

# Crowdsourcing

MICHAEL BERNSTEIN  
CS 376

# Announcements

- Idea brainstorm feedback — tomorrow night
  - We'll identify your four best ideas and focus grading on those
- Abstract v1 due Friday — pick a project!
- If `group_size != 3`, please chat with Rob after class

# Announcements

- Michael @ Yom Kippur + UIST starting Wednesday
  - So, no office hours this week or next
  - But not out of email contact! Please make liberal use of [cs376@cs.stanford.edu](mailto:cs376@cs.stanford.edu) to get feedback.
- Upcoming:
  - Wendy Ju, human-robot interaction
  - Rob Semmens, research methods
  - Jeff Hancock, computational social science

<http://hci.st/wise>

grab your phone, fill it out









**COORDINATION NEGLECT: HOW LAY  
THEORIES OF ORGANIZING  
COMPLICATE COORDINATION IN  
ORGANIZATIONS**

**Out of Sight, Out of Sync: Understanding  
Conflict in Distributed Teams**

**The Mutual Knowledge Problem and Its  
Consequences for Dispersed Collaboration**

**The team scaling fallacy: Underestimating the declining efficiency of larger teams**

**Who's in Charge Here? How Team Authority Structure Shapes Team Leadership**

**Team Familiarity, Role  
Experience, and  
Performance: Evidence from  
Indian Software Services**

**The Influence of Shared Mental Models on Team Process and Performance**

**Some unintended consequences of  
job design**

**Structure and Learning in Self-Managed Teams:  
Why "Bureaucratic" Teams Can Be Better Learners**

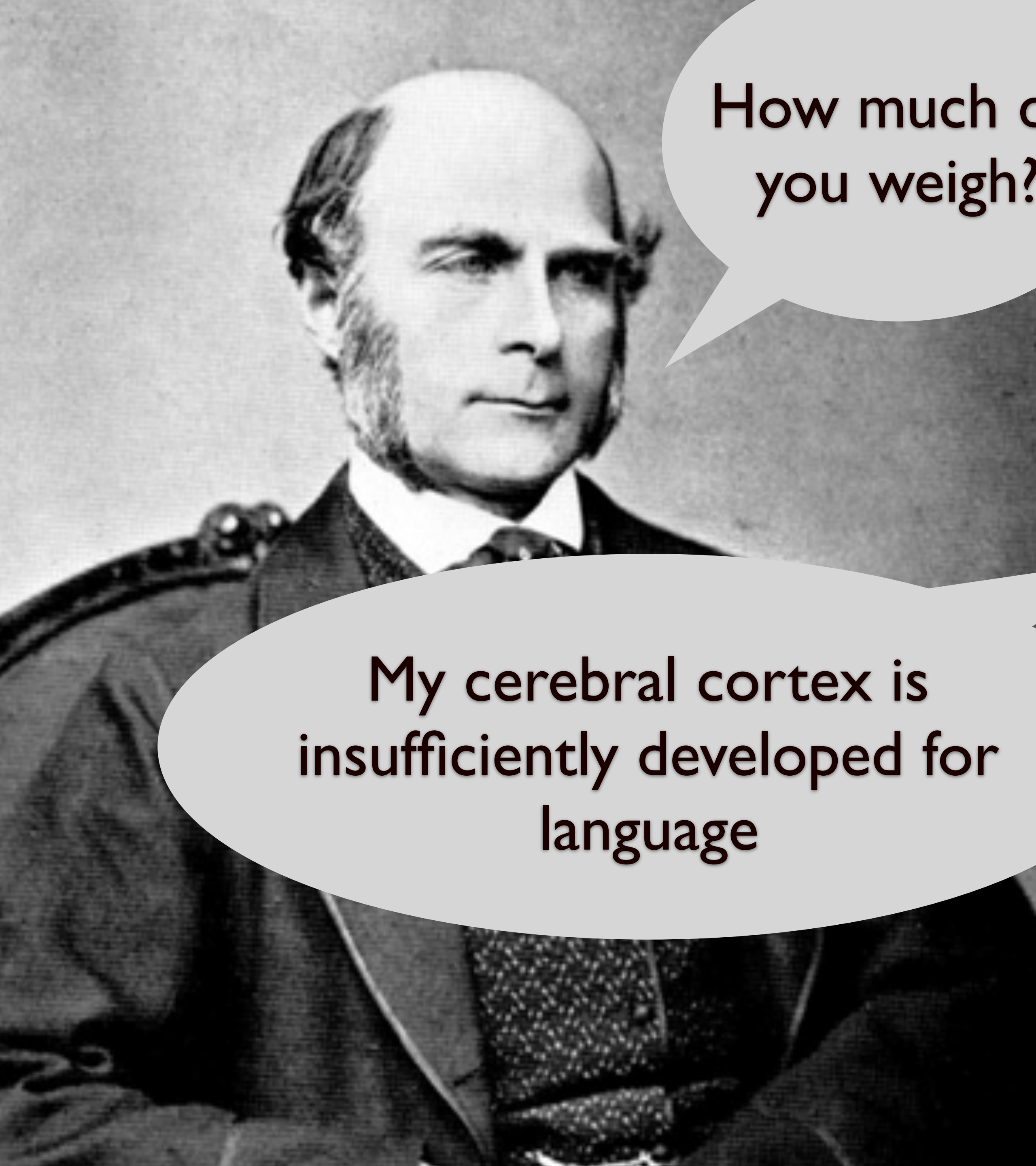






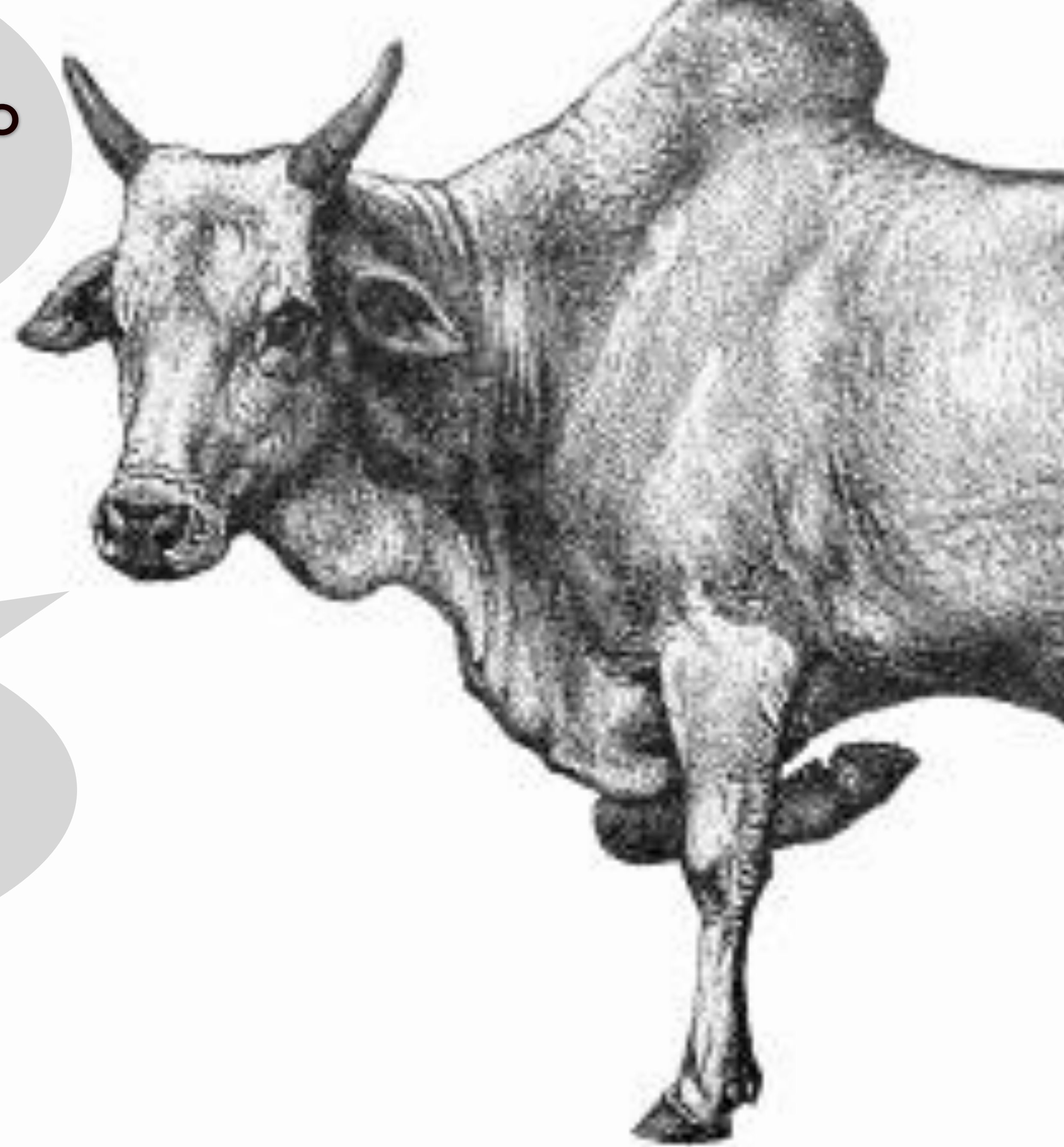
How might computing connect large groups to tackle bigger, harder problems than they could complete in isolation?



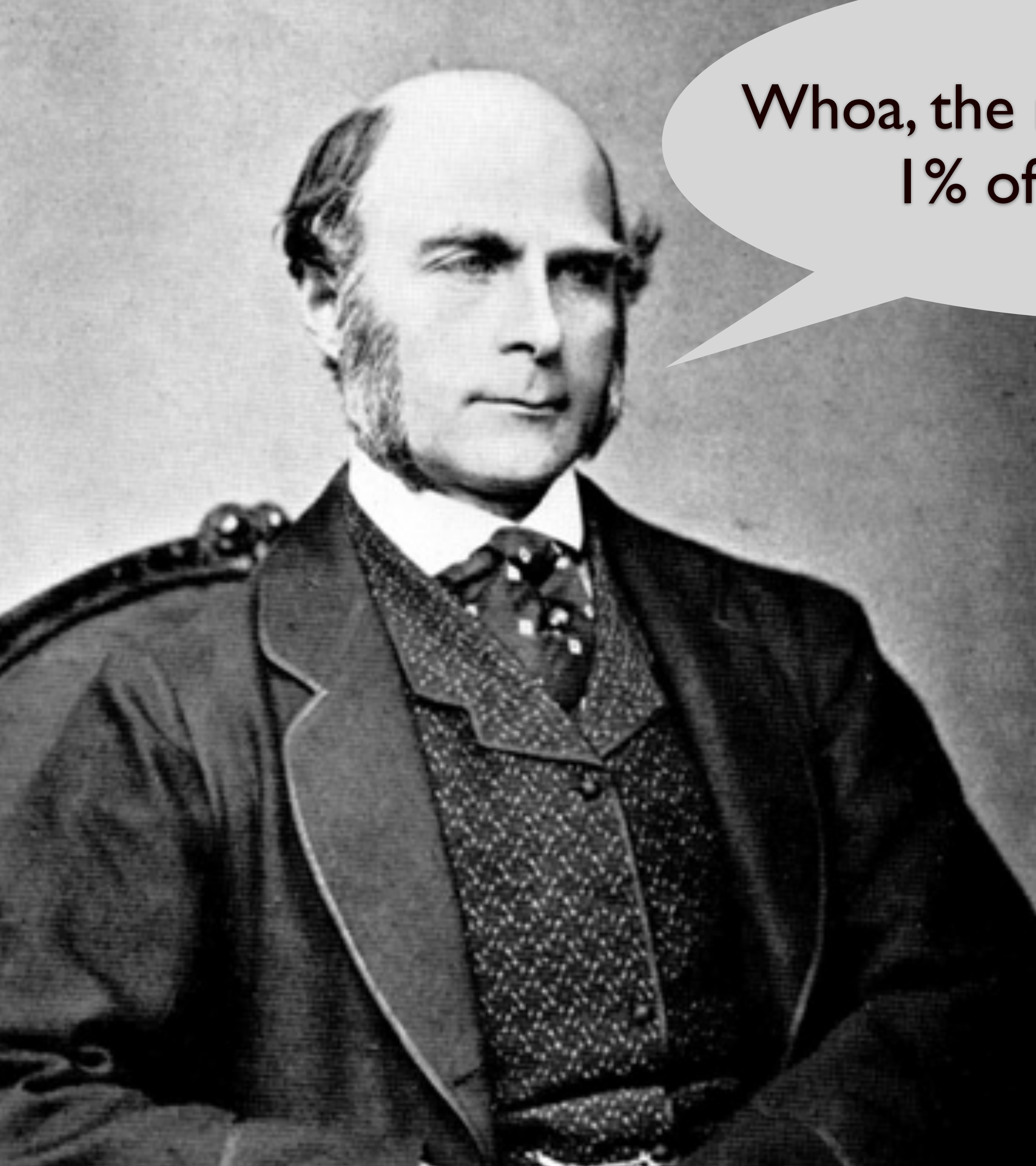


How much do  
you weigh?

My cerebral cortex is  
insufficiently developed for  
language







Whoa, the mean guess is within 1% of the true value

of the dressed weight of a 787 different persons.

		Percentiles			
		Observed deviates from 1207 lbs.	Normal p.e = 37	Excess of Observed over Normal	
	5	1074	- 133	- 90	+ 43
	10	1109	- 98	- 70	+ 28
	15	1126	- 81	- 57	+ 24
	20	1148	- 59	- 46	+ 13
<i>q</i> <sub>1</sub>	25	1162	- 45	- 37	+ 8
	30	1174	- 33	- 29	+ 4
	35	1181	- 26	- 21	+ 5
	40	1188	- 19	- 14	+ 5
	45	1197	- 10	- 7	+ 3
<i>m</i>	50	1207	0	0	0
	55	1214	+ 7	+ 7	0
	60	1219	+ 12	+ 14	- 2
	65	1225	+ 18	+ 21	- 3
	70	1230	+ 23	+ 29	- 6
<i>q</i> <sub>3</sub>	75	1236	+ 29	+ 37	- 8
	80	1243	+ 36	+ 46	- 10
	85	1254	+ 47	+ 57	- 10
	90	1267	+ 52	+ 70	- 18
	95	1293	+ 86	+ 90	- 4

*q*<sub>1</sub>, *q*<sub>3</sub>, the first and third quartiles, stand at 25° and 75° respectively.  
*m*, the median or middlemost value, stands at 50°.



Let's check our  
<http://hci.st/wise>  
results

# Early crowdsourcing research

[Grier 2007]

Two distributed workers work independently, and a third verifier adjudicates their responses

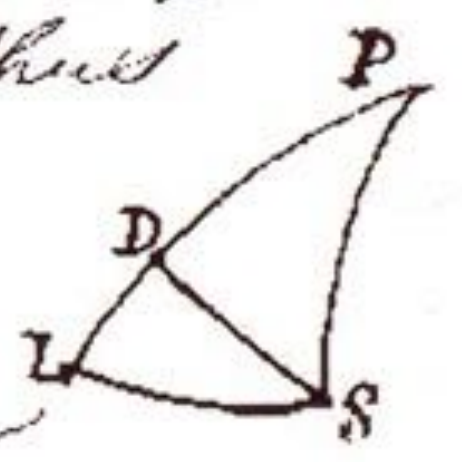


## 1760

British Nautical Almanac  
Neil Maskelyne

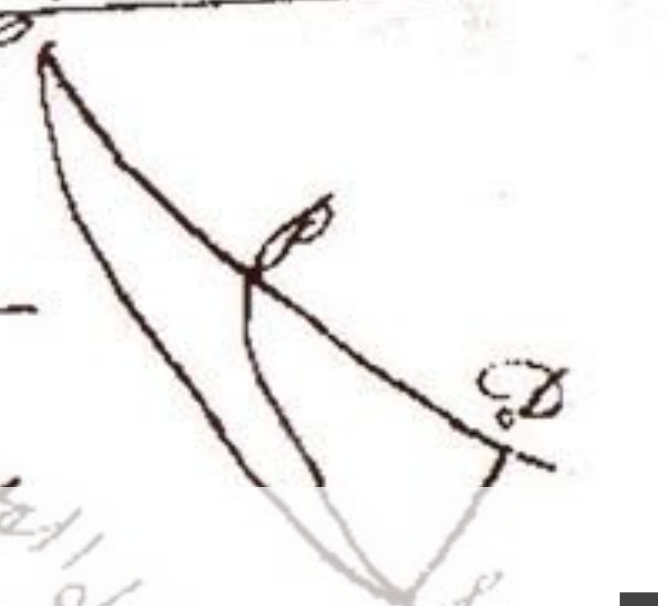


In answer to your letter of the 23. instant  
 Compute D: dist. from a star by Logarithms, thus  
 P the pole of the ecliptic, N or S. P, P, are  
 the stars & m's distances from one of the same pole.  
 & the D, S the star, S D a perpendicular  
 great circle. Let fall from S to PL, perpendicular  
 to it in T. There suppose P to be less than 90.  
 By Log<sup>s</sup>. c, P+t, PS = t, PD & PL - PD = LD  
 c, LS = -c, PD+c, PS+c, LD  
 or = c, P+s, PS-s, PD+c, LD



The latter formula must be used when P:D  
 is large, or near 90; but may be used safely  
 in all cases. —

Example  
~~P 20 30 - c - 9.9715876~~  
~~PS 89 58 30 t - 13.3857588~~  
 P 25 25 30 - c - 9.9557589  
 PS - - - - t 12.0977240 - - - - s - 9.9999879  
 PD 89 29 36 - t 12.0534829 - - - - s. co-ar. - 0.0000170  
 PL 85 0 0 - - - - c - 9.9986631  
 LD - 4 29 36 - - - - c - 9.9544269  
 SL 25 47 30



If P should be greater than 90  
 then PL + PD = LD as in this scheme

As the planets places, excepting the  
 Jun and man are only set down to minutes,  
 there is no occasion to allow for nutation &  
 aberration in computing the distances of  
 stars by the D you must call it the distance  
 to planet 1. and apply nutation & aberration Jan 7 lead to  
 Cambridge. I am

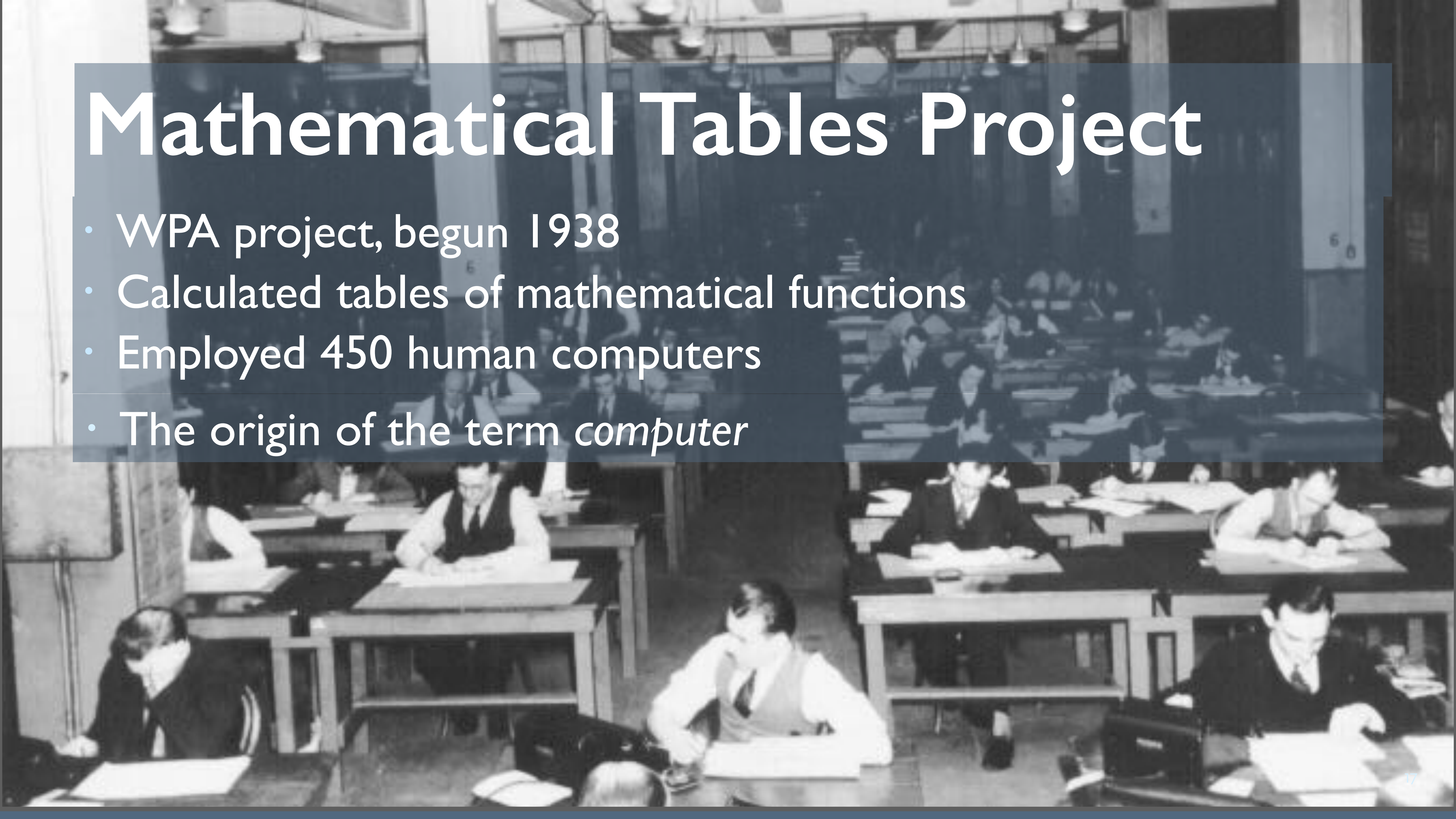
Work distributed via mail

Days	at Noon.	at Noon.	of the	of the
	S. D. M. S.	D. M. S.	of the	of the
1	2 20 0 51	2 58 29 N	1	
2	3 2 45 8	3 48 52	2	
3	3 15 46 23	4 28 29	3	
4	3 29 3 54	4 54 28	4	Spica m.
5	4 12 35 47	5 4 34	5	Id
6	4 26 19 33	4 57 26	6	Id
7	5 10 12 31	4 32 51	7	Id
8	5 24 12 11	3 51 51	8	Antares
9	6 8 16 36	2 56 35	9	Id
10	6 22 24 16	1 50 19	10	Id
11	7 6 34 4 0	37 6 N	11	α aquila
12	7 20 44 43 0	38 25 S	12	Id
13	8 4 54 27 1	51 21	13	Fomalh.
14	8 19 0 45 2	57 1	14	Id
15	9 3 0 17 3	51 19	15	Id
16	9 16 49 18 4	31 12	16	Id
17	10 0 24 8 4	54 54	17	α arictis
18	10 13 41 50 5	1 59	18	Id
19	10 26 40 41 4	53 8	19	Id
20	11 9 20 28 4	29 54	20	Id
21	11 21 42 30 3	54 21	21	Id
22	0 3 49 35 3	8 48	22	α abdo.?
23	0 15 45 40 2	15 32	23	Id
24	0 27 35 32 1	16 47	24	Id
25	0 24 26 2 11	14 11	25	Id
26	0 17 52 0 11	11 11	26	Id
27	2 3 21 8 1	49 51	27	Id
28	9 15 38 50 2	17 21	28	Id



# Mathematical Tables Project

- WPA project, begun 1938
- Calculated tables of mathematical functions
- Employed 450 human computers
- The origin of the term *computer*



# Enter computer science

- Computation allows us to execute these kinds of goals at even larger scale and with even more complexity
- We can design systems that gather evidence, combine estimates, and guide behavior



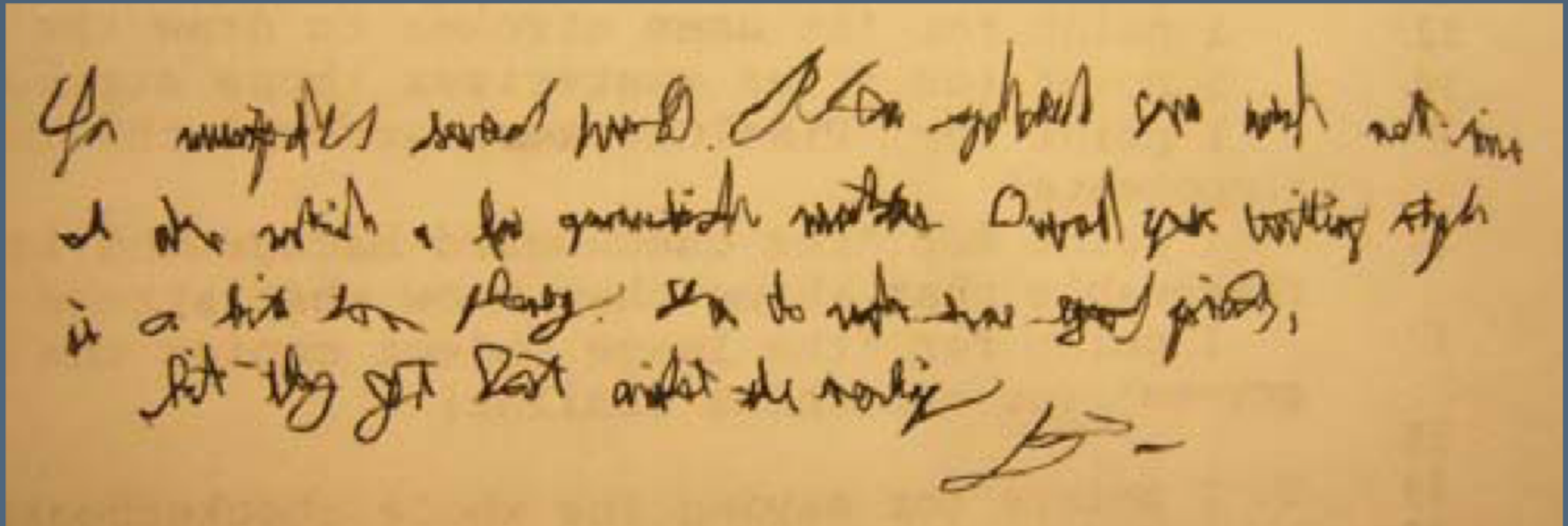
# Iterative crowd algorithm

For multiple server nodes. Also global server with not-in  
it also which a for guaranteed market. Overall your writing style  
is a bit too heavy. So to make sure you find  
it- they get best and it's working





# Iterative crowd algorithm

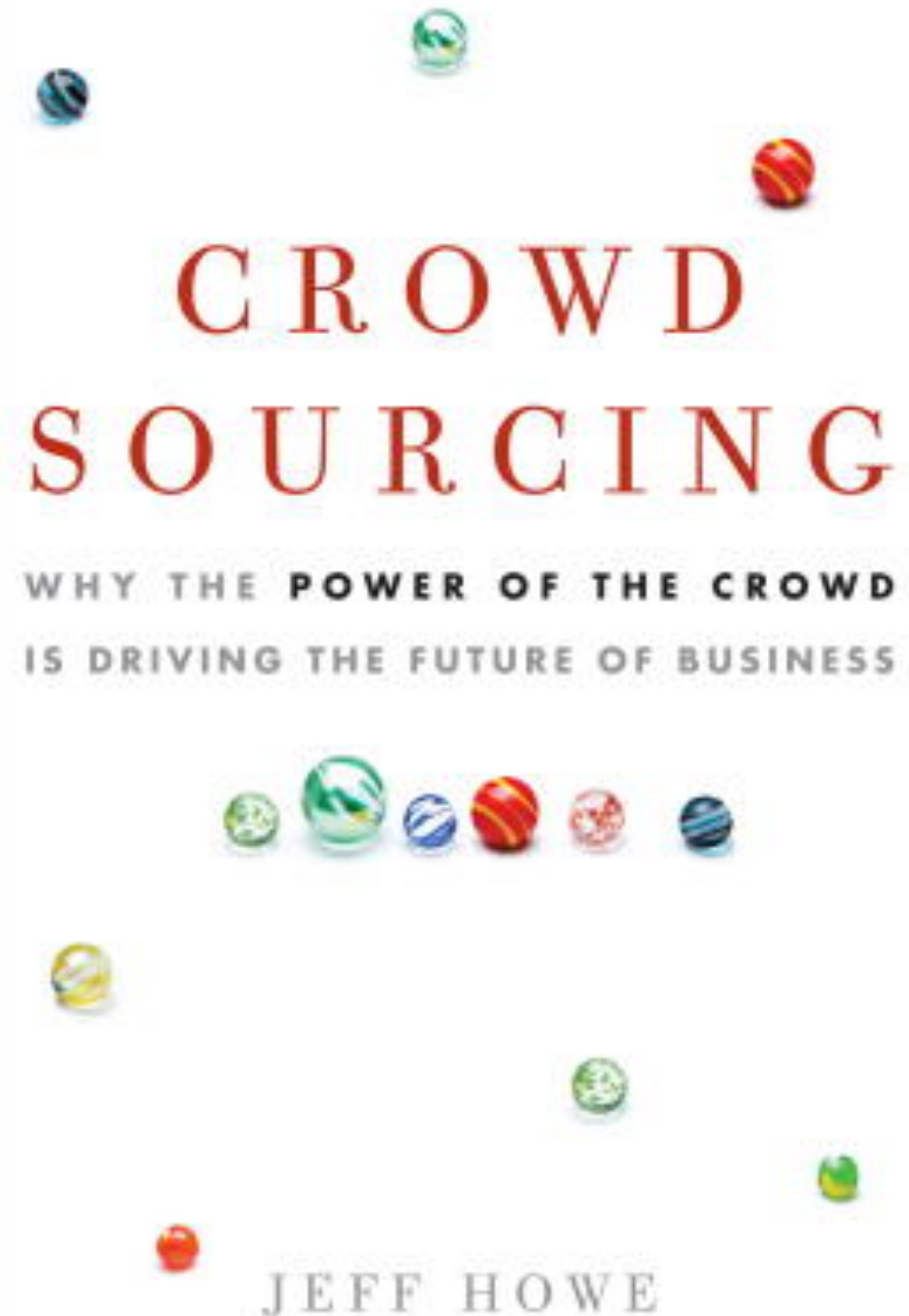


You (misspelled) (several) (words). Please spellcheck your work next time. I also notice a few grammatical mistakes. Overall your writing style is a bit too phoney. You do make some good (points), but they got lost amidst the (writing). (signature)



# Etymology

- Crowdsourcing term coined by Jeff Howe, 2006 in Wired
- “Taking [...] a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.”



# Recall: games with a purpose

Label every image on the internet using a game

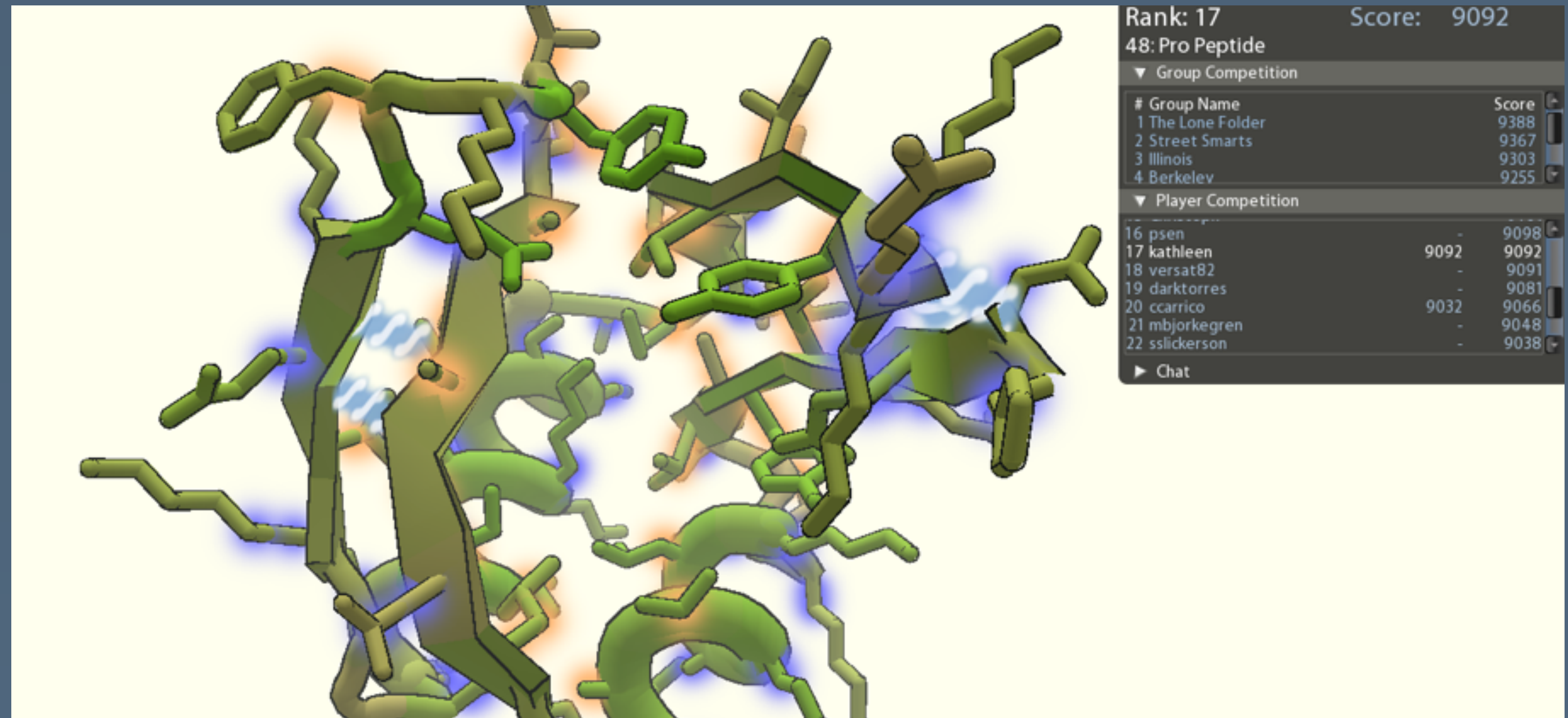
[von Ahn and Dabbish, CHI '06]





# Recall: scientific collaboration

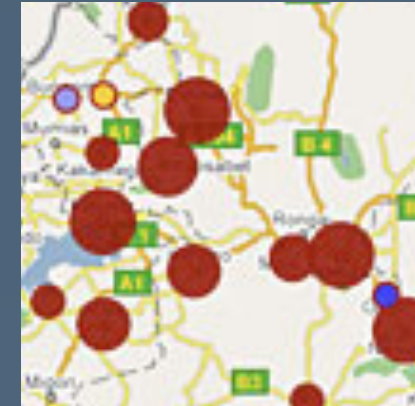
- FoldIt: protein-folding game
- Amateur scientists have found protein configurations that eluded scientists for years



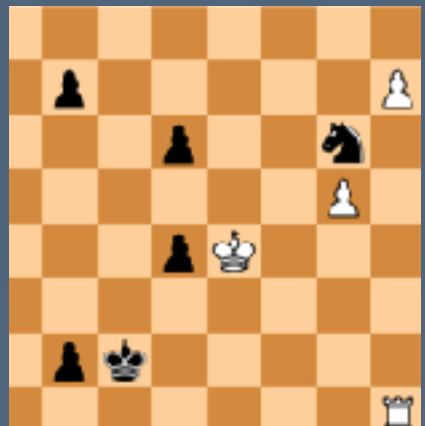
# More successes



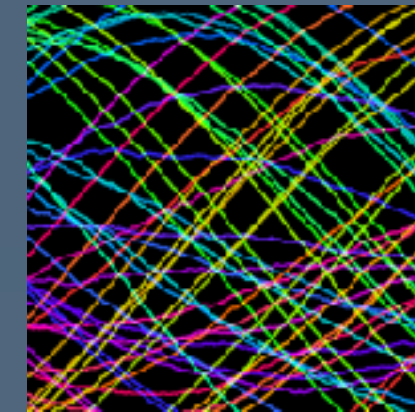
Largest encyclopedia  
in history



Disaster reporting



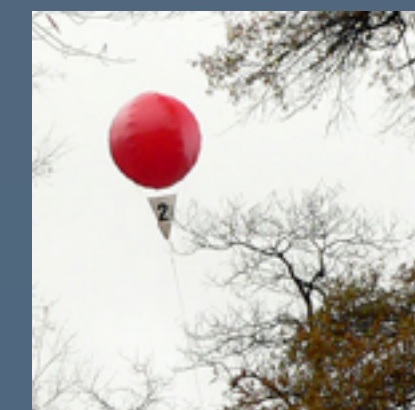
Kasparov vs. the world



Collaborative math proofs



NASA Clickworkers



DARPA Red Balloon Challenge



# Crowd work

- Crowds of online freelancers are now available via API
  - Amazon Mechanical Turk, Upwork, TopCoder, 99 Designs, etc.
  - 600,000 workers are in the United States' digital on-demand economy [Economic Policy Institute 2016]
  - Eventually, this will include 20% of jobs in the U.S. [Blinder 2006], about 45,000,000 full-time workers [Horton 2013]
- The promise: What if the smartest minds of our generation could be brought together with a single click? What if you could flexibly refashion your career with every job you do?
- The peril: what happens when an algorithm is your boss?

# Amazon Mechanical Turk

- Pay small amounts of money for short tasks

Label an image

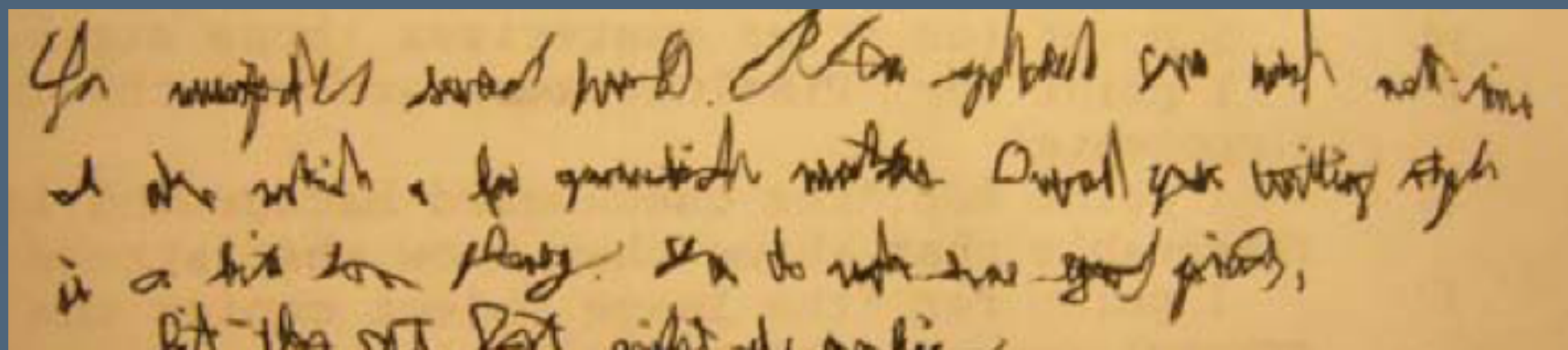
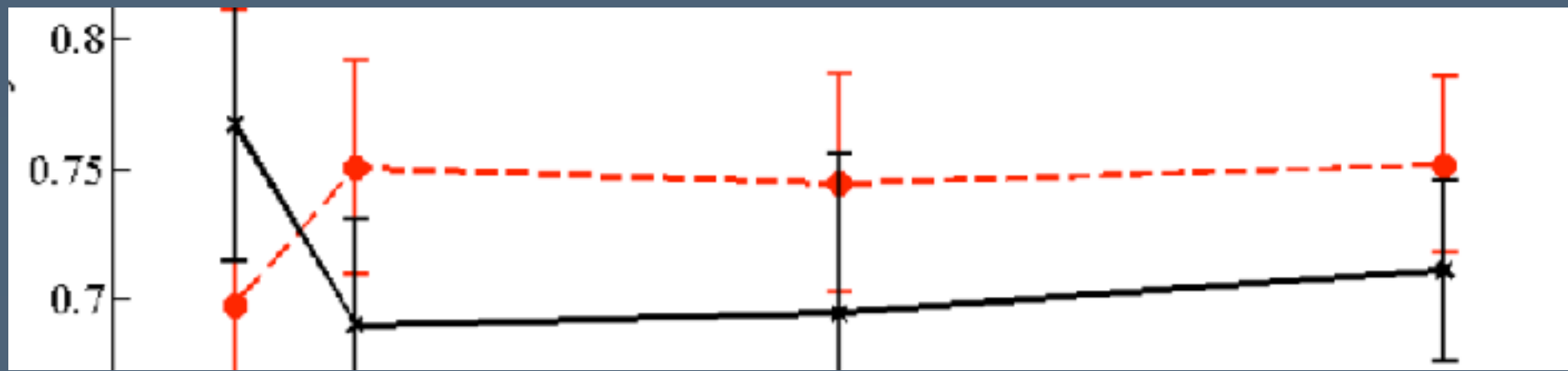
Reward: \$0.02

Transcribe audio clip

Reward: \$0.05



# Major topics of research



Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

## Incentives and Quality

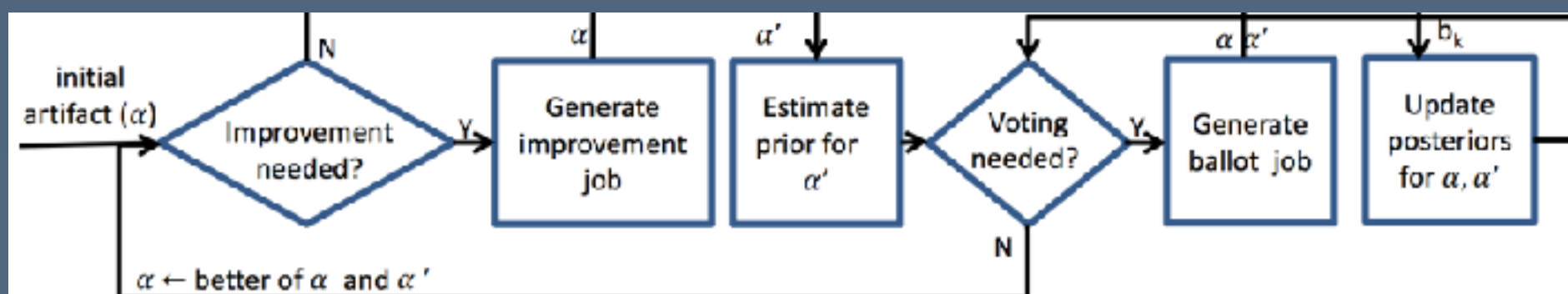
[Mason and Watts, HCOMP 2009]  
[Dow et al., CSCW 2012]

## Crowd algorithms

[Little et al., HCOMP 2009]

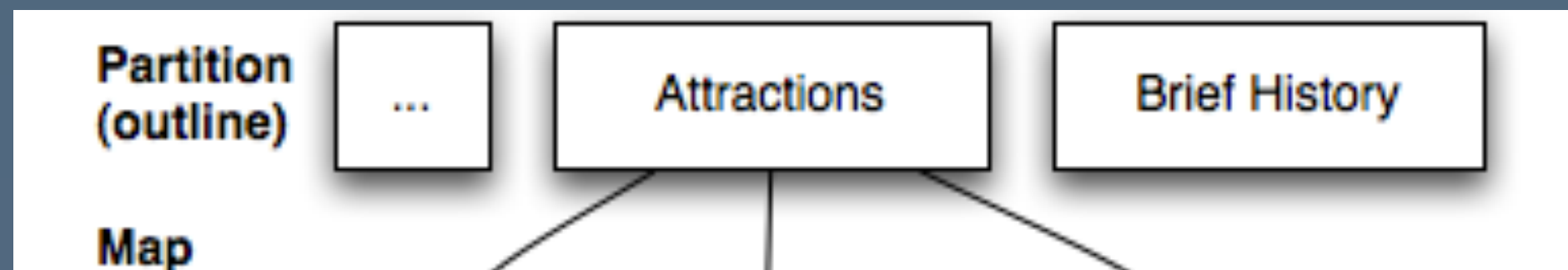
## Crowd-powered systems

[Bernstein et al., UIST 2010]  
[Bigham et al., UIST 2010]



## AI for HCOMP

[Dai, Mausam & Weld, AAAI 2010]



## Complex Work

[Kittur et al., UIST 2011]

# Incentives and quality

# Goal: modularize the task so that anyone can do it

- If done correctly, a decentralized group of workers can accurately complete the task at high quality

### Instructions

You must provide 3 tags for the main subject in this image.

- Each tag must be a single word.
- No tag can be longer than 25 characters.
- The tags must describe the image, the contents of the image,



**Tag 1:**

**Tag 2:**

# Problem: low-quality work

- “These cheap labels may be noisy due to lack of expertise, dedication, [or] interest” [Sheng, Provost, and Ipeirotis 2008]
- “Workers cannot be relied upon to provide high-quality work of the type one might expect from a traditional employee for various reasons including misunderstanding of task directives, laziness, or even maliciousness.” [Lasecki et al. 2011]



# What can we do?

- Does paying more produce better work?
  - More work, but not higher-quality work [Mason and Watts, HCOMP '09]
  - ...Unless the task is designed so that workers can produce higher quality work by exerting more effort [Ho et al., WWW '15]
- Does feedback produce better work?
  - Self-assessment and expert assessment both improve the quality of work [Dow, Kulkarni, Klemmer and Hartmann, CSCW '11]



# Incentives

[Shaw, Horton and Chen, CSCW '11]

- Which of these approaches improve quality?
  - Comparison to other workers
  - Normative claims: “it’s important that you try hard”
  - Solidarity: your team gets a bonus if you are right
  - Humanization: “thanks for working; I’m Aaron.”
  - Reward or punish accuracy with money
  - Reward or punish agreement with money
  - Bayesian truth serum: predict others’ responses
  - Bet payment on the accuracy of your responses

# Incentives

[Shaw, Horton and Chen, CSCW '11]

- Which of these approaches improve quality?
  - Comparison to other workers
  - Normative claims: “it’s important that you try hard”
  - Solidarity: your team gets a bonus if you are right
  - Humanization: “thanks for working; I’m Aaron.”
  - Reward or punish accuracy with money
  - Reward or punish agreement with money
  - **Bayesian truth serum: predict others’ responses**
  - Bet payment on the accuracy of your responses



# Judging quality explicitly

- **Gold standard judgments** [Le et al., SIGIR CSE '10]
  - Include questions with known answers
  - Performance on these “gold standard” questions is used to filter work
- **Get Another Label** [Sheng, Provost, Ipeirotis, KDD '08]
  - Estimate the correct answer and worker quality jointly
  - Try it! <https://github.com/ipeirotis/Get-Another-Label>

# Judging quality implicitly

[Rzeszotarski and Kittur, UIST '12]

- Observe low-level behaviors
  - Clicks
  - Backspaces
  - Scrolling
  - Timing delays
- SVMs on these behaviors predict work quality
- Limitation: models must be built for each task



# Person- vs. process-centric

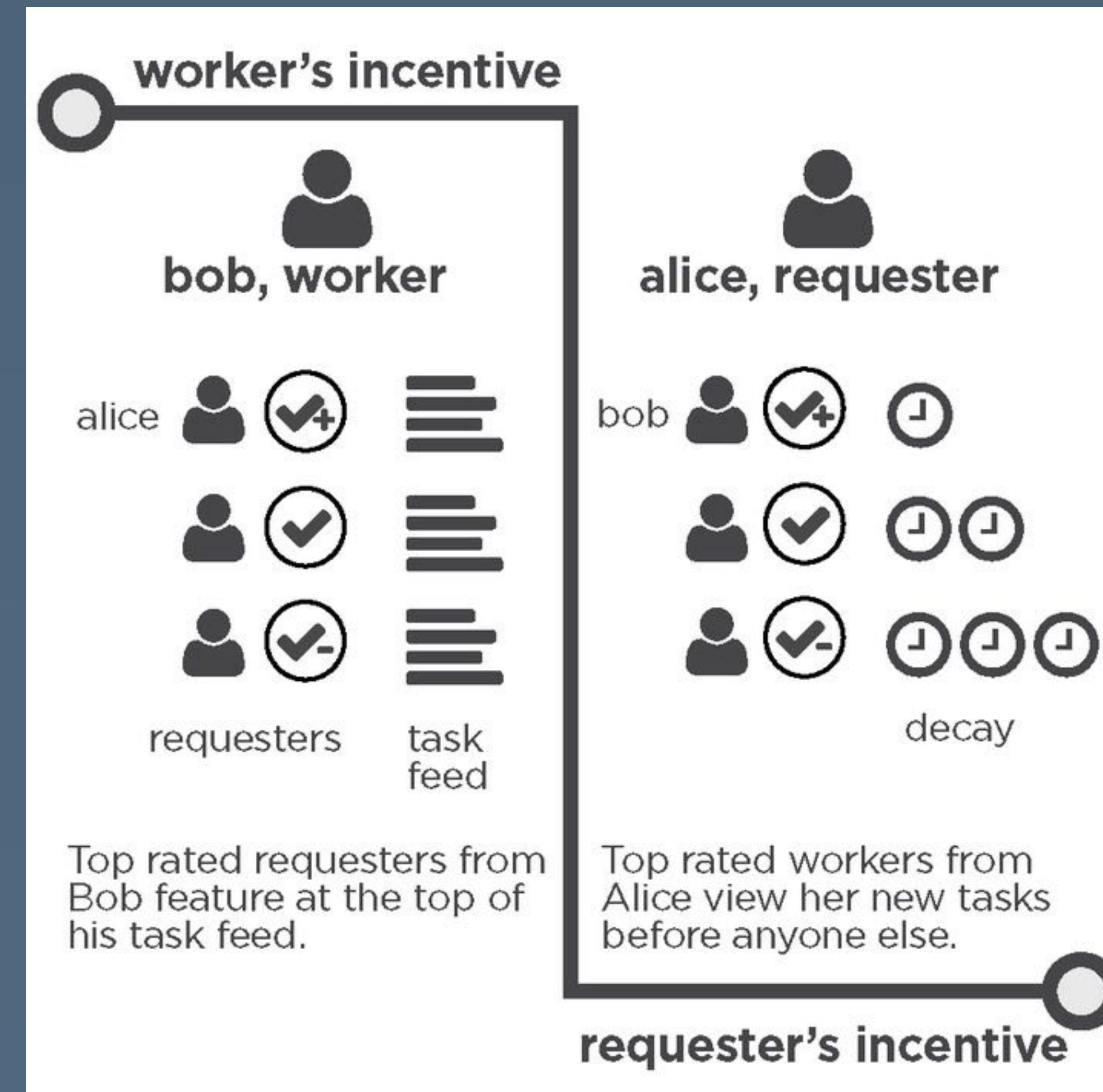
[Mitra, Hutto and Gilbert, CHI '15]

- Person-centric methods: find and filter for high performers
  - Essentially, build up a private reputation measurement
  - e.g., gold standard questions
  - e.g., qualification tests
- Process-centric methods: take all comers and use algorithms
  - e.g., financial incentives
  - e.g., Bayesian Truth Serum
- Result: person-based strategies are most effective

# Boomerang

[Stanford Crowd Research Collective, UIST '16]

- Little incentive to leave accurate feedback
- *Boomerang*: rebound the consequences back onto the rater
  - When I give a worker a high rating, the system gives that worker early access to my future tasks.
  - Example: giving a high rating to a low-quality worker increases the probability that the low-quality worker returns to do more of my work
  - This strategy empirically deflates reputation scores





# Michael's take

- There are two primary causes of quality challenges:
  - **Strategic dishonesty**, where the worker is explicitly seeking to get away with more money and less effort
  - **Mental model misalignment**, where the requester has not clearly communicated their goal to the worker
- My experience is that strategic dishonesty is rare and can be caught, whereas mental model misalignment is ubiquitous
  - (But most of our papers focus on strategic dishonesty)

# Michael's take

- **Quality** isn't the problem with crowdsourcing, per se
- It's actually the **amount of effort** required that drives requesters (buyers) away
  - Authoring tasks
  - Getting rid of bad workers
  - Revising tasks
  - It's a ton of babysitting work
- I now agree with Mitra that finding ways to identify high-quality **workers**, rather than high-quality **work**, is the best way to escape the Mechanical Turk market for lemons



# Crowdsourcing algorithms

# Goal: guide crowds as they work

- Designing crowdsourcing algorithms is often like designing a user interface that will keep a user “in bounds” on your application
- Challenges
  - Taking unexpected action
  - Trying too hard
  - Trying not hard enough



# Crowdsourcing algorithm

- A generalized version of a workflow
- Iterative algorithms [Little et al. 2009]
  - Hand off from one worker to the next



- Most crowdsourcing processes are more parallel, but less interesting algorithmically

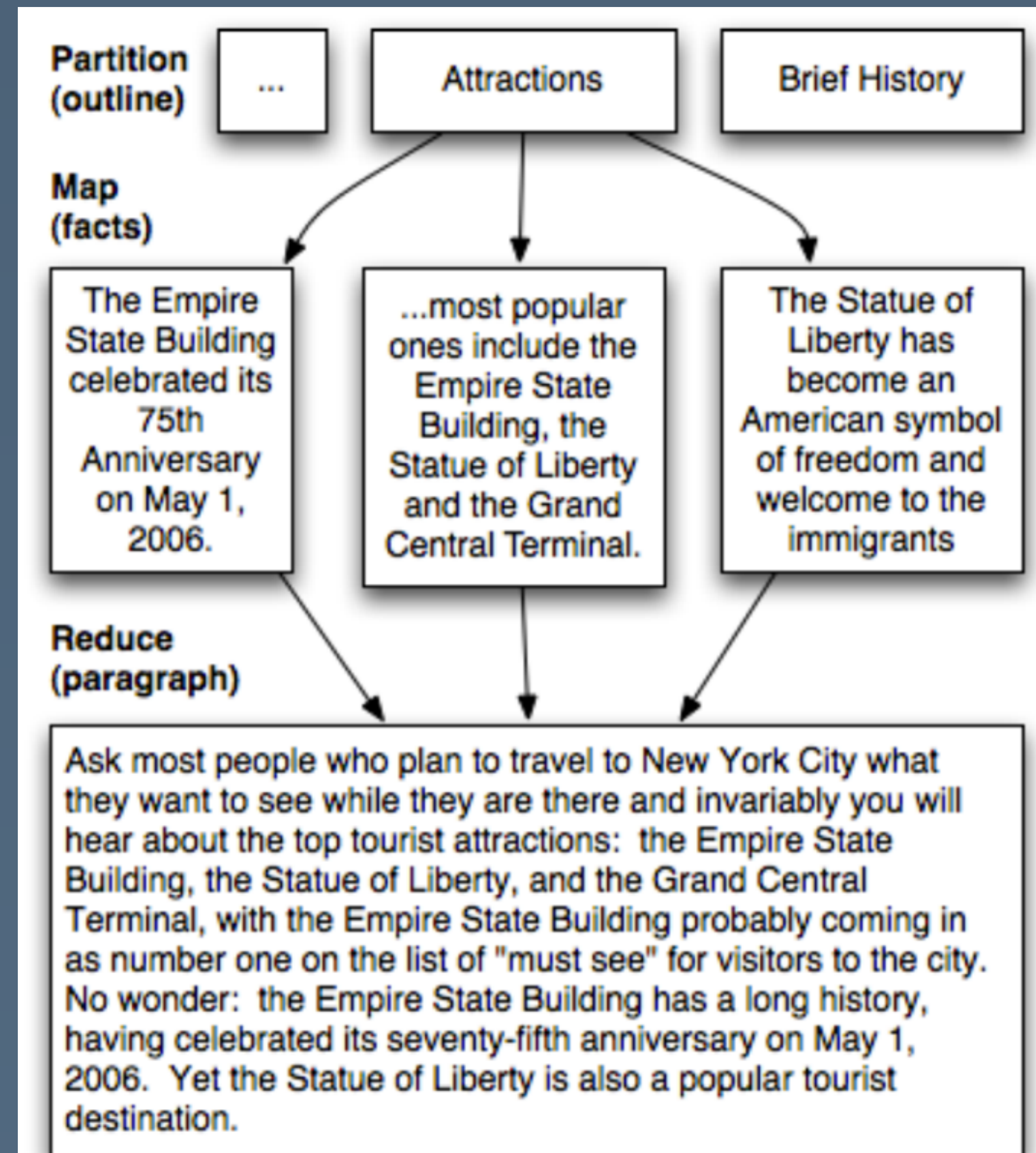
# Crowdsourcing algorithms

- Open-ended editing: Find-Fix-Verify [Bernstein et al., UIST '10]
- Graph search [Parameswaran et al., VLDB '11]
- Clustering [Chilton et al., CHI '13]
- and many more...
  
- When write an algorithm?  
If you tried this in a straightforward way,  
would crowds fail? Why?

# CrowdForge

[Kittur et al., UIST '11]

- Crowdsourcing as a map-reduce process
- To write a wikipedia page, partition on topics, map to find facts and then reduce into a paragraph

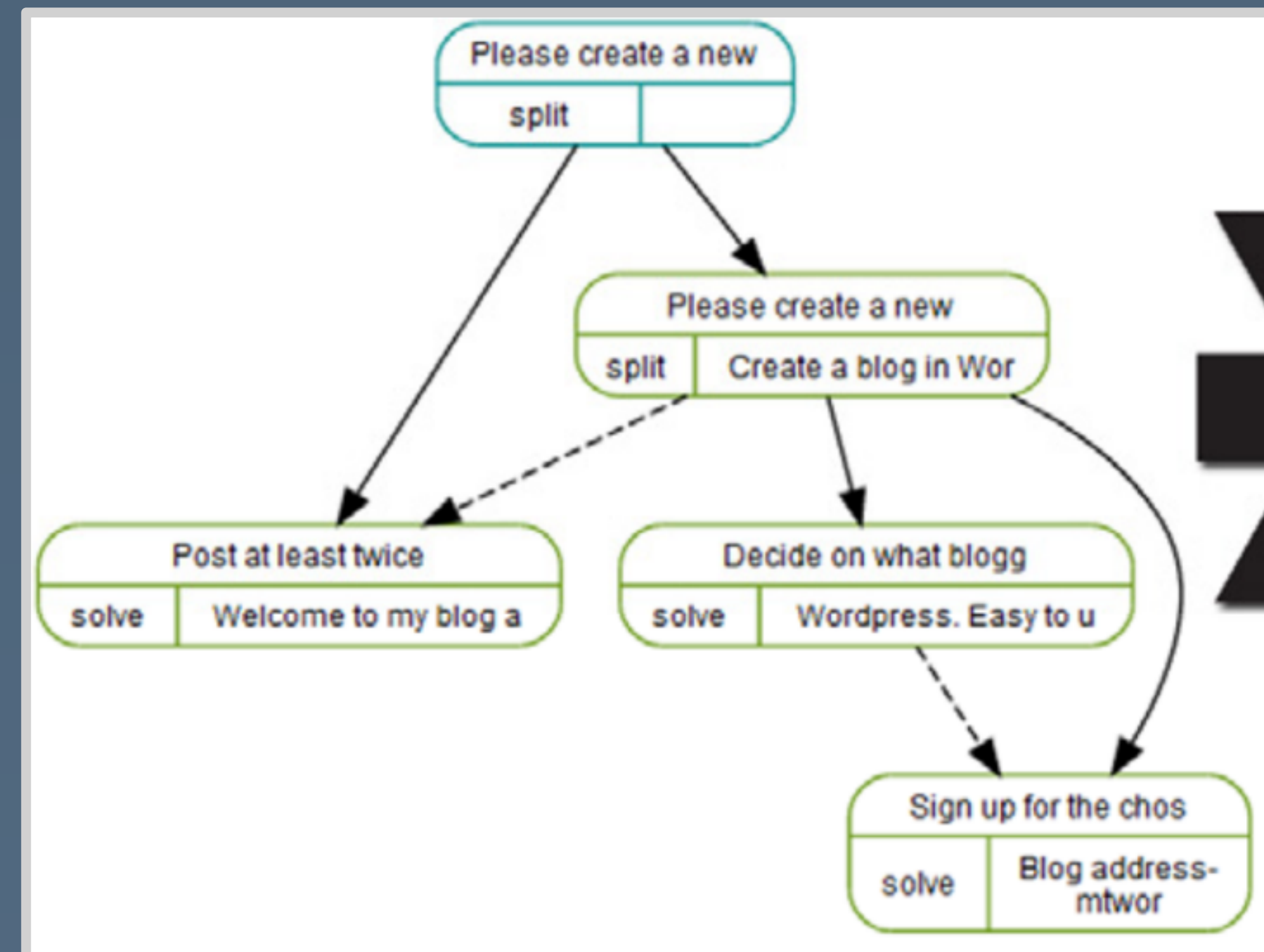




# Turkomatic

[Kulkarni, Can, and Hartmann, CSCW '12]

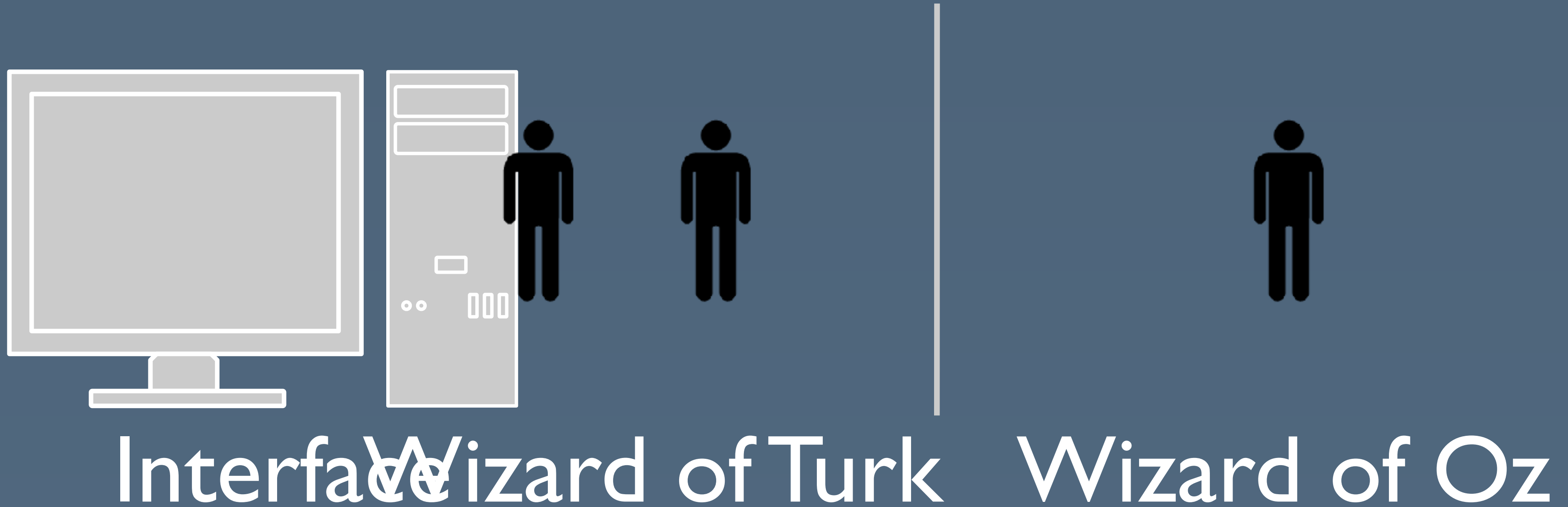
- Let the workers decide on task design
- Is a task too complicated for \$D? If so, ask for sub-tasks and recurse. If not, do it yourself.
- Creating a blog with content:



# Crowd-powered systems

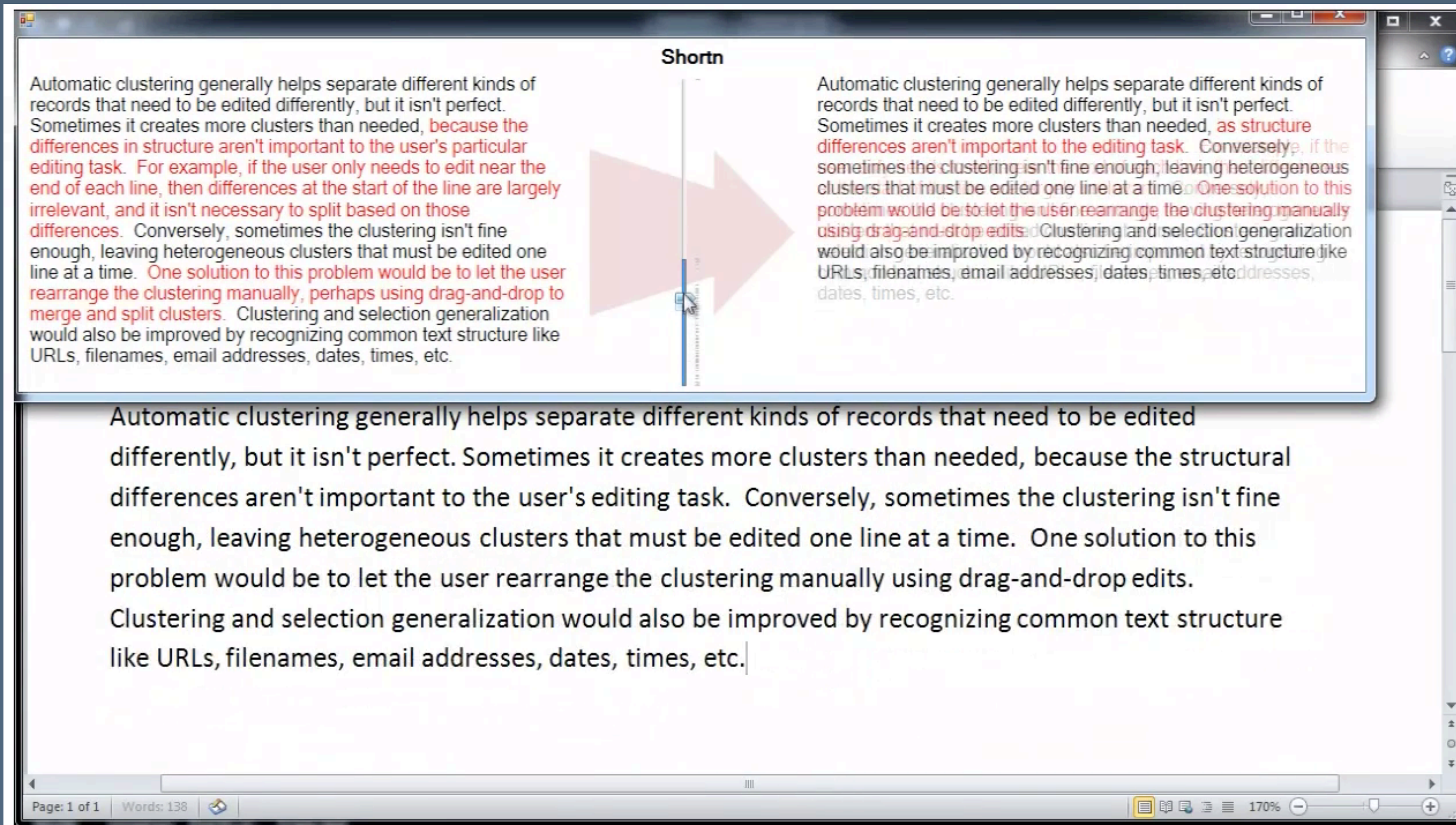
# Why do it?

- Embed crowd intelligence inside of user interfaces and applications we use today






# Soylent



**Shortn**

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.



Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, as structure differences aren't important to the editing task. Conversely, if the sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the structural differences aren't important to the user's editing task. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Page: 1 of 1 Words: 138 170%



# VizWiz

[Bigham et al., UIST '10]

- Visual question answering for the blind

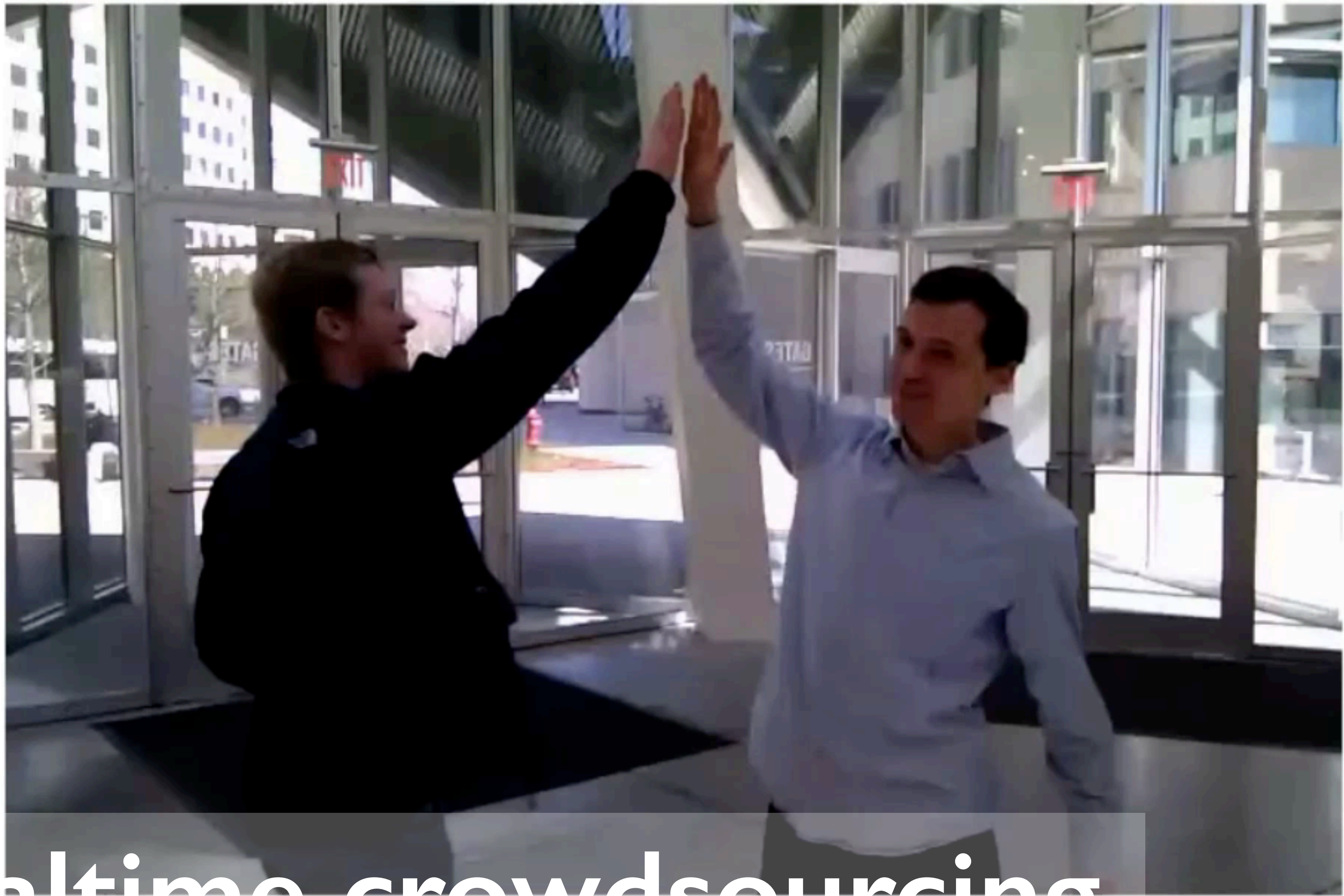
What color is this pillow?	What denomination is this bill?	Do you see picnic tables across the parking lot?	What temperature is my oven set to?	Can you please tell me what this can is?	What kind of drink does this can hold?
					
(89s) I can't tell. (105s) multiple shades of soft green, blue and gold	(24s) 20 (29s) 20	(13s) no (46s) no	(69s) it looks like 425 degrees but the image is difficult to see. (84s) 400 (122s) 450	(183s) chickpeas. (514s) beans (552s) Goya Beans	(91s) Energy (99s) no can in the picture (247s) energy drink

- 1 to 2 minute responses by keeping workers on fake tasks until needed

# Crowd-powered databases

- Database with open-world assumptions:  
`SELECT * FROM ice_cream_flavors`
- Several university flavors
  - Berkeley: CrowdDB [Franklin et al., SIGMOD '11]
  - MIT: Qurk [Marcus et al., CIDR '11]
  - Stanford: Deco [Parameswaran et al. '11]
- Tackling many important optimization questions: e.g., joins, ranking, sorting





# Realtime crowdsourcing

[Bernst...]

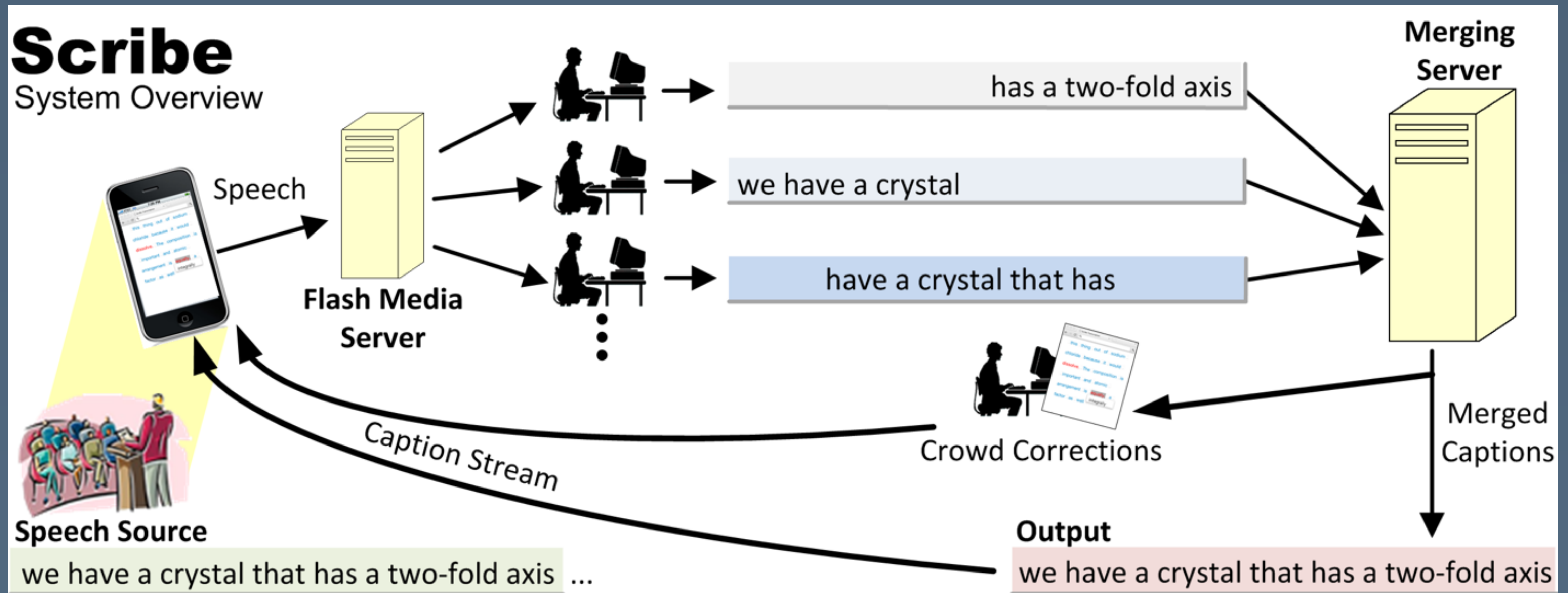
Find photo in this clip





# Realtime crowdsourcing

- Realtime captioning using shotgun gene sequencing techniques



# New forms of crowdsourcing



# Communitysourcing

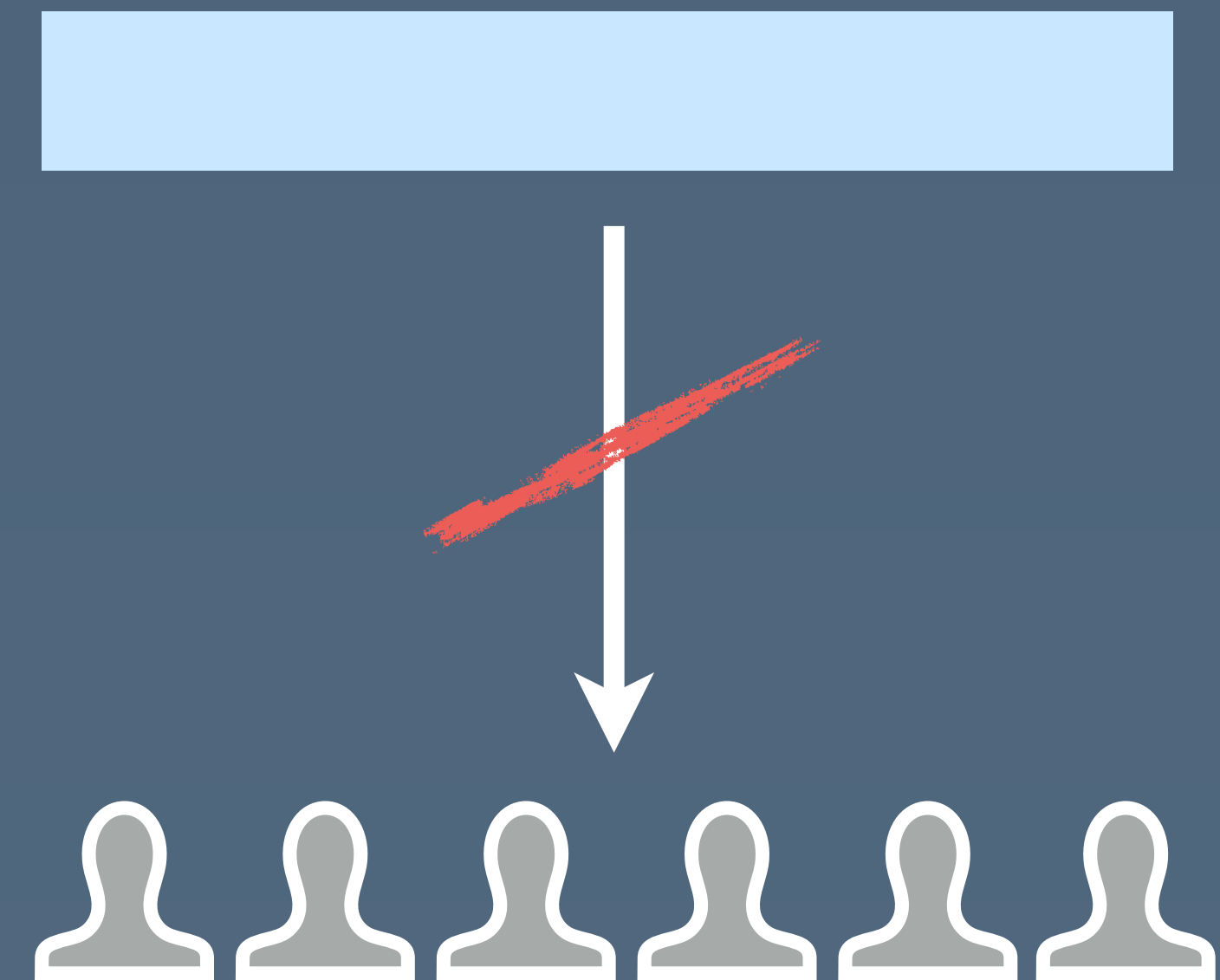
Engaging Local Crowds to Perform  
Expert Work Via Physical Kiosks

Kurtis Heimerl, Brian Gawalt, Kuang Chen  
Tapan Parikh, Björn Hartmann  
University of California, Berkeley

**Hacking motivation** CHI 2012  
[Heimerl et al., CHI '12]

# Microtask crowds struggle with complex tasks

- Design, engineering, writing, video production, music composition  
[Kittur et al. 2013, Kulkarni et al. 2012]



# Crowds of experts

## Mechanical Turk



microtask worker  
microtask worker  
microtask worker  
microtask worker  
microtask worker



## Upwork



programmer  
designer  
video editor  
musician  
statistician

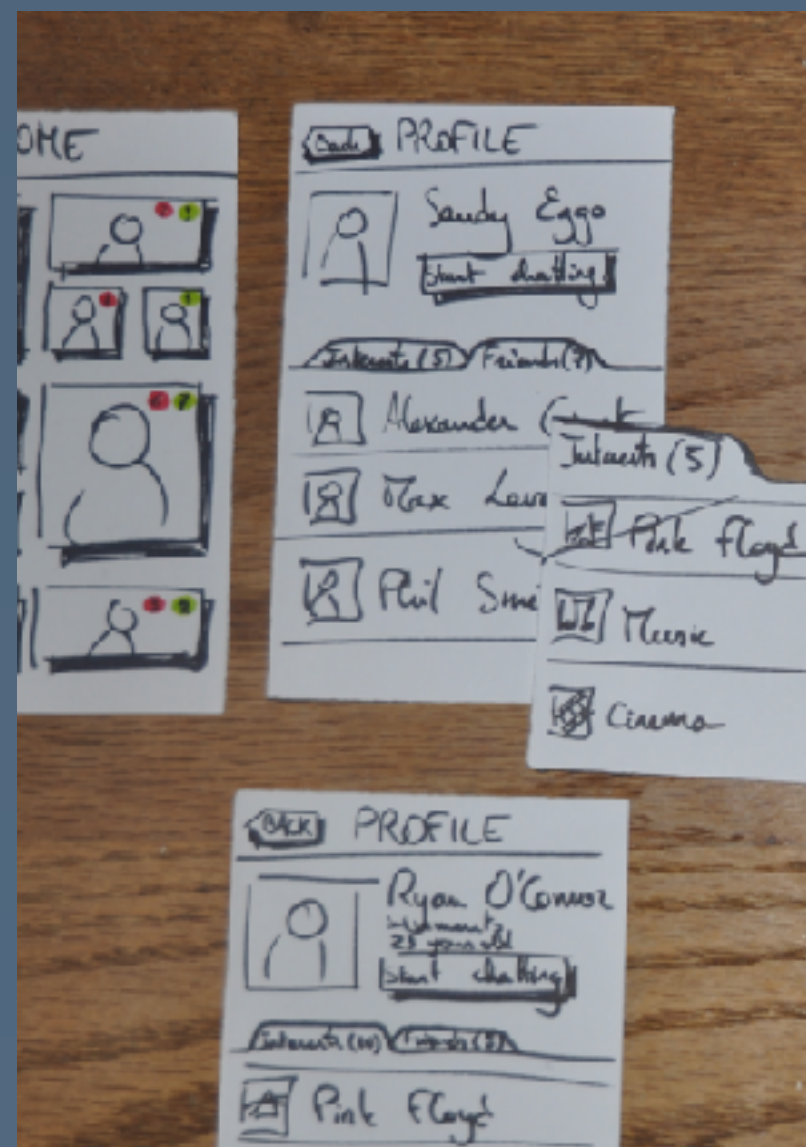


# Flash Teams

[Retelny et al., UIST '14]

Computationally-guided teams of crowd experts supported by lightweight, reproducible and scalable team structures.

Input

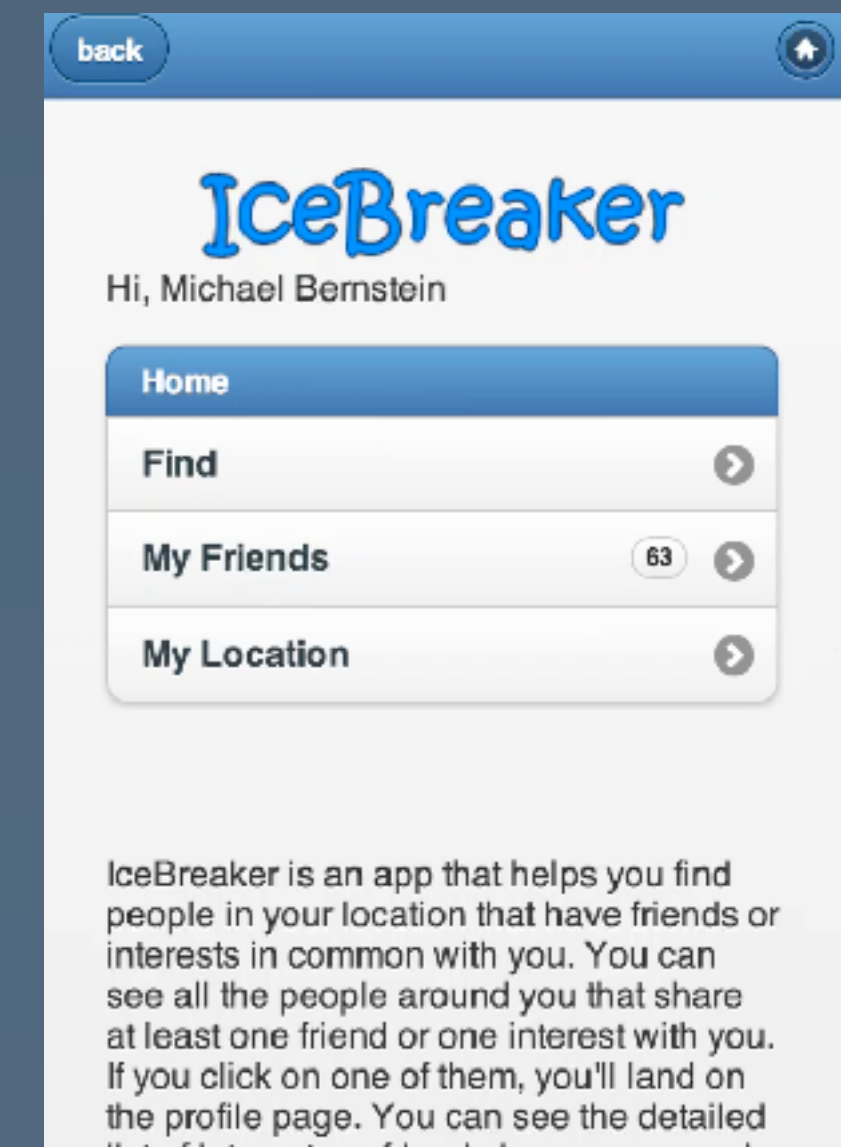


Flash Team



design

Output









# Artificial intelligence for crowds



# TurKontrol: AIs guiding crowds

[Dai, Mausam and Weld, AAAI '10]

- Workflow planning as a decision-theoretic optimization problem
- Trade off quality vs. number of workers required
  - POMDP to decide: do we need a vote? do we need more voters? do we need more improvement?

# The future for crowd workers



A young child, likely a toddler, is the central focus of the image. They are wearing a bright yellow hard hat and an orange safety vest over a light-colored shirt. The child is holding a yellow balloon in their right hand and looking towards the camera with a slight smile. The background is a plain, light-colored wall. The overall tone of the image is warm and positive.

*What would it take for us to be proud of our children growing up to be crowd workers?*



# Careers in crowd work

[Kittur et al., 2013]

- More and more people are engaging in online paid work: programmers, singers, designers, artists, ...
- Would you feel comfortable with your best friend, or your own child, becoming a full-time crowd worker?
- How could we get to that point? What would it take?
  - Education
  - Career advancement
  - Reputation

# Potential or peril?

- Crowdsourcing is a populist form of information work, but the technical infrastructure actively disempowers workers.  
[Irani and Silberman '13]

# Take back the market

- Turkopticon [Irani and Silberman '13]
  - Lets workers (sellers) review requesters (buyers)



- Dynamo [Salehi et al. '15]
  - Lets workers engage in collective action

## Dear Jeff Bezos

We are writing to let you and the rest of the world know about who we are. The intent is for you to see that Turkopticon is not only actual human beings, but people who deserve respect, fair treatment and open communication.

[Donate](#)

Donations will go towards reaching out to more Turkopticon workers via a HIT on Mechanical Turk (3 minute paid vacation). Raised: \$60



# Needed infrastructure

- Support for career growth
  - e.g., micro-internships [Suzuki et al. 2016]
- Training and education
- Longer-term employment
- Decoupling the social safety net from firm-based employment

# Michael's take

- Broadening our worldview from microtasks to a global, digitally–networked expert workforce will reshape our research trajectory
- If Mechanical Turk is the Friendster of online labor, what will be the Facebook of online labor?

# Discussion rooms

<b>Rotation</b>	<b>Littlefield 107</b>	<b>Littlefield 103</b>
a	12	34
b	24	13
c	14	23
d	34	12
e	13	24
f	23	14