

Large-Scale Validation and Analysis of Interleaved Search Evaluation

Olivier Chapelle, Thorsten Joachims
Filip Radlinski, Yisong Yue

Department of Computer Science
Cornell University

Decide between two Ranking Functions

Distribution $P(u,q)$
of users u , queries q

\vdots
 $(t_j, \text{"SVM"})$
 \vdots

Retrieval Function 1

$$f_1(u,q) \rightarrow r_1$$

Which one
is better?

Retrieval Function 2

$$f_2(u,q) \rightarrow r_2$$

1. Kernel Machines
<http://svm.first.gmd.de/>
2. SVM-Light Support Vector Machine
<http://svmlight.joachims.org/>
3. School of Veterinary Medicine at UPenn
<http://www.vet.upenn.edu/>
4. An Introduction to Support Vector Machines
<http://www.support-vector.net/>
5. Service Master Company
<http://www.servicemaster.com/>

\vdots

$U(t_j, \text{"SVM"}, r_1)$

1. School of Veterinary Medicine at UPenn
<http://www.vet.upenn.edu/>
2. Service Master Company
<http://www.servicemaster.com/>
3. Support Vector Machine
<http://jbolivar.freesevers.com/>
4. Archives of SUPPORT-VECTOR-MACHINES
<http://www.jiscmail.ac.uk/lists/SUPPORT...>
5. SVM-Light Support Vector Machine
[http://ais.gmd.de/~thorsten/svm light/](http://ais.gmd.de/~thorsten/svm%20light/)

\vdots

$U(t_j, \text{"SVM"}, r_2)$

Implicit Utility Feedback

- Approach 1: Absolute Metrics
 - Do metrics derived from observed user behavior provide absolute feedback about retrieval quality of f ?
 - For example:
 - $U(f) \sim \text{numClicks}(f)$
 - $U(f) \sim 1/\text{abandonment}(f)$
- Approach 2: Paired Comparison Tests
 - Do paired comparison tests provide relative preferences between two retrieval functions f_1 and f_2 ?
 - For example:
 - $f_1 \succ f_2 \Leftrightarrow \text{pairedCompTest}(f_1, f_2) > 0$

Absolute Metrics: Metrics

Name	Description	Aggregation	Hypothesized Change with Decreased Quality
Abandonment Rate	% of queries with no click	N/A	Increase
Reformulation Rate	% of queries that are followed by reformulation	N/A	Increase
Queries per Session	Session = no interruption of more than 30 minutes	Mean	Increase
Clicks per Query	Number of clicks	Mean	Decrease
Click@1	% of queries with clicks at position 1	N/A	Decrease
Max Reciprocal Rank*	1/rank for highest click	Mean	Decrease
Mean Reciprocal Rank*	Mean of 1/rank for all clicks	Mean	Decrease
Time to First Click*	Seconds before first click	Median	Increase
Time to Last Click*	Seconds before final click	Median	Decrease

(*) only queries with at least one click count

How does User Behavior Reflect Retrieval Quality?

User Study in ArXiv.org

- Natural user and query population
- User in natural context, not lab
- Live and operational search engine
- Ground truth by construction

ORIG \succ SWAP2 \succ SWAP4

- ORIG: Hand-tuned fielded
- SWAP2: ORIG with 2 pairs swapped
- SWAP4: ORIG with 4 pairs swapped

ORIG \succ FLAT \succ RAND

- ORIG: Hand-tuned fielded
- FLAT: No field weights
- RAND : Top 10 of FLAT shuffled

arXiv.org Full Text Search Results

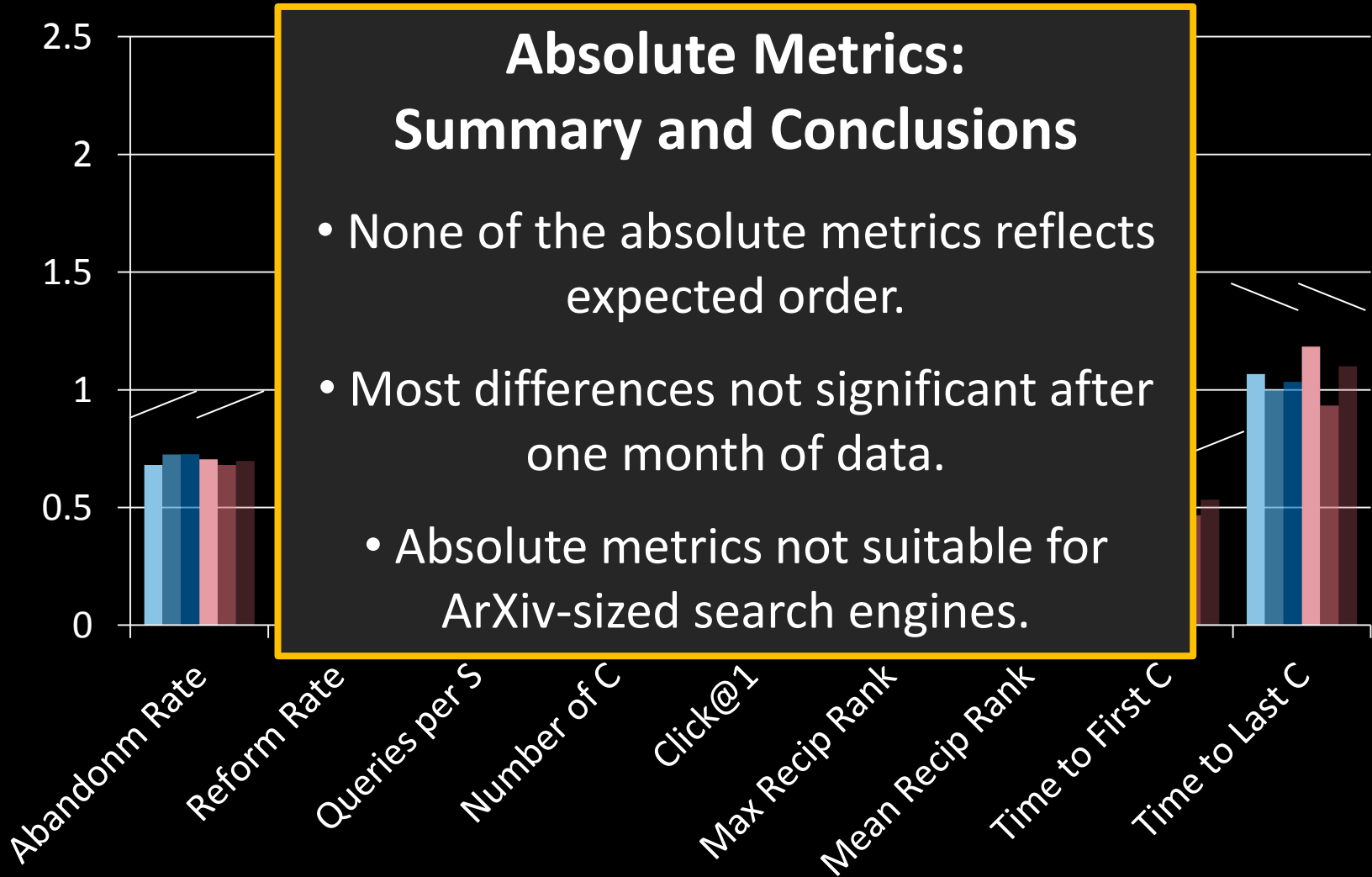
Displaying hits 1 to 10 of 622. [Reorder by date.](#)

- [Emmanuel Monfni, Yann Guemur, A Quadratic Loss Multi-Class SVM \(2008\)](#)
abstract: ... on the leave-one-out error of the pattern recognition SVM have been derived. Among these bounds, the most popular one ... bound. It applies to the hard margin pattern recognition SVM, and by extension to the 2-norm SVM. In this report, we introduce a quadratic loss M-SVM, the M-SVM², as a di ...
<http://arxiv.org/abs/0804.4898>
- [Nathalie Villa, Fabrice Rossi, Un résultat de consistance pour des SVM fonctionnels par interpolation spline \(2007\)](#)
abstract: ... for function classification with Support Vector Machine (SVM). Rather than relying on projection on a truncated ... an implicit spline interpolation that allows us to compute SVM on the derivatives of the studied functions. To that end, w ...
<http://arxiv.org/abs/0705.0210>
- [François Rapoport, Emmanuel Berthet and Jean-Philippe Vert, Classification of arrayCGH data using a fused SVM \(2008\)](#)
abstract: a new method for supervised classification of arrayCGH data. The method is a variant of support vector machine (SVM) that incorporates the biological specificities of DNA copy number variations along the genome as prior knowledge. The ...
<http://arxiv.org/abs/0801.3007>
- [Seonho Wu, Hui Zou, Ming Yuan, Structure variable selection in support vector machines \(2007\)](#)
abstract: When applying the support vector machine (SVM) to high-dimensional classification problems, we often impose a sparse structure in the SVM to eliminate the influences of the irrelevant predictors. ... selection techniques have been successfully used in the SVM to perform automatic variable selection ...
<http://arxiv.org/abs/0710.0508>
- [Marco Frullis, Oriana Mansutti, Praveen Boinse et al., A third level trigger programmable on FPGA for the gamma/hadron separation in a Cherenkov telescope using pseudo-Zernike moments and the SVM classifier \(2005\)](#)
abstract: ... computed Pseudo-Zernike features as classification parameters. We implemented on a FPGA board a kernel function of the SVM and the Pseudo-Zernike features to build a third level trigger for the gamma-hadron separation task of the MAGIC Expen ...
<http://arxiv.org/abs/cs/0602083>
- [Hao Helen Zhang, Yufeng Liu, Yichao Wu et al., Variable selection for the multiclass SVM via adaptive sup-norm regularization \(2008\)](#)
abstract: The Support Vector Machine (SVM) is a popular classification paradigm in machine learning ... great success in real applications. However, the standard SVM can not select variables ... of regularization in the context of the multiclass SVM (MSVM) for simultaneous classification and variable sel ...
<http://arxiv.org/abs/0803.3676>
- [Seung-chan Ahn, Gene Kim and MyungHo Kim, A Note on Applications of Support Vector Machine \(2001\)](#)
abstract: We describe in a rudimentary fashion how SVM (support vector machine) plays the role of classifier in a mathematical setting. We then discuss its application in the ...
<http://arxiv.org/abs/math/0105169>
- [Haoshen Li, J. W. Clerk, E. Mavrommatis et al., Modeling Nuclear Properties with Support Vector Machines \(2005\)](#)
abstract: ... studies of the potential of support vector machines (SVM) for providing statistical models of nuclear systematics with demonstrable predictive power. Using SVM regression and classification procedures, we have created ...
<http://arxiv.org/abs/nuc-th/0506080>
- [Gilles Blanchard, Olivier Bousquet, Pascal Massart, Statistical performance of support vector machines \(2008\)](#)
abstract: The support vector machine (SVM) algorithm is well known to the computer learning community ... builds on the observation made by other authors that the SVM can be viewed as a statistical regularization procedure. Fr ... how does it compare to the penalty actually used in the SVM algorithm; (2) is ...
<http://arxiv.org/abs/0804.0931>
- [Emidio Capriotti and Rita Casadio, The evaluation of protein folding rate constant is improved by predicting the folding kinetic order with a SVM-based method \(2006\)](#)
abstract: ... first we describe a support vector machine-based method (SVM-KO) to predict for a given protein the kinetic order of the ... value can be obtained as a linear regression task with a SVM-based method. In this paper we show that linear correlation ...
<http://arxiv.org/abs/q-bio.BM/0602013>

Absolute Metrics: Experiment Setup

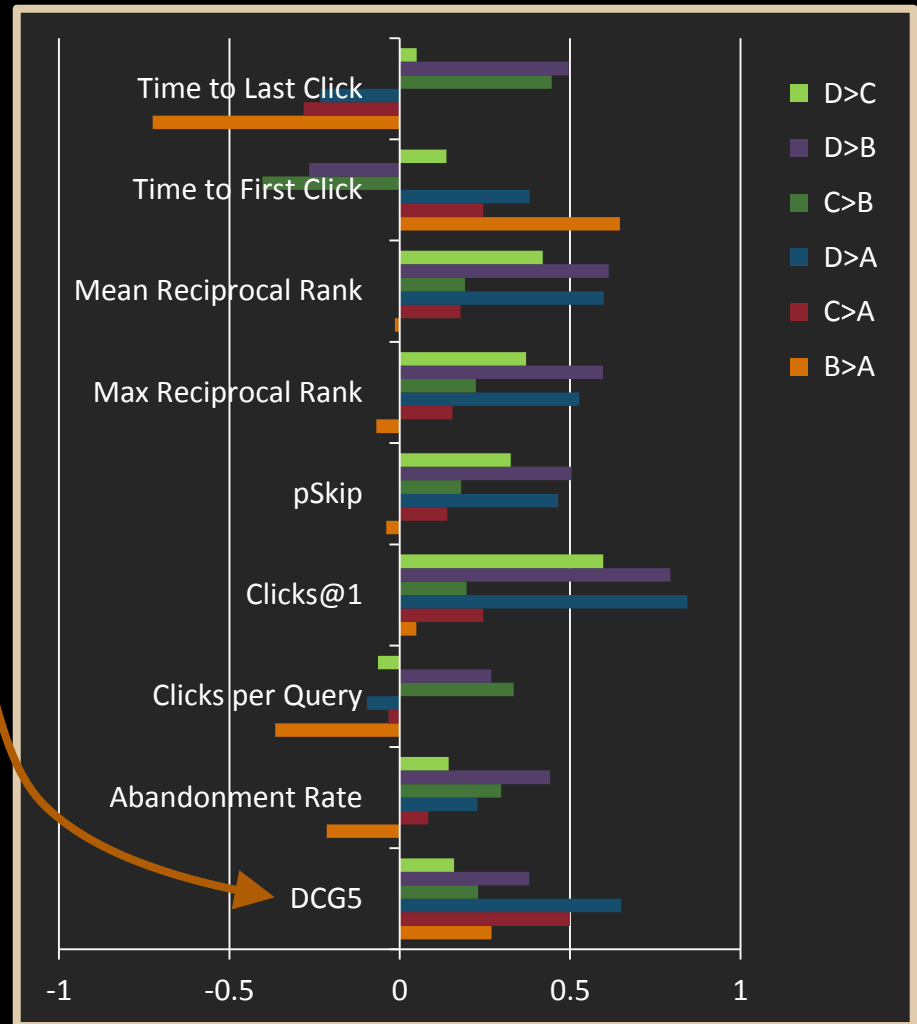
- Experiment Setup
 - Phase I: 36 days
 - Users randomly receive ranking from Orig, Flat, Rand
 - Phase II: 30 days
 - Users randomly receive ranking from Orig, Swap2, Swap4
 - User are permanently assigned to one experimental condition based on IP address and browser.
- Basic Statistics
 - ~700 queries per day / ~300 distinct users per day
- Quality Control and Data Cleaning
 - Test run for 32 days
 - Heuristics to identify bots and spammers
 - All evaluation code was written twice and cross-validated

Absolute Metrics: Results



Yahoo! Search: Results

- Retrieval Functions
 - 4 variants of production retrieval function
- Data
 - 10M – 70M queries for each retrieval function
 - Expert relevance judgments
- Results
 - Still not always significant even after more than 10M queries per function
 - Only Click@1 consistent with DCG@5.



Approaches to Utility Elicitation

- Approach 1: Absolute Metrics
 - Do metrics derived from observed user behavior provide absolute feedback about retrieval quality of f ?
 - For example:
 - $U(f) \sim \text{numClicks}(f)$
 - $U(f) \sim 1/\text{abandonment}(f)$

Approach 2: Paired Comparison Tests

- Do paired comparison tests provide relative preferences between two retrieval functions f_1 and f_2 ?
- For example:
 - $f_1 \succ f_2 \Leftrightarrow \text{pairedCompTest}(f_1, f_2) > 0$

Paired Comparisons: What to Measure?

$(u=tj, q=\text{"svm"})$

$f_1(u,q) \rightarrow r_1$

$f_2(u,q) \rightarrow r_2$

1. Kernel Machines
<http://svm.first.gmd.de/>
2. Support Vector Machine
<http://jbolivar.freesevers.com/>
3. An Introduction to Support Vector Machines
<http://www.support-vector.net/>
4. Archives of SUPPORT-VECTOR-MACHINES ...
<http://www.jiscmail.ac.uk/lists/SUPPORT...>
5. SVM-Light Support Vector Machine
http://ais.gmd.de/~thorsten/svm_light/

1. Kernel Machines
<http://svm.first.gmd.de/>
2. SVM-Light Support Vector Machine
http://ais.gmd.de/~thorsten/svm_light/
3. Support Vector Machine and Kernel ... References
<http://svm.research.bell-labs.com/SVMrefs.html>
4. Lucent Technologies: SVM demo applet
<http://svm.research.bell-labs.com/SVT/SVMsvt.html>
5. Royal Holloway Support Vector Machine
<http://svm.dcs.rhnc.ac.uk>

Interpretation: $(r_1 \succ r_2) \Leftrightarrow \text{clicks}(r_1) > \text{clicks}(r_2)$

Paired Comparison: Balanced Interleaving

$(u=tj, q=\text{"svm"})$

$f_1(u, q) \rightarrow r_1$

$f_2(u, q) \rightarrow r_2$

1. Kernel Machines
<http://svm.first.gmd.de/>
2. Support Vector Machine
<http://jbolivar.freesevers.com/>
3. An Introduction to Support Vector Machines
<http://www.support-vector.net/>
4. Archives of SUPPORT-VECTOR-MACHINES ...
<http://www.jiscmail.ac.uk/lists/SUPPORT...>
5. SVM-Light Support Vector Machine
http://ais.gmd.de/~thorsten/svm_light/

1. Kernel Machines
<http://svm.first.gmd.de/>
2. SVM-Light Support Vector Machine
http://ais.gmd.de/~thorsten/svm_light/
3. Support Vector Machine and Kernel ... References
<http://svm.research.bell-labs.com/SVMrefs.html>
4. Lucent Technologies: SVM demo applet
<http://svm.research.bell-labs.com/SVT/SVMsvt.html>
5. Royal Holloway Support Vector Machine
<http://svm.dcs.rhnc.ac.uk>

Interleaving(r_1, r_2)

- | | | |
|----|---|---|
| 1. | Kernel Machines | 1 |
| | http://svm.first.gmd.de/ | |
| 2. | Support Vector Machine | 2 |
| | http://jbolivar.freesevers.com/ | |
| 3. | SVM-Light Support Vector Machine | 2 |
| | http://ais.gmd.de/~thorsten/svm_light/ | |
| 4. | An Introduction to Support Vector Machines | 3 |
| | http://www.support-vector.net/ | |
| 5. | Support Vector Machine and Kernel ... References | 3 |
| | http://svm.research.bell-labs.com/SVMrefs.html | |
| 6. | Archives of SUPPORT-VECTOR-MACHINES ... | 4 |
| | http://www.jiscmail.ac.uk/lists/SUPPORT... | |
| 7. | Lucent Technologies: SVM demo applet | 4 |
| | http://svm.research.bell-labs.com/SVT/SVMsvt.html | |

Model of User:

Better retrieval functions
is more likely to get more
clicks.

Invariant:

For all k , top k of
balanced interleaving is
union of top k_1 of r_1 and
top k_2 of r_2 with $k_1 = k_2 \pm 1$.

Interpretation: $(r_1 \succ r_2) \Leftrightarrow \text{clicks}(\text{topk}(r_1)) > \text{clicks}(\text{topk}(r_2))$

\rightarrow see also [Radlinski, Craswell, 2012] [Hofmann, 2012]

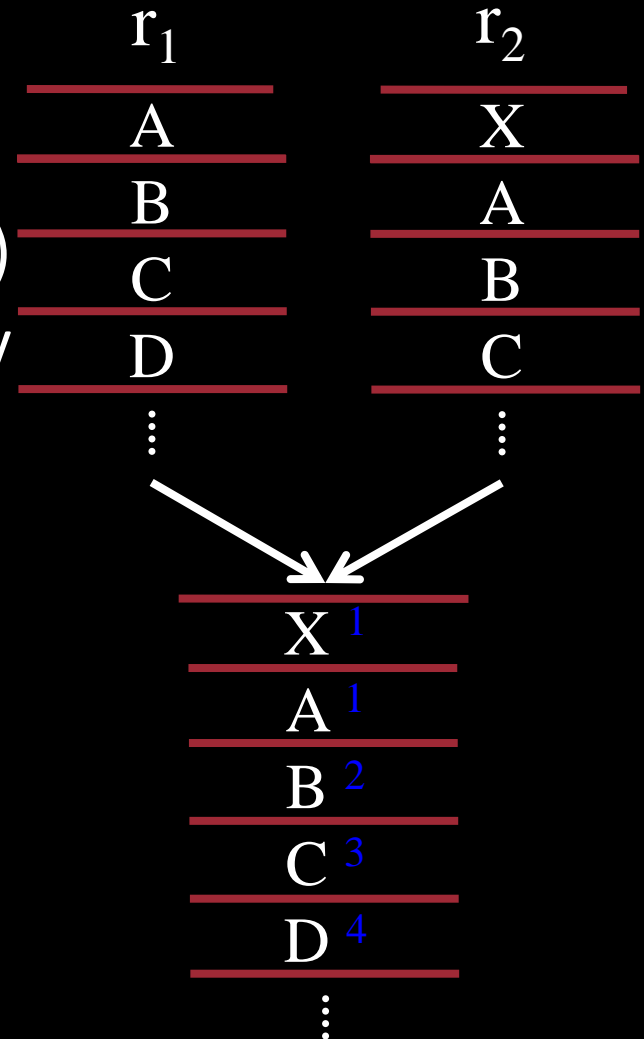
Balanced Interleaving: a Problem

- Example:

- Two rankings r_1 and r_2 that are identical up to one insertion (X)
- “Random user” clicks uniformly on results in interleaved ranking

1. “X” $\rightarrow r_2$ wins
2. “A” $\rightarrow r_1$ wins
3. “B” $\rightarrow r_1$ wins
4. “C” $\rightarrow r_1$ wins
5. “D” $\rightarrow r_1$ wins

\rightarrow biased



Paired Comparisons: Team-Game Interleaving

($u=tj, q="svm"$)

$f_1(u, q) \rightarrow r_1$

$f_2(u, q) \rightarrow r_2$

NEXT
PICK

1. Kernel Machines
~~<http://svm.first.gmd.de/>~~
2. Support Vector Machine
~~<http://jbolivar.freesevers.com/>~~
3. An Introduction to Support Vector Machines
~~<http://www.support-vector.net/>~~
4. Archives of SUPPORT-VECTOR-MACHINES ...
~~<http://www.jiscmail.ac.uk/lists/SUPPORT...>~~
5. SVM-Light Support Vector Machine
~~<http://ais.gmd.de/~thorsten/svm/light/>~~

1. Kernel Machines
~~<http://svm.first.gmd.de/>~~
2. SVM-Light Support Vector Machine
~~<http://ais.gmd.de/~thorsten/svm/light/>~~
3. Support Vector Machine and Kernel ... References
~~<http://svm.research.bell-labs.com/SVMrefs.html>~~
4. Lucent Technologies: SVM demo applet
~~<http://svm.research.bell-labs.com/SVT/SVMsvt.html>~~
5. Royal Holloway Support Vector Machine

Interleaving(r_1, r_2)

- | | | |
|----|--|----|
| 1. | Kernel Machines | T2 |
| | http://svm.first.gmd.de/ | |
| 2. | Support Vector Machine | T1 |
| | http://jbolivar.freesevers.com/ | |
| 3. | SVM-Light Support Vector Machine | T2 |
| | http://ais.gmd.de/~thorsten/svm/light/ | |
| 4. | An Introduction to Support Vector Machines | T1 |
| | http://www.support-vector.net/ | |
| 5. | Support Vector Machine and Kernel ... References | T2 |
| | http://svm.research.bell-labs.com/SVMrefs.html | |
| 6. | Archives of SUPPORT-VECTOR-MACHINES ... | T1 |
| | http://www.jiscmail.ac.uk/lists/SUPPORT... | |
| 7. | Lucent Technologies: SVM demo applet | T2 |
| | http://svm.research.bell-labs.com/SVT/SVMsvt.html | |

Invariant:

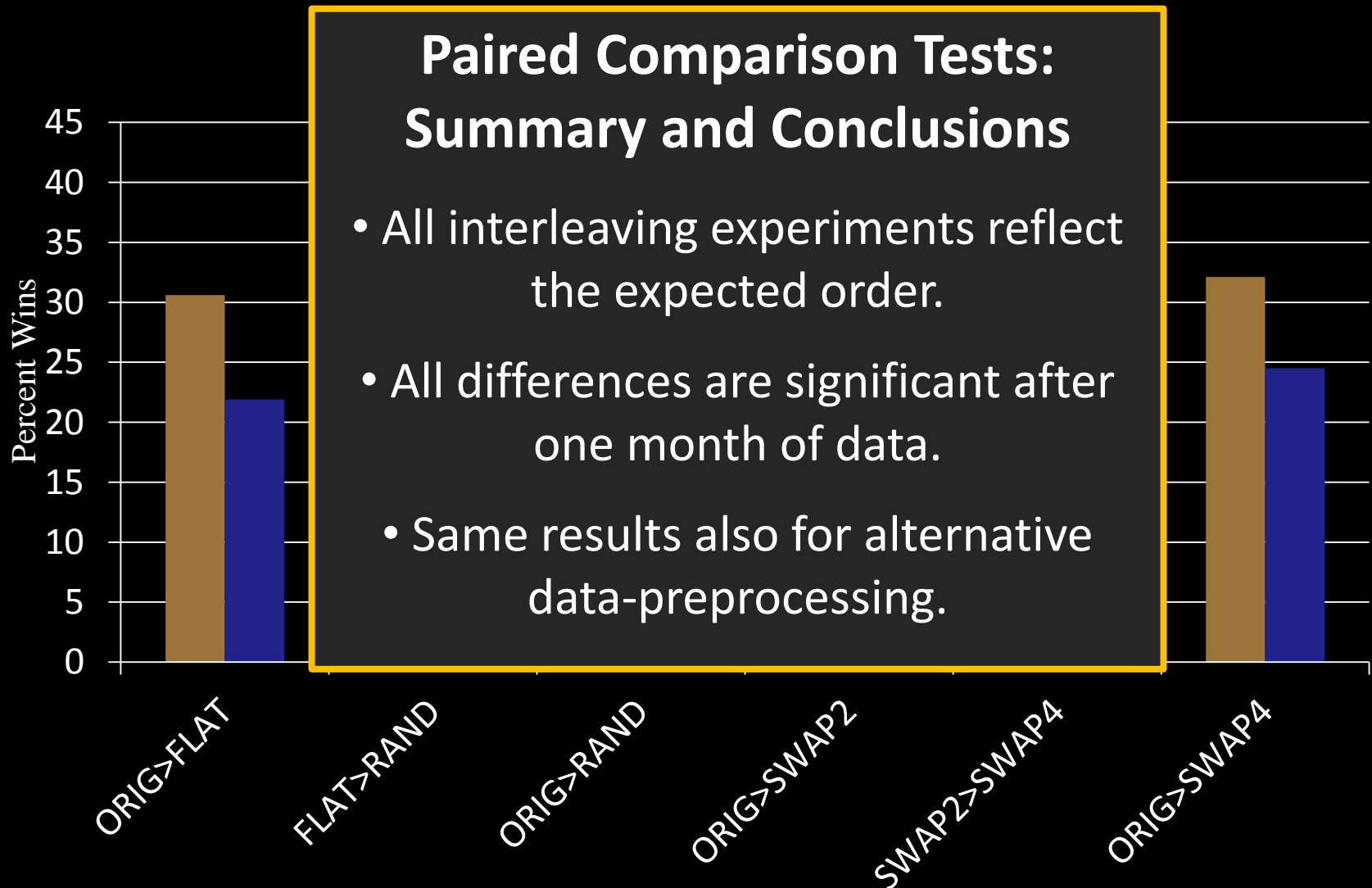
For all k , in expectation
same number of team
members in top k from
each team.

Interpretation: ($r_1 \succ r_2$) \Leftrightarrow clicks(T_1) > clicks(T_2)

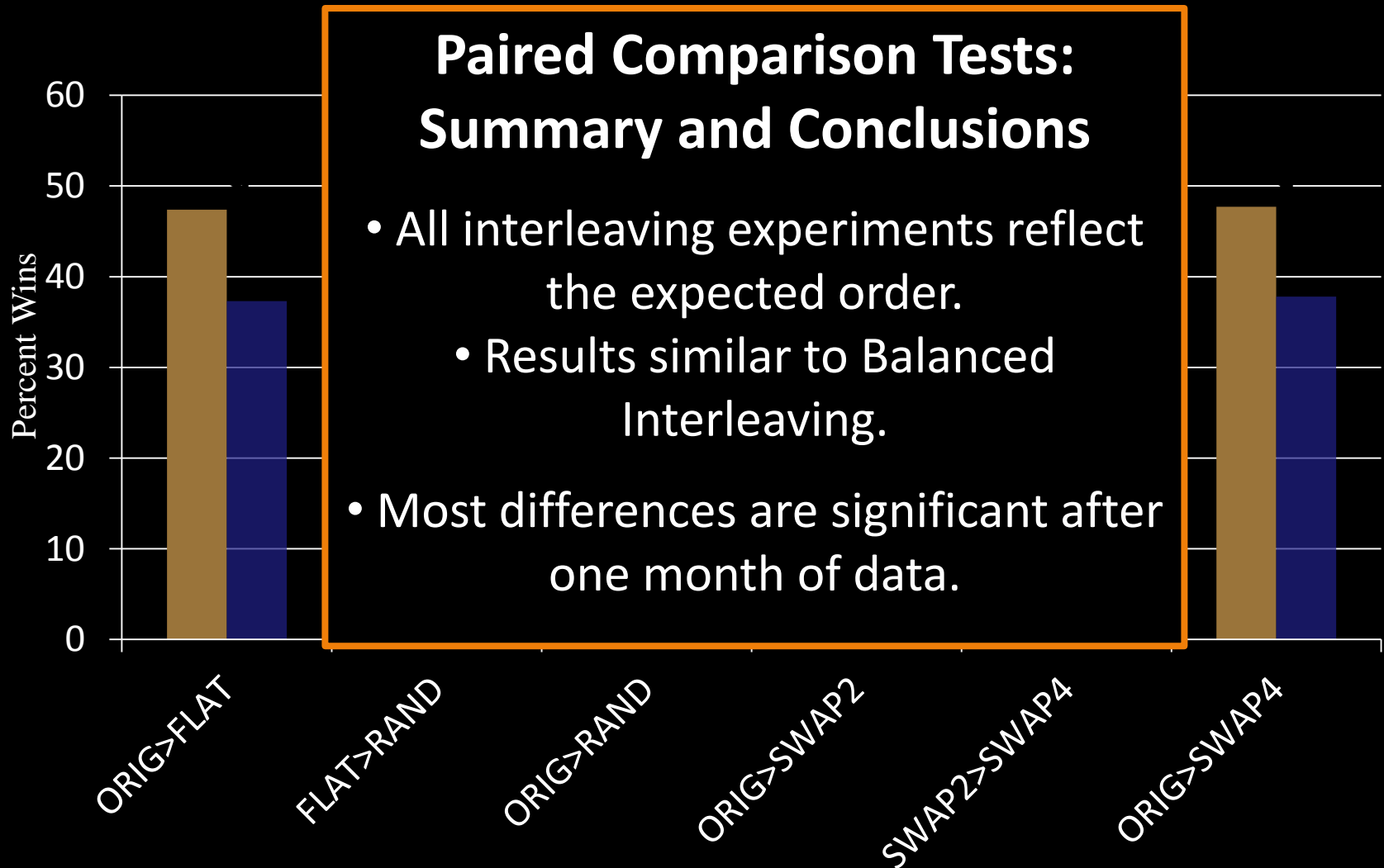
Paired Comparisons: Experiment Setup

- Experiment Setup
 - Phase I: 36 days
 - Balanced Interleaving of (Orig,Flat) (Flat,Rand) (Orig,Rand)
 - Phase II: 30 days
 - Balanced Interleaving of (Orig,Swap2) (Swap2,Swap4) (Orig,Swap4)
- Quality Control and Data Cleaning
 - Same as for absolute metrics

Balanced Interleaving: Results



Team-Game Interleaving: Results

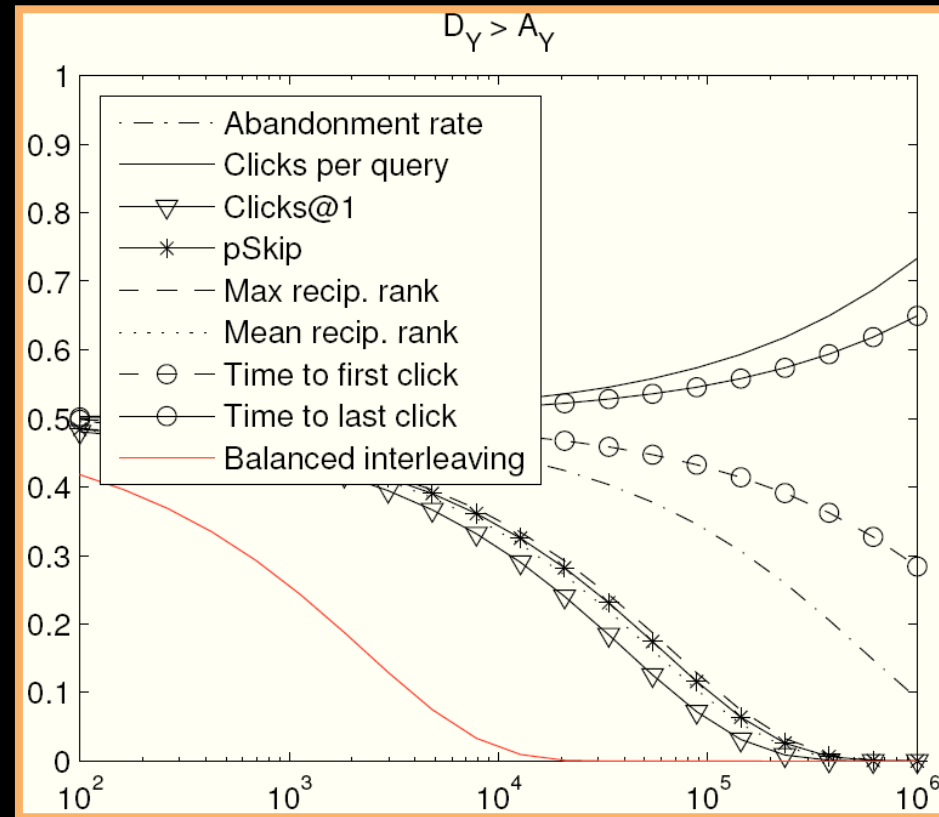


Yahoo and Bing: Interleaving Results

- Yahoo Web Search [Chapelle et al., 2012]
 - Four retrieval functions (i.e. 6 paired comparisons)
 - Balanced Interleaving
 - All paired comparisons consistent with ordering by NDCG.
- Bing Web Search [Radlinski & Craswell, 2010]
 - Five retrieval function pairs
 - Team-Game Interleaving
 - Consistent with ordering by NDGC when NDCG significant.

Efficiency: Interleaving vs. Absolute

- Yahoo Web Search
 - More than 10M queries for absolute measures
 - Approx 700k queries for interleaving
- Experiment
 - REPEAT
 - Draw bootstrap sample S of size x
 - Evaluate metric on S for pair (P, Q) of retrieval functions
 - Estimate $\gamma = P(P >_m Q | x)$



➔ Interleaving by factor ~ 10 more efficient than Click@1.

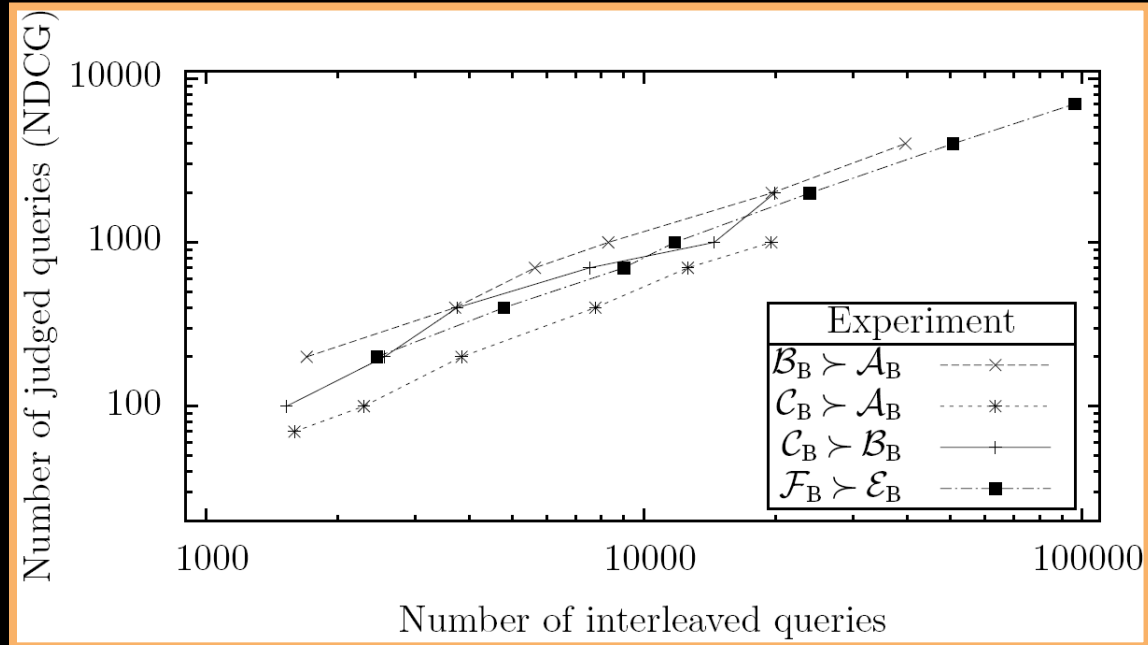
Efficiency: Interleaving vs. Explicit

- Bing Web Search
 - 4 retrieval function pairs
 - ~12k manually judged queries
 - ~200k interleaved queries

- Experiment

- p = probability that NDCG is correct on subsample of size y
- x = number of queries needed to reach same p -value with interleaving

➔ Ten interleaved queries are equivalent to one manually judged query.



Summary and Conclusions

- Interleaving agrees better with expert assessment than absolute metrics
 - Design as pairwise comparison
- All interleaving techniques seem to do roughly equally well
- Efficiency of interleaving compared to expert assessment and Click@1