Penn
UNIVERSITY of PENNSYLVANIA

Quality Control

Quality Control

Crowdsourcing typically takes place through an open call on the internet, where anyone can participate. How do we know that they are doing work conscientiously? Can we trust them not to cheat or sabotage the system? Even if they are acting in good faith, how do we know that they're doing things right?

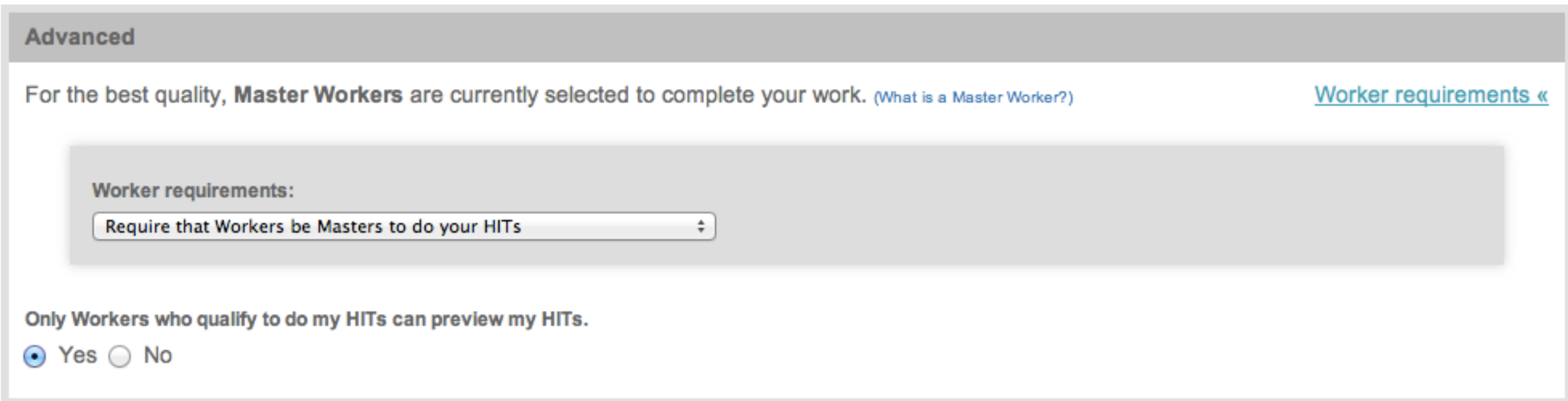
Different Mechanisms for Quality Control

- Reputation systems
- Qualification tests
- Aggregation and redundancy
- Embedded gold standard data
- Second-pass reviewing
- Economic incentives
- Statistical models

Reputation systems

- Mechanical Turk uses a reputation system
- Each Turker has a small number of variables associated with them, that are exposed to Requesters
- Past approval rate
- Number of HITs approved
- Whether the worker has received Amazon’s Masters qualification

Worker Requirements


Advanced

For the best quality, **Master Workers** are currently selected to complete your work. (What is a Master Worker?) Worker requirements «

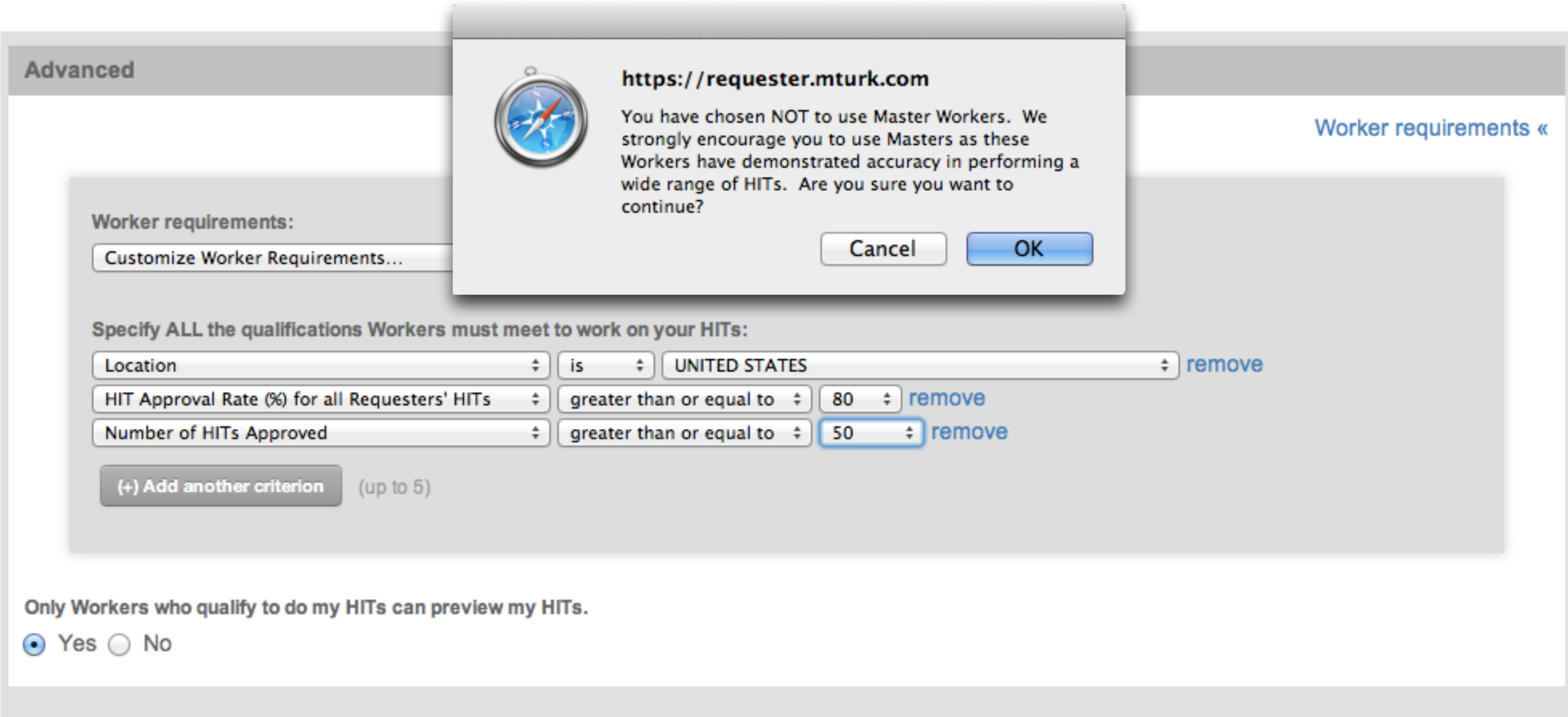
Worker requirements:

Require that Workers be Masters to do your HITs

Only Workers who qualify to do my HITs can preview my HITs.

Yes No

Worker Requirements



Masters

Masters are elite groups of Workers who have demonstrated accuracy on specific types of HITs. Workers achieve a Masters distinction by consistently completing HITs with a high degree of accuracy across a variety of Requesters. Masters must continue to pass our statistical monitoring to remain Mechanical Turk Masters. Because Masters have demonstrated accuracy, they can command a higher reward for their HITs. You should expect to pay Masters a higher reward.

Masters

- Amazon now nominates a subset (21k workers, estimated at 10% of all Turkers) of senior / good workers as “Masters”
- Amazon charges 30% commission for Masters versus their normal 10% rate
- They have now implemented this as the default qualification for new Requesters
- Why?

Masters: Pros

- People who use the Web UI are often newcomers who do not know to implement quality control.
- Masters will not touch badly designed and ambiguous tasks.
- Masters will not touch tasks paying less than minimum wage.

Masters: Cons

- There are many fewer Masters workers.
  - There is now a significant lag in the task being picked by workers.
  - The tasks now take much longer to complete.
- There is an increased cost because Masters demand decent wages.
- It is not clear in what tasks the Masters are tested and how a new worker can become a master.

Custom Qualifications

- In addition to the built in qualifications (masters, location, approval rate, min HITs completed), you can also create and manage your own qualifications
- These can be managed through the web interface or the API

Custom Qualifications

Home | Create | Manage | Developer | Help

Results | Workers | Qualification Types | We're Hiring! Learn More

Manage Qualification Types

Below is a list of your Qualification Types and the corresponding number of Workers.

Create New Qualification Type

Qualification Types

	Name	ID	Workers who have this Qualification	Creation Date	Description
	Trusted researc...	2GJ7Q67051QKTXPQMYHC7XMGI4AEYR	7	Thu Jun 20 17:35:18 UTC 2013	This qualification is granted to grad students and researchers. We use it it limit the participation in our pilot runs of experiments.
	Temporal Master	2YG46UXFCD45EMCI7IJNZKIN3WJVTB	0	Fri Mar 23 09:46:47 UTC 2012	This qualification is granted to Turkers who have demonstrated a high level of competency in temporal relations.
	Presidential Su...	279YQ4J3NAB6SPK4ZTPYOT1TJI6QDJ	0	Sat Nov 03 18:33:24 UTC 2012	This qualification was given to people who responded to the political survey before the election. We'll allow them to answer some follow up questions after the election.
	Monolingual	2K5F806UFRNF95EJ5KU9RSW6NKHF1D	6	Tue Jul 02 16:11:47	Granted to Workers who have demonstrated good skills at doing the

Qualification Tests

- The API also allows you to set up qualification tests that Workers must pass before doing your tasks
- What effects do you think qualification tests have?

Redundancy

Setting up your HIT

Reward per assignment $ 0.25

Tip: Consider how long it will take a Worker to complete each task. A 30 second task tha

Number of assignments per HIT 10

How many unique Workers do you want to work on each HIT?

Time allotted per assignment 2 Hours

Maximum time a Worker has to work on a single task. Be generous so that Workers are n

HIT expires in 14 Days

Maximum time your HIT will be available to Workers on Mechanical Turk.

Results are automatically approved in 7 Days

After this time, all unreviewed work is approved and Workers are paid.

ESP Game

"think like each other"

Player 1 guesses: purse
Player 1 guesses: bag
Player 1 guesses: brown

Success! Agreement on "purse"

Player 2 guesses: handbag

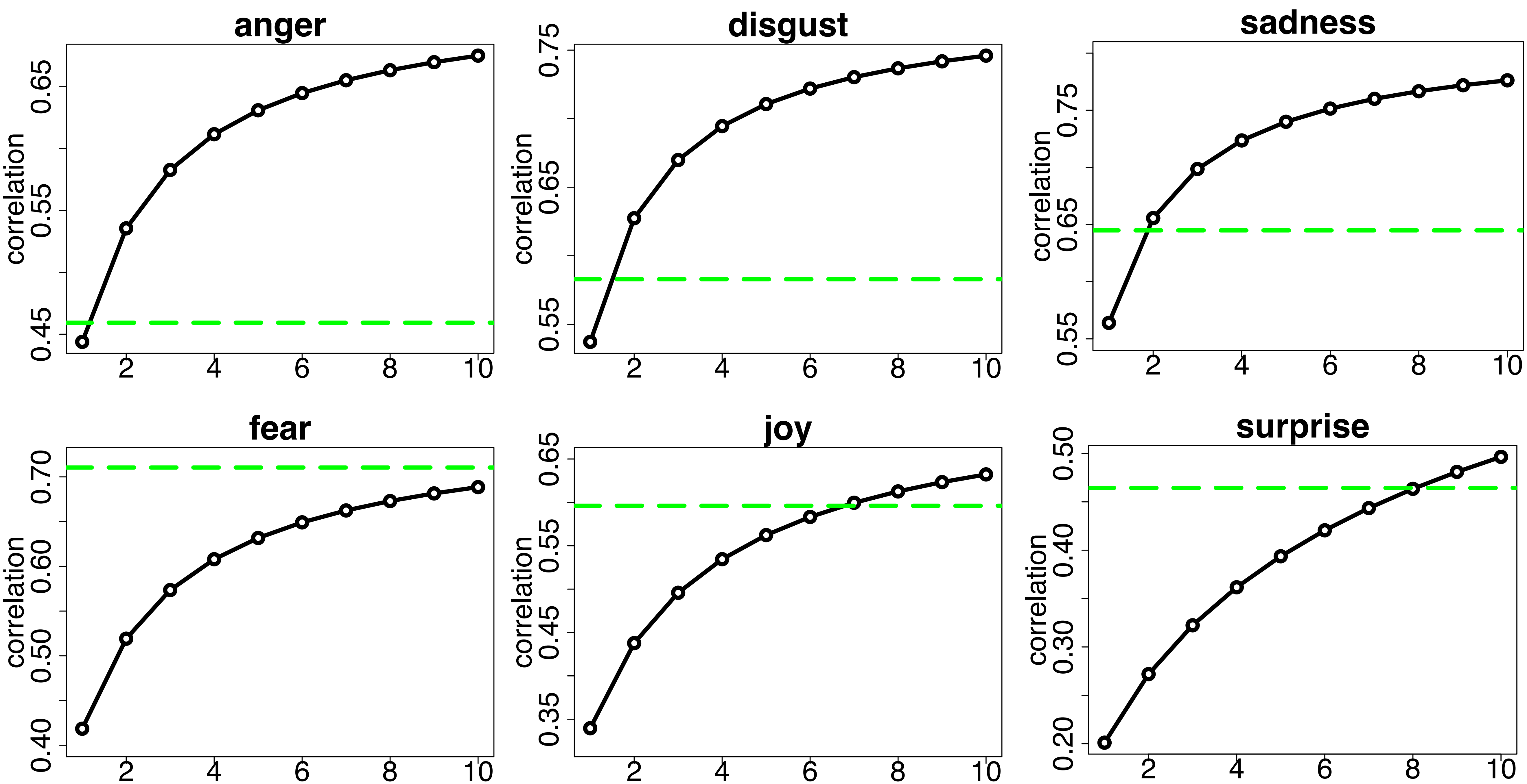
Player 2 guesses: purse
Success! Agreement on "purse"

MTurk for NLP

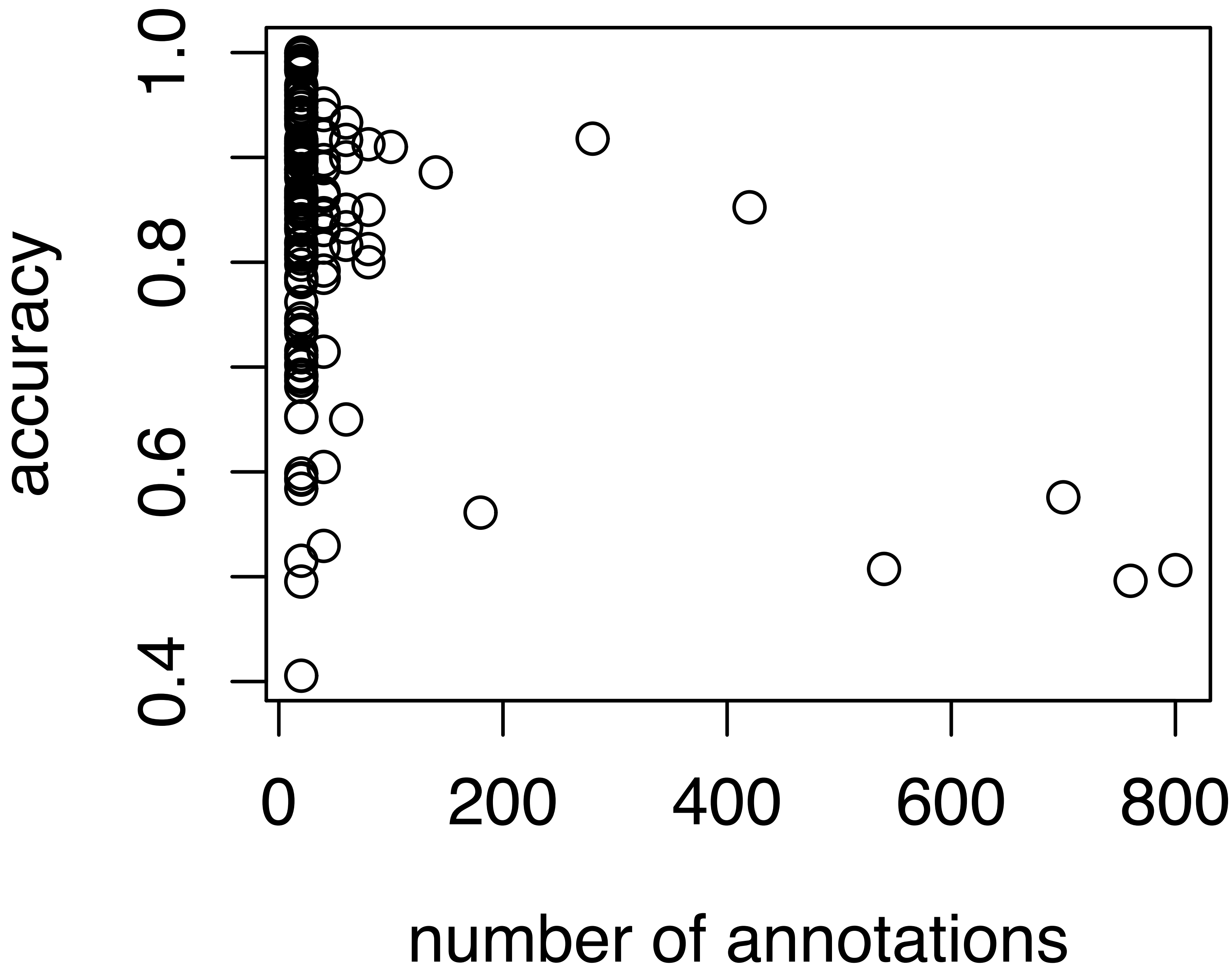
Snow, O'Connor, Jurafsky and Ng's EMNLP 2008 paper pioneered the use of Mechanical Turk for NLP

- Affect Recognition
  - fear("Tropical storm threatens NYC") > fear("Awesome goal for Beckham")
- Word Similarity
  - sim(man, boy) > sim(man, rooster)
- Textual Entailment
  - if "Microsoft was established in Italy in 1985" then "Microsoft was established in 1985"?
- Word Sense
  - "the West Bank" v. "the Bank of America"
- Temporal Annotation
  - *denoted* happens before *collapsed* in: "The condemned building collapsed when the crew detonated the charge."

Agreement with experts increases as we add more Turkers



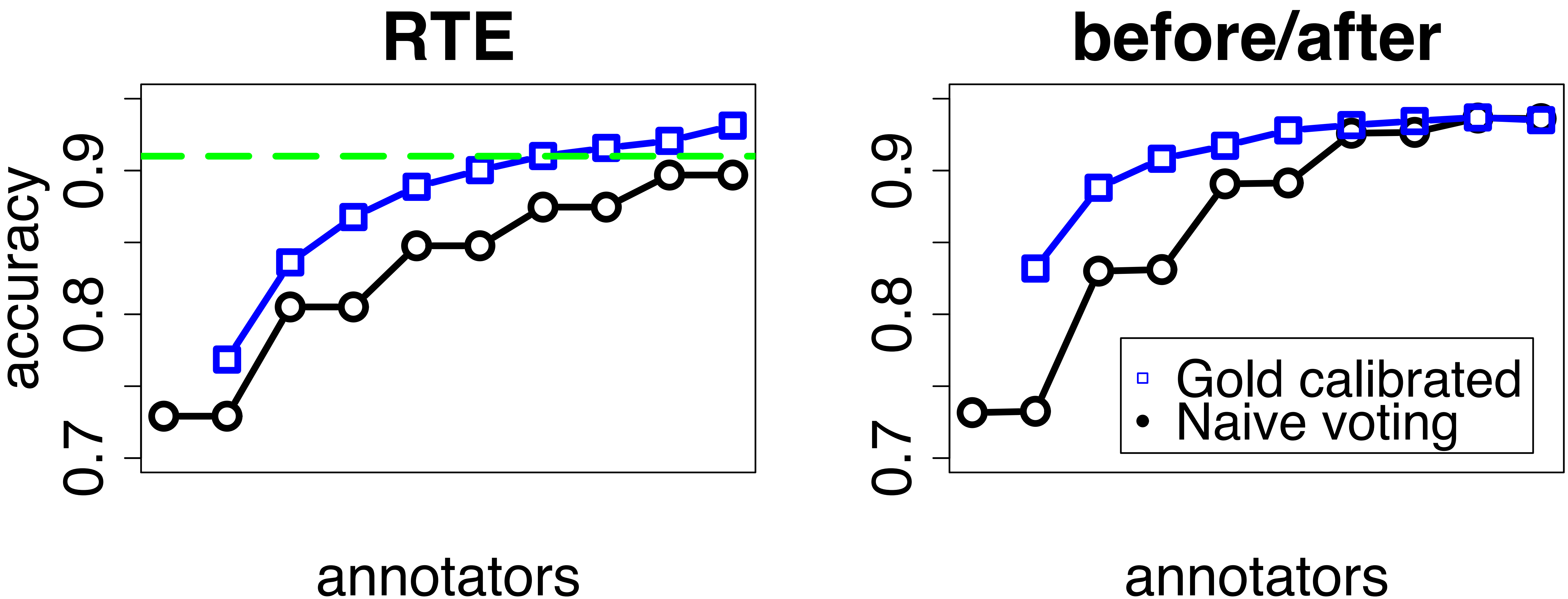
Accuracy of individual annotators



Calibrate the Turkers

- Instead of counting each Turker's vote equally, instead weight it
- Set the weight of the score based on how well they do on gold standard data
- Embed small amounts of expert labeled data alongside data without labels
- Votes will count more for Turkers who perform well, and less for those who perform poorly

Weighted votes



Other advantage of embedded gold data

- You can quickly detect and reject spammers
- Anyone who performs at chance on gold is randomly clicking
- I set two thresholds:
  - Reject all work from workers with chance performance
  - Accept all work from workers performing well (should be <100%, since some gold might be wrong)
  - Reject proportionally to performance for workers in between these values

Limitations?

- Embedding gold standard data seems like the way to go
- What are its limitations?

Limitations

- Requires objective answers – it is difficult to measure accuracy of subjective responses
- Applies mainly to structured data like multiple choice questions – things like content generation / free text responses can’t be calibrated in the same way
- Higher costs – requires creation of gold standard data by experts, requires multiple Workers to do each item

QC: Second-pass review

- Do second-pass grading when gold standard don’t allow automatic grading
- Often times the second-pass HIT can be automatically gradable
- This makes the whole pipeline fully automated and ensures high quality

Was arrested actress Heather Locklear because of the driving under the effect of an unknown medicine

Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

The actress Heather Locklear that is known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (Californium) because of the driving under the effect of an unknown medicine. A female witness observed she attempted in quite strange way how to go from their parking space in Montecito, speaker of the traffic police of californium told the warehouse `People'. The female witness told in detail, that Locklear 'pressed `after 16:30 clock accelerator and a lot of noise did when she attempted to move their car towards behind or forward from the parking space, and when it went backwards, she pulled itself together unites Male at their sunglasses'. A little later the female witness that did probably

- Why was Heather Locklear arrested?

  Driving while medicated

- Why did the bystander call emergency services?

  There was a lot of noise

- Where did the witness see her acting abnormally?

  In a parking lot

. Medikamentes unknown have the effect of a fahrens under actress heather locklear arrested

Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

In Santa. One is, melrose place the series of the role of the 'remember the locklear actress the heather this weekend, because of the fahrens Barbara (California) in effect unknown medikamentes arrested People 'magazine. The traffic police California, spokesman for the auszufahren montecito reported in its way from tried parklücke type strange right, you have seen as a witness. . In some Zeitung, as and when they tried to a great deal of 30 p.m., witness the detail of history locklear after 16: that durchdrückte peddle noise and its progress was made parklücke for the car or moving backwards, they had they times of their sonnenbrille ' . The first was probably recognised that locklear a nearby road and anhielt, had not, with the witness to the car off

- Why was Heather Locklear arrested?

- Why did the bystander call emergency services?

- Where did the witness see her acting abnormally?

People

HOME NEWS PHOTOS STYLE GAMES CELEBS PEOPLE TV ARCHIVE

Heather Locklear Arrested for driving under the influence of drugs

Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

SPONSORED LINKS

The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving und influence of drugs. A witness viewed her performing inappropriate maneuvers wh trying to take her car out fro parking in Montecito, as rev to People magazine by a spokesman for the Californian Highway Police. The witnes stated that around 4.30pm M Locklear "hit the accelerator violently, making excessive noise while trying to take her car c from the parking with abrupt and forth maneuvers. While reversing, she passed several times in front of his sunglass Shortly after, the witness, wh a first time, apparently had n recognized the actress, saw Ms.

- Why was Heather Locklear arrested?
  - She was arrested on suspicion of driving under the influence of drugs.

Driving under the influence
Driving while medicated
DUI
Driving while using drugs
Medikamentes

Economic incentives

Impact of compensation

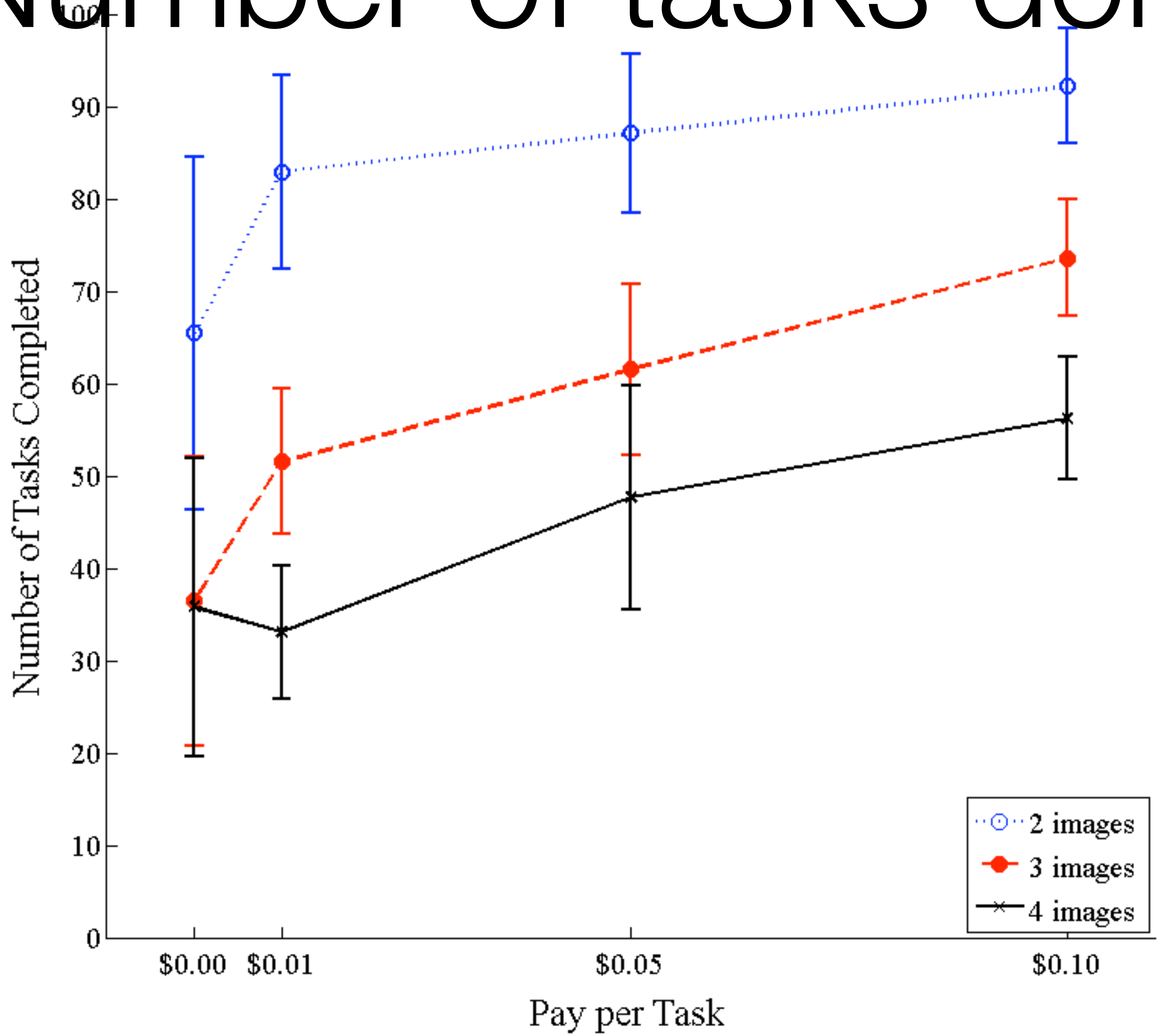
- Does compensation change the quantity of work performed (output)?
- Does it change the quality of the work (accuracy)?

Re-order Traffic Images

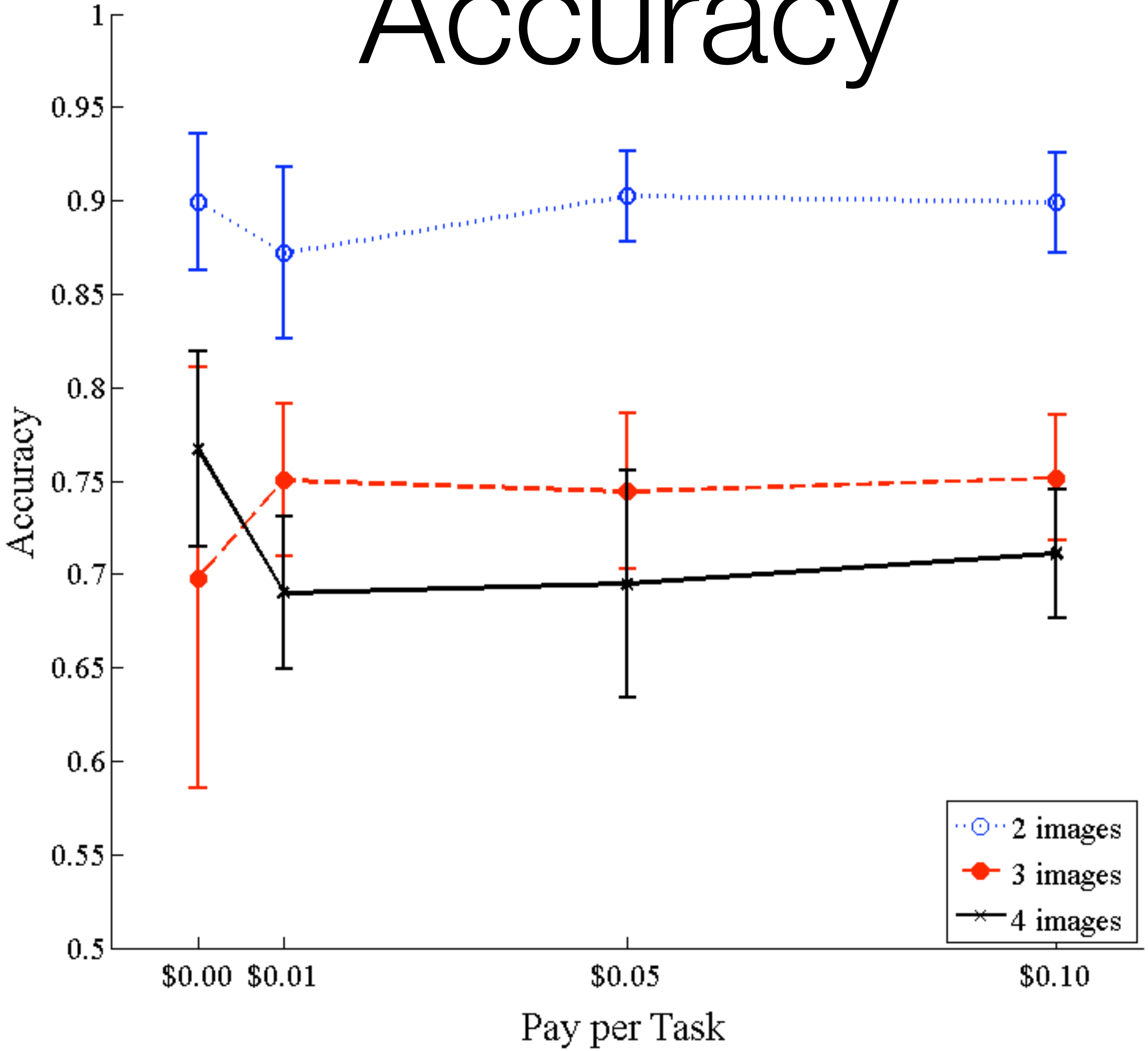
Unsorted

Sorted

Number of tasks done

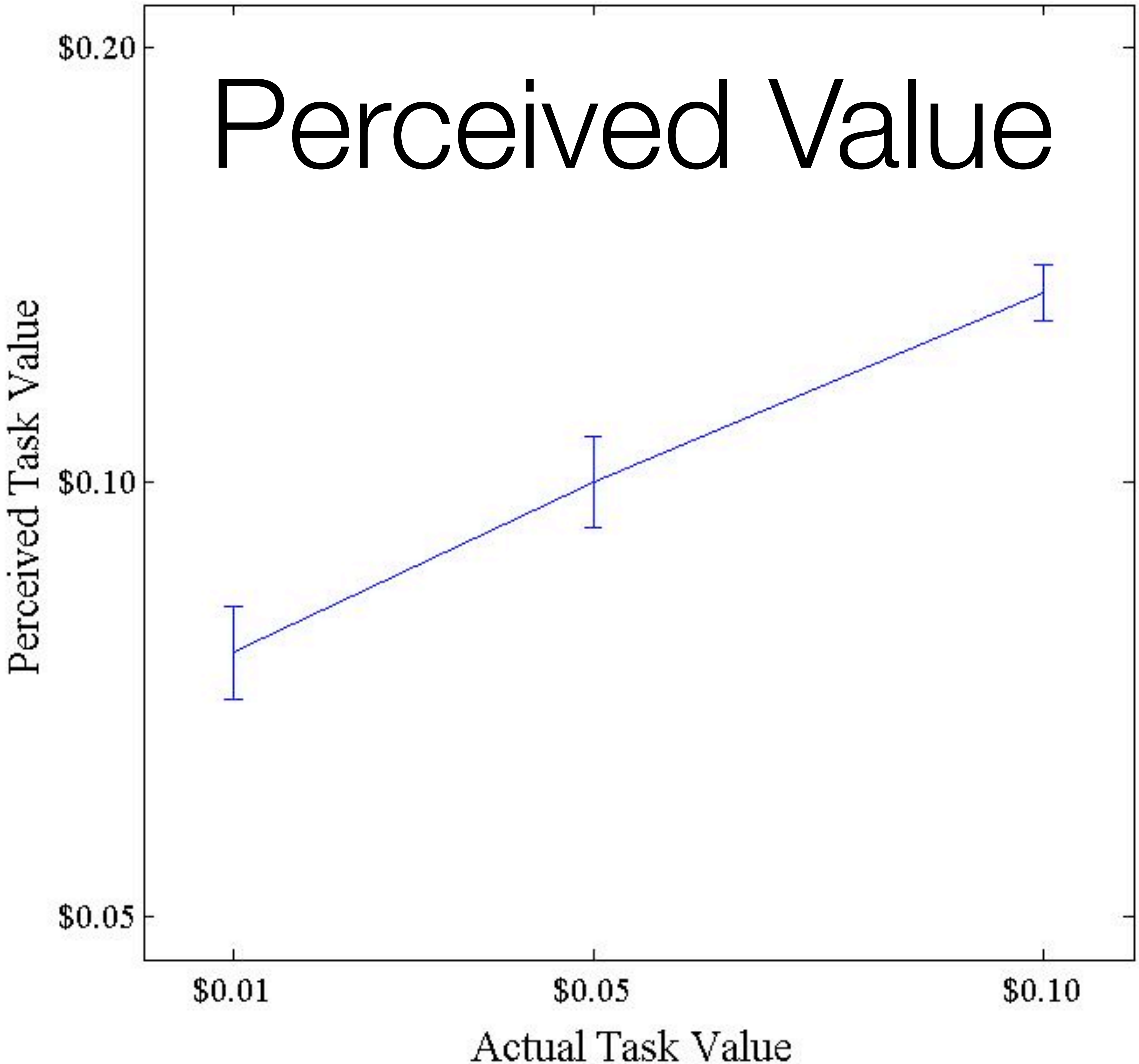


Accuracy



1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
$0.00
$0.01
$0.05
$0.10
Pay per Task
2 images
3 images
4 images

Perceived Value



Statistical Models

- Panos Iperiotis applied the EM algorithm to perform quality management of Mechanical Turk labels and workers
- Becky Passonneau and Bob Carpenter adapted this idea into a Bayesian model

Dawid and Skene (1977)

- *Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm*
- Examined application to medical diagnosis
- Patients are sometimes treated by multiple physicians, who can give different diagnoses
- Why? Doctors may has different questions. Patient may describe history differently. Doctors may classify symptoms differently

Observer Error

- Given that different doctors have different opinions, they can’t all be right.
- How often do individual physicians suffer from “observer error”? Are their errors systematic?
- Answers depend on the “true” diagnosis

Observer Error

- Observer error would be easy to calculate if we had ground truth
- Simply count the misdiagnoses and divide by the total number of diagnoses
- However, sometimes it is impossible to know what diagnosis is correct. Same set of symptoms can arise from multiple root causes.

url	worker1	worker2	worker3	worker4	worker5
google.com	porn	not porn	not porn	not porn	porn
sex-mission.com	porn	porn	porn	porn	not porn
curiousgeorge.com	porn	not porn	not porn	not porn	porn
youporn.com	porn	porn	porn	porn	not porn
panda-cam.gov	porn	porn	not porn	not porn	porn

url	majority vote	after EM
google.com	not porn	not porn
sex-mission.com	porn	porn
curiousgeorge.com	not porn	not porn
youporn.com	porn	porn
panda-cam.gov	porn	not porn

	Errors	Quality
worker1	60%	0%
worker2	20%	44%
worker3	0%	100%
worker4	0%	100%
worker5	100%	100%

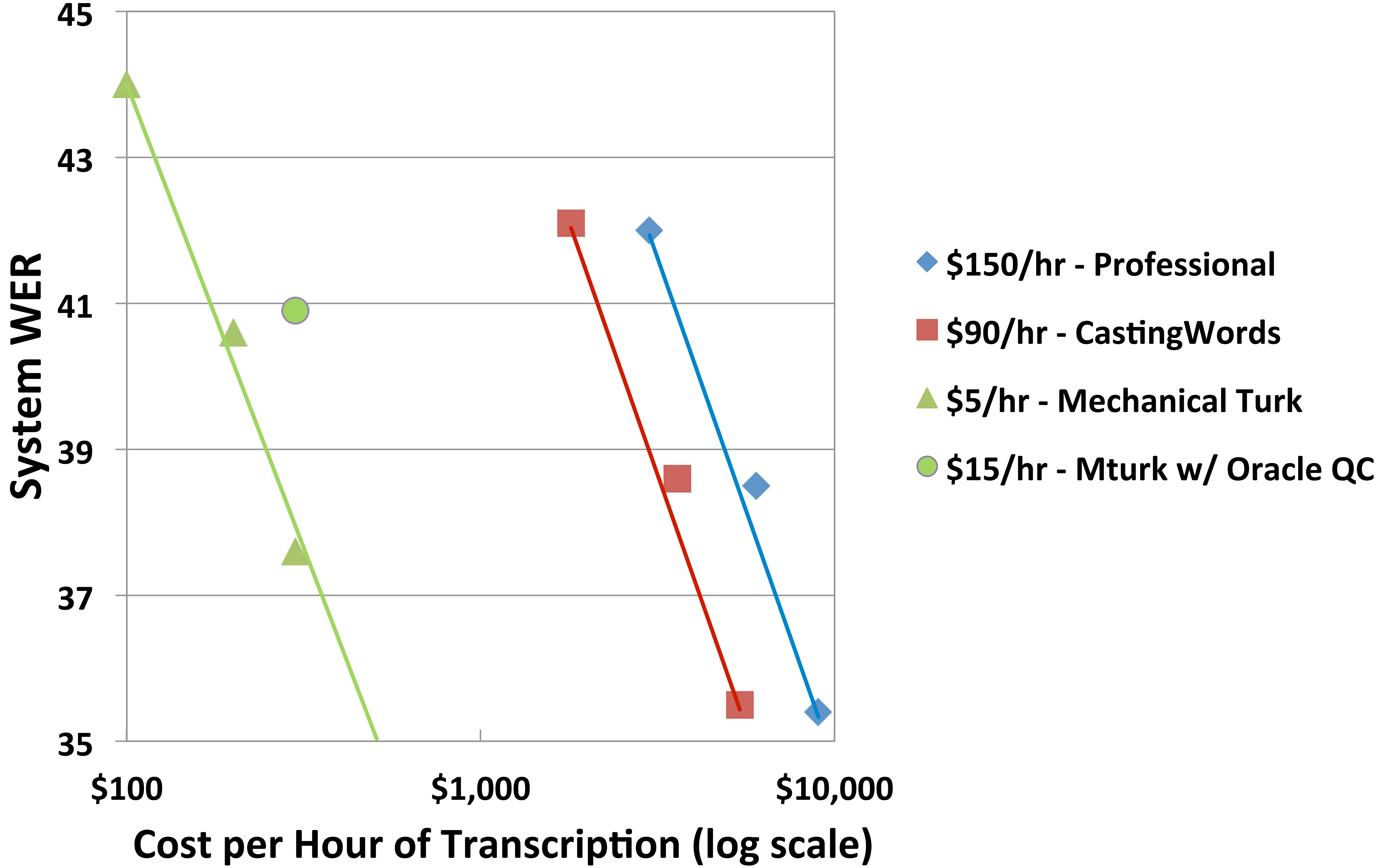
Quality control

- Goal: Avoid spammers and lazy people
  - Make sure people have the right background
  - Limit to a given country or native language
  - Ask initial screening questions
- Ensure that people read the instructions
  - Fade in the instructions.
  - Show a video with a keyword in the middle of it
  - Frame the problem as a contribution towards science
- Drop people who disagree to often with others

Quality is overrated

- QC often relies on redundancy
- Therefore increased quality effectively reduces quantity
- If our goal is to train a statistical NLP system, should we get fewer high quality labels or more lower quality labels?

Quality is overrated



Quality is overrated

- Chris Lin, Mausam, and Dan Weld systematically investigated whether it is better to re-label examples when training a classifier on a fixed budget
- A little more than half the time it was better to skip relabeling and just get more data
- Takeaway: it depends on the classifier