Experimental Evaluation: Agenda

- Why evaluate?
- Relevance: What is it good for?
- "System-oriented" Evaluation
  - Unranked measures
  - Ranked measures
- User-oriented Evaluation

Why evaluate?

THE SIGNIFICANCE OF THE
CRANFIELD TESTS
ON INDEX LANGUAGES
by
Cyril W. Cleverdon

1946 saw the lifting of the security restrictions on large numbers of scientific and technical reports which had been written during World War Two. Pre-war virtually all publication had been in journals, and the report format was strange and unfamiliar, both for the scientific community and for librarians. As such they presented new challenges; the administrative problem of actually being able to obtain copies of the reports was tackled by setting up new government agencies with direct responsibility for collecting and making the reports generally available. The more difficult problem lay in revealing and making accessible the intellectual content of the papers. At that time there were two conventional types of index and two major indexing techniques. An index could be in the form of a card catalogue, as found in most libraries, or alternatively in printed form as, for example, an annual accumulation of an abstract journal. Regarding the techniques of indexing, in Europe there was a tendency to use a classified system, whereas in America the usual practice was to use alphabetical subject headings.

With the deluge of scientific and technical reports, both the physical form of the index and the indexing techniques came under strong attack. While card catalogues and printed indexes still exist, there has been over the past forty years a steady and reasonably placid progress of mechanised systems, culminating now in online systems and CD-ROM, but there was nothing placid about the development of indexing techniques. The early 50s saw many attempts to depart from the conventional systems. In England a small group met regularly to discuss the development of facet classification. This technique breaks away from the conventional enumerative or hierarchical classification, such as the Dewey Decimal Classification, and relies on subject analysis and synthesis by facet principles. However the main thrust of the new methods was in America, from such people as Calvin Mooers with Zatorcording, James Perry with semantic factoring and, in particular, Mortimer Taube. Taube, a government librarian, analysed some 40,000 subject headings used in a major card catalogue and found that the headings were combinations of only some 7,000 different words. He therefore proposed using these individual words as index terms which would be coordinated at the searching stage. This became known as the Uniterm System.

These new techniques generated considerable argument, not only between the proponents of the different systems, but also among the library establish-ment, many of whom saw these new methods as degrading their professional mystiques.

This briefly is the context in which I started my research. In 1946,

"Controversy over the new methods was still raging, with extravagant claims on one side being countered by absurd arguments on the other side, without any firm data being available to justify either viewpoint."

Why evaluate?

Systematic evaluation allows us to make meaningful comparisons between systems and between techniques:

"Is the new weighting scheme 'better' than the old one?"

"How does indexing technique *A* compare to technique *B*?"

"Does my system work as well on medical text as it does on newswire text?"

IR has a long history of empirical evaluation:

The first empirical IR studies began in 1957 with the “Cranfield Studies.”

Cyril Cleverdon
1914-1997

1,400 Documents
225 Queries

This approach became known as the “Cranfield model” of system evaluation...

IR has a long history of empirical evaluation:

The Cranfield studies set a "pro-evaluation" tone within the new field...

... but there was little coordination, which made comparison difficult:

Karen Spärck Jones
1935-2007

"... the most striking feature of the test history of the past two decades is its lack of consolidation..."

Information Retrieval Experiment. 1981

IR has a long history of empirical evaluation:

In 1990, DARPA asked NIST to build a very large test collection for an IR development program...

... NIST realized that this test collection could be useful to the field as a whole, and arranged for its public release.

In 1992, NIST hosted the first TREC conference.

The Text REtrieval Conference (TREC) has been the foundation of modern IR evaluation.

TREC has included a wide variety of tracks:

- Ad-hoc retrieval
- Multi/Cross-lingual retrieval
- Question answering
- Web
- Genome
- Speech
- Medical
- etc.

It happens each year, and anyone can participate!

The Text REtrieval Conference (TREC) has been the foundation of modern IR evaluation.

The TREC collections and topics are all publicly available...

... and so are frequently used as reference collections by IR system developers and evaluators.

TREC-style evaluations generally follow the Cranfield model, and consist of:

A *document collection*;

A set of *information needs* (not queries!) that might be satisfied by documents in the collection;

A set of human-generated *relevance judgments*, indicating which documents are germane to which needs.

A set of *information needs* (not queries!) that might be satisfied by documents in the collection:

Number: 312
Title: Hydroponics

Description: Document will discuss the science of growing plants in water or some substance other than soil.

Narrative: A relevant document will contain specific information on the necessary nutrients, experiments, types of substrates, and/or any other pertinent facts related to the science of hydroponics. Related information includes, but is not limited to, the history of hydroponics, advantages over standard soil agricultural practices, or the approach of suspending roots in a humid enclosure and spraying them periodically with a nutrient solution to promote plant growth.

Example topic from TREC-6 ad-hoc.

Number: 179
Description: Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression.

Example topic from 2012 TREC-Med

Besides TREC, there are many other evaluation campaigns and test collections.

Cross-Language Evaluation Forum (CLEF)

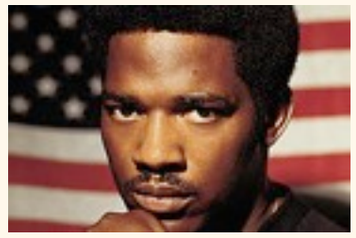
NTCIR

NIST post-TREC corpora (GOV-2, etc.)

RCV1, etc.

Relevance: What is it good for?

(Absolutely Nothing!)



Edwin Starr
1942-

OK, that's an exaggeration...

But this model of evaluation has implicit assumptions:

1. Information needs are static and fully-formed at query time...
2. A document's relevance is binary...
3. A document's relevance can be objectively assessed (and is not person-dependent)...
4. Each document's relevance is independent of any other document's relevance.

Experimental Evaluation: Agenda

- Why evaluate?
- Relevance: What is it good for?
- "System-oriented" Evaluation
  - Unranked measures
  - Ranked measures
- User-oriented Evaluation

Unranked retrieval evaluation:

Precision:

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

Recall:

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

$$P = \frac{tp}{tp + fp} \qquad R = \frac{tp}{tp + fn}$$

Why not just use classification accuracy?

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

$$A = \frac{tp + tn}{tp + fp + fn + tn}$$

"Fraction of classifications that are correct"

Most of the time, the data are extremely skewed (>>>90% "not relevant").

“F-measure” is a good composite measure.

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

The F-measure is the *weighted harmonic mean* of P and R; the weight parameter indicates their relative importance.

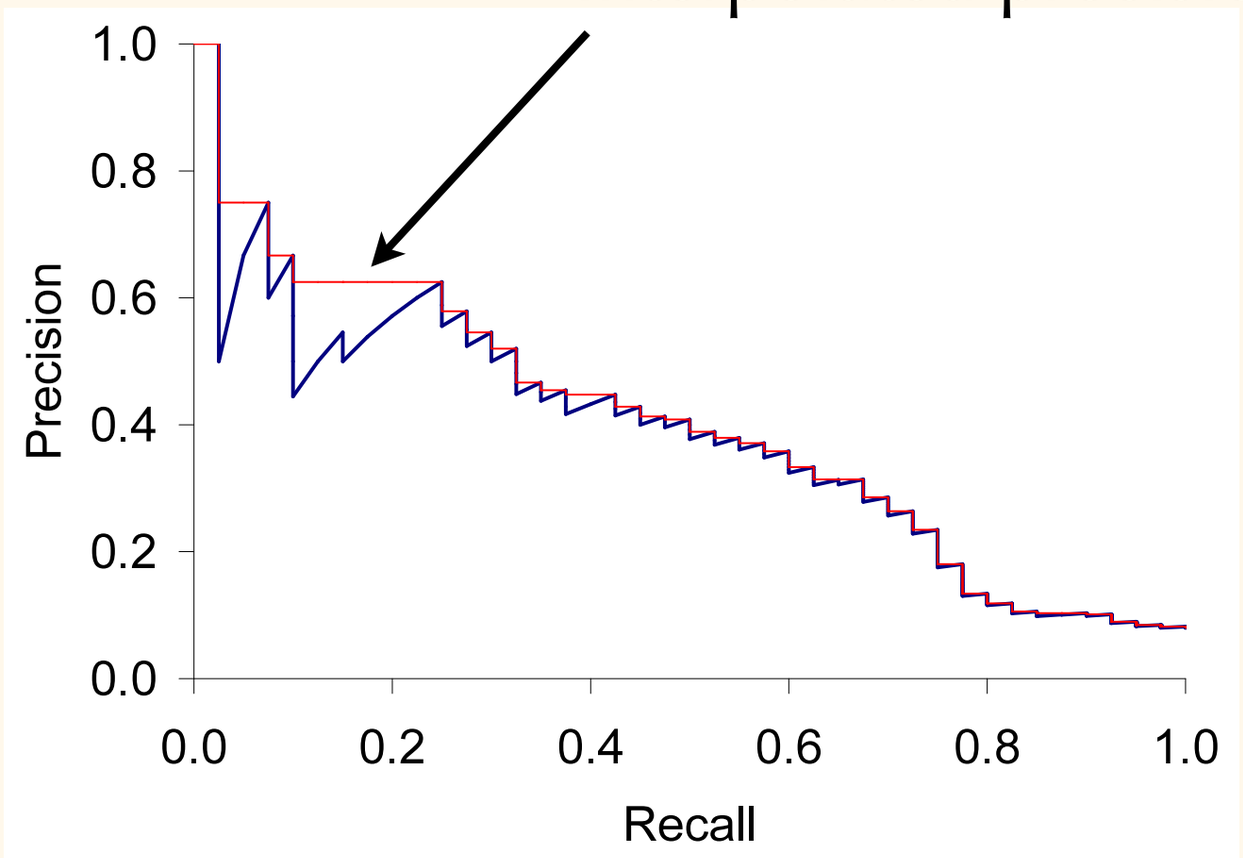
When P and R are equally important, $F_{\beta=1} = \frac{2PR}{P+R}$

The main problem with unranked retrieval:

Most of the time, we want the most relevant results first.

We can extend precision and recall to incorporate ranking information.

"Interpolated precision"



Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

Often reported as an *eleven-point interpolated average precision* (averaged across all topics).

We can extend precision and recall to incorporate ranking information.

Also commonly reported: "precision at *k*" (system performance at one point on the p/r curve).

Advantage: easy to understand; more realistic model for many scenarios (users only usually look at first few results, etc.);

Disadvantage: highly unstable; doesn't average well; is highly influenced by the number of relevant documents.

We can extend precision and recall to incorporate ranking information.

Another popular metric is “Mean average precision”.

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

“Average precision” for a single query is the mean of the precision scores at the rank of each relevant result...

MAP is just the mean of all the AP scores for each topic.

We can extend precision and recall to incorporate ranking information.

One more: Normalized Discounted Cumulative Gain

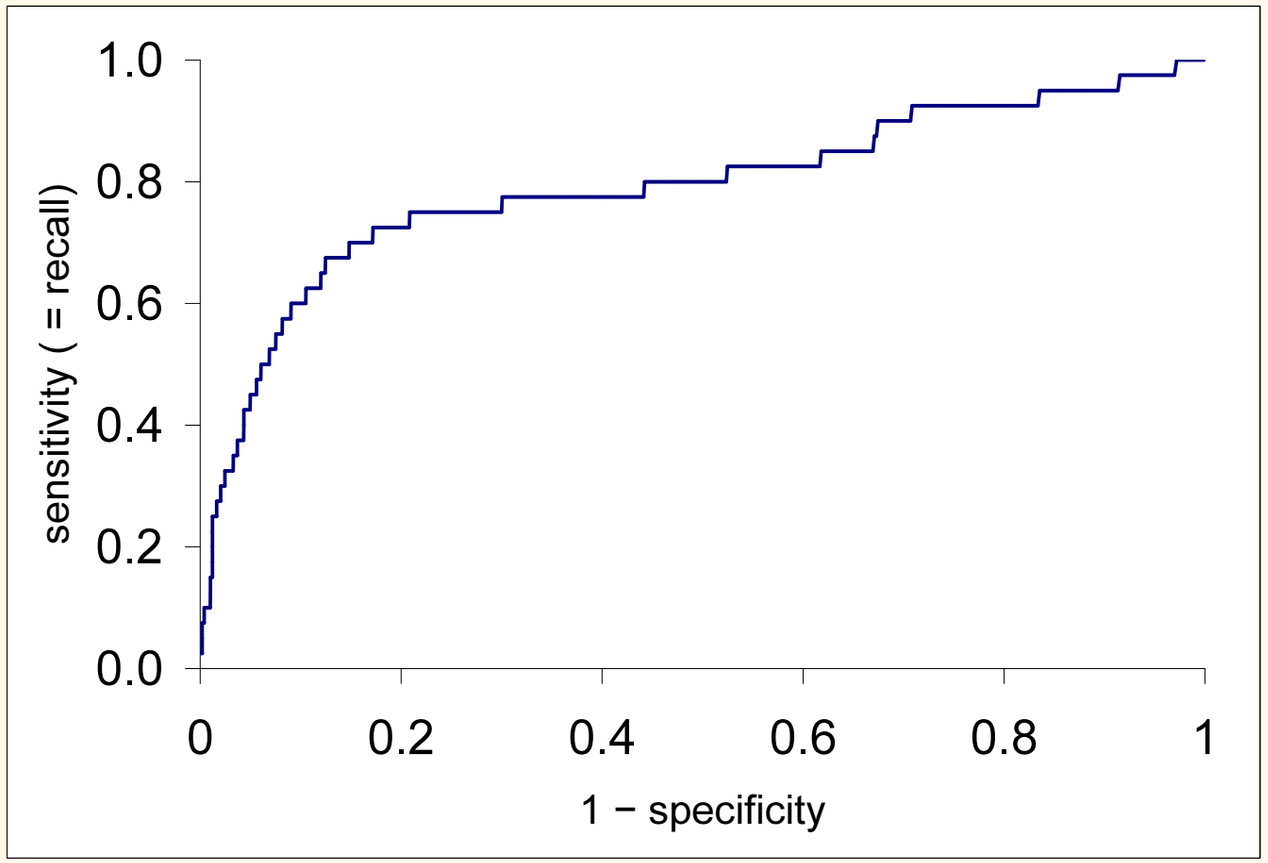
$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{\log_2(1+m)}$$

NDCG works on “graded” (non-binary) relevance judgments...

... the intuition is that highly-relevant articles that score badly should be penalized more than less-relevant articles that score well.

We can extend precision and recall to incorporate ranking information.

ROC curves are also commonly used:



NB: “1-Specificity” == FPR

One typically reports the area under the curve (AUC).

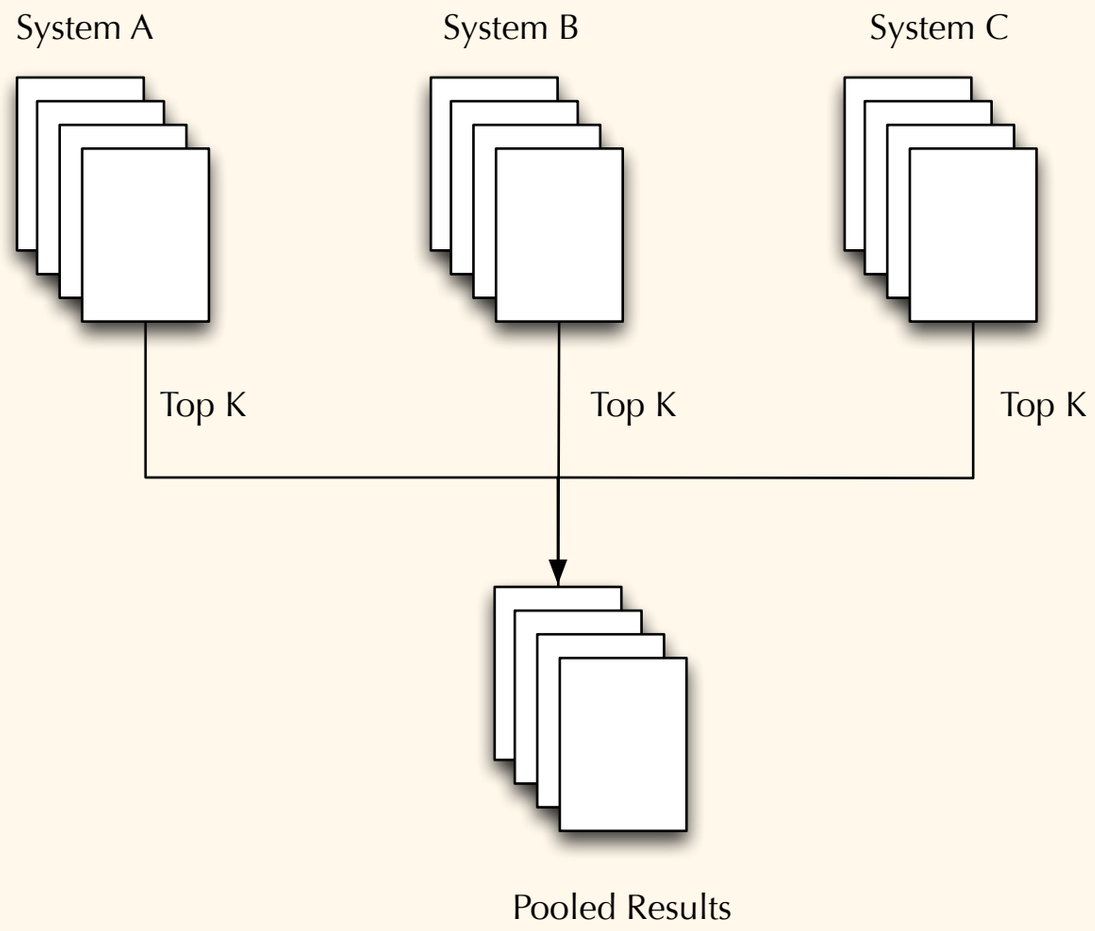
There are many, many other evaluation metrics.

Quick sidebar: assessing relevance

Ideally, judges look at all topics, and all articles, and assign complete pairwise relevance judgments.

For *tiny* collections, this works... but what if your collection isn't tiny?

The solution: pooling

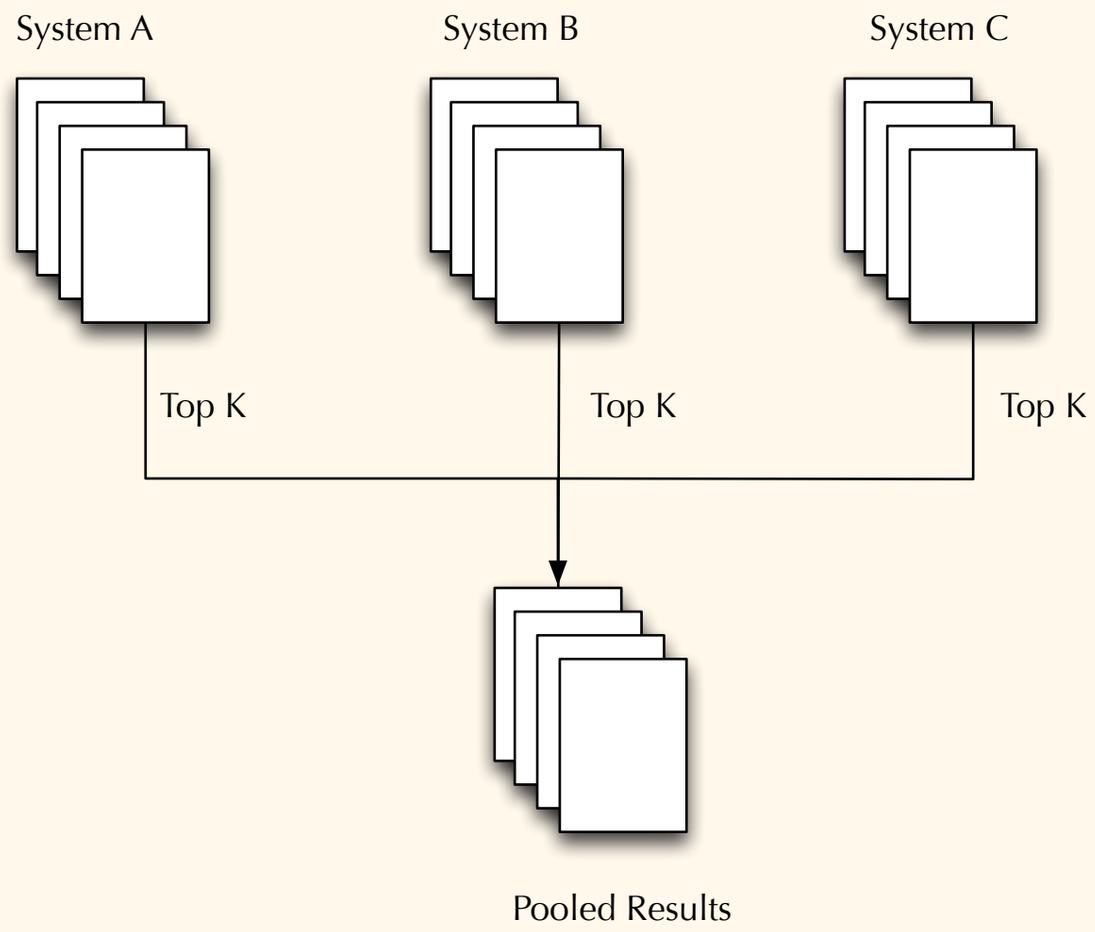


Quick sidebar: assessing relevance

The main problem with pooling is that if one system finds a bunch of unique results...

... they might not get included in the pool...

... thereby penalizing that system.



Quick sidebar: assessing relevance

Another consideration: the performance of human judges.

Judges often disagree about a document's relevance.

One common approach is to calculate *kappa* scores, to measure how well your judges agree with one another.

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

As a rule of thumb, a kappa of >0.8 is "good", 0.67-0.8 is "fair", and <0.67 is "poor".

Changing judges rarely affects relative ranking of systems!

Manning, et al. make a good point:

Measuring ranking performance is really a proxy for what we *really* care about: system utility.

Is the system solving our users' problems?

In other words, are we helping them meet their information needs?

Also: besides ranking, we might care about query service speed, feature set, indexing time and flexibility, etc.

Experimental Evaluation: Agenda

- Why evaluate?
- Relevance: What is it good for?
- "System-oriented" Evaluation
  - Unranked measures
  - Ranked measures
- User-oriented Evaluation

What does it mean to be user-oriented?

Focus on measuring whether users are able to get what they need out of the system.

We might measure:

Clicks: how many, where are they, etc.

Return visits: do they come back and use us again?

Sales (if relevant to our problem area)

Speed: can they finish their task more quickly?

Use amount: do they look at more or different results?

“Satisfaction”: how “happy” are they?

We can also evaluate systems in context: with actual users doing actual tasks.

A formal user study has many considerations:

Task definition: what will you have the subjects do?

Study subjects: who will your subjects be?

Experimental conditions: what perturbation(s) will you introduce to the environment? What will you use as a control?

Measurements: what will you be measuring?

Each of these can introduce bias!

Of special importance: eliminating condition-ordering effects.

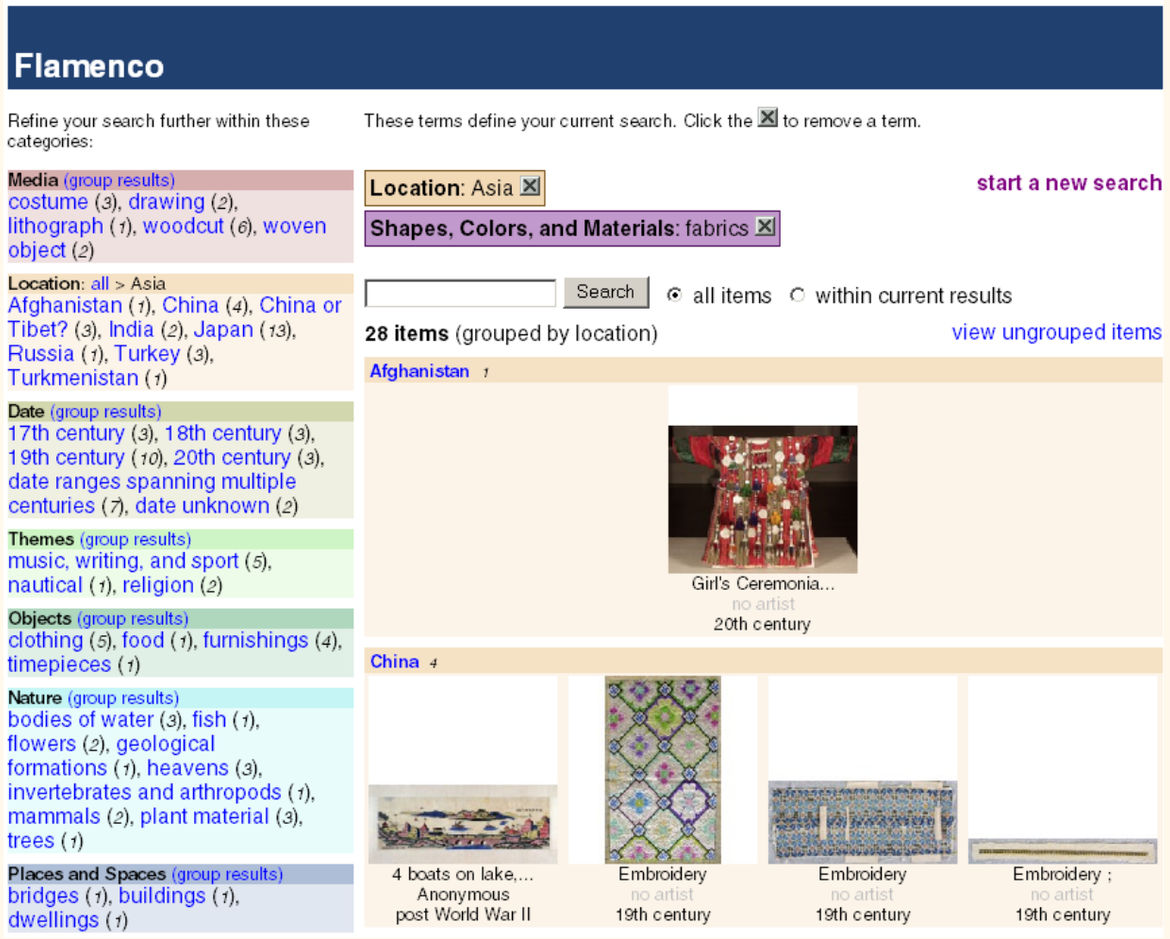
The order that your subjects are exposed to the different conditions can affect their performance!

As subjects perform a task, they typically get better at it ("learning effect");

Sometimes, though, subjects get *slower* as the test goes on ("fatigue effect");

If one condition is "better", the order in which it is seen can affect users' opinions of other conditions.

Of special importance: eliminating condition-ordering effects.



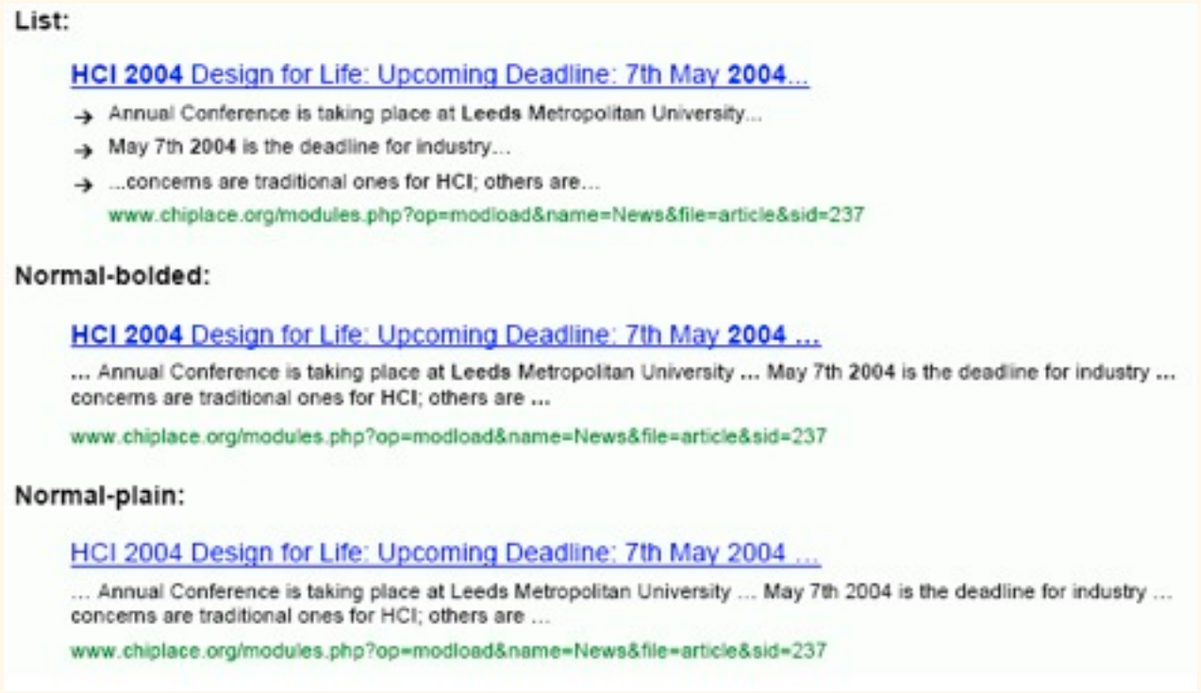
When this “faceted” interface was shown first, users’ subjective ratings of the baseline interface were lower than when the baseline was shown first.

K.-P. Yee, K. Swearingen, K. Li, and M.A. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)*, pages 401–408. ACM New York, NY, USA, 2003

One common way to address this:

Using a “Latin Square” design to balance the number of users exposed to each interface in each position.

This also allows us to use ANOVA to look for interaction effects!



List:

HCI 2004 Design for Life: Upcoming Deadline: 7th May **2004**...

- Annual Conference is taking place at **Leeds** Metropolitan University...
- May 7th **2004** is the deadline for industry...
- ...concerns are traditional ones for **HCI**; others are...

www.chiplace.org/modules.php?op=modload&name=News&file=article&sid=237

Normal-bolded:

HCI 2004 Design for Life: Upcoming Deadline: 7th May **2004** ...

... Annual Conference is taking place at **Leeds** Metropolitan University ... May 7th **2004** is the deadline for industry ... concerns are traditional ones for **HCI**; others are ...

www.chiplace.org/modules.php?op=modload&name=News&file=article&sid=237

Normal-plain:

HCI 2004 Design for Life: Upcoming Deadline: 7th May 2004 ...

... Annual Conference is taking place at Leeds Metropolitan University ... May 7th 2004 is the deadline for industry ... concerns are traditional ones for HCI; others are ...

www.chiplace.org/modules.php?op=modload&name=News&file=article&sid=237

PARTICIPANT	FIRST	SECOND	THIRD
Group	Task Set	Task Set	Task Set
1	Plain (A)	Bold (B)	List (C)
2	Plain (B)	Bold (C)	List (A)
3	Plain (C)	Bold (A)	List (B)
4	Bold (A)	List (B)	Plain (C)
5	Bold (B)	List (C)	Plain (A)
6	Bold (C)	List (A)	Plain (B)
7	List (A)	Plain (B)	Bold (C)
8	List (B)	Plain (C)	Bold (A)
9	List (C)	Plain (A)	Bold (B)

Figures taken from Aula 2004 via Hearst 2009.

When direct user studies are not an option, there are other options:

Log analysis;

A-B testing;

“Crowd-sourced” (Mechanical Turk) tasks;

Etc.

Each has its own experimental design considerations!

Next up: relevance feedback.