Preference, Consensus, and Choice in Crowdsourced Relevance

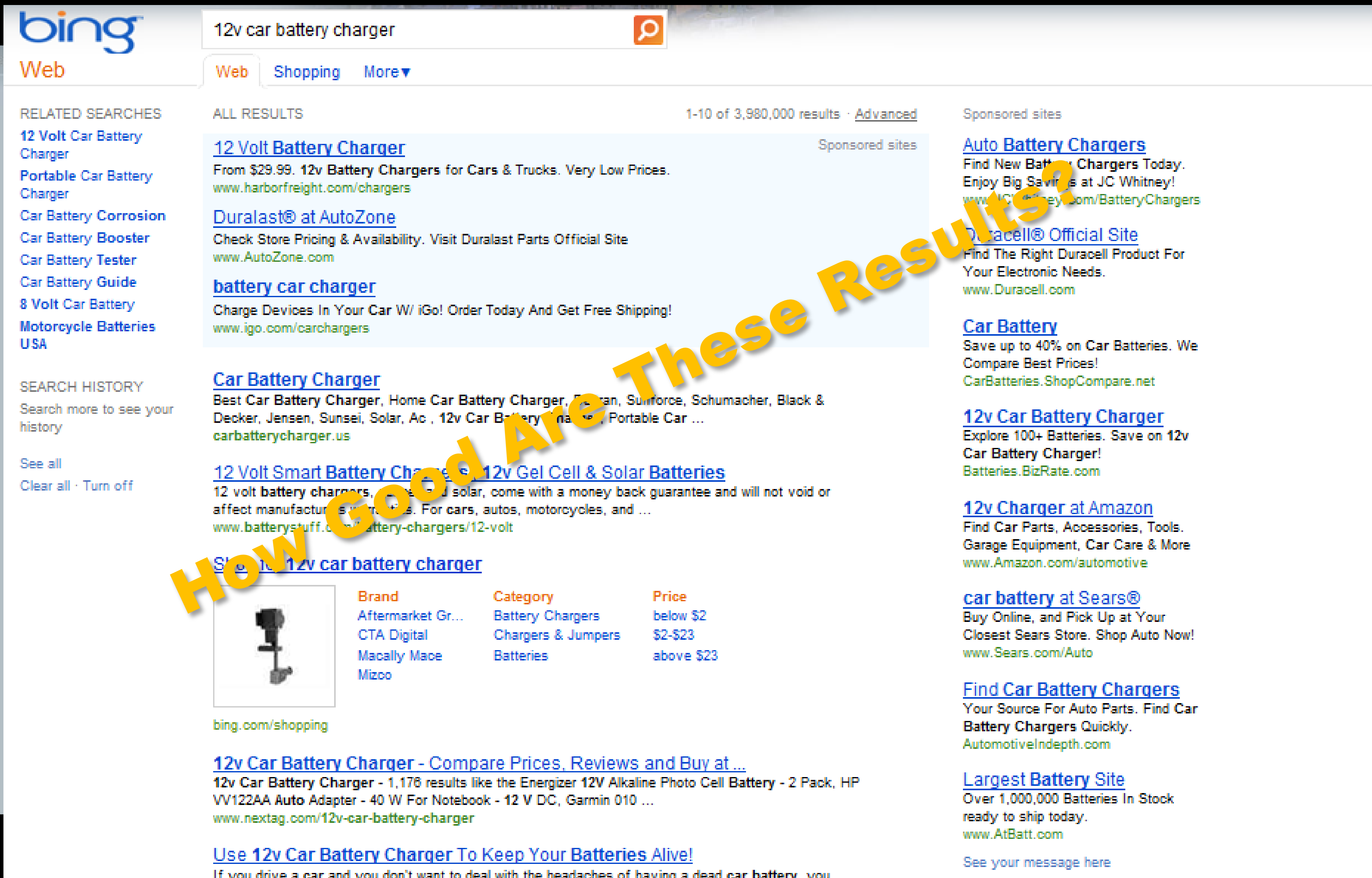
Paul Bennett, Microsoft Research (CLUES)

Joint work with Ben Carterette, Max Chickering, Susan Dumais, Eric Horvitz, Edith Law, and Anton Mityagin.

Road Map

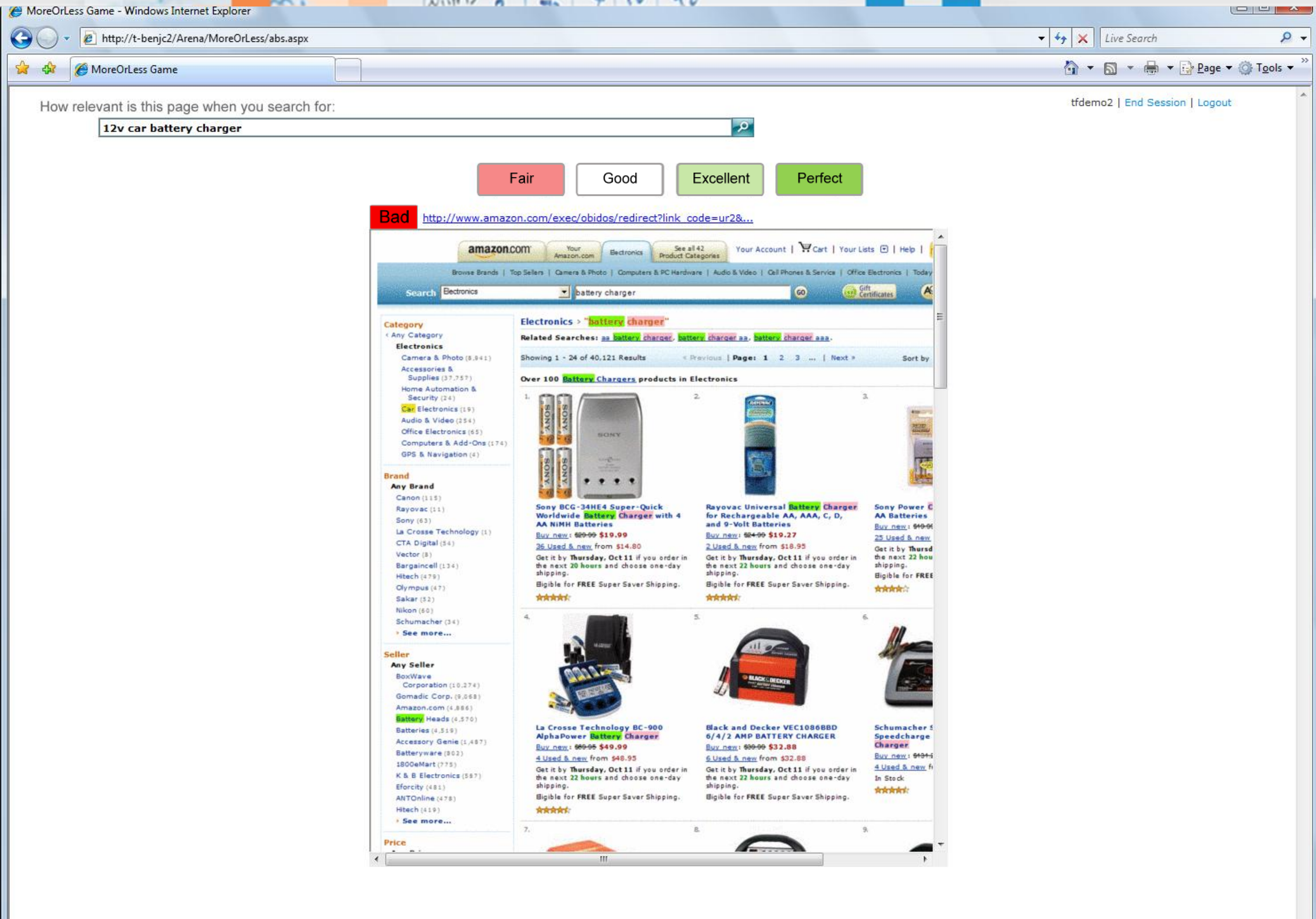
- Why Preferences?
- Learning Consensus from Preferences at Scale
- Beyond Consensus
- Discussion

How Good Are These Results?



Why relevance judgments?

- Used widely in search and advertising to:
  - Train ranking algorithms
  - Measure progress
    - Which system to deploy
      - Is the system better than it previously was?
      - Is it better than alternative models?
      - Is that difference significant?
  - Assess performance against competitors
  - Identify which components of the system need improvement


MoreOrLess Game - Windows Internet Explorer
http://t-benjc2/Arena/MoreOrLess/abs.aspx
Live Search
MoreOrLess Game
Page
Tools
How relevant is this page when you search for:
12v car battery charger
tfdemo2 | End Session | Logout
Fair
Good
Excellent
Perfect
Bad
http://www.amazon.com/exec/obidos/redirect?link_code=ur2&...
amazon.com
Electronics
Your Account | Cart | Your Lists | Help
Search
battery charger
Electronics > "battery charger"
Related Searches: aa battery charger, battery charger aa, battery charger aaa.
Showing 1 - 24 of 40,121 Results
Over 100 Battery Chargers products in Electronics
Category
Any Category
Electronics
Brand
Any Brand
Seller
Any Seller
Price
Sony BCG-34HE4 Super-Quick Worldwide Battery Charger with 4 AA NiMH Batteries
Rayovac Universal Battery Charger for Rechargeable AA, AAA, C, D, and 9-Volt Batteries
La Crosse Technology BC-900 AlphaPower Battery Charger
Black and Decker VEC1086BBD 6/4/2 AMP BATTERY CHARGER

Microsoft Research

From Judgments to Performance

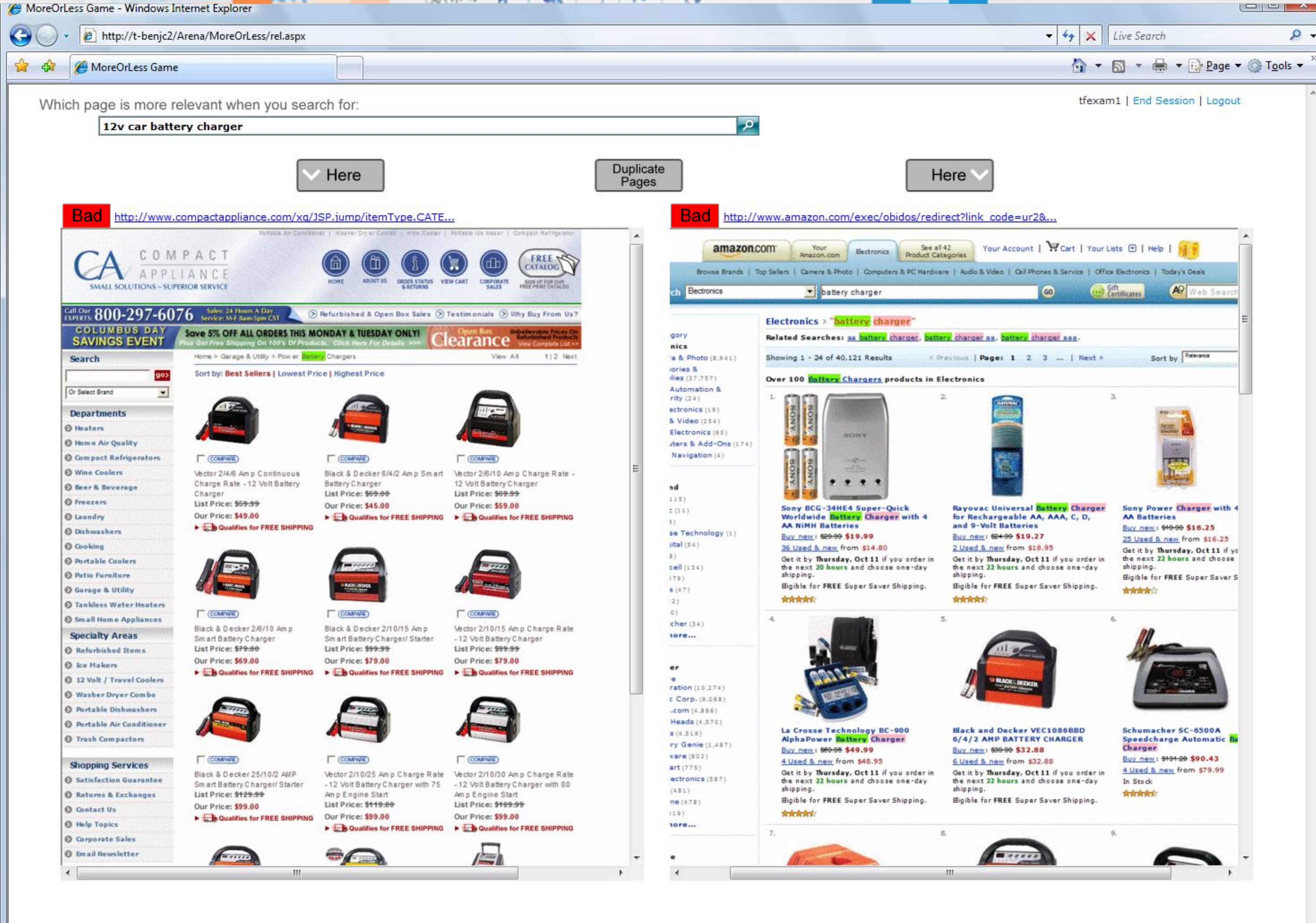
Rank for two systems			Get five-point rel.			Convert to "Gain"			Weight by rank		
1	A	D	1	5	4	1	31	15	1.00	31	15
2	F	B	2	4	5	2	15	31	0.63	15	31
3	B	E	3	5	3	3	31	7	0.50	31	7
4	H	H	4	2	2	4	3	3	0.43	3	3
5	D	A	5	4	5	5	15	31	0.39	15	31
6	E	F	6	3	4	6	7	15	0.36	7	15
7	C	G	7	1	1	7	0	0	0.33	0	0
8	G	C	8	1	1	8	0	0	0.32	0	0

Multiply and Add to get DCG
System 1: *65.55*
System 2: *56.69*

Divide by ideal (68.30) to get NDCG/100
System 1: *96.0*
System 2: *83.0*

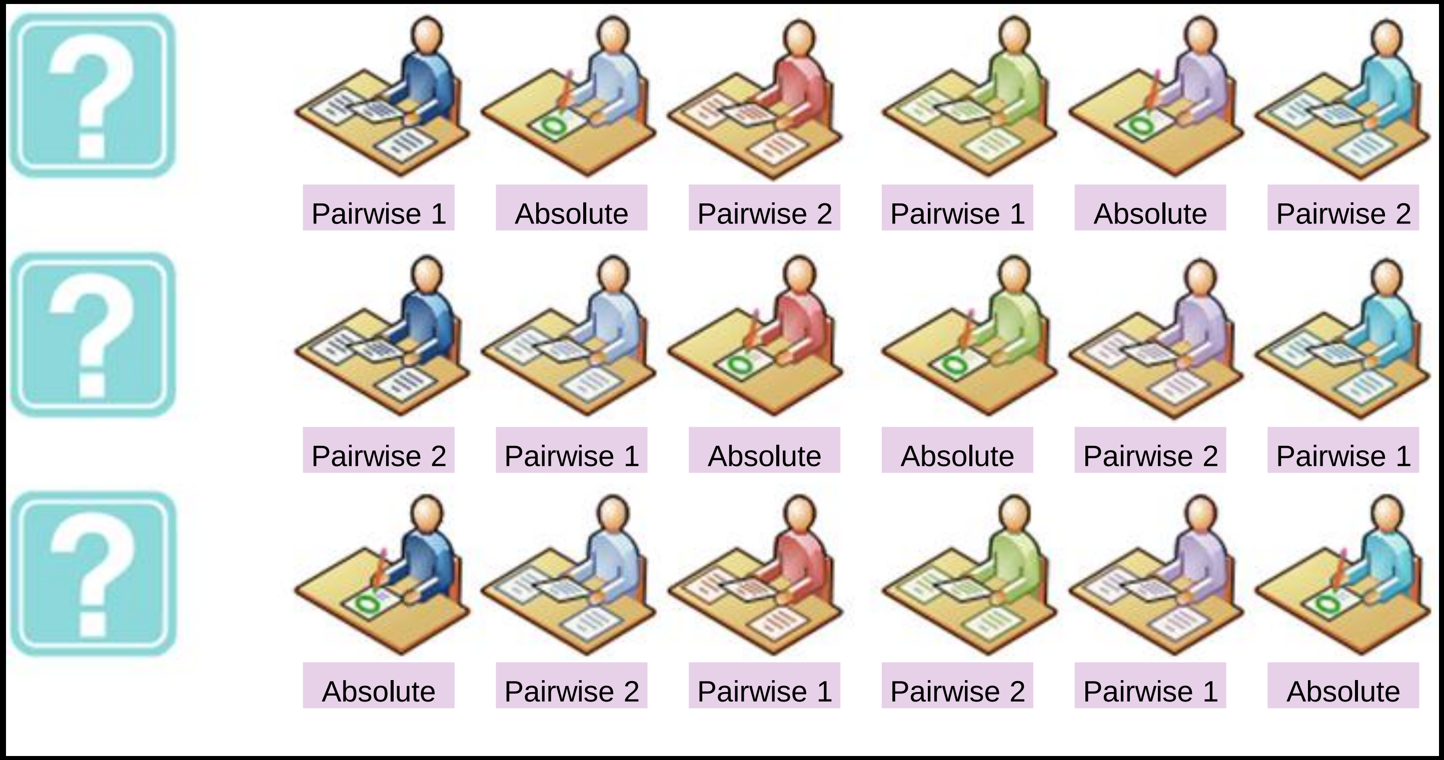
What's wrong with absolute?

- High Variability in Judged Relevance of Page
  - Can lead to incorrect conclusions of system improvement (and deployment of wrong version).
  - Can cause incorrect assessment of performance relative to competitors.
  - Introduces noise in training data for ranker.
- Coarser-grained distinctions
  - May mask real discernible differences in page quality.


MoreOrLess Game - Windows Internet Explorer
http://t-benjc2/Arena/MoreOrLess/rel.aspx
MoreOrLess Game
Which page is more relevant when you search for:
12v car battery charger
tfexam1 | End Session | Logout
Here
Duplicate Pages
Here
Bad
http://www.compactappliance.com/xq/JSP.jump/itemType.CATE...
Bad
http://www.amazon.com/exec/obidos/redirect?link_code=ur2&...
COMPACT APPLIANCE
SMALL SOLUTIONS ~ SUPERIOR SERVICE
800-297-6076
COLUMBUS DAY SAVINGS EVENT
Save 5% OFF ALL ORDERS THIS MONDAY & TUESDAY ONLY!
Clearance
Departments
Specialty Areas
Shopping Services
Electronics › "battery charger"
Over 100 Battery Chargers products in Electronics
Sony BCG-34HE4 Super-Quick Worldwide Battery Charger with 4 AA NiMH Batteries
Rayovac Universal Battery Charger for Rechargeable AA, AAA, C, D, and 9-Volt Batteries
La Crosse Technology BC-900 AlphaPower Battery Charger
Black and Decker VEC1086BBD 6/4/2 AMP BATTERY CHARGER

Comparing Judgment Types

- Judges use absolute vs. relative interfaces
- Two problem domains
  - Search
  - Ads
- Measured
  - Agreement
  - Time
  - Also, studied advanced techniques to reduce number of judgments, calibrate to actual scale, etc.



Carterette *et al.*, ECIR 2008

Summary Findings

Search Inferred Preferences from Absolute

	A < B	A,B Bad	A > B	Total
A < B	**0.657**	[illegible]	[illegible]	[illegible]
A,B Bad	0.297	**0.380**	[illegible]	[illegible]
A > B	0.278	0.053	**0.669**	[illegible]

Search Relative Preferences

	A < B	A,B Bad	A > B	Total
[illegible]	**0.752**	0.033	0.215	2580
[illegible]	[illegible]	**0.567**	0.225	413
[illegible]	[illegible]	[illegible]	**0.765**	2757

Ad Inferred Preferences from Absolute

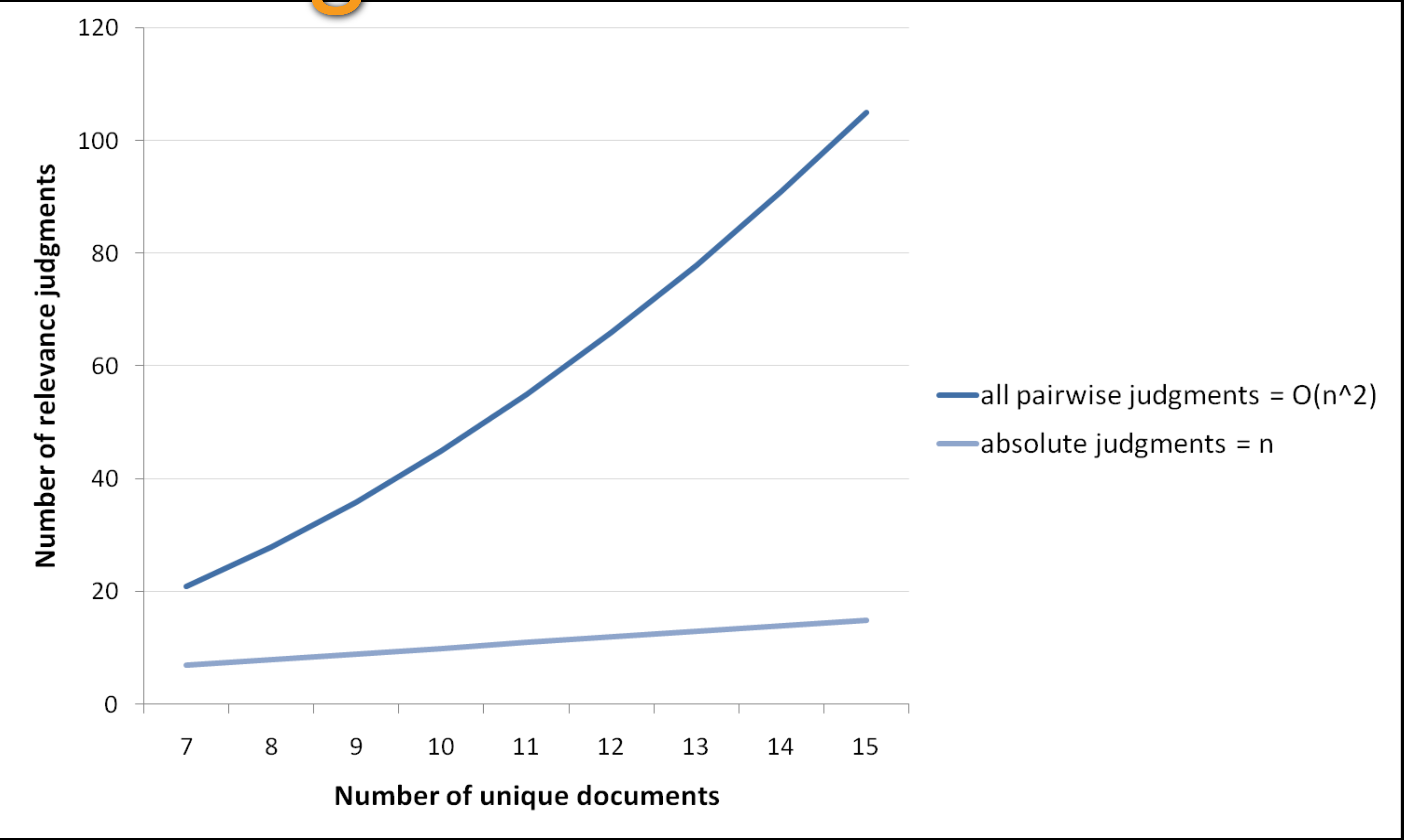
	A < B	A,B Bad	A > B	Total
A < B	**0.635**	[illegible]	[illegible]	[illegible]
A,B Bad	0.263	**0.436**	[illegible]	[illegible]
A > B	0.377	0.009	**0.614**	[illegible]

Ad Relative Preferences

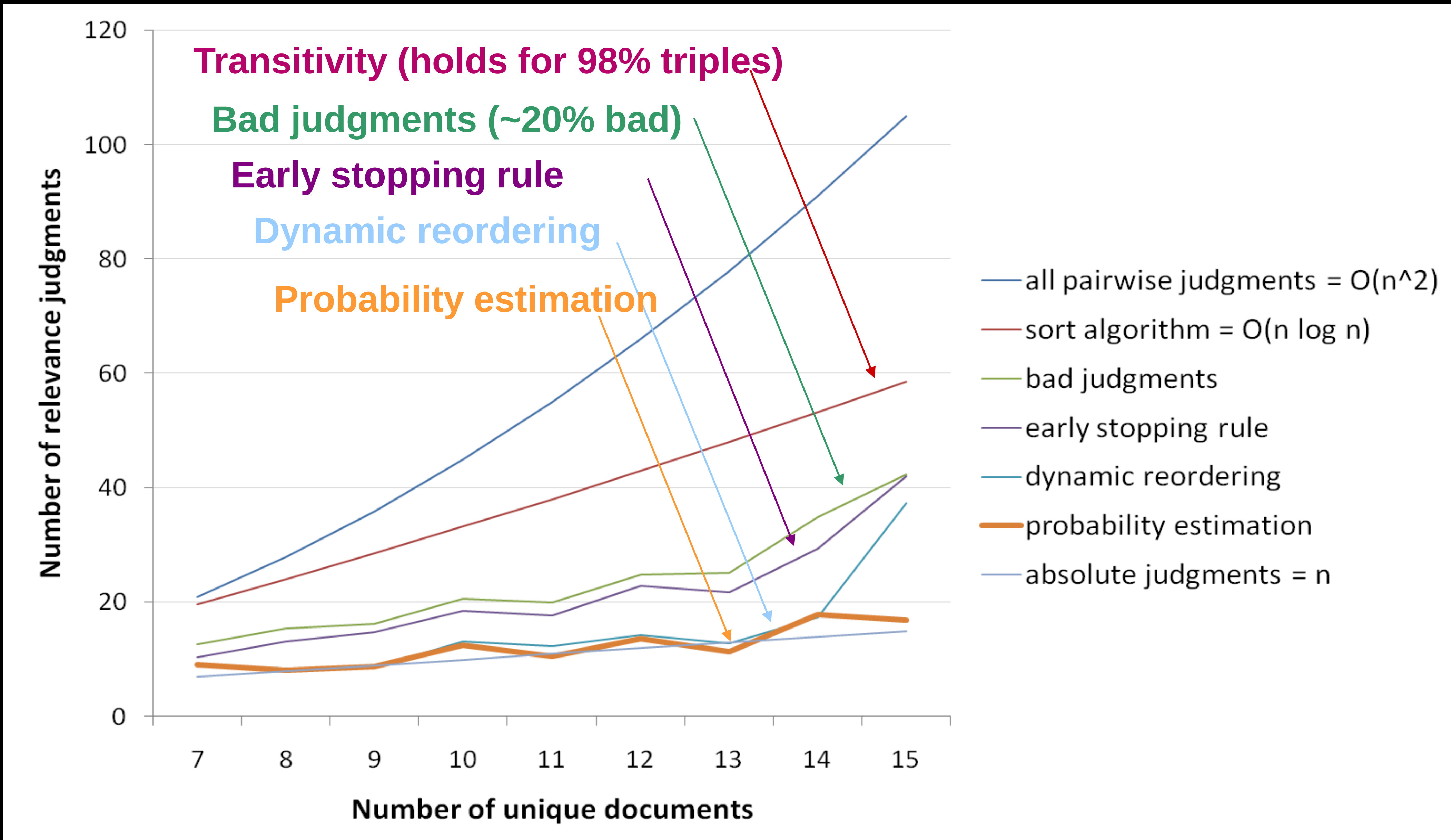
	A < B	A,B Bad or Dup	A > B	Total
[illegible]	**0.700**	0.022	0.277	2691
[illegible]	[illegible]	**0.500**	0.279	272
[illegible]	[illegible]	[illegible]	**0.724**	2985

- Relative Preferences have higher interjudge agreement than absolute.
- Faster per judgment (Two to Three times quicker)
- Other Observations
  - Finer-grained
  - Judges like it better!
- So what are the problems?

The Big O Problem



The Road from n^2 to n



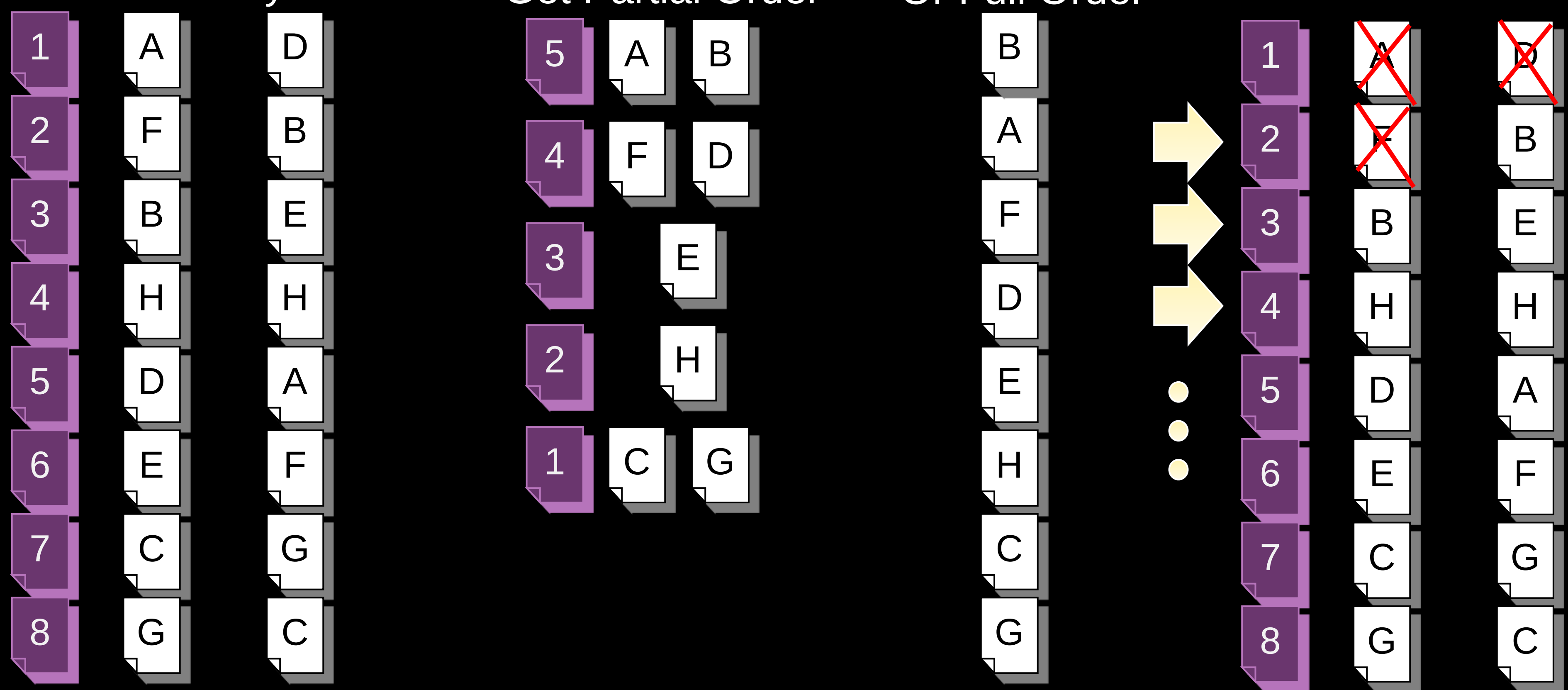
Computing Average Preference Precision

Rank for two systems

Get Partial Order

Or Full Order

Count Out-of-Order Pairs Above Each Rank



Out-Of-Order Count
System 1: 2
System 2: 8

1 – Out-Of-Order/[n(n-1)/2]
System 1: *0.857*
System 2: *0.714*

Carterette & Bennett, SIGIR 2008

Summarizing Preferences for Relevance

- Absolute judgments are noisy and effect system evaluation.
  - Relative judgments have much higher agreement.
  - Relative judgments are faster per judgment.
  - Can reduce total number of relative judgments needed.
  - Relative judgments situate the assessor in a higher degree of context.
- Both learning algorithms and evaluation measures for preference-based judgments are available.

Road Map

- Why Preferences?
- Learning Consensus from Preferences at Scale
- Beyond Consensus
- Discussion

Bennett *et al.*, WWW 2009

Ranking for Search

- Start with a dataset of ground truth
  - An editorial judgment of relevance?
  - A click?
- Apply your favorite ranking learning algorithm
- Problem solved!

Editorial Judgments as Truth

- Pros
  - Control full process
  - Can (somewhat) calibrate judges to a consensus standard
- Cons
  - Ownership of query – what is the user's need?
  - Is relevance *topical* only?
    - Does the quality, authoritativeness, readability matter?
    - How about focus, composition, or artistry for images?

Clicks as Truth

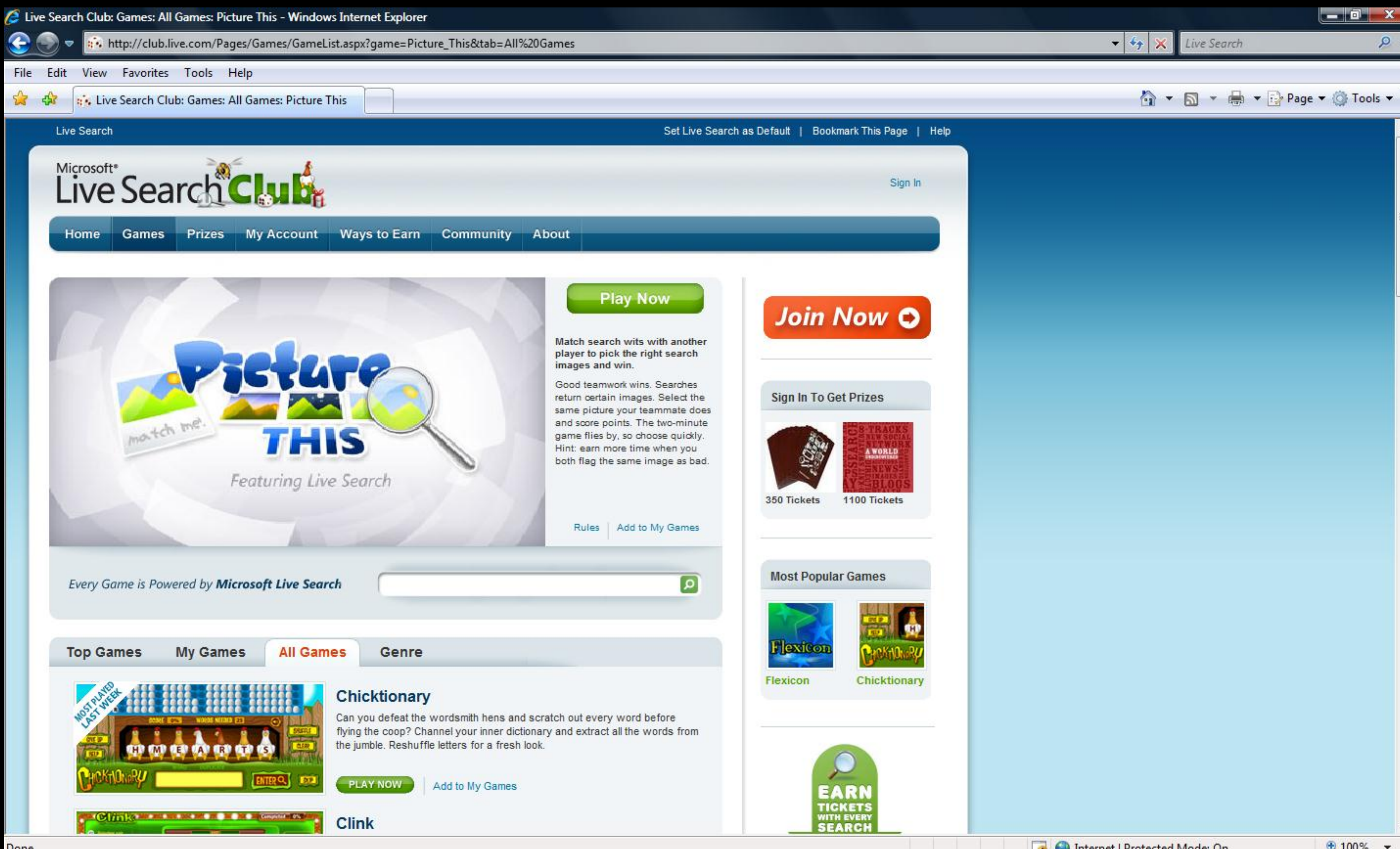
- Pros
  - Whatever properties are rolled into relevance for the user, it all gets wrapped into that click.
- Cons
  - Positional Biases
  - What does "no click" mean?
  - An item that isn't displayed can't be clicked.

Consenus Opinion

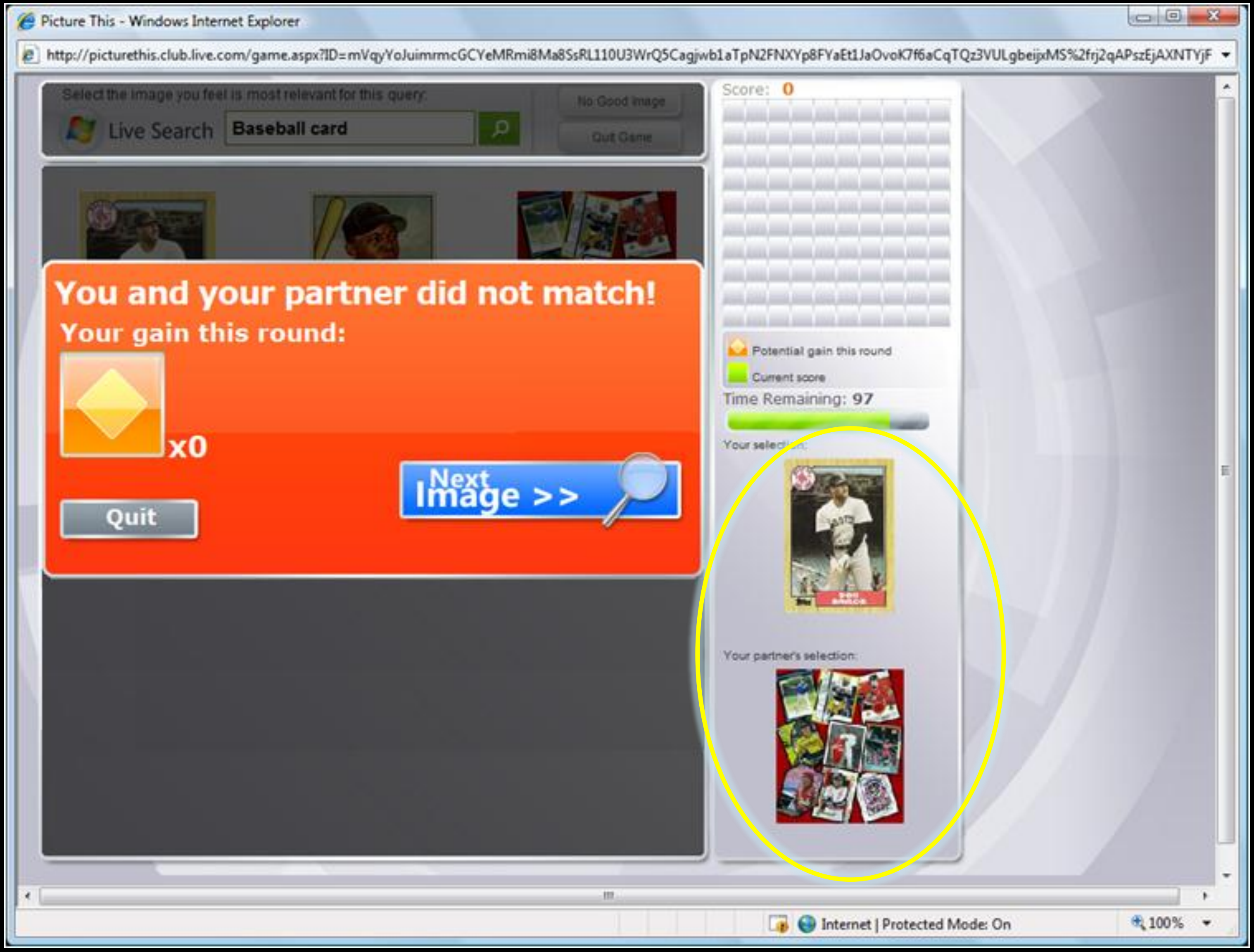
- Desired as ground truth
  - A ranking that is as close to consensus opinion as possible.
- Get rankings from a large number of users
  - Sample the query stream
  - Have a large number of users rank items for many different queries.

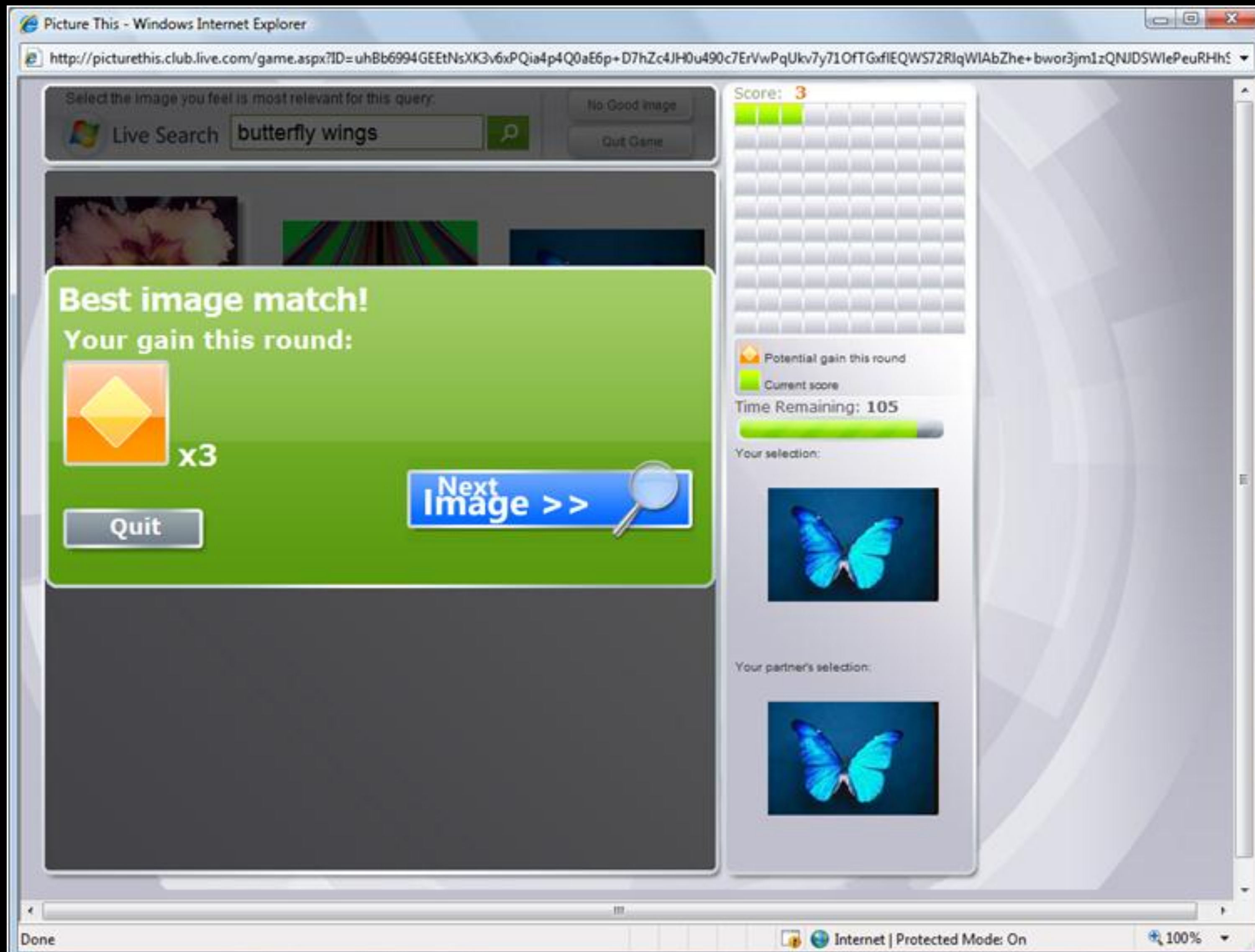
Microsoft Research

Picture This




Picture This - Windows Internet Explorer
http://picturethis.club.live.com/game.aspx?ID=mVqyYoJuimrmcGCYeMRmi8Ma8SsRL110U3WrQ5Cagjwb1aTpN2FNXYp8FYaEt1JaOvoK7f6aCqTQz3VULgbeijxMS%2frj2qAPszEjAXNTYjF
Select the image you feel is most relevant for this query:
Live Search
Baseball card
No Good Image
Quit Game
Score: 0
Potential gain this round
Current score
Time Remaining: 112
Your selection:
Your partner's selection:
Your partner has chosen an image
Done
Internet | Protected Mode: On
100%


Picture This - Windows Internet Explorer
http://picturethis.club.live.com/game.aspx?ID=mVqyYoJuimrmcGCYeMRmi8Ma8SsRL110U3WrQ5Cagjwb1aTpN2FNXYp8FYaEt1JaOvoK7f6aCqTQz3VULgbeijxMS%2frj2qAPszEjAXNTYjF
Select the image you feel is most relevant for this query:
Live Search
Baseball card
No Good Image
Quit Game
Score: 0
You and your partner did not match!
Your gain this round:
x0
Next Image >>
Quit
Potential gain this round
Current score
Time Remaining: 97
Your partner's selection:
Internet | Protected Mode: On
100%


Picture This - Windows Internet Explorer
Select the image you feel is most relevant for this query:
Live Search
butterfly wings
No Good Image
Quit Game
Best image match!
Your gain this round:
x3
Next Image >>
Quit
Score: 3
Potential gain this round
Current score
Time Remaining: 105
Your selection:
Your partner's selection:
Done
Internet | Protected Mode: On
100%

A Different Kind of Click

- Positional Biases
  - Order of choices is randomized.
- What does “no click” mean?
  - Clicks are required to play the game.
  - User can choose “No Good Image” or flag an image as “Bad” (irrelevant or detrimental)
- An item that isn’t displayed can’t be clicked.
  - Can experiment with potentially irrelevant items without risking reputation of search.

Adaptive Game Elements

- Number of images to choose from varies from 2 to 10 (based on performance).
- Comparison set varies from “easy” to “hard”
  - Easy: Images estimated relevance spread across spectrum.
  - Hard: Images close to the same level of relevance.
- Adaptivity
  - Take advantage of more “discerning” partners by increasing game difficulty.
  - Make the game more entertaining.

Partner Robots

- Sometimes no human is available to partner with (*e.g.* odd man out, low system usage)
- Robot partner is used to play the game.
- Robot chooses non-deterministically using basic model of relevance from preferences collected so far.

Fraud Mitigation

- Discussion boards dedicated to how to cheat on Club Bing.
- Random pairing of partners
- No fixed strategy
  - randomization of choices, no score for “no good image”
- Weeding out Bots
  - Players occasionally challenged with a CAPTCHA
  - Game site uses bot detection algorithm
  - Honeypot queries (queries with a highly agreed upon right answer)

Are Responses Random?

- Users permanently routed by ID to a “training” or “testing” server
- Training Set:
  - 49% agreement on human-human rounds.
  - 3.53 actions available on average.
  - Random clicking would yield 28% agreement.
- Testing Set:
  - 50% agreement on human-human rounds.
  - Random clicking would yield 28% agreement.

Game Data

- Simple Experiment
  - Split users into train and test.
  - Use preferences of “train” users to predict preferences of “test” users *when partners agree*.
  - Default hypothesis is user preferences are explained by Image Search’s ranking
- Goal is to predict preferences where two users agree in the test set.
- Agreement only and Raw (Full) used in training.
- 34 days, 427 queries, 95 images/query, 18M (effective) pairwise preferences.

Preprocessing

- “Raw” form
  - Our robots preferences removed.
- “Agreement” form
  - Removed all preferences where partners disagreed.
  - Removed all preferences where human was paired with a robot.
  - Kept only one preference for every pair of preference agreements.

Microsoft Research

Data Overview

- 34 days, 427 queries
- 95 images per query (94 + “neutral”)
  - More than 50 because of churn

	Games	Rounds	Human-Human Rounds
Training	154,060	1,491,206	1,144,409
Testing	155,322	1,522,375	1,159,570

	Preferences	Effective Pairwise Preferences
Raw Training	2,599,531	8,860,418
Raw Testing	2,645,984	9,267,890

	Preferences	Effective Pairwise	No Good Image
Agreement Training	537,651	1,731,317	1,800
Agreement Testing	612,736	1,838,987	1,447

Consensus Ranking

- Goal Reminder: Learn a ground truth ranking.
- Input: Features are “Was image ID X displayed?”
- Label: which ID was chosen as best.
- Model: Gives us consensus ranking.
  - Intuition: a good model predicts what people think is best.

Preference Models

- Progression from most naïve (simplest) to fewest assumptions (most complex)
  - Frequency Model
    - Global win probability of an image
  - Pairwise Probability Model
    - Models pairwise interactions but not comparison set
  - Go Model
    - Interactions in comparison set by conditioning on current set and response when learning and predicting.

Data Representation

- Introduce virtual “Neutral Image” to handle
  - “No Good Image”: Preference for neutral image over every image displayed
  - “Flag as Bad”: Preference for neutral image over image flagged.
- Pairwise preferences:
  - Image *A* preferred to *B, C, and D* can be represented as three pairwise preferences: *A>B, A>C, A>D.*

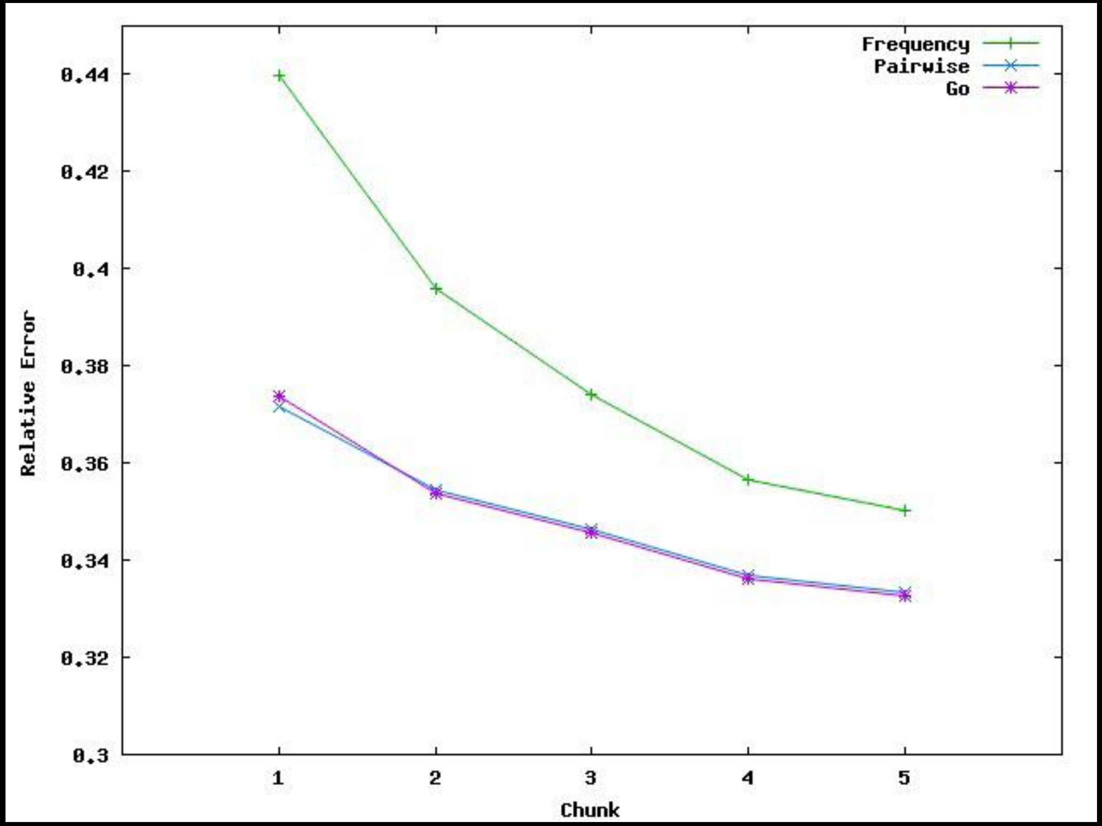
Error Relative to Image Search

	Agreement Training	Raw Training
Frequency	0.3713	0.3504
Pairwise	0.3451	0.3335
Go	**0.3408**	**0.3325**

- All methods perform at least twice and up to three times better than search baseline.
- Given a large enough training sample, simple method works well.
- More structured, complex models can use the (potentially) noisier raw training versions better.

Learning Curve

- Break raw training data into 5 ~7-day chunks
- Each training set is a superset of the previous.



- .

Learning Curve Lessons

- Frequency model performs increasingly worse with less data
  - Higher sparsity relative to number of pairs.
- Worth using complex models if we want to get the minimal number of preferences per query and cover more queries.

Versus Editorial Judgments

- Kendall's Tau:
  (agreements – disagreements) / num_differing_pairs
- Use set of already collected editorial judgments for 60 of 427 queries.
- Examined correlation of Pairwise model (trained on Agree)

	Kendall's Tau	Differing Pairs	Agreements	Disagreements	Raw Pairs
Pairwise	0.6742	27299	22852	4447	67638

Related Work

- Social Labeling: ESP Game (von Ahn & Dabbish, 2004), Peekaboom (von Ahn *et al.*, 2006), TagATune (Law *et al.*, 2007)
  - ESP does not tell you about relevance and tends to least common denominator descriptions.
- Preference Learning
  - Similar in spirit to Joachims (2002) with different source of data. No experimenting with engine itself as in Radlinski & Joachims (2005) or Radlinski *et al.* (2008)

Our Contributions

- Versus other social labeling
  - We use queries drawn from search logs → language distribution seen in practice drives labeling.
- First study of large-scale preference data as a signal to augment other sources of relevance information.
- Unique approach yields click-like relevance data
  - No positional biases, no risk of frustrating users when given non-relevant results, no danger to product reputation.
- Systematic study of a number of preference models
  - For plentiful data, many models are fine.
  - For sparse data, pairwise and Go both work well.
- April 2008 – Jan 2009
  - 2.5M visitors, 60M searches (~90M clicks)
  - Conceivable to collect judgments in academic setting.

Road Map

- Why Preferences?
- Learning Consensus from Preferences at Scale
- Beyond Consensus
- Discussion

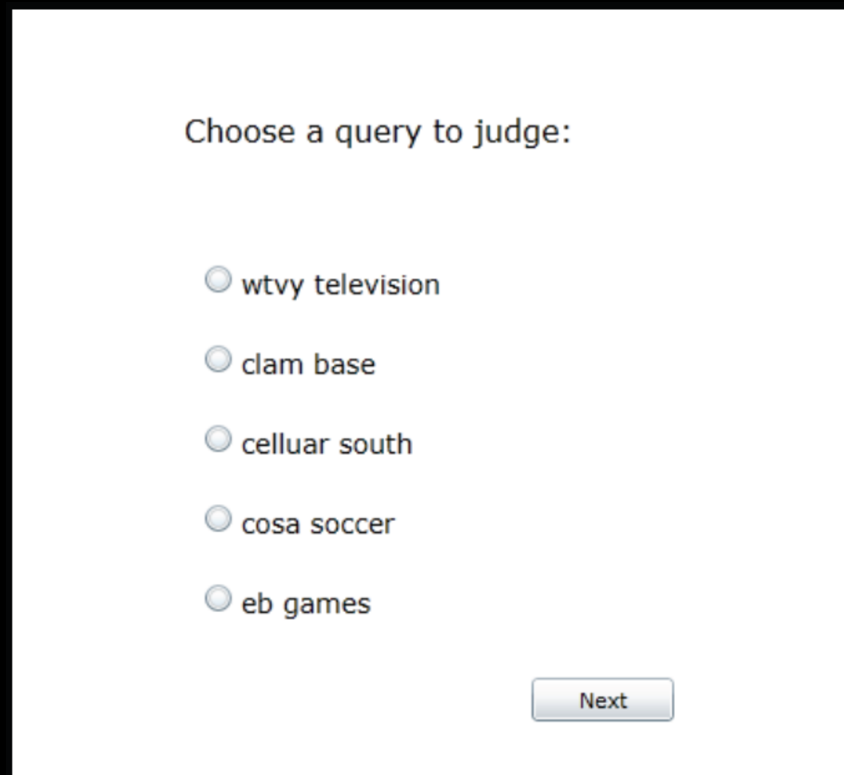
Law *et al.*, *SIGIR 2011*

The need for expertise

The Ideal Assessor

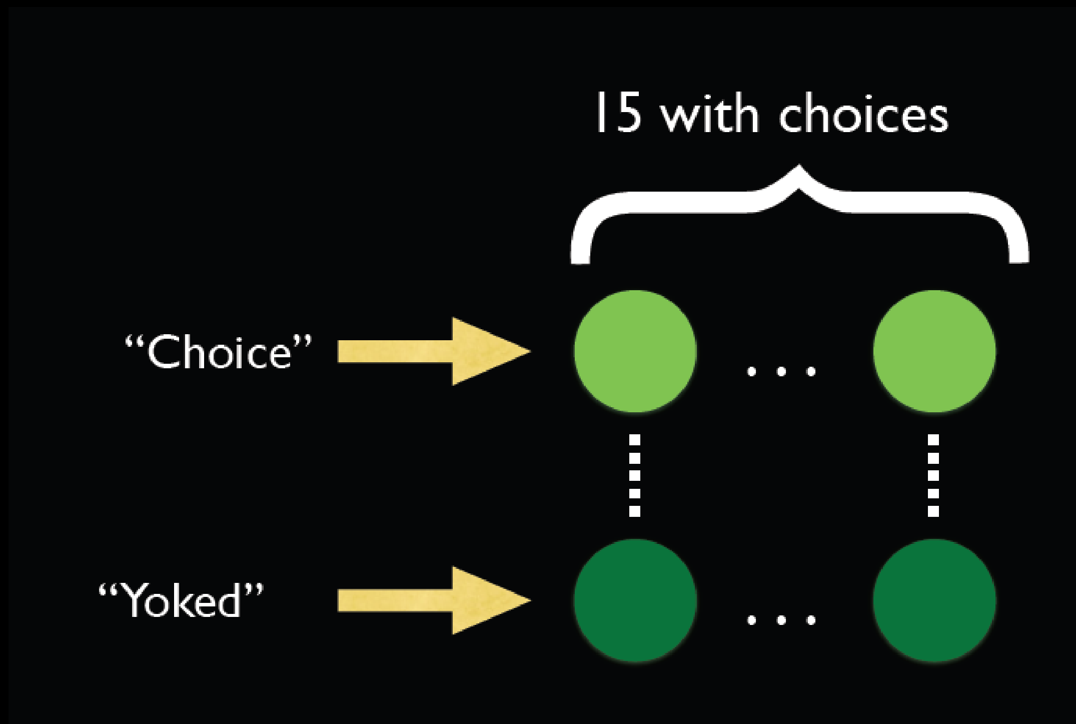
- Interest: Assessor for a query is someone who would likely *issue a* query.
- Expertise: Assessor can *distinguish* relevant, satisfying results from irrelevant ones.
- Confidence: Assessors judgments *agree* with measured utility (high signal to noise).

Choice



- For relevance judgments, can we leverage choice as a signal of expertise, interests and confidence?

Experiment Design



- Balance choices by length, ambiguity, and topic as a proxy for ease of choice selection, difficulty of task, and differentiability of candidates.
- For each query, pre- and post-judgment survey.

Survey

Current Search Query: clam base

[Task 4 of 25] Step 1 of 7

Before you judge the relevance of the webpages for this search query, please answer the following questions:

(1) Specify up to 3 intents (i.e., what the user was looking for) behind the search query and mark the one that is the most likely.

Most Likely

(2) How confident are you that the most likely intent you specified is what the user is looking for?

Very unsure — Very confident

1 2 3 4

(3) How knowledgeable are you in the topic of the search query:

Zero knowledge — Expert

1 2 3 4

(4) How interested are you in the topic of the search query:

Zero interests — Very interested

1 2 3 4

Next

Influence of Choice

- Intent modification: yoked subjects make non-superficial changes to the intentions more often (13.84% of all, and 5.35% of most likely intention).

Summary & Challenges

- Demonstrated
  - Simple mechanism of choice increased self-assessed interest, expertise, and confidence.
  - Also reduces the impact of exposure to the task as measured by intent modification – implying expertise.

- Challenges
  - Objective measurements of expertise beyond intent modification.
  - Given choices, predict interests/expertise and narrow set of choices around them.
  - Balance available experts with demand for tasks.

Road Map

- Why Preferences?
- Learning Consensus from Preferences at Scale
- Beyond Consensus
- Discussion

Specific Challenges

- Modeling player/judgment/worker performance
  - Personalized rankings, discrimination skill, identifying expertise
- Active Learning for Query & Result Selection
  - Which query, which results, and how many opinions?
- Using crowdsourcing to identify noisy *editorial* judgments or difficult to interpret click data.
- Identifying, measuring, and exploiting sources of expertise.

Microsoft Research

A Spectrum of Crowdsourcing

Anyone in crowd can answer

It takes the whole crowd to answer

Much Research

Little Research

Crowd as a cheap way …

Consensus

Expertise

Multistage

Personalization

Microsoft Research

Microsoft®

Your potential. Our passion.™