Empirical research in Information Retrieval

Research Methods and Methodology, 22 November 2006

Djoerd Hiemstra

University of Twente

hiemstra@cs.utwente.nl

http://www.cs.utwente.nl/~hiemstra

Goal

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

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

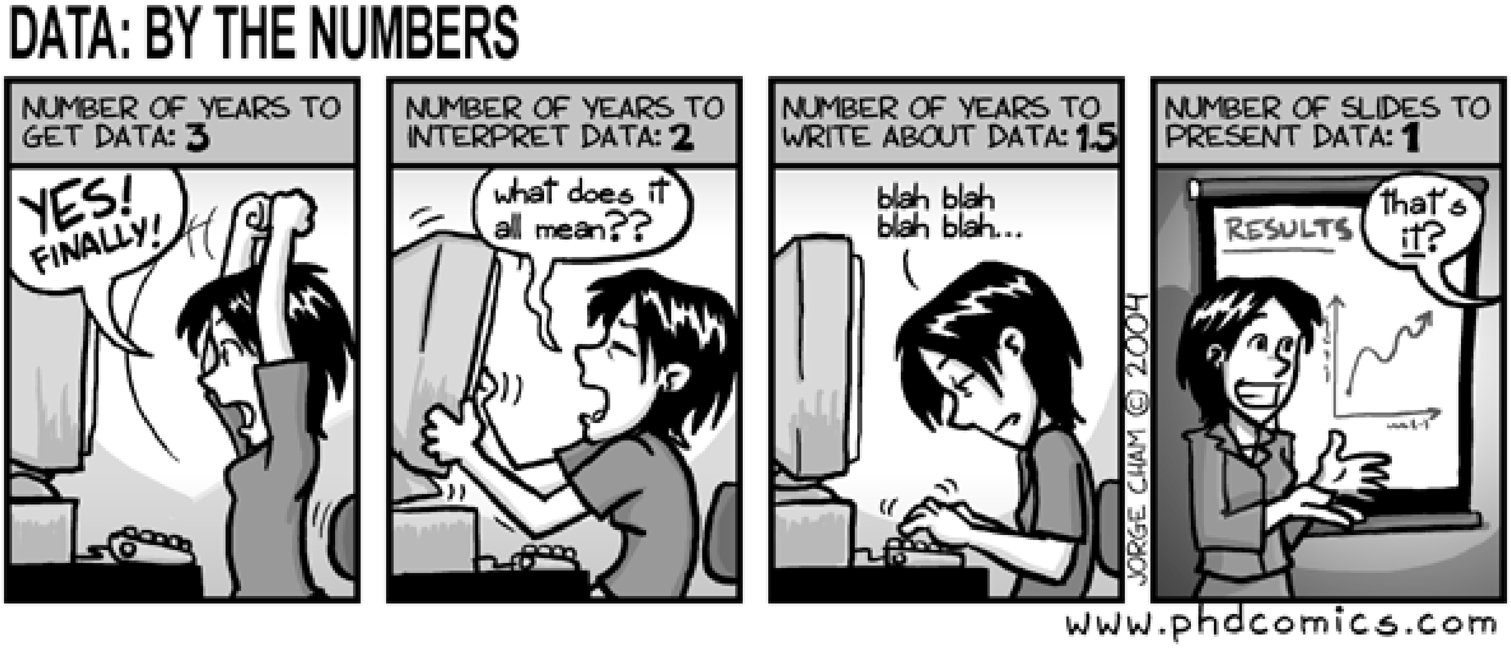
- Compile and present the results.
  - Repeat a number of times

4. Conclusion

- Supporting the hypothesis...

- or rejecting it.

Summary


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?
  - traditionally: a package of information structured by an author
- What is a request?
  - a description of a topic of interest
  - more properly, 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

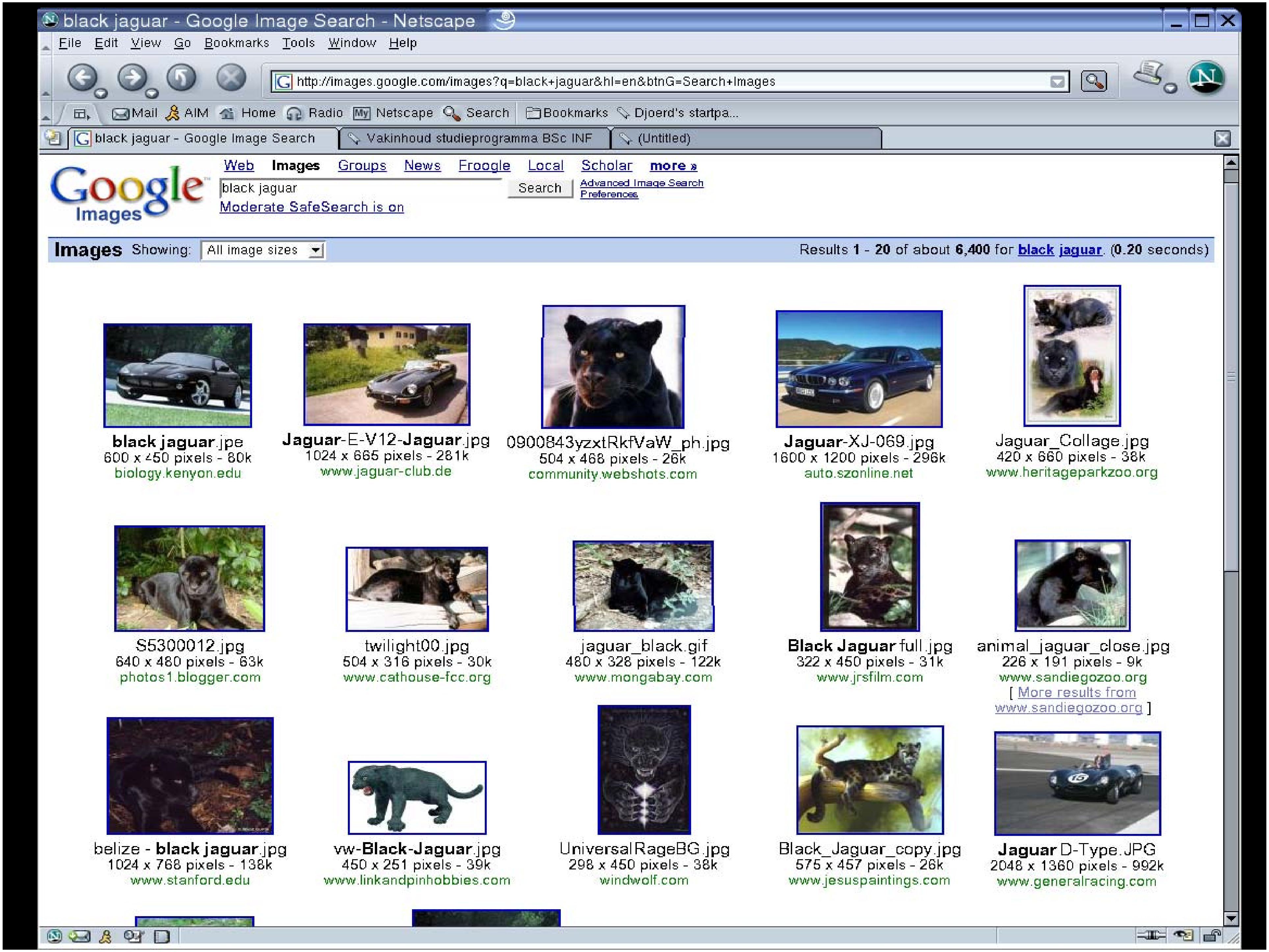
The traditional IR experiment

- 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

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 Image Search - Netscape
File Edit View Go Bookmarks Tools Window Help
Mail AIM Home Radio Netscape Search Bookmarks Djoerd's startpa...
black jaguar - Google Image Search
Vakinhoud studieprogramma BSc INF
(Untitled)
Google Images
Web Images Groups News Froogle Local Scholar more »
Search
Advanced Image Search
Preferences
Moderate SafeSearch is on

http://images.google.com/images?q=black+jaguar&hl=en&btnG=Search+Images

black jaguar

Images Showing: All image sizes

Results **1** - **20** of about **6,400** for **black jaguar**. (**0.20** seconds)

black jaguar.jpe
600 x 450 pixels - 80k
biology.kenyon.edu

Jaguar-E-V12-**Jaguar**.jpg
1024 x 665 pixels - 281k
www.jaguar-club.de

0900843yzxtRkfVaW_ph.jpg
504 x 468 pixels - 26k
community.webshots.com

Jaguar-XJ-069.jpg
1600 x 1200 pixels - 296k
auto.szonline.net

Jaguar_Collage.jpg
420 x 660 pixels - 38k
www.heritageparkzoo.org

S5300012.jpg
640 x 480 pixels - 63k
photos1.blogger.com

twilight00.jpg
504 x 316 pixels - 30k
www.cathouse-fcc.org

jaguar_black.gif
480 x 328 pixels - 122k
www.mongabay.com

Black Jaguar full.jpg
322 x 450 pixels - 31k
www.jrsfilm.com

animal_jaguar_close.jpg
226 x 191 pixels - 9k
www.sandiegozoo.org
[More results from www.sandiegozoo.org]

belize - **black jaguar**.jpg
1024 x 768 pixels - 138k
www.stanford.edu

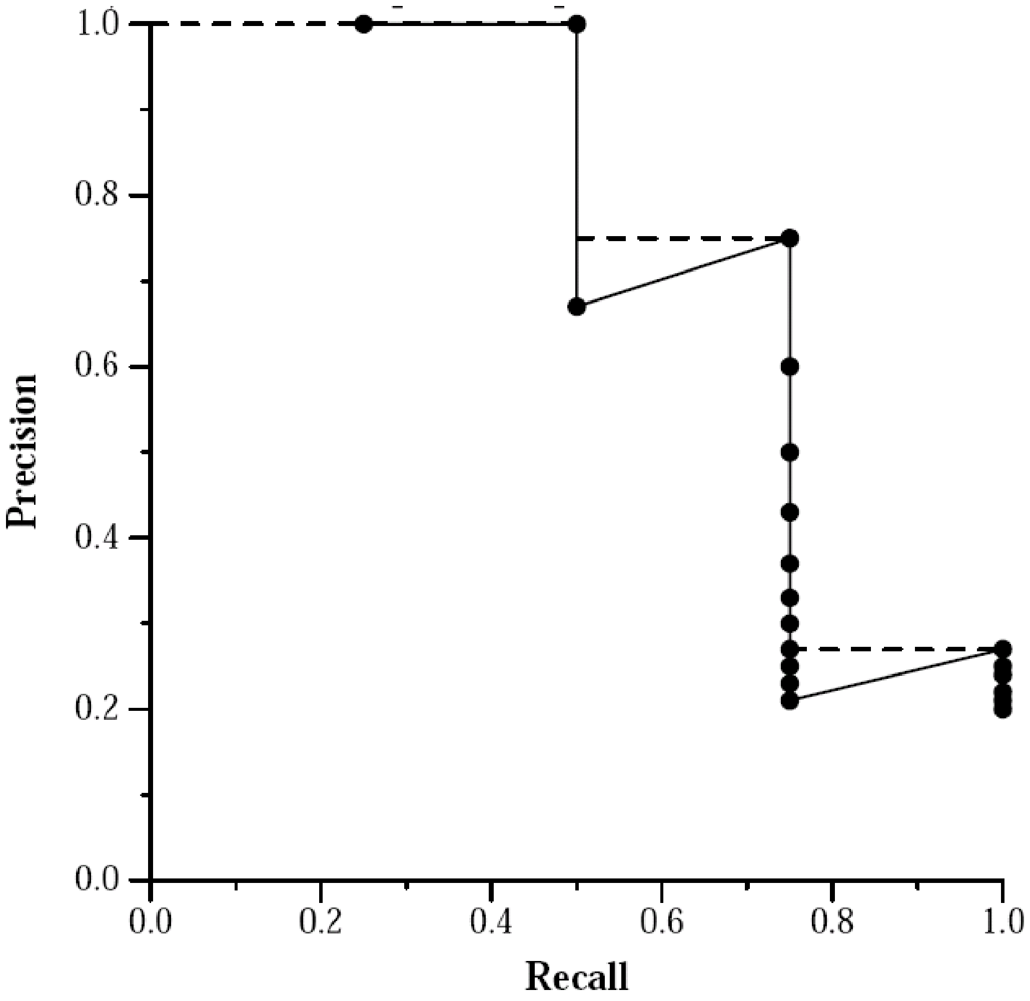
vw-**Black**-**Jaguar**.jpg
450 x 251 pixels - 39k
www.linkandpinhobbies.com

UniversalRageBG.jpg
298 x 450 pixels - 38k
windwolf.com

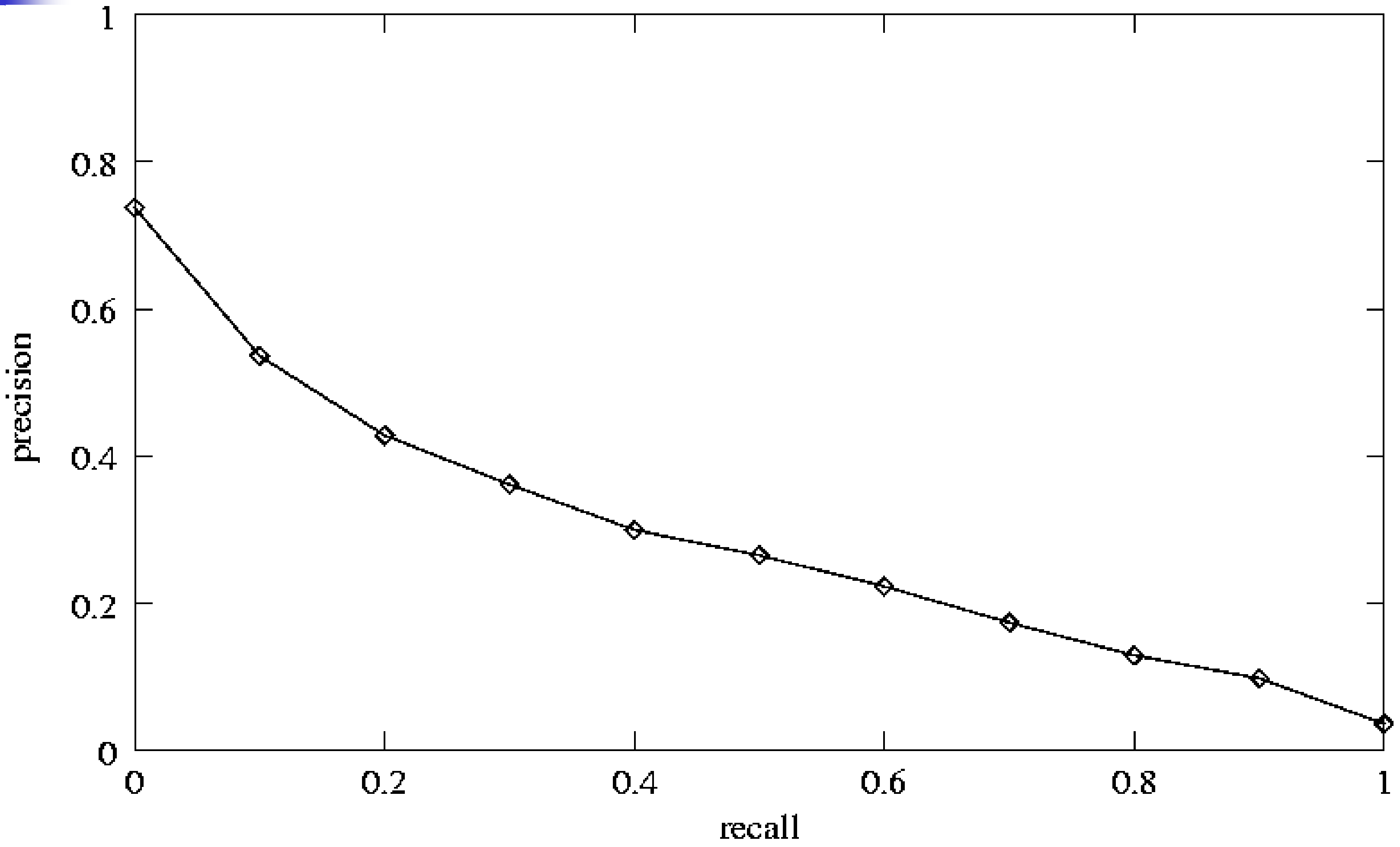
Black_Jaguar_copy.jpg
575 x 457 pixels - 26k
www.jesuspaintings.com

Jaguar D-Type.JPG
2048 x 1360 pixels - 992k
www.generalracing.com

Recall-precision plot



Recall-precision plot



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

The TREC paradigm

doing laboratory tests

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

What is TREC?

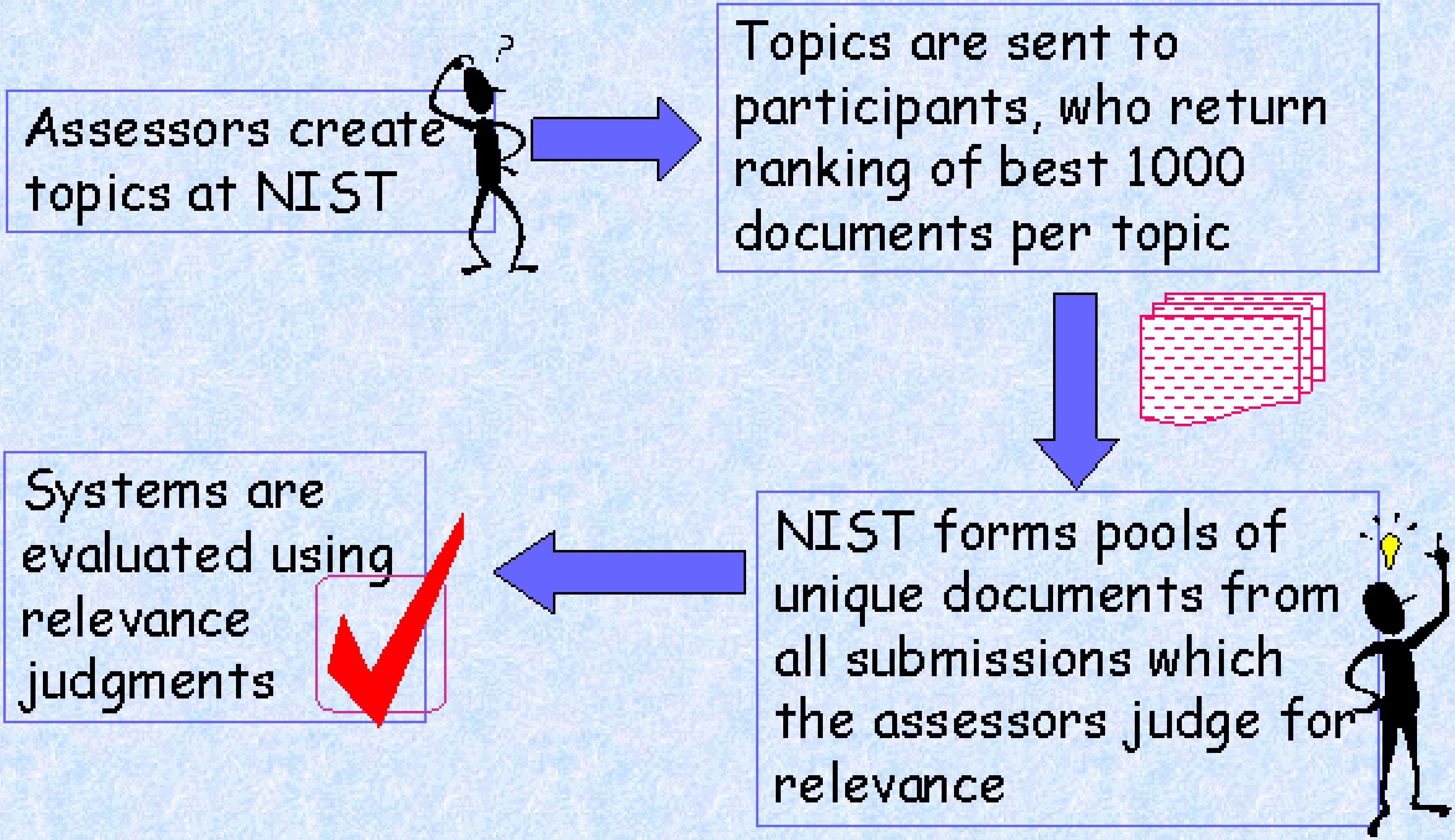
- Competition/collaboration between IR research groups world-wide
- Run by the US National Institute of Standards and Technology (NIST)
- TREC provides:
  - common test collections
  - common tasks
  - common measures
  - common evaluation procedures

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

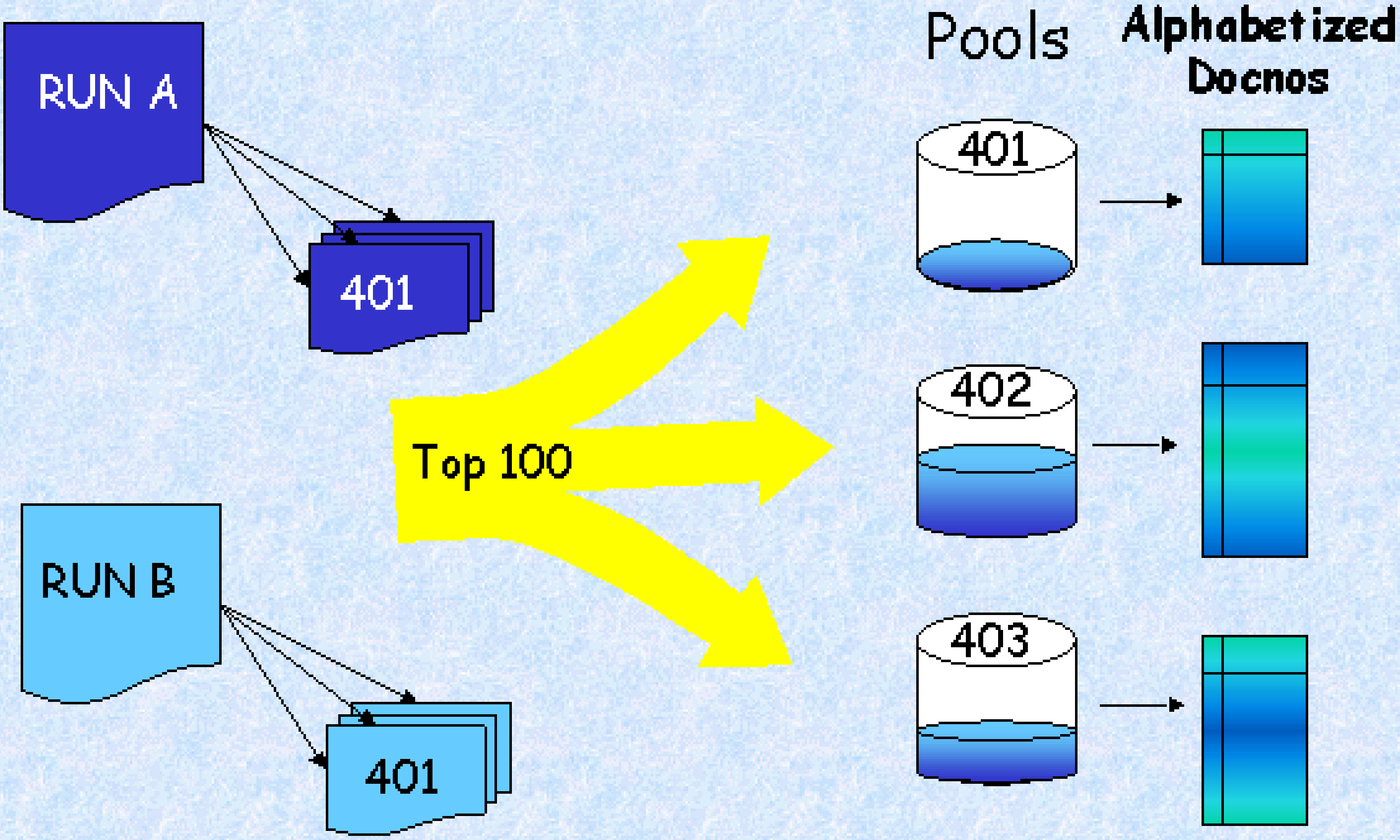
Systems are evaluated using relevance judgments

(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)

TNO

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 of the judgments: What if my run is not judged?

run name	unjudged / judged avg.prec.		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
humNL01x	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
EidNL2001A	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 system 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

Conclusion

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

Acknowledgements

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

Some background reading

- Stephen Robertson, "Evaluation in Information Retrieval", In European Summer School on Information Retrieval 2000, Lecture Notes in Computer Science, Springer-Verlag, pages 81-92, 2000
- David Hull, "Using Statistical Testing in the Evaluation of Retrieval Experiments", In Proceedings of the 16th ACM Conference on Research and Development in Information Retrieval (SIGIR), ACM Press, pages 329-338, 1993
- Donna Harman, "Common Evaluation Measures", In Proceedings of the 13th Text Retrieval Conference, Appendix A, NIST Special Publications, 2005