

# IR evaluation, test collections, and TREC

Ian Soboroff

NIST

[ian.soboroff@nist.gov](mailto:ian.soboroff@nist.gov)

# Questions

- How can I tell if one search engine is better than another?
- What does “better” mean?
  - Better for what purpose? For who? When? How much better?
- How can I measure it?
- Can I measure some other search engine later, and compare it with what I measured before?

**[Jason R. Baron](#)**

Jason R. Baron has served as the National Archives.

[www.archives.gov/legal/bios/jrbaron.html](http://www.archives.gov/legal/bios/jrbaron.html) - 12k - [Cached](#) - [Similar pages](#)

**[JASON BARON | MySpace.com](#)**

MySpace profile for **JASON BARON**. Find friends, share photos, keep in touch with classmates, and meet new people on MySpace.com.

[www.myspace.com/djasonbaron](http://www.myspace.com/djasonbaron) - 99k - [Cached](#) - [Similar pages](#)

**[Jason Baron on MySpace Music - Free Streaming MP3s, Pictures ...](#)**

MySpace Music profile for **Jason Baron**. Download **Jason Baron** Electroacoustic / Lyrical / Indie music singles, watch music videos, listen to free streaming ...

[www.myspace.com/jasonbaronmusic](http://www.myspace.com/jasonbaronmusic) - 107k - [Cached](#) - [Similar pages](#)

**[Jason Baron on Search - How Do You Find Anything When You Have a ...](#)**

**Jason Baron** speaking at the University of Florida on Feb. 26, 2009 You think you've got email problems, **Jason R. Baron** just received 200 million emails from ...

[ralphlosey.wordpress.com/2009/03/04/jason-baron-on-search-how-do-you-find-anything-when-you-have-a-billion-emails/](http://ralphlosey.wordpress.com/2009/03/04/jason-baron-on-search-how-do-you-find-anything-when-you-have-a-billion-emails/) - 69k - [Cached](#) - [Similar pages](#)

**[Industry Insiders: Jason Baron, Dark Knight - BlackBook](#)**

Aug 13, 2008 ... Industry Insiders: **Jason Baron**, Dark Knight **Jason Baron**, owner of two of the Lower East Side's finest music dens, has all the makings of a ...

[www.blackbookmag.com/article/power-players-jason-baron-dark-knight/3780](http://www.blackbookmag.com/article/power-players-jason-baron-dark-knight/3780) - 46k -

[Cached](#) - [Similar pages](#)

**[Jason Baron - LinkedIn](#)**

View **Jason Baron's** professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like **Jason Baron** discover ...

[www.linkedin.com/in/jasonbaron](http://www.linkedin.com/in/jasonbaron) - 17k - [Cached](#) - [Similar pages](#)

**[\[PDF\] Microsoft PowerPoint - jason.baron.PPT](#)**

File Format: PDF/Adobe Acrobat - [View as HTML](#)

Philadelphia, PA. **Jason R. Baron**. Director of Litigation. Office of General Counsel. National Archives and Records Administration. October 26, 2006 ...

[db.cis.upenn.edu/iworkshop/postworkshop/slides/Session1Application/jason.baron.pdf](http://db.cis.upenn.edu/iworkshop/postworkshop/slides/Session1Application/jason.baron.pdf) -

[Similar pages](#)

**[Jason Baron - Burlington, VT | Facebook](#)**

**Jason Baron** (Burlington, VT) is on Facebook. Facebook gives people the power to share and makes the world more open and connected.

[www.facebook.com/people/Jason-Baron/515541581](http://www.facebook.com/people/Jason-Baron/515541581) - [Similar pages](#)

**[OntologWiki: Jason Baron](#)**

Jul 30, 2007 ... **Jason Baron** has served since the year 2000 as Director of Litigation for the National Archives and Records Administration, and is a frequent ...

[ontolog.cim3.net/cgi-bin/wiki.pl?JasonBaron](http://ontolog.cim3.net/cgi-bin/wiki.pl?JasonBaron) - 7k - [Cached](#) - [Similar pages](#)

**[EDD Update: Jason Baron & Carl Sagan](#)**

Mar 5, 2009 ... Remember Carl Sagan? He was probably the most famous astronomer of the modern era, known for the phrase "Billion and Billions" when ...

[www.eddupdate.com/2009/03/jason-baron-and-carl-sagan.html](http://www.eddupdate.com/2009/03/jason-baron-and-carl-sagan.html) - 71k -

[Cached](#) - [Similar pages](#)



Query:  
"jason baron"

Also try: [jason baron md](#), [dr jason baron](#), [jason baron cnn](#), [More...](#)

### [Jason R. Baron](#)

Jason R. Baron has served as the National Archives ... Home General Counsel Staff Biographies  
Jason R. Baron. Staff Biographies. Gary M. Stern ...  
[www.archives.gov/legal/bios/jrbaron.html](http://www.archives.gov/legal/bios/jrbaron.html) - [Cached](#)



### [MySpace.com - JASON BARON - 30 - Male - NEW YORK, New York - www ...](#)

MySpace profile for **JASON BARON** with pictures, videos, personal blog, interests, ... **JASON BARON**  
" LES â€œ NYC â€œ USA â€œ THE ARSENAL â€œ MY DARLING ...  
[www.myspace.com/djasonbaron](http://www.myspace.com/djasonbaron) - 216k - [Cached](#)



### [MySpace.com - Jason Baron - London, UK - Electroacoustic / Lyrical ...](#)

MySpace music profile for **Jason Baron** with tour dates, songs, videos, pictures, blogs, band  
information, ... point really but **Jason Baron** live is pretty unique. ...  
[profile.myspace.com/index.cfm?fuseaction=user.viewProfile&...](http://profile.myspace.com/index.cfm?fuseaction=user.viewProfile&...) - 113k - [Cached](#)



### [Jason Baron \(jetrel\)'s photos- powered by SmugMug](#)

The ultimate in photo sharing. Easily create online photo albums. ... Home | Login | Help | Color: **Jason Baron** (jetrel)'s Home. Gallery Categories. 2008 ...  
[jetrel.smugmug.com](http://jetrel.smugmug.com) - [Cached](#)



### [Jason Baron on Search - How Do You Find Anything When You Have a ...](#)

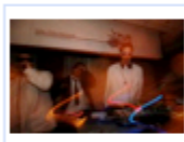
I say **Jason Baron** had time to provide a full overview of his ideas because it ... In addition to **Jason Baron**, the coordinators for the 2008 TREC Legal Track were ...  
[ralphlosey.wordpress.com/2009/03/04/jason-baron-on-search-how-do-yo...](http://ralphlosey.wordpress.com/2009/03/04/jason-baron-on-search-how-do-yo...) - 67k - [Cached](#)



### [Jason Baron on Facebook](#)

[Add friend](#) | [Poke](#) | [Send message](#) | [View friends](#)

**Jason Baron** is on Facebook. Not the **Jason Baron** you were looking for? Visit Facebook to search for friends, family, coworkers.  
[www.facebook.com/people/Jason\\_Baron/515541581](http://www.facebook.com/people/Jason_Baron/515541581) - [Cached](#)



### [Jason Baron on Facebook](#)

[Add friend](#) | [Poke](#) | [Send message](#) | [View friends](#)

**Jason Baron** is on Facebook. Not the **Jason Baron** you were looking for? Visit Facebook to search for friends, family, coworkers.  
[www.facebook.com/people/Jason-Baron/744095277](http://www.facebook.com/people/Jason-Baron/744095277) - [Cached](#)



### [Jason R. Baron](#)

Jason R. Baron. About. About. December 25, 2008 – 4:08 pm ... My email address: [jason.baron@nara.gov](mailto:jason.baron@nara.gov). " Home. Pages. About. Categories. No categories ...  
[jasonrbaron.com](http://jasonrbaron.com) - [Cached](#)



### [Amazon.com: Open All Hours: Barker, Jason, Baron, Cole, Flynn: Video](#)

Actors: Barker, **Jason, Baron**, Cole, Flynn. Format: Color, NTSC. Language: English. Rating: ... such thoughts on the part of his nephew Granville (David **Jason**) ...  
[www.amazon.com/Open-All-Hours-Barker/dp/B00004WG58](http://www.amazon.com/Open-All-Hours-Barker/dp/B00004WG58) - 215k - [Cached](#)



### [Jason Baron : opensubscriber](#)

Home Log In Sign Up Create Group FAQ. Contributors Profile. Name. **Jason Baron**. Possible Location. United States. Groups. Posts. [linux-fsdevel@vger.kernel.org](mailto:linux-fsdevel@vger.kernel.org) ...  
[www.opensubscriber.com/profile.os?individual=1120051](http://www.opensubscriber.com/profile.os?individual=1120051) - [Cached](#)





Web 1-10 of 3,130,000 results · [Advanced](#)

See also: [Images](#), [Video](#), [News](#), [Maps](#), [More](#) ▼

### [Jason R. Baron](#)

I presently serve as Director of Litigation at the U.S. National Archives and Records Administration, a position I have held since May 2000. I also am an Adjunct Professor on the ...

[jasonbaron.com](#) · [Cached page](#)

### [Jason R. Baron](#)

Jason R. Baron has served as the National Archives ... Jason R. Baron. Jason R. Baron has served as the National Archives' Director of Litigation since May 2000.

[www.archives.gov/legal/bios/jrbaron.html](#) · [Cached page](#)

### [Jason Baron Campaign Contributions and Donations -- Huffington Post](#)

Jason Baron donations and other campaign contributions on Huffington Post

[fundrace.huffingtonpost.com/neighbors.php?type=name&lname=Baron&fname=Jason](#) · [Cached page](#)

### [MySpace.com - JASON BARON - 30 - Male - NEW YORK, New York - www ...](#)

MySpace profile for JASON BARON with pictures, videos, personal blog, interests, information about me and more

[www.myspace.com/djasonbaron](#) · [Cached page](#)

### [EDD Update: Jason Baron & Carl Sagan](#)

Remember Carl Sagan? He was probably the most famous astronomer of the modern era, known for the phrase "Billion and Billions" when describing stars and other cosmic phenomena.

[www.eddupdate.com/2009/03/jason-baron-and-carl-sagan.html](#) · [Cached page](#)

### [Jason R. Baron - National Archives](#)

Description. Jason R. Baron has served as the National Archives' Director of Litigation since May 2000. In this position, Mr. Baron is responsible for overseeing all litigation ...

[videlectures.net/jason\\_r\\_baron](#) · [Cached page](#)

### [jason baron: Directory - LinkedIn](#)

Are you jason baron? Get a Profile. jason baron: Directory. Name Location & Industry • Greater Chicago Area • Construction Jason Baron

[www.linkedin.com/pub/dir/jason/baron](#) · [Cached page](#)

### [MySpace.com - Jason Baron - London, UK - Electroacoustic / Lyrical ...](#)

MySpace music profile for Jason Baron with tour dates, songs, videos, pictures, blogs, band information, downloads and more

[profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendid=54223611](#) · [Cached page](#)

### [Song Tank - View Profile: jasonbaron](#)

JASON BARON biography: 'The English Ron Sexsmith', 'a one man Athlete', 'Eels meets Billy Bragg', 'Badly Drawn Boy without the silly hat

[www.song-tank.com/member.php?u=219](#) · [Cached page](#)

### [Jason Baron](#)

Jason Baron. Main Quotes Index. Issue #321, Section #2 (3 Sep 2005: RelayFS Likely To Go Into -mm) Share And Enjoy!

[kerneltraffic.spegulo.be/kernel-traffic/quotes/Jason\\_Baron.html](#) · [Cached page](#)



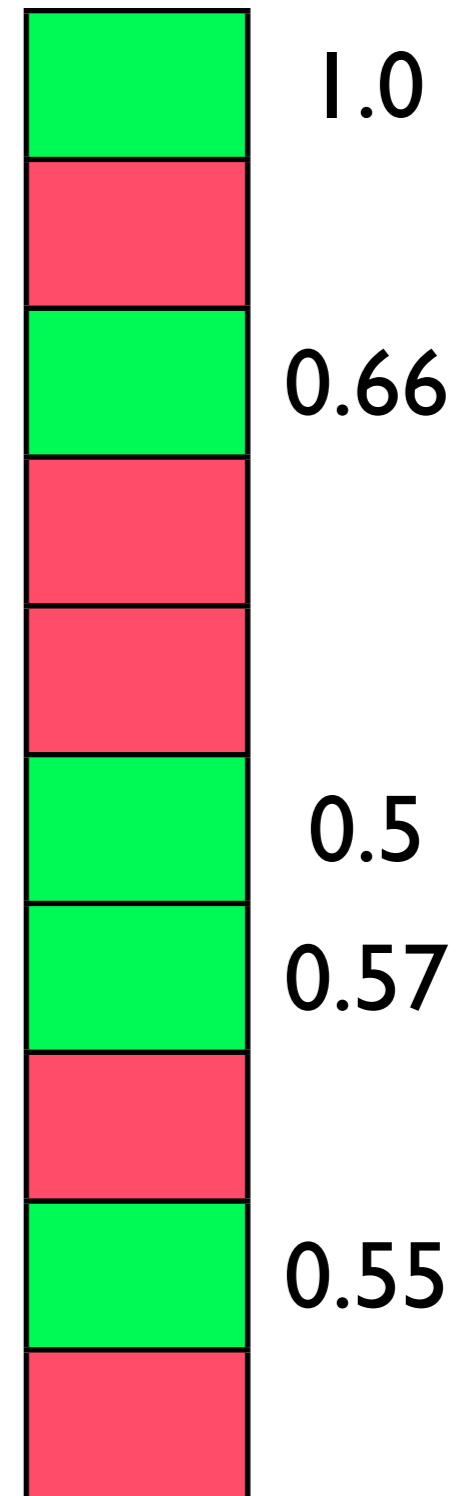
# Precision and Recall

Looking at the set of documents retrieved...

- **Precision:** how many of the answers are right?  
*(how many of the documents are relevant?)*
- **Recall:** how many of the right answers did I find?  
*(how many of the relevant documents did I find?)*

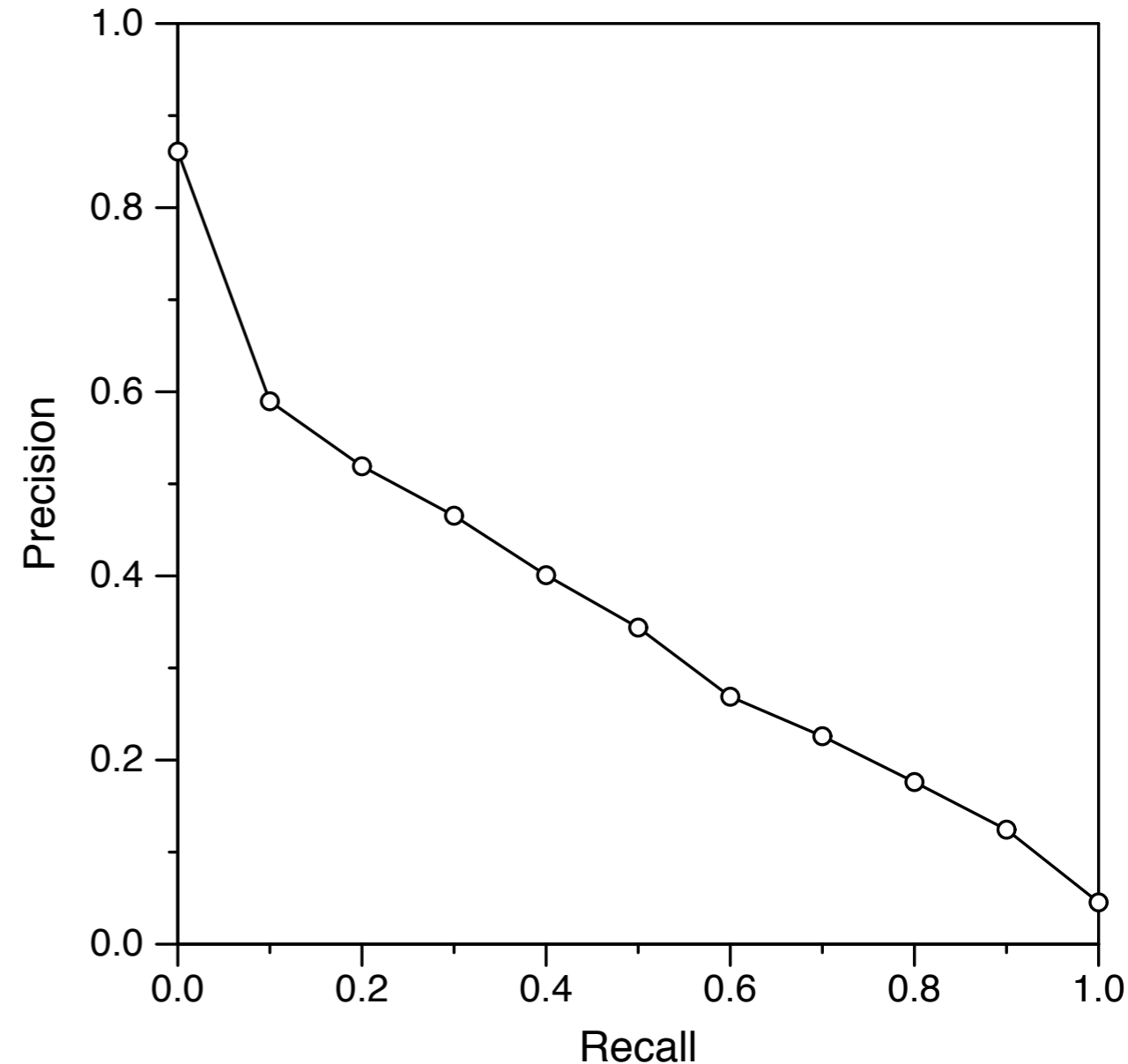
# Measuring a ranked list

- Assume that the searcher looks at the hits in the ranked list in order.
- How good is the ranked list?
- **Average precision:** compute the precision at each rank that a relevant document is found.
- normalize by the number of relevant documents.



# Recall-precision curves

- Plot precision at fixed recall points
- (10% recall, 20%, ..., 100%)
- Precision values are interpolated so that curve is nonincreasing:  
$$prec(r_n) = \max(prec(r_{n+1}))$$
- Avg. precision is the area under the curve.





# What were we looking for?

- Precision and recall measure a set of search results.
- Average precision and RP curves measure a ranked list.
- What were we trying to find, again?
  - A known item? A homepage?
  - A few answers? All the answers?
  - A survey of the information? A bio?
- We can't decide what to measure until we know what the **task** was.

# Tasks and measures

- Known item search, homepage search
  - mean reciprocal rank (MRR) of first relevant document
  - success at rank N
- A few answers: precision at low recall (or early rank)
- All the answers: precision at high recall, average precision
- A survey: aspect recall?

# Are all documents equal?

- Some documents are more useful than others.
  - A “highly relevant” document might capture most of what I need to know.
  - A good summary can tell me what else I should read.
- One document can save me from reading all the others.
- There are measures for “degrees” of relevance.
- But dependencies and more sophisticated navigation of a ranked list is harder to measure.

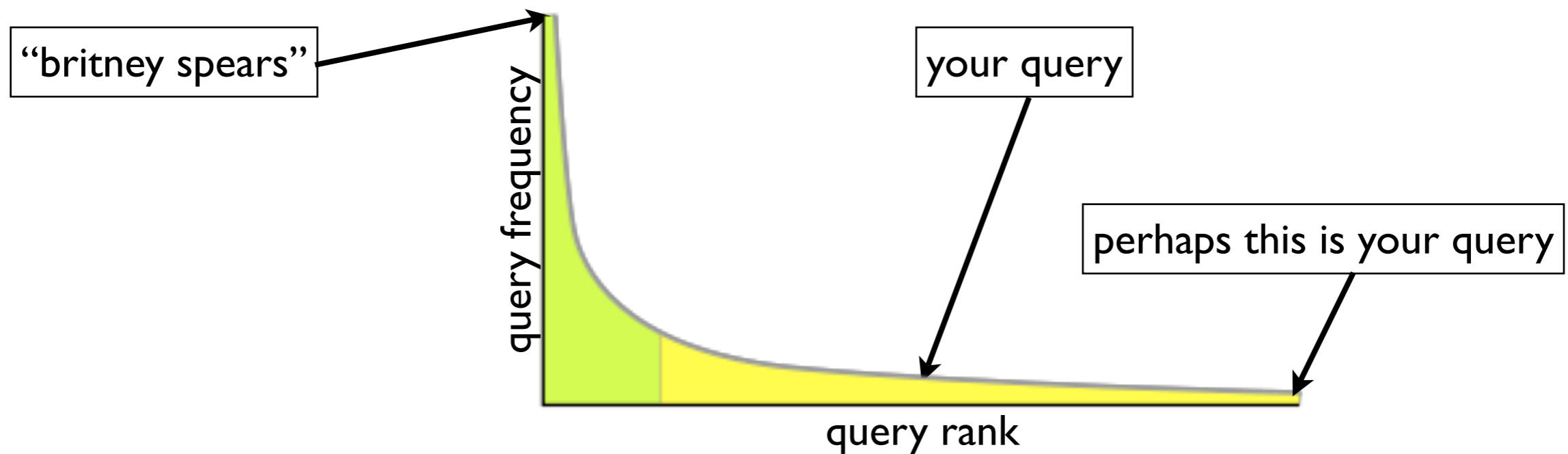
# “It’s only a model”



- So far, we have a simple model for evaluation:
  - information need gets articulated as a query
  - the query yields a ranked list
  - which we read in order
  - noting what’s relevant and what’s not
  - until we’ve read them all
- We can look at some variations in task with measures of the ranked list.
- But this model can’t capture everything.
- Complex interactions and models of relevance are hard to make fit.

# Multiple queries

- So far, we've only considered a single need and query.
- Commercial search engines see millions of queries per day.





britney spears, information retrieval, trec, web search

Tip: Use commas to compare multiple search terms.

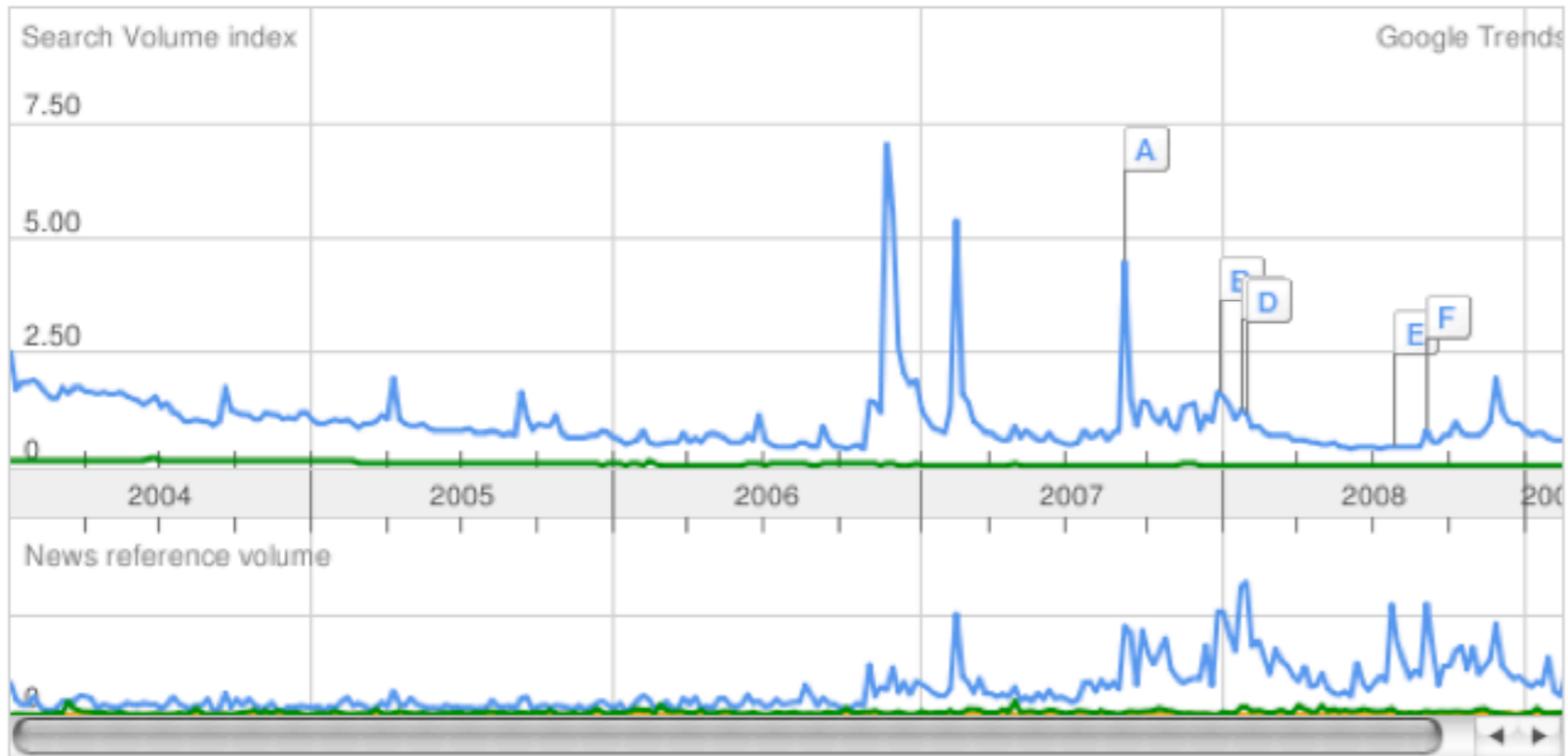
isoboroff@gmail.com | Sign out

Searches Websites

All regions All years

Scale is based on the average worldwide traffic of britney spears in all years. Learn more

britney spears 1.00 information retrieva... 0 trec 0 web search 0.10



- A [Britney Spears MTV Embarrassment](#)  
GameSHOUT - Sep 10 2007
  - B [Police called to Britney Spears' house](#)  
Macon Telegraph - Jan 4 2008
  - C [Britney Spears in hospital](#)  
WIS - Jan 31 2008
  - D [Britney Spears released](#)  
Fashion.ie - Feb 6 2008
  - E [Court extends Britney Spears' conservatorship](#)  
The Ledger - Jul 31 2008
  - F [Britney Spears sweeps at MTV Awards](#)  
Vancouver Sun - Sep 8 2008
- [More news results »](#)

Rank by britney spears

Regions	
1. <a href="#">Mexico</a>	
2. <a href="#">Canada</a>	
3. <a href="#">Venezuela</a>	
4. <a href="#">United States</a>	

Cities	
1. Guadalajara, Mexico	
2. Mexico City, Mexico	
3. Los Angeles, CA, USA	
4. Caracas, Venezuela	

Languages	
1. English	
2. Spanish	
3. Swedish	
4. Danish	




Search Trends

Tip: Use commas to compare multiple search terms.

## Hot Trends (USA)

Mar 23, 2009 - [change date](#)

Updated 49 minutes ago

[iGoogle Gadget](#) New!

[Site Feed](#)

- |   |  |  |  |
|---|--|--|--|
| 1. <a href="#">missingmoney.com</a>         | 26. <a href="#">mt redoubt</a>                     | 51. <a href="#">wydot</a>                                    | 76. <a href="#">jason siegel</a>                     |
| 2. <a href="#">paper cuts</a>               | 27. <a href="#">nicholas hughes</a>                | 52. <a href="#">fedex plane crash</a>                        | 77. <a href="#">obama on 60 minutes</a>              |
| 3. <a href="#">johari smith</a>             | 28. <a href="#">team jonas</a>                     | 53. <a href="#">boston dynamics</a>                          | 78. <a href="#">paul krugman</a>                     |
| 4. <a href="#">jenna bush wedding dress</a> | 29. <a href="#">mt. redoubt</a>                    | 54. <a href="#">ncaa sweet sixteen 2009</a>                  | 79. <a href="#">westconsin credit union</a>          |
| 5. <a href="#">olney texas</a>              | 30. <a href="#">queen bee syndrome</a>             | 55. <a href="#">carbon motors</a>                            | 80. <a href="#">mt redoubt volcano cam</a>           |
| 6. <a href="#">sylvia plath</a>             | 31. <a href="#">anchorage daily news</a>           | 56. <a href="#">irobot roomba 535</a>                        | 81. <a href="#">becky slater</a>                     |
| 7. <a href="#">obama punch drunk</a>        | 32. <a href="#">ted hughes</a>                     | 57. <a href="#">frieda hughes</a>                            | 82. <a href="#">redoubt eruption</a>                 |
| 8. <a href="#">curt schilling</a>           | 33. <a href="#">nikki cox</a>                      | 58. <a href="#">\$2000 car</a>                               | 83. <a href="#">amanda knox</a>                      |
| 9. <a href="#">assia wevill</a>             | 34. <a href="#">bigdog robot</a>                   | 59. <a href="#">octuplet mom admits welfare was her plan</a> | 84. <a href="#">ramapo ny</a>                        |
| 10. <a href="#">c 17</a>                    | 35. <a href="#">jonasbrothersfanclub.com</a>       | 60. <a href="#">magnolia lane plantation</a>                 | 85. <a href="#">c 17 crash texas</a>                 |
| 11. <a href="#">wolf block</a>              | 36. <a href="#">natrona county school district</a> | 61. <a href="#">carbon e7</a>                                | 86. <a href="#">team jonas.com</a>                   |
| 12. <a href="#">ann arbor news</a>          | 37. <a href="#">wyoming road conditions</a>        | 62. <a href="#">jonas brothers presale password</a>          | 87. <a href="#">area code 402</a>                    |
| 13. <a href="#">missing money</a>           | 38. <a href="#">mount redoubt</a>                  | 63. <a href="#">fisher scientific</a>                        | 88. <a href="#">missingmoney.com main search.cfm</a> |
| 14. <a href="#">jonas brothers fan club</a> | 39. <a href="#">hampton jazz festival 2009</a>     | 64. <a href="#">national puppy day</a>                       | 89. <a href="#">k max helicopter</a>                 |
| 15. <a href="#">38 pitches</a>              | 40. <a href="#">dr. erin jacobson</a>              | 65. <a href="#">avo</a>                                      | 90. <a href="#">the bell jar</a>                     |
| 16. <a href="#">wonderlic test</a>          | 41. <a href="#">the amazing jonathan</a>           | 66. <a href="#">george weber</a>                             | 91. <a href="#">38pitches.com</a>                    |
| 17. <a href="#">mark stein</a>              | 42. <a href="#">din thomas</a>                     | 67. <a href="#">hot lake hotel</a>                           | 92. <a href="#">emma hemming</a>                     |
| 18. <a href="#">cisco fatty</a>             | 43. <a href="#">wolf block law firm</a>            | 68. <a href="#">danica mckellar</a>                          | 93. <a href="#">anchorage airport</a>                |
| 19. <a href="#">napa valley register</a>    | 44. <a href="#">fargo forum</a>                    | 69. <a href="#">bruce willis wedding</a>                     | 94. <a href="#">wisconsin dnr</a>                    |
| 20. <a href="#">crystal mckellar</a>        | 45. <a href="#">smiley face murders</a>            | 70. <a href="#">old national bank</a>                        | 95. <a href="#">www.diabetes.org/alert</a>           |
| 21. <a href="#">matt lauer</a>              | 46. <a href="#">hymenoptera</a>                    | 71. <a href="#">mike verta</a>                               | 96. <a href="#">toxic assets</a>                     |
| 22. <a href="#">redoubt volcano</a>         | 47. <a href="#">sweet 16 bracket</a>               | 72. <a href="#">octomom welfare</a>                          | 97. <a href="#">irobot roomba 535 robotic vacuum</a> |

# Test collections

- An IR test collection includes
  - a set of documents (e.g. LTDL),
  - a set of search needs (“topics”),
  - a mapping of which documents are relevant to which needs (“relevance judgments”, or “qrels”).
- A test collection provides a reusable framework for conducting repeatable experiments.



# How to build a test collection

- **Documents:** crawl the web, get news articles, sue somebody
- **Needs:** can be internal or external
  - internal needs come from real users or query logs.
  - external needs are composed based on the data and world knowledge.
  - some minimal number of needs/topics are required.
- **Relevance judgments** can be
  - complete
  - pooled
  - sampled

# Example: TREC 8 Adhoc

- Documents: 500k newswire documents (~2GB).
- Needs/topics: NIST assessors search the collection and created 50 topics on various subjects.
- More than 80,000 documents were pooled from 71 search engine outputs from dozens of research teams. These were judged by the NIST assessors.

# Example: CERC

- CERC: CSIRO Enterprise Research Collection
- (CSIRO is Australia's national science agency)
- Documents: a crawl of the public-facing csiro.au web
- Two sets of needs/topics, from two internal sources
  1. Science communicators (they write the web pages)
  2. CSIRO Enquiries (they answer "contact us" emails)
- Tasks: document search and expert finding
- Relevance judgments collected from CSIRO employees as well as TREC participants

# Pooling

- (Sparck Jones and van Rijsbergen, 1975)
- Get a bunch of different people to search using different search engines.
- Combine their top N results for each need.
- Only judge those documents.

# Pooling

- Pooled relevance judgments let us measure
  - precision up to rank N exactly
  - mean reciprocal rank (usually)
  - an estimate of the number of relevant documents, with which we can compute average precision

# How good is pooling?

- The quality of the pools depends on a lot of factors:
  - how hard is the task?
  - how good are the systems?
  - how deep are the pools?
  - how big is the collection?

Relevant

Not relevant

Unknown

# TREC

- Text REtrieval Conference (1991 - now)
- Sponsored by NIST and \*ARPA
- Goals
  1. To promote IR research based on large test collections.
  2. To provide a forum to share results and ideas.
  3. To bridge academia, government, and industry.
  4. To develop IR evaluation methodology.
- TREC was the first attempt to build test collections using the pooling approach.

# Scaling up test collections

- Originally, we thought that test collections needed to have complete relevance judgments.
- The early collections (60s-80s) were build this way.
- Complete judgments scales to a few thousand documents with a few tens or hundreds of topics.
- Pooling seems to scale into the (low) millions of documents (task: adhoc search in newswire and web pages).
- Beyond that point, some evidence seems to show that relevance judgments can be biased to the pool runs.



# Example: TREC Terabyte Collections

- The TREC terabyte track was an effort to scale the TREC methodology up to much larger collections.
- (and then see what broke)
- Documents: a crawl of .gov websites (426GB, 26m docs)
- Needs: NIST assessors searched the collection to create 50 topics on a variety of subjects.
- Relevance judgments were pooled from search results obtained from 10-20 academic and industry participants in TREC.
- Around 50,000 documents were judged (compared to 80k in TREC 8 adhoc)

# Sampling

- Pooling makes us judge more than we really need to in the smaller collections.
- Pooling may give biased estimates of recall and AP in larger collections.
- Current research is in using sampling.
  - uniform sampling (requires too large a sample)
  - stratified sampling (pooling is one case)
  - measure-optimized sampling (to estimate AP)
  - comparison-optimized sampling (to rank systems)

# Example: TREC MQ 2007

- The million query track's goal is to try to build test collections with many queries and few relevance judgments.
- Collection: same as the Terabyte collection.  
(26 million documents, 426GB)
- Queries were chosen at random from a large query log, and then interpreted into topics.
- Each query/topic got from 8-128 relevance judgments
  - documents selected by one of two sampling strategies
- More than 1600 queries were interpreted and judged.