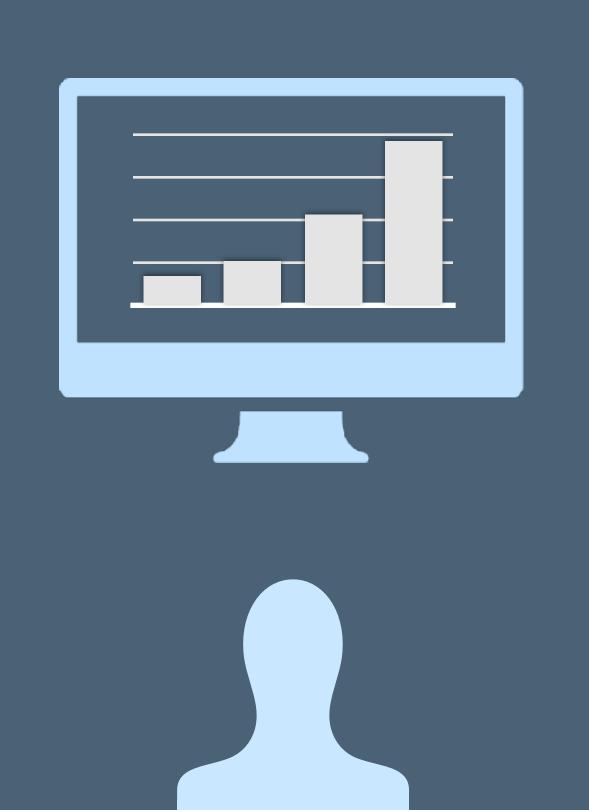
Crowdsourcing

MICHAEL BERNSTEIN CS 376

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?

What is crowdsourcing?

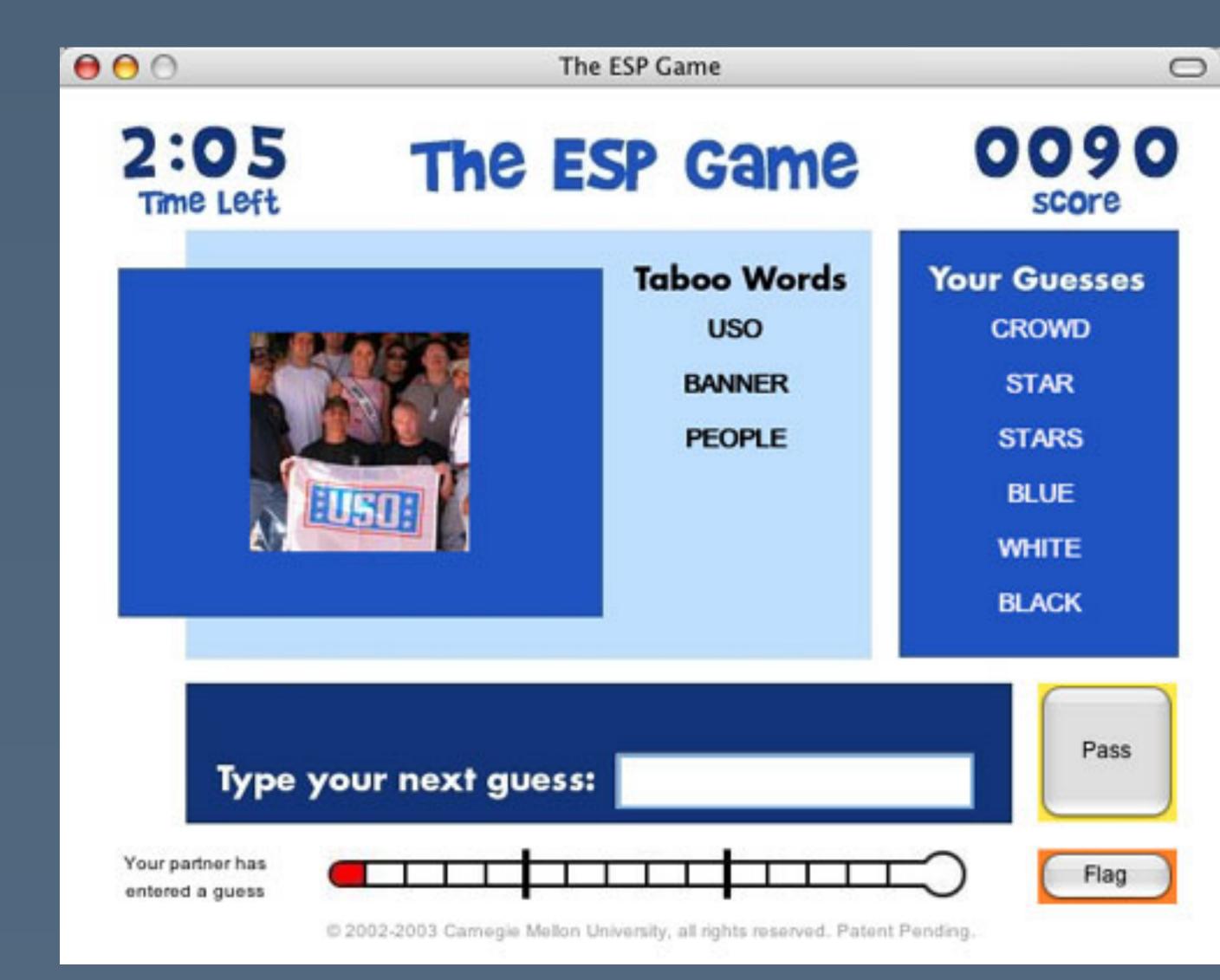
Definitions are great, but...

· Examples are much clearer.

Science: Games with a Purpose

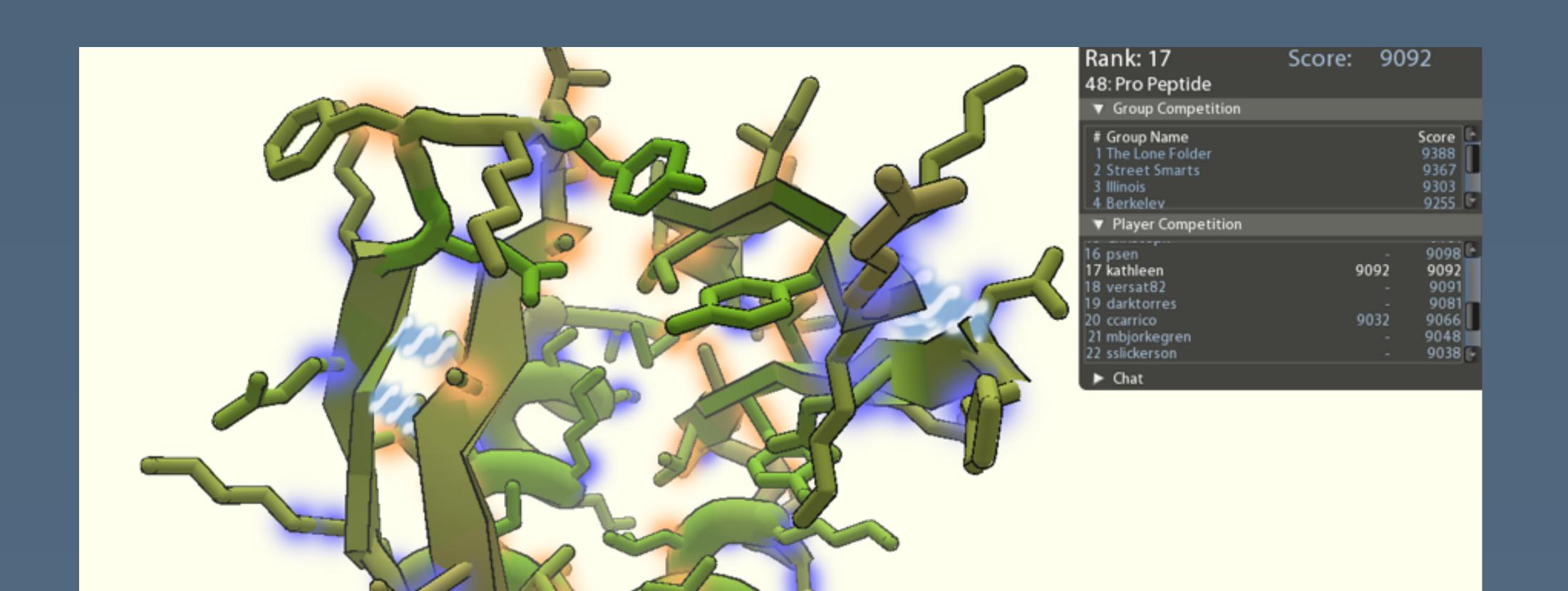
Label every image on the internet using a game

[von Ahn and Dabbish, CHI '06]



Science: Protein Folding

 Amateur scientists have found protein configurations that eluded scientists for years



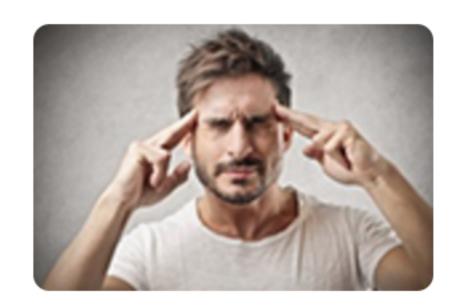
Science: LabInTheWild

· "Buzzfeed-ifying" online studies through narcissism



What is your website aesthetic?

Compare your visual preferences to people around the world. This experiment takes



How good is your implicit memory?

Have you ever had a gut feeling about something? Your implicit memory might have



What is your thinking style?

Find out how your thinking style compares to others. This experiment takes around 5



Quantifying Visual Preferences Around the World

Katharina Reinecke
University of Michigan
Ann Arbor MI 48109

Krzysztof Z. Gajos
Harvard University
33 Oxford St. Cambridge MA

Business: Paid Crowdsourcing

- · Pay small amounts of money for short tasks
- Amazon Mechanical Turk: Roughly five million tasks completed per year
 at I-5¢ each [lpeirotis 2010]

Label an image

Reward: \$0.02

Transcribe audio clip

Reward: \$0.05

- Rough population (needs to be updated):
 40% U.S., 40% India, 20% elsewhere
- · Gender, education and income are close mirrors of overall population distributions

Business: Innovation Competitions

INNOCENTIVE®

Solvers

Our Offering

Challenge Center

Resources

About

Blog

Contact Us

Register | Login

Innovate with InnoCentive

Your partner in enabling creative minds to solve problems that matter to you.

NETFLIX

Netflix Prize



Home

Rules

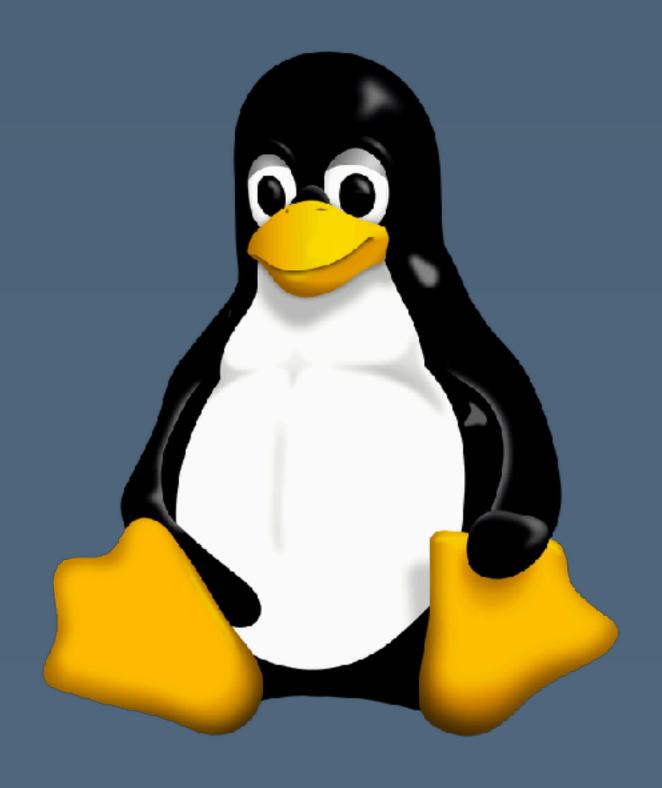
Leaderboard

Update

Public infrastructure



Wikipedia



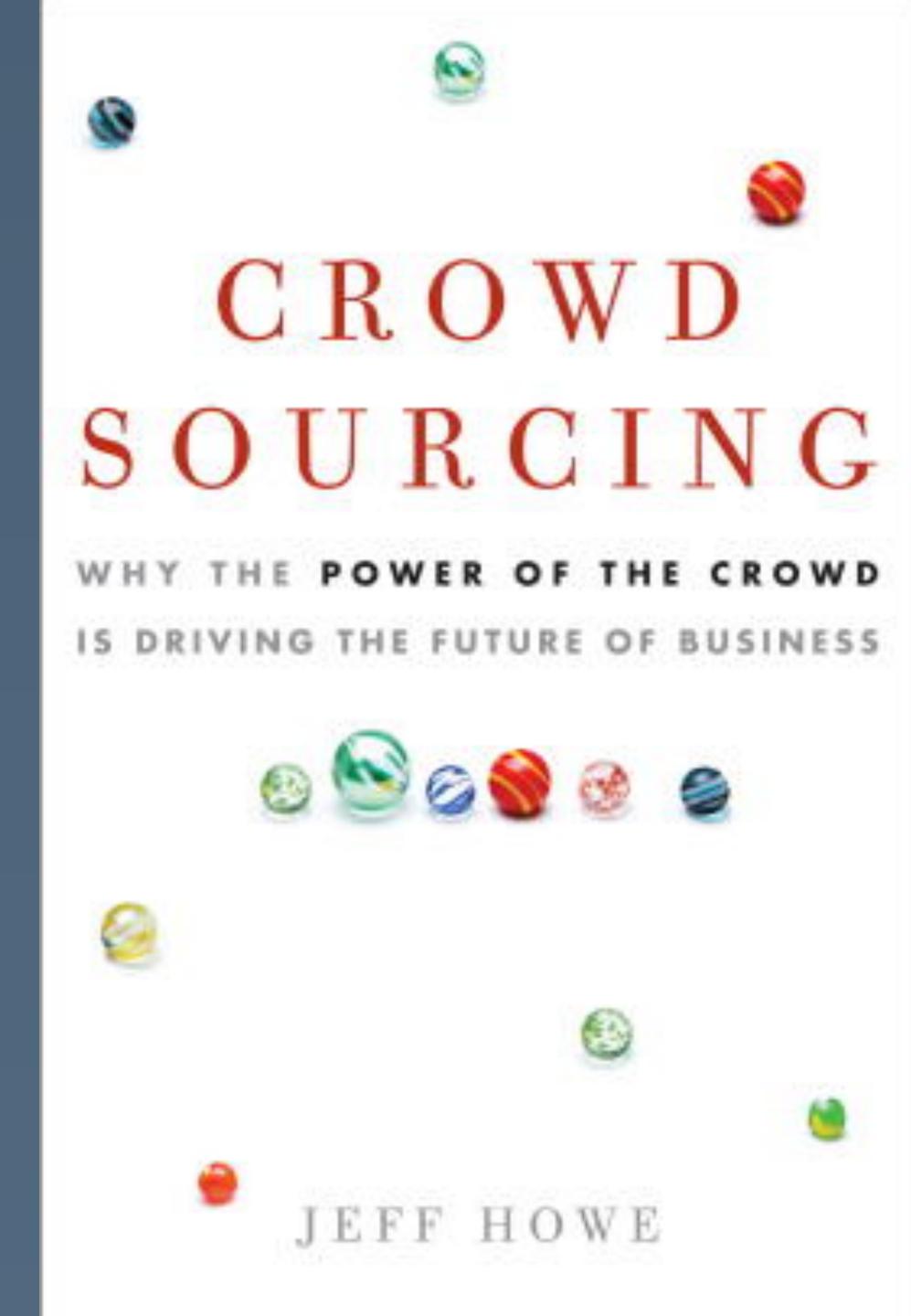
Linux



Ushahidi

Definition

- 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."



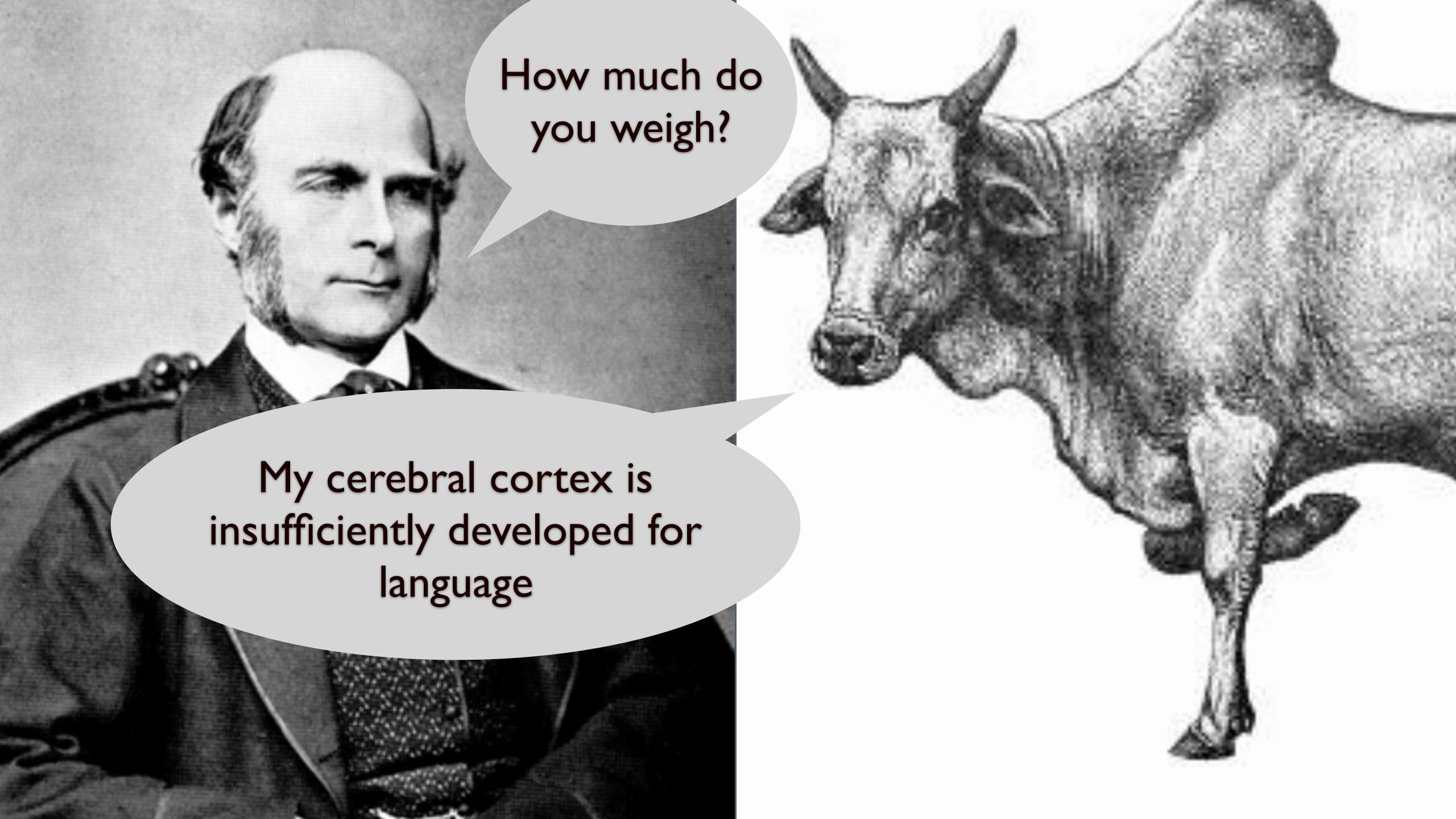
Today

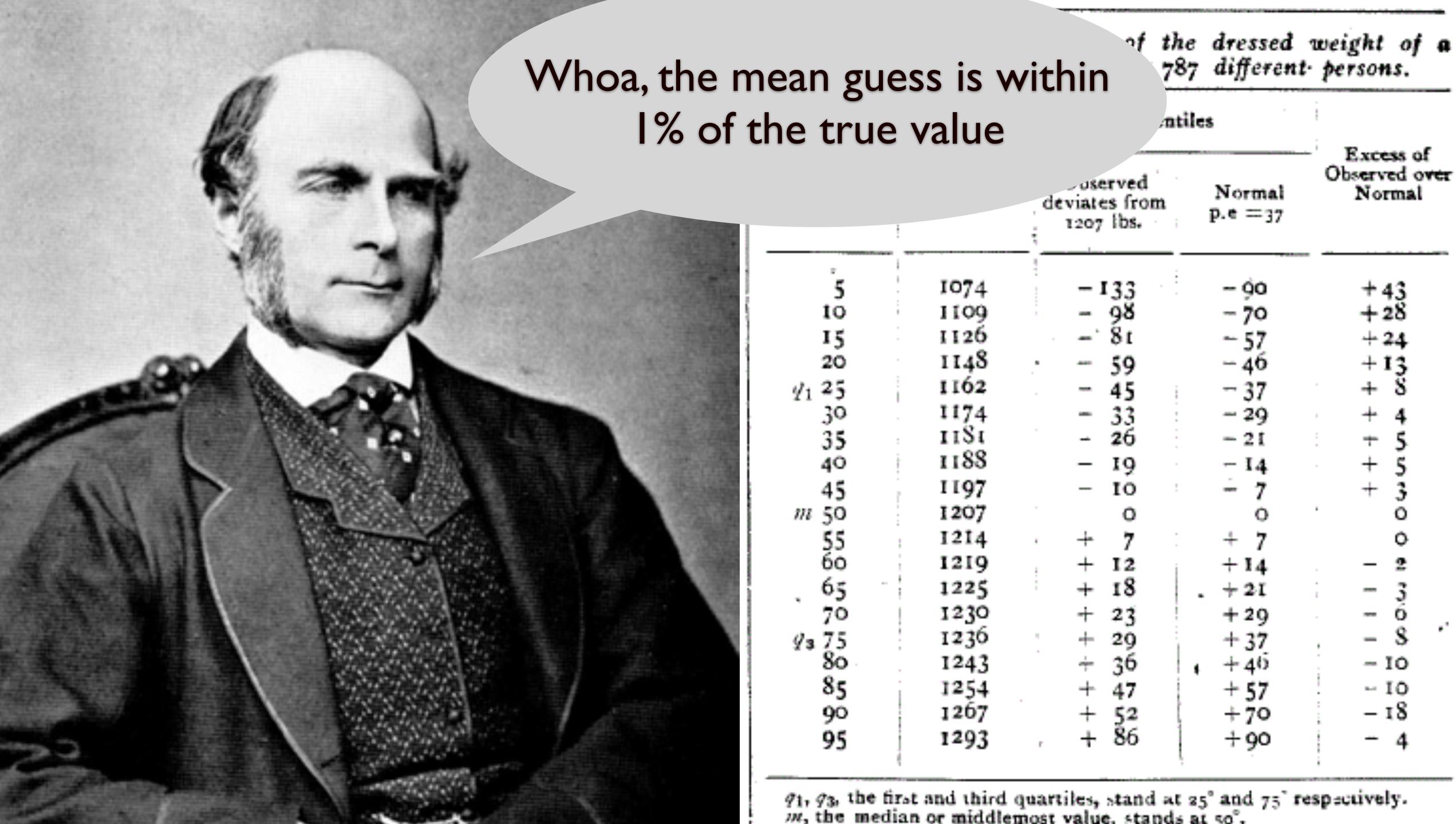
- · The Wisdom of Crowds (and its limits)
- · Crowd algorithms and crowd-powered systems
- · Quality, incentives, motivation
- · The future of work and workers

The Wisdom of Crowds

Crowds are surprisingly accurate at estimation tasks!

- · Who will win the election? How many jelly beans are in the jar? What will the weather be? Is this website a scam?
- · Individually, we all have errors and biases
- · However, in aggregate, we are surprisingly intelligent





let's check our http://hci.st/wise Cesults

What problems can be solved this way?

- · Jeff Howe's definition specified that it required:
 - Diversity of opinion
 - Decentralization
 - · Aggregation of Opinions
- · So any question that has a binary (yes/no), categorical (e.g., win/lose/tie), or interval (e.g., score spread on a football game) outcome

What problems cannot be solved this way?

- Flip the bits!
 - · People all think the same thing
 - · People can communicate
 - No way to combine the opinions
- · For example, writing a short story (is much harder!)

General algorithm

- Ask a large number of people to answer the question
 - Answers must be independent of each other no talking!
 - People must have at least basic understanding of the phenomenon in question.
- Average their responses

Why does this work?

[Simoiu et al. 2017]

- · Independent guesses minimize the effects of social influence
 - · Showing consensus cues such as the most popular guess decrease accuracy
 - · If initial guesses are inaccurate and public, then the crowd never recovers
- · Crowds are more consistent guessers then experts
 - In an experiment, crowds are only at the 67th percentile on average per question...
 - · But at the 90th percentile averaged across questions per domain!

Approaches and algorithms

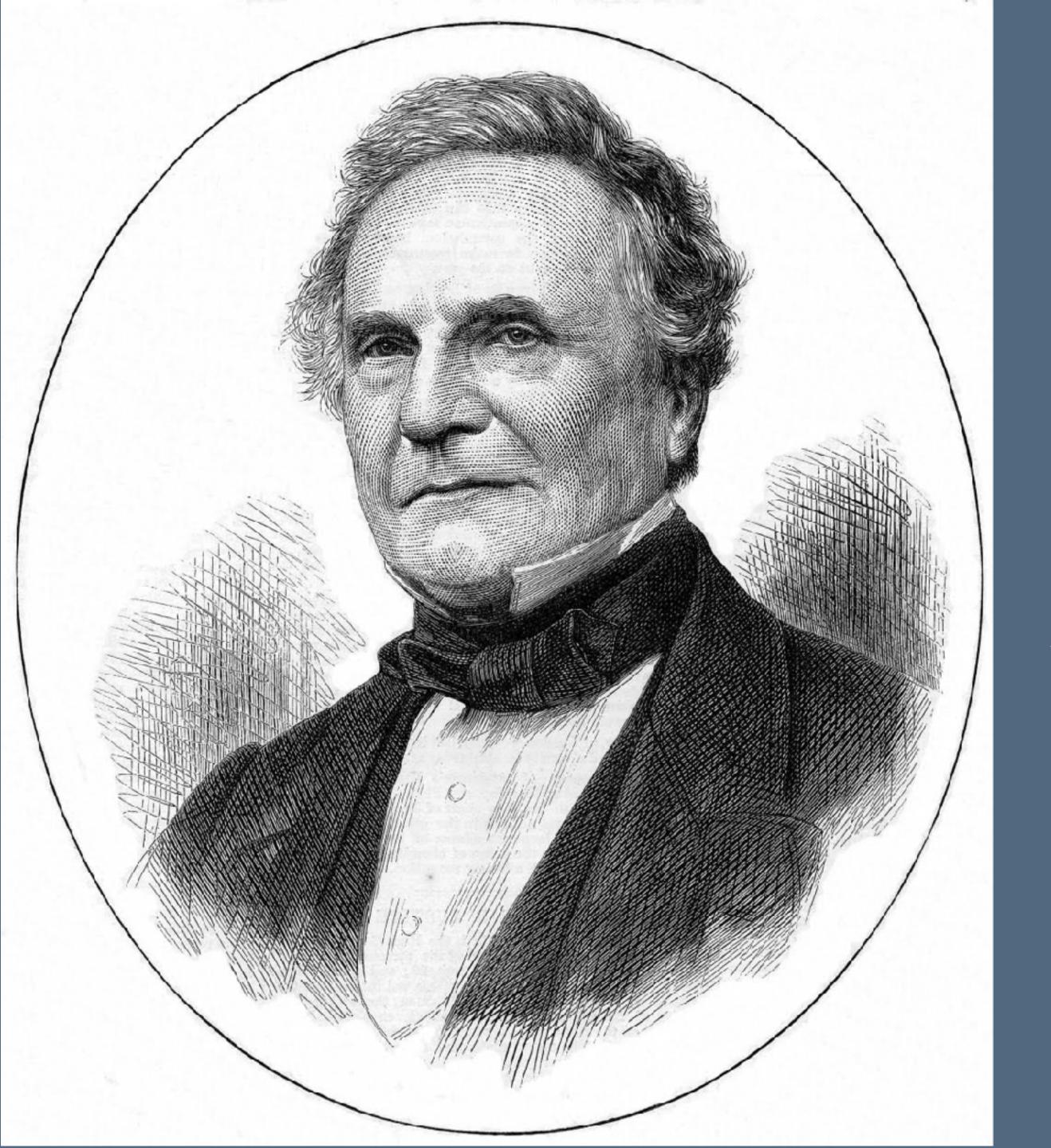
Early crowdsourcing [Grier 2007]

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



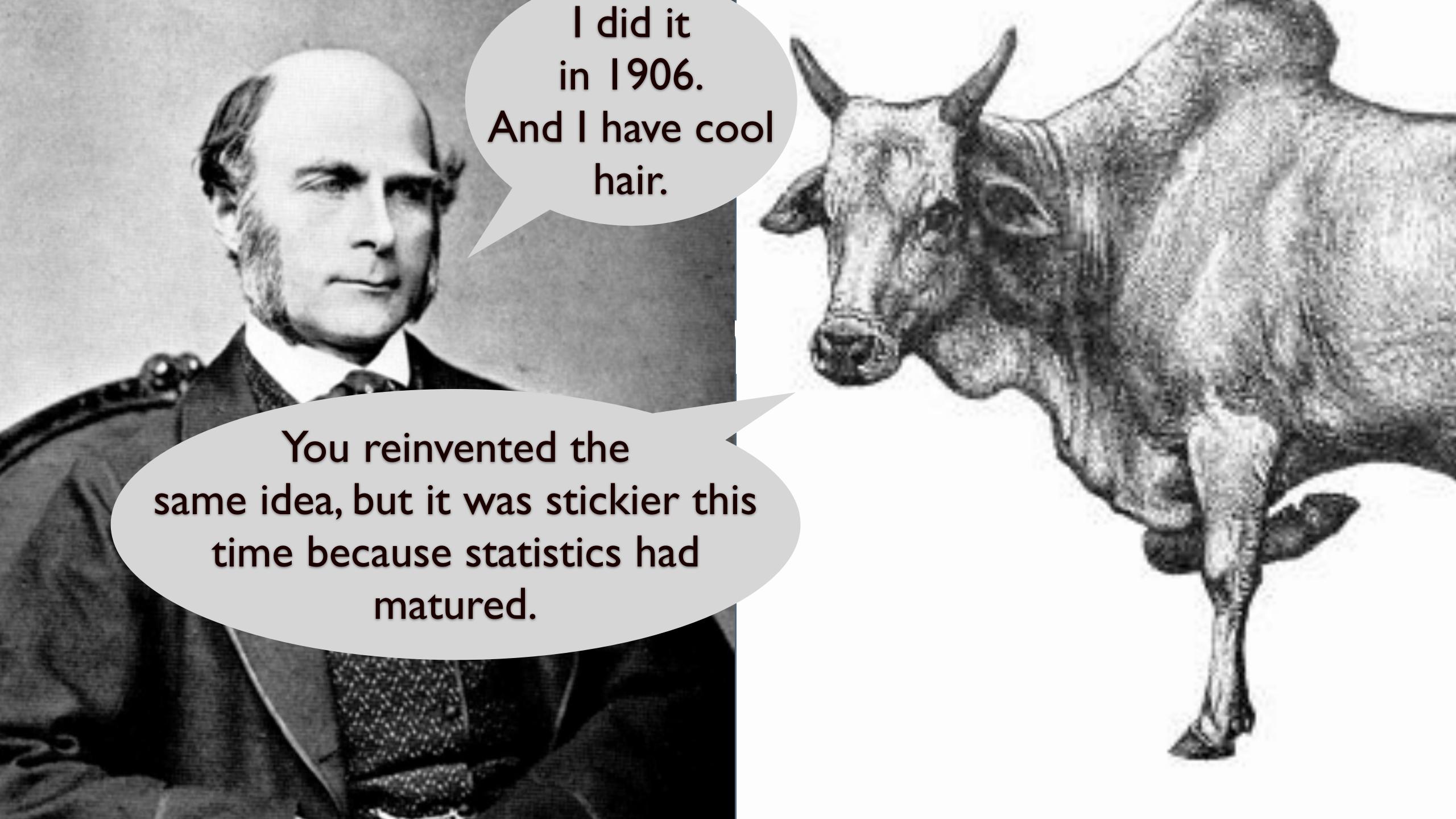
Hevil Maskelyne

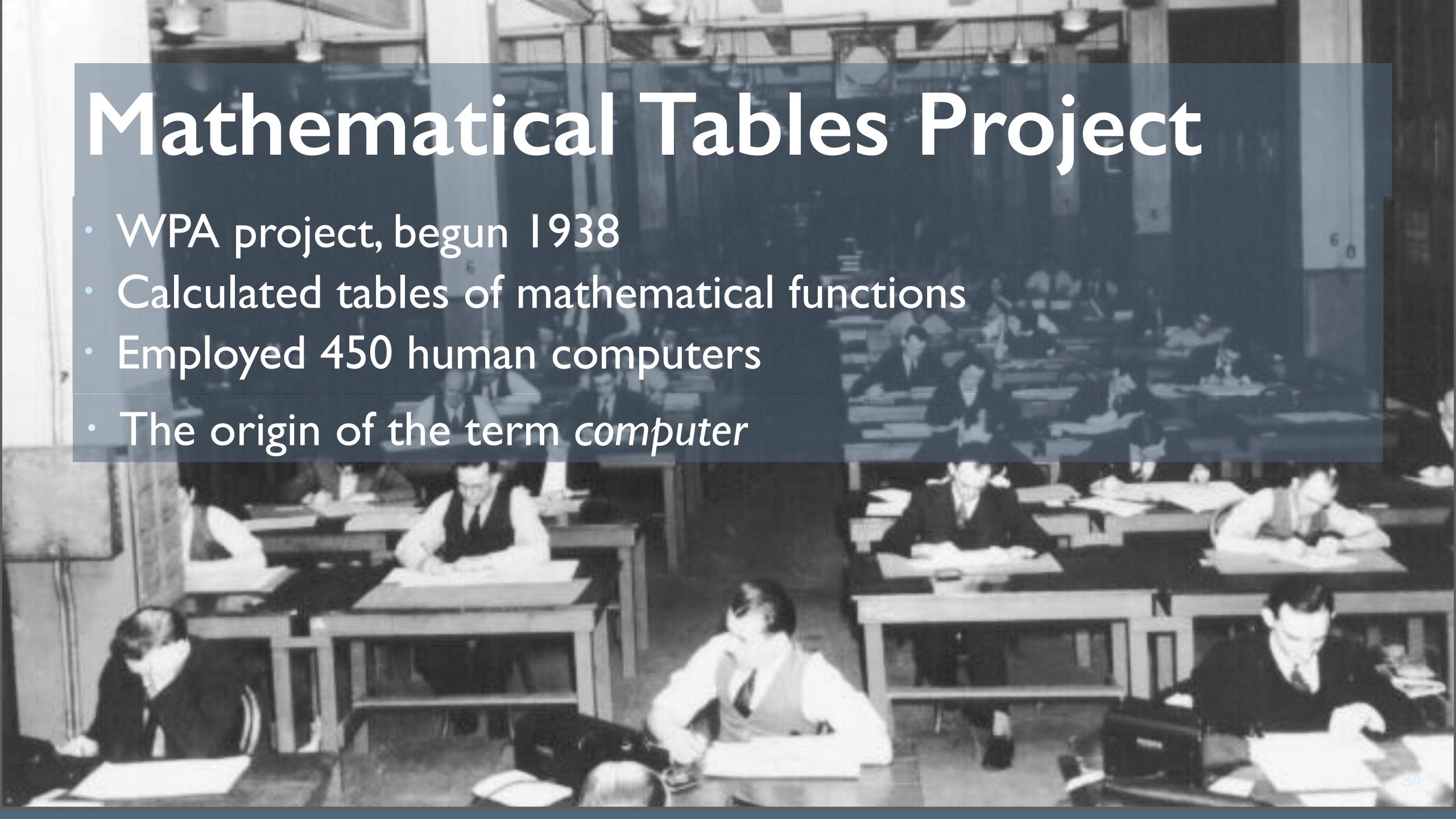
of answer to your Etter of the 23. instant. is conjut. I dantes. Who Dolles at Noon. at Noon ? Compute D'dist. from a star by Cogarithmes, theis of the Jan S. D. M. S. D. M. S. If the hole of the ecliptic Nor of Mis Blave 1 2 20 0 5/ 2 58 291 2 13 2 45 83 48 32 to it in I. There suppose to be to for thein go 3 3 15 46 23 4 28 29 3 4 3 29 3 54 4 54 28 4 By Log. c, P+t, PS=t, PD & PL-PD=LD 5 4 12 35 47 5 4 34 5 Al c, Lf=-c, OD+c, Pf+c, LD 6 4 26 19 33 4 57 26 6 0% 0 = c, 0 + s, PS - s, PD+ c, LD 7 5 10 1231 4 32 51 7 The latter formula must be use I when Pil 5 24 12 11 3 51 51 8 antarce is Cargo, or near go; but may be use da fely 6 8 16 36 2 56 35 9 04 10 6 22 24 16 1 50 19 10 de inall cased. Esample 11 7 6 34 4 0 37 6N 11 a aquila 0 St Visca my 127 20 44 43 0 38 25 112 Vd P. 20-30 - C - - 9. 9715876 13 8 4 54 27 1 51 21 13 Formalhi PS 89 58.30 t- 13.3857588 159 3 0 17 3 51 19 15 8 1 25.25.30 -- C -- 9.9557589 -of --- - t 12.0977246 --- 5. 9.9999879 antazew 89.29.36- + 12.0534829---- 8. co-ar. ---- 0.0000170 16 9 16 49 18 4 31 12 16 24 17:10 0 24 8 4 54 54 17 a arecti 19 10 13 41 505 1 59 18 M 19 10 26 40 41 4 53 8 19 84 If phone of Bo greater than go 2011 9 20 28 4 29 54 20 00 Then At AB = LB asin the sockene 21 11 21 42 30 3 54 21 21 220 3 49 35 3 It the ilanets placed, excepting the



Charles Babbage

Two people doing the same task in the same way will make the same errors.





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

Get Another Label

[Sheng, Provost, Ipeirotis, KDD '08]

https://github.com/ipeirotis/Get-Another-Label

- We need to answer two questions simultaneously: (1) What is the correct answer to each question? and (2) Which participants' answers are most likely to be correct?
- Think of it another way: if people are disagreeing, is there someone who is generally right?
- Get Another Label solves this problem by answering the two questions simultaneously

Get Another Label

[Sheng, Provost, Ipeirotis, KDD '08]

- Inspired by Expectation Maximization (EM) algorithm from artificial intelligence.
- · Use the workers' guesses to estimate the most likely answer for each question. Use those answers to estimate worker quality. Use those estimates of quality to re-weight the guesses and re-compute answers. Loop.

Given worker Initialize worker responses, {r}, to a set parameters, $P_w(r|a)$, to random values. of questions Using the current values of the worker parameters, $P_w(r|a)$, and their responses, estimate the probability of each answer. Iterate until parameters & answer probabilities stop changing Given the relative answer probabilities, compute new parameter values that maximize the likelihood of each worker's response.

Bayesian Truth Serum (BTS)

[Prelec, Seung, and McCoy Nature '04]

- Inspiration: people with accurate meta-knowledge (knowledge of how much other people know) are often more accurate
- · So, when asking for the estimate, also ask for each person's predicted empirical distribution of answers
- · Then, pick the answer that is more popular than people predict

Bayesian Truth Serum (BTS)

[Prelec, Seung, and McCoy Nature '04]

- "When will HBO have its next hit show?"
 I year / 5 years / 10 years
- · "What percentage of people do you think will answer each option?"
 - I year / 5 years / 10 years
- An answer that 10% of people give but is predicted to be only 5% receives a high score

Bayesian Truth Serum (BTS)

[Prelec, Seung, and McCoy Nature '04]

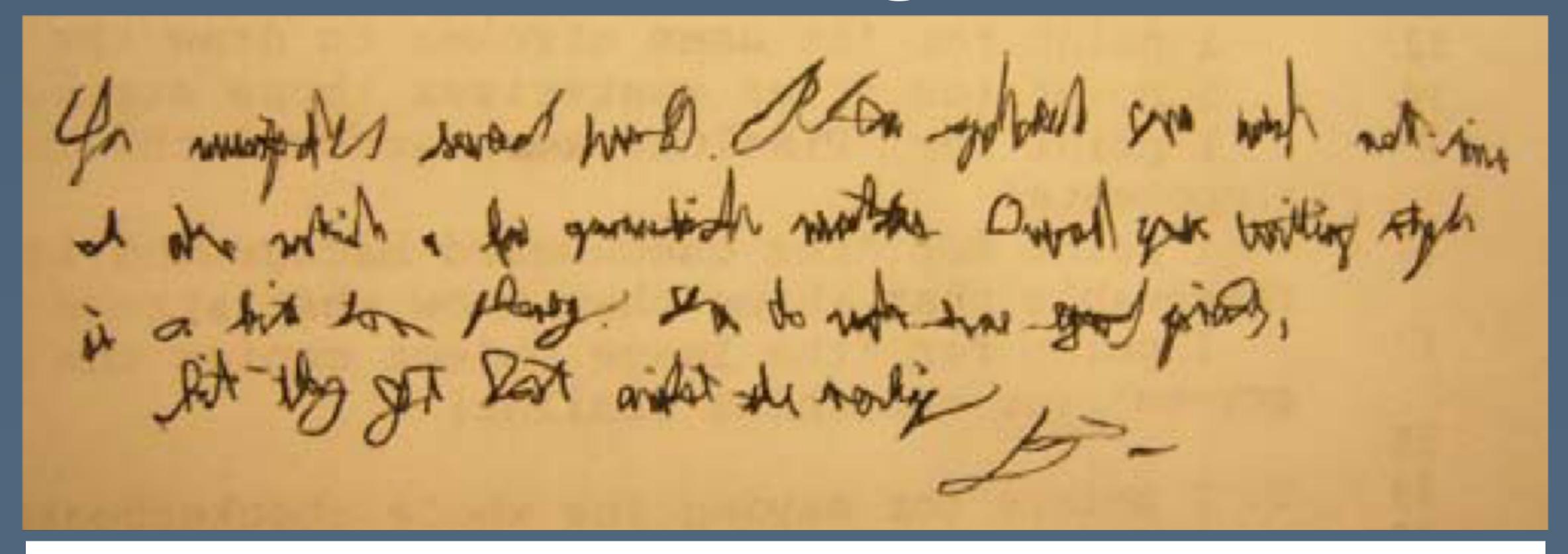
Calculate the population endorsement frequencies $\bar{x_k}$ for each option k and the geometric average of the predicted frequencies $\bar{y_k}$

Evaluate each answer according to its information score:

$$\log \frac{\bar{x}_k}{\bar{y}_k}$$

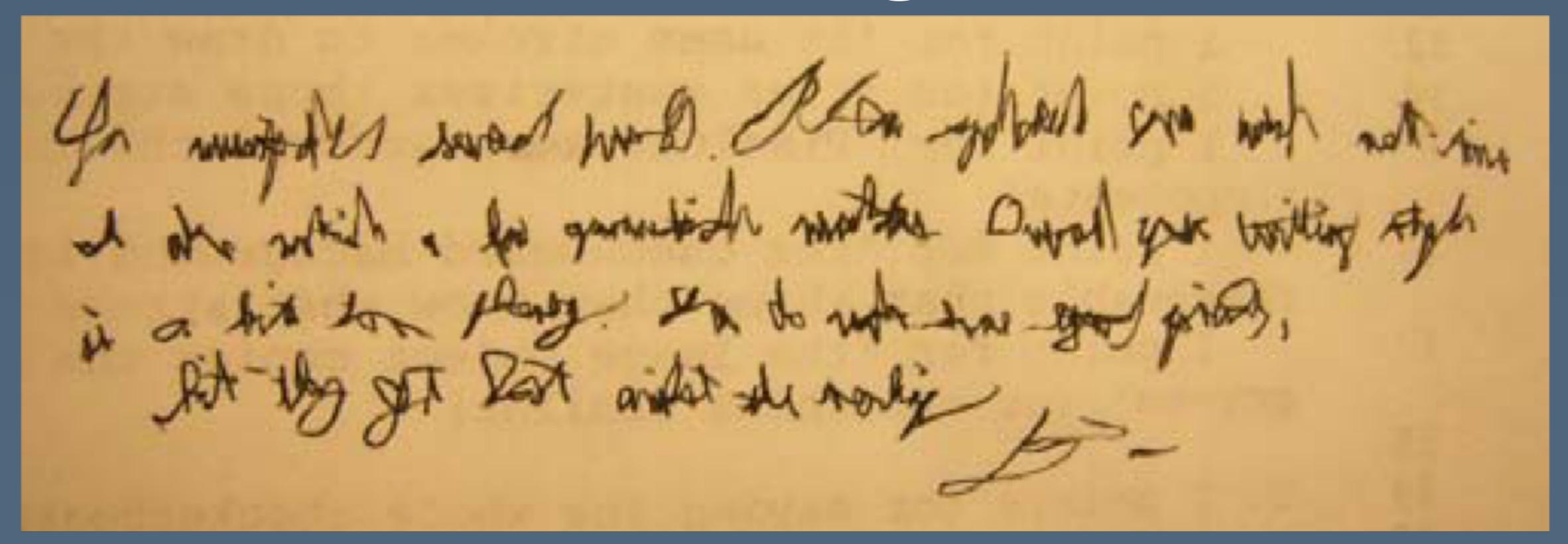
Crowd algorithms

Iterative crowd algorithm





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)

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

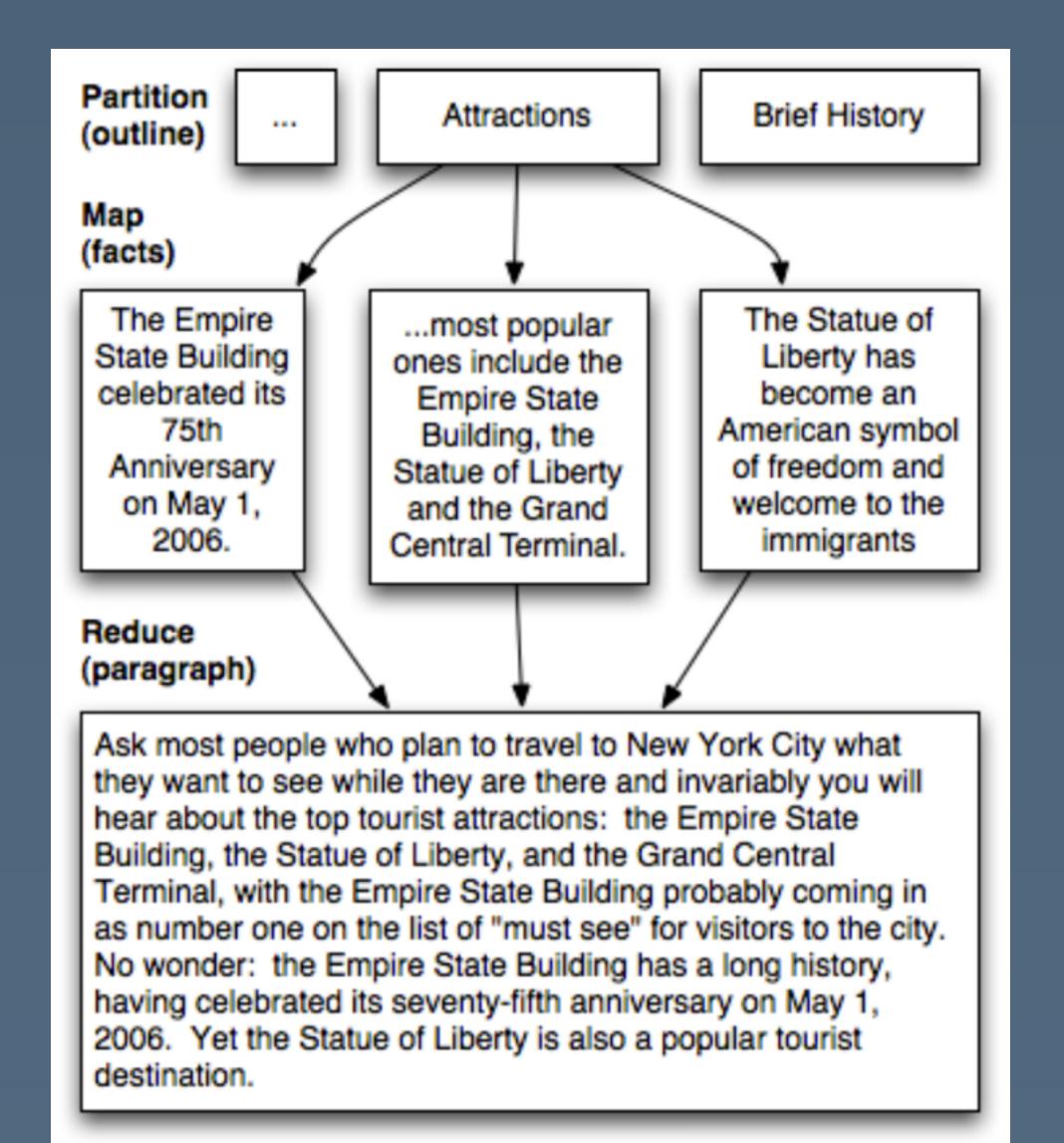
Crowdsourcing algorithms

- 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

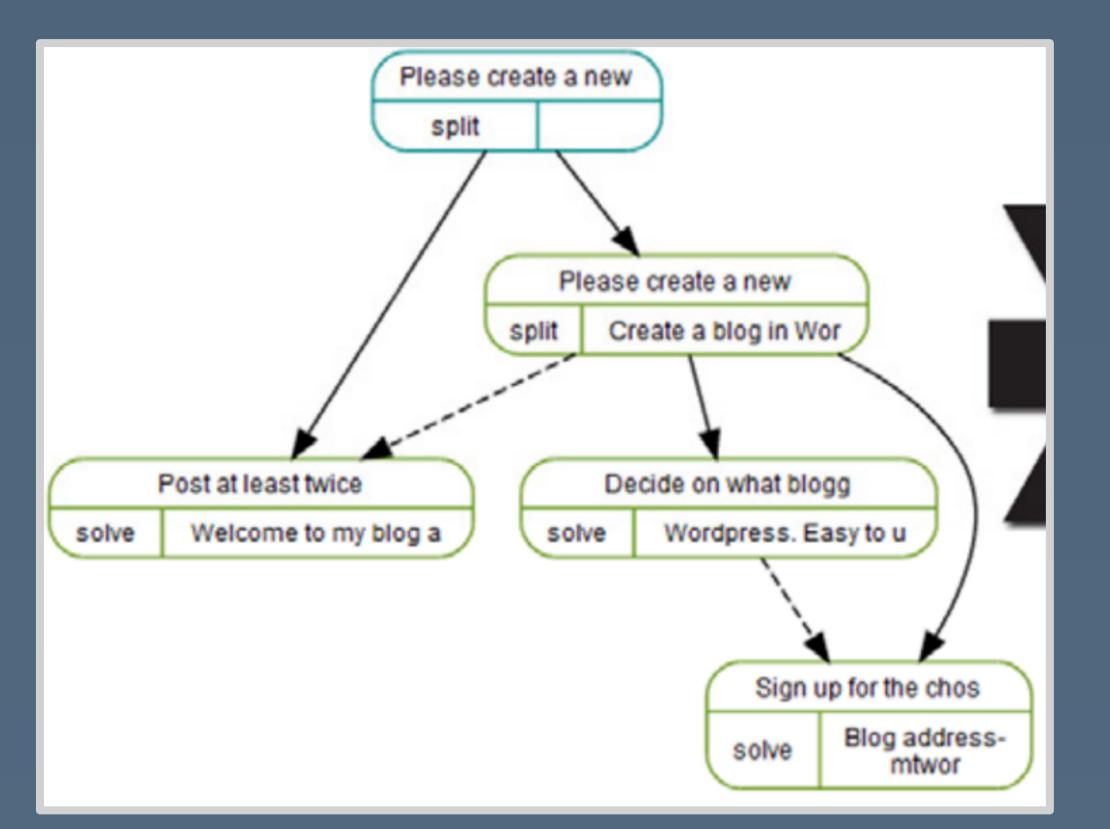


Turkomatic

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

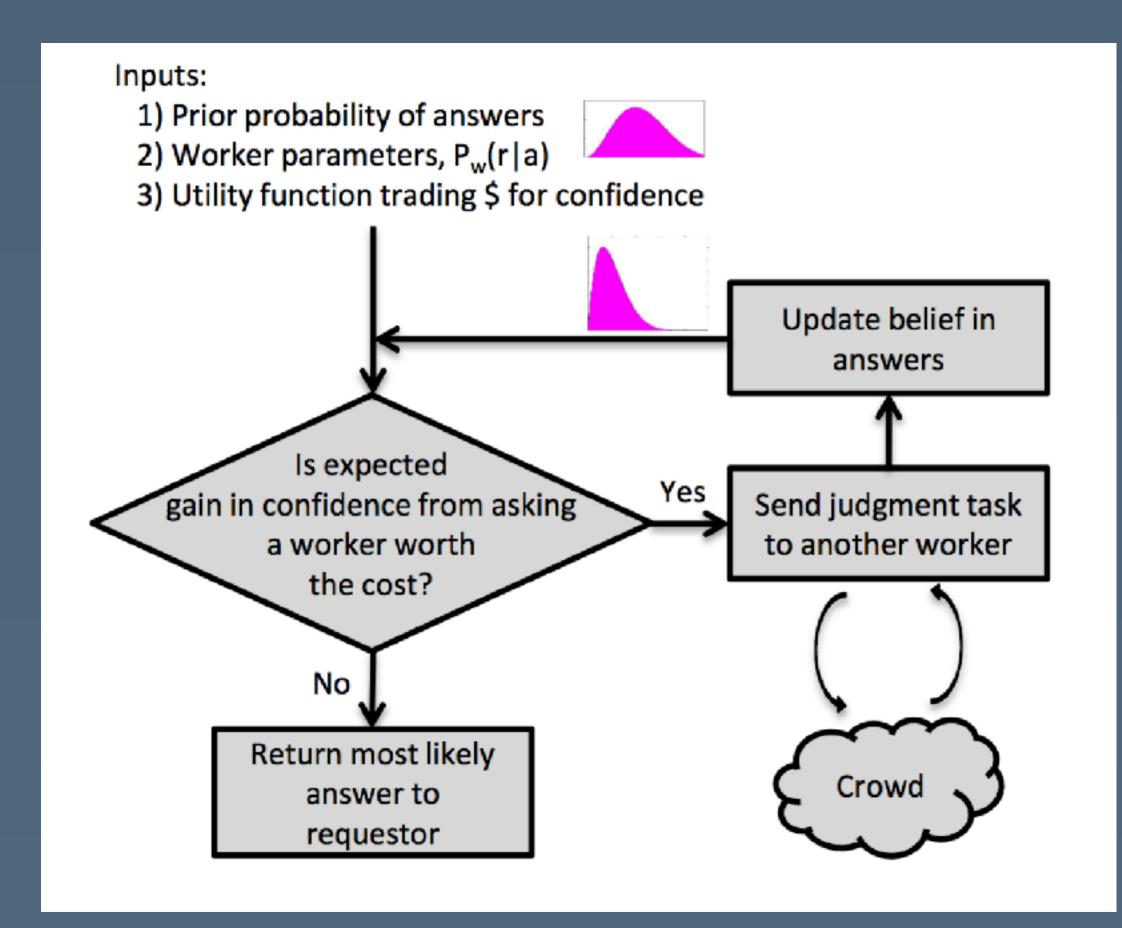
- Let the workers decide on task design
- Is a task too complicated for \$D? If so, ask for subtasks and recurse. If not, do it yourself.

 Creating a blog with content:



Als guiding crowds

- TurKontrol [Dai, Mausam and Weld, '10]: Workflow planning as decision-theoretic optimization
 - 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?
- · Others: Which question should we ask? Do we trust the answer?



Crowds guiding Als

· Crowdsourcing can help artificial intelligence algorithms know where their "unknown unknowns" are: e.g., confidently incorrect predictions

Beat the Machine

Identify pages that contain hate speech on the web

In this task, your goal is to find websites which advocate hostility or aggression toward individuals or groups on the basis of race, religion, gender, nationality, ethnic origin, or other involuntary characteristics.

Your input will be verified by other, trusted humans, and you will receive the bonus payment only if your submission indeed belongs to the correct category.

The URLs that you submit will be used to examine the accuracy of our automatic classifier. You get more bonus points if you submit URLs that are not in our database and trick our classifier to classify the URL into the incorrect category. So, the better you are in "beating the machine", the more bonus points you get.

Remeber 5000 bonus points = 1\$.

Submit 1 urls:

Finish work

Already submited urls:

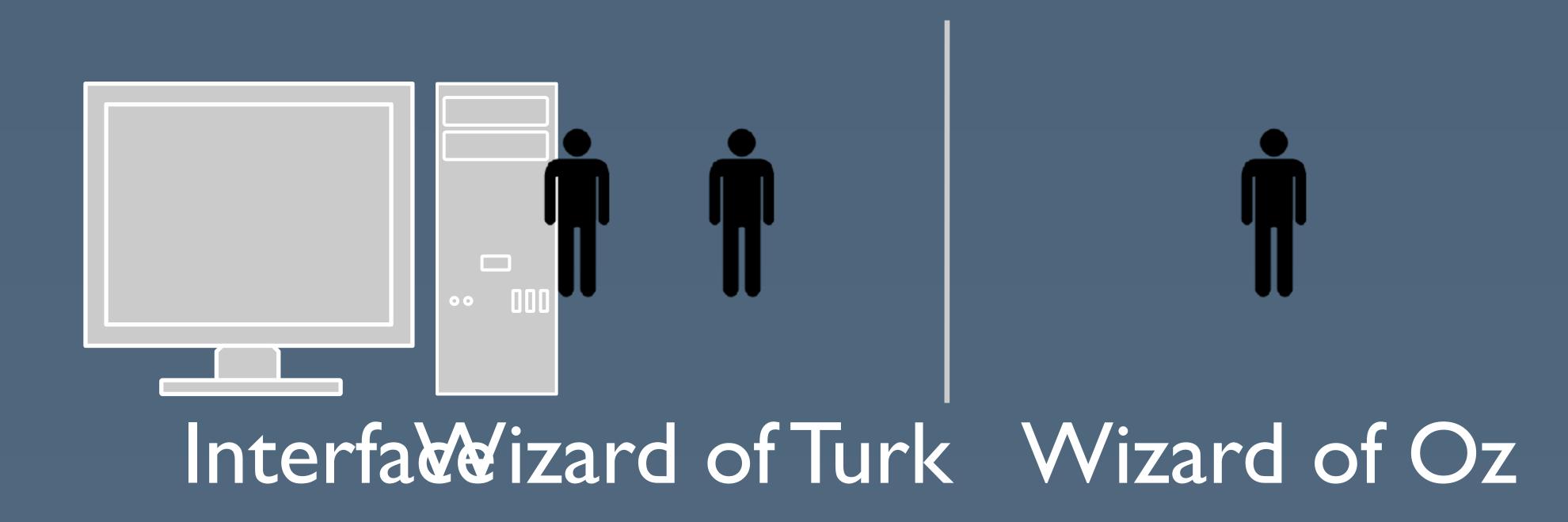
- http://fiber.
- http://pages.stern.nyu.edu/~panos/, We are pretty confident that this is not a hate speech page. If this is a porn page, you will get maximum a bonus of 1000 points
- http://www.ferris.edu/jimcrow/caricature/, We are pretty confident that this
 is a hate speech page, sorry no bonus
- http://www.resist.com/ownersmanual.htm, We are pretty confident that this
 is a hate speech page, sorry no bonus

Maximum possible bonus for this task: 1000

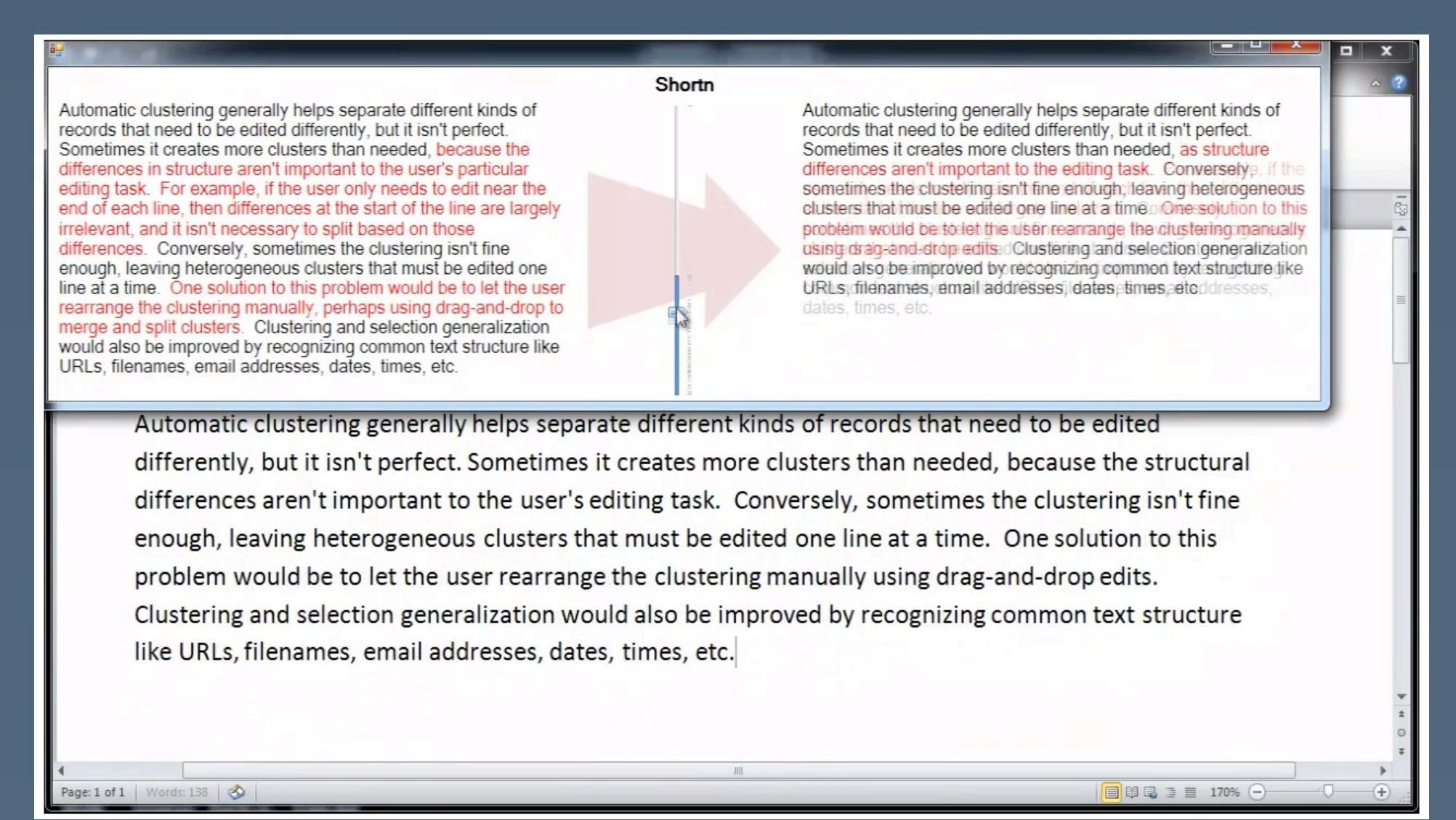
Crowd-powered systems

Why do it?

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



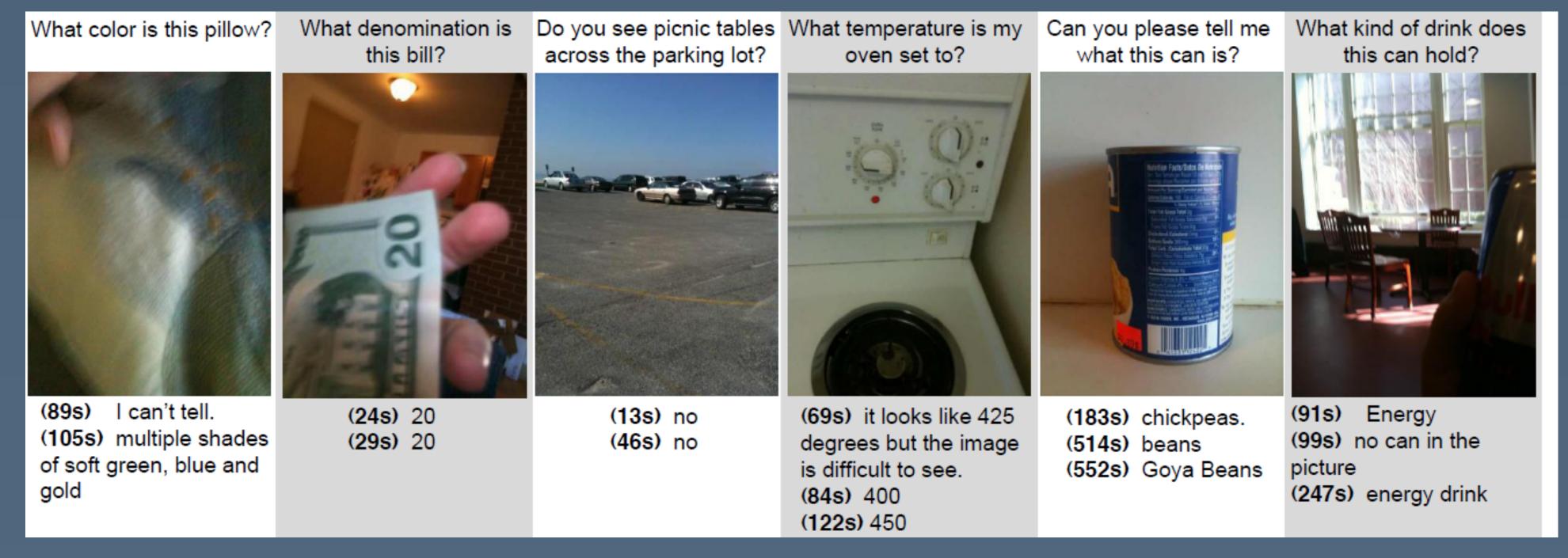
Soylent



VizWiz

[Bigham et al., UIST '10]

Visual question answering for the blind



 I 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 'II]
 - MIT: Qurk [Marcus et al., CIDR 'II]
 - Stanford: Deco [Parameswaran et al. '11]
- Tackling many important optimization questions: e.g., joins, ranking, sorting

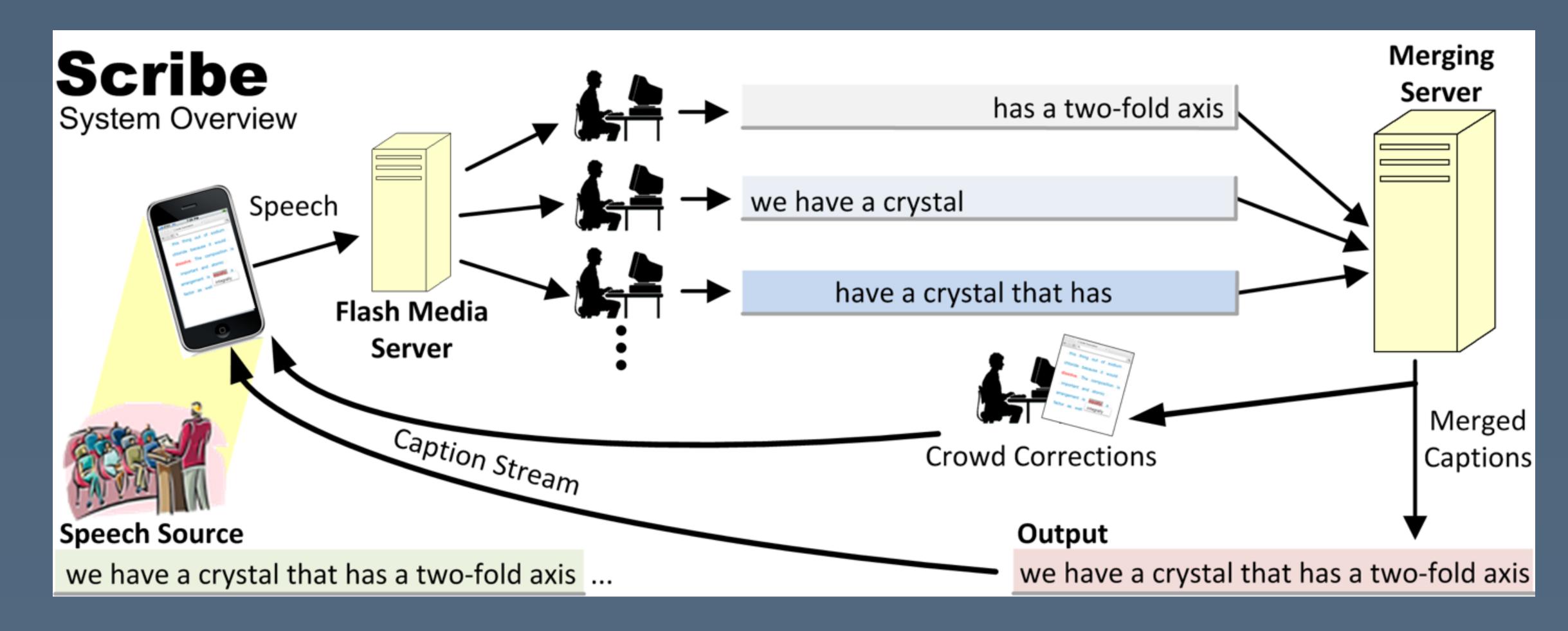


Reatime crewdseurcing

[Bernst

Realtime crowdsourcing

· Realtime captioning using shotgun gene sequencing techniques



Quality, incentives, motivation

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 answers

- "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]

(this is pretty mean to workers! it was an earlier era...)

Problem: manipulation

Boaty McBoatface: What You Get When You Let the Internet Decide



A computer image of the research vessel, which is still being designed and is scheduled to set sail in 2019. The Natural Environment Research Council

TIME

The World's Most Influential Person Is...

By TIME Staff | Monday, Apr. 27, 2009









Read Later

In a stunning result, the winner of the third annual TIME 100 poll and new owner of the title World's Most Influential Person is moot. The 21-year-old college student and founder of the online community 4chan.org, whose real name is Christopher Poole, received 16,794,368 votes and an average influence rating of 90 (out of a possible 100) to handily beat the likes of Barack Obama, Vladimir Putin and Oprah Winfrey. To put the magnitude of the upset in perspective, it's worth noting that everyone moot beat out actually has a job.



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'II]

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

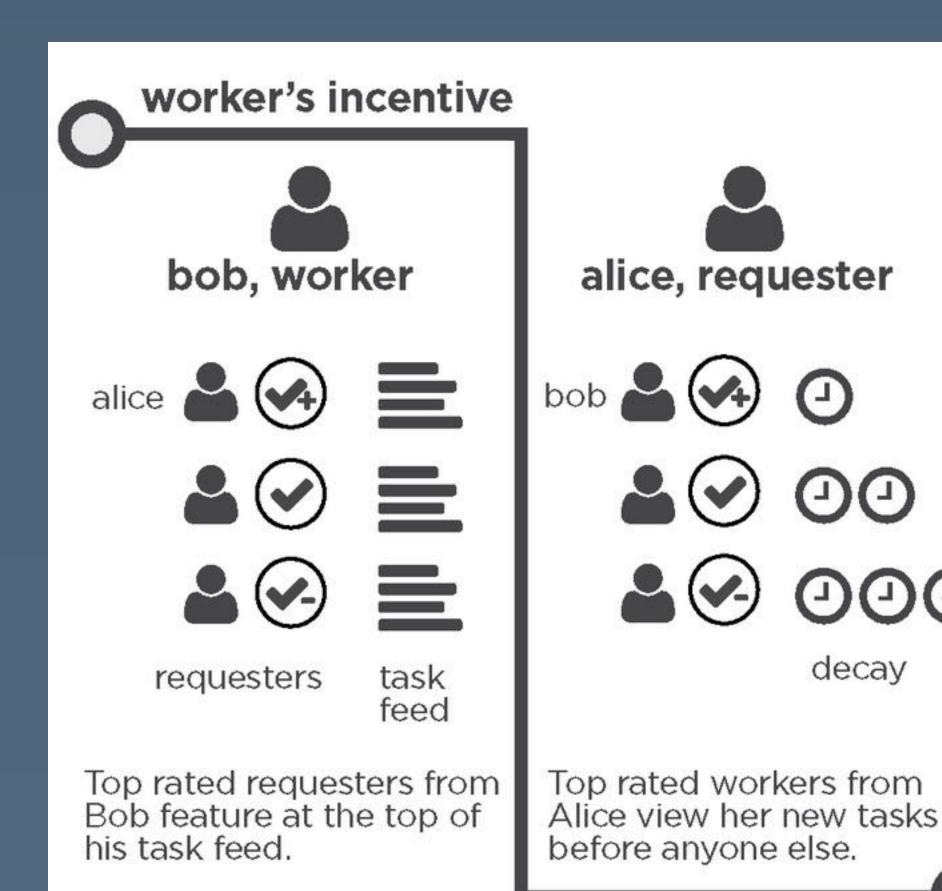
Judging quality implicitly

[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

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



requester's incentive

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

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 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

Future of work and workers

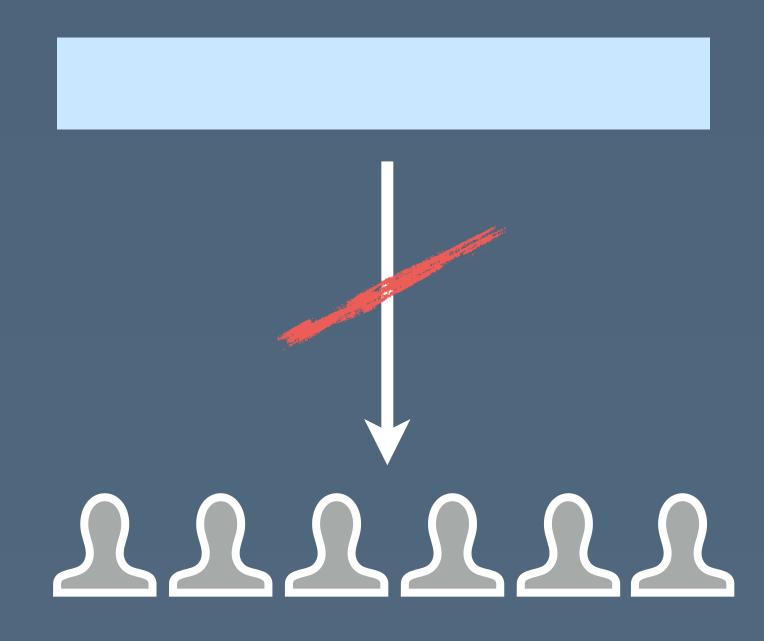
Communitysourcing

Engaging Local Crowds to Perform Expert Work Via Physical Kiosks

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

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

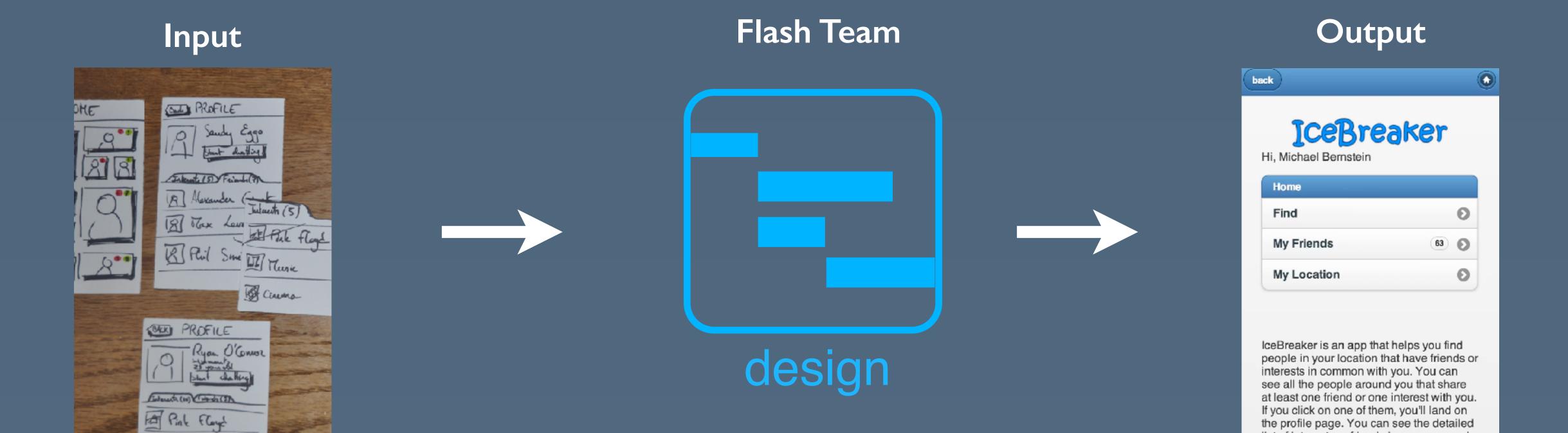


programmer designer video editor musician statistician

Flash Teams

[Retelny et al., UIST '14]

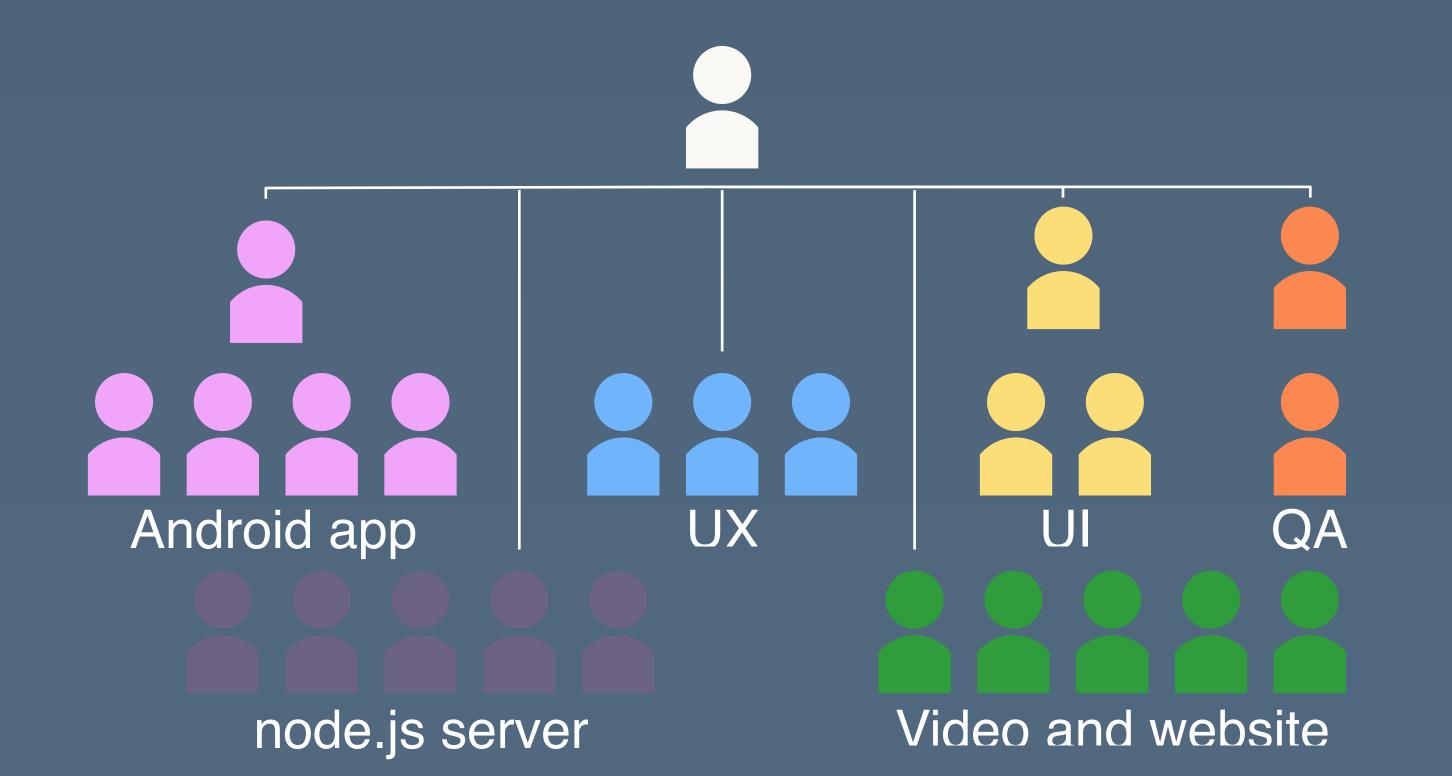
Computationally-guided teams of crowd experts supported by lightweight, reproducible and scalable team structures.

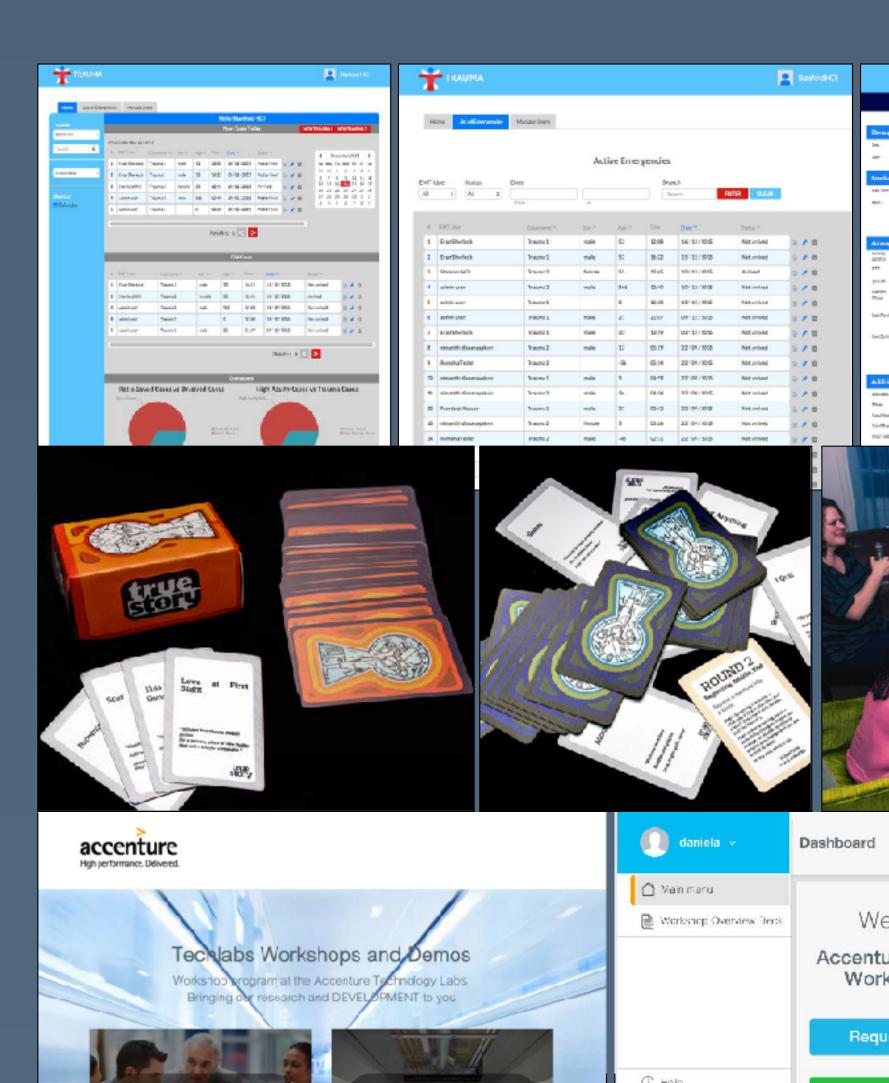




Flash Organizations [Valentine et al., CHI'17]

Achieve complex goals by structuring crowds as organizations, not workflows

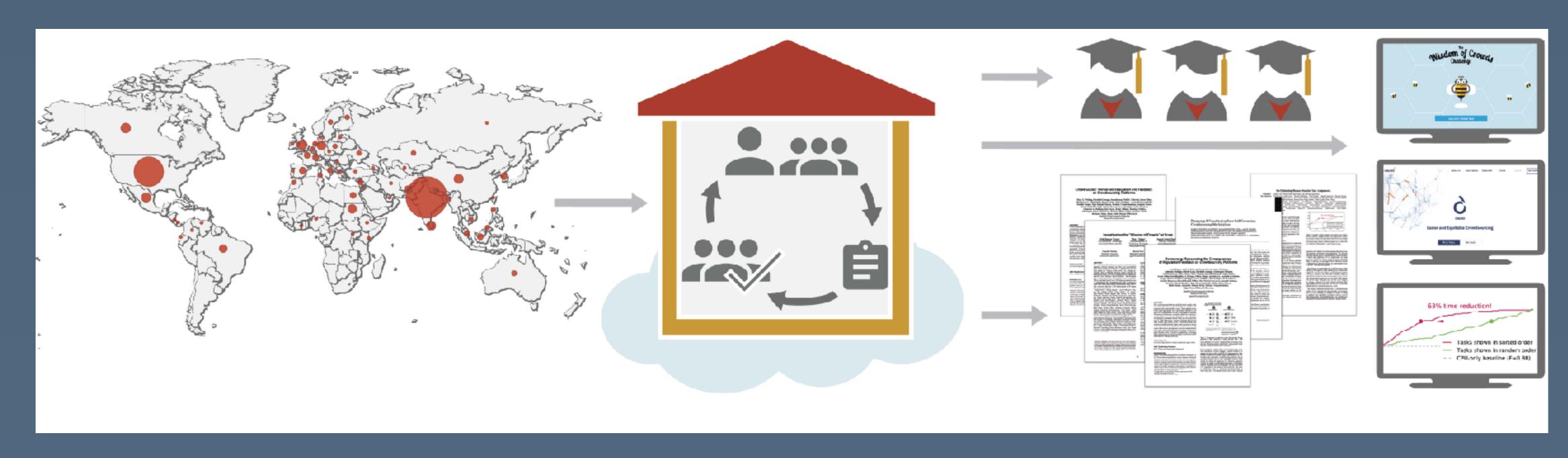


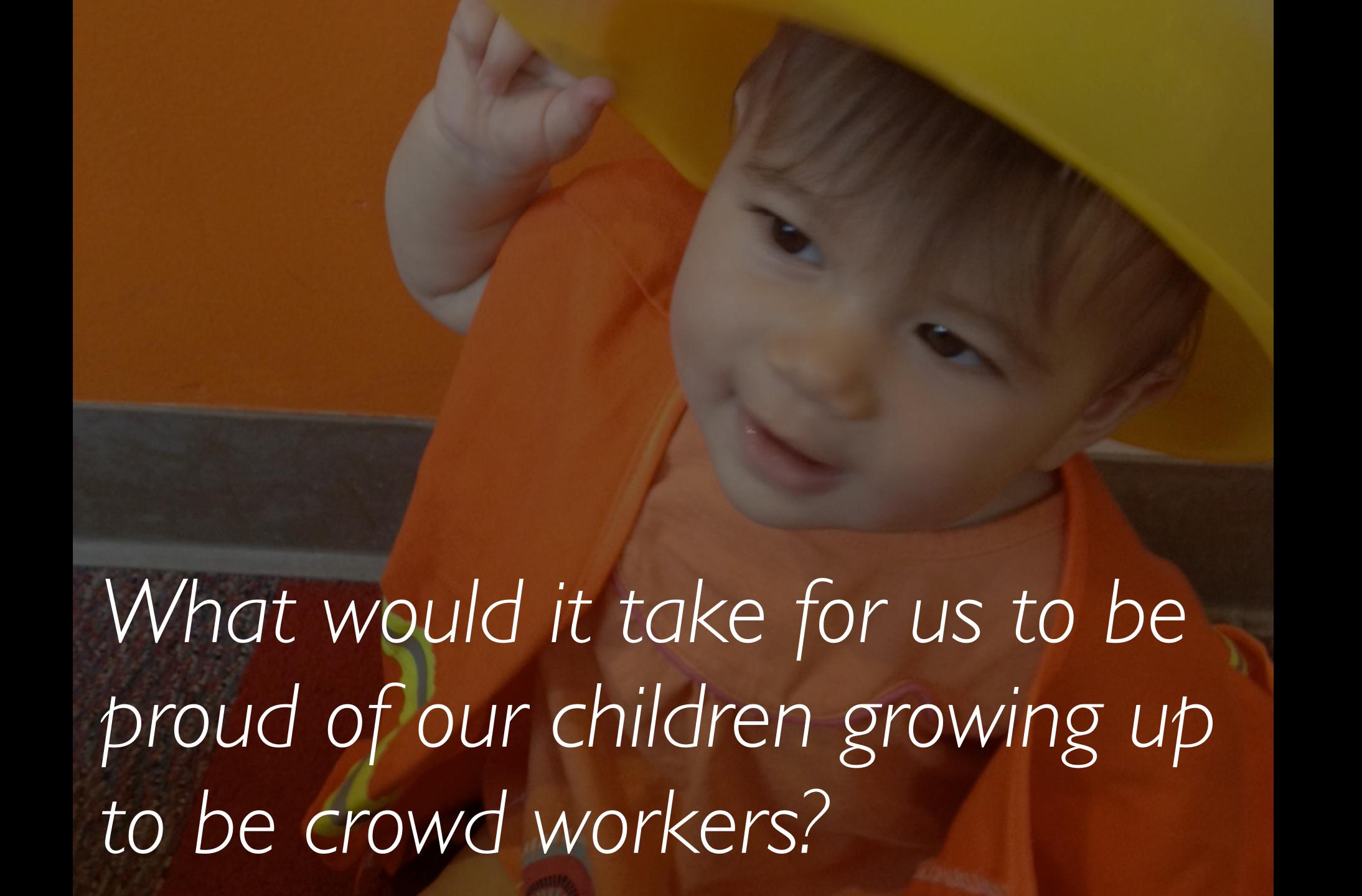


Crowd research

[Vaish et al., UIST '17]

· Crowdsourcing as a route to empower upward career and educational mobility through research experiences





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 Turke not only actual human beings, but people who deserve respect, fair treatment and open communication. Donate Donate Donations will go towards reaching out to more Turkers via a HIT on Manning and 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

Skills for crowdsourcing research

- Grounding in social psychology, organizational behavior, sociology
- · A design/systems orientation or a behavioral orientation
- Earlier focus on microtasking giving rise to a more modern focus on interdependent coordination at scale