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

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

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

Advanced

Worker requirements:
Customize Worker Requirements...

Specify ALL the qualifications Workers must meet to work on your HITs:

Location: United States
HIT Approval Rate (%): 80
Number of HITs Approved: 50

Only Workers who qualify to do my HITs can preview my HITs.
Yes □ No □
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

### Manage Qualification Types

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

[Create New Qualification Type](#)

<table>
<thead>
<tr>
<th>Name</th>
<th>ID</th>
<th>Workers who have this Qualification</th>
<th>Creation Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted research...</td>
<td>2GJ7Q67051QKTXPQMYHC7XMGI4AEYR</td>
<td>7</td>
<td>Thu Jun 20 17:35:18 UTC 2013</td>
<td>This qualification is granted to grad students and researchers. We use it to limit the participation in our pilot runs of experiments.</td>
</tr>
<tr>
<td>Temporal Master</td>
<td>2YG46UXFCD45EMC17IJNZKIN3WJVTB</td>
<td>0</td>
<td>Fri Mar 23 09:46:47 UTC 2012</td>
<td>This qualification is granted to Turkers who have demonstrated a high level of competency in temporal relations.</td>
</tr>
<tr>
<td>Presidential Su...</td>
<td>279YQ4J3NAB6SPK4ZTPYOT1TJ16QDJ</td>
<td>0</td>
<td>Sat Nov 03 18:33:24 UTC 2012</td>
<td>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.</td>
</tr>
<tr>
<td>Monolingual</td>
<td>2K5E806L1ERNEQ55F15KI9BSW6NKH5E1D</td>
<td>6</td>
<td>Tue Jul 02 16:11:47</td>
<td>Granted to Workers who have demonstrated good skills at doing the task in the specified language.</td>
</tr>
</tbody>
</table>
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?
### Setting up your HIT

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reward per assignment</td>
<td>$ 0.25</td>
</tr>
<tr>
<td>Time allotted per assignment</td>
<td>2 hours</td>
</tr>
<tr>
<td>HIT expires in</td>
<td>14 days</td>
</tr>
<tr>
<td>Results are automatically approved in</td>
<td>7 days</td>
</tr>
</tbody>
</table>

**Tip:** Consider how long it will take a Worker to complete each task. A 30 second task typically pays $0.25.

**Number of assignments per HIT:**

How many unique Workers do you want to work on each HIT? 10

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

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

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

![Graphs showing correlation with increased number of annotations for different emotions: anger, disgust, sadness, fear, joy, and surprise.](image-url)
Accuracy of individual annotators

![accuracy vs number of annotations](image.png)
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

The worker response likelihoods $P(y_w|x=Y)$ and $P(y_w|x=N)$ can be directly estimated from frequencies of worker performance on gold standard examples. (If we used maximum likelihood estimation with no Laplace smoothing, then each $y_w|x$ is just the worker's empirical confusion matrix.) For MAP label estimation, the above equation describes a weighted voting rule: each worker’s vote is weighted by their log likelihood ratio for their given response. Intuitively, workers who are more than 50% accurate have positive votes; workers whose judgments are pure noise have zero votes; and anticorrelated workers have negative votes. (A simpler form of the model only considers accuracy rates, thus weighting worker votes by $\log \text{acc}_w - \text{acc}_w.$ But we use the full unconstrained multinomial model here.)

5.1.1 Example tasks: RTE-1 and event annotation

We used this model to improve accuracy on the RTE-1 and event annotation tasks. (The other categorical task, word sense disambiguation, could not be improved because it already had maximum accuracy.) First we took a sample of annotations giving $k$ responses per example. Within this sample, we trained and tested via 20-fold cross-validation across examples. Worker models were fit using Laplace smoothing of 1 pseudocount; label priors were uniform, which was reasonably similar to the empirical distribution for both tasks.

![Figure 7: Gold-calibrated labels versus raw labels](image)

Figure 7 shows improved accuracy at different numbers of annotators. The lowest line is for the naive 50% majority voting rule. (This is equivalent to the model under uniform priors and equal accuracies across workers and labels.) Each point is the data set’s accuracy against the gold labels, averaged across resamplings each of which obtains $k$ annotations per example. RTE has an average +4.0% accuracy increase, averaged across 2 through 10 annotators. We find a +3.4% gain on event annotation.

Finally, we experimented with a similar calibration method for numeric data, using a Gaussian noise model for each worker: $y_w|x \sim N(x + \mu_w, \sigma_w)$. On the affect task, this yielded a small but consistent increases in Pearson correlation at all numbers of annotators, averaging a +0.6% gain.

6 Training a system with non-expert annotations

In this section we train a supervised affect recognition system with expert vs. non-expert annotations.

6.1 Experimental Design

For the purpose of this experiment we create a simple bag-of-words unigram model for predicting affect and valence, similar to the SWAT system (Katz et al., 2007), one of the top-performing systems on the SemEval Affective Text task.

For each token $t$ in our training set, we assign $t_e$ weights for each emotion $e$ equal to the average emotion score observed in each headline $H$ that $t_e$ participates in. i.e., if $H_t$ is the set of headlines containing the token $t$, then:

$$\text{Score}(e, t) = \sum_{H \in H_t} \text{Score}(e, H) |H_t|$$

With these weights of the individual tokens we may then compute the score for an emotion $e$ of a new headline $H$ as the average score over the set of tokens $t \in H$ that we've observed in the training set (ignoring those tokens not in the training set), i.e.:

$$\text{Score}(e, H) = \sum_{t \in H} \text{Score}(e, t) |H|$$

Where $|H|$ is simply the number of tokens in headline $H$, ignoring tokens not observed in the training set.

Unlike the SWAT system we perform no lemmatization, synonym expansion, or any other preprocessing of the tokens; we simply use whitespace-separated tokens within each headline.
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
Heather Locklear Arrested for Driving Under the Influence of Drugs

The actress Heather Locklear, known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (California) 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 there was a lot of noise in a parking lot.
Actress Heather Locklear, known for her role in the series "Melrose Place," was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness observed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4:30 pm, Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses."

Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw her attempting in quite strange way how to go from their parking space in Montecito, according to the traffic police of California, who told the warehouse "People." The witness told in detail that Locklear "pressed the pedal after 16:30 clock and a lot of noise did when she attempted to move towards behind or forward from the parking space, and when she went backwards, she pulled it a few times in their sunglasses." A little later, the female witness, who probably had not recognized Locklear on a nearby road and stopped the car exit.

Questions:

1. Why was Heather Locklear arrested?
2. Why did the bystander call emergency services?
3. Where did the witness see her acting abnormally?
The actress Heather Locklear, known for her role as Amanda in the popular television series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses." Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw her probably had not recognized the actress, saw her.
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

![Graph showing the number of tasks completed at different pay levels for different numbers of images.](image_url)

**Average parameter estimates for the effect of pay in the hierarchical linear model across users.**

<table>
<thead>
<tr>
<th>Pay per Task</th>
<th>Number of Tasks Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.00</td>
<td>0</td>
</tr>
<tr>
<td>$0.01</td>
<td>10</td>
</tr>
<tr>
<td>$0.05</td>
<td>50</td>
</tr>
<tr>
<td>$0.10</td>
<td>100</td>
</tr>
</tbody>
</table>

- **2 images**
- **3 images**
- **4 images**

**Study 1**

Results indicate that across all difficulty levels participants chose to complete more tasks on average when the pay was higher ($0.10 per task). The strong and significant dependence of changes in extrinsic motivation (payment), which can vary by at least $0.10 per task, might have expected that variability in intrinsic motivation (e.g., enjoyment of the task) would have overwhelmed the effect of changes in extrinsic motivation (payment), which can vary by at least $0.10 per task. Nevertheless, the finding is reassuring since, as noted above, one consistent with standard economic theory, which predicts that the more a person is paid to do X, the more of X they will do.
two ways: first, using a simple one confirmed quantitatively what is visually apparent in Figure 3 in the normalized sum of squared differences between the correct order; and second, using Spearman's rank correlation (the proportion of image sets that were sorted into the correct improve accuracy, which we measured in tw.

As Figure 3 indicates, however, increasing compensation did not quality of performance as well.

rates studied ought to be sufficient to observe variability in the output on compens.

most $0.10 per task. The strong and significant dependence of changes in extrinsic motivation (payment), which can vary by at

enjoyment of the task) would have overwhelmed the effect of might have exp.

Nevertheless, the finding is reassuring since, as noted above, one more a person is paid to do X, the more of X they will do consistent with standard economic theory, which predicts that the

than 10 sets. These results, in other words, are completel

proportionately more of the participants paid $0.01 sorted fewer maximum possible than those paid $0.10 or nothing at all, and also found that more of the participants paid $0.10 sorted the thus hereafter we focus on the effect of pay on quantity averaged

across all payment levels, the number of completed tasks increased with increasing difficulty. We also observe, however, that there is no interaction between difficulty and compensation,

decreased with difficult of task

that a similar income and gender distribution, and also recorded 58% female respondents.

http://behind

turk demographics.html

Previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender

titudes and respondents respectively, and
each user. The parameters

between the actual and correct ordering, the model is

correctly, the model is

and when accuracy is defined as Spearman's rank correlation (ρ), which is described a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender

titudes and respondents respectively, and
each user. The parameters

between the actual and correct ordering, the model is

correctly, the model is

and when accuracy is defined as Spearman's rank correlation (ρ), which is described a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.

The previous user surveys of the AMT population subject pool was therefore reasonably diverse, consistent with

$160k, and 3.2

f examine the effects of the wage rate.

Figure 3. Table

3.2

- 1 reveals two main findings: first, that across all difficulty
description a similar income and gender
distribution, and also recorded 58% female respondents.
Regardless of the accuracy measure or analytical method used, we found that the wage rate had no significant effect on the participants' accuracy in sorting the image sets. First, as indicated in Table 1, the parameter estimates in the hierarchical model for the four levels of pay were not reliably different from each other; and second, one-way ANOVAs of wage rate on proportion correct and rank correlation were not statistically reliable (proportion correct: $F(3,607) = 0.66$, $ns$; rank correlation: $F(3,607) = 0.82$, $ns$).

### Discussion

One possible explanation for the absence of an effect of wages on accuracy is that subjects simply assumed they would be paid regardless of performance. This explanation is somewhat unlikely, as AMT's policy is that requestors are only obligated to pay for accurate or useful work, and workers are informed of the policy. Nevertheless, to check the possibility we ran an additional experiment with a single payment level ($0.01) that provided different information to participants regarding the importance of accuracy. In this additional experiment, some participants were given the same instructions as before while others were told that one out of every four image sets was a test image set used to gauge their accuracy. Within this latter condition, we also created four variants: (i) participants only informed that accuracy would be measured; (ii) participants also shown feedback on their accuracy after every fourth image set; (iii) participants also told explicitly that their pay would be contingent on their performance; and (iv) participants shown feedback and also told that pay was contingent. We found that quantity and quality results were indistinguishable in all these conditions, suggesting that participants in all conditions were in fact treating their pay as performance dependent.

Although the differential effect of pay on quantity and quality is at first puzzling, we note that previous studies have also found positive effects of financial incentives on quantity of work performed but no effect on quality [24]. We hypothesize, moreover, that the difference derives from an "anchoring" effect, similar to effects that have been observed in other domains of judgment and decision-making [19-21]. As Figure 4 shows, when surveyed after the completion of their tasks, workers in all conditions generally felt that the appropriate compensation for the work they had just performed was greater than what they had received, but the values they expressed depended significantly ($\chi^2 = 243.61, p < 0.0001$) on their actual compensation: on average, workers paid $0.01 per task felt they should have received $0.05; workers who were paid $0.05 felt they should have received $0.08; and workers who were paid $0.10 felt they should have received $0.13. On the one hand, therefore, paying people more to perform a task makes that task more attractive relative to their available outside options, such as other HITs on AMT; thus subjects in the higher pay conditions stayed longer and completed more tasks than those in low pay condition. On the other hand, because of the anchoring effect, all workers felt like they were being paid less than they deserved; thus were no more motivated to perform better no matter how much they were actually paid.

### STUDY 2: WORD PUZZLES

4.1 Methods

In spite of this explanation, one might suspect that the absence of an effect on accuracy may be an artifact of the task itself—because, for example, it allowed only a small number of potential solutions (in the "easy" condition, for example, only two solutions were possible); or because subjects could not easily improve the quality of their answers with greater effort. To address this possibility, we performed another experiment, using a similar experimental design, but changing the task to finding words hidden in a random array of letters (see Figure 5).

4.1.1 Design

For each puzzle, we provided a list of words that might be found in the puzzle, although only a subset of the list was actually hidden in the word puzzle. As before, this task allowed us to measure quantity (number of puzzles completed) and quality (fraction of words found per puzzle) independently; but because participants did not know how many words from the list could be found, we could infer whether the effect of pay was driven by quantity or quality.

Figure 4. Post-hoc survey shows perceived value of the task increases with the actual pay, but is always slightly greater than the actual pay received.

Figure 5. Screenshot of Study 2. Participants found words hidden in a grid of letters.
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 have 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.
<table>
<thead>
<tr>
<th>url</th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>porn</td>
<td>not porn</td>
<td>not porn</td>
<td>not porn</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not porn</td>
</tr>
<tr>
<td>curiousgeorge.com</td>
<td>porn</td>
<td>not porn</td>
<td>not porn</td>
<td>not porn</td>
<td>porn</td>
</tr>
<tr>
<td>youporn.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not porn</td>
</tr>
<tr>
<td>panda-cam.gov</td>
<td>porn</td>
<td>porn</td>
<td>not porn</td>
<td>not porn</td>
<td>porn</td>
</tr>
<tr>
<td>url</td>
<td>majority vote</td>
<td>after EM</td>
<td>Errors</td>
<td>Quality</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------</td>
<td>----------</td>
<td>--------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>google.com</td>
<td>not porn</td>
<td>not porn</td>
<td>worker1</td>
<td>60%</td>
<td>0%</td>
</tr>
<tr>
<td>sex-mission.com</td>
<td>porn</td>
<td>porn</td>
<td>worker2</td>
<td>20%</td>
<td>44%</td>
</tr>
<tr>
<td>curiousgeorge.com</td>
<td>not porn</td>
<td>not porn</td>
<td>worker3</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>youporn.com</td>
<td>porn</td>
<td>porn</td>
<td>worker4</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>panda-cam.gov</td>
<td>porn</td>
<td>not porn</td>
<td>worker5</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
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

Comparing Cost of Reducing WER

- 35%
- 37%
- 39%
- 41%
- 43%
- 45%

System%WER

Cost per Hour of Transcription (log scale)

- $150/hr - Professional
- $90/hr - CastingWords
- $5/hr - Mechanical Turk
- $15/hr - Mturk w/ Oracle QC

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