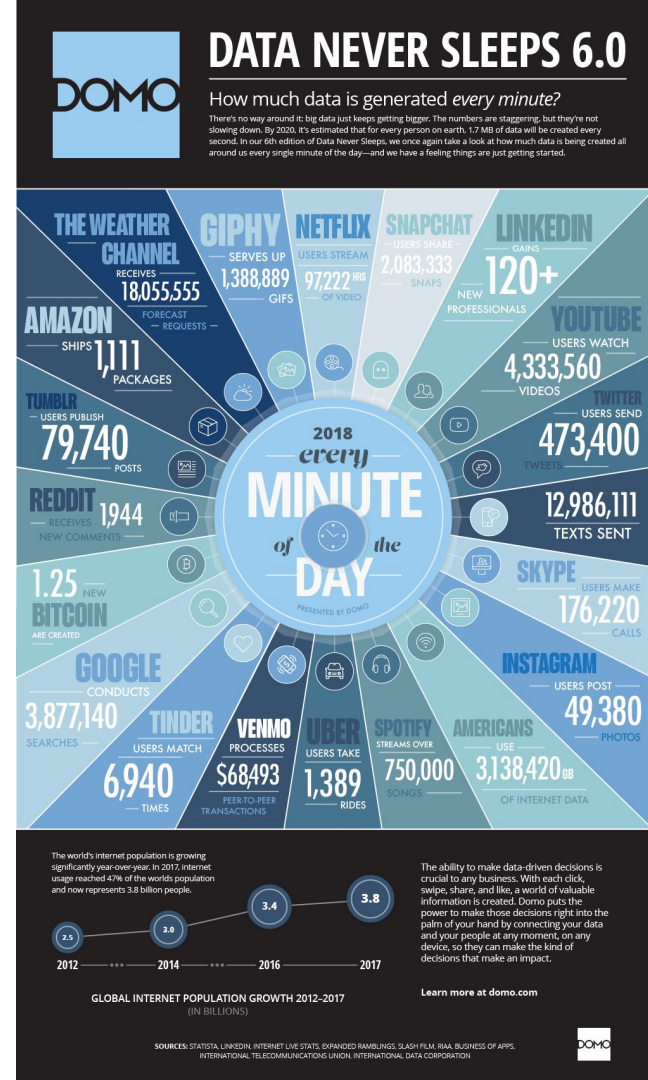
judgments

Activity bias

QUERY MODIFICATION

Data bias

- 90% of existing data was generated in the past 2 years (as of 2017)
- Most new data is unstructured
- There is so much data that no human could possibly interpret it all



Activity Bias online

4% of users provide 50% of Amazon reviews (including paid reviewers)

7% of Facebook users contribute 50% of the content

0.05% Of Twitter's most popular users attract 50% of all users

0.04% of registered Wikipedia editors posted half of Wikipedia's original content

Algorithmic Bias

Motivations behind the behaviors and results ranking of IR systems:

- Precision/recall
- Spam filtration
- Speed
- Sales
- Popularity
- Usability



Decision by algorithm can reflect biases inputted

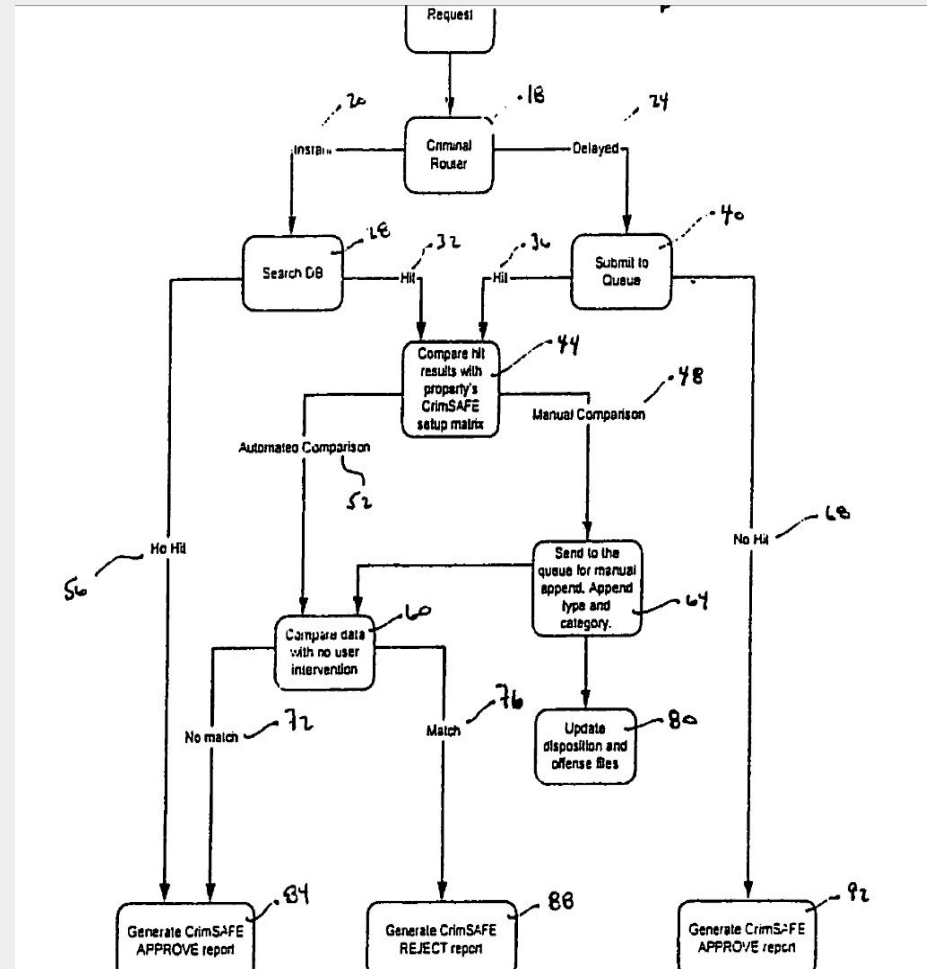
Patent #US20060195351A1:

Internet-based system and method for leasing rental property to a prospective tenant based on criminal history

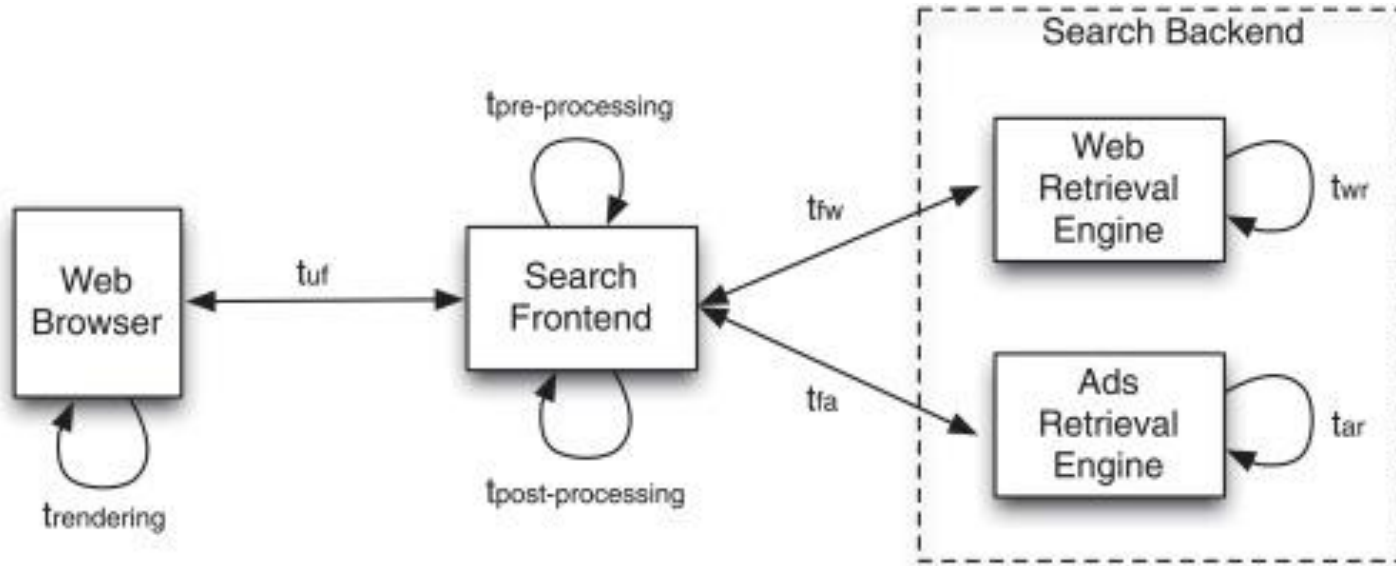
Property manager decides (by crime) whether they would approve an applicant based on specific criminal history.

If African Americans and Hispanics were incarcerated at the same rates as whites, prison and jail populations would decline by almost 40% -NAACP

Algorithms only see the data, not the underlying socioeconomic landscape that creates this disparity



A model of an Search Engine coupled with an 'ads retrieval' system



Google, Bing, Yahoo!, etc. must be effective to be used

Because they are effective they can sell the first result on a search to relevant advertisers

The anchoring bias (and click results) suggests that users will be heavily influenced by the first result



vaccines and autism



All

News

Videos

Images

Maps

More

Settings

Tools

About 47,800,000 results (0.48 seconds)

IR and advertising:

Query: 'vaccines and autism'

IR provides a series of scholarly articles and hospital links demonstrating that vaccines have no statistical link to autism

Ads Retrieval provides a link to a page the 'autism epidemic' and vaccines as the first result

Vaccines And Autism | Mercury & Autism Correlation

(Ad) www.childrenshealthdefense.org/Mercury-Related/Autism ▼

Study Shows The Relation Of Mercury & Autism. Read More. The New Study Shows That A Lone FDA Scientist Could End the Autism Epidemic. Research & Resources. Founded in Malibu, CA.

Scholarly articles for vaccines and autism

Vaccines and autism: a tale of shifting hypotheses - Plotkin - Cited by 269

Vaccines and autism - DeStefano - Cited by 8

Vaccines and autism - Rimland - Cited by 9

Vaccines & Autism – Science-Based Medicine

<https://sciencebasedmedicine.org/reference/vaccines-and-autism/> ▼

Overview of **Vaccines & Autism**. **Vaccines** are generally considered to be the most successful public health intervention ever devised. And yet there have been ...

Vaccines and Autism | Children's Hospital of Philadelphia

<https://www.chop.edu/centers-programs/vaccine...center/vaccines.../vaccines-autism> ▼

Two studies have been cited by those claiming that the MMR **vaccine** causes **autism**. ... Wakefield's hypothesis was that the measles, mumps and rubella (MMR) **vaccine** caused a series of events that include intestinal inflammation, entrance into the bloodstream of proteins harmful to the ...

Are Vaccinations Linked To Autism? The Latest Science Explained

<https://www.webmd.com/brain/autism/do-vaccines-cause-autism> ▼

Data bias in indexing

- The 1998 retracted and debunked study regarding vaccines and autism still permeates the web and has led to refusal of the Measles, Mumps, and Rubella vaccine



"I'M SORRY DOCTOR, BUT AGAIN I HAVE TO DISAGREE."

Fact selection and presentation affects decisions

Data Bias

1983 10 vaccines given to U.S. babies
autism rate 1 in 10,000

2008 36 vaccines given to U.S. babies
autism rate 1 in 150

2013 46 vaccines given to U.S. babies
autism rate 1 in 88

NaturalNews.com
America's truth news bureau, read by 7+ million and growing



Vaccinate with confidence

The rise in autism cases is not from vaccines

PRE-1980 No separate classification for autism diagnosis. "Autistic" only used as a term to describe childhood schizophrenia.

1983 Rigid definitions of autism. Diagnostic criteria not consistent. Actual rates not meaningful or reliable.

1994 Changes in the DSM-IV. Broader definitions. Aspergers syndrome is included. Explosion of "new" diagnoses.

2008 Better broader, earlier diagnosis. Rediagnosing old cases. Statistics catching up

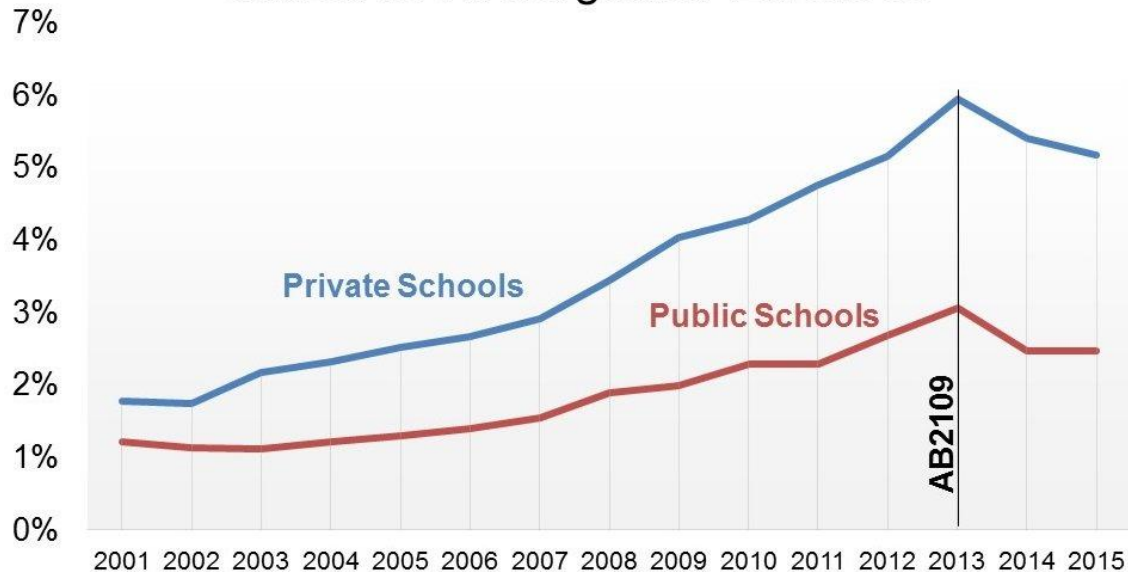
2013 Better awareness, educational resources and interventional funding. More changes of the DSM.

The Vaccine Meme Machine www.unstrange.com Refutatio

Correlation does not equal causation

California limited the rate of increased vaccination exemptions by requiring a medical consult first - forcing people to talk to an expert

Vaccine Mandate Exemptions,
California Kindergarten Enrollees



Possible Solutions

- √ More **awareness** and **education** for users and developers
- √ More **transparency** from Search Engines
- √ Better **checks & balances**
- √ Better data sampling
- √ Clearer representations of data and where it came from

??? Ideas & Questions ???

Bibliography

- Bias on the Web, Ricardo Baeza-Yates - <https://cacm.acm.org/magazines/2018/6/228035-bias-on-the-web/fulltext>
- Assessing Bias in Search Engines, Abbe Mowshowitz, Akira Kawaguchi - <https://www.sciencedirect.com/science/article/pii/S0306457301000206>
- https://web-assets.domo.com/blog/wp-content/uploads/2017/07/17_domo_data-never-sleeps-5-01.png
- How Much Data Do We Create Every Day? The Mind-blowing Stats Everyone Should Read, Bernard Marr - <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#4b62203660ba>
- A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews, Tsur, O. Davidov, D. & A. Rappoport. 2010. ICWSM— Proc. ICWSM 2010, 162- 169
- Impact of response lay on sponsored search - <https://www.sciencedirect.com/science/article/pii/S0306457317302923>
- Criminal Justice Fact Sheet, <https://www.naacp.org/criminal-justice-fact-sheet/>
- Search Engine Market Share <https://www.netmarketshare.com/search-engine-market-share.aspx>
- Vaccine Opt-outs Dropped Slightly When California Added More Hurdles - Eric Hamilton - <https://medicalxpress.com/news/2018-09-vaccine-opt-outs-slightly-california-added.html>
- Bias in algorithmic filtering and personalization, Bozdag, E. <https://link-springer-com.libproxy.lib.unc.edu/article/10.1007%2Fs10676-013-9321-6#enumeration>
- My News Feed is Filtered?, Powers, E. (2017) Digital Journalism, 5:10, 1315-1335,DOI: [10.1080/21670811.2017.1286943](https://doi.org/10.1080/21670811.2017.1286943)
- Using “public interest algorithms” to tackle the problems created by social media algorithms, Wheeler, Tom - <https://www.brookings.edu/blog/techtank/2017/11/01/using-public-interest-algorithms-to-tackle-the-problems-created-by-social-media-algorithms/>