



Purdue Fall 2018

CS59000-CSC: Crowdsourcing and Social Computing

<http://mingyin.org/CS590/Fall2018/index.html>

Class 1: Introduction & Course Overview

2018.8.20

Ming Yin

Who am I?

- <http://mingyin.org/>
- New to Purdue!
 - Assistant Professor in Computer Science
 - PhD, Harvard University
- Research: Social computing and crowdsourcing, HCI, computational social science, applied AI and ML

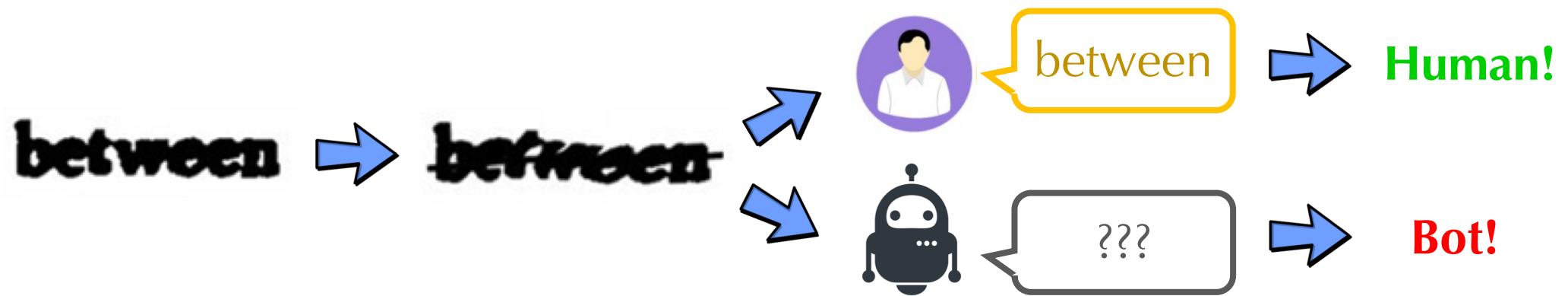
Today's Schedule

- What is Crowdsourcing and Social Computing?
- Getting to Know You
- Course Format and Logistics

What is Crowdsourcing and Social Computing?



An Earlier Version: CAPCHA



Completely Automated Public Turing test to tell Computers and Humans Apart (**CAPTCHA**)

OCR is Hard...

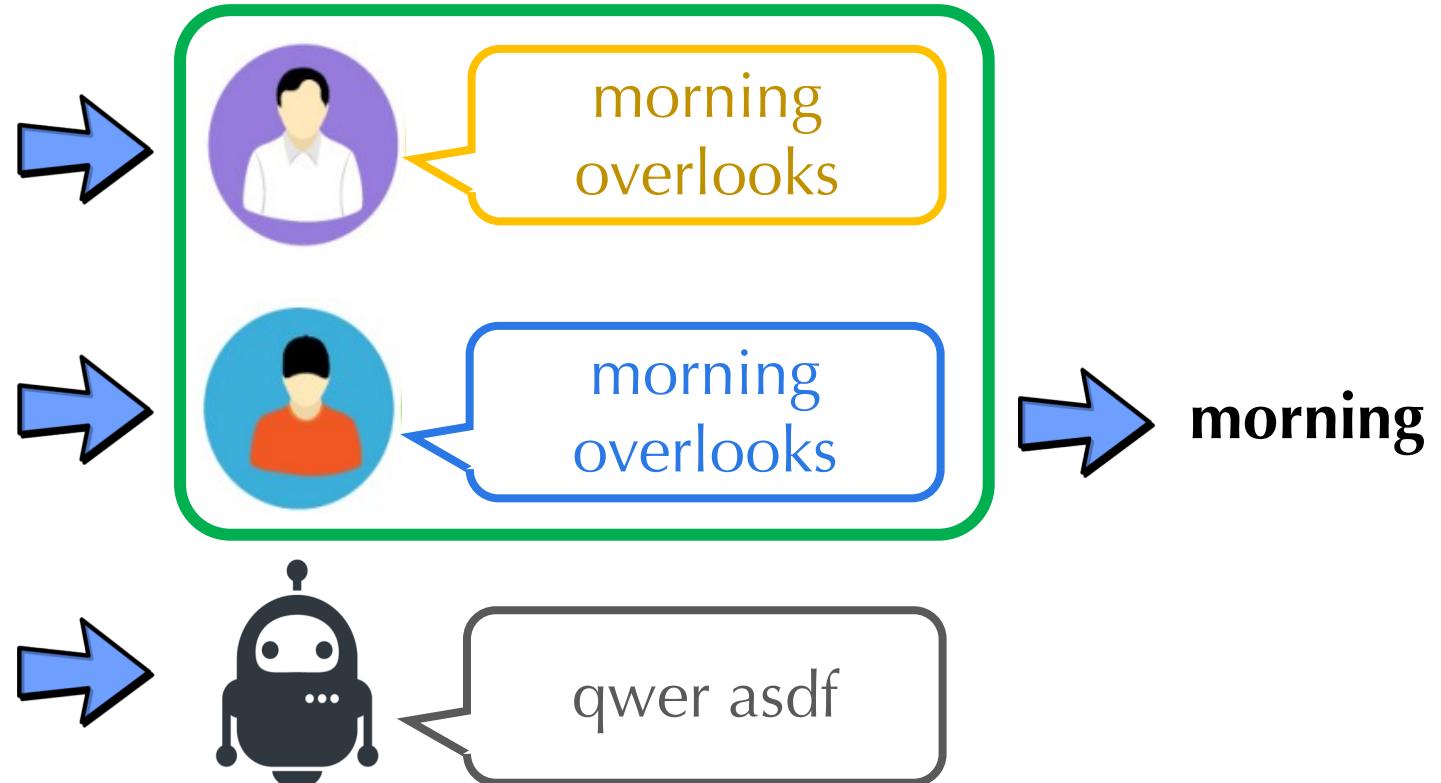
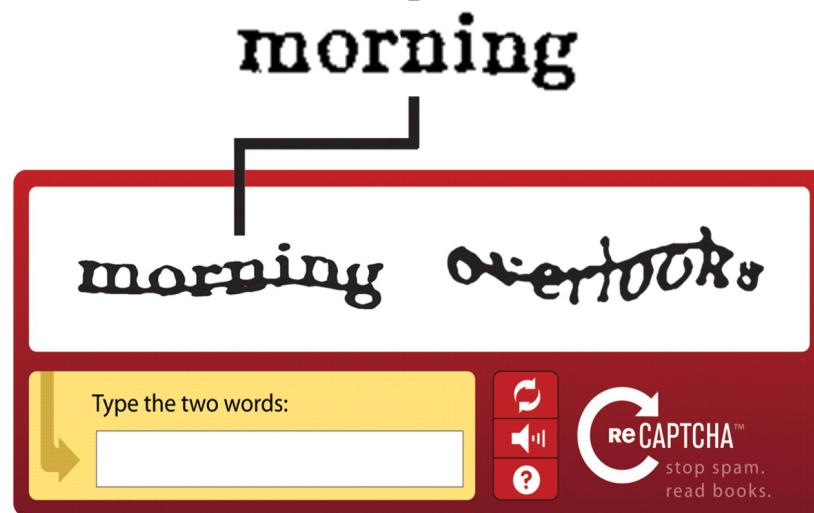
Scanned Text: **This aged portion of society were distinguished from**

OCR Output: “niis aged pntkm at society were distinguished frow”

~30% Error Rate...

From CAPTCHA to reCAPTCHA

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



von Ahn et al. [reCaptcha: Human-based Character Recognition via Web Security Measures](#). Science, September 2008

Document Digitalization with reCAPTCHA



Completed digitizing the archives of the New York Times dating from 1851 to the present day (more than 13 million articles!)

Variations of reCAPTCHA

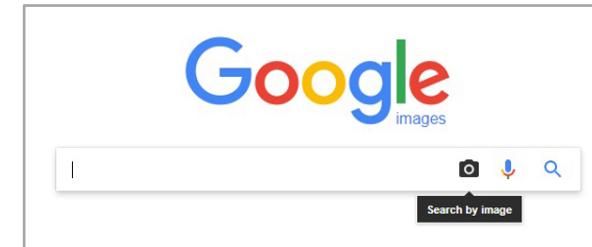
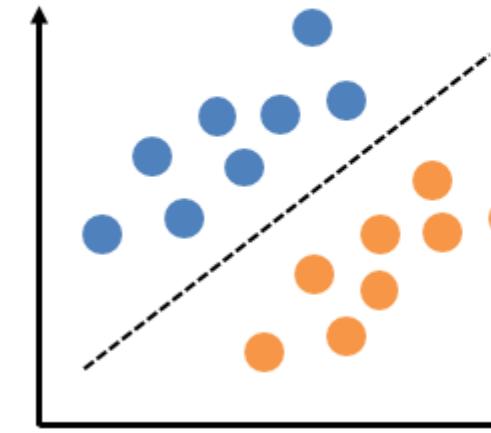


Decode street address for Google Street View

Variations of reCAPTCHA



Training Data
↔
Hard Tasks



Label images to improve image search performance

Another example: Wikipedia



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Crowdsourcing

From Wikipedia, the free encyclopedia

Crowdsourcing is a [sourcing model](#) in which individuals or organizations obtain [goods and services](#). These services include ideas and finances, from a large, relatively open and often rapidly-evolving group of [internet](#) users; it divides work between participants to achieve a cumulative result. The word crowdsourcing itself is a [portmanteau](#) of [crowd](#) and [outsourcing](#), and was coined in 2005.^{[1][2][3][4]} As a mode of sourcing, crowdsourcing existed prior to the digital age (i.e. "[offline](#)").^[5]

There are major differences between crowdsourcing and outsourcing. Crowdsourcing comes from a less-specific, more public group, whereas outsourcing is commissioned from a specific, named group, and includes a mix of bottom-up and top-down processes.^{[6][7][8]} Advantages of using crowdsourcing may include improved costs, speed, quality, flexibility, scalability, or diversity.^{[9][10]}

Some forms of crowdsourcing, such as in "idea competitions" or "innovation contests" provide ways for organizations to learn beyond the "base of minds" provided by their employees (e.g. [LEGO Ideas](#)).^[11] Tedious "microtasks" performed in parallel by large, paid crowds (e.g. [Amazon Mechanical Turk](#)) are another form of crowdsourcing. It has also been used by [not-for-profit](#) organizations and to create [common goods](#) (e.g. [Wikipedia](#)).^[12] The effect of user communication and the platform presentation should be taken into account when evaluating the performance of ideas in crowdsourcing contexts.^[13]

Contents [hide]

- 1 Definitions
- 2 Historical examples
 - 2.1 Timeline of major events
 - 2.2 Early competitions
 - 2.3 In astronomy

Another example: Wikipedia

The screenshot shows a Wikipedia article page for "Editing Crowdsourcing". The page has a red box highlighting the "View history" button in the top navigation bar. The main content area contains a warning message about being logged in and a summary of the crowdsourcing concept.

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit source **View history** Search Wikipedia

Editing Crowdsourcing

You are not logged in. Your IP address will be publicly visible if you make any edits. If you log in or create an account, your edits will be attributed to a user name, among other benefits.

Content that violates any copyrights will be deleted. Encyclopedic content must be verifiable. Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to certain terms and conditions.

B I Advanced Special characters Help Cite

'''Crowdsourcing''' is a [[wikt:sourcing|sourcing model]] in which individuals or organizations obtain [[goods and services]]. These services include ideas and finances, from a large, relatively open and often rapidly-evolving group of [[internet]] users; it divides work between participants to achieve a cumulative result. The word crowdsourcing itself is a [[portmanteau]] of [[crowd]] and [[outsourcing]], and was coined in 2005.<ref name="safire">{{cite news | last= Safire | first= William | url= https://www.nytimes.com/2009/02/08/magazine/08wwln-safire-t.html?_r=3&ref=magazine& | title= On Language | publisher= New York Times Magazine | date= February 5, 2009 | accessdate= May 19, 2013}}</ref><ref>Schenk, Eric; Guittard, Claude (2009). [https://hal.archives-ouvertes.fr/file/index/docid/439256/filename/Crowdsourcing_eng.pdf. Crowdsourcing: What can be Outsourced to the Crowd, and Why ?]</ref><ref>Hirth, Matthias; Hoßfeld, Tobias; Tran-Gia, Phuoc. [http://i3wue.de/staff/matthias.hirth/author_version/papers/conf_410_author_version.pdf Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com]

As a mode of sourcing, crowdsourcing existed prior to the digital age (i.e. "[[offline]]").<ref name = wired2006>{{cite news | url = https://www.wired.com/wired/archive/14.06/crowds.html | title = The Rise of Crowdsourcing | work = ''[[Wired (magazine)|Wired]]'' |

Another example: Wikipedia



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
 Atom
Upload file
Special pages
Page information

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#) Talk Read Edit View history Search Wikipedia

Crowdsourcing: Revision history

[View logs for this page \(view filter log\)](#)

Search for revisions

From year (and earlier): From month (and earlier):

For any version listed below, click on its date to view it. For more help, see [Help:Page history](#) and [Help>Edit summary](#).

External tools: [Revision history statistics](#) · [Revision history search](#) · [Edits by user](#) · [Number of watchers](#) · [Page view statistics](#) · [Fix dead links](#)

(cur) = difference from current version, (prev) = difference from preceding version, m = minor edit, → = section edit, ← = automatic edit summary
(newest | [oldest](#)) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) 16:34, 4 August 2018 Keith D (talk | contribs) . . (118,420 bytes) (+10) . . (Fix cite date errors) ([undo](#))
- (cur | prev) 06:40, 4 August 2018 81.155.21.189 (talk) . . (118,410 bytes) (+1,913) . . (added the role of crowdsourcing in Geography) ([undo](#)) ([Tag: Visual edit](#))
- (cur | prev) 05:06, 4 August 2018 141.168.57.26 (talk) . . (116,497 bytes) (+1,270) . . (added references.) ([undo](#)) ([Tag: Visual edit](#))
- (cur | prev) 09:07, 6 July 2018 Mhonrubia (talk | contribs) m . . (115,227 bytes) (-1) . . (→In engineering: I changed codelcotech.com for codelcotech.cl) ([undo](#)) ([Tag: Visual edit](#))
- (cur | prev) 09:01, 6 July 2018 Mhonrubia (talk | contribs) . . (115,228 bytes) (+787) . . (→Historical examples: I have added the engineering crowdsourcing example to the list of crowdsourcing examples.) ([undo](#)) ([Tag: Visual edit](#))
- (cur | prev) 10:33, 4 July 2018 KH-1 (talk | contribs) m . . (114,441 bytes) (+209) . . (Reverted edits by Webphantoms (talk) to last version by 172.58.100.139) ([undo](#)) ([Tag: Rollback](#))

Another example: Wikipedia

Addition To Modern Methods [\[edit\]](#)

Hello fellow Wikipedians,

Firstly, I'm new to Wikipedia, so please forgive any newbie behavior. Secondly, I want to cite my employer's blog, where we publish research and best practices from our crowdsourcing business. I have read [Wikipedia:Blogs_as_sources](#) and I believe that our company is a source of authority in crowdsourcing and that I am an expert. But I want to call that out from the beginning.

I would like to add a paragraph to the bottom of the [Crowdsourcing#Modern_methods](#) section. It could also go into [Crowdsourcing#Crowdvoting](#) but it is related to modern methods, not examples. My addition:

Crowdsourcing often allows participants to rank each other's contributions, e.g. in answer to the question "What is one thing we can do to make Acme a great company?". One common method for ranking is "like" counting, where the contribution with the most likes ranks first. This method is simple and easy to understand, but it privileges early contributions, which have more time to accumulate likes. In recent years several crowdsourcing companies have begun to use [pairwise comparisons](#), backed by ranking algorithms such as [Elo](#). Ranking algorithms do not penalize late contributions. They also produce results faster. Elo, for example, is at least 10 times faster than manual stack ranking. REF One drawback, however, is that ranking algorithms are more difficult to understand than like counting.

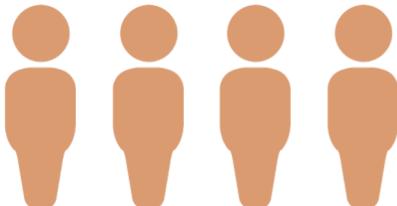
Many thanks, TDQuigley ([talk](#)) 17:50, 26 May 2017 (UTC)

The main concern I have with the content are the advert mentions and referencing source. Comments "backed by ranking algorithms such as Elo" can be simplified to "backed by ranking algorithms"; and "Elo, for example ..." could be more generically stated in terms of ranking algorithm performance improvement over manual methods. Lastly, the only source provided is a primary source; per [WP:RS](#), we should instead use an industry white paper, journal, news article, or similar third-party reliable sources. --- Barek ([talk](#) • [contribs](#)) - 23:26, 30 May 2017 (UTC)

Another example: Wikipedia



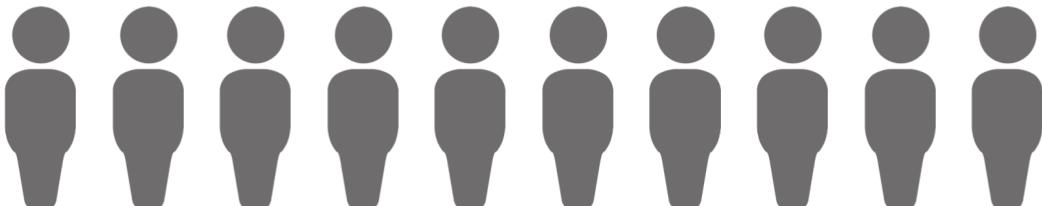
1211 Administrators



>1.4 Million Confirmed editors



>34 Million registered editors!

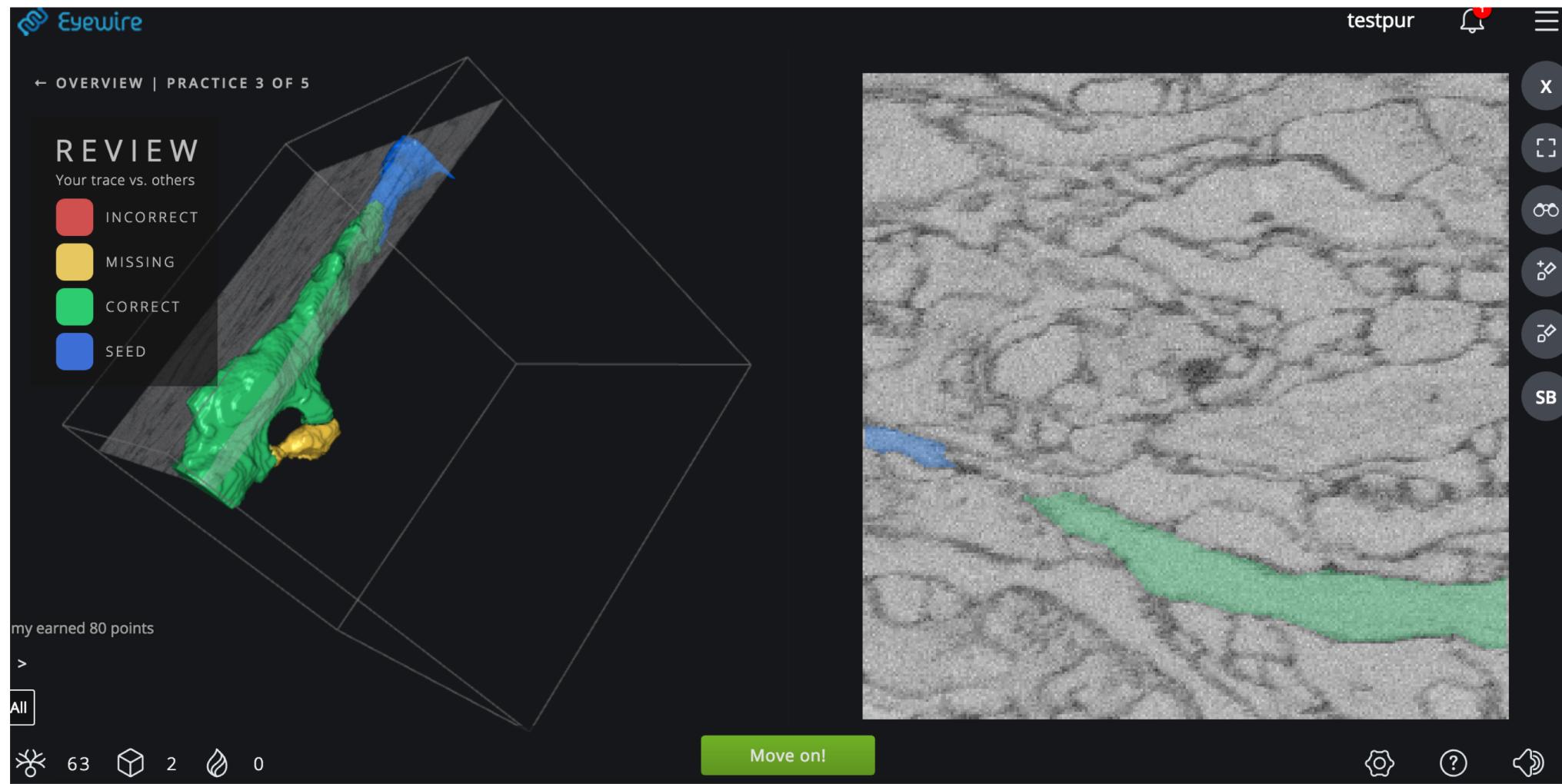


Everyone is editor!

The Wikipedia Article for 2011 Tohoku earthquake

- Within **1 hour**: 3 paragraph article
- Within **24 hours**: **>1,500** edits, an article with maps, photos and 79 references
- Within **10 weeks**: **>5,100** edits by **>1,200** users, an article with 289 references

EyeWire: Reconstruct 3D models of neuron



Inputs



“morning overooks”

Outputs

“Crowdsourcing”



Crowdsourcing
From Wikipedia, the free encyclopedia

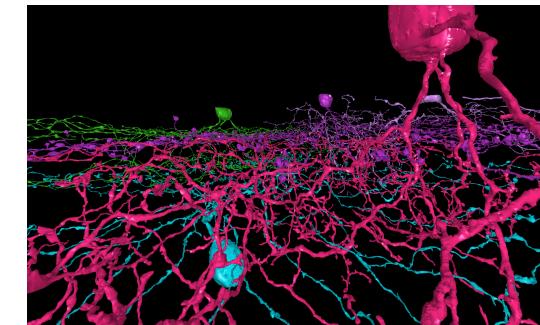
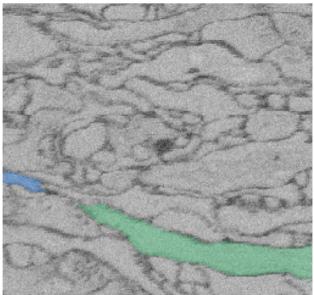
Main page **Contents** **Featured content** **Current events** **Random article** **Donate to Wikipedia** **Wikipedia store**

Interaction **Help** **About Wikipedia** **Community portal** **Recent changes** **Contact page**

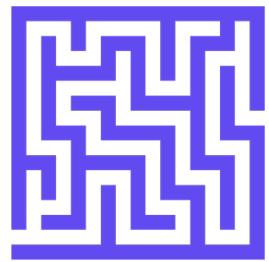
Tools **What links here** **Related changes** **Upload file** **Special pages**

Contents [hide]

- 1 Definitions
- 2 Historical examples
 - 2.1 Timeline of major events
 - 2.2 Early competitions
 - 2.3 In astronomy



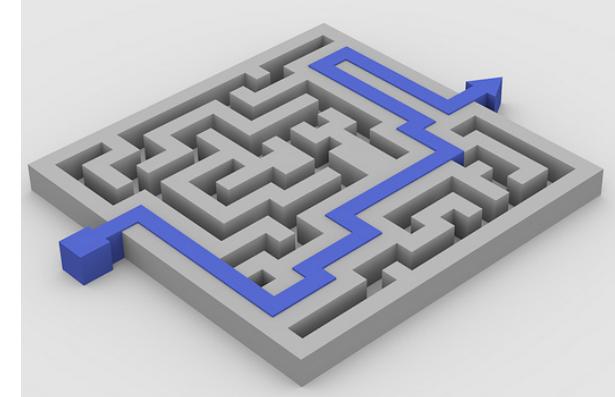
Common Pattern: Humans as the Computing Units



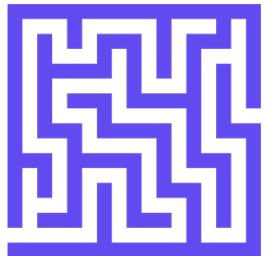
Inputs



Algorithms & Systems



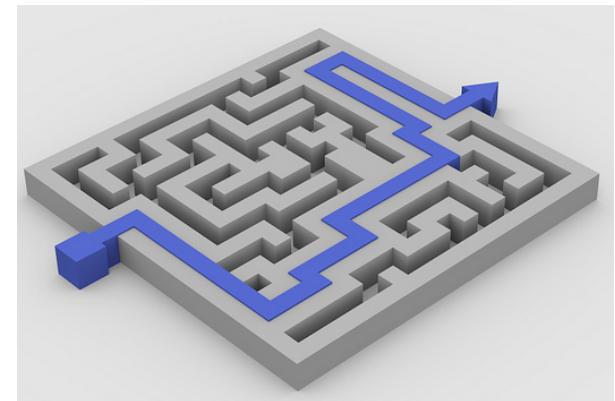
Outputs



Inputs

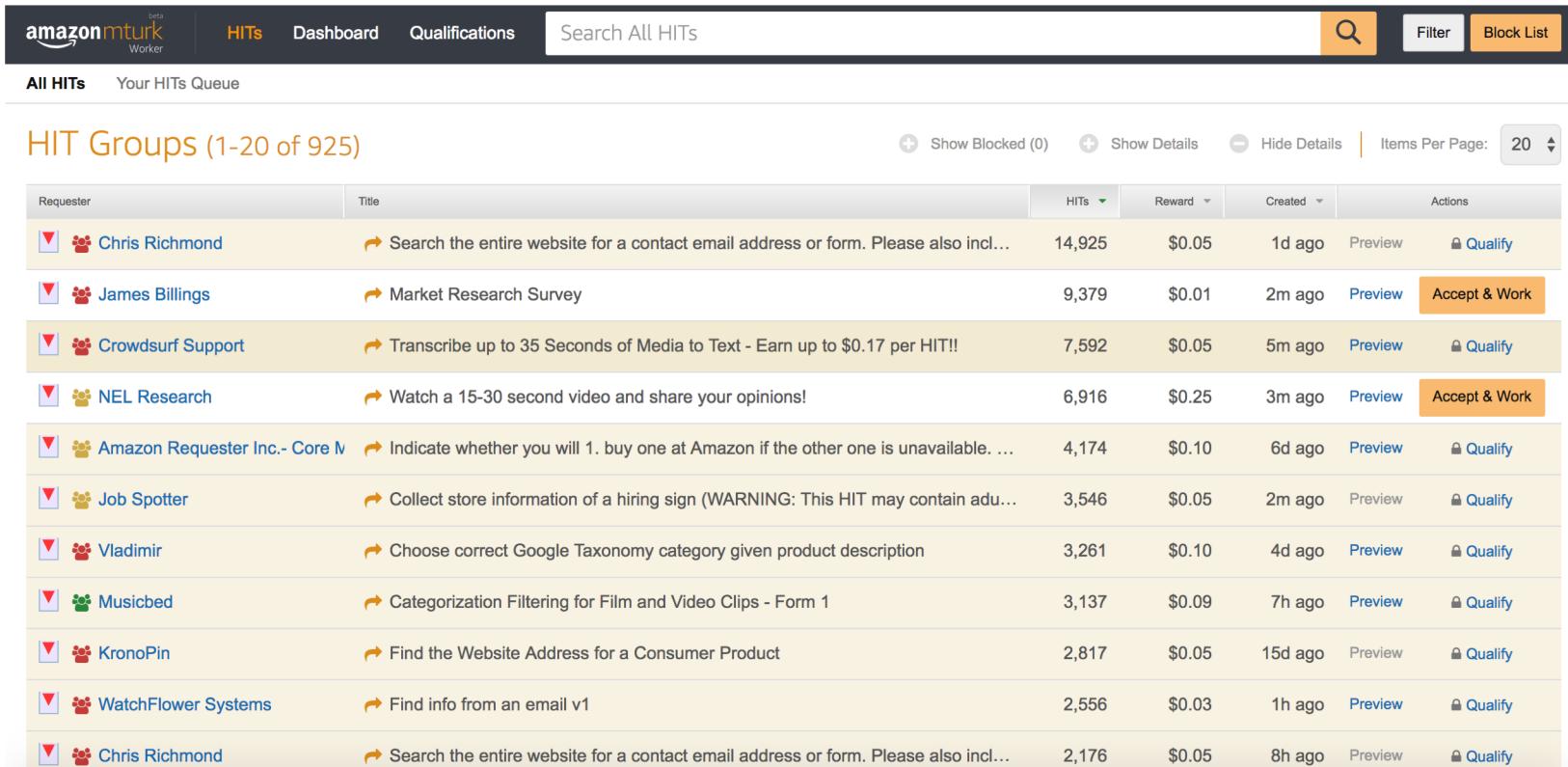


Platforms & Mechanisms



Outputs

Generalization: Crowdsourcing Markets



The screenshot shows the Amazon MTurk Worker interface. At the top, there's a navigation bar with the MTurk logo, 'HITs', 'Dashboard', 'Qualifications', a search bar containing 'Search All HITs' with a magnifying glass icon, and buttons for 'Filter' and 'Block List'. Below the navigation is a sub-navigation bar with 'All HITs' and 'Your HITs Queue'. The main content area is titled 'HIT Groups (1-20 of 925)' and includes filtering options: 'Show Blocked (0)', 'Show Details', 'Hide Details', and 'Items Per Page: 20'. A table lists 10 HIT groups from various requesters:

Requester	Title	HITs	Reward	Created	Actions
Chris Richmond	Search the entire website for a contact email address or form. Please also incl...	14,925	\$0.05	1d ago	Preview Qualify
James Billings	Market Research Survey	9,379	\$0.01	2m ago	Preview Accept & Work
Crowdsurf Support	Transcribe up to 35 Seconds of Media to Text - Earn up to \$0.17 per HIT!!	7,592	\$0.05	5m ago	Preview Qualify
NEL Research	Watch a 15-30 second video and share your opinions!	6,916	\$0.25	3m ago	Preview Accept & Work
Amazon Requester Inc.- Core N	Indicate whether you will 1. buy one at Amazon if the other one is unavailable. ...	4,174	\$0.10	6d ago	Preview Qualify
Job Spotter	Collect store information of a hiring sign (WARNING: This HIT may contain adu...	3,546	\$0.05	2m ago	Preview Qualify
Vladimir	Choose correct Google Taxonomy category given product description	3,261	\$0.10	4d ago	Preview Qualify
Musicbed	Categorization Filtering for Film and Video Clips - Form 1	3,137	\$0.09	7h ago	Preview Qualify
KronoPin	Find the Website Address for a Consumer Product	2,817	\$0.05	15d ago	Preview Qualify
WatchFlower Systems	Find info from an email v1	2,556	\$0.03	1h ago	Preview Qualify
Chris Richmond	Search the entire website for a contact email address or form. Please also incl...	2,176	\$0.05	8h ago	Preview Qualify

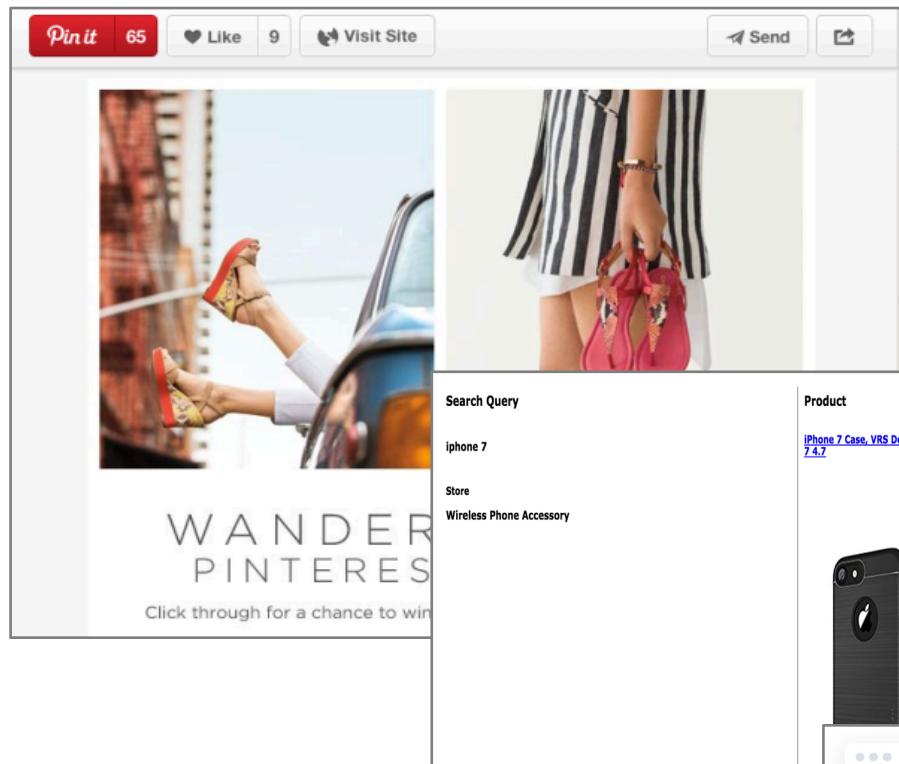
Amazon's Mechanical Turk (MTurk): <https://www.mturk.com/>

Generalization: Crowdsourcing Markets

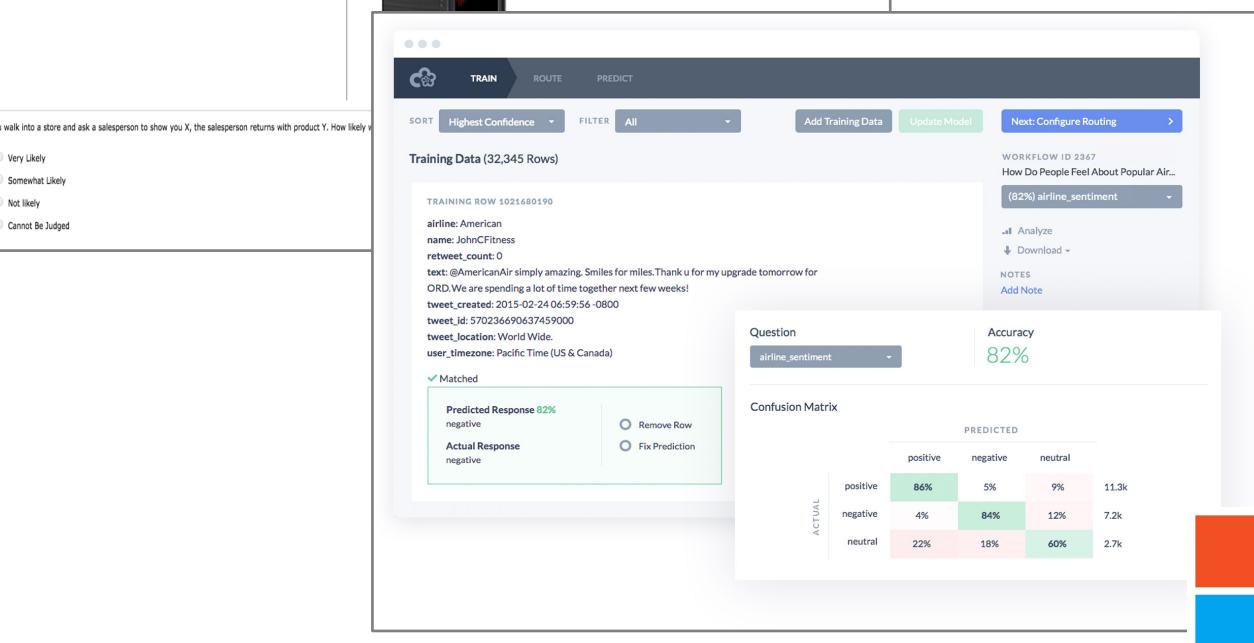
The screenshot shows the CrowdFlower website interface. At the top, there is a navigation bar with links for Home, Tasks, My Account, and Logout. Below the navigation bar, a dark header bar displays the text "Tasks available". Underneath this, there are buttons for "Open task listing in a new window", "View completed tasks", and "Help". A "Available Jobs" section is displayed, featuring a table with columns: JOB, JOB TITLE, LEVEL, PAY / TASK, # OF TASKS, and RATING. The table contains three rows of data:

JOB	JOB TITLE	LEVEL	PAY / TASK	# OF TASKS	RATING
1290384	*Validate Product Name Simplifications *		\$0.01	105	●●●●●
1281169	Find Earning Release Date		\$0.05	16	●●●●●
1294336	Identify Background Noise In Audio		\$0.02	8	●●●●●

Figure Eight (previously known as CrowdFlower): <https://www.figure-eight.com/>



Content Filtering



Product Relevance

Sentiment Analysis

A Lot of Questions to Ask



Who are the people and what does the process look like?

Aug 27, 29, September 5, 10

A Lot of Questions to Ask



How to motivate people?



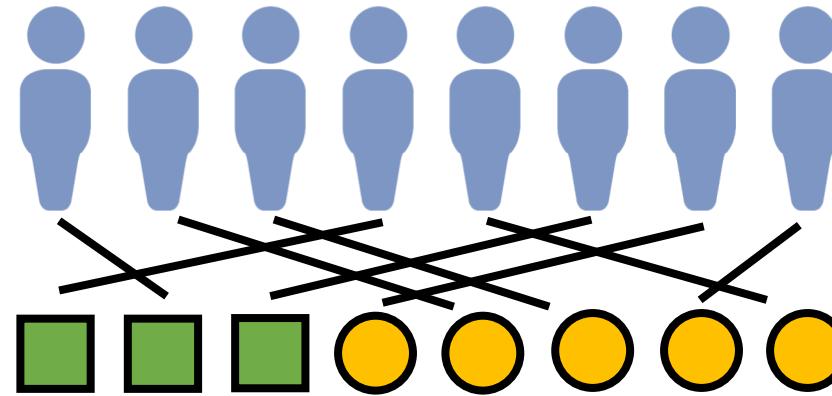
HIT Groups (1-20 of 925)						
Repeater	HITs	Reward	Created	Actions	Details	Block
Chris Richmond	14,925	\$0.05	1d ago	Preview	Accept & Work	
James Billings	9,379	\$0.01	2m ago	Preview	Accept & Work	
Crowdsurf Support	7,592	\$0.05	5m ago	Preview	Accept & Work	
NEL Research	6,916	\$0.25	3m ago	Preview	Accept & Work	
Amazon Requester Inc - Core k	4,174	\$0.10	6d ago	Preview	Accept & Work	
Job Spotter	3,548	\$0.05	2m ago	Preview	Accept & Work	
Vladimir	3,261	\$0.10	4d ago	Preview	Accept & Work	
Musiced	3,137	\$0.09	7h ago	Preview	Accept & Work	
KnozPin	2,817	\$0.05	15d ago	Preview	Accept & Work	
WatchFlower Systems	2,506	\$0.03	1h ago	Preview	Accept & Work	
Chris Richmond	2,176	\$0.05	8h ago	Preview	Accept & Work	

September 12, 17, 19

A Lot of Questions to Ask



How to assign right tasks to right people?



September 24

A Lot of Questions to Ask



**How to aggregate results from multiple people
and control the quality of aggregation?**



Husky!



Malamute!

?



Husky!

September 26, October 1, 3

A Lot of Questions to Ask

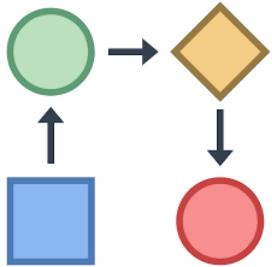


How to engage people?

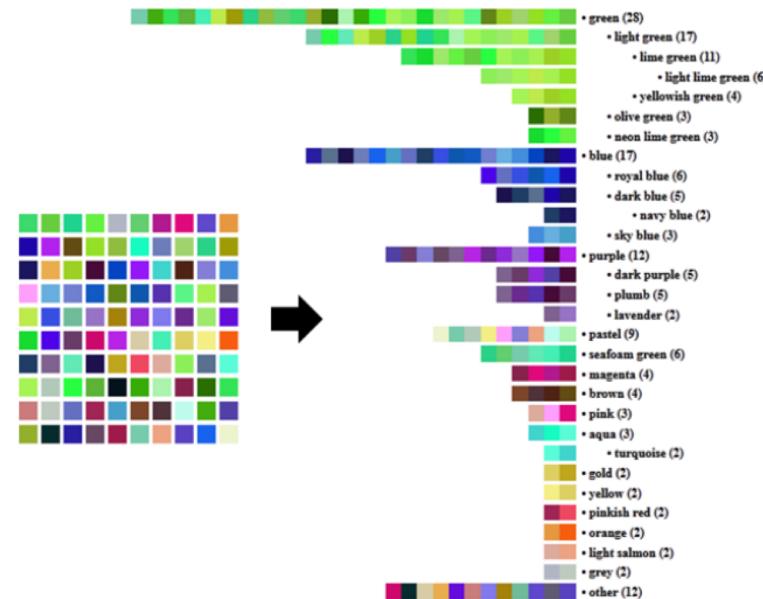
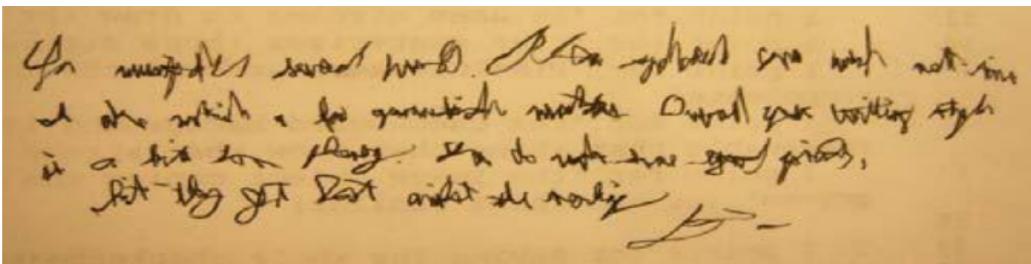


October 15, 17

A Lot of Questions to Ask

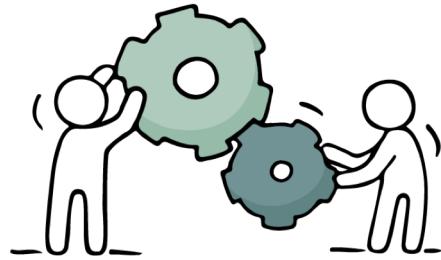


How to design effective workflows?

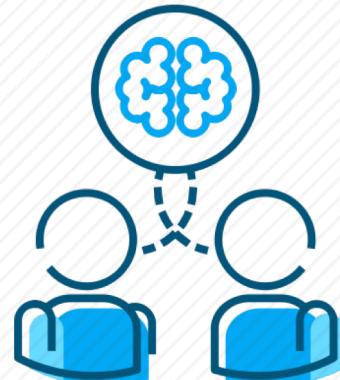


October 22, 24, 29

A Lot of Questions to Ask



**How to facilitate collaborative work
among people?**



October 31, November 12, 14

Should I be interested?

- Want to leverage the power of crowd
- Want to improve the outcome of your crowdsourcing effort
- Want to better understand the phenomenon of collective intelligence and explore innovative crowdsourcing applications
- Interested in research in crowdsourcing, social computing, HCI, applied AI and ML, computational social science

Getting to Know You

- Name/Program
- Why taking this course?
- What topics you are particularly interested in learning about in this course?

Course Format and Logistics

- Time and Location: Mon & Wed, 4:30-5:45pm, LWSN B134
- Course Website: <http://mingyin.org/CS590/Fall2018/index.html>
- Email: crowdsourcing.purdue@gmail.com
- Two major components: paper reading and discussion, final project

Paper Reading and Discussion

Aug 22	Crowdsourcing: Background and applications <i>Lecture</i>	<p>Required</p> <p>Howe. The Rise of Crowdsourcing. WIRED, June 2006.</p> <p>Howe. Crowdsourcing: A Definition. June 2006.</p> <p>Optional</p> <p>von Ahn et al. reCaptcha: Human-based Character Recognition via Web Security Measures. Science, September 2008</p> <p>von Ahn and Dabbish. Labeling Images with a Computer Game. CHI'04</p> <p>Cooper et al. Predicting Protein Structures with a Multiplayer Online Game. Nature, August 2010.</p>	<p>Assignment 1: Due before next class (Aug 26)</p>
--------	---	---	--

- The first 6 classes (until September 10) will be led by me
- Later classes will be led by you! So sign-up for topics you are interested in presenting early!
- Presentation assignment will be out around September 5

Paper Reading and Discussion

- Use Piazza for paper discussion before class: <https://piazza.com/purdue/fall2018/cs59000csc/home>
- First 6 classes
 - I will post reading questions on Piazza
 - Read at least the required paper(s), and answer reading questions on Piazza before midnight of the class day

Paper Reading and Discussion

- Presenters
 - Read all required paper(s) and at least one optional paper for the assigned class
 - Discuss with the instructor (one week before class) about the reading questions, and post them on Piazza
 - Give a presentation, read through classmates' reading question responses and lead the discussion in class
- Non-presenters
 - Read the required paper(s) and answer the reading questions on Piazza before midnight of the class day
 - Participate in the discussion in class!

Final Project

- An opportunity for you to get hands-on experience in crowdsourcing and social computing research. Projects are open-ended, and you are encouraged to connect the final project with your own research.
- (Tentative) key dates:
 - **October 3:** Final project proposal due
 - **October 10:** Final project pitch day
 - **Nov 5-7:** Milestone presentation, getting feedback from peers
 - **Dec 3-5:** Final project presentation
 - **Dec 8:** Final project report (10 pages maximum) due

Grading

- Assignment: 10%
- Reading responses: 15%
- Class discussion: 10%
- Paper presentation: 20%
- Final project: 45% (proposal + pitch: 10%, milestone presentation: 10%, final presentation: 10%, final report: 15%)

Some Rules

- Show up
- Engage
- Think Research

Things to do next...

- Sign up to Piazza: <https://piazza.com/purdue/fall2018/cs59000csc/home>
- Register as a worker of Amazon Mechanical Turk (MTurk)/Figure Eight today
 - MTurk: <https://www.mturk.com/>
 - Figure Eight: <https://www.figure-eight.com/>
 - We will ask you to be a worker yourself and complete tasks on crowdsourcing markets in Assignment 1
 - MTurk: After you successfully register, Amazon will send you invitation for working, which may takes 3 business days...

Next Class

- Crowdsourcing: Background and applications
- Required Reading:
 - Howe. [The Rise of Crowdsourcing](#). WIRED, June 2006.
 - Howe. [Crowdsourcing: A Definition](#). June 2006.
- Optional Reading:
 - von Ahn et al. [reCaptcha: Human-based Character Recognition via Web Security Measures](#). Science, September 2008
 - von Ahn and Dabbish. [Labeling Images with a Computer Game](#). CHI'04
 - Cooper et al. [Predicting Protein Structures with a Multiplayer Online Game](#). Nature, August 2010.
- No reading question for next class!