Making Better Use of the Crowd

Jenn Wortman Vaughan
Microsoft Research
A few disclaimers...
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 NLP and Machine Learning

2. Hybrid Intelligence Systems

3. Large Scale Studies of Human Behavior
Crowd

“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

Part 1:
The Potential of Crowdsourcing
The Potential of Crowdsourcing

1. Direct Applications to NLP and Machine Learning

2. Hybrid Intelligence Systems

3. Large Scale Studies of Human Behavior
Generating Labeled Data
Learner

Aggregation of noisy labels

"dog" "dog" "cat"

"cat" "cat" "cat"
Learner

Aggregation of noisy labels

“dog” “dog” “cat”

“cat” “cat” “cat”

Trained Model

“cat”
Learner

Aggregation of noisy labels

“dog” “dog” “cat”

“cat” “cat” “cat”

Trained Model

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

Inspired huge amounts of algorithmic work on aggregation.
Aggregating Labels with EM

- **Input:** Worker-generated labels for each instance

- Calculate an initial estimate of each instance’s label based on a simple majority vote

- Repeat until convergence:
  - Treating the current label estimates as truth, estimate each worker’s quality
  - Treating the quality estimates as truth, calculate the most likely label for each instance

- **Output:** One aggregated label for each instance

[Dawid and Skene, 1979]
Aggregating Labels with EM

• No guarantees on optimality, but tends to work pretty well in practice

• Many recent variants have been proposed to incorporate the varying difficulty levels of instances, worker expertise, the existence of “gold” tasks, etc.

[Dawid and Skene, 1979]
Beyond Simple Labels: Crowd Translation

• Crowdworkers are asked to
  – Translate sentences from one language to another
  – Edit other workers’ translations to make them more fluent and grammatical
  – Rank the quality of the resulting translations

• Can use machine learning to predict the highest quality translation based on sentence-level features, worker-level features, and ranks

[Zaidan and Callison-Burch, 2011]
Generating Similarity Measures

[Gomes et al., 2011]
Generating Similarity Measures

flags

no flags

[Gomes et al., 2011]
Generating Similarity Measures

Democrats

[Obama]

[Clinton]

Republicans

[Trump]

[Sarah Palin]

[George W. Bush]

[Gomes et al., 2011]
Crowd Clustering

Bayesian model

[Gomes et al., 2011]
Crowdsourcing for Evaluation
Evaluating Topic Models

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*

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

[Chang et al., 2009]
Human Debugging
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

Which component is the weakest link?

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

CRF model

- segment classifier
- supersegment classifier
- scene classifier
- shape prior
- object detector

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

CRF model

- segment classifier
- supersegment classifier
- scene classifier
- shape prior
- object detector

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

CRF model

segment classifier  supersegment classifier  scene classifier  shape prior  object detector

[Parikh & Zitnick, 2011; Mottaghi et al., 2013]
Humans less accurate at task, but system performance **still improved**

[Parikh & Zitnick, 2011; Mottaghi et al., 2013]
The Potential of Crowdsourcing

1. Direct Applications to NLP and 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 Scheduling
Cobi: Communitysourced Scheduling

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

[projectcobi.com]
1. Committeesourcing

2. Authorssourcing

3. Scheduling

4. Attendeesourcing
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

- 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]
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

[Teevan et al., 2016]
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.

[Teevan et al., 2016]
The MicroWriter breaks writing into microtasks. Microtasks can be shared with collaborators. Microtasks can be done while mobile. Collaborative writing 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.

[Teevan et al., 2016]
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.

[Teevan et al., 2016]
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.

[Teevan et al., 2016]
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

[Teevan et al., 2016]
Hybrid Intelligence in Industry
The Potential of Crowdsourcing

1. Direct Applications to NLP and 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

[Ur et al., 2016]
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

[Barrio et al., 2016]
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.

[Barrio et al., 2016]
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.

[Barrio et al., 2016]
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

[Barrio et al., 2016]
User Studies for Online Advertising
The Cost of Annoying Ads

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

VS.

How much do annoying ads cost publishers in dollars?

[Goldstein et al., 2013]
The Cost of Annoying Ads

Step 1: Use the crowd to identify annoying ads.

[Goldstein et al., 2013]
Good Ads

[Goldstein et al., 2013]
Bad Ads

[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?

[Goldstein et al., 2013]
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

[Goldstein et al., 2013]
Summary of Part 1

1. Direct Applications to NLP and 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!
Crowdsourcing Platforms
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

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

Workers

Requesters
Alternative Platforms

**CrowdFlower**

- Offers enterprise solutions for businesses with AI/data needs (search relevance evaluation, sentiment analysis, data classification)

**Clickworker**

- German platform with many European workers offering support for translation and web research plus mobile crowdsourcing
Alternative Platforms

**Prolific**
- UK-based platform focused on connecting researchers with subjects for experiments

**Upwork**
- Marketplace for freelancers with larger jobs like writing articles or designing websites
Crowdworker Demographics
Demographics of Mechanical Turk
Demographics of Mechanical Turk

- 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
- Can be big changes depending on time of day

[mturk-tracker.com]
Are workers dishonest?
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
  
  payment = $0.25 + ($0.25 * roll)

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

[Suri et al., 2011]
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?**

[Note: Diagram showing bar chart for proportions across rolls.]

[Suri et al., 2011]
Dishonesty Can Add Up

• “Are you the parent or guardian of a child with autism?”
  – 4.3% of participants said yes in control
  – 7.8% of participants said yes when told that this was a prescreening test for a further study

• Seems like a small difference, but would lead to (at least) 45% imposters in the subsequent study!

[Chandler and Paolacci, 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.

• Workers may deceive requesters to gain access to work. Prescreening should be done with care, ideally as part of a separate task.
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?

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]

[Ho et al., 2015]
Performance-Based Payments

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

[Ho et al., 2015]
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.

[Ho et al., 2015]
Can PBPs work?

• Task: Proofread an article and find spelling errors.

- We randomly insert 20 typos
  - sufficiently -> sufficiently
  - existence -> exisstance
  - ...

- Useful properties:
  - Quality is measurable
  - Exerting more effort -> better results

[Ho et al., 2015]
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

[Ho et al., 2015]
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

[Ho et al., 2015]
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

[Ho et al., 2015]
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

![Graph showing the relationship between bonus amounts and typos found.](graph.png)

- Could explain Shaw et al., 2011
- Could explain Yin et al., 2013

[Ho et al., 2015]
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 straightforward to guess which tasks are effort-responsive

[Ho et al., 2015]
Which tasks do PBPs work on?

- Proofreading
- Spotting differences

- Handwriting recognition
- Audio transcription

[Ho et al., 2015]
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

[Chandler and Kapelner, 2013]
Zooniverse
Real Science Online

The Zooniverse works

65,081,060
classifications so far by
1,546,928 registered volunteers
Gamification

[von Ahn and Dabbish, 2004]
Gamification

[von Ahn, Kedia, and Blum, 2006]
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 can also increase productivity. Well calibrated timed responses and score keeping (with or without high score lists) can both increase enjoyment.
The Communication Network
Within the Crowd
Implicit 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


[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

[Yin et al., 2016]
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

[Yin et al., 2016]
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.

[Yin et al., 2016]
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?

[Yin et al., 2016]
Measures of Success

<table>
<thead>
<tr>
<th>Property</th>
<th>Connected</th>
<th>Unconnected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Be active &gt; 1 year</td>
<td>55%</td>
<td>46%</td>
</tr>
<tr>
<td>Use forums</td>
<td>83%</td>
<td>56%</td>
</tr>
<tr>
<td>Master</td>
<td>11%</td>
<td>7%</td>
</tr>
<tr>
<td>Approval rate</td>
<td>98.6%</td>
<td>97.4%</td>
</tr>
</tbody>
</table>

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

[Yin et al., 2016]
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
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.

• Strive to be an ethical requester.
Guidelines for Academic Requesters

About the project

Version 2.0

The guidelines are currently going through a phase of editing to release the second edits are finalized and agreed upon the guidelines will be frozen again.

“Treat your workers with respect and dignity. Workers are not numbers and statisti rats. Workers are people and should be treated with respect.” - turker 'T', a Turkop
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

jenn@microsoft.com
http://jennwv.com
@jennwvaughan