Auditing Search Engines for Differential Satisfaction across Demographics

Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, Emine Yilmaz

University College London
Microsoft Research New York

@erishabh

Microsoft Research

WWW 2017

UCL

Fairness across demographics

- Online services - advertised as being available to any user
- Ethical
  - Equal access to everyone
- Practical
  - Equal access helps attract a large and diverse population of users
  - Service providers are scrutinized for seemingly unfair behavior [1,2,3]
- Onus on us
  - develop **fair systems**

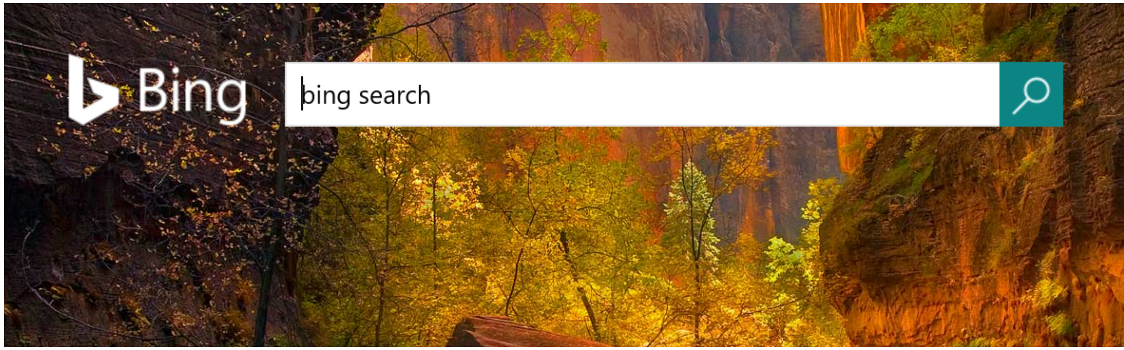
[1] N. Diakopoulos. Algorithmic accountability. Digital Journalism, 3(3):398–415, 2015
[2] S. Barocas and A. D. Selbst. Big data's disparate impact. California Law Review, 104, 2016.
[3] C. Munoz, M. Smith, and D. Patel. Big data: A report on algorithmic systems, opportunity, and civil rights. Technical report, Executive Office of the President of the United States, May 2016.

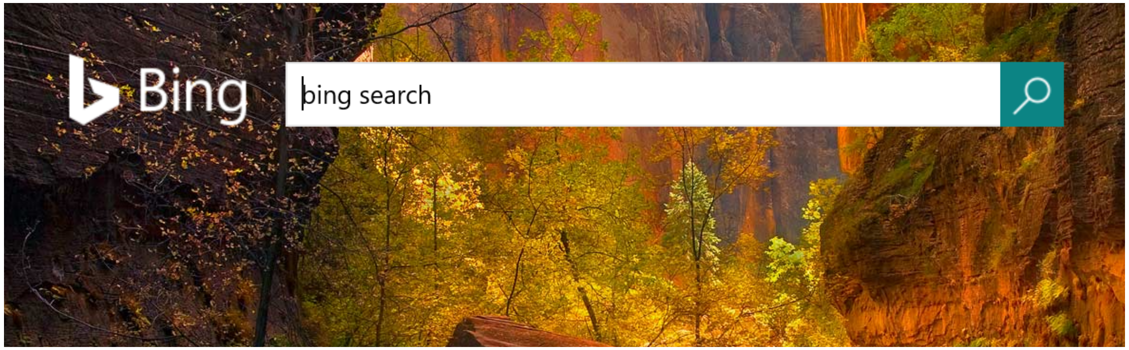
Auditing services for fairness

We offer methods for **auditing a system's** performance for detection of **differences in user satisfaction** across demographics

From public libraries to search engines

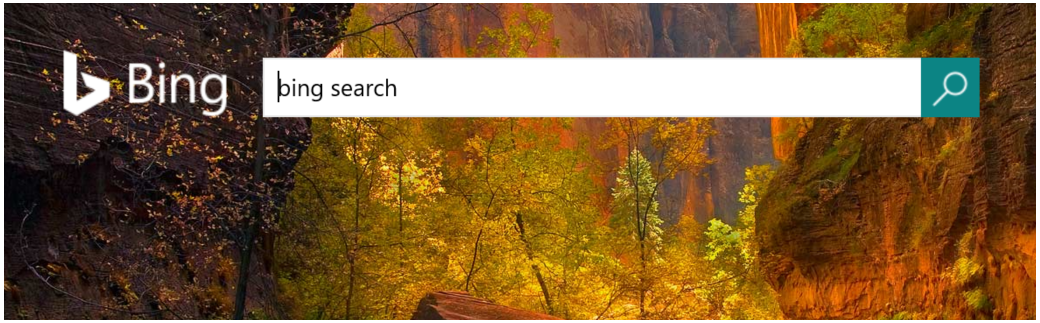
- Modern analogue of public libraries
- Dominant role in information access
- Fairness in **performance**!





Are Search Engines Fair?

From public libraries to search engines



Search Engines:

- Rely on ML models to optimize for **user satisfaction**
- Make use of implicit signals
- **Metric** driven development

… not easy to audit

Tricky: straightforward optimization can lead to differential performance

Goal: estimate difference in user satisfaction between two demographic groups.

- Search engine uses a standard metric: **time spent** on clicked result page as an indicator of satisfaction.

- Suppose older users issue more of *"retirement planning"* queries

1. Aggregate Metrics can be misleading

- Overall metrics can hide differential satisfaction

- **Average user satisfaction for *"retirement planning"* *may be high.***

But,

- Average satisfaction for younger users=**0.7**
- Average satisfaction for older users=**0.2**

2. Query-level metrics can hide differential satisfaction

Younger users

<query-X>
<query-X>
<query-X>
<query-X>
<query-X>
<query-X>
retirement planning
<query-X>
<query-X>

⋮

Older users

retirement planning
retirement planning
<query-X>
retirement planning

Assuming same user satisfaction for *"retirement planning"* for both older and younger users = 0.7

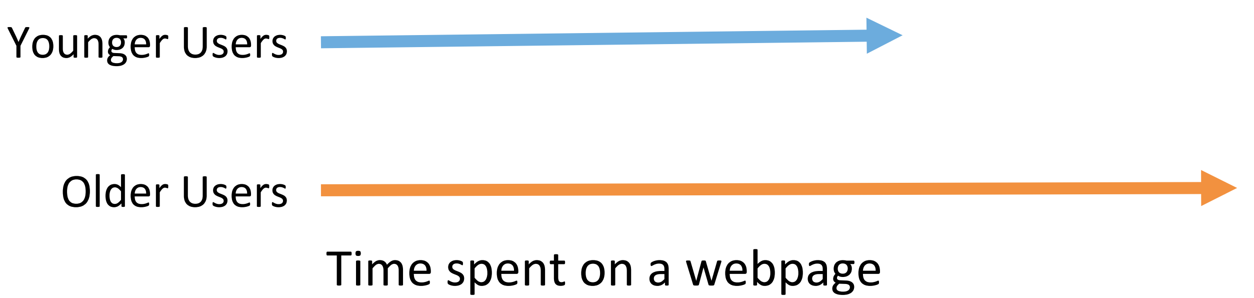
What if average satisfaction for <query-X> = **0.9?**
(e.g. <query-X> = "*facebook*")

Older users still receive more of lower-quality results than younger users.

3. More critically, even individual-level metrics can also hide differential satisfaction

Metric itself could be confounded with demographics

Consider: Reading time for the same webpage result for the same user satisfaction



We must control for natural demographic variation to meaningfully audit for differential satisfaction.

Outline

1 Background

2 Data & metrics

3 Proposed approaches:

  1 Context Matching

  2 Hierarchical Multi-level model

4 From metrics to satisfaction

5 Discussion

Data: Demographic characteristics of search engine users

- Internal logs from Bing.com for two weeks
- 4 M users | 32 M impressions | 17 M sessions
- Demographics: Age & Gender
- Age:
    - post-Millenial: <18
    - Millenial: 18-34
    - Generation X: 35-54
    - Baby Boomer: 55-74

… also perform external auditing using comScore data

Metrics Considered

1. Graded Utility (GU)
   - based on search outcome and user effort

2. **Reformulation Rate (RR)**
   - fraction of queries that were reformulated

3. Successful Click Count (SCC)
   - clicks with significant dwell times

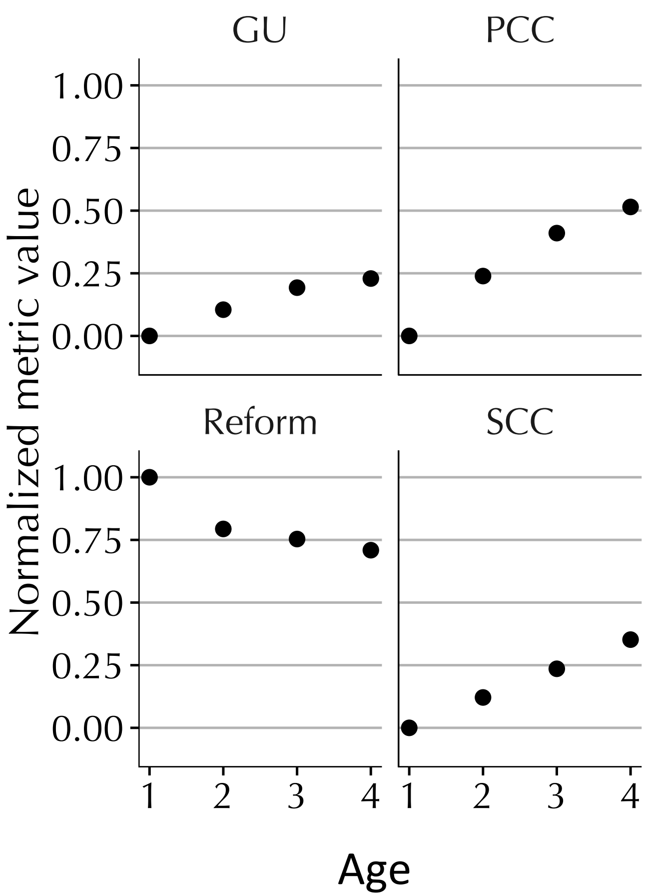
4. Page Click Counts (PCC)
   - total no of clicks on SERP

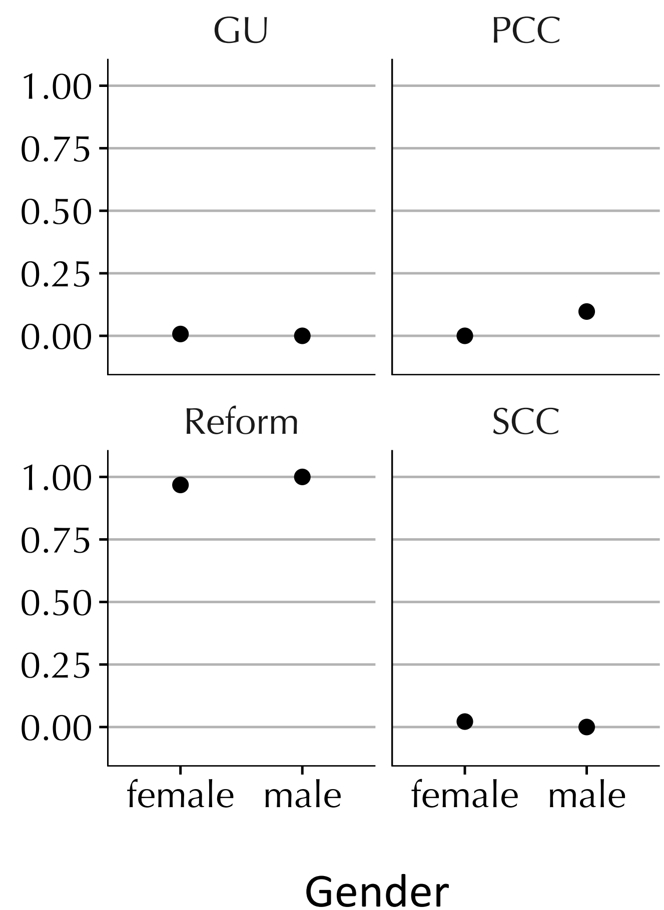
J. Jiang, A. Hassan, Z. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In WSDM, 2015.
Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In CIKM, 2013.
G. Buscher, L. van Elst, and A. Dengel. Segment-level time as implicit feedback: A comparison to eye tracking. In SIGIR, 2009.

Goal: estimate difference in user satisfaction between demographic groups

Obvious solution: **demographic binning**!

Overall metrics across Demographics





- Substantial differences in performance across age
- Gender – not so much

… how true are these?

Pitfalls with Overall Metrics

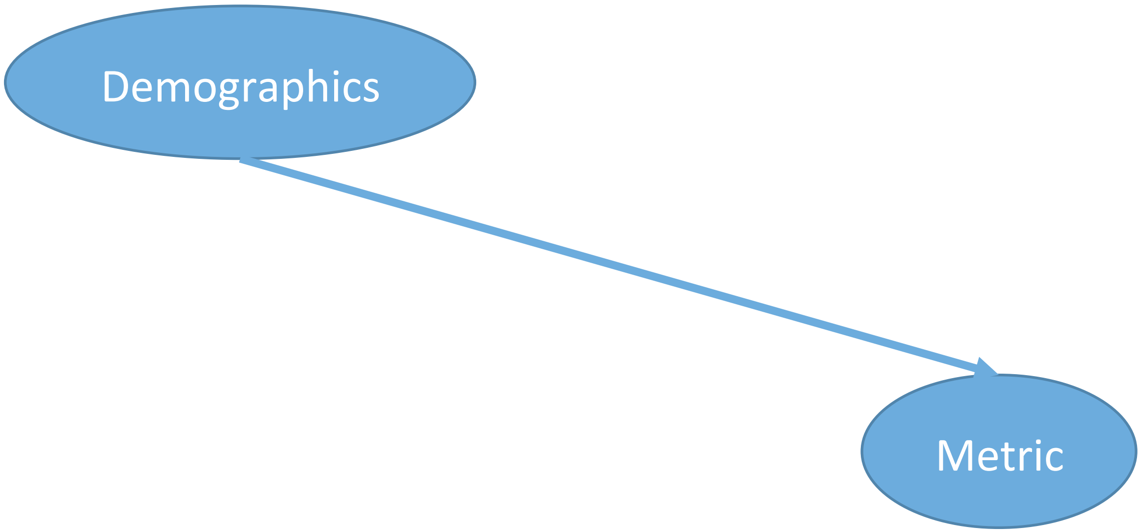
Conflates two separate effects:

- natural **demographic variation** caused by the differing traits among the different demographic groups e.g.
  - Different queries issued
  - Different information need for the same query
  - Even for the same satisfaction, demographic A tends to click more than demographic B

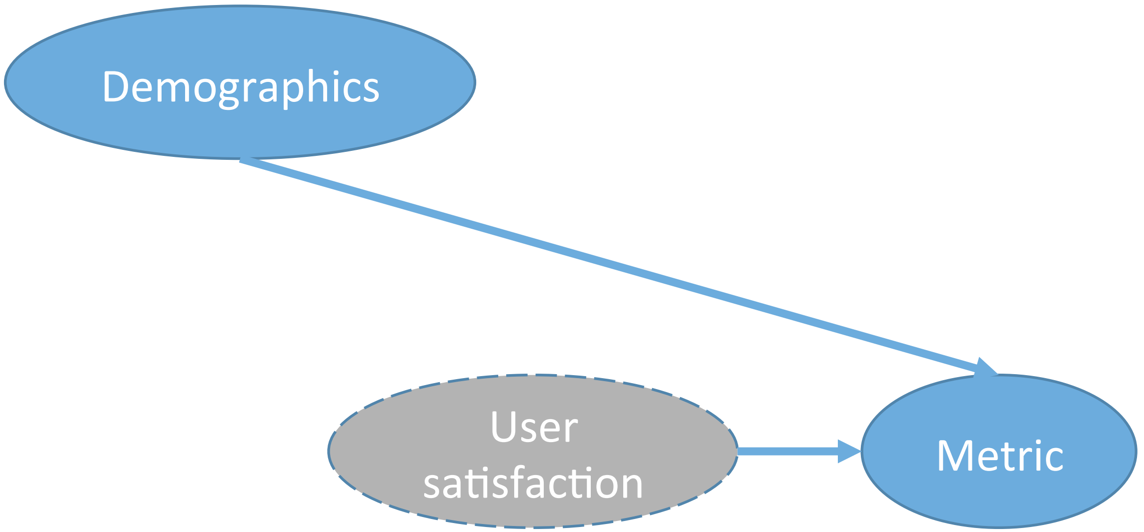
- **Systemic difference** in user satisfaction due to the search engine

… we need to disentangle them!

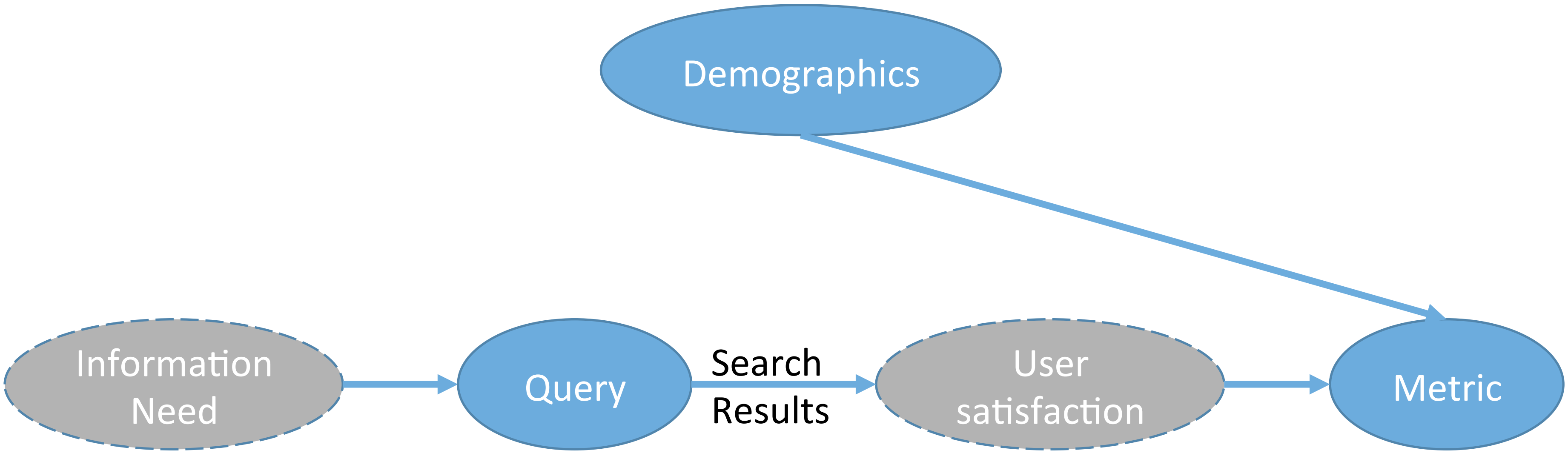
Utilize work from causal inference



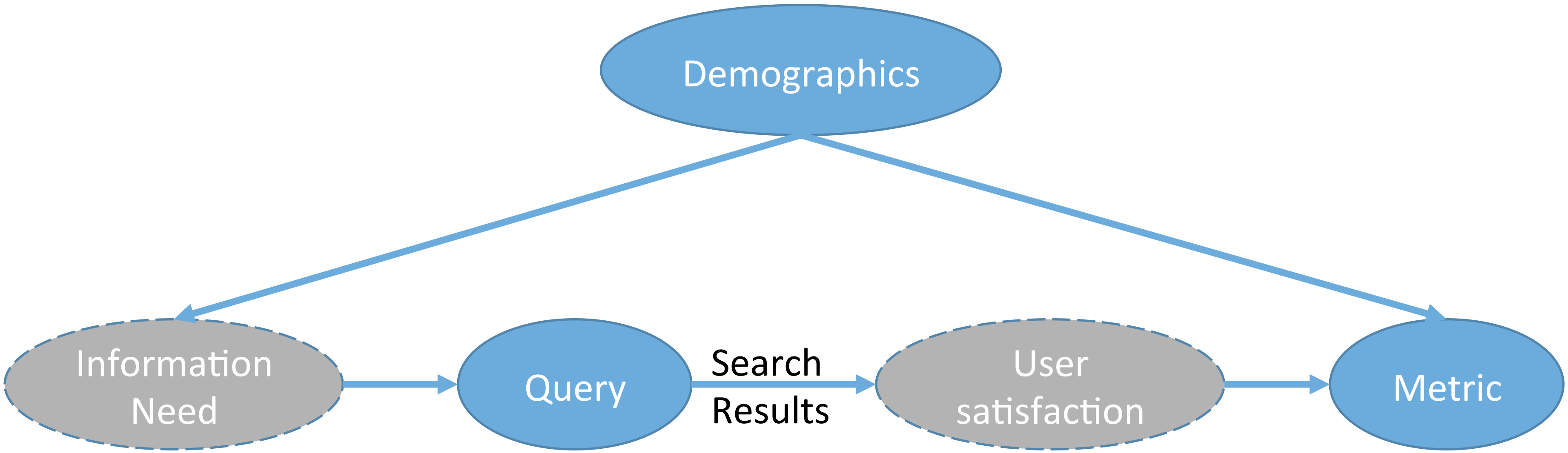
Utilize work from causal inference



Utilize work from causal inference



Utilize work from causal inference



Outline

1 Motivation

2 Problems with naïve auditing

3 Data & Metrics

4 Proposed approaches:

  1 Context Matching

  2 Hierarchical Multi-level model

5 From metrics to satisfaction

6 Discussion

Proposed Approaches

Extremely restrictive
More robust

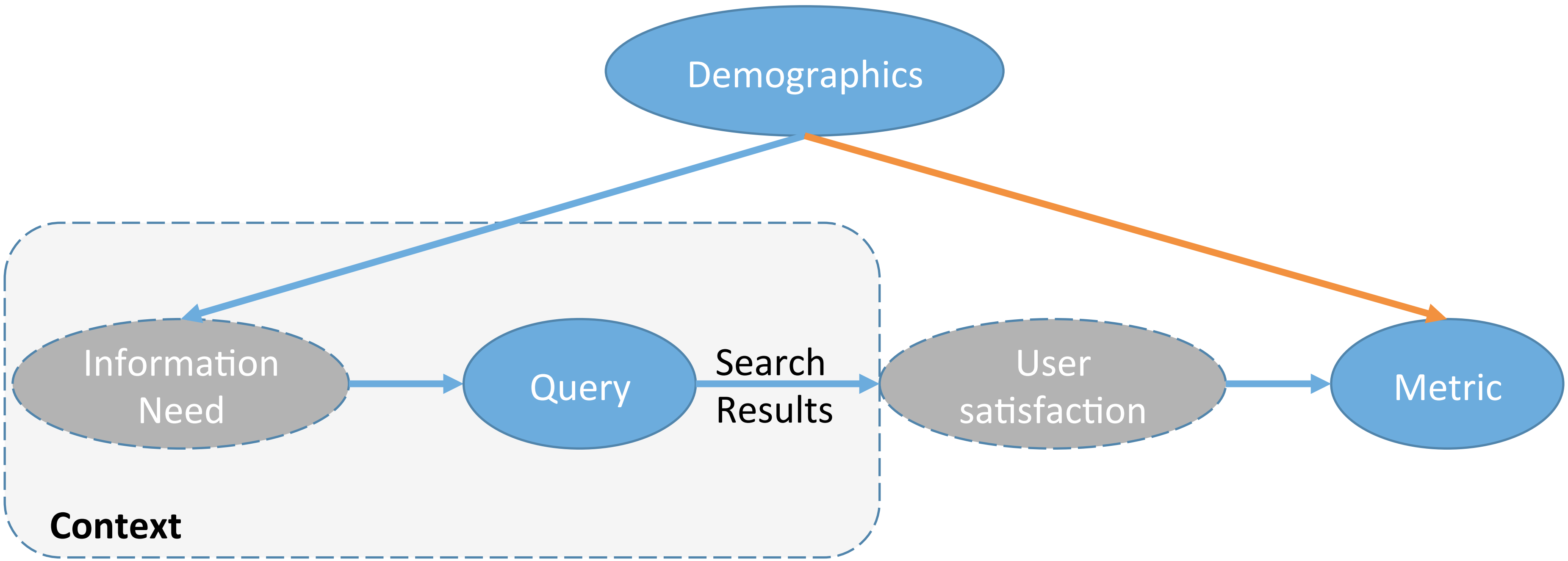
Generalizable
Less Robust

1) Context Matching

2) Multi-level model

I. Context Matching:

selecting for activity with near-identical context

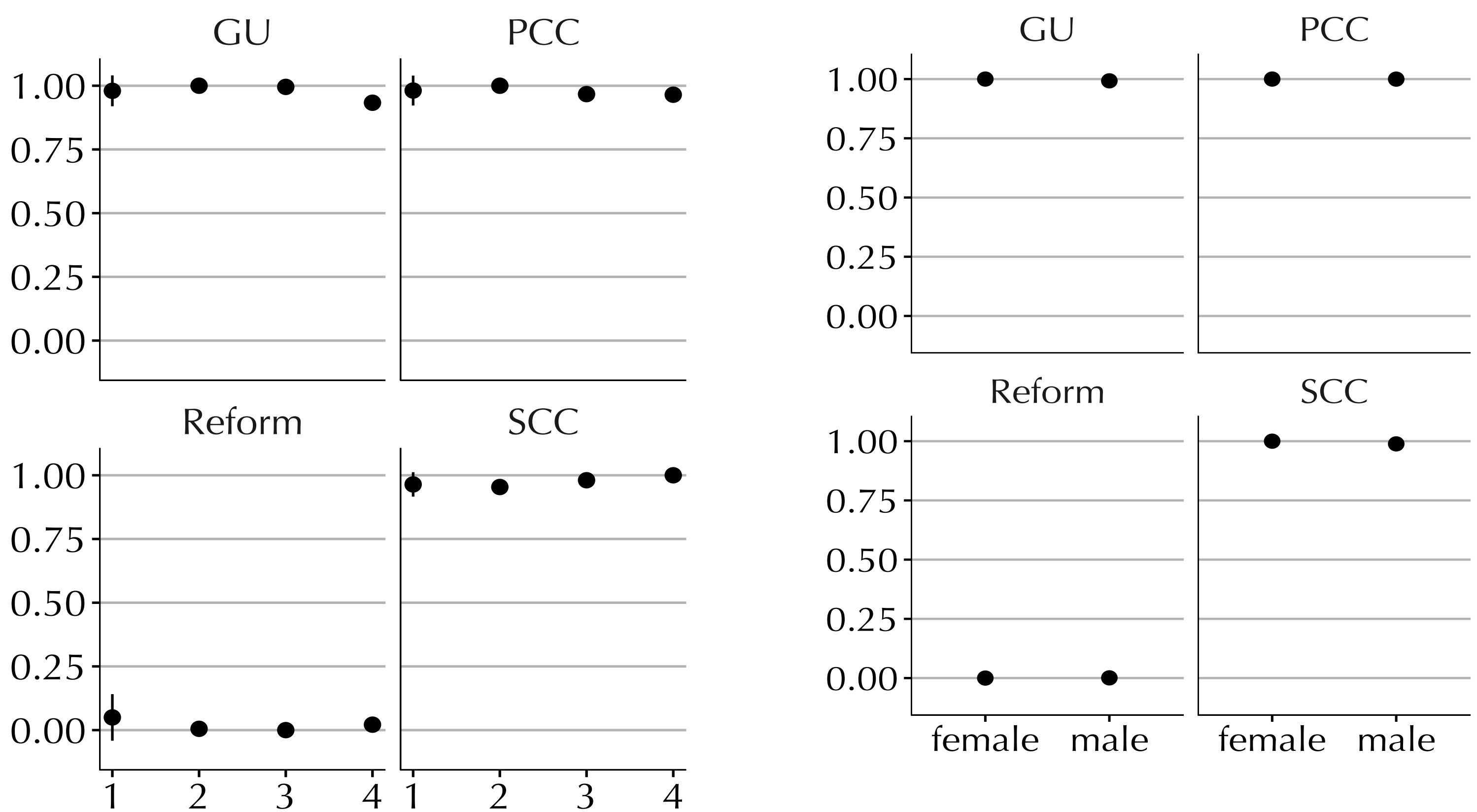


For any two users from different demographics,

1. **Same Query**
2. **Same Information Need:**
    1. Control for user intent: same final SAT click
    2. Only consider navigational queries
3. **Identical top-8 Search Results**

1.2 M impressions
19K unique queries
617K users

Age-wise differences in metrics disappear



- General auditing tool: robust
- Very low coverage across queries
  - Did we control for too much? – lose over 60% of data!

Proposed Approaches

Extremely restrictive
More robust

Generalizable
Less Robust

1) Context Matching

2) Multi-level model

Query-level Multilevel Model

- A hierarchical approach that treats the data as a mixture of distributions based on demographics and queries
- Non-nested multi-level model
  - Users & Queries: nested within non-nested age and gender groups & topics
  - second level captures variation with individual query properties

- Age effects
- Gender effects
- Topic effects
- <age, gender, topic> interaction effects

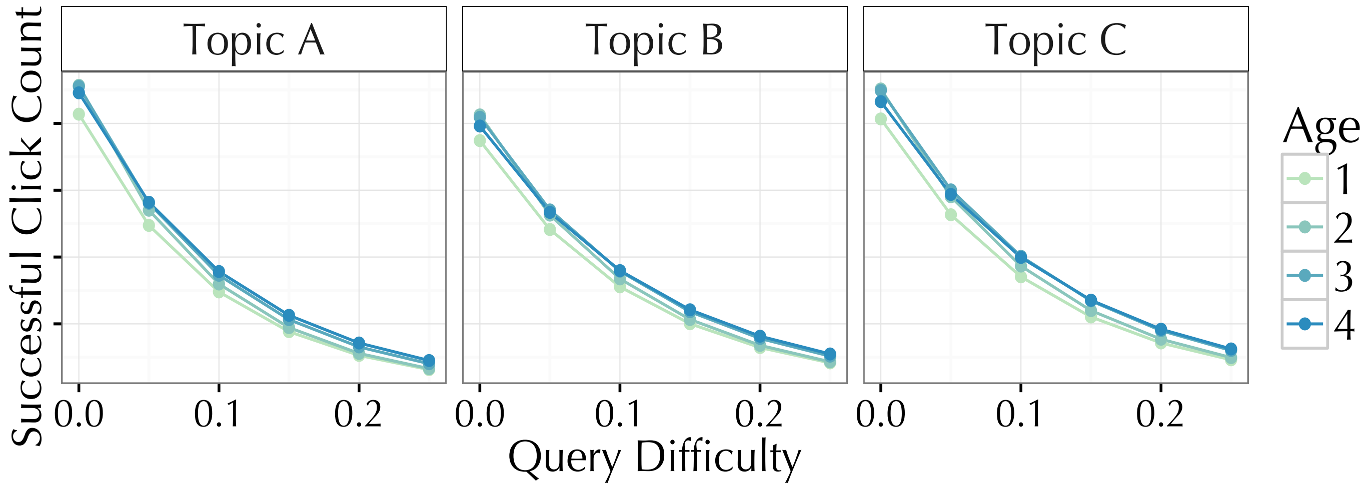
$$E(Y) = f^{-1}(\alpha_{agt} + \beta_{agt} X)$$

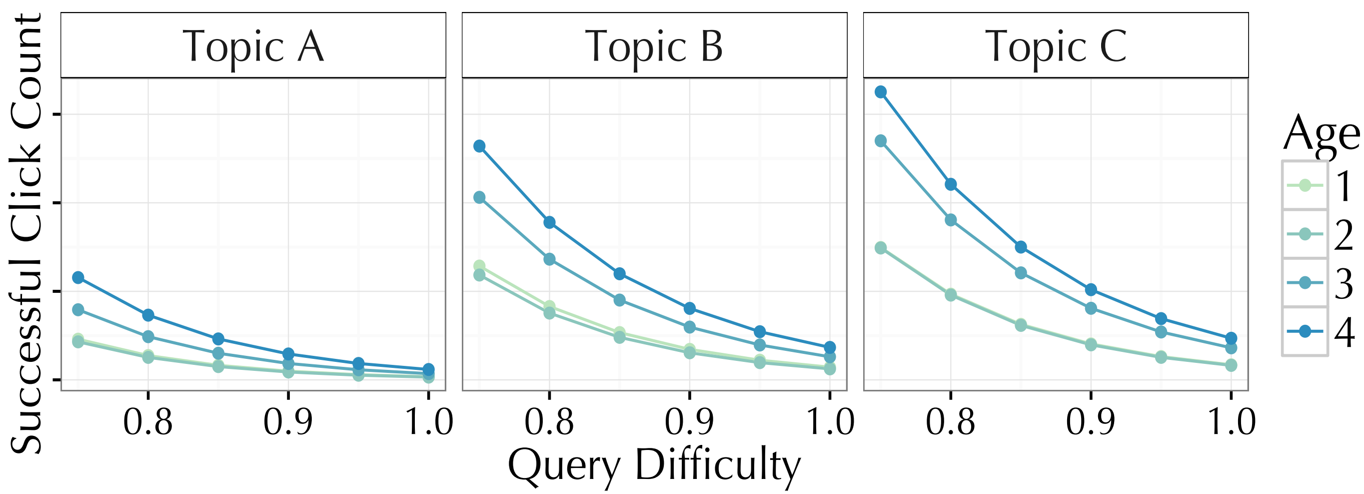
$$\begin{pmatrix} \alpha_{agt} \\ \beta_{agt} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} + \begin{pmatrix} \alpha_a \\ \beta_a \end{pmatrix} + \begin{pmatrix} \alpha_g \\ \beta_g \end{pmatrix} + \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \begin{pmatrix} \alpha_{a \times g \times t} \\ \beta_{a \times g \times t} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_k \right) \quad k \in \{a, g, t\}$$

Specific example: $\mathrm{GU}_i \sim \mathcal{N}(\alpha_{agt} + \beta_{agt} X_i, \sigma_y^2)$

Age-wise differences appear again: bigger differences for harder queries





Outline

From Metric to Satisfaction

- Estimating absolute satisfaction is non-trivial
- We estimate relative satisfaction by considering pairs of impressions
  - which impression led to a higher satisfaction
- Construct a conservative *"high-precision, low-recall"* proxy for pairwise satisfaction
  - by only considering "big" differences in observed metric for the same query

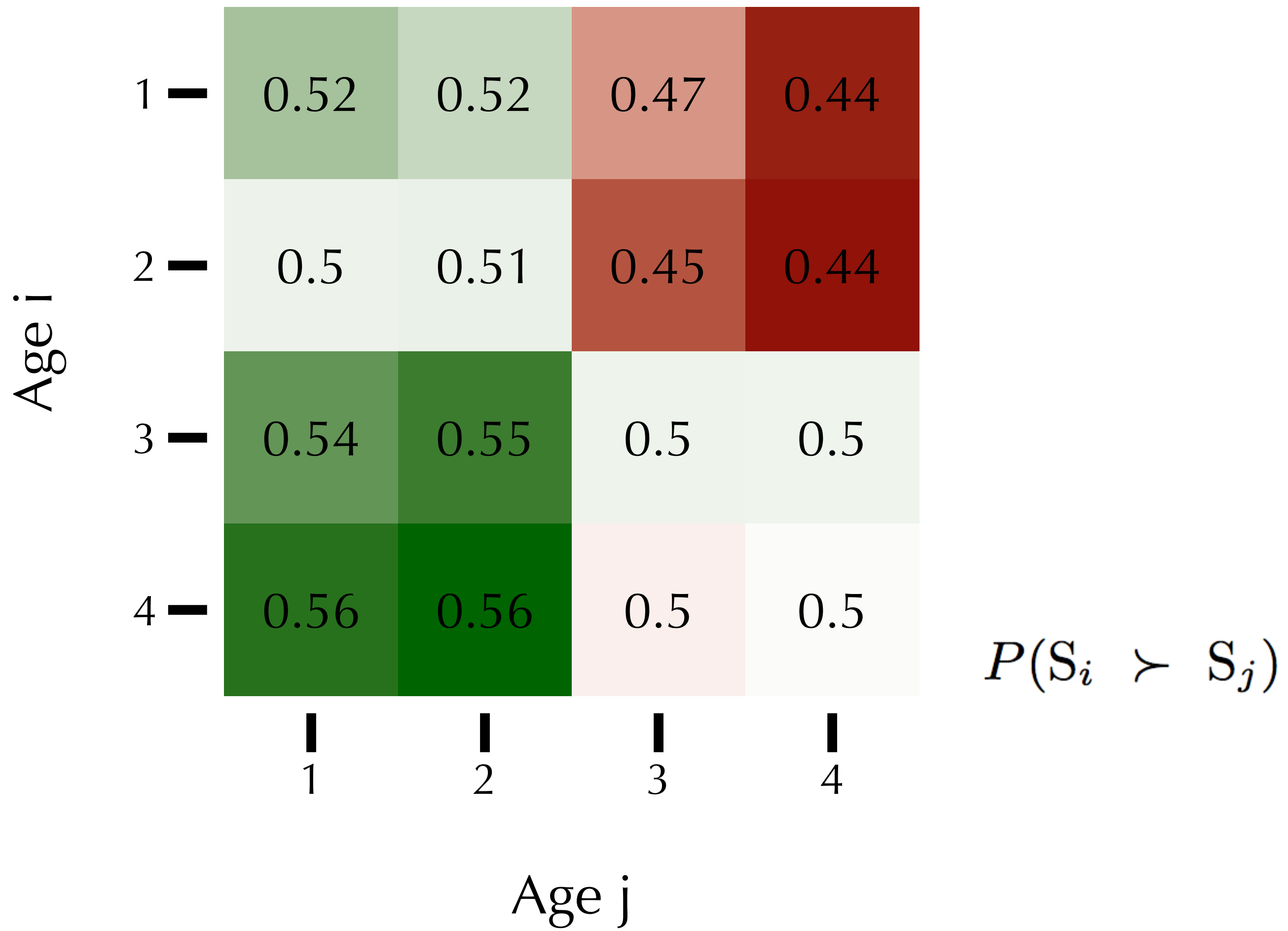
Algorithm 1 Compute satisfaction label

if $\mathrm{RR}_i < \mathrm{RR}_j$ **then return** $+1$
if $\mathrm{RR}_i > \mathrm{RR}_j$ **then return** -1
if $\mathrm{GU}_i - \mathrm{GU}_j > \delta^1_{\mathrm{GU}}$ **then return** $+1$
if $\mathrm{GU}_j - \mathrm{GU}_i > \delta^1_{\mathrm{GU}}$ **then return** -1
if $\mathrm{SCC}_i - \mathrm{SCC}_j > \delta^1_{\mathrm{SCC}}$ **then return** $+1$
if $\mathrm{SCC}_j - \mathrm{SCC}_i > \delta^1_{\mathrm{SCC}}$ **then return** -1
if $\mathrm{GU}_i - \mathrm{GU}_j > \delta^2_{\mathrm{GU}} \wedge \mathrm{SCC}_i - \mathrm{SCC}_j > \delta^2_{\mathrm{SCC}}$ **then return** $+1$
if $\mathrm{GU}_j - \mathrm{GU}_i > \delta^2_{\mathrm{GU}} \wedge \mathrm{SCC}_j - \mathrm{SCC}_i > \delta^2_{\mathrm{SCC}}$ **then return** -1
else return 0

- Logistic regression model for estimating probability of impression i being more satisfied than impression j:

$$P(\mathrm{S}_i \succ \mathrm{S}_j) = \mathrm{logit}^{-1}(\beta_0 + \beta_{a_i} a_i + \beta_{a_j} a_j + \beta_{g_i} g_i + \beta_{g_j} g_j + \beta_{ij} a_i a_j g_i g_j)$$

Again, see a small age-wise difference in satisfaction



- Older users are slightly more satisfied than younger users

Discussion

- Auditing is more nuanced than merely measuring metrics on demographically-binned traffic
  - developed techniques to auditing search engines
- We find light trend towards older users being more satisfied.
- General framework for internally auditing systems
  - Plug-in different metrics
  - Plug-in different demographics/user groups

Future Work

- ➢develop metrics which are not confounded with demographics
- ➢Investigate causes of metric differences
  - ➢Query level analysis
  - ➢SERP level analysis
- ➢Dwell time thresholds for SAT prediction based on demographic information

Auditing is more nuanced than merely measuring metrics on demographically-binned traffic.

General framework for auditing systems

- Plug-in different metrics
- Plug-in different demographics/user groups

Thank You!

Rishabh Mehrotra
PhD candidate @ UCL
http://www.rishabhmehrotra.com

@erishabh
r.mehrotra@cs.ucl.ac.uk

Future Work

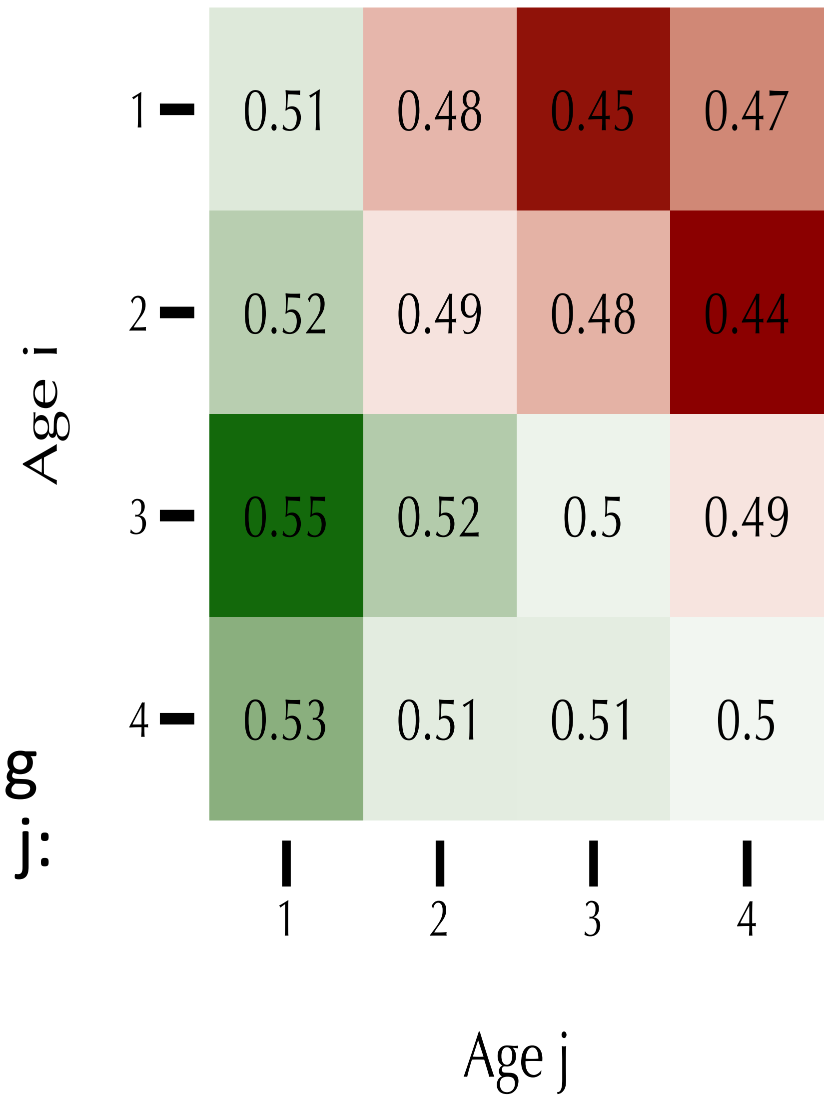
Query	Demographics			Metric Difference
essential oils guide	Female Age 2	vs	Male Age 4	4.5
make your own game	male3	vs	female3	4.25
macbook pro vs macbook air	Female2	vs	male3	3.9
editing software for youtube videos	Male2	vs	male3	3.8333333333333333
emotions	Male2	vs	male4	3.5
avaya phone manual	Female3	vs	male4	3.5
catholic saints	Male4	vs	male3	3.5
futures market	Male3	vs	male5	3.3333333333333333
medal of honor walkthrough ps3	Male3	vs	female2	3.2142857142857144
all wheel drive cars	Male4	vs	female4	3
kob tv albuquerque news 4	Female4	vs	male4-min	3
foods high in iron	Female3	vs	female4	3
478-288-1122	Male3	vs	male4	2.95
cheeseburger dip	Female4	vs	male4	2.8333333333333333
argosy capital	Male3	vs	male4-min	2.5

External Auditing

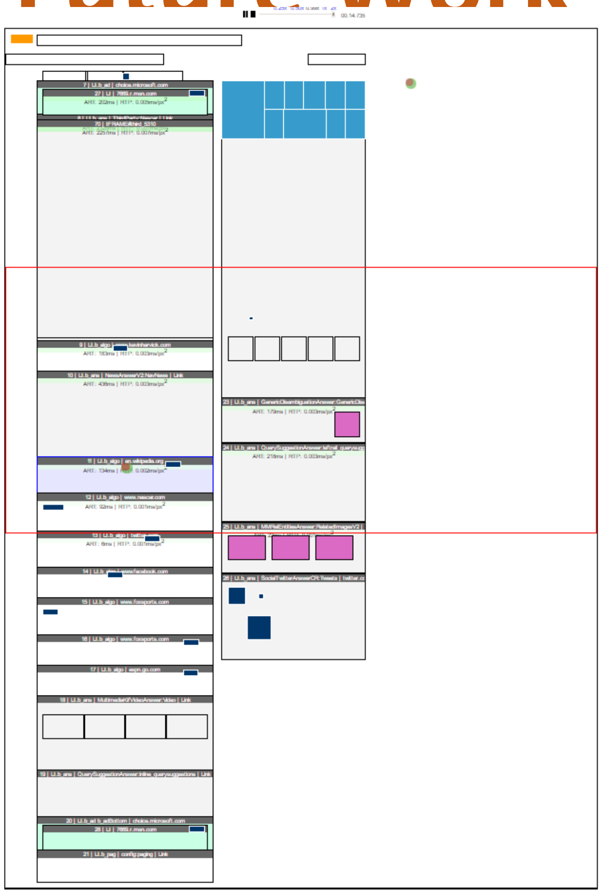
- Experiment on a publicly available dataset
- 2 weeks logs of comScore data
- Use PCC metric to gauge satisfaction
- Probability of impression i being more satisfied than impression j:

$$P(\mathrm{S}_i \succ \mathrm{S}_j) \quad = $$

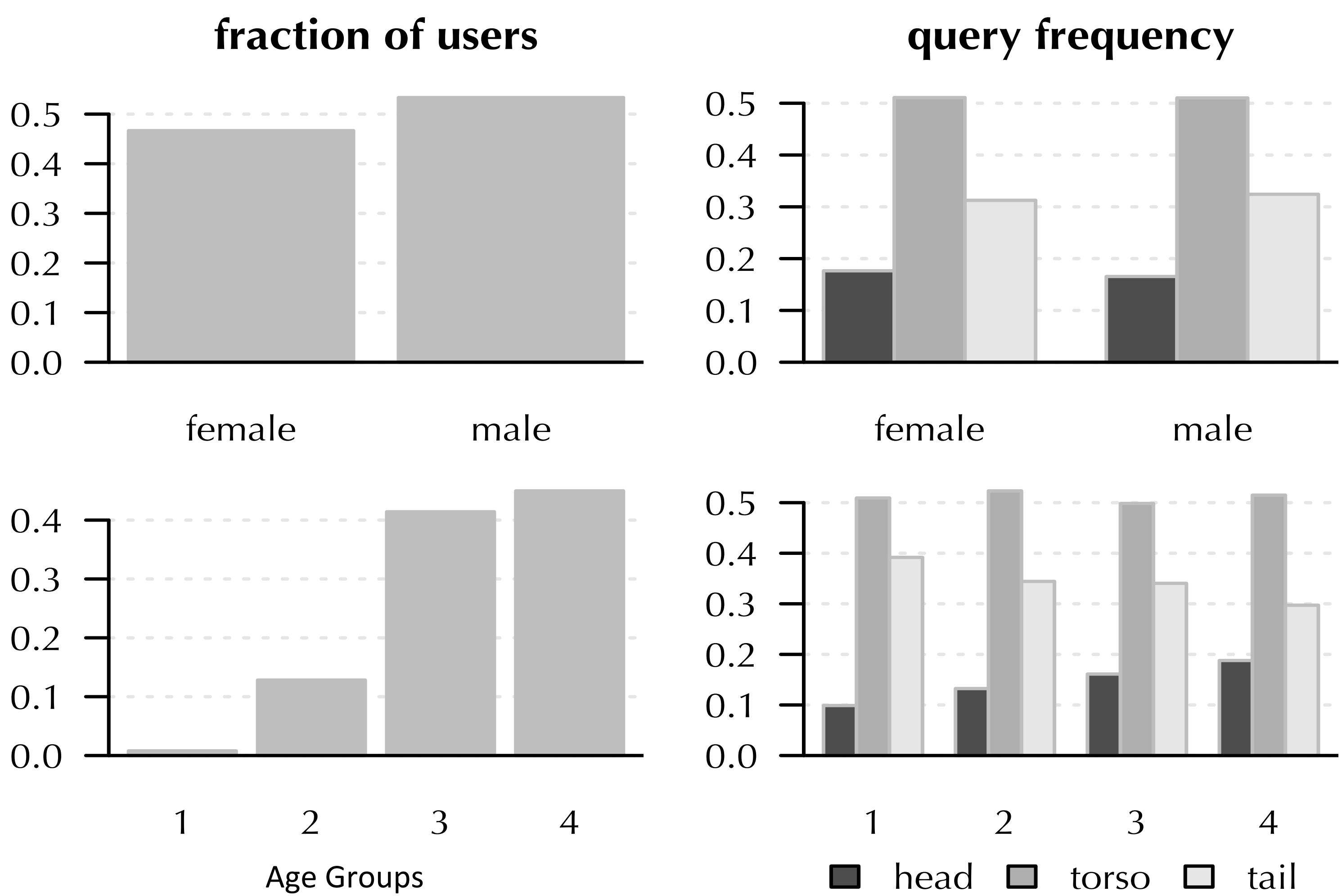
$$\mathrm{logit}^{-1}(\beta_0 + \beta_{a_i} a_i + \beta_{a_j} a_j + \beta_{g_i} g_i + \beta_{g_j} g_j + \beta_{ij} a_i a_j g_i g_j)$$



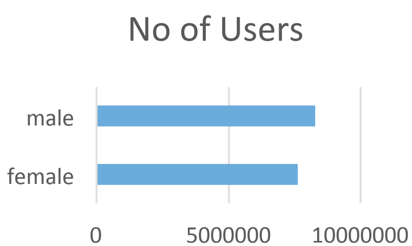
Future Work

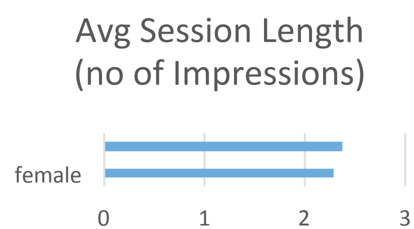


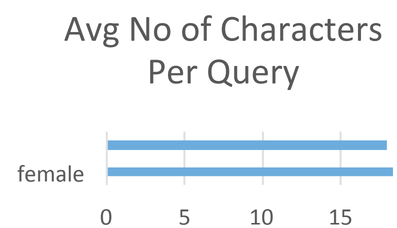
Demographic distribution of user activity

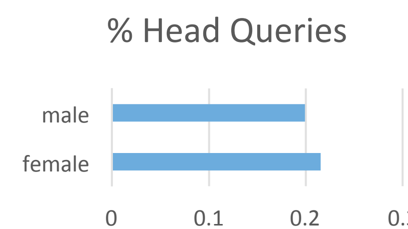


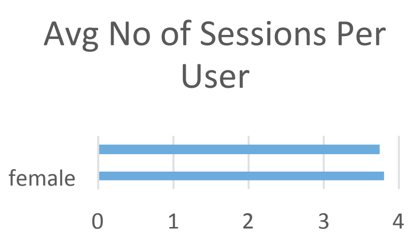
Characterizing Demographics: Gender

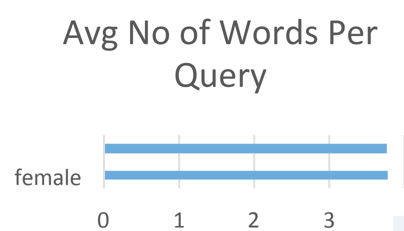


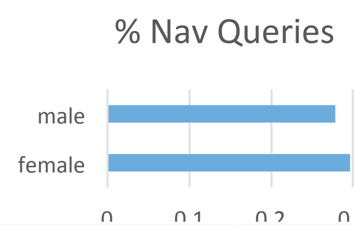


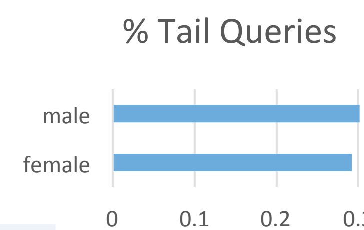












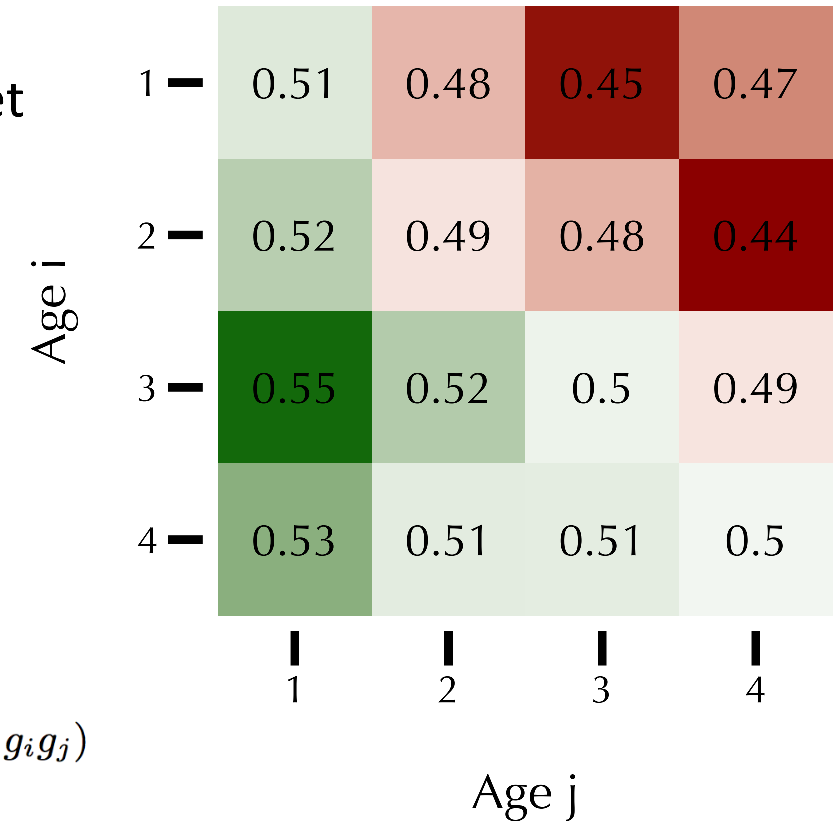
Some highly discriminating queries in terms of P(D|Q):

Male	Female
premier league	pinterest
bbc football	hautelook
watchespn	weight watchers
pirate bay	sephora

External Auditing

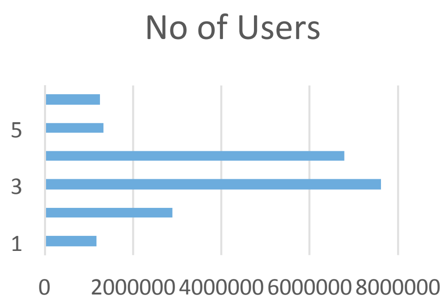
- Experiment on a publicly available dataset
- 2 weeks logs of comScore data
- Use PCC metric to gauge satisfaction
- Probability of impression i being more satisfied than impression j:

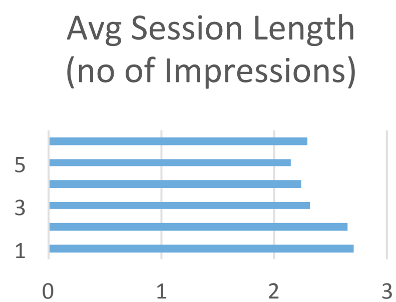
$$P(\mathrm{S}_i \succ \mathrm{S}_j) = \mathrm{logit}^{-1}(\beta_0 + \beta_{a_i} a_i + \beta_{a_j} a_j + \beta_{g_i} g_i + \beta_{g_j} g_j + \beta_{ij} a_i a_j g_i g_j)$$

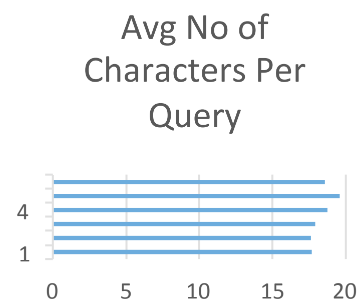


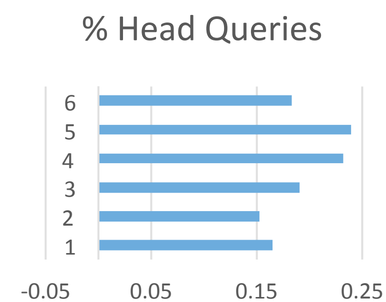
Characterizing Demographics:

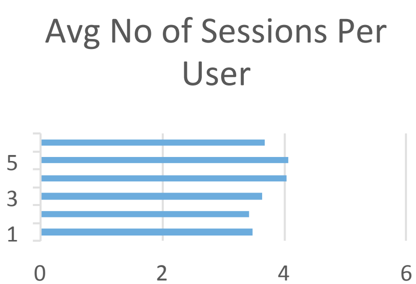
1	<20
2	20-30
3	30-50
4	50-70
5	70-100
6	>100 & NULL

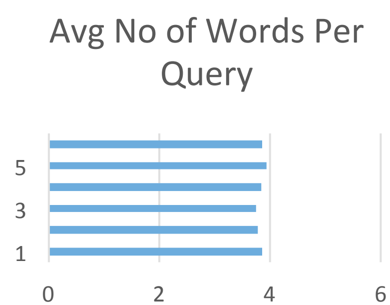


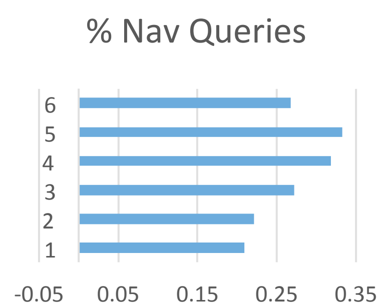


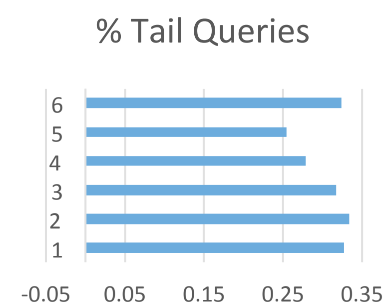












Some highly discriminating queries in terms of P(D|Q):

Age <20	Age: 20-30	Age: 30-50	Age: 50-70
periodic table	debt	spellingcity	ourtime.com dating
mathway	dating	slickdeals	hairstyles women over 50
graphing calculator	school credit	www.linkedin.com	social security benefits

- Young user , Old user
- Issue same query
- See search results
- How satisfied are you?

Query level Difficulty

- X_i: Feature corresponding to inherent difficulty of query
- Typical methods (reformulations, dwell times) employ user behavior – correlated with demographics
- Need a measure unconfounded with demographics
- Method:
  - Per demographic order query by increasing order of avg GU score
  - Compute per demographic percentile of the query (~query's difficulty in each demographic)
  - Mean of percentiles across demographics

Algorithm 1 Compute satisfaction label

if $\mathrm{RR}_i < \mathrm{RR}_j$ **then return** $+1$
if $\mathrm{RR}_i > \mathrm{RR}_j$ **then return** -1
if $\mathrm{GU}_i - \mathrm{GU}_j > \delta^1_{\mathrm{GU}}$ **then return** $+1$
if $\mathrm{GU}_j - \mathrm{GU}_i > \delta^1_{\mathrm{GU}}$ **then return** -1
if $\mathrm{SCC}_i - \mathrm{SCC}_j > \delta^1_{\mathrm{SCC}}$ **then return** $+1$
if $\mathrm{SCC}_j - \mathrm{SCC}_i > \delta^1_{\mathrm{SCC}}$ **then return** -1
if $\mathrm{GU}_i - \mathrm{GU}_j > \delta^2_{\mathrm{GU}} \wedge \mathrm{SCC}_i - \mathrm{SCC}_j > \delta^2_{\mathrm{SCC}}$ **then return** $+1$
if $\mathrm{GU}_j - \mathrm{GU}_i > \delta^2_{\mathrm{GU}} \wedge \mathrm{SCC}_j - \mathrm{SCC}_i > \delta^2_{\mathrm{SCC}}$ **then return** -1
else return 0