

Bias in the Web

Ricardo Baeza-Yates



Wien, June 2017





My greatest fear is that people
will attribute fakes quotes to me
and
millions of people on the Internet will believe it

Fake Content & Bias

- British Prime Minister Benjamin Disraeli (19th century):
 - "There are three kinds of **lies**: **lies**, damned **lies**, and **statistics**."

UTC professor says "Everyone has bias"

BY HANNAH LAWRENCE | FRIDAY, JULY 8TH 2016



We all have biases and preconceptions about certain subjects or groups of people according to one Chattanooga researcher.

Buzzfeed News

TOP POST

173,877 VIEWS



Here Are 50 Of The Biggest Fake News Hits On Facebook From 2016

One fake news entrepreneur says we should expect even more Trump hoaxes in 2017

posted on Dec. 30, 2016, at 2:12 p.m.



Craig Silverman
BuzzFeed News Media Editor

Bias: significant deviation from a prior (unknown) distribution



So (Observational) Human Data has Bias

Goal: Bias Awareness

- Gender
- Racial
- Sexual
- Religious
- Social
- Linguistic
- Geographic
- Political
- Educational
- Economic
- Technological

- from Noise or Spam
- Validity (e.g. temporal)
- Completeness
- Gathering process
-

Attempt of an unbiased (personal) view on bias in the Web

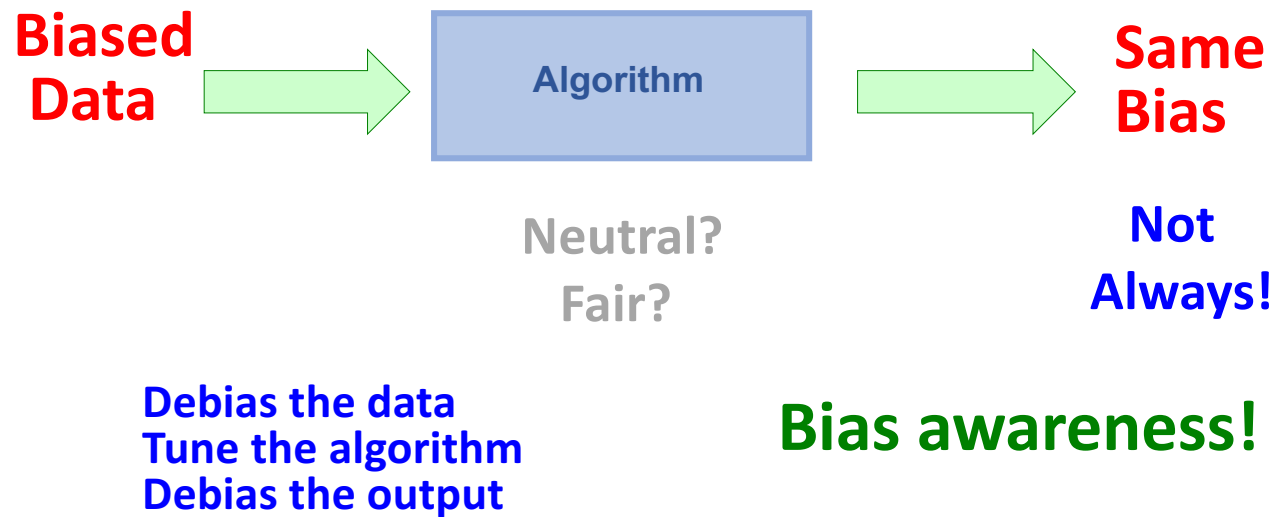
Many people extrapolate results of a sample to the whole population (e.g., social media analysis)

In addition there is bias when measuring bias as well as bias towards measuring it!

A Non-Technical Question



A Non-Technical Question

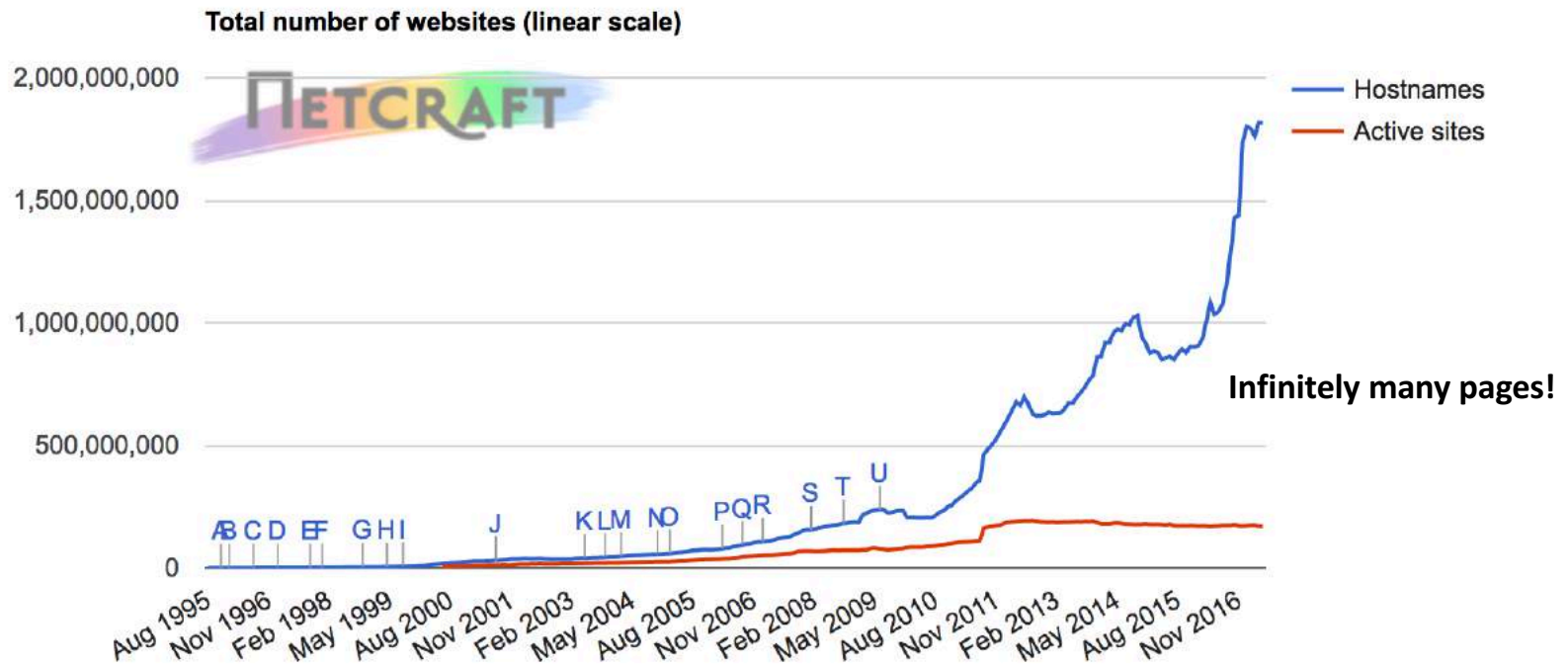


Big Data and Bias

- The quality of any algorithm is bounded by the quality of the data that uses
- Data bias awareness
[Gordon & Desjardins; Provost & Buchanan, MLJ 1995]
- Algorithmic fairness
- Key issues for Machine Learning
 - Uniformity of data properties
 - In the Web, distributions resemble a power law
 - Uniformity of error
 - Data sample methodology
 - E.g., sample size to see infrequent events or sampling bias

WWW

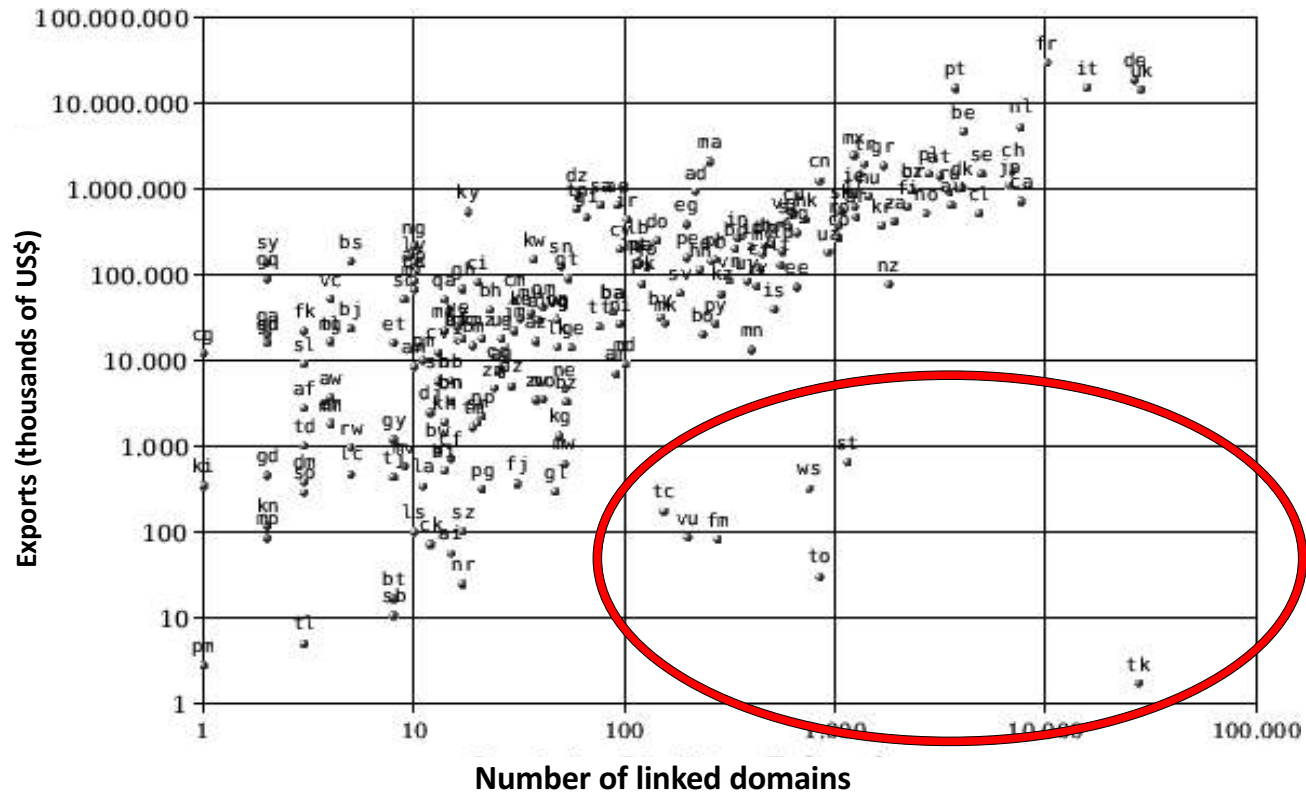
Social Media



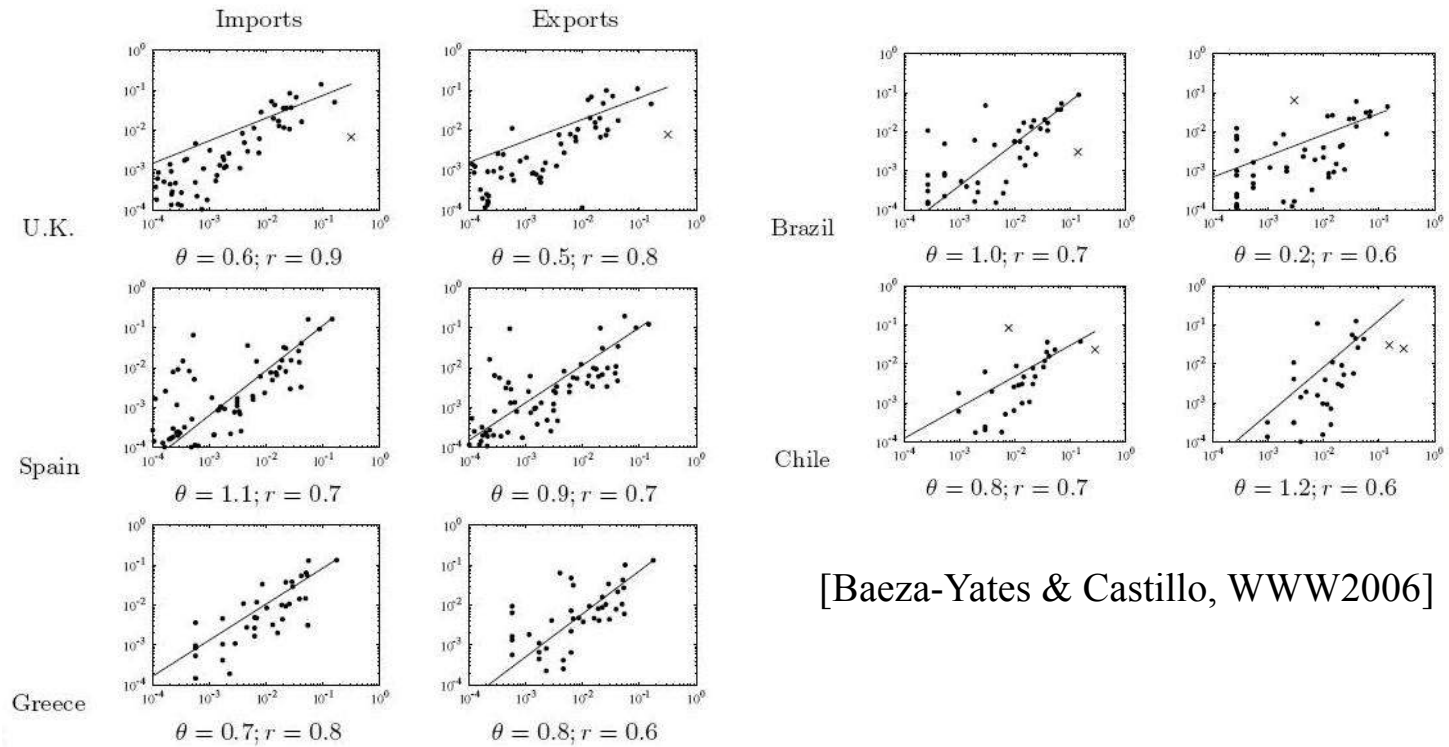
Bias in the Web



Economic Bias in Links



Economic Bias in Links

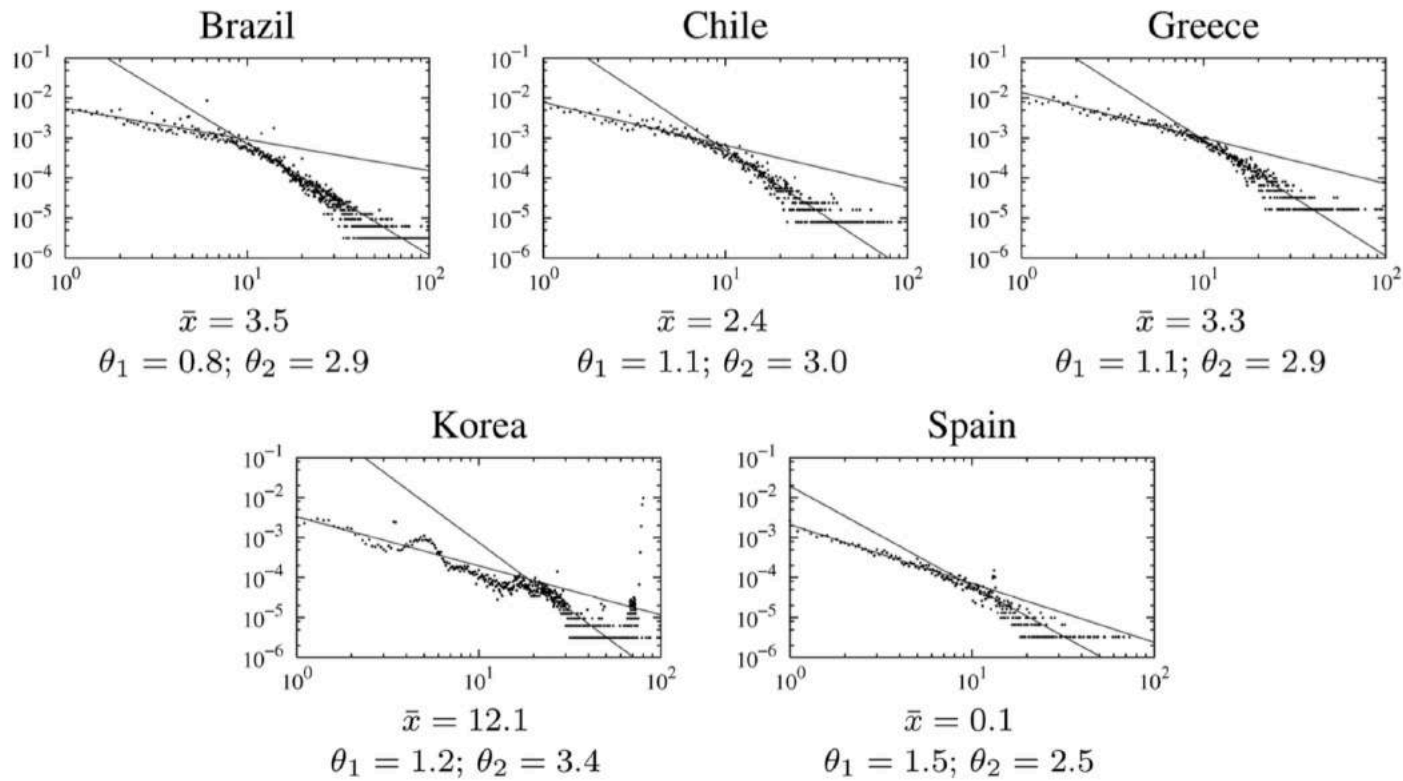


[Baeza-Yates & Castillo, WWW2006]

Cultural Bias in Website Structure

Shame

Minimal effort



[Baeza-Yates, Castillo, Efthimiadis, TOIT 2007]

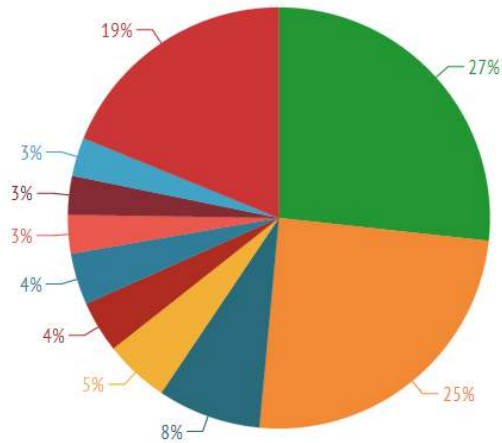
Linguistic Bias in Content

Top 25 World Languages

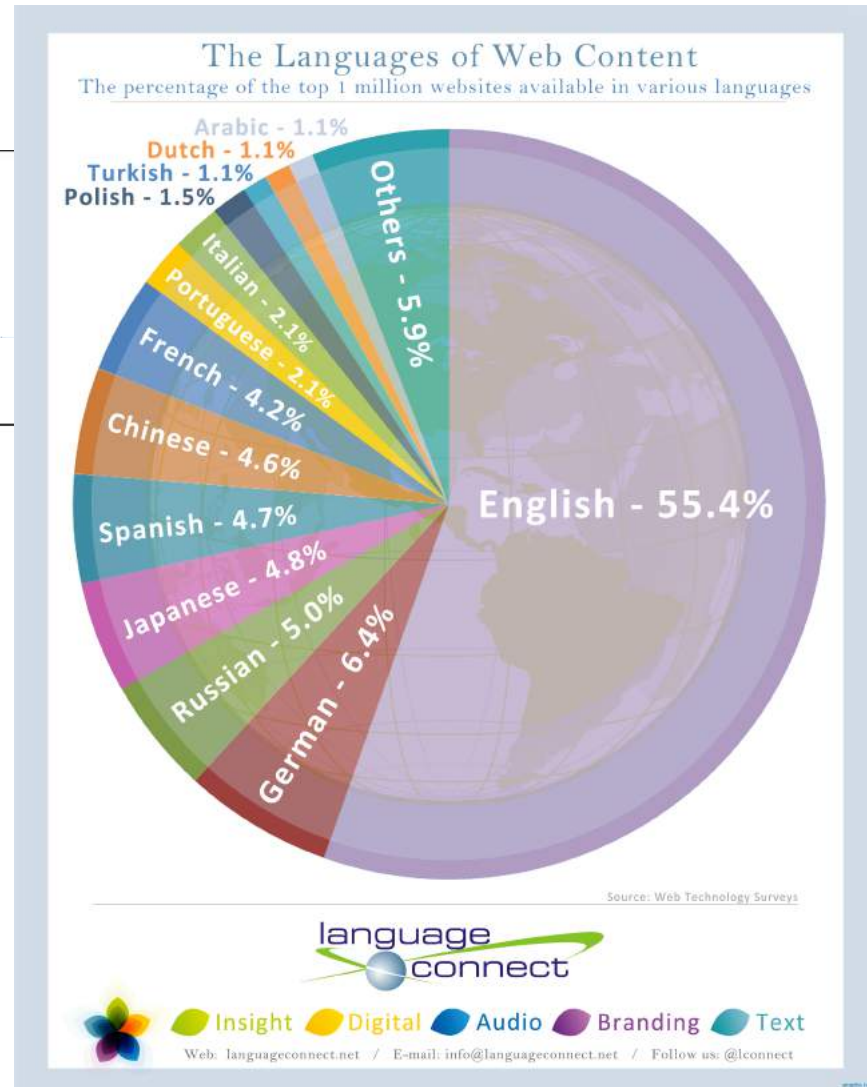
- Chinese, Mandarin
- Spanish
- English
- Hindi

Top Ten Languages in the Internet
in millions of users - November 2015

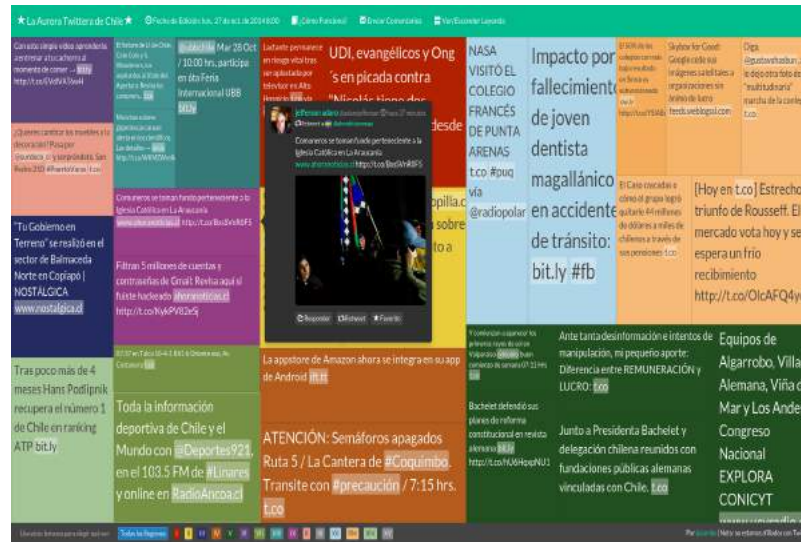
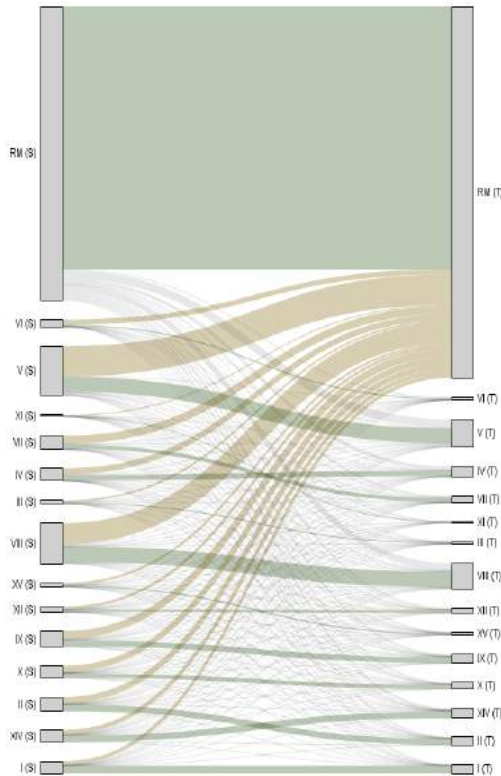
Languages on the Web



- | | | | | |
|---------|------------------|---------|----------|------------|
| English | Chinese Mandarin | Spanish | Japanese | Portuguese |
| German | Arabic | French | Russian | Other |



Geographical Bias in Content



[E. Graells-Garrido and M. Lalmas, “Balancing diversity to counter-measure geographical centralization in microblogging platforms”, ACM Hypertext’14]

Gender Bias in Content

- Word embedding's in w2vNEWS

Gender stereotype *she-he* analogies.

| | | |
|---------------------|-----------------------------|---------------------------|
| sewing-carpentry | register-nurse-physician | housewife-shopkeeper |
| nurse-surgeon | interior designer-architect | softball-baseball |
| blond-burly | feminism-conservatism | cosmetics-pharmaceuticals |
| giggle-chuckle | vocalist-guitarist | petite-lanky |
| sassy-snappy | diva-superstar | charming-affable |
| volleyball-football | cupcakes-pizzas | hairdresser-barber |

Gender appropriate *she-he* analogies.

| | | |
|-----------------|--------------------------------|-------------------|
| queen-king | sister-brother | mother-father |
| waitress-waiter | ovarian cancer-prostate cancer | convent-monastery |

Most journalists are men?

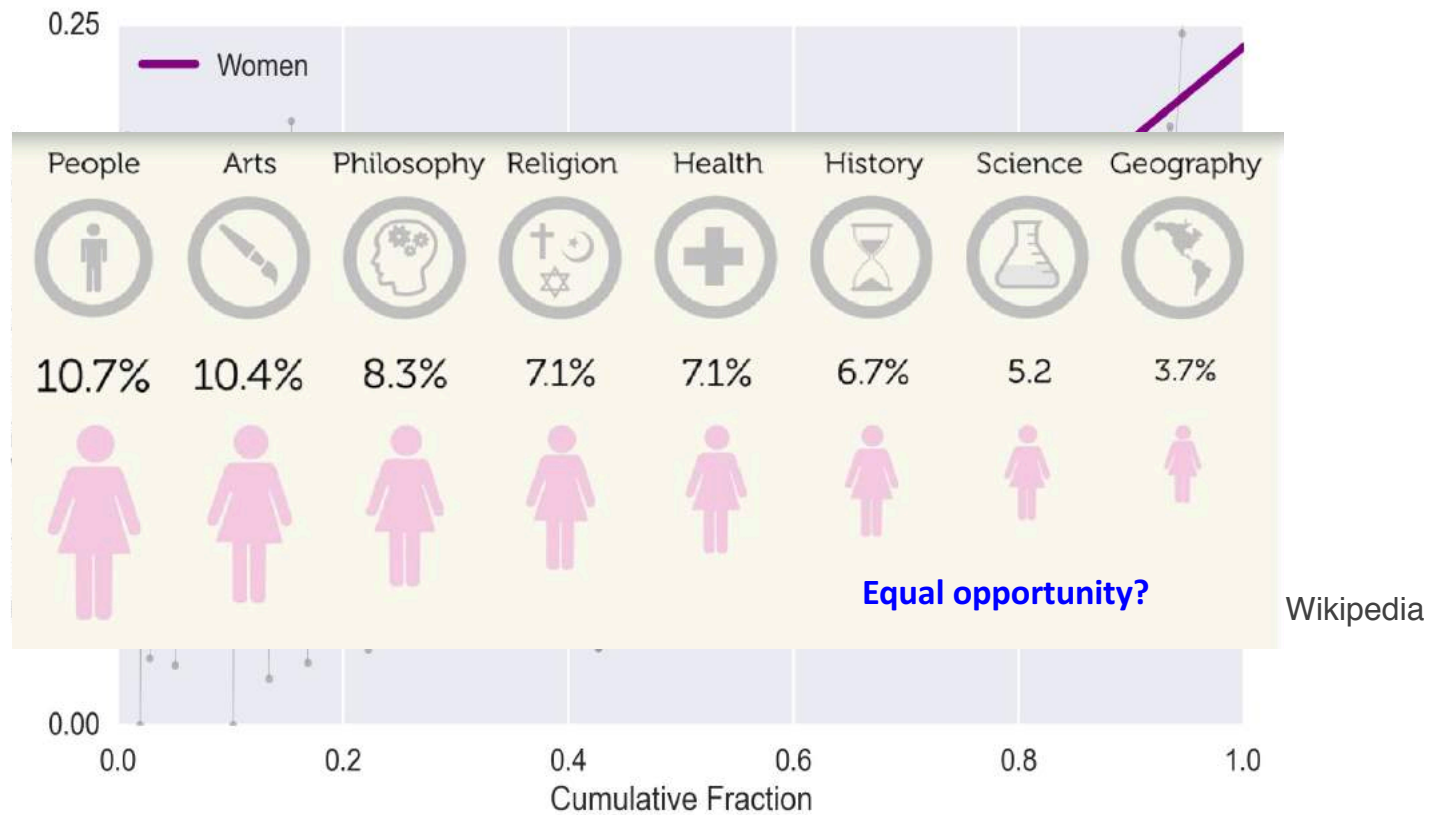
Yes, about 60 to 70% at work
although at college is the inverse

[Bolukbasi et al, NIPS 2016]



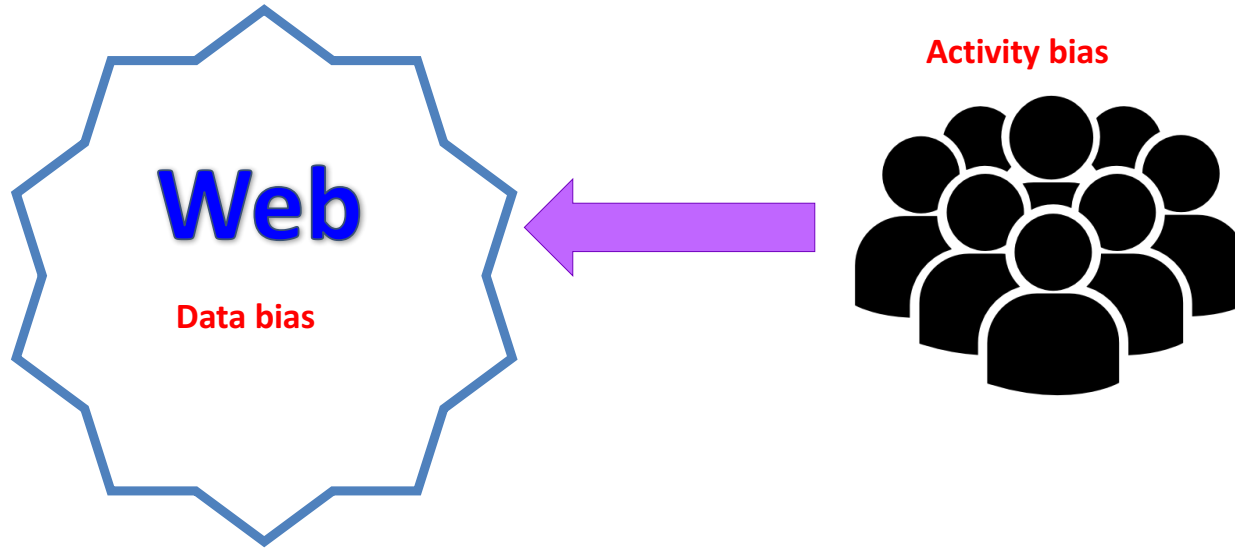
Gender Bias in Content

Systemic bias?



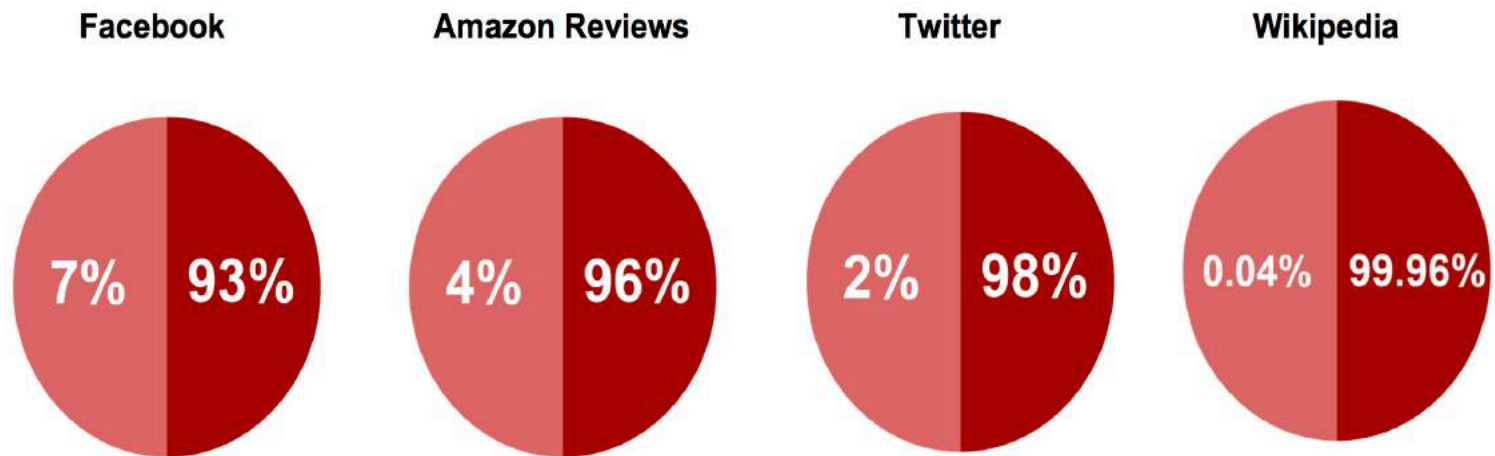
[E. Graells-Garrido et al., "First Women, Second Sex: Gender Bias in Wikipedia", ACM Hypertext'15]

Bias in the Web



Activity Bias

Which percentage of users produce 50% of the content?



[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]

Amazon sues 1,000 'fake reviewers'

October 2015

Online retailer files lawsuit in US against people whose names it says it does not know, claiming they offer reviews for sale

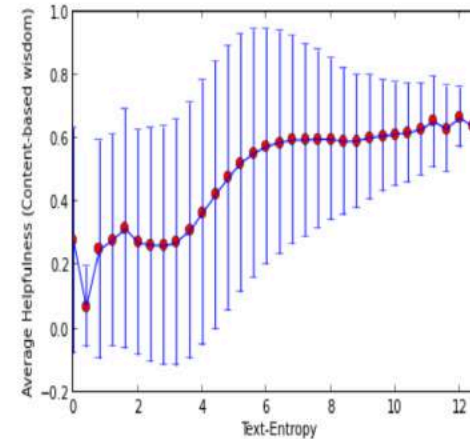
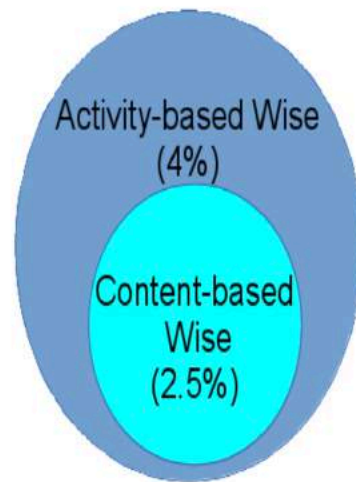
Amazon Continues Their Crusade Against Fake Reviews

By Tyler Lee on 04/26/2016 05:07 PDT



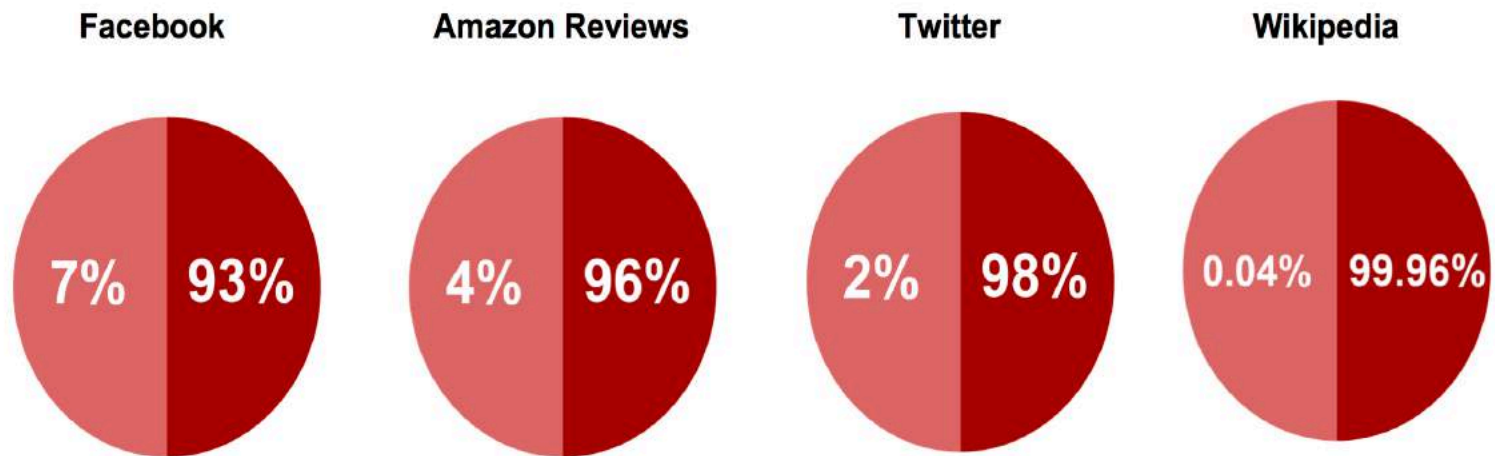
Quality of Content?

- Adding content implies adding wisdom?
- We used Amazon's reviews helpfulness and computed the text entropy
- Content-based-wise users
- How many of those users are being paid?



Activity Bias

Which percentage of users produce 50% of the content?

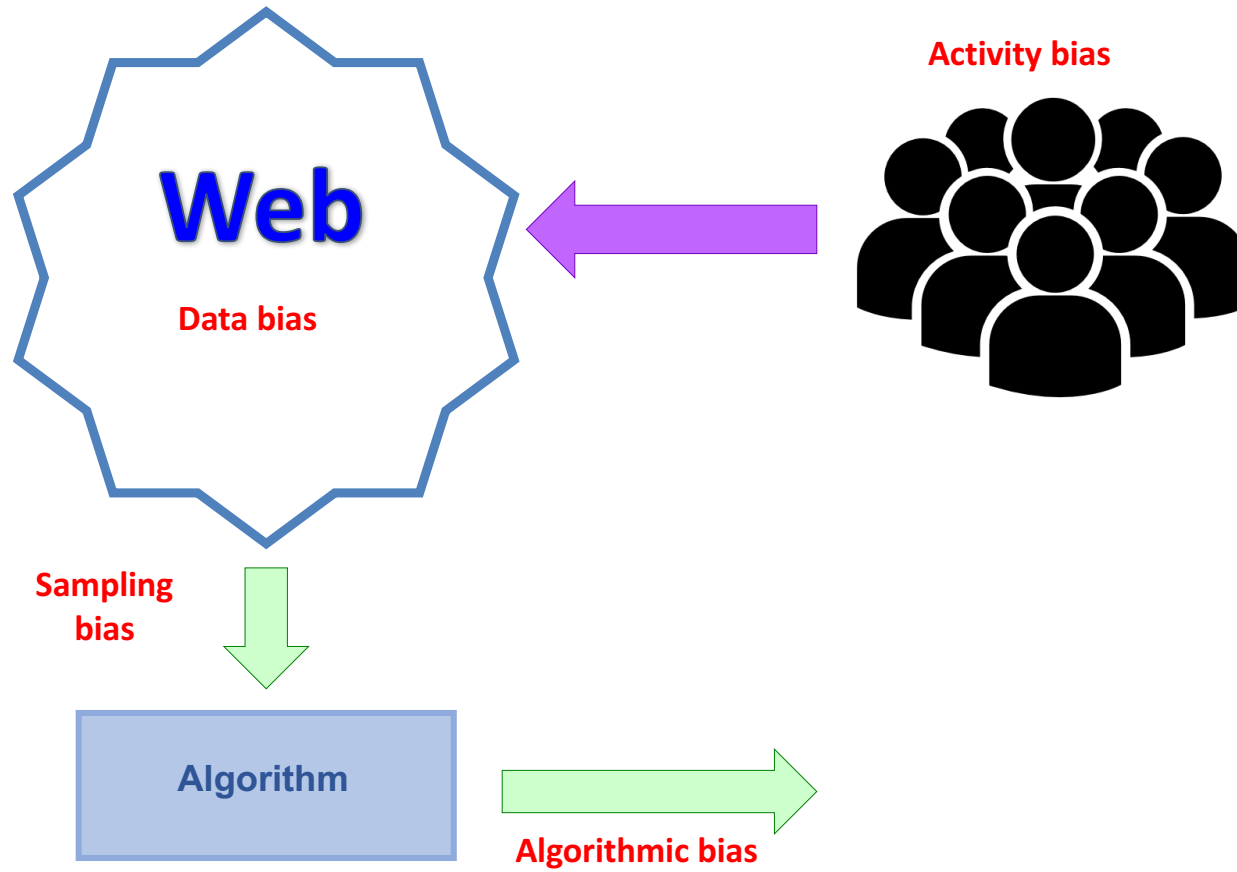


[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]

Content that is never seen: Digital Desert

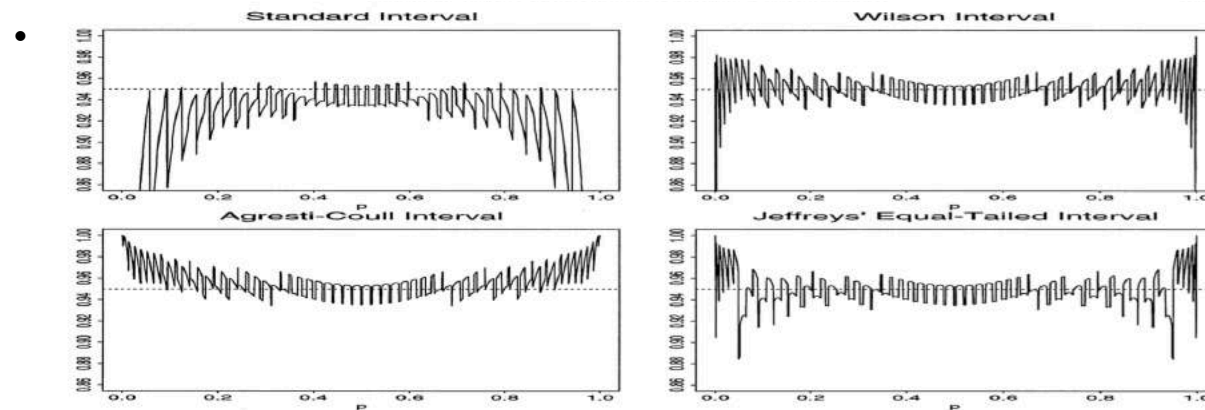
- 1.1% of the Twitter content is never seen.*
- 31% of articles added/edited in May 2014 in wikipedia, were not visited in June.





Sample Size?

- If we want to estimate the frequency of queries that appear with probability at least p with a certain relative error ϵ we can use the standard binomial error formula $\sqrt{(1-p)/np}$ which works well for p near $\frac{1}{2}$ **but not for p near 0**



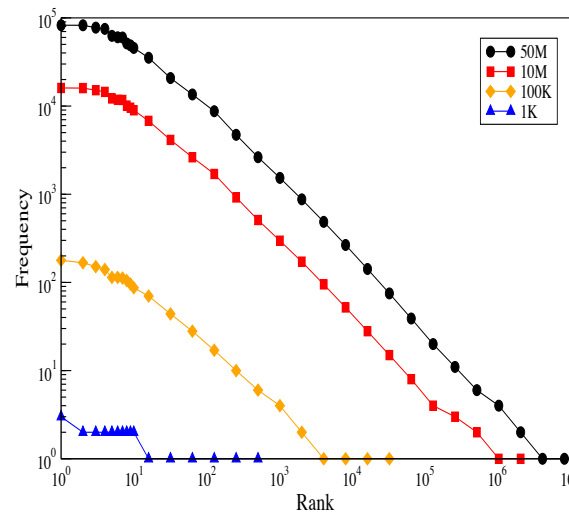
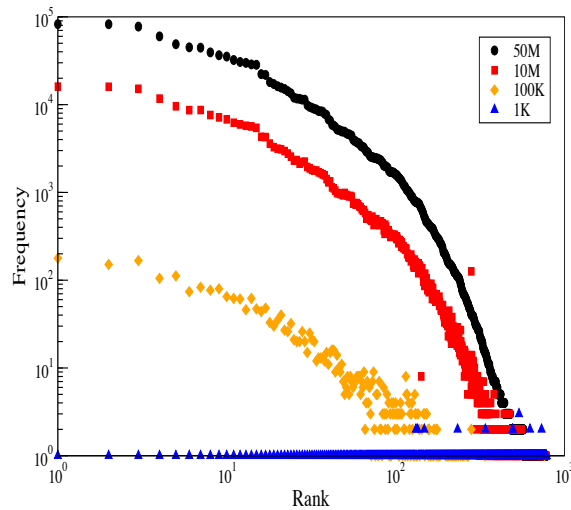
- If $p = 0.1$, $1 - \alpha$ is 90% and ϵ is 10%, we get $n = 2342$.
The standard formula gives $n = 900$!



[Brown, Cai & DasGupta, Statistical Science, 2001]
[Baeza-Yates, SIGIR 2015, Industry track]

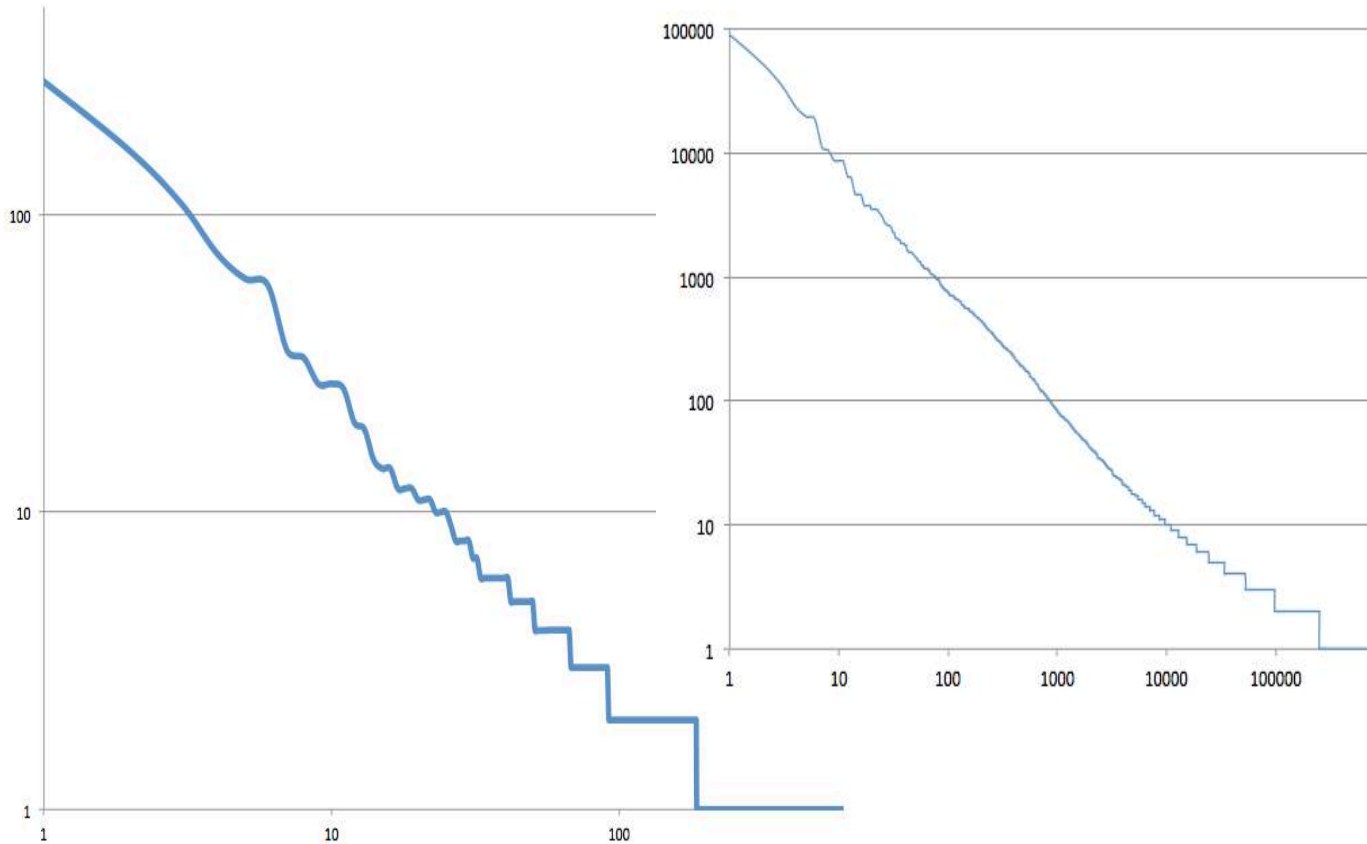
Sampling Techniques

- Standard technique: $p_q \approx \hat{p}_q(\mathcal{S}) = \frac{f_q(\mathcal{S})}{\sum_{q' \in \mathcal{S}} f_{q'}(\mathcal{S})}$
- A good sample should cover well all the query distribution but this does not work with very biased distributions.




[Zaragoza et al, CIKM 2010]

Stratified Sampling Example



Extreme Algorithmic Bias

London Eye



London Eye and Golden Jubilee Bridge seen from Westminister Bridge.

Tag list

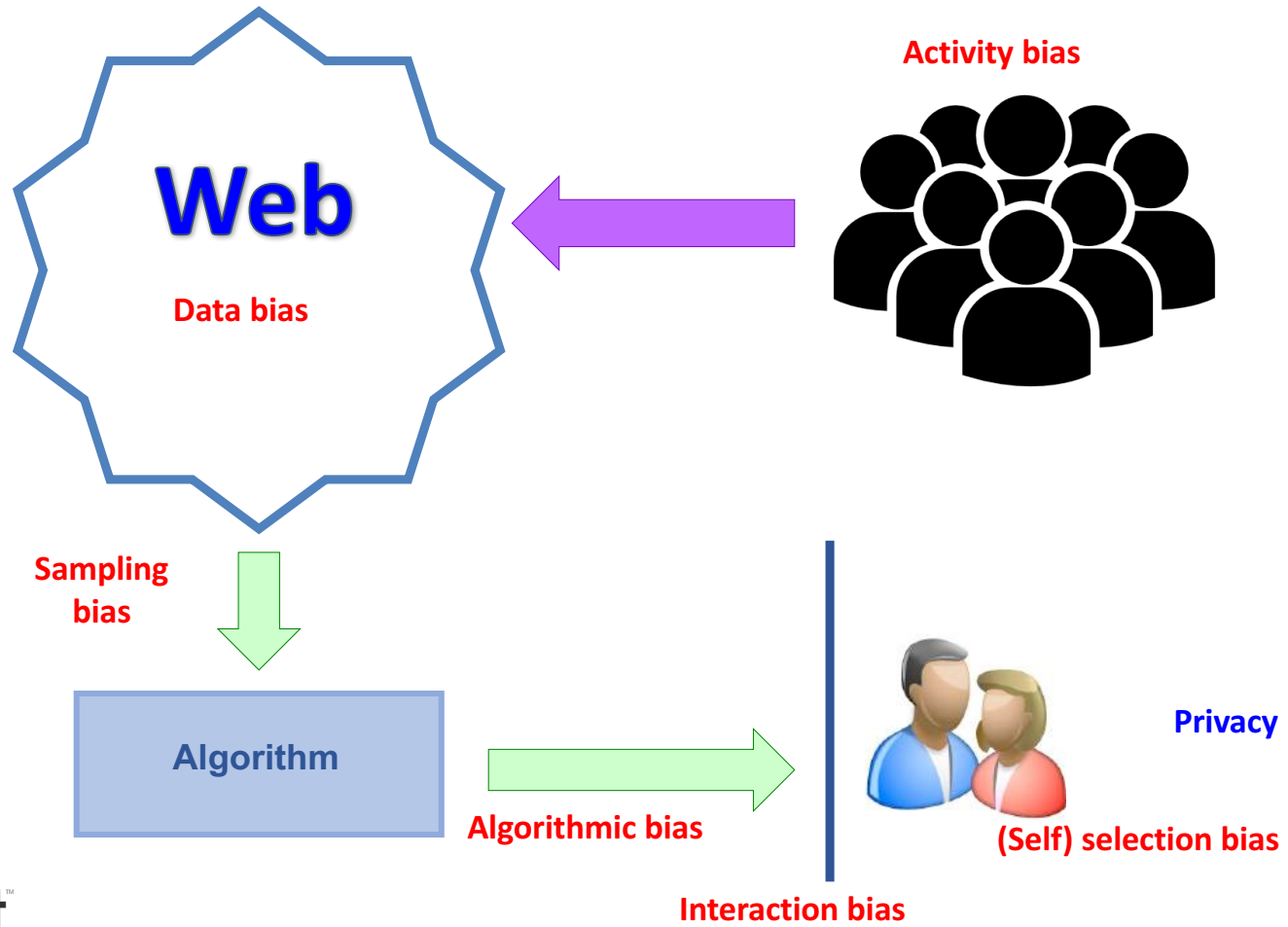
london eye, thames,

Suggested tags

- london
- england
- uk
- river
- eye
- south bank
- big ben
- night
- bridge
- 2006

Update annotation

Bias in the Web



Bias in the Interaction

Related Searches: [tennis racket](#), [tennis shoes](#).

Shop by Category



Position bias
Ranking bias

Presentation bias



Wilson Sporting Goods Championship Extra Duty Tennis Balls (1-Can)

Jun 14, 2012
by Wilson

\$2.79 ~~\$6.99~~ [Add-on Item](#)

Add to a qualifying order to get it by **Tomorrow, May 6.**

More Buying Choices
\$0.99 new (18 offers)
\$7.99 used (2 offers)

[See newer version](#)

★★★★★ 186

Sports & Outdoors: See all 60,449 items

Social bias



Best Seller

Wilson 75 Tennis Ball Pick Up Hopper

by Wilson

\$19.96 [Prime](#)

Get it by **Tomorrow, May 6**

More Buying Choices
\$18.88 new (11 offers)
\$35.00 used (1 offer)

★★★★☆ 319

Product Features
Holds 75 tennis balls with a special no spill lid (Tennis Balls NOT included)

Sports & Outdoors: See all 60,449 items

Interaction bias

Sponsored ⓘ



Tennis Elbow Brace with Gel Comp...

\$24.50 [Prime](#)

★★★★★ 7



DIMANKA Professional Table Tennis...

\$34.99

★★★★★ 9



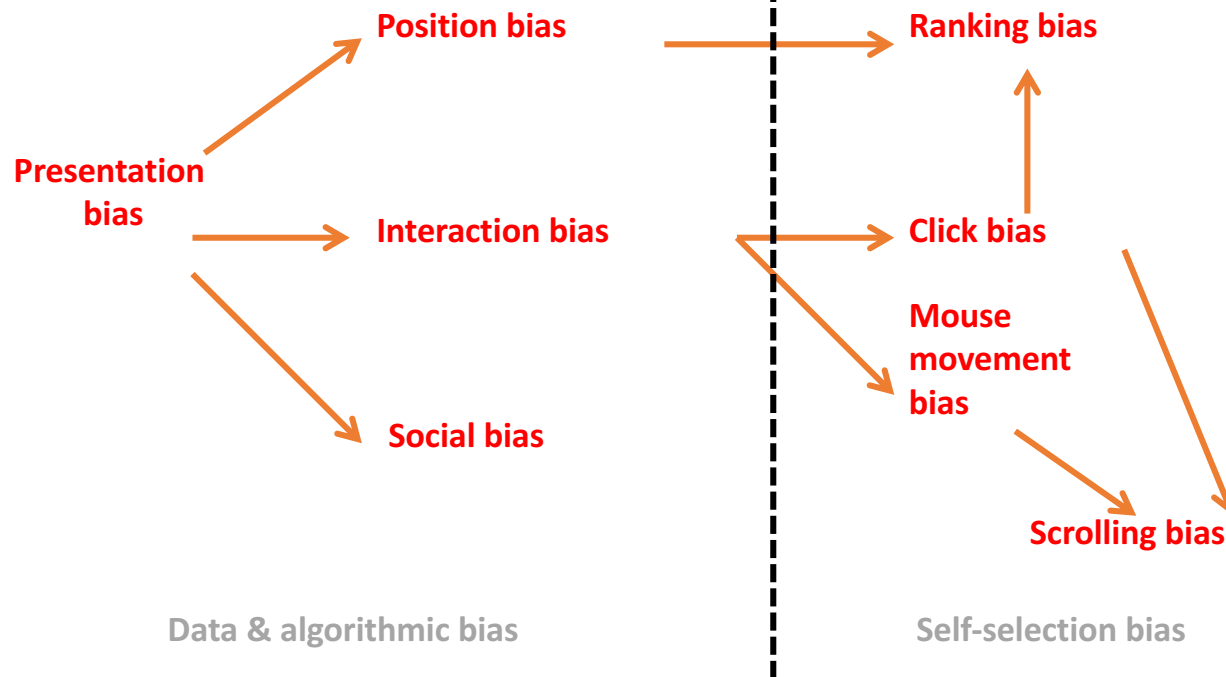
Gamma Quick Kids 78 Ball (12 Pac...

\$19.99 [Prime](#)

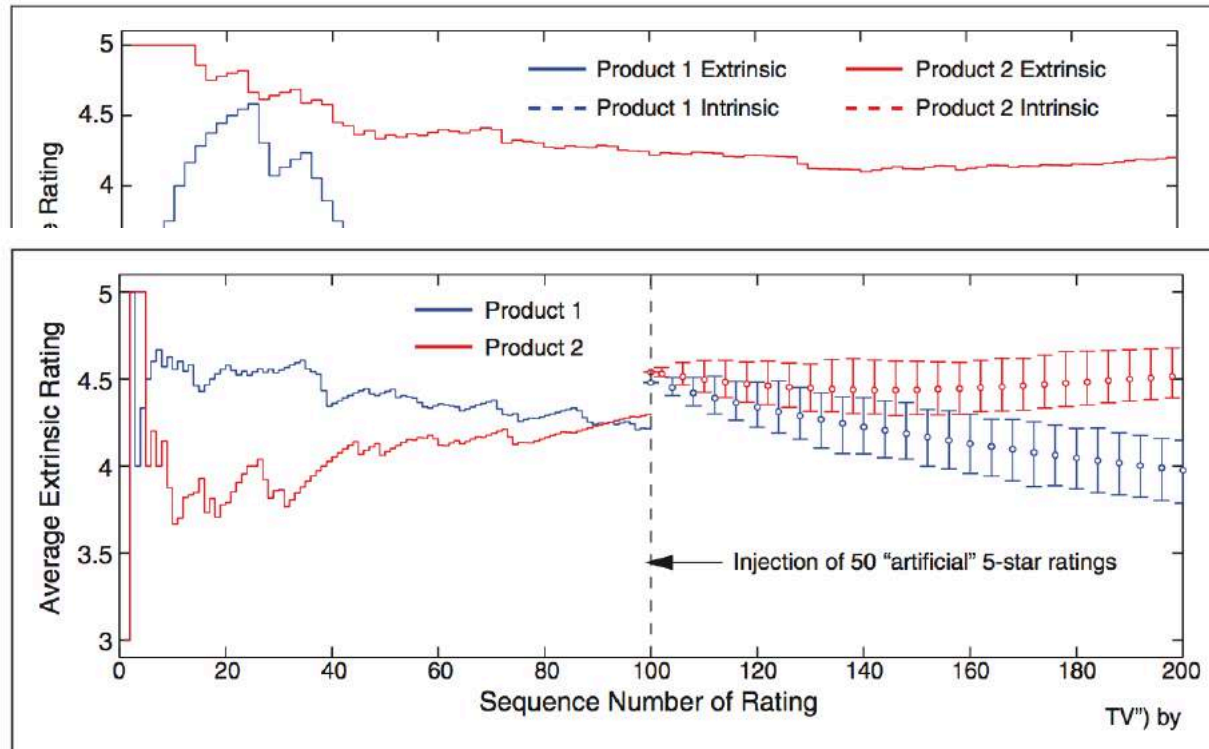
★★★★☆ 44

Amazon.com

Dependencies: A Cascade of Biases!



Social Bias

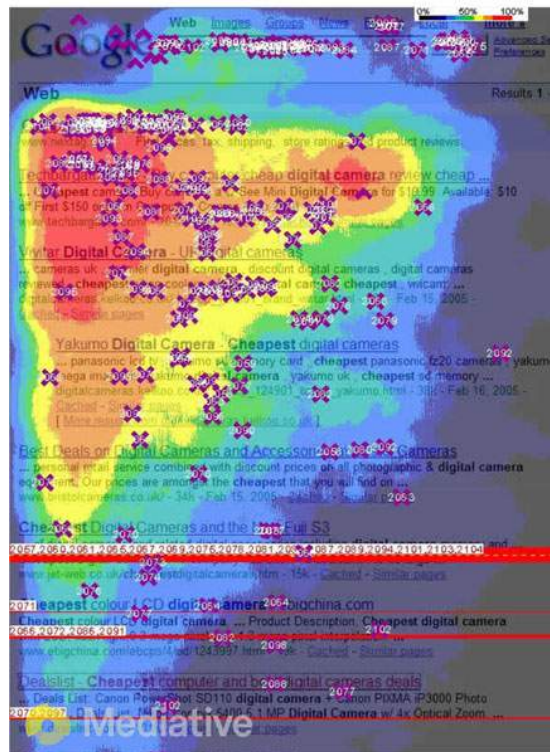


[WHY AMAZON'S RATINGS MIGHT MISLEAD YOU; The Story of Herding Effects
Ting Wang and Dashun Wang, Big Data, 2014]

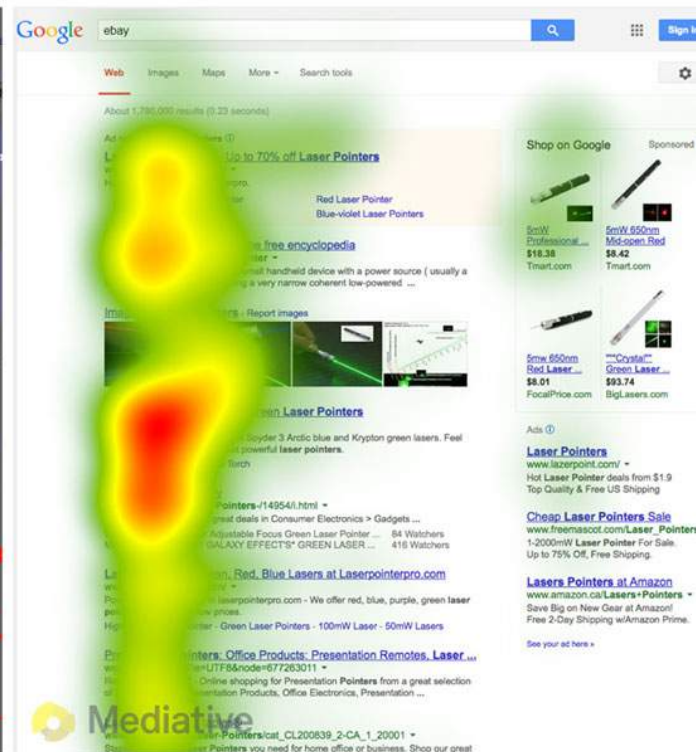


Ranking Bias in Web Search

2005



2014



[Mediative Study, 2014]

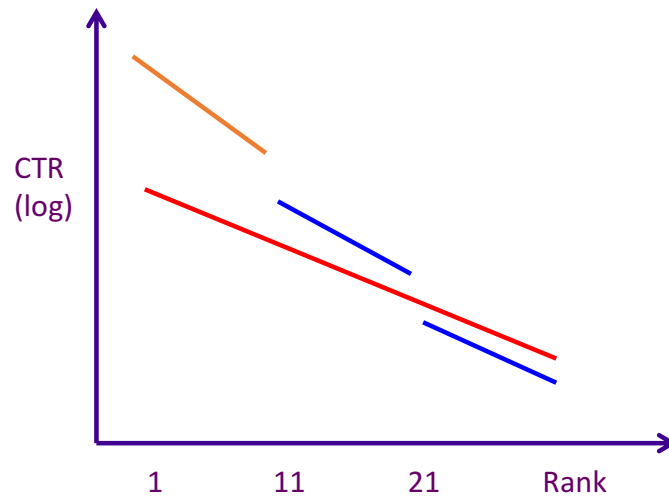
Click Bias in Web Search

- Ranking & next page bias



Debiasing Search Clicks

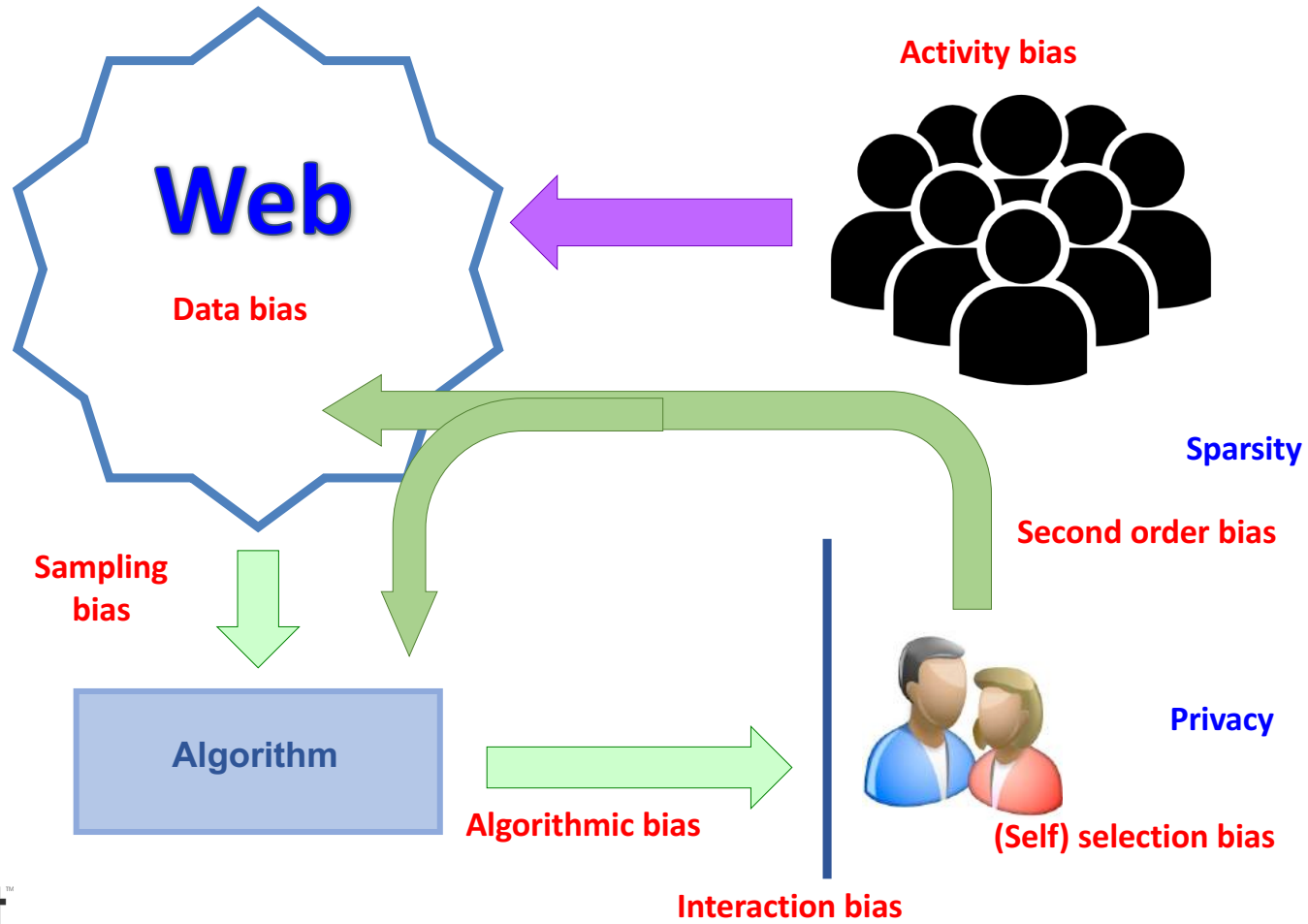
Clicks as implicit positive user feedback



Learning to Rank with bias
[Joachims et al, WSDM 2017, **best paper**]

[Dupret & Piwowarski, SIGIR 2008]
[Chapelle & Zhang, WWW 2009]

Bias in the Web



Avoid Second Order Bias due to Personalization

The Filter “Bubble”, Eli Pariser (2011)

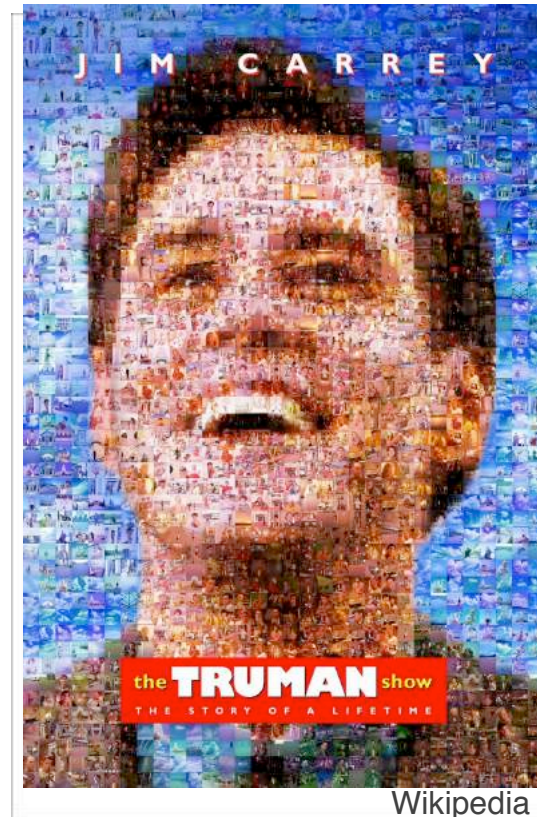
- The effect of self selection bias
- Avoid the poor get poorer syndrome
- Avoid the echo chamber
- Empower the tail

Partial solutions:

- Diversity
- Novelty
- Serendipity
- Show me the dark side

Cold start problem solution: Explore & Exploit

How much exploration is needed for
presentation bias?



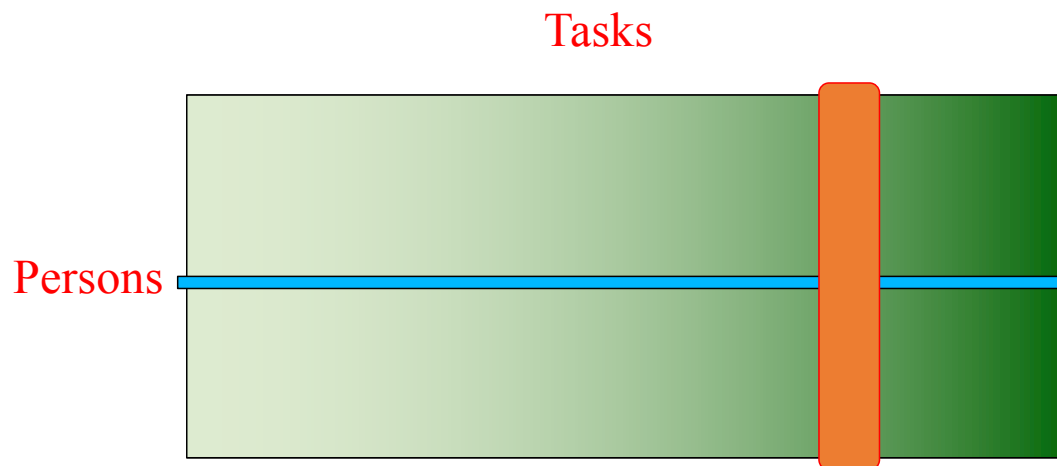
Aggregating in the Tail

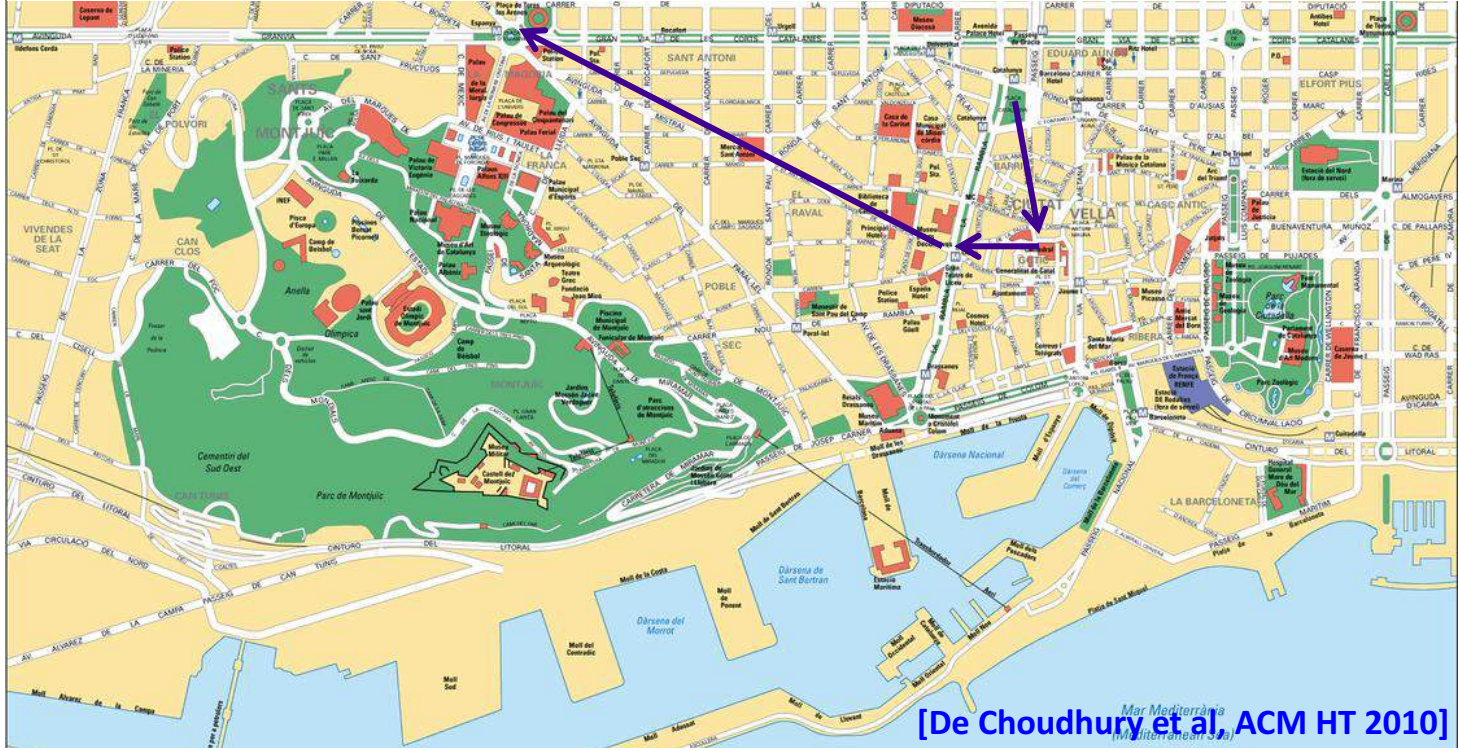
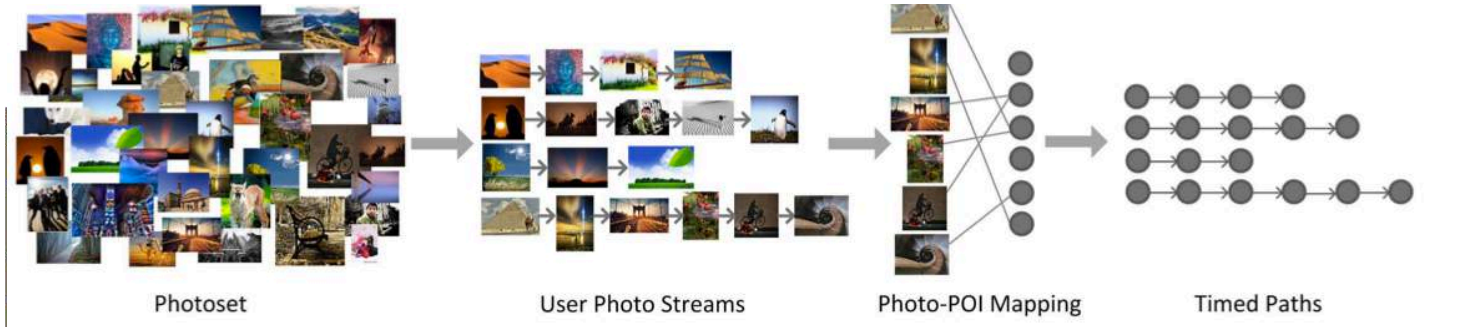
- Exploit the context (and deep learning!)

91% accuracy to predict the next app you will use
[Baeza-Yates et al, WSDM 2015]

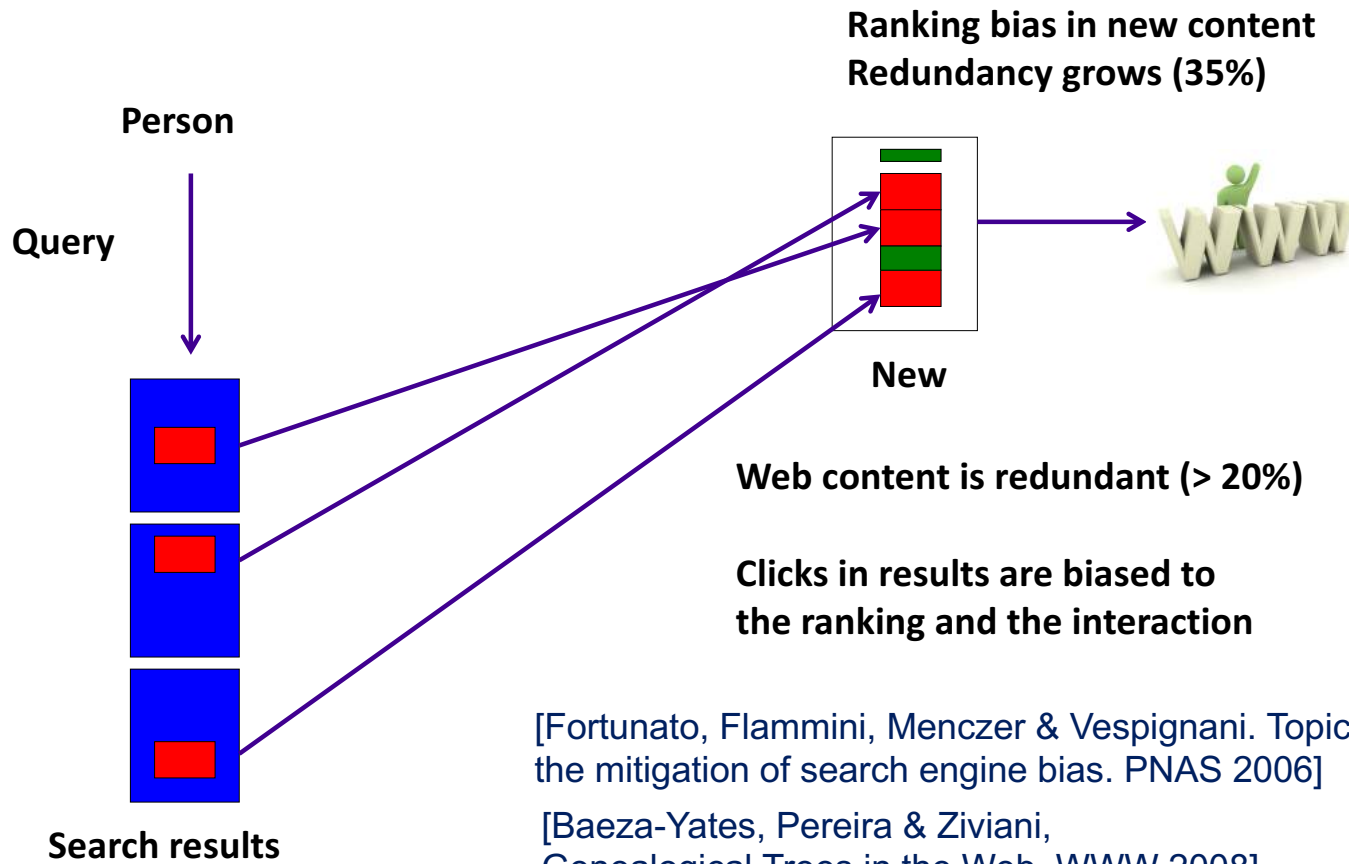
- Personalization vs. **Contextualization**

Recall that user interaction is another long tail





Second Order Bias in Web Content



[Fortunato, Flammini, Menczer & Vespignani. Topical interests and the mitigation of search engine bias. PNAS 2006]

[Baeza-Yates, Pereira & Ziviani, Genealogical Trees in the Web, WWW 2008]

The Web Works Thanks to Bias!

- Web traffic

- Local caching
- Proxy/network caching

Activity bias

- Search engines

- Answer caching
- Essential web pages

(Self) selection bias

- › 25% queries can be answered with less than 1% of the URLs!

[Baeza-Yates, Boldi, Chierichetti, WWW 2015]

- E-Commerce

- Large fraction of revenue comes from few popular items

Take-Home Message

- Web data is a mirror of us, the good, the bad and the ugly
- The Web amplifies everything, but always leaves traces

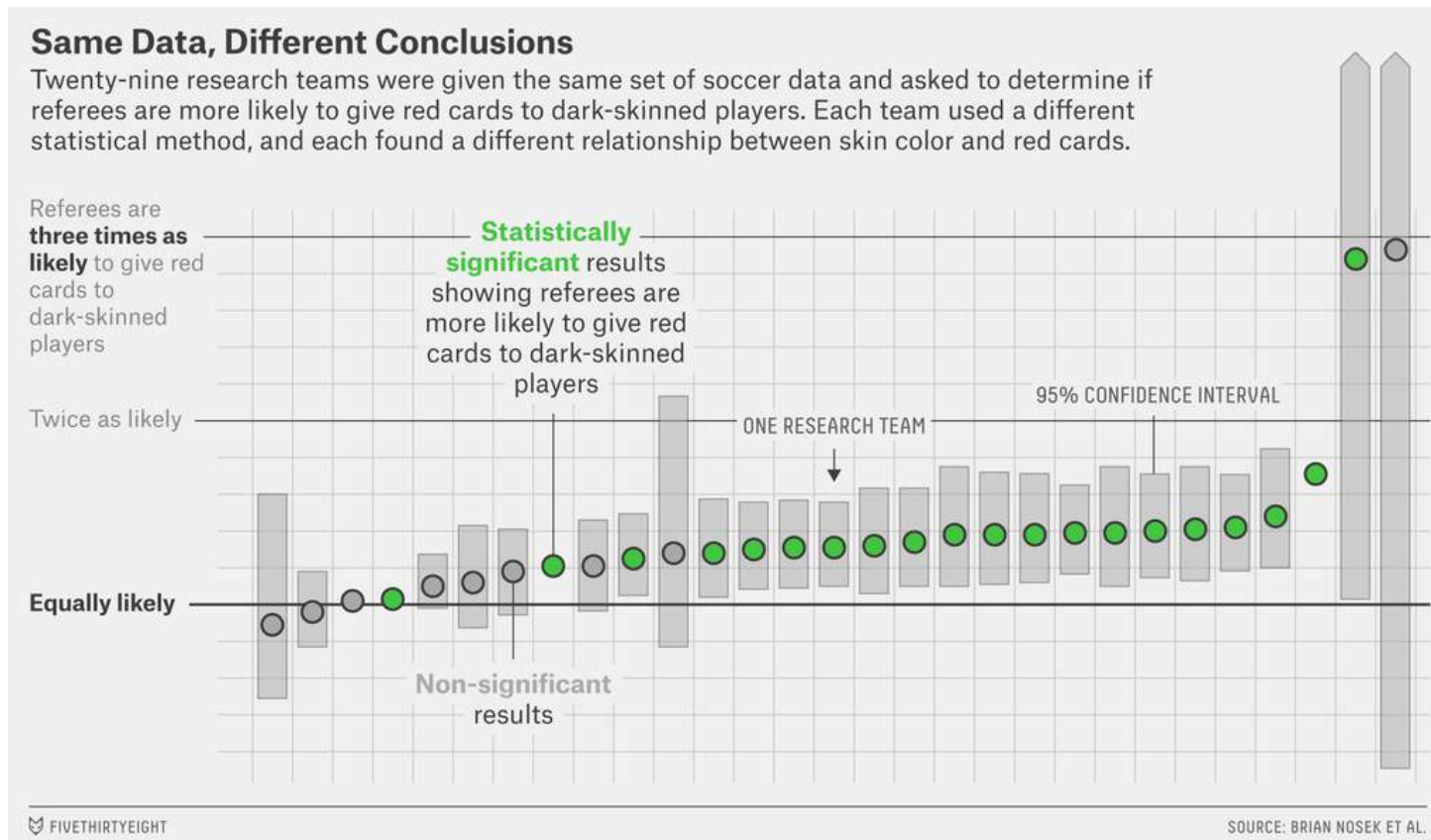
- We need to be aware of our **own bias!**
- We have to be aware of the biases and contrarrest them to stop the **vicious bias cycle**
- We have to be aware of **our privacy**
- **Plenty** of open research problems!

Big Data of People is huge.....
..... but it is tiny compared to the future
Big Data of the Internet of Things (IoT)



No activity bias!

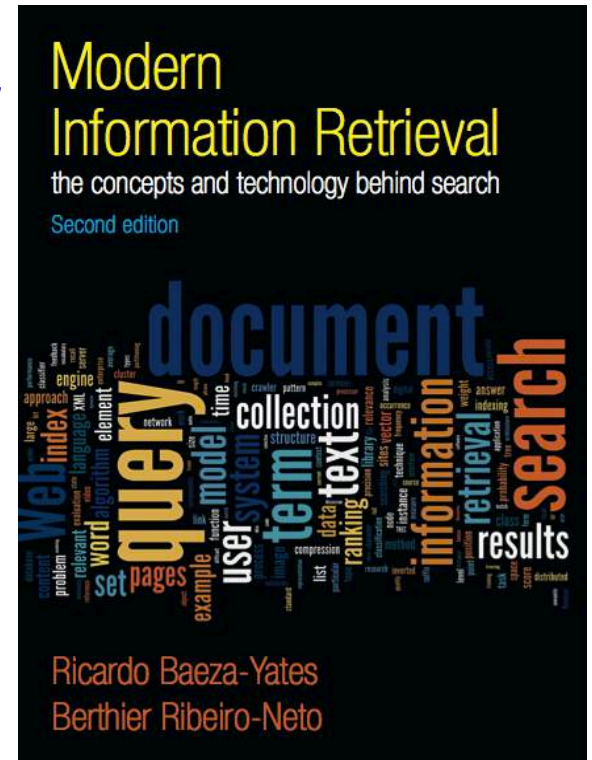
It's Hard to Get the Truth from Data (Professional Bias)



- 61 analysts, 29 teams: 20 yes and 9 no (Univ. of Virginia, COS)
- We need to focus on small data, not big data

Questions?

ASIST 2012
Book of the
Year Award
(Biased Ad)



Contact: rbaeza@acm.org

www.baeza.cl

[@polarbearby](https://twitter.com/polarbearby)



Biased Questions?