Advances in IR Evaluation

Ben Carterette Evangelos Kanoulas Emine Yilmaz

The University Of Sheffield.

Microsoft Research

Course Outline

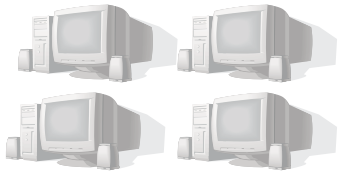
- Intro to evaluation
  - Evaluation methods, test collections, measures, comparable evaluation
- Low cost evaluation
- Advanced user models
  - Web search models, novelty & diversity, sessions
- Reliability
  - Significance tests, reusability
- Other evaluation setups

Judgment Effort

- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results

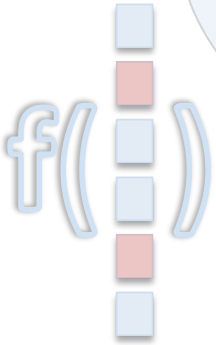
Search Engines



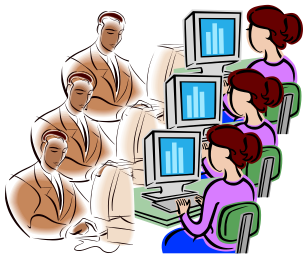
Cost

vs.

Reliability

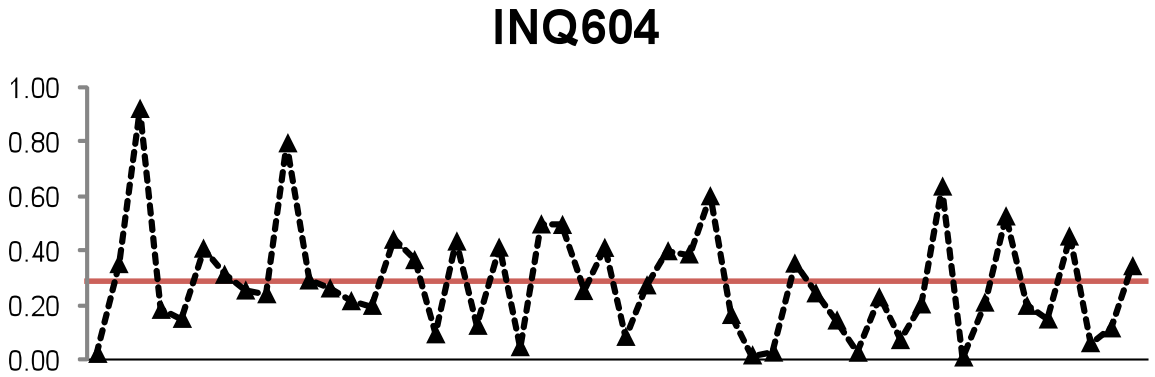


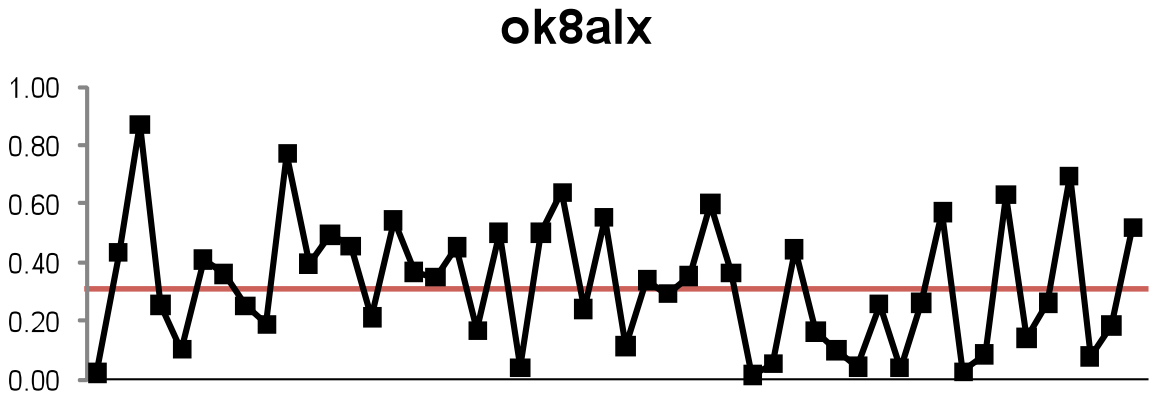
Judges

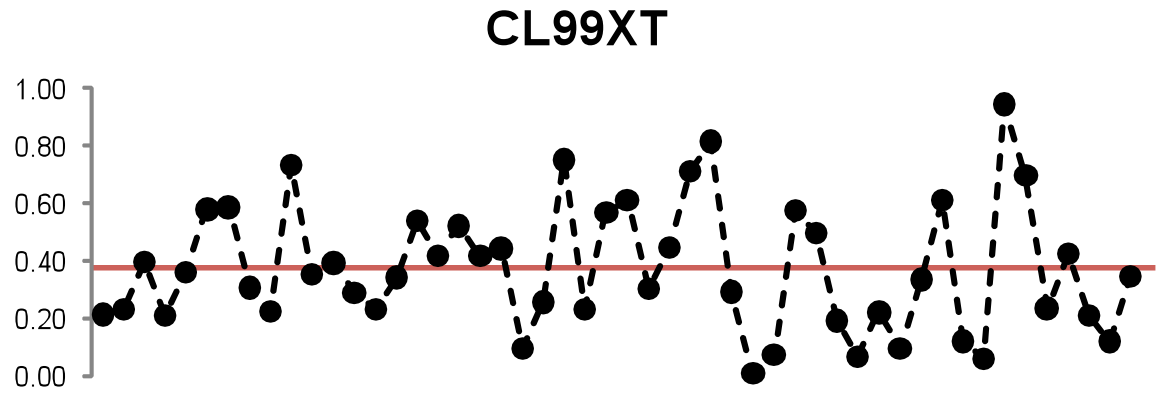


Average?

- INQ604 - 0.281
- ok8alx - 0.324
- CL99XT - 0.373







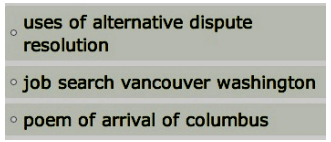
How much is the average

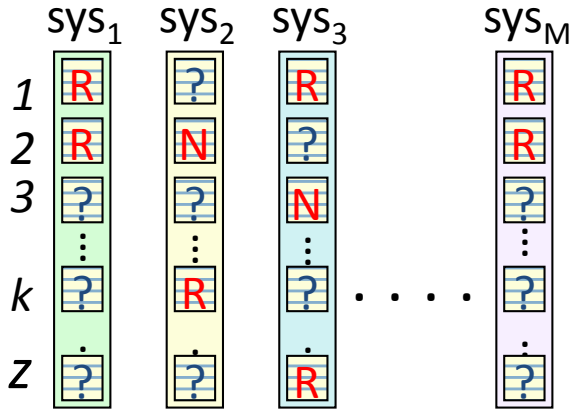
- A product of
  - your IR system
  - chance?
- Slightly different set of topics?
  - Would the average change?

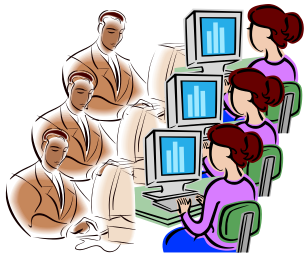
Reliability

- Reliability : the extent to which results reflect real difference (not due to chance)
- Variability in effectiveness scores due to

1. Differentiation in nature of **documents** (corpus)
2. Differentiation in nature of **topics**
3. **Incompleteness** of relevance judgments
4. **Inconsistency** among assessors







Reliability

- Variability in effectiveness scores due to
  1. Differentiation in nature of **documents**
  2. Differentiation in nature of **topics**
  3. **Incompleteness** of judgments
  4. **Inconsistency** among assessors

- Document corpus size

  [Hawking and Robertson J of IR 2003]

- Topics vs. assessors

  [Voorhees SIGIR98, Banks et al. Inf. Retr.99, Bodoff and Li SIGIR07]

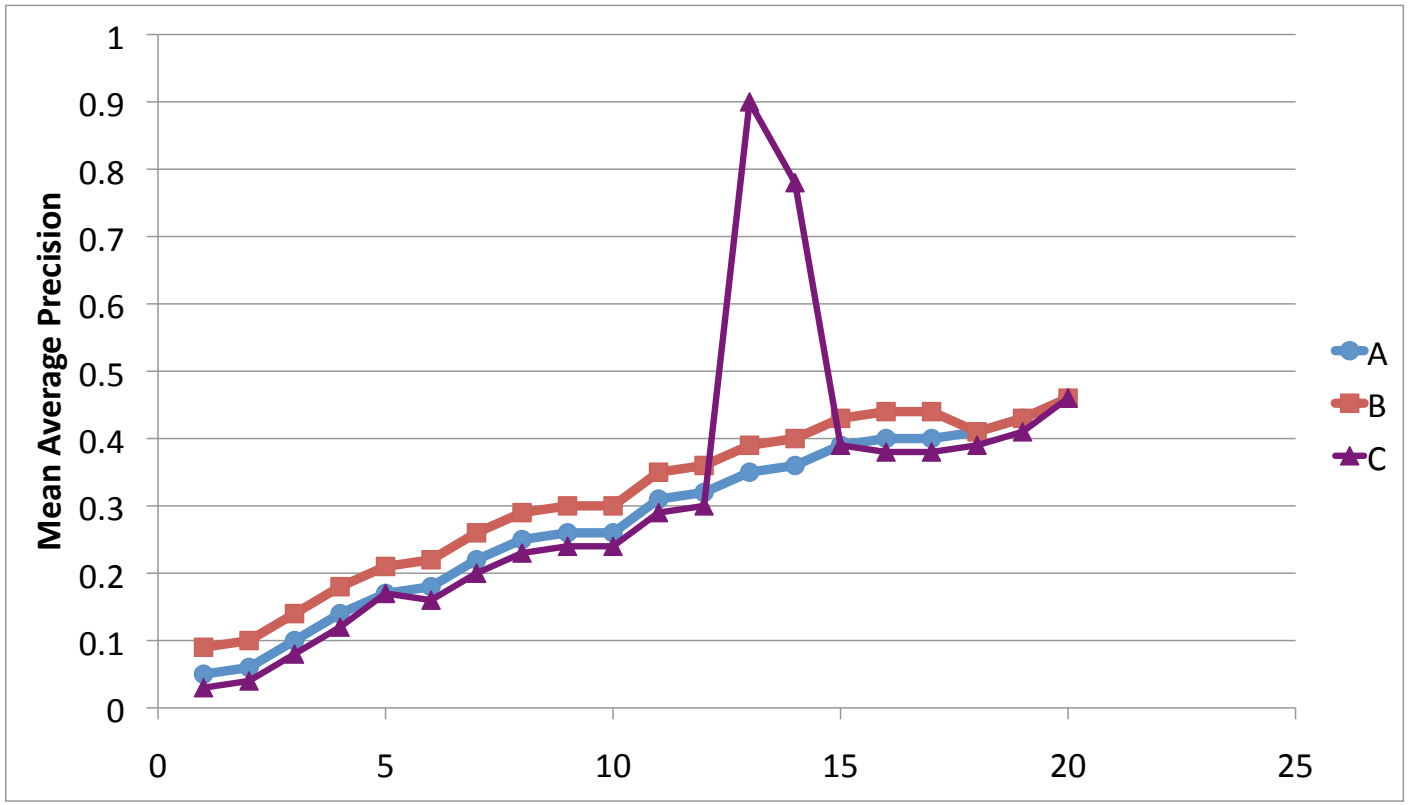
- Effective **size** of the topic set for reliable evaluation

  [Buckley and Voorhees SIGIR00/SIGIR02, Sanderson and Zobel SIGIR05]

- Topics vs. documents (Million Query track)

  [Allan et al. TREC07, Carterette et al. SIGIR08]

Is C really better than A?



Comparison and Significance

- Variability in effectiveness scores
- When observing a difference in effectiveness scores across two retrieval systems
  - Does this difference occur by **random chance**?
- Significance testing
  - Estimates the probability *p* of observing a certain difference in effectiveness given that H_0 is true.
  - In IR evaluation
    - H_0 : the two systems are in effect the same and any difference in scores is by random chance.

Significance Testing

- Significance testing framework:
  - Two hypotheses, e.g.
    H_0: $\mu = 0$
    H_a: $\mu \neq 0$
  - System performance measurements over a sample of topics
  - A test statistic T computed from those measurements
  - The distribution of the test statistic
  - A p-value, which is the probability of sampling T from a distribution obtained by assuming H_0 is true

Student's t-test

- Parametric test
- Assumptions
  1. effectiveness score differences are meaningful
  2. effectiveness score differences follow normal distribution

- Statistic : $t = \frac{\overline{B-A}}{\sigma_{B-A}}.\sqrt{N}$

- t-test performs well even when the normality assumption is violated [Hull SIGIR93]

Student's t-test

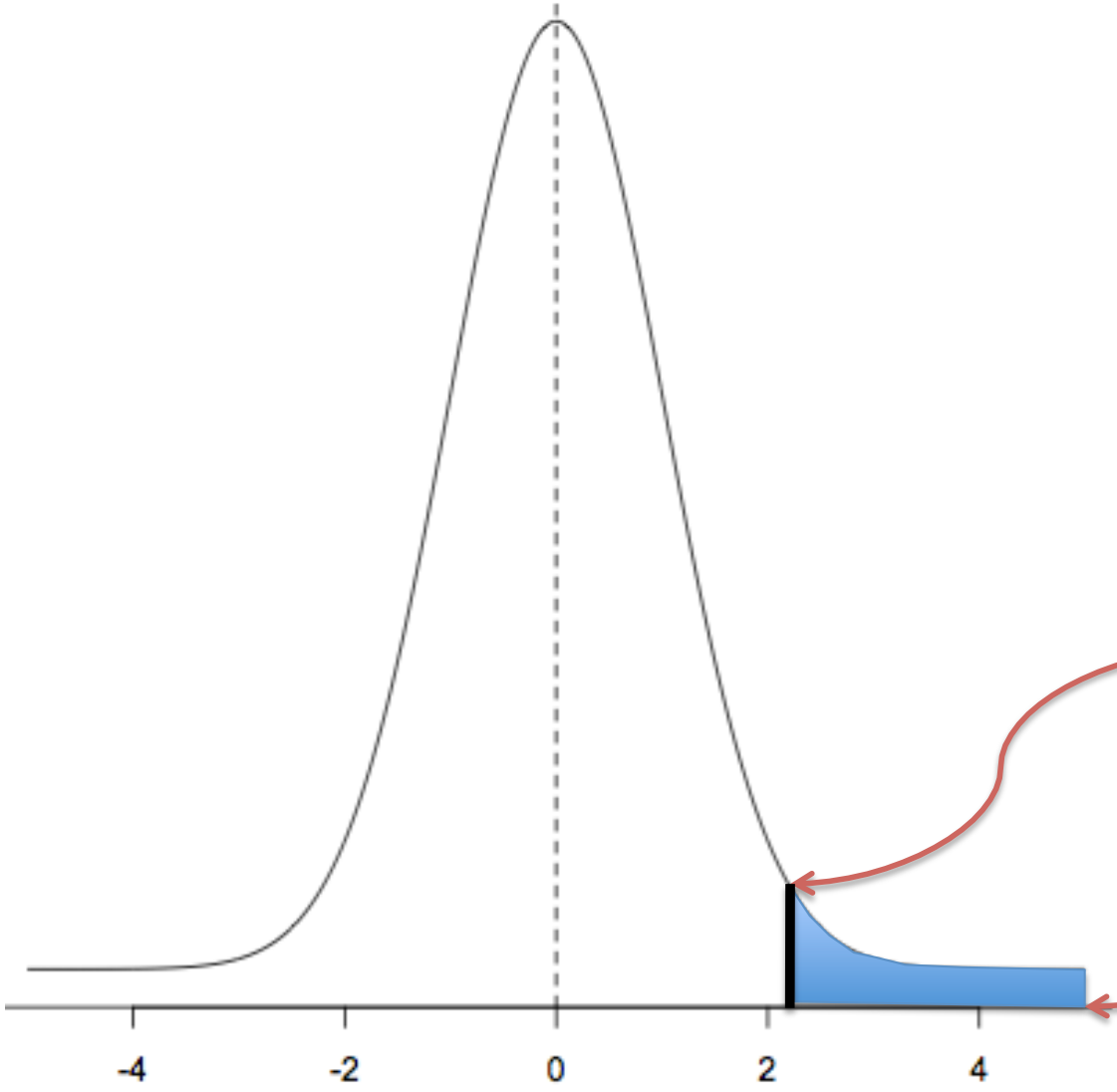
Query	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

$$\overline{B-A} = 0.214$$

$$\sigma_{B-A} = 0.291$$

$$t = \frac{B-A}{\sigma_{B-A}}\sqrt{n} = 2.33$$

Student's t-test



$$\overline{B-A} = 0.214$$

$$\sigma_{B-A} = 0.291$$

$$t = \frac{B-A}{\sigma_{B-A}}\sqrt{n} = 2.33$$

$$p - value = 0.02$$

Sign Test

- Non-parametric test

- Ignores magnitude of differences
- Null hypothesis for this test is that
  - $P(B > A) = P(A > B) = ½$

- Statistic : number of pairs where B>A

Wilcoxon Signed-Ranks Test

- Non-parametric test
- Statistic: $w = \sum_{i=1}^{N} R_i$
  - R_i is a signed rank of absolute differences
  - N is the number of differences ≠ 0

Wilcoxon Signed-Ranks Test

Query	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

Sorted
-.02
+.09
+.10
-.24
+.25
+.25
+.41
+.60
+.70

Signed-rank
-1
+2
+3
-4
+5
+6
+7
+8
+9

$$w = \sum_{i=1}^{N} R_i$$

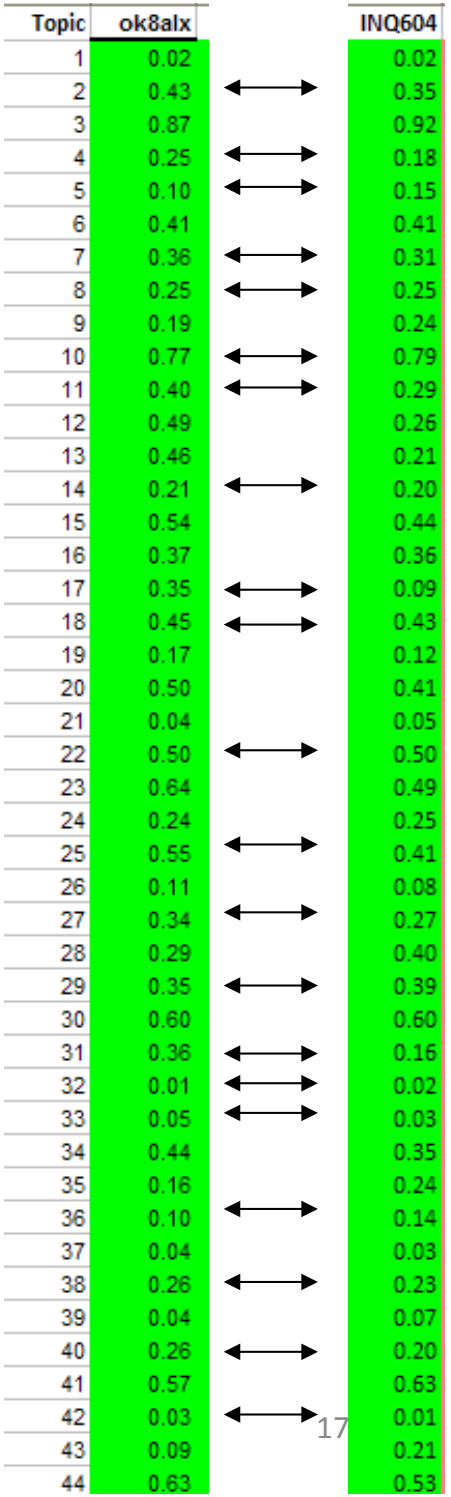
w = 35 =>

p = .025

Randomisation test

- Loop for many times {
  1. Load topic scores for 2 systems
  2. Randomly swap values per topic
  3. Compute average for each system
  4. Compute difference between averages
  5. Add difference to array

  }
- Sort array
- If actual difference outside 95% differences in array
  - Two systems are significantly different



Topic	ok8alx		INQ604
1	0.02		0.02
2	0.43	↔	0.35
3	0.87		0.92
4	0.25	↔	0.18
5	0.10	↔	0.15
6	0.41		0.41
7	0.36	↔	0.31
8	0.25	↔	0.25
9	0.19		0.24
10	0.77	↔	0.79
11	0.40	↔	0.29
12	0.49		0.26
13	0.46		0.21
14	0.21	↔	0.20
15	0.54		0.44
16	0.37		0.36
17	0.35	↔	0.09
18	0.45	↔	0.43
19	0.17		0.12
20	0.50		0.41
21	0.04		0.05
22	0.50	↔	0.50
23	0.64		0.49
24	0.24		0.25
25	0.55	↔	0.41
26	0.11		0.08
27	0.34	↔	0.27
28	0.29		0.40
29	0.35	↔	0.39
30	0.60		0.60
31	0.36	↔	0.16
32	0.01	↔	0.02
33	0.05	↔	0.03
34	0.44		0.35
35	0.16		0.24
36	0.10	↔	0.14
37	0.04		0.03
38	0.26	↔	0.23
39	0.04		0.07
40	0.26	↔	0.20
41	0.57		0.63
42	0.03	↔	0.01
43	0.09		0.21
44	0.63		0.53

Inference from Hypothesis Tests

- We often uses tests to make an inference about the population based on a sample
  - e.g. infer that H_0 is false in the population of topics
- That inference can be wrong for various reasons:
  - Sampling bias
  - Measurement error
  - Random chance
- There are two classes of errors:
  - Type I, or false positives: we reject H_0 even though it is true
  - Type II, or false negatives: we fail to reject H_0 even though it is false

Errors in Inference

- A significance test is basically a classifier

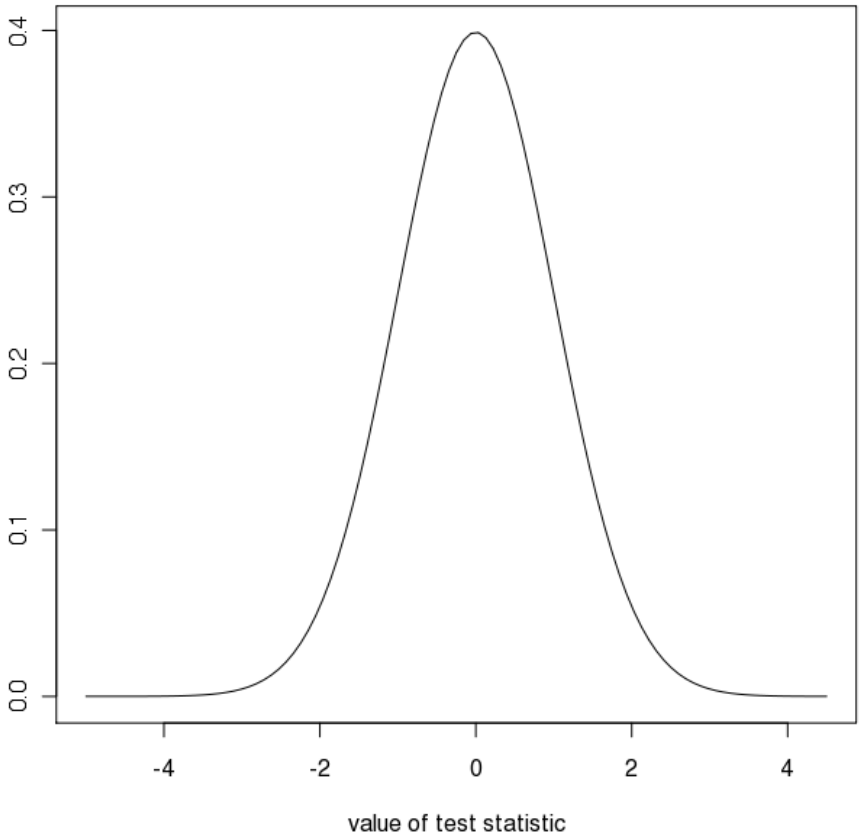
	H_0 false	H_0 true
$p < 0.05$ (reject H_0)	correct	Type I error
$p \geq 0.05$ (do not reject H_0)	Type II error	correct

- We can't actually know whether H_0 is true or not
  - If we could, we wouldn't need the test
- But we can set up the test to control the expected Type I and Type II error rates

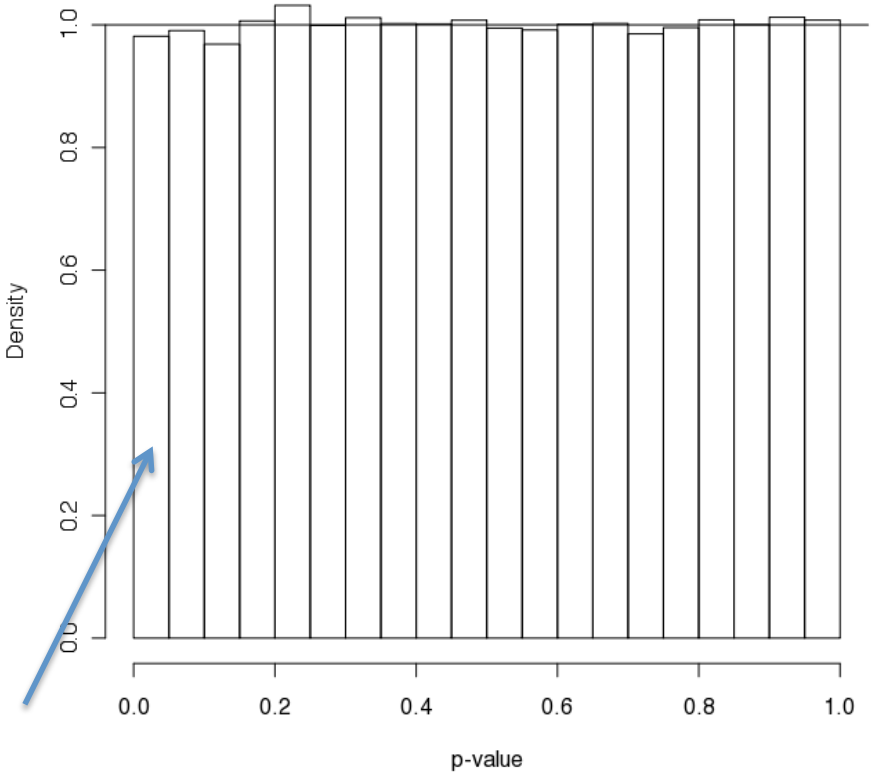
Expected Type I Error Rate

- Test parameter α is used to decide whether to reject H0 or not—if $p < \alpha$, then reject H_0
- Choosing α is equivalent to stating an expected Type I error rate
  - e.g. if $p < 0.05$ is considered significant, we are saying that we expect that we will incorrectly reject H_0 5% of the time
- Why?
  - Because when H_0 is true, every p-value is equally likely to be observed
  - 5% of the time we will observe a p-value less than 0.05… and therefore there is a 5% Type I error rate

Typical distribution of test statistic assuming H_0 true



Distribution of p-values assuming H_0 true

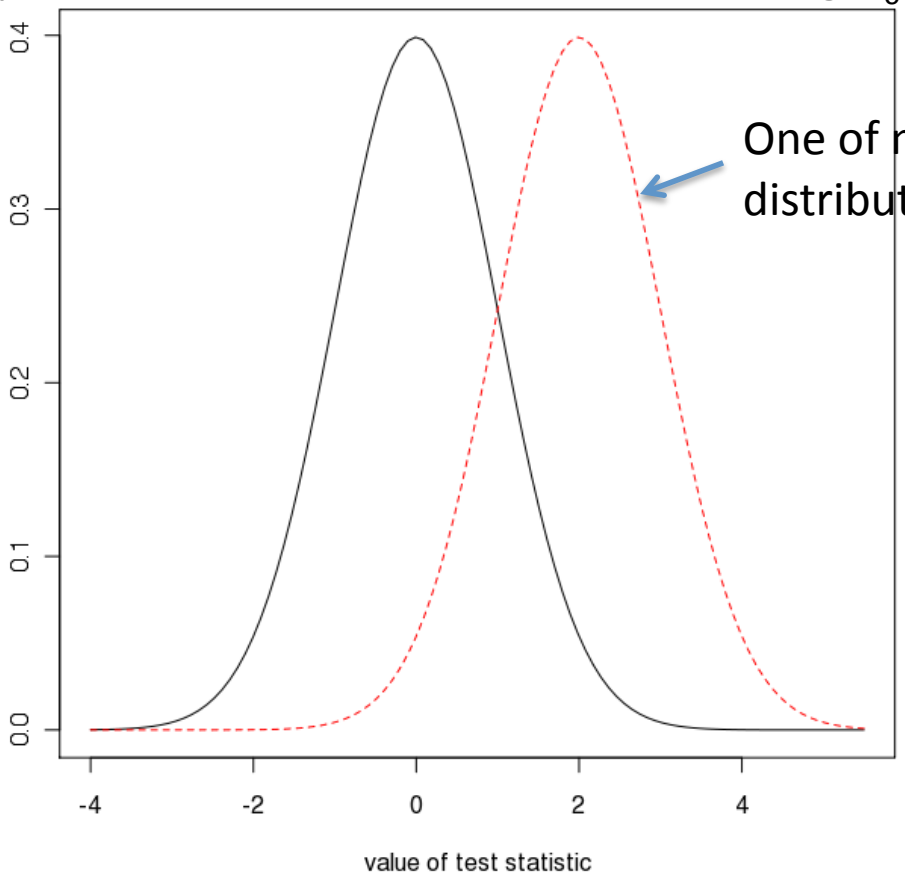


5% chance of incorrectly concluding H_0 false

Expected Type II Error Rate

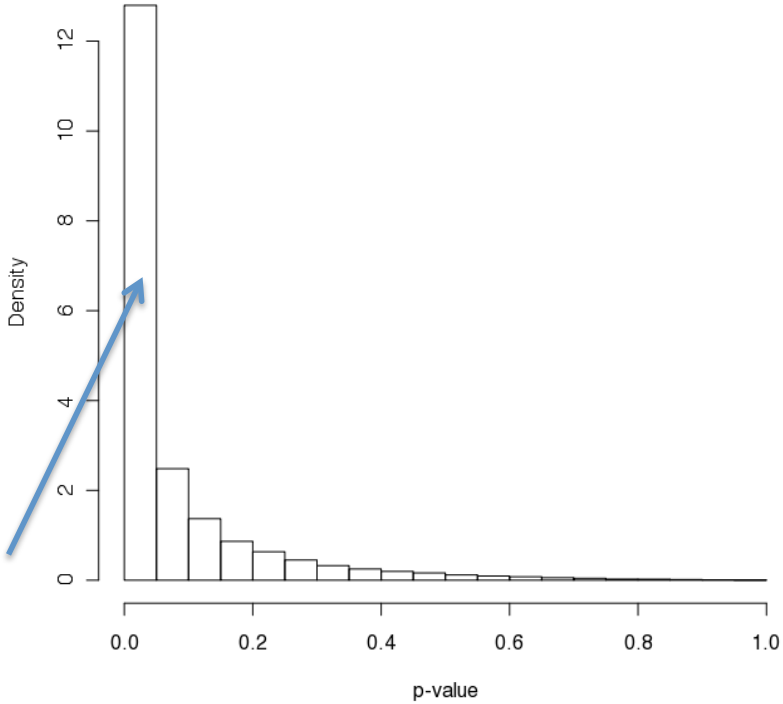
- What about Type II errors?
  - False negatives are bad: if we can't reject H_0 when it's false, we may miss out on interesting results
- What is the distribution of p-values when H_0 is false?
  - Problem: there is only one way H_0 can be true, but there are many ways it can be false

Typical distribution of test statistic assuming H_0 true



One of many possible distributions if H_0 is false

Distribution of p-values for that alternative



0.64 probability of rejecting H_0

⇒ 0.36 Type II error rate

Power and Expected Type II Error Rate

- Power is the probability of rejecting H_0 when it is false
  - Equivalent to (1 – Type II error rate)
- Power parameter is β
  - Choosing a value of β therefore entails setting an expected Type II error rate
- But what does it mean to "choose a value of β"?
  - In the previous slide, β was calculated post hoc, not chosen a priori

Effect Size

- A measure of the magnitude of the difference between two systems
  - Effect size is dimensionless; intuitively similar to % change in performance
  - Bigger population effect size → more likely to find a significant difference in a sample
- Before starting to test, we can say "I want to be able to detect an effect size of h with probability β"
  - e.g. "If there is at least a 5% difference, the test should say the difference is significant with 80% probability"

    ⇒ $h = 0.05$, $\beta = 0.8$

Sample Size

- Once we have chosen α, β, h (Type I error rate, power, and desired effect size), we can determine the sample size needed to make the error rates come out as desired
  - n = f(α, β, h)
  - Usually involves a linear search
  - There are software tools to do this
- Basically:
  - Sample size n increases with β if other parameters held constant
  - If you want more power, you need more samples

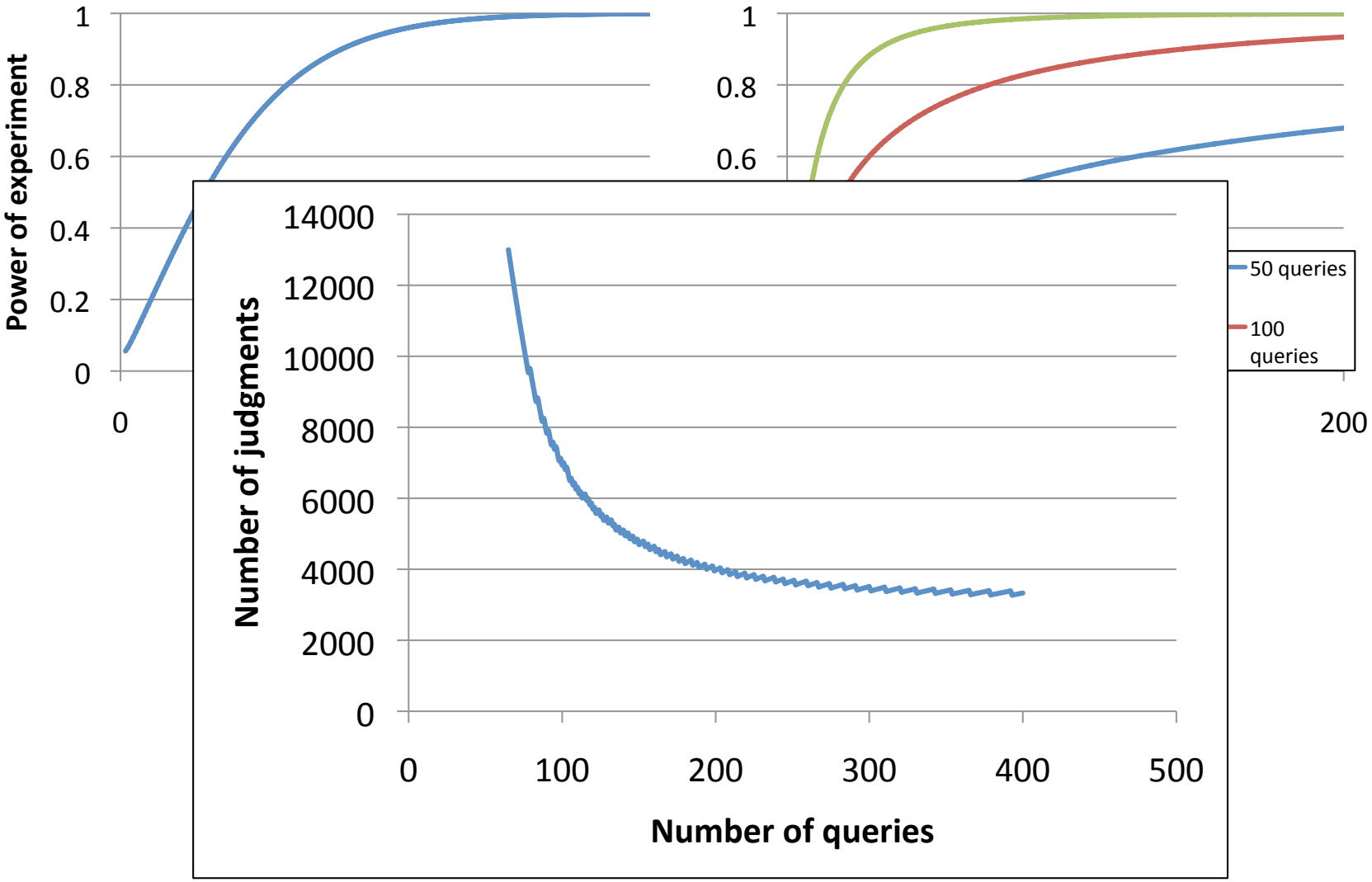
Implications for Low-Cost Evaluation

- First consider how Type I and Type II errors can happen due to experimental design rather than randomness
  - Sampling bias: usually increases Type I error
  - Measurement error: usually increases Type II error
- When judgments are missing, measurements are more errorful
  - And therefore Type II error is higher than expected

Implications for Low-Cost Evaluation

- What is the solution?
  - If Type II error increases, power decreases
  - To get power back up to the desired level, sample size must increase
  - Therefore: deal with reduced numbers of judgments by increasing the number of topics
- Of course, each new topic requires its own judgments
  - Cost-benefit analysis finds the "sweet spot" where power is as desired within available budget for judging

Tradeoffs in Experimental Design



Criticism on Significance Tests

- Significance testing
  - Note: The probability *p* is not the probability that H_0 is true.
    - $p = P(\text{Data} \mid H_0) \neq P(H_0 \mid \text{Data})$

Criticism on Significance Tests

- Are significance tests appropriate for IR evaluation?
- Is there any single best to be used?

  [Saracevic CSL68, Van Rijsbergen 79, Robertson IPM90, Hull SIGIR93, Zobel SIGIR98, Sanderson and Zobel SIGIR05, Cormac and Lynam SIGIR06, Smucker et al SIGIR09 …]
- Are they any useful?
  - With a large number of samples any difference in effectiveness will be statistically significant.
    - e.g. 30,000 queries in new Yahoo! and MSR collections
  - "Strong form" of hypothesis testing [Meehl Phil.Sci.67, Popper 59, Cohen Amer.Psy.94]

Criticism on Significance Tests

- What to do?
  - Improve our data (make them as representative as possible)
  - Report confidence intervals [Cormack and Lynam SIGIR06, Yilmaz et al. SIGIR08, Carterette et al. SIGIR08]
  - Examine whether results are "practically significant" [Allan et al SIGIR05, Thomas and Hawking CIKM06, Turpin and Scholer SIGIR06, Joachims SIGIR02, Radlinski et al CIKM08, Radlinski and Craswell SIGIR10, Sanderson et al. SIGIR10]

Judgment Effort

- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

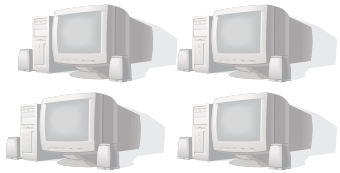
Results

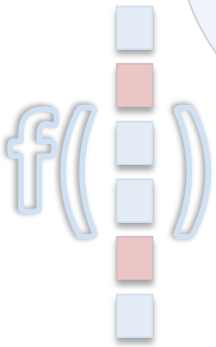
Cost

vs.

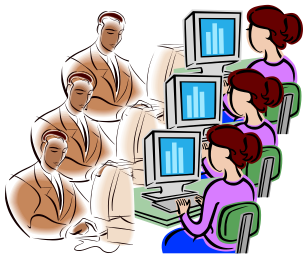
Reusability

Search Engines





Judges



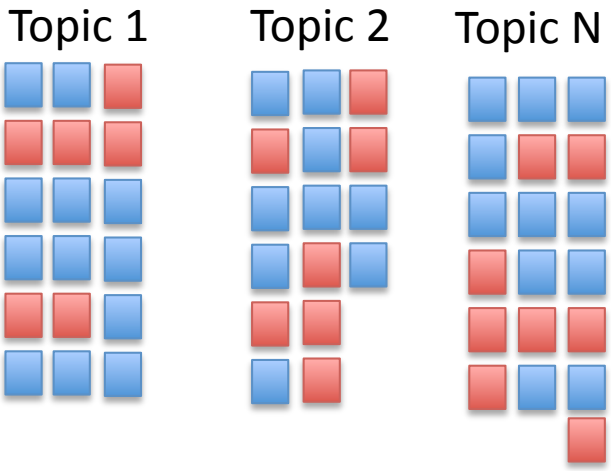
Reusability

- Why reusable test collections?
  - High cost of constructing test collections
    - Amortize cost by reusing test collections
  - Make retrieval system results comparable
    - Test collections used by systems that did not participate while collections were constructed

- Low-cost vs. Reusability
  - If not all relevant documents are identified the effectiveness of a system that did not contribute to the pool may be underestimated

Reusability

Relevance Judgments



New System S'
Ranked List
on Topic1

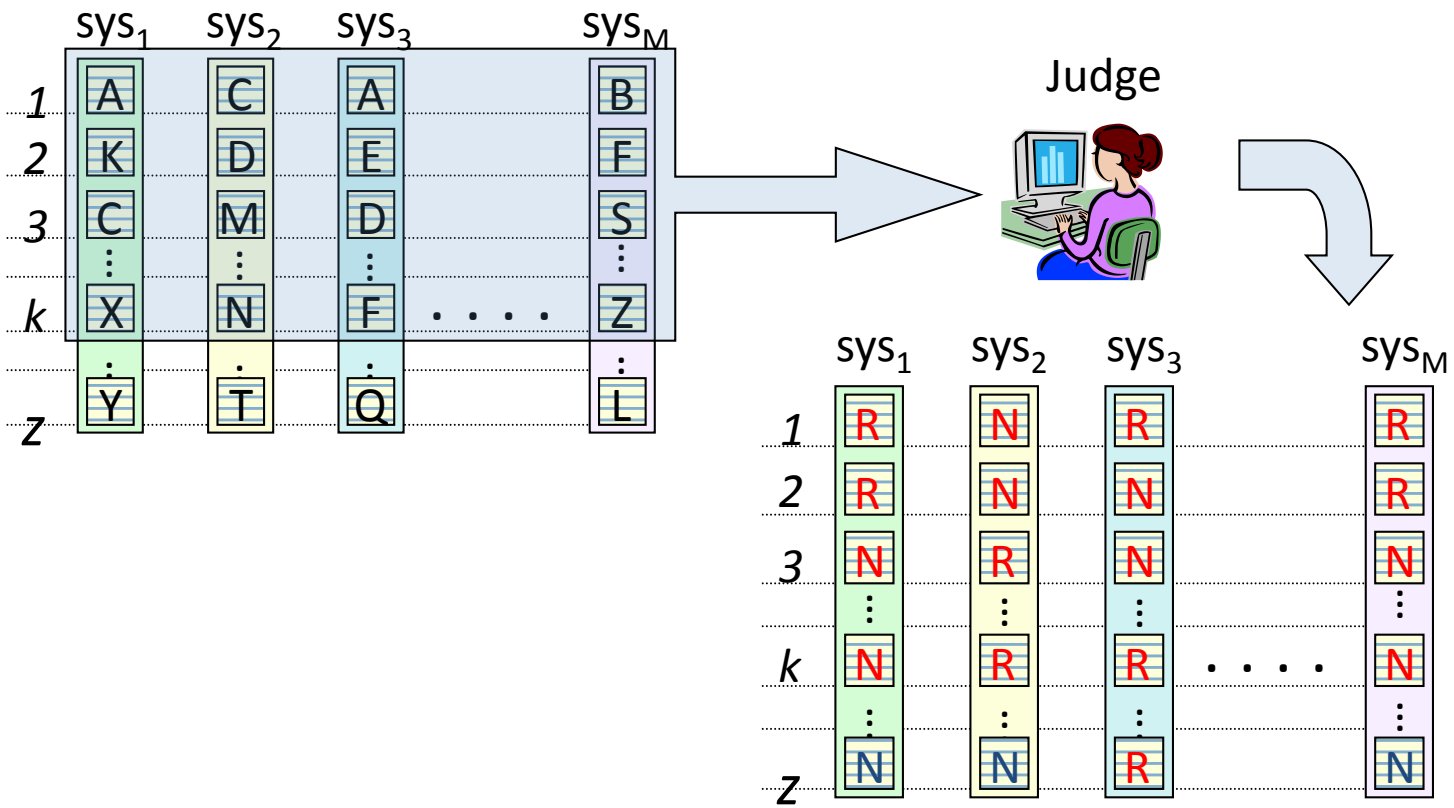
Reusability

- Reusability is hard to guarantee

- Can we test for reusability?
  - Leave-one-out

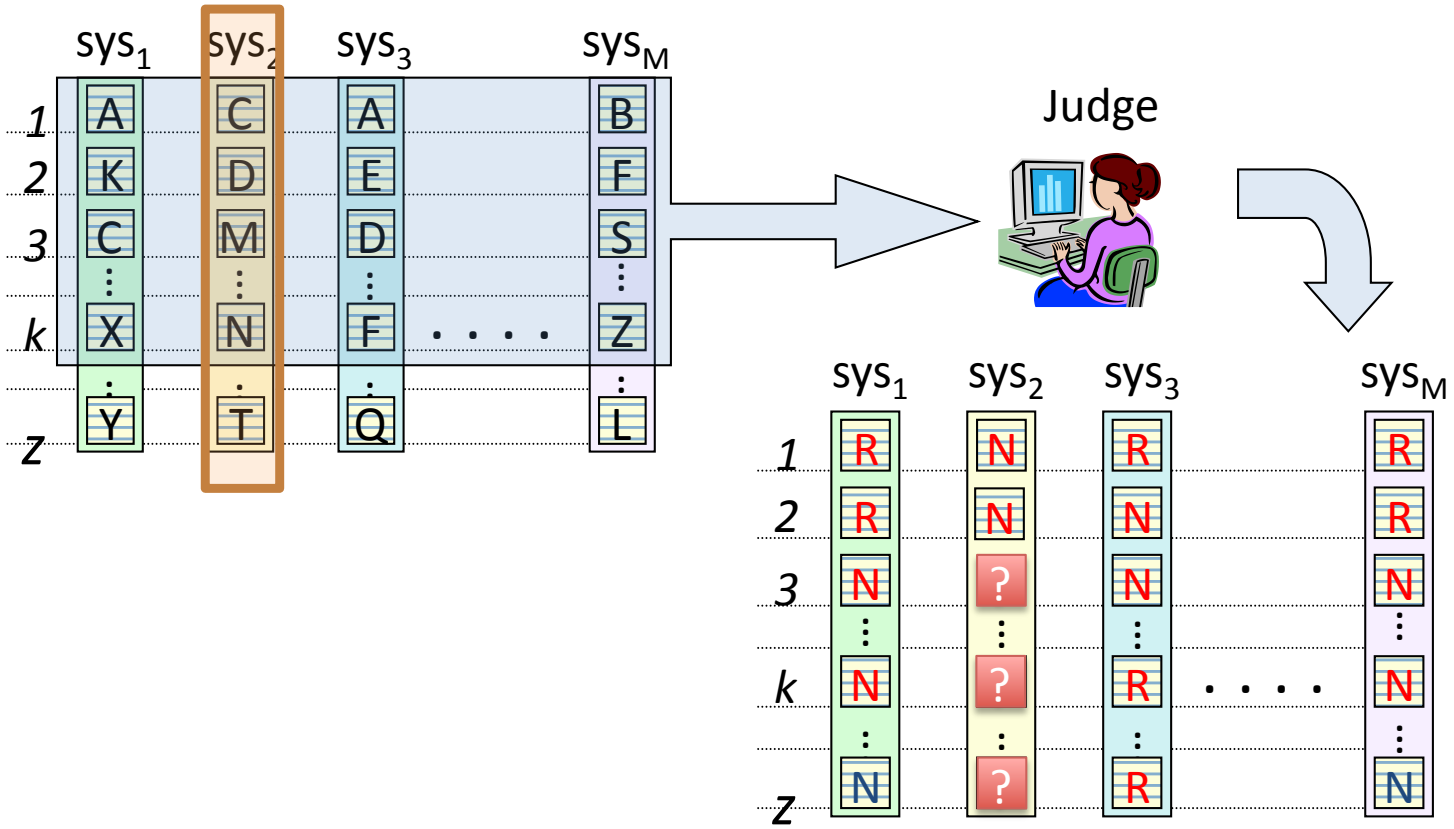
    [Zobel SIGIR98, Voorhees CLEF01, Sanderson and Joho SIGIR04, Sakai CIKM08]
  - Through Experimental Design
    - Carterette et al. SIGIR10

Reusability - Leave-one-out



Reusability - Leave-one-out

- Assume that a systems did not participate
- Evaluate the non-participating system with the rest of the judgments



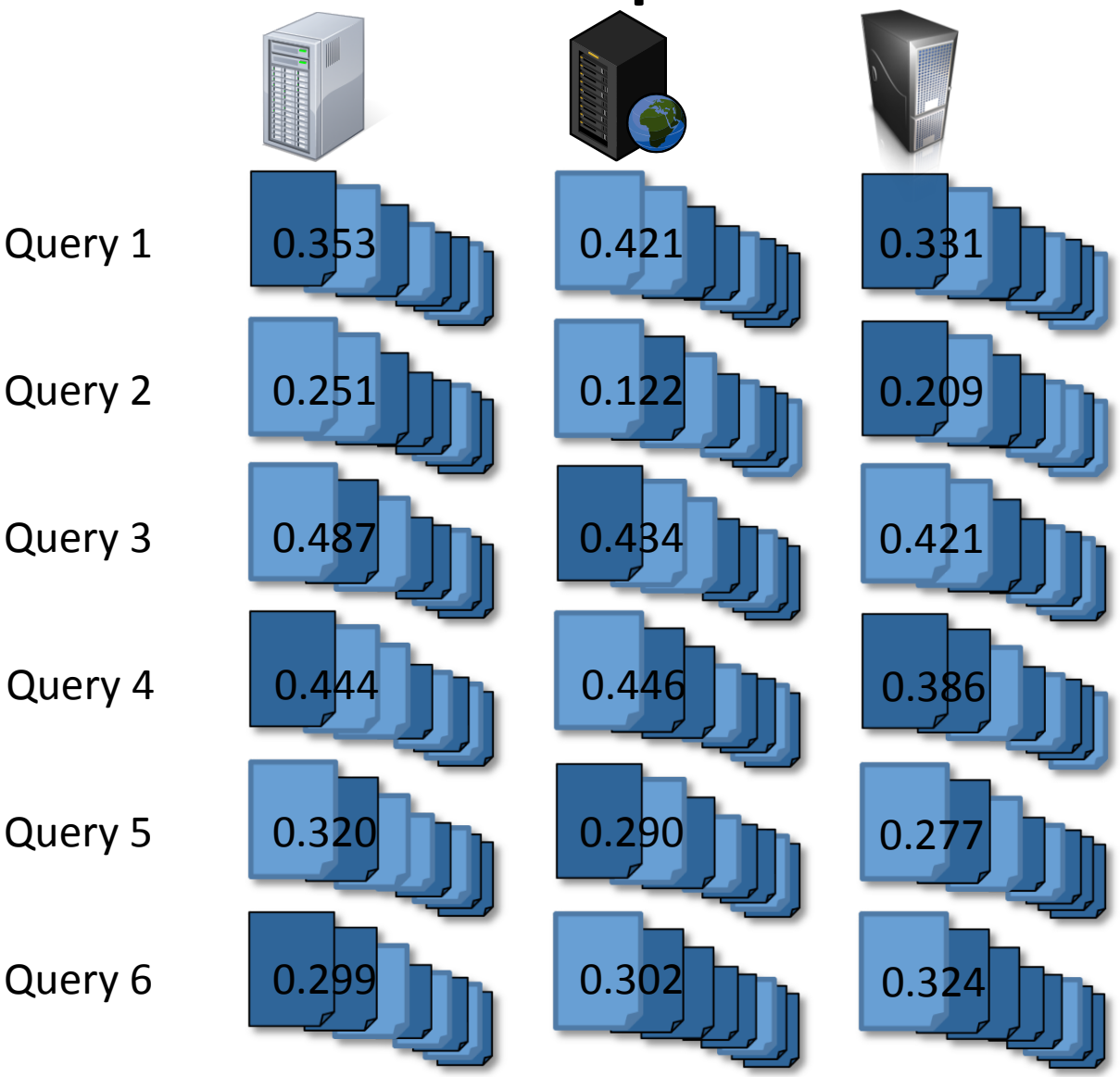
Reusability

- Reusability is hard to guarantee

- Can we test for reusability?
  - Leave-one-out

    [Zobel SIGIR98, Voorhees CLEF01, Sanderson and Joho SIGIR04, Sakai CIKM08]
  - Through Experimental Design
    - Carterette et al. SIGIR10

Low-Cost Experimental Design



Sampling and Reusability

- Does sampling produce a reusable collection?
  - We don't know…
  - … and we can't simulate it
- Holding systems out would produce a different sample
  - Meaning we would need judgments that we don't have

Experimenting on Reusability

- Our goal is to define an experimental design that will allow us to simultaneously:
  - Acquire relevance judgments
  - Test hypotheses about differences between systems
  - Test reusability of the topics and judgments

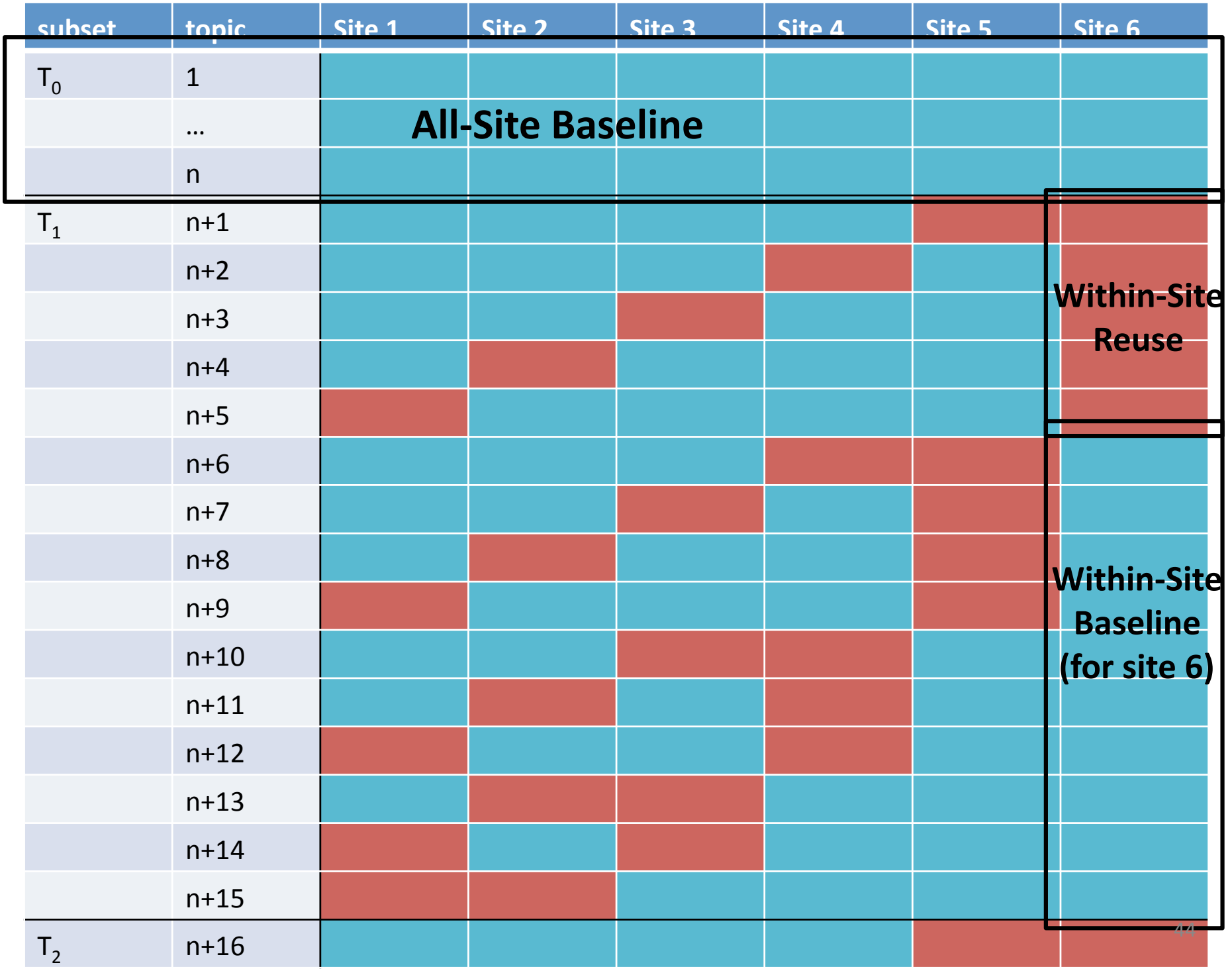
- What does it mean to "test reusability"?
  - Test a null hypothesis that the collection *is* reusable
  - Reject that hypothesis if the data demands it
  - Never *accept* that hypothesis

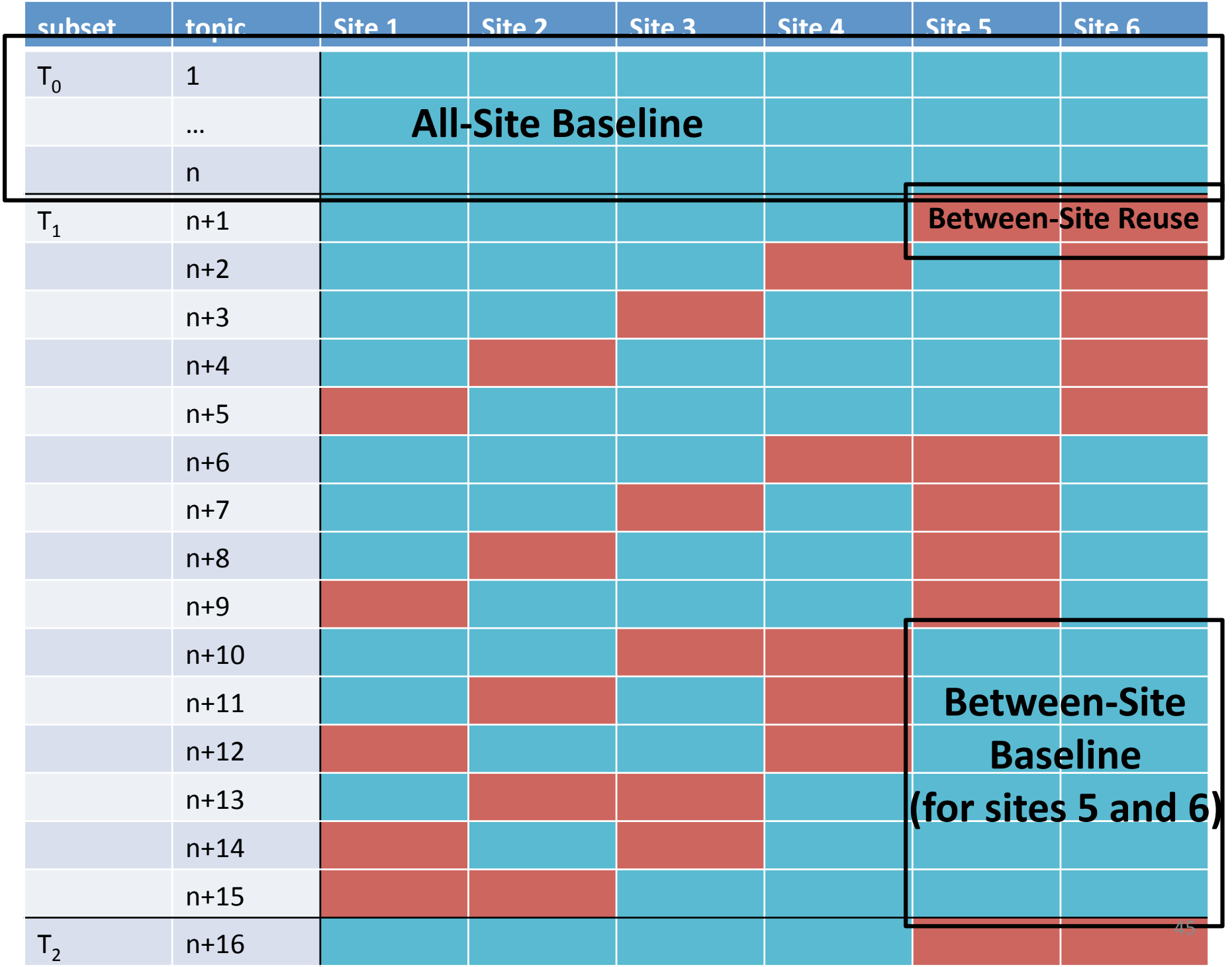
Reusability for Evaluation

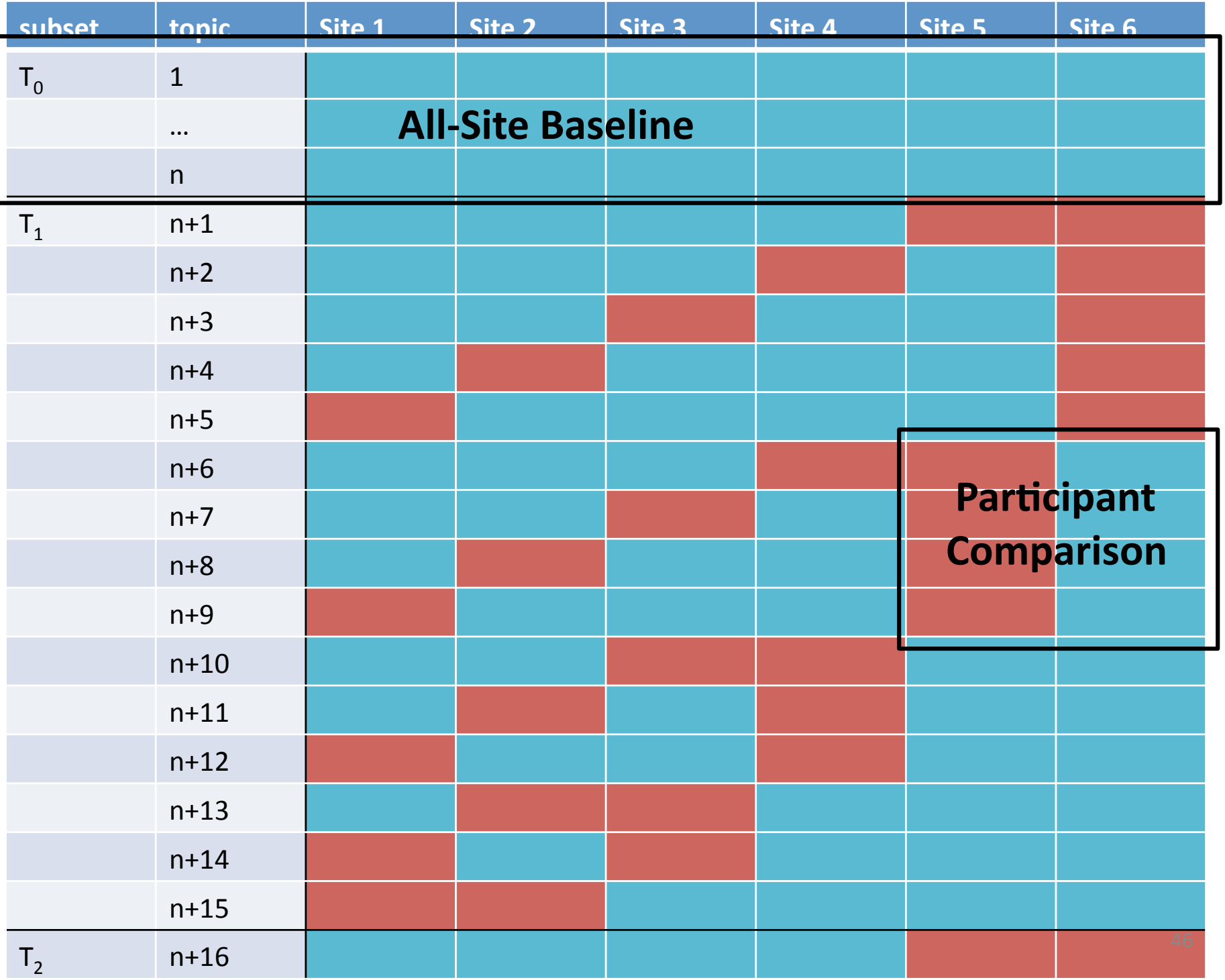
- We focus on evaluation (rather than training, failure analysis, etc)

- Three types of evaluation:
  - **Within-site**: a group wants to internally evaluate their systems
  - **Between-site**: a group wants to compare their systems to those of another group
  - **Participant-comparison**: a group wants to compare their systems to those that participated in the original experiment (e.g. TREC)

- We want data for each of these


subset
topic
Site 1
Site 2
Site 3
Site 4
Site 5
Site 6
T_0
1
…
n
All-Site Baseline
T_1
n+1
n+2
n+3
n+4
n+5
n+6
n+7
n+8
n+9
n+10
n+11
n+12
n+13
n+14
n+15
T_2
n+16
Within-Site Reuse
Within-Site Baseline (for site 6)


subset
topic
Site 1
Site 2
Site 3
Site 4
Site 5
Site 6
T_0
1
…
n
All-Site Baseline
T_1
n+1
n+2
n+3
n+4
n+5
n+6
n+7
n+8
n+9
n+10
n+11
n+12
n+13
n+14
n+15
T_2
n+16
Between-Site Reuse
Between-Site Baseline (for sites 5 and 6)


subset
topic
Site 1
Site 2
Site 3
Site 4
Site 5
Site 6
T_0
1
…
n
All-Site Baseline
T_1
n+1
n+2
n+3
n+4
n+5
n+6
n+7
n+8
n+9
n+10
n+11
n+12
n+13
n+14
n+15
T_2
n+16
Participant Comparison

Statistical Analysis

- Goal of statistical analysis is to try to reject the hypothesis about reusability
  - Show that the judgments are *not* reusable
- Three approaches:
  - Show that measures such as average precision on the baseline sets do not match measures on the reuse sets
  - Show that significance tests in the baseline sets do not match significance tests in the reuse sets
  - Show that rankings in the baseline sets do not match rankings in the reuse sets
- Note: *within confidence intervals!*

Agreement in Significance

- Perform significance tests on:
  - all pairs of systems in a baseline set
  - all pairs of systems in a reuse set

- If the aggregate outcomes of the tests disagree significantly, reject reusability

Within-Site Example

- Some site submitted five runs to the TREC 2004 Robust track

- Within-site baseline: 210 topics
- Within-site reuse: 39 topics

- Perform 5*4/2 = 10 paired t-tests with each group of topics
- Aggregate agreement in a contingency table

Within-Site Example

	baseline tests	
reuse tests	**$p < 0.05$**	**$p \geq 0.05$**
$p' < 0.05$	6	0
$p' \geq 0.05$	3	1

- 3 significant differences in baseline set that are not significant in reuse set
  - ➔ 70% agreement
- … is that bad?

Expected Errors

- Compare observed error rate to expected error rate

- To estimate expected error rate, use *power analysis* (Cohen, 1992)
  - What is the probability that the observed difference over 210 topics would be found significant?
  - What is the probability that the observed difference over 39 topics would be found significant?
  - Call these probabilities q_1, q_2

Expected Errors

- For each pair of runs:
  - q_1 = probability that observed difference is significant over 210 queries
  - q_2 = probability that observed difference is significant over 39 queries

  - Expected number of true positives += $q_1 * q_2$
  - Expected number of false positives += $q_1 * (1-q_2)$
  - Expected number of false negatives += $(1-q_1) * q_2$
  - Expected number of true negatives += $(1-q_1) * (1-q_2)$

Observed vs Expected Errors

- Observed:

	baseline tests	
reuse tests	p < 0.05	p ≥ 0.05
p' < 0.05	6	0
p' ≥ 0.05	3	1

- Expected:

	baseline tests	
reuse tests	p < 0.05	p ≥ 0.05
p' < 0.05	7.098	0.073
p' ≥ 0.05	2.043	0.786

- Perform a X^2 goodness-of-fit test to compare the tables
- p-value = 0.88
- Do not reject reusability (for new systems like these)

Validation of Design and Analysis

- Three tests:
  - Will we reject reusability when it is not true?
  - When reusability is “true”, will the design+analysis be robust to random differences in topic sets?
  - When reusability is “true”, will the design+analysis be robust to random differences in held-out sites?

Differences in Topic Samples

- Set-up: simulate design, but guarantee reusability
  - Randomly choose k sites to hold out
  - Use to define the baseline and reuse sets
  - Performance measure on each system/topic is simply the one calculated using the original judgments

- Reusability is true because all measures are exactly the same as when sites are held out

Observed vs Expected Errors (Within-Site)

- Observed:

	baseline tests	
reuse tests	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	196	2
$p' \geq 0.05$	57	45

- Expected:

	baseline tests	
reuse tests	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	189.5	4.3
$p' \geq 0.05$	62.1	44.1

- Perform a X^2 goodness-of-fit test to compare the tables
- p-value = 0.58
- Do not reject reusability (for new systems like these)

based on TREC 2008 Million Query track data

Differences in Held-Out Sites

- Set-up: simulation of design with TREC Robust data (249 topics, many judgments each)
  - Randomly hold two of 12 submitting sites out
  - Simulate pools of depth 10, 20, 50, 100
  - Calculate average precision over simulated pool
- Previous work suggests reusability is true
- (Only within-site analysis is possible)

Observed vs Expected Errors (Within-Site)

- Observed:

	baseline tests	
reuse tests	p < 0.05	p ≥ 0.05
p' < 0.05	130	17
p' ≥ 0.05	127	160

- Expected:

	baseline tests	
reuse tests	p < 0.05	p ≥ 0.05
p' < 0.05	135.4	13.9
p' ≥ 0.05	121.6	163.1

- Perform a X^2 goodness-of-fit test to compare the tables
- p-value = 0.74
- Do not reject reusability (for new systems like these)

based on TREC 2004 Robust track data

Course Outline

- Intro to evaluation
  - Evaluation methods, test collections, measures, comparable evaluation
- Low cost evaluation
- Advanced user models
  - Web search models, novelty & diversity, sessions
- Reliability
  - Significance tests, reusability
- Other evaluation setups

Evaluation using crowd-sourcing

- Sheng et al KDD 08, Bailey et al SIGIR08, Alonso and Mizarro SIGIR09, Kazai et al. SIGIR09, Yang et al WSDM09, Tang and Sanderson ECIR10, Sanderson et al SIGIR10

- Cheap but noisy judgments
- Large load under a single assessor per topic
  - Can you motivate an Mturker to judge ~1,500 documents
  - Multiple assessors (not MTurkers) per topics works fine [Trotman and Jenkinson ADCS07]
- Inconsistency across assessors
  - Malicious activity?
  - Noise?
  - Diversity in information needs (query aspects)?

Online Evaluation

- Joachims et al SIGIR05, Radlinski et al CIKM08, Wang et al KDD09, …

- Use clicks as indication of relevance?
- Rank bias
  - Users tend to click on documents at the top of the list independent of their relevance
- Quality bias
  - Users tend to click on less relevant documents if the overall quality of the search engine is poor

Online Evaluation

- Evaluate by watching user behaviour:
  - Real user enters a query
  - Record how the users respond
  - Measure statistics about these responses

- Common online evaluation metrics
  - Click-through rate
    - Assumes more clicks means better results
  - Queries per user
    - Assumes users come back more with better results
  - Probability user skips over results they have considered (pSkip)

Online Evaluation

- Interleaving [Radlinski et al CIKM08]
  - A way to compare rankers online
    - Given the two rankings produced by two methods
    - Present a combination of the rankings to users
    - Ranking providing more of the clicked results wins
  - Treat a flight as an active experiment

Team Draft Interleaving

Ranking A

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Valley Wineries - Plan your wine...
   www.napavalley.com/wineries
3. Napa Valley College
   www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley
   www.ivebeenthere.co.u
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

1. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...
   .com

Presented Ranking

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Wineries – Plan your wine...
   www.napavalley.com/wineries
5. Napa Valley Hotels – Bed and Breakfast...
   www.napalinks.com
6. Napa Valley College
   www.napavalley.edu/homex.asp
7. NapaValley.org
   www.napavalley.org

Click

Click

B wins!

Credit Assignment

- The "team" with more clicks wins
  - Randomization removes presentation order bias
- Each impression with clicks gives a preference for one of the rankings (unless there is a tie)
  - By design: If the input rankings are equally good, they have equal chance of winning
- Statistical test to run: ignoring ties, is the fraction of impressions where the new method wins statistically different from 50%?

Course Outline

- Intro to evaluation
  - Evaluation methods, test collections, measures, comparable evaluation
- Low cost evaluation
- Advanced user models
  - Web search models, novelty & diversity, sessions
- Reliability
  - Significance tests, reusability
- Other evaluation setups

By the end of this course…

You will be able to evaluate your retrieval algorithms

A. At low cost
B. Reliably
C. Effectively

Many thanks to Mark Sanderson @RMIT,
for some of the significance testing slides

Many thanks to Filip Radlinski @MSR,
for many of the online evaluation slides