Crowdsourcing for NLP

Matt Lease
School of Information
University of Texas at Austin
ml@ischool.utexas.edu

THE UNIVERSITY OF TEXAS
SCHOOL OF
information

See also: O. Alonso and M. Lease, WSDM 2011 Tutorial

Jane saw the man with the binoculars

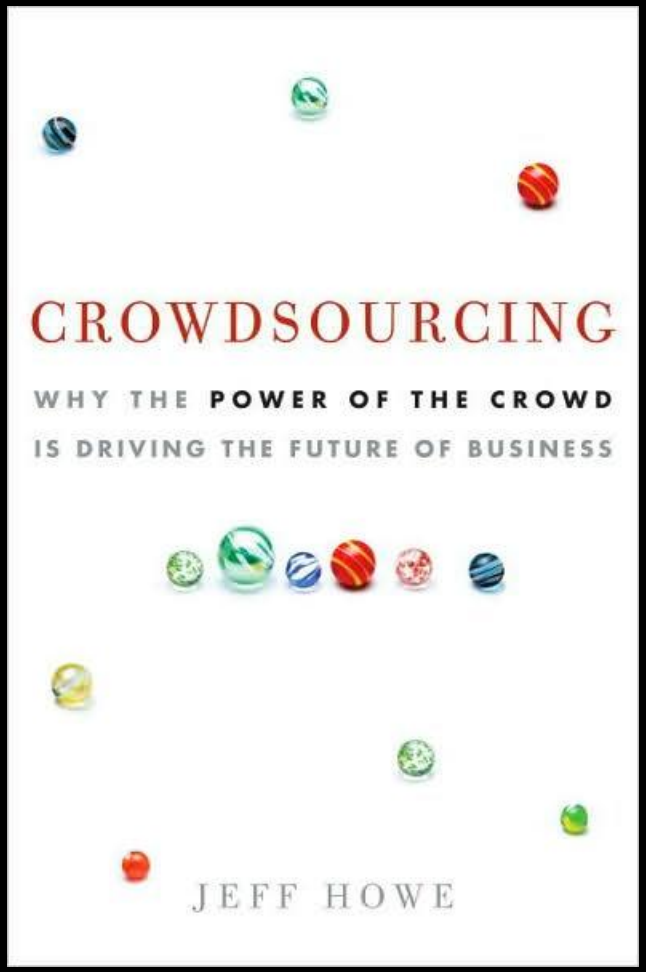
Traditional Annotation / Data Collection

- Setup data collection software / harness
- Recruit volunteers (often undergrads)
- Pay a flat fee for experiment or hourly wage

- Characteristics
  - Slow
  - Expensive
  - Tedious
  - Sample Bias

Crowdsourcing

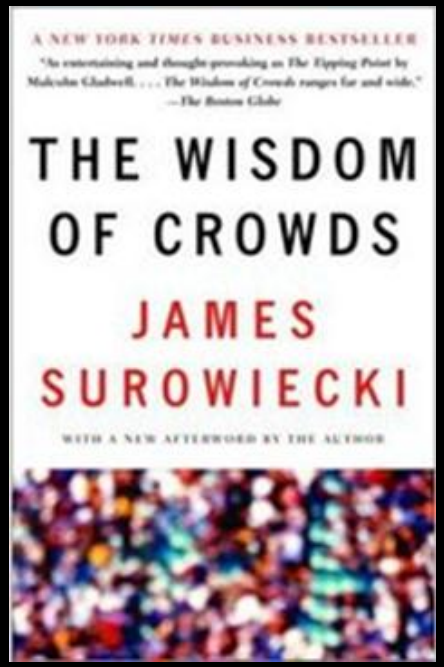
- Take a job traditionally performed by a known agent (often an employee)
- Outsource it to an undefined, generally large group of people via an open call
- New application of principles from open source movement



Wisdom of Crowds

Requires

- Diversity
- Independence
- Decentralization
- Aggregation

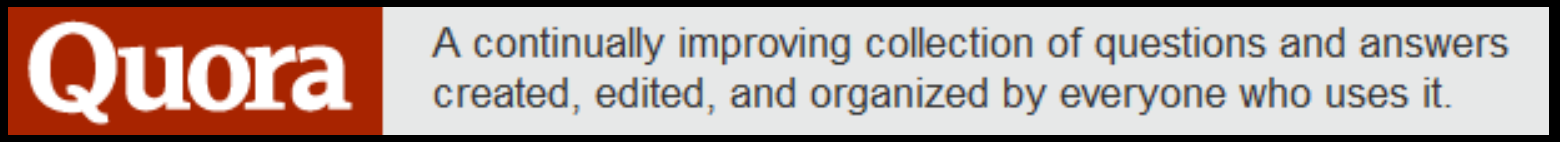


Input: large, diverse sample

(increases likelihood of overall pool quality)

Output: consensus, selection, distribution

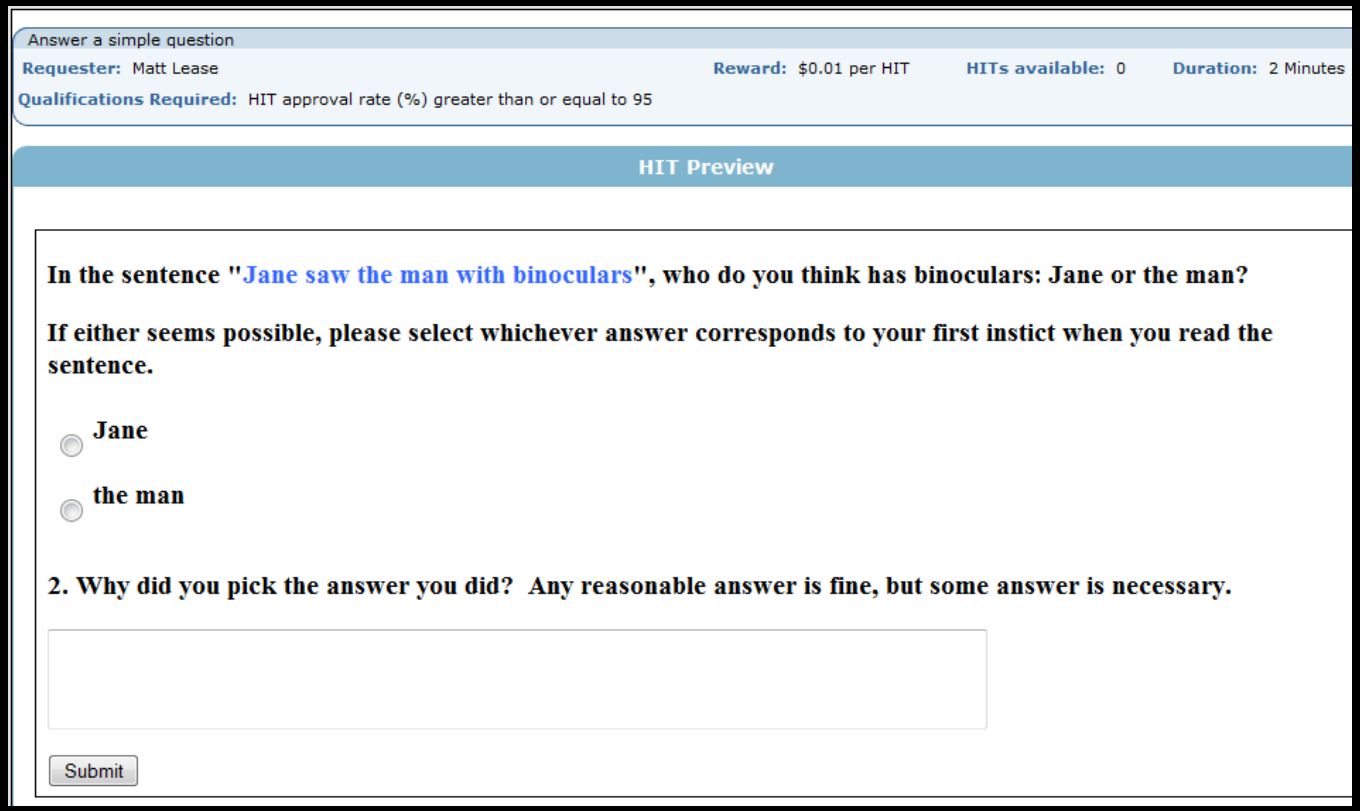
Community Q&A / Social Search / Public Polling



Amazon Mechanical Turk (AMT, Mturk)

- Crowdsourcing platform / marketplace
- On-demand workforce (latency near-real time)
- Went online in 2005
- Progammer's API & "Dashboard" GUI

Human Intelligence Task (HIT)



- Specify # of “assignments” (workers) desired
- Set pay for each assignment (+ 10% surcharge)

Road Map

- Example: annotate PP-attachment
- Crowdsourcing & Human Computation
- Crowdsourcing models and worker incentives
- What's it good for? – Examples
- Crowdsourcing Human Subjects Research
- Quality Control
- Trends, Challenges, and Opportunities

- Bonus: Who are the workers? – Demographics

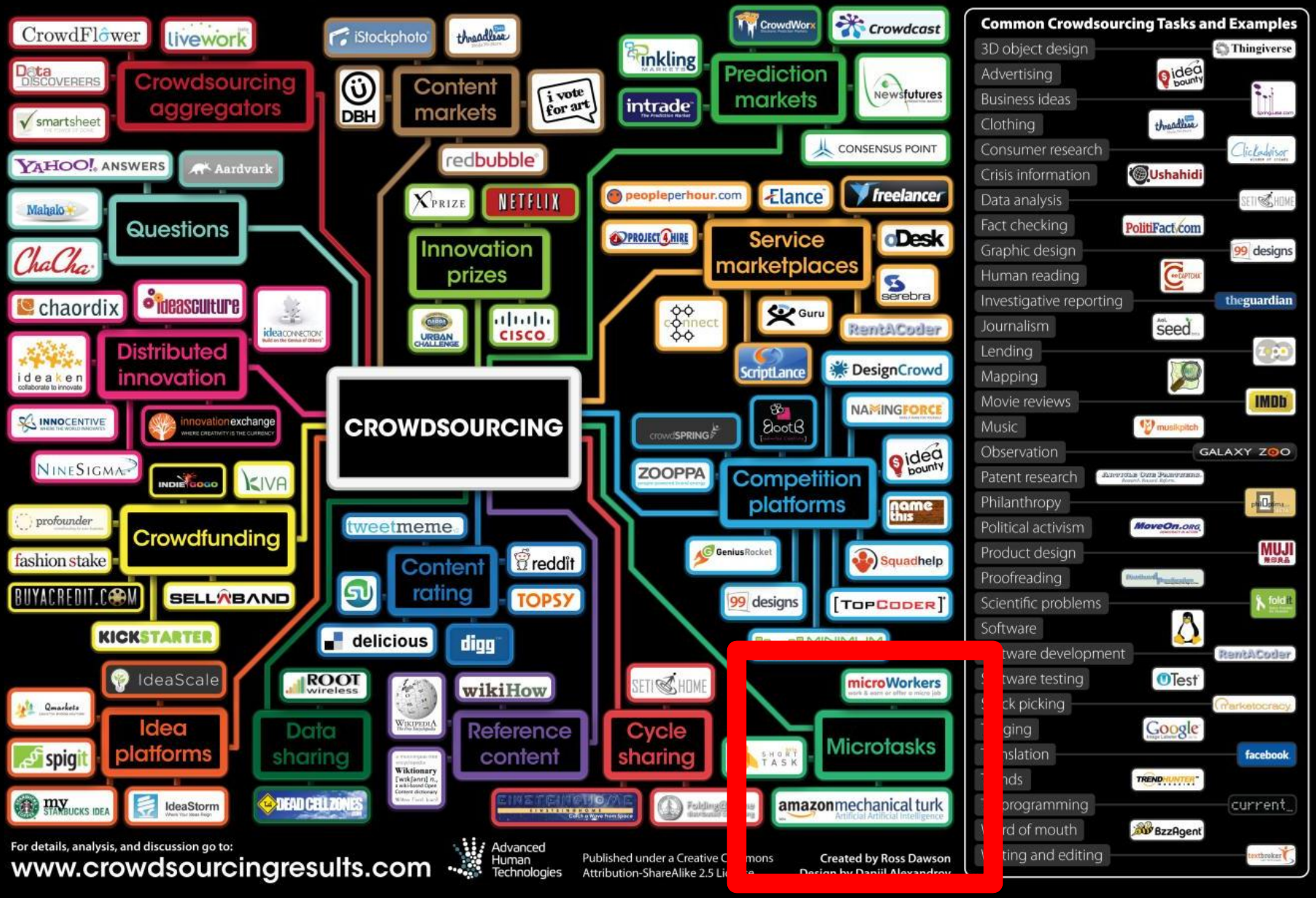
Crowdsourcing & Human Computation

Other Crowdsourcing Examples

INNOCENTIVE®

XPRIZE FOUNDATION

NETFLIX®


CROWDSOURCING LANDSCAPE
Beta v1
CROWDSOURCING
Crowdsourcing aggregators
Content markets
Prediction markets
Questions
Distributed innovation
Innovation prizes
Service marketplaces
Crowdfunding
Competition platforms
Content rating
Idea platforms
Data sharing
Reference content
Cycle sharing
Microtasks
Common Crowdsourcing Tasks and Examples
3D object design
Advertising
Business ideas
Clothing
Consumer research
Crisis information
Data analysis
Fact checking
Graphic design
Human reading
Investigative reporting
Journalism
Lending
Mapping
Movie reviews
Music
Observation
Patent research
Philanthropy
Political activism
Product design
Proofreading
Scientific problems
Software
Thingiverse
theguardian
GALAXY ZOO
IMDb
current_
For details, analysis, and discussion go to:
www.crowdsourcingresults.com
Advanced Human Technologies
Created by Ross Dawson

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?

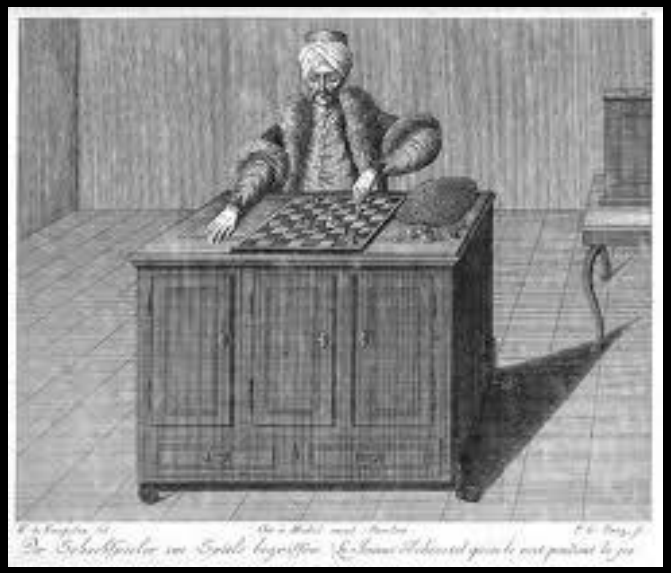
- General graphs - listing of projects, total number of available hits and rewards for every hour
- Top 1000 requesters - sorted according to cumulative rewards
- Top 1000 requesters - sorted according to cumulative rewards
- Arrivals - number of new projects, available hits and rewards computed on daily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

Mechanical Turk Monitor: General Data

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)

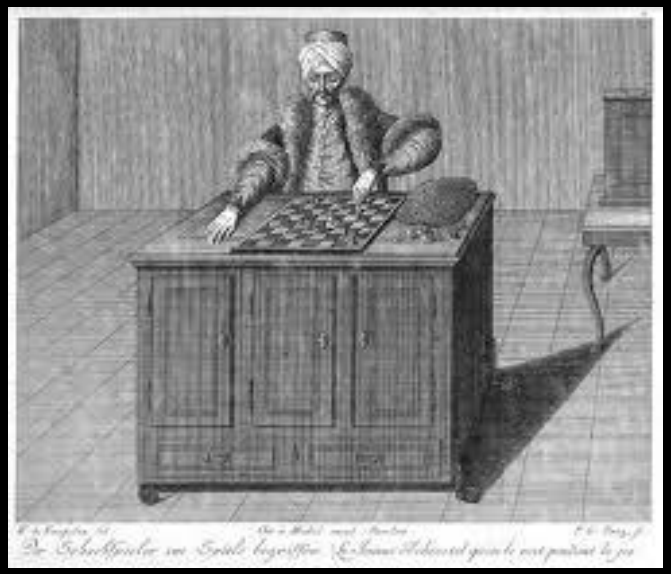
The Mechanical Turk

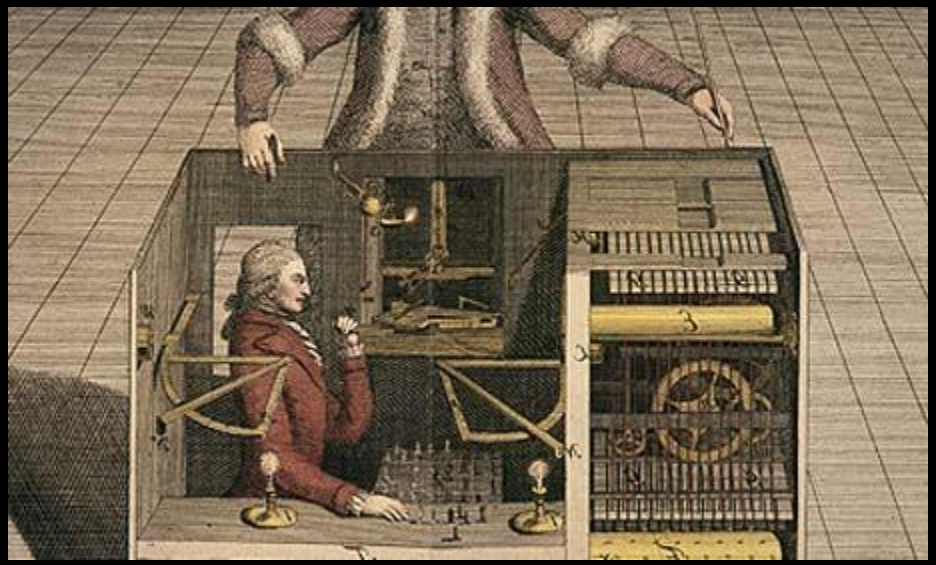


The original, constructed and unveiled in 1770 by Wolfgang von Kempelen (1734–1804)

Gavin Turk, "The Mechanical Turk" (2008)

Artificial Artificial Intelligence

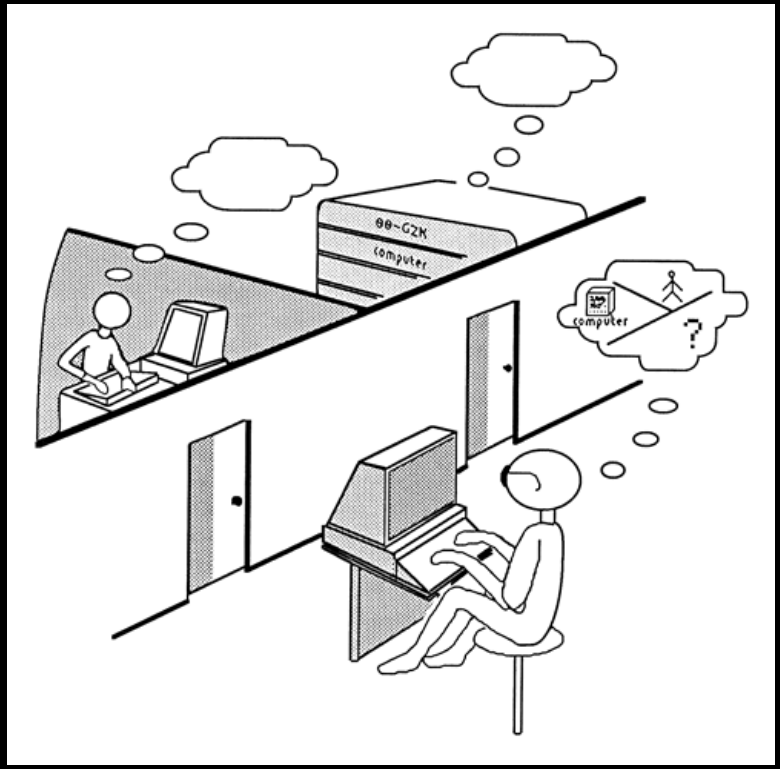




amazon mechanical turk™
Artificial Artificial Intelligence

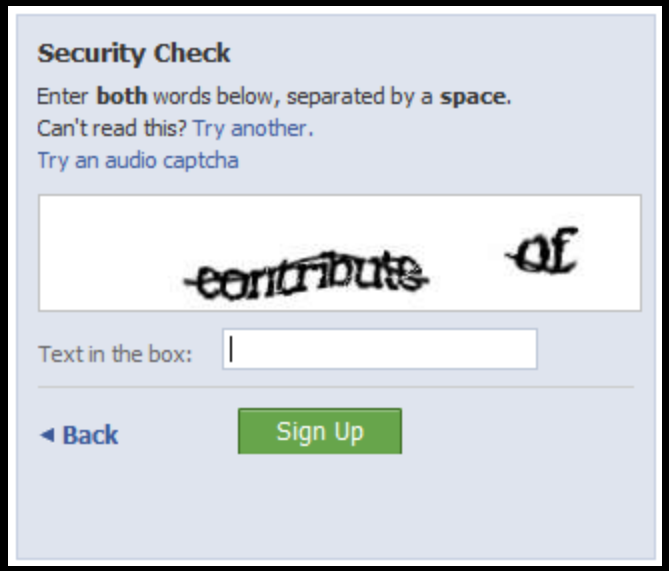
J. Pontin. *Artificial Intelligence, With Help From the Humans*. NY Times (March 25, 2007)

The Turing Test (Alan Turing, 1950)



IBM
GARRY KASPAROV
DEEP BLUE

Jenn
WATSON
Miles


Security Check
Enter both words below, separated by a space.
Can't read this? Try another.
Try an audio captcha
contribute
of
Text in the box:
Back
Sign Up

PREMIER PHARMACY
Lowest VIAGRA, CIALIS, LEVITRA Online Price!
VIAGRA 30 $134.95
CIALIS 30 $169.95
VALIUM 30 $85.45
SOMA 30 $75.95
PROPECIA 30 $64.95
AMBIEN 30 $120.99
XANAX 30 $123.45
VIAGRA SOFT 50 $250.99
New CIALIS SOFT 30 $224.95
Save up to 80% on your prescription Meds!
5RX.ORG
Do not click, type in your browser www.5rx.org

What is a Computer?

com·put·er

[*kuh* m-**pyoo**-ter] Show IPA

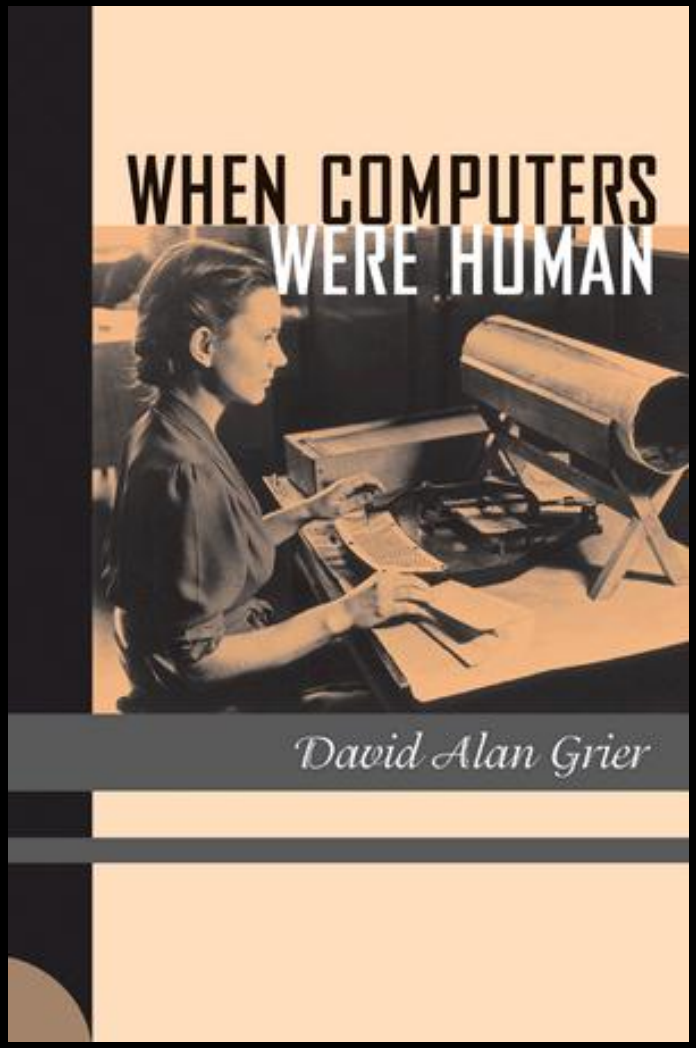
–noun

1. Also called **processor.** an electronic device designed to accept data, perform prescribed mathematical and logical operations at high speed, and display the results of these operations. Compare analog computer, digital computer.

Origin:
1640–50; compute + -er[1]; compare Middle French *computeur*

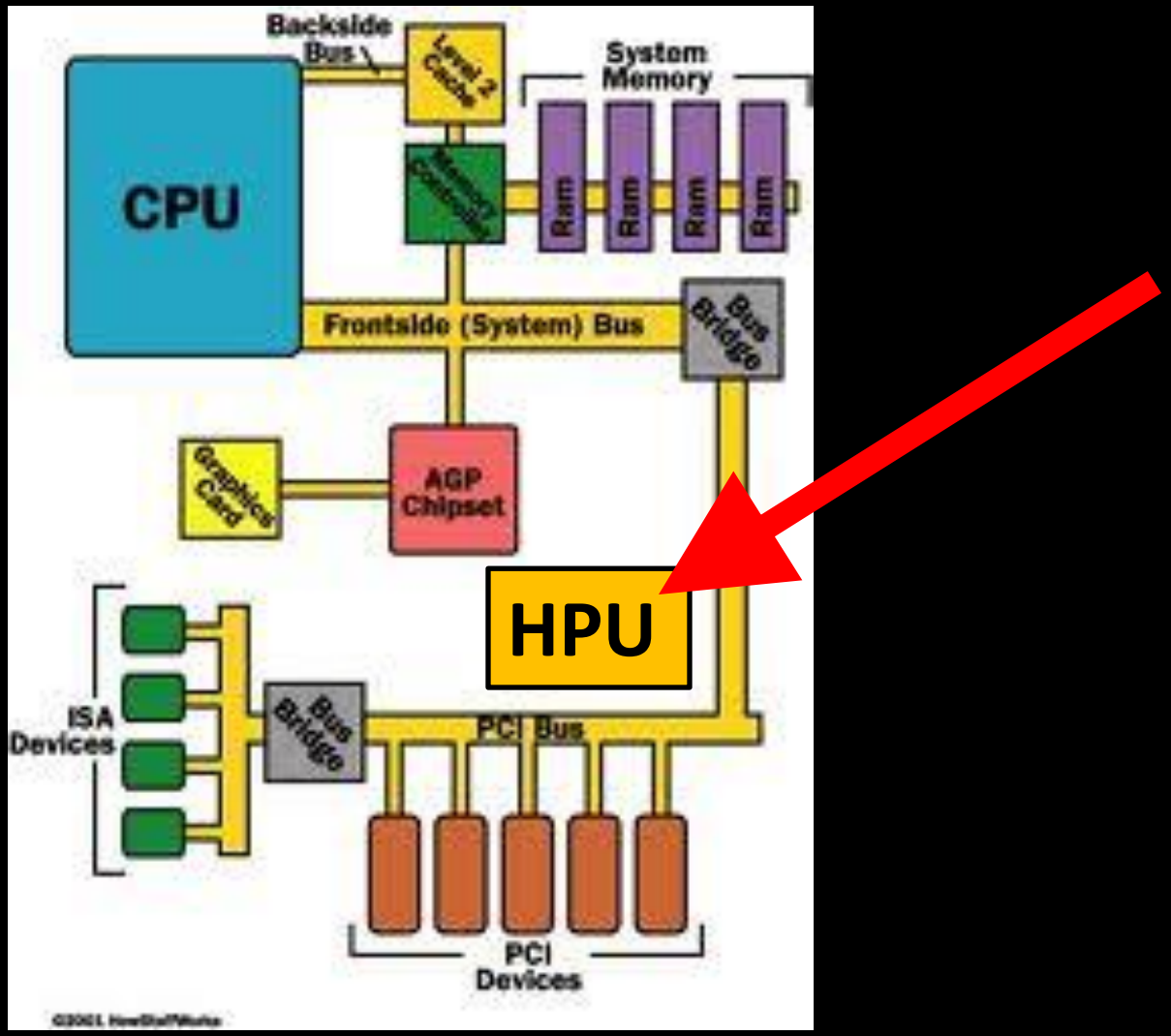
—Related forms
com·put·er·like, *adjective*
non·com·put·er, *adjective*

Dictionary.com Unabridged
Based on the Random House Dictionary, © Random House, Inc. 2011.

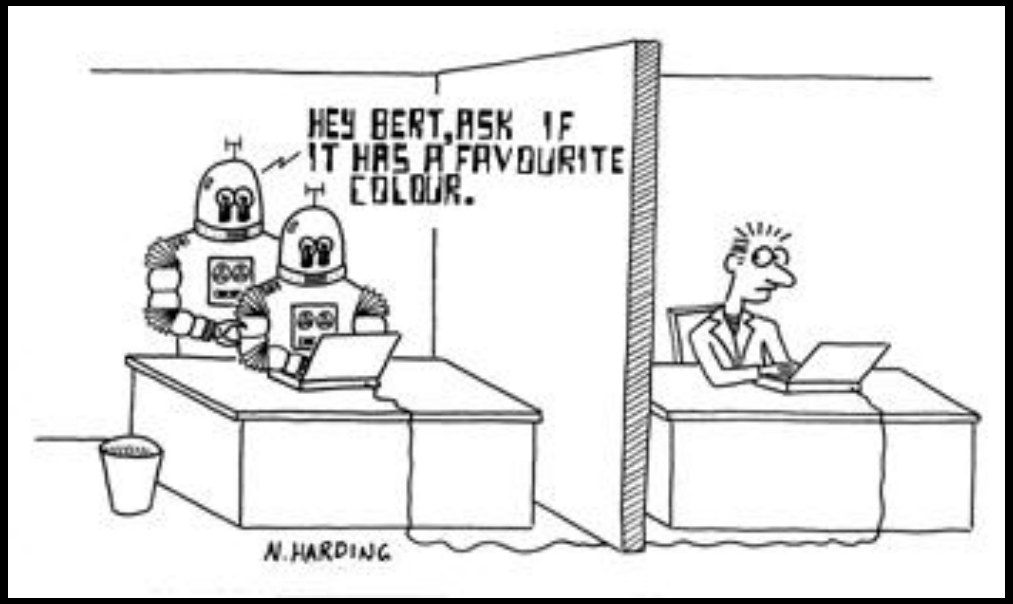


Princeton University Press, 2005

Davis et al. (2010) *The HPU*.



The Turing Test (Alan Turing, 1950)



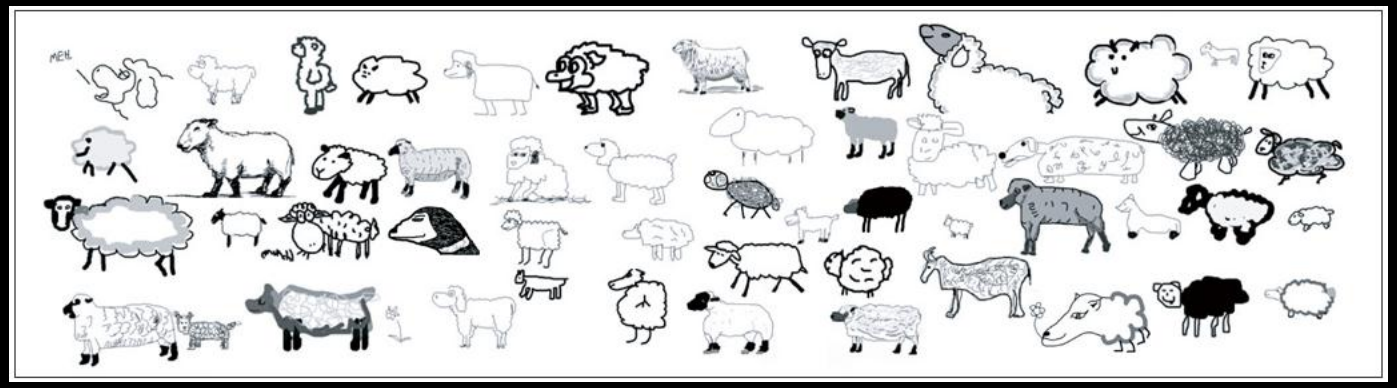
Human Computation

- People become ‘computists’ once more
  - Humans do tasks computers cannot (do well)
  - System makes opaque “external call” to the “HPU”
- Block robots (Captcha – “reverse Turing test”)
- Collect data
  - typically “stupid parallelism”, minimal post-processing
- Deliver new functionality
  - Blend CPU + HPU computation , HPU part of core architecture

Data Collection Examples

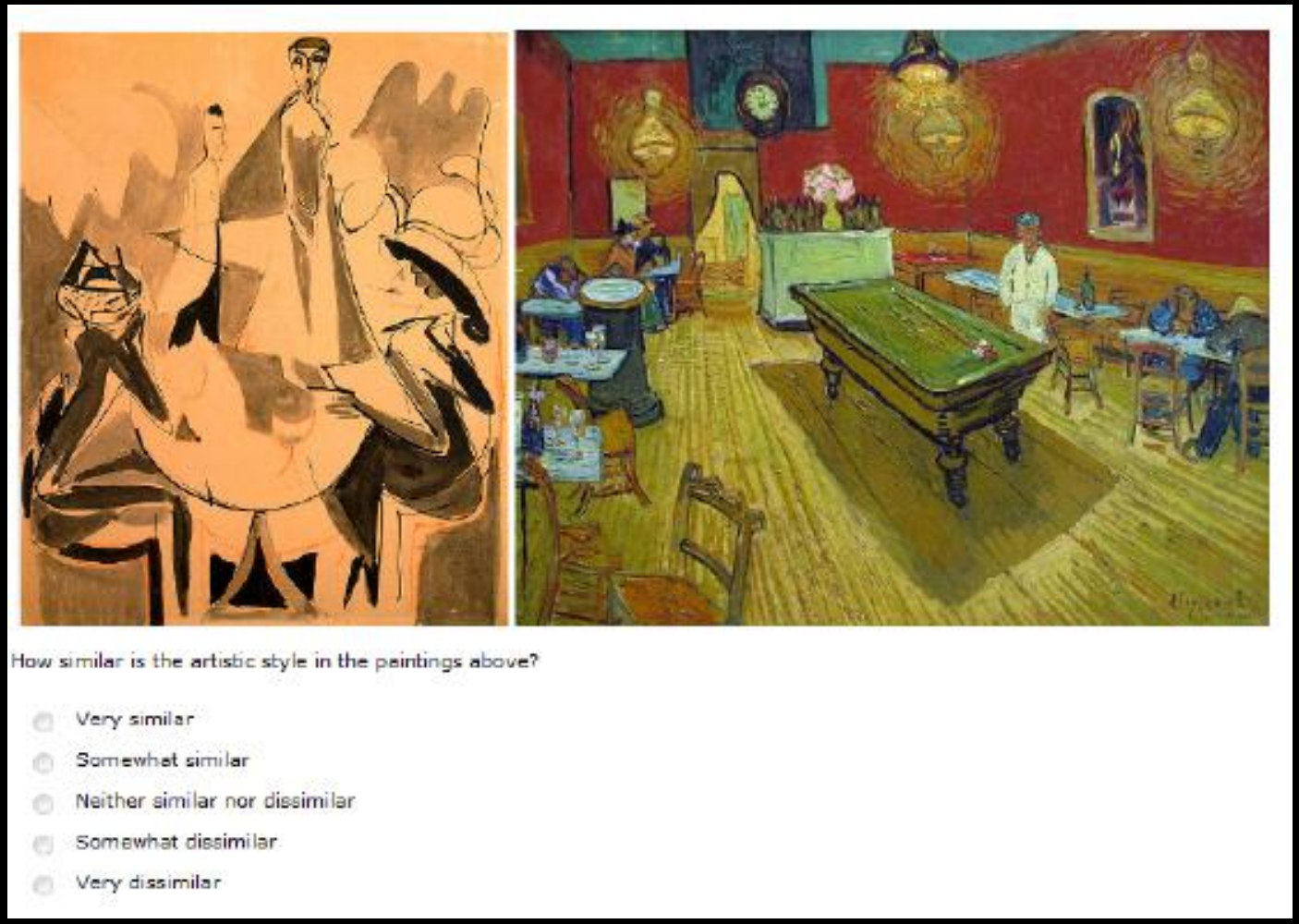
Example – Sheep Market

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



www.thesheepmarket.com

Kovashka & Lease, CrowdConf'10



Example – Dialect Identification

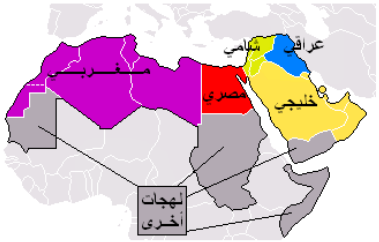
Help Classify Arabic into Dialects!

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 – 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.

A New Class of Applications

Hybrid applications blend automation with human computation to achieve new capabilities

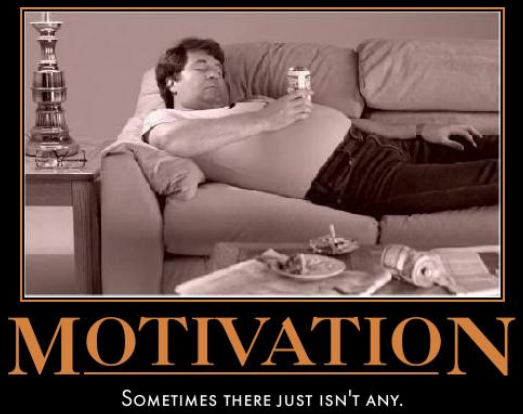
- S. Cooper et al. (2010). Predicting protein structures with a multiplayer online game.
- B. Bederson et al. (2010). Translation by Iterative Collaboration between Monolingual Users.
- T. Yan et al. (2010). CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones.
- M. Bernstein et al. (2010). Soylent: A Word Processor with a Crowd Inside.

Models & Incentives

- Why do workers do it?
- How do I crowdsource effectively?

Models & Incentives

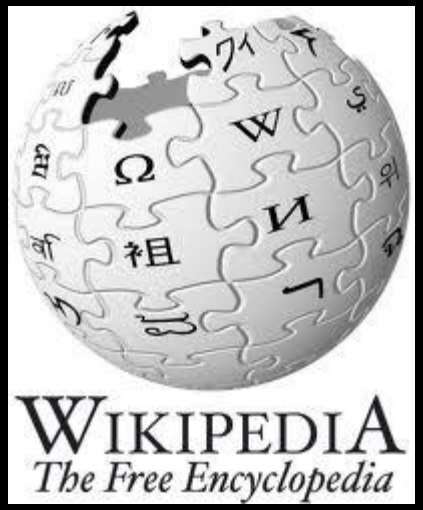
- Pay (e.g. MTurk)
- Fun (or avoid boredom)
- Socialize
- Earn acclaim/prestige
- Altruism
- Learn something new (e.g. English)
- Invisible by-product (e.g. re-Captcha)
- Create self-serving resource (e.g. Wikipedia)



Multiple incentives are often offered in tandem

Altruism

- Contribute knowledge
- Help others (who need knowledge)
- Help workers (e.g. *SamaSource*)
- Charity



YAHOO! ANSWERS
Example of Community Question and Answering

Games with a Purpose (L. von Ahn)

- Players have fun, creators get data as by-product

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

Invisible By-product

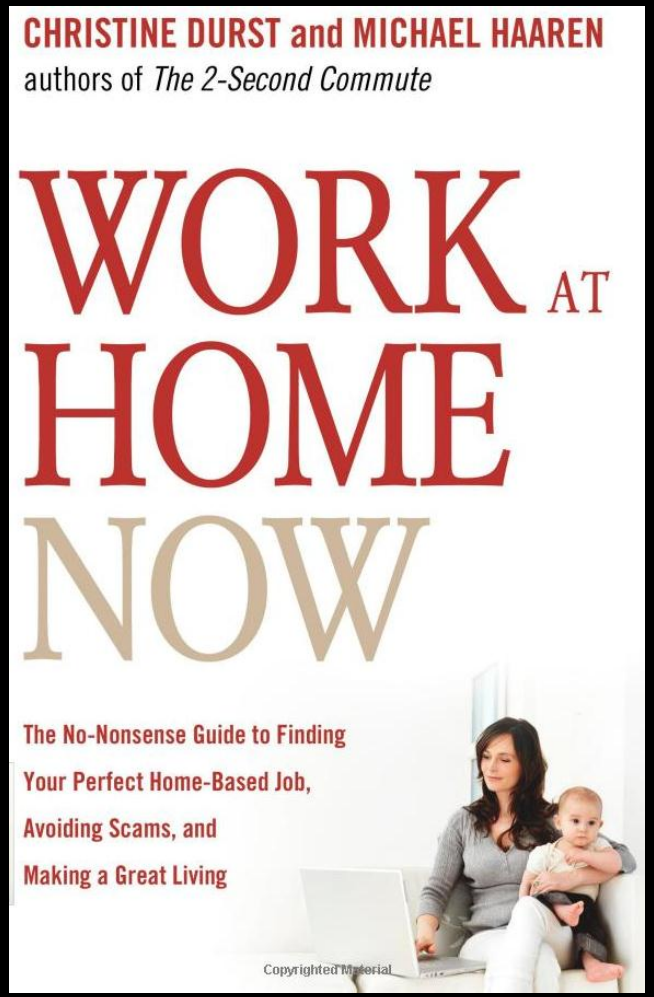
L. von Ahn et al. (2008). *recaptcha…* In *Science*.

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.

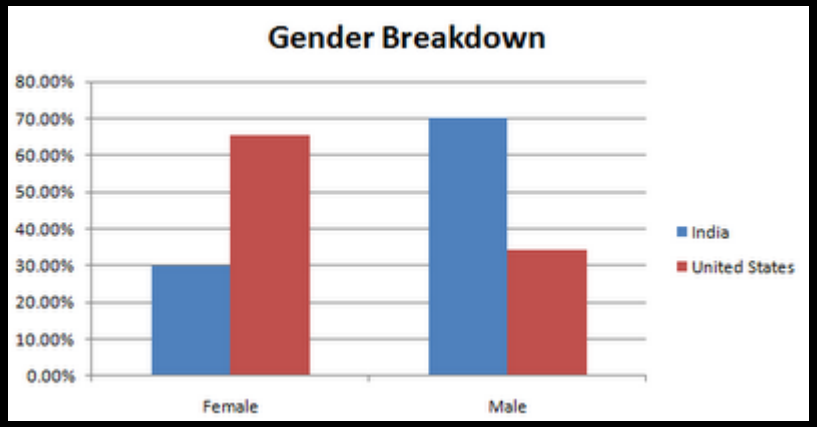
Worker Demographics

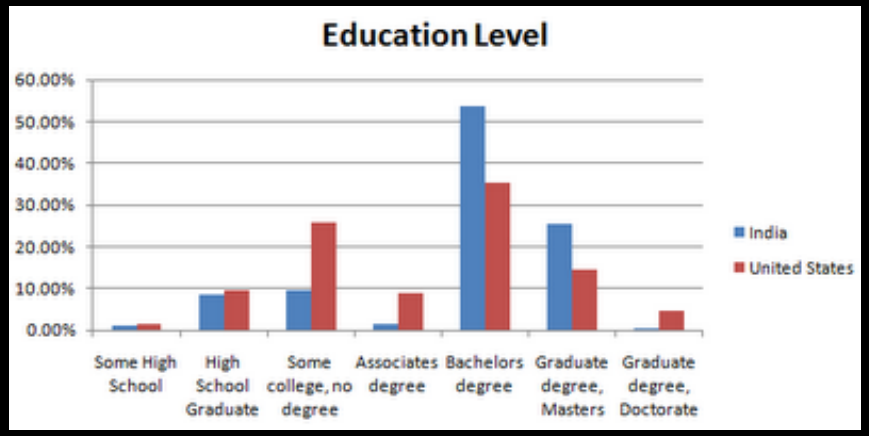
- 2008-2009 studies found less global and diverse than previously thought
  - US
  - Female
  - Educated
  - Bored
  - Money is secondary

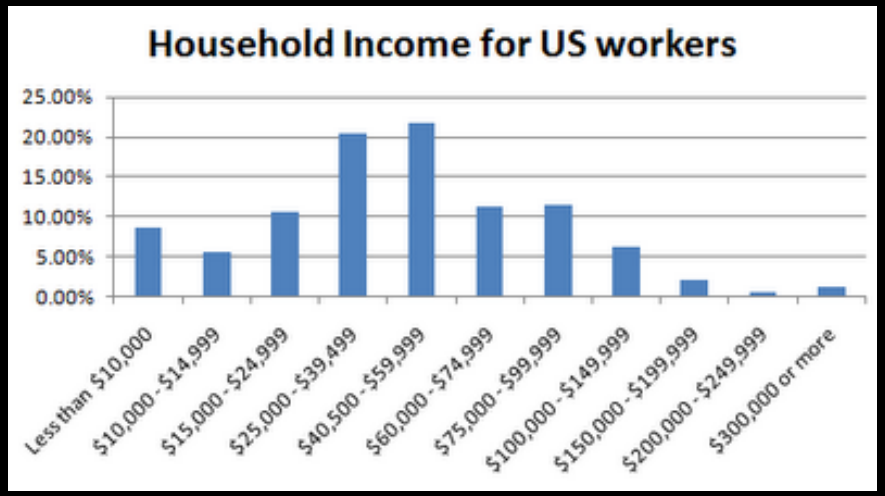


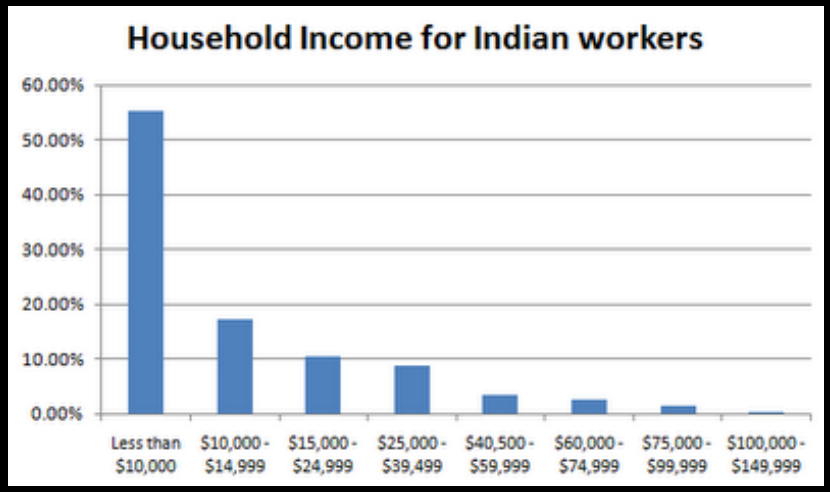
2010 shows increasing diversity

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









Human Subjects Research: When Workers Become Participants

Tutorials & Examples

- W. Mason and S. Suri. Conducting Behavioral Research on Amazon's Mechanical Turk. (October 12, 2010).
- L. Schmidt. Crowdsourcing for Human Subjects Research. CrowdConf'10.
- Michael D. Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? UT Austin, in press.
  - M. Buhrmester's Amazon Mechanical Turk Guide for Social Scientists

MTurk: No Native Support for…

- Verifying participant demographics
- randomizing stimuli (eg question order)
- controlling against same participant joining multiple trials
- "waiting room" for gathering multiple subjects for multi-participant studies
  - See Mason and Suri (2010).

Institutional Review Board (IRB)

- What crowdsourcing studies fall under IRB purvue?
- How to explain crowdsourcing to inexperienced IRBs?

Two NLP examples (with workers rather than participants) :

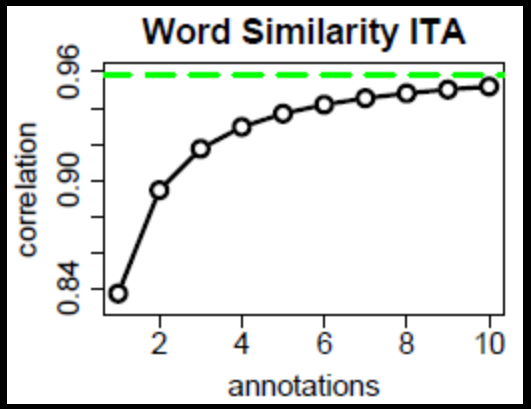
1. Yinon Bentor, UT Austin CS:
   - Determination: work falls outside scope of IRB
   - No direct contact with study participants
   - No collection of confidential data

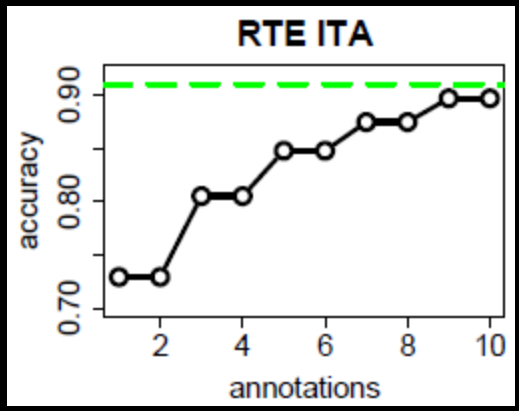
2. Chris Callison-Burch, Johns Hopkins University

Cheap and fast, but is it good?

Snow et al. (2008). EMNLP

- 5 Tasks
  - Affect recognition
  - Word similarity
  - Recognizing textual entailment
  - Event temporal ordering
  - Word sense disambiguation
- 22K labels for $26
- high agreement between Turk annotations and expert “gold” labels





Structuring Labor / Organizing Workers

- Flat: all workers exchangeable
- Hierarchical: laymen vs. experts or multi-level
  - S. Kochar et al., HComp‘10
- Structured (e.g. Find-Verify or Find-Fix-Verify)
  - Quinn and B. Bederson’09, M. Bernstein et al.’10

Consensus

- Majority vote
  - simple, common consensus method
  - What if 2 say yes, 3 say no?
    - Flat: Collect more judgments
    - Hierarchical: Use higher-level workers to break ties
  - How many labels to collect?
    - Fixed N (simple)
    - Vary N based on example difficulty and target confidence
- Better consensus methods exist
  - Various machine learning questions & work

Consensus Questions

- How to measure worker quality
  - P. Ipeitotis. Worker Evaluation in Crowdsourcing: Gold Data or Multiple Workers? September, 2010
  - balanced vs. imbalanced data (e.g. accuracy vs. P/R)
  - How to estimate worker from sparse data
- How to use worker quality for better consensus
- How to maximize labeling effort for learning: improve label accuracy or label new examples?
  - V.S. Sheng et al. KDD'08, Kumar & Lease, CSDM'11
  - SIGIR'11 poster (in review)

Quality control

- Approach as “overall” quality – not just workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester
- Assess worker quality as you go
  - “Trap questions” with known answers (“honey pots”)
  - Measure inner-annotator agreement between workers
    - Risk: confusing valid ambiguity or diversity, “tail” behaviors
- Distinguish bias from noise (e.g. personal scales)
  - Normalize before aggregating across individuals
- Defend against “spammers”
  - Common “attacks”: constant, random, majority label

Quality control - MTurk

- Approval rate: easy to use, easy to defeat
  - P. Ipeirotis. Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes. (10/10)
- Geographic restrictions (e.g. US only)
- Qualification test
  - Pre-screen workers' ability to do task (accurately)
    - Difficult with subject judgment tasks
  - Can get user familiar with task *before* you pay
  - May slow down experiment
- Block worker
  - Affects worker reputation as well
  - Not to be used lightly

Other Issues

- Usability factors (e.g. Grady & Lease, NAACL'10 MTurk Workshop)
  - Provide clear, concise labels that use plain language
  - Avoid unfamiliar jargon and terminology
  - May contradict traditional usability (Kittur et al. CHI'08)
- HR issues: recruiting, selection, & retention
  - Build your reputation (disclose identity?)
  - Tweet tasks, design a better qualification test, give bonuses
  - Experiments go faster once established
- Always request written feedback from workers
  - Often get label justifications (for free / minimal cost)
  - Quasi-captcha, though automatic verification may be difficult

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
- Detect and block spammers

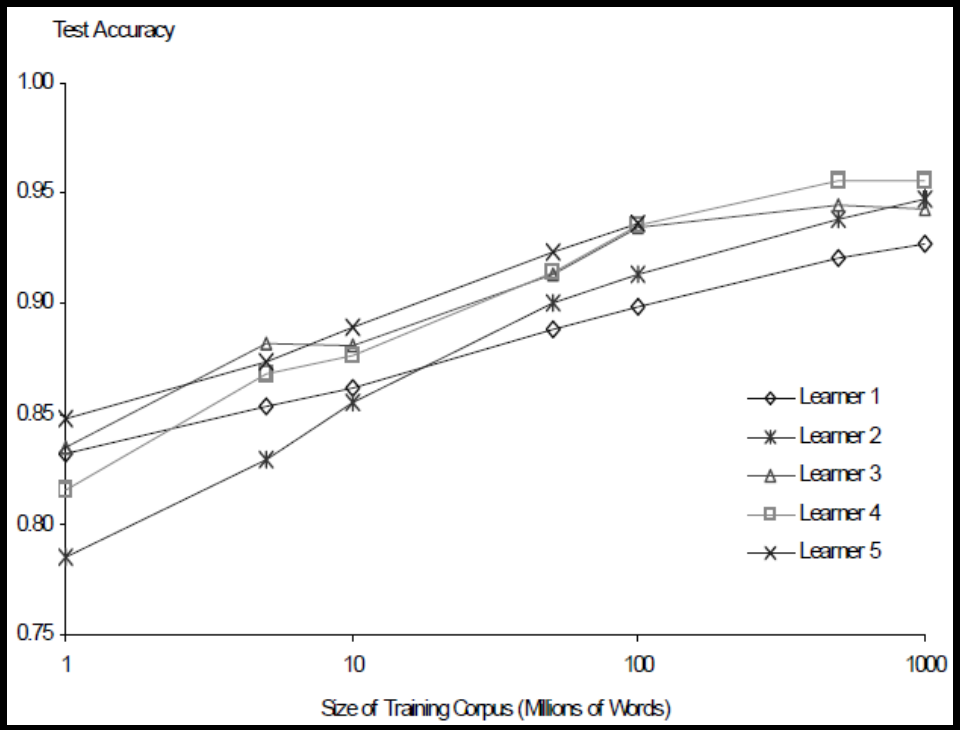
Other practical tips

- Sign up as worker and do some HITs
- "Eat your own dog food"
  - Do it yourself, then have friends do it
- Scale incrementally: first data, then workers
- Monitor discussion forums, address feedback
- Everything counts!
  - Overall design only as strong as weakest link

Unreasonable Effectiveness of Data

Massive free Web data changed how we train learning systems

- Banko and Brill (2001). Human Language Tech.
- Halevy et al. (2009). IEEE Intelligent Systems.

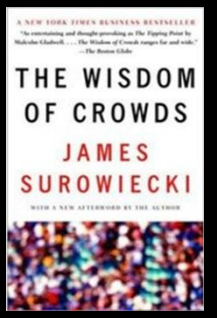


How might access to cheap & plentiful labeled data change the balance again?

Modeling uncertainty

- AI tenet: expose, model, and propagate
- Labeling variance often viewed as "noise" to resolve via consensus algorithms (e.g. majority vote)
- Variance may really reflect inherent ambiguities or distributional nature of the data
- While consensus simplifies, could model uncertainty:
  - Learning (cf. G. Lugosi 1992, P. Smyth, 1995)
  - Evaluation
  - Active learning
    - Example informativeness vs. probability of being labeled correctly
- Temporal uncertainty of HPU: yet another dimension…

Crowd Wisdom & Ensemble Learning



- Combine multiple models to improve performance
  - 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

MapReduce with human computation

- 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

Current MTurk Limitations

- No tools for data analysis
- No integration with databases
- Limited search and browse features
- Limited quality control mechanisms
- No Ratings / recommendations (e.g. books)
- No work routing: who are the right workers given task nature and constraints?
- Human Subjects Research issues noted earlier

Questions

- Usabiliy / human factors / design vs. outcomes?
- How to price tasks automatically?
- How to predict worker quality from observable properties?

- What's the best "mix" of HPU for a task?
- What are the tasks suitable for HPU?
- What level of abstraction for invoking HPU? A new language?
  - `crowdsource(task,5)`
- How to measure inner-annotator agreement when every example labeled by a distinct set of annotators?

Wrap-up

Conclusions

- Crowdsourcing here to stay
  - Shift in practice for conducting research
  - New phenomenon to be studied in its own right
- Fast, cheap, ~ easy, ~ accurate data collection
  - Higher level infrastructure still needed to let researchers focus on tasks, not platform
  - Can collect bad data faster and easier than ever!
    - Still need careful experimental design, & effective design for new crowdsourcing environment
- A new class of NLP applications will integrate automation with HPU to utilize best of both
- "Hot" emerging area with many open problems…

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

Blogs & Sites

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

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…

Past Workshops and Conferences

- 2011: WSDM-CSDM: Crowdsourcing for Search and Data Mining
- 2010
  - AMTA: Collaborative Translation: tech., crowdsourcing, & translator perspective
  - COLING: 2nd People's Web Meets NLP: Collaboratively Constructed … Resources
  - CrowdConf 2010: 1st Conference on the Future of Distributed Work
  - CVPR-ACVHL: Advancing Computer Vision with Humans in the Loop
  - ICWE: Enterprise Crowdsourcing
  - KDD-HCOMP: 2nd Human Computation
  - NAACL: Creating Speech and Language Data With Amazon's Mechanical Turk
  - NIPS: Computational Social Science and the Wisdom of Crowds
  - SIGIR-CSE: Crowdsourcing for Search Evaluation
  - Ubicomp: Ubiquitous Crowdsourcing
  - Maryland Workshop on Crowdsourcing and Translation
- 2009
  - KDD-HCOMP: 1st Human Computation
  - ACL/IJCNLP: 1st People's Web Meets NLP: Collaboratively Constructed … Resources

Resources & Upcoming Events

Special issue of Springer's 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!

- Omar Alonso, Microsoft Bing
- Students
  - Catherine Grady (iSchool)
  - Hyun Joon Jung (ECE)
  - Adriana Kovashka (CS)
  - Abhimanu Kumar (CS)
- Support
  - John P. Commons

ir.ischool.utexas.edu/crowd

UT Mechanical Turk & Crowdsourcing Google Group

Matt Lease - ml@ischool.utexas.edu

Ask a question and I'll find someone to answer
Example questions
Ask someone
1. Send Aardvark a question
2. Aardvark finds the perfect person to answer
3. Get their response in a few minutes
What's a great biking path around Golden Gate Park?
Sent 11:16 AM PDT
My favorite is a secret trail that takes you to the beach...
Sent 11:20 AM PDT

You

In the sentence "Jane touched the man with the stick", who has the stick? Is it (a) Jane, (b) the man, or (c) neither? If multiple answers seem possible, which corresponded to your first instinct? Thanks!

Me K.
F / Minneapolis, MN

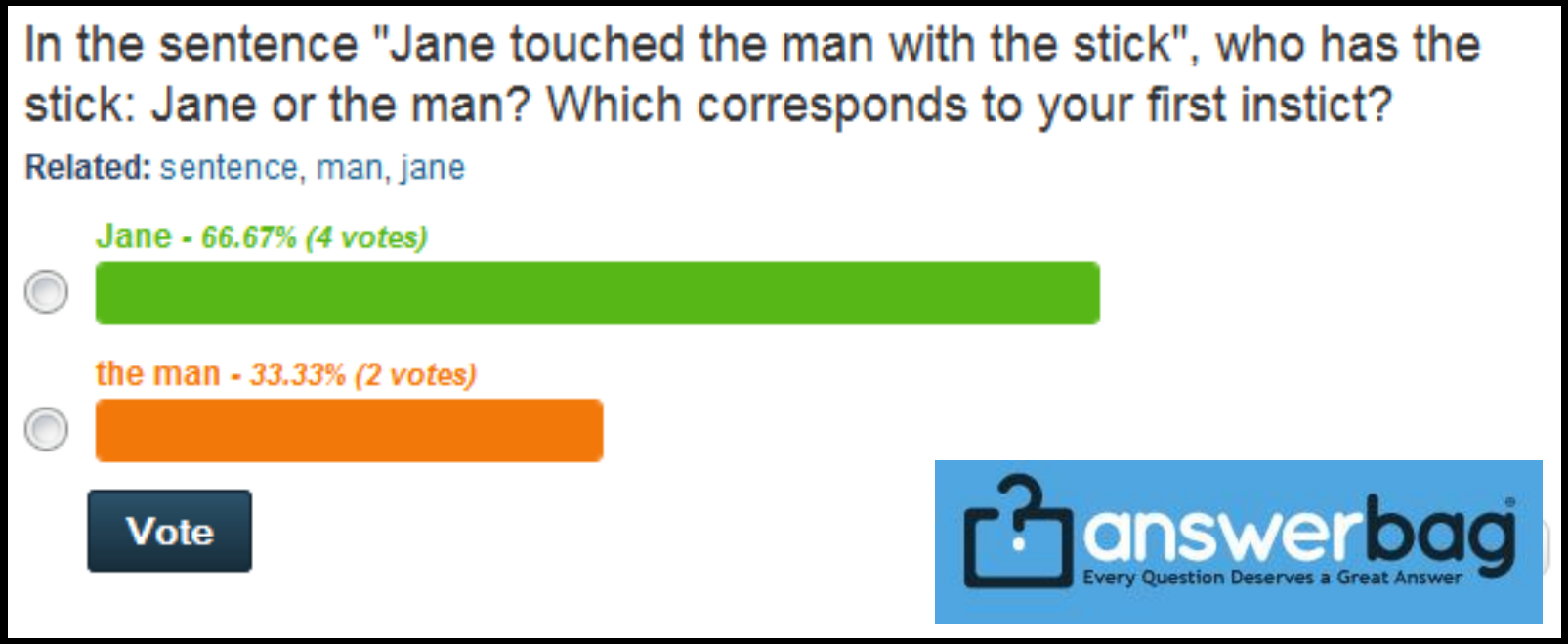
Jane

Jacob R.
30 / M / Montreal, Canada

it's a poorly written sentence in that either Jane or the man could have the stick. (That is to say, "with the stick" could be an adverbial clause modifying "touched" or could be an adjectival clause modifying "the man.") But the answer your second questions is that my initial inclination would be to think it's adverbial--that is, Jane has the stick--but there's no way of knowing the intent of the writer.

Sam F.
20 / M / Rye, NY

The best answer would be (d) either. The sentence is grammatically ambiguous, so it could be Jane, or it could be the man; there is no way to tell without more context. My first thought was that Jane had the stick though, fwiw.



At first glance most will think Jane. it is necessary to have more of the context of the paragraph this sentence was taken from to fully comprehind who had the stick

First instinct, Jane has the stick, because Jane could use the stick to touch him with. If the sentence read, "Jane touched the man with the hat.", I would have guessed that Jane was touching a man wearing a hat.

I was convinced the man had the stick because of "...man with(the) stick"

had it read "...man with (a) stick" I'd probably had concluded Jane had the stick