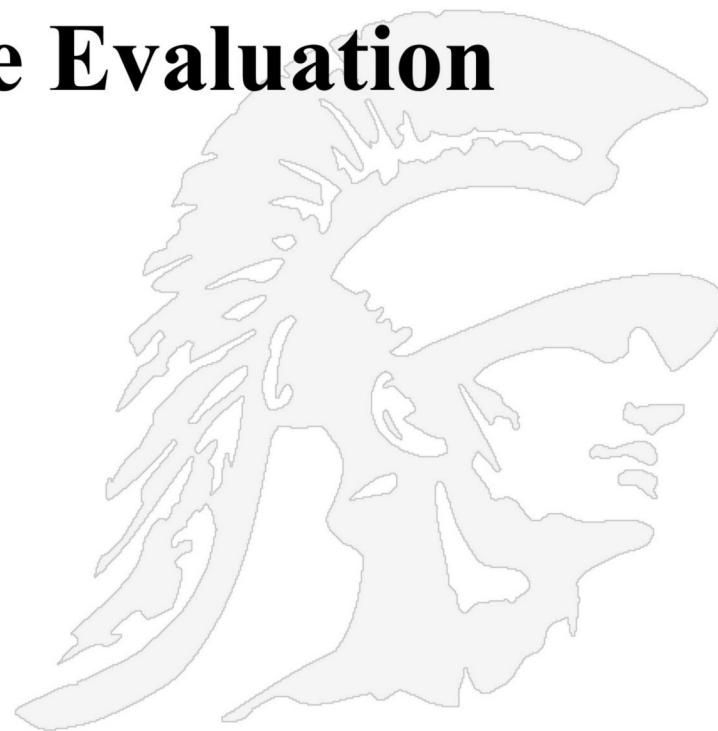


# Search Engine Evaluation



Copyright Ellis Horowitz 2011-2022

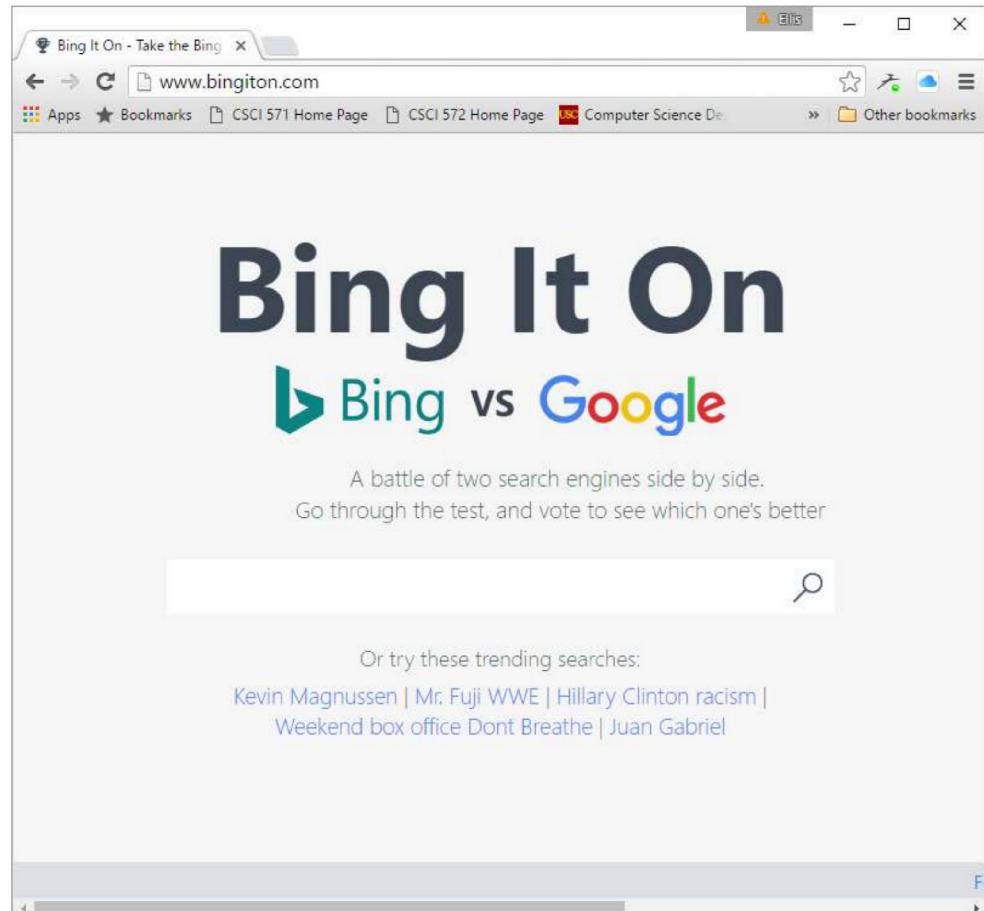
---

# Outline

- Defining precision/recall
- Mean Average Precision
- Harmonic Mean and F Measure
- Discounted Cumulative Gain
- Elements of Good Search Results
- Google's Search Quality Guidelines
- Using log files for evaluation
- A/B Testing



# Comparing Bing and Google



A screenshot of a web browser window titled "Bing It On - Take the Bing". The address bar shows "www.bingiton.com". The page content features a large title "Bing It On" with a subtitle "Bing vs Google". Below this, a subtext reads "A battle of two search engines side by side. Go through the test, and vote to see which one's better". There is a search bar with a magnifying glass icon. Underneath the search bar, there is a section titled "Or try these trending searches:" followed by a list of queries: "Kevin Magnussen | Mr. Fuji WWE | Hillary Clinton racism | Weekend box office Dont Breathe | Juan Gabriel". The browser interface includes standard navigation buttons (back, forward, search) and a toolbar with various icons.

- This site is no longer active, but we can simulate the experiment

## Try it yourself!

here are some queries:

- ac versus dc current
- best bottled water
- worst hotel in Santa Monica
- how many gears do I need on a bicycle
- Clint Eastwood's best movie

- How do we measure the quality of search engines?
- Precision = #(relevant items retrieved)  
divided by  
#(all retrieved items)
- Recall = #(relevant items retrieved)  
divided by  
#(all relevant items)



# Formalizing Precision/Recall

	Relevant	Nonrelevant
Retrieved	True positive (tp)	False positive (fp)
Not retrieved	False negative (fn)	True negative (tn)

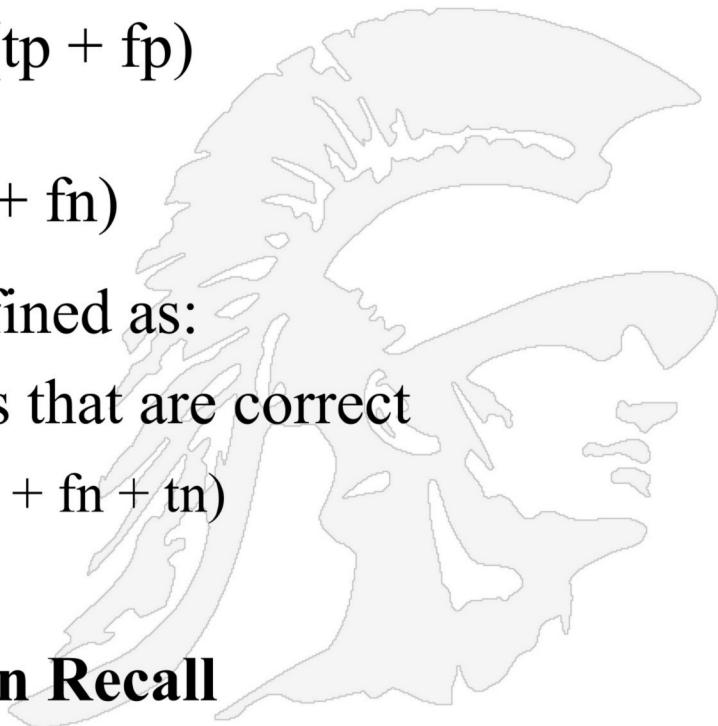
$$\text{Precision} = \text{tp}/(\text{tp} + \text{fp})$$

$$\text{Recall} = \text{tp}/(\text{tp} + \text{fn})$$

- The accuracy of an engine is defined as:  
the fraction of these classifications that are correct

$$(\text{tp} + \text{tn}) / (\text{tp} + \text{fp} + \text{fn} + \text{tn})$$

**For web applications,  
Precision is more important than Recall**



# Precision/Recall Using Set Notation

$A$  is set of relevant documents,  
 $B$  is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

you may not be able  
to see them, but  
 $A$  and  $B$  have a bar  
over them and it  
denotes the  
complement set

$$Recall = \frac{|A \cap B|}{|A|}$$

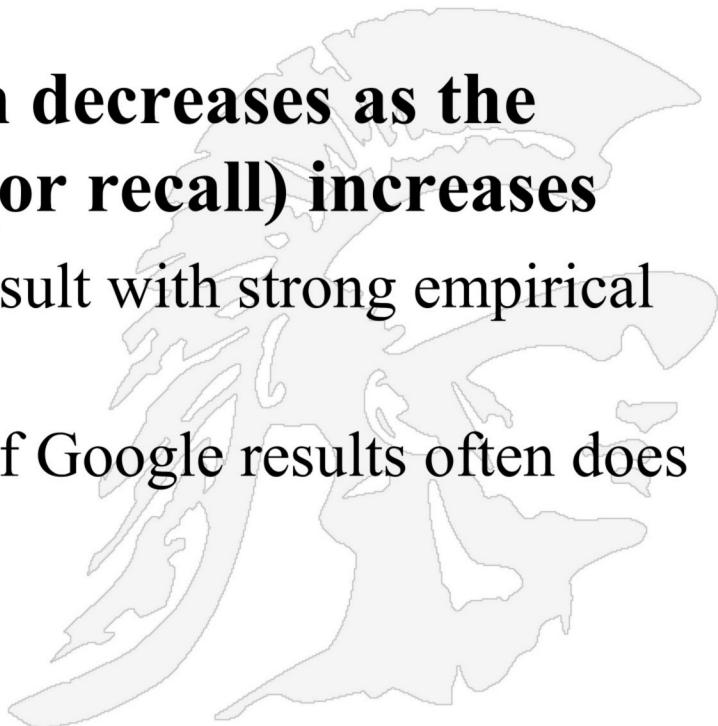
$$Precision = \frac{|A \cap B|}{|B|}$$

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)

# Precision/Recall

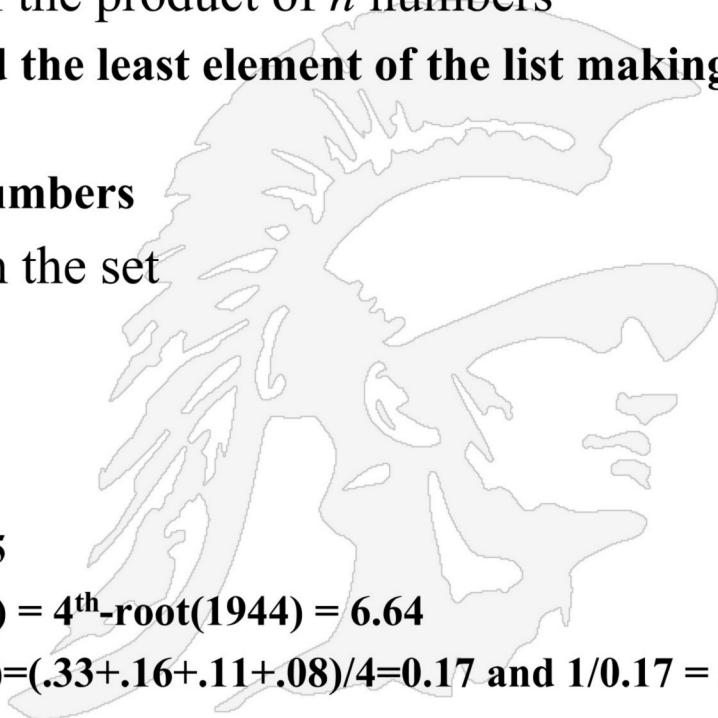
## Two Observations

- You can get high recall (but low precision) by retrieving all docs for all queries!
  - a rather foolish strategy
- In a good system, precision decreases as the number of docs retrieved (or recall) increases
  - This is not a theorem, but a result with strong empirical confirmation
  - E.g. viewing multiple pages of Google results often does not improve precision at all



# Harmonic Mean

- **There are three Pythagorean means**
  - 1. *arithmetic mean*, 2. *geometric mean*, 3. *harmonic mean*
  - of course we all know how to compute the arithmetic mean
  - the geometric mean is the  $n$ th root of the product of  $n$  numbers
- **The harmonic mean tends strongly toward the least element of the list making it useful in analyzing search engine results**
- **To find the harmonic mean of a set of  $n$  numbers**
  1. add the reciprocals of the numbers in the set
  2. divide the sum by  $n$
  3. take the reciprocal of the result
- **e.g. for the numbers 3, 6, 9, and 12**
  - **The arithmetic mean is:**  $(3+6+9+12)/4 = 7.5$
  - **The geometric mean is:**  $\text{nth-root}(3*6*9*12) = \text{4}^{\text{th}}\text{-root}(1944) = 6.64$
  - **The harmonic mean is:**  $(1/3+1/6+1/9+1/12)=(.33+.16+.11+.08)/4=0.17 \text{ and } 1/0.17 = 5.88$



# F Measure

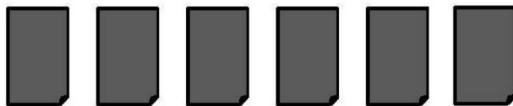
- The harmonic mean of the precision and the recall is often used as an aggregated performance score for the evaluation of algorithms and systems: called the **F-score (or F-measure)**.
- *Harmonic mean* of recall and precision is defined as

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R} + \frac{1}{P})} = \frac{2RP}{(R+P)}$$

- harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large
- More general form of F-Measure
  - $\beta$  is a parameter that controls the relative importance of recall and precision

$$F_\beta = (\beta^2 + 1)RP / (R + \beta^2 P)$$

# Calculating Recall/Precision at Fixed Positions



= the relevant documents

## Steps

Recall:

$$1/6 = 0.17$$

Precision:

$$1/1$$

Recall:

$$1/6 = 0.17$$

Precision:

$$1/2 = 0.5$$

Recall:

$$2/6 = 0.33$$

Precision:

$$2/3 = 0.67$$

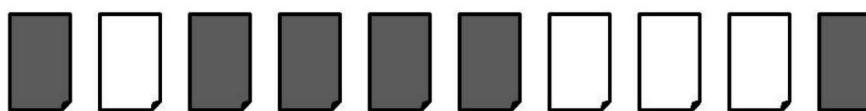
Recall:

$$3/6 = 0.5$$

Precision:

$$3/4 = 0.75$$

Ranking #1

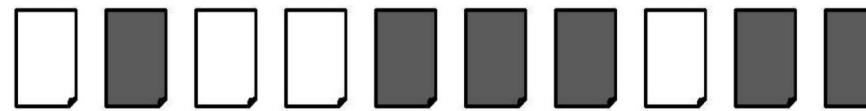


e.g.  
Google  
Result

	Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	1.0
--	--------	------	------	------	-----	------	------	------	------	-----

	Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6
--	-----------	-----	-----	------	------	-----	------	------	------	------	-----

Ranking #2



e.g.  
Bing  
Result

	Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
--	--------	-----	------	------	------	------	-----	------	------	------	-----

	Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6
--	-----------	-----	-----	------	------	-----	-----	------	-----	------	-----

*Recall = #RelevItemsRetr/allRelevItems*

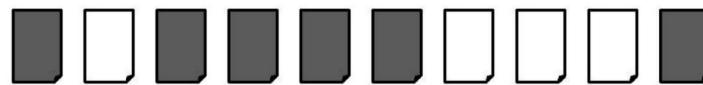
*Prec = #RelevItemsRetr/allItemsRetr*

# Average Precision of the Relevant Documents

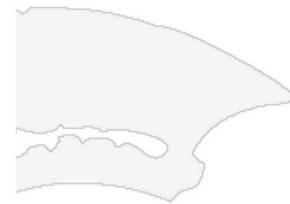


computes the sum of the precisions of the relevant documents

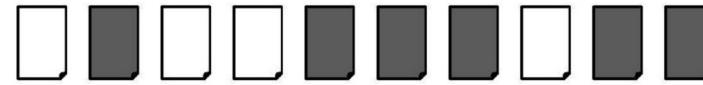
Ranking #1



	Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	1.0	
	Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6



Ranking #2



	Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
	Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

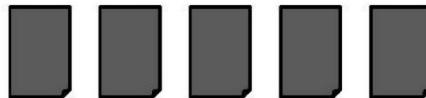


→ Ranking #1:  $(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78$

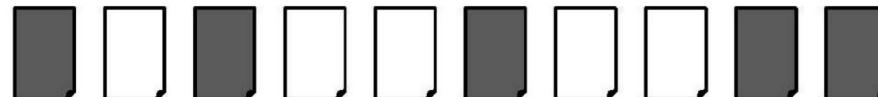
→ Ranking #2:  $(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6) / 6 = 0.52$

Conclusion: Ranking #1 for this query is best

# Averaging Across Queries

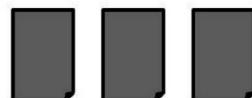
 = relevant documents for query 1

Ranking #1

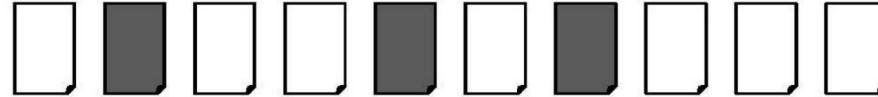


Recall 0.2 0.2 0.4 0.4 0.4 0.6 0.6 0.6 0.8 1.0

Precision 1.0 0.5 0.67 0.5 0.4 0.5 0.43 0.38 0.44 0.5

 = relevant documents for query 2

Ranking #2

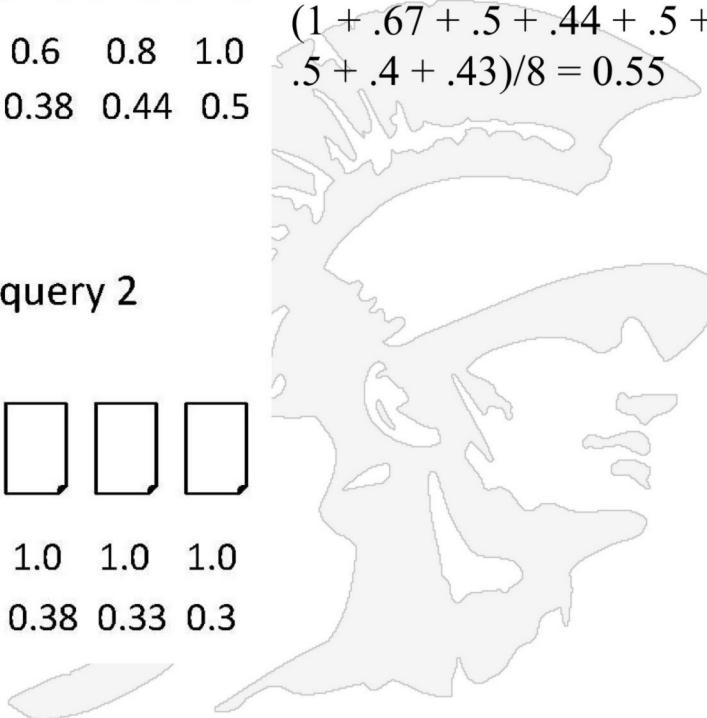


Recall 0.0 0.33 0.33 0.33 0.67 0.67 1.0 1.0 1.0 1.0

Precision 0.0 0.5 0.33 0.25 0.4 0.33 0.43 0.38 0.33 0.3

Average precision across the two queries for relevant docs is:

$$(1 + .67 + .5 + .44 + .5 + .5 + .4 + .43)/8 = 0.55$$



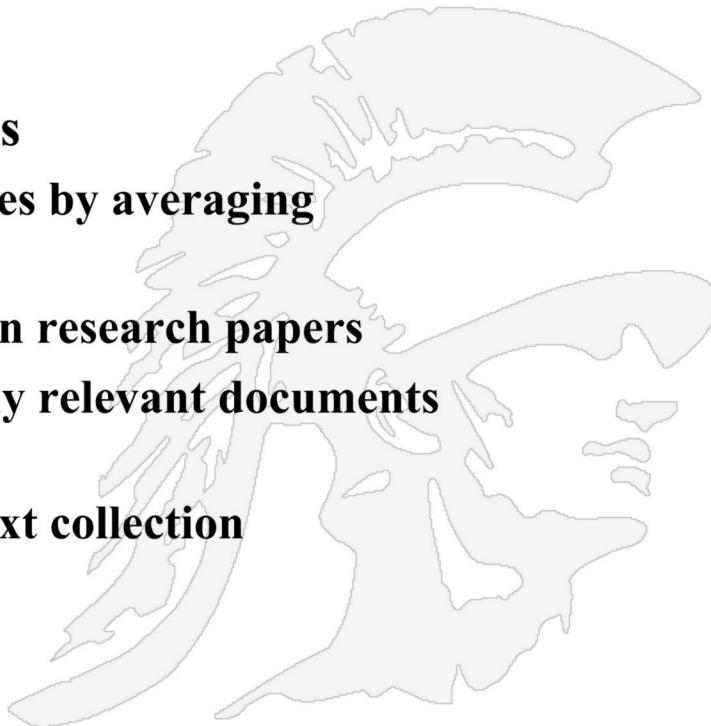
# Averaging Across Queries

- ***Mean average precision (MAP)*** for a set of queries is the mean of the average precision scores for each query.

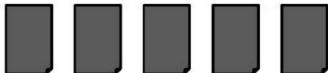
$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q}$$

where  $Q$  is the number of queries

- Summarize rankings from multiple queries by averaging average precision
- This is the *most commonly used* measure in research papers
- Assumes user is interested in finding many relevant documents for each query
- Requires many relevance judgments in text collection



# Mean Average Precision Example

 = relevant documents for query 1

Ranking #1    

	Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
--	--------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

	Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5
--	-----------	-----	-----	------	-----	-----	-----	------	------	------	-----

 = relevant documents for query 2

Ranking #2    

	Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
--	--------	-----	------	------	------	------	------	-----	-----	-----	-----

	Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3
--	-----------	-----	-----	------	------	-----	------	------	------	------	-----

$$\text{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$

$$\text{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

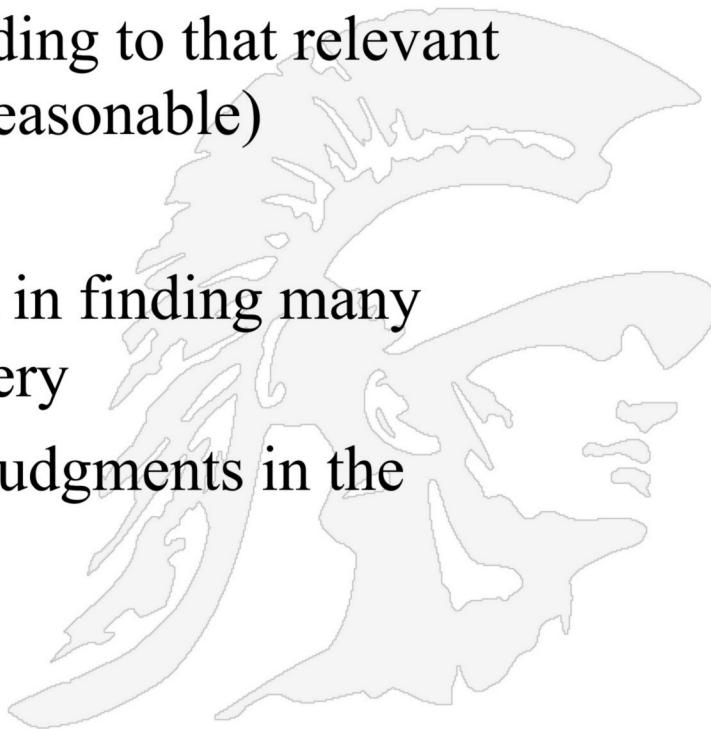
$$\text{mean average precision} = (0.62 + 0.44)/2 = 0.53$$



## More on Mean Average Precision Calculation

### ■ Mean Average Precision (MAP)

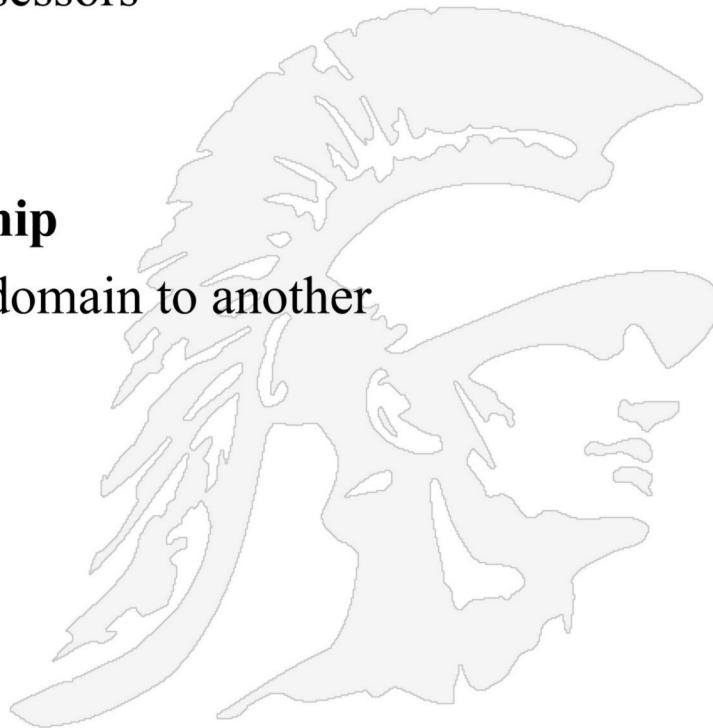
- Some negative aspects
  - If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero (this is actually reasonable)
  - Each query counts equally
  - MAP assumes user is interested in finding many relevant documents for each query
  - MAP requires many relevance judgments in the document collection





# Difficulties in Using Precision/Recall

- Should average over large document collection and query ensembles
- Need human relevance assessments
  - But people aren't always reliable assessors
- Assessments have to be binary
  - Nuanced assessments?
- Heavily skewed by collection/authorship
  - Results may not translate from one domain to another



## A Final Evaluation Measure: Discounted Cumulative Gain

- The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.
- The discounted CG accumulated at a particular rank position  $p$  is defined as

$$\text{DCG}_p = \sum_{i=1}^p \frac{\text{rel}_i}{\log_2(i+1)} = \text{rel}_1 + \sum_{i=2}^p \frac{\text{rel}_i}{\log_2(i+1)}$$

where  $\text{rel}_i$  is the graded relevance of the result at position  $i$

- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is  $1/\log(\text{rank})$ 
  - With base 2, the discount at rank 4 is  $1/2$ , and at rank 8 it is  $1/3$
- An alternative formulation of DCG places stronger emphasis on retrieving relevant documents:

$$\text{DCG}_p = \sum_{i=1}^p \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}$$

# Discounted Cumulative Gain Example

we want high weights for high rank documents, because searchers are likely to inspect them, and low weights for low rank documents that searchers are unlikely to ever see.

The discount factor is commonly chosen as  $\log_2(\text{rank} + 1)$  and is used to divide the relevance grade.

Using a logarithm for the position penalty makes the decay effect more gradual compared to using the position itself.

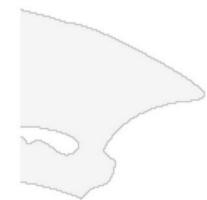
## Discount examples

Rank	Grade	Discount1	Discount2	Discount1	Discount2
		[1/rank]	[\log_2(\text{rank} + 1)]	Grade	Grade
1	4	1.000	1.000	4.000	4.000
2	3	0.500	0.631	1.500	1.893
3	2	0.333	0.500	0.667	1.000
4	1	0.250	0.431	0.250	0.431
5	1	0.200	0.387	0.200	0.387

# Search Engine Evaluation Metrics

## Metrics table

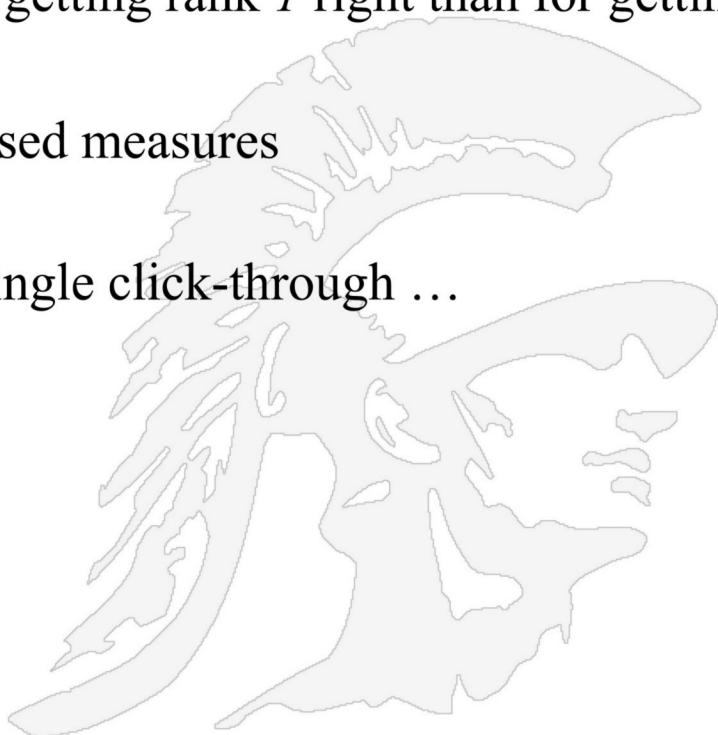
Scale	Metric	Measures	Drawbacks
Binary	<a href="#">Precision (P)</a>	The relevance of the entire results set (gridded results display)	Doesn't account for position
Binary	<a href="#">Average Precision (AP)</a>	Relevance to a user scanning results sequentially	Large impact of low-rank results
Graded	<a href="#">Cumulative Gain (CG)</a>	Information gain from a results set	Same as <i>Precision</i> doesn't factor in position
Graded	<a href="#">Discount Cumulative Gain (DCG)</a>	Information gain with positional weighting	Difficult to compare across queries
Graded	<a href="#">normalized DCG (nDCG)</a>	How close the results are to the best possible	No longer shows information gain



Finally see [https://en.wikipedia.org/wiki/Evaluation\\_measures\\_\(information\\_retrieval\)](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval))  
 Copyright Ellis Horowitz 2011-2022

# How Evaluation is Done at Web Search Engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top  $k$  positions, e.g.,  $k = 10$
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures
  - Click-through on first result
    - Not very reliable if you look at a single click-through ...  
but pretty reliable in the aggregate.
  - Studies of user behavior in the lab
  - A/B testing



# Google's Search Quality Rating Guidelines Document

- Google relies on raters, working in many countries and languages around the world
- The data they generate is rolled up statistically to give
  - a view of the quality of search results and search experience over time, and
  - an ability to measure the effect of proposed changes to Google's search algorithms

## General Guidelines

October 14, 2020

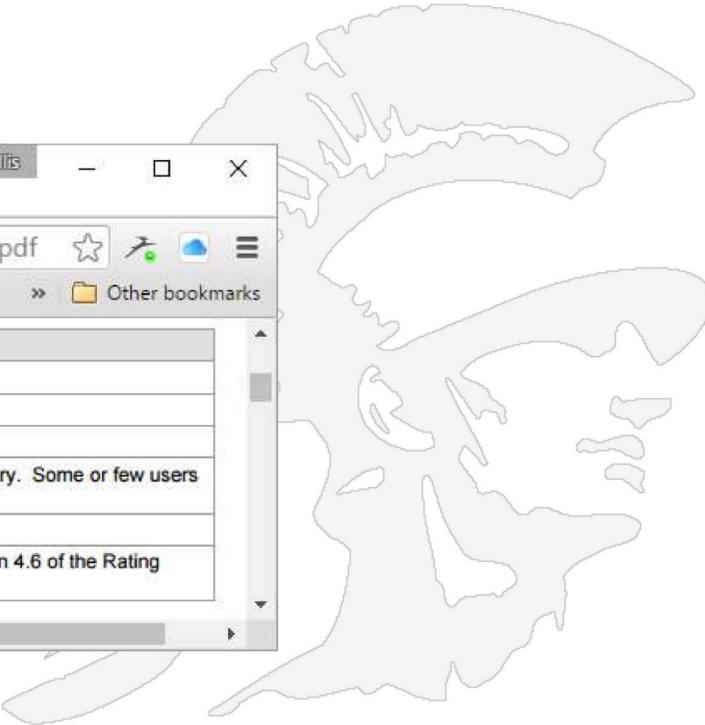
General Guidelines Overview	5
<b>Introduction to Search Quality Rating</b>	6
0.0 The Search Experience	6
0.1 The Purpose of Search Quality Rating	6
0.2 Raters Must Represent People in their Rating Locale	6
0.3 Browser Requirements	7
0.4 Ad Blocking Extensions	7
0.5 Internet Safety Information	7
0.6 The Role of Examples in these Guidelines	7
<b>Part 1: Page Quality Rating Guideline</b>	8
1.0 Introduction to Page Quality Rating	8
2.0 Understanding Webpages and Websites	8
2.1 Important Definitions	8
2.2 What is the Purpose of a Webpage?	9
2.3 Your Money or Your Life (YMLY) Pages	10
2.4 Understanding Webpage Content	10
2.4.1 Identifying the Main Content (MC)	10
2.4.2 Identifying the Supplementary Content (SC)	11
2.4.3 Identifying Advertisements/Monetization (Ads)	11
2.4.4 Summary of the Parts of the Page	12
2.5 Understanding the Website	12
2.5.1 Finding the Homepage	12
2.5.2 Finding Who is Responsible for the Website and Who Created the Content on the Page	14
2.5.3 Finding About Us, Contact Information, and Customer Service Information	14
2.6 Reputation of the Website or Creator of the Main Content	15
2.6.1 Research on the Reputation of the Website or Creator of the Main Content	16
2.6.2 Sources of Reputation Information	16
2.6.3 Customer Reviews of Stores/Businesses	16
2.6.4 How to Search for Reputation Information	16
2.6.5 What to Do When You Find No Reputation Information	18
3.0 Overall Page Quality Rating	19
3.1 Page Quality Rating: Most Important Factors	19
3.2 Expertise, Authoritativeness, and Trustworthiness (E-A-T)	19
4.0 High Quality Pages	20
4.1 Characteristics of High Quality Pages	20
4.2 A Satisfying Amount of High Quality Main Content	21
4.3 Clear and Satisfying Website Information: Who is Responsible and Customer Service	21
4.4 Positive Reputation	21
4.5 A High Level of Expertise/Authoritativeness/Trustworthiness (E-A-T)	22
4.6 Examples of High Quality Pages	22
5.0 Highest Quality Pages	26

[http://csci572.com/papers/2020\\_10searchqualityevaluatorguidelines.pdf](http://csci572.com/papers/2020_10searchqualityevaluatorguidelines.pdf)

# Google's Search Quality Ratings Guidelines Document

- This document gives evaluators examples and guidelines for appropriate ratings.
- the evaluator looks at a search query and a result that could be returned. They rate the relevance of the result for that query on a scale described within the document.

## The six rating scale categories



A screenshot of a web browser window showing the "searchqualityevaluatorguidelines.pdf" document from the USC Computer Science Department. The browser interface includes a title bar, address bar, toolbar, and menu bar. Below the browser is a table titled "The six rating scale categories".

Rating Scale	Description
<i>Vital</i>	A special rating category. See Section 4.1 of the Rating Guidelines.
<i>Useful</i>	A page that is very helpful for most users.
<i>Relevant</i>	A page that is helpful for many or some users.
<i>Slightly Relevant</i>	A page that is not very helpful for most users, but is somewhat related to the query. Some or few users would find this page helpful.
<i>Off-Topic or Useless</i>	A page that is helpful for very few or no users.
<i>Unrateable</i>	A page that cannot be evaluated. A complete description can be found in Section 4.6 of the Rating Guidelines.

# Google's 4-Step Process for Changing Their Search Algorithm

## 1. Precision Evaluations

People use the Guidelines to rate search results

## 2. Side-by-Side Experiments

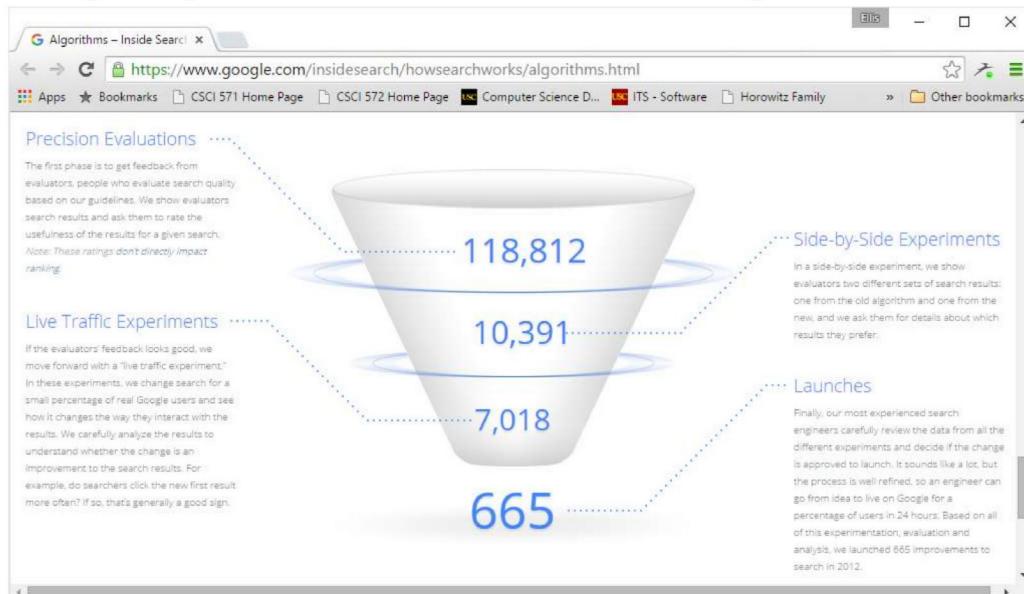
people are shown two different sets of search results and asked which they prefer

## 3. Live Traffic Experiments

the search algorithm is altered for a small number of actual users

## 4. Full Launch

A final analysis by Google engineers and the improvement is released



## A/B Testing at Web Search Engines

- **A/B testing** is comparing two versions of a web page to see which one performs better. You compare two web pages by showing the two variants (let's call them **A** and **B**) to similar visitors at the same time. The one that gives a better conversion rate, wins!

1. **Purpose:** Test a single innovation
  2. **Prerequisite:** You have a large search engine up and running.
  3. **Have most users use old system**
  4. **Divert a small proportion of traffic (e.g., 1%) to an experiment to evaluate an innovation**
  5. **Evaluate with an automatic measure like click through on first result**
- we directly see if the innovation does improve user happiness
  - *This is the evaluation methodology large search engines trust the most*

