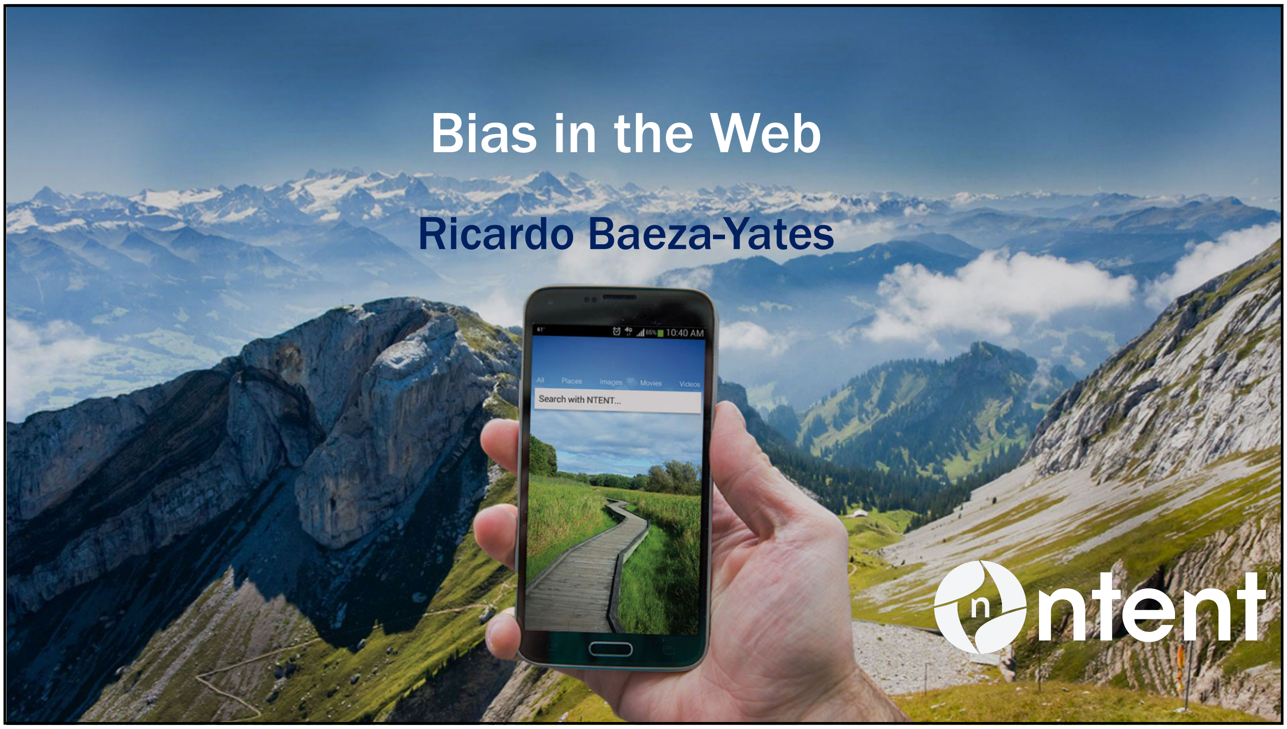


Bias in the Web

Ricardo Baeza-Yates

Fake Content & Bias

- British Prime Minister Benjamin Disraeli:
  - "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

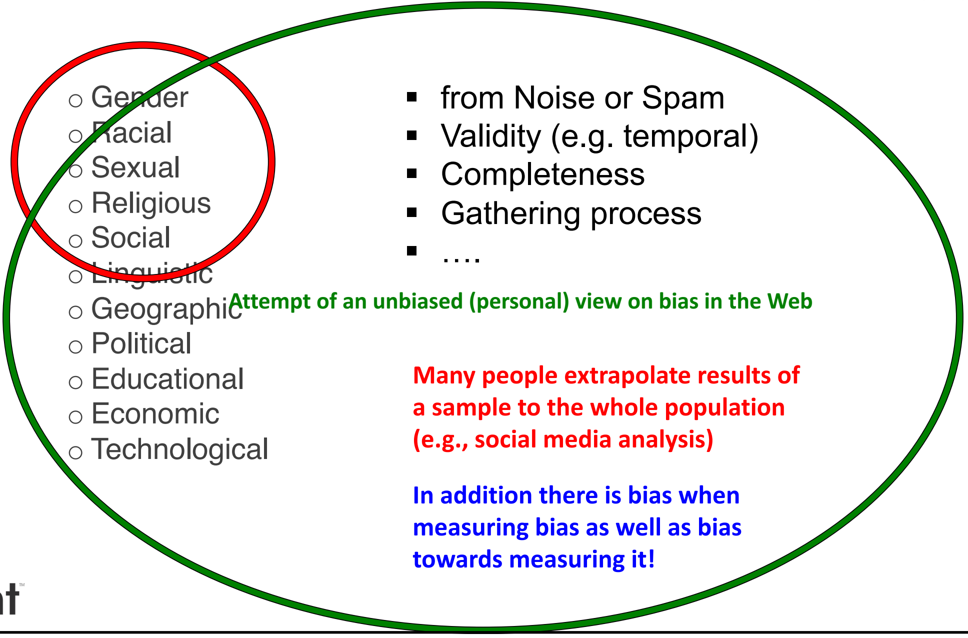
posted on Dec. 30, 2016, at 2:12 p.m.

Craig Silverman
BuzzFeed News Media Editor

Bias: significant deviation from a prior (unknown) distribution

(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
- ….

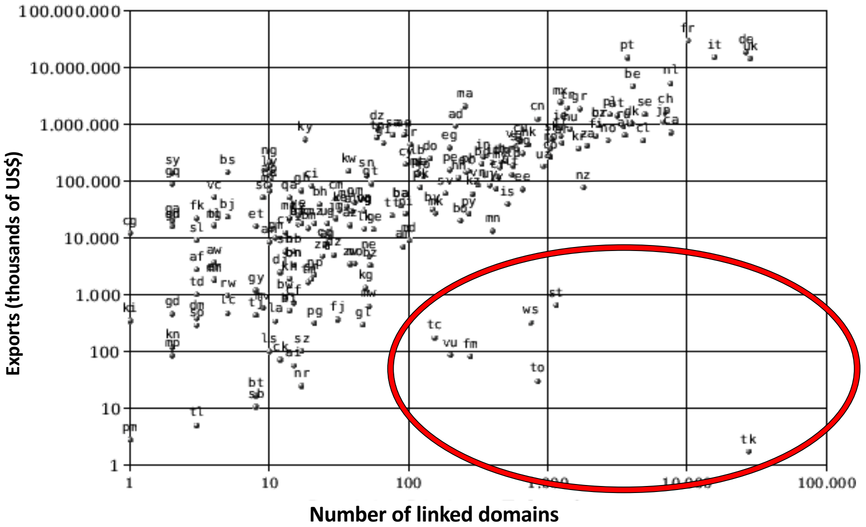
Attempt of an unbiased (personal) view on bias in the Web

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!

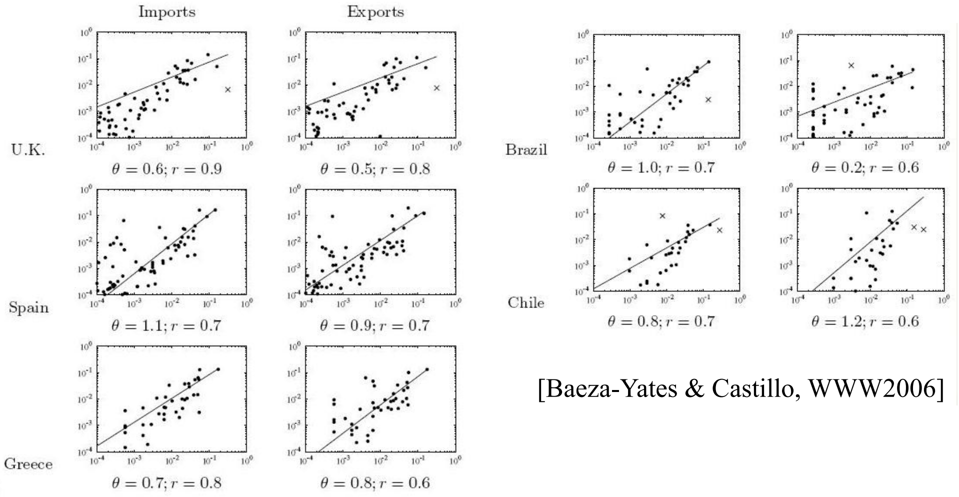
Bias in the Web

Economic Bias in Links



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

Economic Bias in Links

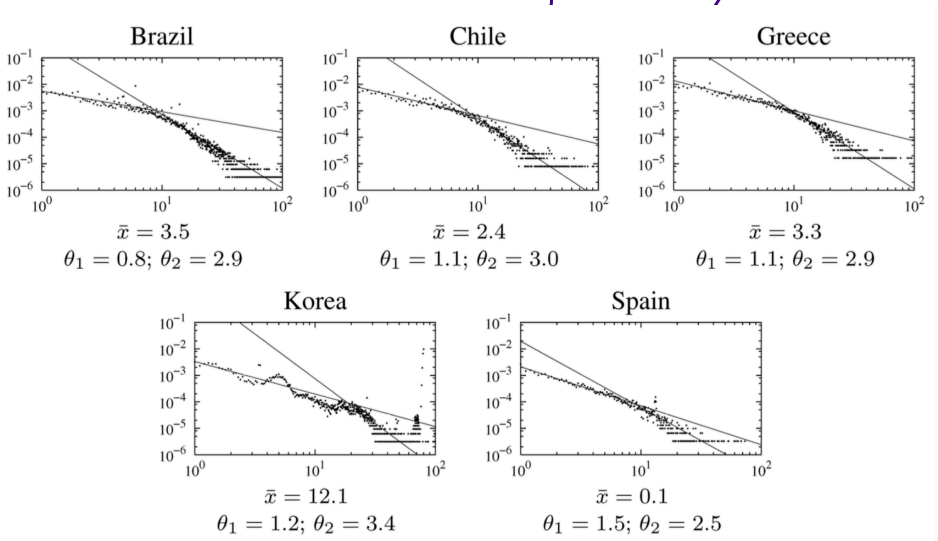


[Baeza-Yates & Castillo, WWW2006]

Website Structure

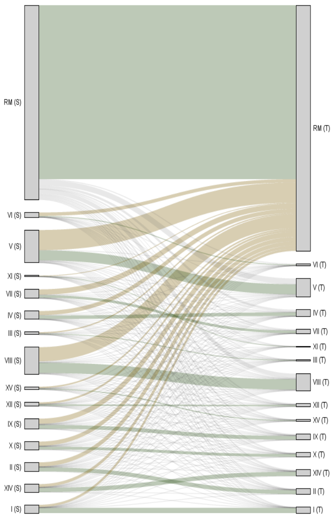
Shame

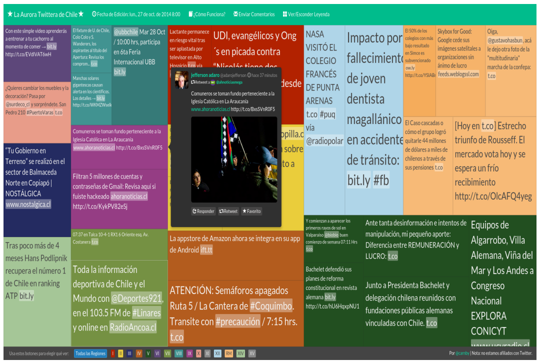
Minimal effort



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

Geographical Bias in Content





★ La Aurora Twittera de Chile ★ Fecha de Edición: lun., 27 de oct. de 2014 8:00 ¿Cómo Funciona? Enviar Comentarios Ver/Esconder Leyenda

Con este simple video aprenderás a entrenar a tu cachorro al momento de comer → bit.ly http://t.co/EVdIVAT6wH

El fixture de U. de Chile, Colo Colo y S. Wanderers, los aspirantes al título del Apertura: Revisa los comprom... t.co

@ubbchile Mar 28 Oct / 10:00 hrs, participa en 6ta Feria Internacional UBB bit.ly

Manchas solares gigantescas causan alerta en los científicos. Los detalles → bit.ly http://t.co/WKMZWselk

Lactante permanece en riesgo vital tras ser aplastada por televisor en Alto Hospicio t.co vía

UDI, evangélicos y Ong ´s en picada contra "Nicolás tiene dos ... desde

jefferson adaro @adarojefferson hace 37 minutos
Retweet a @ahnoticiasmega
Comuneros se toman fundo perteneciente a la Iglesia Católica en La Araucanía www.ahoranoticias.cl http://t.co/BxsSVnR0F5
Responder Retweet Favorito

¿Quieres cambiar los muebles y la decoración? Pasa por @surdeco_cl y sorpréndete. San Pedro 210 #PuertoVaras t.co

Comuneros se toman fundo perteneciente a la Iglesia Católica en La Araucanía www.ahoranoticias.cl http://t.co/BxsSVnR0F5

"Tu Gobierno en Terreno" se realizó en el sector de Balmaceda Norte en Copiapó | NOSTÁLGICA www.nostalgica.cl

Filtran 5 millones de cuentas y contraseñas de Gmail: Revisa aquí si fuiste hackeado ahoranoticias.cl http://t.co/KykPV82eSj

...pilla.c ... sobre ... to a

Tras poco más de 4 meses Hans Podlipnik recupera el número 1 de Chile en ranking ATP bit.ly

07:37 en Talca 10-4-1 RX1 6 Oriente esq. Av. Costanera t.co

Toda la información deportiva de Chile y el Mundo con @Deportes921, en el 103.5 FM de #Linares y online en RadioAncoa.cl

La appstore de Amazon ahora se integra en su app de Android ift.tt

ATENCIÓN: Semáforos apagados Ruta 5 / La Cantera de #Coquimbo. Transite con #precaución / 7:15 hrs. t.co

NASA VISITÓ EL COLEGIO FRANCÉS DE PUNTA ARENAS t.co #puq vía @radiopolar

Impacto por fallecimiento de joven dentista magallánico en accidente de tránsito: bit.ly #fb

El 50% de los colegios con más bajo resultado en Simce es subvencionado ow.ly http://t.co/YSIABr

Skybox for Good: Google cede sus imágenes satelitales a organizaciones sin ánimo de lucro feeds.weblogssl.com

Oiga, @gustavohasbun, acá le dejo otra foto de la "multitudinaria" marcha de la confepa: t.co

El Caso cascadas o cómo el grupo logró quitarle 44 millones de dólares a miles de chilenos a través de sus pensiones t.co

[Hoy en t.co] Estrecho triunfo de Rousseff. El mercado vota hoy y se espera un frío recibimiento http://t.co/OlcAFQ4yeg

Y comienzan a aparecer los primeros rayos de sol en Valparaiso @biobio buen comienzo de semana 07:11 Hrs t.co

Bachelet defendió sus planes de reforma constitucional en revista alemana bit.ly http://t.co/hU6HqxpNU1

Ante tanta desinformación e intentos de manipulación, mi pequeño aporte: Diferencia entre REMUNERACIÓN y LUCRO: t.co

Junto a Presidenta Bachelet y delegación chilena reunidos con fundaciones públicas alemanas vinculadas con Chile. t.co

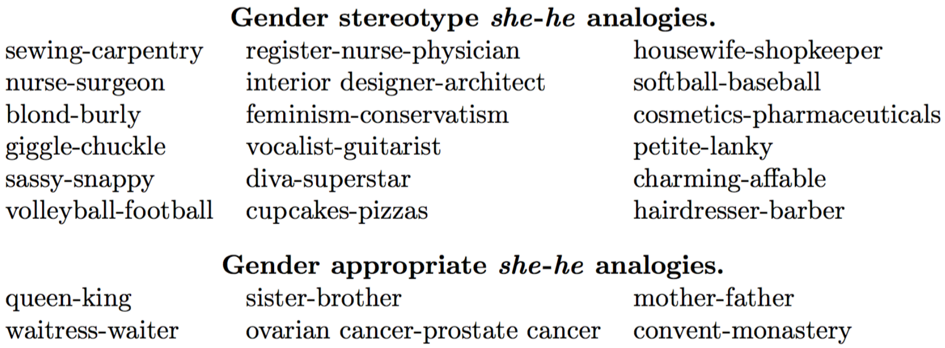
Equipos de Algarrobo, Villa Alemana, Viña del Mar y Los Andes a Congreso Nacional EXPLORA CONICYT

Por @carnby | Nota: no estamos afiliados con Twitter.

[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

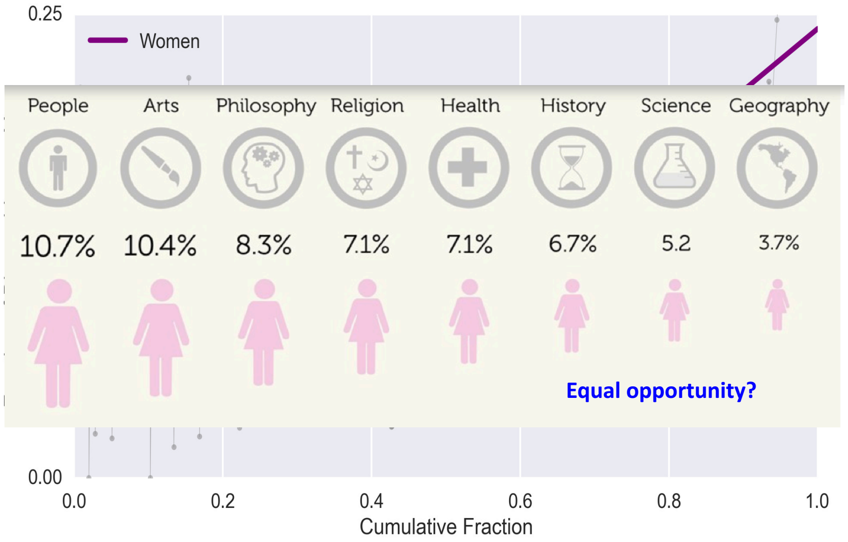
Most journalists are men?

[Bolukbasi at al, ArXiv 2016]

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

Gender Bias in Content

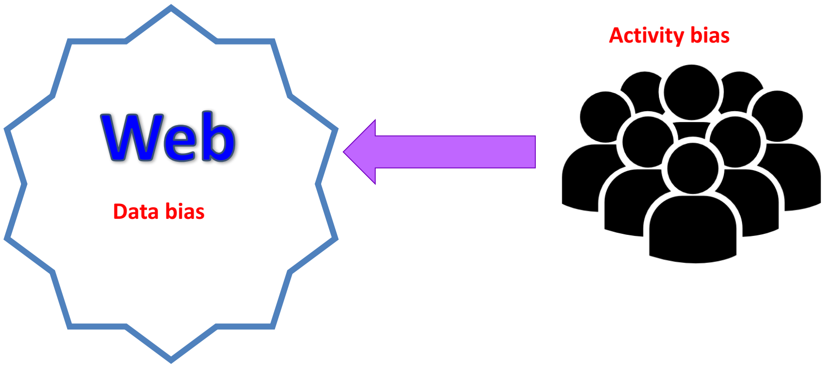
Systemic bias?



Equal opportunity?

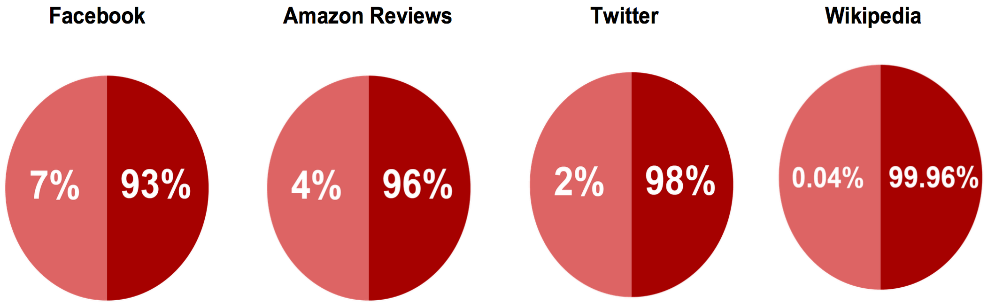
[E. Graells-Garrido et al,. "First Women, Second Sex: Gender Bias in Wikipedia", ACM Hypertext'15]

Bias in the Web

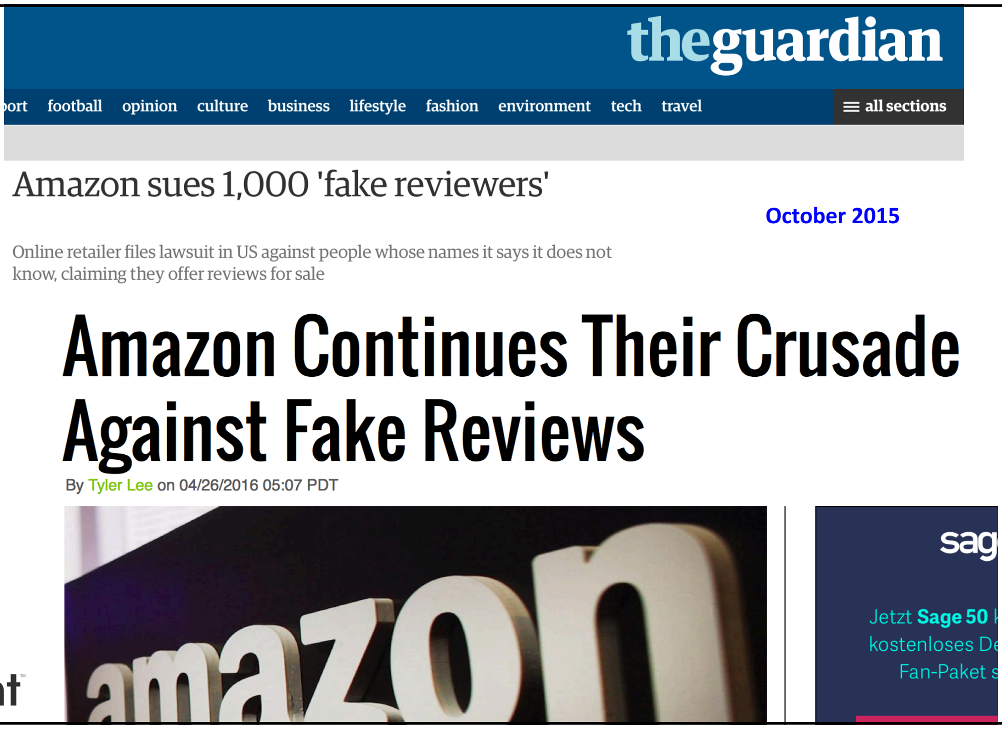


Activity Bias

Which percentage of users produce 50% of the content?



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


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

Amazon sues 1,000 'fake reviewers'

October 2015

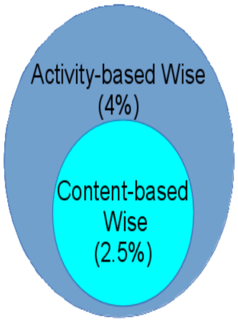
Online retailer files lawsuit in US against people whose names it says it does not know, claiming they offer reviews for sale

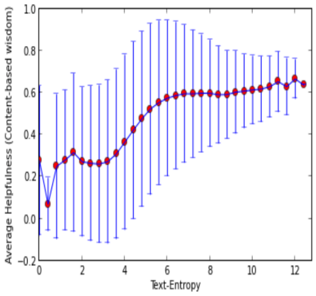
Amazon Continues Their Crusade Against Fake Reviews

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

Quality of Content?

- Adding content 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?



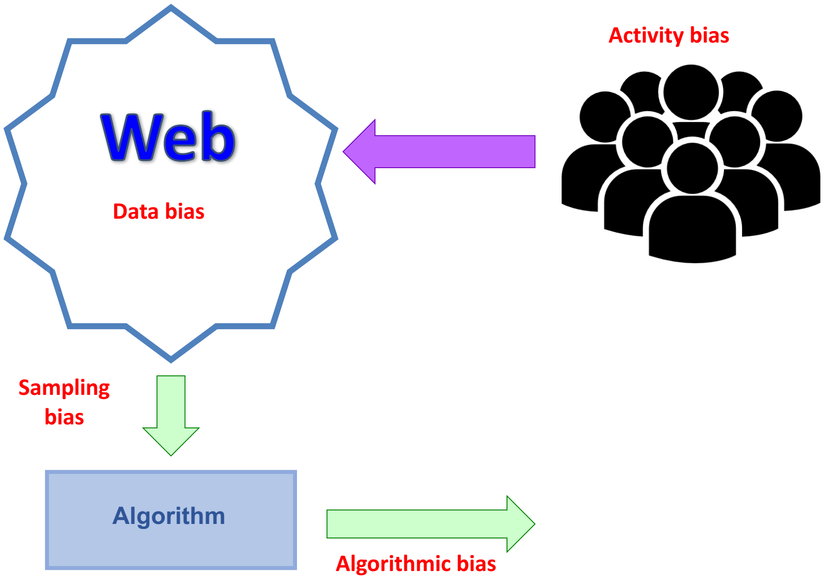


[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]

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 ϵ we can use the standard binomial error formula √(1-p)/*np* which works well for *p* near ½ **but not for p near 0**
- Better is the Agresti-Coull technique (also called *take 2*) which gives:

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

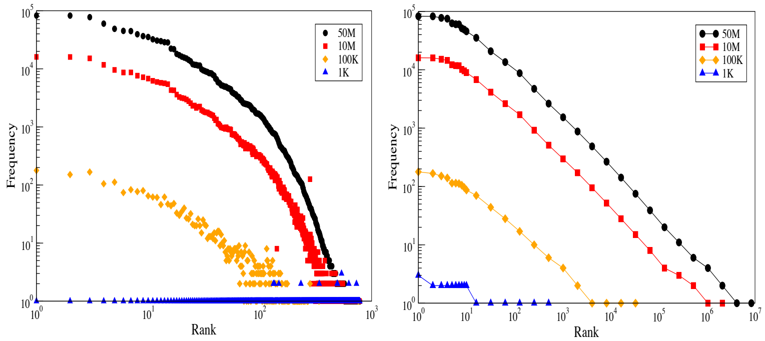
where *Z* is the inverse of the standard normal distribution, $1-\alpha$ is the confidence interval and $p' = p + Z^2/2$

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

[Baeza-Yates, SIGIR 2015, Industry track]

Sampling Techniques

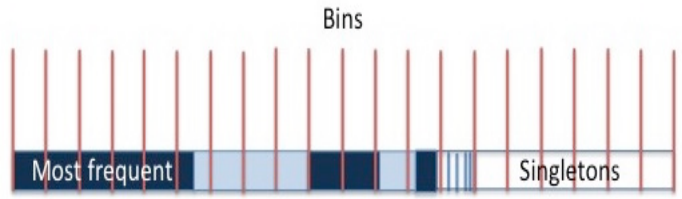
- Standard technique: $p_q \approx \widehat{p}_q(\mathcal{S}) = \dfrac{f_q(\mathcal{S})}{\sum_{q' \in \mathcal{S}} f_{q'}(\mathcal{S})}$
- A good sample should cover well all the query distribution but this does not work with very biased distributions.



[Zaragoza et al, CIKM 2010]

Incremental Stratified Sampling

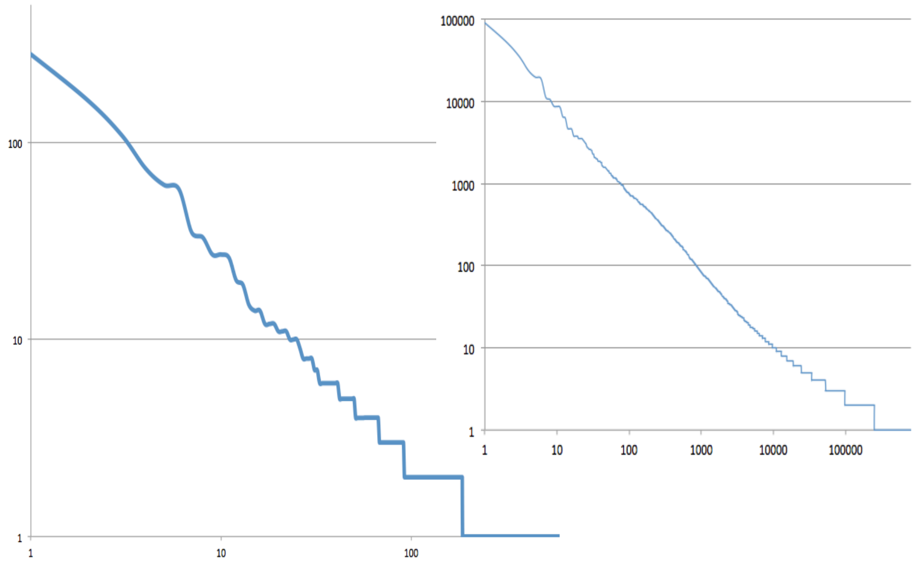
- Main goal: make good samples consistent across time
- Simple idea based in stratified sampling: bins + random start point



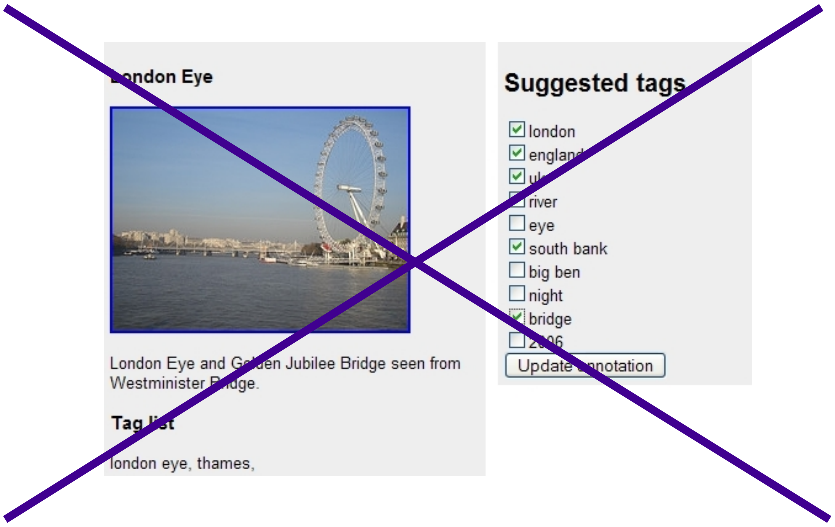
- Bin size can be found by binary search starting with a good approximation if a query frequency model is used (*b* < *V/n*)
- This perfectly mimics the head of the distribution, but not the tail
- Change the bins in the tail to get the right distribution

[Baeza-Yates, SIGIR 2015, Industry track]

Stratified Sampling Example



Extreme Algorithmic Bias



London Eye

London Eye and Golden Jubilee Bridge seen from Westminister Bridge.

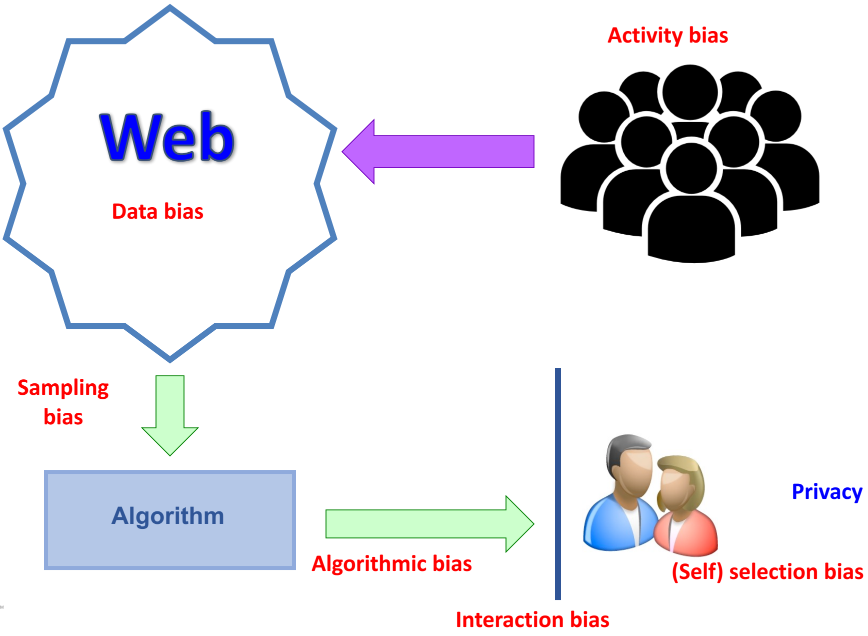
Tag list

london eye, thames,

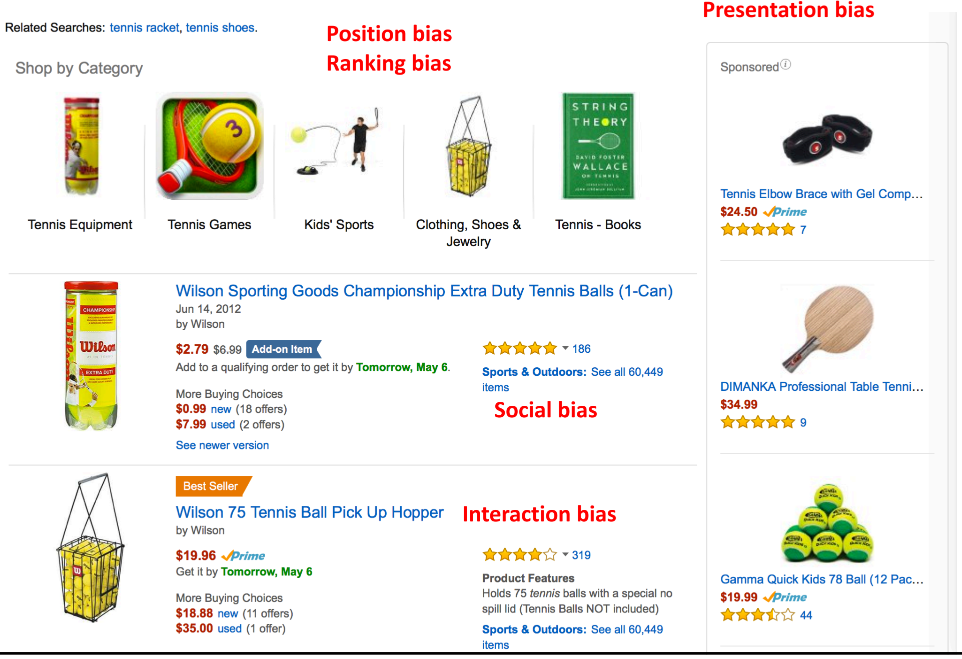
Suggested tags

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

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

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

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

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
★★★½☆ 44

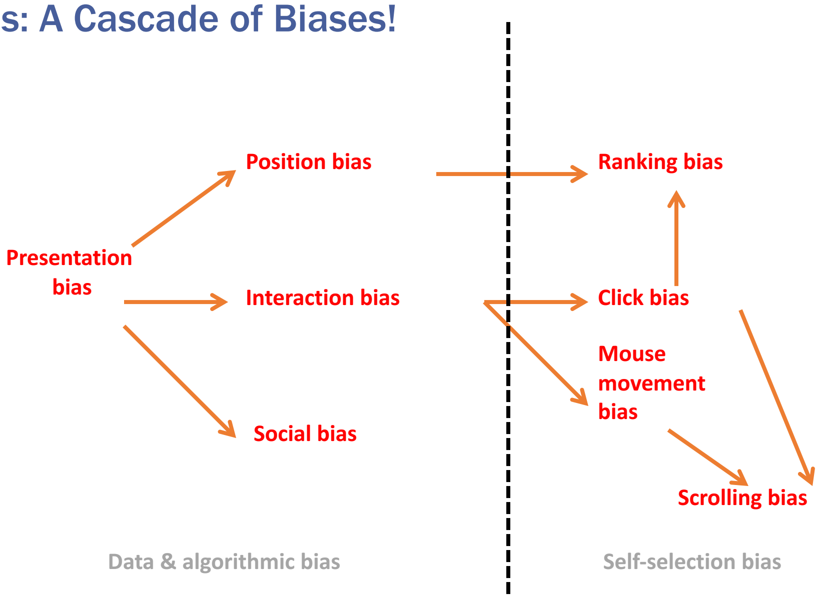
Presentation bias

Position bias
Ranking bias

Social bias

Interaction bias

Dependencies: A Cascade of Biases!



Position bias

Presentation bias

Interaction bias

Social bias

Ranking bias

Click bias

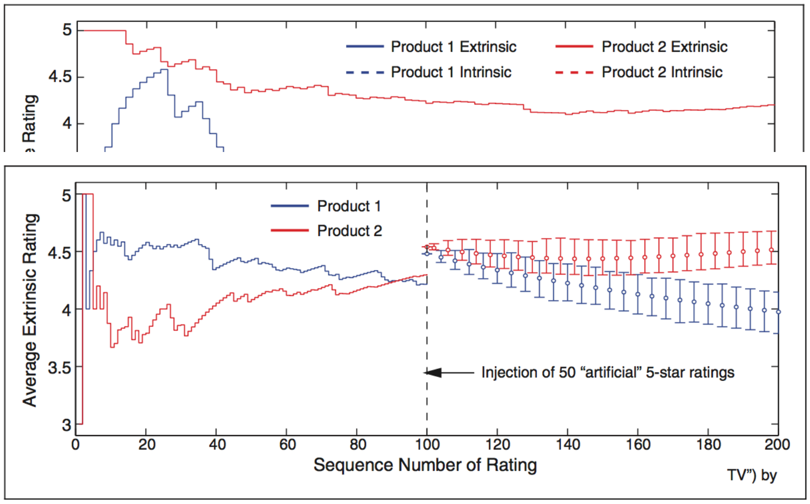
Mouse movement bias

Scrolling bias

Data & algorithmic bias

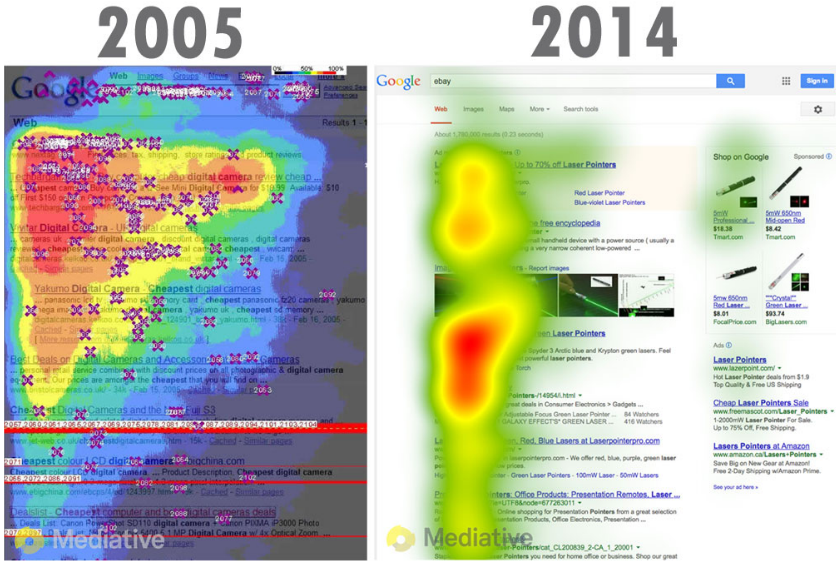
Self-selection bias

Social Bias



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

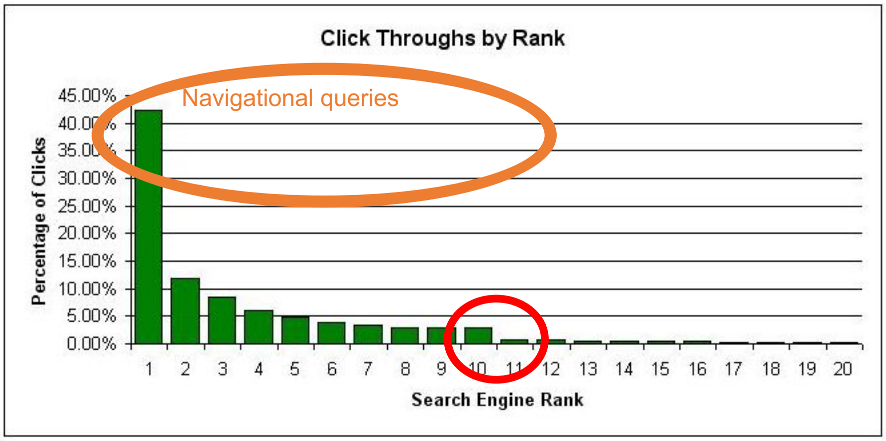
Ranking Bias in Web Search



[Mediative Study, 2014]

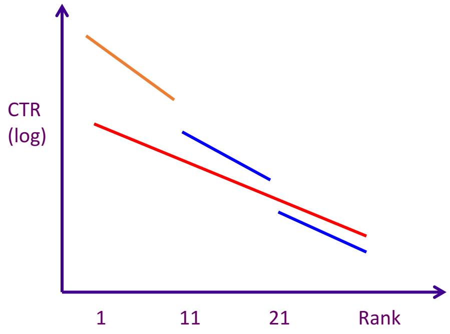
Click Bias in Web Search

- Ranking & next page bias

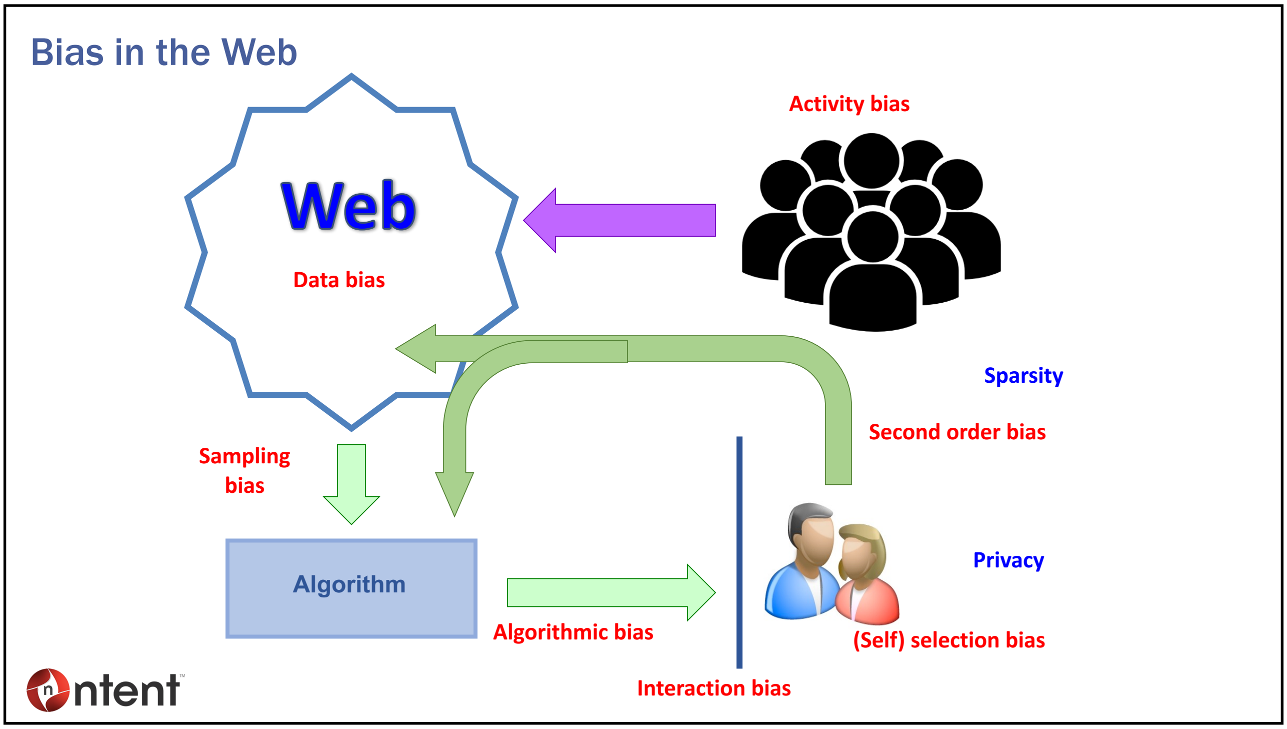


Unbiasing Search Clicks

Clicks as implicit positive user feedback



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



Bias in the Web

Web

Data bias

Activity bias

Sparsity

Second order bias

Sampling bias

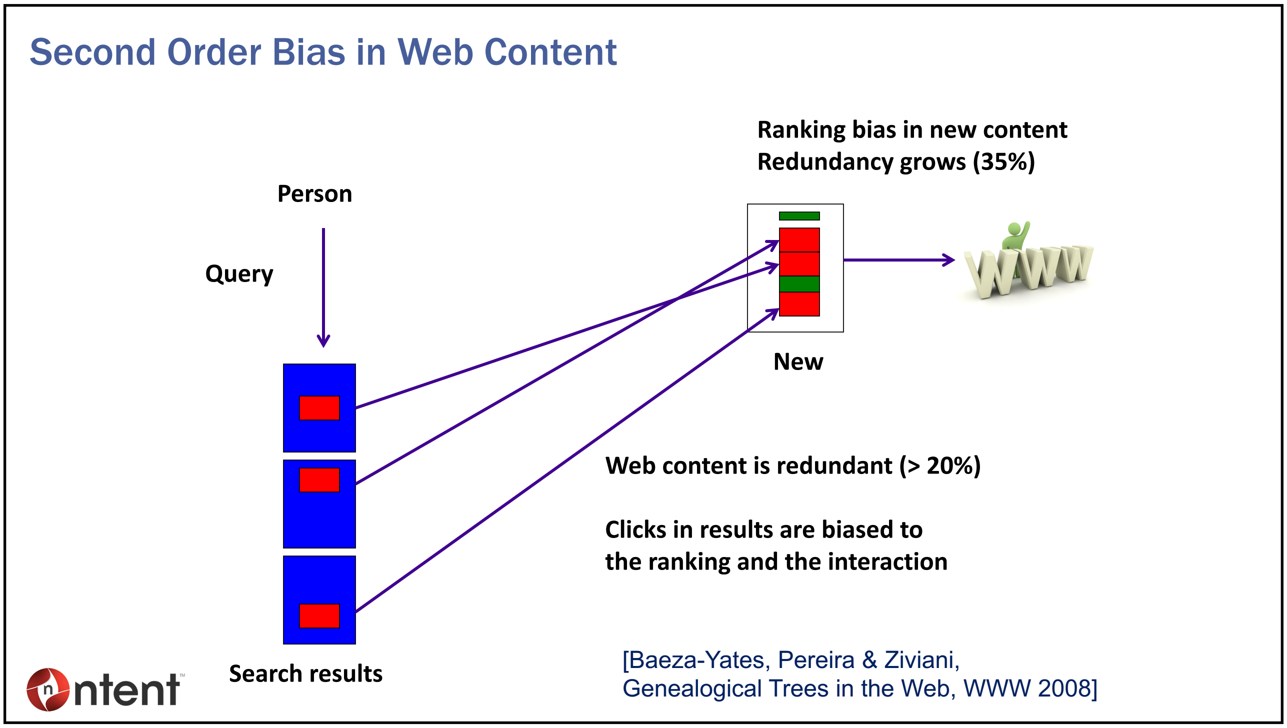
Algorithm

Privacy

Algorithmic bias

(Self) selection bias

Interaction bias



Second Order Bias in Web Content

Ranking bias in new content
Redundancy grows (35%)

Person

Query

New

Web content is redundant (> 20%)

Clicks in results are biased to the ranking and the interaction

Search results

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

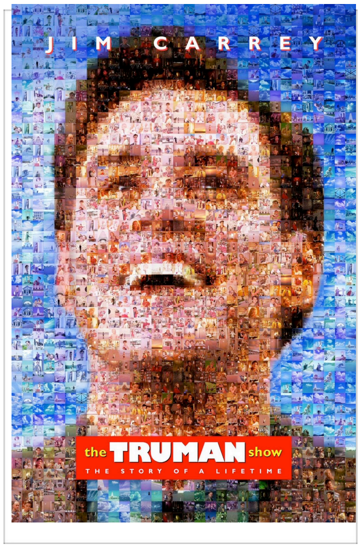
Avoid Second Order Bias due to Personalization

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?

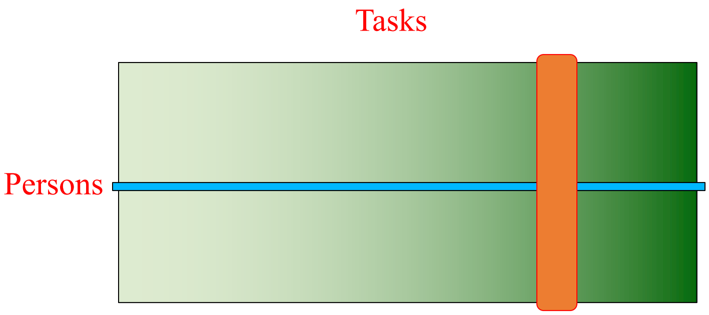
Aggregating in the Tail

- Exploit the context (and deep learning!)

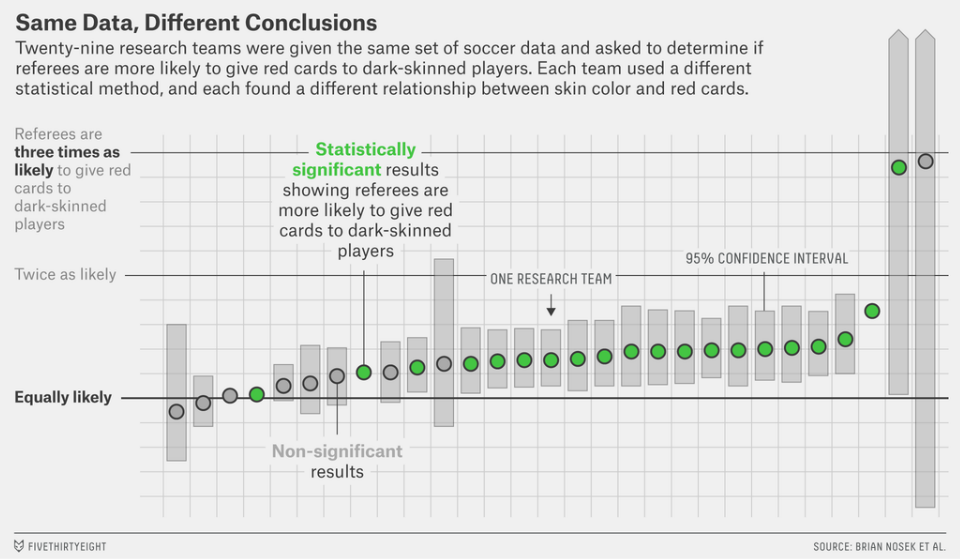
  91% accuracy to predict the next app you will use [Baeza-Yates et al, WSDM 2015]

- Personalization vs. Contextualization

  Recall that user interaction is another long tail



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



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

FIVETHIRTYEIGHT

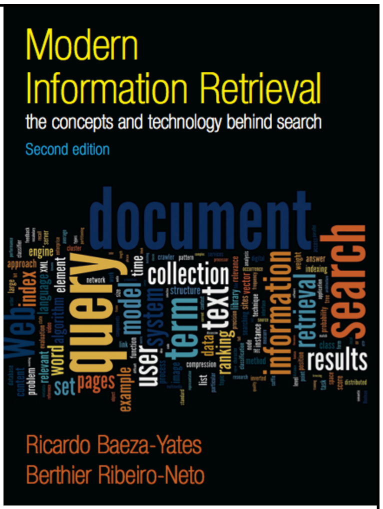
SOURCE: BRIAN NOSEK ET AL.

➔ 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

Biased Questions?

More bias: We are hiring in Barcelona & San Diego!