IR Evaluation

Younghoon Kim
(nongaussian@hanyang.ac.kr)

※ Slides by Chris Manning and Pandu Nayak

Situation

- Thanks to your stellar performance in IR class, you quickly rise to VP of Search at internet retail giant 11st.com. Your boss brings in her/his nephew Sergey, who claims to have built a better search engine for 11st. Do you
  - Laugh derisively and send him to rival Gmarket?
  - Counsel Sergey to go to Stanford and take IR?
  - Try a few queries on his engine and say “Not bad”?
  - … ?

What could you ask Sergey?

- How fast does it index?
  - Number of documents/hour
  - Incremental indexing – 11st adds 10K products/day
- How fast does it search?
  - Latency and CPU needs for 11st's 5 million products
- Does it recommend related products?
- This is all good, but it says nothing about the quality of Sergey's search
  - You want 11st's users to be happy with the search experience

How do you tell if users are happy?

- Search returns products relevant to users
  - How do you assess this at scale?
- Search results get clicked a lot
  - ~~Problem: misleading titles/summaries can cause users to click~~
- Users buy after using the search engine
  - Or, users spend a lot of $ after using the search engine
- Repeat visitors/buyers
  - Do users leave soon after searching?
  - Do they come back within a week/month/… ?

Happiness: elusive to measure

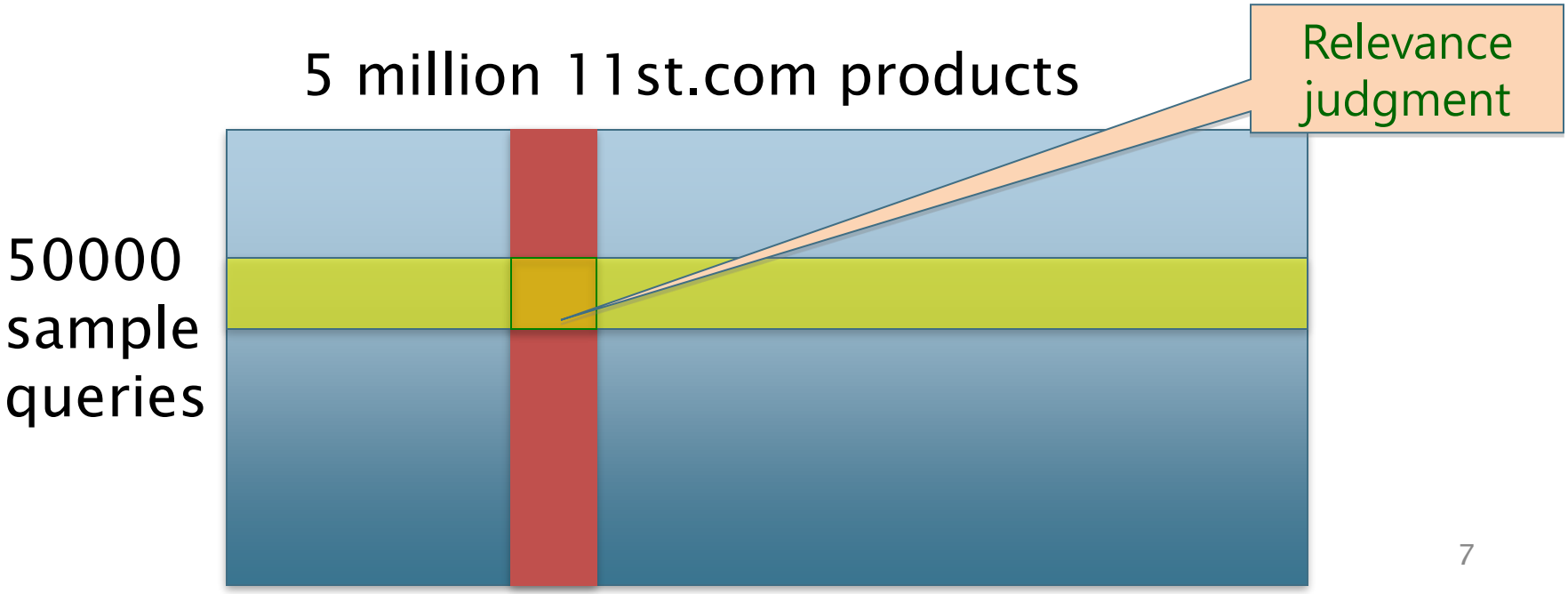
- Most common proxy: relevance of search results
  - But how do you measure relevance?
- Pioneered by Cyril Cleverdon in the Cranfield Experiments
  - http://en.wikipedia.org/wiki/Cranfield_Experiments

Measuring relevance

- Three elements:
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of either Relevant or Non-relevant for each query and each document

So you want to measure the quality of a new search algorithm

- Benchmark documents – 11st.com products
- Benchmark query suite – more on this
- Judgments of document relevance for each query (e.g., rating score)



Relevance judgments

- *Binary* (relevant vs. non-relevant) in the simplest case, more *nuanced* (0, 1, 2, 3 …) in others
- What are some issues already?
- 5 million times 50K takes us into the range of a quarter trillion judgments
  - If each judgment took a human 2.5 seconds, we 'd still need 10^{11} seconds, or nearly $300 million if you pay people $10 per hour to assess
  - 10K new products per day

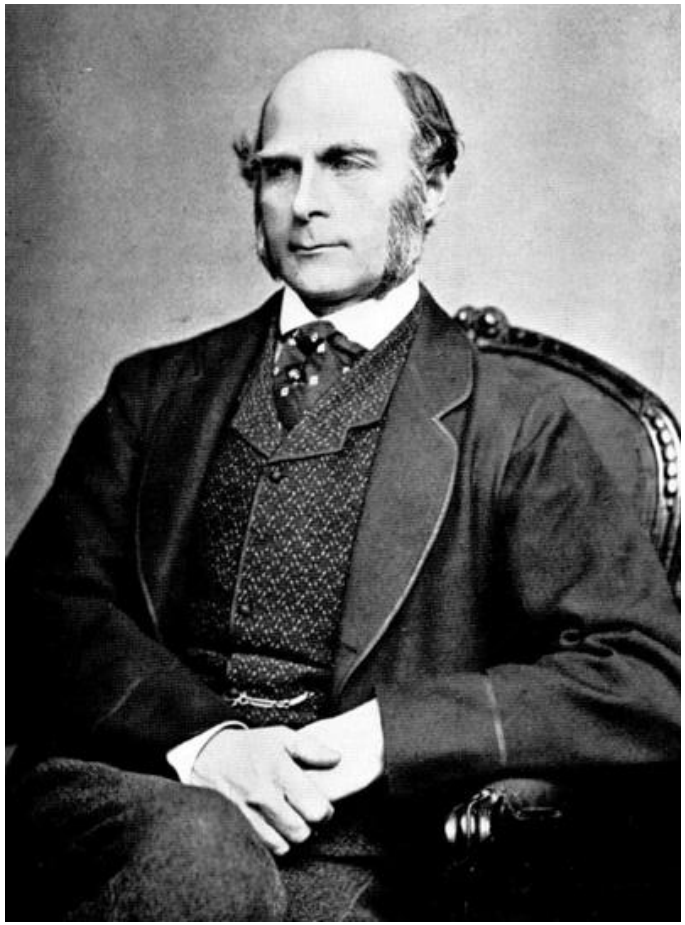
SMALL TALK: CROWDSOURCING

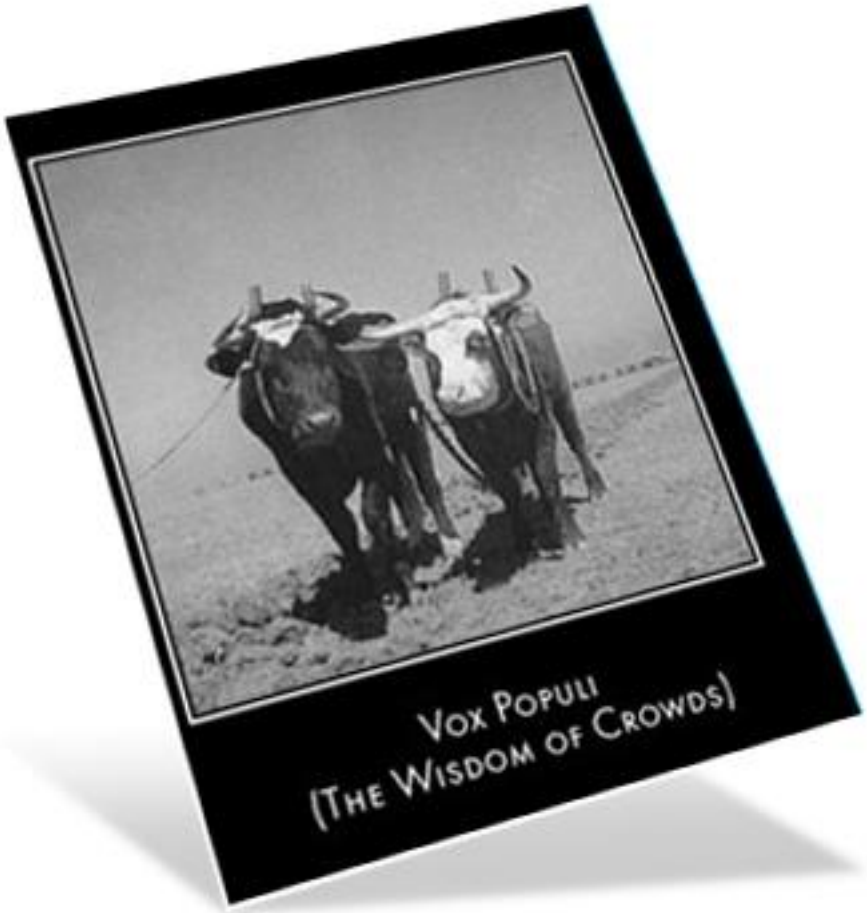
Crowd source relevance judgments?

- Present query-document pairs to *low-cost labor on online* crowd-sourcing platforms
  - Hope that this is cheaper than hiring qualified assessors
- Main takeaway – you get some signal, but the variance in the resulting judgments is very high

Traditional Crowdsourcing

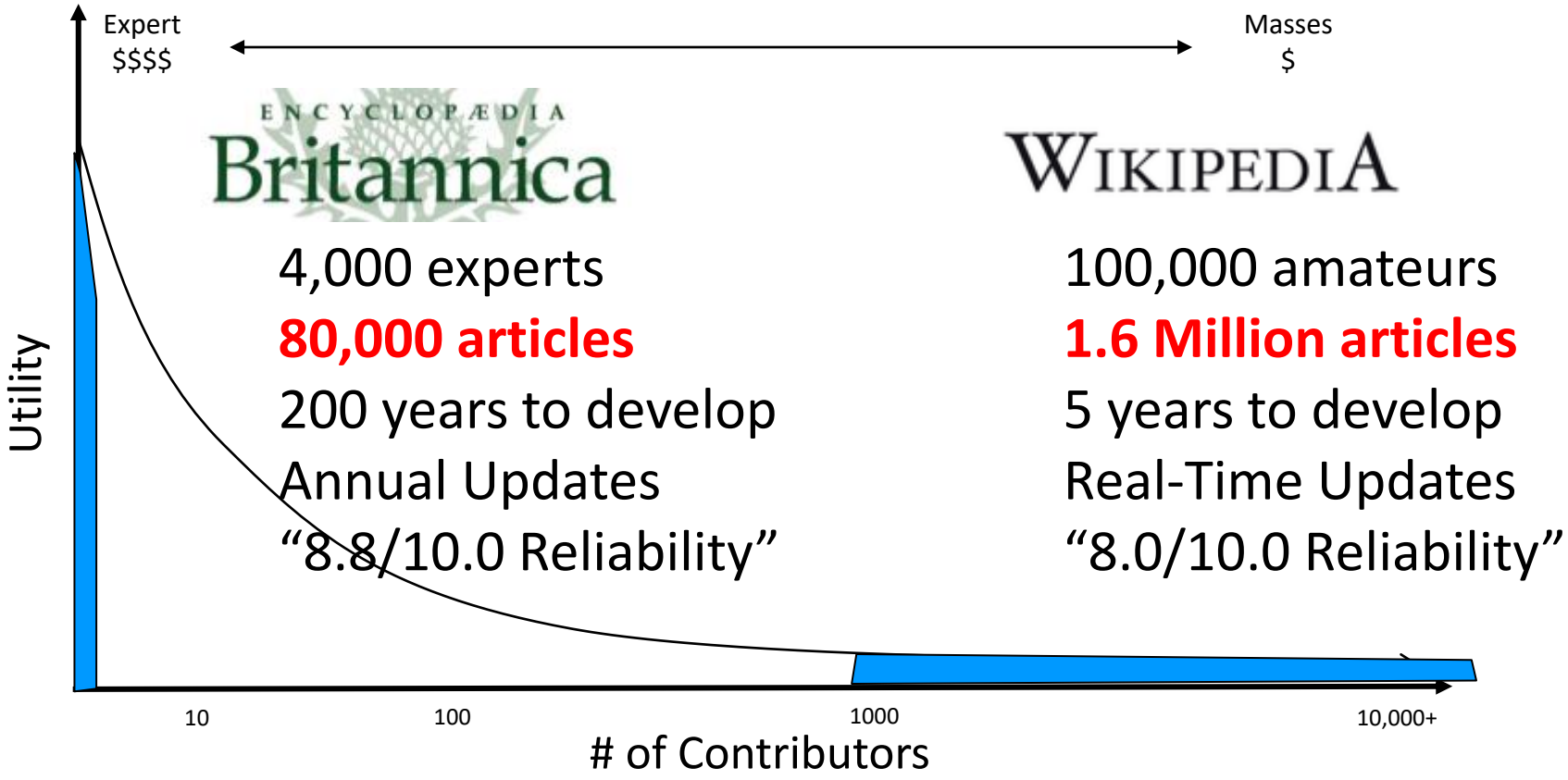
Weight-judging competition (Francis Galton, 1906):
1,197 (mean of 787 crowds) vs. 1,198 pounds (actual measurement)





[Slide by Dongwon Lee, PennState Univ

Economics & Wikinomics



What is Crowdsourcing?

Online

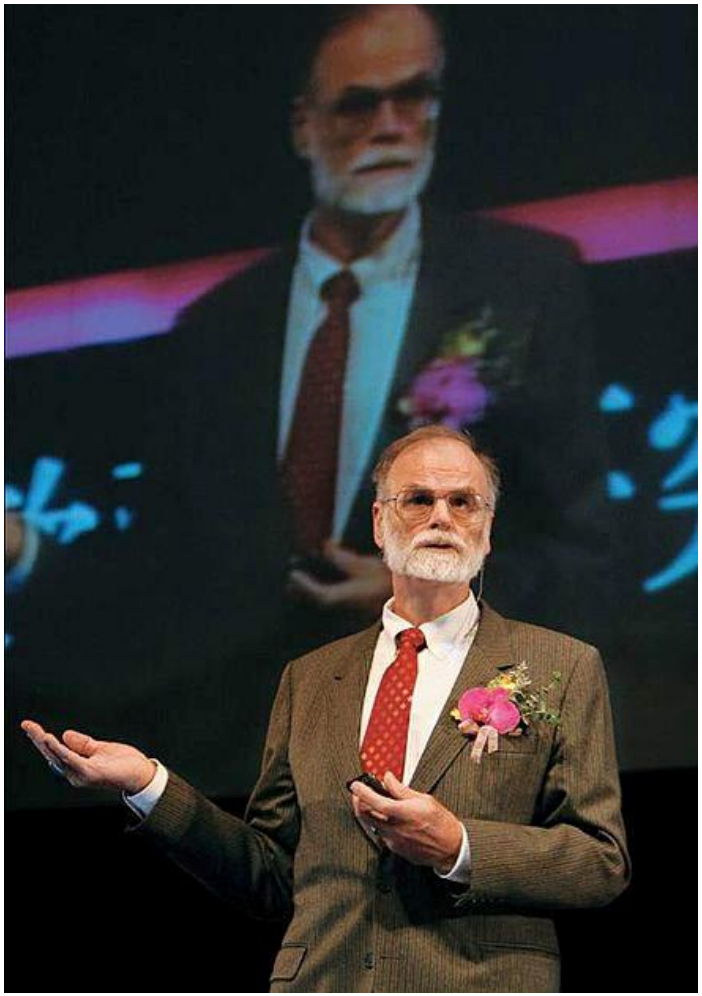
- Crowd typically form into online communities based on the Web site
- The crowd submits solutions to the site or produce its contents

Distributed problem solving and production

A problem is often divided into many micro tasks

Answers from crowd are collected and merged together to derive the final solution

Eg, Finding "Jim Gray", 2007



An American computer scientist who received the **Turing Award** in 1998 for seminal contributions to database and transaction processing research

On Sunday, January 28, 2007, during a short solo sailing trip near San Francisco, Gray and his 40-foot yacht, Tenacious, were reported **missing** by his wife

Eg, Finding "Jim Gray", 2007

COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH

Home / Magazine Archive / July 2011 (Vol. 54, No. 7) / Searching for Jim Gray: A Technical Overview / Full Text

CONTRIBUTED ARTICLES

Searching for Jim Gray: A Technical Overview

By Joseph M. Hellerstein, David L. Tennenhouse
Communications of the ACM, Vol. 54 No. 7, Pages 77-87
10.1145/1965724.1965744
Comments

VIEW AS: SHARE:

- Loosely coupled tean software polytechtur interfaces, decouplin from analysis to enab at a distance.
- The U.S. Coast Guard to aid search and rese interesting potential computer scientists.
- New open-source too could help with grou crowdsourced image volume image proces ocean drift modeling, of open-water satelli

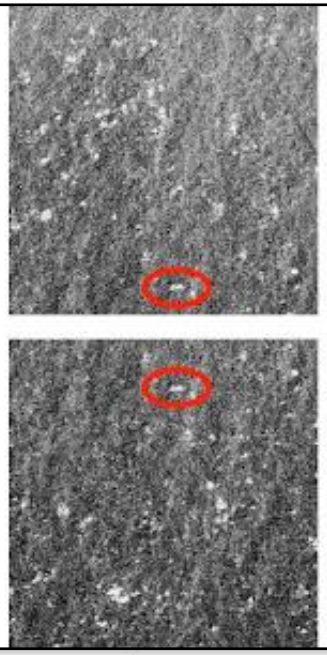
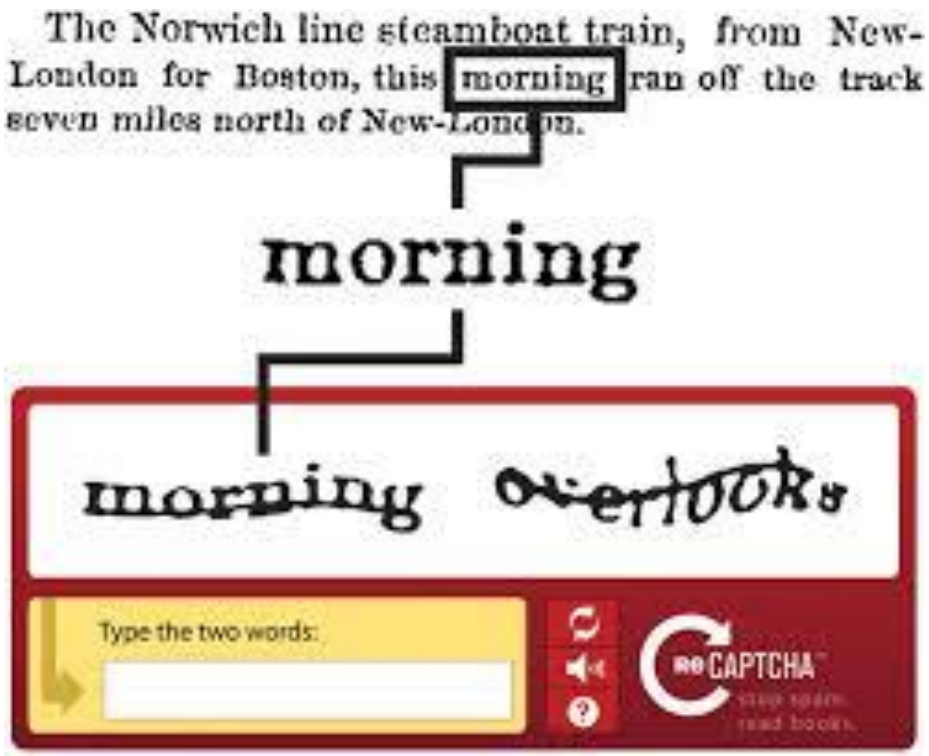


Image A @ time t_1

Image B @ time t_2

On Sunday, January 28, 2007, noted computer scientist Jim Gray

Eg, reCAPCHA



As of 2012

Captcha: 200M every day

ReCaptcha: 750M to date

Eg, reCAPCHA

OCR Transcription

The Hreckinridge' and Lane Democrats, having taken courage at the recent eastern advises, are [xxxxxxxxx] energetically for the campaign: Several prominent Democrats who at first favored DonoLea, are coming out. for the other aide, apparently under the [xxxxxxxx] of Federal [xxxxxxxx]. An address to the National Democracy of ,1ifornia, urging the party to support HaeeslipslDas, has recently been published, which manifestly bss strengthened that aide of the [xxxxxxxx]: It is signed by 65 Democrats, many of whom occupy respectab e and prominent positions in the party, 22 of them are Federal office-holders, [xxxxx] more are recipients of Federal patronage, and the others represent a mass of politicians giving the document [xxxx] [xxxxx] mTheDcuBlas Democrats are also active The Irish and German vote will mostly go with ths# branch of the party, but it is [xxxxxxxx] to [xxxxxxx] [xxxxx] [xxxx] [xx] the stronger. Thus far 17 IT newspapers have declared for DonGres, 13 for BaseS-IealDGS and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equal,- ly balanced as to give the State [xx] LlaCOLV. Same very [xxxxxx] Bell and Everett meetings have been held in different parts of the State, bat thus far that party does not exhibit much rank sad ale air en.

reCAPTCHA Transcription

The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

http://www.google.com/recaptcha/digitizing

Duolingo: Crowdsourcing new languages

- Duolingo utilizes the power of crowds to make learning a language free

GERMAN TEXT:
Falls Pakistans Geschichte ein Indikator ist, so könnte Musharrafs Entscheidung, das Kriegsrecht zu verhängen, jener sprichwörtliche Tropfen sein, der das Fass zum Überlaufen bringt.

PROFESSIONAL HUMAN TRANSLATION (20 cents per word):
If Pakistan's history is any indicator, Musharraf's decision to impose martial law may prove to be the proverbial straw that breaks the camel's back.

GOOGLE TRANSLATE:
If Pakistan's history is any indicator, it could Musharraf's decision to impose martial law, be that proverbial straw that breaks the camel's back.

DUOLINGO:
If Pakistan's history is an indicator, Musharraf's decision to impose martial law could be the straw that breaks the camel's back.

http://www.technologyreview.com/news/506656/the-cleverest-business-model-in-online-education/

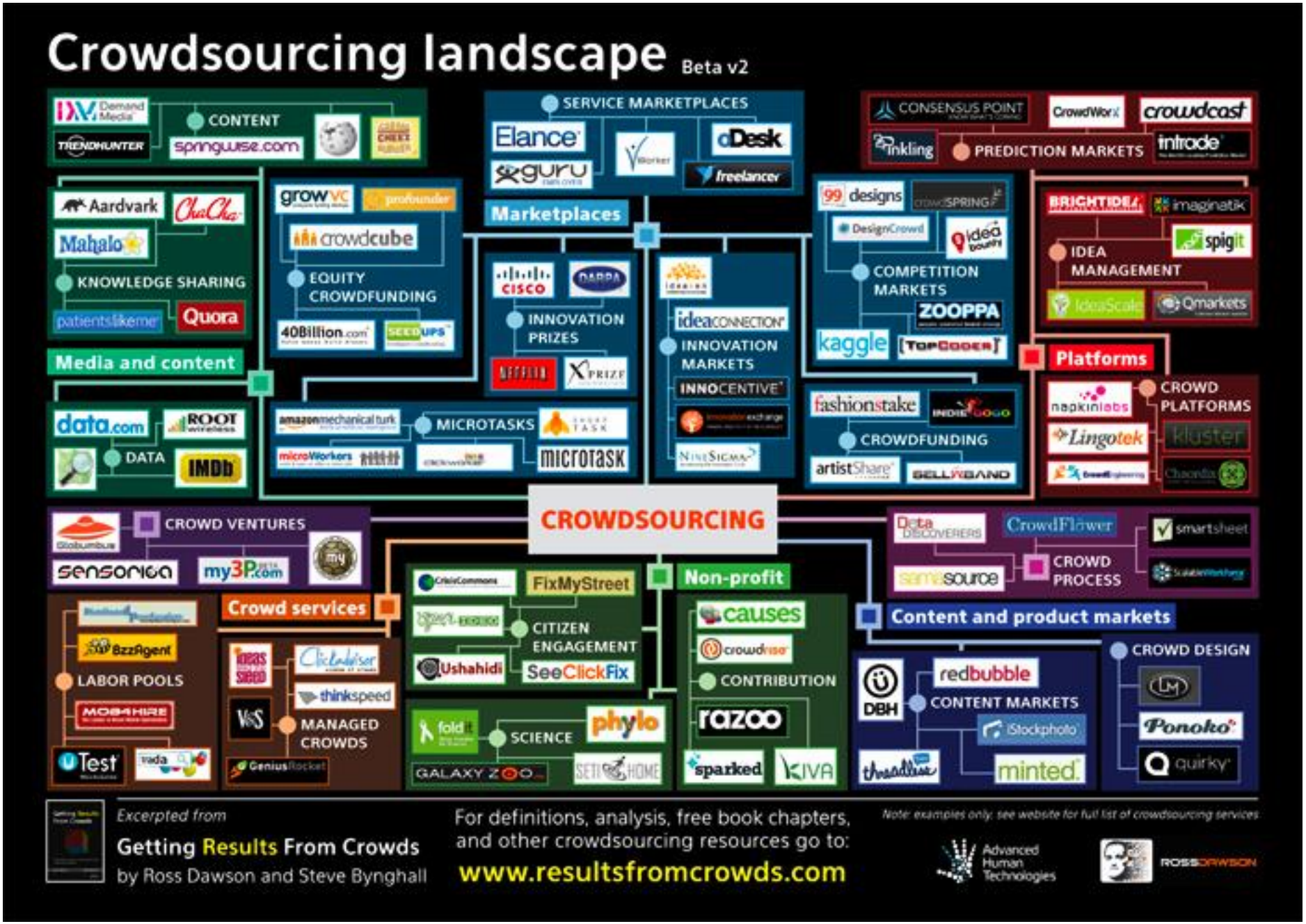
Characteristics of Crowdsourcing

Benefits of Crowdsourcing

- Problems can be explored at comparatively little cost
- Payment is by results
- The organization can tap a wider range of talent than might be present in its own organization
- Turn customers into designers
- Turn customers into marketers

Problems with Crowdsourcing

- Quality
- Intellectual property leakage
- No time constraint
- Not much control over development or ultimate product
- Ill-will with own employees
- Choosing what to crowdsource & what to keep in-house



http://www.resultsfromcrowds.com/features/crowdsourcing-landscape/

Three Participants

- Requesters
  - People submit some tasks
  - Pay rewards to workers
- Marketplaces
  - Provide crowds with tasks
- Crowds
  - Workers perform tasks

[Slide by Dongwon Lee, PennState Univ

AMT: mturk.com

Workers
- Register w. credit account (only US workers can register as of 2013)
- Bid to do tasks for earning money

Requesters
- First deposit money to account
- Post tasks
- Gather results
- Pay to workers if results are satisfactory

Your Account | HITs | Qualificatio

Introduction | Dashboar | Status | Account S

Mechanical Turk is a marketplace f

give businesses and developers access to an on-demand, scalable workforce.

from thousands of tasks and work wh

200,645 HITs available. View the

Workers

Requesters

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task → **Work** → **Earn money**

Find HITs Now

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:

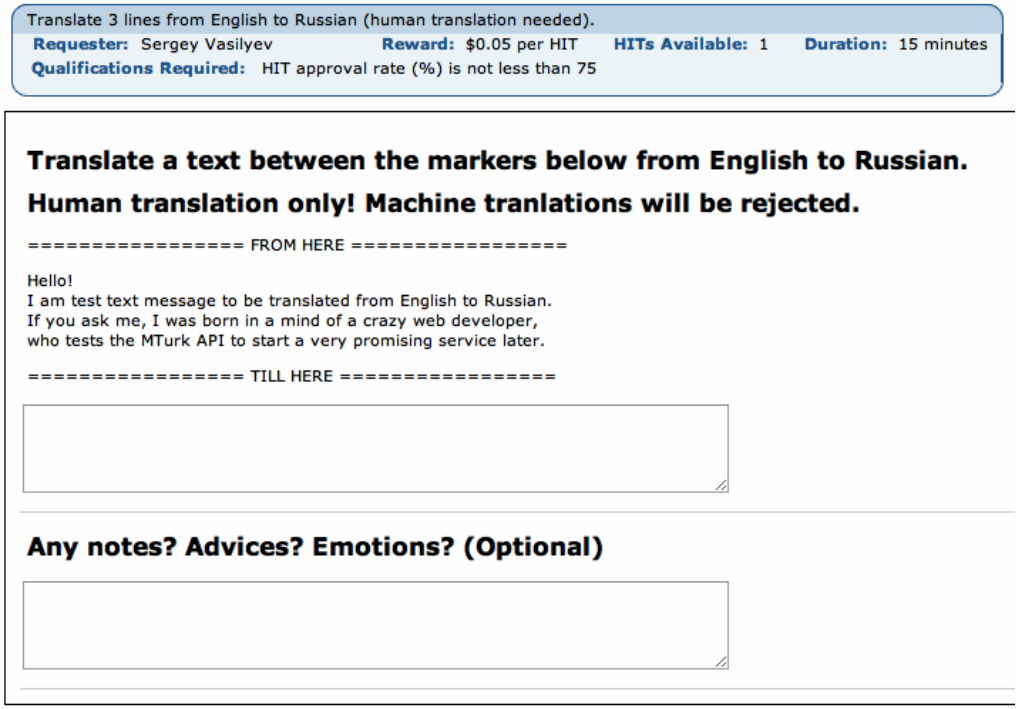
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account → **Load your tasks** → **Get results**

Get Started

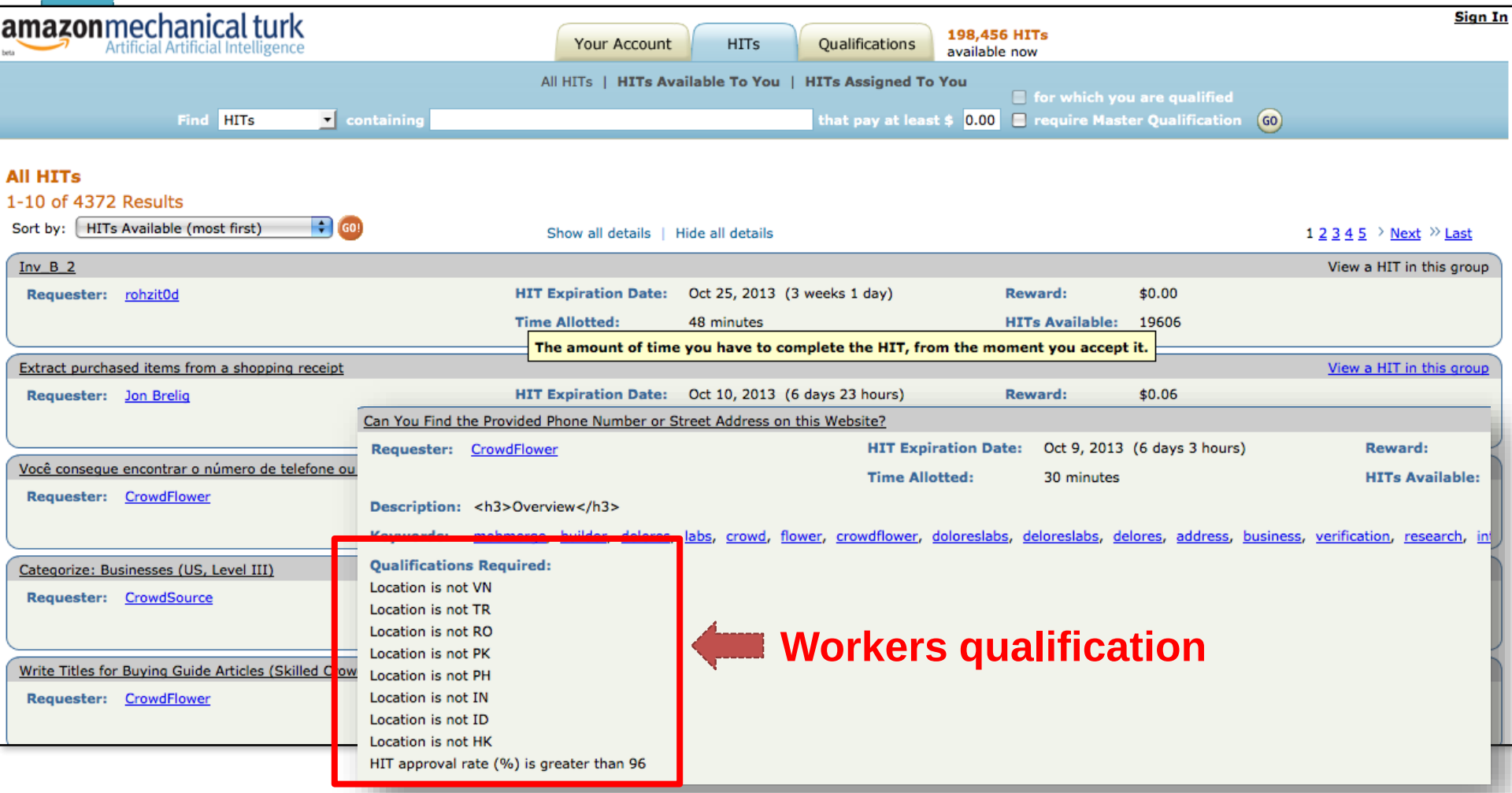
AMT: HIT

- Tasks
  - Called **HIT** (Human Intelligence Task)
  - Micro-task
- Eg
  - Data cleaning
  - Tagging/labeling
  - Sentiment analysis
  - Categorization
  - Surveying
  - Photo moderation
  - Transcription



Translation task

AMT: HIT List


amazonmechanical turk
Artificial Artificial Intelligence
Sign In
Your Account
HITs
Qualifications
198,456 HITs available now
All HITs | HITs Available To You | HITs Assigned To You
Find HITs containing that pay at least $ 0.00
for which you are qualified
require Master Qualification
GO
All HITs
1-10 of 4372 Results
Sort by: HITs Available (most first)
Show all details | Hide all details
1 2 3 4 5 Next Last
Inv_B_2
View a HIT in this group
Requester: rohzit0d
HIT Expiration Date: Oct 25, 2013 (3 weeks 1 day)
Reward: $0.00
Time Allotted: 48 minutes
HITs Available: 19606
The amount of time you have to complete the HIT, from the moment you accept it.
Extract purchased items from a shopping receipt
View a HIT in this group
Requester: Jon Brelig
HIT Expiration Date: Oct 10, 2013 (6 days 23 hours)
Reward: $0.06
Can You Find the Provided Phone Number or Street Address on this Website?
Requester: CrowdFlower
HIT Expiration Date: Oct 9, 2013 (6 days 3 hours)
Reward:
Time Allotted: 30 minutes
HITs Available:
Description: <h3>Overview</h3>
Qualifications Required:
Location is not VN
Location is not TR
Location is not RO
Location is not PK
Location is not PH
Location is not IN
Location is not ID
Location is not HK
HIT approval rate (%) is greater than 96
Você consegue encontrar o número de telefone ou
Requester: CrowdFlower
Categorize: Businesses (US, Level III)
Requester: CrowdSource
Write Titles for Buying Guide Articles (Skilled C
Requester: CrowdFlower
Workers qualification

AMT: HIT Example

Timer: 00:00:00 of 60 minutes

Want to work on this HIT? Want to see other HITs?

Accept HIT Skip HIT

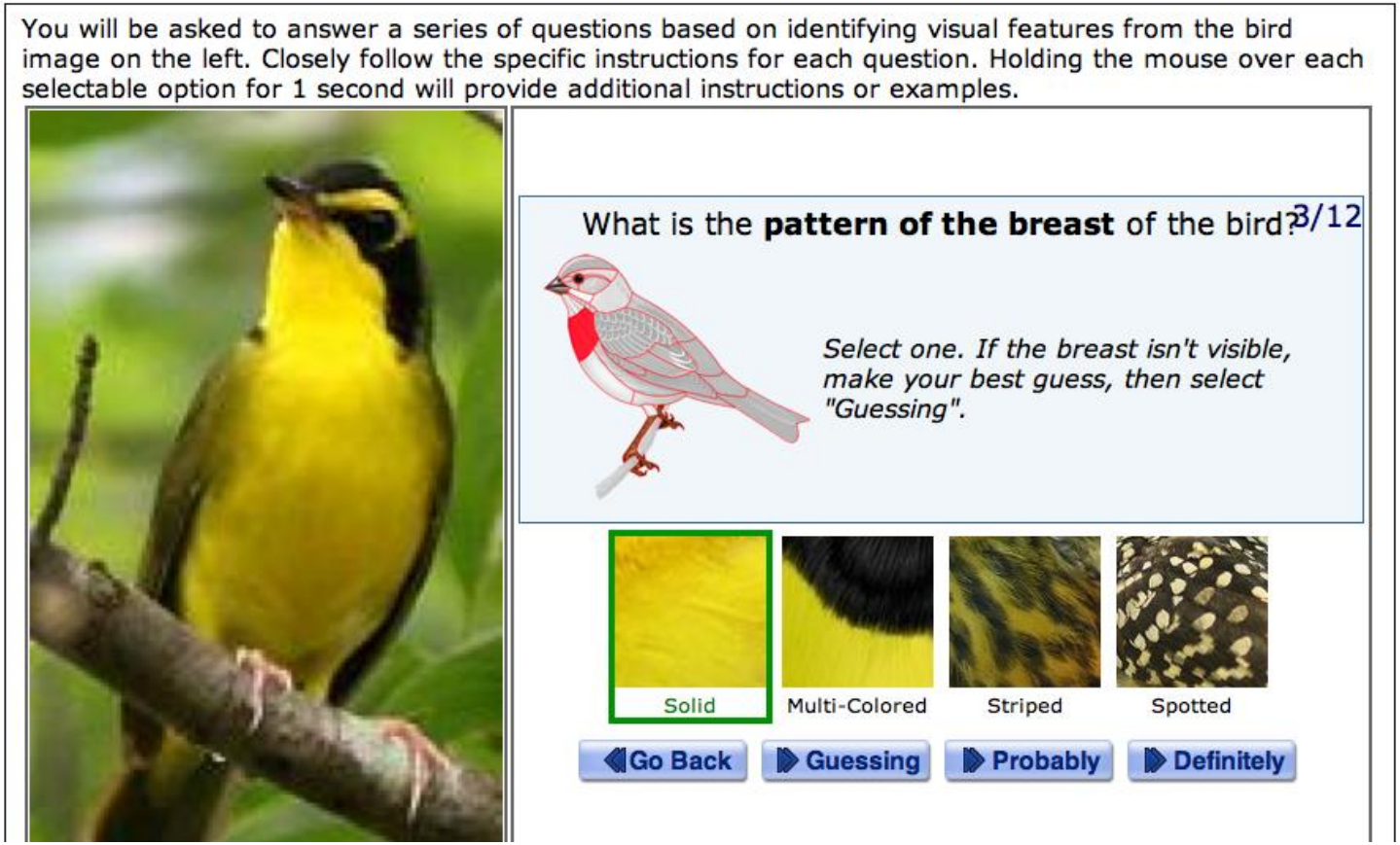
Total Earned: Unavailable
Total HITs Submitted: 0

Store name, date, time, total, location on this receipt

Requester: Vishwanath Kumar **Reward:** $0.03 per HIT **HITs Available:** 71477 **Duration:** 60 minutes

Qualifications Required: Total approved HITs is greater than 1000

AMT: HIT Example



RETURN TO OUR TOPIC: QUALITY MEASURES

What else?

- Still need test queries (how to select them?)
  - Must be *germane* to docs available
  - Must be *representative* of actual user needs
  - Random query terms from the documents generally not a good idea
  - Sample from query logs if available
- Classically (non-Web)
  - Low query rates – not enough query logs
  - Experts hand-craft "user needs"

Some public test Collections

TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Now we have the basics of a benchmark

- Let's review some evaluation measures
  - *Precision*
  - *Recall*
  - DCG
  - …

Evaluating an IR system

- Note: **user need** is translated into a **query**
- Relevance is assessed relative to the **user need**, *not* the **query**
  - E.g., Information need: *My swimming pool bottom is becoming black and needs to be cleaned.*
  - Query: ***pool cleaner***
- Assess whether the doc addresses the underlying need, not whether it has these words

Unranked retrieval evaluation: Precision and Recall

- **Binary assessments**

Precision: fraction of retrieved docs that are relevant = P(relevant|retrieved)

Recall: fraction of relevant docs that are retrieved = P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

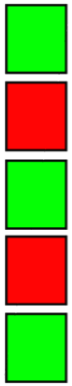
- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Rank-Based Measures

- Binary relevance
  - Precision@K (P@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)
- Multiple levels of relevance
  - Normalized Discounted Cumulative Gain (NDCG)

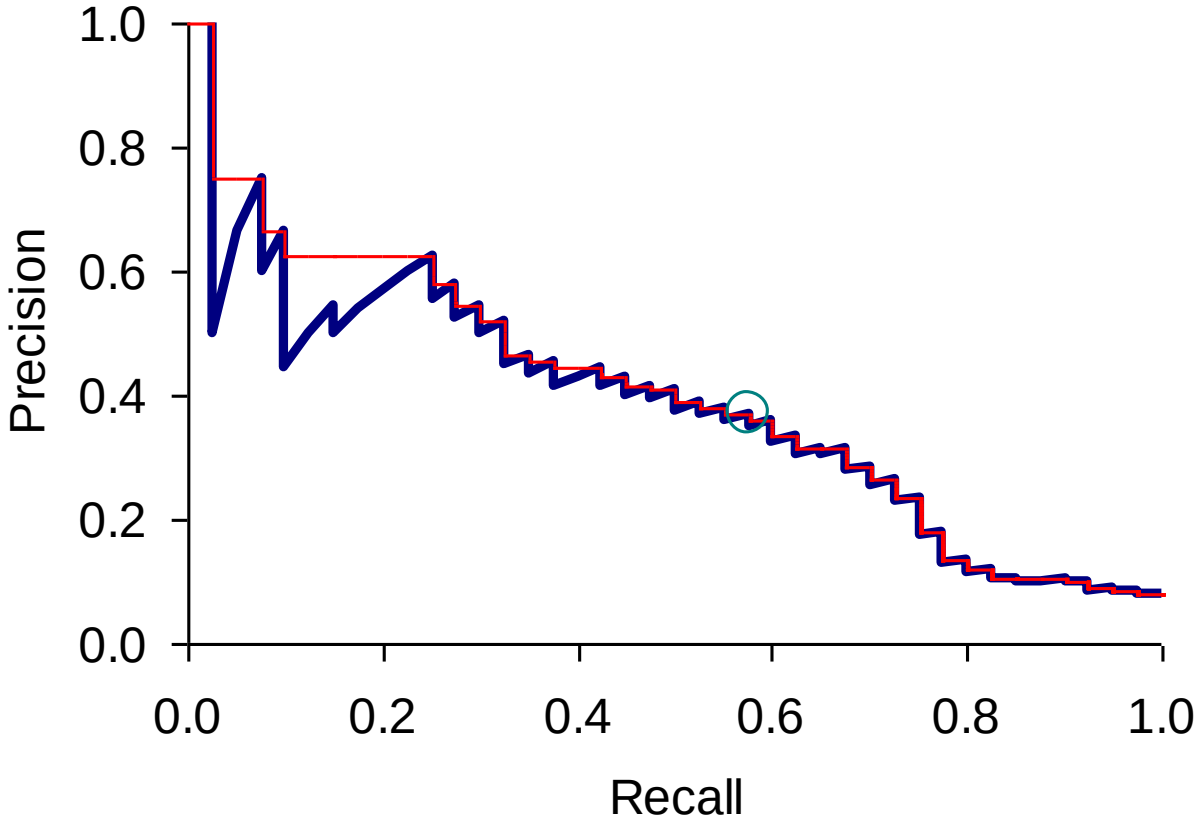
Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K
- Ex:
  - Prec@3 of 2/3
  - Prec@4 of 2/4
  - Prec@5 of 3/5

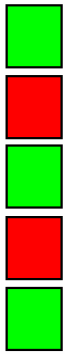


- In similar fashion we have Recall@K

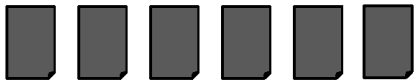
A precision-recall curve



Mean Average Precision

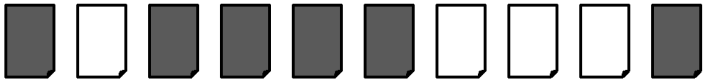
- Consider rank position of each ***relevant*** doc
  - $K_1, K_2, \ldots K_R$
- Compute Precision@K for each $K_1, K_2, \ldots K_R$
- Average precision = average of P@K
  - Ex:  has AvgPrec of $\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$
- MAP is Average Precision *across multiple queries/rankings*

Average Precision



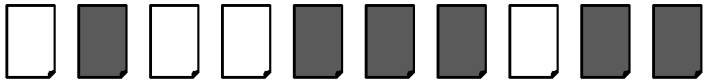
= the relevant documents

Ranking #1



Recall	?		?	?						
Precision	?		?	?						

Ranking #2



Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

MAP

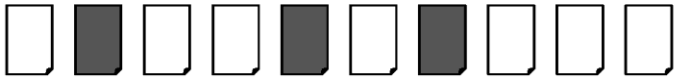
= relevant documents for query 1

Ranking #1

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

= relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

Mean Average Precision

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero
- MAP is macro-averaging: each query counts equally
- Now perhaps most commonly used measure in research papers
- Good for web search?
- MAP assumes user is interested in finding many relevant documents for each query
- MAP requires many relevance judgments in text collection

Mean Reciprocal Rank

- Consider rank position, K, of the first relevant doc
  - Could be – only clicked doc
- Reciprocal Rank score = $\frac{1}{K}$
- MRR is the mean RR across multiple queries

Query 1	R	N	R	N	N	N	N	N	R	R
Query 2	N	R	N	N	R	R	R	N	N	N

MRR = ?

BEYOND BINARY RELEVANCE

YAHOO!

Web Images Video Local Shopping More

Toyota safety

Search Options

Search Pad

SearchScan - On

108,000,000 results for
Toyota safety:

Show All

Toyota

Motor Trend

CarsDirect

Shopping Sites

Also try: toyota safety ratings, toyota safety recall, More...

Sponsored Results

Toyota Recall
Toyota Takes Care of its Customers. Read the FAQs at Toyota.com.
www.Toyota.com/Recall

Toyota Safety
& Latest Prices. Free Info. Toyota Research, Reviews.
www.Toyota.Edmunds.com

fair

TOYOTA | Car Safety Innovation and Technology
Toyota home page for car safety and car technology Prius model.
www.safetytoyota.com - Cached

Toyota home page for car safety and car technology ...
We are presenting Toyota's safety technologies for cars. We clearly explain about car safety and car technology using movies and more.
www.safetytoyota.com/en-gb - Cached

Toyota Safety Ratings - Toyota Safety Features - Motor Trend ...
MotorTrend offers Toyota safety ratings, comprehensive auto safety reports, and more. View a all of the standard Toyota safety features. ...
motortrend.com/new_cars/07/toyota/safety_ratings/index.html - 149k - Cached

Toyota Motor Europe Corporate Site Safety
Our approach. Toyota believes that all stakeholders in the road safety equation share a responsibility to reduce the frequency of road accidents. ...
www.toyota.eu/Safety - Cached

[PDF] pdf European Safety Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a Toyota and/or Lexus brand motor vehicle equipped with the safety systems ...
www.toyota.no/Images/Safety_Brochure_tcm308-344461.pdf

Toyota - Star Safety System
Star Safety System ... Toyota Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. Toyota Newsroom. sign up for info ...
www.toyota.com/vehicles/demos/star-safety.html - 58k - Cached

Toyota Prius Safety Ratings - CarsDirect
Get overall safety ratings and NHTSA crash test results for the Toyota Prius at CarsDirect.

Sponsored Results

Safety for a Toyota
Research Safety Ratings and Reviews For New Car at Kelley Blue Book.
www.kbb.com

Toyota Safety
Find Toyota Safety dealers, new cars, prices, and photos.
www.NewCars.org

fair

Toyota Safety
Toyota safety Discount Prices Save Money Shopping Online Today.
www.smarter.com

Good

Saftey Toyoto
Explore 5,000+ Pro Sports Choices. Save On Saftey Toyoto.
BaseballGear.Shopzilla.com

See your message here...

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks

- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant documents
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of usefulness, or *gain,* from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is 1/log *(rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Summarize a Ranking: DCG

- What if relevance judgments are in a scale of [0,r]? r>2
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)
  - CG = $r_1+r_2+\ldots r_n$
- Discounted Cumulative Gain (DCG) at rank n
  - DCG = $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$
    - We may use any base for the logarithm

The relevance value of the doc ranked at the 1st place

Discounted Cumulative Gain

- *DCG* is the total gain accumulated at a particular rank *p*:

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

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - used by some web search companies
  - emphasis on retrieving highly relevant documents

DCG E

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

- 10 ranked documents judged on 0-3 relevance scale:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0

- Discounted gain:

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

- DCG:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
- Normalization useful for contrasting queries with varying numbers of relevant results
- NDCG is now quite popular in evaluating Web search

NDCG - Example

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

Exercise

- Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1	R	N	R	N	N	N	N	N	R	R
System 2	N	R	N	N	R	R	R	N	N	N

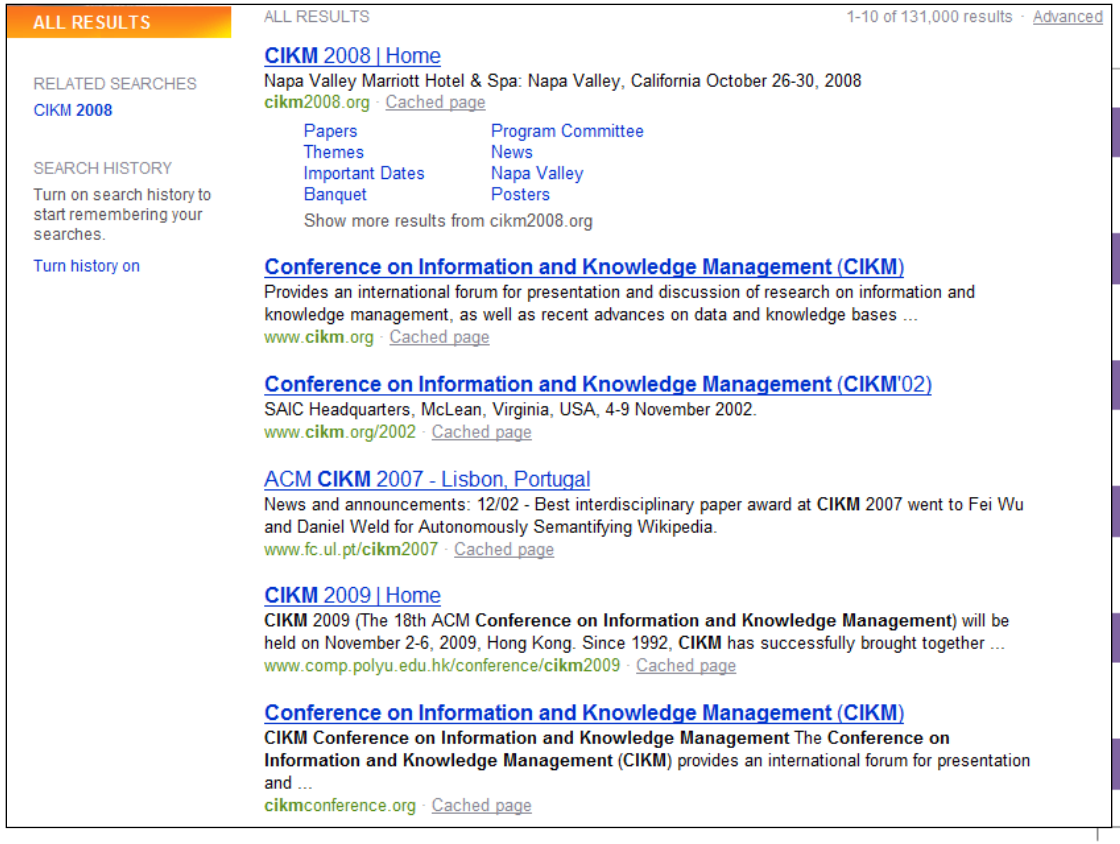
  - What is the MAP of each system? Which has a higher MAP?
  - Compute NDCG of each system
  - Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?

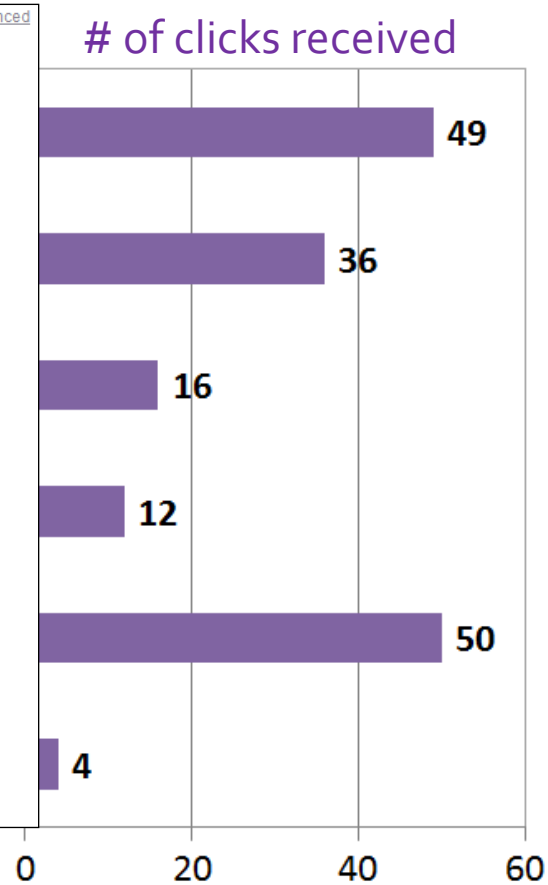
Human judgments are

- Expensive
- Inconsistent
  - Between raters
  - Over time
- Decay in value as documents/query mix evolves
- Not always representative of "real users"
  - Rating vis-à-vis query, vs underlying need
- So – **what alternatives** do we have?

USING USER CLICKS

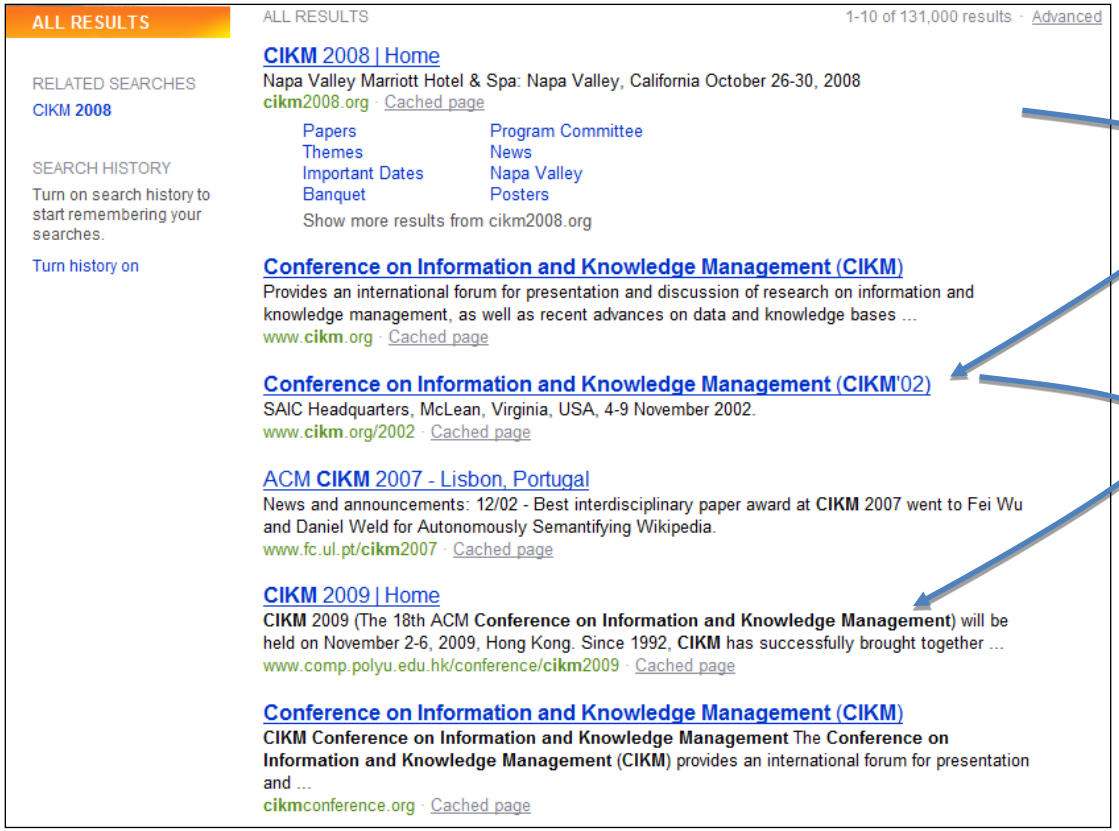
What do clicks tell us?





Strong position bias, so absolute click rates unreliable

Relative vs absolute ratings



User's click sequence

Hard to conclude Result1 > Result3

Probably can conclude Result3 > Result2

Comparing two rankings via clicks (Joachims 2002)

두 Ranking을 섞어 놓고 어느 결과를 사람들이 클릭하는지 관찰!

Query: [support vector machines]

Ranking A	Ranking B
Kernel machines	Kernel machines
SVM-light	SVMs
Lucent SVM demo	Intro to SVMs
Royal Holl. SVM	Archives of SVM
SVM software	SVM-light
SVM tutorial	SVM software

Interleave the two rankings

This interleaving starts with B

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

…

Remove duplicate results

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

…

Count user clicks

Ranking A: 3
Ranking B: 1

	Clicks
Kernel machines	A, B
Kernel machines	
SVMs	
SVM-light	A
Intro to SVMs	
Lucent SVM demo	A
Archives of SVM	
Royal Holl. SVM	
SVM-light	

...

Interleaved ranking

- Present interleaved ranking to users
  - Start randomly with ranking A or ranking B to evens out presentation bias
- Count clicks on results from A versus results from B
  Better ranking will (on average) get more clicks

A/B testing at web search engines

- Purpose: **Test a single innovation**
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to an experiment to evaluate an innovation
  - Interleaved experiment
  - Full page experiment

Comparing two rankings to a baseline pairwise preference

- Given a set of pairwise preferences (baseline) P
- We want to measure two rankings A and B
- Define a proximity measure between A and P
  - And likewise, between B and P
- Want to declare the ranking with better proximity to be the winner
- Proximity measure should reward agreements with P and penalize disagreements

Kendall tau distance

- Let X be the number of agreements between a ranking (say *A*) and *P*
- Let Y be the number of disagreements
- Then the Kendall tau distance between *A* and *P* is

$$(X-Y)/(X+Y)$$

- Say P = {(1,2), (1,3), (1,4), (2,3), (2,4), (3,4))} and A=(1,3,2,4)
- Then X=5, Y=1 …
- (What are the minimum and maximum possible values of the Kendall tau distance?)

Example

Suppose we rank a group of five people by height and by weight:

Person	A	B	C	D	E
Rank by Height	1	2	3	4	5
Rank by Weight	3	4	1	2	5

Assume that people answers that they prefer to:
(A > B), (B > D), (B > E), (C > D), (A > C), (A > E)

Do people like a tall one ? Or heavy one?