Agenda

- WSDM's Birth
- Analysis of Ten Years
- Current & Future Challenges
  - Contextual Multilingual Semantic Search
  - Bias in the Web

WSDM's Birth and Governance (1/2)

- **May 12, 2006**: e-mail from Ziv Bar-Yossef (Google Haifa) and Junghoo Cho (UCLA)
- **Initial SC**: Rakesh Agrawal (Microsoft), Ricardo Baeza-Yates (Yahoo), Krishna Bharat (Google), Andrei Broder (Yahoo), Soumen Chakrabarti (IIT Bombay), Monika Henzinger (EPFL, Google), Jon Kleinberg (Cornell), Rajeev Motwani (Stanford), Prabhakar Raghavan (Yahoo)
- **May 25, 2006**: meeting at WWW 2006 in Edinburgh.
  Most of the SC plus Sue Dumais (Microsoft Research), Ravi Kumar & Andrew Tomkins (Yahoo), Ronny Lempel (IBM), Yoelle Maarek (Google), Mark Manasse (Microsoft), Marc Najork (Microsoft), and Torsten Suel (Polytechnic University)
- **SIGs proposed:** SIGACT, SIGIR, SIGKDD, SIGMOD, and SIGWEB
- **Plan A**: **September 2007**, with the advice from ACM President, Stu Feldman, to request the SIGWEB sponsorship.
- **Early 2007 conversations**: SIGWEB (Ronny) and SIGMOD (Andrew), plus SIGIR & SIGKDD (myself).
- Asked to be conference chair for the first conference with Andrei and Soumen as PC-Chairs, Ravi as treasurer, and Utkarsh Srivastava of Stanford as local organizer

WSDM's Birth and Governance (2/2)

- Our plans changed during my negotiation with SIGIR (Jamie Callan), as the conference was between SIGIR and CIKM, two SIGIR sponsored conferences
- **Plan B**: **second week of February 2008**, with the second conference planned in **Barcelona**.
  As I was the natural chair for 2009, I stepped down for 2008, being replaced by Marc Najork
- **Driving team**: Andrei, Andrew, Junghoo, Marc, Ravi, Ricardo, Ronny, Soumen & Zvi
- **April 2007**: formal support of SIGKDD and SIGMOD
  However SIGIR wanted to have a steering committee with its own representative and clear governance rules
- **May 2007:** Andrei, Andrew, Marc, Ravi, Ricardo, Soumen, and Ziv discussed the governance rules at WWW 2007 in Banff, Canada
- **Steering committee rules**: 8 members, 4 of them representing the sponsoring SIGs, balancing industry and academia members, with periods of 4 years
- With the new rules, SIGIR and SIGWEB became sponsors too

WSDM's Formal Steering Committee

June 2007:

Rakesh Agrawal (Microsoft, KDD rep.)
Ziv Bar-Yossef (Google)
Monika Henzinger (EPFL, Google)
Rajeev Motwani (Stanford)
Ricardo Baeza-Yates (Yahoo, SIGIR rep., chair)
Soumen Chakrabarti (IIT Bombay)
Jon Kleinberg (Cornell)
Prabhakar Raghavan (Yahoo)

January 2010: Marc Najork (Microsoft) replaced Prabhakar, and soon after Hector Garcia-Molina (Stanford) joined representing SIGMOD

May 2013: SC Renewal, almost 2 years late!

Ricardo Baeza-Yates (Yahoo, SIGIR rep., chair)
Andrei Broder (Google)
Nick Koudas (Univ. of Toronto, SIGMOD rep.),
Marc Najork (Microsoft)
Paolo Boldi (Univ. de Milano)
Brian Davison (Lehigh Univ., SIGWEB rep.)
Bing Liu (UIC, SIGKDD rep.)

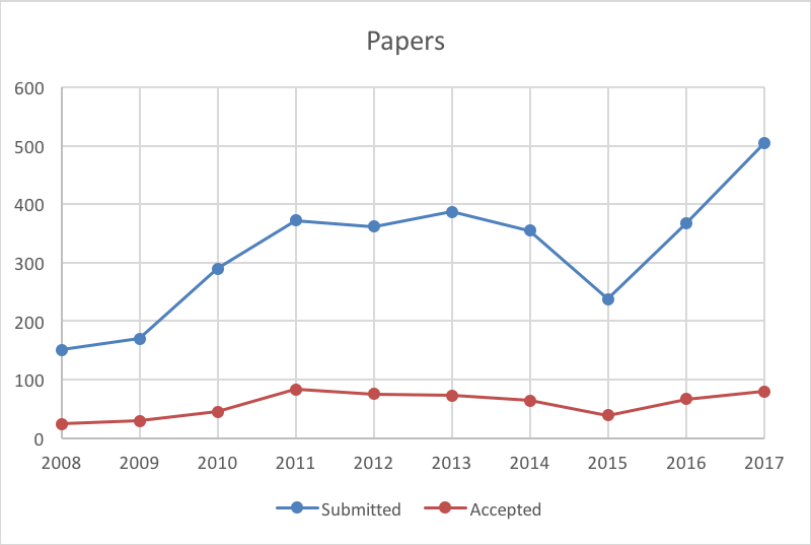
January 2014: Hang Li (Hiawei) joined to complete the SC and soon after I proposed Marc Najork (Google) as new chair, who was elected by acclamation!

2017: SC should be partially renewed

WSDM's History

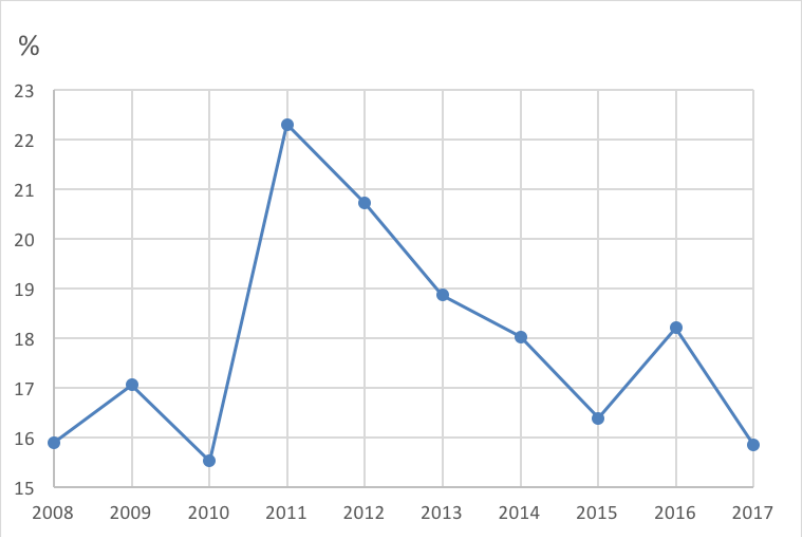
Year	Location	Conference Chair	PC Chairs
2008	Palo Alto	Marc Najork	Andrei Broder; Soumen Chakrabarti
2009	Barcelona	Ricardo Baeza-Yates	Paolo Boldi; Berthier Ribeiro-Neto
2010	New York	Brian Davison; Torsten Suel	Nick Craswell; Bing Liu
2011	Hong Kong	Irwin King	Wolfgang Nejdl; Hang Li
2012	Seattle	Eytan Adar; Jaime Teevan	Eugene Agichtein; Yoelle Maarek
2013	Rome	Stefano Leonardi; Alessandro Panconesi	Paolo Ferragina; Aristides Gionis
2014	New York	Ben Carterette; Fernando Diaz	Carlos Castillo; Donald Metzler
2015	Shanghai	Xueqi Cheng; Hang Li	Evgeniy Gabrilovich; Jie Tang
2016	San Francisco	Paul Bennet; Vanja Josifovski	Jennifer Neville; Filip Radlinski
2017	Cambridge	Milad Shokouhi; Maarten de Rijke	Andrew Tomkins; Min Zhang

WSDM's Growth: Papers



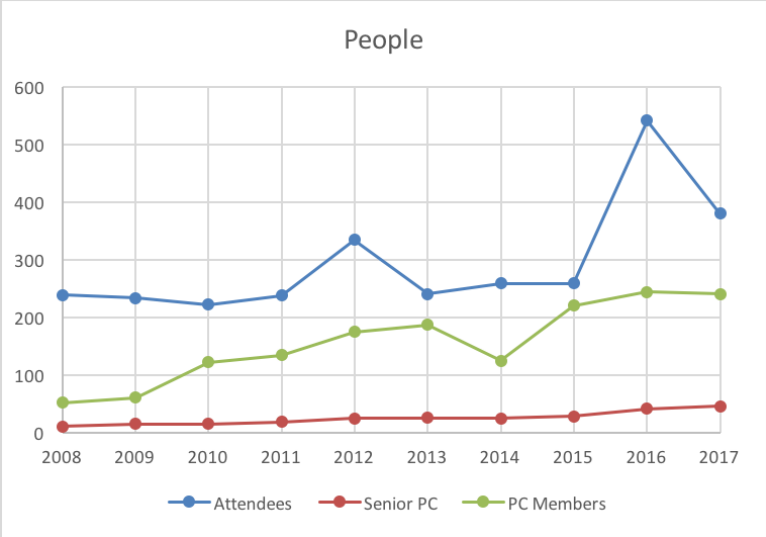
ntent

WSDM's Quality: Acceptance Rate



ntent

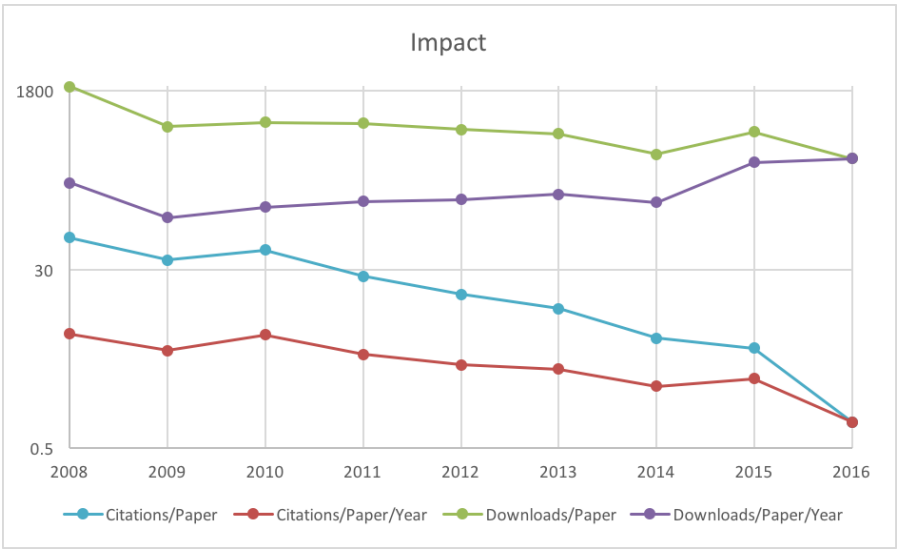
WSDM's Growth: People



ntent

WSDM's Impact: ACM Citations & Downloads (Dec 2016)

100 to 380 downloads per year!



ntent

WSDM's Places

2008	2011	2014	2017
University	University	University	University
Research	Research	Research	Research
Microsoft	Yahoo	Microsoft	Technology
Yahoo	Microsoft	Institute	Microsoft
IBM	Laboratories	Google	Science(s)
Laboratory	Institute	UIUC	California
Indiana	Stanford	Technology	State
Illinois	Technology	Yahoo	Yahoo
Technology	Corporation	Tsinghua	UIUC
State	Chinese	CMU	UI at Chicago

WSDM's Topics

2008	2011	2014	2017
Search	Web	Search	Social Networks
Web	Search	Web	Search
Ranking	Social	Data	Learning
Mining	Mining	Social Networks	Time
Classification	Query	Modeling	IR
Advertising	Analysis	Learning	Modeling
Models	Temporal	Systems	Text
Document	Online	Click	Data Mining
Graph	Learning	Advertising	Recom. Systems
Analysis	Data	Mining	Embedding

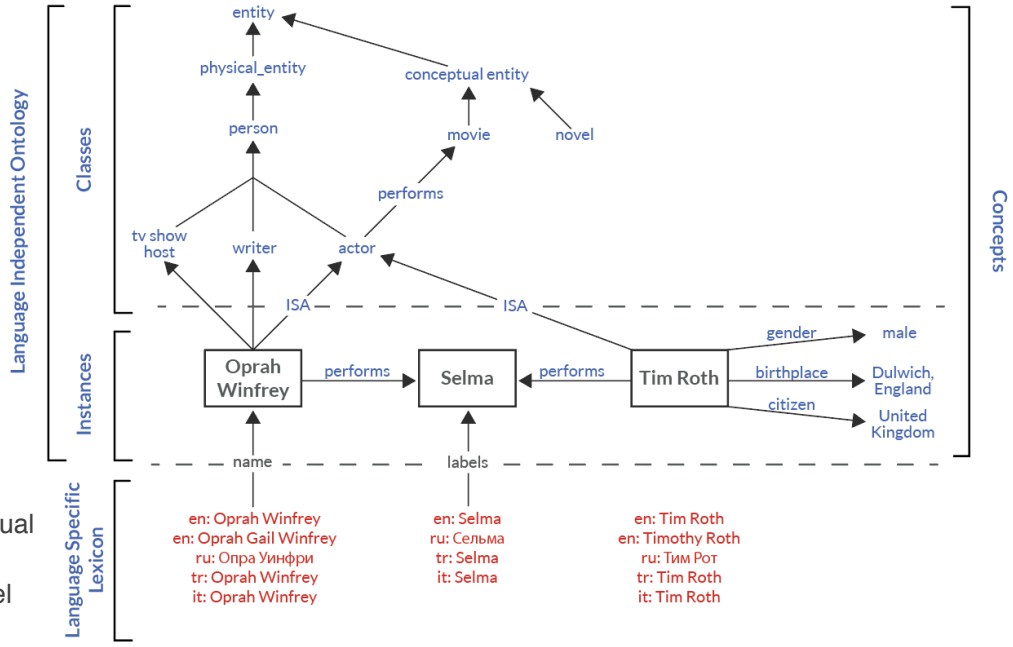
Current & Future Challenges

- Contextual Multilingual Semantic Search
- Fake web content
- Bias in the Web

Contextual Multilingual Semantic Search

- Multilingual Knowledge Base
- Contextual Query Understanding
- Semantic Ranking

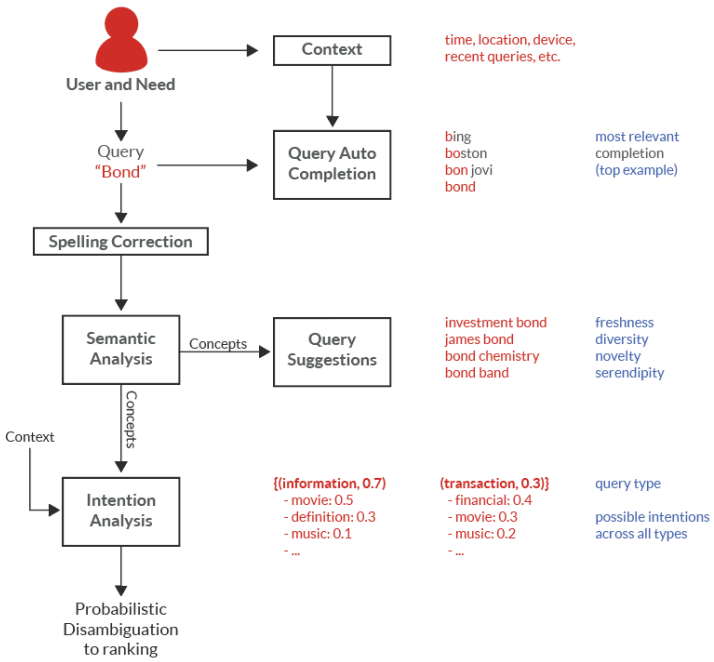
Knowledge Base



Challenge: Multilingual

Only at Lexicon level

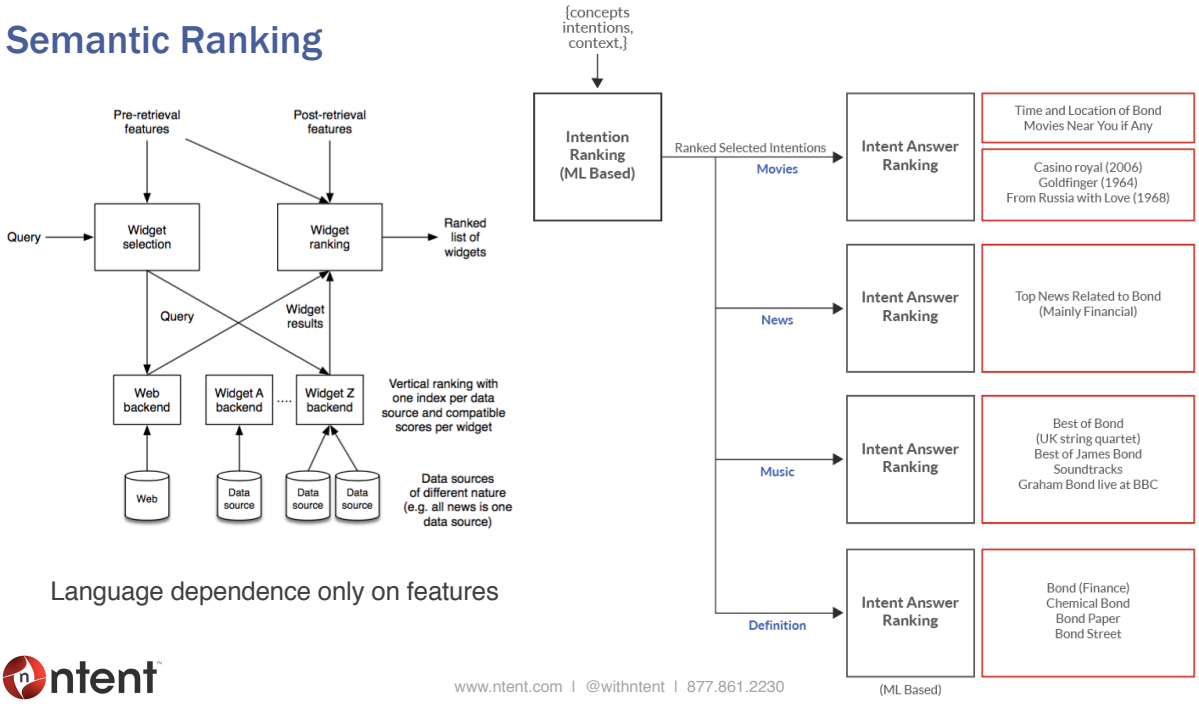
Query Understanding



Challenge: Multilingual

A NLP pipeline might be needed for semantic features

Semantic Ranking



Language dependence only on features

Fake Content & Bias

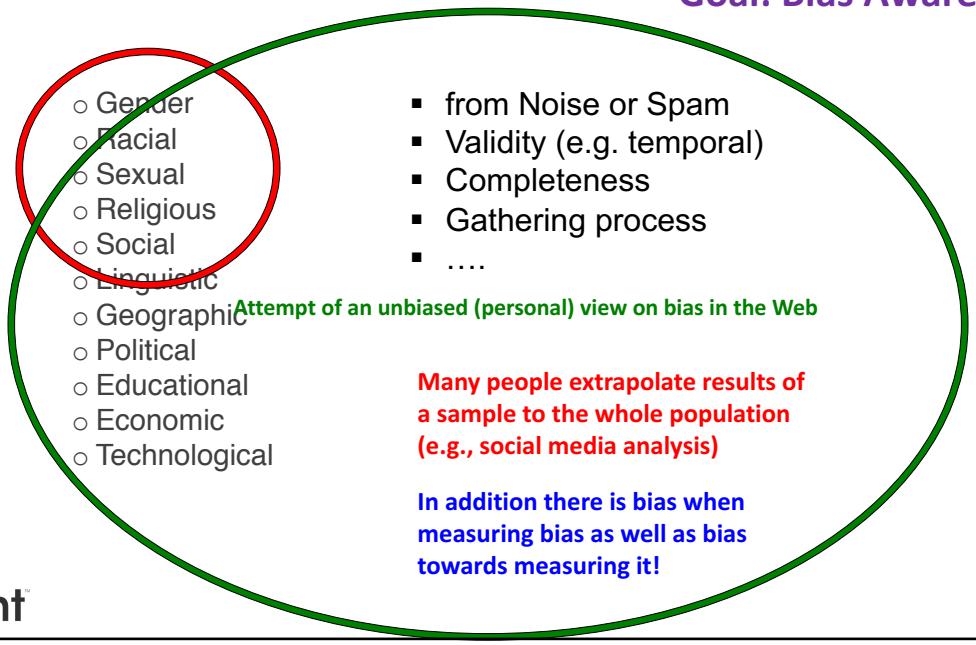
- British Prime Minister Benjamin Disraeli:
  - "There are three kinds of **lies**: **lies**, damned **lies**, and **statistics**.

Buzzfeed News

Bias: significant deviation from a prior (unknown) distribution

(Observational) Human Data has Bias

Goal: Bias Awareness

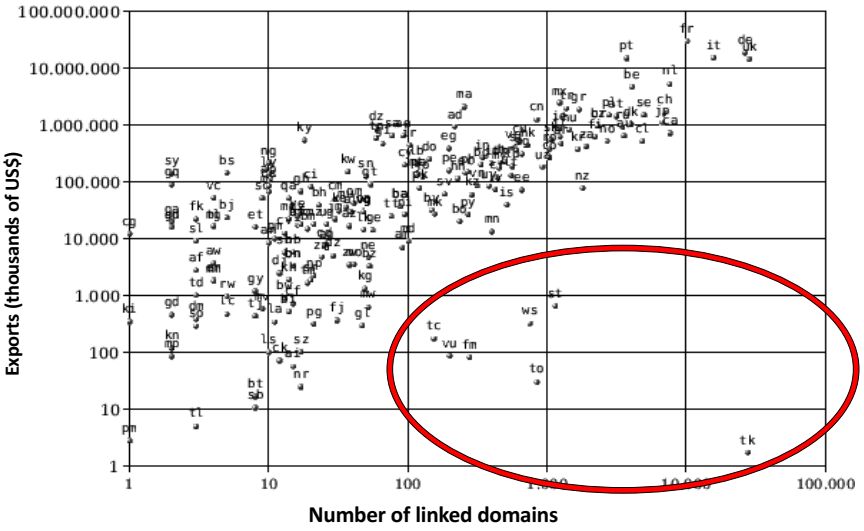


ntent

Bias in the Web

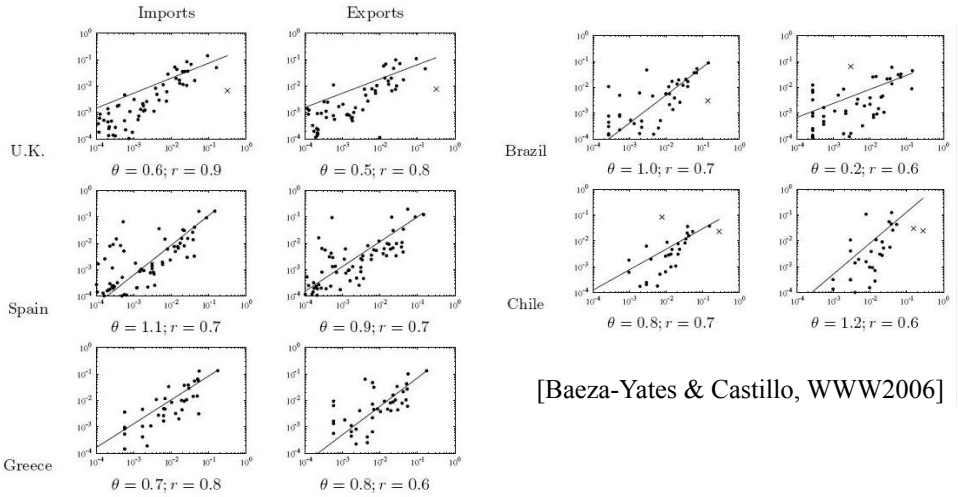
ntent

Economic Bias in Links



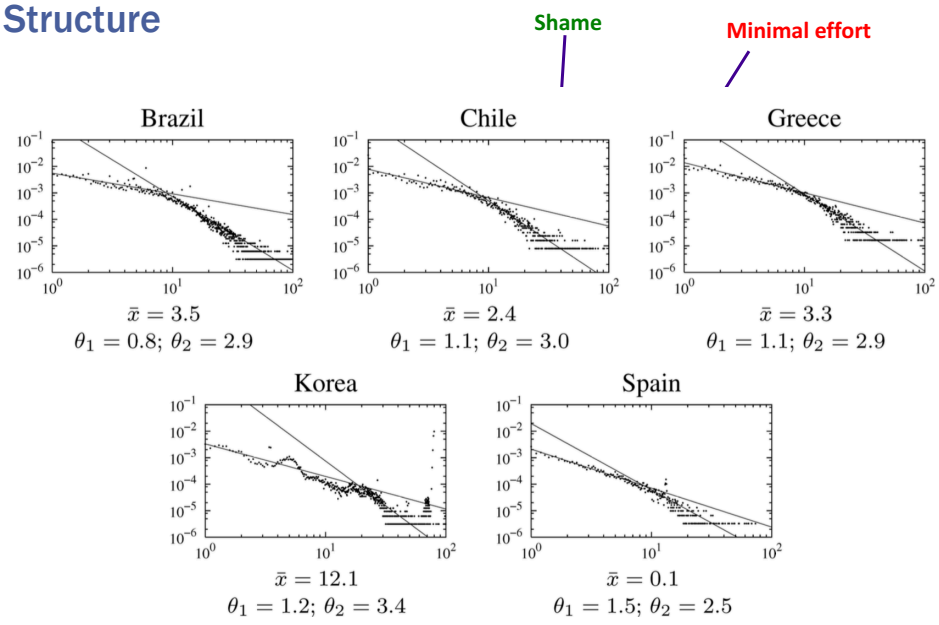
[Baeza-Yates, Castillo & López. Characteristics of the Web of Spain. Cybermetrics, 2005]

Economic Bias in Links



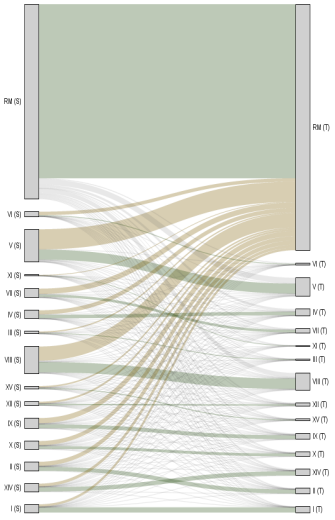
[Baeza-Yates & Castillo, WWW2006]

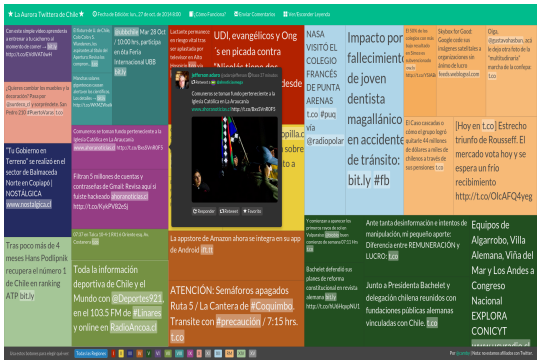
Website Structure



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

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

[Bolukbasi at al, ArXiv 2016]

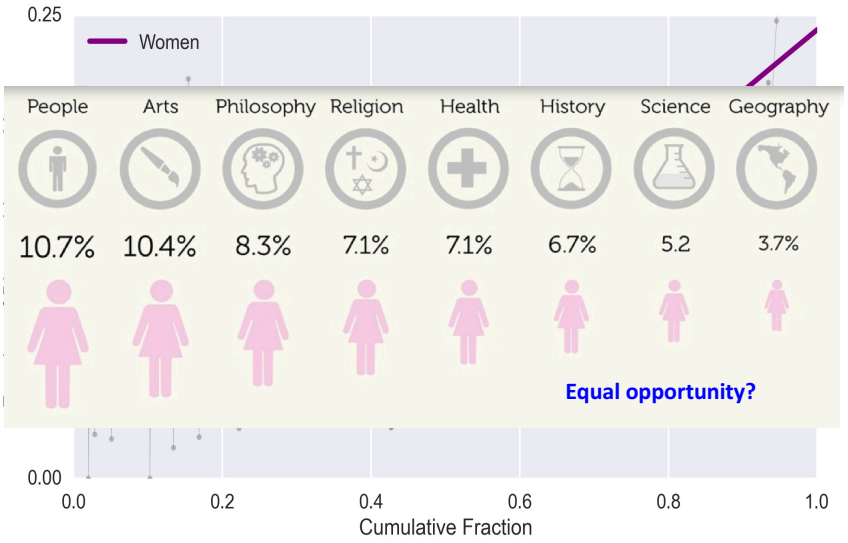
Most journalists are men?

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

ntent

Gender Bias in Content

Systemic bias?

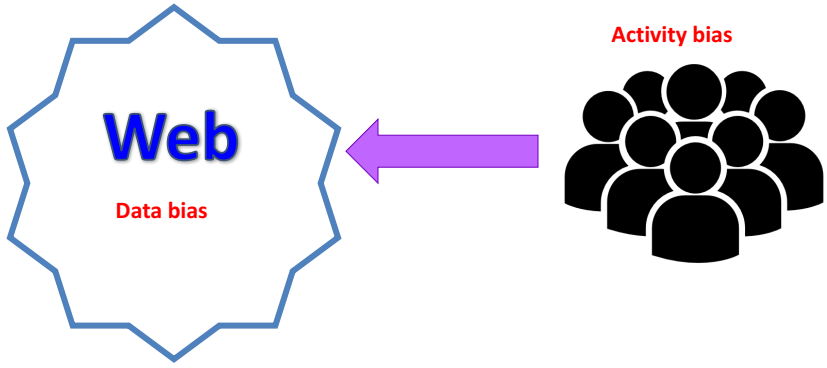


Equal opportunity?

ntent

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

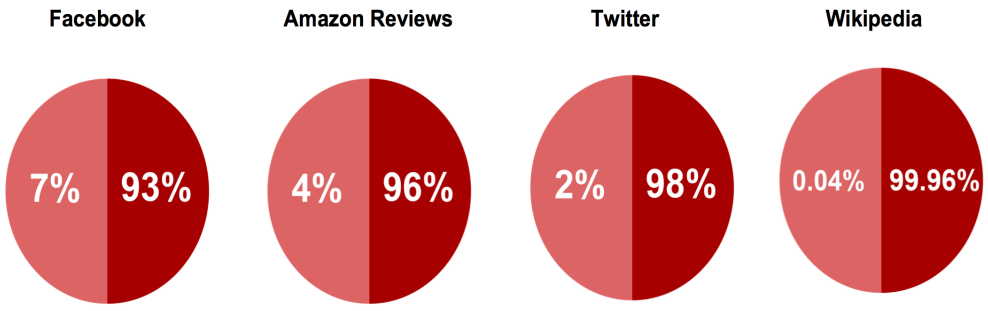
Bias in the Web



ntent

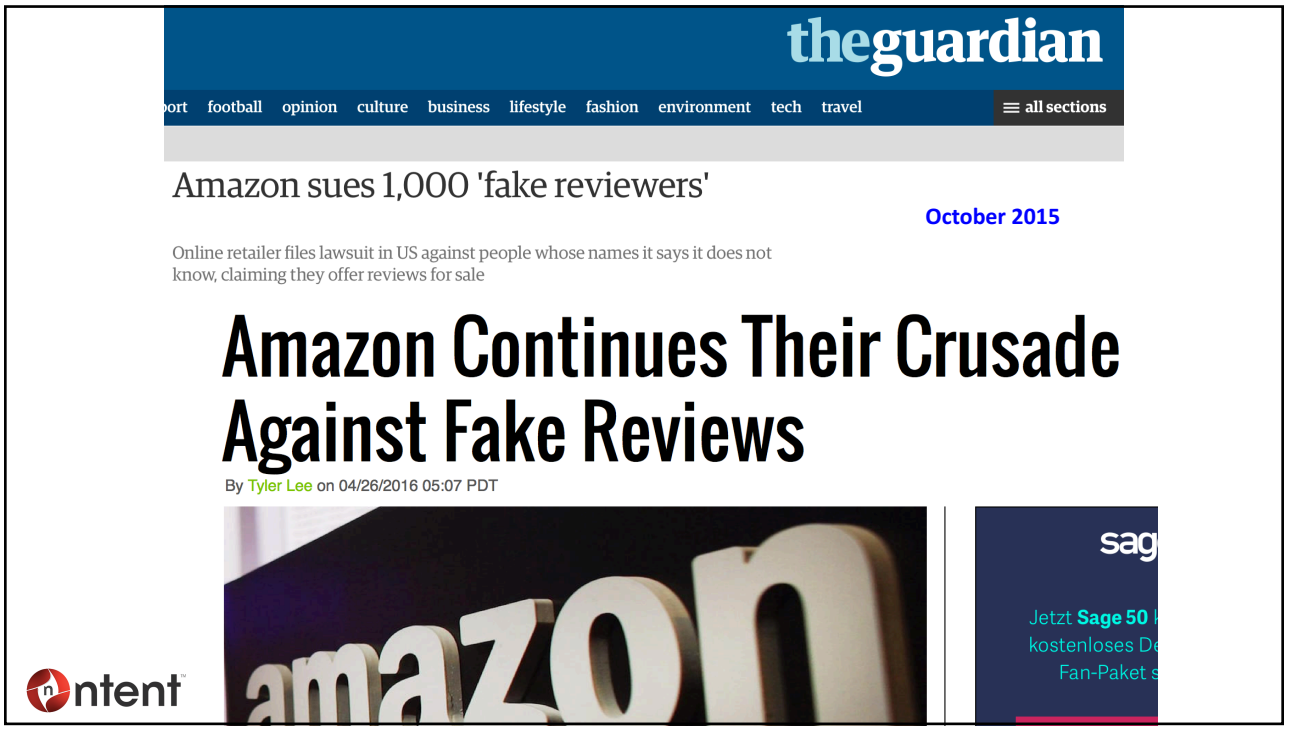
Activity Bias

Which percentage of users produce 50% of the content?



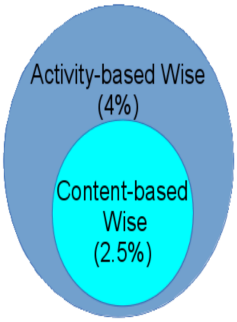
[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]

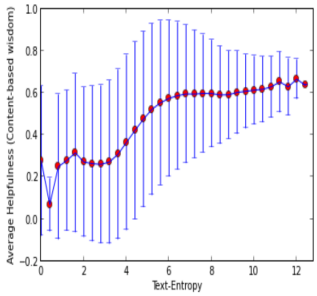
ntent



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?





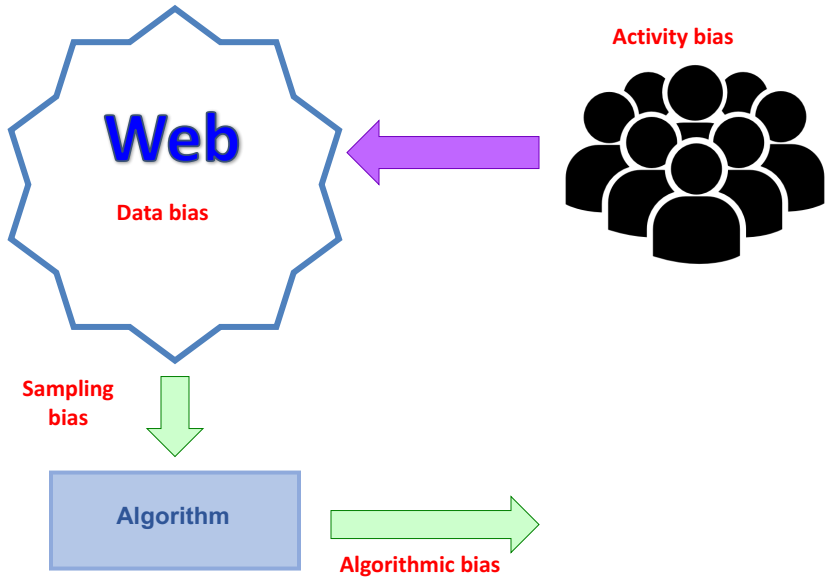
ntent

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

Content that is never seen: Digital Desert

ntent

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



ntent

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 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 $p' = p + Z^2/2$

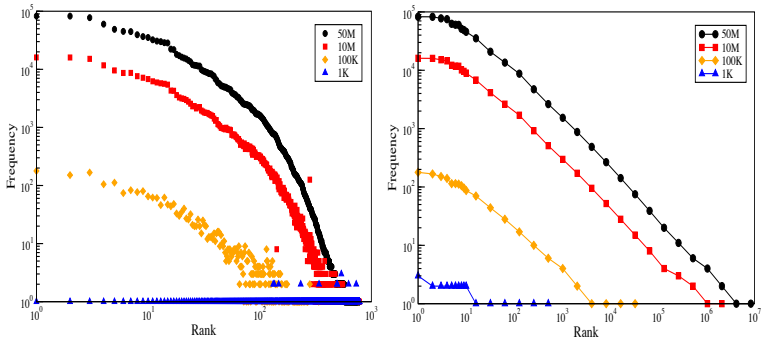
- If p = 0.1, $1-\alpha$ is 90% and ϵ is 10%, we get n = 2342. The standard formula gives n = 900!

ntent

[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.

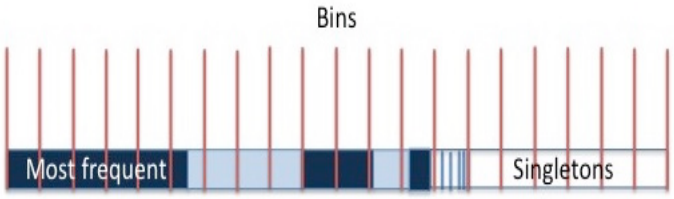


ntent

[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

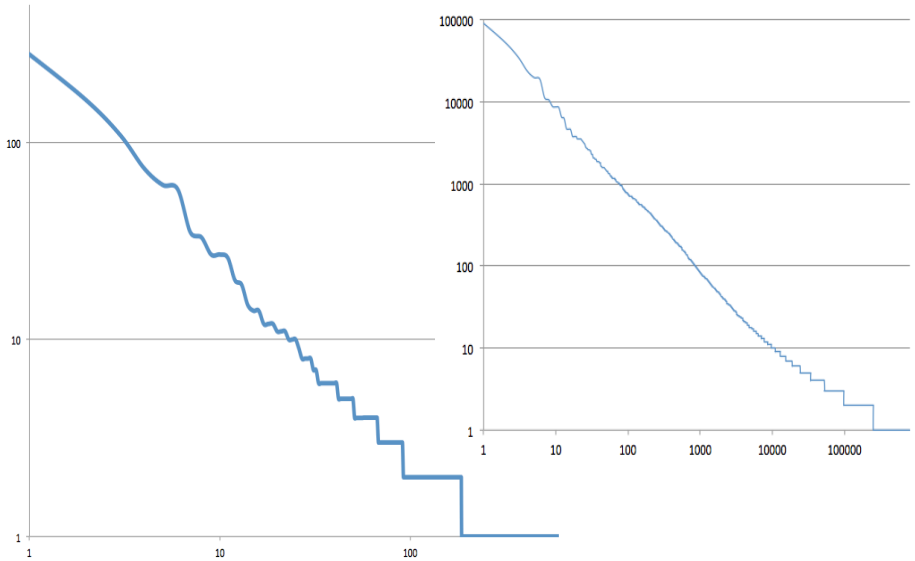


- Bin size can be found by binary search starting with a good approximation if a query frequency model is used ($b < V/n$)
- This perfectly mimics the head of the distribution, but not the tail
- Change the bins in the tail to get the right distribution

ntent

[Baeza-Yates, SIGIR 2015, Industry track] 37

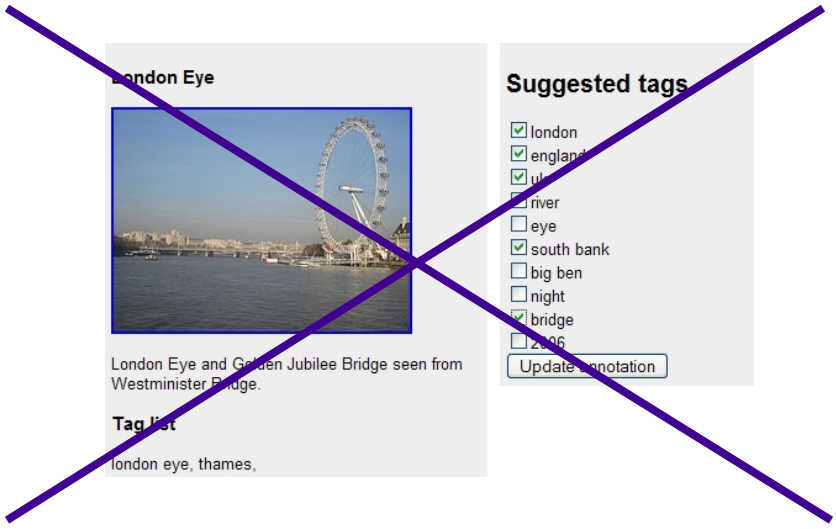
Stratified Sampling Example



ntent

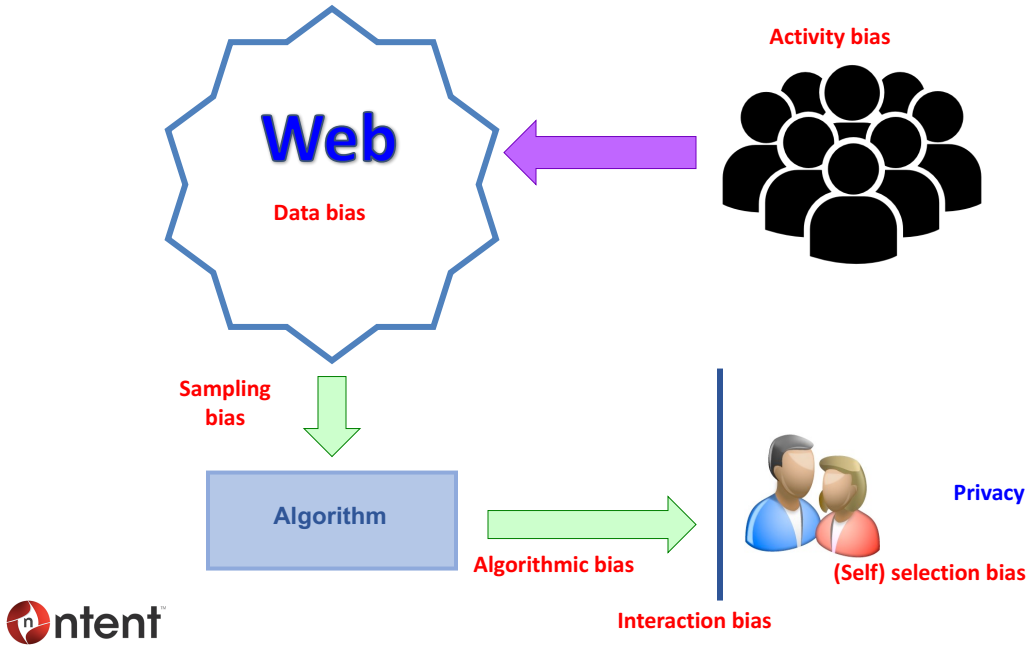
39

Extreme Algorithmic Bias



ntent

Bias in the Web

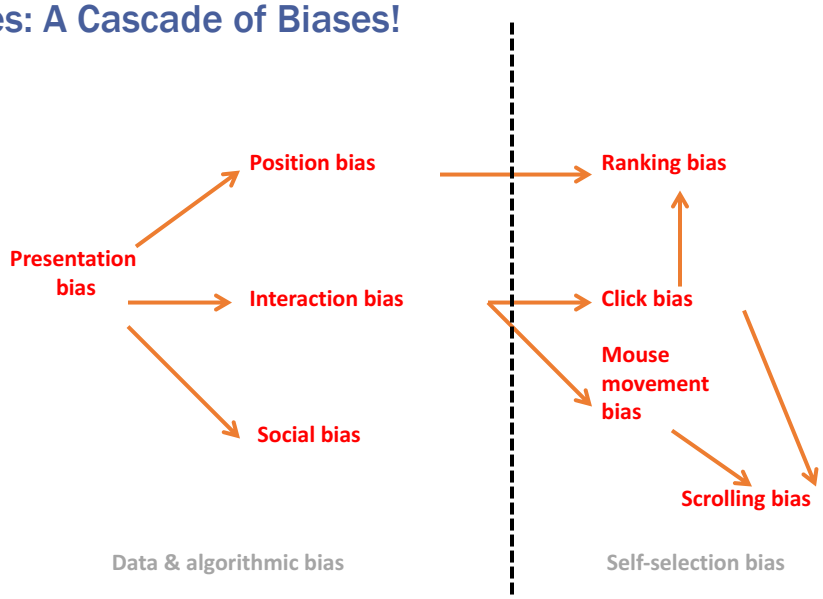


ntent

Bias in the Interaction

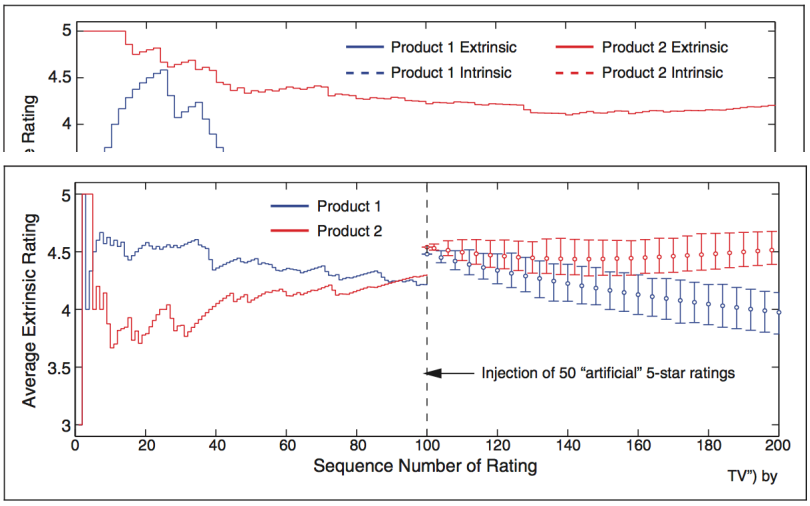
nten

Dependencies: A Cascade of Biases!



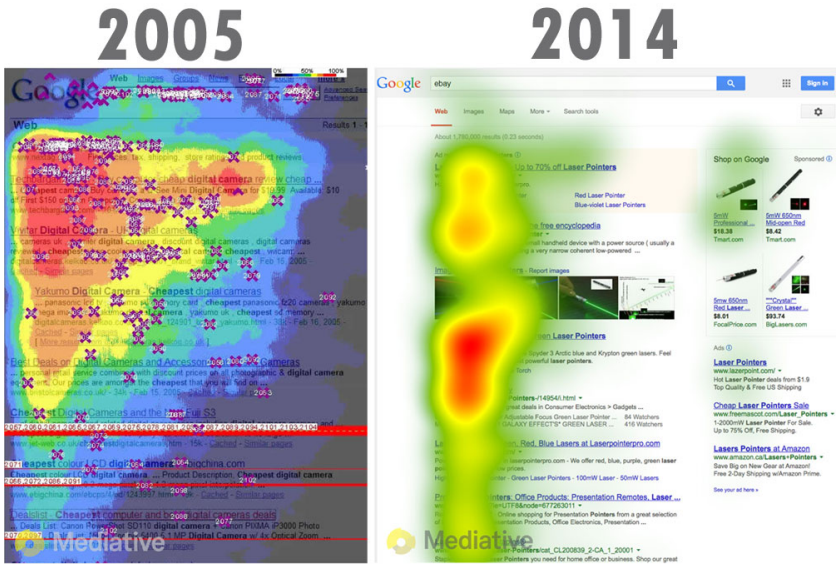
ntent

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]

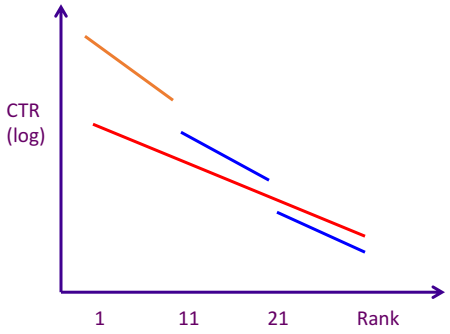
Click Bias in Web Search

- Ranking & next page bias

ntent

Unbiasing Search Clicks

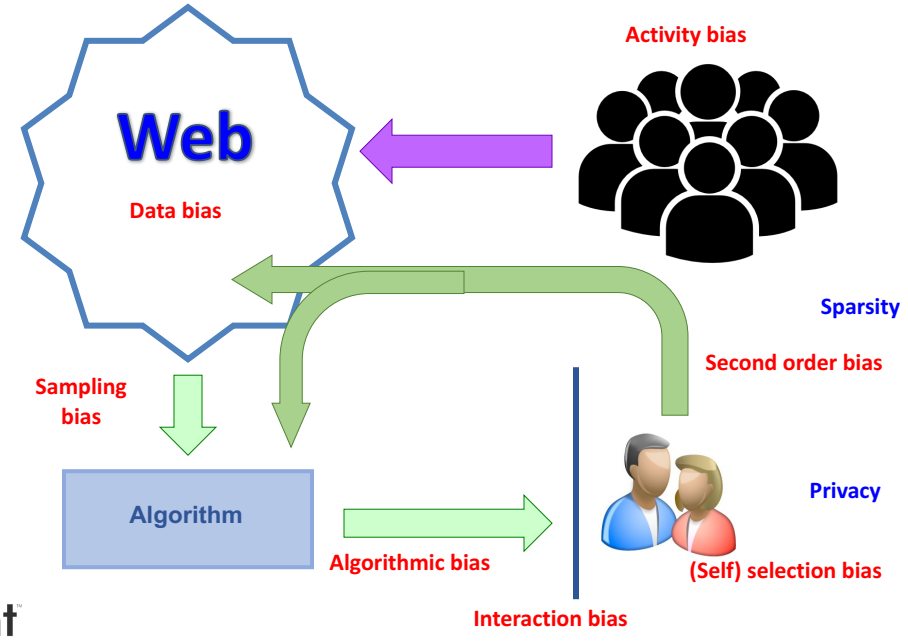
Clicks as implicit positive user feedback



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

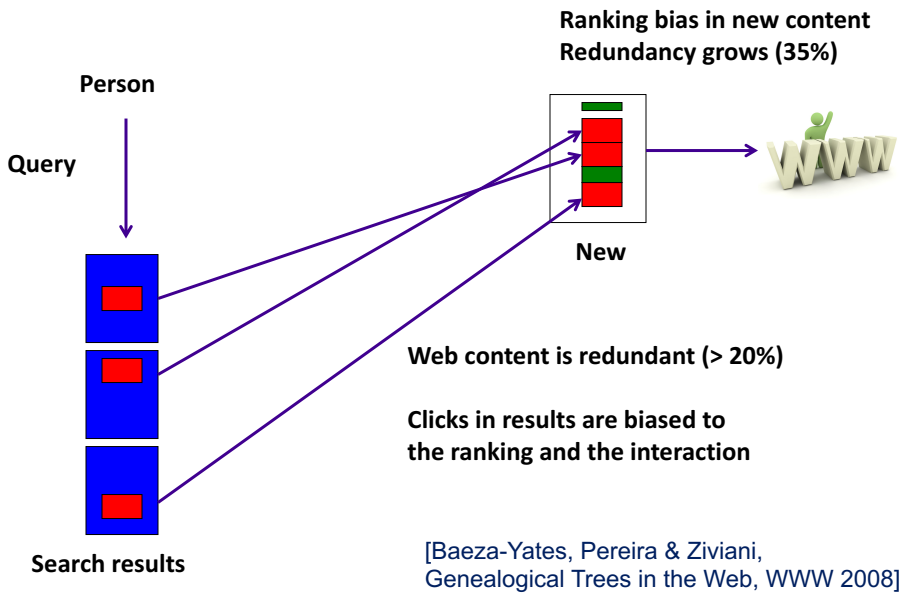
ntent

Bias in the Web



ntent

Second Order Bias in Web Content



Web content is redundant (> 20%)

Clicks in results are biased to the ranking and the interaction

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

ntent

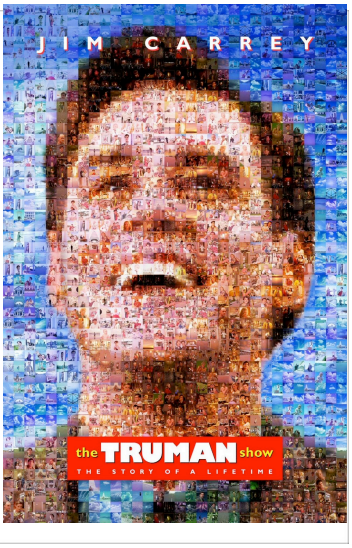
Avoid Second Order Bias due to Personalization

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?

ntent

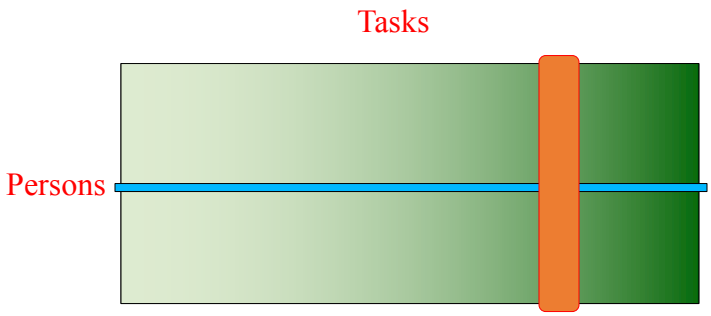
Aggregating in the Tail

- Exploit the context (and deep learning!)

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

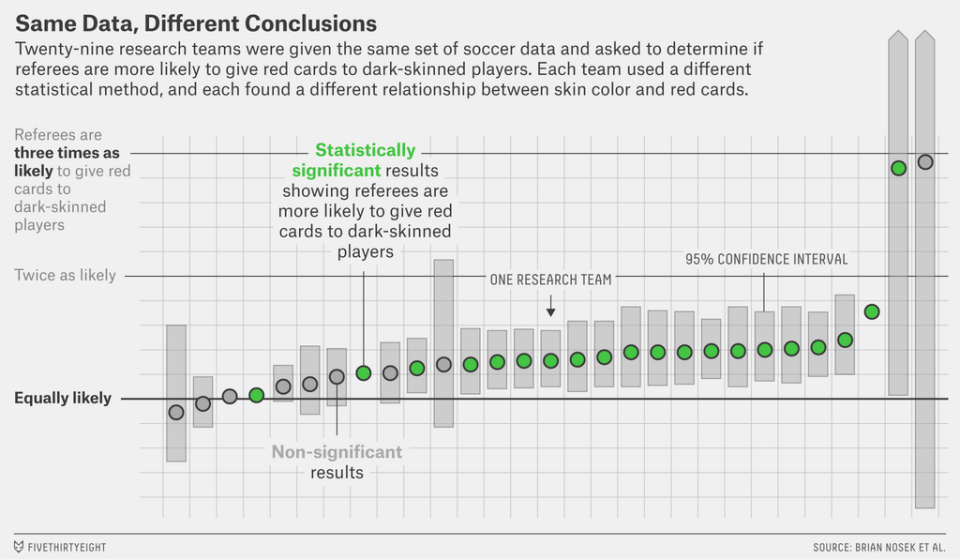
- Personalization vs. Contextualization

  Recall that user interaction is another long tail



ntent

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

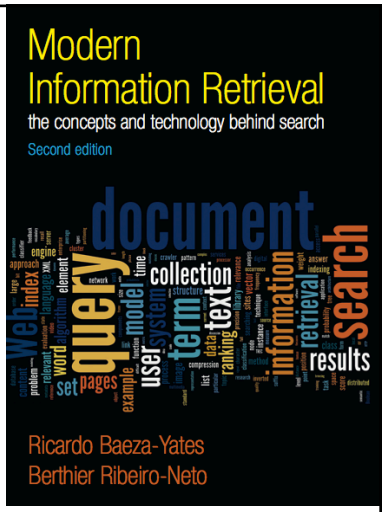


ntent

- ➔ 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

ntent

Biased Questions?

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