Adapted Excerpts from Crowdsourcing 101: Putting the WSDM of of Crowds to Work for You

Omar Alonso
Microsoft

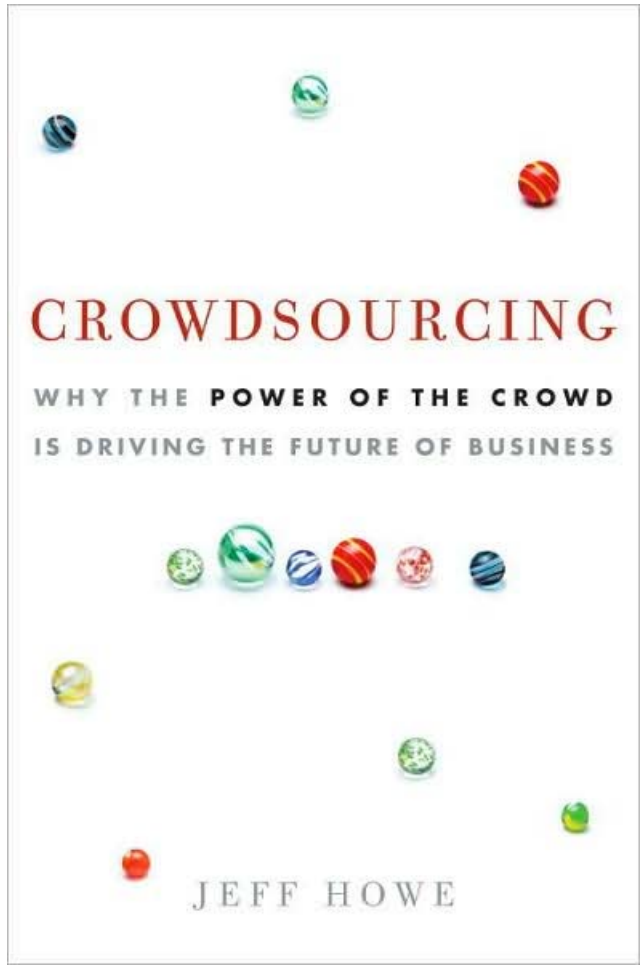
Matthew Lease
University of Texas at Austin

9 February 2011

wsdm 2011 Hong Kong

Crowdsourcing

- Take a job traditionally performed by a designated agent (usually an employee)
  - Outsource it to an undefined, generally large group of people via an open call
- New application of many open source principles

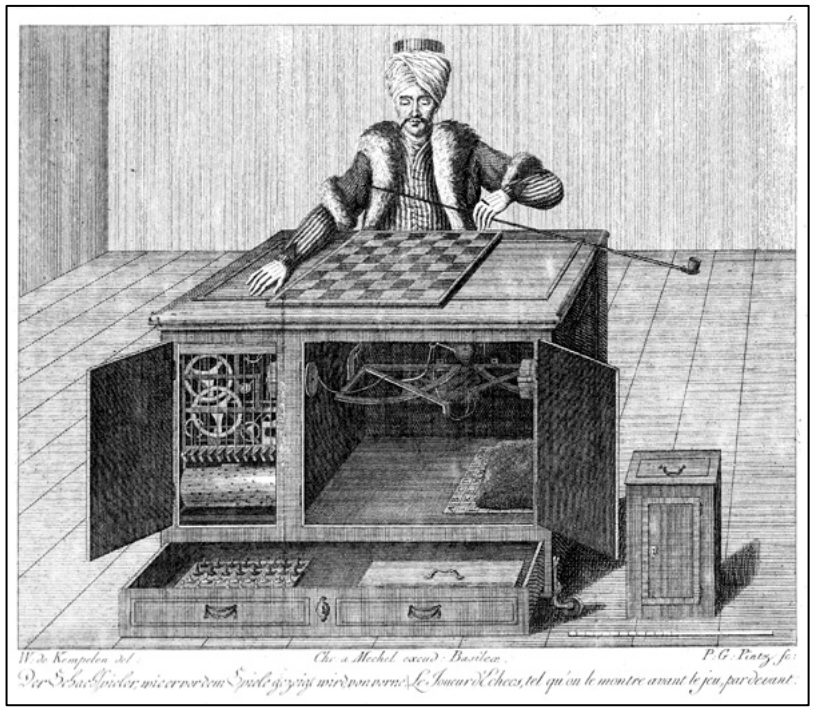


Crowdsourcing

- Outsource micro-tasks
- Success stories
  - Wikipedia
  - Apache
- Power law
- Attention
- Incentives
- Diversity

AMT

- Amazon Mechanical Turk (AMT, `www.mturk.com`)
- Crowdsourcing platform
- On-demand workforce
- “Artificial artificial intelligence”: get humans to do hard part
- Named after “The Turk”, a fake chess playing machine
- Constructed by Wolfgang von Kempelen in 18th C.



Amazon Mechanical Turk

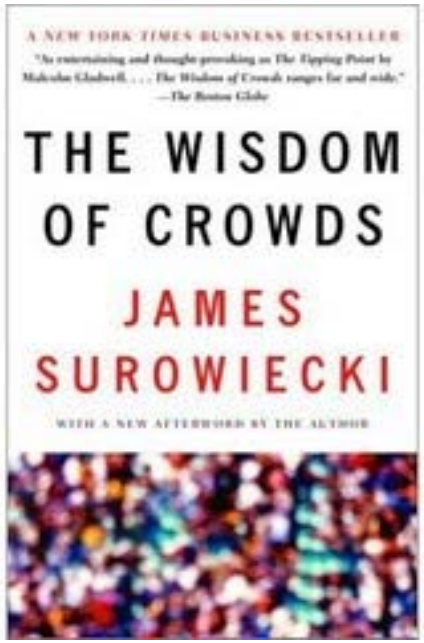
From Wikipedia, the free encyclopedia

- The **Amazon Mechanical Turk** (**MTurk**) is a crowdsourcing Internet marketplace that enables computer programmers (known as *Requesters*) to co-ordinate the use of human intelligence to perform tasks which computers are unable to do.
- The Requesters are able to pose tasks known as **HITs** (Human Intelligence Tasks), such as choosing the best among several photographs of a store-front, writing product descriptions, or identifying performers on music CDs.
- ***Workers*** (called *Providers* in Mechanical Turk's Terms of Service) can then browse among existing tasks and complete them for a monetary payment set by the Requester.
- **To place HITs**, the requesting programs use an open Application Programming Interface, or the more limited Mturk Requester site [1]
- Requesters can ask that Workers fulfill Qualifications before engaging a task, and they can set up a test in order to verify the Qualification.
- They can also accept or reject the result sent by the Worker, which reflects on the Worker's **reputation**.
- **Payments** for completing tasks can be redeemed on Amazon.com via gift certificate or be later transferred to a Worker's U.S. bank account. Requesters, which are typically corporations, pay 10 percent over the price of successfully completed HITs to Amazon.

Wisdom of Crowds

Requires

- Diversity
- Independence
- Decentralization
- Aggregation



Input: large, diverse sample

(to increase likelihood of overall pool quality)

Output: consensus or selection (aggregation)

vs. Ensemble Learning

- Use multiple models to obtain better performance than from any constituent model
- Often combines many *weak learners* to produce a *strong learner*
- Compensate for poor individual learning by performing a lot of extra computation
- Tend to work better when significant diversity
- Using less diverse strong learners has worked better than "*dumbing-down*" models to increase diversity (Gashler et al.'08)

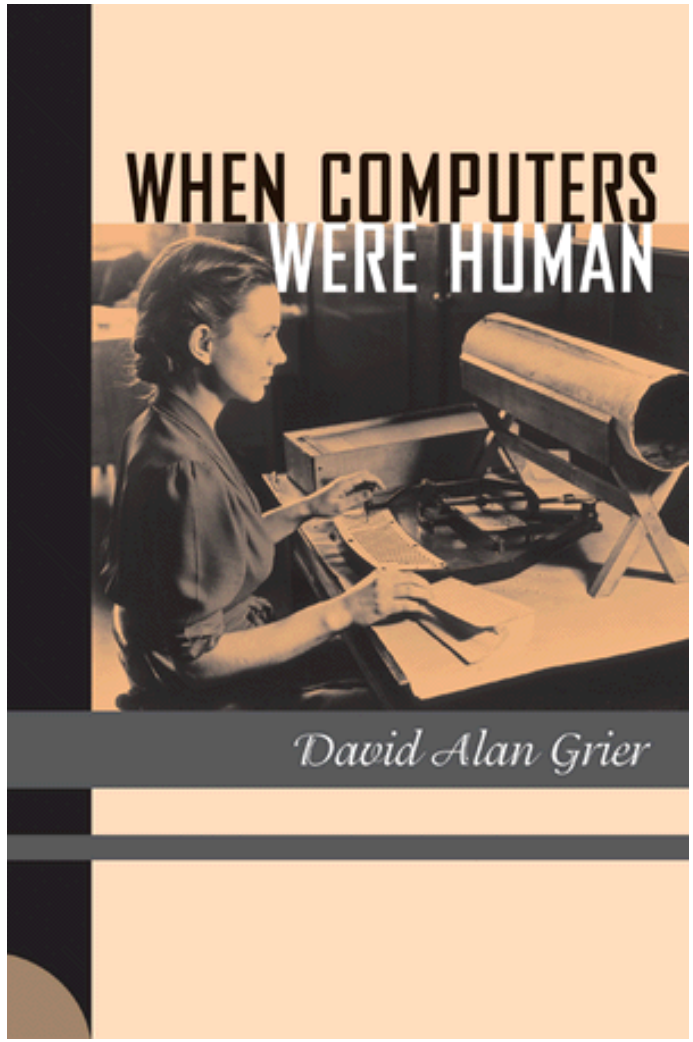
Human Computation

- Use humans as processors in a distributed system
  - Perform tasks computers aren't good at
  - Automated system can make "external calls" to the "HPU"
    - Reverse: identify tasks computers can't do (Captcha)
- Examples
  - Games with a purpose (e.g. ESP game)
  - ReCaptcha

L. von Ahn. "Games with a purpose". Computer, 39 (6), 92–94, 2006.

Human Computation

- Not a new idea
- Computers before computers

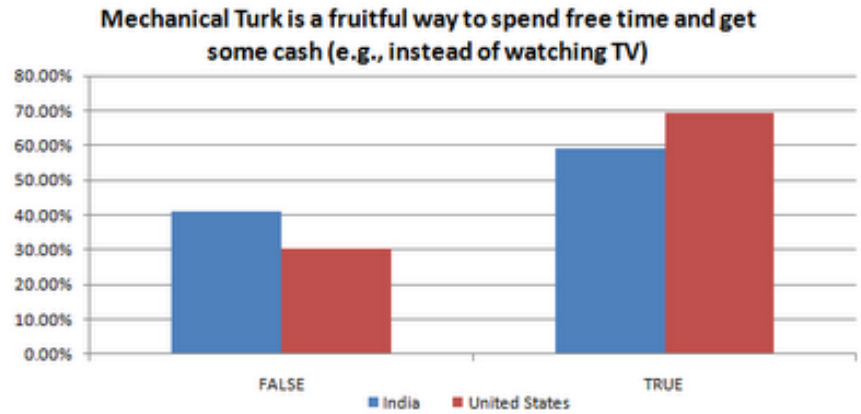


A New World of Application Design

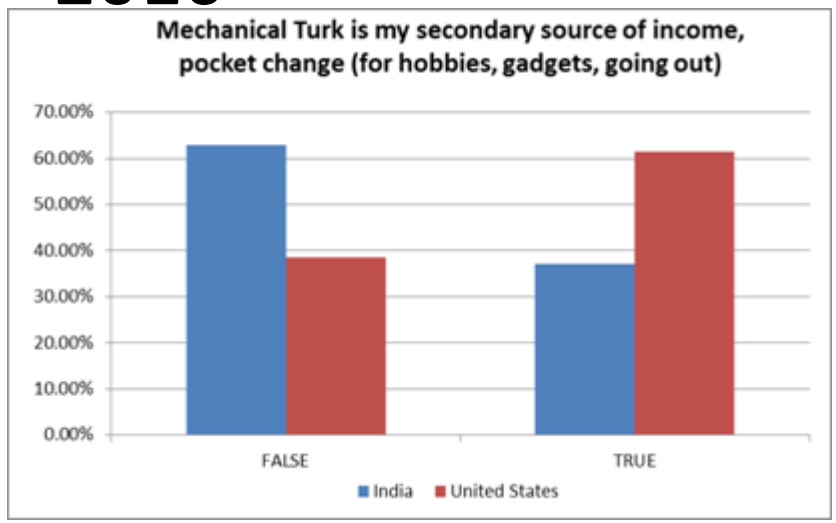
New man+machine hybrid applications blend automation with crowd interaction to achieve new capabilities exceeding components

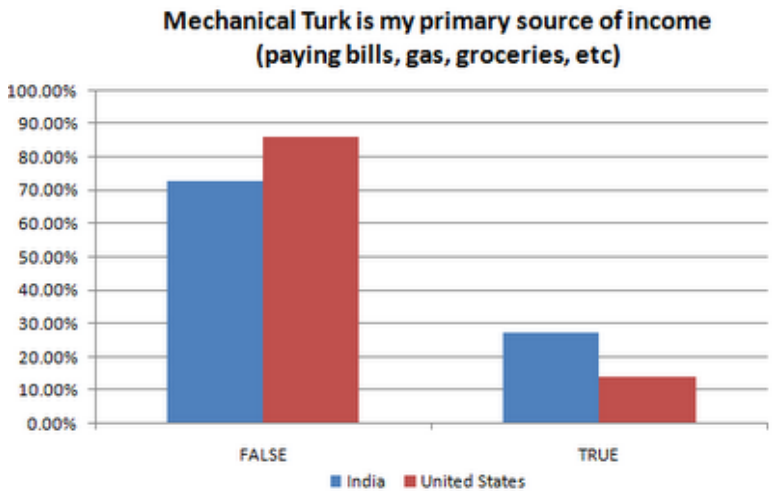
- CrowdSearch (T. Yan et al., MobiSys 2010)
- Soylent: A Word Processor with a Crowd Inside. M. Bernstein et al. UIST 2010.
- Translation by Iteractive Collaboration between Monolingual Users, B. Bederson et al. *GI 2010*

Pay ($$$)



P. Ipeirotis March 2010





Examples

- A closer look at previous work with crowdsourcing
- Includes experiments using AMT
- Subset of current research
  - Check the bibliography section for more references
- Wide range of topics
  - NLP, IR, Machine Translation, etc.

NLP

- AMT to collect annotations
- Five tasks: affect recognition, word similarity, textual entailment, event temporal ordering
- High agreement between workers and gold standard
- Bias correction for non-experts

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. "Cheap and Fast But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". EMNLP-2008.

Machine Translation

- Manual evaluation on translation quality is slow and expensive
- High agreement between non-experts and experts
- $0.10 to translate a sentence

C. Callison-Burch. “Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk”, EMNLP 2009.

B. Bederson et al. Translation by Iteractive Collaboration between Monolingual Users, *GI 2010*

Data quality

- Data quality via repeated labeling
- Repeated labeling can improve label quality and model quality
- When labels are noisy, repeated labeling can preferable to a single labeling
- Cost issues with labeling

V. Sheng, F. Provost, P. Ipeirotis. “Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers” KDD 2008.

Quality control on relevance assessments

- INEX 2008 Book track
- Home grown system (no AMT)
- Propose a game for collecting assessments
- CRA Method

G. Kazai, N. Milic-Frayling, and J. Costello. “Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments”, SIGIR 2009.

Page Hunt

- Learning a mapping from web pages to queries
- Human computation game to elicit data
- Home grown system (no AMT)
- More info: pagehunt.msrlivelabs.com

H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. “Improving Search Engines Using Human Computation Games”, CIKM 2009.

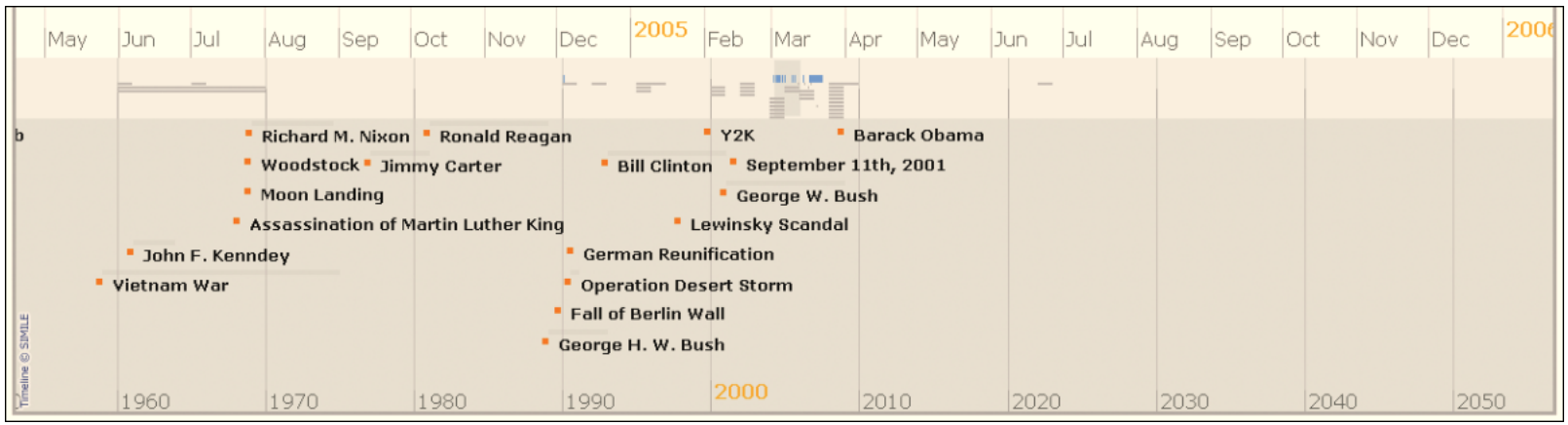
Snippets

- Study on summary lengths
- Determine preferred result length
- Asked workers to categorize web queries
- Asked workers to evaluate the quality of snippets
- Payment between $0.01 and $0.05 per HIT

M. Kaisser, M. Hearst, and L. Lowe. “Improving Search Results Quality by Customizing Summary Lengths”, ACL/HLT, 2008.

Timeline annotation

- Workers annotate timeline on politics, sports, culture
- Given a timex (1970s, 1982, etc.) suggest something
- Given an event (Vietnam, World cup, etc.) suggest a timex



K. Berberich, S. Bedathur, O. Alonso, G. Weikum "A Language Modeling Approach for Temporal Information Needs". ECIR 2010

Twitter

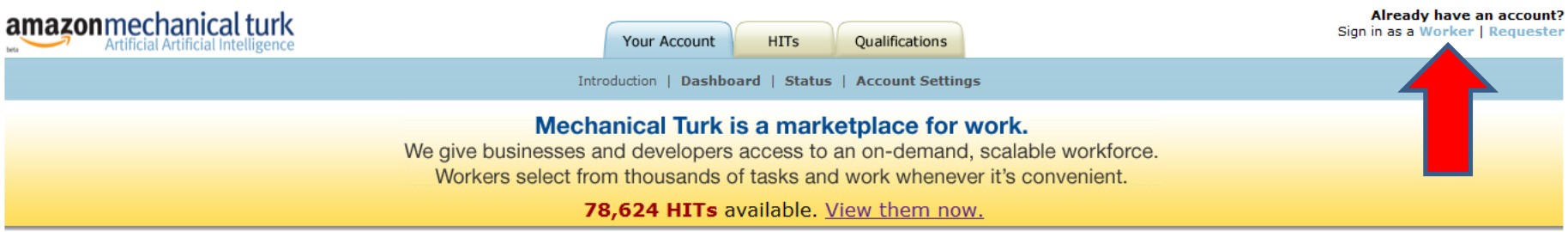
- Detecting uninteresting content text streams
  - Alonso et al. SIGIR 2010 CSE Workshop.
- Is this tweet interesting to the author and friends only?
- Workers classify tweets
- 5 tweets per HIT, 5 workers, $0.02
- 57% is categorically not interesting

AMT – How it works

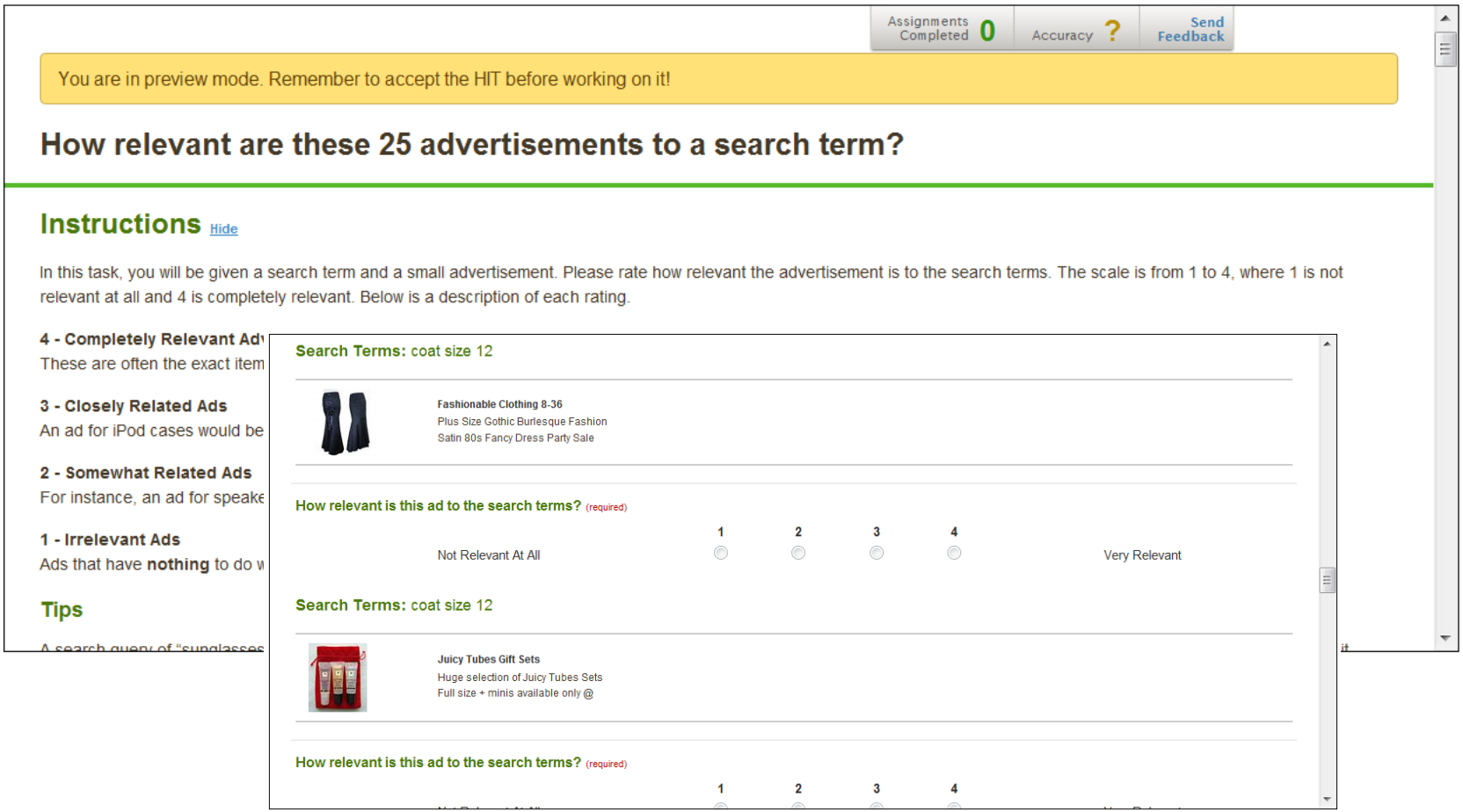
- Requesters create “Human Intelligence Tasks” (HITs) via web services API or dashboard
- Workers (sometimes called “Turkers”) log in, choose HITs, perform them
- Requesters assess results, pay per HIT satisfactorily completed
- Currently >200,000 workers from 100 countries; millions of HITs completed

The Worker

- Sign up with your Amazon account
- Tabs
  - Account: work approved/rejected
  - HIT: browse and search for work
  - Qualifications: browse and search for qualifications



Example – Relevance and ads



Example – Spelling correction

Evaluate a Spelling Correction for a Product Search Query

Instructions

Imagine that a user is searching for products at an online shopping website. When the user searches for a term, the site suggests a spelling correction, such as "Did you mean: XYZ?" Evaluate whether this spelling correction is **GOOD** or **BAD**. If you aren't sure if the suggestion gives the proper spelling or are not familiar with the search terms, select **I DON'T KNOW**.

When evaluating corrections, ignore capitalization. All search terms and corrections are shown in lower case. A correction can be good even if a space is used instead of a hyphen. For example, "blu ray" and "blu-ray" are both good spelling corrections for "blue ray", even though the trademarked term is "Blu-ray".

Sample search results are provided for context. However, you should base your response on the accuracy of the spelling correction, not the relevance of the results.

Note: We pay bonuses for high-quality responses! You will earn a bonus if your answer is consistent with the majority of respondents. However, if you consistently disagree with the majority, you will be blocked from participating in our future experiments. (An answer is considered to be the majority response when it's selected by two-thirds or more of the workers who complete the HIT.)

Instructions

Task

Please evaluate the following spelling correction, using the provided results for context:

User's search query: ***enemax***

Photographic Print of Coloured X-ray of cancer of the colon from Science Photo Library (*kitchen*)
productType: HOME_FURNITURE_AND_DECOR
productGroupID: gl_kitchen
Manufacturer: Science Photo Library
superSaver: false
numberReviews: 0
averageRating: 0.0

Reports on Publications Issued and Registered in the Several Provinces of British India (*Paperback*)
productType: ABIS_BOOK
productGroupID: gl_book
Author: Home Department, Government of India
superSaver: true
numberReviews: 0
averageRating: 0.0
fastTrack: true
fastTrackEndDate:
$escapeUtils.unescapeHtml($highlighter.highlight($misspelling_diff_raw,$engine.get('attributes').get($attrkey)))
fastTrackGuaranteedDeliveryDate:
$escapeUtils.unescapeHtml($highlighter.highlight($misspelling_diff_raw,$engine.get('attributes').get($attrkey)))
listPrice: 18.99 GBP

Suggested correction: ***enema***

Home **enema** kit: (Personal Care)
productType: HEALTH_PERSONAL_CARE
productGroupID: gl_drugstore
Manufacturer: Specialist Supplements Ltd.
superSaver: false
numberReviews: 1
averageRating: 4.0

Enema Kit - 2 litre capacity for home and travel (Misc.)
productType: BEAUTY
productGroupID: gl_beauty
Manufacturer: Manifest Health Limited
superSaver: false
numberReviews: 0
averageRating: 0.0

Is the correction of ***enemax*** to ***enema*** GOOD or BAD?

- ○ GOOD. Yes, the suggested spelling correction corrects a misspelling.
- ○ BAD. No, the suggested spelling correction is incorrect or unnecessary.
- ○ I DON'T KNOW. Not sure if the suggested spelling correction gives the proper spelling, or not familiar with the search terms.

Example - Multilingual

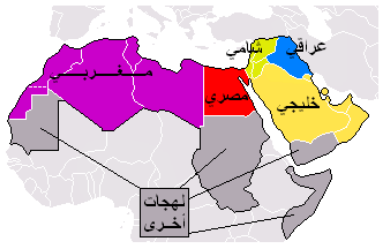
Help Classify Arabic into Dialects!

This task is for Arabic speakers who understand the different local Arabic dialects (اللهجات العامّية، أو الدّارجة), and can distinguish them from *Fusha* Arabic (الفصحى).

Below, you will see several Arabic sentences. For each sentence:

1. Tell us how much dialect (عامّية) is in the sentence, and then
2. Tell us which Arabic dialect the writer intends.

This following map explains the dialects:



PLEASE READ the following. You MUST understand the classifications, otherwise your work might be rejected!!

- **Levantine** (شامي) does NOT mean "Syrian" only. It includes **Syrian**, but ALSO: **Jordanian** is Levantine, **Palestinian** is Levantine, and **Lebanese** is Levantine. That's why all these countries are **green** in the map.
- **Maghrebi** (مغربي) does NOT mean "Moroccan" only. It includes **Moroccan**, but ALSO: **Algerian** is Maghrebi, **Tunisian** is Maghrebi, and **Libyan** is Maghrebi. That's why all these countries are **purple** in the map.
- The word "dialect" (لهجة) does NOT mean "spelling mistake" (خطأ إملائي). If the writer was trying to write in 100% فصحى, classify it as **No dialect**, even if it has some spelling mistakes.

Informed Consent Form

Purpose of research study: We are collecting human annotations to improve automatic translation of Arabic into other languages. These annotations might be class labels, judgments of output quality, or actual translations.

Benefits: Although it will not directly benefit you, this study may benefit society by improving how computers process human languages. This could lead to better translation software, improved web searching, or new user interfaces for computers and mobile devices.

Risks: There are no risks for participating in this study.

Voluntary participation: You may stop participating at any time without penalty by clicking on the "Return HIT" button, or closing your browser window.

We may end your participation if you do not have adequate knowledge of the language, or you are not following the instructions, or your answers significantly deviate from known translations.

Confidentiality: The only identifying information kept about you will be a WorkerID serial number and your IP address. This information may be disclosed to other researchers.

Questions/concerns: You may e-mail questions to the principal investigator, Chris Callison-Burch. If you feel you have been treated unfairly you may contact the Johns Hopkins University Institutional Review Board.

Clicking on the "Accept HIT" button indicates that you understand the information in this consent form. You have not waived any legal rights you otherwise would have as a participant in a research study.

Who are the workers?

A. Baio, November 2008

http://waxy.org/2008/11/the_faces_of_mechanical_turk

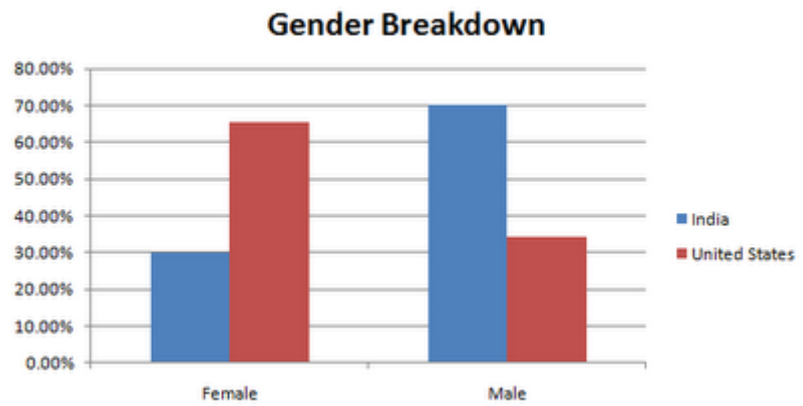
P. Ipeitorotis. March 2010

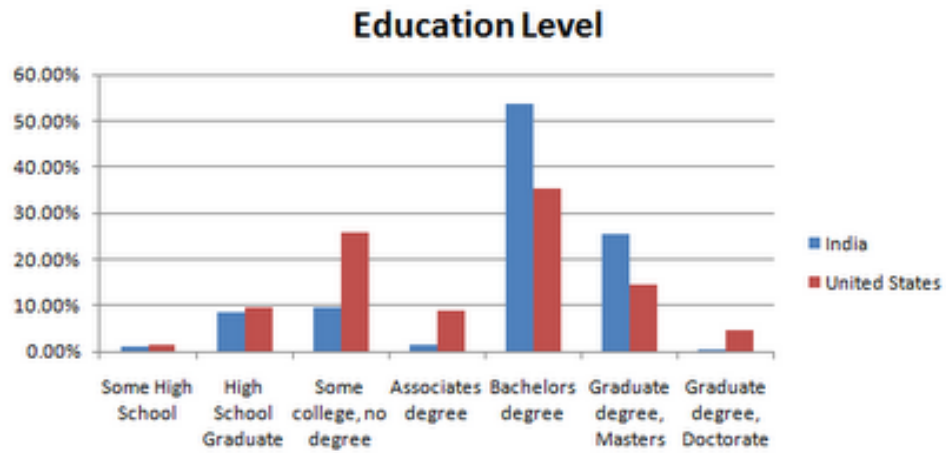
http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html

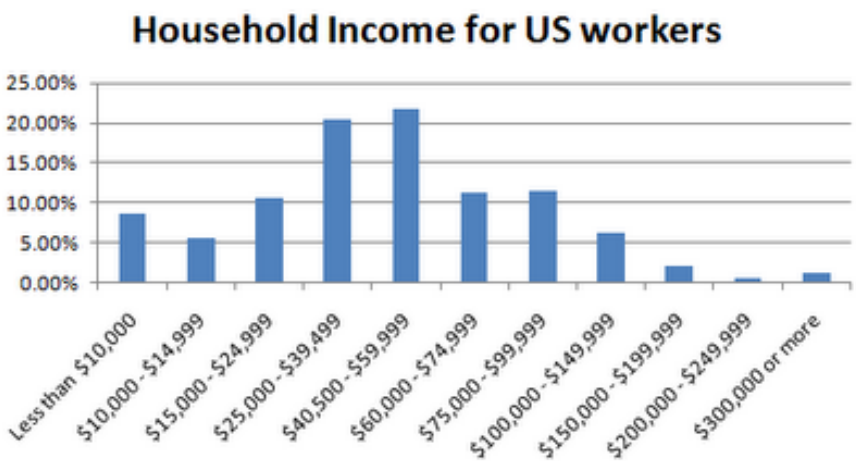
Who are the workers?

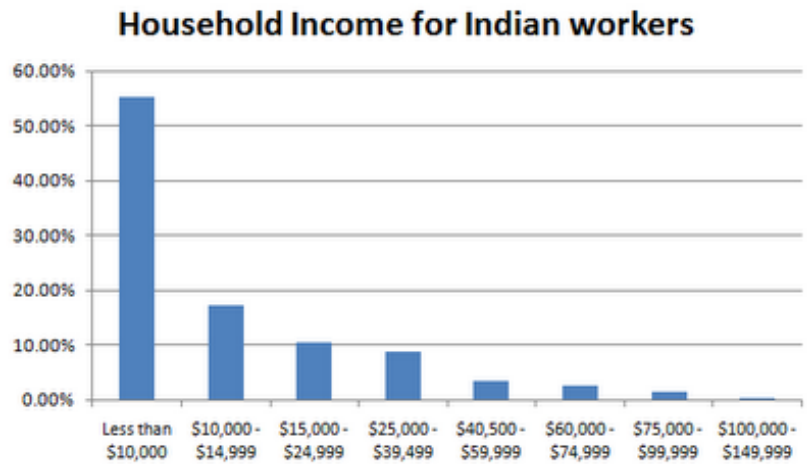
P. Ipeitorotis. March 2010

- 47% US, 34% India, 19% other



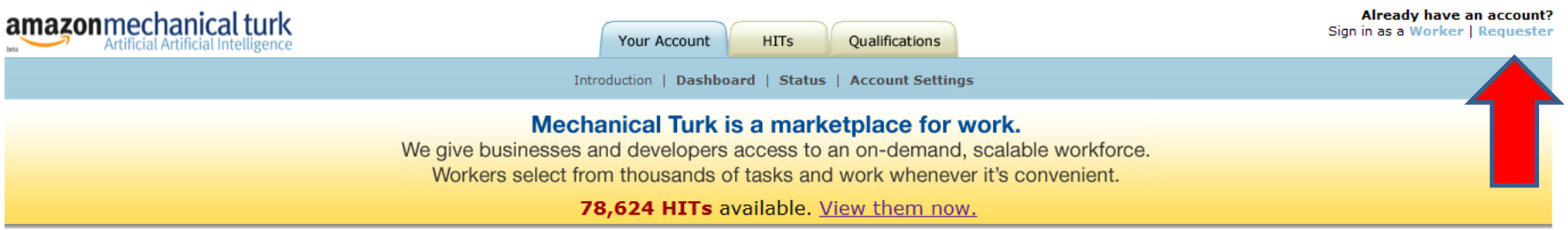




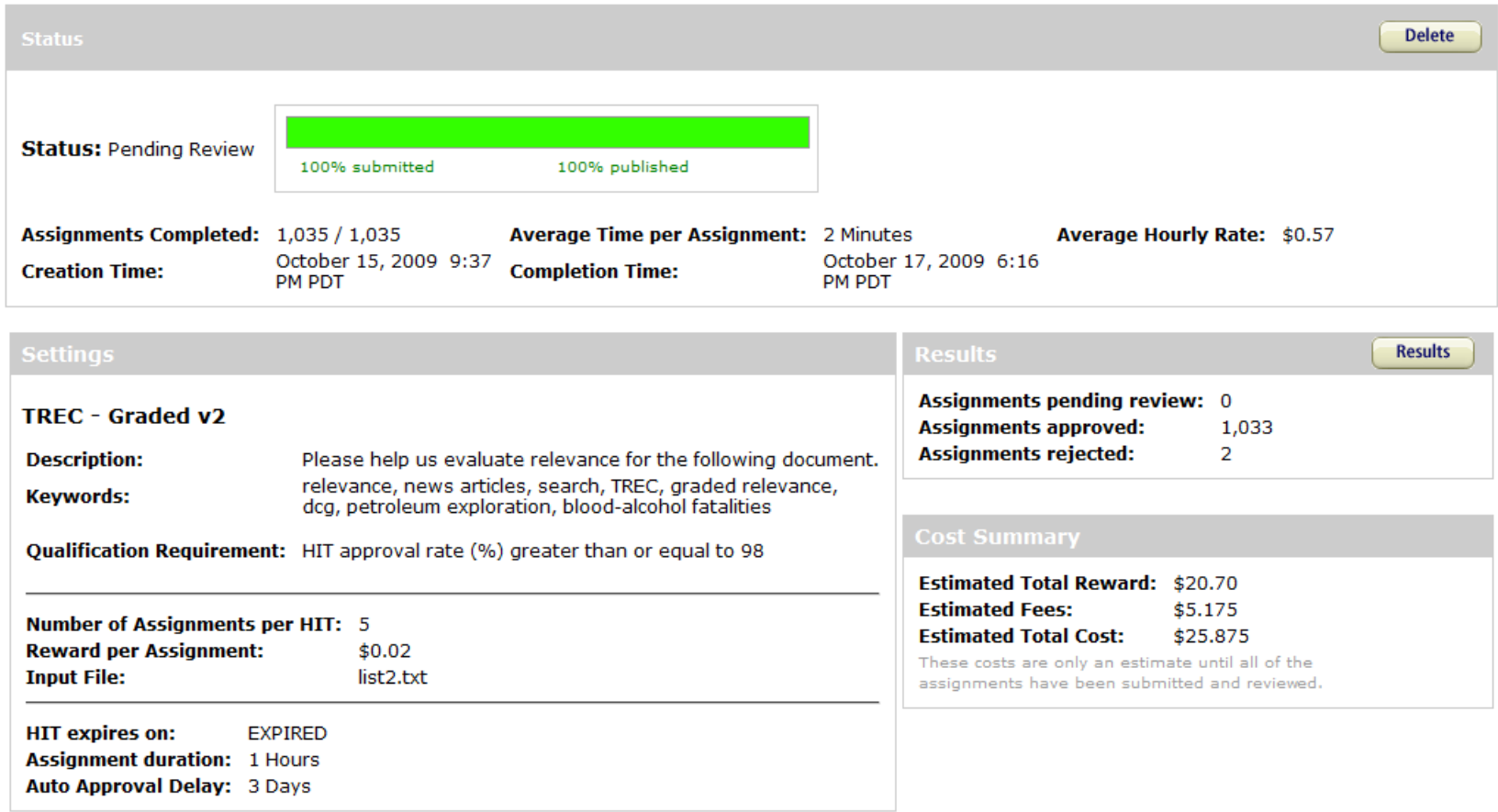


The Requester

- Sign up with your Amazon account
- Amazon payments
- Purchase prepaid HITs
- There is no minimum or up-front fee
- AMT collects a 10% commission
- The minimum commission charge is $0.005 per HIT



Dashboard - II



Quality control - II

- Approval rate
- Qualification test
  - Problems: slows down the experiment, difficult to "test" relevance
  - Solution: create questions on topics so user gets familiar *before* starting the assessment
- Still not a guarantee of good outcome
- Interject gold answers in the experiment
- Identify workers that always disagree with the majority

Filtering bad workers

- Approval rate
- Qualification test
  - Problems: slows down the experiment, difficult to "test" relevance
  - Solution: create questions on topics so user gets familiar *before* starting the assessment
- Still not a guarantee of good outcome
- Interject gold answers in the experiment
- Identify workers that always disagree with the majority

More on quality

- Lots of ways to control quality:
  - Better qualification test
  - More redundant judgments
  - More than 5 workers seems not necessary
- Various methods to aggregate judgments
  - Voting
  - Consensus
  - Averaging

Recent Workshops

- Human Computation: HCOMP 2009 & HCOMP 2010 at KDD

- IR: Crowdsourcing for Search Evaluation at SIGIR 2010
- NLP
  - The People's Web Meets NLP: Collaboratively Constructed Semantic Resources: 2009 at ACL-IJCNLP & 2010 at COLING
  - Creating Speech and Language Data With Amazon's Mechanical Turk. NAACL 2010
  - Maryland Workshop on Crowdsourcing and Translation. June, 2010
- ML: Computational Social Science and the Wisdom of Crowds. NIPS 2010
- Advancing Computer Vision with Humans in the Loop at CVPR 2010

- Conference: CrowdConf 2010 (organized by CrowdFlower)

News and Upcoming Events

New book: Omar Alonso, Gabriella Kazai, and Stefano Mizzaro. Crowdsourcing for Search Engine Evaluation: Why and How. To be published by Springer, 2011.

Special issue of Information Retrieval journal on Crowdsourcing (papers due 4/30)

Upcoming Conferences & Workshops

- HCOMP workshop at AAAI (papers due 4/22)
- SIGIR workshop? (in review)
- CrowdConf 2011 (TBA)

Events & Resources: http://ir.ischool.utexas.edu/crowd

Marketplaces

- Mturk (www.mturk.com)
- Crowdflower (www.crowdflower.com)
- Cloudcrowd, domystuff…
- Other resources:
  - http://blog.turkalert.com/
  - http://www.turkalert.com/
  - http://turkers.proboards.com

Tools and Packages

Common infrastructure layers atop or in place of MTurk or other platforms

- TurkIt
- Get Another Label (& qmturk)
- Turk Surveyor
- Ushandi

Thank You!

For questions about tutorial or crowdsourcing, email:

omar.alonso@microsoft.com

ml@ischool.utexas.edu

Cartoons by Mateo Burtch (buta@mindspring.com)

Bibliography

- O. Alonso, D. Rose, and B. Stewart. “Crowdsourcing for relevance evaluation”, SIGIR Forum, Vol. 42, No. 2 2008.
- O. Alonso and S. Mizzaro. “Can we get rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment”. SIGIR Workshop on the Future of IR Evaluation, 2009.
- O. Alonso, R. Schenkel, and M. Theobald. “Crowdsourcing Assessments for XML Ranked Retrieval”, 32nd ECIR 2010.
- O. Alonso and R. Baeza-Yates. “Design and Implementation of Relevance Assessments using Crowdsourcing, 33rd ECIR 2011.
- J. Barr and L. Cabrera. “AI gets a Brain”, ACM Queue, May 2006.
- K. Berberich, S. Bedathur, O. Alonso, G. Weikum “A Language Modeling Approach for Temporal Information Needs”, ECIR 2010
- Bernstein, M. et al. Soylent: A Word Processor with a Crowd Inside. UIST 2010. Best Student Paper award.
- Bederson, B.B., Hu, C., & Resnik, P. Translation by Iteractive Collaboration between Monolingual Users, Proceedings of Graphics Interface (GI 2010), 39-46.
- N. Bradburn, S. Sudman, and B. Wansink. Asking Questions: The Definitive Guide to Questionnaire Design, Jossey-Bass, 2004.
- C. Callison-Burch. “Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk”, EMNLP 2009.
- P. Dai, Mausam, and D. Weld. “Decision-Theoretic of Crowd-Sourced Workflows”, *AAAI*, 2010.
- J. Davis et al. “The HPU”, IEEE Computer Vision and Pattern Recognition Workshop on Advancing Computer Vision with Human in the Loop (ACVHL), June 2010.
- M. Gashler, C. Giraud-Carrier, T. Martinez. Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous, ICMLA 2008.
- C. Grady and M. Lease. “Crowdsourcing Document Relevance Assessment with Mechanical Turk”. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010.
- D. A. Grief. When Computers Were Human. Princeton University Press, 2005. ISBN 0691091579

Bibliography - II

- JS. Hacker and L. von Ahn. “Matchin: Eliciting User Preferences with an Online Game”, CHI 2009.
- M. Hearst. “Search User Interfaces”, Cambridge University Press, 2009
- J. Heer, M. Bobstock. “Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design”, CHI 2010.
- P. Heymann and H. Garcia-Molina. “Human Processing”, Technical Report, Stanford Info Lab, 2010.
- J. Howe. “Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business”. Crown Business, New York, 2008.
- P. Hsueh, P. Melville, V. Sindhwami. “Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria”. NAACL HLT Workshop on Active Learning and NLP, 2009.
- B. Huberman, D. Romero, and F. Wu. “Crowdsouring, attention and productivity”. Journal of Information Science, 2009.
- M. Kaisser, M. Hearst, and L. Lowe. “Improving Search Results Quality by Customizing Summary Lengths”, ACL/HLT, 2008.
- G. Kazai, N. Milic-Frayling, and J. Costello. “Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments”, SIGIR 2009.
- G. Kazai and N. Milic-Frayling. “On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing”. SIGIR Workshop on the Future of IR Evaluation, 2009.
- D. Kelly. “Methods for evaluating interactive information retrieval systems with users”. Foundations and Trends in Information Retrieval, 3(1-2), 1-224, 2009.
- A. Kittur, E. Chi, and B. Suh. “Crowdsourcing user studies with Mechanical Turk”, SIGCHI 2008.
- K. Krippendorff. "Content Analysis", Sage Publications, 2003
- G. Little, L. Chilton, M. Goldman, and R. Miller. “TurKit: Tools for Iterative Tasks on Mechanical Turk”, KDD-HCOMP 2009.
- H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. “Improving Search Engines Using Human Computation Games”, CIKM 2009.
- T. Malone, R. Laubacher, and C. Dellarocas. Harnessing Crowds: Mapping the Genome of Collective Intelligence. MIT Press, 2009.

Bibliography - III

- W. Mason and D. Watts. “Financial Incentives and the ’Performance of Crowds’”, HCOMP Workshop at KDD 2009.
- S. Mizzaro. Measuring the agreement among relevance judges, MIRA 1999
- J. Nielsen. “Usability Engineering”, Morgan-Kaufman, 1994.
- A. Quinn and B. Bederson. “A Taxonomy of Distributed Human Computation”, Technical Report HCIL-2009-23, University of Maryland, Human-Computer Interaction Lab, 2009
- J. Ross, L. Irani, M. Six Silberman, A. Zaldivar, and B. Tomlinson. “Who are the Crowdworkers? Shifting Demographics in Mechanical Turk”. CHI 2010.
- J. Tang and M. Sanderson. “Evaluation and User Preference Study on Spatial Diversity”, ECIR 2010
- F. Scheuren. “What is a Survey” (http://www.whatisasurvey.info) 2004.
- R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. “Cheap and Fast But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks”. EMNLP-2008.
- V. Sheng, F. Provost, P. Ipeirotis. “Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers” KDD 2008.
- S. Weber. “The Success of Open Source”, Harvard University Press, 2004.
- L. von Ahn. Games with a purpose. Computer, 39 (6), 92–94, 2006.
- L. von Ahn and L. Dabbish. “Designing Games with a purpose”. CACM, Vol. 51, No. 8, 2008.
- T. Yan, V. Kumar, and D. Ganesan. CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. MobiSys pp. 77--90, 2010.