# Exploration of Questioning Strategies for Crowdsourced Labeling

Neil Chainani, Christian Junge, Abhishek Malali

AM 207 Spring 2016

In this project, we contrast two different questioning schemes to assess efficacy in recovering true class labels from noisy labels provided by crowdsourced nonexperts, when there are more than two classes to choose from. We generate data with a confusion matrix for each expert, and an underlying class distribution, and attempt to recover both parameters and the true labels. We use Expectation Maximization (EM), Simulated Annealing, and PyMC to compare efficiency and verify results.

## Introduction

Having reliable labeled data is crucial for any supervised learning task. But given large datasets, manually determining these labels may be intractable. Services like Mechanical Turk can be particularly useful to crowdsource labeling, but it is difficult to guarantee that the workers will reputably provide correct labels, so often the study will aggregate labeling on a particular item from multiple workers.

The task becomes even more difficult with more classes from which to choose, or ambiguity between classes. Take the classic example of lions, tigers, and ligers. An image of a liger may fall closer to a tiger on the lion-tiger spectrum, and thus be hard to classify.

We compare the standard multiclass paradigm (MC), where the worker is presented with an image and has to select a single class, with the Yes-No paradigm (YN), where the worker is presented with the same image K times, is asked "Is this image of class k?", and must answer yes or no.

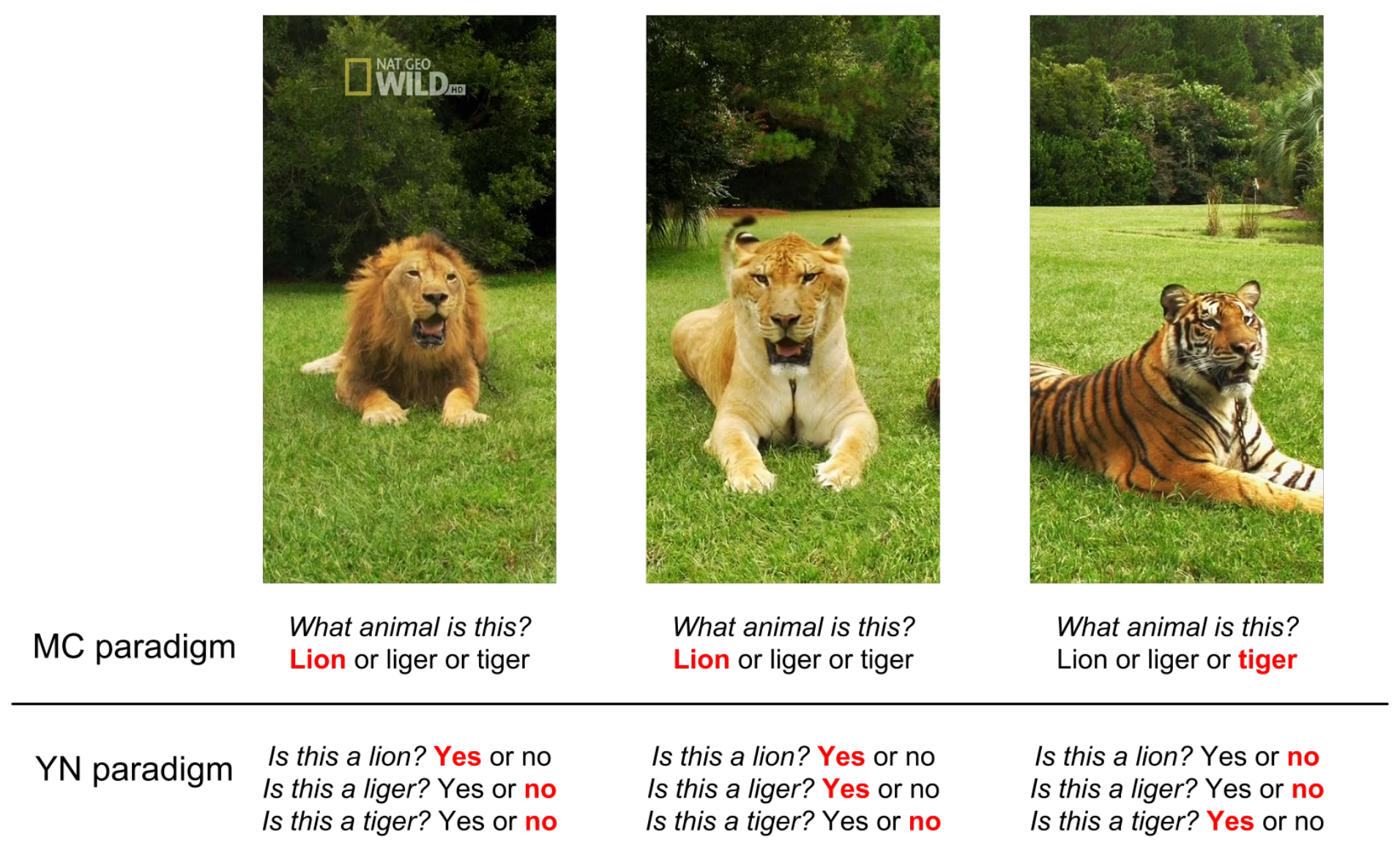
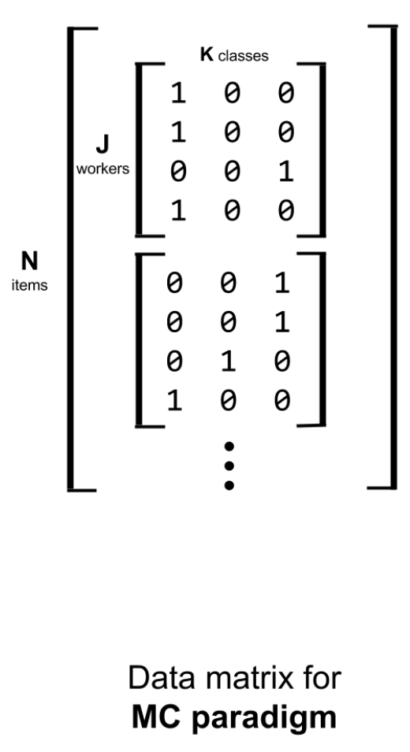

NAT GEO WILD HD
MC paradigm
What animal is this?
Lion or liger or tiger
What animal is this?
Lion or liger or tiger
What animal is this?
Lion or liger or tiger
YN paradigm
Is this a lion? Yes or no
Is this a liger? Yes or no
Is this a tiger? Yes or no
Is this a lion? Yes or no
Is this a liger? Yes or no
Is this a tiger? Yes or no
Is this a lion? Yes or no
Is this a liger? Yes or no
Is this a tiger? Yes or no

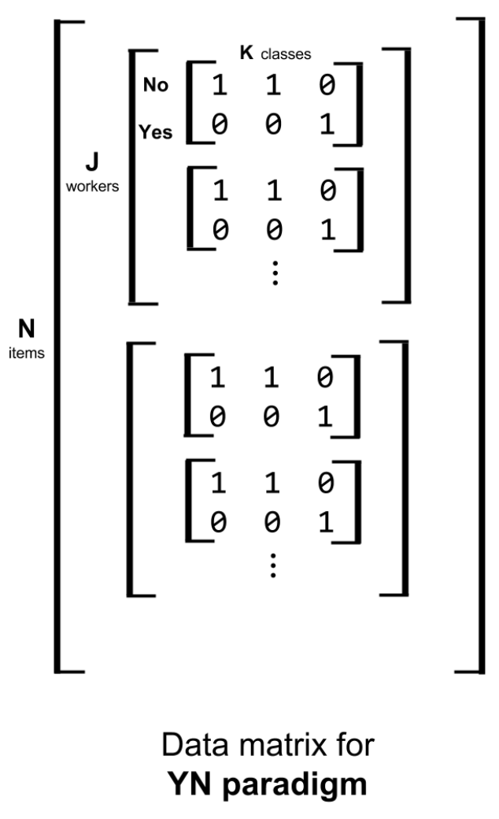
**Figure 1.** Imagine a worker is faced with these three images. They may classify the images of tigers and lions correctly, regardless of the questioning strategy. However, we postulate that their answer with the YN paradigm will allow us to more accurately recover the true class of the liger.

## Data Generation

For both strategies, we begin with a unique confusion matrix of size $K \times K$ per worker, and we predetermine the true class labels for each item. Then, for the MC case, we use a one-hot vector to represent which class a worker picked on a particular item. For the YN strategy, we use a one-hot vector to denote whether they answered yes or no for each class on the item.



Data matrix for **MC paradigm**



Data matrix for **YN paradigm**

## Approach

The complete data log likelihood for the MC case is derived to be:

$$L(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \left( \log \pi_k + \sum_{j=1}^{J} \log \Theta^{(j)}[k, r_{nj}] \right)$$

And for the YN case:

$$L(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \left( \log \pi_k + \sum_{j=1}^{J} \sum_{k'=1}^{K} \left( r_{njk'1} \log \Theta^{(j)}[k, k'] + r_{njk'0} \log(1 - \Theta^{(j)}[k, k']) \right) \right)$$

From here, we first pursued two methods to recover the true class labels:

***Expectation Maximization:***

The E-step in the MC case is:

$$Z_{nk} = \frac{\pi_k \prod_{j=1}^{J} \Theta^{(j)}[k, r_{nj}]}{\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \Theta^{(j)}[k, r_{nj}]}$$

And the M-step to update our estimates of the confusion matrices and class distributions are:

$$\hat{\Theta}^{(j)}[k, k'] = \frac{\sum_{n=1}^{N} Z_{nk} \mathbf{1}(r_{nj} == k')}{\sum_{k'=1}^{K} \sum_{n=1}^{N} Z_{nk} \mathbf{1}(r_{nj} == k')}$$

$$\hat{\pi}_k = \frac{\sum_{n=1}^{N} Z_{nk}}{\sum_{k=1}^{K} \sum_{n=1}^{N} Z_{nk}}$$

We alternate between E and M until convergence.

***Simulated Annealing:***

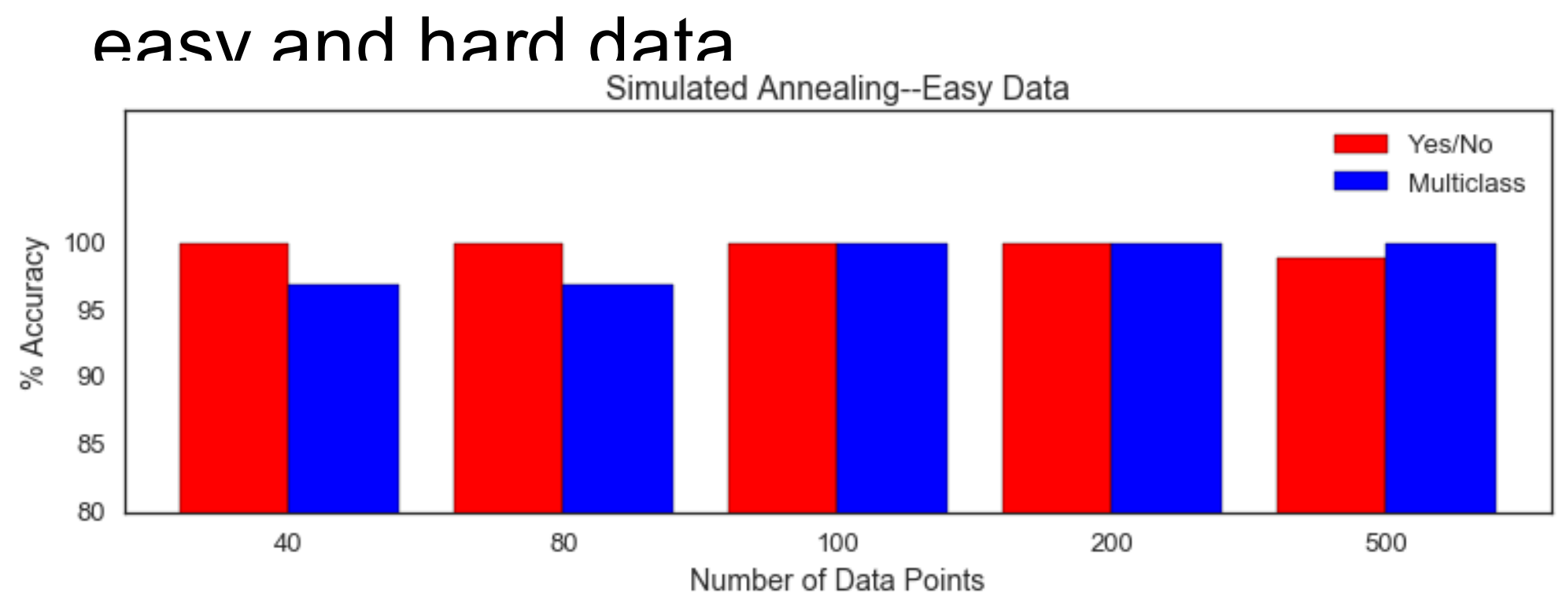
For SA, our states are represented by our estimate of the labels. We initially randomly assign labels to each item, and then reassign labels for a random item. We calculate the likelihood on the set of estimated labels, and determine the confusion matrices based on the counts.

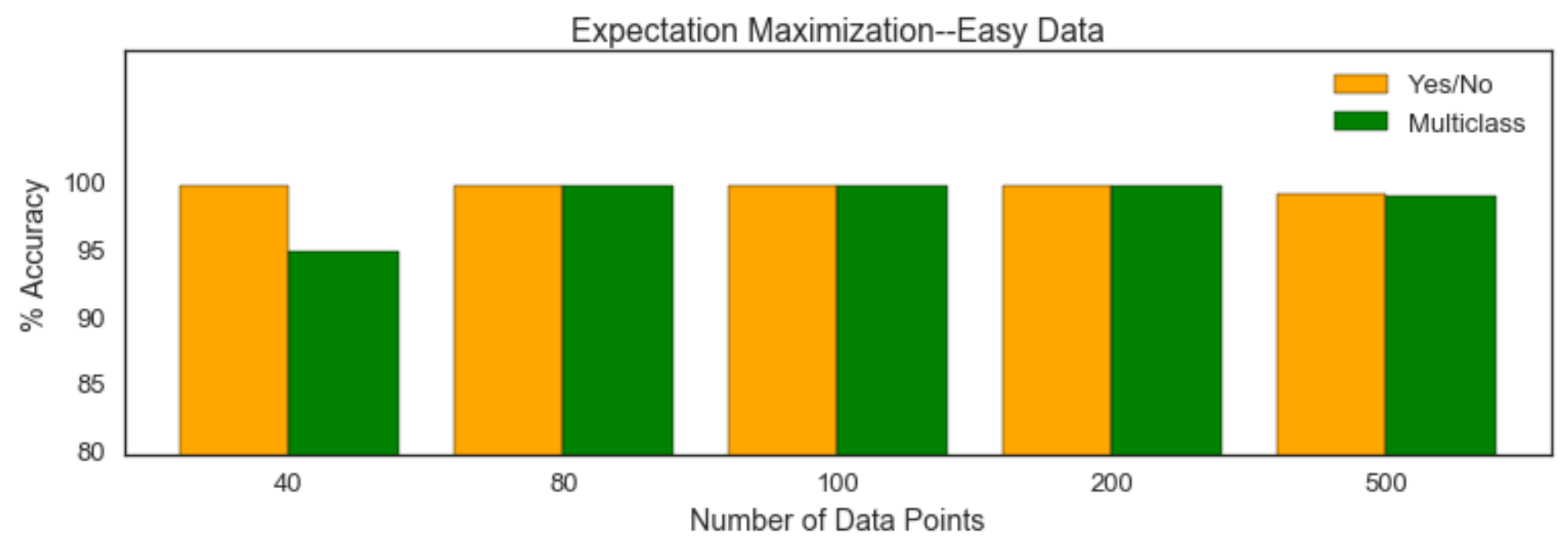
## Conclusions

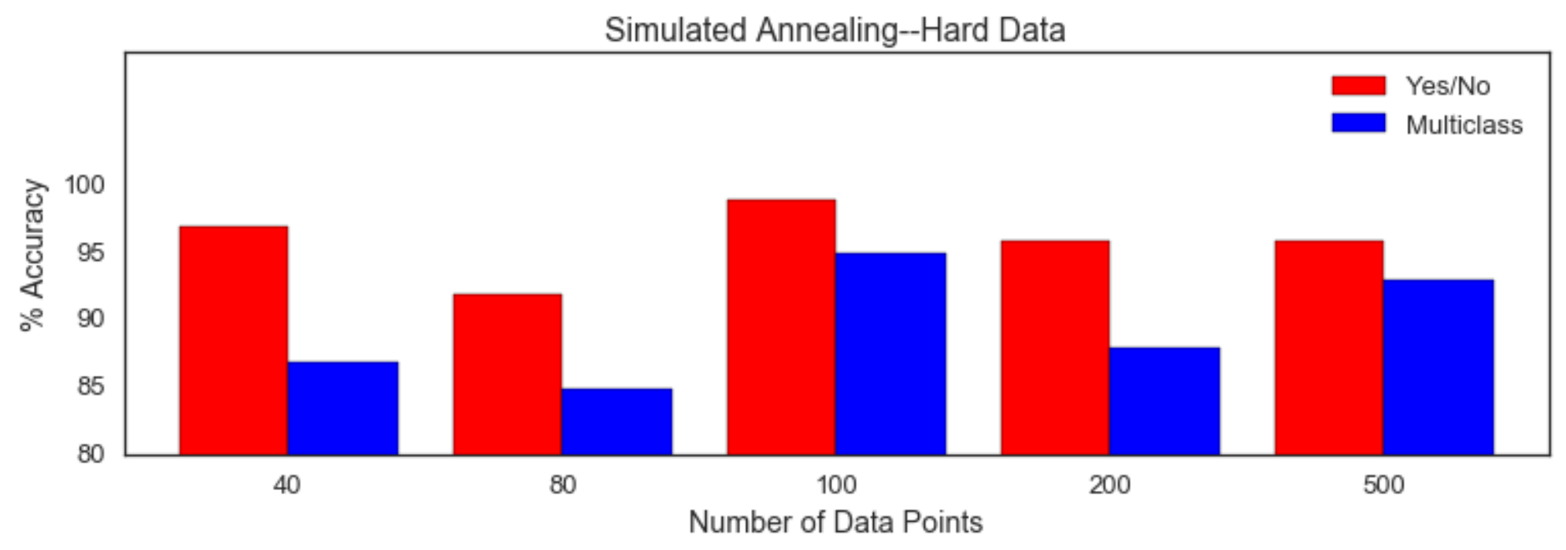
With easy data, regardless of the questioning paradigms we can recover the true labels. However, the differences in accuracy between YN and MC are much more evident on the hard data, where our workers are less precise. Not only can YN consistently recover the true labels with almost 100% accuracy, but with EM it can also capture the confusion matrices for each worker. The implication of this is that we can pick out individual workers who are worse than others. As for differences in methods, EM is much faster than SA; it is able to converge in two iterations (a tenth of a second on average), versus the ten minutes it takes for SA to reach a solution.

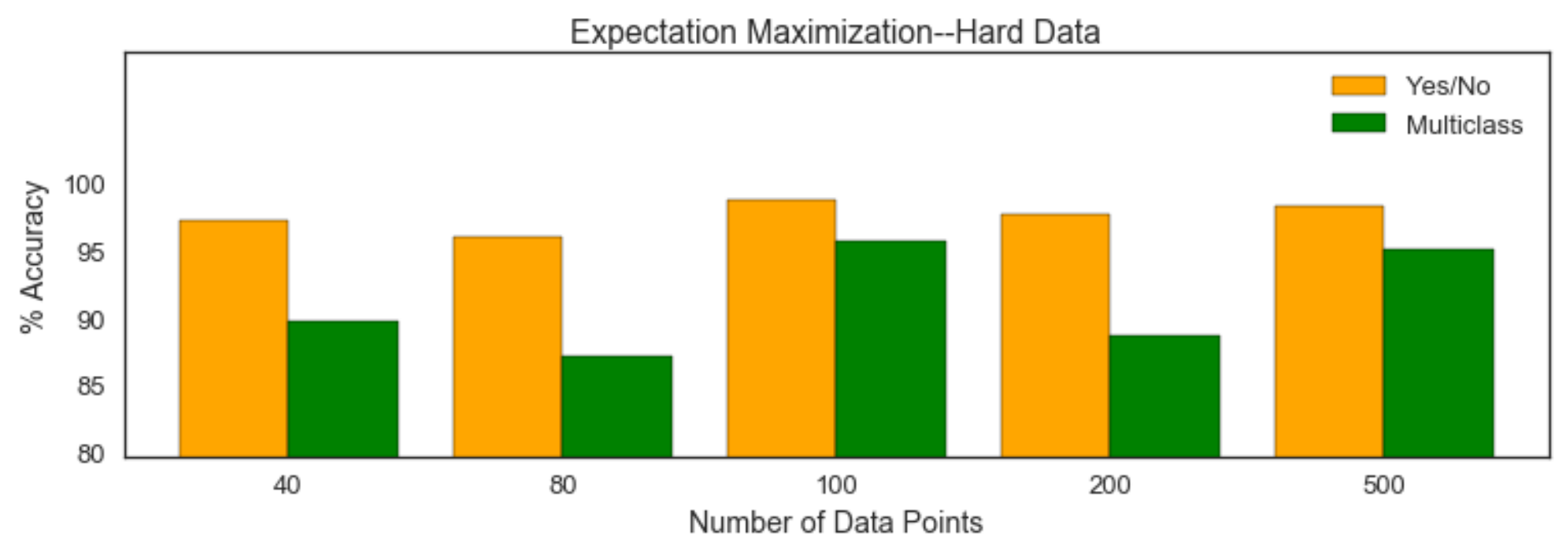
## Results

We used two data sets: an easy data set and a hard data set. Both had 5 workers and 4 classes, but were generated with different confusion matrices. The easy data were created with confusion matrices close to an identity matrix. In contrast, the confusion matrices used to generate the hard data set were much more muddled and varied vastly between workers, representing workers with varying accuracy. We compared overall classification accuracy (the percentage of predicted labels that match the true labels) for 40, 80, 100, 200, and 500 data points, on both the easy and hard data









## Citations and Links

C.Liu and Y.M.Wang, *TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings*, Proceedings of the 29th International Conference on Machine Learning(ICML-12)

A.P.Dawid and A.M.Skene, *Maximum Likelihood Estimation of Observer Error-Rates using EM Algorithm*, Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 28, No. 1(1979), pp. 20-28

School of Engineering and Applied Sciences • Institute for Applied Computational Science