Crowdsourcing 101: Putting the Wisdom of Crowds to Work for You

Omar Alonso
Microsoft

Matthew Lease
University of Texas at Austin

9 February 2011

Tutorial objectives

- What is crowdsourcing?
- How and when to use crowdsourcing?
- How to use Mechanical Turk
- Experimental setup and design guidelines for working with the crowd
- Quality control: issues, measuring, and improving
- Research landscape and open challenges

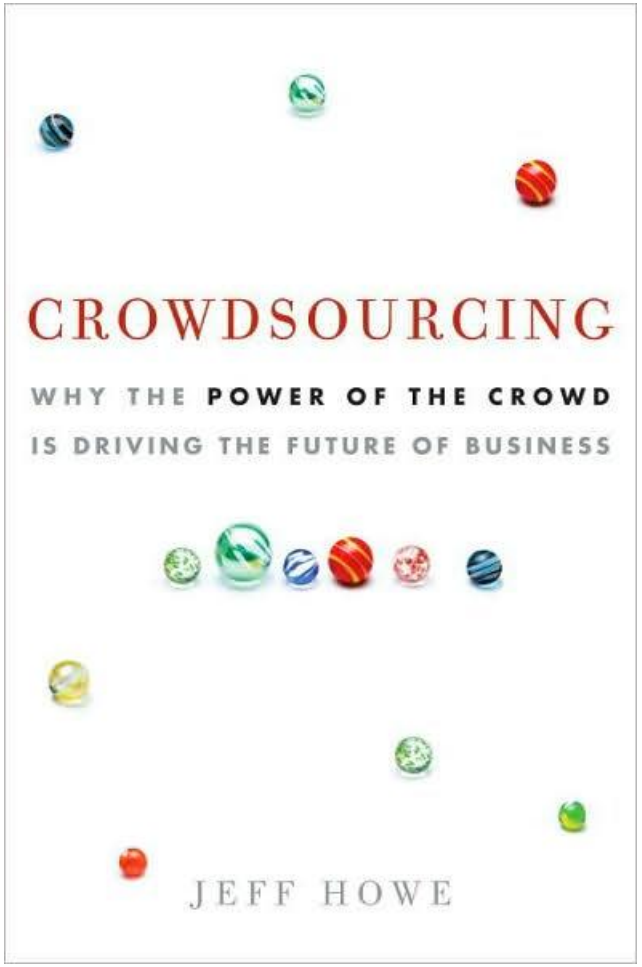
Tutorial Outline

I. Introduction to crowdsourcing
II. Amazon Mechanical Turk (and CrowdFlower)
III. Design of experiments

INTRODUCTION TO CROWDSOURCING

Crowdsourcing

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



Examples

INNOCENTIVE®

XPRIZE FOUNDATION

NETFLIX®

Less Serious Examples

CRAZY OLD MAN LETTERS

Challenge Winners

PhD Challenge

2010 – "I Smoke Crack Rocks"

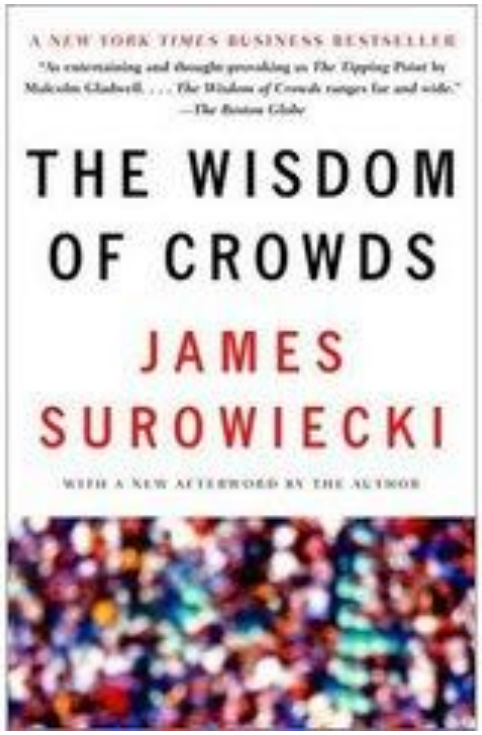
Toward Better Crowd Sourced Transcription: Transcription Of A Year Of The Let's Go Bus Information System Data

By **Gabriel Parent**, Carnegie Mellon University
2010 PhD Challenge Winner
In *Proceedings of IEEE Workshop on Spoken Language Technology* (December 2010)
Announcement • **Call for Participation** • **Eligibility Rules** • **FAQ**

Wisdom of Crowds (WoC)

Requires

- Diversity
- Independence
- Decentralization
- Aggregation



Input: large, diverse sample

(to increase likelihood of overall pool quality)

Output: consensus or selection (aggregation)

WoC vs. Ensemble Learning

- Combine multiple models to improve performance over any constituent model
  - Can use many *weak learners* to make a strong one
  - Compensate for poor models with extra computation
- Tend to work better when significant diversity
  - Using less diverse strong learners better than *dumbing-down* models to promote diversity (Gashler et al.'08)
- cf. NIPS'10 Workshop
  - Computational Social Science & the Wisdom of Crowds

Human Computation

- Use humans as processors in a distributed system
  - Humans do tasks computers cannot (do well)
  - System makes opaque "external call" to the "HPU" ("AAI")
    - HPU has describable functional capabilities (J. Davis et al., ACVHL'10)
- Ex. 1: Detect CPU (Captcha – "reverse Turing test")
- Ex. 2: HPU computation (e.g. labeled data)
  - linear, with minimal post-processing of HPU output
  - E.g. ReCaptcha, ESP game, most Mechanical Turk work
- Ex. 3: Integrate CPU + HPU computation
  - HPU part of core system architecture, many invocations
  - E.g. CrowdSearch, Soylent, Monolingual Translation

***L. von Ahn has pioneered the field. See bibliography for examples of his* work.**

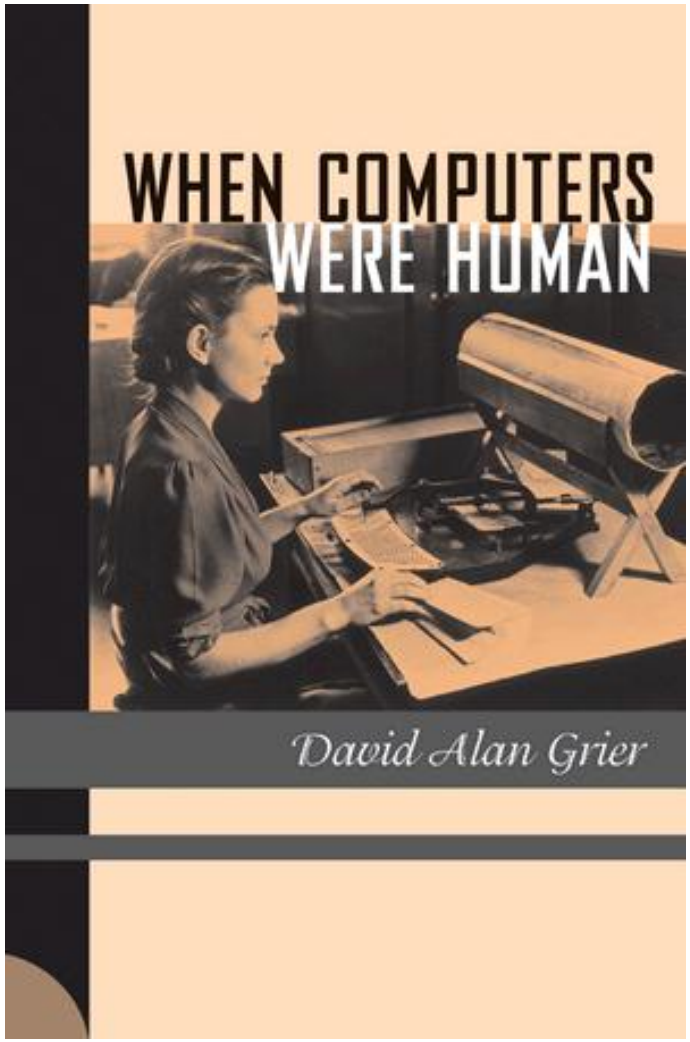
A New Class of Applications

CPU + HPU hybrid applications blend automation with human computation 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*

Human Computation

- Not a new idea
- Computers before computers

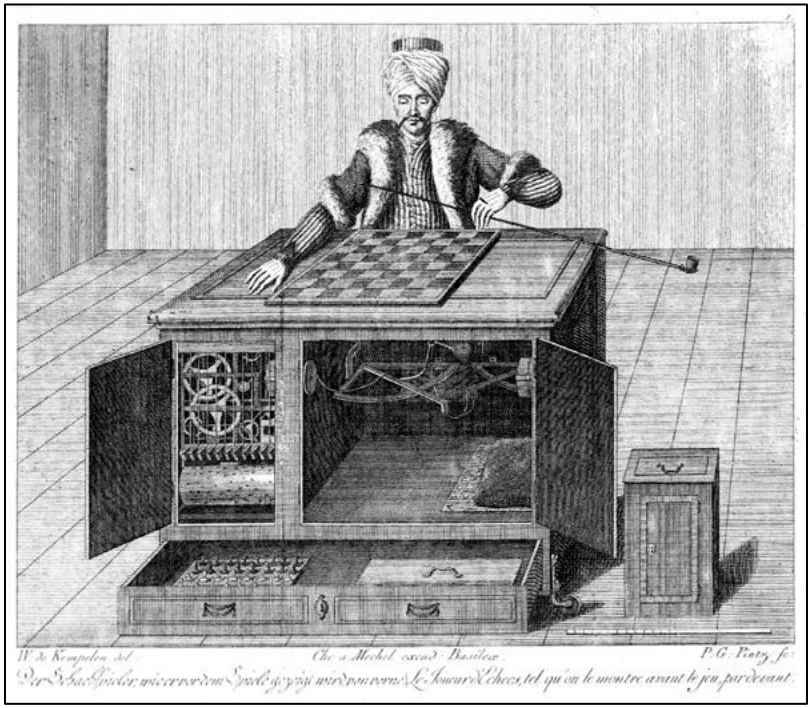


Pay-based Marketplaces / Vendors

- Mechanical Turk (since 2005, www.mturk.com)
- Crowdflower (since 2007, www.crowdflower.com)
- CloudCrowd (cf.
- DoMyStuff
- Livework
- Clickworker
- SmartSheet
- uTest
- Elance
- oDesk
- vWorker (was rent-a-coder)

Amazon Mechanical Turk (AMT, Mturk)

- Crowdsourcing platform
- On-demand workforce
- Went online in 2005
- “Artificial artificial intelligence” (AAI)
- Named after fake chess playing machine by Wolfgang von Kempelen in 18th Century



J. Pontin. Artificial Intelligence, With Help From the Humans (NY Times, March 25, 2007)

AMT Overview

- 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
- Sponsorship: NAMT’10, TREC’10 RF Track…

Mechanical Turk Tracker

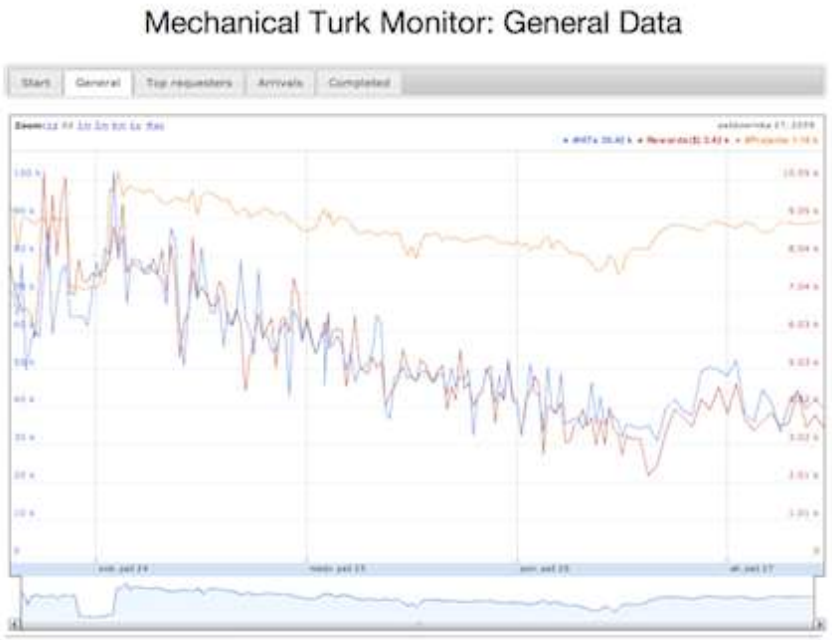
General | Top requesters | Arrivals | Search | About

How does it work?

We use a web crawler written in python to gather all avaliable information from Amazon Mechanical Turk, we store that information PostgreSQL and everything is working on Amazon EC2 instances. We crawl Mechanical Turk each hour, and compute daily statistics for new projects and completed tasks once a day.

What type of graphs?

- General graphs - listing of projects, total number of available hits and rewards for every hour
- Top 1000 requesters - sorted according to cumulative rewards
- Top 1000 requesters - sorted according to cumulative rewards
- Arrivals - number of new projects, available hits and rewards computed on daily basis



http://www.mturk-tracker.com (P. Ipeirotis'10)

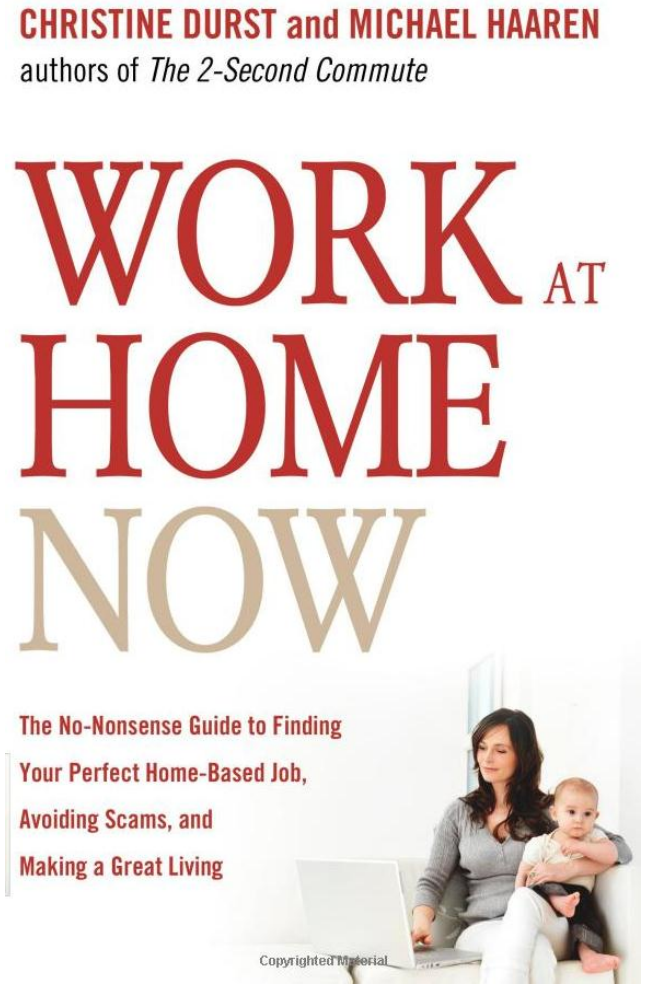
From 1/09 – 4/10, 7M HITs from 10K requestors worth $500,000 USD (assumes only 1 worker/HIT)

Who are the workers?

- A. Baio, November 2008. The Faces of Mechanical Turk.
- P. Ipeitorotis. March 2010. The New Demographics of Mechanical Turk
- J. Ross, et al. Who are the Crowdworkers?... CHI 2010.

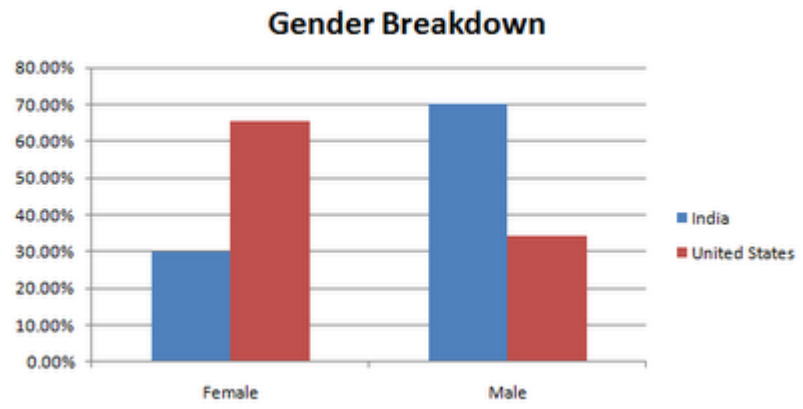
Worker Demographics

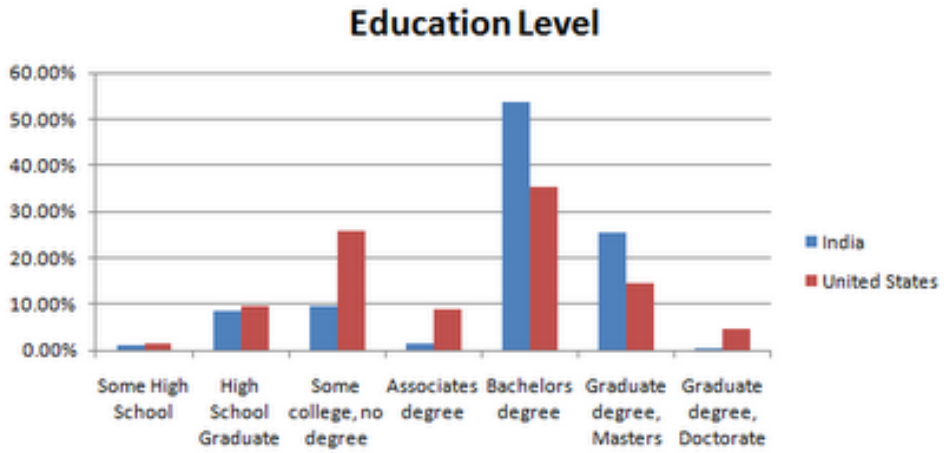
- 2008-2009 studies found less global and diverse than previously thought
  - US
  - Female
  - Educated
  - Bored
  - Money is secondary

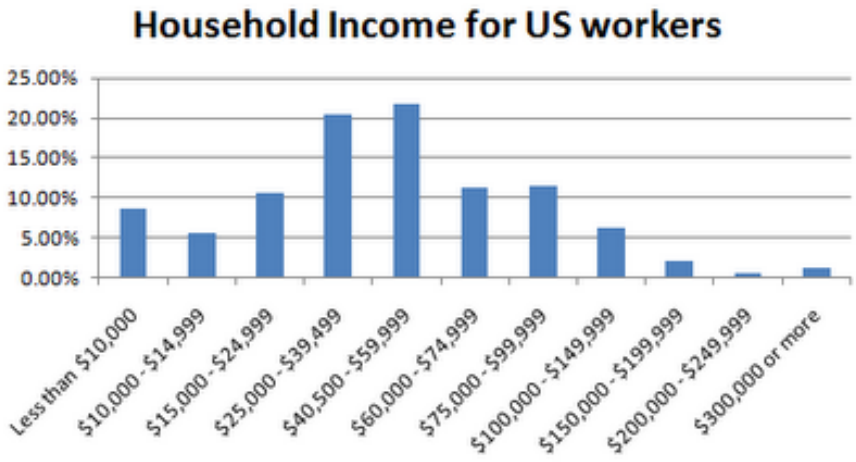


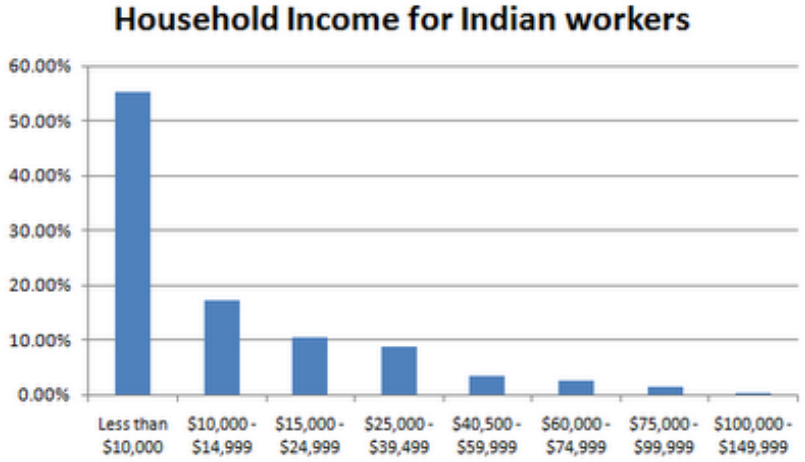
2010 shows increasing diversity

47% US, 34% India, 19% other (P. Ipeitorotis. March 2010)

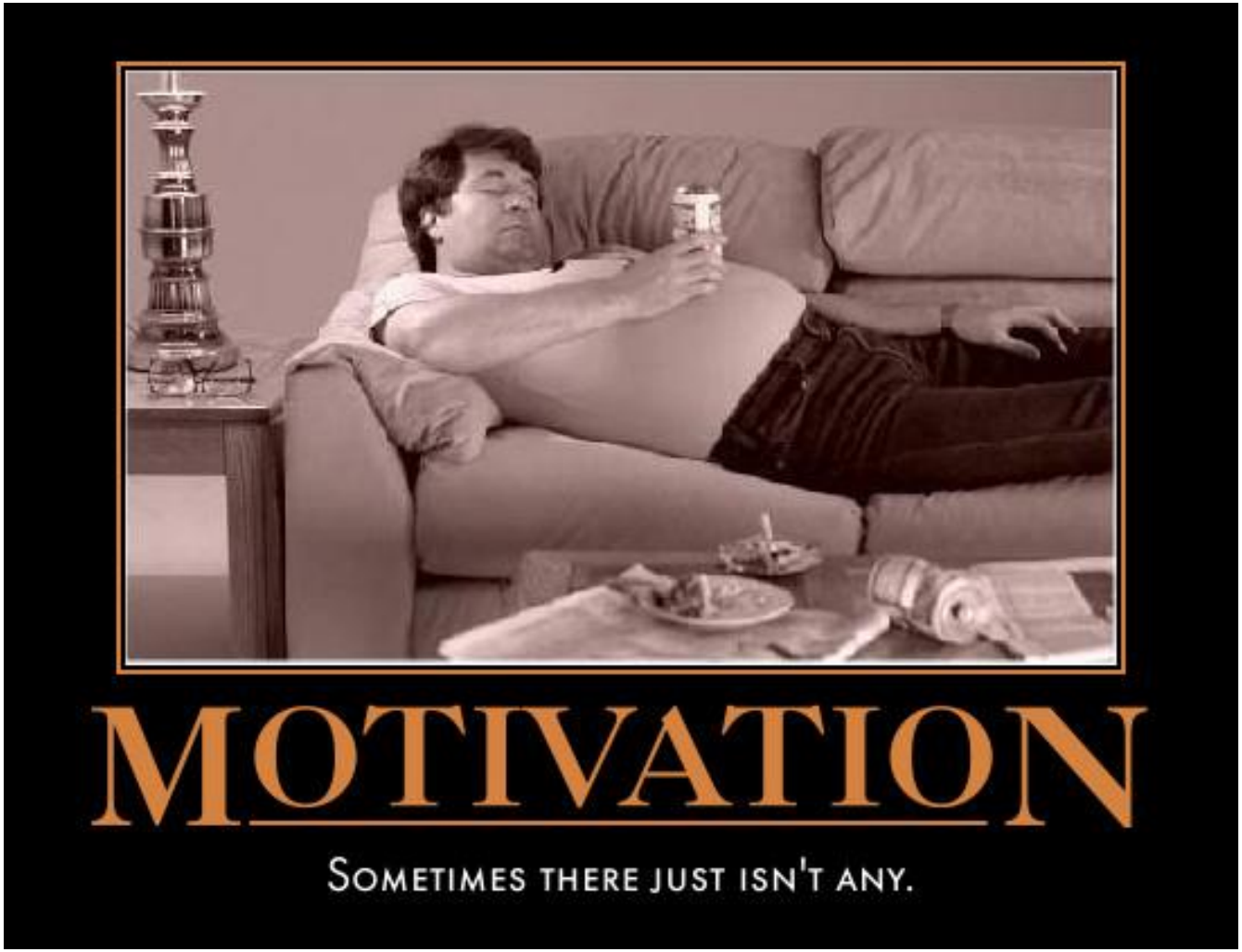






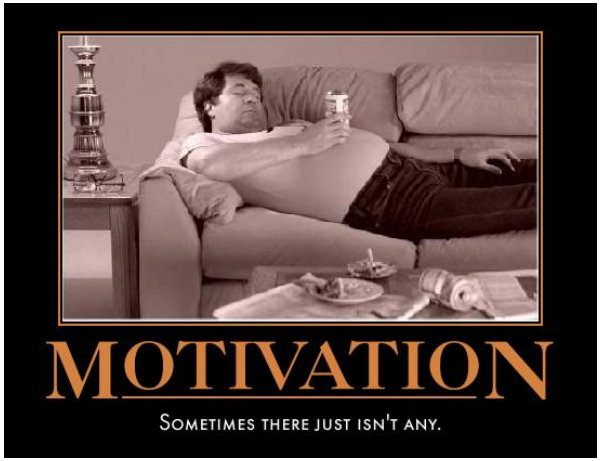


Worker incentives: why do they do it?



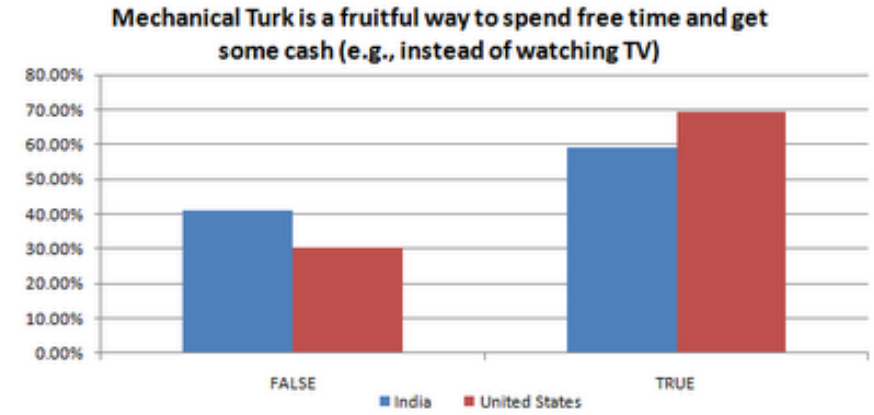
Worker Incentives

- Pay ($$$)
- Fun (or avoid boredom)
- Socialize
- Earn acclaim/prestige
- Altruism
- Learn something new (e.g. English)
- Unintended by-product (e.g. re-Captcha)
- Create self-serving resource (e.g. Wikipedia)

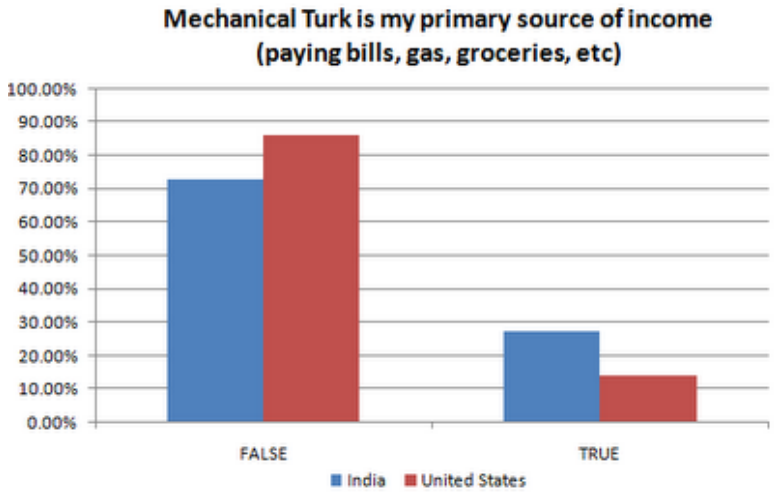


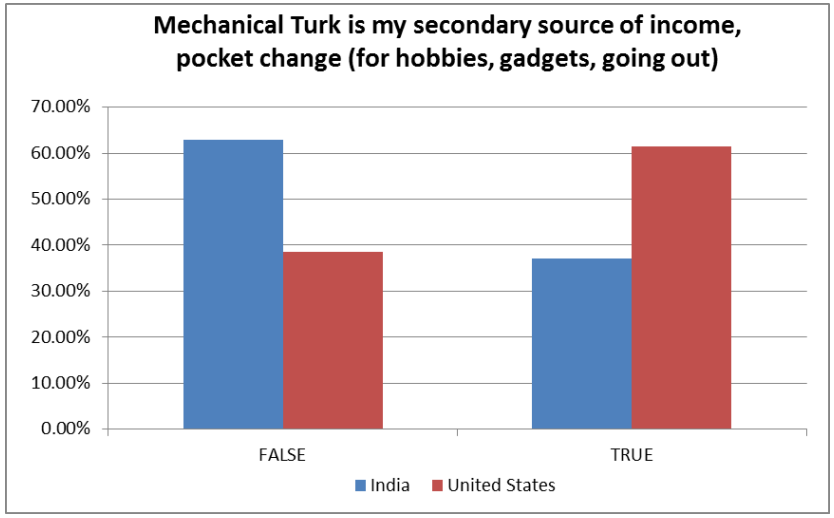
Multiple incentives are typically at work in parallel

Pay ($$$)



P. Ipeirotis March 2010





Pro

- Ready marketplaces (e.g. MTurk, CrowdFlower, …)
- Less need for creativity
- Simple motivation knob

Con: Quality and Quality control required

- Can diminish intrinsic rewards that promote quality:
  - Fun/altruistic value of task
  - Taking pride in doing quality work
  - Self-assessment
- Can attract workers only interested in the pay, fraud
- $$$ (though other schemes cost indirectly)

Pay ($$$)

How much to pay?

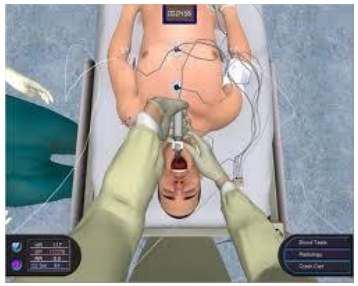
- Mason & Watts 2009: more $ = more work, not better wo…
- Wang et al. 2011: predict from market?
- More later…

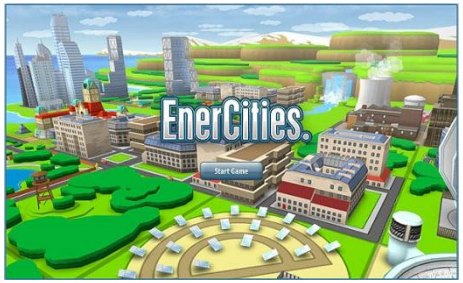
Zittrain 2010: if Encarta had paid for contributions, would we have Wikipedia?

Fun (or avoid boredom)

- *Games with a Purpose* (von Ahn)
  - Data is by-product
  - IR: Law et al. SearchWar. HCOMP 2009.

- distinct from *Serious Gaming / Edutainment*
  - Player learning / training / education is by-product





Fun (or avoid boredom)

- Pro:
  - Enjoyable “work” people want to do (or at least better than anything else they have to do)
  - Scalability potential from involving non-workers
- Con:
  - Need for design creativity
    - some would say this is a plus
    - better performance in game should produce better/more work
  - Some tasks more amenable than others
    - Annotating syntactic parse trees for fun?
    - Inferring syntax implicitly from a different activity?

Socialization & Prestige

wondir
real people, real answers, real fast
www.wondir.com

YAHOO! ANSWERS
Example of Community Question and Answering

WIKIPEDIA
The Free Encyclopedia

WikiAnswers

Facebook Questions
Learn from people in the know and the friends who know you best.
What do you want to know? Ask Question

fluther.com
tap the collective

Quora
A continually improving collection of questions and answers created, edited, and organized by everyone who uses it.

Socialization & Prestige

- Pro:
  - "free"
  - enjoyable for connecting with one another
  - can share infrastructure across tasks
- Con:
  - need infrastructure beyond simple micro-task
  - need critical mass (for uptake and reward)
  - social engineering knob more complex than $$

Altruism

- Contribute knowledge
- Help others (who need knowledge)
- Help workers (e.g. SamaSource)
- Charity (e.g. http://www.freerice.com)

YAHOO! ANSWERS
Example of Community Question and Answering

Altruism

- Pro
  - “free”
  - can motivate quality work for a cause
- Con
  - Seemingly small workforce for pure altruism

What if Mechanical Turk let you donate $$ per HIT?

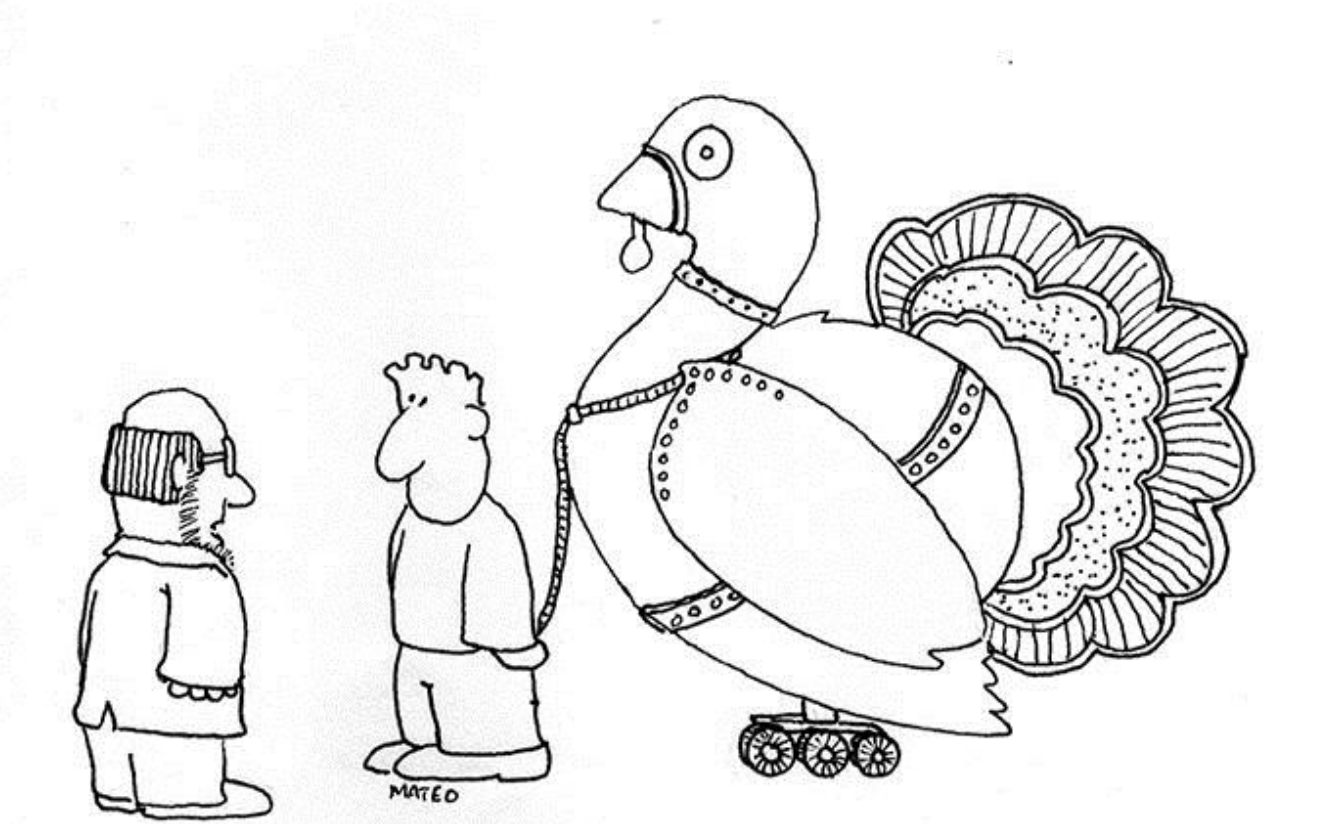
Unintended by-product

- Pro
  - effortless (unnoticed) work
  - Scalability from involving non-workers
- Con
  - Design challenge
    - Given existing activity, find useful work to harness from it
    - Given target work, find or create another activity for which target work is by-product?
  - Maybe too invisible (disclosure, manipulation)

Multiple Incentives

- Ideally maximize all
- Wikipedia, cQA, Gwap
  - fun, socialization, prestige, altruism
- Fun vs. Pay
  - gwap gives Amazon certificates
  - Workers maybe paid in game currency
  - Pay tasks can also be fun themselves
- Pay-based
  - Other rewards: e.g. learn something, socialization
  - altruism: worker (e.g. SamaSource) or task itself
  - social network integration could help everyone (currently separate and lacking structure)

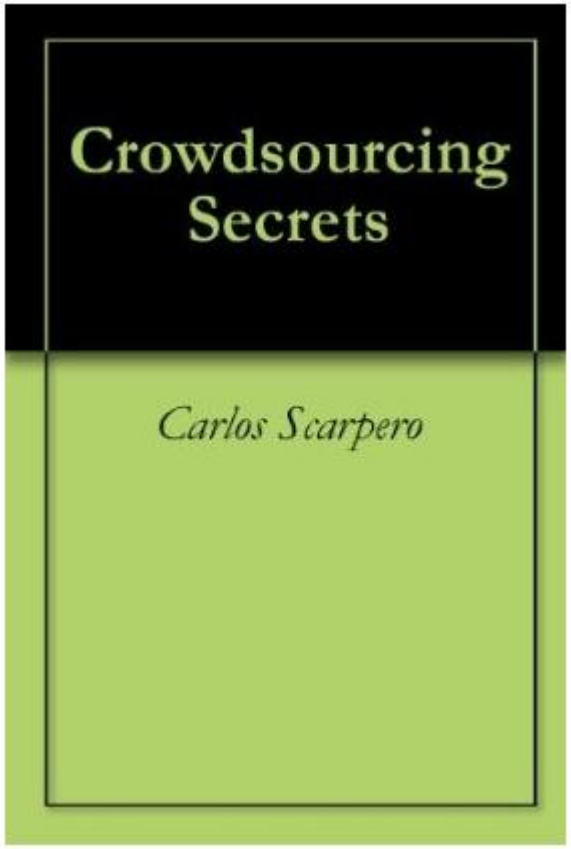
AMAZON MECHANICAL TURK



"No, mechanical ***Turk.****"*

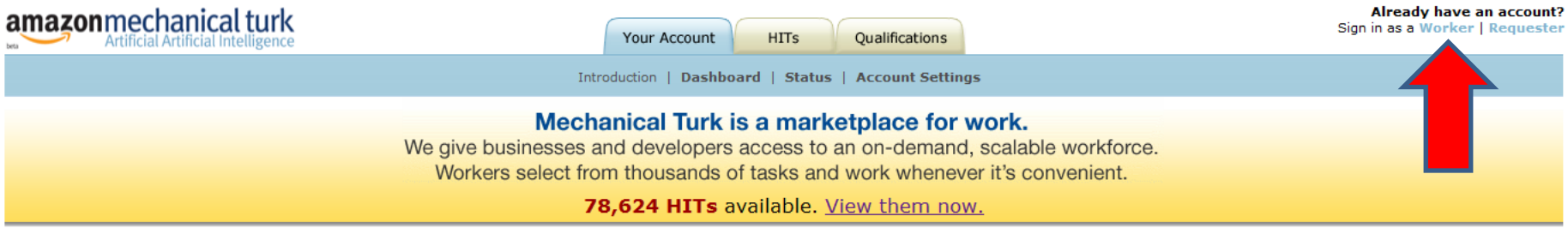
Little *how-to* written for the public

- July 2010, *kindle*-only
- “This book introduces you to the top crowdsourcing sites and outlines step by step with photos the exact process to get started as a requester on Amazon Mechanical Turk.“



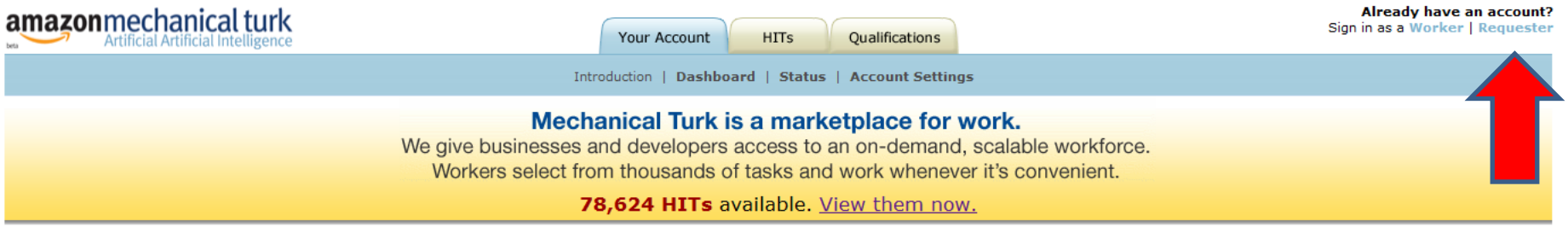
The Worker

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



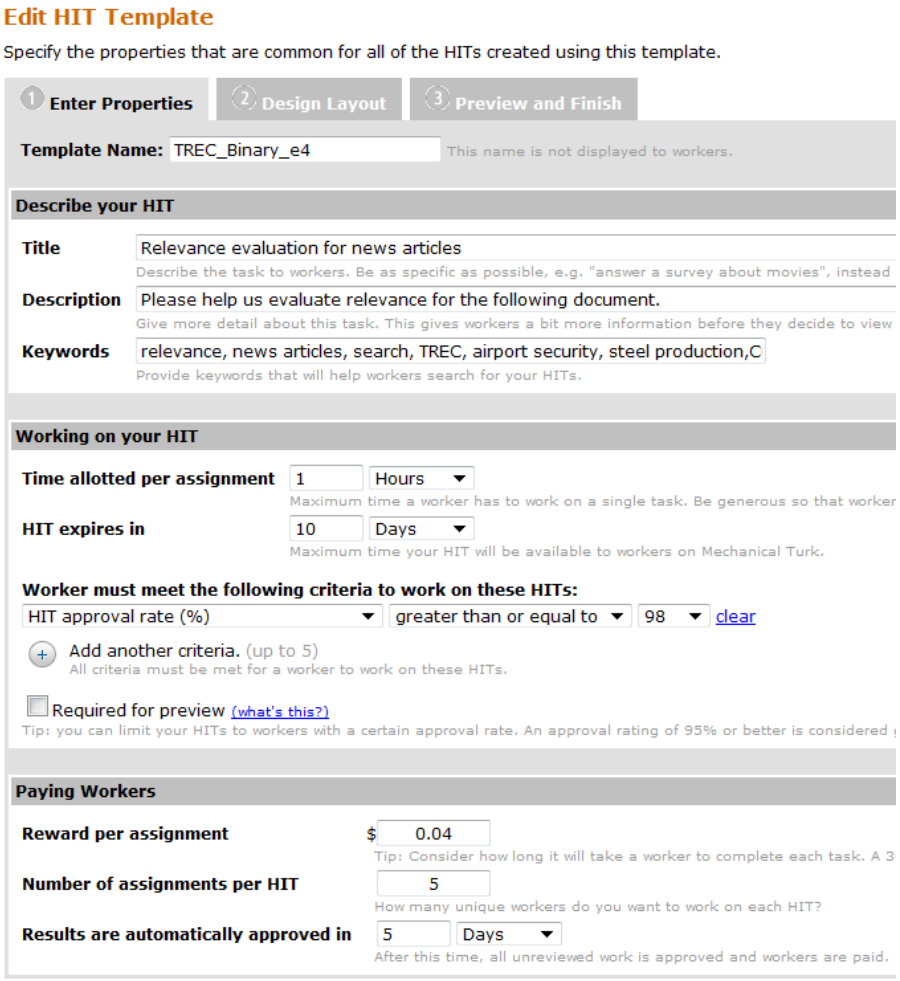
The Requester

- 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

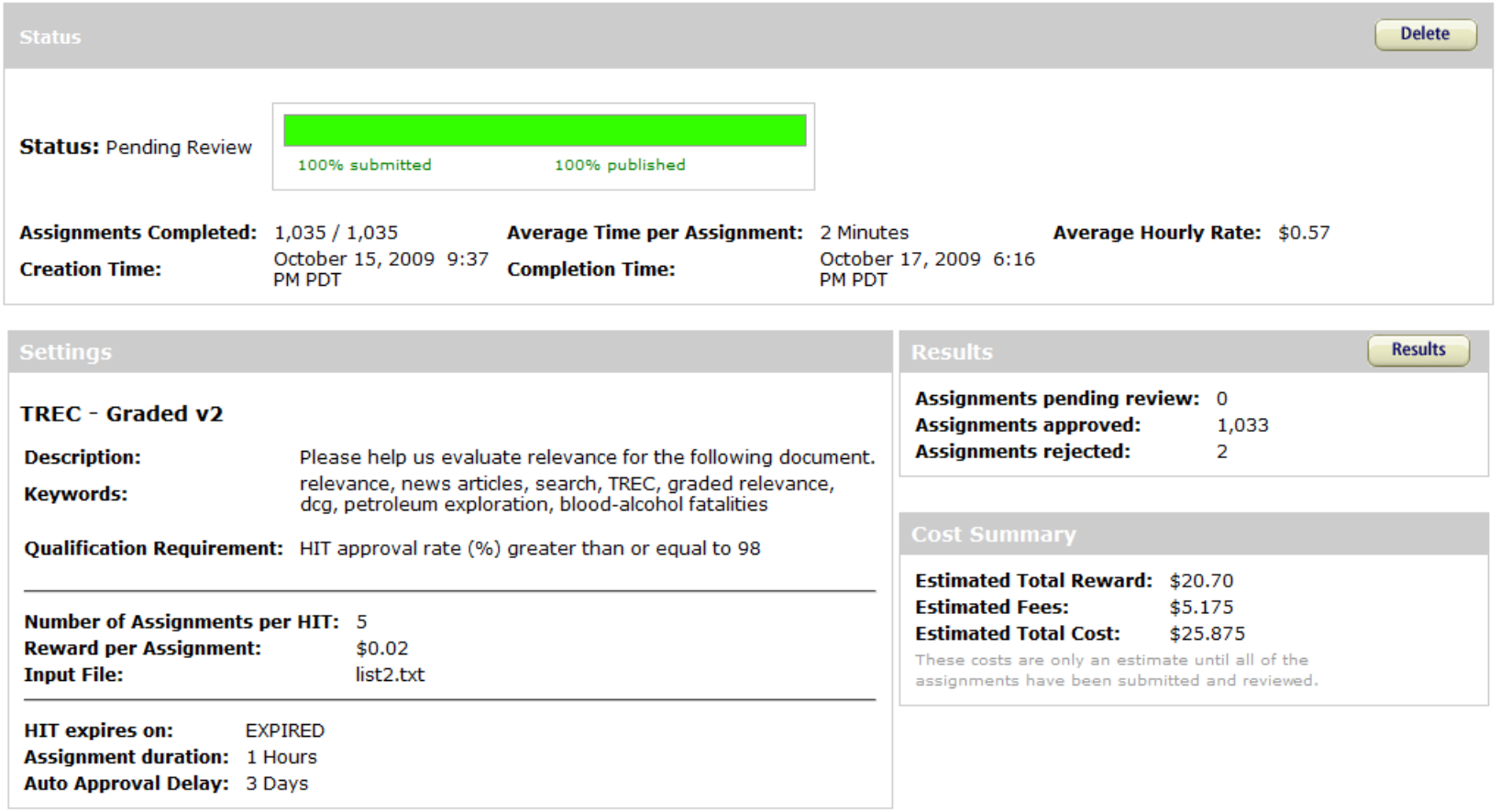


Dashboard

- Three tabs
  - Design
  - Publish
  - Manage
- Design
  - HIT Template
- Publish
  - Make work available
- Manage
  - Monitor progress



Dashboard - II



API

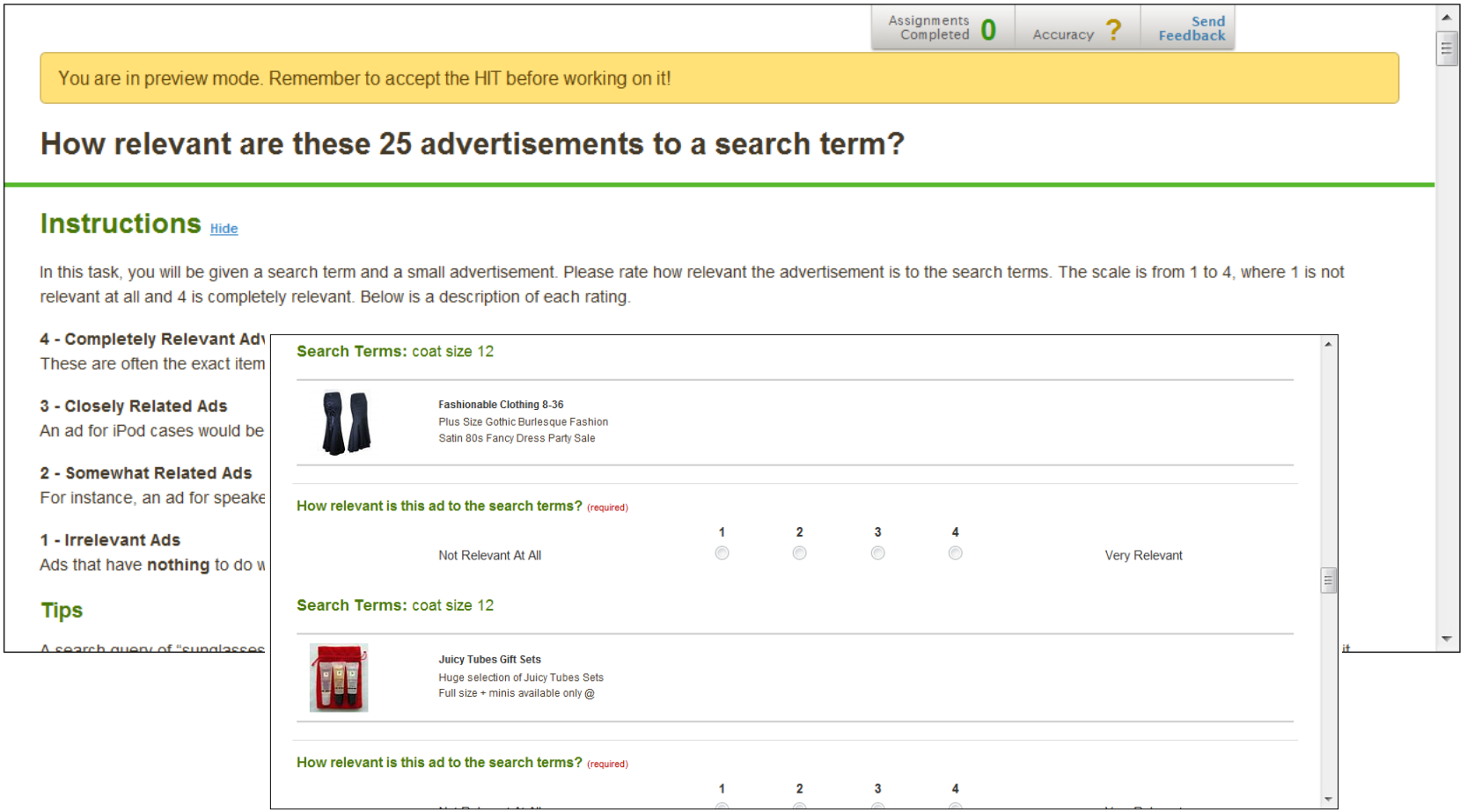
- Amazon Web Services API
- Rich set of services
- Command line tools
- More flexibility than dashboard

Practical discussion

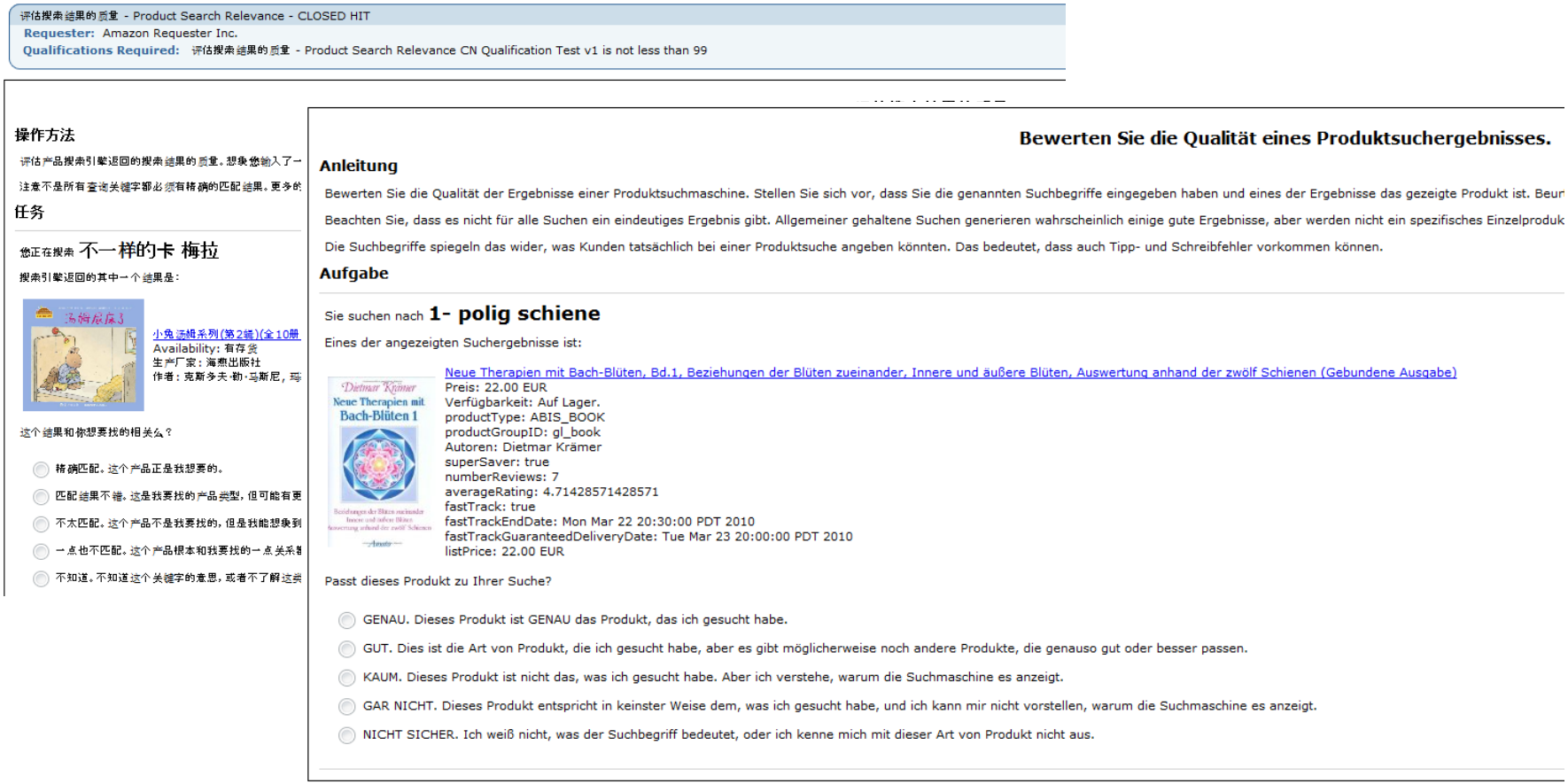
- Dashboard
  - Easy to prototype
  - Setup and launch an experiment in a few minutes
- API
  - Ability to integrate AMT as part of a system
  - Ideal if you want to run experiments regularly
  - Schedule tasks

AMT EXAMPLES

Example – Relevance and ads



Example – Product search



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

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.

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

Example - Multilingual

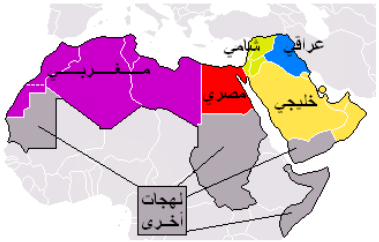
Help Classify Arabic into Dialects!

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.

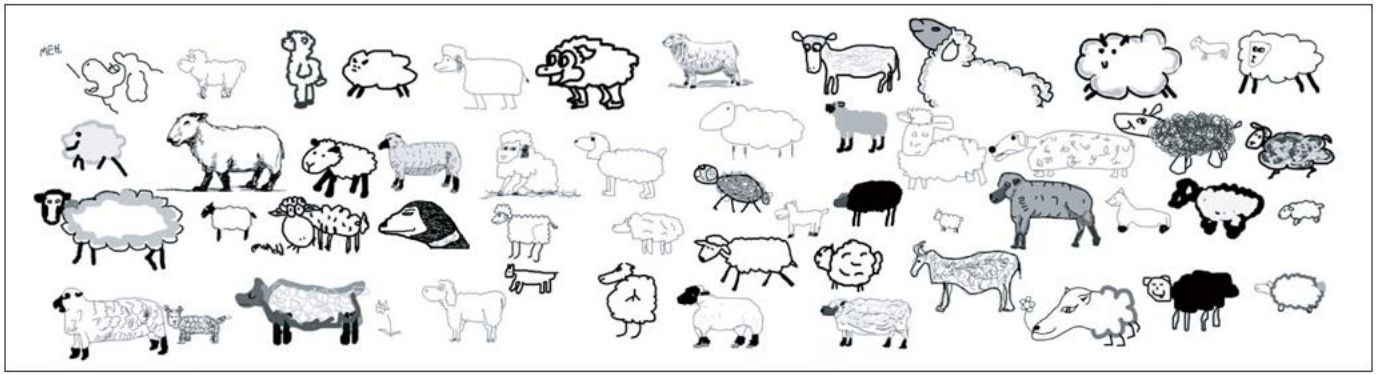
Confidentiality: The only identifying information kept about you will be a WorkerID serial number and your IP address. This information may be disclosed to other researchers.

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.

Example – Sheep Market

- Collection of 10,000 sheep made by workers
- Payment $0.02 to draw a sheep facing left



www.thesheepmarket.com

Example – Ten Thousand Cents

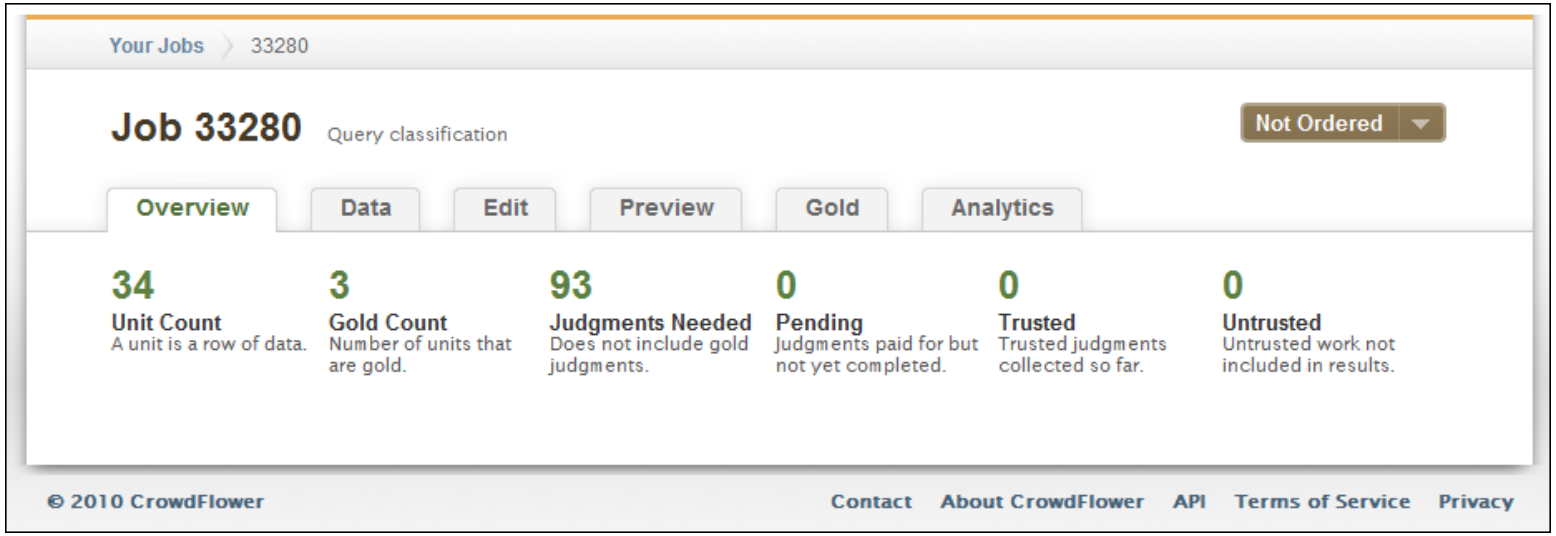
- Creates a representation of a $100 bill
- Workers painted a part of the bill
- Payment $0.01

www.tenthousandcents.com


CrowdFlower

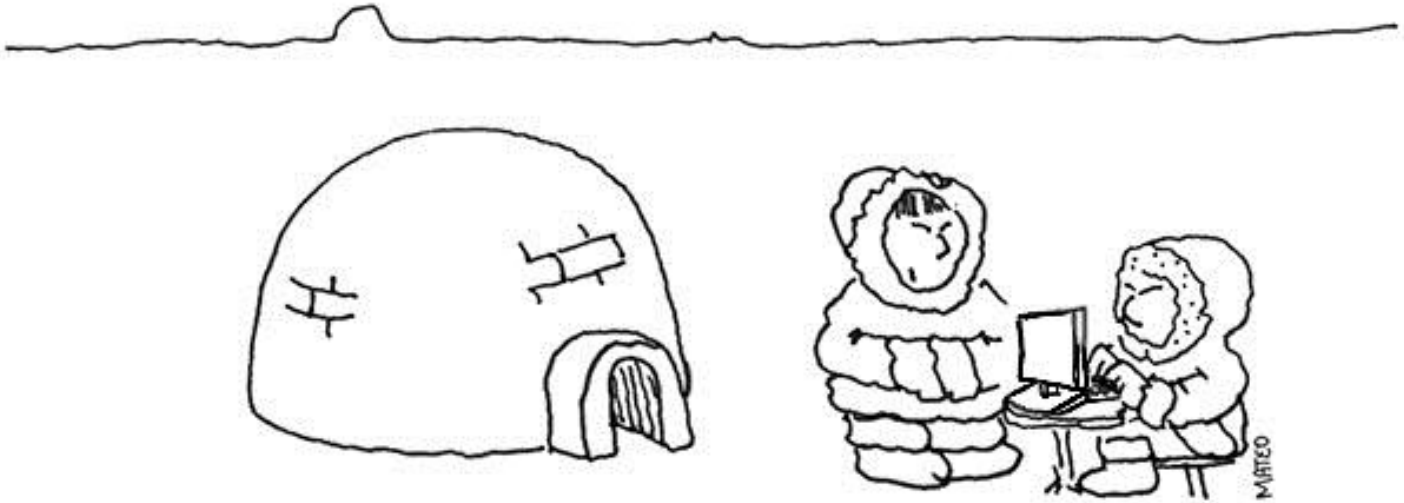
CrowdFlower (founded in 2007)

- Labor on-demand
- Channels
- Quality control features
- Sponsor NAMT'10, CSE'10, CSDM'11, SIGIR'11 workshop…



Next steps

- Evidence from a wide range of projects
- Can I *crowdsource* my experiment?
- How do I start?
- What do I need?



"Snow. Snow is relevant."

Relevance Assessment & Crowdsourcing

Relevance and IR

- What is relevance?
  - Multidimensional
  - Dynamic
  - Complex but systematic and measurable
- Relevance in Information Retrieval
- Frameworks
- Types
  - System or algorithmic
  - Topical
  - Pertinence
  - Situational
  - Motivational

Evaluation

- Relevance is hard to evaluate
  - Highly subjective
  - Expensive to measure
- Click data
- Professional editorial work
- Verticals

You have a new idea

- Novel IR technique
- Don’t have access to click data
- Can’t hire editors
- How to test new ideas?

Crowdsourcing and relevance evaluation

- For relevance, it combines two main approaches
  - Explicit judgments
  - Automated metrics
- Other features
  - Large scale
  - Inexpensive
  - Diversity

Why is this interesting?

- Easy to prototype and test new experiments
- Cheap and fast
- No need to setup infrastructure
- Introduce experimentation early in the cycle
- In the context of IR, implement and experiment as you go
- For new ideas, this is very helpful

Important clarifications

- Trust and reliability
- Spam
- Wisdom of the crowd re-visit
- Adjust expectation
- Crowdsourcing is another data point for analysis
- Complementary to other experiments

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.

- Learning to map from web pages to queries
- Human computation game to elicit data
- Home grown system (no AMT)
- Try it!

pagehunt.msrlivelabs.com

See also:

- H. Ma. et al. "Improving Search Engines Using Human Computation Games", CIKM 2009.
- Law et al. SearchWar. HCOMP 2009.
- Bennett et al. Picture This. HCOMP 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.

TREC

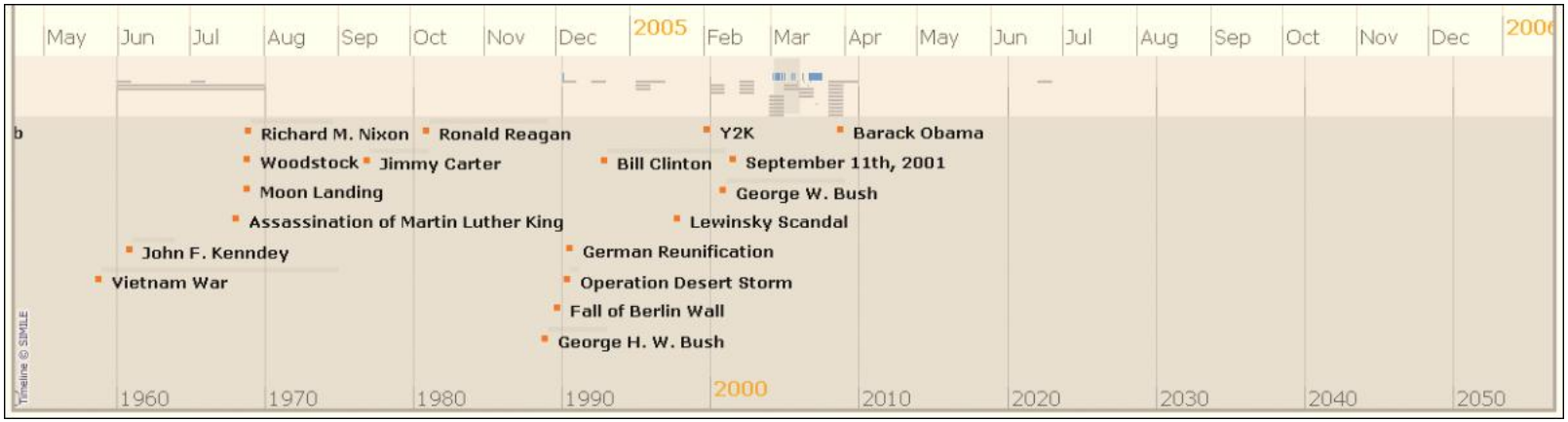
- Can we replace TREC-like relevance assessors with Mechanical Turk?
- Selected topic “space program” (011)
- Modified original 4-page instructions from TREC
- Workers more accurate than original assessors
- 40% provided justification for each answer

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.

"I'm searching on 'precooked meat product', but all I'm getting is spam."

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

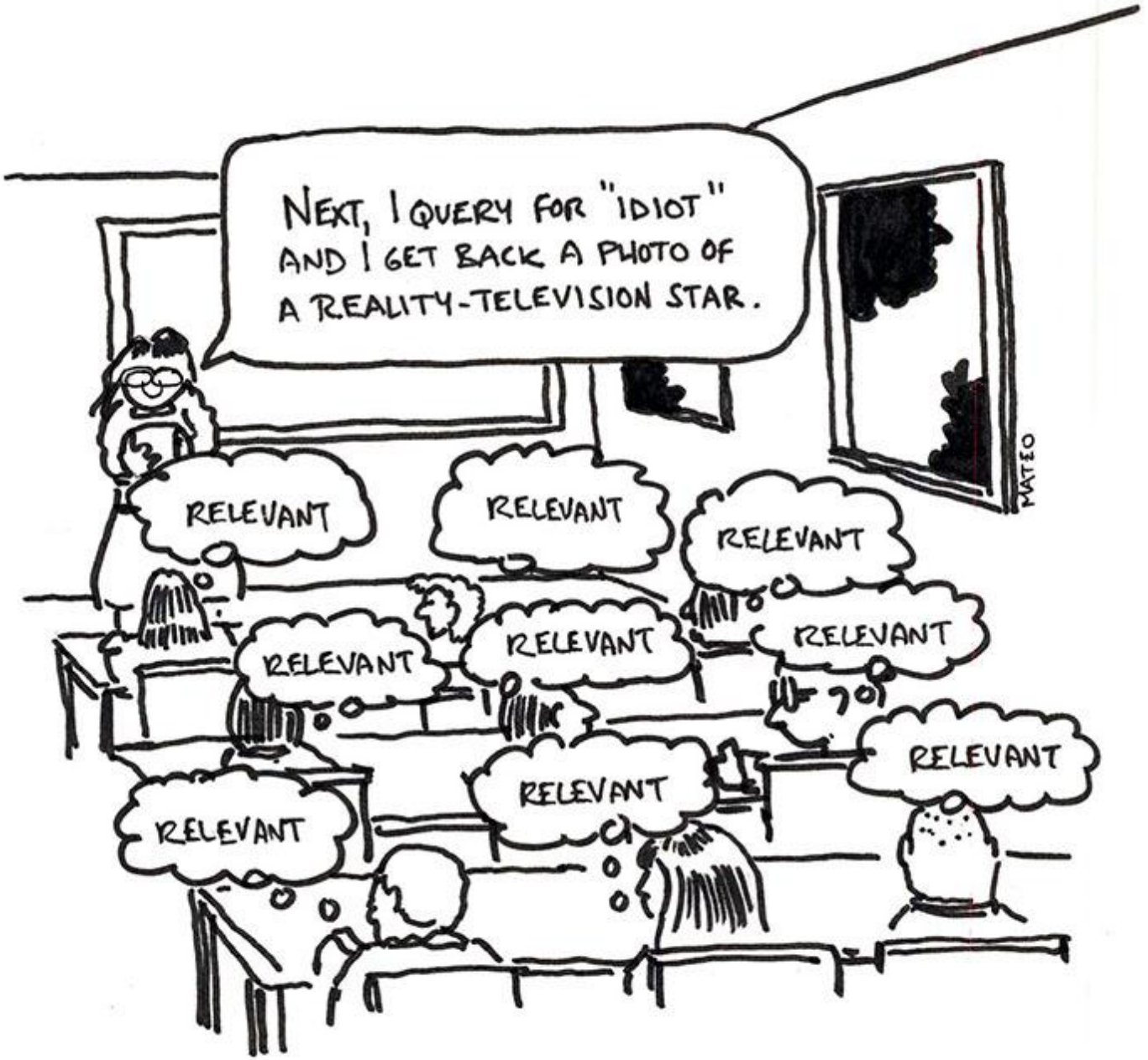
Twitter

- 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

BREAK

Hands on

- Design two experiments
- Show all details
- Launch and monitor progress


NEXT, I QUERY FOR "IDIOT" AND I GET BACK A PHOTO OF A REALITY-TELEVISION STAR.
MATEO
RELEVANT
RELEVANT
RELEVANT
RELEVANT
RELEVANT
RELEVANT
RELEVANT
RELEVANT
RELEVANT

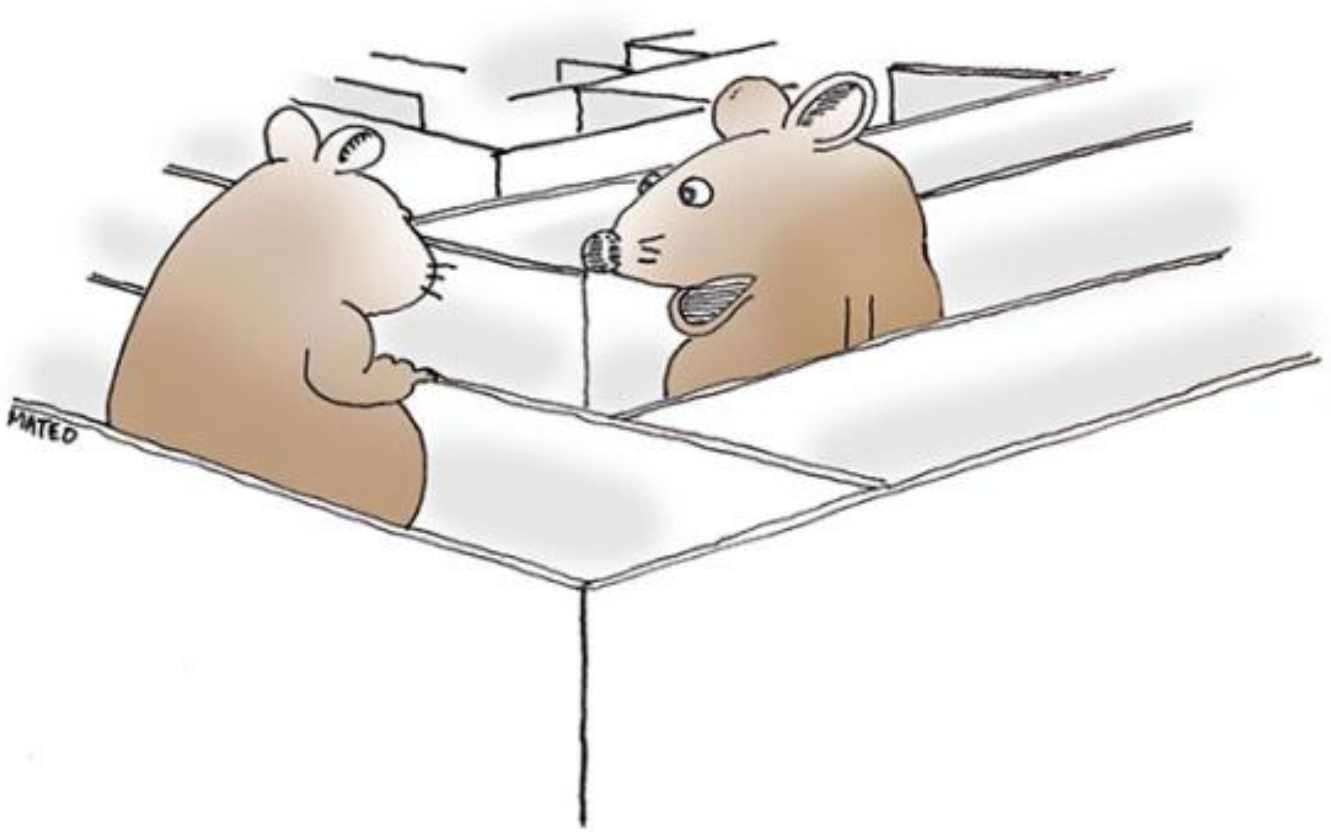
Query classification task

- Ask the user to classify a query
- Show a form that contains a few categories
- Upload a few queries (~20)
- Use 5 workers

Relevance evaluation task

- Relevance assessment task
- Use a few documents from TREC
- Ask user to perform binary evaluation
- Modification: graded evaluation
- Use 5 workers

DESIGN OF EXPERIMENTS



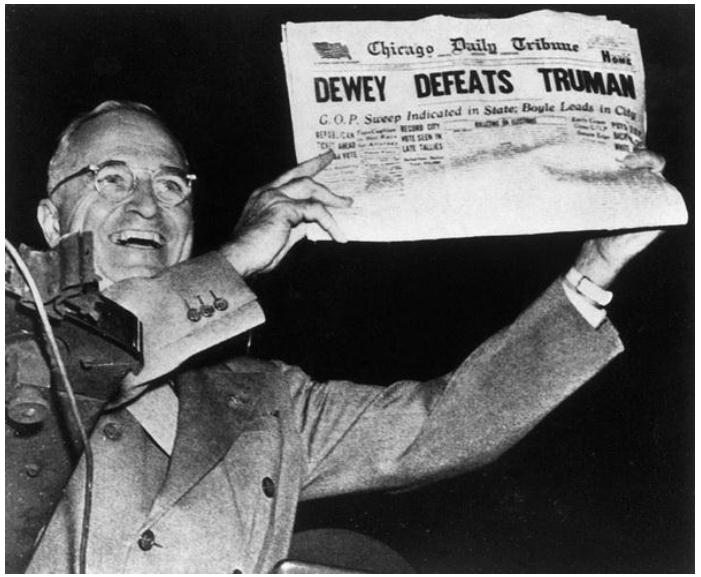
"Whoever designed this place is insane."

Workflow

- Define and design what to test
- Sample data
- Design the experiment
- Run experiment
- Collect data and analyze results
- Quality control

Survey design

- One of the most important parts
- Part art, part science
- Instructions are key
- Prepare to iterate



Questionnaire design

- Ask the right questions
- Workers may not be IR experts so don't assume the same understanding in terms of terminology
- Show examples
- Hire a technical writer
  - Engineer writes the specification
  - Writer communicates

UX design

- Time to apply all those usability concepts
- Generic tips
  - Experiment should be self-contained.
  - Keep it short and simple. Brief and concise.
  - Be very clear with the relevance task.
  - Engage with the worker. Avoid boring stuff.
  - Always ask for feedback (open-ended question) in an input box.

UX design - II

- Presentation
- Document design
- Highlight important concepts
- Colors and fonts
- Need to grab attention
- Localization

Examples - I

- Asking too much, task not clear, “do NOT/reject”
- Worker has to do a lot of stuff

Help us describe How-To Videos! Earn $2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the **Film & Theater** how-to video ==> 332492 Get a 35mm film look with a depth of field adapter
2. Write a description of the video linked in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete **25 video descriptions** in order to earn the $2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is **not necessary** to repeat the headline in your entry. It will **NOT** count toward your word count.
9. Do NOT describe the following: the format, where the video comes from, or how long the video is. This information is **IRRELEVANT**.
10. Do NOT describe the video in the following manner: "She turns around to face the camera. Then she faces left." Follow the examples below.

Current Word Count: 0 Current Character Count: 0 / 2000

Criteria for **REJECTION**:

1. Entries with obvious and multiple spelling or grammatical errors will be **rejected**.
2. Entries with fewer than 100 words will be automatically **rejected**.
3. Text copied from the web or other places will be **rejected**. Multiple plagiarized answers will lead to being **BLOCKED**. You may use a quotation, but the majority of your content must be **ORIGINAL**.
4. Incomplete and blank answers will be rejected. Multiple blank answers will result in being **blocked**.
5. Tasks submitted without descriptions will be **rejected**.
6. Tasks submitted with inaccurate descriptions will be **rejected** as well.
7. Do **NOT** add any personal opinions. Entries with personal opinions or reviews will be automatically **REJECTED**.
8. If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in **rejection**.
9. Entries that transcribe the video will be **REJECTED**.

Example - II

- Lot of work for a few cents
- Go here, go there, copy, enter, count …

Search for a topic and collect details about advertisers

Go to www.ezclout.com. In the Menu on the right side you will find the menu entry "Search" . Click on that Menu Entry which will take you to EZCLOUT's Search Page. Or go here. You must use the Search page provided on EZCLOUT's website or your reply will be rejected

Search for "mustang decal "

1. Copy the url of the search results here

2. Enter the url of the top placed advertiser

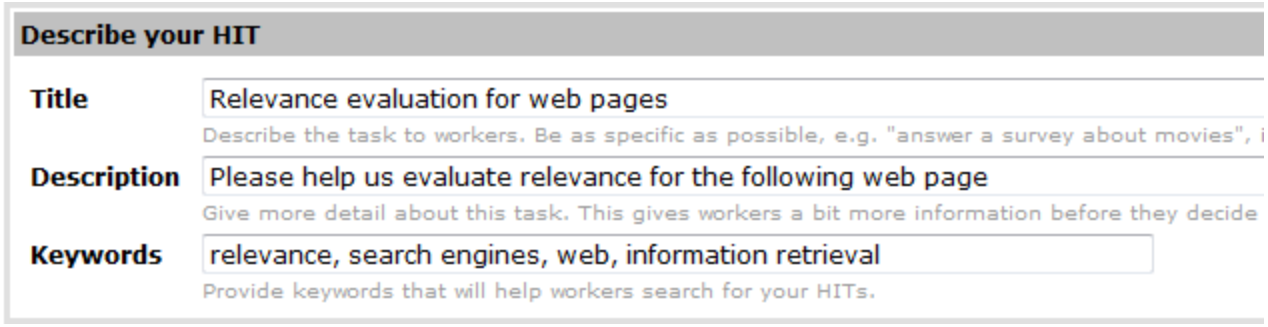
3. Count how many different advertisers are shown on the results page. Include all advertisers (don't forget advertisers at the bottom of the page) If results page does not show advertisers enter "no advertisers". We will verify every answer before we approve your reply.

A better example

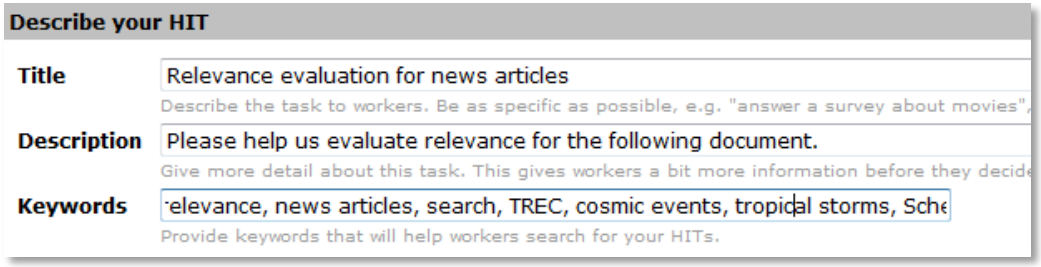
- All information is available
  - What to do
  - Search result
  - Question to answer

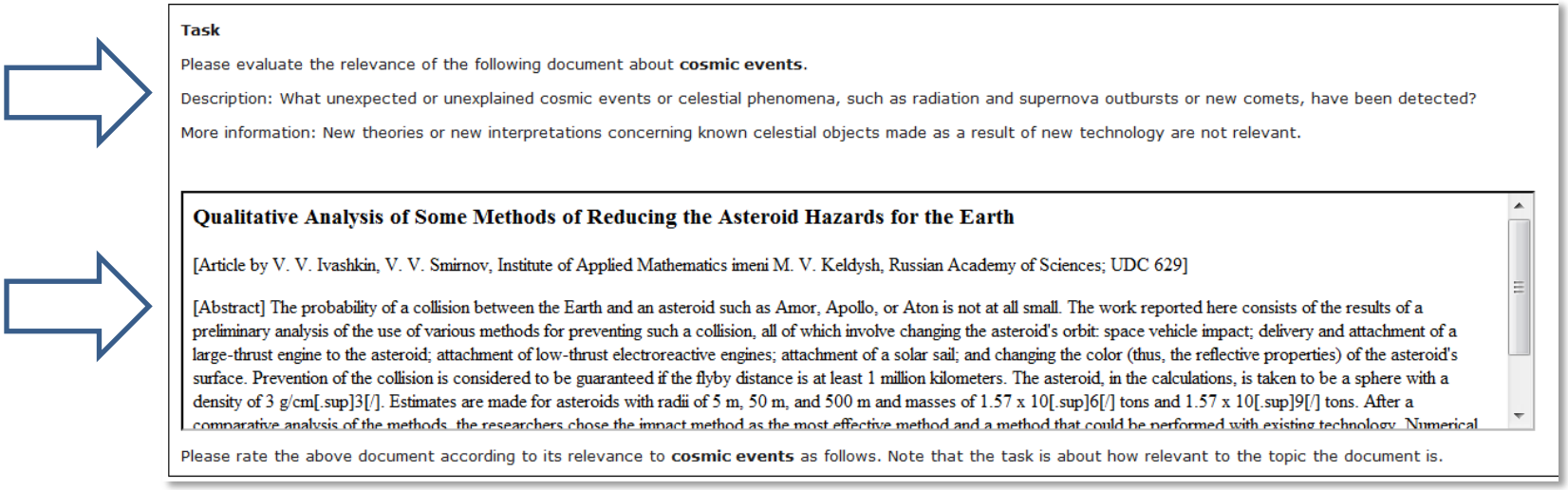
Form and metadata

- Form with a close question (binary relevance) and open-ended question (user feedback)
- Clear title, useful keywords
- Workers need to find your task

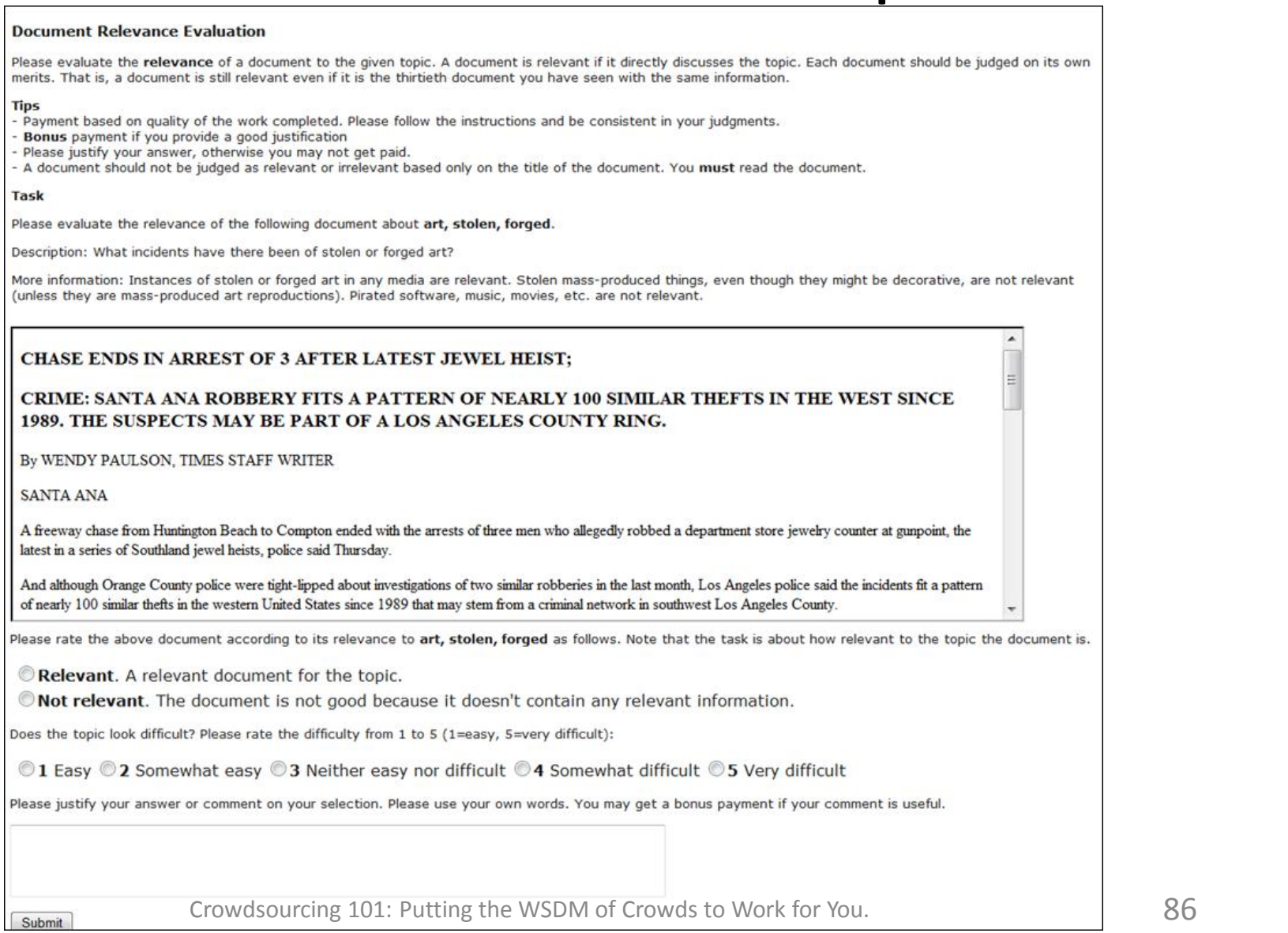


TREC assessment – example I





TREC assessment – example II



Document Relevance Evaluation

Please evaluate the **relevance** of a document to the given topic. A document is relevant if it directly discusses the topic. Each document should be judged on its own merits. That is, a document is still relevant even if it is the thirtieth document you have seen with the same information.

Tips
- Payment based on quality of the work completed. Please follow the instructions and be consistent in your judgments.
- **Bonus** payment if you provide a good justification
- Please justify your answer, otherwise you may not get paid.
- A document should not be judged as relevant or irrelevant based only on the title of the document. You **must** read the document.

Task

Please evaluate the relevance of the following document about **art, stolen, forged**.

Description: What incidents have there been of stolen or forged art?

More information: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass-produced art reproductions). Pirated software, music, movies, etc. are not relevant.

CHASE ENDS IN ARREST OF 3 AFTER LATEST JEWEL HEIST;

CRIME: SANTA ANA ROBBERY FITS A PATTERN OF NEARLY 100 SIMILAR THEFTS IN THE WEST SINCE 1989. THE SUSPECTS MAY BE PART OF A LOS ANGELES COUNTY RING.

By WENDY PAULSON, TIMES STAFF WRITER

SANTA ANA

A freeway chase from Huntington Beach to Compton ended with the arrests of three men who allegedly robbed a department store jewelry counter at gunpoint, the latest in a series of Southland jewel heists, police said Thursday.

And although Orange County police were tight-lipped about investigations of two similar robberies in the last month, Los Angeles police said the incidents fit a pattern of nearly 100 similar thefts in the western United States since 1989 that may stem from a criminal network in southwest Los Angeles County.

Please rate the above document according to its relevance to **art, stolen, forged** as follows. Note that the task is about how relevant to the topic the document is.

- **Relevant**. A relevant document for the topic.
- **Not relevant**. The document is not good because it doesn't contain any relevant information.

Does the topic look difficult? Please rate the difficulty from 1 to 5 (1=easy, 5=very difficult):

1 Easy **2** Somewhat easy **3** Neither easy nor difficult **4** Somewhat difficult **5** Very difficult

Please justify your answer or comment on your selection. Please use your own words. You may get a bonus payment if your comment is useful.

Submit

How much to pay?

- Price commensurate with task effort
  - Ex: $0.02 for yes/no answer + $0.02 bonus for optional feedback
- Ethics & market-factors: W. Mason and S. Suri, 2010.
  - e.g. non-profit SamaSource contracts workers refugee camps
  - Predict right price given market & task: Wang et al. CSDM'11
- Uptake & time-to-completion vs. Cost & Quality
  - Too little $$, no interest or slow – too much $$, attract spammers
  - Real problem is lack of reliable QA substrate
- Accuracy & quantity
  - More pay = more work, not better (W. Mason and D. Watts, 2009)
- Heuristics: start small, watch uptake and bargaining feedback
- Worker retention ("anchoring")

See also: L.B. Chilton et al. KDD-HCOMP 2010.

Implementation

- Similar to a UX
- Build a mock up and test it with your team
  - Yes, you need to judge some tasks
- Incorporate feedback and run a test on AMT with a very small data set
  - Time the experiment
  - Do people understand the task?
- Analyze results
  - Look for spammers
  - Check completion times
- Iterate and modify accordingly

Implementation – II

- Introduce quality control
  - Qualification test
  - Gold answers (honey pots)
- Adjust passing grade and worker approval rate
- Run experiment with new settings and same data set
- Scale on data
- Scale on workers

Experiment in production

- Lots of tasks on AMT at any moment
- Need to grab attention
- Importance of experiment metadata
- When to schedule
  - Split a large task into batches and have 1 single batch in the system
  - Always review feedback from batch *n* before uploading *n*+1

Quality control

- Extremely important part of the experiment
- Approach it as “overall” quality – not just for workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester.

Quality control - II

- Approval rate: easy to use, & just as easily defeated
  - P. Ipeirotis. Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes. Oct. 2010
- Qualification test
  - Pre-screen workers' ability to do the task (accurately)
  - Example and pros/cons in next slides
- Assess worker quality as you go
  - Trap questions with known answers ("honey pots")
  - Measure inner-annotator agreement between workers
- No guarantees

A qualification test snippet

```
<Question>
  <QuestionIdentifier>question1</QuestionIdentifier>
  <QuestionContent>
    <Text>Carbon monoxide poisoning is</Text>
  </QuestionContent>
  <AnswerSpecification>
    <SelectionAnswer>
       <StyleSuggestion>radiobutton</StyleSuggestion>
        <Selections>
         <Selection>
           <SelectionIdentifier>1</SelectionIdentifier>
           <Text>A chemical technique</Text>
         </Selection>
         <Selection>
           <SelectionIdentifier>2</SelectionIdentifier>
           <Text>A green energy treatment</Text>
         </Selection>
         <Selection>
            <SelectionIdentifier>3</SelectionIdentifier>
            <Text>A phenomena associated with sports</Text>
         </Selection>
         <Selection>
            <SelectionIdentifier>4</SelectionIdentifier>
            <Text>None of the above</Text>
         </Selection>
        </Selections>
    </SelectionAnswer>
  </AnswerSpecification>
</Question>
```

Qualification tests: pros and cons

- Advantages
  - Great tool for controlling quality
  - Adjust passing grade
- Disadvantages
  - Extra cost to design and implement the test
  - May turn off workers, hurt completion time
  - Refresh the test on a regular basis
  - Hard to verify subjective tasks like judging relevance
- Try creating task-related questions to get worker familiar with task *before* starting task in earnest

Methods for measuring agreement

- What to look for
  - Agreement, reliability, validity
- Inter-agreement level
  - Agreement between judges
  - Agreement between judges and the gold set
- Some statistics
  - Cohen's kappa (2 raters)
  - Fleiss' kappa (any number of raters)
  - Krippendorff's alpha
- With simple majority vote, what if 2 say relevant, 3 say not?
  - Use expert to break ties
  - Let target confidence threshold drive number of judgment per example; dynamically collect more judgments to reduce uncertainty
- 2-tier approach: Group 1 does task, Group 2 verifies
  - Quinn and B. Bederson'09, Bernstein et al.'10
- Better consensus methods exist…

Inter-rater reliability

- Lots of research
- Statistics books cover most of the material
- Three categories based on the goals
  - Consensus estimates
  - Consistency estimates
  - Measurement estimates

"Plus, we double the accuracy at no extra cost by using our extensive pool of Siamese twins."

Quality Control & Assurance

- Filtering
  - Approval rate (built-in but defeatable)
  - Geographic restrictions (e.g. US only, built-in)
  - Worker blocking
  - Qualification test
    - Con: slows down experiment, difficult to "test" relevance
    - Solution: create questions to let user get familiar *before* the assessment
  - Does not guarantee success
- Assessing quality
  - Interject verifiable/gold answers (trap questions, honey pots)
    - P. Ipeitotis. Worker Evaluation in Crowdsourcing: Gold Data or Multiple Workers? Sept. 2010.
- Identify workers that *always* disagree with the majority
  - Risk: masking cases of ambiguity or diversity, "tail" behaviors

More on quality control & assurance

- HR issues: recruiting, selection, & retention
  - e.g., post/tweet, design a better qualification test, bonuses, …
- Collect more redundant judgments…
  - at some point defeats cost savings of crowdsourcing
  - 5 workers is often sufficient
- Use better aggregation method
  - Voting
  - Consensus
  - Averaging

Scales and labels

- Binary
  - Yes, No
- 5-point Likert
  - Strongly disagree, disagree, neutral, agree, strongly agree
- Graded relevance:
  - DCG: Irrelevant, marginally, fairly, highly (Jarvelin, 2000)
  - TREC: Highly relevant, relevant, (related), not relevant
  - Yahoo/MS: Perfect, excellent, good, fair, bad (PEGFB)
  - The Google Quality Raters Handbook (March 2008)
  - 0 to 10 (0 = totally irrelevant, 10 = most relevant)
- Usability factors
  - Provide clear, concise labels that use plain language
  - Avoid unfamiliar jargon and terminology

Was the task difficult?

- Ask turkers to rate the difficulty of a topic
- 50 topics, TREC; 5 workers, $0.01 per task

Exp	Title	Average
412	airport security	1.4
439	inventions, scientific discoveries	1.6
438	tourism, increase	1.6
428	declining birth rates	1.6
424	suicides	1.6
442	heroic acts	1.8
436	railway accidents	1.8
433	Greek, philosophy, stoicism	3.6
405	cosmic events	3.6
401	foreign minorities, Germany	3.6

Other quality heuristics

- Justification/feedback as quasi-captcha
  - Successfully used at TREC and INEX experiments
  - Should be optional
  - Automatically verifying feedback was written by a person may be difficult (classic spam detection task)
- Broken URL/incorrect object
  - Leave an outlier in the data set
  - Workers will tell you
  - If somebody answers “excellent” on a graded relevance test for a broken URL => *probably* spammer

MTurk QA: Tools and Packages

- QA infrastructure layers atop MTurk promote useful separation-of-concerns from task
  - TurkIt
    - Quik Turkit provides nearly realtime services
  - Turkit-online (??)
  - Get Another Label (& qmturk)
  - Turk Surveyor
  - cv-web-annotation-toolkit (image labeling)
  - Soylent
  - Boto (python library)
    - Turkpipe: submit batches of jobs using the command line.
- More needed…

Dealing with bad workers

- Pay for “bad” work instead of rejecting it?
  - Pro: preserve reputation, admit if poor design at fault
  - Con: promote fraud, undermine approval rating system
- Use bonus as incentive
  - Pay the minimum $0.01 and $0.01 for bonus
  - Better than rejecting a $0.02 task
- If spammer “caught”, block from future tasks

Worker feedback

- Real feedback received via email after **rejection**
- Worker XXX

  I did. If you read these articles most of them have nothing to do with space programs. I'm not an idiot.

- Worker XXX

  As far as I remember there wasn't an explanation about what to do when there is no name in the text. I believe I did write a few comments on that, too. So I think you're being unfair rejecting my HITs.

Real email exchange with worker after **rejection**

WORKER: this is not fair , you made me work for 10 cents and i lost my 30 minutes of time ,power and lot more and gave me 2 rejections atleast you may keep it pending. please show some respect to turkers

REQUESTER: *I'm sorry about the rejection. However, in the directions given in the hit, we have the following instructions: IN ORDER TO GET PAID, you must judge all 5 webpages below *AND* complete a minimum of three HITs.*

Unfortunately, because you only completed two hits, we had to reject those hits. We do this because we need a certain amount of data on which to make decisions about judgment quality. I'm sorry if this caused any distress. Feel free to contact me if you have any additional questions or concerns.

WORKER: I understood the problems. At that time my kid was crying and i went to look after. that's why i responded like that. I was very much worried about a hit being rejected. The real fact is that i haven't seen that instructions of 5 web page and started doing as i do the dolores labs hit, then someone called me and i went to attend that call. sorry for that and thanks for your kind concern.

Exchange with worker

- Worker XXX

  Thank you. I will post positive feedback for you at Turker Nation.

Me: was this a sarcastic comment?

- I took a chance by accepting some of your HITs to see if you were a trustworthy author. My experience with you has been favorable so I will put in a good word for you on that website. This will help you get higher quality applicants in the future, which will provide higher quality work, which might be worth more to you, which hopefully means higher HIT amounts in the future.

Build Your Reputation as a Requestor

- Word of mouth effect
  - Workers trust the requester (pay on time, clear explanation if there is a rejection)
  - Experiments tend to go faster
  - Announce forthcoming tasks (e.g. tweet)
- Disclose your real identity?

Other practical tips

- Sign up as worker and do some HITs
- “Eat your own dog food”
- Monitor discussion forums
- Address feedback (e.g., poor guidelines, payments, passing grade, etc.)
- Everything counts!
  - Overall design only as strong as weakest link

MTurk Worker Forums & Resources

- Turker Nation: http://turkers.proboards.com
- http://www.turkalert.com (and its blog)
- Turkopticon: report/avoid shady requestors
- Amazon Forum for MTurk

Content quality

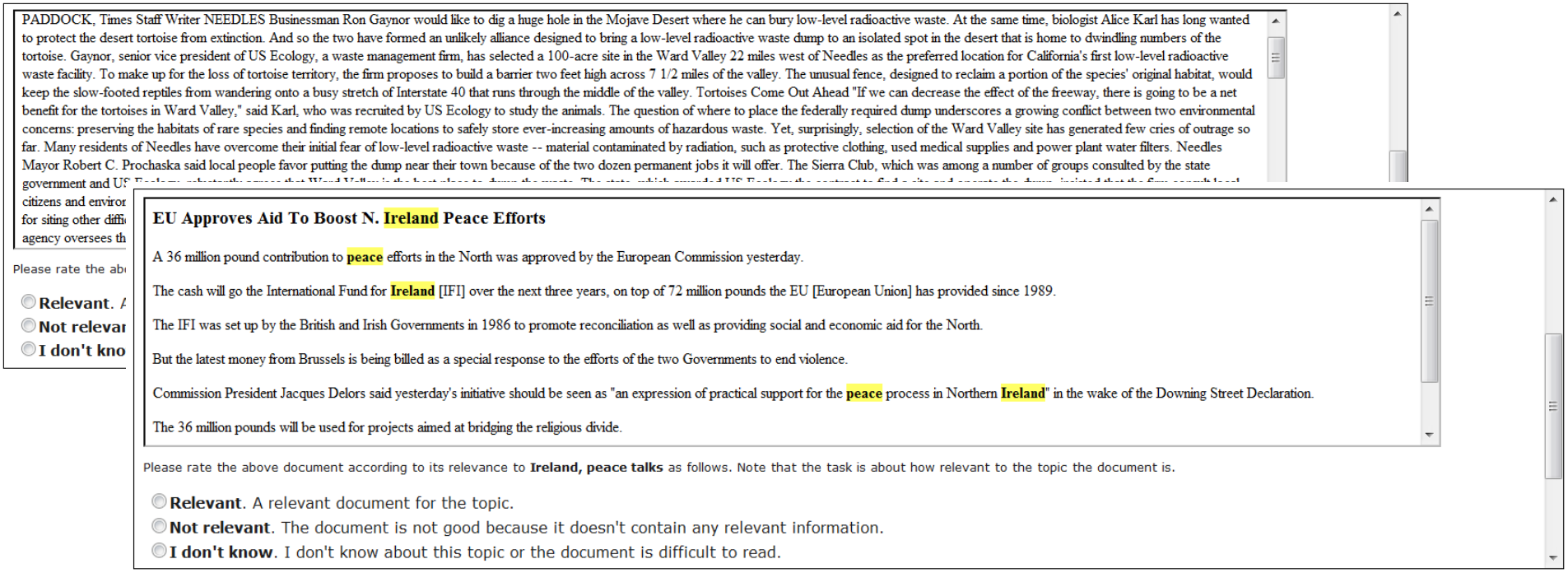
- People like to work on things that they like
- TREC ad-hoc vs. INEX
  - TREC experiments took twice to complete
  - INEX (Wikipedia), TREC (LA Times, FBIS)
- Topics
  - INEX: Olympic games, movies, salad recipes, etc.
  - TREC: cosmic events, Schengen agreement, etc.
- Content and judgments according to modern times
  - Airport security docs are pre 9/11
  - Antarctic exploration (global warming)

Content quality - II

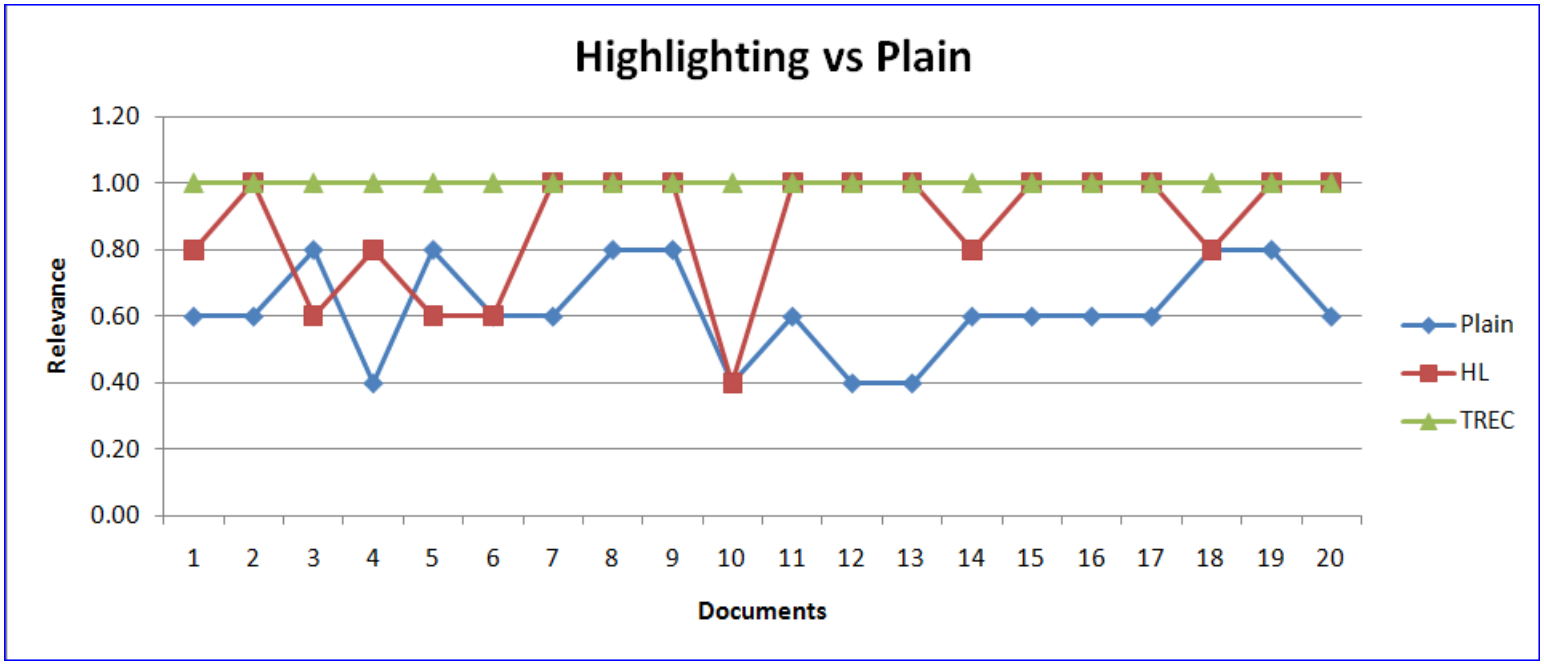
- Document length
- Randomize content
- Avoid worker fatigue
  - Judging 100 documents on the same subject can be tiring

Presentation

- People scan documents for relevance cues
- Document design
- Highlighting no more than 10%



Presentation - II



Relevance justification

- Why settle for a label?
- Let workers justify answers
- INEX
  - 22% of assignments with comments
- Must be optional
- Let's see how people justify

“Relevant” answers

[Salad Recipes]

Doesn't mention the word 'salad', but the recipe is one that could be considered a salad, or a salad topping, or a sandwich spread.
Egg salad recipe
Egg salad recipe is discussed.
History of salad cream is discussed.
Includes salad recipe
It has information about salad recipes.
Potato Salad
Potato salad recipes are listed.
Recipe for a salad dressing.
Salad Recipes are discussed.
Salad cream is discussed.
Salad info and recipe
The article contains a salad recipe.
The article discusses methods of making potato salad.
The recipe is for a dressing for a salad, so the information is somewhat narrow for the topic but is still potentially relevant for a researcher.
This article describes a specific salad. Although it does not list a specific recipe, it does contain information relevant to the search topic.
gives a recipe for tuna salad
relevant for tuna salad recipes
relevant to salad recipes
this is on-topic for salad recipes

“Not relevant” answers

[Salad Recipes]

```
About gaming not salad recipes.
Article is about Norway.
Article is about Region Codes.
Article is about forests.
Article is about geography.
Document is about forest and trees.
Has nothing to do with salad or recipes.
Not a salad recipe
Not about recipes
Not about salad recipes
There is no recipe, just a comment on how salads fit into meal formats.
There is nothing mentioned about salads.
While dressings should be mentioned with salads, this is an article on one specific
    type of dressing, no recipe for salads.
article about a swiss tv show
completely off-topic for salad recipes
not a salad recipe
not about salad recipes
totally off base
```

"All I know is that searching for my own name and then clicking on 'highly relevant' does wonders for my self-esteem."

Other design principles

- Text alignment
- Legibility
- Reading level: complexity of words and sentences
- Attractiveness (worker's attention & enjoyment)
- Multi-cultural / multi-lingual
- Who is the audience (e.g. target worker community)
  - Special needs communities (e.g. simple color blindness)
- Parsimony
- Cognitive load: mental rigor needed to perform task
- Exposure effect

Platform alternatives

- Do I have to use AMT?
- How to build your own crowdsourcing platform
  - Back-end
  - Template language for creating experiments
  - Scheduler
  - Payments?

MapReduce with human computation

- MapReduce meets crowdsourcing
- Commonalities
  - Large task divided into smaller sub-problems
  - Work distributed among worker nodes (turkers)
  - Collect all answers and combine them
  - Varying performance of heterogenous CPUs/HPUs
- Variations
  - Human response latency / size of “cluster”
  - Some tasks are not suitable

Challenges and opportunities

- A back-end perspective
- Problems with the current platform
  - Very rudimentary
  - No tools for data analysis
  - No integration with databases
  - Very limited search and browse features
- Opportunities
  - What is the database model for crowdsourcing?
  - MapReduce with crowdsourcing
  - Can you integrate human-computation into a language?
    - `crowdsource(task,5)`

Research problems – operational

- Methodology
  - Budget, people, document, queries, presentation, incentives, etc.
  - Scheduling
  - Quality
- What's the best "mix" of HC for a task?
- What are the tasks suitable for HC?
- Can I crowdsource my task?
  - Eickhoff and de Vries, WSDM 2011 CSDM Workshop

More problems

- Human factors vs. outcomes
- Editors vs. workers
- Pricing tasks
- Predicting worker quality from observable properties (e.g. task completion time)
- HIT / Requestor ranking or recommendation
- Expert search : who are the right workers given task nature and constraints
- Ensemble methods for Crowd Wisdom consensus

Problems – crowds, clouds and algorithms

- Infrastructure
  - Current platforms are very rudimentary
  - No tools for data analysis
- Dealing with uncertainty (propagate rather than mask)
  - Temporal and labeling uncertainty
  - Learning algorithms
  - Search evaluation
  - Active learning (which example is likely to be labeled correctly)
- Combining CPU + HPU
  - MapReduce with human computation?
  - Human Remote Call?
  - Procedural vs. declarative?
  - Integration points with enterprise systems

Conclusions

- Crowdsourcing for relevance evaluation works
- Fast turnaround, easy to experiment, few dollars to test
- But you have to design the experiments carefully
- Usability considerations
- Worker quality
- User feedback extremely useful

Conclusions - II

- Crowdsourcing is here to stay
- Lots of opportunities to improve current platforms
- Integration with current systems
- AMT is a popular platform and others are emerging
- Open research problems

Recent Events (2010 was big!)

- 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 Mechanical Turk. NAACL 2010
  - Maryland Workshop on Crowdsourcing and Translation. June, 2010
- ML: Computational Social Science and Wisdom of Crowds. NIPS 2010
- Advancing Computer Vision with Humans in the Loop at CVPR 2010

- Conference: CrowdConf 2010 (organized by CrowdFlower)

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

2011 book: Omar Alonso, Gabriella Kazai, and Stefano Mizzaro. Crowdsourcing for Search Engine Evaluation: Why and How.

Special issue of Information Retrieval journal on Crowdsourcing (papers due May 6, 2011)

Upcoming Conferences & Workshops

- AAAI-HCOMP (papers due April 22, 2011)
- CHI 2011 Workshop (May 8)
- CrowdConf 2011 (TBA)
- SIGIR 2011 workshop? (in review)
- TREC 2011 Crowdsourcing Track

Thank You!

For questions about tutorial or crowdsourcing, email:

omar.alonso@microsoft.com

ml@ischool.utexas.edu

Cartoons by Mateo Burtch (buta@mindspring.com)

bing

THE UNIVERSITY OF TEXAS AT AUSTIN

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 Amazon 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.

Bibliography - IV

- P.G. Ipeirotis. The New Demographics of Mechanical Turk. March 9, 2010. PDF and Spreadsheet.
- P.G. Ipeirotis, R. Chandrasekar and P. Bennett. A report on the human computation workshop. ACM SIGKDD Explorations Newsletter v. 11 no . 2 pp. 80-83, 2010.
- P.G. Ipeirotis. **Analyzing the Amazon Mechanical Turk Marketplace. CeDER-10-04 (Sept. 11, 2010)**
- K. Jarvelin, and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. Proceedings of the 23rd annual international ACM SIGIR conference . pp.41—48, 2000.

Blogs

- Behind Enemy Lines (P.G. Ipeirotis, NYU)
- Deneme: a Mechanical Turk experiments blog (Gret Little, MIT)
- CrowdFlower Blog
- http://experimentalturk.wordpress.com
- Jeff Howe

Sites

- The Crowdsortium
- Crowdsourcing.org
- Daily Crowdsource