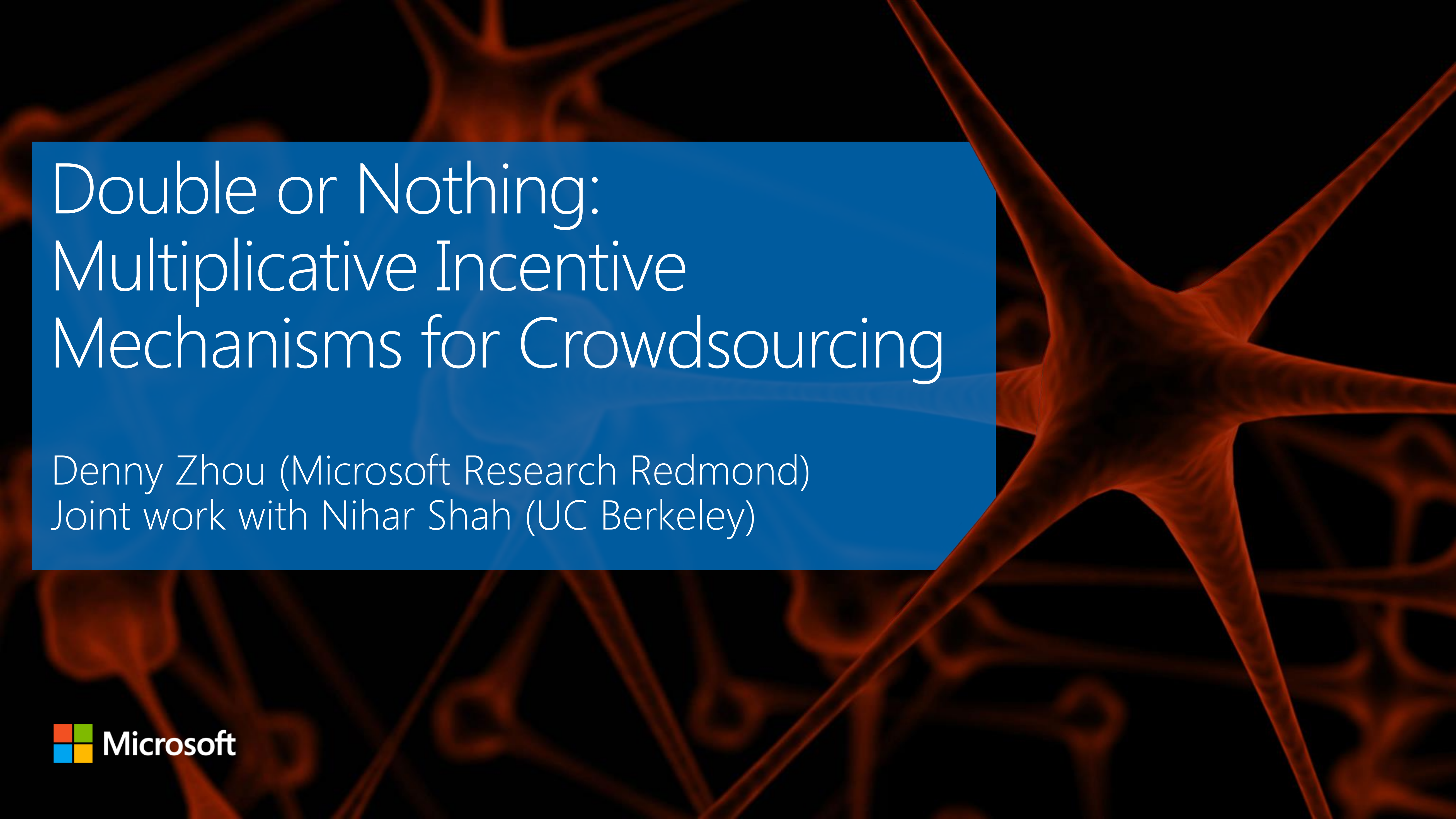


Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing

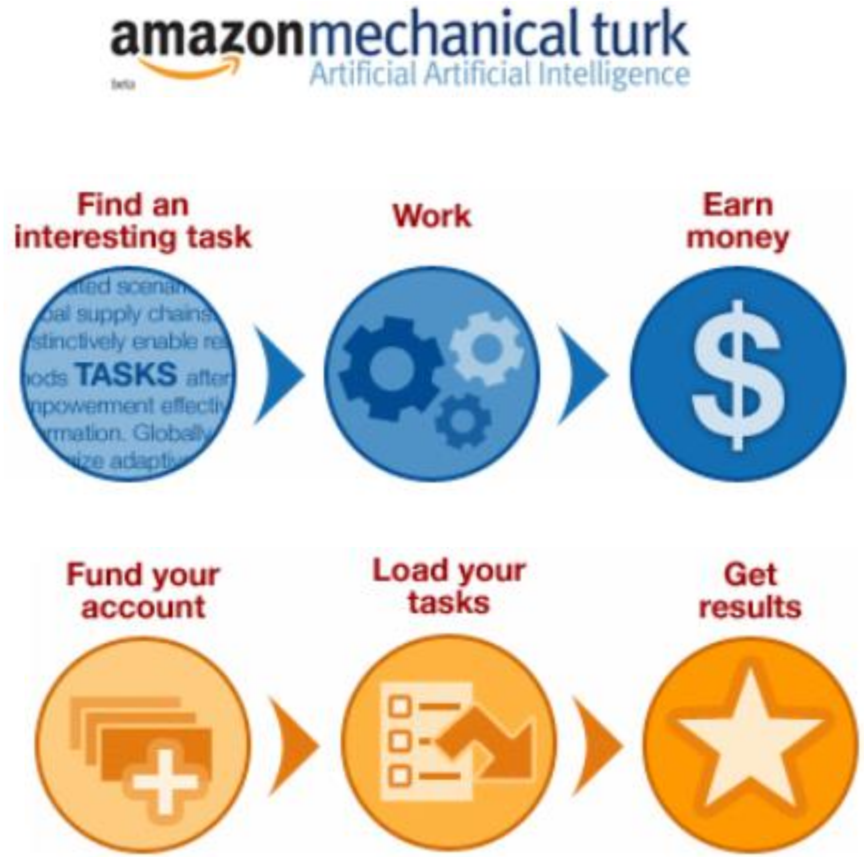
Denny Zhou (Microsoft Research Redmond)
Joint work with Nihar Shah (UC Berkeley)

Microsoft

Collecting labels via crowdsourcing

Is this the Golden Gate Bridge?

○ Yes ○ No

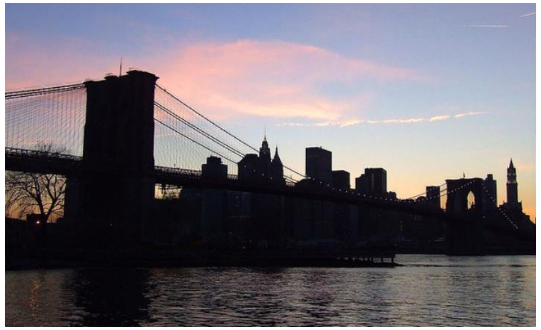


Human intelligence task (HIT)

○ Yes ○ No

○ Yes ○ No

○ Yes ○ No



○ Yes ○ No

○ Yes ○ No

○ Yes ○ No

○ Yes ○ No

○ Yes ○ No

Two fundamental problems

1. Aggregate noisy answers from different workers
2. Incentivize workers to provide high quality answers

Two fundamental problems

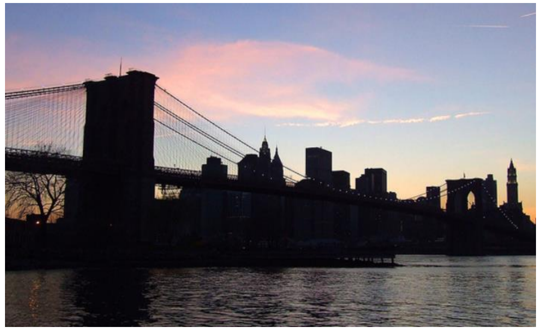
1. Aggregate noisy answers from different workers
2. Incentivize workers to provide high quality answers

Quality control with random gold

○ Yes ○ No

○ Yes ○ No

○ Yes ○ No



○ Yes ○ No

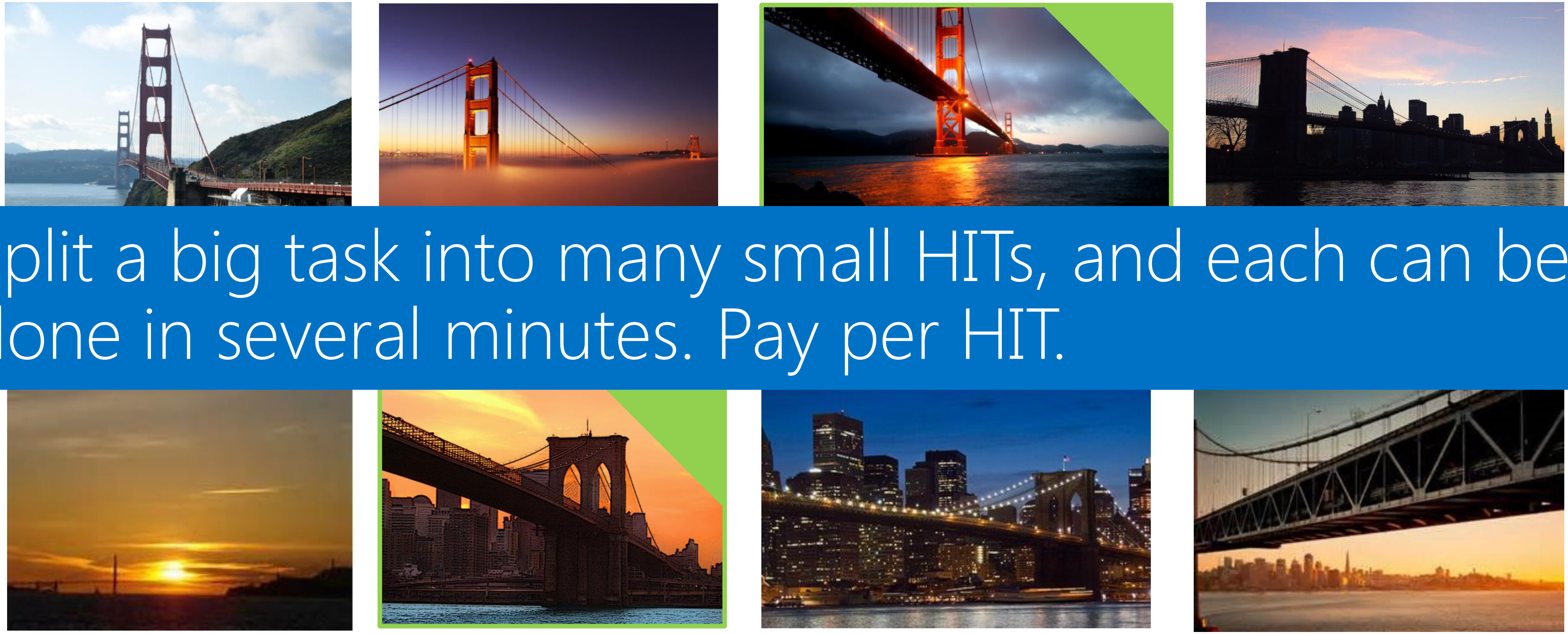
○ Yes ○ No

○ Yes ○ No

○ Yes ○ No

○ Yes ○ No

Quality control with random gold



Split a big task into many small HITs, and each can be done in several minutes. Pay per HIT.

○ Yes ○ No ○ Yes ○ No ○ Yes ○ No ○ Yes ○ No

Our goal

Incentivize human workers to answer questions when they are sure while skip when they are not sure

Everyone can imagine many ways to pay

Case 1: payment proportional worker's accuracy in gold standard questions

Assume 100 images, 4 gold and 1 cent per label. A worker got 1 correct in gold. Then his payment is: $(\mathbf{100 \times 1}) \times \frac{\mathbf{1}}{\mathbf{4}} = \mathbf{25\ cents.}$

Case 2: full payment if accuracy in gold not less than a specified number, and zero otherwise

Assume the number = 60%. Then the above worker will receive 0 payment.

We will show a much better way, which is unique under two basic requirements

Intuition: interest conflict in payment

Crowdsourcing workers want to receive maximum payment using minimum effort

Crowdsourcing requesters want to receive maximum quality work with minimum cost

A good mechanism should resolve the conflict

> “ The best language that mankind has developed for stating things clearly and precisely is mathematics.”
>
> Leslie Lamport
(Turing Award 2013)

Fixed threshold T chosen in (0,1). For every question, we wish to incentivize worker to:

(1) Skip if confidence is less than T

(2) Otherwise, select answer he believes is most likely to be correct

Requirement 1: Incentive Compatible

Fixed threshold T chosen in (0,1). Worker maximizes his expected payment if and only if:

(1) Skip if confidence is less than T

(2) Otherwise, select answer he believes is most likely to be correct

Requirement 1: Incentive Compatible

Is this the Golden Gate Bridge?

○ Yes

○ No

○ I don’t know

Assume choosing T = 60%

I think there’s a 50% chance that I’m correct so I should skip

I think there’s a 90% chance that I’m correct so I should answer

Requirement 1: Incentive Compatible

For any worker, if all his attempted answers to gold are wrong, he should receive zero payment

Requirement 2: No-Free-Lunch

We need to find a mechanism to satisfy the two requirements

Our mechanism: “double-or-nothing”

Let $C =$ number of correct answers, $W =$ number of wrong answers

```
if $W > 0$
        payment $= 0$
else
        payment $= \kappa \frac{1}{T^C}$
```

for some predefined constant $\kappa > 0$, and confidence threshold $T \in (0, 1)$

Our mechanism: an example

Assume: 20 images and 5 gold

Payment rules

- You start with 1 cent (constant $\kappa = 1$)
- For each correct answer, pay doubles (threshold $T = 0.5$)
- If any answer is wrong, becomes zero
- Marking “I don’t know” does not affect the pay

○ Norwich Terrier

○ Norfolk Terrier

○ Irish Wolfhound

○ I don’t know

Our mechanism: an example

Assume: 20 images and 5 gold

constant $\kappa = 1$, threshold $T = 0.5$

Worker A. 2 are correct, and 3 "I don't know"(skip):

$$\text{payment} = 1 \times \underbrace{2 \times 2}_{\text{correct}} \times \underbrace{1 \times 1 \times 1}_{\text{skip}} = 4 \text{ cents}$$

Worker B. 2 are correct, 2 "I don't know", and 1 wrong :

$$\text{payment} = 0 = 1 \times \underbrace{2 \times 2}_{\text{correct}} \times \underbrace{1 \times 1}_{\text{skip}} \times \underset{\text{wrong}}{0}$$

○ Norwich Terrier

○ Norfolk Terrier

○ Irish Wolfhound

○ I don't know

Any other mechanism also satisfying these two requirements ?

Any other mechanism also satisfying these two requirements ?

NO!

Theorem Our mechanism is the only mechanism to be incentive compatible and no-free-lunch

Theorem Among all incentive compatible mechanisms, our mechanism pays the minimum amount to spammers

Choosing parameters in practice

1. Quality requirement (confidence T)
2. Number of gold standard questions
3. Size of HITs
4. Initial payment (constant κ)

Extension: multiple confidence levels

Is this the Golden Gate Bridge?

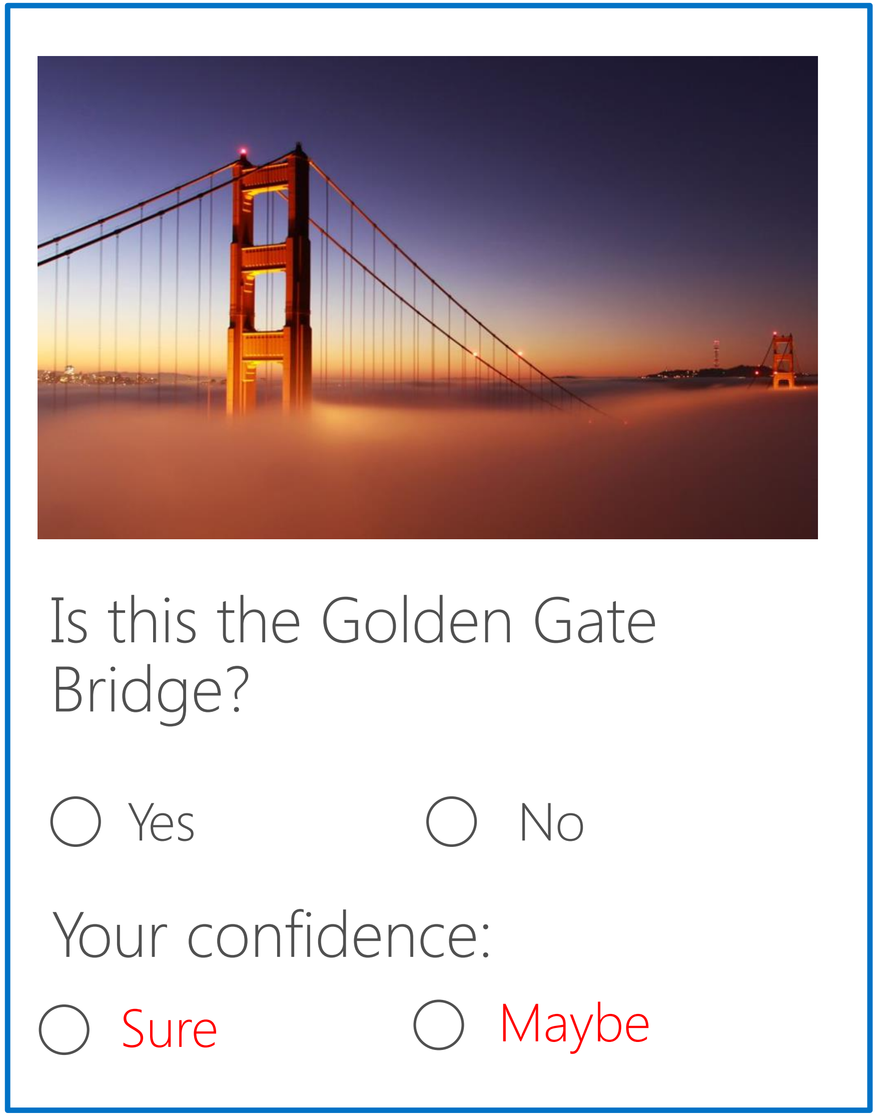
○ Yes ○ No

Your confidence:

○ Sure ○ Maybe

(Shah and Z 2014)

Extension: multiple confidence levels

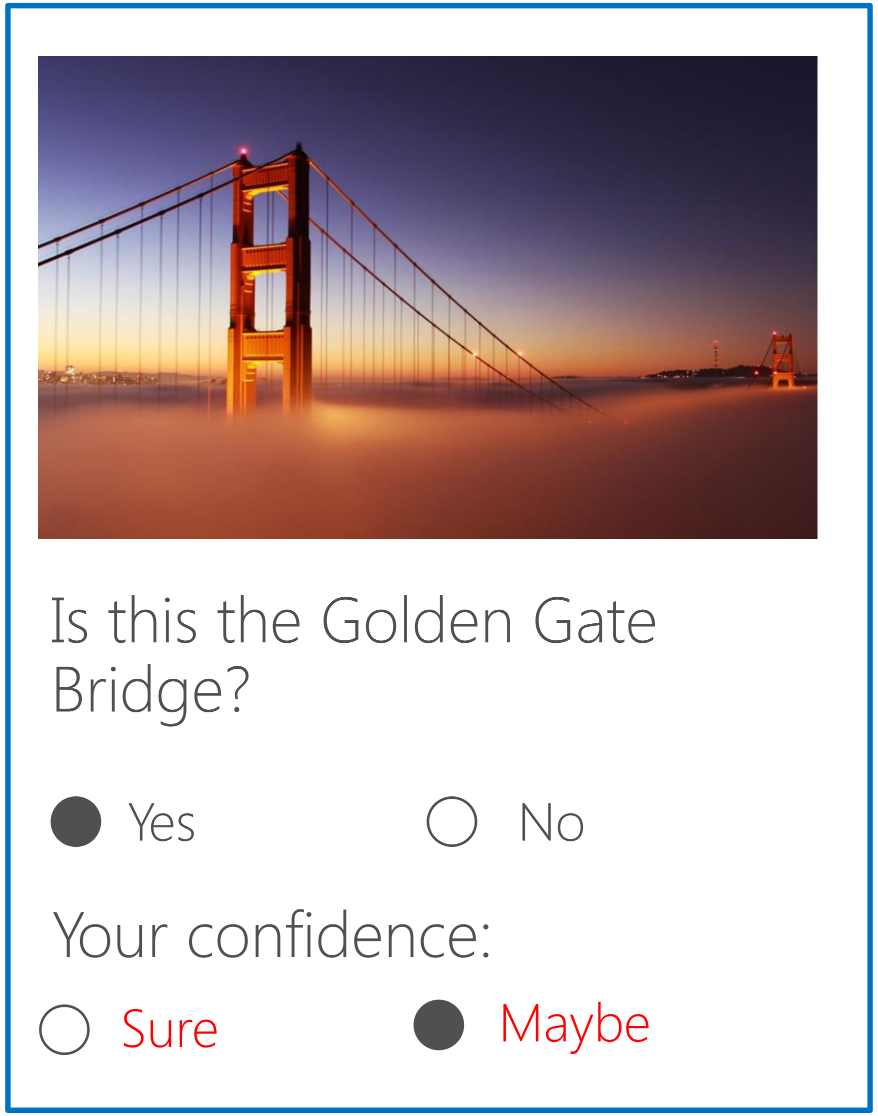


Is this the Golden Gate Bridge?

○ Yes ○ No

Your confidence:

○ Sure ○ Maybe



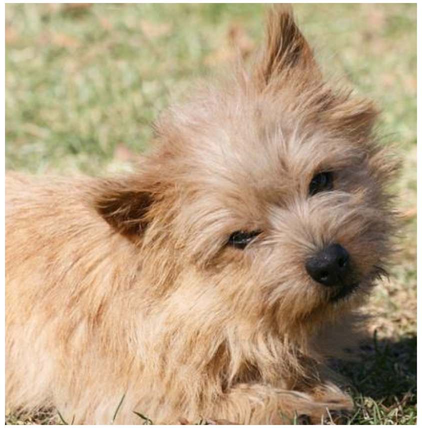
Is this the Golden Gate Bridge?

● Yes ○ No

Your confidence:

○ Sure ● Maybe

Extension: approval voting



○ Norwich Terrier

○ Norfolk Terrier

○ Irish Wolfhound

Note: Mark all possible answers

(Shah, Z and Peres 2015)

Extension: approval voting

○ Norwich Terrier
○ Norfolk Terrier
○ Irish Wolfhound

Note: Mark all possible answers

● Norwich Terrier
● Norfolk Terrier
○ Irish Wolfhound

Note: Mark all possible answers

Experiments

Evaluated mechanisms

Baseline mechanism

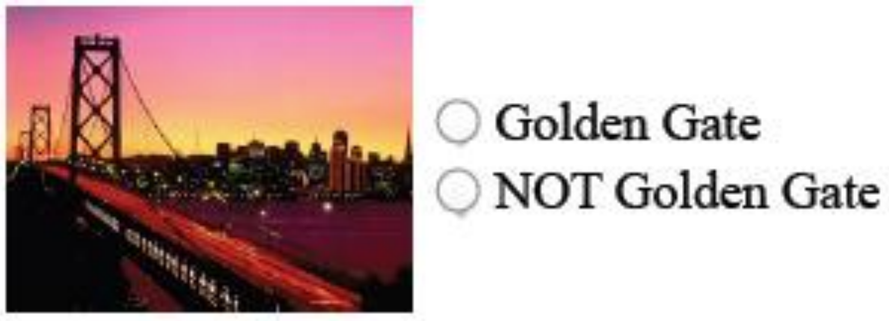
Payment proportional to the number of correct answers to gold

Skip-based mechanism

Confidence-based mechanism

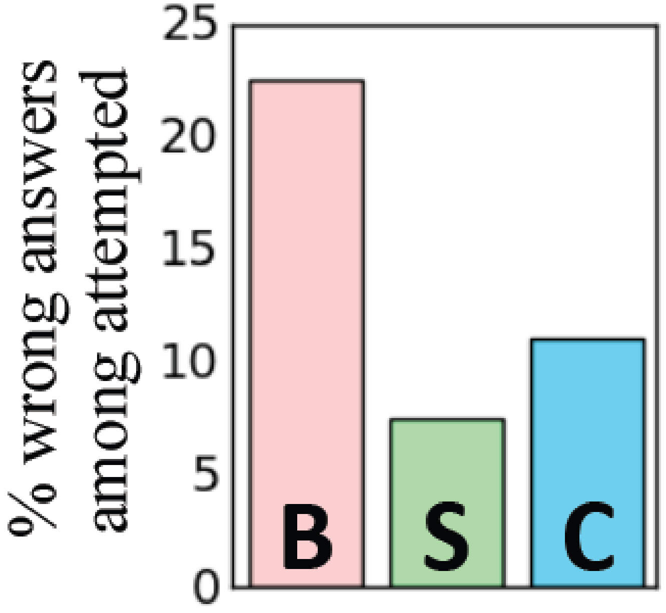
Equal budget for each mechanism

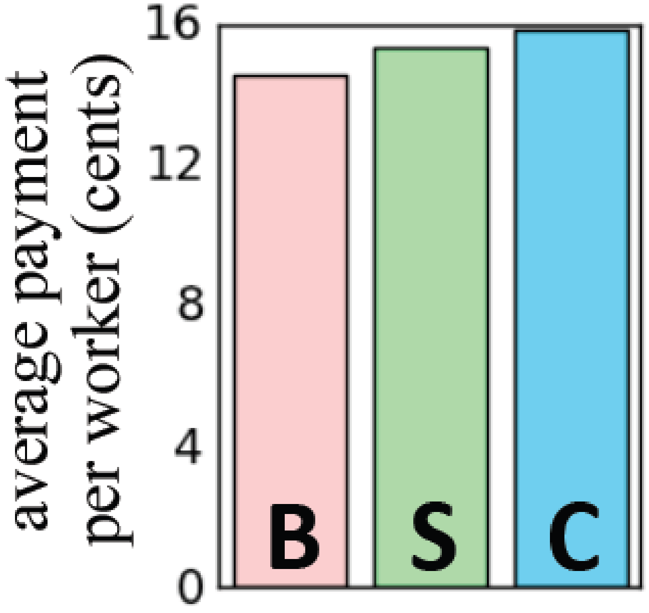
Recognize Godden Gate Bridge



○ Golden Gate

○ NOT Golden Gate

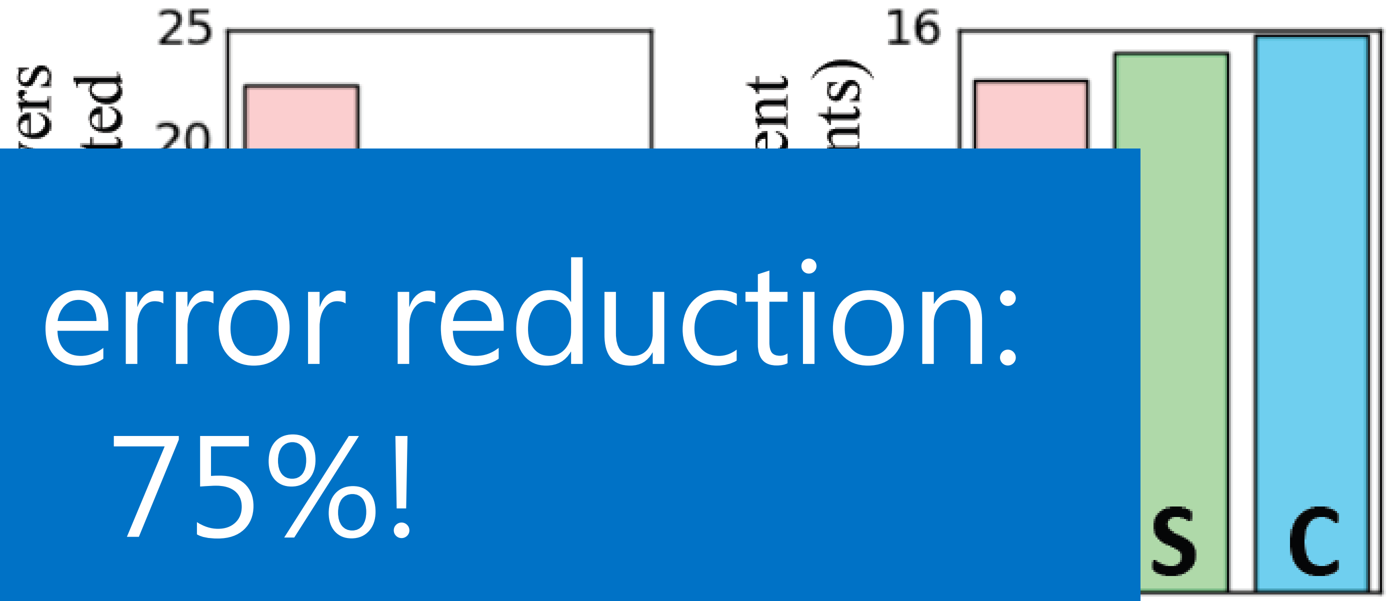




21 images and 3 gold

20% “I don’t know”

Recognize Godden Gate Bridge

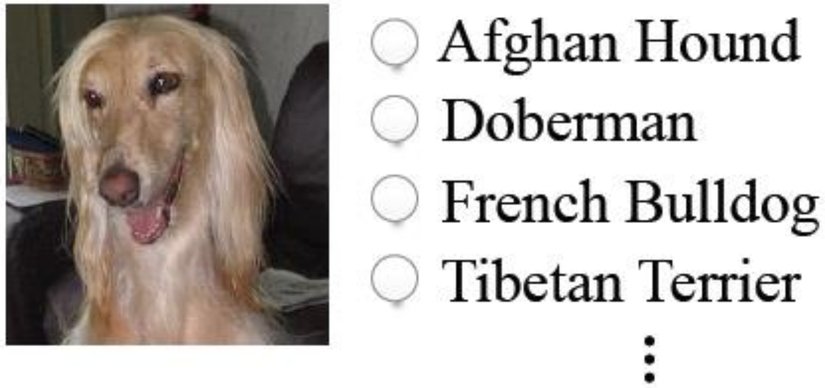


Relative error reduction: 75%!

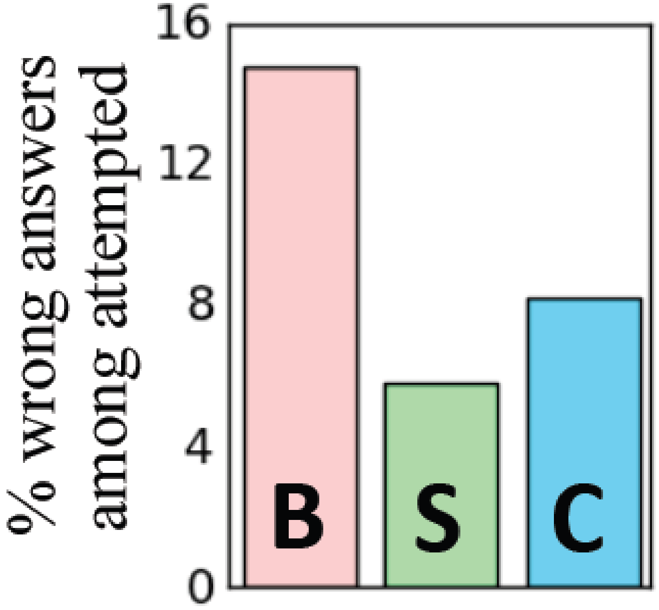
21 images and 3 gold

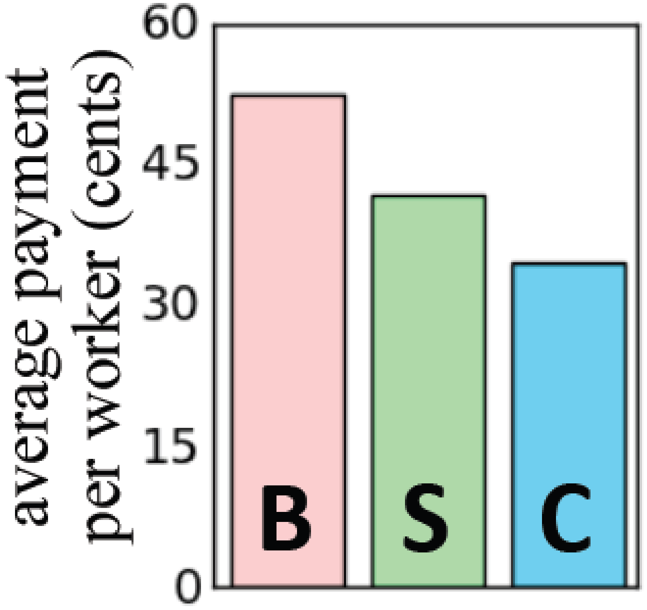
20% “I don’t know”

Mark the breed of the dog



○ Afghan Hound
○ Doberman
○ French Bulldog
○ Tibetan Terrier
⋮

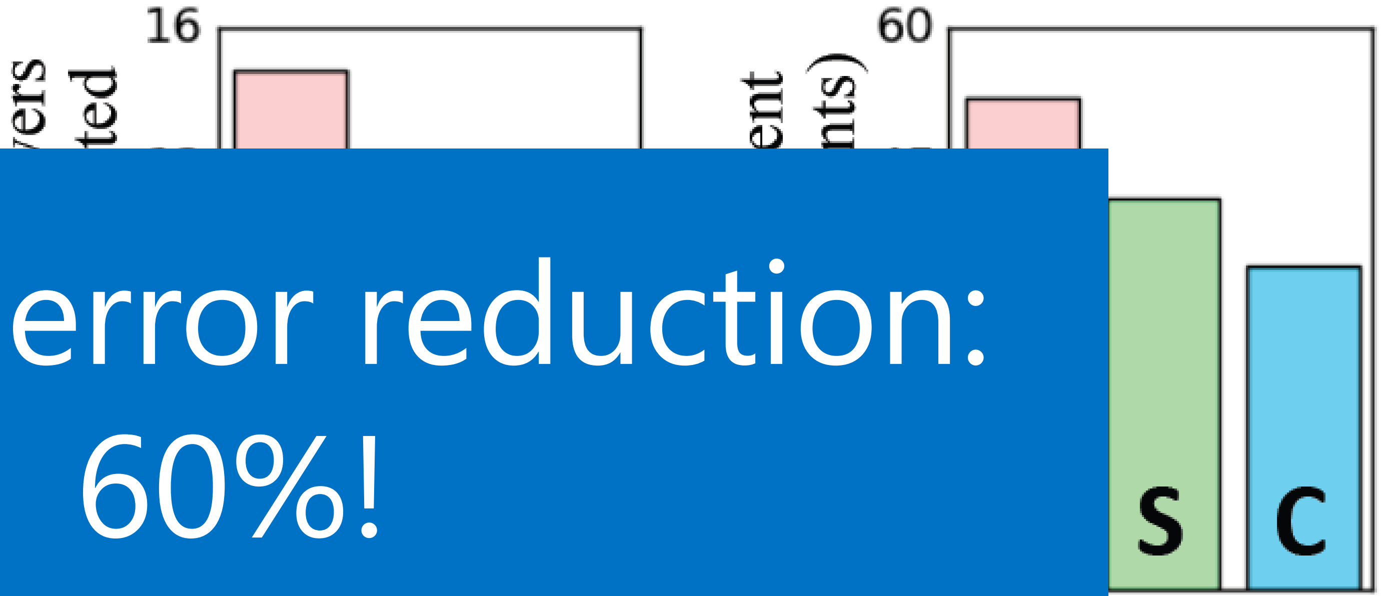




85 images and 7 gold

25% “I don’t know”

Mark the breed of the dog

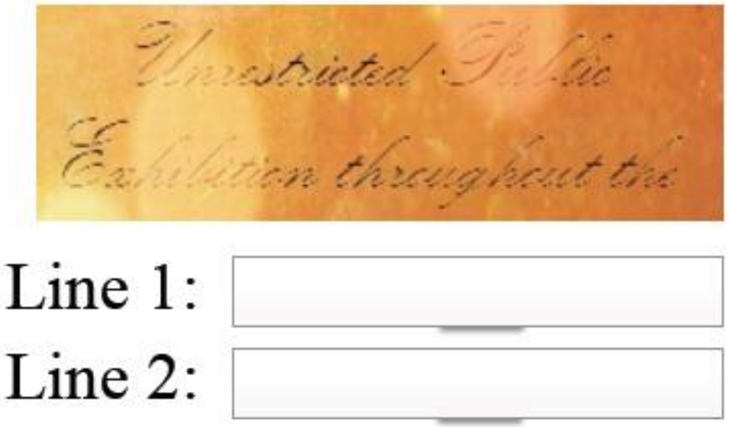


Relative error reduction: 60%!

85 images and 7 gold

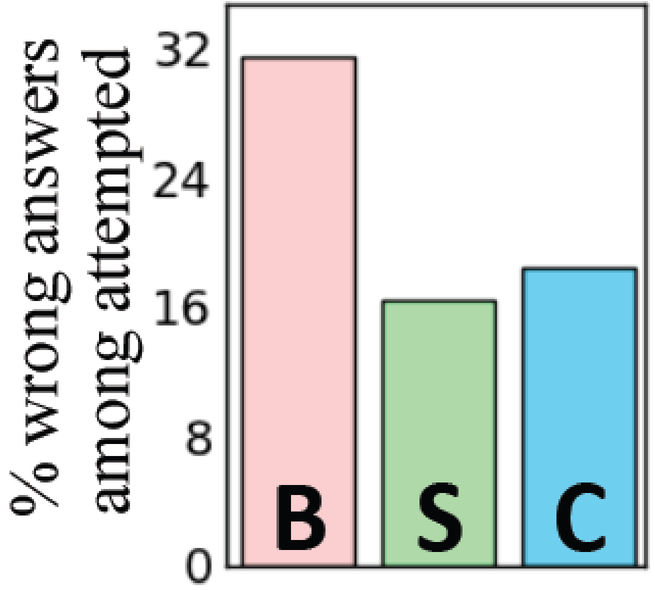
25% “I don’t know”

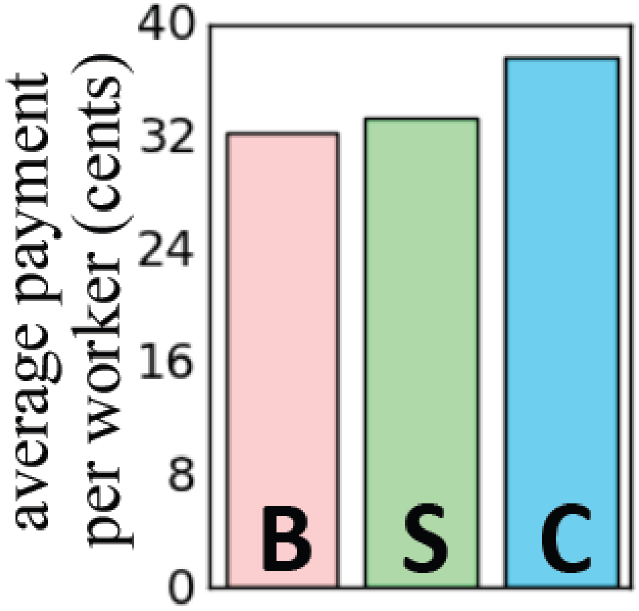
Transcribe text



Line 1:

Line 2:

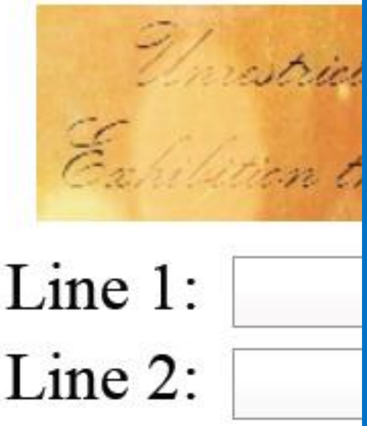




12 images and 2 gold

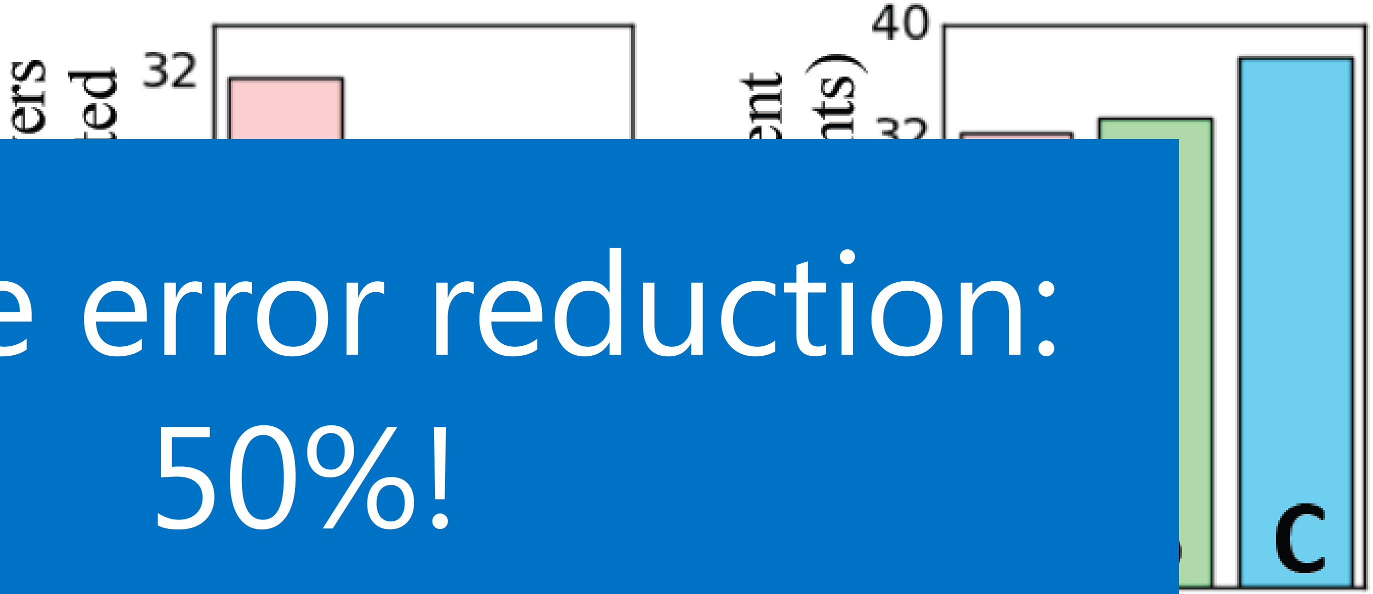
25% “I don’t know”

Transcribe text



Line 1:

Line 2:



Relative error reduction: 50%!

12 images and 2 gold

25% "I don't know"

Conclusion

- Incentive compatible + no-free-lunch = our double-or-nothing mechanism
- Extension: Multi-level confidence, approval voting

Project site: http://research.microsoft.com/en-us/projects/crowd

Microsoft