Lecture 6: Introduction to Techniques

Instructor: Chien-Ju (CJ) Ho

Logistics

- Project Proposal Due **Feb 12 midnight**
  - Team members (2~3 persons)
  - 1~2 paragraph description of the proposal
  - Cite at least one relevant paper
  - Submit through gradescope (Will be set up before early next week)

- You have the chance to change the topic before the first milestone at **March 5**

On Tuesday

- Discuss game theory, scoring rules, peer predictions.
- Designing incentives to ensure workers contribute useful/honest/high-quality information.

- Typical flow of applying game theory in crowdsourcing
  - Formulate users' incentives
  - Describe the game structure
    - Sometime there are no worker interactions. It becomes a simpler optimization problem.
  - Analyze the equilibrium as the prediction of the outcome

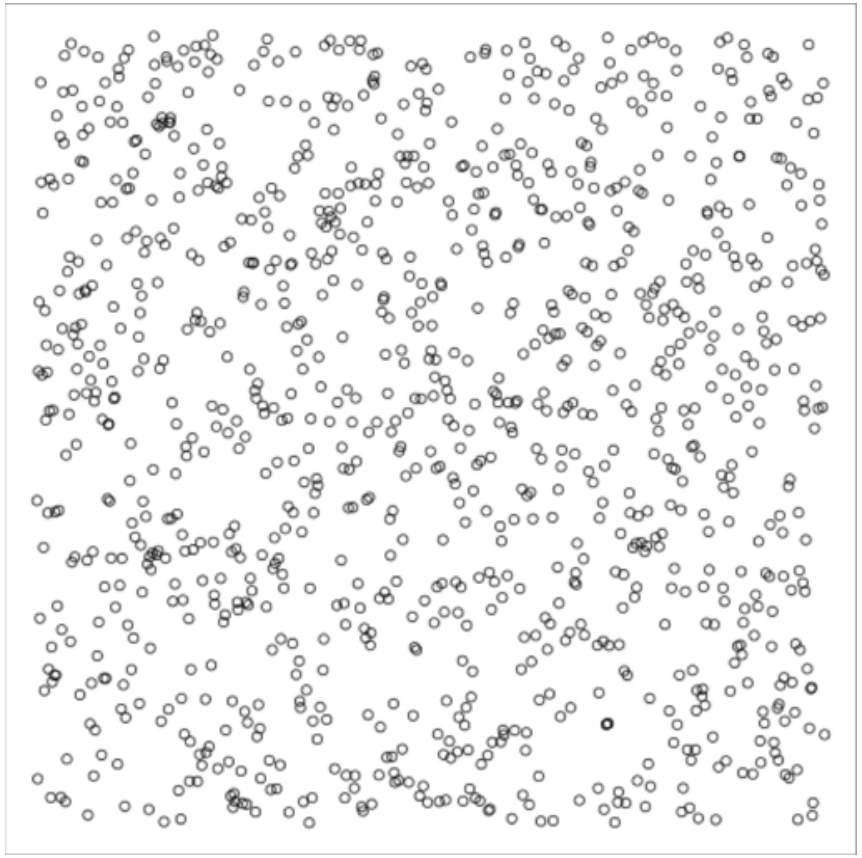
- Potential conference of interests
  - ACM Conference on Economics and Computation (EC)

Today’s Lecture

- Label aggregation in crowdsourcing
  - (Weighted) Majority Voting
  - Maximum likelihood estimation
  - Concentration bounds

Remember this task?

- How many circles are in the image



These are the answers from you!

167	864	1500
187	884	1600
468	960	1600
500	960	1999
600	963	2000
600	999	2500
720	1000	3300
800	1320	10000
800	1500	

How should we combine these numbers to make the final prediction?

Modeling the answer generation process

- A naïve model

  user-answer = true-answer + noise,

  where noise is drawn from some distribution with mean 0

- How should we aggregate if we believe this is how answers are generated.

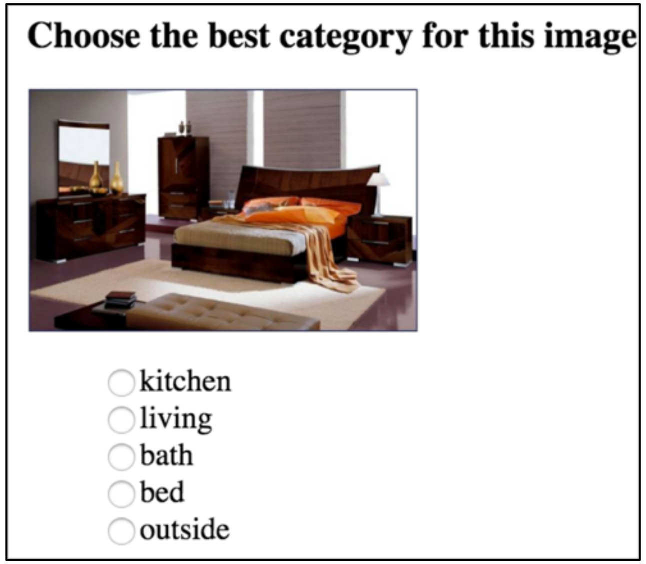
- Can we obtain any sort of theoretical guarantee?
  - how many answers do we need to ensure the aggregation is close to the true answer with high probability?

- Is this a reasonable model?
  - Probably not for your answers from lecture 2

Focus on a common task: categorization

- (Binary) categorization tasks

- Most techniques/results can be extended to multi-label case, but the presentation could be a lot more complicated.



How should we model the label generation process?

A simple model

- Without loss of generality, assume the label is either -1 or +1

- Each worker has the same ability of the giving correct label

$\Rightarrow$Each worker gives the correct label ***with probability p***

- Assume we believe this is how labels are generated, what is your final prediction if we collect the following labels for a task?

{-1, +1, +1, -1, +1}

Majority voting (MV) seems to be the way to go

Q1: Why MV might be a good idea?

Q2: Can we obtain theoretical guarantees for majority voting?

Understanding this simple scenario helps us develop aggregation methods for more complicated scenarios.

Why Majority Voting:
Majority Voting Gives the Maximum-Likelihood Estimation

- Consider a task with true label l^*
- We collect labels $L = \{l_1, l_2, \ldots, l_n\}$ from n workers for this task.
- Each worker gives the correct label with probability $p > 0.5$.

- l^* is the latent variable and L is our observation.

- Maximum likelihood estimation (MLE):
  - Predict +1 if $\Pr[\, L | l^* = +1] \geq \Pr[\, L | l^* = -1]$
  - Predict -1 otherwise

Likelihood: $\Pr[D|\theta]$
D: Observations
θ: latent variables

MLE approach (roughly speaking):
Find $\theta^* = argmax_\theta \Pr[D|\theta]$

Derivation of MLE ⇔ MV (details on the board)

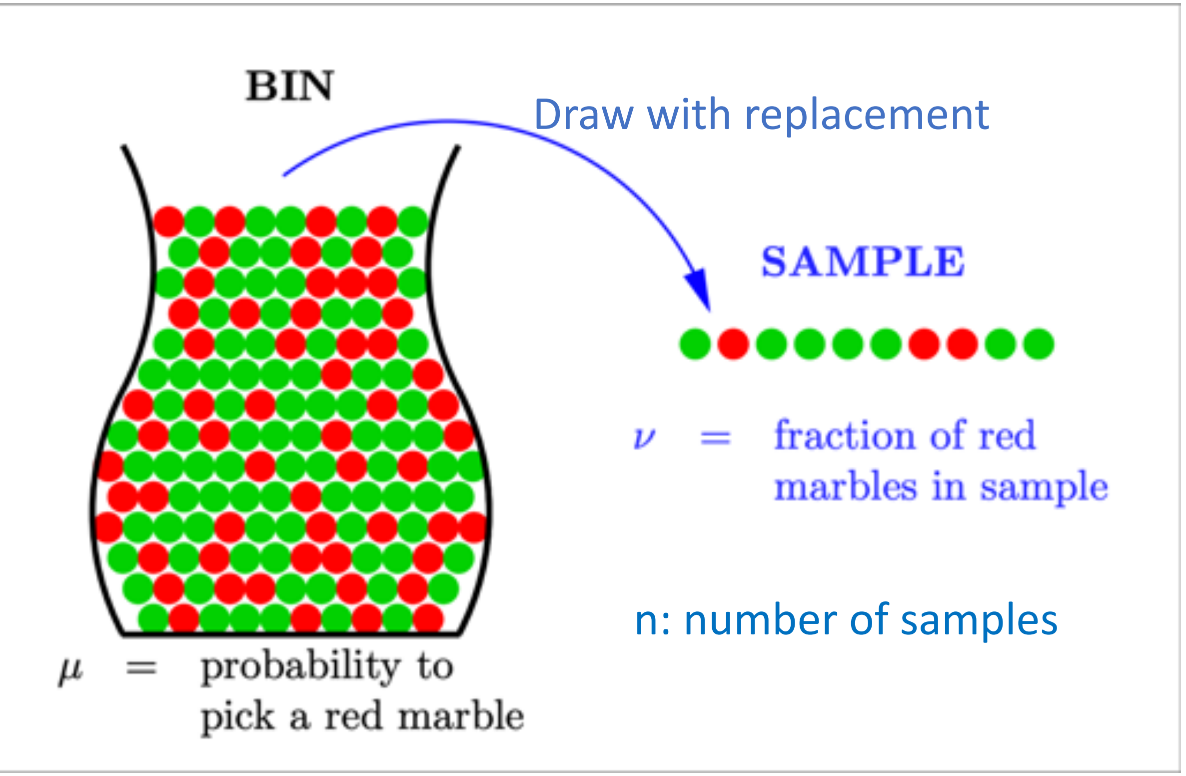
- Let (n_+, n_-) be the number of (+1, -1) labels in L
- $\Pr[\, L | l^* = +1] = p^{n_+}(1-p)^{n_-}$
- $\Pr[\, L | l^* = -1] = p^{n_-}(1-p)^{n_+}$

- MLE rule is equivalent to
  - Predict +1 if $\ln \frac{p^{n_+}(1-p)^{n_-}}{p^{n_-}(1-p)^{n_+}} \geq 0$

  - Predict +1 if $(n_+ - n_-)(\ln p - \ln(1-p)) \geq 0$
  - Predict +1 if $n_+ \geq n_-$
  - This is majority voting

What guarantee can MV achieve?

- Consider a thought experiment



What can we say about μ from ν?

Law of large numbers

- When $n \to \infty, \nu \to \mu$

Hoeffding's Inequality

- $\Pr[|\mu - \nu| > \epsilon] \leq 2e^{-2\epsilon^2 n}$ for any $\epsilon > 0$

Interpretations

$$\Pr[|\mu - \nu| > \epsilon] \leq 2e^{-2\epsilon^2 n}$$

Define $\delta = \Pr[|\mu - \nu| > \epsilon]$: Probably of "bad events"

- Fix $\epsilon, \delta = O(e^{-n})$; Fix n,$\delta = O\left(e^{-\epsilon^2}\right)$; Fix $\delta, \epsilon = O(\sqrt{\frac{1}{n}})$
- n=1000
  - $\mu - 0.05 \leq \nu \leq \mu + 0.05$ with 99% chance
  - $\mu - 0.10 \leq \nu \leq \mu - 0.10$ with 99.9999996% chance

- ν is approximately close to μ with high probability

- ν as an estimate of μ is **p**robably **a**pproximately **c**orrect (P.A.C.)

More general form of Hoeffding's inequality

- Let $X_1, \dots, X_n$ be independent random variables
  - X_i is bounded in the range $[a_i, b_i]$

- Let $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

- (One-sided) Hoeffding's inequality

$$\Pr[\mathbb{E}[\bar{X}] - \bar{X} \geq \epsilon] \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

We get our previous bound by setting $b_i = 1$ and $a_i = 0$

Connection to Our Problem

$$\Pr[\mathbb{E}[\bar{X}] - \bar{X} \geq \epsilon] \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

- Without loss of generality, assume $l^* = +1$
- X_i is the random variable of the label provided by worker i

- $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$: $\mathbb{E}[\bar{X}] = 2p - 1 > 0$
- Majority voting => Predict $sign(\bar{X})$

- Probability of making a wrong prediction

$$\begin{aligned}
\Pr[\bar{X} \leq 0] &= \Pr\big[\mathbb{E}[\bar{X}] - \bar{X} \geq \mathbb{E}[\bar{X}]\big] \\
&\leq \exp\left(-\frac{1}{2} n \left(\mathbb{E}[\bar{X}]\right)^2\right) \\
&= \exp\left(-\frac{1}{2} n \left(2p - 1\right)^2\right)
\end{aligned}$$

Looks like we solved the problem?

if we assume all workers are the same….

What happens if workers are different

- Assume we obtain n labels from n workers.
- Worker $i \in \{1, \ldots, n\}$
  - provides label $l_i \in \{-1, +1\}$
  - correct with probability p_i
  - assume we know p_i

- How should we aggregate?
  - Weighted majority voting?

Predict $sign(\sum_{i=1}^{n} w_i l_i)$

Weighted Majority Voting

- Weighted majority voting Predict $sign(\sum_{i=1}^{n} w_i l_i)$
- Turns out weighted majority voting leads to MLE
  - With weight $\mathrm{w_i} = \ln \frac{p_i}{1-p_i}$ for label l_i
  - Proof on the blackboard
- The weights to minimize the Hoeffding error are different
  - To minimize Hoeffding error, set weights $w_i = 2p_i - 1$ for label l_i
  - Proof on the blackboard (Lemma 1 in Ho et al. ICML 2013)

Can we really know workers' abilities?

What if tasks are different as as well?

Many there are more factors we should consider?

Typical label aggregation approach

Likelihood: $\Pr[D|\theta]$
D: Observations
θ: latent variables

- Propose a model to describe the label generation process
- True labels are the "latent variables" of the process
- Using inference algorithms to learn the latent variables

Feb 26	Label Aggregation: EM-based Algorithms Presenter: Jananthan and Thomas	**Required** Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. Whitehill et al. NIPS 2009. **Optional** Learning from Crowds. Raykar et al. JMLR 2010. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Dawid and Skene. Applied Statistics. 1979.
Feb 28	Label Aggregation: Matrix-based Methods Presenter: Han, James, and Gan	**Required** Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content. Ghosh, Kale, and McAfee. EC 2011. **Optional** Budget-Optimal Crowdsourcing using Low-rank Matrix Approximations. Karger, Oh, and Shah. Allerton 2011. Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing. Zhang et al. JMLR 2016.
Mar 5	Label Aggregation: Belief Propagation and Others Presenter: CJ	**Required** Variational Inference for Crowdsourcing. Liu, Peng, and Ihler. NIPS 2012. **Optional** Iterative Learning for Reliable Crowdsourcing Systems. Karger, Oh, and Shah. NIPS 2011. Learning from the Wisdom of Crowds by Minimax Entropy. Zhou et al. NIPS 2012.

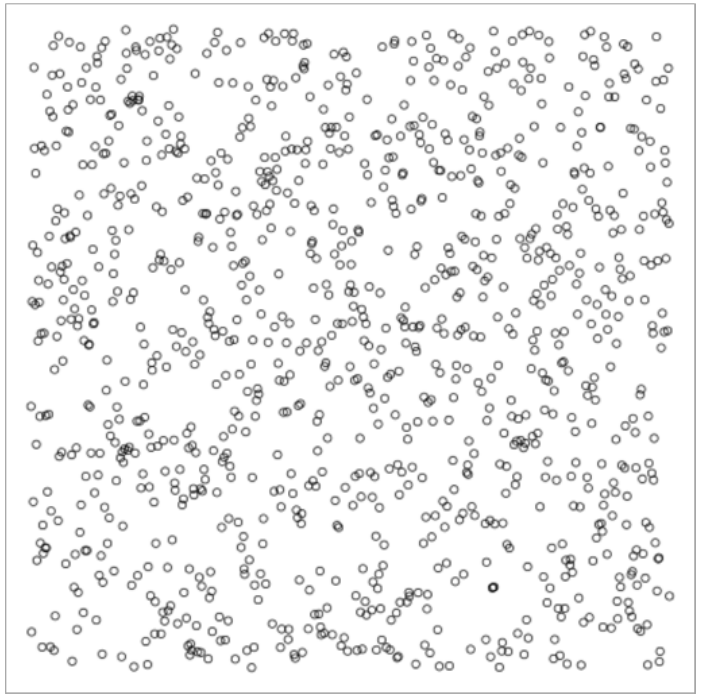
Write down likelihood function
Using EM algorithms to find MLE

Write labels as a matrix (worker by task)
Using low rank matrix approximation

A bunch of other methods

Recall the discussion question in lecture 2

- How should we model the crowdsourcing process for this task?



How many circles are in the image

Modeling the incentive structure

- Tell us users' actions

Modeling label generation process

- Help us aggregate the labels

Most studies have separately discussed them

In case we still have time…

Designing data-elicitation interfaces

Eliciting Categorical Data for Optimal Aggregation

Joint work with

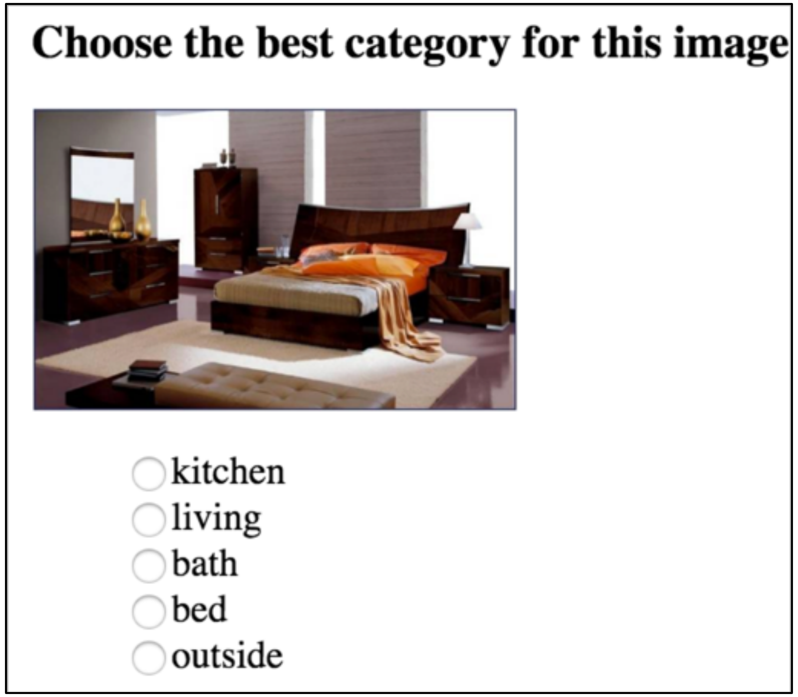
Rafael Frongillo
University of Colorado Boulder

Yiling Chen
Harvard University

In NIPS'16

Label Collection for Classification

d



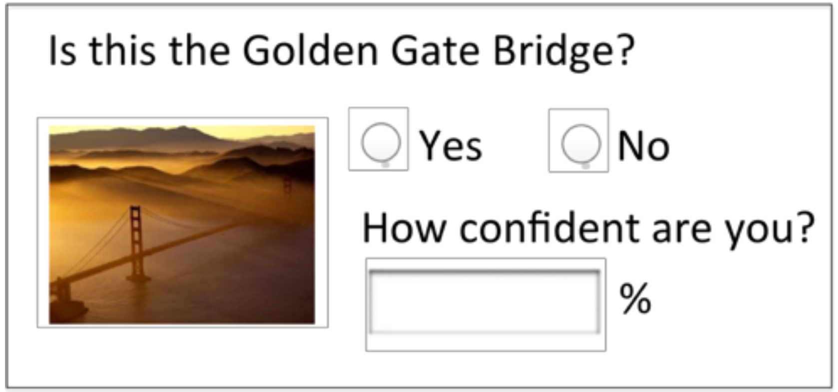
Human Make Mistakes…

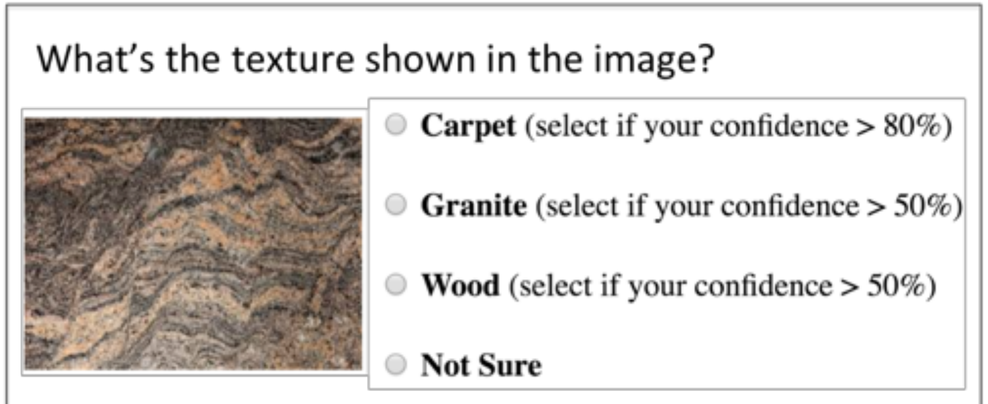
- Repeated sampling
  - Each item is labeled by multiple workers
  - Each worker is asked to label multiple items

- Apply various machine learning techniques to estimate **worker skills** and **item labels**.
  - EM, variational inference, minimax entropy, etc

Workers might know more…

- Workers might know how good their labels are.
- Can we elicit and utilize this information? How?





Research Questions

- How to truthful elicit workers' confidences?

  Proper scoring rules.

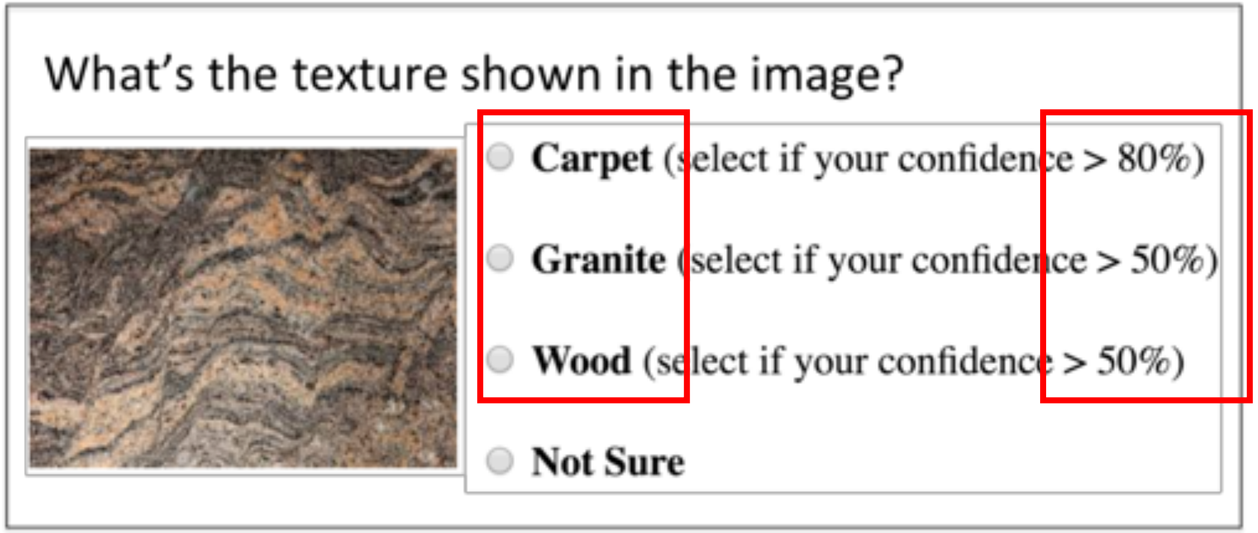
- How to aggregate the labels given confidences?

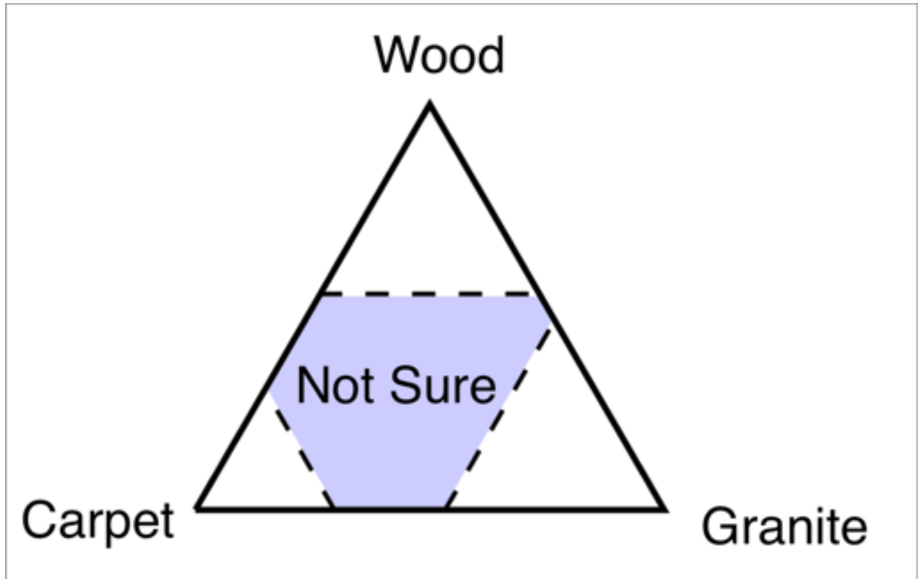
  Bayesian approach, but with similar flavor to MLE

- What's the optimal "interfaces" for eliciting workers' confidences?

  Frame an optimization problem using the above two results.

Threshold Belief Partitions



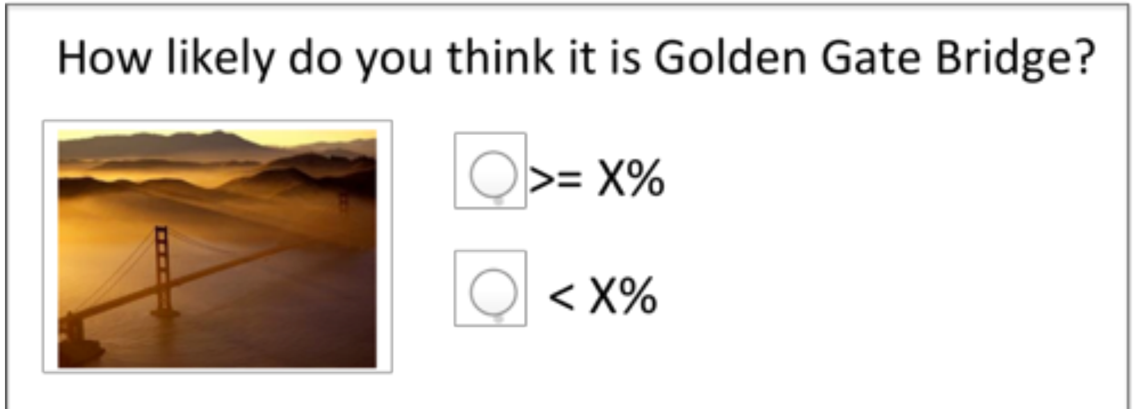


Example: Binary Setting

- Prior: 80% of the images contain Golden Gate Bridge
- Standard design

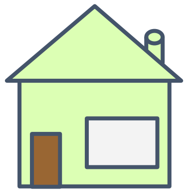
- Threshold belief partition

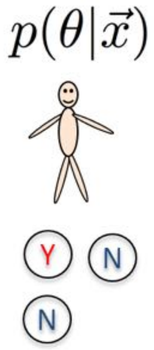
  X=50 ⇔ Standard design

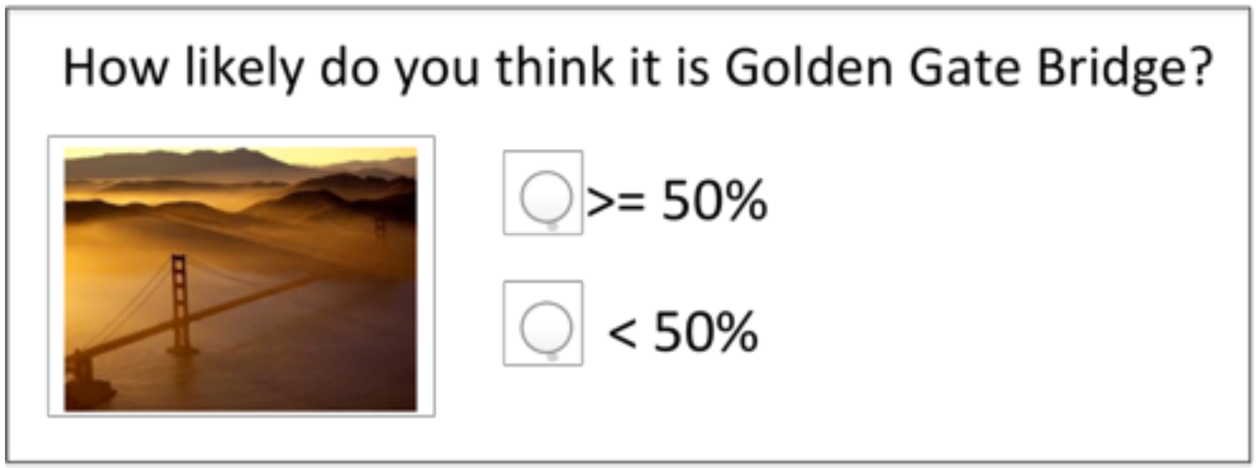


Example: Binary Setting

- Prior: 80% of the images contain Golden Gate Bridge
- If we are eliciting information from **one worker**. What’s the optimal threshold?
  - X = 50
  - Standard design is optimal

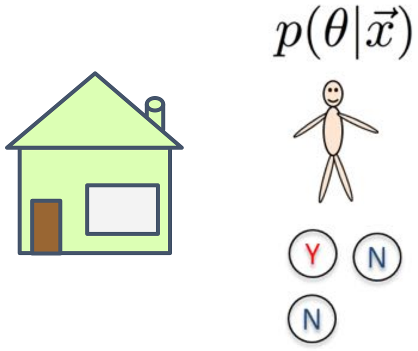






Example: Binary Setting

- Prior: 80% of the images contain Golden Gate Bridge
- If we are eliciting information from **infinitely many workers**. What's the optimal threshold?
  - X = 80
  - Standard design is NOT optimal



$$p(\theta|\vec{x}) \propto p(\theta) \prod p(x_i|\theta)$$

Experiment

- Recruit 200 workers from MTurk
  - Task: identify the texture of blurred images (granite or carpet)
  - Prior: 80% of the images are carpet

- Each worker is given 20 images to label.
  - Bonuses for 5 of the images

Experiment

- Treatments
  - Baseline:
    - 4 cents bonus for answering correctly each of the 5 questions

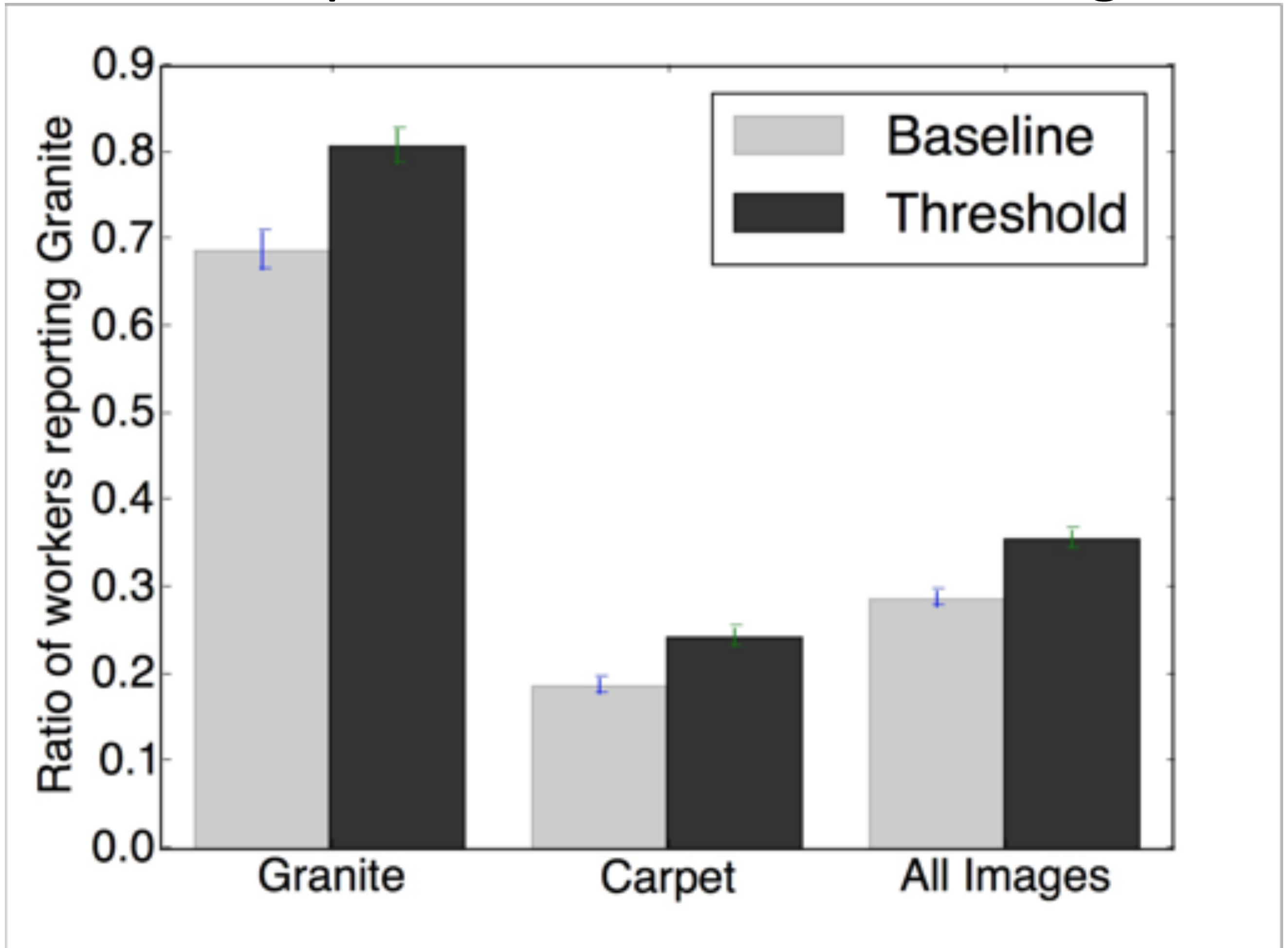
Image	Options
	○ **Carpet** ○ **Granite**

  - Threshold:

How likely do you think the image is **Carpet**
○ **More than 80%** Bonus: 2 cents if the image is **Carpet**, no bonus otherwise ○ **Less than 80%** Bonus: 8 cents if the image is **Granite**, no bonus otherwise

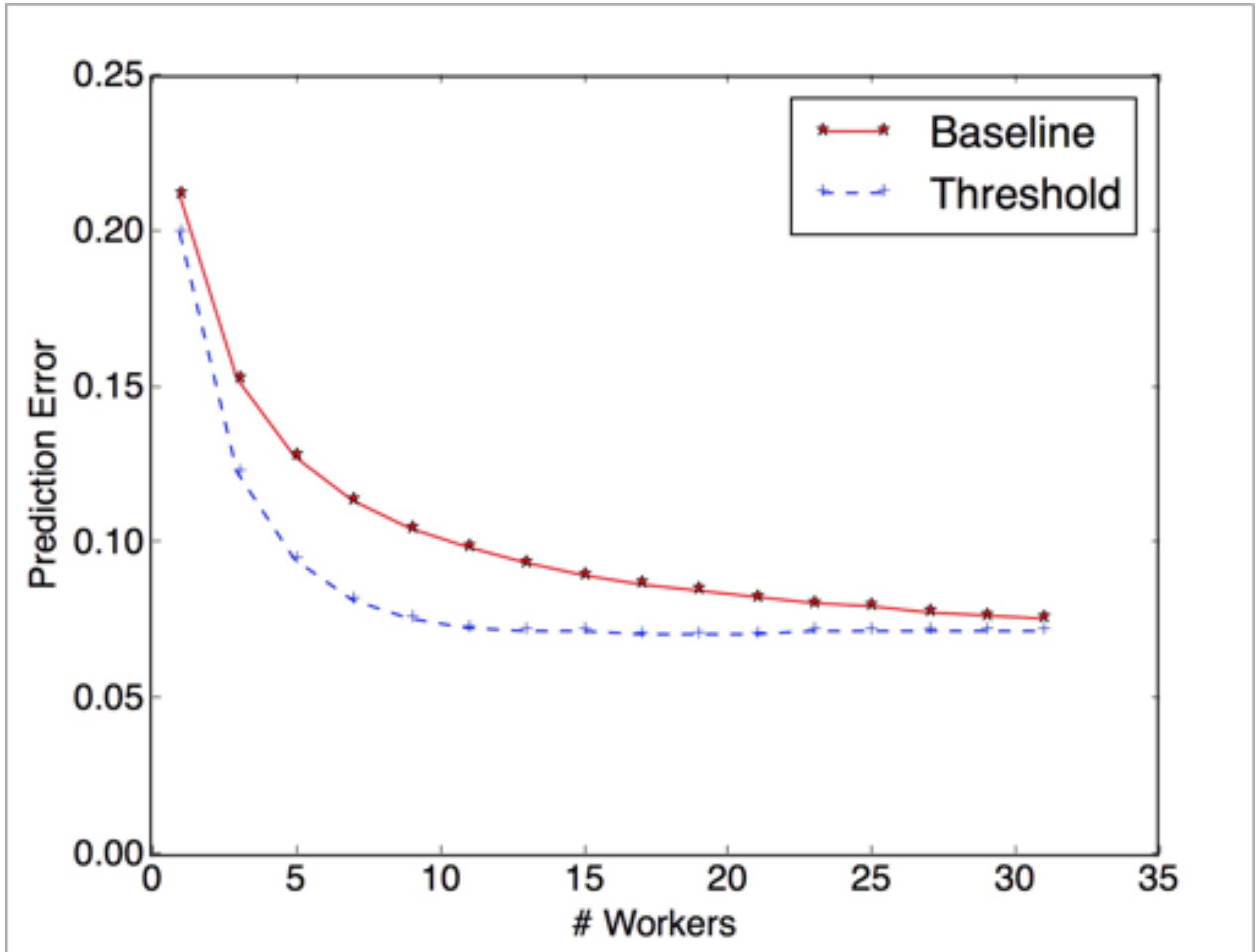
Experiment Results

- Do workers respond to our interface design?



Experiment Results

- Does our interface lead to better aggregation?



Practical Challenge…

- Workers might not respond as expected
  - Details matter!

Respond differently comparing to baseline?

How likely do you think the image is **Carpet**
◯ **More than 80%** Bonus: 2 cents if the image is **Carpet**, no bonus otherwise ◯ **Less than 80%** Bonus: 8 cents if the image is **Granite**, no bonus otherwise

Options
◯ **Carpet** Select this if your confidence is more than 80% Bonus: 2 cents if the image is **Carpet**, no bonus otherwise ◯ **Granite** Select this if your confidence is more than 20% Bonus: 8 cents if the image is **Granite**, no bonus otherwise

Summary

- We propose a Bayesian framework to model the elicitation and aggregation of categorical data.
  - For the full belief setting, we can achieve truthful elicitation and optimal aggregation with an additional sample.
  - For the threshold belief setting, our framework can help find the optimal threshold.

- Gap of theory and practice…