

# Bias on the Web

## Ricardo Baeza-Yates

Appeared in  
CACM of June 2018



**Northeastern University**  
College of Computer and Information Science



Symposium on Fairness and Transparency, Leiden, March 2019

## What is Bias?

- Statistical: significant systematic deviation from a prior (unknown) distribution;
- Cultural: interpretations and judgments phenomena acquired through our life;
- Cognitive: systematic pattern of deviation from norm or rationality in judgment;

### 20 COGNITIVE BIASES THAT SCREW UP YOUR DECISIONS

<p><b>1. Anchoring bias.</b> People are over-reliant on the first piece of information they receive. In a study regarding software licenses, the first offer made was 120 dollars. The license was eventually sold for 120 dollars.</p> 	<p><b>2. Availability heuristic.</b> People overestimate the importance of information that is readily available. A person might judge that smoking is not as harmful because they know someone who smoked and didn't get cancer.</p> 	<p><b>3. Bandwagon effect.</b> The probability of your own behavior being linked to a line or trend. People are more likely to buy a product if they see others buying it.</p> 	<p><b>4. Blind spot bias.</b> Failing to recognize our own biases. People are often unaware of their own biases and overestimate their own ability to avoid them.</p> 
<p><b>5. Choice-supportive bias.</b> After you make a decision, you tend to forget the negative aspects of that choice. People who are in a relationship tend to forget the negative aspects of that relationship.</p> 	<p><b>6. Clustering illusion.</b> The tendency to see patterns in random events. People often see patterns in random events, such as a streak of heads in a coin toss.</p> 	<p><b>7. Confirmation bias.</b> The tendency to search for, interpret, and recall information in a way that confirms one's preexisting beliefs.</p> 	<p><b>8. Conservatism bias.</b> When people face uncertainty, they tend to stick to their previous beliefs. People who are in a relationship tend to stick to their previous beliefs.</p> 
<p><b>9. Information bias.</b> The tendency to seek information when it does not always help. More information is not always better.</p> 	<p><b>10. Ostrich effect.</b> The tendency to ignore negative information. People often ignore negative information, such as a bad grade on a test.</p> 	<p><b>11. Outcome bias.</b> Judging a decision based on the outcome, rather than on the quality of the decision itself.</p> 	<p><b>12. Overconfidence.</b> Overestimating one's own abilities. People often overestimate their own abilities, such as their driving skills.</p> 
<p><b>13. Plausibility effect.</b> When simply believing that something is true makes it seem more likely. People often believe things that are plausible, even if they are not true.</p> 	<p><b>14. Pro-innovator bias.</b> When a person is in a position of authority, they tend to favor new ideas. People often favor new ideas, such as a new product.</p> 	<p><b>15. Reversal.</b> The tendency to weigh the benefits of a decision more heavily than the costs. People often weigh the benefits of a decision more heavily than the costs.</p> 	<p><b>16. Saliency.</b> The tendency to focus on the most salient information. People often focus on the most salient information, such as a red car.</p> 
<p><b>17. Selective perception.</b> The tendency to see what we want to see. People often see what they want to see, such as a bad grade on a test.</p> 	<p><b>18. Shortcutting.</b> The tendency to use shortcuts to make decisions. People often use shortcuts, such as a red light.</p> 	<p><b>19. Survivorship bias.</b> The tendency to focus on the survivors and ignore the failures. People often focus on the survivors, such as a successful business.</p> 	<p><b>20. Zero-risk bias.</b> The tendency to prefer a certain risk over a larger risk. People often prefer a certain risk, such as a small fire.</p> 



www.ntent.com | @withntent | 877.861.2230

© 2018 ntent. All rights reserved. This document is the property of ntent. It is not to be distributed, copied, or reproduced in any form without the prior written permission of ntent. The information contained herein is confidential and intended only for the individual named. If you are not the named individual you should not disseminate the information. If you are not the named individual you should not disseminate the information. If you are not the named individual you should not disseminate the information. If you are not the named individual you should not disseminate the information.

## Motivation 1: Inequality of Content

- First, inequality of Internet access
  - From 98% in Iceland to less than 1% in South Sudan
- Content inequality across languages
  - Most websites are in English (estimated in 52%) while only 13% speaks English
  - On the other hand, only 4% of the websites are in Mandarin (China) while this country has 22% of the users
  - There about 6,900 languages but only 288 of them have an active Wikipedia
  - There are 4 times more Wikipedia entries in English than Spanish although there are more native Spanish speakers than native English speakers
- Content optimized most of the time for local purposes (e.g., business and government) and not for the actual needs of people
- Also there is bias on content quality (later)



www.ntent.com | @withntent | 877.861.2230



Northeastern University  
College of Computer and Information Science

11

## Motivation 2: Impact in Search and Recommender Systems

- Most web systems are optimized by using implicit user feedback
- However, user data is partly biased to the choices that these systems make
  - Clicks can only be done on things that are shown to us
- As those systems are usually based in ML, they learn to reinforce their own biases, yielding self-fulfilled prophecies and/or sub-optimal solutions
  - For example, personalization and the filter bubble
- Moreover, sometimes these systems compete among themselves, learning also biases of other systems rather than real user behavior
- Even more, an improvement in one system might be just a degradation in another system that uses a different (inversely correlated) optimization function
  - For example, user experience vs. monetization



www.ntent.com | @withntent | 877.861.2230



Northeastern University  
College of Computer and Information Science

12

## Motivation 3: 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



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

Buzzfeed News

16



www.ntent.com | @withntent | 877.861.2230



Northeastern University  
College of Computer and Information Science

14

## So (Observational) Human Data has Bias

### Cultural Biases

- Gender
- Racial
- Sexual
- Age
- Religious
- Social

- Linguistic
- Geographic
- Political
- Educational
- Economic
- Technological

### Statistical Biases

- Gathering process
- Sampling process
- Validity (e.g. temporal)
- Completeness
- Noise, spam

### Cognitive Biases

Self-selection

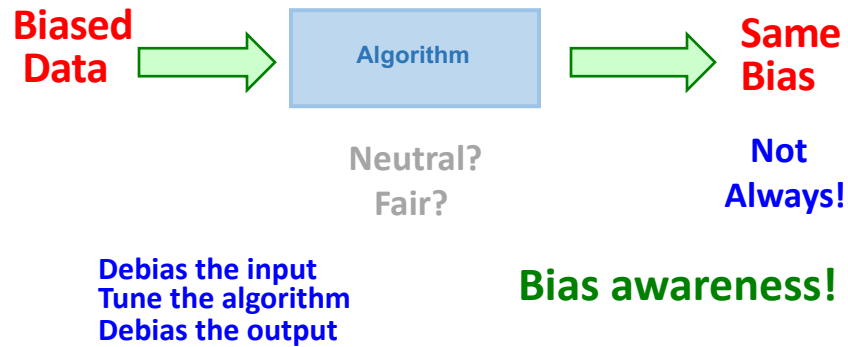
Attempt of an unbiased (personal) view on bias in the Web

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



## ACM US Statement on Algorithm Transparency and Accountability (Jan 2017)

1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data Provenance
6. Auditability
7. Validation and Testing

**They do not need to be perfect,  
they just need to be better than us**



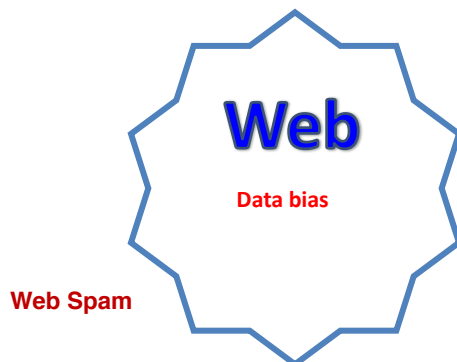
## Big Data and Bias

- The **quality of any algorithm** is bounded by the **quality of the data that uses**
- Data bias awareness
  - [Gordon & Desjardins; Provost & Buchanan, MLJ 1995]
- Bias in computer systems: [Friedman & Nissenbaum 1996]
- 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

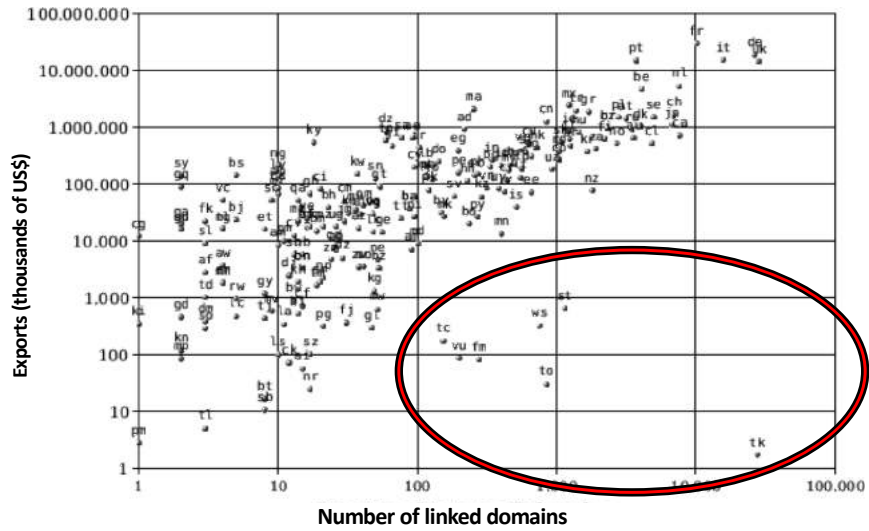


21

## Bias in the Web



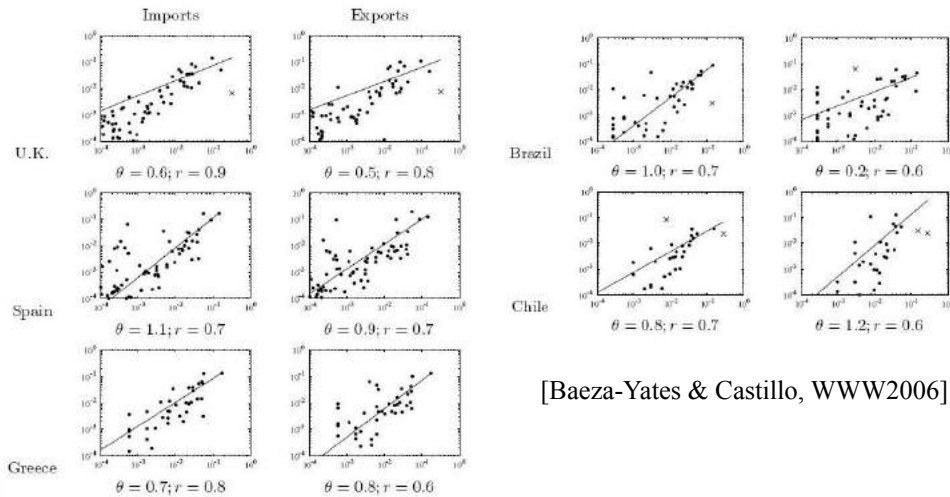
## Economic Bias in Links



[Baeza-Yates, Castillo & López. Characteristics of the Web of Spain. Cybermetrics, 2005]



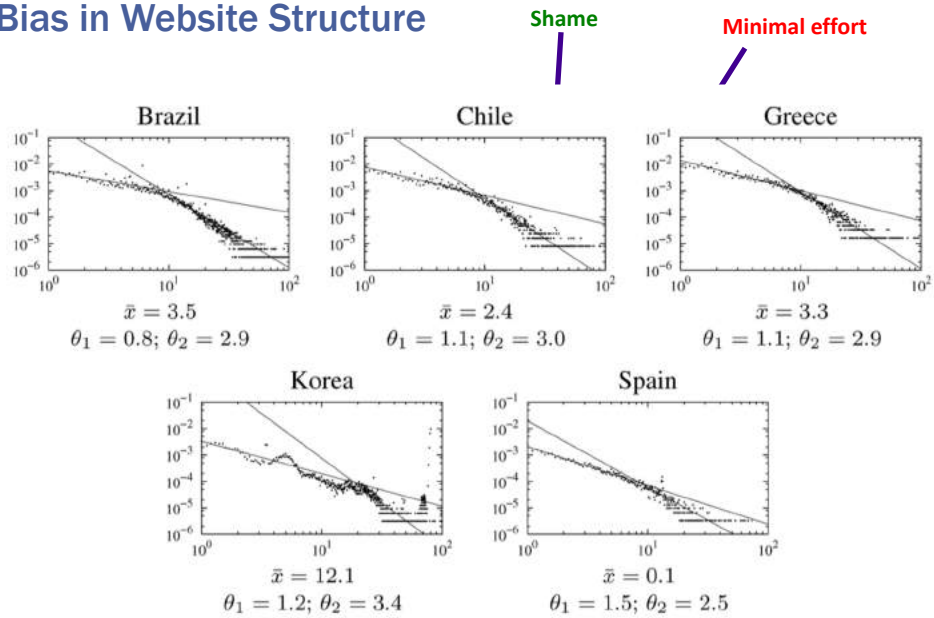
## Economic Bias in Links



[Baeza-Yates & Castillo, WWW2006]



## Cultural Bias in Website Structure



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

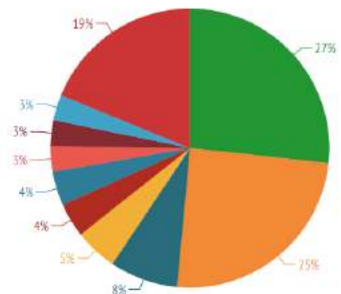
## Linguistic Bias in Content

Top 25 World Languages

- Chinese, Mandarin
- Spanish
- English
- Hindi

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

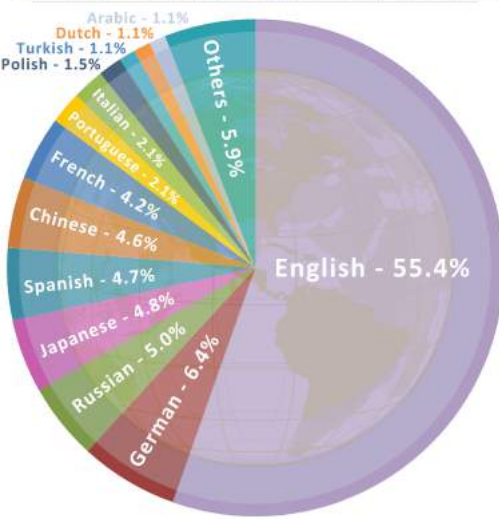
### Languages on the Web



English Chinese Mandarin Spanish Japanese Portuguese German Arabic French Russian Other

### The Languages of Web Content

The percentage of the top 1 million websites available in various languages



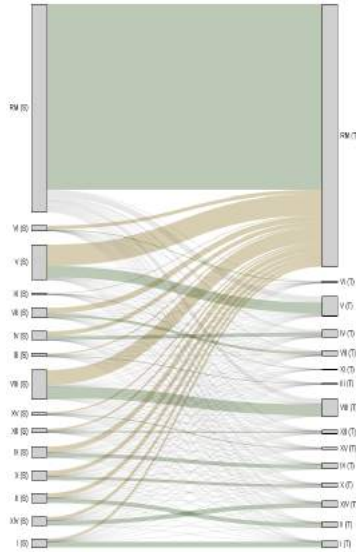
language connect

Insight Digital Audio Branding Text

Web: languageconnect.net / E-mail: info@languageconnect.net / Follow us @lconnect



# 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	hairstresser-barber

### Gender appropriate *she-he* analogies.

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

Most journalists are men?

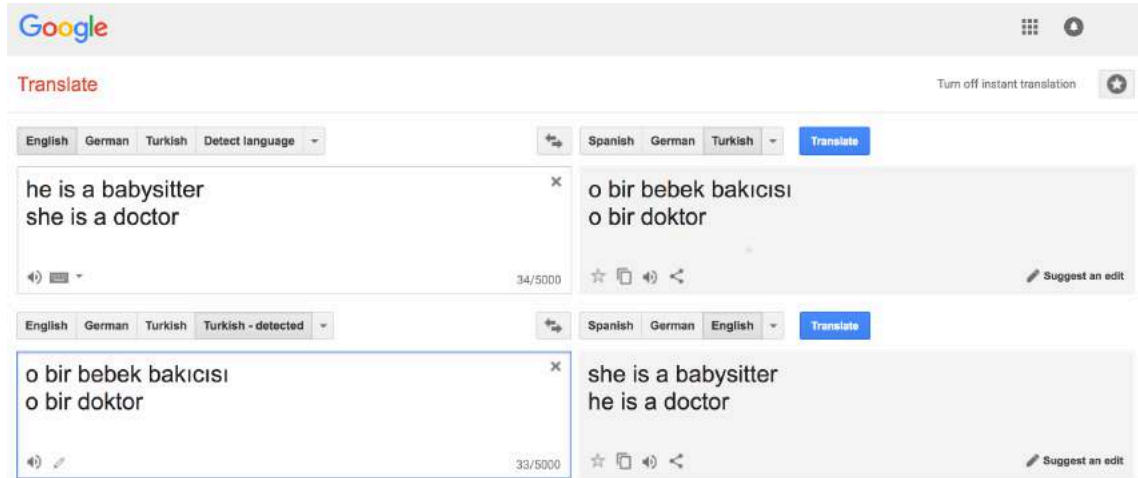
Yes, about 60 to 70% at work  
although at college is the inverse

[Bolukbasi at al, NIPS 2016]

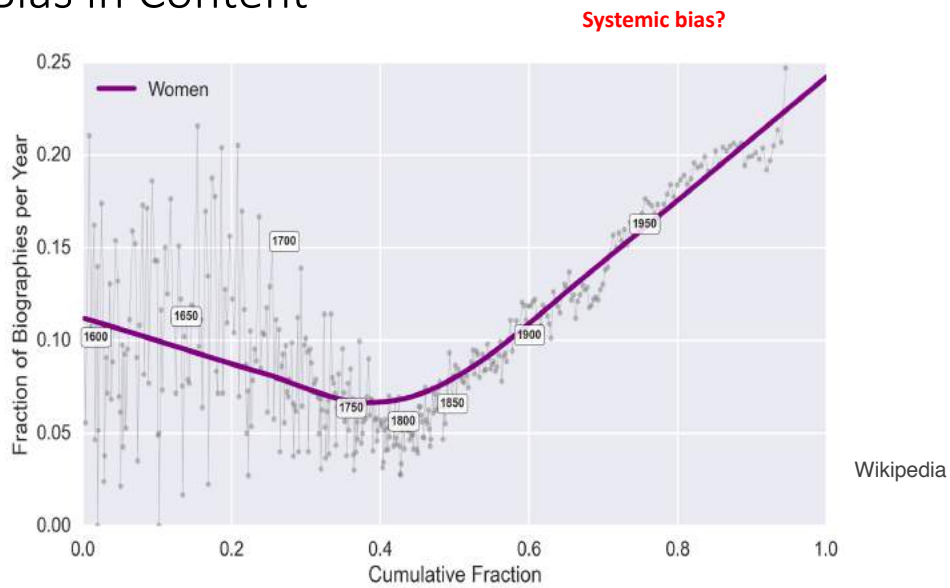




# Gender Bias in Translation

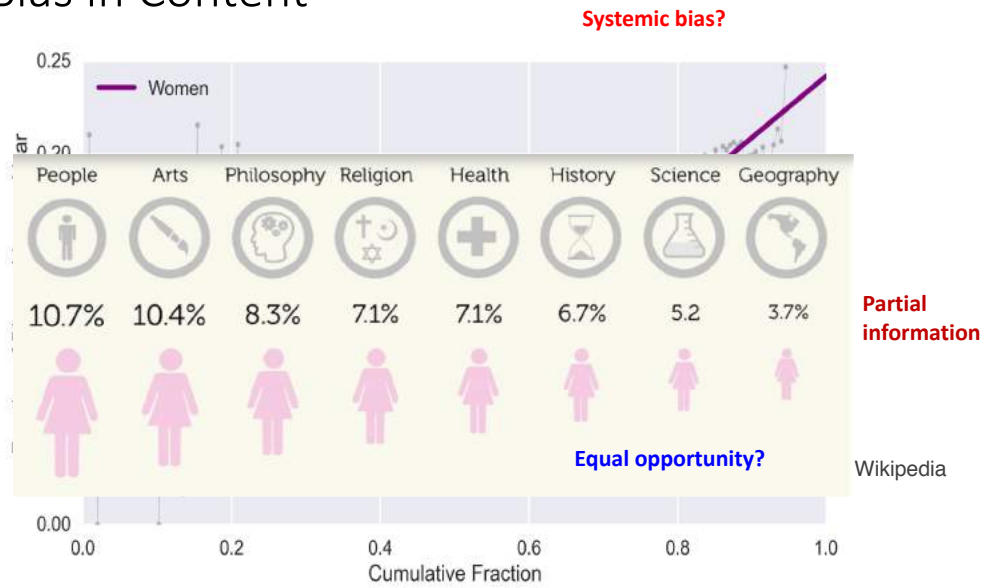


# Gender Bias in Content



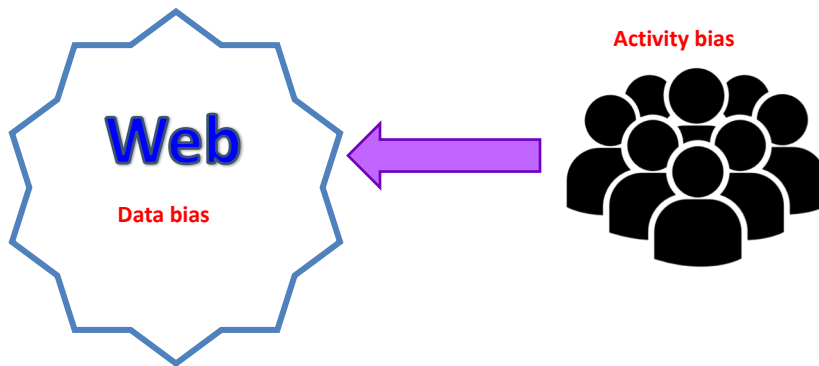
[E. Graells-Garrido et al., "First Women, Second Sex: Gender Bias in Wikipedia", ACM Hypertext'15]

# Gender Bias in Content



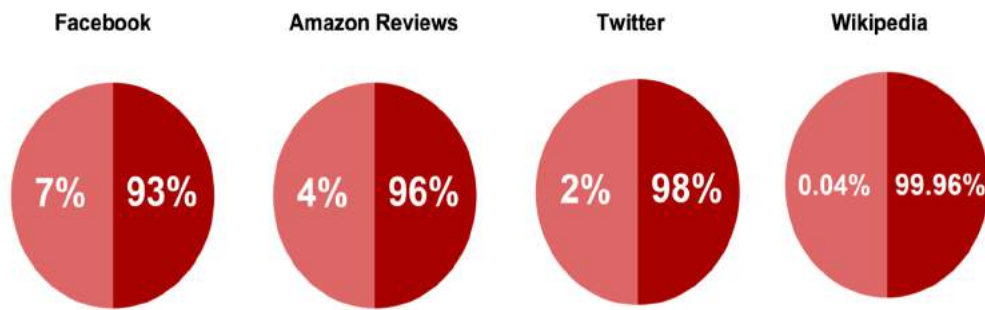
[E. Graells-Garrido et al., "First Women, Second Sex: Gender Bias in Wikipedia", ACM Hypertext'15]

# Bias in the Web



## Activity Bias

Most users are passive (*i.e.*, more than 90%) – wisdom of crowds is a partial illusion  
 Which percentage of **active** users produce 50% of the content?



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



**theguardian**

port football opinion culture business lifestyle fashion environment tech travel ≡ all sections

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

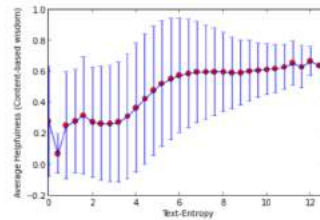
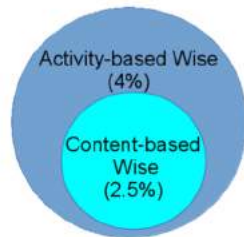
# Amazon Continues Their Crusade Against Fake Reviews

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



## Quality of Content?

- Adding content  $\Rightarrow$  Adding Wisdom ?
- We use Amazon's Reviews helpfulness
- Content-based-wise users



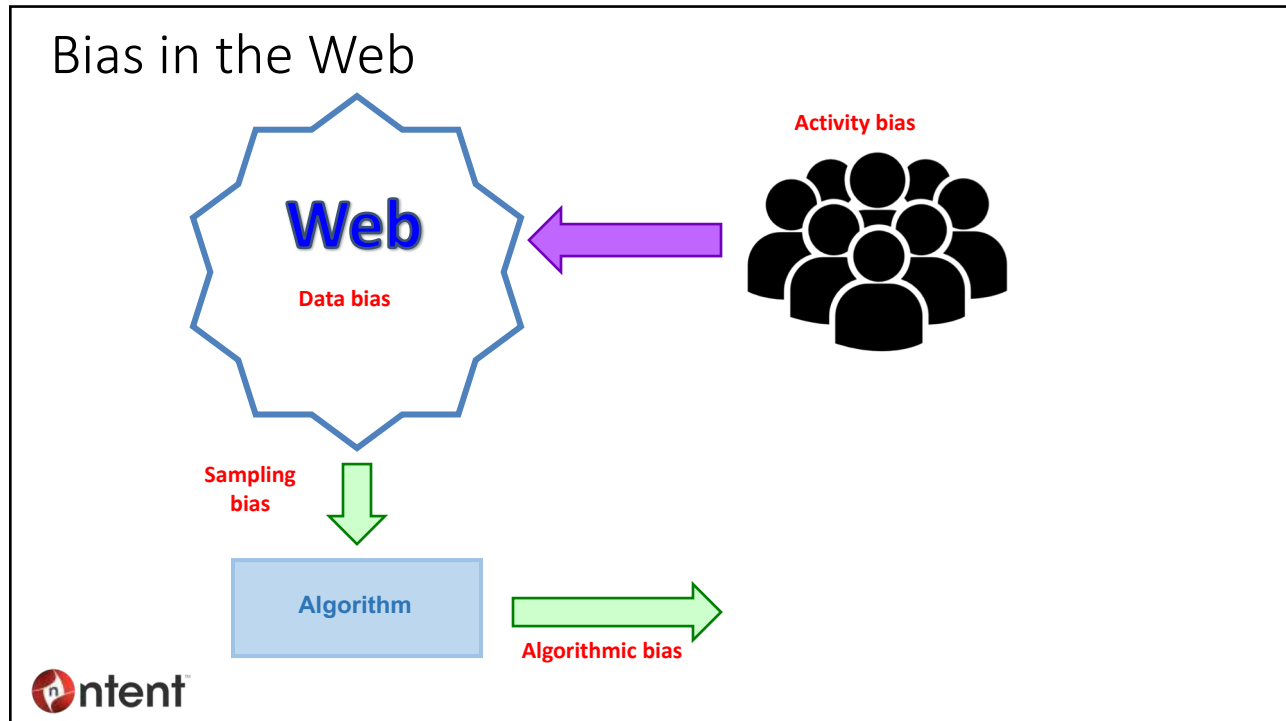
[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]

## 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.



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



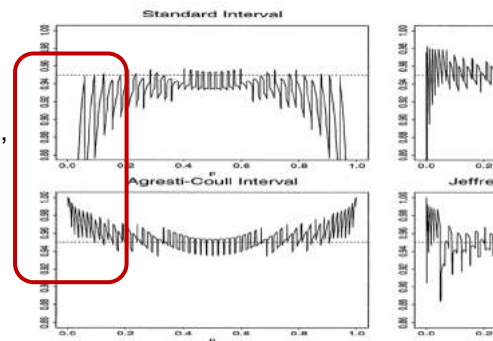
## Sample Size?

- If we want to estimate the frequency of queries that appear with probability at least  $p$  with a certain relative error  $\epsilon$  we can use the standard binomial error formula  $\sqrt{(1-p)/np}$  which works well for  $p$  near  $1/2$  **but not for  $p$  near 0**
- Better is the Agresti-Coull technique (also called *take 2*) which gives:

$$n \geq Z_{1-\alpha/2}^2 \left( \frac{p'(1-p')}{\epsilon^2} - 1 \right)$$

where  $Z$  is the inverse of the standard normal distribution,  $1 - \alpha$  is the confidence interval and  $p' = p + Z^2/2$

- If  $p = 0.1$ ,  $1 - \alpha$  is 80% and  $\epsilon$  is 10%, we get  $n = 2342$ . The standard formula gives  $n = 900$ !

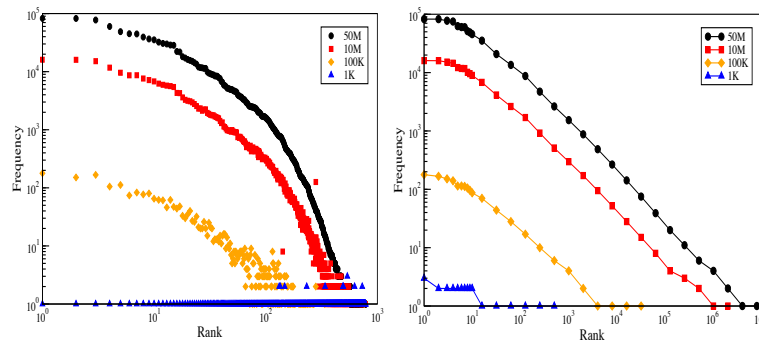


## Sampling Techniques

- Standard technique:

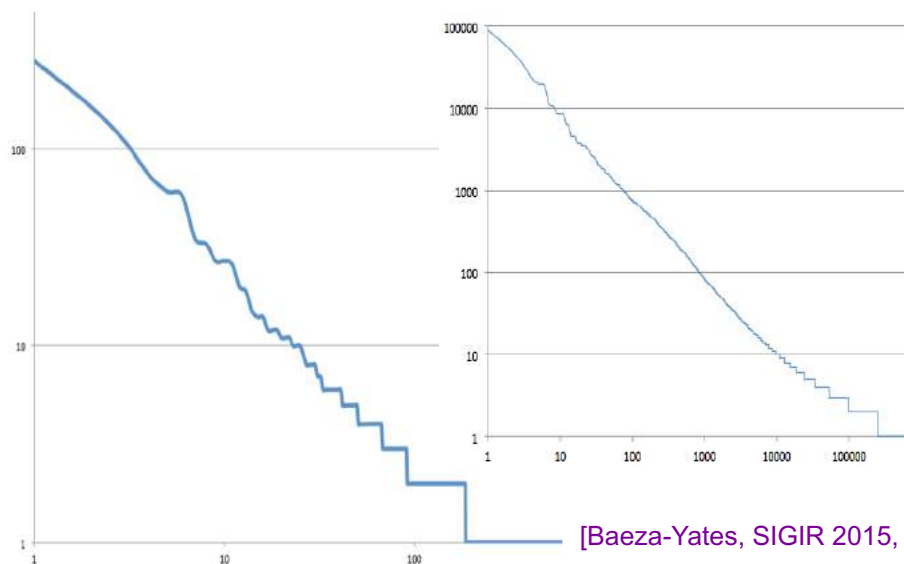
$$p_q \approx \hat{p}_q(\mathcal{S}) = \frac{f_q(\mathcal{S})}{\sum_{q' \in \mathcal{S}} f_{q'}(\mathcal{S})}$$

- A good sample should cover well all the query distribution but this does not work with very skewed distributions.



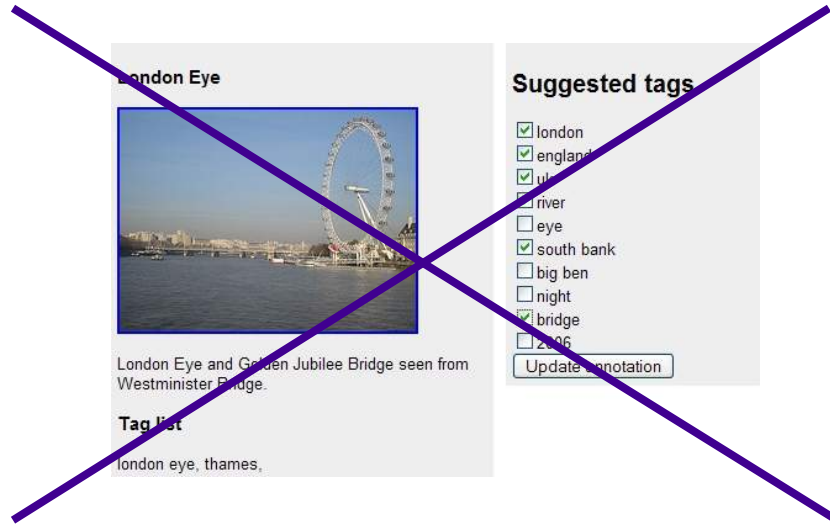
[Zaragoza et al, CIKM 2010]

## Stratified Sampling Example

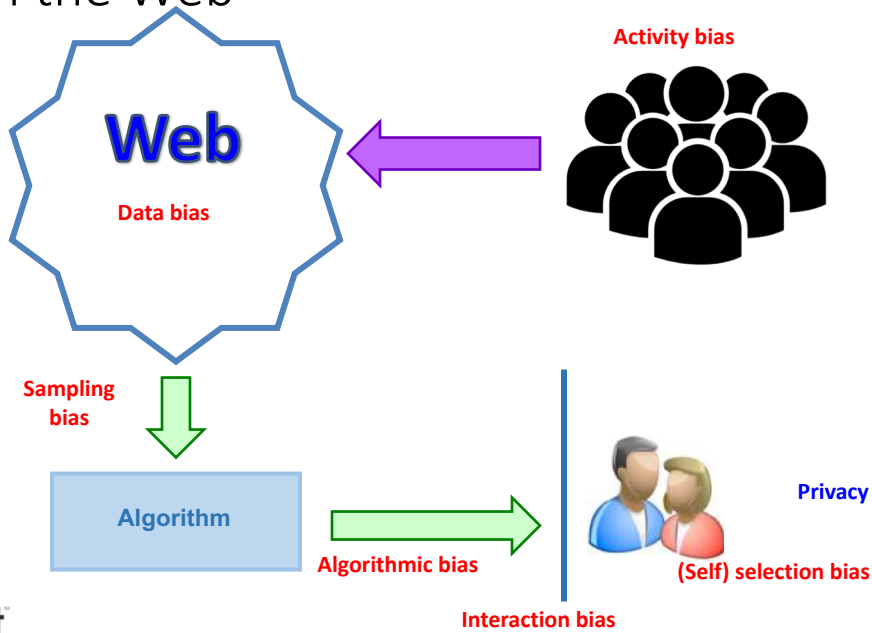


[Baeza-Yates, SIGIR 2015, Industry track]

# Extreme Algorithmic Bias



# Bias in the Web



# Bias in the Interaction

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

Shop by Category

- Tennis Equipment
- Tennis Games
- Kids' Sports
- Clothing, Shoes & Jewelry
- Tennis - Books

**Position bias**  
**Ranking bias**

**Presentation bias**

**Sponsored**

Tennis Elbow Brace with Gel Comp...  
\$24.50 Prime  
★★★★★ 7

DIMANKA Professional Table Tenni...  
\$34.99  
★★★★★ 9

Gamma Quick Kids 78 Ball (12 Pac...  
\$19.99 Prime  
★★★★★ 44

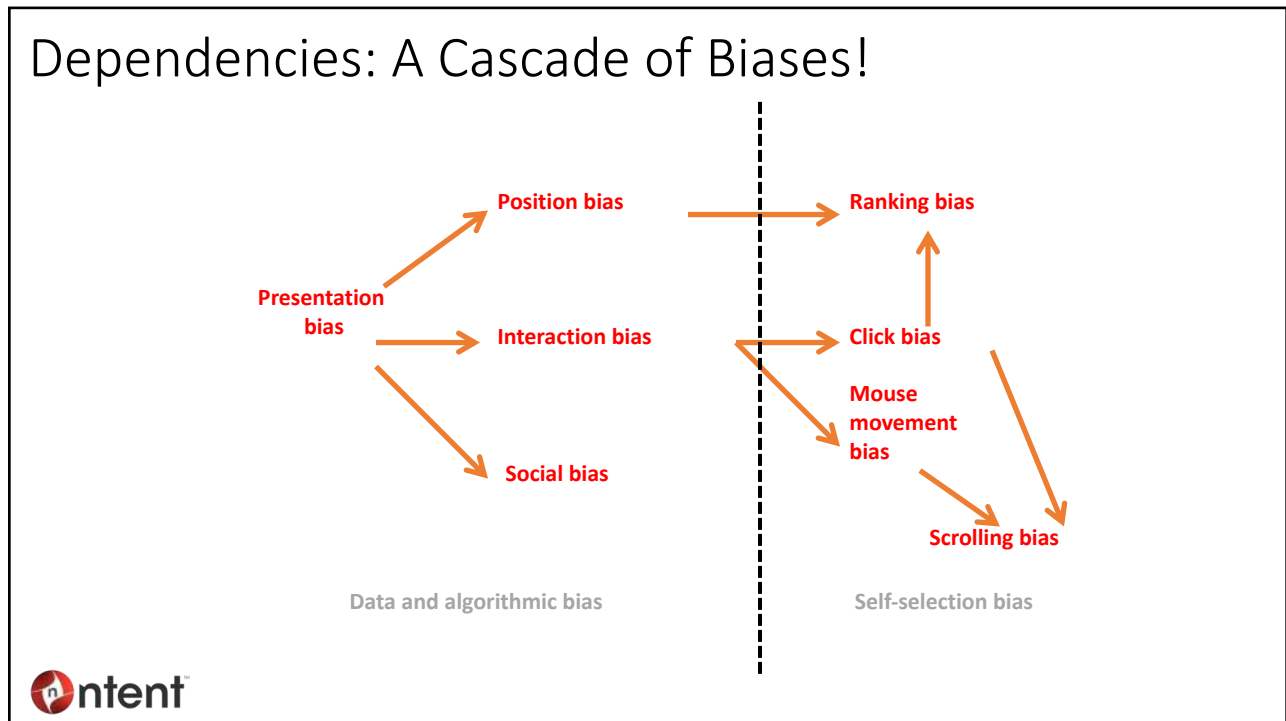
Wilson Sporting Goods Championship Extra Duty Tennis Balls (1-Can)  
Jun 14, 2012  
by Wilson  
\$2.79 \$6.99 Add-on Item  
Add to a qualifying order to get it by **Tomorrow, May 6**.  
More Buying Choices  
\$0.99 new (18 offers)  
\$7.99 used (2 offers)  
See newer version  
★★★★★ + 186  
Sports & 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

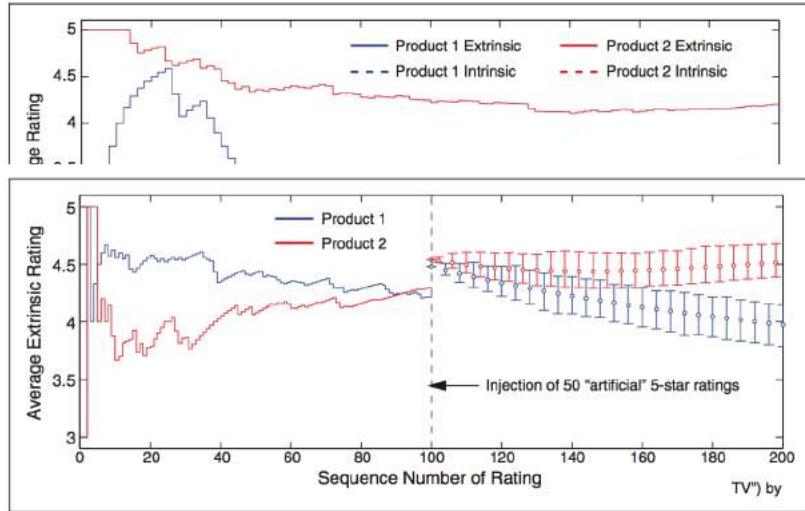
**Interaction bias**

**Amazon.com**





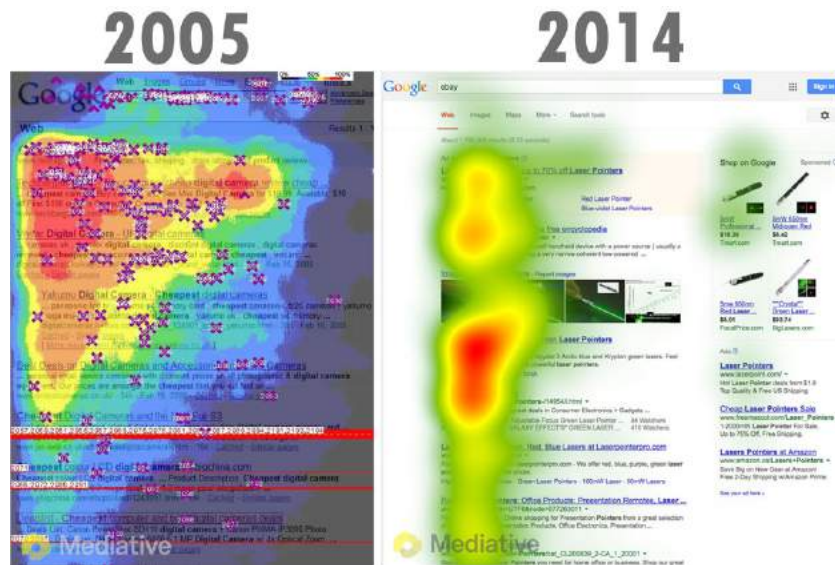
# Social Bias



[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

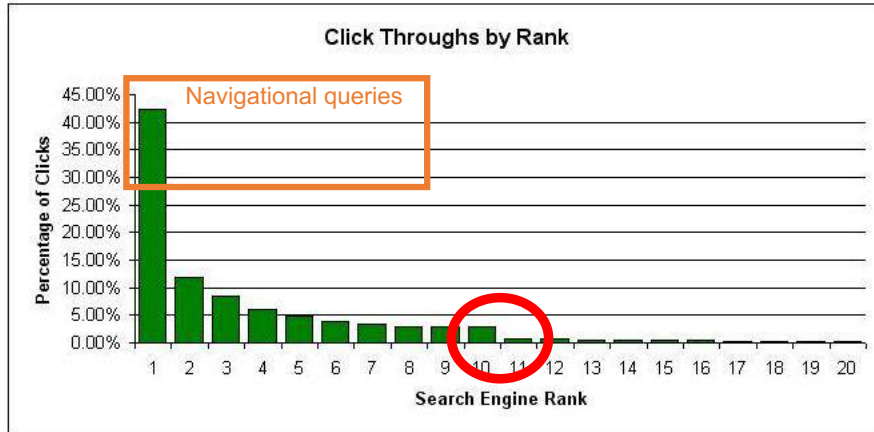


[Mediative Study, 2014]



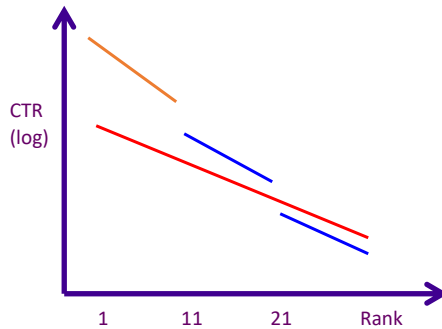
# Click Bias in Web Search

- Ranking & next page bias



# Debiasing Search Clicks

Clicks as implicit positive user feedback

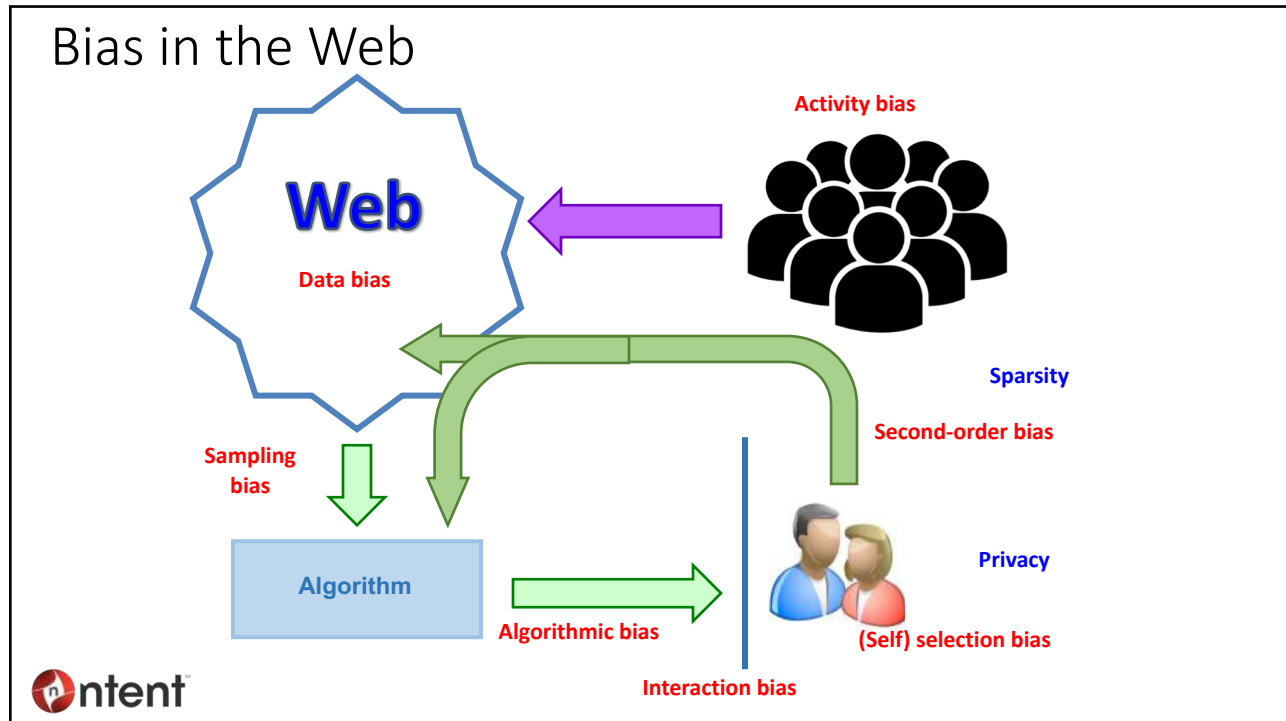


Learning to Rank with bias  
[Joachims et al, WSDM 2017, **best paper**]

Fair rankings  
[Zehlike et al, CIKM 2017]

- [Dupret & Piwowarski, SIGIR 2008]
- [Chapelle & Zhang, WWW 2009]
- [Dupret & Liao, WSDM 2010]





## Second Order Bias due to Personalization

- The effect of self-selection bias
- Avoid the rich get richer and poor get poorer syndrome
- Avoid the echo chamber by empowering the tail

### Partial solutions:

- Diversity
- Novelty
- Serendipity
- My dark side

**Cold start problem solution: Explore & Exploit**

**How much exploration is needed to counteract presentation bias?**



[Eli Pariser, The Filter "Bubble", 2011]

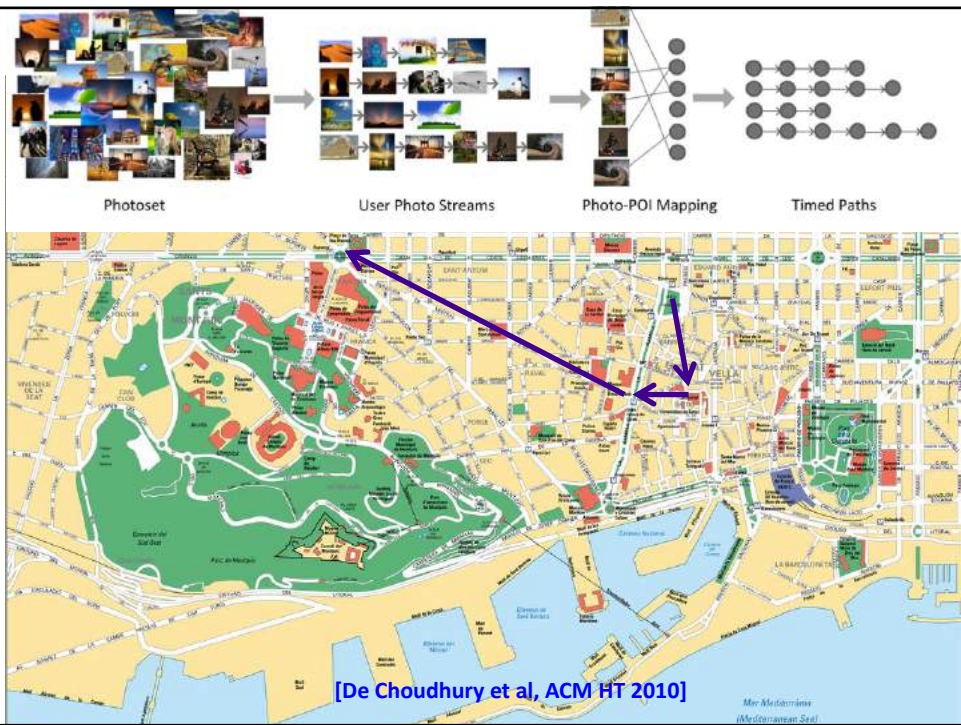
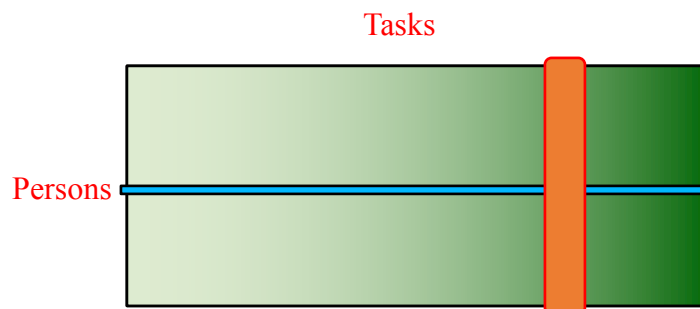
## Aggregating in the Tail

- Exploit the context (and deep learning!)

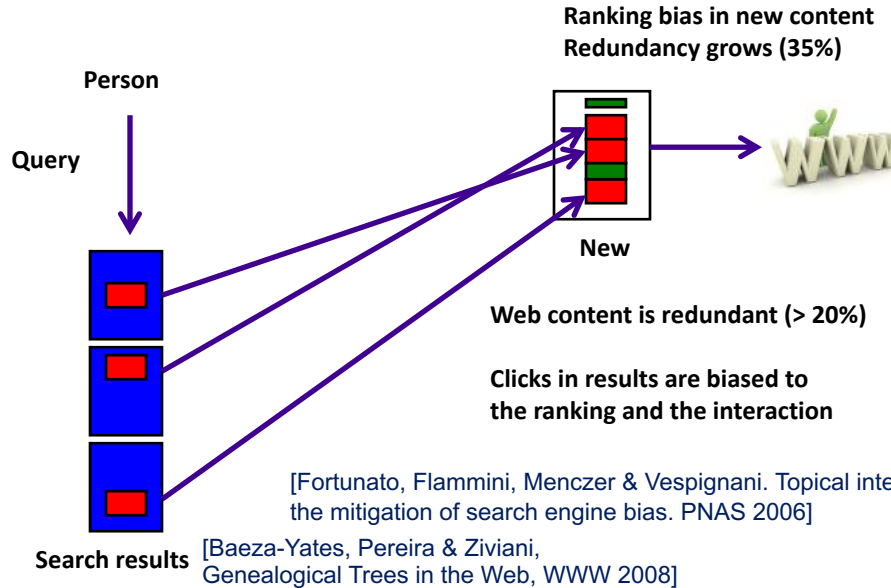
91% accuracy to predict the next app you will use  
[Baeza-Yates et al, WSDM 2015]

- Personalization vs. **Contextualization**

Recall that user interaction has a long tail and we are all in it!  
[Goel et al, WSDM 2010]



## Second Order Bias in Web Content



## The Web Works Thanks to Bias!

- Web traffic
  - Local caching
  - Proxy/network caching
- Search engines
  - Answer caching
  - Essential web pages
    - 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

Activity bias

(Self) selection bias



## 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! (in small data even more!)

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!**

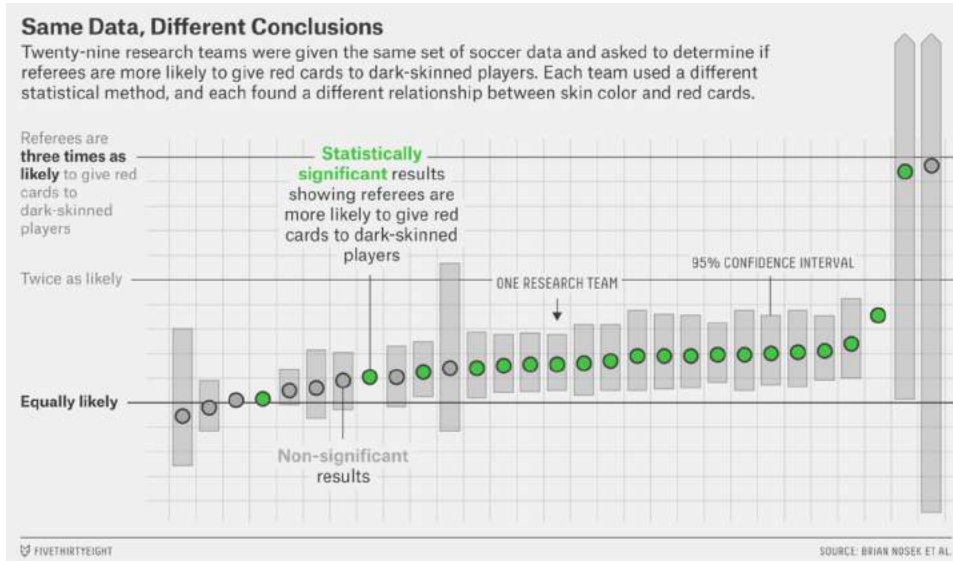


## Recap

Bias \ Type	Statistical	Cultural	Cognitive
Algorithmic	◆	?	?
Presentation	◆		
Position	◆	◆	◆
Data	◆	◆	
Sampling	◆	◆	◆
Activity		◆	
Self-selection		◆	◆
Interaction		◆	◆
Social		◆	◆
Second order	◆	◆	◆



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



→ 61 analysts, 29 teams: 20 yes and 9 no (Univ. of Virginia, COS)

## Questions?

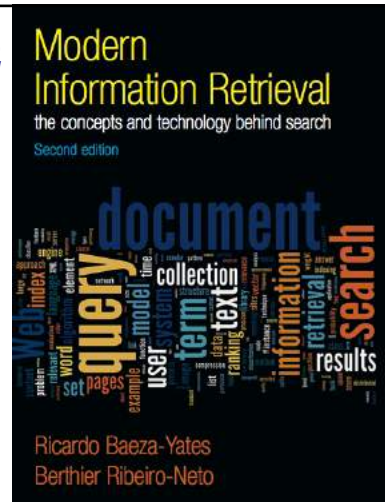
ASIST 2012  
 Book of the  
 Year Award  
 (Biased Ad)

### New Conferences that started in 2018:

AAAI/ACM Conference on AI, Ethics, and Society  
<http://www.aies-conference.com>

Conference on Fairness, Accountability, and Transparency  
<http://fatconference.org>

Resources: <http://fairness-measures.org>



Contact: [rbaeza@acm.org](mailto:rbaeza@acm.org)  
[www.baeza.cl](http://www.baeza.cl)  
 @polarbearby



## Biased Questions?

Northeastern University  
 College of Computer and Information Science