Quality Control

Crowdsourcing and Human Computation
Lecture 12

Instructor: Chris Callison-Burch
TA: Ellie Pavlick

Website: crowdsourcing-class.org

Classification System for Human Computation

- Motivation
- **Quality Control**
- Aggregation
- Human Skill
- Process Order
- Task-request Cardinality

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

- Aggregation and redundancy
- Embedded gold standard data
- Reputation systems
- Economic incentives
- Statistical models

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

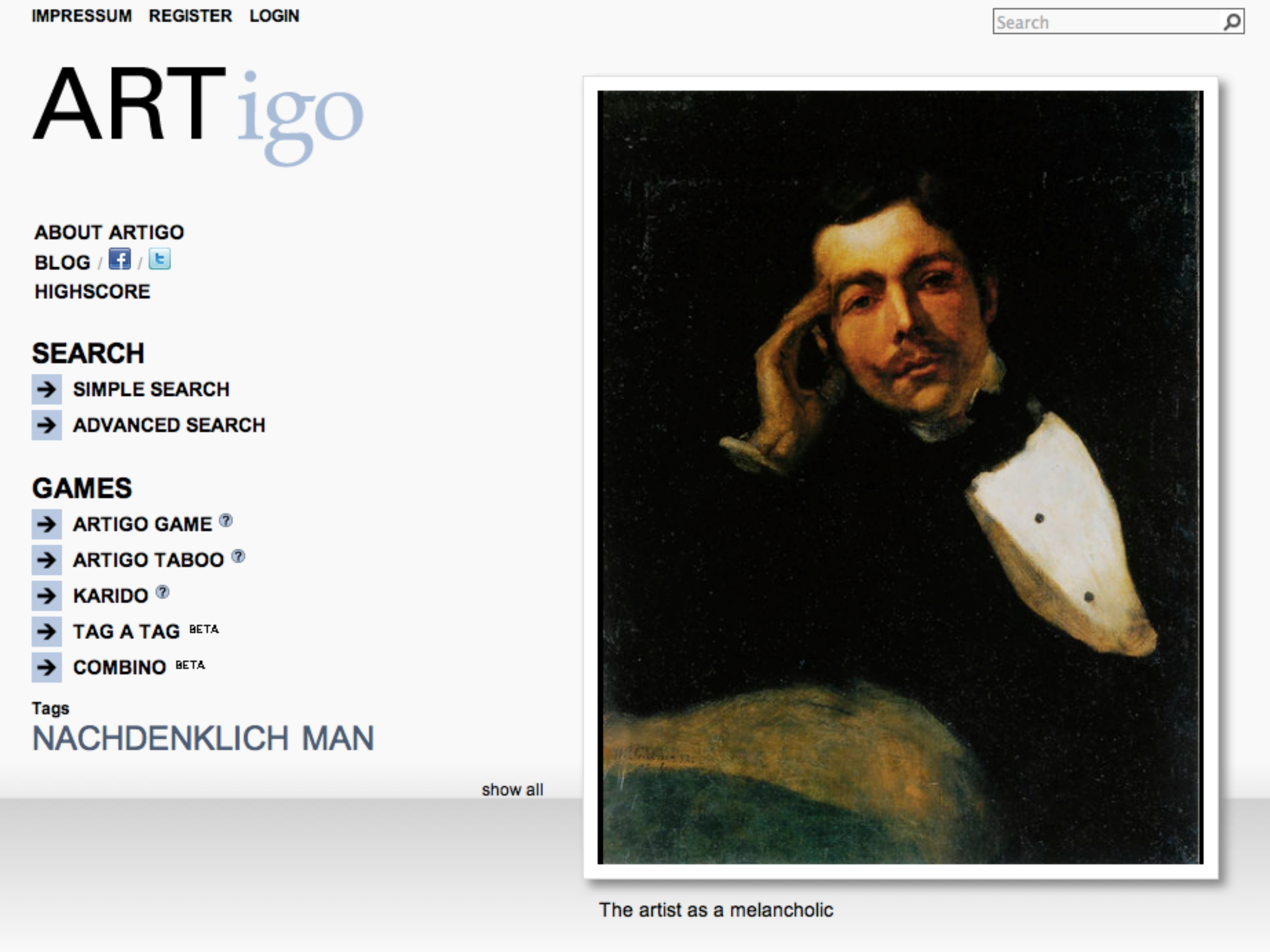
Player 2 guesses: purse
Success! Agreement on “purse”

Rules

- Partners agree on as many images as they can in 2.5 minutes
- Get points for every image, more if they agree on 15 images
- Players can also choose to pass or opt out on difficult image
- If a player clicks the pass button, a message is generated on their partner's screen; a pair cannot pass on an image until both have passed

Taboo words

- Players are not allowed to guess certain words
- Taboo words are the previous set of agreed upon words (up to 6)
- Initial labels for an image are often general ones (like “man” or “picture”)
- Taboo words generate more specific labels and guarantee that images get several different labels



Tags

NACHDENKLICH MAN

The artist as a melancholic

Game stats

- For 4 months in 2003, 13,630 people played the ESP game, generating 1,271,451 labels for 293,760 different images
- 3.89 labels/minute from one pair of players
- At this rate, 5,000 people playing the game 24 hours a day would label all images on Google (425,000,000 images) with 1 label each in 31 days
- In half a year, 6 words could be associated to every image in Google's index

ESP’s Purpose is Good Labels for Search

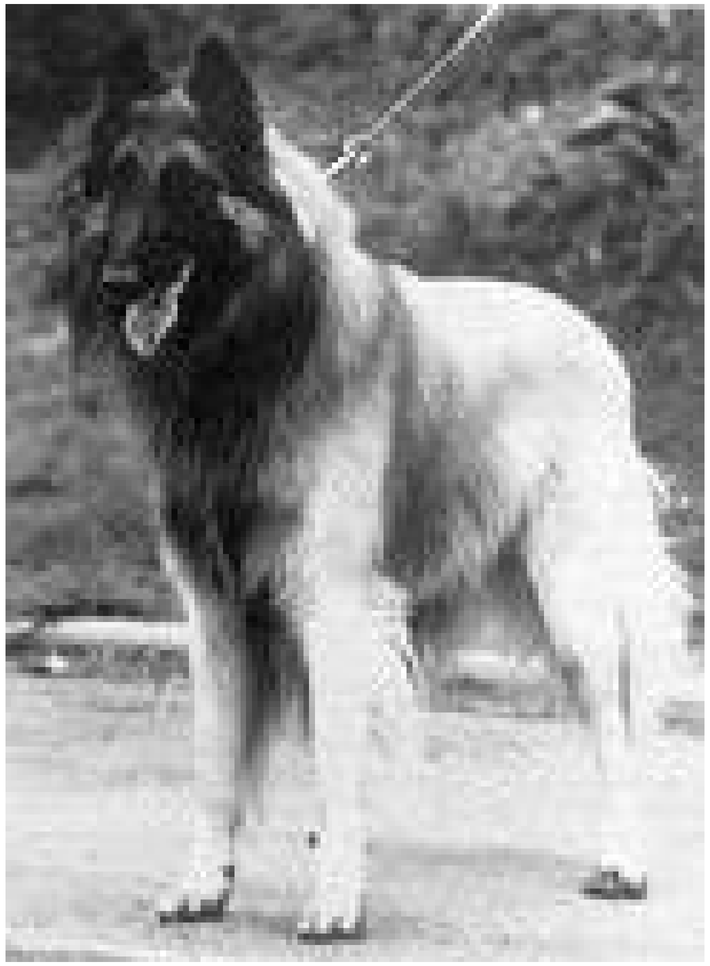
- Labels that players agree on tend to be “better”
- ESP game disregards the labels that players don’t agree on
- Can run the image through many pairs of players
- Establish a threshold for good labels (permissive = 1 pair agrees, strict = 40 agree)

Are they any good?

- Are these labels good for search?
- Is agreement indicative of better search labels?
- Is cheating a problem for the ESP game?
- How do they counter act it?

Original Evaluation

- Pick 20 images at random that have at least 5 labels
- 15 people the images and agreed on labels
- Do these have anything to do with the image?



Dog

Leash

German

Shepard

Standing

Canine

Ground Truth

Ability to produce labels of expert quality

- Measure the quality of labels on an authoritative set
- How good are labels from non-experts compared to labels from experts?

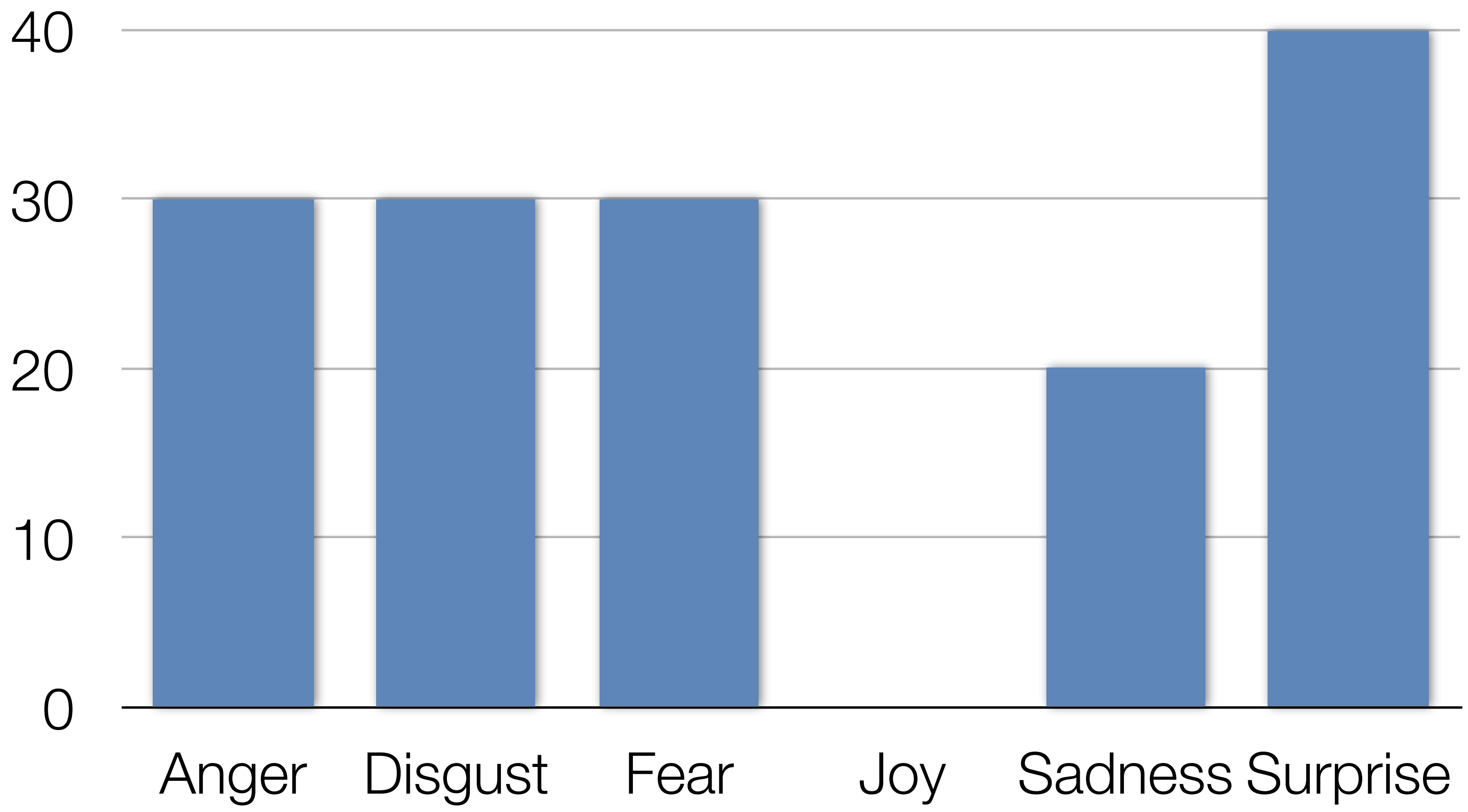
Fast and Cheap – But is it Good?

- Snow, O’Conner, Jurafsky and Ng (2008)
- Can Turkers be used to create data for natural language processing?
- Measured their performance in a series of well-designed experiments

Affect Recognition

- Turkers are shown short headlines
- Given numeric scores to 6 emotions

Outcry at N Korea `nuclear test’



Word Similarity

- Give a subjective numeric score about how similar a pair of words is
- 30 pairs of related words like {boy, lad} and unrelated words like {noon, string}
- Used in psycholinguistic experiments

sim(lad, boy) > sim(rooster, noon)

Word Sense Disambiguation

- Read a paragraph of text, and pick the best meaning for a word
- Robert E. Lyons III was appointed **president** and chief operating officer...
- 1) executive officer of a firm, corporation, or university
  2) head of a country (other than the U.S.)
  3) head of the U.S., President of the United States

Recognizing Textual Entailment

- Decide whether one sentence is implied by another
- Is “Oil prices drop” implied by “Crude Oil Prices Slump”?
- Is “Oil prices drop” implied by “The government announced that it plans to raise oil prices”?

Temporal Annotation

- Did a verb mentioned in a text happen before or after another verb?
- It just blew up in the air, and then we saw two fireballs go down to the water, and there was smoke coming up from that.
- Did *go down* happen before/after *coming up?*
- Did *blew up* happen before/after *saw?*

Experiments

- These data sets have existing labels that were created by experts
- We can therefore measure how well the workers’ labels correspond to experts
- What measurements should we use?

Correlation

Headline	Expert	Non-expert
Beware of peanut butter pathogens	37	15
Experts offer advice on salmonella	23	10
Indonesian with bird flu dies	45	39
Thousands tested after Russian H5N1 outbreak	71	80
Roots of autism more complex than thought	15	20
Largest ever autism study identifies two genetic culprits	12	22

Kendall tau rank correlation coefficient

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{1/2\ n^*(n-1)}$$

Headline	Expert	Non-expert
Beware of peanut butter pathogens	37	15
Experts offer advice on salmonella	23	10

Concordant > >

Kendall tau rank correlation coefficient

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{\text{1/2 n*(n-1)}}$$

Headline	Expert	Non-expert
Experts offer advice on salmonella	23	10
Largest ever autism study identifies two genetic culprits	12	22
discordant	>	<

Kendall tau rank correlation coefficient

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{1/2\ n^*(n-1)}$$

$$\tau = \frac{11 - 4}{15} = 0.46$$

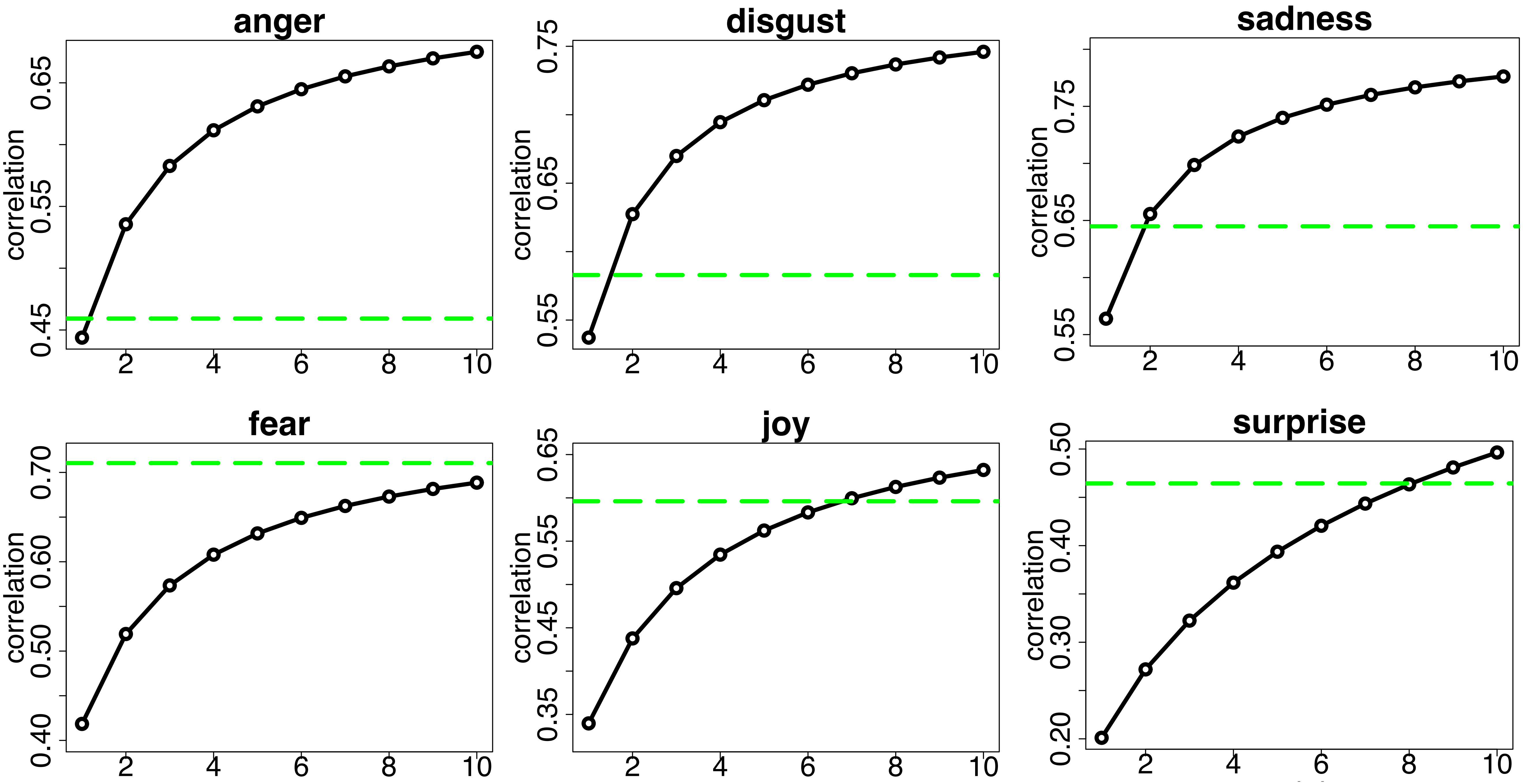
Experiments galore

- Calculate a correlation coefficient for each of the 5 data sets by comparing the non-expert values against expert values
- In most cases there were multiple annotations from different experts – this let’s us establish a topline
- Instead of taking a single Turker, combine multiple Turkers for each judgment

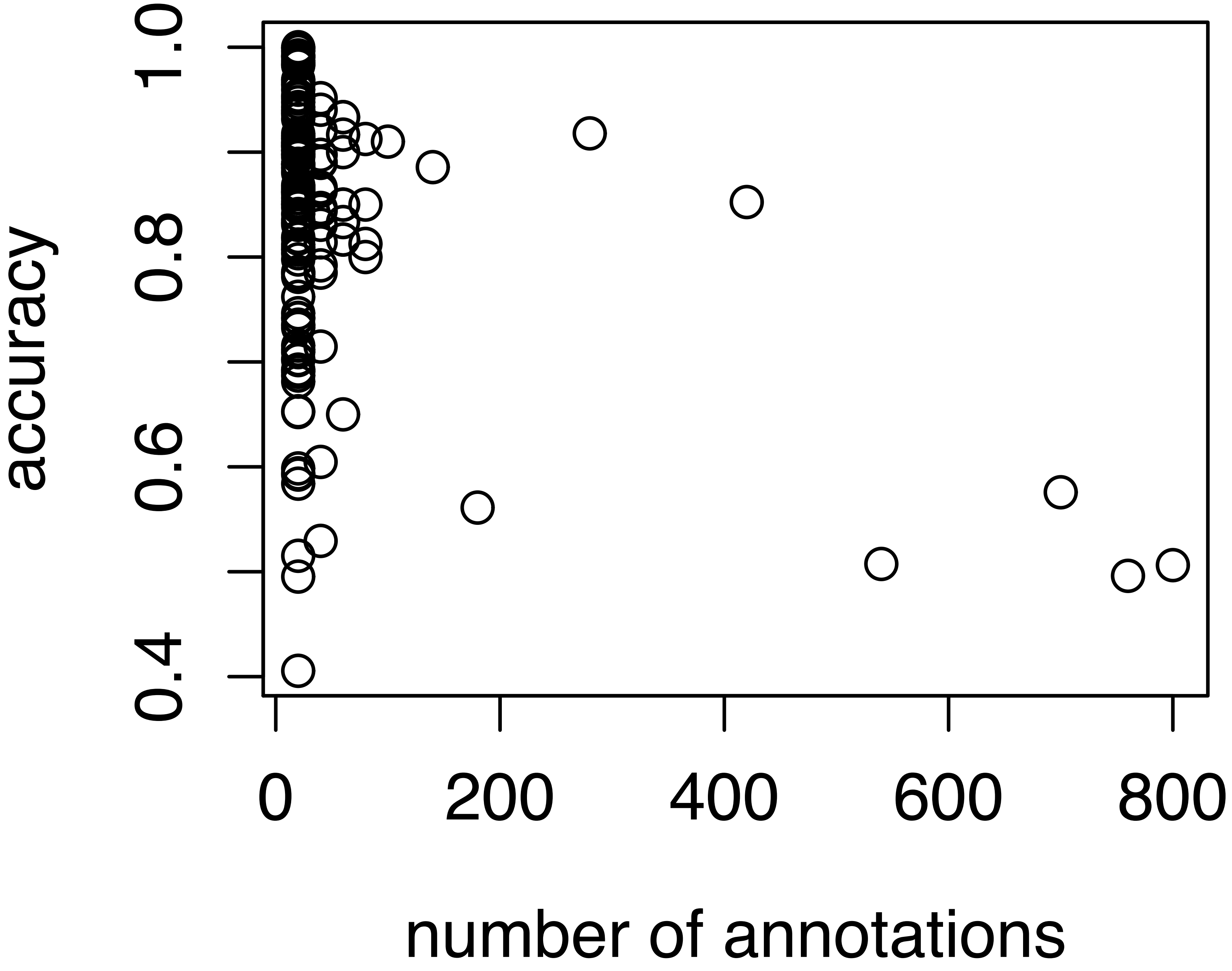
Sample sizes

Task	Labels
Affect Recognition	7000
Word Similarity	300
Recognizing Textual Entailment	8000
Word Sense Disambiguation	1770
Temporal Ordering	4620
Total	**21,690**

Agreement with experts increases as we add more Turkers



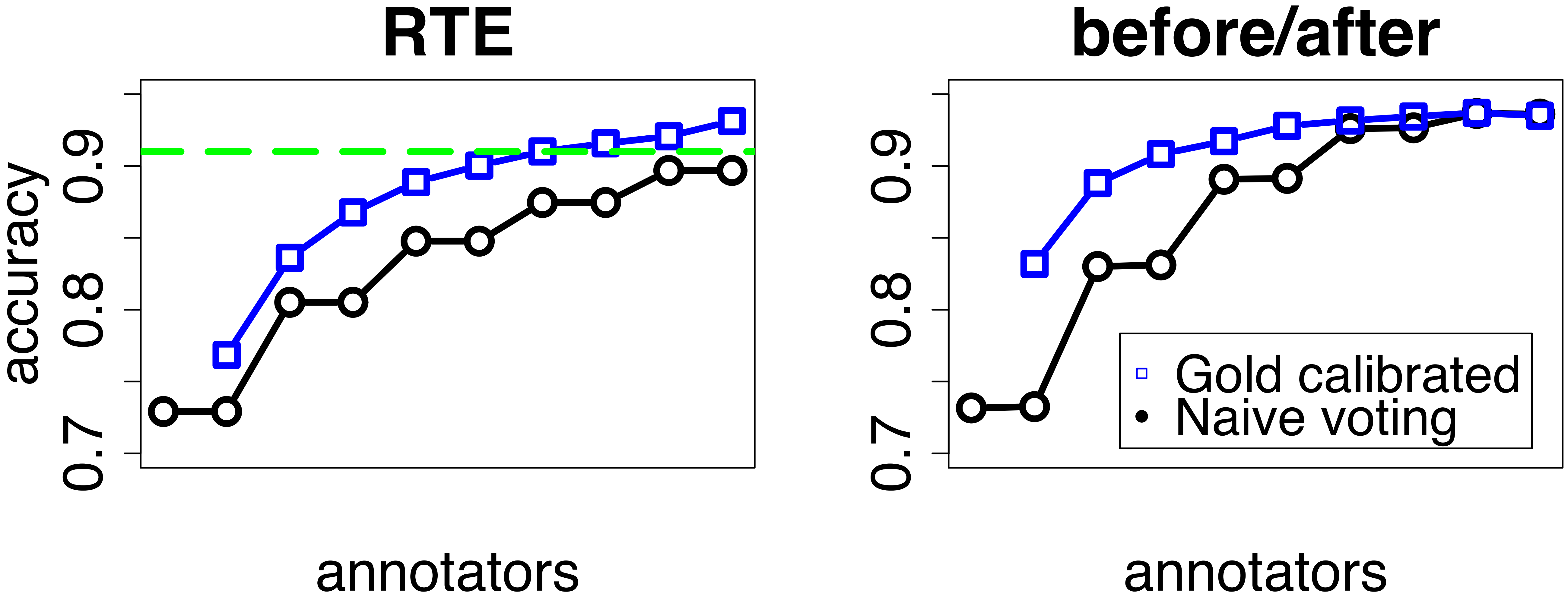
Accuracy of individual annotators



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



Limitations?

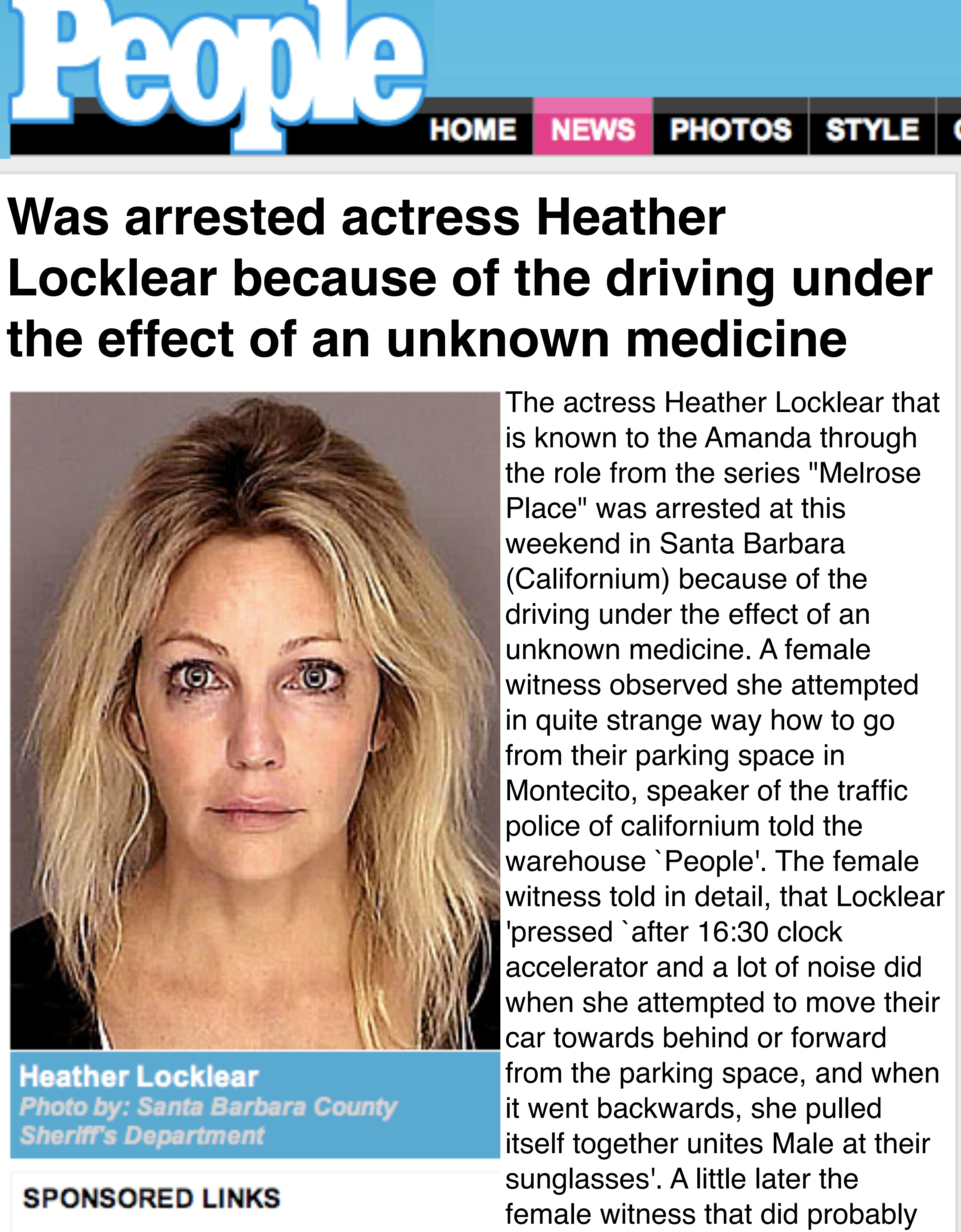
- Embedding gold standard data and weighted voting seems like the way to go
- What are its limitations?

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

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

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

SPONSORED LINKS

- Why was Heather Locklear arrested?

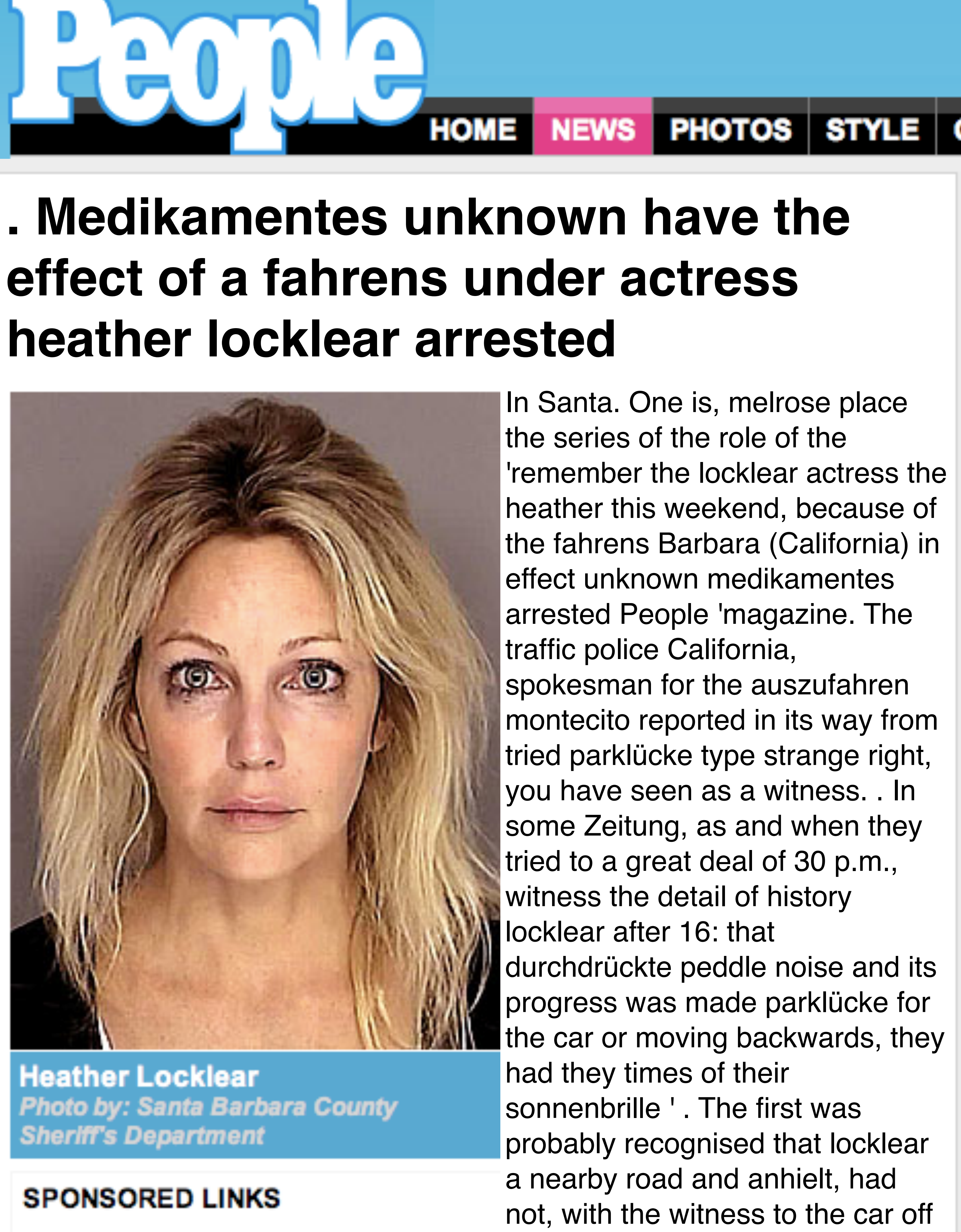
Driving while medicated

- 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

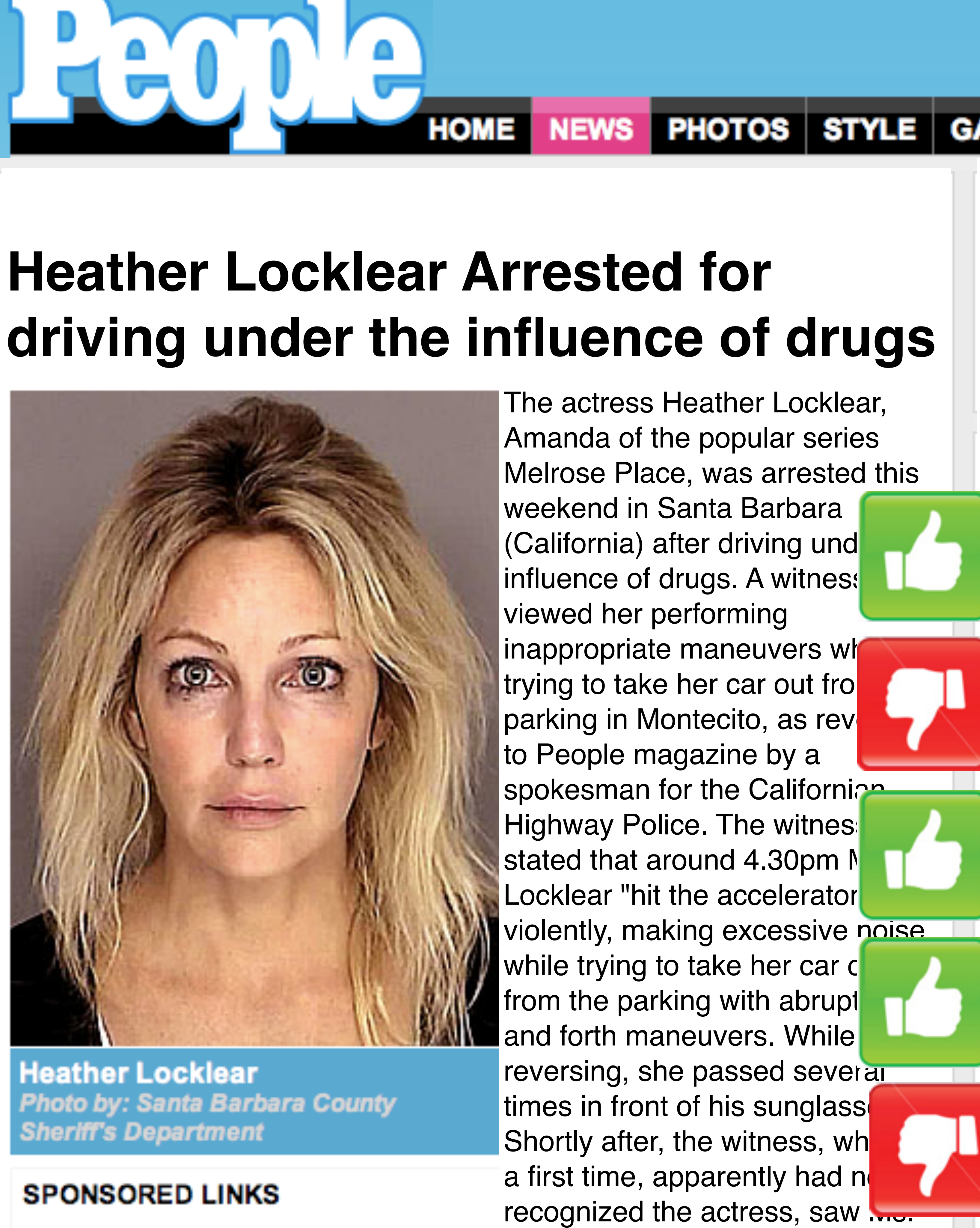
SPONSORED LINKS

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?



Heather Locklear Arrested for driving under the influence of drugs

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 witness stated that around 4.30pm M Locklear "hit the accelerator violently, making excessive noise while trying to take her car o 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.

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

- 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

The art of designing good HITs

Crowdsourcing works for tasks that are

- Natural and easy to explain to non-experts
- Decomposable into simpler tasks that can be joined together
- Parallelizable into small, quickly completed chunks
- Well-suited to quality control (some data has correct gold standard annotations)

Crowdsourcing works for tasks that are

- Robust to some amount of noise/errors (the downstream task is training a statistical model)
- Balanced and each task contains the same amount of work
    - Don’t have tons of work in one assignment but not another
    - Don’t ask Turkers to annotate something occurs in the data <<10% of the time

Guidelines for your own tasks

- Simple instructions are required
- If your task can’t be expressed in one paragraph + bullets, then it may need to be broken into simpler sub-tasks

Guidelines for your own tasks

- Quality control is paramount
    - Measuring redundancy doesn’t work if people answer incorrectly in systematic ways
    - Embed gold standard data as controls
- Qualification tests v. no qualification test
    - Reduce participation, but usually ensures higher quality

Guidelines for your own tasks

- Controls should be
  - Ones that have a unambiguous answer
  - Scorable automatically or with 2nd pass HIT
  - Not easily detectable