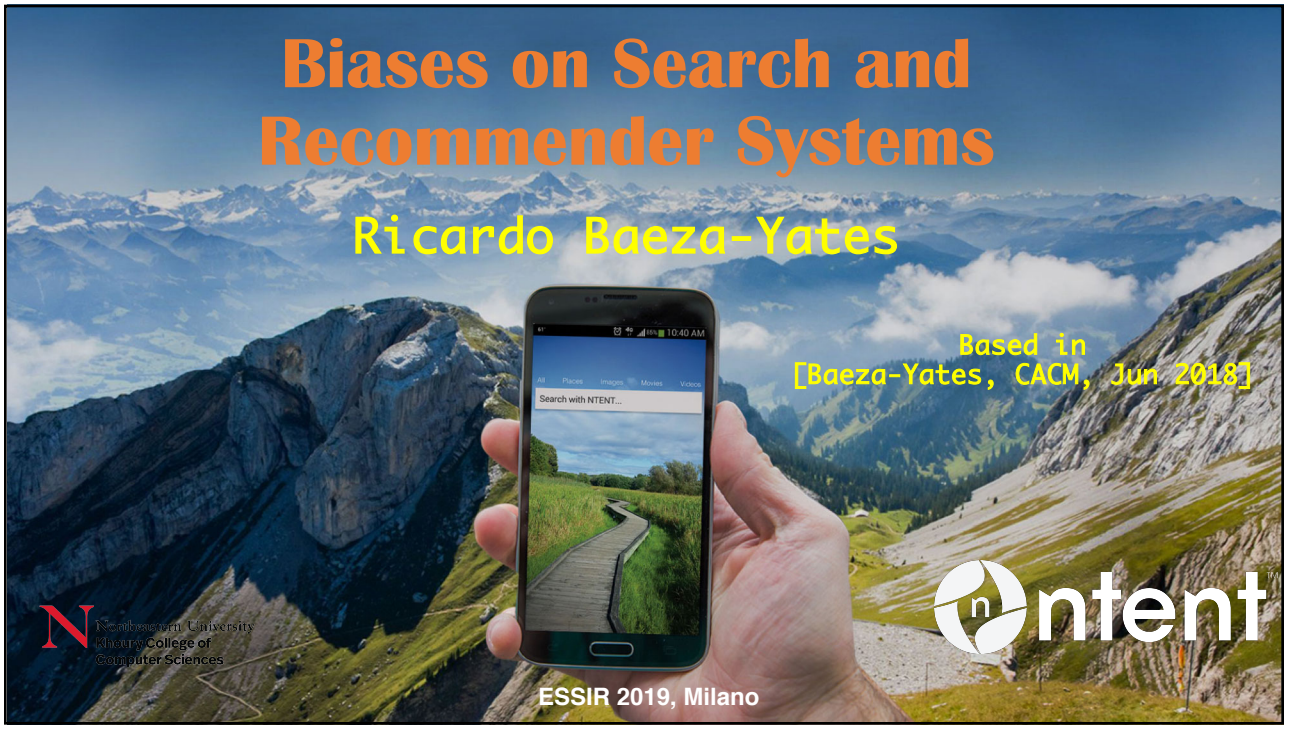


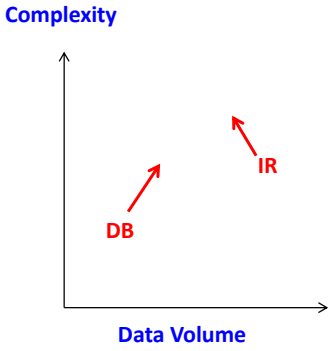
Basic Understanding of Data

Data & Query	Unstructured	Structured
Explicit	Information Retrieval	(Relational) Databases
Implicit	Recommender Systems	
Unknown	Data Mining	

ntent

Northeastern University
Khoury College of Computer and Information Sciences

A Bit of History



Two different points of view for data

A Relational Model of Data for Large Shared Data Banks

E. F. CODD
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on *n*-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity
CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. Introduction

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of *data inconsistency* which are expected to become troublesome even in nondeductive systems.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for noninferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

1.2. Data Dependencies in Present Systems

The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed *without logically impairing some application programs* is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. *Ordering Dependence.* Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

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;

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)

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 filter bubbles for users
  - but also **echo chambers for (recommender) systems**
- 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 (even inversely correlated) optimization function
  - For example, user experience vs. monetization

Motivation 3: Fake Content & Bias

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

Buzzfeed News

ntent

Northeastern University
Khoury College of Computer and Information Sciences

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 systems

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!

ntent

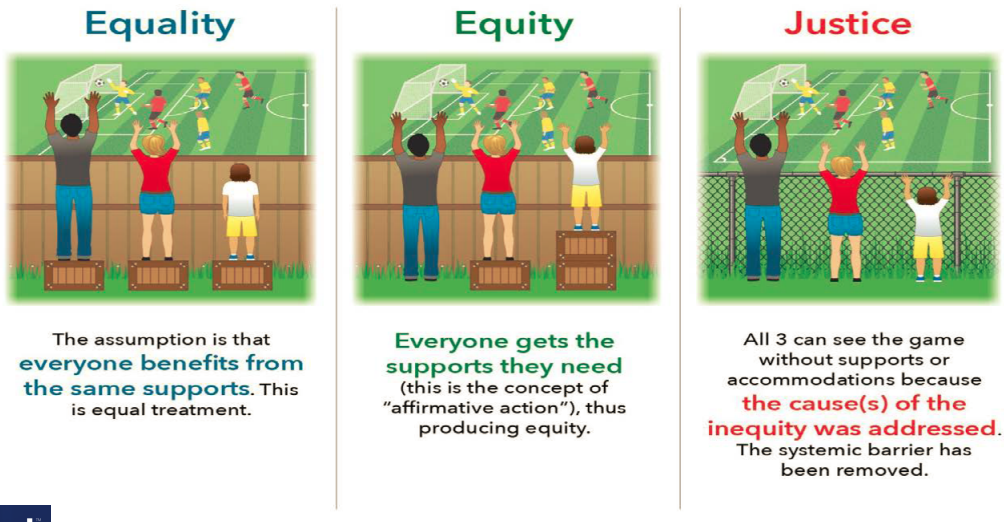
Northeastern University
Khoury College of Computer and Information Sciences

A Non-Technical Question

Neutral?
Fair?

ntent

Northeastern University
Khoury College of Computer and Information Sciences



ntent

A Non-Technical Question

Neutral?
Fair?

Not Always!

Debias the input
Tune the algorithm
Debias the output

Bias awareness!

ntent

Northeastern University
Khoury College of Computer and Information Sciences

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

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

ntent

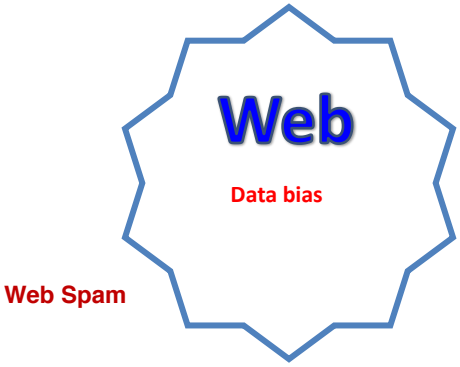
Northeastern University
Khoury College of Computer and Information Sciences

Bias in Computing Systems

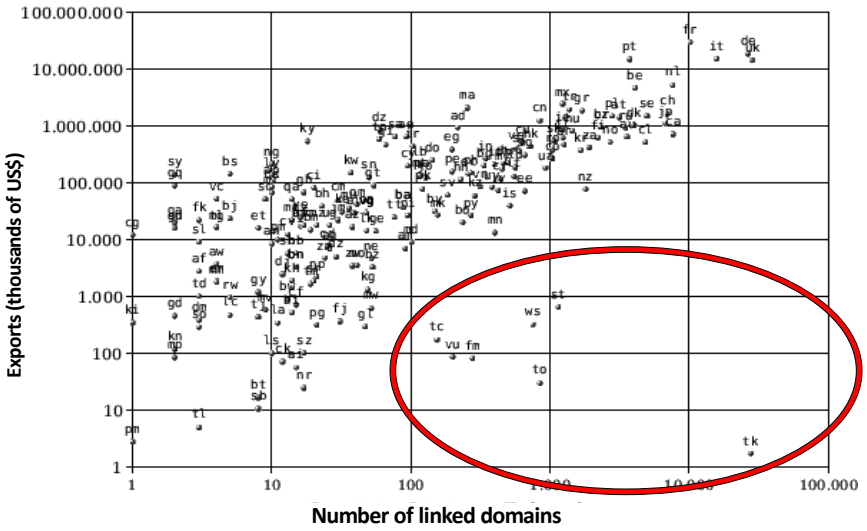
- The **quality of any algorithm** is bounded by the **quality of the data that uses** (and hence of its users)
- Data bias awareness

  [Gordon & Desjardins; Provost & Buchanan, MLJ 1995]
- Bias in computing 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

Biases on Search & RS: Web as a Case Study

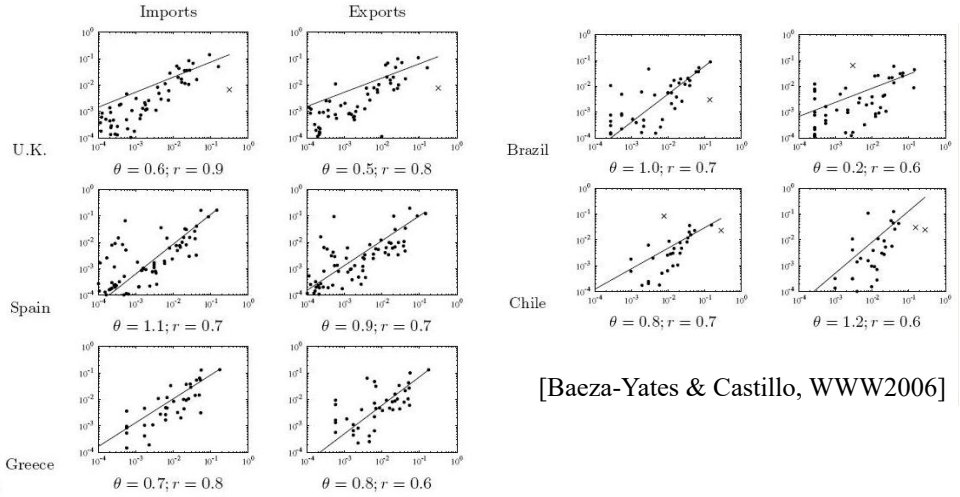


Economic Bias in Links



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

Economic Bias in Links

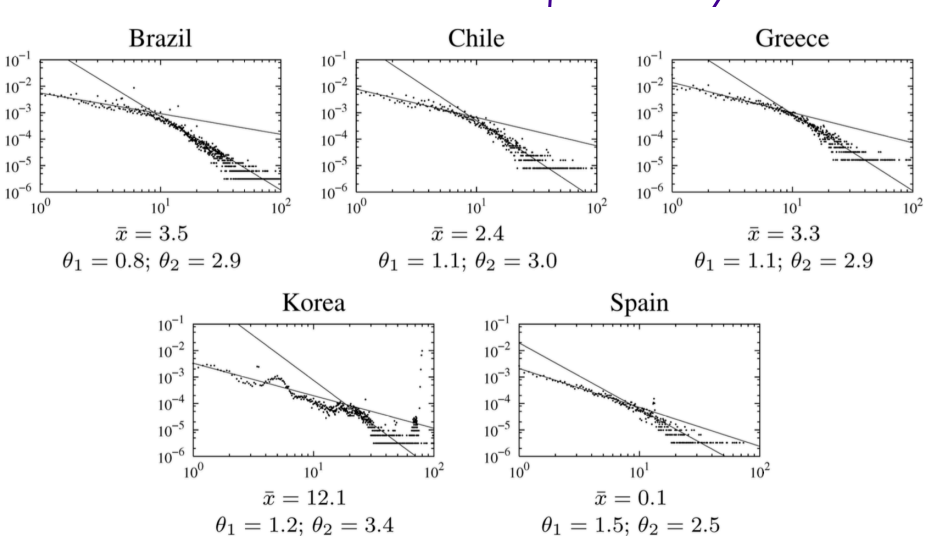


[Baeza-Yates & Castillo, WWW2006]

Cultural Bias in Website Structure

Shame

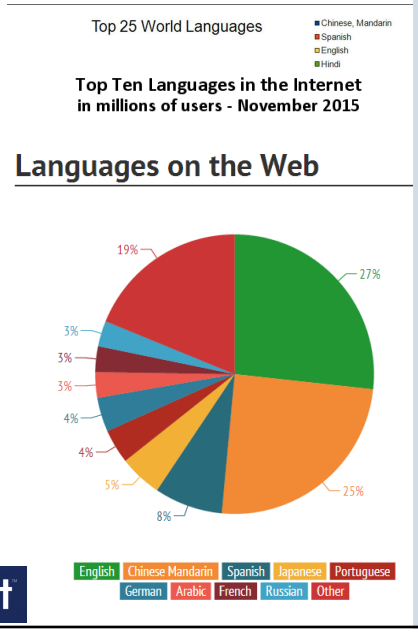
Minimal effort

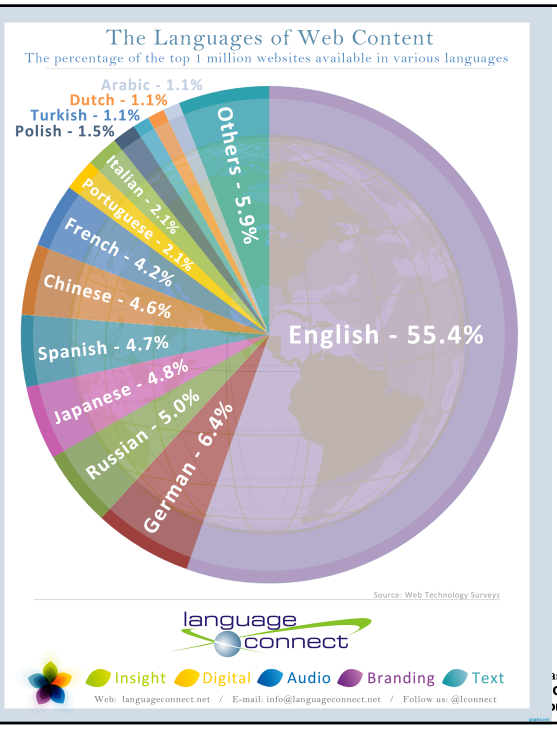


[Baeza-Yates, Castillo, Efthimiadis, TOIT 2007]

Northeastern University
Khoury College of Computer and Information Sciences 18

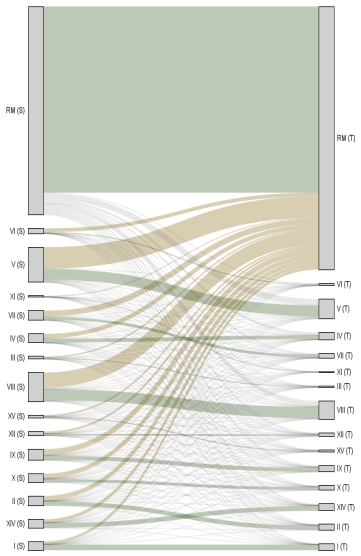
Linguistic Bias in Content

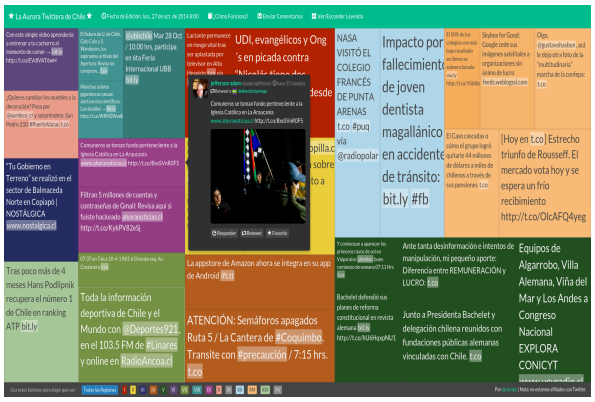




Northeastern University
College of Computer
rmation Sciences

Geographical Bias in Content





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

Northeastern University
Khoury College of Computer and Information Sciences

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

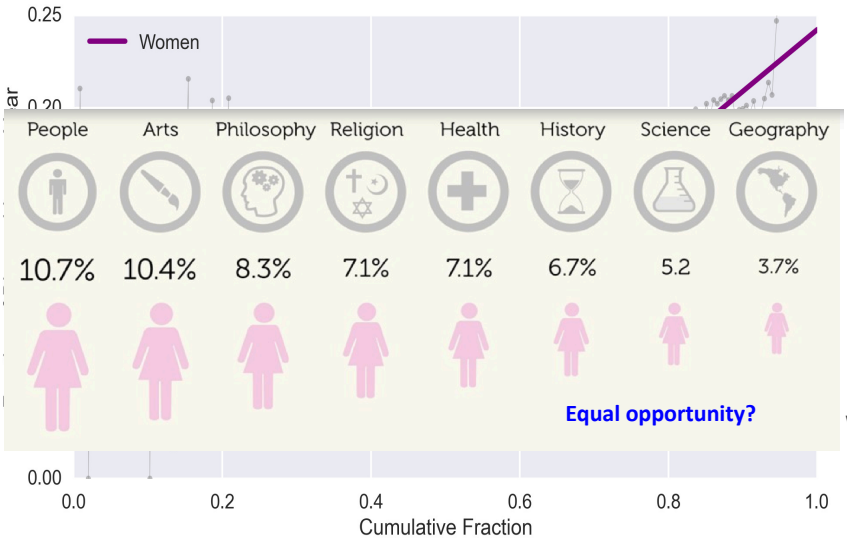
Most journalists are men?

[Bolukbasi at al, NIPS 2016]

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

Northeastern University
Khoury College of Computer and Information Sciences

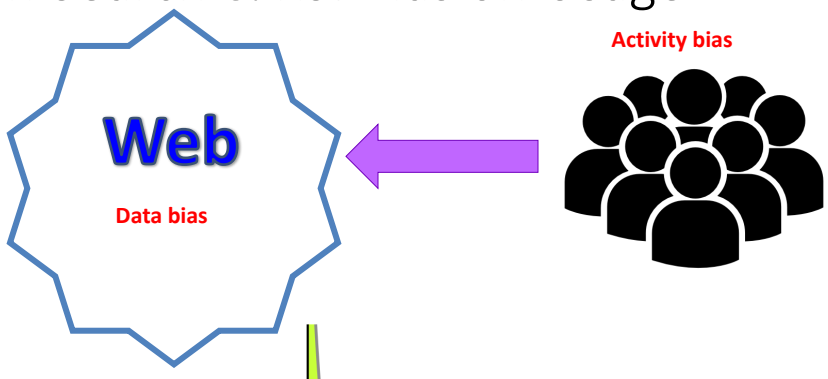
Gender Bias in Content



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

Northeastern University Khoury College of Computer and Information Sciences

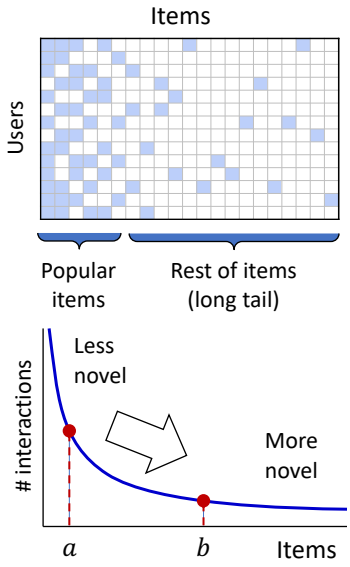
Bias on Search & RS: Bias on Usage



We are all in the long tail!
[Goel et al., WSDM 2010]

Northeastern University Khoury College of Computer and Information Sciences

Popularity Bias in Recommender Systems



- Take care to recommended items that are not too popular
- Metrics

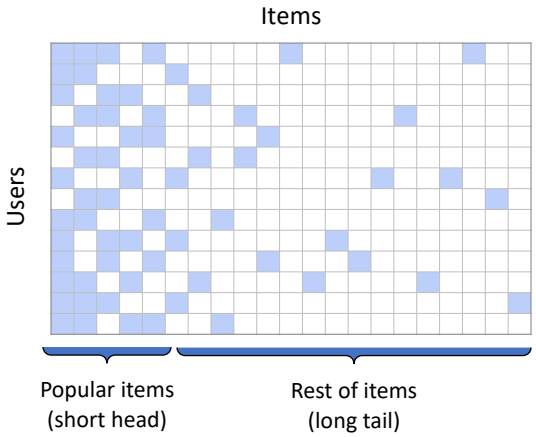
$$nov(i) = 1 - \frac{\#\text{ ratings of } i}{\#\text{ users}}$$

- Novelty enhancement
- Problem solved! …really?

[Vargas & Castells, RecSys 2011]

Northeastern University
Khoury College of Computer and Information Sciences

A self-fulfilling prophecy?

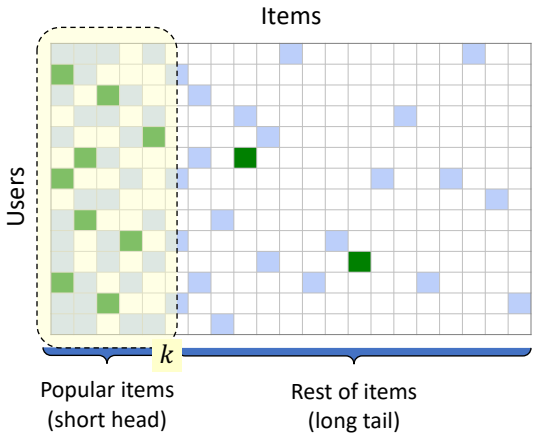


Observed user-item interaction
Unobserved preference

Ratings are missing not at random (MNAR)

Northeastern University
Khoury College of Computer and Information Sciences

A self-fulfilling prophecy?



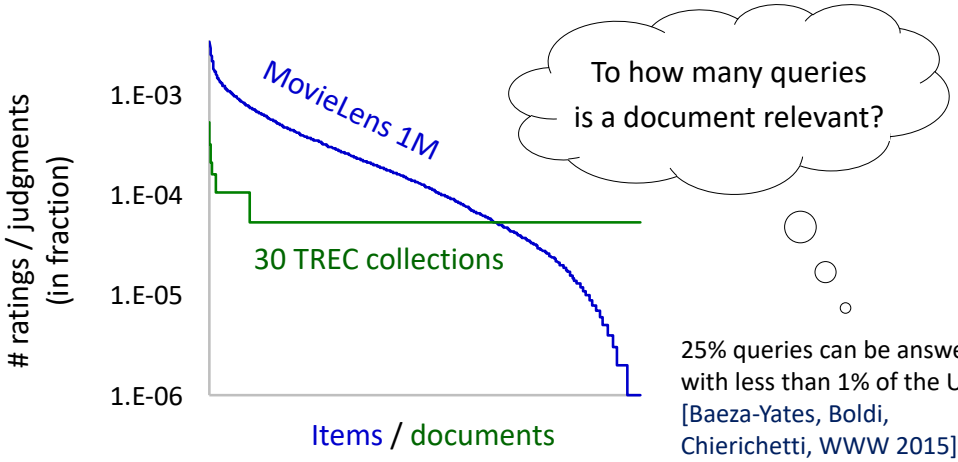
Judgments are missing not at random (MNAR)

$$\text{avg P@}k \sim \frac{\blacksquare}{\blacksquare + \square}$$

Worse yet: user-system reinforcement loop (more later)

[Marlin et al., RecSys 2010]
[Steck, RecSys 2010, 2011]
[Fleder & Hossanagar, Management Sciences 2009]

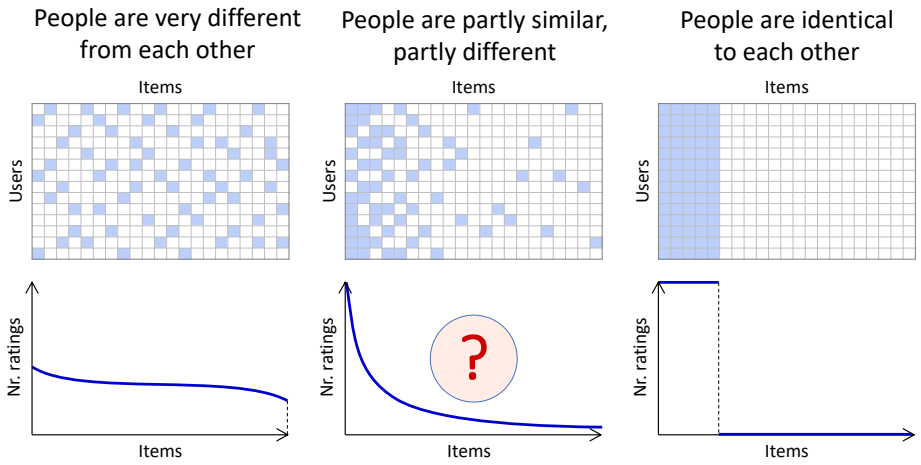
A problem for IR evaluation methodology!



25% queries can be answered with less than 1% of the URLs!
[Baeza-Yates, Boldi, Chierichetti, WWW 2015]

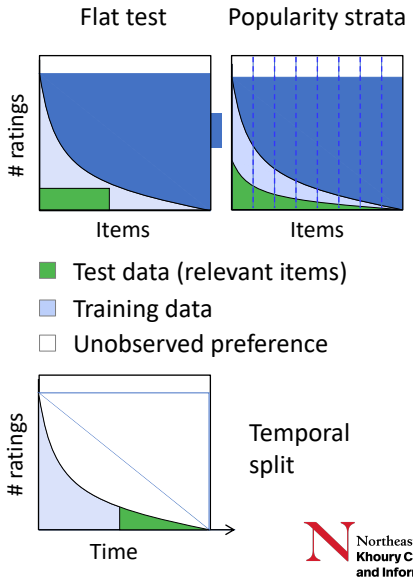
[Bellogín, Castells & Cantador IRJ 2017]

How different or similar are we to each other?



Get rid of the popularity bias!

- In the rating split
  [Bellogín, Castells & Cantador, IRJ 2017]
- In the metrics
  - Stratified recall
    [Steck, RecSys 2011]
  - Importance propensity scoring
    [Yang et al., RecSys 2018]
- In the algorithms
  [Steck, RecSys 2011]
  [Lobato et al., ICML 2014]
  [Jannach et al., UMUAI 2015]
  [Cañamares & Castells, SIGIR 2018, **best paper award**]



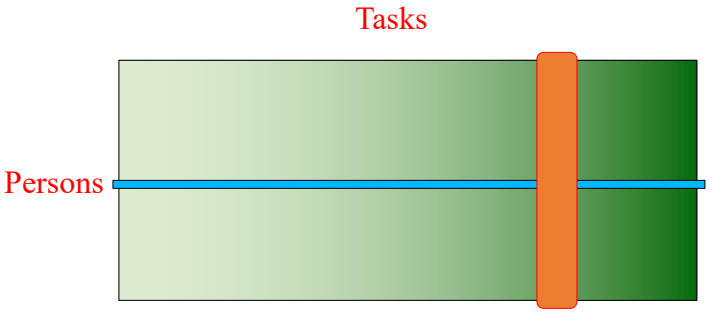
Recommending within the Long Tail

- Exploit the context (and deep learning!)

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

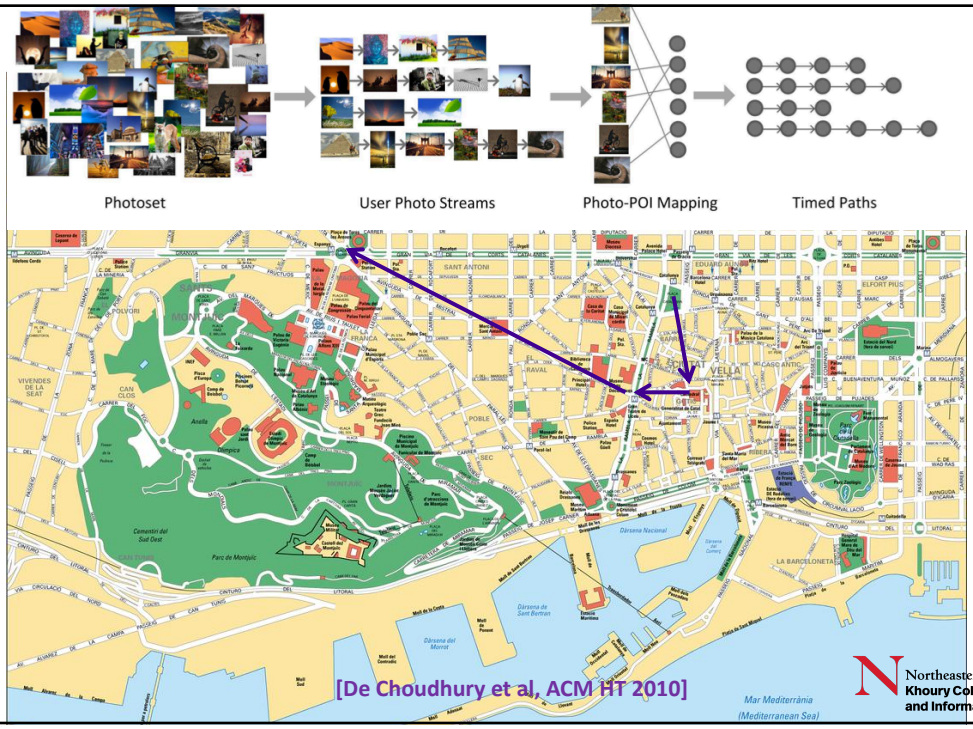
- Personalization vs. Contextualization

  Break the filter bubble! (more later)
  [Goel at al, WSDM 2010]



ntent

Northeastern University
Khoury College of Computer and Information Sciences



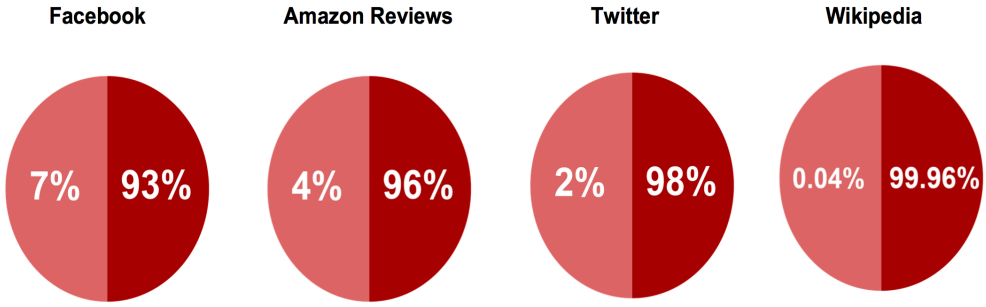
[De Choudhury et al, ACM HT 2010]

ntent

Northeastern University
Khoury College of Computer and Information Sciences

Activity Bias also Affects Content

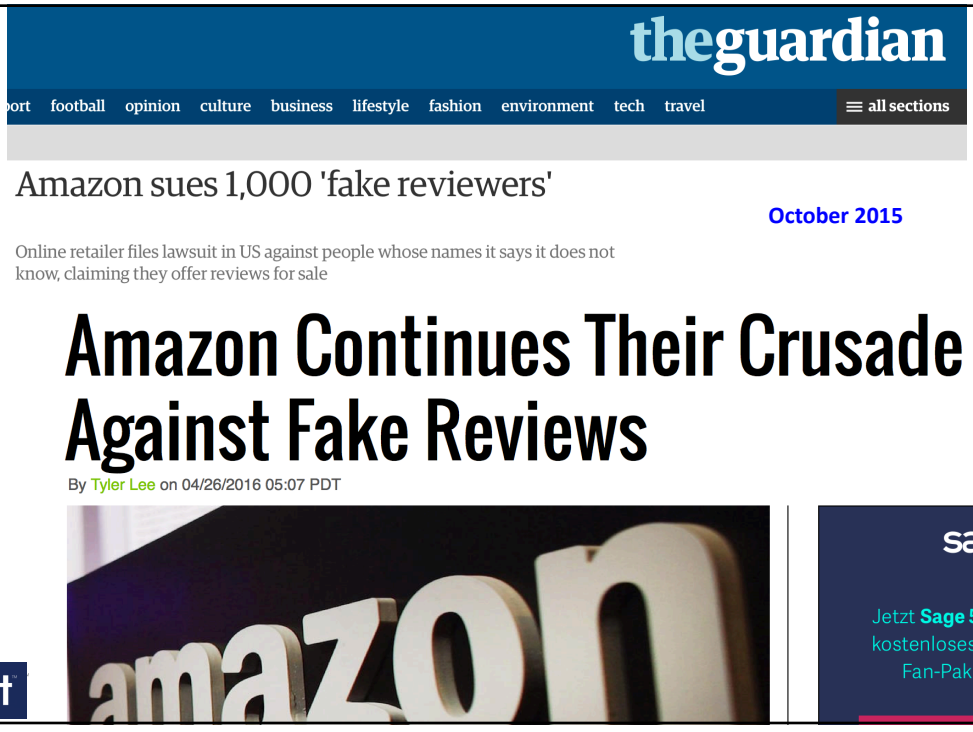
Most users are passive (*i.e.,* more than 90%) – wisdom of crowds is a partial illusion
Hence, which percentage of **active** users produce 50% of the content?



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

ntent

Northeastern University
Khoury College of Computer and Information Sciences

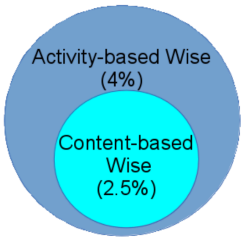


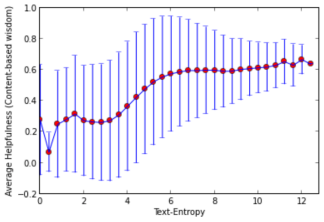
ntent

Northeastern University
College of Computer
ation Sciences

Quality of Content?

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



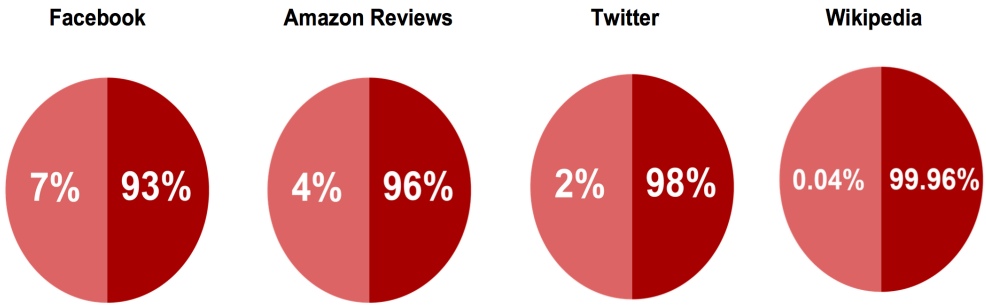


[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]

Northeastern University
Khoury College of Computer and Information Sciences

Activity Bias also Affects Content

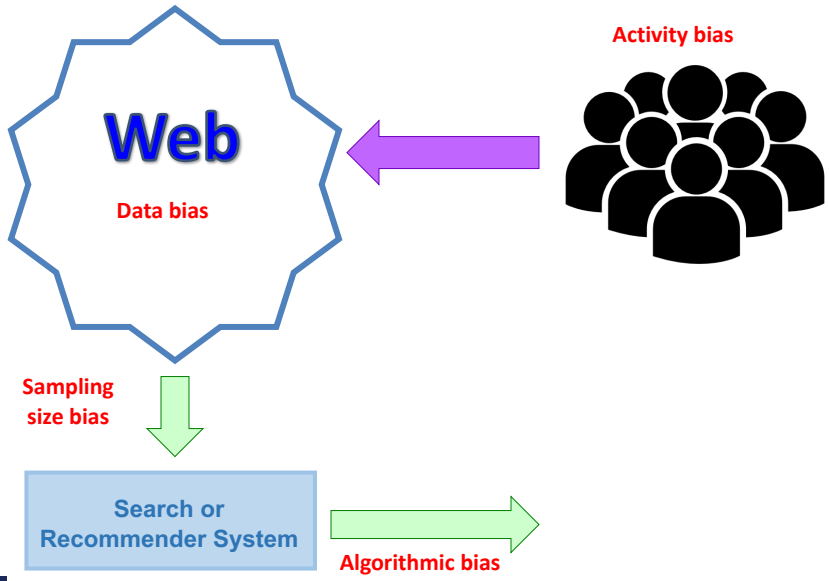
Which percentage of **active** users produce 50% of the content?



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

Northeastern University
Khoury College of Computer and Information Sciences

Bias in the Web



Northeastern University
Khoury College of Computer
and Information Sciences

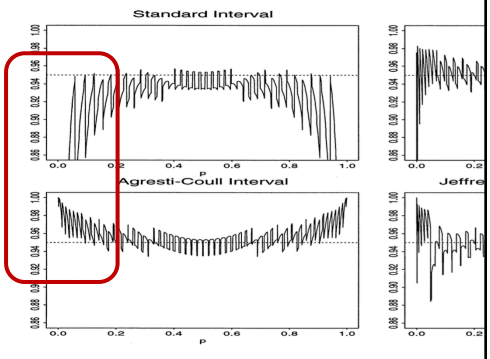
Sample Size?

- If we want to estimate the frequencv 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 80% 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]

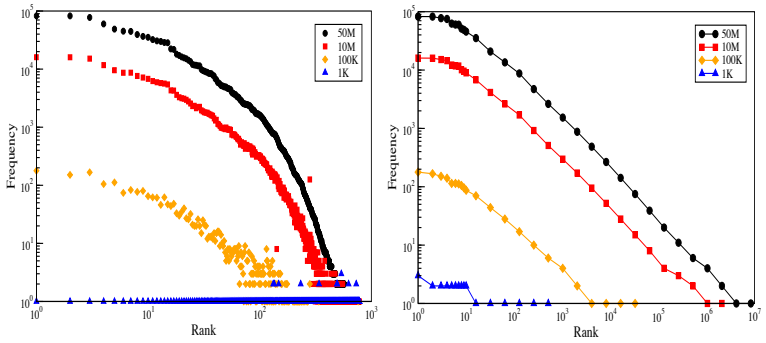
Northeastern University
Khoury College of Computer
and Information Sciences

Sampling Techniques

- Standard technique:

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

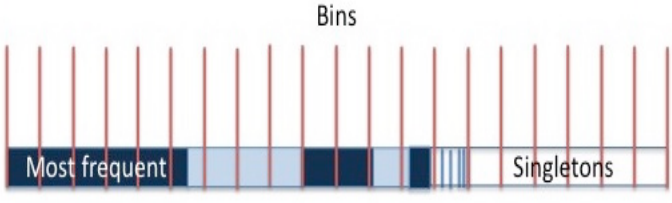
- A good sample should cover well all the items distribution but this does not work with very skewed distributions.



[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]

Fixing the Tail

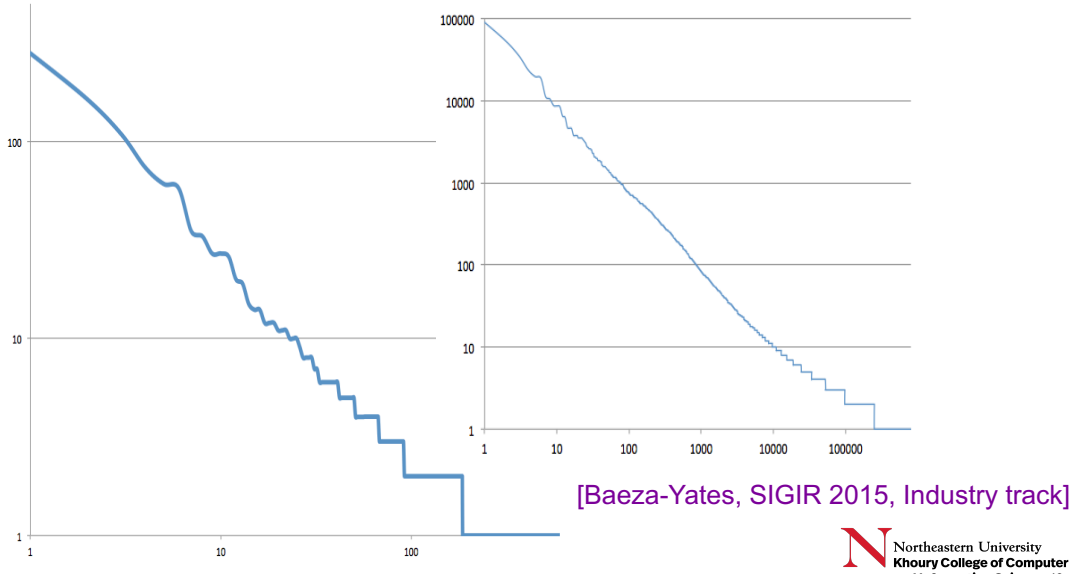
- To mimic the tail we change the binning size when we reach a query frequency of *b*/2
- If we want a singleton ratio of $\beta = S/V$ we recalculate the binning size as

$$b' = (1 - \beta)(Q - Q')/(\beta V')$$

- where Q' and V' are the partial vocabulary size and volume before changing the bin size.

[Baeza-Yates, SIGIR 2015, Industry track]

Stratified Sampling Example



[Baeza-Yates, SIGIR 2015, Industry track]

What About the Index/Ranking? Accessibility Bias

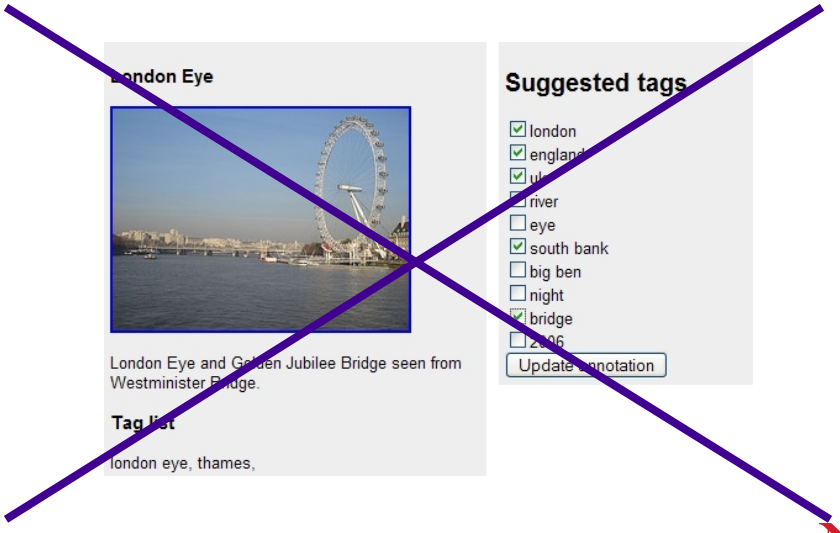
Accessibility bias studies if a document is or is not accessible by the user through the IR system

A measure of accessibility is retrievability, which measures the likelihood of a document to be retrieved by a search engine. Formally:

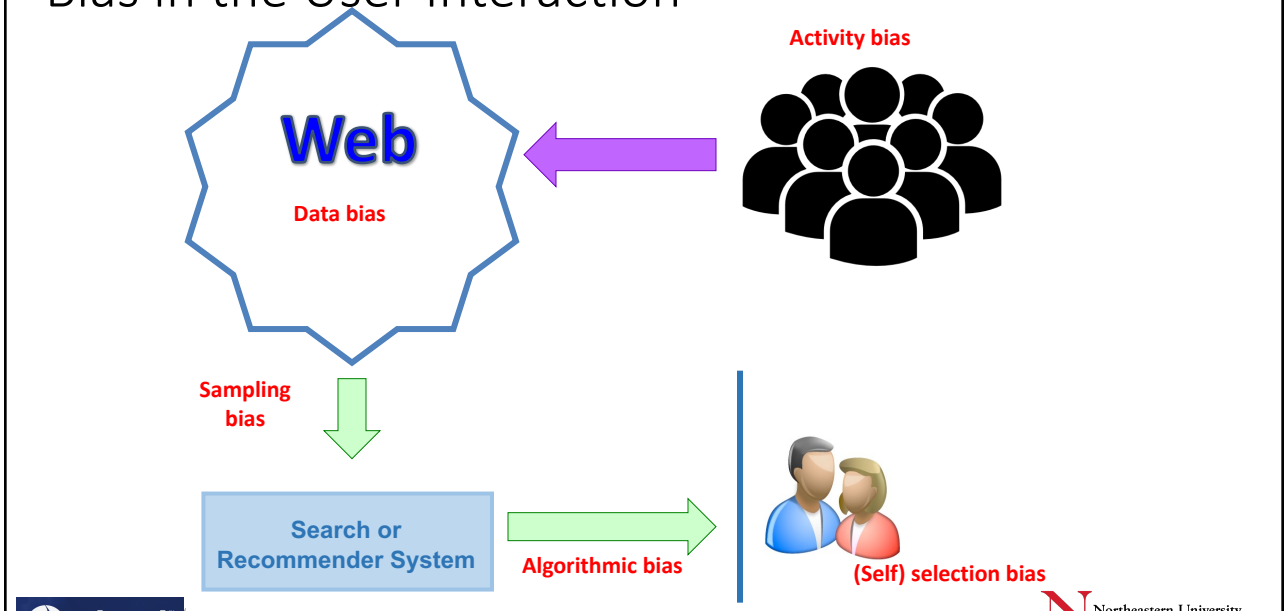
- The retrievability of each document, when combined with a coefficient of distribution unbalance (like the Gini coefficient) tells us the bias of a search engine on retrieving a particular set of documents rather than another
- How to compute retrievability?
  1. Take all the possible queries the users would think of
  2. Issue each query to the search engine
  3. Count how many times each document gets retrieved
- It is possible to compute the retrievability for Boolean Retrieval Models analytically and this is an upper bound for many retrieval models.

[Lipani et al., ICTIR 2015]

Extreme Algorithmic Bias



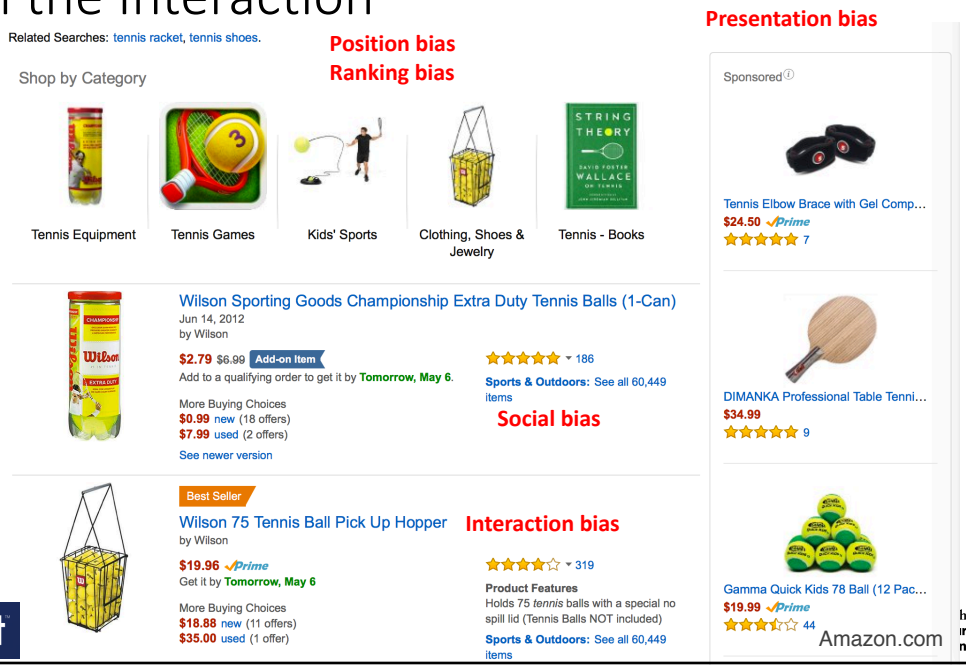
Bias in the User Interaction



ntent

Northeastern University
Khoury College of Computer and Information Sciences

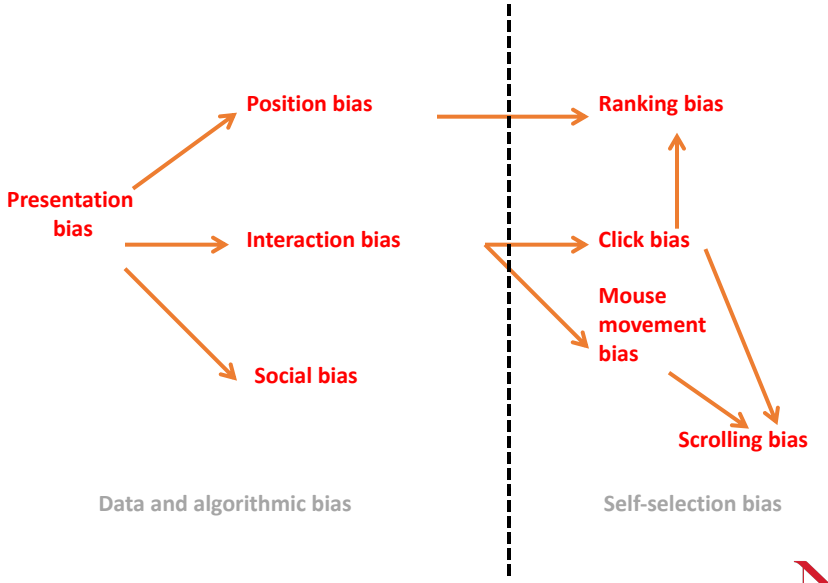
Bias in the Interaction



ntent

heastern University
ry College of Computer
nformation Sciences

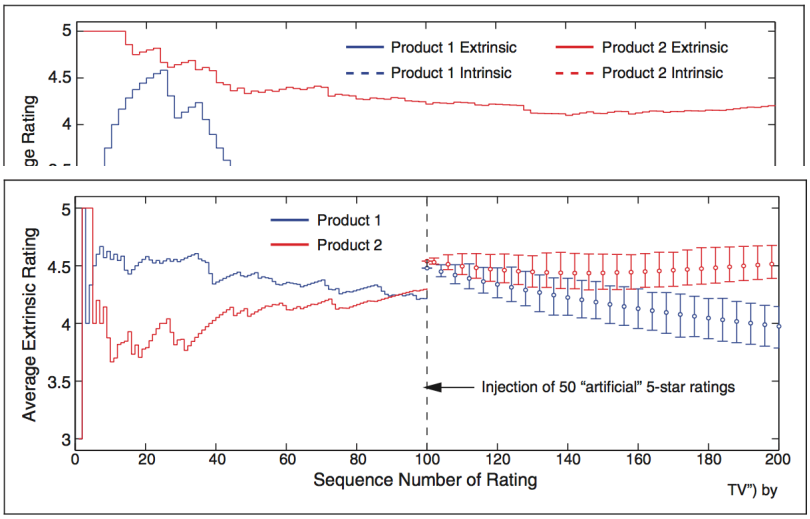
Dependencies: A Cascade of Biases!



ntent

Northeastern University
Khoury College of Computer and Information Sciences

Social Bias

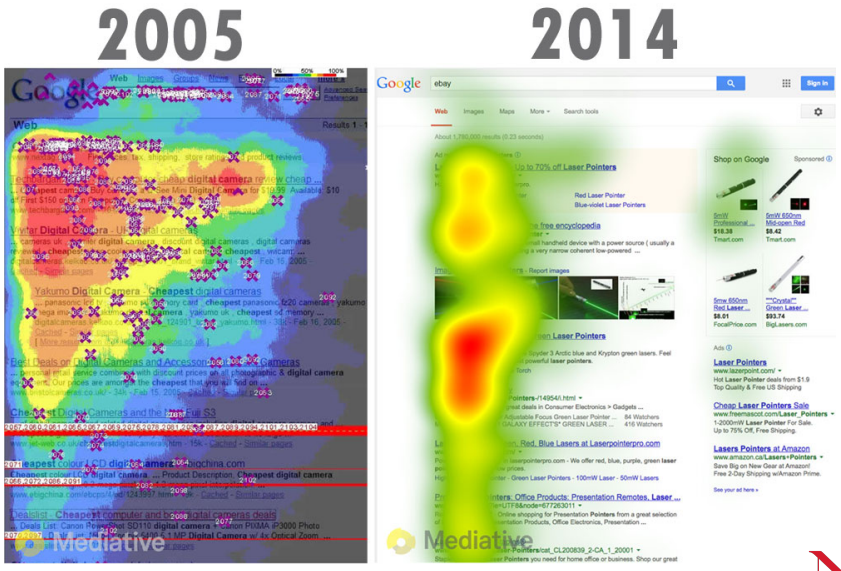


[Why Amazon's Ratings Might Mislead You; The Story of Herding Effects, Ting Wang and Dashun Wang, Big Data, 2014]

ntent

Northeastern University
Khoury College of Computer and Information Sciences

Ranking Bias in Web Search



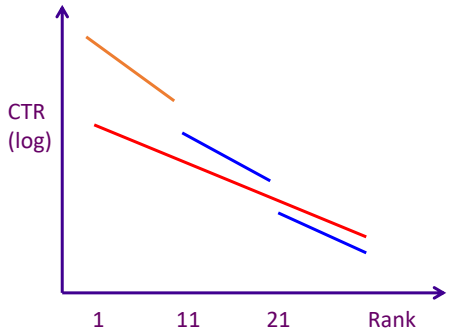
[Mediative Study, 2014]

Ranking Bias: Click Bias in Web Search

- Ranking & next page bias

Debiasing Search Clicks and Other Biases

Clicks as implicit positive user feedback



[Dupret & Piwowarski, SIGIR 2008]
[Chapelle & Zhang, WWW 2009]
[Dupret & Liao, WSDM 2010, **best paper**]

Debias the input

Learning to Rank with bias
[Joachims et al., WSDM 2017, **best paper**]
+ many other papers

Tune the algorithm

Fair rankings
[Zehlike et al., CIKM 2017]

Debias the output

Northeastern University
Khoury College of Computer and Information Sciences

What is a Fair Ranking?

Given:

- Ranking of length k
- Two groups of items → protected, non-protected
- Minimum proportion *p*

A fair ranking implies that the number of protected items does not fall below *p at any point* of the ranking

Example: Expert search

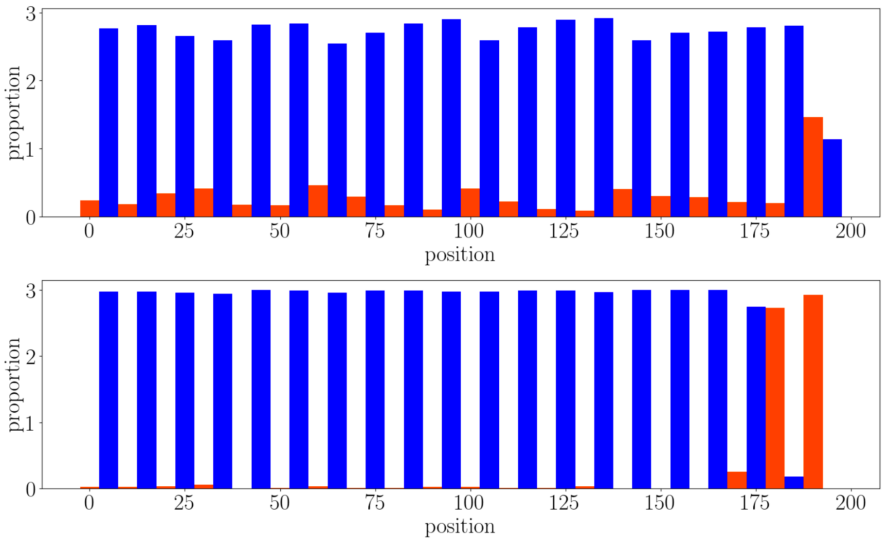
Northeastern University
Khoury College of Computer and Information Sciences

Example: W3C Expert Search

- TREC Enterprise Dataset
- Retrieve a sorted list of experts for a topic given a corpus of Emails written by possible candidates
- Consider women as protected group (13% in the whole dataset, red)
- Training data is sorted such that
  - first all male experts
  - second all female experts
  - third all male non-experts and
  - fourth all female non-experts

Northeastern University
Khoury College of Computer and Information Sciences

Baselines



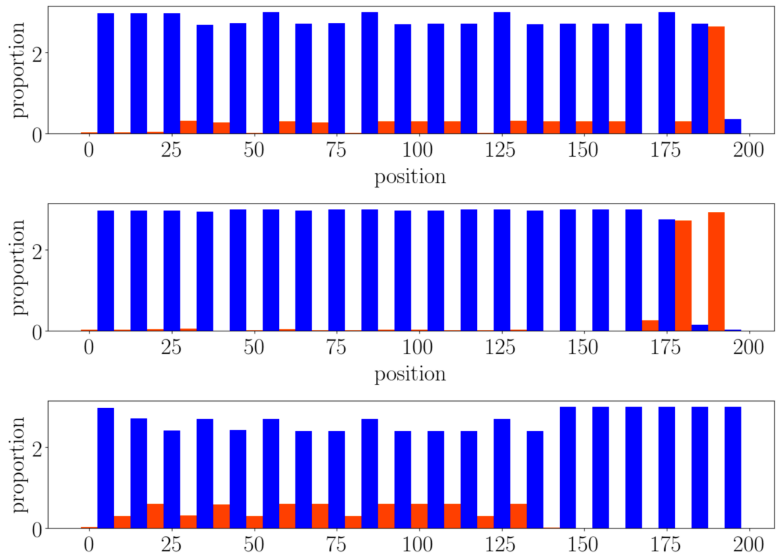
Color blind

Standard Learning to Rank

Northeastern University
Khoury College of Computer and Information Sciences

Post-Processing FA*IR

[Zehlike et al., CIKM 2017]

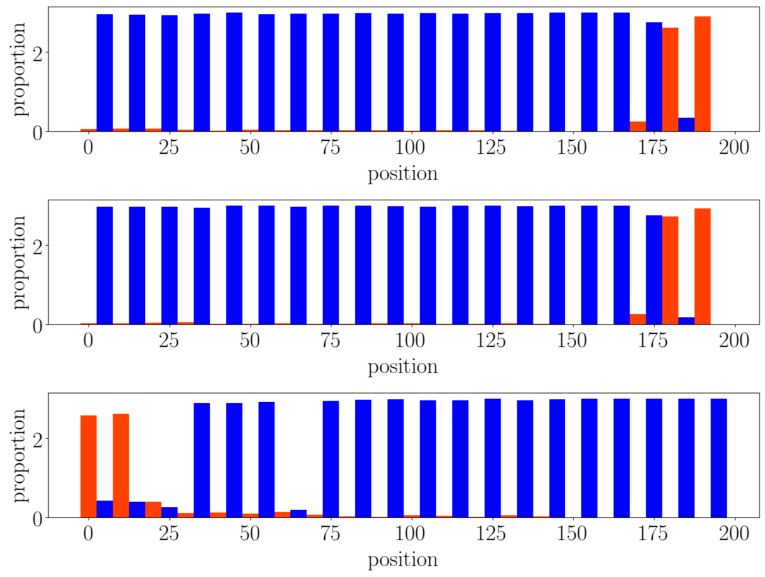


ntent

Northeastern University
Khoury College of Computer and Information Sciences

Pre-Processing FA*IR

[Zehlike & Castillo, ArXiv 2018]
[Zehlike at al., ArXiv 2019]

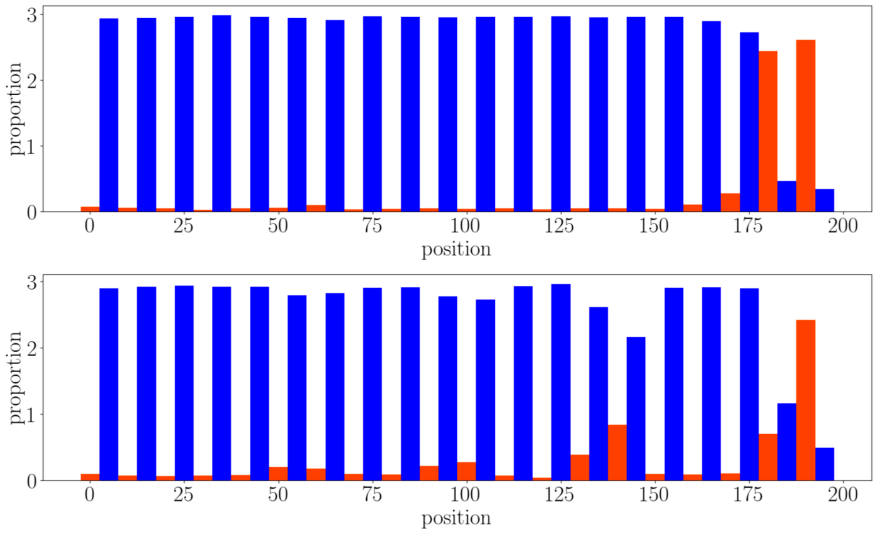


ntent

Northeastern University
Khoury College of Computer and Information Sciences

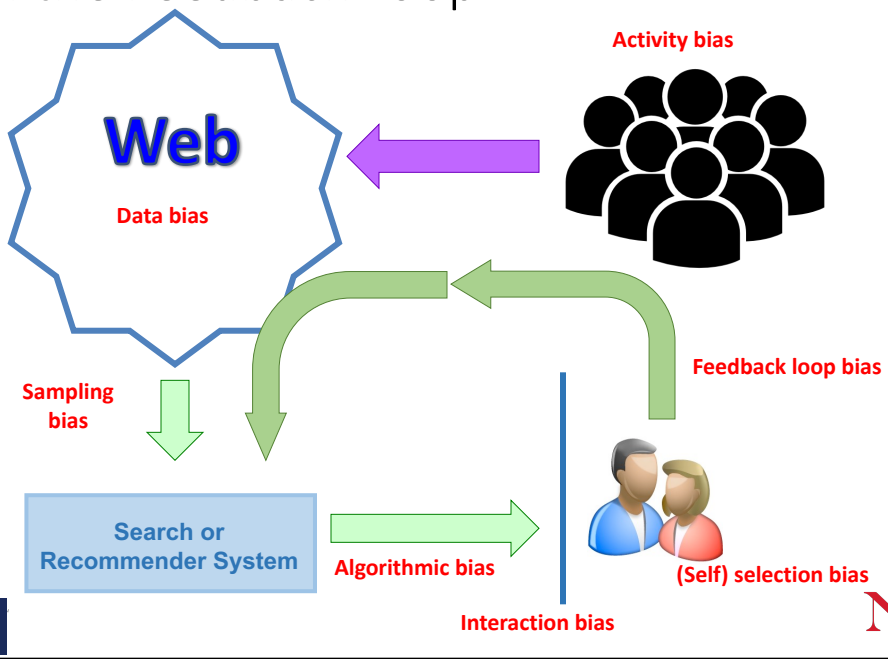
DELTR

[Zehlike & Castillo, ArXiv 2018]
[Zehlike at al., ArXiv 2019]



Northeastern University
Khoury College of Computer and Information Sciences

Bias in the Feedback Loop



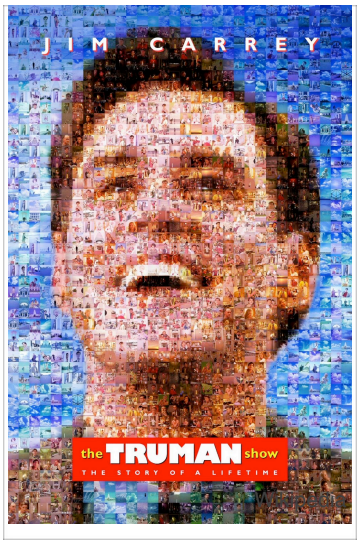
Northeastern University
Khoury College of Computer and Information Sciences

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



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

Cold start problem solution: Explore & Exploit

How much exploration is needed to counteract presentation bias?

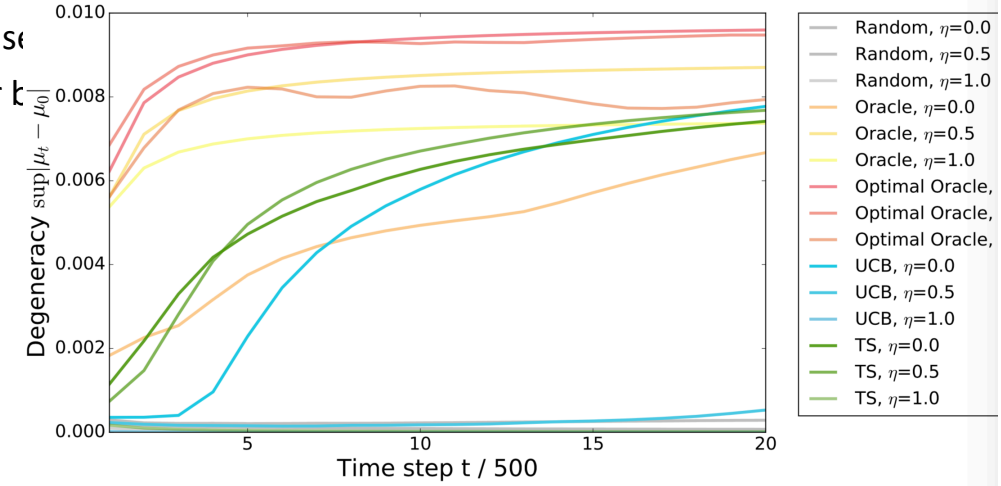
Eco Chambers in Recommender Systems

- For users
  - Filter bubbles
  - Degenerate feedback loops (e.g., YouTube autoplay)
- For systems
  - Short-term greedy optimization
  - The system is partly writing its own future
  - Partial knowledge of the world if not enough exploration/traffic
  - Views from new users should balance the exploration for new items
  - Otherwise the system itself is in a bubble!

Northeastern University
Khoury College of Computer and Information Sciences

Eco Chambers in Recommendation Systems

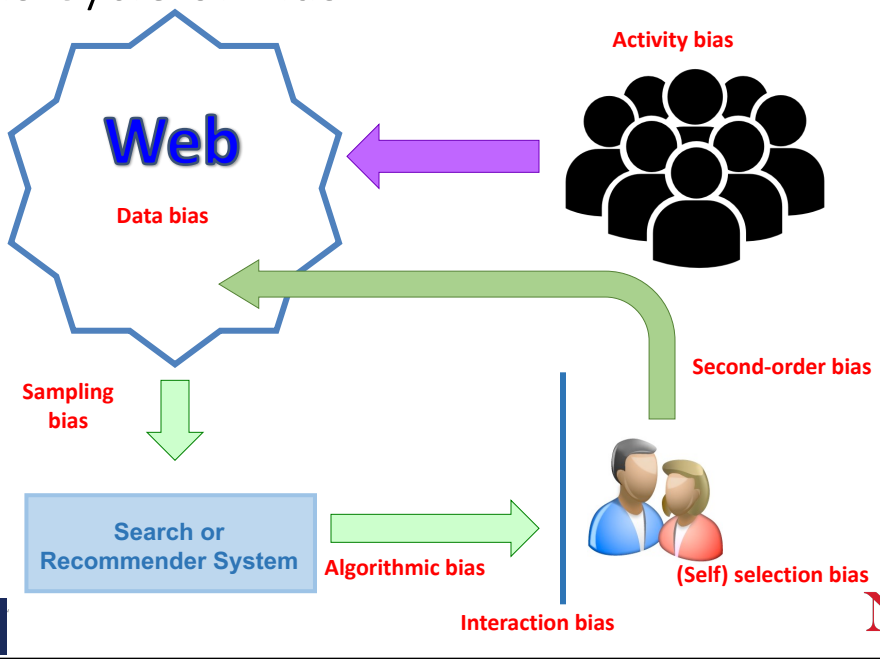
- For use
- Filter b



[Jiang et al. Degenerate Feedback Loops in Recommendation Systems, AAAI 2019]

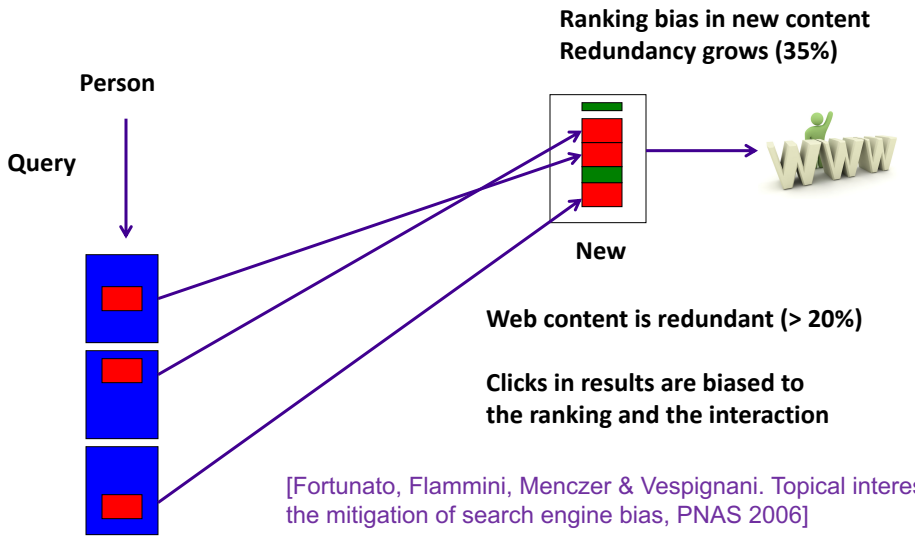
Northeastern University
Khoury College of Computer and Information Sciences

Vicious Cycle of Bias



Northeastern University
Khoury College of Computer and Information Sciences

Second Order Bias in Web Content



Web content is redundant (> 20%)

Clicks in results are biased to the ranking and the interaction

[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]

ntent

Northeastern University
Khoury College of Computer and Information Sciences

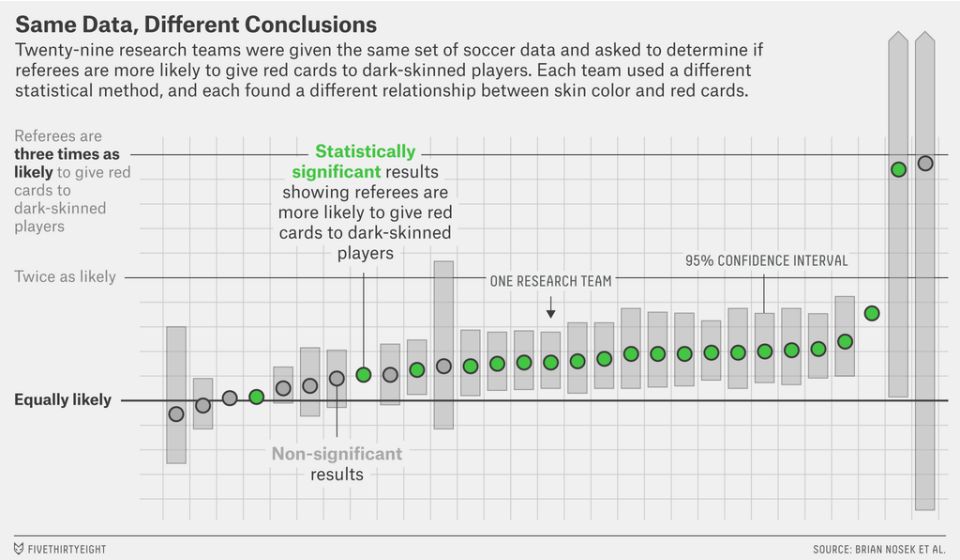
Recap

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

ntent

Northeastern University
Khoury College of Computer and Information Sciences

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



- 61 analysts, 29 teams: 20 yes and 9 no
- [Silberzahn et al., COS, Univ. of Virginia, 2015]

Northeastern University Khoury College of Computer and Information Sciences

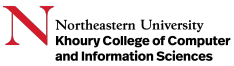
Our Professional Biases

- Design and Implementation
  - Do systems reflect the characteristics of the designers?
  - Do systems reflect the characteristics of the coders?
- Evaluation
  - Choose the right experiment
  - Choose the right test data
    - Pool bias in search test collections [Lipani et al., SIGIR 2015, CIKM 2016]
  - Choose the right metric(s)
  - Choose the right baseline(s)
  - Julio Gonzalo's talk: http://tiny.cc/ESSIR2019-juliogonzalo

Northeastern University Khoury College of Computer and Information Sciences

What we can do?

- Data
  - Analyze for known and unknown biases, debias when possible/needed
  - Recollect more data for sparse regions of the problem
  - Delete attributes associated directly/indirectly with harmful bias
- Interaction
  - Make sure that the user is aware of the biases all the time
  - Give more control to the user
- Design and Implementation
  - Let experts/users/colleagues contest every step of the process
- Evaluation
  - Do not fool yourself!



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

Final Take-Home Message

- Systems are 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 biases**!
- 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!)

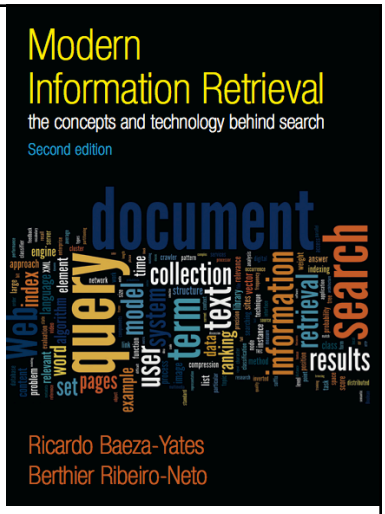
Big Data of **People** is huge…..
….. but it is tiny compared to the future
Big Data of the **Internet of Things (IoT)**

No activity bias!

Northeastern University
Khoury College of Computer and Information Sciences

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
www.baeza.cl
@polarbearby

Biased Questions?

Northeastern University
Khoury College of Computer and Information Sciences