

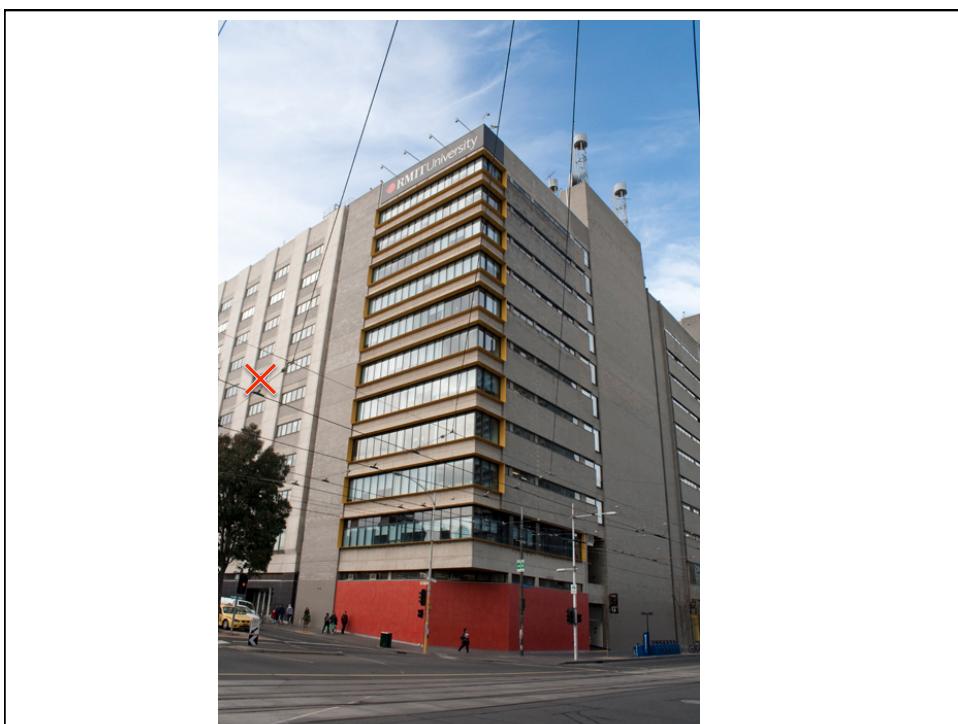
The Psychophysics of Search: Why Cranfield and User Studies (Might) Disagree

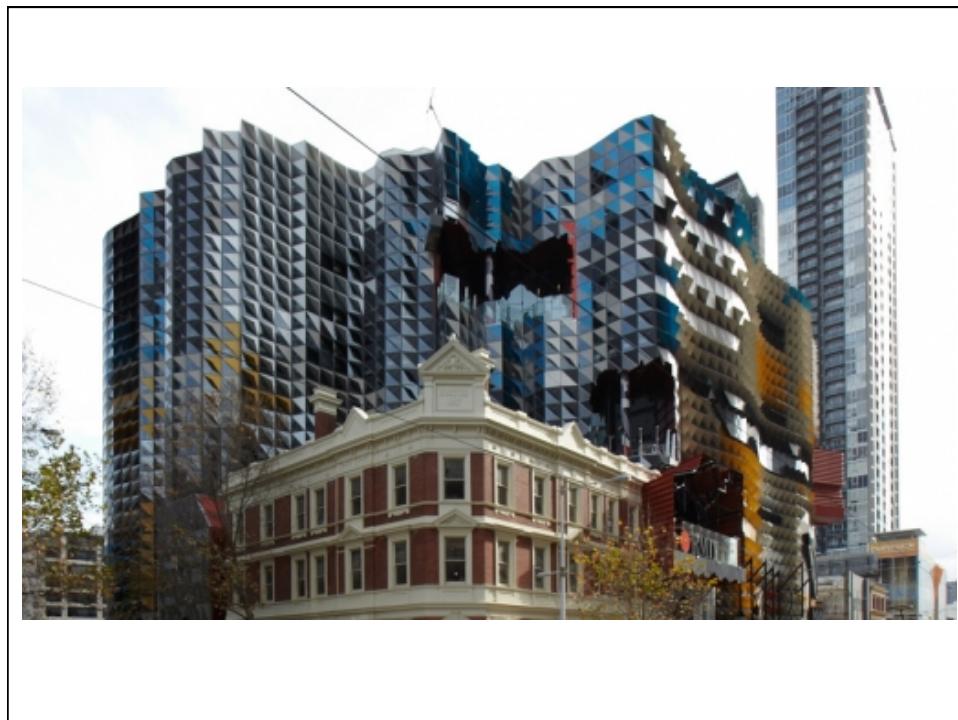
Falk Scholer

School of Computer Science & IT









WEB IMAGES VIDEOS MAPS MORE

bing godzilla sightings usa | 🔎

32,200 RESULTS

X [Godzilla arouses atomic terror - USATODAY.com](#)
www.usatoday.com/life/movies/news/2006-08-28-godzilla-dvd_x.htm ~
 Aug 29, 2006 · USA TODAY's Mike Snider geeks out on four **sightings**: 1962: King Kong vs. **Godzilla** Universal, on DVD with King Kong Escapes, \$20 This battle of ...

✓ [Godzilla - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Godzilla ~
 Name · Attributes · Movie appearances · Television and ... · Cultural impact
 The creature was also depicted as being green in the Hanna-Barbera cartoon and a number of toys in the **United States** prior to the Trendmasters toy ...

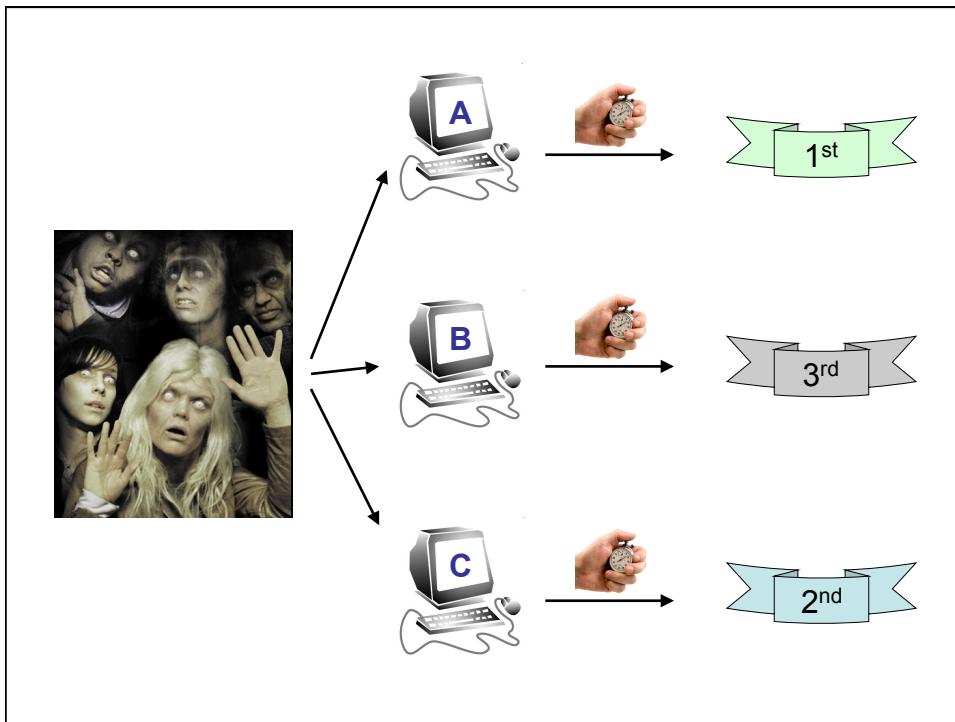
✓ [Buy Amoxicillin Without Prescription » True Life **Godzilla Sighting!**](#)
www.cliffdweller.com/wordpress/2005/07/03>true-life-godzilla-sighting.htm ~
 True Life **Godzilla Sighting**! ... Kim Pullen-Unger on Dollywood, USA (and the rest of The South) Paul on Human ...

✓ [Godzilla - Godzilla Wiki - The wiki of King Kong, **Godzilla** and more](#)
godzilla.wikia.com/wiki/Godzilla ~
 This series described the adventures and confrontations of **Godzilla** in the **United States**. ...
Sightings. Ultra Q-the monster, Gomess, is actually an altered **Godzilla**. ...

X [Varan - Welcome to the Monstrous **Godzilla** section](#)
godzilla.monstrous.com/varan.htm ~
Sightings Movie Trailers Video Games Trailers Music ... Destroy All Monsters living on Monsterland with many other kaiju including **Godzilla**.

Batch System Evaluation

- Inputs:
 - Set of queries
 - Collection of documents
 - Relevance judgements for each query-document pair
- Output:
 - Your favourite system performance metric (MAP, nDCG, P@10, RBP, RR, ...)



User-based Evaluation

- Inputs
 - Group of (human) users
 - Specific search tasks
 - Different retrieval systems
- Output
 - Time to complete task
 - Success/failure
 - Satisfaction
 - ...

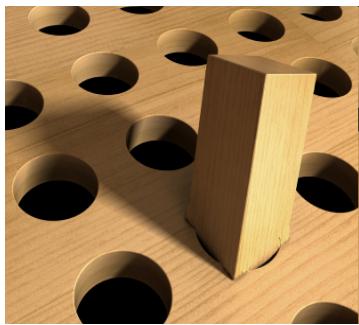
Conflicting Conclusions

- **Batch experiment:** System A is significantly better than System B
- **User evaluation:** No difference between System A and System B
 - Hersh and Turpin 2000, 2001
 - Allan, Carterette and Lewis, 2005
 - Turpin and Scholer 2006
 - Al-Maskari, Sanderson and Clough, 2008
 - Smith and Kantor, 2008
 - ...

The Rest Of This Talk

1. (Mis)matching metrics
 - User study
2. Measuring user relevance behaviour
 - Psychophysics
 - Average split agreement

Mismatching Metrics



- Metrics include assumptions about user behaviour and search tasks
- If not chosen carefully, may not reflect what the user is actually doing

Relevance Judgements

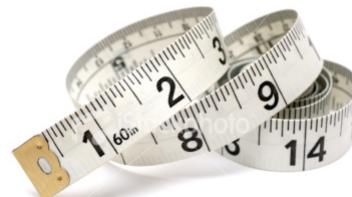
- Binary relevance
 - Used for many batch metrics
- Multi-level relevance
 - Most people can distinguish between the usefulness of documents
- TREC relevance (Web & Terabyte Tracks)
 - Highly relevant (2)
 - Relevant (1)
 - Non-relevant (0)

Some Binary Relevance Metrics

- Precision at cutoff N (P@N):
 - the number of relevant items returned in the top N ranked results
- Mean Reciprocal Rank (MRR):
 - the reciprocal of the rank position at which the (single) relevant item is found

Some Binary Relevance Metrics

- Mean average precision (MAP):
 - The mean of the precision scores at each relevant item returned in a result list
 - Has a recall component (score is normalised by the count of all relevant items in collection)



Multi-level Relevance Metrics

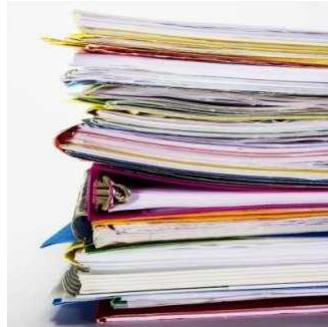
- [Normalised] [Discounted] Cumulative Gain ([n][D]CG):
 - Different gain values are earned depending on the usefulness of each retrieved document (multi-level relevance)
 - Contributions can be discounted to reflect the effort of reading down ranked list
 - The score can be normalised to make it comparable across topics

MAP Versus Simple Search Tasks

- [Turpin and Scholer, SIGIR 2006]
- Users were presented with answer lists at different levels of MAP
- No correlation with simple web search tasks
 - Time to find first relevant document
 - Number of answers found in a set period of time
- But, MAP has a recall component...

A Search User-Study

- 40 users recruited from RMIT University
 - All were CS&IT students
 - Most were familiar with online searching



- Documents and topics from TREC GOV2 collection
 - 426 Gb crawl of .gov domain from 2004
 - 24 topics and corresponding relevance judgements used

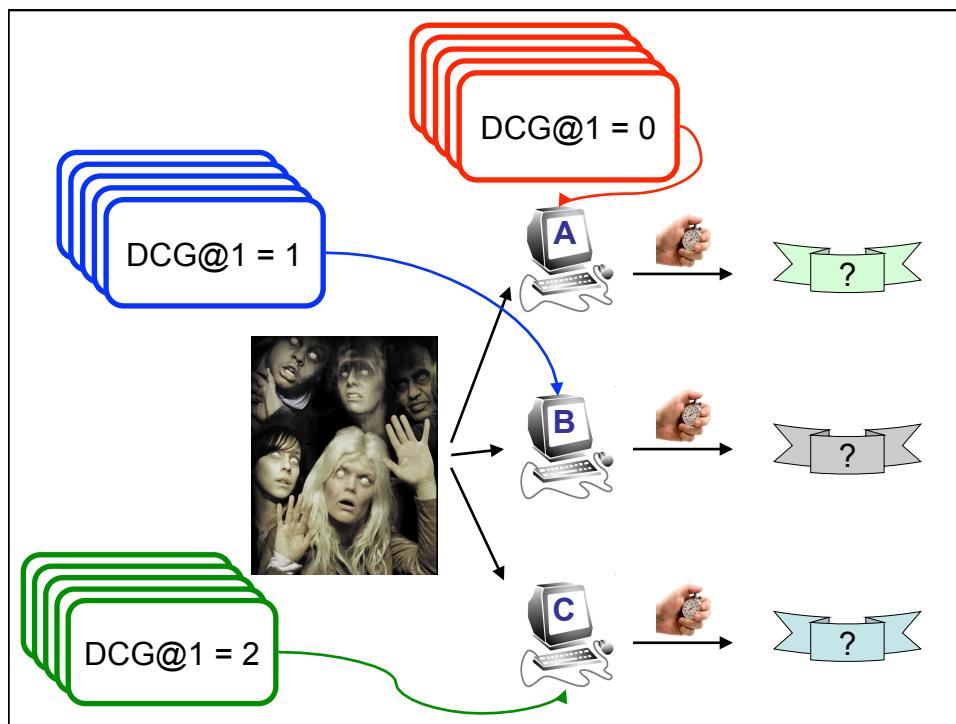
Search Systems

- Answer lists constructed to meet different levels of batch metrics (P@1, DCG@1)
- Since relevance judgements have three levels, there are three possible “systems”
- Rank 1 can vary, $X = \{0, 1, 2\}$
- To reduce variation, the other 9 positions were assigned consistent relevance levels

X, 1, 1, 1, 0, 2, 0, 0, 1, 0

Search Task

- “Imagine that your boss has come running into the room and urgently needs information. He gives you a very quick topic description, and you have only a few minutes to find a document that is useful (that is, contains some information about the requested topic).”



Search Session

1. Information need displayed to user
(description and narrative field from TREC)



“What repairs have been made on the Hubble telescope?

Not relevant are documents such as lists of resources, inquiries or photos of or by Hubble that provide no information on repairs, unless the captions discuss or describe repairs.”

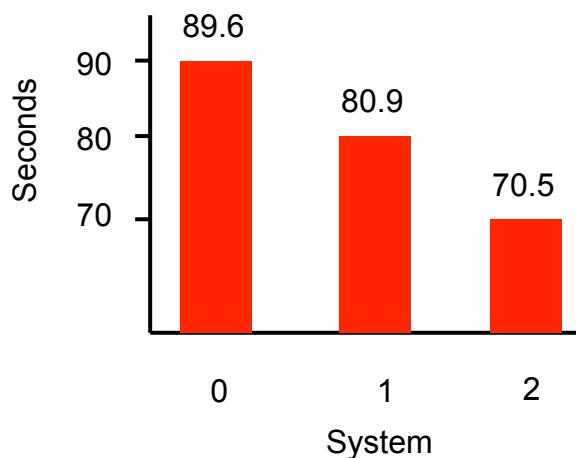
Search Session...

2. User enters query
3. Results list displayed (at a specific metric level)
 - Each item includes the document title, and a short query-biased summary
4. User can select a document for viewing, and then
 - Save, or
 - Close and return to results list

Search Session...

- Saving one document completes the search task for that topic
- 37 users carried out searches on 24 topics, with each of the 3 systems
 - But due to fatigue, only first 48 of 72 searches are analysed

Median Time



($p < 0.0001$, Kruskal-Wallis test)

DCG@1

- Can compare all three levels of metric
 - 0 vs 1
 - $p = 0.0989$
 - 0 vs 2
 - $p < 0.0001$
 - 1 vs 2
 - $p = 0.0046$
- Marginally relevant documents should be combined with non-relevant documents when folding to binary (unlike current practice!)

P@1

- Relevance needs to be folded into binary scale
- Differences are statistically significant
 - 0 vs (1 and 2)
 - $p = 0.0002$
 - (0 and 1) vs 2
 - $p < 0.0001$

Conclusions (Metric Mismatch)

- Metrics need to be matched with tasks carefully
- Reporting results with one metric and expecting this to reflect “general” search behaviour is not appropriate
- Current understanding of “real” matches is limited

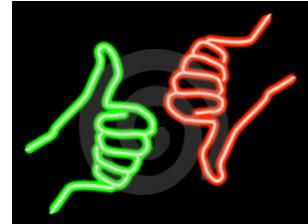
Part 2: Relevance Profiling

- Aim: to account for **relevance mismatch** between batch judgements and users
- Profiles are constructed based on user relevance preferences
- Profiles can be compared against other users, or (batch) relevance judges



Mismatching Relevance Profiles

- Judges are usually given some guidelines to follow
 - “If any **part** of the document contains information that you would **include in a report**, it’s relevant”
- Users in different experiments may have different criteria



Psychophysics

- The study of the relationship between stimuli and perception
 - Enables a *subjective* perceptual experience to be measured by reference to a stimulus
- The method of *constant stimuli* can be used to determine a **threshold** – the intensity of a stimulus required for it to be consciously experienced

Do you see the light?



Do you see the light?

Do you see the light?

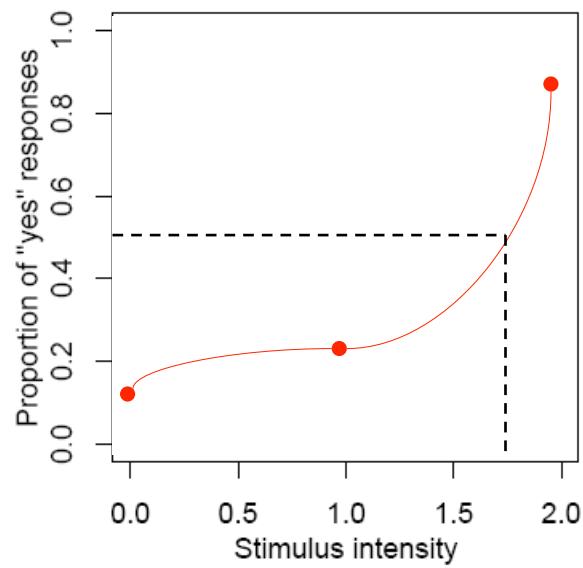


Do you see the light?



Do you see the light?

See the light!

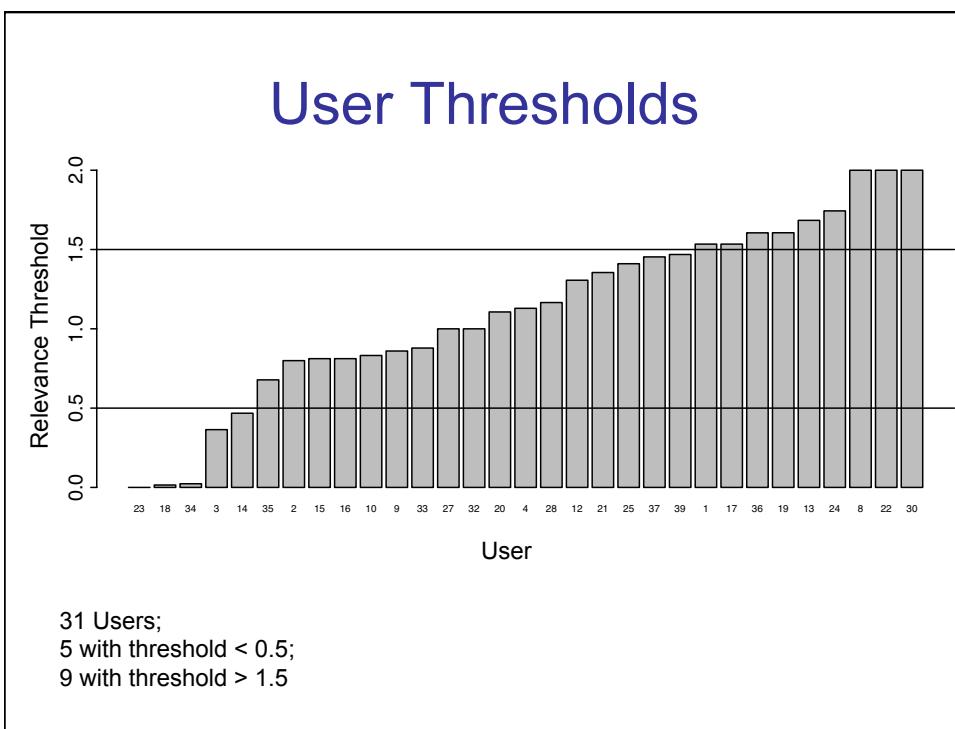
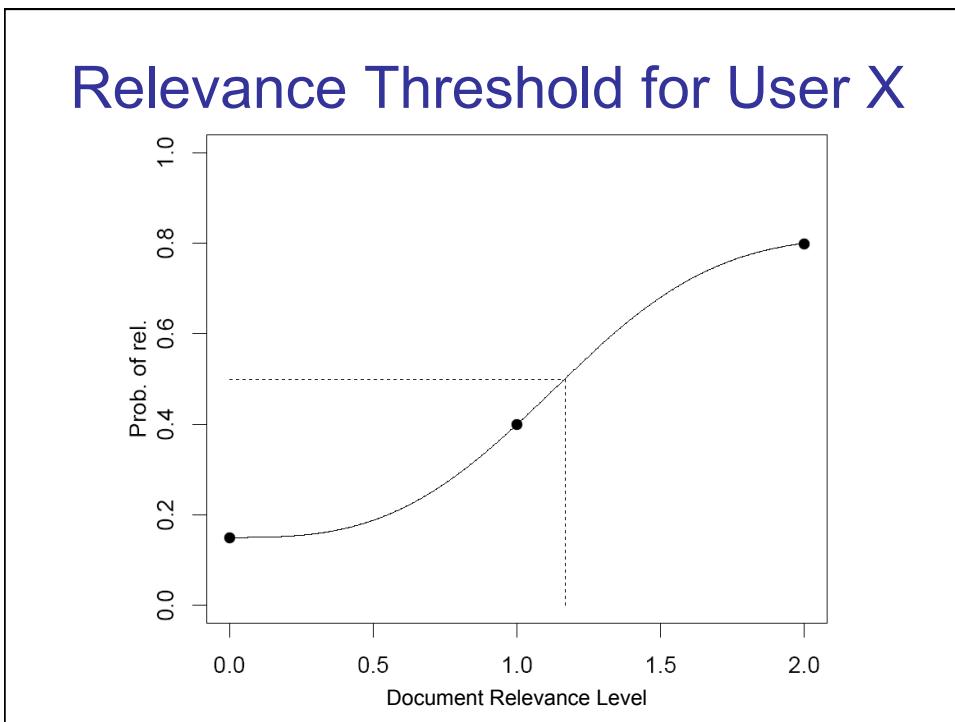


Psychophysics of Relevance

- If relevance can be measured on a scale, then relevance thresholds could be obtained for each user
- TREC judges also have (pre-defined) relevance thresholds
- Differences between these groups might explain the disparate results observed between system metrics and user studies of search performance

Judgement Process

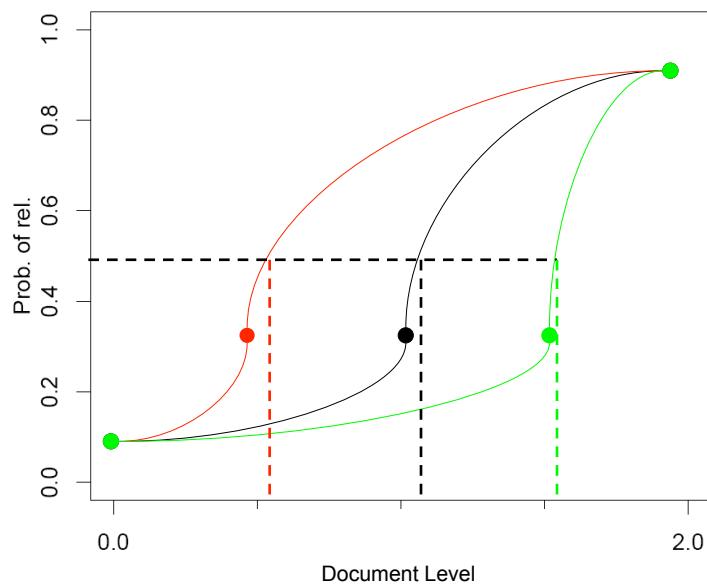
1. Show an information need to user
2. Present a series of stimuli (documents), one at a time
 - Each document has a “true” relevance level (the TREC judgement, unknown to user)
 - For each document, user makes a *binary* decision about whether the document meets the given information need or not
3. After the presentation of many documents at different relevance levels, a response proportion can be calculated



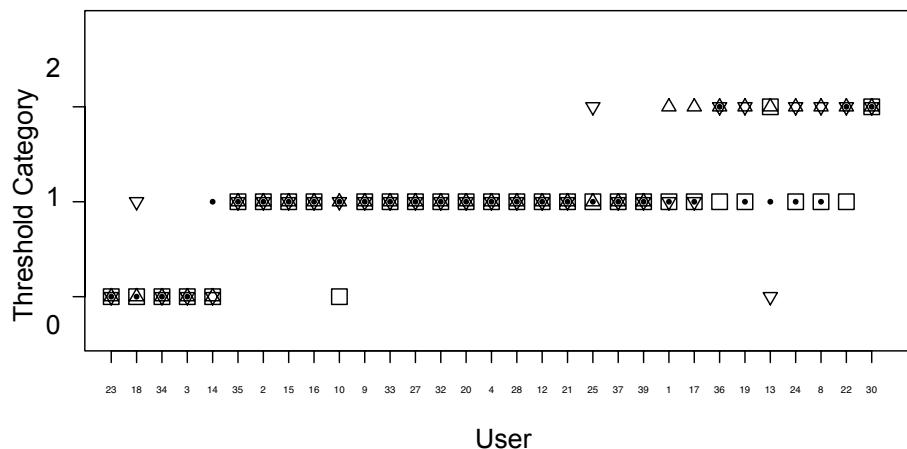
Do You Really See the Light?

- 9 users had “abnormal” response functions
- Construction of thresholds assumes that the underlying (TREC) relevance scale can be interpreted numerically (as a ratio scale)
 - But TREC judges are not instructed that a level 2 document should be “twice as relevant” as a level 1 document

Varying “Level 1”



Threshold Confusion



Thresholds: 19 Consistent, 11 dual, 1 triple

Split Agreement Approach

- The underlying relevance “scale” derived from TREC relevance judgements is not suitable for defining psychophysical response functions
- Simpler approach: define user classes based on relevance behaviour

User Classes

- **TREC-like:** follow the assumed batch relevance profile
 - Level 1 and level 2 documents are marked relevant more than 50% of the time



User Classes...

- **Generous:** have lower criteria for relevance than TREC judges, tend to like even level 0 documents
 - Level 0 documents are marked relevant **more** than 50% of the time
- **Parsimonious:** have stricter criteria than TREC judges, tend to like only level 2 documents
 - Level 1 documents are marked relevant **less** than 50% of the time

Example

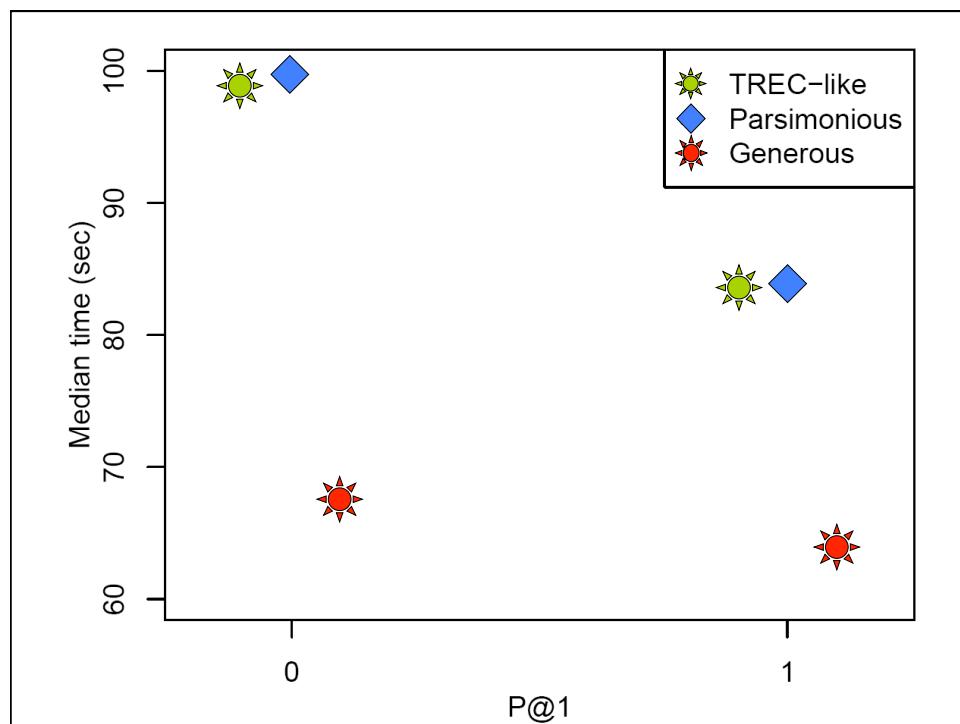
TREC Level	0	1	2
Proportion “saved”	6%	63%	94%

- A TREC-like user...



Implications for Searching

- Consider the *time* that a user needs to find a relevant document when using batch systems at different levels of P@1:
 - For **TREC-like** users, expect to see a difference in time when using a P@1=0 batch system versus a P@1=1 system
 - **Generous** users should be faster than others, no matter which batch system they use
 - **Parsimonious** users should be slower with either system



Differences in Search Times

Class	Time difference	p-value	Number
TREC-like	15.2	0.0770	8
Parsimonious	16.2	0.0029	11
Generous	3.43	0.2209	19

Conclusions

- Past work has shown problems in transferring batch results to real user tasks
- When metrics are chosen appropriately, to match the user search tasks, it appears that results *can* be comparable between approaches

Conclusions

- Differences in user performance can vary for different relevance levels
 - Level 0 and 2: significant
 - Level 1 and 2: significant
 - Level 0 and 1: weakly significant
- Marginally relevant (level 1) documents should probably be grouped with non-relevant documents when considering binary relevance

Conclusions (Relevance Profiling)

- Relevance profiles can help to explain the transferability of results between batch and user evaluation
 - Generous users don't reflect the batch assumptions

Future Work

- How to estimate relevance profiles with minimum effort
 - How many topics, documents need to be assessed?
- How to devise a *ratio* relevance scale
 - Underlying click behaviour in query logs might represent “average” users
 - Thresholds/split agreement can inform how individual users deviate from this behaviour
 - Magnitude estimation

How To Find Me

- At SILS until 1 December
- Sitting in the Interaction Design Lab
- Email: falk.scholer@rmit.edu.au