

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

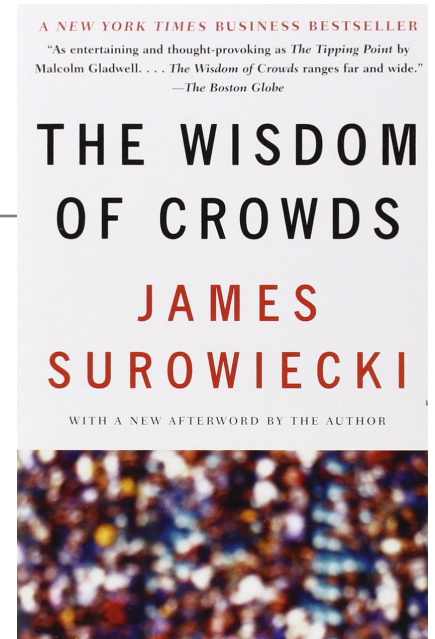
The Wisdom of Crowds

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 for requesters. At the top, there's a navigation bar with 'Your Account', 'HITS', and 'Qualifications' tabs. Below this, a yellow 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. 694,300 HITS available. View them now.'

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

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



or learn more about being a Worker

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

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



The screenshot shows the Amazon Mechanical Turk interface for a specific HIT. At the top, there's a navigation bar with 'Your Account', 'HITS', and 'Qualifications' tabs. Below this, a yellow banner states: '473,182 HITs available now'. The main content area shows the HIT details: 'Identify if two receipts are the same', 'Requester: Jon Brellig', 'Qualifications Required: None', 'Reward: \$0.01 per HIT', 'HITS Available: 1', 'Duration: 2 minutes'. There are buttons for 'YES' and 'NO' to answer the question. Below the question, there are two receipts labeled 'Receipt 1' and 'Receipt 2' for comparison.

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

Learning New Knowledge



duolingo

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

Playing Games

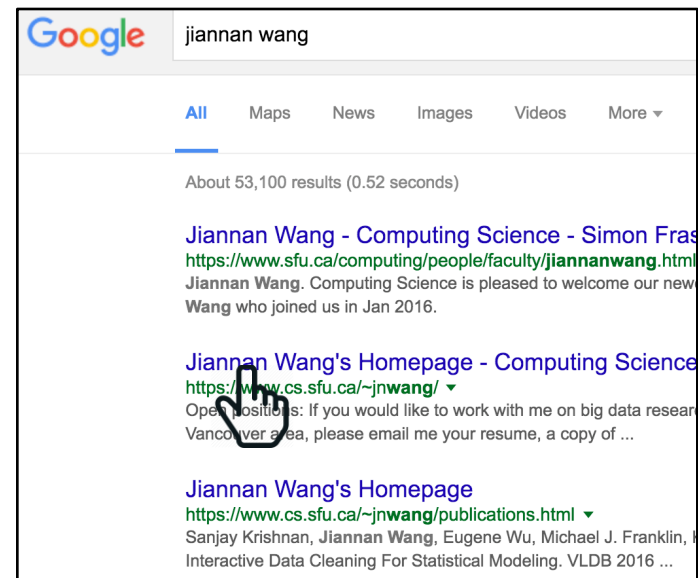
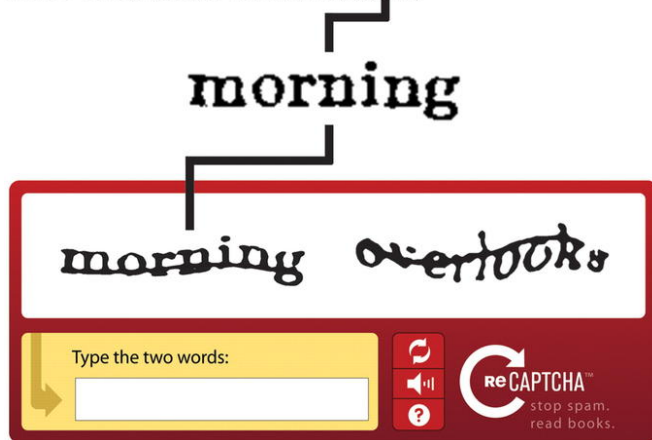


Playing a game
while helping to **Labeling**
images

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



Leader Board

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



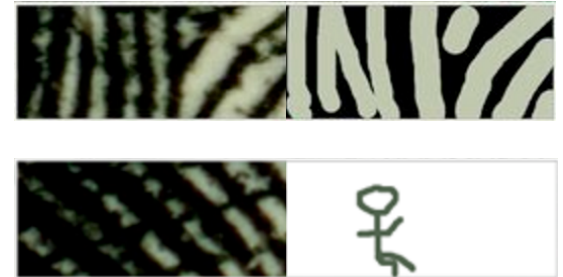
What contributions can crowds make?

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

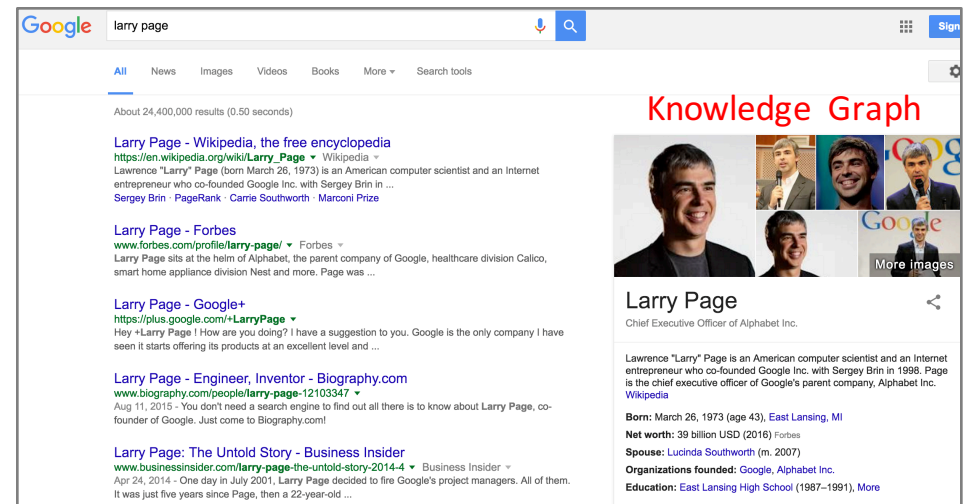
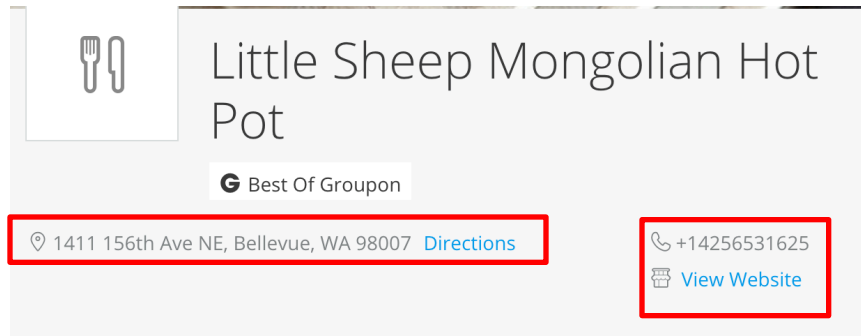
Simon Fraser U.



Data Cleaning

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

- Cleaning knowledge graph
- Verifying business information



Ranking

Examples

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

Spam Detection

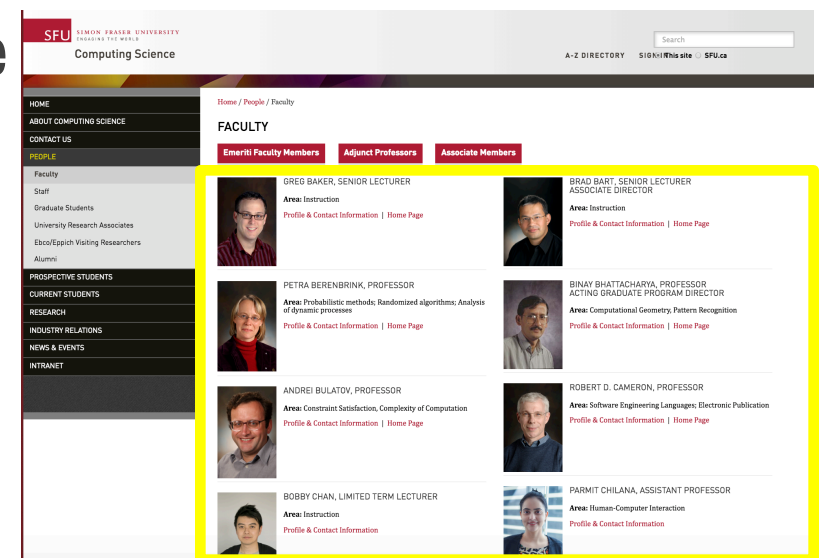
Examples

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

Data Extraction

Examples

- Digitizing paper forms
- Extracting structured data from Web page



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 TurKit:	Turk	Sima [slides] Han Shen
Wed 9/21		CrowdForge: crowdsourcing complex work		[slides] Han Bao [slides]

Learning the basics of Python by Oct 1