More Data Cleaning; Crowdsourcing

February 11, 2020
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Josh Levin, Diane Mutako, Sol Zitter

(Some slides stolen from Chris Callison-Burch and Kristy Milland. Thank you!)

Fill out the Brown Computer Science Survey you got in your email!

Only takes 5 min!

All multiple choice!

% The Percentage Project 2020

percentageproject.com

If you didn't receive the survey, email litofish@cs.brown.edu

Today

- Basic Bash Commands
- Crowdsourcing (as much as we get through)

Code-along!

```
cat data.txt | cut -f 2,4 | sort | uniq -c | sort -nr | head
```

Bash Scripting

https://cs.brown.edu/people/epavlick/articles.txt

1. ID
2. City
3. State
4. Date (YYYY-MM-DD)
5. Time
6. Victim Age
7. Shooter Age
8. Url
9. Title
10. Article Text

- **head –n {K} blah.txt** # first K lines
- **tail –n {K} blah.txt** # last K lines
- **shuf** # shuffle lines
- **wc blah.txt** # print number of bytes, chars, lines
- **wc –l blah.txt** # print number of lines
- **{cmd1} | {cmd2}** # run cmd1 on the output of cmd2
- **{cmd1} ; {cmd2}** # run cmd1 then cmd2
- **{cmd1} > {file}** # write output of cmd1 to file
- **cut –f {K} –d {D}** # split on delimiter D and select the Kth column
- **sort** # sort the lines by default ordering
- **sort –n** # sort numerically
- **sort –r** # reverse sort
- **uniq** # remove adjacent duplicate lines
- **uniq –c** # remove duplicates but count how many times each occurred
- **uniq –d** # print just the duplicated lines
- **grep** “{exp}” # print only lines matching exp
- **sed “s/{exp1}/{exp2}/g”** # replace exp1 with exp2

cat, less, head, tail

- what does this data even look like?

```
# first 10 lines of file
$ head articles.txt

# first line of file
$ head –n 1 articles.txt

# random 10 lines from file
$ cat articles.txt | shuf | head
```

wc

- how many articles are there

```
# how many bytes, words, and lines are
there?
$ wc articles.txt

# how many lines are there?
$ wc –l articles.txt
```

pipe (|), redirect (>)

```
$ head articles.txt | wc –l
      10

# write output to file called “tmp”
$ head articles.txt > tmp
$ wc –l tmp
      10 tmp

$ head articles.txt | wc –l > tmp
$ cat tmp
      10
```

Clicker Question!

Clicker Question!

What is city listed on line 817 of the file?

Clicker Question!

Which command will print just line 817 to the terminal?

(a) `$ head -n 817 articles.txt | tail -n 1`

(b) `$ cat articles.txt | head -n 817 | tail -n 1`

(c) `$ tail -n 817 articles.txt  | head -n 1`

cut

```
$ cat articles.txt | cut -f 1 | head -n
3
Antioch
Greeley
Bridgeport

$ cat articles.txt | cut -f 4 | cut -f 1
-d '-' | head -n 3
2016
2015
2014
```

sort, uniq

```
# print the lowest 3 values (includes duplicates)
$ cat articles.txt | cut -f 4 | cut -f 1 -d '-'| sort | head -n 3
1929
1932
1932

# print lowest three values (remove duplicates but count how many
occurrences of each
$ cat articles.txt | cut -f 4 | cut -f 1 -d '-'| sort | uniq -c |
head -n 3
   1 1929
   2 1932
   3 1942
```

Clicker Question!

Clicker Question!

Find the most frequent value for year

(a) `2015`

(b) `2016`

(c) `NA`

```
$ cat articles.txt | cut -f 4 | cut -f 1 -d '-'| sort | uniq -c |
sort -r | head -n 3
5091 2015
1821 2016
1784 NA
```

Find the most frequent value for year

(a) 2015

(b) 2016

(c) NA

sort, uniq

How many duplicated entries are there (using url as the uniq id)?

```
# total number of urls (lines)
$ cat articles.txt | cut -f 8 | wc -l
    9584

# number of unique urls
$ cat articles.txt | cut -f 8 | sort | uniq | wc -l
    7990

# number of duplicated urls
$ cat articles.txt | cut -f 8 | sort | uniq -d | wc -l
    981
```

regex (grep, sed, awk)

```
$ cat articles.txt | cut -f 2 | grep "NY" | head -n 5
NY
HOMINY
NYC
NY
NY

$ cat articles.txt | cut -f 2 | grep "^NY$" | head
NY
NY
NY
NY

$ cat articles.txt | cut -f 2 | grep "^NY[.]*" | head
NY
NYC
NY
NY
NY
```

regex (grep, sed, awk)

```
# mask numbers to look at formats
$ cat articles.txt | cut -f 4 | sed "s/[0-9]/#/g" | head -n 3
####-##-##
####-##-##
####-##-##

# remove the leading abbreviations
$ cat articles.txt | cut -f 3 | sed "s/[A-Z][A-Z] - //g" |
grep -v Unclear | head -n 3
Minnesota
North Carolina
Michigan

# lowercase everything
$ cat articles.txt | cut -f 3 | sed "s/.*/\L&/g"

# replace all non-numeric characters with blanks
$ cat articles.txt | cut -f 6 | sed "s/[^0-9]//g" | head
```

Clicker Question!

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

How many unique values are there for "city" in our data?

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

How many unique values are there for "city" in our data?

(a)

```
$ cat articles.txt | cut -f 2 | uniq | wc -l
```

(b)

```
$ cat articles.txt | sort | uniq | cut -f 2 | wc -l
```

(c)

```
$ cat articles.txt | cut -f 2  |sort | uniq | wc -l
```

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

How many unique values are there for "city" in our data?

(a) `$ cat articles.txt | cut -f 2 | uniq | wc -l`

(b) `$ cat articles.txt | sort | uniq | cut -f 2 | wc -l`

[✓ **(c)**] `$ cat articles.txt | cut -f 2  |sort | uniq | wc -l`

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

Find the 10 titles that appear with the largest number of unique urls.

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

Find the 10 titles that appear with the largest number of unique urls.

(a)
```
$ cat articles.txt | cut -f 9 | sort | uniq -c |
sort -nr | head
```

(b)
```
$ cat articles.txt | cut -f 8,9 | sort | uniq |
cut -f 2 | sort | uniq -c | sort -nr | head
```

(c)
```
$ cat articles.txt | sort | uniq -f 9 | sort -nr |
head
```

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

Find the 10 titles that appear with the largest number of unique urls.

(a)

```
$ cat articles.txt | cut -f 9 | sort | uniq -c |
sort -nr | head
```

[✓ (b)]

```
$ cat articles.txt | cut -f 8,9 | sort | uniq |
cut -f 2 | sort | uniq -c | sort -nr | head
```

(c)

```
$ cat articles.txt | sort | uniq -f 9 | sort -nr |
head
```

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

How many different cities are there for the article titled "Suspect arrested in Memphis cop killing"

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

How many different cities are there for the article titled "Suspect arrested in Memphis cop killing"

(a)

```
$ cat articles.txt | cut -f 2 | grep "Suspect
arrested in Memphis cop killing" | sort | uniq -c
```

(b)

```
$ cat articles.txt | grep "Suspect arrested in
Memphis cop killing" | cut -f 2 | sort | uniq -c
```

(c)

```
$ cat articles.txt | sort | grep "Suspect arrested
in Memphis cop killing" | cut -f 2 | uniq -c
```

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

How many different cities are there for the article titled "Suspect arrested in Memphis cop killing"

(a)

```
$ cat articles.txt | cut -f 2 | grep "Suspect
arrested in Memphis cop killing" | sort | uniq -c
```

[✓ (b)]

```
$ cat articles.txt | grep "Suspect arrested in
Memphis cop killing" | cut -f 2 | sort | uniq -c
```

(c)

```
$ cat articles.txt | sort | grep "Suspect arrested
in Memphis cop killing" | cut -f 2 | uniq -c
```

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

Print out all the victim ages that contain no numeric characters.

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

Print out all the victim ages that contain no numeric characters.

(a)

```
$ cat articles.txt | cut -f 6 | grep -e "^[^0-9]*$"
| sort | uniq -c | sort -nr | head
```

(b)

```
$ cat articles.txt | cut -f 6 | grep -e "^[0-9]*$"
| sort | uniq -c | sort -nr | head
```

(c)

```
$ cat articles.txt | cut -f 6 | grep -e "^[^0-9]*"
| sort | uniq -c | sort -nr | head
```

Clicker Question!

Hint: Columns are ID=1, City=2, State=3, Date=4, Time=5, Victim Age=6, Shooter Age=7, Url=8, Title=9, Text=10

Print out all the victim ages that contain no numeric characters.

[✓ (a)]

```
$ cat articles.txt | cut -f 6 | grep -e "^[^0-9]*$"
| sort | uniq -c | sort -nr | head
```

(b)

```
$ cat articles.txt | cut -f 6 | grep -e "^[0-9]*$"
| sort | uniq -c | sort -nr | head
```

(c)

```
$ cat articles.txt | cut -f 6 | grep -e "^[^0-9]*"
| sort | uniq -c | sort -nr | head
```

Being all fancy…

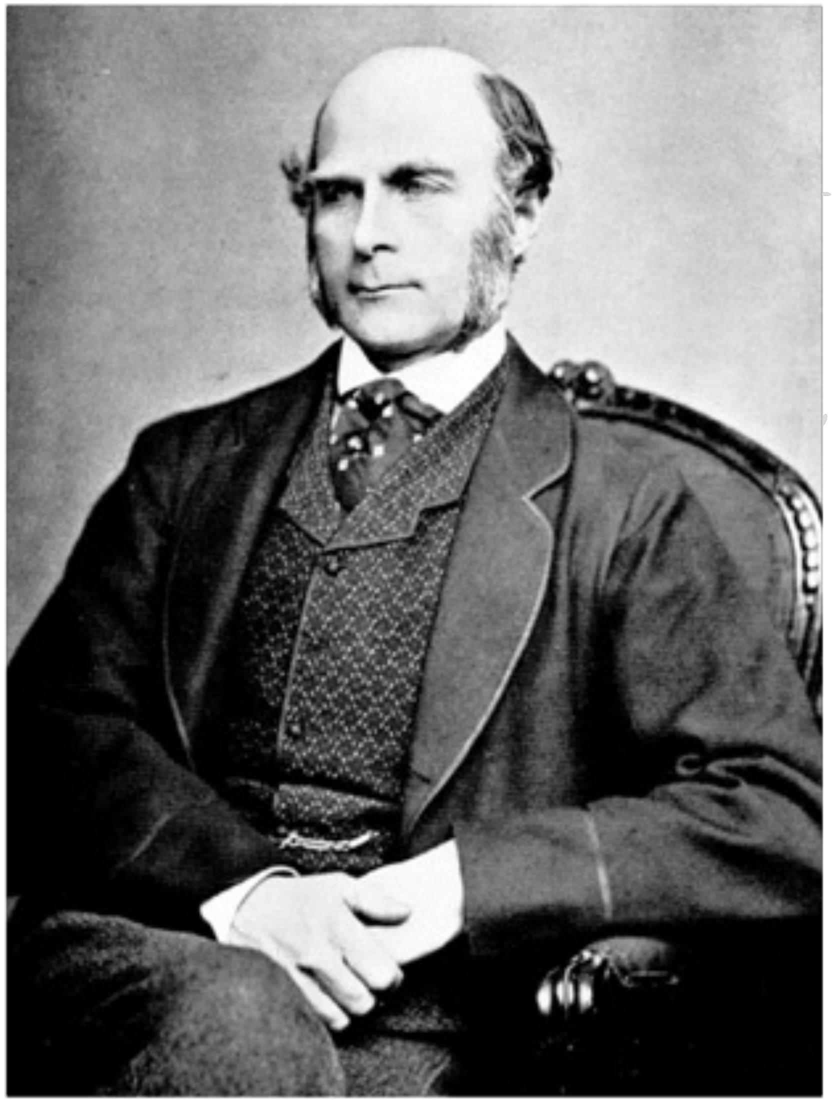
```
# plot a histogram of all ages
cat articles.txt | cut -f 6 | sed "s/[^0-9]//
g" | grep -v "^$" | pythonw -c "import sys,
matplotlib.pyplot as plt; plt.hist([int(i)
for i in sys.stdin]); plt.show()”

# plot a histogram of all ages, removing
outliers
cat articles.txt | cut -f 6 | sed "s/[^0-9]//
g" | grep -v "^$" | pythonw -c "import sys,
matplotlib.pyplot as plt;
plt.hist([min(int(i), 100) for i in
sys.stdin]); plt.show()"
```

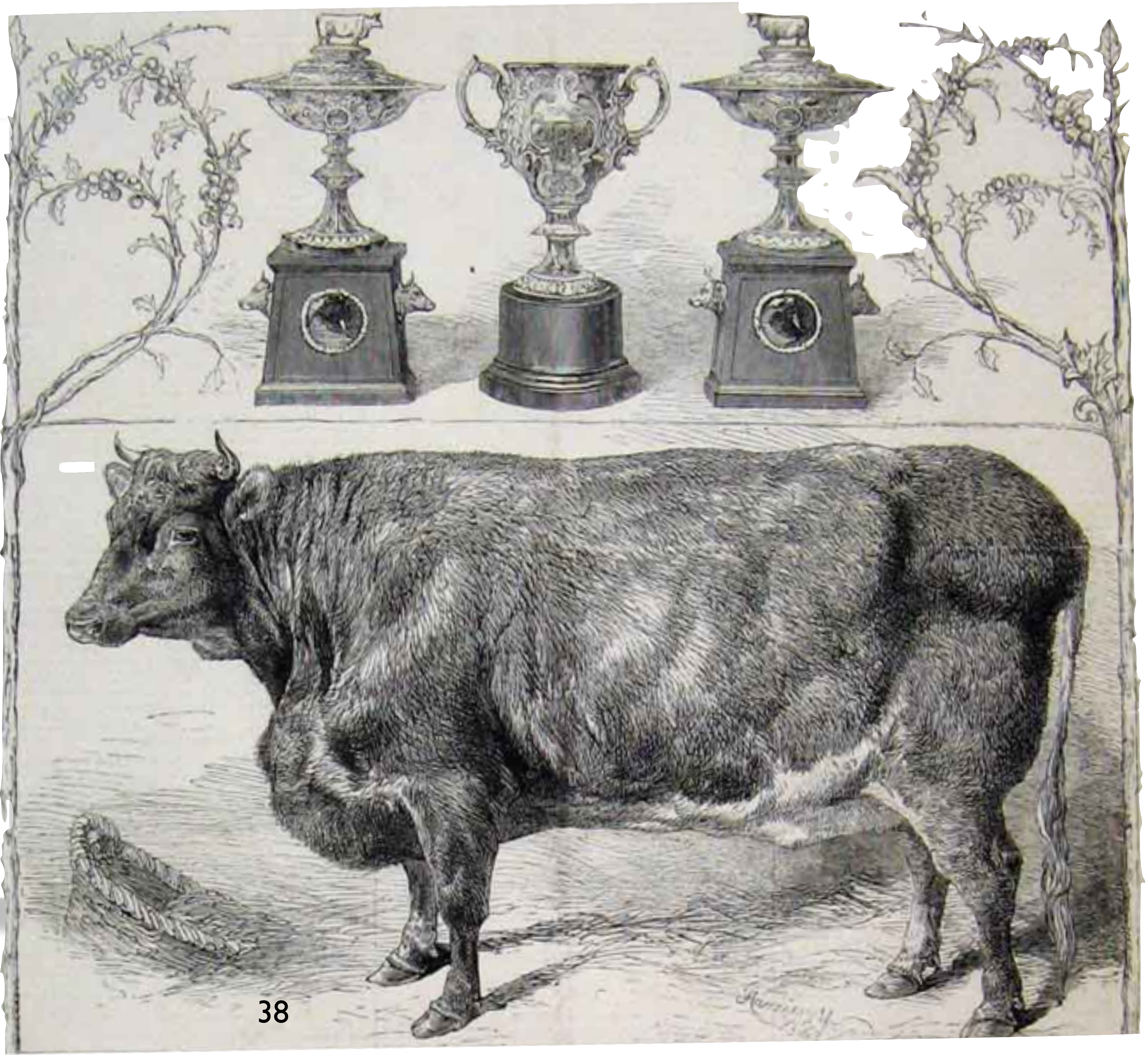
Crowdsourcing!

reddit
twitter
UBER
coursera
amazon mechanical turk™
Artificial Artificial Intelligence
Quora
stack overflow
KICKSTARTER
craigslist

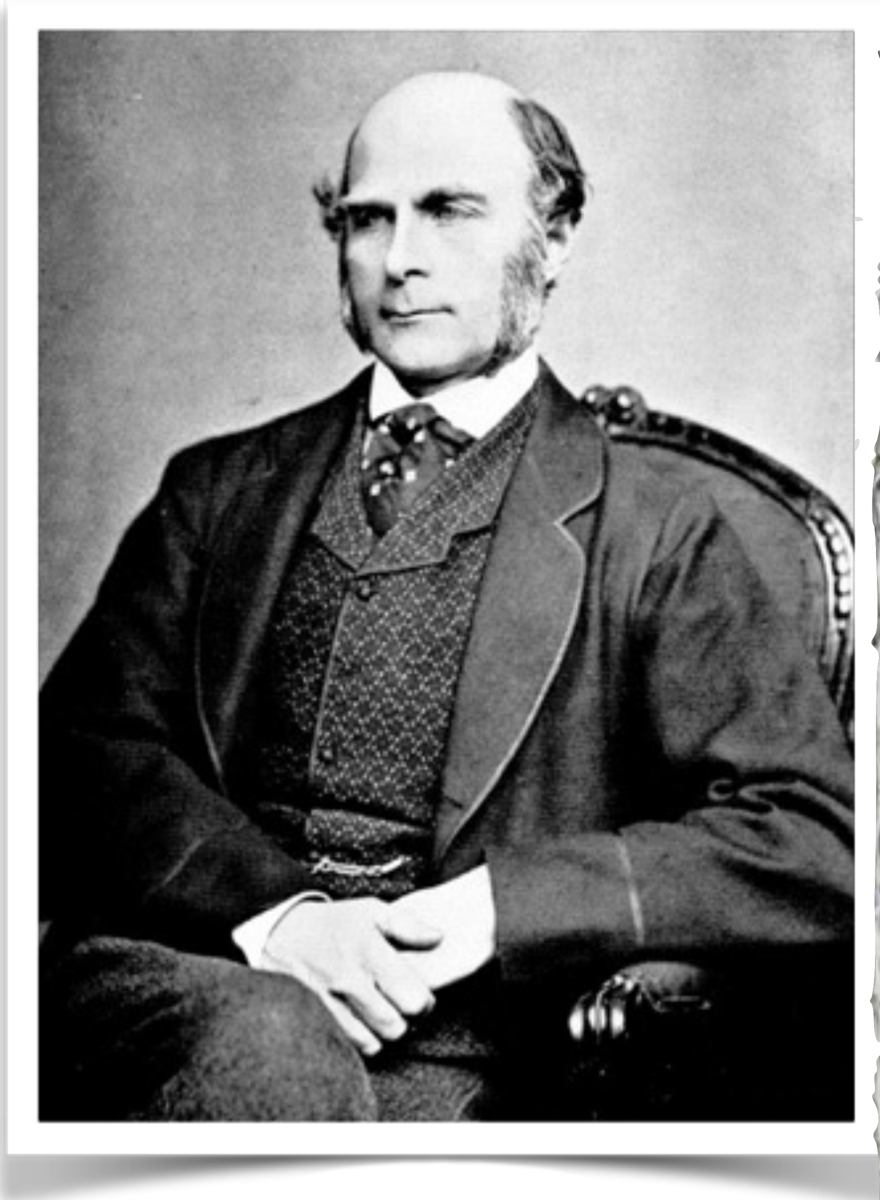
Wisdom of the Crowd

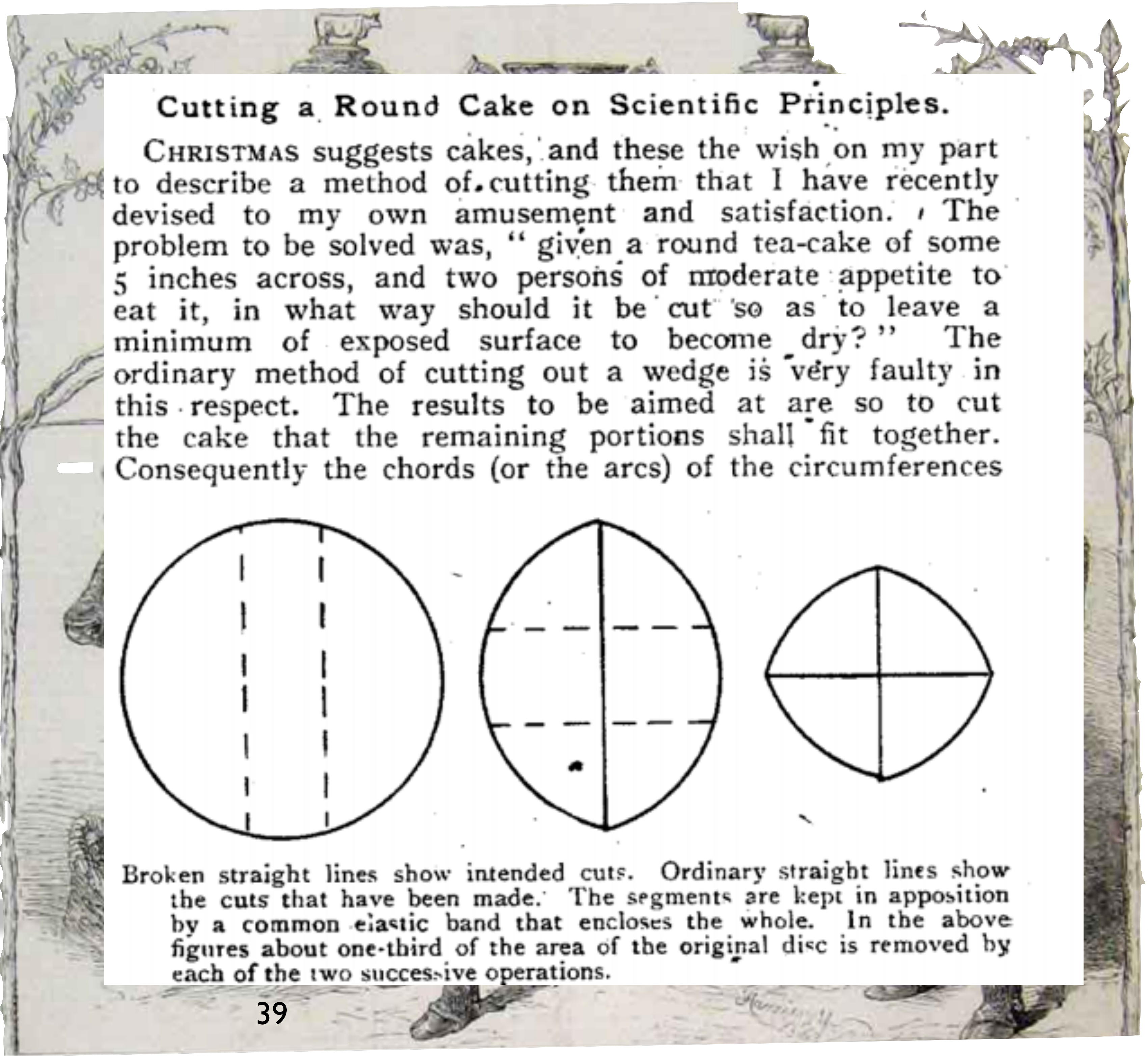
Wisdom of the Crowd





Wisdom of the Crowd





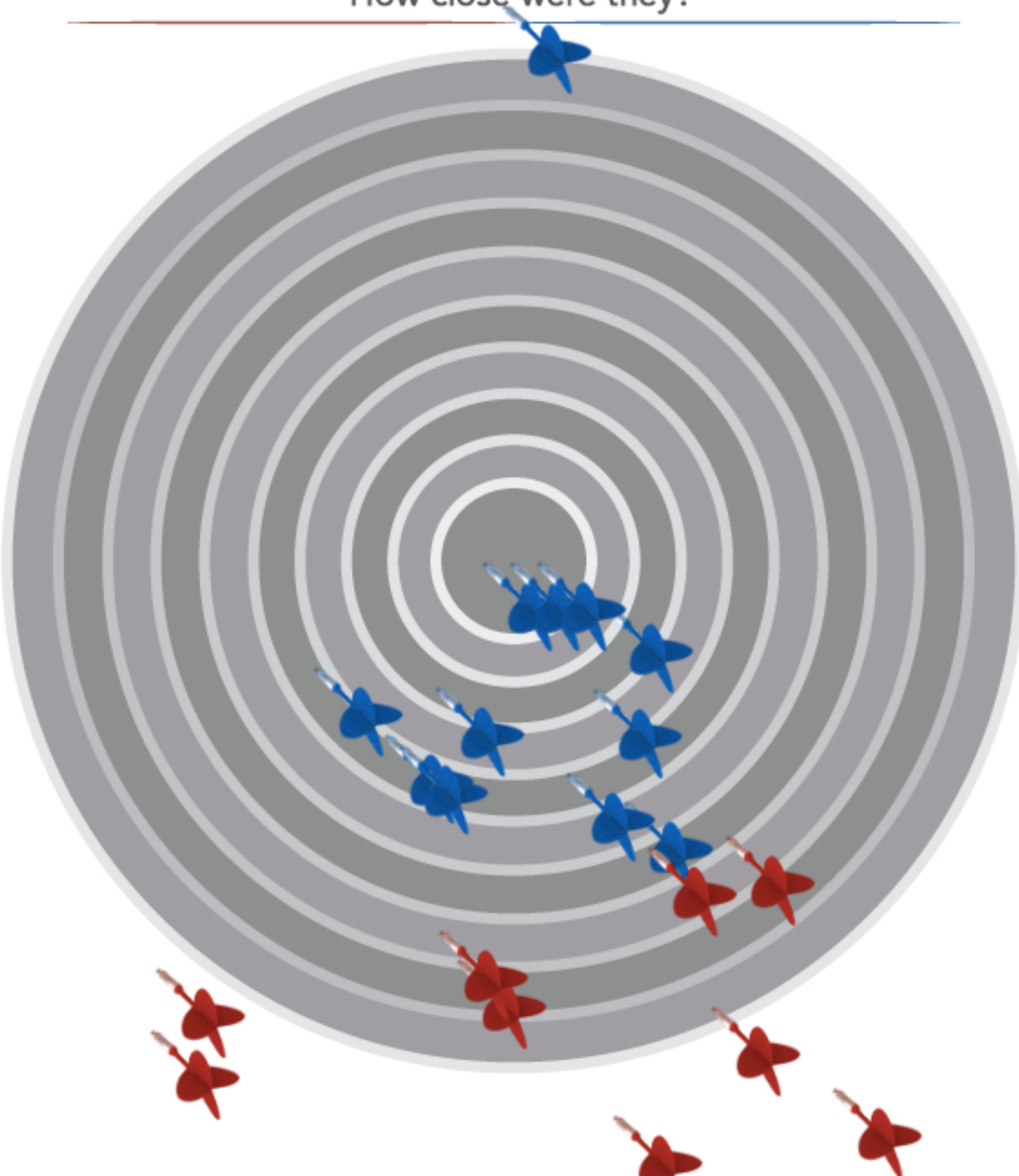
Cutting a Round Cake on Scientific Principles.

CHRISTMAS suggests cakes, and these the wish on my part to describe a method of cutting them that I have recently devised to my own amusement and satisfaction. The problem to be solved was, "given a round tea-cake of some 5 inches across, and two persons of moderate appetite to eat it, in what way should it be cut so as to leave a minimum of exposed surface to become dry?" The ordinary method of cutting out a wedge is very faulty in this respect. The results to be aimed at are so to cut the cake that the remaining portions shall fit together. Consequently the chords (or the arcs) of the circumferences

Broken straight lines show intended cuts. Ordinary straight lines show the cuts that have been made. The segments are kept in apposition by a common elastic band that encloses the whole. In the above figures about one-third of the area of the original disc is removed by each of the two successive operations.

PUNDARTS

How close were they?

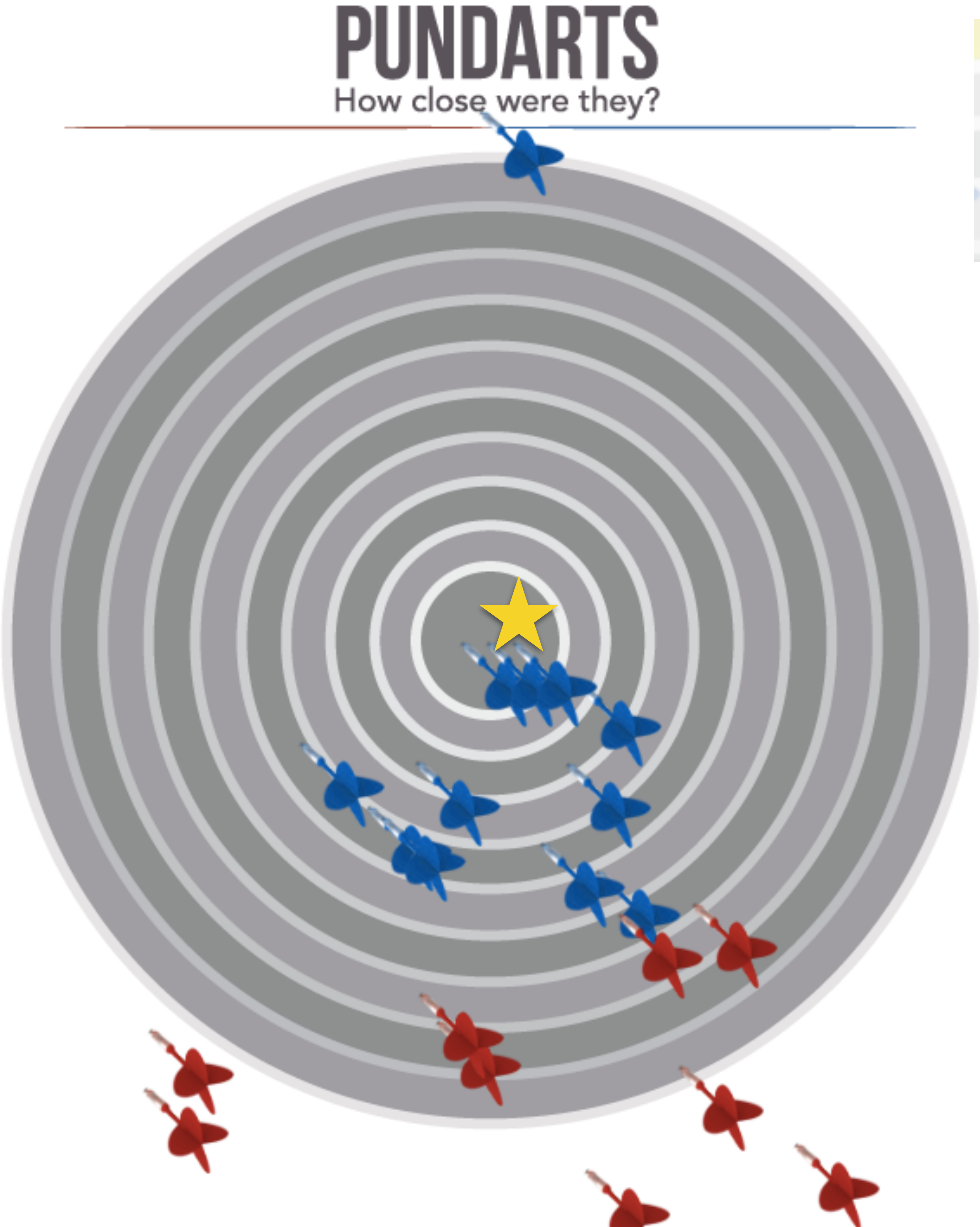


Nate Silver
The New York Times
Obama: 332
Romney: 206
Difference: 0

Donna Brazile
Vice Chairwoman of the Democratic National Committee
Obama: 313
Romney: 225
Difference: -19

Ann Coulter
Fox News
Obama: 265
Romney: 273
Difference: -67

Karl Rove
Obama: 259
Romney: 279
Difference: -73



Nate Silver
The New York Times
Obama: 332
Romney: 206
Difference: 0

amazonmechanical turk
beta
Artificial Artificial Intelligence ★

Obama: 335, Romney: 189
Error: 3, Unallocated: 14

Donna Brazile
Vice Chairwoman of the Democratic National Committee
Obama: 313
Romney: 225
Difference: -19

Ann Coulter
Fox News
Obama: 265
Romney: 273
Difference: -67

Karl Rove
Obama: 259
Romney: 279
Difference: -73

reddit

twitter

UBER

coursera

amazon mechanicalturk™

Artificial Artificial Intelligence

Quora

stackoverflow

KICKSTARTER

craigslist

Motivation

Why do people contribute?

What do workers gain from contributing?

Motivation

Why do people contribute?

What do workers gain f
contributing?

Quality Control

How to we make sure the contributions are good?

How to we identify and incentivize good work?

Motivation

Why do people contribute?

What do workers gain f

contributing?

Quality Control

How to we make sure the

tributions are good?

v to we identify and

entivize good work?

Aggregation

How to we combine many small or distributed contributions into one final answer/result/product?

Motivation

Why do people contribute?

What do workers gain f

contributing?

Quality Control

How to we make sure the

tributions are good?

Aggregation

How to we combine m

small or distributed

contributions into one

answer/result/produc

Skill

Do workers need specialized skills?

How to we find or train workers to match the skill sets we need?

Motivation

Why do people contribute?

What do workers gain f

contributing?

Quality Control

How to we make sure the

tributions are good?

Aggregation

How to we combine m

c

Skill

workers need

cialized skills?

o we find or train

s to match the skill

ts we need?

Decomposition

How is the task decomposed into subtasks?

How many subtasks are required to get from input to output?

Motivation

Motivation

airbnb

START HOSTING AND EARN A $250 CASH BONUS

SURGE PRICING ×

Demand is off the charts! Fares have increased to get more Ubers on the road.

2.2X
THE NORMAL FARE

$7.60 MINIMUM FARE

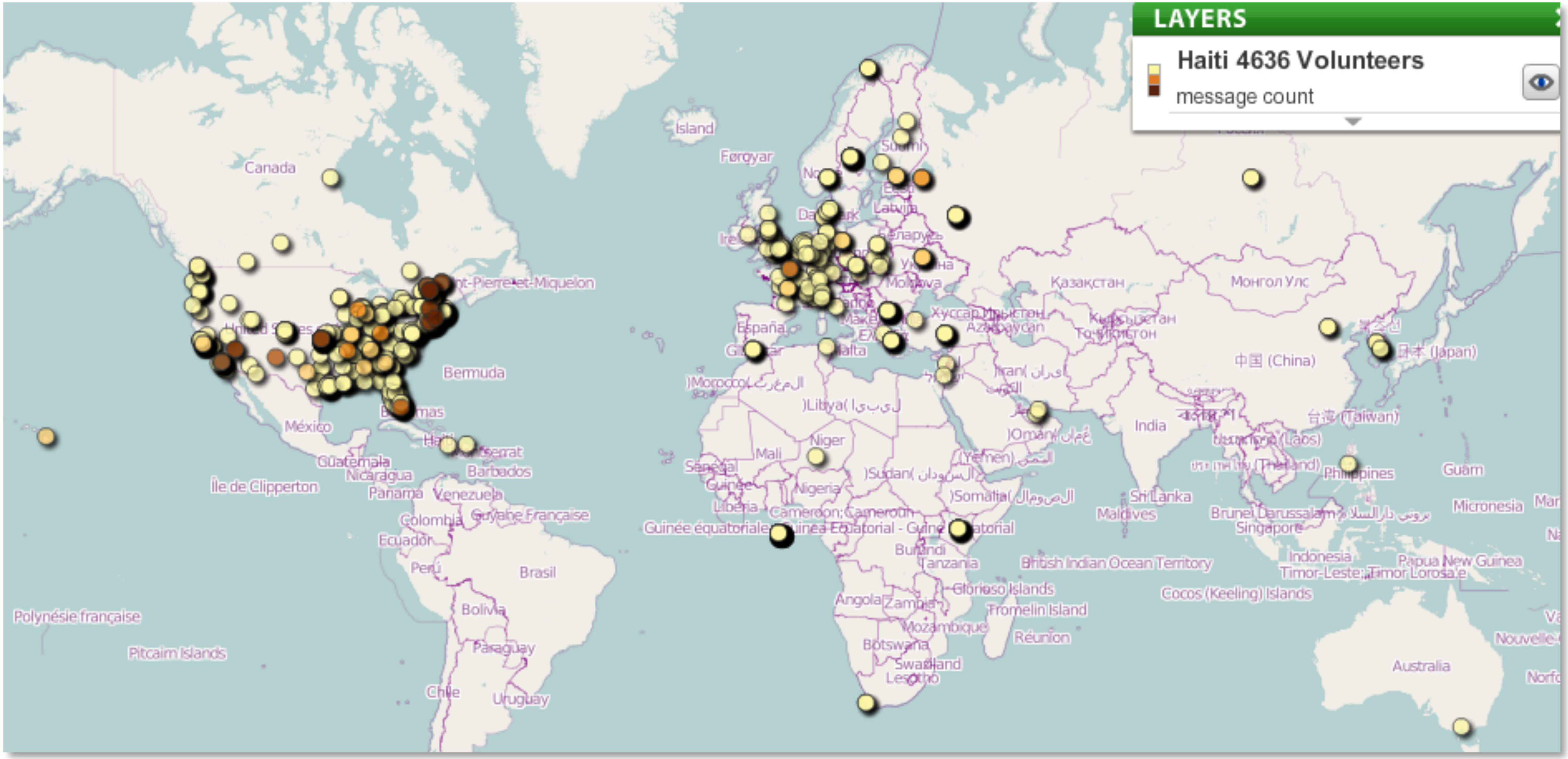
$0.35 / MIN $2.09 / MILE

Total Earnings (What's this?)

Rewards You Have Earned		Value
Approved HITs		$1.16
Bonuses		$0.81
Total Earnings	Show earnings details	$1.97

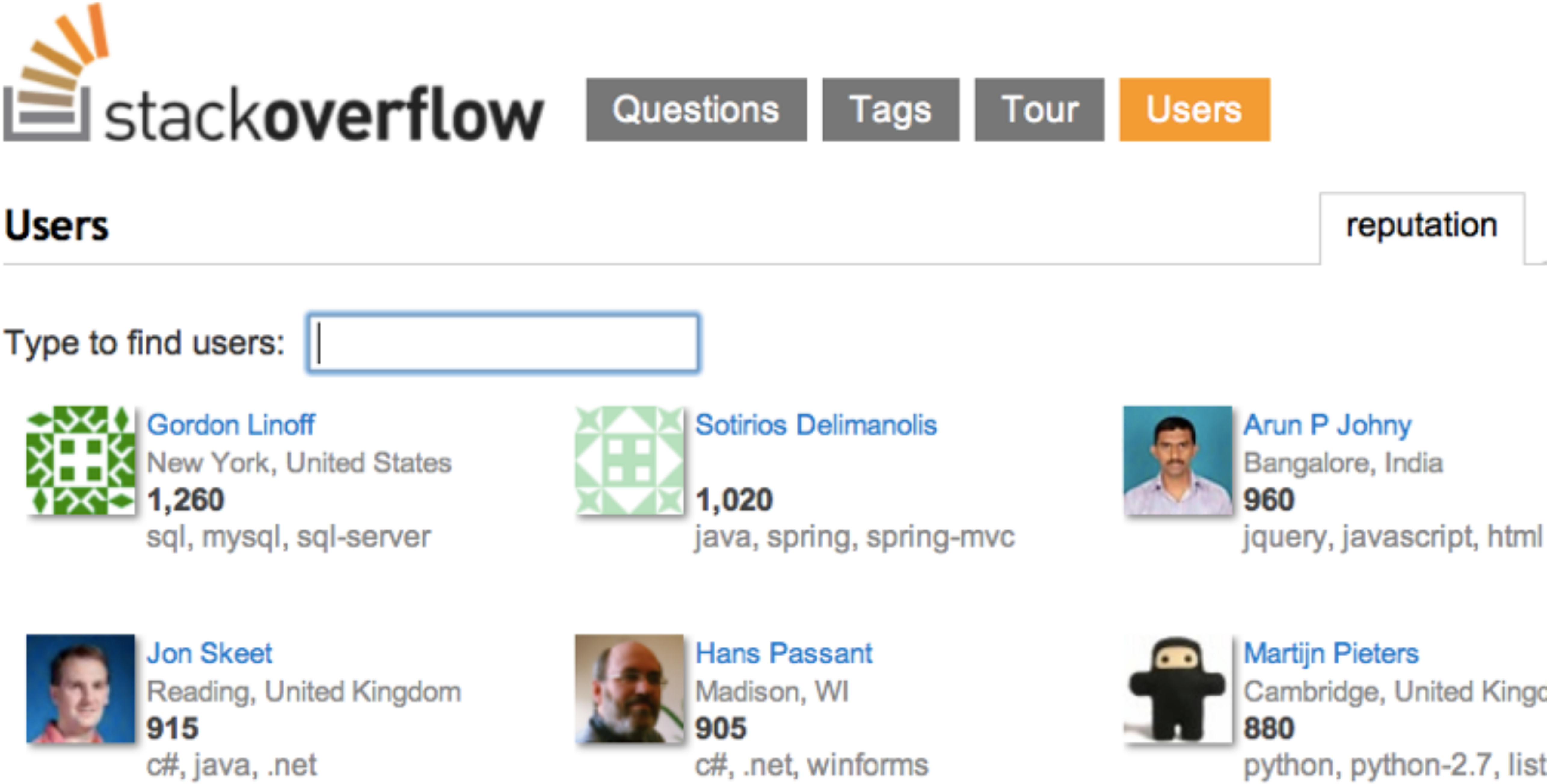
Pay

Motivation



Altruism

Motivation



Users

reputation

Type to find users:

Gordon Linoff
New York, United States
1,260
sql, mysql, sql-server

Sotirios Delimanolis

1,020
java, spring, spring-mvc

Arun P Johny
Bangalore, India
960
jquery, javascript, html

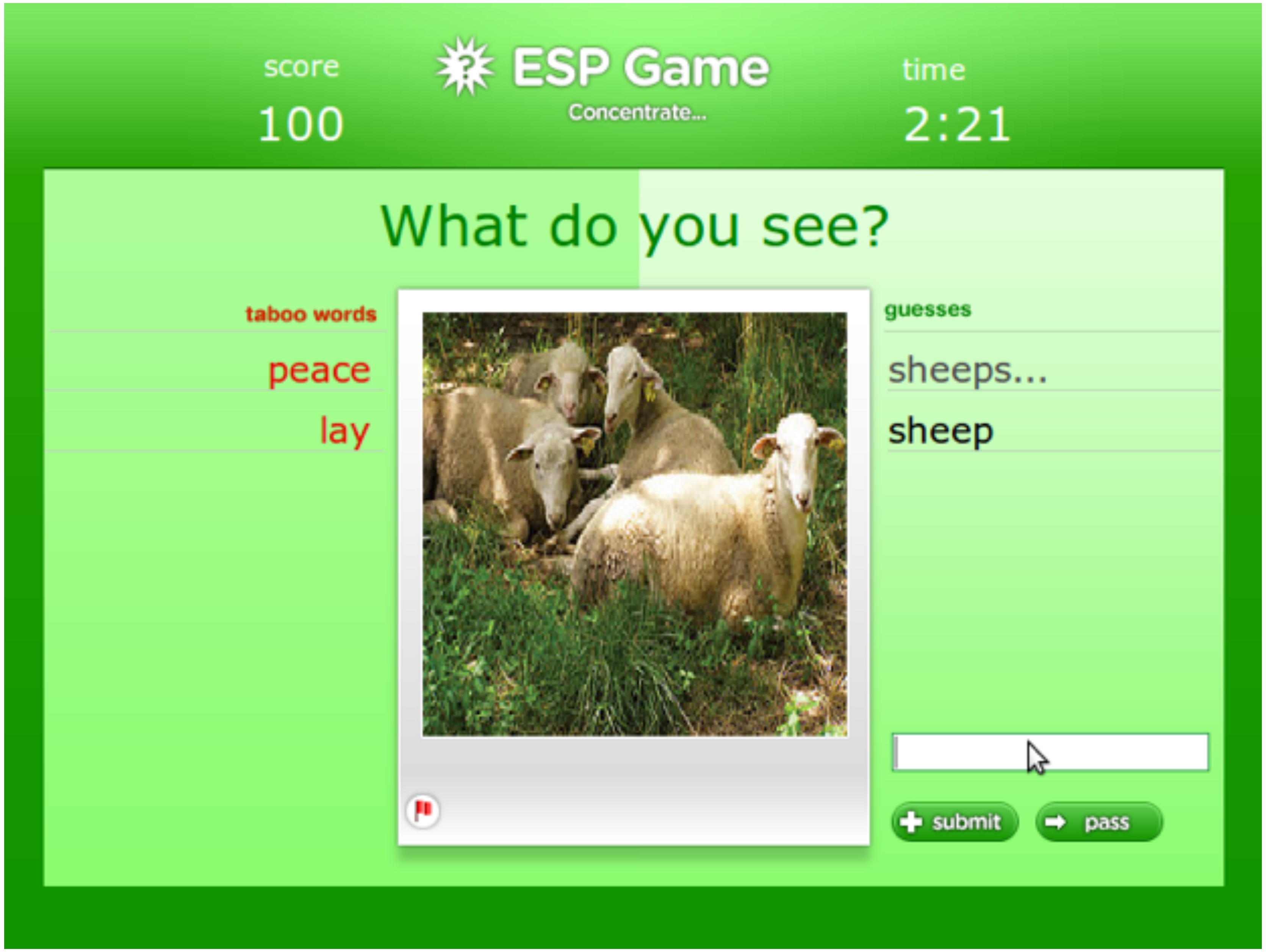
Jon Skeet
Reading, United Kingdom
915
c#, java, .net

Hans Passant
Madison, WI
905
c#, .net, winforms

Martijn Pieters
Cambridge, United Kingc
880
python, python-2.7, list

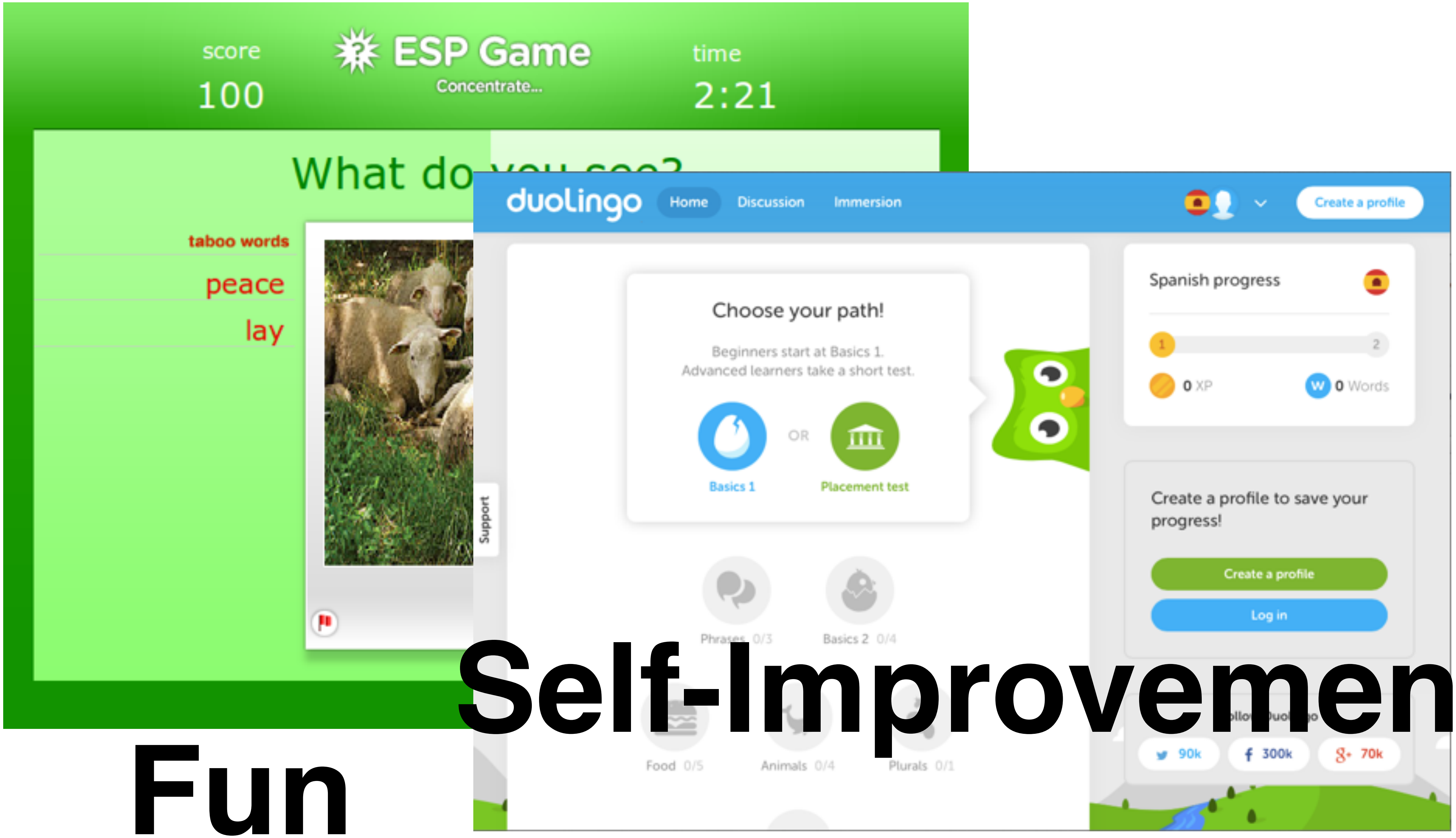
Reputation

Motivation



Fun

Motivation



Choose your path!

Beginners start at Basics 1.
Advanced learners take a short test.

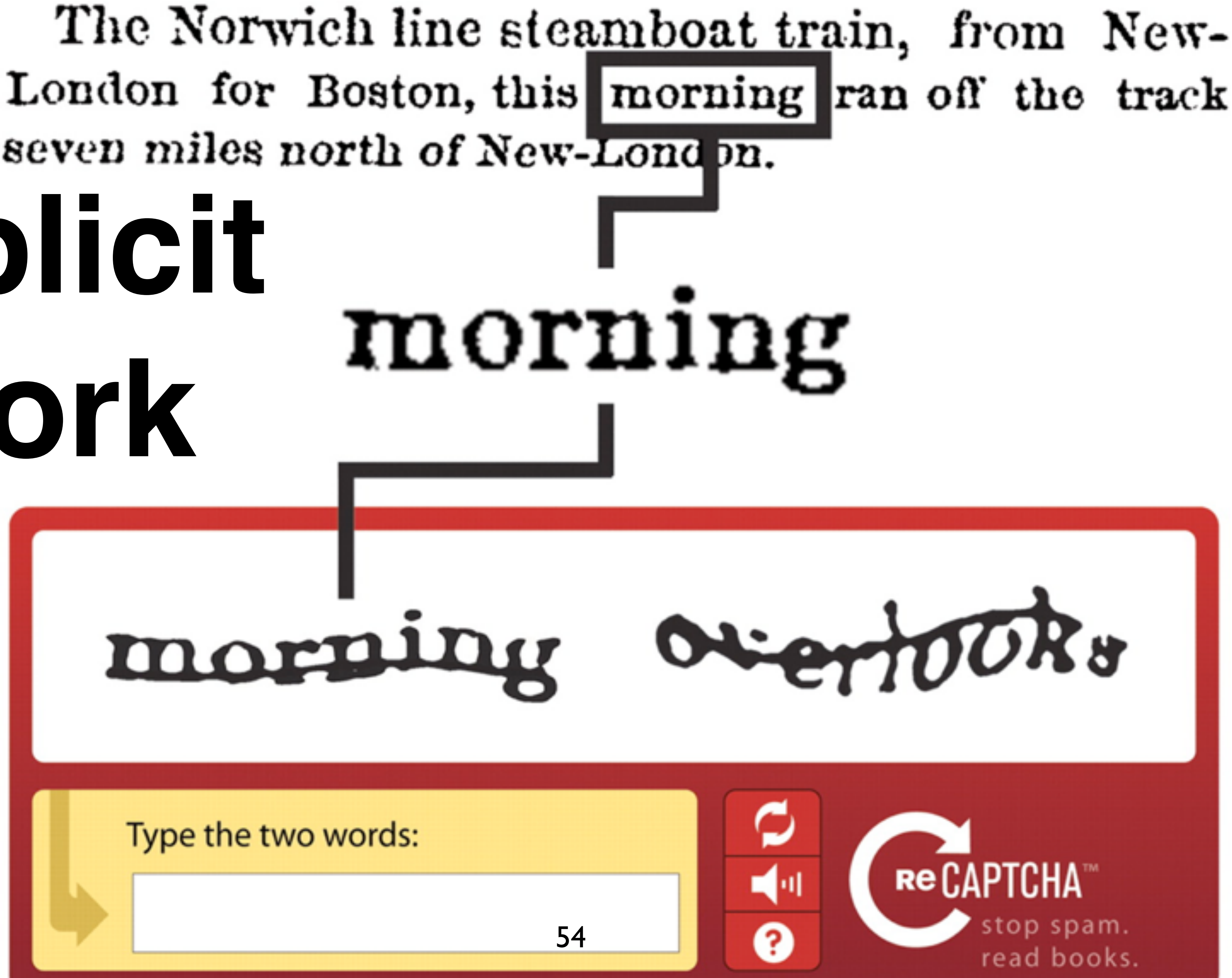
Create a profile to save your progress!

Fun

Self-Improvement

Motivation

Implicit Work

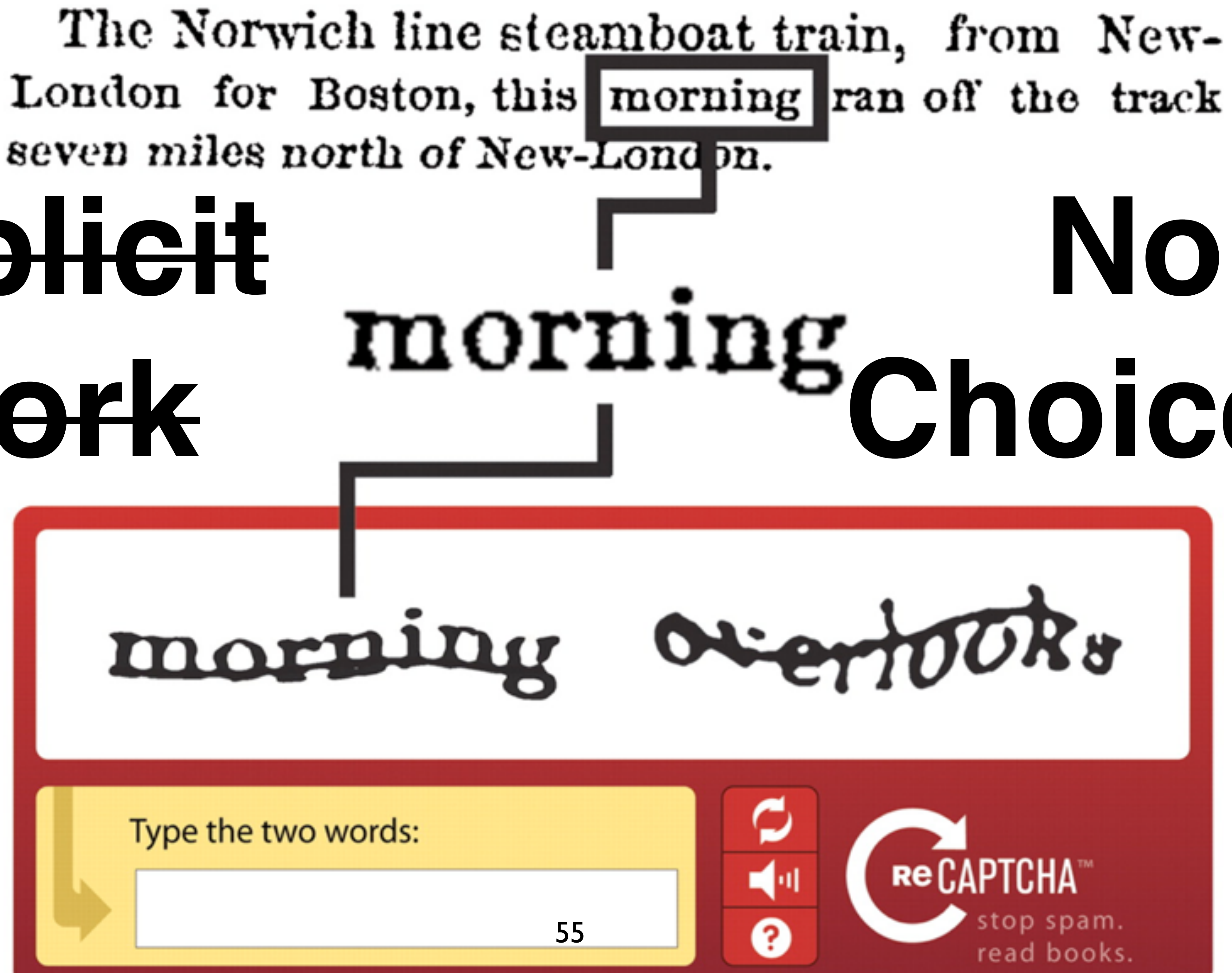


The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

Motivation

~~**Implicit Work**~~

No Choice :)

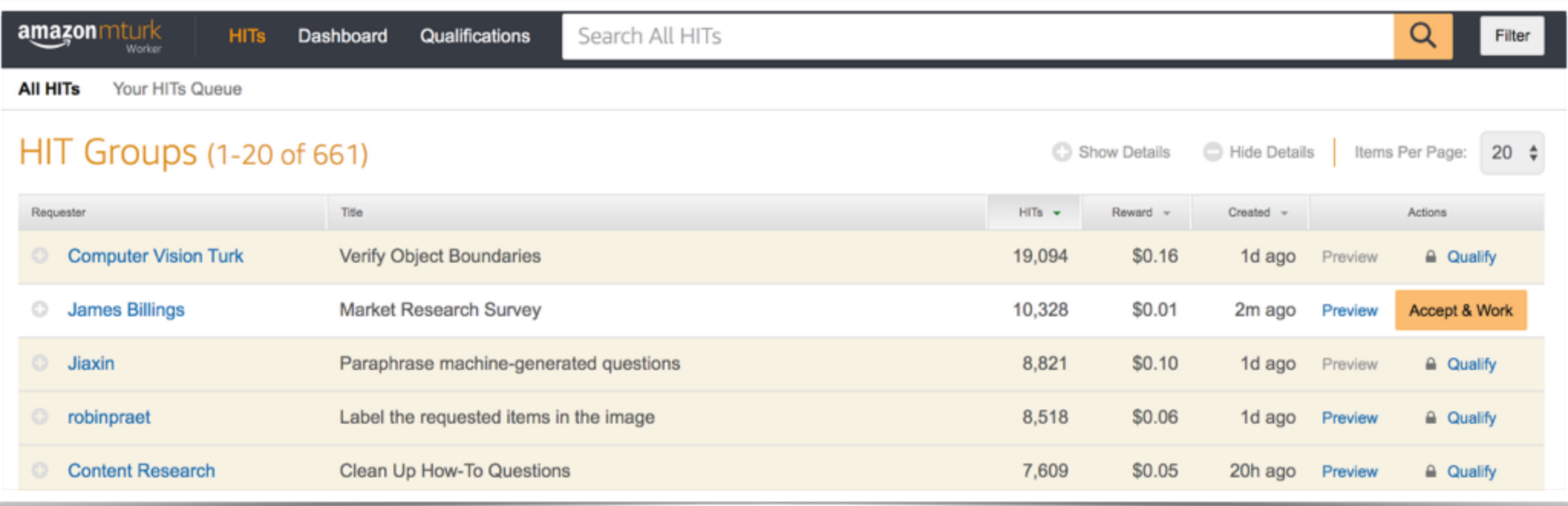


The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

Focus of today:

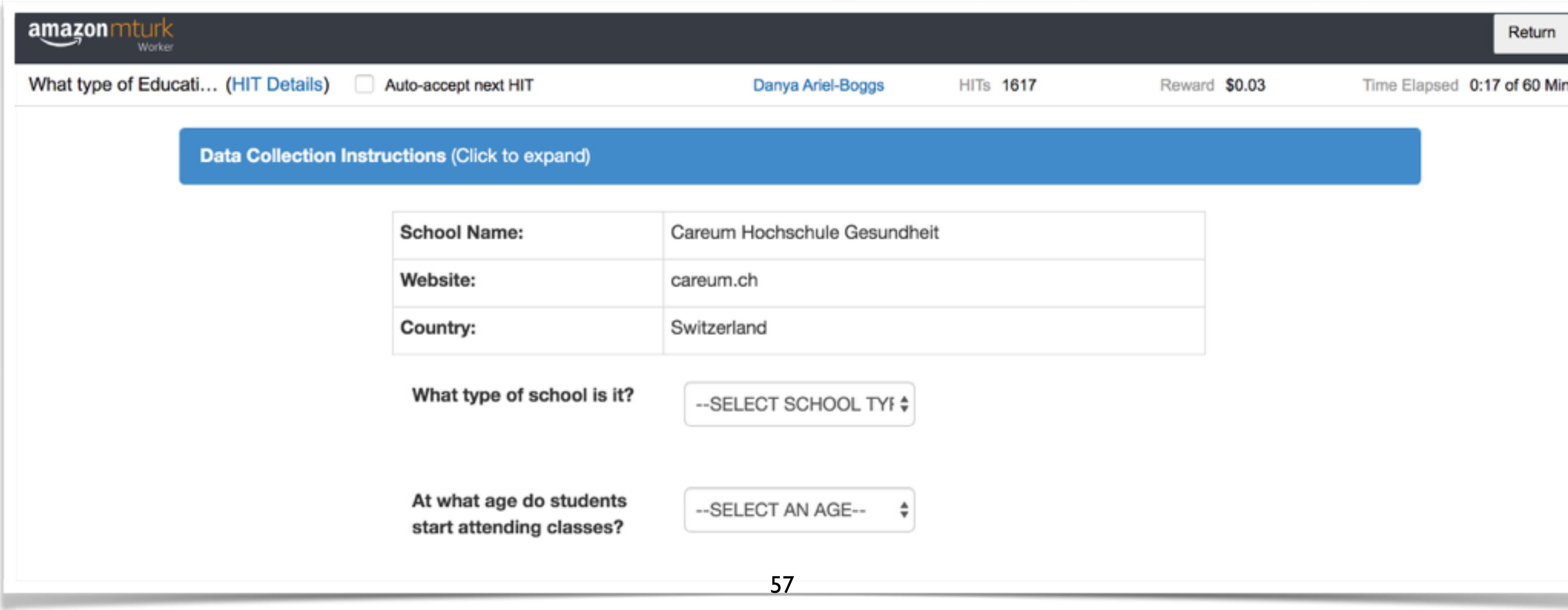
amazon mechanical turk™
Artificial Artificial Intelligence

https://worker.mturk.com/



HIT Groups (1-20 of 661)

Requester	Title	HITs	Reward	Created	Actions	
Computer Vision Turk	Verify Object Boundaries	19,094	$0.16	1d ago	Preview	Qualify
James Billings	Market Research Survey	10,328	$0.01	2m ago	Preview	Accept & Work
Jiaxin	Paraphrase machine-generated questions	8,821	$0.10	1d ago	Preview	Qualify
robinpraet	Label the requested items in the image	8,518	$0.06	1d ago	Preview	Qualify
Content Research	Clean Up How-To Questions	7,609	$0.05	20h ago	Preview	Qualify



What type of Educati… (HIT Details) □ Auto-accept next HIT Danya Ariel-Boggs HITs 1617 Reward $0.03 Time Elapsed 0:17 of 60 Min

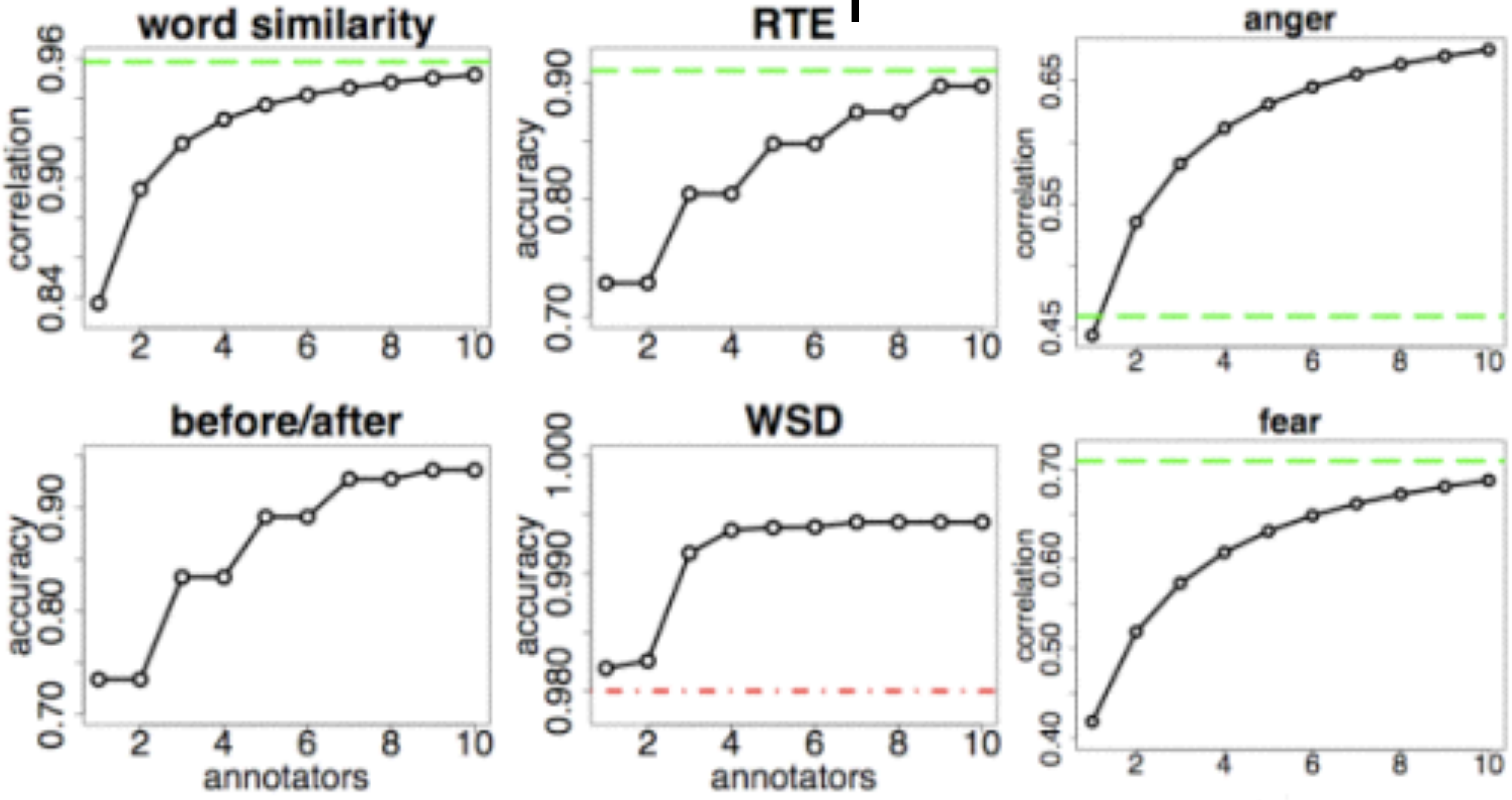
Data Collection Instructions (Click to expand)

School Name:	Careum Hochschule Gesundheit
Website:	careum.ch
Country:	Switzerland

What type of school is it? --SELECT SCHOOL TYF

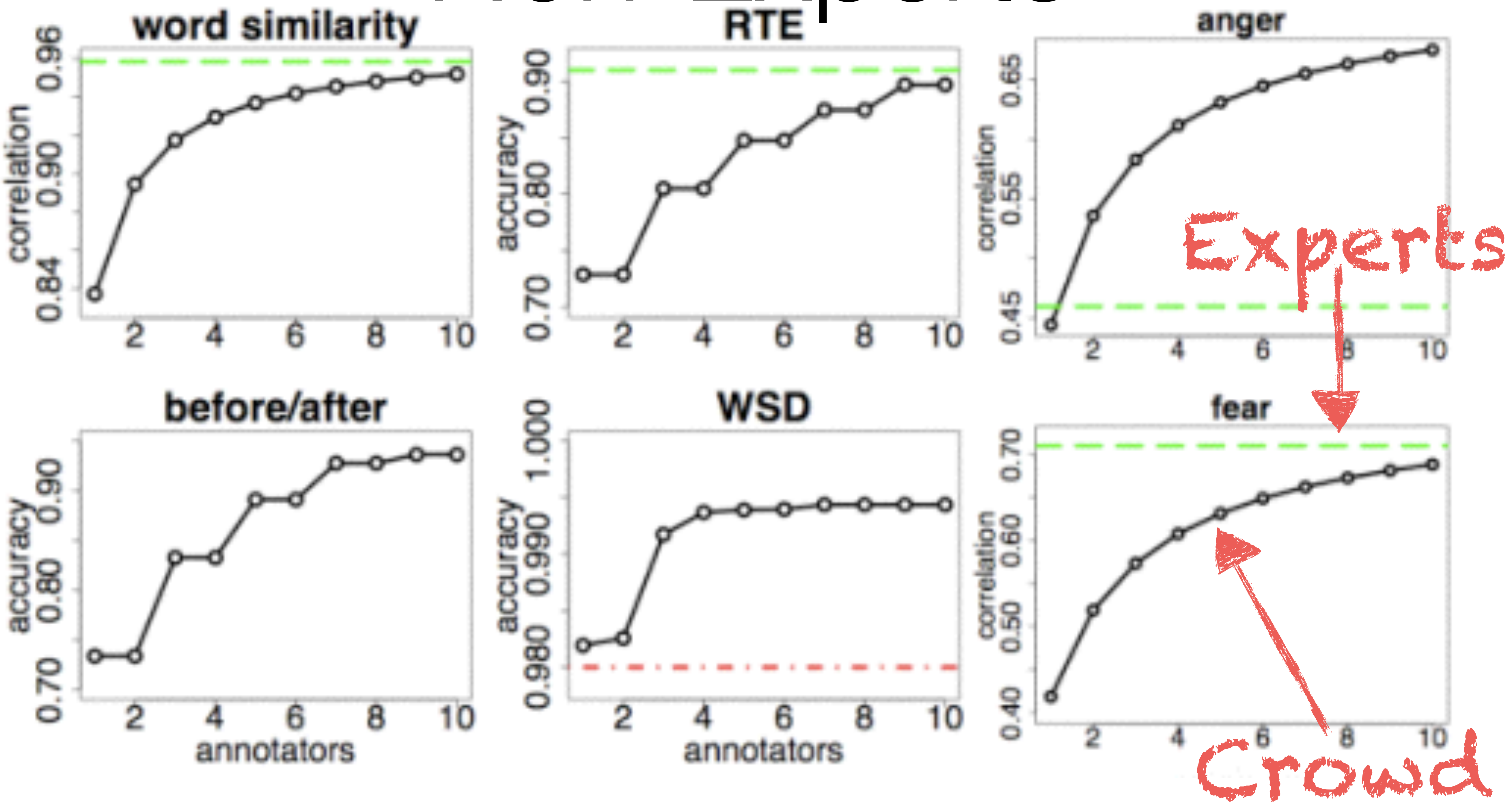
At what age do students start attending classes? --SELECT AN AGE--

Expert Annotation from Non-Experts



- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)

Expert Annotation from Non-Experts

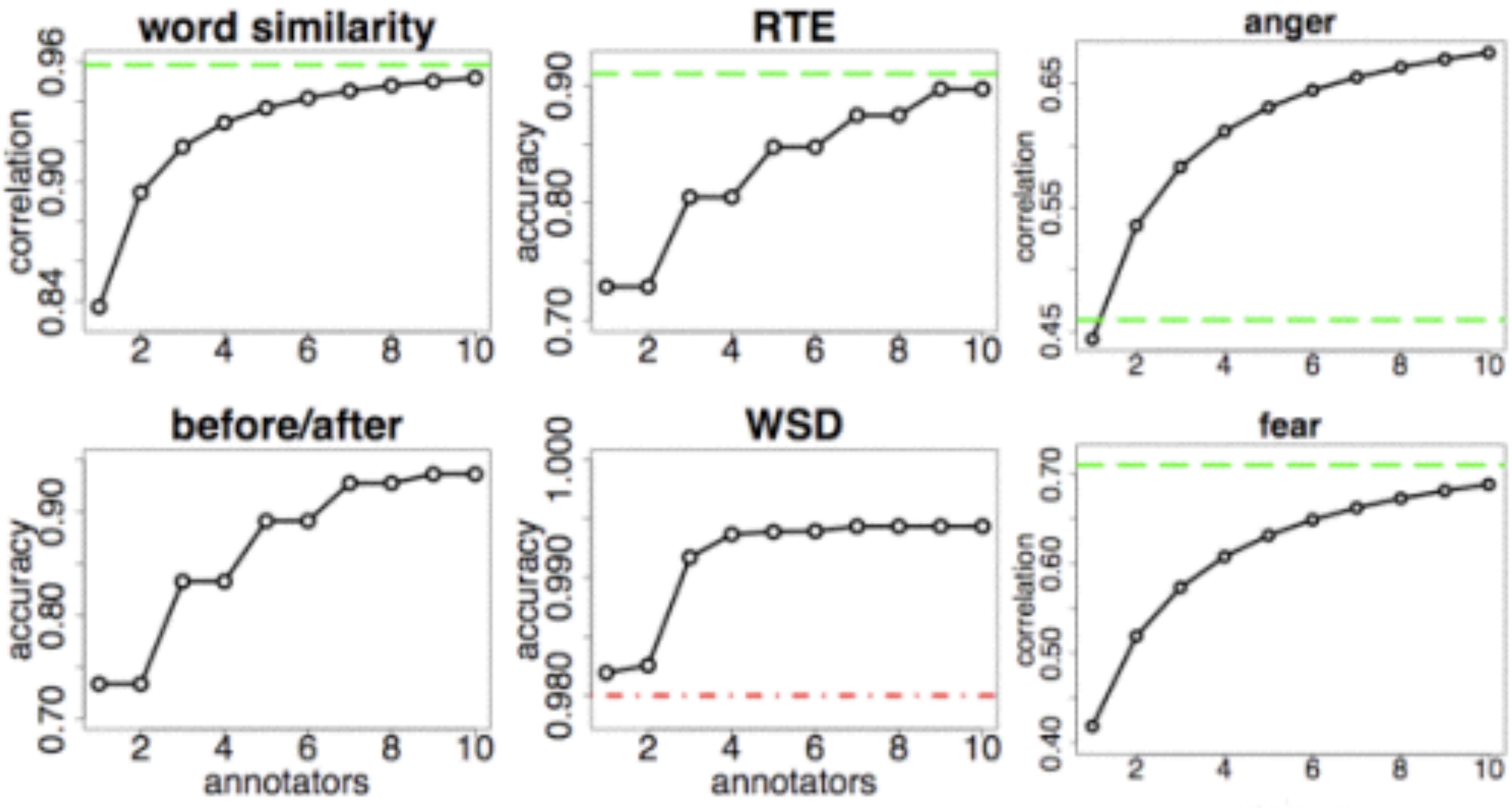


- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)

Quality Control

- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)

Quality Control

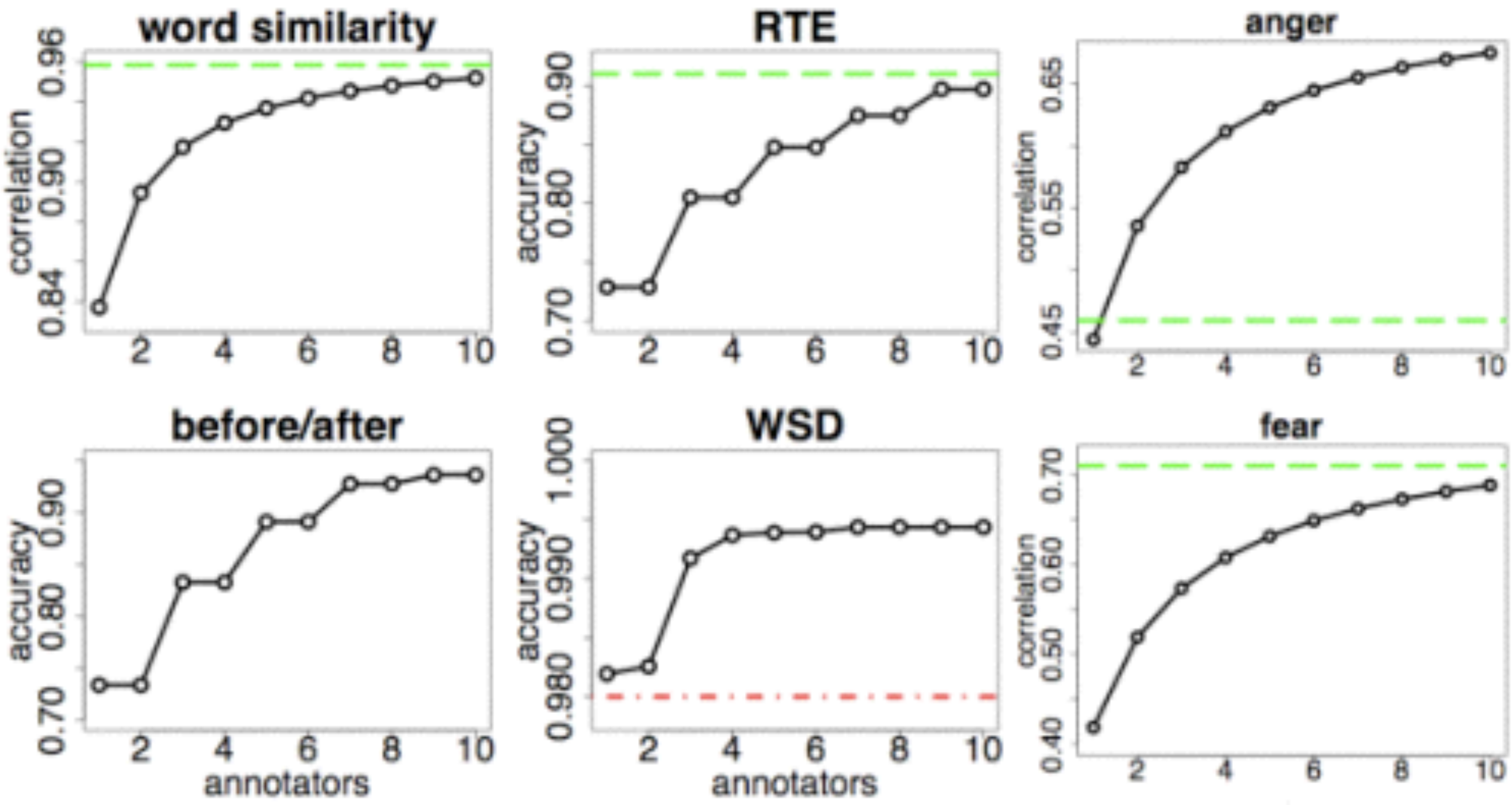


Agreement/Redundancy

- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)

Malicious workers?

Quality Control



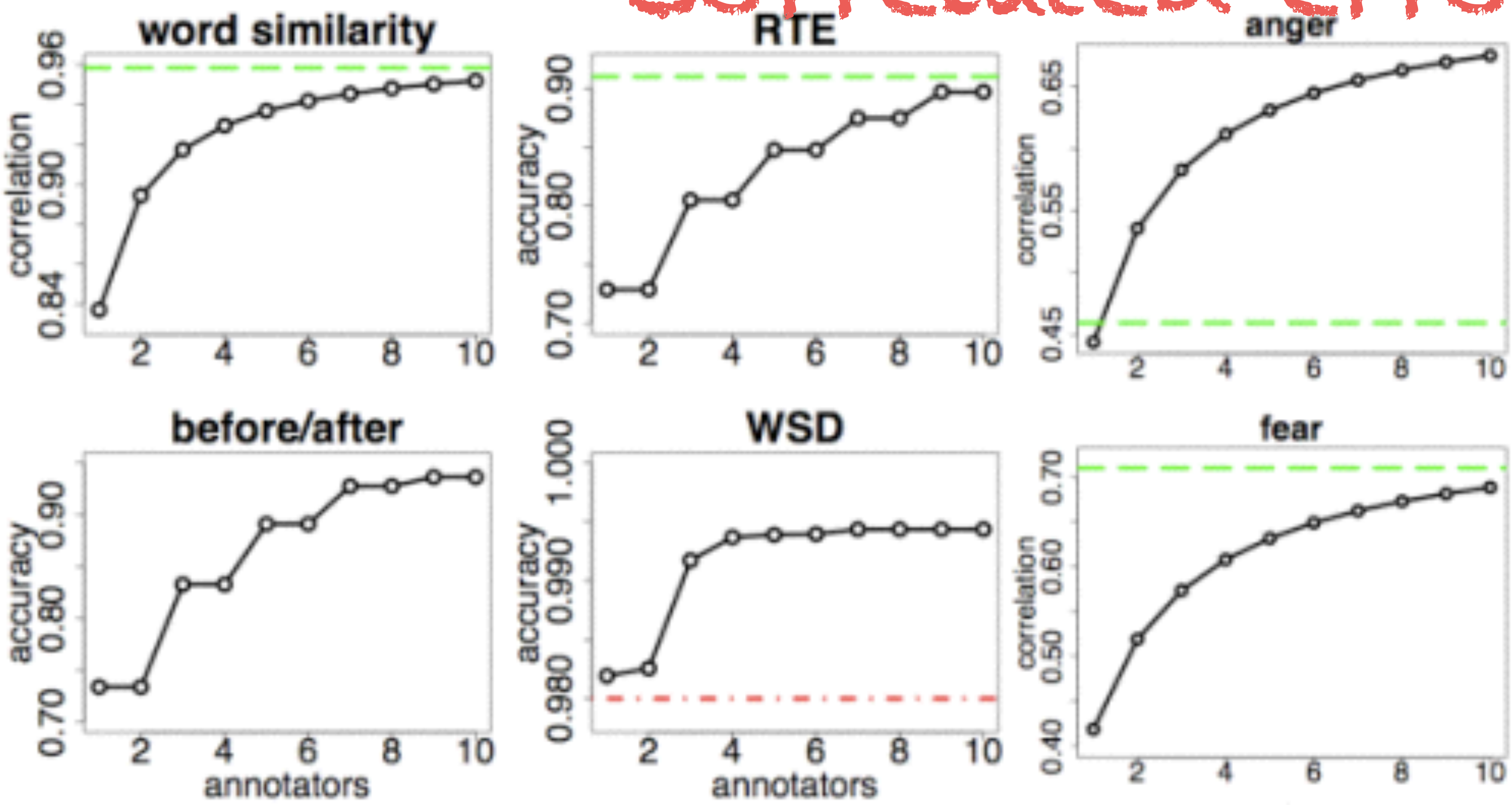
Agreement/Redundancy

- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)

Malicious workers?

Quality Control

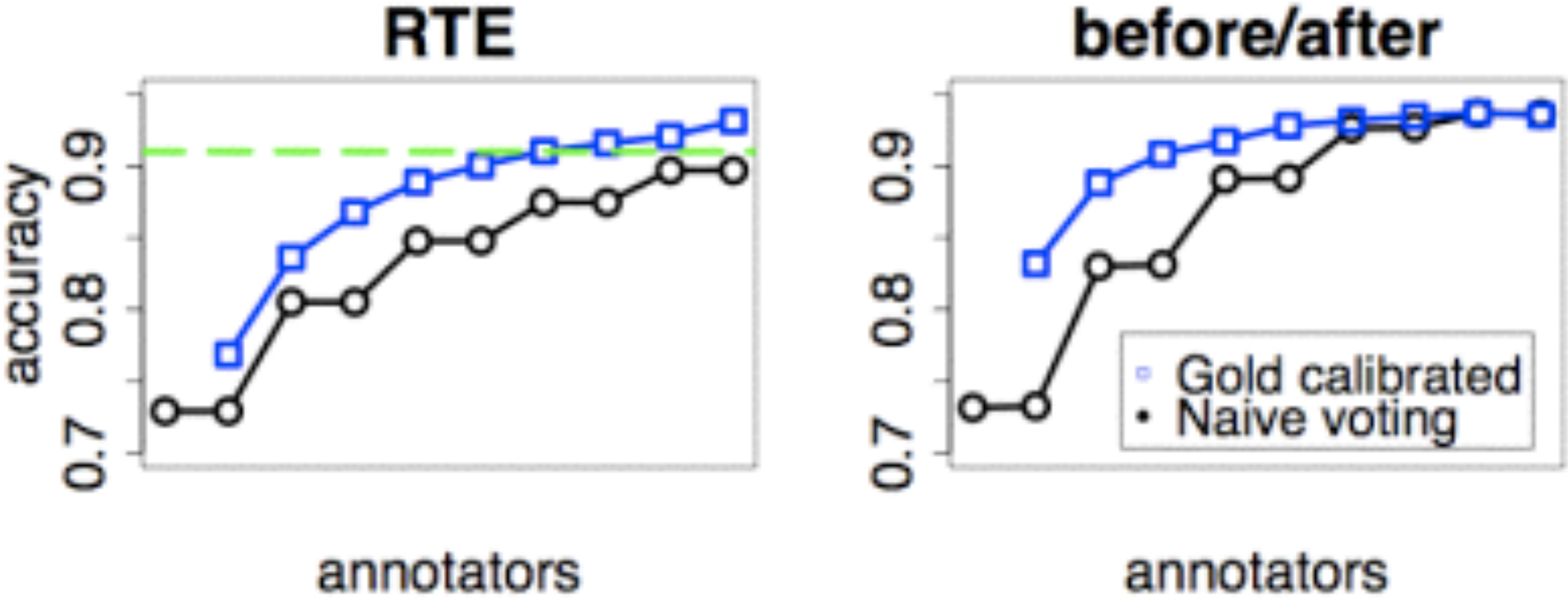
Correlated errors?



Agreement/Redundancy

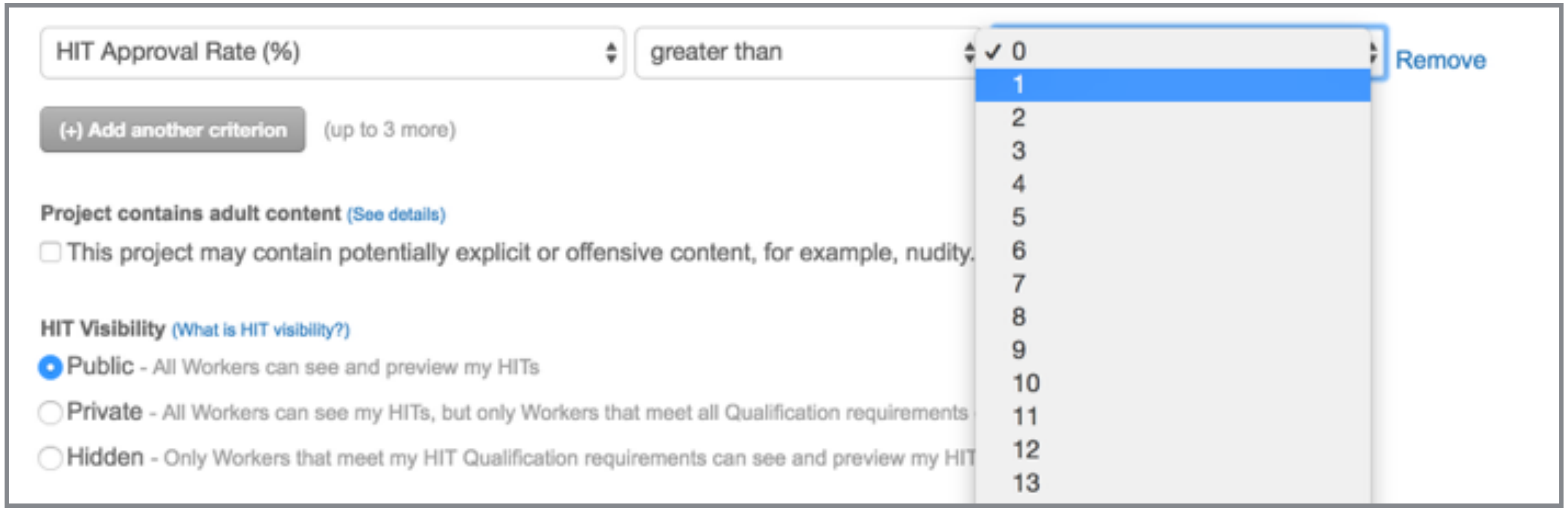
- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)

Quality Control



Confidence estimates on workers improve accuracy.

- Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Snow et al (2008)



HIT Approval Rate (%) | greater than | ✓ 0 | Remove

(up to 3 more)

Project contains adult content (See details)

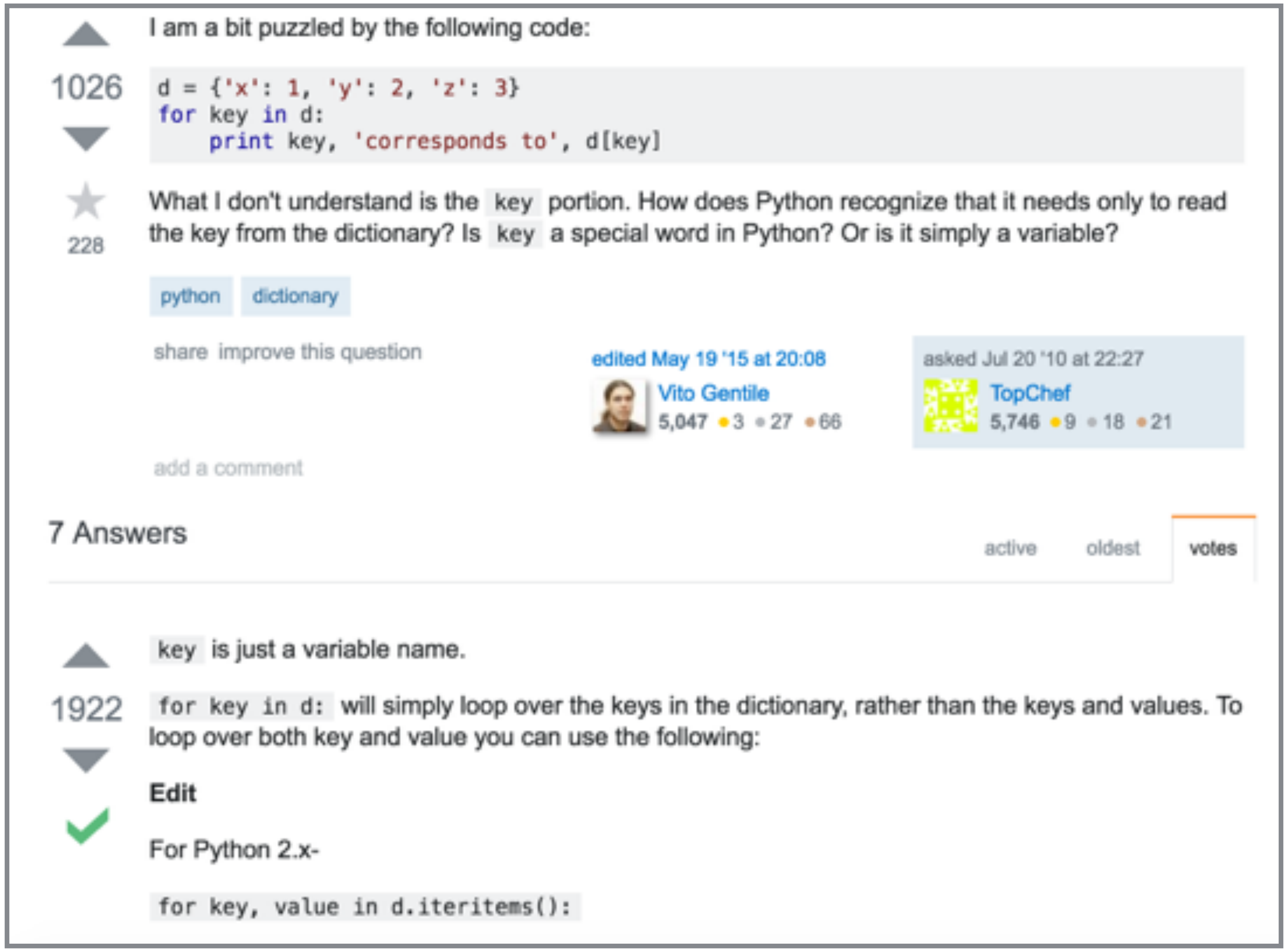
□ This project may contain potentially explicit or offensive content, for example, nudity.

HIT Visibility (What is HIT visibility?)

● Public - All Workers can see and preview my HITs

○ Private - All Workers can see my HITs, but only Workers that meet all Qualification requirements

○ Hidden - Only Workers that meet my HIT Qualification requirements can see and preview my HIT



I am a bit puzzled by the following code:

```
d = {'x': 1, 'y': 2, 'z': 3}
for key in d:
    print key, 'corresponds to', d[key]
```

What I don't understand is the `key` portion. How does Python recognize that it needs only to read the key from the dictionary? Is `key` a special word in Python? Or is it simply a variable?

edited May 19 '15 at 20:08
Vito Gentile
5,047 ●3 ●27 ●66

asked Jul 20 '10 at 22:27
TopChef
5,746 ●9 ●18 ●21

7 Answers

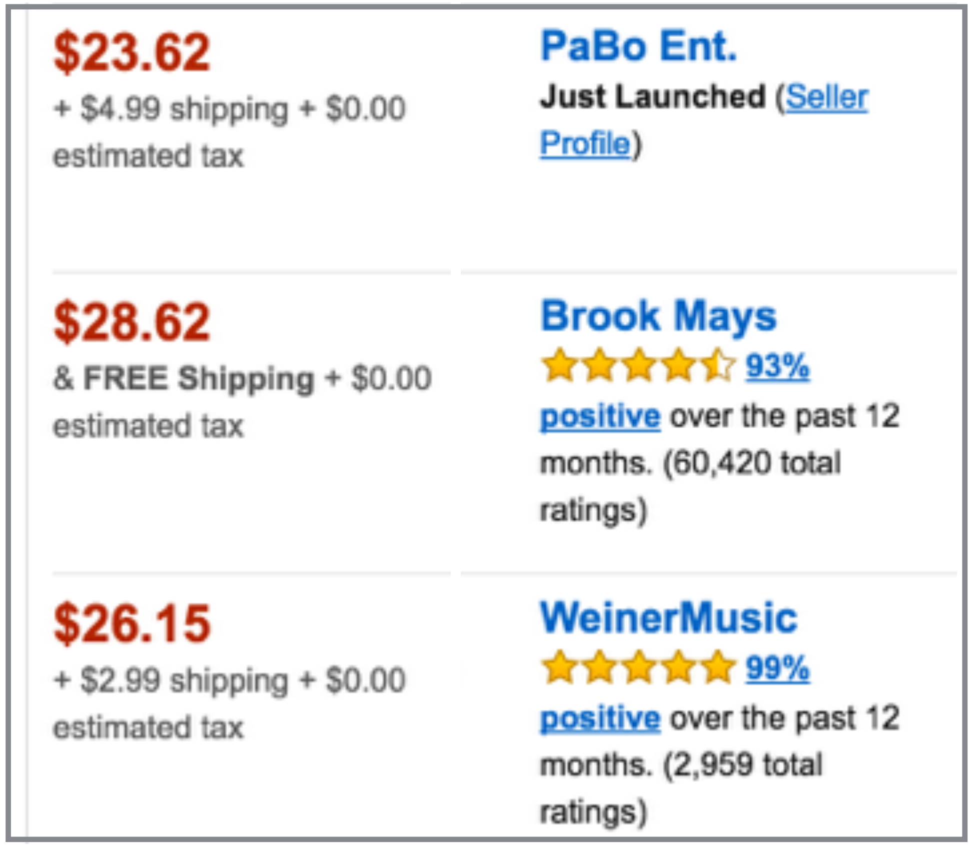
`key` is just a variable name.

`for key in d:` will simply loop over the keys in the dictionary, rather than the keys and values. To loop over both key and value you can use the following:

Edit

For Python 2.x-

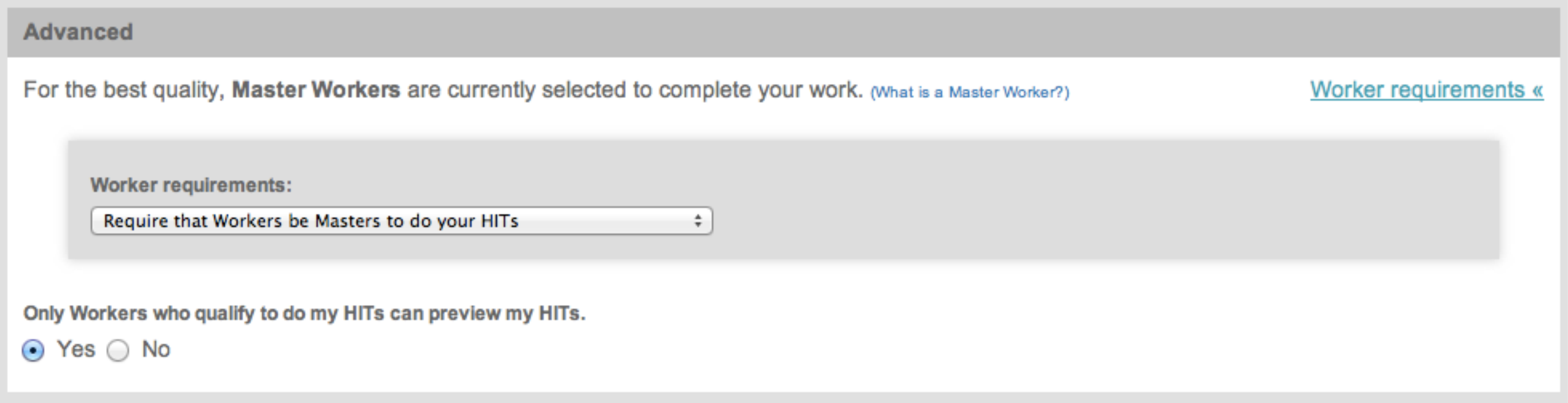
```
for key, value in d.iteritems():
```



$23.62 + $4.99 shipping + $0.00 estimated tax	**PaBo Ent.** **Just Launched** (Seller Profile)
$28.62 & FREE Shipping + $0.00 estimated tax	**Brook Mays** ★★★★☆ 93% positive over the past 12 months. (60,420 total ratings)
$26.15 + $2.99 shipping + $0.00 estimated tax	**WeinerMusic** ★★★★★ 99% positive over the past 12 months. (2,959 total ratings)

Reputation Systems

Quality Control



Advanced

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

Worker requirements:

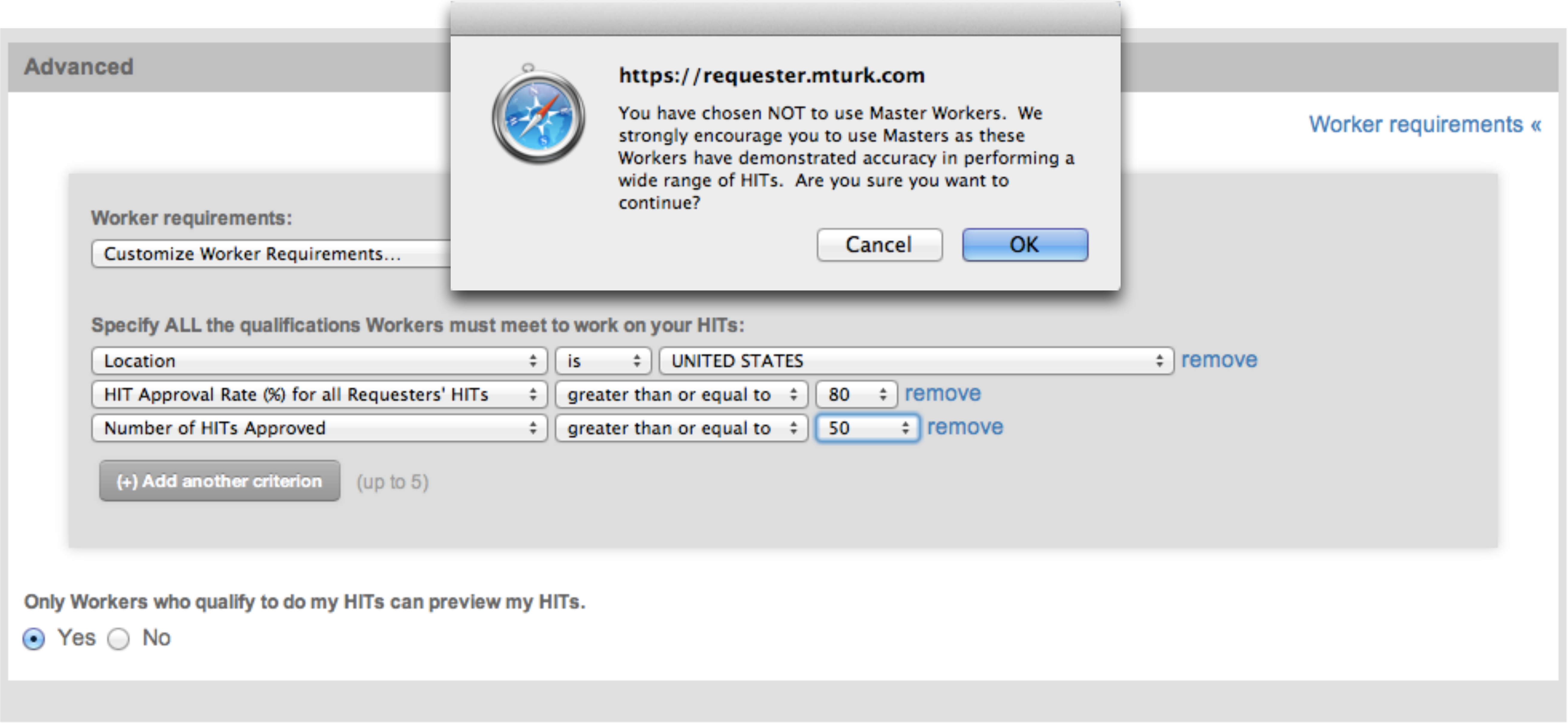
Require that Workers be Masters to do your HITs

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

● Yes ○ No

Pre-Vetted Workers

Quality Control



Advanced

Worker requirements «

Worker requirements:

Customize Worker Requirements...

https://requester.mturk.com

You have chosen NOT to use Master Workers. We strongly encourage you to use Masters as these Workers have demonstrated accuracy in performing a wide range of HITs. Are you sure you want to continue?

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

Location	is	UNITED STATES	remove
HIT Approval Rate (%) for all Requesters' HITs	greater than or equal to	80	remove
Number of HITs Approved	greater than or equal to	50	remove

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

● Yes ○ No

Pre-Vetted Workers

Quality Control

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.

Quality Control

- Amazon now nominates a subset (21k workers, estimated at 10% of all Turkers) of senior / good workers as “Masters”
- Amazon charges 25% commission for Masters versus their normal 20% rate
- They have now implemented this as the default qualification for new Requesters
- Why?

Pre-Vetted Workers

Quality Control

Pros

- Easier for new requesters who do not know to implement quality control.
- Masters will not touch badly designed or low-paying tasks

et (21k
ll Turkers) of
ers”

- Amazon charges 25% commission for Masters versus their normal 20% rate
- They have now implemented this as the default qualification for new Requesters
- Why?

Pre-Vetted Workers

Quality Control

Pros

- Easier for new requesters who do not know to implement quality control.
- Masters will not touch badly designed or low-paying tasks

et (21k

ll Turkers) of

ers”

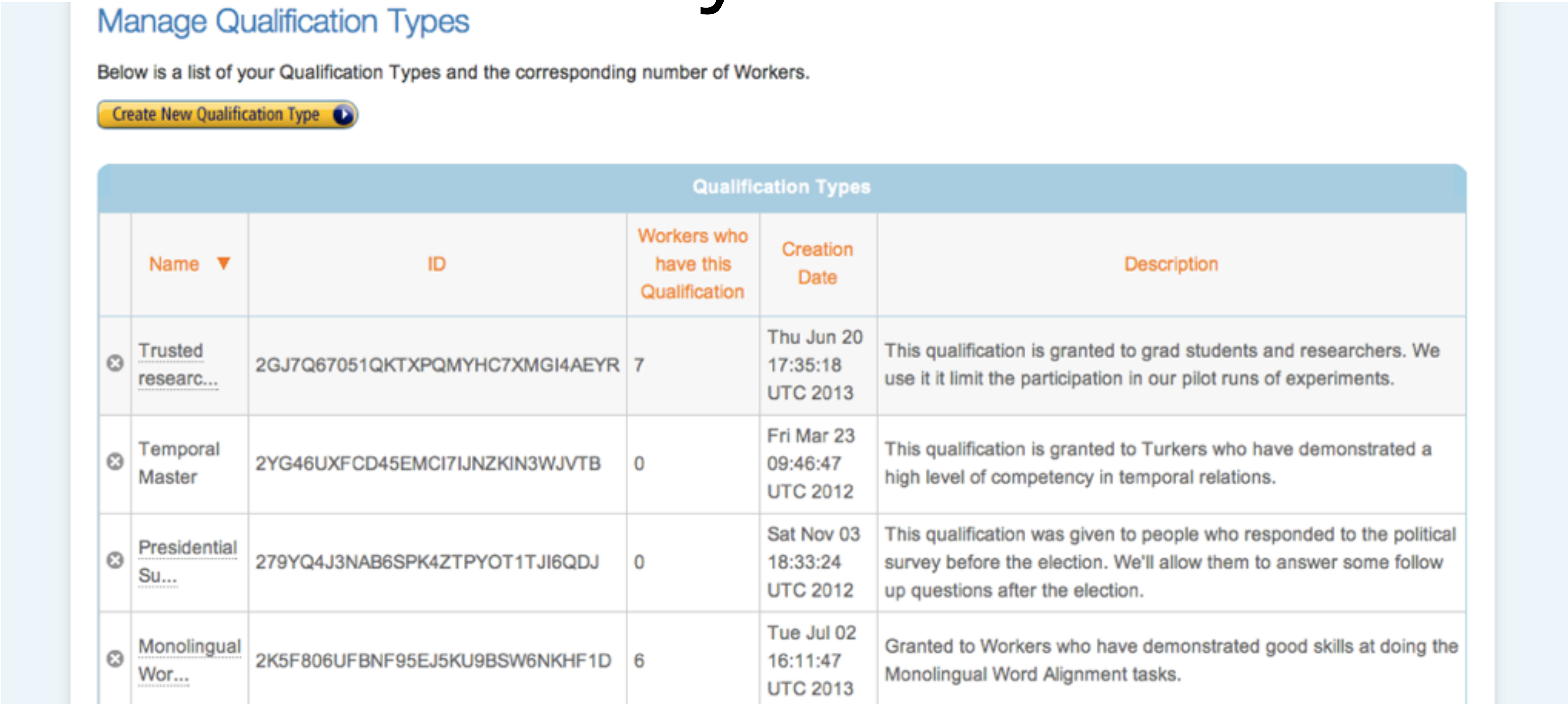
- Amazon charges 25% commission for Master
- They h default
- Why?

Cons

- Fewer Masters workers -> significant lag in taks being picked up
- More expensive
- Not clear in what tasks the Masters are tested and how a new worker can become a master.

P

Quality Control



Manage Qualification Types

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

Qualification Types

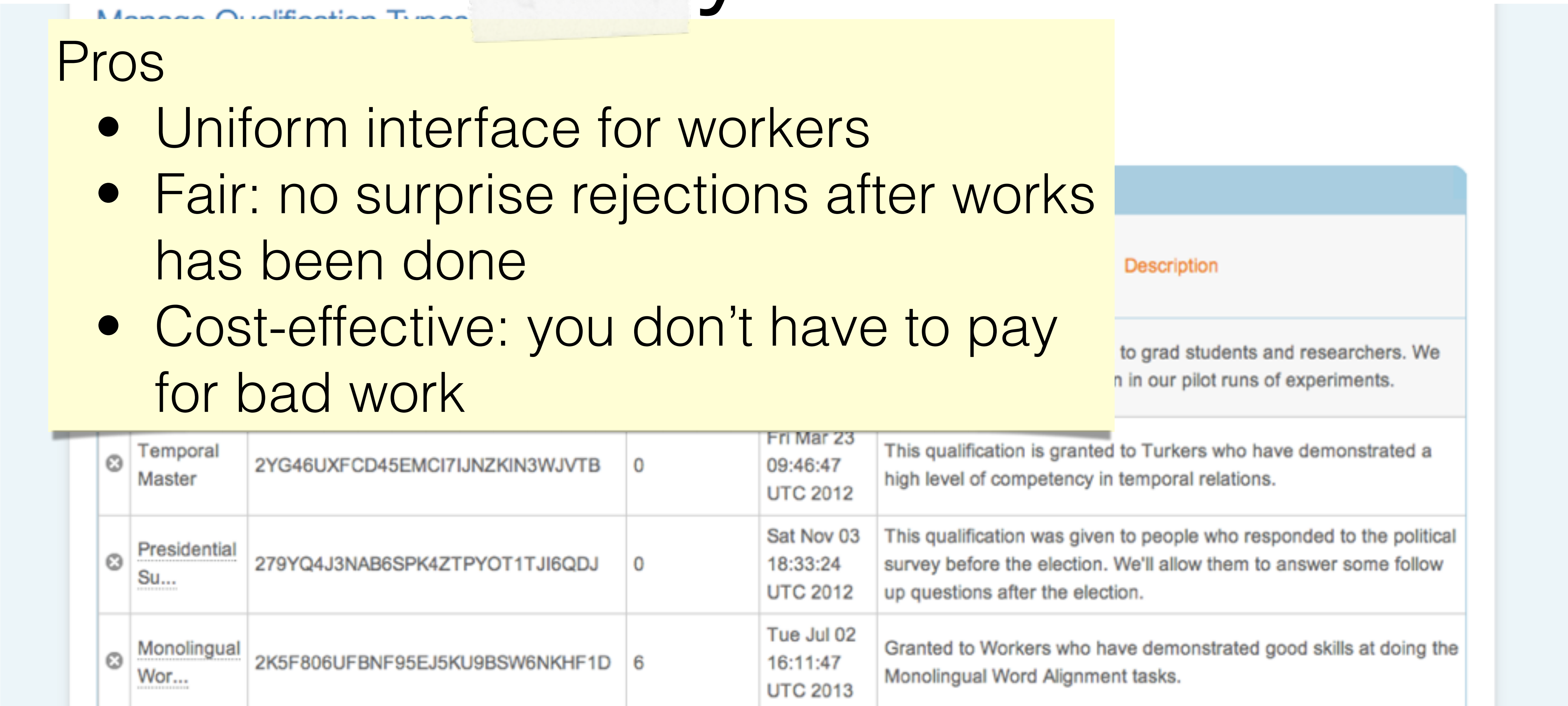
	Name ▼	ID	Workers who have this Qualification	Creation Date	Description
⊗	Trusted researc...	2GJ7Q67051QKTXPQMYHC7XMGI4AEYR	7	Thu Jun 20 17:35:18 UTC 2013	This qualification is granted to grad students and researchers. We use it it limit the participation in our pilot runs of experiments.
⊗	Temporal 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 Wor...	2K5F806UFBNF95EJ5KU9BSW6NKHF1D	6	Tue Jul 02 16:11:47 UTC 2013	Granted to Workers who have demonstrated good skills at doing the Monolingual Word Alignment tasks.

Qualification Tests

Quality Control

Pros

- Uniform interface for workers
- Fair: no surprise rejections after works has been done
- Cost-effective: you don't have to pay for bad work

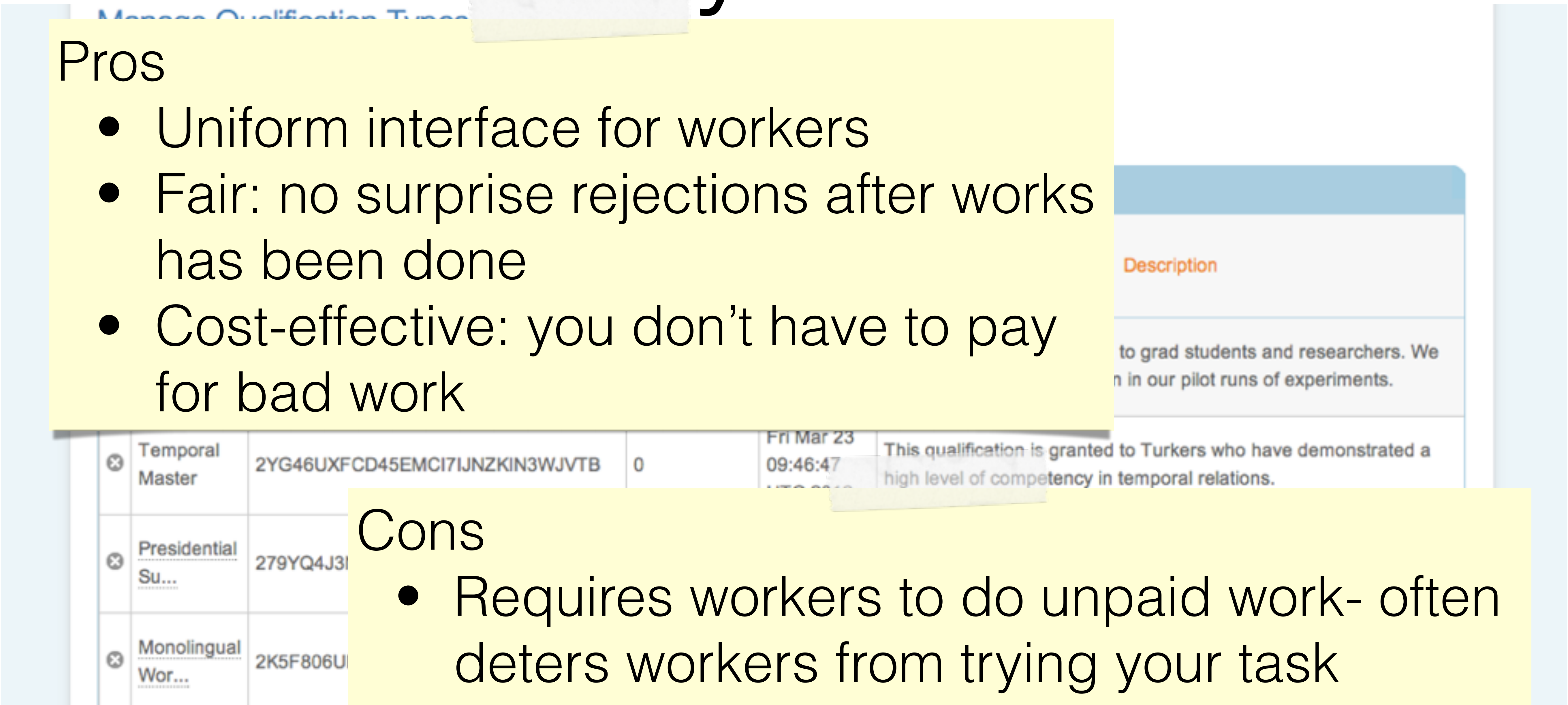


Manage Qualification Types

					Description
					to grad students and researchers. We n 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 Wor...	2K5F806UFBNF95EJ5KU9BSW6NKHF1D	6	Tue Jul 02 16:11:47 UTC 2013	Granted to Workers who have demonstrated good skills at doing the Monolingual Word Alignment tasks.

Qualification Tests

Quality Control



Manage Qualification Types

Description

to grad students and researchers. We
n in our pilot runs of experiments.

Temporal Master	2YG46UXFCD45EMCI7IJNZKIN3WJVTB	0	Fri Mar 23 09:46:47	This qualification is granted to Turkers who have demonstrated a high level of competency in temporal relations.
Presidential Su...	279YQ4J3			
Monolingual Wor...	2K5F806U			

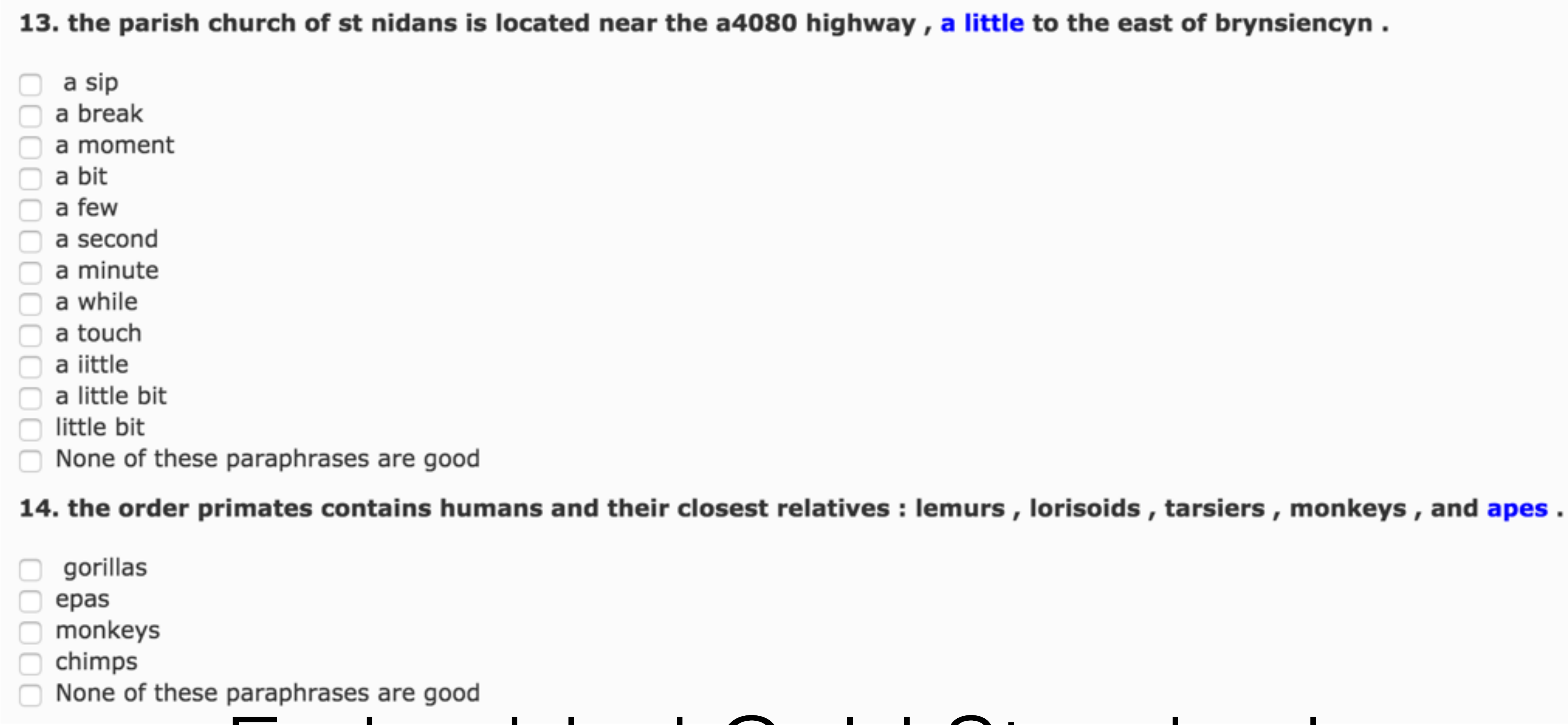
Pros

- Uniform interface for workers
- Fair: no surprise rejections after works has been done
- Cost-effective: you don't have to pay for bad work

Cons

- Requires workers to do unpaid work- often deters workers from trying your task
- Turkers knows when they are being evaluated, so their performance on the test might not reflect performance on the task

Quality Control



13. the parish church of st nidans is located near the a4080 highway , a little to the east of brynsienc yn .

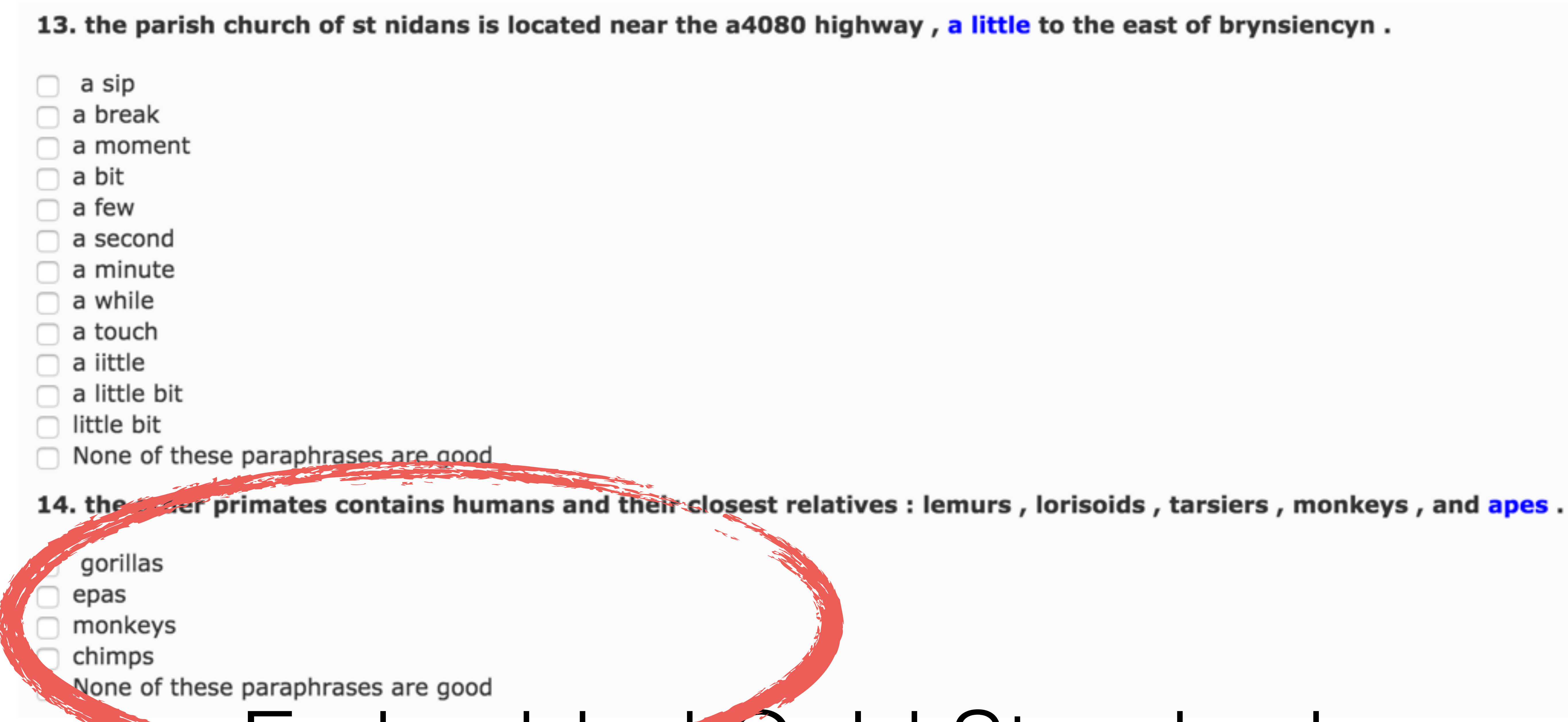
□ a sip
□ a break
□ a moment
□ a bit
□ a few
□ a second
□ a minute
□ a while
□ a touch
□ a iittle
□ a little bit
□ little bit
□ None of these paraphrases are good

14. the order primates contains humans and their closest relatives : lemurs , lorisoids , tarsiers , monkeys , and apes .

□ gorillas
□ epas
□ monkeys
□ chimps
□ None of these paraphrases are good

Embedded Gold Standard

Quality Control



13. the parish church of st nidans is located near the a4080 highway , a little to the east of brynsiencyn .

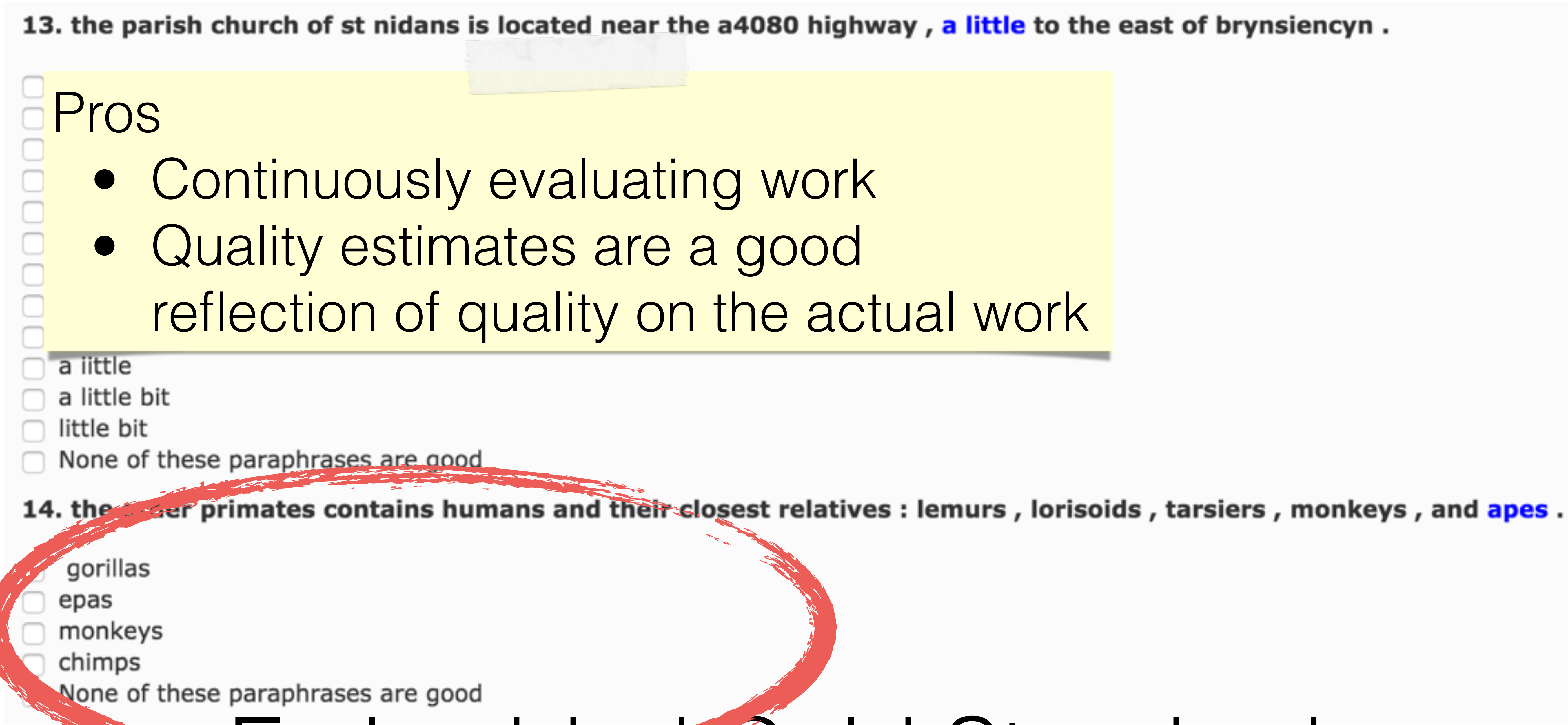
□ a sip
□ a break
□ a moment
□ a bit
□ a few
□ a second
□ a minute
□ a while
□ a touch
□ a iittle
□ a little bit
□ little bit
□ None of these paraphrases are good

14. the order primates contains humans and their closest relatives : lemurs , lorisoids , tarsiers , monkeys , and apes .

□ gorillas
□ epas
□ monkeys
□ chimps
□ None of these paraphrases are good

Embedded Gold Standard

Quality Control



13. the parish church of st nidans is located near the a4080 highway , a little to the east of brynsiencyn .

Pros

- Continuously evaluating work
- Quality estimates are a good reflection of quality on the actual work

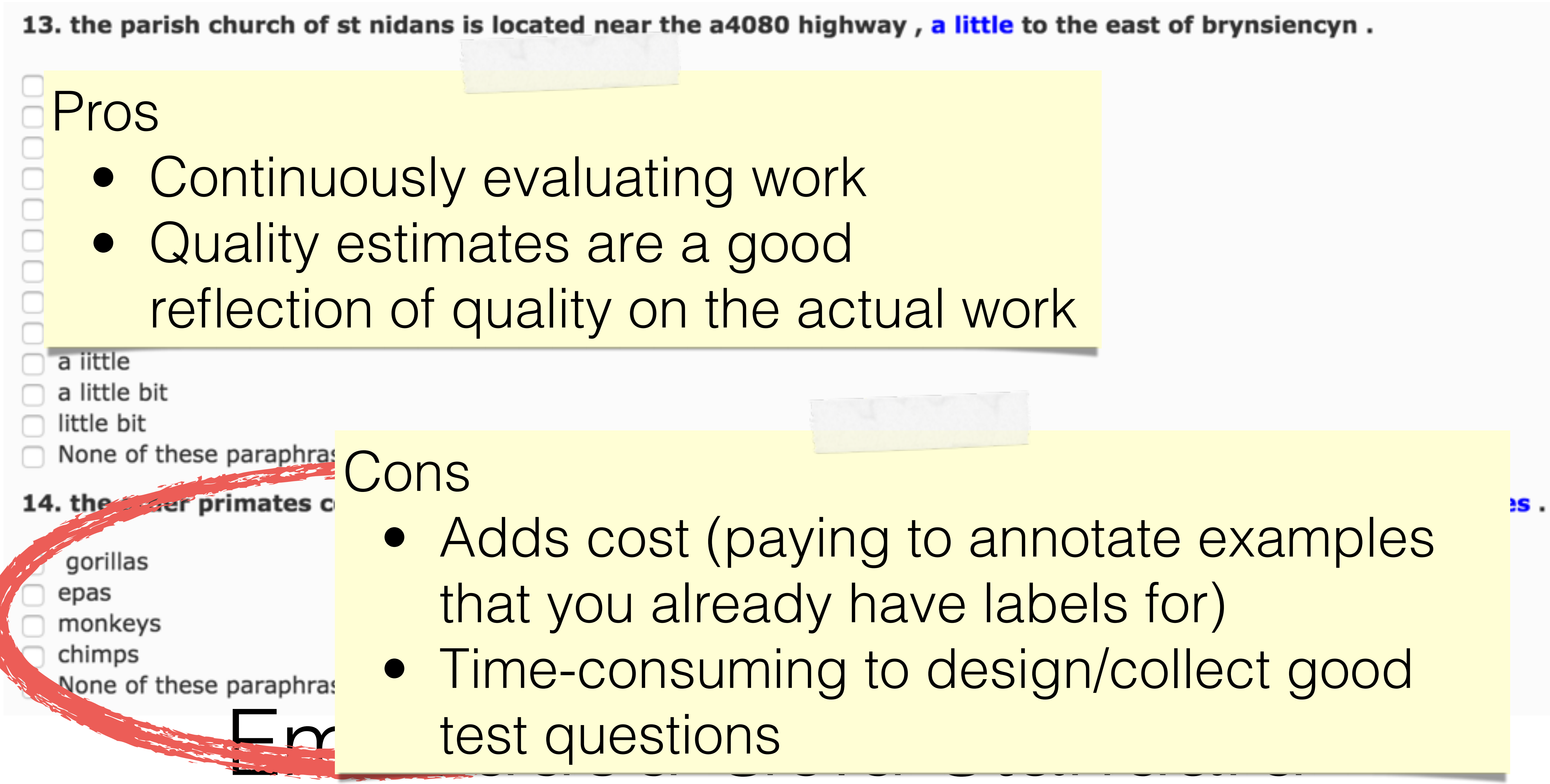
□ a little
□ a little bit
□ little bit
□ None of these paraphrases are good

14. the order primates contains humans and their closest relatives : lemurs , lorisoids , tarsiers , monkeys , and apes .

□ gorillas
□ epas
□ monkeys
□ chimps
□ None of these paraphrases are good

Embedded Gold Standard

Quality Control



13. the parish church of st nidans is located near the a4080 highway , a little to the east of brynsiencyn .

☐ a little
☐ a little bit
☐ little bit
☐ None of these paraphras

14. the ...er primates c ...es .

☐ gorillas
☐ epas
☐ monkeys
☐ chimps
☐ None of these paraphras

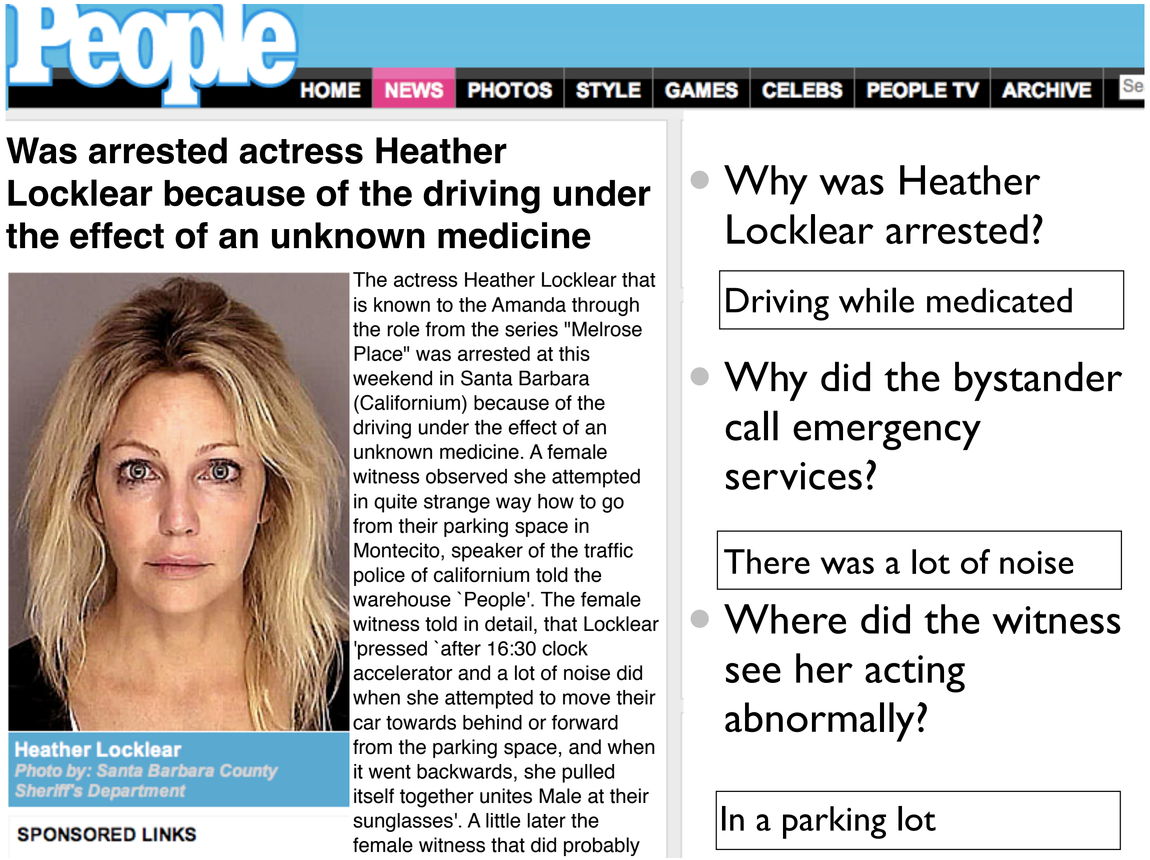
Pros

- Continuously evaluating work
- Quality estimates are a good reflection of quality on the actual work

Cons

- Adds cost (paying to annotate examples that you already have labels for)
- Time-consuming to design/collect good test questions

Quality Control



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?

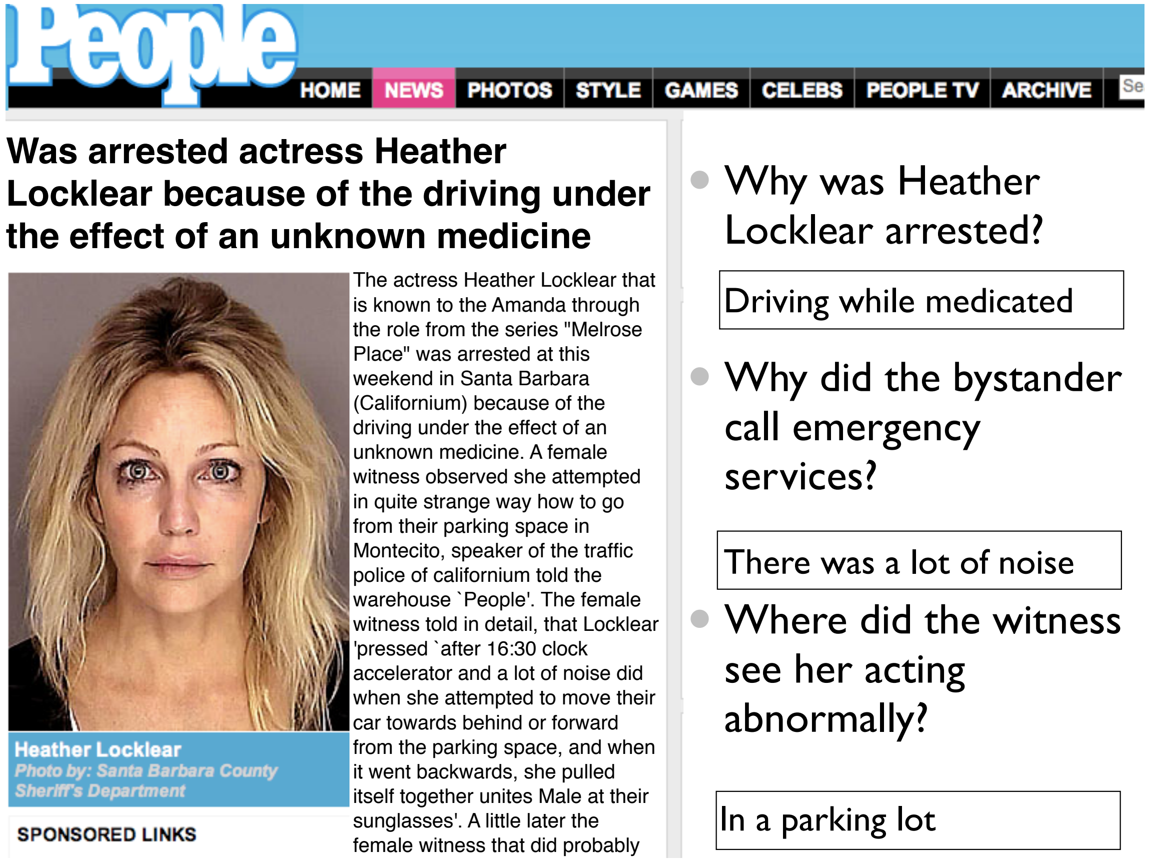
There was a lot of noise

- Where did the witness see her acting abnormally?

In a parking lot

Second-Pass HIT

Quality Control



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

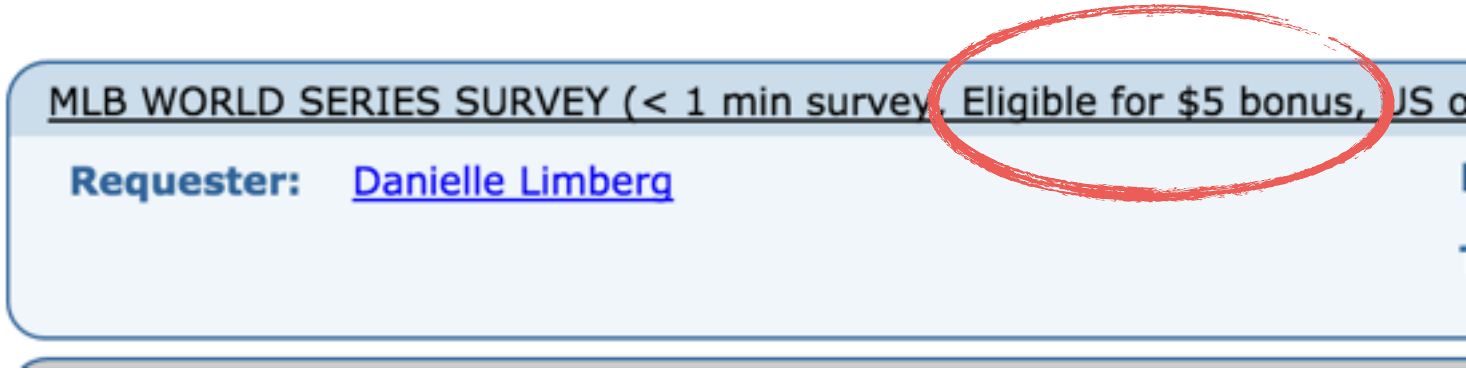
- Why did the bystander call emergency services?

There was a lot of noise

- Where did the witness see her acting abnormally?

In a parking lot

Second-Pass HIT

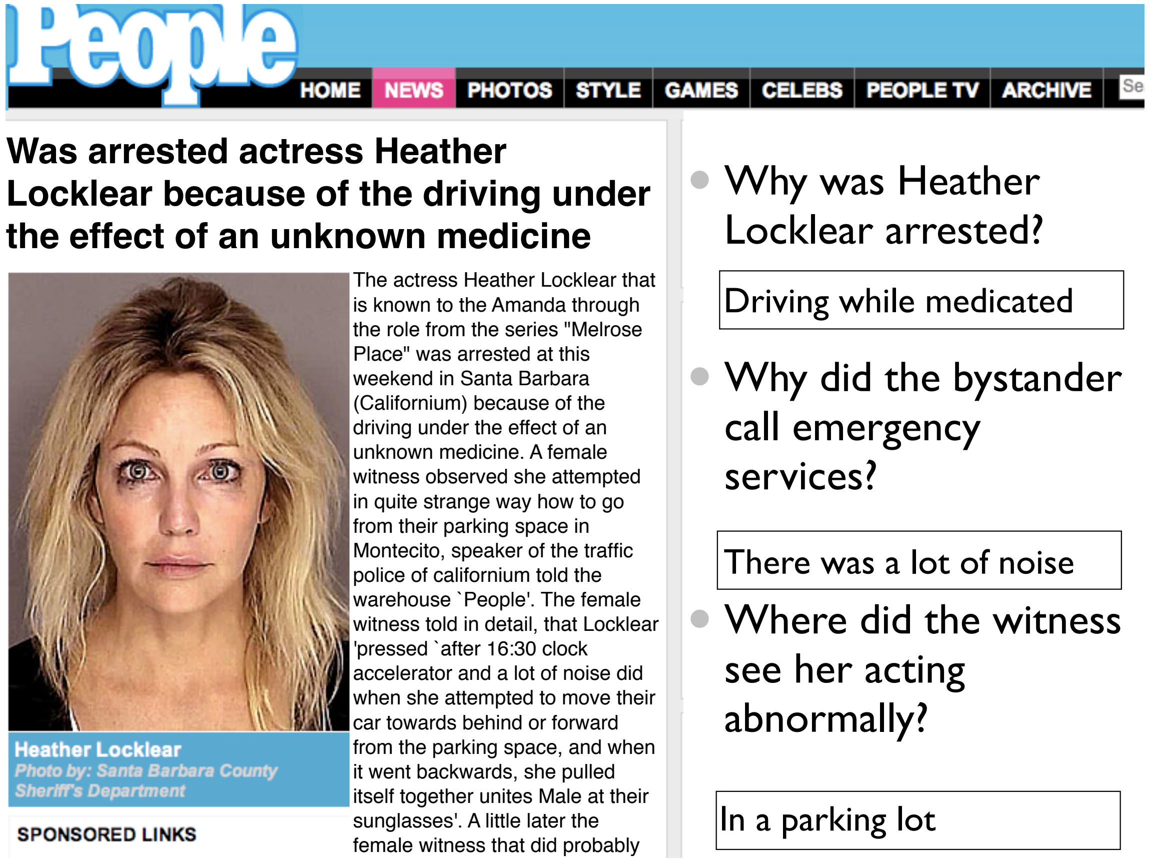


MLB WORLD SERIES SURVEY (< 1 min survey, Eligible for $5 bonus, US o

Requester: Danielle Limberg

Incentive Pay

Quality Control



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

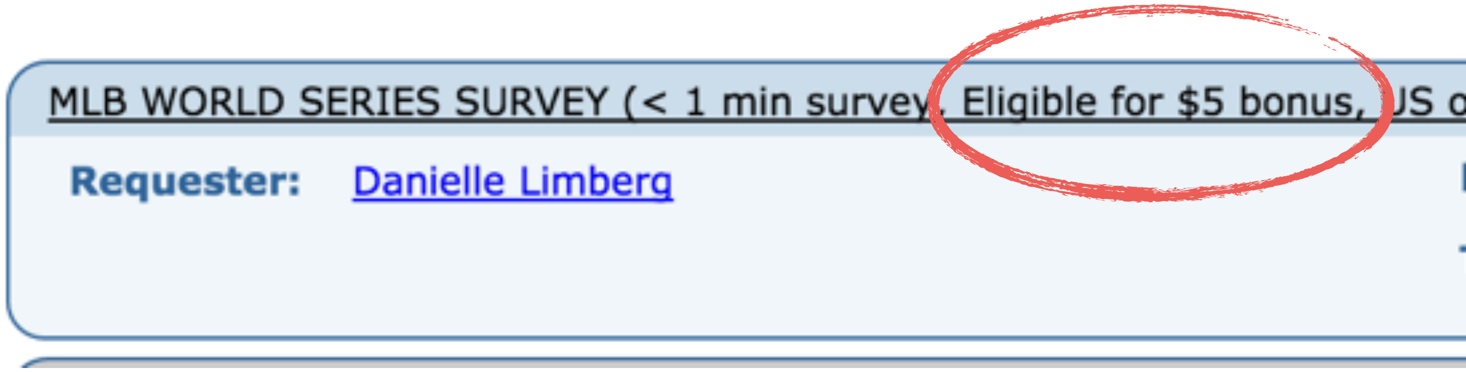
- Why did the bystander call emergency services?

There was a lot of noise

- Where did the witness see her acting abnormally?

In a parking lot

Second-Pass HIT


MLB WORLD SERIES SURVEY (< 1 min survey, Eligible for $5 bonus, US o

Requester: Danielle Limberg

Incentive Pay

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Statistical Models

Common Misconceptions:

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled
- Work for $1/hour, doing it for fun in our PJs, unemployed

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled
- Work for $1/hour, doing it for fun in our PJs, unemployed
- Isolated, anti-social

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled
- Work for $1/hour, doing it for fun in our PJs, unemployed
- Isolated, anti-social
- Cheaters, lazy, satisficers, inattentive

Comm onceptions:

ntries, non-

speal

unskilled

tentive

I sat contemplating that question over and over. I wanted to say not necessarily true or false but at the last moment decided to change my mind about it.

I did an awful lot of these HITs. For my part, it was because they pay very well and I enjoy them quite a bit--**finally a productive use for my hitherto underutilized English degree!**

I currently have a 98.4% approval ra got the extensions that warned me a requesters who mass reject I unfortunately was victim to many of them who dropped my approval rating. I was wondering if you could make an exception to your rule for me…**Obviously if you aren't happy with my work you could take away my qualification** to work on them. Thank you for your consideration.

How Turkers Work

- 10-20% of workers do 80% of the work
- Want large batches with high throughput
- Often dislike one-off HITs, e.g. surveys

- Musthag, M., & Ganesan, D. (2013). Labor dynamics in a mobile micro-task market. Proceedings of the SIGCHI Conference on …, 641. http://doi.org/10.1145/2470654.2470745
- Chandler, J., Mueller, P. A., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behavior Research Methods, 46, 112–130. http://doi.org/10.3758/s13428-013-0365-7

How Turkers Work

- Online communities: Turkopticon, TurkerNation, Reddit, Facebook
- Scripts: IndiaTurkers, GreasyFork, HitDB, TurkMaster, HIT Scraper
- Websites and plugins: Turk Alert, mTurk List, CrowdWorkers

$

This is funny because it is a regex joke. Please laugh and validate me. I will wait.

$

*yes, this joke is recycled from last year. people didn't laugh then, but this time will be different. I can feel it.