Intern Presentation

A/B Testing by Interleaving

Sida Wang

My Project

- Evaluating search relevance by interleaving results and collecting user data
  - Interleaving Framework
    - Generic, Extensible
  - Experiments to evaluate relevance by interleaving
    - Based on the paper How Does Clickthrough Data Reflect Retrieval Quality? by F. Radlinski et al

Evaluating Search Relevance

- Without Interleaving
  - Full time human judges -> precision, recall, NDCG
  - Compare Search

Compare Search

TM
Google
Mini

Result S1

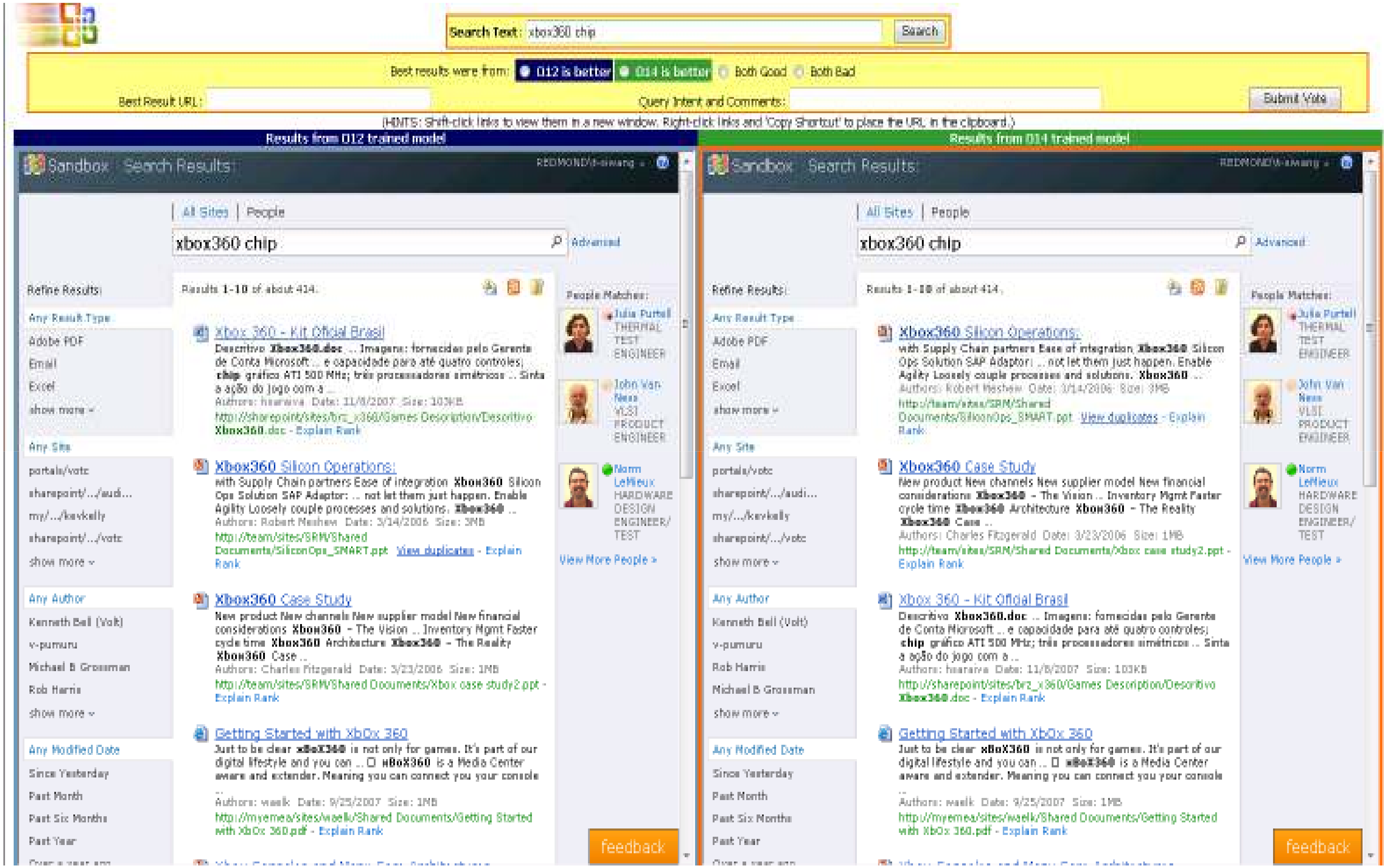
Result S2

Result S3

Result G1

Result G2

Result G3



Search Text: xbox360 chip [Search]

Best results were from: ○ 012 is better ○ 014 is better ○ Both Good ○ Both Bad

Best Result URL:

Query Intent and Comments:

[Submit Vote]

(HINTS: Shift-click links to view them in a new window. Right-click links and 'Copy Shortcut' to place the URL in the clipboard.)

Results from 012 trained model

Sandbox Search Results REDMOND\t-siwang

All Sites | People

xbox360 chip Advanced

Refine Results:

Any Result Type
Adobe PDF
Email
Excel
show more

Any Site
portals/vatc
sharepoint/.../audi...
my/.../kevkelly
sharepoint/.../vatc
show more

Any Author
Kenneth Bell (Volt)
v-pumuru
Michael B Grossman
Rob Harris
show more

Any Modified Date
Since Yesterday
Past Month
Past Six Months
Past Year
Over a year ago

Results **1-10** of about 414.

Xbox 360 - Kit Oficial Brasil
Descritivo **Xbox360.doc** ... Imagens: fornecidas pelo Gerente de Conta Microsoft ... e capacidade para até quatro controles; **chip** gráfico ATI 500 MHz; três processadores simétricos ... Sinta a ação do jogo com a ...
Authors: hsaraiva Date: 11/8/2007 Size: 103KB
http://sharepoint/sites/brz_x360/Games Description/Descritivo **Xbox360**.doc - Explain Rank

Xbox360 Silicon Operations:
with Supply Chain partners Ease of integration **Xbox360** Silicon Ops Solution SAP Adaptor: ... not let them just happen. Enable Agility Loosely couple processes and solutions. **Xbox360** ...
Authors: Robert Meshew Date: 3/14/2006 Size: 3MB
http://team/sites/SRM/Shared Documents/SiliconOps_SMART.ppt View duplicates - Explain Rank

Xbox360 Case Study
New product New channels New supplier model New financial considerations **Xbox360** – The Vision ... Inventory Mgmt Faster cycle time **Xbox360** Architecture **Xbox360** – The Reality **Xbox360** Case ...
Authors: Charles Fitzgerald Date: 3/23/2006 Size: 1MB
http://team/sites/SRM/Shared Documents/Xbox case study2.ppt - Explain Rank

Getting Started with XbOx 360
Just to be clear **xBoX360** is not only for games. It's part of our digital lifestyle and you can ... □ **xBoX360** is a Media Center aware and extender. Meaning you can connect you your console ...
Authors: waelk Date: 9/25/2007 Size: 1MB
http://myemea/sites/waelk/Shared Documents/Getting Started with XbOx 360.pdf - Explain Rank

Xbox Compiler and Many Core Architectures

People Matches:
Julie Purtell THERMAL TEST ENGINEER
John Van Ness VLSI PRODUCT ENGINEER
Norm LeMieux HARDWARE DESIGN ENGINEER/ TEST
View More People »

feedback

Results from 014 trained model

Sandbox Search Results REDMOND\t-siwang

All Sites | People

xbox360 chip Advanced

Refine Results:

Any Result Type
Adobe PDF
Email
Excel
show more

Any Site
portals/vatc
sharepoint/.../audi...
my/.../kevkelly
sharepoint/.../vatc
show more

Any Author
Kenneth Bell (Volt)
v-pumuru
Rob Harris
Michael B Grossman
show more

Any Modified Date
Since Yesterday
Past Month
Past Six Months
Past Year
Over a year ago

Results **1-10** of about 414.

Xbox360 Silicon Operations:
with Supply Chain partners Ease of integration **Xbox360** Silicon Ops Solution SAP Adaptor: ... not let them just happen. Enable Agility Loosely couple processes and solutions. **Xbox360** ...
Authors: Robert Meshew Date: 3/14/2006 Size: 3MB
http://team/sites/SRM/Shared Documents/SiliconOps_SMART.ppt View duplicates - Explain Rank

Xbox360 Case Study
New product New channels New supplier model New financial considerations **Xbox360** – The Vision ... Inventory Mgmt Faster cycle time **Xbox360** Architecture **Xbox360** – The Reality **Xbox360** Case ...
Authors: Charles Fitzgerald Date: 3/23/2006 Size: 1MB
http://team/sites/SRM/Shared Documents/Xbox case study2.ppt - Explain Rank

Xbox 360 - Kit Oficial Brasil
Descritivo **Xbox360.doc** ... Imagens: fornecidas pelo Gerente de Conta Microsoft ... e capacidade para até quatro controles; **chip** gráfico ATI 500 MHz; três processadores simétricos ... Sinta a ação do jogo com a ...
Authors: hsaraiva Date: 11/8/2007 Size: 103KB
http://sharepoint/sites/brz_x360/Games Description/Descritivo **Xbox360**.doc - Explain Rank

Getting Started with XbOx 360
Just to be clear **xBoX360** is not only for games. It's part of our digital lifestyle and you can ... □ **xBoX360** is a Media Center aware and extender. Meaning you can connect you your console ...
Authors: waelk Date: 9/25/2007 Size: 1MB
http://myemea/sites/waelk/Shared Documents/Getting Started with XbOx 360.pdf - Explain Rank

Xbox Compiler and Many Core Architectures

People Matches:
Julie Purtell THERMAL TEST ENGINEER
John Van Ness VLSI PRODUCT ENGINEER
Norm LeMieux HARDWARE DESIGN ENGINEER/ TEST
View More People »

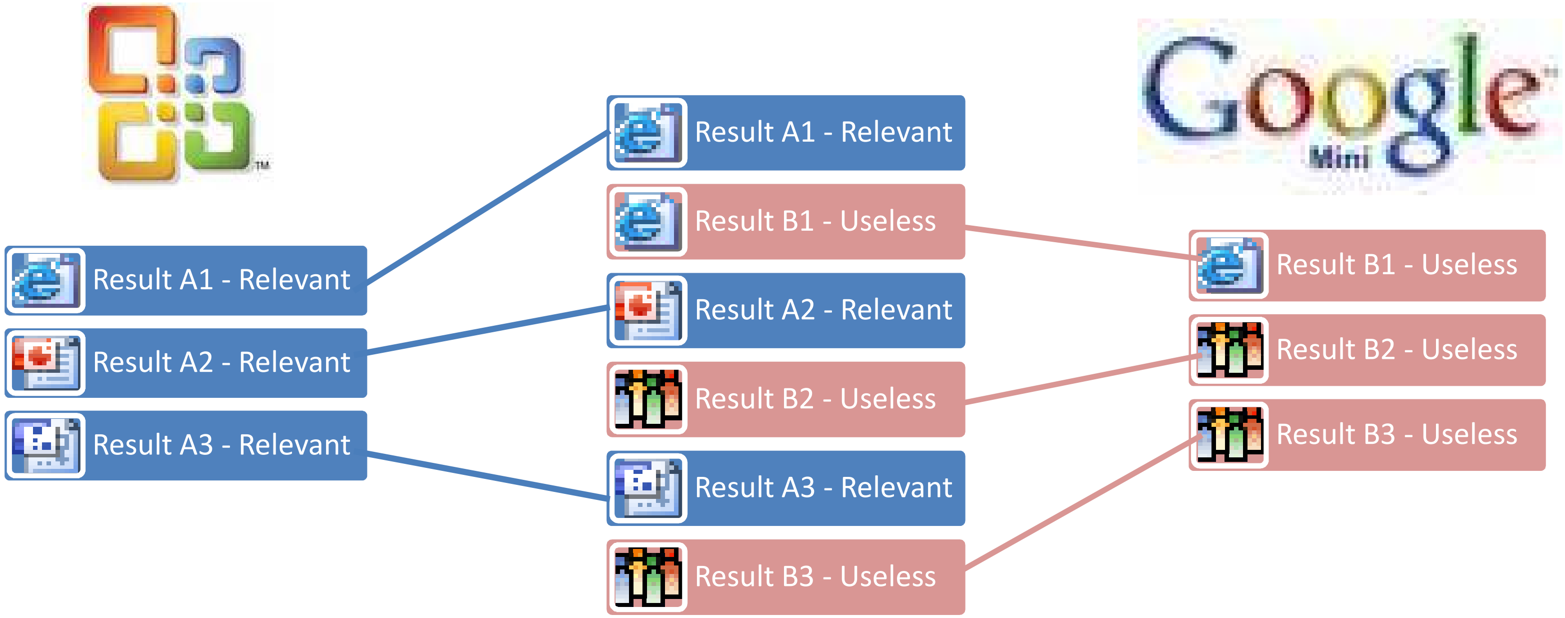
feedback

Issues

-

- But do Microsoft people pick O14 Search or Google Mini?
- Maybe people tend to pick the left?
- **Alters the search experience**
  - Can never collect a lot of data using this method

By Interleaving



Result A1 - Relevant

Result A2 - Relevant

Result A3 - Relevant

Result A1 - Relevant

Result B1 - Useless

Result A2 - Relevant

Result B2 - Useless

Result A3 - Relevant

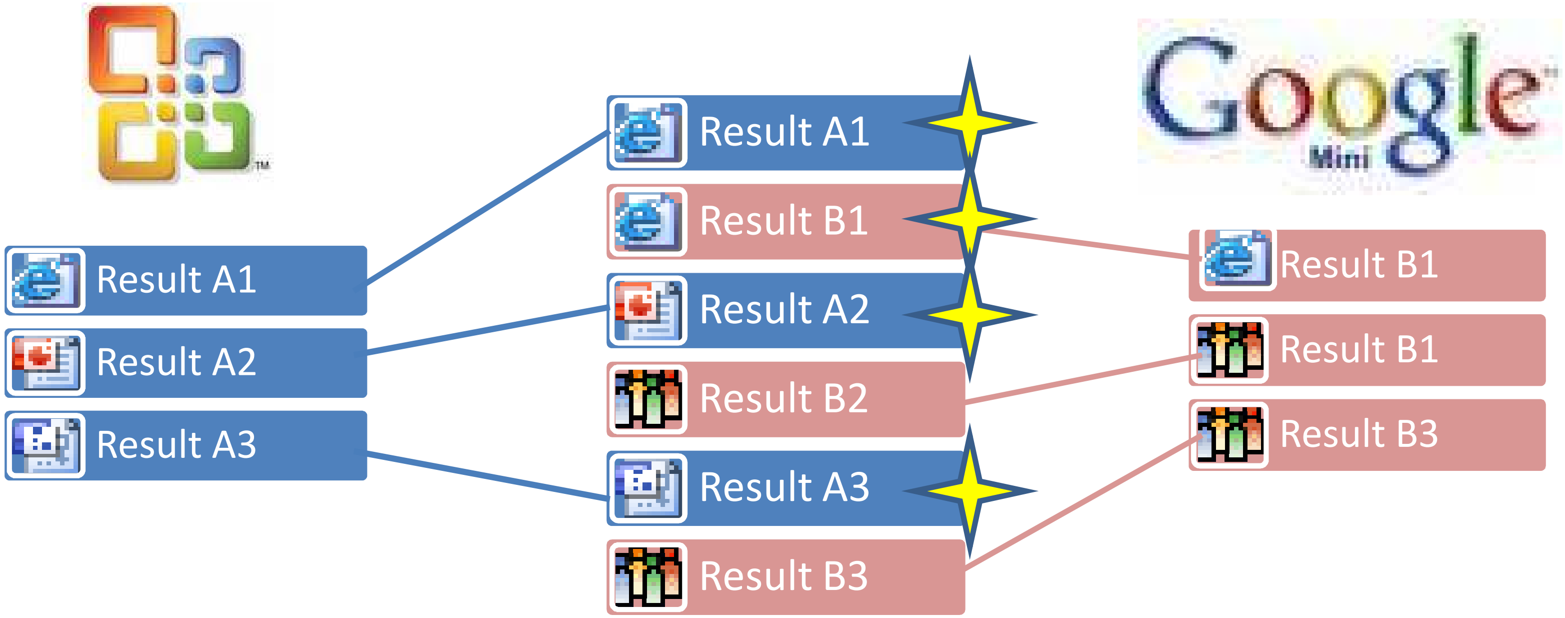
Result B3 - Useless

Result B1 - Useless

Result B2 - Useless

Result B3 - Useless

By Interleaving



Result A1
Result A2
Result A3

Result A1
Result B1
Result A2
Result B2
Result A3
Result B3

Result B1
Result B1
Result B3

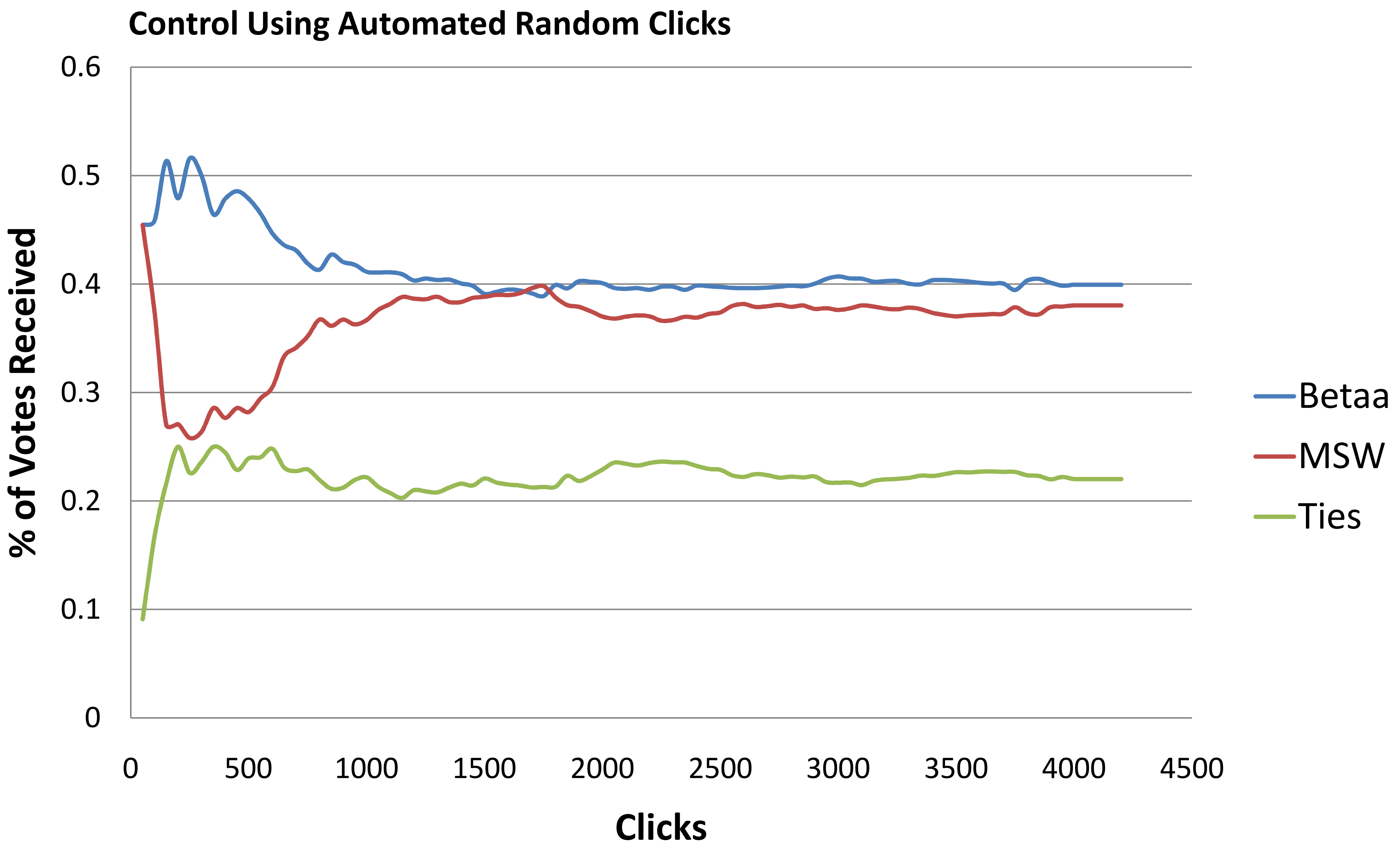
Considerations

- Minimize impact to UX
  - So no demo, it looks exactly like normal search
- Minimize Bias
  - Summary normalization
  - Interleaving algorithms
- Reliability / performance / and the usual

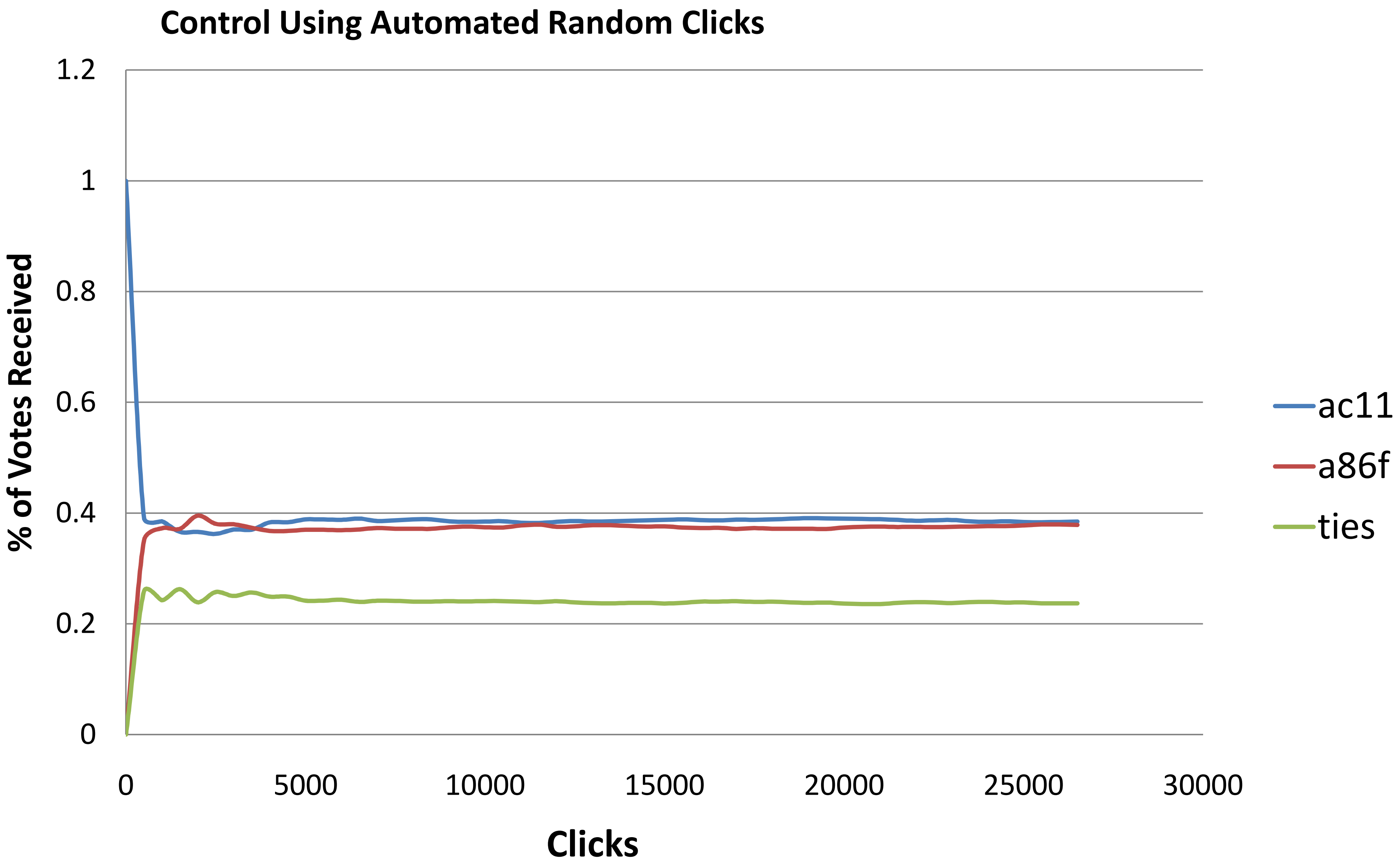
Experiments I did

- **Automated random clicks**
- Automated clicks according to relevance judgments
- Clicks from real people

Random Clicks



A Lot of Random Clicks

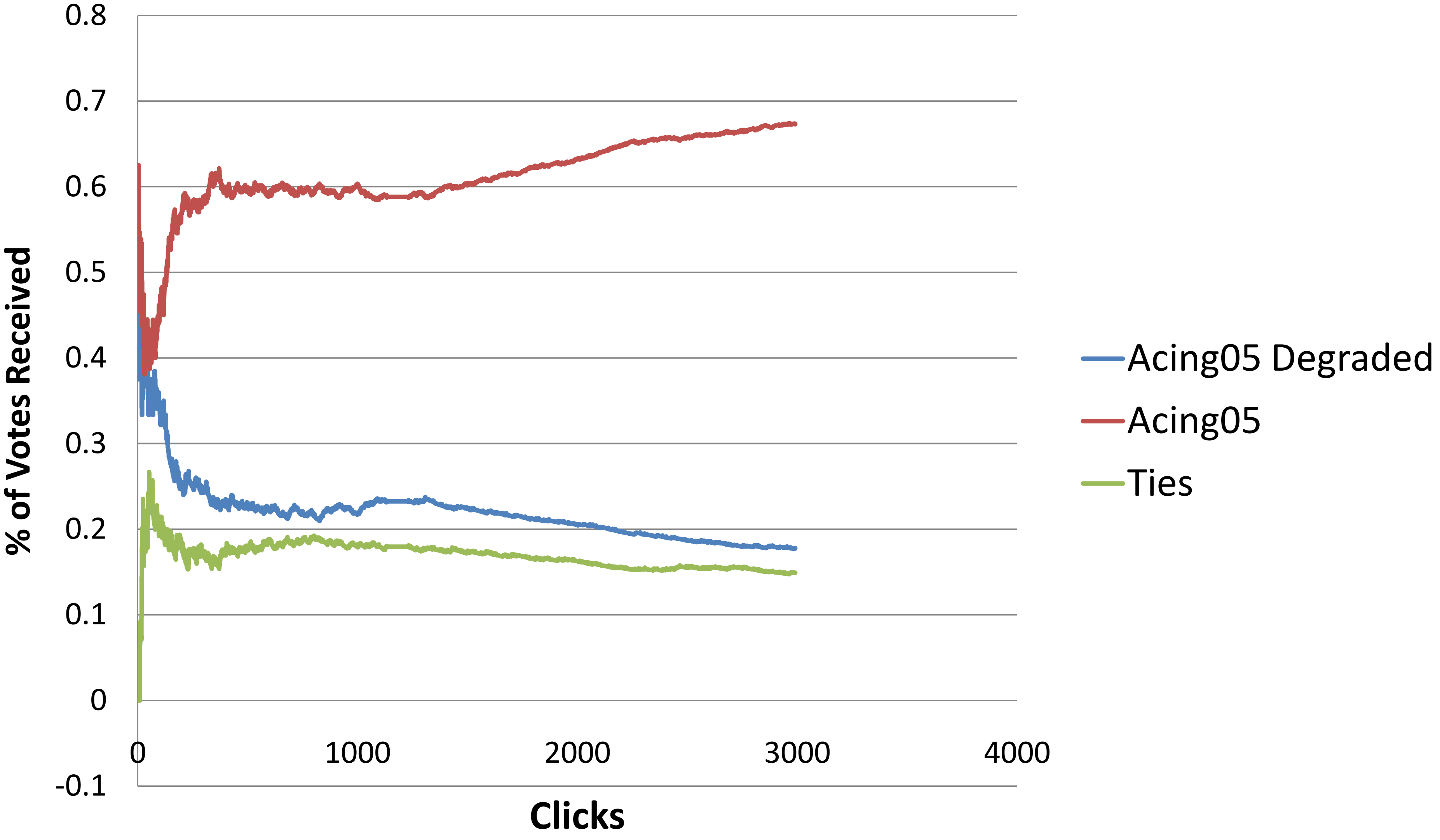


Experiments I did

- Automated random clicks
- **Automated clicks according to relevance judgments**
- Clicks from real people

O12 vs. O14

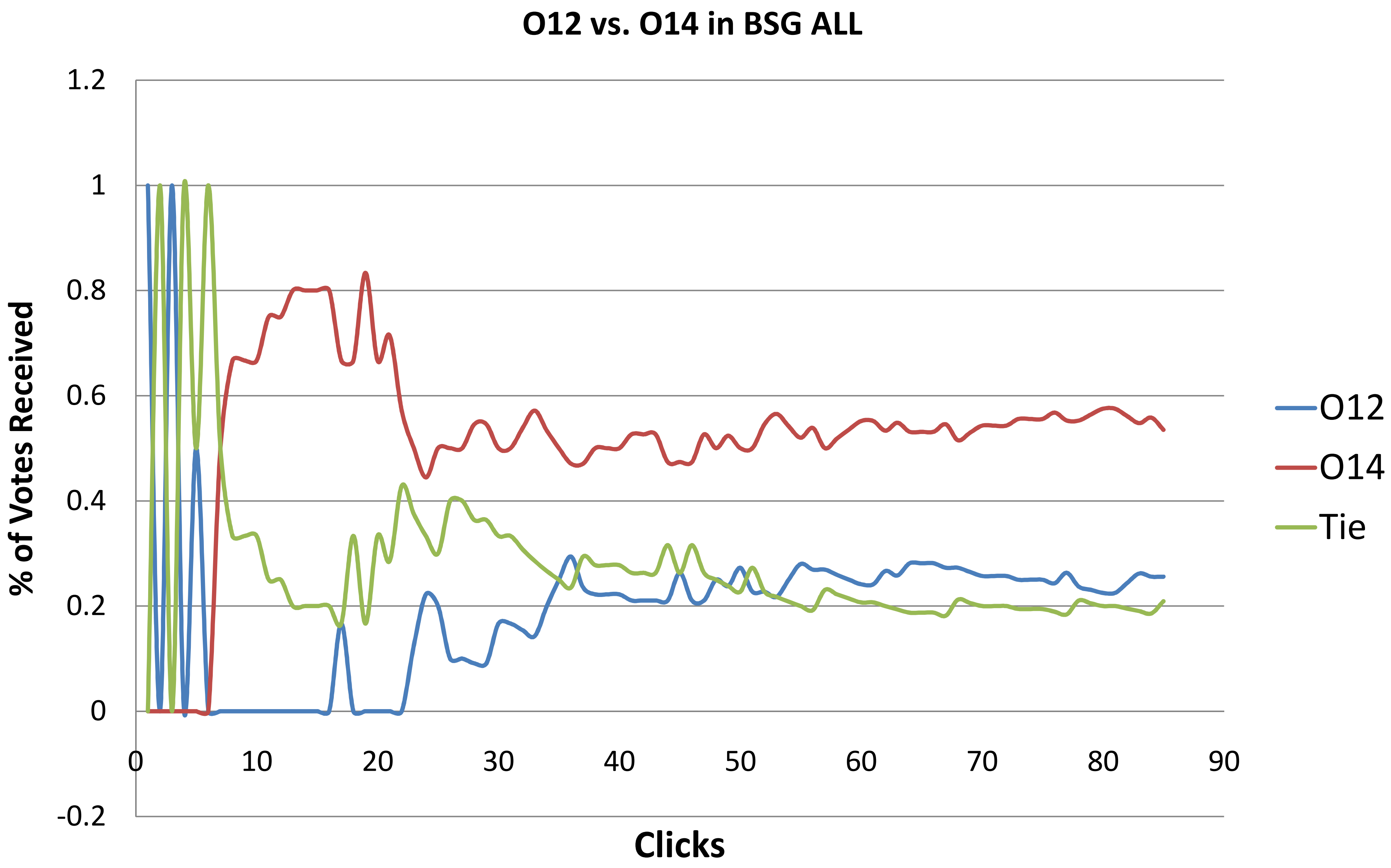
Automated Clicks Using Relevence Judgments



Experiments I did

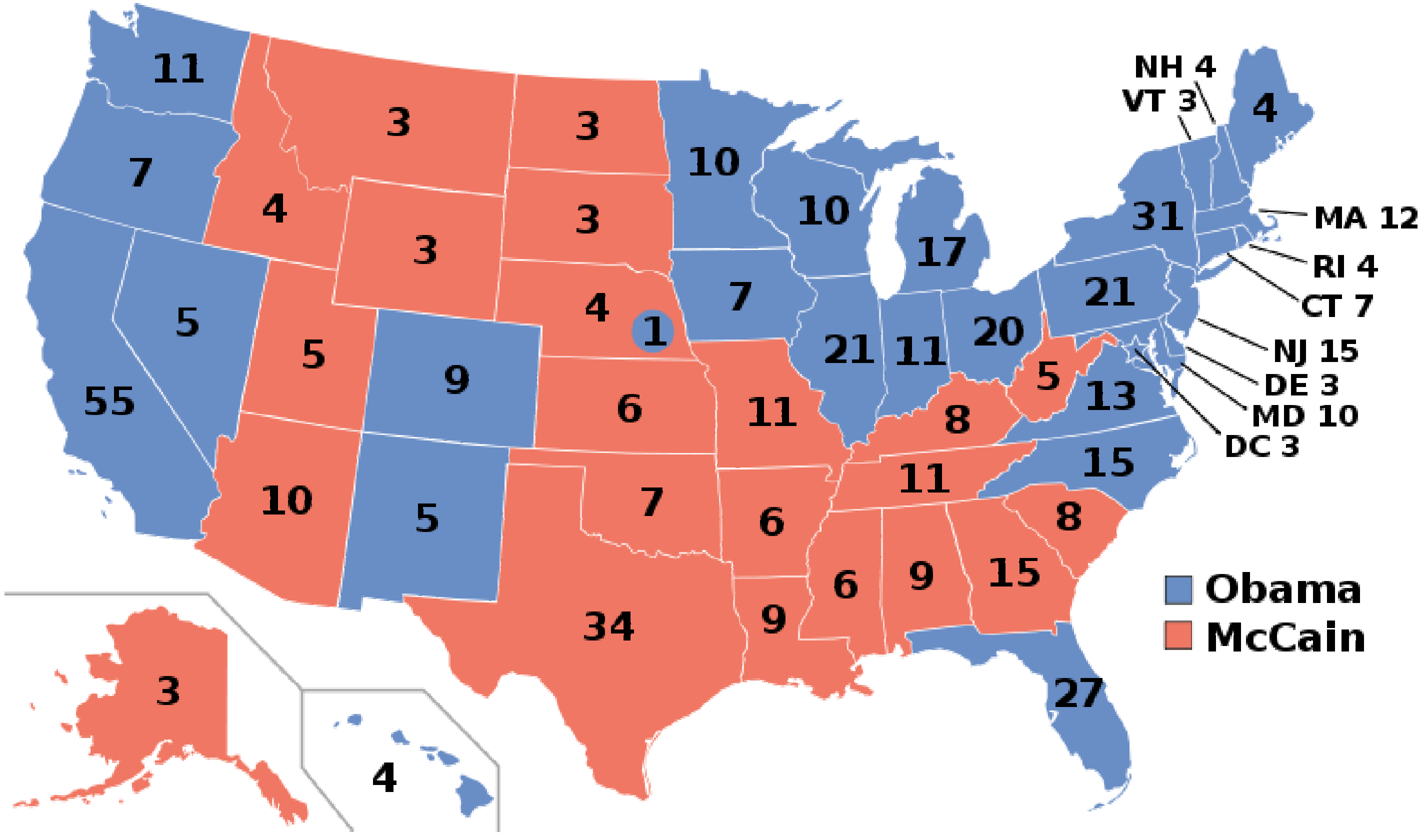
- Automated random clicks
- Automated clicks according to relevance judgments
- **Clicks from real people**

O12 vs. O14


O12 vs. O14 in BSG ALL
% of Votes Received
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
10
20
30
40
50
60
70
80
90
Clicks
O12
O14
Tie

Method of Analysis (election)

- Vote by query, by user, by session etc.
- query = person, user = state



Summary of Results

Method of Voting	**O12 vs. O14**
by queries (direct election):	12 vs. 24
by users (1 vote per state):	4 vs. 9
by sessions (~electoral votes):	5 vs. 11

- System does not seem to matter much, but too little clicks (85) to draw significant conclusion

What Logically Follows

- Google Mini vs. O14 (after fixing Google Mini)
- FAST vs. O14 (after fixing RSS in fssearchoffice)
- I’d love to see the results

What can interleaving do?

- **Give relevance team more confidence**
- Use interleaving for displaying results
- Use interleaving to automatically tune the search engine

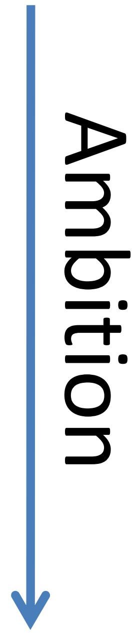
Ambition

Add Confidence

- In addition to very traditional measures like NDCG, Precision and Recall. It is nice to have another independent metric.
- Automatic
  - Does not require human judgments
- Scalable
  - Small impact to UX

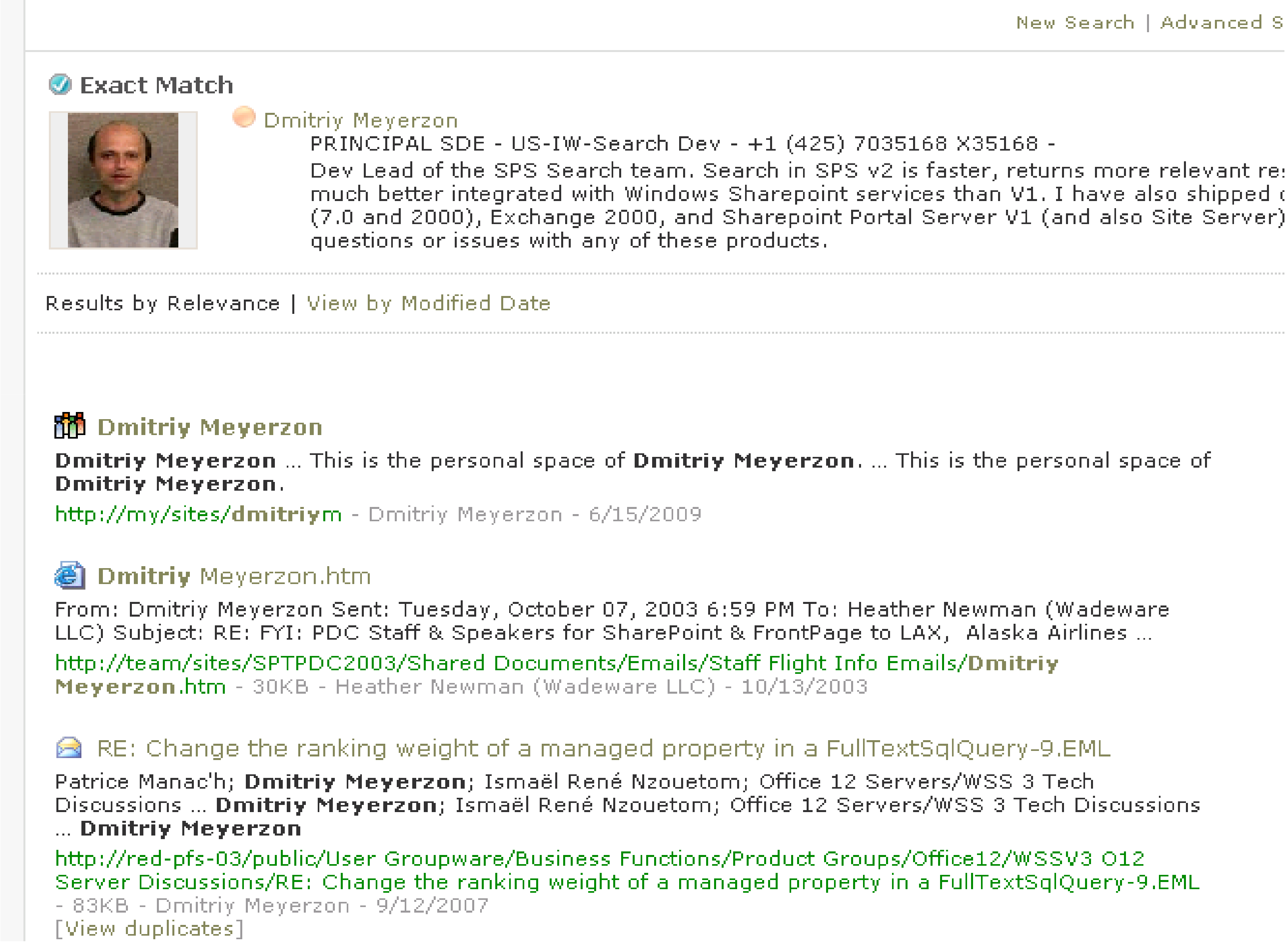
What can interleaving do?

- Give relevance team more confidence
- **Use interleaving for displaying results**
- Use interleaving to automatically tune the search engine



Ambition

Display



Exact Match

Dmitriy Meyerzon

PRINCIPAL SDE - US-IW-Search Dev - +1 (425) 7035168 X35168 -

Dev Lead of the SPS Search team. Search in SPS v2 is faster, returns more relevant re: much better integrated with Windows Sharepoint services than V1. I have also shipped ((7.0 and 2000), Exchange 2000, and Sharepoint Portal Server V1 (and also Site Server) questions or issues with any of these products.

Results by Relevance | View by Modified Date

Dmitriy Meyerzon

Dmitriy Meyerzon ... This is the personal space of **Dmitriy Meyerzon**. ... This is the personal space of **Dmitriy Meyerzon**.

http://my/sites/**dmitriym** - Dmitriy Meyerzon - 6/15/2009

Dmitriy Meyerzon.htm

From: Dmitriy Meyerzon Sent: Tuesday, October 07, 2003 6:59 PM To: Heather Newman (Wadeware LLC) Subject: RE: FYI: PDC Staff & Speakers for SharePoint & FrontPage to LAX, Alaska Airlines ...

http://team/sites/SPTPDC2003/Shared Documents/Emails/Staff Flight Info Emails/**Dmitriy Meyerzon**.htm - 30KB - Heather Newman (Wadeware LLC) - 10/13/2003

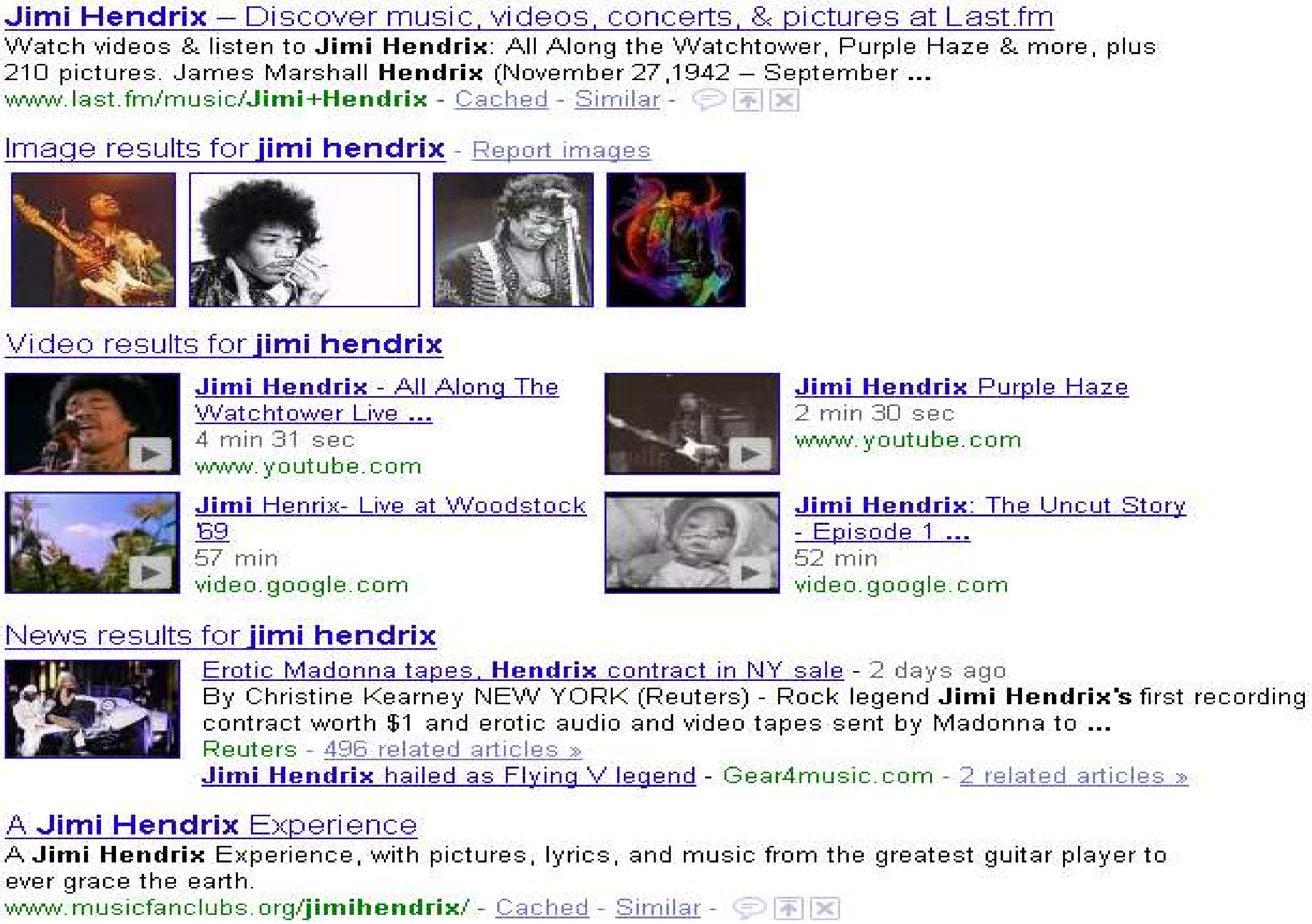
RE: Change the ranking weight of a managed property in a FullTextSqlQuery-9.EML

Patrice Manac'h; **Dmitriy Meyerzon**; Ismaël René Nzouetom; Office 12 Servers/WSS 3 Tech Discussions ... **Dmitriy Meyerzon**; Ismaël René Nzouetom; Office 12 Servers/WSS 3 Tech Discussions ... **Dmitriy Meyerzon**

http://red-pfs-03/public/User Groupware/Business Functions/Product Groups/Office12/WSSV3 O12 Server Discussions/RE: Change the ranking weight of a managed property in a FullTextSqlQuery-9.EML - 83KB - Dmitriy Meyerzon - 9/12/2007

[View duplicates]

Display



Jimi Hendrix – Discover music, videos, concerts, & pictures at Last.fm
Watch videos & listen to **Jimi Hendrix**: All Along the Watchtower, Purple Haze & more, plus 210 pictures. James Marshall **Hendrix** (November 27,1942 – September **...**
www.last.fm/music/**Jimi+Hendrix** - Cached - Similar -

Image results for **jimi hendrix** - Report images

Video results for **jimi hendrix**

Jimi Hendrix - All Along The Watchtower Live **...**
4 min 31 sec
www.youtube.com

Jimi Hendrix Purple Haze
2 min 30 sec
www.youtube.com

Jimi Henrix- Live at Woodstock '69
57 min
video.google.com

Jimi Hendrix: The Uncut Story - Episode 1 **...**
52 min
video.google.com

News results for **jimi hendrix**

Erotic Madonna tapes, **Hendrix** contract in NY sale - 2 days ago
By Christine Kearney NEW YORK (Reuters) - Rock legend **Jimi Hendrix's** first recording contract worth $1 and erotic audio and video tapes sent by Madonna to **...**
Reuters - 496 related articles »
Jimi Hendrix hailed as Flying V legend - Gear4music.com - 2 related articles »

A **Jimi Hendrix** Experience
A **Jimi Hendrix** Experience, with pictures, lyrics, and music from the greatest guitar player to ever grace the earth.
www.musicfanclubs.org/**jimihendrix**/ - Cached - Similar -

What can we do?

- Give relevance team more confidence
- Use interleave for displaying results
- **Use interleaving to automatically tune the search engine**

Ambition

Automatic Tuning

- Many relevance models, each is good for a particular type of corpora (specs, user data, academic articles, product catalog, websites)

- Use interleaving in 10% of searches

- Use user click data to:
  - Automatically and dynamically decide on the best model, or tweak model parameters

Thank you!

- Dmitriy, Eugene, Puneet
- Jamie, Jessica, Ping, Victor, Relevance Team

- Russ, Jon
- Search Team

- Hope to see you again in the future!

Extra Slides

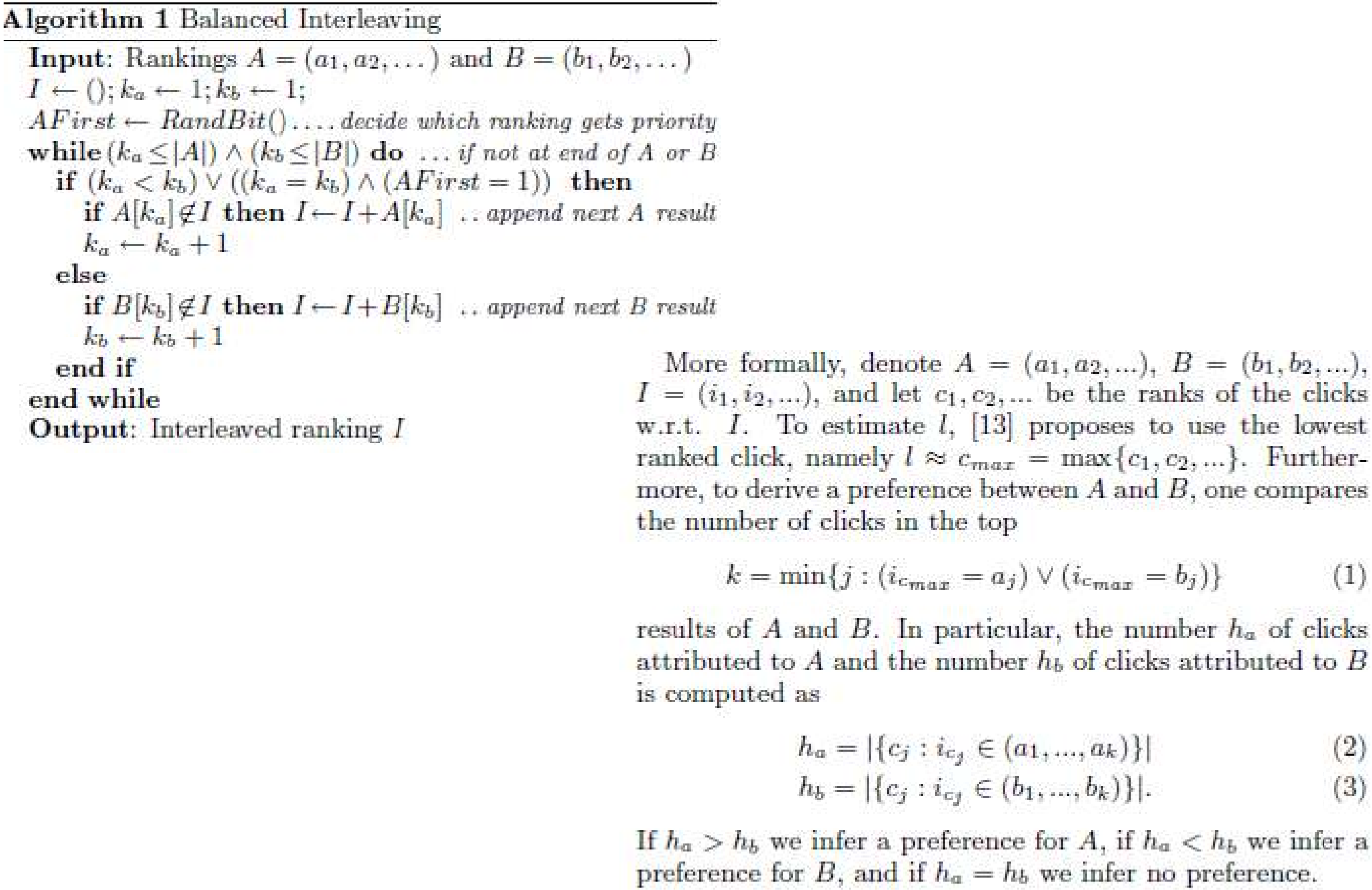
Automatic Tuning – Pair wise?

- Pair wise comparisons scales poorly
- But there seems to be “strong stochastic transitivity”
  - Given locations A, B ,C
  - If A > B > C then $\Delta_{AC} > \text{Max}(\Delta_{AB}, \Delta_{BC})$

How to Interleave

- Balanced
- Team Draft

Balanced Interleaving



Algorithm 1 Balanced Interleaving

```
Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
$I \leftarrow ()$; $k_a \leftarrow 1$; $k_b \leftarrow 1$;
$AFirst \leftarrow RandBit()$ .... decide which ranking gets priority
while $(k_a \leq |A|) \wedge (k_b \leq |B|)$ do ... if not at end of A or B
  if $(k_a < k_b) \vee ((k_a = k_b) \wedge (AFirst = 1))$ then
    if $A[k_a] \notin I$ then $I \leftarrow I + A[k_a]$ .. append next A result
    $k_a \leftarrow k_a + 1$
  else
    if $B[k_b] \notin I$ then $I \leftarrow I + B[k_b]$ .. append next B result
    $k_b \leftarrow k_b + 1$
  end if
end while
Output: Interleaved ranking $I$
```

More formally, denote $A = (a_1, a_2, ...)$, $B = (b_1, b_2, ...)$, $I = (i_1, i_2, ...)$, and let $c_1, c_2, ...$ be the ranks of the clicks w.r.t. I. To estimate l, [13] proposes to use the lowest ranked click, namely $l \approx c_{max} = \max\{c_1, c_2, ...\}$. Furthermore, to derive a preference between A and B, one compares the number of clicks in the top

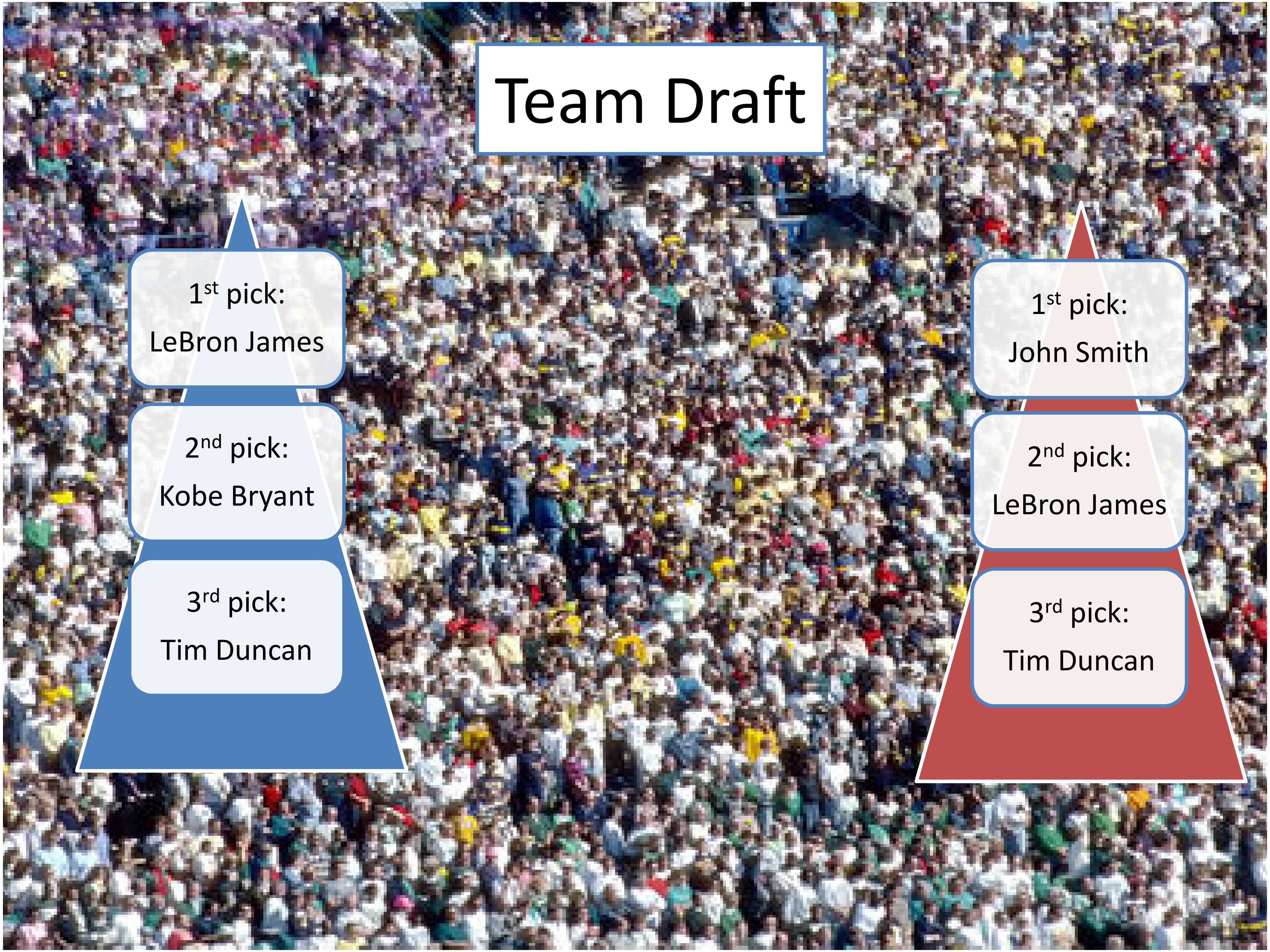
$$k = \min\{j : (i_{c_{max}} = a_j) \vee (i_{c_{max}} = b_j)\} \tag{1}$$

results of A and B. In particular, the number h_a of clicks attributed to A and the number h_b of clicks attributed to B is computed as

$$h_a = |\{c_j : i_{c_j} \in (a_1, ..., a_k)\}| \tag{2}$$
$$h_b = |\{c_j : i_{c_j} \in (b_1, ..., b_k)\}|. \tag{3}$$

If $h_a > h_b$ we infer a preference for A, if $h_a < h_b$ we infer a preference for B, and if $h_a = h_b$ we infer no preference.

Team Draft



1st pick:
LeBron James

2nd pick:
Kobe Bryant

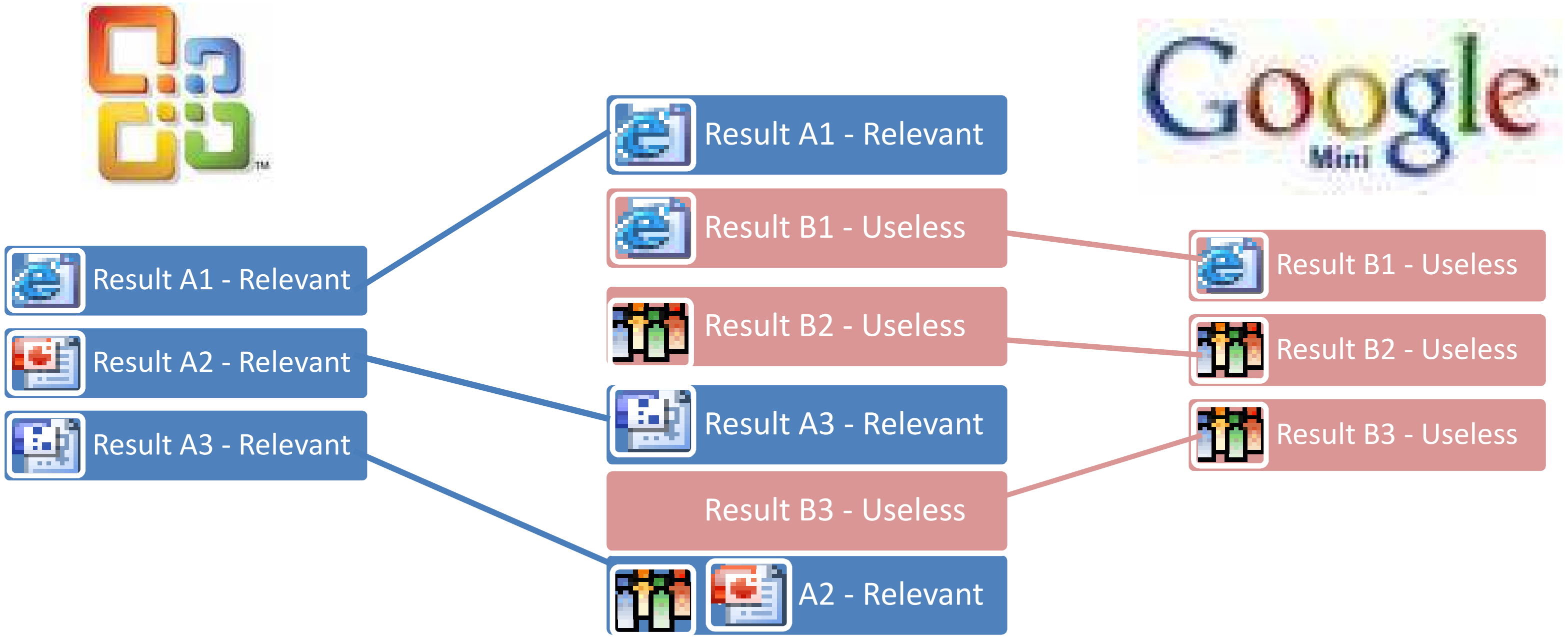
3rd pick:
Tim Duncan

1st pick:
John Smith

2nd pick:
LeBron James

3rd pick:
Tim Duncan

A/B Testing By Interleaving



Result A1 - Relevant

Result A2 - Relevant

Result A3 - Relevant

Result A1 - Relevant

Result B1 - Useless

Result B2 - Useless

Result A3 - Relevant

Result B3 - Useless

A2 - Relevant

Result B1 - Useless

Result B2 - Useless

Result B3 - Useless