

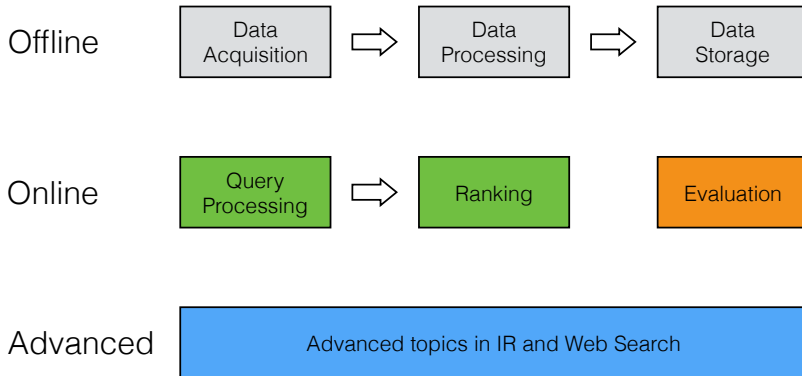
# Information Retrieval

## Online Evaluation and Statistical Testing

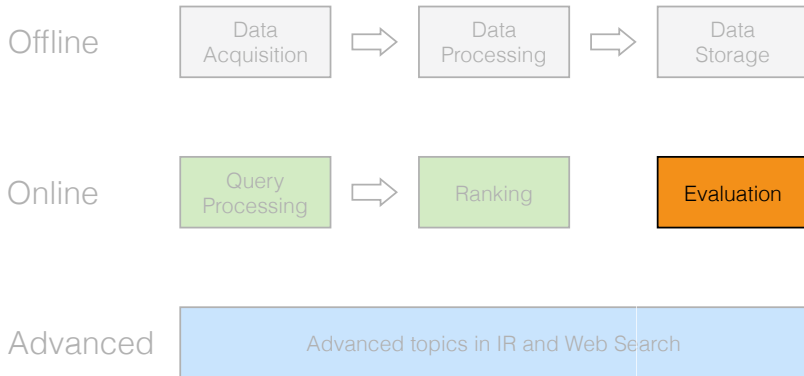
**Ilya Markov**  
i.markov@uva.nl

University of Amsterdam

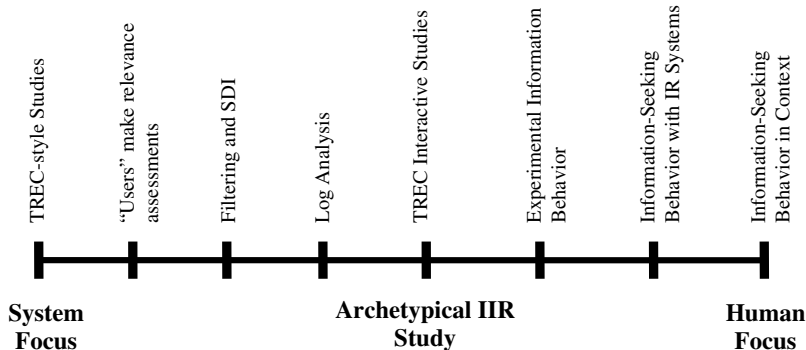
# Course overview



# This lecture



# Taxonomy of evaluation approaches



Diane Kelly, "Methods for Evaluating Interactive Information Retrieval Systems with Users"

# Online scenario

Yandex  ✕ ↵ Search

Web **W** **Information retrieval - Wikipedia, the free encyclopedia**  
 en.wikipedia.org > **Information retrieval** ▾  
 Information retrieval (IR) is the activity of obtaining **information** resources relevant to an information need from a collection of **information** resources. Searches can be based on full-text or on content-based indexing.

Images

Video

Translate

More

**I** **Information retrieval**  
 en.academic.ru > dic.nsf/enwiki/9176 ▾  
 1950: The term "**information retrieval**" appears to have been coined by Calvin Mooers. 1951: Philip Bagley conducted the earliest experiment in computerized document **retrieval** in a...

**P** **Information retrieval - Psychology Wiki - Wikia**  
 psychology.wikia.com > wiki/**Information\_retrieval** ▾  
 Assessment | Biopsychology | Comparative | Cognitive | Developmental | Language | Individual differences | Personality | Philosophy | Social | Methods | Statistics | Clinical | Educational | Industrial | Professional items | World psychology |.

**I** **Introduction to Information Retrieval**  
 nlp.stanford.edu > IR-book/ ▾  
 A list of **information retrieval** resources is also available. Introduction to **Information Retrieval** Contents.

**W** **Information retrieval - Wikiquote**  
 en.wikiquote.org > wiki/**Information\_retrieval** ▾  
**Information retrieval** is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources, and the part of **information** science, which studies of these activity.

# Outline

- 1 Online evaluation
  - Online metrics
  - Between-subject experiments
  - Within-subject experiments
  - Summary
- 2 Hypothesis testing

# Outline

- 1 Online evaluation
  - Online metrics
  - Between-subject experiments
  - Within-subject experiments
  - Summary

# Which user-generated signals indicate search quality?

Yandex  ✕ ↶ ↷ Search

Web **W** **Information retrieval - Wikipedia, the free encyclopedia**  
 en.wikipedia.org > **Information retrieval** ▾  
 Information retrieval (IR) is the activity of obtaining **information** resources relevant to an information need from a collection of **information** resources. Searches can be based on full-text or keyword-based indexing.

Images

Video

Translate

More

**I** **Information retrieval**  
 en.academic.ru > dic.nsf/enwiki/9176 ▾  
 1950: The term "**information retrieval**" appears to have been coined by Calvin Mooers. 1951: Philip Bagley conducted the earliest experiment in computerized document **retrieval** in a...

**P** **Information retrieval - Psychology Wiki - Wikia**  
 psychology.wikia.com > wiki/**Information\_retrieval** ▾  
 Assessment | Biopsychology | Comparative | Cognitive | Developmental | Language | Individual differences | Personality | Philosophy | Social | Methods | Statistics | Clinical | Educational | Industrial | Professional items | World psychology |.

**I** **Introduction to Information Retrieval**  
 nlp.stanford.edu > IR-book/ ▾  
 A list of **information retrieval** resources is also available. Introduction to **Information Retrieval** Contents.

**W** **Information retrieval - Wikiquote**  
 en.wikiquote.org > wiki/**Information\_retrieval** ▾  
**Information retrieval** is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources, and the part of **information** science, which studies of these activity.



# Classification of online measures

Evaluation Method	Absolute	Relative
Item Level	Click-through rate, ...	Click-skip, ...
SERP Level	Abandonment rate, ...	A/B Testing, Interleaving, ...

Evangelos Kanoulas, "A Short Survey on Online and Offline Methods for Search Quality Evaluation"

# Online metrics

Type of interaction	Metric	Good	Bad
Clicks	Click-through rate	↑	↓
	Click rank (reciprocal rank)	↓	↑
	Abandonment	↓	↑
Time	Dwell time	↑	↓
	Time to first click	↓	↑
	Time to last click	↑	↓
Queries	Number of reformulations	↓	↑
	Number of abandoned queries	↓	↑

# Outline

- 1 Online evaluation
  - Online metrics
  - **Between-subject experiments**
  - Within-subject experiments
  - Summary

## A/B testing

Control

Treatment

Google what is information retrieval

Web Videos Images Maps News More Search tools

About 16,000,000 results (0.52 seconds)

**Information retrieval - Wikipedia, the free encyclopedia**  
 en.wikipedia.org/wiki/Information\_retrieval -  
 Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. Standard Boolean model - Category:Information retrieval - Relevance

**[PDF] Introduction to Information Retrieval - The Stanford NLP**  
 nlp.stanford.edu/IR-book/pdf/01bool.pdf -  
 Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

**Information retrieval - Merriam-Webster Online**  
 www.merriam-webster.com/dictionary/information%20retrieval -  
 the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system ...

**CS533: Information Retrieval**  
 www.cs.cmu.edu/~david/cs533/cs533.html -  
 22 Dec 2014 Information Retrieval (IR) is about the process of providing answers to client information needs! It is thus concerned with the collection, ...

**Information retrieval - Wikiquote**  
 en.wikiquote.org/wiki/Information\_retrieval -  
 Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources, and the part of ...



Google what is information retrieval

Web Videos Images Maps News More Search tools

About 16,000,000 results (0.52 seconds)

**Information retrieval - Wikipedia, the free encyclopedia**  
 en.wikipedia.org/wiki/Information\_retrieval -  
 Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Overview - Model types - Performance and ... - Awards in the field

**What is information retrieval? - Yahoo Answers**  
 https://answers.yahoo.com/question/index?qid=1006041603723 -  
 Resolved · 4 posts · 3 total answers  
 16/04/2006 · Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources ...

**Information Retrieval definition of Information Retrieval ...**  
 encyclopaedia2.thefreedictionary.com/information+retrieval -  
 Information retrieval [in-fair-má-shan ri-bré-val] (computer science) The technique and process of searching, recovering, and interpreting information ...

**What Is Data Retrieval? | eHow | eHow | How to - Discover ...**  
 www.ehow.com | Internet | On the Web | Online Research -  
 28/08/2014 · What Is Data Retrieval?. "Data retrieval" refers to various processes, including recovery of lost information, gathering information on an unknown ...

**Introduction to Information Retrieval - Stanford University**  
 nlp.stanford.edu/IR-book -  
 Introduction to Information Retrieval. This is the companion website for the following book. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze ...

Google what is information retrieval

Web Videos Images Maps News More Search tools

About 16,000,000 results (0.52 seconds)

**Information retrieval - Wikipedia, the free encyclopedia**  
 en.wikipedia.org/wiki/Information\_retrieval -  
 Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Overview - Model types - Performance and ... - Awards in the field

**[PDF] Introduction to Information Retrieval - The Stanford NLP**  
 nlp.stanford.edu/IR-book/pdf/01bool.pdf -  
 Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

**Information retrieval - Merriam-Webster Online**  
 www.merriam-webster.com/dictionary/information%20retrieval -  
 the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system ...

**CS533: Information Retrieval**  
 www.cs.cmu.edu/~david/cs533/cs533.html -  
 22 Dec 2014 Information Retrieval (IR) is about the process of providing answers to client information needs! It is thus concerned with the collection, ...

**Information retrieval - Wikiquote**  
 en.wikiquote.org/wiki/Information\_retrieval -  
 Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources, and the part of ...

# A/B testing

- ① Set the current search system as control
- ② Set an alternative search system as treatment
- ③ Assign 0.5%-1.0% of users to each of these systems
- ④ Record user interactions with these systems during time period  $T$
- ⑤ Compare the systems using online metrics
- ⑥ Choose a winner based on one or several metrics

# A/B testing, practical considerations

- Choosing metrics
- Control extraneous factors
- Estimate adequate sample size
- Novelty impact
- Etc.

# A/B testing discussion

- Pros
  - Can evaluate anything
  - Using any online metric
- Cons
  - High variance between users
  - Not very sensitive
  - Needs lots of observations

# Outline

- 1 Online evaluation
  - Online metrics
  - Between-subject experiments
  - Within-subject experiments
  - Summary

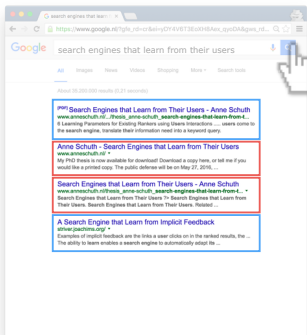


# Interleaving

- ① Given a user's query, produce two rankings (current and alternative)
- ② Merge the rankings into a single ranking using a **mixing policy**
- ③ Present the merged ranking to a user and collect interactions (see online metrics)
- ④ Choose a winning ranking using a **scoring rule**
- ⑤ Repeat steps 1–4 until a clear winner is identified

# Team draft interleaving

Google A



Google B

1 Search Engines that Learn from Implicit Feedback - Depar...  
www.cs.uq.ni/~te/courses/OnIml/Jacobs-Search-Engines.pdf \*  
by T. Jacobs - Cited by 161 - Related articles  
Aug 2, 2007 - search engines present results heavily biased a user's. Search engine logs provide a ... that employees who search their company interest for.

5 Alternative Search Engines That Respect Your Privacy - Ho...  
www.howlogick.com/.../5-alternative-search-engines-that-respect-your-priv...  
May 8, 2012 - Google, Bing, Yahoo - all the major search engines track your search ... you, and it discards user agents and IP addresses from its server logs.

5 Alternative Search Engines That Respect Your Privacy - Ho...  
www.howlogick.com/.../5-alternative-search-engines-that-respect-your-priv...  
May 8, 2012 - Google, Bing, Yahoo - all the major search engines track your search ... you, and it discards user agents and IP addresses from its server logs.

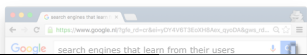
1 Search Engines that Learn from Implicit Feedback - Depar...  
www.cs.uq.ni/~te/courses/OnIml/Jacobs-Search-Engines.pdf \*  
by T. Jacobs - Cited by 161 - Related articles  
Aug 2, 2007 - search engines present results heavily biased a user's. Search engine logs provide a ... that employees who search their company interest for.

Search Engines that Learn from Their Users - Anne Schuth  
www.amschuth.nl/files\_anna-schuth\_search-engines-that-learn-from-4...  
Search Engines that Learn from Their Users. Search Engines that Learn from Their Users. Related.

1 Search Engines that Learn from Implicit Feedback - Depar...  
www.cs.uq.ni/~te/courses/OnIml/Jacobs-Search-Engines.pdf \*  
by T. Jacobs - Cited by 161 - Related articles  
Aug 2, 2007 - search engines present results heavily biased a user's. Search engine logs provide a ... that employees who search their company interest for.

A Search Engine that Learn from Implicit Feedback  
drive.google.com/...  
Examples of implicit feedback are the links a user clicks on in the ranked results, the ... The ability to learn enables a search engine to automatically adapt its ...

5 Alternative Search Engines That Respect Your Privacy - Ho...  
www.howlogick.com/.../5-alternative-search-engines-that-respect-your-priv...  
May 8, 2012 - Google, Bing, Yahoo - all the major search engines track your search ... you, and it discards user agents and IP addresses from its server logs.



# Team draft interleaving

- **Mixing policy:** each ranker selects its highest ranked document that is not yet in the combined list
- **Scoring rule:** a ranker is preferred if its results get more clicks

# Other interleaving methods

- Probabilistic interleaving
- Optimized interleaving
- Multileaving

# Interleaving discussion

- Pros
  - No variance due to different users
  - Highly sensitive
  - Needs much fewer observations compared to A/B testing
- Cons
  - Can only use document-level metrics

# Outline

- 1 Online evaluation
  - Online metrics
  - Between-subject experiments
  - Within-subject experiments
  - Summary

# Online evaluation summary

- Online metrics
- Between-subject experiments – A/B testing
- Within-subject experiments – interleaving

# What are the advantages of online evaluation?

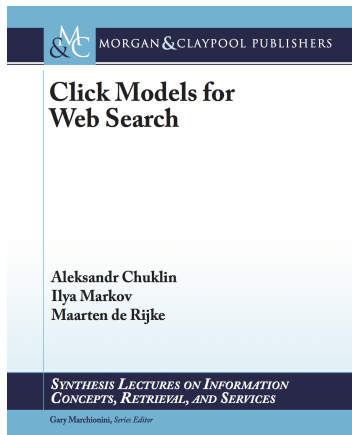
- Based on real users
- Cheap as it uses a running search system



# What are the disadvantages of online evaluation?

- Online metrics are difficult to interpret
- May disturb users
- Cannot run too many experiments in parallel
- User search interactions are biased

# Click models



<http://clickmodels.weebly.com/the-book.html>

# Materials

K. Hofmann, L. Li, F. Radlinski

**Online Evaluation for Information Retrieval**

Foundations and Trends in Information Retrieval, 2016

# Evaluating efficiency

Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.
Query throughput	Number of queries processed per second.
Query latency	The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.
Indexing temporary space	Amount of temporary disk space used while creating an index.
Index size	Amount of storage necessary to store the index files.

Croft et al., "Search Engines, Information Retrieval in Practice"

# Outline

- 1 Online evaluation
- 2 Hypothesis testing
  - Basics
  - Hypothesis testing in IR

# Example

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	-2
9	49	58	9
10	50	75	25

- How can we be sure that B is better than A?
- Test statistical significance (hypothesis testing)

Croft et al., "Search Engines, Information Retrieval in Practice"

# Outline

- 2 Hypothesis testing
  - Basics
  - Hypothesis testing in IR

# Test procedure

- ① Set null hypothesis  $H_0$
- ② Set alternative hypothesis  $H_1$
- ③ Collect sample data  $\mathbf{X} = \{X_1, \dots, X_n\}$ 
  - $\mathbf{X}$  is **unlikely** under  $H_0 \implies$  reject  $H_0$  in favor of  $H_1$
  - $\mathbf{X}$  is **not unlikely** under  $H_0 \implies$  no evidence against  $H_0$
  - ... but this is not an evidence in favor of  $H_0$ !



# Example

- ①  $H_0 : p_{head} = 0.8$
- ②  $H_1 : p_{head} \neq 0.8$
- ③ Perform 10 tosses, observe 4 heads

$$\frac{n!}{h!(n-h)!} p^h (1-p)^{n-h} = \frac{10!}{4!6!} 0.8^4 0.2^6 = 0.005$$

- ④ Perform 10 tosses, observe 7 heads

$$\frac{10!}{7!3!} 0.8^7 0.2^3 = 0.201$$

## Test procedure (cont'd)

- 1 Consider a statistical model
- 2 Set  $H_0$  and  $H_1$
- 3 Choose a test statistics  $T(X_1, \dots, X_n)$
- 4 Choose a critical region  $C$  (discussed next)
- 5 Decision rule
  - $T \in C \implies$  reject  $H_0$  in favor of  $H_1$
  - $T \notin C \implies$  fail to reject  $H_0$

## Example (cont'd)

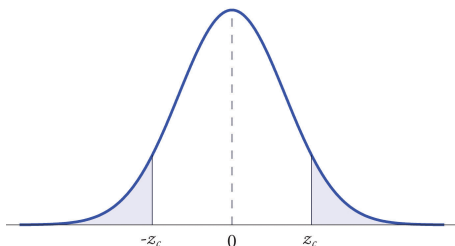
- $H_0 : p_{head} = 0.8$
- $H_1 : p_{head} \neq 0.8$
- Sample mean  

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$
- For large enough  $n$   

$$\bar{X} | H_0 \sim \mathcal{N}(0.8, \sigma^2/n)$$
- Test statistics  

$$T = \frac{\bar{X} - 0.8}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$
- Critical region  

$$C = (-\infty, -z_c] \cup [z_c, \infty)$$



- $P(T \in C | H_0) \leq \alpha$
- $\alpha$  - size, usually  $\in \{0.01, 0.05\}$
- For  $\alpha = 0.05$ ,  $z_{\alpha/2} = 1.96$

Picture taken from <http://2012books.lardbucket.org/books/beginning-statistics/s09-04-areas-of-tails-of-distribution.html>

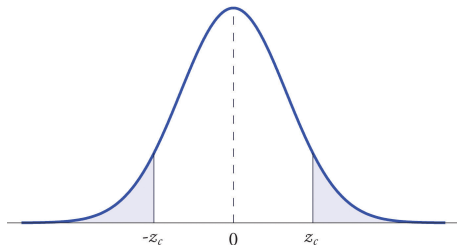
## Example (cont'd)

Reject  $H_0$  if

- $T \leq -z_{\alpha/2}, T \geq z_{\alpha/2}$
- $\bar{X} \leq -z_{\alpha/2} \cdot \sigma / \sqrt{n} + 0.8,$   
 $\bar{X} \geq z_{\alpha/2} \cdot \sigma / \sqrt{n} + 0.8$

Alternatively, reject  $H_0$  if

- $p = P(|T| > T_{obs} | H_0) \leq \alpha$
- $p$ -value



Test statistics and  $p$ -value can be calculated  
using any statistical software, e.g., R

# Errors

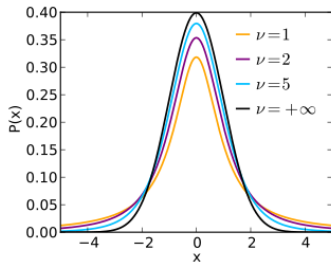
	$H_0$ is false	$H_0$ is true
Reject $H_0$	power	type I error ( $\alpha$ )
Not reject $H_0$	type II error	

# Outline

- 2 Hypothesis testing
  - Basics
  - Hypothesis testing in IR

# T-test

- ① Get measurements for systems A and B  
 $M(A) \sim \mathcal{N}(\mu_A, \sigma^2)$ ,  $M(B) \sim \mathcal{N}(\mu_B, \sigma^2)$
- ②  $H_0 : \mu_A = \mu_B$
- ③  $H_1 : \mu_A \neq \mu_B$
- ④  $T = \frac{\bar{A} - \bar{B}}{\hat{\sigma} / \sqrt{n}} \sim \mathcal{T}^{(n-1)}$  –  
 Student's t-distribution
- ⑤ Use standard hypothesis testing procedure  
 with  $\alpha \in \{0.01, 0.05\}$



Picture taken from [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution](https://en.wikipedia.org/wiki/Student%27s_t-distribution)

# T-test example

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	-2
9	49	58	9
10	50	75	25

- $\bar{B} - \bar{A} = 21.4$
- $\hat{\sigma} = 29.1$
- $T = \frac{21.4}{29.1/\sqrt{10}} = 2.33$
- $p = P(|T| > 2.33 \mid H_0) = 0.02$
- If  $\alpha = 0.05$ , reject  $H_0$
- If  $\alpha = 0.01$ , do not reject  $H_0$

Croft et al., "Search Engines, Information Retrieval in Practice"



# Wilcoxon signed-ranks test

- 1 Get measurements for systems A and B
- 2 For each item  $i$ , compute  $|m_{A,i} - m_{B,i}|$  and  $\text{sgn}(m_{A,i} - m_{B,i})$
- 3 Exclude items with  $|m_{A,i} - m_{B,i}| = 0$
- 4 Order the remaining  $N_{nz}$  items based on  $|m_{A,i} - m_{B,i}|$
- 5 Assign ranks  $R_i$  from smallest to largest
- 6 Compute the test statistics

$$W = \sum_{i=1}^{N_{nz}} [\text{sgn}(m_{A,i} - m_{B,i}) \cdot R_i]$$

- 7 For large  $N_{nz}$ ,  $W \sim \mathcal{N}$
- 8 Use standard hypothesis testing procedure with  $\alpha \in \{0.05, 0.01\}$

# Wilcoxon test example

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	-2
9	49	58	9
10	50	75	25

- Ranked non-zero differences  
2, 9, 10, 24, 25, 25, 41, 60, 70
- Signed ranks  
-1, +2, +3, -4, +5.5, +5.5, +7, +8, +9
- $W = 35$
- $p = P(|W| > 35 \mid H_0) = 0.025$
- If  $\alpha = 0.05$ , reject  $H_0$
- If  $\alpha = 0.01$ , do not reject  $H_0$

Croft et al., "Search Engines, Information Retrieval in Practice"

# Hypothesis testing summary

- IR must use statistical testing
- The most common and one of the most powerful is the **paired t-test**

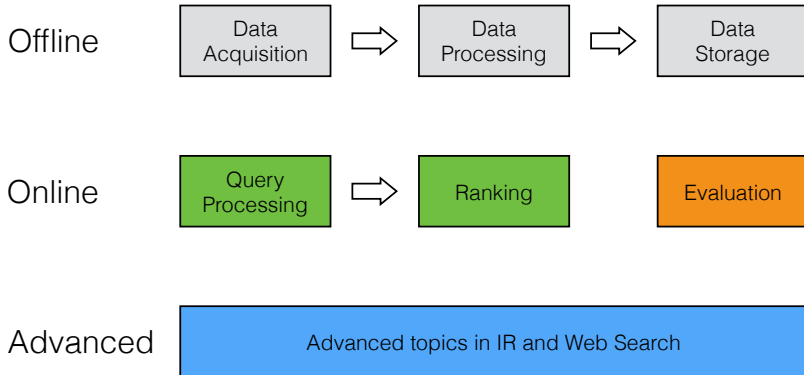
# Materials

M. Smucker, J. Allan, B. Carterette

## **A Comparison of Statistical Significance Tests for Information Retrieval Evaluation**

Proceedings of CIKM, pages 623–632, 2007

# Course overview



# See you tomorrow at 14:30

