

Crowdsourcing in Real World

CMPT 884, FALL 2016

JIANNAN WANG

<https://sfu-db.github.io/cmpt884-fall16>

What is Crowdsourcing?

Outsourcing

- Allocates work to a **defined** organizational entity

Crowdsourcing

- Allocates work to an **unorganized** collection of individuals

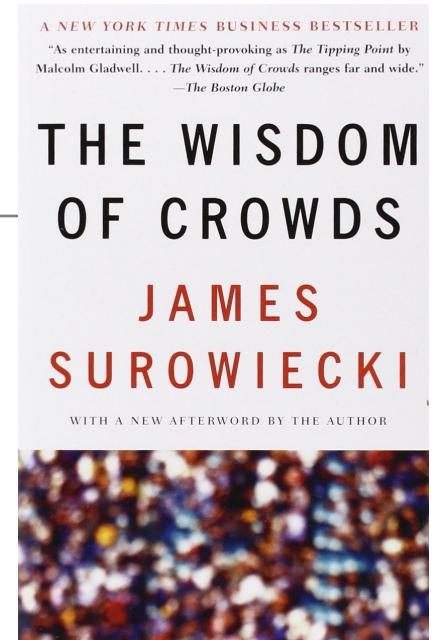
The Wisdom of Crowds

What does it mean?

- Two heads are better than one

A famous example: Wikipedia

But it does not mean it always works ([video](#)).



Challenges To Manage Crowds

How to recruit crowds?

How to retain crowds?

What contributions can crowds make?

How to combine crowd contributions?

How to evaluate crowds and contributions?

How To Recruit Crowds?

An equivalent question:

What benefits can the crowds get?

- Making money
- Learning new knowledge
- Playing games
- Being able to use other services

Making Money

Amazon Mechanical Turk (500K+ workers*)

The screenshot shows the Amazon Mechanical Turk homepage. At the top, there are navigation links for "Your Account", "HITS", and "Qualifications". A message "Already have an account? Sign in as a Worker | Requester" is displayed. Below this, a banner states "Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient." It also displays "694,300 HITs available. View them now.". The main section is divided into two columns: "Make Money by working on HITs" and "Get Results from Mechanical Turk Workers". The "Make Money" column features icons for finding tasks, working, and earning money, along with a "Find HITs Now" button. The "Get Results" column features icons for funding your account, loading tasks, and getting results, along with a "Get Started" button.

The screenshot shows the Amazon Mechanical Turk requester interface. At the top, it says "473,182 HITs available now". Below this, a search bar allows filtering by "Find HITs containing" and "that pay at least \$ 0.00". A timer indicates "00:00:00 of 2 minutes". A "Want to work on this HIT?" button is present. The main area displays a HIT titled "Identify if two receipts are the same". It shows "Requester: Jon Breig" and "Qualifications Required: None". Two receipts are shown side-by-side: "Receipt 1" and "Receipt 2", both from Walgreens. The receipts show purchases of alkaline batteries. A question "Are these two receipts the same?" with "YES" and "NO" buttons is displayed below the receipts.

* <https://requester.mturk.com/tour>

Learning New Knowledge



duolingo

**Learning a new language
while helping to translate the web**

Playing Games

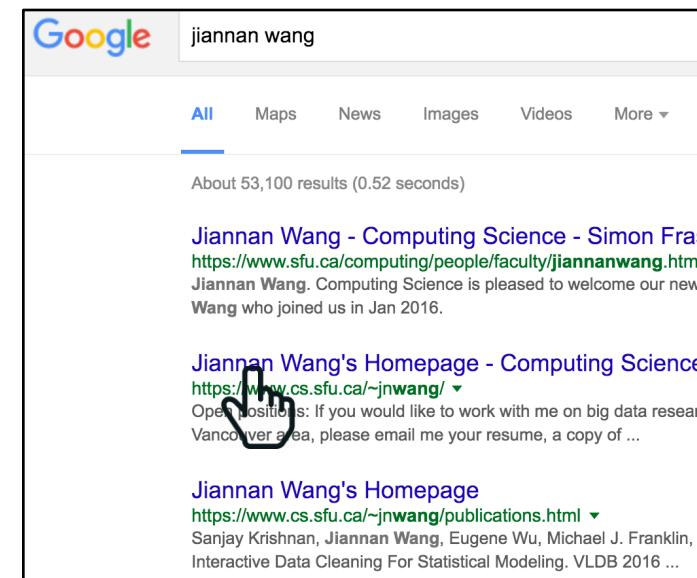
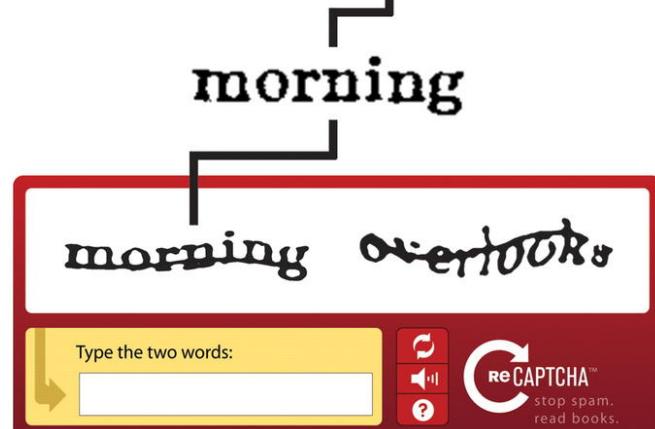


Playing a game
while helping to Labeling
images

For Using Other Services

Everyone has been recruited!

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.



How To Retain Crowds?

Ownership

- E.g., co-founder, co-contributor

Time-constrained Bonus

- E.g., \$100 for staying ten days

Reputation

- Beginner → Skilled → Experienced → Expert



stackoverflow

Leader Board

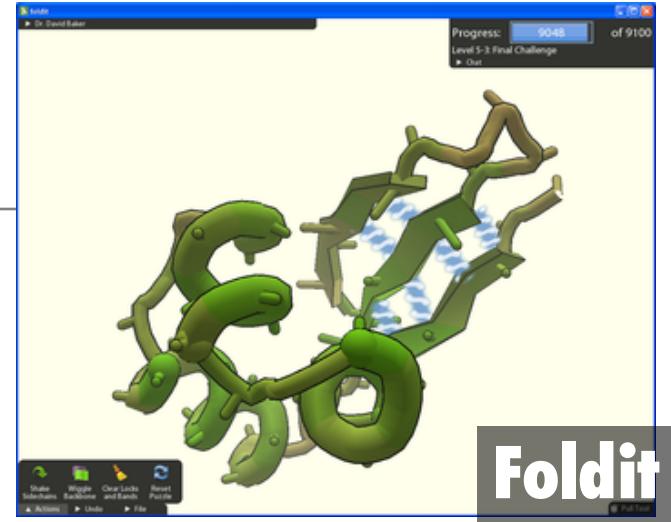
- #1 Mike, #2 John, #3 Tom, ...



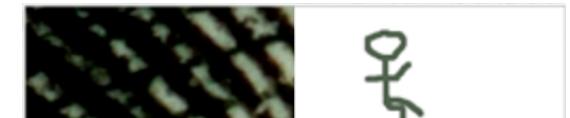
What contributions can crowds make?

Far beyond the imagination

- Protein structure prediction
- Creating digital artworks



10,000 Cents



How to Combine Crowd Contributions?

Automatic

- E.g., Majority Vote

Manual

- E.g., Report to a higher level



How to Evaluate Crowds and Contributions?

Task Redundancy

- Multiple answers + EM algorithm

Hidden Ground Truth

Peer Evaluation

- Asking crowds themselves to evaluate each other

Crowdsourcing + Data Management

Industrial Survey

Company	Team	Persona
Amazon	Product classification	Largely single-case user
Captricity	Focus of large part of company	Largely single-case user
Dropbox	Single person consulting several teams	Multi-case user / Internal provider
Facebook	Entities team	Multi-case user
Flipora	Startup CTO	Multi-case user
GoDaddy	Small business data extraction	Multi-case user
Groupon	Merchant data team	Multi-case user
Google	Internal crowdsourcing team	Internal provider
Google	Web knowledge discovery team	Multi-case user
LinkedIn	Single person consulting several teams	Multi-case user / Internal provider
Microsoft	Internal crowdsourcing team	Internal provider
Microsoft	Search relevance team	Multi-case user
Youtube	Crowdsourcing team	Largely single-case user



Use Cases

Use Cases	# Participants
Classification	12
Entity resolution	6
Data cleaning	5
Ranking	5
Spam detection	5
Data extraction	5
Text generation	5

Classification

Examples

- Infer a user's gender or job from a social media profile
- Infer a business's category (e.g., dry cleaner or restaurant)
- Sentiment analysis on product reviews

Entity Resolution

Finding different records that refer to the same real-world entity

Simon Fraser University

Simon Fraser U.



Data Cleaning

Detecting and removing data errors (missing values, inaccurate values, etc.)

- o Cleaning knowledge graph
- o Verifying business information

A screenshot of a Groupon listing for "Little Sheep Mongolian Hot Pot". The listing includes a restaurant icon, the name "Little Sheep Mongolian Hot Pot", a "Best Of Groupon" badge, the address "1411 156th Ave NE, Bellevue, WA 98007", a "Directions" link, a phone number "+14256531625", and a "View Website" button.

A screenshot of a Google search results page for "larry page". The search bar shows the query. Below it, there's a "Knowledge Graph" panel on the right side. The panel features a large photo of Larry Page, several smaller images of him, and his name. It also contains text about his role as Chief Executive Officer of Alphabet Inc., his birth date (March 26, 1973), and his net worth (\$39 billion USD). The main search results list includes links to Wikipedia, Forbes, Google+, Biography.com, and Business Insider articles about Larry Page.

Ranking

Examples

- Evaluating / tuning search engine
- Identify high-quality articles to recommend to users

Spam Detection

Examples

- Web spam, email spam, comment spam
- Illicit search engine optimization (seo) schemes

Data Extraction

Examples

- Digitizing paper forms
- Extracting structured data from Web page



The screenshot shows the Simon Fraser University Computing Science faculty page. The page has a dark header with the SFU logo and the text "SIMON FRASER UNIVERSITY" and "Computing Science". Below the header is a navigation menu with links like "HOME", "ABOUT COMPUTING SCIENCE", "CONTACT US", "PEOPLE", "PROSPECTIVE STUDENTS", "CURRENT STUDENTS", "RESEARCH", "INDUSTRY RELATIONS", "NEWS & EVENTS", and "INTRANET". The main content area is titled "FACULTY" and contains three tabs: "Emeriti Faculty Members", "Adjunct Professors", and "Associate Members". A yellow box highlights the "Associate Members" section, which lists several faculty members with their names, areas of instruction, and links to their profiles and home pages. The highlighted section includes:

- GREG BAKER, SENIOR LECTURER
Area: Instruction
[Profile & Contact Information](#) | [Home Page](#)
- BRAD BART, SENIOR LECTURER
ASSOCIATE DIRECTOR
Area: Instruction
[Profile & Contact Information](#) | [Home Page](#)
- PETRA BERENBRINK, PROFESSOR
ACTING GRADUATE PROGRAM DIRECTOR
Area: Probabilistic methods; Randomized algorithms; Analysis of dynamic processes
[Profile & Contact Information](#) | [Home Page](#)
- BINAY BHATTACHARYA, PROFESSOR
Area: Computational Geometry; Pattern Recognition
[Profile & Contact Information](#) | [Home Page](#)
- ANDREI BULATOV, PROFESSOR
Area: Constraint Satisfaction; Complexity of Computation
[Profile & Contact Information](#) | [Home Page](#)
- ROBERT D. CAMERON, PROFESSOR
Area: Software Engineering Languages; Electronic Publication
[Profile & Contact Information](#) | [Home Page](#)
- BOBBY CHAN, LIMITED TERM LECTURER
Area: Instruction
[Profile & Contact Information](#)
- PARMIT CHILANA, ASSISTANT PROFESSOR
Area: Human-Computer Interaction
[Profile & Contact Information](#)

Text Generator

Examples

- Researching a company or product and writing a blurb about it
- Summarizing news articles
- Rewriting existing content

Conclusion

What is crowdsourcing?

- Allocates work to an **unorganized** collection of individuals

Challenges and solutions to manage crowd workers

- Crowd recruitment/retainment/contributions/combination/evaluation

Crowdsourced data-management problems

- Classification, Entity Resolution, Data Cleaning, Ranking, Spam Detection, Data Extraction, Text Generation

Requirements for next week

For presenters

- Assume it's your own paper
- Do (at least) one practice talk
- Give enough background knowledge
- Pay attention to your time (35mins + 15mins Q&A)

For the audience

- Do (at least) one pass over each paper
- Prepare (at least) one question

Two more things

Mon 9/19	Systems and Programming Models	CrowdDB: Answering Queries Using Crowdsourcing <u>TurKit:</u> Read the Turkit paper in class	Turk	Sima [slides] Han Shen [slides]
Wed 9/21		CrowdForge: crowdsourcing complex work		Han Bao [slides]

Learning the basics of Python by Oct 1