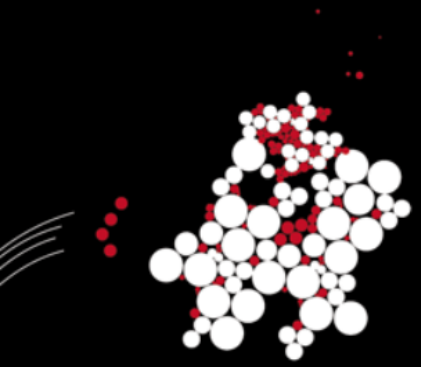
UNIVERSITY OF TWENTE.



SEARCH EVALUATION

Foundations of Information Retrieval 2018

Djoerd Hiemstra
Dolf Trieschnigg
Theo Huibers

Congratulations!

Is it any good?

BACKGROUND READING

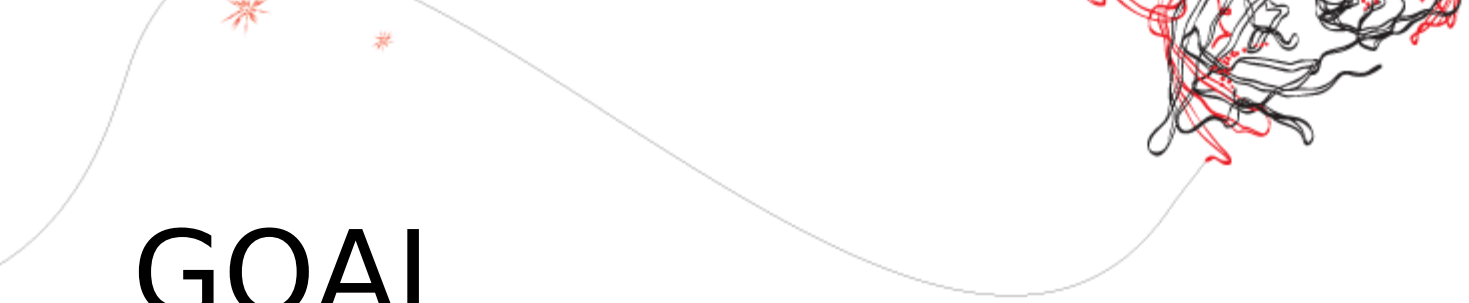
- Chapter 8, Evaluation in Information Retrieval, of: Christopher Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

  http://informationretrieval.org

- *Alternative:* Djoerd Hiemstra and Wessel Kraaij, *Evaluation of Multimedia Retrieval Systems*, In Multimedia Retrieval, Springer, pages 347-365, 2007

  http://www.cs.utwente.nl/~hiemstra/papers/mmbook-eval.pdf

NOTABLE PEOPLE



GOAL

- An introduction to doing *real* (measurable, repeatable) research
- Getting acquainted with the “TREC paradigm”

THE EMPIRICAL STUDY

- Clearly laid out sequence of steps:
  1. hypothesis;
  2. method;
  3. results;
  4. conclusion.
- The environment must be carefully controlled if the results of an evaluation are to be trusted.

1. YOUR HYPOTHESIS

- System *A* outperforms system *B* on task *C*
  - e.g. Google's Page Rank outperforms the vector space model with *tf.idf* weighting for searching home pages on the web

2. WHAT METHOD?

- Identify the techniques that will be used to establish the hypothesis.
  - choose data
  - choose suitable evaluation measures: assign values to results of your system
  - choose a statistical methodology: determine whether observed differences are significant
- The ability to repeat an experiment is a key feature of empirical research.

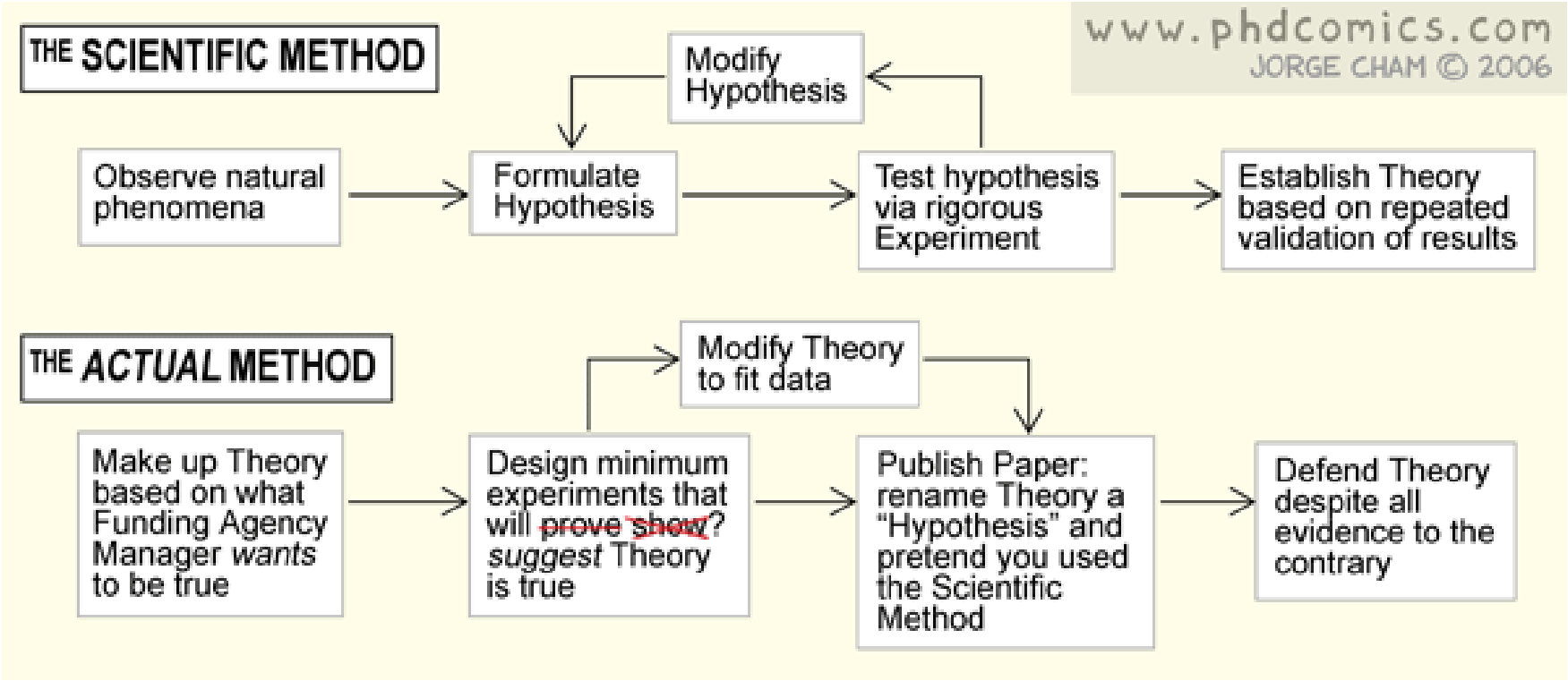
3. RESULTS

- Compile and present the results.
  - Repeat a number of times

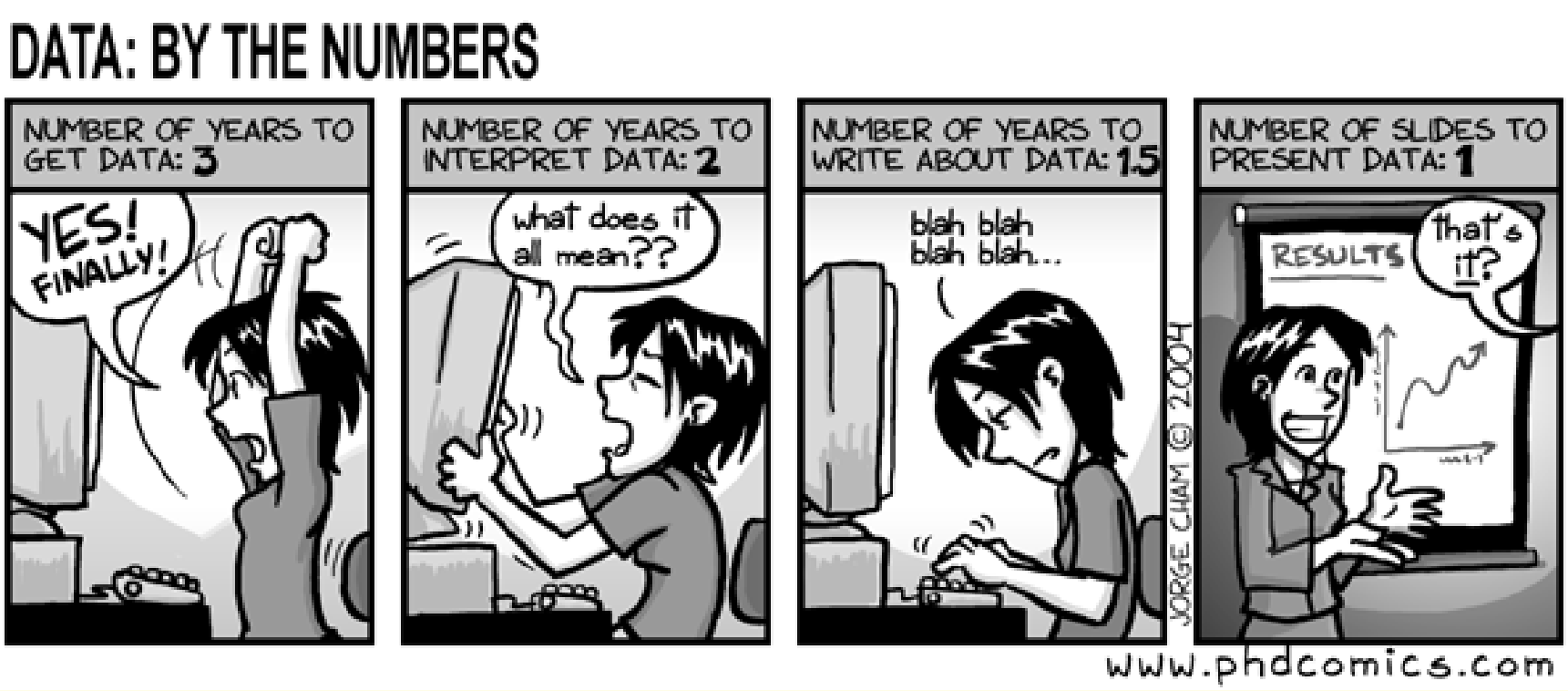
4. CONCLUSION

- Supporting the hypothesis...
- or rejecting it.

SUMMARY


THE SCIENTIFIC METHOD
Modify Hypothesis
Observe natural phenomena
Formulate Hypothesis
Test hypothesis via rigorous Experiment
Establish Theory based on repeated validation of results
www.phdcomics.com
JORGE CHAM © 2006
THE ACTUAL METHOD
Modify Theory to fit data
Make up Theory based on what Funding Agency Manager wants to be true
Design minimum experiments that will prove show? suggest Theory is true
Publish Paper: rename Theory a "Hypothesis" and pretend you used the Scientific Method
Defend Theory despite all evidence to the contrary

SUMMARY


DATA: BY THE NUMBERS
NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
JORGE CHAM © 2004
www.phdcomics.com

EMPIRICAL COMPUTER SCIENCE RESEARCH

- “3.7 % of computer science journal papers use the *laboratory experiment* as the primary research method”
- ACM Transactions on Information Systems was the only journal in which comparative studies of systems (laboratory experiment) was used as the primary research method (14.3 %)

V. Ramesh et al. “Research in computer science: an empirical study”, Journal of Systems and Software 70 (2004) 165-176

THE TRADITIONAL IR EXPERIMENT

- To start with you need
  - A system (or two)
  - A collection of documents / data
  - A collection of queries / requests
- Then you run your experiment
  - Input (index) the documents
  - Put each query to the system
  - Collect the output

(thanks to Stephen Robertson)

THE TRADITIONAL IR EXPERIMENT

- Then you need to
  - Evaluate the output, document by document
  - Discover (??) the good documents your system has missed
  - Analyse the results

(thanks to Stephen Robertson)

THE TRADITIONAL IR EXPERIMENT

- What is a document?
  - package of information structured by an author
- What is a request?
  - a description of a topic of interest
  - a partial representation of an underlying information need
- What is a system?
  - A device that accepts a request and delivers of identifies documents
  - "device" may be an organisation: involve people(!)

(thanks to Stephen Robertson)

THE TRADITIONAL IR EXPERIMENT

- Assuming that documents are either relevant or not, the objective is:
  - To retrieve relevant documents
  - Not to retrieve non-relevant documents

THE TRADITIONAL IR EXPERIMENT

- Evaluation measures
  - precision = r/n : fraction of retrieved documents that is relevant
  - recall = r/R : fraction of relevant documents that is retrieved

  r : number of relevant documents retrieved

  n : number of documents retrieved

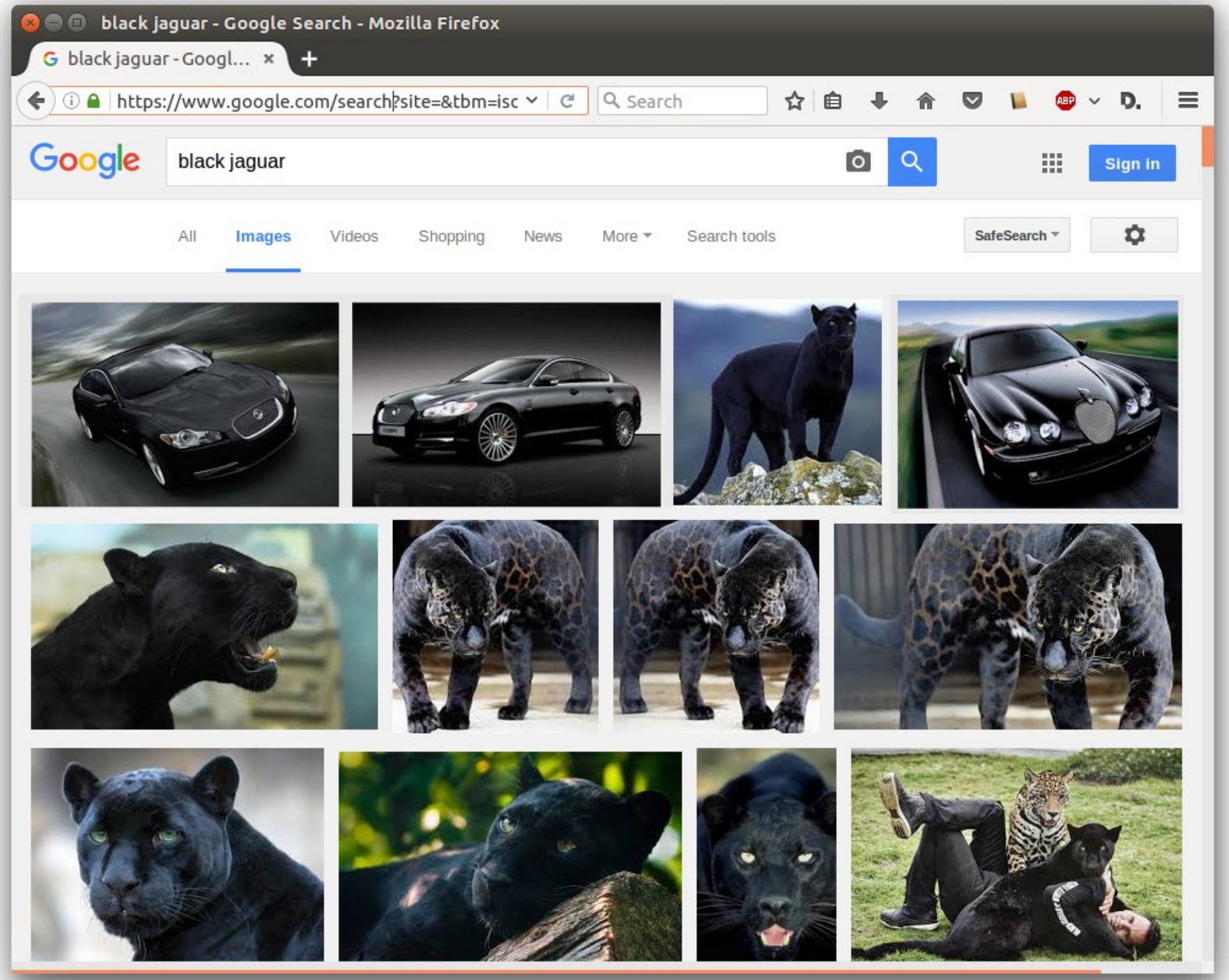
  R : number of relevant documents

HOW TO DECIDE?

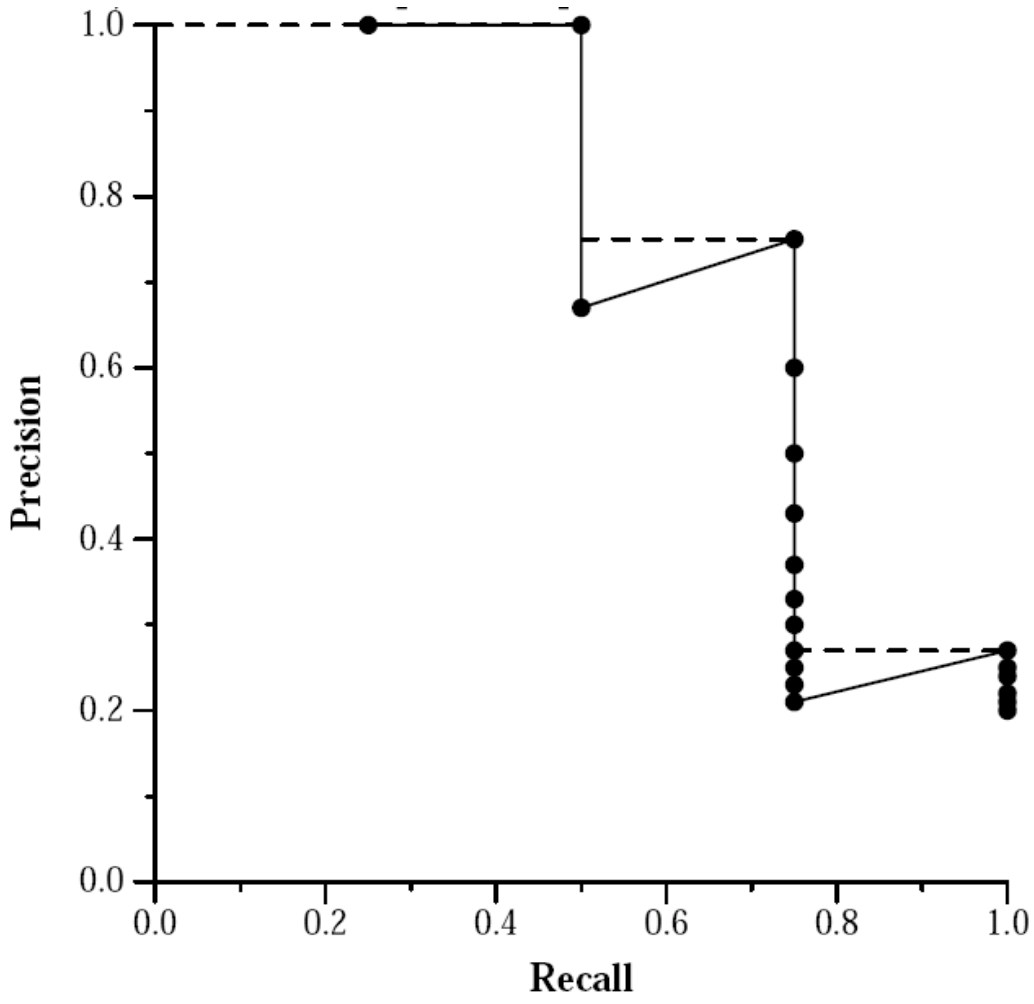
- We need a single measure:
- $F = 2 \cdot \text{Precision} \cdot \text{Recall} / \text{Precision} + \text{Recall}$

WHAT ABOUT RANKED OUTPUT?

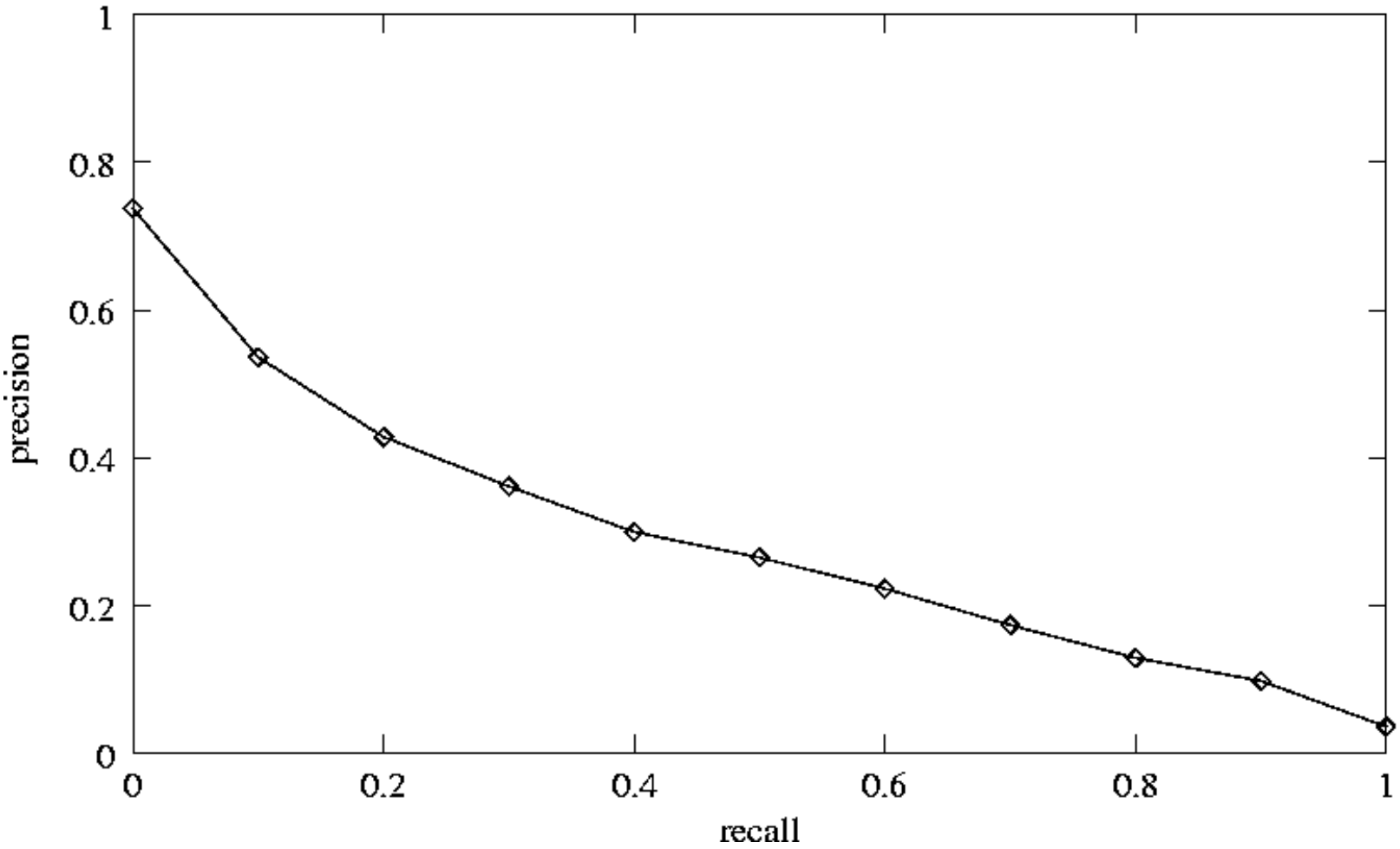
- Report precision for positions in the ranked list
  - 5, 10, 20 document retrieved
- Report precision for some recall levels
  - precision at 0.1, 0.2, etc.


black jaguar - Google Search - Mozilla Firefox
black jaguar - Googl...
https://www.google.com/search?site=&tbm=isc
Search
Google
black jaguar
Sign in
All
Images
Videos
Shopping
News
More
Search tools
SafeSearch

RECALL-PRECISION PLOT



RECALL-PRECISION PLOT



Average precision

- Calculates a trade-off between precision and recall
- Average precision at recall points
  - Average P@k for relevant documents (at rank k)

$$AP = \frac{\sum_{k=1}^{n} P(k)rel(k)}{\text{num rel docs}}$$

- Calculate the AP (assume num rel docs=6)
  - Relevant docs at ranks: 1, 7, 8, 10, 11

$$\frac{\frac{1}{1} + \frac{2}{7} + \frac{3}{8} + \frac{4}{10} + \frac{5}{11}}{6} = 0.42$$

HOW TO DECIDE?

- Mean Average Precision

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

Reciprocal rank

- Inverse of rank of first relevant hit
  - First relevant hit at rank 1: 1/1
  - First relevant hit at rank 10: 1/10
- Useful for evaluating
  - Known-item search
  - Navigational queries

Normalized discounted cumulative gain (nDCG)

- Some documents are more important than others
- Uses graded relevance judgments

$$CG_p = \sum_{i=1}^{p} rel_i$$

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

rank	rel_i	$\log_2(i)$	$\frac{rel_i}{\log_2(i)}$
1	1	0.00	n/a
2	0	1.00	0.00
3	3	1.58	1.89
4	2	2.00	1.00
5	1	2.32	0.43

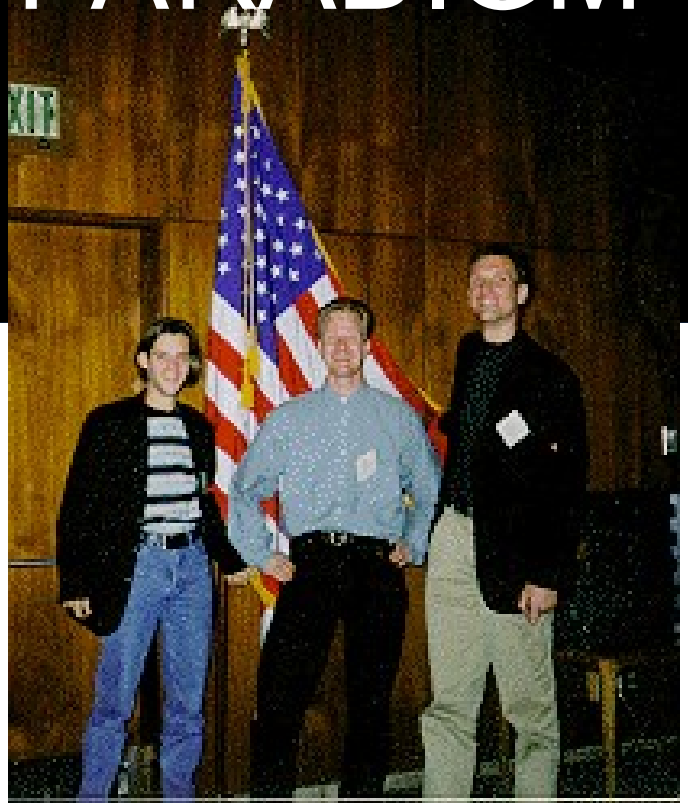
DCG_5	4.32
$IDCG_5$	6.13
$NDCG_5$	0.71

THE TRADITIONAL IR EXPERIMENT

- Problems with IR system evaluation
  - costly (involves users)
  - which documents did the system miss?
  - hard to repeat in same settings (learning / fatigue effects)
  - we need a complete system(!) we do not in general know how to evaluate components

UNIVERSITY OF TWENTE.

THE TREC PARADIGM




Certificate of Appreciation
In recognition of your 15 years of
loyal dedication to the
Text Retrieval Conference (TREC)
University of Twente
November 16, 2011

BENCHMARK COLLECTIONS

- Consists of three parts:
  - documents (realistic contents and size)
  - requests (textual description of information need; realistic, "real" application)
  - relevance assessments: how useful is the retrieved document?
- How to design?
  - Cranfield → TREC → CLEF, NTCIR, INEX

CRANFIELD EXPERIMENTS

- Librarian at Cranfield College of Aeronautics
- First empirical IR experiments
- (maybe the first empirical research in computer science...)

Cyril Cleverdon

CRANFIELD EXPERIMENTS

- Text search beats manual classification!

“This conclusion is so controversial and so unexpected that it is bound to throw considerable doubt on the methods which have been used (...) A complete recheck has failed to reveal any discrepancies (...) there is no other course except to attempt to explain the results which seem to offend against every canon on which we were trained as librarians.”

(Cleverdon & Keen 1966)

WHAT IS TREC?

- Competition/collaboration between IR research groups world-wide
- Run by the US National Institute of Standards and Technology (NIST)
- TREC provides:
  - common test collections
  - common tasks
  - common measures
  - common evaluation procedures

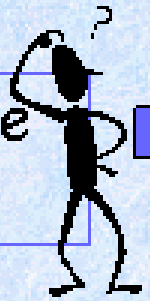
What is TREC?

- **A workshop series that provides the infrastructure for large-scale testing of text retrieval technology**
  - realistic test collections
  - uniform, appropriate scoring procedures
  - a forum for the exchange of research ideas and for the discussion of research methodology

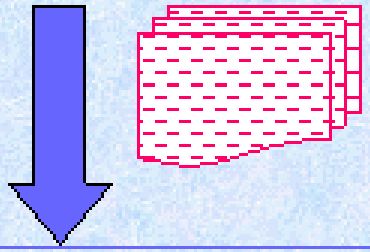
(thanks to Ellen Voorhees)

TREC approach

Assessors create topics at NIST



Topics are sent to participants, who return ranking of best 1000 documents per topic



NIST forms pools of unique documents from all submissions which the assessors judge for relevance

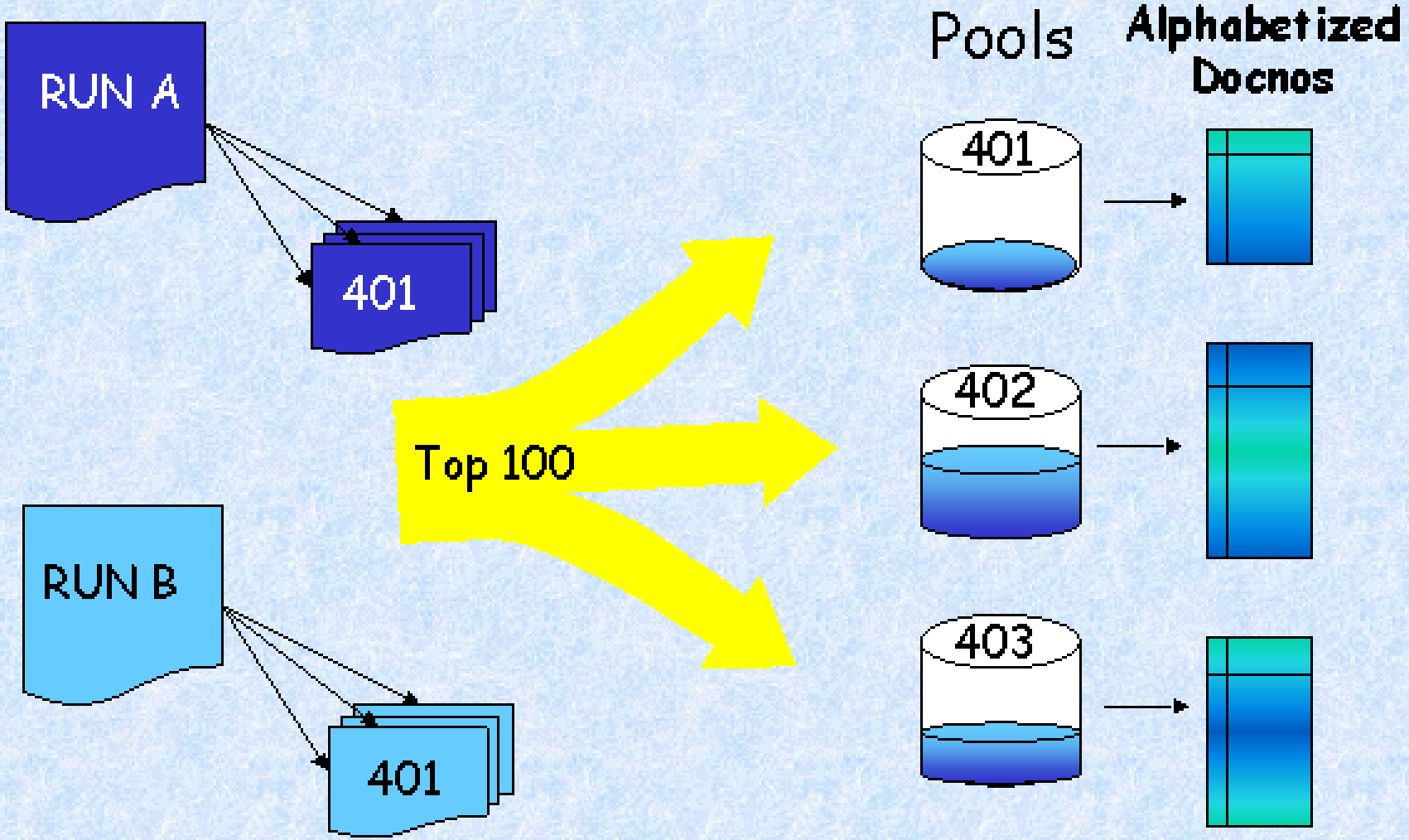
Systems are evaluated using relevance judgments

(thanks to Ellen Voorhees)

AN EXAMPLE TREC TOPIC

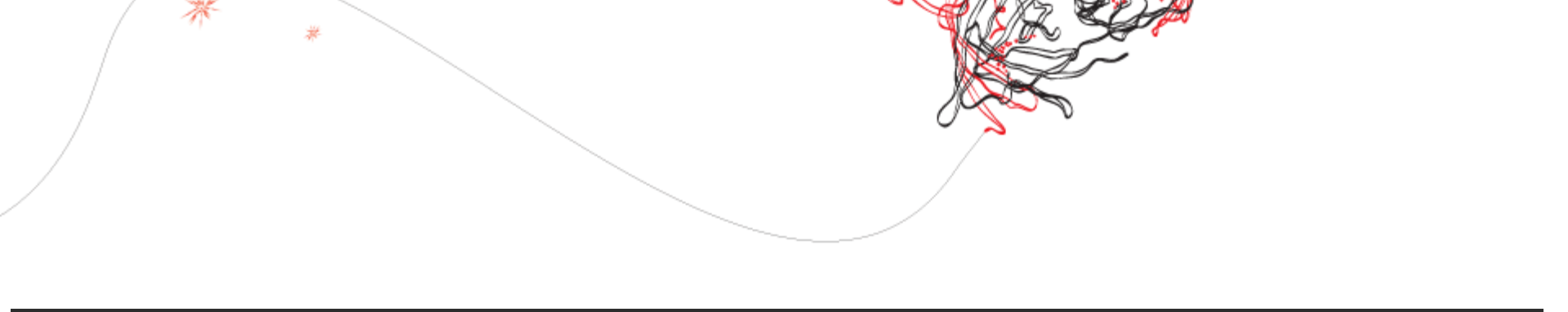
```
<top>
<num> 405
<title> cosmic events
<desc> What unexpected or unexplained cosmic
  events or celestial phenomena, such as
  radiation and supernova outbursts or new
  comets, have been detected?
<narr> New theories or new interpretations
  concerning known celestial objects made as a
  result of new technology are not relevant.
</top>
```

Creating Relevance Judgments



(thanks to Ellen Voorhees)

(thanks to Ellen Voorhees)



Home Tasks Extra logout Labels Guidelines

"haskell hash string"

You want to find out how to create a hashed value like MD5 from a string, in haskell.

(none)

Overview | Previous | neXt | next neW | next probleM | Update

(•) Non ( ) Rel ( ) Hrel ( ) Key ( ) naV ( ) Answered

Options:

[ ] There is a problem with displaying the snippet content (reference: FW13-e050-7152-01).

done Estimate the relevance of the result page, based only on the search result snippet below

A Library of Software written in C++ with full source code.

In some languages (e.g. Lisp, Prolog, and Haskell) the use of memoization can be .... This is an implementation of a string search algorithm in C++ that is a couple of times .... pHash - a Perceptiual Hash to help identify similar multimedia files ...

http://cplus.about.com/od/codelibraryfor1/A_Library_of_Software_written_in_C_with_full_source_code.htm

Home Tasks Extra logout Labels Guidelines

"cystic fibrosis treatment"

You want to know what kind of treatments are available for cystic fibrosis

You are currently studying for your retake on heart, long, and kidneys. During the part on respiratory fysiology, you come across the desease 'cystic fibrosis', and you want to know which are the possibilities for therapy. Since you are going to be a physician, results with sufficient scientific depth are welcome.

Overview | Previous | neXt | next neW | next probleM | Update

○Non ○Rel ◉Hrel ○Key ○naV

Options:

open new tab in case no content is displayed and the 'update' button does not help, or to open online video/audio.

☐I watched the video / listened to the audio online

☐There is a problem with displaying the content of this page (reference: FW13-e185-7103-01).

done Judge the relevance of the page below

TREC ASSUMPTIONS ABOUT RELEVANCE

- Relevance of one element does not affect the relevance of another element
- Relevance is a binary decision, i.e., a document is either relevant or not
- A document is relevant if it would help in writing an article about the subject
  - relevant? topicality? clarity? recency? accuracy? trustworthiness?

TREC ASSUMPTIONS ABOUT SYSTEMS

- A system is a programme
  - the user is outside the system
- A system is an input-output device
  - query in, documents out
  - although... most real searches involve interaction

HOW ABOUT THE QUALITY OF A TEST COLLECTION?

- Two concerns:
  - Consistency of the judgments: *do the results of the experiments critically depend on the particular choices of human judges*?
  - Completeness of the judgments: *do the results critically depend on the pool construction process, i.e. on the systems that participated in TREC*?

CONSISTENCY OF THE JUDGEMENTS

- Experiment: 10 topics assessed twice by two different assessors
- Dutch CLEF collection, overlap: 0.465
- TREC: overlap between: 0.421 and 0.494

  (Overlap = size of intersection of the relevant document sets divided by the size of the union of the relevant document sets.)
- (Overall agreement 93.4 %)

COMPLETENESS OF JUDGMENTS

- Can we use the collection for future experiments?
- What if my run is not judged?
- Experiment: recompute for each official run the average precision as if it was not in the pool, i.e. ignoring the relevant documents uniquely found by that run

COMPLETENESS:
WHAT IF MY RUN IS NOT JUDGED?

name	unjudged	judged	difference	unique rel.	
ut1	0.4222	0.4230	0.0008	0.2 %	55
aplmonla	0.3943	0.4002	0.0059	1.5 %	29
tnonn3	0.3914	0.3917	0.0003	0.1 %	2
humNL0	0.3825	0.3831	0.0006	0.2 %	5
tlrnltd	0.3760	0.3775	0.0015	0.4 %	10
tnoen1	0.3246	0.3336	0.0090	2.8 %	32
AmsNlM	0.2770	0.2833	0.0063	2.3 %	32
aplbiennl	0.2692	0.2707	0.0015	0.6 %	7
oce2	0.2363	0.2405	0.0042	1.8 %	21
glaenl	0.2113	0.2123	0.0010	0.5 %	8
oce1	0.2024	0.2066	0.0042	2.1 %	23
medialab	0.1600	0.1640	0.0040	2.5 %	23
EidNL2	0.1339	0.1352	0.0013	1.0 %	8 .
		mean:	0.0031	1.2 %	20
	standard deviation:		0.0027	1.0 %	15

SIGNIFICANCE TESTING

- When is one system better than another?
  - Maybe the average difference can be contributed to chance?
  - Need a reasonable amount of queries (e.g. 50), which should be a random sample of all possible queries for a given task

SIGNIFICANCE TESTING

- Two hypotheses
  - null-hypothesis H_0: there is no difference between system *A* and system *B*
  - alternative hypothesis H_1: either system *A* consistently outperforms system *B*, or sys-tem *B* consistently outperforms system *A*
- Show that, given the evaluation results, H_0 is indefensible

SIGNIFICANCE TESTING

- Test statistics should behave differently under H_0 than under H_1:
  - Paired tests: for each query the performance difference between system A and B consist of a mean difference μ and some error.

    $H_0 : \mu = 0; H_1 : \mu \neq 0;$
  - Paired t-test: assumes that errors are normally distributed. Under H_0 the distribution is Student's t
  - Paired sign test: assumes equal probability of positive and negative error. Under H_0 the distribution is binomial

A significant change ≠ a substantial change

Topic	AP System 1	AP System 2
MAP (ALL)	0.22	0.23

Unsubstantial, but significant (sign test)

A significant change ≠ a substantial change

Topic	AP System 1	AP System 2
MAP (ALL)	0.16	0.46

Substantial (on average), but insignificant (sign test)

Evaluation in the wild

- Obtaining relevance judgments
  - Using panels
  - Crowd sourcing
  - Based on clicks
- Relevance judgments are used for learning to rank
  - Crown jewels of a web search engine

CONCLUSION

- To evaluate your system, use a benchmark collection.
- Choose appropriate evaluation measures
- Base your conclusions on statistical tests

BACKGROUND READING

- Cyril Cleverdon and Michael Keen, Factors Determining the Performance of Indexing Systems, Volume 2, The College of Aeronautics, Cranfield, 1966

ACKNOWLEDGEMENTS

- Thanks to the following people for making their slides available
  - Stephen Robertson (Microsoft Research)
  - Ellen Voorhees (NIST)