GE 461
Crowdsourcing

Spring 2022

What is Crowdsourcing?

"the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers."
Merriam-Webster definition

The term was coined by Jeff Howe in 2006 in the following WIRED article (recommended reading):
https://www.wired.com/2006/06/crowds/

Talks about several examples of *crowd* generated *cheap* data sources.

So diverse, you might not have noticed an application is related to crowdsourcing.

Early Crowdsourcing Examples by Howe

Museum director needs images of sick people for an exhibition.
Professional photographer charges a bargain price of $150 per piece.
She gets 56 pics from iStockPhoto each $1 per piece.

Today Wikimedia Commons is a widely used platform for public images.

Repackaging of the Internet content.

VH1 Web Junk 20: popular videos from YouTube on TV.

Low-Cost-High-Rating.

OynatBakalim on TV8 today!

Company as a Platform

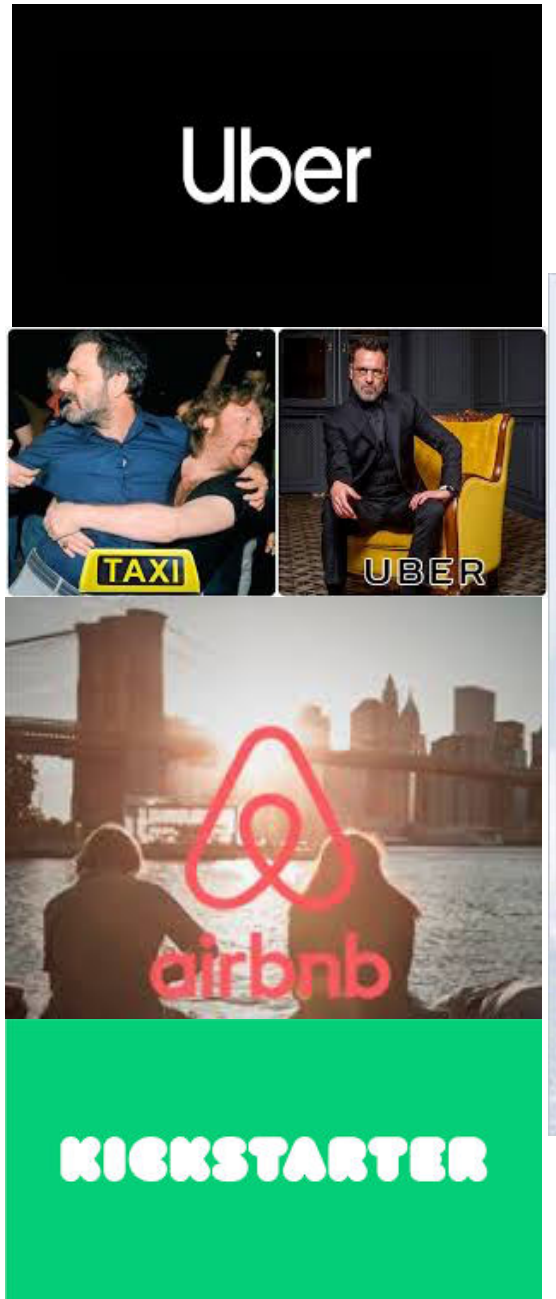
Uber is a taxi company that does not own a single cab.

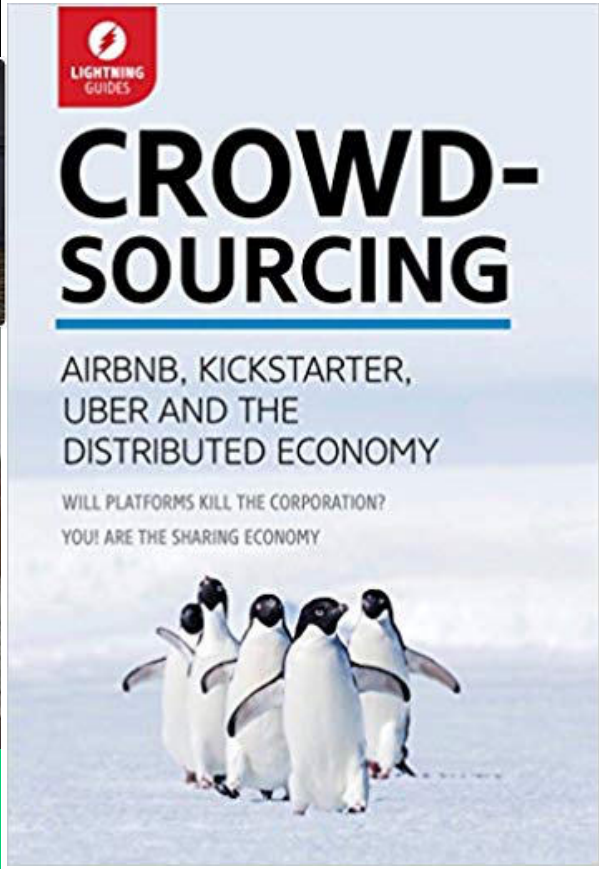
- Everyone with a car can be a cab driver in their off times.
- This brings up questions about liability.
- Now has research centers for autonomous driving.

AirBNB is a company that does not own a single room.

Kickstarter provides a platform for inventors to get their products funded by enthusiasts.

- Make the product possible and earn discounts





Company as a Platform – cont'd

Crowdsourced marketplace

Craigslist: Old but gold, since 1996. Inspired by garage sales

eBay

LetGo

gittigidiyor.com

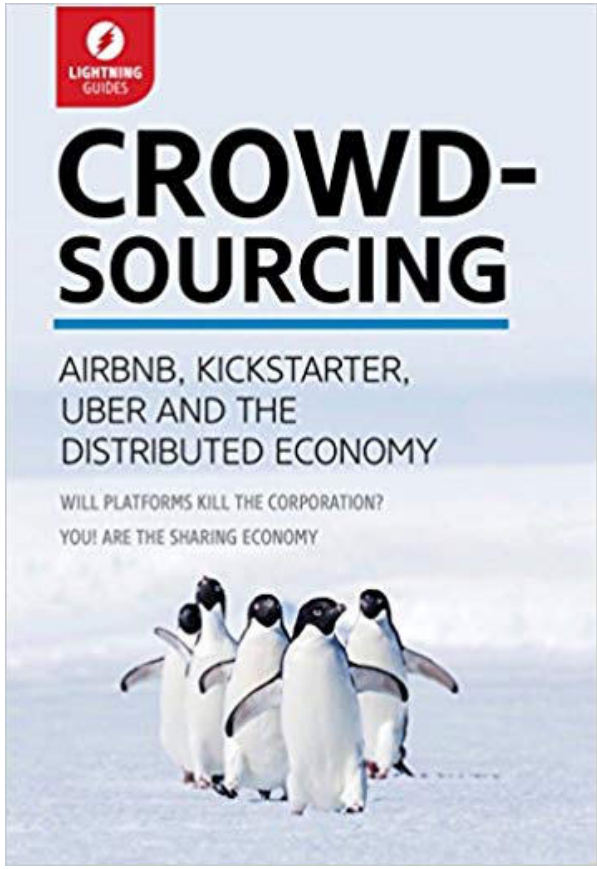
sahibinden.com

Shopping websites now also provide a marketplace for individuals.

Amazon

Alibaba

n11.com



Collaborative Effort

WikiPedia

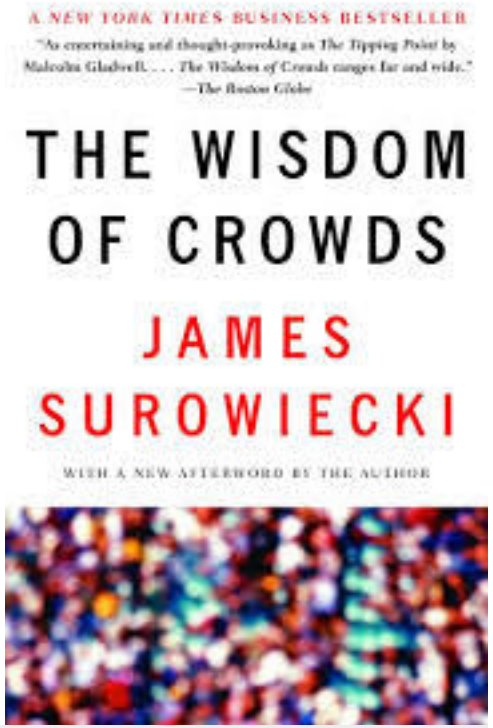
A living encyclopedia.

wiki + encyclopedia.

Written collaboratively by volunteers. No pay.

As of January 2020, active in 309 languages.

Roughly 68k contributors and 48m articles. 6m in English.



Collaborative Effort – cont'd

Waze

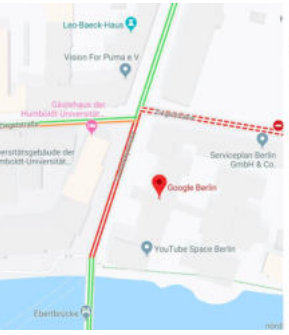
community-driven (Wazers) GPS navigation app.

turn-by-turn navigation information + uses user-submitted information

can report accidents, traffic jams, speed and police traps.

can modify the map data.

Now a Google company.



An Artist Used 99 Phones to Fake a Google Maps Traffic Jam. B. Barrett. wired.com

GADGETS

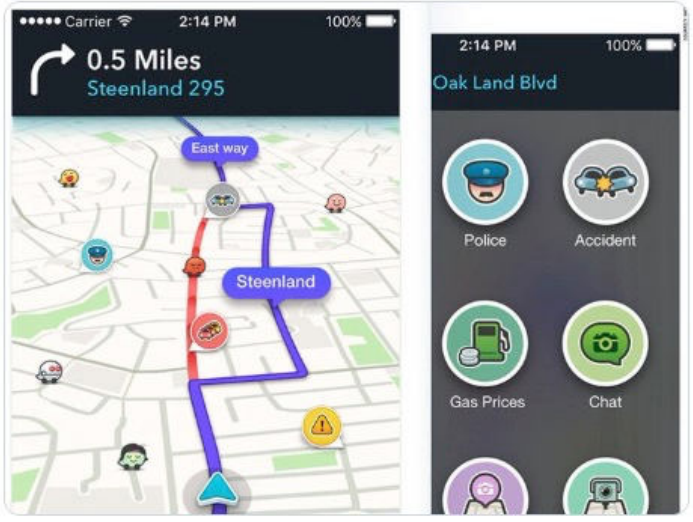
Israeli Students Spoof Waze App With Fake Traffic Jam

The future of cyber attacks is mildly annoying.

By Kelsey D. Atherton
March 31, 2014

CNN
@CNN

The NYPD, the nation's largest police force, is demanding Google stop allowing users to post DUI checkpoint data on its live traffic and navigation application, Waze cnn.it/2DmDsDv



11:49 AM · Feb 8, 2019 · SocialFlow

Crowdsourcing for Science

Protein folding problem

Predicting the 3D structure of a protein using 1D sequence.

Have people play a game instead [http://fold.it]

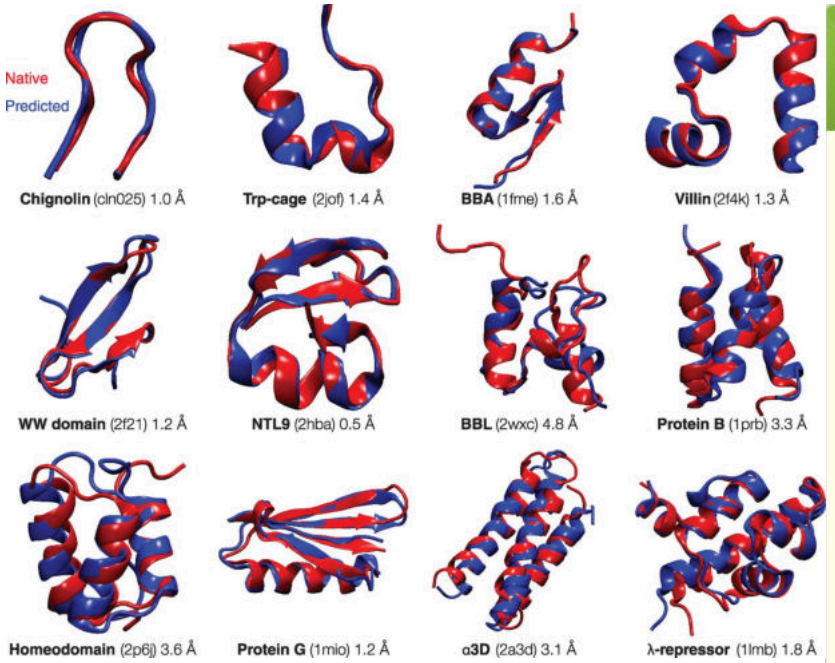
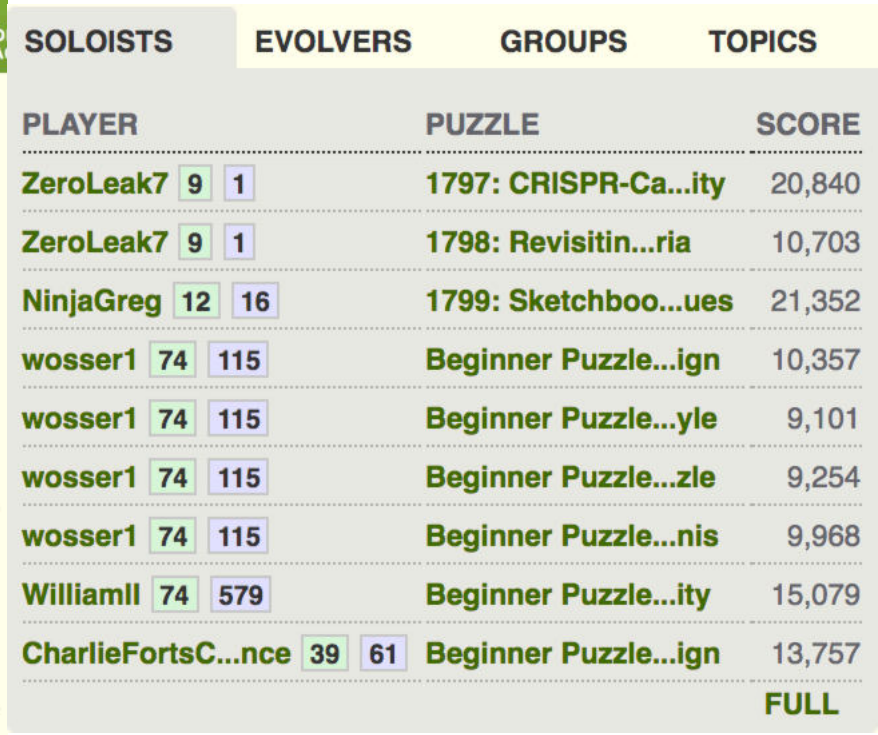


Image: Dill et al., Science 23 Nov 2012:
Vol. 338, Issue 6110, pp. 1042-1046
DOI: 10.1126/science.121902



Crowdsourcing for Science – cont'd

Antibiotic resistant gene curation.

5k genes are in current databases.

Very hard to curate and have a consensus in distributed databases.

ARGminer provides a central platform for all scentists to collaboratively work and have a consensus database.

ARGminer: A web platform for crowdsourcing based curation of antibiotic resistance genes

G. A. Arango-Argoty[1], G. K. P. Guron[2,3], E. Garner[2], M.V. Riquelme[2], L. S. Heath[1], A. Pruden[2], P. J. Vikesland[2], and L. Zhang[1*]

Bioinformatics, btaa095, https://doi.org/10.1093/bioinformatics/btaa095

http://bench.cs.vt.edu/argminer

Crowdsourcing **of** Science

Seti@Home

Help searching for extraterrestial life by donating your CPU.

".. uses Internet-connected computers in the Search for Extraterrestrial Intelligence (SETI). You can participate by running a free program that downloads and analyzes radio telescope data."

Now, you can also search for new planets by going through telescope images of NASA. [https://blog.backyardworlds.org/]

Unconventional Examples

Jim Gray was a noted computer scientist working for Microsoft. He disappeared while sailing alone near Farallon Islands off the CA shore.

Democratization of the search and rescue

Blog named *Tenacious Search* for coordination.

A network of friends have analyzed thousands of satellite (Quickbird) and plane (ER-2 of NASA) imagery.

Redundant viewing of 300x300 pixel sub images

A subgroup of experts with remote sensing filtered false-positives.

Joseph M. Hellerstein, David L. Tennenhouse
Communications of the ACM, July 2011, Vol. 54 No. 7, Pages 77-87

OK, crowdsourcing is nice, but how does it relate to Data Science?

Crowdsourcing for Data Science Problems

We are data rich, but really?

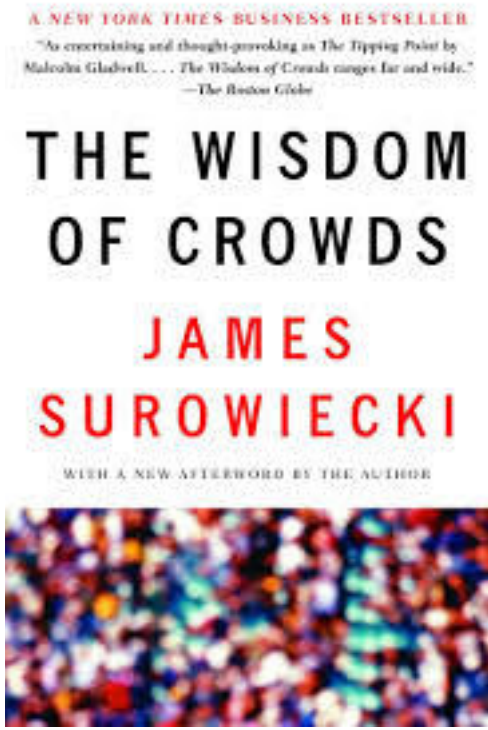
Label crisis:

10 billion images on Google in 2010*!

The largest labeled dataset is ImageNet with 14m labeled images.

How to collect data?

- Collect data via software (e.g., crawl)
  - Unstructured.
  - Hard to associate with a label.
- Set up an experiment, collect participants, pay a fee.
  - Slow
  - Sample bias
  - Small data size.
- Can crowdsourcing help?



*Google VP https://techcrunch.com/2010/07/20/google-image-search/

Crowdsourcing for Data Collection & Labeling

Mechanical Turk [https://www.mturk.com]

> "… is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs."
>
> Various Human Intelligence Tasks (HITs)

amazon mechanical turk

Image: brittannica.com

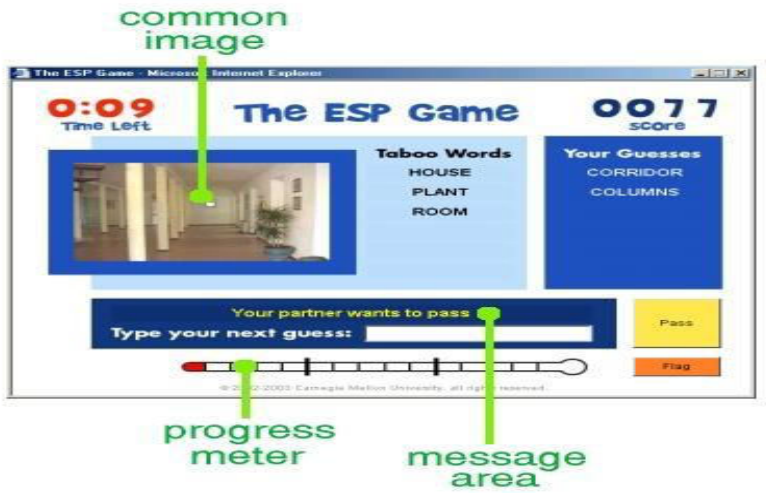
Crowdsourcing for Data Collection & Labeling – cont'd

A career on crowdsourcing, Luis von Ahn.

PhD in Computer Science at Carnegie Mellon University in 2005

Game with a purpose: have a game/puzzle that is easy for humans to solve but hard for computers. Also entertaining!

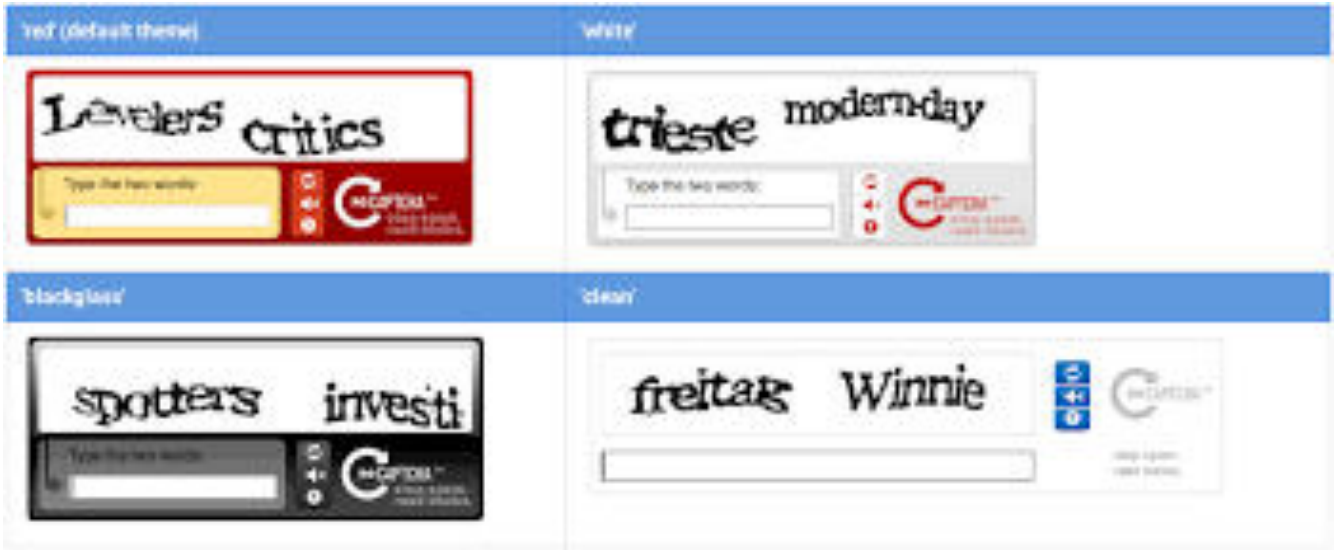
ESP game: Two people try to assign the same label to an image.



Crowdsourcing for Data Collection & Labeling – cont'd

Captcha: are you human?

ReCaptcha: are you human? I can make use of your proof.



Crowdsourcing for Data Collection & Labeling – cont’d

DuoLingo: You teach me while I play a game, I will teach you by translating sentences from CNN and BuzzFeed.

94 courses in 23 languages as of today. 300m users.

Crowdsourcing **of** Data Scientists

InnoCentive [https://www.innocentive.com]

Kaggle [https://www.kaggle.com]

Firms set up challenges to freelance experts to solve R&D problems.

Expertise from within and outside of the industry.

Prizes to winners, IPs retained by the company.

A community of ML experts.

Datasets available to play with (in Kaggle).

Crowdsourcing **of** Data Scientists – cont'd

FeatureHub

Hire experts to identify useful features to solve a problem.

Models built using these features are very close to Kaggle winners in just a matter of days.

Smith, M.J., Wedge, R. and Veeramachaneni, K., 2017, October. FeatureHub: Towards collaborative data science. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 590-600). IEEE.

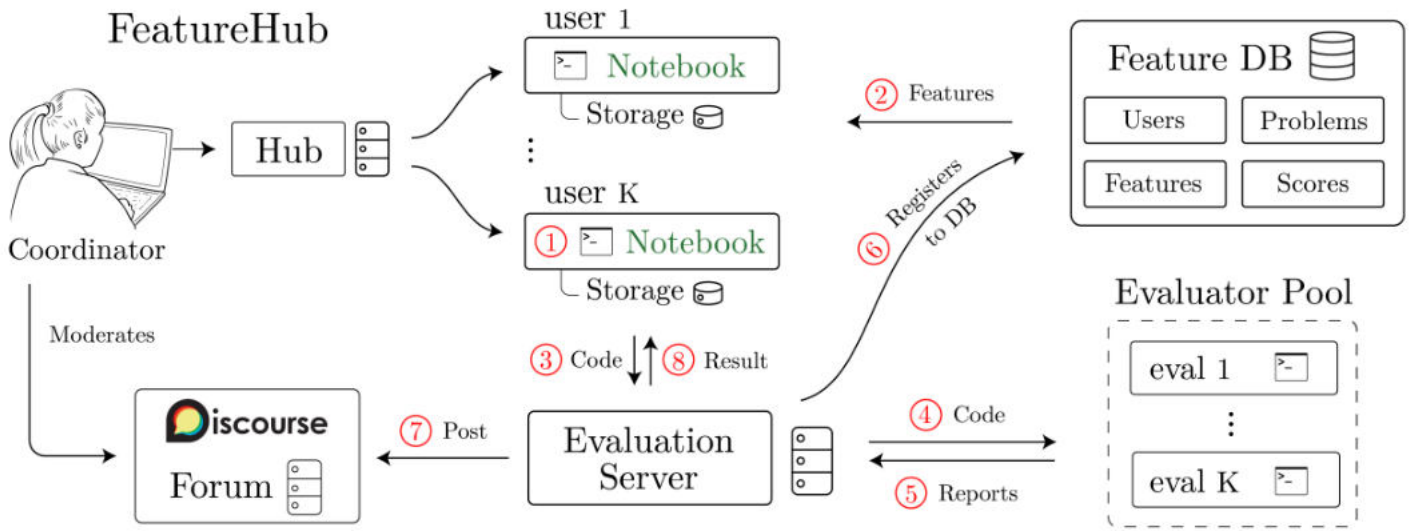


Fig. 4. Overview of FeatureHub platform architecture, comprising the FeatureHub computing platform and the Discourse-based discussion platform. A user writes a feature and evaluates it locally on training data. The user then submits their code to the Feature Evaluation server. The code is validated to satisfy some constraints and to ensure that it can be integrated without bugs into the predictive model. The corresponding feature is extracted and an automated machine learning module selects and trains a predictive model using the candidate features and other previously-registered features. The results are registered to the database, posted to the forum, and returned to the user.

Incentives for Crowdsourcing

- Earn money (Mechanical Turk)
- Have fun (ESP, fold.it)
- Socialization (Kaggle, Innocentive)
- Recognition (Kaggle, Innocentive ,Wikipedia)
- Do good (wikipedia, search for Jim Gray, Seti@home)
- Learn something new (Kaggle, Innocentive, Duolingo)
- Science (fold.it, seti@home, ARGminer)

Slide based on M. Lease, UT Austin