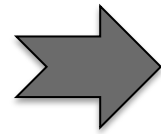


Crowdsourcing: Beyond Label Generation

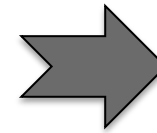
Jenn Wortman Vaughan

Microsoft Research

What do you think of when you think
of crowdsourcing?



“Crowd”



guitar

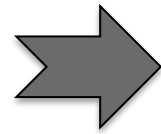
man

Are there better ways to make use of the crowd?

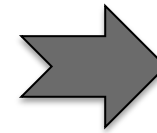
What other problems can the crowd solve?

Part 1: The Potential of Crowdsourcing

1. Direct Applications to Machine Learning
2. Hybrid Intelligence Systems
3. Large Scale Studies of Human Behavior



“Crowd”



guitar

man

Part 2: The Crowd is Made of People

- What motivates workers?
- Are workers independent?
- Are workers honest?



What does this teach us about how to effectively interact with crowd?

Hint: Be respectful. Be responsive. Be clear.

Extensive notes, slides, and eventually
video at

[http://www.jennwv.com/projects/
crowdtutorial.html](http://www.jennwv.com/projects/crowdtutorial.html)

Part 1:

The Potential of Crowdsourcing

The Potential of Crowdsourcing

1. Direct Applications to Machine Learning
2. Hybrid Intelligence Systems
3. Large Scale Studies of Human Behavior

Generating Labeled Data

Learner



Learner

Aggregation
of noisy
labels



“dog” “dog” “cat”



“cat” “cat” “cat”

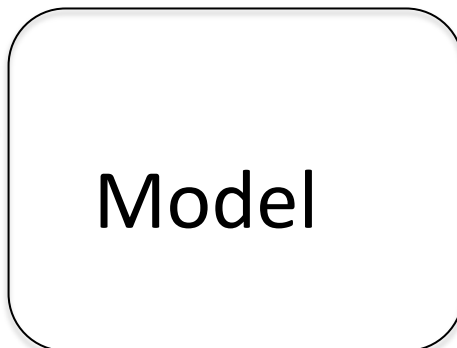




“dog” “dog” “cat”



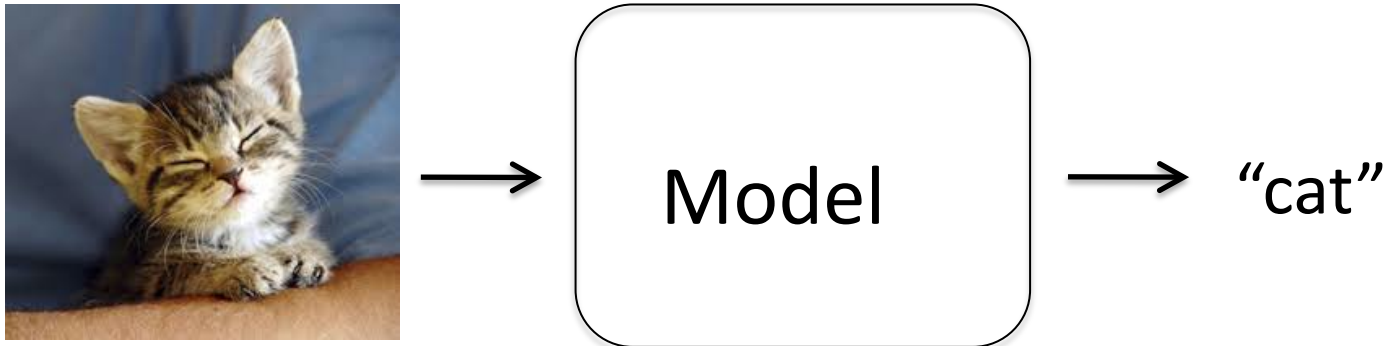
“cat” “cat” “cat”



Used to annotate
medical images, label
text, extract and label
features of scenes.

Inspired huge amounts
of algorithmic work on
aggregation.

The ultimate goal is to take humans out of the loop.



Crowdsourcing for Evaluation

Evaluating Topic Models



cheese
kale
bread
steak
mushroom
pizza
...

election
senate
bill
delegate
president
proposal
...

To be useful for data exploration or summarization, topics must be **human-interpretable!**

[Chang et al., 2009]

Evaluating Topic Models

Word intrusion task:

mushroom, kale, cheese, bread, election, steak

worker
accuracy

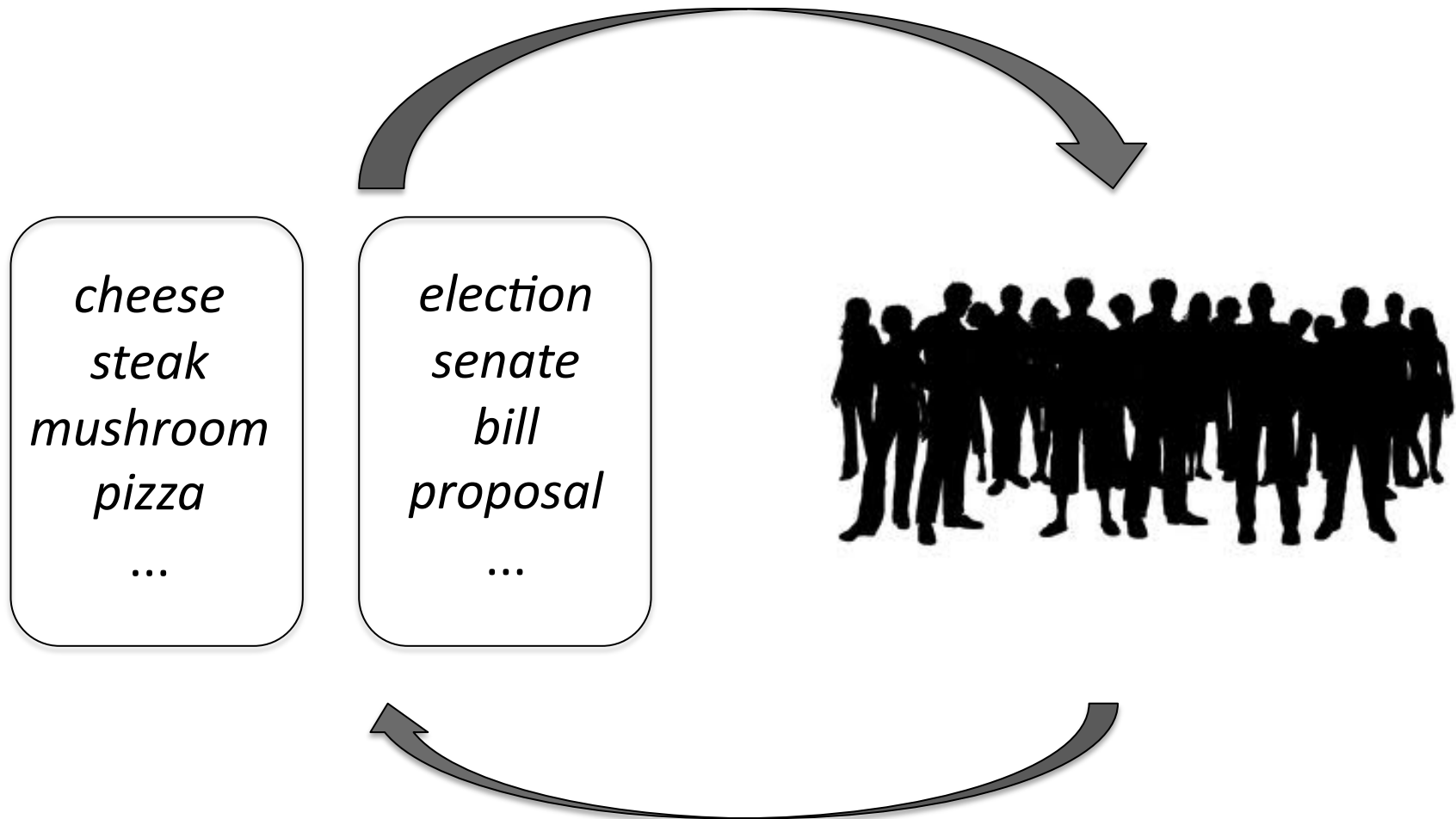


human-
interpretability

Previous measures of success (e.g., log likelihood of held-out data) do not imply interpretability!

[Chang et al., 2009]

Evaluating Topic Models



Human Debugging of Machine Learning Models

Human Debugging

- Semantic segmentation: partition an image into semantically meaningful parts, label each part

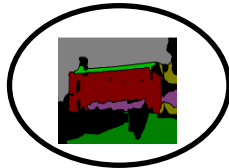


[Parikh & Zitnick, 2011; Mottaghi et al., 2013]

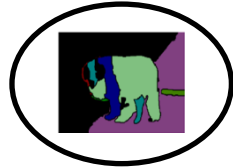
Human Debugging

- Semantic segmentation: partition an image into semantically meaningful parts, label each part

CRF model



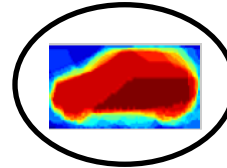
segment
classifier



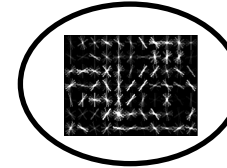
supersegment
classifier



scene
classifier



shape
prior

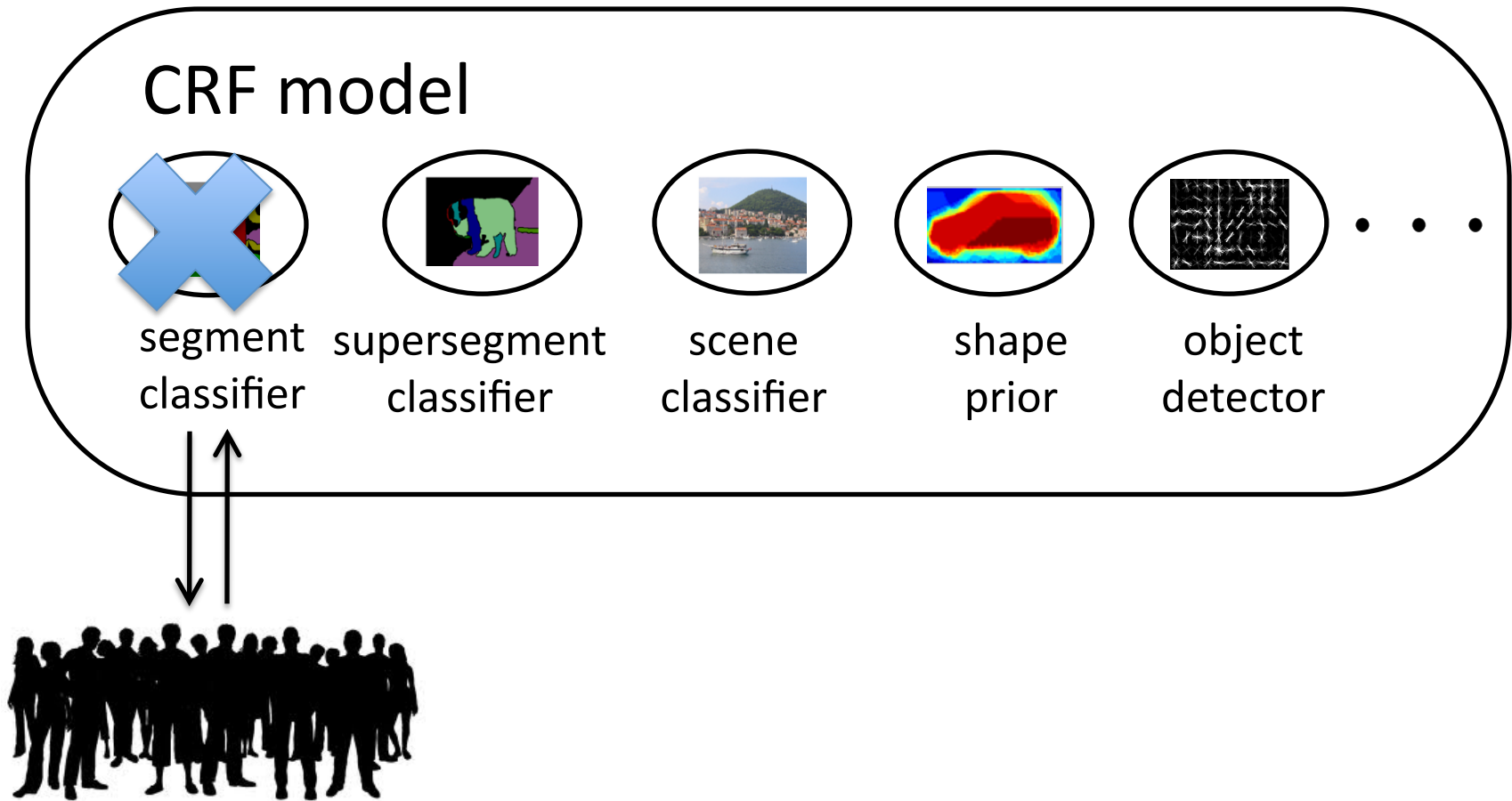


object
detector



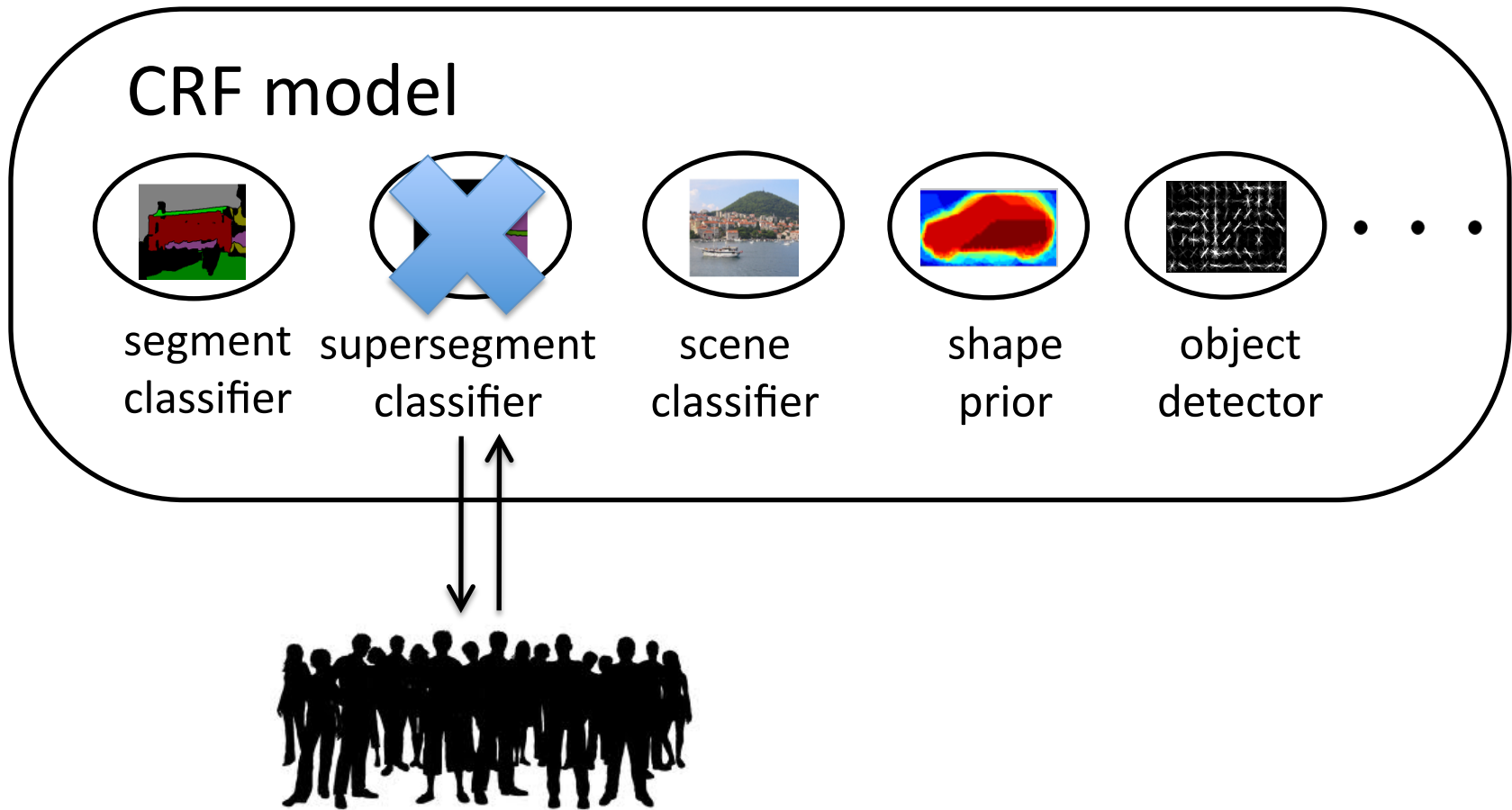
Which component is the weakest link?

Human Debugging



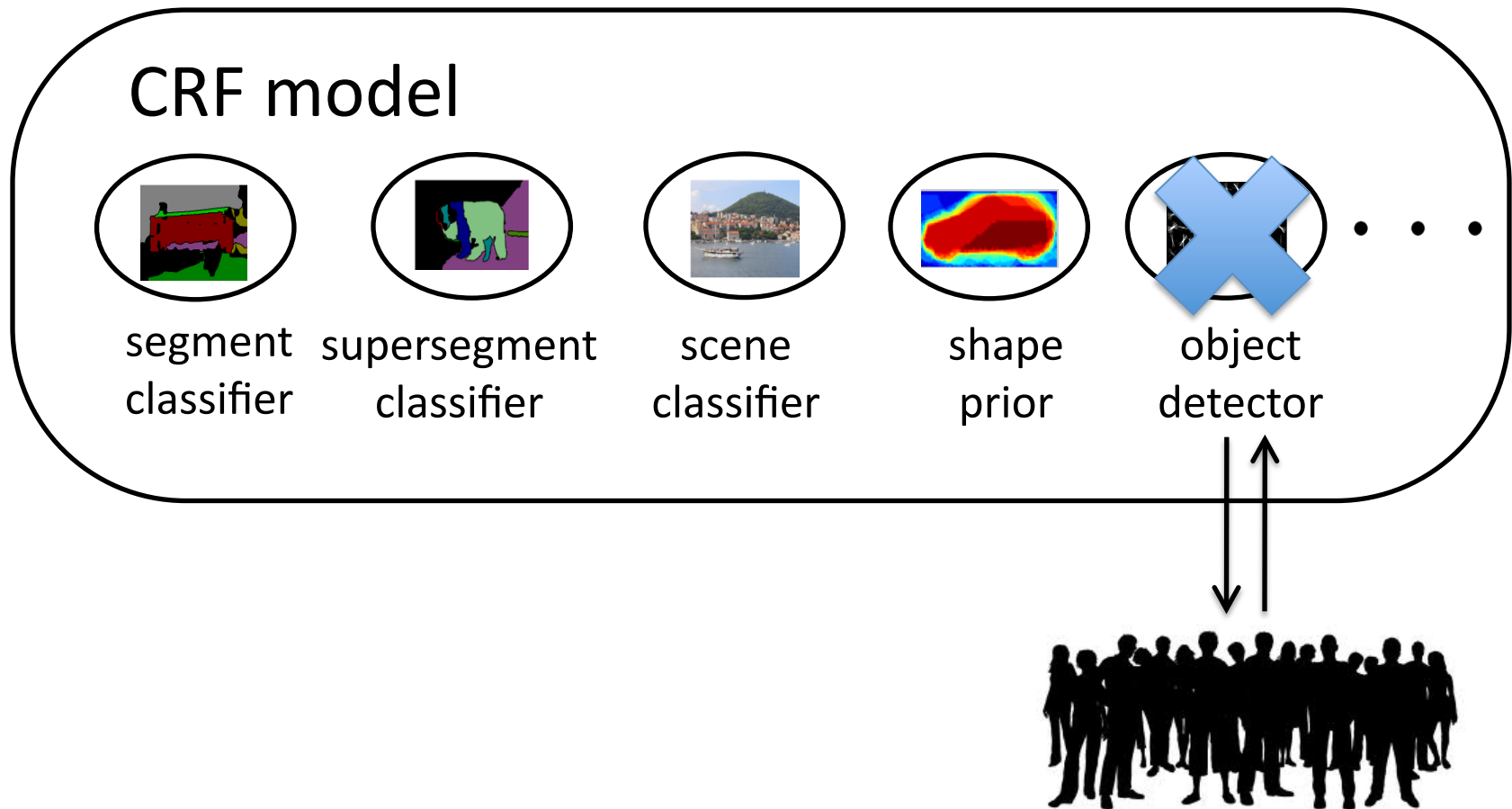
[Parikh & Zitnick, 2011; Mottaghi et al., 2013]

Human Debugging



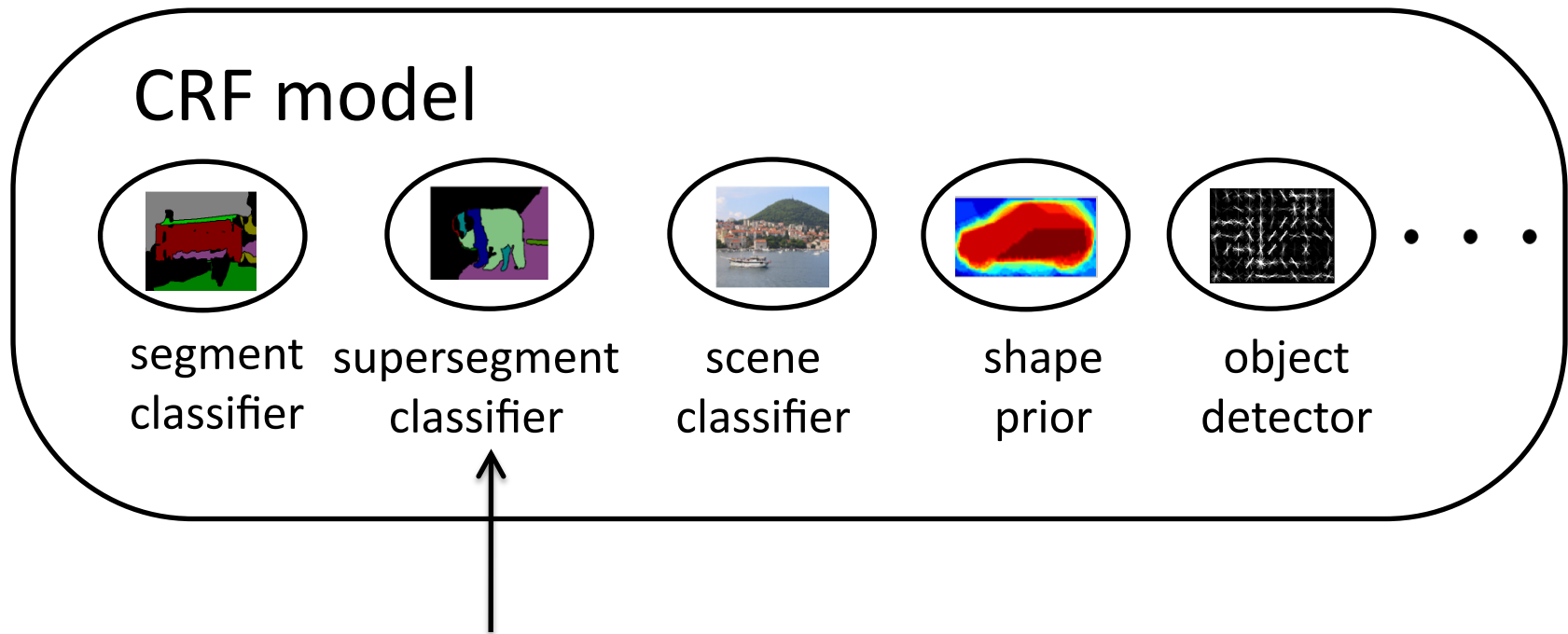
[Parikh & Zitnick, 2011; Mottaghi et al., 2013]

Human Debugging



[Parikh & Zitnick, 2011; Mottaghi et al., 2013]

Human Debugging



Humans less accurate at task, but
system performance **still improved**

[Parikh & Zitnick, 2011; Mottaghi et al., 2013]

Crowdsourcing Similarity

Human Clustering



[Gomes et al., 2011]

Human Clustering



flags



no flags

[Gomes et al., 2011]

Human Clustering

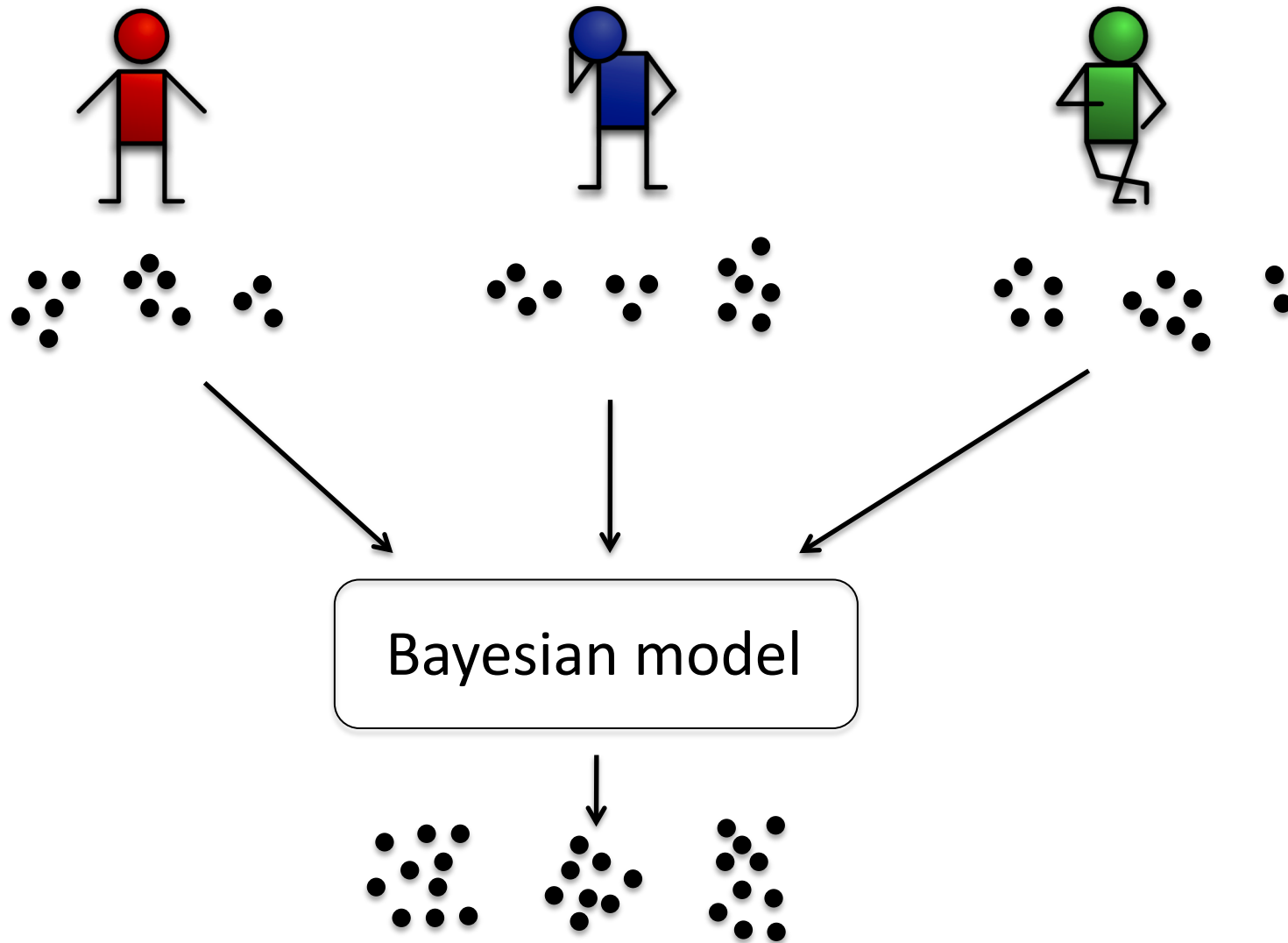


Democrats

Republicans

[Gomes et al., 2011]

Crowd Clustering



[Gomes et al., 2011]

The Potential of Crowdsourcing

1. Direct Applications to Machine Learning
2. Hybrid Intelligence Systems
3. Large Scale Studies of Human Behavior

Hybrid Intelligence for Speech Recognition

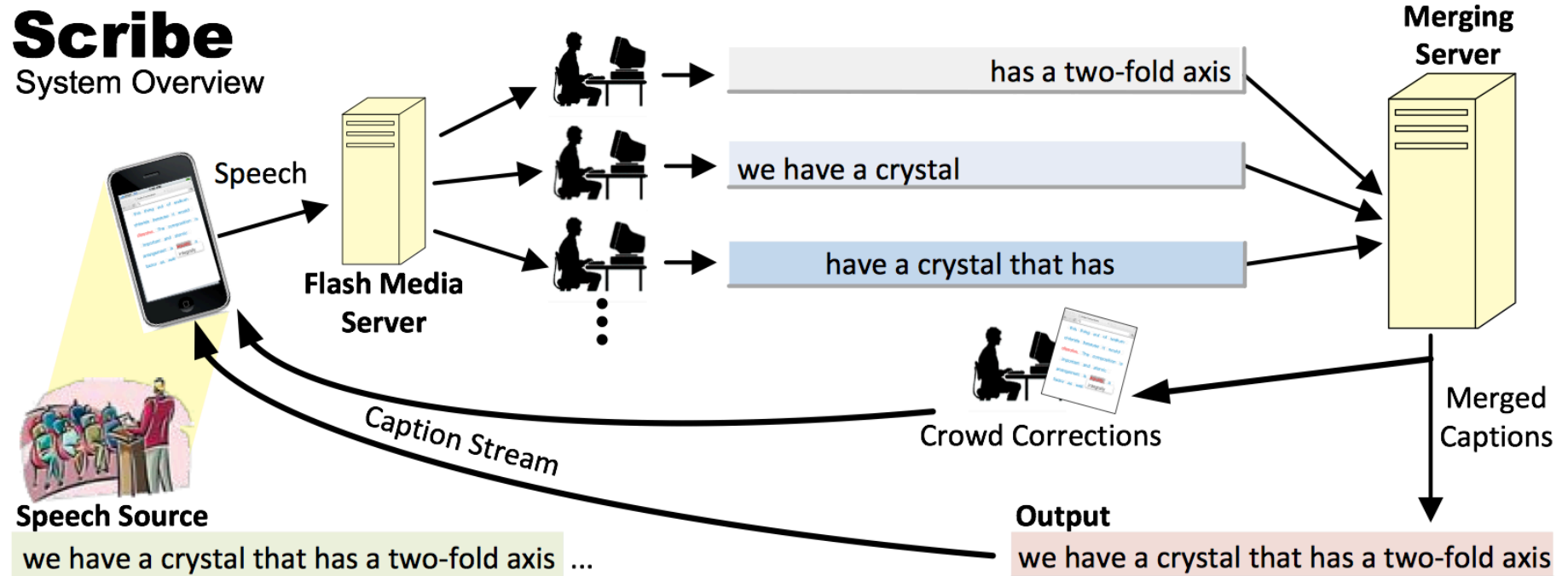
Crowd-Based Closed Captioning



Is it possible to provide real-time closed captioning of lectures, meetings, or other day-to-day conversations?

[Lasecki et al., 2012]

Crowd-Based Closed Captioning



The system merges **real-time partial inputs** from **dynamic, untrained crowds** to outperform individuals

[Lasecki et al., 2012]

Hybrid Intelligence for Constrained Optimization

Cobi: Communitysourced Scheduling



A big constrained optimization problem with no access to the constraints!

[projectcobi.com]

1. Committeesourcing

pn1171 (Paper)
Investigating the Long-Term Use of Exergames in the Home with Elderly Fallers
 Stephen Uzor, Glasgow Caledonian University
 Lynne Baillie, Glasgow Caledonian University

Abstract: Rehabilitation has been shown to significantly reduce the risk of fall... [\(more\)](#)

In Categories

- Older Adults (0)
- Motivation (1) +3
- Exergames (2) +1
- health and behavior change (1) +0
- Health Care (4) +1
- Home (2) +0
- User Studies (0)
- Rehabilitation (2) +1
- SC_Applications-V (28) +0

add a category +

2. Authorsourcing

Your Paper: **A Pilot Study of Using Crowds in the Classroom**

1. Tell us your name: (as it appears in the paper)

2. We've identified 10 papers that may be similar to yours. Tell us how they would fit in a session with your paper:

- Crowdfunding inside the Enterprise: Employee-Initiatives for Innovation and Collaboration** [\[abstract\]](#)
- Great in same session
 - Okay in same session
 - Not sure if it should be in same session
 - Should not be in same session

3. Scheduling

Touch -5 -5 -1	Social Impact Award	Shopping and Tagging -4 -4 -1	Place meets Engagement -4 -4 -1	Authenticator -4 -4 -1	Automated Usability / Evaluation -4 -4 -1	Reflection and Evaluation
Haptics -4 -4 -1	Collaborative Technology: I share, you	Pointing and Fitts Law -4 -4 -1	Studies of the Use of Digital -4 -4 -1	unused session 1	Evaluation Methods 2 -4 -4 -1	Blindness and Design -4 -4 -1
Fabrication -2 -1 -1 -2	Search and Find +2 +2 -1	Mobile keyboard / text entry +2 +2 -1	Hedonism, narrative, materiality & +2 +2 -1	Consent and Integrity +2 +2 -1	Novel Programming	Dising in a Psychiatric Setting +2 +2 -1
Touch, Tangibles, Touch -4 -4 -1	Mobiles and more -4 -4 -1	Mobile 1: Mobile Phones -6 -3 -3 -1	Case Studies in the wild	Privacy -7 -4 -3 -1	Nature and Nurture	ICT4D -4 -4 -1

4. Attendeesourcing

Monday, 11:00–12:20

Managing Social Media **SCJ**
 recommended ux management
 Paper Room: Blue

Enhancing Access **STJ**
 recommended HCI4D health ux design
 Paper Room: 242A

[projectcobi.com]

Authorsourcing

Your Paper: **A Pilot Study of Using Crowds in the Classroom**

1. **Tell us your name:** (as it appears in the paper)

crowdsourced clustering!

2. **We've identified 10 papers that may be similar to yours.**
Tell us how they would fit in a session with your paper:

Crowdfunding inside the Enterprise: Employee-Initiatives for Innovation and Collaboration

[\[abstract\]](#)

- Great in same session
- Okay in same session
- Not sure if it should be in same session
- Should not be in same session

87% response rate!

[projectcobi.com]

Scheduling

Cobi Charles Carmichael

Select a session for scheduling options and more information.

Conflicts 121

High severity (83)

- papers of mutual interests in opposing sessions (37)
- authors with papers in opposing sessions (1)
- chairs with papers in opposing sessions (5)
- chairs with papers in their own sessions (19)

Medium severity (58)

- papers that don't fit well in the same session (48)
- topics of interest to a persons in opposing sessions (2)
- chairs who don't fit well in their session (1)
- chairs and their papers of interest in opposing sessions (7)

Preferences 343

View Options

- Conflict
- Preference
- Session Chair Conflict
- Session Chair Names
- Session Type
- Number of Papers
- Duration
- Best Paper
- Honorable Mention

Session Types

Personas

Communities

History 0

Unscheduled Sessions 8

Unscheduled Papers 16

Unscheduled Chairs 65

Room/Time	Blue	Bordeaux	252B	352AB	Havana	241	342A	251	351	242A	242B	243	253	343	252A	361	362/363	221/221M
Mon 11:00-12:20	Navigating Data	Text Visualization	Call All Game Changers: BYOD (Bring)	MultiTouch and Gestures	Lifetime Research Award	Power to the People: Utilizing Crowdsourcing	Design and Design Lessons	Learning	Teaching Experiences: Tangible	Contact, Creation, and Health		User Interface Design and Adaptation for	Six Steps to Successful UX in an Agile	Rapid Design Labs—A Tool to Turbocharge	Body, Why? & Videotape: Applying	Designing Interactive Secure Systems	Human Computer Interaction for	Birds of a feather - session 1
Mon 14:00-15:20	Language	Game	Will Massive Online Open Courses		Enterprise and online communities	Holkeys / Touch keyboards	Brain Interfaces	Design for the Classroom	Co-Design: involving perspective	Technologies for Life		Practical Statistics for User Experience	Agile User Experience and UCD 1/2	Rapid Design Labs—A Tool to Turbocharge	Speech-based Interaction: Myths		The Role of Engineering Work in CHI	Birds of a feather - session 2
Mon 16:00-17:20	Management of Knowledge and Collaboration	Video	Theory and Practice in UX Research	Table and Floors	Smart Tools for Smart Work Environments	Large and public Displays	Case Studies in Innovating UCD Process	unusual session 8	Mobile 2: Very Moving: reflection in	Novelty Games		Practical Statistics for User Experience	Agile User Experience and UCD 2/2	Rapid Design Labs—A Tool to Turbocharge	Speech-based Interaction: Myths	unusual session 2	Enhancing the Research Infrastructure	Birds of a feather - session 3
Tue 9:00-10:20	Classrooms	Social Face: creativity unleashed	CHI at the Barcodes - an Activist	Interaction around Devices	Lifetime Practice Award	Gestures studies / empirical	Communities of practice	Embodied Interaction (and Thinking)	Evaluation Methods 1	Technologies for Life 2		User Experience Evaluation Methods	Choice and Decision Making for HCI	Cognitive Crash Durmies: Predicting	Analyzing Social Media Data 1/2	SIG: NVI (Non Visual Interaction)	Managing UX Teams	Birds of a feather - session 4
Tue 11:00-12:20	Crowds and activism	Visualization 1	Gamification @ Work	Mobile Gestures and Grasp	Invited talk - Don Norman	Ceasing and Authoring	Design Ideation Methods	Online Classrooms	Ethics	Impairment and Rehabilitation		User Experience Evaluation Methods	Choice and Decision Making for HCI	Cognitive Crash Durmies: Predicting	Analyzing Social Media Data 2/2	Research-Practice Interaction	Digital Art: Challenging Perspectives	Birds of a feather - session 5
Tue 14:00-15:20	cross-over work	Bodies Matter	UX Management: Current and	Multi-device Interaction	Design and Time: Long-term User	3D Use	Case Studies in Novel Settings	Game Design	HCI Ethics	Health, Information, and		Practical Statistics for User Experience	Expert Reviews - For Experts	Make This! Introduction to Electronics	Test Submission 1/2	Consumer Engagement in Health	Changing Perspectives on Sustainability	Birds of a feather - session 6
Tue 16:00-17:20	Energy / Sustainability	Interaction Design for Social	Is My Doctor Listening to Me? Impact of	Bendable, Flexible	Design Research, Paradigm and	Displays in public space	Case Study of Changing the Way We Work	Exergames, Inclusion	Food	The Clinical Setting		Practical Statistics for User Experience	Expert Reviews - For Experts	Make This! Introduction to Electronics	Test Submission 2/2	HCI with Sports	SIG NIME: Music, Technology,	Birds of a feather - session 7
Wed 9:00-10:20	Autism	Crowdsourcing Activism: Volunteering	Exploring the Representation of Woman	Touch	Social Impact Award	Shopping and Tagging	Place meets Engagement	Authentication	Automated Usability / Evaluation	Reflection and Evaluation		So-Fi and CHI in the Movies and Television	Interactive Walking in Virtual	Designing with and for Children in the 21st	Student Design Competition	unusual session 4	CHI 2013 Human Work Interaction	Birds of a feather - session 8
Wed 11:00-12:20	Crime, Conflicts, and Revolution	How We Fail About Websites	Leveraging the Progress of Women in the	Haptics	Collaborative Technology: I share, you	Pointing and Fits Law	Studies of the Use of Digital Artifacts	unusual session 1	Evaluation Methods 2	Blindness and Design		So-Fi and CHI in the Movies and Television	Interactive Walking in Virtual	Designing with and for Children in the 21st	Student Research Competition		On Top of the User Experience Wave - How is	Birds of a feather - session 9

The system solves an optimization problem to propose a schedule, but chairs retain control.

Hybrid Intelligence for Writing

The Selfsourcing Process

1. Collect content
2. Organize content
3. Turn content into writing

Collect Content

The MicroWriter breaks writing into microtasks.

Microtasks can be shared with collaborators.

Microtasks can be done while mobile.

Collaborative writing typically requires coordination.

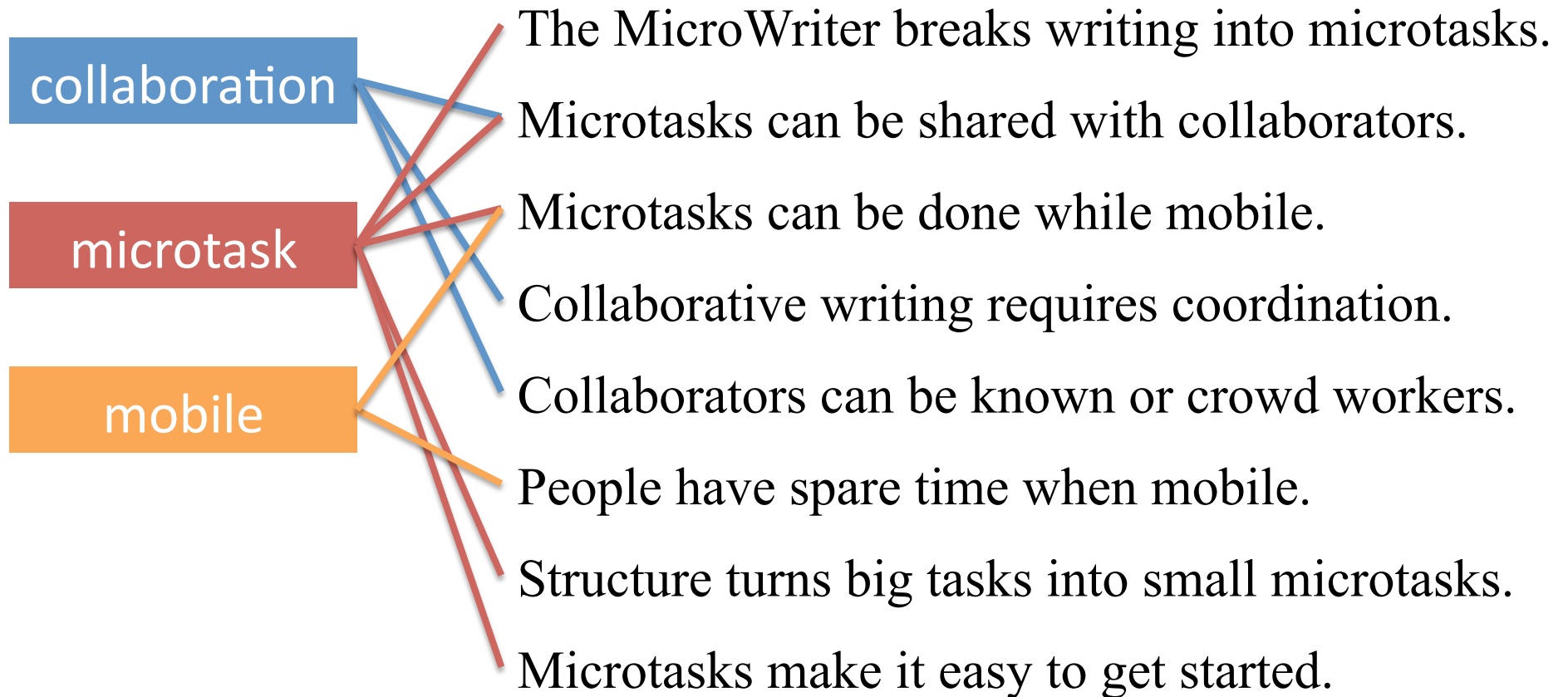
Collaborators can be known or crowd workers.

People have spare time when mobile.

Structure turns big tasks into small microtasks.

Microtasks make it easy to get started.

Organize Content



Turn Content into Writing

collaboration



Microtasks can be shared with collaborators.

Collaborative writing requires coordination.

Collaborators can be known or crowd workers.

Collaborative writing typically requires coordination, but microtasks are easy to share with collaborators without the need for coordination. The collaborators can be known colleagues or paid crowd workers.

Turn Content into Writing

Structure makes it possible to turn big tasks into a series of smaller microtasks. For example, the MicroWriter breaks writing into microtasks. These microtasks make the larger task easier to start.

Collaborative writing typically requires coordination, but microtasks are easy to share with collaborators without the need for coordination. The collaborators can be known colleagues or paid crowd workers.

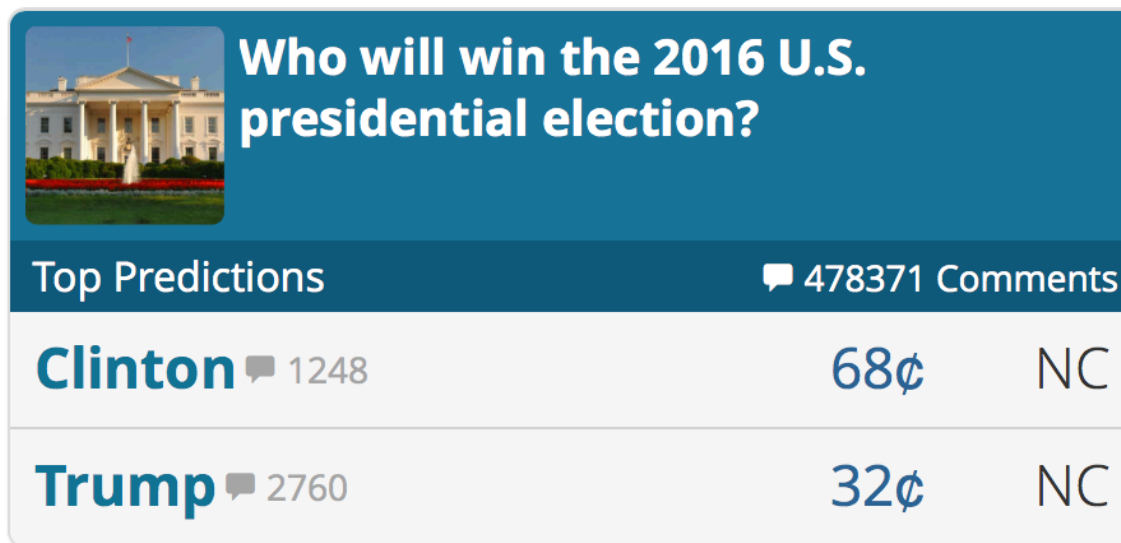
People have spare time when mobile, and these micromoments are ideal for doing microtasks.

~~The Selfsourcing Process~~ Crowdsourcing

1. Collect content
 2. Organize content
 3. Turn content into writing
- Steps 2 & 3 could be done by crowdworkers, traditional ML/AI approaches, or a combination
 - Author takes final pass, no need for perfection

Hybrid Intelligence for Information Aggregation

Combinatorial Prediction Markets



The screenshot shows a prediction market interface for the 2016 U.S. presidential election. The title is "Who will win the 2016 U.S. presidential election?" with a small image of the White House. Below the title, it says "Top Predictions" and "478371 Comments". The table lists two candidates: Clinton and Trump. Clinton has 1248 comments, a price of 68¢, and a status of NC. Trump has 2760 comments, a price of 32¢, and a status of NC.

Top Predictions		478371 Comments	
Clinton	1248	68¢	NC
Trump	2760	32¢	NC

source:
PredictIt.org

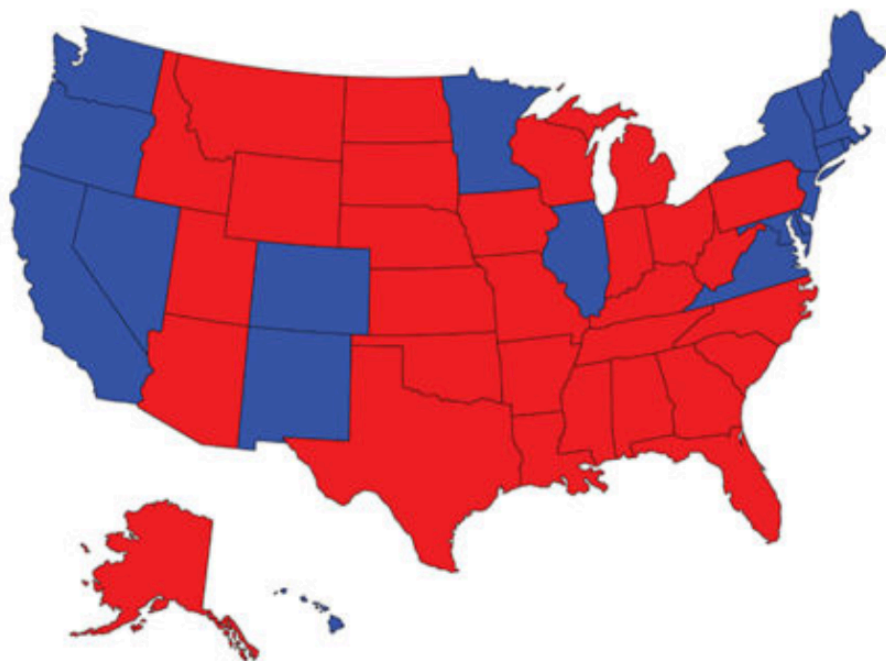
Payoff would have been \$1 if Clinton won. If probability of Clinton winning was x , I should have

- Bought at any price less than $\$x$
- Sold at any price greater than $\$x$

Market price captures crowd's collective belief

[Abernethy, Chen, Vaughan, 2013]

Combinatorial Prediction Markets



*Chance of Democrat
winning North
Carolina?*

*Chance of Republican
winning Ohio or
Pennsylvania?*

Challenges: liquidity, computational issues, ...

Can combine optimization techniques with human input to generate **coherent prices** (and therefore **coherent predictions**) over large outcome spaces

Hybrid Intelligence in Industry

The Potential of Crowdsourcing

1. Direct Applications to Machine Learning
2. Hybrid Intelligence Systems
3. Large Scale Studies of Human Behavior

User Studies for Security Research

How well do Internet users understand security risks?

p@ssw0rd vs. pAsswOrd

Who tries to guess passwords?

Only 14% mentioned both strangers *and* familiar people as threats

User Studies to Improve the Communication of Numbers

Q: How many times larger is a trillion than a million? Would you say...

- One Thousand Times- **18%**
- Ten Thousand Times- **12%**
- One Hundred Thousand Times- **21%**
- One Million Times- **21%**
- Ten Million Times- **17%**
- Don't Know- **12%**

This report presents the findings of a telephone survey conducted among a national probability sample of 1,001 adults comprising 501 men and 500 women 18 years of age and older, living in private households in the continental United States.

Interviewing for this CARAVAN® Survey was completed during the period April 23-26, 2009.

[Barrio et al., 2016]

Perspectives

- Is a **one hundred billion dollar** cut to the US federal budget big or small?
- One hundred billion dollars is about...
 - 3% of the 2015 US federal budget
 - 1/6 of annual US spending on military
 - 30% of the net worth of Beyoncé
 - \$5 for every person in New York state

Step 1: Perspective Generation



Six months of New York Times front page articles

64 quotes with measurements

370 crowd-generated perspectives
with incentives for quality

Workers rated other workers'
perspectives for helpfulness

Chose the highest-rated perspectives

[Barrio et al., 2016]

Perspective Examples

- The Ohio National Guard brought **33,000 gallons** of drinking water to the region.
- To put this into perspective, 33,000 gallons of water is about equal to the amount of water it takes to fill 2 average swimming pools.

Perspective Examples

- They also recommended safety programs for the nation's gun owners; Americans own almost **300 million firearms**.
- To put this into perspective, 300 million firearms is about 1 firearm for every person in the United States.

Step 2: Perspective Experiments

- Randomized experiments run on 3200+ subjects on AMT to test three proxies of comprehension
 - Recall
 - Estimation
 - Error detection
- Support found for the benefits of perspectives across all experiments
 - Example: 55% remembered number of firearms in US with perspective, only 40% without

User Studies for Online Advertising

The Cost of Annoying Ads



The Fast, Simple, Safe Browser

VS.



Advertisers pay publishers to display ads, but annoying ads cost publishers page views.

How much do annoying ads cost publishers in dollars?

[Goldstein et al., 2013]

The Cost of Annoying Ads



The Fast, Simple, Safe Browser

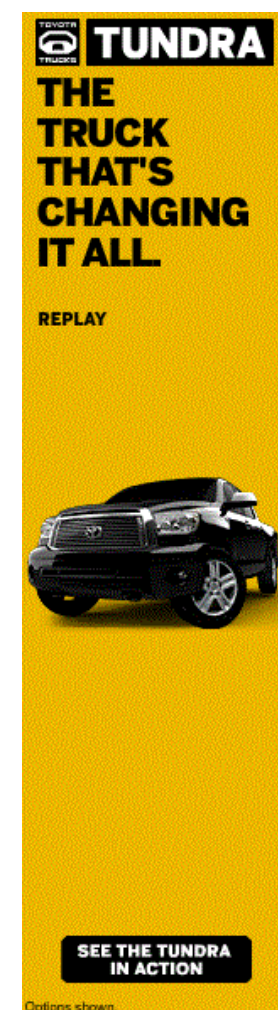
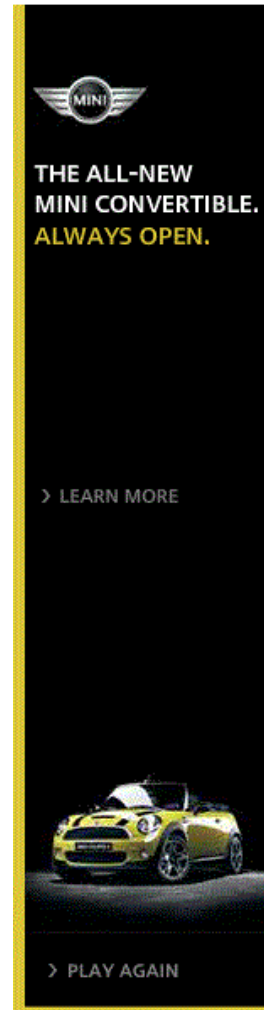
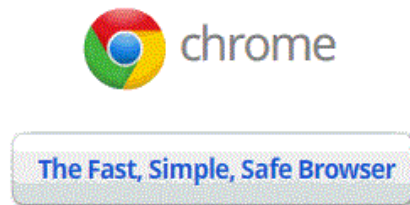
VS.

Step 1: Use the crowd to identify annoying ads.



[Goldstein et al., 2013]

Good Ads



Bad Ads

I Need a Degree In...
Click Your Career

- Business Education
- Nursing
- Health Care
- Criminal Justice
- Other Programs

classesUSA™

Think You're Too Busy to Go Back to School?
Graduate Online in as Fast as 13 Months!

Select Your State
Alabama

See Degrees Now

classesUSA™

ALERT FREE* SCREENSAVERS!

Screensavers

- Angelina Jolie
- Britney Spears
- Paris Hilton
- Jessica Alba
- Jessica Simons

Preview Settings

Click the "OK" button now to get your FREE* Screensavers!

OK

*See Details

Would You Go Back To School If You Qualified For A Grant? See If You Qualify!

Click Your Age:

- Under 18
- 19-25
- 26-35
- 36-45
- 46-55
- 56-65
- 66-75
- Over 75

classesUSA™

House Payments Fall Again!
Think You Pay Too Much for Your Mortgage? Find Out!
Click Your State

Estimate New Payment

LowerMyBills.com

What is your Credit Score?

- Excellent
750 - 840
- Good
660 - 749
- Fair
620 - 659
- Poor
340 - 619
- I Don't Know
????

Find out INSTANTLY!

FreeScore.com

[Goldstein et al., 2013]

Step 2: Estimate the Cost

- Workers asked to label email as spam or not
- Shown good, bad, or no ads; paid varying amounts per email
- *How much more must a worker be paid to do the same tasks when shown bad ads?*

What is your Credit Score?
Excellent 750 - 840
Good 660 - 749
Fair 620 - 659
Poor 340 - 619
I Don't Know ????
Find out INSTANTLY!
FreeScore.com

Hi!

We have a new product that we offer to you, C_I_A_L_I_S soft tabs,

Cialis Soft Tabs is the new impotence treatment drug that everyone is talking about. Soft Tabs acts up to 36 hours, compare this to only two or three hours of Viagra action! The active ingredient is Tadalafil, same as in brand Cialis.

Simply dissolve half a pill under your tongue, 10 min before sex, for the best erections you've ever had!

Soft Tabs also have less sidebacks (you can drive or mix alcohol drinks with them).

You can get it at: <http://onlinegenericrx.com/soft/>

No thanks: <http://onlinegenericrx.com/tr.php>

Mortgage Rates Hit Record Lows!
Rate 3% low!
Click Your State
Bad Credit OK
LowerMyBills.com

AL
AK
AZ
AR
CA
CO
CT
DE
DC
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MA
MI
MN
MS
MO
MT
NE
NV
NH
NJ
NM
NY
NC
ND
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY

Step 2: Estimate the Cost

- Good ads lead to about the same number of views (emails classified) as no ads
- Costs **more than \$1 extra** to generate 1000 views of bad ads instead of no ads or good ads
- Takeaway: Publishers **lose money** by showing bad ads unless they are paid significantly more to show them

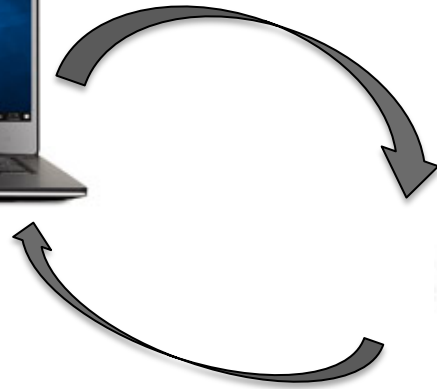
Summary of Part 1

1. Direct Applications to Machine Learning
2. Hybrid Intelligence Systems
3. Large Scale Studies of Human Behavior

Part 2:

The Crowd is Made of People

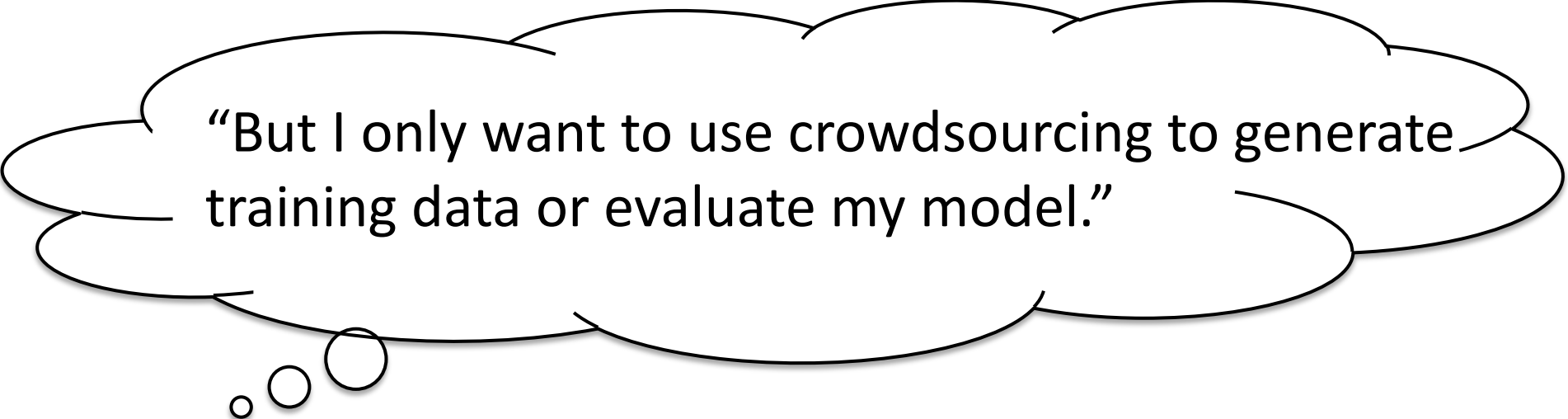
Traditional computer science tools let us reason about programs run on machines (runtime, scalability, correctness, ...)



What happens when there are humans in the loop?

Need a **model of human behavior**. (Are they accurate? Honest? Do they respond rationally to incentives?)

Wrong assumptions lead to suboptimal systems!



“But I only want to use crowdsourcing to generate training data or evaluate my model.”

Understanding the crowd can teach you

- How much to pay for your tasks and what payment structure to use
- How much you really need to worry about spam
- How and why to communicate with workers
- Whether your labels/evaluations are independent
- How to avoid common pitfalls

The Crowd is Made of People

- Crowdworker demographics
- Honesty of crowdworkers
- Monetary incentives
- Intrinsic motivation
- The network within the crowd

Best practices! Tips and tricks!

Amazon Mechanical Turk

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



Workers

Get Results from Mechanical Turk Workers

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

As a Mechanical Turk Requester you:

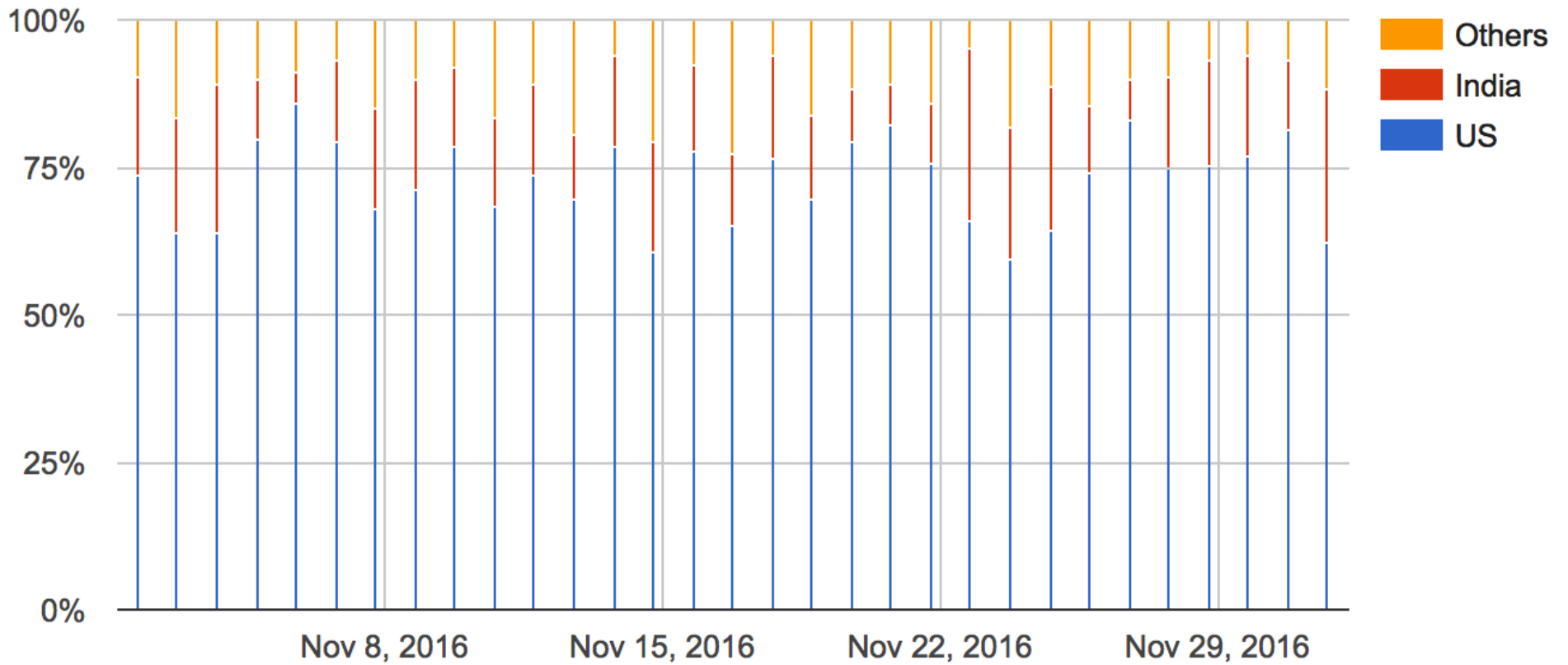
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Requesters

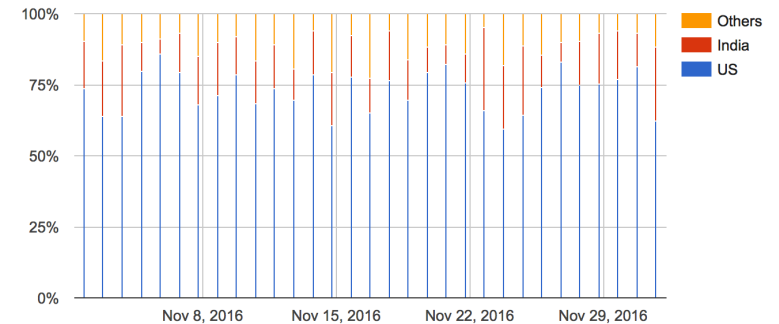
Crowdworker Demographics

Basic Demographics



Basic Demographics

- 70-80% US, 10-20% India
- Roughly equal gender split
- Median (reported) household income:
 - \$40K-\$60K for US workers
 - Less than \$15K for Indian workers



Spammers Aren't Such a Big Problem

Experimental Paradigm

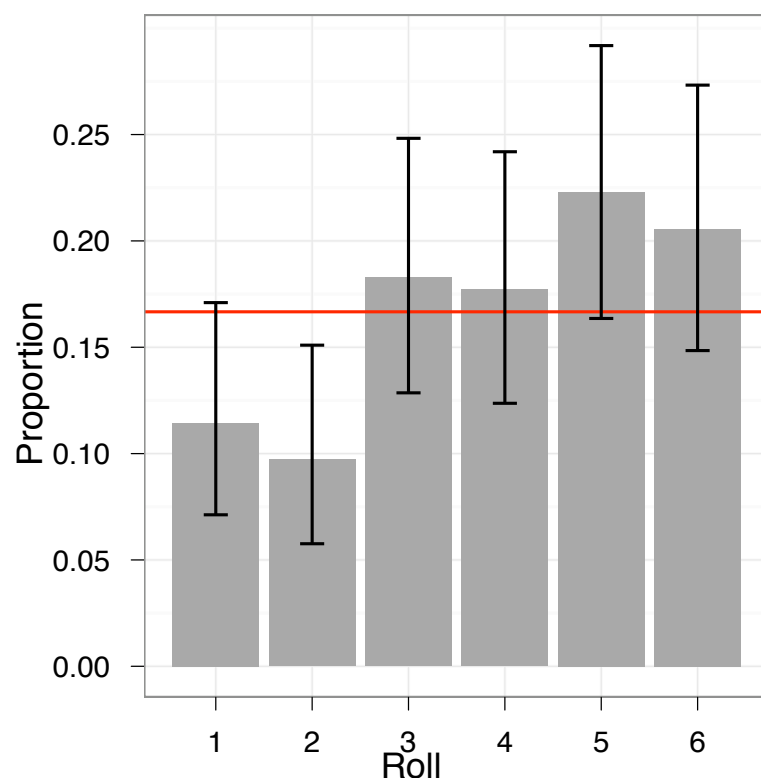
- Ask participants about demographics
 - Sex, Age, Location, Income, Education
- Ask participants to **privately** roll a die (or simulate it on an external website) and report the outcome

$$\text{payment} = \$0.25 + (\$0.25 * \text{roll})$$

- If workers honest, mean reported roll should be about 3.5... **What do you think the mean was?**

Baseline

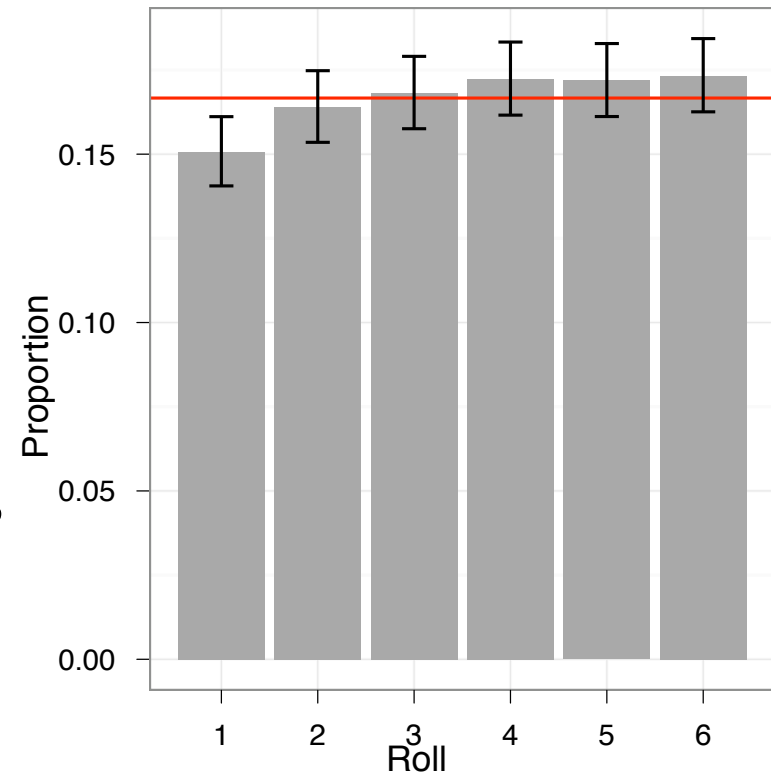
- Average reported roll higher than expectation
 - $M = 3.91, p < 0.0005$
- Players under-reported ones and twos and over-reported fives
- **But many workers were honest!**
- Similar to Fischbacher & Huesi lab study



[Suri et al., 2011]

Thirty rolls

- Overall, **much less dishonesty**
- Average reported roll much closer to expectation
 - $M = 3.57, p < 0.0005$
- Only 3 of 232 reported significantly unlikely outcomes
- Only 1 was fully income maximizing (all sixes)
- **Why is this the case?**



[Suri et al., 2011]

Takeaways & Related Best Practices

- Most workers are honest most of the time.
- But some are not. You should still use care to avoid attacks.

Monetary Incentives

How much should you pay?

A useful trick:

- Pilot your task on students, colleagues, or a few workers to see how long it generally takes.
- Use that to make sure your payments work out to at least the US minimum wage.

Benefits:

- It's the decent thing to do!
- It helps maintain good relationships with workers.

Can performance-based payments improve the quality of crowdwork?



1: Nearly every group of animals has its giants, its species which
2: their fellows as Goliath of Gath stood head and shoulders above
3: hosts; and while some of these are giants only in comparison
4: fellows, belonging to families whose members are short of
5: sufficiently great to be called giants under any circumstance
6: giants live to-day, some have but recently passed away, and
7: long ages before man trod this earth. The most gigantic of
8: whales—still survive, and the elephant of to-day suffers but little in
9: comparison with the mammoth of yesterday; the monstrous Dinosaurs, greatest of
10: all reptiles—greatest, in fact, of all animals that have walked the
11: earth—flourished thousands upon thousands of years ago. As for birds, some of
12: the giants among them are still living, some existed long geologic periods ago,
13: and a few have so recently vanished from the scene that their
14: lingers amid the haze of tradition. The best known among
15: most recent in point of time, are the Moas of New Zealand,
16: notice by the Rev. W. Colenso, later on Bishop of New Zealand,
17: missionaries to whom Science is under obligations. Early in
18: Colenso, while on a missionary visit to the East Cape region,
19: natives of Waiapu tales of a monstrous bird, called Moa, had
20: man, that inhabited the mountain-side some eighty miles
21: the last of his race, was said to be attended by two equally
22: kept guard while he slept, and on the approach of man
23: immediately rushed upon the intruders and trampled them to
24: Maoris had seen this bird, but they had seen and somewhat
25: making parts of their fishing tackle, bones of its extinct relatives, and these
26: bones they declared to be as large as those of an ox.
27:
28: About the same time another missionary, the Rev. Richard Taylor, found a bone
29: ascribed to the Moa, and met with a very similar tradition among the natives of
30: a near-by district, only, as the foot of the rainbow moves away as we move
31: toward it, in his case the bird was said to dwell in quite a different locality
32: from that given by the natives of East Cape. While, however, the Maoris were

Proofread this text, earn \$0.50

Earn an extra \$0.10 for every typo found

[Ho et al., 2015]

Prior Work on Crowd Payments



- Paying more increases the quantity of work, but not the quality [MW09, RK+11, BKG11, LRR14]
- PBPs improve quality [H11, YCS14]
- PBPs do not improve quality [SHC11]
- Bonus sizes don't matter [YCS13]

Performance-Based Payments



We explore **when**, **where**, and **why** performance-based payments improve the quality of crowdwork on Amazon Mechanical Turk.

Can PBPs work?

- Warm-up to verify that PBPs can lead to higher quality crowdwork on some task.
- Test whether there exists an **implicit PBP effect**: workers have **subjective beliefs** on the quality of work they must produce to receive the base payment, and so already behave as if payments are (implicitly) performance-based.

Can PBP's work?

- Task: Proofread an article and find spelling errors.

1: Nearly every group of animals has its giants, its species which tower above
2: their fellows as Goliath of Gath stood head and shoulders above the Philistine
3: hosts; and while some of these are giants only in comparison with their
4: fellows, belonging to families whose members are short of stature, others are
5: sufficiently great to be called giants under any circumstances. Some of these
6: giants live to-day, some have but recently passed away, and some ceased to be
7: long ages before man trod this earth. The most gigantic of mammals—the
8: whales—still survive, and the elephant of to-day suffers but little in
9: comparison with the mammoth of yesterday; the monstrous Dinosaurs, greatest of
10: all reptiles—greatest, in fact, of all animals that have walked the
11: earth—flourished thousands upon thousands of years ago. As for birds, some of
12: the giants among them are still living, some existed long geologic periods ago,
13: and a few have so recently vanished from the scene that their memory still
14: lingers amid the haze of tradition. The best known among these, as well as the
15: most recent in point of time, are the Moas of New Zealand, first brought to
16: notice by the Rev. W. Colenso, later on Bishop of New Zealand, one of the many
17: missionaries to whom Science is under obligations. Early in 1838, Bishop
18: Colenso, while on a missionary visit to the East Cape region, heard from the
19: natives of Waiapu tales of a monstrous bird, called Moa, having the head of a
20: man, that inhabited the mountain-side some eighty miles away. This mighty bird,
21: the last of his race, was said to be attended by two equally huge lizards that
22: kept guard while he slept, and on the approach of man wakened the Moa, who
23: immediately rushed upon the intruders and trampled them to death. None of the
24: Maoris had seen this bird, but they had seen and somewhat irreverently used for
25: making parts of their fishing tackle, bones of its extinct relatives, and these
26: bones they declared to be as large as those of an ox.
27:
28: About the same time another missionary, the Rev. Richard Taylor, found a bone
29: ascribed to the Moa, and met with a very similar tradition among the natives of
30: a near-by district, only, as the foot of the rainbow moves away as we move
31: toward it, in his case the bird was said to dwell in quite a different locality
32: from that given by the natives of East Cape. While, however, the Maoris were

- We randomly insert 20 typos
 - sufficiently -> sufficently
 - existence -> existance
 - ...
- Useful properties:
 - Quality is measurable
 - Exerting more effort -> better results

Can PBPs work?

Base payment: \$0.50; Bonus payment: \$1.00

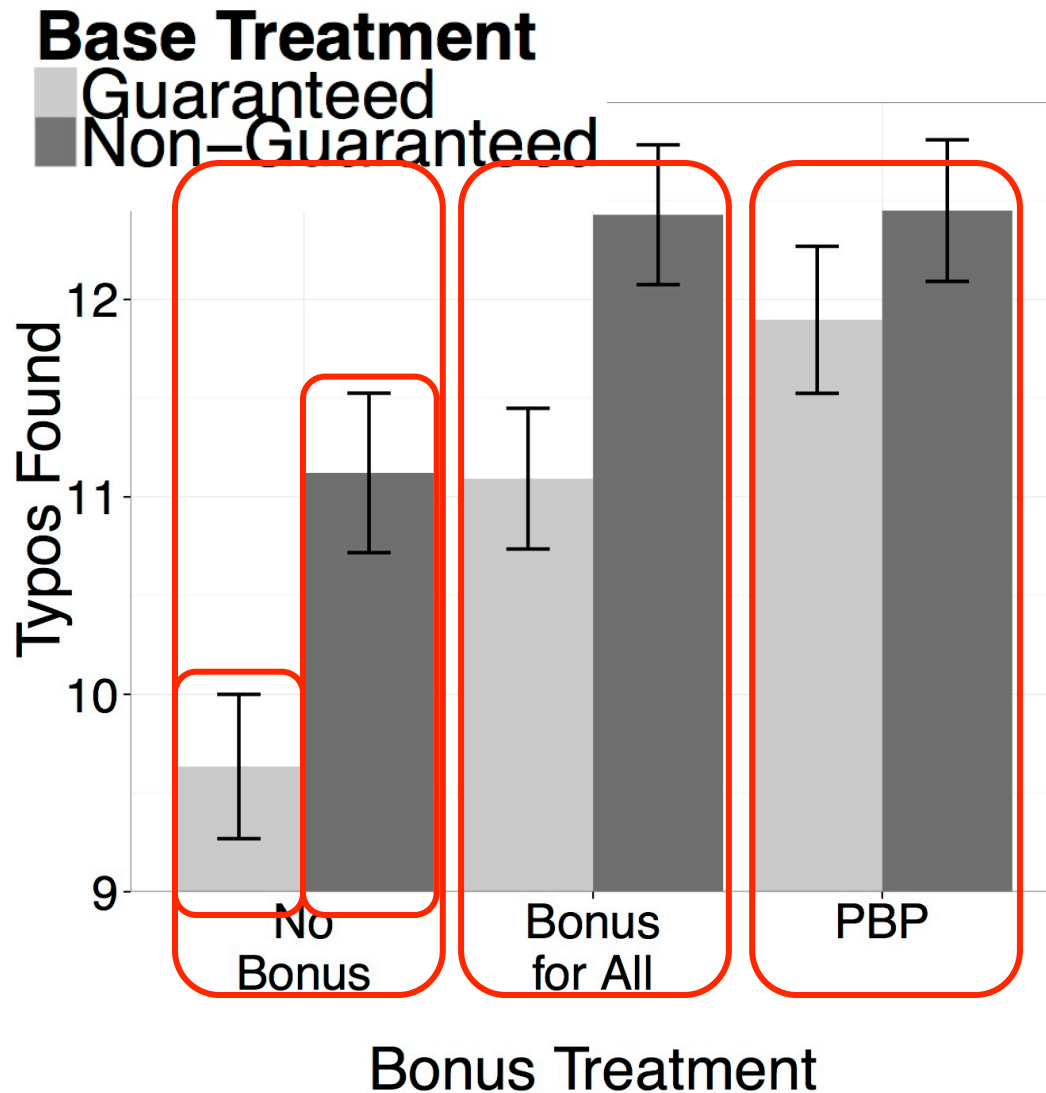
Three Bonus Treatments:

- *No Bonus:* no bonus or mention of a bonus
- *Bonus for All:* get the bonus unconditionally
- *PBP:* get the bonus if you find 75% of the typos found by others

Two Base Treatments:

- *Guaranteed:* guaranteed to get paid
- *Non-Guaranteed:* no mention of a guarantee

Can PBPs work?

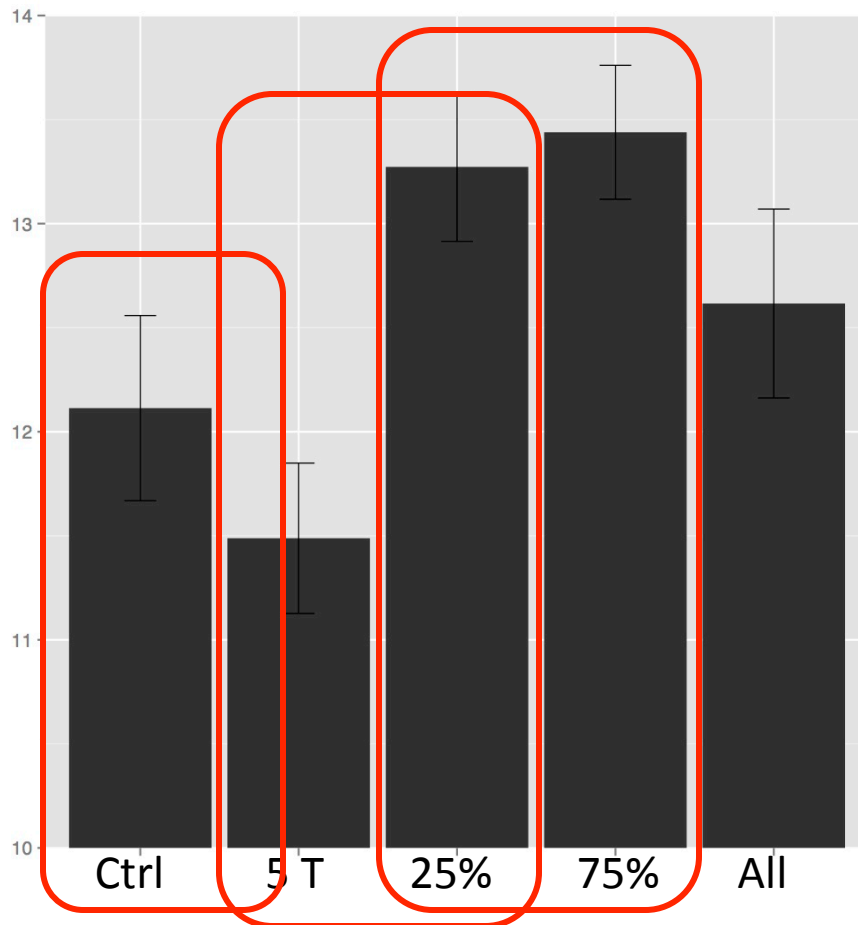


- Results from 1000 unique workers
- Guaranteed payments hurt (**implicit PBP**)
- PBPs improve quality
- Unlike in prior work, paying more also improves quality

Under what conditions do PBPs work?

Bonus threshold (585 unique workers)

- \$0.50 base + \$1.00 bonus for finding X typos

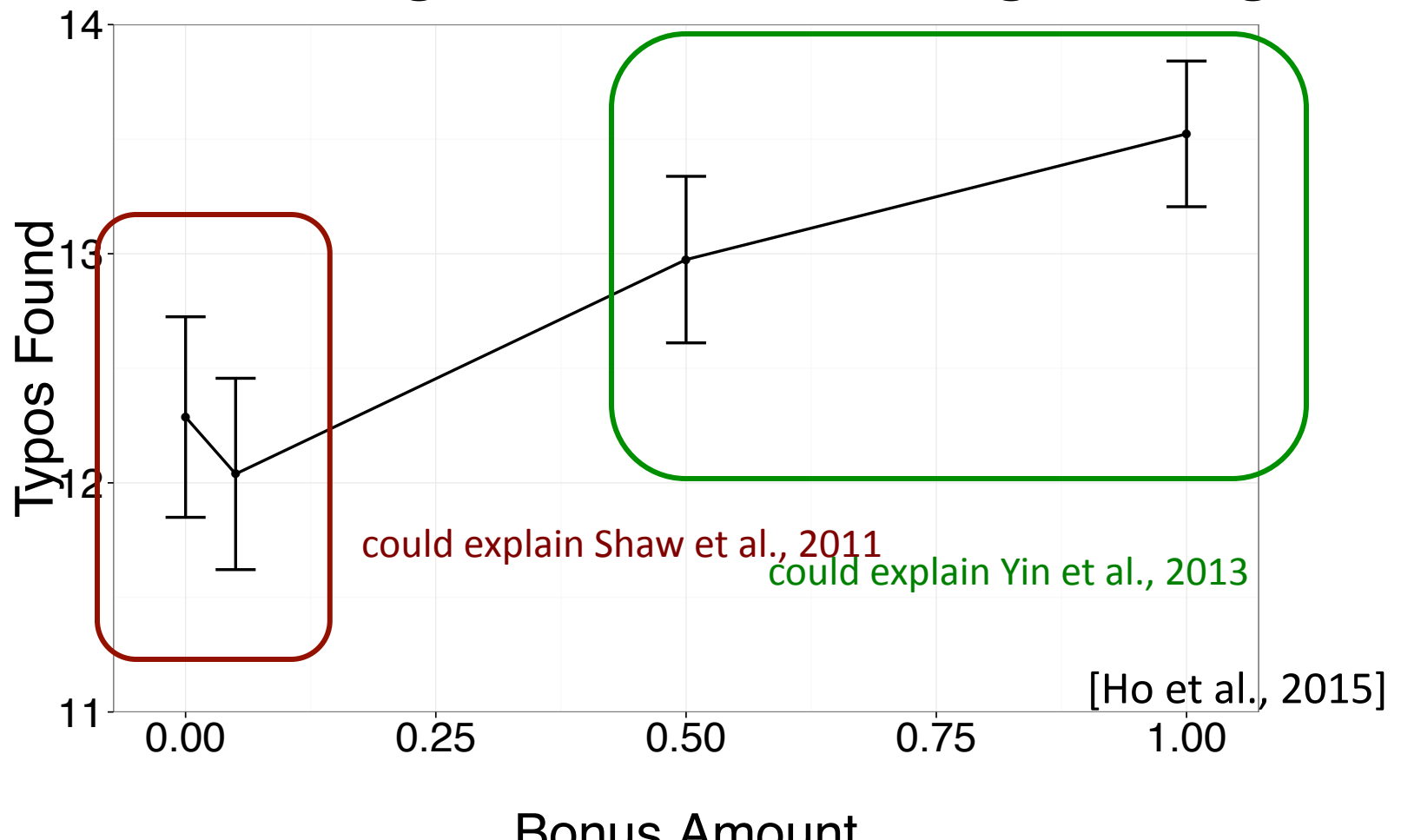


- PBPs work for a wide range of thresholds
- Subjective beliefs (5 typos vs. 25% of typos) can improve quality

Under what conditions do PBPs work?

Bonus amounts (451 unique workers)

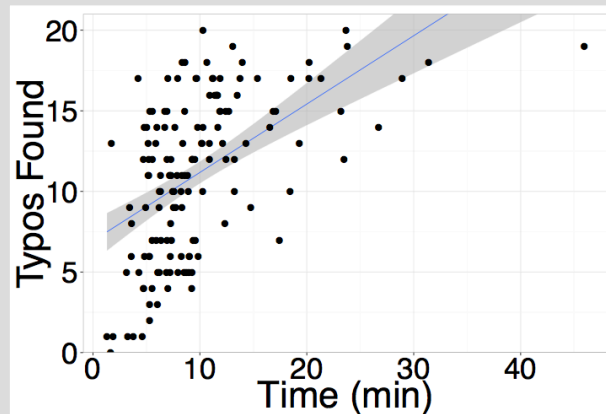
- \$0.50 base + \$X bonus for finding 75% of typos
- PBPs work as long as the bonus is large enough



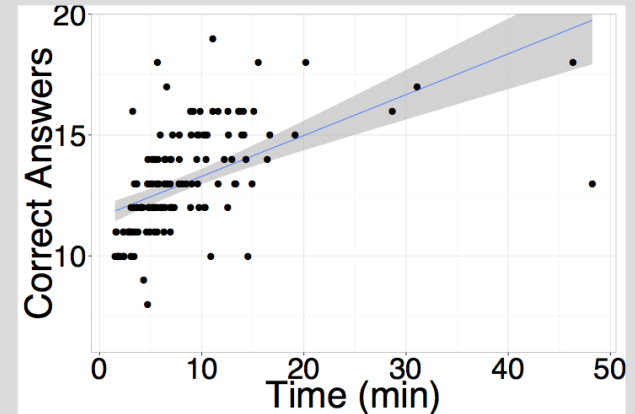
Which tasks do PBPs work on?

- What properties of a task lead to quality improvements from performance-based pay?
- Some pilot experiments on audio transcription suggested that
 - PBPs improve quality for **effort-responsive** tasks
 - It is not always straight-forward to guess which tasks are effort-responsive

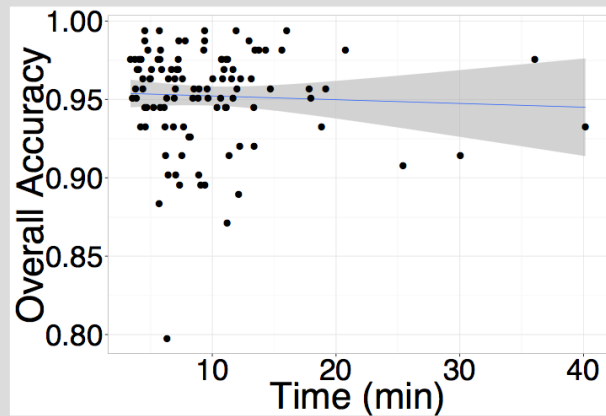
Which tasks do PBPs work on?



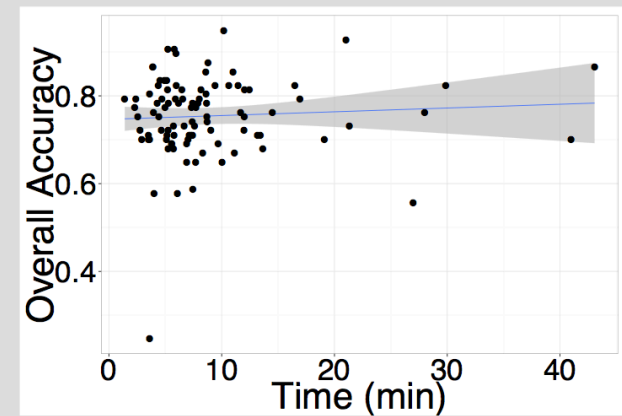
proofreading



Spotting differences



Handwriting recognition



Audio transcription

Takeaways & Related Best Practices

- Aim to pay at least US minimum wage. Pilot your task to find out how long it takes.
- Performance-based payments can improve quality for effort-responsive tasks. Pilot to check the relationship between time and quality.
- Bonus payments should be large relative to the base. The precise amount and precise criteria for receiving the bonus don't matter too much.

Intrinsic Motivation

Work That Matters

- Three treatments:
 - **control**: no context given
 - **meaningful**: told they were labeling tumor cells to assist medical researchers
 - **shredded**: no context, told work would be discarded
- Meaningful -> **quantity** up, but **quality** similar
- Shredded -> **quality** down, but **quantity** similar



Takeaways & Related Best Practices

- Workers produce more work when they know they are performing a meaningful task.
- But the quality of their work might not improve.
- Gamification and explicitly stoking workers' curiosity can also increase productivity.

The Communication Network Within the Crowd

Assumption: Crowdworkers are independent



[Yin et al., 2016]

In reality workers talk and collaborate

Ethnographic field studies show that crowdworkers...



Help each other with
administrative
overhead



Share tasks and
reputable
employers



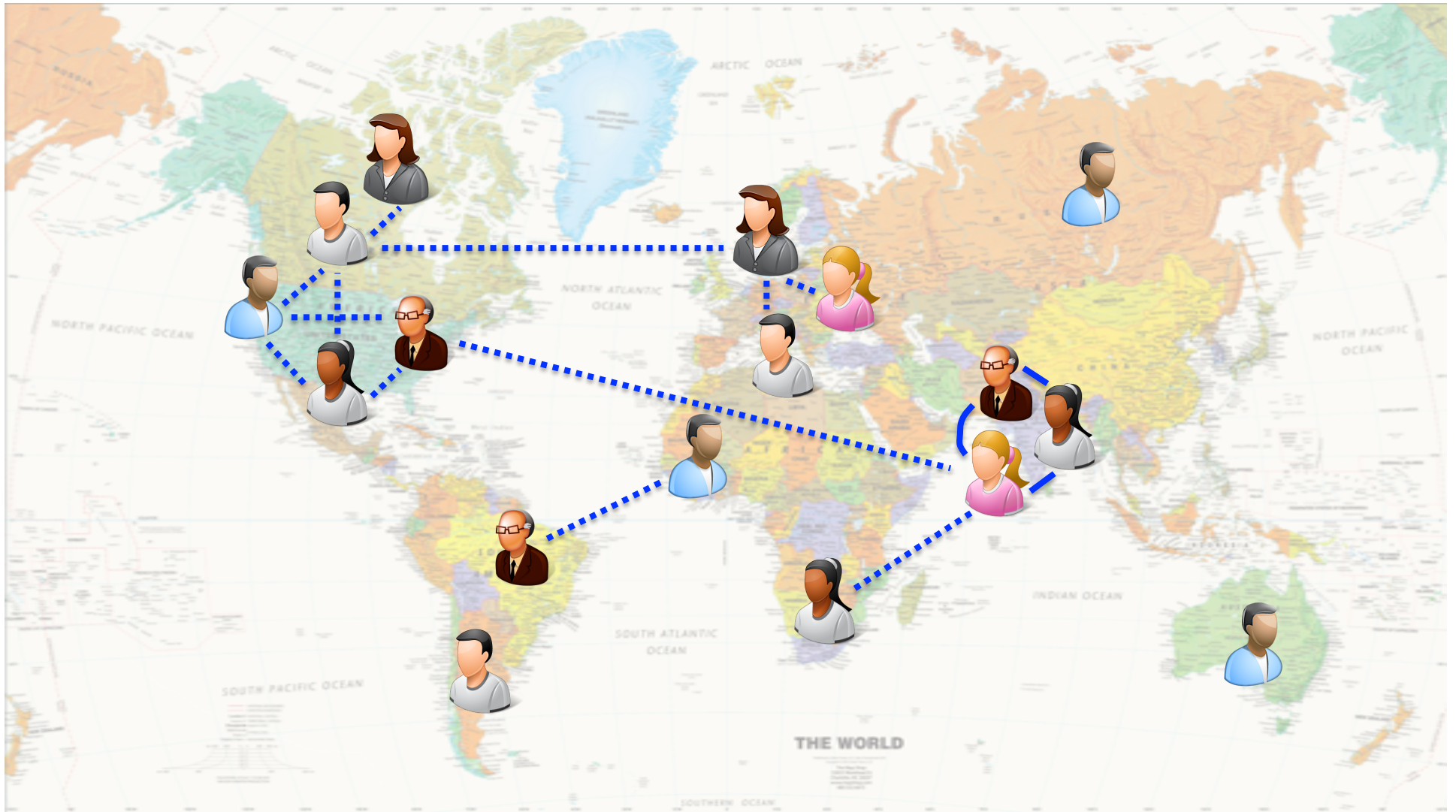
Recreate social
connections and
support

M.L. Gray, S. Suri, S.S. Ali and D. Kulkarni. The Crowd is a Collaborative Network. *CSCW* 2016

N. Gupta, D. Martin, B.V. Hanrahan and J. O'Neil. Turk-life in India. *Group* 2014

[Yin et al., 2016]

A Communication Network



What is the scale?

What is the structure?

How is it used?

[Yin et al., 2016]

Our goal: Open the black box of
crowdsourcing to **map the
communication network** of
crowdworkers

Why is it challenging?

The network is not accessible from the API so we can't simply download, crawl, or scrape it!

Want to map the network in a way that

#1 Elicits only “true” edges

#2 Elicits as many true edges as possible

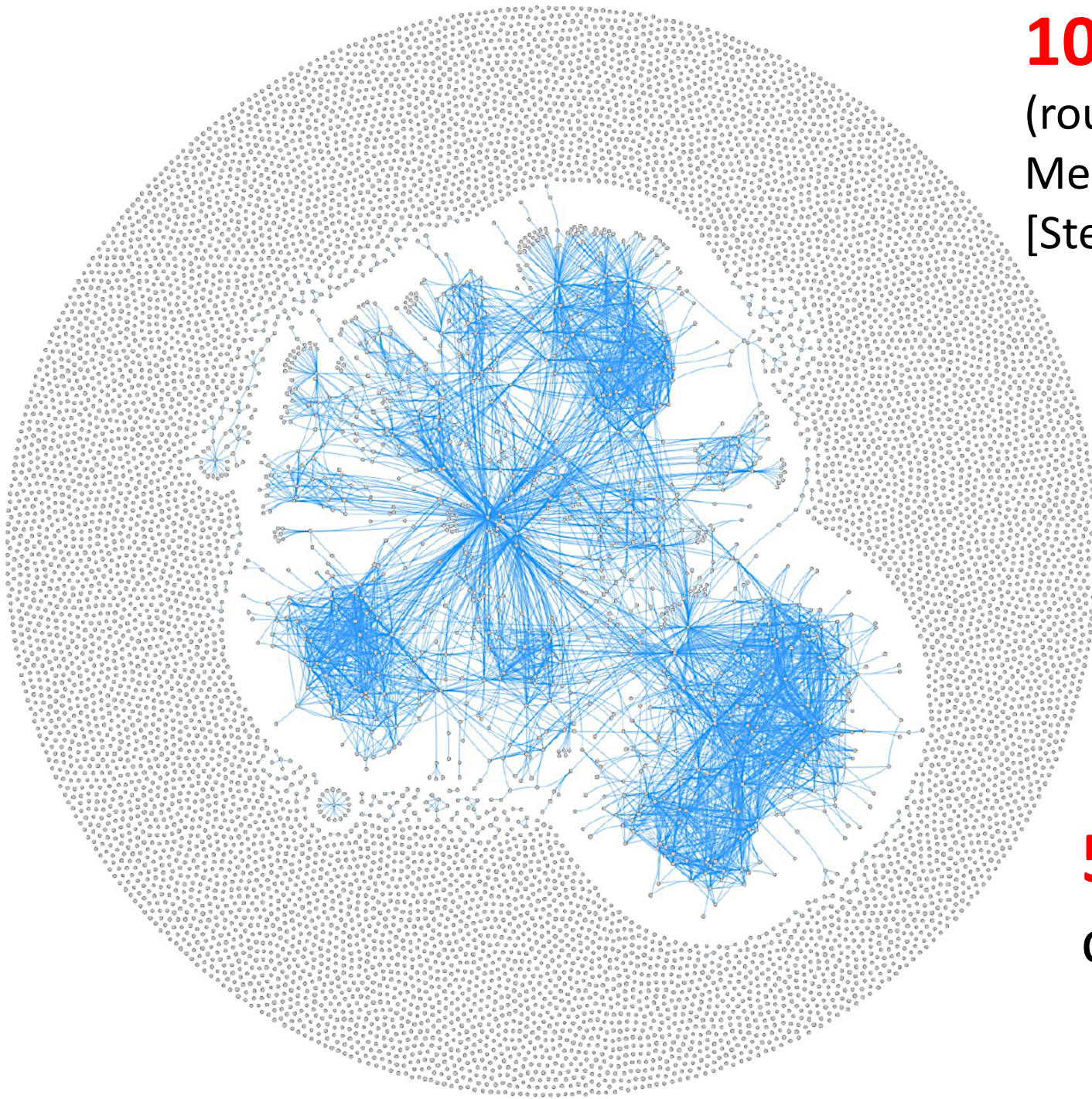
#3 Preserves workers' privacy

A Web App

- Workers **self-report** their connections
- Provides some **value back** to the workers so that it's in their best interest to report as many true connections as possible



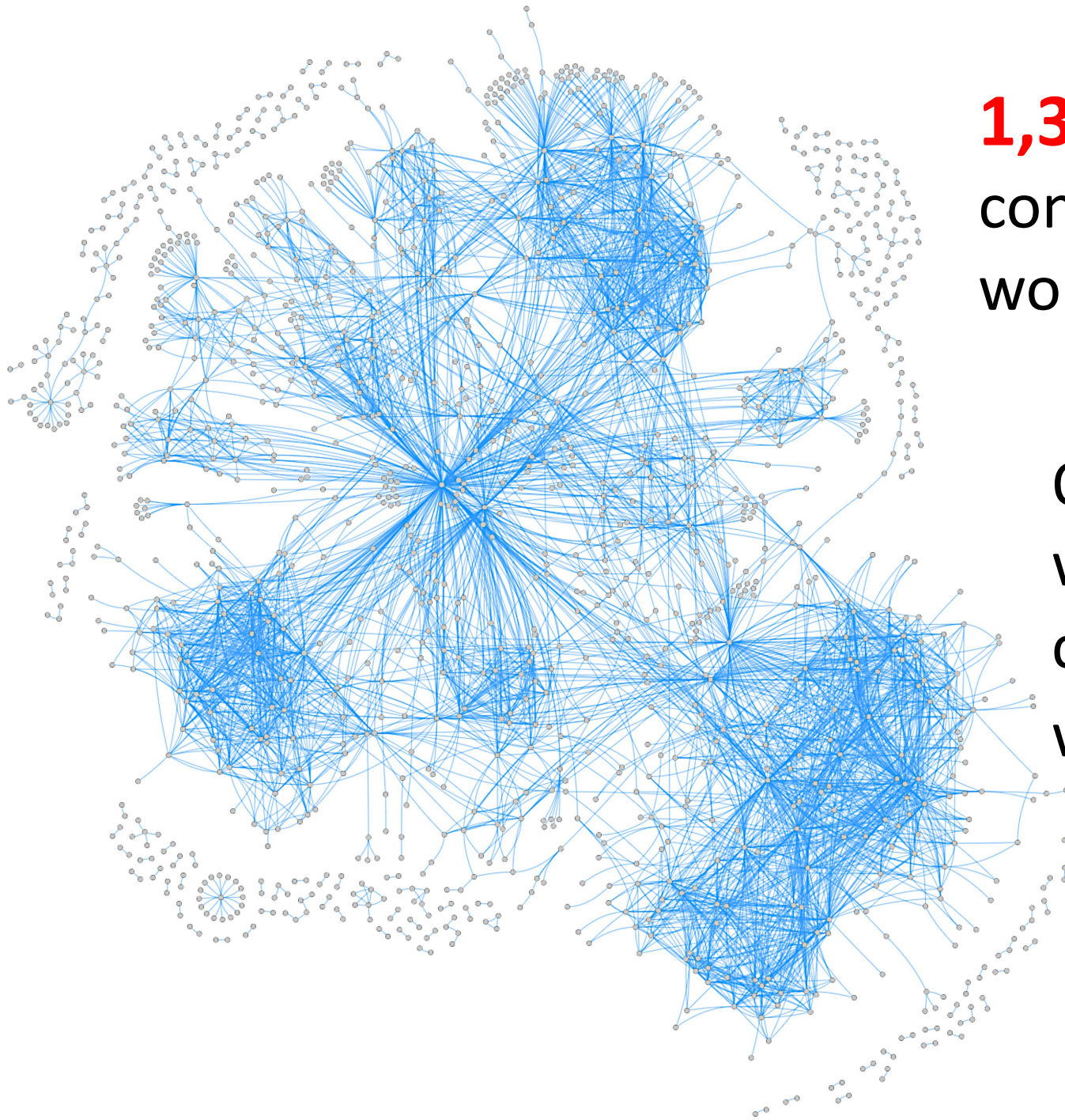
[Yin et al., 2016]



10,354 workers
(roughly a census of
Mechanical Turk
[Stewart et al. 2015])

5268
connections

[Yin et al., 2016]



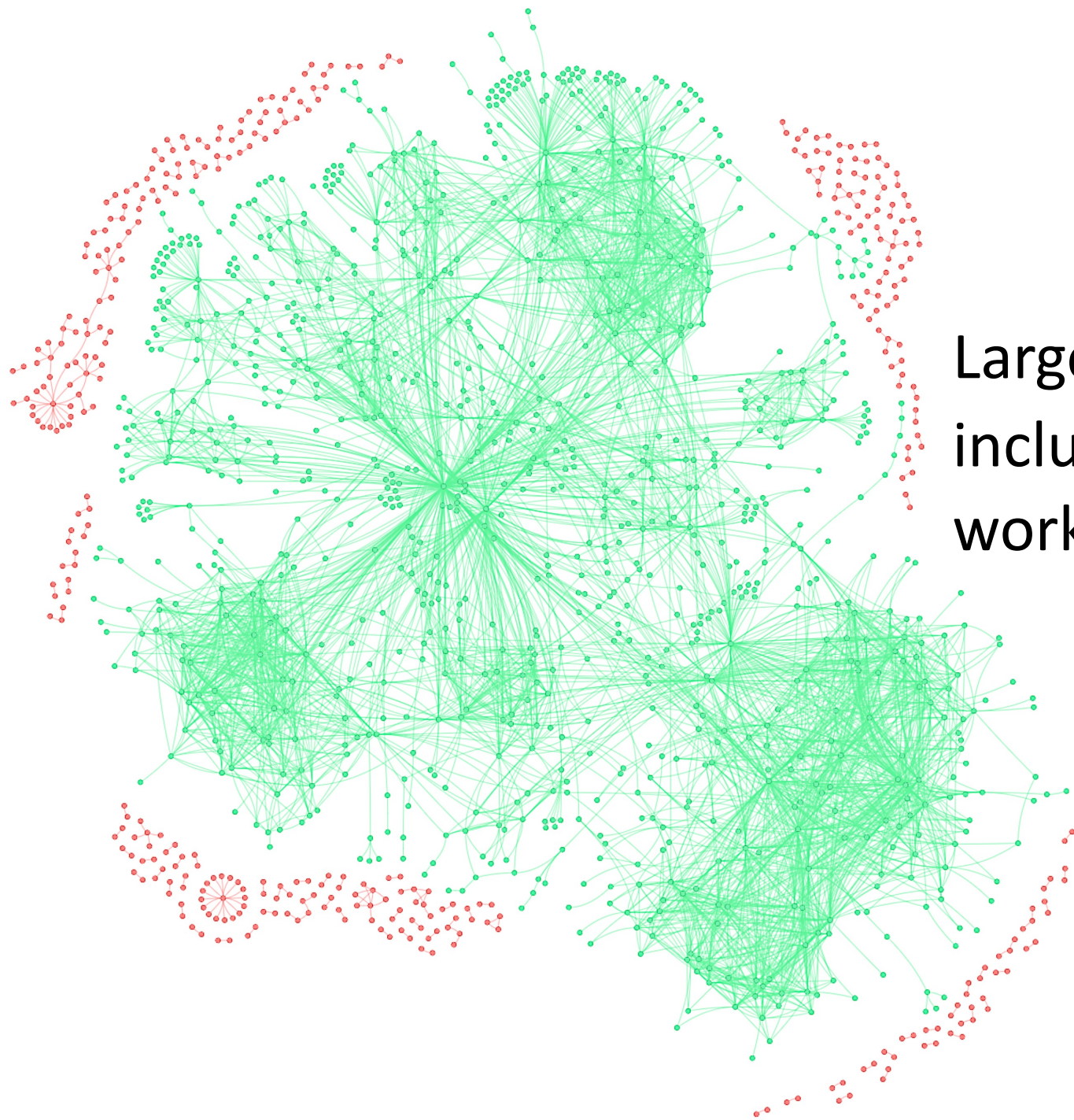
1,389 (13%)

connected
workers

On average,
workers
communicate
with **7.6** others

Max degree
is **321**

[Yin et al., 2016]



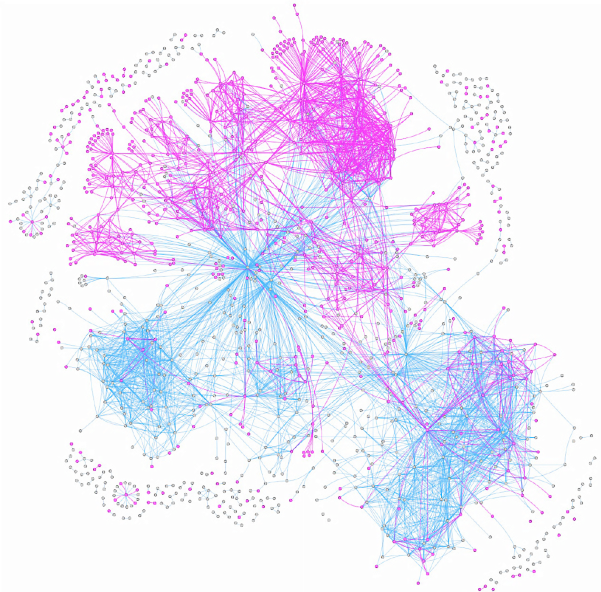
Largest component
includes **994 (72%)**
workers

[Yin et al., 2016]

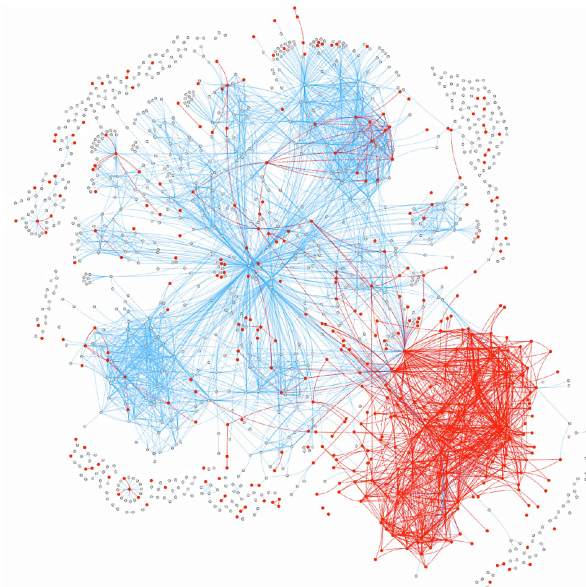
A Network Enabled By Forums

- **59%** of all workers and **83%** of connected workers reported using at least one forum.
- **90%** of all edges are between pairs of workers who communicate via forums, and **86%** are between pairs who communicate *exclusively* through forums.

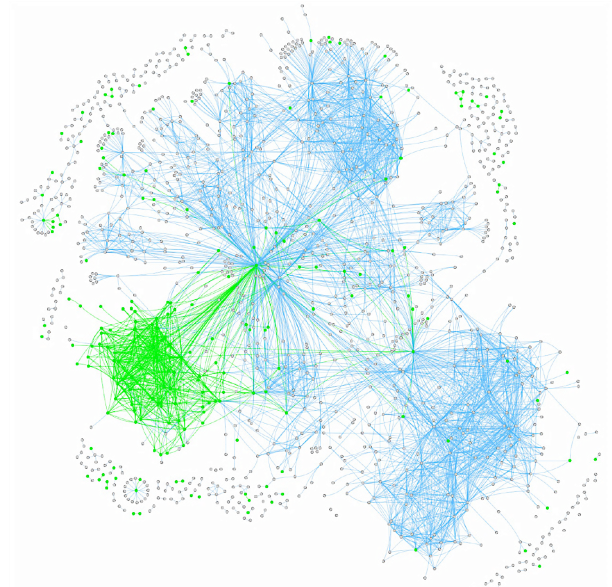
Forums Create Subcommunities



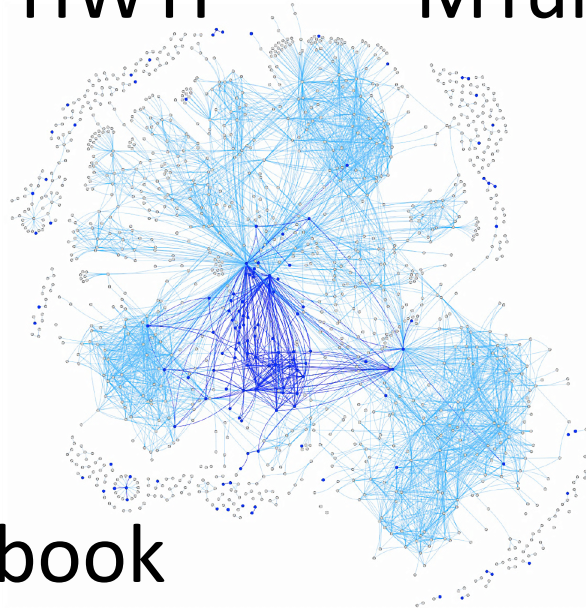
Reddit HWTF



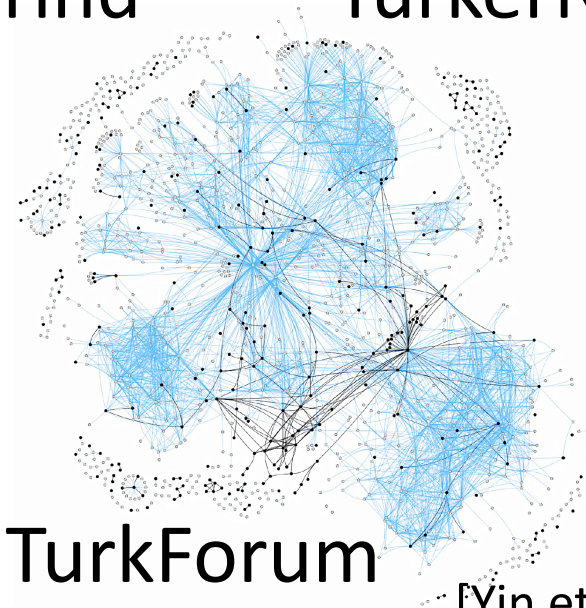
MTurkGrind



TurkerNation



Facebook



MTurkForum

[Yin et al., 2016]

Subcommunities Are Different



Topological Structure: How tightly connected is each subcommunity?



Temporal Dynamics: Do relationships endure over time?



Communication Content: Is communication social or strictly business?

Measures of Success

Property	Connected	Unconnected
Be active > 1 year	55%	46%
Use forums	83%	56%
Master	11%	7%
Approval rate	98.6%	97.4%

Connected workers were also **more likely** than unconnected workers to find our task **early**.

Takeaways and Related Best Practices

- Forum usage is widespread. Forums are the virtual “water coolers” of crowdworkers.
- Engage with workers on forums. Introduce yourself. Introduce your tasks.
- Actively monitor forum discussion about your task. When appropriate, request that workers do not discuss your task. Monitor anyway.
- Be careful about assuming independence!

Additional Best Practices

Behav Res (2012) 44:1–23
DOI 10.3758/s13428-011-0124-6

Conducting behavioral research on Amazon's Mechanical Turk

Winter Mason • Siddharth Suri

Maintain Good Relationships with Workers

- Set aside time to actively monitor your requester email account and respond to questions.
- Approve work quickly.
- Avoid rejecting work except in the most extreme of circumstances.

Tips to Make Your Project Run Smoothly

- Pilot, pilot, pilot! Test your task on your collaborators, other colleagues, and eventually small batches of workers.
- Iterate as many times as needed.

**If you remember one slide from this talk,
remember this!**

Tips to Make Your Project Run Smoothly

- Create clear instructions. Include quiz questions if needed. Pilot them and collect feedback.
- Create an attractive and easy-to-use interface. Pilot this too!
- Ask workers for feedback. Ask them to report bugs. Conduct exit surveys when appropriate. Workers generally want to help!

Thanks...

To Chien-Ju Ho, Andrew Mao, Joelle Pineau, Sid Suri, Hanna Wallach, and especially Ming Yin for extensive discussions and feedback

To Dan Goldstein, Chien-Ju Ho, Jake Hofman, Roozbeh Mottaghi, Sid Suri, Jaime Teevan, Ming Yin, Haoqi Zhang, and all of their collaborators for the use of material from their slides

And to all the people who sent me pointers to cool research... this tutorial was a crowdsourced effort!

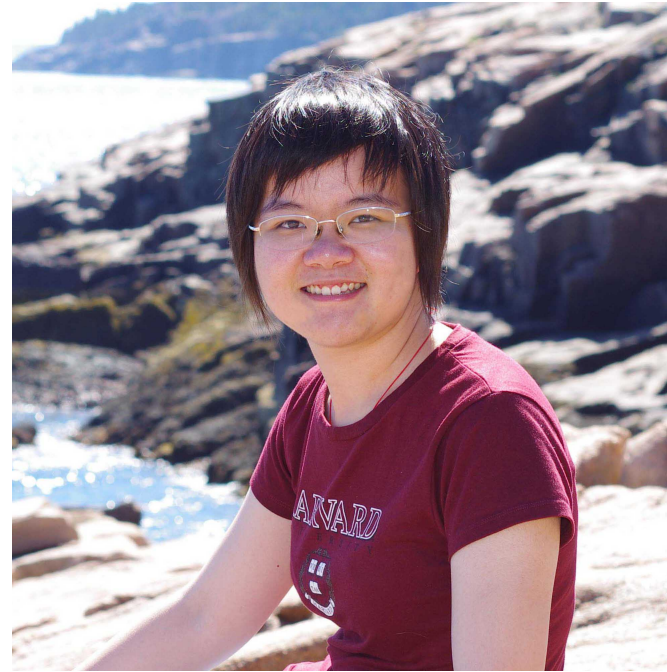
Extensive notes, slides, and eventually
video at

[http://www.jennwv.com/projects/
crowdtutorial.html](http://www.jennwv.com/projects/crowdtutorial.html)

On the Market



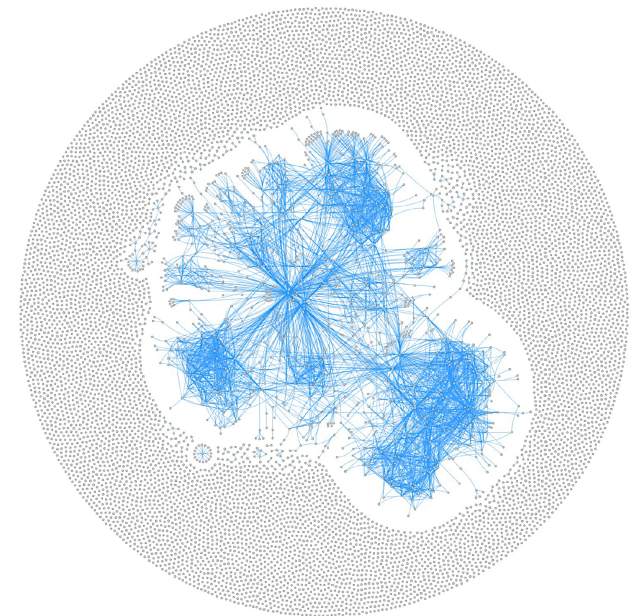
Chien-Ju Ho
Cornell
(Tuesday poster)



Ming Yin
Harvard

NIPS Workshop on Crowdsourcing
and Machine Learning, this Friday

<http://crowdml.cc/nips2016/>



HCOMP 2017

A nighttime photograph of Quebec City, Canada, featuring the illuminated Fairmont Le Château Frontenac hotel. The city lights and the St. Lawrence River are visible in the background.

October 24-26 in Quebec

Deadline in May

Chairs: Adam Kalai and Steven Dow

WIMML

Women in Machine Learning

Poster session **1:30-3:30pm** today, open to all!



jenn@microsoft.com

<http://jennwv.com>

@jennwvaughan