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

(tj,"SVM")

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/

$U(tj,"SVM",r_1)$

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.freeservers.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/

$U(tj,"SVM",r_2)$

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) ~ numClicks(f)
    - U(f) ~ 1/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	Aggre-gation	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

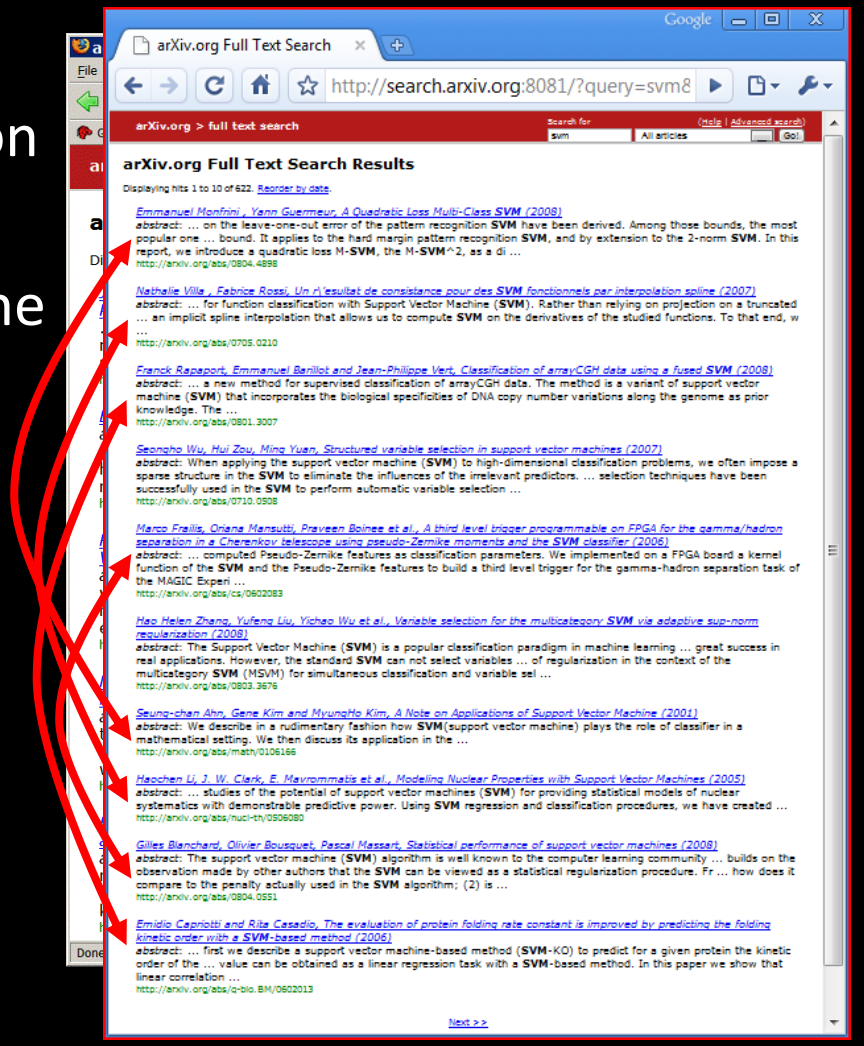
- Natural user and query population
- User in natural context, not lab
- Live and operational search engine
- Ground truth by construction

ORIG ≻ SWAP2 ≻ SWAP4

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

ORIG ≻ FLAT ≻ RAND

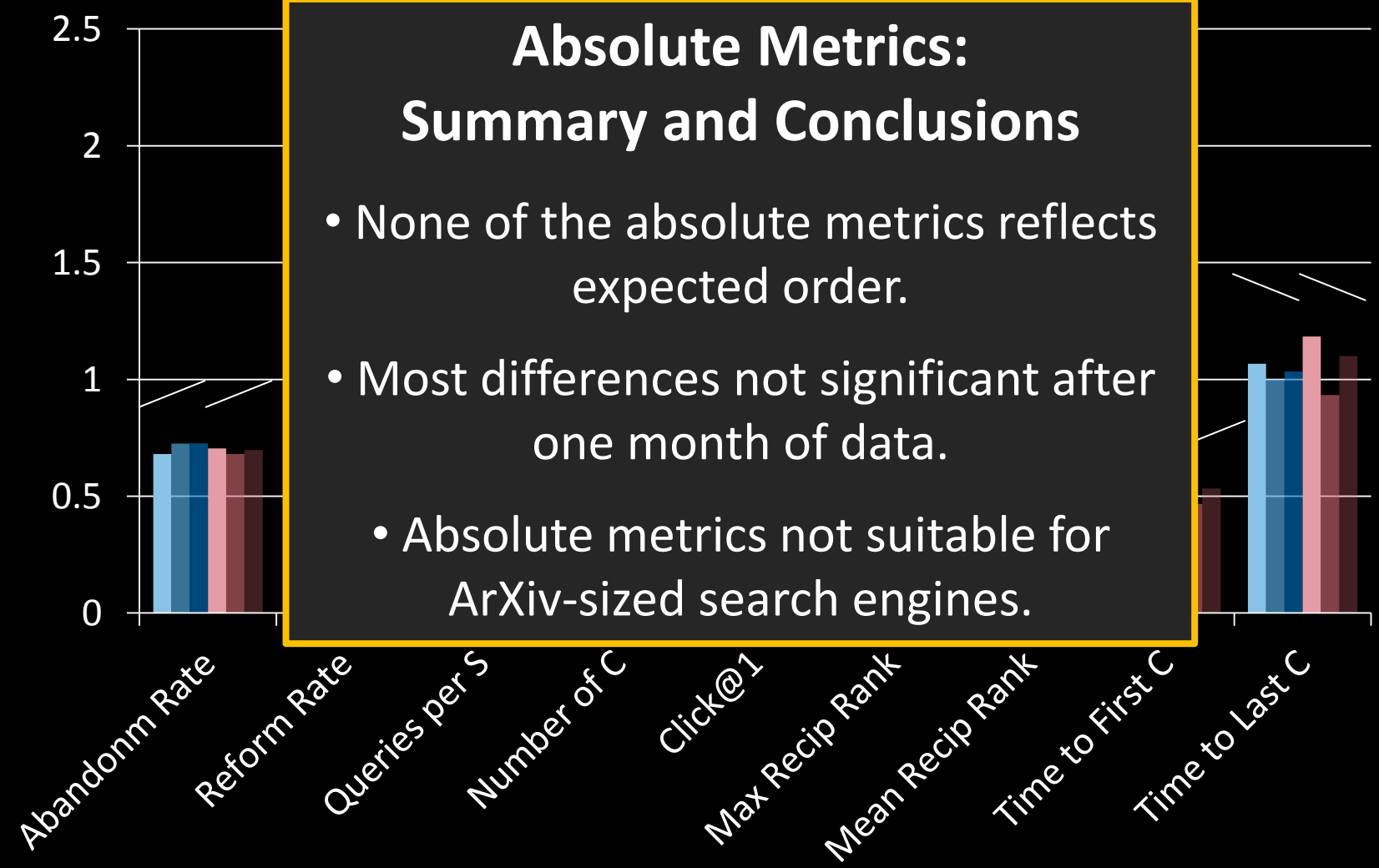
- ORIG: Hand-tuned fielded
- FLAT: No field weights
- RAND : Top 10 of FLAT shuffled



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

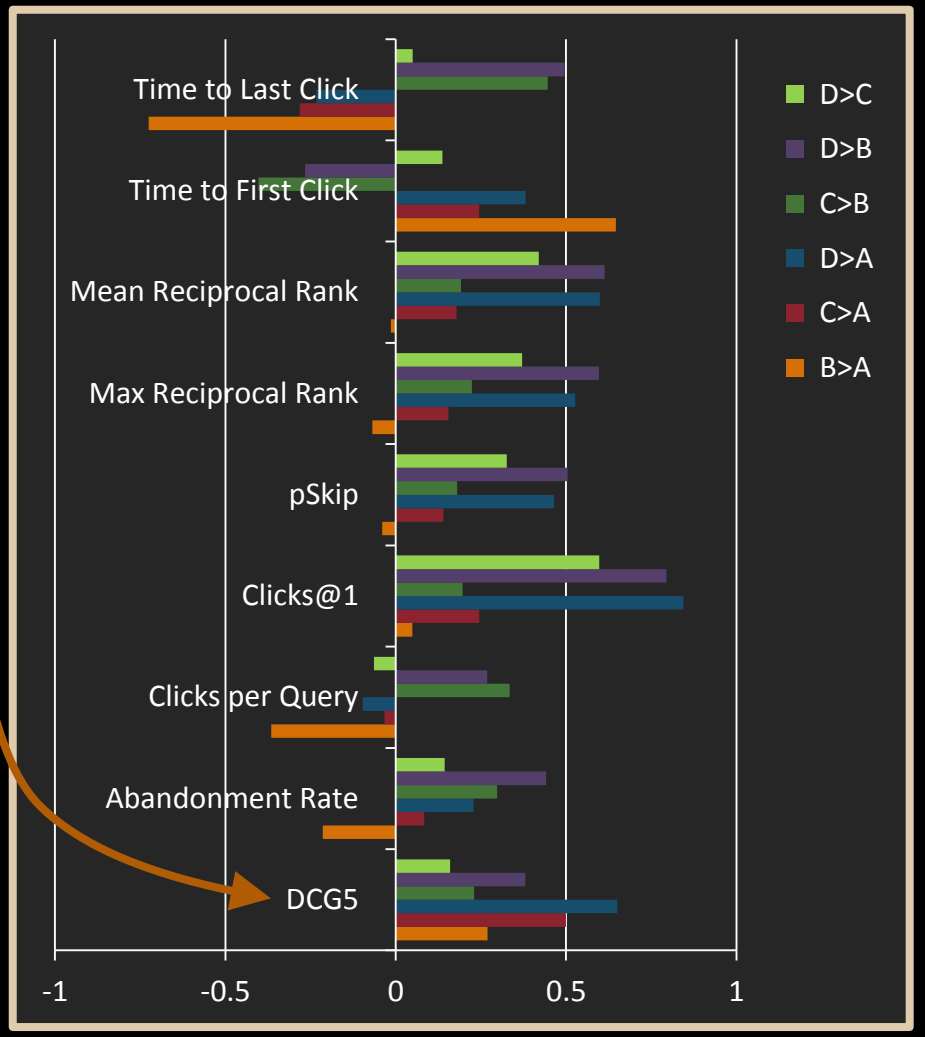


Absolute Metrics: Summary and Conclusions

- None of the absolute metrics reflects expected order.
- Most differences not significant after one month of data.
- Absolute metrics not suitable for ArXiv-sized search engines.

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.



[Chapelle et al., 2012]

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) ~ numClicks(f)
    - U(f) ~ 1/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="svm")

$f_1(u,q) \rightarrow r_1$

1. Kernel Machines
   http://svm.first.gmd.de/
2. Support Vector Machine
   http://jbolivar.freeservers.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/

$f_2(u,q) \rightarrow r_2$

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.rhbnc.ac.uk

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

Paired Comparison: Balanced Interleaving

(u=tj, q="svm")

$f_1(u,q) \rightarrow r_1$

1. Kernel Machines
   http://svm.first.gmd.de/
2. Support Vector Machine
   http://jbolivar.freeservers.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/

$f_2(u,q) \rightarrow r_2$

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.rhbnc.ac.uk

Interleaving(r_1,r_2)

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

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: $(r_1 \succ r_2) \leftrightarrow$ clicks(topk(r_1)) > clicks(topk(r_2))

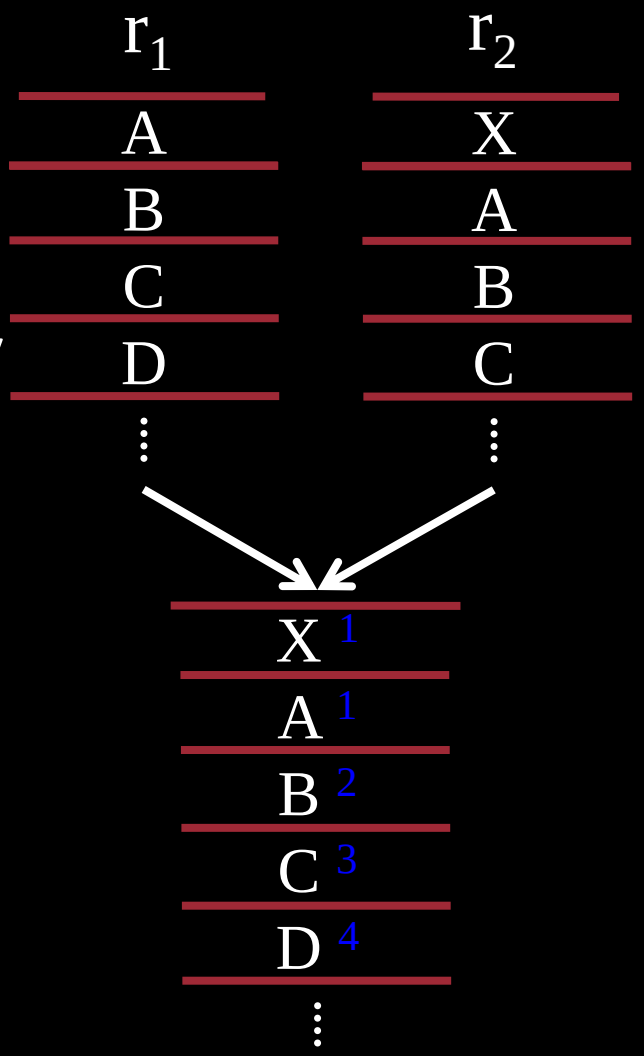
→ see also [Radlinski, Craswell, 2012] [Hofmann, 2012]

[Joachims, 2001] [Radlinski et al., 2008]

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" → r_2 wins
    2. "A" → r_1 wins
    3. "B" → r_1 wins
    4. "C" → r_1 wins
    5. "D" → r_1 wins

  → biased



Paired Comparisons: Team-Game Interleaving

(u=tj,q="svm")

$f_1(u,q) \rightarrow r_1$

1. Kernel Machines
   http://svm.first.gmd.de/
2. Support Vector Machine
   http://jbolivar.freeservers.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/

NEXT PICK

$f_2(u,q) \rightarrow r_2$

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

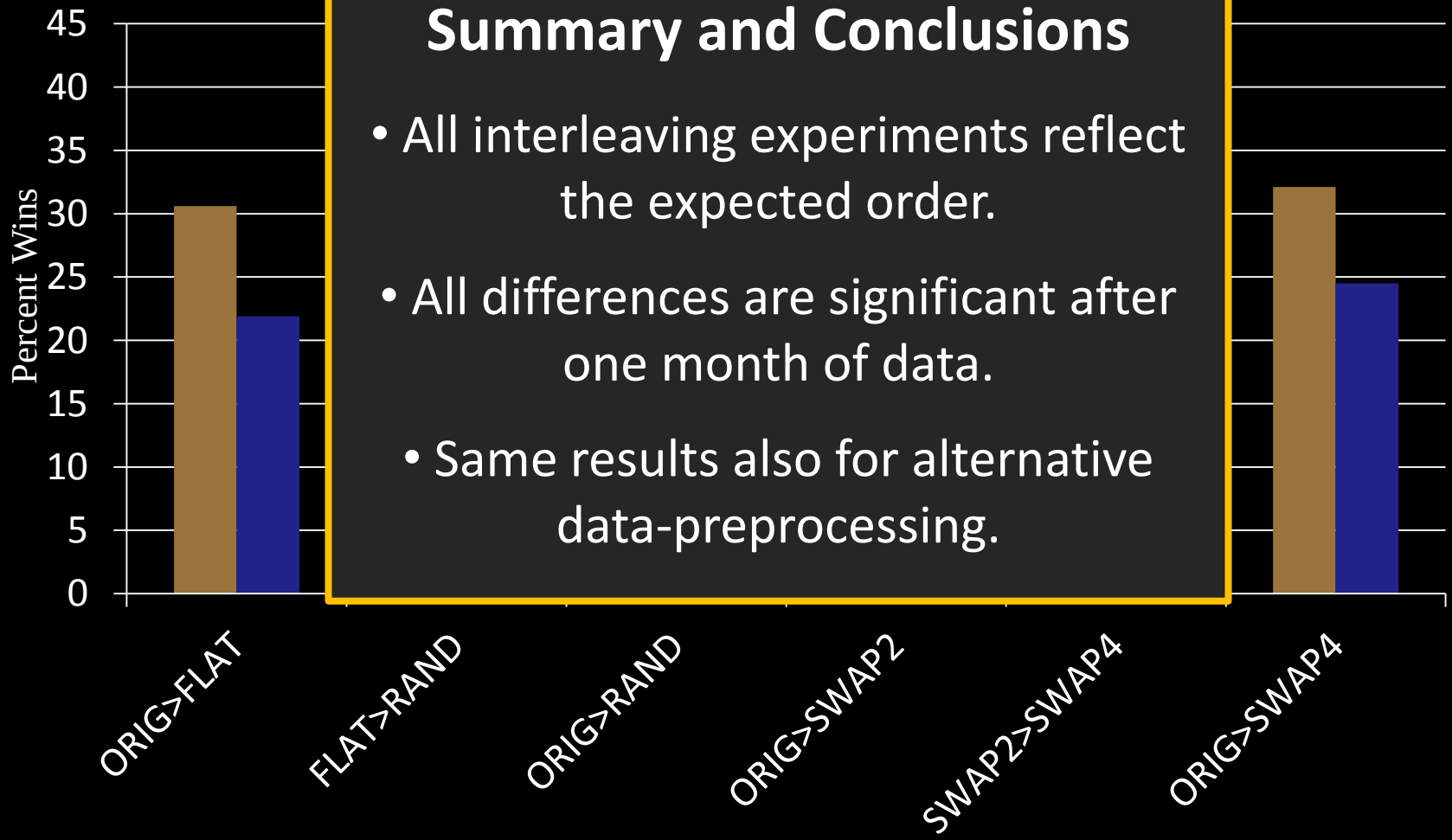
Invariant:
For all k, in expectation same number of team members in top k from each team.

Interpretation: $(r_1 \succ r_2) \leftrightarrow \text{clicks}(T_1) > \text{clicks}(T_2)$

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

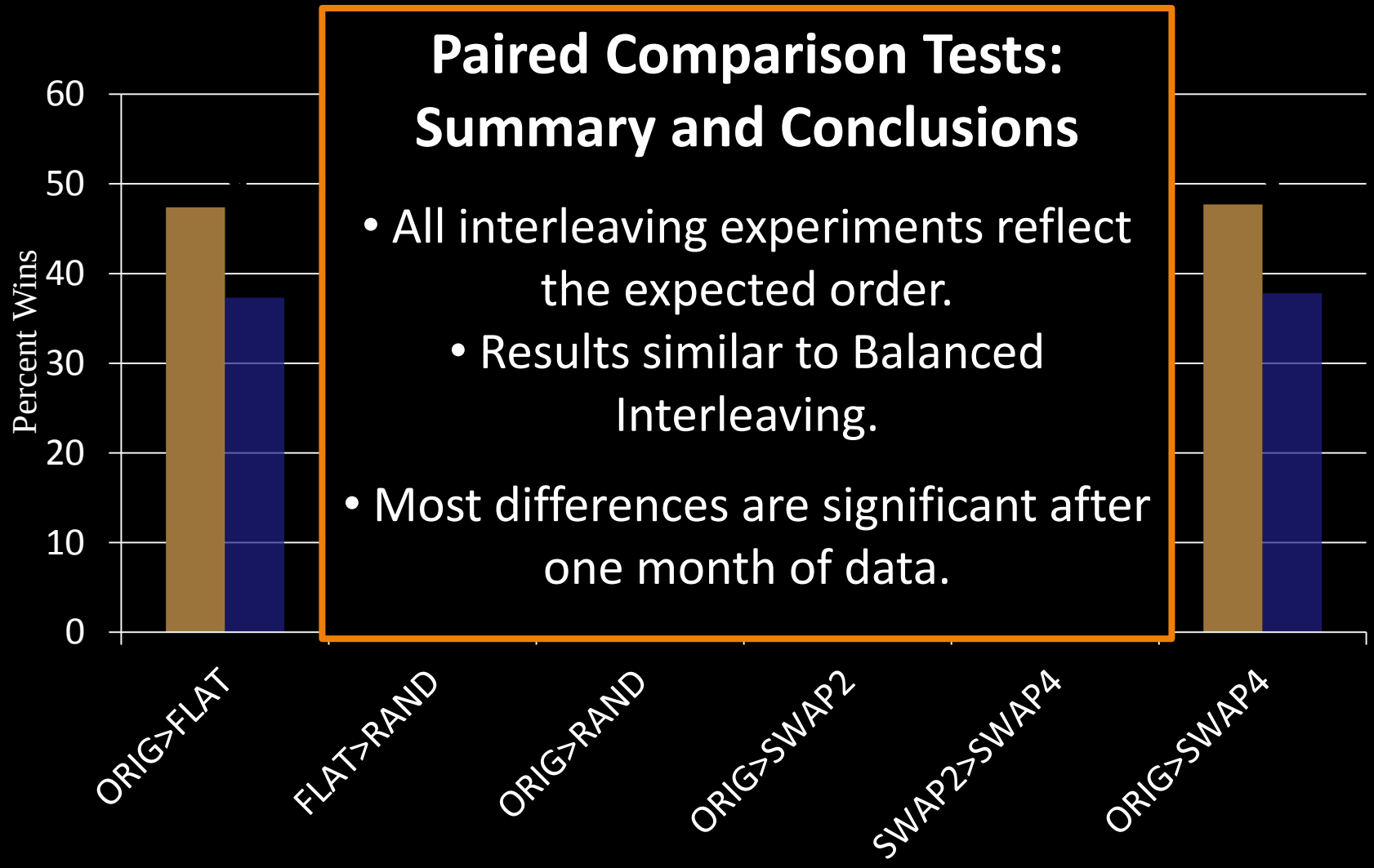
Balanced Interleaving: Results



Paired Comparison Tests: Summary and Conclusions

- All interleaving experiments reflect the expected order.
- All differences are significant after one month of data.
- Same results also for alternative data-preprocessing.

Team-Game Interleaving: Results



Paired Comparison Tests: Summary and Conclusions

- All interleaving experiments reflect the expected order.
- Results similar to Balanced Interleaving.
- Most differences are significant after one month of data.

Yahoo and Bing: Interleaving Results

- Yahoo Web Search [Chapelle et al., 2012]
  - Four retrieval functions (i.e. 6 paired comparisons)
  - Balanced Interleaving

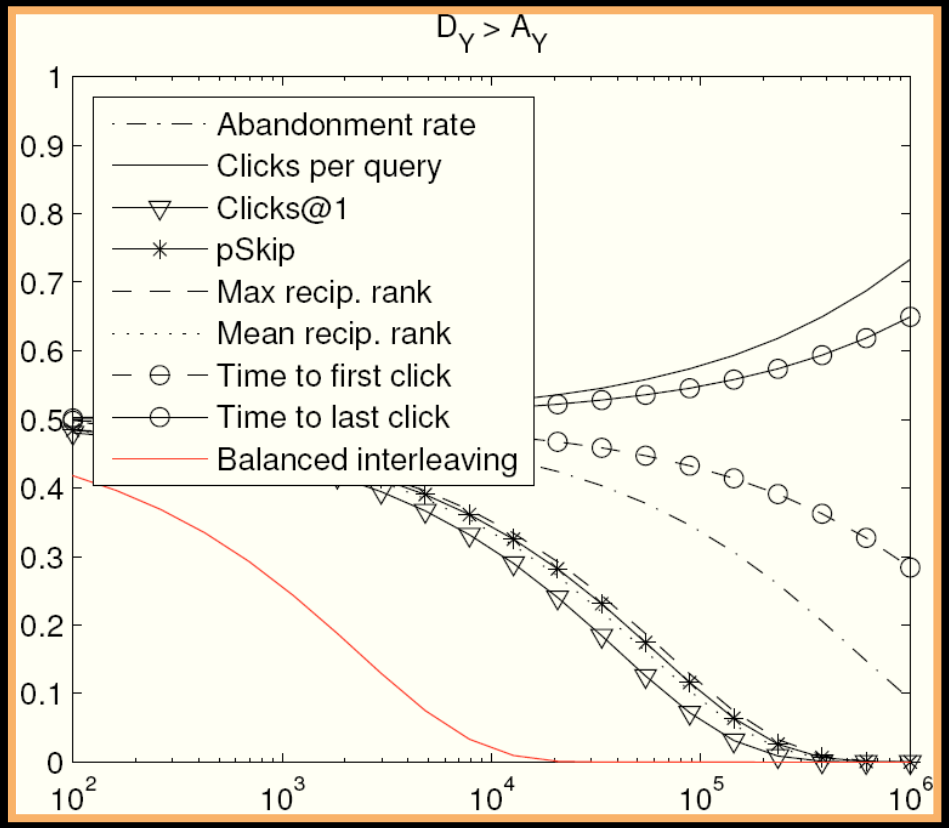
    → All paired comparisons consistent with ordering by NDCG.

- 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 $y = P(P >_m Q | x)$



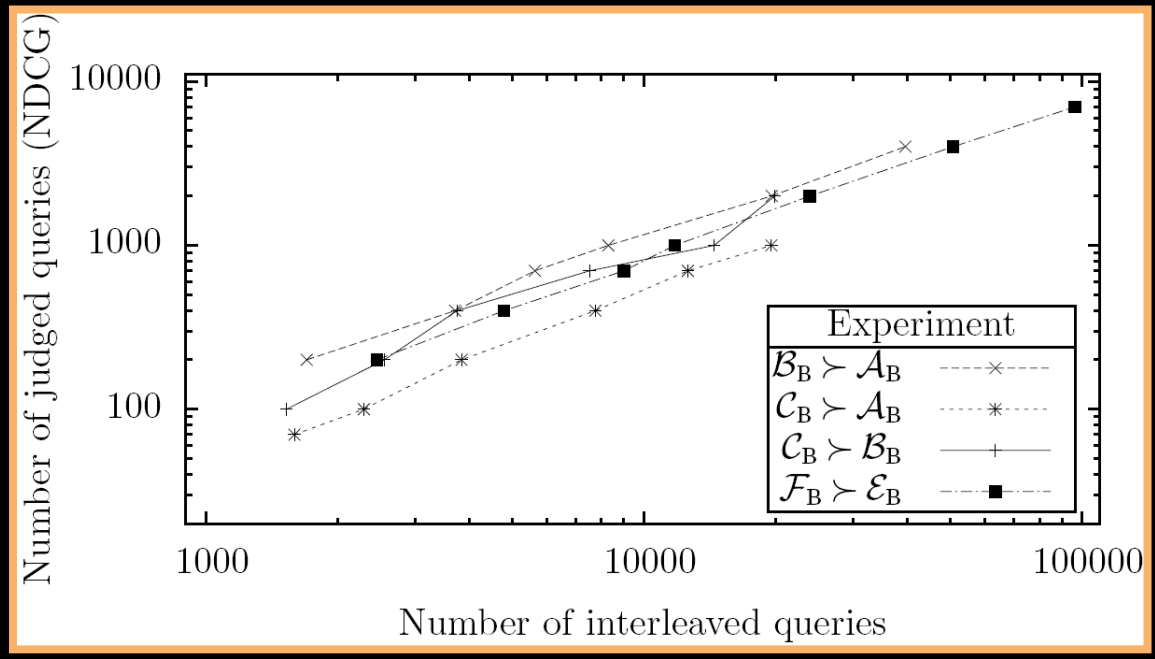
➔ Interleaving by factor ~10 more efficient than Click@1.

[Chapelle, Joachims, Radlinski, Yue, to appear]

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.



[Radlinski & Craswell, 2010]

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